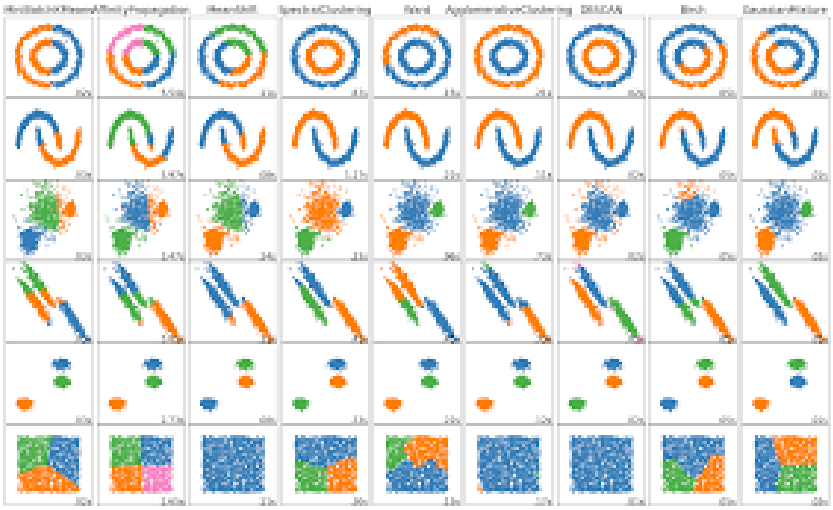
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

Μεταπτυχιακό Πρόγραμμα Σπουδών στην Εφαρμοσμένη Στατιστική

ΤΕΧΝΙΚΕΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ ΜΕ ΕΦΑΡΜΟΓΗ ΣΤΟΝ ΧΩΡΙΣΜΟ ΠΕΛΑΤΩΝ ΣΕ ΟΜΑΔΕΣ



Διατριβή

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

ΜΠΙΡΜΠΑΣ ΓΕΩΡΓΙΟΣ

Ιούλιος 2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

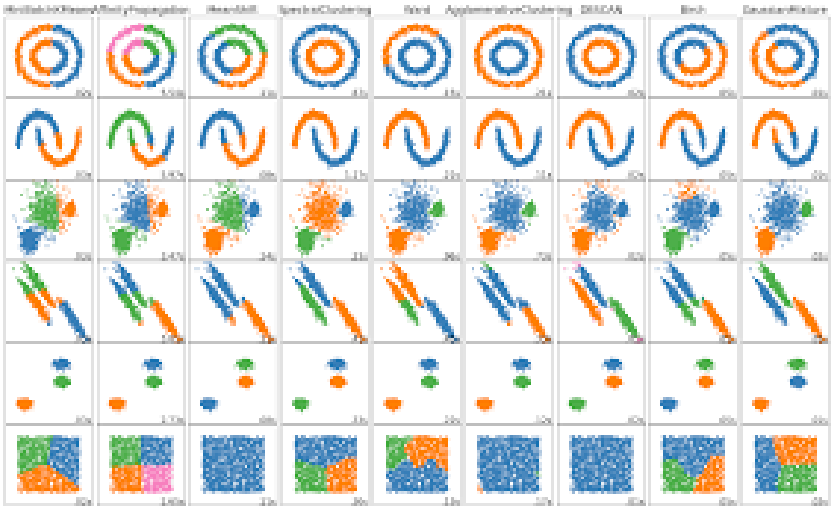
UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

Postgraduate Program in Applied Statistics

Clustering algorithms with application in customer segmentation



Thesis

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

BIRMPAS GEORGIOS

July 2020

Σελίδα έγκρισής της ΔΕ

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών.

Τα μέλη της Επιτροπής ήσαν:

- Κούτρας Μάρκος (Επιβλέπων)

- Μπούτσικας Μιχαήλ

- Ευαγγελάρας Χαράλαμπος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

Αφιέρωση - Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά του γονείς μου μου με βοήθησαν και με στήριξαν καθ' όλη την πορεία των σπουδών μου και ήταν αρωγοί της προσπάθειάς μου όλα αυτά τα χρόνια. Ένα ξεχωριστό ευχαριστώ στον αδερφό μου, ο οποίος με βοήθησε σε πολλές δύσκολες καταστάσεις ψυχολογικά αλλά ακόμα και με τις γνώσεις του καθώς ασχολείται με το ίδιο επιστημονικό αντικείμενο. Επίσης θα ήθελα να ευχαριστήσω και τον καθηγητή μου, κύριο Μάρκο Κούτρα, ο οποίος είναι επιβλέπων της παρούσας διπλωματικής αλλά και εξαίρετος επιστήμονας που με τις γνώσεις του με βοήθησε να χτίσω τις απαραίτητες βάσεις για την είσοδο μου στην αγορά εργασίας.

Περίληψη

Σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι να εφαρμοστούν και να αξιολογηθούν τεχνικές συσταδοποίησης σε ένα πραγματικό σύνολο δεδομένων συναλλαγών ενός ομίλου σούπερ μάρκετ με σκοπό την ομαδοποίηση πελατών. Η αξιολόγηση θα γίνει με βάση τα αποτελέσματα των τεχνικών που θα εφαρμοστούν αλλά και μέσω βιβλιογραφικής ανασκόπησης των συγκεκριμένων τεχνικών.

Ως σημείο αναφοράς θα χρησιμοποιηθεί η μέθοδος *k*-means ενώ θα εφαρμοστούν και οι τεχνικές ιεραρχικής ομαδοποίησης, ο DBSCAN και ο fuzzy *C*-means και στη συνέχεια θα γίνει σύγκριση των αποτελεσμάτων τους με αυτά της *k*-means. Θα αναφερθούν και θα προταθούν λύσεις για τυχόν προβλήματα ή αστοχίες κάθε μίας από τις παραπάνω μεθόδους, όπως για παράδειγμα ευαισθησίες που δεν υπάρχουν σε άλλες μεθόδους, και θα επισημανθούν τα δυνατά σημεία τους, όπως για παράδειγμα η καλή εφαρμογή σε πολυδιάστατα δεδομένα.

Ο απώτερος σκοπός της διαδικασίας της ομαδοποίησης, δηλαδή του χωρισμού των πελατών σε ομάδες με βάση διαθέσιμα δεδομένα που αφορούν τις αγοραστικές τους συνήθειες είναι η αξιοποίηση αυτών των ομάδων για την εφαρμογή διαφορετικών και πιο κατάλληλων στρατηγικών μάρκετινγκ σε κάθε ομάδα ξεχωριστά.

Στόχος είναι να διερευνηθεί ποια μέθοδος ομαδοποίησης αποδίδει καλύτερα για τα συγκεκριμένα δεδομένα αλλά και να εξεταστεί η δυνατότητα χρήσης των αποτελεσμάτων της διπλωματικής για περεταίρω έρευνα των τεχνικών μάρκετινγκ μου μπορούν να εφαρμοστούν σε κάθε μία από τις παραγόμενες ομάδες.

Abstract

The purpose of this dissertation is to apply and evaluate clustering techniques in a real set of transaction data of a supermarket group for the purpose of grouping customers. The evaluation will be based on the results of the techniques that will be applied but also through a bibliographic review of the specific techniques.

The k-means method will be used as a reference point, while hierarchical grouping techniques, DBSCAN and fuzzy C-means, will be applied, and their performance will be compared with those of k-means. For each of the above methods we will try to point out issues that cause malfunctions during the modelling process and we will try to propose some solutions. We may face malfunctions due to sensitivities that do not exist in other methods. Also we will highlight the strengths of each method, such as the good application to multidimensional data.

The goal of the grouping process, i.e. customer classification into groups based on available data on their buying habits, is to use these groups to implement different and more appropriate marketing strategies for each group individually.

The aim is to draw conclusions about which grouping method works best for the specific data but also the possible use of the results of the dissertation for further research of marketing techniques that can be applied to each of the produced groups.

Περιεχόμενα

1. Εισαγωγή στο πρόβλημα της συσταδοποίησης

Ο όρος συσταδοποίηση ή αλλιώς ομαδοποίηση αναφέρεται σε τεχνικές με τις οποίες ένα σύνολο δεδομένων χωρίζεται σε ομάδες με σκοπό την αντιμετώπιση κάθε ομάδας ξεχωριστά ανάλογα με τα επιμέρους χαρακτηριστικά της. Μια επιτυχημένη συσταδοποίηση θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις σε κάθε μία από αυτές να είναι όσο γίνεται πιο όμοιες, ενώ παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο. Φυσικά για να συμβεί αυτό θα πρέπει να ορίσουμε τι σημαίνει η ομοιότητα δύο παρατηρήσεων.

Αυτό είναι ένα πολύ δύσκολο ερώτημα το οποίο θα πρέπει να απαντηθεί κάθε φορά από τον εκάστοτε ερευνητή ανάλογα με τη φύση του προβλήματος που υπάρχει προς επίλυση καθώς και το σύνολο των δεδομένων που έχουν συλλεχθεί για τον σκοπό αυτό. Κατά την ανάπτυξη των τεχνικών ομαδοποίησης έχει προταθεί η χρήση είτε ενός μέτρου απόστασης είτε ενός μέτρου ομοιότητας. Παρατηρήσεις οι οποίες είναι όμοιες πρέπει να δίνουν μικρές τιμές στα μέτρα απόστασης και μεγάλες τιμές στα μέτρα ομοιότητας, ενώ ανόμοιες παρατηρήσεις θα πρέπει να δίνουν τα αντίθετα αποτελέσματα.

Η ομαδοποίηση ανήκει σε μία ευρύτερη ομάδα αλγορίθμων, αυτή των *αλγορίθμων μηχανικής μάθησης χωρίς εποπτεία* (*unsupervised learning algorithms*). Οι αλγόριθμοι αυτοί ονομάζονται *αλγόριθμοι χωρίς εποπτεία* καθώς δεν υπάρχει κάποια μεταβλητή στόχος με βάση την οποία προσπαθούμε να κατατάξουμε τα δεδομένα σε ομάδες, κάτι που σημαίνει πως τα δεδομένα δεν έχουν κάποιο είδος ετικέτας που να τα διαχωρίζει. Αυτό έχει ως αποτέλεσμα να μην μπορούμε να αξιολογήσουμε την ακρίβεια των αποτελεσμάτων που παράγει ένας αλγόριθμος. Στην πραγματικότητα ο στόχος αυτών των μεθόδων είναι η εύρεση διαφόρων ομοιοτήτων ανάμεσα στα δεδομένα οι οποίες να τα κατατάσσουν σε ομάδες.

Το πρόβλημα της ομαδοποίησης μπορεί να προσεγγιστεί με διάφορους τρόπους. Μία πρώτη προσέγγιση, η οποία είναι κάπως εμπειρική, είναι με τη χρήση διαφόρων γραφικών παραστάσεων των παρατηρήσεων. Στη συνέχεια έχουμε τις ιεραρχικές και μη ιεραρχικές μεθόδους οι οποίες χρησιμοποιούνται πιο συχνά καθώς είναι οι πιο αποδοτικές μέθοδοι που υπάρχουν. Τέλος υπάρχουν και οι λιγότερο διαδεδομένες προσεγγίσεις της ομαδοποίησης όπως η μπευζιανή ομαδοποίηση, η ομαδοποίηση με βάση την πυκνότητα κ.α.

Η μέθοδος προσέγγισης αυτών των προβλημάτων μέσω γραφικών αναπαραστάσεων είναι απλοϊκή αλλά παράλληλα όχι και τόσο αξιόπιστη καθώς τα αποτελέσματα που παράγει είναι υποκειμενικά, αφού σε μεγάλο βαθμό ο αναλυτής είναι αυτός που κατατάσσει τα δεδομένα σε ομάδες με έναν εμπειρικό τρόπο. Η κύρια ιδέα πίσω από τις μεθόδους αυτές είναι ότι δημιουργείται κάποια γραφική παράσταση των πολυδιάστατων δεδομένων, π.χ. καμπύλες Andrews, sun ray plots, Chernoff faces

κ.α., και στη συνέχεια τοποθετούνται στην ίδια ομάδα παρατηρήσεις οι οποίες παρουσιάζουν παρόμοια εικόνα.

Ένας άλλος τρόπος προσέγγισης είναι με την χρήση μη ιεραρχικών μεθόδων. Οι τεχνικές αυτές έχουν ως βάση έναν επαναληπτικό αλγόριθμο, ο οποίος διαφοροποιείται ανάλογα με την μέθοδο που χρησιμοποιείται, και μέσω αυτού τοποθετούνται τα δεδομένα σε ομάδες με χρήση ενός μέτρου απόστασης ή ένα μέτρο ομοιότητας. Ένα μεγάλο μειονέκτημα αυτών των μεθόδων είναι ότι θα πρέπει ο ερευνητής να γνωρίζει εκ των προτέρων τον αριθμό των ομάδων (*k*) που θα δημιουργηθούν. Αυτό μπορεί να είναι γνωστό από πρότερη επιστημονική γνώση και εμπειρία στο αντικείμενο της έρευνας. Στην περίπτωση όμως που δεν υπάρχει αυτή η εμπειρία θα πρέπει να γίνουν δοκιμές και με βάση αυτές θα πρέπει να αποφασιστεί ποιος θα είναι ο πλέον κατάλληλος αριθμός των ομάδων που θα δημιουργηθούν. Οι βασικές τεχνικές λειτουργίας των μη ιεραρχικών μεθόδων είναι οι εξής:

- Επιλέγονται *k* συγκεκριμένες παρατηρήσεις ως μητρικά (αρχικά) σημεία και γύρω από αυτά τοποθετούνται οι υπόλοιπες παρατηρήσεις με βάση ένα μέτρο απόστασης ή ένα μέτρο ομοιότητας μέχρι να δημιουργηθούν οι ομάδες (μπορεί να γίνει επαναληπτικά περισσότερες της μίας φοράς) ή
- Δημιουργείται ένας αρχικός διαμερισμός των παρατηρήσεων σε *k* ομάδες και στη συνέχεια μετακινούνται τα άτομα μεταξύ των ομάδων μέχρι να επιτευχθεί ο καλύτερος διαμερισμός.

Στην κατηγορία αυτή ανήκει ο ευρύτατα διαδεδομένος αλγόριθμος *k*-means ο οποίος έχει ένα ευρύ φάσμα εφαρμογών σχεδόν σε όλους τους κλάδους που χρειάζεται να γίνει διαχωρισμός ατόμων, αντικειμένων ή αγαθών σε ομάδες καθώς είναι εύκολος στην εφαρμογή του και συνήθως πολύ αποδοτικός. Ο συγκεκριμένος αλγόριθμος έχει όμως κάποια μειονεκτήματα:

1. Ο ερευνητής πρέπει να καθορίσει εκ των προτέρων τον αριθμό των ομάδων *k*
2. Ο αλγόριθμος είναι ευαίσθητος στις ακραίες τιμές
3. Εξαρτάται από τα αρχικά σημεία ή την αρχική διαμέριση και
4. Μπορεί να εφαρμοστεί μόνο σε ποσοτικά δεδομένα.

Αξίζει να σημειωθεί ότι σε περίπτωση που δεν υπάρξει η σωστή αντιμετώπιση στα παραπάνω προβλήματα, ενέχει ο κίνδυνος τα αποτελέσματα που θα εξαχθούν να είναι λάθος και να υπάρξουν ανεπιθύμητες παρενέργειες γι' αυτούς που θα τα χρησιμοποιήσουν. Επίσης είναι αρκετά πιθανό δύο ερευνητές με τα ίδια δεδομένα να οδηγηθούν σε τελείως διαφορετικά αποτελέσματα, κάτι το οποίο θα πρέπει να αποφευχθεί.

Οι ιεραρχικές μέθοδοι, όπως φαίνεται και από το όνομά τους, δημιουργούν ιεραρχία από ομάδες. Υπάρχουν δύο διαφορετικοί τρόποι για να επιτευχθεί αυτό. Ο πρώτος τρόπος είναι ο συσσωρευτικός ο οποίος ξεκινάει με την κάθε παρατήρηση να αποτελεί μία ομάδα μόνη της και συγχωνεύει σταδιακά παρατηρήσεις μέχρις ότου ενωθούν όλες οι παρατηρήσεις σε μία ομάδα. Ο δεύτερος τρόπος είναι ο διαιρετικός ο

οποίος ξεκινάει με όλο το σύνολο δεδομένων σαν μία ομάδα και στη συνέχεια διασπάται σταδιακά έως ότου η κάθε παρατήρηση αποτελεί από μόνη της μία ομάδα. Το κυριότερο πρόβλημα αυτών των μεθόδων είναι ότι σε περίπτωση που το σύνολο το δεδομένων είναι αρκετά μεγάλο τότε απαιτείται πολύ μνήμη, χρόνος και υπολογιστική ισχύς και γι' αυτό συνήθως αποφεύγονται. Άλλο ένα πρόβλημα είναι ότι σε περίπτωση που δύο παρατηρήσεις ή δύο ομάδες παρατηρήσεων χωριστούν (διαιρετική μέθοδος) είτε ενωθούν (συσσωρευτική μέθοδος) αυτό δεν μπορεί να αλλάξει, το οποίο σημαίνει ότι και να εντοπιστεί κάποιο λάθος δεν υπάρχει τρόπος διόρθωσής του, άρα τα αποτελέσματα της μεθόδου δεν είναι τα καλύτερα δυνατά. Τα αποτελέσματα των συγκεκριμένων αλγορίθμων συνήθως απεικονίζονται σε μορφή δενδρογράμματος τα οποία είναι ευρέως κατανοητά.

Σε όλες τις παραπάνω μεθόδους είναι σύνηθες να χρησιμοποιείται αρχικά η μέθοδος *PCA* (Ανάλυση Κύριων Συνιστωσών). Η μέθοδος αυτή σκοπεύει στην ελάττωση των διαστάσεων του προβλήματος σε περίπτωση που στα δεδομένα είναι υπάρχει μεγάλο πλήθος μεταβλητών. Είναι μια μέθοδος η οποία χρησιμοποιείται μόνο σε αριθμητικά δεδομένα και δημιουργεί ένα πλήθος μεταβλητών (συνήθως πολύ μικρότερο από το αρχικό πλήθος) οι οποίες αποτελούν γραμμικό συνδυασμό των αρχικών μεταβλητών, έτσι ώστε να είναι ασυσχέτιστες μεταξύ τους και να διατηρούν όσο το δυνατό μεγαλύτερο ποσοστό πληροφορίας του αρχικού συνόλου δεδομένων.

Μία σημαντική εφαρμογή των παραπάνω τεχνικών αφορά χωρισμό των πελατών μίας επιχείρησης σε ομάδες με βάση τις αγοραστικές τους συνήθειες. Αυτή η διαδικασία γίνεται με σκοπό την διαφορετική αντιμετώπιση κάθε πελάτη με βάση την ομάδα που ανήκει είτε μέσω μίας στοχευμένης διαφήμισης, είτε για να γίνει η κατάλληλη προσπάθεια ικανοποίησης του πελάτη έτσι ώστε αυτός να παραμείνει πιστός στην εταιρία. Αυτός ο τρόπος διαφήμισης έχει χαμηλότερο κόστος και καλύτερο ποσοστό ανταπόκρισης.

Ο χωρισμός ξεκινάει συχνά χρησιμοποιώντας τις ήδη υπάρχουσες ομαδοποιήσεις π.χ. άνδρες-γυναίκες, οι οποίες εν συνεχεία σπάνε σε μικρότερες ομάδες που απαρτίζονται από πελάτες οι οποίοι έχουν ένα κοινό χαρακτηριστικό. Τα χαρακτηριστικά επιλέγονται πολύ προσεκτικά, καθώς είναι εκείνα τα οποία είναι πιθανό να επηρεάσουν την συμπεριφορά και τις αγοραστικές συνήθειες των πελατών.

2. Μέθοδοι συσταδοποίησης

Στο κεφάλαιο αυτό θα γίνει μία εκτενής θεωρητική παρουσίαση των μοντέλων που θα χρησιμοποιηθούν στην ανάλυση που θα ακολουθήσει. Οι μέθοδοι που θα παρουσιαστούν είναι οι:

- *k*-means
- Ιεραρχικές Μέθοδοι Ομαδοποίησης (Συσσωρευτικός, Διαιρετικός και Ενισχυμένος)
- DBSCAN
- Fuzzy *C*-means

Θα αναπτυχθεί η μεθοδολογία τους και θα αναφερθούν τα πλεονεκτήματα αλλά και τα μειονεκτήματα της κάθε μεθόδου.

2.1 k-means

Η μέθοδος k-means προτάθηκε από τον MacQueen (1967), και είναι η πιο γνωστή μέθοδος συσταδοποίησης, καθώς είναι πολύ απλή και κατανοητή. Στόχος της μεθόδου είναι να διαιρέσει το σύνολο των δεδομένων σε *k* (προκαθορισμένο από τον ερευνητή) ομάδες. Ο αλγόριθμος περιλαμβάνει μια επαναληπτική διαδικασία, και χρησιμοποιεί την έννοια του κέντρου (centroid) το οποίο είναι η μέση τιμή των παρατηρήσεων κάθε ομάδας, για κάθε μεταβλητή.

Υπάρχουν δύο βασικές προσεγγίσεις της μεθόδου k-means. Στην πρώτη προσέγγιση σε κάθε επανάληψη τοποθετείται ένα στοιχείο από τα δεδομένα σε μία ομάδα και γίνεται επανυπολογισμός του κέντρου της αντίστοιχης ομάδας. Στην δεύτερη προσέγγιση πρώτα τοποθετούνται όλα τα άτομα στις ομάδες με τα πιο κοντινά κέντρα και στη συνέχεια υπολογίζονται εκ νέου τα κέντρα.

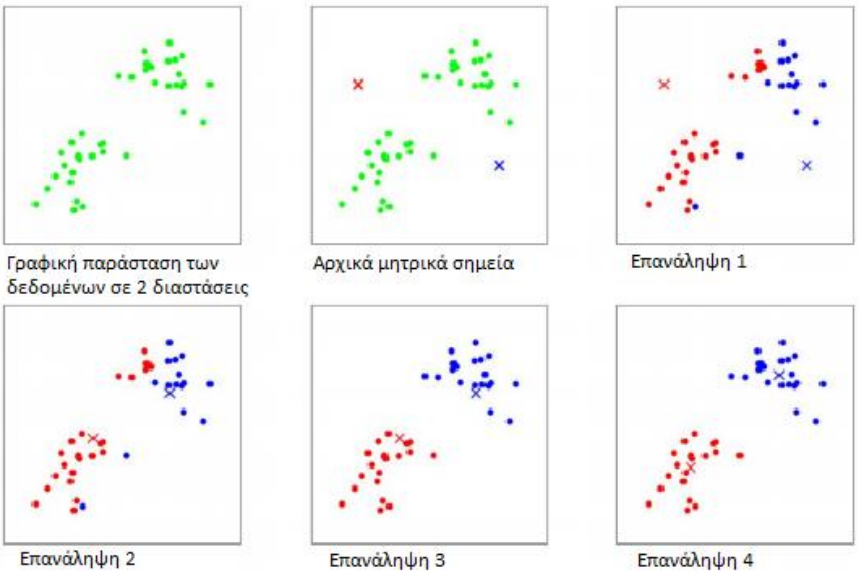
Αναλυτικότερα, ο αλγόριθμος της μεθόδου k-means με την πρώτη προσέγγιση έχει ως εξής:

- Επιλέγονται *k* μητρικά σημεία και θεωρούνται κέντρα των ομάδων
- Κατατάσσεται κάθε άτομο σε μία ομάδα υπολογίζοντας τις αποστάσεις του ατόμου από κάθε κέντρο βάρους και επιλέγοντας την μικρότερη απόσταση από αυτές. Στη συνέχεια υπολογίζουμε το κέντρο βάρους της αλλαγμένης πλέον ομάδας μετά από κάθε εισαγωγή ενός σημείου σε κάποια ομάδα
- Το προηγούμενο βήμα επαναλαμβάνεται μέχρις ότου όλα τα άτομα του συνόλου των δεδομένων ενταχθούν σε μία ομάδα
- Τα κέντρα που έχουν δημιουργηθεί θεωρούνται ως μητρικά σημεία και γίνεται μία τελική σάρωση στα δεδομένα τοποθετώντας κάθε άτομο στο πλησιέστερο κέντρο χωρίς να γίνεται επανυπολογισμός του κέντρου

Ο αλγόριθμος της μεθόδου *k*-means με την δεύτερη προσέγγιση έχει ως εξής:

- Διαμερίζεται το σύνολο των δεδομένων σε *k* ομάδες και υπολογίζονται τα κέντρα βάρους.
- Υπολογίζονται οι αποστάσεις κάθε ατόμου από κάθε κέντρο και στη συνέχεια κατατάσσεται κάθε άτομο στο κέντρο με την μικρότερη απόσταση. Στη συνέχεια υπολογίζονται εκ νέου τα κέντρα βάρους.
- Επαναλαμβάνουμε το παραπάνω βήμα έως ότου συμβεί επανάληψη η οποία δεν θα επιφέρει καμία αλλαγή στις ομάδες της προηγούμενης επανάληψης.

Η παρακάτω εικόνα δείχνει τα αποτελέσματα των τεσσάρων επαναλήψεων ενός παραδείγματος, που οδηγούν στην τελική συσταδοποίηση στην περίπτωση που θέλουμε να φτιάξουμε δύο ομάδες με τα συγκεκριμένα δεδομένα. Στις τέσσερεις επαναλήψεις με *x* έχουν σημειωθεί τα κέντρα των ομάδων που δημιουργούνται σε κάθε μία από αυτές.



Εικόνα 1: Απεικόνιση αποτελεσμάτων τεσσάρων επαναλήψεων της μεθόδου *k*-means (http://bennymachinelearning.blogspot.com/)

Η μέθοδος *k*-means θεωρείται ιδιαίτερα γρήγορη αφού σε περιπτώσεις που χρησιμοποιείται για ομαδοποίηση μεγάλων δεδομένων τερματίζει μετά από λίγες επαναλήψεις, όμως και στις περιπτώσεις που απαιτούνται αρκετές επαναλήψεις έχει παρατηρηθεί ότι κατά την εφαρμογή των τελευταίων επαναλήψεων οι διαφοροποιήσεις που πραγματοποιούνται είναι ελάχιστες, και αυτές σε άτομα τα οποία βρίσκονται στο όριο ανάμεσα σε δύο ομάδες. Αυτό σημαίνει ότι όσο αυξάνεται ο αριθμός των επαναλήψεων οι διαφοροποιήσεις δεν είναι σημαντικές και άρα μπορούν να παραληφθούν. Επίσης ο τρόπος που σκανάρει τα δεδομένα είναι δομημένος έτσι ώστε να μην κρατά πολλά στοιχεία στην μνήμη και έτσι οι απαιτήσεις σε υπολογιστική ισχύ και μνήμη είναι πολύ χαμηλές. Ένα μειονέκτημα της μεθόδου

είναι ότι οι ομάδες που δημιουργούνται είναι συνήθως ισοπληθείς και αυτό καθιστά τη μέθοδο ακατάλληλη για δεδομένα με διαφορετικό πλήθος ατόμων ανά ομάδα.

Όπως αναφέρθηκε και στην εισαγωγή, η μέθοδος *k*-means αντιμετωπίζει κάποια προβλήματα τα οποία καλείται να λύσει ο ερευνητής. Ένα από τα κυριότερα προβλήματα είναι ότι τα αποτελέσματα που θα προκύψουν επηρεάζονται πολύ από την επιλογή των μητρικών σημείων. Σε πολλές περιπτώσεις η μέθοδος δεν οδηγεί σε ολική βέλτιστη λύση ομαδοποίησης, αλλά εγκλωβίζεται σε μία τοπικά βέλτιστη λύση. Αν λοιπόν σε μία εκτέλεση του αλγορίθμου τα αρχικά σημεία επιλεγούν τυχαία μεν αλλά με τρόπο τέτοιο ώστε να οδηγούν σε μία τοπικά βέλτιστη λύση δεν υπάρχει τρόπος να ξεπεράσουμε αυτό το πρόβλημα. Σε περίπτωση που είναι γνωστό εκ των προτέρων περίπου πως θα σχηματιστούν οι ομάδες βάση παλαιότερων παρόμοιων μελετών το πρόβλημα αντιμετωπίζεται δηλώνοντας ποια μητρικά σημεία θα χρησιμοποιήσει η μέθοδος. Στην περίπτωση όμως που δεν κατέχει ο ερευνητής αυτή την γνώση ο τρόπος αντιμετώπισης του προβλήματος είναι η επαναλαμβανόμενη εφαρμογή της μεθόδου αρκετές φορές έτσι ώστε με διάφορα κριτήρια που έχουν προταθεί να επιλεγεί η βέλτιστη λύση.

Άλλο ένα πρόβλημα των συγκεκριμένων αλγορίθμων είναι η ευαισθησία τους στα outliers ή αλλιώς στις ακραίες τιμές. Στην περίπτωση που το σύνολο των δεδομένων περιέχει ακραίες τιμές , και ειδικά στην περίπτωση που το δείγμα είναι μικρό, τα κέντρα των ομάδων επηρεάζονται πολύ από αυτές τις τιμές και οδηγούν σε λάθος αποτελέσματα. Το πρόβλημα με τις ακραίες τιμές έγκειται στην εύρεσή τους καθώς σε περίπτωση που εντοπισθούν ο πιο σύνηθες τρόπος αντιμετώπισής τους είναι η διαγραφή τους. Όπως γίνεται αντιληπτό η μέθοδος μπορεί να εφαρμοστεί μόνο σε ποσοτικά δεδομένα και όχι σε κατηγορικά.

Ίσως το μεγαλύτερο πρόβλημα αυτής της μεθόδου είναι η επιλογή του *k*, του αριθμού των ομάδων που θα δημιουργηθούν. Στην περίπτωση που ο αριθμός αυτός είναι γνωστός από παλαιότερες μελέτες, είναι πιθανό να μην αντιπροσωπεύεται κάποια ομάδα στο δείγμα και έτσι αυτή η ομάδα να μην εντοπίζεται από τον αλγόριθμο. Αυτό οδηγεί σε αποτελέσματα που δεν ανταποκρίνονται στην πραγματικότητα όπως π.χ. την διάσπαση μίας συμπαγούς ομάδας. Η πιο συνηθισμένη αντιμετώπιση αυτού του προβλήματος είναι η εφαρμογή της μεθόδου για διαφορετικές τιμές του *k* και η επιλογή εκείνου που μας δίνει την καλύτερη τιμή σε κάποιον δείκτη. Οι πιο συνηθισμένοι δείκτες που έχουν προταθεί είναι οι παρακάτω:

- Ο δείκτης silhouette:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}},$$

όπου $a(i)$ είναι η μέση απόσταση του σημείου i με όλα τα υπόλοιπα σημεία μέσα στην ίδια ομάδα και $b(i)$ είναι η ελάχιστη μέση απόσταση του σημείου i με όλα τα σημεία κάθε άλλης κλάσης. Στη συνέχεια υπολογίζεται ο μέσος όρος για κάθε ομάδα και ο

τελικός δείκτης βρίσκεται από τον μέσο όρο των δεικτών των ομάδων. Ο δείκτης παίρνει τιμές από το -1 μέχρι το 1 και επιθυμητές είναι οι τιμές κοντά στο 1.

- Το Ch index :

$$CH(k) = \frac{B(k)}{W(k)} \times \frac{n-k}{k-1},$$

όπου $W(k)$ είναι η διακύμανση μέσα στις κλάσεις, $B(k)$ είναι η διακύμανση ανάμεσα στις κλάσεις, n είναι ο αριθμός των σημείων του συνόλου δεδομένων και k είναι ο αριθμός των κλάσεων που επιλέγεται να δημιουργηθούν. Η τιμή του k που παράγει τον μέγιστο Ch index είναι και αυτή που επιλέγουμε.

2.2 Ιεραρχικές Μέθοδοι Ομαδοποίησης

Αυτή η κατηγορία ομαδοποίησης χωρίζεται σε δύο μεγάλες υποκατηγορίες, τις συσσωρευτικές και τις διαιρετικές μεθόδους. Οι συσσωρευτικές ξεκινούν με κάθε στοιχείο σαν μία ομάδα και καταλήγουν με όλα τα στοιχεία σε μια ομάδα, ενώ οι διαιρετικές λειτουργούν με τον αντίθετο ακριβώς τρόπο. Η ανάλυση που θα ακολουθήσει αφορά αποκλειστικά τις συσσωρευτικές μεθόδους καθώς είναι πιο διαδεδομένες. Σκοπός της μεθόδου είναι να δημιουργήσει μία ιεραρχία ομάδων όχι απαραίτητα ίσου μεγέθους και στη συνέχεια ο ερευνητής με διάφορα κριτήρια επιλέγει την κατάλληλη ομαδοποίηση.

Αρχικά θα πρέπει να οριστούν τα μέτρα απόστασης και τα μέτρα ομοιότητας. Μία συνάρτηση $d_{ij} = d(x_i, x_j)$, όπου x_i και x_j είναι δύο στοιχεία του συνόλου δεδομένων, καλείται απόσταση όταν πληρούνται οι τρεις παρακάτω ιδιότητες :

1) $d_{ij} \geq 0 \; \forall \; i, j \; και \; d_{ij} = 0 \; \Leftrightarrow \; i = j$
2) $d_{ij} \leq d_{is} + d_{sj} \; \forall \; s$
3) $d_{ij} = d_{ji}$

Η πιο σημαντική ιδιότητα από τις τρεις είναι η δεύτερη, δηλαδή η τριγωνική, η οποία όμως δεν ικανοποιείται από κάποια μέτρα που χρησιμοποιούνται στην πράξη. Οι πιο γνωστές αποστάσεις είναι οι εξής :

Ευκλείδεια απόσταση : $d_{ij} = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$

Απόσταση του Pearson : $d_{ij} = \sqrt{\sum_{r=1}^{p} \left(\frac{x_{ir} - x_{jr}}{s_r}\right)^2}$

Απόσταση Manhattan : $d_{ij} = \sum_{r=1}^{p} |x_{ir} - x_{jr}|$

όπου s_r συμβολίζει την τυπική απόκλιση της r μεταβλητής και p είναι το πλήθος των μεταβλητών του συνόλου δεδομένων.

Μέτρο ομοιότητας είναι μία αριθμητική μέτρηση για το πόσο όμοια είναι δύο στοιχεία του συνόλου των δεδομένων. Για τα μέτρα ομοιότητας εύκολα αποδεικνύεται ότι αν έχει δημιουργηθεί ένα μέτρο απόστασης, τότε προκύπτει ένα αντίστοιχο μέτρο ομοιότητας s_{ij} (και αντίστροφα) μέσω του τύπου:

$$s_{ij} = \frac{1}{1 + d_{ij}}$$

Τα βασικά βήματα των συσσωρευτικών μεθόδων είναι τα έξης :

- Αναθέτεται κάθε στοιχείο του συνόλου σε μία ομάδα και δημιουργείται ο πίνακας αποστάσεων όλων των στοιχείων μεταξύ τους
- Εντοπίζεται το κοντινότερο (πιο όμοιο) ζευγάρι ομάδων και συνενώνονται σε μία ομάδα
- Επανυπολογίζεται ο πίνακας αποστάσεων διαγράφοντας τις γραμμές και τις στήλες του παλιού πίνακα που αντιστοιχούν στις ομάδες που ενώθηκαν και προσθέτοντας μία γραμμή και μία στήλη που θα περιέχει τις αποστάσεις της καινούριας ομάδας από όλες τις υπόλοιπες.
- Επαναλαμβάνονται τα δύο παραπάνω βήματα μέχρις ότου όλα τα δεδομένα τοποθετηθούν σε μία ομάδα. Σε κάθε επανάληψη καταγράφονται οι αποστάσεις στις οποίες έγιναν οι συνενώσεις.

Με τον ακριβώς αντίθετο τρόπο από αυτόν που παρουσιάστηκε παραπάνω μπορούν να υλοποιηθούν οι διαιρετικές μέθοδοι.

Όμως ενώ έχει οριστεί η απόσταση (ομοιότητα) ανάμεσα σε δύο σημεία, δεν έχει γίνει το ίδιο και για την απόσταση δύο ομάδων. Αυτό μπορεί να γίνει με διάφορους τρόπους. Ο πιο γνωστός τρόπος είναι αυτός του κοντινότερου γείτονα, όπου η απόσταση μεταξύ δύο ομάδων ορίζεται να είναι ίση με την απόσταση των δύο κοντινότερων σημείων των ομάδων. Αντίστοιχα και η μέθοδος του μακρινότερου γείτονα που ορίζεται ως η απόσταση ανάμεσα στα δύο πιο μακρινά σημεία των δύο ομάδων. Παρόμοια μέθοδος προκύπτει αν αντί για την χρήση των κοντινότερων ή μακρύτερων σημείων των δύο ομάδων χρησιμοποιηθούν τα κέντρα βάρους των ομάδων. Ένας πιο περίπλοκος τρόπος σύνδεσης των ομάδων είναι η μέθοδος του Wald. Σκοπός της μεθόδου είναι στις ομάδες που δημιουργούνται να ελαχιστοποιείται η διακύμανση ανάμεσα στα σημεία με σκοπό να βρίσκονται όσο πιο κοντά στο κέντρο βάρους τους. Σε κάθε βήμα της μεθόδου δοκιμάζονται όλες οι δυνατές συγχωνεύσεις και συγχωνεύονται οι ομάδες που δίνουν την μικρότερη αύξηση των συνολικών αποκλίσεων.

Το μεγαλύτερο πλεονέκτημα των συγκεκριμένων μεθόδων είναι ότι δεν χρειάζονται σαν είσοδο τον αριθμό των ομάδων που θα δημιουργηθούν καθώς δεν δημιουργούν έναν συγκεκριμένο αριθμό ομάδων, αλλά μία ιεραρχία ομάδων, και στο τέλος ο ερευνητής καλείται να επιλέξει είτε μέσω δενδρογράμματος, είτε μέσω κάποιας άλλης τεχνικής ποιες και πόσες ομάδες θα κρατήσει. Ένα άλλο πλεονέκτημα

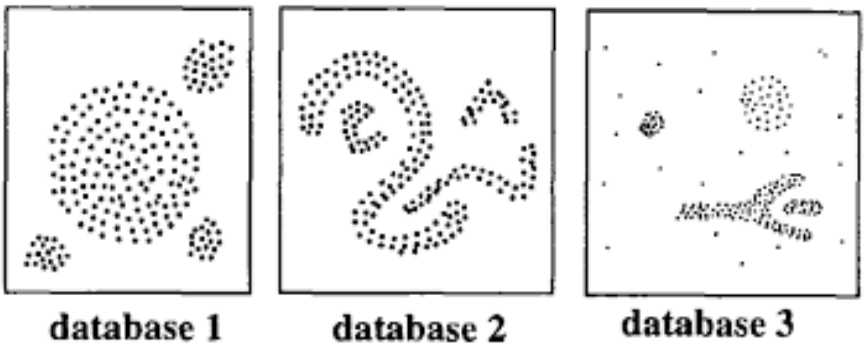
είναι ότι η εφαρμογή τους είναι πολύ εύκολη και σε κάποιες περιπτώσεις τα αποτελέσματα που παράγουν είναι τα καλύτερα από οποιαδήποτε άλλης μεθόδου.

Παρ' όλα αυτά οι μέθοδοι ιεραρχικής ομαδοποίησης έχουν αρκετά μειονεκτήματα. Βασικότερο όλων είναι ότι σε περίπτωση που υλοποιηθεί κάποιο βήμα του αλγορίθμου δεν μπορεί να αλλάξει. Δεν υπάρχει τρόπος αξιολόγησης κάποιας ένωσης που έγινε στον αλγόριθμο κάτι που σημαίνει ότι δεν μπορεί και να διορθωθεί. Κάποιες φορές είναι δύσκολο μέσω του δενδρογράμματος να εντοπιστεί ο σωστός αριθμός των κλάσεων που θα χωριστεί το σύνολο των δεδομένων. Ένα άλλο πρόβλημα που υπάρχει είναι ότι ανάλογα με την συνάρτηση σύνδεσης που χρησιμοποιείται, δηλαδή τον τρόπο που θα συγχωνεύονται ή θα διαιρούνται δύο ομάδες, μπορεί οι αλγόριθμοι να γίνουν ευαίσθητοι στον θόρυβο και στις ακραίες τιμές ή να σπάνε μεγάλες συμπαγείς κλάσεις ή να υπάρχει δυσκολία στον χειρισμό κλάσεων διαφορετικού μεγέθους ή σχήματος.

Παρότι οι ιεραρχικές μέθοδοι είναι σχεδόν συνυφασμένες με τις συσσωρευτικές μεθόδους όπως προαναφέραμε υπάρχει και άλλη μια κατηγορία, οι διαιρετικές μέθοδοι. Ξεκινούν με μία μόνο ομάδα η οποία περιέχει όλα τα στοιχεία και στην συνέχεια διαιρείται σε μικρότερες ομάδες. Η λογική είναι στις υπάρχουσες ομάδες να βρίσκουν υποσύνολα στοιχείων που είναι απομακρυσμένα μεταξύ τους και εκεί να τις διαχωρίζουν. Σε κάθε επανάληψη του αλγορίθμου χωρίζεται μία ομάδα σε δύο υποομάδες μέχρι κάθε στοιχείο να βρίσκεται μία ομάδα μόνο του. Ο βασικότερος λόγος που οι διαιρετικές μέθοδοι είναι λιγότερο διαδεδομένες από τις συσσωρευτικές είναι ότι απαιτούν πολύ περισσοτέρους υπολογισμούς.

2.3 DBSCAN

Ο αλγόριθμος DBSCAN (Density Based Spatial Clustering of Applications with Noise) ανήκει στην κατηγορία αλγορίθμων οι οποίοι βασίζονται στην πυκνότητα των δεδομένων και δημιουργούν ομάδες με βάση τη συγκέντρωση ή όχι των σημείων γύρω από μία περιοχή. Η βασική ιδέα είναι ότι για κάθε σημείο μίας συστάδας που δημιουργείται πρέπει να υπάρχει μία γειτονιά, με δεδομένη ακτίνα, μέσα στην οποία θα πρέπει να περιέχονται κατ' ελάχιστο ένας προκαθορισμένος αριθμός σημείων. Το σχήμα της γειτονιάς καθορίζεται από την επιλογή της συνάρτησης απόστασης ανάμεσα σε δύο σημεία.



Εικόνα 2: Τυχαία δείγματα δεδομένων από τους Ester, M., Kriegel, H. P., Sander, J. και Xu, X (1996)

Στην παραπάνω εικόνα μπορούμε εύκολα να διακρίνουμε τις κλάσεις που σχηματίζονται αλλά και τα σημεία θορύβου που δεν ανήκουν σε καμία κλάση. Αυτό συμβαίνει καθώς οι κλάσεις αποτελούνται από ομάδες σημείων οι οποίες έχουν πυκνότητα πολύ μεγαλύτερη από τα σημεία εκτός ομάδων. Ως πυκνότητα ορίζουμε το πλήθος των σημείων σε μία μονάδα όγκου. Σε αυτό το σημείο θα δοθούν κάποιοι ορισμοί που θα βοηθήσουν στην περεταίρω ανάλυση του αλγορίθμου.

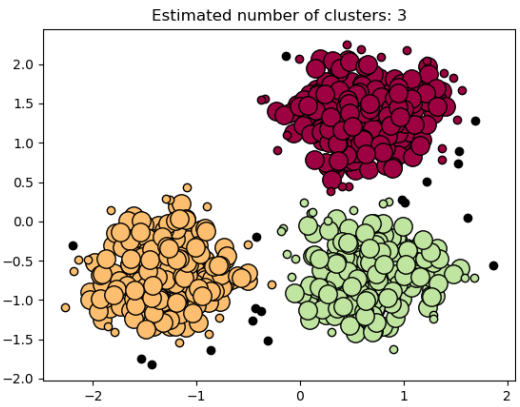
- MinPts: ελάχιστος αριθμός σημείων μέσα σε μία συστάδα (απαιτείται σαν είσοδος από τον αλγόριθμο)
- Eps: δηλώνει απόσταση και για κάθε σημείο της συστάδας θα πρέπει να υπάρχει ένα άλλο σημείο συστάδας με απόσταση μικρότερη από Eps (απαιτείται σαν είσοδος από τον αλγόριθμο)
- Eps-γειτονιά ενός σημείου: Το σύνολο των σημείων μέσα σε Eps απόσταση από το σημείο
- Πυρήνας (core point): ένα σημείο με Eps-γειτονιά αρκετά πυκνή ώστε πληθυσμός γειτονιάς ≥ minPts
- Σύνορο (border point): ένα σημείο με Eps-γειτονιά όχι αρκετά πυκνή ώστε να ονομαστεί πυρήνας, το οποίο όμως ανήκει στην Eps-γειτονιά ενός πυρήνα
- Θόρυβος (noise): ένα σημείο που δεν είναι ούτε πυρήνας ούτε σύνορο
- Γειτονιά του σημείου p : $N_{Esp}(p) = \{q \in D \mid \text{dist}(p,q) \leq Esp\}$
- Άμεσα προσβάσιμο: Ένα σημείο p είναι άμεσα προσβάσιμο από ένα σημείο q αν:
    1) $p \in N_{Esp}(q)$
    2) $| N_{Esp}(q) | \geq$ MinPts
- Προσβάσιμο: Ένα σημείο p είναι προσβάσιμο από ένα σημείο q αν υπάρχει μία αλυσίδα σημείων $p_1, p_2, \ldots, p_n$, $p_1 = q$, $p_n = p$ τέτοια ώστε το p_{i+1} να είναι άμεσα προσβάσιμο από το p_i
- Συνδεδεμένο: Ένα σημείο p είναι συνδεδεμένο με ένα σημείο q αν υπάρχει ένα σημείο r τέτοιο ώστε τα p και q να είναι προσβάσιμα από το r

Οι αλγόριθμοι που βασίζονται στην πυκνότητα έχουν παίξει καθοριστικό ρόλο στην ομαδοποίηση δεδομένων που περιέχουν ομάδες με ακανόνιστο σχήμα που σε πολλές περιπτώσεις δεν μπορεί να προσδιοριστεί. Ο αλγόριθμος DBSCAN είναι ο πιο

γνωστός αλγόριθμος σε αυτή την κατηγορία. Τα βήματα του αλγορίθμου είναι τα εξής (οι μεταβλητές Eps και MinPts δίνονται σαν είσοδοι στον αλγόριθμο ή βρίσκονται οι βέλτιστες τιμές τους με διάφορες τεχνικές βελτιστοποίησης):

- Επιλέγεται ένα τυχαίο αρχικό σημείο
- Δημιουργείται η Eps-γειτονιά του σημείου
- Αν δημιουργείται μία επαρκής γειτονιά γύρω από το σημείο, δηλαδή μέσα στην Eps-γειτονία του σημείου υπάρχουν τουλάχιστον MinPts σημεία, η διαδικασία συνεχίζει και το σημείο σημειώνεται ως διαβασμένο, σε αντίθετη περίπτωση το σημείο σημειώνεται ως ακραία τιμή
- Αν ένα σημείο αποτελεί μέρος της κλάσης τότε όλα τα σημεία στην γειτονιά του ανήκουν στην κλάση και επαναλαμβάνεται η παραπάνω διαδικασία για κάθε σημείο της γειτονιάς μέχρις ότου να μην υπάρχουν καινούρια σημεία που μπορούν να μπουν σε αυτή την κλάση
- Επιλέγεται ένα καινούριο σημείο που δεν έχει σημειωθεί ως αναγνωσμένο, δηλαδή δεν έχει επεξεργαστεί ακόμα από τον αλγόριθμο, και ξεκινάει η διαδικασία από την αρχή με σκοπό την δημιουργία μιας νέας κλάσης
- Η διαδικασία ολοκληρώνεται όταν όλα τα σημεία έχουν σημειωθεί ως διαβασμένα

Η παρακάτω εικόνα δείχνει τα αποτελέσματα της εφαρμογής της μεθόδου σε ένα σύνολο δεδομένων



Εικόνα 3: Παράδειγμα συσταδοποίησης με την μέθοδο DBSCAN (https://scikit-learn.org/)

Όπως βλέπουμε δημιουργούνται τρεις ομάδες για τα συγκεκριμένα δεδομένα, καθεμία από τις οποίες έχει σημειωθεί με διαφορετικό χρώμα. Τα σημεία με μεγάλο κύκλο υποδηλώνουν ότι είναι σημεία πυρήνες, τα σημεία με μικρότερο κύκλο αλλά με χρώμα υποδηλώνουν σύνορα και τέλος τα σημεία που είναι με μαύρο χρώμα υποδηλώνουν ακραίες τιμές.

Ένα πολύ μεγάλο προσόν του συγκεκριμένου αλγορίθμου, αλλά και γενικά των αλγορίθμων αυτής της κατηγορίας, είναι ότι δεν υπάρχει η ανάγκη του εκ των προτέρων προσδιορισμού του αριθμού των συστάδων που θα δημιουργηθούν. Αυτό

σημαίνει ότι ο ερευνητής δεν είναι απαραίτητο να έχει γνώση πάνω στο θέμα ή να υπάρχει αντίστοιχη μελέτη στο παρελθόν. Άλλο ένα πολύ μεγάλο πλεονέκτημα του συγκεκριμένου αλγορίθμου είναι ότι προσδιορίζει τις ακραίες τιμές των δεδομένων κατά την διάρκεια της ομαδοποίησης. Μάλιστα σε πολλές περιπτώσεις ο συγκεκριμένος αλγόριθμος χρησιμοποιείται σαν προεργασία για άλλες μεθόδους μόνο και μόνο για την συγκεκριμένη του ιδιότητα. Το γεγονός ότι ο αλγόριθμος δουλεύει αποκλειστικά χρησιμοποιώντας την πυκνότητα των δεδομένων, τον κάνει ικανό να βρίσκει κλάσεις οι οποίες διαφέρουν σε σχήμα και σε μέγεθος, κάτι που είναι αρκετά σύνηθες σε πραγματικά δεδομένα. Τέλος άλλο ένα συγκριτικό πλεονέκτημα της μεθόδου είναι το γεγονός ότι για συγκεκριμένο σύνολο δεδομένων και τις ίδιες τιμές στις μεταβλητές που χρησιμοποιούνται σαν είσοδος παράγονται πάντα τα ίδια αποτελέσματα όσες φορές και αν τρέξει ο αλγόριθμος.

Η συγκεκριμένη μέθοδος έχει κάποια προβλήματα τα οποία έχουν αποτέλεσμα σε πολλές περιπτώσεις να μην είναι αρκετά αποδοτική. Το κύριο πρόβλημα έγκειται στο γεγονός ότι δεν αποδίδει καλά στην περίπτωση των πολυδιάστατων δεδομένων με μεγάλες διαστάσεις κάτι το οποίο είναι αρκετά σύνηθες στα σύγχρονα προβλήματα. Άλλο ένα αρνητικό της μεθόδου είναι ότι αποτυγχάνει στην περίπτωση που το πρόβλημα περιέχει κλάσεις με διαφορετικές πυκνότητες. Αυτό συμβαίνει γιατί δεν μπορεί να προσδιοριστεί ο αριθμός Eps έτσι ώστε να είναι λειτουργικός για κάθε κλάση.

2.4 Fuzzy *C*-means

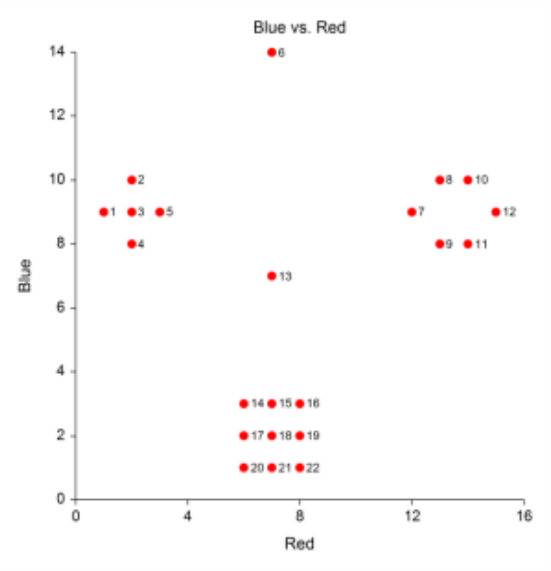
Το fuzzy clustering θεωρείται μια τεχνική αδρής ομαδοποίησης (*soft clustering*) καθώς δεν είναι υποχρεωτικό τα στοιχεία που ομαδοποιούνται σε μία και μόνο ομάδα. Κάθε στοιχείο μπορεί να ανήκει σε δύο ή και παραπάνω ομάδες με διαφορετικές πιθανότητες για κάθε ομάδα. Σε αυτή την ομάδα αλγορίθμων ένα σημείο κοντά στο κέντρο της ομάδας ανήκει σε μεγαλύτερο βαθμό στην ομάδα απ' ότι ένα σημείο μακριά από το κέντρο.

Ο αλγόριθμος *Fuzzy C-means* προτάθηκε από τον Jim Bezdek το 1981 ως βελτίωση για την μέθοδο k-means και είναι ο πιο γνωστός αυτής της κατηγορίας. Υπολογίζει το κέντρο κάθε ομάδας ως σταθμισμένο μέσο όλων των σημείων του συνόλου δεδομένων, δίνοντας ανάλογη βαρύτητα σε κάθε σημείο. Η βαρύτητα που αναλογεί σε κάθε σημείο προσδιορίζεται από το ποσοστό συμμετοχής του στην ομάδα, δηλαδή την πιθανότητα που έχει το κάθε σημείο να ανήκει σε κάθε μία από τις ομάδες. Σκοπός του αλγορίθμου είναι η ελαχιστοποίηση της συνάρτησης:

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{k}(\mu_{ij})^m \left\|\boldsymbol{x}_i - \boldsymbol{c}_j\right\|^2$$

όπου μ_{ij} είναι η πιθανότητα η i παρατήρηση να ανήκει στην j κλάση, $m \geq 1$ είναι ένας ακέραιος αριθμός που προσδιορίζει πόσο ασαφής θα είναι η ομαδοποίηση, $\|\boldsymbol{x}_i - \boldsymbol{c}_j\|^2$ είναι η ευκλείδεια απόσταση του i σημείου από το κέντρο της j κλάσης, και τέλος n είναι το πλήθος των δεδομένων και k είναι το πλήθος των ομάδων που θα δημιουργηθούν.

Για να γίνουν κατανοητοί οι λόγοι εισαγωγής αυτής της μεθόδου, παρατίθεται το γράφημα ενός απλού δισδιάστατου παραδείγματος:



Εικόνα 4: Δισδιάστατο παράδειγμα δεδομένων

Στην παραπάνω εικόνα φαίνεται ξεκάθαρα η ύπαρξη τριών κλάσεων και δύο ακραίων τιμών, (6 και 13). Η μέθοδος k-means θα κατέτασσε αυτές τις δύο τιμές σε κάποια κλάση. Με την μέθοδο fuzzy *C*-means θα ανατεθούν στα συγκεκριμένα σημεία μεγάλες πιθανότητες στο να ανήκουν σε κάθε κλάση. Στη συνέχεια ορίζεται ένα κατώφλι όπου αν η πιθανότητα το σημείο να ανήκει σε όλες τις κλάσεις είναι πάνω από αυτό το κατώφλι τότε το σημείο θεωρείται ως ακραία τιμή και έτσι γίνεται εύκολος ο εντοπισμός των ακραίων τιμών του συνόλου.

Τα βήματα του αλγορίθμου Fuzzy C-means είναι τα εξής:

- Επιλέγονται τυχαία τα C κέντρα των ομάδων
- Υπολογίζονται οι πιθανότητες συμμετοχής μ_{ij} του i στοιχείου στην j ομάδα μέσω του τύπου:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} (d_{ij}/d_{ik})^{(2/(m-1))}}$$

όπου d συμβολίζει της απόσταση ενός σημείου από το κέντρο μίας ομάδας.

- Επανυπολογίζονται τα κέντρα των ομάδων μέσω του τύπου:

$$\boldsymbol{v_j} = \frac{\sum_{i=1}^{N} (\boldsymbol{\mu_{ij}})^m \boldsymbol{x_i}}{\sum_{i=1}^{N} (\boldsymbol{\mu_{ij}})^m} \; \forall j = 1,2, \ldots, C$$

- Η διαδικασία τερματίζεται αν η αλλαγή σε κάποιο κέντρο είναι κάτω από ένα κατώφλι, αλλιώς επανάλαβε τα δύο παραπάνω βήματα

Όπως αναφέρθηκε και πιο πάνω, το μεγαλύτερο πλεονέκτημα της μεθόδου είναι ότι κάθε παρατήρηση δεν ανήκει αποκλειστικά σε μία ομάδα αλλά σε όλες, με κάποια πιθανότητα για κάθε μία. Αυτό την κάνει πολύ ισχυρή σε σύνολα δεδομένων όπου οι ομάδες επικαλύπτονται μεταξύ τους και είναι καλύτερη από τον αλγόριθμο k-means.

Όπως και στον *k*-means έτσι και εδώ πρέπει να προσδιοριστεί εκ των προτέρων το πλήθος των ομάδων που θα δημιουργηθούν. Το γεγονός ότι δεν υπάρχει ξεκάθαρος διαχωρισμός ανάμεσα στις ομάδες, κάνει τα κέντρα αυτών να μετατοπίζονται πιο αργά και έτσι χρειάζονται πολλές επαναλήψεις για να επιτευχθεί η βέλτιστη λύση. Φυσικά ρόλο στο πλήθος των επαναλήψεων παίζει και το κατώφλι που θα επιλεγεί ως κριτήριο τερματισμού του αλγορίθμου. Ένα ακόμα μειονέκτημα του αλγορίθμου είναι ότι η χρήση της ευκλείδειας απόστασης μπορεί να δώσει λανθασμένα λιγότερο βάρος σε κάποιο παράγοντα, δηλαδή μεταβλητή των δεδομένων, ο οποίος στην πραγματικότητα θα μπορούσε να καθορίζει σε μεγάλο βαθμό τον σχηματισμό των ομάδων.

3. Τεχνικές Προ-επεξεργασία Δεδομένων

Πριν από οποιαδήποτε μορφή στατιστικής ανάλυσης θα πρέπει να ελέγχονται τα δεδομένα ως προς την ποιότητα και την καταλληλότητά τους έτσι ώστε να διασφαλίζεται η εξαγωγή εγκύρων αποτελεσμάτων. Είναι ένα στάδιο της έρευνας το οποίο δεν πρέπει να παραλείπεται και βοηθάει πολύ στην ομαλότερη πορεία της έρευνας. Η προ-επεξεργασία των δεδομένων χωρίζεται σε δύο μεγάλες κατηγορίες, τον καθαρισμό των δεδομένων και την αντιμετώπιση των ακραίων τιμών. Παρακάτω θα αναλυθούν διάφορες τεχνικές των δύο αυτών κατηγοριών.

3.1 Καθαρισμός δεδομένων

Ο καθαρισμός των δεδομένων είναι μία πολύ σημαντική διαδικασία η οποία καθορίζει σε μεγάλο βαθμό το πόσο ορθολογική και εύκολη θα είναι η πορεία της ανάλυσης καθώς και την αξιοπιστία των αποτελεσμάτων τα οποία θα εξαχθούν. Υπάρχουν πολλοί μέθοδοι που χρησιμοποιούνται ώστε να παραχθούν "καθαρά" δεδομένα . Στην περίπτωση που δεν ακολουθηθεί η διαδικασία του καθαρισμού των δεδομένων υπάρχει μεγάλη πιθανότητα παραχθούν συμπεράσματα τα οποία να μην εξυπηρετούν τους σκοπούς της έρευνας και έτσι να οδηγηθούμε σε αντίθετα αποτελέσματα από αυτά για τα οποία πραγματοποιήθηκε η έρευνα. Μερικές από αυτές τις μεθόδους θα προσπαθήσουμε να αναπτύξουμε θεωρητικά παρακάτω και κάποιες θα χρησιμοποιηθούν στην ανάλυση που θα ακολουθήσει στα επόμενα κεφάλαια.

Τα δύο βασικά χαρακτηριστικά που αναζητά κάθε ερευνητής από τα δεδομένα είναι να διευκολύνουν την ροή των εργασιών που θα ακολουθηθούν στην έρευνα και να είναι ποιοτικά. Ποιοτικά θεωρούνται τα δεδομένα τα οποία χαρακτηρίζονται από εγκυρότητα, ακρίβεια, συνέπεια, πληρότητα και ομοιομορφία, ενώ όσον αφορά την διευκόλυνση της έρευνας χρησιμοποιούνται διάφορες τεχνικές για την εύρεση και αντιμετώπιση τυχών προβληματικών υποθέσεων, όπως για παράδειγμα οι ελλιπείς τιμές, μέσα στο σύνολο των δεδομένων.

Για την ποιότητα των δεδομένων είναι λίγα τα πράγματα που μπορούν να γίνουν αν η συλλογή αυτών δεν γίνεται από τον ίδιο τον ερευνητή, όπως στην περίπτωση των δεδομένων για την συγκεκριμένη διπλωματική εργασία, και λαμβάνονται από κάποιον πάροχο δεδομένων, το οποίο είναι και το πιο σύνηθες τα τελευταία χρόνια. Χρειάζεται πολύ μεγάλος όγκος δεδομένων ώστε να θεωρηθεί αξιόπιστη μία επιστημονική έρευνα, κάτι που κάνει την χρήση συμβατικών τρόπων συλλογής δεδομένων με ερωτηματολόγια κ.λ.π. ουσιαστικά αδύνατη. Θα πρέπει λοιπόν ο εκάστοτε ερευνητής να αξιολογεί τα παραπάνω πέντε βασικά προ απαιτούμενα για τα δεδομένα του καθώς να διασφαλίζει ότι η πηγή των δεδομένων είναι αξιόπιστη και τα δεδομένα του ποιοτικά.

Για την εγκυρότητα των δεδομένων ο ερευνητής πρέπει να ελέγξει κατά πόσο τα δεδομένα του ταιριάζουν με τους σκοπούς της έρευνας που θέλει να εκπονήσει ή τους περιορισμούς που μπορεί να υπάρχουν. Θα πρέπει να ελεγχθεί κατά πόσο οι τιμές των δεδομένων είναι οι προβλεπόμενες με βάση τον τύπο της εκάστοτε μεταβλητής, π.χ. η ημερομηνία και η ώρα πρέπει να έχουν συγκεκριμένο εύρος ψηφίων. Θα πρέπει να ελεγχθεί το κατά πόσο υπάρχουν κενές (ελλιπείς) τιμές και ειδικά σε βασικές μεταβλητές οι οποίες δεν μπορούν να αντικατασταθούν με μία εκτίμηση. Υπάρχουν περιπτώσεις που κάποιες μεταβλητές μπορούν να πάρουν συγκεκριμένες τιμές όπως π.χ. το φύλο ενός ανθρώπου. Οι τιμές στην συγκεκριμένη μεταβλητή πρέπει να περιορίζονται στο "άνδρας", "γυναίκα" ή έστω να υπάρχει και η επιλογή "χωρίς φύλο".

Για την ακρίβεια των δεδομένων ο ερευνητής πρέπει να ελέγξει το κατά πόσο τα δεδομένα που παρέχονται είναι όσο πιο κοντά γίνεται στις πραγματικές τιμές. Το γεγονός ότι τα δεδομένα είναι έγκυρα δεν σημαίνει ότι είναι και ακριβή. Για παράδειγμα στην ερώτηση χρώμα ματιών μπορούμε να βρούμε σαν απάντηση το "μπλε", που είναι μία έγκυρη τιμή αλλά δεν είναι ακριβής σε περίπτωση που το άτομο που αναφερόμαστε έχει "καφέ" μάτια. Όπως γίνεται κατανοητό και από το παράδειγμα η ακρίβεια των δεδομένων είναι δύσκολο να ελεγχθεί και γι' αυτό εμπιστευόμαστε τον πάροχο των δεδομένων.

Τα δεδομένα θα πρέπει να είναι συνεπή και ανάμεσα στις εγγραφές ενός συνόλου δεδομένων αλλά και σε περίπτωση που διασταυρωθούν με ένα άλλο σύνολο δεδομένων που ενδεχομένως περιέχουν παρόμοια πληροφορία. Θα πρέπει να είναι πλήρη και σε περίπτωση που δεν είναι να αντιμετωπίζεται. Επίσης θα πρέπει να είναι ομοιόμορφα δηλαδή για παράδειγμα αν υπάρχει μία μέτρηση θα πρέπει σε όλο το σύνολο δεδομένων να υπάρχει η ίδια μονάδα μέτρησης.

Στη συνέχεια θα ασχοληθούμε με τις διεργασίες που χρειάζεται να γίνουν ώστε να είναι πιο ομαλή η ροή των εργασιών που θα ακολουθηθούν στην έρευνα. Οι διεργασίες αυτές χωρίζονται σε τρεις βασικούς πυλώνες. Τον έλεγχο των δεδομένων, τον καθαρισμό τους και την επαλήθευση. Στην αρχή γίνεται ένας έλεγχος για το κατά πόσο τα δεδομένα είναι ολοκληρωμένα και δεν περιέχουν λάθη. Στη συνέχεια γίνεται ο καθαρισμός των δεδομένων και τέλος γίνεται η επαλήθευση ότι οι τεχνικές που ακολουθήθηκαν στα παραπάνω βήματα ήταν αποδοτικές και έφεραν τα επιθυμητά αποτελέσματα. Συνήθως στο τέλος των παραπάνω διαδικασιών γίνεται και μια μικρή αναφορά στις αλλαγές που έγιναν και στην ποιότητα των τελικών δεδομένων που παράχθηκαν και θα χρησιμοποιηθούν για την ανάλυση.

Οι τεχνικές που χρησιμοποιούνται για τον έλεγχο των δεδομένων είναι η εξαγωγή διαφόρων στατιστικών μέτρων για τα δεδομένα και η γραφική αναπαράσταση αυτών. Τα διάφορα μέτρα που θα επιλεγούν μπορούν να δώσουν μία πολύ καλή εικόνα στον ερευνητή για την ποιότητα των δεδομένων. Για παράδειγμα δεν μπορεί μία μεταβλητή που μετράει ύψος να έχει αρνητική μέση τιμή. Η γραφική αναπαράσταση των δεδομένων είναι η πιο συνηθισμένη μέθοδος για τον έλεγχο των

δεδομένων καθώς είναι εύκολα αντιληπτή από το ευρύ κοινό. Επίσης μέσω αυτής είναι εύκολη η εύρεση ακραίων τιμών οι οποίες δεν είναι απαραίτητο να εξαιρεθούν από το μοντέλο σε πολλές περιπτώσεις αλλά χρήζουν περεταίρω έρευνας.

Η διαδικασία του καθαρισμού των δεδομένων θεωρείται η πιο σημαντική διαδικασία. Σε κάποιες περιπτώσεις εφαρμόζεται το συγκεκριμένο βήμα χωρίς να έχουν προηγηθεί τα βήματα που έχουμε αναφέρει παραπάνω. Αυτό μπορεί να επιφέρει τα αντίθετα από τα επιθυμητά αποτελέσματα καθώς αν δεν γνωρίζεις τα προβλήματα και την φύση των δεδομένων είναι δύσκολο να χρησιμοποιήσεις την σωστή μέθοδο καθαρισμού για τα δεδομένα σου.

Πρώτο και κύριο βήμα είναι ο καθαρισμός της βάσης δεδομένων από τυχόν αχρείαστα δεδομένα. Μόνο σε περίπτωση που ο ερευνητής είναι σίγουρος ότι κάποια πληροφορία δεν είναι σημαντική πρέπει να την διαγράψει. Στην συνέχεια διαγράφονται οι διπλοεγγραφές. Διπλές εγγραφές είναι εύκολο να συμβούν σε περίπτωση που συνδυαστούν δεδομένα από δύο διαφορετικές πηγές. Το πιο σύνηθες είναι να υπάρχουν διπλές εγγραφές στη μεταβλητή που υποδηλώνει την ταυτότητα της κάθε εγγραφής στο σύνολο των δεδομένων.

Στη συνέχεια γίνεται έλεγχος των τιμών που περιέχονται στις μεταβλητές και σβήνονται τυχών έξτρα κενά ή ειδικοί χαρακτήρες που δεν μπορεί να επεξεργαστεί το μοντέλο. Καλό είναι, όσες μεταβλητές πρέπει να έχουν συγκεκριμένο εύρος και μεταβλητές που περιέχουν συγκεκριμένες τιμές, να επεξεργαστούν και να διορθωθούν σε περίπτωση που περιέχουν κάτι διαφορετικό από το προβλεπόμενο.

Μία επιλογή ακόμα είναι η μετατροπή κάποιων μεταβλητών έτσι ώστε να βρίσκονται μέσα σε ένα συγκεκριμένο εύρος. Αυτή η μέθοδος είναι αρκετά διαδεδομένη σε περίπτωση που υπάρχουν δεδομένα που χρησιμοποιούν διαφορετική μονάδα μέτρησης. Σε περίπτωση που έχουμε για παράδειγμα βαθμολογίες γραπτών με άριστα το 20 και άλλες βαθμολογίες με άριστα το 10 εφαρμόζεται η παραπάνω μέθοδος σε μία από τις δύο μεταβλητές και έτσι έχουμε την ίδια μονάδα μέτρησης και στις δύο μεταβλητές. Επίσης υπάρχει η δυνατότητα της κανονικοποίησης των δεδομένων δηλαδή η αφαίρεση της μέσης τιμής και η διαίρεση με την τυπική απόκλιση. Αυτή η μέθοδος επιλέγεται στην περίπτωση που το μοντέλο που θα χρησιμοποιηθεί βασίζεται στην κανονικότητα των δεδομένων. Αν εφαρμοστεί η μέθοδος της κανονικοποίησης γίνεται και μετατροπή των δεδομένων έτσι ώστε να βρίσκονται μέσα σε ένα συγκεκριμένο εύρος. Το εύρος τιμών που θα παραχθεί εξαρτάται από την μέθοδο κανονικοποίησης που θα χρησιμοποιηθεί.

Τελευταίο και πολύ σημαντικό βήμα είναι η αντιμετώπιση των ελλιπών τιμών. Ο πιο εύκολος τρόπος αντιμετώπισης είναι η διαγραφή των παρατηρήσεων που περιέχουν κενά. Είναι μία πολύ γρήγορη μέθοδος αλλά χάνεται αρκετή πληροφορία από τα δεδομένα και σε περίπτωση που το σύνολο των δεδομένων περιέχει αρκετές ελλιπείς τιμές δεν γίνεται να εφαρμοστεί.

Δεύτερη και αρκετά διαδεδομένη μέθοδος είναι η συμπλήρωση των ελλιπών τιμών με τιμές που θα υπολογίσει ο ερευνητής. Η συμπλήρωση γίνεται με την εισαγωγή κάποιου στατιστικού μέτρου όπως π.χ. η μέση τιμή ή η διάμεσος της μεταβλητής, είτε με την χρήση μία γραμμικής παλινδρόμησης χρησιμοποιώντας το γεγονός ότι η συγκεκριμένη παρατήρηση περιέχει πληροφορία απλά όχι για αυτή την μεταβλητή.

Αρκετοί διαφωνούν με την μέθοδο της συμπλήρωσης καθώς πιστεύουν ότι όποια μέθοδος και να χρησιμοποιηθεί δεν ανακτάται η χαμένη πληροφορία λόγω της ελλιπής τιμής και ότι με την συμπλήρωση απλά επαναλαμβάνεται το μοτίβο που ήδη έχουν τα δεδομένα. Αυτό είναι αρκετά σημαντικό στην περίπτωση που οι ελλιπείς τιμές δεν συμβαίνουν τυχαία. Για παράδειγμα όταν σε μία έρευνά υπάρχει μία φυλή αρνείται να απαντήσει σε μία συγκεκριμένη ερώτηση συστηματικά δεν μπορούν να συμπληρωθούν τα δεδομένα με την μέση τιμή των υπολοίπων φυλών. Ο τρόπος αντιμετώπισης που προτείνεται είναι η επισήμανσή των ελλιπών τιμών έτσι ώστε να μην ληφθούν υπόψιν κατά την εφαρμογή του μοντέλου. Η επισήμανση μπορεί α γίνει με την τιμή “0” για παράδειγμα αλλά χρειάζεται προσοχή καθώς το “0” δεν μπορεί να επισημαίνει τις ελλιπείς τιμές και παράλληλα να είναι και πιθανή τιμή της μεταβλητής.

Αφού πραγματοποιηθούν όλα τα παραπάνω βήματα ο ερευνητής καλείται να ελέγξει πως οι μέθοδοι που εφάρμοσε είχαν τα επιθυμητά αποτελέσματα. Για παράδειγμά μπορεί μετά την συμπλήρωση των ελλιπών τιμών να παραβιάζεται κάποιος κανόνας η περιορισμός των δεδομένων. Υπάρχει περίπτωση να χρειαστεί κάποια διόρθωση με το χέρι αν δεν μπορεί αν διορθωθεί με άλλον τρόπο.

Τέλος καλό είναι να παραχθεί μία αναφορά η οποία να συνοψίζει την ποιότητα των τελικών δεδομένων. Αυτό βοηθάει αρκετά τον αναλυτή να σιγουρευτεί ότι η τελική μορφή των δεδομένων είναι η επιθυμητή. Συνήθως χρησιμοποιούνται διάφορα γραφήματα ώστε να αποτυπωθούν τα δεδομένα και να επαληθευτεί οπτικά ότι δεν υπάρχει κάποιο πρόβλημα.

3.2 Εύρεση και αντιμετώπιση ακραίων τιμών

Οι ακραίες τιμές στα δεδομένα είναι ένα από τα μεγαλύτερα προβλήματα που αντιμετωπίζει ένας ερευνητής κατά την ανάλυση. Ο λόγος ύπαρξης τους μπορεί να διαφέρει. Μπορεί να είναι αποτέλεσμα μίας ακραίας συμπεριφοράς κατά τη διαδικασία συλλογής των δεδομένων, μπορεί να είναι μία λάθος εγγραφή είτε μπορεί να έχει προέλθει από την συμπλήρωση μίας ελλιπούς τιμής, με λάθος τρόπο, και έτσι η συμπλήρωση να μην αντικατοπτρίζει την πραγματικότητα. Το σίγουρο είναι ότι οι τιμές αυτές πρέπει να βρεθούνε και να αντιμετωπιστούνε, διαφορετικά τα αποτελέσματα της έρευνας δεν θα είναι αξιόπιστα.

Ο πρώτος ορισμός της ακραίας τιμής, αν και όχι τόσο ακριβής, έγινε από τον Hawkins (1980) και είναι ο εξής:

> *«Ακραία τιμή είναι μία παρατήρηση που αποκλίνει τόσο πολύ από τις άλλες παρατηρήσεις ώστε να μας εγείρει τις υποψίες ότι δημιουργήθηκε από διαφορετικό μηχανισμό.»*

Ο παραπάνω ορισμός έχει σαν παραδοχή ότι οι τυπικές, μη ακραίες, τιμές παράγονται μέσω ενός προκαθορισμένου μηχανισμού, κάτι που στην πραγματικότητα δεν είναι πάντα αληθές.

Υπάρχουν αρκετοί τρόποι εύρεσης των ακραίων τιμών, οι οποίοι μπορούν να κατηγοριοποιηθούν με βάση το μοντέλο στο οποίο βασίζεται η λειτουργεία τους. Δεν μπορούν να χρησιμοποιηθούν όλοι οι τρόποι σε όλες τις περιπτώσεις. Υπάρχουν τεχνικές οι οποίες δεν είναι λειτουργικές με βάση την φύση των δεδομένων, τις διαστάσεις τους αλλά και το πλήθος τους. Μία ακόμα σημαντική παράμετρος είναι η ευκολία στην ερμηνεία των αποτελεσμάτων που παράγονται. Είναι σημαντικό για έναν ερευνητή να γνωρίζει για ποιο λόγο μια παρατήρηση είναι ακραία έτσι ώστε να μπορεί να αξιολογήσει τα αποτελέσματα. Το μεγαλύτερο αρνητικό της εφαρμογής μίας τέτοιας μεθόδου είναι οι ενδεχόμενοι μετασχηματισμοί που μπορεί να χρειαστούν κάτι που δυσκολεύει την ερμηνεία του μοντέλου που θα χρησιμοποιηθεί αργότερα και γι' αυτό είναι σημαντικό να εφαρμοστεί η κατάλληλη μέθοδος σε κάθε σύνολο δεδομένων έτσι ώστε να αποφευχθούν περιττοί μετασχηματισμοί.

Οι ακραίες τιμές μπορούν να χωρισθούν σε δύο μεγάλες κατηγορίες. Τις μεμονωμένες ακραίες τιμές καθώς και τις συλλογικές ακραίες τιμές. Η πρώτη περίπτωση, που είναι και η πιο συνηθισμένη, αναφέρεται σε μεμονωμένες περιπτώσεις παρατηρήσεων που θεωρούνται ακραίες σε σύγκριση με τα υπόλοιπα δεδομένα. Οι συλλογικές ακραίες τιμές αναφέρονται στην ύπαρξη συνόλων (ομάδων) δεδομένων που αποκλίνουν σημαντικά από το υπόλοιπο σύνολο των δεδομένων. Είναι σημαντικό να ελεγχθεί η ύπαρξη τέτοιων ομάδων καθώς οι μεμονωμένες παρατηρήσεις τέτοιων ομάδων είναι πολύ πιθανό να μην αναγνωριστούν ως ακραίες τιμές.

Η πιο απλή μέθοδος εύρεσης των ακραίων τιμών είναι με την χρήση περιγραφικής στατιστικής. Μέσω διαφόρων γραφικών αναπαραστάσεων των δεδομένων σε μία, δύο ή και παραπάνω διαστάσεις προσπαθούν να προσδιοριστούν οι τιμές που συμπεριφέρονται με διαφορετικό τρόπο από το υπόλοιπο σύνολο δεδομένων. Κυρίως χρησιμοποιούνται τα ιστογράμματα και τα θηκογράμματα καθώς είναι ευρέως γνωστά και εύκολα στην κατανόηση τους.

- Ιστόγραμμα είναι η γραφική απεικόνιση στατιστικών συχνοτήτων περιοχών τιμών. Σχηματίζεται από παρακείμενα ορθογώνια. Η επιφάνεια κάθε ορθογωνίου είναι μέτρο της συχνότητας εμφάνισης της συγκεκριμένης περιοχής τιμών ενώ το ύψος του ισούται με το λόγο της συχνότητας προς το εύρος των τιμών που αντιπροσωπεύει το ορθογώνιο

- Θηκόγραμμα (box-plot) είναι ένας γραφικός τρόπος παρουσίασης πέντε περιληπτικών μέτρων μιας κατανομής δεδομένων, της διαμέσου, του πρώτου και τρίτου τεταρτημόριου και των δύο οριακών τιμών

Εφόσον κατασκευαστούν τα παραπάνω γραφήματα για τις μεταβλητές του μοντέλου, στη συνέχεια παρατηρούμε τυχόν τιμές που βρίσκονται εκτός των οριακών τιμών του θηκογράμματος. Οι οριακές τιμές του θηκογράμματος ορίζονται ως εξής:

$$x_{3/4} + 1.5\,(x_{3/4} - x_{1/4})$$

$$x_{1/4} - 1.5\,(x_{3/4} - x_{1/4})$$

όπου $\chi_{3/4}$ και $\chi_{1/4}$ είναι το τρίτο και πρώτο τεταρτημόριο των δεδομένων αντίστοιχα. Τα παραπάνω γραφήματα φτιάχνονται συνήθως για να έχει ο ερευνητής μία εικόνα των δεδομένων και δεν λαμβάνεται κάποια απόφαση για την αντιμετώπιση των ακραίων τιμών μέσω αυτών.

Μια μέθοδος που σχετίζεται αρκετά με το θηκόγραμμα είναι η μέθοδος των τεταρτημόριων. Είναι η πιο απλή μη παραμετρική μέθοδος αντιμετώπισης των ακραίων τιμών η οποία λειτουργεί για μία διάσταση και σε περίπτωση που υπάρχουν παραπάνω μεταβλητές εφαρμόζεται σε κάθε μία, ή στις κυριότερες, ξεχωριστά. Για οποιοδήποτε σημείο χ_i ισχύει:

$$x_i > x_{3/4} + k\left(x_{3/4} - x_{1/4}\right) \text{ ή } x_i < x_{1/4} - k\,(x_{3/4} - x_{1/4})$$

Η μέθοδος το χαρακτηρίζει ως ακραία τιμή, όπου k είναι μία τιμή που επιλέγει ο ερευνητής με βάση το πόσο αυστηρά θέλει να ορίσει τις ακραίες τιμές. Αν επιλεγεί το k να είναι 1,5 τότε ακραία τιμή είναι ότι βρίσκεται εκτός των οριακών τιμών του θηκοράμματος. Είναι εφαρμόσιμη σχεδόν σε όλα τα αριθμητικά δεδομένα και γι' αυτό επιλέγεται αρκετά συχνά.

Άλλος ένας τρόπος αντιμετώπισης είναι μέσω του *Z*-σκορ (*Z*-score). Η μέθοδος *Z*-σκορ εφαρμόζεται και αυτή σε μονοδιάστατα δεδομένα και είναι μία παραμετρική μέθοδος καθώς χρησιμοποιεί την μέση τιμή και την τυπική απόκλιση της μεταβλητής. Για κάθε παρατήρηση δημιουργείται μία τιμή η οποία υπολογίζεται ως εξής:

$$z = \frac{x - \mu}{\sigma}$$

και μετρά ουσιαστικά το πόσες τυπικές αποκλίσεις απέχει μία παρατήρηση από την μέση τιμή. Η μέθοδος πρέπει να εφαρμοστεί αφού έχουν γίνει τυχόν μετασχηματισμοί στα δεδομένα. Πριν την εφαρμογή της μεθόδου θα πρέπει ο ερευνητής να έχει αποφασίσει από ποιες τιμές του z και πάνω θεωρείται μία τιμή ως ακραία. Συνήθως χρησιμοποιείται μία τιμή ανάμεσα στο 2,5 και 3,5. Τα μειονεκτήματα της μεθόδου είναι ότι δεν λειτουργεί καλά σε πολυδιάστατα και πολύ μεγάλα δεδομένα και επίσης

δεν αποδίδει καλά αν τα δεδομένα δεν ακολουθούν κάποια κανονική κατανομή. Είναι πάρα πολύ αποδοτική μέθοδος στην περίπτωση που τα δεδομένα έχουν λίγες διαστάσεις και είναι αρκετά εύκολη στην εφαρμογή της. Το πρόβλημα της κανονικής κατανομής μπορεί να λυθεί μέσω την κανονικοποίησης των δεδομένων έτσι ώστε να μπορεί να εφαρμοστεί η μέθοδος, δηλαδή τον μετασχηματισμό της κατανομής ώστε να έχει παρόμοιες στατιστικές ιδιότητες με την κανονική κατανομή.

Όπως αναφέραμε και πιο πάνω, άλλος ένας τρόπος για την αντιμετώπιση των ακραίων τιμών είναι μέσω του αλγορίθμου DBSCAN. Είναι μια μέθοδος συσταδοποίησης αλλά κατά την διαδικασία υλοποίησης της εντοπίζει τις ακραίες τιμές. Δημιουργεί ομάδες δεδομένων με βάση την πυκνότητά τους και σαν αποτέλεσμα όσες παρατηρήσεις δεν πληρούν τα κριτήρια ώστε να ενταχθούν σε μία ομάδα θεωρούνται ως ακραίες τιμές. Δεν θα επεκταθούμε περισσότερο στην συγκεκριμένη μέθοδο και δεν θα αναλύσουμε τον τρόπο λειτουργίας της καθώς έχει γίνει εκτενής αναφορά στην προηγούμενη ενότητα. Το πλεονέκτημα της μεθόδου όσον αφορά την εύρεση ακραίων τιμών και όχι για την συσταδοποίηση είναι ότι η μέθοδος αποδίδει αρκετά καλά σε μεγάλες διαστάσεις. Παράλληλα είναι και μειονέκτημα της μεθόδου όμως αφού για να λειτουργήσει πρέπει να εφαρμοστεί και κάποια μέθοδος μείωσης των διαστάσεων.

Η αντιμετώπιση των ακραίων τιμών γίνεται κυρίως με τρεις τρόπους. Είτε διαγράφονται από το σύνολο των δεδομένων, είτε σημειώνονται (flag), μόνο σε αλγόριθμους που υπάρχει η δυνατότητα, ώστε να λαμβάνονται υπόψιν αλλά με μικρότερη σημαντικότητα από τις μη ακραίες τιμές στην λήψη των αποφάσεων. Συνήθως επιλέγεται η μέθοδος της διαγραφής καθώς είναι πολύ πιο εύκολα υλοποιήσιμη και είναι πολύ λίγοι οι αλγόριθμοι που μπορούν να χρησιμοποιηθούν σημειωμένα (flagged) δεδομένα. Τέλος υπάρχουν και τεχνικές αντικατάστασης των ακραίων τιμών με κάποια αντιπροσωπευτική τιμή όπως η μέση τιμή ή η διάμεσος.

4. Ομαδοποίηση πελατών και επεξεργασία δεδομένων

4.1 Περιγραφή του προβλήματος

Η μεγάλη πλειονότητα των σούπερ μάρκετ έχει κατανοήσει την ανάγκη για χωρισμό των πελατών τους σε κατηγορίες (ομάδες), έτσι ώστε να διαχειρίζεται κάθε κατηγορία με διαφορετικό τρόπο. Ο διαχωρισμός αυτός έχει πάρα πολλά οφέλη. Το κύριο πλεονέκτημα είναι η διαφορετική προσέγγιση όσον αφορά την διαφημιστική καμπάνια που επιλέγεται για κάθε κατηγορία καταναλωτών. Για παράδειγμα, ένα σούπερ μάρκετ που αποφασίζει να δημιουργήσει μία διαφημιστική εκστρατεία για προϊόντα παιδικής φροντίδας είναι λογικό να θέλει να προβάλει αυτή την προωθητική ενέργεια σε άτομα που έγιναν πρόσφατα γονείς. Αυτό μπορεί να εντοπισθεί μέσω της αγοραστικής συνήθειας συγκεκριμένων καταναλωτών οι οποίοι για κάποιο χρονικό διάστημα πριν την έρευνα θα έχουν αρχίσει να προσθέτουν στο καλάθι τους προϊόντα παιδικής φροντίδας. Με τον τρόπο αυτό μπορεί επίσης να κατευθύνει τις προσφορές στις αντίστοιχες ομάδες ανθρώπων που πιστεύει ότι θα τις χρησιμοποιήσουν.

Η παρούσα ενότητα της διπλωματικής έχει ως σκοπό την εφαρμογή τεχνικών συσταδοποίησης σε πραγματικά δεδομένα και συγκεκριμένα στον χωρισμό πελατών σε ομάδες. Τα δεδομένα για αυτή την εργασία αφορούν αγορές καταναλωτών από σούπερ μάρκετ (αποδείξεις συναλλαγών) και ως εκ τούτου ο σκοπός της εργασίας είναι ο χωρισμός των πελατών σε ομάδες με βάση τις αγοραστικές τους συνήθειες και της συμπεριφοράς τους σαν καταναλωτές. Το σύνολο των δεδομένων αποτελείται από 451.943 εγγραφές, που αφορούν αγορές 5130 νοικοκυριών για 2 χρόνια, και 12 μεταβλητές για την κάθε εγγραφή οι οποίες είναι οι εξής: αριθμός νοικοκυριού, αριθμός καταστήματος, ώρα συναλλαγής, ημερομηνία συναλλαγής, αριθμός συναλλαγής, κόστος συναλλαγής, αριθμός προϊόντων, αριθμός κατηγοριών προϊόντων, αριθμός διαφορετικών μαρκών προϊόντων, αριθμός διαφορετικών προϊόντων, αριθμός προϊόντων σε κάποια μορφή προωθητικής ενέργειας και αριθμός προϊόντων ιδιωτικής ετικέτας.

Η τεχνική που χρησιμοποιείται συνήθως σε αυτά τα προβλήματα είναι η μέθοδος *k*-means καθώς με βάση προγενέστερες μελέτες φαίνεται να είναι η πιο αποδοτική μέθοδος και σε θέμα χρόνου αλλά και στα αποτελέσματα της ομαδοποίησης. Θα εφαρμόσουμε λοιπόν την συγκεκριμένη τεχνική και θα την αξιολογήσουμε, μόνο μέσω στατιστικών τεχνικών αφού δεν υπάρχει η δυνατότητα να αξιολογηθεί στην καθημερινότητα μίας επιχείρησης λιανικού εμπορίου. Επίσης θα εφαρμόσουμε άλλες τρείς μεθόδους, Ιεραρχική, DBSCAN και Fuzzy *C*-means και θα προσπαθήσουμε να συγκρίνουμε τις συγκεκριμένες τεχνικές ώστε να αξιολογήσουμε ποια είναι η πιο αποδοτική.

Η ανάλυση των δεδομένων θα πραγματοποιηθεί με χρήση της γλώσσας προγραμματισμού *R* καθώς είναι η καταλληλότερη γλώσσα προγραμματισμού για

στατιστική ανάλυση αφού διαθέτει τις περισσότερες βιβλιοθήκες για αυτόν τον σκοπό σε σύγκρισή με οποιαδήποτε άλλη γλώσσα προγραμματισμού.

4.2 Προ-επεξεργασία των δεδομένων

Τα δεδομένα για την παρούσα διπλωματική εργασία παραχωρήθηκαν από μία εταιρία στατιστικών ερευνών στα πλαίσια της συνεργασίας όσον αφορά την εκπόνηση της διπλωματικής αλλά και το θέμα της εργασίας. Τα δεδομένα στην αρχική τους μορφή δεν περιείχαν ελλιπείς τιμές καθώς η αντιμετώπισή τους είχε γίνει από την ίδια την εταιρία. Ως αποτέλεσμα παρακάτω θα εφαρμοστούν τεχνικές μόνο για την διόρθωση κάποιων μεταβλητών που δεν είχαν την κατάλληλη μορφή καθώς και για την εύρεση ακραίων τιμών στο σύνολο των δεδομένων.

4.2.1 Διόρθωση – Δημιουργία μεταβλητών

Το πρώτο που παρατηρήθηκε κατά το άνοιγμα των δεδομένων ήταν πως η ώρα και η ημερομηνία των δεδομένων δεν ήταν "σωστή". Όπως φαίνεται παρακάτω στην περίπτωση της ώρας η μεταβλητή περιείχε εγγραφές με άνω και κάτω τελείες καθώς και εγγραφές που ήταν απλοί αριθμοί, ενώ στην περίπτωση της ημερομηνίας όλες οι εγγραφές ήταν καταχωρημένες σαν αριθμοί και η σειρά γραφής της χρονολογίας, του μήνα και της ημέρας δεν ήταν η συνηθισμένη για τα Ελληνικά δεδομένα.

```
> data[c(1,3),3:4]
   transaction_time transaction_date
1:         18:42:00         20170227
2:            90100         20180111
```

Εικόνα 5: Μορφή ώρας και ημερομηνίας πριν την επεξεργασία

Χρησιμοποιώντας εντολές όπως οι *sprintf*, όπου κρατάει συγκεκριμένο εύρος ψηφίων και *substring*, όπου χωρίζει μία ακολουθία αριθμών σε μέρη φέραμε την μορφή της ώρας σε "HH:MM:SS" και της ημερομηνία σε "DD-MM-YYYY".

```
> data[c(1,3),3:4]
   transaction_time transaction_date
1:         18:42:00       27/02/2017
2:         09:01:00       11/01/2018
```

Εικόνα 6: Μορφή ώρας και ημερομηνίας μετά την επεξεργασία

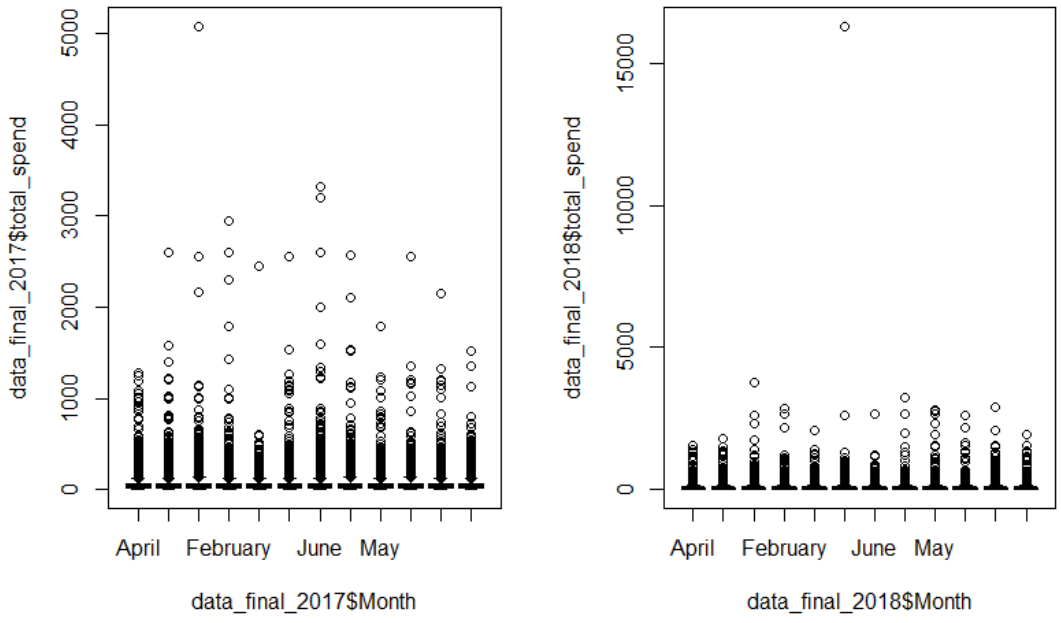
Στη συνέχεια δημιουργήθηκε η μεταβλητή TXn_gap η οποία για κάθε εγγραφή υπολογίζει τον χρόνο που πέρασε από την προηγούμενη αγορά του εκάστοτε πελάτη με την απλή αφαίρεση των ημερομηνιών δύο διαδοχικών εγγραφών για κάθε πελάτη. Η συγκεκριμένη μεταβλητή δημιουργήθηκε για να μετρηθεί με κάποιο τρόπο

η αφοσίωση των πελατών στην συγκεκριμένη εταιρία αλλά θα χρησιμοποιηθεί και ως κριτήριο για το ποιοι πελάτες θα χρησιμοποιηθούν στην ανάλυση.

4.2.2 Εύρεση Ακραίων τιμών – Ομαδοποίηση εγγραφών

Πριν την εφαρμογή των τεχνικών εύρεσης ακραίων τιμών προτάθηκαν από την εταιρία και εφαρμόστηκαν δύο κανόνες για την διαγραφή συγκεκριμένων πελατών, οι οποίοι θεωρήθηκαν πως δεν εξυπηρετούν στην ανάλυση αφού δεν θεωρούνται πελάτες με βάση την αγοραστική τους συμπεριφορά. Οι δύο αυτοί κανόνες είναι η διαγραφή όσων πελατών εμφανίζονται κάτω από τρεις φορές στο σύνολο των δεδομένων, δηλαδή μέχρι δύο αγορές στα δύο χρόνια των δεδομένων, και όσων πελατών για τους οποίους υπήρξε διάστημα μεγαλύτερο του ενός χρόνου χωρίς καμία αγορά. Τα κριτήρια αυτά θεωρήθηκαν επαρκή ώστε να διαγραφούν οι αντίστοιχες εγγραφές. Για την διαγραφή των δεδομένων έγινε χρήση της SQL καθώς και εντολές της βιβλιοθήκης data.table της R. Από την εφαρμογή των δύο αυτών κανόνων αφαιρέθηκαν 317 νοικοκυριά από το σύνολο των δεδομένων.

Στη συνέχεια δημιουργήσαμε τα θηκογράμματα για την μεταβλητή *total spend* (κόστος συναλλαγής) για κάθε έτος και κάθε μήνα ξεχωριστά:



Διάγραμμα 1: Θηκογράμματα της μεταβλητής κόστους ανά μήνα πριν την εφαρμογή μεθόδων εύρεσης ακραίων τιμών

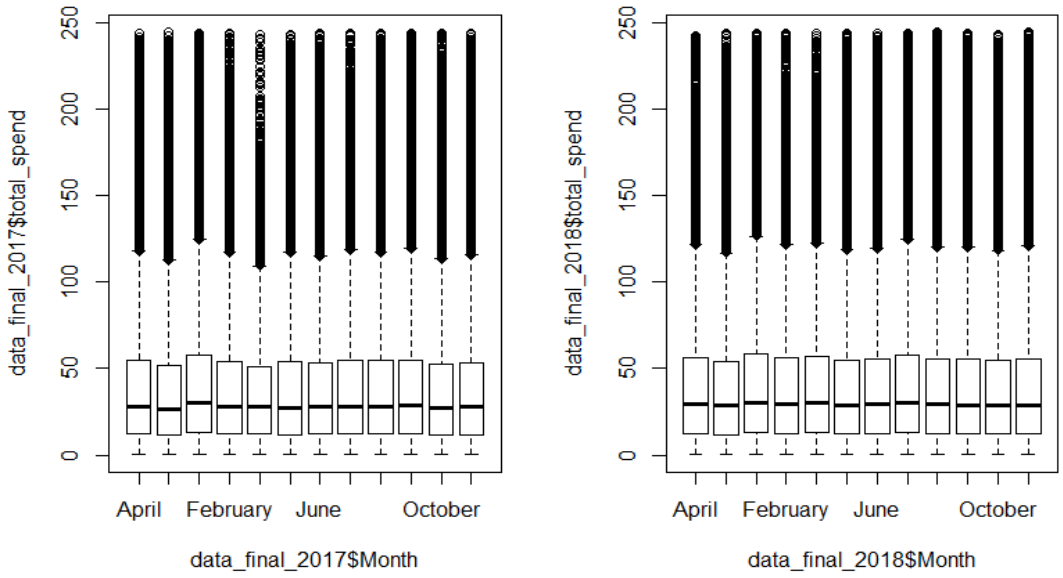
Όπως παρατηρείται η εικόνα των δεδομένων δεν είναι πολύ καλή καθώς υπάρχουν αρκετές τιμές εκτός των οριακών τιμών του θηκογράμματος. Οι τιμές αυτές ή τουλάχιστον το μεγαλύτερο μέρος αυτών πρέπει να θεωρηθούν ως ακραίες τιμές.

Η τεχνική που θα εφαρμοστεί για τον εντοπισμό των ακραίων τιμών είναι η μέθοδος των τεταρτημόριων που, όπως αναφέραμε παραπάνω σχετίζεται με τα θηκογράμματα που παρουσιάστηκαν παραπάνω. Οι συνήθεις τιμές του k για την συγκεκριμένη μέθοδο είναι από 2,5 έως 3,5 αλλά στην συγκεκριμένη περίπτωση το σύνολο τιμών των μεταβλητών δεν περιέχει αρνητικές τιμές οπότε θα χρησιμοποιηθούν μεγαλύτερες τιμές. Αυτό το κάνουμε καθώς δεν υπάρχουν ακραίες τιμές στο κάτω μέρος του θηκογράμματος.

Η μέθοδος των τεταρτημόριων όπως αναφέραμε παραπάνω εφαρμόζεται σε μονοδιάστατα δεδομένα, άρα δεν γίνεται να εφαρμοστεί μαζικά σε όλο το σύνολο δεδομένων. Αυτό που θα γίνει είναι να εφαρμοστεί σε κάθε μεταβλητή ξεχωριστά και στη συνέχεια συνδυάζοντας τις τιμές που παρήχθησαν ως ακραίες τιμές για κάθε μεταβλητή θα σημειώσουμε τα αντίστοιχα δεδομένα ως ακραίες τιμές. Οι τιμές του k που θα χρησιμοποιηθούν είναι 4 για τις δύο βασικότερες μεταβλητές, συνολικό κόστος και αριθμός προϊόντων, και 4.5 για όλες τις υπόλοιπες μεταβλητές. Αυτή η διαφοροποίηση γίνεται καθώς οι δύο συγκεκριμένες μεταβλητές περιέχουν κατά κάποιο τρόπο μαζεμένο ένα μέρος της πληροφορίας που περιέχεται στις υπόλοιπες μεταβλητές. Γι' αυτόν τον λόγο είναι δύσκολο να χαρακτηριστεί μία παρατήρηση ως ακραία και έτσι επιλέξαμε να εφαρμοστούν πιο αυστηρά κριτήρια για τις συγκεκριμένες μεταβλητές.

Για την μεταβλητή κόστος ανά μονάδα προϊόντος εφαρμόστηκε η ίδια τεχνική (θηκόγραμμα) και τα αποτελέσματα ήταν να επισημανθούν σαν ακραίες τιμές πάνω από το 15% του συνόλου των δεδομένων. Αυτό συνέβη γιατί υπήρχε πάρα πολύ μεγάλη συγκέντρωση παρατηρήσεων ανάμεσα στις τιμές 2 και 4 (πάνω από 70%) με αποτέλεσμα το 1ο και 3ο τεταρτημόριο να βρίσκονται πολύ κοντά και να μην δουλεύει καλά η συγκεκριμένη μέθοδος. Στην περίπτωση αυτή με βάση την εμπειρία και σχετική πληροφορία για τα πιο ακριβά και πιο φθηνά προϊόντα σε σούπερ μάρκετ προσδιορίσαμε σαν ακραία οποιαδήποτε τιμή ήταν πάνω από 70 ή κάτω από 0,05.

Η αντιμετώπιση των ακραίων τιμών έγινε με διαγραφή των παρατηρήσεων από το σύνολο των δεδομένων αφού δεν υπάρχει η δυνατότητα στις μεθόδους που θα εφαρμοστούν στην συνέχεια να επισημανθούν συγκεκριμένες παρατηρήσεις ως ακραίες. Στη συνέχεια εμφανίζονται τα ίδια θηκογράμματα που παρουσιάστηκαν παραπάνω μετά την διαγραφή των ακραίων τιμών.



Διάγραμμα 2: Θηκογράμματα της μεταβλητής κόστους ανά μήνα μετά την εφαρμογή μεθόδων εύρεσης ακραίων τιμών

Όπως φαίνεται και από τα θηκογράμματα η εικόνα των δεδομένων είναι εμφανώς καλύτερη και έχουν διαγραφεί αρκετές τιμές οι οποίες έκαναν τα γραφήματα να μην μπορούν να διαβαστούν. Μία τελευταία διαδικασία πριν ξεκινήσουμε την συσταδοποίηση των δεδομένων ήταν η συγχώνευσή τους. Ομαδοποιήσαμε τις εγγραφές του συνόλου δεδομένων από το επίπεδο των αποδείξεων σε επίπεδο νοικοκυριού καθώς η συσταδοποίηση θα γίνει σε επίπεδο νοικοκυριού. Άλλο ένα θετικό της διαδικασίας αυτής ήταν η αισθητή μείωση του απαιτούμενου χρόνου για να τρέξουν οι παρακάτω διαδικασίες. Η συγχώνευση έγινε με βάση την διάμεση τιμή, δηλαδή για κάθε νοικοκυριό και για κάθε μεταβλητή χρησιμοποιήσαμε την διάμεσο όλων των αντίστοιχων εγγραφών στο σύνολο των εγγραφών. Χρησιμοποιήθηκε η διάμεσος καθώς στην βιβλιογραφία παρουσιάζεται σαν το πιο αντιπροσωπευτικό μέτρο για αντίστοιχες περιπτώσεις. Το τελικό σύνολο δεδομένων περιέχει 4405 νοικοκυριά.

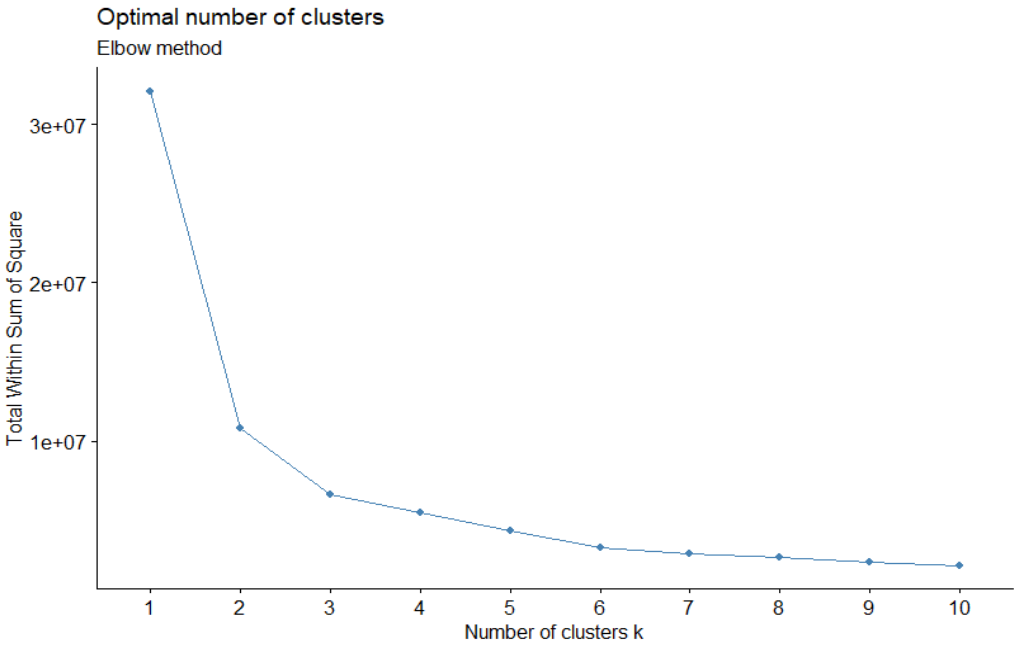
4.3 Δημιουργία ομάδων

Στην ενότητα αυτή θα γίνει εφαρμογή των μεθόδων συσταδοποίησης στα τελικά δεδομένα που παρήχθησαν παραπάνω. Όπως έχουμε αναφέρει προηγουμένως η μέθοδος *k*-means χρειάζεται σαν είσοδο τον αριθμό των ομάδων που θα δημιουργήσει. Γι' αυτόν τον λόγο πριν την εφαρμογή των μεθόδων θα προσπαθήσουμε να προσδιορίσουμε την τιμή αυτή.

4.3.1 Εύρεση βέλτιστου αριθμού συστάδων

Δυστυχώς δεν υπάρχει κάποιος συγκεκριμένος αλγόριθμος εύρεσης του βέλτιστου αριθμού συστάδων ο οποίος να λειτουργεί σε κάθε περίπτωση. Για την εύρεση του βέλτιστου αριθμού συστάδων θα εφαρμόσουμε δύο διαφορετικές μεθόδους. Για την ακρίβεια θα χρησιμοποιήσουμε την "μέθοδο του αγκώνα" (elbow method) και μία ρουτίνα της η οποία εφαρμόζει πολλές διαφορετικές μεθόδους και εμφανίζει τα αποτελέσματα για κάθε μία από αυτές και επίσης συνοψίζει και το πόσες μέθοδοι επιλέγουν τον κάθε αριθμό κλάσεων. Σε αυτό το σημείο να επισημάνουμε ότι η εταιρία που παρείχε τα δεδομένα πρότεινε την δημιουργία πέντε ομάδων αλλά επειδή η επιλογή αυτή έγινε με βάση διάφορα επιχειρησιακά κριτήρια δεν θα την ακολουθήσουμε. Χωρίς λοιπόν να γνωρίζουμε τα αποτελέσματα των μεθόδων αποφασίστηκε να δημιουργηθούν τόσες ομάδες όσες προταθούν περισσότερες φορές με βάση την ρουτίνα "NbClust" της R που θα χρησιμοποιηθεί.

Η μέθοδος του αγκώνα είναι ουσιαστικά μία γραφική παράσταση που στον οριζόντιο άξονα έχει τον αριθμό των ομάδων που πρόκειται να δημιουργηθούν και στον κάθετο έχει το άθροισμα των τετραγώνων των σφαλμάτων που παράγονται για κάθε αριθμό κλάσεων. Τα σφάλματα υπολογίζονται ως η απόστασή κάθε σημείου μίας κλάσης από το κέντρο της αντίστοιχης κλάσης. Για να παραχθεί το συγκεκριμένο γράφημα εφαρμόζονται διαδοχικές επαναλήψεις της μεθόδου *k*-means μέσα σε ένα εύρος για τις τιμές του *k* και στη συνέχεια υπολογίζεται το άθροισμα των τετραγώνων των σφαλμάτων για κάθε επανάληψη. Στην περίπτωση που το γράφημα που παράγεται μοιάζει με χέρι τότε χρησιμοποιούμε σαν *k* την τιμή που αντιστοιχεί στον αγκώνα του χεριού. Το ζητούμενο είναι να χρησιμοποιήσουμε έναν αριθμό ομάδων ώστε να ελαχιστοποιηθεί το άθροισμα των τετραγώνων των σφαλμάτων, αλλά όσο μεγαλώνει ο αριθμός των ομάδων, τόσο πέφτει το άθροισμα των τετραγώνων των σφαλμάτων. Επειδή είναι ζητούμενο και η, όσο το δυνατόν, ελαχιστοποίηση του αριθμού των ομάδων έτσι ώστε να είναι κατανοητά τα αποτελέσματα, η ιδέα πίσω από την συγκεκριμένη τεχνική είναι να επιλέγεται όσο το δυνατόν μικρότερος αριθμός ομάδων με όσο το δυνατόν μικρότερο άθροισμα σφαλμάτων. Έτσι επιλέγουμε το σημείο του γραφήματος που η κλίση της καμπύλης μειώνεται αισθητά. Εφαρμόζοντας την μέθοδο για τα δεδομένα της εργασίας το γράφημα που παράγεται είναι το εξής:



Διάγραμμα 3: Διάγραμμα της μεθόδου «elbow method»

Στο συγκεκριμένο γράφημα οι δύο πιο λογικές τιμές για τον αριθμό των συστάδων είναι το δύο και το τρία. Παρότι από το σημείο δύο στο τρία η κλίση είναι πολύ μικρότερη από αυτή ανάμεσα στο ένα και στο δύο βλέπουμε ότι υπάρχει αρκετή μείωση του αθροίσματος και στην πρώτη περίπτωση και γι' αυτόν τον λόγο αν χρησιμοποιούσαμε αυτή την μέθοδο θα επιλέγαμε να χωρίσουμε τα δεδομένα σε τρεις ομάδες.

Στη συνέχεια εφαρμόσαμε την εντολή «NbClust" της R η οποία στην ουσία εφαρμόζει είκοσι έξι διαφορετικές μεθόδους προσδιορισμού του αριθμού των συστάδων όπως οι Hartigan method, Silhouette κλπ και εμφανίζει τα αποτελέσματα κάθε μεθόδου. Η εντολή που εφαρμόστηκε ήταν η εξής:

```
NbClust(data = data_final[, -c(1,9)], diss = NULL, distance = "euclidean", min.nc = 2, max.nc = 7, method = "kmeans")
```

Η ρουτίνα ζητά από τον χρήστη να υποδείξει τις τιμές του αριθμού των συστάδων που θα ελεγχθούν. Επιλέχθηκε να χρησιμοποιηθούν τιμές από το δύο ως το εφτά. Για κάθε μέθοδο και κάθε τιμή του αριθμού των συστάδων υπολογίζεται ο αντίστοιχος δείκτης και με βάση τα κριτήρια του κάθε δείκτη επιλέγεται η βέλτιστη τιμή του αριθμού των συστάδων. Τέλος παράγεται ένας πίνακας που συνοψίζει το πόσες από τις μεθόδους προτείνουν κάθε αριθμό συστάδων. Τα αποτελέσματα στην συγκεκριμένη περίπτωση εμφανίζονται στον παρακάτω πίνακα.

```
Among all indices:
7 proposed 2 as the best number of clusters
11 proposed 3 as the best number of clusters
3 proposed 4 as the best number of clusters
2 proposed 7 as the best number of clusters
```

Εικόνα 7: Αποτελέσματα της ρουτίνας "NbClust" για την εύρεση του βέλτιστου αριθμού συστάδων για τα δεδομένα

Όπως φαίνεται οι περισσότεροι δείκτες προτείνουν την χρήση τριών συστάδων κάτι που επαληθεύει και τα αποτελέσματα της μεθόδου του αγκώνα που χρησιμοποιήσαμε νωρίτερα. Τελικά ο αριθμός των συστάδων που θα χρησιμοποιήσουμε για να χωρίσουμε τα δεδομένα μας θα είναι τρεις.

Στη συνέχεια θα παρουσιάσουμε πιο αναλυτικά τρεις δείκτες που χρησιμοποίησε η συγκεκριμένη ρουτίνα για να παραχθούν τα συγκεκριμένα αποτελέσματα. Θα παρουσιάσουμε τους δύο δείκτες Silhouette και Ch index που έχουμε αναφέρει παραπάνω αλλά και τον δείκτη Scott. Είναι πιθανό να μην προτείνουν και οι τρεις μέθοδοι σαν βέλτιστο αριθμό συστάδων τον αριθμό 3.

Πρώτο θα δούμε τον δείκτη Silhouette. Όπως έχουμε αναφέρει και παραπάνω ο τύπος υπολογισμού του δείκτη είναι:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}},$$

όπου $a(i)$ είναι η μέση απόσταση του σημείου i με όλα τα υπόλοιπα σημεία μέσα στην ίδια ομάδα και $b(i)$ είναι η ελάχιστη μέση απόσταση του σημείου i με όλα τα σημεία κάθε άλλης κλάσης. Στη συνέχεια υπολογίζεται ο μέσος όρος για κάθε ομάδα και ο τελικός δείκτης βρίσκεται από τον μέσο όρο των δεικτών των ομάδων.

Clustrers	Silhouette
2	0.6195
3	0.5212
4	0.4747
5	0.4261
6	0.3972
7	0.3760

Εικόνα 8: Αποτελέσματα δείκτη Silhouette για την εύρεση του βέλτιστου αριθμού συστάδων

Οι τιμές που αναζητάμε για τον συγκεκριμένο δείκτη είναι τιμές όσο πιο κοντά στο 1 γίνεται, οπότε σε αυτήν την περίπτωση ο δείκτης Silhouette προτείνει την δημιουργία 2 ομάδων για τα συγκεκριμένα δεδομένα.

Στη συνέχεια θα δούμε τον δείκτη Ch index. Ο τύπος του δείκτη είναι:

$$CH(k) = \frac{B(k)}{W(k)} \times \frac{n-k}{k-1},$$

όπου $W(k)$ είναι η διακύμανση μέσα στις κλάσεις, $B(k)$ είναι η διακύμανση ανάμεσα στις κλάσεις, n είναι ο αριθμός των σημείων του συνόλου δεδομένων και k είναι ο αριθμός των κλάσεων που επιλέγεται να δημιουργηθούν.

Clustrers	CH
2	6565.649
3	7711.533
4	8697.676
5	8659.578
6	8509.888
7	8419.805

Εικόνα 9: Αποτελέσματα δείκτη CH Index για την εύρεση του βέλτιστου αριθμού συστάδων

Αναζητούμε την μέγιστη τιμή του δείκτη για να προσδιορίσουμε τον βέλτιστο αριθμό συστάδων οπότε σε αυτήν την περίπτωση ο δείκτης προτείνει την δημιουργία 4 ομάδων για τα συγκεκριμένα δεδομένα.

Τέλος θα αναφέρουμε τον δείκτη των Scott και Symons. Ο συγκεκριμένος δείκτης υπολογίζεται μέσω του τύπου:

$$Scott(k) = n \times \log(\frac{|T(k)|}{|W(k)|}) ,$$

όπου *n* είναι το πλήθος των δεδομένων του συνόλου των δεδομένων, *T(k)* είναι η διακύμανση όλων των δεδομένων και *W(k)* είναι η διακύμανση μέσα στις κλάσεις. Στη συνέχεια υπολογίζεται η διαφορά του δείκτη που παράγεται για κάθε τιμή του αριθμού των ομάδων και από την προηγούμενη τιμή του αριθμού ομάδων.

Clustrers	Scott
2	10105.40
3	13212.38
4	15749.60
5	17468.57
6	18851.99
7	20181.45

Εικόνα 10: Αποτελέσματα δείκτη Scott και Symons για την εύρεση του βέλτιστου αριθμού συστάδων

Και σε αυτήν την περίπτωση αναζητάμε την μέγιστη τιμή του δείκτη για την επιλογή του βέλτιστου αριθμού των ομάδων. Βλέπουμε ότι η διαφορά ανάμεσα στον αριθμό 2 και 3 είναι 3106,98 που είναι η μέγιστη, άρα ο δείκτης προτείνει την δημιουργία 3 ομάδων για τα συγκεκριμένα δεδομένα.

4.3.2 Εφαρμογή της μεθόδου *k*-means

Η εφαρμογή της μεθόδου έγινε με την χρήση της εντολής "kmeans" της R για την δημιουργία τριών ομάδων στις παρατηρήσεις. Οι ομάδες που παρήχθησαν από τον αλγόριθμο είχαν μέγεθος 1121, 527 και 2757 αντίστοιχα. Παρακάτω παραθέτουμε τις μέσες τιμές των παρατηρήσεων ανά μεταβλητή ανά ομάδα έτσι ώστε να επιχειρήσουμε να προσδιορίσουμε τα χαρακτηριστικά της κάθε ομάδας.

Cluster means:

	unique_categories	unique_brands	unique_upcs	unique_promo_items	total_units	unique_private_label	total_spend	max_gap
1	5.747101	5.198037	7.206512	1.983051	9.30107	1.535236	32.10023	129.34166
2	5.706831	5.100569	7.206831	1.993359	9.40038	1.569260	32.18717	259.43833
3	6.588683	5.775481	8.155604	2.067465	10.29960	1.693689	37.36058	34.01197

Εικόνα 11: Μέσες τιμές των μεταβλητών ανά ομάδα μετά την εφαρμογή της μεθόδου *k*-means

Όπως βλέπουμε η εμφανής διαφορά εντοπίζεται στην μεταβλητή "max_gap" η οποία υπολογίζει την μεγαλύτερη διάρκεια σε ημέρες που έχει να επισκεφθεί ένας πελάτης κάποιο κατάστημα. Με βάση την συγκεκριμένη μεταβλητή η τρίτη ομάδα θα μπορούσε να χαρακτηριστεί ως η ομάδα με τους συχνούς πελάτες η οποία με βάση το

μέγεθος είναι και αυτή με τους περισσότερους πελάτες. Η δεύτερη ομάδα, η οποία αριθμεί και τους λιγότερους πελάτες, θα μπορούσε να αποτελείται από ανθρώπους που επισκέπτονται τα καταστήματα μόνο όταν υπάρχει ανάγκη, ενώ στην πραγματικότητα είναι συχνοί πελάτες σε κάποιο άλλο σούπερ μάρκετ. Η πρώτη ομάδα στην συγκεκριμένη μεταβλητή έχει μέση τιμή η οποία βρίσκεται ενδιάμεσα στις άλλες δύο ομάδες κάτι που θα μπορούσε να σημαίνει ότι περιέχει συχνούς πελάτες οι οποίοι για κάποιο λόγο δεν εμπιστεύονται απόλυτα την συγκεκριμένη αλυσίδα ή άτομα τα οποία χρησιμοποιούν συχνά παραπάνω από ένα σούπερ μάρκετ για τις αγορές τους.

Άλλο ένα στοιχείο που φαίνεται να ξεχωρίζει στα παραπάνω αποτελέσματα είναι το γεγονός ότι η μεταβλητή που δηλώνει το πόσα ξοδεύει ο κάθε καταναλωτής σε κάθε επίσκεψη είναι μεγαλύτερη για την ομάδα των συχνών πελατών. Αυτό είναι λογικό και αναμενόμενο καθώς συχνός πελάτης σημαίνει ότι πραγματοποιεί κάθε φορά όλα του τα ψώνια σε ένα συγκεκριμένο κατάστημα και σαν αποτέλεσμα έχει να δημιουργεί μεγάλα καλάθια αγορών. Επίσης βλέπουμε ότι οι συχνοί πελάτες δεν έχουν τα καλύτερα στατιστικά μόνο στο συνολικό κόστος των αγορών αλλά και γενικότερα σε όλες τις μεταβλητές του μοντέλου μας, με μικρότερη βέβαια απόκλιση απ' ότι στο συνολικό κόστος.

4.3.3 Εφαρμογή των Ιεραρχικών μεθόδων

Η ιεραρχική μέθοδος εφαρμόστηκε με τρεις διαφορετικούς τρόπους, τον συσσωρευτικό, τον διαιρετικό και τον ενισχυμένο. Η ενισχυμένη μέθοδος που θα χρησιμοποιήσουμε έχει δύο βασικές διαφορές με τους δύο άλλους αλγόριθμους. Η πρώτη διαφορά είναι ότι μετατρέπει τις μεταβλητές έτσι ώστε να βρίσκονται μέσα σε ένα συγκεκριμένο εύρος. Με αυτόν τον τρόπο επιχειρείται να περιοριστούν οι ακραία μεγάλες αποστάσεις μεταξύ των παρατηρήσεων. Η δεύτερη και κυριότερη διαφορά του ενισχυμένου αλγόριθμού είναι ότι ο πίνακας αποστάσεων που χρησιμοποιείται δημιουργείται μόνο μία φορά στην αρχή και σε κάθε επανάληψη του αλγόριθμου απλά ανανεώνεται μόνο για τα δεδομένα ή τις ομάδες δεδομένων που ομαδοποιούνται. Αυτό βοηθάει αρκετά στον χρόνο εκτέλεσης του αλγορίθμου καθώς μειώνονται δραματικά οι υπολογισμοί που απαιτούνται.

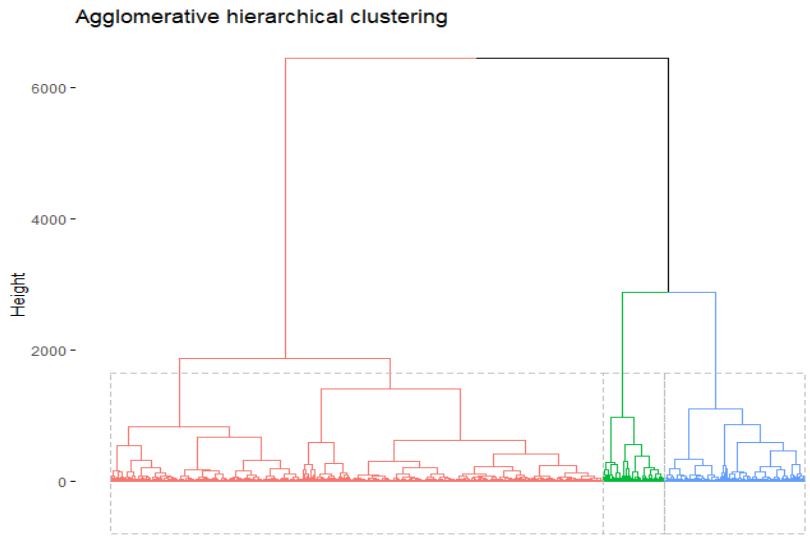
Η μέθοδος σύνδεσης που θα χρησιμοποιηθεί καθορίζει τον τρόπο με τον οποίο υπολογίζονται οι αποστάσεις τον ομάδων σε βάση τις αποστάσεις κατά ζεύγη μεταξύ των παρατηρήσεων. Για την επιλογή της δημιουργήσαμε μία ρουτίνα που υπολογίζει έναν συντελεστή για κάθε μία μέθοδο σύνδεσης. Ο συντελεστής αυτός μετρά το κατά πόσο η δομή των δεδομένων υποστηρίζει την δημιουργία ομάδων με βάση την εκάστοτε μέθοδο σύνδεσης. Τιμές του συντελεστή που προσεγγίζουν το ένα είναι επιθυμητές. Τα αποτελέσματα της ρουτίνας παρατίθενται παρακάτω:

```
> map_dbl(m, ac)
  average    single  complete      ward
0.9693546 0.8810644 0.9840703 0.9978895
```

Εικόνα 12: Τιμές των συντελεστών ανά μέθοδο σύνδεσης

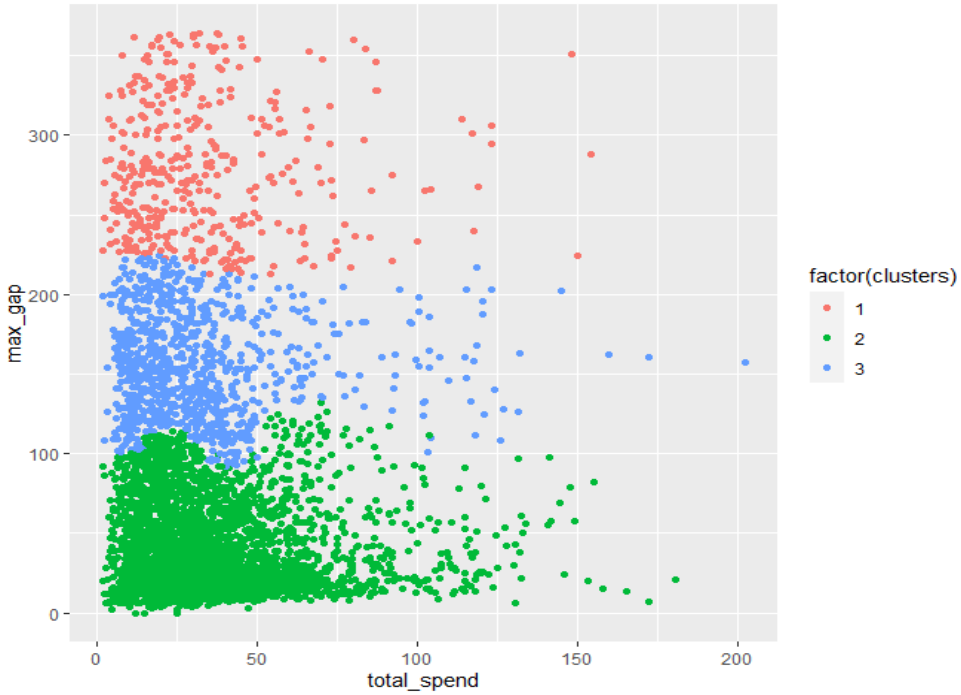
Όπως είναι φανερό η μέθοδος σύνδεσης του "Ward" είναι η πλέον κατάλληλη για τα συγκεκριμένα δεδομένα και είναι αυτή που θα χρησιμοποιήσουμε και στις τρεις μεθόδους ιεραρχικής ομαδοποίησης που θα εφαρμόσουμε.

Στη συνέχεια εφαρμόζουμε τον συσσωρευτικό αλγόριθμο του οποίου τα αποτελέσματα εμφανίζονται παρακάτω:



Διάγραμμα 4: Δενδρόγραμμα της συσσωρευτικής ιεραρχικής μεθόδου.

Όπως φαίνεται και από το παραπάνω γράφημα, η επιλογή των τριών κλάσεων για τα δεδομένα μας είναι μία αρκετά καλή επιλογή. Επειδή το παραπάνω γράφημα δεν μας δίνει πολλές πληροφορίες για την ομοιογένεια των ομάδων θα προσπαθήσουμε να παρουσιάσουμε διαγραμματικά την διαίρεση των ομάδων σε συνάρτηση με τις μεταβλητές που χρησιμοποιήθηκαν. Για τον σκοπό αυτό παρουσιάζουμε παρακάτω το διάγραμμα των μεταβλητών "συνολικό κόστος αγορών" με την μεταβλητή "μέγιστο διάστημα χωρίς καμία αγορά από τον πελάτη". Με διαφορετικό χρώμα έχουν επισημανθεί οι παρατηρήσεις που έχουν συσταδοποιηθεί σε κάθε διαφορετική κλάση.



Διάγραμμα 5: Διάγραμμα των μεταβλητών συνολικό κόστος αγορών με το μέγιστο διάστημα χωρίς καμία αγορά από τον πελάτη.

Είναι φανερό πως και σε αυτή την περίπτωση η μεταβλητή του μέγιστου διαστήματος χωρίς καμία αγορά από τον πελάτη έπαιξε πολύ σημαντικό ρόλο στην δημιουργία των κλάσεων αφού οι κλάσεις είναι σχεδόν τέλεια διαχωρισμένες με βάση την συγκεκριμένη μεταβλητή. Διαισθητικά και σε αυτή την περίπτωση, όπως και στην περίπτωση του *k*-means, φαίνεται πως η ομάδα με τους πιο συχνούς πελάτες είναι αυτή με το μεγαλύτερο μέγεθος, δεύτερη έρχεται η ομάδα που περιέχει συχνούς πελάτες οι οποίοι για κάποιο λόγο δεν εμπιστεύονται απόλυτα την συγκεκριμένη αλυσίδα ή άτομα τα οποία χρησιμοποιούν παραπάνω από ένα σούπερ μάρκετ για τις αγορές τους και το μικρότερο πλήθος παρατηρείται στην ομάδα που αποτελείται από ανθρώπους που επισκέπτονται τα καταστήματα μόνο όταν υπάρχει ανάγκη. Το γεγονός πως οι δύο μέθοδοι εμφανίζουν παρόμοια συμπεριφορά, τουλάχιστον στα σημεία που ελέγχθηκαν, είναι αρκετά ενθαρρυντικό αφού και τα πραγματικά δεδομένα φαίνεται να συμπεριφέρονται με τον ίδιο τρόπο.

Παρακάτω παραθέτουμε, όπως και στην περίπτωση της μεθόδου *k*-means, τις μέσες τιμές των παρατηρήσεων ανά ομάδα για κάθε μεταβλητή έτσι ώστε να παρατηρήσουμε αν και κατά πόσο διαφέρουν τα αποτελέσματα των δύο μεθόδων.

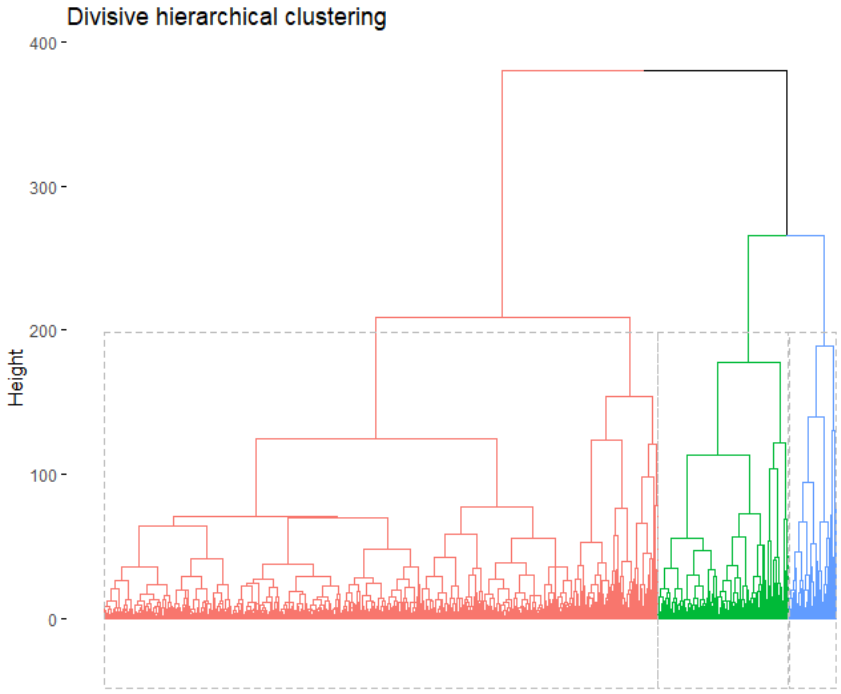
unique_categories	unique_brands	unique_upcs	unique_promo_items	total_units	unique_private_label	total_spend	max_gap
5.910256	5.251282	7.461538	2.055128	9.748718	1.650000	33.50984	278.1051
6.489430	5.708040	8.054132	2.060698	10.186259	1.680493	36.74919	41.2255
5.655095	5.117021	7.062150	1.946809	9.152296	1.486562	31.52329	154.8936

Εικόνα 9: Μέσες τιμές των μεταβλητών ανά ομάδα μετά την εφαρμογή της συσσωρευτικής ιεραρχικής μεθόδου

Όπως φαίνεται οι μέσοι των ομάδων ακολουθούν την ίδια λογική με τους συχνούς πελάτες να ξοδεύουν τα περισσότερα αλλά και να έχουν το μεγαλύτερο

καλάθι ανάμεσα στις τρεις ομάδες, με μικρότερες όμως διαφορές απ' ότι στην περίπτωση του *k*-means. Η μεγαλύτερη διαφορά ανάμεσα στις δύο μεθόδους παρατηρείται στο μέγεθος των ομάδων που δημιουργούνται με την συσσωρευτική ιεραρχική μέθοδο να περιέχει 3122 πελάτες στην ομάδα με τους συχνούς πελάτες, δηλαδή σχεδόν 450 πελάτες παραπάνω κάτι που σημαίνει ότι αφήνει την μικρότερη ομάδα με σχεδόν το μισό πλήθος πελάτων.

Στη συνέχεια θα εφαρμόσουμε την διαιρετική μέθοδο χωρίς να αναμένουμε μεγάλες διαφορές από τα αποτελέσματα της συσσωρευτικής μεθόδου. Ακολουθεί το δενδρόγραμμα της μεθόδου:



Διάγραμμα 6: Δενδρόγραμμα της διαιρετικής ιεραρχικής μεθόδου.

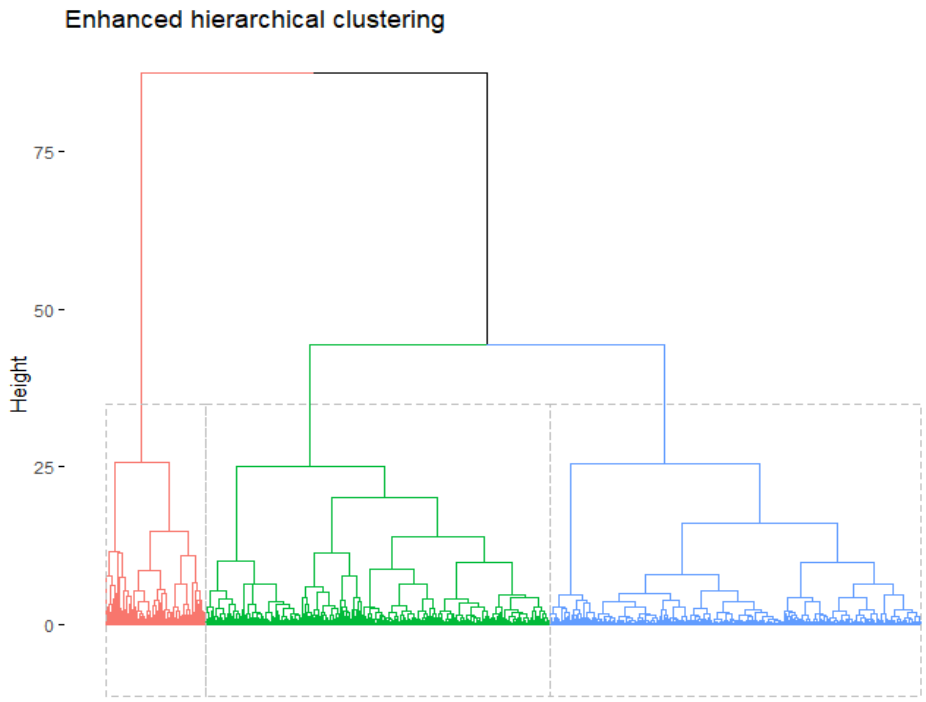
Για άλλη μία φορά θα δημιουργήσουμε τον πίνακα με τις μέσες τιμές των παρατηρήσεων ανά ομάδα για κάθε μεταβλητή έτσι ώστε να συγκρίνουμε τα αποτελέσματα των μεθόδων.

unique_categories	unique_brands	unique_upcs	unique_promo_items	total_units	unique_private_label	total_spend	max_gap
6.473446	5.734456	8.053109	2.108808	10.350389	1.678109	37.45434	45.41839
5.644828	5.137931	7.120690	1.900000	9.448276	1.586207	32.79003	169.64138
5.554217	4.861446	6.891566	1.939759	9.367470	1.620482	31.61458	289.54217

Εικόνα 13: Μέσες τιμές των μεταβλητών ανά ομάδα μετά την εφαρμογή της διαιρετικής ιεραρχικής μεθόδου

Και σε αυτήν την περίπτωση το πλήθος των πελατών που απαρτίζουν την ομάδα των συχνών πελατών είναι πολύ μεγαλύτερη σε σχέση με την μέθοδο *k*-means. Στη συγκεκριμένη μέθοδο το πλήθος της πρώτης ομάδας είναι 3401 πελάτες, της δεύτερης, 639 και της τρίτης 365. Όπως αναμενόταν η τάση σχεδόν όλων των μέσων των ομάδων διατηρείται ίδια ανάμεσα σε όλες τις μεθόδους.

Τέλος θα εφαρμόσουμε την ενισχυμένη ιεραρχική μέθοδο η οποία θα διαφέρει από την συσσωρευτική μέθοδο μόνο στο ότι τα δεδομένα θα είναι σε συγκεκριμένο εύρος και δεν θα είναι τα αρχικά δεδομένα. Το δενδρόγραμμα της μεθόδου εμφανίζεται παρακάτω:



Διάγραμμα 7: Δενδρόγραμμα της ενισχυμένης ιεραρχικής μεθόδου.

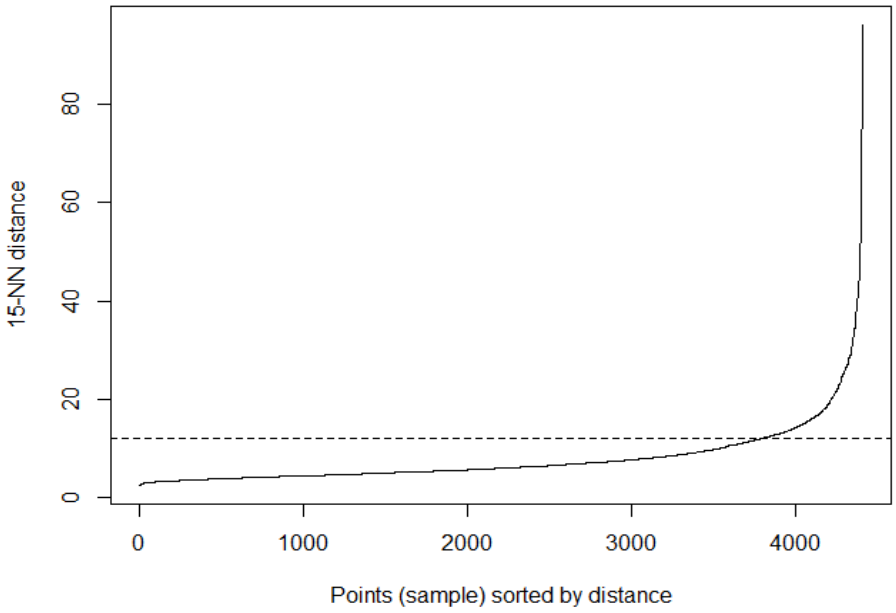
Όπως αναμέναμε η μετατροπή των δεδομένων δεν έπαιξε σημαντικό ρόλο στην δημιουργία των ομάδων και έτσι τα αποτελέσματα είναι σχεδόν ίδια και δεν θεωρήθηκε σημαντικό να παρουσιαστούν.

4.3.4 Εφαρμογή της μεθόδου DBSCAN

Όπως έχουμε αναφέρει και στην θεωρητική παρουσίαση του αλγορίθμου DBSCAN, απαιτούνται δύο τιμές σαν είσοδος στον αλγόριθμο. Η μία είναι το MinPts και η άλλη το Eps. Η λογική για την επιλογή του MinPts είναι πως όσο μεγαλύτερο είναι ένα σύνολο δεδομένων, τόσο μεγαλύτερο MinPts πρέπει να επιλεγεί. Για τον προσδιορισμό της τιμής αυτής θα εφαρμόσουμε τον εμπειρικό κανόνα που αναφέρεται στην βιβλιογραφία, που είναι δύο φορές ο αριθμός των μεταβλητών που μοντελοποιούνται μείον ένα. Στην περίπτωσή μας χρησιμοποιούμε 8 μεταβλητές στα μοντέλα αρά θα επιλέξουμε ως MinPts την τιμή 15.

Στη συνέχεια χρειάζεται και η επιλογή της τιμής Eps. Για την επιλογή αυτή θα υπολογίσουμε την απόσταση των σημείων του συνόλου των δεδομένων από τους MinPts κοντινότερους γείτονες τους. Στη συνέχεια υπολογίζεται η μέση απόσταση για κάθε τιμή και δημιουργήσουμε το γράφημα βάζοντας τα σημεία σε σειρά από την

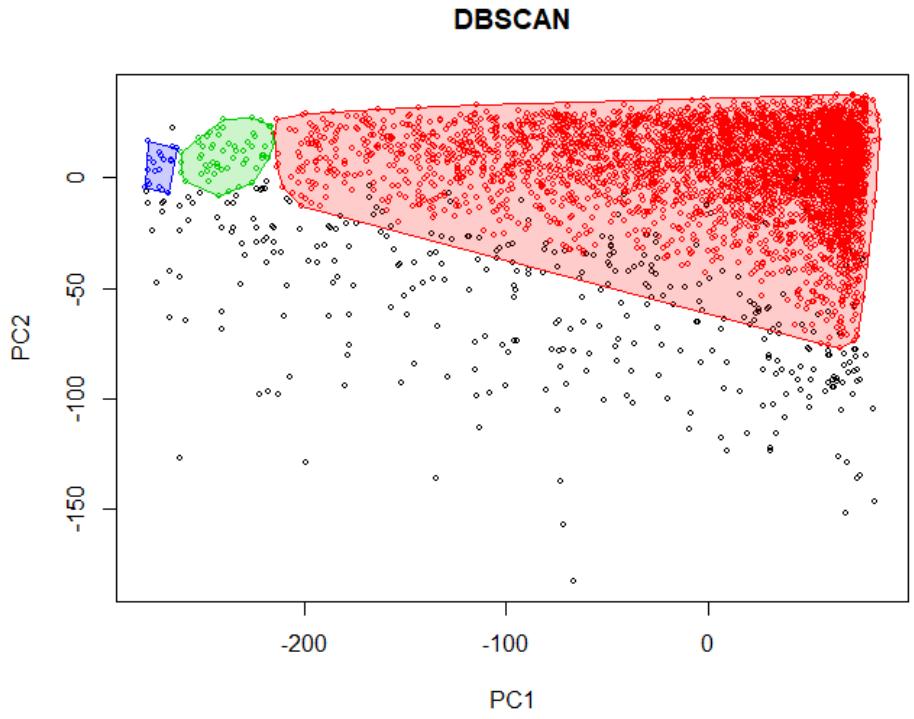
μικρότερη προς την μεγαλύτερη μέση απόσταση. Περιμένουμε να δημιουργηθεί ένα γράφημα που θα έχει την μορφή γόνατου.



Διάγραμμα 8: Διάγραμμα των μέσων των αποστάσεων των σημείων από τους MinPts κοντινότερους γείτονες τους.

Όπως παρατηρούμε όντως δημιουργείται ένα γράφημα που έχει την μορφή γόνατου. Σαν Eps επιλέγουμε την τιμή στην οποία το γράφημα αρχίζει να αλλάζει δραματικά την κλίση του. Η τιμή που επιλέξαμε με βάση το συγκεκριμένο γράφημα είναι η τιμή 12 και αυτή θα χρησιμοποιήσουμε για την εφαρμογή της μεθόδου DBSCAN.

Να τονίσουμε σε αυτό το σημείο πως, όπως έχουμε αναφέρει και στην θεωρητική περιγραφή του αλγορίθμου, η μέθοδος αυτή δεν μπορεί να λειτουργήσει ικανοποιητικά για μεγάλο αριθμό μεταβλητών. Για το λόγο αυτό η εντολή της R που χρησιμοποιήσαμε εφαρμόζει πρώτα την μέθοδο PCA, η οποία είναι μία μέθοδος μείωσης των διαστάσεων, και χρησιμοποίει τις δύο πρώτες κύριες συνιστώσες για την εφαρμογή της μεθόδου.



Διάγραμμα 9: Δισδιάστατη απεικόνιση των αποτελεσμάτων της μεθόδου DBSCAN αφού έχει προηγηθεί η εφαρμογή της μεθόδου PCA.

Εκ πρώτης όψεως, διαγραμματικά, η ομαδοποίηση που πραγματοποιείται φαίνεται να είναι καλή αφού στην παραπάνω απεικόνιση φαίνεται η κάθε ομάδα να βρίσκεται εκτός των υπολοίπων. Τα σημεία που είναι μακριά από τα κέντρα των ομάδων έχουν σημειωθεί σαν ακραίες τιμές. Το πλήθος των στοιχείων σε κάθε ομάδα είναι 4004, 55 και 19 με 327 σημεία να αποτελούν την ομάδα των ακραίων τιμών. Το γεγονός πως αυτοί οι αριθμοί είναι αρκετά διαφορετικοί από όλες τις υπόλοιπες μεθόδους που εφαρμόσαμε μέχρι στιγμής, και το γεγονός ότι η πρώτη ομάδα περιέχει το 98% των δεδομένων που τελικά μπήκαν σε μία ομάδα δημιουργούν ερωτηματικά και υποψίες πως η μέθοδος δεν έχει λειτουργήσει ικανοποιητικά για τα συγκεκριμένα δεδομένα. Για το λόγο αυτό παρουσιάζουμε παρακάτω τον πίνακα των μέσων τιμών ανά ομάδα για κάθε μεταβλητή. Όπως βλέπουμε δημιουργείται ένας πίνακας με τέσσερεις γραμμές κάτι που είναι αναμενόμενο καθώς η πρώτη γραμμή αναφέρετε στα δεδομένα που έχουν σημειωθεί σαν ακραίες τιμές.

unique_categories	unique_brands	unique_upcs	unique_promo_items	total_units	unique_private_label	total_spend	max_gap
14.966361	13.006116	19.665138	5.068807	25.264526	3.798165	86.71157	161.12844
5.601149	4.973901	6.885240	1.803072	8.749875	1.470280	31.49730	74.49076
3.636364	3.400000	4.563636	1.272727	6.290909	1.090909	18.82019	322.74545
4.947368	4.342105	5.868421	1.394737	7.078947	1.500000	23.42803	357.15789

Εικόνα 14: Μέσες τιμές των μεταβλητών ανά ομάδα μετά την εφαρμογή της μεθόδου DBSCAN

Στον παραπάνω πίνακα είναι φανερό πως οι ομάδες που δημιουργούνται είναι πολύ διαφορετικές από όλες τις υπόλοιπες μεθόδους ομαδοποίησης. Βλέπουμε ότι οι

πελάτες που ξοδεύουν τα περισσότερα χρήματα έχουν σημειωθεί ως ακραίες τιμές, ενώ οι δύο μικρότερες ομάδες αποτελούνται από άτομα για τα οποία έχει υπάρξει διάστημα απουσίας από τα καταστήματα σχεδόν ενός χρόνου. Ένα τρίτο σημείο το οποίο δεν υπάρχει σε καμία άλλη μέθοδο είναι ότι έχουν σημειωθεί αρκετά σημεία ως ακραίες τιμές. Το πλήθος τους είναι 327 δηλαδή σχεδόν το πλήθος της μικρότερης ομάδας στις προηγούμενες μεθόδους. Κύριος λόγος όμως που μας κάνει να χαρακτηρίζουμε την ομαδοποίηση ως κακή είναι το πλήθος των πελατών που υπάρχουν σε κάθε ομάδα, δημιουργώντας ουσιαστικά μία ομάδα η οποία περιλαμβάνει την συντριπτική πλειοψηφία των πελατών και οι άλλες δύο να είναι τόσο μικρές που δεν αξίζουν περεταίρω μελέτης. Συμπερασματικά θα λέγαμε ότι η μέθοδος DBSCAN δεν είναι η πλέον κατάλληλη για τα συγκεκριμένα δεδομένα και θα προσπαθήσουμε στο επόμενο κεφάλαιο να τα αναλύσουμε λίγο περισσότερο.

4.3.5 Εφαρμογή της μεθόδου Fuzzy *C*-means

Η εφαρμογή της μεθόδου έγινε με χρήση της εντολής “cmeans” της R για την δημιουργία τριών ομάδων στις παρατηρήσεις. Να τονίσουμε σε αυτό το σημείο πως και η μέθοδος “cmeans” της R κάνει χρήση της μεθόδου PCA πριν την εφαρμογή της, κάτι που αναμένουμε να διαφοροποιήσει άλλα όχι σε μεγάλο βαθμό τα αποτελέσματα της μεθόδου σε σύγκριση με τα αποτελέσματα της *k*-means καθώς οι δύο μέθοδοι είναι παρόμοιες και για τα συγκεκριμένα δεδομένα δεν περιμένουμε να έχουν μεγάλες διαφορές. Οι ομάδες που παρήχθησαν από τον αλγόριθμο είχαν μέγεθος 1213, 573 και 2619 αντίστοιχα. Όπως παρατηρούμε δεν χρησιμοποιήθηκε κάποιο κατώφλι για τα ποσοστά συμμετοχής σε κάποια ομάδα έτσι ώστε να χαρακτηρίζονται τα δεδομένα σαν ακραίες τιμές μιας και έχει γίνει εκ των προτέρων αφαίρεση των ακραίων τιμών και θέλαμε όλα τα δεδομένα να κατανεμηθούν σε μία από τις ομάδες που θα δημιουργηθούν. Παρακάτω παραθέτουμε τις μέσες τιμές των παρατηρήσεων ανά μεταβλητή ανά ομάδα έτσι ώστε να συγκρίνουμε τα αποτελέσματα της μεθόδου με τις προηγούμενες.

unique_categories	unique_brands	unique_upcs	unique_promo_items	total_units	unique_private_label	total_spend	max_gap
5.722589	5.171063	7.176834	1.964551	9.253504	1.519373	31.98708	121.12448
6.638221	5.814051	8.218977	2.082474	10.375716	1.710004	37.69543	31.72012
5.738220	5.128272	7.208551	1.983421	9.385689	1.563700	32.15610	253.90576

Εικόνα 15: Μέσες τιμές των μεταβλητών ανά ομάδα μετά την εφαρμογή της μεθόδου *C*-means

Όπως αναμέναμε τα αποτελέσματα είναι σχεδόν ίδια με αυτά της μεθόδου *k*-means χωρίς να παρατηρούμε μεγάλες διαφορές ούτε και στο πλήθος των δεδομένων που περιέχει κάθε ομάδα. Για τον παραπάνω λόγο δεν θα επεκταθούμε άλλο στην ανάλυση της συγκεκριμένης μεθόδου και θα προσπαθήσουμε στην επόμενη παράγραφο να προσδιορίζουμε κατά πόσο θα συνέφερε να την χρησιμοποιήσουμε για την ανάλυση.

4.4 Αξιολόγηση και σύγκριση των αποτελεσμάτων

Η αξιολόγηση και η σύγκριση των μεθόδων συσταδοποίησης είναι ένα πολύ δύσκολο κομμάτι της ανάλυσης, εξίσου δύσκολο με την διαδικασία της συσταδοποίηση. Υπάρχουν δύο τρόποι αντιμετώπισης του συγκεκριμένου προβλήματος. Ο πρώτος τρόπος αφορά την "εσωτερική" αξιολόγηση, δηλαδή χρησιμοποιώντας μόνο τα στοιχεία και τα αποτελέσματα που παράγει η μέθοδος συσταδοποίησης που έχει χρησιμοποιηθεί, και η "εξωτερική" αξιολόγηση όπου η τελική συσταδοποίηση συγκρίνεται με μία ήδη υπάρχουσα η οποία είναι επιστημονικά αποδεκτή ως "καλή" συσταδοποίηση.

Για την διενέργεια εσωτερικής αξιολόγησης συνήθως υπολογίζονται κάποιοι δείκτες οι οποίοι αναφέρονται στην επιστημονική βιβλιογραφία ως κριτήρια αξιολόγησης των μεθόδων συσταδοποίησης. Πολλές φορές χρησιμοποιούνται στατιστικά και δείκτες οι οποίοι έχουν χρησιμοποιηθεί και κατά την εφαρμογή της ομαδοποίησης, για παράδειγμα ο δείκτης Silhouette μπορεί να χρησιμοποιηθεί και για την δημιουργία αλλά και για την αξιολόγηση των παραγόμενων ομάδων. Άλλο ένα αρνητικό των μεθόδων εσωτερικής αξιολόγησης είναι ότι στην πραγματικότητα αξιολογείται η αποδοτικότητα του αλγόριθμου και κατά πόσο τα αποτελέσματα που παράγει είναι τα βέλτιστα και όχι το κατά πόσο οι παραγόμενες ομάδες εξυπηρετούν τους σκοπούς της έρευνας και πόσο χρήσιμες θα είναι στην επίλυση του αρχικού προβλήματος. Παρόλα αυτά χρησιμοποιούνται αρκετά συχνά καθώς δεν απαιτούν βαθιά γνώση στο αντικείμενο απ' όπου έχουν προέλθει τα δεδομένα.

Αναλυτικά εσωτερική ονομάζεται η αξιολόγηση για την οποία χρησιμοποιούνται τα ίδια δεδομένα με αυτά που χρησιμοποιήθηκαν κατά την εφαρμογή των αλγορίθμων ομαδοποίησης. Οι μέθοδοι αυτοί στις περισσότερες περιπτώσεις αξιολογούν σαν καλές ομαδοποιήσεις αυτές που παράγουν ομάδες με μεγάλη ομοιότητα στα δεδομένα κάθε κλάσης και μικρή ομοιότητα ανάμεσα στις διαφορετικές ομάδες. Όπως αναφέραμε μεγάλο μειονέκτημα αυτού είναι πως μία καλή αξιολόγηση χρησιμοποιώντας εσωτερικά κριτήρια δεν συνεπάγεται και εξαγωγή ορθών πληροφοριών από τα δεδομένα.

Απ' την άλλη πλευρά η εσωτερική αξιολόγηση είναι καλή στην περίπτωση που χρειάζεται να αξιολογήσουμε και να συγκρίνουμε την αποτελεσματικότητα δύο ή παραπάνω αλγορίθμων. Αυτό βέβαια εξαρτάται από τις μεθόδους και τα δεδομένα που έχουν χρησιμοποιηθεί σε κάθε περίπτωση. Αν ένας αλγόριθμος έχει δημιουργηθεί για ένα συγκεκριμένο τύπο δεδομένων, πχ δισδιάστατα, και χρησιμοποιείται για έναν όλο, πχ πολυδιάστατα, τα αποτελέσματα δεν γίνεται να είναι αξιόλογα. Αντίστοιχα αν ένας δείκτης εσωτερικής αξιολόγησης μετράει κάτι διαφορετικό από αυτό που στοχεύει ο αλγόριθμος δεν μπορεί να παράγει ασφαλή αποτελέσματα. Για παράδειγμα ο αλγόριθμος *k*-means βρίσκει ομάδες σε κυκλική μορφή και πολλά μέτρα εσωτερικής αξιολόγησης έχουν σαν προϋπόθεση την κυκλική μορφή των ομάδων. Σε ένα σύνολο δεδομένων που δεν περιέχει κυκλικές ομάδες ούτε η μέθοδος *k*-means αλλά ούτε και πολλά από τα κριτήρια μπορούν να έχουν καλά

αποτελέσματα. Ενδεικτικά κάποιοι δείκτες εσωτερικής αξιολόγησης είναι η συνολική διακύμανση των δεδομένων μέσα στις ομάδες, ο δείκτης silhouette και ο δείκτης CH στους οποίους έχουμε αναφερθεί σε προηγούμενα κεφάλαια.

Το βασικό πρόβλημα της εξωτερικής αξιολόγησης είναι η παραδοχή ύπαρξης αποτελεσμάτων τα οποία είναι "αντικειμενικά" και επιστημονικά "σωστά". Στην πραγματικότητα αν υπήρχαν τέτοια αποτελέσματα δεν θα υπήρχε λόγος χρήσης τεχνικών μηχανικής μάθησης χωρίς εποπτεία αλλά θα χρησιμοποιούσαμε τεχνικές όπως τα δέντρα απόφασης, τα νευρωνικά δίκτυα και οι παλινδρομήσεις.

Στην περίπτωση της εξωτερικής αξιολόγησης τα κριτήρια που χρησιμοποιούνται δεν σχετίζονται με τα δεδομένα που χρησιμοποιήθηκαν για την ομαδοποίηση των δεδομένων. Για παράδειγμα μπορεί να χρησιμοποιηθεί ένα αντίστοιχο σύνολο δεδομένων το οποίο έχει ήδη ομαδοποιηθεί και τα δεδομένα του περιέχουν για κάθε παρατήρηση την ομάδα που ανήκουν. Τα δεδομένα αυτά προέρχονται από έρευνες ατόμων που θεωρούνται ειδικοί στον αντίστοιχο τομέα της έρευνας. Αυτές οι μέθοδοι αξιολογούν το κατά πόσο η ομαδοποίηση που έχει παραχθεί είναι κοντά σε αυτήν των ειδικών. Παρ' όλα αυτά υπάρχει συζήτηση στο κατά πόσο μια ομαδοποίηση κοντά σε αυτή των ειδικών είναι επαρκής καθώς δεν είναι αντικειμενικός ο χαρακτηρισμός σαν παρόμοια τα δεδομένα που θα χρησιμοποιηθούν για να εκτελεστούν δύο διαφορετικές έρευνες. Ακόμα τέτοιες τεχνικές έχουν ως αποτέλεσμα την εξαγωγή συμπερασμάτων που είναι ήδη γνωστά κάτι το οποίο δεν είναι επιθυμητό για την εξέλιξη ενός κλάδου. Επίσης στην περίπτωση που χρησιμοποιούμε το ίδιο σύνολο δεδομένων για εκπαίδευση του μοντέλου και αξιολόγηση των αποτελεσμάτων, χάνεται μεγάλο μέρος της πληροφορίας, το κομμάτι των δεδομένων δηλαδή που χρησιμοποιείται για αξιολόγηση, αφού δεν χρησιμοποιείται στην εκπαίδευση του μοντέλου.

Υπάρχουν αρκετά μέτρα εξωτερικής αξιολόγησης. Αυτά υπολογίζονται μετά την διαδικασία της επαλήθευσης των αποτελεσμάτων. Αφού λοιπόν εφαρμόσουμε το μοντέλο μας στα ήδη ομαδοποιημένα δεδομένα εξετάζουμε πόσα από αυτά ομαδοποιήθηκαν σωστά. Άλλο ένα μειονέκτημα αυτών των μεθόδων είναι ότι προϋποθέτουν δύο και όχι παραπάνω κλάσεις δεδομένων ώστε να ταξινομήσουν σαν ορθός και λάθος ταξινομημένα τα δεδομένα σε αυτές τις ομάδες. Σε περίπτωση παραπάνω ομάδων εφαρμόζουμε τις μεθόδους εξωτερικής αξιολόγησης θεωρώντας κάθε φορά μία ομάδα και όλο το υπόλοιπο σύνολο των δεδομένων σαν μία ομάδα κάτι που δημιουργεί σύγχυση στην παρουσίαση των αποτελεσμάτων.

Για τις ανάγκες της συγκεκριμένης διπλωματικής εφαρμόσαμε δύο δείκτες εσωτερικής αξιολόγησης για την σύγκριση των αποτελεσμάτων των μεθόδων που χρησιμοποιήθηκαν. Αυτή η επιλογή έγινε για δύο βασικούς λόγους. Ο πρώτος είναι ότι δεν υπήρχε η δυνατότητα εφαρμογής εξωτερικής αξιολόγησης καθώς το σύνολο των δεδομένων που είχαμε στην διάθεσή μας δεν περιείχε προγενέστερο διαχωρισμό των δεδομένων σε ομάδες οπότε δεν υπήρχε πληροφορία ώστε να ελέγξουμε την παραγόμενη ομαδοποίηση. Ο δεύτερος λόγος είναι η αποτελεσματικότητα που έχουν

οι εσωτερικοί δείκτες αξιολόγησης στην σύγκριση αποτελεσμάτων ανάμεσα σε αλγορίθμους που χρησιμοποιούν τα ίδια δεδομένα σαν είσοδο. Οι δείκτες που χρησιμοποιήσαμε τελικά για την αξιολόγηση είναι ο δείκτης silhouette και η διακύμανση των δεδομένων μέσα στις κλάσεις που δημιουργήθηκαν. Τα αποτελέσματα παρουσιάζονται στο παρακάτω γράφημα:

	skm	shc1	shc2	sdb	scm
within.cluster.ss	" 6644936"	" 7179278"	" 7885326"	"24105611"	" 6680964"
avg.silwidth	"0.582748"	"0.5424153"	"0.552042"	"0.3856607"	"0.5115881"

Εικόνα 16: Τιμές των δεικτών εσωτερικής αξιολόγησης ανά μέθοδο

Όπως αναμέναμε και με βάση αποτελέσματα άλλων ερευνών αλλά και τα αποτελέσματα που είδαμε αναλυτικά για κάθε μέθοδο στα προηγούμενα κεφάλαια, καλύτερη μέθοδος για τα συγκεκριμένα δεδομένα φαίνεται να είναι η μέθοδος *k*-means και χειρότερη η μέθοδος DBSCAN όσον αφορά τις δύο μεθόδους αξιολόγησης που χρησιμοποιήθηκαν.

Όπως βλέπουμε η μέθοδος *k*-means παράγει την μικρότερη διακύμανση μέσα στις κλάσεις σε σύγκριση με τις υπόλοιπες μεθόδους, με την μέθοδο *C*-means να είναι πολύ κοντά, κάτι που ήταν αναμενόμενο αν αναλογιστούμε ότι οι δύο μέθοδοι έχουν σχεδόν τα ίδια τελικά αποτελέσματα. Η βασική διαφορά τους όπως έχουμε αναφέρει είναι ότι ο *C*-means δίνει σε κάθε σημείο πιθανότητες συμμετοχής σε κάθε ομάδα που δημιουργείται οι οποίες όμως, όπως είναι λογικό, είναι μεγαλύτερες στην ομάδα που ανήκει το στοιχείο και στην μέθοδο *k*-means.

Όσον αφορά τον δείκτη silhouette βλέπουμε και σε αυτήν την περίπτωση την μέθοδο *k*-means να παράγει τα καλύτερα αποτελέσματα, όμως βλέπουμε ότι στην συνέχεια έρχονται οι ιεραρχικές μέθοδοι.

Τέλος θα παρουσιάσουμε κάποιους πίνακες σύγκρισης ανάμεσα στις μεθόδους (ανά δύο) όπου θα εμφανίζεται το πλήθος των δεδομένων που ταξινομήθηκε σε κάθε ομάδα. Πρώτα θα συγκρίνουμε τις δύο μεθόδους που αναμένουμε να είναι παρόμοιες με βάση τα παραπάνω αποτελέσματα, δηλαδή την *k*-means και την Fuzzy *C*-means.

km / cm	1	2	3
1	2,619	-	-
2	138	1,075	-
3	-	46	527

Εικόνα 17: Σύγκριση ταξινόμησης για τις μεθόδους *k*-means και Fuzzy *C*-means

Όπως βλέπουμε ελάχιστες παρατηρήσεις διαφοροποιούνται ανάμεσα στις δύο μεθόδους. Αυτό επιβεβαιώνει τα αποτελέσματα που παρατηρήσαμε κατά την εφαρμογή της μεθόδου Fuzzy C-means όπου οι μέσες τιμές ανά μεταβλητή ήταν παρόμοιες με αυτές της *k*-means. Αυτό που παρατηρούμε στον παραπάνω πίνακα είναι ότι τα κριτήρια για την ένταξη μιας παρατήρησης στην ομάδα των συχνών

πελατών είναι λίγο πιο αυστηρά για την μέθοδο Fuzzy *C*-means και αυτό έχει σαν αποτέλεσμα κάποιοι πελάτες να ταξινομούνται σε πιο "κάτω" ομάδες σε σύγκριση με την μέθοδο *k*-means.

Στη συνέχεια θα παράξουμε τον ίδιο πίνακα για σύγκριση των μεθόδων *k*-means και Ιεραρχικό (συσσωρευτικό).

ha \ km	1	2	3
1	2,757	365	-
2	-	756	137
3	-	-	390

Εικόνα 18: Σύγκριση ταξινόμησης για τις μεθόδους *k*-means και Ιεραρχικό (συσσωρευτικό)

Παρατηρούμε ότι τα αποτελέσματα έχουν παρόμοια συμπεριφορά με την παραπάνω σύγκριση, με την μεγαλύτερη διαφορά να παρατηρείται στο γεγονός ότι περισσότεροι πελάτες έχουν μετατοπιστεί από την πρώτη στην δεύτερη ομάδα και από την δεύτερη στην τρίτη.

Επίσης θα συγκρίνουμε και τις μεθόδους Fuzzy *C*-means και Ιεραρχικό (συσσωρευτικό) ώστε να επιβεβαιώσουμε την εγκυρότητα των αποτελεσμάτων.

ha \ cm	1	2	3
1	2,619	503	-
2	-	710	183
3	-	-	390

Εικόνα 19: Σύγκριση ταξινόμησης για τις μεθόδους Fuzzy *C*-means και Ιεραρχικό (συσσωρευτικό)

Όπως είναι λογικό τα αποτελέσματα του παραπάνω πίνακα είναι ένας συνδυασμός των δύο προηγούμενων συγκρίσεων. Αρκετές από τις παρατηρήσεις της πρώτης ομάδας του Ιεραρχικού έχουν ταξινομηθεί στην δεύτερη του *C*-means μεγαλώνοντας έτσι αρκετά την δεύτερη ομάδα σε ποσοστό συμμετοχής στο σύνολο των δεδομένων.

Τέλος θα συγκρίνουμε την μέθοδο DBSCAN με τις υπόλοιπες τρείς που παρουσιάστηκαν παραπάνω.

DBSCAN \ km	1	2	3
0	98	108	121
1	2,659	1,013	332
2	-	-	55
3	-	-	19

DBSCAN \ ha	1	2	3
0	132	87	108
1	2,990	806	208
2	-	-	55
3	-	-	19

DBSCAN \ cm	1	2	3
0	92	109	126
1	2,527	1,104	373
2	-	-	55
3	-	-	19

Εικόνα 20: Σύγκριση ταξινόμησης για τις μεθόδους DBSCAN με τις υπόλοιπες μεθόδους

Όπως βλέπουμε στην μέθοδο DBSCAN υπάρχει η τάση να ταξινομεί όλα τα δεδομένα σε μία ομάδα. Χαρακτηριστικό είναι ότι στην σύγκριση με όλες τις υπόλοιπες μεθόδους τα στοιχεία των δύο πρώτων ομάδων έχουν ταξινομηθεί όλα

στην πρώτη ομάδα στην περίπτωση του DBSCAN ενώ πολύ λίγα δεδομένα υπάρχουν στην δεύτερη και τρίτη ομάδα. Επίσης βλέπουμε ότι ο αλγόριθμος DBSCAN έχει σημειώσει και 327 παρατηρήσεις σαν ακραίες τιμές (ομάδα 0) κάτι που δεν υπάρχει στις υπόλοιπες μεθόδους.

Σαν ένα γενικό συμπέρασμα από τους παραπάνω πίνακες είναι ότι υπάρχει μία συνέπεια στα αποτελέσματα όλων των μεθόδων αφού δεν υπάρχει μεγάλη διασταύρωση των ομάδων. Για παράδειγμα δεν υπάρχει στοιχείο, εκτός από την περίπτωση του DBSCAN, που να έχει ταξινομηθεί σε μία μέθοδο στην πρώτη ομάδα (των συχνών πελατών) και να ταξινομηθεί σε κάποια άλλη μέθοδο στην τρίτη ομάδα (των πελατών που επισκέπτονται σπάνια κάποιο κατάστημα).

Συμπερασματικά η μέθοδος που θα επιλέγαμε για την ομαδοποίηση παρόμοιων δεδομένων θα ήταν η *k*-means βάση των αποτελεσμάτων της παραπάνω έρευνας αλλά και στηριζόμενοι στις προτάσεις της παγκόσμιας επιστημονικής κοινότητας. Σε καμία περίπτωση δεν θα χρησιμοποιούσαμε την μέθοδο DBSCAN, που απ' ότι είδαμε είχε την χειρότερη απόδοση, κάτι που αναμέναμε λόγο της πολυδιάστατης μορφής του συνόλου των δεδομένων αλλά και του γεγονότος ότι ο αλγόριθμος DBSCAN βασίζεται στην πυκνότητα για τον διαχωρισμό των ομάδων.

5. Συμπεράσματα και προτάσεις για περεταίρω έρευνα

Υπάρχουν πολλά είδη έρευνας τα οποία μπορεί να εφαρμόσει κάποιος που θέλει να ασχοληθεί πιο εμπεριστατωμένα με τα δεδομένα της συγκεκριμένης διπλωματικής και να εμβαθύνει πιο πολύ στα συμπεράσματα που μπορούν να εξαχθούν από αυτά.

Μία πρώτη ιδέα είναι με την χρήση των παραπάνω αποτελεσμάτων να επιχειρήσει κάποιος την ανάπτυξη ενός νέου μοντέλου ομαδοποίησης, το οποίο ενδεχομένως να εφαρμόζεται καλύτερα στα συγκεκριμένα δεδομένα, και εφόσον θα δημιουργηθεί για τον συγκεκριμένο σκοπό θα μπορούσε να παράγει και άλλα αποτελέσματα όπως για παράδειγμα την συσχέτιση συγκεκριμένων προϊόντων με την καταναλωτική συμπεριφορά της κάθε ομάδας.

Μία άλλη ιδέα είναι ο ενδιαφερόμενος ερευνητής να χρησιμοποιήσει τις παραγόμενες ομάδες των δεδομένων και να προσπαθήσει να αποκωδικοποιήσει τις καταναλωτικές συνήθειες της κάθε ομάδας ξεχωριστά. Με αυτόν τον τρόπο δεν θα έχουμε απλά χωρισμένους τους πελάτες σε ομάδες αλλά θα παράγουμε δεδομένα για κάθε ομάδα, που σημαίνει καλύτερη κατανόηση των χαρακτηριστικών τους και ευκολότερη διαχείριση.

Αυτό θα μπορούσε να οδηγήσει σε περεταίρω έρευνα για τεχνικές διαφήμισης αλλά και διαφορετικές προωθητικές ενέργειες σε κάθε καταναλωτή με βάση την ομάδα που ανήκει. Μπορεί ο συνηθισμένος τρόπος διαφήμισης μέσω της τηλεόρασης να λειτουργεί για την μεγάλη μάζα των καταναλωτών αλλά για παράδειγμα να μην λειτουργεί στα άτομα που ξοδεύουν πολύ πάνω από τον μέσο όρο σε κάθε επίσκεψη σε ένα κατάστημα.

Τέλος μια καλή ιδέα θα ήταν η χρήση των συγκεκριμένων ομάδων για την εκπόνηση μίας κοινωνιολογικής έρευνας όπου με την χρήση διάφορων ερωτηματολογίων να προσπαθήσει να προσδιορίσει τα κοινωνικά χαρακτηριστικά των ατόμων που αποτελούν την κάθε ομάδα. Αυτό εκτός από την βοήθεια που θα προσφέρει στην εκάστοτε εταιρία, μέσω της γνώσης των καταναλωτικών συνήθειών των κοινωνικών ομάδων με στόχο την δημιουργία ενός κοινωνικού προφίλ της εταιρίας, ενδεχομένως να έχει και επιστημονικό ενδιαφέρον για τους ερευνητές των αντιστοίχων κλάδων.

6. Βιβλιογραφία

1. Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning*, Springer, California USA, 2017.
2. Ester, M., Kriegel, H. P., Sander, J. and Xu, X. *Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, Munich Germany, 1996
3. Bora, D. J. and Gupta, A. K. *A Comparative Study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm*, Bhopal India, 2014
4. Elgabry, O. *The Ultimate Guide to Data Cleaning*, https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4
5. Μπούρου Δ. *Στατιστική Ανάλυση Δεδομένων με Ακραίες και Ελλιπούσες Τιμές*, Διπλωματική εργασία, Πάτρα Ελλάδα. 2016
6. Santoyo, S. *A Brief Overview of Outlier Detection Techniques*, https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561
7. Pathak, M. *Hierarchical Clustering in R*, https://www.datacamp.com/community/tutorials/hierarchical-clustering-R
8. Rand, W. *Objective Criteria for Evaluation of Clustering Methods*, USA, 2019