ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Συγκριτική Μελέτη Εργαλείων Οπτικής Αναπαράστασης και Ανάλυσης Δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΠΑΝΤΕΛΗ - ΠΑΡΙ ΝΑΤΣΗ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2018

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Συγκριτική Μελέτη Εργαλείων Οπτικής Αναπαράστασης και Ανάλυσης Δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΠΑΝΤΕΛΗ – ΠΑΡΙ ΝΑΤΣΗ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Ιουλίου 2018.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Ασκούνης	Ιωάννης Ψαρράς	Χρυσόστομος Δούκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2018

(Υπογραφή)

.....

ΠΑΝΤΕΛΗΣ – ΠΑΡΙΣ ΝΑΤΣΗΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στη σημερινή αγορά λογισμικού, υπάρχει πληθώρα εργαλείων ανάλυσης δεδομένων και οπτικής αναπαράστασης των αποτελεσμάτων της. Τα εργαλεία αυτά απευθύνονται τόσο σε επιστήμονες ανάλυσης δεδομένων, όσο και σε χρήστες με ελάχιστο προγραμματιστικό ή θεωρητικό υπόβαθρο πάνω στην επεξεργασία και ανάλυση δεδομένων. Κατά συνέπεια, συμβάλλουν καταλυτικά στην ανάπτυξη της επιστήμης των visual analytics, της επιχειρηματικής νοημοσύνης αλλά και τη δυνατότητα αξιοποίησης δεδομένων από εργαζόμενους όπως ερευνητές και δημοσιογράφους, στην εποχή των δεδομένων μεγάλης κλίμακας (Big Data).

Το θέμα της παρούσας διπλωματικής εργασίας, είναι η μελέτη και η σύγκριση εργαλείων ανάλυσης και οπτικής αναπαράστασης δεδομένων.

Σε πρώτο στάδιο κατηγοριοποιούνται ορισμένα εργαλεία και περιγράφεται η λειτουργία τους. Για όσα από αυτά δίνουν τη δυνατότητα διαδραστικής αναπαράστασης δεδομένων, γίνεται μία σύγκριση βασισμένη στα χαρακτηριστικά που τα διακρίνουν. Εκτενώς αναλύονται τα χαρακτηριστικά των πιο δημοφιλών εργαλείων, συγκεκριμένα των Tableau, MS Power BI και Qlik Sense. Για αυτά, γίνεται αξιολόγηση των δυνατοτήτων τους μέσω της υλοποίησης σεναρίων διαφορετικών αναγκών και αντικειμένου. Με βάση τα συμπεράσματα που προκύπτουν από τις συγκρίσεις, γίνεται μία προσπάθεια ανάλυσης της αλληλεπίδρασης μεταξύ του χρήστη και των εργαλείων και προτείνονται πιθανές βελτιώσεις και προσθήκες.

Λέξεις Κλειδιά: <<visual analytics tools,ανάλυση δεδομένων, οπτική αναπαράσταση δεδομένων, Tableau, MS Power BI, Qlik Sense >>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

Nowadays, in the era of Big Data, there is a large variety of visual analytics tools in the software market. These tools are widely used by data scientists, as well as users with minimal theoretical or hands-on experience on data analysis. Thus, they convey a substantial role in the development of visual analytics and business intelligence, as well as the ability of users, such as researchers and journalists, to benefit from data.

The subject of this study is the research and comparison of a series of visual analytics tools.

At first, various such tools are identified and categorized, and their operation is briefly described. Afterwards, we compare the ones that are able to interactively visualize data, based on their features. For the most popular among these tools, namely Tableau, MS Power BI and Qlik Sense we extensively study their features and evaluate their abilities through various scenarios of different purpose. Based on the conclusions that we derive from the above comparisons, a gap analysis between the user's needs and the visual analytics tools abilities is attempted, and possible additions and improvements are proposed.

Keywords: << visual analytics tools, data analysis, data visualization, Tableau, MS Power BI, Qlik Sense >>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

Με αφορμή τη διεκπεραίωση της εργασίας αυτής, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Δημήτρη Ασκούνη για την ευκαιρία που μου έδωσε, καθώς και την Ευμορφία Μπιλίρη για την καθοδήγηση και το αδιάκοπο ενδιαφέρον που έδειξε κατά την εκπόνηση της.

Επίσης, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου, και ιδιαίτερα τη μητέρα μου Εύη, για τη συνεχή τους στήριξη καθ' όλο το διάστημα της φοίτησής μου.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

1 Εισαγωγή

1.1 Οπτική Αναπαράσταση Δεδομένων

Με τον όρο αυτό αναφερόμαστε στην αναπαράσταση δεδομένων, με σχηματική μορφή διαγράμματος ή εικόνας (data visualization). Μέσω της οπτικής αναπαράστασης αποσκοπούμε στην οπτικοποίηση της πληροφορίας. Τα δεδομένα έχουν αξία εφόσον προσφέρουν χρήσιμες πληροφορίες για τους ενδιαφερόμενους, με τρόπο ώστε να γίνονται άμεσα αντιληπτές και κατανοητές.

Η οπτική αναπαράσταση πληροφορίας είναι σαφώς η πιο εύκολα ερμηνεύσιμη για τον άνθρωπο, ως μια διαδικασία αποτυπωμένη στη φύση του, απ' όσο μας επιτρέπει να γνωρίζουμε η προϊστορία. Η χάραξη χαρτών σε λίθινες επιφάνειες ήταν από τις πρώτες προσπάθειες του ανθρώπου για μετάδοση πληροφορίας, πέραν της ομιλίας. Εκτιμάται ότι μεγάλο μέρος (περίπου τα 2/3) των νευρώνων του εγκεφάλου συμμετέχουν στην οπτική επεξεργασία [1]. Το μάτι είναι άμεσα αποκρινόμενο στην πληροφορία – και απαιτεί την ελάχιστη δυνατή επεξεργασία από τον ανθρώπινο νου · εύστοχη αναλογία η διαφορά της ανάγνωσης ενός χάρτη από τη λήψη οδηγιών για ένα μέρος (ακουστική περιγραφή).

Με το χρόνο, ο άνθρωπος χρειάστηκε να μοιράζεται εύκολα την πληροφορία και χρησιμοποίησε κωδικοποίηση – ως επί το πλείστον αριθμητική (ψηφία), ευρέως αποδεκτή, πέρα από όποια διαφορά στη γλώσσα- για την εύκολη διακίνηση της. Η έκταση της γης περιγράφεται με στρέμματα, η απόσταση με χιλιόμετρα, ο πληθυσμός με πολλών ψηφίων νούμερα, τα χρέη με πολλά μηδενικά, η θέση στον πλανήτη με συντεταγμένες. Η κωδικοποίηση αποδείχθηκε ιδιαίτερα χρήσιμη και προσέφερε ακρίβεια και εξοικονόμηση πόρων. Ιδιαίτερα τον περασμένο αιώνα η αύξηση του όγκου πληροφοριών απέδειξε τη χρησιμότητα αυτή, με την υιοθέτηση και περεταίρω κωδικοποιήσεων που να συμβαδίζουν με τους ηλεκτρονικούς υπολογιστές– δυαδική πληροφορία, κώδικες γραμματοσειρών κ.ά. Ωστόσο, η κωδικοποιημένη πληροφορία απαιτεί ανθρώπινη επεξεργασία για να γίνει αντιληπτή. Αναλογιζόμαστε πόσο περίπου είναι και σε τι αντιστοιχεί ένα οικονομικό μέγεθος, την κατανομή μίας στατιστικής μελέτης, τη διαφορά μεταξύ δύο χωρικών μεγεθών. Συνήθως, προσπαθούμε να αντιστοιχίσουμε το νούμερο με την εικόνα που φανταζόμαστε ότι περιγράφει. Αντιλαμβανόμαστε, επομένως, ότι πράγματι μια εικονική αναπαράσταση είναι όσο το δυνατόν πιο κοντά στην ανθρώπινη αντίληψη. Τα διάφορα δεδομένα όταν αποτυπώνονται οπτικά μεταδίδουν πληροφορία «εύπεπτη» για το νου.

Εδώ και μερικές δεκαετίες, τα διαγράμματα αποτελούν μια πολύ συνηθισμένη και ευρέως χρησιμοποιούμενη μέθοδο οπτικοποίησης δεδομένων. Συναντούμε κυρίως ραβδογράμματα, διαγράμματα χρονοσειρών, διαγράμματα «πίτας», ιστογράμματα και διαγράμματα διασποράς για τη διαχείριση ποσοτικών και στατιστικών μεγεθών.

- Τα ραβδογράμματα (bar charts) χρησιμοποιούνται για σύγκριση μεγεθών κατ' επιλογή.
- Το ιστόγραμμα (histogram) συνήθως για μέτρηση μεγεθών/ποσοστών για τις τιμές μιας παραμέτρου.

- Τα διαγράμματα διασποράς (scatter plots) απεικονίζουν τη σχέση μεταξύ δύο μεταβλητών σε πολλαπλές χρονικές περιόδους, πχ συσχέτιση.

- Τα διαγράμματα πίτας (pie charts) χρησιμοποιήθηκαν κυρίως για στατιστική και γενικά για την απεικόνιση του μεριδίου ενός μεγέθους επί του συνόλου.

- Τα γραφήματα χρονοσειρών (line charts) δείχνουν τα μεγέθη και τη μεταβολή τους στον άξονα του χρόνου.

Τα τελευταία χρόνια, με την κυριαρχία φορητών συσκευών με πρόσβαση στο internet στη ζωή των περισσότερων ανθρώπων (τουλάχιστον στις αναπτυγμένες τεχνολογικά χώρες) και την ψηφιοποίηση σε μεγάλο βαθμό των περισσότερων επαγγελματικών εργασιών, το μέγεθος των δεδομένων που χρήζουν διαχείρισης έχει εκτοξευτεί. Εν αναμονεί της πολυσυζητημένης εξάπλωσης του Internet of Things (IoT), δηλαδή σύνδεση κάθε συσκευής με το διαδίκτυο, προβλέπουμε με σιγουριά πως ο ρυθμός αύξησης των παραχθέντων δεδομένων όχι απλά θα διατηρηθεί υψηλός αλλά θα εκτιναχθεί. Εύλογα, επομένως οι πιο trendy λέξεις στο χώρο της τεχνολογίας είναι «Big Data». Η έλευση της τεράστιας μάζας δεδομένων έφερε την οπτική αναπαράσταση στο προσκήνιο, καθιστώντας τη συμβολή της αναγκαία.

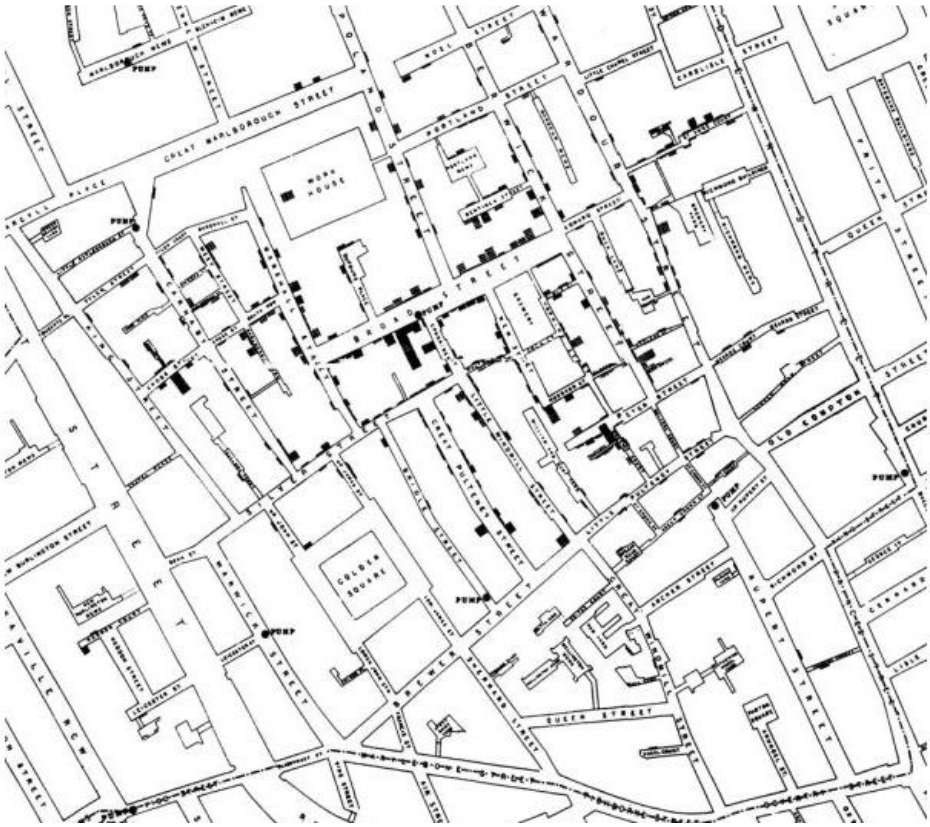
Μέχρι το πρόσφατο παρελθόν, τα δεδομένα που καλούμαστε να ερμηνεύσουμε ήταν συνήθως διαχειρίσιμα από πλευράς μεγέθους, για το σκοπό που χρειαζόταν, π.χ. οικονομικά μεγέθη για επιχειρήσεις και στατιστικές έρευνες. Ωστόσο, τα δεδομένα ήταν υποπολλαπλάσια των σημερινών, αλλά και σε περιπτώσεις μεγάλων datasets προβλεπόταν αρκετός χρόνος επεξεργασίας και συνήθως περισσότεροι υπολογιστικοί πόροι. Όταν, όμως, τα δεδομένα προς επεξεργασία φτάνουν σε ένα μέγεθος που δεν είναι κάποιος σε θέση να έχει ούτε την ελάχιστη εποπτεία και κατανόηση αυτών είναι αναγκαία η οπτική προσέγγιση [2]. Αναλογιζόμαστε πόσο πιο εύπεπτο είναι ένα διάγραμμα σύγκρισης χαρακτηριστικών για μια βάση δεδομένων εκατομμυρίων εγγραφών, πόσο πιο κατανοητό ένα χρονοδιάγραμμα για την εξέλιξη ενός μεγέθους στο χρόνο, πόσο πιο παραστατικός ένας χάρτης με διαβάθμιση χρωμάτων ανά μέγεθος σε σχέση με έναν πίνακα δεδομένων.

Σίγουρα, είναι κοινός τόπος πως η οπτική αναπαράσταση προσφέρει τη γενική εποπτεία ενός συνόλου δεδομένων στιγμιαία. Επιπλέον, είναι δυνατόν εστιάζοντας σε συγκεκριμένα σημεία να έχουμε και καλύτερη αντίληψη για λεπτομερή στοιχεία. Αυτό, όμως, που είναι το μεγαλύτερο προτέρημα είναι η δυνατότητα συγκρίσεων. Η σύγκριση είναι διαδικασία που οδηγεί στην πληρέστερη δυνατή κατανόηση ενός συνόλου δεδομένων, όχι μόνον αυτοτελώς, αλλά και σε ευρύτερα πλαίσια. Υποθέτουμε πως έχουμε τα οικονομικά στοιχεία για το ελληνικό χρέος ανά έτος. Αμέσως, δημιουργώντας ένα διάγραμμα με τα χρέη των κρατών-μελών της ευρωπαϊκής ένωσης αποκτούμε μια σχετικιστική αίσθηση του ποσού με άλλα όμοια. Μπορούμε να τα αποτυπώσουμε σε ένα pie chart, ή σε ένα bar chart, μιας και υποστηρίζεται ότι ο άνθρωπος έχει καλύτερη αντίληψη μήκους απ' ότι επιφάνειας. Έπειτα μπορούμε να αναπαραστήσουμε το ΑΕΠ της χώρας αντιπαραβαλλόμενο με το σημερινό χρέος σε ένα διάγραμμα «treemap» - μέθοδο απεικόνισης ιεραρχιών μεγέθους, με εμφωλευμένα ορθογώνια, για να αναπαραστήσουμε επιφανειακά τι μέγεθος του χρέους καλύπτει. Ένας γεωγραφικός χάρτης με χρωματική διαβάθμιση για το ΑΕΠ που παράγει κάθε περιφέρεια και ένα bar chart με το παραγόμενο ΑΕΠ ανά εργασιακό κλάδο, θα έδινε άμεση πληροφορία για τις πηγές εσόδων. Ένα line chart θα έδειχνε τις μεταβολές στην αξία του στην επιθυμητή χρονική κλίμακα, με χρωματικές διακρίσεις που θα αντιστοιχούν σε κάθε κυβέρνηση. Επίσης, σε άλλο treemap θα απεικονίζαμε τις ετήσιες πολεμικές δαπάνες μιας

υπερδύναμης σε σχέση με το ελληνικό χρέος. Δεδομένο είναι ότι το ελληνικό χρέος είναι 319,739,888,684 €, όταν γράφεται αυτό το κείμενο. Κάποιος γνωρίζοντας απλά αυτό τον αριθμό δεν κατανοεί κάτι για την οικονομική κατάσταση της χώρας. Μαθαίνοντας ότι το γαλλικό χρέος είναι 2,155,947,940,964 € έχει ένα μέτρο σύγκρισης, και την πληροφορία ότι το ελληνικό αντιστοιχεί στο 14.8 % του γαλλικού χρέους. Όταν όμως βλέπει σε ένα γράφημα στοιχειοθετημένα μήκη, ενστικτωδώς κατανοεί τη διαφορά τους – παρακάμπτουμε τη διαδικασία αριθμητικής κωδικοποίησης και δίνουμε την ήδη επεξεργασμένη πληροφορία στο μάτι, τη σημαντικότερη πηγή εισόδου για το νου, χωρίς να χρειάζεται διερμηνεία.

Παραθέτοντας και τα διάφορα γραφήματα για τα ΑΕΠ και τις δαπάνες , σε σύγκριση με το χρέος, μπορούμε όχι απλά να αντιληφθούμε τα μεγέθη και τις συγκρίσεις, αλλά και αίτια, άστοχους χειρισμούς και πιθανούς τρόπους ελάφρυνσης. Η πληροφορία των γραφημάτων αυτών θα μπορούσε αλλιώς να μεταδοθεί μέσω πλήρους κειμένου – σίγουρα όχι μόνο με αριθμούς- σε κάποιες εκατοντάδες λέξεις που θα χρειαζόταν χρόνο και φαιά ουσία για την κατανόηση τους . «Καν’ το εικόνα!» . Η εικόνα είναι αυτό που προσπαθεί να δημιουργήσει ο άνθρωπος με τη φαντασία όταν διαβάζει φράσεις και αριθμούς. Με την οπτική αναπαράσταση παρακάμπτουμε αυτή τη διαδικασία – δίνουμε άμεσα το αποτέλεσμά της, και μάλιστα όσο το δυνατόν πιο συμπυκνωμένο. Ο «data to ink» ratio [3] (λόγος δεδομένων προς μελάνι) είναι ο μέγιστος δυνατός. Κατά την οπτικοποίηση των δεδομένων στόχος είναι η ελαχιστοποίηση του θορύβου, των περιττών στοιχείων και λεπτομερειών, αλλά και της πολυπλοκότητας. Εάν μια απλή εικόνα αντιστοιχεί σε χίλιες λέξεις, τότε εύκολα προκύπτει αναλογικά σε πόσες αντιστοιχούν κάποια συγκοινωνούντα, δυναμικά διαγράμματα.

Το ίδιο ευκολότερα κατανοητά γίνονται και τα γεωγραφικά και χωρικά δεδομένα που αναπαρίστανται σε διαγράμματα. Με την απεικόνιση σε χάρτες, είναι δυνατή η αξιοποίηση των πληροφοριών και η εξαγωγή συμπερασμάτων, στα οποία δεν θα μπορούσε κάποιος να καταλήξει υπό διαφορετικό πρίσμα. Χαρακτηριστικότατο παράδειγμα ο προσδιορισμός των αιτίων της χολέρας [4] από το γιατρό John Snow (Εικόνα 1-1) και η ανακάλυψη των μικροβίων από τον γιατρό Ignaz Semmelweis .



Εικόνα 1-1 Χάρτης του Λονδίνου με έντονα σκιασμένα τα μέρη που εμφανίστηκε η επιδημία της χολέρας

Σημαντικό, όσον αφορά τη διαδικασία οπτικής αναπαράστασης, είναι ότι το αποτέλεσμά της αφορά – και θα συνεχίσει τουλάχιστον στο κοντινό μέλλον – τον ανθρώπινο τρόπο σκέψης και όχι κάποιον τεχνητό. Η υπολογιστική μηχανή έχει ως θεμελιώδη λίθο το ψηφίο – ό,τι δεδομένο δέχεται το μετατρέπει σε αυτό. Ο άνθρωπος λειτουργεί αντίστροφα ως προς το δίπολο εικόνα – ψηφίο. Είναι στη φύση του να σκέπτεται με εργαλείο την εικόνα. Η τεχνητή νοημοσύνη και η μηχανική μάθηση, που αποτελούν ίσως το επίκεντρο της έρευνας του σύγχρονου τεχνολογικού τομέα, έχουν πολλά να προσφέρουν και στη διαδικασία της οπτικοποίησης των δεδομένων Η κατηγοριοποίηση, η παλινδρόμηση και το clustering μπορούν να φανούν χρήσιμα, όμως δε φανερώνουν κάποια δυνατότητα ερμηνείας εικόνων. Δεν θα περιμέναμε ποτέ από κάποιο εργαλείο τεχνητής νοημοσύνης να ερμηνεύσει ένα διάγραμμα (δε χρειάζεται αφετέρου, επεξεργάζεται πολύ πιο εύκολα τους όγκους αριθμών και συμβόλων) και να εξάγει συμπεράσματα. Εφόσον η οπτική αναπαράσταση δεδομένων (data visualization) είναι ένα εργαλείο αποκλειστικά απευθυνόμενο στον άνθρωπο και δύναται να αξιοποιηθεί αποκλειστικά από αυτόν, είναι σημαντικό να συνεχίσει να αναπτύσσεται. Απαλλάσσοντας μας από τη διαδικασία αποκωδικοποίησης και σύνθεσης δεδομένων, εξασφαλίζοντας μας πλείστο χρόνο και αποκαλύπτοντας μας δυσδιάκριτες πληροφορίες ανοίγει το δρόμο για νέες ιδέες, λύσεις και προτάσεις. Φορώντας τους φακούς του data visualization για να παρατηρήσουμε τα δεδομένα, βρισκόμαστε σε μια πολύ πλεονεκτική θέση!

1.2 Οργάνωση της οπτικής αναπαράστασης

Τα δεδομένα που καλούμαστε να χρησιμοποιήσουμε και να αναπαραστήσουμε διακρίνονται σε δύο κατηγορίες ˙ τα ποσοτικά δεδομένα (quantitative data) και τα κατηγορικά δεδομένα (categorical data). Τα ποσοτικά είναι αριθμοί, και μπορούν να εκτελεστούν με αυτά πράξεις και υπολογισμοί. Τα κατηγορικά είναι «ετικέτες», ή απλά πληροφορία επεξηγηματική ως προς τα ποσοτικά μεγέθη. Ο διαχωρισμός των κατηγορικών έχει ως εξής: τα ονομαστικά (nominal), που αναφέρονται π.χ. σε χώρες, εταιρίες και πρόσωπα, τα ordinal που αναφέρονται σε μεγέθη που υπάρχει κάποια καθορισμένη σειρά, π.χ. μήνες, ημέρες, και τέλος interval που αντιστοιχούν σε τιμές, οπότε είναι απολύτως ταξινομήσιμα.

Συνήθως τα πιο δημοφιλή διαγράμματα αποτελούνται από έναν άξονα κάθε κατηγορίας. Γνωστή εξαίρεση το scatter plot , που έχει και στους δύο άξονες ποσοτικά δεδομένα.

Οι πιο κοινές σχέσεις ποσοτικών δεδομένων είναι οι εξής [5] :

- **Χρονοσειρές**

Οι ποσοτικές αξίες εκφράζονται ως ακολουθίες τιμών ληφθεισών ανά ίσα χρονικά διαστήματα. Η χρονική συχνότητα κυμαίνεται συνήθως από έτη μέχρι ώρες. Χαρακτηριστικό είναι ότι στις επιχειρήσεις, το 75% των διαγραμμάτων αφορούν χρονοσειρές.

- **Σχέσεις Κατάταξης**

Εύκολη σύγκριση ποσοτικών μεγεθών με διαδοχική παράταξη. Ειδικά τα διαγράμματα μήκους, όπως τα bar charts, είναι τα πλέον κατάλληλα για συγκρίσεις. Η διαφορά μήκους είναι εύπεπτη για το ανθρώπινο μυαλό.

- **Μέρος προς Όλον**

Οι ποσοτικές αξίες εμφανίζονται ως τιμή τμήματος επί ενός συνόλου (ποσοστιαία), π.χ. τα έξοδα ανά τμήμα σε μια επιχείρηση.

- **Σχέσεις Απόκλισης**

Οι ποσοτικές τιμές παρουσιάζονται έτσι ώστε να αναδειχτεί το πώς διαφέρει μια τιμή επί του συνόλου, π.χ. απόκλιση εξόδων από τον προϋπολογισμό.

- **Σχέσεις Κατανομής**

Εδώ το σύνολο των ποσοτικών τιμών εκτείνεται σε όλο το εύρος, σε μια κατανομή. Μπορούμε από το γράφημα να διαπιστώσουμε εάν υπάρχει κλίση προς κάποια πλευρά, κενά, συγκέντρωση ή συμμετρία.

- **Σχέσεις Συνδιακύμανσης (Correlation)**

Σύγκριση μεταξύ δύο μεταβλητών, για να προσδιορίσουμε εάν κινούνται στην ίδια ή διαφορετική κατεύθυνση. Κυρίως χρησιμοποιούμε scatter plots.

- **Ονομαστική Σύγκριση**

Εδώ απλά παραθέτουμε τα μεγέθη ονομαστικών κατηγορικών στοιχείων που δε σχετίζονται μεταξύ τους, π.χ. τους πληθυσμούς των χωρών.

- **Γεωγραφικά**

Σύγκριση μεγεθών σε γεωγραφικό επίπεδο, ανά χώρες, περιοχές, ταχ. κώδικες**.**

Συνήθως, χρησιμοποιούνται 4 τύποι σχημάτων για την αναπαράσταση των ποσοτικών μεγεθών στα διαγράμματα.

1. Σημεία (points)

Τα σημεία είναι δυνατόν να λάβουν τη μορφή οποιουδήποτε σχήματος, όπως ευθείες, παραλληλόγραμμα, τρίγωνα και οτιδήποτε άλλο. Μπορούν να αναπαραστήσουν ποσοτικές τιμές σε διάγραμμα δύο ποσοτικών αξόνων ταυτόχρονα (scatter plot). Η θέση τους στο χώρο και η αποτύπωσή τους είναι ευέλικτη, για κάθε τύπο άξονα, που έχει διαφορετική διαμόρφωση στο εύρος του (π.χ. ξεκινά από το μηδέν ή άλλον αριθμό). Η χρήση τους είναι ιδανική όταν θέλουμε να δώσουμε έμφαση στην αναπαράσταση μεμονωμένων τιμών και τη διάταξή τους.

2. Ευθείες (lines)

Χρησιμοποιούνται για interval scale δεδομένα (κλίμακας καθορισμένων διαστημάτων), κυρίως χρονικών. Σε ονομαστικές και διατεταγμένες κλίμακες δεν έχει νόημα η χρήση τους. Καθιστούν εμφανή τα μοτίβα , τις τάσεις , τις μεταβολές και τις εξαιρέσεις , κατά μήκος του άξονα.

3. Ράβδοι (Bars)

Κατάλληλες για την ανάδειξη μεμονωμένων τιμών, κυρίως για συγκρίσεις. Οι δύο διαστάσεις τους προσδίδουν άνεση στην όψη, και η σύγκριση γίνεται στη διάσταση του μήκους, όπου είναι και πιο εύκολη για τον άνθρωπο. Μπορούν να τοποθετούνται κάθετα ή και οριζόντια σε ένα διάγραμμα. Κατά τη χρήση τους, η ποσοτική κλίμακα πρέπει να ξεκινά από το μηδέν, ώστε η σύγκριση να είναι ρεαλιστική επί του συνολικού μεγέθους.

4. Boxes

Και οι δύο ακμές του «κουτιού» αντιπροσωπεύουν τιμές. Γνωστός τύπος είναι και το box-and-whisker plot, κουτί που εμφανίζεται το 25% της τιμής του στην κάτω ακμή, το 75 % στην άνω, και με σημεία το 0% και το 100% .

Σαν γενική αρχή, όταν προσπαθούμε να μεταβιβάσουμε πληροφορίες, θεωρούμε πως δεν πρέπει να παραθέτουμε περιττά στοιχεία. Ό,τι υπάρχει σε μια παρουσίαση πρέπει είτε να αναπαράγει είτε να επεξηγεί τις επιθυμητές πληροφορίες.

Πώς, λοιπόν, καταφέρνουμε αυτή τη σωστή μεταβίβαση; Αρχικά πρέπει να αποφασίσουμε αν χρειάζεται γράφημα, εάν αρκούν πίνακες ή χρειάζεται συνδυασμός αμφοτέρων. Ένας πίνακας είναι ιδανικός για να παραθέσουμε συγκεκριμένες, ακριβείς τιμές, οργανωμένες ανά πεδία και εγγραφές. Δεν μπορούμε όμως να αποκομίσουμε τίποτα ως προς τη γενική εικόνα ή να κάνουμε συγκρίσεις και να προσεγγίσουμε σχετικιστικά τα μεγέθη, εξαιρουμένων μόνον πολύ μικρών συνόλων δεδομένων. Επίσης, μπορούμε να χρησιμοποιήσουμε πίνακες συμπληρωματικά των γραφημάτων, συνδυάζοντας τα πλεονεκτήματά τους, εάν ωφελεί.

Μεγάλη σημασία έχει τι επιλέγουμε να τοποθετήσουμε στους άξονες των διαγραμμάτων. Συνηθίζεται π.χ. τα χρονικά μεγέθη να τοποθετούνται στον οριζόντιο άξονα, ή όταν έχουμε πολλές τιμές σε ένα bar chart, τις τοποθετούμε στον κάθετο άξονα και τους αριθμούς στον οριζόντιο. Χρησιμοποιούμε, δηλαδή, τέτοια διάταξη που να είναι όσο το δυνατόν πιο κατανοητή και εύκολα παρατηρήσιμη. Όσον αφορά την κλίμακα των αξόνων, την επιλέγουμε σύμφωνα με τα διαστήματα που θέλουμε να εστιάσουμε και το επίπεδο της λεπτομέρειας. Εάν, π.χ., έχουμε να αναπαραστήσουμε 3 μεταβλητές, πιο αποτελεσματικό είναι να χρησιμοποιήσουμε χρώμα σε ένα διάγραμμα αντί τρισδιάστατων διαγραμμάτων. Καταλήγουμε ότι πρέπει να έχουμε καθορισμένα:

- Το μήνυμα και τα δεδομένα που θέλουμε να μεταδώσουμε
- Τον τρόπο και το μέσο που θα το χρησιμοποιήσουμε γι' αυτό, και αν θα χρειαστούμε εικονική αναπαράσταση δεδομένων
- Εάν ναι, τι γραφήματα θα επιλέξουμε
- Που θα τοποθετηθεί η κάθε μεταβλητή – τιμή
- Τη διαμόρφωση του γραφήματος , όσον αφορά άξονες, κλίμακα, χρώματα και μεγέθη
- Τα επεξηγηματικά κείμενα, ετικέτες, επισημασμένα σημεία που τυχόν θα χρειαστούν, για να γίνει πιο εύκολα αντιληπτό το μήνυμα.

1.3 Visual Analytics

Τα visual analytics ορίζονται ως η επιστήμη του αναλυτικού συλλογισμού, βασισμένη σε διαδραστικές οπτικές διεπαφές [6]. Ο αυξανόμενος ρυθμός παραγωγής δεδομένων είναι μεγαλύτερος από τη δυνατότητα ανάλυσης αυτών, με τις κλασσικές μεθόδους. Τα visual analytics μπορούν να αντιμετωπίσουν προβλήματα που το μέγεθος, η πολυπλοκότητα και η ανάγκη τόσο για ανθρώπινη όσο και μηχανική ανάλυση τα καθιστούν, ειδάλλως, δυσεπίλυτα.

Η «ερευνητική agenda» των visual analytics εμπεριέχει διάφορες επιστημονικές και τεχνολογικές κοινότητες, όπως επιστήμη των υπολογιστών, διαδραστικό και γραφικό σχεδιασμό, γνωστική επιστήμη (cognitive & perpetual sciences), κοινωνικές επιστήμες καθώς και την επιστήμη της οπτικοποιήσης πληροφοριών (information visualization). Όσον αφορά την οπτικοποίηση πληροφοριών, τα όρια με τα visual analytics είναι δυσδιάκριτα, σε βαθμό που οι όροι πλέον τείνουν να ταυτιστούν. Η επιστήμη οπτικοποιήσης πληροφοριών (information visualization) έχει προκύψει από την έρευνα στην επιστήμη των υπολογιστών, στην αλληλεπίδραση ανθρώπου – μηχανής, την ψυχολογία, στις μεθόδους των επιχειρήσεων και τις γραφικές αναπαραστάσεις. Ορίζεται ως η μελέτη της οπτικής αναπαράστασης μεγάλης κλίμακας μεγέθους πληροφοριών, αριθμητικών και μη. Από τα visual analytics και την επιστήμη οπτικοποίησης πληροφοριών στοχεύουμε στην εξαγωγή συμπερασμάτων, βασιζόμενοι στην ανθρώπινη λογική και αντιληπτική ικανότητα, και όχι σε κάποια σύνθετη μέθοδο. Διαφορετικά, όταν στοχεύουμε σε αναπαράσταση φυσικών σχημάτων και γεωμετρικών δομών, κινούμαστε στα πλαίσια της επιστημονικής οπτικοποίησης (scientific visualization), π.χ. σε MRI, στη ροή του ανέμου, βαρομετρικό πεδίο κ.α.

Μέσω των οπτικών αναλυτικών στοιχείων, οι ανθρώπινες γνωστικές ικανότητες μπορούν να ενισχυθούν ως εξής:

- Αύξηση των γνωστικών πόρων, όπως χρήση οπτικών μέσων για υποβοήθηση της λειτουργίας της ανθρώπινης μνήμης.
- Μείωση των υπεράριθμων αναζητήσεων, με την αναπαράσταση μεγάλου όγκου δεδομένου σε μικρές επιφάνειες.
- Ενίσχυση της αναγνώρισης προτύπων, π.χ. όταν οι πληροφορίες οργανώνονται σε διαστήματα ανάλογα με τις χρονικές σχέσεις.
- Επιτρέπουν την εύκολη εξαγωγή συμπερασμάτων κάποιων σχέσεων, που είναι, ειδάλλως, δύσκολο να επαχθούν.
- Κατάλληλες προβολές για αντίληψη μεγάλου αριθμού πιθανών γεγονότων.
- Με την παροχή ενός εύχρηστου μέσου, που σε αντίθεση με τα στατικά διαγράμματα, επιτρέπει την εξερεύνηση του χώρου των παραμετρικών τιμών.

Η διαδικασία της παραγωγής των οπτικών αναλυτικών στοιχείων έχει ως εξής [7] :

Τα σύνολα δεδομένων για τα οπτικά αναλυτικά στοιχεία προέρχονται από ετερογενείς πηγές δεδομένων (αρχεία, βάσεις δεδομένων, ιστοσελίδες). Από τις πηγές αυτές επιλέγονται τα σύνολα δεδομένων $S = S_1, ..., S_m$ όπου το κάθε ένα αποτελείται από διάφορα χαρακτηριστικά $A_{i1}, ..., A_{ik}$. Ο σκοπός ή η έξοδος της διαδικασίας είναι η πληροφορία I .

Η συνάρτηση παραγωγής οπτικών αναλυτικών στοιχείων εκφράζεται από το μετασχηματισμό $F: S \rightarrow I$, όπου $f \in \{D_W, V_X, H_Y, U_Z\}$. Η D_W είναι η συνάρτηση προεπεξεργασίας των δεδομένων $D_W : S \rightarrow S$, που εκφράζεται από μια αλληλουχία συναρτήσεων, $W \in \{T, C, SL, I\}$, που αφορούν το μετασχηματισμό, τον καθαρισμό, την επιλογή και την ενσωμάτωση των δεδομένων αντίστοιχα. Συνήθως η συνάρτηση προεπεξεργασίας ορίζεται ως εξής $D_w = D_T(D_I(D_C(S_1, ..., S_n)))$. Η προεπεξεργασία είναι απαραίτητη για να γίνουν εφαρμόσιμες οι συναρτήσεις ανάλυσης των δεδομένων.

Μετά το στάδιο της επεξεργασίας δεδομένων, η ζητούμενη πληροφορία μπορεί να προσεγγίζεται μέσω κάποιας υπόθεσης H της αυτοματοποιημένης μεθόδου ανάλυσης. Η υπόθεση H_y ,Y $\in \{S, V\}$, μπορεί να παράγεται από τα δεδομένα $H_S : S \rightarrow H$, με εφαρμογή στατιστικών συναρτήσεων, ανάλυσης μέσω μηχανικής μάθησης κ.α., ή από δημιουργηθείσα οπτική αναπαράσταση $H_V : V \rightarrow H$. Η ζητούμενη πληροφορία προσεγγίζεται, φυσικά, και με οπτικές αναπαραστάσεις V_W, $W \in \{S, H\}$. Η οπτική αναπαράσταση απεικονίζει δεδομένα $V_S : S \rightarrow V$ ή κάποια υπόθεση $V_H : H \rightarrow V$.

1.4 Εργαλεία Οπτικής Αναπαράστασης και Ανάλυσης Δεδομένων

Η «εισβολή» των δεδομένων στο χώρο των επιχειρήσεων όπως και η εκθετική αύξηση του όγκου προς διαχείριση στον ακαδημαϊκό τομέα δημιούργησαν νέες ανάγκες. Η ταχύτητα στην επεξεργασία και στην αναπαράσταση των δεδομένων είναι σίγουρα μία από αυτές. Η γρήγορη και εμπρόθεσμη διεκπεραίωση είναι ζωτικής σημασίας για τους αναλυτές δεδομένων, τα τμήματα ανάλυσης και οικονομικών των επιχειρήσεων και τα ερευνητικά

προγράμματα. Ανάγκη αποτελεί σίγουρα και η ενασχόληση πολύ περισσότερων εργαζομένων και ερευνητών με τα δεδομένα. Δηλαδή εργαζόμενοι σε τμήματα όπως μάρκετινγκ, οικονομικά, προμήθειες και ερευνητές σε διαφορετικά αντικείμενα, συχνά άσχετα με την τεχνολογία, καλούνται να επεξεργαστούν δεδομένα και να δημιουργήσουν γραφήματα.

Οι παραπάνω ανάγκες έφεραν στο προσκήνιο τα εργαλεία ανάλυσης δεδομένων και οπτικών αναπαραστάσεων. Τα εργαλεία αυτά δίνουν τη δυνατότητα για γρήγορη αποτύπωση του μηνύματος των δεδομένων, χωρίς τη χρήση κάποιας γλώσσας προγραμματισμού και χωρίς να είναι απαιτούμενη η γνώση στατιστικής και ανάλυσης δεδομένων.

Τα εργαλεία αποτελούν «όπλο στη φαρέτρα» και για τους αναλυτές δεδομένων (data analysts), όπως είναι αναμενόμενο. Για να είναι χρήσιμο το εργαλείο στον αναλυτή, πρέπει πρωτίστως να παρέχει αποτελεσματικά διαγράμματα, δηλαδή «οπτικά εύπεπτα» όσον αφορά το σχήμα, τα χρώματα και τη διαβάθμιση [8]. Το ίδιο αποτελεσματικό βέβαια πρέπει να είναι όσον αφορά την αλληλεπίδραση (επεξεργασία) με τα δεδομένα. Συγκεκριμένα, να έχει εύκολη και πλήρη διαδικασία φιλτραρίσματος, ταξινόμησης και σχολιασμού – επεξήγησης. Τα δύο αυτά χαρακτηριστικά είναι τα σημαντικότερα για κάθε εργαλείο. Όσον αφορά την εισαγωγή των αρχείων και το διάβασμα των δεδομένων, ο αναλυτής έχει ανάγκη από εύχρηστο data extraction, cleansing, transformation και loading (ECTL). Η παροχή σωστών στατιστικών μεγεθών, προβλέψεων και τάσεων (trend lines) είναι επίσης σημαντική, συμπληρώνοντας και επεξηγώντας την πληροφορία των διαγραμμάτων, παρ' ότι δεν είναι ο βασικός σκοπός του εργαλείου.

Τα εργαλεία προσφέρουν λοιπόν πολλές διευκολύνσεις. Ωστόσο, εάν έχουμε κατά νου συγκεκριμένα διαγράμματα δεν είναι απαραίτητα, αφού μπορούμε, με περισσότερο κόπο, γνώση και χρόνο, να τα σχεδιάσουμε με μια γλώσσα (όπως R,Python). Είναι όμως απολύτως απαραίτητα στον αναλυτή δίνοντας τη δυνατότητα για Εξερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis -EDA). Δεν είναι δυνατή η EDA όταν ο αναλυτής πρέπει να ασχοληθεί με τα «μηχανικά» μιας γλώσσας ή προγράμματος, βιβλιοθήκες, εντολές, διαχείριση μνήμης, αρχείων δεδομένων κτλ. Η προσοχή του πρέπει να είναι απολύτως εστιασμένη στα διαγράμματα, την πληροφορία που μεταβιβάζουν και την ερμηνεία τους. Αναλογιζόμενοι ότι αυτή η διαδικασία χρειάζεται για πολλά γραφήματα και σύνολα δεδομένων, γίνεται κατανοητό πώς δεν είναι πρακτικά δυνατό να πειραματιστείς και να τα επεξηγήσεις όταν απαιτείται να συντάσσεις εντολές για το κάθε ένα. Η διαδικασία πρέπει να είναι άμεση για να τελεσφορήσει η ερευνητική ανάλυση των δεδομένων.

Και κάτι που δεν αφορά μόνον τους αναλυτές και τους επαγγελματίες: η ταχύτητα και κυρίως η ευκολία, που επιτρέπει στους χρήστες την αναπαράσταση των δεδομένων υπό μορφή σχημάτων, έφερε μία επανάσταση στον τομέα της πληροφορίας. Δεδομένου ότι πάρα πολλά σύνολα δεδομένων είναι πλέον διαθέσιμα, ακόμα και από επίσημους φορείς και οργανισμούς, ο κάθε ένας μπορεί πλέον να τους δώσει αξία και να βγάλει συμπεράσματα. «Κατακτά» ο ίδιος μία πληροφορία – δεν είναι αποδέκτης των, συχνά αβάσιμων, συμπερασμάτων τρίτων. Δεν είναι υπερβολή, λοιπόν, να αποδώσουμε τον «εκδημοκρατισμό των δεδομένων» στα εργαλεία επεξεργασίας και αναπαράστασης.

Πολλά από τα εργαλεία οπτικής αναπαράστασης απευθύνονται πρωτίστως σε επιχειρήσεις και έχουν σχεδιαστεί για να καλύπτουν τις ανάγκες τους στην οργάνωση και την ερμηνεία των δεδομένων. Αξίζει, επομένως, να αναφερθούμε και στην «Επιχειρηματική Νοημοσύνη»

(Business Intelligence), που πολλά από τα εργαλεία αποτελούν μέρος της, και όλα έχουν στοιχεία εμπνευσμένα από αυτή.

1.5 Επιχειρηματική Νοημοσύνη

Η επιχειρηματική νοημοσύνη περιλαμβάνει τεχνολογίες και εφαρμογές για τη συλλογή, την οργάνωση, την ανάλυση και την παρουσίαση των δεδομένων που αφορούν την επιχείρηση. Θα μπορούσε να χαρακτηριστεί ως η εξέλιξη των συστημάτων λήψης αποφάσεων, συνυφασμένη πλέον με τη χρήση υπολογιστικών πόρων. Με τη βοήθειά της στοχεύουμε στη βελτίωση της διαδικασίας λήψης αποφάσεων, βασισμένης σε δεδομένα και γεγονότα.

Τα αναλυτικά στοιχεία επιχειρήσεων (business analytics) [9] αποτελούνται από τρεις υποκατηγορίες, τα περιγραφικά, τα προγνωστικά και τα «συμβουλευτικά» αναλυτικά στοιχεία . Τα περιγραφικά αναλυτικά (descriptive analytics) στοιχεία βασίζονται σε δεδομένα του παρόντος και του παρελθόντος. Προσδίδουν διορατικότητα πάνω σε αυτά, μέσω των reports, των γραφημάτων, του clustering και άλλων μεθόδων. Τα προγνωστικά αναλυτικά στοιχεία (predictive analytics), αφορούν μοντέλα προβλέψεων, με τη χρήση στατιστικών μεγεθών και διαφόρων τεχνικών μηχανικής μάθησης. Τα συμβουλευτικά αναλυτικά στοιχεία (prescriptive analytics), χρησιμοποιούνται για να προταθούν αποφάσεις, που μπορούν να λάβουν οι ενδιαφερόμενοι. Διαδικασίες όπως η βελτιστοποίηση και η προσομοίωση είναι απαραίτητες για την εξαγωγή των στοιχείων αυτών. Γενικά, τα prescriptive analytics συμπεριλαμβάνουν μηχανική μάθηση, επιχειρησιακή έρευνα, εφαρμοσμένη στατιστική, επεξεργασία φυσικής γλώσσας, επεξεργασία σημάτων και εικόνας, και μετα-ευρετικών.

Η επιχειρησιακή νοημοσύνη «εξάγει» κυρίως περιγραφικά αναλυτικά στοιχεία, μιας και εστιάζει στη δημιουργία αναφορών (reports) που αξιοποιούν τα δεδομένα, διερμηνεύουν γεγονότα και παραθέτουν λεπτομερείς πληροφορίες. Ωστόσο, τελευταία, φαίνεται όλο και περισσότερο η τάση των BI tools να ενσωματώνουν predictive και prescriptive analytics και να καλύπτουν το πεδίο της υποστήριξης αποφάσεων μέσω προγνωστικών – κάτι που αφορούσε άλλα analytical tools στο παρελθόν. Η εξέλιξη αυτή δημιουργήθηκε από την ανάγκη των επιχειρήσεων να συνδυάζουν πλέον τις πληροφορίες και τα insights (διορατικότητα) των reports, με προβλέψεις και προσομοιώσεις ώστε να έχουν όσο το δυνατόν πιο ολοκληρωμένη υποστήριξη στις αποφάσεις που λαμβάνουν. Χαρακτηριστικό είναι ότι πλέον γίνεται λόγος για τα **augmented analytics**, όπου εμφανίζεται μια σειρά από ερωτήσεις και απαντήσεις-reports, απαλλάσσοντας τον ενδιαφερόμενο από την όλη αυτή διαδικασία. Ο χρήστης των εργαλείων θα έχει απλά να επιλέξει ποιες από τις ερωτήσεις τον ενδιαφέρουν.

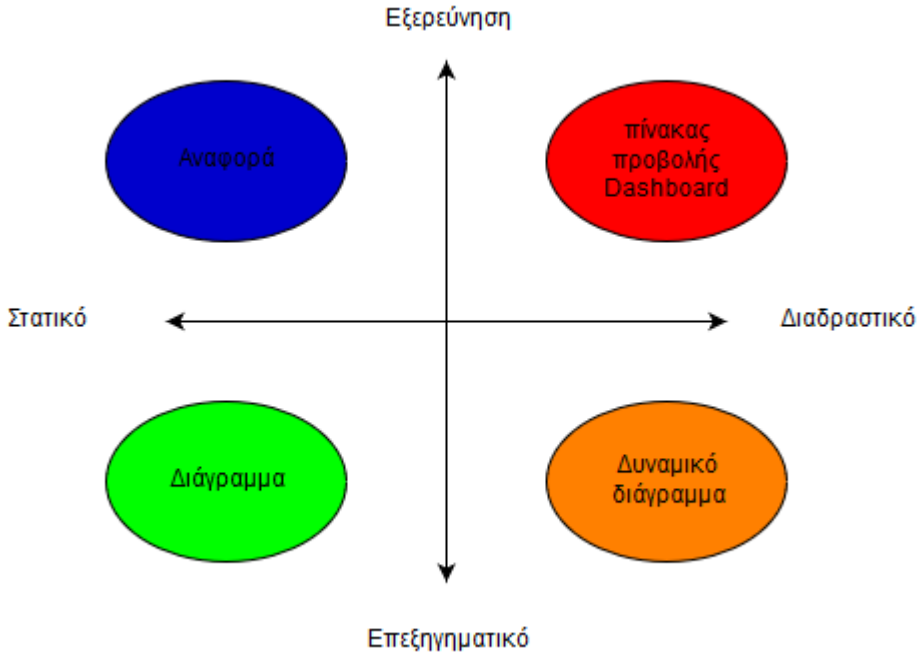
Είναι κατάδηλο πως τα εργαλεία που πρωταγωνιστούν στο χώρο του BI, κάνουν το ίδιο και στον τομέα του data visualization. Εδώ πρέπει να γίνει ξεκάθαρο ότι αναφερόμαστε σε BI εργαλεία που εξάγουν αναλυτικά στοιχεία μέσω οπτικών αναπαραστάσεων, και όχι εργαλεία οργάνωσης της διοίκησης, των λογιστικών και της αλληλογραφίας. Αναδεικνύεται λοιπόν το πόσο σημαντικό είναι για την επιχειρησιακή νοημοσύνη η ακριβής, αποτελεσματική και εύκολη οπτική αναπαράσταση.

1.6 Οργάνωση της μελέτης

Η Εργασία αυτή είναι οργανωμένη σε τέσσερα κεφάλαια. Το πρώτο κεφάλαιο αποτελεί την εισαγωγή στο θέμα της εργασίας. Στο δεύτερο, αφού δοθεί μία τυπική οργάνωση των εργαλείων οπτικής αναπαράστασης και ανάλυσης δεδομένων και γίνει μία κατηγοριοποίηση αυτών, περιγράφονται συνοπτικά κάποια εργαλεία και σχολιάζονται οι βασικές λειτουργίες τους. Για όσα από αυτά η οπτικοποίηση των δεδομένων είναι προτεραιότητα και συμβαδίζει με τη μη εξειδικευμένη χρήση, τα βασικά χαρακτηριστικά τους παρατίθενται σε έναν κοινό πίνακα. Έπειτα, για τα επικρατέστερα από αυτά γίνεται μια αναλυτική περιγραφή των δυνατοτήτων τους. Στο τρίτο κεφάλαιο, γίνεται μία «hands-on» σύγκριση πάνω στα εργαλεία αυτά. Η αποτελεσματικότητά τους και οι λύσεις που δίνουν, ελέγχονται μέσω της εκτέλεσης κάποιων κοινών σεναρίων, τα οποία αφορούν χρονοσειρές, αριθμητικά δεδομένα μεγάλου μεγέθους και γεωγραφικά δεδομένα. Στο τέταρτο κεφάλαιο παρατίθενται τα συμπεράσματα της σύγκρισης μεταξύ των επικρατέστερων εργαλείων, των διαφορών που εντοπίστηκαν στην οπτική αναπαράσταση δεδομένων μέσω εργαλείων και μέσω εντολών. Καταλήγοντας, εντοπίζονται ορισμένα περιθώρια βελτίωσης, και προτείνονται μελλοντικές λειτουργίες που θα αναβάθμιζαν τα εργαλεία.

2 Εργαλεία Οπτικής Αναπαράστασης και Ανάλυσης Δεδομένων

Κατά την εισαγωγή, γίνεται αναφορά στη χρησιμότητα των εργαλείων, στους χρήστες που προσελκύουν, και σε κάποιες από τις λειτουργίες τους. Δεδομένου όμως ότι εργαλεία σχεδιασμού διαγραμμάτων και επεξεργασίας δεδομένων υπάρχουν εδώ και δεκαετίες, με πιο γνωστό και επιτυχημένο το Ms Excel, πρέπει να γίνει εμφανής η ειδοποιός διαφορά τους με τη νέα γενιά εργαλείων οπτικής αναπαράστασης δεδομένων. Αυτή έγκειται στη δυνατότητα εξερευνητικής ανάλυσης των δεδομένων και στη δυνατότητα εξαγωγής οπτικών αναλυτικών στοιχείων. Στα «παραδοσιακά» εργαλεία, η διαδικασία δημιουργίας γραφημάτων είναι σταδιακή, δηλαδή ρυθμίζονται τα δεδομένα και οι όποιες παράμετροι σε κατάλληλο παράθυρο, και έπειτα στο «φύλλο» εργασίας εμφανίζεται το διάγραμμα. Στα εργαλεία οπτικής αναπαράστασης ο σχεδιασμός και η προβολή του γραφήματος γίνονται παράλληλα, επιτρέποντας στο χρήστη να κάνει αλλαγές και να προσαρμόζει τις παραμέτρους, ανάλογα με αυτά που παρατηρεί στο παράθυρο εργασίας. Συνεπώς, ο σχεδιασμός, δεν είναι αποκομμένος από την εποπτεία και η διαμόρφωση του διαγράμματος γίνεται με τρόπο διαδραστικό, δεν αποφασίζεται εκ των προτέρων.



Εικόνα 2-1 – Τρόποι έκθεσης των αποτελεσμάτων της επεξεργασίας δεδομένων

Η εξερευνητική ανάλυση λαμβάνει χώρα και στην προβολή των διαγραμμάτων. Τα διαγράμματα δεν είναι στατικά, αλλά αλληλοεπιδρούν μεταξύ τους, με την εμφάνιση των κοινών τους στοιχείων, τη «ζωντανή» διαδικασία φιλτραρίσματος σε ένα κλικ, την εστίαση σε δεδομένα που εμφανίζουν περισσότερο ενδιαφέρον και φυσικά περισσότερες κατηγορίες γραφικών, όπως χάρτες (Εικόνα 2-1). Με τα κλασσικά εργαλεία (τύπου Excel), η εξερευνητική ανάλυση γίνεται με αναφορές στατικών διαγραμμάτων, και, συνεπώς, είναι πολύ περιορισμένη, και εξ αρχής οριοθετημένη σε ότι επιλέγει ο σχεδιαστής των διαγραμμάτων. Συν τοις άλλοις, από τα εργαλεία οπτικής αναπαράστασης απαιτείται να

καλύπτουν τις ανάγκες που εμφανίζονται σήμερα στον τομέα διαχείρισης των δεδομένων. Βέβαια, όπως γνωστοποιήθηκε και θα διαπιστωθεί και στη συνέχεια, αυτοί είναι οι κύριοι άξονες διαφοροποίησης.

2.1 Τυπική οργάνωση λειτουργιών των εργαλείων

Όπως προαναφέρθηκε, τα εργαλεία περιλαμβάνουν διάφορες λειτουργίες, παράλληλα με την κεντρική της οπτικής αναπαράστασης. Ο βασικός κορμός της οργάνωσης των λειτουργιών είναι:

- Εισαγωγή του αρχείου δεδομένων ή σύνδεση με την πηγή δεδομένων
- «Καθαρισμός» (cleaning-cleansing), δηλαδή προσαρμογή των δεδομένων σε κατάλληλη προς χρήση μορφή
- Μετασχηματισμοί των δεδομένων
- «Φόρτωση» των δεδομένων, για τη διαδικασία της ανάλυσης
- Οπτική Αναπαράσταση, με τη δημιουργία διαγραμμάτων και υπολογισμό αναλυτικών στοιχείων
- Παρουσίαση των visual analytics

Στο στάδιο της εισαγωγής, το εργαλείο εισάγει ένα αρχείο όταν δεν αναμένονται μεταβολές σε αυτό, ή δημιουργεί μια σύνδεση με το αρχείο-πηγή για την ανανέωση των δεδομένων, σε αντίθεση περίπτωση. Συνήθως, τα αρχεία-πηγές είναι βάσεις δεδομένων. Σημαντικό λοιπόν είναι η εισαγωγή να γίνει σωστά, να διαβαστούν δηλαδή όλα τα στοιχεία και τα μετα-δεδομένα (metadata), αλλά και οι συνδέσεις να γίνονται με τρόπο που να μην επιβραδύνουν τη λειτουργία του εργαλείου και να μην θέτουν σε κίνδυνο την πηγή, με αναίτιες παρεμβάσεις στα δεδομένα της.

Ο καθαρισμός των δεδομένων ξεκινά με την αναγνώριση των τύπων των πεδίων δεδομένων, όπως ακέραιοι, πραγματικοί, ακολουθίες χαρακτήρων κ.τ.λ. και τη σωστή αναπαράστασή τους (όπως στίξη του διαχωρισμού δεκαδικών ψηφίων). Έπειτα γίνεται αντιστοίχιση σε κατηγορίες δεδομένων, διαδικασία που αφορά κυρίως τα γεωγραφικά στοιχεία, όπως συντεταγμένες, ονόματα περιοχών και ταχυδρομικοί κώδικες. Κατά τον καθαρισμό γίνεται και η διαχείριση των κενών τιμών (null values) και η αναπαράστασή τους σύμφωνα με το πώς είναι προδιαγεγραμμένο στο πρόγραμμα. Μπορούν, επίσης, να εντοπιστούν και να αναφερθούν τυχόν σφάλματα, όπως διαφορετικός τύπος δεδομένου, κάποιου στοιχείου, από τα υπόλοιπα σε ένα πεδίο. Η εισαγωγή και ο καθαρισμός είναι διαδικασίες μεγάλης σημασίας, γιατί όσα καλά και να είναι τα αναλυτικά στοιχεία που δημιουργούνται είναι άχρηστα, ή ακόμα χειρότερα παραπλανητικά, όταν βασίζονται σε λάθος δεδομένα.

Οι μετασχηματισμοί στα δεδομένα αφορούν αλλαγές και προσθήκες που ο χρήστης επιθυμεί. Μερικές από αυτές είναι η περιστροφή στηλών σε γραμμές, η συγχώνευση ή σύνδεση πινάκων, η δημιουργία και η προσθήκη νέου πεδίου και ο διαχωρισμός πεδίων. Στο σημείο αυτό είναι σημαντικό να αναφέρουμε πως τα δεδομένα είναι οργανωμένα, κατά κανόνα, σε δομή πίνακα.

Στο στάδιο της φόρτωσης, τα δεδομένα καθίστανται έτοιμα για την ανάλυση και τη δημιουργία διαγραμμάτων. Συνήθως το βήμα αυτό δεν είναι φανερό στο χρήστη. Θεωρώντας

το cleaning μέρος των μετασχηματισμών, τα τέσσερα στάδια αυτά αποτελούν την «ETL» διαδικασία (Extract- Transform- Load).

Στο στάδιο της οπτικοποίησης, δημιουργούνται τα κατάλληλα διαγράμματα και παράλληλα υπολογίζονται αναλυτικά στοιχεία που αφορούν αριθμητικά μεγέθη και ίσως στατιστικά μεγέθη, προβλέψεις, κ.α. Ο περιορισμός των πεδίων και των τιμών που εμφανίζονται (φιλτράρισμα), όπως και οι ρυθμίσεις εμφάνισης είναι μέρη της διαδικασίας.

Το σύνολο των διαγραμμάτων και πληροφοριών, οργανώνονται σε «αναφορές» (reports), όπου τα γραφήματα έχουν την τελική, μορφή με την οποία θα εκτεθούν. Σε πολλά εργαλεία υπάρχουν επιλογές για την κατάλληλη διαμόρφωση των αναφορών και των διαγραμμάτων, όπως προσθήκη σχολίων κειμένου και συνημμένων αρχείων, με σκοπό την παρουσίαση σε τρίτους.

2.2 Κατηγοριοποίηση

Σε όλο το εύρος της μελέτης, εξετάζουμε εργαλεία που βασίζονται, ή τουλάχιστον εμπεριέχουν, την προαναφερθείσα τυπική οργάνωση. Η ενστικτώδης δημιουργία διαγραμμάτων και η διαδραστικότητα αυτών είναι το κύριο χαρακτηριστικό που τα διακρίνει. Ωστόσο, τα εργαλεία διαφέρουν στον τρόπο επεξεργασίας των δεδομένων, στην ποιότητα των διαγραμμάτων και στην πληρότητα παρουσίασης τους, και φυσικά στις δυνατότητες που εμπεριέχουν. Με βάση τις λειτουργίες στις οποίες εστιάζουν και τη φιλοσοφία οργάνωσής τους ακολουθεί η κατηγοριοποίηση τους.

Αριθμητικών και στατιστικών συναρτήσεων

Στα εργαλεία που εστιάζουν στα αριθμητικά αναλυτικά στοιχεία πρωταρχικό ρόλο έχουν τα στατιστικά μεγέθη. Για στατιστικά όπως η συσχέτιση και η αυτοσυσχέτιση, και η ανάλυση διακύμανσης παρέχονται ως έτοιμες επιλογές. Το ίδιο ισχύει και για στατιστικά μοντέλα όπως αναδρομική ανάλυση (π.χ. μέθοδος ελαχίστων τετραγώνων). Συνήθως, τα εργαλεία αυτά μπορούν να κάνουν και προβλέψεις (forecasting) με βάση τις δεδομένες τιμές, με τη χρήση συναρτήσεων όπως την εκθετική εξομάλυνση. Επίσης, η χάραξη ευθειών τάσης με διαστήματα εμπιστοσύνης είναι δυνατότητα που συνήθως υποστηρίζουν.

Επιστήμης δεδομένων (με χρήση δυνατοτήτων μηχανικής μάθησης)

Ορισμένα εργαλεία υποστηρίζουν τη χρήση μηχανικής μάθησης για την εξαγωγή αναλυτικών στοιχείων. Είναι λοιπόν δυνατά το clustering, η κατηγοριοποίηση και η ανάδραση. Συγκεκριμένα, ο χρήστης μπορεί να εκμεταλλευτεί τις έτοιμες επιλογές, μεταξύ των οποίων είναι τα δέντρα αποφάσεων (decision trees) με σύνολα εκπαίδευσης, βαθμολογητών για επιβλεπόμενη μάθηση (scorers), δυνατότητα επεξεργασίας κειμένου και προβλέψεων μέσω μηχανικής μάθησης (όχι απλά με συναρτήσεις).

Μεγάλου όγκου δεδομένων

Η κατάκλιση του πεδίου ανάλυσης δεδομένων από τους «μεγάλους όγκους» καθιστά αδύνατη την επεξεργασία κάποιων συνόλων δεδομένων από υπολογιστικούς πόρους τυπικών δυνατοτήτων. Τα εργαλεία που εστιάζουν στη διαχείριση των «Big Data», μπορούν να

επεξεργαστούν μεγάλου μεγέθους σύνολα περιορίζοντας τη χρήση πόρων όπως μνήμη και χώρο στο δίσκο και ισχύ επεξεργαστή. Σημαντική είναι και η καλή σύνδεση με τις μεγάλες βάσεις δεδομένων, ώστε να γίνεται απρόσκοπτα η λήψη τους .

Ειδικής κατηγορίας διαγραμμάτων

Υπάρχουν εργαλεία που είτε εξειδικεύονται είτε παρέχουν αποκλειστικά επιλογές για την οπτικοποίηση γεωγραφικών δεδομένων μέσα από χάρτες διαφορετικών τύπων και εμφάνισης. Αντίστοιχα, υπάρχουν πολλά εργαλεία που εστιάζουν σε άλλους τύπους γραφημάτων, όπως γράφους και δίκτυα.

Απευθυνόμενα σε αναλυτές και προγραμματιστές

Όπως έγινε εμφανές, τα εξεταζόμενα εργαλεία δεν απαιτούν για τη χρήση τους καμία γνώση προγραμματισμού. Ωστόσο, ορισμένα από αυτά αναπτύχθηκαν αρχικά, στις πρώτες εκδόσεις τους, απευθυνόμενα σε αναλυτές, προγραμματιστές και επαγγελματίες στο χώρο της πληροφορικής. Απότοκο αυτής της προέλευσης είναι η παροχή πάρα πολλών επιπλέον δυνατοτήτων για εξειδικευμένα διαγράμματα, υπολογισμό στατιστικών και επεμβάσεων στην εμφάνιση, μέσω της χρήσης συναρτήσεων από διάφορες βιβλιοθήκες, είτε ευρέως γνωστών γλωσσών, είτε γλωσσών ανεπτυγμένων από το εργαλείο.

Όλα τα εργαλεία που εξετάζουμε, έχουν ως σημείο αναφοράς την οπτική αναπαράσταση των δεδομένων με διαδραστικό τρόπο και τη δημιουργία διαγραμμάτων χωρίς την ανάγκη της χρήσης εντολών και την επεξεργασία δεδομένων- είναι δηλαδή visual analytics tools. Ένα εργαλείο μπορεί, φυσικά, να μην ανήκει αποκλειστικά σε μία από τις παραπάνω κατηγορίες. Μάλιστα, σπάνια υπάρχει κάποιο να καλύπτεται αποκλειστικά από μία.

2.3 Περιγραφή Εργαλείων

Στην ενότητα αυτή, γίνεται σύντομη περιγραφή ορισμένων ευρέως χρησιμοποιούμενων visual analytics εργαλείων, εξαιρουμένων των 3 εμπορικά δημοφιλέστερων Tableau, MS Power BI και QlikSense, για τα οποία ακολουθεί εκτενής ανάλυση των δυνατοτήτων και της λειτουργίας τους στην ενότητα 2.5. Για κάθε εργαλείο, επισημαίνονται τα ιδιαίτερα χαρακτηριστικά του, η πιθανή κατηγοριοποίηση του και οι δυνατότητες του.

Tibco Sportfire

Ένα εργαλείο [10] που δίνει έμφαση στα αναλυτικά στοιχεία, ακόμα και κατά τη δημιουργία των διαγραμμάτων. Εστιάζει στα προγνωστικά αναλυτικά στοιχεία, στην ανάλυση κέρδους, στους δείκτες τάσεων, και στην ανάλυση γεγονότων (event analytics). Παρέχει αυτοματοποιημένες υπηρεσίες, όπως προκαθορισμένη αποστολή e-mail. Έχει ενσωματωμένες τις συναρτήσεις από R, Matlab, S+ και SAS. Τα διαγράμματα είναι διαδραστικά στα πλαίσια των αναφορών αλλά το περιβάλλον του πολύ απλουστευμένο και παρωχημένο.

Plot.ly

Το plot.ly [11] είναι ένα open source εργαλείο, με πραγματικά πολλές και, κυρίως, εντυπωσιακές επιλογές διαγραμμάτων και πολύ καλή υποστήριξη γλωσσών προγραμματισμού. Εκτός από τους βασικούς/ συνηθισμένους τύπους διαγραμμάτων,

συμπεριλαμβάνει τρισδιάστατα γραφήματα, όπως επιφάνειες και scatter plots, τρισδιάστατους χάρτες, ειδικά διαγράμματα στατιστικών μεγεθών κ.α., αντίστοιχα των matplotlib και ggplot2. Οι δυνατότητες κατάλληλης διαμόρφωσης των διαγραμμάτων είναι σίγουρα επαρκείς, αφού επιτρέπουν την παρέμβαση σε άξονες, χρώματα, φόντο και σημειώσεις. «Δυνατό σημείο» του εργαλείου είναι, επίσης, η ανάλυση των δεδομένων. Βασικά στατιστικά μεγέθη, "best fit" ευθείες, ανάλυση διακύμανσης (ANOVA) , στατιστικό τεστ δείκτη χι εις το τετράγωνο (chi-squared test) είναι μερικά από αυτά. Είναι απόλυτα συμβατό με Python, R, D3.js βιβλιοθήκες (επίσης Matlab, Arduino, NodeJS), αλλά και απόλυτα διαχειρίσιμο από το γραφικό περιβάλλον του , χωρίς κώδικα. Μετά την εισαγωγή τους, τα δεδομένα εμφανίζονται υπό τη μορφή λογιστικού φύλλου (spreadsheet), με δυνατή επεξεργασία αντίστοιχη του excel (πλήρης μεν, αλλά όχι τόσο φιλική στο χρήστη- αναχρονιστική). Επίσης, είναι συμβατό με ποικίλους τύπους αρχείων. Τα προτερήματα του εργαλείου είναι πράγματι εντυπωσιακά, όσον αφορά το κομμάτι της οπτικής αναπαράστασης και των υπολογισμών. Ωστόσο, η απουσία δυνατότητας αυτόματης σύνδεσης με διάφορες βάσεις δεδομένων (connectors), η περιορισμένη διαδραστικότητα, και κυρίως η αδυναμία διαχείρισης και καθαρισμού αρχείων και δεδομένων μεγάλου όγκου στη «Big Data» εποχή, είναι σημαντικά μειονεκτήματα.

Kibana

Το Kibana [12] είναι μία από τις τρεις εφαρμογές (υπο-εργαλεία) του elastic stack (ή ELK Stack), που καλύπτει μόνο το κομμάτι της οπτικής αναπαράστασης δεδομένων. Η εισαγωγή των δεδομένων γίνεται μέσω του Logstash, εφαρμογής όπου λαμβάνει χώρα η εισαγωγή- ο μετασχηματισμός- και η «φόρτωση» των δεδομένων (ETL). Η τρίτη εφαρμογή, elasticsearch, έχει το ρόλο μηχανής αναζήτησης δεδομένων και υπολογισμού αναλυτικών στοιχείων (analytics). Επομένως, θα συγκρίνουμε ολιστικά τις δυνατότητες που προσφέρει το εργαλείο (Πίνακας 2-1), εστιάζοντας, όμως, στο Kibana.

Το Kibana, open source (ανοικτού λογισμικού) όπως όλο το ELK Stack, περιλαμβάνει όλους τους βασικούς τύπους διαγραμμάτων, γράφους και διάφορες δυνατότητες, όπως forecasting, με τη συμβολή της μηχανικής μάθησης. Η οργάνωσή του είναι απλή, με τα «παράθυρα» Discover (εποπτεία δεδομένων), Visualize (οπτικοποιήση), Dashboard (πίνακας παρουσίασης) και Settings (ρυθμίσεις). Το φιλτράρισμα των δεδομένων είναι εύχρηστο, με άμεση επιλογή «include» και «exclude», τόσο για τα ονοματικά πεδία όσο και τις ποσοτικές τιμές. Υπάρχει επιλογή για εξειδικευμένες συναρτήσεις συνόλων (aggregation functions), ως JSON (input), από βιβλιοθήκες. Το Kibana αναπτύχθηκε αρχικά ως εργαλείο για άτομα στο χώρο της ανάπτυξης και χρήσης του λογισμικού και επαγγελματίες στο χώρο της πληροφορικής (DevOps και IT Ops). Παρ' όλο που πλέον έχει γενική απεύθυνση, η φιλοσοφία του το καθιστά δύσκολο προς εκμάθηση από απλούς χρήστες. Στα διαγράμματα, επίσης, δεν είναι δυνατό να συμπεριλάβουμε περισσότερα του ενός ονοματικά πεδία ως διαστάσεις. Η επεξεργασία των δεδομένων δεν καλύπτει πολλές ανάγκες (όπως Pivot, Joins) και, τέλος, η αλλαγή μεταξύ τριών εργαλείων για την επίτευξη των επιθυμητών αποτελεσμάτων σίγουρα δεν διευκολύνει το χρήστη.

Chartio

Το Chartio [13] είναι ένα ολοκληρωμένο διαδικτυακό πρόγραμμα που προσφέρει πάρα πολλές επιλογές για την επεξεργασία των δεδομένων και χειρίζεται σύνθετα αιτήματα πολύ καλά, με βελτιστοποίηση σε γλώσσα SQL. Μπορεί να συνδεθεί με πολλές πηγές δεδομένων,

με ρυθμιζόμενη ή χειροκίνητη ανανέωση, αλλά όχι «ζωντανή» σύνδεση. Προσφέρει μάλιστα και, εκτός από την απλή, κρυπτογραφημένη σύνδεση με την πηγή, μέσω SSH tunnel. Για ό,τι δημιουργεί διαδραστικά ο χρήστης στο εργαλείο, μπορεί να δει τις SQL εντολές που αντιστοιχούν, ή να δημιουργήσει εξ αρχής με κώδικα SQL. Γενικά τα διαθέσιμα διαγράμματα είναι χρηστικά και επαρκή, και οι αριθμητικές και στατιστικές συναρτήσεις επίσης πολλές και επαρκείς. Το εργαλείο αξιοποιεί επίσης τη μηχανική μάθηση. Ωστόσο, το γραφικό περιβάλλον του εργαλείου είναι πολύ βασικό, και δεν προσφέρει καλή ποιότητα οπτικής αναπαράστασης. Επίσης, δεν είναι κατάλληλο για άπειρους και νέους χρήστες, έχοντας δύσκολη διαδικασία μάθησης.

Sisense

Το Sisense [14] είναι ένα εργαλείο πλήρων δυνατοτήτων σε διαδραστικά διαγράμματα, συναρτήσεις και αναλυτικά στοιχεία, και επεξεργασία δεδομένων, με απλουστευμένο όμως περιβάλλον χρήσης. Υποστηρίζει αιτήματα σε φυσική γλώσσα. Διαθέτει μάλιστα και bot που απαντά σε φυσική γλώσσα, σε διάφορες εφαρμογές επικοινωνίας, όπως το Skype. Δεν είναι ιδανικό για νέους χρήστες και η εκμάθησή του απαιτεί, σχετικά, χρόνο.

Domo

Ένα διαδικτυακό εργαλείο [15] (SaaS- λειτουργεί μέσω του browser), που παρέχει μεγάλο αριθμό διαγραμμάτων (85 για την ακρίβεια) αλλά και τη δυνατότητα σύνδεσης με πάρα πολλές πηγές δεδομένων. Έχει παράθυρο για την πλατφόρμα επεξεργασίας δεδομένων, την «ETL Magic» . Μπορεί να αναβαθμιστεί με πρόσθετες εφαρμογές από το «App Store» του εργαλείου, ανάλογα με τις ανάγκες του χρήστη. Δίνει έμφαση στην ασφάλεια των δεδομένων, αλλά και στο διαμοιρασμό των διαγραμμάτων και γενικά, των αναλυτικών στοιχείων - υπάρχει η δυνατότητα «μηνύματος» (internal message), για την άμεση αποστολή διαγραμμάτων σε άλλους χρήστες του προγράμματος. Στο Domo η δημιουργία των διαγραμμάτων δεν είναι ενστικτώδης, και συνεπώς δεν μπορεί να χρησιμοποιηθεί από εντελώς άπειρους χρήστες.

Watson IBM

Το Watson [16], της IBM, βασιζόμενο στην τεχνητή νοημοσύνη, έχει τυπική οργάνωση της διαδρομής δεδομένων: εισαγωγή -> διαμόρφωση -> διερεύνηση -> έκθεση. Για την εισαγωγή, είναι δυνατή η σύνδεση με όλους τους συνήθεις τύπους αρχείων. Αυτό που ξεχωρίζει είναι η δυνατότητα άμεσης διασύνδεσης με μέσα κοινωνικής δικτύωσης και blogs. Η διαμόρφωση των δεδομένων πριν την εξαγωγή των αναλυτικών στοιχείων γίνεται στο παράθυρο «Shape Before». Χρησιμοποιεί τη μηχανική μάθηση και για την εξαγωγή αναλυτικών στοιχείων (π.χ. προβλέψεις), εκτός από την επικοινωνία με το χρήστη. Το API του εργαλείου είναι διαθέσιμο σε σχεδιαστές λογισμικού για την ανάπτυξη εφαρμογών σε τομείς όπως η υγεία και το εμπόριο, π.χ. εικονικός σύμβουλος εργαζομένων στο χώρο της υγεία. Εδώ είναι σημαντικό να υπογραμμιστεί ότι, το εργαλείο πρωτίστως απαντά στις ερωτήσεις φυσικής γλώσσας με διαγράμματα και άλλα αναλυτικά στοιχεία, και δεν προτάσσει τη δημιουργία διαγραμμάτων από τον ίδιο το χρήστη, αν και είναι δυνατή.

Το εργαλείο, είναι ουσιαστικά ένα υπολογιστικό σύστημα ερωταπαντήσεων, που δίνει απαντήσεις και σε φυσική γλώσσα. Πρόεκυψε από το project DeepQA της IBM, για να συμμετάσχει στο τηλεοπτικό παιχνίδι γνώσεων «Jeopardy!» . Εμπίπτει δηλαδή στα πλαίσια αλληλεπίδρασης ανθρώπου - μηχανής. Το Watson κατακερματίζει τις ερωτήσεις, αναλύοντας

τις σε λέξεις κλειδιά για να εντοπίσει στατιστικά συσχετιζόμενες φράσεις. Η καινοτομία του έγκειται στην ικανότητα ταυτόχρονης εκτέλεσης εκατοντάδων αλγορίθμων γλωσσικής ανάλυσης, και όχι σε κάποιον νέο αλγόριθμο.

SAS Visual Analytics

Το εργαλείο της SAS [17] που εστιάζει στην οπτική αναπαράσταση δεδομένων είναι το «Visual Analytics» . Είναι από τα κορυφαία εργαλεία όσον αφορά τη σύνδεση με πηγές δεδομένων, παρέχει καλή επεξεργασία δεδομένων (ETL), μπορεί να διαχειριστεί μεγάλα αρχεία, και έχει πολλές αριθμητικές και στατιστικές συναρτήσεις. Σε συνδυασμό με άλλες εφαρμογές της SAS, παρέχει δυνατότητες όπως η πρόβλεψη, η ομαδοποίηση (clustering) κ.α., επιστρατεύοντας τη μηχανική μάθηση. Επομένως, είναι ένα πολύ «ισχυρό» εργαλείο, που υστερεί όμως στην ποιότητα των διαγραμμάτων, τα οποία φαντάζουν ξεπερασμένα και η διαδραστικότητά τους είναι περιορισμένη.

Pentaho

Το Pentaho [18] αποτελείται από μια σειρά εξειδικευμένων εφαρμογών, για κάθε λειτουργία. Η Data Integration εφαρμογή ονομάζεται «Kettle» και αφορά τις διεργασίες ETL. Η Big Data εστιάζει στην ενσωμάτωση αρχείων μεγάλων δεδομένων . Συνδέεται άμεσα με το Apache Hadoop και υποστηρίζει NoSQL πηγές, όπως το MongoDB. Η εφαρμογή Report Designer αφορά την οπτική αναπαράσταση των δεδομένων και τη δημιουργία αναφορών. Το Data Mining είναι η εφαρμογή που φέρνει το εργαλείο κοντά στο πεδίο της επιστήμης δεδομένων. Με την επιστράτευση του εργαλείου «Weka», παρέχει δυνατότητες της μηχανικής μάθησης όπως προβλέψεις και αναζήτηση μοτίβου στα σύνολα δεδομένων (data mining. Επίσης, υπάρχει εφαρμογή του εργαλείου για το IoT. Εκτός από τα προαναφερθέντα, το Pentaho διαθέτει και προϊόντα που απευθύνονται σε εταιρικό επίπεδο και server εφαρμογές. Τα διαγράμματα του εργαλείου δεν είναι καλής ποιότητας και η διαδραστικότητά τους περιορίζεται στο φιλτράρισμα, χωρίς να έχουν αλληλεπίδραση μεταξύ τους. Ομοίως, και το περιβάλλον χρήσης του εργαλείου είναι παρωχημένο σε σχέση με άλλα εργαλεία.

Ενδιαφέρον παρουσιάζουν και ορισμένα εργαλεία που προσδιορίζονται πιο αυστηρά και τοποθετούνται στο πεδίο της επιστήμης δεδομένων. Σε αυτά υπάρχει η δυνατότητα δημιουργίας διαγραμμάτων διαφορετικού τύπου, χωρίς τη χρήση εντολών. Ωστόσο, το εργαλείο δεν εστιάζει στην ποιότητα, τη διαδραστικότητα και την «μεταδοτικότητα» πληροφορίας των διαγραμμάτων. Κύριο χαρακτηριστικό τους είναι η χρήση της μηχανικής μάθησης για την ανάλυση δεδομένων, την ανάλυση κειμένου τη διενέργεια προβλέψεων αλλά και την προετοιμασία των δεδομένων μετά την εισαγωγή. Όλα αυτά φυσικά μέσω του περιβάλλοντος του εργαλείου, χωρίς να χρειάζονται εντολές. Η ροή εργασιών (workflow) είναι χαρακτηριστική : σε ένα παράθυρο εργασίας τοποθετούνται κόμβοι , οι οποίοι σχηματίζουν τη διαδρομή των δεδομένων. Ο κόμβος μπορεί να αφορά τη διαδικασία εισαγωγής δεδομένων, την επεξεργασία τους, εφαρμογές μηχανικής μάθησης όπως δέντρα αποφάσεων, βαθμολογητές για επιβλεπόμενη μάθηση, και την εξαγωγή των αποτελεσμάτων σε άλλα αρχεία.

Αντιπροσωπευτικά εργαλεία αυτής της κατηγορίας είναι το **Rapidminer** [19] και το **Knime** [20]. Αμφότερα είναι υλοποιημένα σε γλώσσα Java. Έχουν πολλές δυνατότητες για την προετοιμασία δεδομένων, την ανάλυσή τους, σε πολλές υποκατηγορίες της μηχανικής

μάθησης (deep learning, meta learning), και τη διαχείριση δεδομένων μεγάλου όγκου. Επιδέχονται προεκτάσεις από scripts γραμμένα σε Python, R, Java κ.α. Το Knime, μάλιστα, είναι εργαλείο ανοιχτού κώδικα (open source) και μπορεί να ενσωματώσει επεκτάσεις και άλλα project ανοικτού κώδικα για διάφορες λειτουργίες όπως π.χ. image mining. Τα εργαλεία αυτά συνδυάζονται πολλές φορές με όσα έχουν καλύτερη οπτική αναπαράσταση, για βέλτιστα αποτελέσματα (π.χ. εισαγωγή και επεξεργασία δεδομένων στο Rapidminer, με δέντρο αποφάσεων για ανάλυση συναισθήματος σε κάποιες φράσεις, και οπτική αναπαράσταση στο Tableau).

Εάν τα εργαλεία πιο ενστικτώδους χρήσης και εύκολης μάθησης εντασσόταν σε μία κατηγορία των «εκδημοκρατισμένων εργαλείων», τότε το SAS και το Pentaho βρίσκονται στο μεταίχμιο της επιστήμης των δεδομένων και των «εκδημοκρατισμένων» εργαλείων.

Carto DB

Ένα open source, διαδικτυακό εργαλείο [21] οπτικής αναπαράστασης, που εξειδικεύεται στην αναπαράσταση γεωγραφικών δεδομένων και τη δημιουργία χαρτών. Παρέχονται διάφορες βάσεις χαρτών, που επιδέχονται τροποποιήσεις, και δυνατότητες όπως μέτρηση αποστάσεων και βελτιστοποίηση διαδρομής. Οι χάρτες παραμένουν ενήμεροι, σε πραγματικό χρόνο. Εκτός από χάρτες, στο εργαλείο είναι διαθέσιμα και άλλα διαγράμματα, για την ανάλυση και επεξήγηση των στοιχείων που απεικονίζονται γεωγραφικά. Ως πηγές είναι δεκτά αρχεία όπως xls, csv, ή χωρικών δεδομένων όπως shapefiles.Το εργαλείο μπορεί να διαχειριστεί και αρχεία μεγάλου μεγέθους. Συγκεκριμένα το Carto Builder απευθύνεται σε κάθε χρήστη, χωρίς να απαιτείται τεχνικό υπόβαθρο σε γεωγραφικά πληροφοριακά συστήματα (GIS). Το Carto Engine, ένα σύνολο από APIs και βιβλιοθήκες, απευθύνεται σε σχεδιαστές λογισμικού.

Άλλα γνωστά εργαλεία σχεδιασμένα για την απεικόνιση γεωγραφικών δεδομένων είναι το ArcGIS Esri, που ενσωματώνεται σε κάποιες εκδόσεις του MS Power BI, και το Mapbox.

Gephi

Ενδεικτικά, γίνεται αναφορά και σε ένα εργαλείο εξειδικευμένων διαγραμμάτων, εκτός των γεωγραφικών. Το Gephi [22] εστιάζει στο σχεδιασμό γράφων και δικτύων αποκλειστικά. Παρέχει μια σειρά από κατάλληλες για κοινωνικά δίκτυα στατιστικές συναρτήσεις (betweenness centrality, clustering coefficient), δυναμικό φιλτράρισμα, δυνατότητα δημιουργίας χρονομεταβλητών δικτύων και δυνατότητα εξαγωγής των αποτελεσμάτων σε μορφή εγγράφου.

Επιπρόσθετα, αξίζει να αναφέρουμε απλώς και κάποια εργαλεία που στα οποία είναι απαραίτητη η σύνταξη εντολών, και συνεπώς δεν εξετάζουμε περεταίρω τις δυνατότητες τους. Αυτά απευθύνονται σε προγραμματιστές και αναλυτές δεδομένων, και αφορούν μόνο τα διαγράμματα και όχι τη διαδικασία σύνδεσης με τα δεδομένα. Γνωστά είναι τα **FusionCharts**, High Charts, Google Charts κ.α. Είναι ουσιαστικά βιβλιοθήκες διαγραμμάτων υλοποιημένων σε Javascript, που ο χρήστης τα ενσωματώνει με εντολές σε html και css. Προτιμητέο από πολλούς είναι το εργαλείο **D3.js**, [23] το οποίο είναι μία βιβλιοθήκη γραφικών στοιχείων υλοποιημένων σε Javascript , σχεδιασμένων από χρήστες μιας και είναι ανοιχτού κώδικα. Ο χρήστης μπορεί να συνδυάσει όλα αυτά τα, τεράστιας ποικιλίας, γραφικά μέρη για να δημιουργήσει το διάγραμμα που επιθυμεί. Συνεπώς, δίνει απεριόριστες δυνατότητες σχεδιασμού στο χρήστη, αλλά η εκμάθησή του είναι πολύ δύσκολη.

2.4 Χαρακτηριστικά Σύγκρισης

Έχοντας αποκτήσει μία γενική εικόνα διαφόρων εργαλείων και των δυνατοτήτων που παρέχουν, μπορούμε πλέον να επικεντρωθούμε σε όσα έχουν ως προτεραιότητα την οπτική αναπαράσταση των δεδομένων. Για την καλύτερη αξιολόγησή τους, θα συγκρίνουμε τα εργαλεία σε έναν πίνακα (Πίνακας 2-1), ως προς κάποια βασικά χαρακτηριστικά, που συναντούμε στα περισσότερα και είναι σημαντικά για τον μέσο χρήστη.

Εκδόσεις

Αρχικά, οι εκδόσεις (versions) ενός εργαλείου μπορεί να είναι οι εξής:

- Υπολογιστή (desktop)
- Κινητή (mobile)
- Software-as-a-Service (cloud)

Η έκδοση υπολογιστή αφορά τη χρήση του εργαλείου από έναν χρήστη – ή από κάθε χρήστη ξεχωριστά όταν αφορά περισσότερα πακέτα. Η κινητή έκδοση αφορά φορητές συσκευές, και συνήθως είναι συμπληρωματική των άλλων. Χρησιμεύει, κυρίως, για προβολή και μικρές παρεμβάσεις σε reports. Η SaaS, είναι μια cloud έκδοση, όπου το εργαλείο αποθηκεύει τα δεδομένα του σε απομακρυσμένους πόρους μνήμης και εκτελείται στο browser.

Προετοιμασία δεδομένων

Το κομμάτι της προετοιμασίας δεδομένων (data preparation), έχει μεγάλη σημασία στα εργαλεία αναπαράστασης, ίσως ίση με αυτό της γραφικής αναπαράστασης. Η μορφή των δεδομένων προς επεξεργασία, όπως αυτά λαμβάνονται από το χρήστη, είναι , συνήθως, σε ακατάλληλη προς επεξεργασία μορφή. Επομένως, σύμφωνα με το ETL «πρωτόκολλο», ελέγχουμε τα εξής χαρακτηριστικά :

Extract

Συμβατότητα με διάφορες πηγές δεδομένων

Ιδιαίτερη διευκόλυνση αποτελεί η δυνατότητα για άμεση και αυτοματοποιημένη εξαγωγή δεδομένων από ποικίλες βάσεις, αρχεία και ιστοσελίδες. Αποφεύγεται έτσι μία χρονοβόρα διαδικασία που είναι η δημιουργία κατάλληλης σύνδεσης για κάθε τύπο πηγής.

Δυνατότητα ταυτόχρονης λήψης δεδομένων από διαφορετικές πηγές

Είναι, τις περισσότερες φορές, αναγκαίο να αντλούνται σύνολα δεδομένων από περισσοτέρων της μίας πηγής, ώστε να συγκρίνουμε και να αναπαραστήσουμε τα επιθυμητά στοιχεία.

Transform

Προσθήκη δεδομένων από το χρήστη

Δυνατότητα προσθήκης νέων δεδομένων χειροκίνητα, όπως π.χ. κάποιον μικρό πίνακα. Διευκολύνει το χρήστη, ώστε να μη χρειαστεί να χρησιμοποιήσει άλλη εφαρμογή, και έπειτα να συνδέσει το αρχείο ως πηγή.

Αλλαγή τύπου δεδομένων

Το εργαλείο προσφέρει τη δυνατότητα αλλαγής ανάμεσα σε ακέραιο, δεκαδικό αριθμό, σειρά χαρακτήρων (string), γεωγραφικού ρόλου κ.α. Η αναγνώριση τύπου γίνεται συνήθως αυτόματα, αλλά συχνά δεν είναι η επιθυμητή .

Ταξινόμηση

Η ταξινόμηση, αύξουσα ή φθίνουσα, μπορεί να γίνει σε όποια κατηγορία – στήλη είναι χρήσιμη. Μπορεί να δώσει μια κατατοπιστική εικόνα στο χρήστη κατά την επεξεργασία των δεδομένων, όσον αφορά τα μεγέθη, ώστε να κάνει εύστοχες οπτικές αναπαραστάσεις. Επίσης, προσφέρει εύκολο έλεγχο της μέγιστης και ελάχιστης τιμής.

Περιστρεφόμενος Πίνακας (Pivot Table)

Η επιθυμητή γραφική απεικόνιση των δεδομένων, κάποιες φορές, προϋποθέτει τη μεταβολή του πίνακα. Γενικά, η διαδικασία της περιστροφής πίνακα «συστήθηκε» από το Lotus Improv [24], το 1991, και έγινε ευρέως γνωστή από τις εκδόσεις του MS Excel που ακολούθησαν. Πολλά εργαλεία προσφέρουν μία περιορισμένη εκδοχή του pivot table, όπως, π.χ. τις στήλες σε γραμμές (αφότου δημιουργηθεί και μια ακόμα στήλη προσδιορισμού).

Region	2014	2015	2016
North	500	450	150
East	150	300	225
South	325	300	375
West	200	200	150
Central	300	200	250

Abc Pivot Pivot Field Names	# Pivot Pivot Field Values	Abc Sheet1 Region
2014	150	South
2014	325	East
2014	200	West
2014	300	Central
2015	450	North
2015	300	South
2015	300	East
2015	200	West
2015	200	Central
2016	150	North

Επομένως, κατά τη σύγκριση των εργαλείων, θα επισημαίνεται εάν το υποστηρίζεται πλήρως ή περιορισμένα η περιστροφή πίνακα.

Συνδέσεις και Ενώσεις

Στους διάφορους πίνακες των συνόλων δεδομένων, πιθανόν να υπάρχουν κοινά πεδία. Τα κοινά πεδία θα είναι κατά κανόνα (σε μία βάση δεδομένων) οι στήλες, αφού στις γραμμές βρίσκονται οι εγγραφές. Όταν θα δημιουργήσουμε μία σύνδεση θα επιλέξουμε το πεδίο της τομής στα δύο σύνολα. Η σύνδεση μπορεί να είναι κλασσική τομή (inner join) σύμφωνα με

τη θεωρία συνόλων, εξ αριστερών σύνδεση (left outer join), όπου υπάρχουν όλα τα στοιχεία του αριστερού συνόλου και όσα εκ του δεξιού είναι κοινά, αντίστοιχη εκ δεξιών σύνδεση (right outer join), και πλήρης σύνδεση με όλα τα στοιχεία των συνόλων, κατάλληλα αντιστοιχισμένων (full outer join).Όταν υπάρχουν κενές/μηδενικές (null) εγγραφές, μπορούμε να επιλέξουμε εάν θα συμπεριληφθούν ή όχι στη σύνδεση.Η αντιστοίχιση γίνεται για τα κοινά πεδία, ενώ εμφανίζονται και τα μη κοινά, σε κάθε εγγραφή.

Table 1

ID	First Name	Last Name	Publisher Type
20034	Adam	Davis	Independent
20165	Ashley	Garcia	Big
20233	Susan	Nguyen	Small/medium

Table 2

Book Title	Price	Royalty	ID
Weather in the Alps	19.99	5,000	20165
My Physics	8.99	3,500	20800
The Magic Shoe Lace	15.99	7,000	20034

Join στο πεδίο ID

ID	First Name	Last Name	Publisher Type	Book Title	Price	Royalty
20034	Adam	Davis	Independent	The Magic Shoe Lace	15.99	7,000
20165	Ashley	Garcia	Big	Weather in the Alps	19.99	5,000

Για την ένωση απαιτούνται μόνο κοινά πεδία στους πίνακες, δηλαδή ίδιες στήλες. Οι εγγραφές- σειρές εμφανίζονται όλες σε έναν κοινό πίνακα. Εάν υπάρχει η ίδια εγγραφή εμφανίζεται μία φορά.

May2016

Day	Customer	Purchases	Type
4	Lane	5	Credit
10	Chris	6	Credit
28	Juan	1	Credit

June2016

Day	Customer	Purchases	Type
1	Lisa	3	Credit
28	Isaac	4	Cash
28	Sam	2	Credit

July2016

Day	Customer	Purchases	Type
2	Mario	2	Credit
15	Wei	1	Cash
21	Jim	7	Cash

Union

Day	Customer	Purchases	Type
4	Lane	5	Credit
10	Chris	6	Credit
28	Juan	1	Credit
1	Lisa	3	Credit
28	Isaac	4	Cash
28	Sam	2	Credit
2	Mario	2	Credit
15	Wei	1	Cash
21	Jim	7	Cash

Διαχωρισμός Στήλης (Split)

Η εισαγωγή των δεδομένων γίνεται ορισμένες φορές μη ιδανικά. Συχνά παρατηρείται διαφορετικά πεδία δεδομένων να βρίσκονται στην ίδια στήλη. Τότε είναι απαραίτητη η διαδικασία διαχωρισμού, σε όσες στήλες είναι επιθυμητό.

Αναζήτηση Στοιχείου

Η δυνατότητα εντοπισμού συγκεκριμένων στοιχείων σε στήλη- πεδίο του πίνακα. Εκτός από την προφανή χρησιμότητά της, δηλαδή όταν εστιάζουμε στην τιμή που αναζητούμε, είναι αποτελεσματική και για τον εντοπισμό σφαλμάτων, λάθος εγγραφών, διπλοτύπων και τον προσδιορισμό του τύπου ενός πεδίου, π.χ. σε ένα μεγάλο σύνολο > 1.000.000 εγγραφών, να αναζητήσουμε εάν υπάρχουν δεκαδικοί αριθμοί.

Loading

Live Loading

Αφού έχει προηγηθεί η διαδικασία εξαγωγής από τα αρχεία και μετατροπής των δεδομένων, απομένει η «φόρτωση» για την οπτικοποιήση. Στο σημείο αυτό ενδιαφέρον έχει να εξετάσουμε τη δυνατότητα για «ζωντανή», άμεση και διαρκή φόρτωση των δεδομένων. Αυτό σημαίνει ότι για κάθε αλλαγή που λαμβάνει χώρα στο αρχείο εισαγωγής, θα ενημερώνονται και τα δεδομένα του εργαλείου, σχεδόν ταυτόχρονα, χωρίς να χρειάζεται ο χρήστης να κάνει κάποια ανανέωση. Ο όρος «ζωντανά» αντιστοιχεί εν προκειμένω σε ελάχιστα δευτερόλεπτα, και η καθυστέρηση έχει να κάνει με την ταχύτητα της σύνδεσης αλλά και την ταχύτητα απόκρισης της πηγής.

Αναλυτικά Στοιχεία

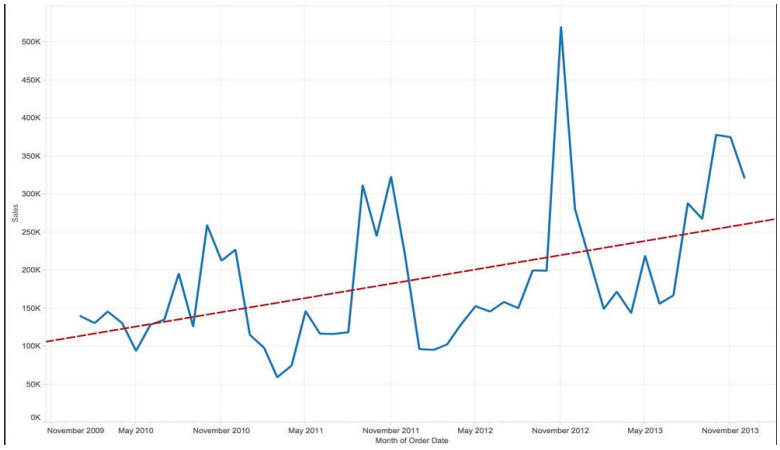
Στατιστικά Στοιχεία

Τα στατιστικά απαιτούνται τόσο για γραφική απεικόνιση, όσο και για την άμεση σύγκριση μεταξύ των πεδίων και των κατηγοριών των δεδομένων. Βασικά στατιστικά μεγέθη, που

γίνεται συχνότατη χρήση τους, είναι το μέγιστο και ελάχιστο, μέσος όρος, διάμεσος, τυπική απόκλιση, διακύμανση και percentile. Μέσω αυτών μπορούν να υπολογιστούν και άλλα στατιστικά μεγέθη, όπως , π.χ., ο συντελεστής διακύμανσης r.

Trend Lines (Ευθείες Τάσης)

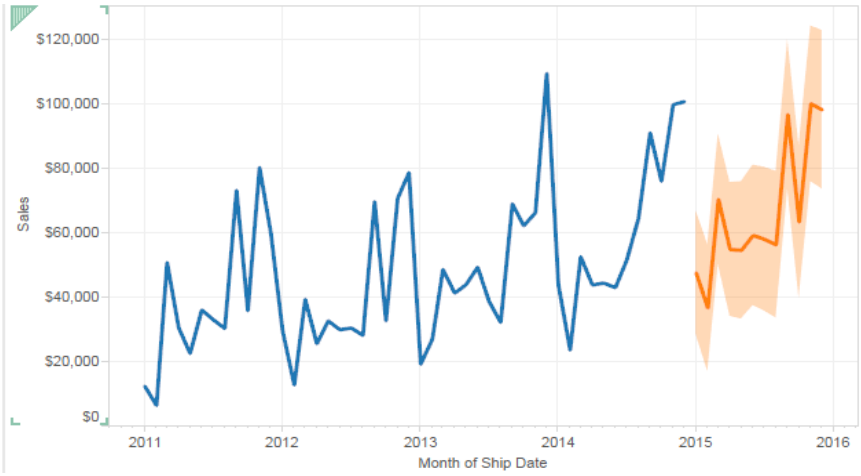
Οι ευθείες τάσης σχεδιάζονται σε χρονοσειρές (διαγράμματα με άξονα χρόνου) για ανάλυση οικονομικών μεγεθών κυρίως, αλλά είναι και χρήσιμες γενικά για τη διεξαγωγή προβλέψεων (Εικόνα 2-2). Μια ευθεία τάσης χαράσσεται μεταξύ δύο σημείων και επαληθεύεται όταν ένα τρίτο σημείο (ή και περισσότερα) την τέμνει. Η ευθεία τάσης μπορεί να είναι γραμμική, λογαριθμική, εκθετική ή πολυωνυμική.



Εικόνα 2-2 – παράδειγμα διαγράμματος με ευθείες τάσης

Forecasting (Πρόβλεψη)

Η πρόβλεψη αφορά τις μελλοντικές τιμές μίας μεταβλητής, για το επιθυμητό χρονικό διάστημα, βασισμένες στις παρελθοντικές τιμές (Εικόνα 2-3). Βέβαια, αναφερόμαστε σε χρονοσειρές, και οι δεδομένες παρατηρήσεις πρέπει να είναι περισσότερες από κάποιο ελάχιστο αριθμό, ανάλογα τη μέθοδο πρόβλεψης. Οι περισσότερες μέθοδοι που ακολουθούν τα εργαλεία, βασίζονται στην εκθετική εξομάλυνση (simple exponential smoothing).



Εικόνα 2-3 – παράδειγμα διαγράμματος με πρόβλεψη

Άλλα χαρακτηριστικά

Γεωγραφικά Διαγράμματα

Εξετάζουμε εάν το εργαλείο υποστηρίζει γεωγραφικά διαγράμματα, όπως χάρτες, χάρτες με διακεκριμένα σημεία, χάρτες με χρωματική διαβάθμιση, με λεπτομέρειες σε δρόμους και οδικό δίκτυο και ό,τι έχει να κάνει με τη γεωγραφική απεικόνιση.

Φίλτρα

Τα φίλτρα των δεδομένων εφαρμόζονται τόσο σε ποσοτικά - αριθμητικά όσο και σε ποιοτικά περιγραφικά δεδομένα. Όσον αφορά τα ποσοτικά μεγέθη, φιλτράρουμε επιλέγοντας άνω ή κάτω όριο, επιτρεπτό διάστημα, ενώ για τα ποιοτικά να περιλαμβάνει, να ξεκινάει ή να καταλήγει με συγκεκριμένη ακολουθία χαρακτήρων. Επίσης , μπορούμε να φιλτράρουμε ως προς την κατάταξη, π.χ. 10 στοιχεία με το μεγαλύτερο άθροισμα, ή κατά συνθήκη, π.χ. μέγιστο αριθμό εγγραφών ανά μεταβλητή και ταυτόχρονα άθροισμα μικρότερο κάποιας προκαθορισμένης τιμής. Τα φίλτρα εφαρμόζονται είτε σε συγκεκριμένο διάγραμμα, σε ένα σύνολο διαγραμμάτων ή σε πίνακα παρουσίασης (dashboard).

Πλαίσιο Παρουσίασης (Dashboard)

Το πλαίσιο παρουσίασης συγκεντρώνει τα όσα επιλεγμένα διαγράμματα, επεξηγήσεις κειμένου και εικόνες κρίνεται ότι συντελούν στην μετάδοση των επιθυμητών πληροφοριών. Στο dashboard τα διαγράμματα αλληλεπιδρούν μεταξύ τους, δηλαδή το φιλτράρισμα και η προβολή συγκεκριμένων στοιχείων έχουν ταυτόχρονη εφαρμογή σε όποια διαγράμματα είναι επιθυμητό. Η παρουσίαση σε ειδικά διαμορφωμένο για το σκοπό αυτό παράθυρο, διαχωρισμένο από αυτό της δημιουργίας επιτρέπει να εστιάσουμε στο τελικό προϊόν και την «ουσία» του.

Πεδία μέσω εντολών

Όσες δυνατότητες και να προσφέρει ένα εργαλείο, δεν είναι δυνατόν να καλύπτει πάντα όλες τις ανάγκες ενός χρήστη, όπως γίνεται με τη χρήση κώδικα. Έτσι τα περισσότερα εργαλεία έχουν στη φαρέτρα τους τη χρήση συναρτήσεων και εντολών, SQL ή δικών τους βιβλιοθηκών (στα πρότυπα των SQL συνήθως), για τη δημιουργία νέων πεδίων στους πίνακες .

Αιτήματα σε φυσική γλώσσα

Ενδιαφέρουσα και φέρελπις είναι η ενσωμάτωση φυσικής γλώσσας στη λειτουργία των εργαλείων. Η αναγνώριση και επεξεργασία φυσικής γλώσσας έλκει το ενδιαφέρον σήμερα.Ορισμένα εργαλεία χρησιμοποιούν φυσική γλώσσα για να επικοινωνήσουν με το χρήστη-άνθρωπο, έστω και σε απλή μορφή.

Εμπειρία Χρήστη

Εύκολη εκμάθηση

Χωρίς αμφιβολία, ο τρόπος λειτουργίας του κάθε εργαλείου είναι διαφορετικός, παρά τις όποιες ομοιότητες που παρουσιάζονται μεταξύ τους. Είναι εύλογο, επομένως, η καμπύλη εκμάθησης να καλύπτει μεγάλο εύρος και οι διαφορές στο χρόνο που απαιτείται για την εξοικείωση να είναι μεγάλες. Τα εργαλεία που μπορούν να χρησιμοποιηθούν από μη έμπειρους χρήστες σε εύλογο χρονικό διάστημα, τα χαρακτηρίζουμε ως «εύκολης εκμάθησης».

Εύκολη δημοσίευση

Τα διαγράμματα και ό,τι τα συνοδεύει, αποτελέσματα των εργαλείων, τις περισσότερες περιπτώσεις προορίζονται για κοινοποίηση σε τρίτους ή δημοσίευση. Η εύκολη δημοσίευση σε ιστοσελίδες (ως εμφωλευμένα στοιχεία), blogs, servers υπό μορφή ευανάγνωστων γραφημάτων και αναφορών είναι μεγάλης σημασίας, τόσο για την εύκολη πρόσβαση και κατανόηση από τους αποδέκτες, όσο και για την εξοικονόμηση χρόνου και αποφυγή περιττών ενεργειών από το χρήστη.

Κοινότητα χρηστών – Forum

Για τη σωστή λειτουργία των εργαλείων, η συμβολή των διαδικτυακών κοινοτήτων είναι πολύ σημαντική, και πιθανόν απαραίτητη. Και αυτό γιατί είναι αδύνατο να επεξηγείται πλήρως η λειτουργία και να καλύπτονται οι όποιες ανάγκες από ένα εγχειρίδιο των κατασκευαστών. Άλλωστε οι αναβαθμισμένες εκδόσεις των εργαλείων προκύπτουν από τα προβλήματα και τις ιδέες που αναφέρονται στα fora, γεγονός που αποδεικνύει το πόσο αναγκαία είναι. Μία πολυπληθής και ενεργή κοινότητα χρηστών αναβαθμίζει δραστικά και το αντίστοιχο εργαλείο, αφού οι συμμετέχοντες συνεργάζονται για λύσεις των δικών τους, συγκεκριμένων προβλημάτων, αλληλοϋποστηρίζονται και έχουν ενεργή πρόσβαση σε μία συνεχώς εμπλουτιζόμενη δεξαμενή γνώσεων.

Ακολουθεί ο πίνακας σύγκρισης:

	SiSense	Qlik Sense	Spotfire	Domo	Pentaho	Tableau	Power BI	Plotly	Chartio	Kibana ELK
Versions										
Desktop	✓	✗	✓	✗	✓	✓	✓	✗	✗	✓
Cloud SaaS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mobile	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
Data Preparation										
varying source extraction	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
Simultaneous extraction - dif sources	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Transform										
Manually Add Data	✓	✓	✗	✗	✓	✗	✓	✓	✗	✗
Convert Type	✗	✗	✓	✗	✗	✓	✓	✓	✗	✗
Sort	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓
Pivot	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗
Joins&Unions	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗
Split	✗	✓	✓	✗	✓	✓	✓	✓	✓	✗
Search Data	✗	✓	✗	✗	✓	✗	✓	✓	✗	✓
.....										
Live Loading	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
Analytics										
Statistics	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓
Trend Line / Regression Line	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗
Forecasting	✗	✗	✓	✓	✓	✓	✗	✗	✗	✓
Other traits										
Maps	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
filters	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
calculated fields	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
natural language queries	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗
dashboard	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
User exp										
easy to learn	✗	✓	✓	✗	✗	✗	✓	✓	✗	✗
easy to publish	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
community/ forum	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Πίνακας 2-1 – πίνακας σύγκρισης βασικών λειτουργιών των εργαλείων

2.5 Συχνότερα χρησιμοποιούμενα εργαλεία visual analytics

Οι χρήστες ενδιαφέρονται κατά σειρά για τα εξής χαρακτηριστικά των εργαλείων: τις οπτικές αναπαραστάσεις, τις αναφορές, τους πίνακες αναπαράστασης, τα προγνωστικά και στατιστικά αναλυτικά στοιχεία, το ETL και τη διασύνδεση με τις πηγές δεδομένων.

Από την πληθώρα των διαθέσιμων εργαλείων, παρατηρούμε ότι οι χρήστες επιλέγουν συχνότερα τα εξής [25]: Tableau 24.44 %, Power BI 11.37%, Qlik Sense 8.79%. Δεν συμπεριλαμβάνουμε το εργαλείο SAS, γιατί είναι μεν ένα εργαλείο επιχειρηματικής νοημοσύνης, αλλά δεν εστιάζει στην οπτική αναπαράσταση και στα διαγράμματα. Είναι, επομένως, ενδιαφέρον να παρουσιαστούν αναλυτικά οι δυνατότητες και η λειτουργία αυτών των τριών. Η περιγραφή των εργαλείων βασίζεται μεν στην τυπική σειρά εισαγωγής-επεξεργασίας –οπτικοποιήσης -αναλυτικών και αναφοράς, αλλά έχει ξεχωριστή προσέγγιση στο κάθε ένα, ώστε να μεταφέρει επιτυχώς τη δομή και τον τρόπο λειτουργίας του.

2.5.1 Tableau 10.4

Εάν θέλουμε να αποδώσουμε ένα κύριο χαρακτηριστικό στο Tableau [26], αυτό είναι η προσήλωση στη δημιουργία διαγραμμάτων - με απλό «σύρσιμο» drag & drop των πεδίων- που καλύπτουν όσο το δυνατόν περισσότερες ανάγκες των χρηστών για οπτική αναπαράσταση, και στην εύστοχη παρουσίασή τους μέσω αναφορών. Η «φιλοσοφία» του εργαλείου αποτυπώνεται με τον διαχωρισμό των πεδίων σε ποσοτικά και ποιοτικά μεγέθη και, στη συνέχεια, στην παροχή πλείστων επιλογών για τη δημιουργία και τη διαμόρφωση των γραφημάτων. Η δυνατότητα σύνδεσης με μεγάλο αριθμό «πηγών» εισόδου δεδομένων, η έγκαιρη ενημέρωση για τις αλλαγές αυτών και η διαχείριση μεγάλου όγκου αρχείων κάνουν, επίσης, το εργαλείο να ξεχωρίζει.

Η ροή των δεδομένων στο Tableau είναι τυπική : εισαγωγή, επεξεργασία, «φόρτωση» προς χρήση, οπτική αναπαράσταση και αναφορές.

Για την εισαγωγή (extract) των δεδομένων γίνεται σύνδεση με ένα ή περισσότερα αρχεία πολλών τύπων. Σε αυτά συμπεριλαμβάνονται αρχεία χωρικών δεδομένων (spatial files με κατάληξη *.kml, *.shp (ESRI),*.geojson κ.α.), στατιστικών δεδομένων (statistical files με κατάληξη *.rda (R) *.sav (SPSS), *.sas7bdat (SAS) κ.α.), αλλά και αρχεία PDF. Η εισαγωγή δεδομένων από αρχείων pdf είναι ιδιαίτερη, αφού λίγα εργαλεία παρέχουν τη δυνατότητα αυτή. Το Tableau μπορεί να διαβάσει οργανωμένα δεδομένα, όπως πίνακες, είτε από προκαθορισμένες σελίδες είτε από ολόκληρο το έγγραφο. Η διαδικασία είναι εντυπωσιακή, αφού τα αρχεία pdf δεν περιέχουν metadata για την περιγραφή τους, και το εργαλείο καταφέρνει να διαβάσει τα δεδομένα, τα οποία πιθανόν να χρειάζονται μόνο ελάχιστες τροποποιήσεις. Η Σύνδεση είναι εφικτή και με μια πληθώρα βάσεων δεδομένων από servers (πίνακας σύνδεσης δεδομένων). Ενώ στα αρχεία είναι απρόσκοπτη η ταυτόχρονη σύνδεση με πολλά και διαφορετικού τύπου, στις βάσεις δεδομένων που προέρχονται από servers, δεν γίνεται πάντα ομαλά, με σωστή εισαγωγή.

Τα εισαχθέντα δεδομένα παρουσιάζονται σε ένα πλαίσιο, για την επισκόπηση των στοιχείων, ανά στήλες. Φυσικά, υπάρχει εποπτεία των αρχείων που έχουν συνδεθεί. Σε άλλο πλαίσιο, αναπαρίστανται οι πίνακες των δεδομένων. Εκεί μπορούν να πραγματοποιηθούν συνδέσεις μεταξύ των πινάκων (joins) και ενώσεις (unions). Οι συνδέσεις και οι ενώσεις τελούνται

ιδανικά μεταξύ πινάκων της ίδιας βάσης δεδομένων- ειδάλλως, πρέπει να δημιουργηθεί μια πηγή πολλαπλής σύνδεσης. Το εργαλείο παρέχει και εξειδικευμένες επιλογές, όπως σύνδεση με τομή σε πεδία με κενές τιμές (null values). Χρήσιμη είναι και η επιλογή pivot- διαμόρφωση των εγγραφών στις γραμμές, ώστε να αντικαταστήσουν την ύπαρξη μίας στήλης. Επίσης, είναι δυνατή η μετονομασία στηλών, ο διαχωρισμός τους, η ομαδοποίησή τους, η δημιουργία νέας στήλης μέσω εντολών και χρήσης συναρτήσεων και η αλλαγή του τύπου των δεδομένων (αριθμός, ακολουθία χαρακτήρων κ.τ.λ.). Εκτός των μετασχηματισμών που επιτρέπονται στο χρήστη, υπάρχει και η δυνατότητα αυτόματης προσαρμογής των στοιχείων, μέσω του «Data Interpreter». Επιτελεί λειτουργίες όπως διαχωρισμός στηλών, αντιστοίχιση επικεφαλίδων και μετατροπή τύπων των δεδομένων. Βέβαια, μπορεί να χρησιμοποιηθεί μόνο σε ορισμένα σύνολα δεδομένων, και δεν είναι πάντα αποτελεσματικός.

Καθοριστικής σημασίας, ώστε να παραχθεί το επιθυμητό αποτέλεσμα, αλλά και να λειτουργεί γρήγορα και απρόσκοπτα το εργαλείο, είναι ο τρόπος «φόρτωσης» (loading), καθιστώντας τα δεδομένα έτοιμα προς ανάλυση και οπτικοποιήση. Υπάρχουν δύο επιλογές «φόρτωσης», η εξαγωγή «extract» των δεδομένων από την πηγή στο εργαλείο και η ζωντανή επικοινωνία «live» loading. Η εξαγωγή δεδομένων αποθηκεύει μόνιμα τα δεδομένα, όπου είναι συνεχώς διαθέσιμα. Στη ζωντανή επικοινωνία το εργαλείο αντλεί από την πηγή, με queries τα δεδομένα που χρειάζονται. Για τη «live» σύνδεση, εμφανίζεται συχνά πρόβλημα στην ανανέωση δεδομένων που ήδη βρίσκονται στην μνήμη cache του Tableau.

Η επιλογή του loading έχει να κάνει κυρίως με το αρχείο. Εάν, π.χ. είναι επιθυμητή η εισαγωγή ενός αρχείου excel, που ανανεώνεται ανά ορισμένα χρονικά διαστήματα, είναι προτιμητέο το «extract». Μπορεί, μάλιστα, να ρυθμιστεί και ανανέωση του αρχείου ανά τακτά διαστήματα, με ελάχιστα τα 15 λεπτά. Για το loading από μία βάση δεδομένων, που καταγράφει, π.χ. , κινήσεις και απαιτείται άμεση επεξεργασία, προτιμούμε τη «ζωντανή φόρτωση». Η όποια πιθανή καθυστέρηση στη «ζωντανή φόρτωση» οφείλεται στο χρόνο που απαιτείται για τον εντοπισμό των δεδομένων που ζητούνται από τα queries, από τις βάσεις. Συνεπώς, έχει να κάνει και με τη δομή της βάσης, το μέγεθος του αρχείου και τη σύνδεση στο δίκτυο. Επομένως, όταν δεν είναι απαραίτητη η κυριολεκτικά άμεση ενημέρωση, προτιμούμε το «extract», με το οποίο το εργαλείο λειτουργεί χωρίς καθυστερήσεις και επιβάρυνση για τους υπολογιστικούς πόρους του συστήματος που χρησιμοποιείται. Η επιλογή extract είναι σχεδιασμένη για την καλύτερη δυνατή απόδοση, ώστε και να είναι δυνατή η επεξεργασία μεγάλων αρχείων σε εύλογο χρονικό διάστημα αλλά και η έγκαιρη απόκριση στο περιβάλλον του εργαλείου (θα αναλυθεί εκτενέστερα η λειτουργία αυτή). Η «ζωντανή» σύνδεση, ωστόσο, είναι εφικτή μόνο σε ορισμένους τύπους βάσεων δεδομένων, και είναι, ουσιαστικά, σχεδιασμένη να λειτουργεί βέλτιστα στο Tableau Server.

Ακολουθεί η δημιουργία των διαγραμμάτων, όπου λαμβάνει χώρα η οπτική αναπαράσταση των δεδομένων. Για τη δημιουργία των γραφημάτων, αρκεί να «συρθούν» τα επιθυμητά πεδία στη θέση των σειρών και στηλών. Ο χρήστης επιλέγει τα δεδομένα και το Tableau επιλέγει το καταλληλότερο διάγραμμα, ανάλογα την περίσταση. Υπάρχει και παράθυρο πρότασης διαγραμμάτων, όπου επεξηγείται τι τύπου πεδία προϋποθέτει το κάθε ένα. Η «φιλοσοφία» του εργαλείου για τη διαχείριση των δεδομένων στο παράθυρο της οπτικοποιήσης έγκειται στο διαχωρισμό αυτών σε ποσοτικά και ποιοτικά δεδομένα. Τα ποσοτικά, «measures», αναπαρίστανται σε πράσινο πλαίσιο (green pill) και τα ποιοτικά, «dimensions», σε μπλε πλαίσιο (blue pill). Στα ποσοτικά ανήκουν τα δεδομένα με τα οποία γίνονται αριθμητικές πράξεις, ενώ στα ποιοτικά ονόματα, ημερομηνίες και άλλες πληροφορίες, (υπάρχει, όμως, η δυνατότητα να αλλάξει η κατηγορία, εάν χρειαστεί).

Δημιουργούνται, επίσης, δύο νέα «ψεύδο-πεδία», τα «Measure Values» ποσοτικό και «Measure Values» ποιοτικό, τα οποία δεν αντιπροσωπεύουν συγκεκριμένες τιμές, αλλά χρησιμοποιούνται στη χάραξη των διαγραμμάτων. Γενικότερα, η δημιουργία των διαγραμμάτων στηρίζεται στο διαχωρισμό ποιοτικών και ποσοτικών μεγεθών σε κατηγορίες. Η επιλογή των πεδίων στα διαγράμματα δεν είναι απλή, και ίσως δυσχεραίνει το νέο χρήστη, ωστόσο επιτρέπει τη δημιουργία πιο απαιτητικών και πολύπλοκων διαγραμμάτων. Διαγράμματα μπορούν να αναπαρασταθούν και από scripts σε γλώσσα R.

Συν τοις άλλοις, είναι δυνατή η επιλογή των σχημάτων απεικόνισης μεταξύ ευθειών, σημείων, απεικόνισης με ράβδους, με κάλυψη επιφάνειας, με κύκλους και με τετράγωνα, σε όσα διαγράμματα μπορούν να δεχτούν τέτοια αλλαγή, π.χ. διαγράμματα χρονοσειρών. Οι επιλογές μορφοποίησης είναι πολλές και καλύπτουν τη διαμόρφωση των αξόνων, την επιλογή και αντιστοίχιση χρωμάτων, το φόντο εργασίας, τις σκιάσεις και τα περιθώρια, το μέγεθος των σημείων αναπαράστασης και τις ετικέτες πληροφοριών.

Για τα γραφήματα γεωγραφικών δεδομένων, το εργαλείο αναγνωρίζει ή δημιουργεί (εάν δεν υπάρχουν) αυτόματα τα πεδία γεωγραφικού πλάτους και μήκους. Τα ονόματα χωρών, περιοχών και πόλεων, οι ταχυδρομικοί κώδικες και άλλα συναφή στοιχεία αντιστοιχίζονται στην περιοχή του χάρτη, χωρίς να χρειάζεται κάποια ενέργεια από το χρήστη. Η δυνατότητα πλοήγησης και εστίασης στους χάρτες είναι ικανοποιητική. Προσφέρονται διάφορες επιλογές επιπέδων φόντου, όπως δρόμοι, ακτογραμμές κ.α. Σημαντικό είναι ότι τα χωρικά αρχεία (spatial files), που δίνουν γεωγραφική πληροφορία επιπλέον, π.χ. όρια των δήμων σε μία χώρα, συγχωνεύονται επιτυχώς με το βασικό χάρτη.

Όσον αφορά το φιλτράρισμα των πληροφοριών, το Tableau δίνει πολλές επιλογές και καλύπτει όλο το φάσμα των επιθυμητών περιορισμών σε ποσοτικά δεδομένα, διαστάσεις και ημερομηνίες (ποιοτικά). Τα φίλτρα των ποσοτικών δεδομένων είναι :

- Καθορισμένου εύρους
- Άνω ορίου
- Κάτω ορίου
- Ειδικό, για φιλτράρισμα κενών τιμών

Για τα ποιοτικά δεδομένα:

- Γενικό (general), με επιλογή ή απόκλιση συγκεκριμένων τιμών
- Περιεχομένου, οπουδήποτε, στην αρχή, ή στο τέλος μίας ακολουθίας χαρακτήρων
- Υπό όρους, συνήθως για κάποιο πεδίο ποσοτικών τιμών (π.χ. άθροισμα >100)
- Κορυφής, όπου εμφανίζονται τα στοιχεία με τις μέγιστες ή ελάχιστες τιμές σε ποσοτικό πεδίο

Τα φίλτρα, μπορούν να διαμορφωθούν και από εντολές και συναρτήσεις, παρ' ότι καλύπτονται σχεδόν όλες οι περιπτώσεις από τις δυνατότητες του εργαλείου. Δίνεται και η επιλογή γρήγορου φιλτραρίσματος, άμεσα με δεξί κλικ στο επιθυμητό πεδίο. Πολύ σημαντική είναι η ρύθμιση για την επίδραση των φίλτρων, όπου γίνεται επιλογή εάν το φίλτρο ισχύει για όλο το project ή για συγκεκριμένες σελίδες (sheets) και πίνακες αναπαράστασης (dashboards).

Μέσω του Tableau, μπορούν να υπολογιστούν διάφορα αναλυτικά στοιχεία. Για τα στατιστικά μεγέθη, άμεσα μπορούν να υπολογιστούν η τυπική απόκλιση, η διακύμανση το percentile, ο μ.ο. και η διάμεσος με 95% διαστήματα εμπιστοσύνης. Επίσης, εύκολα σχεδιάζονται ευθείες κάποιας σταθεράς, ευθεία μ.ο. , box plots, διάμεσος με τεταρτημόρια. Σε κατάλληλες χρονοσειρές, με αρκετά στοιχεία και ημερομηνίες, μπορούν να πραγματοποιηθούν και προβλέψεις (forecasting). Το forecasting γίνεται μέσω κάποιου τύπου εκθετικής εξομάλυνσης, για το επιθυμητό μελλοντικό χρονικό διάστημα. Διαθέσιμη είναι και η σχεδίαση διαστημάτων εμπιστοσύνης των προβλέψεων. Οι ευθείες τάσης (trend lines), που επίσης συμπεριλαμβάνει το εργαλείο, μπορεί να σχεδιαστούν γραμμικά, λογαριθμικά, εκθετικά και πολυωνυμικά. Αυτές περιγράφονται σε ξεχωριστό «παράθυρο», όπου αναγράφεται η εξίσωση, το μέσο τετραγωνικό σφάλμα, probability value, συντελεστής προσδιορισμού (R- squared) κ.α. Επίσης, ο χρήστης μπορεί να κάνει ομαδοποίησης (clustering) στα δεδομένα, με το κριτήριο Calinski- Harabasz.

Η παρουσίαση των διαγραμμάτων και άλλων πληροφοριών γίνεται στους πίνακες αναπαράστασης (dashboards). Εύκολα προστίθενται τα διαγράμματα, τα σχόλια, και όποια, επιπλέον, φίλτρα χρειάζεται. Η παρουσίαση πολλών διαγραμμάτων σε έναν πίνακα δεν είναι κατάλληλη για τη σωστή μετάδοση της πληροφορίας. Χρήσιμη δυνατότητα είναι η εμφάνιση κάποιας διαδικτυακής σελίδας, παράλληλα με τα διαγράμματα στους πίνακες αναπαράστασης. Προσφέρεται, επίσης, υποθετική ανάλυση, όπου αλλάζουν τα παραχθέντα αποτελέσματα, ανάλογα με την υποθετική μεταβολή μιας τιμής. Στο Tableau, η αναφορά παρουσιάζεται στο κατάλληλο πλαίσιο «Story», έτοιμο για παρουσίαση στους ενδιαφερόμενους.

2.5.1.1 Αρχιτεκτονική δεδομένων στο Tableau

Η γλώσσα στην οποία εκτελείται το Tableau είναι η ιδιόκτητη VizQL [27]. Όπως υποδηλώνει και το όνομά της, είναι μία βασισμένη στην SQL γλώσσα, που εστιάζει στην οπτική αναπαράσταση. Ό,τι η SQL κάνει με κείμενο, η VizQL κάνει με γραφήματα. Η VizQL αναπαριστά τα δεδομένα πολύ πιο γρήγορα από τις συμβατικές μεθόδους. Η επιλογή «φόρτωσης» των δεδομένων «extract», λειτουργεί σύμφωνα με το σχεδιασμό του εργαλείου, αρα και της VizQL. Στη «ζωντανή» επικοινωνία με βάσεις δεδομένων, λειτουργεί σύμφωνα με την αρχιτεκτονική και τη γλώσσα της πηγής. Για το λόγο αυτό, η πρόσβαση στα δεδομένα μέσω του «extract» είναι πολύ γρηγορότερη και, ταυτόχρονα, επιτρέπει στο εργαλείο να λειτουργεί απρόσκοπτα κατά την ανάλυση και αναπαράσταση των δεδομένων.

Το «Tableau Data Extract» είναι ένα συμπιεσμένο στιγμιότυπο δεδομένων, αποθηκευμένων στο δίσκο, ενώ αυτό είναι «φορτωμένο» στη μνήμη. Η οργάνωση της μνήμης του εργαλείου είναι «column store», δηλαδή αποθήκευση των δεδομένων σε στήλες, και όχι σε μορφή πίνακα. Συνεπώς, η αναζήτηση στοιχείου απαιτεί προσπέλαση μόνο σε μία διάσταση, αντί δύο σε πίνακα. Το ίδιο πλεονέκτημα ισχύει και για αριθμητικές πράξεις σε σύνολα, όπως άθροισμα και μ.ο. , όταν γίνεται κατά στήλη. Δε χρειάζεται λοιπόν η προσπέλαση όλων των σειρών, και έπειτα η μετάβαση στη στήλη του πεδίου, και ο χρόνος που απαιτείται είναι συγκριτικά ελάχιστος.

Το Tableau, αρχικά, προσδιορίζει τη δομή του «extract» και κάθε στήλη προς «φόρτωση» τοποθετείται σε ξεχωριστό αρχείο (αυτός είναι και ο λόγος που το εργαλείο επιβαρύνεται όταν ο αριθμός των στηλών είναι μεγάλος, πράγμα που δεν ισχύει για τις γραμμές). Οι στήλες

ταξινομούνται και συμπιέζονται, όχι ως συμπιεσμένα αρχεία, αλλά με τεχνικές όπως dictionary compression, run length encoding, frame of reference encoding και delta encoding, ώστε ο προσωρινά καταλαμβανόμενος χώρος (στη μνήμη) να είναι ο ελάχιστος δυνατός. Για τις τεχνικές αυτές, δεν απαιτείται αποσυμπίεση. Με το συνδυασμό των αρχείων όπου βρίσκονται οι στήλες και των κατάλληλων metadata, δημιουργείται το memory-mapped αρχείο. Το «Tableau Data Extract, είναι δηλαδή memory-mapped αρχείο που βρίσκεται μόνιμα στη μνήμη κατά τη λειτουργία του εργαλείου, δεν απαιτείται άνοιγμα και αποσυμπίεση του από το λειτουργικό σύστημα. Με το σχεδιασμό αυτό, τα αρχεία των στηλών δε χρειάζεται να βρίσκονται μόνιμα στη μνήμη, αλλά φέρονται από το δίσκο, όποτε ζητούνται, έτοιμα προς χρήση. Ως εκ τούτου, το Tableau μπορεί να διαχειριστεί μεγάλα σύνολα δεδομένων, που ξεπερνούν το μέγεθος της μνήμης, με τυπικές απαιτήσεις υλισμικού και χωρίς καθυστερήσεις.

2.5.2 MS Power BI

*η περιγραφή αφορά την έκδοση του MS Power BI για τον Νοέμβριο του 2017

Το Power BI [28], εργαλείο της Microsoft, χαρακτηρίζεται από τη γρήγορη δημιουργία διαγραμμάτων και την εύκολη επεξεργασία των δεδομένων. Το εργαλείο αναπτύχθηκε ως μία εξέλιξη του Excel, με σκοπό να έχει την ίδια μεγάλη δυνατότητα επεξεργασίας δεδομένων, αλλά να εμπλουτιστεί με διαδραστικά διαγράμματα, αναφορές, αμεσότητα και, εν τέλει να προσφέρει ολοκληρωμένες παροχές εξερευνητικής ανάλυσης δεδομένων (exploratory data analysis). Η service (διαδικτυακή) έκδοση διαφέρει – συγκεκριμένα, παρέχει κάποιες επιπλέον δυνατότητες- από τη desktop, οπότε οι όποιες διαφορές θα επισημαίνονται.

Η εισαγωγή των δεδομένων γίνεται εύκολα, επιλέγοντας είτε από τα πιο συνηθισμένα αρχεία εισόδου, όπως excel, csv, SQL server, είτε από μια μεγαλύτερη λίστα πολλών βάσεων δεδομένων και άλλων αρχείων (πίνακας δεδομένων). Όσον αφορά τα αρχεία του Excel, στην service έκδοση, μπορεί να γίνει είτε εισαγωγή των δεδομένων, είτε σύνδεση, που σημαίνει αλληλεπίδραση με την πηγή, ζωντανά. Η «Ζωντανή» σύνδεση με την πηγή, για βάσεις δεδομένων παρέχεται μόνο στους χρήστες της «Pro» έκδοσης. Κατά την εισαγωγή των δεδομένων, η αναγνώριση και η μετατροπή τους σε κατάλληλη μορφή μπορεί να γίνει αυτόματα, βασισμένη στις 200 πρώτες σειρές δεδομένων. Μπορούν να εισαχθούν δεδομένα από διαφορετικές πηγές (περιορισμένου αριθμού στη desktop έκδοση). Εάν είναι επιθυμητό να συνδυαστούν τα δεδομένα αυτών για κάποια διαγράμματα ή αναλυτικά στοιχεία, τότε συγχωνεύουμε (merge) τους πίνακες.

Το Power BI χωρίζεται σε τρία «παράθυρα», το κεντρικό των διαγραμμάτων, αυτό της εποπτείας των δεδομένων, και αυτό της καταγραφής των εισαχθέντων ή δημιουργηθέντων πινάκων. Στο παράθυρο της εποπτείας φαίνονται όλα τα δεδομένα, όπου μπορούν να ταξινομηθούν, να ομαδοποιηθούν, να αντιγραφούν, να ανανεωθούν ή να δημιουργηθούν άμεσα νέα πεδία/ στήλες.

Υπάρχουν γραμμές εργαλείων, στα πρότυπα του Excel, που έχουν διάφορα «πλήκτρα», όπως εισαγωγής νέου πίνακα, εισαγωγής εικόνας και κειμένου, διαχείρισης των συνδέσεων μεταξύ των πινάκων, εισαγωγή προτύπων, δημοσίευσης, βοήθειας και άλλων συντομεύσεων.

Για το μετασχηματισμό και την επεξεργασία των δεδομένων, εκτός των προαναφερθέντων συντομεύσεων, χρησιμοποιείται το υπο-πρόγραμμα «Query Editor». Μέσω αυτού, οι δυνατότητες επεξεργασίας είναι ποικίλες και τελούνται πολύ εύκολα, ακόμα και από νέους χρήστες. Μετασχηματισμοί όπως αλλαγή τύπου δεδομένων, επικεφαλίδων, διαχωρισμοί στηλών, ομαδοποιήσεις (group by), κατάργηση στηλών, συγχώνευση (merge), συνδέσεις και προσαρτήσεις πινάκων (append, αντίστοιχο του join), περιστροφή πίνακα (pivot), στήλες μετρητών κ.α. είναι εύκολα εκτελέσιμοι. Χρήσιμη είναι η επιλογή αντικατάστασης επιλεγμένων τιμών «replace values», ειδικά για περιπτώσεις που υπάρχει κάποιο σφάλμα προς αντικατάσταση στα αρχικά δεδομένα. Επίσης, η δημιουργία νέας στήλης, γίνεται απλά, μόνο με επιλογές, χωρίς να χρειάζεται ο χρήστης να γράψει εντολές και να γνωρίζει συναρτήσεις. Ενδεικτικό είναι ότι υπάρχει ακόμα και έτοιμη επιλογή δημιουργίας νέας στήλης υπό συνθήκη, π.χ. ανάλογα την τιμή του στοιχείου της σειράς σε ένα πεδίο. Υπάρχει η επιλογή «formula bar», για να εμφανίζονται οι συναρτήσεις που αντιστοιχούν στα νεοδημιουργηθέντα πεδία, και γενικά στον πίνακα. Εάν ο χρήστης επιθυμεί να συντάξει ο ίδιος εντολές και συναρτήσεις, το κάνει μέσω του «Advanced Editor», στη γλώσσα αιτημάτων «M» της Microsoft. Το υπο-πρόγραμμα μπορεί να επεξεργαστεί ταυτόχρονα περισσότερους του ενός πίνακες, καταγράφοντας την κάθε αλλαγή σε μία λίστα βημάτων. Για να εφαρμοστούν οι αλλαγές απαιτείται έγκριση από το χρήστη. Σε γενικές γραμμές, ο «Query Editor», προσφέρει όλες τους συνήθεις μετασχηματισμούς, αλλά και προηγμένες επιλογές επεξεργασίας, χωρίς την ανάγκη χρήσης εντολών σε κάποια γλώσσα, και σε όλες τις πιθανές μορφές που μπορεί να ζητηθούν. Η επεξεργασία δεδομένων είναι σίγουρα από τα πληρέστερα μέρη του εργαλείου, και ένα σημείο υπεροχής έναντι των υπολοίπων εργαλείων.

Η δημιουργία διαγραμμάτων στο Power BI είναι πολύ απλή διαδικασία. Επιλέγεται το επιθυμητό διάγραμμα από τα διαθέσιμα, και «σύρονται» τα πεδία προς αναπαράσταση στις θέσεις «αξία», «λεζάντα», «άξονα», «εργαλείο κείμενου» (κ.α. , ανάλογα τον τύπο του διαγράμματος). Τα πεδία προς απεικόνιση φέρουν ένα χαρακτηριστικό σύμβολο που διευκρινίζει το είδος τους, όπως το σύμβολο του αθροίσματος για αριθμητικά δεδομένα, και της υδρογείου για γεωγραφικά. Η διαδικασία μπορεί να γίνει και αντίστροφα, δηλαδή να επιλεχθούν τα πεδία, και έπειτα να εμφανίσει ο χρήστης όποιο διάγραμμα επιθυμεί. Η δυναμική μορφή των διαγραμμάτων, δηλαδή ότι δεν είναι σταθερά και με τα ίδια πεδία μπορούν να δοκιμαστούν όλα τα γραφήματα, είναι η πεμπτουσία της εξερευνητικής ανάλυσης δεδομένων, για τον εντοπισμό χρήσιμων πληροφοριών. Η εύκολη σχεδίαση διαγραμμάτων αντισταθμίζεται ωστόσο από την αδυναμία δημιουργίας πολύπλοκων γραφημάτων.

Το εργαλείο παρέχει έτοιμα τα βασικά διαγράμματα, όπως διαγράμματα μπάρας (stacked και clustered), και σε ποσοστιαία απεικόνιση, διαγράμματα ευθειών, επιφάνειας, πίττας, treemaps, gauge κ.α. Υπάρχουν, επίσης, βοηθήματα για την αναπαράσταση στοιχείων με την αριθμητική τους μορφή. Για παράδειγμα, ο χρήστης θέλει να εμφανίσει στην αναφορά του την τυπική απόκλιση και τους μέσους όρους κάποιων δεδομένων, και κάποιο ταξινομημένο πεδίο, που θα συνοδεύουν τα διαγράμματα. Τέτοια είναι οι κάρτες (μίας ή πολλών σειρών) και οι πίνακες table και matrix (ο matrix είναι ένας περιστρεφόμενος πίνακας). Και οι δύο αυτοί πίνακες μπορούν είτε να αντιστοιχιστούν με τα υπάρχοντα πεδία, είτε να δημιουργηθούν υπό συνθήκη οι τιμές τους. Φυσικά, όλα τα βοηθήματα αυτά είναι διαδραστικά, και όταν επιλέγεται μία μεταβλητή, αυτή τονίζεται σε όλα τα διαγράμματα της αναφοράς. Ενδιαφέρον έχει και το γράφημα δείκτη επίδοσης «KPI» (key performance indicator), χρήσιμο για την αξιολόγηση επιχειρήσεων συνολικά αλλά και συγκεκριμένων

τομέων. Πάρα πολύ σημαντικό βοήθημα διαγραμμάτων είναι και το «slicer», με το οποίο μπορούμε να επιλέξουμε τι εμφανίζουμε στα διαγράμματα, ανάλογα την τιμή κάθε εγγραφής για ένα πεδίο, συνήθως ποιοτικών δεδομένων. Είναι δηλαδή μια μορφή κεντρικού φιλτραρίσματος της αναφοράς, κατάλληλη για συγκρίσεις και απομόνωση δεδομένων. Πέρα από αυτές τις επιλογές, μπορεί να χρησιμοποιηθεί και κάποιος εισαγόμενος τύπος διαγράμματος «custom visual».

Για την οπτική αναπαράσταση γεωγραφικών δεδομένων, διατίθενται το διάγραμμα χάρτη (σημείων), το διάγραμμα σκιασμένου χάρτη (πολύγωνων). Δεν είναι απαραίτητο να δώσουμε συντεταγμένες για το σχεδιασμό, όταν είναι διαθέσιμα ονόματα χωρών και περιοχών, γίνεται αυτόματα η αντιστοίχιση στα πολύγωνα που καταλαμβάνουν στην επιφάνεια του χάρτη. Στο Power BI, είναι δυνατή η εισαγωγή νέων χαρτών, από χωρικά αρχεία, μόνο αφότου μετατραπούν σε topojson, μία προέκταση του geojson, που χρησιμοποιείται από το εργαλείο. Υπάρχει δυνατότητα επιλογής από διαφορετικούς τύπους φόντου για την εμφάνιση του χάρτη, όπως απλός, γεωγραφικός κ.α. Σε ορισμένες εκδόσεις, είναι διαθέσιμοι και οι χάρτες της πλατφόρμας «ArcGis», οι οποίοι είναι ποικίλοι θεματικοί χάρτες, π.χ. χάρτες που αναπαριστούν την κίνηση, τις καιρικές συνθήκες ή τη βλάστηση. Σε συνδυασμό με τις δυνατότητες του Power BI προσφέρουν μια ολοκληρωμένη προσέγγιση, ώστε η ανάλυση των γεωγραφικών δεδομένων να είναι λεπτομερής αλλά και διαδραστική, πλαισιωμένη από τα κατάλληλα διαγράμματα.

Τα διαγράμματα μπορούν να υποστούν μορφοποιήσεις, όχι όμως σε λεπτομερή βαθμό. Συμπεριλαμβάνονται αλλαγές στη θέση και το μήκος των αξόνων, στο μέγεθος των σημείων και ευθειών του γραφήματος, στα όρια, στις λεζάντες και στο φόντο. Ειδικά για τα χρώματα που χρησιμοποιούνται από τα διαγράμματα, η επιλογή γίνεται από διάφορες διαβαθμίσεις βασικών χρωμάτων, είτε από την παλέτα, με πολλούς κωδικούς χρωμάτων. Η αναπαράσταση των χρωμάτων, ωστόσο, δεν είναι ευδιάκριτη για τους αποδέκτες των πληροφοριών, τουλάχιστον για την αυτόματη αντιστοίχιση χρωμάτων.

Τα διαγράμματα απεικονίζονται σε «σελίδες» του εργαλείου. Μπορούμε να ρυθμίσουμε το χώρο που καταλαμβάνουν με χαρακτηριστική ευκολία, χωρίς να χάνεται η ευκρίνεια των σχημάτων. Συνεπώς, κάθε σελίδα μπορεί να οργανωθεί κατά βούληση, π.χ. σύμφωνα με ένα θέμα αναφοράς. Κάθε «πλακάκι» (tile) διαγράμματος έχει πλήκτρο εστίασης, ώστε να απεικονίζεται μεμονωμένα στην οθόνη, σε πλήρεις διαστάσεις. Το σύνολο των σελίδων απαρτίζουν την αναφορά. Στη διαδικτυακή έκδοση «service» υπάρχει και ο πίνακας οπτικής αναπαράστασης «dashboard», όπου μπορούμε να εισάγουμε διαγράμματα. Η διαφορά σε σχέση με την «αναφορά» είναι ότι το «dashboard» προσφέρεται μόνο για ανάγνωση, πλην κάποιων τροποποιήσεων στην εμφάνιση των διαγραμμάτων. Οπότε, ανάλογα με το εάν είναι επιθυμητές αλλαγές και παρεμβάσεις, επιλέγουμε και πεδίο παρουσίασης.

Όσον αφορά τις επιλογές φιλτραρίσματος, ο χρήστης μπορεί να βάλει περιορισμούς σε επίπεδο διαγράμματος, σε ολόκληρη τη σελίδα ή στην αναφορά. Καινοτομία αποτελούν τα φίλτρα «drillthrough», που δίνουν τη δυνατότητα εστίασης και απομόνωσης μίας οντότητας, π.χ. ενός πελάτη ή προμηθευτή, σε ξεχωριστή σελίδα. Αναπαρίστανται, προσαρμοσμένα, όλα τα σχήματα που αφορούν πεδία όπου εμπεριέχεται η επιλεγμένη οντότητα. Για τα ποιοτικά μεγέθη, τα φίλτρα είναι «βασικά», όπου επιλέγονται οι τιμές που θα παρουσιαστούν, μεγίστων και ελαχίστων τιμών, για κάποιο ποσοτικό μέγεθος που τους αντιστοιχεί, ή «προηγμένα» όπου φιλτράρονται με την αρχή, κατάληξη ή περιεχόμενο χαρακτήρων στην τιμή τους (όνομα) όπου υπάρχει δυνατότητα διάζευξης και σύζευξης με άλλο «προηγμένο»

φίλτρο. Για τα ποσοτικά μεγέθη, εισάγονται περιορισμοί τιμών με τελεστές σύγκρισης – και εδώ υπάρχει δυνατότητα διάζευξης και σύζευξης με άλλες εντολές. Για τις ημερολογιακές τιμές χρονοσειρών, το Power BI λειτουργεί καλύτερα με το διαδραστικό φιλτράρισμα, με επιλογή χρονικής κλίμακας άξονα (χρόνος έως ημέρα) και την εστίαση σε συγκεκριμένα τμήματα του άξονα.

Παρέχονται έτοιμες βασικές στατιστικές συναρτήσεις, όπως τυπική απόκλιση, διακύμανση, διάμεσο των παρατηρήσεων, μέγιστο και ελάχιστο. Επίσης, υπάρχει η δυνατότητα χάραξης γραμμικών ευθειών τάσεων (trend lines). Άξια αναφοράς είναι και τα υποθετικά σενάρια (what if), όπου τα διαγράμματα και τα αναλυτικά στοιχεία μεταβάλλονται ανάλογα με την υποθετική τιμή μίας μεταβλητής.

Στο Power BI υπάρχουν κάποιες ξεχωριστές λειτουργίες, που αφορούν τη διαδικτυακή service έκδοση. Η εξαγωγή των «quick insights», είναι η αυτόματη διαδικασία δημιουργίας διαγραμμάτων και υπολογισμού αναλυτικών στοιχείων, που εικάζει το πρόγραμμα ότι πιθανώς να ενδιαφέρουν το χρήστη (augmented analytics). Οι επιλογές για τα διαγράμματα που θα παρασταθούν γίνονται μέσω της τεχνητής νοημοσύνης του εργαλείου. Εντυπωσιακά είναι και τα αιτήματα σε φυσική γλώσσα (natural language queries). Γράφοντας σε ένα ειδικό πλαίσιο, ονόματα πεδίων και απλές προτάσεις, π.χ. « which product has the highest revenue», εμφανίζονται διαγράμματα που απαντούν αυτήν την ερώτηση. Κατά την πληκτρολόγηση, εμφανίζονται προτάσεις προς επιλογή. Εκτός από έναν πολύ γρήγορο τρόπο δημιουργίας διαγραμμάτων, αφήνει το περιθώριο για οπτική αναπαράσταση δεδομένων, όχι με χειρισμό ενός προγράμματος, συμβατικά, μπροστά από το πληκτρολόγιο, αλλά με ερωτήσεις που θα κάναμε σε έναν συνάδελφο του τμήματος ανάλυσης δεδομένων. Έτσι, ο χρήστης μπορεί να επικεντρώνεται στην εξαγωγή συμπερασμάτων, χωρίς να τον απασχολούν ζητήματα τεχνικής φύσεως.

2.5.3 Qlik Sense

*η περιγραφή αφορά την έκδοση Qlik Sense του Σεπτέμβρη του 2017

Με την πρώτη επαφή του χρήστη με το περιβάλλον του εργαλείου, γίνεται αντιληπτό πως πυρήνας της φιλοσοφίας σχεδιασμού του εργαλείου είναι η γρήγορη δημιουργία διαγραμμάτων και η καλή οργάνωση, ώστε να μην αποσπάται από την διερμηνεία των δεδομένων. Το Qlik Sense [29], «απόγονος» του Qlik View, είναι ένα «online» εργαλείο, με αναγκαία τη σύνδεση στο διαδίκτυο για την είσοδο του χρήστη.

Είναι οργανωμένο σε «εφαρμογές» (apps), όπου κάθε εφαρμογή να αντιστοιχεί σε ξεχωριστό project, ώστε να μην εμπλέκονται χωρίς λόγο τα διαγράμματα, τα σύνολα δεδομένων και οι παρουσιάσεις, αλλά και να εμφανίζονται στην κεντρική σελίδα του εργαλείου έτοιμες προς επιλογή. Στις εφαρμογές, είναι δυνατόν να δημιουργηθούν διαγράμματα και πίνακες αναπαράστασης με συγκεκριμένη διάταξη πεδίων, έτοιμα προς προβολή μετά την είσοδο νέων πληροφοριών. Έτσι, η δημιουργία του ίδιου τύπου διαγράμματος για διαφορετικά δεδομένα δεν απαιτεί το σχεδιασμό του, παρά μόνο μία φορά. Ενδεικτικά το Qlik Sense έχει τέσσερις έτοιμες εφαρμογές: πωλήσεις αγαθών, πλατφόρμα ελέγχου επιχειρήσεων, ανάλυσης πωλήσεων και διεύθυνσης υποστήριξης. Οι έτοιμες φόρμες επεξεργασίας δεδομένων που προσφέρουν οι «εφαρμογές», μπορούν να εξοικονομήσουν πολύ χρόνο από το σχεδιασμό.

Οι επιφάνειες εργασίας του ιεραρχούνται ως εξής: στην επιφάνεια εισόδου εμφανίζονται όλες οι εφαρμογές, οι οποίες προβάλλονται με εικόνα και σύντομη περιγραφή. Στο επίπεδο της εφαρμογής, υπάρχουν τα «φύλλα» (sheets) όπου παρουσιάζονται τα διαγράμματα, και οι «ιστορίες» που είναι ουσιαστικά παρουσιάσεις.

Κατά την εκκίνηση μίας εφαρμογής, γίνεται η προσθήκη δεδομένων είτε από αρχείο, είτε με τη σύνδεση σε κάποια πηγή δεδομένων. Μπορούν να εισαχθούν ταυτόχρονα δεδομένα από πολλαπλά αρχεία. Μία ξεχωριστή δυνατότητα που παρέχεται, είναι η εισαγωγή δεδομένων χειροκίνητα από το χρήστη. Στο ξεχωριστό παράθυρο ρύθμισης της εισαγωγής δεδομένων «Data Load Editor», εμφανίζεται το σενάριο (script) εισαγωγής, όπου ρυθμίζεται ουσιαστικά το στάδιο του «data cleaning». Ο χρήστης μπορεί να ρυθμίζει παραμέτρους όπως στίξη διαχωρισμού δεκαδικών και χιλιάδων, τα ονόματα ημερών και μηνών και η κατάταξη τους, τη μορφή των ακολουθιών για ημερομηνία, ώρα, χρηματικά ποσά, και τις τοπικές ρυθμίσεις. Είναι δυνατόν να γίνει παρέμβαση και στο σενάριο εισαγωγής, που επιβάλλει όμως να συνεχιστεί η επεξεργασία στο ETL στάδιο μόνο με script.

Στο παράθυρο διαχείρισης δεδομένων «Data Manager», εμφανίζονται τα σύνολα δεδομένων (πίνακες), καθώς και οι μεταξύ τους σχέσεις (associations), οι αντίστοιχες συνδέσεις και ενώσεις δηλαδή υλοποιημένες έτσι ώστε να δεσμεύουν λιγότερη μνήμη, που εκτελούνται εύκολα μέσα στο γραφικό περιβάλλον. Επιλέγοντας έναν πίνακα, δίνεται η εποπτεία όλων των στοιχείων του, καθώς και η δυνατότητα μετασχηματισμών. Η διαδικασία περιστροφής στήλης σε σειρές (pivot ή unpivot) γίνεται άμεσα μέσω πλήκτρου. Επίσης, τα πεδία μπορούν να ταξινομηθούν, να μετονομαστεί η στήλη, και να επιλεγεί η κατηγορία που ανήκουν τα δεδομένα τους (γενικά, ημερολογιακά, γεωγραφικά). Δίνεται η δυνατότητα δημιουργίας νέων πεδίων «calculated fields», μέσω συναρτήσεων όμως, χωρίς υποβοήθηση. Η ανανέωση των δεδομένων από την πηγή γίνεται σε αυτό το παράθυρο, με πλήκτρο. Προγραμματισμένη ανανέωση γίνεται μόνο στη server έκδοση, ενώ «ζωντανή» ανανέωση δεν υπάρχει. Η φόρτωση των δεδομένων γίνεται χειροκίνητα, με πλήκτρο, ώστε να υπάρχει έλεγχος και επικύρωση των μετασχηματισμών. Ιδιαίτερο χαρακτηριστικό του διαχειριστή δεδομένων είναι τα διαγράμματα που δημιουργεί αυτόματα κατά την επιλογή ενός πεδίου από το χρήστη (augmented analytics). Τα διαγράμματα που σχεδιάζονται επιλέγονται, ώστε να προσφέρουν διορατικότητα, και μπορεί ο χρήστης να ρυθμίσει τη διάταξή τους και να αποκρύψει πεδία.

Για τη δημιουργία διαγραμμάτων, επιλέγεται το επιθυμητό διάγραμμα από μία λίστα, που εκτός από τα κλασσικά γραφήματα περιλαμβάνει διάγραμμα κατανομής, διάγραμμα «καταρράκτη», συνδυαστικό ραβδογράμματος – ευθείας, δείκτη επίδοσης (KPI), μετρητή λόγου «gauge». Ο χρήστης αντιστοιχίζει με ακρίβεια το χώρο που θα καταλαμβάνει το διάγραμμα στο φύλλο, και στα κατάλληλα πλαίσια σύρει τα πεδία που θέλει να απεικονίσει. Σε κάθε διάγραμμα είναι προκαθορισμένο τι τύπου πεδία δεδομένων (ποσοτικά, μετά από πράξη, ή κατηγορικά) μπορούν να χρησιμοποιηθούν και σε τι πλήθος. Για τα ποσοτικά μεγέθη, που μπορούν να αναπαρασταθούν μόνο σαν τιμές αριθμητικών πράξεων, οι πράξεις που διατίθενται είναι άθροισμα, αρίθμηση, μ.ο., μέγιστο και ελάχιστο. Οποιοσδήποτε άλλος υπολογισμός, αριθμητικό ή στατιστικό μέγεθος, δημιουργείται ως «Master Item» χρήσει συναρτήσεων και εντολών. Ό,τι ισχύει για την κατασκευή των διαγραμμάτων, ισχύει και για τη δημιουργία περιστρεφόμενου πίνακα.

Όσον αφορά τη διαμόρφωση των διαγραμμάτων, είναι δυνατή η ταξινόμηση των απεικονισθέντων μεγεθών, η διαμόρφωση των αξόνων και της θέσης τους, η διαχείριση των κενών τιμών, η χάραξη γραμμών (μέσω συναρτήσεων), ο προσανατολισμός οριζόντιος, ή,

κάθετος. Η αντιστοίχιση χρωμάτων γίνεται αυτόματα, ή με βάση την τιμή κάποιου ποσοτικού μεγέθους- διαφορετικά απαιτεί εντολές και συναρτήσεις για την αντιστοίχιση κατ' επιλογή.

Οι δυνατότητες φιλτραρίσματος είναι ανύπαρκτες κατά την κατασκευή του διαγράμματος, εξαιρουμένης της δυνατότητας περιορισμού κάποιων ποσοτικών μεγεθών. Το φιλτράρισμα γίνεται με το βοήθημα «filter pane» -εργαλείο αντίστοιχο με το «slicer tool» του MS Power BI- κατά την παρατήρηση των διαγραμμάτων στα «φύλλα». Το φιλτράρισμα που γίνεται με την επιλογή τιμών ξεχωριστά, είναι εξυπηρετικό μόνο για κατηγορικές τιμές, και μάλιστα περιορισμένου αριθμού.

Στα «φύλλα» του εργαλείου, εμφανίζονται τα διαγράμματα, μαζί με ότι έχουμε επιλέξει να τα συνοδεύουν, όπως πίνακες, κείμενο και εικόνες. Με την επιλογή κάποιου πεδίου, είτε από τις πλευρικές λεζάντες, είτε πάνω στο διάγραμμα, αυτό τονίζεται σε όλα τα σχήματα του φύλλου. Τα «φύλλα» μπορούν να εξαχθούν άμεσα από το εργαλείο σε pdf. Στο Qlik Sense, όλα τα αναλυτικά στοιχεία μπορούν να παρουσιαστούν μέσω των «stories», παρουσιάσεων οργανωμένων σε διαφάνειες. Μπορούν να προστεθούν σχόλια κειμένου, σχήματα, φωτογραφίες και «snapshots» στιγμιότυπα μέσα από την εφαρμογή. Εξαγωγή των «stories» μπορεί να γίνει σε pdf και σε powerpoint.

Σε γενικές γραμμές, το εργαλείο είναι απλό στη χρήση και παράγει γρήγορα αποτελέσματα, όσον αφορά όμως απλές περιπτώσεις αναλυτικών, χωρίς ιδιαίτερες απαιτήσεις. Για τη δημιουργία πιο πολύπλοκων διαγραμμάτων, νέων πεδίων δεδομένων, λεπτομερούς μορφοποίησης, επαρκούς φιλτραρίσματος και προβλέψεων απαιτούνται εντολές και χρήση των συμβατών με το εργαλείο συναρτήσεων. Μάλιστα, υπάρχει και η πλατφόρμα ανάπτυξης «Qlik Analytics Platform», που επιτρέπει την κατασκευή εφαρμογών, βασισμένων σε front-end και back-end APIs, δίνοντας πλήρη πρόσβαση στις δυνατότητες του εργαλείου. Επίσης, η σύνδεση με επιπλέον επεκτάσεις (extensions) που έχουν σχεδιαστεί για το εργαλείο επιτρέπει λειτουργίες όπως νέα διαγράμματα, δυνατότητες εξεργασίας δεδομένων, μεταβολή των τιμών των στοιχείων, προσθήκη βίντεο του «YouTube» στις παρουσιάσεις «stories» και βιβλιοθήκες συναρτήσεων.

2.5.4 Πίνακας Επιπρόσθετων Χαρακτηριστικών

	Tableau Public 10.4	MS PowerBI	Qlik Sense
forecasting			
forecast length	auto/exactly/until	✓	✗
source data	auto/ years-seconds	seasonality points	✗
forecast model	exp.smoothing	✗	✗
model options	auto/without season./custom	✗	✗
% prediction intervals	90,95,99 %	✓	✗
trend lines			
linear	✓	✓	✗
logarithmic	✓	✗	✗
exponential	✓	✗	✗
polynomial	✓	✗	✗
p value	✓	✗	✗
r-squared value	✓	✗	✗
described trend line	✓	✗	✗
charts			
bar chart (horizontal)	✓	✓	✓
stacked bars	✓	✓	✓
100% stacked bars	✗	✓	✗
side-by-side bars/ clustered columns	✓	✓	✓
text tables	✓	✓	✓
highlight tables	✓	✗	✗
heat maps	✓	✗	✗
pie chart	✓	✓	✓
donut chart	✗	✓	✗
symbol maps	✓	✗	✗
treemaps	✓	✓	✓
circle views	✓	✗	✗
side-by-side circles	✓	✗	✗
lines (continuous)	✓	✗	✓
lines (discrete)	✓	✓	✓
dual lines (non-synchronized axis)	✓	✗	✗
area charts (continuous)	✓	✗	✓
area charts (discrete)	✓	✓	✓
dual combination /line and column chart	✓	✓	✓
stacked area chart	✗	✓	✓
waterfall chart	✗	✓	✓
scatter plots	✓	✓	✓
histogram	✓	✗	✓
box-and-whisker plots	✓	✗	✓
funnel	✗	✓	✗
gantt	✓	✗	✗
gauge	✗	✓	✓
bullet graphs	✓	✗	✗
packed bubbles	✓	✗	✗
card	✗	✓	✓
multi-row card	✗	✓	✓
KPI (key performance indicator)	✗	✓	✓
matrix	✗	✓	✗
distribution plot	✗	✗	✓

filters			
dimension			
specific values (general)	✓	✓	✓
starting with (wildcard)	✓	✓	✗
does not start/exclude	✓	✓	✗
conditional (e.g. sum)	✓	✓	✗
top filter (certain top values)	✓	✓	✓
contains	✓	✓	✗
does not contain/exclude	✓	✓	✗
measure			
at least	✓	✓	✓
at most	✓	✓	✓
range	✓	✓	✗
special for null values	✓	✓	✓
relative value filter	✗	✗	✓
date	✓	✓	✗
filter level			
visual	_	✓	_
page/worksheet	✓	✓	✓
report/dashboard	✓	✓	✓

Πίνακας 2-2 – λεπτομερής σύγκριση χαρακτηριστικών για τα 3 επικρατέστερα εργαλεία

Στον παραπάνω πίνακα (Πίνακας 2-2), συγκρίνονται τα τρία επικρατέστερα προγράμματα ως προς κάποια πιο λεπτομερή χαρακτηριστικά και εμβαθύνουμε στις προβλέψεις, τις ευθείες τάσης, τους τύπους των διαγραμμάτων και τις δυνατότητες φιλτραρίσματος.

3 Σενάρια Σύγκρισης

3.1 Στόχος των σεναρίων

Όπως έγινε φανερό από το προηγούμενο κεφάλαιο, τα κυρίαρχα εργαλεία στην κατηγορία επεξεργασίας και οπτικοποίησης δεδομένων έχουν αφενός πολλές κοινές δυνατότητες και παρόμοια χαρακτηριστικά, και αφετέρου το ίδιο πολλά χαρακτηριστικά, τρόπους προσέγγισης εργασιών και περιβάλλον, που τα καθιστούν πολύ διαφορετικά. Επομένως, για να συγκρίνουμε τα εργαλεία, τα προτερήματα και τις περιπτώσεις που υστερούν, επιλέξαμε να τα χρησιμοποιήσουμε για τη διεκπεραίωση κοινών εργασιών. Με τη σύγκριση στην επεξεργασία των ίδιων δεδομένων και τη δημιουργία ίδιων διαγραμμάτων, στοχεύουμε στον πρακτικό έλεγχο των δυνατοτήτων τους, καθώς και στις αδυναμίες και μειονεκτήματα που εμφανίζουν, τόσο αυτοτελώς όσο και σε σχέση με τα υπόλοιπα εργαλεία. Σε ορισμένα σενάρια, επιλέξαμε να συγκρίνουμε και τη χρήση των εργαλείων σε σχέση με μια υλοποίηση με χρήση μιας γλώσσας προγραμματισμού, της Python.

Έγινε, επομένως, η επιλογή κάποιων σεναρίων, με συγκεκριμένα ζητούμενα αποτελέσματα από τα εργαλεία, με χρήση των ίδιων αρχείων δεδομένων. Γίνεται έτσι λεπτομερώς αντιληπτός ο τρόπος λειτουργίας για το στάδιο της σύνδεσης με την πηγή των δεδομένων, την εισαγωγή, τον μετασχηματισμό και τη φόρτωση των δεδομένων, την υλοποίηση των διαγραμμάτων με την επιθυμητή μορφή και την παρουσίαση αυτών. Οι μεταξύ τους διαφορές για κάθε στάδιο, μας διευκολύνουν στον εντοπισμό ελλείψεων στη λειτουργία των εργαλείων.

Η επιλογή των σεναρίων έγινε με στόχο την εξέταση της απόδοσης των εργαλείων για διάφορους τύπους δεδομένων. Τα σενάρια περιλαμβάνουν αριθμητικά δεδομένα, χρονοσειρές, και γεωγραφικά δεδομένα. Κατά την υλοποίηση τους, βέβαια, μπορεί να περιλαμβάνονται οποιαδήποτε άλλα στοιχεία που θεωρείται σημαντική η επεξεργασία και απεικόνισή τους.

3.2 Σενάριο χρονοσειράς

Το σενάριο αυτό επικεντρώνεται σε τιμές που αντιστοιχούν σε άξονες ημερομηνιών, δηλαδή χρονοσειρές. Για τα διαγράμματα χρονοσειρών, μας ενδιαφέρει η δυνατότητα διαμόρφωσης του άξονα σε χρονική διαβάθμιση εύρους από ημέρες μέχρι και έτη. Σημαντικό είναι επίσης εάν ο άξονας μπορεί να αναπαρασταθεί με συνεχείς και διακριτές χρονολογικές τιμές. Κατά την υλοποίηση του σεναρίου θα καταγράψουμε και τις όποιες επιλογές οργάνωσης των ημερομηνιών προσφέρει το κάθε εργαλείο, π.χ. χρονικός άξονας εβδομάδων. Περιέχει ωστόσο και αριθμητικά δεδομένα, όπως ποσοστά.

Το σενάριο, αφορά τις δημοσκοπήσεις για τις προεδρικές εκλογές των Η.Π.Α για το 2012, μεταξύ των δύο υποψηφίων. Οι παρατηρήσεις αφορούν την περίοδο από τον Ιανουάριο του 2009 μέχρι και το Νοέμβριο του 2012. Έχει σημασία να επισημάνουμε ότι το σενάριο είναι έργο ανάλυσης δεδομένων που αλλιεύτηκε από το http://nbviewer.jupyter.org /github/jmportilla/Udemy-notes/blob/master/Data%20Project%20-%20Election%20Analysis.

ipynb ενώ τα δεδομένα του είναι διαθέσιμα στο HuffPost *Pollster* [30]. Είναι υλοποιημένο με τη χρήση της γλώσσας Python, και κατάλληλες βιβλιοθήκες, όπως η python pandas. Ο κώδικας του σεναρίου υπάρχει στο παράρτημα της εργασίας (Εντολές σεναρίου χρονοσειράς). Εδώ γίνεται περιγραφή του σεναρίου, ενώ παρατίθενται και κάποιες αντιπροσωπευτικές εντολές.

Αρχικά, γίνεται η εισαγωγή των απαραίτητων βιβλιοθηκών για την επεξεργασία δεδομένων και τη δημιουργία διαγραμμάτων. Στη συνέχεια γίνεται η εισαγωγή των δεδομένων και μία επισκόπηση αυτών (Εικόνα 3-1).

```
source = requests.get(url).text
```

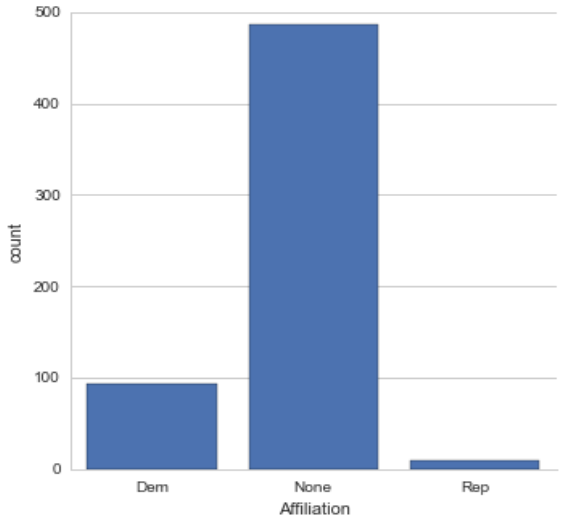
	Pollster	Start Date	End Date	Entry Date/Time (ET)	Number of Observations	Population	Mode
0	Politico/GWU /Battleground	2012-11-04	2012-11-05	2012-11-06 2000-01-01 08:40:26 UTC	1000	Likely Voters	Live Phone
1	UPI/CVOTER	2012-11-03	2012-11-05	2012-11-05 2000-01-01 18:30:15 UTC	3000	Likely Voters	Live Phone
2	Gravis Marketing	2012-11-03	2012-11-05	2012-11-06 2000-01-01 09:22:02 UTC	872	Likely Voters	Automated Phone
3	JZ Analytics/Newsmax	2012-11-03	2012-11-05	2012-11-06 2000-01-01 07:38:41 UTC	1041	Likely Voters	Internet
4	Rasmussen	2012-11-03	2012-11-05	2012-11-06 2000-01-01 08:47:50 UTC	1500	Likely Voters	Automated Phone

Εικόνα 3-1 – επισκόπηση δεδομένων

Το πρώτο διάγραμμα για το πεδίο «affiliation» , σε σχέση με την αρίθμηση (Εικόνα 3-2) :

```
# Factorplot the affiliation
sns.factorplot('Affiliation',data=poll_df)
```

```
In [609]: # Factorplot the affiliation
          sns.factorplot('Affiliation',data=poll_df)

Out[609]: <seaborn.axisgrid.FacetGrid at 0xb9bd37b8>
```

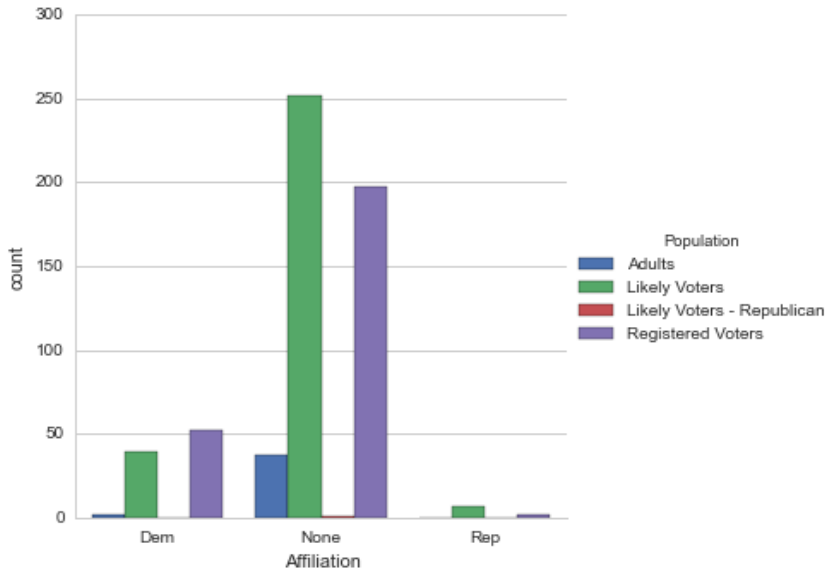


Εικόνα 3-2 – ραβδόγραμμα για το πεδίο υπαγωγής σε παράταξη «affiliation»

Φαίνεται ότι οι ψηφοφόροι που λαμβάνουν μέρος είναι επί το πλείστον κομματικά ουδέτεροι. Έπειτα, δημιουργείται συνδυαστικό ραβδόγραμμα συστοιχισμένων στηλών για τα πεδία «affiliation» και «population» (Εικόνα 3-3).

```
sns.factorplot('Affiliation',data=poll_df,hue='Population')
```

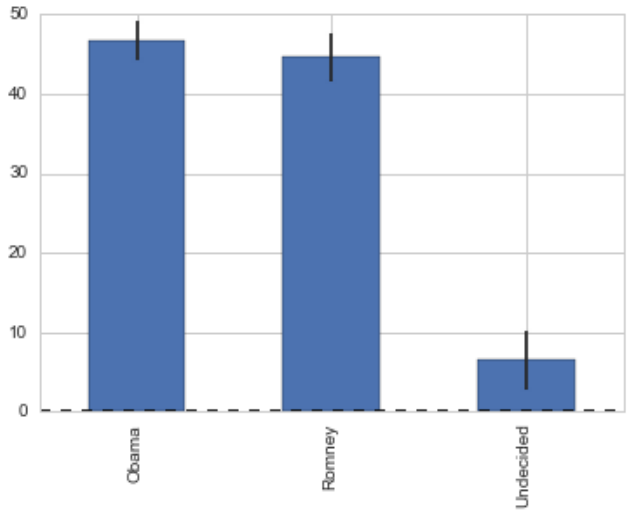
```
In [610]: # Factorplot the affiliation by Population
          sns.factorplot('Affiliation',data=poll_df,hue='Population')

Out[610]: <seaborn.axisgrid.FacetGrid at 0xcbc90fd0>
```



Εικόνα 3-3- διάγραμμα συστοιχισμένων στηλών για δύο διαφορετικά πεδία

Οι κατηγορίες που εμφανίζονται συχνότερα είναι οι πιθανοί και οι «δηλωμένοι» ψηφοφόροι. Έπειτα, θα αναπαραστήσουμε σε διάγραμμα τον μ.ο. των ποσοστών που προβλέπουν οι δημοσκοπήσεις (Εικόνα 3-4).

```
avg = pd.DataFrame(poll_df.mean())
```



Εικόνα 3-4 – διάγραμμα μ.ο πρόθεσης ψήφου. ανά υποψήφιο

Είναι ενδιαφέρον το πόσο κοντά είναι οι μ.ο. των δύο υποψηφίων, δεδομένου του μικρού μ.ο. των αναποφάσιστων. Στη συνέχεια, υπολογίζουμε τους μ.ο. και την τυπική απόκλιση των τριών πεδίων (Πίνακας 3-1).

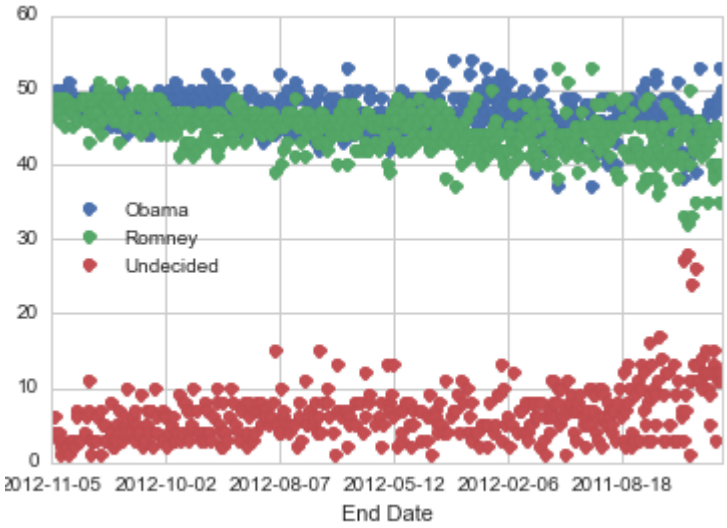
```
poll_avg = pd.concat([avg,std],axis=1)
```

	Average	STD
Obama	46.772496	2.448627
Romney	44.573854	2.927711
Undecided	6.549763	3.702235

Πίνακας 3-1 – μ.ο. και τυπική απόκλιση ανά υποψήφιο

Ακολουθεί μία χρονοσειρά με τα αποτελέσματα των δημοσκοπήσεων, από τον Αύγουστο του 2011 μέχρι το Νοέμβρη του 2012, με ανεστραμμένο άξονα (Εικόνα 3-5).

```
poll_df.plot(x='End Date',y=['Obama','Romney','Undecided'],marker='o',linestyle='')
```



Εικόνα 3-5 – διάγραμμα αναπαράστασης των αποτελεσμάτων των δημοσκοπήσεων στον άξονα του χρόνου

Το διάγραμμα που απεικονίζει όλα τα ποσοστά που αντιστοιχούν στις τρεις επιλογές της δημοσκόπησης δεν είναι ευκατανόητο. Ίσως η αναπαράσταση της διαφοράς μεταξύ των υποψηφίων να οδηγεί σε πιο ξεκάθαρη εικόνα. Θα δημιουργήσουμε μία νέα στήλη «Difference», με την ποσοστιαία διαφορά των δύο υποψηφίων (Εικόνα 3-6). Το θετικό πρόσημο στη δημιουργηθείσα στήλη, υποδηλώνει προβάδισμα του Obama.

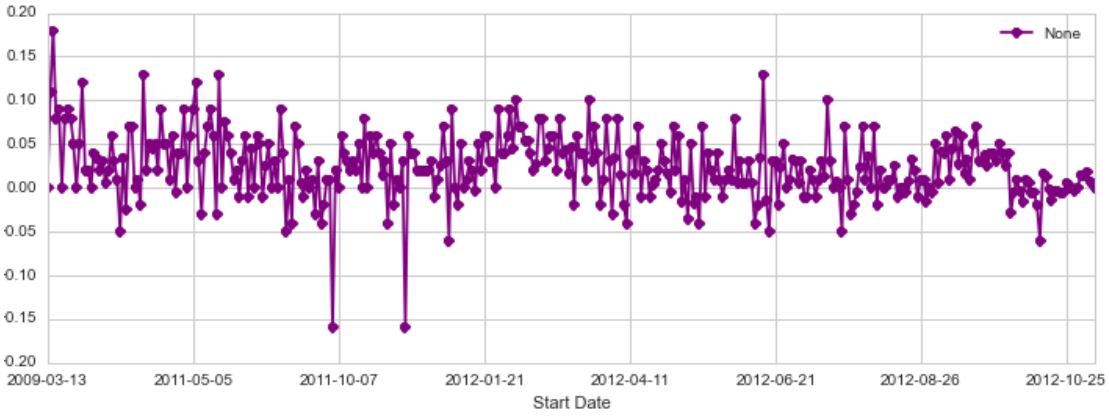
```
poll_df['Difference'] = (poll_df.Obama - poll_df.Romney)/100
```

	Start Date	Number of Observations	Obama	Romney	Undecided	Difference
0	2009-03-13	1403	44	44	12	0.00
1	2009-04-17	686	50	39	11	0.11
2	2009-05-14	1000	53	35	12	0.18
3	2009-06-12	638	48	40	12	0.08
4	2009-07-15	577	49	40	11	0.09

Εικόνα 3-6 – νέα στήλη Difference, για την ποσοστιαία διαφορά μεταξύ των υποψηφίων

Ο σχεδιασμός της χρονοσειράς για το πεδίο «Difference» που έχει υπολογισθεί (Εικόνα 3-7):

```
poll_df = poll_df.groupby(['Start Date'],as_index=False).mean()
```



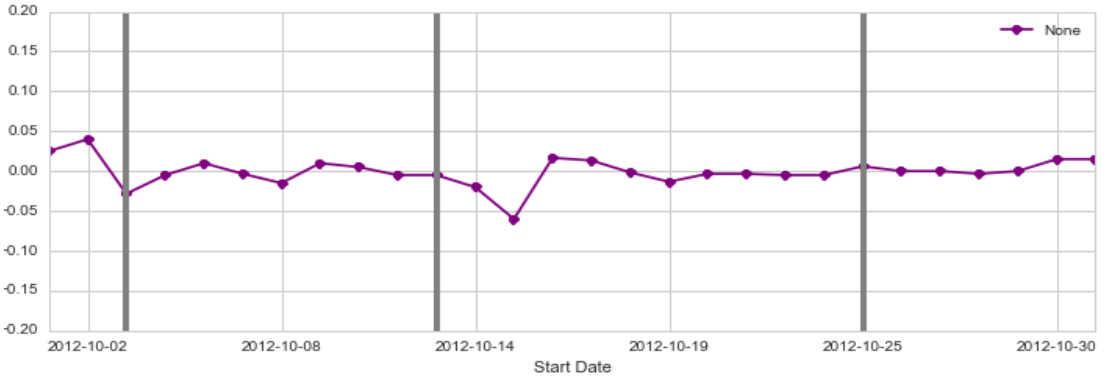
Εικόνα 3-7 – χρονοσειρά για το πεδίο Difference

Θα είχε ενδιαφέρον να σχεδιαστούν ευθείες που εφιστούν την προσοχή στις ημερομηνίες που διεξήχθησαν «debates», στις 3, 11 και 22 Οκτώβρη του 2012.Απαιτείται ο εντοπισμός του μήνα, στην στήλη του πεδίου τους , με αναζήτηση μέσω επανάληψης «for loop».

```
for date in poll_df['Start Date']:
    if date[0:7] == '2012-10':
        xlimit.append(row_in)
        row_in +=1
    else:
        row_in += 1
```

Αφού βρεθεί η πρώτη ημερομηνία που αντιστοιχεί στον Οκτώβρη του 2012, προστίθενται οι τιμές +2, +10 και +21, ώστε να χαρακτούν οι ευθείες που σηματοδοτούν τα «debates». Έπεται ο σχεδιασμός του διαγράμματος με τις ευθείες (Εικόνα 3-8):

```
fig = poll_df.plot('Start Date','Difference',figsize=(12,4),marker='o',linestyle='-
',color='purple',xlim=(329,356))
```



Εικόνα 3-8 – εστίαση στο μήνα Οκτώβρη του 2012

Το διάγραμμα αποκαλύπτει μία «βύθιση» για τον Obama, επακόλουθη του δεύτερου debate, παρόλο που τα πήγε πιο άσχημα στο πρώτο.

3.2.1 Tableau

Βήμα 1ο: Εισαγωγή αρχείου csv
source = requests.get(url).text
* οι εντολές περιγραφής κάτω από τον υπότιτλο κάθε βήματος είναι ενδεικτικές, συνήθως απαιτούνται πολύ περισσότερες για την κάθε υλοποίηση, όπως διακρίνεται και από το script της python.

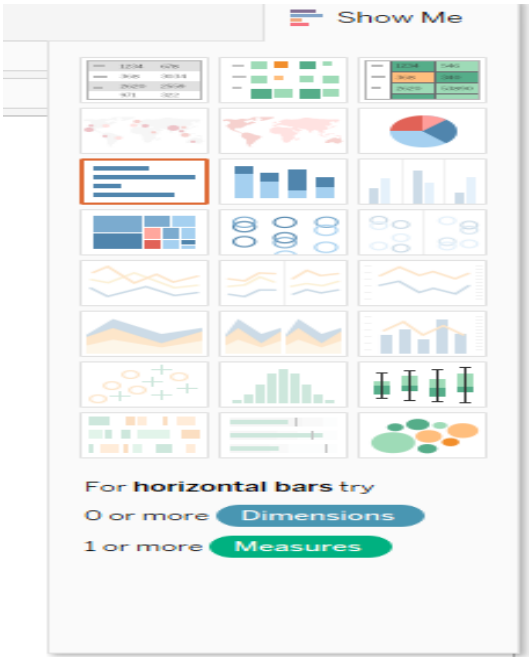
Από το παράθυρο εισαγωγής δεδομένων, επιλέγουμε εισαγωγή από αρχείο κειμένου. Κατά την εισαγωγή των δεδομένων, ο Data Interpreter δεν μπορεί να επεξεργαστεί εν τη περιπτώσει το csv, οπότε με ρύθμιση των text file properties γίνεται ο διαχωρισμός των στοιχείων και η οργάνωσή τους σε στήλες, και αυτόματα αναγνωρίζεται ότι η πρώτη σειρά αποτελεί επικεφαλίδα. Εδώ οι σειρές είναι 586, όπως αναγράφεται στον καταμετρητή. Γενικά, εμφανίζονται μέχρι 1000 σειρές αυτόματα, και μπορούμε να ρυθμίσουμε να εμφανίζονται μέχρι 1.000.000 .

Επιλέγουμε Extracted Data, ώστε να γίνονται στο βέλτιστο χρόνο οι ενέργειες που επιθυμούμε. Εάν τα δεδομένα ανανεώνονται συνεχώς, μπορούμε να επιλέξουμε live connection (δεν ανανεώνονται αυτόματα και στο dashboard στο Tableau Desktop).

Βήμα 2ο: Διάγραμμα κομματικής προτίμησης συμμετεχόντων
sns.factorplot('Affiliation',data=poll_df)

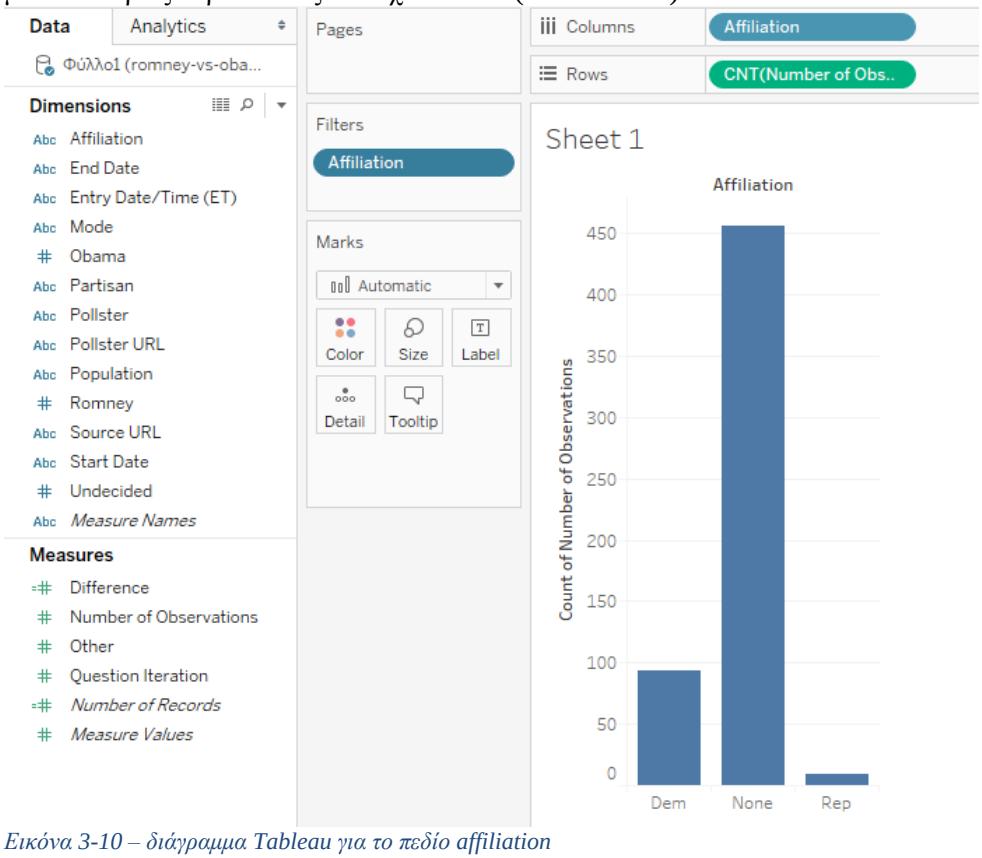
Μεταβαίνουμε στα sheets, όπου χρησιμοποιούμε ένα για κάθε γράφημα. Τα headers των δεδομένων είναι διαχωρισμένα αυτόματα ανάλογα με τον τύπο τους (string, whole number, etc.), στις κατηγορίες Dimensions (blue pill) –τοποθετούνται τα διακριτά μεγέθη και Measures (green pill) – τα συνεχή. Η κατηγοριοποίηση έχει μεγάλη σημασία για την υλοποίηση των γραφημάτων. Δίνεται η δυνατότητα να αλλάζει η κατηγορία που ανήκει κάθε στήλη δεδομένων.

Η δημιουργία γραφημάτων βασίζεται στο απλό drag&drop. Στο παράθυρο του εργαλείου, εμφανίζεται ένα toolbar, που εμφανίζει τις επιλογές γραφημάτων, τα ελάχιστα δεδομένα από κάθε κατηγορία που είναι απαραίτητα για κάθε γράφημα, και προτείνει και ένα ή περισσότερα από αυτά εάν έχουμε «σύρει» κάποια δεδομένα προς απεικόνιση (Εικόνα 3-9).



Εικόνα 3-9 – παράθυρο προτεινόμενων γραφημάτων

Έτσι, από τα measures σύρουμε στα rows τη στήλη Number of Observations, με επιλογή αρίθμησης count (μεταξύ sum, count, average, min, max, median, variance, std.deviation, percentile). Σύρουμε το affiliation στα columns και filters, όπου επιλέγουμε να εμφανίζονται μόνον οι τρεις περιπτώσεις και όχι «other» (Εικόνα 3-10).



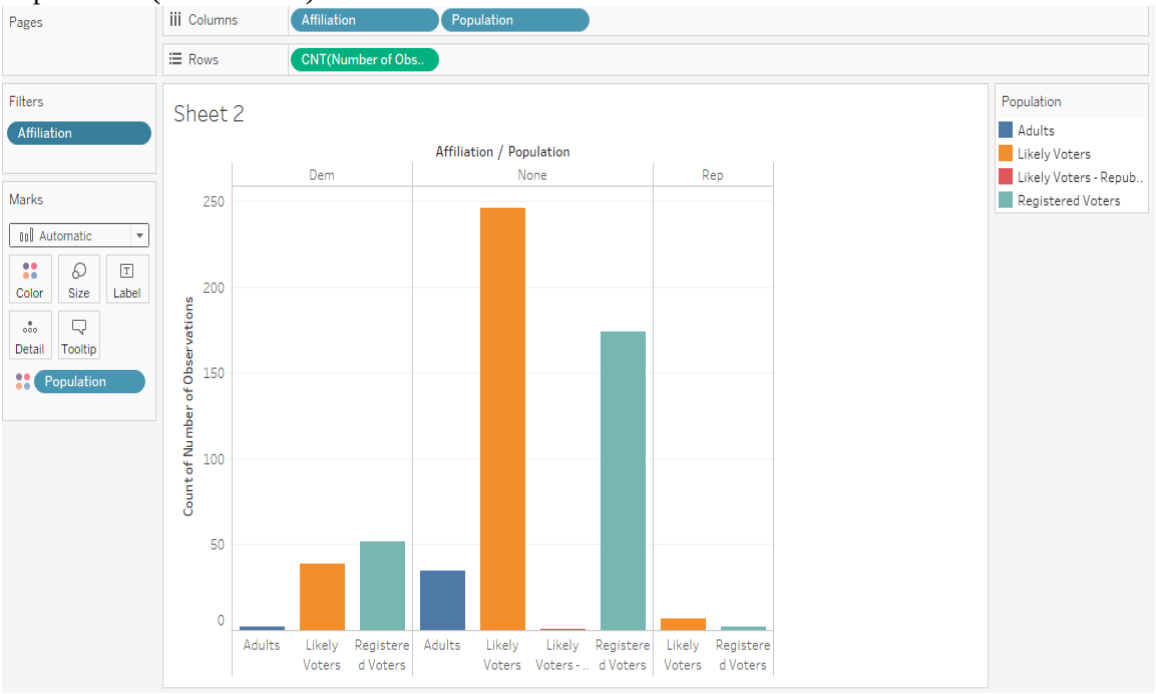
Εικόνα 3-10 – διάγραμμα Tableau για το πεδίο affiliation

Τα φίλτρα διαφέρουν για τα measures, όπου είναι ποσοτικά και επιλέγουμε range, ελάχιστο ή μέγιστο. Στα φίλτρα για τα dimensions μπορούμε να κάνουμε συγκεκριμένη επιλογή (general), να εμφανίσουμε τα κορυφαία στο πεδίο που θέλουμε (top), να εμφανίσουμε όταν ικανοποιείται κάποια προϋπόθεση για όποιο πεδίο ενδιαφερόμαστε (condition), και τέλος να περιέχει /ξεκινάει/καταλήγει με συγκεκριμένα στοιχεία (wildcard).

Βήμα 3ο: Διάγραμμα κομματικής προτίμησης ανά πληθυσμιακή κατηγορία

```
sns.factorplot('Affiliation',data=poll_df,hue='Population')
```

Το ίδιο εύκολα και το διάγραμμα , που συνδυάζει στις στήλες (ποιοτικές τιμές) και το πεδίο « Population» (Εικόνα 3-11):

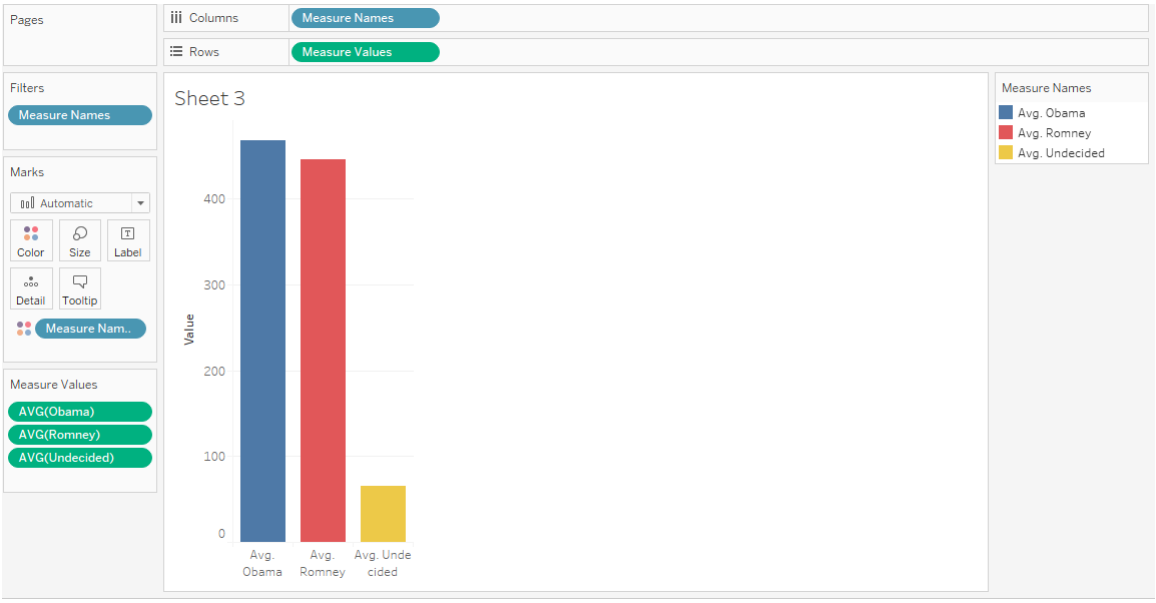


Εικόνα 3-11 – διάγραμμα συστοιχισμένων στηλών του Tableau για τα πεδία affiliation και population

Βήμα 4ο: Διάγραμμα μ.ο. των τριών πεδίων

```
avg = pd.DataFrame(poll_df.mean())
```

Σε αυτή την περίπτωση, στο Tableau, τα πεδία που πρέπει να επιλέξουμε φαίνεται να είναι πιο περίπλοκα. Ωστόσο, τοποθετώντας τα 3 πεδία (Obama, Romney, Undecided) στις columns, δημιουργούνται 3 διαφορετικοί άξονες-γραφήματα, που, μετά από drag&drop- σε παράλληλη στοίχιση, ενσωματώνονται σε ένα σχήμα. Στη συνέχεια, σύρουμε και τα Measure Names στο σύμβολο των χρωμάτων, ώστε να επιλέξουμε και το χρώμα κάθε στήλης (Εικόνα 3-12).

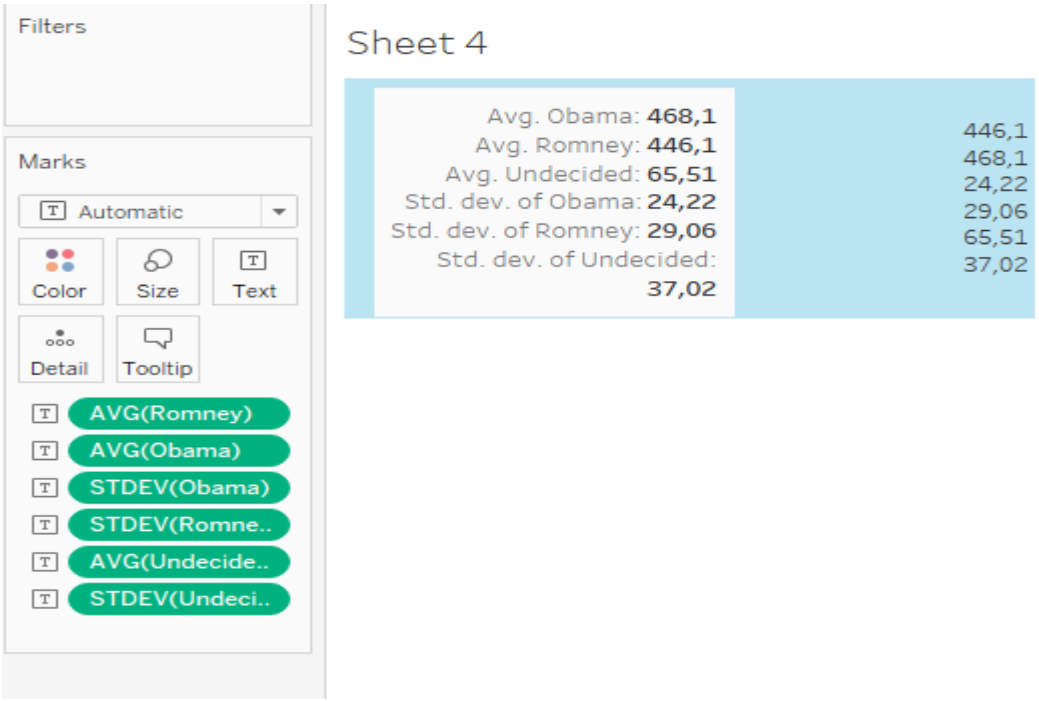


Εικόνα 3-12 – διάγραμμα Tableau για τους μ.ο. των υποψηφίων

Βήμα 5ο: Υπολογισμός μ.ο. και τυπικών αποκλίσεων

```
poll_avg = pd.concat([avg,std],axis=1)
```

Εδώ σύρουμε τα average και std.deviation measures στο σύμβολο του κειμένου, για να εμφανιστούν σαν κείμενο (Εικόνα 3-13). Με right-click -> view data, μπορούμε να εξάγουμε τα αποτελέσματα αυτά σε csv αρχείο.

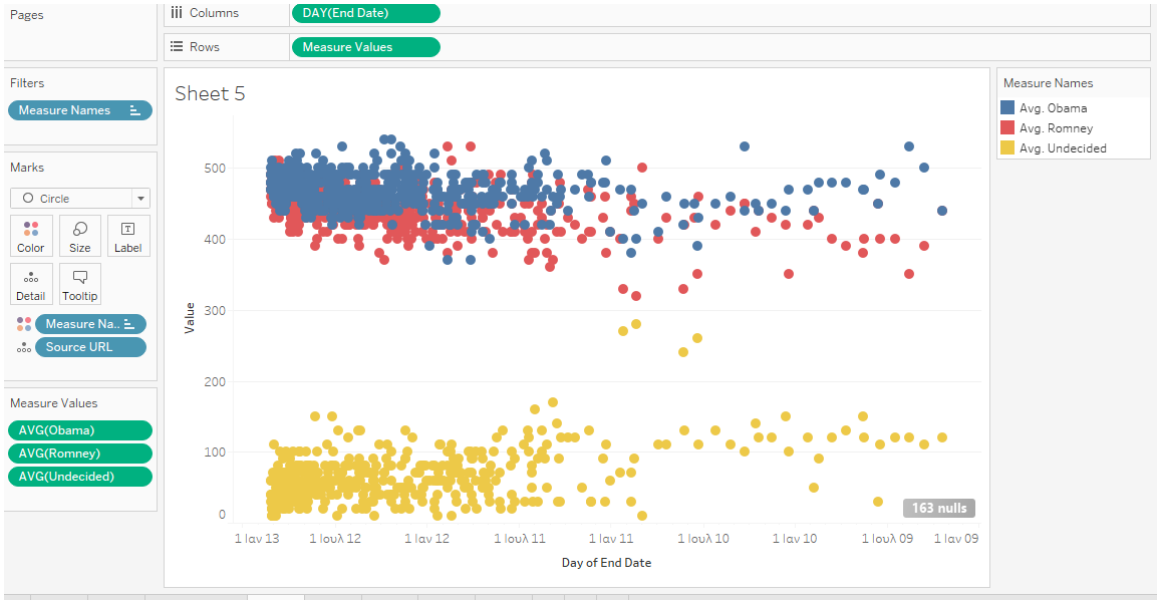


Εικόνα 3-13 – υπολογισμός μ.ο. και τυπικών αποκλίσεων

Βήμα 6ο: Γράφημα τιμών των προβλέψεων, ανεστραμμένου χρονικού άξονα

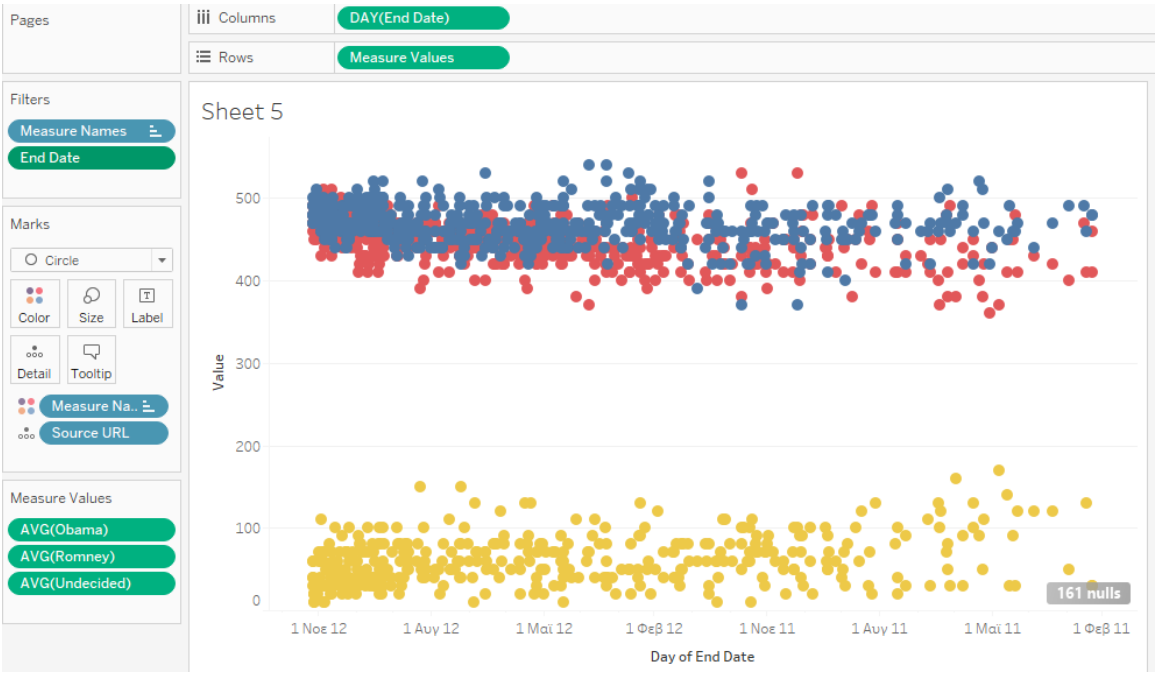
```
poll_df.plot(x='End Date',y=['Obama','Romney','Undecided'],marker='o',linestyle='')
```

Εδώ τοποθετήσαμε πάλι τα Measure Values στις σειρές και στις στήλες, στον άξονα X δηλαδή, τις ημερομηνίες End Dates. Την ημερομηνία επιλέγουμε να την προβάλλουμε συνεχή, στο επίπεδο των ημερών για όλο το χρονικό διάστημα. Επιλέγουμε «average» στα Measure Values. Έπειτα θα κάνουμε ένα τέχνασμα, αφού δεν προσφέρεται επιλογή απεικόνισης των τιμών. Σύρουμε στο σύμβολο του detail, το source url, κάτι που θα εμφανίζει μοναδικά κάθε «άθροισμα», αρκεί να είναι από διαφορετικό url για την ίδια ημερομηνία. Έτσι, εμφανίζονται οι τιμές και όχι το άθροισμά τους (Εικόνα 3-14), πλην μίας περίπτωσης, όπου δεδομένα είναι ανεβασμένα από κοινό url, οπότε θα απεικονίζει εκεί τον μ.ο. (κάτι που ουσιαστικά δεν αλλάζει το επιθυμητό γράφημα).



Εικόνα 3-14 – διάγραμμα Tableau για τα αποτελέσματα των δημοσκοπήσεων σε άξονα χρόνου

Οπότε, φιλτράροντας τις ημερομηνίες, να εμφανίζονται από το 2011 κ.ε., όπως στο πρωτότυπο (Εικόνα 3-15):



Εικόνα 3-15 – διάγραμμα Tableau για τα αποτελέσματα των δημοσκοπήσεων με φίλτρο ημερομηνιών

Το τέχνασμα αυτό προτάθηκε έπειτα από ερώτηση σε forum της community του tableau. Η ερώτηση απαντήθηκε άμεσα και βρέθηκε κάποιου είδους λύση. Η κοινότητα του tableau έχει μεγάλη συμμετοχή και μέλη που συστηματικά απαντούν σε ερωτήσεις.

Βήμα 7ο: Δημιουργία νέας στήλης «Difference»

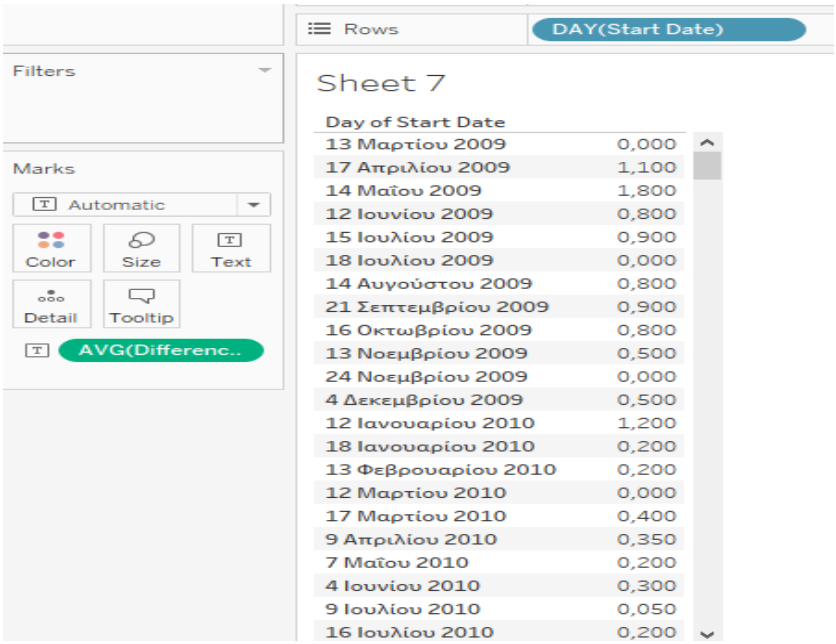
```
poll_df['Difference'] = (poll_df.Obama - poll_df.Romney)/100
```

Στο πεδίο Data Source, δημιουργούμε Calculated Field με τη διαφορά Obama – Romney προς 100 και ονομάζουμε difference τη νέα στήλη.

Βήμα 8ο: Γραφική αναπαράσταση των μ.ο.της «Difference» , στη χρονοσειρά με άξονα τη start date

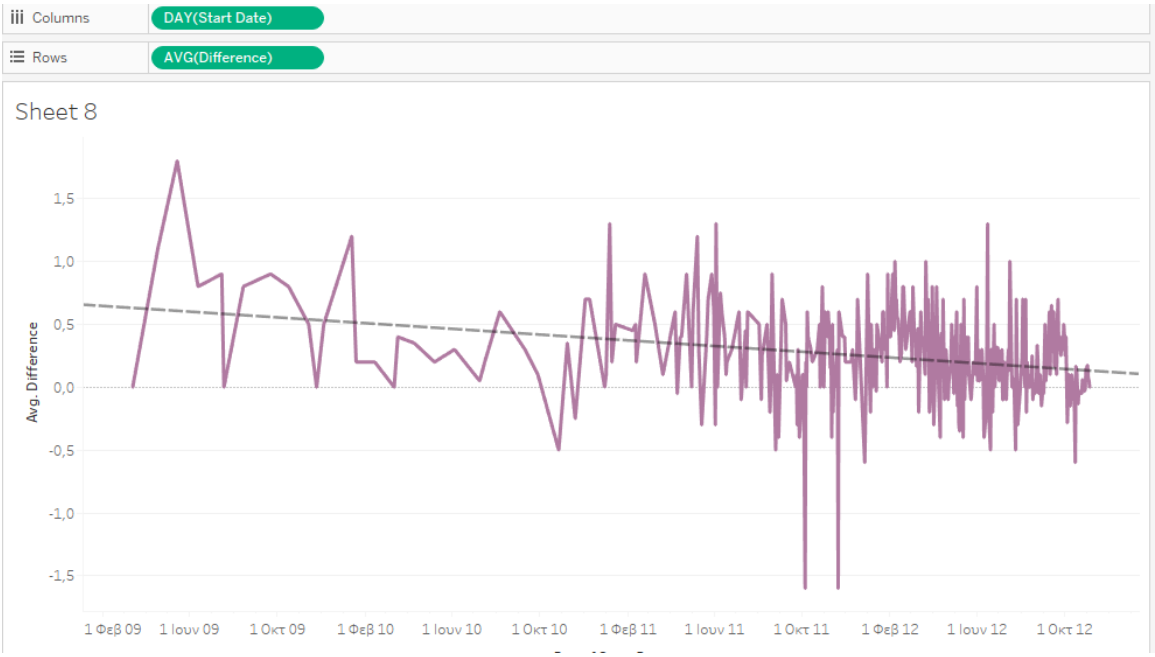
```
poll_df = poll_df.groupby(['Start Date'],as_index=False).mean()
```

Εδώ δε χρειάζεται κάποιο groupby στα δεδομένα. Υπολογίζουμε τον Μ.Ο. της Difference στο σύμβολο text, με Rows τα Start Dates (Εικόνα 3-16).



Εικόνα 3-16- Υπολογισμός του πεδίου Difference στο Tableau

Το διάγραμμα του πρωτοτύπου δεν έχει αναλογικά διατεταγμένες τις ημερομηνίες, αλλά βάσει των μετρήσεων που τους αντιστοιχούν. Χαράζουμε και την trend line στο γράφημα (Εικόνα 3-17).

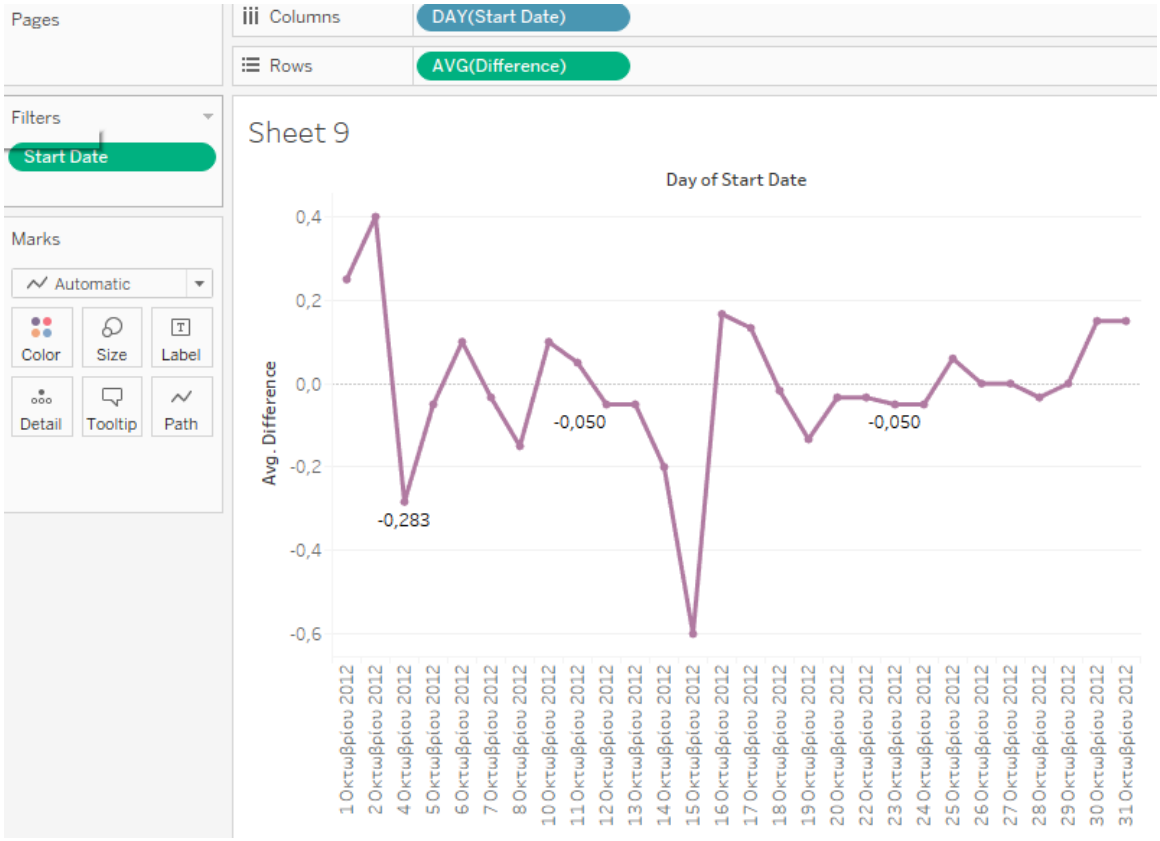


Εικόνα 3-17 Χρονοσειρά για το πεδίο Difference (μ.ο.) με ευθεία τάσης , στο Tableau

Βήμα 9ο: Εστίαση σε συγκεκριμένες ημερομηνίες, αλλαγή επιπέδου στην κλίμακα χρονοσειράς

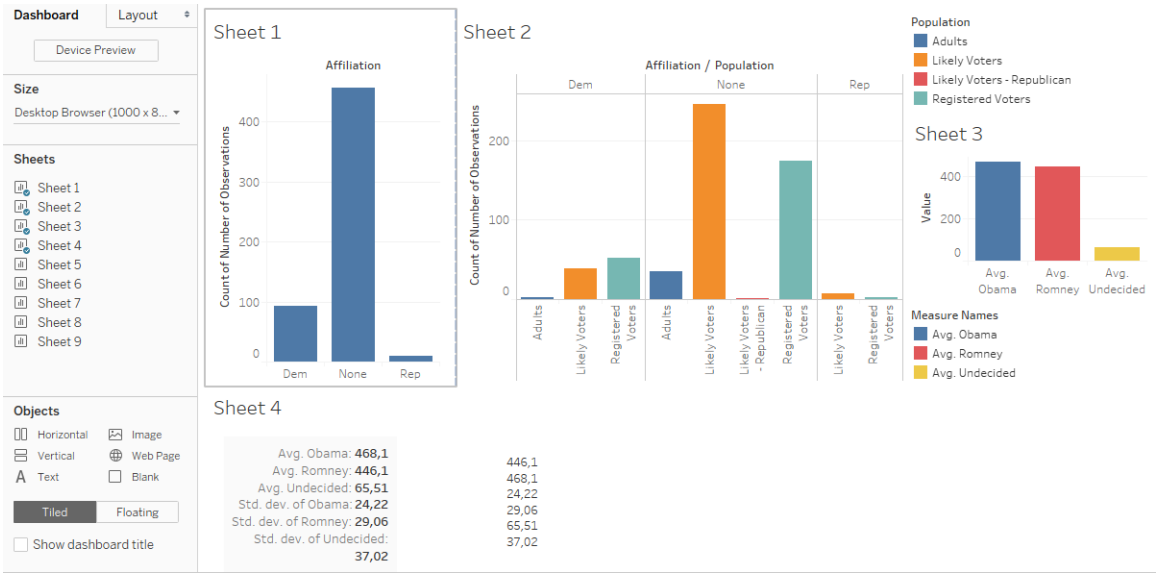
```
fig = poll_df.plot('Start Date','Difference',figsize=(12,4),marker='o',linestyle='-
',color='purple',xlim=(329,356))
```

Με τη χρήση του φίλτρου στις ημερομηνίες, εστιάζουμε στο μήνα Οκτώβριο του 2012. Δεν είναι δυνατόν να χαράξουμε τις κάθετες γραμμές επισήμανσης. Αντί αυτών, θα χρησιμοποιήσουμε την επιλογή «always show Mark Label» στα τρία σημεία που επιθυμούμε, ώστε να ξεχωρίζουν από τα υπόλοιπα (Εικόνα 3-18).

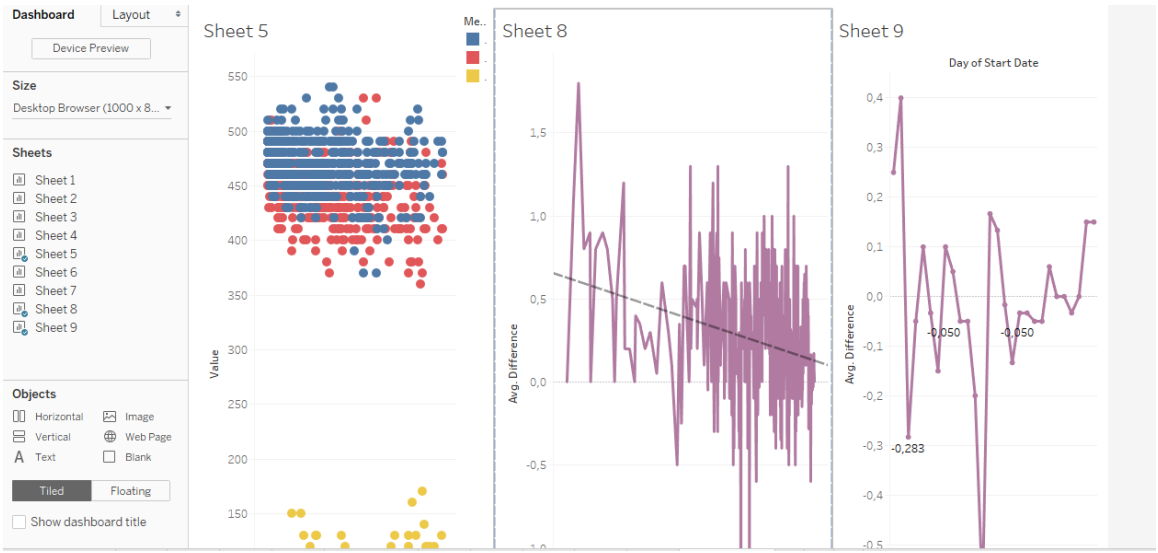


Εικόνα 3-18 – χρονοσειρά για το πεδίο Difference στο Tableau, με εστίαση στο μήνα Οκτώβριο

Τελικά, τα δύο dashboards που συμπεριλαμβάνουν τα παραπάνω γραφήματα (Εικόνα 3-19,Εικόνα 3-20).



Εικόνα 3-19 – dashboard 1 του Tableau, για το σενάριο χρονοσειράς



Εικόνα 3-20 – dashboard 2 του Tableau, για το σενάριο χρονοσειράς

3.2.2 MS Power BI

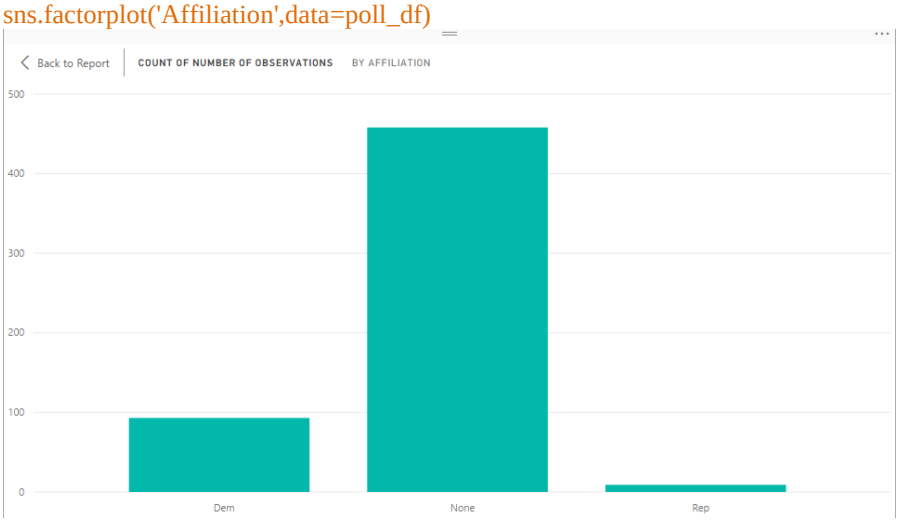
Βήμα 1^{ο}: Εισαγωγή αρχείου csv

source = requests.get(url).text

Στο MS Power BI δεν υπάρχει προφανώς κάποια ανάγκη για εγκατάσταση βιβλιοθηκών επεξεργασίας δεδομένων ή γραφημάτων, μιας και είναι ένα εργαλείο συγκεκριμένης χρήσης. Οπότε αρκεί το get data, επιλογή τύπου text/csv και επιλογή τύπου μεταβλητών «based on 200 first rows», μιας και δεν υπάρχει διαφορά με τις επόμενες σειρές (πχ ακέραιοι στις πρώτες, ενώ μετά και δεκαδικοί). Στο παράθυρο Data εμφανίζεται το data set.

Βήμα 2ο: Διάγραμμα κομματικής προτίμησης συμμετεχόντων

sns.factorplot('Affiliation',data=poll_df)



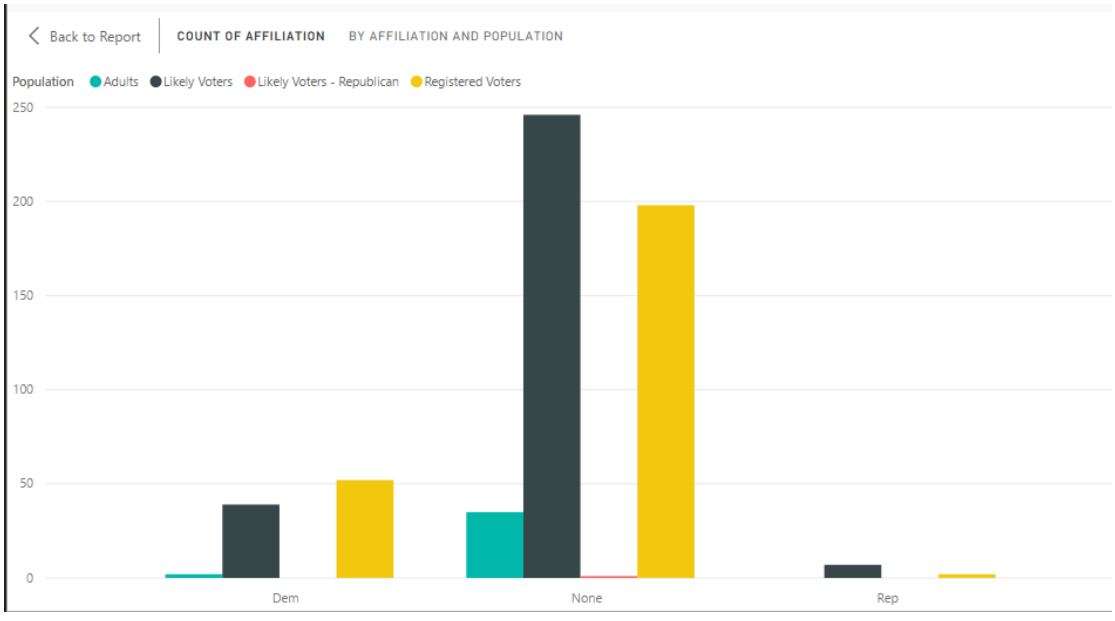
Εικόνα 3-21 - διάγραμμα Power BI για το πεδίο affiliation

Με το stacked column chart (Εικόνα 3-21), με το affiliation στον άξονα, όπου επιλέγουμε να μην εμφανίζεται η κατηγορία other (φίλτρο), και count του number of observations στο values.

Βήμα 3ο: Διάγραμμα κομματικής προτίμησης ανά πληθυσμιακή κατηγορία

sns.factorplot('Affiliation',data=poll_df,hue='Population')

Χρησιμοποιούμε clustered column charts, με το affiliation στον άξονα, το count του affiliation στα values, και το population στο legend, ώστε να εμφανίζονται και το προφίλ του πληθυσμού σε σχέση με την ψήφο τους (Εικόνα 3-22).

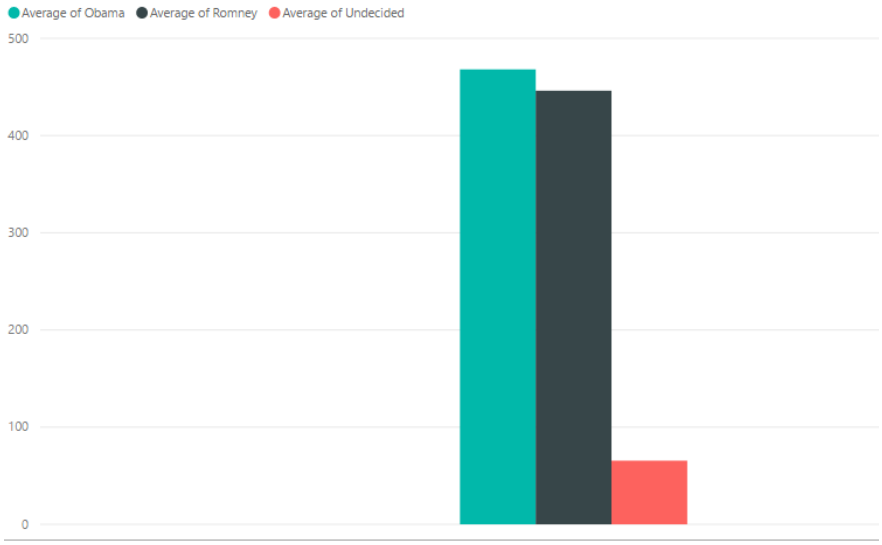


Εικόνα 3-22 - διάγραμμα συστοιχισμένων στηλών του Power BI για τα πεδία affiliation και population

Βήμα 4ο: Διάγραμμα μ.ο. των τριών πεδίων

```
avg = pd.DataFrame(poll_df.mean())
```

Για το average κάνουμε ένα clustered column chart με τους μ.ο. στα values (Εικόνα 3-23).



Εικόνα 3-23 - διάγραμμα Power BI για τους μ.ο. των υποψηφίων

Βήμα 5ο: Υπολογισμός μ.ο. και τυπικών αποκλίσεων

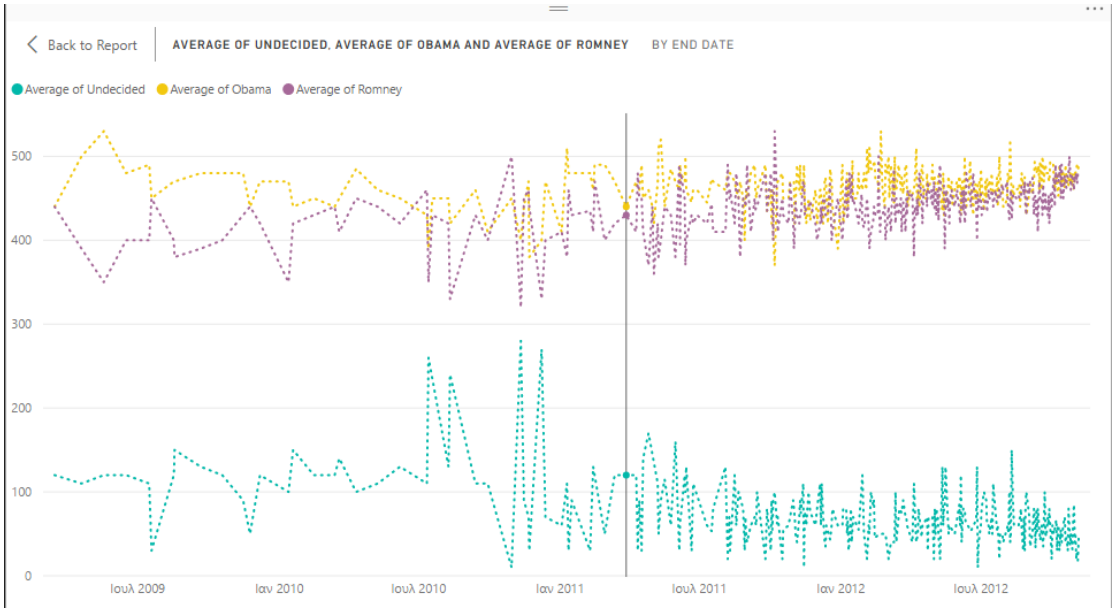
```
poll_avg = pd.concat([avg,std],axis=1)
```

Και παρουσιάζουμε σε δύο multi row cards τους μ.ο. και την τυπική απόκλιση, όπου ο μ.ο. για τον Obama είναι 468.05 ενώ για τον Romney 446.14. Αντίστοιχα, η τυπική απόκλιση 24.20 και 29.04 .

Βήμα 6ο: Γράφημα τιμών των προβλέψεων, ανεστραμμένου χρονικού άξονα

```
poll_df.plot(x='End Date',y=['Obama','Romney','Undecided'],marker='o',linestyle='')
```

Δημιουργούμε ένα linechart με άξονα τα end dates και τους μ.ο. των 3 μεταβλητών (Εικόνα 3-24).



Εικόνα 3-24 - διάγραμμα Power BI με τους μ.ο. των αποτελεσμάτων των δημοσκοπήσεων σε άξονα χρόνου

Στο πρωτότυπο, οι ημερομηνίες είναι ανάποδα (2012-2011), βλέπουμε πως η τάση είναι ίδια, αλλά φαίνονται όλα τα σημεία κάθε ημερομηνίας, ενώ στο Power BI δεν δίνεται η επιλογή απεικόνισης όλων των σημείων.

Βήμα 7ο: Δημιουργία νέας στήλης «Difference»

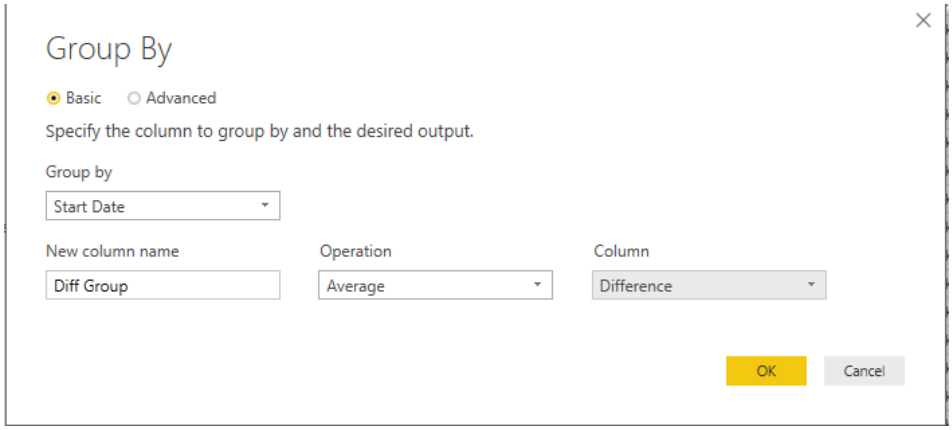
```
poll_df['Difference'] = (poll_df.Obama - poll_df.Romney)/100
```

Από το query editor δημιουργούμε νέα στήλη «difference», ίση με τη διαφορά των δύο στηλών προς 100.

Βήμα 8ο: Γραφική αναπαράσταση των μ.ο.της «Difference» , στη χρονοσειρά με άξονα τη start date

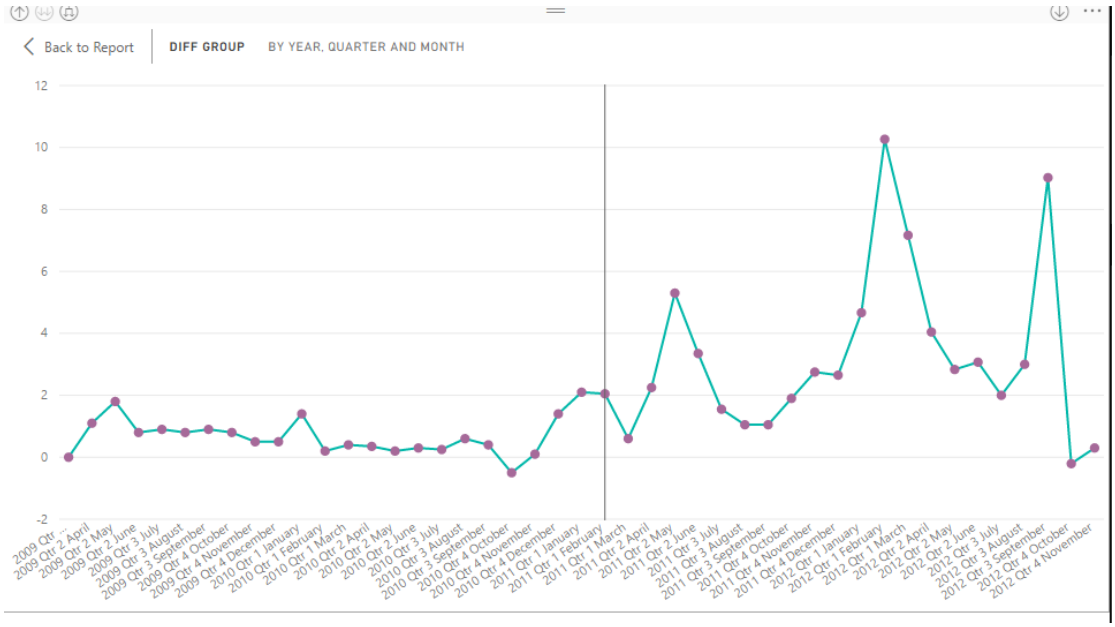
```
poll_df = poll_df.groupby(['Start Date'],as_index=False).mean()
```

Στη συνέχεια, η εντολή poll_df = poll_df.groupby(['Start Date'],as_index=False).mean(), πραγματοποιείται στο query editor, με την επιλογή group by, σε ανάλογο παράθυρο (Εικόνα 3-25).



Εικόνα 3-25 – η επιλογή Group by, του Query Editor στο Power BI

Όμως, το group by πρέπει να εκτελεστεί σε ένα duplicate data set, αφού θα αλλάξει ο αριθμός των σειρών. Εμφανίζουμε ένα line chart με values τη στήλη diff group (Εικόνα 3-26).



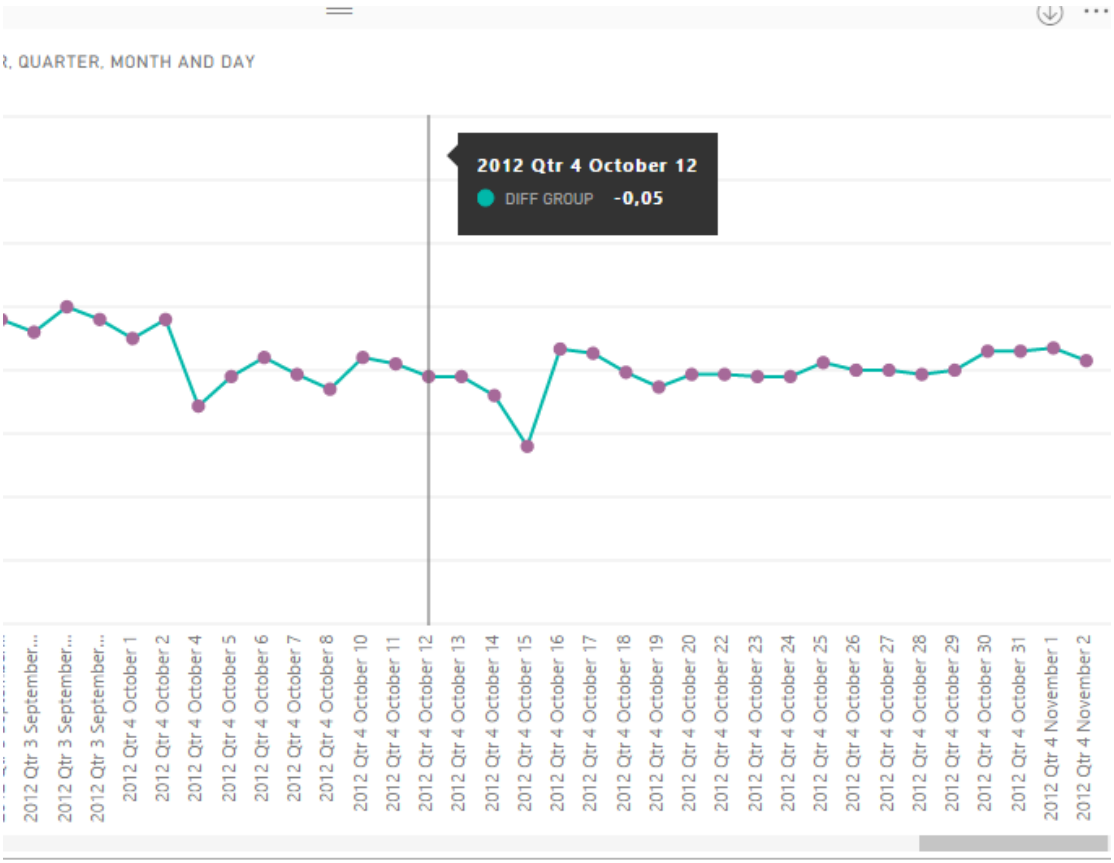
Εικόνα 3-26 - Χρονοσειρά για το πεδίο Difference (μ.ο.) με ευθεία τάσης , στο Power BI

Στα ημερολογιακά line charts, μπορούμε να εναλλάσσουμε τον άξονα σε έτη, τρίμηνα, μήνες και ημέρες μέσω των «drill up», «expand all down one level in hierarchy» και με το «drill down» να εστιάζουμε σε χαμηλότερη ιεραρχία.

Βήμα 9ο: Εστίαση σε συγκεκριμένες ημερομηνίες , αλλαγή επιπέδου στην κλίμακα χρονοσειράς

```
fig = poll_df.plot('Start Date','Difference',figsize=(12,4),marker='o',linestyle='-
',color='purple',xlim=(329,356))
```

Το Power BI, ενώ εστιάζει εύκολα στις επιθυμητές ημερομηνίες του άξονα (Εικόνα 3-27), δεν προσφέρει δυνατότητα προσθήκης σχημάτων-διακριτικών πάνω σε διάγραμμα (μόνο σε ολόκληρο το report μπορούμε να εισάγουμε).



Εικόνα 3-27 - χρονοσειρά για το πεδίο Difference στο Power BI, με εστίαση στο μήνα Οκτώβριο

3.2.3 Qlik Sense

Βήμα 1ο: Εισαγωγή αρχείου csv

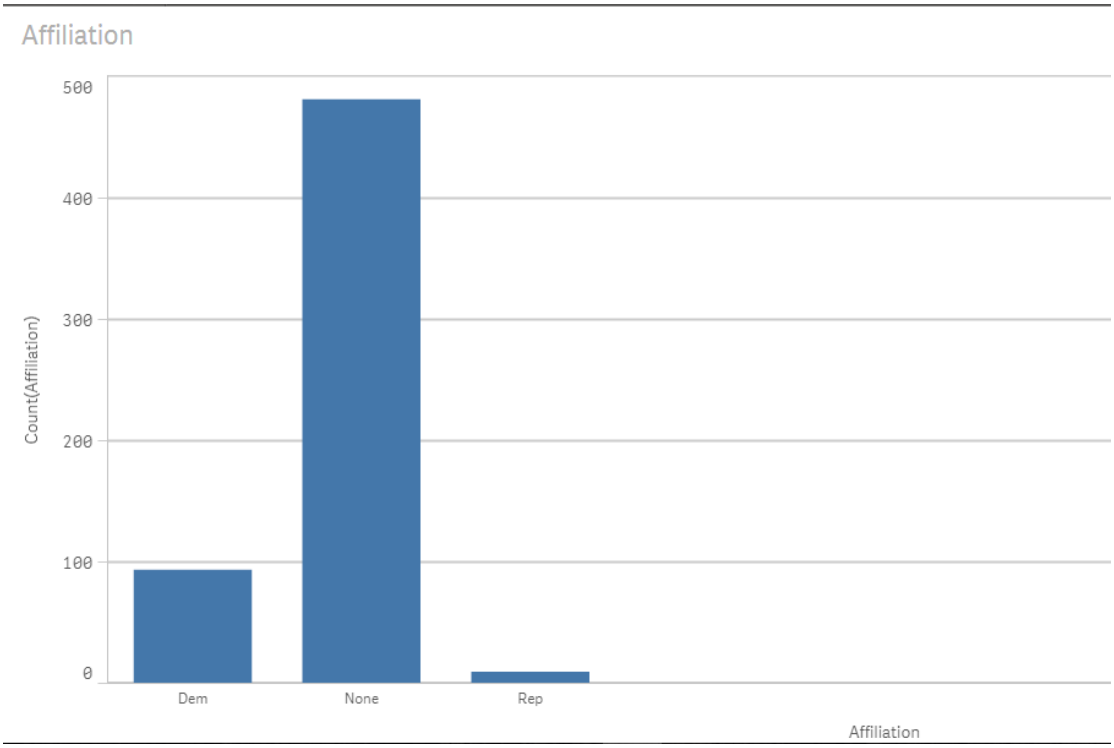
```
source = requests.get(url).text
```

Η εισαγωγή των δεδομένων από το csv αρχείο γίνεται άμεσα, και δίνεται η δυνατότητα εξαίρεσης στηλών εάν είναι επιθυμητό. Η εισαγωγή γίνεται σύμφωνα με το script του load editor. Εν προκειμένω, επειδή είναι ρυθμισμένο αυτόματα σύμφωνα με το χρησιμοποιηθέν λειτουργικό σύστημα, πρέπει να αντιστρέψουμε ‘,’ και ‘.’ στη στίξη δεκαδικών και χιλιάδων.

Βήμα 2ο:Διάγραμμα κομματικής προτίμησης συμμετεχόντων

```
sns.factorplot('Affiliation',data=poll_df)
```

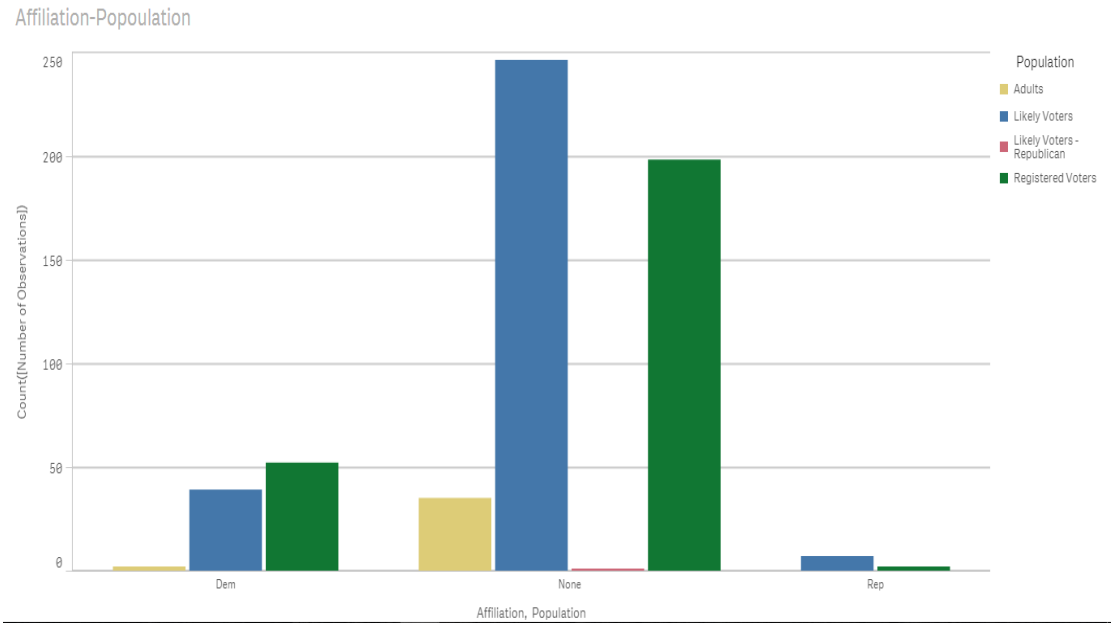
Με διάγραμμα μπάρας, drag & drop του dimension άξονα χ και των ποσοτικών μεγεθών (Εικόνα 3-28). Στη συγκεκριμένη περίπτωση ταυτίζονται, στο πεδίο affiliation. Η λεπτομέρεια του άξονα y μπορεί να ρυθμιστεί σε wide, medium, narrow, από την πιο «αραιή» στην πιο πυκνή διαβάθμιση.



Εικόνα 3-28 - διάγραμμα Qliksense για το πεδίο affiliation

Βήμα 3ο: Διάγραμμα κομματικής προτίμησης ανά πληθυσμιακή κατηγορία
sns.factorplot('Affiliation',data=poll_df,hue='Population')

Για να υλοποιήσουμε το διάγραμμα affiliation – population στον άξονα (Εικόνα 3-29), θα «σύρουμε» τα δύο πεδία ως dimensions (διαστάσεις). Δύο είναι το μέγιστο που μας επιτρέπει το εργαλείο. Έπειτα, στο παράθυρο της εμφάνισης (appearance), διαλέγουμε clustered (ή άλλη επιλογή είναι stacked).



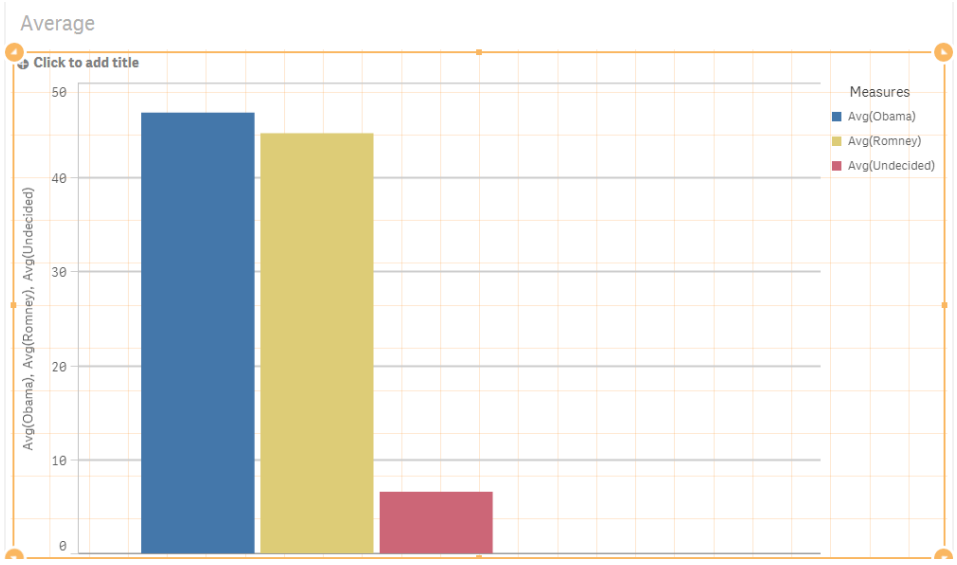
Εικόνα 3-29 - διάγραμμα συστοιχισμένων στηλών του Qliksense για τα πεδία affiliation και population

Σχετικά με τα χρώματα, μπορούμε να επιλέξουμε ανάμεσα σε μία βασική παλέτα 12 χρωμάτων και μία με ποικιλία 100. Ωστόσο, με συμβατική χρήση, χωρίς συναρτήσεις, δεν μπορεί να γίνει η αντιστοίχιση των χρωμάτων κατά βούληση. Έτσι το χρώμα από τα ίδια πεδία μπορεί να διαφέρει στα διαγράμματα.

Βήμα 4ο: Διάγραμμα μ.ο. των τριών πεδίων

```
avg = pd.DataFrame(poll_df.mean())
```

Εδώ, θέτουμε ως measures τους μ.ο. των τριών πεδίων που μας ενδιαφέρουν (Εικόνα 3-30). Για τη dimension επιλέγουμε ένα μέγεθος «ουδέτερο», που απλά να μην επηρεάζει το διάγραμμα. Ιδανικό για την περίσταση είναι το « Questions Iteration», που είναι ίσο με 1 σε όλες τις εγγραφές και των τριών πεδίων, παντού κοινό δηλαδή.



Εικόνα 3-30 - διάγραμμα Qliksense για τους μ.ο. των υποψηφίων

Βήμα 5ο: Υπολογισμός μ.ο. και τυπικών αποκλίσεων

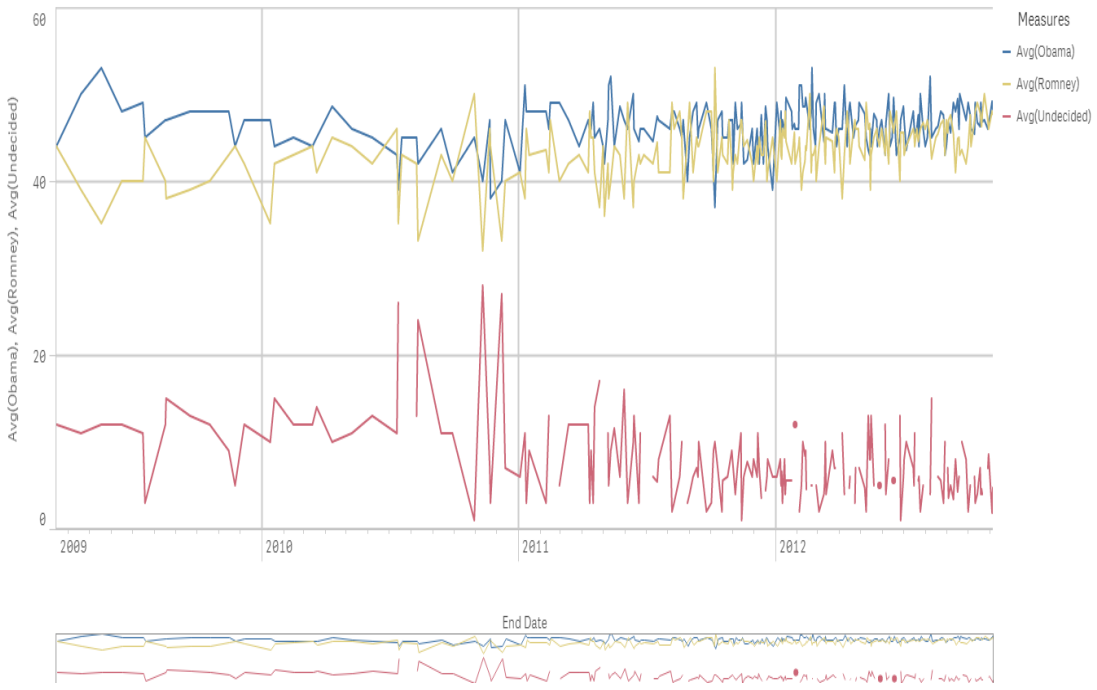
```
poll_avg = pd.concat([avg,std],axis=1)
```

Για να υπολογιστούν ο μέσος όρος και η τυπική απόκλιση, καλούμε τις επιθυμητές συναρτήσεις, με τα κατάλληλα ορίσματα. Υπολογίζονται στο 46.81 ο μ.ο. για τον Obama και στο 44.61 για τον Romney. Αντίστοιχα 2.42 και 2.91 οι τυπικές αποκλίσεις.

Βήμα 6ο: Γράφημα τιμών των προβλέψεων, ανεστραμμένου χρονικού άξονα

```
poll_df.plot(x='End Date',y=['Obama','Romney','Undecided'],marker='o',linestyle='')
```

Και σε αυτό το εργαλείο δεν είναι εφικτό να αναπαραστήσουμε το διάγραμμα του πρωτοτύπου, οπότε δημιουργούμε μία χρονοσειρά, με τους μ.ο. των 3 πεδίων (Εικόνα 3-31) .



Εικόνα 3-31 – διάγραμμα Qliksense με τους μ.ο. των αποτελεσμάτων των δημοσκοπήσεων σε άξονα χρόνου

Πολύ εύκολη εστίαση (drill through) στις επιθυμητές ημερομηνίες, κατά τη λειτουργία προβολής/παρουσίασης, απλά με το scroll του ποντικιού.

Βήμα 7ο: Δημιουργία νέας στήλης «Difference»

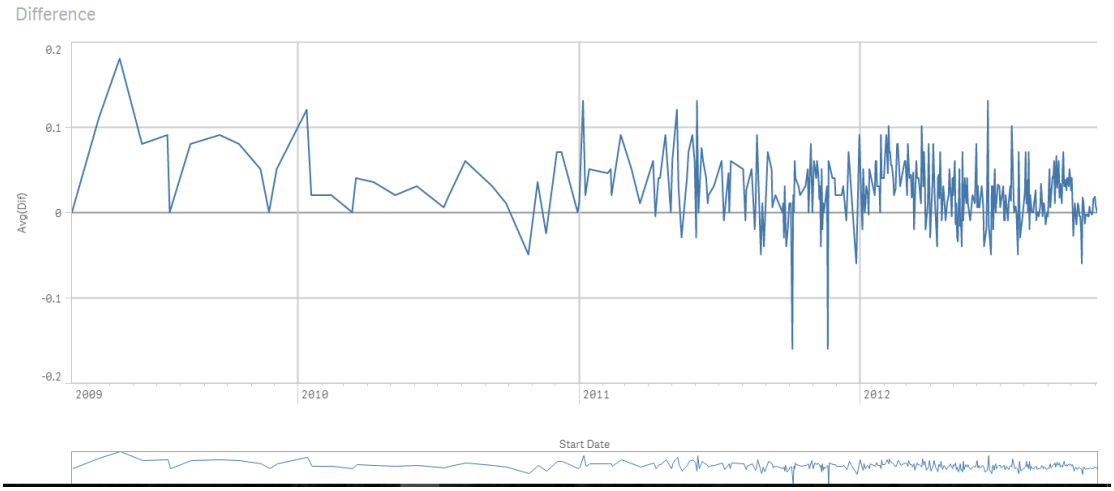
```
poll_df['Difference'] = (poll_df.Obama - poll_df.Romney)/100
```

Δημιουργούμε μία νέα στήλη «Difference» , όπως ορίζεται πιο πάνω, ως calculated field. Η διαδικασία αυτή λαμβάνει χώρα στο data manager, όπου υπάρχει και παράθυρο προεπισκόπησης (preview) του αποτελέσματος.

Βήμα 8ο: Γραφική αναπαράσταση των μ.ο. της «Difference» , στη χρονοσειρά με άξονα τη start date

```
poll_df = poll_df.groupby(['Start Date'],as_index=False).mean()
```

Επειδή, το group by date, που απαιτείται στη συνέχεια μπορεί να γίνει μόνο μέσω script, στο data load editor, το αποφεύγουμε. Απλά, υπολογίζουμε το μ.ο. της διαφοράς ανά ημέρα (ουσιαστικά η ίδια διαδικασία). Το διάγραμμα (Εικόνα 3-32) έχει ως εξής, σε συνολικό εύρος:

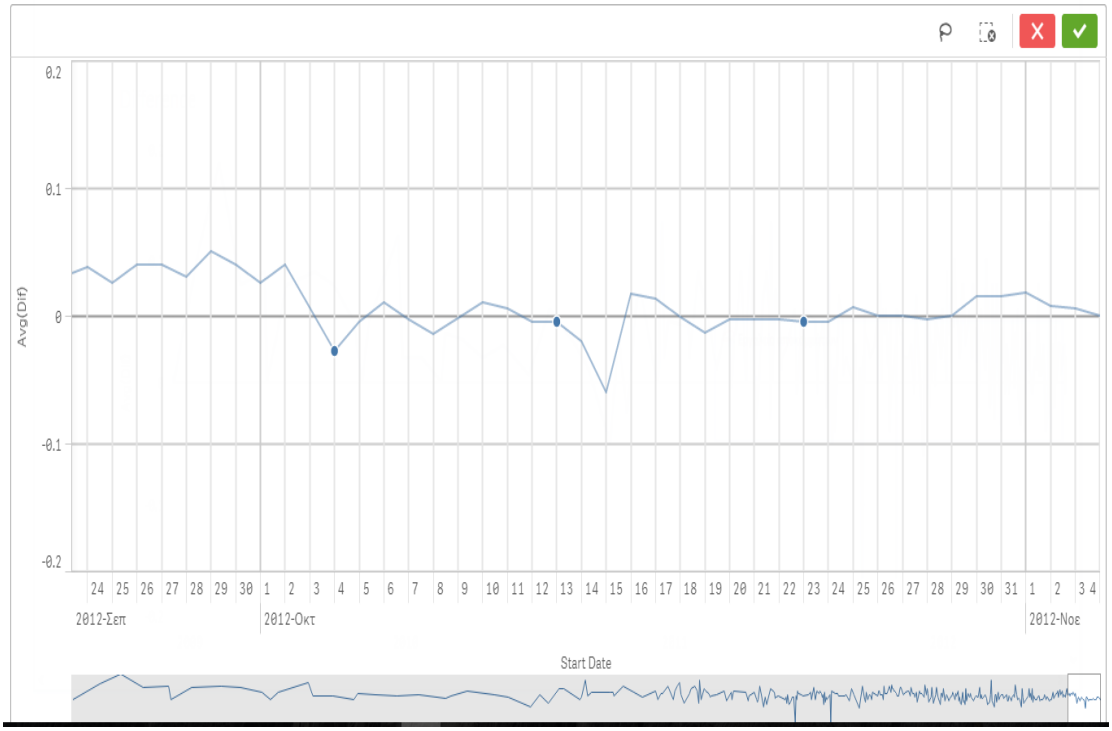


Εικόνα 3-32 - Χρονοσειρά για το πεδίο Difference (μ.ο.), στο Qliksense

Βήμα 9ο: Εστίαση σε συγκεκριμένες ημερομηνίες , αλλαγή επιπέδου στην κλίμακα χρονοσειράς

```
fig = poll_df.plot('Start Date','Difference',figsize=(12,4),marker='o',linestyle='-
',color='purple',xlim=(329,356))
```

Και στο μήνα Οκτώβρη 2012, με σημειωμένες τις ημερομηνίες που μας ενδιαφέρουν (Εικόνα 3-33):



Εικόνα 3-33- χρονοσειρά για το πεδίο Difference στο Qliksense, με εστίαση στο μήνα Οκτώβριο

3.3 Σενάριο αριθμητικών δεδομένων

Στο σενάριο αριθμητικών δεδομένων, εστιάζουμε στη διαχείριση μεγάλου αριθμού δεδομένων, που αφορούν χρηματικά ποσά. Εξετάζουμε επίσης και την αποτελεσματικότητα των εργαλείων σε ένα μεγάλου μεγέθους αρχείο, 154 MB, με περισσότερες από ένα εκατομμύριο εγγραφές.

Και αυτό το σενάριο (http://nbviewer.jupyter.org/github/jmportilla/Udemy-notes/blob/master/Data%20Project%20-%20Election%20Analysis.ipynb) αφορά τις προεδρικές εκλογές του 2012 στις Η.Π.Α., και συγκεκριμένα τις δωρεές που έλαβαν οι υποψήφιοι για το αξίωμα. Μέσω των διαγραμμάτων θα προσπαθήσουμε να δώσουμε κάποιες πληροφορίες - πόσα χρήματα δόθηκαν, ποιος ο μ.ο. αυτών, τι ποσό πήρε ο κάθε υποψήφιος, τι ποσό αντιστοιχεί σε κάθε κόμμα, την ύπαρξη μοτίβου στις συνεισφορές, και τα δημογραφικά στοιχεία των δωρητών. Γίνεται περιγραφή του σεναρίου και παρατίθενται ορισμένες αντιπροσωπευτικές εντολές, ενώ ο κώδικας βρίσκεται στο παράρτημα (Εντολές σεναρίου αριθμητικών στοιχείων).

Αρχικά εισάγεται το csv αρχείο και γίνεται μία επισκόπηση, ενδεικτικά, των πρώτων στοιχείων του (Εικόνα 3-34).

	cmte_id	cand_id	cand_nm	contbr_nm	contbr_city	contbr_st	contbr_zip	co
0	C00410118	P20002978	Bachmann, Michelle	HARVEY, WILLIAM	MOBILE	AL	3.660103e+08	
1	C00410118	P20002978	Bachmann, Michelle	HARVEY, WILLIAM	MOBILE	AL	3.660103e+08	
2	C00410118	P20002978	Bachmann, Michelle	SMITH, LANIER	LANETT	AL	3.686334e+08	
3	C00410118	P20002978	Bachmann, Michelle	BLEVINS, DARONDA	PIGGOTT	AR	7.245483e+08	
4	C00410118	P20002978	Bachmann, Michelle	WARDENBURG, HAROLD	HOT SPRINGS NATION	AR	7.190165e+08	

Εικόνα 3-34 – επισκόπηση των δεδομένων

Έπειτα, εμφανίζονται μέσω εντολών, η συχνότητα εμφάνισης ανά ποσό δωρεάς, καθώς και ο μ.ο. και η τυπική απόκλιση αυτών.

```
donor_df['contb_receipt_amt'].value_counts()
don_mean = donor_df['contb_receipt_amt'].mean()
don_std = donor_df['contb_receipt_amt'].std()
```

Ο μέσος όρος υπολογίζεται σε 298.24 δολάρια , και η τυπική απόκλιση 3749.67. Η τυπική απόκλιση είναι πού μεγάλη! Μετά από ταξινόμηση για καλύτερη εποπτεία, διαπιστώνουμε ότι υπάρχουν κάποια τεράστια ποσά, αλλά και κάποιες αρνητικές τιμές που αντιστοιχούν σε αποζημιώσεις. Θα ασχοληθούμε μόνο με τις θετικές τιμές.

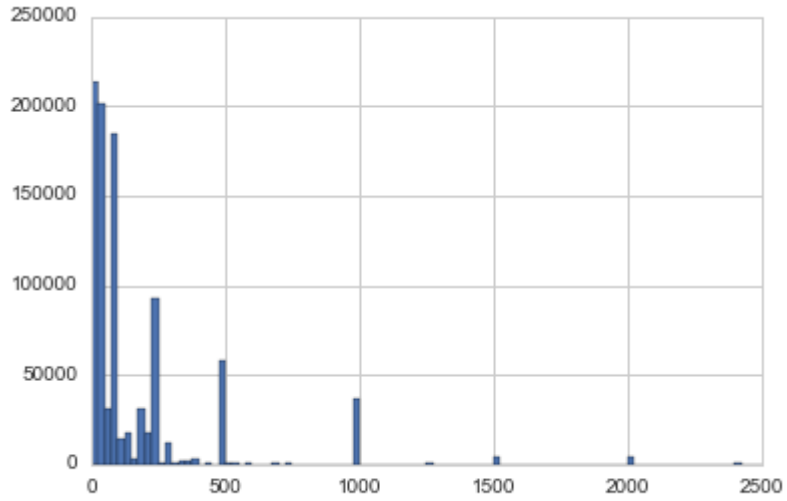
```
top_donor = top_donor[top_donor >0]
top_donor.sort()
```

Τυπώνονται τα δέκα πιο συχνά ποσά των συνεισφορών:

```
top_donor.value_counts().head(10)
```

Παρατηρούμε ότι οι 10 συχνότερες συνεισφορές είναι μέχρι 2500 δολάρια. Εύλογη ερώτηση είναι εάν γίνονται οι δωρεές αυτές σε «στρογγυλά» νούμερα. Ας το διαπιστώσουμε, μέσω ενός ιστογράμματος (Εικόνα 3-35) :

```
com_don = top_donor[top_donor < 2500]
com_don.hist(bins=100)
```



Εικόνα 3-35 – ιστόγραμμα για τα ποσά συνεισφορών, μέχρι τις 2500 δολ.

Πράγματι, τα spikes αντιστοιχούν σε «στρογγυλά» ποσά. Ας κατατάξουμε, τώρα, τις εγγραφές ανά κόμμα, σε μια νέα στήλη. Αυτό θα εξαρτηθεί από το όνομα του υποψηφίου.
Η αντιστοίχιση του κόμματος για κάθε υποψήφιο μπορεί να γίνει είτε με «mapping» (άμεση αντιστοίχιση), είτε με «for loop» μιας και ο Barack Obama είναι ο μοναδικός υποψήφιος των Δημοκρατικών. Η μέθοδος του «for loop» είναι απλούστερη, αλλά σαφώς πιο αργή. Επίσης, απαλλασσόμαστε από τα αρνητικά ποσά των συνεισφορών που αντιστοιχούν σε χρέη. Εμφανίζεται ο αριθμός των συνεισφορών ανά υποψήφιο :

```
donor_df['Party'] = donor_df.cand_nm.map(party_map)
donor_df.groupby('cand_nm')['contb_receipt_amt'].count()
```

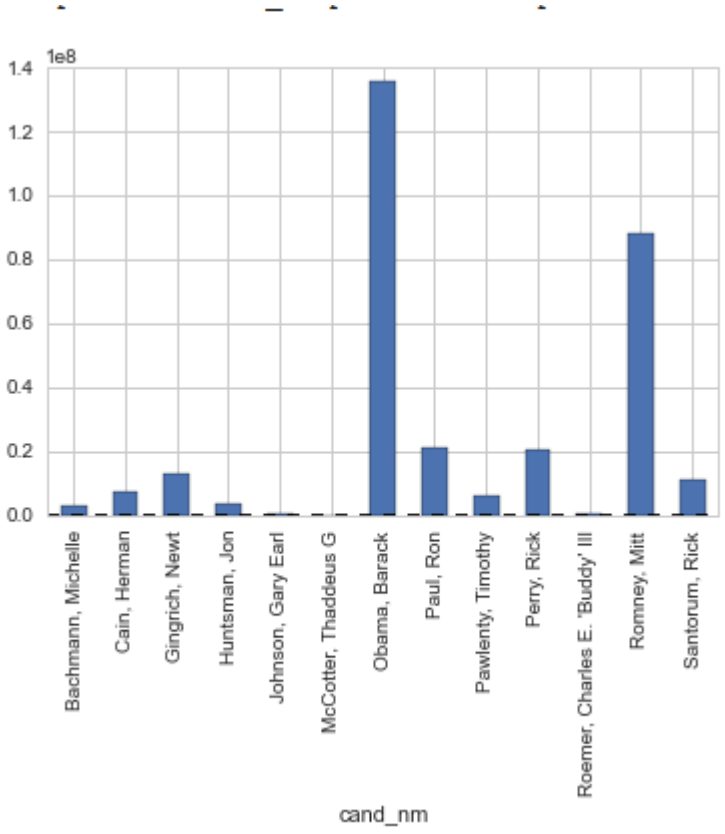
```
cand_nm
Bachmann, Michelle              13082
Cain, Herman                    20052
Gingrich, Newt                  46883
Huntsman, Jon                    4066
Johnson, Gary Earl               1234
McCotter, Thaddeus G               73
Obama, Barack                  589127
Paul, Ron                      143161
Pawlenty, Timothy                3844
Perry, Rick                     12709
Roemer, Charles E. 'Buddy' III   5844
Romney, Mitt                   105155
Santorum, Rick                  46245
Name: contb_receipt_amt, dtype: int64
```

Ο Obama, ως ο μόνος υποψήφιος των δημοκρατικών, συγκεντρώνει τις περισσότερες δωρεές. Στη συνέχεια, υπολογίζεται το άθροισμα των συνεισφορών ανά υποψήφιο :

```
donor_df.groupby('cand_nm')['contb_receipt_amt'].sum()
```

```
Bachmann, Michelle              2.711439e+06
Cain, Herman                    7.101082e+06
Gingrich, Newt                  1.283277e+07
Huntsman, Jon                   3.330373e+06
Johnson, Gary Earl              5.669616e+05
McCotter, Thaddeus G            3.903000e+04
Obama, Barack                   1.358774e+08
Paul, Ron                       2.100962e+07
Pawlenty, Timothy               6.004819e+06
Perry, Rick                     2.030575e+07
Roemer, Charles E. 'Buddy' III  3.730099e+05
Romney, Mitt                    8.833591e+07
Santorum, Rick                  1.104316e+07
Name: contb_receipt_amt, dtype: float64
```

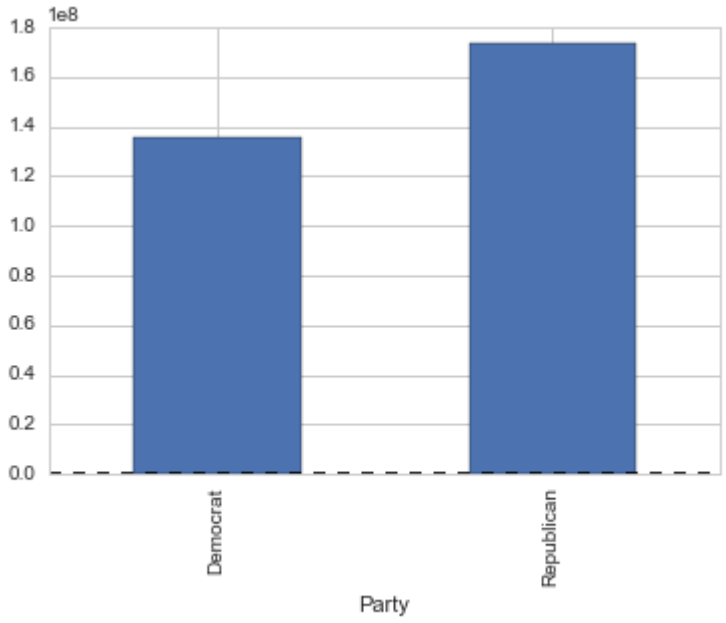
Οι τιμές είναι δυσανάγνωστες, όπως παρουσιάζονται και είναι δύσκολο να συγκριθούν. Επομένως, κατασκευάζουμε ένα ραβδόγραμμα (Εικόνα 3-36) για την καλύτερη και ευκολότερη αντίληψη των μεγεθών.



Εικόνα 3-36 – διάγραμμα με το άθροισμα του ποσού των συνεισφορών για κάθε υποψήφιο

Και ένα γράφημα (Εικόνα 3-37) για τα ποσά εμφανίζονται συγκεντρωτικά για κάθε κόμμα:

```
donor_df.groupby('Party')['contb_receipt_amt'].sum().plot(kind='bar')
```



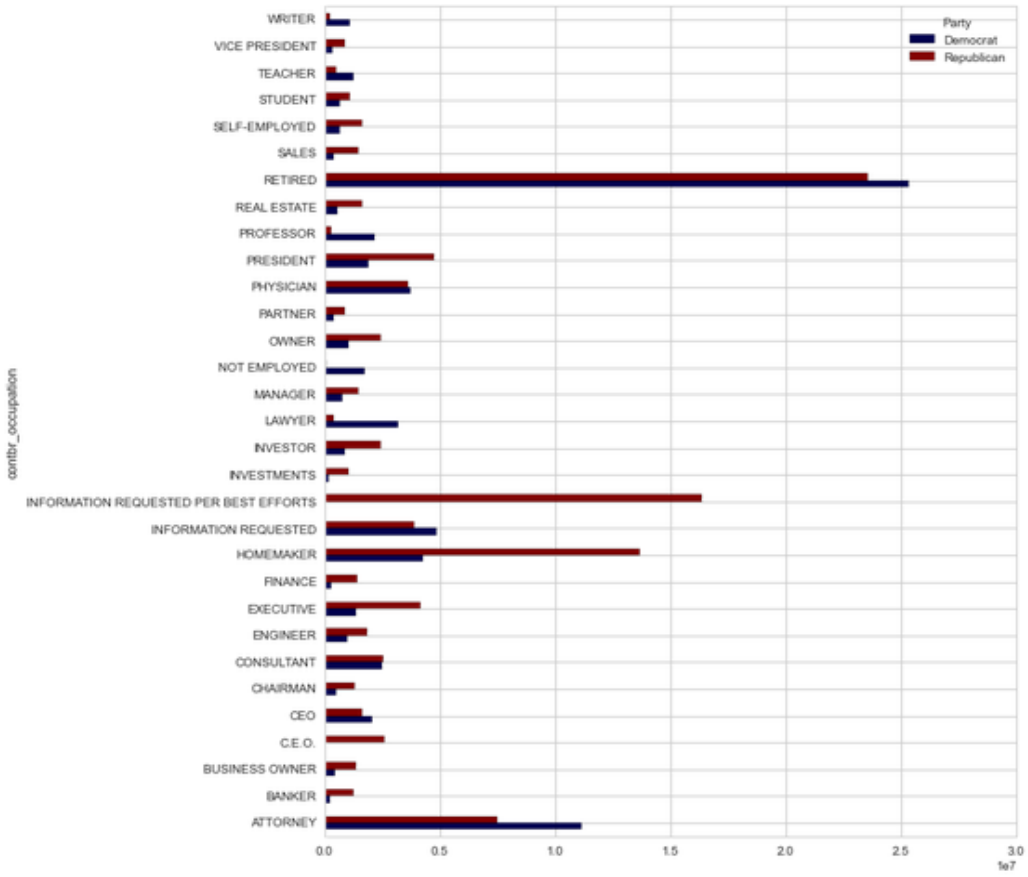
Εικόνα 3-37 – διάγραμμα με το άθροισμα συνεισφορών για κάθε κόμμα

Παρά το μεγάλο ποσό που συγκεντρώνει ο Barrack Obama, δεν καταφέρνει να ξεπεράσει το άθροισμα των ρεπουμπλικάνων. Κλείνοντας, θέλουμε να έχουμε εικόνα από το που προήλθαν οι δωρεές, όσον αφορά το επάγγελμα των ατόμων, με τη χρήση ενός πίνακα περιστροφής (pivot table):

```
# Use a pivot table to extract and organize the data by the donor occupation
occupation_df = donor_df.pivot_table('contb_receipt_amt',
                        index='contbr_occupation',
                        columns='Party', aggfunc='sum')
```

Τα επαγγέλματα είναι πάρα πολλά, για να αναπαρασταθούν οπτικά. Επομένως, θα εμφανίσουμε αυτά που, αθροιστικά και γα τα δύο κόμματα, έχουν προσφέρει ποσό μεγαλύτερο του ενός εκατομμυρίου δολαρίων (Εικόνα 3-38). Στο σημείο αυτό πρέπει να αποκλειστούν από το διάγραμμα στοιχεία που δεν έχουν πληροφορία στο πεδίο του επαγγέλματος («information requested»). Επίσης, πρέπει οι τιμές που αφορούν το ίδιο επάγγελμα, εδώ συγκεκριμένα, C.E.O. και CEO, με κοινή τιμή «CEO».

```
occupation_df = occupation_df[occupation_df.sum(1) > 1000000]
```



Εικόνα 3-38 – Οριζόντιο διάγραμμα για τα επαγγέλματα των συνεισφερόντων

3.3.1 Tableau

Βήμα 1ο: Εισαγωγή του csv αρχείου

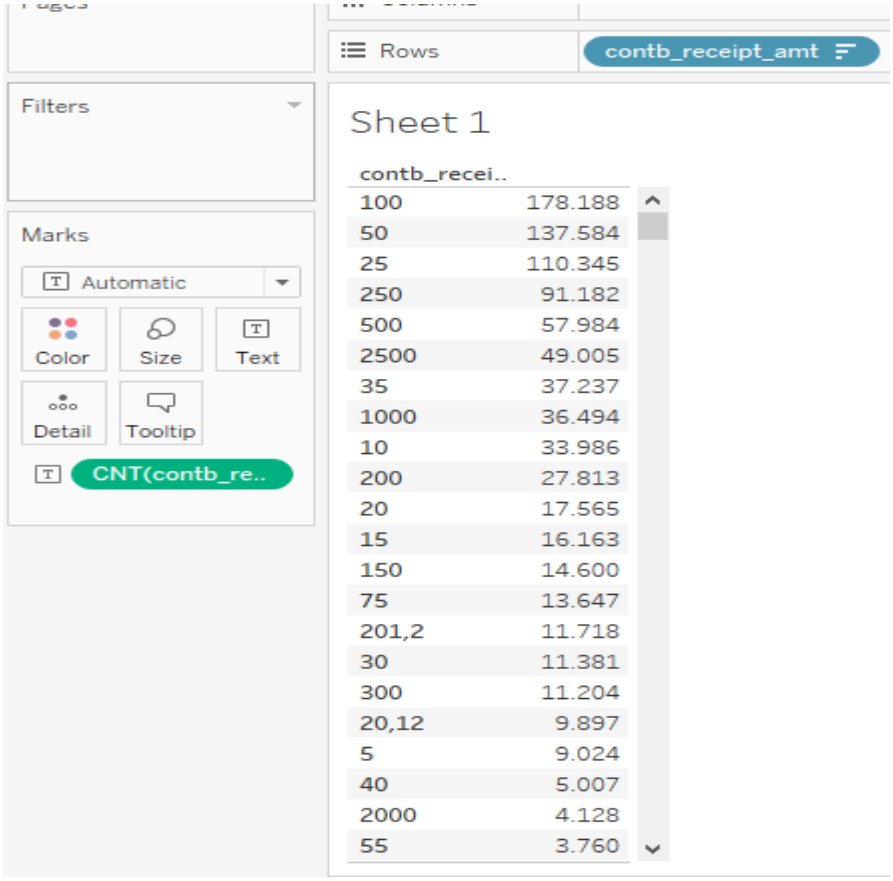
```
donor_df = pd.read_csv('Election_Donor_Data.csv')
```

Λόγω του μεγέθους του αρχείου, η εισαγωγή ήταν σχετικά χρονοβόρα. Με το πέρας της, ρυθμίστηκε η αναγνώριση των συμβόλων σύμφωνα με την αγγλική γλώσσα, δηλαδή η «.» ως διαχωριστικό του δεκαδικού αριθμού.

Βήμα 2ο: Εμφάνιση συχνότητας των ποσών

```
donor_df['contb_receipt_amt'].value_counts()
```

Τοποθετώντας τα ποσά, ως ποιοτική μεταβλητή στο πεδίο των γραμμών, εμφανίζουμε τη συχνότητα, υπό τη μορφή κειμένου (Εικόνα 3-39).

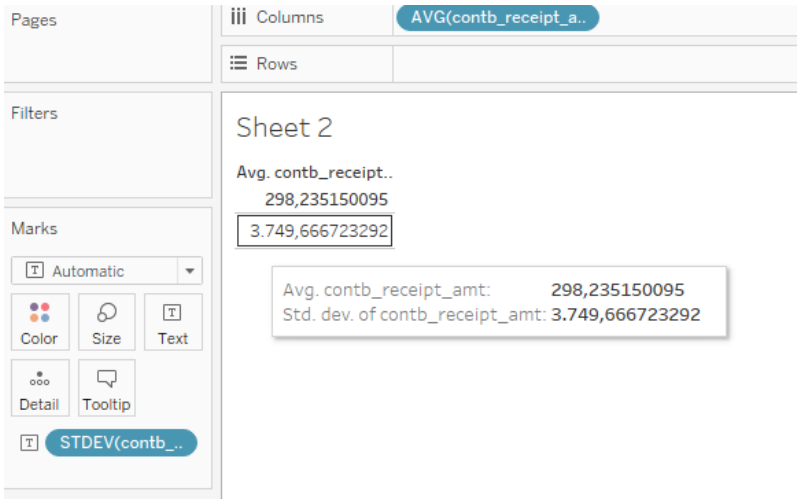


Εικόνα 3-39 – επισκόπηση της συχνότητας εμφάνισης ανά ποσό συνεισφοράς στο Tableau

Βήμα 3ο: Υπολογισμός της μέσης συνεισφοράς και της τυπικής απόκλισης

```
don_mean = donor_df['contb_receipt_amt'].mean()
don_std = donor_df['contb_receipt_amt'].std()
```

Με τις συναρτήσεις που παρέχονται αυτόματα, υπολογίζουμε τη μέση συνεισφορά και την τυπική απόκλιση (Εικόνα 3-40).

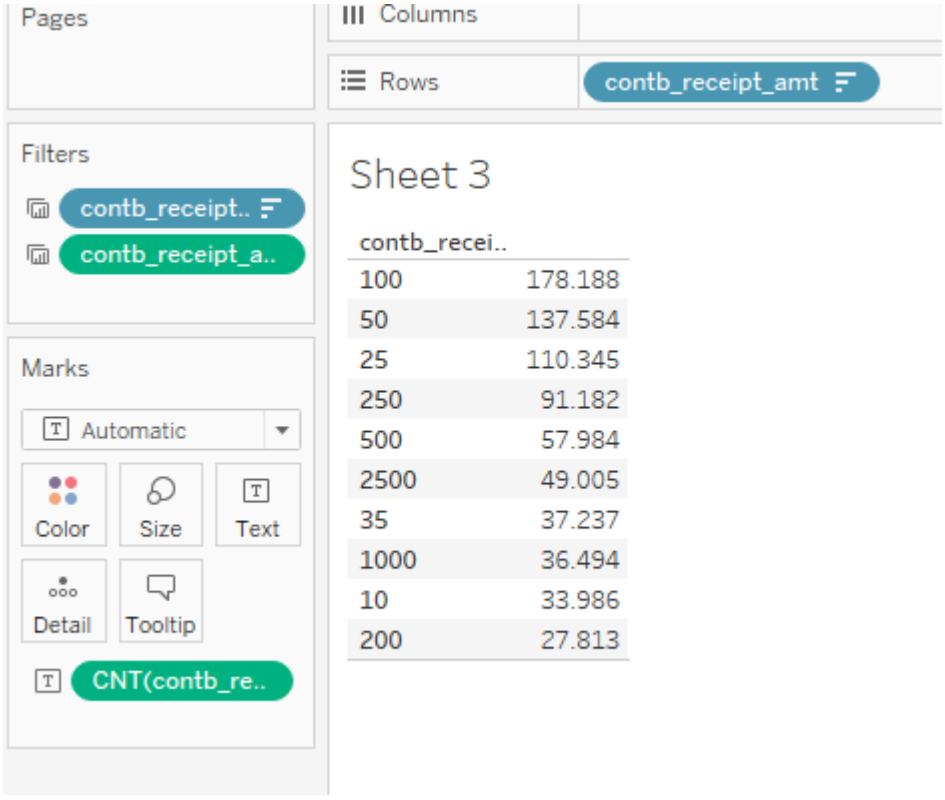


Εικόνα 3-40 – Υπολογισμός μέσης συνεισφοράς και τυπικής απόκλισης στο Tableau

Βήμα 4ο: Απαλλοιφή των αρνητικών τιμών, ταξινόμηση και 10 συχνότερα ποσά δωρεάς

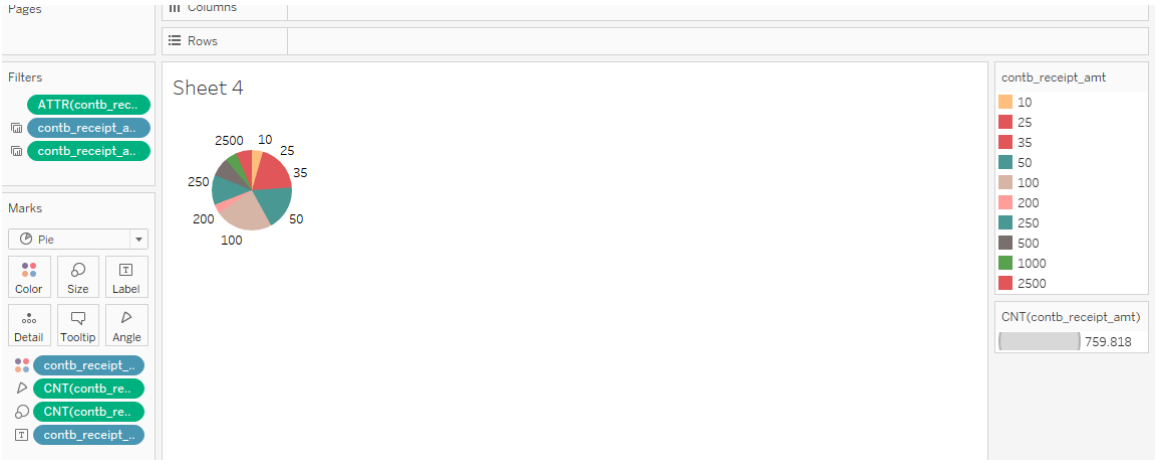
```
top_donor = top_donor[top_donor >0]
top_donor.sort()
top_donor.value_counts().head(10)
```

Εφαρμόζουμε φίλτρα, ώστε να εμφανίζονται τα δέκα συχνότερα εμφανιζόμενα θετικά ποσά, και ταξινομούμε (Εικόνα 3-41). Για την καλύτερη αντίληψη των μεγεθών, αναπαριστούμε σε διάγραμμα πίττας (Εικόνα 3-42).



Εικόνα 3-41- Φίλτρο στα Tableau για τα συχνότερα εμφανιζόμενα ποσά

Το φίλτρο για θετικές συνεισφορές (contb_receipt_amount >0)επιλέγουμε να εφαρμόζεται και σε όποια επόμενα sheets θέλουμε.

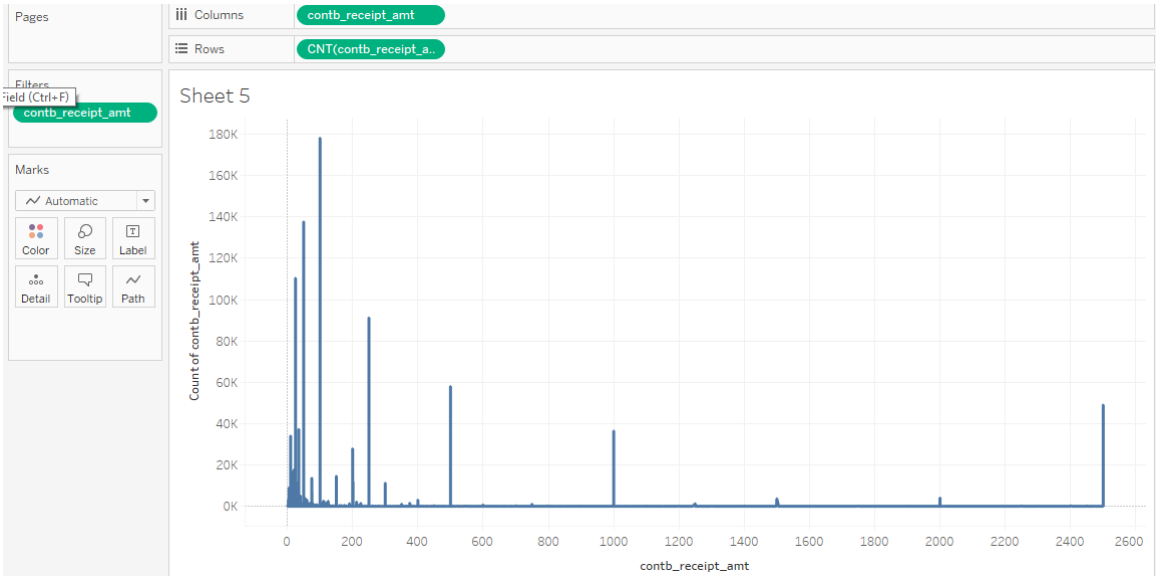


Εικόνα 3-42 – Διάγραμμα «πίτας» με τα συχνότερα εμφανιζόμενα ποσά

Βήμα 5ο: Περιορισμός ποσών στα 2500 δολάρια και ιστόγραμμα

```
com_don = top_donor[top_donor < 2500]
com_don.hist(bins=100)
```

Το ιστόγραμμα απεικονίζει τη συχνότητα για τα ποσά των συνεισφορών (Εικόνα 3-43).

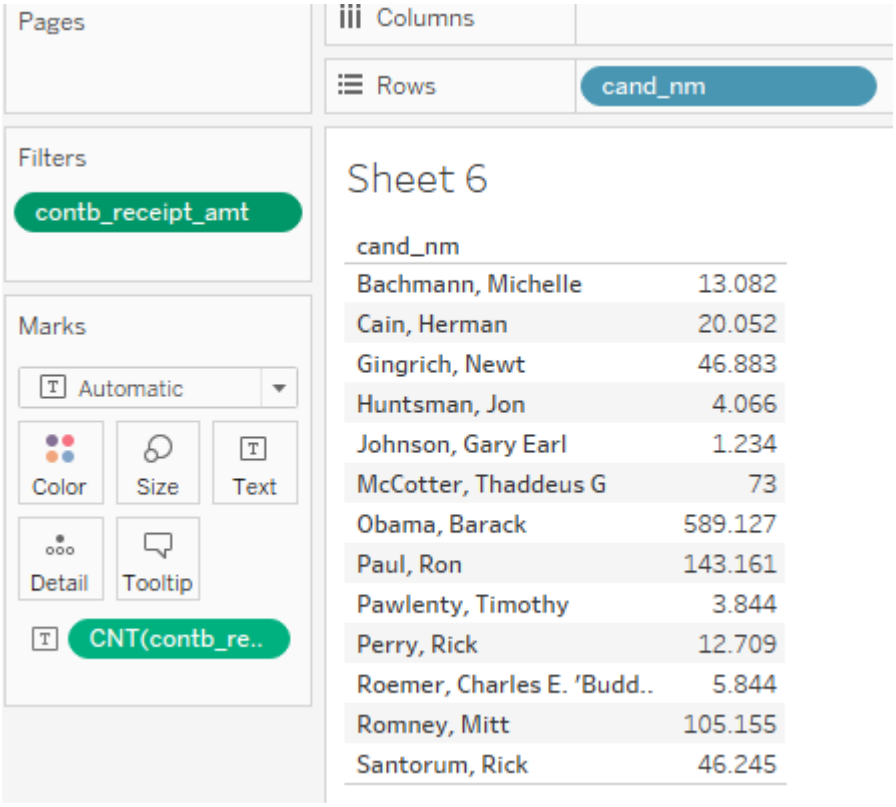


Εικόνα 3-43 – ιστόγραμμα στο Tableau για τα ποσά συνεισφορών, μέχρι τις 2500 δολ.

Βήμα 6ο: Δημιουργία πεδίου αντιστοίχισης κομμάτων και αρίθμηση συνεισφορών ανά υποψήφιο

```
donor_df['Party'] = donor_df.cand_nm.map(party_map)
donor_df.groupby('cand_nm')['contb_receipt_amt'].count()
```

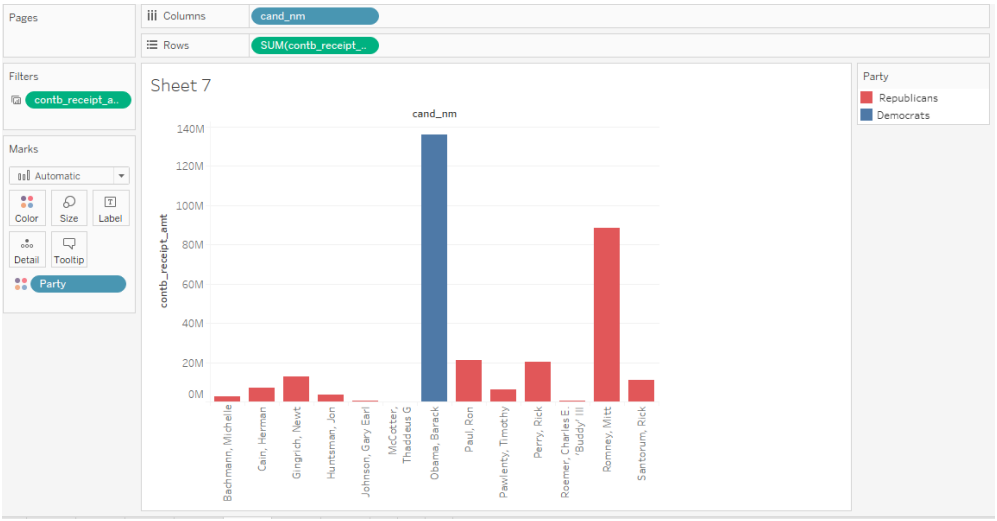
Δημιουργούμε αρχικά calculated field με όνομα Party, όπου οι τιμές βασίζονται στο πεδίο cand_nm. Υπολογίζουμε και την αρίθμηση των συνεισφορών ανά υποψήφιο (Εικόνα 3-44) .



Εικόνα 3-44 – η αρίθμηση των συνεισφορών ανά υποψήφιο στο Tableau

Βήμα 7ο: Διάγραμμα με τα συνολικά αθροιστικά ποσά ανά υποψήφιο
donor_df.groupby('cand_nm')['contb_receipt_amt'].sum()

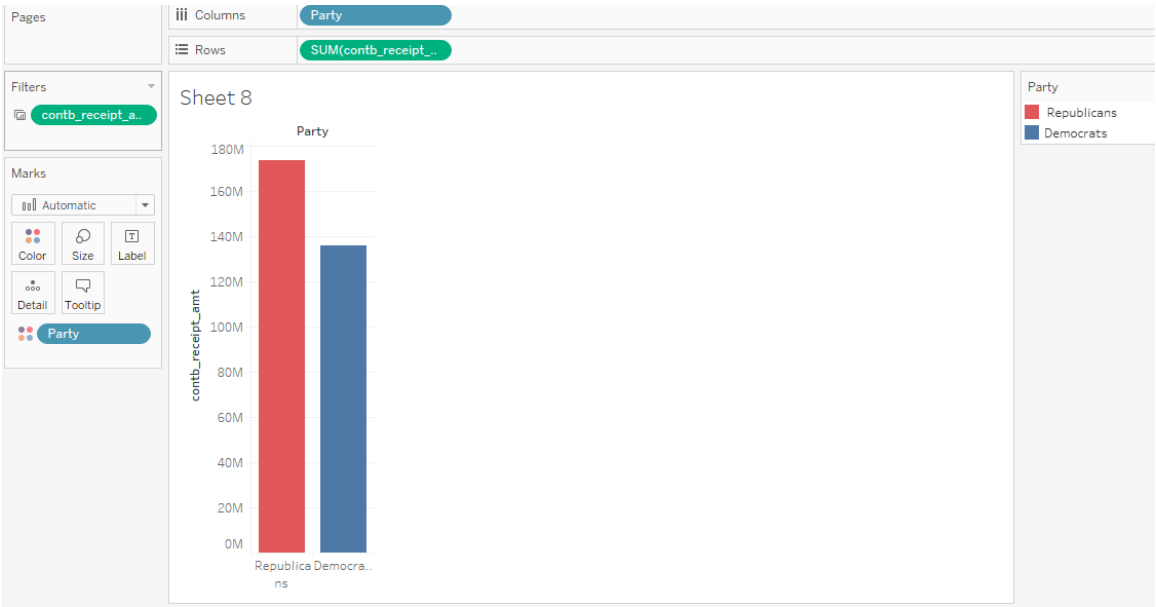
Στο ραβδόγραμμα (Εικόνα 3-45) εμφανίζονται τα ποσά ως άθροισμα ανά υποψήφιο και διακρίνονται χρωματικά, όσον αφορά το κόμμα που ανήκει ο κάθε υποψήφιος.



Εικόνα 3-45 – ραβδόγραμμα του Tableau με το άθροισμα των συνεισφορών ανά υποψήφιο

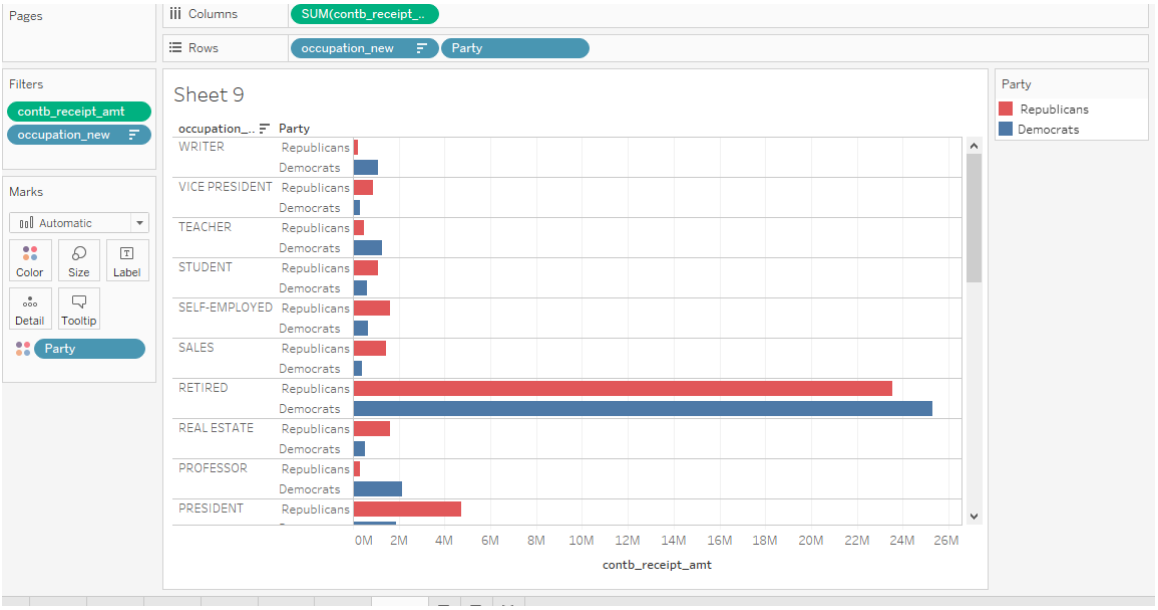
Βήμα 8ο: Διάγραμμα αθροιστικού ποσού συνεισφορών ανά κόμμα
donor_df.groupby('Party')['contb_receipt_amt'].sum().plot(kind='bar')

Σε αυτό το ραβδόγραμμα (Εικόνα 3-46), εμφανίζεται το άθροισμα των συνεισφορών για τα δύο κόμματα.



Εικόνα 3-46 – διάγραμμα του Tableau, με το άθροισμα των συνεισφορών για κάθε παράταξη

Βήμα 9ο: Ελάχιστο κατώφλι 1,000,000 και οριζόντιο clustered διάγραμμα
occupation_df = occupation_df[occupation_df.sum(1) > 1000000]



Εικόνα 3-47 - Οριζόντιο διάγραμμα του Tableau για τα επαγγέλματα των συνεισφερόντων

Δημιουργήσαμε τη στήλη occupation_new όπου συγχωνεύσαμε CEO και C.E.O., με κοινή τιμή «CEO» (Εικόνα 3-47).Στη συνέχεια, με το φίλτρο στο occupation_new επιλέγουμε να μη συμπεριλαμβάνονται τα πεδία «information required» etc. και να εμφανίζονται μόνο όσες τιμές του (επαγγέλματα δηλαδή) έχουν sum >1.000.000. Το φιλτράρισμα εδώ λειτουργεί όπως ακριβώς θέλουμε, χωρίς δηλαδή να αντιλαμβάνεται ως διαφορετικές εγγραφές π.χ. το Vice-President Republicans και Vice-President Democrats, όπου ξεχωριστά κανένα δεν υπερέβαινε το 1εκ. (Συνέβαινε αυτό στο Power BI). Δηλαδή, μας επιτρέπει να φιλτράρουμε συγκεκριμένο πεδίο ξεχωριστά, και όχι απλά να εφαρμόσουμε φίλτρο σε όλες τις εγγραφές του γραφήματος.

3.3.2 MS Power BI

Βήμα 1ο: Εισαγωγή του csv αρχείου

Με χρήση της βιβλιοθήκης pandas της python, η ανάγνωση του csv γίνεται άμεσα, με την εντολή pd.read_csv ('Election_Donor_Data.csv'), και με την .head() , διαπιστώνουμε ότι έγινε σωστά (και αυτόματα) η φόρτωση των δεδομένων. Το «διάβασμα» των δεδομένων είναι ίσως το σημαντικότερο σημείο της επεξεργασίας, ώστε να εξάγουμε συμπεράσματα με βάση τα σωστά δεδομένα, τόσο στις τιμές, στους τύπους δεδομένων κ.α.

Στο MS Power BI, κατά το διάβασμα, εμφανίζεται το αντίστοιχο παράθυρο όπου επιλέγουμε τύπο αρχείου text/csv, επιλέγουμε το διαχωριστικό (delimeter), και στο data type detection αλλάζουμε την προεπιλογή based on 200 first rows, μιας και στις περισσότερες του 1,000,000

σειρές υπάρχουν και δεκαδικοί αριθμοί, ενώ τυχαίνει στις 200 πρώτες να υπάρχουν μόνον ακέραιες τιμές. Προσφέρεται η επιλογή based on entire data set, αλλά αργεί χαρακτηριστικά (με τους υπολογιστικούς πόρους της δοκιμής δεν ολοκληρώνεται).

Επομένως, θα επιλέξουμε do not detect data type, ώστε να διαμορφώσουμε ex nihilo το σύνολο των δεδομένων στο πρόγραμμα query editor, χωρίς αυτόματη προεπεξεργασία. Είναι σημαντικό εδώ να ρυθμίσουμε τα regional settings ώστε να αναγνωρίζονται οι δεκαδικοί με ‘ . ‘ (παρατηρούμε όμως ότι οι δεκαδικοί διαχωρίζονται με ‘,’ ,παρότι η αλλαγή τύπου γίνεται απρόσκοπτα, χωρίς errors). Στο Power BI, έχουμε την ευχέρεια να εποπτεύουμε και να ταξινομούμε τα δεδομένα άμεσα στο παράθυρο Data (Εικόνα 3-48). Οπότε δε χρειάζονται εντολές για εποπτεία, όπως η head() .

cmte_id	cand_id	cand_nm	contbr_nm	contbr_city	contbr_st	contbr_zip	contbr_employer	contbr_occupation	contb_receipt_amt	contb_receipt_dt	receipt_desc	memo_cd	memo_text	form_tp	file_num	Party	occupation
C00431445	P80003338	Obama, Barack	CAMPISI, CATHERINE	SACRAMENTO	CA	958182935	RETIRED	RETIRED	100	Τετάρτη, 30 Νοεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	COLLINSWORTH, JEAN	CLAREMONT	CA	91711	RETIRED	RETIRED	100	Παρασκευή, 14 Οκτωβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	SCHMIDT, ILENE	RLLNG HLS EST	CA	902741530	RETIRED	RETIRED	100	Τετάρτη, 19 Οκτωβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	SPELBOS, BARBARA	VALLEJO	CA	945917139	RETIRED	RETIRED	100	Παρασκευή, 30 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	BECKER, CAROL A.	TORRANCE	CA	905052031	RETIRED	RETIRED	100	Πέμπτη, 29 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	SUDA, ROBERT	MONTE SERENO	CA	950304231	RETIRED	RETIRED	100	Παρασκευή, 30 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	ALPERT, KATHLEEN M	CRESCENT CITY	CA	95531	RETIRED	RETIRED	100	Τρίτη, 22 Νοεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	HALL, THOMAS L.	SAN FRANCISCO	CA	941223524	RETIRED	RETIRED	100	Τετάρτη, 7 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	MENO, MIKE	FULLERTON	CA	92832	RETIRED	RETIRED	100	Πέμπτη, 29 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	JOHNSON, J. DONALD	OCEANSIDE	CA	920565105	RETIRED	RETIRED	100	Τετάρτη, 7 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	TRAN, LUU	SAN DIEGO	CA	921392014	RETIRED	RETIRED	100	Δευτέρα, 19 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	GENELIZA, DARLENE	SAN JOSE	CA	951252270	RETIRED	RETIRED	100	Παρασκευή, 23 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	COSTA, MARGARET	HUNTINGTON BEACH	CA	92648	RETIRED	RETIRED	100	Πέμπτη, 1 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	MYERS, KATHY	SACRAMENTO	CA	958183528	RETIRED	RETIRED	100	Παρασκευή, 2 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	ZWERN, RITA	BURBANK	CA	915011128	RETIRED	RETIRED	100	Τετάρτη, 16 Νοεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	HUMA, BEVERLY J.	REDLANDS	CA	923745348	RETIRED	RETIRED	100	Παρασκευή, 30 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	WALLACH, VERA	TARZANA	CA	913564027	RETIRED	RETIRED	100	Παρασκευή, 28 Οκτωβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	BROWN, CHARLES E	SACRAMENTO	CA	958182923	RETIRED	RETIRED	100	Παρασκευή, 14 Οκτωβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	MCGINTY, SHIRLEY	RANCHO MIRAGE	CA	922705314	RETIRED	RETIRED	100	Τρίτη, 13 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	KWEI, GLORIA	KENSINGTON	CA	947071320	RETIRED	RETIRED	100	Δευτέρα, 17 Οκτωβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	CRISMAN, AL	CARLSBAD	CA	920114708	RETIRED	RETIRED	100	Δευτέρα, 19 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	BURRIS, BETTY M.	MONTEREY PARK	CA	917554003	RETIRED	RETIRED	100	Παρασκευή, 28 Οκτωβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	SANKORE, SHELBY L	POMONA	CA	917664840	RETIRED	RETIRED	100	Δευτέρα, 12 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	LOUIS, JERRY	ENCINITAS	CA	920245205	RETIRED	RETIRED	100	Δευτέρα, 3 Οκτωβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	ABRAHAMSE, DEE	LONG BEACH	CA	908153503	RETIRED	RETIRED	100	Τετάρτη, 28 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	JOHNSON, STANLEY	FULLERTON	CA	928351407	RETIRED	RETIRED	100	Τετάρτη, 23 Νοεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	BALSAM, DAVID	OAKLAND	CA	946181329	RETIRED	RETIRED	100	Πέμπτη, 1 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	KRESS, WENDY	LAFAYETTE	CA	94549	RETIRED	RETIRED	100	Σάββατο, 31 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	FELSON, JAMES A	WOODLAND	CA	956954342	RETIRED	RETIRED	100	Παρασκευή, 21 Οκτωβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	RIDDER, LEONARD H.	LAGUNA BEACH	CA	926512840	RETIRED	RETIRED	100	Πέμπτη, 22 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	THREATT, BETH	REDLANDS	CA	92373	RETIRED	RETIRED	100	Τετάρτη, 14 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	FINKELSTEIN, RUTH	SAN MATEO	CA	944034619	RETIRED	RETIRED	100	Τρίτη, 27 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	HENRY, CHARLES R	DANVILLE	CA	945065804	RETIRED	RETIRED	100	Δευτέρα, 12 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	SPALSBURY, MARY ANNE	IDYLLWILD	CA	925490202	RETIRED	RETIRED	100	Τετάρτη, 21 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	MAUDLIN, CHARLES	SAN DIEGO	CA	921265630	RETIRED	RETIRED	100	Παρασκευή, 30 Δεκεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	CORDO, LAWRENCE D	LOS ANGELES	CA	900463408	RETIRED	RETIRED	100	Δευτέρα, 7 Νοεμβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	GORMAN, EILY	SANTA MONICA	CA	904054549	RETIRED	RETIRED	100	Παρασκευή, 21 Οκτωβρίου 2011				SA17A	772372	Democrats	RETIRED
C00431445	P80003338	Obama, Barack	GRAY, DONNA S	TIBURON	CA	949200974	RETIRED	RETIRED	100	Τετάρτη, 19 Οκτωβρίου 2011				SA17A	772372	Democrats	RETIRED

Εικόνα 3-48 – επισκόπηση των δεδομένων από το παράθυρο Data του Power BI

Βήμα 2ο: Εμφάνιση συχνότητας των ποσών

donor_df['contb_receipt_amt'].value_counts()

Έπειτα, για την εμφάνιση των ποσών δωρεάς και τη συχνότητα επιλέγουμε απεικόνιση matrix (Εικόνα 3-49), ώστε να είναι αντίστοιχο της donor_df_['contb_receipt_amt']. value_counts(). Το πρόγραμμα μας επιτρέπει να επιλέξουμε ανάμεσα από άθροισμα, μέσον όρο, αρίθμηση (count), διακριτή αρίθμηση (count distinct), διακύμανση, τυπική απόκλιση, διάμεσο, minimum και maximum.Για την περίπτωση επιλέγουμε το count.

contb_receipt_amt	Count of contb_receipt_amt
100,00	178188
50,00	137584
25,00	110345
250,00	91182
500,00	57984
2.500,00	49005
35,00	37237
1.000,00	36494
10,00	33986
200,00	27813
20,00	17565
15,00	16163
150,00	14600
75,00	13647
Total	**1001731**

Εικόνα 3-49 – υπολογισμός της συχνότητας εμφάνισης του κάθε ποσού συνεισφοράς στο Power BI

Βήμα 3ο: Υπολογισμός της μέσης συνεισφοράς και της τυπικής απόκλισης

```
don_mean = donor_df['contb_receipt_amt'].mean()
don_std = donor_df['contb_receipt_amt'].std()
```

Εδώ εύκολα επιλέγουμε average και standard deviation για τη στήλη contb_receipt_amount και τα εμφανίζουμε με multi-row cards, ώστε να φαίνεται και η περιγραφή τους, πράγμα που δε συμβαίνει στην απλή card. Η μ.ο. των συνεισφορών υπολογίζεται σε 298.24 και η τυπική απόκλιση σε 3749,66 .

Βήμα 4ο: Απαλλοιφή των αρνητικών τιμών, ταξινόμηση και 10 συχνότερα ποσά δωρεάς

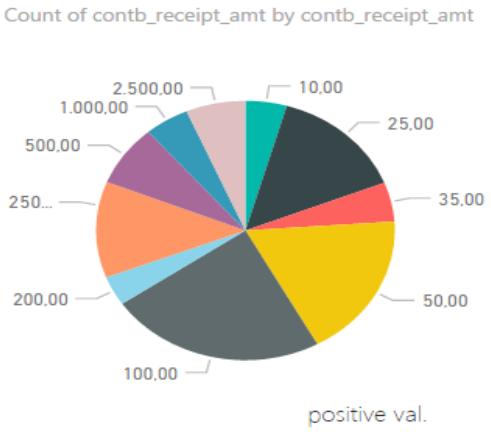
```
top_donor = top_donor[top_donor >0]
top_donor.sort()
top_donor.value_counts().head(10)
```

cmte_id	cand_id	cand_nm	contbr_nm	contbr_city	contbr_st	contbr_zip	contbr_employer	contbr_occupation	contb_receipt_amt
C00431445	P80003338	Obama, Barack	MURPHY, CYNTHIA C.	LITTLE ROCK	AR	722075462			-30800
C00431445	P80003338	Obama, Barack	DAVIS, STEPHEN JAMES	SAN FRANCISCO	CA	941151123			-25800
C00431171	P80003353	Romney, Mitt	KNIGHT, GLADE	RICHMOND	VA	23219			-7500
C00431445	P80003338	Obama, Barack	JEFFERSON, ULYSESE MR.	UPPER MARLBORO	MD	207748578			-5500
C00431445	P80003338	Obama, Barack	RICHARDSON, EUDORA	WASHINGTON	DC	20007			-5455
C00431171	P80003353	Romney, Mitt	LOEFFLER, KELLY	ATLANTA	GA	30305			-5414,31
C00496067	P00003608	Cain, Herman	LUCERO, JEAN	WALESKA	GA	301832895			-5115
C00431445	P80003338	Obama, Barack	KING, HENRY L.	CINCINNATI	OH	452291624	GOVERNMENT	CONSTRUCTION	-5000
C00431445	P80003338	Obama, Barack	FELDMAN, GABRIEL E	NEW YORK	NY	100242349	NYCHSRO	MD	-5000
C00500587	P20003281	Perry, Rick	MORAN, JOHN D. JR	LEWISBURG	PA	17837			-5000
C00496497	P60003654	Gingrich, Newt	WHEELESS, CLIFFORD JR.	BALTIMORE	MD	21210			-5000
C00500587	P20003281	Perry, Rick	ADNAN, AWAD	ELGIN	TX	78621			-5000
C00494393	P20002556	Pawlenty, Timo	LEATHERDALE, LOUISE	LONG LAKE	MN	55356			-5000
C00500587	P20003281	Perry, Rick	WHITED, MICHAEL	FRANKLIN	TN	37064			-5000
C00500587	P20003281	Perry, Rick	MCCALMON, RODGAR SR	NASHVILLE	TN	37220			-5000
C00431171	P80003353	Romney, Mitt	EFRON, PAUL	LARCHMONT	NY	10538			-5000
C00431171	P80003353	Romney, Mitt	GALLEY, RICHARD	PALM BEACH	FL	33480			-5000
C00431171	P80003353	Romney, Mitt	TOWER, SHIRLEY	NORTH PALM BEACH	FL	33408			-5000
C00431171	P80003353	Romney, Mitt	DEPAOLA, THOMAS	LEBANON	NJ	8833			-5000
C00431171	P80003353	Romney, Mitt	CICIRELLI, MARK	NEW YORK	NY	10023			-5000
C00431171	P80003353	Romney, Mitt	CHRISTENSEN, CLAYTON	BELMONT	MA	2478			-5000

Εικόνα 3-50 – αύξουσα ταξινόμηση για τα ποσά των συνεισφορών στο παράθυρο Data του Power BI

Εδώ (Εικόνα 3-50) η αρχή της αύξουσας ταξινόμησης της στήλης contb_receipt_amount, για ολόκληρο το σύνολο.

Στο σημείο αυτό, χρησιμοποιώντας το query editor δημιουργούμε ένα νέο πανομοιότυπο (duplicate) σύνολο δεδομένων, ώστε να μην αλλοιωθούν τα γράφημα που δημιουργήσαμε, στο οποίο θα εμφανίζονται μόνο οι θετικές τιμές του contb_receipt_amount μετά από την εφαρμογή φίλτρου αριθμών (number filters). Ονομάζουμε το νέο σύνολο Donor_Data_positive, και μετά την εφαρμογή των αλλαγών, το έχουμε στη διάθεση μας για το Power BI. Εκεί, το ταξινομούμε άμεσα στο παράθυρο Data. Τις 10 συχνότερα εμφανιζόμενες θετικές τιμές, επιλέγουμε να τις απεικονίσουμε με ένα pie chart (Εικόνα 3-51), στο οποίο εμφανίζονται οι αριθμήσεις (counts) ως legend.

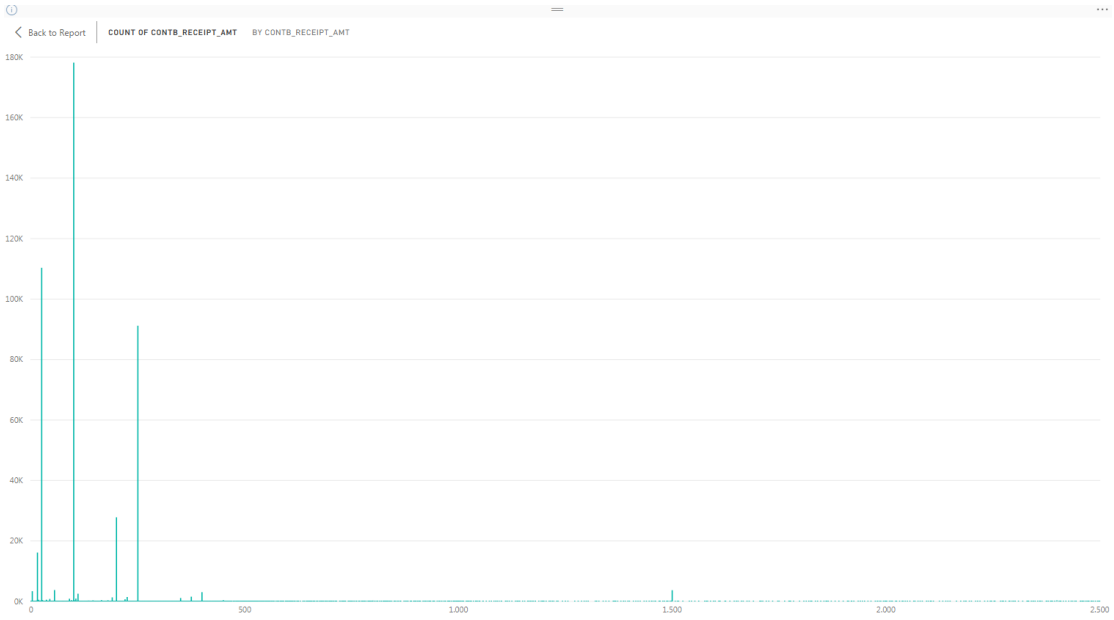


Εικόνα 3-51 – διάγραμμα «πίττας» του Power BI, για τα 10 συχνότερα εμφανιζόμενα θετικά ποσά

Βήμα 5^{ο} :Περιορισμός ποσών στα 2500 δολάρια και ιστόγραμμα

```
com_don = top_donor[top_donor < 2500]
com_don.hist(bins=100)
```

Εδώ, επιλέγοντας το stacked column chart, όπου στο πεδίο axis τοποθετούμε την contb_receipt_amount με φίλτρο μέχρι την τιμή 2500 και στο πεδίο values το count contb_receipt_amount (Εικόνα 3-52). Ωστόσο, παρατηρούμε πως εμφανίζονται κάποιες μόνο τιμές, και οι περισσότερες λείπουν (οι τιμές που εμφανίζονται διαφέρουν ανάλογα με το κατώφλι του φίλτρου) .Το power BI σε αυτή την περίπτωση δε φαίνεται να προσφέρει κάποια εύχρηστη επιλογή.

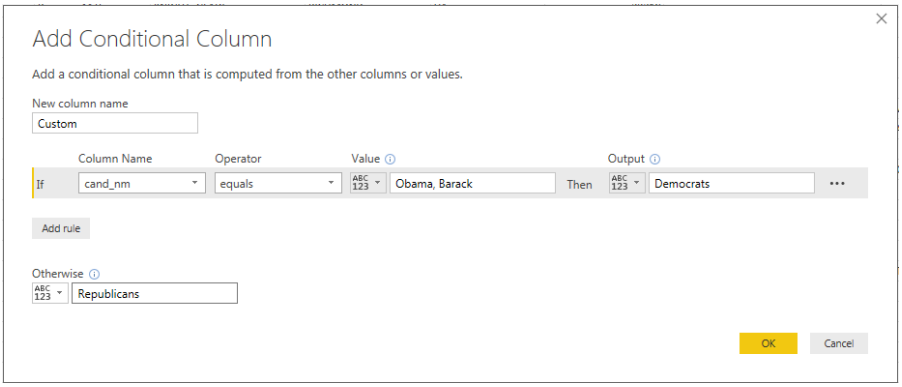


Εικόνα 3-52 – διάγραμμα του Power BI για τη συχνότητα εμφάνισης κάθε ποσού συνεισφοράς

Βήμα 6ο: Δημιουργία πεδίου αντιστοίχισης κομμάτων και αρίθμηση συνεισφορών ανά υποψήφιο

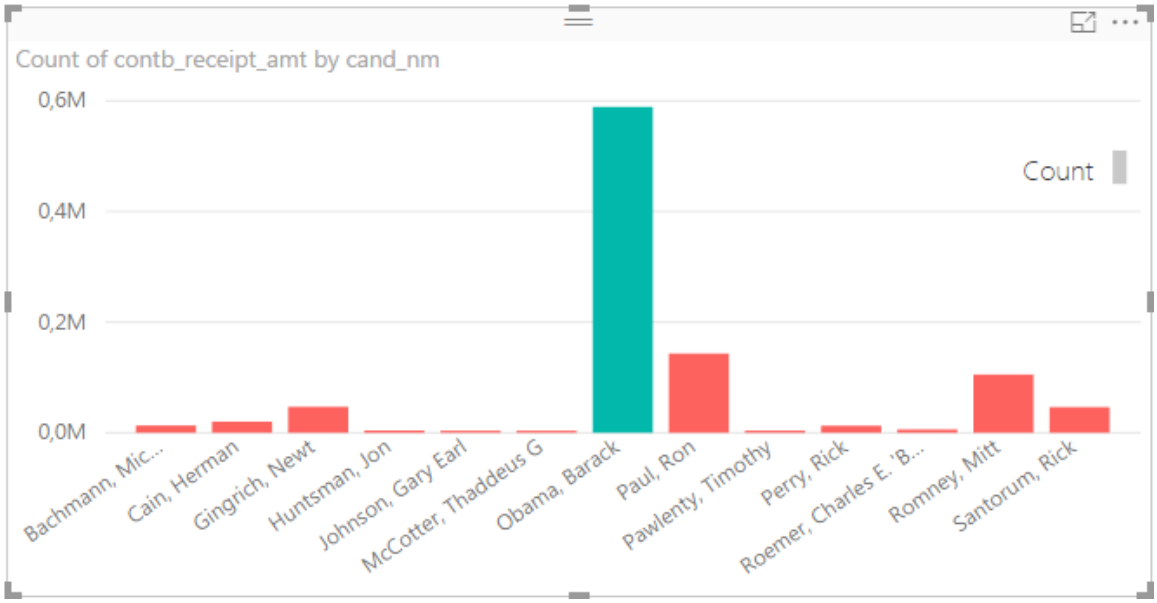
```
donor_df['Party'] = donor_df.cand_nm.map(party_map)
donor_df.groupby('cand_nm')['contb_receipt_amt'].count()
```

Εύκολα μπορούμε να προσθέσουμε μια νέα στήλη «Party», στο query editor (Εικόνα 3-53). Επιλέγουμε Add conditional column, όπου δημιουργούμε νέα στήλη με τιμές που εξαρτώνται από άλλη στήλη, για την αντίστοιχη εγγραφή.



Εικόνα 3-53 – δημιουργία νέας στήλης υπό συνθήκη στον Query Editor του Power BI

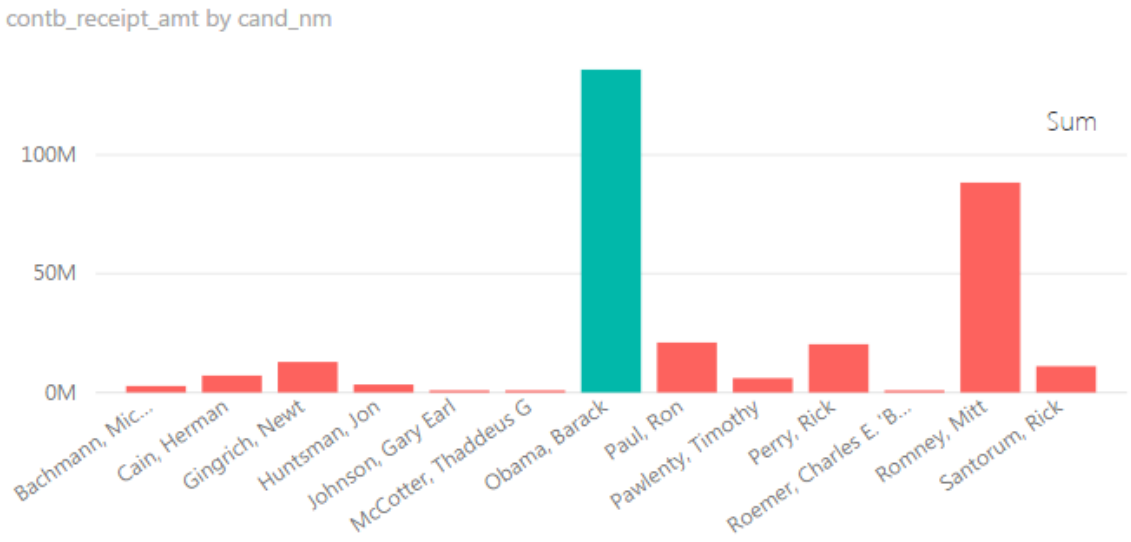
Άμεσα, χρησιμοποιώντας το γράφημα stacked column chart, στον άξονα cand_nm και contb_receipt_amount count στις τιμές (values), δημιουργούμε διάγραμμα (Εικόνα 3-54) με την αρίθμηση των συνεισφορών σε κάθε υποψήφιο, χωρίς να χρειαστεί να προηγηθεί του διαγράμματος εντολή groupby cand_nm .



Εικόνα 3-54 – διάγραμμα με τον αριθμό των συνεισφορών για κάθε υποψήφιο στο Power BI

Βήμα 7ο: Διάγραμμα με τα συνολικά αθροιστικά ποσά ανά υποψήφιο
donor_df.groupby('cand_nm')['contb_receipt_amt'].sum()

Ακριβώς την ίδια διαδικασία ακολουθούμε για το άθροισμα των συνεισφορών ανά υποψήφιο (Εικόνα 3-55).

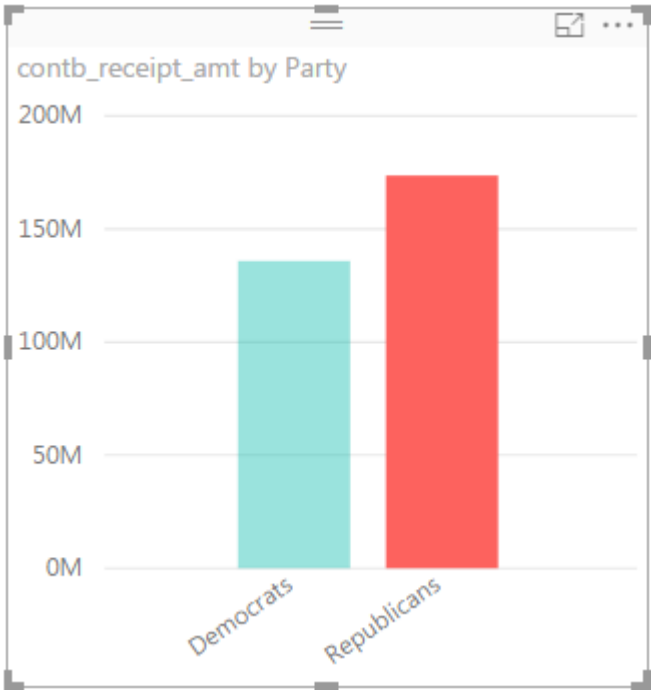


Εικόνα 3-55 – διάγραμμα του Power BI με το άθροισμα των συνεισφορών ανά υποψήφιο

Βήμα 8ο: Διάγραμμα αθροιστικού ποσού συνεισφορών ανά κόμμα

```
donor_df.groupby('Party')['contb_receipt_amt'].sum().plot(kind='bar')
```

Το ίδιο εύκολα δημιουργείται και το ραβδόγραμμα που αφορά τα ποσά συνεισφοράς για τα δύο κόμματα (Εικόνα 3-56).



Εικόνα 3-56 – διάγραμμα του Power BI με το άθροισμα των συνεισφορών ανά παράταξη

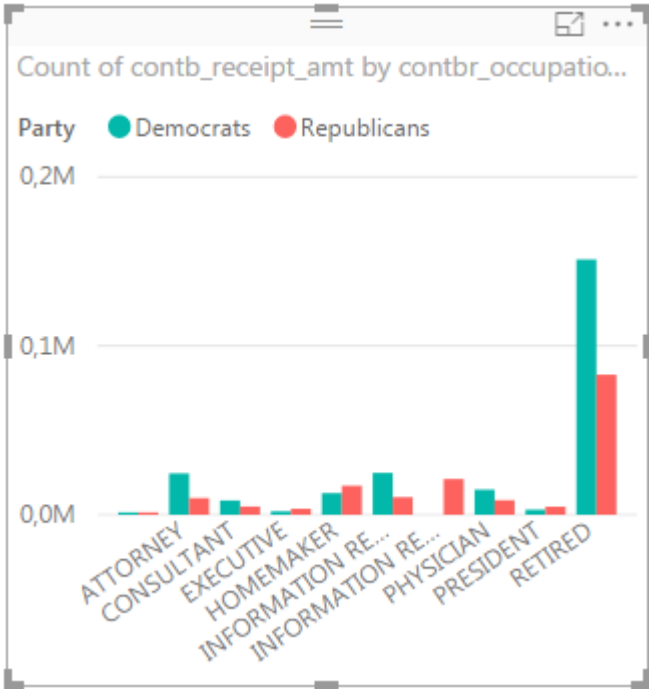
Βήμα 9ο: Ελάχιστο κατώφλι 1,000,000 και οριζόντιο clustered διάγραμμα

```
occupation_df = occupation_df[occupation_df.sum(1) > 1000000]
```

Χρησιμοποιώντας πίνακα τύπου matrix, με σειρές τα επαγγέλματα των συνεισφερόντων, στήλες τα κόμματα που υποστηρίζουν, τιμές την αρίθμηση (count) και φίλτρο τις top 10 μεγαλύτερες αριθμήσεις (Εικόνα 3-57), προκύπτει χωρίς προεπεξεργασία το επιθυμητό αποτέλεσμα. Για τα δεδομένα που εμφανίζονται, δημιουργούμε ραβδόγραμμα συστοιχισμένων στηλών (Εικόνα 3-58).

contbr_occupation	Democrats	Republicans	**Total**
RETIRED	151115	82875	**233990**
INFORMATION REQUESTED	24747	10360	**35107**
ATTORNEY	24451	9835	**34286**
HOMEMAKER	12773	17158	**29931**
PHYSICIAN	14845	8587	**23432**
INFORMATION REQUESTED PER BEST EFFORTS		21138	**21138**
ENGINEER	5424	8910	**14334**
TEACHER	11163	2827	**13990**
CONSULTANT	8430	4843	**13273**
PROFESSOR	11545	1010	**12555**
Total	**264493**	**167543**	**432036**

Εικόνα 3-57 – πίνακας matrix του Power BI για τα επαγγέλματα των δωρητών

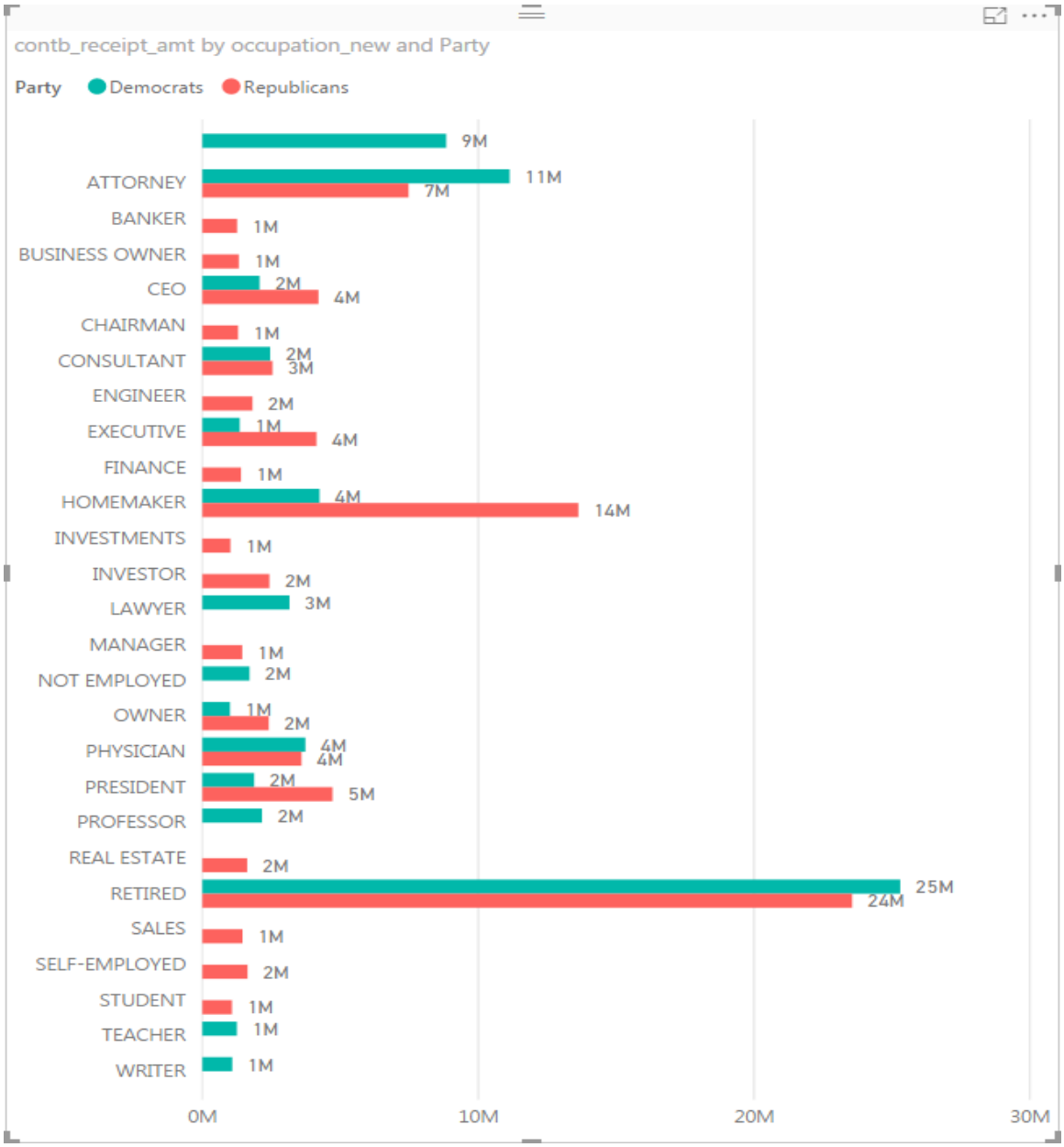


Εικόνα 3-58 – διάγραμμα συστοιχισμένων στηλών του Power BI για τα επαγγέλματα των δωρητών

Εμφανίζουμε τον αριθμό των επαγγελμάτων σε μια multi-row card, όπου υπολογίζονται σε 45071 διαφορετικά.

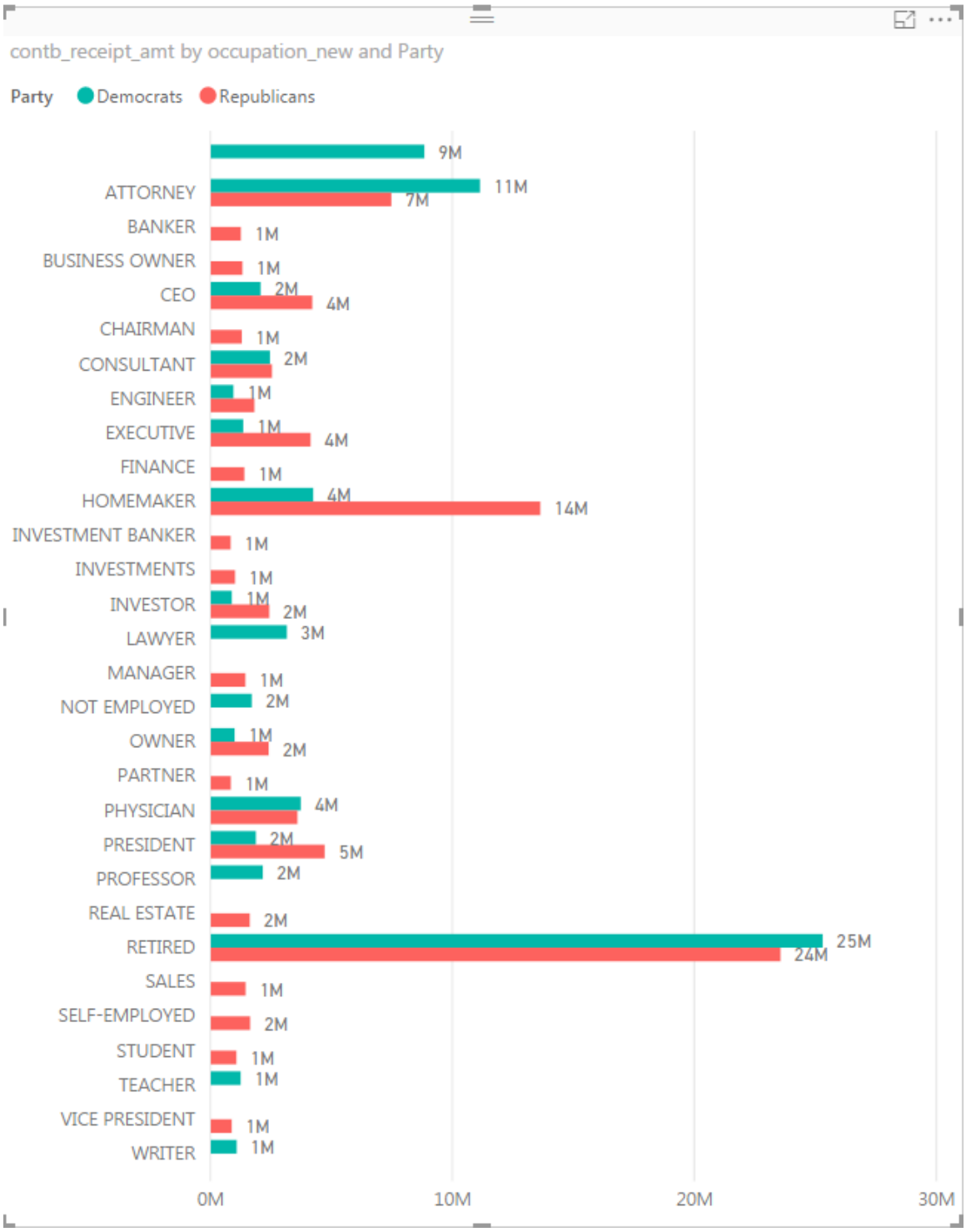
Δημιουργούμε το clustered bar chart, με φίλτρο 1,000,000 minimum συνεισφορές ανά επάγγελμα. Για να απαλείψουμε τα info requested προσθέτουμε φίλτρο στον άξονα «does not contain: information». Για να συγχωνεύσουμε CEO και C.E.O. , δημιουργούμε στο power

query conditional column, if contbr_occupation is C.E.O. then CEO , else contbr_occupation και την ονομάζουμε occupation_new.



Εικόνα 3-59 – 1ο οριζόντιο διάγραμμα του Power BI για τα επαγγέλματα των δωρητών (κατόπιν φιλτραρίσματος)

Εδώ (Εικόνα 3-60) εμφανίζεται το εξής πρόβλημα: εφαρμόζοντας φίλτρο 1,000,000 , παρατηρούμε ότι εφαρμόζεται στην τιμή μετά από τον κομματικό διαχωρισμό (ενώ είναι επιθυμητό συνολικά, ανά επάγγελμα), και έτσι δεν εμφανίζονται κατηγορίες, π.χ. vice president. Τυπικά, φιλτράρουμε στις 800,000, απλά για να δείξουμε τις κατηγορίες επαγγελμάτων όπως στον αρχικό κώδικα.



Εικόνα 3-60 - 2ο οριζόντιο διάγραμμα του Power BI για τα επαγγέλματα των δωρητών (κατόπιν φιλτραρίσματος)

3.3.3 Qlik Sense

Εισαγωγή του csv αρχείου

Η εισαγωγή δεδομένων γίνεται εύκολα, εν προκειμένω του csv, με προεπισκόπηση των δεδομένων και δυνατότητα αφαίρεσης στηλών (πεδίων) εφόσον είναι επιθυμητό. Το πρώτο στάδιο της αλυσίδας επεξεργασίας των δεδομένων, data load editor, είναι ένα script που επιτρέπει τον καθορισμό των παραμέτρων εισαγωγής. Στο σενάριο αυτό αλλάξαμε την προκαθορισμένη στίξη δεκαδικών σε ‘ . ‘ και των χιλιάδων σε ‘,’ . Το αρχείο εισήχθηκε πολύ γρήγορα για το μέγεθος του, σε σύγκριση με τα άλλα εργαλεία, στους ίδιους υπολογιστικούς πόρους.

Βήμα 1^{ο}: Εισαγωγή του csv αρχείου

Ο επόμενος «κρίκος» στην επεξεργασία δεδομένων, αντιστοιχεί στο «παράθυρο» του data manager (Εικόνα 3-61). Έχουμε συγκεντρωμένα τους όποιους πίνακες έχουν εισαχθεί, και την εποπτεία αυτών.

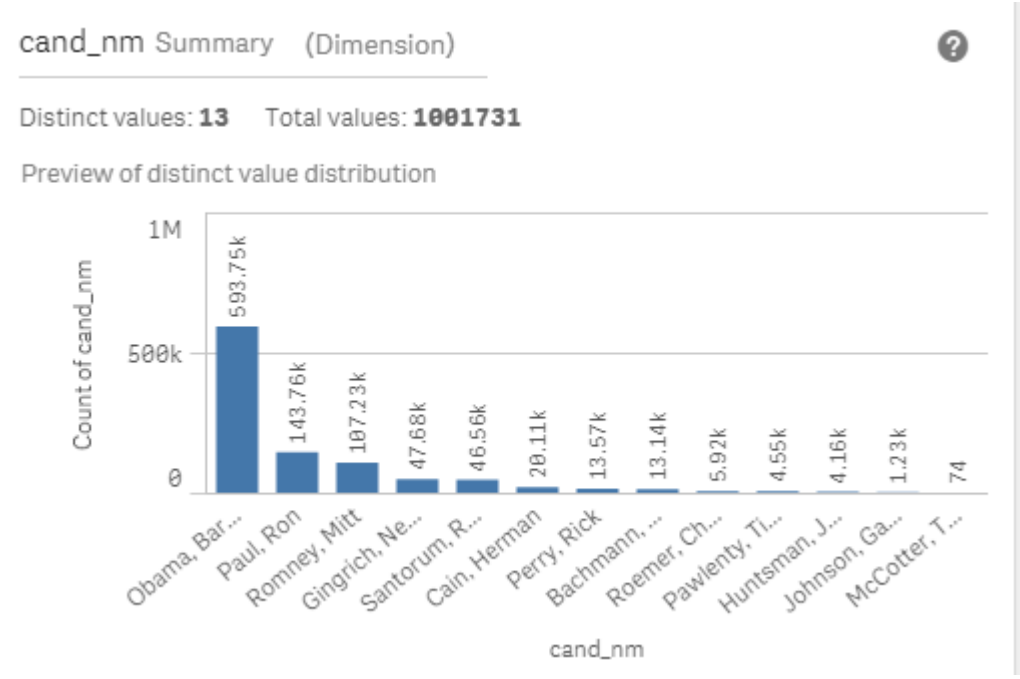
Δεν μπορούμε να αλλάξουμε τον τύπο των δεδομένων μέσω του εργαλείου (number,string,decimal…) παρά μόνο να χαρακτηρίσουμε ως general, date,timestamp και geo data. Οπότε, είναι ιδιαίτερα σημαντικό η εισαγωγή τους να γίνει σωστά (κατάλληλο script στο load manager).

cmte_id	cand_id	cand_nm	contbr_nm	contbr_city	contbr_st	contbr_zip	contbr_employer
C00410118	P20002978	Bachmann, Michelle	ADAMS, JAY	DAYTONA BEACH	FL	321184948	NONE
C00410118	P20002978	Bachmann, Michelle	ADAMS, KENNETH	JACKSONVILLE	FL	322255412	SELF
C00410118	P20002978	Bachmann, Michelle	AGER, GREGORY	HAMILTON	VA	201583311	UPDATA CAPITAL INC
C00410118	P20002978	Bachmann, Michelle	AHRENS, THOMAS	BUFFALO	MN	553134455	BUFFALO CLINIC PA
C00410118	P20002978	Bachmann, Michelle	ALBERSSCHONBERG, JOYCE	PRINCETON	NJ	085406764	NONE

Εικόνα 3-61 – εποπτεία των δεδομένων στο παράθυρο data manager του Qlik sense

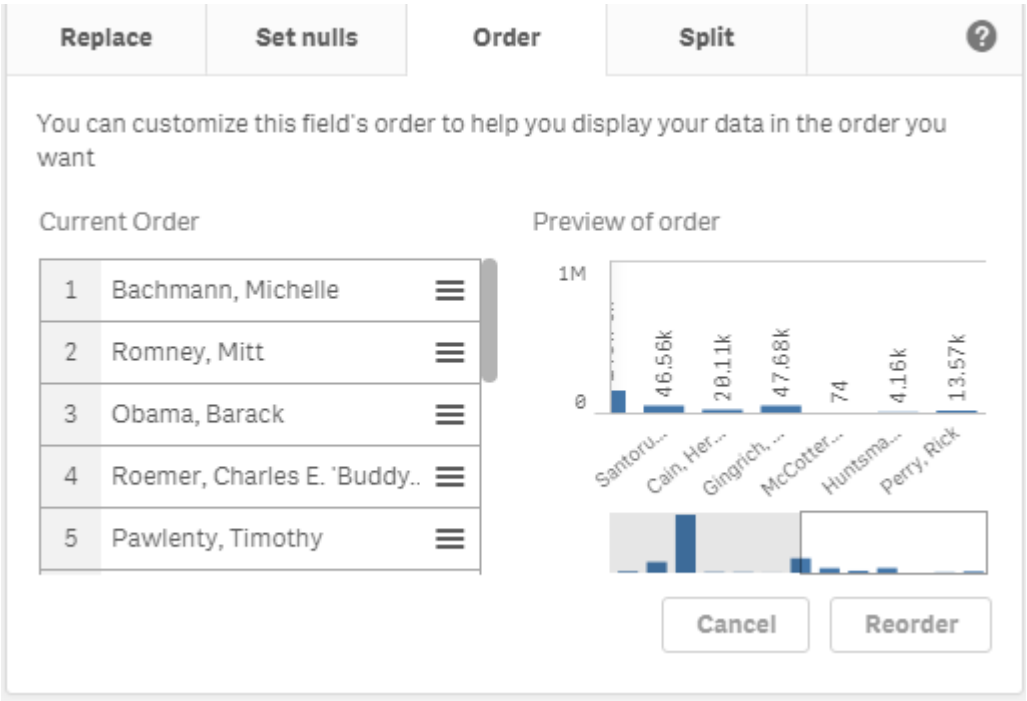
Μία ιδιαίτερη παροχή του qlik sense είναι τα αυτόματα εμφανιζόμενα insights στο παράθυρο data manager, που δίνουν τις απαντήσεις σε πολλά από τα ερωτήματα που τίθενται, ανάλογα με το επιλεχθέν πεδίο.

Επιλέγοντας τη στήλη με τα ονόματα των υποψηφίων (Εικόνα 3-62), cand_nm (εμφανίζει την αρίθμιση του ίδιου μεγέθους):



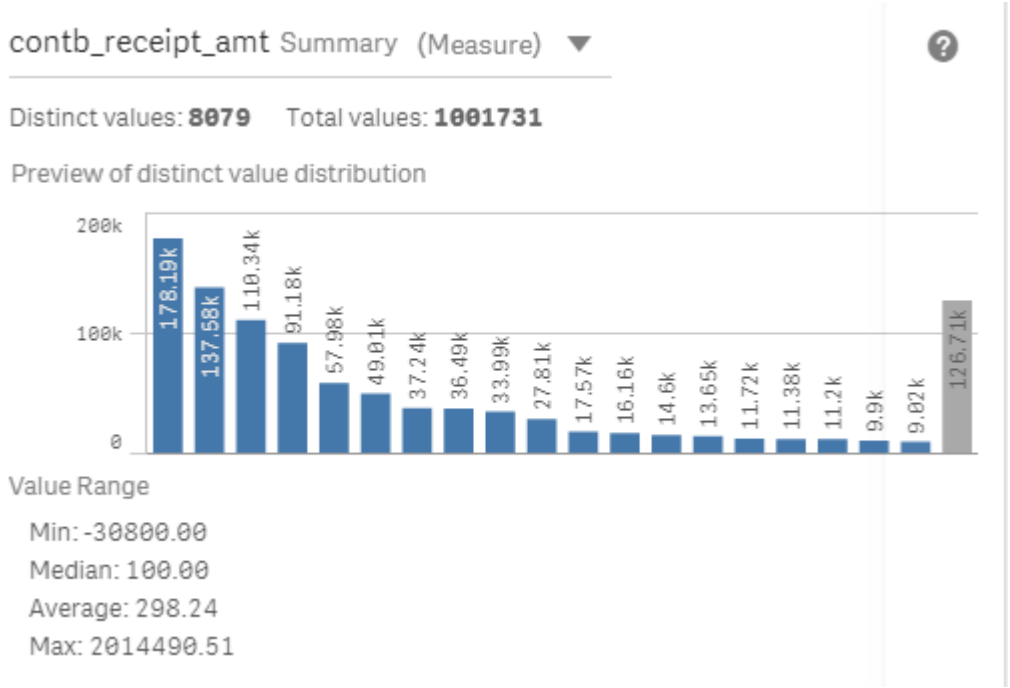
Εικόνα 3-62 – αρίθμηση της εμφάνισης κάθε ονόματος υποψηφίου, insight του data manager

Όπως και αντίστοιχο πλαίσιο επεξεργασίας (Εικόνα 3-63)



Εικόνα 3-63 – πλαίσιο επεξεργασίας των insights του data manager

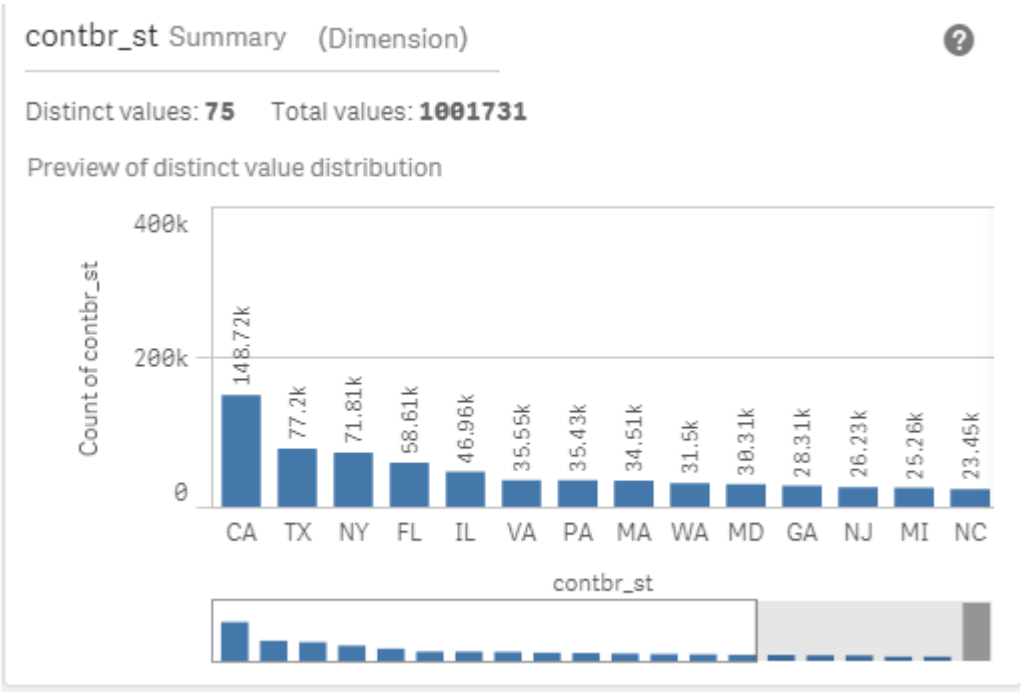
Για την στήλη των συνεισφορών (Εικόνα 3-64) εμφανίζει άθροισμα:



Εικόνα 3-64 – συχνότητα εμφάνισης ανά ποσό συνεισφοράς, insight του data manager

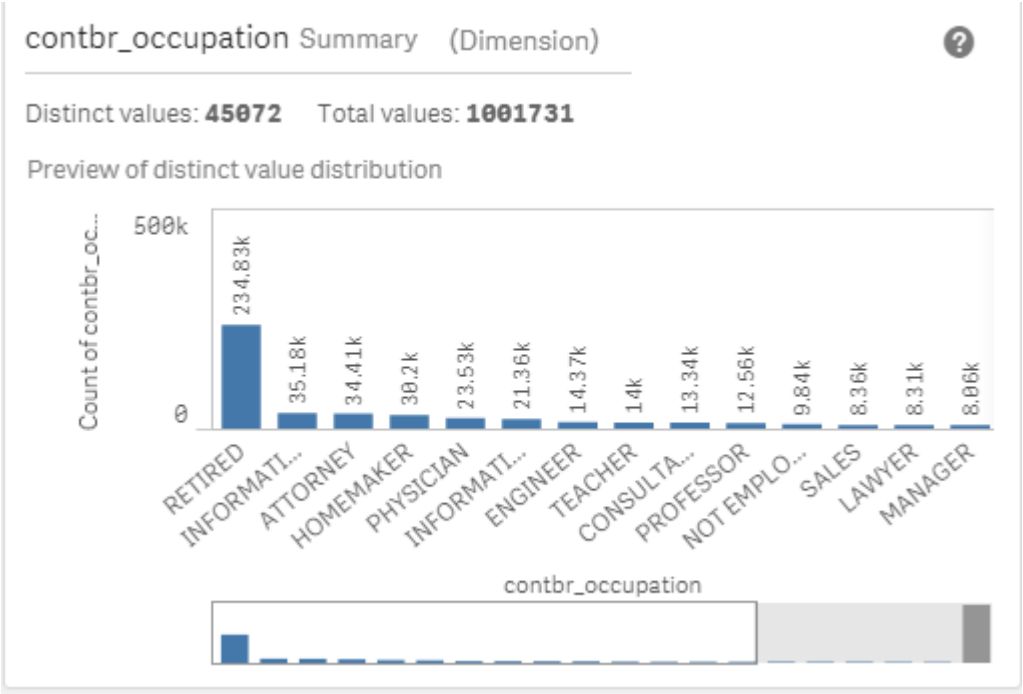
Όπου το ποσό φαίνεται με roll over.

Επίσης, states με τις πιο μεγάλες δωρεές (Εικόνα 3-65) :



Εικόνα 3-65 – αρίθμηση συνεισφορών ανά πολιτεία, insight του data manager

Και επαγγέλματα (Εικόνα 3-66):



Εικόνα 3-66 αρίθμηση συνεισφορών ανά επάγγελμα, insight του data manager

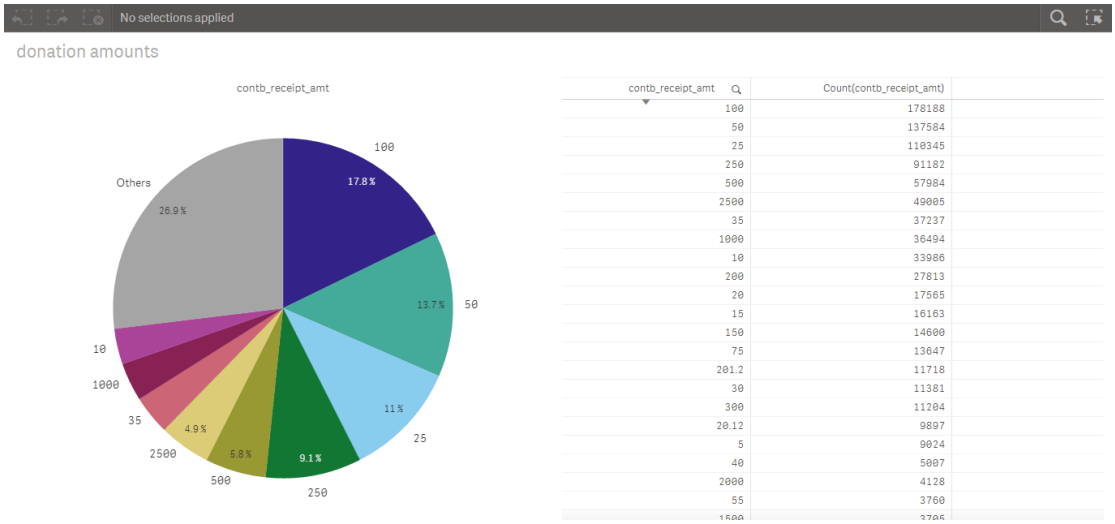
Εμφανίζονται αυτόματα, όλα με summaries.

Τα διαγράμματα αυτά, καλύπτουν ήδη μεγάλο μέρος της πληροφορίας που θέλουμε να μεταδώσουμε, και μάλιστα μερικά είναι σχεδόν ίδια με αυτά που σκοπεύουμε να δημιουργήσουμε.

Βήμα 2^{ο}: Εμφάνιση συχνότητας των ποσών

donor_df['contb_receipt_amt'].value_counts()

Για τη γραφική αναπαράσταση των ποσών συνεισφοράς, επιλέγουμε ένα «διάγραμμα πίττας» (Εικόνα 3-67), συνοδευόμενο από πίνακα table. Στο διάγραμμα εμφανίζονται τα 9 συχνότερα εμφανιζόμενα ποσά- με τα ποσοστά τους, καθώς και τι ποσοστό καταλαμβάνουν τα υπόλοιπα (others). Στον πίνακα εμφανίζεται ο αριθμός της συχνότητας εμφάνισης, με φθίνουσα ταξινόμηση.



Εικόνα 3-67 – διάγραμμα «πίττας» και πίνακας του Qlik Sense, για τα συχνότερα εμφανιζόμενα ποσά

Βήμα 3ο: Υπολογισμός της μέσης συνεισφοράς και της τυπικής απόκλισης

```
don_mean = donor_df['contb_receipt_amt'].mean()
don_std = donor_df['contb_receipt_amt'].std()
```

Τα δύο μεγέθη υπολογίζονται ως «master items». Για το μέσο όρο η συνάρτηση είναι έτοιμη ως aggregation, σε πλήκτρο, και υπολογίζεται σε 298.2. Για την τυπική απόκλιση, η οποία υπολογίζεται 3749.7, χρειάζεται να πληκτρολογήσουμε τη συνάρτηση, ενώ στα άλλα εργαλεία είναι έτοιμη, ως επιλογή.

Βήμα 4ο: Ταξινόμηση

Εικόνα 3-68 – ταξινόμηση, φθίνουσα και αύξουσα, για τις τιμές των ποσών στο Qlik Sense

Εκ των δεξιών φθίνουσα και εκ των αριστερών αύξουσα ταξινόμηση (Εικόνα 3-68). Για λόγους εποπτείας του δημιουργού, θα μπορούσε να είχε γίνει απευθείας στο data manager.

Βήμα 5ο: Απαλλοιφή των αρνητικών τιμών, ταξινόμηση και 10 συχνότερα ποσά δωρεάς

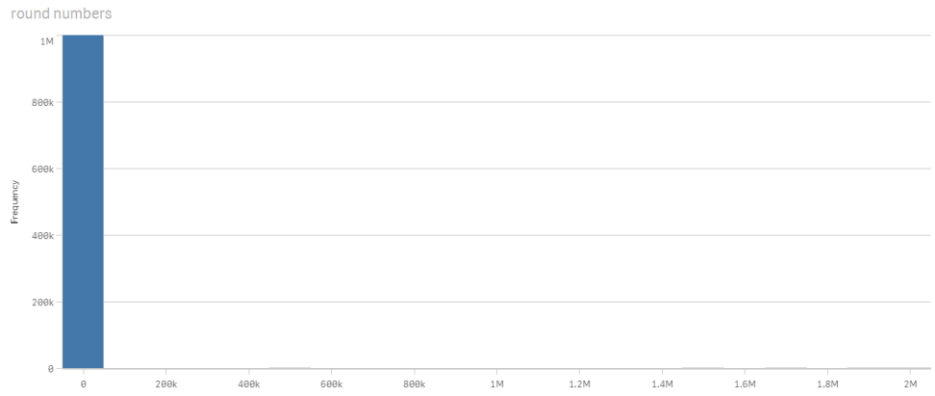
```
top_donor = top_donor[top_donor >0]
top_donor.sort()
top_donor.value_counts().head(10)
```

Εδώ, δεν μπορούμε να φιλτράρουμε, σε κεντρικό επίπεδο, τα δεδομένα μας, ώστε να κρατήσουμε μόνο τα θετικά. Οι δέκα συχνότερες συνεισφορές, εμφανίζονται ήδη πιο πάνω.

Περιορισμός ποσών στα 2500 δολάρια και ιστόγραμμα

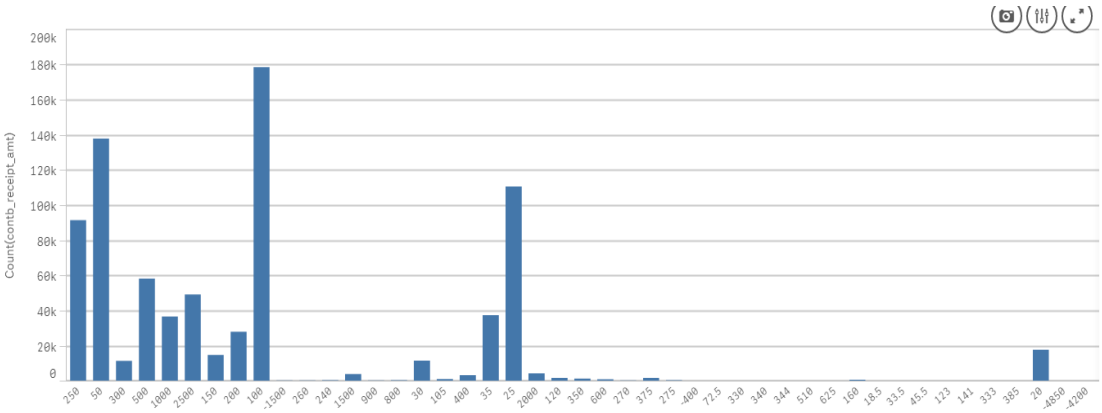
```
com_don = top_donor[top_donor < 2500]
com_don.hist(bins=100)
```

Σε αυτό το διάγραμμα (Εικόνα 3-69) δεν καλυπτόμαστε από το εργαλείο. Τυπικά, θα χρησιμοποιούσαμε ιστόγραμμα, για να εμφανίσουμε τη συχνότητα εμφάνισης των ποσών. Ωστόσο, δεν έχουμε δυνατότητα φιλτραρίσματος στο ιστόγραμμα, με αυτό να εκτίνεται σε πολύ μεγάλο άξονα (αναμενόμενο από την τεράστια τυπικά απόκλιση).



Εικόνα 3-69 – ιστόγραμμα του Qlik Sense, για τη συχνότητα εμφάνισης των ποσών

Όμως, το φίλτρο δεν εφαρμόζεται σωστά ούτε στο ραβδόγραμμα (Εικόνα 3-70), αφού δεν μπορούμε να το εφαρμόσουμε στα μεγέθη του χ άξονα (dimension), παρά μόνο στα ποσοτικά μεγέθη (measures). Ακόμα και στα φίλτρα που είναι δυνατό να εφαρμόσουμε, περιοριζόμαστε σε μία συνθήκη και δε θα μπορούσαμε να συνδυάσουμε την απαίτηση για θετικά και μικρότερα-ίσα του 2500.



Εικόνα 3-70 – ραβδόγραμμα του Qlik Sense, για τη συχνότητα εμφάνισης των ποσών με εφαρμογή φίλτρου

Βήμα 6ο: Δημιουργία πεδίου αντιστοίχισης κομμάτων και αρίθμηση συνεισφορών ανά υποψήφιο

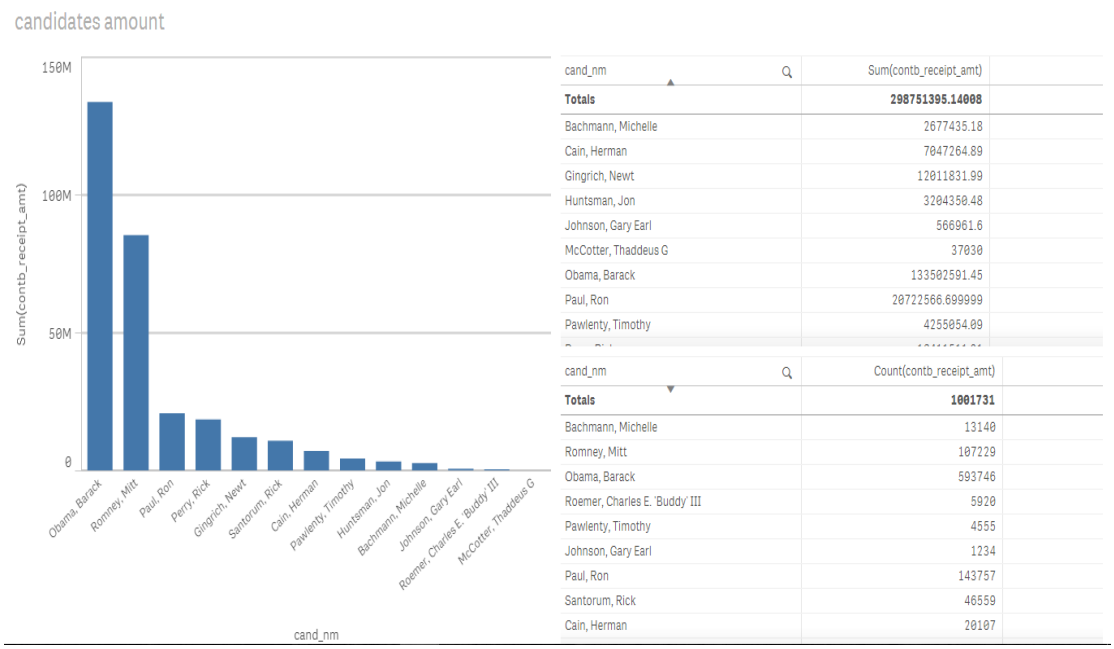
```
donor_df['Party'] = donor_df.cand_nm.map(party_map)
donor_df.groupby('cand_nm')['contb_receipt_amt'].count()
```

Στο παράθυρο data manager, δημιουργούμε μία νέα στήλη, add calculated field. Με εντολή, χρησιμοποιούμε την συνάρτηση συνθήκης if και τη συνάρτηση εντοπισμού χαρακτήρων «wildmatch», του qlik sense αμφότερες. Οι εντολές και οι συναρτήσεις του qlik είναι αρκετά «ιδιότροπα» και λιγότερο άμεσα και εύχρηστα.

Βήμα 7ο: Διάγραμμα με τα συνολικά αθροιστικά ποσά ανά υποψήφιο

```
donor_df.groupby('cand_nm')['contb_receipt_amt'].sum()
```

Το group by στο qlik sense γίνεται με τη χρήση εντολών, δεν υπάρχει δηλαδή πλήκτρο ή άλλη διευκόλυνση. Δεν είναι απαραίτητο ωστόσο για τα επιθυμητά γραφήματα.



cand_nm	Sum(contb_receipt_amt)
Totals	**298751395.14008**
Bachmann, Michelle	2677435.18
Cain, Herman	7047264.89
Gingrich, Newt	12011831.99
Huntsman, Jon	3204350.48
Johnson, Gary Earl	566961.6
McCotter, Thaddeus G	37030
Obama, Barack	133592591.45
Paul, Ron	20722566.699999
Pawlenty, Timothy	4255054.09

cand_nm	Count(contb_receipt_amt)
Totals	**1001731**
Bachmann, Michelle	13140
Romney, Mitt	107229
Obama, Barack	593746
Roemer, Charles E. 'Buddy' III	5920
Pawlenty, Timothy	4555
Johnson, Gary Earl	1234
Paul, Ron	143757
Santorum, Rick	46559
Cain, Herman	20107

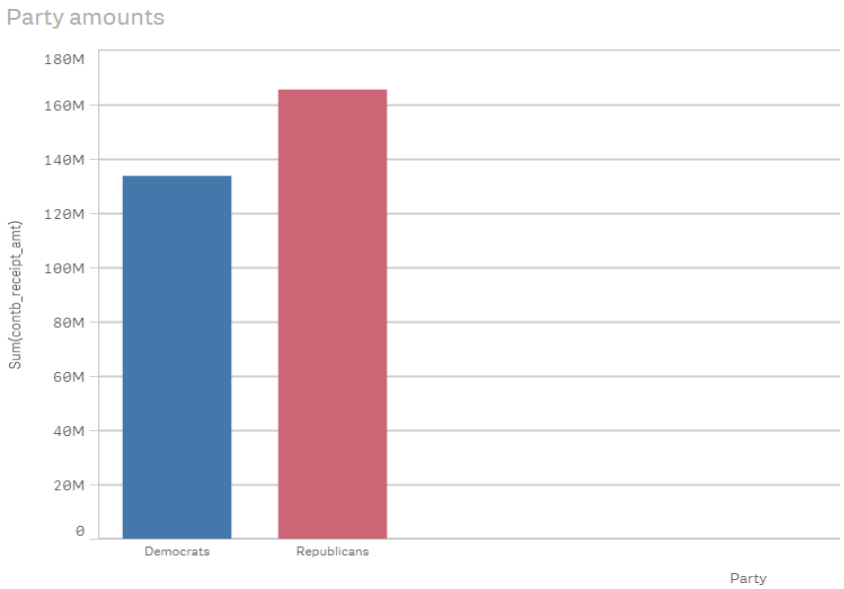
Εικόνα 3-71 – διάγραμμα και πίνακες του Qlik Sense, με το άθροισμα των συνεισφορών ανά υποψήφιο

Για την καίρια απεικόνιση, χρησιμοποιούμε ραβδόγραμμα με τα αθροίσματα, τα οποία αναλυτικά παραθέτουμε σε ταξινομημένο πίνακα (Εικόνα 3-71). Πίνακα επίσης δημιουργούμε και για το αριθμήσιμα (count) των λαμβανόμενων ποσών από κάθε υποψήφιο.

Βήμα 8ο: Διάγραμμα αθροιστικού ποσού συνεισφορών ανά κόμμα

```
donor_df.groupby('Party')['contb_receipt_amt'].sum().plot(kind='bar')
```

Στο ραβδόγραμμα εμφανίζεται το άθροισμα των συνεισφορών για τα δύο κόμματα (Εικόνα 3-72).

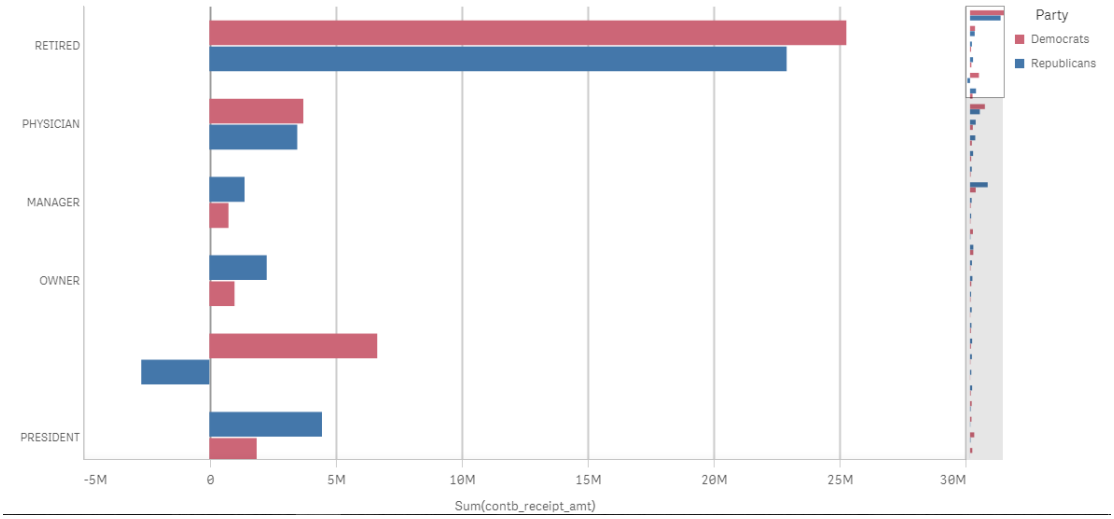


Εικόνα 3-72 – διάγραμμα του Qlik Sense, με το άθροισμα των συνεισφορών για κάθε παράταξη

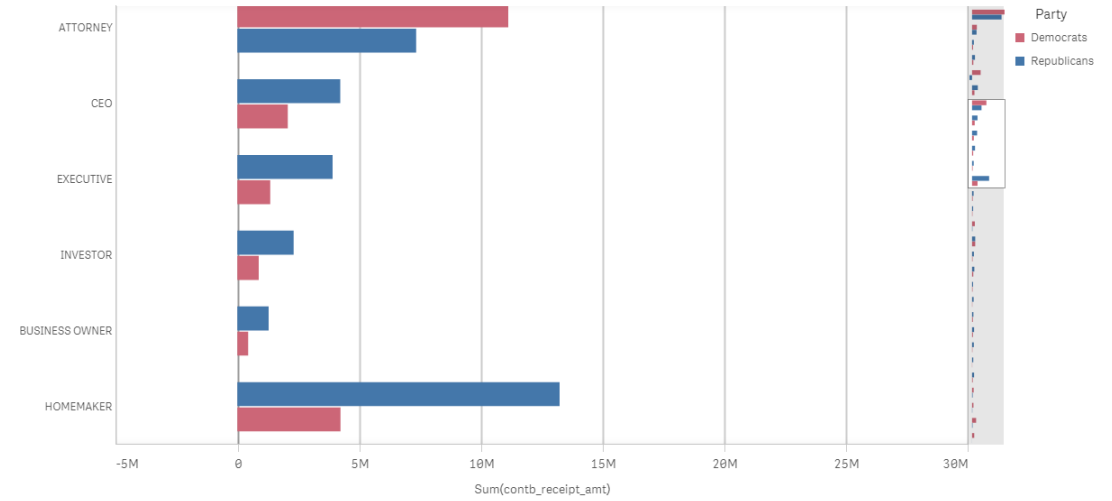
Βήμα 9ο: Ελάχιστο κατώφλι 1,000,000 και οριζόντιο clustered διάγραμμα

```
occupation_df = occupation_df[occupation_df.sum(1) > 1000000]
```

Δημιουργούμε μία στήλη occupation, όπου θέτουμε κενές (null) (Εικόνα 3-73) τις εγγραφές που ξεκινάν με «information» και, επίσης, συγχωνεύουμε τα «CEO» και «C.E.O» ως «CEO» (Εικόνα 3-74). Επιλέγουμε στις ρυθμίσεις απεικόνισης να μην εμφανίζονται οι κενές τιμές. Παρατηρούμε ότι υπάρχει μία εγγραφή χωρίς όνομα. Αυτή οφείλεται σε καταχωρήσεις με κενό χαρακτήρα στο πεδίο των επαγγελμάτων. Δεν ήταν δυνατή η μετατροπή αυτών σε κενές (nulls), που συμβολίζονται με « - ». Ο οριζόντιος προσανατολισμός των διαγραμμάτων έγινε εύκολα, με πλήκτρο, στο παράθυρο appearance.



Εικόνα 3-73 -1ο οριζόντιο διάγραμμα του Qlik Sense για τα επαγγέλματα των δωρητών (κατόπιν φιλτραρίσματος)



Εικόνα 3-74 - 2ο οριζόντιο διάγραμμα του Qlik Sense για τα επαγγέλματα των δωρητών (κατόπιν φιλτραρίσματος)

3.4 Σενάρια Γεωγραφικών Δεδομένων

Τα γεωγραφικά δεδομένα είναι είτε συντεταγμένες σημείων, είτε ονόματα χωρών, περιοχών, πόλεων και ταχυδρομικοί κώδικες που αντιστοιχούν σε πολύγωνα πάνω στο χάρτη. Τα εργαλεία διαθέτουν αναπαραστάσεις χάρτη για σημεία και για περιοχές, με διαβάθμιση χρωμάτων για να τις αντιστοιχίσουν σε διάφορα μεγέθη. Κάθε εργαλείο χρησιμοποιεί συγκεκριμένο χάρτη, όπως google maps, open street maps κ.α. Επίσης, διαθέτει διάφορα φόντα, όπως, π.χ., οδικό δίκτυο, ακτογραμμή, δασικές εκτάσεις.

Για τη σύγκριση των εργαλείων όσον αφορά τα γεωγραφικά δεδομένα, επιλέξαμε τρία σενάρια. Στο πρώτο από αυτά εξετάζεται η απεικόνιση σημείων σε θέσεις εντός πόλης, και παροχή περεταίρω σχετικών πληροφοριών. Στο επόμενο δημιουργούμε παγκόσμιους χάρτες, όπου με χρωματικές διαβαθμίσεις αναπαρίστανται κάποια αριθμητικά μεγέθη. Επίσης, στο σενάριο αυτό, για πληρέστερη κατανόηση και εμβάθυνση στις πληροφορίες, χρησιμοποιούνται χρονοσειρές και treemaps. Στο τρίτο σενάριο , σε έναν χάρτη των Η.Π.Α, αξιολογούμε τη δυνατότητα εστίασης (zoom), από την κλίμακα μίας ηπείρου στους δρόμους μίας πόλης, και τη δυνατότητα αναπαράστασης μεγάλου αριθμού δεδομένων (σημείων).

3.4.1 Χάρτης Πανεπιστημιακών Κτηρίων Αθηνών

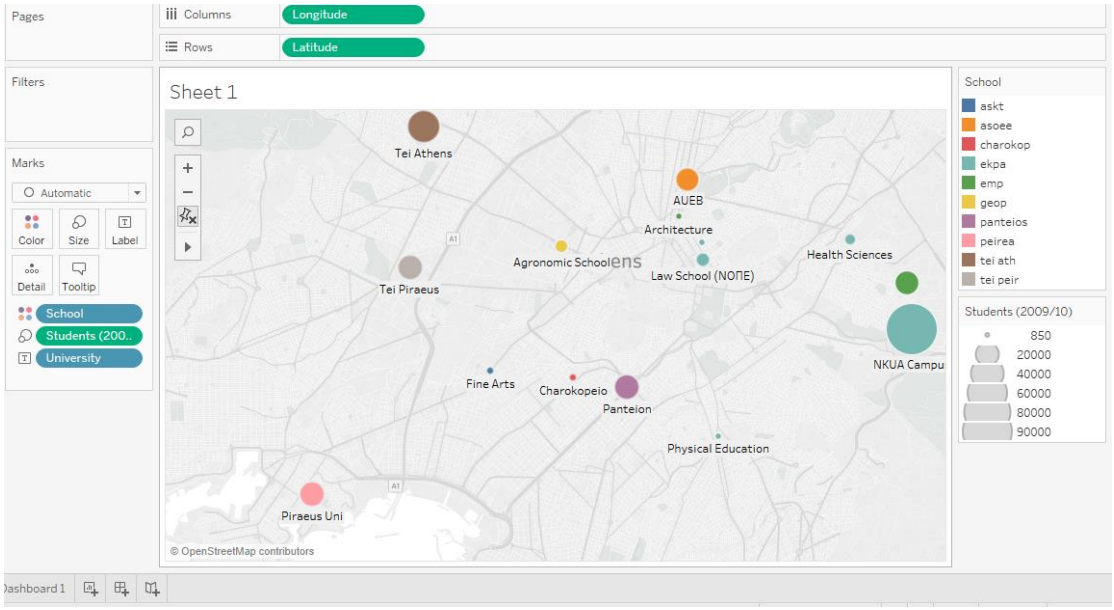
Στόχος αυτού του σεναρίου, είναι η δημιουργία ενός χάρτη, σε επίπεδο πόλεως, με λεπτομέρεια επιπέδων (layers), ώστε να είναι διακριτές οι οδοί και οι δρόμοι. Θα δημιουργήσουμε έναν χάρτη με τα πανεπιστημιακά κτήρια των Αθηνών. Οι πληροφορίες που θέλουμε να μοιραστούμε και να αναπαραστήσουμε οπτικά είναι η ακριβής θέση κάθε κτηρίου, με συντεταγμένες, τη σχολή που στεγάζει και το πανεπιστήμιο στο οποίο αυτή ανήκει, καθώς και το πλήθος των φοιτητών που φιλοξενεί. Τα δεδομένα αυτά συντάχθηκαν σε αρχείο excel, το οποίο αποτελεί την πηγή των δεδομένων.

University	Latitude	Longtitude	Students (2009/10)	School
Law School (ΝΟΠΕ)	37.980718	23.735583	5500	ekpa
AUEB	37.994067	23.732355	17427	asoee
pedagogy	37.983629	23.735336	1200	ekpa
NTUA	37.976746	23.778727	18530	emp
NKUA Campus	37.96911	23.779812	90000	ekpa
Piraeus Uni	37.941682	23.652684	19760	peirea
Tei Athens	38.002795	23.676359	35000	tei ath
Tei Piraeus	37.97936	23.673424	20000	tei peir
Charokopeio	37.961114	23.708061	1379	charokop
Panteion	37.959443	23.719391	20129	panteios

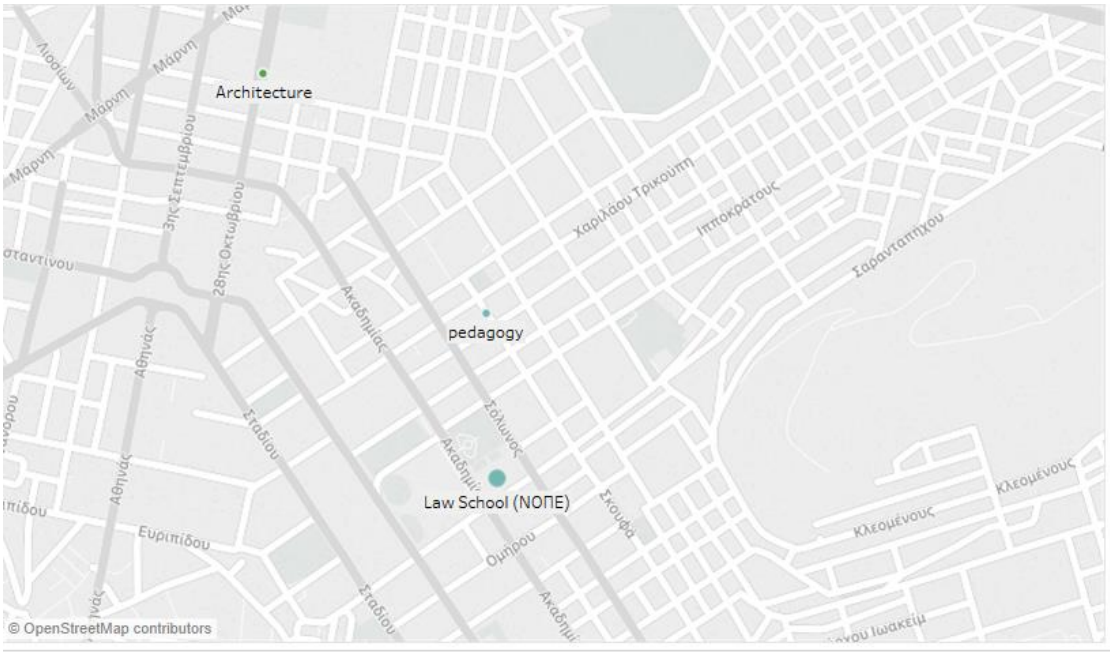
Health Sciences	37.983999	23.766647	3500	ekpa
Agronomic School	37.982906	23.705605	4731	geop
Physical Education	37.951282	23.738767	850	ekpa
Fine Arts	37.962164	23.690459	1445	askt
Architecture	37.987844	23.730365	1000	emp

Το πλήθος των φοιτητών θα αντιστοιχεί στο μέγεθος του σημείου στο χάρτη, το πανεπιστημιακό ίδρυμα στο χρώμα, και θα είναι εμφανής η ονομασία της σχολής.

3.4.1.1 Tableau

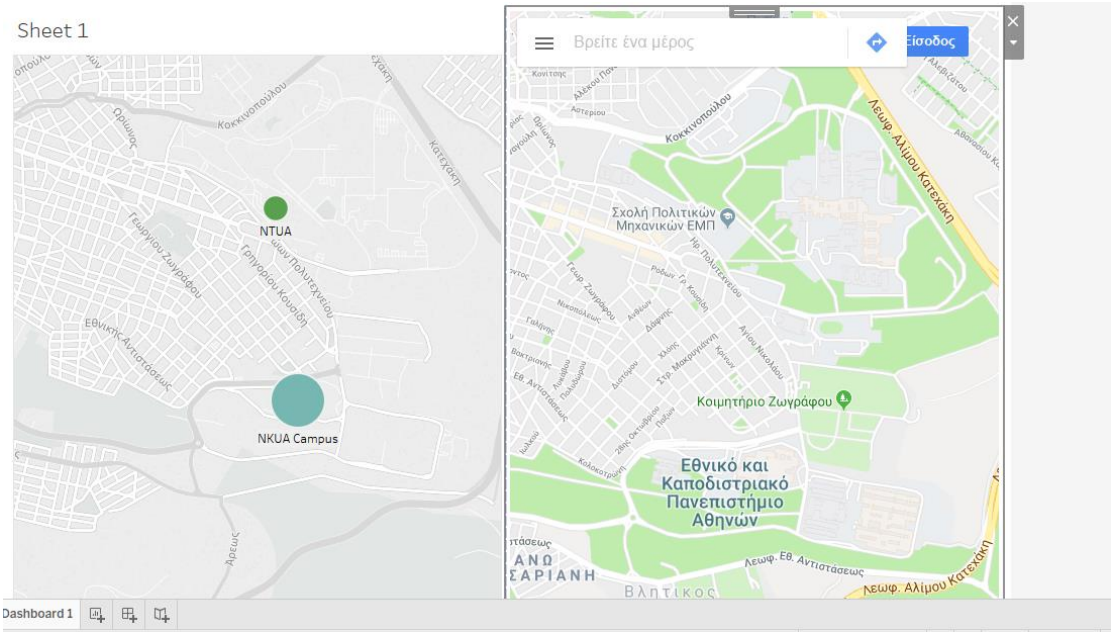


Εικόνα 3-75 – 1ος χάρτης Tableau, για τα πανεπιστημιακά κτήρια των Αθηνών



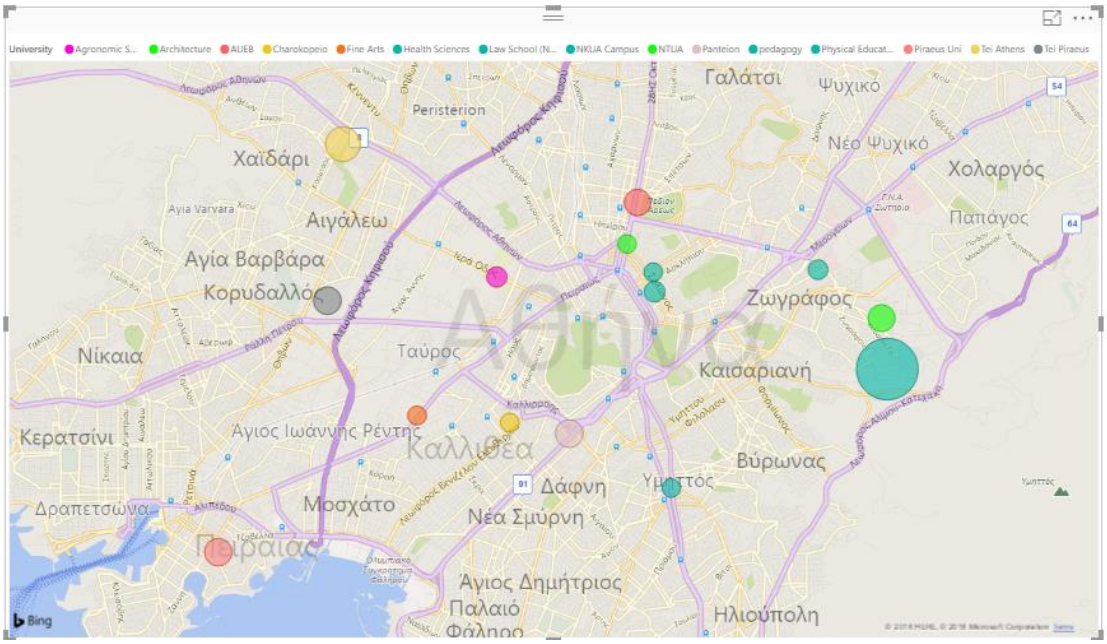
Εικόνα 3-76 - 2ος χάρτης Tableau, για τα πανεπιστημιακά κτήρια των Αθηνών

Επιλέγουμε να εμφανίζεται το όνομα της κάθε σχολής που στεγάζεται σε ξεχωριστές εγκαταστάσεις, κάθε ίδρυμα να αντιστοιχεί και σε διαφορετικό χρώμα, π.χ. ΕΜΠ με πράσινο, (Εικόνα 3-75), και το μέγεθος του κάθε σημείου να είναι ανάλογο των φοιτητών της σχολής. Παρατηρούμε πως η θέση των κτηρίων είναι ακριβής, σύμφωνα με τις συντεταγμένες που έχουμε δώσει (Εικόνα 3-76). Επιλέγουμε να εμφανίζεται το «Streets and Highways», στα map layers – στην καρτέλα Map, ώστε να εμφανίζονται οι οδοί στον χάρτη. Το tableau χρησιμοποιεί το OpenStreetMap, ωστόσο μπορούμε και να αλλάξουμε το χάρτη με άλλον συμβατό – π.χ. το Google Maps δεν είναι συμβατό για ενσωμάτωση. Μπορούμε όμως να το χρησιμοποιήσουμε βοηθητικά, δίπλα από το χάρτη μας στο dashboard, ως embedded url (Εικόνα 3-77).



Εικόνα 3-77 - 3ος χάρτης Tableau, για τα πανεπιστημιακά κτήρια των Αθηνών και εμφωλευμένο « Google Maps»

3.4.1.2 MS Power BI

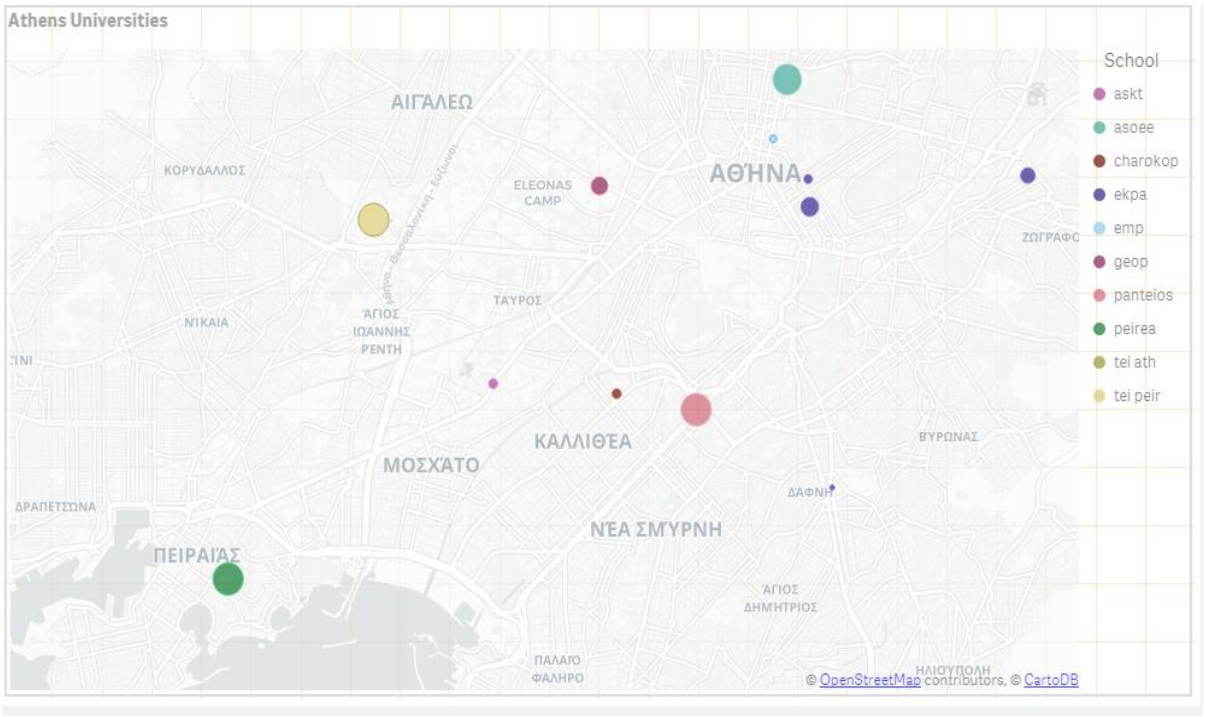


Εικόνα 3-78 - 1ος χάρτης Power BI, για τα πανεπιστημιακά κτήρια των Αθηνών

Το PowerBI χρησιμοποιεί τους χάρτες του Bing, οι οποίοι είναι επίσης ακριβείς, και ίσως πιο φιλικοί στην όψη. Δεν μπορέσαμε να εμφανίσουμε χρώμα ανά ίδρυμα στο χάρτη (Εικόνα 3-78), και επιλέξαμε χρώμα για κάθε κτήριο ξεχωριστά- μη αυτοματοποιημένα. Το όνομα κάθε σχολής εδώ αντιστοιχίζεται με χρώμα, σε μία λεζάντα ακριβώς πάνω από το χάρτη. Στο Tableau, εμφανίζεται πάνω στο χάρτη, δίπλα στο κάθε σημείο.

Αρχικά, τα δεδομένα που δώσαμε στα εργαλεία δεν ήταν πλήρη. Αφού δώσαμε την τελική μορφή στο αρχείο, στο Tableau η ανανέωση των δεδομένων γίνεται αυτόματα εάν έχουμε επιλέξει «live» σύνδεση, ή αρκεί η επαναφόρτωση του αρχείου για «extract». Αντίθετα, στο Power BI, απαιτείται να «φορτώσουμε» εκ νέου το αρχείο, και να αντιστοιχίσουμε τα δεδομένα, με τα πεδία του χάρτη.

3.4.1.3 Qlik Sense



Εικόνα 3-79 - 1ος χάρτης Qlik Sense, για τα πανεπιστημιακά κτήρια των Αθηνών

Αρχικά, η εισαγωγή των δεδομένων γίνεται από το data load editor. Στο «παράθυρο» του data manager, θέτουμε τον τύπο των latitude (γεωγραφικό πλάτος) και longitude (γεωγραφικό μήκος) ως geo data. Αυτόματα δημιουργείται νέο πεδίο - στήλη Longitude_Latitude, με το ζεύγος των εκάστοτε συντεταγμένων. Ενώ η στίξη των συντεταγμένων έχει γραφθεί με κόμμα, δεν υπάρχει πρόβλημα στην αναγνώριση και στη στήλη ζευγών η στίξη είναι ‘ . ‘ τελεία. Το qlik sense, για τη δημιουργία χαρτών δέχεται μόνον μία διάσταση. Επομένως, οι συντεταγμένες πρέπει να βρίσκονται σε ένα πεδίο.

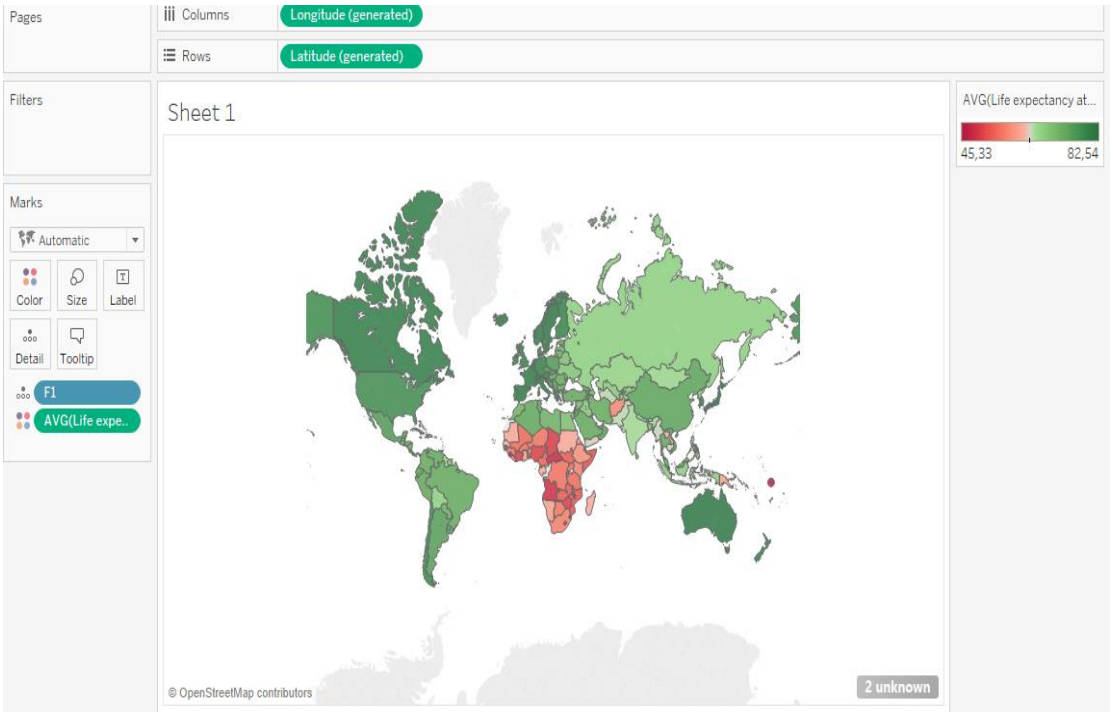
Σαν ποσοτική μεταβλητή (measure) εκχωρείται το πεδίο students, που καθορίζει το μέγεθος των points (Εικόνα 3-79). Το πεδίο school αντιστοιχίζεται στην επιλογή color by. Το «bubble size» ρυθμίζεται από έναν άξονα στο παράθυρο data. Τα χρώματα αντιστοιχίζονται αυτόματα – έχουμε επιλέξει τη βασική παλέτα 12 χρωμάτων. Πλευρικά του διαγράμματος, εμφανίζονται τα labels των πανεπιστημίων, με τα χρώματά τους. Ο χάρτης στο φόντο (background) είναι Open Street Maps, υποστηριζόμενος από το λογισμικό της carto.

3.4.2 Περιγραφικός Παγκόσμιος Χάρτης του μέσου προσδόκιμου ζωής

Μέσα από μία σειρά χαρτών θα αποτυπώσουμε το μ.ο. προσδόκιμο ζωής ανά τον πλανήτη – σε επίπεδο κρατών, και όποια άλλη ενδιαφέρουσα πληροφορία. Το αρχείο των δεδομένων αλιεύθηκε από τον ιστότοπο του παγκόσμιου οργανισμού υγείας World Health Organization, who.int, και συγκεκριμένα του παγκοσμίου παρατηρητηρίου υγείας [31]. Είναι αρχείο τύπου csv. Τα δεδομένα αφορούν 200 χώρες, για χρονική περίοδο 16 ετών, από το 2000 έως το 2015.

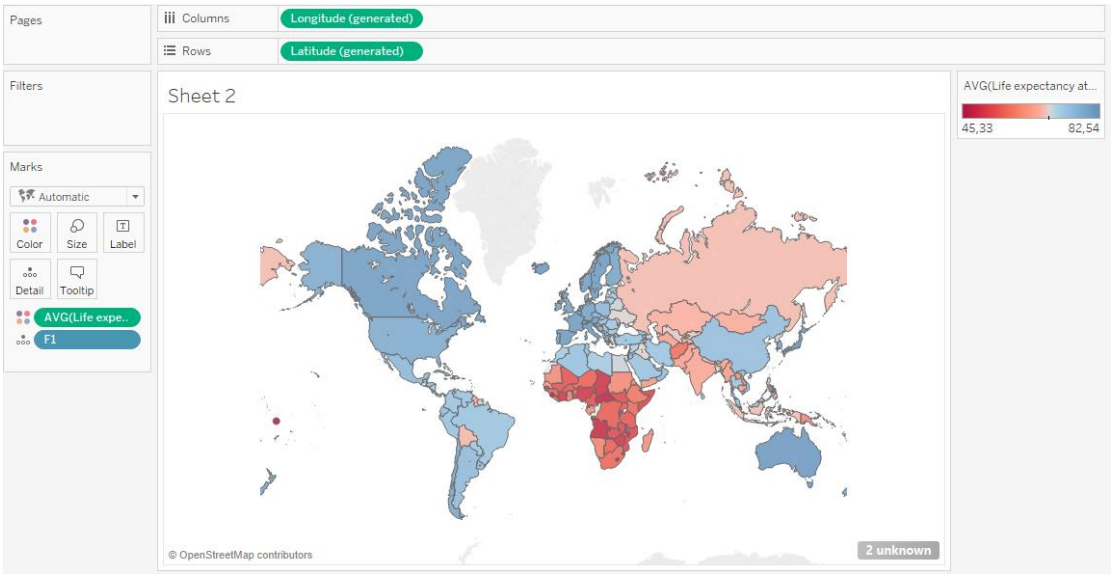
3.4.2.1 Tableau

Αρχικά, δημιουργούμε ένα χάρτη (Εικόνα 3-80), με χρωματική διαφοροποίηση ανάλογη του μέσου όρου του προσδόκιμου ζωής για τα έτη 2000-2015, ανά χώρα. Στο αρχείο δεδομένων υπάρχει στήλη με ονόματα των χωρών. Αυτόματα αναγνωρίζονται ως γεωγραφικά δεδομένα, χωρίς να χρειάζεται να δώσουμε συντεταγμένες, και σέρνοντας εμφανίζονται οι χώρες στο χάρτη. Παρατηρούμε πώς το λευκό χρώμα (ουδέτερο) αντιστοιχεί στη διάμεσο τιμή, 63,94 έτη.



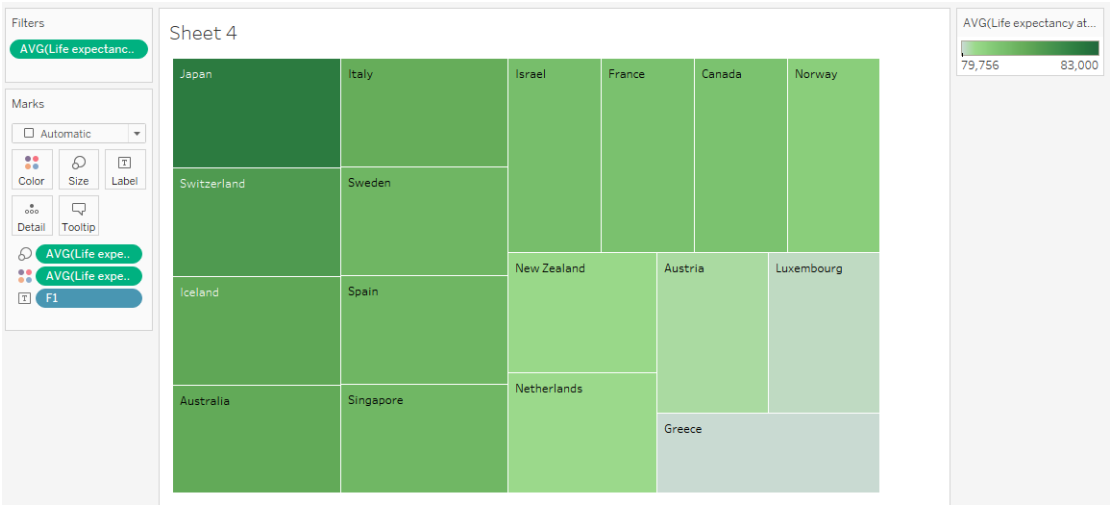
Εικόνα 3-80 – 1ος χάρτης Tableau για το μ.ο. του προσδόκιμου ζωής ανά χώρα

Έχει ενδιαφέρον να δούμε ένα χάρτη (Εικόνα 3-81) με χρωματικό ουδέτερο τον Μ.Ο. του προσδόκιμου. Στο «κουμπί» των χρωμάτων δίνεται η δυνατότητα αλλαγής του κέντρου, το οποίο ορίζουμε στα 68,79 έτη, όπως υπολόγισε το εργαλείο τον Μ.Ο.

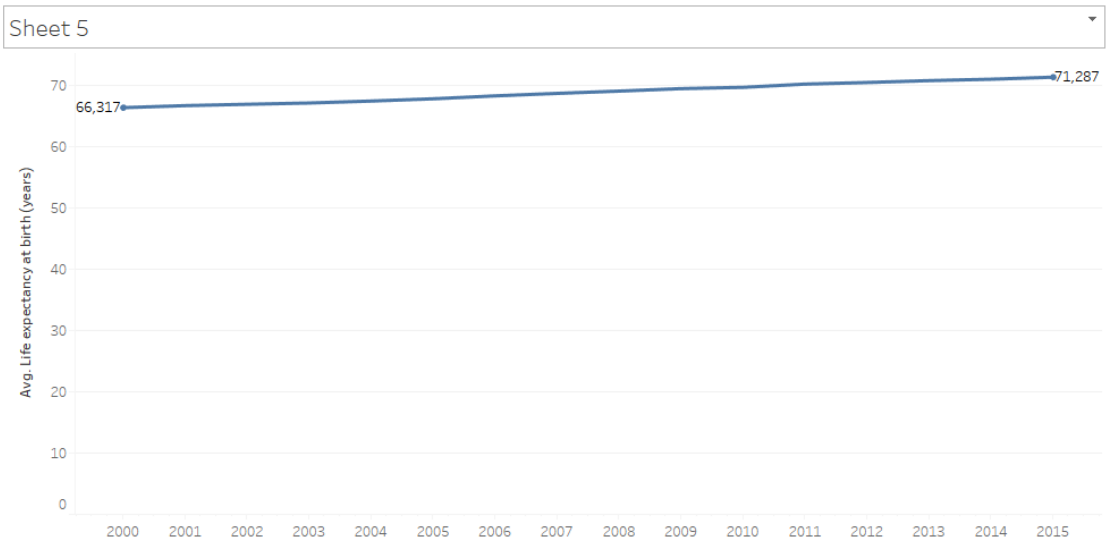


Εικόνα 3-81 - 2ος χάρτης Tableau για το μ.ο. του προσδόκιμου ζωής ανά χώρα

Έπειτα, σχεδιάζουμε έναν treemap (Εικόνα 3-82), όπου εμφανίζονται μόνον όσες χώρες έχουν μεγαλύτερο ή ίσο προσδόκιμο ζωής με την Ελλάδα. Η επιφανειακή αναπαράσταση, σε συνδυασμό με την ένταση του χρώματος, διευκολύνει να γίνει κατανοητή η διαφορά.

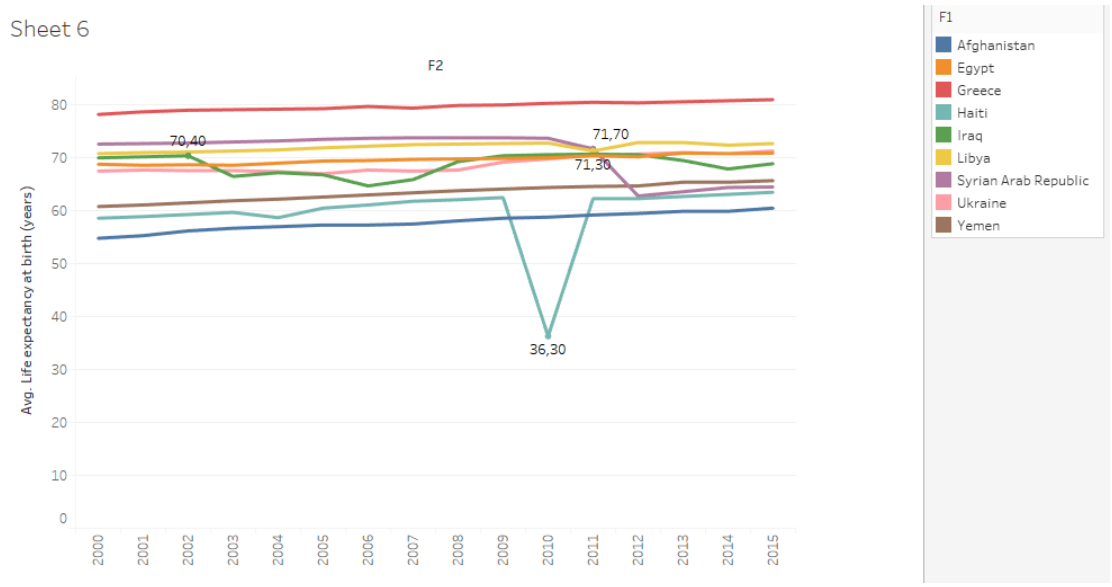


Εικόνα 3-82 – «treemap» του Tableau, με τις χώρες με μεγαλύτερο μ.ο. προσδόκιμου ζωής από την Ελλάδα

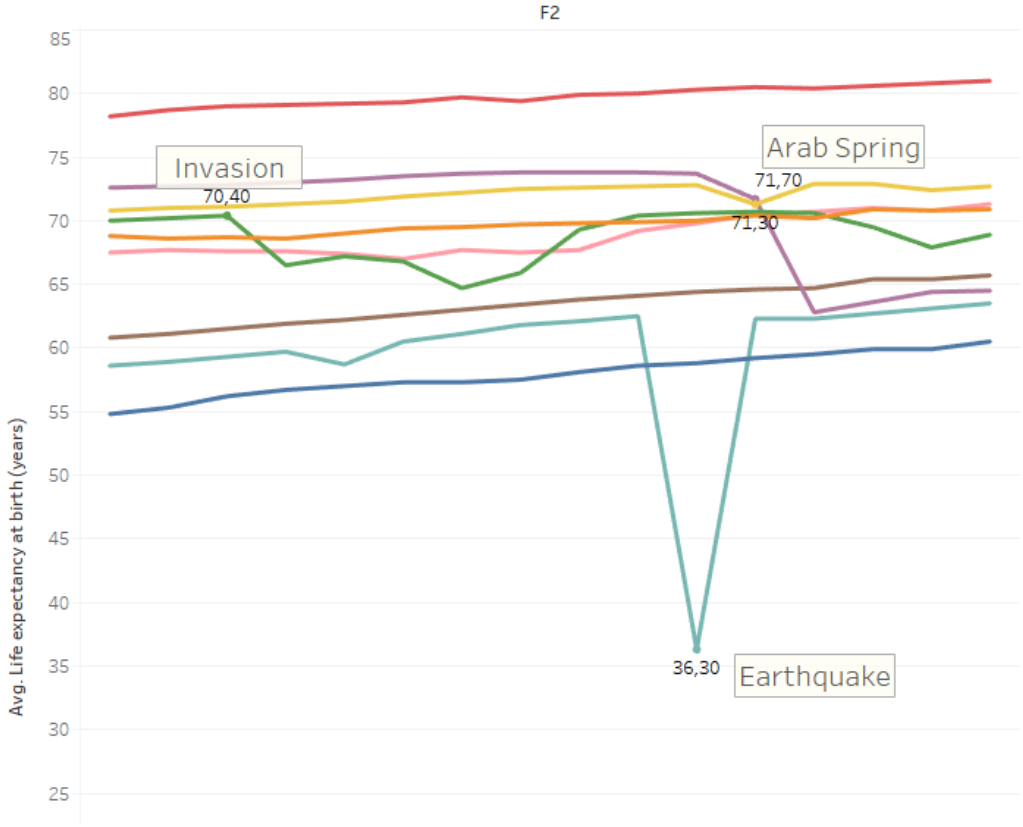


Εικόνα 3-83 – χρονοσειρά του Tableau, για τη μεταβολή του μέσου προσδόκιμου ζωής κατά το διάστημα 2000- 2015

Έπειτα, σχεδιάζουμε ένα line chart (Εικόνα 3-83), που απεικονίζει τη μεταβολή του μέσου προσδόκιμου παγκοσμίως, ανά έτος. Ακολουθεί ένα ακόμα line chart (Εικόνα 3-84), με χώρες που στο διάστημα 2000-2015, συνέβησαν γεγονότα που πιθανόν να απείλησαν τη ζωή των πολιτών. Ιράκ, Αφγανιστάν, Αιτή, Ουκρανία, Λιβύη, Υεμένη, Συρία, Αίγυπτος είναι οι χώρες που συμπεριλαμβάνουμε. Συμπεριλαμβάνουμε και την Ελλάδα ως σημείο αναφοράς, αλλά και ως χώρα πιθανής απότομης μεταβολής του προσδόκιμου εξαιτίας της οικονομικής ύφεσης.



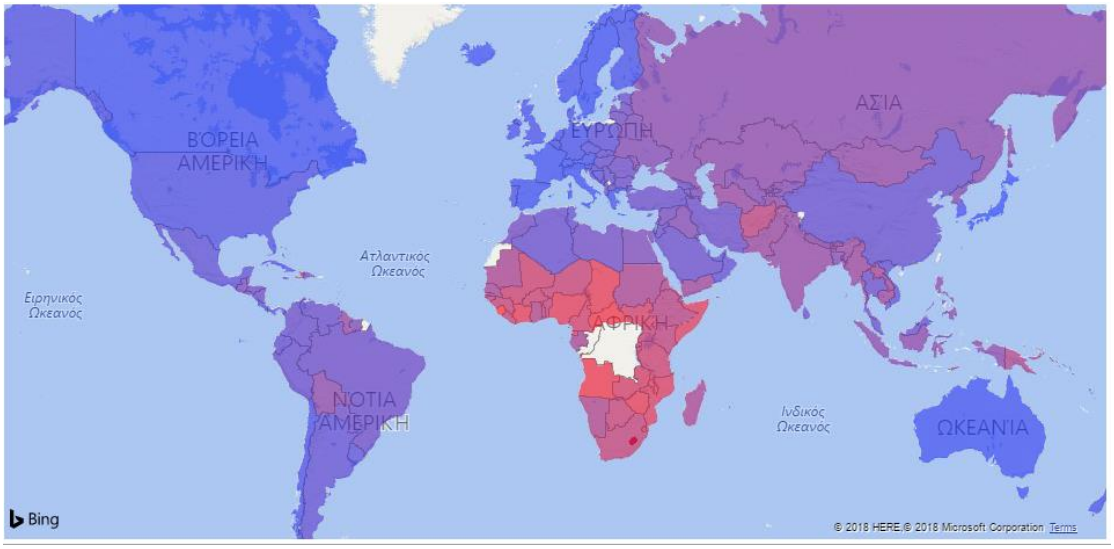
Εικόνα 3-84 – 1η χρονοσειρά του Tableau, για τη μεταβολή του μ.ο. προσδόκιμου ζωής για χώρες σε έκτακτη κατάσταση για το χρονικό διάστημα 2000- 2015



Εικόνα 3-85 -2η χρονοσειρά του Tableau, για τη μεταβολή του μ.ο. προσδόκιμου ζωής για χώρες σε έκτακτη κατάσταση για το χρονικό διάστημα 2000- 2015

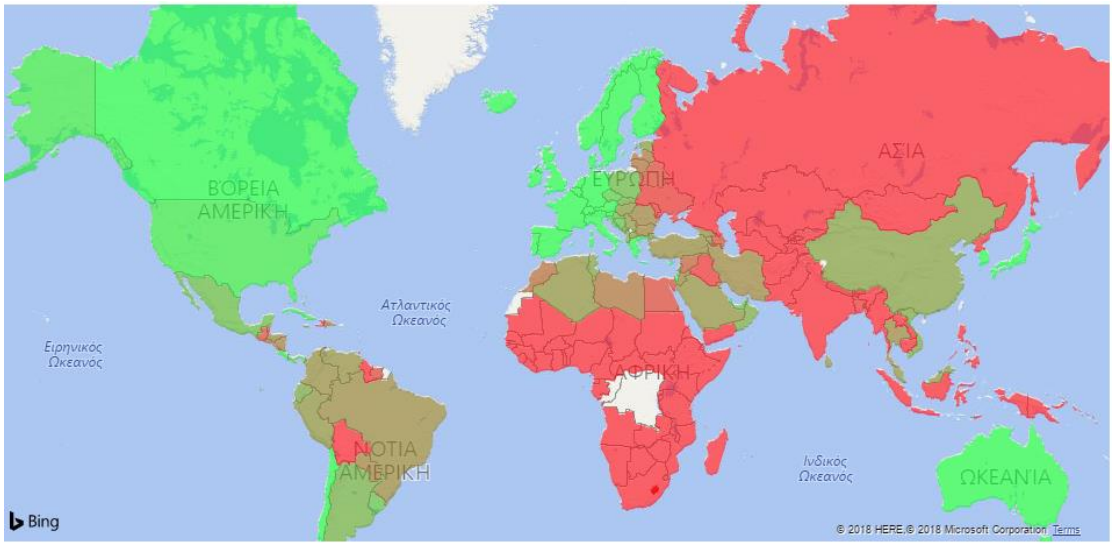
Ο πιο οφθαλμοφανής μεταβολή είναι στην Αιτή για το 2010, από τα 62,55 στα 36,3. Λόγος ο τεράστιος αριθμός θυμάτων που προξένησε ο σεισμός. Μπορούμε εδώ (Εικόνα 3-85) να εμφανίζουμε συνεχώς τα σημεία ενδιαφέροντος (Marks : always show), αλλά και να προσθέτουμε περιγραφή κειμένου για τα γεγονότα που προξένησαν τη μεταβολή πάνω στο γράφημα – το μέγεθος της γραμματοσειράς όμως δεν ήταν δυνατό να προσαρμοστεί επιθυμητά στην πράξη. Επίσης, εμφανείς πτώσεις έχουμε στο Ιράκ το 2002 και έπειτα (πόλεμος με Η.Π.Α), στη Λιβύη μόνον το 2011 (Αραβική Άνοιξη), και στη Συρία από το 2011 και έπειτα (εμφύλιος πόλεμος).

3.4.2.2 MS Power BI



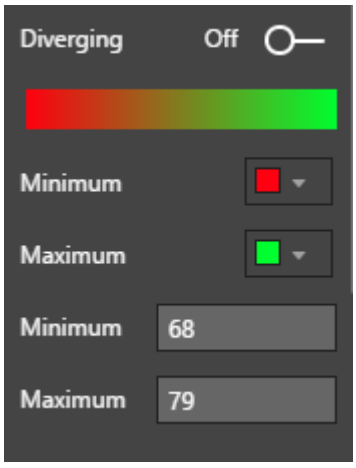
Εικόνα 3-86 -1ος χάρτης τουPower BI για το μ.ο. του προσδόκιμου ζωής ανά χώρα

Αντίστοιχα, ο χάρτης (Εικόνα 3-86) του μέσου προσδόκιμου ζωής για τα 2000-2015 ανά χώρα. Τα χρώματα του εργαλείου δε διευκολύνουν ιδιαίτερα τον παρατηρητή.



Εικόνα 3-87 - 2ος χάρτης τουPower BI για το μ.ο. του προσδόκιμου ζωής ανά χώρα

Δεν μας δίνεται η δυνατότητα αλλαγής του μέσου στη διαβάθμιση των χρωμάτων. Μπορούμε όμως να ορίσουμε μια τιμή που από αυτή και για μεγαλύτερες θα εμφανίζεται το max χρώμα και το ίδιο για το min.

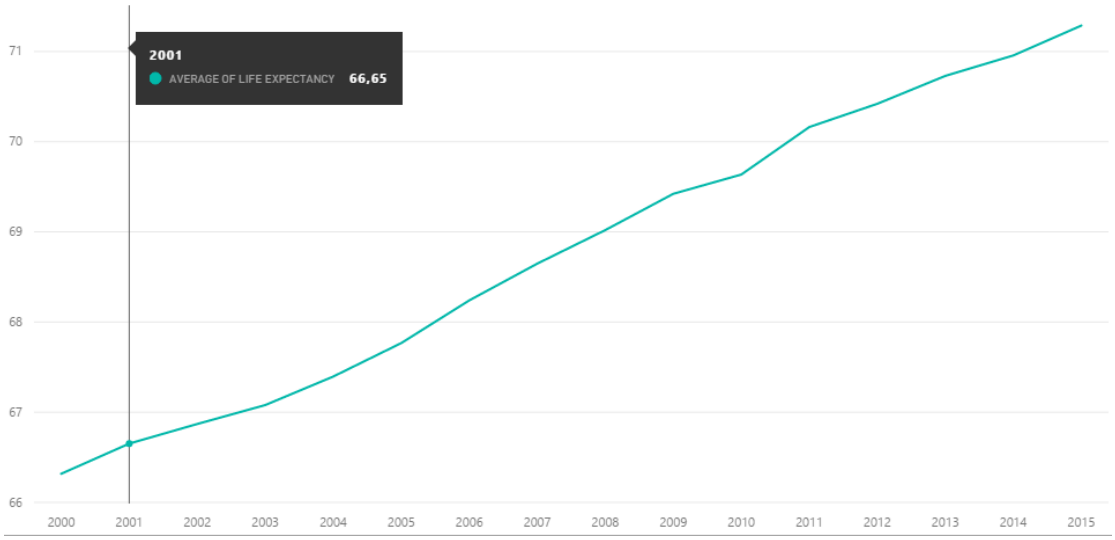


Θέτουμε τα 68 έτη ως ελάχιστο, για να δούμε άμεσα ποιες χώρες είναι κάτω του παγκοσμίου μ.ο., και τα 79 ως max για να έχουμε την Ελλάδα ως σημείο αναφοράς.

Το treemap (*Εικόνα 3-88*)

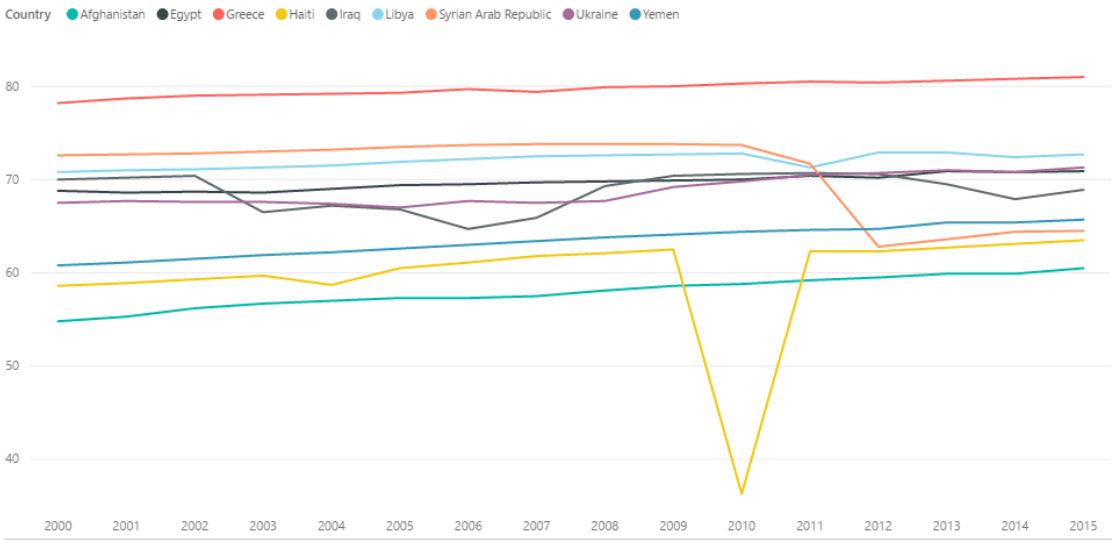
Εικόνα 3-88 - «treemap» του Power BI, με τις χώρες με μεγαλύτερο μ.ο. προσδόκιμου ζωής από την Ελλάδα

Το linechart (*Εικόνα 3-89*) για τα έτη 2000 – 2015



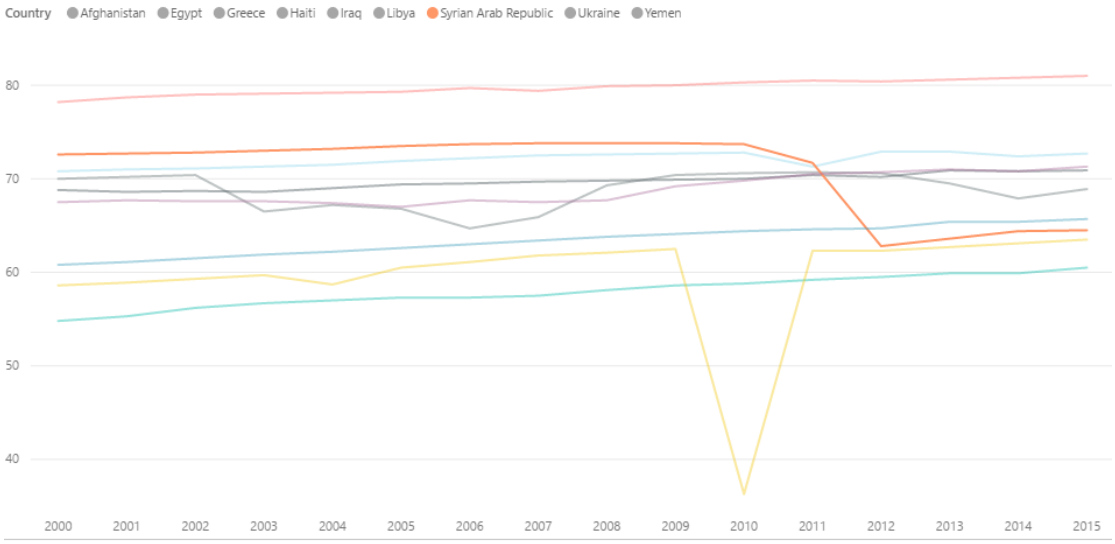
Εικόνα 3-89 - χρονοσειρά του Power BI, για τη μεταβολή του μέσου προσδόκιμου ζωής κατά το διάστημα 2000-2015

Το εργαλείο εδώ αυτόματα θέτει ως αρχή στον άξονα των τιμών το ελάχιστο προσδόκιμο. Αυτή η αναπαράσταση είναι πιο διαφωτιστική όσον αναφορά τις μεταβολές στο συγκεκριμένο χρονικό διάστημα , ενώ η αρχή του άξονα από το μηδέν δείχνει πόση είναι η αύξηση σε σχέση με ολόκληρη τη διάρκεια της ζωής. Και στα δύο εργαλεία βέβαια μπορεί να ρυθμιστεί η αρχή του κάθε άξονα. Το line chart για συγκεκριμένες χώρες (Εικόνα 3-90 και Εικόνα 3-91):



Εικόνα 3-90 - 1η χρονοσειρά του Power BI, για τη μεταβολή του μ.ο. προσδόκιμου ζωής για χώρες σε έκτακτη κατάσταση για το χρονικό διάστημα 2000- 2015

Η λεζάντα αντιστοίχισης χρωμάτων- χωρών εδώ είναι πραγματικά πολύ χρήσιμη. Δεν μπορούμε να εμφανίζουμε μόνιμα τα σημεία που μας ενδιαφέρουν στο διάγραμμα.

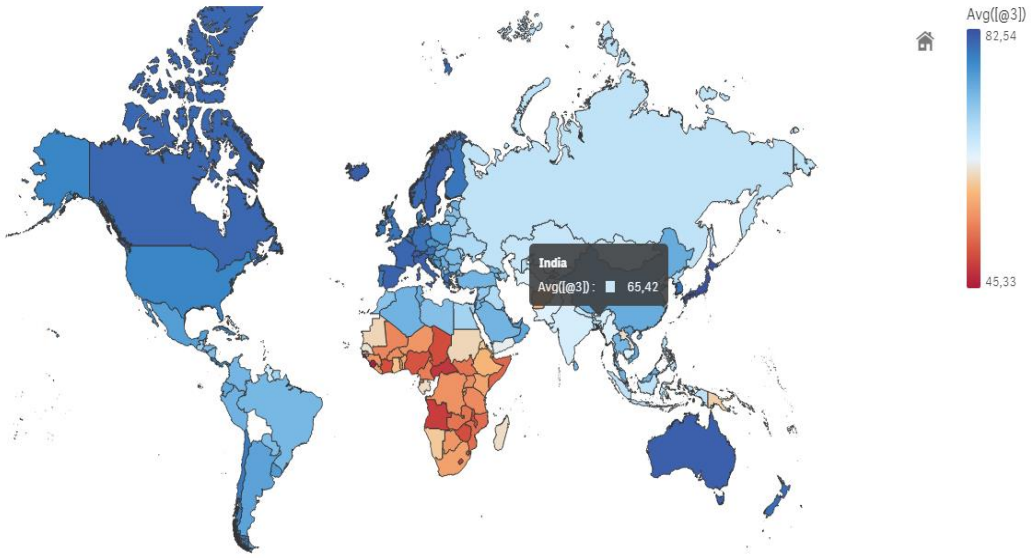


Εικόνα 3-91 -21η χρονοσειρά του Power BI, για τη μεταβολή του μ.ο. προσδόκιμου ζωής για χώρες σε έκτακτη κατάσταση για το χρονικό διάστημα 2000- 2015

3.4.2.3 Qlik Sense

Και σε αυτό το εργαλείο, μετά την εισαγωγή των δεδομένων, το πεδίο των χωρών αναγνωρίζεται ως «geo data» και η αντιστοίχιση στα σχηματικά πλαίσια των χωρών γίνεται αυτόματα.

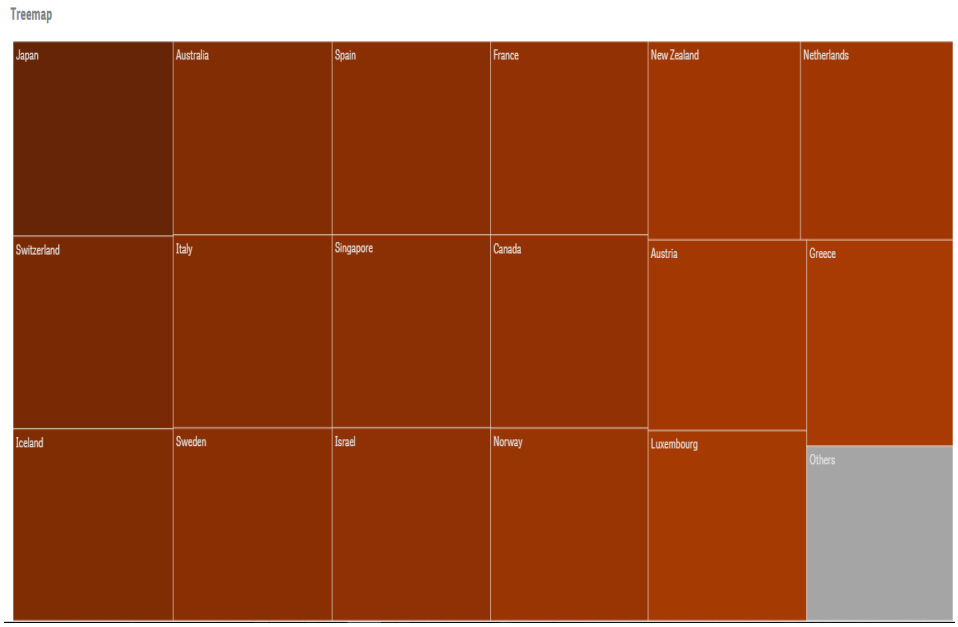
Στο χάρτη αναπαράστασης του μέσου προσδόκιμου (Εικόνα 3-92), μπορούμε να επιλέξουμε μόνο το μοτίβο του χρωματισμού (ένα ή δύο χρώματα, με επίπεδα ή συνεχή) και όχι τα χρώματα. Το ουδέτερο λευκό αντιστοιχεί και εδώ στη διάμεσο.



Εικόνα 3-92 - χάρτης του Qlik Sense για το μ.ο. του προσδόκιμου ζωής ανά χώρα

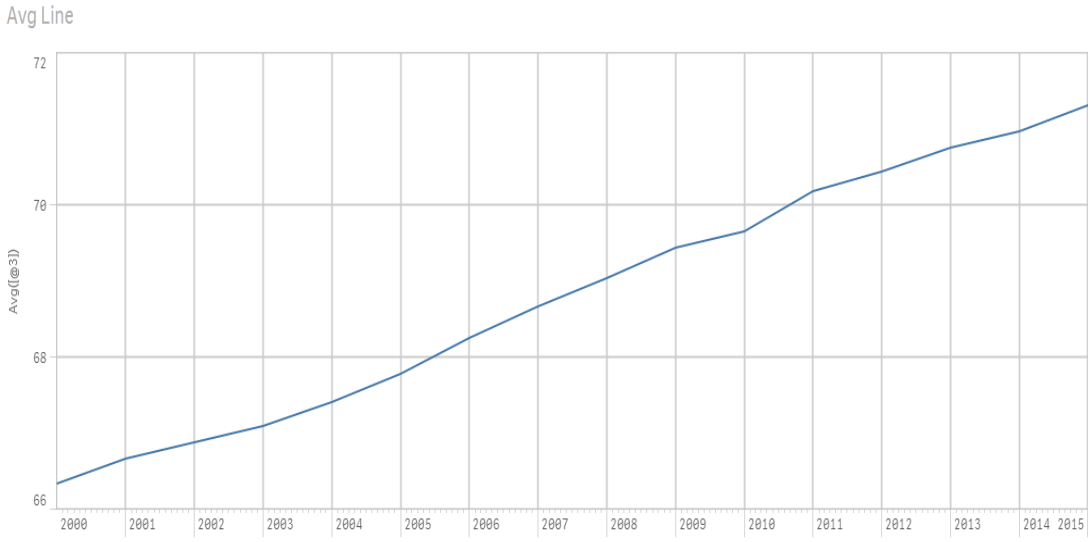
Δεν είναι δυνατόν, να αλλάξουμε το μέσον στη διαβάθμιση των χρωμάτων, ώστε να ορίσουμε σε αυτό τον μ.ο, ούτε κάποια άλλη διαφοροποίηση χωρίς τη χρήση εντολών και συναρτήσεων.

Ακολουθεί το treemap (*Εικόνα 3-93*):



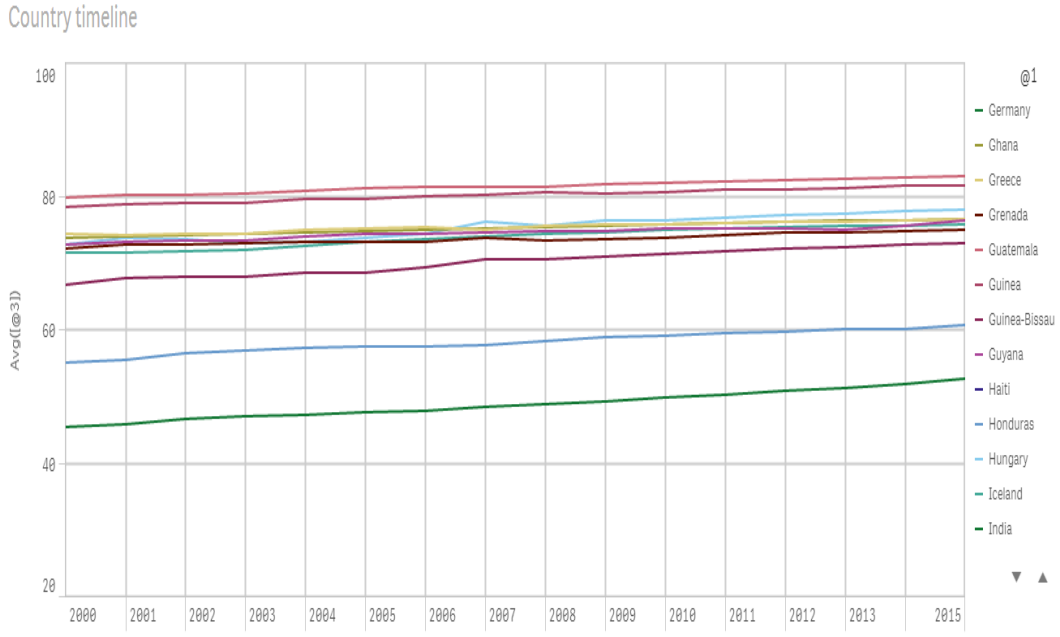
Εικόνα 3-93 - «treemap» του Qlik Sense, με τις χώρες με μεγαλύτερο μ.ο. προσδόκιμου ζωής από την Ελλάδα

Εδώ συμπεριλαμβάνεται αυτόματα και ο μ.ο. των υπολοίπων κρατών, «Others».



Εικόνα 3-94 - χρονοσειρά του Qlik Sense, για τη μεταβολή του μέσου προσδόκιμου ζωής κατά το διάστημα 2000-2015

Και εδώ (Εικόνα 3-94) η αρχή των τιμών στον y άξονα τίθεται στον ελάχιστο προσδόκιμο μ.ο. Όσον αφορά το διάγραμμα επιλεγμένων χωρών, δε μπορεί να γίνει χωρίς τη χρήση εντολών, αφού δεν υπάρχει δυνατότητα φίλτρου για τις διαστάσεις.



Εικόνα 3-95 - χρονοσειρά τουQlik Sense, για τη μεταβολή του μ.ο. προσδόκιμου ζωής για χώρες σε έκτακτη κατάσταση για το χρονικό διάστημα 2000- 2015

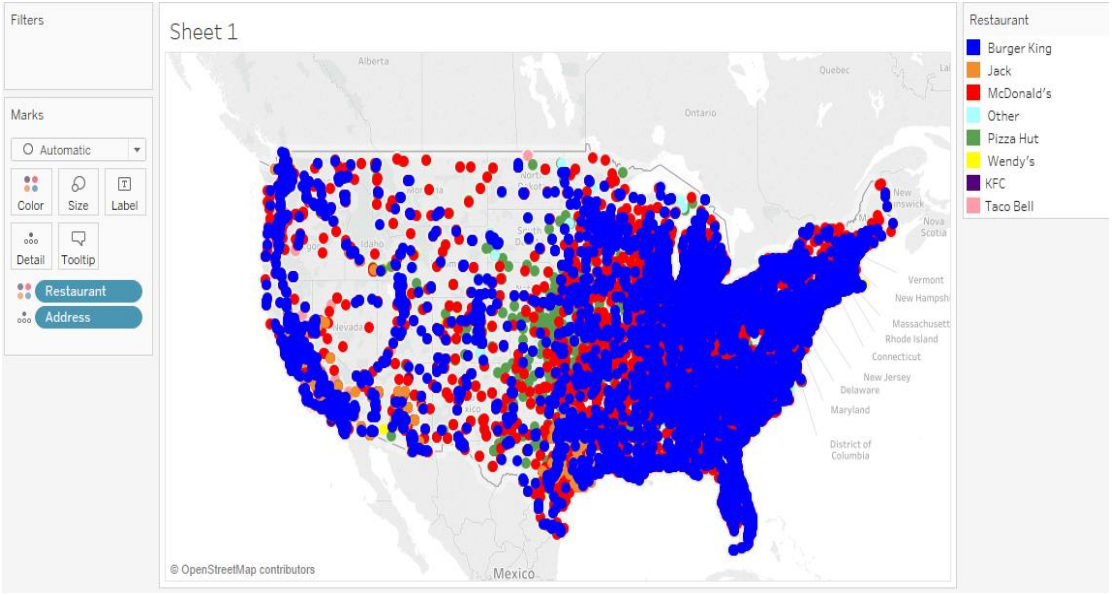
Στη χρονοσειρά (Εικόνα 3-95), εμφανίζονται μόνο κάποιες χώρες (οι πρώτες στην αλφαβητική κατάταξη), όπως είναι λογικό, μιας και είναι αδύνατο να συμπεριληφθούν από το εργαλείο όλες.

3.4.3 Χάρτης Εστιατορίων “Fast Food” στις Η.Π.Α

Με τη δημιουργία του χάρτη αυτού θέλουμε να αξιολογήσουμε τη δυνατότητα εστίασης και μεγέθυνσης περιοχών, και τη σωστή διαχείριση και αποτύπωση μεγάλου όγκου γεωγραφικών δεδομένων. Τα σημεία προς δημιουργία είναι περισσότερα από 50,000 , και εισάγονται από αρχείο csv [32]. Αναπαριστούν την ακριβή θέση όλων των εστιατορίων των αλυσίδων fast food στις Η.Π.Α., μέσω συντεταγμένων.

3.4.3.1 Tableau

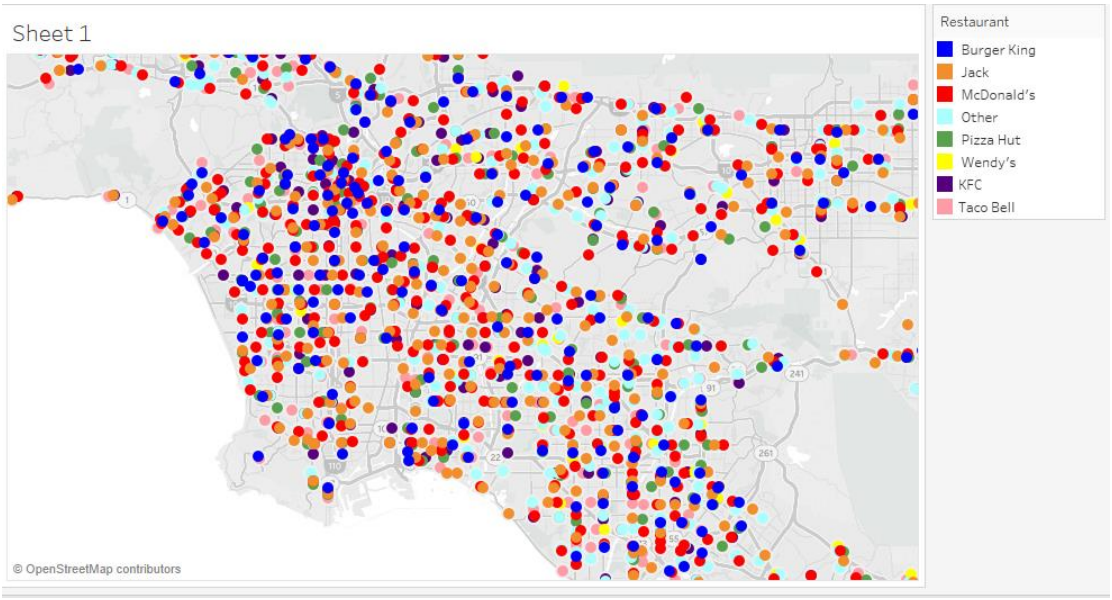
Με την εισαγωγή του αρχείου csv, γίνεται αυτόματα η αναγνώριση των γεωγραφικών δεδομένων. Κατά την επεξεργασία των δεδομένων, ο χρήστης πρέπει να δημιουργήσει μία νέα στήλη με το πλήρες όνομα του εστιατορίου, μέσω των calculated fields, με εντολή Case.



Εικόνα 3-96 – χάρτης του Tableau με τα εστιατόρια «fast food» στις Η.Π.Α.

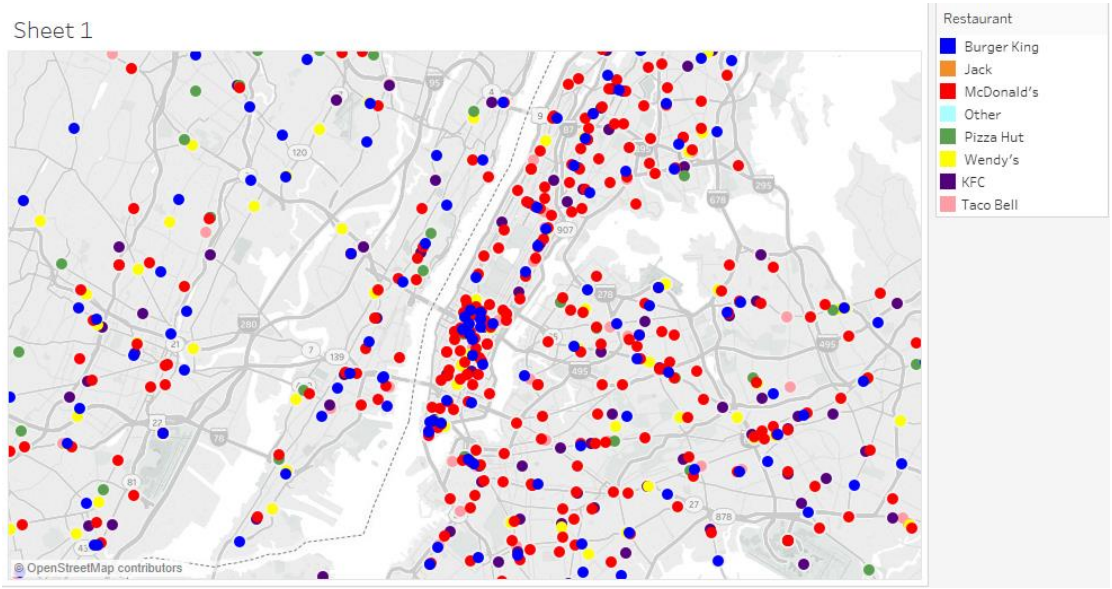
Το εργαλείο αποτυπώνει όλα τα σημεία στο χάρτη (Εικόνα 3-96), αντιστοιχίζει με το χρώμα της επιλογής μας την αλυσίδα του εστιατορίου και εμφανίζει τη διεύθηνση του καθενός σαν λεπτομέρεια, με roll over.

Σε κλίμακα πολιτειών , στη Δυτική Ακτή (Εικόνα 3-97):



Εικόνα 3-97 – χάρτης του Tableau, με τα εστιατόρια «fast food» στη Δυτική Ακτή των Η.Π.Α.

Σε κλίμακα πόλης, στη Νέα Υόρκη (Εικόνα 3-98):



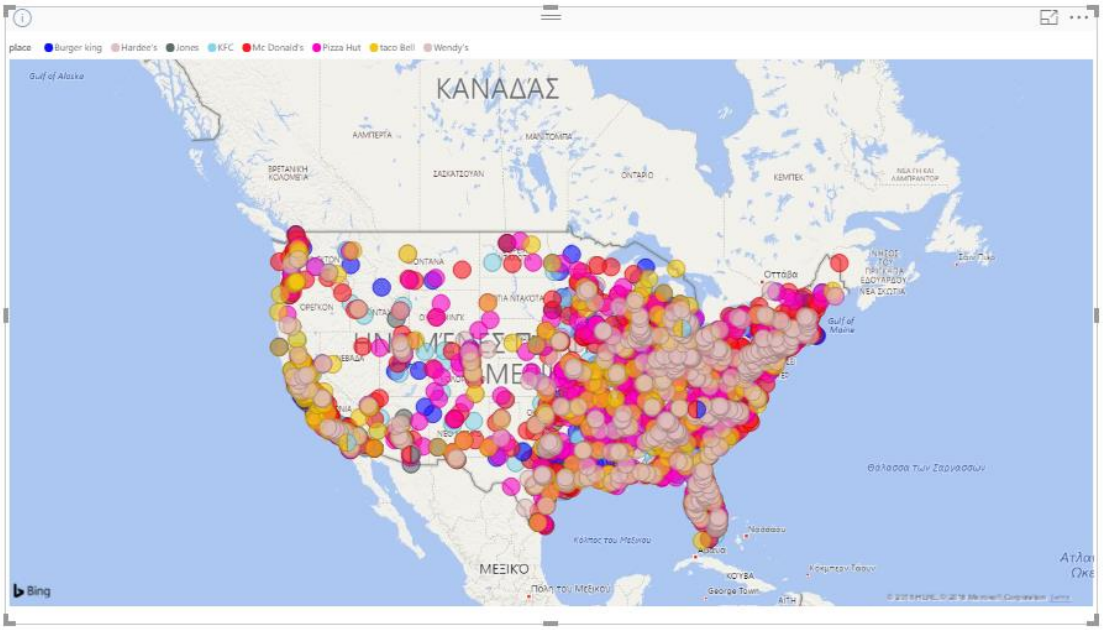
Εικόνα 3-98 - χάρτης του Tableau, με τα εστιατόρια «fast food» στην πόλη της Νέας Υόρκης

Εικόνα 3-99 - χάρτης του Tableau, για τα εστιατόρια «fast food», με εστίαση zoom σε οδούς πόλης

Το επίπεδο εστίασης στο χάρτη (zoom level), είναι λεπτομερές, με εμφάνιση των δρόμων (Εικόνα 3-99).

3.4.3.2 MS Power BI

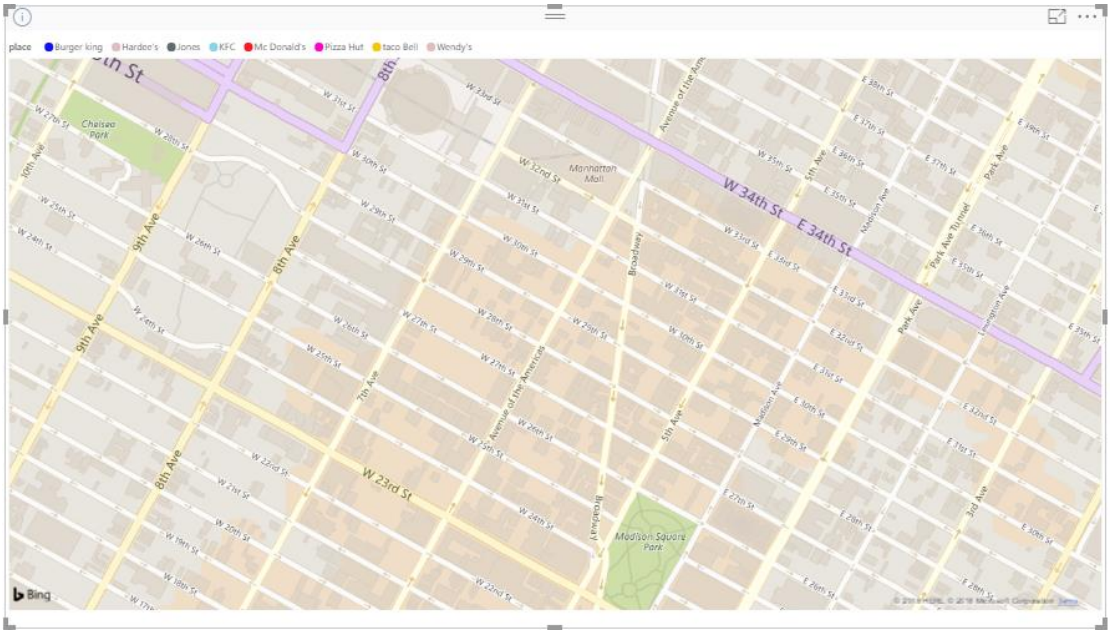
Δημιουργούμε αρχικά τη στήλη με τα πλήρη ονόματα στο query editor, που κάνει τη διαδικασία ευκολότερη από τα άλλα εργαλεία. Όταν επιλέγουμε τις παραμέτρους του χάρτη, καλούμαστε να επιλέξουμε ανάμεσα από κάποια map styles. Μας βολεύει το «road» (Εικόνα 3-100), με λεπτομέρειες για το οδικό δίκτυο, διευθύνσεις κ.α.



Εικόνα 3-100 - χάρτης του Power BI με τα εστιατόρια «fast food» στις Η.Π.Α.

Τα διαθέσιμα χρώματα, ίσως, καθιστούν τα σημεία λίγο δυσδιάκριτα. Πολλά γεωγραφικά ονόματα είναι, εξ' αρχής, σε ελληνική απόδοση.

Σε κλίμακα πόλης (Εικόνα 3-101):

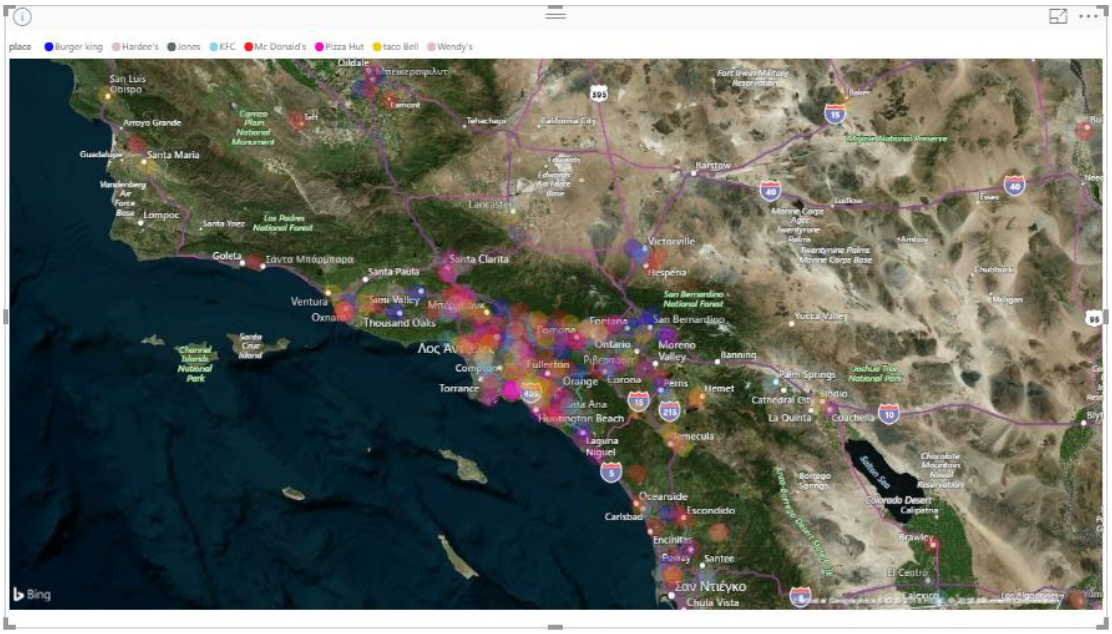


Εικόνα 3-101 - χάρτης του Power BI, με τα εστιατόρια «fast food» στην πόλη της Νέας Υόρκης

Παρά το καλό επίπεδο εστίασης στο χάρτη, και την ακρίβεια και ποιότητα αυτού, παρατηρούμε πως πολλά σημεία λείπουν. Αυτό συμβαίνει γιατί το εργαλείο δεν μπορεί να αποτυπώσει απεριόριστο αριθμό σε γράφημα, αλλά περίπου 3500 σημεία. Η αναπαράσταση αυτών ωστόσο είναι όσο το δυνατόν πιο αντιπροσωπευτική του συνόλου δεδομένων, αφού καλύπτουν ολόκληρη τη χώρα, και δεν γίνονται απλά, με τη σειρά κατάταξης.

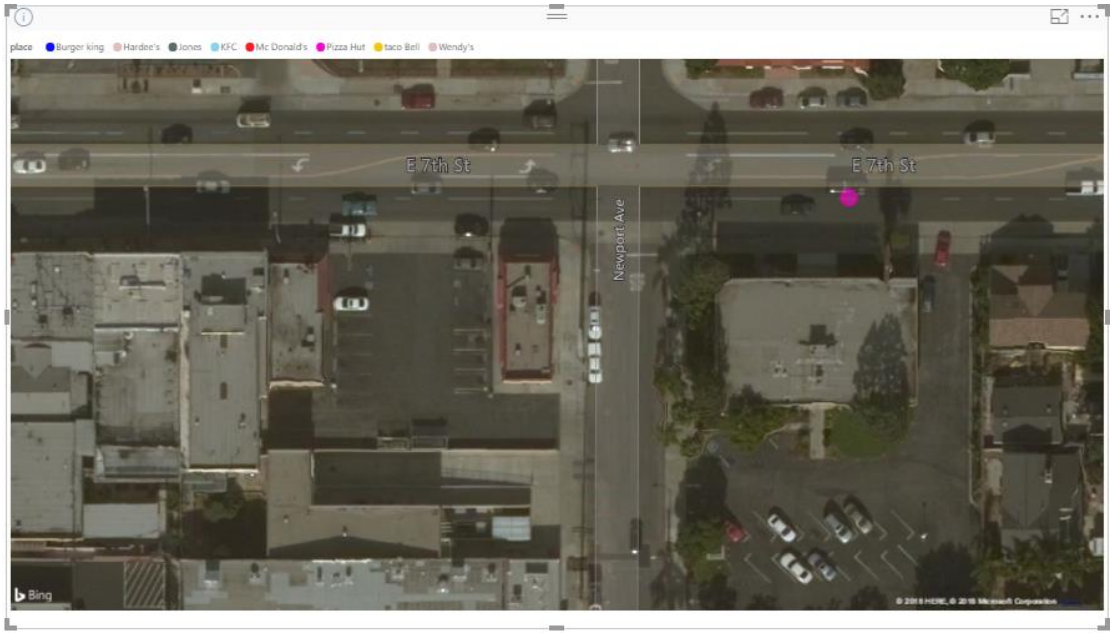
Ένα πολύ ενδιαφέρον θέμα του «map style» είναι το aerial, όπου γίνεται δορυφορική αναπαράσταση των εκάστοτε περιοχών, με τη μέγιστη δυνατή λεπτομέρεια. Το χαρακτηριστικό αυτό του Power BI είναι πάρα πολύ χρήσιμο και απουσιάζει από τα άλλα εργαλεία (προς σύγκριση).

Λος Άντζελες δορυφορική απεικόνιση (Εικόνα 3-102):



Εικόνα 3-102 – χάρτης του Power BI, με δορυφορική απεικόνιση για την περιοχή του Λος Άντζελες

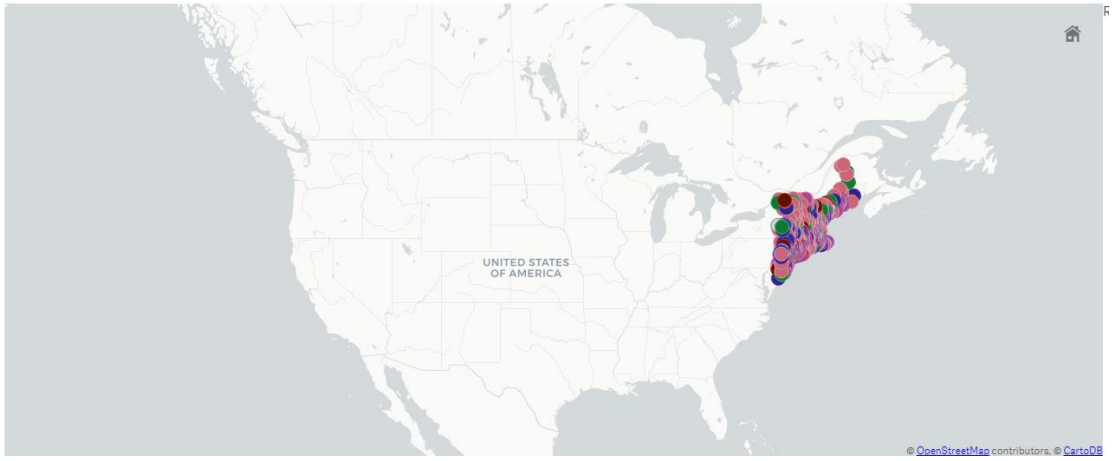
Pizza Hut σε γειτονία του Λος Άντζελες (Εικόνα 3-103):



Εικόνα 3-103 - χάρτης του Power BI, με εστίαση – zoom σε δορυφορική απεικόνιση για εστιατόριο σε δρόμο του Λος Άντζελες

3.4.3.3 Qlik Sense

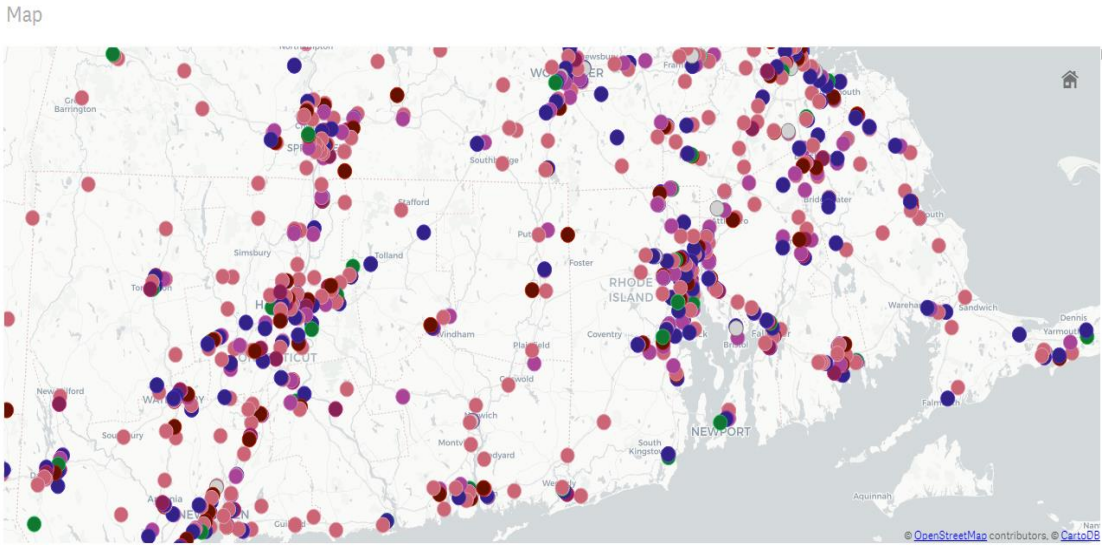
Και στο εργαλείο αυτό δημιουργούμε νέα στήλη με τα πλήρη ονόματα των εστιατορίων, ως calculated field στο παράθυρο data manager.



Εικόνα 3-104 - χάρτης του Qlik Sense με τα εστιατόρια «fast food» στις Η.Π.Α.

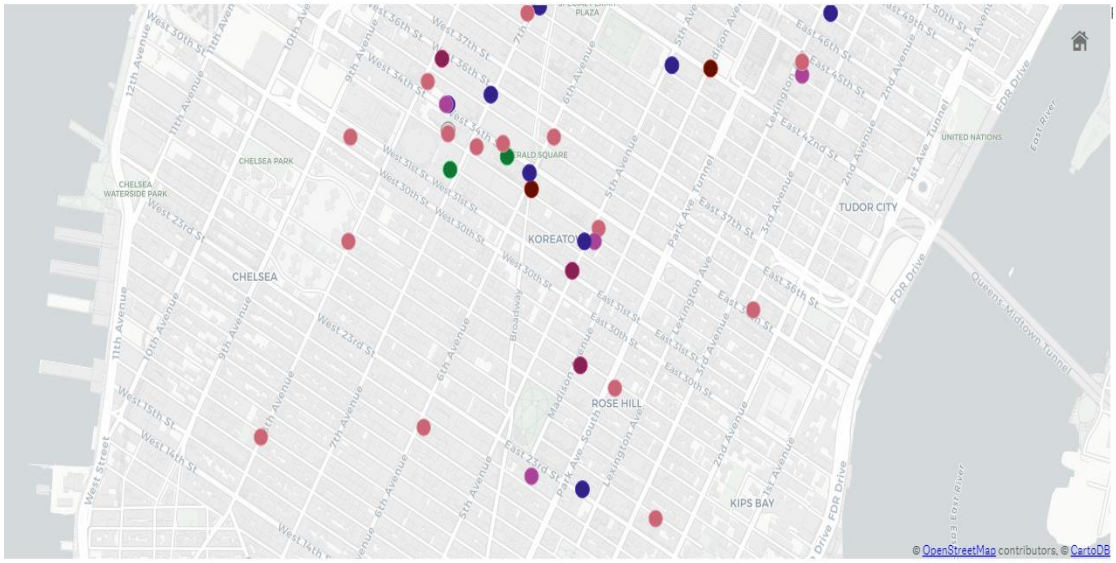
Στο χάρτη (Εικόνα 3-104) εμφανίζονται μόνο ορισμένα σημεία. Αυτό συμβαίνει γιατί το Qlik Sense μπορεί να αναπαραστήσει μέχρι και 3,333 σημεία σε χάρτη, αριθμός πολύ μικρότερος από τον ζητούμενο. Επίσης, δεν μπορούμε να επιλέξουμε εμείς τα χρώματα και την αντιστοιχία αυτών.

Σε κλίμακα πολιτειών (Εικόνα 3-105):



Εικόνα 3-105 - χάρτης του Qlik Sense, με τα εστιατόρια «fast food» στην ευρύτερη πολιτεία της Νέας Υόρκης

Σε κλίμακα πόλης (Εικόνα 3-106):



Εικόνα 3-106 - χάρτης του Qlik Sense, με τα εστιατόρια «fast food» στην πόλη της Νέας Υόρκης

3.5 Αποτελέσματα Συγκρίσεων- Προτερήματα και Μειονεκτήματα

3.5.1 Tableau

Το Tableau είναι ένα εργαλείο πολλών δυνατοτήτων και λύσεων, ακόμα και «ανορθόδοξων», ωστόσο πιο πολύπλοκο στη χρήση. Η δημιουργία «απαιτητικών» διαγραμμάτων χωρίς περιορισμούς σε διαστάσεις και τρόπους εμφάνισης, η απολύτως προσαρμόσιμη στις ανάγκες διαδικασία φιλτραρίσματος με ποικίλες επιλογές για όλα τα μεγέθη και η δυνατότητα ρύθμισης της εμφάνισης (format) κάθε άξονα, διαγράμματος και φόντου, είναι τα στοιχεία που ξεχώρισαν στο Tableau.

Τα υπόλοιπα προτερήματα του εργαλείου που παρατηρήθηκαν :

- Ο «data interpreter», που μορφοποιεί κατάλληλα τα εισαχθέντα δεδομένα σε ορισμένες περιπτώσεις)
- Η δυνατότητα «live connection» του αρχείου προέλευσης με τα δεδομένα στο εργαλείο (όχι όμως με τα διαγράμματα στη desktop έκδοση)
- Η ευκολία στη δημιουργία joins και unions
- Τα εύχρηστα παράθυρα εργασίας και η γρήγορη μετάβαση από αυτό του ETL σε αυτά των διαγραμμάτων
- Το παράθυρο (toolbar) προτάσεων διαγραμμάτων και των προαπαιτούμενών τους
- Καλή οπτικοποίηση σχηματικά και χρωματικά, «οικεία» στο μάτι
- Εξαγωγή των επεξεργασθέντων δεδομένων (ETL) σε csv
- Δυνατότητα ελεύθερης αντιστοίχισης χρωμάτων, μεγάλης ποικιλίας
- Εύχρηστος τρόπος χάραξης trend lines
- Επιλογή κλίμακας στον άξονα των ημερομηνιών, στις χρονοσειρές
- Εύκολη εστίαση στις επιθυμητές ημερομηνίες
- Λύση προβλημάτων και άμεση απάντηση ερωτήσεων στο community forum
- Τα απεριόριστα, σχετικά, σημεία αναπαράστασης σε διαγράμματα και χάρτες
- Η ρύθμιση του μέσου (κέντρου), άνω και κάτω ορίου του άξονα διαβάθμισης χρωμάτων σε χάρτες περιοχών

Στα μειονεκτήματα του εργαλείου κατατάσσουμε τη δυσκολία εκμάθησης από το νέο χρήστη, με τη «φιλοσοφία» green και blue pills για τα ποσοτικά και ποιοτικά μεγέθη.

Επιπλέον αρνητικά στοιχεία:

- Περιορισμοί στην εισαγωγή αρχείων λόγω της θέσης του αρχείου (directory)
- Χρησιμότητα του data interpreter σε περιορισμένα αρχεία, δεν καταφέρνει να κάνει τις απαραίτητες μεταβολές
- Έλλειψη κατάλληλων πλαισίων για απεικόνιση αριθμητικών μεγεθών
- «Στενόχωρα» dashboard (πίνακες αναπαράστασης)
- Αδύνατη η παρέμβαση στο σύνολο δεδομένων, περιορισμένες δυνατότητες μετασχηματισμών των δεδομένων

3.5.2 MS Power BI

Όπως προέκυψε από τη διαδικασία των σεναρίων, το Power BI είναι ένα εργαλείο με εξαιρετικές δυνατότητες και διευκολύνσεις στο transformation των δεδομένων, με γρήγορη δημιουργία διαγραμμάτων και χαρτών με ακρίβεια και εύκολο χειρισμό. Το query editor αποδείχτηκε πολύ χρήσιμο για τις αναγκαίες μεταβολές στα δεδομένα και τα πεδία των πινάκων, με πλατφόρμες επεξεργασίας εύχρηστες από όλους τους χρήστες. Εντυπωσιακή ήταν και η δυνατότητα για δορυφορική εικόνα στους χάρτες, που δεν προσφέρεται στα άλλα εργαλεία.

Τα υπόλοιπα θετικά στοιχεία που αποκομίσαμε:

- Load δεδομένων με προεπισκόπηση
- Τα εύχρηστα, έτοιμα προς χρήση, παράθυρα δημιουργίας των calculated fields, με περιπτώσεις για διπλότυπα, κατά συνθήκη, αριθμητικών υπολογισμών κ.α.
- Άμεση μετονομασία πεδίων
- Λίστα ενεργειών, με δυνατότητα αναίρεσης, στο query editor
- Δημιουργία γραφημάτων με χειροκίνητη αντιστοίχιση των πεδίων στους άξονες
- Αναζήτηση στοιχείου σε στήλη
- Αντικατάσταση στοιχείου χειροκίνητα
- Χρήσιμα τα cards και multi-row cards για την παράθεση αριθμητικών στοιχείων
- Έτοιμο πλήκτρο group by στο query editor
- Δυνατότητα αντιγράφου του συνόλου δεδομένων
- Εύκολο drill up και drill down στην ιεραρχία των χρονοσειρών (έτος έως ημέρα)
- Εύκολη εστίαση στα επιθυμητά σημεία στις χρονοσειρές

Μειονεκτήματα που εντοπίσαμε:

- Αδύνατη η αναπαράσταση χωρίς aggregation, σε αριθμητικά δεδομένα
- Μη ικανοποιητική λειτουργία του φιλτραρίσματος όταν αφορά δύο ή περισσότερες διαστάσεις, ή δεν περιορίζεται σε ένα διάγραμμα
- Περιορισμός των σημείων που μπορούν να αναπαρασταθούν γραφικά στα 3500
- Η διαχείριση των σφαλμάτων (errors) των δεδομένων δεν παρέχει, συνήθως, λύσεις
- Τα χρώματα που αναπαρίστανται δεν συμβάλουν στην καλύτερη οπτική αντίληψη της πληροφορίας
- Η εισαγωγή δεδομένων βασισμένη στα 200 πρώτα στοιχεία εγκυμονεί τον κίνδυνο λανθασμένης αναγνώρισης τύπου δεδομένων
- Αργή «φόρτωση» των δεδομένων σε μεγάλα αρχεία
- Δεσμεύει πολλούς υπολογιστικούς πόρους, «βαρύ» πρόγραμμα

3.5.3 Qlik Sense

Το Qlik Sense ξεχώρισε για την αμεσότητα στη χρήση του, την ταχύτητα όσον αφορά απλά διαγράμματα, και τη δυνατότητα ρύθμισης και παρέμβασης στην ETL διαδικασία. Ιδιαίτερα χρήσιμα φάνηκαν σε πρακτικό επίπεδο τα αυτόματα δημιουργημένα διαγράμματα στο «παράθυρο» του data manager. Τα διαγράμματα αυτά εμφανίζονται με την επιλογή ενός πεδίου και δίνουν κάποιες πληροφορίες γι' αυτό, συνήθως ζητούμενες και χρήσιμες. Άλλο ένα πολύ θετικό στοιχείο ήταν η γρήγορη εισαγωγή των δεδομένων, ακόμα και από μεγάλα αρχεία, όπου υστερούσαν τα υπόλοιπα εργαλεία, όπως και η ανάγκη για λίγους

υπολογιστικούς πόρους (ελαφρύ πρόγραμμα). Επίσης το περιβάλλον με τα ξεχωριστά «apps» προσφέρει καλύτερη οργάνωση αρχείων και διαγραμμάτων.

Κατά την υλοποίηση των σεναρίων προέκυψαν τα εξής, επιπλέον, προτερήματα και θετικά στοιχεία του:

- Γρήγορη διαδικασία δημιουργίας απλών διαγραμμάτων
- επιλογές εμφάνισης "top ten" σε πλήκτρο με εμφάνιση των υπολοίπων στη στήλη «others»
- δυνατότητα χειροκίνητης εισαγωγής δεδομένων, άμεσα από τον χρήστη
- επεξεργασία / μετονομασία του τίτλου κάθε πεδίου
- δυνατότητα απόρριψης πεδίων κατά την εισαγωγή δεδομένων
- ρύθμιση παραμέτρων εισαγωγής με το script του load editor
- ευθείες αναφοράς στα διαγράμματα
- εύκολη τοποθέτηση πλαισίων περιγραφής πάνω στα διαγράμματα, στους τίτλους αυτών και στις εφαρμογές "apps" του εργαλείου.
- Λεπτομερής μεγέθυνση στον άξονα των χρονοσειρών, με απλό scroll
- Απλή, με πλήκτρο επιλογή διαβάθμισης των αξόνων των διαγραμμάτων – που όμως υστερεί σε λεπτομερή ρύθμιση
- Μενού εξερεύνησης των γραφημάτων, ώστε ο αποδέκτης να κάνει τις επιθυμητές αλλαγές σε διάταξη, ταξινόμηση και έλεγχο διαστάσεων, προσωρινά, για την διευκόλυνσή του.
- Η επαναφόρτωση των δεδομένων, σε περίπτωση αλλαγών από το data manager, γίνεται με πλήκτρο, δίνοντας πλήρη έλεγχο
- Επιλογή φιλτραρίσματος κατά την προβολή του διαγράμματος, από το «filter pane», λειτουργική όμως για ποιοτικές τιμές μόνο.

Αρνητική εντύπωση κατά τη χρήση έκανε η περιορισμένη δυνατότητα υλοποίησης διαγραμμάτων, τουλάχιστον χωρίς τη χρήση εντολών. Επίσης ελλιπείς ή δυσλειτουργικές οι διαδικασίες φιλτραρίσματος, ταξινόμησης και επιλογής χρωμάτων. Τα μειονεκτήματα που προέκυψαν:

- Κατά τη δημιουργία του διαγράμματος, για φιλτράρισμα (περιορισμούς) επιτρέπεται μόνο μία συνθήκη περιορισμού, δεν επιτρέπεται το φιλτράρισμα διαστάσεων, ούτε και η χειροκίνητη επιλογή τιμών
- Η ταξινόμηση είναι δυσλειτουργική, δεν εφαρμόζεται πάντα σύμφωνα με τις εκάστοτε ρυθμίσεις
- Αδύνατη η αναπαράσταση τιμών πεδίου απλά ως μεμονωμένα στοιχεία, χωρίς aggregation
- Δύσχρηστα calculated fields και «ιδιότροπες» συναρτήσεις
- Χρονοβόρες οι εναλλαγές μεταξύ των παραθύρων του εργαλείου
- Απουσία έτοιμων πλήκτρων για τα στατιστικά μεγέθη τυπικής απόκλισης, διακύμανσης και percentile
- Αδύνατη, χωρίς τη χρήση συναρτήσεων, η επιθυμητή αντιστοίχιση χρωμάτων στα στοιχεία του διαγράμματος
- Περιορισμός απεικόνισης στα 3330 περίπου σημεία, σε κάθε διάγραμμα
- Το group by γίνεται μόνο με συνάρτηση

- Δεν γίνεται αλλαγή του τύπου των δεδομένων, π.χ. από string (ακολουθία χαρακτήρων) σε αριθμό
- Δυσκολία στην αναίρεση κινήσεων

4 Συμπεράσματα

4.1 Αποτελέσματα Συγκρίσεων

Σύγκριση με γλώσσες προγραμματισμού

Η δημιουργία διαγραμμάτων μέσω εργαλείων είναι σαφώς πιο εύκολη και γρήγορη σε σχέση με τα scripts εντολών σε κάποια γλώσσα προγραμματισμού, καθώς απαιτούν σαφώς λιγότερη τεχνική κατάρτιση στο κομμάτι της ανάλυσης και απεικόνισης δεδομένων μέσω κώδικα. Οι θέσεις αυτές επαληθεύτηκαν και κατά τα σενάρια σύγκρισης για τα αριθμητικά δεδομένα και τις χρονοσειρές.

Τα διαγράμματα στα εργαλεία δημιουργούνται χωρίς καμία εντολή, μόνο με απλές επιλογές. Σε αντίθεση, με χρήση γλώσσας προγραμματισμού, απαιτείται η εισαγωγή των κατάλληλων βιβλιοθηκών για σχεδίαση διαγραμμάτων (π.χ. η matplotlib), εντολές για τις αριθμητικές πράξεις (aggregations), και για την υλοποίηση του γραφήματος. Ειδικά για τη δημιουργία δυναμικών διαγραμμάτων και dashboards, που είναι αναπόσπαστα στοιχεία των visual analytics, απαιτούνται ακόμα περισσότερος χρόνος και τεχνικές γνώσεις προγραμματισμού. Για τα διαγράμματα γεωγραφικών δεδομένων πρέπει να γίνει χειροκίνητα η αντιστοίχιση των συντεταγμένων και των ονομάτων χωρών, μέσω βιβλιοθηκών. Επίσης, πρέπει να εισαχθούν τα χωρικά αρχεία (spatial files), για τη βάση του επιθυμητού χάρτη.

Για την εισαγωγή δεδομένων από αρχείο και έπειτα επεξεργασία τους χρειάζεται πληθώρα εντολών. Μεγάλη δυσκολία εντοπίζεται στη σύνδεση με κάποια βάση δεδομένων ή κάποιο cloud αποθήκευσης, αφού κάθε τύπος πηγής δεδομένων χρειάζεται διαφορετικό σενάριο σύνδεσης, και πιθανόν στη γλώσσα που υποστηρίζει η πηγή των δεδομένων. Επιπλέον, οι ρυθμίσεις εμφάνισης γίνονται χωρίς να φαίνονται άμεσα τα αποτελέσματά τους, και να προσαρμόζονται αναλόγως. Το ίδιο ισχύει και για το φιλτράρισμα.

Σε τελικό απολογισμό, η προσκόλληση σε τεχνικά ζητήματα και ο περισσότερος χρόνος που απαιτείται για τη δημιουργία των διαγραμμάτων μέσω εντολών, δεν διευκολύνουν την εξαγωγή των visual analytics. Για την εξερευνητική ανάλυση δεδομένων, ο χρήστης ιδανικά πρέπει να εστιάζει στην επιλογή των δεδομένων, των στοιχείων και των μεγεθών που θα αναπαρασταθούν, τα κατάλληλα διαγράμματα αναπαράστασης και, φυσικά, την ερμηνεία τους. Από την άλλη, το προτέρημα των εντολών σε γλώσσα, είναι, προφανώς, οι απεριόριστες δυνατότητες για το σχεδιασμό διαγραμμάτων, ακόμα και έμπνευσης του χρήστη, και ο υπολογισμός οποιωνδήποτε αριθμητικών και στατιστικών συναρτήσεων.

Συνοπτική σύγκριση των εργαλείων

Από την εξέταση των δυνατοτήτων που προσφέρουν τα εργαλεία, προέκυψε ότι έχουν κοινή προσέγγιση σχετικά με τη διαδραστικότητα των διαγραμμάτων, όπως και στις βασικές δυνατότητες επεξεργασίας και της παρουσίασης των δεδομένων. Βέβαια, τα χαρακτηριστικά των εργαλείων και ο τρόπος λειτουργίας τους έχουν μεγάλες διαφορές, που έχουν μεγάλη σημασία για το χρήστη.

Η εισαγωγή των δεδομένων στο εργαλείο, που ως η βάση της οπτικής αναπαράστασης πρέπει να είναι έγκυρη και σωστή, έχει να κάνει με τη σωστή σύνδεση με τις πηγές. Πέραν των μεμονωμένων αρχείων, το πλεονέκτημα που προσφέρουν τα εργαλεία έγκειται στην άμεση

αποστολή αιτημάτων (queries) στις βάσεις δεδομένων, απαλλάσσοντας το χρήστη από τη διαδικασία δημιουργίας σύνδεσης. Κριτήριο επιλογής ενός εργαλείου λοιπόν είναι η δυνατότητα σύνδεσης με την πηγή δεδομένων που ο χρήστης επιθυμεί. Για τα τρία εργαλεία που εξετάσαμε, παρατίθεται στο παράρτημα ο αναλυτικός πίνακας των συμβατών πηγών.

Για την επεξεργασία των δεδομένων, κάθε εργαλείο έχει διαφορετική προσέγγιση. Το Tableau δίνει μεγαλύτερο βάρος στη δομή του πίνακα δεδομένων και στις συσχετίσεις μεταξύ των συνόλων. Το Power BI έχει πολύ μεγάλη γκάμα επιλογών για τους μετασχηματισμούς των αρχικών δεδομένων, απαλλαγμένων από τη χρήση εντολών. Το Qlik Sense επιτρέπει τις συσχετίσεις μεταξύ των πινάκων μέσω του γραφικού περιβάλλοντος και αυτόματα δημιουργεί, στο παράθυρο εποπτείας πίνακα, κατατοπιστικά διαγράμματα ανά πεδίο. Στα αρνητικά του η δημιουργία νέων πεδίων, που απαιτεί τη γνώση συναρτήσεων του εργαλείου. Μεγάλη διαφορά υπάρχει στο μέγιστο αριθμό σημείων που μπορούν να αναπαρασταθούν σε ένα διάγραμμα. Το Power BI μπορεί να απεικονίσει μέχρι και 3500 σημεία, όπως και το Qlik Sense 3330 περίπου σημεία. Το Tableau δεν έχει θεωρητικά κάποιο περιορισμό.

Όσον αφορά τη «φόρτωση» των δεδομένων εύκολη διαδικασία αυτόματης ανανέωσης των δεδομένων έχουν το Tableau και το Power BI, ενώ στο Qlik Sense είναι δυνατή μόνο μέσω σεναρίου (script). «Ζωντανή» ανανέωση, άμεση ταύτιση δηλαδή των δεδομένων της πηγής και του εργαλείου, παρέχει το Tableau και το Power BI Pro.

Στην οπτική αναπαράσταση και στην εξαγωγή αναλυτικών στοιχείων, το Tableau επιτρέπει τη δημιουργία πολύπλοκων διαγραμμάτων, που ο χρήστης ρυθμίζει τη μορφή τους, χωρίς περιορισμούς. Η ρύθμιση εμφάνισης και το φιλτράρισμα εμβαθύνουν σε όλες τις λεπτομέρειες. Επίσης παρέχεται η δυνατότητα προβλέψεων, για ορισμένες χρονοσειρές. Από την άλλη το Power BI και το Qlik Sense παρέχουν συγκεκριμένους τύπους διαγραμμάτων, αλλά πιο εύκολα υλοποιήσιμων. Η κλιμάκωση ευκολίας είναι αντίστροφη, δηλαδή, από αυτή των δυνατοτήτων. Το Power BI έχει επαρκείς δυνατότητες φιλτραρίσματος και ρυθμίσεις εμφάνισης, ενώ το Qlik Sense αρκετά περιορισμένες. Τέλος, σχετικά με τις παρουσιάσεις των αναλυτικών στοιχείων και τα τρία εργαλεία δεν υστερούν. Τα Tableau και Qlik Sense παρέχουν τη μορφή παρουσίασης «Story», ενώ το Power BI ευπαρουσίαστες αναφορές.

Όσον αφορά τις απαιτήσεις σε υπολογιστικούς πόρους, αλλά και την ταχύτητα λειτουργίας των εργαλείων, κατατάσσονται, μετά από τις παρατηρήσεις κατά τη διάρκεια των συγκριτικών σεναρίων, από το ελαφρύτερο στο πιο απαιτητικό και αργό ως εξής : Qlik Sense, Tableau, Power BI. Εκτός από τη χρηστικότητα και την εξοικονόμηση χρόνου, η κατάταξη αυτή είναι σημαντική και για τη διαχείριση μεγάλων αρχείων και, συνεπώς, μεγάλου όγκου δεδομένων.

4.2 Περιθώρια Βελτίωσης

Η βελτίωση και, φυσικά, η επέκταση των δυνατοτήτων είναι υψηλά στις προτεραιότητες των εταιρειών λογισμικού, οι οποίες μάλιστα δημοσιεύουν συχνά τις τάσεις στον τομέα αυτό [33]. Μία περιεκτική προσέγγιση των περιθωρίων βελτίωσης που έχουν τα εργαλεία οπτικής αναπαράστασης και ανάλυσης των δεδομένων, περιλαμβάνει τόσο τα ιδανικά

χαρακτηριστικά που μπορεί να αντιστοιχούν σε ένα εργαλείο με βάση τις δεδομένες δυνατότητες, όσο και τις μελλοντικές προοπτικές εξέλιξης.

Βασικά Χαρακτηριστικά

Μεγάλη σημασία για την αποτελεσματικότητα ενός εργαλείου έχει η δυνατότητα διαδραστικής δημιουργίας των όποιων επιθυμητών διαγραμμάτων, συνηθισμένων ή πιο πολύπλοκων, με απλή διαδικασία σχεδιασμού, όπου θα αρκεί ο χρήστης να σύρει τα πεδία προς απεικόνιση. Σίγουρα, μία λίστα επιλογής διαγραμμάτων είναι βοηθητική, αλλά θα πρέπει να μην περιορίζει μόνο στα διαγράμματα που παρέχει. Είναι λογικό, επίσης, το ιδανικό εργαλείο να δίνει όσο το δυνατόν περισσότερες επιλογές σύνδεσης με πηγές, επιλογής «ζωντανής σύνδεσης» με τα δεδομένα ή εισαγωγής, δυνατότητας επεξεργασίας μεγάλων συνόλων δεδομένων, υπολογισμό στατιστικών συναρτήσεων, εύκολης επεξεργασίας δεδομένων και συνδυασμού πινάκων. Όλα αυτά φυσικά χωρίς την ανάγκη χρήσης εντολών και συναρτήσεων. Όσον αφορά τις αναφορές, οι δυνατότητες που προσφέρουν τα εργαλεία μέχρι σήμερα κρίνονται αρκετά ικανοποιητικές ως προς το συνδυασμό πολλών διαγραμμάτων και αναλυτικών στοιχείων, στους πίνακες dashboards. Όσον αφορά τις παρουσιάσεις τύπου «stories», με απλούς χειρισμούς μπορεί να δημιουργηθεί μόνο κλασσική παρουσίαση σε διαφάνειες. Η δυνατότητα δημιουργίας πιο εντυπωσιακών stories με εύκολο τρόπο είναι σίγουρα κάτι που θα αναβαθμίσει την αποτελεσματική μετάδοση των μηνυμάτων και των πληροφοριών.

Διαδικτυακή σύνδεση και APIs

Μία χρήσιμη προσθήκη είναι η δυνατότητα σύνδεσης των διαγραμμάτων με υλικό από διαδικτυακές σελίδες, ιδιαίτερα ωφέλιμη για τα διαγράμματα χαρτών (π.χ. εμφάνιση στιγμιότυπου «streetview» με την επιλογή αντίστοιχου σημείου στο χάρτη). Ακόμα πιο χρήσιμα είναι τα εμφωλευμένα γραφήματα και αναλυτικά που εξάγονται από τα εργαλεία σε άλλες εφαρμογές ή ιστοσελίδες. Για να γίνουν οι διασυνδέσεις αυτές, χρειάζονται τα κατάλληλα APIs.

Καθαρισμός Δεδομένων και διασύνδεση με πηγές δεδομένων

Ίσως το σημαντικότερο στάδιο της διαδικασίας, που αφορά την ποιότητα των δεδομένων, ο καθαρισμός, επιδέχεται βελτιώσεων. Σχεδόν όλα τα αρχεία δεδομένων απέχουν από την ιδανική μορφή για διάβασμα των δεδομένων. Η επεξεργασία πινάκων με κακή οργάνωση, όπως τίτλους και σχόλια σε ενδιάμεσες σειρές, η διαχείριση των metadata και των λανθασμένων στοιχείων (π.χ. συμβολοσειρά σε αριθμητικά δεδομένα) είναι ζητούμενα. Μία πρόκληση θα ήταν να μπορεί το εργαλείο να διαβάζει και να ταξινομεί δεδομένα από αρχεία κειμένου (δεν αναφερόμαστε σε φυσική γλώσσα), που να μην είναι οργανωμένα σε πίνακες, αλλά απλά να παρατίθενται. Η ανάγνωση πινάκων απλής μορφής από pdf, που παρέχει το Tableau, είναι ένας πρόδρομος αυτής της διαδικασίας. Γενικά, για να εισάγονται σωστά τα δεδομένα, πρέπει να γίνεται σωστά η σύνδεση με την πηγή. Οπότε η δυνατότητα σύνδεσης με όσον το δυνατόν περισσότερων τύπων πηγές, οδηγεί και σε μικρότερη ανάγκη καθαρισμού των δεδομένων. Ιδιαίτερη έμφαση πρέπει να δοθεί στη σύνδεση με πηγές που σχετίζονται με τα μέσα κοινωνικής δικτύωσης, καθώς από τα δεδομένα τους μπορούν να εξαχθούν πληροφορίες, από τεράστιο στατιστικό δείγμα.

Αξιοποίηση Μηχανικής Μάθησης

Με την εισβολή του μεγάλου όγκου δεδομένων στους τομείς πολλών επαγγελμάτων, φαίνεται πως για τους χρήστες των εργαλείων – ακόμα και για αυτούς που δεν είναι ειδικοί στην ανάλυση δεδομένων, όπως εργαζόμενοι σε επιχειρήσεις ή σε ερευνητικά προγράμματα - θα είναι επιτακτική η ανάγκη για την ενσωμάτωση της μηχανικής μάθησης στις υπηρεσίες τους. Αυτό θα επιτρέψει την επεξεργασία κειμένου, τις ακριβέστερες προβλέψεις, την ταξινόμηση κατηγορικών δεδομένων σε ομάδες με επίβλεψη (classification) ή και χωρίς (clustering). Είναι, λοιπόν, σημαντικό να συνδυαστούν οι δυνατότητες της μηχανικής μάθησης με αυτές της εξερευνητικής ανάλυσης δεδομένων σε ολόκληρο το εύρος τους, ώστε να καλύπτονται οι ανάγκες των χρηστών από ένα μόνο εργαλείο.

Custom επιλογές μέσω scripts

Μία ακόμα επιθυμητή προσθήκη που, αυτή τη φορά αφορά τους εξειδικευμένους χρήστες, όπως ερευνητές με τεχνολογικό υπόβαθρο, είναι η παροχή μιας πλατφόρμας τροποποιήσεων μέσω εντολών σε κάποια γλώσσα προγραμματισμού. Η χρησιμότητα στην περίπτωση αυτή έγκειται στην ανάγκη αυτής της (μικρής) ομάδας χρηστών να διαμορφώσουν τα διαγράμματά τους μέσω κάποιου script εντολών, αλλά να εκμεταλλευτούν τη διαδραστική παρουσίαση και την αλληλεπίδραση μεταξύ των διαγραμμάτων, που προσφέρουν τα εργαλεία.

Ζωντανή σύνδεση υψηλής ακρίβειας

Στο κοντινό μέλλον, με την αναμενόμενη διασύνδεση πλήθους συσκευών, επαγγελματικής ή προσωπικής χρήσης, και αισθητήρων με το διαδίκτυο, η ανάγκη για πραγματικά «ζωντανή» σύνδεση με πολλές πηγές δεδομένων θα είναι υψίστης σημασίας. Για την αναπαράσταση δεδομένων όπως οικονομικών στοιχείων ή δημογραφικών η προγραμματισμένη ανανέωση σε ελάχιστα δευτερόλεπτα μπορεί να είναι υπέρ αρκετή. Όμως για την παρακολούθηση της κατάστασης ενός ασθενούς σε ένα χειρουργείο ή την παρακολούθηση της θερμοκρασίας κατά τη διάρκεια ενός πειράματος απαιτείται συνεχής επεξεργασία των τιμών, ώστε τα αναλυτικά στοιχεία που προκύπτουν να είναι έγκυρα. Ως εκ τούτου, όλα τα εργαλεία για να καλύψουν τις απαιτήσεις που προκύπτουν, θα πρέπει να παρέχουν ζωντανή σύνδεση με τις πηγές δεδομένων. Σήμερα, συναντάται σε ελάχιστα και με δυνατότητα σύνδεσης για λίγους τύπους πηγών

Ασφάλεια

Τα περισσότερα εργαλεία λαμβάνουν μέτρα για την ασφάλεια των δεδομένων, ειδικά όταν πρόκειται για εκδόσεις server ή εταιρικές. Ο μεγάλος αριθμός των βάσεων δεδομένων προς σύνδεση, όπως και γενικά οι προκλήσεις στην κρυπτογραφία, απαιτούν συνεχείς ανανεώσεις και προσθήκες στον τομέα της ασφάλειας. Όμως, ιδιαίτερο βάρος θα μπορούσε να δοθεί στην προστασία των ευαίσθητων δεδομένων, κατά την λήψη συμβουλών είτε από επαγγελματίες είτε από άλλους χρήστες. Χαρακτηριστικό παράδειγμα που συναντήθηκε πολλές φορές στις κοινότητες των εργαλείων, η αδυναμία σωστής περιγραφής διαφόρων προβλημάτων, λόγω ευαίσθητων δεδομένων.

Αισθητική και περιβάλλον χρήσης

Δεν πρέπει βέβαια να αμεληθεί η αισθητική των εργαλείων, τόσο ως προς τα διαγράμματα που δημιουργούνται όσο και ως προς το περιβάλλον εργασίας. Εξ’ άλλου η σωστή μετάδοση

της πληροφορίας επιτυγχάνεται με ελκυστικές αναπαραστάσεις. Είναι γεγονός άλλωστε ότι τα εργαλεία που συγκεντρώνουν την προτίμηση των χρηστών, υπερτερούν αισθητικά και γραφικά, και επιτρέπουν τον εκδημοκρατισμό των δεδομένων.

Όλα τα παραπάνω, που έχουν παρατεθεί, αποτελούν συγκεκριμένες βελτιώσεις στα πλαίσια των ήδη υπαρχόντων προτερημάτων του συνόλου των υφιστάμενων εργαλείων. Το πως θα μοιάζουν τα εργαλεία οπτικής αναπαράστασης στο επόμενο διάστημα δεν μπορεί να προβλεφθεί, ωστόσο τα «δειλά βήματα» σε κάποια καινοτόμα χαρακτηριστικά δείχνουν την κατεύθυνση.

Augmented Analytics - Αυτόματη εξαγωγή αναλυτικών στοιχείων και δημιουργία διαγραμμάτων

Τα χαρακτηριστικά αυτά αφορούν τα επαυξημένα αναλυτικά στοιχεία (augmented analytics) και την κατανόηση της φυσικής γλώσσας για τη λήψη εντολών. Σε ένα εργαλείο augmented analytics, λοιπόν, ο χρήστης διατυπώνει σε φυσική γλώσσα τις όποιες ερωτήσεις του. Η λεπτομέρεια αυτή είναι σημαντική, γιατί το εργαλείο προσεγγίζει ολιστικά ένα ερώτημα, αναδεικνύοντας χρήσιμα δεδομένα που πιθανώς ο απλός χρήστης δεν θα αξιοποιούσε. Μέσω των αναλυτικών στοιχείων αυτών, ιδανικά, εξαλείφεται η ανάγκη για καθοδήγηση και υποστήριξη από επαγγελματίες αναλυτές δεδομένων, καθώς τα επιθυμητά στοιχεία και διαγράμματα πλαισιώνονται αυτόματα και από τα όποια απαραίτητα δεδομένα που θα συμβάλλουν στην εξαγωγή συμπερασμάτων από το χρήστη. Βέβαια, η επαρκής αντικατάσταση του αναλυτή δεδομένων από τα augmented analytics, αφορά, στην καλύτερη περίπτωση, το κοντινό μέλλον και απαιτεί φυσικά προηγμένο επίπεδο τεχνητής νοημοσύνης. Λεπτό σημείο για την αυτόματη επιλογή διαγραμμάτων είναι να μην κατευθύνει το εργαλείο το χρήστη και να μην «σκέφτεται» στη θέση του. Αντίθετα, πρέπει να του δίνει τις διαθέσιμες επιλογές απαλλάσσοντας τον από οτιδήποτε περιττό πέραν των οπτικών αναλυτικών στοιχείων. Όσο εξελιγμένες και να είναι οι δυνατότητες παροχής αυτόματων διαγραμμάτων και ανάλυσης δεδομένων, δεν μπορούν να καλύψουν την κάθε μία ιδιαίτερη περίπτωση και να αντικαταστήσουν την ανθρώπινη κρίση- μπορούν όμως να αναδείξουν όλες τις χρήσιμες, μη προφανείς, πληροφορίες.

Εντολές σε Φυσική Γλώσσα

Πέρα από τη χρησιμότητα των ερωτημάτων σε φυσική γλώσσα για τα augmented analytics, δεν πρέπει να παραβλεφθούν οι δυνατότητες για πλήρη χειρισμό των εργαλείων που προσφέρονται. Χαρακτηριστικά, ο χρήστης θα μπορούσε να σχεδιάζει τα διαγράμματα και ζητά υπολογισμούς με φωνητικές εντολές, ή, ακόμα καλύτερα, καθοδηγώντας έναν εικονικό βοηθό του [34]. Η επικοινωνία σε φυσική γλώσσα μπορεί, συνεπώς, να αλλάξει τη διαδικασία αλληλεπίδρασης ανθρώπου και λογισμικού, καθιστώντας την άμεση και ελάχιστα απαιτητική από πλευράς προσπάθειας.

Φορητότητα εργαλείων

Με τις προσθήκες αυτές, η πλήρης λειτουργία των εργαλείων, για σχεδίαση και επιλογή των διαγραμμάτων, θα είναι δυνατή και στις φορητές συσκευές, οι οποίες αποτελούν πλέον το μεγαλύτερο μέρος των ηλεκτρονικών συσκευών πρόσβασης στο διαδίκτυο. Η φορητότητα όλων των δυνατοτήτων των εργαλείων, σε συνδυασμό με τον ελάχιστο απαιτούμενο χρόνο για το σχεδιασμό των διαγραμμάτων, θα καταστήσει την εξερευνητική ανάλυση δεδομένων εφικτή από ακόμα περισσότερα άτομα, αίροντας τους περιορισμούς στο χρόνο, στην τοποθεσία και στις τεχνικές δεξιότητες. Μάλιστα, στην ενίσχυση της φορητότητας, θα συντελούσε και η δυνατότητα μερικής λειτουργίας των εργαλείων χωρίς σύνδεση στο διαδίκτυο.

Εν κατακλείδι, στη διαδικασία συνεχούς εξέλιξης των εργαλείων, ο άνθρωπος θα έχει κεντρικό ρόλο. Στην ανάπτυξή τους θα συμβάλλουν και ειδικότητες όπως ψυχολόγοι, κοινωνικοί επιστήμονες, ανθρωπολόγοι και καλλιτέχνες. Η αλληλεπίδραση ανθρώπου-μηχανής έχει ανάγκη και αυτές τις επιστήμες, ίσως ακόμα και στον ίδιο βαθμό με την τεχνολογία. Σε κάθε περίπτωση, τα εργαλεία οπτικής αναπαράστασης θα υποβοηθούν την ανθρώπινη αντίληψη, «ρίχνοντας φως» σε ό,τι κρύβουν τα δεδομένα και επιτρέποντας στους χρήστες να εστιάσουν στα πιο ενδιαφέροντα σημεία.

5 Βιβλιογραφία

[1] M. Posner, S. Petersen, “The attention system of the human brain”, Annual Review of Neuroscience ,Vol 13:25-42, 1995

[2] Cavanillas J., Curry E., Wahlster W “ The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches”, New Horizons for a Data-Driven Economy. Springer, Cham, 2016.

[3] E. Tufte, “The visual display of quantitative information”, 1983, pp. 223-224.J. Snow, “On the mode of communication of cholera”, 1855S. Few, “Selecting the right graph for your message”, Perceptual Edge, 2004.

[4] J.Snow, “On the mode of communication of cholera”, 1855.

[5] S.Few, “Selecting the right graph for your message”, Perceptual Edge, 2004.

[6] J.Thomas, K.Cook, “Illuminating the path: The research and development agenda for visual analysis”, National Visualization and Analytics Center, 2005, pp. 7-11

[7] Visual Analytics”, Wikipedia. [Online]. Διαθέσιμο στο: https://en.wikipedia.org/ wiki/ Visual_analytics#cite_note-4. [Πρόσβαση: 20- Dec- 2017].

[8] S. Few, “What do data analysts most need from their tools”, Perceptual Edge, 2015.

[9] “Business Analytics”, Wikipedia. [Online]. Διαθέσιμο στο: https://en.wikipedia.org /wiki/Business_analytics [Πρόσβαση : 9-Jan-2018]

[10]Tibco Spotfire Documentation. [Online]. Διαθέσιμο στο : https://docs.tibco.com. [Πρόσβαση: 15- Jan – 2018]

[11]Plot.ly Documentation. [Online]. Διαθέσιμο στο : https://plot.ly/python/user-guide/ [Πρόσβαση: 15- Jan – 2018]

[12]Kibana Documentation. [Online]. Διαθέσιμο στο : https://www.elastic.co/guide/en/kibana/index.html [Πρόσβαση : 16- Jan - 2018]

[13]Chartio Documentation. [Online]. Διαθέσιμο στο : https://support.chartio.com/docs/index. [Πρόσβαση: 16- Jan – 2018]

[14]Sisense Documentation. [Online]. Διαθέσιμο στο : https://documentation.sisense.com/. [Πρόσβαση: 16- Jan – 2018]

[15]Domo Documentation. [Online]. Διαθέσιμο στο : https://www.domo.com/product. [Πρόσβαση: 16- Jan – 2018]

[16]Watson IBM Documentation. [Online]. Διαθέσιμο στο : https://console.bluemix.net/developer/watson/documentation. [Πρόσβαση: 16- Jan – 2018]

[17]SAS Visual Analytics. [Online]. Διαθέσιμο στο : https://support.sas.com/documentation/onlinedoc/va/. [Πρόσβαση: 16- Jan – 2018]

[18]Pentaho Documentation. [Online]. Διαθέσιμο στο : https://help.pentaho.com/Documentation/7.1 . [Πρόσβαση: 17- Jan – 2018]

[19]Rapidminer Documentation. [Online]. Διαθέσιμο στο : https://docs.rapidminer.com/. [Πρόσβαση: 17- Jan – 2018]

[20]Knime Documentation. [Online]. Διαθέσιμο στο : https://www.knime.com/documentation. [Πρόσβαση: 17- Jan – 2018]
[21]CartoDB Documentation. [Online]. Διαθέσιμο στο : https://carto.com/docs/. [Πρόσβαση: 20- Jan – 2018]
[22]Gephi Documentation. [Online]. Διαθέσιμο στο : https://gephi.org/users/. [Πρόσβαση: 5- Oct – 2017]]
[23]D3.js Documentation. [Online]. Διαθέσιμο στο : https://github.com/d3/d3/wiki [Πρόσβαση: 19- Jan – 2018]
[24]"Pivot Table", Wikipedia. [Online]. Διαθέσιμο στο: https://en.wikipedia.org/ wiki/Pivot_table . [Πρόσβαση: 18- Jan- 2018].
[25]"Market Share", Datanyze. [Online]. Διαθέσιμο στο: https://www.datanyze.com/ market-share/business-intelligence. [Πρόσβαση: 11-Jan-2018].
[26]"Tableau Desktop", Tableau. [Online]. Διαθέσιμο στο: https://www.tableau.com/ products/desktop. [Πρόσβαση: 8 – Oct- 2017].
[27]"Tableau Technology", Tableau. [Online]. Διαθέσιμο στο: https://www.tableau.com /products/technology. [Πρόσβαση: 2 – Nov – 2017].
[28]"Ms Power BI Desktop", Microsoft. [Online]. Διαθέσιμο στο: https://powerbi.microsoft.com/en-us/ desktop/. [Πρόσβαση : 4 – Nov – 2017]
[29]"Qlik Sense", Qlik. [Online]. Διαθέσιμο στο: https://www.qlik.com/us/products/qlik-sense. [Πρόσβαση: 20 – Nov – 2017].
[30]"2012 General Election: Romney vs Obama", Huffington Post. [Online]. Διαθέσιμο στο: http://elections.huffingtonpost.com/pollster/2012-general-election-romney-vs-obama. [Πρόσβαση : 20- Nov – 2017].
[31]"Life expectancy and Healthy life expectancy ", World Health Organization. [Online]. Διαθέσιμο στο: http://apps.who.int/gho/data/view.main.SDG2016LEXv?%20. [Πρόσβαση: 22 – Jan – 2018].
[32]"Fast food maps data", fastfoodmaps.com [Online]. Διαθέσιμο στο: http://www.fastfoodmaps.com/data.html. [Πρόσβαση: 10 – Dec -2017].
[33]"Top 10 Business Intelligence trends 2018", Tableau. [Online]. Διαθέσιμο στο: https://www.tableau.com/reports/business-intelligence-trends. [Πρόσβαση: 2- Mar- 2018]
[34]T. Chandler, M. Cordeil, T. Czauderna, T. Dwyer, J. Glowacki, C. Goncu, "Immersive Analytics", 2015

6 Παράρτημα

6.1 Εντολές σεναρίου χρονοσειράς

```
# For data
import pandas as pd
from pandas import Series,DataFrame
import numpy as np

# For visualization
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('whitegrid')
%matplotlib inline

from __future__ import division


# Use to grab data from the web(HTTP capabilities)
import requests

# We'll also use StringIO to work with the csv file, the DataFrame will require a .read() method
from StringIO import StringIO
# This is the url link for the poll data in csv form
url = "http://elections.huffingtonpost.com/pollster/2012-general-election-romney-vs-obama.csv"

# Use requests to get the information in text form
source = requests.get(url).text

# Use StringIO to avoid an IO error with pandas
poll_data = StringIO(source)
```

Γίνεται έπειτα μία επισκόπηση στα δεδομένα :

```
# Set poll data as pandas DataFrame
poll_df = pd.read_csv(poll_data)

# Let's get a glimpse at the data
poll_df.info()
```

Out:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 589 entries, 0 to 588
Data columns (total 14 columns):
Pollster                589 non-null object
Start Date              589 non-null object
End Date                589 non-null object
Entry Date/Time (ET)    589 non-null object
Number of Observations  567 non-null float64
Population              589 non-null object
Mode                    589 non-null object
Obama                   589 non-null int64
Romney                  589 non-null int64
Undecided               422 non-null float64
Pollster URL            589 non-null object
Source URL              587 non-null object
Partisan                589 non-null object
Affiliation             589 non-null object
dtypes: float64(2), int64(2), object(10)
```

```
memory usage: 69.0+ KB


# Preview DataFrame
poll_df.head()


# Factorplot the affiliation
sns.factorplot('Affiliation',data=poll_df)

# First we'll get the average
avg = pd.DataFrame(poll_df.mean())
avg.drop('Number of Observations',axis=0,inplace=True)

# After that let's get the error
std = pd.DataFrame(poll_df.std())
std.drop('Number of Observations',axis=0,inplace=True)

# now plot using pandas built-in plot, with kind='bar' and yerr='std'
avg.plot(yerr=std,kind='bar',legend=False)


# Concatenate our Average and Std DataFrames
poll_avg = pd.concat([avg,std],axis=1)

#Rename columns
poll_avg.columns = ['Average','STD']

#Show
poll_avg


# Quick plot of sentiment in the polls versus time.
poll_df.plot(x='End Date',y=['Obama','Romney','Undecided'],marker='o',linestyle='')

# For timestamps
from datetime import datetime
# Create a new column for the difference between the two candidates
poll_df['Difference'] = (poll_df.Obama - poll_df.Romney)/100
# Preview the new column
poll_df.head()
# Set as_index=Flase to keep the 0,1,2,... index. Then we'll take the mean of the polls on that day.
poll_df = poll_df.groupby(['Start Date'],as_index=False).mean()
# Plotting the difference in polls between Obama and Romney
fig = poll_df.plot('Start Date','Difference',figsize=(12,4),marker='o',linestyle='-',color='purple')


# Set row count and xlimit list
row_in = 0
xlimit = []

# Cycle through dates until 2012-10 is found, then print row index
for date in poll_df['Start Date']:
    if date[0:7] == '2012-10':
        xlimit.append(row_in)
        row_in +=1
    else:
        row_in += 1

print min(xlimit)
```

```
print max(xlimit)

Out:
329
356

# Start with original figure
fig = poll_df.plot('Start Date','Difference',figsize=(12,4),marker='o',linestyle='-
',color='purple',xlim=(329,356))

# Now add the debate markers
plt.axvline(x=329+2, linewidth=4, color='grey')
plt.axvline(x=329+10, linewidth=4, color='grey')
plt.axvline(x=329+21, linewidth=4, color='grey')
```

6.2 Εντολές σεναρίου αριθμητικών στοιχείων

```
# Set the DataFrame as the csv file
donor_df = pd.read_csv('Election_Donor_Data.csv')

# Get a quick overview
donor_df.info()

Out:
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1001731 entries, 0 to 1001730
Data columns (total 16 columns):
cmte_id             1001731 non-null object
cand_id             1001731 non-null object
cand_nm              1001731 non-null object
contbr_nm            1001731 non-null object
contbr_city          1001712 non-null object
contbr_st            1001727 non-null object
contbr_zip           1001620 non-null object
contbr_employer      988002 non-null object
contbr_occupation    993301 non-null object
contb_receipt_amt    1001731 non-null float64
contb_receipt_dt     1001731 non-null object
receipt_desc         14166 non-null object
memo_cd              92482 non-null object
memo_text            97770 non-null object
form_tp              1001731 non-null object
file_num             1001731 non-null int64
dtypes: float64(1), int64(1), object(14)
memory usage: 129.9+ MB



# let's also just take a glimpse
donor_df.head()
# Get a quick look at the various donation amounts
donor_df['contb_receipt_amt'].value_counts()

Out:
100.0    178188
50.0     137584
25.0     110345
250.0     91182
```

```
500.0      57984
2500.0     49005
35.0       37237
1000.0     36494
10.0       33986
200.0      27813
20.0       17565
15.0       16163
150.0      14600
75.0       13647
201.2      11718
...
0.88       1
19.35      1
58.18      1
71.20      1
70.68      1
163.90     1
14.97      1
264.39     1
162.60     1
81.15      1
45.15      1
106.56     1
62.20      1
58.82      1
73.83      1
Length: 8079, dtype: int64

# Get the mean donation
don_mean = donor_df['contb_receipt_amt'].mean()

# Get the std of the donation
don_std = donor_df['contb_receipt_amt'].std()

print 'The average donation was %.2f with a std of %.2f' %(don_mean,don_std)

# Let's make a Series from the DataFrame, use .copy() to avoid view errors
top_donor = donor_df['contb_receipt_amt'].copy()

# Now sort it
top_donor.sort()

# Then check the Series
top_donor


Out:
114604   -30800.00
226986   -25800.00
101356    -7500.00
398429    -5500.00
250737    -5455.00
33821     -5414.31
908565    -5115.00
456649    -5000.00
574657    -5000.00
30513     -5000.00
562267    -5000.00
30584     -5000.00
```

```
86268     -5000.00
708920    -5000.00
665887    -5000.00
...
90076      10000.00
709859     10000.00
41888      10000.00
65131      12700.00
834301     25000.00
823345     25000.00
217891     25800.00
114754     33300.00
257270    451726.00
335187    512710.91
319478    526246.17
344419   1511192.17
344539   1679114.65
326651   1944042.43
325136   2014490.51
Name: contb_receipt_amt, Length: 1001731, dtype: float64
```

```
 # Get rid of the negative values
top_donor = top_donor[top_donor >0]

# Sort the Series
top_donor.sort()

# Look at the top 10 most common donations value counts
top_donor.value_counts().head(10)
```

```
Out:
100    178188
50     137584
25     110345
250     91182
500     57984
2500    49005
35      37237
1000    36494
10      33986
200     27813
dtype: int64
```

```
# Create a Series of the common donations limited to 2500
com_don = top_donor[top_donor < 2500]

# Set a high number of bins to account for the non-round donations and check histogram for spikes.
com_don.hist(bins=100)
# Grab the unique object from the candidate column
candidates = donor_df.cand_nm.unique()
#Show
candidates
```

```
array(['Bachmann, Michelle', 'Romney, Mitt', 'Obama, Barack',
       "Roemer, Charles E. 'Buddy' III", 'Pawlenty, Timothy',
       'Johnson, Gary Earl', 'Paul, Ron', 'Santorum, Rick', 'Cain, Herman',
       'Gingrich, Newt', 'McCotter, Thaddeus G', 'Huntsman, Jon',
       'Perry, Rick'], dtype=object)
```

```
# Dictionary of party affiliation
party_map = {'Bachmann, Michelle': 'Republican',
          'Cain, Herman': 'Republican',
          'Gingrich, Newt': 'Republican',
          'Huntsman, Jon': 'Republican',
          'Johnson, Gary Earl': 'Republican',
          'McCotter, Thaddeus G': 'Republican',
          'Obama, Barack': 'Democrat',
          'Paul, Ron': 'Republican',
          'Pawlenty, Timothy': 'Republican',
          'Perry, Rick': 'Republican',
          "Roemer, Charles E. 'Buddy' III": 'Republican',
          'Romney, Mitt': 'Republican',
          'Santorum, Rick': 'Republican'}

# Now map the party with candidate
donor_df['Party'] = donor_df.cand_nm.map(party_map)
for i in xrange(0,len(donor_df)):
   if donor_df['cand_nm'][i] == 'Obama,Barack':
      donor_df['Party'][i] = 'Democrat'
   else:
      donor_df['Party'][i] = 'Republican'

# Clear refunds
donor_df = donor_df[donor_df.contb_receipt_amt >0]

# Preview DataFrame
donor_df.head()
# Groupby candidate and then displayt the total number of people who donated
donor_df.groupby('cand_nm')['contb_receipt_amt'].count()

cand_nm
Bachmann, Michelle              13082
Cain, Herman                    20052
Gingrich, Newt                  46883
Huntsman, Jon                    4066
Johnson, Gary Earl               1234
McCotter, Thaddeus G               73
Obama, Barack                  589127
Paul, Ron                      143161
Pawlenty, Timothy                3844
Perry, Rick                     12709
Roemer, Charles E. 'Buddy' III      5844
Romney, Mitt                   105155
Santorum, Rick                   46245
Name: contb_receipt_amt, dtype: int64
# Groupby candidate and then displayt the total amount donated
donor_df.groupby('cand_nm')['contb_receipt_amt'].sum()


cand_nm
Bachmann, Michelle             2.711439e+06
Cain, Herman                   7.101082e+06
Gingrich, Newt                 1.283277e+07
Huntsman, Jon                  3.330373e+06
Johnson, Gary Earl             5.669616e+05
McCotter, Thaddeus G           3.903000e+04
Obama, Barack                  1.358774e+08
Paul, Ron                      2.100962e+07
Pawlenty, Timothy              6.004819e+06
```

```
Perry, Rick                    2.030575e+07
Roemer, Charles E. 'Buddy' III    3.730099e+05
Romney, Mitt                   8.833591e+07
Santorum, Rick                  1.104316e+07
Name: contb_receipt_amt, dtype: float64

# Plot out total donation amounts
cand_amount.plot(kind='bar')
# Use a pivot table to extract and organize the data by the donor occupation
occupation_df = donor_df.pivot_table('contb_receipt_amt',
                        index='contbr_occupation',
                        columns='Party', aggfunc='sum')


# Check size
occupation_df.shape
(45067, 2)

# Set a cut off point at 1 milllion dollars of sum contributions
occupation_df = occupation_df[occupation_df.sum(1) > 1000000]
# Now let's check the size!
occupation_df.shape

Out:
(31, 2)

# plot out with pandas
occupation_df.plot(kind='bar')
# Horizontal plot, use a convienently colored cmap
occupation_df.plot(kind='barh',figsize=(10,12),cmap='seismic')
```

6.3 Πίνακας πηγών εισαγωγής δεδομένων

	Tableau Public 10.4	MS PowerBI	Tableau Desktop Professional	Qlik Sense
DATA INPUT				
browse organization(company) data		✓		
content packs of online services		✓		
files				
excel	✓	✓	✓	✓
text/csv	✓	✓	✓	✓
xml		✓		✓
json	✓	✓	✓	✓
folder	✓	✓	✓	✓
sharepoint folder		✓	✓	
PDF file	✓		✓	
database				
SQL Server		✓	✓	✓
MySQL		✓	✓	
Access		✓	✓	
Teradata		✓	✓	✓
SQL Server Analysis Services		✓	✓	
Oracle		✓	✓	✓
Impala		✓	✓	✓
IBM DB2		✓	✓	✓
IBM Infomix		✓		
IBM Netezza		✓		
PostgreSQL		✓	✓	✓
Sybase		✓	✓	✓
SAP HANA		✓	✓	
SAP Business Warehouse Server		✓	✓	
Amazon Redshift		✓	✓	✓
Google BigQuery		✓	✓	✓
Snowflake		✓	✓	
Azure SQL database		✓	✓	
Azure SQL Data Warehouse		✓	✓	
Azure Analysis Services		✓		
Azure Blob Storage		✓		
Azure Cosmos DB		✓		
Azure Data Lake Store		✓	✓	
Azure HDInsight		✓		
Azure HDInsight Spark		✓		
Online Services				
Power BI service		✓		
Sharepoint Online List		✓		
Microsoft Exchange Online		✓		
Dynamics 365		✓		
Dynamics 365 for Financials		✓		
Common Data Service		✓		
Azure Enterprise		✓		
Visual Studio Team Services		✓		
Salesforce Objects		✓		
Salesforce Reports		✓		✓
Google Sheets	✓	✓	✓	
appFigures		✓		
comScore Digital Analytix		✓		
Dynamics 365 for Customer Insights		✓		
Facebook		✓		
GitHub		✓		
MailChimp		✓		✓
Marketo		✓	✓	
Mixpanel		✓		
Planview Enterprise		✓		
Projectplace		✓		
QuickBooks Online		✓	✓	
Smartsheet		✓		
SparkPost		✓		
SQL Sentry		✓		
Stripe		✓		
SweetIQ		✓		
Troux		✓		
Twilio		✓		
tyGraph		✓		
Webtrends		✓		
Zendesk		✓		
.....				
Web	✓	✓	✓	✓
SharePoint list		✓	✓	
O Data Feed	✓	✓	✓	
Active Directory		✓		
Microsoft Exchange		✓		
Hadoop File		✓		
Spark		✓	✓	
R script	✓	✓	✓	
ODBC		✓	✓	✓
OLE DB		✓		✓
Blank Query		✓		
Tableau Data Extract	✓		✓	
statistical files				
SPSS (*.sav)	✓		✓	
SAS (*.sas7bdat)	✓		✓	
R (*.rda , *.rdata)	✓	✓	✓	

spatial files				
KML (*.kml)	✓		✓	
ESRI (*.shp)	✓	✓	✓	
MapInfo Tables (*.tab)	✓		✓	
MapInfo Interchange format (*.mif)	✓		✓	
*.geojson	✓		✓	
*.topojson		✓		
.....				
Actian Matrix			✓	
Actian Vector			✓	
Amazon Athena			✓	
Amazon Aurora			✓	
Amazon Elastic MapReduce			✓	
Apache Drill			✓	
Aster Database			✓	
Cisco Information Server			✓	
Datastax enterprise			✓	
Denodo			✓	
Dropbox			✓	✓
HP Vertica			✓	
IBM Biginsights			✓	
IBM PDA			✓	
Kognitio			✓	
MarkLogic			✓	
MemSQL			✓	
MonetDB			✓	
MongoDB			✓	✓
MongoDB BI			✓	
Oracle Eloqua			✓	
Oracle Essabase			✓	
Pivotal Greenplum Database			✓	
Presto			✓	
SAP Sybase ASE			✓	
SAP Sybase IQ			✓	
ServiceNow ITSM			✓	
Splunk			✓	
Teradata OLAP Connector			✓	✓
firebird			✓	
exasol			✓	
Google Cloud SQL			✓	
Google Analytics		✓	✓	✓
MS OneDrive*		✓	✓	✓
MS PowerPivot		✓	✓	
Progress OpenEdge			✓	
Salesforce.com			✓	✓
Adobe Analytics				✓
Amazon S3				✓
bitly				✓
Blue Yonder				✓
Box				✓
Exacto Online				✓
File Transfer (FTP/SFTP)				✓
FreeAgent				✓
JIRA				✓
IMAP/POP3 Mailboxes				✓
Klout				✓
Mashape				✓
Microsoft Dynamics CRM				✓
Notification Connector				✓
Odata				✓
SugarCRM				✓
SurveyMonkey				✓
Azure Data Marketplace				✓
Xero				✓
YouTube Analytics				✓
YouTube Data				✓
Google AdSense				✓
Google AdWards				✓
Google Calendar				✓
Google Prediction				✓
Google Webmaster Tools				✓
Google DoubleClick for Advertisers				✓
Google DoubleClick for Publishers				✓
Google+				✓
Twitter				✓