ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ & ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Συγκριτική Μελέτη Δεδομένων Πολυτροπικών Συστημάτων Μεταφοράς της Ευρώπης και της Αμερικής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΝΙΚΟΛΑΟΥ ΓΑΛΑΝΗ

Επιβλέπουσα : Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ
& ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Συγκριτική Μελέτη Δεδομένων Πολυτροπικών Συστημάτων Μεταφοράς της Ευρώπης και της Αμερικής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΝΙΚΟΛΑΟΥ ΓΑΛΑΝΗ

Επιβλέπουσα : Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Σεπτεμβρίου 2024.

.....	.....	.....
Θεοδώρα Βαρβαρίγου	Εμμανουήλ Βαρβαρίγος	Συμεών Παπαβασιλείου
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2024

.....

ΝΙΚΟΛΑΟΣ ΓΑΛΑΝΗΣ

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

Περίληψη

Ο σκοπός της διπλωματικής εργασίας ήταν αρχικά η περιγραφή των προβλημάτων και προκλήσεων στην σύγχρονη μετακίνηση σε μεγάλες πόλεις στην προσπάθεια να γίνουν πιο αποδοτικές και περιβαλλοντολογικά βιώσιμες. Στην συνέχεια αναφέρθηκαν τα βασικά εργαλεία και τεχνολογίες της επιστήμης ανάλυσης δεδομένων που χρησιμοποιούνται και αναπτύσσονται στην προσπάθεια αυτή.

Ακολούθησε μελέτη, προεπεξεργασία, οπτικοποίηση και ανάλυση των αποτελεσμάτων με τα εργαλεία αυτά μίας ευρείας γκάμας πολυτροπικών συστημάτων μεταφοράς συνόλων δεδομένων σε διάφορες πόλεις της Ευρώπης και της Αμερικής. Βασική πόλη σύγκρισης για όλα τα μέσα μεταφοράς ήταν η Νέα Υόρκη και έγινε έρευνα για να συμπεριληφθεί η μεγαλύτερη δυνατή γκάμα τρόπων μεταφορών μεταφοράς στην ανάλυση.

Η μεθοδολογία αυτή μπορεί να χρησιμοποιηθεί ως υπόδειγμα για την προεπεξεργασία και την περιγραφική και διαγνωστική ανάλυση παρόμοιων συνόλων δεδομένων. Επίσης αποτελεί τα πρώτα βήματα πριν την προγνωστική ανάλυση και εφαρμογή των τεχνικών μηχανικής μάθησης και τεχνητής νοημοσύνης για την πρόβλεψη διαφόρων παραμέτρων του ταξιδιού με κάθε μέσο μεταφοράς. Επόμενα στάδια ανάπτυξης είναι ο συνδυασμός της ανάλυσης και των συνόλων δεδομένων για την δημιουργία Συστήματος Συστάσεων διαδρομών με βασικό κριτήριο σύστασης την πράσινη μετακίνηση.

Λέξεις Κλειδιά: Πολυτροπική Μεταφορά, Προεπεξεργασία Συνόλου Δεδομένων, Βελτιστοποίηση Μεταφοράς, Μηχανική Μάθηση, Συστήματα Συστάσεων, Περιγραφική Στατιστική, Οπτικοποίηση Δεδομένων

Abstract

The purpose of this thesis was initially to describe the problems and challenges in modern transportation in large cities in an effort to make them more efficient and environmentally sustainable. Then, the basic tools and technologies of Data Science that are used and developed in this effort were mentioned.

Subsequently, preprocessing, visualization, and analysis of the results were conducted using these tools on a wide range of multimodal transportation datasets in various cities in Europe and America. The primary city for comparison for all modes of transportation was New York City, and research was conducted to include the widest possible range of transportation modes in the analysis.

This methodology can serve as a guide for the preprocessing and descriptive and diagnostic analysis of similar data sets. It also constitutes the first steps before predictive analysis and the application of Machine Learning and Artificial Intelligence techniques for predicting various travel parameters for each mode of transportation. Next steps for development include the combination of analysis and data sets to create a route Recommendation System with the main recommendation criterion being “green” transportation.

Keywords: Multimodal Transport, Dataset Preprocessing, Transport Optimization, Machine Learning, Recommendation Systems, Descriptive Statistics, Data Visualization

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω την κ. Θεοδώρα Βαρβαρίγου, Καθηγήτρια ΕΜΠ και επιβλέπουσα της παρούσας εργασίας που με εμπιστεύτηκε με την ανάθεση της συγκεκριμένης διπλωματικής εργασίας και μου έδωσε την ευκαιρία να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα.

Επίσης θα ήθελα να ευχαριστήσω τον κ. Ευθύμιο Χονδρογιάννη για τον χρόνο που αφιέρωσε και την καθοδήγηση που μου παρείχε κατά την εκπόνηση αυτής της εργασίας.

Τέλος θέλω να ευχαριστήσω τους γονείς μου, Βασίλη και Μαρία για την αμέριστη στήριξη τους και την ελευθερία που μου έδωσαν να κυνηγήσω τα όνειρα μου. Τον Νικηφόρο που μου έδειξε πως να αφήνεις τα προσωπικά προβλήματα να χαθούν στην σκιά της ακαταμάχητης δίψας σου για ζωή. Την Μυρτώ που με στήριξε όσο κανένας κατά την διάρκεια των σπουδών μου και όχι μόνο και που (παρ)ακολούθησε από την αρχή έως το τέλος. Χωρίς την ανεκτίμητη βοήθεια τους δεν θα είχα φτάσει έως εδώ.

Πίνακας περιεχομένων

1 Εισαγωγή 1
1.1 Εισαγωγή στο αντικείμενο της Διπλωματικής Εργασίας 3
1.2 Οργάνωση κειμένου 4
2 Θεωρητικό υπόβαθρο 7
2.1 Εισαγωγή στις Μετακινήσεις και τα Μέσα Μαζικής Μεταφοράς 9
2.1.1 Σύντομη Ιστορική Αναδρομή των Μέσων Μαζικής Μεταφοράς 9
2.1.2 Σημερινή Κατάσταση και Προκλήσεις 10
2.1.3 Πολυτροπική Μεταφορά (Multimodal Transport) 12
2.2 Μηχανική Μάθηση και Ανάλυση Δεδομένων 13
2.2.1 Βασικές Έννοιες της Μηχανικής Μάθησης 14
2.2.2 Ανάλυση Δεδομένων 16
2.2.3 Εργαλεία και Τεχνικές Ανάλυσης Δεδομένων 19
2.3 Εφαρμογές Μηχανικής Μάθησης στις Μετακινήσεις 22
2.3.1 Πρόβλεψη Ζήτησης 22
2.3.2 Βελτιστοποίηση Διαδρομών 23
2.3.3 Προγνωστική Συντήρηση Οχημάτων και Υποδομών 23
2.3.4 Πρόβλεψη Κυκλοφορίας 24
2.3.5 Προσωποποιημένες Υπηρεσίες Επιβατών 24
2.4 Recommendation Systems 25
2.4.1 Προτάσεις Διαδρομών 25
2.4.2 Ενημερώσεις σε Πραγματικά Χρόνο 26
2.4.3 Προτάσεις για Εναλλακτικούς Τρόπους Μετακίνησης 27
2.5 Τεχνητή Νοημοσύνη (AI) 28
2.5.1 Αλγόριθμοι Τεχνητής Νοημοσύνης 28
2.5.2 Εφαρμογές Τεχνητής Νοημοσύνης στις Μεταφορές 29
2.5.3 Προκλήσεις και Μέλλον της Τεχνητής Νοημοσύνης στις Μεταφορές 30
3 Μεθοδολογία Ανάλυσης 33
3.1 Περιγραφή των Συνόλων Δεδομένων και των Χαρακτηριστικών τους 35
3.1.1 Περιγραφή Dataset Μισθωμένα οχήματα – Ταξί 36
3.1.2 Περιγραφή Dataset Ενοικιαζόμενα ποδήλατα 41
3.1.3 Περιγραφή Dataset Μέσα Μαζικής Μεταφοράς 45
3.2 Μεθοδολογία Επεξεργασίας και Ανάλυσης Δεδομένων 48
3.2.1 Βήμα 1ο – Διάβασμα του συνόλου δεδομένων (dataset) και εξέταση χαρακτηριστικών (features) 49
3.2.2 Βήμα 2ο – Έλεγχος μηδενικών και ακραίων (outliers) τιμών 52

3.2.3 Βήμα 3ο – Exploratory Data Analysis – Descriptive Statistics 55
3.2.4 Βήμα 4ο – Δημιουργία νέων features 56
3.2.5 Βήμα 5ο – Εξαγωγή Γραφημάτων και Οπτικοποίηση Δεδομένων 58
3.2.6 Βήμα 6ο – Εξαγωγή Πίνακα Συσχετίσεων και Θερμικού Χάρτη (heatmap) 60
3.2.7 Βήμα 7ο – Κωδικοποίηση Αριθμητικών και Κατηγορηματικών Χαρακτηριστικών (Features) 60
3.2.8 Βήμα 8ο – Αλγόριθμος FP-Growth για εύρεση μοτίβων 65
3.3 Εργαλεία που χρησιμοποιήθηκαν 68

4 Εφαρμογή και Αποτελέσματα 71
4.1 Περιγραφική Στατιστική (Descriptive Statistics) 72
4.2 Οπτικοποίηση Αποτελεσμάτων 80
4.2.1 Μισθωμένα οχήματα – Taxi 80
4.2.2 Ενοικιαζόμενα Ποδήλατα 90
4.2.3 Μέσα Μαζικής Μεταφοράς 96
4.3 Συσχέτιση και Συχνά Μοτίβα 102
4.3.1 Μισθωμένα οχήματα – Taxi 102
4.3.2 Ενοικιαζόμενα Ποδήλατα 106
4.3.3 Μέσα Μαζικής Μεταφοράς 110

5 Επίλογος 113
5.1 Σύνοψη και συμπεράσματα 115
5.2 Μελλοντικές επεκτάσεις 116

6 Βιβλιογραφία 119

Ευρετήριο εικόνων

Εικόνα 1 Αξία εναντίον Δυσκολίας στην Ανάλυση Δεδομένων..... 18
Εικόνα 2 Μεθοδολογία και Βήματα Επεξεργασίας συνόλων δεδομένων..... 49
Εικόνα 3 Ιστόγραμμα της διάρκειας του ταξιδιού με ταξί σε λεπτά..... 80
Εικόνα 4 Ιστόγραμμα της απόστασης του ταξιδιού με ταξί σε μίλια..... 81
Εικόνα 5 Ιστόγραμμα του συνολικού κόστους με ταξί σε δολάρια..... 81
Εικόνα 6 Κοινά ιστόγραμμα των 3 χαρακτηριστικών μεταξύ Σικάγο και Νέα Υόρκης..... 82
Εικόνα 7 Συχνότητες ταξιδιών ανά ώρα με ταξί..... 83
Εικόνα 8 Συχνότητες δεδομένων ανά εταιρεία ταξί στο Σικάγο..... 83
Εικόνα 9 Συχνότητα τρόπου πληρωμής για ταξί στο Σικάγο..... 84
Εικόνα 10 Οι δέκα δημοφιλέστερες γειτονιές του ταξιδιού με ταξί στο Σικάγο..... 85
Εικόνα 11 Συχνότητες δεδομένων ανά εταιρεία ταξί στην Νέα Υόρκη..... 85
Εικόνα 12 Συχνότητες τρόπου πληρωμής για ταξί στην Νέα Υόρκη..... 86
Εικόνα 13 Συχνότητες επιβατών για ταξί στην Νέα Υόρκη..... 86
Εικόνα 14 Συχνότητες ταξιδιών Uber ανά ημέρα και μήνα..... 87
Εικόνα 15 Διαδρομές ανά διαμέρισμα της Νέας Υόρκης..... 87
Εικόνα 16 Θερμικός Χάρτης διαδρομών Uber στο Manhattan ανά ώρα..... 88
Εικόνα 17 Θερμικός χάρτης συχνότητας διαδρομών ανά ώρα και ημέρα του μήνα..... 89
Εικόνα 18 Συχνότητα ταξιδιών Uber ανά ώρα και μήνα..... 89
Εικόνα 19 Ιστόγραμμα της διάρκεια ταξιδιού με ποδήλατο σε λεπτά..... 90
Εικόνα 20 Ιστόγραμμα της ηλικίας του χρήστη με ποδήλατο στην Νέα Υόρκη..... 91
Εικόνα 21 Διάγραμμα συχνοτήτων χαρ. χρήστη με ποδήλατο στην Νέα Υόρκη..... 92
Εικόνα 22 Διάγραμμα συχνοτήτων σταθμών ποδηλάτων της Νέας Υόρκης..... 92
Εικόνα 23 Διάγραμμα συχνοτήτων σταθμών ποδηλάτων του Λονδίνου..... 93
Εικόνα 24 Ιστόγραμμα των υπόλοιπων χαρακτηριστικών για τα ποδήλατα του Ελσίνκι..... 94
Εικόνα 25 Δέκα πιο δημοφιλείς γειτονιές ποδηλάτου Ελσίνκι..... 94
Εικόνα 26 Διάγραμμα συχνοτήτων σταθμών ποδηλάτων του Ελσίνκι..... 95
Εικόνα 27 Είκοσι πιο δημοφιλείς γραμμές λεωφορείου στην Νέα Υόρκη..... 96
Εικόνα 28 Είκοσι πιο δημοφιλείς στάσεις λεωφορείου στην Νέα Υόρκη..... 96
Εικόνα 29 Διάγραμμα συχνοτήτων γεωγραφικών διαμερισμάτων στην Νέα Υόρκη..... 96
Εικόνα 30 Διάγραμμα συχνοτήτων απόστασης από την στάση Νέα Υόρκη..... 97
Εικόνα 31 Θερμικός Χάρτης θέσεων λεωφορείων στην Νέα Υόρκη ανά ώρα..... 98
Εικόνα 32 Είκοσι πιο δημοφιλείς γραμμές λεωφορείου στο Ρίο Ντε Τζανέιρο..... 98
Εικόνα 33 Θερμικός Χάρτης θέσεων λεωφορείων στο Ρίο Ντε Τζανέιρο ανά ώρα..... 99
Εικόνα 34 Ιστόγραμμα των εισόδων/εξόδων στο μετρό στην Νέα Υόρκη..... 100
Εικόνα 35 Διαγράμματα συχνοτήτων γεωγραφικών χαρακτηριστικών μετρό..... 100
Εικόνα 36 Διάγραμμα των δέκα πιο δημοφιλών στάσεων για κάθε ημέρα..... 101
Εικόνα 37 Πίνακας συσχέτισης χαρακτηριστικών για τα ταξί του Σικάγο..... 102

Εικόνα 38 Πίνακας συσχέτισης χαρακτηριστικών για τα ταξί της Νέας Υόρκης……..... 104
Εικόνα 39 Πίνακας συσχέτισης για τα ποδήλατα στην Νέα Υόρκη……... 106
Εικόνα 40 Πίνακας συσχέτισης για τα ποδήλατα στο Λονδίνο……..... 107
Εικόνα 41 Πίνακας συσχέτισης για τα ποδήλατα στο Ελσίνκι……... 108
Εικόνα 42 Πίνακας συσχέτισης για το μετρό της Νέας Υόρκης…… 111

Ευρετήριο πινάκων

Πίνακας 1 Χαρακτηριστικά του συνόλου δεδομένων ταξί του Σικάγο……..... 38
Πίνακας 2 Χαρακτηριστικά του συνόλου δεδομένων ταξί της Νέας Υόρκης…… 40
Πίνακας 3 Χαρακτηριστικά του συνόλου δεδομένων Uber της Νέας Υόρκης το 2014……. 41
Πίνακας 4 Χαρακτηριστικά του συνόλου δεδομένων Uber της Νέας Υόρκης το 2015……. 41
Πίνακας 5 Χαρακτηριστικά του συνόλου δεδομένων ποδηλάτων της Νέας Υόρκης…… 42
Πίνακας 6 Χαρακτηριστικά του συνόλου δεδομένων ποδηλάτων του Λονδίνου……..... 44
Πίνακας 7 Χαρακτηριστικά του συνόλου δεδομένων ποδηλάτων του Ελσίνκι…… 45
Πίνακας 8 Χαρακτηριστικά του συνόλου δεδομένων λεωφορείων της Νέας Υόρκης……... 46
Πίνακας 9 Χαρακτηριστικά του συνόλου δεδομένων λεωφορείων του Ρίο Ντε Τζανέιρο…… 47
Πίνακας 10 Χαρακτηριστικά του συνόλου δεδομένων μετρό της Νέας Υόρκης……... 48
Πίνακας 11 Χαρακτηριστικά συγκρίσιμα για την κατηγορία μισθωμένα οχήματα-ταξί……... 51
Πίνακας 12 Χαρακτηριστικά συγκρίσιμα για την κατηγορία ενοικιαζόμενα ποδήλατα…… 51
Πίνακας 13 Κατηγοριοποίηση χαρακτηριστικών για ταξί του Σικάγο……... 61
Πίνακας 14 Κατηγοριοποίηση χαρακτηριστικών για ταξί της Νέας Υόρκης……. 62
Πίνακας 15 Κατηγοριοποίηση χαρακτηριστικών για Uber της Νέας Υόρκης……... 62
Πίνακας 16 Κατηγοριοποίηση χαρακτηριστικών για ποδήλατα της Νέας Υόρκης…… 63
Πίνακας 17 Κατηγοριοποίηση χαρακτηριστικών για ποδήλατα του Λονδίνου……..... 63
Πίνακας 18 Κατηγοριοποίηση χαρακτηριστικών για ποδήλατα του Ελσίνκι……. 64
Πίνακας 19 Κατηγοριοποίηση χαρακτηριστικών για λεωφορεία της Νέας Υόρκης……..... 64
Πίνακας 20 Κατηγοριοποίηση χαρακτηριστικών για λεωφορεία του Ρίο Ντε Τζανέιρο……... 65
Πίνακας 21 Κατηγοριοποίηση χαρακτηριστικών για μετρό της Νέας Υόρκης……..... 65
Πίνακας 22 Χαρακτηριστικά για εύρεση μοτίβων στην κατηγορία μισθωμένα οχήματα-ταξί..... 67
Πίνακας 23 Χαρακτηριστικά για εύρεση μοτίβων στην κατηγορία ενοικιαζόμενα ποδήλατα..... 67
Πίνακας 24 Χαρακτηριστικά για εύρεση μοτίβων στην κατηγορία μέσα μαζικής μεταφοράς..... 68
Πίνακας 25 Αποτελέσματα στατιστικών για τα χαρακτηριστικά της κατηγορίας μισθωμένα οχήματα-ταξί……... 73
Πίνακας 26 Αποτελέσματα στατιστικών για τα χαρακτηριστικά της κατηγορίας ενοικιαζόμενα ποδήλατα……. 76
Πίνακας 27 Αποτελέσματα στατιστικών για τα χαρακτηριστικά της κατηγορίας μέσα μαζικής μεταφοράς……..... 78
Πίνακας 28 Αποτελέσματα FP-Growth για ταξί στο Σικάγο…….. 103
Πίνακας 29 Αποτελέσματα FP-Growth για ταξί στην Νέα Υόρκη……. 105
Πίνακας 30 Αποτελέσματα FP-Growth για Uber στην Νέα Υόρκη……... 105
Πίνακας 31 Αποτελέσματα FP-Growth για ποδήλατα στην Νέα Υόρκη……..... 107
Πίνακας 32 Αποτελέσματα FP-Growth για ποδήλατα στο Λονδίνο……... 108
Πίνακας 33 Αποτελέσματα FP-Growth για ποδήλατα στο Ελσίνκι……..... 109
Πίνακας 34 Αποτελέσματα FP-Growth για λεωφορεία στην Νέα Υόρκη……..... 110

Πίνακας 35 Αποτελέσματα FP-Growth για λεωφορεία στο Ρίο Ντε Τζανέιρο……..... 110
Πίνακας 36 Αποτελέσματα FP-Growth για το μετρό της Νέας Υόρκης…… 111

1

Εισαγωγή

1.1 Εισαγωγή στο αντικείμενο της Διπλωματικής Εργασίας

Η χρήση της Επιστήμης Δεδομένων και της Μηχανικής Μάθησης στον τομέα των δημόσιων αλλά και γενικότερα των εναλλακτικών μεταφορών αποτελεί ένα δυναμικό και συνεχώς εξελισσόμενο πεδίο έρευνας και εφαρμογής. Με την αυξανόμενη αστικοποίηση και την ανάγκη για βιώσιμες λύσεις μετακίνησης, η αξιοποίηση των σύγχρονων τεχνολογιών γίνεται όλο και πιο απαραίτητη. Η Επιστήμη Δεδομένων και η Μηχανική Μάθηση μπορούν να προσφέρουν σημαντικές βελτιώσεις στη διαχείριση και τον σχεδιασμό των συστημάτων μεταφορών, οδηγώντας σε μεγαλύτερη αποτελεσματικότητα και μειωμένες περιβαλλοντικές επιπτώσεις.

Σήμερα, τα συστήματα δημόσιων μεταφορών αντιμετωπίζουν διάφορα προβλήματα, όπως η συμφόρηση, η αναξιοπιστία και οι περιβαλλοντικές επιπτώσεις από τις εκπομπές ρύπων. Η κακή διαχείριση των δρομολογίων, η έλλειψη προβλέψεων ακριβείας για την κίνηση και η ανισοκατανομή των πόρων είναι μερικά από τα ζητήματα που καθιστούν τα δημόσια μέσα μεταφοράς μη ελκυστικά για τους πολίτες. Παράλληλα, η κλιματική αλλαγή και η ανάγκη για μείωση των εκπομπών άνθρακα ωθούν τις κυβερνήσεις και τις εταιρείες προς την υιοθέτηση "πράσινων" μεταφορικών λύσεων.

Η Επιστήμη Δεδομένων συμβάλλει στη βελτίωση της λειτουργίας των μεταφορών μέσω της ανάλυσης μεγάλου όγκου δεδομένων, όπως τα δεδομένα κυκλοφορίας, οι τάσεις μετακίνησης και οι συνήθειες των επιβατών. Η Μηχανική Μάθηση, από την άλλη, μπορεί να εφαρμοστεί για τη δημιουργία προηγμένων αλγορίθμων πρόβλεψης και βελτιστοποίησης, επιτρέποντας την καλύτερη κατανομή των πόρων και τη βελτίωση της αξιοπιστίας των υπηρεσιών. Μέσω της ανάλυσης και της αξιοποίησης των δεδομένων, μπορούν να αναπτυχθούν λύσεις που βελτιώνουν την αποδοτικότητα, μειώνουν τις εκπομπές ρύπων και προάγουν την βιωσιμότητα. Ωστόσο, για να επιτευχθούν αυτοί οι στόχοι, είναι απαραίτητη η συνεχής έρευνα, η επένδυση σε τεχνολογίες και η συνεργασία μεταξύ κυβερνήσεων, επιστημονικής κοινότητας και βιομηχανίας.

Η σημερινή κατάσταση στον χώρο των μεταφορών δείχνει ότι υπάρχει σημαντικό περιθώριο βελτίωσης. Παρόλο που έχουν γίνει βήματα προς την κατεύθυνση της ενσωμάτωσης των νέων τεχνολογιών, οι περισσότερες πόλεις ακόμα αντιμετωπίζουν

προκλήσεις στη διαχείριση της κυκλοφορίας και στην εφαρμογή βιώσιμων λύσεων. Τα έργα "έξυπνων πόλεων" και η ανάπτυξη ηλεκτρικών και αυτόνομων οχημάτων είναι παραδείγματα των προσπαθειών που γίνονται για να αντιμετωπιστούν τα προβλήματα αυτά, αλλά η πλήρης εφαρμογή τους απαιτεί χρόνο και συντονισμένη προσπάθεια.

Στο πλαίσιο της παρούσας εργασίας χρησιμοποιήσαμε ελεύθερα δεδομένα μετακίνησης διαφορετικών κοινόχρηστων μέσων μεταφοράς για διαφορετικές πόλεις της Ευρώπης και της Αμερικής με σκοπό την μελέτη του τρόπου χρήσης τους και την εξαγωγή συμπερασμάτων για τις ιδιαιτερότητες του κάθε μέσου, τις προτιμήσεις ως προς την χρήση τους και τις ομοιότητες ή διαφορές που παρατηρούνται από περιοχή σε περιοχή. Επίσης μέσω της ανάλυσης αυτών των δεδομένων, προσπαθήσαμε να εντοπίσουμε πιθανά μοτίβα στην μετακίνηση, που σχετίζονται με κοινωνικούς, οικονομικούς ή γεωγραφικούς παράγοντες. Τα αποτελέσματα της μελέτης μπορούν να συμβάλουν στην ανάπτυξη στρατηγικών βελτίωσης των υπηρεσιών αυτών και στην προώθηση βιώσιμων λύσεων μετακίνησης στις αστικές περιοχές.

1.2 Οργάνωση κειμένου

Η παρούσα διπλωματική εργασία διαρθρώνεται σε πέντε Κεφάλαια.

Στο Κεφάλαιο 1 κάναμε μία σύντομη εισαγωγή στο πρόβλημα της αποδοτικής μετακίνησης σε μεγάλες πόλεις και τον ρόλο της Επιστήμης Δεδομένων στην αντιμετώπιση αυτού.

Στο Κεφάλαιο 2 αναπτύσσουμε το θεωρητικό υπόβαθρο που είναι απαραίτητο για την κατανόηση της παρούσας εργασίας. Παρουσιάζουμε τις βασικές έννοιες και πληροφορίες σχετικά με την πολυτροπική μεταφορά και τις σημερινές προκλήσεις στις σύγχρονες μεταφορές. Ταυτόχρονα παρουσιάζουμε βασικές έννοιες της Επιστήμης Δεδομένων και Μηχανικής Μάθησης στο πλαίσιο του τομέα της μετακίνησης και των μεταφορών.

Στο Κεφάλαιο 3 παρουσιάζουμε την μεθοδολογία ανάλυσης που ακολουθήσαμε για τα σύνολα δεδομένων. Αναφέρουμε αναλυτικά πληροφορίες για τα δεδομένα που συλλέξαμε και τα βήματα και τεχνικές που ακολουθήσαμε για την επεξεργασία τους.

Στο Κεφάλαιο 4 παρουσιάζουμε τα αποτελέσματα της ανάλυσης δεδομένων που πραγματοποιήθηκε, συγκρίνουμε με τα αναμενόμενα και σχολιάζουμε επιμέρους σημεία ενδιαφέροντος.

Στο Κεφάλαιο 5 καταγράφουμε τα βασικά συμπεράσματα που προέκυψαν και οι μελλοντικές κατευθύνσεις στις οποίες θα μπορούσε να επεκταθεί η παρούσα εργασία.

2

Θεωρητικό υπόβαθρο

2.1 Εισαγωγή στις Μετακινήσεις και τα Μέσα Μαζικής Μεταφοράς

Οι μετακινήσεις αποτελούν βασικό στοιχείο της καθημερινής ζωής των ανθρώπων και της οικονομικής δραστηριότητας. Τα μέσα μαζικής μεταφοράς (ΜΜΜ) περιλαμβάνουν διάφορα είδη μεταφορικών μέσων όπως λεωφορεία, τρένα, μετρό, τραμ, και πλοία, που έχουν ως στόχο να μεταφέρουν μεγάλους αριθμούς επιβατών με ασφάλεια και αποτελεσματικότητα. Οι πολίτες εξαρτώνται από τα ΜΜΜ για τις καθημερινές τους μετακινήσεις από και προς εργασία, σχολείο, ψυχαγωγία και άλλες δραστηριότητες. Η βελτίωση της ποιότητας και της απόδοσης των ΜΜΜ μπορεί να έχει σημαντικά οφέλη στην ποιότητα ζωής, τη μείωση της κυκλοφοριακής συμφόρησης και τη μείωση των εκπομπών ρύπων.

2.1.1 Σύντομη Ιστορική Αναδρομή των Μέσων Μαζικής Μεταφοράς

Η ιστορία των μέσων μαζικής μεταφοράς ξεκινά από την αρχαιότητα, όπου οι άνθρωποι χρησιμοποιούσαν βασικά μέσα για τη μετακίνησή τους. Στην αρχαία Ελλάδα και Ρώμη, άμαξες και άλογα χρησιμοποιούνταν για τις μεγάλες αποστάσεις και τη μεταφορά εμπορευμάτων. Στην Κίνα, ο τροχός και το άλογο χρησιμοποιήθηκαν επίσης εκτενώς, ενώ τα ποτάμια αποτελούσαν σημαντικές οδούς μεταφοράς με τη χρήση πλοίων.

Κατά τη μεσαιωνική περίοδο, η χρήση των αμαξών εξελίχθηκε και βελτιώθηκε. Η μεταφορά εντός πόλεων γινόταν κυρίως με τα πόδια, ενώ οι άμαξες και τα άλογα χρησιμοποιούνταν για μεγαλύτερες αποστάσεις. Στην Αναγέννηση, η ανάπτυξη των δρόμων και των γεφυρών διευκόλυνε τη μετακίνηση, και η χρήση των αμαξών έγινε πιο διαδεδομένη.

Με την Βιομηχανική Επανάσταση, κατά τον 18ο και 19ο αιώνα, τα μέσα μαζικής μεταφοράς μεταβλήθηκαν ραγδαία. Οι σιδηρόδρομοι εισήχθησαν στην καθημερινότητα και επέτρεψαν τη γρήγορη και μαζική μεταφορά ανθρώπων και αγαθών. Η εμφάνιση των ατμοκίνητων πλοίων και, αργότερα, των πετρελαιοκίνητων πλοίων, βελτίωσε τις θαλάσσιες μεταφορές και “μίκρυνε” τον κόσμο. Οι πόλεις

άρχισαν να αναπτύσσουν δίκτυα δημόσιας συγκοινωνίας, όπως τα πρώτα λεωφορεία και τα τραμ.

Στον 20ό αιώνα, η εφεύρεση του αυτοκινήτου και του αεροπλάνου έφερε την επανάσταση στις μεταφορές. Οι αστικές συγκοινωνίες επεκτάθηκαν με την εισαγωγή των μετρό και των ηλεκτρικών τραμ, ενώ τα αεροπλάνα επέτρεψαν ταχύτερες διεθνείς μετακινήσεις. Σήμερα, οι τεχνολογικές εξελίξεις συνεχίζουν να μεταμορφώνουν τα μέσα μεταφοράς, με τα ηλεκτρικά αυτοκίνητα, τα αυτόνομα οχήματα και τις νέες μορφές δημόσιας συγκοινωνίας να αναπτύσσονται ραγδαία. Οι αστικές περιοχές επενδύουν σε βιώσιμες και φιλικές προς το περιβάλλον λύσεις μεταφοράς, όπως οι ποδηλατόδρομοι και οι ζώνες για πεζούς.

2.1.2 Σημερινή Κατάσταση και Προκλήσεις

Η αστική κινητικότητα σε μεγάλες πόλεις αποτελεί ένα από τα πιο σύνθετα και απαιτητικά ζητήματα της σύγχρονης αστικής διαχείρισης. Η αύξηση του πληθυσμού, η ανάπτυξη της οικονομίας και η ταχεία αστικοποίηση έχουν οδηγήσει σε σημαντικές προκλήσεις για τα συστήματα μεταφορών.

Οι μεγάλες πόλεις, παγκοσμίως αντιμετωπίζουν κοινά προβλήματα στις μεταφορές, όπως η συμφόρηση, η ατμοσφαιρική ρύπανση και η έλλειψη υποδομών. Η συμφόρηση είναι ένα από τα πιο κρίσιμα ζητήματα, καθώς επηρεάζει την καθημερινή ζωή των κατοίκων και μειώνει την αποδοτικότητα των μεταφορών. Ο αυξημένος αριθμός οχημάτων στους δρόμους οδηγεί σε συχνές καθυστερήσεις και αυξημένο χρόνο μετακινήσεων.

Η ατμοσφαιρική ρύπανση είναι ένα άλλο σοβαρό πρόβλημα, που συνδέεται άμεσα με την υπερβολική χρήση οχημάτων με κινητήρες εσωτερικής καύσης. Η ποιότητα του αέρα σε πολλές μεγάλες πόλεις είναι κάτω από τα επιτρεπτά όρια, προκαλώντας προβλήματα υγείας στους κατοίκους και επιβαρύνοντας το περιβάλλον. Επομένως η ανάγκη για μείωση των εκπομπών αερίου του θερμοκηπίου για την αντιμετώπιση της κλιματικής αλλαγής και της βελτίωσης της ποιότητας ζωής στον αστικό ιστό βελτιώνοντας την ποιότητα του αέρα είναι ένας από τους πιο κρίσιμους και άμεσους στόχους που έχουν τεθεί παγκοσμίως. Τα Ηνωμένα Έθνη έχουν θέσει ως οδηγία/στόχο μέχρι το 2030 να έχουν μειωθεί οι εκπομπές σε διοξειδίου του άνθρακα (CO_2) κατά 55% σε σχέση με τα επίπεδα του 1990, και μέχρι το 2050 οι καθαρές εκπομπές να είναι μηδενικές.

Η έλλειψη υποδομών, όπως επαρκές δίκτυο δημόσιων συγκοινωνιών και ποδηλατοδρόμων, εμποδίζει τη βιώσιμη ανάπτυξη των πόλεων. Οι επενδύσεις σε υποδομές συχνά δεν ακολουθούν τον ρυθμό αύξησης του πληθυσμού, οδηγώντας σε υπερφόρτωση των υπαρχόντων συστημάτων μεταφορών.

Η επίλυση των προβλημάτων αυτών απαιτεί ολοκληρωμένες και καινοτόμες προσεγγίσεις. Στην προσπάθεια αυτή, αντιμετωπίζουμε διάφορες προκλήσεις, όπως:

- Προώθηση της βιώσιμης κινητικότητας: Χρήση εναλλακτικών μέσων μεταφοράς, όπως ποδήλατα, ηλεκτρικά οχήματα και μέσα δημόσιας συγκοινωνίας για τη μείωση της συμφόρησης και της ρύπανσης.
- Ευφυή Συστήματα Μεταφορών (ITS): Η εφαρμογή τεχνολογιών για την καλύτερη διαχείριση της κυκλοφορίας, όπως τα Ευφυή Συστήματα Μεταφορών, μπορεί να βελτιώσει σημαντικά την αποδοτικότητα και την ασφάλεια των μεταφορών.
- Συμμετοχή των Πολιτών: Η ενεργή συμμετοχή των πολιτών στη λήψη αποφάσεων και στον σχεδιασμό των μεταφορών μπορεί να εξασφαλίσει ότι οι λύσεις είναι προσαρμοσμένες στις ανάγκες της κοινωνίας.
- Επενδύσεις σε Υποδομές: Απαιτούνται σημαντικές επενδύσεις για τη βελτίωση και την επέκταση των δικτύων δημόσιων συγκοινωνιών και άλλων υποδομών, προκειμένου να υποστηριχθεί η αυξανόμενη ζήτηση.

Επάνω στο πλαίσιο αυτό, το World Economic Forum (WEF) [10] έχει αναπτύξει κάποιες οδηγίες για να βοηθήσει τις πόλεις να αντιμετωπίσουν τις προκλήσεις τις αστικής κινητικότητας. Στην τελευταία έκδοση του οι οδηγίες αυτές χωρίζονται συνοπτικά σε οχτώ κατηγορίες που είναι οι εξής: κοινή χρήση δεδομένων μεταξύ πόλεων για βελτίωση των συστημάτων, χρήση του δημόσιου χώρου και επιπτώσεις στις υποδομές, ασφάλεια, ένταξη και ισότητα, δίκαιη δουλειά, κοινή κινητικότητα και ομαδοποίηση, καθαρή μετάβαση και ενσωμάτωση της πολυτροπικής (multimodal) λογικής. Οι κατευθυντήριες αυτές γραμμές αποτελούν ένα θεμελιώδες υπόβαθρο για την ανάπτυξη βιώσιμων, τεχνολογικά προηγμένων και πολιτοκεντρικών συστημάτων μεταφοράς.

2.1.3 Πολυτροπική Μεταφορά (Multimodal Transport)

Η ενσωμάτωση διαφορετικών τύπων μεταφορών, γνωστή και ως Πολυτροπική Μεταφορά (Multimodal Transport), προσφέρει μια ολιστική προσέγγιση για την αντιμετώπιση αυτών των προκλήσεων. Η πολυτροπική μεταφορά συνδυάζει διάφορους τρόπους μετακίνησης στην ίδια διαδρομή, όπως ποδήλατο, περπάτημα, αυτοκίνητο, και ΜΜΜ, προκειμένου να παρέχει ευέλικτες και βέλτιστες λύσεις για τους χρήστες. Σκοπός είναι να βελτιωθεί η αποδοτικότητα στην μετακίνηση και να μειωθεί η εξάρτηση από τα ιδιωτικά αυτοκίνητα, ειδικά στις μεγάλες πόλεις.

Η υιοθέτηση της πολυτροπικής μεταφοράς στον αστικό ιστό έχει πολλαπλά πλεονεκτήματα. Τα πιο βασικά είναι η μείωση της συμφόρησης και τα περιβαλλοντικά οφέλη. Η ενσωμάτωση διαφόρων μέσων μεταφοράς μπορεί να μειώσει την εξάρτηση από ιδιωτικά οχήματα με μηχανές εσωτερικής καύσης (αυτοκίνητα, μηχανές) μειώνοντας έτσι και την συμφόρηση στους δρόμους. Επίσης η χρήση μέσων μεταφοράς χαμηλών εκπομπών, όπως τα ποδήλατα και οι δημόσιες συγκοινωνίες, συμβάλλουν στη μείωση της ατμοσφαιρικής ρύπανσης και των εκπομπών διοξειδίου του άνθρακα. Ενδεικτικά αναφέρουμε ότι η μείωση του προσωπικού αποτυπώματος άνθρακα είναι 42% με την χρήση λεωφορείου και 73% με την χρήση τρένου αντί για αυτοκίνητο. Άλλο πλεονέκτημα είναι ότι οι πολυτροπικές λύσεις μπορούν να είναι πιο οικονομικές τόσο για τους χρήστες όσο και για τις πόλεις, καθώς μειώνουν την ανάγκη για δαπανηρές επενδύσεις σε υποδομές οδικής κυκλοφορίας και για να γίνει αποσυμφόρηση. Ένα τελευταίο βασικό πλεονέκτημα είναι ότι διευκολύνεται η πρόσβαση σε απομακρυσμένες περιοχές και βελτιώνεται η συνδεσιμότητα μεταξύ διαφόρων περιοχών της πόλης.

Εκτός όμως από πλεονεκτήματα, οι πόλεις αντιμετωπίζουν και προκλήσεις για την υιοθέτηση των πολυτροπικών μεταφορών. Πρώτον για να επιτευχθεί η ενσωμάτωση αυτής της νοοτροπίας στο δίκτυο μίας πόλης θα πρέπει να συντονιστούν διάφοροι φορείς μεταξύ τους και να ενσωματώσουν διάφορα συστήματα. Επίσης θα πρέπει να αναπτυχθούν οι απαραίτητες υποδομές για την υποστήριξη της πολυτροπικής μεταφοράς, όπως διασυνδεδεμένοι σταθμοί (πχ λεωφορείων και τρένων) και ποδηλατόδρομοι, οι οποίες είναι δαπανηρές και χρονοβόρες. Σημαντική επένδυση αλλά και τεχνογνωσία απαιτεί και η εφαρμογή καινοτόμων τεχνολογιών όπως τα Ευφυή Συστήματα Μεταφορών (ITS) [11] που αναφέραμε και παραπάνω. Τέλος, για να θεωρηθεί επιτυχημένη αυτή η μετάβαση, και οι επενδύσεις επάνω σε αυτή, θα

πρέπει το ποσοστό υιοθέτησης από τους πολίτες να είναι υψηλό, γεγονός που δεν συμβαίνει πολλές φορές καθώς οι οποιεσδήποτε αλλαγές στις συνήθειες μετακίνησης αντιμετωπίζονται με αντίσταση από το κοινό.

Ενδεικτικά αναφέρουμε τις παρακάτω πόλεις που έχουν υιοθετήσει με επιτυχία μοντέλα Πολυτροπικής Μεταφοράς:

- Κοπεγχάγη, Δανία: Έχει γίνει επένδυση σε εκτεταμένο δίκτυο ποδηλατοδρόμων και ενθαρρύνει τη χρήση ποδηλάτων σε συνδυασμό με τα μέσα μαζικής μεταφοράς.
- Άμστερνταμ, Ολλανδία: Ομοίως, το δίκτυο μετακίνησης ποδηλάτων είναι εκτενέστατο και το ποδήλατο αποτελεί το βασικό μέσω μετακίνησης στην πόλη.
- Σιγκαπούρη: Η πόλη-κράτος έχει ενσωματώσει ευφυή συστήματα μεταφορών για την διαχείριση της κυκλοφορίας και την προώθηση της χρήσης δημοσίων συγκοινωνιών.
- Τόκιο, Ιαπωνία: Το εκτεταμένο δίκτυο τρένων και λεωφορείων επιτρέπει την εύκολη μετάβαση μεταξύ διαφορετικών μέσων μεταφοράς, βελτιώνοντας την αποδοτικότητα των μετακινήσεων.

Συμπερασματικά, βλέπουμε ότι η πολυτροπική μεταφορά αποτελεί μία πολλά υποσχόμενη λύση για την αντιμετώπιση των προκλήσεων που αντιμετωπίζουν οι μεγάλες πόλεις στις μεταφορές, ωστόσο η επιτυχής εφαρμογή της απαιτεί συντονισμένες προσπάθειες, κατάλληλες υποδομές και την ευρεία αποδοχή από το κοινό.

2.2 Μηχανική Μάθηση και Ανάλυση Δεδομένων

Η Μηχανική Μάθηση (Machine Learning) αποτελεί έναν κλάδο της Τεχνητής Νοημοσύνης (Artificial Intelligence), ο οποίος επικεντρώνεται στην ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να μαθαίνουν από δεδομένα και να λαμβάνουν αποφάσεις ή να κάνουν προβλέψεις και να ανακαλύπτουν μοτίβα χωρίς να είναι ρητά προγραμματισμένοι για την εκτέλεση συγκεκριμένων εργασιών. Η ανάλυση δεδομένων και η μηχανική μάθηση παίζουν κρίσιμο ρόλο στην κατανόηση και βελτίωση των συστημάτων μεταφορών.

2.2.1 Βασικές Έννοιες της Μηχανικής Μάθησης

Η Μηχανική Μάθηση [12] ορίζεται ως η διαδικασία κατά την οποία τα υπολογιστικά συστήματα βελτιώνουν τις επιδόσεις τους σε μία εργασία μέσω της εκμάθησης και εμπειρίας. Ο βασικός στόχος της είναι να αναπτύξει μεθόδους που επιτρέπουν στους υπολογιστές να μάθουν από δεδομένα και να σχηματίσουν νέες προβλέψεις ή αποφάσεις βασισμένες σε αυτά τα δεδομένα.

Η Μηχανική Μάθηση μπορεί να διακριθεί σε διάφορες κατηγορίες, με τις κυριότερες να είναι οι εξής:

- **Επιβλεπόμενη Μάθηση (Supervised Learning)**: Σε αυτή την κατηγορία, ο αλγόριθμος μαθαίνει από ένα σύνολο δεδομένων που περιλαμβάνει τόσο τις εισόδους όσο και τις αντίστοιχες εξόδους. Στόχος είναι η ανάπτυξη ενός μοντέλου που μπορεί να προβλέψει την έξοδο για νέες εισόδους. Οι τεχνικές σε αυτή την κατηγορία έχουν εφαρμογή σε προβλήματα όπως η αναγνώριση φωνής, η ανάλυση κειμένου, και η ανίχνευση ανωμαλιών, στο πεδίο της Αναγνώρισης Προτύπων.

  Από τις πιο γνωστές τεχνικές είναι η Ταξινόμηση (classification) που είναι η διαδικασία κατηγοριοποίησης δεδομένων σε διαφορετικές κατηγορίες ή κλάσεις, που ακολουθεί την διαδικασία εκπαίδευσης σε ένα σύνολο δεδομένων με γνωστές ετικέτες για κάθε κατηγορία ή κλάση. Κατά τη διαδικασία αυτή, ο αλγόριθμος μάθησης εκπαιδεύεται να αναγνωρίζει τα χαρακτηριστικά των δεδομένων που διαφοροποιούν τις κλάσεις μεταξύ τους και στη συνέχεια εφαρμόζει αυτήν την κατανόηση σε νέα, μη επισημειωμένα δεδομένα προκειμένου να τα κατηγοριοποιήσει.

- **Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)**: Εδώ ο αλγόριθμος προσπαθεί να βρει υποκείμενα μοτίβα ή δομές στα δεδομένα χωρίς να υπάρχει συγκεκριμένη έξοδος που να χρησιμοποιείται ως οδηγός. Είναι αποδοτική μέθοδος σε περιπτώσεις που θέλουμε να εξερευνήσουμε δεδομένα, στην εύρεση προτύπων, στην ανίχνευση ανωμαλιών και την προεπεξεργασία δεδομένων για περαιτέρω ανάλυση.

  Μία από τις πιο χρήσιμες μεθόδους Unsupervised Learning είναι η Συσταδοποίηση (Clustering), όπου είναι η διαδικασία ομαδοποίησης δεδομένων σε συστάδες, ή "clusters", βάσει κοινών χαρακτηριστικών. Κατά

την διαδικασία αυτή, η μέθοδος αυτή επιδιώκει να βρει μοτίβα ή σχέσεις μεταξύ των δεδομένων χωρίς πρότερη γνώση για τις κατανομές των χαρακτηριστικών. Συγκεκριμένα, ελαχιστοποιεί το κριτήριο απόστασης μεταξύ των μελών της κάθε συστάδας, ενώ μεγιστοποιεί το κριτήριο απόστασης των συστάδων μεταξύ τους.

- **Ημι-Επιβλεπόμενη Μάθηση (Semi-Supervised Learning)**: Η τεχνική αυτή συνδυάζει στοιχεία από την Επιβλεπόμενη και Μη Επιβλεπόμενη μάθηση, χρησιμοποιώντας ένα μικρό όγκο επισημειωμένων δεδομένων και ένα μεγάλο όγκο μη επισημειωμένων δεδομένων.
- **Ενισχυτική Μάθηση (Reinforcement Learning)**: Ο αλγόριθμος, εδώ μαθαίνει αλληλοεπιδρώντας με το περιβάλλον μέσω δοκιμών και λαθών, επιβραβεύοντας τις σωστές ενέργειες και τιμωρώντας τις λανθασμένες. Η μέθοδος αυτή είναι ιδιαίτερα αποτελεσματική σε περιπτώσεις όπου πρέπει να βρούμε την βέλτιστη λύση χωρίς εποπτεία, όπως πχ. στην ρομποτική, στα παιχνίδια κτλ.

Παρακάτω αναφέρουμε κάποιες βασικές έννοιες και όρους της Μηχανικής Μάθησης:

- **Δεδομένα (Data)**: Το σύνολο της πληροφορίας που χρησιμοποιείται για την εκπαίδευση των μοντέλων. Μπορεί να περιλαμβάνει χαρακτηριστικά (features) και ετικέτες (labels).
- **Χαρακτηριστικά (Features)**: Οι ανεξάρτητες μεταβλητές που χρησιμοποιούνται για την πρόβλεψη της εξόδου. Κάθε χαρακτηριστικό αντιπροσωπεύει μία συγκεκριμένη ιδιότητα των δεδομένων.
- **Μοντέλο (Model)**: Η μαθηματική αναπαράσταση που παράγεται από τον αλγόριθμο της Μηχανικής Μάθησης μετά την εκπαίδευση. Το μοντέλο χρησιμοποιείται για την πρόβλεψη τιμής ή την κατηγοριοποίηση των δεδομένων.
- **Εκπαίδευση (Training)**: Η διαδικασία κατά την οποία ο αλγόριθμος Μηχανικής Μάθησης μαθαίνει από τα δεδομένα για να βελτιώσει τις επιδόσεις του. Περιλαμβάνει τη βελτιστοποίηση των παραμέτρων του μοντέλου με βάση τα δεδομένα εκπαίδευσης.
- **Αλγόριθμοι Μηχανικής Μάθησης (Machine Learning Algorithms)**: Τα σύνολα κανόνων και διαδικασιών που χρησιμοποιούνται για την εκπαίδευση

των μοντέλων. Κάθε αλγόριθμος έχει τα δικά του πλεονεκτήματα και μειονεκτήματα ανάλογα με το είδος των δεδομένων και την εφαρμογή.

Η Μηχανική Μάθηση έχει τεράστια σημασία στον σύγχρονο κόσμο, καθώς επιτρέπει στους υπολογιστές να μαθαίνουν και να βελτιώνονται από την εμπειρία χωρίς να χρειάζονται ρητή προγραμματιστική καθοδήγηση. Αυτή η τεχνολογία βρίσκεται στην καρδιά πολλών καινοτομιών και εφαρμογών, όπως οι μηχανές αναζήτησης, οι προσωπικοί ψηφιακοί βοηθοί, τα συστήματα συστάσεων, τα οχήματα αυτόνομης οδήγησης, και η ανάλυση μεγάλων δεδομένων (big data). Επιπλέον, παίζει κρίσιμο ρόλο στην έρευνα και την ανάπτυξη νέων τεχνολογιών, όπως η Τεχνητή Νοημοσύνη και τα Νευρωνικά Δίκτυα, επηρεάζοντας σημαντικά το μέλλον της τεχνολογίας και της κοινωνίας. Με τη συνεχή εξέλιξη και την αυξανόμενη υιοθέτησή της, η Μηχανική Μάθηση είναι κεντρικός πυλώνας στη διαμόρφωση του ψηφιακού μέλλοντος.

2.2.2 Ανάλυση Δεδομένων

Η ανάλυση δεδομένων αποτελεί ένα κρίσιμο στάδιο στη διαδικασία διαχείρισης των πληροφοριών και αφορά την επεξεργασία των δεδομένων με σκοπό την εξαγωγή χρήσιμης πληροφορίας. Η ανάλυση δεδομένων χωρίζεται σε τέσσερα βασικά στάδια/τεχνικές:

Περιγραφική Στατιστική (Descriptive Statistics)

Η Περιγραφική Στατιστική είναι η διαδικασία που περιλαμβάνει τη συλλογή, οργάνωση και περιγραφή των δεδομένων. Στόχος της είναι η περιγραφή των βασικών χαρακτηριστικών των δεδομένων μέσω συνοπτικών και σαφών μέτρων. Τα βασικά αυτά μέτρα είναι τα εξής:

- **Μέση Τιμή (Mean)**: Είναι ο αριθμητικός μέσος ενός δείγματος και αποτελεί ένα δείκτη κεντρικής τάσης.
- **Διάμεση Τιμή (Median)**: Ο μεσαίος αριθμός σε ένα διατεταγμένο σύνολο δεδομένων, που χωρίζει τα δεδομένα σε δύο ίσα μέρη.
- **Τυπική Απόκλιση (Standard Deviation)**: Ένα μέτρο διασποράς που δείχνει πόσο αποκλίνουν τα δεδομένα από τη μέση τιμή.

Η Περιγραφική Στατιστική είναι απαραίτητη για την κατανόηση της δομής των δεδομένων και την ανίχνευση των βασικών τάσεων και προτύπων. Δίνει μια πρώτη

εικόνα για τα δεδομένα, που είναι σημαντική για τη μετέπειτα ανάλυση και λήψη αποφάσεων.

Διαγνωστική Ανάλυση (Diagnostic Analytics)

Η Διαγνωστική Ανάλυση επικεντρώνεται στην εξερεύνηση σχέσεων αιτιότητας στα χαρακτηριστικά των δεδομένων. Μέσω αυτής της ανάλυσης, οι ερευνητές μπορούν να αναγνωρίσουν τις αιτίες πίσω από συγκεκριμένα φαινόμενα και να κατανοήσουν τις σχέσεις μεταξύ διαφόρων μεταβλητών. Οι κύριες τεχνικές που χρησιμοποιούνται περιλαμβάνουν:

- **Ανάλυση Συσχέτισης (Correlation Analysis)**: Εξετάζει τις σχέσεις μεταξύ δύο ή περισσοτέρων μεταβλητών και καθορίζει το βαθμό στο οποίο αυτές συνδέονται.
- **Ανάλυση Παλινδρόμησης (Regression Analysis)**: Χρησιμοποιείται για να εκτιμήσει τις σχέσεις μεταξύ μιας εξαρτημένης μεταβλητής και μίας ή περισσοτέρων ανεξαρτήτων μεταβλητών. Γίνεται πρόβλεψη της εξαρτημένης μεταβλητής βάσει των ανεξαρτήτων.

Η Διαγνωστική Ανάλυση παρέχει βαθύτερη κατανόηση των δεδομένων, βοηθώντας τους ερευνητές να αναγνωρίσουν τους παράγοντες που επηρεάζουν τα δεδομένα και να λάβουν πιο ενημερωμένες αποφάσεις.

Προγνωστική Ανάλυση (Predictive Analytics)

Η Προγνωστική Ανάλυση χρησιμοποιεί δεδομένα και μοντέλα για την πρόβλεψη μελλοντικών αποτελεσμάτων. Αυτή η ανάλυση βασίζεται στη χρήση προηγμένων αλγορίθμων και τεχνικών μηχανικής μάθησης για την αναγνώριση προτύπων και τάσεων στα δεδομένα. Κύριες τεχνικές περιλαμβάνουν:

- **Αλγόριθμοι Μηχανικής Μάθησης (Machine Learning Algorithms):** Χρησιμοποιούνται για την αναγνώριση προτύπων στα δεδομένα και την πρόβλεψη μελλοντικών αποτελεσμάτων.
- **Χρονοσειρές (Time Series Analysis):** Εστιάζει στην ανάλυση δεδομένων που έχουν συλλεχθεί διαχρονικά για την πρόβλεψη μελλοντικών τιμών.

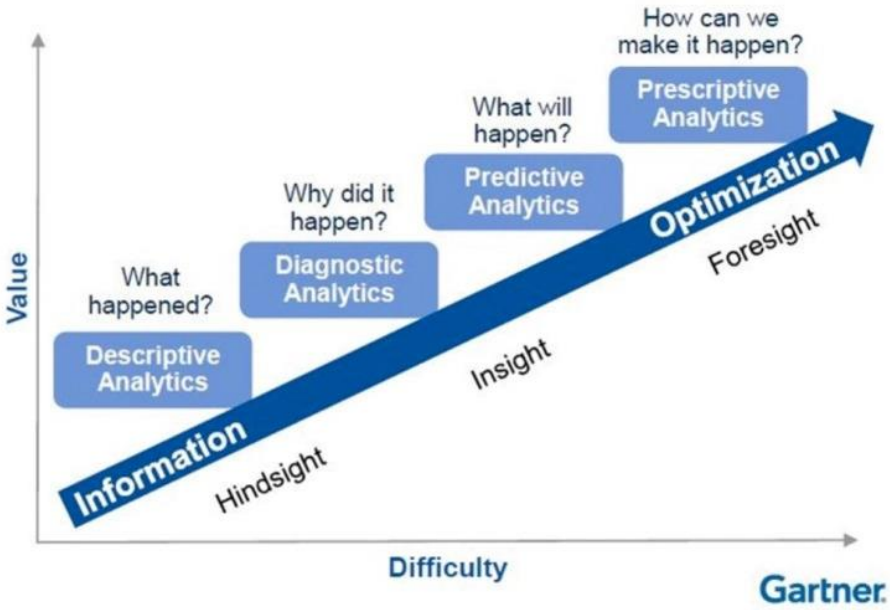
Η Προγνωστική Ανάλυση είναι ιδιαίτερα χρήσιμη για τη λήψη προληπτικών μέτρων και τη στρατηγική σχεδίαση, καθώς επιτρέπει την πρόβλεψη μελλοντικών τάσεων και γεγονότων.

Καθοδηγητική Ανάλυση (Prescriptive Analytics)

Η Καθοδηγητική Ανάλυση είναι η διαδικασία που προτείνει συγκεκριμένες ενέργειες με βάση τα δεδομένα και τις προβλέψεις. Συνδυάζει πληροφορίες από την Περιγραφική, Διαγνωστική και Προγνωστική ανάλυση για να παρέχει συστάσεις και να υποστηρίζει τη λήψη αποφάσεων. Οι κύριες τεχνικές περιλαμβάνουν:

- **Μοντέλα Βελτιστοποίησης (Optimization Models):** Γίνεται μεγιστοποίηση κριτηρίου στόχου ή ελαχιστοποίηση συνάρτησης κόστους για την εύρεση βέλτιστης λύσης σε πρόβλημα με συγκεκριμένους περιορισμούς.
- **Προσομοίωση (Simulation):** Χρησιμοποιείται για να εξετάσει διάφορα σενάρια και να εκτιμήσει τα αποτελέσματα τους, επιτρέποντας την διερεύνηση των πιθανών εκβάσεων και ανάλυση ρίσκου.

Η Καθοδηγητική Ανάλυση οδηγεί στη λήψη τεκμηριωμένων αποφάσεων με βάση τα δεδομένα, βελτιστοποιώντας τα κριτήρια απόδοσης και ελαχιστοποιώντας τους κινδύνους.



Εικόνα 1. Αξία εναντίον Δυσκολίας στην Ανάλυση Δεδομένων [21]

Στο πλαίσιο της Μηχανικής Μάθησης, το πρώτο στάδιο της Ανάλυσης Δεδομένων είναι η Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis – EDA) όπου εξετάζονται οι βασικές ιδιότητες και χαρακτηριστικά του συνόλου δεδομένων. Εδώ

χρησιμοποιούμε την Περιγραφική Στατιστική που αναφέραμε και παραπάνω. Επίσης κάνουμε οπτικοποίηση των δεδομένων κάνοντας χρήση γραφημάτων όπως διαγράμματα διασποράς (scatter plots), ιστογράμματα, θηκογράμματα (boxplots) και γραφήματα πυκνότητας (density plots) για την ανίχνευση τάσεων και ανωμαλιών στα δεδομένα. Επίσης σε αυτό το βήμα της ανάλυσης εντοπίζουμε τις τιμές που είναι μηδενικές και τις ακραίες ή έκτοπες τιμές (outliers) τις οποίες διαχειριζόμαστε κατάλληλα, διότι μπορούν να επηρεάσουν αρνητικά την επίδοση των μοντέλων και τα αποτελέσματα μας.

Το δεύτερο και πολύ σημαντικό στάδιο είναι η προεπεξεργασία των δεδομένων, η οποία είναι απαραίτητη για την προετοιμασία των δεδομένων σε μορφή κατάλληλη για να τροφοδοτηθούν στα μοντέλα Μηχανικής Μάθησης. Τα βασικά στάδια της προεπεξεργασίας είναι:

- **Καθαρισμός Δεδομένων**: Διαχείριση κενών τιμών μέσω τεχνικών όπως η απαλοιφή (deletion) ή η αντικατάσταση (μέσω απόδοσης) (imputation).
- **Μετασχηματισμοί Χαρακτηριστικών**: Κανονικοποίηση (normalization) ή τυποποίηση (standardization) των χαρακτηριστικών για να αντιμετωπιστούν οι διαφορές στις κλίμακες δεδομένων.
- **Δημιουργία Νέων Χαρακτηριστικών**: Κατασκευή νέων χαρακτηριστικών με χρήση τεχνικών όπως η Πολυωνυμική Επέκταση (polynomial expansion) ή η Ανάλυση Κυρίων Συνιστωσών (PCA)
- **Κωδικοποίηση κατηγορικών δεδομένων**: Μετασχηματισμός των κατηγορικών δεδομένων σε κωδικοποιημένη μορφή με χρήση τεχνικών όπως η one-hot encoding ή label encoding.

2.2.3 Εργαλεία και Τεχνικές Ανάλυσης Δεδομένων

Για την ανάλυση δεδομένων και την εφαρμογή τεχνικών μηχανικής μάθησης, χρησιμοποιούνται διάφορα εργαλεία που επιτρέπουν την αποθήκευση, την επεξεργασία και την ανάλυση μεγάλου όγκου δεδομένων. Αυτά τα εργαλεία και οι τεχνικές περιλαμβάνουν γλώσσες προγραμματισμού, βάσεις δεδομένων και πλατφόρμες για την επεξεργασία δεδομένων.

Python και R

Οι γλώσσες προγραμματισμού Python και R είναι ιδιαίτερα δημοφιλείς στην Ανάλυση Δεδομένων και τις εφαρμογές Μηχανικής Μάθησης λόγω της ευελιξίας και της ισχύος τους.

Η Python είναι μια ευρέως χρησιμοποιούμενη γλώσσα προγραμματισμού στην επιστήμη των δεδομένων λόγω της ευκολίας στη χρήση και της μεγάλης ποικιλίας βιβλιοθηκών της. Ορισμένες από τις πιο δημοφιλείς βιβλιοθήκες περιλαμβάνουν:

- Pandas: Χρησιμοποιείται για τον χειρισμό και την ανάλυση δεδομένων., καθώς περιλαμβάνει απλές συναρτήσεις για την εισαγωγή, εξαγωγή και μετατροπή δεδομένων.
- NumPy: Παρέχει υποστήριξη για μεγάλους, πολυδιάστατους πίνακες και συστοιχίες, καθώς και για μαθηματικές συναρτήσεις υψηλής απόδοσης.
- Scikit-learn: Είναι μια βιβλιοθήκη μηχανικής μάθησης που περιλαμβάνει μια ποικιλία αλγορίθμων για Ταξινόμηση, Παλινδρόμηση και Ομαδοποίηση.
- TensorFlow και Keras: Χρησιμοποιούνται για την κατασκευή και την εκπαίδευση νευρωνικών δικτύων, επιτρέποντας τη δημιουργία σύνθετων μοντέλων μηχανικής μάθησης.

Η R είναι μια γλώσσα προγραμματισμού και περιβάλλον λογισμικού για στατιστική ανάλυση και γραφήματα. Η R χρησιμοποιείται ευρέως στην ακαδημαϊκή έρευνα και τη βιομηχανία για τη στατιστική ανάλυση και την οπτικοποίηση δεδομένων. Ορισμένες από τις πιο χρήσιμες βιβλιοθήκες είναι:

- dplyr και tidyr: Παρέχουν εργαλεία για τη διαχείριση και τον μετασχηματισμό δεδομένων.
- ggplot2: Χρησιμοποιείται για τη δημιουργία απλών έως πολύπλοκων γραφημάτων.
- caret: Παρέχει ένα συνεκτικό πλαίσιο για τη δημιουργία, την εκπαίδευση και την αξιολόγηση μοντέλων μηχανικής μάθησης.

SQL και NoSQL Βάσεις Δεδομένων

Η αποθήκευση και η ανάκτηση δεδομένων είναι κρίσιμες για την ανάλυση δεδομένων και οι βάσεις δεδομένων έχουν κεντρικό ρόλο σε αυτές τις διαδικασίες.

Οι SQL (Structured Query Language) Βάσεις Δεδομένων χρησιμοποιούνται κυρίως για τη διαχείριση δομημένων δεδομένων που αποθηκεύονται σε πίνακες. Οι SQL βάσεις δεδομένων προσφέρουν δυναμικές δυνατότητες αναζήτησης και φιλτραρίσματος δεδομένων με χρήση ερωτημάτων. Παραδείγματα SQL βάσεων δεδομένων περιλαμβάνουν:

- MySQL και PostgreSQL: Είναι δύο από τις πιο δημοφιλείς ανοιχτού κώδικα SQL βάσεις δεδομένων, γνωστές για την απόδοση, την αξιοπιστία και την ευελιξία τους.
- Microsoft SQL Server και Oracle Database: Είναι εμπορικές λύσεις που προσφέρουν επιπλέον δυνατότητες και υποστήριξη για μεγάλες επιχειρήσεις.

Οι NoSQL (Not Only SQL) Βάσεις Δεδομένων σχεδιάστηκαν για την αποθήκευση και διαχείριση μη δομημένων δεδομένων, όπως έγγραφα, γραφήματα και δεδομένα κλειδιών-τιμών. Οι NoSQL βάσεις δεδομένων είναι ιδανικές για την αποθήκευση μεγάλου όγκου δεδομένων με δομή που ποικίλλει και για την κλιμάκωση σε μεγάλες εφαρμογές. Παραδείγματα NoSQL βάσεων δεδομένων περιλαμβάνουν:

- MongoDB: Μια βάση δεδομένων εγγράφων που αποθηκεύει δεδομένα σε μορφή εγγράφων όμοια με τα JSON έγγραφα.
- Cassandra: Μια κατανεμημένη βάση δεδομένων κλειδιών-τιμών που σχεδιάστηκε για να χειρίζεται μεγάλα σύνολα δεδομένων σε πολλές τοποθεσίες.

Hadoop και Spark

Για την επεξεργασία μεγάλου όγκου δεδομένων, χρησιμοποιούνται πλατφόρμες όπως το Hadoop και το Spark που επιτρέπουν την κατανεμημένη επεξεργασία δεδομένων.

Το Hadoop είναι ένα πλαίσιο ανοιχτού κώδικα που επιτρέπει την κατανεμημένη αποθήκευση και επεξεργασία μεγάλων συνόλων δεδομένων σε ομάδες υπολογιστών χρησιμοποιώντας απλά προγραμματιστικά μοντέλα. Το Hadoop αποτελείται από διάφορα στοιχεία:

- HDFS (Hadoop Distributed File System): Ένα κατανεμημένο σύστημα αρχείων που αποθηκεύει δεδομένα σε πολλούς κόμβους.
- MapReduce: Ένα προγραμματιστικό μοντέλο για την επεξεργασία μεγάλων συνόλων δεδομένων με κατανεμημένο τρόπο.

- YARN (Yet Another Resource Negotiator): Ένα πλαίσιο για τη διαχείριση των πόρων και των εφαρμογών σε ένα cluster.

Το Spark είναι μια πλατφόρμα επεξεργασίας μεγάλων δεδομένων που προσφέρει ταχύτητα και ευελιξία. Σε σύγκριση με το Hadoop, το Spark μπορεί να επεξεργαστεί δεδομένα πολύ πιο γρήγορα λόγω της δυνατότητάς του να εκτελεί επεξεργασία στη μνήμη. Το Spark υποστηρίζει διάφορες λειτουργίες όπως:

- Spark SQL: Χρησιμοποιείται για την επεξεργασία δομημένων δεδομένων με χρήση SQL.
- Spark Streaming: Επιτρέπει την επεξεργασία ροών δεδομένων σε πραγματικό χρόνο.
- MLlib: Μια βιβλιοθήκη μηχανικής μάθησης για το Spark.
- GraphX: Ένα API για την ανάλυση δεδομένων γραφημάτων.

2.3 Εφαρμογές Μηχανικής Μάθησης στις Μετακινήσεις

Η Μηχανική Μάθηση έχει εισέλθει δυναμικά στον τομέα της μετακίνησης, προσφέροντας πολυάριθμες δυνατότητες για τη βελτίωση της αποτελεσματικότητας και της εμπειρίας των επιβατών. Οι εφαρμογές της καλύπτουν ένα ευρύ φάσμα προβλημάτων από την πρόβλεψη ζήτησης μέχρι την προγνωστική συντήρηση, συμβάλλοντας στη δημιουργία ενός πιο αποδοτικού και αξιόπιστου συστήματος μεταφορών.

2.3.1 Πρόβλεψη Ζήτησης

Η πρόβλεψη της ζήτησης είναι μια κρίσιμη εφαρμογή της Μηχανικής Μάθησης που βοηθά τους οργανισμούς δημόσιων συγκοινωνιών να προγραμματίσουν και να βελτιστοποιήσουν τα δρομολόγια και τις υπηρεσίες τους [12][13][15]. Η Ανάλυση Χρονοσειρών (Time Series Analysis) είναι μια από τις τεχνικές που χρησιμοποιούνται για την πρόβλεψη της μελλοντικής ζήτησης. Αυτή η τεχνική περιλαμβάνει τη συλλογή και ανάλυση ιστορικών δεδομένων για την ανίχνευση προτύπων και τάσεων.

Για παράδειγμα, τα ιστορικά δεδομένα από τις κάρτες επιβίβασης, τα δεδομένα των αισθητήρων που μετρούν την κίνηση των επιβατών και οι δημογραφικές πληροφορίες μπορούν να αναλυθούν για να προβλέψουν την ζήτηση σε διαφορετικές ώρες της

ημέρας, ημέρες της εβδομάδας ή ακόμα και εποχές του έτους. Οι αλγόριθμοι Μηχανικής Μάθησης όπως τα Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM) και τα Αναδρομικά Νευρωνικά Δίκτυα (RNNs) είναι ιδιαίτερα αποτελεσματικοί στην ανάλυση τέτοιων δεδομένων και στην πρόβλεψη μελλοντικών αναγκών.

Μία πιο ακριβής πρόβλεψη της ζήτησης επιτρέπει στους φορείς μεταφορών να προσαρμόσουν τις υπηρεσίες τους αναλόγως, μειώνοντας την περιττή κίνηση των οχημάτων ΜΜΜ και βελτιώνοντας την εμπειρία των επιβατών μέσω της μείωσης του χρόνου αναμονής και της υπερφόρτωσης.

2.3.2 Βελτιστοποίηση Διαδρομών

Η βελτιστοποίηση των διαδρομών είναι ένα άλλο κρίσιμο πεδίο εφαρμογής της Μηχανικής Μάθησης στις μετακινήσεις. Οι Αλγόριθμοι Δρομολόγησης Οχημάτων (Vehicle Routing Algorithms) βοηθούν στον καθορισμό των βέλτιστων διαδρομών για τα οχήματα των ΜΜΜ, λαμβάνοντας υπόψη διάφορους παράγοντες όπως η ζήτηση, η κυκλοφοριακή συμφόρηση, και οι χρονικοί περιορισμοί.

Η χρήση αλγορίθμων όπως οι Γενετικοί Αλγόριθμοι (GA) [17], Βελτιστοποίηση με την μέθοδο Αποικίας Μυρμηγκιών (ACO) [18] και Βελτιστοποίηση Σμήνους Σωματιδίων (PSO) επιτρέπει την εύρεση των βέλτιστων διαδρομών με βάση τις δεδομένες συνθήκες. Οι αλγόριθμοι αυτοί μπορούν να βελτιώσουν σημαντικά την αποδοτικότητα των δρομολογίων, μειώνοντας το κόστος καυσίμων, τις εκπομπές ρύπων και τον χρόνο μετακίνησης.

Επιπλέον, με την ενσωμάτωση δεδομένων σε πραγματικό χρόνο από GPS και αισθητήρες κυκλοφορίας, οι αλγόριθμοι μπορούν να προσαρμόσουν τις διαδρομές δυναμικά για να αποφύγουν κυκλοφοριακή συμφόρηση και άλλες καθυστερήσεις. Αυτό έχει ως αποτέλεσμα την αύξηση της αξιοπιστίας των υπηρεσιών και την καλύτερη ικανοποίηση των επιβατών.

2.3.3 Προγνωστική Συντήρηση Οχημάτων και Υποδομών

Η προγνωστική συντήρηση αποτελεί μια εξαιρετικά σημαντική εφαρμογή της Μηχανικής Μάθησης, καθώς επιτρέπει την πρόβλεψη και την πρόληψη βλαβών στα οχήματα και τις υποδομές μεταφορών. Αυτή η τεχνολογία χρησιμοποιεί δεδομένα από αισθητήρες που παρακολουθούν την απόδοση και την κατάσταση των οχημάτων, καθώς και προηγμένους αλγόριθμους για την ανάλυση των δεδομένων αυτών.

Αλγόριθμοι όπως τα Νευρωνικά Δίκτυα και οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) είναι ιδιαίτερα αποτελεσματικοί στην ανίχνευση ανωμαλιών και στην πρόβλεψη βλαβών. Για παράδειγμα, τα Νευρωνικά Δίκτυα μπορούν να εκπαιδευτούν να αναγνωρίζουν μοτίβα στα δεδομένα αισθητήρων που υποδηλώνουν επικείμενες βλάβες, επιτρέποντας στους τεχνικούς να επεμβαίνουν προληπτικά πριν την εμφάνιση σοβαρών προβλημάτων.

Η προγνωστική συντήρηση μειώνει τον χρόνο παύσης λειτουργίας των οχημάτων, μειώνει το κόστος συντήρησης και βελτιώνει την αξιοπιστία των υπηρεσιών μεταφορών. Επίσης, συμβάλλει στη μείωση των ατυχημάτων που προκαλούνται από μηχανικές βλάβες, εξασφαλίζοντας έτσι μεγαλύτερη ασφάλεια για τους επιβάτες και τους οδηγούς.

2.3.4 Πρόβλεψη Κυκλοφορίας

Η πρόβλεψη κυκλοφορίας είναι μια εφαρμογή της Μηχανικής Μάθησης στον τομέα των μετακινήσεων, που χρησιμοποιεί ιστορικά δεδομένα και δεδομένα σε πραγματικό χρόνο για την πρόβλεψη των κυκλοφοριακών συνθηκών. Με τη χρήση αλγορίθμων όπως τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs) και τα Νευρωνικά Δίκτυα Γράφων (GNNs), μπορεί να γίνει ανάλυση μεγάλου όγκου δεδομένων από αισθητήρες κυκλοφορίας, κάμερες και άλλα μέσα παρακολούθησης.

Οι προβλέψεις αυτές βοηθούν στην καλύτερη διαχείριση της κυκλοφορίας, επιτρέποντας στους διαχειριστές να λαμβάνουν ενημερωμένες αποφάσεις για την ανακατεύθυνση της κυκλοφορίας και τη ρύθμιση των φωτεινών σηματοδοτών. Επιπλέον, μία σχετικά ακριβής πρόβλεψη κυκλοφορίας μέσω συστημάτων πλοήγησης επιτρέπει στους οδηγούς να επιλέξουν μία λιγότερο χρονοβόρα διαδρομή καθοδόν.

Η βελτιστοποίηση της διαχείρισης της κυκλοφορίας όχι μόνο μειώνει τις καθυστερήσεις και την κατανάλωση καυσίμων, αλλά επίσης μειώνει τις εκπομπές ρύπων, συμβάλλοντας σε ένα πιο βιώσιμο περιβάλλον.

2.3.5 Προσωποποιημένες Υπηρεσίες Επιβατών

Η Μηχανική Μάθηση μπορεί επίσης να χρησιμοποιηθεί για την παροχή προσωποποιημένων υπηρεσιών στους επιβάτες. Μέσω της ανάλυσης δεδομένων από τις μετακινήσεις των επιβατών, τις προτιμήσεις τους και τα μοτίβα συμπεριφοράς

τους, οι αλγόριθμοι μπορούν να προσαρμόσουν τις υπηρεσίες ώστε να ανταποκρίνονται καλύτερα στις ανάγκες των επιβατών.

Για παράδειγμα, τα Συστήματα Συστάσεων Διαδρομών (Route Recommendation Systems) μπορούν να προτείνουν τις καλύτερες διαδρομές με βάση τις προτιμήσεις των επιβατών, ενώ οι προσωποποιημένες ειδοποιήσεις μπορούν να ενημερώνουν τους επιβάτες για τυχόν αλλαγές στα δρομολόγια ή καθυστερήσεις που έχουν αντίκτυπο στην δική τους μετακίνηση. Τέτοια συστήματα ενισχύουν την ικανοποίηση των επιβατών και προάγουν την εμπιστοσύνη στις δημόσιες συγκοινωνίες.

Συμπεραίνουμε λοιπόν ότι οι τεχνικές της Μηχανικές Μάθησης βρίσκουν εφαρμογή σε σημαντικά προβλήματα στον τομέα της μετακίνησης, παρέχοντας λύσεις που βελτιώνουν την αποδοτικότητα, την αξιοπιστία και την εμπειρία των επιβατών. Με τη συνεχή πρόοδο της τεχνολογίας και την αυξανόμενη διαθεσιμότητα δεδομένων, οι δυνατότητες της Μηχανικής Μάθησης στον τομέα των μεταφορών αναμένεται να επεκταθούν ακόμη περισσότερο, οδηγώντας σε μια νέα εποχή έξυπνων και πιο αποδοτικών συστημάτων μεταφορών.

2.4 Συστήματα Συστάσεων (Recommendation Systems)

Τα Συστήματα Συστάσεων (Recommendation Systems) τα οποία αναφέραμε στο τέλος του ανωτέρω υποκεφαλαίου χρησιμοποιούνται σε διάφορους τομείς της τεχνολογίας και της πληροφορικής, παρέχοντας εξατομικευμένες προτάσεις στους χρήστες. Αυτά τα συστήματα χρησιμοποιούν αλγόριθμους και δεδομένα για να προσφέρουν προτάσεις που βελτιώνουν την εμπειρία του χρήστη, αυξάνουν την ικανοποίησή του και προωθούν τη χρήση συγκεκριμένων υπηρεσιών ή προϊόντων. Εφαρμόζονται ευρέως σε τομείς όπως το ηλεκτρονικό εμπόριο, η ψυχαγωγία, οι πλατφόρμες κοινωνικής δικτύωσης καθώς και στον τομέα των μεταφορών. Εδώ θα εστιάσουμε στην χρήση τους στον τομέα της μετακίνησης και των μεταφορών [19][20].

2.4.1 Προτάσεις Διαδρομών

Τα Συστήματα Συστάσεων Διαδρομών στοχεύουν στην πρόταση βέλτιστων διαδρομών στους επιβάτες, αξιοποιώντας τις δύο βασικές προσεγγίσεις και τεχνικές

των συστημάτων συστάσεων, το συνεργατικό φιλτράρισμα (collaborative filtering) και το φιλτράρισμα βάση περιεχομένου (content-based filtering).

Το συνεργατικό φιλτράρισμα βασίζεται στη συλλογή και ανάλυση δεδομένων από την αλληλεπίδραση των χρηστών με το σύστημα. Συγκρίνει τις προτιμήσεις των χρηστών και βρίσκει κοινά μοτίβα συμπεριφοράς για να προτείνει διαδρομές που έχουν χρησιμοποιήσει όμοιοι χρήστες. Για παράδειγμα, αν ένας χρήστης έχει επιλέξει συγκεκριμένες διαδρομές στο παρελθόν, το σύστημα μπορεί να προτείνει αυτές τις διαδρομές σε άλλους χρήστες με παρόμοιες προτιμήσεις.

Το φιλτράρισμα βάση περιεχομένου εστιάζει στις ιδιότητες των διαδρομών και των προτιμήσεων του χρήστη. Χρησιμοποιεί δεδομένα όπως η διάρκεια της διαδρομής, ο αριθμός των στάσεων, και άλλες πληροφορίες για να δημιουργήσει ένα προφίλ για τον χρήστη και να προτείνει διαδρομές που ταιριάζουν με αυτό το προφίλ.

Οι προτάσεις διαδρομών όχι μόνο βοηθούν τους επιβάτες να βρουν την πιο γρήγορη ή βολική διαδρομή, αλλά μπορούν επίσης να προσαρμοστούν στις προσωπικές προτιμήσεις του χρήστη, όπως η αποφυγή πολυσύχναστων περιοχών ή η επιλογή διαδρομών με συγκεκριμένα αξιοθέατα.

2.4.2 Ενημερώσεις σε Πραγματικό Χρόνο

Όπως αναφέραμε και παραπάνω στο πρόβλημα της Πρόβλεψης Κυκλοφορίας οι ενημερώσεις σε πραγματικό χρόνο είναι κρίσιμες για την αποτελεσματική λειτουργία των συστημάτων μεταφορών. Χρησιμοποιούνται δεδομένα από αισθητήρες και δίκτυα για να παρέχονται έγκαιρα πληροφορίες στους επιβάτες σχετικά με αλλαγές στις διαδρομές, καθυστερήσεις, και άλλες κρίσιμες παραμέτρους. Οι βασικές πτυχές αυτής της προσέγγισης είναι:

Τεχνολογία και Δεδομένα: Οι αισθητήρες τοποθετούνται σε διάφορα σημεία, όπως στάσεις λεωφορείων, σιδηροδρομικούς σταθμούς, και μέσα μεταφοράς, καταγράφοντας δεδομένα που αφορούν την ταχύτητα, τη θέση, και την πληρότητα των οχημάτων. Αυτά τα δεδομένα μεταδίδονται σε κεντρικά συστήματα διαχείρισης όπου γίνεται ανάλυση και παρέχεται πληροφορία στους χρήστες μέσω εφαρμογών ή οθονών σε πραγματικό χρόνο.

Προβλέψεις και Ανάλυση: Η ανάλυση δεδομένων σε πραγματικό χρόνο μπορεί να διευκολύνει την πρόβλεψη καθυστερήσεων και την ανίχνευση συμβάντων πριν γίνουν

αντιληπτά από τους επιβάτες. Για παράδειγμα, εάν ένα λεωφορείο καθυστερεί λόγω κυκλοφοριακής συμφόρησης, το σύστημα μπορεί να ενημερώσει τους επιβάτες και να προτείνει εναλλακτικές διαδρομές ή μέσα μεταφοράς.

Εμπειρία Χρήστη: Οι ενημερώσεις σε πραγματικό χρόνο βελτιώνουν την εμπειρία του χρήστη, προσφέροντας πληροφορίες που μειώνουν το άγχος και την αβεβαιότητα σχετικά με την αναμονή και τη μετακίνηση. Επιπλέον, οι ενημερώσεις αυτές μπορούν να προσαρμοστούν στις προσωπικές προτιμήσεις του χρήστη, όπως η επιλογή της πιο γρήγορης διαδρομής ή η αποφυγή στάσεων με υψηλή πληρότητα.

2.4.3 Προτάσεις για Εναλλακτικούς Τρόπους Μετακίνησης

Τα Συστήματα Συστάσεων (Recommendation Systems) δεν περιορίζονται μόνο στις προτάσεις διαδρομών, αλλά μπορούν επίσης να χρησιμοποιηθούν για την πρόταση εναλλακτικών τρόπων μετακίνησης, όπως ποδήλατο ή περπάτημα, λαμβάνοντας υπόψη τις καιρικές συνθήκες και άλλους παράγοντες. Εδώ οι βασικές πτυχές είναι:

Αλγόριθμοι και Δεδομένα: Οι αλγόριθμοι των συστημάτων αυτών αξιολογούν δεδομένα όπως οι καιρικές συνθήκες, η κατάσταση των υποδομών, και οι προσωπικές προτιμήσεις του χρήστη για να προτείνουν εναλλακτικούς τρόπους μετακίνησης. Για παράδειγμα, σε μια ηλιόλουστη μέρα, το σύστημα μπορεί να προτείνει το περπάτημα ή το ποδήλατο αντί για τη χρήση μέσων μαζικής μεταφοράς.

Οφέλη για την Υγεία και το Περιβάλλον: Η προώθηση εναλλακτικών τρόπων μετακίνησης μπορεί να έχει θετικές επιπτώσεις στην υγεία των χρηστών, καθώς προάγεται η σωματική άσκηση. Επίσης, μειώνεται η περιβαλλοντική επιβάρυνση από την υπερβολική χρήση των μηχανοκίνητων μέσων μεταφοράς, συμβάλλοντας στη μείωση των εκπομπών αερίων του θερμοκηπίου και στην προστασία του περιβάλλοντος.

Προσαρμογή στις Ανάγκες του Χρήστη: Οι προτάσεις μπορούν να προσαρμοστούν στις ατομικές ανάγκες του χρήστη, λαμβάνοντας υπόψη παράγοντες όπως η απόσταση, η διαθεσιμότητα με βάση την τοποθεσία του χρήστη, και οι προσωπικές προτιμήσεις για σωματική δραστηριότητα. Έτσι, το σύστημα μπορεί να προτείνει την καλύτερη δυνατή επιλογή για κάθε χρήστη ξεχωριστά.

Συμπερασματικά, τα Συστήματα Συστάσεων (Recommendation Systems) έχουν αναδειχθεί ως ένα ισχυρό εργαλείο για τη βελτίωση της εμπειρίας του χρήστη στις

μεταφορές , καθώς παρέχουν εξατομικευμένες προτάσεις διαδρομών, ενημερώσεις σε πραγματικό χρόνο καθώς και προτάσεις για εναλλακτικούς τρόπους μετακίνησης. Αυτές οι τεχνολογίες όχι μόνο διευκολύνουν την καθημερινή μετακίνηση, αλλά συμβάλλουν επίσης στην υγεία των χρηστών και στην προστασία του περιβάλλοντος. Με την περαιτέρω ανάπτυξη και εξέλιξη των συστημάτων αυτών, οι δυνατότητες για βελτίωση της ποιότητας ζωής των πολιτών και της απόδοσης των συστημάτων μεταφορών θα συνεχίσουν να αυξάνονται.

2.5 Τεχνητή Νοημοσύνη (Artificial Intelligence)

Η Τεχνητή Νοημοσύνη (Artificial Intelligence) είναι ένα πεδίο της επιστήμης των υπολογιστών που επικεντρώνεται στην ανάπτυξη συστημάτων ικανών να εκτελούν εργασίες που απαιτούν ανθρώπινη αντίληψη, όπως η αναγνώριση ομιλίας, η όραση υπολογιστών, η λήψη αποφάσεων και η μετάφραση γλωσσών. Οι τεχνολογίες της Τεχνητής Νοημοσύνης έχουν σημειώσει τεράστια πρόοδο τα τελευταία χρόνια, αλλάζοντας ριζικά πολλούς τομείς της καθημερινής ζωής και των βιομηχανιών.

2.5.1 Αλγόριθμοι Τεχνητής Νοημοσύνης

Η Τεχνητή Νοημοσύνη περιλαμβάνει διάφορους αλγόριθμους που συμβάλλουν στην ανάπτυξη ευφυών συστημάτων. Οι κυριότεροι αλγόριθμοι είναι οι εξής:

Νευρωνικά Δίκτυα και Βαθιά Μάθηση: Τα Νευρωνικά Δίκτυα είναι εμπνευσμένα από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου. Η Βαθιά Μάθηση, μια υποκατηγορία της Μηχανικής Μάθησης, χρησιμοποιεί νευρωνικά δίκτυα με πολλαπλά επίπεδα για την επίλυση πολύπλοκων προβλημάτων. Τα Νευρωνικά Δίκτυα μαθαίνουν από μεγάλο όγκο δεδομένων να αναγνωρίζουν μοτίβα και έχουν εφαρμογές σε διάφορους τομείς, όπως η Αναγνώριση Εικόνας και η Επεξεργασία Φυσικής Γλώσσας.

Αλγόριθμοι Βελτιστοποίησης: Αυτοί οι αλγόριθμοι χρησιμοποιούνται για την εύρεση βέλτιστων λύσεων σε προβλήματα με πολλούς περιορισμούς και στόχους. Είναι βασικοί για την επίλυση προβλημάτων όπως η δρομολόγηση οχημάτων, ο σχεδιασμός διαδρομών και η διαχείριση πόρων.

Αλγόριθμοι Ενισχυτικής Μάθησης: Οι αλγόριθμοι Ενισχυτικής Μάθησης μαθαίνουν μέσω της αλληλεπίδρασης με το περιβάλλον τους, βελτιστοποιώντας τις δράσεις τους

βάσει των ανταμοιβών που λαμβάνουν. Αυτοί οι αλγόριθμοι είναι ιδιαίτερα χρήσιμοι για την ανάπτυξη αυτόνομων συστημάτων, όπως τα ρομπότ και τα αυτόνομα οχήματα.

2.5.2 Εφαρμογές Τεχνητής Νοημοσύνης στις Μεταφορές

Η Τεχνητή Νοημοσύνη έχει βρει πολλές εφαρμογές στις μεταφορές [14], προσφέροντας καινοτόμες λύσεις και βελτιώσεις στην αποδοτικότητα, την ασφάλεια και την εμπειρία των χρηστών. Οι πιο σημαντικές και δημοφιλείς είναι οι παρακάτω:

Αυτόνομα Οχήματα: Τα αυτόνομα οχήματα είναι ίσως η πιο γνωστή εφαρμογή της Τεχνητής Νοημοσύνης στις μεταφορές. Χρησιμοποιούν διάφορους αλγόριθμους και τεχνολογίες για την πλοήγηση και τη λήψη αποφάσεων σε πραγματικό χρόνο. Οι αισθητήρες, όπως οι κάμερες, οι αισθητήρες Light Detection and Ranging (LiDAR) και τα ραντάρ, συλλέγουν δεδομένα από το περιβάλλον, τα οποία επεξεργάζονται τα νευρωνικά δίκτυα για να αναγνωρίσουν αντικείμενα, να προβλέψουν κινήσεις και να σχεδιάσουν ασφαλείς διαδρομές. Οι αλγόριθμοι ενισχυτικής μάθησης επιτρέπουν στα οχήματα να βελτιώνουν τις επιδόσεις τους με την πάροδο του χρόνου.

Διαχείριση Κυκλοφορίας: Η Τεχνητή Νοημοσύνη χρησιμοποιείται για τη διαχείριση και τη βελτιστοποίηση της ροής της κυκλοφορίας. Τα συστήματα διαχείρισης κυκλοφορίας χρησιμοποιούν δεδομένα σε πραγματικό χρόνο από αισθητήρες και κάμερες για να παρακολουθούν τις συνθήκες κυκλοφορίας και να προσαρμόζουν τα σήματα των φαναριών και τους περιορισμούς ταχύτητας. Αυτή η δυναμική διαχείριση μπορεί να μειώσει τη συμφόρηση, να βελτιώσει την ασφάλεια και να μειώσει τις εκπομπές ρύπων. Ένα παράδειγμα είναι τα έξυπνα συστήματα σηματοδότησης που προσαρμόζονται στις συνθήκες κυκλοφορίας για να μειώσουν τους χρόνους αναμονής.

Ανάλυση Συμπεριφοράς Επιβατών: Η Τεχνητή Νοημοσύνη χρησιμοποιείται για την ανάλυση των προτύπων συμπεριφοράς των επιβατών. Αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν για την παροχή εξατομικευμένων υπηρεσιών και τη βελτίωση της εμπειρίας των επιβατών. Για παράδειγμα, οι αεροπορικές εταιρείες μπορούν να χρησιμοποιούν τεχνικές Μηχανικής Μάθησης για να προβλέπουν τις προτιμήσεις των επιβατών και να προσφέρουν προσωποποιημένες προτάσεις. Οι εταιρείες μέσων μαζικής μεταφοράς μπορούν να αναλύουν τα δεδομένα επιβατών για να βελτιστοποιήσουν τα δρομολόγια και να μειώσουν τους χρόνους αναμονής.

Συστήματα Υποστήριξης Οδηγού (ADAS): Τα συστήματα υποστήριξης οδηγού χρησιμοποιούν τεχνητή νοημοσύνη για να παρέχουν στον οδηγό ειδοποιήσεις και βοήθεια κατά την οδήγηση. Παραδείγματα περιλαμβάνουν το αυτόματο φρενάρισμα, τη διατήρηση λωρίδας και την προσαρμοστική ταχύτητα. Αυτά τα συστήματα χρησιμοποιούν κάμερες και αισθητήρες για να παρακολουθούν το περιβάλλον και να εντοπίζουν κινδύνους, βελτιώνοντας έτσι την ασφάλεια. Η τεχνολογία αυτή εισέρχεται σε όλες τις κατηγορίες οχημάτων σιγά σιγά και έχει αρχίσει να γίνεται ιδιαίτερα προσιτή.

Προγνωστική Συντήρηση: Η Τεχνητή Νοημοσύνη επιτρέπει την προγνωστική συντήρηση οχημάτων και υποδομών μεταφοράς όπως αναφέραμε και παραπάνω. Οι αλγόριθμοι Μηχανικής Μάθησης χρησιμοποιούν δεδομένα από αισθητήρες για να προβλέψουν πότε ένα όχημα ή μια υποδομή θα χρειαστεί συντήρηση, προλαμβάνοντας έτσι τις βλάβες και μειώνοντας το κόστος.

Λειτουργία και Διαχείριση Στόλου: Οι εταιρείες μεταφορών χρησιμοποιούν την Τεχνητή Νοημοσύνη για τη διαχείριση των στόλων οχημάτων τους. Η ανάλυση δεδομένων σε πραγματικό χρόνο επιτρέπει την παρακολούθηση της απόδοσης των οχημάτων, τη βελτιστοποίηση των δρομολογίων και τη μείωση της κατανάλωσης καυσίμου. Οι αλγόριθμοι τεχνητής νοημοσύνης βοηθούν επίσης στη διαχείριση της ζήτησης και της προσφοράς, εξασφαλίζοντας ότι τα οχήματα είναι διαθέσιμα όταν και όπου χρειάζονται.

2.5.3 Προκλήσεις και Μέλλον της Τεχνητής Νοημοσύνης στις Μεταφορές

Εκτός από τα πολλά οφέλη που προσφέρουν οι εφαρμογές τις Τεχνητής Νοημοσύνης στην καθημερινότητα και ειδικότερα στον χώρο των μετακινήσεων υπάρχουν και διάφορες προκλήσεις που πρέπει να αντιμετωπιστούν για μία πιο καθολική ενσωμάτωση και υιοθέτηση αυτών των τεχνολογιών. Οι βασικότερες είναι:

Ασφάλεια και Εμπιστοσύνη: Η ανάπτυξη ασφαλών και αξιόπιστων συστημάτων τεχνητής νοημοσύνης είναι κρίσιμη. Τα αυτόνομα οχήματα πρέπει να είναι ικανά να λαμβάνουν αποφάσεις σε απρόβλεπτες συνθήκες και να διαχειρίζονται σωστά τους κινδύνους. Η ανάπτυξη εμπιστοσύνης από των χρήστη είναι επίσης σημαντική, καθώς πολλοί άνθρωποι είναι διστακτικοί στη χρήση αυτόνομων οχημάτων.

Ηθικά και Νομικά Ζητήματα: Η Τεχνητή Νοημοσύνη φέρνει στο προσκήνιο πολλά ηθικά και νομικά ζητήματα. Ποιος είναι υπεύθυνος σε περίπτωση ατυχήματος με

αυτόνομο όχημα; Πώς προστατεύονται τα δεδομένα των επιβατών; Αυτά είναι ερωτήματα που πρέπει να απαντηθούν καθώς η χρησιμοποιούμενη τεχνολογία εξελίσσεται.

Ενσωμάτωση με Υφιστάμενα Συστήματα: Η ενσωμάτωση της Τεχνητής Νοημοσύνης με τα υπάρχοντα συστήματα μεταφορών μπορεί να είναι ιδιαίτερα δύσκολη, κυρίως λόγω παλαιότητας των υφιστάμενων συστημάτων και τεχνολογιών. Τα νέα συστήματα πρέπει να είναι συμβατά με τις υπάρχουσες υποδομές και να μπορούν να συνεργάζονται με άλλες τεχνολογίες.

Εκπαίδευση και Ανάπτυξη Δεξιοτήτων: Η ανάπτυξη και η διαχείριση συστημάτων Τεχνητής Νοημοσύνης απαιτεί εξειδικευμένες γνώσεις και δεξιότητες. Η προετοιμασία του ανθρώπινου δυναμικού για την επιτυχή εφαρμογή των τεχνολογιών αυτών περιλαμβάνει διάφορες πτυχές. Τα πανεπιστήμια και τα εκπαιδευτικά ιδρύματα πρέπει να προσαρμόσουν τα προγράμματα σπουδών τους, ενώ είναι αναγκαία η συνεχής επαγγελματική κατάρτιση και η διεπιστημονική συνεργασία.

3

Μεθοδολογία Ανάλυσης

3.1 Περιγραφή των Συνόλων Δεδομένων και των Χαρακτηριστικών τους

Στο πλαίσιο της παρούσας εργασίας έγινε εκτεταμένη έρευνα για σύνολα δεδομένων (datasets) που να αφορούν τους πολυτροπικούς (multimodal) τρόπους μεταφοράς. Οι τρόποι μεταφοράς που μελετήθηκαν ήταν ταξί, λεωφορεία, τραμ, μετρό, ποδήλατα, περπάτημα. Σκοπός μας ήταν η αναζήτηση συνόλων δεδομένων που να έχουν τα εξής χαρακτηριστικά:

- Να αφορούν μεγάλο αστικό ιστό – πόλη όπου να υπάρχει πληθώρα διαφορετικών τρόπων μετακίνησης.
- Να διαθέτουν πληροφορίες για πάνω από ένα τρόπο μεταφοράς για το ίδιο χρονικό διάστημα.
- Ο αριθμός των χαρακτηριστικών (features) να είναι όσο το δυνατόν μεγαλύτερος, ώστε να περιέχουν δεδομένα τοποθεσίας αλλά και αρκετή πληροφορία για τον χρήστη και το ταξίδι του.
- Το μέγεθος του συνόλου δεδομένων (dataset) να είναι επαρκές τόσο από πλευράς όγκου δεδομένων όσο και του χρονικού διαστήματος που καλύπτει ώστε να μπορούμε να εξάγουμε αξιόπιστα συμπεράσματα.
- Να είναι σε μορφή κατάλληλη για επεξεργασία πχ. αρχείο csv.

Η αναζήτηση έγινε σε διάφορους ιστότοπους που είναι είτε αποθετήρια συνόλων δεδομένων για ελεύθερη χρήση και επεξεργασία (πχ. kaggle.com) είτε από δημόσια αποθετήρια των εκάστοτε πόλεων για διάφορα δεδομένα σχετικά με αυτές (πχ. NYC Open data).

Καταλήξαμε σε εννέα σύνολα δεδομένων μετακίνησης που κατηγοριοποιούνται με βάση την τοποθεσία και τον τρόπο μετακίνησης ως εξής:

1. Μισθωμένα οχήματα – Ταξί
    - Taxi Database – Chicago
    - Taxi Database– New York City
    - Uber Database – New York City

2. Ενοικιαζόμενα ποδήλατα
    - Bike Database – New York City
    - Bike Database – London
    - Bike Database – Helsinki
3. Μέσα Μαζικής Μεταφοράς
    - Bus Database – New York City
    - Bus Database – Rio De Janeiro
    - Metro Database – New York City

3.1.1 Περιγραφή Dataset Μισθωμένα οχήματα – Ταξί

Taxi Database – Chicago

Το σύνολο δεδομένων Taxi Trips Chicago 2024 [9] παρέχει ολοκληρωμένες πληροφορίες για τα ταξίδια με ταξί που πραγματοποιήθηκαν στην πόλη του Σικάγο κατά τη διάρκεια του έτους 2024, πιο συγκεκριμένα για το χρονικό διάστημα από 01-01-2024 έως 01-03-2024. Περιλαμβάνει μια λεπτομερή καταγραφή διαφόρων χαρακτηριστικών που σχετίζονται με κάθε ταξίδι, προσφέροντας πολύτιμες πληροφορίες για ανάλυση μεταφοράς, πολεοδομικό σχεδιασμό και επιχειρηματική ευφυΐα. Το dataset αποθηκεύεται σε ένα αρχείο CSV μεγέθους 340 mb που περιέχει περίπου 865 χιλιάδες έγγραφες. Αναλυτικά τα χαρακτηριστικά (features) του είναι:

Όνομα	**Περιγραφή**
Trip ID	Ένας μοναδικός αναγνωριστικός αριθμός για κάθε ταξίδι, διευκολύνοντας την εύκολη παρακολούθηση και ανάλυση.
Taxi ID	Ένας αναγνωριστικός αριθμός για το ταξί που συμμετέχει σε κάθε ταξίδι, επιτρέποντας την παρακολούθηση των μεμονωμένων οχημάτων.
Trip Start Timestamp	Ημερομηνία και Ώρα, περιγράφει πότε ξεκίνησε το ταξίδι, στρογγυλοποιημένο στο πλησιέστερο τέταρτο της ώρας.
Trip End Timestamp	Ημερομηνία και Ώρα, περιγράφει πότε τελείωσε το ταξίδι, στρογγυλοποιημένο στο πλησιέστερο τέταρτο της ώρας.
Trip Seconds	Διάρκεια του ταξιδιού σε δευτερόλεπτα.

Trip Miles	Απόσταση του ταξιδιού σε μίλια.
Pickup Census Tract	Η περιοχή απογραφής όπου ξεκίνησε το ταξίδι. Για λόγους ιδιωτικότητας, αυτή η περιοχή απογραφής δεν εμφανίζεται για ορισμένα ταξίδια. Αυτή η στήλη συχνά θα είναι κενή για τοποθεσίες εκτός του Σικάγο.
Dropoff Census Tract	Η περιοχή απογραφής όπου τελείωσε το ταξίδι. Για λόγους ιδιωτικότητας, αυτή η περιοχή απογραφής δεν εμφανίζεται για ορισμένα ταξίδια. Αυτή η στήλη συχνά θα είναι κενή για τοποθεσίες εκτός του Σικάγο.
Pickup Community Area	Η περιοχή κοινότητας όπου ξεκίνησε το ταξίδι. Αυτή η στήλη θα είναι κενή για τοποθεσίες εκτός του Σικάγο.
Dropoff Community Area	Η περιοχή κοινότητας όπου τελείωσε το ταξίδι. Αυτή η στήλη θα είναι κενή για τοποθεσίες εκτός του Σικάγο.
Fare	Ο ναύλος για το ταξίδι σε δολάρια.
Tips	Το φιλοδώρημα για το ταξίδι. Τα φιλοδωρήματα με μετρητά γενικά δεν θα καταγράφονται. Οι τιμές είναι σε δολάρια.
Tolls	Τα διόδια για το ταξίδι. Οι τιμές είναι σε δολάρια.
Extras	Επιπλέον χρεώσεις για το ταξίδι. Οι τιμές είναι σε δολάρια.
Trip Total	Το συνολικό κόστος του ταξιδιού. Το συνολικό άθροισμα των προηγούμενων χαρακτηριστικών. Οι τιμές είναι σε δολάρια.
Payment Type	Τύπος πληρωμής για το ταξίδι.
Company	Η εταιρεία ταξί.
Pickup Centroid Latitude	Το γεωγραφικό πλάτος του κέντρου της περιοχής απογραφής παραλαβής ή της περιοχής κοινότητας όπου ξεκίνησε το ταξίδι, αν η περιοχή απογραφής έχει κρυφτεί για λόγους ιδιωτικότητας. Αυτή η στήλη συχνά θα είναι κενή για τοποθεσίες εκτός του Σικάγο.
Pickup Centroid Longitude	Το γεωγραφικό μήκος του κέντρου της περιοχής απογραφής παραλαβής ή της περιοχής κοινότητας όπου ξεκίνησε το ταξίδι, αν η περιοχή απογραφής έχει κρυφτεί για λόγους ιδιωτικότητας. Αυτή η στήλη συχνά θα είναι κενή για τοποθεσίες εκτός του Σικάγο.
Pickup Centroid	Η τοποθεσία του κέντρου της περιοχής απογραφής παραλαβής ή της

Location	περιοχής κοινότητας όπου ξεκίνησε το ταξίδι, αν η περιοχή απογραφής έχει κρυφτεί για λόγους ιδιωτικότητας. Αυτή η στήλη συχνά θα είναι κενή για τοποθεσίες εκτός του Σικάγο.
Dropoff Centroid Latitude	Το γεωγραφικό πλάτος του κέντρου της περιοχής απογραφής αποβίβασης ή της περιοχής κοινότητας όπου τελείωσε το ταξίδι, αν η περιοχή απογραφής έχει κρυφτεί για λόγους ιδιωτικότητας. Αυτή η στήλη συχνά θα είναι κενή για τοποθεσίες εκτός του Σικάγο.
Dropoff Centroid Longitude	Το γεωγραφικό μήκος του κέντρου της περιοχής απογραφής αποβίβασης ή της περιοχής κοινότητας όπου τελείωσε το ταξίδι, αν η περιοχή απογραφής έχει κρυφτεί για λόγους ιδιωτικότητας. Αυτή η στήλη συχνά θα είναι κενή για τοποθεσίες εκτός του Σικάγο.
Dropoff Centroid Location	Η τοποθεσία του κέντρου της περιοχής απογραφής αποβίβασης ή της περιοχής κοινότητας όπου τελείωσε το ταξίδι, αν η περιοχή απογραφής έχει κρυφτεί για λόγους ιδιωτικότητας. Αυτή η στήλη συχνά θα είναι κενή για τοποθεσίες εκτός του Σικάγο.

Πίνακας 1. Χαρακτηριστικά του συνόλου δεδομένων ταξί του Σικάγο.

Taxi Database – New York City

Το σύνολο δεδομένων Taxi trip date NYC [3] περιέχει δεδομένα για τα ταξίδια με ταξί που πραγματοποιήθηκαν στον δήμο της Νέας Υόρκης, όπως αυτά έχουν συγκεντρωθεί από δύο εταιρίες ταξί, για διάστημα τρεισήμισι εβδομάδων από 14/12/20 έως και 08/01/21. Και σε αυτό το database όπως και στο παραπάνω έχουμε χαρακτηριστικά (features) που αφορούν το ταξίδι, όπως κόστος, τοποθεσία και διάρκεια διαδρομής. Εδώ έχουμε μικρότερη όγκο δεδομένων, γύρω στις 84.000 εγγραφές και το αρχείο μας έχει μέγεθος 7 mb. Τα χαρακτηριστικά αναλυτικά είναι:

Όνομα	**Περιγραφή**
Vendor ID	Ένας κωδικός που υποδεικνύει τον πάροχο ταξί που παρείχε την εγγραφή. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.
tpep_pickup_datetime	Η ημερομηνία και η ώρα που ενεργοποιήθηκε ο μετρητής.

tpep_dropoff_datetime	Η ημερομηνία και η ώρα που απενεργοποιήθηκε ο μετρητής.
passenger_count	Ο αριθμός των επιβατών στο όχημα.
trip_distance	Η απόσταση του ταξιδιού σε μίλια που αναφέρεται στο ταξίμετρο.
PULocationID	Η ζώνη ταξί Taxi and Limousine Commission (TLC) στην οποία ενεργοποιήθηκε ο ταξιμετρητής.
DOLocationID	Η ζώνη ταξί Taxi and Limousine Commission (TLC) στην οποία απενεργοποιήθηκε ο ταξιμετρητής.
RateCodeID	Ο τελικός κωδικός τιμολόγησης που ίσχυε στο τέλος του ταξιδιού. 1= Τυπική τιμή 2= JFK 3= Newark 4= Nassau ή Westchester 5= Διαπραγματευόμενος ναύλος 6= Ομαδική μεταφορά
Store_and_fwd_flag	Μία σημαία (μεταβλητή flag) που μας δείχνει αν η εγγραφή του ταξιδιού κρατήθηκε στη μνήμη του οχήματος πριν αποσταλεί στον πάροχο, γνωστή ως "αποθήκευση και αποστολή", επειδή το όχημα δεν είχε σύνδεση με τον διακομιστή. Y= ταξίδι αποθήκευσης και αποστολής N= όχι ταξίδι αποθήκευσης και αποστολής
Payment_type	Ένας αριθμητικός κωδικός που υποδηλώνει τον τρόπο πληρωμής του επιβάτη για το ταξίδι. 1= Πιστωτική κάρτα 2= Μετρητά 3= Χωρίς χρέωση 4= Διαφωνία 5= Άγνωστο 6= Ακυρωμένο ταξίδι
Fare_amount	Ο ναύλος που υπολογίζεται από τον μετρητή βάσει του χρόνου και της απόστασης. Ο τιμές είναι σε δολάρια.

Extra	Διάφορα επιπλέον και πρόσθετες χρεώσεις. Επί του παρόντος, αυτό περιλαμβάνει μόνο τις χρεώσεις $0.50 και $1 για ώρες αιχμής και κατά τη διάρκεια της νύχτας.
MTA_tax	$0.50 φόρος Metropolitan Transportation Authority (MTA) που ενεργοποιείται αυτόματα βάσει του τιμολογημένου ναύλου που χρησιμοποιείται.
Improvement_surcharge	$0.30 χρέωση βελτίωσης που επιβάλλεται στα ταξίδια κατά την έναρξη του ναύλου. Η χρέωση βελτίωσης άρχισε να επιβάλλεται το 2015.
Tip_amount	Ποσό φιλοδωρήματος – Αυτή η στήλη συμπληρώνεται αυτόματα για τα φιλοδωρήματα με πιστωτική κάρτα. Τα φιλοδωρήματα με μετρητά δεν περιλαμβάνονται. Οι τιμές είναι σε δολάρια.
Tolls_amount	Το συνολικό ποσό όλων των διοδίων που πληρώθηκαν στο ταξίδι. Οι τιμές είναι σε δολάρια.
Total_amount	Το συνολικό ποσό που χρεώθηκε στους επιβάτες. Δεν περιλαμβάνει τα φιλοδωρήματα με μετρητά. Οι τιμές είναι σε δολάρια.
Congestion_surcharge	Το συνολικό ποσό που χρεώθηκε στο ταξίδι για το τέλος συμφόρησης της Νέας Υόρκης. Οι τιμές είναι σε δολάρια.
Airport_fee	$1.25 για παραλαβή μόνο στα αεροδρόμια LaGuardia και John F. Kennedy.

Πίνακας 2. Χαρακτηριστικά του συνόλου δεδομένων ταξί της Νέας Υόρκης.

Uber Database – New York City

Το σύνολο δεδομένων Uber Pickups [2] περιέχει δεδομένα ταξιδιού Uber στην περιοχή της Νέας Υόρκης, αποτελείται από 20 εκατομμύρια εγγραφές και αποθηκεύεται σε αρχεία χωρητικότητας 200 mb συνολικά. Το σύνολο δεδομένων αποτελείται από δύο υποσύνολα με βάση τον παρακάτω διαχωρισμό:

- Δεδομένα ταξιδιού Uber για το χρονικό διάστημα Απρίλιος 2014 – Σεπτέμβριος 2014, διαχωρισμένα ανά μήνα με ακριβείς πληροφορίες τοποθεσίας. Τα χαρακτηριστικά (features) είναι:

Όνομα	Περιγραφή
Date/Time	Ημερομηνία και ώρα παραλαβής
Lat	Γεωγραφικό πλάτος ξεκινήματος ταξιδιού
Long	Γεωγραφικό μήκος ξεκινήματος ταξιδιού
Base	Βασικός εταιρικός κωδικός TLC που σχετίζεται με την παραλαβή Uber

Πίνακας 3. Χαρακτηριστικά του συνόλου δεδομένων Uber της Νέας Υόρκης το 2014.

- Δεδομένα ταξιδιού Uber για το χρονικό διάστημα Ιανουάριος 2015 – Ιούνιος 2015, με λιγότερο ακριβείς πληροφορίες τοποθεσίας. Τα χαρακτηριστικά (features) είναι:

Όνομα	Περιγραφή
Dispatching base num	Ο βασικός κωδικός εταιρείας TLC της βάσης που απέστειλε το Uber
Pickup date	Ημερομηνία και ώρα παραλαβής
Affiliated base num	Ο βασικός κωδικός εταιρείας TLC που σχετίζεται με την παραλαβή Uber
location ID	Το αναγνωριστικό τοποθεσίας παραλαβής που σχετίζεται με την παραλαβή Uber

Πίνακας 4. Χαρακτηριστικά του συνόλου δεδομένων Uber της Νέας Υόρκης το 2015.

3.1.2 Περιγραφή Dataset Ενοικιαζόμενα ποδήλατα

Bike Database – New York City

Το σύνολο δεδομένων Bike Trips [6] περιέχει δεδομένα ταξιδιού με ποδήλατο που ανήκει σε εταιρεία κοινής χρήσης ποδηλάτων στην Νέα Υόρκη, για το χρονικό διάστημα από 01/05/2018 έως και 31/05/2018. Αποτελείται από 1.6 εκατομμύρια εγγραφές, 11 χαρακτηριστικά (features) και αποθηκεύεται σε αρχείο χωρητικότητας 180 mb συνολικά. Η υπηρεσία κοινόχρηστων ποδηλάτων είναι μια υπηρεσία κοινής μεταφοράς κατά την οποία τα ποδήλατα διατίθενται για κοινή χρήση σε ιδιώτες σε

βραχυπρόθεσμη βάση για μια συγκεκριμένη τιμή ή δωρεάν. Πολλά συστήματα κοινής χρήσης ποδηλάτων επιτρέπουν στους ανθρώπους να δανειστούν ένα ποδήλατο από έναν σταθμό και να το επιστρέψουν σε άλλο σταθμό που ανήκει στο ίδιο σύστημα, αλλά δεν έχουμε στοιχεία για την ενδιάμεση διαδρομή (αν υπήρξε στάση, παράκαμψη μέχρι να παραδοθεί στον σταθμό κτλ). Εδώ αναλυτικά τα χαρακτηριστικά που έχουμε είναι:

Όνομα	**Περιγραφή**
Start Time	Η ημερομηνία και ώρα που ξεκινάει το ταξίδι (σε τοπική ώρα Νέας Υόρκης)
Stop Time	Η ημερομηνία και ώρα που τελειώνει το ταξίδι (σε τοπική ώρα Νέας Υόρκης)
Start Station ID	Ένας μοναδικός κωδικός για την ταυτοποίηση του σταθμού όπου ξεκινά το ταξίδι.
Start Station Name	Το όνομα του σταθμού όπου ξεκινά το ταξίδι.
End Station ID	Ένας μοναδικός κωδικός για την ταυτοποίηση του σταθμού όπου τελειώνει το ταξίδι.
End Station Name	Το όνομα του σταθμού όπου τελειώνει το ταξίδι.
User Type	Ο τύπος χρήστη ποδηλάτου (Subscriber/Customer)
Bike ID	Ένας μοναδικός κωδικός για την ταυτοποίηση του χρήστη ποδηλάτου.
Gender	Το φύλο του χρήστη.
Age	Η ηλικία του χρήστη.
Trip Duration	Η διάρκεια του ταξιδιού σε λεπτά.

Πίνακας 5. Χαρακτηριστικά του συνόλου δεδομένων ποδηλάτων της Νέας Υόρκης.

Bike Database – London City

Το σύνολο δεδομένων London Bike-Share Usage [7] περιέχει δεδομένα ταξιδιού με ποδήλατο από το σύστημα μίσθωσης ποδηλάτων Transport for London (TfL) στο Λονδίνο, για το χρονικό διάστημα από 01/08/2023 έως και 31/08/2023. Αποτελείται περίπου από 776 χιλιάδες εγγραφές και αποθηκεύεται σε αρχείο χωρητικότητας 105 mb συνολικά. Η πρωτοβουλία TfL Cycle Hire παρέχει δημόσια προσβάσιμα ποδήλατα προς ενοικίαση σε όλο το Λονδίνο, προωθώντας τις βιώσιμες μεταφορές και την ανάπτυξη της φυσικής κατάστασης στην κοινωνία. Αυτό το ολοκληρωμένο σύνολο δεδομένων καταγράφει μεμονωμένα δεδομένα ταξιδιού, τα οποία μπορούν να χρησιμοποιηθούν για την ανάλυση των προτύπων αστικής κινητικότητας, την βελτίωση της απόδοσης των σταθμών και την εύρεση προτιμήσεων ποδηλασίας μεταξύ του διαφορετικού πληθυσμού του Λονδίνου. Σε αντιστοιχία με το παραπάνω σύνολο δεδομένων της Νέας Υόρκης έχουμε δεδομένα για την μετάβαση από σταθμό σε σταθμό αλλά δεν έχουμε στοιχεία για την ενδιάμεση διαδρομή (αν υπήρξε στάση, τυχόν παράκαμψη μέχρι να παραδοθεί στον σταθμό κτλ). Τα χαρακτηριστικά (features) που έχουμε εδώ είναι:

Όνομα	**Περιγραφή**
Number	Ένας μοναδικός αναγνωριστικός αριθμός για κάθε ταξίδι.
Start Date	Η ημερομηνία και ώρα όπου ξεκίνησε το ταξίδι.
Start Station Number	Ο αναγνωριστικός αριθμός του σταθμού όπου ξεκίνησε το ταξίδι.
Start Station	Το όνομα του σταθμού όπου ξεκίνησε το ταξίδι.
End Date	Η ημερομηνία και ώρα όπου τελείωσε το ταξίδι.
End Station Number	Ο αναγνωριστικός αριθμός του σταθμού όπου τελείωσε το ταξίδι.
End Station	Το όνομα του σταθμού που τελείωσε το ταξίδι.
Bike Number	Ένας μοναδικός αναγνωριστικός αριθμός για το ποδήλατο που χρησιμοποιήθηκε.
Bike Model	Το μοντέλο του ποδηλάτου που χρησιμοποιήθηκε.

Total Duration	Η συνολική διάρκεια του ταξιδιού (σε μορφή ευανάγνωστη από τον άνθρωπο, HH:MM:SS)
Total Duration (ms)	Η συνολική διάρκεια του ταξιδιού σε χιλιοστά του δευτερολέπτου.

Πίνακας 6. Χαρακτηριστικά του συνόλου δεδομένων ποδηλάτων του Λονδίνου.

Bike Database – Helsinki City

Το σύνολο δεδομένων Helsinki City Bikes [8] περιέχει δεδομένα ταξιδιού με κοινόχρηστο ποδήλατο στις μητροπολιτικές περιοχές του Ελσίνκι και του Έσποο (Espoo), για τα έτη 2016-2020. Αποτελείται περίπου από 12.1 εκατομμύρια εγγραφές και αποθηκεύεται σε αρχείο χωρητικότητας 1.8 Gb συνολικά. Ο κύριος στόχος του συστήματος ποδηλάτων πόλης του Ελσίνκι είναι να αντιμετωπίσει το λεγόμενο πρόβλημα του τελευταίου μιλίου που υπάρχει σε όλα τα δίκτυα διανομής. Τα ποδήλατα πόλης εισήχθησαν το 2016 ως πιλοτικό έργο με μόνο 46 σταθμούς ποδηλάτων διαθέσιμους στο Ελσίνκι. Αφού το μέσο έγινε δημοφιλές στους πολίτες, η πόλη του Ελσίνκι αποφάσισε να επεκτείνει σταδιακά το δίκτυο ποδηλάτων. Την περίοδο μεταξύ 2017 και 2019, περίπου 100 σταθμοί προστέθηκαν στο δίκτυο κάθε χρόνο, ενώ το έτος 2020 προστέθηκαν μόνο 7 σταθμοί ακόμα για να ολοκληρωθεί το δίκτυο ποδηλάτων. Έχουμε αναλυτικά τα παρακάτω χαρακτηριστικά (features):

Όνομα	**Περιγραφή**
departure	Η ημερομηνία και ώρα έναρξης του ταξιδιού
return	Η ημερομηνία και ώρα τέλους του ταξιδιού
departure_id	Ο αναγνωριστικός αριθμός του σταθμού όπου ξεκίνησε το ταξίδι.
departure_name	Το όνομα του σταθμού όπου ξεκίνησε το ταξίδι.
return_id	Ο αναγνωριστικός αριθμός του σταθμού όπου τελείωσε το ταξίδι.
return_name	Το όνομα του σταθμού όπου τελείωσε το ταξίδι.
distance (m)	Η απόσταση του ταξιδιού σε μέτρα.
duration (sec)	Η διάρκεια του ταξιδιού σε δευτερόλεπτα.

avg_speed (km/h)	Η μέση ταχύτητα του ταξιδιού σε χλμ/ώρα.
departure_latitude	Το γεωγραφικό πλάτος του σημείου όπου ξεκίνησε το ταξίδι.
departure_longitute	Το γεωγραφικό μήκος του σημείου όπου ξεκίνησε το ταξίδι.
return_latitude	Το γεωγραφικό πλάτος του σημείου όπου τελείωσε το ταξίδι.
return_longitute	Το γεωγραφικό μήκος του σημείου όπου τελείωσε το ταξίδι.
Air temperature (degC)	Η θερμοκρασία του αέρα σε βαθμούς Κελσίου.

Πίνακας 7. Χαρακτηριστικά του συνόλου δεδομένων ποδηλάτων του Ελσίνκι.

3.1.3 Περιγραφή Dataset Μέσα Μαζικής Μεταφοράς

Bus Database – New York City

Το σύνολο δεδομένων NYC MTA Buses [5] περιέχει δεδομένα ταξιδιού με λεωφορείο από την υπηρεσία ροής δεδομένων NYC Metropolitan Transportation Authority (MTA) για τους μήνες Ιούνιο, Αύγουστο, Οκτώβριο και Δεκέμβριο του έτους 2017. Αποτελείται περίπου από 26.5 εκατομμύρια εγγραφές και αποθηκεύεται σε 4 αρχεία χωρητικότητας 5.16 Gb συνολικά. Περιλαμβάνει πληροφορία για βήματα περίπου 10 λεπτών, την τοποθεσία του λεωφορείου, την διαδρομή, την στάση καθώς και την προγραμματισμένη ώρα άφιξης σύμφωνα με το πρόγραμμα των λεωφορείων, για να δοθεί μία ένδειξη για το πού πρέπει να βρίσκεται το λεωφορείο (πόσο καθυστερεί, αν φτάνει έγκαιρα, ή πριν από το χρονοδιάγραμμα). Αναλυτικά τα χαρακτηριστικά (features) είναι:

Όνομα	**Περιγραφή**
Recorded At time	Η ημερομηνία και ώρα κατά την οποία έγινε η καταγραφή.
Direction Ref	Δυαδική κωδικοποίηση της κατεύθυνσης του λεωφορείου (0/1).
Published Line Name	Ο αριθμός γραμμής
Origin Name	Το όνομα της αφετηρίας της γραμμής.

Origin Lat	Το γεωγραφικό πλάτος της αφετηρίας της γραμμής.
Origin Long	Το γεωγραφικό μήκος της αφετηρίας της γραμμής.
Destination Name	Το όνομα του τέρματος της γραμμής.
Destination Lat	Το γεωγραφικό πλάτος του τέρματος της γραμμής.
Destination Long	Το γεωγραφικό πλάτος του τέρματος της γραμμής.
Vehicle Ref	Ο αναγνωριστικός αριθμός του οχήματος.
Vehicle Location Lat	Το γεωγραφικό πλάτος του οχήματος την ώρα καταγραφής.
Vehicle Location Long	Το γεωγραφικό μήκος του οχήματος την ώρα καταγραφής.
Next Stop Point Name	Το όνομα της επόμενης στάσης.
Arrival Proximity Test	Η κατηγοριοποίηση του πόσο κοντά βρίσκεται το όχημα στην επόμενη στάση (πχ. approaching, at stop, <1 stop away κτλ.)
Distance from Stop	Η απόσταση από την επόμενη στάση σε γιάρδες.
Expected Arrival Time	Η ημερομηνία και ώρα που αναμένεται η άφιξη στην επόμενη στάση.
Scheduled Arrival Time	Η ημερομηνία και ώρα που είναι προγραμματισμένη η άφιξη στην επόμενη στάση.

Πίνακας 8. Χαρακτηριστικά του συνόλου δεδομένων λεωφορείων της Νέας Υόρκης.

Bus Database – Rio De Janeiro

Το σύνολο δεδομένων GPS Data from Rio De Janeiro Buses [1] περιέχει δεδομένα ταξιδιού με λεωφορείο στην περιοχή του Ρίο Ντε Τζανέιρο για το χρονικό διάστημα 25/01/2019 έως 21/03/2019. Αποτελείται περίπου από 59 εκατομμύρια εγγραφές και αποθηκεύεται σε αρχείο χωρητικότητας 3.5 Gb συνολικά. Περιλαμβάνει τα δεδομένα GPS όλων των διαθέσιμων λεωφορείων στον δήμο της πόλης και μπορεί να παρέχει χρήσιμη πληροφορία για την δυναμική της πόλης και την εύρεση μοτίβων. Τα χαρακτηριστικά (features) εδώ είναι:

Όνομα	Περιγραφή
Date	Η ημερομηνία καταγραφής (MM-DD-YYYY).
Time	Η ώρα καταγραφής (HH:MM:SS).
Order	Ο αναγνωριστικός αριθμός του οχήματος.
Line	Το όνομα της γραμμής/διαδρομής.
Latitude	Το γεωγραφικό πλάτος του οχήματος την ώρα καταγραφής.
Longitude	Το γεωγραφικό μήκος του οχήματος την ώρα καταγραφής.
Speed	Η ταχύτητα του οχήματος σε χλμ./ώρα.

Πίνακας 9. Χαρακτηριστικά του συνόλου δεδομένων λεωφορείων του Ρίο Ντε Τζανέιρο.

Metro Database – New York City

Το σύνολο δεδομένων Normal and New Normal: NYC Subway Traffic 2017-21 [4] περιέχει δεδομένα ταξιδιού με μετρό από 469 σταθμούς μετρό στην περιοχή της Νέας Υόρκης για το χρονικό διάστημα από 04/02/2017 έως 13/08/2021. Αποτελείται περίπου από 4.6 εκατομμύρια εγγραφές και αποθηκεύεται σε αρχείο χωρητικότητας 701 mb. Περιλαμβάνει πληροφορία που εξάγεται από τον αριθμό εισόδων και εξόδων στους σταθμούς του μετρό, όπως υπολογίζονται από τον αριθμό ατόμων που διέρχονται από τα τουρνικέ στους σταθμούς ανά διαστήματα τεσσάρων ωρών. Αναλυτικά τα χαρακτηριστικά (features) είναι:

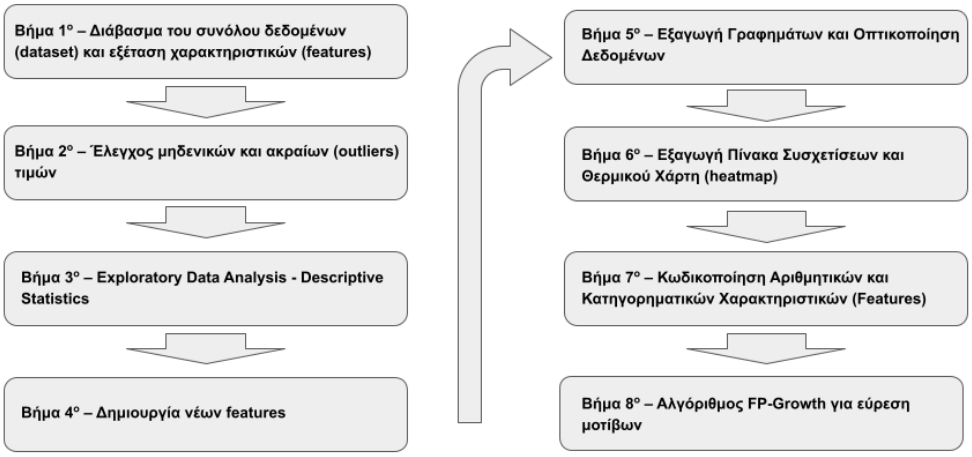
Όνομα	Περιγραφή
Unique ID	Ο αναγνωριστικός αριθμός του σταθμού μετρό.
Datetime	Η ημερομηνία και ώρα καταγραφής.
Stop Name	Το όνομα του σταθμού.
Remote Unit	Αναγνωριστικό ιεραρχικό ανώτερου επιπέδου.
Line	Το όνομα της γραμμής στο οποίο ανήκει η στάση.
Connecting Lines	Οι γραμμές που περνάνε από αυτό τον σταθμό.

Daytime Routes	Οι πρωινές γραμμές που περνάνε από αυτό τον σταθμό.
North Direction Label	Προορισμός των γραμμών που πηγαίνουν βόρεια.
South Direction Label	Προορισμός των γραμμών που πηγαίνουν νότια.
Division	Κάθε γραμμή ανήκει σε ένα από τρία τμήματα του μετρό (IRT, IND, Other)
Structure	Κατασκευή του σταθμού (Υπόγειος, Πάνω από τον δρόμο, κτλ.).
Borough	Προάστιο της Νέας Υόρκης όπου βρίσκεται ο σταθμός (Manhattan, Brooklyn, κτλ.).
Neighborhood	Γειτονιά όπου βρίσκεται ο σταθμός.
Latitude	Γεωγραφικός πλάτος του σταθμού.
Longitude	Γεωγραφικό μήκους του σταθμού.
Entries	Ο αριθμός των ανθρώπων που μπαίνουν στο σταθμό στο διάστημα τεσσάρων ωρών.
Exits	Ο αριθμός των ανθρώπων που βγαίνουν από το σταθμό στο διάστημα τεσσάρων ωρών.

Πίνακας 10. Χαρακτηριστικά του συνόλου δεδομένων μετρό της Νέας Υόρκης.

3.2 Μεθοδολογία Επεξεργασίας και Ανάλυσης Δεδομένων

Σε αυτή την ενότητα θα παρουσιάσουμε την μεθοδολογία και αναλυτικά τα βήματα που ακολουθήσαμε για την επεξεργασία όλων των συνόλων δεδομένων (dataset). Σε περίπτωση που υπάρχουν διαφοροποιήσεις στον χειρισμό θα αναφέρουμε την κάθε υποκατηγορία χωριστά.



Εικόνα 2. Μεθοδολογία και Βήματα Επεξεργασίας συνόλων δεδομένων

3.2.1 Βήμα 1ο – Διάβασμα του συνόλου δεδομένων (dataset) και εξέταση χαρακτηριστικών (features)

Για το διάβασμα και την επεξεργασία θα χρησιμοποιήσουμε το περιβάλλον Jupyter Notebook και την γλώσσα R. Όλα τα σύνολα δεδομένων που αναφέραμε στο παρόν κεφάλαιο είναι σε ένα ή πολλαπλά αρχεία CSV και επομένως θα τα φορτώσουμε με την αντίστοιχη εντολή της Python. Στην περίπτωση που αποθηκεύονται σε πολλαπλά αρχεία θα τα ενοποιήσουμε σε ένα ενιαίο σύνολο δεδομένων αφού φορτωθούν.

Στις περιπτώσεις που τα αρχεία των συνόλων δεδομένων έχουν χωρητικότητα κάτω του 1Gb έχουμε χρησιμοποιήσει την Python και τις βιβλιοθήκες της, ενώ για τα μεγάλα σύνολα δεδομένων έχουμε χρησιμοποιήσει το Apache Pyspark. Αυτά είναι τα σύνολα που αφορούν τα λεωφορεία του Ρίο Ντε Τζανέιρο και της Νέας Υόρκης και τα ποδήλατα του Helsinki. Αυτό γίνεται λόγω του υψηλού κόστους που έχει σε μνήμη και ταχύτητα επεξεργασίας των δεδομένων ένα μεγάλο σύνολο δεδομένων (database), τα οποία αντιμετωπίζονται αποτελεσματικά με χρήση του Pyspark, χάρη στην κατανεμημένη αρχιτεκτονική του για επεξεργασία δεδομένων σε κλίμακα. Ένα άλλο πλεονέκτημα του είναι ότι επιτρέπει την χρήση της γλώσσας R ώστε να έχουμε παρόμοια επεξεργασία και ανάλυση σε όλα τα σύνολα δεδομένων ανά κατηγορία.

Η δομή δεδομένων στην οποία διαβάζουμε τα δεδομένα για την μετέπειτα επεξεργασία και ανάλυση είναι το πλαίσιο δεδομένων (DataFrame). Η δομή δεδομένων αυτή είναι ένα από τα πιο σημαντικά και ευέλικτα εργαλεία στην ανάλυση

δεδομένων, ειδικά στη βιβλιοθήκη Pandas της Python. Ένα DataFrame είναι ουσιαστικά ένας δισδιάστατος πίνακας δεδομένων που μοιάζει με ένα λογιστικό φύλλο ή έναν πίνακα βάσης δεδομένων, και μπορεί να περιέχει ετερογενή δεδομένα. Κάθε στήλη μπορεί να περιλαμβάνει δεδομένα διαφορετικών τύπων (αριθμούς, κείμενο, ημερομηνίες, κ.λπ.), και κάθε γραμμή αντιπροσωπεύει μία εγγραφή ή ένα δείγμα. Οι στήλες και οι γραμμές είναι κωδικοποιημένες, επιτρέποντας την εύκολη πρόσβαση και χειρισμό των δεδομένων. Το DataFrame υποστηρίζει πληθώρα λειτουργιών, όπως φιλτράρισμα, ομαδοποίηση, συγχώνευση και στατιστική ανάλυση, καθιστώντας το απαραίτητο εργαλείο για επιστήμονες δεδομένων και αναλυτές.

Μετά την δημιουργία των Dataframes, για κάθε σύνολο δεδομένων (dataset) ελέγχουμε αν όλα τα χαρακτηριστικά (features) υπάρχουν στο DataFrame και ότι έχουν το σωστό τύπο δεδομένων (float, datetime, etc.). Αν κάτι δεν έχει διαβαστεί σωστά επαναλαμβάνουμε την διαδικασία και μετατρέπουμε όποιο feature χρειάζεται στον σωστό τύπο δεδομένων. Εδώ κοιτάμε ποια χαρακτηριστικά είναι χρήσιμα, ποια θα αναλύσουμε και μεταξύ των κατηγοριών μετακίνησης ποια είναι όμοια και επομένως συγκρίσιμα μεταξύ των διαφορετικών συνόλων δεδομένων. Για κάθε κατηγορία μετακίνησης για τα σύνολα δεδομένων έχουμε:

- *Μισθωμένα οχήματα – Ταξί*

Για τα σύνολα δεδομένων ταξιδιού με ταξί στην Νέα Υόρκη και το Σικάγο, θα επικεντρωθούμε στα χαρακτηριστικά που αφορούν το κόστος της διαδρομής και την χρονική διάρκεια του ταξιδιού και την απόσταση που διανύεται. Επίσης θα χρησιμοποιήσουμε τα χαρακτηριστικά που αφορούν την ημερομηνία και ώρα εκκίνησης και τέλους του ταξιδιού καθώς και θα χρησιμοποιήσουμε στις περιπτώσεις που υπάρχουν χαρακτηριστικά τοποθεσίας όπως γεωγραφικό πλάτος και μήκος για να γίνει περαιτέρω επεξεργασία και εξαγωγή συμπερασμάτων.

Για το σύνολο δεδομένων των ταξιδιών Uber έχουμε μικρό αριθμό χαρακτηριστικών και θα επικεντρωθούμε στα δεδομένα που αφορούν το 2014 ώστε να χρησιμοποιήσουμε περισσότερη πληροφορία όπως τα χαρακτηριστικά της ημερομηνίας και ώρας εκκίνησης του ταξιδιού και τα γεωγραφικά χαρακτηριστικά (γεωγραφικό πλάτος, γεωγραφικό μήκος).

Τα χαρακτηριστικά που είναι “1-1” συγκρίσιμα ανάμεσα στα τρία σύνολα δεδομένων που αφορούν ταξίδι με Μισθωμένα οχήματα – Ταξί φαίνονται στον παρακάτω πίνακα.

Taxi Database - Chicago	**Taxi Database – New York**	**Uber Database – New York**
Trip Start Timestamp	lpep_pickup_datetime	Date/Time
Trip Miles	trip_distance	
Fare	fare_amount	
Extras	extra	
Tips	tip_amount	
Tolls	tolls_amount	
Trip Total	total_amount	
Payment Type	payment_type	

Πίνακας 11. Χαρακτηριστικά συγκρίσιμα για την κατηγορία μισθωμένα οχήματα-ταξί.

- *Ενοικιαζόμενα ποδήλατα*

Σε αυτή την κατηγορία μετακίνησης τα τρία σύνολα δεδομένων που έχουμε έχουν αρκετά παρόμοια χαρακτηριστικά. Θα χρησιμοποιήσουμε τα χαρακτηριστικά που αφορούν την ημερομηνία και ώρα έναρξης και λήξης του ταξιδιού σε παρακάτω ανάλυση. Επίσης τα χαρακτηριστικά που αφορούν τα ονόματα των σταθμών εκκίνησης και λήξης του ταξιδιού θα τα αναλύσουμε για την εύρεση των πιο δημοφιλή σταθμών. Τέλος, θα κρατήσουμε όλα τα αριθμητικά χαρακτηριστικά που αφορούν το ταξίδι όπως απόσταση, διάρκεια, θερμοκρασία περιβάλλοντος και τον χρήστη όπως φύλο και ηλικία.

Εδώ τα συγκρίσιμα χαρακτηριστικά είναι:

Bike Database – New York	**Bike Database - London**	**Bike Database - Helsinki**
start time	Start date	departure
stop_time	End date	return
trip_duration	Total duration (ms)	duration (sec)

Πίνακας 12. Χαρακτηριστικά συγκρίσιμα για την κατηγορία ενοικιαζόμενα ποδήλατα.

Σημείωση: Για την σύγκριση των χαρακτηριστικών όπως η διάρκεια του ταξιδιού θα πρέπει να μετατρέψουμε όλα στην ίδια μονάδα μέτρησης (λεπτά – min).

- *Μέσα Μαζικής Μεταφοράς*

Σε αυτή την κατηγορία μετακίνησης τα σύνολα δεδομένων που έχουμε δεν έχουν πολλά χαρακτηριστικά τα οποία μπορούμε να συγκρίνουμε απευθείας για να βγάλουμε συμπεράσματα. Τα μόνα χαρακτηριστικά που μπορούμε να χρησιμοποιήσουμε για απευθείας σύγκριση είναι αυτά που αφορούν την ημερομηνία και ώρα για να εξάγουμε πληροφορία για τις ώρες αιχμής.

Εκτός από αυτά θα χρησιμοποιήσουμε για τα σύνολα δεδομένων των λεωφορείων της Νέας Υόρκης και του Ρίο Ντε Τζανέιρο τα χαρακτηριστικά που αφορούν τις γεωγραφικές θέσεις του λεωφορείου (γεωγραφικό πλάτος και γεωγραφικό μήκος) κατά την διάρκεια του ταξιδιού. Επίσης τα χαρακτηριστικά που αφορούν την γραμμή του λεωφορείου και των οχημάτων θα χρησιμοποιηθούν για την ανάλυση των πιο δημοφιλών γραμμών. Τέλος, θα κρατήσουμε όλα τα αριθμητικά χαρακτηριστικά όπως η ταχύτητα του λεωφορείου, η απόσταση από την στάση κτλ.

Για το σύνολο δεδομένων του μετρό της Νέας Υόρκης θα χρησιμοποιήσουμε επίσης τα αριθμητικά χαρακτηριστικά που αφορούν τον αριθμό εισόδου και εξόδου από κάθε στάση του μετρό και το χαρακτηριστικό που αφορά το όνομα της στάσης. Επίσης θα χρησιμοποιήσουμε τα γεωγραφικά χαρακτηριστικά όπως η γειτονιά και το προάστιο του σταθμού. Τέλος θα χρησιμοποιήσουμε για την ανάλυση μας τον αριθμό της κάθε γραμμής του μετρό και τον τρόπο κατασκευής του σταθμού (αν είναι υπόγειο ή όχι).

3.2.2 Βήμα 2ο – Έλεγχος μηδενικών και ακραίων (outliers) τιμών

Αφού έχουμε φορτώσει τα δεδομένα μας για κάθε σύνολο δεδομένων (dataset) ελέγχουμε αν έχουμε κενές τιμές σε κάποιο χαρακτηριστικό (feature) και τους δίνουμε την τιμή NaN (Not A Number). Για κάθε χαρακτηριστικό, γίνεται υπολογισμός του ποσοστού που έχουμε σε τιμές NaN ώστε να το διαχειριστούμε αναλόγως.

Μετά από αυτή την διαδικασία ελέγχουμε το εύρος τιμών που έχουμε σε βασικά αριθμητικά χαρακτηριστικά (features) όπως πχ. η απόσταση του ταξιδιού ώστε να αφαιρέσουμε πιθανόν ακραίες τιμές που επηρεάζουν τα αποτελέσματα μας και πιθανόν να οφείλονται σε σφάλματα ή διάφορα τεστ κατά την διάρκεια συλλογής των

δεδομένων. Συγκεκριμένα για κάθε κατηγορία μετακίνησης των συνόλων δεδομένων έχουμε:

- *Μισθωμένα οχήματα – Ταξί*

Taxi Database – Chicago: Ο αριθμός των τιμών NaN για τα feature που εξετάζουμε στατιστικά δηλαδή τις μεταβλητές Trip Seconds, Trip Miles, Fare, Tips, Tolls, Extras, Trip Total είναι αρκετά μικρός (<0,01%) σε σχέση με το μέγεθος του συνόλου των δεδομένων οπότε θα διαγράψουμε αυτές τις εγγραφές.

Όσον αφορά τα features που είναι γεωγραφικές συντεταγμένες του σημείου εκκίνησης και λήξης του ταξιδιού που θέλουμε να χρησιμοποιήσουμε για να γίνει ανάλυση για τις γειτονιές όπου βρίσκονται αυτά τα σημεία, το ποσοστό των τιμών NaN κυμαίνεται από 9,2%-62,17% ανάλογα το χαρακτηριστικό, τις εγγραφές αυτές δεν θα τις διαγράψουμε διότι θα μειωθεί αρκετά το δείγμα των δεδομένων, αλλά θα εισάγουμε μία επιλογή 'None' στην γειτονιά και θα προχωρήσουμε με την διερευνητική ανάλυση, όπως θα δούμε σε παρακάτω βήμα.

Σε αυτό το σύνολο δεδομένων, ελέγχουμε για ακραίες τιμές στο χαρακτηριστικό Trip Seconds και αφαιρούμε τις εγγραφές για τις οποίες το ταξίδι είναι πολύ μεγάλο σε διάρκεια, δηλαδή πάνω από 60 λεπτά (1,45% επί του συνόλου). Αντίστοιχα για το χαρακτηριστικό Trip Miles αφαιρούμε τις εγγραφές που αντιστοιχούν σε ταξίδι με πάνω από 40 μίλια (1,44% επί του συνόλου). Μετά τον χειρισμό αυτόν για την απαλοιφή εγγραφών με ακραίες τιμές, παρατηρείται ότι εξομαλύνονται και τα χαρακτηριστικά του κόστους, με το χαρακτηριστικό Trip Total σε ελάχιστες περιπτώσεις να υπερβαίνει τα 100 δολάρια.

Taxi Database – New York City: Εδώ βλέπουμε ότι το feature ehail_fee δεν μπορούμε να το χρησιμοποιήσουμε γιατί έχει τιμές NaN σε όλο το δείγμα, καθώς δεν έχουν συμπληρωθεί οι τιμές στο σύνολο δεδομένων. Τα features VendorID, store_and_fwd_flag, RatecodeID, passenger_count, payment_type, trip_type και congestion_surcharge έχουν τιμές που λείπουν, NaN, σε αρκετά μεγάλο ποσοστό των παρατηρήσεων στο σύνολο δεδομένων, περίπου 38. 85%, επομένως θα κρατήσουμε τις εγγραφές στο σύνολο δεδομένων και θα τις συμπεριλάβουμε στην ανάλυσή μας.

Κατά τον έλεγχο για τις ακραίες τιμές σε αντιστοιχία με το σύνολο δεδομένων ταξιδιού με ταξί στο Σικάγο, το χαρακτηριστικό trip_duration, που δημιουργήσαμε αφαιρώντας το χαρακτηριστικό lpep_dropoff_datetime από το lpep_pickup_datetime,

είχε ακραίες τιμές άνω των 60 λεπτών διάρκειας ταξιδιού σε ποσοστό 3.61% επί του συνόλου δεδομένων, επομένως αφαιρέσαμε τις εγγραφές που αντιστοιχούσαν σε αυτές τις ακραίες τιμές. Το χαρακτηριστικό total_amount είχε ακραίες τιμές, αρνητικό αριθμό και τιμή μεγαλύτερη των 100 δολαρίων σε ποσοστό 0.2% επί του συνόλου, επομένως αφαιρέσαμε τις αντίστοιχες εγγραφές από το σύνολο δεδομένων. Τέλος, αντίστοιχα με το σύνολο δεδομένων ταξί στο Σικάγο, αφαιρέσαμε εγγραφές με τιμή μεγαλύτερη των 40 μιλίων για την μεταβλητή trip_distance, που ήταν στο 0.2% του συνόλου.

Uber Database – New York: Σε αυτό το σύνολο δεδομένων δεν είχαμε καθόλου μηδενικές/NaN τιμές και δεν χρειάστηκε να αφαιρέσουμε ακραίες τιμές.

- *Ενοικιαζόμενα Ποδήλατα*

Bike Database – New York City: Σε αυτό το σύνολο δεδομένων δεν είχαμε καθόλου μηδενικές/NaN τιμές να χειριστούμε.

Το χαρακτηριστικό trip_duration είχε ακραίες τιμές άνω των 60 λεπτών διάρκειας ταξιδιού σε ποσοστό 0.6% επί του συνόλου δεδομένων, επομένως αφαιρέσαμε τις εγγραφές που αντιστοιχούσαν στις τιμές αυτές.

Bike Database – London: Σε αυτό το σύνολο δεδομένων δεν είχαμε καθόλου μηδενικές/NaN τιμές να χειριστούμε.

Το χαρακτηριστικό Total duration (ms) είχε ακραίες τιμές άνω των 60 λεπτών διάρκειας ταξιδιού σε ποσοστό 3.59% επί του συνόλου δεδομένων, επομένως αφαιρέσαμε τις εγγραφές που αντιστοιχούσαν σε αυτές τις ακραίες τιμές.

Bike Database – Helsinki: Παρατηρούμε ότι τα χαρακτηριστικά avg_speed και Air temperature είχαν τιμές NaN σε ποσοστό 0.003 και 0.12% επί του συνόλου δεδομένων, αντίστοιχα, επομένως αφαιρέσαμε τις αντίστοιχες εγγραφές.

Κατά τον έλεγχο για ακραίες τιμές, παρατηρούμε ότι το χαρακτηριστικό duration (sec) είχε ακραίες τιμές άνω των 60 λεπτών διάρκειας ταξιδιού σε ποσοστό 1.32% επί του συνόλου δεδομένων, επομένως αφαιρέσαμε τις αντίστοιχες εγγραφές. Επίσης, κρατήσαμε εγγραφές που είχαν τιμές για το χαρακτηριστικό avg_speed (km/h) στο εύρος [0, 40] και για το χαρακτηριστικό distance (m) στο εύρος [0, 10], καθώς θεωρήσαμε ότι δεν είναι έκτοπες.

- *Μέσα Μαζικής Μεταφοράς*

Bus Database – New York City: Σε αυτό το σύνολο δεδομένων δεν είχαμε καθόλου μηδενικές/NaN τιμές να χειριστούμε. Επίσης ούτε ακραίες τιμές υπήρχαν που χρειάστηκαν να χειριστούμε.

Bus Database – Rio De Janeiro: Ομοίως σε αυτό το σύνολο δεδομένων δεν είχαμε καθόλου μηδενικές/NaN τιμές να χειριστούμε και δεν ανιχνεύθηκαν ακραίες τιμές για κάποιο χαρακτηριστικό.

Metro Database – New York City: Τα χαρακτηριστικά North Direction Label, South Direction Label έχουν αρκετές τιμές που λείπουν στο σύνολο δεδομένων, αλλά επειδή αφαιρέθηκαν από το δείγμα ως στατιστικά μη σημαντικές μεταβλητές, κρατάμε τις αντίστοιχες εγγραφές του δείγματος.

Τα χαρακτηριστικά Entries και Exit έχουν τιμές για κάποιους σταθμούς οι οποίες είναι πολύ υψηλές, τις οποίες δεν έχουμε συμπεριλάβει σε κάποια γραφήματα για λόγους ομαλότητας και για να είναι διαισθητικά αντιληπτά, όπως θα δούμε σε επόμενο κεφάλαιο, αλλά επειδή είναι πραγματικές τιμές δεν έχουμε αφαιρέσει τις αντίστοιχες εγγραφές από το σύνολο δεδομένων και τις χρησιμοποιούμε στους υπολογισμούς μας.

3.2.3 Βήμα 3ο – Exploratory Data Analysis – Descriptive Statistics

Στο επόμενο βήμα της ανάλυσης, για κάθε σύνολο δεδομένων (dataset), γίνεται υπολογισμός των βασικών στατιστικών μέτρων Περιγραφικής Στατιστικής για τις κατανομές που ακολουθούν τα αριθμητικά χαρακτηριστικά (features) στο δείγμα του κάθε συνόλου. Τα μέτρα αυτά μας βοηθούν να κατανοήσουμε την ποιότητα των δεδομένων μας και να συγκρίνουμε την στατιστική σημαντικότητα όμοιων χαρακτηριστικών μεταξύ των διαφορετικών συνόλων δεδομένων για την ίδια κατηγορία μετακίνησης.. Τα στατιστικά μέτρα που χρησιμοποιήσαμε είναι τα παρακάτω χωρισμένα σε τρεις κατηγορίες:

- *Μέτρα Κεντρικής Τάσης (Central Tendency)*

1. Μέσος όρος (Mean): Το άθροισμα όλων των τιμών των παρατηρήσεων μίας μεταβλητής του συνόλου δεδομένων διαιρούμενο με το πλήθος των παρατηρήσεων. Μας δίνει μια ιδέα της "κεντρικής" τιμής γύρω από την οποία βρίσκονται οι περισσότερες παρατηρήσεις της μεταβλητής του συνόλου δεδομένων.

2. Διάμεσος (Median): Η μεσαία τιμή των τιμών των παρατηρήσεων μίας μεταβλητής συνόλου δεδομένων όταν οι παρατηρήσεις είναι ταξινομημένες κατά αύξουσα ή φθίνουσα σειρά. Είναι η κεντρική τιμή που χωρίζει το δείγμα σε δύο ίσα μέρη.
3. Επικρατούσα τιμή (Mode): Η τιμή που εμφανίζεται με τη μεγαλύτερη συχνότητα στις τιμές των παρατηρήσεων μίας μεταβλητής στο σύνολο δεδομένων, δηλαδή μας δείχνει την πιο κοινή τιμή που παίρνει η μεταβλητή στο σύνολο δεδομένων.
- *Μέτρα Διασποράς (Varibility)*
1. Τυπική Απόκλιση (Standard Deviation): Είναι η τετραγωνική ρίζα της διακύμανσης. Η τυπική απόκλιση παρέχει μια πιο άμεσα κατανοητή μέτρηση της διασποράς, καθώς είναι στην ίδια μονάδα μέτρησης με τα δεδομένα. Δείχνει πόσο οι τιμές των παρατηρήσεων τείνουν να αποκλίνουν από τον μέσο όρο.
2. Ενδοτεταρτημοριακό εύρος (Interquartile Range - IQR): Έχουμε τις τιμές για κάθε εκατοστημόριο δηλαδή για το 25%, 50% και 75%. Η διαφορά μεταξύ του πρώτου (25ο εκατοστημόριο) και του τρίτου τεταρτημόριου (75ο εκατοστημόριο), δείχνει την κεντρική τάση των παρατηρήσεων και πόσο διασκορπισμένες είναι οι μεσαίες τιμές.
3. Ελάχιστη και Μέγιστη τιμή (min και max): Η διαφορά των μέτρων αυτών μας δίνει το εύρος τιμών στις παρατηρήσεις της μεταβλητής στο σύνολο δεδομένων.
4. Πλήθος δεδομένων (count): Μας δείχνει το μέγεθος του δείγματος στο σύνολο δεδομένων μας.
- *Μέτρα Κατανομής*
1. Συμμετρία (Skewness): Δείχνει την ασυμμετρία της κατανομής μίας μεταβλητής στο σύνολο δεδομένων. Αν η τιμή είναι θετική, η κατανομή είναι ασύμμετρη προς τα δεξιά, ενώ αν είναι αρνητική, η κατανομή είναι ασύμμετρη προς τα αριστερά.
2. Κύρτωση (Kurtosis): Δείχνει το "ύψος" της κατανομής μίας μεταβλητής. Υψηλή κύρτωση σημαίνει ότι η κατανομή έχει αιχμηρή κορυφή και βαριές ουρές, ενώ χαμηλή κύρτωση σημαίνει ότι η μεταβλητή έχει πιο επίπεδη κατανομή.

3.2.4 Βήμα 4ο – Δημιουργία νέων features

Για να μπορέσουμε να προχωρήσουμε την ανάλυση και να ενοποιήσουμε τα σύνολα δεδομένων ανά κατηγορία μετακίνησης ώστε να έχουμε συγκρίσιμα αποτελέσματα και διαγράμματα χρειάστηκε να δημιουργήσουμε καινούργια χαρακτηριστικά (features) χρησιμοποιώντας τα υπάρχοντα, με τα οποία μπορούμε να εξάγουμε πιο χρήσιμες πληροφορίες. Τα χαρακτηριστικά (features) που δημιουργήσαμε είναι:

- Στο σύνολο δεδομένων Ταξί στην Νέα Υόρκη δεν έχουμε χαρακτηριστικό για την χρονική διάρκεια του ταξιδιού ώστε να το συγκρίνουμε με το αντίστοιχο χαρακτηριστικό στο σύνολο δεδομένων ταξιδιού με ταξί στο Σικάγο. Έχοντας όμως την πληροφορία για την ημερομηνία και ώρα εκκίνησης και λήξης το δημιουργούμε με την διαφορά τους με μονάδα μέτρησης τα λεπτά.
- Και στα 9 σύνολα δεδομένων έχουμε χαρακτηριστικά που δηλώνουν την ημερομηνία και ώρα που λαμβάνει χώρα το ταξίδι (εκκίνηση ταξιδιού, λήξη ταξιδιού, στιγμή που έχει καταγραφεί το όχημα) και από αυτά μπορούμε να εξάγουμε νέα χαρακτηριστικά για πιο λεπτομερή ανάλυση. Συγκεκριμένα δημιουργούμε το χαρακτηριστικό Hour που δηλώνει την ώρα της μέρας που έχουμε (από το 0 έως το 24), το χαρακτηριστικό Day of the Week που δηλώνει την ημέρα της εβδομάδας που έχουμε (Δευτέρα, Τρίτη, κτλ.) και το χαρακτηριστικό Month που δηλώνει το μήνα που έχουμε (Ιανουάριος, Φεβρουάριος, κτλ.).
- Στα σύνολα δεδομένων που έχουμε δεδομένα τοποθεσίας (γεωγραφικό πλάτος, γεωγραφικό μήκος) σε μία πόλη τα χρησιμοποιήσουμε για δημιουργήσουμε ένα καινούργιο feature που να δηλώνει την γειτονιά/δήμο που βρισκόμαστε ώστε να μπορούμε να εξάγουμε και πληροφορίες για την τοπολογία των μετακινήσεων.

  Για να το καταφέρουμε αυτό έχουμε βρει κατάλληλα shapefile που περιέχουν τα πολύγωνα/όρια των γειτονιών για τις πόλεις που μελετούμε. Ένα αρχείο shapefile είναι μια δημοφιλής μορφή αποθήκευσης γεωγραφικών δεδομένων που αναπτύχθηκε από την ESRI (Environmental Systems Research Institute). Χρησιμοποιείται ευρέως σε Γεωγραφικά Συστήματα Πληροφοριών (GIS) για την αναπαράσταση γεωχωρικών δεδομένων και την ανάλυση χωρικών πληροφοριών. Το shapefile δεν είναι ένας μοναδικός τύπος αρχείου, αλλά μια συλλογή από τουλάχιστον τρία αρχεία με την ίδια βασική ονομασία και διαφορετικές επεκτάσεις. Εμάς μας ενδιαφέρει η επέκταση .shp που περιέχει τη γεωμετρία των γεωχωρικών χαρακτηριστικών (σημεία, γραμμές, πολύγωνα).

  Επίσης για να πετύχουμε μεγάλη μείωση στο χρόνο τρεξίματος αυτών των υπολογισμών θα κάνουμε χρήση της βιβλιοθήκης shapely για την Python και της μεθόδου spatial indexing. Το spatial indexing είναι μια τεχνική που

χρησιμοποιείται για τη βελτιστοποίηση της αποθήκευσης, ανάκτησης και διαχείρισης γεωχωρικών δεδομένων σε βάσεις δεδομένων, προσφέροντας ταχύτατη πρόσβαση και αποτελεσματική εκτέλεση χωρικών ερωτημάτων.0

Ο αλγόριθμος που ακολουθήσαμε σε απλά βήματα περιγράφεται ως εξής:

1. Φορτώνουμε το shapefile των γειτονιών σε ένα GeoDataFrame και το μετατρέπουμε στο κατάλληλο σύστημα αναφοράς συντεταγμένων.
2. Ορίζουμε μια συνάρτηση που παίρνει το γεωγραφικό πλάτος και μήκος, και επιστρέφει το όνομα της γειτονιάς.
3. Σε αυτήν ελέγχουμε αν το πλάτος και το μήκος είναι έγκυρα.
4. Δημιουργούμε ένα σημείο από τις συντεταγμένες.
5. Εντοπίζουμε τα πολύγωνα που ενδέχεται να περιέχουν το σημείο χρησιμοποιώντας το χωρικό ευρετήριο.
6. Ελέγχουμε κάθε πιθανό ταίριασμα για να δούμε αν περιέχει το σημείο και επιστρέφουμε το όνομα της γειτονιάς αν βρεθεί.
7. Αν δεν βρεθεί κανένα ταίριασμα, επιστρέφουμε `None`.

3.2.5 Βήμα 5ο – Εξαγωγή Γραφημάτων και Οπτικοποίηση Δεδομένων

Για την οπτικοποίηση των δεδομένων, τα είδη γραφημάτων που δημιουργήσαμε ανά κατηγορία μετακίνησης για τα σύνολα δεδομένων είναι:

- *Μισθωμένα οχήματα – Ταξί*

Σε αυτή την κατηγορία για τα χαρακτηριστικά που αφορούν το ταξίδι με ταξί (Σικάγο και Νέα Υόρκη) και συγκεκριμένα για αυτά που μας δείχνουν την χρονική διάρκεια του ταξιδιού, την απόσταση του ταξιδιού και το συνολικό κόστος του ταξιδιού, δημιουργούμε ιστογράμματα που μας δείχνουν την κατανομή των τιμών αυτών. Επίσης δημιουργούμε διάγραμμα συχνοτήτων για το χαρακτηριστικό της ώρα της ημέρας εκκίνησης του ταξιδιού ώστε να δούμε τις ώρες αιχμής. Διάγραμμα συχνοτήτων φτιάχνουμε και για τους τρόπους πληρωμής και την εταιρεία ταξί για αυτά τα δύο σύνολα δεδομένων. Για το σύνολο δεδομένων ταξιδιού με ταξί της Νέας Υόρκης έχουμε και διάγραμμα συχνοτήτων του αριθμού των επιβατών, ενώ για αυτό του Σικάγο που περιέχει δεδομένα τοποθεσίας έχουμε διάγραμμα συχνοτήτων των 10 πιο δημοφιλών γειτονιών στις οποίες ξεκινούν και τελειώνουν ταξίδια.

Στο σύνολο δεδομένων ταξιδιού με Uber της Νέας Υόρκης έχουμε αντίστοιχα διάγραμμα συχνοτήτων για το χαρακτηριστικό της ώρας της ημέρας εκκίνησης του ταξιδιού ώστε να δούμε τις ώρες αιχμής. Σε συνέχεια της χρονικής ανάλυσης που πραγματοποιείται ένα ταξίδι δημιουργούμε επίσης διάγραμμα συχνοτήτων για τον μήνα, την ημέρα της εβδομάδας και ανά ώρα ξεχωριστά για τον κάθε μήνα. Τέλος έχουμε ένα θερμικό χάρτη συσχέτισης (heatmap) για την ώρα εκκίνησης του ταξιδιού ανά ημερομηνία (δηλαδή 1,2,3 κτλ κάθε μήνα).

- *Ενοικιαζόμενα Ποδήλατα*

Σε αυτή την κατηγορία μετακίνησης για τα τρία σύνολα δεδομένων δημιουργούμε ιστόγραμμα που μας δείχνει την κατανομή των τιμών για ένα χαρακτηριστικό για την διάρκεια του ταξιδιού. Στην περίπτωση του συνόλου δεδομένων ταξιδιού με ποδήλατο της Νέας Υόρκης έχουμε αντίστοιχο ιστόγραμμα κατανομής της ηλικίας των χρηστών και διάγραμμα συχνοτήτων για το φύλο και τον τύπο του χρήστη (συνδρομητής ή όχι). Για το σύνολο δεδομένων ταξιδιού με ποδήλατο του Ελσίνκι έχουμε επίσης ιστόγραμμα κατανομής για την απόσταση του ταξιδιού, την μέση ταχύτητα του ταξιδιού και την θερμοκρασία του αέρα. Επειδή εδώ έχουμε γεωγραφικά χαρακτηριστικά μπορούμε να εξάγουμε και διάγραμμα συχνοτήτων για τις 10 πιο δημοφιλείς γειτονιές για εκκίνηση ταξιδιού. Τέλος και στις τρεις περιπτώσεις εξάγουμε διαγράμματα συχνοτήτων των 10 πιο δημοφιλών σταθμών εκκίνησης, τέλους και αθροιστικά , δηλαδή είτε εκκίνησης είτε τέλους για ένα ταξίδι με ποδήλατο.

- *Μέσα Μαζικής Μεταφοράς*

Σε αυτή την κατηγορία μετακίνησης στις περιπτώσεις των λεωφορείων (Νέα Υόρκη και Ρίο Ντε Τζανέιρο) έχουμε διάγραμμα συχνοτήτων του χαρακτηριστικού της γραμμής του λεωφορείου και για πιο ευανάγνωστη πληροφορία των 20 πιο δημοφιλών γραμμών. Για το σύνολο δεδομένων ταξιδιού με λεωφορείο της Νέας Υόρκης έχουμε ακόμα διάγραμμα των 20 πιο πολυσύχναστων στάσεων με βάση την επισκεψιμότητα των γραμμών, διάγραμμα συχνοτήτων της απόστασης του λεωφορείου από την στάση και της κατηγοριοποίησης αυτής της απόστασης και τέλος διάγραμμα συχνοτήτων του προαστίου που βρίσκεται το λεωφορείο , όπως αυτή έχει εξαχθεί με χρήση των γεωχωρικών χαρακτηριστικών στο σύνολο δεδομένων.

Για το σύνολο δεδομένων ταξιδιού με μετρό της Νέας Υόρκης εξάγουμε διαγράμματα συχνοτήτων για τον τύπο, του σταθμού (υπόγεια, πάνω από τον δρόμο, κτλ.), του προαστίου σε σχέση με τον αριθμό στάσεων, της γειτονιάς σε σχέση με τον αριθμό στάσεων και των δέκα πιο δημοφιλών στάσεων με βάση τον συνολικό αριθμό των εισερχομένων επιβατών σε σχέση με την μέρα της εβδομάδας. Τέλος, δημιουργούμε ιστόγραμμα της κατανομής του αριθμού των εισερχομένων επιβατών και των εξερχομένων επιβατών από τους σταθμούς.

3.2.6 Βήμα 6ο – Εξαγωγή Πίνακα Συσχετίσεων και Θερμικού Χάρτη (heatmap)

Σε κάθε σύνολο δεδομένων που περιέχει πάνω από ένα αριθμητικό χαρακτηριστικό, δημιουργούμε το πίνακα συσχέτισης ώστε να παρατηρήσουμε τυχόν συσχετίσεις μεταξύ αυτών. Στον πίνακα συσχέτισης εμφανίζονται οι τιμές των συντελεστών συσχέτισης μεταξύ κάθε ζεύγους μεταβλητών, όπου οι τιμές κυμαίνονται από -1 έως 1, δείχνοντας την κατεύθυνση (θετική ή αρνητική) και την ένταση της γραμμικής σχέσης μεταξύ τους. Οι τιμές κοντά στο 1 ή -1 υποδηλώνουν ισχυρή συσχέτιση (θετική ή αρνητική αντίστοιχα), ενώ οι τιμές κοντά στο 0 δείχνουν αδύναμη ή ανύπαρκτη γραμμική σχέση.

Στα σύνολα δεδομένων του Uber της Νέας Υόρκης, των λεωφορείων της Νέας Υόρκης και του Ρίο Ντε Τζανέιρο, καθώς και του μετρό της Νέας Υόρκης στα οποία έχουμε γεωγραφικά χαρακτηριστικά για να προχωρήσουμε την ανάλυση μας δημιουργούμε γεωγραφικό θερμικό χάρτη που δείχνει που βρίσκονται τα οχήματα στον χάρτη ανά ώρα. Αυτή η απεικόνιση μας δείχνει τη γεωγραφική κατανομή δεδομένων χρησιμοποιώντας χρώματα για να απεικονίσει την ένταση ή τη συχνότητα και βοηθά στον γρήγορο εντοπισμό των περιοχών με υψηλή ή χαμηλή συγκέντρωση του φαινομένου που μελετάται, στην περίπτωση αυτή την μετακίνηση, κάνοντας έτσι πιο εύκολο να εντοπίσουν μοτίβα και τάσεις στον χώρο, βοηθώντας στη λήψη αποφάσεων και στον σχεδιασμό στρατηγικών.

3.2.7 Βήμα 7ο – Κωδικοποίηση Αριθμητικών και Κατηγορηματικών Χαρακτηριστικών (Features)

Σε αυτό το βήμα κάνουμε την προετοιμασία ώστε να μπορέσουμε να τρέξουμε τον αλγόριθμο FP-Growth για να την εύρεση συχνών μοτίβων στα δεδομένα μας. Τα αριθμητικά χαρακτηριστικά που έχουμε επιλέξει να συμπεριλάβουμε στον αλγόριθμο

θα πρέπει να τα μετατρέψουμε σε κατηγορικά. Αναλυτικά λοιπόν για κάθε κατηγορία μετακίνησης και κάθε σύνολο δεδομένων παρουσιάζουμε ένα πίνακα που αναφέρει το αριθμητικό χαρακτηριστικό που κατηγοριοποιήθηκε και το πώς.

- *Μισθωμένα οχήματα – Ταξί*

Taxi Database - Chicago	
Χαρακτηριστικό	**Κατηγοριοποίηση**
Trip Seconds	< 5 λεπτά: short trip duration 5 ≤ λεπτά <20: regular trip duration 20 ≤ λεπτά <35: medium trip duration ≥ 35 λεπτά: long trip duration
Trip Miles	< 5 μίλια: short trip 5 ≤ μίλια <10: regular trip 10 ≤ μίλια <20: medium trip ≥ 20 μίλια: long trip
Extras	< 5 δολάρια: low extras 5 ≤ δολάρια <10: regular extras ≥ 10 δολάρια: lots of extras
Trip Total	< 10 δολάρια: low fair 10 ≤ δολάρια <20: regular fair 20 ≤ δολάρια <30: high fair ≥ 30 δολάρια: very high fair
Hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night

Πίνακας 13. Κατηγοριοποίηση χαρακτηριστικών για ταξί του Σικάγο.

Taxi Database – New York City	
Χαρακτηριστικό	**Κατηγοριοποίηση**
trip_duration_minutes	< 5 λεπτά: short trip duration 5 ≤ λεπτά <20: regular trip duration 20 ≤ λεπτά <35: medium trip duration ≥ 35 λεπτά: long trip duration

trip_distance	< 5 μίλια: short trip 5 ≤ μίλια <10: regular trip 10 ≤ μίλια <20: medium trip ≥ 20 μίλια: long trip
extra	< 5 δολάρια: low extras 5 ≤ δολάρια <10: regular extras ≥ 10 δολάρια: lots of extras
total_amount	< 10 δολάρια: low fair 10 ≤ δολάρια <20: regular fair 20 ≤ δολάρια <30: high fair ≥ 30 δολάρια: very high fair
hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night
passenger_count	< 2 άτομα: solo trip 2 ≤ άτομα < 3: duo trip ≥ 3 άτομα: party trip

Πίνακας 14. Κατηγοριοποίηση χαρακτηριστικών για ταξί της Νέας Υόρκης.

Uber Database – New York City	
Χαρακτηριστικό	**Κατηγοριοποίηση**
hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night

Πίνακας 15. Κατηγοριοποίηση χαρακτηριστικών για Uber της Νέας Υόρκης.

- *Ενοικιαζόμενα Ποδήλατα*

Bike Database – New York City	
Χαρακτηριστικό	**Κατηγοριοποίηση**
trip_duration	< 5 λεπτά: short trip duration 5 ≤ λεπτά <20: regular trip duration

	20 ≤ λεπτά <35: medium trip duration ≥ 35 λεπτά: long trip duration
Start Hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night
End Hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night
age	< 18 ηλικία: teenagers 18 ≤ ηλικία <30: young adults 30 ≤ ηλικία <40: adults 40≤ ηλικία <60: middle age ≥ 60 ηλικία: seniors

Πίνακας 16. Κατηγοριοποίηση χαρακτηριστικών για ποδήλατα της Νέας Υόρκης.

Bike Database – London	
Χαρακτηριστικό	**Κατηγοριοποίηση**
Total duration (ms)	< 5 λεπτά: short trip duration 5 ≤ λεπτά <20: regular trip duration 20 ≤ λεπτά <35: medium trip duration ≥ 35 λεπτά: long trip duration
Start Hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night
End Hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night

Πίνακας 17. Κατηγοριοποίηση χαρακτηριστικών για ποδήλατα του Λονδίνου.

Bike Database – Helsinki	
Χαρακτηριστικό	**Κατηγοριοποίηση**
duration (sec)	< 5 λεπτά: short trip duration 5 ≤ λεπτά <20: regular trip duration 20 ≤ λεπτά <35: medium trip duration ≥ 35 λεπτά: long trip duration
Start Hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night
End Hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night
distance (m)	< 0,5 χλμ.: very short distance 0,5 ≤ χλμ. <1,5: short distance 1,5 ≤ χλμ. <3,0: regular distance ≥ 3,0 χλμ.: long distance
avg_speed (km/h)	< 3 χλμ./ώρα: very slow 3 ≤ χλμ./ώρα <10: moderate ≥ 10 χλμ./ώρα: fast
Air temperature (degC)	< 5 °C: very cold 5 ≤ °C < 15: cold 15 ≤ °C < 25: regular temp 25 ≤ °C < 32: warm ≥ 32 °C: very warm

Πίνακας 18. Κατηγοριοποίηση χαρακτηριστικών για ποδήλατα του Ελσίνκι.

- *Μέσα Μαζικής Μεταφοράς*

Bus Database – New York City	
Χαρακτηριστικό	**Κατηγοριοποίηση**
Hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night

Πίνακας 19. Κατηγοριοποίηση χαρακτηριστικών για λεωφορεία της Νέας Υόρκης.

Bus Database – Rio De Janeiro	
Χαρακτηριστικό	**Κατηγοριοποίηση**
Hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night
speed	< 5 χλμ/ώρα: very slow 5 ≤ χλμ/ώρα < 15: slow 15 ≤ χλμ/ώρα < 30: regular speed 30 ≤ χλμ/ώρα < 45: fast ≥ 45 χλμ/ώρα: very fast

Πίνακας 20. Κατηγοριοποίηση χαρακτηριστικών για λεωφορεία του Ρίο Ντε Τζανέιρο.

Metro Database – New York City	
Χαρακτηριστικό	**Κατηγοριοποίηση**
Hour	< 7 ώρα: night 7 ≤ ώρα <17: working hours 17 ≤ ώρα <21: afternoon ≥ 21 ώρα: night
Entries	< 400 άτομα: low people 400 ≤ άτομα < 1000: regular people 1000 ≤ άτομα < 2000: high traffic-people ≥ 2000 άτομα: very high traffic-people
Exits	< 400 άτομα: low people 400 ≤ άτομα < 1000: regular people 1000 ≤ άτομα < 2000: high traffic-people ≥ 2000 άτομα: very high traffic-people

Πίνακας 21. Κατηγοριοποίηση χαρακτηριστικών για μετρό της Νέας Υόρκης.

3.2.8 Βήμα 8ο – Αλγόριθμος FP-Growth για εύρεση μοτίβων

Σε αυτό το βήμα που είναι και το τελευταίο της ανάλυσης θα τρέξουμε τον αλγόριθμο FP-Growth για να δούμε ποια είναι τα συχνά μοτίβα στα δεδομένα μας για τα χαρακτηριστικά που έχουμε επιλέξει. Ο αλγόριθμος FP-Growth (Frequent Pattern

Growth) είναι ένας αποδοτικός και επεκτάσιμος αλγόριθμος για την εξαγωγή συχνών μοτίβων από δεδομένα. Χρησιμοποιείται κυρίως στην εξόρυξη δεδομένων για να αναγνωρίσει συσχετίσεις και να δημιουργήσει κανόνες συσχέτισης (association rules). Η βασική ιδέα του FP-Growth είναι να αποφύγει την επαναλαμβανόμενη σάρωση της βάσης δεδομένων, κάτι που κάνει άλλους αλγόριθμους, όπως ο Apriori, λιγότερο αποδοτικούς.

Κύρια βήματα του αλγόριθμου FP-Growth:

1. Κατασκευή του FP-Tree:
   - Σάρωση της βάσης δεδομένων: Μετράει τη συχνότητα εμφάνισης κάθε στοιχείου στη βάση δεδομένων.
   - Φιλτράρισμα και ταξινόμηση: Απορρίπτει τα στοιχεία που δεν πληρούν το ελάχιστο κατώφλι υποστήριξης (support) και ταξινομεί τα υπόλοιπα σε φθίνουσα σειρά συχνότητας.
   - Δημιουργία του FP-Tree: Δημιουργεί το δέντρο FP-Tree με μία δεύτερη σάρωση της βάσης δεδομένων. Τα στοιχεία εισάγονται στο δέντρο με βάση την ταξινόμησή τους ως προς την συχνότητα εμφάνισης, ενώ οι κοινές διαδρομές μοιράζονται κόμβους.
2. Εξαγωγή συχνών μοτίβων από το FP-Tree:
   - Ανίχνευση των προτύπων από το FP-Tree: Εξετάζει τις διαδρομές του FP-Tree για να εξάγει τα συχνά μοτίβα.
   - Διαίρεση του προβλήματος: Για κάθε στοιχείο που εξάγεται, δημιουργείται ένα μικρότερο, εξειδικευμένο υποδέντρο (conditional FP-Tree), και επαναλαμβάνεται η διαδικασία.

Πλεονεκτήματα του FP-Growth:

- Αποδοτικότητα: Ο αλγόριθμος FP-Growth είναι γενικά πιο αποδοτικός από τον Apriori γιατί δεν επαναλαμβάνει συνεχείς σαρώσεις της βάσης δεδομένων και περιορίζει το πρόβλημα δημιουργώντας μικρότερα δέντρα.
- Κλιμακωσιμότητα: Λειτουργεί καλά και με μεγάλες βάσεις δεδομένων και με μεγάλα σύνολα στοιχείων.

Συνολικά, τα χαρακτηριστικά για τα οποία τρέξαμε τον αλγόριθμο για κάθε κατηγορία των συνόλων δεδομένων είναι:

- *Μισθωμένα οχήματα – Ταξί*

Taxi Database – Chicago	**Taxi Database – New York City**	**Uber Database – New York City**
Trip Seconds	passenger_count	Weekday
Trip Miles	trip_distance	Hour
Extras	trip_duration_minutes	Pickup Borough
Trip Total	extra	
Payment Type	total_amount	
Company	hour	
Pickup Neighborhood	payment_type	
Dropoff Neighborhood		
hour		

Πίνακας 22. Χαρακτηριστικά για εύρεση μοτίβων στην κατηγορία μισθωμένα οχήματα-ταξί.

- *Ενοικιαζόμενα Ποδήλατα*

Bike Database – New York City	**Bike Database – London**	**Bike Database – Helsinki**
gender	Total duration (ms)	distance (m)
age	Start Hour	duration (sec)
trip_duration	Start Weekday	avg_speed (km/h)
Start Hour	End Hour	Air temperature (degC)
Start Weekday	End Weekday	Start Hour
End Hour		Start Weekday
End Weekday		End Hour
		End Weekday

Πίνακας 23. Χαρακτηριστικά για εύρεση μοτίβων στην κατηγορία ενοικιαζόμενα ποδήλατα.

- *Μέσα Μαζικής Μεταφοράς*

Bus Database – New York City	**Bus Database – Rio De Janeiro**	**Metro Database – New York City**
ArrivalProximityText	speed	Stop Name
Hour	Hour	Line
Weekday	Weekday	Structure
Pickup Borough	Neighborhood	Borough
		Neighborhood
		Entries
		Exits
		Month
		Weekday
		Hour

Πίνακας 24. Χαρακτηριστικά για εύρεση μοτίβων στην κατηγορία μέσα μαζικής μεταφοράς.

3.3 Εργαλεία που χρησιμοποιήθηκαν

Για την επεξεργασία των δεδομένων και την εξαγωγή των αποτελεσμάτων χρησιμοποιήσαμε τα παρακάτω εργαλεία και βιβλιοθήκες:

Python: Η Python είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου, γνωστή για τη σαφή και ευανάγνωστη σύνταξή της, η οποία διευκολύνει την ταχεία ανάπτυξη λογισμικού. Η ευρεία γκάμα βιβλιοθηκών της Python, όπως η NumPy, Pandas και Matplotlib, την καθιστά ιδανική για εφαρμογές σε Επιστήμη Δεδομένων και Μηχανική Μάθηση.

Jupyter Notebook: Το Jupyter Notebook είναι ένα ανοιχτού κώδικα περιβάλλον που επιτρέπει στους χρήστες να δημιουργούν και να μοιράζονται έγγραφα που περιέχουν ζωντανό κώδικα, εξισώσεις, οπτικοποιήσεις και αφήγηση κειμένου. Είναι ιδιαίτερα χρήσιμο για την εκπαίδευση και την έρευνα, καθώς επιτρέπει τη διαδραστική εξερεύνηση δεδομένων και την οπτικοποίηση των αποτελεσμάτων με δυναμικό τρόπο.

Apache Spark: Το Apache Spark είναι μια πλατφόρμα ανοιχτού κώδικα για κατανεμημένο υπολογισμό, σχεδιασμένη για να επεξεργάζεται μεγάλους όγκους δεδομένων με ταχύτητα και ευελιξία. Παρέχει APIs υψηλού επιπέδου για Java, Scala, Python και R, και υποστηρίζει επεξεργασία δεδομένων σε πραγματικό χρόνο μέσω του Spark Streaming, καθιστώντας το ιδανικό για εφαρμογές big data.

scikit-learn: Μια δημοφιλής βιβλιοθήκη για Μηχανική Μάθηση στη γλώσσα Python, που προσφέρει μια μεγάλη γκάμα αλγορίθμων για ταξινόμηση, παλινδρόμηση, συμπίεση δεδομένων και ανάλυση συστάδων.

pandas: Μια ισχυρή βιβλιοθήκη για διαχείριση και ανάλυση δεδομένων, που παρέχει δομές δεδομένων όπως DataFrame για αποδοτική αποθήκευση και χειρισμό πινάκων και σειρών δεδομένων.

numpy: Βασική βιβλιοθήκη για αριθμητικούς υπολογισμούς στην Python, που υποστηρίζει για πολυδιάστατους πίνακες και έναν μεγάλο αριθμό μαθηματικών συναρτήσεων.

matplotlib: Μια βιβλιοθήκη για δημιουργία στατικών, διαδραστικών και κινούμενων γραφημάτων στην Python, εξαιρετικά ευέλικτη και χρήσιμη για την οπτικοποίηση δεδομένων.

seaborn: Μια βιβλιοθήκη βασισμένη στη matplotlib, σχεδιασμένη για τη δημιουργία στατιστικών γραφημάτων με πιο ελκυστικά και ενημερωτικά σχέδια, που διευκολύνει την ανάλυση και οπτικοποίηση δεδομένων.

mlxtend.frequent_patterns: Υπο-βιβλιοθήκη του mlxtend, προσφέρει εργαλεία για εξόρυξη συχνών μοτίβων σε δεδομένα, όπως αλγορίθμους για ανάλυση συνάφειας (association rule mining).

math: Ενσωματωμένη βιβλιοθήκη της Python που παρέχει βασικές μαθηματικές συναρτήσεις και σταθερές, όπως τριγωνομετρικές, λογαριθμικές και εκθετικές συναρτήσεις.

time: Ενσωματωμένη βιβλιοθήκη της Python για διαχείριση και χειρισμό χρόνου, επιτρέποντας λειτουργίες όπως μέτρηση χρόνου, καθυστερήσεις και διαμόρφωση χρονικών αντικειμένων.

geopandas: Επέκταση της pandas που προσθέτει υποστήριξη για γεωχωρικά δεδομένα, επιτρέποντας την ανάγνωση, γραφή και ανάλυση γεωχωρικών δεδομένων σε μορφή DataFrame.

shapely.geometry: Βιβλιοθήκη για χειρισμό και ανάλυση γεωμετρικών σχημάτων στη Python, που παρέχει εργαλεία για τη δημιουργία και την εκτέλεση γεωμετρικών πράξεων.

folium: Βιβλιοθήκη για την κατασκευή διαδραστικών χαρτών στη Python, που χρησιμοποιεί την τεχνολογία Leaflet.js για τη δημιουργία χαρτών από δεδομένα γεωχωρικού τύπου.

4

Εφαρμογή και Αποτελέσματα

4.1 Περιγραφική Στατιστική (Descriptive Statistics)

Εδώ θα παρουσιάσουμε τα αποτελέσματα που έχουμε στα διάφορα μέτρα Περιγραφική Στατιστικής που υπολογίσαμε και παρουσιάσαμε στο υποκεφάλαιο 3.2.3 *Βήμα 3* της ανάλυσης μας. Επίσης θυμίζουμε ότι τα στατιστικά αυτά υπολογίστηκαν μετά τον χειρισμό των μηδενικών και ακραίων τιμών που έγινε στο *Βήμα 2*. Έτσι, έχουμε για κάθε κατηγορία μετακίνησης:

- *Μισθωμένα Οχήματα – Ταξί*

Εδώ για το σύνολο δεδομένων ταξιδιού με Uber της Νέας Υόρκης δεν έχουμε αριθμητικές μεταβλητές και επομένως δεν έχουμε τα αντίστοιχα στατιστικά. Τα σύνολα δεδομένων ταξιδιού με Ταξί του Σικάγο και της Νέας Υόρκης περιγράφονται στατιστικά στον παρακάτω πίνακα με τα χαρακτηριστικά (features) που είναι ίδια να παρουσιάζονται πρώτα.

Στατιστικό μέτρο / **Χαρακτηριστικό**	*Πλήθος δεδομένων (count)*	*Μέσος όρος (Mean)*	*Διάμεσος (Median)*	*Επικρατούσα τιμή (Mode)*	*Τυπική Απόκλιση (Standard Deviation)*	*Ελάχιστο (Min.)*	*25%*	*50%*	*75%*	*Μέγιστο (max)*	*Συμμετρία (Skewness)*	*Κύρτωση (Kurtosis)*
Chicago: Trip Miles	842.119	6,264	2,94	0,0	6,521	0,0	0,90	2,94	11,48	39,86	0,904	- 0,181
New York: trip_distance	80.364	4,326	2.64	0,0	4,594	0,0	1,33	2,64	5,75	38,75	2,038	4,923
Chicago: Trip Seconds	842.119	17,70	14,80	0,0	12,79	0,0	7,57	14,8	25,88	60,0	0,830	0,143
New York: trip_duration_minutes	80.364	18,13	14,73	14,0	12,87	0,0	8,51	14,7	24,97	60,0	1,067	0,587
Chicago: Fare	842.119	21,07	15,00	9,0	15,48	0,0	8,0	15,0	32,75	100,0	0,810	0,005
New York: fare_amount	80.364	19,12	15,50	7,0	13,37	0,0	9,0	15,5	25,46	100,0	1,338	1,922
Chicago: Extras	842.119	1,66	0,0	0,0	4,07	0,0	0,0	0,0	2,0	40,0	4,852	28,811
New York: extra	80.364	1,15	0,50	0,0	1,36	0,0	0,0	0,5	2,75	8,25	0,985	0,502
Chicago: Tips	842.119	2,63	0,0	0,0	3,77	0,0	0,0	0,0	4,0	80,0	1,915	7,954

New York: tip_amount	80.364	1,05	0,0	0,0	2,15	0,0	0,0	0,0	1,75	87,71	5,616	102,137
Chicago: Tolls	842.119	0,02	0,0	0,0	0,35	0,0	0,0	0,0	0,0	39,0	53,341	3711,71
New York: tolls_amount	80.364	0,52	0,0	0,0	1,80	0,0	0,0	0,0	0,0	27,5	3,503	12,873
Chicago: Trip Total	842.119	25,54	18,0	3,25	19,68	0,0	10,0	18,0	38,25	100,0	0,971	0,122
New York: total_amount	80.364	22,84	19,44	7,8	14,87	0,0	11,4	19,4	29,80	100,0	1,295	1,820
New York: passenger _count	49.247	1,31	1	1	0,99	0	1	1	1	32	4,131	30,441
New York: Mta_tax	80.364	0,30	0,50	0,50	0,24	0,0	0,0	0,50	0,50	0,50	- 0,375	- 1,859
New York: Improvement _surcharge	80.364	0,30	0,30	0,30	0,02	0,0	0,30	0,30	0,30	0,30	- 16,193	260,247

Πίνακας 25. Αποτελέσματα στατιστικών για τα χαρακτηριστικά της κατηγορίας μισθωμένα οχήματα-ταξί.

Σημείωση: Έχουμε διατηρήσει τα αρχικά ονόματα των χαρακτηριστικών αλλά έχουμε μετατρέψει τα συγκρίσιμα χαρακτηριστικά στις ίδιες μονάδες μέτρησης. Για παράδειγμα τα χαρακτηριστικά του χρόνου του ταξιδιού είναι όλα σε λεπτά.

Για τα χαρακτηριστικά (features) που αφορούν την διαδρομή δηλαδή την απόσταση και το χρόνο του ταξιδιού σε λεπτά μπορούμε να παρατηρήσουμε από τα αποτελέσματα ότι όσον αφορά το χρόνο του ταξιδιού είναι αρκετά κοντά στην περίπτωση της Νέας Υόρκης και του Σικάγο. Ο μέσος όρος και στις δύο περιπτώσεις είναι γύρω στα 18 λεπτά με τυπική απόκλιση γύρω στα 12,8 λεπτά. Τα μέτρα των τεταρτημόριων είναι επίσης πολύ κοντά και όσον αφορά την κατανομή βλέπουμε ότι και στις δύο περιπτώσεις όσον αφορά την συμμετρία έχουμε μία ελαφριά ασυμμετρία προς τα δεξιά και όσον αφορά την κύρτωση στην περίπτωση της Νέας Υόρκης έχουμε μία πιο υψηλή κορυφή και πιο απλωμένες ουρές. Η μόνη σημαντική διαφοροποίηση είναι ότι στο Σικάγο η επικρατούσα τιμή είναι το 0 που σημαίνει ότι έχουμε πολλά δεδομένα με αυτή την τιμή. Τώρα όσον αφορά το χαρακτηριστικό της απόστασης που διανύθηκε στο ταξίδι βλέπουμε από τα πρώτα 3 στατιστικά μέτρα

μετά από αυτό του πλήθους των δεδομένων ότι ενώ η διάμεσος και επικρατούσα τιμή έχουν παρόμοια τιμή ο μέσος όρος σε μίλια για το σύνολο δεδομένων του Σικάγο είναι αισθητά μεγαλύτερος, κατά περίπου 1,5 μίλια από αυτόν την Νέας Υόρκης. Αντίστοιχα, παρατηρείται και για την τυπική απόκλιση ενώ τα υπόλοιπα μέτρα είναι παρόμοια. Από τα μέτρα κατανομής παρατηρούμε ότι και στις δύο περιπτώσεις έχουμε ασύμμετρη κατανομής προς τα δεξιά, η οποία όμως είναι πιο έντονη στο σύνολο δεδομένων της Νέα Υόρκης, ενώ όσον αφορά την κύρτωση, η κατανομή του χαρακτηριστικού στο σύνολο δεδομένων του Σικάγο είναι σχετικά επίπεδη, ενώ η αντίστοιχη κατανομή στο σύνολο δεδομένων της Νέας Υόρκης παρουσιάζει μία σχετικά ψηλή κορυφή. Ένα αρχικό συμπέρασμα που μπορούμε να βγάλουμε εδώ είναι ότι τα ταξίδια με ταξί είναι παρόμοια ως προς τον χρόνο στις μεγάλες πόλεις της Αμερικής. Η διαφορά που υπάρχει στην μέση απόσταση για ένα ταξίδι με ταξί στις δύο αυτές πόλεις οφείλεται στην φύση αυτών, δηλαδή μία πιο πυκνοκατοικημένη πόλη όπως η Νέα Υόρκη έχει περισσότερη κίνηση και ένα μικρότερο ταξίδι κατά μέσο όρο διαρκεί περισσότερο.

Για τα χαρακτηριστικά που αφορούν το κόστος της διαδρομής παρατηρούμε ομοίως ότι σε γενικές γραμμές βρίσκονται αρκετά κοντά τα στατιστικά μέτρα στις δύο πόλεις. Τα χαρακτηριστικά του καθαρού κόστος (fare, fare_amount) και του συνολικού κόστους της διαδρομής (Trip Total, total_amount) ακολουθούν κατανομές με παρόμοιες τιμές στα στατιστικά μέτρα, δηλαδή μοιάζουν. Το μέσο κόστος του ταξιδιού είναι ελαφρώς υψηλότερο για το Σικάγο καθώς παρατηρούμε ότι το μέσο καθαρό κόστος της διαδρομής είναι περίπου δύο δολάρια υψηλότερο από το μέσο καθαρό κόστος της διαδρομής με ταξί στην Νέα Υόρκη και το μέσο συνολικό κόστος αφού έχουν συμπεριληφθεί όλες οι υπόλοιπες χρεώσεις, είναι περίπου 3 δολάρια υψηλότερο από το μέσο συνολικό κόστος της διαδρομής με ταξί στην Νέα Υόρκη. Και στις δύο πόλεις, το μέσο κόστος είναι κοντά στα 20 δολάρια που είναι σχετικά υψηλό, όμως το καθαρό και συνολικό κόστος ταξιδιού με ταξί στο Σικάγο έχει μεγαλύτερη τυπική απόκλιση σε σχέση με την Νέα Υόρκη, επομένως οι τιμές έχουν μεγαλύτερη διακύμανση εκεί. Εδώ η κατανομή είναι ασύμμετρη προς τα δεξιά που σημαίνει ότι οι διαδρομές με πιο μικρό κόστος είναι συχνότερες, με την περίπτωση της Νέας Υόρκης να είναι λίγο πιο έντονη αυτή η συμμετρία. Επίσης όσον αφορά την κύρτωση το καθαρό και συνολικό κόστος ταξιδιού με ταξί στο Σικάγο ακολουθεί μία σχετικά επίπεδη κατανομή, σε σχέση με την Νέα Υόρκη, όπου τα αντίστοιχα

χαρακτηριστικά παρουσιάζουν πιο ψηλές κορυφές.. Στα υπόλοιπα χαρακτηριστικά που αφορούν τις επιμέρους χρεώσεις παρατηρούμε σε γενικές γραμμές ότι οι τιμές είναι σχετικά χαμηλές κοντά στα 1-2 δολάρια και σε μεγάλα ποσοστά του συνόλου δεδομένων είναι κοντά στο μηδέν, ενώ σε χαμηλά ποσοστά του συνόλου δεδομένων, έχουν υψηλές τιμές με αποτέλεσμα να ανεβάζουν την μέση τιμή (διόδια, tips, extras). Επομένως παρατηρείται έντονη ασυμμετρία και κυρτότητα στις κατανομές που ακολουθούν τα χαρακτηριστικά αυτά. Τα συμπεράσματα που μπορούμε να βγάλουμε είναι ότι οι διαδρομές επειδή είναι σχετικά κοντινές δεν βγαίνουν εκτός ορίων της πόλης συχνά για αυτό και τα διόδια στις περισσότερες διαδρομές είναι μηδενικά ενώ σε λίγες είναι αρκετά υψηλά. Επίσης βλέπουμε ότι οι τιμές των φιλοδωρημάτων είναι κατά μέσο όρο πολύ χαμηλές καθώς στις περισσότερες διαδρομές είναι μηδενικές ενώ σε χαμηλά ποσοστά του συνόλου δεδομένων, είναι πολύ υψηλές. Αυτό είναι αναμενόμενο διότι γνωρίζουμε ότι κατά την συλλογή των δεδομένων τα φιλοδωρήματα που δίνονταν με μετρητά δεν καταγράφηκαν. Άρα συμπεραίνουμε ότι πρώτον το χαρακτηριστικό των φιλοδωρημάτων δεν ανταποκρίνεται πολύ στην πραγματικότητα και δεύτερον ότι όταν ο τρόπος πληρωμής είναι με κάρτα τα φιλοδωρήματα είναι αρκετά υψηλά.

Τα χαρακτηριστικά Mta_tax και Improvement_surcharge που αφορούν μόνο το σύνολο δεδομένων ταξιδιού με ταξί της Νέας Υόρκης, περιγράφουν κάποιες συγκεκριμένες χαμηλές χρεώσεις που είτε υπάρχουν είτε όχι στην διαδρομή και για αυτό και τα αντίστοιχα στατιστικά είναι σχετικά flat. Για το πλήθος των επιβατών (passenger_count) για την Νέα Υόρκη παρατηρούμε τα αποτελέσματα των μέτρων κεντρικής τάσης και διασποράς που τείνουν προς την τιμή 1, καθώς και την υψηλή ασυμμετρία και κύρτωση που παρουσιάζει η κατανομή του χαρακτηριστικού, και συμπεραίνουμε εύκολα ότι τα περισσότερα ταξίδια με ταξί έχουν μόνο έναν επιβάτη, κατά συντριπτική πλειοψηφία.

- *Ενοικιαζόμενα Ποδήλατα*

Σε αυτή την κατηγορία μετακίνησης για τα τρία σύνολα δεδομένων έχουμε ένα κοινό αριθμητικό χαρακτηριστικό το οποίο είναι η διάρκεια του ταξιδιού το οποίο θα αναφέρουμε πρώτο. Όπως αναφέραμε και στην προηγούμενη κατηγορία τα χαρακτηριστικά έχουν μετατραπεί όλα σε λεπτά, ώστε να μπορούμε εύκολα να τα συγκρίνουμε. Τα αποτελέσματα είναι στον παρακάτω πίνακα.

Στατιστικό μέτρο / **Χαρακτηριστικό**	*Πλήθος δεδομένων (count)*	*Μέσος όρος (Mean)*	*Διάμεσος (Median)*	*Επικρατούσα τιμή (Mode)*	*Τυπική Απόκλιση (Standard Deviation)*	*Ελάχιστο (Min.)*	*25%*	*50%*	*75%*	*Μέγιστο (max)*	*Συμμετρία (Skewness)*	*Κύρτωση (Kurtosis)*
New York: trip_duration	$1{,}58 \cdot 10^6$	13,09	10,2	5,1	9,58	1,02	6,1	10,2	17,47	60	1,414	2,034
London: Total duration (ms)	749658	16,36	13,51	4,91	11,41	0,02	7,9	13,5	21,91	60	1,272	1,555
Helsinki: duration (sec)	$1{,}96 \cdot 10^7$	11,54	9,63	0,28	8,05	0,02	5,68	9,63	15,82	60	1,361	3,071
New York: age	$1{,}58 \cdot 10^6$	37,87	35	30	11,02	16	29	35	46	65	0,561	- 0,632
Helsinki: distance (m)	$1{,}96 \cdot 10^7$	2080	1725	0	1520	0	995	1725	2830	10000	1,179	1,576
Helsinki: avg_speed (km/h)	$1{,}96 \cdot 10^7$	0,178	0,187	0,0	0,076	0,0	0,148	0,187	0,220	37,33	44,799	14025,9
Helsinki: Air temperature (degC)	$1{,}96 \cdot 10^7$	15,6	16,4	17,1	5,49	- 5,2	12,3	16,4	19,3	32,9	- 0,386	- 0,022

Πίνακας 26. Αποτελέσματα στατιστικών για τα χαρακτηριστικά της κατηγορίας ενοικιαζόμενα ποδήλατα.

Τα χαρακτηριστικά που αφορούν την διάρκεια του ταξιδιού στα τρία σύνολα δεδομένων (dataset) παρουσιάζουν αρκετά όμοιες κατανομές. Η διάρκεια του ταξιδιού με ποδήλατο έχει παρόμοια μέση τιμή και διάμεσο για το Ελσίνκι και την Νέα Υόρκη, ενώ είναι κατά 25% περίπου μεγαλύτερη στην περιοχή του Λονδίνου. Άρα μία πρώτη παρατήρηση που μπορούμε να κάνουμε ότι στο Λονδίνο γίνονται μεγαλύτερα ταξίδια που μπορεί να οφείλεται στο ότι δεν είναι τόσο πυκνοκατοικημένο όσο η Νέα Υόρκη και ο μέσος χρήστης διανύει λίγο μεγαλύτερη απόσταση για να φτάσει στον προορισμό του. Κάτι ανάλογο με το Λονδίνο θα περιμέναμε και για το Ελσίνκι που είναι πιο αραιοκατοικημένο και άρα είναι λίγο μεγαλύτερες οι αποστάσεις, κάτι που δεν συμβαίνει, πιθανόν λόγω των καιρικών συνθηκών που επικρατούν το μεγαλύτερο μέρος του χρόνου, που ενδεχομένως κάνει τους κατοίκους της περιοχής να επιλέγουν να μένουν σε περιοχή κοντινή με την

εργασία τους. Επιπλέον, ο τουρισμός στην περιοχή του Ελσίνκι είναι σημαντικά χαμηλότερος σε σχέση με το Λονδίνο, επομένως οι έκτακτοι ταξιδιώτες στην πόλη που ταξιδεύουν μεγαλύτερες αποστάσεις για να επισκεφθούν αξιοθέατα πιθανόν να επηρεάζουν την κατανομή της διάρκειας ταξιδιού ανάλογα. Τα υπόλοιπα στατιστικά μέτρα δείχνουν μία παρόμοια κατανομή για το χαρακτηριστικό της διάρκειας του ταξιδιού με ποδήλατο και για τα τρία σύνολα δεδομένων, όπου όσον αφορά την συμμετρία έχουμε ασυμμετρία προς τα δεξιά αρκετά κοντά και στις τρεις περιπτώσεις. Η κύρτωση είναι σχετικά υψηλή και στις τρεις , όπου η κορυφή στην κατανομή που ακολουθεί η διάρκεια ταξιδιού στο Ελσίνκι είναι υψηλότερη, με την Νέα Υόρκη και τέλος το Λονδίνο να ακολουθούν.

Στο σύνολο δεδομένων της Νέας Υόρκης όπου έχουμε χαρακτηριστικό για την ηλικία παρατηρούμε μέση τιμή στα 35 χρόνια και τυπική απόκλιση στο 11, επομένως ο μεγαλύτερος αριθμός των χρηστών βρίσκεται ανάμεσα στα 25 με 45 χρόνια. Αποτέλεσμα που είναι θετικό διότι θα περιμέναμε η υιοθέτηση ενός μέσου μεταφοράς όπως το ποδήλατο να είναι πιο δημοφιλής μόνο σε πολύ μικρές ηλικίες έως και 25-28. Από τα υπόλοιπα στατιστικά μέτρα βλέπουμε ότι έχουμε μία ομαλή καμπύλη χωρίς πολύ μεγάλη κορυφή και με μια σχετική συμμετρία.

Τέλος, για το σύνολο δεδομένων του Ελσίνκι έχουμε και χαρακτηριστικό που αφορά την απόσταση που διανύθηκε σε μέτρα. Έχουμε μέσο όρο τα 2χλμ. με τυπική απόκλιση το 1.5χλμ. περίπου, άρα οι περισσότερες τιμές κυμαίνονται από τα 500μ. έως τα 3.5χλμ. Παρατηρούμε από τα στατιστικά μέτρα τις κατανομής, ότι είναι παρόμοια με αυτήν της διάρκειας του ταξιδιού με ποδήλατο για την ίδια περιοχή του Ελσίνκι, όπως είναι αναμενόμενο. Για το χαρακτηριστικό της μέσης ταχύτητας ταξιδιού με ποδήλατο, από τα μέτρα Περιγραφικής Στατιστικής, βλέπουμε ότι έχουμε πολύ χαμηλές τιμές για την ταχύτητα, κάτω από 1 χλμ./ώρα, για το μεγαλύτερο ποσοστό των εγγραφών στο σύνολο δεδομένων, το οποίο φαίνεται και από τις πολύ υψηλές τιμές για την συμμετρία και την κύρτωση. Το μόνο συμπέρασμα που μπορούμε να εξάγουμε εδώ είναι ότι είτε οι χρήστες κάνουν μεγάλες στάσεις αφού νοικιάσουν το ποδήλατο, που παρεμβάλλονται της επιθυμητής διαδρομής, που οδηγεί σε μικρότερη μέση ταχύτητα για την ίδια απόσταση, είτε ότι πολλοί νοικιάζουν το ποδήλατο δοκιμαστικά διανύοντας διαδρομές σχετικά τοπικά με πολύ μικρές ταχύτητες. Αξίζει να σημειωθεί ότι εφόσον το μέτρο της ταχύτητας προκύπτει από την διανυόμενη απόσταση προς την διάρκεια του ταξιδιού του ποδηλάτου, που

μετρείται με στοιχεία GPS και στοιχεία αλλαγής ποδηλάτου ανά σταθμό, οποιοδήποτε συμπέρασμα εξάγεται με επιφύλαξη για την ποιότητα των δεδομένων ταχύτητας. Τέλος, το χαρακτηριστικό της εξωτερικής θερμοκρασίας αέρα κατά την διαδρομή του ταξιδιού με ποδήλατο ακολουθεί μία σχεδόν κανονική κατανομή με κορυφή γύρω στους 15 °C.. Τα αποτελέσματα αυτά θα επιβεβαιώσουμε και στο επόμενο υποκεφάλαιο όπου θα δούμε τα αντίστοιχα γραφήματα.

- *Μέσα Μαζικής Μεταφοράς*

Εδώ στην τελευταία μας κατηγορία, δεν έχουμε κάποια χαρακτηριστικά που να είναι όμοια για όλα τα σύνολα δεδομένων ταξιδιού με ΜΜΜ, επομένως να μπορούν να συγκριθούν για κάθε περιοχή. Έχουμε υπολογίσει τα μέτρα Περιγραφικής Στατιστικής για τα διαφορετικά αριθμητικά χαρακτηριστικά του κάθε συνόλου δεδομένων που φαίνονται στον παρακάτω πίνακα.

Στατιστικό μέτρο / **Χαρακτηριστικό**	*Πλήθος δεδομένων (count)*	*Μέσος όρος (Mean)*	*Διάμεσος (Median)*	*Επικρατούσα τιμή (Mode)*	*Τυπική Απόκλιση (Standard Deviation)*	*Ελάχιστο (Min.)*	*25%*	*50%*	*75%*	*Μέγιστο (max)*	*Συμμετρία (Skewness)*	*Κύρτωση (Kurtosis)*
New York Subway: Entries	$4{,}59 \cdot 10^6$	1219	528	0	1997	0	164	528	1400	40374	4,299	29,725
New York Subway: Exits	$4{,}59 \cdot 10^6$	971	397	0	1683	0	140	397	1040	40697	4,853	41,139
Rio Bus: speed	$5{,}92 \cdot 10^7$	17,80	14,0	0,0	18,43	0,0	0,0	14,0	31,0	79,82	0,782	- 0,332
New York Bus: DistanceFromStop	$2{,}65 \cdot 10^6$	225,5	89,0	0,0	1003,77	0,0	22,0	89,0	199,0	9999	18,831	449,341

Πίνακας 27. Αποτελέσματα στατιστικών για τα χαρακτηριστικά της κατηγορίας μέσα μαζικής μεταφοράς.

Εδώ για το σύνολο δεδομένων του μετρό της Νέας Υόρκης έχουμε παρόμοια αποτελέσματα για τα χαρακτηριστικά των ατόμων που μπαίνουν και βγαίνουν από μία στάση. Έχουμε παρόμοια στατιστικά μέτρα τόσο όσον αφορά τα μέτρα διασποράς όσο και τα μέτρα που χαρακτηρίζουν την καμπύλη. Για τα μέτρα κεντρικής τάσης υπάρχει μία μικρή διαφοροποίηση με το χαρακτηριστικό που

μετράει την είσοδο στην στάση του μετρό να έχει μέση τιμή υψηλότερη από αυτήν του χαρακτηριστικού που μετράει την έξοδο, κατά περίπου 20%. Αυτά τα αποτελέσματα δεν είναι αναμενόμενα καθώς λογικό είναι ότι όσοι μπαίνουν στο μετρό τόσοι πρέπει και να βγαίνουν συνολικά. Λογικό είναι επίσης ότι σε πολυσύχναστους κεντρικούς σταθμούς, μπορεί να εισέρχονται περισσότεροι επιβάτες απ' όσους εξέρχονται, αν έχουν χρησιμοποιήσει διαφορετικό μέσο μετακίνησης για την άφιξή τους στους σταθμούς αυτούς, και όχι το μετρό. Οι μικρές διαφοροποιήσεις που έχουμε ανάμεσα στην είσοδο και την έξοδο μπορεί να οφείλονται στο ότι ένα ποσοστό λόγω βιασύνης/αδιαφορίας μπορεί να μην χτυπάει το εισιτήριο για την έξοδο εφόσον έχει πληρώσει και το έχει χτυπήσει στην είσοδο περνώντας πίσω από κάποιον άλλο χρήστη. Δεν υπάρχει πρόσθετη χρέωση κατά την έξοδο, και ενώ αν δεν χτυπήσεις εισιτήριο βγαίνοντας αυτό μπορεί να θεωρηθεί παραβίαση των κανόνων, ο MTA τυπικά δεν επιβάλει πρόστιμα γι’ αυτό.

Για το σύνολο δεδομένων λεωφορείων του Ρίο μελετούμε το χαρακτηριστικό “speed” που δηλώνει την ταχύτητα του λεωφορείου την στιγμή της καταγραφής της θέσης του. Από το αποτελέσματα βλέπουμε ότι τα λεωφορεία κινούνται με ταχύτητες που κυμαίνονται στο μεγαλύτερο ποσοστό από 0 έως 40χλμ/ώρα γεγονός που είναι και αναμενόμενο για ένα τέτοιο μέσο μεταφοράς. Η καμπύλη της ταχύτητας έχει μία ασυμμετρία προς τα δεξιά χωρίς πολύ έντονες διακυμάνσεις στις κορυφές, κάτι που μας επιβεβαιώνει και το παραπάνω συμπέρασμα.

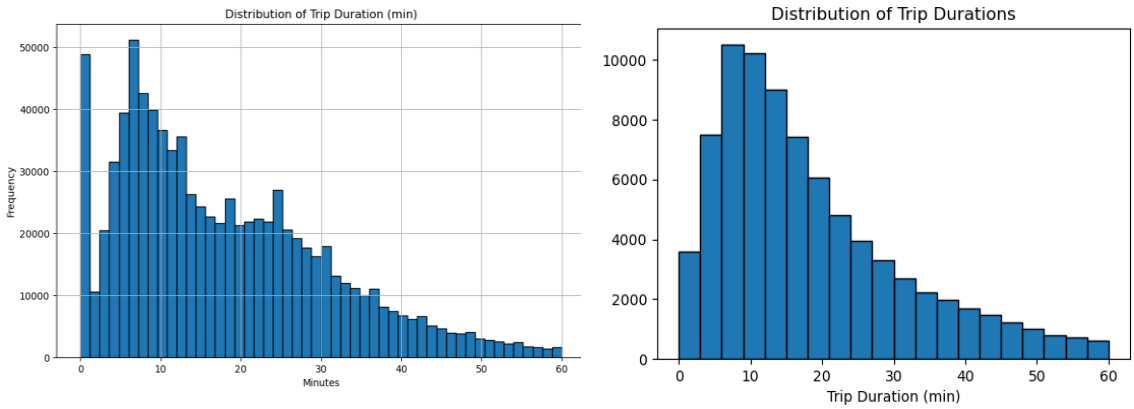
Τέλος για το σύνολο δεδομένων των λεωφορείων της Νέας Υόρκης έχουμε το χαρακτηριστικό “DistanceFromStop” το οποίο μας δηλώνει την απόσταση του λεωφορείου από την επόμενη στάση σε σχέση με τον προγραμματισμό. Η πολύ μεγάλη ασυμμετρία προς τα δεξιά και κύρτωση μας δείχνει ότι η πλειοψηφία των τιμών βρίσκονται στις χαμηλές τιμές το οποίο φαίνεται και από την μέση τιμή και την τυπική απόκλιση, που δείχνουν ότι οι περισσότερες τιμές βρίσκονται μεταξύ 100 και 200 γιαρδών, όπου 1 γιάρδα είναι περίπου 1 μέτρο. Μπορούμε να συμπεράνουμε λοιπόν ότι τα λεωφορεία στην Νέα Υόρκη στο μεγαλύτερο ποσοστό είναι στην ώρα τους ή έχουν μία μικρή καθυστέρηση 1-2 λεπτών, που αντιστοιχεί στην κάλυψη της μικρής αυτής απόστασης.

4.2 *Οπτικοποίηση Αποτελεσμάτων*

Σε αυτό το κεφάλαιο θα παρουσιάσουμε τα γραφήματα, διαγράμματα συχνοτήτων και θερμικούς χάρτες συσχέτισης που δημιουργήθηκαν στην ανάλυση των δεδομένων, για τα οποία συζητήσαμε στο υποκεφάλαιο 3.2.5 *Βήμα 5 και 3.2.6 Βήμα 6*. Παρακάτω έχουμε τα αποτελέσματα για κάθε σύνολο δεδομένων χωριστά.

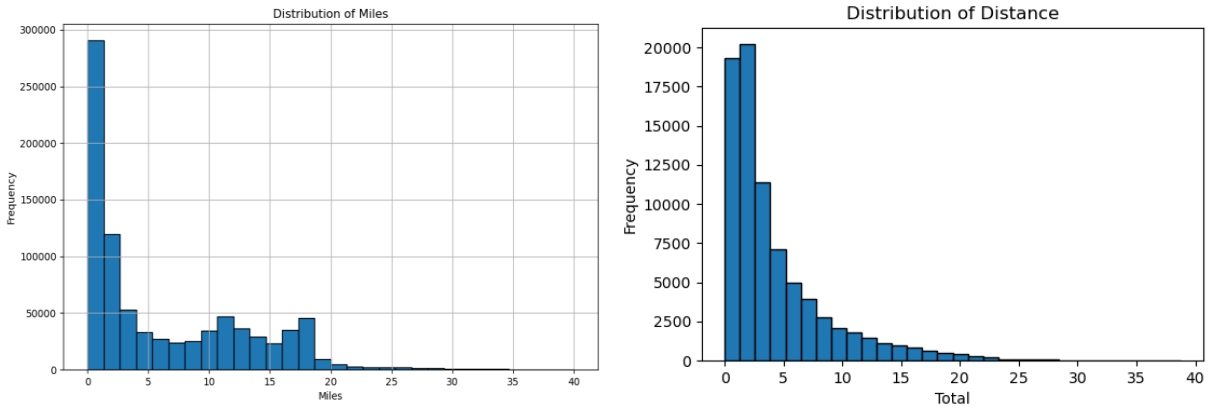
4.2.1 Μισθωμένα οχήματα – Taxi

- *Κοινά χαρακτηριστικά και αποτυπώσεις μεταξύ των συνόλων δεδομένων*



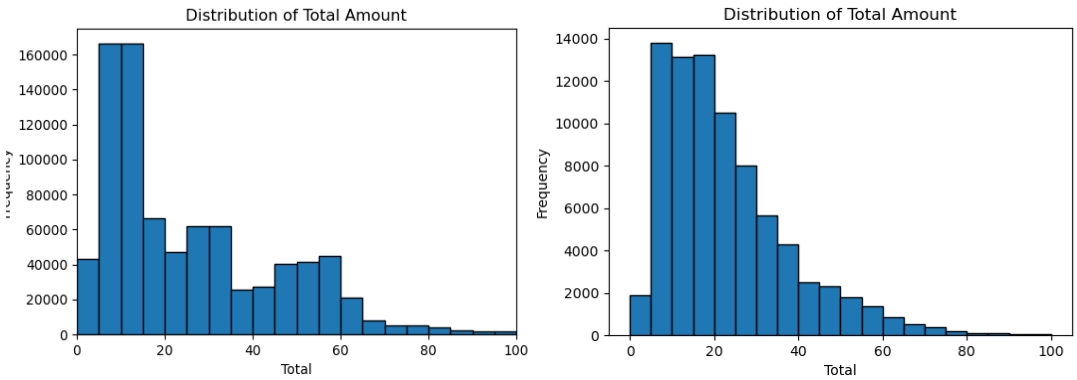
Εικόνα 3. Ιστόγραμμα του χαρακτηριστικού της διάρκειας του ταξιδιού με ταξί σε λεπτά. Αριστερά Σικάγο – Δεξιά Νέα Υόρκη

Στα ανωτέρω διαγράμματα βλέπουμε οπτικοποιημένα τα αποτελέσματα που αναλύσαμε στο προηγούμενο κεφάλαιο με χρήση των μέτρων Περιγραφικής Στατιστικής. Οι περισσότερες διαδρομές με ταξί βρίσκονται γενικά στο διάστημα μεταξύ 5-20 λεπτών. Στην περίπτωση του Σικάγο βλέπουμε ότι έχουμε έναν λοβό υψηλών συχνοτήτων κοντά στο 6ο λεπτό και έναν λοβό κοντά στο 20ο λεπτό, ενώ στην Νέα Υόρκη το διάγραμμα συχνοτήτων παρουσιάζει μία λιγότερο στενή και ομαλή κορυφή στις κατηγορίες που αντιστοιχούν από 6 έως 20 λεπτά ως οι πιο συχνές. Στην περίπτωση του Σικάγο όπως αναφέραμε και στο παραπάνω κεφάλαιο έχουμε πολύ υψηλή συχνότητα στις τιμές κοντά στο 0, που έχει ως αποτέλεσμα η κατηγορία του 1ου λεπτού να είναι η δεύτερη πιο συχνή κατηγορία στο ιστόγραμμα, κάτι που πρέπει να ληφθεί υπόψιν σε περαιτέρω ανάλυση του συγκεκριμένου συνόλου δεδομένων καθώς αυτό μπορεί να οφείλεται σε σφάλματα κατά την συλλογή δεδομένων.



Εικόνα 4. Ιστόγραμμα του χαρακτηριστικού της απόστασης του ταξιδιού με ταξί σε μίλια. Αριστερά Σικάγο – Δεξιά Νέα Υόρκη

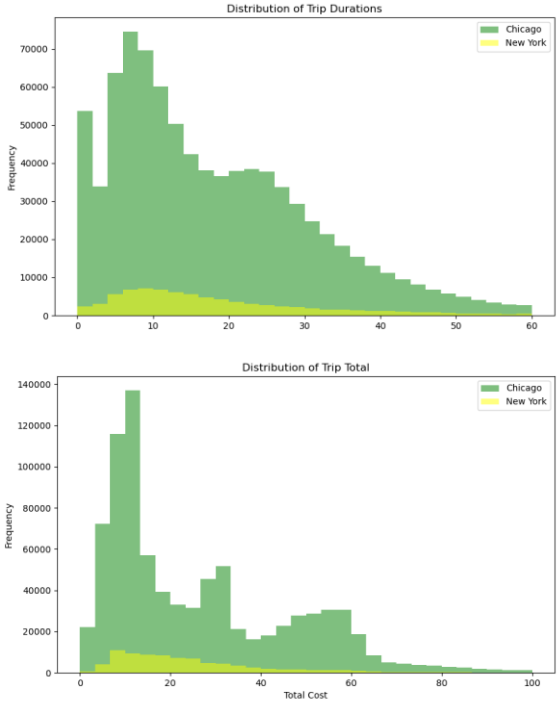
Εδώ παρατηρούμε και στις δύο πόλεις τα ταξίδια με ταξί είναι για μικρές αποστάσεις μέχρι 5 μίλια, με την συντριπτική πλειοψηφία να είναι για 0-3 μίλια. Στα αριστερά για το Σικάγο βλέπουμε ότι μετά τα 5 μίλια έχουμε μια σχετικά σταθερή συχνότητα μέχρι τα 18 μίλια περίπου, η οποία είναι υψηλότερη από τις συχνότητες στις αντίστοιχες κατηγορίες αποστάσεων στο σύνολο της Νέας Υόρκης. Αντίθετα για την Νέα Υόρκη όσο μεγαλύτερη είναι η απόσταση του ταξιδιού, τόσο η συχνότητα είναι πτωτική. Όπως είδαμε και στα αποτελέσματα των στατιστικών μέτρων, παρόλο που η συμπεριφορά των διαδρομών φαίνεται να είναι όμοια μεταξύ των δύο πόλεων, στο Σικάγο έχουμε ένα σχετικά πιο υψηλό μέσο όρο για την απόσταση των διαδρομών.



Εικόνα 5. Ιστόγραμμα του χαρακτηριστικού του συνολικού κόστους με ταξί σε δολάρια. Αριστερά Σικάγο – Δεξιά Νέα Υόρκη

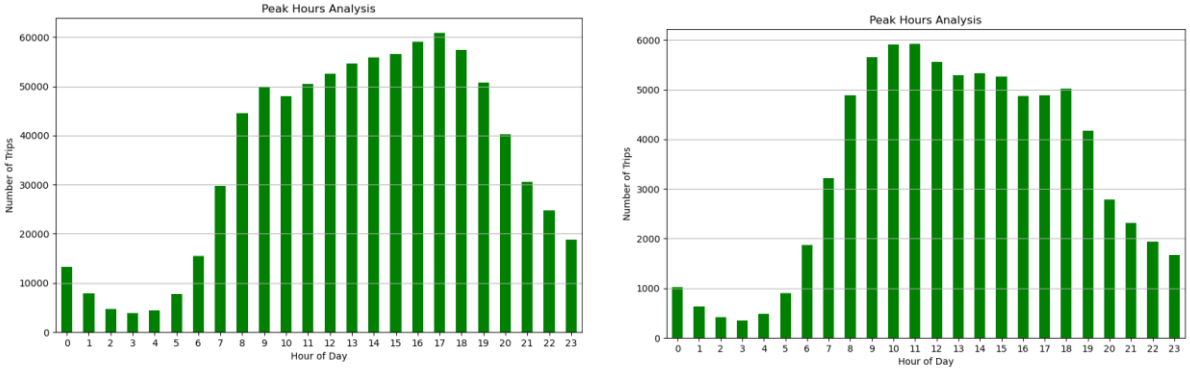
Εδώ παρατηρούμε ότι τα διαγράμματα αυτά είναι παρόμοια με αυτά της διάρκειας του ταξιδιού σε λεπτά το οποίο είναι αναμενόμενο καθώς οι χρεώσεις των ταξί γίνονται κυρίως με βάση τον χρόνο και λιγότερο με την απόσταση. Όπως είδαμε στα αποτελέσματα των στατιστικών μέτρων, το καθαρό και συνολικό κόστος ταξιδιού με ταξί στο Σικάγο έχει μεγαλύτερη διακύμανση σε σχέση με την Νέα Υόρκη, επομένως

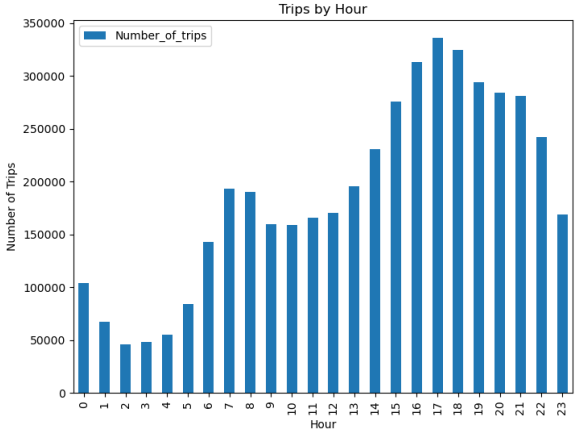
οι μεγάλες διαδρομές, που έχουν μεγαλύτερη συχνότητα στην περιοχή αυτή, ανεβάζουν την μέση τιμή του συνολικού κόστους για την διαδρομή. Στην Νέα Υόρκη το πιο κοινό κόστος από το διάγραμμα φαίνεται να είναι μεταξύ 5 και 10 δολαρίων, ενώ ακολουθούν με κοντινή συχνότητα οι κατηγορίες 10-15 και 15-20 δολάρια. Ενώ στο Σικάγο φαίνεται να είναι μεταξύ 5 και 15 δολαρίων.



Εικόνα 6. Κοινά ιστογράμματα των 3 αυτών χαρακτηριστικών μεταξύ Σικάγο και Νέα Υόρκης.

Για την καλύτερη οπτικοποίηση των ανωτέρω παραθέτουμε και τα κοινά ιστογράμματα αυτών των χαρακτηριστικών. Η διαφορά που βλέπουμε μεταξύ του ύψους των διαγραμμάτων συχνοτήτων οφείλεται στον πολύ μεγαλύτερο όγκο και άρα συχνότητας δεδομένων στο σύνολο δεδομένων του Σικάγο.

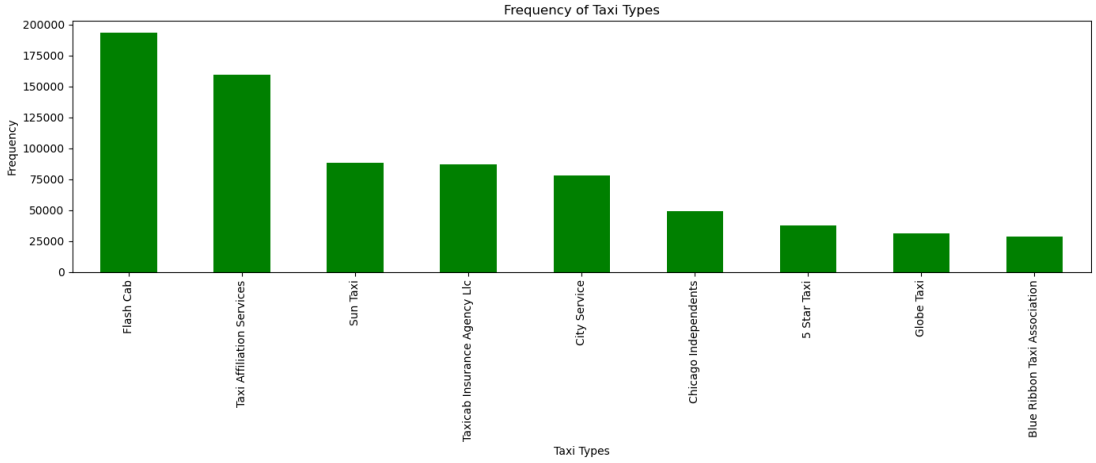




Εικόνα 7. Συχνότητες ταξιδιών ανά ώρα.
Επάνω αριστερά Ταξί Σικάγο, επάνω δεξιά Ταξί Νέα Υόρκη, Κάτω Uber Νέα Υόρκη

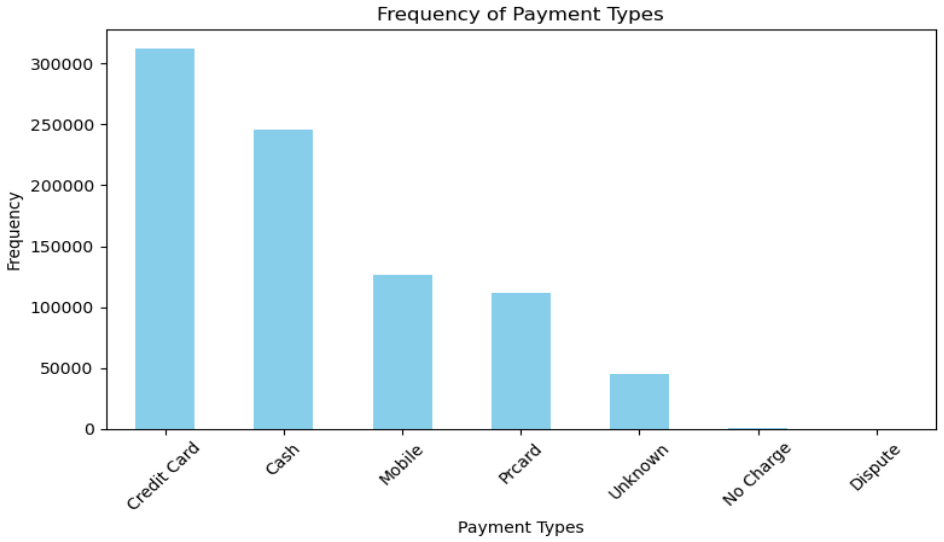
Σε αυτά τα διαγράμματα παρατηρούμε ότι έχουμε παρόμοια αποτελέσματα όσον αφορά την χρήση των ταξί ανεξαρτήτου πόλης (Σικάγο – Νέας Υόρκης), με πολύ μικρή χρήση τις πρώτες πρωινές ώρες και μετά σταθερή υψηλή δραστηριότητα κατά τις εργάσιμες ώρες μέχρι και δύο ώρες μετά (8 το πρωί έως 6 το απόγευμα). Αντίθετα, στην περίπτωση του Uber για την Νέα Υόρκη ενώ τις πρώτες πρωινές ώρες έχουμε της παρόμοια δραστηριότητα με τα ταξί, παρατηρούμε ότι η δραστηριότητα αυξάνεται από τις 4 το απόγευμα και μετά, ενώ υπάρχουν αρκετά μεγάλες συχνότητες ταξιδιού με Uber μέχρι και τις 9-10 το βράδυ. Ένα σημαντικό συμπέρασμα που μπορούμε να βγάλουμε από αυτό είναι ότι τα ταξί χρησιμοποιούνται κατά τις ώρες εργασίας, επομένως κατά πάσα πιθανότητα για μετάβαση από και προς την εργασία ή άλλα καθήκοντα σχετικά με την απασχόληση, ενώ το Uber χρησιμοποιείται και κατά τις ώρες εκτός τυπικού ωραρίου εργασίας, άρα πιθανόν και σε άλλες δραστηριότητες, όπως έξοδοι, νυχτερινή ζωή κλπ.

- *Taxi Database – Chicago*



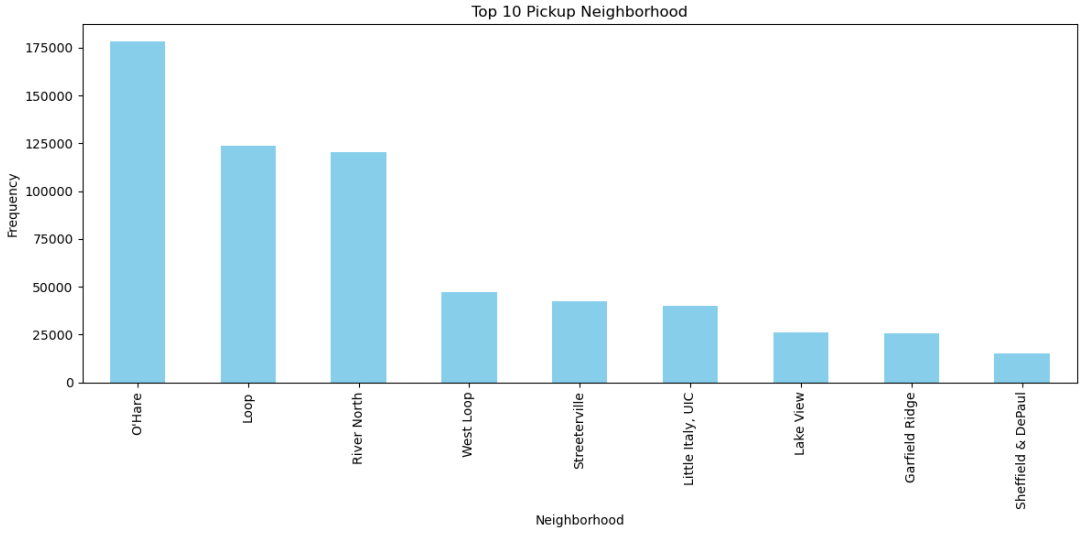
Εικόνα 8. Διάγραμμα συχνοτήτων ανά εταιρεία ταξί στο Σικάγο

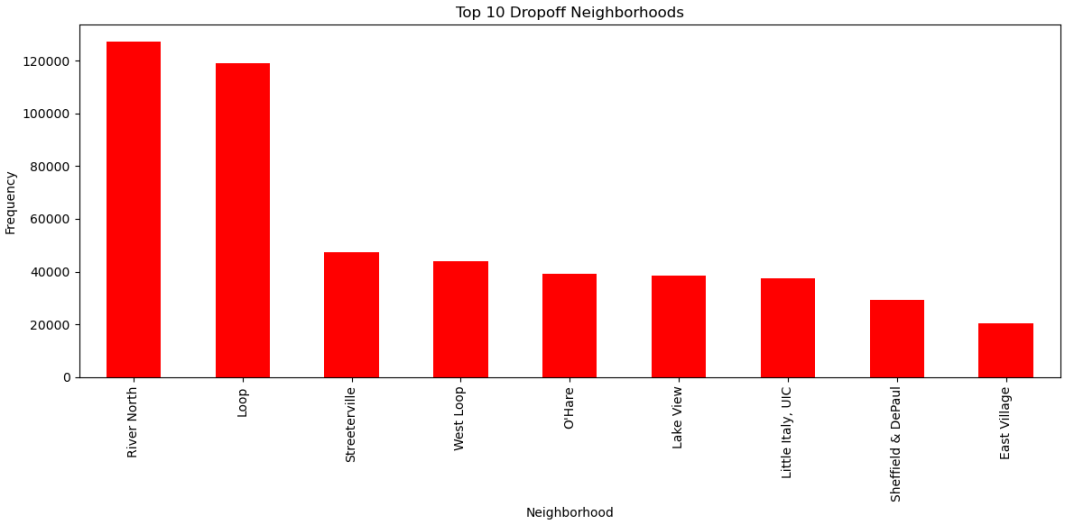
Εδώ έχουμε την συχνότητα ανά εταιρεία ταξί στο Σικάγο όπου παρατηρούμε ότι υπάρχουν αρκετές εταιρείες στο σύνολο δεδομένων το οποίο μας δείχνει ότι υπάρχει ανταγωνισμός στην συγκεκριμένη αγορά και ότι έχουμε καλή και αξιόπιστη ποιότητα δεδομένων αφού προέρχονται από πολλές διαφορετικές πηγές.



Εικόνα 9. Διάγραμμα συχνοτήτων για τον τρόπο πληρωμής για ταξίδι με ταξί στο Σικάγο

Από το παραπάνω διάγραμμα παρατηρούμε ότι οι περισσότερες διαδρομές με ταξί στο Σικάγο πληρώνονται με τραπεζική κάρτα και αμέσως μετά με μετρητά. Παρόμοια αποτελέσματα θα δούμε παρακάτω και στην περίπτωση των ταξί της Νέα Υόρκης. Αυτό μας δείχνει την συνεχή άνοδο που έχουν οι ηλεκτρονικές πληρωμές την τελευταία δεκαετία, αφού σε ένα μέσο όπου πληρωνόταν η διαδρομή σχεδόν αποκλειστικά με μετρητά, πλέον είναι η δεύτερη δημοφιλέστερη επιλογή.

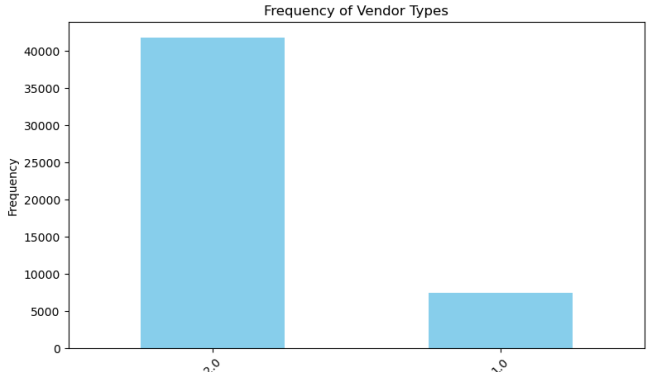




Εικόνα 10. Οι δέκα δημοφιλέστερες γειτονιές για το ξεκίνημα (μπλε) και τελείωμα (κόκκινο) του ταξιδιού με ταξί στο Σικάγο.

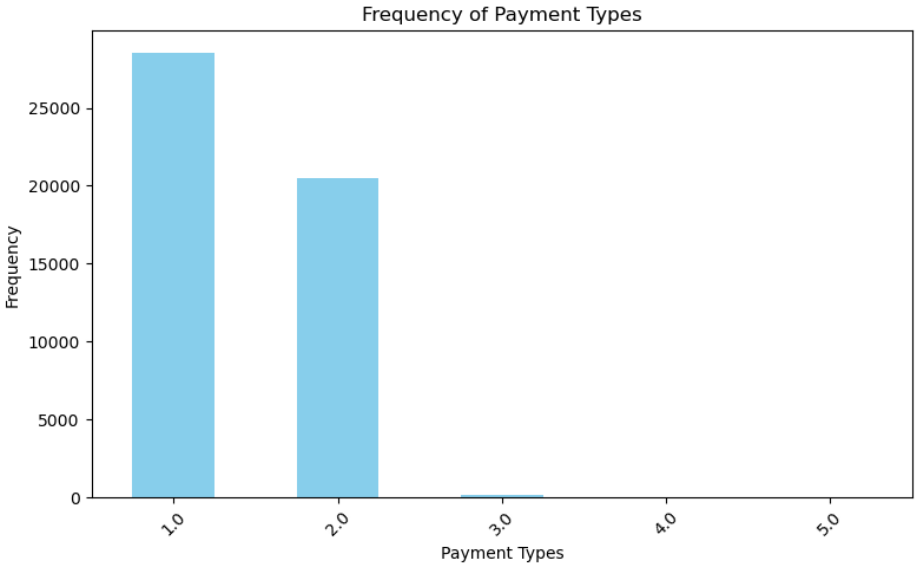
Εδώ έχουμε απεικονίσει τις δέκα πιο δημοφιλείς περιοχές για το ξεκίνημα και για το τελείωμα του ταξιδιού με ταξί στο Σικάγο. Παρατηρούμε ότι ενώ οι δέκα πιο δημοφιλείς περιοχές σχεδόν ταυτίζονται με μικρές διαφοροποιήσεις και στα δύο διαγράμματα, αυτές δεν βρίσκονται στην ίδια θέση στην πρώτη δεκάδα. Για παράδειγμα η περιοχή "O'Hare" είναι η πιο δημοφιλής περιοχή για ξεκίνημα, ενώ για το τελείωμα του ταξιδιού βρίσκεται στην 5η θέση. Στην περιοχή αυτή βρίσκεται το αεροδρόμιο όποτε συμπεραίνουμε ότι πολλοί άνθρωποι παίρνουν ταξί όταν θέλουν να φύγουν από το αεροδρόμιο ενώ όταν είναι να φθάσουν σε αυτό μετακινούνται και με άλλα μέσα. Παρατηρώντας την διαφοροποίηση των χαρακτηριστικών επί του συνόλου του database μπορούμε να επιβεβαιώσουμε και να βγάλουμε και αντίστοιχα συμπεράσματα και για άλλες περιοχές πχ. financial district.

- *Taxi Database – New York City*



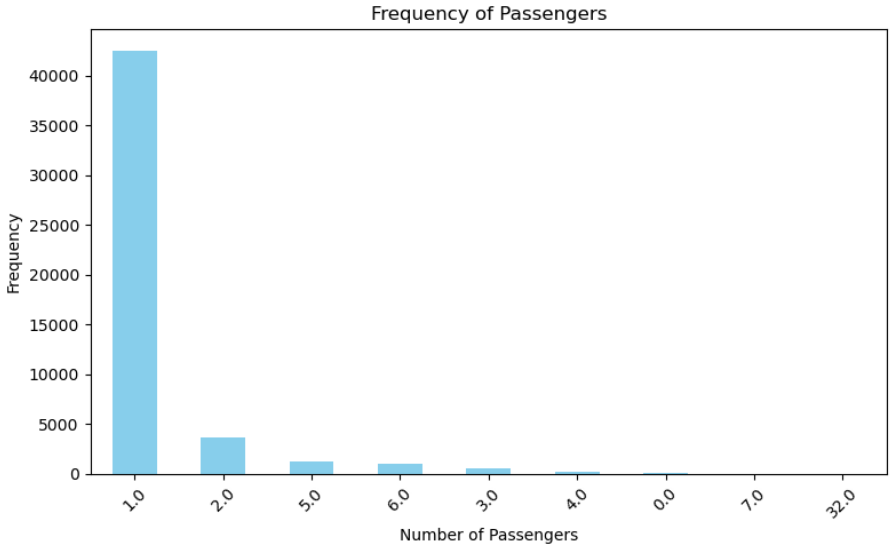
Εικόνα 11. Διάγραμμα συχνοτήτων δεδομένων ανά εταιρεία ταξί στην Νέα Υόρκη 1=Creative mobile technologies, 2=Verifone Inc.

Εδώ σε αντίθεση με το σύνολο δεδομένων ταξιδιού με ταξί του Σικάγο, έχουμε δεδομένα κυρίως από μία εταιρεία ταξί και αρκετά λιγότερα από μία δεύτερη. Επομένως δεν έχουμε μεγάλη ποικιλία των δεδομένων όσον αφορά τις πηγές τους.



Εικόνα 12. Διάγραμμα συχνοτήτων για τον τρόπο πληρωμής για ταξίδι με ταξί στην Νέα Υόρκη
1=Credit card, 2=Cash, 3=No charge, 4=Dispute, 5=Unknown

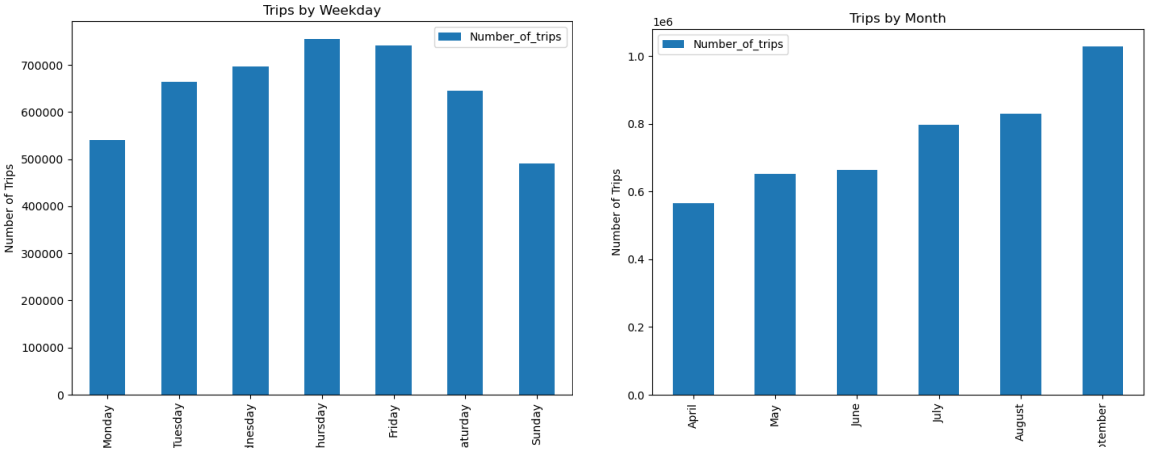
Ως προς τον τρόπο πληρωμής έχουμε όμοια αποτελέσματα με τα ταξί του Σικάγο, καθώς η πλειοψηφία των πληρωμών για το ταξίδι με ταξί στην έγινε με χρήση κάρτας, ενώ αμέσως μετά έρχεται η χρήση μετρητών.



Εικόνα 13. Διάγραμμα συχνοτήτων για το πλήθος επιβατών σε ταξί στην Νέα Υόρκη

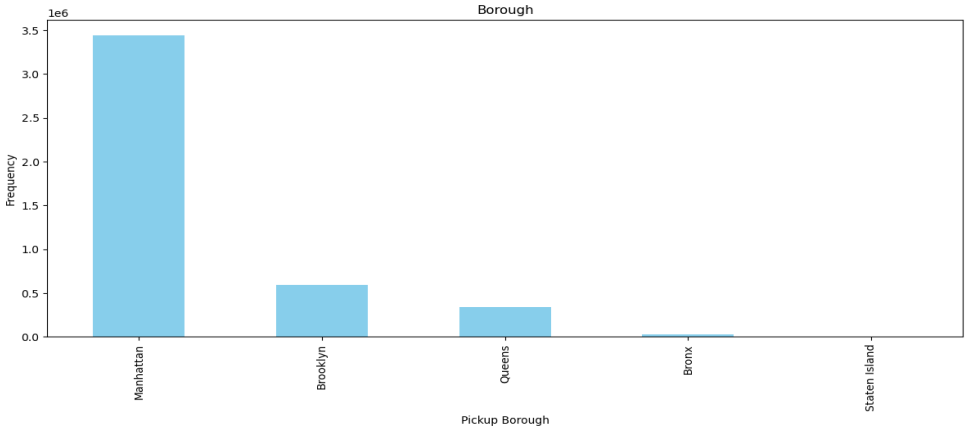
Εδώ, παρατηρούμε ότι η διαδρομή με ταξί στην Νέα Υόρκη είναι «ατομική» υπόθεση, ενώ σε ένα ποσοστό γύρω στο 10% το χρησιμοποιούν δύο άτομα.

- *Uber Database – New York City*



Εικόνα 14. Διάγραμμα συχνοτήτων ταξιδιών Uber ανά ημέρα της εβδομάδας αριστερά και ανά μήνα δεξιά.

Από την συχνότητα των ταξιδιών Uber στην Νέα Υόρκη ανά μήνα βλέπουμε μία σταθερή αύξηση των διαδρομών ανά μήνα κατά το χρονικό διάστημα στο οποίο αναφέρεται το σύνολο δεδομένων, από τον Απρίλιο έως τον Σεπτέμβριο. Αυτό είναι πιθανόν να συμβαίνει λόγω της σταδιακής εξάπλωσης του Uber στην αγορά, κατά το διάστημα αυτό, ενώ επίσης μπορεί να οφείλεται στην μετάβαση από την άνοιξη προς την τουριστική σεζόν του καλοκαιριού. Από την συχνότητα ταξιδιών Uber στην Νέα Υόρκη ανά ημέρα, παρατηρούμε ότι ξεκινώντας από την Δευτέρα έχουμε μία μικρή αύξηση μέχρι την μέση της εβδομάδας, όπου παρουσιάζει μέγιστο την Πέμπτη, και μετά μία συμμετρική μικρή μείωση έως την Κυριακή.

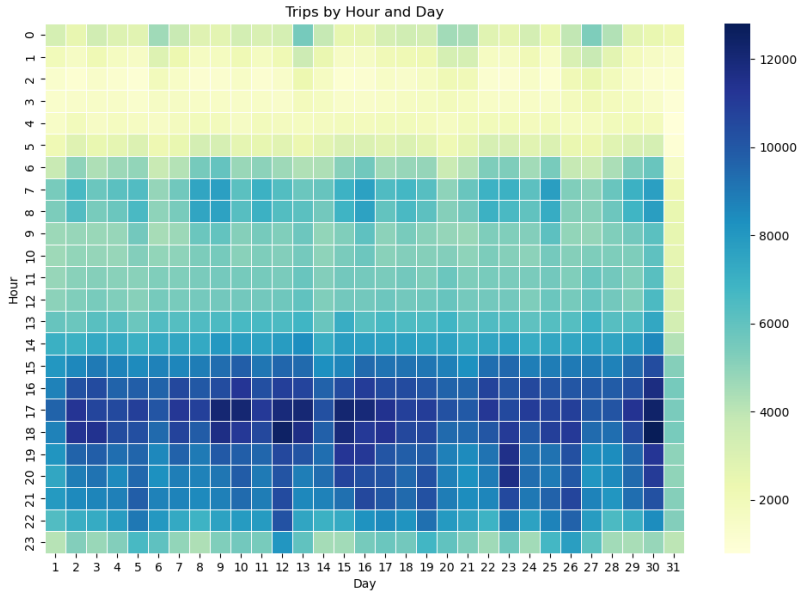


Εικόνα 15. Διάγραμμα Συχνοτήτων για το πλήθος ταξιδιών με Uber ανά προάστιο της Νέας Υόρκης

Από εδώ παρατηρούμε ότι οι διαδρομές του Uber στην Νέα Υόρκη αφορούν κυρίως το προάστιο του Manhattan. Αυτό φαίνεται από τον γεωγραφικό θερμικό χάρτη για τα ταξίδια με Uber ανά ώρα στην περιοχή της Νέας Υόρκης, όπως αυτός παρουσιάζεται παρακάτω. Λόγω της υψηλής πυκνότητα δεδομένων τμηματικά στον χάρτη, θα παρουσιάσουμε τα αποτελέσματα για τέσσερις διαφορετικές ώρες της ημέρας, εντός και εκτός των ωρών αιχμής, κοντά στο προάστιο του Manhattan.

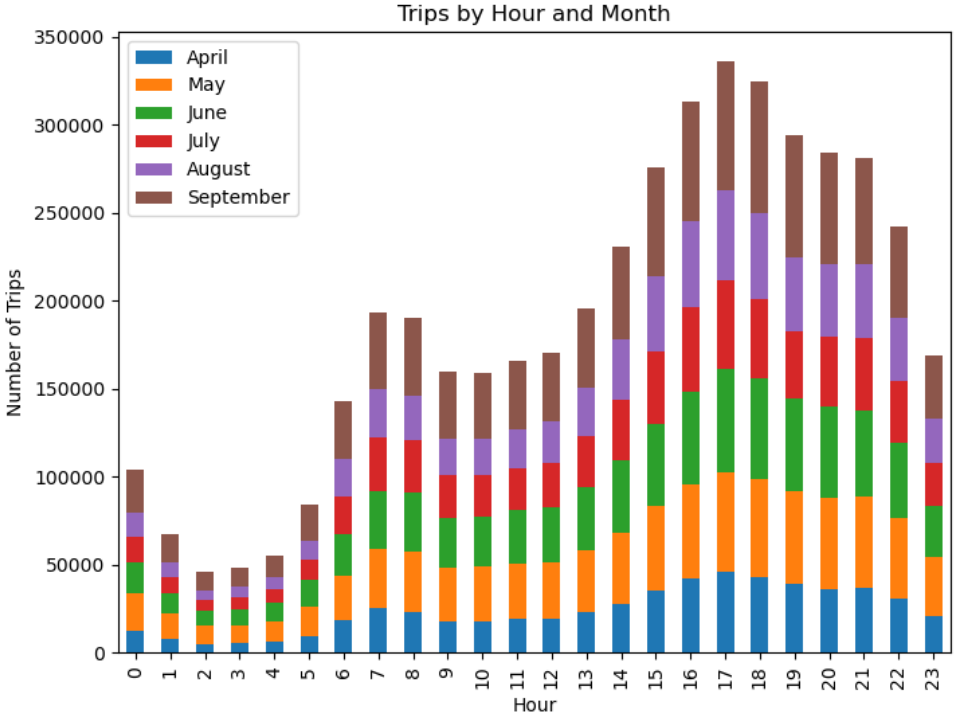
Εικόνα 16. Γεωγραφικός Θερμικός Χάρτης διαδρομών Uber στο Manhattan ανά ώρα. Επάνω αριστερά 4 π.μ., Επάνω δεξιά 8 π.μ., κάτω αριστερά 3 μ.μ., κάτω δεξιά 9 μ.μ.

Στα παραπάνω διαγράμματα φαίνονται πιο καθαρά τα συμπεράσματα που βγάλαμε από το διάγραμμα συχνοτήτων για ταξίδια με Uber ανά ώρα, αλλά ειδικότερα για το προάστιο του Manhattan. Παρατηρείται ότι η δραστηριότητα αυξάνεται μετά τις 8 το πρωί και γίνεται πιο έντονη μετά τις 3-4 το απόγευμα.



Εικόνα 17. Δισδιάστατος θερμικός χάρτης συχνότητας διαδρομών ανά ώρα και ημέρα του μήνα

Από τον Δισδιάστατο Θερμικό χάρτη συχνοτήτων, φαίνεται ότι η συχνότητα ανά ώρα είναι σταθερή σε σχέση με την ημέρα του μήνα, κάτι που μας δείχνει ότι αυτό το φαινόμενο είναι καθημερινό και δεν επηρεάζεται σημαντικά από την ημερομηνία του μήνα, αλλά πιθανόν κάποιους μήνες να είναι λιγότερο ή περισσότερο αισθητό. Τα συμπεράσματα που έχουμε βγάλει παραπάνω επαληθεύονται και εδώ για τις ώρες αιχμής και υψηλής δραστηριότητας των διαδρομών με Uber.

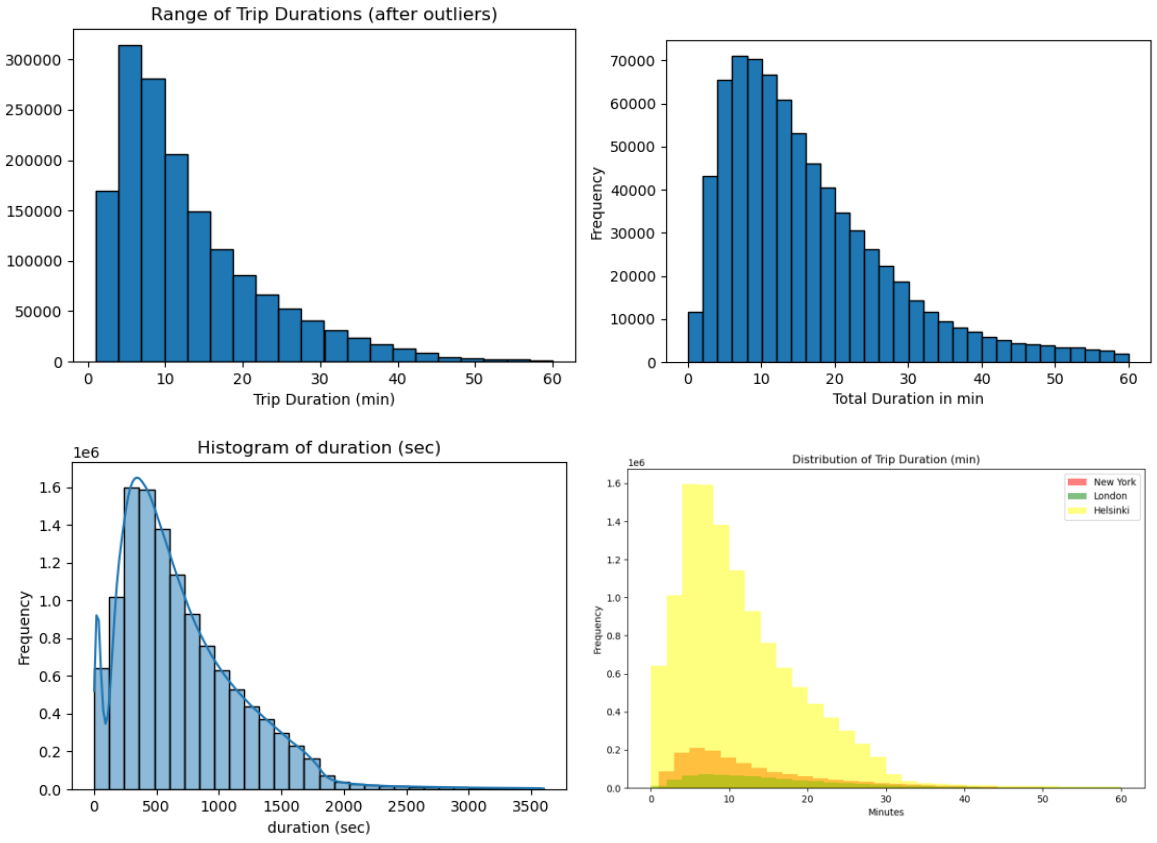


Εικόνα 18. Ραβδόγραμμα Συχνότητα ταξιδιών με Uber στην Νέα Υόρκη ανά ώρα και μήνα

Εδώ παρουσιάζουμε το ραβδόγραμμα συχνότητας ταξιδιών με Uber στην Νέα Υόρκη, ανά ώρα της ημέρας, με διαχωρισμό για κάθε μήνα κατά το χρονικό διάστημα στο οποίο αναφέρεται το σύνολο δεδομένων. Με αυτό το τρόπο μπορούμε να παρατηρήσουμε πιο εύκολα την αυξανόμενη μετακίνηση με Uber στην Νέα Υόρκη στο συγκεκριμένο χρονικό διάστημα, από τον Απρίλιο έως και τον Σεπτέμβριο του 2014, ανά μήνα. Παρατηρείται ότι η χρήση Uber ακολουθεί την ίδια κατανομή με αύξηση κατά τις ώρες αιχμής και τις απογευματινές ώρες, με μικρές διαφοροποιήσεις για κάθε μήνα κατά το χρονικό διάστημα μελέτης, παρά το ότι περιλαμβάνονται οι μήνες Ιουλίου, Αυγούστου όπου μειώνεται η δραστηριότητα στην πόλη λόγω διακοπών, αλλά κορυφώνεται η τουριστική περίοδος.

4.2.2 Ενοικιαζόμενα Ποδήλατα

- *Κοινά χαρακτηριστικά και αποτυπώσεις μεταξύ των συνόλων δεδομένων*

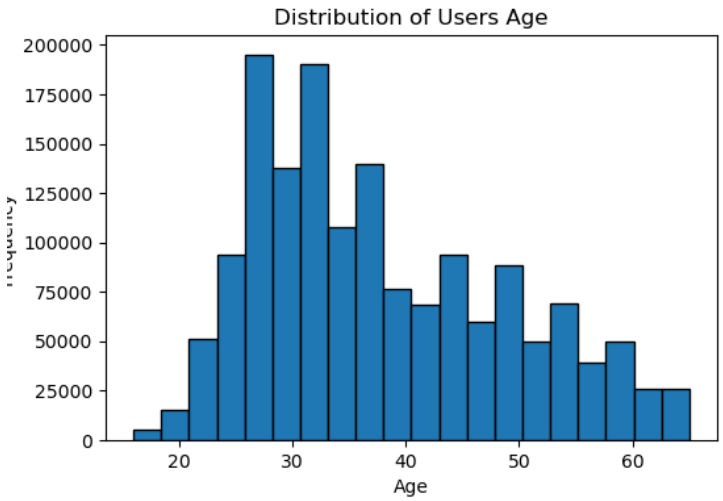


Εικόνα 19. Ιστόγραμμα του χαρακτηριστικού της διάρκειαq ταξιδιού με ποδήλατο σε λεπτά. Επάνω αριστερά Νέα Υόρκη, επάνω δεξιά Λονδίνο, κάτω αριστερά Ελσίνκι, κάτω δεξιά απεικόνιση σε κοινό διάγραμμα

Παρατηρούμε ότι όπως και στην προηγούμενη κατηγορία μετακίνησης, η παρατήρηση των ιστογραμμάτων οδηγεί σε συμπεράσματα παρόμοια με αυτά που

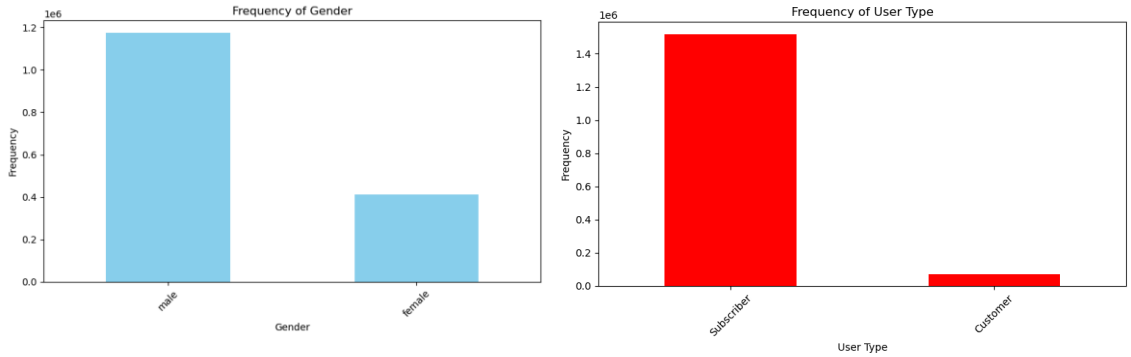
βγάλαμε για τα δεδομένα από τα αποτελέσματα των μέτρων Περιγραφικής Στατιστικής. Το ιστόγραμμα συχνοτήτων της διάρκειας ταξιδιού με ποδήλατο για τις τρεις πόλεις προσεγγίζεται από μία κατανομή που είναι ασύμμετρη δεξιά, με κορυφή στα 10 λεπτά, όπου η πλειοψηφία των διαδρομών βρίσκονται μεταξύ των 5 και 15 λεπτών. Ένα σημαντικό συμπέρασμα που βγάζουμε είναι ότι φαίνεται και στις τρεις πόλεις παρόλο που είναι πολύ διαφορετικές και σε διαφορετικά σημεία του πλανήτη η χρήση που γίνεται στα ενοικιαζόμενα ποδήλατα είναι παρόμοια και για ίδιους τύπου μετακινήσεις (μικρές αποστάσεις εντός πόλης). Στο τέταρτο διάγραμμα όπου φαίνεται το ιστόγραμμα συχνοτήτων της διάρκειας ταξιδιού με ποδήλατο για τις τρεις πόλεις σε κοινό άξονα, παρατηρούμε ότι η μεγάλη διαφορά στις συχνότητες και στο ύψος των διαγραμμάτων οφείλεται στο πολύ μεγαλύτερο όγκο δεδομένων που έχουμε στην περίπτωση του Ελσίνκι.

- *Bike Database – New York City*



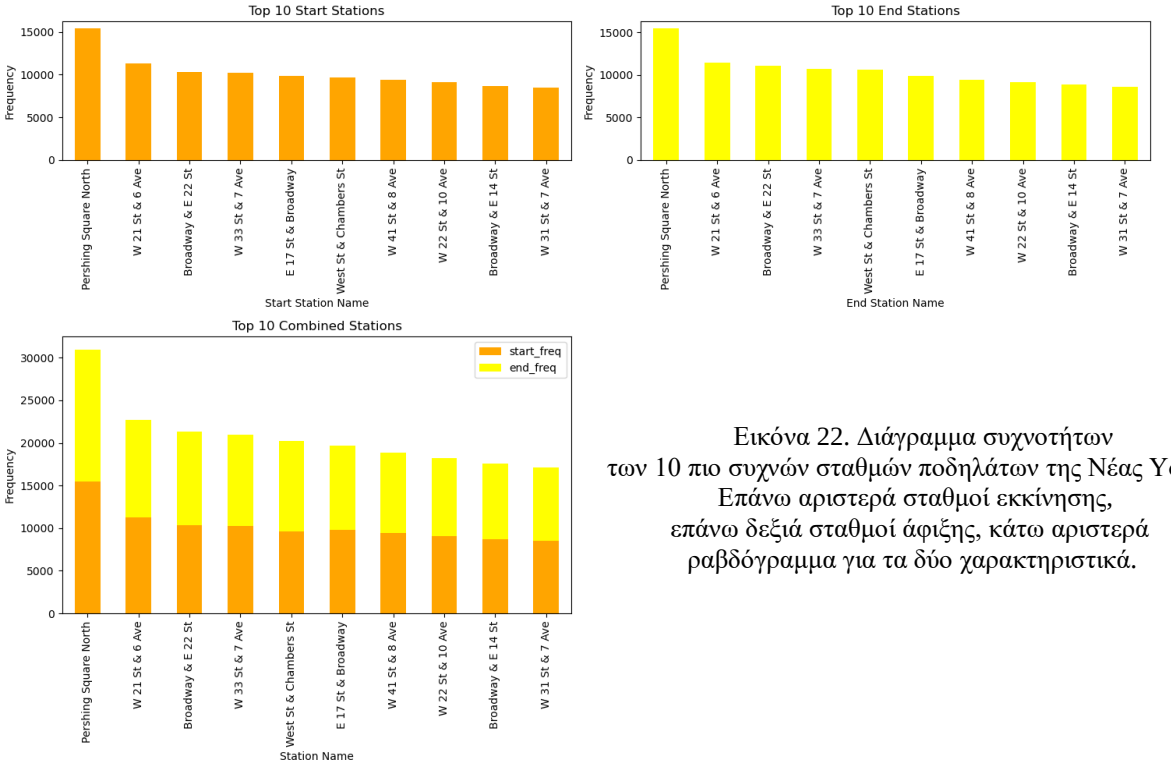
Εικόνα 20. Ιστόγραμμα του χαρακτηριστικού της ηλικίας του χρήστη με ποδήλατο στην Νέα Υόρκη

Το ιστόγραμμα συχνοτήτων του χαρακτηριστικού της ηλικίας χρήστη με ποδήλατο στην Νέα Υόρκη προσεγγίζεται από μία καμπύλη που έχει κεντρικό λοβό κεντραρισμένο στις ηλικίες 28-32 και είναι ασύμμετρη δεξιά, αλλά παρουσιάζει τοπικά μέγιστα, δηλαδή κορυφές μικρότερου ύψους στις τιμές 38, 44 και 50. Αυτό επιβεβαιώνει και το συμπέρασμα που βγάλαμε από τα μέτρα Περιγραφικής Στατιστικής για την διείσδυση του μέσου αυτού και σε μεγαλύτερες ηλικίες.



Εικόνα 21. Διάγραμμα συχνοτήτων του φύλου του χρήστη (μπλε) και του είδους του χρήστη (κόκκινο) με ποδήλατο στην Νέα Υόρκη

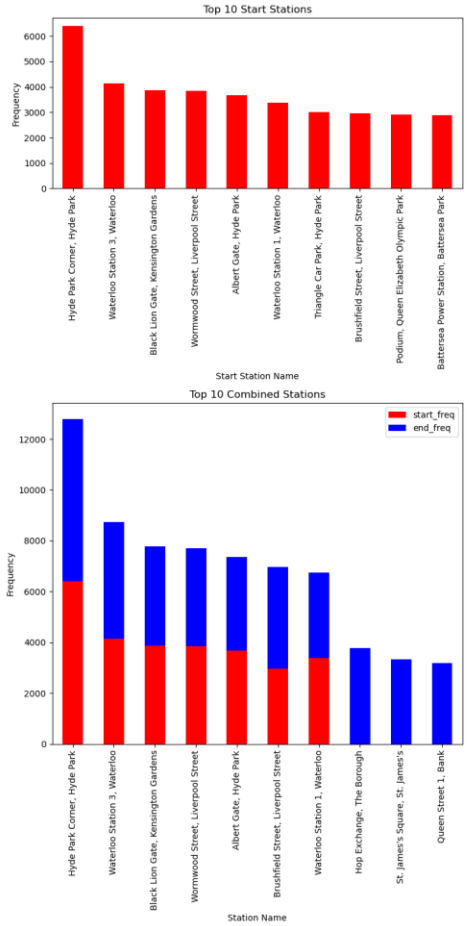
Από εδώ βλέπουμε ότι οι χρήστες των ενοικιαζόμενων ποδηλάτων στην Νέα Υόρκη είναι κατά πλειοψηφία άνδρες με αναλογία 3-1. Μία άλλη σημαντική πληροφορία που εξάγουμε είναι ότι οι χρήστες του μέσου αυτού είναι συνδρομητές της υπηρεσίας κατά ποσοστό άνω του 95%, και όχι μεμονωμένοι χρήστες. Επομένως, οι χρήστες του ποδηλάτου στην Νέα Υόρκη είναι χρήστες κατ' εξακολούθηση, δηλαδή τακτικοί πελάτες, γεγονός πολύ θετικό για ένα «πράσινο» εναλλακτικό μέσω μετακίνησης.



Εικόνα 22. Διάγραμμα συχνοτήτων των 10 πιο συχνών σταθμών ποδηλάτων της Νέας Υόρκης Επάνω αριστερά σταθμοί εκκίνησης, επάνω δεξιά σταθμοί άφιξης, κάτω αριστερά ραβδόγραμμα για τα δύο χαρακτηριστικά.

Στα παραπάνω διαγράμματα παρατηρούμε ότι ταυτίζονται οι πιο δημοφιλείς στάσεις για την εκκίνηση και την άφιξη του ταξιδιού με το ποδήλατο στην Νέα Υόρκη. Αυτό μας δείχνει ότι οι στάσεις αυτές είναι μεγάλα “hubs” που ίσως βρίσκονται και κοντά σε δημοφιλείς στάσεις μετρό ώστε να συνεχίζεται το ταξίδι.

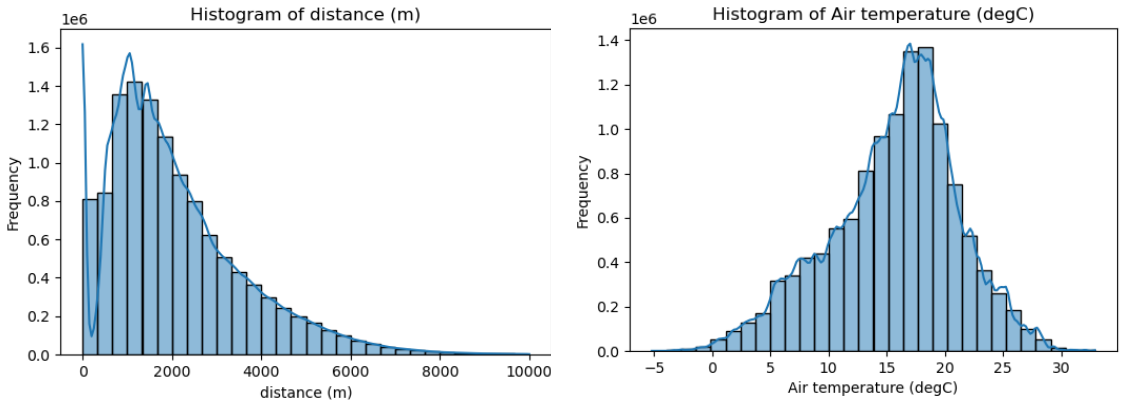
- *Bike Database – London*

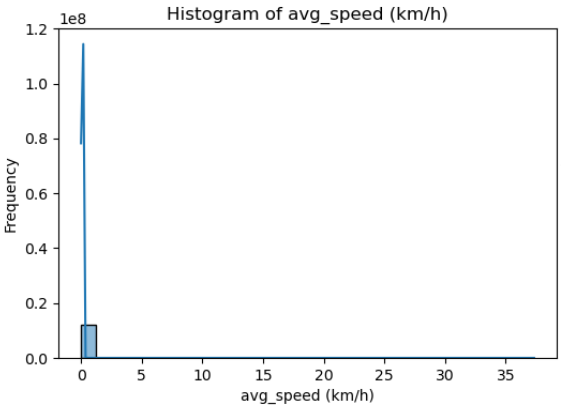


Εικόνα 23. Διάγραμμα συχνοτήτων
των 10 πιο συχνών σταθμών ποδηλάτων του Λονδίνου
Επάνω αριστερά σταθμοί εκκίνησης,
επάνω δεξιά σταθμοί άφιξης, κάτω αριστερά
ραβδόγραμμα για τα δύο χαρακτηριστικά.

Για το σύνολο ταξιδιών με ποδήλατο στο Λονδίνο, παρατηρείται ότι οι 10 πιο δημοφιλείς σταθμοί εκκίνησης δεν συμπίπτουν με τους 10 πιο δημοφιλείς σταθμούς άφιξης, καθώς σε κάθε δεκάδα υπάρχουν τρεις σταθμοί που δεν υπάρχουν στην άλλη, αλλά οι υπόλοιποι επτά σταθμοί είναι κοινοί, αλλά σε διαφορετικές θέσεις δημοφιλίας. Αυτό φαίνεται και στο παραπάνω ραβδόγραμμα όπου τρεις σταθμοί άφιξης δεν ανήκουν στην αντίστοιχη λίστα με τους πιο δημοφιλείς σταθμούς εκκίνησης, επομένως η συχνότητα εμφάνισης είναι 0.

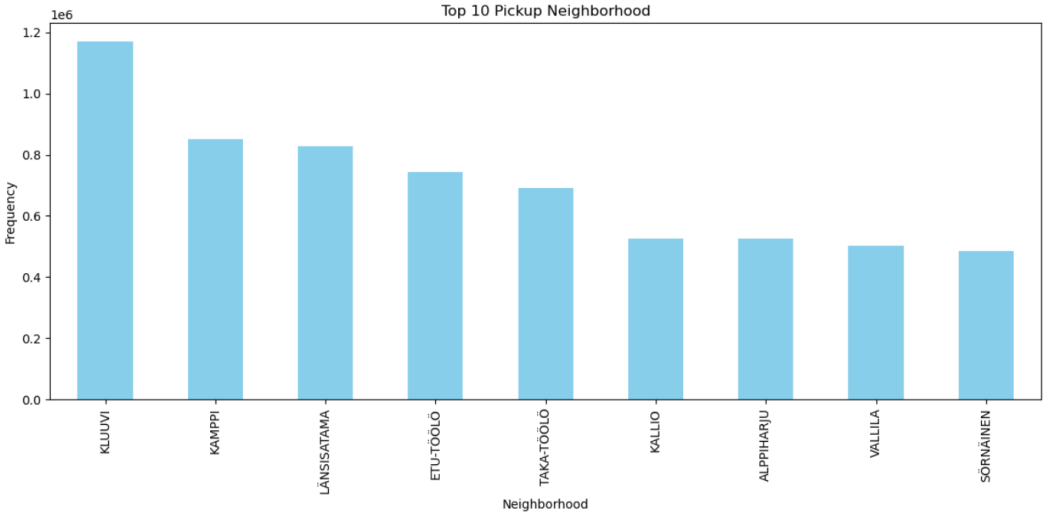
- *Bike Database – Helsinki*





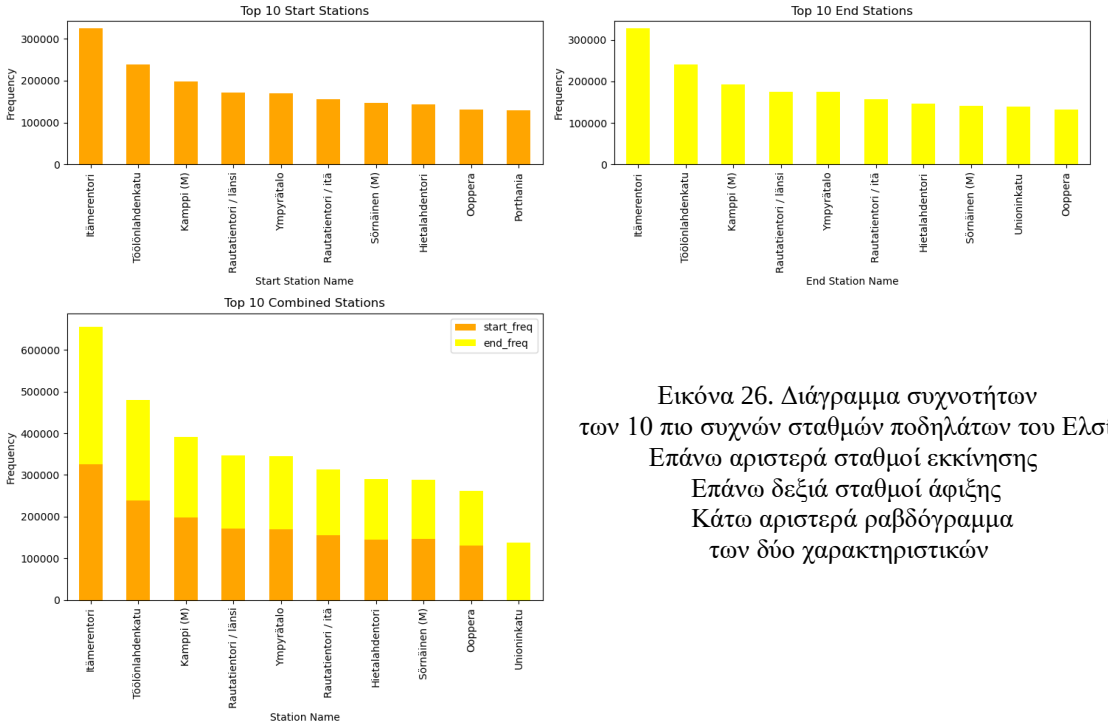
Εικόνα 24. Ιστόγραμμα των υπόλοιπων χαρακτηριστικών για ταξίδι με ποδήλατα στο Ελσίνκι. Επάνω αριστερά απόσταση σε μέτρα. Επάνω δεξιά θερμοκρασία αέρα σε °C. Κάτω αριστερά μέση ταχύτητα σε χλμ./ώρα.

Παρατηρούμε ότι το ιστόγραμμα της απόστασης προσεγγίζεται από κατανομή που είναι παρόμοια με αυτή της κατανομής της διάρκειας που είδαμε παραπάνω, όπως είναι και αναμενόμενο, καθώς η διάρκεια του ταξιδιού είναι ανάλογη με την απόσταση. Για το ιστόγραμμα της θερμοκρασίας αέρα, βλέπουμε μία σχεδόν συμμετρική κατανομή που το προσεγγίζει, με μία ελαφριά ασυμμετρία προς τα αριστερά, και με κορυφή κοντά στους 18 βαθμούς °C, το οποίο μας δείχνει ότι στο Ελσίνκι η χρήση των ποδηλάτων λαμβάνει χώρα όλο το χρόνο και σε εποχές με ακραίες θερμοκρασίες για το τοπικό κλίμα, αλλά με μία μεγαλύτερη χρήση σε πιο θερμό κλίμα όπως είναι και αναμενόμενο. Στο ιστόγραμμα της μέσης ταχύτητας του ταξιδιού με ποδήλατο, παρατηρείται μόνο μια πολύ ψηλή κορυφή κοντά στο 0, καθώς όλες οι τιμές ταχύτητας στο σύνολο δεδομένων ήταν πολύ χαμηλές σε μέσο όρο. Τα διαγράμματα αυτά επαληθεύουν και τα συμπεράσματα που βγάλαμε από τα αποτελέσματα των μέτρων Περιγραφικής Στατιστικής στο προηγούμενο υποκεφάλαιο.



Εικόνα 25. Διάγραμμα συχνοτήτων των 10 πιο δημοφιλών γειτονιών εκκίνησης ταξιδιού

Στην περίπτωση του συνόλου δεδομένων ταξιδιού με ποδήλατο στο Ελσίνκι, όπως έχουμε αναφέρει, έχουμε χαρακτηριστικά τοποθεσίας οπότε μπορέσαμε να εξάγουμε πληροφορία για την γειτονιά από όπου ξεκινάει το ταξίδι. Στο παραπάνω διάγραμμα παρουσιάζουμε τις δέκα πιο δημοφιλείς γειτονιές εκκίνησης ταξιδιού και την συχνότητα με την οποία εμφανίζονται στο σύνολο δεδομένων.

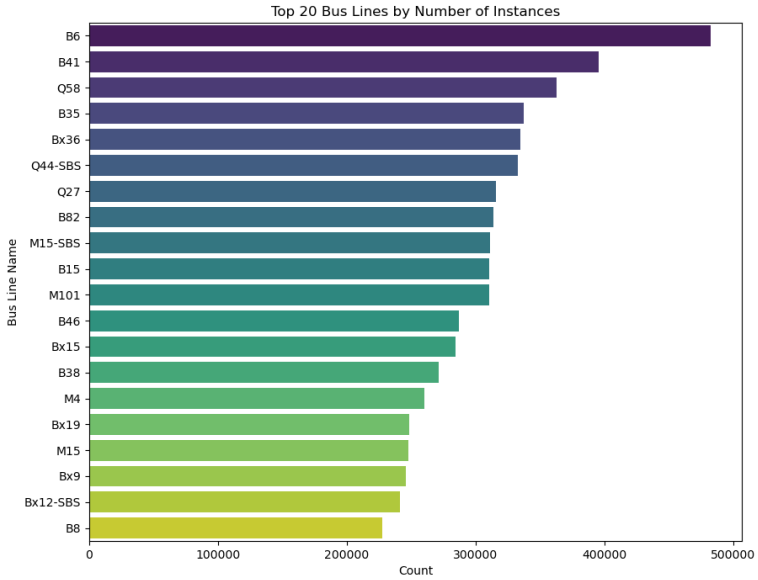


Εικόνα 26. Διάγραμμα συχνοτήτων
των 10 πιο συχνών σταθμών ποδηλάτων του Ελσίνκι
Επάνω αριστερά σταθμοί εκκίνησης
Επάνω δεξιά σταθμοί άφιξης
Κάτω αριστερά ραβδόγραμμα
των δύο χαρακτηριστικών

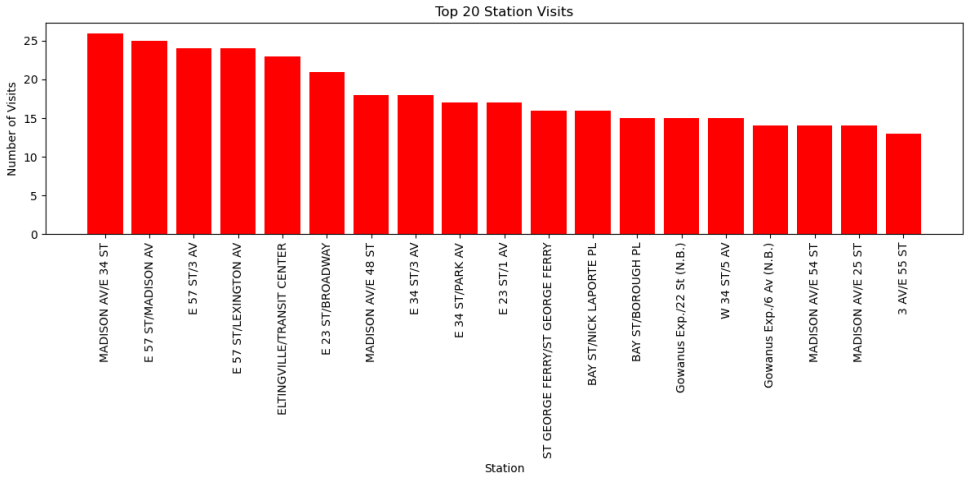
Για το σύνολο ταξιδιών με ποδήλατο στο Ελσίνκι, παρατηρείται ότι οι 6 πιο δημοφιλείς σταθμοί εκκίνησης είναι κοινοί με τους 6 πιο δημοφιλείς σταθμούς άφιξης, αλλά σε διαφορετικές θέσεις δημοφιλίας, ενώ σε κάθε δεκάδα υπάρχει ένας σταθμός που δεν υπάρχει στην άλλη. Αυτό φαίνεται και στο παραπάνω ραβδόγραμμα όπου ένας σταθμός άφιξης δεν ανήκει στην αντίστοιχη λίστα με τους πιο δημοφιλείς σταθμούς εκκίνησης, επομένως η συχνότητα εμφάνισης είναι 0. Αυτό μας δείχνει ότι οι στάσεις αυτές είναι μεγάλα “hubs” που ίσως βρίσκονται και κοντά σε δημοφιλείς στάσεις μετρό ώστε να συνεχίζεται το ταξίδι.

4.2.3 Μέσα Μαζικής Μεταφοράς

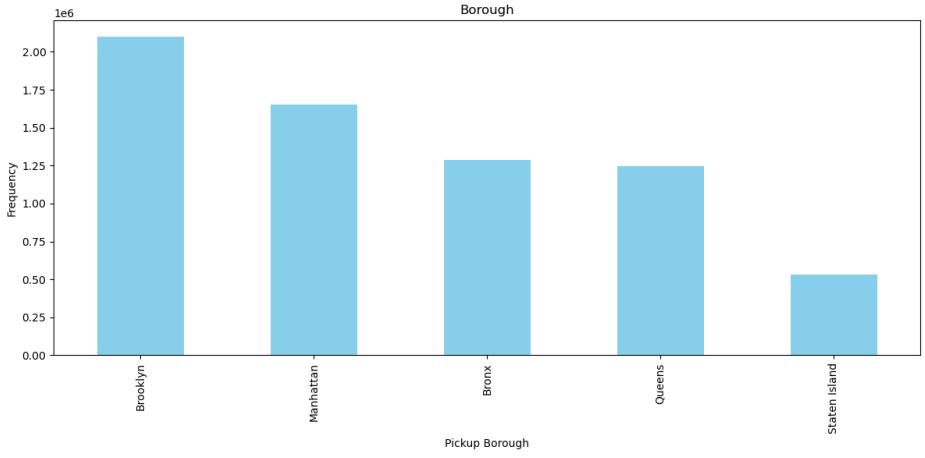
- *Bus Database – New York City*



Εικόνα 27. Είκοσι πιο δημοφιλείς γραμμές λεωφορείου στην Νέα Υόρκη

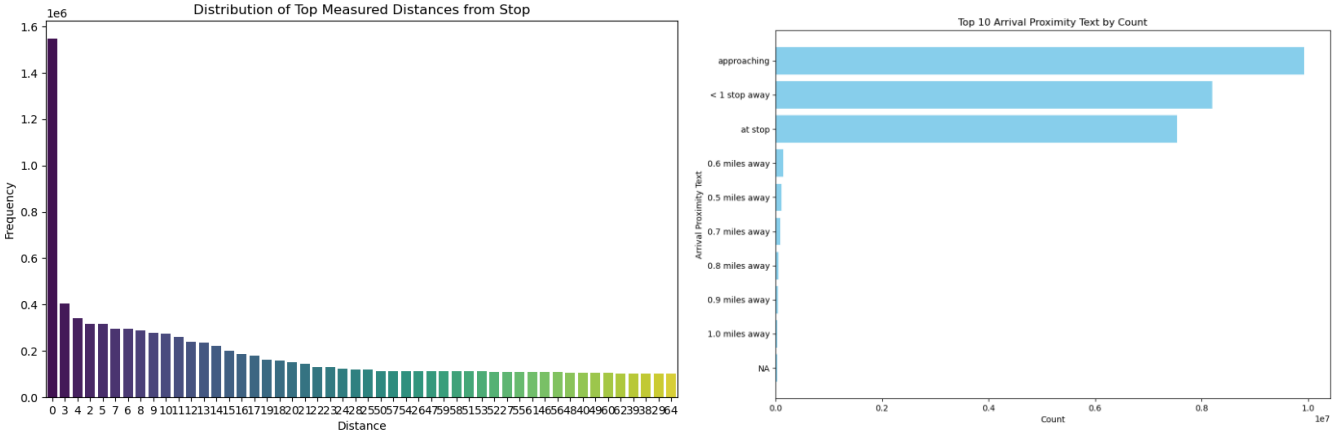


Εικόνα 28. Είκοσι πιο δημοφιλείς στάσεις λεωφορείου στην Νέα Υόρκη



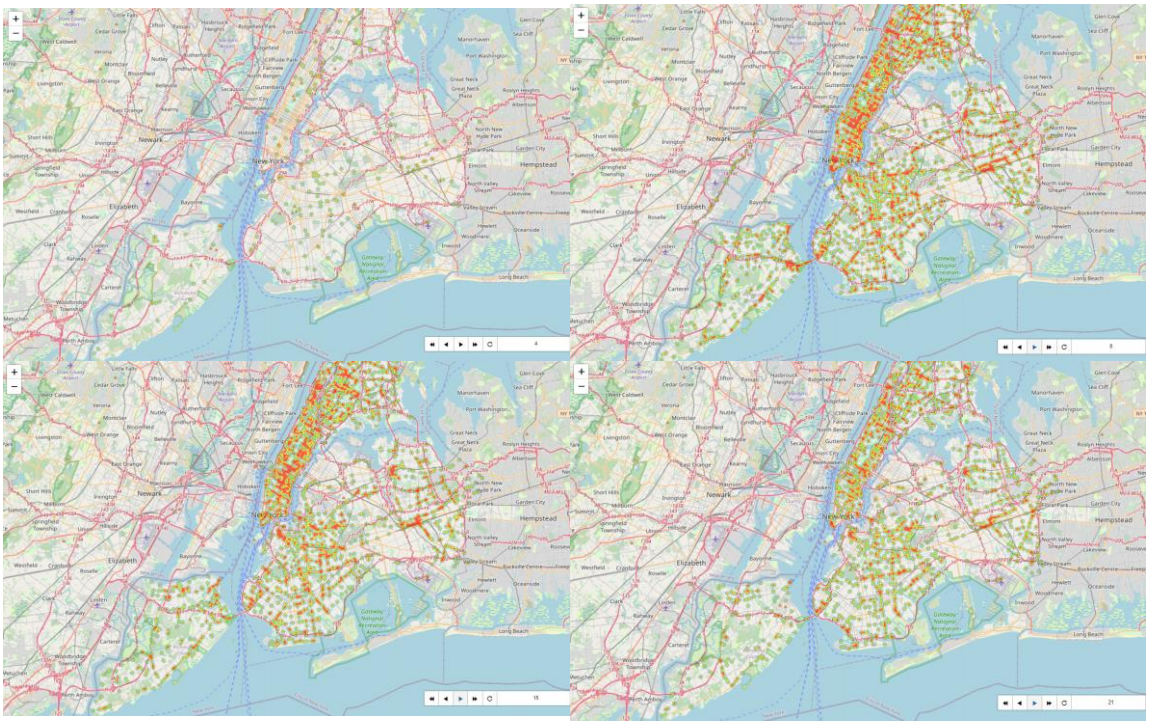
Εικόνα 29. Διάγραμμα συχνοτήτων ταξιδιού με λεωφορείο ανά προάστιο της Νέας Υόρκης

Στα τρία αυτά διαγράμματα παρουσιάζουμε τις πιο συχνές κατηγορίες για τα κατηγορικά χαρακτηριστικά με πληροφορία τοποθεσίας, δηλαδή το γεωγραφικό διαμέρισμα, τις στάσεις και τις γραμμές λεωφορείου ώστε να σχηματιστεί μια πρώτη εικόνα των πιο πυκνοκατοικημένων και δημοφιλών περιοχών.



Εικόνα 30. Διάγραμμα συχνοτήτων των χαρακτηριστικών που μετράνε την απόσταση από την στάση. Αριστερά σχετική απόσταση από στάση, Δεξιά κατηγορική σχετική απόσταση από στάση.

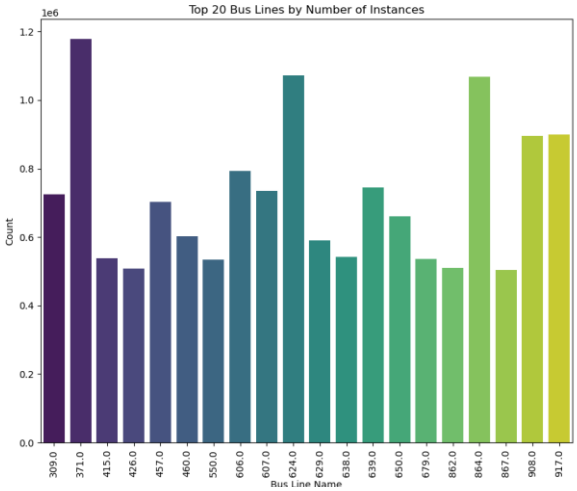
Το αριστερό διάγραμμα συχνοτήτων αφορά το χαρακτηριστικό της σχετικής απόστασης του λεωφορείου από την στάση, σε σχέση με αυτήν όπου θα έπρεπε να βρίσκεται, μετρημένη σε γιάρδες. Ειδικότερα παρουσιάζονται οι συχνότητες των πιο συχνά εμφανιζόμενων τιμών σχετικής απόστασης. Στο δεξί διάγραμμα παρουσιάζονται οι συχνότητες των πιο συχνά εμφανιζόμενων κατηγοριών ενός παρόμοιου χαρακτηριστικού που μας δείχνει την σχετική απόσταση από την στάση περιγραφικά, έχοντας χωρίσει σε διάφορες κατηγορίες την απόσταση. Και από τα δύο αυτά διαγράμματα φαίνεται ότι τα λεωφορεία στην Νέα Υόρκη δεν έχουν πολύ μεγάλες καθυστερήσεις καθώς κατά συντριπτική πλειοψηφία βρίσκονται είτε στην στάση στην ώρα τους, είτε πλησιάζουν όντας λίγα μέτρα μακριά, είτε βρίσκονται μία στάση πριν.



Εικόνα 31. Γεωγραφικός Θερμικός Χάρτης θέσεων λεωφορείων στην Νέα Υόρκη ανά ώρα. Επάνω αριστερά 4 π.μ., Επάνω δεξιά 8 π.μ., κάτω αριστερά 3 μ.μ., κάτω δεξιά 9 μ.μ.

Από την παρατήρηση του γεωγραφικού θερμικού χάρτη της θέσης των λεωφορείων ανά ώρα βλέπουμε ότι τα λεωφορεία στην Νέα Υόρκη λειτουργούν όλο το 24ώρο. Τις πρώτες πρωινές ώρες μέχρι τις 7-8 το πρωί δεν υπάρχει μεγάλη δραστηριότητα, η οποία αυξάνεται έντονα μετά τις 8 και μένει σταθερή μέχρι τις 6 το απόγευμα. Μετά έχουμε σταδιακή πτώση της δραστηριότητας μέχρι τις 9-10 το βράδυ όπου πάλι έχουμε μικρή δραστηριότητα. Μεγαλύτερη κινητικότητα λεωφορείων κατά τις ώρες αιχμής έχει το Μανχάταν το οποίο είναι λογικό μιας και είναι μια πυκνοκατοικημένη περιοχή με μικρή σχετικά επιφάνεια.

- *Bus Database – Rio De Janeiro*

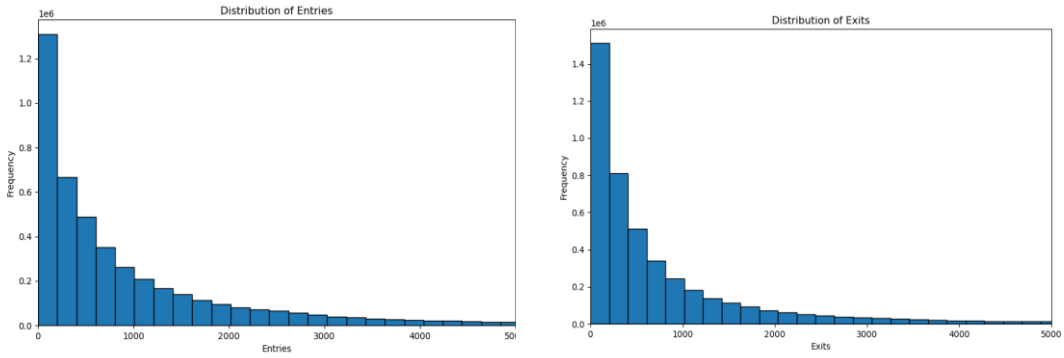


Εικόνα 32. Είκοσι πιο δημοφιλείς γραμμές λεωφορείου στο Ρίο Ντε Τζανέιρο

Εικόνα 33. Γεωγραφικός Θερμικός Χάρτης θέσεων λεωφορείων στο Ρίο Ντε Τζανέιρο ανά ώρα. Επάνω αριστερά 4 π.μ., Επάνω δεξιά 8 π.μ., κάτω αριστερά 3 μ.μ., κάτω δεξιά 9 μ.μ.

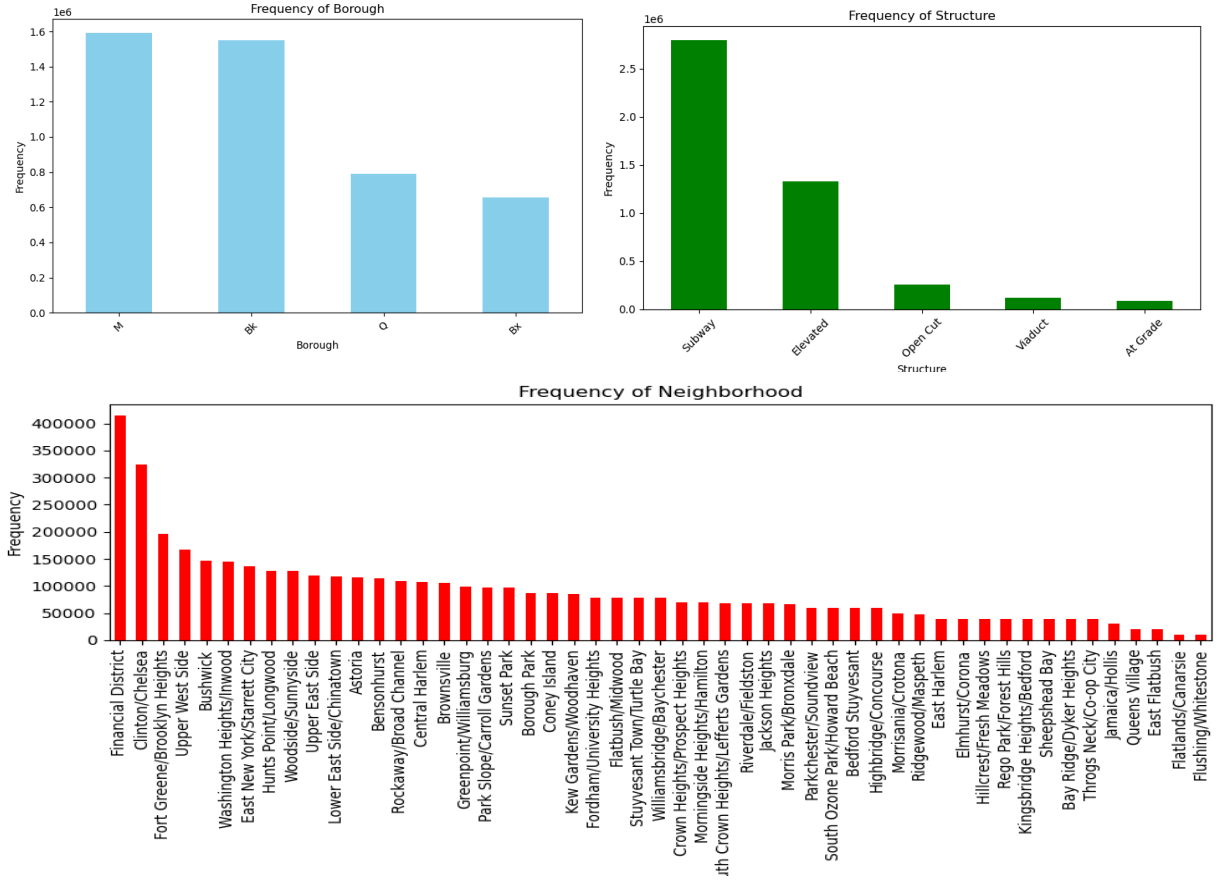
Εδώ παρουσιάζουμε το διάγραμμα συχνοτήτων των είκοσι πιο συχνών ή δημοφιλών γραμμών λεωφορείων στο Ρίο ντε Τζανέιρο, αντίστοιχα όπως κάναμε για το σύνολο δεδομένων λεωφορείων της Νέας Υόρκης. Στον γεωγραφικό θερμικό χάρτη ανά ώρα, όπου φαίνεται η πυκνότητα θέσης λεωφορείου ανά μονάδα έκτασης, παρατηρούμε ότι δεν έχουμε 24ώρα δρομολόγια. Τα δρομολόγια ξεκινούν από τις 5-6 το πρωί με την δραστηριότητα να αγγίζει το μέγιστο από τις 8 το πρωί έως τις 8 το βράδυ, ενώ μετά έχουμε σταδιακή αραίωση της δραστηριότητας μέχρι τις 12 το βράδυ. Επίσης παρατηρούμε ότι στην περίπτωση του Ρίο τα λεωφορεία κινούνται σε συγκεκριμένα σημεία της πόλης και στις βασικές κεντρικές αρτηρίες. Αυτό μας αποτυπώνει και τον γεωγραφικό και οικονομικό διαχωρισμό που υπάρχει στην πόλη ανάμεσα στις αναπτυγμένες, τουριστικά και οικονομικά, περιοχές και τις φαβέλες.

- *Metro Database – New York City*



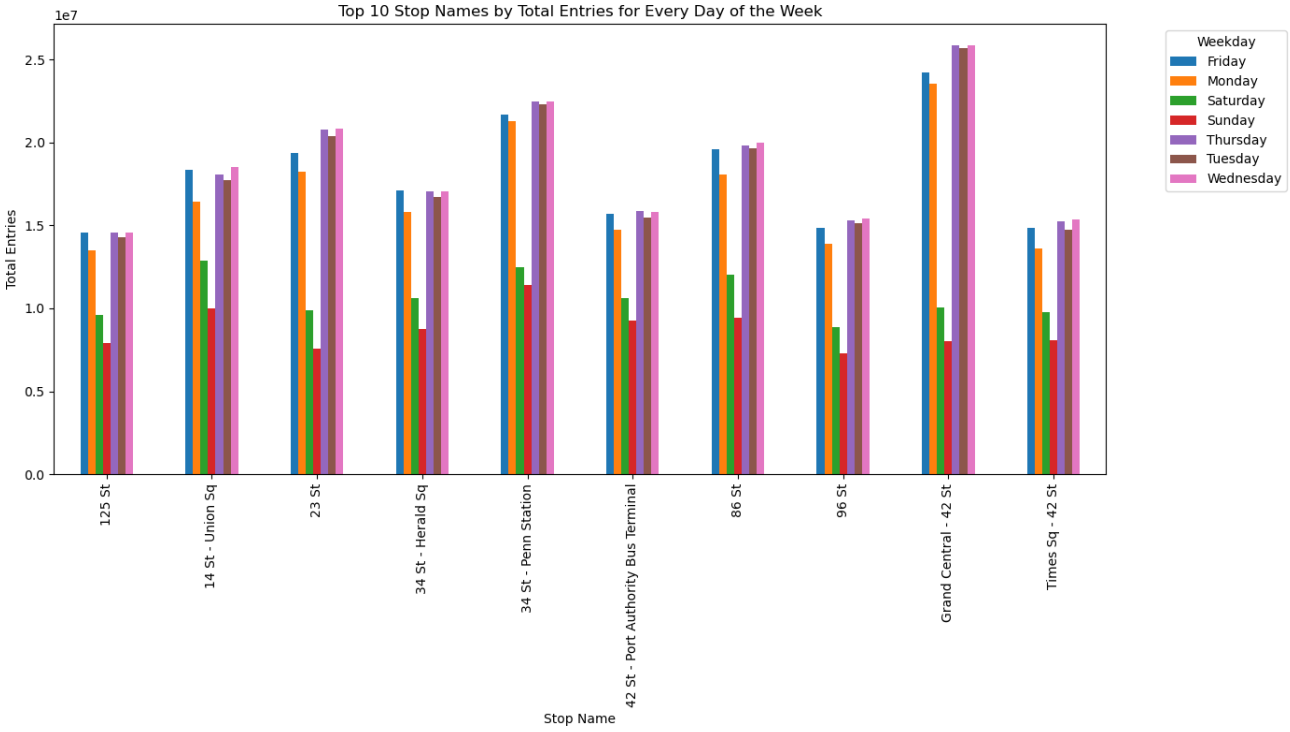
Εικόνα 34. Ιστόγραμμα των εισόδων (αριστερά) και των εξόδων (δεξιά) από τις στάσεις του μετρό στην Νέα Υόρκη.

Από τα ιστογράμματα συχνοτήτων του αριθμού εισερχομένων και εξερχομένων από τις στάσεις του μετρό της Νέας Υόρκης ανά τέσσερις ώρες, παρατηρούμε ότι τα χαρακτηριστικά ακολουθούν παρόμοια κατανομή, κάτι που είναι αναμενόμενο, ενώ η πιο συχνή τιμή βρίσκεται στο εύρος από 0 έως 500 είσοδοι/έξοδοι από μία στάση μετρό. Τα αποτελέσματα αυτά συμφωνούν και με τα συμπεράσματα που βγήκαν από τα μέτρα Περιγραφικής Στατιστικής που υπολογίστηκαν.



Εικόνα 35. Διαγράμματα συχνοτήτων γεωγραφικού προαστίου (επάνω αριστερά), τύπου δομής στάσης (επάνω δεξιά) και γειτονιάς (κάτω)

Παρουσιάζουμε το διάγραμμα συχνοτήτων για τα γεωγραφικά χαρακτηριστικά των ταξιδιών με μετρό στην Νέα Υόρκη, δηλαδή το προάστιο και την γειτονιά, ταξινομημένο σε φθίνουσα σειρά. Το μεγαλύτερο και πιο ανεπτυγμένο δίκτυο μετρό βρίσκεται στο Μανχάταν και στο Μπρούκλιν. Από το διάγραμμα συχνοτήτων του τύπου δομής του σταθμού, παρατηρούμε ότι ο πιο συχνός τύπος είναι η υπόγεια στάση, και αμέσως μετά έρχεται η υπερυψωμένη στάση, που όμως έχει μισή συχνότητα εμφάνισης από αυτήν του υπόγειου σταθμού.



Εικόνα 36. Διάγραμμα των δέκα πιο δημοφιλών στάσεων για κάθε ημέρα της εβδομάδας.

Από το διάγραμμα του πλήθους των εισόδων ανά ημέρα της εβδομάδας για τις δέκα πιο δημοφιλείς στάσεις του μετρό της Νέας Υόρκης, παρατηρούμε ότι τις καθημερινές έχουμε σχετικά σταθερή χρήση του μετρό ανά ημέρα, με μία μειωμένη συχνότητα την Δευτέρα, η οποίο το Σαββατοκύριακο είναι αισθητά μειωμένη. Συμπεραίνουμε ότι το μετρό χρησιμοποιείται ως μέσο μεταφοράς πιο συχνά για την μετακίνηση για εργασία ενώ οι κάτοικοι της Νέας Υόρκης τα Σαββατοκύριακα χρησιμοποιούν περισσότερο άλλα μέσα για την μετακίνησή τους, ή δεν μετακινούνται, κάτι που δεν φαίνεται πιθανό σε μία τόσο ζωντανή πόλη. Η μείωση χρήσης του μετρό την Δευτέρα σε σχέση με τις άλλες εργάσιμες μπορεί να οφείλεται σε υβριδικό μοντέλο απασχόλησης, ή στο ότι μεγάλες επιχειρήσεις της πόλης είναι κλειστές την ημέρα αυτή, όπως κομμωτήρια και εστιατόρια. Παρατηρείται ότι όσο

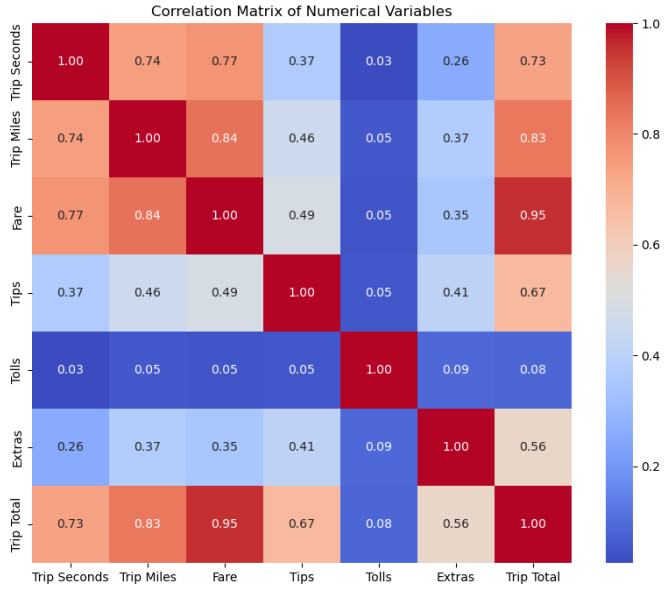
πιο κεντρικός είναι ο σταθμός του μετρό, τόσο μεγαλύτερη είναι η διαφορά μεταξύ του αριθμού εισερχομένων κατά τις εργάσιμες μέρες σε σχέση με το Σαββατοκύριακο. Για παράδειγμα στον σταθμό του Grand Central, ο αριθμός των εισερχομένων αυτός φτάνει ως το 1/3 σε σχέση με τις καθημερινές.

4.3 Συσχέτιση και Συχνά Μοτίβα

Σε αυτό το κεφάλαιο θα παρουσιάσουμε τα αποτελέσματα που αφορούν το πώς σχετίζονται τα διαφορετικά χαρακτηριστικά των δεδομένων μας μεταξύ τους. Σύμφωνα με το υποκεφάλαιο *3.2.6 Βήμα 6 και 3.2.8 Βήμα 8*, για κάθε σύνολο δεδομένων, υπολογίζουμε τον πίνακα συσχέτισης των αριθμητικών features και εφαρμόζουμε τον αλγόριθμο FP-Growth. Παρακάτω έχουμε τα αποτελέσματα για κάθε σύνολο δεδομένων χωριστά.

4.3.1 Μισθωμένα οχήματα – Taxi

- *Taxi Database – Chicago*



Εικόνα 37. Πίνακας συσχέτισης αριθμητικών χαρακτηριστικών για τα ταξί του Σικάγο

Από τον πίνακα συσχέτισης παρατηρούμε ότι υπάρχει υψηλή συσχέτιση μεταξύ των χαρακτηριστικών του χρόνου και της απόστασης του ταξιδιού και του καθαρού και συνολικού κόστους. Το αποτέλεσμα αυτό είναι λογικό καθώς όσο περισσότερο διαρκεί ένα ταξίδι με ταξί τόσο μεγαλύτερη απόσταση αναμένεται να διανύσουμε και τόσο περισσότερο αναμένεται να κοστίσει η διαδρομή. Ανάλογα αποτελέσματα αναμένουμε να δούμε και στο σύνολο δεδομένων ταξιδιού με ταξί στην Νέας

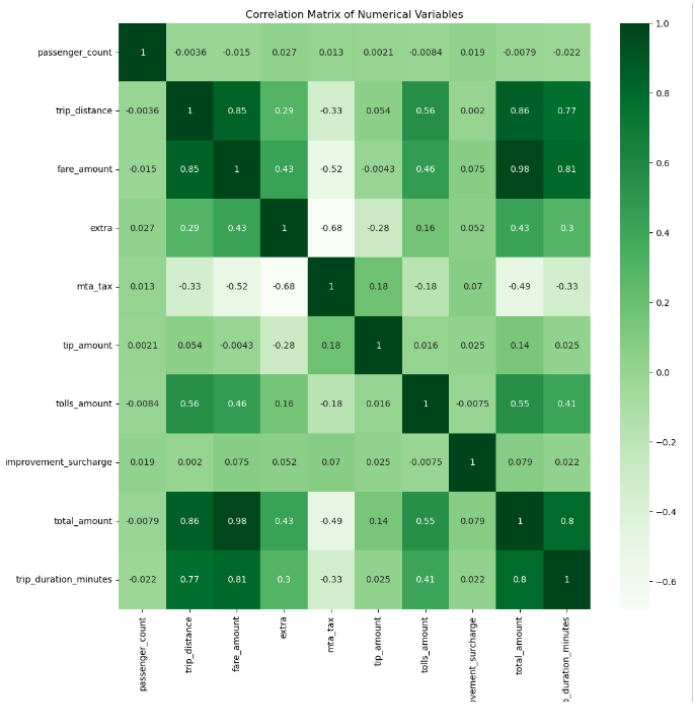
Υόρκης. Το χαρακτηριστικό του κόστους των διοδίων δεν φαίνεται να έχει κάποια συσχέτιση με τα υπόλοιπα χαρακτηριστικά. Τέλος το χαρακτηριστικό της έξτρα χρέωσης έχει μία συσχέτιση με το χαρακτηριστικό του φιλοδωρήματος, η οποία όμως είναι λιγότερο έντονη σε σχέση με την συσχέτισή του με το συνολικό κόστος της διαδρομής.

Υποστήριξη (Support)	**Σύνολο δεδομένων (Itemsets)**
0.899419	(low extras)
0.595283	(working hours)
0.575495	(short trip)
0.555451	(low extras, short trip)
0.550630	(working hours, low extras)
0.481377	(regular trip duration)
0.461208	(low extras, regular trip duration)
0.385054	(short trip, regular trip duration)
0.376719	(low extras, short trip, regular trip duration)

Πίνακας 28. Αποτελέσματα FP-Growth για ταξί στο Σικάγο

Εδώ παραθέτουμε τα αποτελέσματα του αλγορίθμου FP-Growth για αναγνώριση μοτίβων στα χαρακτηριστικά του συνόλου δεδομένων. Όσο πιο κοντά στο 1 βρίσκεται το αποτέλεσμα της υποστήριξης, τόσο πιο κοινό είναι το συγκεκριμένο υποσύνολο στο οποίο παρατηρείται ένας συνδυασμός τιμών για τα χαρακτηριστικά, στα δεδομένα μας. Από εδώ παρατηρούμε ότι έχουμε πολλές διαδρομές ταξί στο Σικάγο που λαμβάνουν χώρα κατά τις ώρες εργασίας και έχουν χαμηλές έξτρα χρεώσεις. Ένας συνδυασμός τιμών για τα χαρακτηριστικά της έξτρα χρέωσης, της απόστασης και της διάρκειας του ταξιδιού με ταξί, είναι το πιο συχνό μοτίβο τριπλέτας (όχι ζευγάρι ή μόνο ένα χαρακτηριστικό). Συγκεκριμένα, στο σύνολο δεδομένων ταξιδιού με ταξί στο Σικάγο, μία σύντομη χρονικά διαδρομή που έχει κανονική απόσταση έχει συχνά χαμηλές έξτρα χρεώσεις.

- *Taxi Database – New York City*



Εικόνα 38. Πίνακας συσχέτισης αριθμητικών χαρακτηριστικών για τα ταξί της Νέας Υόρκης

Όπως και στο σύνολο δεδομένων ταξιδιού με ταξί στο Σικάγο, στην Νέα Υόρκη, παρατηρείται υψηλή συσχέτιση μεταξύ των χαρακτηριστικών διάρκεια ταξιδιού, απόσταση ταξιδιού, καθαρό κόστος και συνολικό κόστος για το ταξίδι. Τα χαρακτηριστικά της έξτρα χρέωσης και της χρέωσης διοδίων έχουν μία συσχέτιση με τα χαρακτηριστικά της διάρκειας, απόστασης και του καθαρού και συνολικού κόστους για το ταξίδι με ταξί. Παρατηρείται επίσης ότι το χαρακτηριστικό της έξτρα χρέωσης είναι αρνητικά συσχετισμένο με τον MTA φόρο και το φιλοδώρημα, που σημαίνει ότι υπάρχει μία αντιστρόφως ανάλογη σχέση. Ο MTA φόρος έχει μία ήπια αρνητική συσχέτιση με την διάρκεια, την απόσταση, την έξτρα χρέωση, όπως αναφέρθηκε, την χρέωση διοδίων και τα κόστη του ταξιδιού, ενώ έχει μία μικρή θετική συσχέτιση με το φιλοδώρημα. Αυτό σημαίνει ότι όταν χρεώνεται ο φόρος αυτός στο ταξίδι με ταξί είναι πιο πιθανό ο επιβάτης να δώσει φιλοδώρημα για την διαδρομή του. Επίσης βλέπουμε ότι ο αριθμός των επιβατών έχει αμελητέα συσχέτιση, θετική ή αρνητική κατά περίπτωση, με τα υπόλοιπα χαρακτηριστικά το οποίο είναι ενδιαφέρον αποτέλεσμα καθώς θα περιμέναμε ότι περισσότερα άτομα θα επηρέαζαν το κόστος και την απόσταση του ταξιδιού, ενώ φαίνεται ότι υπάρχει μία αμελητέα αντίστροφη σχέση μεταξύ των μεγεθών αυτών και του πλήθους επιβατών. Είναι πιθανόν ότι μία διαδρομή με πάνω από έναν επιβάτη συνήθως αφορά τον ίδιο

προορισμό και χρεώνεται μόνο στον έναν, επομένως το κόστος δεν μοιράζεται και δεν επηρεάζεται από τον αριθμό επιβατών.

Υποστήριξη (Support)	**Σύνολο δεδομένων (Itemsets)**
0.980924	(low extras)
0.713379	(short trip)
0.705129	(short trip, low extras)
0.645886	(working hours)
0.631029	(working hours, low extras)
0.551528	(regular trip duration)
0.545033	(low extras, regular trip duration)
0.529441	(solo trip)
0.529441	(solo trip, low extras)
0.507740	(short trip, regular trip duration)
0.502252	(short trip, low extras, regular trip duration)

Πίνακας 29. Αποτελέσματα FP-Growth για ταξί στην Νέα Υόρκη

Στα αποτελέσματα του FP-Growth για την Νέα Υόρκη παρατηρούμε παρόμοια αποτελέσματα με αυτά που έδωσε για το σύνολο δεδομένων ταξί στο Σικάγο. Ένα χαρακτηριστικό συχνό μοτίβο είναι το σύντομο ταξίδι με χαμηλές έξτρα χρεώσεις και κανονική απόσταση, όπως είδαμε και στα ταξίδια με ταξί στην Νέα Υόρκη. Επίσης εδώ που έχουμε το χαρακτηριστικό του πλήθους των επιβατών, ένα συχνό μοτίβο είναι το ταξίδι με έναν μόνο επιβάτη, αλλά και ο συνδυασμός του ταξιδιού με έναν επιβάτη και χαμηλές έξτρα χρεώσεις. Παρατηρούμε και εδώ λοιπόν τα πιο συχνά μοτίβα περιέχουν έναν συνδυασμό των ωρών εργασίας, σύντομο ταξίδι, κανονική απόσταση και χαμηλά έξτρα.

- *Uber Database – New York City*

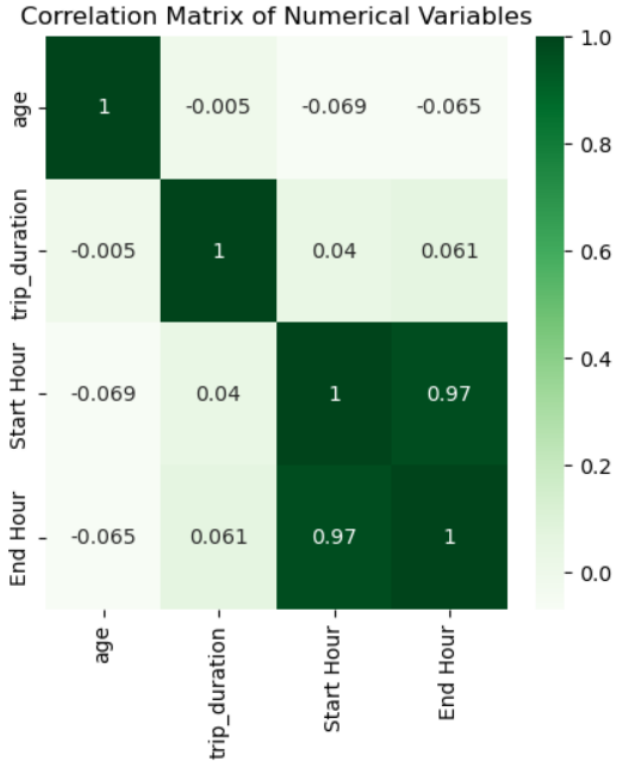
Υποστήριξη (Support)	**Σύνολο δεδομένων (Itemsets)**
0.75942	(Manhattan)

Πίνακας 30. Αποτελέσματα FP-Growth για Uber στην Νέα Υόρκη

Στην περίπτωση του Uber της Νέας Υόρκης δεν έχουμε πολλά features και βλέπουμε ότι από τα χαρακτηριστικά δεν προκύπτει κάποιο μοτίβο πέραν του ότι τα περισσότερα ταξίδια αφορούν το Manhattan όπως είδαμε και σε προηγούμενα αποτελέσματα.

4.3.2 Ενοικιαζόμενα Ποδήλατα

- *Bike Database – New York City*



Εικόνα 39. Πίνακας συσχέτισης αριθμητικών χαρακτηριστικών για τα ποδήλατα στην Νέα Υόρκη

Από τον πίνακα συσχέτισης των αριθμητικών χαρακτηριστικών για τα ποδήλατα στην Νέα Υόρκη παρατηρούμε υψηλή συσχέτιση μόνο μεταξύ της ώρας εκκίνησης και της ώρας λήξης του ταξιδιού, το οποίο είναι και αναμενόμενο καθώς η διάρκεια του ταξιδιού είναι μικρή και δεν ξεπερνάει την μία ώρα στο σύνολο δεδομένων μετά την αφαίρεση των ακραίων τιμών που έχουμε κάνει. Βλέπουμε επίσης ότι η διάρκεια του ταξιδιού δεν έχει συσχέτιση με την ώρα εκκίνησης, αλλά ούτε και με την ηλικία του χρήστη.

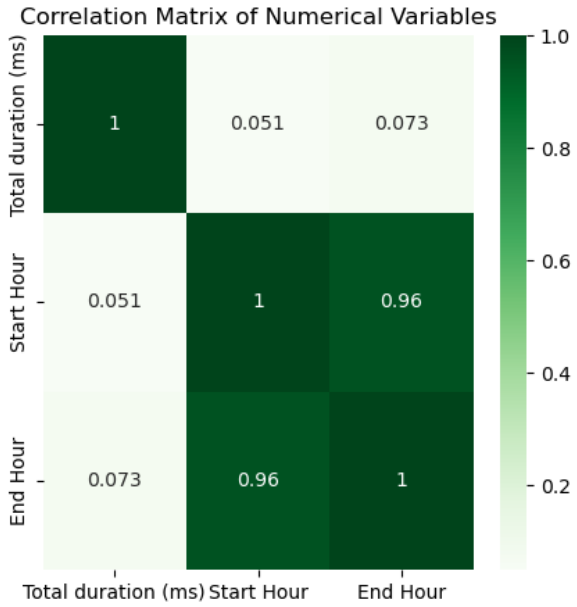
Υποστήριξη (Support)	**Σύνολο δεδομένων (Itemsets)**
0.740270	(male)
0.628766	(regular trip)
0.570394	(working hours)
0.466041	(regular trip, male)
0.415771	(working hours, male)
0.361215	(regular trip, working hours)
0.345709	(adults)

0.342102	(middle age)
0.331215	(afternoon)

Πίνακας 31. Αποτελέσματα FP-Growth για ποδήλατα στην Νέα Υόρκη

Από τα αποτελέσματα του αλγορίθμου FP-Growth στο σύνολο δεδομένων ταξιδιού με ποδήλατο στην Νέα Υόρκη, παρατηρούμε ότι όλοι οι συνδυασμοί ζευγαριού που σχηματίζουν οι τιμές των χαρακτηριστικών: άνδρας επιβάτης, ταξίδι κατά τις εργάσιμες ώρες, ταξίδι με κανονική απόσταση, αποτελούν τα πιο συχνά ζευγάρια στο σύνολο δεδομένων. Αυτά επιβεβαιώνονται και από την προηγούμενη ανάλυση που έχουμε κάνει.

- *Bike Database – London*



Εικόνα 40. Πίνακας συσχέτισης αριθμητικών χαρακτηριστικών για τα ποδήλατα στο Λονδίνο

Από τον πίνακα συσχέτισης των αριθμητικών χαρακτηριστικών για το σύνολο δεδομένων ταξιδιού με ποδήλατο στο Λονδίνο, βλέπουμε τα ίδια αποτελέσματα σε σχέση με την Νέα Υόρκη, δηλαδή υψηλή συσχέτιση μεταξύ της ώρας εκκίνησης και της ώρας που τελειώνει το ταξίδι καθώς και πολύ χαμηλή συσχέτιση της διάρκειας του ταξιδιού με αυτά τα δύο χαρακτηριστικά.

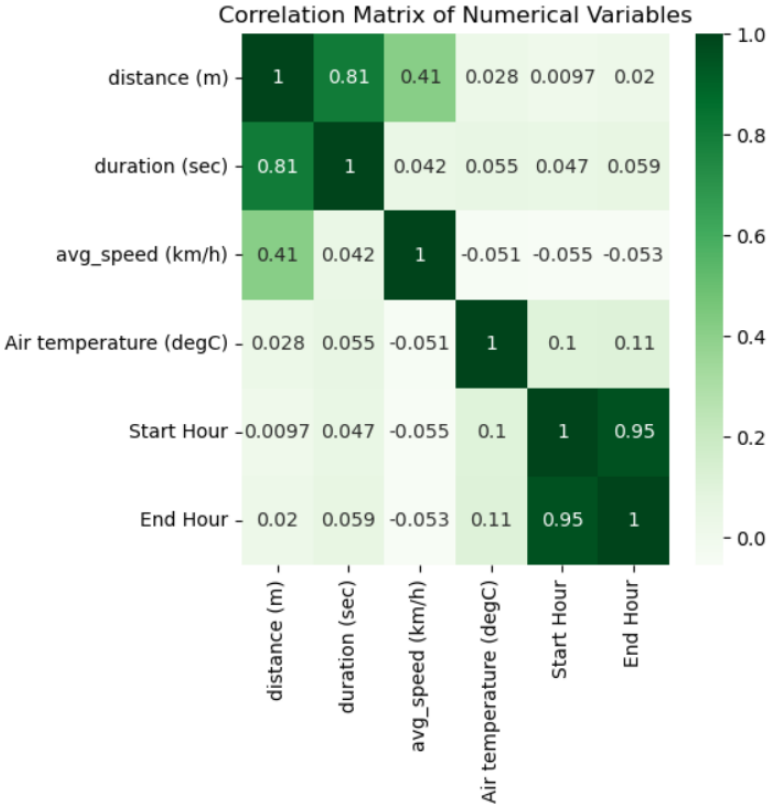
Υποστήριξη (Support)	**Σύνολο δεδομένων (Itemsets)**
0.590612	(regular trip)
0.554534	(working hours)
0.345991	(afternoon)

0.328645	(regular trip, working hours)

Πίνακας 32. Αποτελέσματα FP-Growth για ποδήλατα στο Λονδίνο

Ο αλγόριθμος FP-Growth δεν έδωσε μοτίβα με υποστήριξη άνω του χρησιμοποιούμενου κατωφλίου στο σύνολο δεδομένων ταξιδιού με ποδήλατο στο Λονδίνο. Ένα αρκετά συχνό μοτίβο είναι ο συνδυασμός ενός ταξιδιού με κανονική απόσταση που λαμβάνει χώρα κατά τις ώρες εργασίας, το οποίο είναι ένα μοτίβο που το είδαμε και στην Νέα Υόρκη και θα το δούμε στο Ελσίνκι.

- *Bike Database – Helsinki*



Εικόνα 41. Πίνακας συσχέτισης αριθμητικών χαρακτηριστικών για τα ποδήλατα στο Ελσίνκι

Από τον πίνακα συσχέτισης των αριθμητικών χαρακτηριστικών για το σύνολο δεδομένων ταξιδιού με ποδήλατο στο Ελσίνκι, παρατηρούμε υψηλή συσχέτιση μεταξύ της ώρας εκκίνησης και λήξης του ταξιδιού. Το χαρακτηριστικό της απόστασης έχει αρκετά υψηλή συσχέτιση με την διάρκεια του ταξιδιού και μία σχετικά χαμηλότερη συσχέτιση με την μέση ταχύτητα του ποδηλάτου κατά το ταξίδι. Τα υπόλοιπα χαρακτηριστικά έχουν αμελητέα συσχέτιση μεταξύ τους, ενώ παρατηρείται ειδικότερα ότι η θερμοκρασία αέρα έχει χαμηλή αλλά θετική συσχέτιση με την ώρα εκκίνησης και την ώρα λήξης, και μία μικρή αρνητική συσχέτιση με την

μέση ταχύτητα, κάτι που είναι λογικό καθώς έχουν αντιστρόφως ανάλογη σχέση κατά την κίνηση.

Υποστήριξη (Support)	**Σύνολο δεδομένων (Itemsets)**
0.999981	(very slow)
0.649851	(regular trip)
0.649851	(very slow, regular trip)
0.568280	(very slow, regular temp)
0.548950	(working hours)
0.548939	(very slow, working hours)
0.368188	(regular temp, regular trip)
0.368188	(very slow, regular temp, regular trip)
0.359552	(working hours, regular trip)
0.359552	(very slow, working hours, regular trip)

Πίνακας 33. Αποτελέσματα FP-Growth για ποδήλατα στο Ελσίνκι

Από τα αποτελέσματα του αλγορίθμου FP-Growth στο σύνολο δεδομένων ταξιδιού με ποδήλατο στο Ελσίνκι, παρατηρούμε ότι η πολύ χαμηλή τιμή για το χαρακτηριστικό της μέσης ταχύτητας και η κανονική θερμοκρασία κατά το ταξίδι βρίσκονται σε αρκετούς συνδυασμούς συχνών μοτίβων στο σύνολο δεδομένων. Αυτό μπορεί να εξηγηθεί από το ότι οι ταχύτητες είναι πολύ χαμηλές σε συντριπτικό ποσοστό στο σύνολο δεδομένων, ενώ λογικό είναι η χαμηλή θερμοκρασία που διευκολύνει το ταξίδι με ποδήλατο να εμφανίζεται συχνά σε συνδυασμό με άλλες τιμές χαρακτηριστικών, όπως την κανονική απόσταση ταξιδιού. Ένας συνδυασμός τιμών για τα χαρακτηριστικά της μέσης ταχύτητας, της απόστασης και της θερμοκρασίας του ταξιδιού με ποδήλατο, είναι το πιο συχνό μοτίβο τριπλέτας (όχι ζευγάρι ή μόνο ένα χαρακτηριστικό). Συγκεκριμένα, στο σύνολο δεδομένων ταξιδιού με ποδήλατο στο Ελσίνκι, μία διαδρομή που έχει κανονική απόσταση γίνεται συχνά σε πολύ χαμηλή ταχύτητα και σε κανονική θερμοκρασία. Αμέσως μετά σε συχνότητα ακολουθεί η διαδρομή που έχει κανονική απόσταση και γίνεται σε πολύ χαμηλή ταχύτητα, κατά τις ώρες εργασίας.

4.3.3 Μέσα Μαζικής Μεταφοράς

- *Bus Database – New York City*

Υποστήριξη (Support)	**Σύνολο δεδομένων (Itemsets)**
0.487285	(working hours)
0.373113	(approaching)
0.317434	(night)
0.310771	(< 1 stop away)
0.306049	(Brooklyn)

Πίνακας 34. Αποτελέσματα FP-Growth για λεωφορεία στην Νέα Υόρκη

Ο αλγόριθμος FP-Growth δεν έδωσε μοτίβα με υποστήριξη άνω του χρησιμοποιούμενου κατωφλίου στο σύνολο δεδομένων ταξιδιού με λεωφορείο στην Νέα Υόρκη, καθώς δεν έχουμε πολλά χαρακτηριστικά. Οι τιμές των χαρακτηριστικών που βλέπουμε μεμονωμένα εδώ σαν αποτελέσματα, όπως το ταξίδι κατά τις ώρες εργασίας ή την νύχτα, η τοποθεσία επιβίβασης που είναι το Μπρούκλιν, ή η σχετική απόσταση του λεωφορείου απ' την στάση με βάση τον προγραμματισμό, συμφωνούν με τις πιο συχνές κατηγορίες των χαρακτηριστικών, όπως αναφέρθηκε στο προηγούμενο υποκεφάλαιο.

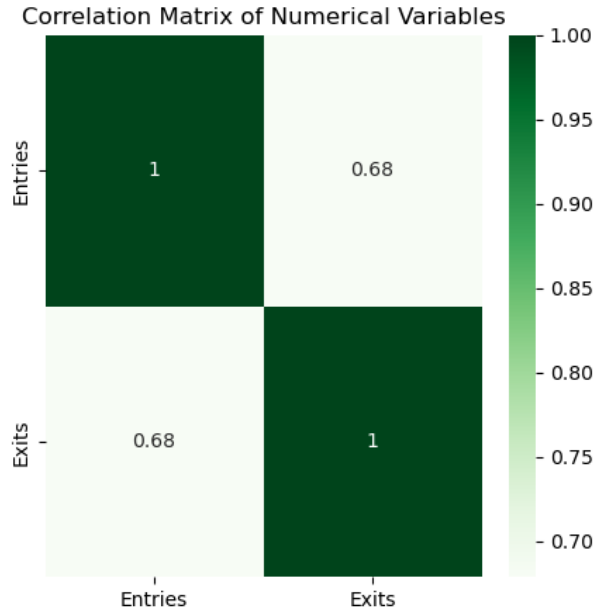
- *Bus Database – Rio De Janeiro*

Υποστήριξη (Support)	**Σύνολο δεδομένων (Itemsets)**
0.605770	(working hours)
0.389338	(very slow)

Πίνακας 35. Αποτελέσματα FP-Growth για λεωφορεία στο Ρίο Ντε Τζανέιρο

Ο αλγόριθμος FP-Growth δεν έδωσε συχνά ζευγάρια με υποστήριξη άνω του χρησιμοποιούμενου κατωφλίου στο σύνολο δεδομένων ταξιδιού με λεωφορείο στο Ρίο Ντε Τζανέιρο, καθώς δεν έχουμε πολλά χαρακτηριστικά. Το ταξίδι με λεωφορείο στην πόλη αυτή λαμβάνει συχνά χώρα κατά τις ώρες εργασίας και έχει πολύ χαμηλή μέση ταχύτητα.

- *Metro Database – New York City*



Εικόνα 42. Πίνακας συσχέτισης αριθμητικών χαρακτηριστικών
για το μετρό της Νέας Υόρκης

Από τον πίνακα συσχέτισης των αριθμητικών χαρακτηριστικών για το σύνολο δεδομένων ταξιδιού με μετρό στην Νέα Υόρκη, βλέπουμε ότι υπάρχει συσχέτιση μεταξύ των εισόδων και των εξόδων από μία στάση του μετρό όπως έχουμε σχολιάσει και στην ανάλυση παραπάνω αλλά δεν είναι όσο υψηλή θα περιμέναμε γεγονός που μας δείχνει ότι κάποιες στάσεις είναι πιο δημοφιλείς για αναχώρηση από ότι για άφιξη και το αντίστροφο.

Υποστήριξη (Support)	**Σύνολο δεδομένων (Itemsets)**
0.610004	(Subway)
0.575298	(low people)
0.503536	(working hours)
0.387027	(regular people)
0.346620	(Monday)
0.338048	(Brooklyn)
0.335911	(Monday, Subway)
0.306565	(working hours, Subway)

Πίνακας 36. Αποτελέσματα FP-Growth για το μετρό της Νέας Υόρκης

Ο αλγόριθμος FP-Growth δεν έδωσε συχνά ζευγάρια ή τριπλέτες, με υποστήριξη άνω του χρησιμοποιούμενου κατωφλίου, στο σύνολο δεδομένων ταξιδιού με μετρό στην

Νέα Υόρκη, ενώ περιέχει αρκετά χαρακτηριστικά. Ένας συνδυασμός τιμών χαρακτηριστικών που αφορούν την ημέρα της εβδομάδας και τον τύπο κατασκευής του μετρό, εμφανίζεται στα δεδομένα με υποστήριξη 33.5%, ενώ ακολουθεί ένας συνδυασμός τιμών για την ώρα της ημέρας και τον τύπο της στάσης με υποστήριξη 30.65%. Ειδικότερα, ένα ταξίδι με υπόγειο μετρό στην Νέα Υόρκη συχνά λαμβάνει χώρα την Δευτέρα και ένα ταξίδι με μετρό στην Νέα Υόρκη κατά τις εργάσιμες ώρες είναι συχνά σε υπόγεια στάση. Αυτό είναι αναμενόμενο αν σκεφτεί κανείς ότι το μοτίβο αυτό συσχετίζει την υπόγεια στάση του μετρό, που έχει αναλογία 2 προς 1 σε σχέση με τους άλλους τύπους κατασκευής, με τις πλέον συχνές κατηγορίες ώρας και ημέρας για το ταξίδι με μετρό στην Νέα Υόρκη.

5

Επίλογος

5.1 Σύνοψη και συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήθηκαν τα Πολυτροπικά Συστήματα Μεταφοράς σε μεγάλες πόλεις της Ευρώπης και της Αμερικής με στόχο την κατανόηση των προκλήσεων και την εύρεση τρόπων για την βελτιστοποίηση των μετακινήσεων. Χρησιμοποιήθηκαν προηγμένα εργαλεία και τεχνικές ανάλυσης δεδομένων για την προεπεξεργασία και ανάλυση μεγάλων συνόλων δεδομένων μεταφορών.

Χρησιμοποιήθηκαν δεδομένα από πόλεις με διαφορετικά χαρακτηριστικά, γεωγραφικά και κοινωνικο-οικονομικά, όπως Λονδίνο, Ρίο Ντε Τζανέιρο, Ελσίνκι κ.ά., κάνοντας κύρια αναφορά στη Νέα Υόρκη, για την ανίχνευση των μοτίβων μετακίνησης και την βελτιστοποίηση της χρήσης διαφόρων μέσων μεταφοράς. Συγκεκριμένα εξετάσαμε τα ταξί, λεωφορεία, ενοικιαζόμενα ποδήλατα και μετρό. Η ανάλυση αυτή υπέδειξε διάφορες συνήθειες στην χρήση του κάθε μέσου μετακίνησης, όπως την μετακίνηση κατά τις ώρες αιχμής, την χρήση τους από μέρος του πληθυσμού με συγκεκριμένα δημογραφικά χαρακτηριστικά,, τις διανυόμενες αποστάσεις και τι είδους διαδρομές πραγματοποιούνται με κάθε μέσο. Με βάση τα αποτελέσματα αυτά, μπορεί κανείς να πει ότι η χρήση συνδυαστικών μέσων μεταφοράς μπορεί να συμβάλει στη μείωση της κυκλοφοριακής συμφόρησης και στη βελτίωση της αποδοτικότητας των αστικών μετακινήσεων.

Κάποια ενδιαφέροντα συμπεράσματα από τα δεδομένα που επεξεργάστηκαν είναι:

- Στα ταξίδια με ταξί και τρόπο πληρωμής με κάρτα όταν υπήρχε φιλοδώρημα ήταν πολύ υψηλό.
- Τα ταξί χρησιμοποιούνται κατά το τυπικό ωράριο εργασίας, έως και δύο ώρες νωρίτερα και αργότερα, επομένως κατά πάσα πιθανότητα για μετάβαση από και προς την εργασία. Σε αντίθεση το Uber χρησιμοποιείται κατά τις ώρες αιχμής αλλά και κατά τις απογευματινές ώρες, μέχρι τις 10 το βράδυ.
- Η πλειοψηφία των διαδρομών με ποδήλατο για τις τρεις πόλεις είχαν διάρκεια μεταξύ των 5 και 15 λεπτών.
- Το μετρό της Νέας Υόρκης χρησιμοποιείται σχετικά σταθερά υψηλά τις εργάσιμες ημέρες της εβδομάδας με μία αισθητή μείωση το Σαββατοκύριακο.

- Τα λεωφορεία στο Ρίο δεν λειτουργούν όλο το 24ώρο και κινούνται σε συγκεκριμένα σημεία της πόλης και στις βασικές κεντρικές αρτηρίες, κάτι που αποτυπώνει και τον γεωγραφικό και οικονομικό διαχωρισμό που υπάρχει στην πόλη.

Τα συγκριτικά αποτελέσματα μεταξύ πόλεων, για το ίδιο μέσο μετακίνησης, μας έδειξαν κάποιες μικροδιαφορές μεταξύ τους αλλά και πολλά κοινά χαρακτηριστικά για το πως χρησιμοποιείται ένα μέσο μεταφοράς σε αρκετά διαφορετικές πόλεις. Αυτό μας δείχνει ότι πολλά αποτελέσματα μπορούν να γενικευθούν και να εφαρμοστούν σε άλλες πόλεις για την ανάπτυξη Πολυτροπικών Συστημάτων Μεταφορών, παρέχοντας στους χρήστες προσωποποιημένες προτάσεις διαδρομών βασισμένες σε πραγματικά δεδομένα και προγνωστικά μοντέλα.

Τέλος, η μελέτη αυτή έθεσε τα θεμέλια για περαιτέρω έρευνα στην εφαρμογή των τεχνικών Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης για να επιτευχθεί η πρόγνωση ζήτησης και ατυχημάτων, και η βελτιστοποίηση των αστικών μετακινήσεων.

5.2 Μελλοντικές επεκτάσεις

Η παρούσα διπλωματική εργασία ανέδειξε πολυάριθμες δυνατότητες για μελλοντικές επεκτάσεις και βελτιώσεις στις πολυτροπικές αστικές μεταφορές. Μία από τις κύριες κατευθύνσεις για μελλοντική έρευνα είναι η ανάπτυξη και εφαρμογή προηγμένων αλγορίθμων Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης για την πρόγνωση της ζήτησης και τη βελτιστοποίηση των διαδρομών σε πραγματικό χρόνο.

Ένα άλλο σημαντικό πεδίο επέκτασης είναι η ενσωμάτωση περισσότερων και ποικιλόμορφων συνόλων δεδομένων διαφορετικών μέσων μεταφοράς από την ίδια πόλη, κατά το ίδιο χρονικό διάστημα, όπως δεδομένα από αισθητήρες IoT, κοινωνικά δίκτυα και εφαρμογές κινητών συσκευών. Η ανάλυση αυτών των δεδομένων μπορεί να προσφέρει βαθύτερες και πιο ολοκληρωμένες γνώση για τις προτιμήσεις και τις ανάγκες των επιβατών.

Επιπλέον, μία άλλη επέκταση είναι η ανάπτυξη Συστήματος Συστάσεων (Recommendation System) που μπορεί να συνδυάζει δεδομένα από διαφορετικούς παρόχους μεταφορών και να παρέχει προσωποποιημένες ολοκληρωμένες προτάσεις μετακίνησης για διαδρομές εντός πόλης που να περιλαμβάνουν πράσινες επιλογές

μετακίνησης όπως ποδήλατα, ηλεκτρικά πατίνια, περπάτημα και να λαμβάνουν υπόψιν και τις προσωπικές προτιμήσεις του χρήστη (πχ. αν είναι αθλητικός, αν προτιμάει το λεωφορείο κτλ.) και άλλους εξωτερικούς παράγοντες όπως ο καιρός.

6

Βιβλιογραφία

[1] "GPS data from Rio De Janeiro Buses" https://www.kaggle.com/datasets/igorbalteiro/gps-data-from-rio-de-janeiro-buses

[2] "Uber Pickups in New York City" https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city/data

[3] "Taxi Trip Data NYC" https://www.kaggle.com/datasets/anandaramg/taxi-trip-data-nyc?select=taxi_tripdata.csv

[4] "NYC Subway Traffic 2017-2021" https://www.kaggle.com/datasets/eddeng/nyc-subway-traffic-data-20172021

[5] "New York City Bus Data" https://www.kaggle.com/datasets/stoney71/new-york-city-transport-statistics/data

[6] "Bike Trips" https://www.kaggle.com/datasets/gabrielramos87/bike-trips/data

[7] "London Bike-Share Usage Dataset" https://www.kaggle.com/datasets/kalacheva/london-bike-share-usage-dataset

[8] "Helsinki City Bikes" https://www.kaggle.com/datasets/geometrein/helsinki-

city-bikes

[9] "Taxi Trips Chicago 2024" https://www.kaggle.com/datasets/adelanseur/taxi-trips-chicago-2024/data

[10] "Guidelines for City Mobility" WEF 2020 https://www3.weforum.org/docs/WEF_Guidelines_for_City_Mobility_2020.pdf

[11] Wikipedia: Intelligent Transportation System https://en.wikipedia.org/wiki/Intelligent_transportation_system

[12] "Mining of Massive Datasets" Jure Leskovec, Anand Rajaraman, Jeff Ullman p.319-353

[13] "Benefits of Public Transport" Net Zero Scotland https://www.netzeronation.scot/take-action/travel-less-car/benefits-public-transport#:~:text=Replacing%20car%20journeys%20with%20public,73%25%20if%20travelling%20by%20train.

[14] "How AI and Machine Learning Enchases the safety, efficiency and passenger comfort of public transport" Jan Haegeman, John Wright https://www.intelligenttransport.com/transport-articles/150030/ai-machine-learning-enhances-public-transport/

[15] "Machine Learning Applied to Public Transportation by Bus: A Systematic Literature Review" Alexandre, T., Bernardini, F., Viterbo, J., & Pantoja

[16] "Use of Machine Learning in understanding transport dynamics of land use and public transportation in a developing city" M.Dorosan, D.Dailisan, J.F.Valemzula, C.Monterola https://www.sciencedirect.com/science/article/pii/S0264275123003992

[17] Wikipedia: Genetic Algorithm https://en.wikipedia.org/wiki/Genetic_algorithm

[18] Wikipedia: Ant colony optimization algorithms https://en.wikipedia.org/wiki/Ant_colony_optimization_algorithms

[19] "A personalized recommendation system for multi-modal transportation systems" F. Wu, C. Lyu, Y. Liu https://www.sciencedirect.com/science/article/pii/S2772586322000168

[20] “A Hybrid Knowledge-based Recommender for Mobility-as-a-Service” K.Arnaoutaki, B.Magoutas, E.Bothos, G.Mentzas https://www.scitepress.org/PublishedPapers/2019/79214/79214.pdf

[21] Business Intelligence and Analytics https://thehamsterwheelinmyhead.wordpress.com/case-study-3-business-intelligence-and-analytics/