ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Ανάλυση κατά Συστάδες και Εφαρμογές σε Τραπεζικά δεδομένα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΑΤΙΑΜΠΑ ΜΑΡΘΑ

Επιβλέπων : Χρήστος Κουκουβίνος
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2014

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Ανάλυση κατά Συστάδες και Εφαρμογές σε Τραπεζικά δεδομένα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΑΤΙΑΜΠΑ ΜΑΡΘΑ

Επιβλέπων : Χρήστος Κουκουβίνος
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2014

ΠΕΡΙΛΗΨΗ

Η δημιουργία και ανάλυση του προφίλ των πελατών πολλών οργανισμών και επιχειρήσεων απαιτούν τη χρήση μεθόδων, όπως η ανάλυση κατά συστάδες, προκειμένου να ενισχύσουν τα κέρδη τους με την αποτελεσματική διαχείριση των προβλημάτων των πελατών τους. Η παρούσα διπλωματική εργασία ασχολείται με την εφαρμογή της ανάλυσης κατά συστάδες σε τραπεζικά δεδομένα των πελατών και τον διαχωρισμό τους σε ομάδες που καθορίζουν το πιστοληπτικό προφίλ τους. Με βάση τα αποτελέσματα της ανάλυσης, προτείνεται η συστάδα που συγκεντρώνει την υψηλότερη δυνητική ικανότητα αποπληρωμής, ώστε να λάβουν τις απαραίτητες διευκολύνσεις από την τράπεζα.

Η ανάλυση των δεδομένων, η οποία διεξήχθη χρησιμοποιώντας τα χαρακτηριστικά γνωρίσματα των πελατών που συμβάλλουν σημαντικά στη δημιουργία των συστάδων και με την εφαρμογή της μεθόδου TwoStep Clustering, είχε ως λύση τη διαμόρφωση τεσσάρων συστάδων. Ως αποτέλεσμα δημιουργήθηκε το προφίλ των πελατών για κάθε συστάδα και συνιστάται η συστάδα που πληροί τις προϋποθέσεις για επαναδιαπραγμάτευση ενός σχεδίου αποπληρωμής των χρεών τους προκειμένου να αυξηθούν τα κέρδη της τράπεζας.

ABSTRACT

The creation and analysis of customer profiles of many organizations require the use of methods such as cluster analysis in order to enhance their profits by effectively managing existing or upcoming customer problems. This thesis deals with the application of the cluster analysis technique in customer banking data by separating them into groups that determine the appropriate credit profiles. Based on the results of the analysis, the team concentrating the highest potential repayment capacity is being suggested so as to receive the appropriate facilitations by the bank.

The data analysis, which was conducted by using the customer features that contribute significantly to the creation of groups and by implementing the TwoStep Clustering method, resolved in four groups. As a result, a customer profile was created for each group and the group that qualified for renegotiating a repayment plan for their debts in order to grow the bank's profits was recommended.

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας πρωτίστως οφείλω να ευχαριστήσω θερμά τον Καθηγητή του Εθνικού Μετσοβίου Πολυτεχνείου κ. Χ. Κουκουβίνο τόσο για την εμπιστοσύνη που μου έδειξε με την ανάθεσή της καθώς μου έδωσε την ευκαιρία να ασχοληθώ με ένα εξαιρετικά ενδιαφέρον θέμα και να αποκομίσω ουσιαστικά προσόντα μέσα από αυτή την εργασία, αλλά και για την υπομονή του σε όλες τις δυσκολίες κατά τη διάρκεια εκπόνησης της εργασίας.

Παράλληλα, ευχαριστώ πολύ την υποψήφια διδάκτορα Χ. Παρπούλα για τις εποικοδομητικές τις παρατηρήσεις και την πολύτιμη βοήθεια που μου προσέφερε για την ολοκλήρωση της διπλωματικής εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω την Ε. Λυγκώνη για τις συμβούλες, τις υποδείξεις και τη βοήθειά της σε σημαντικά θέματα της διπλωματικής εργασίας.

Τέλος, ένα μεγάλο ευχαριστώ στην οικογένειά μου, στους γονείς μου και στα αδέρφια μου, στην ξαδέρφη μου Ο. Κωνσταντακοπούλου, στους φίλους μου, για την πολύτιμη βοήθεια, την κατανόηση, την ηθική στήριξη και τη συμπαράστασή τους κατά τη διάρκεια των σπουδών μου.

Στον παππού μου

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1. Εισαγωγή 12

1.1 Γενικά 12

1.2 Οργάνωση κειμένου 14

2. Η ανάλυση κατά συστάδες 15

2.1 Ορισμός 15

2.2 Είδη συσταδοποίησης – Τύποι συστάδων 15

2.3 Διαφορές της ανάλυσης κατά συστάδες με άλλες τεχνικές 20

2.4 Σκοπός- Χρησιμότητα 21

2.5 Εφαρμογές 22

3. Μέτρα εγγύτητας-Κριτήρια Συσταδοποίησης 24

3.1 Μέτρα εγγύτητας 24

3.1.1 Μέτρα απόστασης ή ανομοιότητας 25

3.1.2 Μέτρα ομοιότητας 29

3.2 Κριτήρια συσταδοποίησης 33

4. Τα στάδια της ανάλυσης κατά συστάδες 36

4.1 Επιλογή γνωρισμάτων (μεταβλητών) για το σχηματισμό ομάδων 37

4.2 Επιλογή μεθόδου ομαδοποίησης και εφαρμογή της 38

4.2.1 Ιεραρχικές μέθοδοι ομαδοποίησης 38

4.2.1.1 Συσσωρευτικές μέθοδοι ομαδοποίησης 41

4.2.1.2 Διαιρετικές μέθοδοι ομαδοποίησης 46

4.2.2 Μη ιεραρχικές μέθοδοι ομαδοποίησης 47

4.2.3 TwoStep Ανάλυση κατά συστάδες (TwoStep Cluster) 50

4.3 Έλεγχος εγκυρότητας ομαδοποίησης 52

4.4 Ερμηνεία αποτελέσματος 54

5. Εφαρμογή και ανάλυση δεδομένων 55

5.1 Εισαγωγή 55
5.2 Το δείγμα και οι μεταβλητές 56
5.3 Περιγραφικά μέτρα και Γραφήματα 57
5.4 Συσχετίσεις μεταξύ μεταβλητών 67
5.5 TwoStep Συσταδοποίηση: Αποτελέσματα αλγορίθμου- Προσαρμογή μοντέλου- Αριθμός ομάδων 78
5.5.1 Τα χαρακτηριστικά γνωρίσματα κάθε ομάδας και η διερεύνηση της σύστασής τους 81
5.5.2 Το προφίλ των συστάδων 91
5.5.3 Αποτελέσματα 96
6. Συμπεράσματα 97
ΒΙΒΛΙΟΓΡΑΦΙΑ 99

1. Εισαγωγή

1.1 Γενικά

«Οι τράπεζες είναι ενδιάμεσοι χρηματοοικονομικοί οργανισμοί που διαμεσολαβούν ανάμεσα στις πλεονασματικές και ελλειμματικές μονάδες της οικονομίας, βελτιώνοντας την αποτελεσματικότητα του χρηματοδοτικού της συστήματος.

Σύμφωνα με το νομοθετικό πλαίσιο της Ελλάδας και ειδικότερα το Νόμο 2076/1992, οι τράπεζες είναι πιστωτικά ιδρύματα που, ύστερα από την Άδεια Λειτουργίας που λαμβάνουν από την Τράπεζα της Ελλάδος, λειτουργούν ως επιχειρήσεις και δραστηριοποιούνται στην αποδοχή καταθέσεων ή άλλων επιστρεπτέων κεφαλαίων από το κοινό και στη χορήγηση πιστώσεων για λογαριασμό τους.» (Σαπουντζόγλου & Πεντότης, 2009)

Ανάμεσα στις ποικίλες λειτουργίες που διενεργούνται από τις τράπεζες είναι και η χορήγηση τραπεζικών προϊόντων που διακρίνονται στις ακόλουθες κατηγορίες:

1. Τα δάνεια (μακροπρόθεσμα, μεσοπρόθεσμα)
2. Τις πιστώσεις ανοικτού αλληλόχρεου λογαριασμού (βραχυπρόθεσμες δανειακές χορηγήσεις προς επιχειρήσεις)
3. Τις πιστωτικές κάρτες

Οι τράπεζες διατρέχονται από ένα πλήθος κινδύνων και μία από τις βασικές μορφές κινδύνων που αντιμετωπίζουν είναι ο πιστωτικός κίνδυνος. Ο πιστωτικός κίνδυνος αναφέρεται στο ενδεχόμενο της μη εκπλήρωσης των υποχρεώσεων φυσικών ή νομικών προσώπων προς την τράπεζα. Η τράπεζα λοιπόν ως χρηματοπιστωτικό ίδρυμα οδηγείται σε τρόπους αντιμετώπισης αθέτησης των υποχρεώσεων των πελατών τους. Ένας από τους τρόπους αντιμετώπισης αποτελεί και η ανάλυση του προφίλ των υπαρχόντων πελατών για την κατηγοριοποίηση και ταξινόμησή τους σε ομάδες που αντικατοπτρίζουν τις πιθανές ενέργειες τους και τις μελλοντικές τους αποφάσεις. Με βάση τα αποτελέσματα των αναλύσεων, οι τράπεζες λαμβάνουν αποφάσεις που αφορούν είτε τη χορήγηση προϊόντων σε νέους πελάτες είτε τη διαδικασία «χειρισμού» των υπαρχόντων πελατών. Αρωγός στη λήψη αποφάσεων των τραπεζών συνιστάται η εφαρμογή τεχνικών της Πολυμεταβλητής Στατιστικής Ανάλυσης.

Η Πολυμεταβλητή Στατιστική Ανάλυση αποτελεί ένα σημαντικό μέρος της Στατιστικής Ανάλυσης καθώς αφορά δεδομένα που περιέχουν μετρήσεις πολλών χαρακτηριστικών γνωρισμάτων (τα οποία χαρακτηριστικά εκπροσωπούνται στο σύνολο των δεδομένων υπό τη μορφή των μεταβλητών) για ένα σύνολο αντικειμένων τα οποία βρίσκονται υπό μελέτη (άτομα, οικονομικές μονάδες κ.λπ.). Στην πραγματικότητα, αποτελεί ένα ευρέως διαδεδομένο εργαλείο για τη διερεύνηση της φύσης και της σχέσης των δεδομένων αλλά και για την εξαγωγή συμπερασμάτων χρησιμοποιώντας πολλές μεταβλητές ανεξάρτητες και εξαρτημένες, οι οποίες όλες παρουσιάζουν ένα βαθμό συσχέτισης μεταξύ τους. Ακριβώς αυτό το γνώρισμα αποτελεί και το πλεονέκτημά τους έναντι των μεθόδων της μονομεταβλητής στατιστικής ανάλυσης.

Οι προηγμένες τεχνικές και μέσα αποθήκευσης μεγάλου όγκου δεδομένων πολλές φορές παρουσιάζουν εκτός από το συνεχώς αυξανόμενο πλήθος τους και μεγάλη ασάφεια και ποικιλότητα ως προς το είδος, τη σχέση και τη συνοχή τους. Αυτά τα χαρακτηριστικά των συνόλων δεδομένων (που συνήθως προκύπτουν από έρευνες με ερωτηματολόγια ή από βάσεις δεδομένων που συσσωρεύουν συνεχώς νέα δεδομένα) έχουν δημιουργήσει την ανάγκη αλλά και την προοπτική της ανάλυσής τους μέσω τεχνικών ταξινόμησης και ανάκτησης δεδομένων, καθώς η έλλειψη δομής των δεδομένων επιτάσσει μεθοδολογίες κατανόησης, επεξεργασίας και ομαδοποίησης των δεδομένων.

Μια ευρεία ταξινόμηση των τεχνικών ανάλυσης δεδομένων αναφέρεται στην ταξινόμηση τους σε δύο μεγάλες κατηγορίες:

(Α) Διερευνητικές ή περιγραφικές, δηλαδή χρησιμεύουν και χρησιμοποιούνται όταν δεν υπάρχουν εκ των προτέρων υποθέσεις ή υποδείγματα που περιγράφουν ή βοηθούν στην κατανόηση των χαρακτηριστικών γνωρισμάτων της δομής των δεδομένων.

(Β) Επιβεβαιωτικές ή Επαγωγικές, δηλαδή χρησιμεύουν και χρησιμοποιούνται όταν υπάρχουν εξαρχής υποθέσεις για τα χαρακτηριστικά των παρατηρήσεων και σκοπός των ερευνητών είναι ο έλεγχος και η επικύρωση αυτών των υποθέσεων δεδομένου του συνόλου των δεδομένων.

Στη βιβλιογραφία, συναντώνται πολλές τεχνικές πολυμεταβλητής στατιστικής ανάλυσης δεδομένων όπως η πολυμεταβλητή ανάλυση παλινδρόμησης, η λογιστική παλινδρόμηση, η

διακριτική ανάλυση, η ανάλυση επιβίωσης, η ανάλυση κατά κύριες συνιστώσες, η παραγοντική ανάλυση και η ανάλυση κατά συστάδες/ομάδες (Tabachnick & Fidell, 2007).

1.2 Οργάνωση κειμένου

Η διπλωματική εργασία παρουσιάζεται σύμφωνα με τα παρακάτω κεφάλαια:

Στο **δεύτερο** κεφάλαιο, αναλύεται η τεχνική της ανάλυσης κατά συστάδες με έμφαση στην λογική βάση στην οποία στηρίζεται η εφαρμογή της, τα είδη συσταδοποίησης (ιεραρχική, διαμεριστική κλπ), οι τύποι συστάδων, οι βασικότερες διαφορές της με άλλες τεχνικές και ειδικότερα οι έννοιες που τις διαφοροποιούν, ο σκοπός της και η χρησιμότητά της πέραν των αντικειμενικών της στόχων. Επίσης, δίνεται μια συνοπτική περιγραφή των σημαντικών εφαρμογών που βρίσκει αυτή η τεχνική σε πολλαπλά επιστημονικά πεδία.

Στο **τρίτο** κεφάλαιο, γίνεται εκτενής αναφορά στα μέτρα εγγύτητας, καθώς αποτελούν θεμελιώδη έννοια για την ανάλυση κατά συστάδες, καθώς η ομαδοποίηση των παρατηρήσεων γίνεται με βάση την προσεγγισιμότητά τους και πιο συγκεκριμένα στο μέτρο απόστασης ή μέτρο ανομοιότητας και το μέτρο ομοιότητας, με στόχο την κατάταξη στην ίδια ομάδα των παρατηρήσεων που έχουν μικρότερη απόσταση ή μεγαλύτερη ομοιότητα. Επίσης, γίνεται αναφορά και στα κριτήρια δημιουργίας συστάδων, σύμφωνα με τα οποία συγκεκριμένες μέθοδοι ομαδοποίησης αξιολογούν την ομοιογένεια και τη διαφορετικότητα μεταξύ των συστάδων.

Στο **τέταρτο** κεφάλαιο, αναλύονται τα 4 βασικά στάδια της διαδικασίας ομαδοποίησης των παρατηρήσεων όπως αυτή γίνεται με την ανάλυση κατά συστάδες.

Στο **πέμπτο** κεφάλαιο, αναλύονται τα δεδομένα μελετώντας το δείγμα και τα χαρακτηριστικά γνωρίσματά του (μεταβλητές) και αναλύονται οι συσχετίσεις τους. Γίνεται η εφαρμογή της ανάλυσης κατά συστάδες με τη μέθοδο ομαδοποίησης TwoStep που οδηγεί στη δημιουργία των ομάδων και του αντίστοιχου προφίλ τους.

Στο **έκτο** κεφάλαιο, παρουσιάζονται τα συμπεράσματα της ανάλυσης κατά συστάδες και προτείνεται η κατάλληλη ομάδα που ικανοποιεί τις αρχικές απαιτήσεις του προβλήματος.

2. Η ανάλυση κατά συστάδες

2.1 Ορισμός

Η τεχνική της ανάλυσης κατά συστάδες, σαν μέρος της πολυμεταβλητής στατιστικής ανάλυσης αναφέρεται στην κατάταξη σε ομάδες παρατηρήσεων προερχόμενες από ένα σύνολο δεδομένων (που περιέχουν πληροφορίες από ένα σύνολο μεταβλητών). Ουσιαστικά βάσει αυτών των μεταβλητών, εξετάζεται κατά πόσο οι παρατηρήσεις παρουσιάζουν ομοιότητα μεταξύ τους, δηλαδή απαρτίζουν ομάδες με τέτοιο βαθμό ομοιογένειας ώστε να θεωρούνται μέλη μόνο μίας συστάδας-ομάδας. Συνήθως στην ανάλυση κατά συστάδες δεν υπάρχει a priori γνώση για το πλήθος των ομάδων στις οποίες διαχωρίζονται οι παρατηρήσεις του δείγματος και ποιες παρατηρήσεις ταξινομούνται σε κάθε ομάδα. Αποτελεί μία ιδιαίτερη στατιστική τεχνική λόγω της διερευνητικής της φύσης καθώς είναι σχεδιασμένη με τέτοιο τρόπο ώστε να εντοπίζονται εσωτερικές σχέσεις ανάμεσα στα δεδομένα. Η λογική βάση πάνω στην οποία στηρίζεται η ανάλυση κατά συστάδες είναι, ότι η απόσταση μεταξύ των δεδομένων μπορεί να αποτελέσει έναν επαρκή δείκτη για την κατάταξή του σε ομάδες (η έννοια της απόστασης – ομοιότητας θα εξηγηθεί αναλυτικότερα σε επόμενη ενότητα). Οι ομάδες που επιδιώκουμε να δημιουργήσουμε θα απαρτίζονται από άτομα, μονάδες ή αντικείμενα με ομοειδή χαρακτηριστικά και συμπεριφορές.

2.2 Είδη συσταδοποίησης – Τύποι συστάδων

Στην ανάλυση κατά συστάδες διακρίνονται διάφορα είδη συσταδοποίησης (σύνολα από συστάδες) τα οποία συνοπτικά αναφέρονται παρακάτω (Tan, Steinbach, & Kumar, 2006):

- **Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)**
  Η ιεραρχική συσταδοποίηση αποτελεί ένα σύνολο εμφωλευμένων(nested) συστάδων, επιτρέποντας σε κάθε συστάδα να έχει υποσυστάδες (subclusters) οργανωμένες σε ένα ιεραρχικό δέντρο.

- **Διαμεριστική Συσταδοποίηση (Partitional Clustering)**

  Η διαμεριστική συσταδοποίηση αποτελεί τη διαμέριση του συνόλου των αντικειμένων των δεδομένων σε μη επικαλυπτόμενα υποσύνολα, τέτοια ώστε κάθε αντικείμενο να ανήκει ακριβώς σε ένα υποσύνολο.

- **Επικαλυπτόμενη – Μη Επικαλυπτόμενη Συσταδοποίηση (Exclusive - Non Exclusive or Overlapping Clustering)**

  Επικαλυπτόμενη καλείται η συσταδοποίηση που εκχωρεί κάθε αντικείμενο σε μία ενιαία συστάδα ενώ η μη επικαλυπτόμενη συσταδοποίηση αντικατοπτρίζει την ύπαρξη αντικειμένων σε διαφορετικές από μία συστάδες.

- **Ασαφής Συσταδοποίηση (Fuzzy Clustering)**

  Στην ασαφή συσταδοποίηση οι συστάδες αντιμετωπίζονται σαν ασαφή σύνολα, κάθε αντικείμενο ανήκει σε κάθε συστάδα με βάρος μεταξύ του 0 (δεν ανήκει απολύτως) και του 1 (ανήκει απολύτως). Επειδή τα βάρη για κάθε αντικείμενο έχουν άθροισμα τη μονάδα, η ασαφής συσταδοποίηση αδυνατεί να αντιμετωπίσει τις περιπτώσεις που κάθε αντικείμενο ανήκει σε περισσότερες από μία κατηγορίες.

- **Πλήρης – Μερική Συσταδοποίηση (Partial – Complete Clustering)**

  Η πλήρης συσταδοποίηση εκχωρεί κάθε αντικείμενο σε μία συστάδα ενώ σύμφωνα με τη μερική συσταδοποίηση δεν είναι απαραίτητο όλα τα αντικείμενα να ανήκουν σε καλώς καθορισμένες συστάδες, καθώς πολλά αντικείμενα σε ένα σύνολο δεδομένων πιθανόν να αντιπροσωπεύουν θόρυβο ή και ακραίες τιμές.

- **Ετερογενής – Ομογενής Συσταδοποίηση (Heterogeneous – Homogeneous Clustering)**

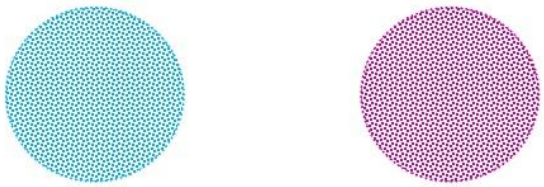
  Στην ετερογενή συσταδοποίηση εμφανίζονται συστάδες με διαφορές στα μεγέθη στο σχήμα και στην πυκνότητα.

Βάσει των παραπάνω, στόχος της συσταδοποίησης είναι η εύρεση χρήσιμων συστάδων (ομάδων αντικείμενων) που επιδιώκουν να ικανοποιήσουν τον σκοπό της ανάλυσης κατά

συστάδες. Στοχεύοντας σε μια επιτυχημένη ομαδοποίηση, οι συστάδες κατατάσσονται ανάλογα με τα χαρακτηριστικά τους στις παρακάτω κατηγορίες (Tan, Steinbach, & Kumar, 2006):

- **Καλώς διαχωρισμένες συστάδες (Well-separated clusters)**

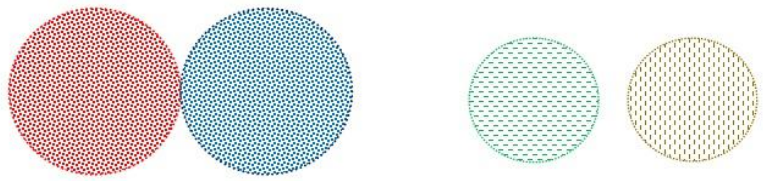
  Στις καλώς διαχωρισμένες συστάδες τα αντικείμενα εντός της συστάδας είναι πιο κοντά μεταξύ τους (περισσότερο όμοια). Παράλληλα η απόσταση μεταξύ δύο οποιονδήποτε αντικειμένων διαφορετικών συστάδων είναι μεγαλύτερη των αποστάσεων που υπάρχουν μεταξύ δύο οποιονδήποτε αντικειμένων εντός των συγκεκριμένων συστάδων. Εκτός από το σφαιρικό σχήμα, απεικονίζονται και με άλλα σχήματα. Στην εικόνα 2.1 παρουσιάζεται ένα παράδειγμα καλώς διαχωρισμένων συστάδων που αποτελείται από δύο ομάδες σημείων σε διδιάστατο χώρο.



Εικόνα 2.1 Καλώς διαχωρισμένες συστάδες

- **Συστάδες που βασίζονται στο «κέντρο» ή πρότυπο (Center-based clusters)**

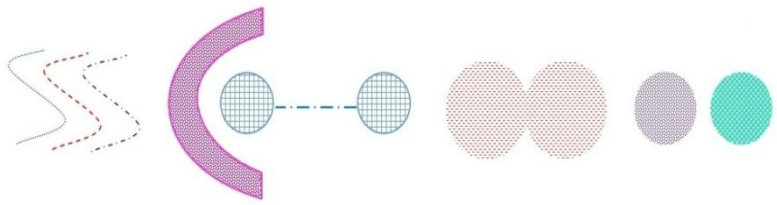
  Στην κατηγορία των συστάδων που βασίζονται στο «κέντρο» ή πρότυπο, κάθε συστάδα αποτελεί ένα σύνολο αντικειμένων τέτοιο ώστε κάθε αντικείμενο είναι πιο κοντά (ή περισσότερο όμοιο) με το «κέντρο» ή πρότυπο της συστάδας σε σύγκριση με την απόσταση (ομοιότητα) από το «κέντρο» οποιασδήποτε άλλης συστάδας. Οι συνήθεις μορφές του κέντρου είναι: α) centroid, δηλαδή του μέσου όρου των σημείων της συστάδας, β) του medoid, δηλαδή του πιο «αντιπροσωπευτικού» σημείου της συστάδας και γ) για αρκετούς τύπους δεδομένων το πιο κεντρικό σημείο της συστάδας. Το σχήμα τους είναι σφαιρικό. Στην εικόνα 2.2 παρουσιάζονται τέσσερις ομάδες σημείων που βασίζονται στο «κέντρο» ή πρότυπο.



Εικόνα 2.2 Συστάδες που βασίζονται στο «κέντρο» ή πρότυπο

- **Συστάδες που βασίζονται στην εγγύτητα (Contiguous clusters: Nearest Neighbor or Transitive Clustering)**

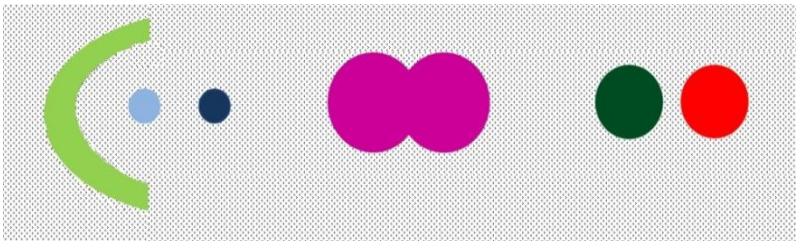
  Σ' αυτήν την κατηγορία συστάδων, κάθε συστάδα αποτελεί ομάδα αντικειμένων τέτοια ώστε εντός της συστάδας υπάρχει σύνδεση των αντικειμένων μεταξύ τους αλλά δεν παρουσιάζεται σύνδεση αντικειμένων μεταξύ διαφορετικών συστάδων. Κάθε αντικείμενο είναι πιο κοντά (ή περισσότερο όμοιο) σε ένα ή περισσότερα σημεία της συστάδας από ό, τι σε οποιοδήποτε αντικείμενο εκτός συστάδας Στην εικόνα 2.3 παρουσιάζονται οχτώ ομάδες σημείων που βασίζονται στην εγγύτητα.



Εικόνα 2.3 Συστάδες που βασίζονται στην εγγύτητα

- **Συστάδες που βασίζονται στην πυκνότητα (Density - based clusters)**

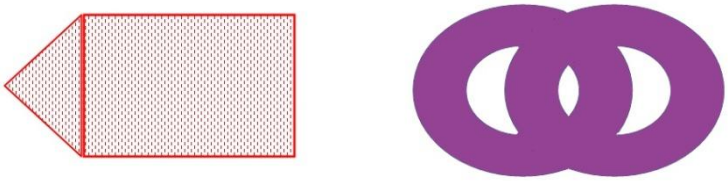
  Στην κατηγορία των συστάδων που βασίζονται στην πυκνότητα, κάθε συστάδα αποτελεί μία πυκνή περιοχή αντικειμένων η οποία περιβάλλεται από χαμηλής πυκνότητας περιοχές. Ο ορισμός των συστάδων με βάση την πυκνότητα χρησιμοποιείται συνήθως σε περιπτώσεις συστάδων με ακανόνιστο και αλληλένδετο σχήμα αλλά και όταν εμφανίζονται ακραίες τιμές και θόρυβος. Στην εικόνα 2.4 παρουσιάζονται έξι ομάδες σημείων που βασίζονται στην πυκνότητα.



Εικόνα 2.4 Συστάδες που βασίζονται στην πυκνότητα

- **Συστάδες που βασίζονται σε κοινές ιδιότητες ή έννοιες (Shared - Property or Conceptual clusters)**

  Σύμφωνα με αυτή την κατηγορία, οι συστάδες αποτελούν σύνολα αντικειμένων με κοινές ιδιότητες. Ο ορισμός των συστάδων με βάση τις κοινές ιδιότητες ή έννοιες εκτός ότι περιλαμβάνει όλους τους παραπάνω ορισμούς (π.χ. τα αντικείμενα σε συστάδες που βασίζονται στο «κέντρο» ή πρότυπο έχουν την ιδιότητα ότι ισαπέχουν από το ίδιο «κέντρο»), εμπεριέχει και νέους τύπους συστάδων. Στην εικόνα 2.5 παρουσιάζονται μία τριγωνική και μία ορθογώνια συστάδα καθώς και δύο αλληλένδετες κυκλικές συστάδες.



Εικόνα 2.5 Συστάδες που βασίζονται σε κοινές ιδιότητες ή έννοιες

2.3 Διαφορές της ανάλυσης κατά συστάδες με άλλες τεχνικές

Οι βασικές διαφορές της τεχνικής της ανάλυσης κατά συστάδες με τις υπόλοιπες τεχνικές της πολυμεταβλητής στατιστικής ανάλυσης που χρησιμοποιούνται για το διαχωρισμό και κατάταξη των παρατηρήσεων αντικειμένων σε ομάδες αφορούν τις παρακάτω παρατηρήσεις:

- Η ανάλυση κατά συστάδες δεν αποτελεί μία μέθοδο ταξινόμησης με την κλασσική έννοια, δηλαδή «αντλεί» τον αριθμό και την «ονομασία» της ομάδας από τα δεδομένα χωρίς να υπάρχει εκ προοιμίου μία γνωστή κατηγοριοποίηση των παρατηρήσεων. Η διάκριση γίνεται ανάμεσα σε προβλήματα εκμάθησης «υπό επίβλεψη» (supervised) που έχουν ως στόχο την ταξινόμηση (classification) και αφορά την εξόρυξη κάποιου είδους σχηματισμού των δεδομένων (pattern) από ένα σετ δεδομένων με γνωστές ετικέτες κατηγοριών και σε «χωρίς επίβλεψη» (unsupervised) που έχουν ως στόχο την ομαδοποίηση/συσταδοποίηση (clustering) και αφορά την εξόρυξη κάποιου είδους σχηματισμού των δεδομένων από ένα σετ δεδομένων με άγνωστες ετικέτες κατηγορίας. Η ομαδοποίηση/συσταδοποίηση σε συστάδες αποτελεί μια πιο περίπλοκη και δύσκολη διαδικασία σε σύγκριση με την κλασσική έννοια της ταξινόμησης των δεδομένων. Αυτό αποτελεί και το κύριο χαρακτηριστικό το οποίο κατατάσσει την τεχνική της ανάλυσης κατά συστάδες στις διερευνητικές μεθόδους ταξινόμησης.

- Μία από τις βασικότερες διαφορές είναι η έμφαση στην έννοια της ομοιογένειας καθώς οι ομάδες στις οποίες επιδιώκει να καταλήξει η ανάλυση κατά συστάδες θα πρέπει να χαρακτηρίζονται από αντικείμενα με παρόμοια χαρακτηριστικά ή γνωρίσματα.

- Κατά τη χρήση της τεχνικής της ανάλυσης κατά συστάδες δεν είναι απαραίτητο να γίνουν υποθέσεις για την κατανομή που ακολουθούν τα δεδομένα σε αντίθεση με την διακριτική ανάλυση, παραγοντική ανάλυση και την ανάλυση παλινδρόμησης.

2.4 Σκοπός- Χρησιμότητα

Όπως φαίνεται και από τον ορισμό που δόθηκε στην προηγούμενη ενότητα ο σκοπός της τεχνικής της ανάλυσης κατά συστάδες είναι η ταξινόμηση των παρατηρήσεων σε συστάδες/ομάδες που αποτελούνται από αντικείμενα με ομοιογενή χαρακτηριστικά και πιο συγκεκριμένα αυτή η τμηματοποίηση θα πρέπει να γίνεται με τέτοιο τρόπο ώστε :

Πρώτον, κάθε ομάδα να παρουσιάζει ομοιογένεια ως προς κάποια χαρακτηριστικά γνωρίσματα έτσι ώστε οι παρατηρήσεις της ομάδας να είναι όμοιες μεταξύ τους και

Δεύτερον, κάθε ομάδα, βάσει των χαρακτηριστικών της γνωρισμάτων, να είναι διαφορετική από τις άλλες.

Βάσει των παραπάνω, για να έχουμε μία επιτυχημένη ομαδοποίηση θα πρέπει να έχουμε καταλήξει σε ομάδες όπου οι παρατηρήσεις εντός αυτών να είναι όσο περισσότερο γίνεται ομοιογενείς, αλλά οι παρατηρήσεις που ανήκουν σε διαφορετικές ομάδες να διαφέρουν μεταξύ τους στο μέγιστο βαθμό. Άρα επιζητούμε μικρές αποκλίσεις εντός των ομάδων (ομοιογένεια μέσα στην ομάδα) και μεγάλες αποκλίσεις μεταξύ τους (ετερογένεια μεταξύ των ομάδων)

Πολλές φορές όμως, η ανάλυση κατά συστάδες, πέρα από την ομαδοποίηση των παρατηρήσεων που παρουσιάζουν ομοιότητες μεταξύ τους, μπορεί να δίνει και άλλα χρήσιμα αποτελέσματα που μπορούν να συνοψιστούν στους παρακάτω σκοπούς-άξονες :

- Η εξόρυξη γνώσης από τα δεδομένα μας ως προς τον βαθμό ομοιότητας, τη διακριτική ικανότητα κάποιων μεταβλητών κ.ο.κ.

- Τη μελέτη ύπαρξης σχέσεων ανάμεσα στα δεδομένα, το είδος και την ένταση των σχέσεων αυτών, καθώς τα μη επεξεργασμένα σύνολα δεδομένων πολλές φορές παρουσιάζουν χαοτική μορφή.

- Η αποφυγή επικαλύψεων μεταβλητών που δεν παρουσιάζουν συνάφεια με τους ερευνητικούς σκοπούς με στόχο τη μείωση του όγκου του προβλήματος και την εστίαση στις μεταβλητές που παρουσιάζουν ενδιαφέρον όπως αυτό προκύπτει από την ομαδοποίηση των δεδομένων.

- Ο έλεγχος θεωρητικών υποθέσεων ιδιαίτερα όταν υπάρχει υποψία για ύπαρξη ομάδων βάσει θεωρητικών μοντέλων.

- Η πρόβλεψη και κατάταξη νέων παρατηρήσεων βάσει της ομαδοποίησης των παρατηρήσεων σε ομοιογενείς ομάδες που έχει ήδη προηγηθεί. Ένα τέτοιο παράδειγμα θα μπορούσε να είναι η ένταξη των νέων πελατών μίας τράπεζας σε βαθμίδες χαμηλής, μέτριας και υψηλής πιστοληπτικής ικανότητας βάσει των χαρακτηριστικών γνωρισμάτων τους.

2.5 Εφαρμογές

Η τεχνική της ανάλυσης κατά συστάδες βρίσκει σημαντικές εφαρμογές σε πολλαπλά επιστημονικά πεδία όπως η οικονομική επιστήμη (έρευνα μάρκετινγκ-έρευνα αγοράς, επιχειρησιακή έρευνα, οικονομική των επιχειρήσεων), οι επιστήμες υγείας (γενετική, βιολογία, μικροβιολογία), οι φυσικές επιστήμες (χημεία, γεωλογία, αστρονομία), οι κοινωνικές-ανθρωπιστικές επιστήμες (ψυχολογία, εκπαίδευση), η μηχανική, η επιστήμη των υπολογιστών κ.λπ.

Μερικά παραδείγματα εφαρμογών στα άνω επιστημονικά πεδία είναι τα εξής:

- Οικονομική επιστήμη: Εφαρμογές που αφορούν την ταξινόμηση πελατών βάσει των χαρακτηριστικών τους, το αγοραστικό τους προφίλ και συνήθειες, την ταξινόμηση των επιχειρήσεων βάσει της οικονομικής τους θέσης, την ανάλυση της χρονικής τάσης των περιουσιακών τίτλων κ.λπ.

- Επιστήμες υγείας: Εφαρμογές που αφορούν την ταξινόμηση των ειδών του φυτικού και ζωικού βασιλείου σύμφωνα με τα κοινά τους γνωρίσματα, ταξινόμηση των γονιδίων, διάγνωση και θεραπεία ασθενειών με ταξινόμηση κοινών συμπτωμάτων κ.λπ.

- Φυσικές επιστήμες: Εφαρμογές που αφορούν την ταξινόμηση στοιχείων περιοδικού πίνακα, ταξινόμηση γεωλογικών σχηματισμών ομαδοποίηση αστερισμών και πλανητών κ.λπ.

- Κοινωνικές-Ανθρωπιστικές επιστήμες: Εφαρμογές που αφορούν την ανάλυση προτύπων συμπεριφοράς, εγκληματική ψυχολογία, κατάταξη σχολείων βάσει των χαρακτηριστικών τους κ.λπ.

- Μηχανική: Εφαρμογές που αφορούν τη βιομετρική και φωνητική αναγνώριση, ανάλυση ανίχνευσης σήματος κ.λπ.

- Επιστήμη υπολογιστών: Εφαρμογές που αφορούν την ανάκτηση δεδομένων, την εξόρυξη γνώσης από βάσεις δεδομένων, την τμηματοποίηση εικόνας κ.λπ.

3. Μέτρα εγγύτητας-Κριτήρια Συσταδοποίησης

3.1 Μέτρα εγγύτητας

Σε ένα σετ δεδομένων που αποτελείται από ένα δείγμα n ατόμων ή αντικειμένων παρατηρούνται p χαρακτηριστικά, οι τυχαίες μεταβλητές. Ο πίνακας $X = [x_{ij}]$ που συγκεντρώνει τις *nxp* παρατηρήσεις, καλείται πίνακας δεδομένων ή πίνακας πρωτογενών δεδομένων (raw data table):

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & & x_{nj} & & x_{np} \end{bmatrix},$$

όπου x_{ij} η τιμή της μεταβλητής j (χαρακτηριστικό j) που παρατηρείται στο i άτομο ή αντικείμενο.

Η ανάλυση κατά συστάδες ταξινομεί αυτές τις παρατηρήσεις σε ομάδες που αποτελούνται από αντικείμενα που παρουσιάζουν ομοιογένεια ως προς τα χαρακτηριστικά τους, ενώ οι παρατηρήσεις μεταξύ των ομάδων διαφέρουν όσο γίνεται περισσότερο. Συνεπώς, το μέτρο εγγύτητας (proximity measure) μεταξύ ενός ζεύγους αντικειμένων, ενός αντικειμένου και μίας ομάδας ή ακόμα ενός ζεύγους ομάδων, αποτελεί θεμελιώδης έννοια για την ανάλυση κατά συστάδες, καθώς η ομαδοποίηση των παρατηρήσεων γίνεται με βάση την προσεγγισιμότητά τους.

Το μέτρο εγγύτητας αποτελείται από το μέτρο απόστασης (distance measure) ή μέτρο ανομοιότητας (dissimilarity measure) και το μέτρο ομοιότητας (similarity measure). Παρατηρώντας τις έννοιες της απόστασης και της ομοιότητας διαπιστώνουμε πως οι δύο έννοιες είναι αντίθετες. Αν ένα μέτρο εγγύτητας αναπαριστά την απόσταση, η τιμή του μέτρου μειώνεται όταν τα αντικείμενα είναι περισσότερο όμοια μεταξύ τους ενώ αν το μέτρο εγγύτητας αναπαριστά την ομοιότητα, η τιμή του μέτρου αυξάνεται. Επομένως, παρατηρήσεις που έχουν μικρότερη απόσταση ή μεγαλύτερη ομοιότητα κατατάσσονται στην ίδια ομάδα. Τα ζεύγη ανομοιοτήτων (dissimilarities) – ομοιοτήτων (similarities) μεταξύ των αντικειμένων των δεδομένων απεικονίζονται από ένα τετραγωνικό, συμμετρικό πίνακα που καλείται πίνακας αποστάσεων (proximity matrix). Για ένα σετ δεδομένων που αποτελείται

από n άτομα ή αντικείμενα το (i,j) στοιχείο του nxn πίνακα αποστάσεων συμβολίζει το μέτρο απόστασης ή το μέτρο ομοιότητας για τα i και j άτομα ή αντικείμενα (i,j=1,…,n).

Ο προσδιορισμός του κατάλληλου μέτρου εγγύτητας μεταξύ των δεδομένων αποτελεί ουσιαστικό βήμα στην εφαρμογή της μεθόδου της ανάλυσης κατά συστάδες, διότι διαφορετικά μέτρα απόστασης ή ομοιότητας μπορούν να καταλήξουν σε διαφορετικές λύσεις άρα και σε διαφορετική ερμηνεία των αποτελεσμάτων. Αν και θα ήταν εξαιρετικά χρήσιμο να γνωρίζουμε εκ των προτέρων ποιο μέτρο εγγύτητας οδηγεί στη βέλτιστη λύση, η επιλογή του κατάλληλου μέτρου εγγύτητας, βασίζεται στο είδος των δεδομένων, στην κλίμακα μέτρησης, στον τύπο των μεταβλητών αλλά και στη διαίσθηση και εμπειρία του ερευνητή.

3.1.1 Μέτρα απόστασης ή ανομοιότητας

Η έννοια της απόστασης ή ανομοιότητας χρησιμοποιείται για να προσδιορίσει πόσο απέχουν οι παρατηρήσεις μεταξύ τους, δηλαδή αν μοιάζουν ή όχι ώστε να κατηγοριοποιηθούν στην κατάλληλη ομάδα. Δοθέντων δύο αντικειμένων x_i και x_j σ' ένα p-διάστατο χώρο, το μέτρο απόστασης ορίζεται ως μία συνάρτηση όπου ικανοποιούνται οι ακόλουθες ιδιότητες (Xu & Wunsch, 2009):

1. $D(x_i, x_j) \geq 0$ για όλα τα x_i και x_j (3.1)
2. $D(x_i, x_j) = D(x_j, x_i)$ (3.2)
3. $D(x_i, x_j) \leq D(x_i, x_k) + D(x_j, x_k)$ για όλα τα x_i, x_j και x_k (3.3)
   (τριγωνική ανισότητα)
4. $D(x_i, x_j) = 0$ αν και μόνο αν $x_i = x_j$ (3.4)

Η πρώτη ιδιότητα αναφέρει πως το μέτρο απόστασης δεν είναι ποτέ αρνητικό, η δεύτερη πως το μέτρο είναι συμμετρικό ενώ η τέταρτη ιδιότητα απαιτεί το μέτρο απόστασης να είναι μηδέν όταν τα αντικείμενα είναι πανομοιότυπα και σε καμία άλλη περίπτωση. Όταν ικανοποιούνται όλες οι ιδιότητες τότε η απόσταση καλείται μετρική (metric). Αν δεν ικανοποιείται η τρίτη ιδιότητα της τριγωνικής ανισότητας, τότε το μέτρο απόστασης καλείται ημιμετρικό (semimetric measure).

Για **<u>συνεχείς μεταβλητές (continuous variables)</u>**, τα μέτρα απόστασης που χρησιμοποιούνται είναι τα εξής:

- **Ευκλείδεια απόσταση (Euclidean distance)**

  Η Ευκλείδεια απόσταση ή L_2 νόρμα (L_2 norm) ορίζεται ως :

$$D(x_i, x_j) = \left(\sum_{l=1}^{p} |x_{il} - x_{jl}|^{1/2} \right)^2 , \tag{3.5}$$

  όπου x_i και x_j είναι n-διάστατα αντικείμενα δεδομένων. Η Ευκλείδεια απόσταση είναι το πιο απλό και σύνηθες μέτρο απόστασης που χρησιμοποιείται κι επειδή ικανοποιεί όλες τις ιδιότητες της απόστασης καλείται μετρική απόσταση. Παρουσιάζει όμως και κάποια μειονεκτήματα. Εξαρτάται από τις μονάδες μέτρησης των μεταβλητών και από την κλίμακα μέτρησής τους με αποτέλεσμα αλλάζοντας την κλίμακα μέτρησης να έχουμε διαφορετικές αποστάσεις και συνεπώς διαφορετικές ομάδες παρατηρήσεων. Η τακτική αντιμετώπισης του συγκεκριμένου προβλήματος είναι η τυποποίηση των δεδομένων, όμως χρησιμοποιώντας τις αποστάσεις των τυποποιημένων δεδομένων δε λαμβάνεται υπόψη η σημαντικότητα κάποιων μεταβλητών.

- **Απόσταση Manhattan (Manhattan distance)**

  Η απόσταση Manhattan ή απόσταση city-block (city-block distance), ή L_1 νόρμα (L_1 norm) βασίζεται επίσης στις απόλυτες τιμές των διαφορών των παρατηρήσεων και ορίζεται ως :

$$D(x_i, x_j) = \sum_{l=1}^{p} |x_{il} - x_{jl}| . \tag{3.6}$$

Εξαιτίας της ομοιότητάς της με την Ευκλείδεια απόσταση, παρουσιάζει παρόμοια αποτελέσματα αλλά στην περίπτωση που υπάρχουν ακραίες παρατηρήσεις (outliers) δίνοντάς τους μικρότερο βάρος λόγω της απόλυτης τιμής οδηγούμαστε σε πιο ανθεκτικά αποτελέσματα.

- **Απόσταση Minkowski (Minkowski distance)**

Η γενικότερη μορφή της Ευκλείδειας απόστασης και της απόστασης Manhattan είναι η απόσταση Minkowski ή L_p νόρμα (L_p norm) και ορίζεται ως:

$$D(x_i, \mathrm{x}_j) = \left(\sum_{\mathrm{l}=1}^{\mathrm{p}} \left| x_{il} - \mathrm{x}_{jl} \right|^q \right)^{1/\mathrm{q}} . \tag{3.7}$$

Προφανώς αν q = 1 προκύπτει η απόσταση Manhattan ενώ αν q = 2 προκύπτει η Ευκλείδεια απόσταση.

- **Απόσταση Power (Power distance)**

Η απόσταση Power αποτελεί γενίκευση της απόστασης Minkowski και ορίζεται ως:

$$D(x_i, \mathrm{x}_j) = \left[\sum_{\mathrm{l}=1}^{\mathrm{p}} \left(x_{il} - \mathrm{x}_{jl} \right)^{\mathrm{q}} \right]^{1/\mathrm{r}} , \tag{3.8}$$

όπου q και r παράμετροι που ορίζει ο ερευνητής.

- **Απόσταση Chebychev (Chebychev distance)**

Σε αντίθεση με τις παραπάνω αποστάσεις, η απόσταση Chebychev χρησιμοποιεί τη μέγιστη απόλυτη διαφορά των τιμών των παρατηρήσεων ως προς το σύνολο των μεταβλητών και ορίζεται ως :

$$D(x_i, \mathrm{x}_j) = \max\left\{ \left| x_{il} - \mathrm{x}_{jl} \right|, l = 1, \dots, \mathrm{p} \right\} . \tag{3.9}$$

Η απόσταση Chebychev είναι κατάλληλη όταν εξετάζουμε τη διαφορά δύο παρατηρήσεων ως προς μία μεταβλητή, όμως λόγω της εξάρτησής της από τις διαφορές στην κλίμακα των μεταβλητών στην περίπτωση που οι κλίμακες είναι διαφορετικές θα παρουσιάζει τη διαφορά στη μεταβλητή με τη μεγαλύτερη κλίμακα.

- **<u>Απόσταση Mahalanobis (Mahalanobis distance)</u>**

  Σε αντίθεση με όλα τα παραπάνω μέτρα απόστασης, η απόσταση Mahalanobis βασίζεται σε στατιστικές έννοιες, λαμβάνει υπόψη διακυμάνσεις και συνδιακυμάνσεις και ορίζεται ως :

$$\mathrm{D}^2\left(x_i, \mathrm{x}_j\right) = \left(x_i - \mathrm{x}_j\right)' \mathrm{S}^{-1}\left(x_i - \mathrm{x}_j\right), \tag{3.10}$$

  όπου S ο δειγματικός πίνακας συνδιακύμανσης ή συνδιασποράς και ορίζεται ως: $S = E[(\chi - \mu)(\chi - \mu)']$ (3.11), όπου μ είναι η μέση τιμή και $E[.]$ η εκτιμώμενη τιμή τυχαίας μεταβλητής.

- **Μέτρα συσχέτισης**

  Το μέτρο απόστασης μπορεί να προέρχεται από το συντελεστή συσχέτισης (correlation coefficient), όπως τον συντελεστή συσχέτισης Pearson (Pearson correlation coefficient) που ορίζεται ως :

$$r_{\mathrm{ij}} = \frac{\sum_{\mathrm{l}=1}^{\mathrm{p}} (\mathrm{x}_{\mathrm{il}} - \overline{\mathrm{x}_{\mathrm{i}}})\left(\mathrm{x}_{\mathrm{jl}} - \overline{\mathrm{x}_{\mathrm{j}}}\right)}{\sqrt{\sum_{\mathrm{l}=1}^{\mathrm{p}} (\mathrm{x}_{\mathrm{il}} - \overline{\mathrm{x}_{\mathrm{i}}})^2 \sum_{\mathrm{l}=1}^{\mathrm{p}} \left(\mathrm{x}_{\mathrm{jl}} - \overline{\mathrm{x}_{\mathrm{j}}}\right)^2}}, \tag{3.12}$$

  όπου $\overline{\mathrm{x}_{\mathrm{i}}} = \frac{1}{\mathrm{p}} \sum_{\mathrm{l}=1}^{\mathrm{p}} \mathrm{x}_{il}$ (3.13). Τότε το μέτρο απόστασης ορίζεται ως :

  $\mathrm{D}\left(x_i, \mathrm{x}_j\right) = \left(1 - r_{\mathrm{ij}}\right)/2$ (3.14). Χρησιμοποιώντας τα μέτρα συσχέτισης δεν αθροίζουμε ως προς όλες τις παρατηρήσεις αλλά ως προς όλες τις μεταβλητές.

3.1.2 Μέτρα ομοιότητας

Δοθέντων δύο αντικειμένων x_i και x_j σ' ένα p-διάστατο χώρο το μέτρο ομοιότητας ορίζεται ως η συνάρτηση ομοιότητας που ικανοποιεί τις ακόλουθες ιδιότητες (Xu & Wunsch, 2009):

1. $0 \leq S(x_i, x_j) \leq 1$ για όλα τα x_i και x_j (3.15)
2. $S(x_i, x_j) = S(x_j, x_i)$ (3.16)
3. Για όλα τα x_i, x_j, x_k
   $S(x_i, x_j) S(x_j, x_k) \leq [S(x_i, x_j) + S(x_j, x_k)]\, S(x_i, x_k)$ (3.17)
4. $S(x_i, x_j) = 1$ αν και μόνο αν $x_i = x_j$ (3.18)

Όταν ικανοποιούνται όλες οι ιδιότητες τότε η ομοιότητα καλείται μετρική (similarity metric). Δεδομένου ότι το μέτρο ομοιότητας ικανοποιεί τις παραπάνω ιδιότητες μπορούμε να ορίσουμε το μέτρο ανομοιότητας σύμφωνα με τη σχέση $D(x_i, x_j) = 1 - S(x_i, x_j)$ (3.19), ωστόσο παρατηρείται πως αυτό το νέο μέτρο απόστασης δεν αποτελεί μετρική απόσταση. Ένας άλλος ορισμός του μέτρου ανομοιότητας, σύμφωνα με τον Gower (1966), είναι η σχέση $D(x_i, x_j) = \sqrt{1 - S(x_i, x_j)}$ (3.20). Αντίστροφα, δοθέντος ενός μέτρου ανομοιότητας μπορούμε να κατασκευάσουμε ένα μέτρο ομοιότητας σύμφωνα με τη σχέση $S(x_i, x_j) = 1/[1 + D(x_i, x_j)]$ (3.21). Προφανώς, το νέο μέτρο ομοιότητας δεν αποκτά ποτέ την τιμή μηδέν καθώς το μέτρο ανομοιότητας είναι απεριόριστο.

Για **<u>συνεχείς μεταβλητές</u>**, το μέτρο ομοιότητας που χρησιμοποιείται είναι το **συνημίτονο ομοιότητας (cosine similarity)** και ορίζεται ως :

$$S(x_i, x_j) = \cos a = \frac{x_i^T x_j}{\|x_i\| \|x_j\|}, \quad (3.22)$$

σύμφωνα με το οποίο όταν τα αντικείμενα παρουσιάζουν μεγαλύτερη ομοιογένεια ως προς τα χαρακτηριστικά τους τότε αυξάνεται και η τιμή του συνημίτονου. Το μέτρο απόστασης ορίζεται από τη σχέση (3.19)

Για **δίτιμες μεταβλητές (binary variables)**, υπάρχει μια ποικιλία μέτρων ομοιότητας που μπορεί ο ερευνητής να χρησιμοποιήσει όπου τα χαρακτηριστικά των αντικειμένων των παρατηρήσεων λαμβάνουν δύο τιμές : την τιμή 1 στην παρουσία του χαρακτηριστικού του αντικειμένου και την τιμή 0 στην απουσία του. Πιο συγκεκριμένα, για να υπολογίσουμε ένα μέτρο ομοιότητας ή ένα μέτρο ανομοιότητας (σύμφωνα με τη σχέση 3.19) ανάμεσα στις παρατηρήσεις x_i και x_j χρησιμοποιούμε τον παρακάτω **πίνακα συνάφειας (contingency table)** (Πίνακας 3.1).

		Παρατήρηση x_j		
		1	**0**	**Άθροισμα**
Παρατήρηση x_i	**1**	a	b	a+b
	0	c	d	c+d
	Άθροισμα	a+c	b+d	p=a+b+c+d

Πίνακας 3.1 Πίνακας συνάφειας

Στον Πίνακα 3.1 τα a,b,c,d αναπαριστούν το πλήθος των συνδυασμών (1,1),(1,0),(0,1),(0,0) αντίστοιχα. Το κελί (1,1) υποδηλώνει την παρουσία των χαρακτηριστικών και στις δύο παρατηρήσεις, το κελί (1,0) τα χαρακτηριστικά που είναι παρόντα στην παρατήρηση x_i και απόντα στην παρατήρηση x_j, το κελί (0,1) τα χαρακτηριστικά που είναι απόντα στην παρατήρηση x_i και παρόντα στην παρατήρηση x_j ενώ το κελί (0,0) υποδηλώνει την απουσία των χαρακτηριστικών και στις δύο παρατηρήσεις.

Οι μεταβλητές χωρίζονται σε δύο κατηγορίες: τις συμμετρικές μεταβλητές και τις ασύμμετρες μεταβλητές. Στην περίπτωση των συμμετρικών μεταβλητών και οι δύο τιμές των χαρακτηριστικών είναι εξίσου σημαντικές ενώ στις ασύμμετρες μεταβλητές μια εκ των δύο τιμών των χαρακτηριστικών δεν είναι χρήσιμη για την εξαγωγή συμπερασμάτων. Βάσει της κατηγοριοποίησης των μεταβλητών σε συμμετρικές και ασύμμετρες διαχωρίζονται και τα μέτρα ομοιότητας σε δύο κατηγορίες αντίστοιχα όπως απεικονίζονται στους παρακάτω

πίνακες (Πίνακας 3.2 και Πίνακας 3.3), όπου χρησιμοποιώντας τη συμπληρωματική σχέση (3.19) απεικονίζονται και τα μέτρα απόστασης.

Ονομασία συντελεστή	$\mathbf{S(x_i, x_j)}$	$\mathbf{D(x_i, x_j)}$
Simple Matching Coefficient	$\frac{a+d}{a+b+c+d}$	$\frac{b+c}{a+b+c+d}$
Double Matching Coefficient	$\frac{2(a+d)}{2(a+d)+(b+c)}$	$\frac{2(b+c)}{2(a+d)+(b+c)}$
Rogers-Tarimoto	$\frac{a+d}{(a+d)+2(b+c)}$	$\frac{b+c}{(a+d)+2(b+c)}$

Πίνακας 3.2 Μέτρα ομοιότητας και ανομοιότητας χρήσιμα για συμμετρικές δίτιμες μεταβλητές

Στην κατηγορία των συμμετρικών μεταβλητών τα ζεύγη των χαρακτηριστικών των αντικειμένων που συμφωνούν θεωρούνται εξίσου σημαντικά (κελιά (0,0) και (1,1)) ενώ τα βάρη που δίνονται στα ζεύγη που δε συμφωνούν βασίζονται ανάλογα με τη συνεισφορά τους στην ομοιότητα. Πιο αναλυτικά, παρατηρούμε στον Πίνακα 3.2 πως το πρώτο μέτρο, ο συντελεστής ομοιότητας(Simple Matching Coefficient), μετράει τον αριθμό των μεταβλητών για τις οποίες οι δύο παρατηρήσεις συμφωνούν (κελιά συμφωνίας (0,0) και (1,1)) και τα δύο επόμενα μέτρα Double Matching Coefficient και Rogers-Tarimoto χρησιμοποιούν την ίδια πληροφορία αλλά δίνουν διαφορετικό βάρος στα κελιά συμφωνίας από ότι στα κελιά ασυμφωνίας.

Στην περίπτωση που η παράβλεψη του συντελεστή της κοινής απουσίας των χαρακτηριστικών των παρατηρήσεων (κελί (0,0)) καθόλου δεν επηρεάζει το τελικό αποτέλεσμα τότε τα μέτρα ομοιότητας για ασύμμετρες δυαδικές μεταβλητές, όπως αυτά του Πίνακα 3.3, είναι ιδιαίτερα χρήσιμα.

Ονομασία συντελεστή	$\mathbf{S(x_i, x_j)}$	$\mathbf{D(x_i, x_j)}$
Jaccard Coefficient (Jacard 1908)	$\frac{a+d}{a+b+c+d}$	$\frac{b+c}{a+b+c+d}$
Chekanowski(1932),Dice(1945), Sørensen (1948)	$\frac{2(a+d)}{2(a+d)+(b+c)}$	$\frac{2(b+c)}{2(a+d)+(b+c)}$
Socal και Sneath (1963)	$\frac{a+d}{(a+d)+2(b+c)}$	$\frac{b+c}{(a+d)+2(b+c)}$

Πίνακας 3.3 Μέτρα ομοιότητας και ανομοιότητας χρήσιμα για ασύμμετρες δίτιμες μεταβλητές

Το κύριο βάρος των συντελεστών εστιάζεται στην κοινή παρουσία κάποιων χαρακτηριστικών (κελί (1,1)), δε λαμβάνουν υπόψη τη συχνότητα του κελιού της κοινής απουσίας χαρακτηριστικών και δίνουν βάρος στα ζεύγη χαρακτηριστικών που δε συμφωνούν ανάλογα με τη σπουδαιότητά τους.

Οι συντελεστές ομοιότητας που απεικονίζονται στους Πίνακες 3.2 και 3.3 παίρνουν τιμές στο διάστημα (0,1) και για το ίδιο σετ δεδομένων παρουσιάζουν διαφορετικές τιμές.

Για **κατηγορικές μεταβλητές που αναφέρονται σε ονομαστική κλίμακα**, η δυσκολία μέτρησης της ομοιότητας ή απόστασης αντιμετωπίζεται είτε με τη μετατροπή τους σε δίτιμες μεταβλητές (τρόπος που παρουσιάζει αρκετά μειονεκτήματα) ή ακολουθώντας μία καλύτερη μέθοδο χρησιμοποιώντας τον συντελεστή ομοιότητας (Simple Matching Coefficient) και ανομοιότητας αντίστοιχα :

$$S(x_i, x_j) = \frac{u}{p} \text{ και } D(x_i, x_j) = \frac{p-u}{p} \ , \qquad (3.23)$$

όπου u είναι ο αριθμός των μεταβλητών που έχουν την ίδια τιμή και p συνολικός αριθμός των μεταβλητών.

Στην περίπτωση **κατηγορικών μεταβλητών σε κλίμακα κατάταξης**, ο ερευνητής είτε θεωρεί τις μεταβλητές ως συνεχείς και χρησιμοποιεί μία κατάλληλη απόσταση ή μετασχηματίζει την κλίμακα για να παίρνει τιμές στο διάστημα (0,1).

Στην ανάλυση κατά συστάδες τα περισσότερα προβλήματα που αντιμετωπίζει ο ερευνητής δεν αφορούν σετ δεδομένων που περιέχουν μεταβλητές που ανήκουν αποκλειστικά στην ίδια κατηγορία αλλά αποτελούνται από συνεχείς, αλλά και κατηγορικές (δίτιμες συμμετρικές και ασύμμετρες μεταβλητές, σε ονομαστική κλίμακα ή κλίμακα κατάταξης). Δηλαδή στην πράξη αφορούν σετ δεδομένων που περιέχουν **μεταβλητές μεικτού τύπου** (mixed mode variables). Στην περίπτωση μεταβλητών μεικτού τύπου ουσιαστική σημασία έχει η επιλογή ενός κατάλληλου μέτρου ομοιότητας που συνυπολογίζει κάθε τύπο μεταβλητής. Ο συντελεστής αυτός ορίστηκε από τον Gower (1971) και είναι της μορφής:

$$S(\mathrm{x}_i, \mathrm{x}_j) = \frac{\sum_l^p \mathrm{w}_l (\mathrm{x}_i, \mathrm{x}_j) \mathrm{S}_l (\mathrm{x}_i, \mathrm{x}_j)}{\sum_l^p \mathrm{w}_l (\mathrm{x}_i, \mathrm{x}_j)},$$

(3.24)

όπου $\mathrm{S}_l(\mathrm{x}_i, \mathrm{x}_j)$ η ομοιότητα ανάμεσα στις παρατηρήσεις x_i και x_j για την μεταβλητή l και $\mathrm{w}_l(\mathrm{x}_i, \mathrm{x}_j)$ το βάρος της μεταβλητής l και παίρνει την τιμή 0 ή 1 ανάλογα με το αν σύγκριση ανάμεσα στις παρατηρήσεις θεωρείται πως είναι απαραίτητη για τη μεταβλητή l. Όταν η μεταβλητή είναι κατηγορική (δίτιμη, σε ονομαστική κλίμακα ή κλίμακα κατάταξης) και ταυτίζονται οι τιμές των παρατηρήσεων($\mathrm{x}_{il} = \mathrm{x}_{jl}$),τότε $\mathrm{S}_l(\mathrm{x}_i, \mathrm{x}_j) = 1$, διαφορετικά $\mathrm{S}_l(\mathrm{x}_i, \mathrm{x}_j) = 0$. Αν όμως η μεταβλητή είναι συνεχής τότε $\mathrm{S}_l(\mathrm{x}_i, \mathrm{x}_j) = 1 - \frac{|\mathrm{x}_{il} - \mathrm{x}_{jl}|}{R_l}$(3.25), όπου R_l το εύρος της l –οστής μεταβλητής που ορίζεται ως: $R_l = {}^{max}_{\mathrm{m}}\, \mathrm{x}_{ml} - {}^{min}_{\mathrm{m}}\, \mathrm{x}_{ml}$ (3.26)

3.2 Κριτήρια συσταδοποίησης

Μέσω των μεθόδων ομαδοποίησης, η ανάλυση κατά συστάδες στοχεύει στο διαχωρισμό των αντικειμένων των δεδομένων και την αντιστοίχησή τους σε ομάδες που χαρακτηρίζονται για την ομοιογένειά τους εντός των ομάδων καθώς και τη διαφορετικότητα των χαρακτηριστικών τους μεταξύ των ομάδων. Για συγκεκριμένες μεθόδους ομαδοποίησης (και πιο συγκεκριμένα για τις μη ιεραρχικές μεθόδους ομαδοποίησης στις οποίες θα αναφερθούμε

στο επόμενο κεφάλαιο), η ομοιογένεια και η διαφορετικότητα, αξιολογούνται από συναρτήσεις κριτηρίων. Πιο συγκεκριμένα, δοθέντος ενός συνόλου δεδομένων $x=(x_1, x_2, \dots x_N,)$, με N το συνολικό αριθμό των παραμέτρων, οι μέθοδοι ομαδοποίησης έχουν ως στόχο την αντιστοίχησή του σε K συστάδες ($C_1,\dots C_K$) μεγιστοποιώντας ή ελαχιστοποιώντας μία συνάρτηση κριτηρίου ομαδοποίησης J. Τα κριτήρια ομαδοποίησης που χρησιμοποιούνται είναι τα εξής (Xu & Wunsch, 2009):

- **Το κριτήριο του αθροίσματος τετραγώνων των σφαλμάτων (Sum of Squares Error)**

  Το κριτήριο του αθροίσματος τετραγώνων των σφαλμάτων ορίζεται ως :

$$\mathbf{J}_s(\mathbf{\Gamma}, \mathbf{M}) = \sum_{i=1}^{K}\sum_{j=1}^{N} \boldsymbol{\gamma}_{ij} \left\|\mathbf{x}_j - \mathbf{m}_i\right\|^2 = \sum_{i=1}^{K}\sum_{j=1}^{N} \boldsymbol{\gamma}_{ij} \left(\mathbf{x}_j - \mathbf{m}_i\right)^T \left(\mathbf{x}_j - \mathbf{m}_i\right), \tag{3.27}$$

  όπου $\mathbf{\Gamma}=\{\boldsymbol{\gamma}_{ij}\}$ είναι ένας πίνακας διαμέρισης, $\boldsymbol{\gamma}_{ij}=\begin{cases}1 & \text{αν } \mathbf{x}_j \in \textit{στην } i \textit{ συστάδα} \\ 0 & \textit{διαφορετικά}\end{cases}$ με $\sum_{\mathrm{i}=1}^{K} \boldsymbol{\gamma}_{ij} = 1 \; \forall j$, $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \dots \mathbf{m}_\mathrm{K}]$ είναι ο πίνακας κεντροειδών (μέσων) και $\mathbf{m}_i = \frac{1}{\boldsymbol{N}_\mathrm{i}} \sum_{\mathrm{j}=1}^{\mathrm{N}} \boldsymbol{\gamma}_{ij} \mathbf{x}_\mathrm{j}$ (3.28) είναι το μέσο δείγμα για τη i-οστή συστάδα με $\mathrm{N_i}$ αντικείμενα. Βέλτιστη θεωρείται η διαμέριση που ελαχιστοποιεί το κριτήριο του αθροίσματος τετραγώνων των σφαλμάτων και καλείται η ελάχιστη διαμέριση της διασποράς (the minimum variance partition). Το κριτήριο του αθροίσματος τετραγώνων των σφαλμάτων είναι κατάλληλο για συμπαγείς συστάδες αλλά συνήθως είναι επιρρεπές στην ύπαρξη ακραίων παρατηρήσεων που πιθανώς να οδηγήσουν στο διαχωρισμό μίας μεγάλης συστάδας σε αρκετά μικρότερες ομάδες. Το κριτήριο του αθροίσματος τετραγώνων των σφαλμάτων επίσης προέρχεται και από τους πίνακες διασποράς $\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$(2.29), όπου $\mathbf{S}_T = \sum_{j=1}^{N}\left(\mathbf{x}_j - \mathbf{m}\right)\left(\mathbf{x}_j - \mathbf{m}\right)^T$(3.30) είναι ο συνολικός πίνακας διασποράς, $\mathbf{S}_W = \sum_{i=1}^{K}\sum_{j=1}^{N} \boldsymbol{\gamma}_{ij} \left(\mathbf{x}_j - \mathbf{m}_i\right)\left(\mathbf{x}_j - \mathbf{m}_i\right)^T$(3.31) είναι ο πίνακας διασποράς εντός της συστάδας, $\mathbf{S}_B = \sum_{i=1}^{K} N_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$ (3.32) είναι ο πίνακας διασποράς μεταξύ των ομάδων και $\mathbf{m} = \frac{1}{N}\sum_{i=1}^{K} N_i \mathbf{m}_i$ (3.33) είναι το συνολικό διάνυσμα των μέσων του σετ δεδομένων. Ο συνολικός πίνακας διασποράς $\mathbf{S}_T$ δεν εξαρτάται από τη διαμέριση, ενώ οι πίνακες διασποράς $\mathbf{S}_W$ και $\mathbf{S}_B$ είναι αλληλοεξαρτώμενοι.

- **Το κριτήριο του ίχνους (trace)**

  Για n-διάστατα δεδομένα, n=2,…, βέλτιστη θεωρείται η διαμέριση που ελαχιστοποιεί το ίχνος του πίνακα διασποράς S_W που ορίζεται σύμφωνα με τον τύπο:

$$\mathbf{tr}(S_W) = \sum_{i=1}^{K}\sum_{j=1}^{N} \boldsymbol{\gamma}_{ij} \left(\mathbf{x}_j - \mathbf{m}_i\right)^T \left(\mathbf{x}_j - \mathbf{m}_i\right) = \mathbf{J}_s(\boldsymbol{\Gamma}, \mathbf{M}),$$

(3.34),

  Η ελαχιστοποίηση του ίχνους του S_W είναι ισοδύναμη με την ελαχιστοποίηση του κριτηρίου του αθροίσματος τετραγώνων των σφαλμάτων το οποίο είναι ισοδύναμο και με τη μεγιστοποίηση του πίνακα διασποράς S_B, καθώς προέρχεται από τον τύπο $\mathbf{tr}(S_T) = \mathbf{tr}(S_W) + \mathbf{tr}(S_B)$ (3.35).

- **Το κριτήριο της ορίζουσας (determinant)**

  Για το κριτήριο που βασίζεται στις ορίζουσες δε λαμβάνεται υπόψη ο πίνακας διασποράς S_B (γιατί όταν ο αριθμός των συστάδων είναι ίσος ή μικρότερος της διάστασης των δεδομένων ο $\mathbf{S}_B$ μετατρέπεται σε μη ιδιάζων πίνακα), δεν εξαρτάται από την κλιμακοποίηση και δε θέτει ως περιορισμό να παρουσιάζουν οι συστάδες σφαιρική δομή. Ορίζεται ως:

$$\mathbf{J}_d(\boldsymbol{\Gamma}, \mathbf{M}) = |\mathbf{S}_W| = \left| \sum_{i=1}^{K}\sum_{j=1}^{N} \boldsymbol{\gamma}_{ij} \left(\mathbf{x}_j - \mathbf{m}_i\right)\left(\mathbf{x}_j - \mathbf{m}_i\right)^T \right|.$$

(3.36)

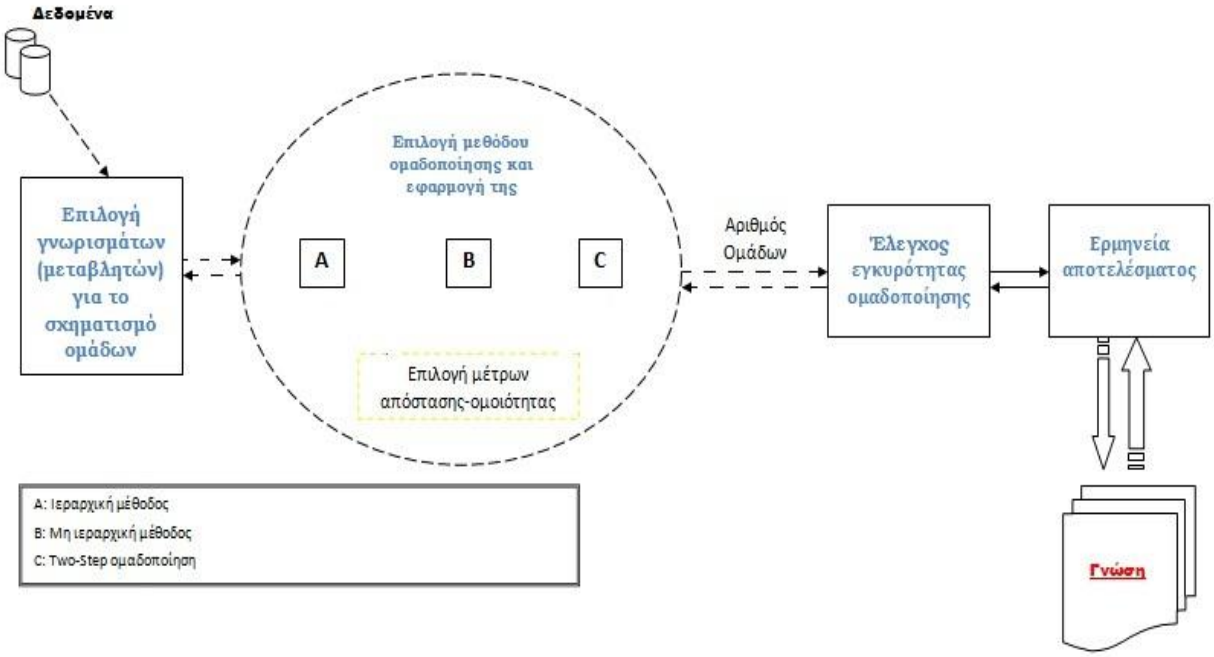
  Βέλτιστη θεωρείται η διαμέριση που ελαχιστοποιεί το κριτήριο της ορίζουσας.

4. Τα στάδια της ανάλυσης κατά συστάδες

Η διαδικασία ομαδοποίησης των παρατηρήσεων όπως αυτή γίνεται με την ανάλυση κατά συστάδες περιλαμβάνει 4 βασικά στάδια:

1. Επιλογή γνωρισμάτων (μεταβλητών) για το σχηματισμό ομάδων
2. Επιλογή μεθόδου ομαδοποίησης και εφαρμογή της
3. Έλεγχος εγκυρότητας της ομαδοποίησης
4. Ερμηνεία του αποτελέσματος

Στην παρακάτω εικόνα (Εικόνα 4.1), φαίνονται σχηματικά τα 4 βασικά στάδια της διαδικασίας της ανάλυσης κατά συστάδες, τα οποία χρησιμοποιούν τη πληροφορία που προκύπτει από το δείγμα των παρατηρήσεων, ενσωματώνουν μηχανισμούς επανατροφοδότησης (δηλαδή πραγματοποιούνται συνεχώς δοκιμές και επαναλήψεις), συνδέονται πολύ στενά μεταξύ τους και τελικά καθορίζουν τις ομάδες/συστάδες των παρατηρήσεων.



Εικόνα 4.1 Τα στάδια της ανάλυσης κατά συστάδες

4.1 Επιλογή γνωρισμάτων (μεταβλητών) για το σχηματισμό ομάδων

Το πρώτο στάδιο αποτελεί ίσως την πιο καθοριστική φάση του καθορισμού των ομάδων των παρατηρήσεων καθώς η επιλογή των σωστών και κατάλληλων γνωρισμάτων που παρουσιάζουν τα δεδομένα όπως αυτά διαφαίνονται και εκπροσωπούνται από τις αντίστοιχες μεταβλητές μπορεί να οδηγήσει με περισσότερη ασφάλεια στην παραγωγή/εξαγωγή από το δείγμα των σωστών ομάδων που το απαρτίζουν. Το τελευταίο σημείο έχει όχι μόνο σημασία για την ορθότερη τμηματοποίηση των παρατηρήσεων αλλά και για την απόφαση λήψης αποτελεσματικών στρατηγικών διοίκησης που βασίζονται στη σωστή τμηματοποίηση. Έτσι, αποφεύγονται ομαδοποιήσεις που στηρίζονται στη διαίσθηση ή απλά στην εμπειρία. Η αρχική επιλογή των μεταβλητών καθορίζει τα χαρακτηριστικά γνωρίσματα για τη δημιουργία των υποομάδων και η τυχόν απόκλιση σημαντικών μεταβλητών, θεωρώντας πως δεν έχουν καμία διακριτική ικανότητα, πιθανόν να οδηγήσει σε μη επαρκή αποτελέσματα που αλλοιώνουν την αξία της ομαδοποίησης.

Ο αριθμός των μεταβλητών που θα χρησιμοποιηθούν εξαρτάται από το μέγεθος του δείγματος, όπως και η ύπαρξη συσχετίσεων μεταξύ των μεταβλητών έχει παρατηρηθεί πως εξαρτάται από το σχήμα των δεδομένων. Η διαφορές στις κλίμακες μέτρησης των μεταβλητών επιφέρουν σημαντικές αλλαγές σε όλη την ανάλυση, καθώς μεταβλητές με αρκετά μεγάλες τιμές συμβάλλουν περισσότερο στο μέτρο εγγύτητας απ’ ότι μεταβλητές με σχετικά μικρές τιμές. Επίσης, η τυποποίηση των μεταβλητών, ώστε να έχουν ίδια μεταβλητότητα, μπορεί να οδηγήσει σε μη επιθυμητά αποτελέσματα εξαιτίας των χρήσιμων πληροφοριών που παρατηρείται πως χάνονται κατά τη διάρκεια της ανάλυσης.

Οι μεταβλητές ομαδοποίησης μπορούν να κατηγοριοποιηθούν σε διάφορους τύπους σύμφωνα με το κριτήριο της σχέσης/εξάρτησης της παρατήρησης (άτομο) με την υπηρεσία και το κριτήριο της παρατηρησιμότητας. Σύμφωνα με το κριτήριο της σχέσης/εξάρτησης της παρατήρησης (άτομο) με την υπηρεσία, οι μεταβλητές διακρίνονται σε γενικές (αυτές που αφορούν πολιτιστικά, δημογραφικά, γεωγραφικά, κοινωνικοοικονομικά χαρακτηριστικά και στοιχεία ψυχογραφικά, κ.λπ.) και ειδικές (αυτές που αφορούν την κατάσταση του πελάτη και στοιχεία που σχετίζονται με τις αντιλήψεις και στάσεις του. Σύμφωνα με το κριτήριο της παρατηρησιμότητας, οι μεταβλητές διακρίνονται σε παρατηρήσιμες, δηλαδή σε άμεσα μετρήσιμες (αυτές που αφορούν πολιτιστικά, δημογραφικά, γεωγραφικά, κοινωνικοοικονομικά χαρακτηριστικά, την κατάσταση του πελάτη κ.ά.) και σε μη

παρατηρήσιμες, δηλαδή σε έμμεσα εξαγόμενες από τις προηγούμενες (αυτές που αφορούν στοιχεία ψυχογραφικά, αντιλήψεις και στάσεις).

Πριν την επιλογή της μεθόδου ομαδοποίησης είναι εξαιρετικά χρήσιμη και η διαγραμματική απεικόνιση και μελέτη των δεδομένων με την χρήση των διαγραμμάτων διασποράς (scatter diagrams) και των profile diagrams, καθώς προσφέρουν κατευθύνσεις (γενικές βέβαια) πριν από τη χρήση των μέτρων των αποστάσεων, για τον αριθμό και το είδος των ομάδων που «κρύβονται» στο δείγμα. Και τα δύο είδη γραφικών απεικονίσεων είναι πολύ εύχρηστα για μικρό αριθμό μεταβλητών και για τον έλεγχο ύπαρξης ακραίων παρατηρήσεων, στις οποίες οι αποστάσεις είναι ευαίσθητες.

4.2 Επιλογή μεθόδου ομαδοποίησης και εφαρμογή της

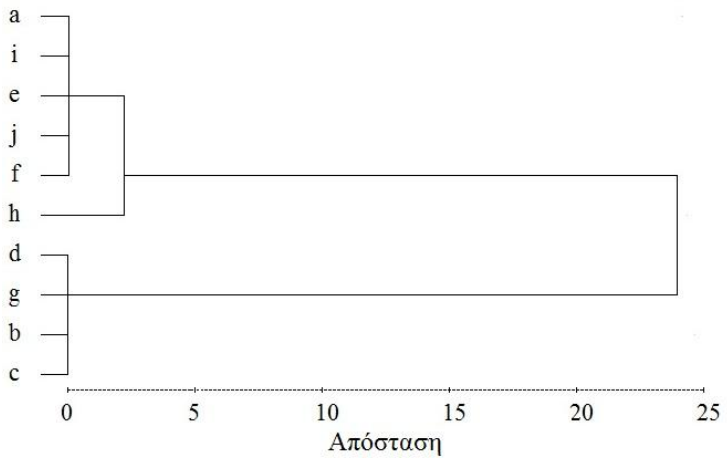
Η επιλογή της μεθόδου ομαδοποίησης αφορά στον καθορισμό του τρόπου βάσει του οποίου θα διαμορφωθούν οι ομάδες με τη χρήση κριτηρίων βελτιστοποίησης όπως η ελαχιστοποίηση της διακύμανσης εντός της ομάδας ή η μεγιστοποίηση της απόστασης μεταξύ των ομάδων. Ο ερευνητής επιλέγει την κατάλληλη μέθοδο ομαδοποίησης (συσταδοποίησης), σύμφωνα με τον τύπο των διαθέσιμων δεδομένων, τον τύπο των παραγόμενων συστάδων, αλλά και τον ιδιαίτερο σκοπό της ανάλυσης κατά συστάδες. Καθώς η ανάλυση κατά συστάδες είναι κυρίως περιγραφικό ή διερευνητικό εργαλείο, είναι επιτρεπτό ο ερευνητής να δοκιμάσει διάφορες μεθόδους με τα ίδια δεδομένα.

Ο συνήθης διαχωρισμός των διαφορετικών μεθόδων ομαδοποίησης, που θα αναλύσουμε και στην παρούσα διπλωματική εργασία, τις κατατάσσει σε ιεραρχικές **(hierarchical) και μη ιεραρχικές (non - hierarchical) ή διαμέρισης (partitioning).**

4.2.1 Ιεραρχικές μέθοδοι ομαδοποίησης

Ο στόχος των ιεραρχικών μεθόδων ομαδοποίησης (Hierarchical clustering) είναι να δημιουργήσουν μια ιεραρχία ομάδων που αποτελούν είτε μικρές ομάδες με ομοιογενή αντικείμενα ή μεγάλες ομάδες που περιλαμβάνουν πιο ανόμοια στοιχεία. Βασίζονται σ’ έναν αλγόριθμο που χρησιμοποιεί έναν πίνακα αποστάσεων και τα αποτελέσματα αναπαρίστανται από ένα **δενδρόγραμμα (dendrogram)**, ένα διάγραμμα δύο διαστάσεων, στο οποίο παρουσιάζονται οι συγχωνεύσεις και οι διαχωρισμοί των ομάδων καθώς και η σειρά με την

οποία πραγματοποιούνται. Έχοντας τη δομή ενός δέντρου, περιέχει μία κλίμακα αποστάσεων και τις παρατηρήσεις οι οποίες απεικονίζονται με τέτοιο τρόπο που διευκολύνεται η διαγραμματική απεικόνιση των βημάτων ομαδοποίησης. Το δενδρόγραμμα ενώνει με γραμμές τις παρατηρήσεις που ομαδοποιούνται παρουσιάζοντας τις τελικές ομάδες αλλά και τις υποομάδες που δημιουργήθηκαν κατά τη διαδικασία της ιεραρχικής μεθόδου ομαδοποίησης (Εικόνα 4.2). Κατασκευασμένο με την κλίμακα απόστασης είτε κάθετα είτε οριζόντια, ανάλογα με τη μέθοδο ιεραρχικής ομαδοποίησης που θα επιλέξει ο ερευνητής, το δενδρόγραμμα παρουσιάζει διαφοροποιήσεις στη μορφή των εξαχθεισών ομάδων.



Εικόνα 4.2 Δενδρόγραμμα

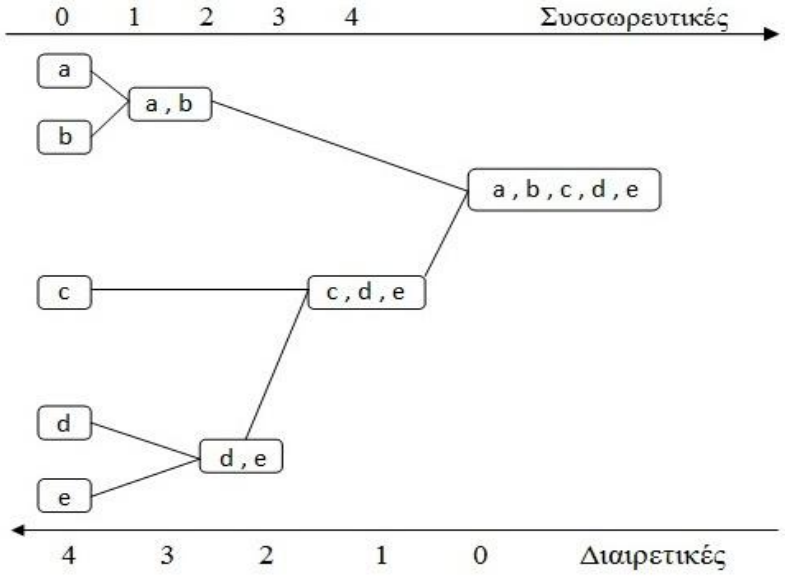
Διαγραμματική παρουσίαση των ομάδων σε κάθε στάδιο των ιεραρχικών μεθόδων ομαδοποίησης δίνει και το **icicle γράφημα (icicle plot)** φανερώνοντας σε κάθε στάδιο ποιες παρατηρήσεις ενώνονται μεταξύ τους, όπως εμφανίζεται και στην παρακάτω εικόνα (Εικόνα 4.3). Το icicle γράφημα διαβάζεται από κάτω προς τα πάνω, με τις σειρές να αντιστοιχούν στον αριθμό των ομάδων ενώ οι στήλες αντιστοιχούν στις παρατηρήσεις που ομαδοποιούνται.

Αριθμός των συστάδων	**c**		**f**		**e**		**d**		**b**		**a**
1	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X
3	X		X	X	X	X	X		X	X	X
4	X		X	X	X		X		X	X	X
5	X		X	X	X		X		X		X

Εικόνα 4.3 Icicle Γράφημα

Κατά την εφαρμογή των ιεραρχικών μεθόδων δεν απαιτείται η προεπιλογή του αριθμού των ομάδων που θα εξαχθούν για την υλοποίησή τους και από την στιγμή που έχει δημιουργηθεί μία συγχώνευση ή διαίρεση ομάδων δεν είναι δυνατή η διάσπασή τους, παρά μόνο ο συνδυασμός τους με άλλες προϋπάρχουσες. Δηλαδή, οι ιεραρχικές μέθοδοι ομαδοποίησης δεν επιτρέπουν στις παρατηρήσεις να διαχωριστούν από τις ομάδες στις οποίες ήδη έχουν ενταχθεί. Επιπλέον, δεν είναι δυνατή η εφαρμογή τους σε πολύ μεγάλα σύνολα δεδομένων .

Οι ιεραρχικές μέθοδοι χωρίζονται σε δύο κατηγορίες: **τις συσσωρευτικές μεθόδους (agglomerative) και τις διαιρετικές μεθόδους (divisive)**.Οι συσσωρευτικές μέθοδοι ξεκινούν με κάθε παρατήρηση να απαρτίζει μία ομάδα και στη συνέχεια με διαδοχικά βήματα ενώνονται οι ομάδες που παρουσιάζουν μικρότερη απόσταση (μεγαλύτερη ομοιότητα) έχοντας ως τελικό στάδιο την ένωση όλων των παρατηρήσεων σε μία ενιαία ομάδα. (Εικόνα 4.4). Η συγχώνευση ενός ζεύγους συστάδων εξαρτάται από τον ορισμό της συνάρτησης της απόστασης με αποτέλεσμα να έχουν δημιουργηθεί κριτήρια συνένωσης τα οποία γενικεύονται στον επαναληπτικό τύπο των Lance και Williams (1967), για τον οποίο υπάρχει εκτενής ανάλυση παρακάτω. Αντιθέτως, ο διαιρετικές μέθοδοι ξεκινούν με όλες τις παρατηρήσεις να απαρτίζουν μία ενιαία ομάδα και σε κάθε βήμα αποκολλώνται οι παρατηρήσεις με τη μεγαλύτερη απόσταση (μικρότερη ομοιότητα) δημιουργώντας με αυτό τον τρόπο μικρότερες υποομάδες ώσπου στο τελικό στάδιο κάθε παρατήρηση αποτελεί μία ομάδα (Εικόνα 4.4).



Εικόνα 4.4 Διάκριση μεταξύ των συσσωρευτικών και διαιρετικών μεθόδων ομαδοποίησης

4.2.1.1 Συσσωρευτικές μέθοδοι ομαδοποίησης

Στις συσσωρευτικές μεθόδους ομαδοποίησης, έχουμε την εξής αλγοριθμική διαδικασία (Xu & Wunsch, 2009):

1. Ξεκίνα με N συστάδες, κάθε μία να περιέχει μία παρατήρηση και υπολόγισε τον πίνακα αποστάσεων για τις N συστάδες.
2. Αναζήτησε στον πίνακα αποστάσεων το ζεύγος παρατηρήσεων C_i και C_j με τη μικρότερη απόσταση $D(C_i , C_j)= \{\min D(C_m , C_l), l \leq m , l \leq N, m \neq l\}$, ένωσε τις C_i και C_j και δημιούργησε μία νέα συστάδα C_{ij} .
3. Ενημέρωσε τον πίνακα αποστάσεων υπολογίζοντας τις αποστάσεις μεταξύ της συστάδας C_{ij} και των υπολοίπων συστάδων.
4. Επανέλαβε τα βήματα 2 και 3 μέχρι να παραμείνει μόνο μία συστάδα.

Για τον σχηματισμό των ομάδων αρχικά θα πρέπει να προσδιοριστεί το κατάλληλο μέτρο απόστασης (ομοιότητας) που θα χρησιμοποιηθεί για την ομαδοποίηση των δεδομένων. Στη συνέχεια, σημαντικό σημείο αποτελεί ο υπολογισμός της απόστασης μεταξύ των ομάδων σύμφωνα με την οποία οι συσσωρευτικές ιεραρχικές μέθοδοι ομαδοποίησης κατηγοριοποιούνται στις εξής μεθόδους:

- **Η μέθοδος της απλής συνένωσης (single linkage)**

  Σε αυτή τη μέθοδο η απόσταση μεταξύ δύο ομάδων C_i και C_j ορίζεται ως η μικρότερη απόσταση (μεγαλύτερη ομοιότητα) από όλα τα ζεύγη παρατηρήσεων που περιέχουν μία παρατήρηση της C_i και μία παρατήρηση της C_j (Εικόνα 4.5). Γι' αυτό το λόγο είναι γνωστή και ως η μέθοδος του κοντινότερου γείτονα (nearest neighbor). Η μέθοδος της απλής συνένωσης συνήθως παράγει μη συμπαγείς, επιμήκεις και διαφορετικού μεγέθους ομάδες, λόγω του φαινομένου της αλυσίδας (chaining effect).Δεν έχει την ικανότητα να ξεχωρίσει τον θόρυβο και να βρει τις πραγματικές ομάδες με αποτέλεσμα ανόμοια αντικείμενα να καταλήγουν στην ίδια συστάδα.

- **Η μέθοδος της πλήρους συνένωσης (complete linkage)**

  Σε αντίθεση με τη μέθοδο της απλής συνένωσης, η μέθοδος της πλήρους συνένωσης υπολογίζει την απόσταση μεταξύ δύο ομάδων C_i και C_j ως τη μεγαλύτερη απόσταση (μικρότερη ομοιότητα) από όλα τα ζεύγη παρατηρήσεων που περιέχουν μία παρατήρηση της C_i και μία παρατήρηση της C_j (Εικόνα 4.5). Γι' αυτό το λόγο είναι γνωστή και ως η μέθοδος του μακρινότερου γείτονα (farthest neighbor). Συνήθως δημιουργεί συμπαγείς ομάδες αλλά παρατηρείται συχνά αδυναμία δημιουργίας πολύ μικρών συμπαγών ομάδων (Καρλής, 2005).

- **Η μέθοδος των μέσων μεταξύ των ομάδων (group average linkage)**

  Η μέθοδος των μέσων μεταξύ των ομάδων είναι γνωστή και ως η μέθοδος μη σταθμισμένων ζευγών των ομάδων με τη χρησιμοποίηση αριθμητικών μέσων (unweighted pair group method average, UPGMA). Σ' αυτή την περίπτωση η απόσταση μεταξύ δύο ομάδων C_i και C_j είναι ο μέσος της απόστασης μεταξύ όλων των ζευγών των παρατηρήσεων που περιέχουν μία παρατήρηση της C_i και μία παρατήρηση της C_j (Εικόνα 4.5). Λόγω της χρησιμοποίησης όλων των παρατηρήσεων, διαφέρει από τις προηγούμενες μεθόδους και παρατηρείται πως προτιμάται περισσότερο από τους ερευνητές αλλά επηρεάζεται στην περίπτωση ακραίων παρατηρήσεων.

- **Η μέθοδος των μέσων εντός των ομάδων (average within groups)**

  Η μέθοδος των μέσων εντός των ομάδων αποτελεί παραλλαγή της μεθόδου των μέσων μεταξύ των ομάδων και είναι γνωστή και ως η μέθοδος σταθμισμένων ζευγών των ομάδων με τη χρησιμοποίηση αριθμητικών μέσων (weighted pair group method average, WPGMA). Σε αυτή την περίπτωση η απόσταση μεταξύ δύο ομάδων C_i και C_j ορίζεται ως ο μέσος όρος όλων των αποστάσεων των παρατηρήσεων που προκύπτουν ύστερα από την ένωση των ομάδων C_i και C_j.

- **Η μέθοδος κέντρου βάρους (centroid method)**

  Σύμφωνα με τη μέθοδο των κέντρων των ομάδων (UPGMC), η απόσταση μεταξύ δύο ομάδων C_i και C_j υπολογίζεται ως η απόσταση των κέντρων των ομάδων C_i και C_j (Εικόνα 4.5). Συνήθως παράγει συμπαγείς και ελλειπτικές ομάδες και στην περίπτωση ακραίων παρατηρήσεων παραμένει ανεπηρέαστη. Δε χρησιμοποιεί τον

πίνακα αποστάσεων καθώς όταν τα στοιχεία μας δεν είναι συνεχή χρησιμοποιείται η κορυφή ή η διάμεσος των ομάδων και για κάποιες μορφές δεδομένων είναι προτιμότερη η αποφυγή της χρήσης της μεθόδου καθώς αδυνατεί να χρησιμοποιήσει τον συγκεκριμένο εναλλακτικό τρόπο.

- **Η μέθοδος της διαμέσου (median)**

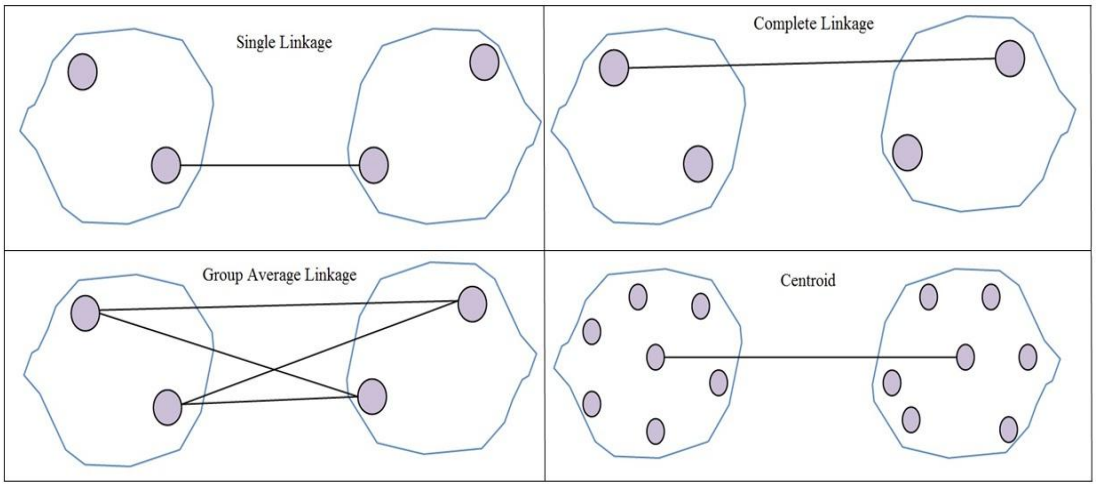
  Η μέθοδος των διαμέσων των ομάδων είναι γνωστή και ως ο σταθμισμένος συνδυασμός των κέντρων δύο ξεχωριστών ομάδων (weighted pair group method centroid WPGMC). Η διαφορά της με την μέθοδο των κέντρων των ομάδων είναι οι συστάδες σταθμίζονται ισοδύναμα προκειμένου να παραχθεί το νέο κέντρο της συγχωνευμένης συστάδας. Στην περίπτωση ακραίων παρατηρήσεων παραμένει ανεπηρέαστη.

- **Η μέθοδος του Ward (Ward' s method)**

  Η μέθοδος του Ward (1963) είναι γνωστή και ως η μέθοδος που ελαχιστοποιεί τις διακυμάνσεις μέσα στις ομάδες δημιουργώντας τελικά μία νέα ομάδα που θα έχει τη μικρότερη δυνατή διακύμανση. Αρχικά, υπολογίζουμε για κάθε παρατήρηση την απόστασή της από το κέντρο της ομάδας. Δηλαδή, έχοντας k ομάδες ορίζουμε ως SSE_k το άθροισμα των τετραγωνικών αποκλίσεων κάθε παρατήρησης από το μέσο της ομάδας και τότε το συνολικό άθροισμα ορίζεται $SSE=SSE_1+SSE_2+\ldots+SSE_k$. Προφανώς, επειδή στην αρχή κάθε αντικείμενο απαρτίζει και μία ομάδα, η απόστασή του από το κέντρο κάθε ομάδας είναι μηδέν άρα και το συνολικό άθροισμα είναι μηδέν. Δηλαδή για N παρατηρήσεις, $SSE_k=0$, $k=1,2,\ldots,N$ και $SSE=0$. Σε κάθε βήμα η ένωση των ομάδων πρέπει να οδηγεί στη μικρότερη αύξηση του συνολικού αθροίσματος των αποστάσεων. Στο τελικό στάδιο, που τα αντικείμενα πια αποτελούν μία ενιαία ομάδα, το συνολικό άθροισμα τετραγώνων των σφαλμάτων SSE (Sum of Squares Error) ορίζεται ως εξής :

$$\mathrm{SSE} = \sum_{i=1}^{N} (x_i - \bar{x})^2 \qquad (4.1)$$

Σε κάθε στάδιο επιδιώκεται η κατάταξη των παρατηρήσεων σε ομάδες που οδηγούν στο μικρότερο SSE μέσα στις ομάδες.



Εικόνα 4.5 Απεικόνιση των μεθόδων υπολογισμού των αποστάσεων μεταξύ των ομάδων

Η απόσταση μεταξύ της ομάδας C_l και C_{ij}, που δημιουργείται από τη συγχώνευση των C_i και C_j ομάδων, ορίζεται από τον **επαναληπτικό τύπο των Lance και Williams** ως εξής (Xu & Wunsch, 2009):

$$\mathrm{D}\big(C_l,(C_i,C_j)\big)=\alpha_i \mathrm{D}(C_l,C_i)+\alpha_j \mathrm{D}\big(C_l,C_j\big)+\beta D\big(C_i,C_j\big)+\gamma\left|\mathrm{D}(C_l,C_i)-\mathrm{D}\big(C_l,C_j\big)\right|, \qquad (4.2)$$

όπου $\mathrm{D}(,)$ είναι η συνάρτηση απόστασης και α_i, α_j, β και γ οι παράμετροι των οποίων οι τιμές αλλάζουν ανάλογα με τη τεχνική μεθόδου ομαδοποίησης που χρησιμοποιείται, όπως παρουσιάζεται και στον παρακάτω πίνακα (Πίνακας 4.1)**.**

Μέθοδοι ομαδοποίησης	α_i	α_j	β	γ
Single linkage	$1/2$	$1/2$	0	$-1/2$
Complete linkage	$1/2$	$1/2$	0	$1/2$
Group average linkage	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0
Weighted average	$1/2$	$1/2$	0	0
Centroid	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$\frac{-n_i n_j}{(n_i + n_j)^2}$	0
Median linkage	$1/2$	$1/2$	$-1/4$	0
Ward's method	$\frac{n_i + n_l}{n_i + n_{j+n_l}}$	$\frac{n_j + n_l}{n_i + n_j}$	$\frac{-n_l}{(n_i + n_j)^2}$	0

Πίνακας 4.1 Οι τιμές των παραμέτρων των Lance και Williams για τις συσσωρευτικές ιεραρχικές μεθόδους

Σύμφωνα λοιπόν με τον παραπάνω πίνακα τιμών των παραμέτρων ο επαναληπτικός τύπος των Lance και Williams έχει τη μορφή (Xu & Wunsch, 2009):

- Για τη μέθοδο απλής συνένωσης :

$$\mathrm{D}\big(C_l, (C_i, C_j)\big) = min\left(\mathrm{D}(C_l, C_i), \mathrm{D}\big(C_l, C_j\big)\right) \tag{4.3}$$

- Για τη μέθοδο απλής συνένωσης :

$$\mathrm{D}\big(C_l, (C_i, C_j)\big) = max\left(\mathrm{D}(C_l, C_i), \mathrm{D}\big(C_l, C_j\big)\right) \tag{4.4}$$

- Για τη μέθοδο των μέσων μεταξύ των ομάδων

$$\mathrm{D}\big(C_l, (C_i, C_j)\big) = \frac{1}{2}\left(\mathrm{D}(C_l, C_i), \mathrm{D}\big(C_l, C_j\big)\right) \tag{4.5}$$

- Για τη μέθοδο των μέσων εντός των ομάδων

$$\mathrm{D}\big(C_l, (C_i, C_j)\big) = \frac{n_i}{n_i + n_j}\mathrm{D}(C_l, C_i) + \frac{n_j}{n_i + n_j}\mathrm{D}\big(C_l, C_j\big) \tag{4.6}$$

- Για τη μέθοδο των κέντρων των ομάδων

$$\mathrm{D}\Big(C_l,\big(C_i,C_j\big)\Big)=\frac{n_i}{n_i+n_j}\mathrm{D}(C_l,C_i)+\frac{n_j}{n_i+n_j}\mathrm{D}\big(C_l,C_j\big)-\frac{n_in_j}{\big(n_i+n_j\big)^2}\mathrm{D}\big(C_i,C_j\big) \quad (4.7)$$

- Για τη μέθοδο των διαμέσων των ομάδων

$$\mathrm{D}\Big(C_l,\big(C_i,C_j\big)\Big)=\frac{1}{2}\mathrm{D}(C_l,C_i)+\frac{1}{2}\mathrm{D}\big(C_l,C_j\big)-\frac{1}{4}\mathrm{D}\big(C_i,C_j\big) \quad (4.8)$$

- Για τη μέθοδο του Ward

$$\mathrm{D}\Big(C_l,\big(C_i,C_j\big)\Big)=\frac{n_i+n_l}{n_i+n_{j+n_l}}\mathrm{D}(C_l,C_i)+\frac{n_j+n_l}{n_i+n_j}\mathrm{D}\big(C_l,C_j\big)-\frac{n_l}{\big(n_i+n_j\big)^2}\mathrm{D}\big(C_i,C_j\big) \quad (4.9)$$

όπου n_i, n_j και n_l οι τιμές των δεδομένων στις συστάδες C_i, C_j και C_l αντίστοιχα.

Εκτός από τις παραπάνω μεθόδους υπολογισμού αποστάσεων μεταξύ ομάδων υπάρχουν επιπλέον τεχνικές όπως η μέθοδος του αθροίσματος τετραγώνων (Jambu, 1978, Podani,1989) η οποία είναι παρόμοια με τη μέθοδο του Ward αλλά βασίζεται στο άθροισμα των τετραγώνων εντός κάθε ομάδας.

4.2.1.2 Διαιρετικές μέθοδοι ομαδοποίησης

Σε σχέση με τις συσσωρευτικές μεθόδους ομαδοποίησης, οι **διαιρετικές μέθοδοι** λειτουργούν, όπως αναφέρθηκε και παραπάνω, με την ακριβώς αντίστροφη διαδικασία. Δεν είναι αρκετά διαδεδομένες, καθώς ο υπολογιστικός φόρτος είναι μεγαλύτερος από ότι στις συσσωρευτικές. Για N παρατηρήσεις ένας διαιρετικός αλγόριθμος ξεκινά τη διαδικασία διάσπασης των δεδομένων της αρχικής ενιαίας ομάδας σε δύο υποομάδες, εκτελώντας $2^{N-1}-1$ υπολογισμούς ενώ στο αρχικό στάδιο της αλγοριθμικής διαδικασίας των συσσωρευτικών μεθόδων εκτελούνται n(n-1)/2 υπολογισμοί.

Οι διαιρετικές μέθοδοι χωρίζονται επιπλέον στις **μονοθετικές διαιρετικές μεθόδους (monothetic divisive methods)** και στις **πολυθετικές διαιρετικές μεθόδους (polythetic divisive methods)**. Στις μονοθετικές στηριζόμαστε στη χρήση μίας μεταβλητής προκειμένου να πραγματοποιηθεί η διάσπαση σε κάθε στάδιο. Οι μονοθετικές μέθοδοι είναι απλές και υπολογιστικά αποδοτικές και χρησιμοποιούνται για δεδομένα που αποτελούνται από δίτιμες

μεταβλητές. Χαρακτηριστικός αλγόριθμος των μονοθετικών μεθόδων είναι ο MONA (Monothetic Analysis) . Στις πολυθετικές μεθόδους γίνεται χρήση όλων των μεταβλητών σε κάθε στάδιο διαχωρισμού των ομάδων. Χαρακτηριστικός αλγόριθμος των πολυθετικών διαιρετικών μεθόδων είναι ο DIANA (Divisive Analysis) (Kaufman και Rousseeuw, 1990). Ο αλγόριθμος DIANA με επαναληπτικά βήματα στοχεύει να τοποθετήσει σε κάθε στάδιο της διαδικασίας διαχωρισμού τα αντικείμενα που είναι πιο κοντά σε μία ομάδα αποστατών (splinter group), στην οποία ουσιαστικά τοποθετείται το αντικείμενο που είναι το πιο απομακρυσμένο από τα υπόλοιπα αντικείμενα της ομάδας που θα διασπασθεί. Η συστάδα που επιλέγεται για περαιτέρω διαίρεση είναι εκείνη με τη μεγαλύτερη διάμετρο (τη μεγαλύτερη απόσταση μεταξύ ενός οποιουδήποτε ζεύγους αντικειμένων).

4.2.2 Μη ιεραρχικές μέθοδοι ομαδοποίησης

Σε αντίθεση με τις ιεραρχικές μεθόδους ομαδοποίησης που παρέχουν ένα διαδοχικό επίπεδο συστάδων με επαναληπτικές συγχωνεύσεις ή διασπάσεις, οι **μη ιεραρχικές μέθοδοι ομαδοποίησης (non - hierarchical clustering) ή μέθοδοι διαμέρισης (partitional clustering)** διαμερίζουν το πολυεπίπεδο του συνόλου των δεδομένων σε περιοχές κατατάσσοντας κάθε περιοχή και σε μια ομάδα χωρίς συγκεκριμένη ιεραρχική δομή. Κατηγοριοποιούν τα δεδομένα σε k συστάδες, οι οποίες ικανοποιούν τους εξής περιορισμούς:

Πρώτον, κάθε συστάδα πρέπει να αποτελείται τουλάχιστον από ένα αντικείμενο.

Δεύτερον, κάθε αντικείμενο πρέπει να ανήκει ακριβώς σε μία ομάδα.

Σύμφωνα με τους παραπάνω περιορισμούς, παρατηρείται πως οι συστάδες που θα δημιουργηθούν είναι στην καλύτερη περίπτωση όσα και τα αντικείμενα των δεδομένων. Ο αριθμός των συστάδων στις μεθόδους διαμέρισης είναι καθορισμένος από την αρχή, χωρίς ουσιαστικά να θεωρείται ότι είναι και ο κατάλληλος που απαιτείται για μια επιτυχημένη ομαδοποίηση. Οι μη ιεραρχικές μέθοδοι ομαδοποίησης ή μέθοδοι διαμέρισης υλοποιούνται από αλγορίθμους που οι μεταξύ τους διαφορές κατά τη διαδικασία ομαδοποίησης παρουσιάζονται στην ταξινόμηση των παρατηρήσεων που απομένουν στις συστάδες και στον υπολογισμό των νέων κέντρων των ομάδων (υπάρχει παρακάτω σχετική αναφορά).

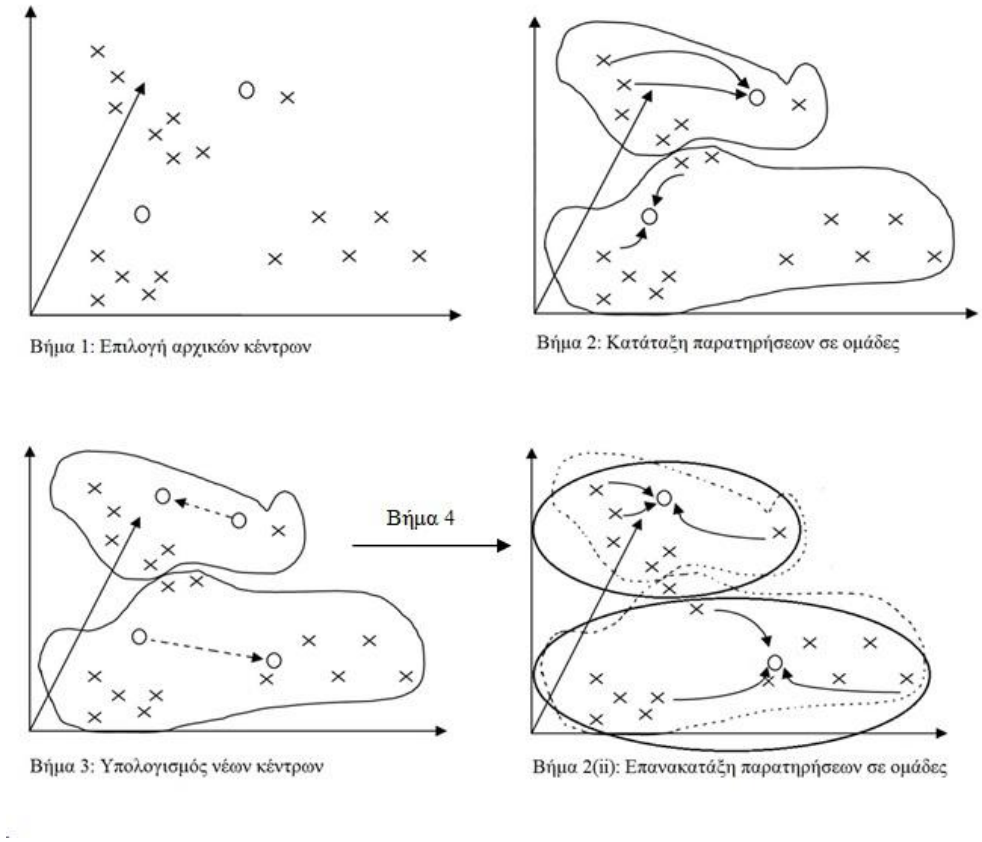
Η πιο δημοφιλής από τις μεθόδους διαμέρισης είναι η **μέθοδος K-means** (Forgy, 1965; MacQueen, 1967). Η μέθοδος, μέσω μιας επαναληπτικής διαδικασίας βελτιστοποίησης, επιδιώκει τη βέλτιστη διαμέριση των δεδομένων. Χαρακτηρίζεται για την απλότητα και την ευκολία στην εφαρμογή της και επιδιώκει τη βέλτιστη διαμέριση των συστάδων ώστε να ελαχιστοποιείται το σφάλμα του αθροίσματος των τετραγωνικών αποστάσεων των παρατηρήσεων από τα κέντρα των ομάδων στις οποίες ανήκουν. Ο αριθμός των ομάδων είναι γνωστός εκ των προτέρων και σε κάθε βήμα επιτρέπει την επανεξέταση και των διαχωρισμό των παρατηρήσεων από τις ομάδες στις οποίες έχουν ενταχθεί. Η αλγοριθμική διαδικασία που ακολουθείται κατά τη μέθοδο K-means είναι η εξής (Xu & Wunsch, 2009):

1. Επέλεξε μία αρχική αυθαίρετη διαμέριση των παρατηρήσεων σε K αριθμό συστάδων. Υπολόγισε τον πίνακα των κεντροειδών $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \dots \mathbf{m}_\mathrm{K}]$
2. Αντιστοίχισε κάθε παρατήρηση του συνόλου δεδομένων στη πλησιέστερη συστάδα C_l, δηλαδή $\mathbf{x}_j \in C_l$, αν $\left\|\mathbf{x}_j - \mathbf{m}_l\right\| < \left\|\mathbf{x}_j - \mathbf{m}_i\right\|$ για j=1,…N, i≠l, και i=1,…K
3. Υπολόγισε το νέο πίνακα των κεντροειδών με βάση την τρέχουσα διαμέριση,

$$\mathbf{m}_i = \frac{1}{\boldsymbol{N}_\mathrm{i}} \sum_{\mathrm{x_j} \epsilon C_l} \mathbf{x}_\mathrm{j}$$

4. Επανέλαβε τα βήματα 2 και 3 μέχρι καμία παρατήρηση να μην αλλάζει συστάδα και δεν υπάρχουν αλλαγές σε κάθε συστάδα.

Η εικόνα 4.6 απεικονίζει τη διαδικασία του αλγορίθμου K-means, όπου οι παρατηρήσεις συμβολίζονται με «x» και τα κέντρα των ομάδων με «o» (Καρλής, 2005).



Εικόνα 4.6 Η διαδικασία της μεθόδου K-means

Η μέθοδος K-means είναι εξαιρετικά χρήσιμη σε περιπτώσεις μεγάλων σετ δεδομένων καθώς δε χρειάζεται ιδιαίτερα μεγάλη υπολογιστική ισχύ , είναι γρήγορη μέθοδος και δεν απαιτούνται αρκετές επαναλήψεις για την ολοκλήρωσή της. Απαιτεί όμως να προσδιοριστεί εξ' αρχής, η αρχική διαμέριση των παρατηρήσεων σε Κ συστάδες, το μέτρο απόστασης και το πιο κρίσιμο ο αριθμός των Κ συστάδων. Διαφορετικές αρχικές διαμερίσεις των παρατηρήσεων μπορούν να οδηγήσουν σε διαφορετικές τελικά ομαδοποιήσεις. Ένας τρόπος που μπορεί να χρησιμοποιήσει ο ερευνητής για να αντιμετωπίσει το συγκεκριμένο πρόβλημα είναι να τρέξει τον αλγόριθμο K-means με πολλαπλές αρχικές διαμερίσεις, επιλέγοντας τελικά τη διαμέριση με το μικρότερο σφάλμα τετραγώνων. Ένα άλλο μειονέκτημα της μεθόδου k-means είναι πως εφαρμόζεται καλύτερα χρησιμοποιώντας συνεχείς μεταβλητές. Στην περίπτωση των κατηγορικών μεταβλητών, η μέθοδος K-means τις μετατρέπει σε dummy μεταβλητές (dummy attributes) και το σύνολο των dummy μεταβλητών αντικαθιστούν τις αρχικές κατηγορικές μεταβλητές και ο αλγόριθμος τις εφαρμόζει στην ανάλυση θεωρώντας τις ως συνεχείς μεταβλητές.

4.2.3 TwoStep Ανάλυση κατά συστάδες (TwoStep Cluster)

Οι περιορισμοί και τα προβλήματα των κλασσικών μεθόδων οδήγησαν στην ανάπτυξη πολλών άλλων μεθόδων που αποτελούν παραλλαγές και επεκτάσεις τους και επιδιώκουν τη βέλτιστη διαμέριση των παρατηρήσεων, αντιμετωπίζοντάς τες διαφορετικά. Κάποιες από τις εξειδικευμένες μεθόδους που έχουν αναπτυχθεί χρησιμοποιούν την ιεραρχική μέθοδο ομαδοποίησης με ένα στάδιο δημιουργίας πρωταρχικών συστάδων προ ομαδοποίησης (pre clustering). Μία από αυτές τις μεθόδους είναι και η TwoStep ανάλυση κατά συστάδες (Everitt, Landau, Leese, & Stahl, 2011).

Η TwoStep ανάλυση συστάδων αποτελεί μία μέθοδο ομαδοποίησης που αν και αρχικά σχεδιάστηκε για την ανάλυση μεγάλων σετ δεδομένων, παρέχει περισσότερα και ουσιώδη πλεονεκτήματα από τις ιεραρχικές και μη ιεραρχικές μεθόδους που αναλύσαμε παραπάνω. Η μέθοδος TwoStep έχει τη δυνατότητα να ορίζει αυτόματα ένα βέλτιστο αριθμό ομάδων, εφαρμόζεται σε μεικτές μεταβλητές (συνεχείς ή κατηγορικές) κατηγοριοποιώντας τα δεδομένα σε ομάδες με τη χρήση κριτηρίων προσεγγισιμότητας. Για τον υπολογισμό της απόστασης μεταξύ των συστάδων χρησιμοποιούνται η Ευκλείδεια απόσταση και η log-likelihood απόσταση. Στην περίπτωση ομαδοποίησης δεδομένων που χαρακτηρίζονται από συνεχείς μεταβλητές χρησιμοποιείται η Ευκλείδεια απόσταση (έχει οριστεί σε προηγούμενη ενότητα) ενώ όταν χαρακτηρίζονται από το συνδυασμό συνεχών και κατηγορικών μεταβλητών (μεικτές μεταβλητές) χρησιμοποιείται η log-likelihood απόσταση μεταξύ των συστάδων *i* και *j* σύμφωνα με τον τύπο (Chiu, Fang, Chen, Wang, & Jeris, 2001):

$$d(i,j) = \xi_i + \xi_j - \xi_{<i,j>} \tag{4.10}$$

$$\xi_s = -\mathrm{N_s}\left(\sum_{\mathrm{k}=1}^{\mathrm{K^A}} \frac{1}{2}\log\left(\widehat{\sigma_k^2} + \widehat{\sigma_{sk}^2}\right) + \sum_{\mathrm{k}=1}^{\mathrm{K}^B} \widehat{\mathrm{E_{sk}}}\right) \ (4.11) \ \kappa\alpha\iota \ \mathrm{E_{sk}} = -\sum_{l=1}^{\mathrm{L_k}} \frac{\mathrm{N}_{sk1}}{\mathrm{N}_s} \log \frac{\mathrm{N}_{sk1}}{\mathrm{N}_s} \quad (4.12)$$

όπου:

- d(i,j) η απόσταση μεταξύ των συστάδων i και j
- <i,j> αναπαριστά τη συστάδα που προκύπτει ύστερα από τη συγχώνευση των συστάδων i και j
- $\mathrm{K^A}$ ο αριθμός των συνεχών μεταβλητών
- $\mathrm{K^B}$ ο αριθμός των συνεχών μεταβλητών
- $\mathrm{L_k}$ ο αριθμός των κατηγοριών της k-οστής κατηγορικής μεταβλητής
- $\mathrm{N_s}$ το πλήθος των παρατηρήσεων της *s*-οστής συστάδας

- N_{skl} το πλήθος των παρατηρήσεων στην *s*-οστή συστάδα που ανήλουν στην *l*-οστή κατηγορία της *k*-οστής κατηγορικής μεταβλητής
- σ_k^2 η εκτίμηση της διασποράς της k-οστής συνεχούς μεταβλητής στο σύνολο των δεδομένων
- σ_{sk}^2 η εκτίμηση της διασποράς της k-οστής συνεχούς μεταβλητής στη συστάδα s

Ο αλγόριθμος που υλοποιεί τη μέθοδο περιλαμβάνει δύο βήματα: α)το βήμα της προ–ομαδοποίησης (pre-clustering step) και β)το τελικό βήμα ομαδοποίησης (clustering step) (Chiu, Fang, Chen, Wang, & Jeris, 2001).

- Το βήμα της προ-ομαδοποίησης αφορά τη δημιουργία πρωταρχικών συστάδων ακολουθώντας μία διαδοχική διαδικασία συσταδοποίησης. Ο αλγόριθμος σαρώνει μία προς μία τις εγγραφές από το σύνολο δεδομένων και αποφασίζει κάθε φορά αν η τρέχουσα παρατήρηση πρέπει να συγχωνευτεί με οποιαδήποτε ήδη διαμορφωμένη πρωταρχική συστάδα ή θα δημιουργήσει μία καινούρια με βάση του κριτηρίου της απόστασης.

- Στο τελικό βήμα της ομαδοποίησης, ο αλγόριθμος έχοντας τις πρωταρχικές συστάδες που δημιουργήθηκαν στο προηγούμενο βήμα τις ομαδοποιεί ώστε να δημιουργήσει τον επιθυμητό αριθμό συστάδων. Για να καταλήξει στις τελικές ομάδες χρησιμοποιεί μία συσσωρευτική ιεραρχική μέθοδο ομαδοποίησης στις ήδη υπάρχουσες συστάδες.

Η συνεισφορά των μεταβλητών στη δημιουργία των τελικών συστάδων ελέγχεται και για τις συνεχείς μεταβλητές και για τις κατηγορικές μεταβλητές. Για τις συνεχείς μεταβλητές η σημαντικότητα βασίζεται στον υπολογισμό της τιμής $t = \frac{\hat{\mu}_k - \hat{\mu}_{jk}}{\hat{\sigma}_{jk}/\sqrt{N_k}}$ (4.13), ενώ για τις κατηγορικές μεταβλητές υπολογίζεται η τιμή ελέγχου $\chi^2 = \sum_{l=1}^{L_k} \left(\frac{N_{jkl} - N_{kl}}{N_{kl}}\right)^2$ (4.14).

4.3 Έλεγχος εγκυρότητας ομαδοποίησης

Λόγω του γεγονότος πως διαφορετικές προσεγγίσεις συσταδοποίησης μπορεί να οδηγούν σε διαφορετικό είδος ή/και αριθμό ομάδων (ακόμα και όταν αυτές δεν υπάρχουν!), είναι αναγκαία ή ύπαρξη κριτηρίων για την εγκυρότητα και την αξιοπιστία των αποτελεσμάτων που παράγουν οι μέθοδοι. Αυτά τα κριτήρια θα πρέπει να είναι αντικειμενικά και αμερόληπτα ως προς την προτίμηση σε κάποια συγκεκριμένη μέθοδο και να απαντούν σε ερωτήματα που αφορούν στον αριθμό των ομάδων που εγκολπώνουν τα δεδομένα, στο αν οι εξαγόμενες ομάδες έχουν πρακτική σημασία ή στο γιατί να πρέπει να χρησιμοποιηθεί μια μέθοδος έναντι μίας άλλης.

Ανάλογα με τη μέθοδο συσταδοποίησης που θα επιλεχθεί υπάρχουν διαφορετικές κατηγορίες κριτηρίων ελέγχου της εξαχθείσας δομής της συσταδοποίησης:

- Για την ιεραρχική μέθοδο ομαδοποίησης ένα από τα κριτήρια είναι ο συντελεστής συσχέτισης cophenetic (Cophenetic Correlation Coefficient,CPCC). Δοθέντος ενός πίνακα αποστάσεων $\mathbf{P}\{q_{ij}\}$ ενός συνόλου δεδομένων **X**, το κριτήριο CPCC μετρά το βαθμό ομοιότητας μεταξύ του πίνακα αποστάσεων **P** και ενός cophenetic πίνακα $\mathbf{Q}\{q_{ij}\}$,του οποίου τα στοιχεία αναπαριστούν το επίπεδο της εγγύτητας μεταξύ των ζευγών των αντικειμένων των δεδομένων που ομαδοποιούνται στην ίδια συστάδα για πρώτη φορά. Με $\mu_P = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} p_{ij}$ και $\mu_Q = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} q_{ij}$ τους μέσους των **P** και **Q**, όπου M=N(N-1)/2, το CPCC ορίζεται ως

$$CPCC = \frac{\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} p_{ij}q_{ij} - \mu_P\mu_Q}{\sqrt{\left(\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} p_{ij}^2 - \mu_P^2\right)}\left(\left(\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} q_{ij}^2 - \mu_Q^2\right)\right)} \quad (4.12)$$

με τιμές στο διάστημα [-1,1]. Όταν η τιμή του δείκτη είναι κοντά στην τιμή 1 τότε παρουσιάζεται σημαντική ομοιότητα μεταξύ των πινάκων P και Q και μία ικανοποιητική ιεραρχική δομή των δεδομένων (Xu & Wunsch, 2009).

- Στην περίπτωση που η μέθοδος ομαδοποίησης απαιτεί την εισαγωγή από τον ερευνητή του αριθμού των συστάδων (όπως η k-means), ένα από τα κριτήρια είναι

το VRC (Variance Ratio Criterion) (Calinski and Harabasz,1974). Με N τον αριθμό των αντικειμένων και K διαμερίσεις το κριτήριο VRC ορίζεται ως

$$\mathrm{CH(K)} = \frac{\mathrm{T_r(S_B)}}{\mathrm{K-1}} \Big/ \frac{\mathrm{T_r(S_W)}}{\mathrm{N-K}} \tag{4.13}$$

Η τιμή του K που μεγιστοποιεί το κριτήριο CH(K) υποδεικνύει και την εκτίμηση του αριθμού των συστάδων K (Xu & Wunsch, 2009).

- Στην περίπτωση που ο ερευνητής επιλέξει την TwoStep ανάλυση συσταδοποίησης, η μέθοδος προσδιορίζει αυτόματα το βέλτιστο αριθμό των ομάδων σε δύο στάδια. Στο πρώτο στάδιο χρησιμοποιεί είτε το κριτήριο BIC (Bayesian Information Criterion), είτε το κριτήριο AIC (Akaike's Information Criterion). Η τιμή των K διαμερίσεων που μεγιστοποιεί το κριτήριο BIC και ελαχιστοποιεί το κριτήριο AIC αποτελεί και την εκτίμηση του βέλτιστου αριθμού K (Xu & Wunsch, 2009). Στο δεύτερο στάδιο χρησιμοποιεί τη μεταβολή των R(K) αποστάσεων των K συστάδων όπου $R(K)= d_{K-1}/d_K$, με d_{K-1} την απόσταση στην περίπτωση που οι K συστάδες συγχωνευτούν σε K-1 συστάδες. Όμοια ορίζεται και η d_K.. Όταν αυξάνεται η μεταβολή λαμβάνεται και ο αριθμός των συστάδων. Η μεταβολή των αποστάσεων ορίζεται ως $R(K_1)/ R(K_2)$ για τις δύο μεγαλύτερες τιμές της R(K) ($K=1,2,\ldots,K_{max}$). Αν η πρώτη μεγαλύτερη τιμή μεταβολής είναι κατά 1.15 φορές μεγαλύτερη από τη δεύτερη μεγαλύτερη μεταβολή, επιλέγεται ως βέλτιστος αριθμός συστάδων η περίπτωση με το μεγαλύτερο δείκτη R(K). Διαφορετικά, από τις δύο περιπτώσεις με τις δύο μεγαλύτερες τιμές R(K) επιλέγεται εκείνη με τις περισσότερες (Bacher, Wenzig, & Vogler, 2004).

Για να εξασφαλιστεί η εγκυρότητα της συσταδοποίησης είναι απαραίτητο να επαναληφθεί αρκετές φορές η διαδικασία ανάλυσης στα ίδια δεδομένα, εφαρμόζοντας παραπάνω από μία μεθόδους ομαδοποίησης-κριτήρια σύνδεσης-μέτρα εγγύτητας, εξετάζοντας τις ομοιότητες και τις διαφορές των τελικών αποτελεσμάτων. Μια άλλη μέθοδος, αφορά το χωρισμό των δεδομένων σε δύο ομάδες και την εφαρμογή σε κάθε ομάδα της διαδικασίας της ανάλυσης με τις ίδιες παραμέτρους. Αν τα τελικά κέντρα των συστάδων δε διαφέρουν σημαντικά τότε εξασφαλίζεται η σταθερότητα της αρχικής συσταδοποίησης.

Στην περίπτωση όμως που διαφέρουν επιβάλλεται ο έλεγχος ύπαρξης ακραίων τιμών που ενδεχομένως να επηρεάζουν το τελικό αποτέλεσμα (Mooi & Sarstedt, 2011).

Η αξιοπιστία των αποτελεσμάτων της συσταδοποίησης αναφέρεται ουσιαστικά στην απαίτηση οι προκύπτουσες ομάδες να παραμένουν σταθερές στο χρόνο. Για την ανάπτυξη στρατηγικών πρέπει να διασφαλίζεται η σταθερότητα της σύστασης των τελικών ομάδων ή η συμπεριφορά των ατόμων-αντικειμένων ώστε να αποφευχθεί η παραγωγή ανεπαρκών αποτελεσμάτων. Απαραίτητη λοιπόν είναι η επανάληψη της διαδικασίας της ανάλυσης μετά από κάποια χρονική περίοδο (κυρίως σε έρευνες αγοράς), ώστε να διασφαλιστεί η εγκυρότητα των αποτελεσμάτων (Mooi & Sarstedt, 2011).

Επίσης, μετά την ανάλυση των δεδομένων, θα πρέπει να ελεγχτεί και η αναγκαιότητα χρήσης όλων των μεταβλητών στην ανάλυση για την εξαγωγή συμπερασμάτων για τον αριθμό των ομάδων με την αφαίρεση των μη χρήσιμων μεταβλητών (μια μεταβλητή δε θα είναι χρήσιμη αν, όταν αφαιρεθεί και εφαρμοστεί ξανά η μέθοδος ομαδοποίησης, τα αποτελέσματα δε θα αλλάξουν.

4.4 Ερμηνεία αποτελέσματος

Στο τελικό στάδιο της ανάλυσης κατά συστάδες εξετάζονται κατά πόσο οι τελικές συστάδες είναι εννοιολογικά διαχωρίσιμες. Βασιζόμενοι στις παρατηρήσιμες μεταβλητές (δημογραφικά, γεωγραφικά, κοινωνικοοικονομικά χαρακτηριστικά, την κατάσταση του πελάτη κ.ά.) εξάγεται ένα προφίλ για κάθε τελική συστάδα που διευκολύνει την εισαγωγή νέων ατόμων-αντικειμένων στις υπάρχουσες συστάδες, αποφεύγοντας τη διεξαγωγή εκ νέου της διαδικασία συσταδοποίησης (Mooi & Sarstedt, 2011).

Ο στόχος της ομαδοποίησης είναι η εξαγωγή χρήσιμων και με σημασία συμπερασμάτων για την ύπαρξη ή μη μιας δομής που μπορεί να διατρέχει τα αρχικά δεδομένα με απώτερο σκοπό την διαμόρφωση κάποιων υποθέσεων προς έλεγχο από τους ερευνητές, οι οποίοι μπορούν να χρησιμοποιήσουν τα αποτελέσματα της ομαδοποίησης σε συνδυασμό με τις ειδικές γνώσεις και εμπειρία τους για να βρουν απαντήσεις στο υπό μελέτη ερώτημα.

5. Εφαρμογή και ανάλυση δεδομένων

5.1 Εισαγωγή

Στην παρούσα διπλωματική εργασία θα μελετήσουμε την ανάλυση των πελατών μιας τράπεζας και την κατηγοριοποίησή τους σε ομάδες που αναδεικνύουν τη φερεγγυότητά τους ή μη στην εξόφληση των τραπεζικών προϊόντων που τους έχουν χορηγηθεί.

Οι πελάτες διαχωρίζονται σε κατηγορίες ανάλογα με τις μέρες καθυστέρησης αποπληρωμής των τραπεζικών προϊόντων που τους έχουν χορηγηθεί. Οι μέρες καθυστέρησης παρουσιάζονται σε ένα πίνακα δείκτη καθυστέρησης πληρωμών (bucket) (Πίνακας 5.1).

Δείκτης	Ημέρες καθυστέρησης
0	0
1	1-29
2	30-59
3	60-89
4	90-119
5	120-149
6	150-179
7	180-209
8	210-239
…	…

Πίνακας 5.1Δείκτης καθυστέρησης πληρωμών

Στην πρώτη στήλη παρουσιάζονται οι τιμές του δείκτη καθυστέρησης πληρωμών και στη δεύτερη στήλη οι ημέρες καθυστέρησης των πληρωμών. Με δείκτη 0 είναι οι πελάτες που δεν έχουν καθυστερήσει την αποπληρωμή των προϊόντων τους. Με δείκτη 1 οι πελάτες που έχουν καθυστερήσει 1-29 ημέρες την αποπληρωμή, με δείκτη 2 έχουν καθυστερήσει 30-59 ημέρες κτλ. Κάθε μήνα ελέγχεται η τιμή του δείκτη για κάθε πελάτη της τράπεζας, εντάσσοντας τον πελάτη και σε διαφορετική κατηγορία. Δηλαδή, οι πελάτες με δείκτη 0-6

αντιμετωπίζονται διαφορετικά από τους πελάτες με δείκτη από 7 και περισσότερο. Στην πρώτη περίπτωση ανήκουν οι «καλοί» πελάτες και στη δεύτερη περίπτωση οι «κακοί» πελάτες. Όταν ένας πελάτης ξεπερνάει το δείκτη 6 γίνεται write off (κατάσταση διαγραφής), εντάσσεται σε διαφορετική κατηγορία και αντιμετωπίζεται με διαφορετικό τρόπο από την τράπεζα που αθέτησε τις υποχρεώσεις του, αλλά και από τους υπόλοιπους χρηματοπιστωτικούς οργανισμούς της χώρας. Πιο συγκεκριμένα, δε μπορεί να λάβει τα προνόμια (όπως χρησιμοποίηση πιστωτικών καρτών κ.ά.) ούτε και προσφορές σε προϊόντα που λαμβάνουν οι πελάτες έως τον δείκτη 6.

Θα επικεντρωθούμε στους πελάτες με δείκτες 0 έως 6 και στοχεύουμε στο διαχωρισμό τους σε ομάδες, προκειμένου να κατανοήσουμε ποιοι πελάτες είναι εκείνοι που πλησιάζουν τους επόμενους μήνες την κατάσταση του write off. Η τράπεζα βασιζόμενη στις ομάδες πελατών που θα προκύψουν, αποφασίζει για ποια ομάδα πελατών θα προχωρήσει σε κατάλληλα μέτρα αποφυγής ένταξής τους στην κατάσταση του write off μέσω προτάσεων και διακανονισμών επιδιώκοντας την επίτευξη αποπληρωμής των υποχρεώσεων τους.

5.2 Το δείγμα και οι μεταβλητές

Τα δεδομένα που χρησιμοποιούνται στην ανάλυση προήλθαν από Ελληνική τράπεζα στην Αθήνα και συλλέχτηκαν τους μήνες : Ιανουάριος, Φεβρουάριος και Μάρτιος του έτους 2011.

Η βάση δεδομένων περιέχει **21956 εγγραφές**-πελάτες και **11 μεταβλητές** οι οποίες περιγράφονται στον παρακάτω πίνακα (Πίνακας 5.2):

	Ονομασία Μεταβλητής	Ερμηνεία
1	**write_off_status**	Αν ο πελάτης γίνει write off (0) ή όχι (1) μετά από 3 μήνες από τη στιγμή της παρατήρησης
2	**pay_last3months**	Το άθροισμα των πληρωμών τους τελευταίους 3 μήνες προς το ποσό που οφείλει ο πελάτης (ποσοστό)
3	**months6_sumbucket**	Το άθροισμα του δείκτη των καθυστερήσεων των πληρωμών τους τελευταίους 6 μήνες (Δείκτης Παραβατικότητας)
4	**customer_worst_bucket**	Η χειρότερη καθυστέρηση όλων των προϊόντων του πελάτη (5, 6, 7)
5	**last_payment_months**	Οι μήνες που έχουν περάσει από την τελευταία πληρωμή του πελάτη (0,1,2,3,4,5,6)
6	**property**	Αν ο πελάτης έχει περιουσιακά στοιχεία (0) ή όχι (1)
7	**segment**	Το είδος των τραπεζικών προϊόντων που έχει ο πελάτης (1=Cards, 2=Loans, 3=Cards and Loans)
8	**amount_pastdue_onus**	Το ποσό καθυστέρησης του πελάτη
9	**balance**	Το υπόλοιπο του πελάτη τη στιγμή της καθυστέρησης
10	**Tiresias**	Αν ο πελάτης ανήκει στον Τειρεσία (0) ή όχι (1)
11	**income**	Το εισόδημα του πελάτη

Πίνακας 5.2 Πίνακας μεταβλητών

5.3 Περιγραφικά μέτρα και Γραφήματα

Οι **ποσοτικές – συνεχείς μεταβλητές** που χρησιμοποιούμε για την ανάλυσή μας είναι οι pay_last3months, months6_sumbucket, amount_pastdue_onus, balance και income (το άθροισμα των πληρωμών τους τελευταίους 3 μήνες προς το ποσό που οφείλει ο πελάτης ως ποσοστό, το άθροισμα του δείκτη καθυστερήσεων των πληρωμών τους τελευταίους 6 μήνες - Δείκτης Παραβατικότητας, το ποσό καθυστέρησης του πελάτη, το υπόλοιπο του πελάτη τη στιγμή της καθυστέρησης και το εισόδημα του πελάτη), των οποίων η μέση τιμή (κέντρο), ελάχιστη τιμή, μέγιστη τιμή, το εύρος και η τυπική απόκλιση φαίνονται στον παρακάτω πίνακα (Πίνακας 5.3):

Descriptive Statistics

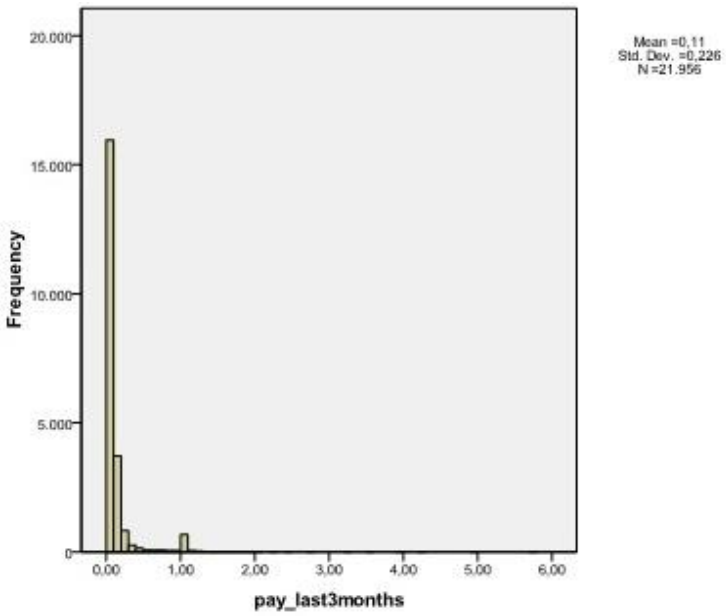
	N	Range	Minimum	Maximum	Mean	Std. Deviation
pay_last3months	21956	5,72	,00	5,72	,1066	,22572
months6_sumbucket	21956	32	4	36	18,63	6,163
amount_pastdue_onus	21956	40565,92	,00	40565,92	1597,4256	2471,03975
balance	21956	58754,69	,25	58754,94	7549,5163	8271,09890
income	20507	298619,92	1380,08	300000,00	29164,6928	23945,56698
Valid N (listwise)	20507					

Πίνακας 5.3

Μεταβλητή pay_last3months

Το άθροισμα των πληρωμών τους τελευταίους 3 μήνες προς το ποσό που οφείλει ο πελάτης δείχνει το ποσοστό αποπληρωμής του χρέους του κάθε πελάτη σε τριμηνιαία βάση Κυμαίνεται από 0 έως 5,72% μονάδες με μέση τιμή το 0,1066%. Επίσης, η μεταβλητότητα του ποσοστού αποπληρωμής, ως προς το μέσο ποσοστό, είναι μεγάλη (0,22572%).

Όπως φαίνεται από το επόμενο γράφημα (Γράφημα 5.1), που απεικονίζει την κατανομή του ποσοστού αποπληρωμής του χρέους του κάθε πελάτη σε τριμηνιαία βάση, παρουσιάζει αριστερή ασυμμετρία ως προς το μέσο ποσοστό αποπληρωμής, δηλαδή οι μεγάλες συχνότητες συγκεντρώνονται στο αριστερό άκρο της κατανομής που αντιστοιχεί στα χαμηλότερα ποσοστά αποπληρωμής σε σχέση με το μέσο.



Γράφημα 5.1

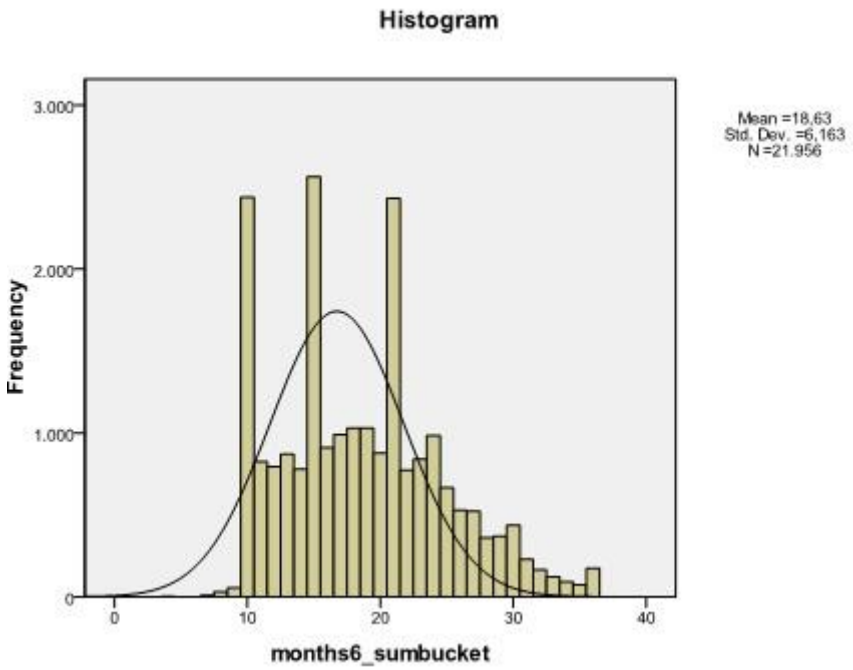
Μεταβλητή months6_sumbucket

Στον επόμενο πίνακα (Πίνακας 5.4) καταγράφονται οι κατανομές συχνοτήτων του βαθμού παραβατικότητας του πελάτη όπως αυτός αντανακλάται από το άθροισμα του δείκτη των καθυστερήσεων των πληρωμών τους τελευταίους 6 μήνες. Ο βαθμός παραβατικότητας κυμαίνεται από 4 έως 36 μονάδες με μέση τιμή τις 18,63. Οι μισοί πελάτες εμφανίζουν βαθμό παραβατικότητας μικρότερο των 18 μονάδων ενώ οι συχνότεροι βαθμοί παραβατικότητας είναι οι 10, 15, 21 με 11,1%, 11,7% και 11,1% αντίστοιχα. Επίσης, η μεταβλητότητα των βαθμών παραβατικότητας, ως προς τη μέση παραβατικότητα, είναι 6,163 βαθμοί και ο μέση τιμή συμπίπτει με τη διάμεσο.

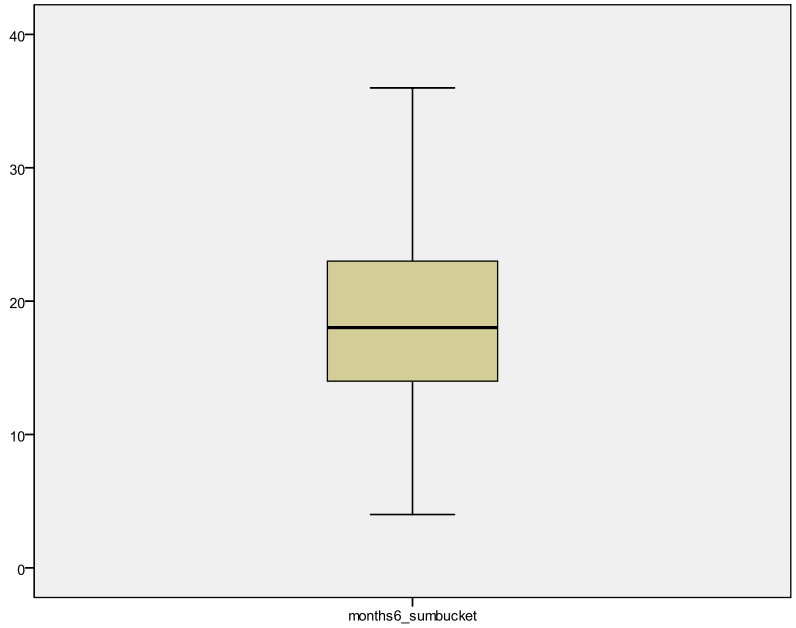
months6_sumbucket

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	4	2	,0	,0	,0
	6	1	,0	,0	,0
	7	10	,0	,0	,1
	8	31	,1	,1	,2
	9	54	,2	,2	,4
	10	2436	11,1	11,1	11,5
	11	823	3,7	3,7	15,3
	12	793	3,6	3,6	18,9
	13	869	4,0	4,0	22,9
	14	778	3,5	3,5	26,4
	15	2563	11,7	11,7	38,1
	16	910	4,1	4,1	42,2
	17	990	4,5	4,5	46,7
	18	1029	4,7	4,7	51,4
	19	1029	4,7	4,7	56,1
	20	877	4,0	4,0	60,1
	21	2430	11,1	11,1	71,2
	22	773	3,5	3,5	74,7
	23	839	3,8	3,8	78,5
	24	984	4,5	4,5	83,0
	25	665	3,0	3,0	86,0
	26	527	2,4	2,4	88,4
	27	521	2,4	2,4	90,8
	28	361	1,6	1,6	92,4
	29	370	1,7	1,7	94,1
	30	437	2,0	2,0	96,1
	31	230	1,0	1,0	97,2
	32	165	,8	,8	97,9
	33	122	,6	,6	98,5
	34	91	,4	,4	98,9
	35	73	,3	,3	99,2
	36	173	,8	,8	100,0
	Total	21956	100,0	100,0	

Πίνακας 5.4



Γράφημα 5.2



Γράφημα 5.3

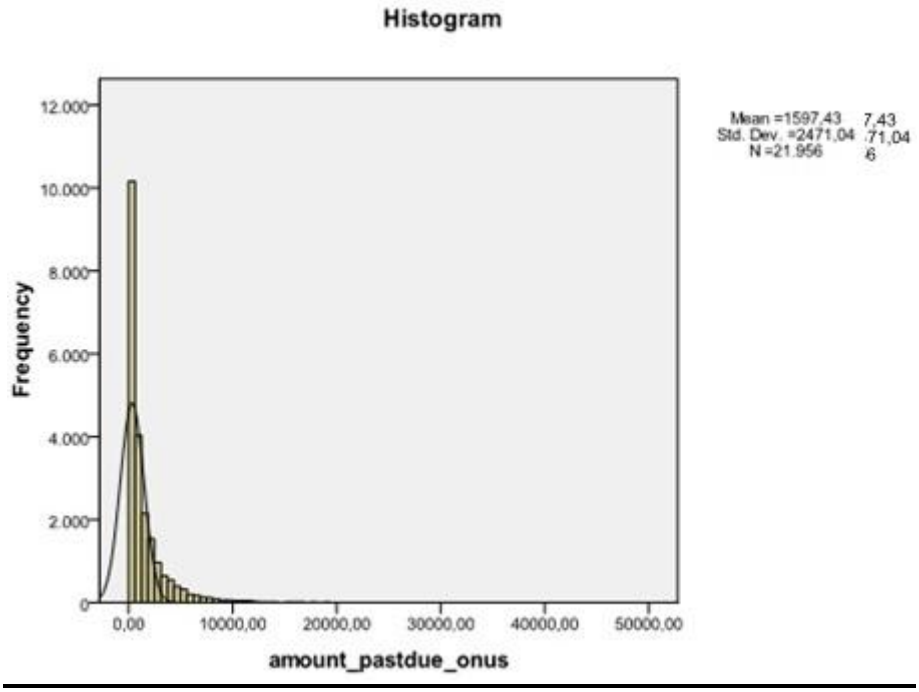
Όπως φαίνεται από τα δύο προηγούμενα γραφήματα (Γράφημα 5.2 και Γράφημα 5.3), που απεικονίζουν την κατανομή του βαθμού παραβατικότητας, το κεντρικό 50% των βαθμών παραβατικότητας κυμαίνεται από 14 έως 25 βαθμούς ενώ τα άνω και κάτω 25% των βαθμών παραβατικότητας είναι από 25 έως 36 βαθμούς, και από 4 έως 14, αντίστοιχα. Μπορούμε να συμπεράνουμε πως η μεταβλητότητα είναι μεγάλη, δεν υπάρχουν ακραίοι βαθμοί και

υπάρχουν 3 βαθμοί παραβατικότητας (10, 15, 21) που καταλαμβάνουν αθροιστικά το 33,9%, δηλαδή το 1/3 του συνόλου με την ίδια συνεισφορά.

Μεταβλητή amount_pastdue_onus

Το ποσό καθυστέρησης του πελάτη κυμαίνεται από 0 έως 40565,92 € με μέση τιμή τα 1597,4 €. Επίσης, η μεταβλητότητα του ποσού καθυστέρησης, ως προς το μέσο ποσό, είναι μεγάλη (2471,03 €).

Όπως φαίνεται από το επόμενο γράφημα (Γράφημα 5.4), που απεικονίζει την κατανομή του ποσού καθυστέρησης, παρουσιάζει αριστερή ασυμμετρία ως προς το μέσο ποσό καθυστέρησης, δηλαδή οι μεγάλες συχνότητες συγκεντρώνονται στο αριστερό άκρο της κατανομής που αντιστοιχεί στα χαμηλότερα ποσά καθυστέρησης σε σχέση με το μέσο.

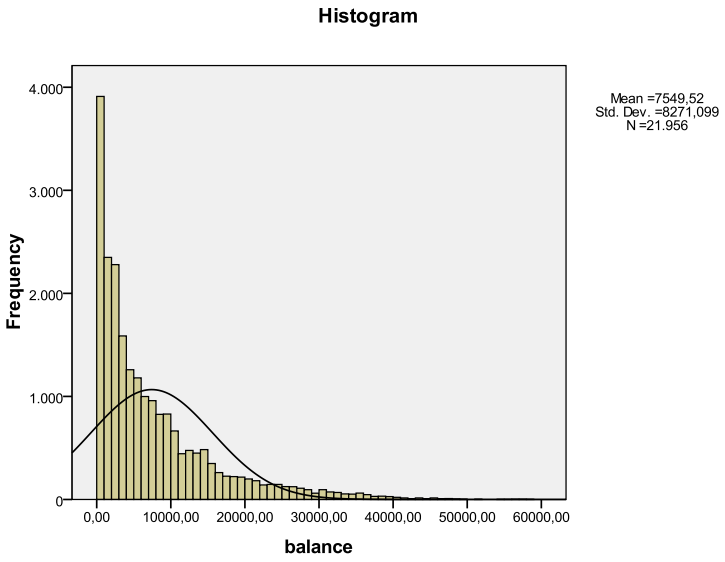


Γράφημα 5.4

Μεταβλητή balance

Το υπόλοιπο του πελάτη τη στιγμή της καθυστέρησης κυμαίνεται από 0,25 έως 58754,94 € με μέση τιμή τα 7549,51 €. Επίσης, η μεταβλητότητα του υπολοίπου του πελάτη, ως προς το μέσο υπόλοιπο, είναι μεγάλη (8271,09 €).

Όπως φαίνεται από το επόμενο γράφημα (Γράφημα 5.5), που απεικονίζει την κατανομή του υπολοίπου του πελάτη, παρουσιάζει αριστερή ασυμμετρία ως προς το μέσο υπόλοιπο, δηλαδή οι μεγάλες συχνότητες συγκεντρώνονται στο αριστερό άκρο της κατανομής που αντιστοιχεί στα χαμηλότερα υπόλοιπα σε σχέση με το μέσο.

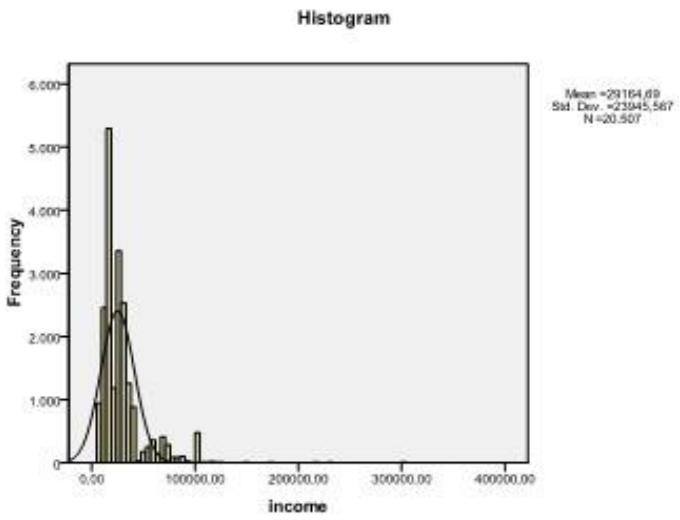


Γράφημα 5.5

Μεταβλητή income

Το εισόδημα του πελάτη κυμαίνεται από 1380,08 έως 300000,00 € με μέση τιμή τα 29164,69 €. Επίσης, η μεταβλητότητα του εισοδήματος του πελάτη, ως προς το μέσο εισόδημα, είναι μεγάλη (23945,56 €).

Όπως φαίνεται από το επόμενο γράφημα (Γράφημα 5.6), που απεικονίζει την κατανομή του εισοδήματος του πελάτη, παρουσιάζει αριστερή ασυμμετρία ως προς το μέσο εισόδημα, δηλαδή οι μεγάλες συχνότητες συγκεντρώνονται στο αριστερό άκρο της κατανομής που αντιστοιχεί στα χαμηλότερα εισοδήματα σε σχέση με το μέσο.

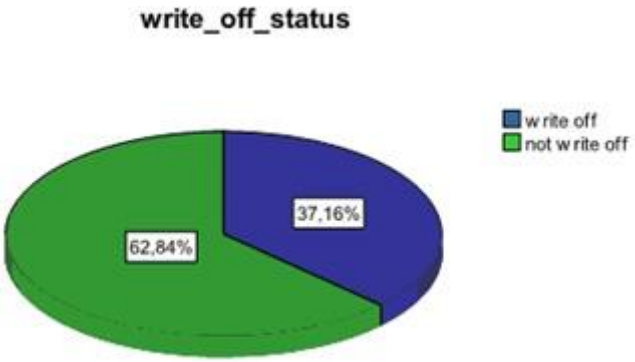


Γράφημα 5.6

Οι **ποιοτικές – κατηγορικές μεταβλητές** που χρησιμοποιούμε για την ανάλυσή μας είναι οι write_off_status, customer_worst_bucket, last_payment_months, property, segment και Tiresias (αν ο πελάτης γίνει written off (0) ή όχι (1) μετά από 3 μήνες από τη στιγμή της παρατήρησης, η χειρότερη καθυστέρηση όλων των προϊόντων του πελάτη (5, 6, 7), οι μήνες που έχουν περάσει από την τελευταία πληρωμή του πελάτη (0,1,2,3,4,5,6), αν ο πελάτης έχει περιουσιακά στοιχεία (0) ή όχι (1) , το είδος των τραπεζικών προϊόντων που έχει ο πελάτης (1=Cards, 2=Loans, 3=Cards and Loans) και αν ο πελάτης ανήκει στον Τειρεσία (0) ή όχι (1)) των οποίων φαίνονται στον παρακάτω πίνακα:

Μεταβλητή Write_off_status

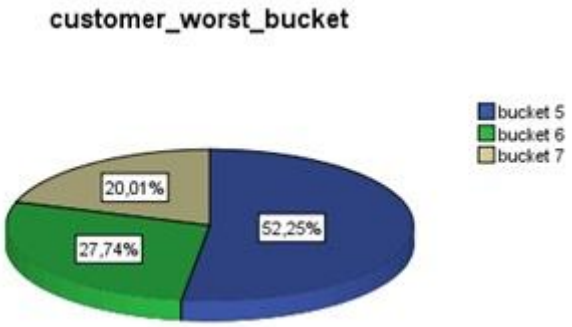
Η μεταβλητή αυτή παίρνει την τιμή 0 (δηλαδή ότι ο πελάτης ενδέχεται να γίνει written off μέσα στους οποίους 3 μήνες) αν ο πελάτης έχει δείκτη καθυστέρησης (bucket) 4, 5, 6 και 1 (δηλαδή ότι ο πελάτης δεν ενδέχεται να γίνει written off μέσα στους οποίους 3 μήνες) αν έχει δείκτη καθυστέρησης 0, 1, 2, 3. Από το παρακάτω γράφημα (Γράφημα 5.7), παρατηρούμε πως το 37,16% των πελατών ενδέχεται να γίνει written off μετά από 3 μήνες από τη στιγμή της παρατήρησης και το 62,84% δεν ενδέχεται να γίνει.



Γράφημα 5.7

Μεταβλητή Customer_worst_bucket

Από το παρακάτω γράφημα (Γράφημα 5.8), παρατηρούμε πως το 52,25% των πελατών είχε κατά το παρελθόν ανέλθει στη χειρότερη καθυστέρηση αποπληρωμής χρεών δείκτη 5, το 27,74% δείκτη 6 και το 20,01% δείκτη 7.

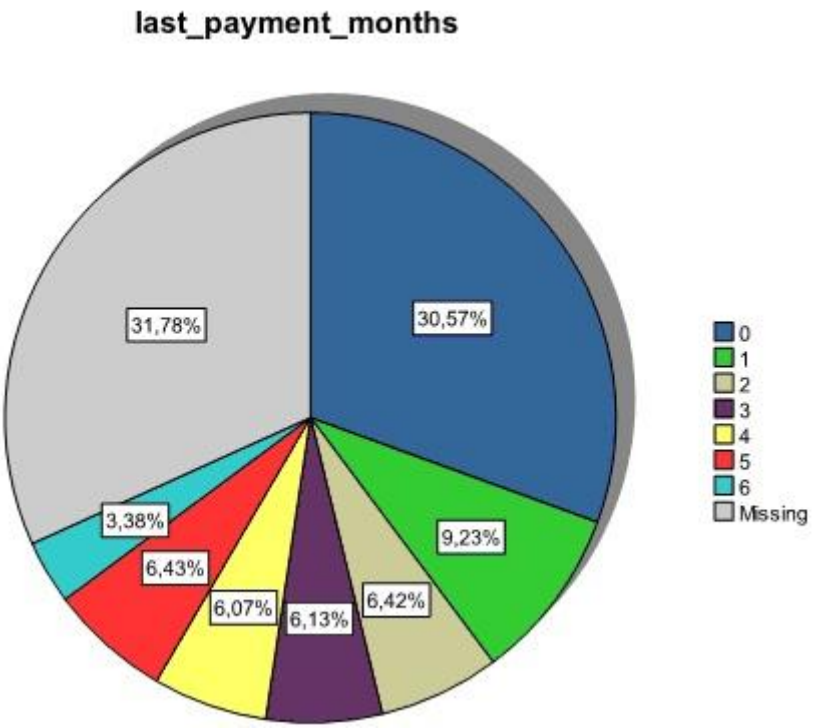


Γράφημα 5.8

Μεταβλητή Last_payment_months

Από το παρακάτω γράφημα (Γράφημα 5.9), παρατηρούμε πως για το 30,57% των πελατών έχουν περάσει 0 μήνες από την τελευταία πληρωμή τους, για το 9,23% των πελατών 1 μήνας, για το 6,42% των πελατών 2 μήνες, για το 6,13% των πελατών 3 μήνες, για το 6,07% των πελατών 4 μήνες, για το 6,43% των πελατών 5 μήνες και για το 3,38% των πελατών 6 μήνες.

Επίσης, για το 31,78% των πελατών δεν έχουμε διαθέσιμη αυτήν την πληροφορία.

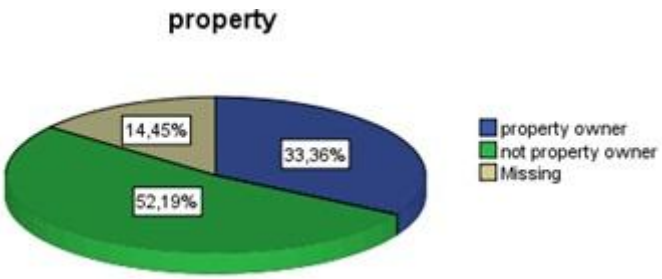


Γράφημα 5.9

Μεταβλητή Property

Από το παρακάτω γράφημα (Γράφημα 5.10), παρατηρούμε πως το 33,36% των πελατών έχουν ιδιοκτησία και το 52,19% των πελατών δεν έχει.

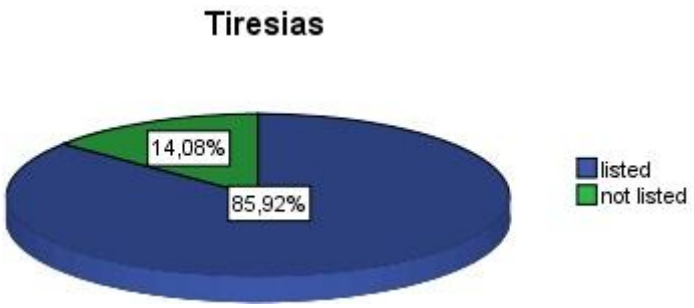
Επίσης, για το 14,45% των πελατών δεν έχουμε διαθέσιμη αυτήν την πληροφορία.



Γράφημα 5.10

Μεταβλητή Tiresias

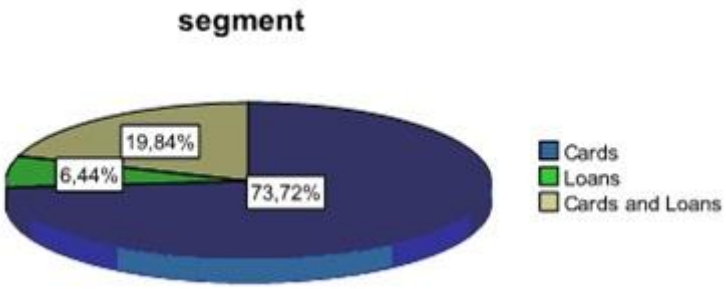
Από το παρακάτω γράφημα (Γράφημα 5.11), παρατηρούμε πως το 85,92% των πελατών είναι εγγεγραμμένοι στα αρχεία δεδομένων του Τειρεσία και το 14,08% των πελατών δεν είναι.



Γράφημα 5.11

Μεταβλητή Segment

Από το παρακάτω γράφημα (Γράφημα 5.12), παρατηρούμε πως το 73,72% των πελατών έχει λάβει ως τραπεζικό προϊόν Πιστωτικές Κάρτες, το 6,44% Δάνειο και το 19,84% και τα δύο.



Γράφημα 5.12

5.4 Συσχετίσεις μεταξύ μεταβλητών

Ο έλεγχος $\mathbf{X^2}$ (Pearson Chi-square) αποτελεί επαγωγική διαδικασία η οποία διερευνά τη σχέση δύο κατηγορικών μεταβλητών που υπεισέρχονται στη δομή ενός πίνακα συνάφειας. Ειδικότερα, ελέγχει την υπόθεση ότι οι δύο μεταβλητές δεν σχετίζονται (είναι ανεξάρτητες) μεταξύ τους ή, με πιο απλή διατύπωση, ότι δεν επιδρά η μία στην άλλη.

Η **μηδενική υπόθεση** ($\mathbf{H_0}$) αφορά τη μη συσχέτιση (ή μη συνάφεια ή ανεξαρτησία) των 2 αυτών μεταβλητών. Αν το p-value είναι μικρότερο του 0,05, τότε η μηδενική υπόθεση της «ανεξαρτησίας» απορρίπτεται και επομένως, οι δύο μεταβλητές συσχετίζονται (ή έχουν συνάφεια ή εξαρτώνται) η μία με την άλλη.

Α. Θέλουμε να διερευνήσουμε αν ο δείκτης της χειρότερης καθυστέρησης σε τραπεζικό προϊόν που έχει σημειωθεί στο ιστορικό του πελάτη επιδρά στο αν ο πελάτης ενδέχεται να διαγραφεί μετά από 3 μήνες ή όχι.

customer_worst_bucket * write_off_status Crosstabulation

			write_off_status		
			write off	not write off	Total
customer_worst_bucket	bucket 5	Count	5461	6010	11471
		% within customer_worst_bucket	47,6%	52,4%	100,0%
		% within write_off_status	66,9%	43,6%	52,2%
	bucket 6	Count	2087	4004	6091
		% within customer_worst_bucket	34,3%	65,7%	100,0%
		% within write_off_status	25,6%	29,0%	27,7%
	bucket 7	Count	611	3783	4394
		% within customer_worst_bucket	13,9%	86,1%	100,0%
		% within write_off_status	7,5%	27,4%	20,0%
Total		Count	8159	13797	21956
		% within customer_worst_bucket	37,2%	62,8%	100,0%
		% within write_off_status	100,0%	100,0%	100,0%

Πίνακας 5.5

Από την ανάγνωση του παραπάνω πίνακα (Πίνακας 5.5), προκύπτει πως στο σύνολο των πελατών που ενδέχεται να γίνουν written off στους επόμενους 3 μήνες, το 66,9% παρουσιάζουν δείκτη 5 ως τη χειρότερη καθυστέρηση που είχε σημειωθεί στο ιστορικό τους, το 25,6% παρουσιάζουν δείκτη 6 ως τη χειρότερη καθυστέρηση που είχε σημειωθεί στο ιστορικό τους και το 7,5% παρουσιάζουν δείκτη 7 ως τη χειρότερη καθυστέρηση που είχε σημειωθεί στο ιστορικό τους.

Επίσης, προκύπτει πως στο σύνολο των πελατών που έχουν παρουσιάσει δείκτη 5 ως τη χειρότερη καθυστέρηση στο παρελθόν, το ποσοστό αυτών που ενδέχεται να διαγραφούν είναι μεγαλύτερο του συνολικού δειγματικού ποσοστού, 47,6% έναντι 37,2%, ενώ στο σύνολο των πελατών που έχουν παρουσιάσει δείκτη 6 ως τη χειρότερη καθυστέρηση στο παρελθόν, το ποσοστό αυτών που ενδέχεται να διαγραφούν είναι μικρότερο του συνολικού δειγματικού ποσοστού, 34,3% έναντι 37,2%, και στο σύνολο των πελατών που έχουν παρουσιάσει δείκτη 7 ως τη χειρότερη καθυστέρηση στο παρελθόν, το ποσοστό αυτών που ενδέχεται να διαγραφούν είναι πολύ μικρότερο του συνολικού δειγματικού ποσοστού, 13,9% έναντι 37,2%.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1575,584[a]	2	,000
Likelihood Ratio	1,723,581	2	,000
Linear-by-Linear Association	1,553,820	1	,000
N of Valid Cases	21956		

Πίνακας 5.6

Η μηδενική υπόθεση (H_0) αφορά την ανεξαρτησία μεταξύ των 2 αυτών μεταβλητών. Το p-value είναι μικρότερο του 0,05 (0,0<0,05) (Πίνακας 5.6), επομένως η μηδενική υπόθεση της ανεξαρτησίας δεν γίνεται αποδεκτή κάτι που σημαίνει πως οι δύο μεταβλητές εξαρτώνται η μία από την άλλη, δηλαδή η χειρότερη καθυστέρηση όλων των προϊόντων του πελάτη επιδρά στο ενδεχόμενο ο πελάτης να γίνει written off μετά από 3 μήνες.

B. Θέλουμε να διερευνήσουμε αν οι μήνες που έχουν περάσει από την τελευταία πληρωμή του πελάτη (0,1,2,3,4,5,6) επιδρούν στο αν θα διαγραφεί μετά από 3 μήνες ή όχι.

last_payment_months * write_off_status Crosstabulation

			write_off_status		
			write off	not write off	Total
last_payment_months	0	Count	4085	2627	6712
		% within last_payment_months	60,9%	39,1%	100,0%
		% within write_off_status	50,3%	38,3%	44,8%
	1	Count	1175	851	2026
		% within last_payment_months	58,0%	42,0%	100,0%
		% within write_off_status	14,5%	12,4%	13,5%
	2	Count	788	621	1409
		% within last_payment_months	55,9%	44,1%	100,0%
		% within write_off_status	9,7%	9,1%	9,4%
	3	Count	737	608	1345
		% within last_payment_months	54,8%	45,2%	100,0%
		% within write_off_status	9,1%	8,9%	9,0%
	4	Count	670	663	1333
		% within last_payment_months	50,3%	49,7%	100,0%
		% within write_off_status	8,3%	9,7%	8,9%
	5	Count	658	753	1411
		% within last_payment_months	46,6%	53,4%	100,0%
		% within write_off_status	8,1%	11,0%	9,4%
	6	Count	5	738	743
		% within last_payment_months	,7%	99,3%	100,0%
		% within write_off_status	,1%	10,8%	5,0%
Total		Count	8118	6861	14979
		% within last_payment_months	54,2%	45,8%	100,0%
		% within write_off_status	100,0%	100,0%	100,0%

Πίνακας 5.7

Από την ανάγνωση του παραπάνω πίνακα (Πίνακας 5.7), προκύπτει πως στο σύνολο των πελατών που ενδέχεται να γίνουν written off στους επόμενους 3 μήνες, το 50,3% πλήρωσε για τελευταία φορά πριν 0 μήνες, το 14,5% πλήρωσε για τελευταία φορά πριν 1 μήνα, το 9,7% πλήρωσε για τελευταία φορά πριν 2 μήνες, το 9,1% πλήρωσε για τελευταία φορά πριν 3 μήνες, το 8,3% πλήρωσε για τελευταία φορά πριν 4 μήνες, το 8,1% πλήρωσε για τελευταία φορά πριν 5 μήνες και το 0,1% πλήρωσε για τελευταία φορά πριν 6 μήνες.

Επίσης, προκύπτει πως στο σύνολο των πελατών που έχουν πληρώσει για τελευταία φορά πριν 0 μήνες, το ποσοστό αυτών που ενδέχεται να διαγραφούν είναι μεγαλύτερο του συνολικού δειγματικού ποσοστού, 60,9% έναντι 54,2%, με το ίδιο να ισχύει και για αυτούς που έχουν να πληρώσουν από 1 έως 3 μήνες, ενώ στο σύνολο των πελατών που έχουν πληρώσει για τελευταία φορά πριν 4, 5 και 6 μήνες, τα ποσοστά αυτών που ενδέχεται να διαγραφούν είναι μικρότερα του συνολικού δειγματικού ποσοστού.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1032,042[a]	6	,000
Likelihood Ratio	1,274,386	6	,000
Linear-by-Linear Association	606,629	1	,000
N of Valid Cases	14979		

Πίνακας 5.8

Το p-value είναι μικρότερο του 0,05 (0,0<0,05) (Πίνακας 5.8), επομένως η μηδενική υπόθεση της ανεξαρτησίας δεν γίνεται αποδεκτή κάτι που σημαίνει πως οι δύο μεταβλητές εξαρτώνται η μία από την άλλη, δηλαδή ο αριθμός των μηνών που έχουν περάσει από την τελευταία πληρωμή από τον πελάτη επιδρά στο ενδεχόμενο να γίνει written off μετά από 3 μήνες.

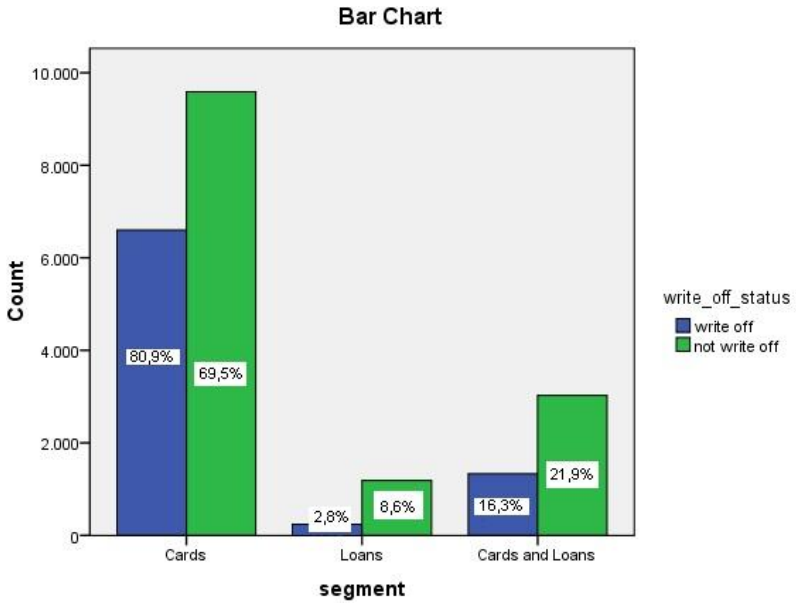
Γ. Θέλουμε να διερευνήσουμε αν το είδος των τραπεζικών προϊόντων που έχει ο πελάτης (1=Cards, 2=Loans, 3=Cards and Loans) επιδρούν στο αν θα διαγραφεί μετά από 3 μήνες ή όχι.

segment * write_off_status Crosstabulation

			write_off_status		
			write off	not write off	Total
segment	Cards	Count	6598	9587	16185
		% within segment	40,8%	59,2%	100,0%
		% within write_off_status	80,9%	69,5%	73,7%
	Loans	Count	228	1186	1414
		% within segment	16,1%	83,9%	100,0%
		% within write_off_status	2,8%	8,6%	6,4%
	Cards and Loans	Count	1333	3024	4357
		% within segment	30,6%	69,4%	100,0%
		% within write_off_status	16,3%	21,9%	19,8%
Total		Count	8159	13797	21956
		% within segment	37,2%	62,8%	100,0%
		% within write_off_status	100,0%	100,0%	100,0%

Πίνακας 5.9

Από την ανάγνωση του παραπάνω πίνακα (Πίνακας 5.9) και όπως απεικονίζεται και στο επόμενο γράφημα (Γράφημα 5.13)., προκύπτει πως στο σύνολο των πελατών που ενδέχεται να γίνουν written off στους επόμενους 3 μήνες, το 80,9% έχει Πιστωτικές Κάρτες το 2,8% έχει Δάνειο και το 16,3% έχει Πιστωτικές Κάρτες και Δάνειο.



Γράφημα 5.13

Επίσης, προκύπτει πως στο σύνολο των πελατών που έχουν Πιστωτικές Κάρτες, το ποσοστό αυτών που ενδέχεται να διαγραφούν είναι μεγαλύτερο του συνολικού δειγματικού ποσοστού, 40,8% έναντι 37,2% ενώ στο σύνολο των πελατών που έχουν Δάνειο και Πιστωτικές Κάρτες και Δάνειο, τα ποσοστά αυτών που ενδέχεται να διαγραφούν είναι μικρότερα του συνολικού δειγματικού ποσοστού

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	438,505[a]	2	,000
Likelihood Ratio	475,941	2	,000
Linear-by-Linear Association	228,542	1	,000
N of Valid Cases	21956		

Πίνακας 5.10

Το p-value είναι μικρότερο του 0,05 (0,0<0,05) (Πίνακας 5.10), επομένως η μηδενική υπόθεση της ανεξαρτησίας δεν γίνεται αποδεκτή κάτι που σημαίνει πως οι δύο μεταβλητές εξαρτώνται η μία από την άλλη, δηλαδή το είδος του τραπεζικού προϊόντος που έχει λάβει ο πελάτης επιδρά στο ενδεχόμενο να γίνει written off μετά από 3 μήνες.

Δ. Θέλουμε να διερευνήσουμε αν ο πελάτης ανήκει στον Τειρεσία επιδρά στο αν θα διαγραφεί μετά από 3 μήνες ή όχι.

Tiresias * write_off_status Crosstabulation

			write_off_status		Total
			write off	not write off	
Tiresias	listed	Count	6283	12581	18864
		% within Tiresias	33,3%	66,7%	100,0%
		% within write_off_status	77,0%	91,2%	85,9%
	not listed	Count	1876	1216	3092
		% within Tiresias	60,7%	39,3%	100,0%
		% within write_off_status	23,0%	8,8%	14,1%
Total		Count	8159	13797	21956
		% within Tiresias	37,2%	62,8%	100,0%
		% within write_off_status	100,0%	100,0%	100,0%

Πίνακας 5.11

Από την ανάγνωση του παραπάνω πίνακα (Πίνακας 5.11), προκύπτει πως στο σύνολο των πελατών που ενδέχεται να γίνουν written off στους επόμενους 3 μήνες, το 77% είναι στον Τειρεσία και το 23% δεν είναι.

Επίσης, προκύπτει πως στο σύνολο των πελατών που είναι στον Τειρεσία, το ποσοστό αυτών που ενδέχεται να διαγραφούν είναι μικρότερο του συνολικού δειγματικού ποσοστού, 33,3% έναντι 37,2%, ενώ στο σύνολο των πελατών που δεν είναι στον Τειρεσία, το ποσοστό αυτών που ενδέχεται να διαγραφούν είναι πολύ μεγαλύτερο του συνολικού δειγματικού ποσοστού.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	851,969[a]	1	,000		
Continuity Correction[b]	850,797	1	,000		
Likelihood Ratio	821,450	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	851,930	1	,000		
N of Valid Cases	21956				

Πίνακας 5.12

Το p-value είναι μικρότερο του 0,05 (0,0<0,05) (Πίνακας 5.12), επομένως η μηδενική υπόθεση της ανεξαρτησίας δεν γίνεται αποδεκτή κάτι που σημαίνει πως οι δύο μεταβλητές εξαρτώνται η μία από την άλλη, δηλαδή το αν ο πελάτης είναι στον Τειρεσία επιδρά στο ενδεχόμενο να γίνει written off μετά από 3 μήνες.

Ε. Θέλουμε να διερευνήσουμε αν το αν ο πελάτης έχει ιδιοκτησία επιδρά στο αν θα διαγραφεί μετά από 3 μήνες ή όχι.

property * write_off_status Crosstabulation

			write_off_status		
			write off	not write off	Total
property	property owner	Count	2795	4529	7324
		% within property	38,2%	61,8%	100,0%
		% within write_off_status	37,5%	40,0%	39,0%
	not property owner	Count	4663	6796	11459
		% within property	40,7%	59,3%	100,0%
		% within write_off_status	62,5%	60,0%	61,0%
Total		Count	7458	11325	18783
		% within property	39,7%	60,3%	100,0%
		% within write_off_status	100,0%	100,0%	100,0%

Πίνακας 5.13

Από την ανάγνωση του παραπάνω πίνακα (Πίνακας 5.13), προκύπτει πως στο σύνολο των πελατών που ενδέχεται να γίνουν written off στους επόμενους 3 μήνες, το 37,5% έχει ιδιοκτησία και το 62,5% δεν έχει.

Επίσης, προκύπτει πως στο σύνολο των πελατών που έχουν ιδιοκτησία, το ποσοστό αυτών που ενδέχεται να διαγραφούν είναι μικρότερο του συνολικού δειγματικού ποσοστού, 38,2% έναντι 39,7%, ενώ στο σύνολο των πελατών που δεν έχουν ιδιοκτησία, το ποσοστό αυτών που ενδέχεται να διαγραφούν είναι λίγο μεγαλύτερο του συνολικού δειγματικού ποσοστού.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	11,953[a]	1	,001		
Continuity Correction[b]	11,848	1	,001		
Likelihood Ratio	11,974	1	,001		
Fisher's Exact Test				,001	,000
Linear-by-Linear Association	11,952	1	,001		
N of Valid Cases	18783				

Πίνακας 5.14

Το p-value είναι μικρότερο του 0,05 (0,0<0,05) (Πίνακας 5.14), επομένως η μηδενική υπόθεση της ανεξαρτησίας δεν γίνεται αποδεκτή κάτι που σημαίνει πως οι δύο μεταβλητές εξαρτώνται η μία από την άλλη, δηλαδή η κατοχή ιδιοκτησίας επιδρά στο ενδεχόμενο ο πελάτης να γίνει written off μετά από 3 μήνες.

Correlations

		write_off_status	pay_last3months	months6_sumbucket	customer_worst_bucket	last_payment_months	property	segment	amount_pastdue_onus	balance	Tiresias	income
write_off_status	Pearson Correlation	1	-,376**	-,096**	,266**	,201**	-,025**	,102**	,069**	,134**	-,197**	-,038**
	Sig. (2-tailed)		,000	,000	,000	,000	,001	,000	,000	,000	,000	,000
	N	21956	21956	21956	21956	14979	18783	21956	21956	21956	21956	20507
pay_last3months	Pearson Correlation	-,376**	1	,020**	-,092**	,181**	,030**	,013	-,035**	-,096**	,194**	,017*
	Sig. (2-tailed)	,000		,003	,000	,000	,000	,052	,000	,000	,000	,013
	N	21956	21956	21956	21956	14979	18783	21956	21956	21956	21956	20507
months6_sumbucket	Pearson Correlation	-,096**	,020**	1	,638**	-,063**	,012	-,089**	,148**	,074**	-,004	,029**
	Sig. (2-tailed)	,000	,003		,000	,000	,110	,000	,000	,000	,507	,000
	N	21956	21956	21956	21956	14979	18783	21956	21956	21956	21956	20507
customer_worst_bucket	Pearson Correlation	,266**	-,092**	,638**	1	,037**	-,044**	-,242**	,174**	,032**	-,124**	,001
	Sig. (2-tailed)	,000	,000	,000		,000	,000	,000	,000	,000	,000	,836
	N	21956	21956	21956	21956	14979	18783	21956	21956	21956	21956	20507
last_payment_months	Pearson Correlation	,201**	,181**	-,063**	,037**	1	-,035**	,003	-,020*	-,039**	-,024**	-,003
	Sig. (2-tailed)	,000	,000	,000	,000		,000	,668	,015	,000	,003	,724
	N	14979	14979	14979	14979	14979	13665	14979	14979	14979	14979	13894
property	Pearson Correlation	-,025**	,030**	,012	-,044**	-,035**	1	,107**	,110**	,159**	,065**	,129**
	Sig. (2-tailed)	,001	,000	,110	,000	,000		,000	,000	,000	,000	,000
	N	18783	18783	18783	18783	13665	18783	18783	18783	18783	18783	17708
segment	Pearson Correlation	,102**	,013	-,089**	-,242**	,003	,107**	1	,056**	,321**	,438**	,008
	Sig. (2-tailed)	,000	,052	,000	,000	,668	,000		,000	,000	,000	,270
	N	21956	21956	21956	21956	14979	18783	21956	21956	21956	21956	20507
amount_pastdue_onus	Pearson Correlation	,069**	-,035**	,148**	,174**	-,020*	,110**	,056**	1	,628**	,032**	,116**
	Sig. (2-tailed)	,000	,000	,000	,000	,015	,000	,000		,000	,000	,000
	N	21956	21956	21956	21956	14979	18783	21956	21956	21956	21956	20507
balance	Pearson Correlation	,134**	-,096**	,074**	,032**	-,039**	,159**	,321**	,628**	1	,078**	,115**
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000		,000	,000
	N	21956	21956	21956	21956	14979	18783	21956	21956	21956	21956	20507
Tiresias	Pearson Correlation	-,197**	,194**	-,004	-,124**	-,024**	,065**	,438**	,032**	,078**	1	,012
	Sig. (2-tailed)	,000	,000	,507	,000	,003	,000	,000	,000	,000		,079
	N	21956	21956	21956	21956	14979	18783	21956	21956	21956	21956	20507
income	Pearson Correlation	-,038**	,017*	,029**	,001	-,003	,129**	,008	,116**	,115**	,012	1
	Sig. (2-tailed)	,000	,013	,000	,836	,724	,000	,270	,000	,000	,079	
	N	20507	20507	20507	20507	13894	17708	20507	20507	20507	20507	20507

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Πίνακας 5.15

Όσον αφορά τις **συσχετίσεις** των μεταβλητών που διαθέτουμε, κάνουμε έλεγχο για τον δειγματικό συντελεστή συσχέτισης ο οποίος εκτιμά τον πληθυσμιακό συντελεστή συσχέτισης. Ο συντελεστής συσχέτισης είναι ένας δείκτης της γραμμικής εξάρτησης μεταξύ 2 μεταβλητών ο οποίος λαμβάνει τιμές στο διάστημα [-1,1] (όταν είναι κοντά στο 1 και -1 σημαίνει ότι οι αλλαγές (αύξηση ή μείωση) στην τιμή της 1 μεταβλητής συνοδεύονται με αλλαγές προς την αντίθετη κατεύθυνση και για τις τιμές της άλλης μεταβλητής και προς την ίδια κατεύθυνση αντίστοιχα). Η ένταση της συσχέτισης εξαρτάται από το ύψος του συντελεστή. Όταν λαμβάνει την τιμή 0, δεν υπάρχει καμία σχέση μεταξύ των δύο μεταβλητών. Έτσι μπορούμε, βάσει του πίνακα των συσχετίσεων (Πίνακας 5.15), να καταλήξουμε στα εξής συμπεράσματα:

- Οι μεταβλητές write_off_status, property, amount_pastdue_onus και balance δηλαδή το ενδεχόμενο ο πελάτης να γίνει written off μετά από 3 μήνες, η κατοχή ιδιοκτησίας, το ποσό καθυστέρησης και το υπόλοιπο του πελάτη τη στιγμή της καθυστέρησης παρουσιάζουν **στατιστικά σημαντικές συσχετίσεις με όλες τις υπόλοιπες μεταβλητές**.

- **Υψηλή και στατιστικά σημαντική συσχέτιση** παρουσιάζουν τα εξής ζεύγη μεταβλητών:

    - write_off_status και pay_last3months (το ενδεχόμενο ο πελάτης να γίνει written off μετά από 3 μήνες και το άθροισμα των πληρωμών τους τελευταίους 3 μήνες προς το ποσό που οφείλει ο πελάτης ως ποσοστό)

    - months6_sumbucket και customer_worst_bucket (το άθροισμα των καθυστερήσεων των πληρωμών τους τελευταίους 6 μήνες -Δείκτης Παραβατικότητας και η χειρότερη καθυστέρηση όλων των προϊόντων του πελάτη)

    - customer_worst_bucket και segment (η χειρότερη καθυστέρηση όλων των προϊόντων του πελάτη και το είδος του τραπεζικού προϊόντος που έχει λάβει ο πελάτης)

    - segment και balance (το είδος του τραπεζικού προϊόντος που έχει λάβει ο πελάτης και το υπόλοιπό του)

- segment και Tiresias (το είδος του τραπεζικού προϊόντος που έχει λάβει ο πελάτης και το αν είναι στο Τειρεσία)
- amount_pastdue_onus και balance (το ποσό καθυστέρησης και το υπόλοιπο)

- Τα παρακάτω ζεύγη μεταβλητών **δεν παρουσιάζουν στατιστικά σημαντική συσχέτιση μεταξύ τους**:
  - pay_last3months και segment (το άθροισμα των πληρωμών τους τελευταίους 3 μήνες προς το ποσό που οφείλει ο πελάτης ως ποσοστό και το είδος του τραπεζικού προϊόντος που έχει λάβει)
  - months6_sumbucket και Tiresias (το άθροισμα των καθυστερήσεων των πληρωμών τους τελευταίους 6 μήνες -Δείκτης Παραβατικότητας και το αν είναι στον Τειρεσία)
  - customer_worst_bucket και income (η χειρότερη καθυστέρηση όλων των προϊόντων του πελάτη και το εισόδημά του)
  - last_payment_months και segment (οι μήνες που έχουν περάσει από την τελευταία πληρωμή του πελάτη και το είδος του τραπεζικού προϊόντος)
  - last_payment_months και income (οι μήνες που έχουν περάσει από την τελευταία πληρωμή του πελάτη και το εισόδημά του)
  - segment και income (το είδος του τραπεζικού προϊόντος και το εισόδημά του)
  - Tiresias και income (το αν είναι στον Τειρεσία και το εισόδημά του)

5.5 TwoStep Συσταδοποίηση: Αποτελέσματα αλγορίθμου-Προσαρμογή μοντέλου-Αριθμός ομάδων

Για την εφαρμογή της ανάλυσης κατά συστάδες με τη βοήθεια του στατιστικού πακέτου SPSS και τη μέθοδο TwoStep Cluster, προχωρήσαμε στην υλοποίηση της διαδικασίας της ανάλυσης κατά συστάδες.

Η TwoStep Cluster προσδιόρισε αυτόματα το βέλτιστο αριθμό των συστάδων με την επιλογή του κριτηρίου BIC (Bayesian Information Criterion). Όμοια αποτελέσματα επέδειξε η μέθοδος και με τη χρήση του κριτηρίου AIC (Akaike's Information Criterion). Ο αυτόματος προσδιορισμός των ομάδων παρουσιάζεται στον παρακάτω πίνακα αυτόματης συσταδοποίησης (Auto Clustering) (Πίνακας 5.16). Ύστερα από πολλές δοκιμές επιλέχθηκαν τρεις μεταβλητές που συμβάλλουν σημαντικά στο διαχωρισμό των συστάδων. Χρησιμοποιήθηκαν οι συνεχείς μεταβλητές months6_sumbucket και pay_last3months και η κατηγορική μεταβλητής customer_worst_bucket. Με μέγιστο αριθμό συστάδων να είναι 15 (default επιλογή του στατιστικού πακέτου), ο αλγόριθμος της TwoStep ορίζει τον αριθμό των συστάδων λαμβάνοντας υπόψη τη χαμηλή πληροφορία του κριτηρίου BIC και της υψηλότερης αναλογίας των μέτρων αποστάσεων. Εξαιτίας της χαμηλής τιμής BIC (16605,042) και της υψηλότερης αντίστοιχης τιμής της αναλογίας των μέτρων αποστάσεων (3,372) επιλέχτηκαν τέσσερις ομάδες.

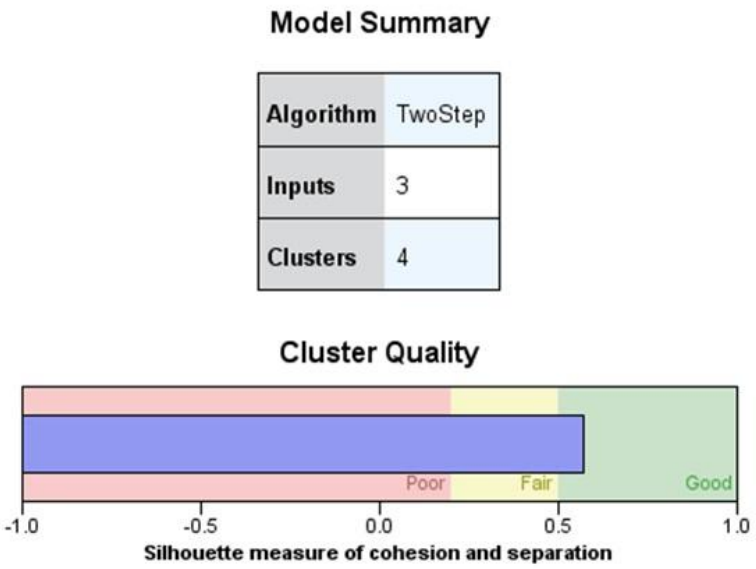
Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	75149,069			
2	42058,636	-33090,433	1,000	2,285
3	27609,983	-14448,653	,437	1,311
4	16605,042	-11004,942	,333	3,372
5	13383,384	-3221,658	,097	1,609
6	11404,063	-1979,321	,060	1,043
7	9508,612	-1895,451	,057	1,649
8	8383,013	-1125,599	,034	1,610
9	7706,398	-676,614	,020	1,230
10	7167,524	-538,874	,016	1,014
11	6636,639	-530,885	,016	1,088
12	6153,681	-482,958	,015	1,122
13	5729,802	-423,879	,013	1,282
14	5412,231	-317,571	,010	1,098
15	5128,410	-283,821	,009	1,082

Πίνακας 5.16

Η συσταδοποίηση των ομάδων ελέγχεται για την ποιότητά της ως προς τη συνοχή και το διαχωρισμό των συστάδων (Γράφημα 5.14) Με βάση τις αποστάσεις μεταξύ των

αντικειμένων, το μέτρο συνοχής και διαχωρισμού κυμαίνεται μεταξύ -1 και 1. Υποδεικνύει «κακή» ποιότητα αν είναι μικρότερο του 0,20, «δίκαιη» ποιότητα αν είναι μεταξύ του 0,20 και 0,50 και «καλή» ποιότητα άνω του 0,50. Στην περίπτωσή μας το αποτέλεσμα είναι μια ικανοποιητικά «καλή» ποιότητα συσταδοποίησης.



Γράφημα 5.14 Παρουσίαση αρχικών αποτελεσμάτων της συσταδοποίησης

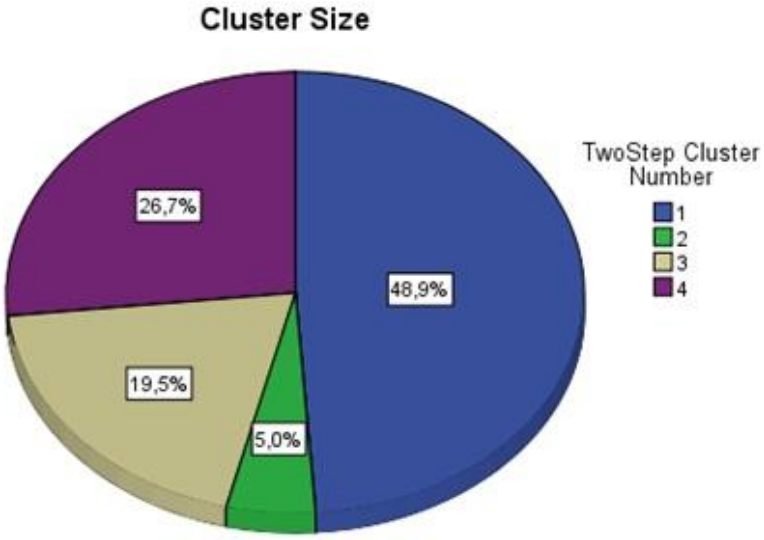
Πιο αναλυτικά, όπως παρατηρούμε στον πίνακα κατανομής των συστάδων (Cluster Distribution) (Πίνακας 5.17) από τις συνολικά 21956 παρατηρήσεις η πρώτη συστάδα διαθέτει 10732 παρατηρήσεις και το 48,9% των συνολικών παρατηρήσεων, η δεύτερη διαθέτει 1094 παρατηρήσεις και το 5% των συνολικών παρατηρήσεων, η τρίτη 4274 παρατηρήσεις και το 19,5% του συνολικού και η τέταρτη διαθέτει 5856 παρατηρήσεις και το 26,7% των συνολικών παρατηρήσεων.

Cluster Distribution

		N	% of Combined	% of Total
Cluster	1	10732	48,9%	48,9%
	2	1094	5,0%	5,0%
	3	4274	19,5%	19,5%
	4	5856	26,7%	26,7%
	Combined	21956	100,0%	100,0%
Total		21956		100,0%

Πίνακας 5.17

Η διαγραμματική παρουσίαση των αποτελεσμάτων της κατανομής των συστάδων παρουσιάζεται στο παρακάτω γράφημα (Γράφημα 5.15).



Γράφημα 5.15

5.5.1 Τα χαρακτηριστικά γνωρίσματα κάθε ομάδας και η διερεύνηση της σύστασής τους

Για τις συνεχείς μεταβλητές months6_sumbucket και pay_last3months ο πίνακας των κέντρων (Centroids) παρέχει τις μέσες τιμές και τις τυπικές αποκλίσεις στις τέσσερις συστάδες που δημιουργήθηκαν (Πινακας 5.18).

Παρατηρούμε πως στην πρώτη συστάδα ανήκουν οι παρατηρήσεις με το μικρότερο άθροισμα των τελευταίων 6 μηνών των δεικτών καθυστέρησης αποπληρωμής των χρεών, δηλαδή αυτή η ομάδα παρουσιάζει πελάτες με το χαμηλότερο δείκτη παραβατικότητας. Παράλληλα παρουσιάζουν χαμηλή δυνατότητα αποπληρωμής του χρέους σε τριμηνιαία βάση. Τον δεύτερο χαμηλότερο δείκτη παραβατικότητας παρουσιάζει η δεύτερη συστάδα ενώ έχει την υψηλότερη από όλες τις ομάδες δυνατότητα αποπληρωμής του χρέους. Αντίθετα η τρίτη συστάδα αν και έχει τον υψηλότερο δείκτη παραβατικότητας από όλες τις υπόλοιπες συστάδες, παρουσιάζει την χαμηλότερη δυνατότητα αποπληρωμής του χρέους. Τέλος, στην τέταρτη συστάδα ανήκουν οι παρατηρήσεις με υψηλό δείκτη παραβατικότητας και τη χαμηλότερη από όλες τις ομάδες δυνατότητα αποπληρωμής του χρέους.

Centroids

		months6_sumbucket		pay_last3months	
		Mean	Std. Deviation	Mean	Std. Deviation
Cluster	1	15,23	4,661	,0673	,07856
	2	16,94	6,005	,9751	,33731
	3	25,08	4,467	,0463	,07623
	4	20,47	5,084	,0603	,07532
	Combined	18,63	6,163	,1066	,22572

Πίνακας 5.18

Στη συνέχεια παρουσιάζουμε τον πίνακα συχνοτήτων (Πίνακας 5.19) της κατηγορικής μεταβλητής customer_worst_bucket ανά συστάδα που αφορά το χειρότερο δείκτη καθυστέρησης (bucket) που είχε ο πελάτης της τράπεζας σε οποιοδήποτε προϊόν.

Frequencies — customer_worst_bucket

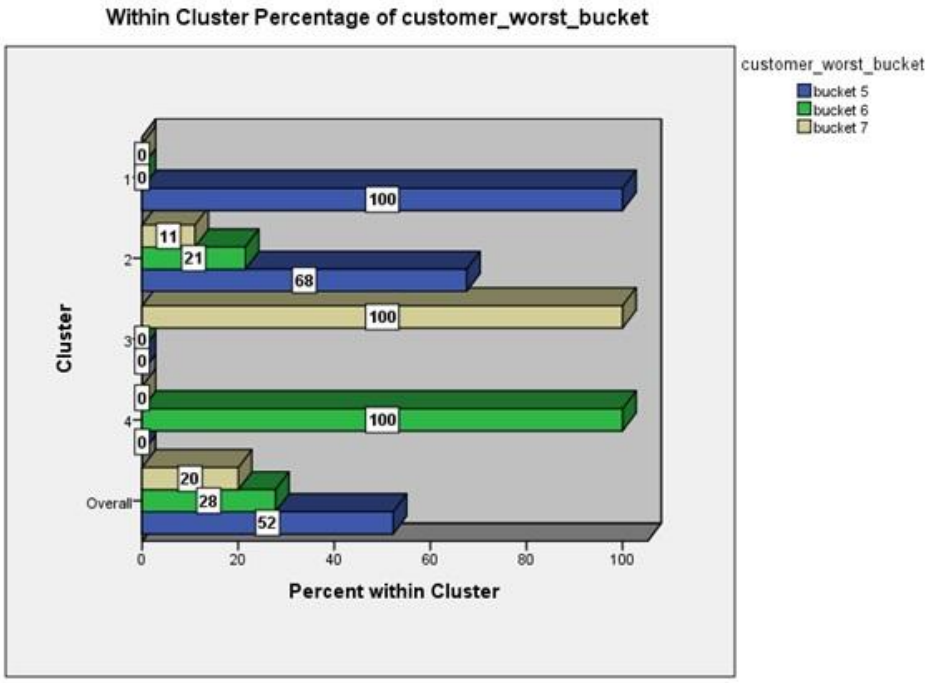
		bucket 5		bucket 6		bucket 7	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
Cluster	1	10732	93,6%	0	,0%	0	,0%
	2	739	6,4%	235	3,9%	120	2,7%
	3	0	,0%	0	,0%	4274	97,3%
	4	0	,0%	5856	96,1%	0	,0%
	Combined	11471	100,0%	6091	100,0%	4394	100,0%

Πίνακας 5.19

Από τον παραπάνω πίνακα συμπεραίνουμε πως:

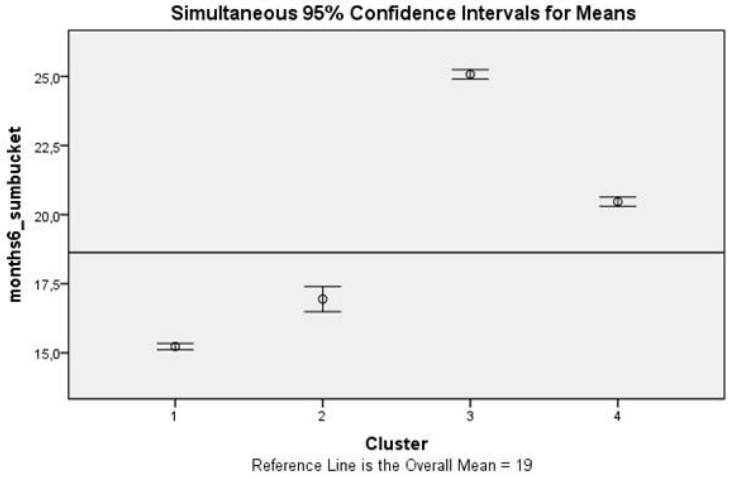
- Οι παρατηρήσεις με χειρότερο δείκτη καθυστέρησης 5 διαχωρίζονται σε 93,6% επί του συνόλου στην πρώτη συστάδα ενώ μόλις το 6,4% στη δεύτερη.
- Οι παρατηρήσεις με χειρότερο δείκτη καθυστέρησης 6 διαχωρίζονται σε 96,1% επί του συνόλου στην τέταρτη συστάδα ενώ μόλις το 3,9% στη δεύτερη.
- Οι παρατηρήσεις με χειρότερο δείκτη καθυστέρησης 7 διαχωρίζονται σε 97,3% επί του συνόλου στην τρίτη συστάδα ενώ μόλις το 2,7% στη δεύτερη.

Παρατηρείται λοιπόν από το διαχωρισμό της μεταβλητής πως οι παρατηρήσεις και των τριών χειρότερων δεικτών καθυστέρησης τοποθετούνται σχεδόν εξ' ολοκλήρου στις τρεις ομάδες, ενώ η δεύτερη συστάδα περιλαμβάνει ελάχιστο ποσοστό τοποθέτησης παρατηρήσεων και από τους τρεις δείκτες καθυστέρησης. Τα παραπάνω συμπεράσματα μπορούν να συναχθούν και από το επόμενο γράφημα (Γράφημα 5.16) :

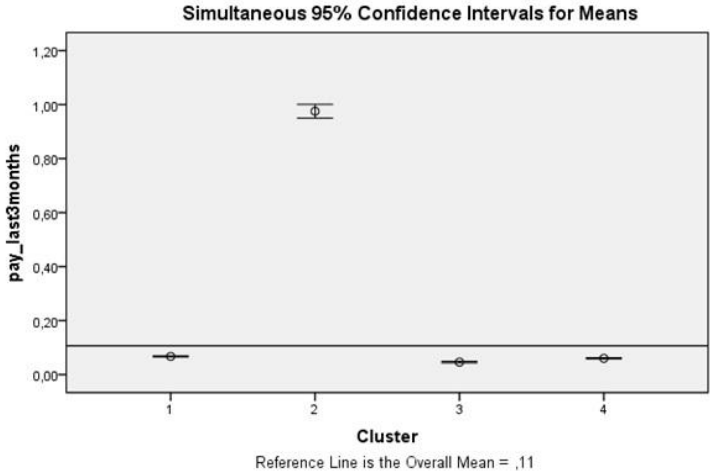


Γράφημα 5.16

Στα παρακάτω γραφήματα (Γράφημα 5.17) και (Γράφημα 5.18) απεικονίζονται διαγραμματικά οι μέσοι κάθε ομάδας με διάστημα εμπιστοσύνης 95% για τις συνεχείς μεταβλητές months6_sumbucket και pay_last3months αντίστοιχα. Παρατηρείται λοιπόν πως ο μέσος χειρότερος δείκτης καθυστέρησης είναι υψηλότερος για την τρίτη συστάδα, η τέταρτη συστάδα έχει μέσο χειρότερο δείκτη πάνω από το δειγματικό μέσο ενώ η δεύτερη και η πρώτη συστάδα είναι κάτω από το δειγματικό μέσο με την τελευταία να παρουσιάζει το χαμηλότερο μέσο χειρότερο δείκτη καθυστέρησης. Αντίστοιχα, παρατηρείται πως πάνω από το δειγματικό μέσο του δείκτη παραβατικότητας είναι ο μέσος δείκτης παραβατικότητας της δεύτερης ομάδας και αποτελεί και τον υψηλότερο μέσο δείκτη παραβατικότητας, ενώ οι υπόλοιπες τρεις συστάδες έχουν μέσους χειρότερους δείκτες παραβατικότητας κάτω από το δειγματικό μέσο, με το μικρότερο να ανήκει στην τρίτη ομάδα.



Γράφημα 5.17

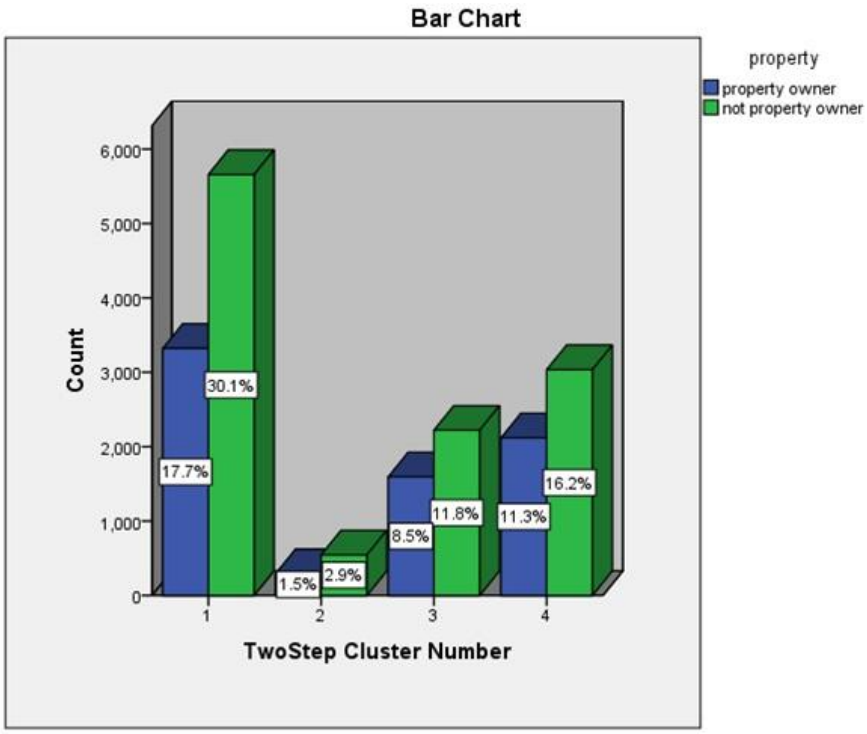


Γράφημα 5.18

Στη συνέχεια , θα παρουσιάσουμε τα χαρακτηριστικά γνωρίσματα κάθε ομάδας με σκοπό να μας βοηθήσουν στον προσδιορισμό της σύνθεσης των ομάδων και να συμβάλλουν στη διαμόρφωση των προφίλ των ομάδων.

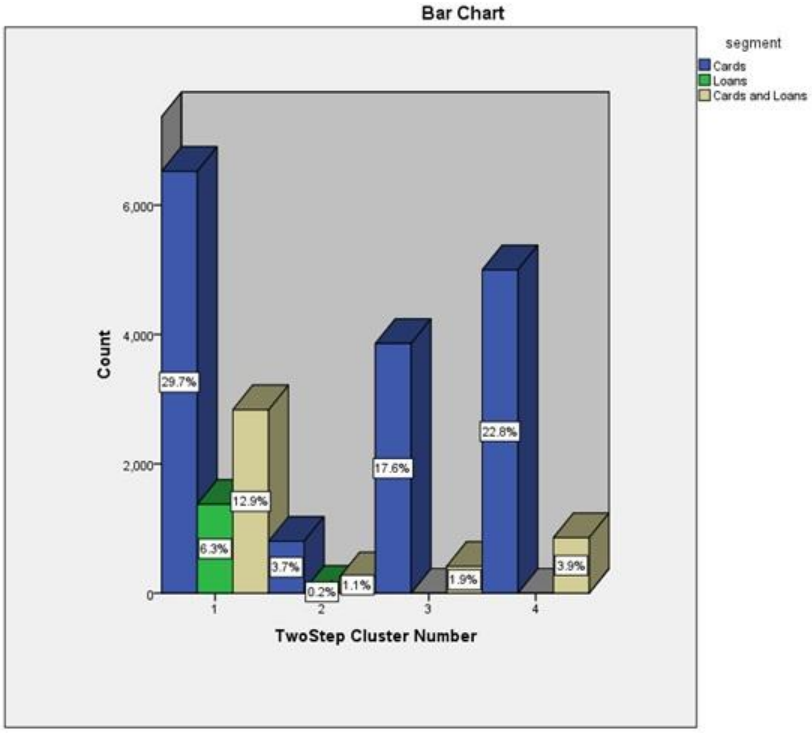
Στο επόμενο γράφημα (Γράφημα 5.19) παρατηρούμε τα ποσοστά της κατοχής περιουσιακών στοιχείων σε κάθε ομάδα. Πιο συγκεκριμένα, από το 39% επί του δειγματικού συνόλου που έχουν στην κατοχή τους περιουσιακά στοιχεία το μεγαλύτερο ποσοστό, 17,7%, ανήκει στην πρώτη συστάδα και τα υπόλοιπα ποσοστά διαχωρίζονται στις άλλες ομάδες με το μικρότερο ποσοστό κατοχής περιουσιακών στοιχείων να ανήκει στη δεύτερη ομάδα, μόλις 1,5%. Αντίθετα, από το 61% επί του δειγματικού συνόλου που δεν έχουν στην κατοχή τους περιουσιακά στοιχεία, παρατηρείται πως το 30,1%, που αποτελεί και το μεγαλύτερο ποσοστό, ανήκει στην πρώτη συστάδα και τα υπόλοιπα ποσοστά διαχωρίζονται στις

υπόλοιπες συστάδες με το μικρότερο ποσοστό και πάλι να ανήκει στην δεύτερη συστάδα, μόλις 2,9%.



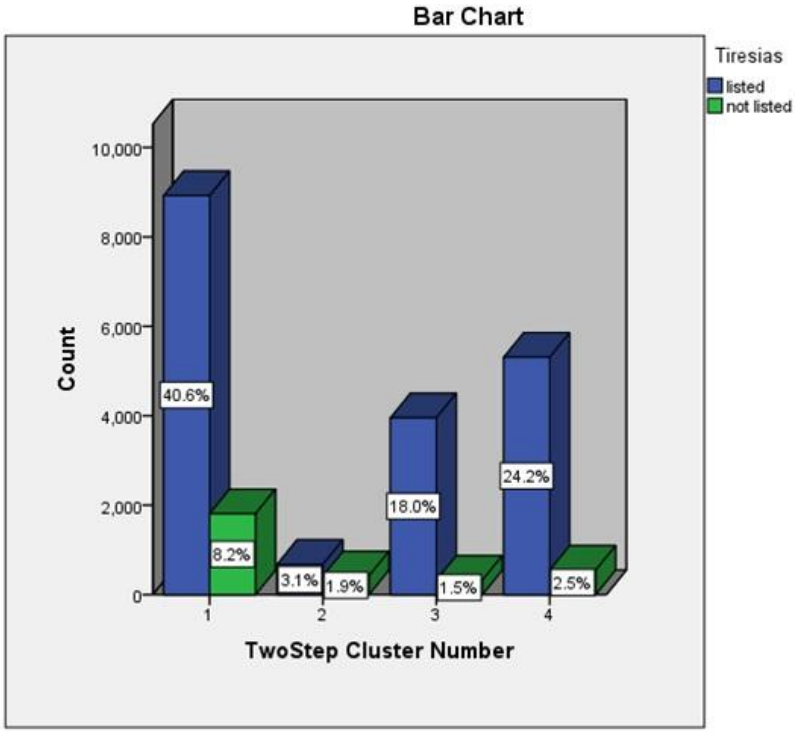
Γράφημα 5.19

Από το 73,7% επί του δειγματικού συνόλου που τους έχουν χορηγηθεί Πιστωτικές Κάρτες, το μεγαλύτερο ποσοστό, 29,7% ανήκει στην πρώτη ομάδα, ενώ το χαμηλότερο ποσοστό παρατηρήσεων με Πιστωτικές Κάρτες ανήκει στη δεύτερη ομάδα, μόλις 3,7%. Παράλληλα, από το 6,5% επί του δειγματικού συνόλου των παρατηρήσεων που τους έχουν χορηγηθεί Δάνεια, παρατηρείται πως έχει διαχωριστεί σχεδόν εξ' ολοκλήρου στην πρώτη συστάδα με ποσοστό 6,3% και το υπόλοιπο μόλις 0,2% στην δεύτερη συστάδα. Από το συνδυασμό παρατηρήσεων που τους έχουν χορηγηθεί Πιστωτικές Κάρτες και Δάνεια από το 19,85 επί του δειγματικού συνόλου, το 12,9% ανήκει στην πρώτη συστάδα και τα υπόλοιπα ποσοστά διαχωρίζονται στις υπόλοιπες συστάδες με το χαμηλότερο ποσοστό να ανήκει στην δεύτερη συστάδα, μόλις 1,1%. (Γράφημα 5.20)



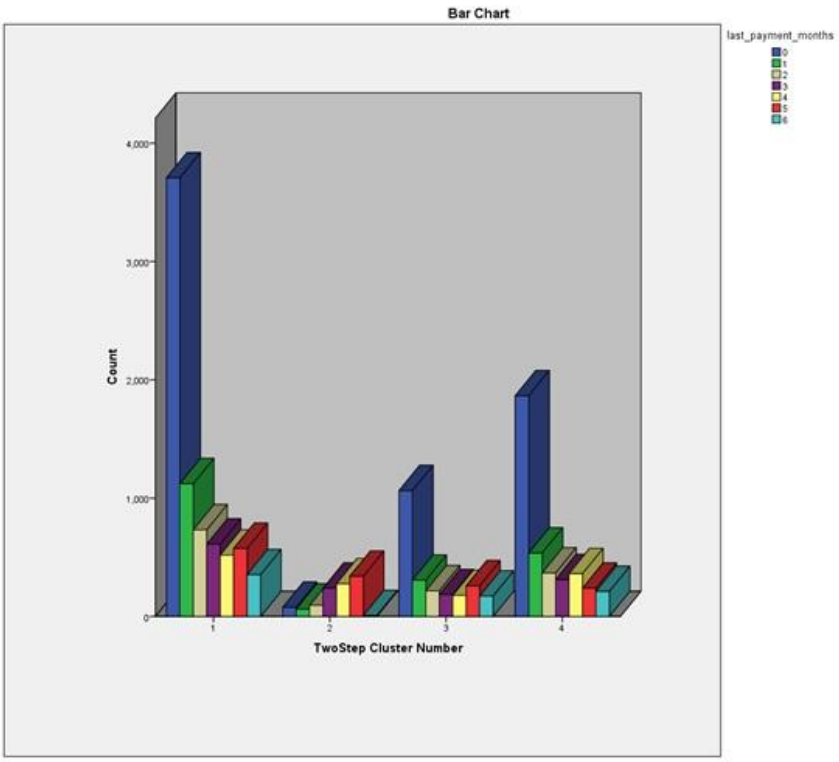
Γράφημα 5.20

Από το 85,9% επί του δειγματικού συνόλου που ανήκουν στο αρχείο δεδομένων οικονομικής συμπεριφοράς του Τειρεσία, το μεγαλύτερο ποσοστό, 40,6%, ανήκει στην πρώτη συστάδα ενώ παράλληλα η δεύτερη συστάδα παρουσιάζει το μικρότερο ποσοστό εγγραφής παρατηρήσεων στο αρχείο του Τειρεσία, μόλις 3,1%. Το υπόλοιπο 14,15% των παρατηρήσεων επί του δειγματικού συνόλου δεν είναι εγγεγραμμένοι στο αρχείο του Τειρεσία και το μικρότερο ποσοστό παρουσιάζεται στην τρίτη συστάδα, μόλις 1,5% (Γράφημα 5.21).



Γράφημα 5.21

Στο παρακάτω γράφημα (Γράφημα 5.22) παρουσιάζεται ο διαχωρισμός της μεταβλητής last_payment_months που αντιστοιχεί στον αριθμό των μηνών από την τελευταία πληρωμή. Πιο αναλυτικά, παρατηρούμε πως ένα αρκετά μεγάλο ποσοστό παρατηρήσεων που έχουν περάσει μηδέν μήνες από την τελευταία πληρωμή των χρεών τους ανήκει στην πρώτη συστάδα ενώ ένα αρκετά μικρότερο ποσοστό ανήκει στη δεύτερη συστάδα. Παράλληλα, το μεγαλύτερο ποσοστό σε σχέση με τις υπόλοιπες ομάδες, που από την τελευταία τους πληρωμή έχουν περάσει έξι μήνες, ανήκει στην πρώτη συστάδα ενώ ένα ελάχιστο ποσοστό ανήκει στην δεύτερη συστάδα.



Γράφημα 5.22

Από τον επόμενο πίνακα (Πίνακας) παρατηρούμε πως και οι τέσσερις ομάδες παρουσιάζουν ίδιο μέσο εισόδημα που κυμαίνεται γύρω στις 30.000 Ευρώ.

income

TwoStep Cluster Number	Mean	N	Std. Deviation
1	29004.8346	10189	23917.74555
2	30239.7619	789	24254.54637
3	29298.4621	4022	23966.36511
4	29208.7360	5507	23939.07907
Total	29164.6928	20507	23945.56698

Πίνακας 5.20

Από τις τέσσερις ομάδες που έχουν διαμορφωθεί το μικρότερο μέσο υπόλοιπο αντιστοιχεί στη δεύτερη συστάδα, ενώ οι υπόλοιπες τρεις συστάδες παρουσιάζουν μεγαλύτερο και περίπου ίδιο μέσο υπόλοιπο χρέους (Πίνακας 5.21).

balance

TwoStep Cluster Number	Mean	N	Std. Deviation
1	7495.2965	10732	7802.28821
2	5269.2736	1094	7090.15461
3	8036.2769	4274	8956.46389
4	7719.6079	5856	8709.05725
Total	7549.5163	21956	8271.09890

Πίνακας 5.21

Από τον παρακάτω πίνακα (Πίνακας 5.22) παρατηρούμε πως το μικρότερο μέσο ποσό καθυστέρησης αντιστοιχεί στην πρώτη συστάδα ενώ η τρίτη συστάδα παρουσιάζει σχεδόν το διπλάσιο μέσο ποσό καθυστέρησης.

amount_pastdue_onus

TwoStep Cluster Number	Mean	N	Std. Deviation
1	1204.5104	10732	1934.29034
2	1423.1361	1094	2502.96735
3	2281.1020	4274	3154.78797
4	1851.0811	5856	2629.02283
Total	1597.4256	21956	2471.03975

Πίνακας 5.22

Στον παρακάτω πίνακα (Πίνακας 5.23) παρατηρούμε τα ποσοστά διαφόρων χαρακτηριστικών σε κάθε ομάδα, όπως η κατοχή περιουσιακών στοιχείων, το είδος του τραπεζικού προϊόντος που τους έχει χορηγηθεί, πληροφορίες για την εγγραφή τους στο αρχείο του Τειρεσία και τους μήνες που έχουν περάσει από την τελευταία πληρωμή των χρεών τους.

Ως προς την κατοχή περιουσιακών στοιχείων, το μεγαλύτερο ποσοστό, 41,8%, αντιστοιχεί στην τρίτη συστάδα ενώ το μεγαλύτερο ποσοστό παρατηρήσεων που δεν έχουν στην κατοχή τους περιουσιακά στοιχεία αντιστοιχεί στη δεύτερη συστάδα με ποσοστό 65,2%.

Ως προς το τραπεζικό προϊόν, το μεγαλύτερο ποσοστό παρατηρήσεων που έχουν λάβει ως τραπεζικό προϊόν Πιστωτική Κάρτα αντιστοιχεί στην τρίτη συστάδα με ποσοστό 90,3%,

το μεγαλύτερο ποσοστό, μόλις 12,7%, που έχουν λάβει Δάνειο ανήκει στην πρώτη συστάδα ενώ το μεγαλύτερο ποσοστό που έχουν λάβει ως τραπεζικό προϊόν Πιστωτική Κάρτα και Δάνειο αντιστοιχεί στην πρώτη συστάδα με ποσοστό 26,4%.

Ως προς την εγγραφή τους στο αρχείο του Τειρεσία, παρουσιάζονται και στις τέσσερις συστάδες υψηλά ποσοστά στην περίπτωση που είναι εγγεγραμμένοι με υψηλότερο ποσοστό, 92,5%, να αντιστοιχεί στην τρίτη συστάδα ενώ παράλληλα το μεγαλύτερο ποσοστό που δεν είναι εγγεγραμμένοι στο αρχείο του Τειρεσία αντιστοιχεί στη δεύτερη συστάδα με ποσοστό 38,1%.

Τέλος, ως προς τους μήνες που έχουν περάσει από την τελευταία πληρωμή των χρεών τους, παρατηρούμε πως στην περίπτωση που έχουν περάσει 0 μήνες από την τελευταία πληρωμή το μεγαλύτερο ποσοστό αντιστοιχεί στην πρώτη συστάδα με 48,7% με δεύτερο μεγαλύτερο ποσοστό να αντιστοιχεί στην τέταρτη συστάδα με ποσοστό 48% ενώ στην τρίτη συστάδα αντιστοιχεί το 44,7%. Για την περίπτωση που έχει περάσει 1 μήνας από την τελευταία πληρωμή το μεγαλύτερο ποσοστό, 14,7%, αντιστοιχεί στην πρώτη συστάδα ενώ όταν έχουν περάσει 2 μήνες το μεγαλύτερο ποσοστό, 9,6%, αντιστοιχεί στην πρώτη ομάδα. Επίσης, όταν έχουν περάσει 3 μήνες από την τελευταία πληρωμή το μεγαλύτερο ποσοστό, 21,8% αντιστοιχεί στη δεύτερη συστάδα, ενώ στην περίπτωση που έχουν περάσει 4 μήνες το μεγαλύτερο ποσοστό, 25,2%, αντιστοιχεί στη δεύτερη συστάδα. Στην περίπτωση που έχουν περάσει 5 μήνες από την τελευταία πληρωμή το μεγαλύτερο ποσοστό, 31,1%, αντιστοιχεί στη δεύτερη συστάδα ενώ όταν έχουν περάσει 6 μήνες από την τελευταία πληρωμή των χρεών το μεγαλύτερο ποσοστό, 7,2%, αντιστοιχεί στην τρίτη συστάδα. Θα πρέπει να σημειωθεί, όπως παρατηρούμε και από τον Πίνακα 5.23, πως υπάρχουν αρκετές παρατηρήσεις (σε σύνολο 6977) για τις οποίες δεν υπάρχει διαθέσιμη καταγραφή των μηνών που έχουν περάσει από την τελευταία πληρωμή και η μόνη συστάδα για την οποία υπάρχει πλήρης διαθέσιμη πληροφορία είναι η δεύτερη συστάδα.

		ΣΥΣΤΑΔΕΣ ΑΠΟ TWO STEP ΣΥΣΤΑΔΟΠΟΙΗΣΗ			
ΜΕΤΑΒΛΗΤΗ	ΠΕΡΙΓΡΑΦΗ	ΣΥΣΤΑΔΑ 1	ΣΥΣΤΑΔΑ 2	ΣΥΣΤΑΔΑ 3	ΣΥΣΤΑΔΑ 4
Property	Property owner	37,0%	34,8%	41,8%	41,1%
	Not propertyowner	63,0%	65,2%	58,2%	58,9%
Missing values	Total 3173	1752	258	459	704
Segment	Cards	60,8%	73,3%	90,3%	85,4%
	Loans	12,8%	3,7%	,0%	,0%
	Cards and Loans	26,4%	22,9%	9,7%	14,6%
Missing values	Total -	-	-	-	-
Tiresias	Listed	83,2%	61,9%	92,5%	90,7%
	Not Listed	16,8%	38,1%	7,5%	9,3%
Missing values	Total -	-	-	-	-
Last payment months	0 months since last payment	48,7%	7,0%	44,7%	48,0%
	1 months since last payment	14,7%	5,7%	12,9%	13,8%
	2 months since last payment	9,6%	8,6%	9,1%	9,4%
	3 months since last payment	8,0%	21,8%	7,8%	8,0%
	4 months since last payment	6,8%	25,2%	7,5%	9,3%
	5 months since last payment	7,5%	31,1%	10,8%	6,1%
	6 months since last payment	4,6%	,6%	7,2%	5,4%
Missing values	Total 6977	3114	-	1986	1972

Πίνακας 5.23

5.5.2 Το προφίλ των συστάδων

<u>Συστάδα 1</u>

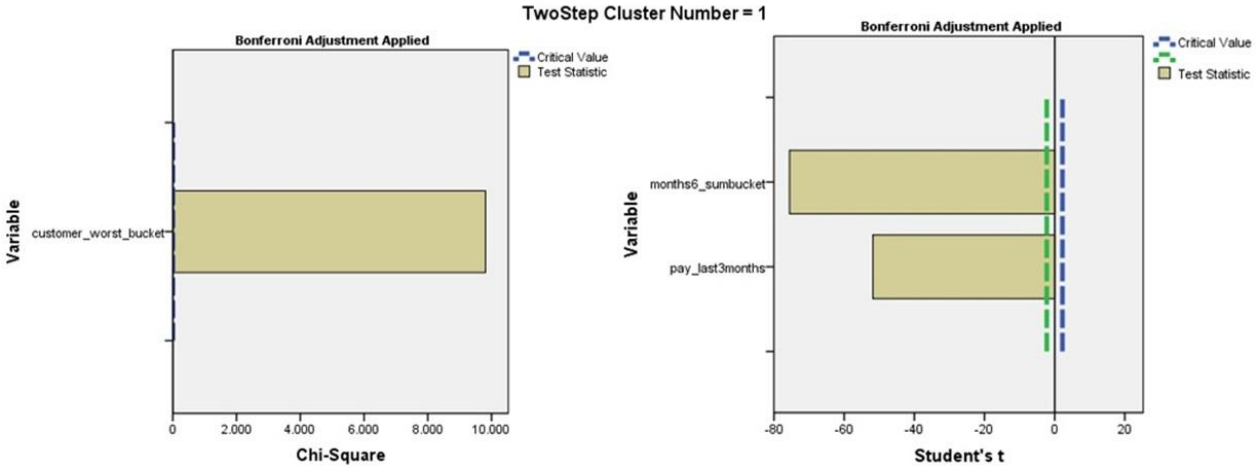
Μέγεθος συστάδας: 48,9%

Στη πρώτη συστάδα παρατηρούμε πως οι πελάτες χαρακτηρίζονται κυρίως για το παρελθόν τους. Έχουν δείκτη 5 ως το χειρότερο δείκτη που είχαν στο παρελθόν σε τραπεζικό προϊόν, χαμηλότερο από το συνολικό μέσο άθροισμα των δεικτών καθυστερήσεων των

πληρωμών τους τελευταίους 6 μήνες και χαμηλότερο από το συνολικό μέσο ποσοστό αποπληρωμής του χρέους σε τριμηνιαία βάση (Γράφημα 5.23).

Το 48,7% των πελατών παρατηρείται πως πληρώνει κάθε μήνα τις υποχρεώσεις του προς την τράπεζα, έχοντας ως τραπεζικό προϊόν κυρίως Πιστωτικές Κάρτες. Με μέσο εισόδημα περίπου 29.000 € έχουν μέσο υπόλοιπο χρέους περίπου 5.200 € και το μέσο ποσό καθυστέρησης αποπληρωμής ανέρχεται περίπου στα 1200 €. Το 63% των πελατών δεν έχει στην κατοχή του περιουσιακά στοιχεία και παρατηρείται πως το 83% των πελατών είναι εγγεγραμμένοι στο αρχείο δεδομένων οικονομικής συμπεριφοράς του Τειρεσία.

Συνεπώς, στην πρώτη συστάδα αντιστοιχούν πελάτες που χαρακτηρίζονται για την «καλή εικόνα» του παρελθόντος στη συγκεκριμένη τράπεζα και «προσπαθούν να τη διατηρήσουν» πληρώνοντας κάθε μήνα ένα μικρό ποσοστό του χρέους.



Γράφημα 5.23

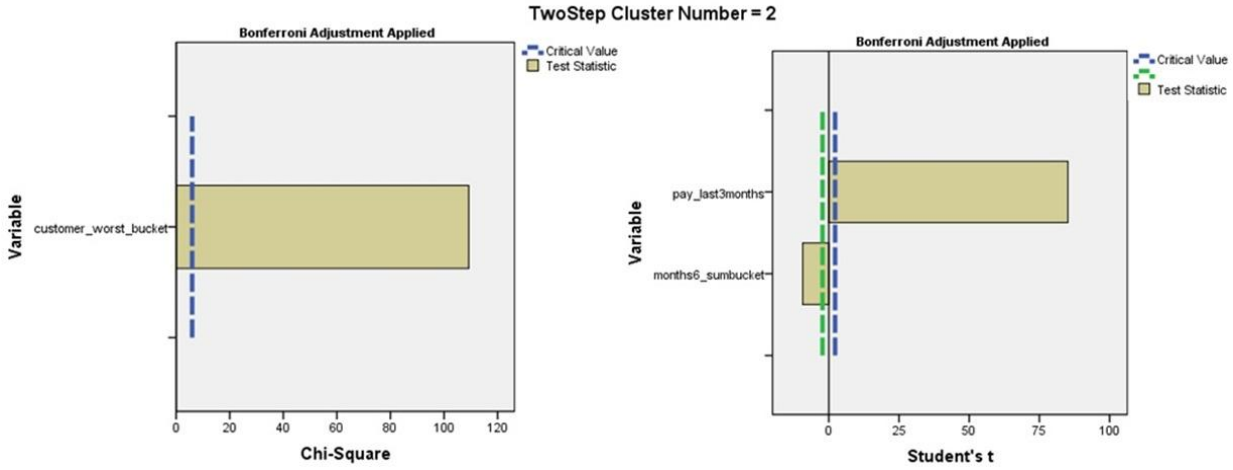
Συστάδα 2

Μέγεθος συστάδας: 5%

Η δεύτερη συστάδα αποτελείται από τους λιγότερους πελάτες σε σχέση με τις υπόλοιπες συστάδες. Χαρακτηρίζονται κυρίως για την αποπληρωμή των χρεών τους προς την τράπεζα.. Έχουν δείκτη 5 ως το χειρότερο δείκτη που είχαν στο παρελθόν σε τραπεζικό προϊόν, χαμηλότερο από το συνολικό μέσο άθροισμα των δεικτών καθυστερήσεων των πληρωμών τους τελευταίους 6 μήνες και η μέση τιμή του ποσοστού αποπληρωμής του χρέους σε τριμηνιαία βάση είναι υψηλότερη του συνολικού μέσου (Γράφημα 5.24).

Το 31,1% των πελατών παρατηρείται πως πλήρωσε τις υποχρεώσεις του προς την τράπεζα πριν 5 μήνες έχοντας ως τραπεζικό προϊόν κυρίως Πιστωτικές Κάρτες. Με μέσο εισόδημα περίπου 30.200 € έχουν μέσο υπόλοιπο χρέους περίπου 7.500 € και το μέσο ποσό καθυστέρησης αποπληρωμής ανέρχεται περίπου στα 1400 €. Το 65,2% των πελατών δεν έχει στην κατοχή του περιουσιακά στοιχεία και παρατηρείται πως το 61,9% των πελατών είναι εγγεγραμμένοι στο αρχείο δεδομένων οικονομικής συμπεριφοράς του Τειρεσία.

Συνεπώς, η δεύτερη συστάδα προσδιορίζεται για το μεγάλο ποσοστό αποπληρωμής των χρεών τους στην τράπεζα το τελευταίο τρίμηνο. Φαίνεται πως οι πελάτες της δεύτερης συστάδας προσπαθούν να ανακάμψουν από το «κακό» παρελθόν τους.



Γράφημα 5.24

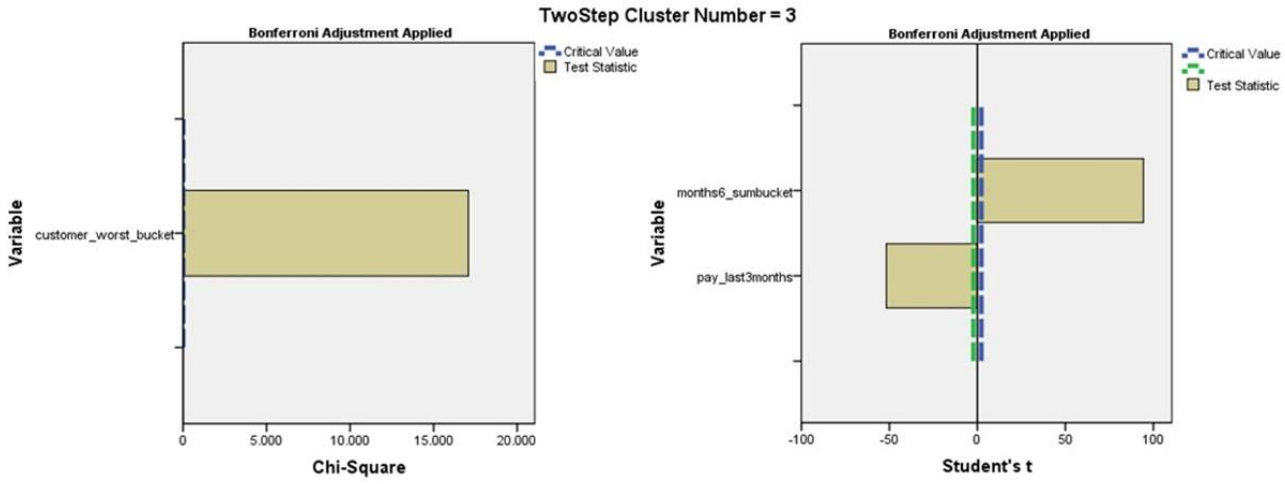
Συστάδα 3

Μέγεθος συστάδας: 19,5%

Οι πελάτες της τρίτης συστάδας προσδιορίζονται ως η «πιο παραβατική» συστάδα σε σχέση με τις υπόλοιπες. Έχουν δείκτη 7 ως το χειρότερο δείκτη που είχαν στο παρελθόν σε τραπεζικό προϊόν, που αντικατοπτρίζει το «κακό» τους παρελθόν και την πιθανή εξαίρεση που έλαβαν από την τράπεζα για να μην ενταχθούν στη διαδικασία του write off. Έχουν υψηλότερη μέση τιμή αθροίσματος των δεικτών καθυστερήσεων των πληρωμών τους τελευταίους 6 μήνες από το συνολικό μέσο και η μέση τιμή του ποσοστού αποπληρωμής του χρέους σε τριμηνιαία βάση είναι χαμηλότερη του συνολικού μέσου (Γράφημα 5.25).

Το 44,7% των πελατών παρατηρείται πως πληρώνει κάθε μήνα έχοντας ως τραπεζικό προϊόν σχεδόν εξ' ολοκλήρου Πιστωτικές Κάρτες με ποσοστό 90,3%. Με μέσο εισόδημα περίπου 29.000 € έχουν μέσο υπόλοιπο χρέους περίπου 8.000 € και το μεγαλύτερο σε σχέση με τις υπόλοιπες συστάδες μέσο ποσό καθυστέρησης αποπληρωμής που ανέρχεται περίπου στα 2200 €. Το 58,2% των πελατών δεν έχει στην κατοχή του περιουσιακά στοιχεία και παρατηρείται πως σχεδόν εξ' ολοκλήρου με ποσοστό το 92,5% των πελατών είναι εγγεγραμμένοι στο αρχείο δεδομένων οικονομικής συμπεριφοράς του Τειρεσία.

Συνεπώς, η τρίτη συστάδα χαρακτηρίζεται για την υψηλή «παραβατικότητα» και του παρελθόντος και του παρόντος. Φαίνεται πως με την μηνιαία αποπληρωμή των χρεών τους καταβάλλουν προσπάθειες για να μην γίνουν write off .



Γράφημα 5.25

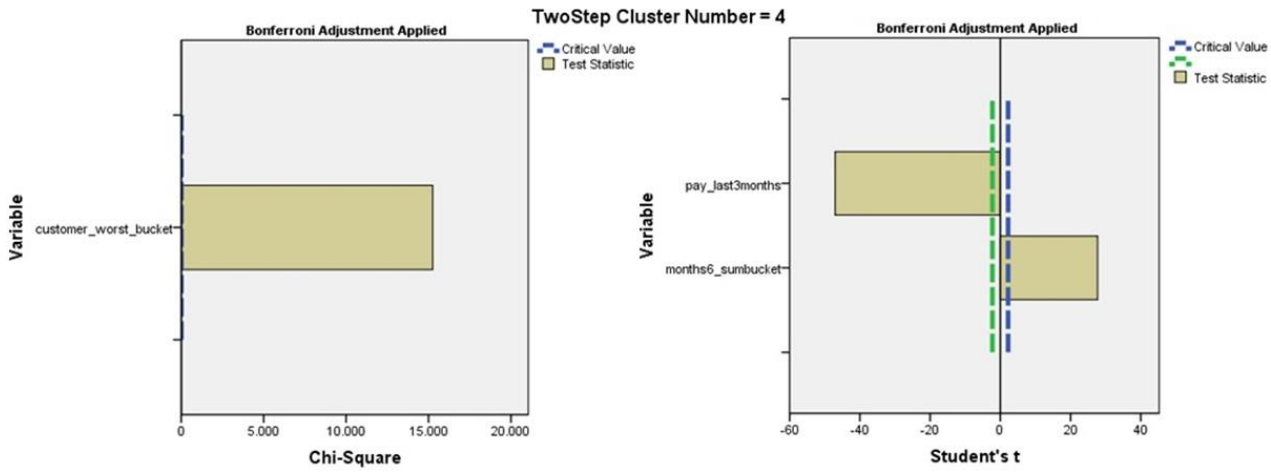
Συστάδα 4

Μέγεθος συστάδας: 26,7%

Οι πελάτες της τέταρτης συστάδας έχουν δείκτη 6 ως το χειρότερο δείκτη που είχαν στο παρελθόν σε τραπεζικό προϊόν, υψηλότερη μέση τιμή αθροίσματος των δεικτών καθυστερήσεων των πληρωμών τους τελευταίους 6 μήνες από το συνολικό μέσο και η μέση τιμή του ποσοστού αποπληρωμής του χρέους τους σε τριμηνιαία βάση είναι χαμηλότερη του συνολικού μέσου (Γράφημα 5.26).

Το 48% των πελατών παρατηρείται πως πληρώνει κάθε μήνα έχοντας ως τραπεζικό προϊόν σχεδόν εξ' ολοκλήρου Πιστωτικές Κάρτες με ποσοστό 85,4% . Με μέσο εισόδημα περίπου 29.000 € έχουν μέσο υπόλοιπο χρέους περίπου 7.700 € και το μεγαλύτερο σε σχέση με τις υπόλοιπες συστάδες μέσο ποσό καθυστέρησης αποπληρωμής που ανέρχεται περίπου στα 1850 €. Το 58,9% των πελατών δεν έχει στην κατοχή του περιουσιακά στοιχεία και παρατηρείται πως σχεδόν εξ' ολοκλήρου με ποσοστό το 90,7% των πελατών είναι εγγεγραμμένοι στο αρχείο δεδομένων οικονομικής συμπεριφοράς του Τειρεσία.

Συνεπώς, οι πελάτες της τέταρτης συστάδας αν και παρουσιάζουν παρόμοια χαρακτηριστικά με τους πελάτες της τρίτης συστάδας διαφέρουν ως προς την παρελθοντική συμπεριφορά τους. Αποτελούν σίγουρα λοιπόν μια «παραβατική» συστάδα που αποπληρώνει ένα μικρό ποσοστό των χρεών της σε τριμηνιαία βάση.



Γράφημα 5.26

5.5.3 Αποτελέσματα

Θα εξετάσουμε το προφίλ των πελατών με βάση τη μεταβλητή write_off_status και θα παρατηρήσουμε πως κατανέμονται οι πελάτες κάθε συστάδας ανάλογα με την κατάσταση της πιθανής διαγραφής τους. Στον Πίνακα 5.24 παρατηρούμε ποιοι πελάτες κάθε ομάδας θα γίνουν written off ή όχι τρεις μήνες μετά από τη στιγμή της παρατήρησης.

Πιο αναλυτικά, από τους 10732 πελάτες της πρώτης συστάδας που χαρακτηρίζονται για την «καλή εικόνα του παρελθόντος» τους, οι 4828 ενδέχεται να γίνουν written off με ποσοστό 45%. Από τους 1094 πελάτες της δεύτερης συστάδας που χαρακτηρίζονται για το μεγάλο ποσοστό αποπληρωμής των χρεών τους στην τράπεζα το τελευταίο τρίμηνο, οι 867 ενδέχεται να γίνουν written off με ποσοστό 79,3%. Από τους 4274 πελάτες της τρίτης συστάδας, οι 559 ενδέχεται να γίνουν written off με ποσοστό 13,1%. Από τους 5856 πελάτες της τέταρτης συστάδας, οι 1905 ενδέχεται να γίνουν written off με ποσοστό 32,5%.

write_off_status * TwoStep Cluster Number Crosstabulation

			TwoStep Cluster Number				Total
			1	2	3	4	
write_off_status	write off	Count	4828	867	559	1905	8159
		% within TwoStep Cluster Number	45.0%	79.3%	13.1%	32.5%	37.2%
	not write off	Count	5904	227	3715	3951	13797
		% within TwoStep Cluster Number	55.0%	20.7%	86.9%	67.5%	62.8%
Total		Count	10732	1094	4274	5856	21956
		% within TwoStep Cluster Number	100.0%	100.0%	100.0%	100.0%	100.0%

Πίνακας 5.24

6. Συμπεράσματα

Από το δείγμα των 21956 πελατών της τράπεζας που είχαμε στη διάθεση μας εφαρμόσαμε TwoStep ανάλυση κατά συστάδες και καταλήξαμε σε 4 συστάδες. Διερευνώντας τις συσταδοποιήσεις της μεθόδου TwoStep διαμορφώθηκαν τα προφίλ των τεσσάρων συστάδων:

- Οι πελάτες της πρώτης συστάδας με ποσοστό 48,9% χαρακτηρίζονται για την «καλή εικόνα» του πελάτη που είχαν στο παρελθόν, που στην πλειοψηφία τους πληρώνουν κάθε μήνα τις υποχρεώσεις του προς την τράπεζα κυρίως για Πιστωτικές Κάρτες. Όμως επειδή η αποπληρωμή του χρέους σε τριμηνιαία βάση είναι αρκετά χαμηλή, το 45% των πελατών της ομάδας τους επόμενους 3 μήνες θα γίνει written off.

- Οι πελάτες της δεύτερης συστάδας με ποσοστό μόλις 5% χαρακτηρίζονται για το μεγάλο ποσοστό αποπληρωμής των χρεών τους στην τράπεζα το τελευταίο τρίμηνο κυρίως για Πιστωτικές Κάρτες. Αλλά, η τελευταία εμφάνιση πληρωμής των χρεών προς την τράπεζα σε πλειοψηφία είναι 5 μήνες πριν από τη στιγμή της παρατήρησης, γεγονός που δικαιολογεί το ποσοστό του 79,3% που θα γίνει written off.

- Οι πελάτες της τρίτης συστάδας με ποσοστό 19,5% χαρακτηρίζονται για την «κακή εικόνα» του παρελθόντος αλλά και του παρόντος. Αν και στην πλειοψηφία τους πληρώνουν κάθε μήνα τις υποχρεώσεις του προς την τράπεζα για Πιστωτικές Κάρτες, η αποπληρωμή του χρέους τους προς την τράπεζα σε τριμηνιαία βάση είναι αρκετά χαμηλή. Από τους πελάτες της τρίτης συστάδας θα γίνει written off μόνο το 13,1%, ένα ποσοστό που αντιβαίνει στο προφίλ της ομάδας.

- Οι πελάτες της τέταρτης συστάδας με ποσοστό 26,7% αποτελούν μια «παραβατική» συστάδα σε σχέση με τις υποχρεώσεις τους προς την τράπεζα. Η πλειοψηφία των πελατών πληρώνει κάθε μήνα έχοντας ως τραπεζικό προϊόν κυρίως Πιστωτικές Κάρτες. Με χαμηλή όμως την αποπληρωμή του χρέους σε τριμηνιαία βάση το 32,5% θα γίνει written off.

Από τα αποτελέσματα της ανάλυσης κατά συστάδες και τις διαθέσιμες πληροφορίες για τους πελάτες της τράπεζας συμπεραίνουμε πως η τράπεζα θα πρέπει να επικεντρωθεί στους πελάτες της πρώτης συστάδας. Αποτελούν το μεγαλύτερο ποσοστό πελατών σε σχέση με τις υπόλοιπες συστάδες και δεν έχουν ξεπεράσει ποτέ στο δείκτη καθυστέρησης αποπληρωμής των χρεών τον δείκτη 5. Αν και παρατηρείται πως πληρώνουν αρκετά λιγότερα από όσα απαιτούνται κάθε μήνα, στην πλειοψηφία τους πληρώνουν κάθε μήνα. Συνεπώς, με βάση όλα τα προηγούμενα στοιχεία, η τράπεζα θα αποφασίσει πως ένας διακανονισμός με τους πελάτες της πρώτης συστάδας ή μια ενδεχόμενη πρόταση θα έχει ως σίγουρο αποτέλεσμα την αποπληρωμή των χρεών τους και άμεσα την αύξηση των κερδών της τράπεζας.

Παρόλα αυτά, θα ήταν προτιμότερο η τράπεζα πριν προβεί σε οποιαδήποτε προσφορά προς τους πελάτες της πρώτης συστάδας να επανεξετάσει τα δεδομένα με σκοπό τη δημιουργία ενός ολοκληρωμένου πιστοληπτικού προφίλ όλων των πελατών της. Μη έχοντας συγκεντρωμένες όλες τις πληροφορίες που συνθέτουν εξ' ολοκλήρου το προφίλ των πελατών δεν μπορούμε να αποκλείσουμε το γεγονός πως το αποτέλεσμα της ανάλυσης κατά συστάδες ίσως να μην ανταποκρίνεται στο έπακρο στην ουσιαστική κατάσταση των πελατών. Προτείνεται λοιπόν, να συμπληρωθούν στις ήδη υπάρχουσες μεταβλητές στοιχεία όπως οι πηγές των εισοδημάτων των πελατών από τις οποίες θα αντιλαμβάνονταν την οικονομική δυνατότητα των πελατών να ανταποκριθούν στις υποχρεώσεις τους. Ακόμα, τα χρέη προς άλλες τράπεζες ή και το Δημόσιο και οι λόγοι που τους χορηγήθηκαν τα τραπεζικά προϊόντα θα αποτελούσαν σημαντικές προσθήκες στην ολοκλήρωση του προφίλ και στη λήψη αποφάσεων που πάνω από όλα αφορούν το κέρδος της τράπεζας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΕΛΛΗΝΙΚΗ

[1] Καρλής, Δ. (2005). *Πολυμεταβλητή Στατιστική Ανάλυση.* Αθήνα: Εκδόσεις Σταμούλης Α.

[2] Κουκουβίνος, Χ. (2005). Γραμμικά Μοντέλα και Σχεδιασμοί. Αθήνα

[3] Μπεχράκης, Θ. Ε. (1999). *Ανάλυση Δεδομένων Μέθοδοι και Εφαρμογές.* Αθήνα: Εκδόσεις Νέα Σύνορα, Εκδοτικός Οργανισμός Λιβάνη.

[4] Πανάρετος, Ι., & Ξεκαλάκη, Ε. (1995). *Εισαγωγή στην Πολυμεταβλητή Στατιστική Ανάλυση.* Αθήνα.

[5] Σαπουντζόγλου, Γ. Γ., & Πεντότης, Χ. Ν. (2009). *Τραπεζική Οικονομική* (Τόμ. Α'). Αθήνα: Εκδόσεις Γ. Μπένου.

[6] Σιάρδος, Γ. Κ. (1999). *Μέθοδοι πολυμεταβλητής στατιστικής ανάλυσης - Με την επίλυση ασκήσεων μέσω του προγράμματος SPSS: Διερεύνηση εξάρτησης μεταξύ μεταβλητών.* Θεσσαλονίκη: Εκδόσεις Ζήτη.

[7] Σιώμκος, Γ., & Βασιλικοπούλου, Α. (2005). *Εφαρμογή Μεθόδων Ανάλυσης στην Έρευνα Αγοράς.* Αθήνα: Εκδόσεις Σταμούλης Α.

ΞΕΝΟΓΛΩΣΣΗ

[8] Altman, E. I., Avery, R. B., Eisenbeis, R. A., & Sinkey, J. F. (1981). *Application of Classification Techniques in Business,* Banking and Finance. Jai Press.

[9] Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.). New Jersey: John Wiley & Sons Inc.

[10] Bacher, J., Wenzig, K., & Vogler, M. (2004). SPSS TwoStep cluster: A first evaluation. *Lehrstuhl für Soziologie*.

[11] Bardos, M. (1998). Detecting the risk of company failure at the Banque de France. *Journal of Banking & Finance* , 22 (10), 1405-1419.

[12] Bartholomew, J. D., Moustaki, I., & Galbraith, I. J. (2002). *The analysis and interpretation of multivariate data for social scientists.* London: Champman and Hall.

[13] Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. *In Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining* (σσ. 263-268). ACM.

[14] Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th ed.). West Sussex: John Wiley & Sons, Ltd.

[15] Giudici, P. (2003). *Applied Data Mining : Statistical Methods for Business and Industry.* West Sussex: John Wiley & Sons Ltd.

[16] Hardle, W., & Simar, L. (2003). *Applied Multivariate Statistical Analysis.* Berlin: Springer.

[17] Jain, A. K. (2009). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters , 31* (8), 651-666.

[18] Jajuga, K., Sokolowski, A., & Bock, H.-H. (2002). *Classification, Clustering, and Data Analysis: Recent Advances and Applications.* Berlin: Springer.

[19] Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley & Sons,Inc.

[20] Mooi, E., & Sarstedt, M. (2011). *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics.* Berlin Heidelber: Springer-Verlag .

[21] Myatt, G. J. (2007). *Making Sence of Data: A Practical Guide to Exploratory Data Analysis and Data Mining.* New Jersey: John Wiley & Sons.

[22] Park, H.-S., & Baik, D.-K. (2006). A study for control of client value using cluster analysis. *Journal of Network and Computer Applications , 29* (4), 262-276.

[23] Şchiopu, D. (2010). Applying TwoStep cluster analysis for identifying bank customers' profile. *Buletinul , 62*, 66-75.

[24] Shih, M. Y., Jheng, J. W., & Lai, L. F. (2010). A two-step method for clustering mixed categroical and numeric data. *Tamkang Journal of Science and Engineering , 13* (1), 11-19.

[25] *SPSS 15.0 Algorithms.* (2006). SPSS, Inc.

[26] Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics.* Boston: Pearson Education,Inc.

[27] Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Addison-Wesley.

[28] Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting , 16* (2), 149-172.

[29] Timm, N. H. (2002). *Applied Multivariate Analysis.* New York: Springer.

[30] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function . *Journal of the American Statistical Association , 58* (301), 236-244.

[31] Xu, R., & Wunsch, D. C. (2009). *Clustering.* New Jersey: John Wiley & Sons,Inc.

[32] Yap, B. W., Ong, S. H., & Husain, N. H. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications* , pp. 13274-13283.

[33] Zadeh, R. B., Faraahi, A., & Mastali, A. (2011, January). Profiling bank customers behaviour using cluster analysis for profitability. *In International Conference on Industrial Engineering and Operations Management Kuala Lumpur,Malaysia.*