Πρόγραμμα Μεταπτυχιακών Σπουδών
στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

<u>Διπλωματική εργασία</u>

Τμηματοποίηση πελατών με χρήση αλγορίθμων ανάλυσης κατά συστάδες για μεικτού τύπου δεδομένα

της

Αναστασίας Ντογραματζίδου του Στεφάνου

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Επιβλέπων Καθηγητής: Άγγελος Μάρκος

Φεβρουάριος 2023

Αφιερώσεις

Η εργασία αυτή είναι αφιερωμένη στον Άρη, στη Μαριάννα, στην Έλλη και στον Γιάννη στους φίλους-οικογένεια που είναι δίπλα μου σε κάθε προσωπικό και επαγγελματικό μου βήμα καθώς και στην οικογένεια μου για την παντοτινή υποστήριξη που μου παρέχουν.

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής διπλωματικής μου εργασίας, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου, κύριο Άγγελο Μάρκο, που χωρίς την σημαντική βοήθεια και καθοδήγηση του δεν θα μπορούσε να ολοκληρωθεί εν ευθέτω χρόνο αυτή η εργασία.

Πίνακας Σχημάτων

Πίνακας Πινάκων

Περίληψη

Η τμηματοποίηση πελατών (customer segmentation) είναι σημαντική πρακτική για να υποστηρίξει τον στρατηγικό σχεδιασμό μιας ομάδας μάρκετινγκ με απώτερο σκοπό την βελτιστοποίηση στοχευμένης διαφημιστικής καμπάνιας σε μια μερίδα του καταναλωτικού κοινού. Η τμηματοποίηση των πελατών στηρίζεται στην αντίληψη ότι δεν είναι εφικτό ένα προϊόν να είναι αρεστό στο σύνολο του καταναλωτικού κοινού. Ο τρόπος ζωής, η γνώση πάνω στο καταναλωτικό αγαθό ακόμη και η αγοραστική συμπεριφορά διαφέρει από άτομο σε άτομο. Συνεπώς μια επιτυχημένη επιχείρηση προσαρμόζει τα προϊόντα της για να ικανοποιήσει κάθε ομάδα πελατών. Η τμηματοποίηση αγοράς είναι μια κοστοβόρα και χρονοβόρα διαδικασία για μια επιχείρηση και θα πρέπει να κάνει ενδελεχή έρευνα αγοράς προτού προβεί στην υλοποίηση της. Στην παρούσα εργασία, αρχικά παρατίθενται δέκα βήματα που χρειάζεται να λάβει υπόψη της μια επιχείρηση ώστε να αποφασίσει βάσει κοινής λογικής αλλά και εμπειρικών δεδομένων αν η τμηματοποίηση της αγοράς είναι η ορθή απόφαση για την υλοποίηση ενός επιτυχημένου μείγματος μάρκετινγκ. Στη συνέχεια, παρουσιάζονται οι πιο αντιπροσωπευτικές στατιστικές μέθοδοι τμηματοποίησης για μεικτού τύπου δεδομένα, όπως αυτές που βασίζονται σε μετρικές απόστασης, οι ιεραρχικές μέθοδοι, οι μέθοδοι διαμέρισης και υβριδικές προσεγγίσεις. Οι μέθοδοι εφαρμόζονται για την τμηματοποίηση των δυνητικών πελατών μιας αυτοκινητοβιομηχανίας. Η εργασία μπορεί να αποτελέσει σημείο αναφοράς για τους αναλυτές δεδομένων στην επιλογή της καταλληλότερης μεθόδου τμηματοποίησης της αγοράς.

Πίνακας Περιεχομένων

1 Εισαγωγή

Ο σκοπός του μάρκετινγκ είναι να συνδυάσει τις πραγματικές ανάγκες και επιθυμίες του καταναλωτή με την προσφορά αγαθών από την επιχείρηση που είναι η κατάλληλη για να ικανοποιήσει τη ζήτηση του συγκεκριμένου ατόμου. Αυτή η διαδικασία αντιστοίχισης ωφελεί τόσο τον καταναλωτή όσο και την επιχείρηση ενώ παράλληλα διαμορφώνει τη διαδικασία σχεδιασμού στρατηγικής μάρκετινγκ (Dolnicar, Grün, & Leisch, 2018).

Η διαδικασία σχεδιασμού στρατηγικής μάρκετινγκ είναι μια λογική ακολουθία και μια σειρά από διαδικασίες που οδηγούν στον καθορισμό των στόχων μάρκετινγκ και της διαμόρφωσης πλάνου για την επίτευξή τους (McDonald and Wilson 2011, όπως αναφ. στο Dolnicar, Grün, & Leisch (2018)). Έτσι ένα πλάνο στρατηγικής μάρκετινγκ αποτελείται από δύο μέρη: τον στρατηγικό σχεδιασμό που περιγράφει τα μακροπρόθεσμα σχέδια της επιχείρησης και τον τακτικό σχεδιασμό που αφορά στην εφαρμογή του μακροπρόθεσμου σχεδιασμού στις τρέχουσες δράσεις μάρκετινγκ της εταιρίας. Πιο συγκεκριμένα ο στρατηγικός σχεδιασμός αφορά στο που θέλει να κατευθυνθεί η επιχείρηση και γιατί, ενώ ο τακτικός σχεδιασμός δίνει τις τεχνικές βάσεις για το τι θα χρειαστεί η επιχείρηση για να καθιερωθεί στον χάρτη της αγοράς. Με άλλα λόγια ο στρατηγικός σχεδιασμός θέτει τις βάσεις για το ποιος είναι ο στόχος της επιχείρησης και γιατί, ενώ ο τακτικός σχεδιασμός παρέχει τις οδηγίες που χρειάζεται να ακολουθήσει η επιχείρηση για να κατακτήσει τον στόχο της.

Σε ένα στρατηγικό σχεδιασμό για μια διαφημιστική καμπάνια λαμβάνεται υπόψη το απλό, αλλά πολύ σημαντικό πλαίσιο «Βλέπω - Σκέφτομαι - Πράττω - Ενδιαφέρομαι» (See - Think - Do - Care) που χρησιμοποιείται κυρίως από την Google και καθιερώθηκε από τον Avinash Kaushik, Marketing Evangelist της Google (Vollrath & Villegas, 2021). Ο Kaushik (2015) αναφέρει πως οι περισσότεροι marketers βλέπουν μηχανικά τον σχεδιασμό μια καμπάνιας. Τους ενδιαφέρει η απόδοση της επένδυσης (Return on Investment - ROI), και οι μετρήσεις απόδοσης (Key Performance Indicator - KPI), αλλά δεν λαμβάνουν υπόψη τα χαρακτηριστικά της κάθε ομάδας καταναλωτών και την εμπειρίας τους στο καταναλωτικό τους ταξίδι. Το μάρκετινγκ, όμως, επαναπροσδιορίζεται, δίνοντας έμφαση στις σχέσεις επιχειρήσεων και καταναλωτών, χτίζοντας μια πιο πελατοκεντρική οπτική. Πλέον δείκτης επιτυχίας αποτελεί η αξία του πελάτη (Customer Lifetime Value - CLV) (Rust, Moorman, & Bhalla, 2009).

1.1 Ορισμός της τμηματοποίησης πελατών

Η τμηματοποίηση της αγοράς είναι από τις πιο σημαντικές πρακτικές στον στρατηγικό σχεδιασμό του μάρκετινγκ και στην κατανόηση της καταναλωτικής συμπεριφοράς, γεγονός που την καθιστά απαραίτητη για την επιτυχία της επιχείρησης στην αγορά (Morgan, Whitler, Feng, & Chari, 2019). Αφενός μελετώντας την καταναλωτική συμπεριφορά, δημιουργούνται τα θεμέλια για αποδοτική τμηματοποίηση, αφετέρου στοχεύοντας τη κατάλληλη ομάδα καταναλωτών, κρίνεται επιτυχής μια ενέργεια μάρκετινγκ (Peter & Olson, 2010). Η τμηματοποίηση αγοράς βασίζεται στην απλή λογική πως δεν είναι εφικτό ένα προϊόν να είναι αρεστό σε ολόκληρο το καταναλωτικό κοινό, επομένως χρειάζεται ο σχεδιασμός ενός στρατηγικού μείγματος μάρκετινγκ. Ο τρόπος ζωής του κάθε καταναλωτή, η γνώση του για το προϊόν, η αγοραστική του συμπεριφορά, διαφέρει από άνθρωπο σε άνθρωπο. Συνεπώς οι επιτυχημένες επιχειρήσεις προσαρμόζουν το προϊόν τους για να ικανοποιεί κάθε τύπο καταναλωτή. Για παράδειγμα, ένα αναψυκτικό τύπου κόλα διαφοροποιεί τα προϊόντα του σύμφωνα με τον τύπο του καταναλωτή, έτσι υπάρχει η κανονική έκδοση για το ευρύ κοινό, η light έκδοση για τα άτομα που προσέχουν τη διατροφή τους και κάποιες εκδόσεις με πρόσθετες γεύσεις για τα άτομα που ενθουσιάζονται με το να δοκιμάζουν κάτι διαφορετικό.

Όπως αναφέρθηκε προηγουμένως, κριτήριο για την τμηματοποίηση του αγοραστικού κοινού, μπορεί να είναι ένα κοινό χαρακτηριστικό του κοινού όπως το φύλο, η ηλικία, η χώρα κατοικίας, ή η οικογενειακή κατάσταση. Επιπλέον, θα μπορούσαν να είναι κριτήρια, πέρα των δημογραφικών στοιχείων του κοινού, καταναλωτικές συμπεριφορές, όπως τι είδους πλεονεκτήματα αναζητούνται κατά την αγορά ενός αγαθού, τι είδους δραστηριότητες προτιμώνται κατά τις διακοπές του ατόμου, την περιβαλλοντικές του συνήθειες ή κάποιο μοτίβο δαπανών.

Ένα μειονέκτημα της τμηματοποίησης της αγοράς είναι η επένδυση χρόνου και ανθρώπινου δυναμικού από την επιχείρηση. Για να πραγματοποιηθεί μια ενδελεχής έρευνα για την τμηματοποίηση του καταναλωτικού κοινού απαιτείται ένας μεγάλος αριθμός ατόμων που θα αφιερώσουν σημαντικό χρόνο για έρευνα. Αν η επιχείρηση αποφασίσει να ακολουθήσει μια τέτοια στρατηγική, απαιτούνται μεγαλύτεροι ανθρώπινοι και οικονομικοί πόροι για την ανάπτυξη και εφαρμογή ενός προσαρμοσμένου μείγματος μάρκετινγκ. Πέρα από αυτό, η αξιολόγηση της επικείμενης επιτυχίας αυτής της στρατηγικής, και η παρακολούθηση της δυναμικής της αγοράς, που μπορεί να υποδεικνύει

και πιθανές τροποποιήσεις, προϋποθέτουν συνεχή δέσμευση πόρων της επιχείρησης. Οπότε για να είναι αυτό ωφέλιμο για την κάθε επιχείρηση, είναι προϋπόθεση πως θα αξίζει η αρχική επένδυση καθώς θα αποδώσει το πλέον ωφέλιμο για την επιχείρηση.

Στην περίπτωση που η αρχική επένδυση δεν αποδώσει λόγω κακής εφαρμογής, ή όλη υλοποίηση της τμηματοποίησης αποτελεί σπατάλη πόρων. Αυτή η αποτυχημένη στρατηγική μπορεί να οδηγήσει σε σημαντικές δαπάνες που δεν αποφέρουν καμία προστιθέμενη αξία στην υλοποίηση διαφημιστικής καμπάνιας και απογοητεύουν το εμπλεκόμενο ανθρώπινο δυναμικό.

Συνεπώς, προτείνεται μια επιχείρηση να συγκεντρώσει όλες τις πληροφορίες που χρειάζονται για να αξιολογήσει αν η στρατηγική τμηματοποίησης της αγοράς είναι συμφέρουσα επιλογή και να λάβει μια τεκμηριωμένη απόφαση ώστε να ξεκινήσει έναν στρατηγικό σχεδιασμό τμηματοποίησης του καταναλωτικού κοινού και να υλοποιήσει την στρατηγική αυτή για την επίτευξη του στόχου που έχει θέσει.

1.2 Στάδια της ανάλυσης τμηματοποίησης πελατών

Η ανάλυση τμηματοποίησης των πελατών, όπως έχει αναφερθεί παραπάνω, είναι η διαδικασία ομαδοποίησης του καταναλωτικού κοινού το οποίο μοιράζεται κοινές καταναλωτικές προτιμήσεις και χαρακτηριστικά, σε ομάδες.

Στον πυρήνα αυτής της διαδικασίας βρίσκεται η εφαρμογή κατάλληλων μεθόδων από τις οποίες προκύπτουν τα τμήματα του καταναλωτικού κοινού. Οι αποφάσεις που λαμβάνονται από τον αναλυτή σε αυτή τη διαδικασία επηρεάζουν την τελική απόφαση της στρατηγικής. Συνεπώς, για να είναι η ανάλυση τμηματοποίησης χρήσιμη σε μια επιχείρηση, χρειάζεται τόσο ένας ικανός αναλυτής όσο μια επιχείρηση η οποία κατανοεί το καταναλωτικό προφίλ των τμημάτων πελατών που εξάγονται από τα δεδομένα των καταναλωτών.

Για να διασφαλιστεί η διεξαγωγή υψηλής ποιότητας ανάλυσης απαιτείται μια σειρά από προϋποθέσεις, όπως η συλλογή χρήσιμων δεδομένων, η καλή διερεύνηση αυτών των δεδομένων, η σωστή τμηματοποίηση του καταναλωτικού κοινού και η περιγραφή αυτών των τμημάτων. Η συλλογή χρήσιμων δεδομένων είναι πολύ σημαντική κατά τους Dolnicar et al. (2018) καθώς όσο καλή και να είναι η στατιστική μέθοδος που εφαρμόζεται τον εντοπισμό των τμημάτων αγοράς, δεν θα μπορέσει να αντισταθμίσει τα κακής ποιότητας δεδομένα. Με άλλα λόγια, η ομαδοποίηση των καταναλωτών θα είναι

πάντα τόσο καλή όσο και τα δεδομένα που έχουν χρησιμοποιηθεί για να σχηματιστούν οι ομάδες αυτές.

Αν το στάδιο συλλογής δεδομένων ολοκληρωθεί, θα πρέπει να διερευνηθούν τα δεδομένα που προκύπτουν για να αποκτηθεί μια πρώτη εικόνα του προφίλ των ομάδων. Αφού ταξινομηθεί το καταναλωτικό κοινό στις ομάδες αυτές, θα πρέπει να σκιαγραφηθούν πλήρως οι ομάδες για να μπορέσουν οι επιχειρήσεις να κατανοήσουν τα χαρακτηριστικά της κάθε ομάδας και να επιλέξουν πόσες και ποιες θα στοχεύσουν, σχεδιάζοντας έτσι το εξατομικευμένο μείγμα μάρκετινγκ που θα ακολουθήσουν.

Με την επιτυχή ολοκλήρωση των πρώτων δύο σταδίων της ανάλυσης της τμηματοποίησης το αποτέλεσμα είναι ένας άψογος θεωρητικά σχεδιασμός για να ξεκινήσει η διαδικασία στόχευσης. Όμως καμία άψογη λύση δεν μπορεί να έχει νόημα αν οι επιχειρήσεις δεν μπορούν να μετατρέψουν αυτή την ανάλυση σε στρατηγικές αποφάσεις και τακτικές δράσεις μάρκετινγκ. Γι’ αυτό χρειάζεται ένα τρίτο στάδιο που θα περικλείει μέσα του τα άλλα δύο. Αυτό το στάδιο περιλαμβάνει μη τεχνικές εργασίες οι οποίες δεν ακολουθούν διαδοχικά τα προηγούμενα στάδια. Περιλαμβάνει οργανωτικά θέματα όπως την απόφαση για τις ομάδες, να οριστεί η ιδανική ομάδα, να επιλεχθούν οι ομάδες που θα στοχεύσει το μείγμα μάρκετινγκ καθώς να εκτιμηθεί η αποτελεσματικότητα και να παρακολουθούνται οι αγοραστικές αλλαγές.

Μετά την ολοκλήρωση της εργασίας εξαγωγής τμημάτων, οι επιχειρήσεις πρέπει να αξιολογήσουν τις ομάδες του καταναλωτικού κοινού που προέκυψαν και να επιλέξουν μία ή περισσότερες ομάδες για περαιτέρω ανάλυση. Οι αναλυτές δεδομένων μπορούν να παρέχουν στοιχεία για αυτά τις ομάδες, αλλά δεν μπορούν να επιλέξουν τις πλέον κατάλληλες. Η επιλογή αυτή καθοδηγείται, εν μέρει, τόσο από τα δυνατά σημεία και τις ευκαιρίες της επιχείρησης όσο και την ευθυγράμμισή της με τις βασικές ανάγκες των ομάδων της αγοράς. Τέλος, αφού επιλεχθούν οι ομάδες-στόχοι, οι επιχειρήσεις θα πρέπει να αναπτύξουν ένα σχέδιο μάρκετινγκ γι’ αυτές και να σχεδιάσουν ένα εξατομικευμένο μείγμα μάρκετινγκ.

Μια προσέγγιση τμηματοποίησης των πελατών είναι η τμηματοποίηση με γνώμονα τα υπάρχοντα δεδομένα σύμφωνα με τους Dolnicar et al. (2018). Η τμηματοποίηση που καθοδηγείται από τα δεδομένα (data driven) χρησιμοποιεί την ανάλυση δεδομένων, ώστε να προκύψουν οι τελικές ομάδες. Κατά την εφαρμογή αυτού του τύπου τμηματοποίησης η επιχείρηση έχει διατυπώσει συγκεκριμένες υποθέσεις σχετικά με τα χαρακτηριστικά των καταναλωτών, τα οποία είναι χρήσιμα για την

προώθηση προϊόντων σε ομάδες καταναλωτών, αλλά δεν γνωρίζουν τα προφίλ αυτών των ομάδων. Άρα ο στόχος είναι διττός: να διερευνηθούν οι ομάδες καταναλωτών που προέκυψαν από την τμηματοποίηση και να επιτευχθεί ένα λεπτομερέστερο προφίλ των ομάδων (Dolnicar, Grün, & Leisch, 2018). Η τμηματοποίηση των πελατών δημιουργεί τις ομάδες και καθορίζει ποιες μπορούν να επιλεχθούν για τις διαφημιστικές καμπάνιες. Η επιχείρηση με αυτά τα στοιχεία, πρέπει να αναπτύξει ένα μείγματος μάρκετινγκ που να εξυπηρετεί κάθε ομάδα, στα πλαίσια του δεοντολογικού κανόνα (Peter & Olson, 2010).

1.3 Τμηματοποίηση πελατών σε δέκα βήματα

Οι Dolnicar et al. (2018) παρουσιάζουν μια προσέγγιση δέκα βημάτων για την ανάλυση της τμηματοποίησης των πελατών. Η βασική ιδέα αυτών των βημάτων είναι βασισμένη τόσο στην κοινή λογική όσο και στην τμηματοποίηση που βασίζεται στα δεδομένα. Αρχικά, η επιχείρηση πρέπει να αξιολογήσει τόσο τα πλεονεκτήματα όσο και τα μειονεκτήματα μιας στρατηγικής τμηματοποίησης του καταναλωτικού κοινού και να αποφασίσει αν θα την υιοθετήσει ή όχι (1^{ο} βήμα). Έπειτα η επιχείρηση χρειάζεται να προσδιορίσει το προφίλ της ομάδας του καταναλωτικού κοινού που θεωρεί ιδανικό (2^{ο} βήμα). Μόνο αφού ολοκληρωθούν αυτά τα προκαταρκτικά βήματα, μπορούν να συλλεχθούν και να συγκεντρωθούν τα εμπειρικά δεδομένα από τις υπάρχουσες πηγές (3^{ο} βήμα). Αυτά τα δεδομένα πρέπει να εξεταστούν μέσω διερευνητικής ανάλυσης (4^{ο} βήμα) πριν να εξαχθούν οι τελικές ομάδες των καταναλωτών (5^{ο} βήμα). Οι ομάδες που θα προκύψουν πρέπει να είναι ομοιογενείς (6^{ο} βήμα) και να περιγραφούν ενδελεχώς (7^{ο} βήμα). Η επιχείρηση θα επιλέξει μία ή έναν μικρό αριθμό ομάδων στις οποίες θα επικεντρωθεί (8^{ο} βήμα). Πάνω σε αυτή την επιλογή θα βασιστεί το εξατομικευμένο μείγμα μάρκετινγκ που θα αναπτυχθεί (9^{ο} βήμα). Τέλος, αφού ολοκληρωθεί αυτή η ανάλυση, η επιτυχία εφαρμογής μιας στρατηγικής βασισμένης στην τμηματοποίηση της αγοράς, χρειάζεται διαρκής αξιολόγηση και παρακολούθηση των ομάδων καταναλωτών για πιθανές αλλαγές στο μέγεθος ή και στα χαρακτηριστικά τους (10^{ο} βήμα). Αυτές οι αλλαγές μπορεί να προκαλέσουν τροποποιήσεις στην στρατηγική μείγματος μάρκετινγκ που θα υιοθετηθεί.

Μολονότι αυτά τα δέκα βήματα είναι τα ίδια τόσο σύμφωνα με την κοινή λογική όσο και σύμφωνα με την τμηματοποίηση που βασίζεται σε εμπειρικά δεδομένα, η διαδικασία που ακολουθείται είναι διαφορετική σε κάθε μια από τις παραπάνω

προσεγγίσεις. Συνήθως η τμηματοποίηση που βασίζεται στα δεδομένα απαιτεί τη λήψη πρόσθετων αποφάσεων, όπως θα παρουσιαστεί σε επόμενη ενότητα.

1.4 Σκοπός και ειδικοί στόχοι της εργασίας

Σκοπός αυτής της εργασίας είναι η ανασκόπηση μεθόδων τμηματοποίησης πελατών για μεικτού τύπου δεδομένα, δηλαδή δεδομένα που περιγράφονται τόσο από ποιοτικές όσο και από ποσοτικές μεταβλητές, και η εφαρμογή των μεθόδων αυτών πάνω σε πραγματικά δεδομένα τμηματοποίησης πελατών. Στο 2^{ο} κεφάλαιο παρουσιάζονται οι βασικές μέθοδοι τμηματοποίησης, όπως αυτές που βασίζονται σε μετρικές απόστασης, οι ιεραρχικές μέθοδοι, οι μέθοδοι διαμέρισης και υβριδικές προσεγγίσεις. Στο 3^{ο} κεφάλαιο γίνεται εφαρμογή των μεθόδων αυτών για την τμηματοποίηση των δυνητικών πελατών μιας αυτοκινητοβιομηχανίας. Ο στόχος της συγκεκριμένης επιχείρησης είναι να προωθήσει τα προϊόντα της σε νέο καταναλωτικό κοινό με στοχευμένη διαφημιστική καμπάνια ανά ομάδα ή τμήμα πελατών. Για κάθε μέθοδο τμηματοποίησης που εφαρμόζεται, γίνεται μια περιγραφή του προφίλ των πελατών ανά ομάδα που προέκυψε. Τέλος, στα συμπεράσματα της παρούσας εργασίας παρουσιάζονται τα επιθυμητά χαρακτηριστικά ενός κατάλληλου αλγόριθμου για την τμηματοποίηση πελατών

2 Μέθοδοι τμηματοποίησης για μεικτού τύπου δεδομένα

2.1 Μέθοδοι που βασίζονται σε μετρικές απόστασης

Ένας τυπικός πίνακας δεδομένων αποτελείται από γραμμές και στήλες. Κάθε γραμμή ενός πίνακα δεδομένων αντιστοιχεί σε μια παρατήρηση ενώ, αντίστοιχα, κάθε στήλη αντιστοιχεί σε μια μεταβλητή. Ως προς τη μαθηματική του μορφή, ένας πίνακας δεδομένων μπορεί να αναπαρασταθεί ως ένας πίνακας $n \times p$, όπου το n συμβολίζει τον αριθμό των παρατηρήσεων (γραμμές) και p τον αριθμό των μεταβλητών (στήλες).

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Το διάνυσμα που αντιστοιχεί στην i-οστή γραμμή του πίνακα **X** συμβολίζεται με $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$, έτσι ώστε το $X = \{x_1, x_2, \ldots . x_p\}$ να αναπαριστά το σύνολο όλων των παρατηρήσεων.

Υπάρχουν πολυάριθμες προσεγγίσεις για τη μέτρηση της απόστασης μεταξύ δύο διανυσμάτων. Η πλειοψηφία αυτών χρησιμοποιείται συνήθως στην ανάλυση κατά συστάδες (clustering) και στην τμηματοποίηση πελατών. Η απόσταση είναι μια συνάρτηση $d(\cdot, \cdot)$ με δύο ορίσματα: τα δύο διανύσματα x και y των οποίων η απόσταση θα υπολογιστεί. Το αποτέλεσμα είναι η μεταξύ τους απόσταση (η οποία είναι πάντα μία μη αρνητική τιμή). Προκειμένου να γίνει πιο κατανοητό το παραπάνω, δίνεται ένα παράδειγμα στο πλαίσιο της γεωγραφίας. Αν επιδιώκεται να υπολογιστεί η απόσταση μεταξύ δύο πόλεων, οι θέσεις των πόλεων είναι τα δύο διανύσματα και το μήκος της αεροπορικής διαδρομής σε χιλιόμετρα είναι η απόσταση μεταξύ τους. Ωστόσο, ακόμα και στο ίδιο πλαίσιο, αυτό δηλαδή της γεωγραφικής απόστασης, μπορούν να χρησιμοποιηθούν και άλλες, εξίσου έγκυρες, μετρικές. Ένα τέτοιο παράδειγμα είναι η απόσταση που πρέπει να διανύσει ένα αυτοκίνητο οδικά για να φτάσει από τη μία πόλη στην άλλη.

Μία μετρική απόστασης πρέπει να πληροί ορισμένα κριτήρια. Το πρώτο κριτήριο είναι η συμμετρία, η οποία αναπαρίσταται μαθηματικά ως εξής:

$$d(x,y) = d(y,x)$$

Το δεύτερο κριτήριο είναι ότι η απόσταση ενός διανύσματος από τον εαυτό του είναι πάντα μηδενική:

$$d(x,y) = 0 <=> x = y$$

Επιπλέον, τα περισσότερα μέτρα απόστασης πληρούν τη λεγόμενη τριγωνική ανισότητα:

$$d(x, z) \leq d(x,y) + d(y,z)$$

Με άλλα λόγια, η τριγωνική ανισότητα δείχνει ότι αν κάποιος πηγαίνει από το σημείο *x* στο σημείο *z* με μια ενδιάμεση στάση στο *y*, η συνδυασμένη απόσταση είναι τουλάχιστον τόσο μεγάλη όσο η απευθείας μετάβαση από το *x* στο *z*.

Έστω $x = (x_1, \ldots, x_p)'$ και $y = (y_1, \ldots, y_p)'$ δύο *p*-διάστατα διανύσματα. Οι πιο συχνά χρησιμοποιούμενες μετρικές απόστασης στην ανάλυση τμηματοποίησης είναι οι παρακάτω:

Ευκλείδεια απόσταση:

$$d(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$$

απόσταση Manhattan:

$$d(x, y) = \sum_{j=1}^{p} |x_j - y_j|$$

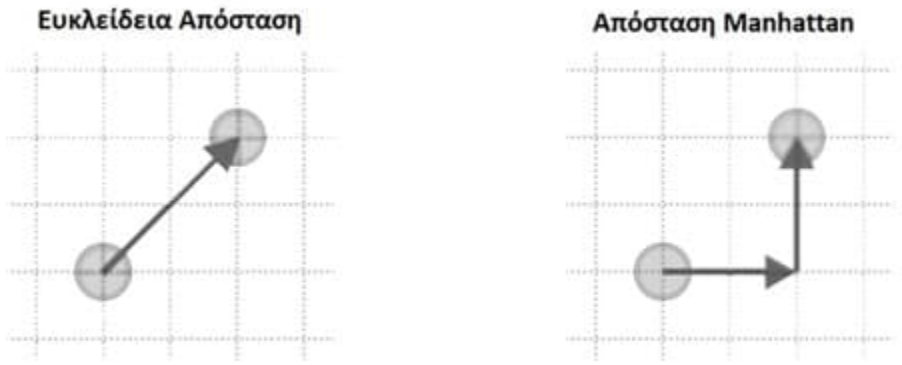
Ασύμμετρη δυαδική απόσταση: Η εν λόγω μετρική μπορεί να εφαρμοστεί μόνο σε δυαδικά διανύσματα, όπου όλα τα x_j και y_j λαμβάνουν τιμές είτε 0 είτε 1.

$$d(x,y) = \begin{cases} 0, & x = y = 0 \\ \left(\#\{j | x_j = 1 \text{ και } y_j = 1\}\right) / \left(\#\{\{j | x_j = 1 \text{ } or \text{ } y_j = 1\}\right) \end{cases}$$

Το δεύτερο μέρος της ισότητας μπορεί να περιγραφεί λεκτικά ως εξής: η απόσταση μεταξύ δύο σημείων *x* και *y* ισούται με το πλήθος των περιπτώσεων στις οποίες και οι δύο τιμές είναι ίσες με 1 διαιρεμένο με το πλήθος των περιπτώσεων όπου τουλάχιστον μία από αυτές είναι 1.

Η Ευκλείδεια απόσταση είναι η πλέον χρησιμοποιούμενη μετρική απόστασης στην ανάλυση τμηματοποίησης. Η Ευκλείδεια απόσταση αντιστοιχεί στην άμεση "ευθεία" απόσταση μεταξύ δύο σημείων στο δισδιάστατο χώρο, όπως φαίνεται στο Σχήμα 2-1.

Από την άλλη πλευρά, η απόσταση Manhattan πήρε το όνομά της από το γεγονός ότι προκειμένου να υπολογιστεί η απόσταση μεταξύ δύο σημείων χρησιμοποιούνται οδοί που σχηματίζουν τετράγωνα (όπως π.χ. στην πόλη του Manhattan). Η απόσταση Manhattan απεικονίζεται στο Σχήμα 2-1 στα δεξιά.



Σχήμα 2-1: Σύγκριση της Ευκλείδιας απόστασης και της απόστασης Manhattan [Πηγή: Προσαρμοσμένο από τους (Dolnicar, Grün, & Leisch, 2018)

Σε κάθε περίπτωση, τόσο η Ευκλείδεια απόσταση όσο και η απόσταση Manhattan χρησιμοποιούν όλες τις διαστάσεις των διανυσμάτων, υπολογίζοντας το άθροισμα των τετραγώνων των τιμών (στην περίπτωση της Ευκλείδειας) ή των απολύτων τιμών των διαφορών (στην περίπτωση της Manhattan). Εάν οι τιμές των δεδομένων σε κάθε διάσταση ή μεταβλητή δεν είναι στην ίδια κλίμακα μέτρησης (για παράδειγμα, η πρώτη μεταβλητή δείχνει την ύπαρξη ή όχι μιας κατάστασης η οποία μετράται σε 0 και 1, και η δεύτερη μεταβλητή είναι σε χρηματικές μονάδες), η διάσταση με τος μεγαλύτερες τιμές θα κυριαρχήσει στον υπολογισμό της απόστασης μεταξύ δύο παρατηρήσεων. Σε τέτοιες περιπτώσεις τα δεδομένα πρέπει να κανονικοποιηθούν πριν από τον υπολογισμό της μεταξύ τους απόστασης.

Τέλος, η ασύμμετρη δυαδική απόσταση δεν χρησιμοποιεί όλες τις διαστάσεις των διανυσμάτων, αλλά μόνο αυτές όπου τουλάχιστον ένα από τα δύο διανύσματα έχει τιμή 1. Για τον λόγο αυτό, άλλωστε, χαρακτηρίζεται και ως ασύμμετρη καθώς αντιμετωπίζει με διαφορετικό τρόπο τις τιμές 0 και 1. Η ομοιότητα μεταξύ δύο παρατηρήσεων συνάγεται μόνο αν έχουν κοινά 1, αλλά όχι αν έχουν κοινά 0. Επιπλέον, η ανομοιότητα μεταξύ δύο παρατηρήσεων αυξάνεται αν η μία έχει 1 και η άλλη όχι. Αυτό έχει επιπτώσεις στην ανάλυση τμηματοποίησης. Με άλλα λόγια, η ασύμμετρη δυαδική απόσταση αντιστοιχεί στο ποσοστό των κοινών 1 σε όλες τις διαστάσεις όπου τουλάχιστον ένα διάνυσμα περιέχει ένα 1.

2.2 Ιεραρχικές μέθοδοι

Οι μέθοδοι ιεραρχικής ταξινόμησης (hierarchical clustering) αποτελούν έναν από τους πιο διαισθητικούς τρόπους ομαδοποίησης δεδομένων, καθώς μιμούνται τον τρόπο με τον οποίο ένας άνθρωπος θα επέλεγε να διαχωρίσει και να ομαδοποιήσει ένα σύνολο n παρατηρήσεων (π.χ., καταναλωτές) σε k ομάδες (π.χ. τμήματα αγοράς). Εάν στόχο αποτελεί η δημιουργία ενός και μόνο μεγάλου τμήματος αγοράς (επομένως $k = 1$), η μόνη δυνατή λύση είναι η δημιουργία ενός συνόλου το οποίο θα εμπεριείχε το σύνολο των καταναλωτών που ανήκουν στο σύνολο δεδομένων X. Από την άλλη πλευρά, το ακριβώς αντίθετο οριακό σενάριο αποτελεί η περίπτωση κατά την οποία στοχεύεται η δημιουργία ισάριθμων τμημάτων με το πλήθος των καταναλωτών που εμπεριέχονται στο σύνολο δεδομένων X (επομένως $k = n$). Σε αυτήν την περίπτωση, κάθε τμήμα περιέχει μόνο έναν καταναλωτή, ο οποίος και αντιπροσωπεύει τη δική του ομάδα. Η διαδικασία ανάλυσης τμηματοποίησης πελατών πραγματοποιείται μεταξύ αυτών των δύο οριακών καταστάσεων.

Οι μέθοδοι φθίνουσας ιεραρχικής ταξινόμησης ξεκινούν με το πλήρες σύνολο δεδομένων X ($k = 1$). Στο πρώτο βήμα, το αρχικό σύνολο διαιρείται σε δύο τμήματα - ομάδες. Στη συνέχεια, κάθε ένα από τα τμήματα διαχωρίζεται με τη σειρά του σε δύο επιμέρους τμήματα. Η διαδικασία αυτή συνεχίζεται έως ότου κάθε μοναδικός καταναλωτής αντιπροσωπεύει το δικό του τμήμα της αγοράς ($k = n$).

Μια διαφορετική προσέγγιση αποτελούν οι μέθοδοι αύξουσας ιεραρχικής ταξινόμησης. Το σημείο εκκίνησης του αλγορίθμου είναι η κατάσταση κατά την οποία κάθε καταναλωτής αντιπροσωπεύει τη δική του ομάδα (n ομάδες ή με άλλα λόγια $k = n$). Βήμα προς βήμα, τα δύο τμήματα πελατών που βρίσκονται πιο κοντά το ένα στο άλλο συγχωνεύονται έως ότου σχηματιστεί το αρχικό πλήρες σύνολο δεδομένων (δηλαδή $k = 1$).

Και οι δύο προσεγγίσεις οδηγούν σε μια ακολουθία εμφωλευμένων κατατμήσεων. Κάθε κατάτμηση αποτελεί ένα ομαδοποιημένο σύνολο παρατηρήσεων, έτσι ώστε κάθε παρατήρηση να ανήκει σε μία και μοναδική ομάδα. Επιπλέον, κάθε κατάτμηση μπορεί να περιέχει μόνο μία ομάδα (τμήμα αγοράς) έως n ομάδες (τμήματα αγοράς). Αυτές χαρακτηρίζονται ως εμφωλευμένες, καθώς η κατάτμηση με $k + 1$ ομάδες (τμήματα) προκύπτει από την κατάτμηση με k ομάδες αφού αποσπαστεί μία εκ των ομάδων.

Στη σχετική βιβλιογραφία, έχουν προταθεί αρκετοί αλγόριθμο για τις δύο προαναφερθείσες προσεγγίσεις. Ένα γενικό πλαίσιο για την αύξουσα ιεραρχική

ταξινόμηση - το οποίο αναπτύχθηκε στην πρωτοποριακή εργασία των Lance και Williams (Lance & Williams 1967, όπως αναφ. στο Dolnicar, Grün, & Leisch (2018)), - περιλαμβάνει τις περισσότερες μεθόδους που χρησιμοποιούνται ακόμη και σήμερα. Σε κάθε βήμα, οι αλγόριθμοι ιεραρχικής ταξινόμησης εκτελούν το βέλτιστο βήμα. Αυτό οδηγεί κάθε φορά σε έναν ντετερμινιστικό αλγόριθμο, δηλαδή κάθε φορά που εφαρμόζεται στο ίδιο σύνολο δεδομένων η ιεραρχική ταξινόμηση, προκύπτει ακριβώς η ίδια ακριβώς ακολουθία εμφωλευμένων κατατμήσεων.

Τόσο η φθίνουσα όσο και η αύξουσα ιεραρχική ταξινόμηση χρησιμοποιούν μια μετρική απόστασης μεταξύ των παρατηρήσεων (τμημάτων). Η μετρική αυτή καθορίζεται με τον προσδιορισμό 1) του τρόπου μέτρησης της απόστασης $d(x, y)$ μεταξύ των παρατηρήσεων (καταναλωτών) x και y, και (2) της μεθόδου συνένωσης των ομάδων. Η μέθοδος συνένωσης προσδιορίζει τον τρόπο με τον οποίο υπολογίζονται οι αποστάσεις μεταξύ των ομάδων παρατηρήσεων, δεδομένης μιας απόστασης μεταξύ ζευγών παρατηρήσεων. Έστω δύο σύνολα X και Y παρατηρήσεων (καταναλωτών). Οι μέθοδοι συνένωσης για τη μέτρηση της απόστασης l(X, Y) μεταξύ αυτών των δύο συνόλων παρατηρήσεων είναι οι ακόλουθες:

Απλή (single): η απόσταση μεταξύ των δύο πλησιέστερων παρατηρήσεων των δύο επιμέρους συνόλων είναι

$$l(\mathrm{X,Y}) = \min_{x \in X, y \in Y} d(x, y)$$

Πλήρης (complete): η απόσταση μεταξύ των δύο πιο απομακρυσμένων παρατηρήσεων των δύο επιμέρους συνόλων είναι

$$l(\mathrm{X,Y}) = \max_{x \in X, y \in Y} d(x, y)$$

Μέση (average): η μέση απόσταση μεταξύ όλων των παρατηρήσεων των δύο επιμέρους συνόλων δίνεται από

$$l(\mathrm{X,Y}) = \frac{1}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y)$$

όπου |X| = το πλήθος των στοιχείων του X.

Όλες οι παραπάνω μέθοδοι συνένωσης μπορούν να συνδυαστούν με οποιαδήποτε μετρική απόστασης. Δεν υπάρχει σωστός ή λάθος συνδυασμός μετρικής απόστασης και μεθόδου σύνδεσης. Εν γένει, η διαδικασία της ταξινόμησης, και ιδιαίτερα η ιεραρχική ταξινόμηση, είναι διερευνητικές μέθοδοι. Διαφορετικοί συνδυασμοί μπορούν να αποκαλύψουν διαφορετικά χαρακτηριστικά του συνόλου δεδομένων.

Η απλή συνένωση αξιοποιεί την τεχνική του “πλησιέστερου γείτονα” προκειμένου να ενώσει σύνολα δεδομένων. Με άλλα λόγια, βάσει αυτής της προσέγγισης ενώνονται οι δύο πλησιέστεροι καταναλωτές. Κατά συνέπεια, η ιεραρχική ταξινόμηση με απλή συνένωση έχει την ικανότητα να αναδεικνύει μη κυρτές και μη γραμμικές ομάδες παρατηρήσεων.

Σε περιπτώσεις όπου οι συστάδες δεν είναι καλά διαχωρισμένες - και αυτό συμβαίνει στις περισσότερες περιπτώσεις δεδομένων που αφορούν σε καταναλωτές- η προσέγγιση του πλησιέστερου γείτονα μπορεί να οδηγήσει στο μη επιθυμητό φαινόμενο κατά το οποίο δύο ομάδες καταναλωτών σχηματίζουν μια ομάδα μόνο και μόνο επειδή δύο καταναλωτές που ανήκουν σε κάθε μία από αυτές τις επιμέρους ομάδες βρίσκονται κοντά ο ένας στον άλλο. Αντίθετα, η πλήρης και η μέση συνένωση εξάγουν πιο ομοιογενείς ομάδες.

Μια πολύ δημοφιλής μέθοδος ιεραρχικής ταξινόμησης πήρε το όνομά της από τον Ward (1963, όπως αναφ. στο Dolnicar, Grün, & Leisch (2018)) και βασίζεται στις τετραγωνικές Ευκλείδειες αποστάσεις. Η συσταδοποίηση Ward ενώνει τα δύο σύνολα παρατηρήσεων (καταναλωτές) με την ελάχιστη σταθμισμένη τετραγωνική Ευκλείδεια απόσταση μεταξύ των κέντρων των ομάδων. Τα κέντρα των ομάδων αντιστοιχούν στα κέντρα βάρους της κάθε ομάδας. Προκύπτουν από τον υπολογισμό του μέσου όρου των παρατηρήσεων της ομάδας. Βάσει αυτού, μπορούν να θεωρηθούν ως εκπρόσωποι κάθε ομάδας. Αυτό που χρειάζεται ιδιαίτερη προσοχή κατά την εφαρμογή της συσταδοποίησης κατά Ward είναι η επιλογή της κατάλληλης μετρικής απόστασης (συνήθως απλή Ευκλείδεια ή τετραγωνική Ευκλείδεια).

Το αποτέλεσμα της ιεραρχικής ταξινόμησης παρουσιάζεται συνήθως ως ένα δενδρικό διάγραμμα ή δενδρόγραμμα. Η ρίζα του δέντρου αντιπροσωπεύει τη συστάδα εκείνη όπου ένα τμήμα της αγοράς περιλαμβάνει όλους τους καταναλωτές. Τα φύλλα του δέντρου είναι οι μεμονωμένες παρατηρήσεις (καταναλωτές), και τα κλαδιά στο ενδιάμεσο αντιστοιχούν στα ιεραρχία των τμημάτων της αγοράς που σχηματίζονται σε κάθε βήμα της διαδικασίας. Το ύψος του των κόμβων αντιστοιχεί στην απόσταση μεταξύ των συστάδων. Οι κόμβοι που βρίσκονται σε μεγαλύτερο ύψος στο δενδρόγραμμα υποδεικνύουν πιο διακριτά τμήματα της αγοράς. Τα δενδρογράμματα χρησιμοποιούνται συχνά ως οδηγός για την επιλογή του κατάλληλου αριθμού των τμημάτων της αγοράς.

2.3 Μέθοδοι διαμέρισης

Οι μέθοδοι ιεραρχικής ταξινόμησης θεωρούνται ιδιαίτερα κατάλληλες για την ανάλυση μικρών συνόλων δεδομένων, τα οποία έχουν έως και μερικές εκατοντάδες παρατηρήσεις. Όσον αφορά σε μεγαλύτερα σύνολα δεδομένων, τα δενδρογράμματα είναι δύσκολο να διαβαστούν και ο πίνακας των αποστάσεων μεταξύ των παρατηρήσεων συχνά δεν χωράει στη μνήμη του υπολογιστή.

Για σύνολα δεδομένων τα οποία περιέχουν περισσότερες από 1000 παρατηρήσεις (στην προκειμένη περίπτωση καταναλωτές), οι μέθοδοι τμηματοποίησης που οδηγούν σε μία μόνο διαμέριση (partition) του συνόλου δεδομένων με προκαθορισμένο τον αριθμό των ομάδων, είναι πιο κατάλληλες από μια εμφωλευμένη ακολουθία διαδοχικών διαμερίσεων. Αυτό σημαίνει ότι - αντί να υπολογίζονται όλες οι αποστάσεις μεταξύ όλων των ζευγών παρατηρήσεων του συνόλου δεδομένων στην αρχή της διαδικασίας της ταξινόμησης - υπολογίζονται μόνο οι αποστάσεις μεταξύ κάθε καταναλωτή που ανήκει στο σύνολο δεδομένων και του κέντρου κάθε ομάδας.

Για να γίνει καλύτερα κατανοητή αυτή η διαφορά, μπορεί κανείς να αναλογιστεί ένα σύνολο δεδομένων με 1000 καταναλωτές. Ένας αλγόριθμος αύξουσας ιεραρχικής ταξινόμησης χρειάζεται να υπολογίσει (1000×999)/2 = 499.500 διαφορετικές αποστάσεις προκειμένου να υπολογιστεί ο πίνακας αποστάσεων μεταξύ όλων των καταναλωτών που ανήκουν στο σύνολο δεδομένων. Αντίθετα, ένας αλγόριθμος ο οποίος επιδιώκει τη διαμέριση του συνόλου δεδομένων σε πέντε τμήματα αγοράς, χρειάζεται να υπολογίσει μόνο 5 έως 5000 αποστάσεις σε κάθε βήμα της επαναληπτικής διαδικασίας (ο ακριβής αριθμός εξαρτάται από τον αλγόριθμο που χρησιμοποιείται). Επιπλέον, εάν επιδιώκεται να εξαχθεί ένας μικρός αριθμός ομάδων, είναι προτιμότερο να βελτιστοποιηθεί ο αλγόριθμος ειδικά για το συγκεκριμένο πλήθος αντί να κατασκευαστεί το δενδρόγραμμα και στη συνέχεια να επιλεγεί ο επιθυμητός αριθμός των ομάδων.

2.3.1 Αλγόριθμοι k-Means και k-Centroid

Ο πιο δημοφιλής αλγόριθμος διαμέρισης είναι ο k-Means. Έστω $X = \{x_1, \ldots, x_n\}$ είναι ένα σύνολο παρατηρήσεων (καταναλωτών) ενός συνόλου δεδομένων. Οι μέθοδοι διαμέρισης ουσιαστικά διαχωρίζουν αυτούς τους καταναλωτές σε συγκεκριμένα υποσύνολα (τμήματα της αγοράς) με τέτοιο τρόπο ώστε οι καταναλωτές που ανήκουν στο ίδιο τμήμα της αγοράς να είναι όσο το δυνατόν πιο όμοιοι μεταξύ τους. Αντίστοιχα, οι καταναλωτές που ανήκουν σε διαφορετικά τμήματα της αγοράς να είναι κατά το δυνατόν

πιο ανόμοιοι. Ο αντιπρόσωπος (representative) κάθε τμήματος της αγοράς αναφέρεται σε πολλές αλγορίθμους συσταδοποίησης και ως κεντροειδές. Ειδικότερα, για τον αλγόριθμο k-means, ο οποίος βασίζεται στην τετραγωνική Ευκλείδεια απόσταση, το κεντροειδές υπολογίζεται από τον μέσο όρο των παρατηρήσεων που ανήκουν στο συγκεκριμένο τμήμα της αγοράς. Υπενθυμίζεται ότι τα σύνολα δεδομένων που αφορούν σε ανάλυση της αγοράς περιέχουν παρατηρήσεις (καταναλωτές) σε γραμμές και μεταβλητές (πληροφορίες συμπεριφοράς ή απαντήσεις σε έρευνα ερωτήσεων) σε στήλες.

Ο ακόλουθος αλγόριθμος αντιπροσωπεύει έναν ευρετικό τρόπο βελτιστοποίησης του προβλήματος διαχωρισμού των καταναλωτών σε ένα δεδομένο πλήθος τμημάτων αγοράς, έτσι ώστε οι καταναλωτές να είναι όσο το δυνατό πιο όμοιοι με τα υπόλοιπα μέλη του τμήματος στο οποίο ανήκουν, αλλά και όσο το δυνατό πιο ανόμοιοι με τα μέλη των υπόλοιπων τμημάτων. Πρόκειται για έναν επαναληπτικό αλγόριθμο ο οποίος βελτιώνει τον διαχωρισμό σε κάθε βήμα, συγκλίνοντας σε κάποιο βέλτιστο μιας συνάρτησης απώλειας (loss function), χωρίς ωστόσο αυτό να είναι απαραίτητα το ολικό βέλτιστο.

Ο αλγόριθμος αποτελείται από πέντε επιμέρους βήματα:

1. Καθορισμός του επιθυμητού πλήθους τμημάτων της αγοράς (k)
2. Τυχαία επιλογή k παρατηρήσεων (για παράδειγμα, αν στόχος είναι ο διαχωρισμός των καταναλωτών σε 5 τμήματα της αγοράς, τότε το 2ο βήμα ξεκινάει με την επιλογή 5 τυχαίων καταναλωτών) οι οποίοι θα αποτελέσουν τους αντιπροσώπους των k συστάδων, αποτελώντας το σύνολο $C = \{c_1, \dots, c_k\}$. Είναι φανερό πως εφόσον η αρχική επιλογή γίνεται με τυχαίο τρόπο, η επιλογή των k αυτών αντιπροσώπων δεν θα είναι η βέλτιστη δυνατή. Ωστόσο θα αποτελέσει την αρχή της επαναληπτικής διαδικασίας.
3. Αντιστοίχιση κάθε παρατήρησης x_i με τον εγγύτερο αντιπρόσωπο προκειμένου να σχηματιστεί μία συστάδα δεδομένων, όπου για k συστάδες $S_1, \dots, S_k$ θα ισχύει:

$$S_j = \{x \in X | \, d(x, c_j) \leq d(x, c_h), 1 \leq h \leq k\}$$

Το παραπάνω πρακτικά σημαίνει ότι κάθε καταναλωτής στο σύνολο δεδομένων κατατάσσεται σε έναν από τους αρχικούς k αντιπροσώπους που επιλέχθηκαν τυχαία. Αυτό επιτυγχάνεται με τον υπολογισμό της απόστασης μεταξύ κάθε καταναλωτή και κάθε αντιπροσώπου των k τμημάτων. Στη συνέχεια ο κάθε καταναλωτής αντιστοιχίζεται με τον πιο όμοιό του αντιπρόσωπο. Εάν δύο ή περισσότεροι αντιπρόσωποι διαφορετικών τμημάτων βρίσκεται σε ίδια απόσταση με τον εξεταζόμενο καταναλωτή, τότε η αντιστοίχιση πραγματοποιείται με τυχαίο τρόπο. Το αποτέλεσμα αυτού του βήματος είναι

μια αρχική - σε κάθε περίπτωση μη βέλτιστη - λύση τμηματοποίησης. Όλοι οι καταναλωτές στο σύνολο δεδομένων αντιστοιχίζονται σε ένα και μόνο τμήμα. Ωστόσο ακόμα δεν έχει επιτευχθεί ο στόχος που ορίζει ότι όλα τα μέλη του ίδιου τμήματος είναι όσο το δυνατόν πιο όμοια μεταξύ τους, και τα μέλη διαφορετικών τμημάτων είναι όσο το δυνατόν πιο ανόμοια.

4. Επαναϋπολογισμός του κέντρου κάθε συστάδας διατηρώντας σταθερά τα μέλη από τα οποία απαρτίζεται η συστάδα και ελαχιστοποιώντας την απόσταση του κάθε καταναλωτή από το κέντρο:

$$\text{cj} = \arg\, min_C \sum_{x \in S_j} d(x, c)$$

Όταν χρησιμοποιείται η τετραγωνική Ευκλείδεια απόσταση, τα βέλτιστα κέντρα υπολογίζονται με τους - κατά συστάδες - μέσους όρους, ενώ όταν χρησιμοποιείται η απόσταση Manhattan με τις - κατά συστάδες - διαμέσους. Αυτό αντικατοπτρίζεται, άλλωστε, και στην ορολογία, όπου η πρώτη διαδικασία αποκαλείται k-means ενώ η δεύτερη k-medians. Λαμβάνοντας υπόψη ότι η διαδικασία του 2ου βήματος οδηγεί σε μη βέλτιστη λύση, πρέπει να εντοπιστούν και να επαναϋπολογιστούν οι αντιπρόσωποι κάθε ομάδας (τμήματος πελατών). Αυτό, άλλωστε, είναι ακριβώς αυτό που επιτυγχάνεται σε αυτό το βήμα· χρησιμοποιώντας την αρχική λύση τμηματοποίησης, "εκλέγεται" ένας νέος εκπρόσωπος για καθεμιά από τις ομάδες. Όταν ως μετρική απόστασης χρησιμοποιείται η τετραγωνική Ευκλείδεια απόσταση, η διαδικασία πραγματοποιείται με τον υπολογισμό του μέσου όρου για όλα τα μέλη της ομάδας. Με τον τρόπο αυτό “ανακηρύσσεται” νέος αντιπρόσωπος για την εκάστοτε ομάδα.

5. Επανάληψη της διαδικασίας των βημάτων 3 και 4 μέχρις ότου επιτευχθεί η επιθυμητή σύγκληση ή επιτευχθεί ένα προκαθορισμένο πλήθος βημάτων. Πρακτικά, αυτό σημαίνει ότι τα στάδια της ανάθεσης των παρατηρήσεων στον πλησιέστερο αντιπρόσωπο καθώς και η εκλογή νέων αντιπροσώπων επαναλαμβάνεται έως ότου οι εκπρόσωποι των ομάδων παραμένουν οι ίδιοι. Στο σημείο αυτό ο αλγόριθμος σταματά και η λύση τμηματοποίησης στην οποία έχει καταλήξει λαμβάνεται ως η τελική.

Σε κάθε περίπτωση, ο αλγόριθμος θα συγκλίνει. Η διαδικασία κατάτμησης που χρησιμοποιείται σε έναν αλγόριθμο συσταδοποίησης θα οδηγεί πάντα σε μια λύση. Η επίτευξη της λύσης μπορεί να χρειαστεί μεγαλύτερο χρόνο και περισσότερους

υπολογιστικούς πόρους για μεγάλα σύνολα δεδομένων και μεγάλο επιδιωκόμενο πλήθος τμημάτων της αγοράς.

Το σημείο εκκίνησης της διαδικασίας είναι τυχαίο. Όπως περιεγράφηκε παραπάνω, τυχαίοι αρχικοί αντιπρόσωποι τμημάτων επιλέγονται προκειμένου να ξεκινήσει η διαδικασία των πέντε βημάτων. Διαφορετικοί αρχικοί αντιπρόσωποι (κεντροειδή) θα οδηγήσουν αναπόφευκτα και σε διαφορετικές λύσεις τμηματοποίησης του καταναλωτικού κοινού. Λαμβάνοντας το παραπάνω υπόψη, οδηγείται κανείς στο συμπέρασμα ότι η εκτέλεση ενός πειράματος οδηγεί σε μία από όλες τις πιθανές πρακτικές τμηματοποίησης πελατών. Έτσι, αν θέλει ο εκάστοτε αναλυτής να διενεργήσει υψηλού επιπέδου ανάλυση της αγοράς, θα πρέπει να επαναλάβει το πείραμα περισσότερες από μία φορές. Το κλειδί για μια υψηλής ποιότητας ανάλυση τμηματοποίησης είναι η συστηματική επανάληψη, που επιτρέπει στον αναλυτή δεδομένων να ξεχωρίσει τις λιγότερο χρήσιμες λύσεις και να παρουσιάσει στους χρήστες της λύσης τμηματοποίησης - για παράδειγμα, στα διευθυντικά στελέχη της επιχείρησης που θέλουν να υιοθετήσουν στοχευμένες πολιτικές μάρκετινγκ - το καλύτερο διαθέσιμο τμήμα ή σύνολο τμημάτων του κοινού.

Επιπλέον, όπως φάνηκε και στο 1ο βήμα της προαναφερθείσας διαδικασίας, απαιτείται ο εκ των προτέρων προσδιορισμός του πλήθους των τμημάτων αγορών, κάτι το οποίο δεν είναι πάντα εύκολο. Διαχρονικά, πολλοί τρόποι και δείκτες έχουν προταθεί από την επιστημονική κοινότητα για να μπορέσει να αποφασίσει ο αναλυτής, βάσει των δεδομένων που έχει κάθε φορά, τον κατάλληλο αριθμό συστάδων (Davies & Bouldin, 1979; Rousseeuw, 1987; Huang, et al., 2013; Kolesnikov, Trichina, & Kauranne, 2015; Ling, Wu, Zhou, & Zheng, 2016; Rojas-Thomas, Santos, & Mora, 2017; Guo, Chen, Ye, & Jiang, 2017; Zhang, Mańdziuk, Hiok Quek, & Wooi Goh, 2017; Manochandar, Punniyamoorthy, & Jeyachitra, 2020; Abdalameer, Alswaitti, Alsudani, & Isa, 2022). Μία διαφορετική εναλλακτική αποτελεί η επαναληπτική διαδικασία διαχωρισμού των καταναλωτών σε διαφορετικού πλήθους τμήματα κάθε φορά. Με τον τρόπο αυτό, και ανάλογα τα αποτελέσματα που προκύπτουν κάθε φορά, ο αναλυτής μπορεί να επιλέξει τον κατάλληλο αριθμό συστάδων που θα δημιουργηθούν βάσει είτε την πιο σταθερή συνολική λύση τμηματοποίησης είτε μεμονωμένα τα πιο σταθερά τμήματα.

Το σύνολο της προαναφερθείσας διαδικασίας αφορά σε μια γενικευμένη μορφή αλγορίθμων συσταδοποίησης. Στα πλαίσια της μηχανικής μάθησης, οι αλγόριθμοι συσταδοποίησης αναφέρονται ως αλγόριθμοι μη εποπτευόμενης μάθησης καθώς ο διαχωρισμός σε συστάδες δεν πραγματοποιείται βάσει μίας εξαρτημένης μεταβλητής.

Αντίθετα, στην περίπτωση της εποπτευόμενης μηχανικής μάθησης, υπάρχει μία εξαρτημένη μεταβλητή κάθε φορά, οδηγώντας στη διαδικασία είτε της παλινδρόμησης (αν η εξαρτημένη μεταβλητή είναι αριθμητική) είτε της κατηγοριοποίησης (αν η εξαρτημένη μεταβλητή είναι κατηγορική).

Σε κάθε περίπτωση, όποιος αλγόριθμος συσταδοποίησης και αν τελικά χρησιμοποιηθεί, η πραγματοποίηση υπολογισμών αποστάσεων είναι η απαραίτητη. Για τον λόγο αυτό, η επιλογή της της μετρικής απόστασης μπορεί να επηρεάσει σημαντικά τα παραγόμενα αποτελέσματα. Στην πραγματικότητα, η επιλογή της μετρικής απόστασης, συνήθως επηρεάζει περισσότερο το τελικό αποτέλεσμα συγκριτικά με την επιλογή του αλγορίθμου συσταδοποίησης (Leich, 2006).

2.3.2 "Βελτιωμένος" Αλγόριθμος k-Means

Στη βιβλιογραφία συναντάμε πολλές προσπάθειες βελτίωσης του αλγορίθμου συσταδοποίησης k-means. Η απλούστερη βελτίωση αφορά στην "έξυπνη" - και όχι απόλυτα τυχαία - αρχικοποίηση των *k* πρώτων καταναλωτών, οι οποίοι όπως αναφέρθηκε παραπάνω, αποτελούν τους πρώτους αντιπροσώπους των *k* συστάδων. Πιο συγκεκριμένα, η τυχαία επιλογή των αντιπροσώπων πολλές φορές ελλοχεύει τον κίνδυνο ανάδειξης *k* αντιπροσώπων οι οποίοι βρίσκονται πολύ κοντά μεταξύ τους και, για τον λόγο αυτό, δεν είναι αντιπροσωπευτικοί του συνόλου δεδομένων. Το παραπάνω, επιπρόσθετα, μπορεί να εγκλωβίσει τον αλγόριθμο σε κάποιο τοπικό βέλτιστο (δηλαδή σε μία καλή μεν αλλά όχι στην απόλυτα βέλτιστη λύση δε) (Jain, 2010), απαιτώντας έτσι πολλαπλές εκτελέσεις του αλγορίθμου προκειμένου να αποφευχθεί αυτό το γεγονός (Ikotun, Ezugwu, Abualigah, Abuhaija, & Heming, 2023; Bai, Liang, & Cao, 2020; Zhou, Wu, Luo, & Mohamed, 2019). Έτσι, επιλέγοντας τους αρχικούς *k* αντιπροσώπους από όλη την έκταση του δειγματικού χώρου μπορεί να αποτρέψει την εμφάνιση των δύο προαναφερθέντων προβλημάτων.

Οι Steinley και Brusco (2007) συγκρίνουν 12 διαφορετικές στρατηγικές που προτείνονται για την αρχικοποίηση τον αλγόριθμο k-means. Έχοντας εκπονήσει μια εκτεταμένη μελέτη προσομοίωσης με τη χρήση προσομοιωμένων δεδομένων με γνωστή δομή, οι ερευνητές καταλήγουν στο συμπέρασμα ότι η καλύτερη προσέγγιση είναι η τυχαία επιλογή πολλών διαφορετικών αρχικών αντιπροσώπων, η εφαρμογή της μεθόδου k-menas και η επιλογή της καλύτερης δυνατής ομαδοποίησης. Η καλύτερη δυνατή ομαδοποίηση είναι αυτή η οποία περιγράφει με τον καλύτερο τρόπο το σύνολο δεδομένων. Οι "καλοί" αντιπρόσωποι βρίσκονται κοντά στα μέλη των τμημάτων που

αντιπροσωπεύουν. Έτσι, η συνολική απόσταση όλων των μελών των τμημάτων με τους αντιπροσώπους τους είναι μικρή. Αντίθετα, οι “κακοί” αντιπρόσωποι βρίσκονται μακριά από τα μέλη των τμημάτων που αντιπροσωπεύουν και, συνεπακόλουθα, η συνολική απόσταση όλων των μελών του τμήματος με τους εκάστοτε αντιπροσώπους είναι μεγάλη.

2.3.3 Αλγόριθμος ανταγωνιστικής μάθησης

Ο αλγόριθμος ανταγωνιστικής μάθησης διαφέρει από τον τυπικό αλγόριθμο k-means ως προς τον τρόπο εξαγωγής των επιμέρους τμημάτων στα οποία διαχωρίζονται τα δεδομένα προκειμένου να διαμορφώσουν τις συστάδες. Αν και η ανταγωνιστική μάθηση ελαχιστοποιεί, επίσης, το άθροισμα των αποστάσεων από κάθε καταναλωτή που περιέχεται στο σύνολο δεδομένων προς τον πλησιέστερο αντιπρόσωπό του (κεντροειδές), η διαδικασία με την οποία επιτυγχάνεται αυτό είναι ελαφρώς διαφορετική. Πιο συγκεκριμένα, ο k-means χρησιμοποιεί το σύνολο των καταναλωτών που εμπεριέχονται στο σύνολο δεδομένων προκειμένου σε κάθε επανάληψη να επιλέξει τους νέους αντιπροσώπους. Η ανταγωνιστική μάθηση, αντίθετα, επιλέγει τυχαία έναν καταναλωτή και μετακινεί τον πλησιέστερο αντιπρόσωπο κατά ένα μικρό βήμα προς την κατεύθυνση του τυχαία επιλεγμένου καταναλωτή.

Ως συνέπεια αυτής της διαδικαστικής διαφοράς, διαφορετικές λύσεις τμηματοποίησης μπορεί να προκύψουν, ακόμη και αν χρησιμοποιούνται τα ίδια σημεία εκκίνησης για την αρχικοποίηση του αλγορίθμου. Είναι, επίσης, πιθανό η ανταγωνιστική μάθηση να βρει την ολικά βέλτιστη λύση ενώ ο k-means να έχει εγκλωβιστεί σε ένα τοπικό βέλτιστο (ή και αντίστροφα). Καμία από τις δύο μεθόδους δεν μπορεί να χαρακτηριστεί ανώτερη της άλλης λόγω του διαφορετικού τρόπου με τον οποίον καταλήγουν στην τελική τμηματοποίηση της αγοράς.

2.3.4 Αλγόριθμος Neural Gas και δίκτυα τοπολογικής αναπαράστασης

Μία παραλλαγή του αλγορίθμου ανταγωνιστικής μάθησης αποτελεί ο αλγόριθμος Neural Gas, ο οποίος προτάθηκε από τους (Martinez et al 1993, όπως αναφ. στο Dolnicar, Grün, & Leisch (2018)). Στην περίπτωση αυτή, όχι μόνο ο εγγύτερος αντιπρόσωπος αλλά και ο δεύτερος εγγύτερος αντιπρόσωπος μετακινείται προς τον τυχαία επιλεγμένο καταναλωτή. Ωστόσο, η θέση του δεύτερου εγγύτερου αντιπροσώπου αναπροσαρμόζεται σε μικρότερο βαθμό από ό,τι εκείνη του κοντινότερου αντιπροσώπου.

Μία προέκταση της παραπάνω προσέγγισης αποτελούν τα δίκτυα τοπολογικής αναπαράστασης (topology representing networks). Αυτό που προστίθεται στη

συγκεκριμένη περίπτωση είναι ότι ο αλγόριθμος μετρά πόσο συχνά ένα ζεύγος αντιπροσώπων (centroids) είναι το εγγύτερο και το δεύτερο εγγύτερο στον τυχαία επιλεγμένο καταναλωτή. Η πληροφορία αυτή αξιοποιείται για τη δημιουργία ενός εικονικού χάρτη στον οποίο "παρόμοιοι" αντιπρόσωποι - εκείνοι των οποίων οι τιμές προσαρμόστηκαν συχνά ταυτόχρονα - τοποθετούνται ο ένας δίπλα στον άλλο. Σχεδόν οι ίδιες πληροφορίες - οι οποίες είναι βασικές για την κατασκευή του χάρτη στα δίκτυα τοπολογικής αναπαράστασης - μπορούν να ληφθούν από οποιουσδήποτε άλλους αλγορίθμους συσταδοποίησης μετρώντας πόσοι καταναλωτές έχουν συγκεκριμένους αντιπροσώπους ως εγγύτερους και δεύτερους εγγύτερους στην τελική λύση τμηματοποίησης. Και σε αυτήν την περίπτωση, δεν μπορεί κάποιος να επικαλεστεί την ανωτερότητα ή μη των δύο προαναφερθέντων αλγορίθμων συγκριτικά με τον k-means ή τον αλγόριθμο ανταγωνιστικής μάθησης καθώς - και σε αυτήν την περίπτωση - ο τρόπος που καταλήγει στην τελική λύση είναι διαφορετικός.

2.3.5 Αλγόριθμος Self-Organizing Maps

Μία ακόμα παραλλαγή του αλγορίθμου ανταγωνιστικής μάθησης αποτελούν οι αυτό-οργανωτικοί χάρτες (Self-Organizing Maps – SOM). Σε αυτήν την περίπτωση οι αντιπρόσωποι αναπαρίστανται πάνω σε ένα ορθογώνιο ή εξαγωνικό πλέγμα.

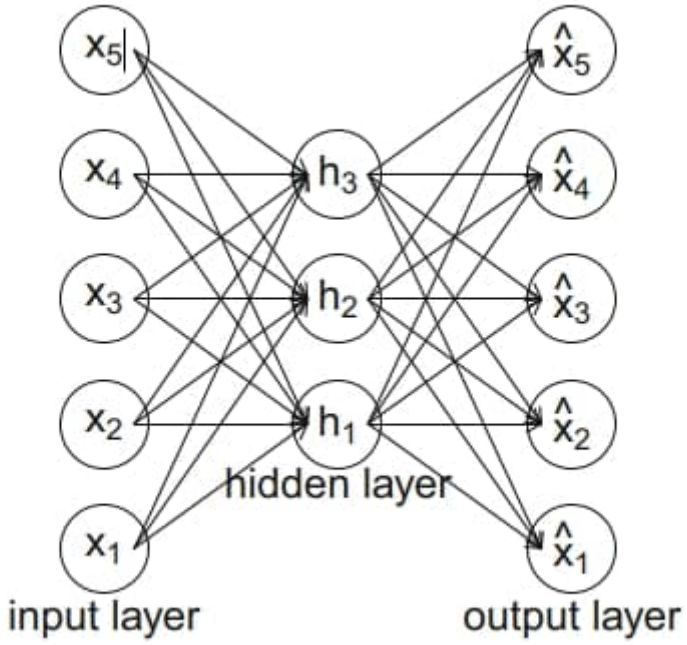
Πιο συγκεκριμένα, στα πλαίσια αυτού του αλγορίθμου, επιλέγεται και πάλι τυχαία ένας καταναλωτής από το σύνολο δεδομένων και ο εγγύτερος σε αυτόν αντιπρόσωπος μετακινείται ένα μικρό βήμα προς τον τυχαίο καταναλωτή. Επιπλέον, οι αντιπρόσωποι που είναι άμεσοι γείτονες του πλησιέστερου αντιπροσώπου μετακινούνται προς την κατεύθυνση του επιλεγμένου τυχαίου καταναλωτή. Η διαδικασία επαναλαμβάνεται πολλές φορές. Με τον τρόπο αυτό, κάθε καταναλωτής του συνόλου δεδομένων επιλέγεται τυχαία πολλές φορές και χρησιμοποιείται για την προσαρμογή της θέσης των κεντροειδών στο χάρτη SOM. Αυτό, ωστόσο, που αλλάζει κατά τη διάρκεια των επαναλήψεων είναι το πόσο επιτρέπεται στον αντιπροσώπους να μετακινηθούν. Οι επιτρεπόμενες αναπροσαρμογές γίνονται όλο και μικρότερες μέχρις ότου επιτευχθεί μία τελική λύση.

Το πλεονέκτημα της συγκεκριμένης προσέγγισης συγκριτικά με άλλους αλγορίθμους συσταδοποίησης είναι το γεγονός ότι το πλήθος των τμημάτων αγοράς που θα προκύψουν δεν είναι τυχαίο. Αντίθετα, η απαρίθμηση ευθυγραμμίζεται με το πλέγμα κατά μήκος του οποίου τοποθετούνται όλοι οι αντιπρόσωποι των τμημάτων. Το τίμημα αυτού του πλεονεκτήματος είναι ότι το άθροισμα των αποστάσεων μεταξύ των μελών ενός

τμήματος και των αντιπροσώπων των τμημάτων μπορεί να είναι μεγαλύτερο από ό,τι σε άλλους αλγορίθμους συσταδοποίησης. Αυτό οφείλεται στο γεγονός ότι η θέση των αντιπροσώπων δεν μπορεί να επιλεγεί ελεύθερα αλλά περιορίζεται από το πλέγμα.

2.3.6 Αλγόριθμος νευρωνικών δικτύων

Η συσταδοποίηση με τη χρήση νευρωνικών δικτύων λειτουργεί με διαφορετικό μαθηματικό τρόπο συγκριτικά με τους υπόλοιπους αλγορίθμους συσταδοποίησης, οι οποίοι αναλύθηκαν μέχρι στιγμής. Η πιο δημοφιλής μέθοδος αυτής της οικογενείας αλγορίθμων είναι αυτή που χρησιμοποιεί ένα κρυφό στρώμα μεταξύ του στρώματος εισόδου και του στρώματος εξόδου. Η αρχιτεκτονική ενός τέτοιου δικτύου αναπαρίσταται στην εικόνα που ακολουθεί.



Σχήμα 2-2: Σχηματική αναπαράσταση ενός νευρωνικού δικτύου [Πηγή: (Dolnicar, Grün, & Leisch, 2018)]

Πιο συγκεκριμένα, στην παραπάνω περίπτωση, το δίκτυο αποτελείται από συνολικά τρία στρώματα. Το πρώτο στρώμα δέχεται ως είσοδο τα δεδομένα. Το τρίτο στρώμα εξάγει το αποτέλεσμα της ανάλυσης. Μεταξύ των δύο αυτών στρωμάτων βρίσκεται το κρυφό στρώμα. Ονομάζεται κρυφό καθώς συνδέεται μόνο με κόμβους εντός του δικτύου. Το στρώμα εισόδου έχει τόσους κόμβους όσο και το πλήθος των τμημάτων

της αγοράς. Οι τιμές των κόμβων του κρυφού στρώματος h_1, h_2, h_3 είναι γραμμική συνάρτηση των δεδομένων εισόδου

$$hj = f_j\left(\sum_{i=1}^{5} \alpha_{ij} x_i\right)$$

για μη γραμμική εξίσωση f_j. Κάθε στάθμιση a_{ij} αναπαρίσταται στο παραπάνω γράφημα από ένα βέλος το οποίο ενώνει τους κόμβους του στρώματος εισόδου με αυτούς του κρυφού στρώματος. Οι εξισώσεις f_j επιλέγονται με τέτοιο τρόπο ώστε $h_j \leq 1$, και το άθροισμά τους να ισούται με μονάδα, δηλαδή $h_1 + h_2 + h_3 = 1$.

Στην απλούστερη περίπτωση, τα εξαγόμενα $x\hat{}_i$ θα είναι σταθμισμένοι συνδυασμοί των κρυφών κόμβων.

$$x\hat{}i = \sum_{j=1}^{3} \beta_{ji} h_i$$

Όπου οι συντελεστές β_{ij} αντιστοιχούν στα βέλη μεταξύ των κρυφών κόμβων και των κόμβων του τελευταίου στρώματος. Κατά την εκπαίδευση του δικτύου, οι παράμετροι α_{ij} και β_{ji} αναπροσαρμόζονται και επιλέγονται έτσι ώστε η τετραγωνική ευκλείδεια απόσταση μεταξύ εισόδων και εξόδων να είναι όσο το δυνατόν μικρότερη για τα διαθέσιμα δεδομένα εκπαίδευσης (δηλαδή τους καταναλωτές προς τμηματοποίηση). Η ορολογία "εκπαίδευση" χρησιμοποιείται για να δείξει τη διαδικασία εκτίμησης και επιλογής των δύο παραπάνω παραμέτρων.

Αφού εκπαιδευτεί το δίκτυο, οι παράμετροι που συνδέουν το κρυφό στρώμα με το στρώμα εξόδου ερμηνεύονται με τον ίδιο τρόπο όπως οι αντιπρόσωποι των τμημάτων που προκύπτουν από τους παραδοσιακούς αλγορίθμους συσταδοποίησης. Οι παράμετροι που συνδέουν το επίπεδο εισόδου στρώμα με το κρυφό στρώμα μπορούν να ερμηνευθούν με τον ακόλουθο τρόπο: έστω ότι για έναν συγκεκριμένο καταναλωτή h1 = 1, και συνεπώς h2 = h3 = 0. Στην περίπτωση αυτή $x\hat{}i = \beta_{1i}$ για i = 1, . . . , 5. Αυτό ισχύει για όλους τους καταναλωτές όπου το h1 είναι 1 ή κοντά στο 1. Το δίκτυο προβλέπει την ίδια τιμή για όλους τους καταναλωτές με h1 ≈ 1. Όλοι αυτοί οι καταναλωτές είναι μέλη του τμήματος αγοράς 1 με εκπρόσωπο β_{1i}. Όλοι οι καταναλωτές με $h_2 \approx 1$, είναι μέλη του τμήματος 2, και ούτω καθεξής.

Οι καταναλωτές που δεν έχουν τιμή h_j κοντά στο 1 μπορούν να θεωρηθούν ως ενδιάμεσα τμήματα. Η διαδικασία συσταδοποίησης με τον αλγόριθμο k-means και ο αλγόριθμος ανταγωνιστικής μάθησης παράγουν σαφείς τμηματοποιήσεις, όπου κάθε καταναλωτής ανήκει σε ακριβώς ένα τμήμα. Η συσταδοποίηση με τη χρήση νευρωνικών δικτύων είναι ένα παράδειγμα της λεγόμενης ασαφούς τμηματοποίησης με τιμές συμμετοχής μεταξύ 0 (δεν είναι μέλος αυτού του τμήματος) και 1 (μέλος μόνο αυτού του

τμήματος). Οι τιμές συμμετοχής μεταξύ 0 και 1 υποδηλώνουν συμμετοχή σε πολλαπλά τμήματα.

2.4 Υβριδικές προσεγγίσεις

Αρκετές προσεγγίσεις συνδυάζουν ιεραρχικούς αλγορίθμους και αλγορίθμους κατάτμησης σε μια προσπάθεια να αντισταθμίσουν τις αδυναμίες της μιας μεθόδου με τα πλεονεκτήματα της άλλης. Τα σημεία στα οποία υπερτερούν οι αλγόριθμοι ιεραρχικής ομαδοποίησης είναι ότι ο αριθμός των τμημάτων αγοράς που πρέπει να εξαχθούν δεν χρειάζεται να καθοριστεί εκ των προτέρων καθώς και το γεγονός ότι οι ομοιότητες των τμημάτων της αγοράς μπορούν να απεικονιστούν με τη χρήση δενδρογράμματος. Από την άλλη πλευρά, το μεγαλύτερο μειονέκτημα των αλγορίθμων ιεραρχικής συσταδοποίησης είναι ότι οι τυπικές υλοποιήσεις απαιτούν σημαντική χωρητικότητα μνήμης, περιορίζοντας έτσι το πιθανό μέγεθος του δείγματος των δεδομένων για την εφαρμογή αυτών των μεθόδων. Επίσης, η ερμηνεία και η ανάγνωση των δενδρογραμμάτων καθίσταται δύσκολη όταν το μέγεθος του συνόλου δεδομένων είναι μεγάλο.

Το πλεονέκτημα των αλγορίθμων συσταδοποίησης με κατάτμηση είναι ότι απαιτούν ελάχιστη μνήμη κατά τον υπολογισμό, και επομένως είναι κατάλληλοι για την τμηματοποίηση μεγάλων συνόλων δεδομένων. Αντίθετα, το μειονέκτημα των αλγορίθμων συσταδοποίησης με κατάτμησης είναι ότι ο αριθμός των τμημάτων αγοράς που πρόκειται να εξαχθούν πρέπει να καθοριστεί εκ των προτέρων. Οι αλγόριθμοι κατάτμησης δεν επιτρέπουν, επίσης, στον αναλυτή δεδομένων να παρακολουθεί τις αλλαγές στη συμμετοχή των καταναλωτών σε διαφορετικές λύσεις τμηματοποίησης.

Η βασική ιδέα πίσω από τις υβριδικές προσεγγίσεις τμηματοποίησης της αγοράς είναι η εκτέλεση αρχικά ενός αλγορίθμου κατάτμησης, καθώς είναι σε θέση να χειριστεί σύνολα δεδομένων οποιουδήποτε μεγέθους. Ωστόσο ο αλγόριθμος κατάτμησης που χρησιμοποιείται αρχικά δεν παράγει τον αριθμό ομάδων που επιδιώκεται αλλά έναν πολύ μεγαλύτερο αριθμό. Για τον λόγο αυτό, στη συνέχεια, τα αρχικά δεδομένα απορρίπτονται και διατηρούνται μόνο τα κέντρα των τμημάτων που προκύπτουν, οι αντιπρόσωποι και το μέγεθος δηλαδή κάθε τμήματος πελατών. Τα δεδομένα αυτά χρησιμοποιούνται ως είσοδος στους ιεραρχικούς αλγορίθμους συσταδοποίησης. Στο σημείο αυτό, το σύνολο των δεδομένων είναι αρκετά μικρό για τους ιεραρχικούς αλγορίθμους συσταδοποίησης και το δενδρόγραμμα μπορεί να δείξει το πλήθος των τμημάτων που πρέπει να εξαχθούν.

2.4.1 Αλγόριθμος συσταδοποίησης δύο βημάτων

Το λογισμικό SPSS προσφέρει μία έτοιμη υλοποιημένη διαδικασία συσταδοποίησης, η οποία αναφέρεται ως συσταδοποίηση δύο βημάτων. Το πρώτο βήμα αυτής της διαδικασία αποτελείται από μια διεργασία κατάτμησης ενώ το δεύτερο από μία ιεραρχική διεργασία.

2.4.2 Αλγόριθμος Bagged Clustering

Ο αλγόριθμος Bagged Clustering (Leisch 1998,1999, όπως αναφ. στο Dolnicar, Grün, & Leisch (2018)), συνδυάζει, επίσης, τόσο αλγορίθμους διαμέρισης όσο και αυτούς της ιεραρχικής ταξινόμησης. Αυτό που κάνει τον συγκεκριμένο αλγόριθμο να διαφέρει από τους υπόλοιπους υβριδικούς αλγορίθμους είναι ότι ενσωματώνει την μέθοδο bootstrapping (Efron & Tibshirani 1993 όπως αναφ. στο Dolnicar, Grün, & Leisch (2018)). Το bootstrapping μπορεί να υλοποιηθεί με την τυχαία επιλογή παρατηρήσεων από το σύνολο δεδομένων με επανατοποθέτηση. Αυτό σημαίνει ότι η διαδικασία εξαγωγής τμημάτων επαναλαμβάνεται πολλές φορές με τυχαία επιλεγμένα (bootstrapped) σύνολα δεδομένων, ίδιου μεγέθους με το αρχικό. Το bootstrapping έχει το πλεονέκτημα ότι καθιστά την τελική λύση τμηματοποίησης λιγότερο εξαρτημένη από τα εκάστοτε συγκεκριμένα άτομα που συγκαταλέγονται στα επιλεγμένα σύνολα δεδομένων.

Στον αλγόριθμο Bagged Clustering (Leisch 1998,1999, όπως αναφ. στο Dolnicar, Grün, & Leisch (2018)), πρώτα πραγματοποιείται ομαδοποίησης των δεδομένων στα οποία έχει εφαρμοστεί bootstrapping με αλγόριθμο διαμέρισης. Το πλεονέκτημα της εκκίνησης με έναν αλγόριθμο διαμέρισης είναι ότι δεν υπάρχουν περιορισμοί στο μέγεθος του δείγματος. Στη συνέχεια, απορρίπτονται το αρχικό σύνολο δεδομένων και όλα τα σύνολα δεδομένων στα οποία έχει εφαρμοστεί bootstrapping και αποθηκεύονται μόνο τα κεντροειδή των συστάδων που προκύπτουν από τις επαναλαμβανόμενες διαμερίσεις με τον αλγόριθμο διαμέρισης. Αυτά τα κεντροειδή χρησιμεύουν ως το σύνολο δεδομένων σε ένα δεύτερο βήμα, το οποίο είναι η ιεραρχική ταξινόμηση. Το πλεονέκτημα της χρήσης της ιεραρχικής ταξινόμησης στο δεύτερο βήμα είναι ότι το δενδρόγραμμα που προκύπτει μπορεί να βοηθήσει στην επιλογή του πλήθους των ομάδων που πρέπει να εξαχθούν.

Ο αλγόριθμος Bagged Clustering (Leisch 1998,1999, όπως αναφ. στο Dolnicar, Grün, & Leisch (2018)), ενδείκνυται στις παρακάτω περιστάσεις:

- Εάν υπάρχει υποψία ύπαρξης εξειδικευμένων αγορών

- Εάν υπάρχει ο κίνδυνος οι κλασικοί αλγόριθμοι συσταδοποίησης να εγκλωβιστούν σε τοπικά βέλτιστα
- Εάν υπάρχει σαφής προτίμηση προς τους ιεραρχικούς αλγόριθμους αλλά τα δεδομένα είναι πολλά

Ο αλγόριθμος Bagged Clustering έχει την ικανότητα να εντοπίσει τμήματα εξειδικευμένων αγορών, καθώς οι αλγόριθμοι ιεραρχικής συσταδοποίησης καταγράφουν τις εξειδικευμένες αγορές ως μικρά διακριτά κλαδιά στο δενδρόγραμμα. Η αυξημένη πιθανότητα να καταλήξουμε σε μια καλή λύση τμηματοποίησης προκύπτει από:

1. τη λήψη πολλών δειγμάτων bootstrap από το αρχικό σύνολο δεδομένων
2. την επανάληψη της ανάλυσης k-means - ή οποιουδήποτε άλλου αλγορίθμου κατάτμησης - πολλές φορές προκειμένου να αποφευχθεί μια μη βέλτιστη αρχικοποίηση (τυχαία επιλογή των αρχικών αντιπροσώπων των τμημάτων)
3. τη χρήση μόνο των κεντροειδών που προκύπτουν από την ανάλυση με τον αλγόριθμο k-means στο δεύτερο (ιεραρχικό) βήμα της ανάλυσης
4. τη χρήση της ντετερμινιστικής ιεραρχικής ανάλυσης στο τελευταίο βήμα.

Έστω ένα σύνολο δεδομένων X μεγέθους *n*. Ο αλγόριθμος Bagged Clustering αποτελείται από πέντε βήματα:

1. Δημιουργία *b* συνόλων δεδομένων μεγέθους *n* με τυχαία επιλογή και επανατοποθέτηση (bootstrapping) των καταναλωτών από το αρχικό σύνολο δεδομένων.
2. Πραγματοποίηση επαναληπτικής διαδικασίας με έναν αλγόριθμο κατάτμησης και τη δημιουργία b × k αντιπροσώπους, όπου k είναι το πλήθος των τμημάτων αγοράς.
3. Χρησιμοποίηση όλων των αντιπροσώπων των συστάδων που προκύπτουν από τις επαναλαμβανόμενες διαδικασίες κατάτμησης προκειμένου να δημιουργηθεί ένα νέο σύνολο δεδομένων. Τα αρχικά δεδομένα απορρίπτονται. Στα επόμενα βήματα, τα αρχικά δεδομένα αντικαθίστανται με το νέο αυτό σύνολο δεδομένων που περιέχει τα κέντρα συστάδων. Για το λόγο αυτό, ο συγκεκριμένος αλγόριθμος μπορεί να ανταπεξέλθει σε μεγάλα σύνολα δεδομένων.
4. Πραγματοποίηση ιεραρχικής συσταδοποίησης χρησιμοποιώντας το νέο σύνολο δεδομένων.
5. Καθορισμός της τελικής λύσης τμηματοποίησης επιλέγοντας ένα σημείο “κοπής” του δενδρογράμματος. Στη συνέχεια, κάθε αρχική παρατήρηση (καταναλωτής στο

σύνολο δεδομένων) ανατίθεται στο τμήμα της αγοράς του οποίου ο εκπρόσωπος είναι πλησιέστερα στο συγκεκριμένο καταναλωτή.

3 Εφαρμογή σε πραγματικό σύνολο δεδομένων

Τα δεδομένα που χρησιμοποιούνται στην παρούσα εργασία έχουν αντληθεί από το Kaggle.com, είναι ανωνυμοποιημένα και αφορούν 6.665 πιθανούς πελάτες μιας αυτοκινητοβιομηχανίας. Η αυτοκινητοβιομηχανία που αναφέρεται έχει ως στόχο να προωθήσει τα αυτοκίνητα της σε νέο καταναλωτικό κοινό. Έχοντας ολοκληρώσει μια ενδελεχή έρευνα αγοράς, κατέληξαν στο συμπέρασμα πως οι πελάτες τους μπορούν να ταξινομηθούν σε 4 ομάδες και να στοχεύονται με ένα ξεχωριστό στρατηγικό μείγμα μάρκετινγκ ανάλογα με την κατηγορία που ανήκουν. Για τους 6.665 πιθανούς πελάτες έχουν καταγραφεί τα παρακάτω χαρακτηριστικά - μεταβλητές (στήλες):

- *ID*: Μοναδικός αναγνωριστικός αριθμός του πελάτη.
- *Gender*: Ποιοτική μεταβλητή για το φύλο του πελάτη (Άντρας/Γυναίκα)
- *Ever_Married*: Ποιοτική μεταβλητή για την οικογενειακή κατάσταση του πελάτη. (Ναι αν είναι έγγαμος/ Όχι αν είναι άγαμος)
- *Age*: Ποσοτική μεταβλητή για την ηλικία του πελάτη.
- *Graduated*: Ποιοτική μεταβλητή για αν ο πελάτης έχει αποφοιτήσει από κάποιο πανεπιστημιακό ίδρυμα. (Ναι αν είναι απόφοιτος/Όχι αν δεν είναι απόφοιτος πανεπιστημίου)
- *Profession*: Το επάγγελμα του πελάτη.
- *Work_Experience*: Ποσοτική μεταβλητή που απεικονίζει την επαγγελματική εμπειρία του πελάτη σε έτη.
- *Spending_Score*: Ποιοτική μεταβλητή που αφορά την καταναλωτική συμπεριφορά του πελάτη. (Χαμηλό για πελάτες με λίγες δαπάνες, Μεσαίο για πελάτες με μέτρια καταναλωτική δραστηριότητα και Υψηλό για πελάτες με έντονη καταναλωτική συμπεριφορά)
- *Family_Size*: Ποσοτική μεταβλητή που δηλώνει τα μέλη της οικογένειας του πελάτη συμπεριλαμβανομένου αυτού.
- *Var_1*: Ανώνυμη ποιοτική μεταβλητή για την κατηγοριοποίηση του πελάτη.
- *Segmentation*: Ποιοτική μεταβλητή που αφορά τις 4 κατηγορίες του καταναλωτικού κοινού, όπως προέκυψαν από την έρευνα αγοράς της αυτοκινητοβιομηχανίας.

Όλες οι αναλύσεις που ακολουθούν έχουν πραγματοποιηθεί με τη χρήση της γλώσσας R και ο κώδικας έχει παρατεθεί στο Παράρτημα Α.

3.1 Απόσταση του Gower & Μέθοδος Partitioning Around Medoids

Η πρώτη ανάλυση που εφαρμόστηκε στα δεδομένα είναι ο συνδυασμός της απόστασης του Gower και της μεθόδου partitioning around medoids. Η μέθοδος του Gower χρησιμοποιείται για μεικτού τύπου δεδομένα και μετρά την ανομοιότητα 2 ομάδων με έναν αριθμό από το 0 (όμοια) μέχρι 1 (εντελώς ανόμοια) (Anand, 2020). Η μέθοδος PAM επιλέγει το πρώτο medoid βάσει του μικρότερου αθροίσματος αποστάσεων (ανομοιότητας) μεταξύ των άλλων σημείων. Τα υπόλοιπα medoids επιλέγονται με επανάληψη. Μόλις βρεθούν τα k medoids, τα υπόλοιπα αντικείμενα του συνόλου δεδομένων ταξινομούνται βάσει του κοντινότερου medoid (Helm, 2021). Γι' αυτή την εφαρμογή χρησιμοποιήθηκε το πακέτο ***cluster*** της **R** (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2022). Με την συνάρτηση ***daisy()*** δημιουργήθηκε ο πίνακας ανομοιότητας μεταξύ των πιθανών πελατών της αυτοκινητοβιομηχανίας. Με την χρήση της συνάρτησης ***pam()*** εξετάστηκαν οι 4 ομάδες πελατών που έχουν προκύψει από την έρευνα αγοράς. Για να αξιολογηθεί η εγκυρότητα των ομάδων αυτών, έγινε χρήση του δείκτη εσωτερικής εγκυρότητας τμηματοποίησης Average Silhouette Width (ASW) (Rousseeuw, 1987). Ο δείκτης αυτός παίρνει τιμές από το -1 έως το +1, με το +1 να δηλώνει την συμπάγεια της ομάδας. Με άλλα λόγια, όσο η τιμή του δείκτη τείνει προς το +1, τόσο πιο συμπαγής είναι η ομάδα των παρατηρήσεων, ενώ οι τιμές κάτω του 0 δηλώνουν πως η συγκεκριμένη παρατήρηση θα μπορούσε να βρίσκεται σε άλλη ομάδα με περισσότερα κοινά χαρακτηριστικά. Σε επίπεδο ομάδας υπολογίστηκαν οι μέσοι όροι των τιμών ASW του κάθε αντικειμένου και στην λύση με τις 4 ομάδες πελατών η τιμή ASW αντιστοιχεί σε 0,222. Για την ερμηνεία των 4 ομάδων χρησιμοποιήθηκε το πακέτο FactoMineR της R και πιο συγκεκριμένα η συνάρτηση catdes() (Lê, Josse, & Husson, 2008). Οι δείκτες της συνάρτησης αυτής Cla/Mod, Mod/Cla και Global δηλώνουν το ποσοστό των αντικειμένων μιας κατηγορίας που βρίσκονται εντός της ομάδας, το ποσοστό των αντικειμένων που χαρακτηρίζονται από τη συγκεκριμένη κατηγορία και το ποσοστό της κατηγορίας στο σύνολο των αντικειμένων αντίστοιχα. Με τον δείκτη ***adjustedRandIndex()*** υποδεικνύεται η ομοιότητα μεταξύ των ομάδων που σχηματίστηκαν και οι τιμές κυμαίνονται από 0 έως 1 με τη μονάδα να δηλώνει συμπαγείς και ομογενείς ομάδες και τιμές άνω του 0.6 να θεωρούνται αποδεκτές. Στην περίπτωση αυτή, ο

προσαρμοσμένος δείκτης Rand είναι περίπου στο 0,1103 γεγονός που υποδεικνύει πως αυτός ο αλγόριθμος δεν δίνει ομοιογενείς ομάδες στην περίπτωση των δεδομένων που εξετάζονται.

Πίνακας 3-1: Αποτελέσματα ποιοτικών μεταβλητών Gower & PAM

Description of each cluster by the categories

$`1`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Spending_Score=Low	39.30982746	99.80952381	60.000000	0.000000e+00	Inf
Profession=Healthcare	76.97307335	52.63492063	16.159040	0.000000e+00	Inf
Graduated=No	54.01490066	82.85714286	36.249062	0.000000e+00	Inf
Ever_Married=No	53.25248071	92.00000000	40.825206	0.000000e+00	Inf
Segmentation=D	54.58167331	60.88888889	26.361590	5.573137e-254	34.040695
Gender=Female	33.43373494	63.42857143	44.831208	1.150968e-64	16.980198
Profession=Marketing	47.21030043	6.98412698	3.495874	1.313422e-15	7.993330
Var_1=Cat_4	33.68669022	18.15873016	12.738185	9.931210e-13	7.131457
Var_1=Cat_2	38.67403315	8.88888889	5.431358	5.138683e-11	6.566861
Profession=Engineer	33.84879725	12.50793651	8.732183	4.884790e-09	5.851050
Var_1=Cat_3	31.23028391	12.57142857	9.512378	4.178875e-06	4.602282
Profession=Doctor	30.40540541	11.42857143	8.882221	7.444716e-05	3.961606
Var_1=Cat_5	39.18918919	1.84126984	1.110278	2.809934e-03	2.987800
Profession=Homemaker	31.42857143	3.49206349	2.625656	1.694662e-02	2.387864
Profession=Entertainment	15.32756489	7.87301587	12.138035	6.748369e-10	-6.171878
Segmentation=A	16.83168317	17.26984127	24.246062	3.317325e-14	-7.585295
Profession=Lawyer	5.00000000	1.58730159	7.501875	1.642533e-31	-11.678455
Var_1=Cat_6	18.92314567	53.77777778	67.156789	5.160087e-37	-12.710654
Profession=Executive	2.57425743	0.82539683	7.576894	2.706290e-43	-13.795669
Segmentation=B	10.36895674	10.34920635	23.585896	1.806327e-51	-15.092793
Segmentation=C	10.52325581	11.49206349	25.806452	4.457705e-56	-15.777312
Gender=Male	15.66494425	36.57142857	55.168792	1.150968e-64	-16.980198
Spending_Score=High	0.09960159	0.06349206	15.063766	3.393384e-127	-23.991717
Spending_Score=Average	0.12033694	0.12698413	24.936234	1.532526e-224	-31.989372
Profession=Artist	1.91605839	2.66666667	32.888222	1.261399e-245	-33.471294
Graduated=Yes	6.35443634	17.14285714	63.750938	0.000000e+00	-Inf
Ever_Married=Yes	3.19472617	8.00000000	59.174794	0.000000e+00	-Inf

$`2`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Spending_Score=Average	87.665463	68.403756	24.936234	0.000000e+00	Inf
Ever_Married=Yes	54.006085	100.000000	59.174794	0.000000e+00	Inf
Graduated=Yes	45.045893	89.859155	63.750938	1.047630e-229	32.358672
Profession=Artist	54.151460	55.727700	32.888222	1.169357e-158	26.837847
Segmentation=C	57.325581	46.291080	25.806452	2.187237e-144	25.585944
Gender=Female	45.214190	63.427230	44.831208	1.657607e-97	20.955901
Segmentation=B	42.366412	31.267606	23.585896	2.185522e-23	9.964234
Var_1=Cat_6	33.422699	70.234742	67.156789	2.321364e-04	3.681203
Profession=Engineer	38.659794	10.563380	8.732183	3.478965e-04	3.576747
Profession=Homemaker	40.000000	3.286385	2.625656	2.317588e-02	2.270523
Profession=Doctor	28.378378	7.887324	8.882221	4.900903e-02	-1.968513
Profession=Entertainment	26.946848	10.234742	12.138035	9.754022e-04	-3.297527
Var_1=Cat_4	25.912839	10.328638	12.738185	4.090481e-05	-4.102309
Profession=Marketing	11.587983	1.267606	3.495874	1.857095e-13	-7.358700
Segmentation=A	22.277228	16.901408	24.246062	1.077705e-22	-9.804414
Profession=Executive	6.534653	1.549296	7.576894	1.029413e-46	-14.352381
Gender=Male	21.185749	36.572770	55.168792	1.657607e-97	-20.955901
Profession=Healthcare	3.899721	1.971831	16.159040	1.468227e-134	-24.687188
Segmentation=D	6.715993	5.539906	26.361590	4.899941e-185	-29.010150
Graduated=No	8.940397	10.140845	36.249062	1.047630e-229	-32.358672
Spending_Score=Low	9.352338	17.558685	60.000000	0.000000e+00	-Inf
Ever_Married=No	0.000000	0.000000	40.825206	0.000000e+00	-Inf

```
$`3`
                          Cla/Mod     Mod/Cla    Global       p.value     v.test
Spending_Score=High      68.027888  60.4424779 15.063766  0.000000e+00        Inf
Ever_Married=Yes         28.651116 100.0000000 59.174794 3.452659e-292  36.531931
Profession=Executive     85.346535  38.1415929  7.576894 9.540759e-282  35.868287
Gender=Male              27.821594  90.5309735 55.168792 5.882321e-175  28.199790
Graduated=No             30.960265  66.1946903 36.249062 1.938631e-112  22.533743
Profession=Lawyer        44.000000  19.4690265  7.501875  6.989698e-50  14.849684
Var_1=Cat_6              18.163539  71.9469027 67.156789  1.455390e-04   3.798560
Var_1=Cat_4              21.319199  16.0176991 12.738185  4.034246e-04   3.537833
Segmentation=A           19.863861  28.4070796 24.246062  4.116452e-04   3.532503
Segmentation=B           19.783715  27.5221239 23.585896  7.405809e-04   3.374098
Profession=Entertainment 20.766378  14.8672566 12.138035  2.565035e-03   3.015562
Var_1=Cat_5               6.756757   0.4424779  1.110278  1.145791e-02  -2.528414
Profession=Doctor        12.162162   6.3716814  8.882221  7.652289e-04  -3.365075
Profession=Homemaker      8.000000   1.2389381  2.625656  5.746589e-04  -3.443301
Profession=Engineer      11.340206   5.8407080  8.732183  8.233935e-05  -3.937487
Var_1=Cat_3              10.725552   6.0176991  9.512378  4.071570e-06  -4.607695
Var_1=Cat_2               6.906077   2.2123894  5.431358  8.785961e-09  -5.752642
Spending_Score=Average   12.214200  17.9646018 24.936234  1.000862e-09  -6.109273
Segmentation=D           12.179852  18.9380531 26.361590  1.725024e-10  -6.384017
Profession=Healthcare     3.156917   3.0088496 16.159040  2.041431e-52 -15.235948
Profession=Artist         4.105839   7.9646018 32.888222 6.459221e-103 -21.540793
Graduated=Yes             8.990351  33.8053097 63.750938 1.938631e-112 -22.533743
Gender=Female             3.580991   9.4690265 44.831208 5.882321e-175 -28.199790
Spending_Score=Low        6.101525  21.5929204 60.000000 2.142180e-184 -28.959315
Ever_Married=No           0.000000   0.0000000 40.825206 3.452659e-292 -36.531931

$`4`
                          Cla/Mod    Mod/Cla    Global       p.value     v.test
Spending_Score=Low       45.236309 98.8524590 60.000000  0.000000e+00        Inf
Graduated=Yes            39.609320 91.9672131 63.750938 8.039460e-224  31.937569
Ever_Married=No          46.747519 69.5081967 40.825206 1.317639e-188  29.291838
Gender=Male              35.327713 70.9836066 55.168792  5.665122e-59  16.192841
Profession=Artist        39.826642 47.7049180 32.888222  9.099254e-55  15.585748
Segmentation=A           41.027228 36.2295082 24.246062  1.272838e-42  13.683572
Profession=Entertainment 36.959209 16.3387978 12.138035  3.078127e-10   6.294792
Var_1=Cat_6              29.490617 72.1311475 67.156789  8.118624e-08   5.364472
Profession=Homemaker     20.571429  1.9672131  2.625656  3.513394e-02  -2.106811
Var_1=Cat_5              14.864865  0.6010929  1.110278  1.097014e-02  -2.543649
Profession=Lawyer        19.000000  5.1912568  7.501875  5.434590e-06  -4.547271
Var_1=Cat_4              19.081272  8.8524590 12.738185  1.596054e-09  -6.034348
Profession=Engineer      16.151203  5.1366120  8.732183  2.054345e-11  -6.702108
Profession=Healthcare    15.970288  9.3989071 16.159040  5.424175e-22  -9.639898
Profession=Executive      5.544554  1.5300546  7.576894  4.533869e-39 -13.075697
Segmentation=C           15.639535 14.6994536 25.806452  4.331181e-40 -13.253070
Gender=Female            17.771084 29.0163934 44.831208  5.665122e-59 -16.192841
Spending_Score=High       2.091633  1.1475410 15.063766 2.317783e-118 -23.129879
Ever_Married=Yes         14.148073 30.4918033 59.174794 1.317639e-188 -29.291838
Graduated=No              6.084437  8.0327869 36.249062 8.039460e-224 -31.937569
Spending_Score=Average    0.000000  0.0000000 24.936234 3.497775e-275 -35.444724
```

Πίνακας 3-2: Αποτελέσματα ποσοτικών μεταβλητών Gower & PAM

```
Description of each cluster by quantitative variables
=====================================================
$`1`
              v.test Mean in category Overall mean sd in category Overall sd      p.value
Family_Size  23.79423          3.64000      2.84111       1.666644   1.524628 3.83185e-125
Age         -40.66676         28.73905     43.53608      10.150854  16.522814  0.00000e+00

$`2`
                   v.test Mean in category Overall mean sd in category Overall sd      p.value
Age             19.603947        49.325822    43.536084      13.358156   16.52281 1.430842e-85
Work_Experience -2.167514         2.497183     2.629107       3.213563    3.40511 3.019567e-02

$`3`
                   v.test Mean in category Overall mean sd in category Overall sd       p.value
Age             27.573032        55.887611    43.536084      17.558920  16.522814 2.343685e-167
Family_Size      2.430944         2.941593     2.841110       1.412381   1.524628  1.505954e-02
Work_Experience -5.971048         2.077876     2.629107       3.026600   3.405110  2.357342e-09

$`4`
                    v.test Mean in category Overall mean sd in category  Overall sd       p.value
Work_Experience   7.573862     3.142623e+00 2.629107e+00       3.732868    3.405110  3.622885e-14
ID               -2.740816     4.633798e+05 4.635198e+05    2651.955271 2566.239202  6.128684e-03
Age              -4.956367     4.190546e+01 4.353608e+01      13.535960   16.522814  7.182363e-07
Family_Size     -23.872570     2.116393e+00 2.841110e+00       1.359486    1.524628 5.904188e-126
```

- **1η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 1η ομάδα χαρακτηρίζεται από άτομα που έχουν χαμηλό προφίλ δαπανών,

- δουλεύουν κατά κύριο λόγο στον τομέα της υγείας,
- δεν έχουν αποφοιτήσει από κάποιο πανεπιστημιακό ίδρυμα και
- είναι άγαμοι.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο ανήκουν σε 4μελείς οικογένειες και
- Κατά μέσο όρο είναι νεαροί με μέση ηλικία τα 29 έτη.

❖ **2η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 2η ομάδα χαρακτηρίζεται από άτομα που έχουν μεσαίο προφίλ δαπανών,
- είναι έγγαμοι,
- είναι απόφοιτοι πανεπιστημίου και
- είναι κατά κύριο λόγο καλλιτέχνες.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο είναι περίπου 49 ετών και
- έχουν 2,5 έτη επαγγελματικής εμπειρίας.

❖ **3η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 3η ομάδα χαρακτηρίζεται από άτομα με υψηλό προφίλ δαπανών,
- είναι έγγαμοι,
- είναι στελέχη επιχειρήσεων,
- είναι κατά κύριο λόγο άντρες και
- δεν έχουν αποφοιτήσει από κάποιο πανεπιστήμιο.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο είναι περίπου 56 ετών και
- ανήκουν σε 3μελείς οικογένειες.

❖ **4η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 4η ομάδα χαρακτηρίζεται από άτομα με χαμηλό προφίλ δαπανών,
- είναι απόφοιτοι πανεπιστημίου,
- είναι άγαμοι και
- είναι κατά κύριο λόγο άντρες.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο έχουν 3 χρόνια επαγγελματικής εμπειρίας.

3.2 Αλγόριθμος K-prototypes

Ο αλγόριθμος k-prototypes δημιουργήθηκε από τον Huang Z. για να μπορεί να χειριστεί την συσταδοποίηση μεικτού τύπου δεδομένων. Ο συγκεκριμένος αλγόριθμος είναι μέθοδος συσταδοποίησης που στηρίζεται στον διαμερισμό των δεδομένων σε κατηγορίες και αποτελεί βελτίωση των μεθόδων k-means και k-mode που επίσης απευθύνονται σε μεικτού τύπου δεδομένα. (Aprilliant, 2021) Η εφαρμογή της μεθόδου αυτής πραγματοποιήθηκε με τη χρήση του πακέτου ***clustMixType*** της **R** (Szepannek, 2018) και της συνάρτησης ***kproto()***. Ο προσαρμοσμένος δείκτης Rand δίνει τιμή 0,1146, τιμή που είναι αρκετά χαμηλή.

Πίνακας 3-3: Αποτελέσματα ποιοτικών μεταβλητών K-prototypes

Description of each cluster by the categories
==========

$`1`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Spending_Score=Low	38.8097024	97.7945810	60.000000	0.000000e+00	Inf
Ever_Married=No	45.0937155	77.3156900	40.825206	2.677396e-254	34.062206
Graduated=Yes	33.2078136	88.9098929	63.750938	4.330631e-144	25.559273
Segmentation=A	36.7574257	37.4291115	24.246062	4.056750e-42	13.599049
Profession=Artist	32.4361314	44.8015123	32.888222	4.826343e-30	11.387501
Profession=Entertainment	31.7676143	16.1940769	12.138035	3.580383e-08	5.510391
Gender=Male	25.6187109	59.3572779	55.168792	1.175675e-04	3.851144
Profession=Doctor	29.3918919	10.9640832	8.882221	1.072936e-03	3.270668
Profession=Engineer	20.2749141	7.4354127	8.732183	3.394266e-02	-2.120752
Var_1=Cat_4	19.7879859	10.5860113	12.738185	2.783225e-03	-2.990718
Gender=Female	21.5863454	40.6427221	44.831208	1.175675e-04	-3.851144
Profession=Healthcare	17.1773445	11.6572149	16.159040	9.366001e-09	-5.741828
Profession=Executive	7.5247525	2.3944549	7.576894	2.957000e-23	-9.934146
Segmentation=C	12.4418605	13.4845621	25.806452	1.839170e-41	-13.488043
Profession=Lawyer	1.6000000	0.5040958	7.501875	3.625364e-49	-14.738920
Spending_Score=High	2.4900398	1.5752993	15.063766	1.663597e-91	-20.287339
Graduated=No	7.2847682	11.0901071	36.249062	4.330631e-144	-25.559273
Spending_Score=Average	0.6016847	0.6301197	24.936234	8.992989e-210	-30.909175
Ever_Married=Yes	9.1277890	22.6843100	59.174794	2.677396e-254	-34.062206

$`2`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Spending_Score=Average	88.3273165	61.8105263	24.936234	0.000000e+00	Inf
Ever_Married=Yes	59.7109533	99.1578947	59.174794	0.000000e+00	Inf
Profession=Artist	57.0255474	52.6315789	32.888222	2.716902e-141	25.306387
Segmentation=C	59.0116279	42.7368421	25.806452	1.597638e-118	23.145930
Graduated=Yes	44.1280301	78.9473684	63.750938	6.683428e-86	19.642638
Segmentation=B	48.4096692	32.0421053	23.585896	7.447174e-33	11.938594
Gender=Male	39.7878705	61.6000000	55.168792	3.360160e-15	7.876746
Profession=Executive	52.0792079	11.0736842	7.576894	4.158180e-15	7.850066
Spending_Score=High	40.0398406	16.9263158	15.063766	1.679929e-03	3.141660
Profession=Entertainment	38.9369592	13.2631579	12.138035	3.730811e-02	2.082376
Var_1=Cat_2	30.3867403	4.6315789	5.431358	3.080054e-02	-2.159640
Profession=Doctor	31.0810811	7.7473684	8.882221	1.469508e-02	-2.439808
Profession=Marketing	14.1630901	1.3894737	3.495874	9.647925e-14	-7.445635
Gender=Female	30.5220884	38.4000000	44.831208	3.360160e-15	-7.876746
Segmentation=A	27.4133663	18.6526316	24.246062	9.182762e-16	-8.037316
Profession=Lawyer	1.6000000	0.3368421	7.501875	7.410557e-86	-19.637393
Graduated=No	20.6953642	21.0526316	36.249062	6.683428e-86	-19.642638
Profession=Healthcare	4.5496750	2.0631579	16.159040	1.450169e-152	-26.310608
Segmentation=D	8.8787706	6.5684211	26.361590	7.320773e-191	-29.468382
Spending_Score=Low	12.6281570	21.2631579	60.000000	0.000000e+00	-Inf
Ever_Married=No	0.7350239	0.8421053	40.825206	0.000000e+00	-Inf

```
$`3`
                          Cla/Mod   Mod/Cla    Global       p.value     v.test
Profession=Healthcare    77.623027 49.791543 16.159040  0.000000e+00        Inf
Graduated=No             57.408940 82.608696 36.249062  0.000000e+00        Inf
Ever_Married=No          52.186696 84.574151 40.825206  0.000000e+00        Inf
Segmentation=D           58.338076 61.048243 26.361590 4.965652e-281  35.822304
Spending_Score=Low       38.534634 91.780822 60.000000 2.972236e-242  33.238708
Gender=Female            33.935743 60.393091 44.831208  1.329467e-49  14.806520
Var_1=Cat_4              37.691402 19.058964 12.738185  5.091421e-18   8.651304
Profession=Marketing     49.785408  6.908874  3.495874  2.294942e-16   8.205573
Var_1=Cat_2              39.502762  8.516974  5.431358  7.422878e-10   6.156802
Profession=Doctor        34.628378 12.209649  8.882221  7.897604e-08   5.369452
Profession=Homemaker     42.857143  4.466945  2.625656  2.554148e-07   5.153686
Profession=Engineer      33.676976 11.673615  8.732183  1.614338e-06   4.796535
Var_1=Cat_3              32.334385 12.209649  9.512378  2.127389e-05   4.251083
Var_1=Cat_5              40.540541  1.786778  1.110278  3.709217e-03   2.901888
Profession=Entertainment 18.788628  9.053008 12.138035  4.420313e-06  -4.590573
Segmentation=A           19.492574 18.761167 24.246062  6.564754e-10  -6.176237
Profession=Executive      8.316832  2.501489  7.576894  1.019214e-23 -10.039760
Var_1=Cat_6              20.129580 53.662895 67.156789  7.438214e-41 -13.384615
Gender=Male              18.085396 39.606909 55.168792  1.329467e-49 -14.806520
Segmentation=B           10.687023 10.005956 23.585896  9.214779e-59 -16.162884
Profession=Lawyer         0.800000  0.238237  7.501875  6.122190e-59 -16.188067
Spending_Score=High       5.876494  3.513996 15.063766  2.994145e-66 -17.193040
Segmentation=C            9.941860 10.184634 25.806452  4.411868e-73 -18.082072
Spending_Score=Average    4.753309  4.705182 24.936234 5.501546e-136 -24.819649
Profession=Artist         2.417883  3.156641 32.888222 4.664015e-254 -34.045921
Graduated=Yes             6.872205 17.391304 63.750938  0.000000e+00       -Inf
Ever_Married=Yes          6.566937 15.425849 59.174794  0.000000e+00       -Inf

$`4`
                           Cla/Mod    Mod/Cla    Global       p.value     v.test
Profession=Lawyer        96.0000000 46.8750000  7.501875  0.000000e+00        Inf
Spending_Score=High      51.5936255 50.5859375 15.063766 1.733897e-201  30.286355
Ever_Married=Yes         24.5943205 94.7265625 59.174794 6.056585e-173  28.035176
Var_1=Cat_6              19.3476318 84.5703125 67.156789  3.293363e-42  13.614289
Profession=Executive     32.0792079 15.8203125  7.576894  8.324717e-23   9.830449
Segmentation=C           18.6046512 31.2500000 25.806452  2.082844e-05   4.255820
Segmentation=B           18.7659033 28.8085938 23.585896  2.659908e-05   4.200785
Gender=Male              16.5080228 59.2773438 55.168792  3.988091e-03   2.879102
Profession=Marketing      9.4420601  2.1484375  3.495874  7.507912e-03  -2.673434
Gender=Female            13.9558233 40.7226562 44.831208  3.988091e-03  -2.879102
Profession=Homemaker      6.2857143  1.0742188  2.625656  2.251670e-04  -3.688966
Profession=Entertainment 10.5067985  8.3007812 12.138035  2.139388e-05  -4.249824
Var_1=Cat_7               3.6144578  0.5859375  2.490623  1.061194e-06  -4.879938
Profession=Engineer       8.5910653  4.8828125  8.732183  4.501690e-07  -5.046418
Var_1=Cat_2               6.3535912  2.2460938  5.431358  7.985758e-08  -5.367450
Var_1=Cat_4               8.2449941  6.8359375 12.738185  5.075064e-11  -6.568717
Var_1=Cat_3               6.4668770  4.0039062  9.512378  8.615810e-13  -7.150984
Profession=Doctor         4.8986486  2.8320312  8.882221  9.385875e-17  -8.312308
Segmentation=D            8.2527035 14.1601562 26.361590  3.322565e-24 -10.149737
Profession=Artist         8.1204380 17.3828125 32.888222  2.978939e-33 -12.014576
Spending_Score=Average    6.3176895 10.2539062 24.936234  3.807761e-37 -12.734397
Spending_Score=Low       10.0275069 39.1601562 60.000000  1.718784e-48 -14.633438
Profession=Healthcare     0.6499536  0.6835938 16.159040  1.512892e-73 -18.140987
Ever_Married=No           1.9845645  5.2734375 40.825206 6.056585e-173 -28.035176
```

Πίνακας 3-4: Αποτελέσματα ποσοτικών μεταβλητών K-prototypes

```
Description of each cluster by quantitative variables
======================================================
$`1`
                v.test Mean in category Overall mean sd in category Overall sd      p.value
Age            20.220803        50.992871    43.536084       7.741904  16.522814 6.422958e-91
Family_Size     1.983948         2.908620     2.841110       1.388807   1.524628 4.726165e-02
Work_Experience -5.259160         2.229423     2.629107       2.998896   3.405110 1.447146e-07

$`2`
                  v.test Mean in category Overall mean sd in category Overall sd       p.value
Family_Size      21.308995         3.446544     2.841110       1.697836   1.524628 9.368416e-101
Work_Experience   5.155302         2.956241     2.629107       3.644042   3.405110  2.532222e-07
Age             -52.005344        27.523123    43.536084       6.127344  16.522814  0.000000e+00

$`3`
                v.test Mean in category Overall mean sd in category Overall sd      p.value
Age             59.73138        72.516194    43.536084       8.737092  16.522814 0.000000e+00
Work_Experience -13.35755         1.293522     2.629107       2.221281   3.405110 1.070227e-40
Family_Size     -15.17047         2.161943     2.841110       1.125743   1.524628 5.547660e-52

$`4`
                  v.test Mean in category Overall mean sd in category Overall sd      p.value
Work_Experience   9.869755         3.231276     2.629107       3.685048   3.405110 5.630189e-23
Family_Size     -11.219422         2.534621     2.841110       1.384288   1.524628 3.274030e-29
Age             -12.626774        39.797927    43.536084       7.368898  16.522814 1.503160e-36
```

- **1η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 1η ομάδα χαρακτηρίζεται από άτομα που έχουν χαμηλό προφίλ δαπανών,

- είναι άγαμοι,
- είναι απόφοιτοι πανεπιστημίου και
- είναι καλλιτέχνες.

➢ Ποσοτικές μεταβλητές:

- Κατά μέσο όρο είναι 51 ετών και
- ανήκουν σε 3μελείς οικογένειες.

❖ **2η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 2η ομάδα χαρακτηρίζεται από άτομα που έχουν μεσαίο προφίλ δαπανών,
- είναι έγγαμοι και
- είναι κατά κύριο λόγο καλλιτέχνες.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο ανήκουν σε 3μελείς οικογένειες και
- Έχουν 3 χρόνια επαγγελματικής εμπειρίας.

❖ **3η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 3η ομάδα χαρακτηρίζεται από άτομα που εργάζονται στον χώρο της υγείας,
- δεν είναι απόφοιτοι πανεπιστημίου και
- είναι άγαμοι.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο είναι περίπου 73 ετών.

❖ **4η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 4η ομάδα χαρακτηρίζεται από δικηγόρους,
- έχουν υψηλό σκορ δαπανών και
- είναι παντρεμένοι.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο έχουν 3 χρόνια επαγγελματικής εμπειρίας.

3.3 Μέθοδος Mixed K-means

Η μέθοδος K-means for mixed data είναι η τροποποιημένη έκδοση της k-means που λαμβάνει υπόψη την ευκλείδεια απόσταση μεταξύ τιμών ποσοτικών δεδομένων. Στη mixed k-means οι τιμές χωρίζονται στο δοσμένο πλήθος ομάδων τυχαία, κάθε ομάδα έχει

ένα κέντρο το οποίο χρησιμοποιείται ώστε κάθε στοιχείο να βρει το κοντινότερο κέντρο και να ταξινομηθεί στην αντίστοιχη ομάδα. Όταν τελειώσει αυτή η διαδικασία επαναυπολογίζεται το κέντρο της ομάδας με τα νέα πλέον ταξινομημένα στοιχεία και επαναλαμβάνεται η διαδικασία υπολογισμού του κέντρου και ταξινόμησης στοιχείων μέχρι να μην αλλάζουν οι ομάδες περαιτέρω. (Ahmad & Dey, 2007) Για την εφαρμογή της mixed k-means έγινε χρήση του πακέτου ***kmed*** της **R**. Ο προσαρμοσμένος δείκτης Rand έδωσε τιμή 0,1044 για την εφαρμογή αυτής της μεθόδου στα δεδομένα που εξετάζονται.

Πίνακας 3-5: Αποτελέσματα ποιοτικών μεταβλητών Mixed K-means

Description of each cluster by the categories
==========

$`1`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Ever_Married=Yes	38.235294	86.916427	59.174794	3.550766e-183	28.862307
Spending_Score=Average	45.487365	43.573487	24.936234	2.816012e-90	20.147754
Segmentation=C	43.662791	43.285303	25.806452	2.337170e-78	18.739980
Profession=Artist	40.693431	51.412104	32.888222	2.744503e-78	18.731429
Graduated=Yes	32.713580	80.115274	63.750938	5.296604e-65	17.025688
Segmentation=B	34.860051	31.585014	23.585896	4.053664e-19	8.935493
Profession=Executive	38.811881	11.296830	7.576894	5.865641e-11	6.547122
Var_1=Cat_6	27.703307	71.469741	67.156789	7.468428e-06	4.479882
Spending_Score=High	31.075697	17.982709	15.063766	9.931172e-05	3.892267
Gender=Male	27.495241	58.270893	55.168792	2.492560e-03	3.024243
Profession=Entertainment	29.542645	13.775216	12.138035	1.628484e-02	2.402469
Var_1=Cat_2	20.994475	4.380403	5.431358	2.254073e-02	-2.281130
Profession=Homemaker	18.285714	1.844380	2.625656	1.513630e-02	-2.429101
Gender=Female	24.230254	41.729107	44.831208	2.492560e-03	-3.024243
Profession=Marketing	17.596567	2.363112	3.495874	2.029839e-03	-3.085831
Var_1=Cat_4	20.259128	9.913545	12.738185	2.795221e-05	-4.189539
Profession=Doctor	18.243243	6.224784	8.882221	3.037242e-06	-4.668285
Segmentation=A	19.616337	18.270893	24.246062	5.582613e-12	-6.889914
Profession=Lawyer	6.800000	1.959654	7.501875	2.537226e-30	-11.443407
Graduated=No	14.279801	19.884726	36.249062	5.296604e-65	-17.025688
Spending_Score=Low	16.679170	38.443804	60.000000	2.046423e-99	-21.164098
Profession=Healthcare	2.414113	1.498559	16.159040	5.321215e-114	-22.692432
Segmentation=D	6.772908	6.858790	26.361590	5.058606e-122	-23.490853
Ever_Married=No	8.342521	13.083573	40.825206	3.550766e-183	-28.862307

$`2`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Profession=Healthcare	83.194058	45.4822335	16.159040	0.000000e+00	Inf
Ever_Married=No	59.757442	82.5380711	40.825206	0.000000e+00	Inf
Segmentation=D	65.395561	58.3248731	26.361590	9.180713e-305	37.315558
Spending_Score=Low	44.111028	89.5431472	60.000000	1.440243e-252	33.945109
Graduated=No	51.738411	63.4517766	36.249062	4.595678e-193	29.639763
Profession=Doctor	43.074324	12.9441624	8.882221	2.544396e-13	7.316544
Var_1=Cat_4	39.340400	16.9543147	12.738185	6.311884e-11	6.536159
Profession=Marketing	48.927039	5.7868020	3.495874	2.786207e-10	6.310230
Var_1=Cat_2	41.988950	7.7157360	5.431358	2.369343e-07	5.167746
Var_1=Cat_3	36.277603	11.6751269	9.512378	1.270381e-04	3.832131
Var_1=Cat_5	44.594595	1.6751269	1.110278	6.128632e-03	2.740819
Profession=Homemaker	38.857143	3.4517766	2.625656	7.684631e-03	2.665621
Var_1=Cat_7	36.746988	3.0964467	2.490623	4.359721e-02	2.017943
Profession=Entertainment	24.474660	10.0507614	12.138035	6.114613e-04	-3.426479
Segmentation=A	24.628713	20.2030457	24.246062	4.520326e-07	-5.045629
Profession=Executive	10.495050	2.6903553	7.576894	1.852009e-26	-10.644386
Var_1=Cat_6	25.044683	56.9035533	67.156789	3.718749e-30	-11.410200
Segmentation=B	12.977099	10.3553299	23.585896	4.171363e-68	-17.439024
Profession=Lawyer	0.800000	0.2030457	7.501875	2.485148e-72	-17.986512
Segmentation=C	12.732558	11.1167513	25.806452	1.435253e-78	-18.765908
Spending_Score=High	6.374502	3.2487310	15.063766	4.041905e-85	-19.551041
Spending_Score=Average	8.543923	7.2081218	24.936234	3.214376e-122	-23.510114
Profession=Artist	10.492701	11.6751269	32.888222	2.235950e-141	-25.314072
Graduated=Yes	16.945164	36.5482234	63.750938	4.595678e-193	-29.639763
Ever_Married=Yes	8.722110	17.4619289	59.174794	0.000000e+00	-Inf

```
$`3`
                           Cla/Mod    Mod/Cla    Global       p.value     v.test
Profession=Lawyer        91.4000000 51.7553794  7.501875  0.000000e+00        Inf
Ever_Married=Yes         21.4249493 95.6964892 59.174794 4.373768e-157  26.702744
Spending_Score=High      41.2350598 46.8856172 15.063766 2.924536e-136  24.845055
Var_1=Cat_6              16.9347632 85.8437146 67.156789  1.480734e-41  13.504019
Profession=Executive     21.5841584 12.3442809  7.576894  7.382898e-08   5.381594
Segmentation=C           15.7558140 30.6908267 25.806452  4.533018e-04   3.506934
Segmentation=B           15.7124682 27.9728199 23.585896  1.175633e-03   3.244727
Profession=Marketing      6.4377682  1.6987542  3.495874  7.727736e-04  -3.362367
Profession=Homemaker      3.4285714  0.6795017  2.625656  9.376704e-06  -4.431069
Var_1=Cat_7               2.4096386  0.4530011  2.490623  1.031520e-06  -4.885528
Var_1=Cat_2               4.9723757  2.0385051  5.431358  1.139161e-07  -5.302998
Profession=Entertainment  7.4165637  6.7950170 12.138035  2.804041e-08  -5.553252
Var_1=Cat_4               7.3027091  7.0215176 12.738185  5.624209e-09  -5.827565
Profession=Engineer       6.0137457  3.9637599  8.732183  4.368234e-09  -5.869609
Var_1=Cat_3               4.2586751  3.0577576  9.512378  2.851756e-15  -7.897227
Profession=Doctor         3.7162162  2.4915062  8.882221  3.265142e-16  -8.163107
Spending_Score=Average    7.3405535 13.8165345 24.936234  5.519346e-18  -8.642091
Segmentation=D            7.3420603 14.6092865 26.361590  3.393896e-19  -8.955112
Profession=Artist         8.0291971 19.9320498 32.888222  7.214975e-20  -9.124377
Spending_Score=Low        8.6771693 39.2978482 60.000000  1.658003e-40 -13.324925
Profession=Healthcare     0.2785515  0.3397508 16.159040  1.374076e-67 -17.370754
Ever_Married=No           1.3965454  4.3035108 40.825206 4.373768e-157 -26.702744

$`4`
                           Cla/Mod    Mod/Cla    Global      p.value     v.test
Graduated=Yes            36.76159 75.2046221 63.750938 1.924551e-40  13.313794
Profession=Artist        40.78467 43.0428503 32.888222 7.453305e-32  11.745441
Segmentation=A           41.15099 32.0173327 24.246062 1.103553e-22   9.802021
Spending_Score=Average   38.62816 30.9099663 24.936234 6.858726e-14   7.490547
Segmentation=B           36.45038 27.5878671 23.585896 2.917902e-07   5.128674
Profession=Entertainment 38.56613 15.0216659 12.138035 1.820885e-06   4.772350
Profession=Engineer      39.17526 10.9773712  8.732183 1.789012e-05   4.289717
Gender=Female            33.70147 48.4833895 44.831208 5.613423e-05   4.028498
Profession=Homemaker     39.42857  3.3220992  2.625656 1.897972e-02   2.345929
Profession=Doctor        34.96622  9.9662975  8.882221 3.788836e-02   2.076060
Var_1=Cat_6              30.31725 65.3346172 67.156789 3.347652e-02  -2.126321
Segmentation=C           27.84884 23.0621088 25.806452 5.298853e-04  -3.465176
Gender=Male              29.09981 51.5166105 55.168792 5.613423e-05  -4.028498
Spending_Score=High      21.31474 10.3033221 15.063766 5.550546e-14  -7.518271
Segmentation=D           20.48947 17.3326914 26.361590 7.523026e-31 -11.548374
Graduated=No             21.31623 24.7953779 36.249062 1.924551e-40 -13.313794
Profession=Healthcare    14.11328  7.3182475 16.159040 1.757963e-44 -13.991448
Profession=Lawyer         1.00000  0.2407318  7.501875 8.696724e-76 -18.422336
```

Πίνακας 3-6: Αποτελέσματα ποσοτικών μεταβλητών Mixed K-means

```
Description of each cluster by quantitative variables
======================================================
$`1`
                  v.test Mean in category Overall mean sd in category Overall sd      p.value
Family_Size      22.652965         3.491671     2.841110       1.631206   1.524628 1.304307e-113
Work_Experience   5.995033         3.013629     2.629107       3.652956   3.405110  2.034436e-09
Age             -56.397811        25.983342    43.536084       4.251463  16.522814  0.000000e+00

$`2`
                  v.test Mean in category Overall mean sd in category Overall sd      p.value
Work_Experience  14.235853         3.514756     2.629107       3.773478   3.405110 5.488906e-46
Family_Size      -8.798173         2.596033     2.841110       1.437908   1.524628 1.390616e-18
Age             -15.183296        38.952588    43.536084       3.535113  16.522814 4.562657e-52

$`3`
                  v.test Mean in category Overall mean sd in category Overall sd       p.value
Age              28.54026        53.277393    43.536084       5.115256   16.52281 3.711041e-179
Work_Experience -10.78842         1.870242     2.629107       2.715435    3.40511  3.904282e-27

$`4`
                  v.test Mean in category Overall mean sd in category Overall sd      p.value
Age              59.81293        74.515289    43.536084       7.281785  16.522814 0.000000e+00
Work_Experience -13.54373         1.183465     2.629107       2.093045   3.405110 8.628914e-42
Family_Size     -16.31979         2.061155     2.841110       1.032154   1.524628 7.138636e-60
```

- **1η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 1η ομάδα χαρακτηρίζεται από άτομα που είναι παντρεμένοι,
    - έχουν μεσαίο προφίλ δαπανών και
    - είναι καλλιτέχνες.
  - Ποσοτικές μεταβλητές:

- Ανήκουν σε 3μελείς οικογένειες και
- έχουν περίπου 3 χρόνια επαγγελματικής εμπειρίας.

❖ **2η ομάδα**

- Ποιοτικές μεταβλητές:
  - Η 2η ομάδα χαρακτηρίζεται από άτομα που εργάζονται στον τομέα της υγείας,
  - είναι άγαμοι και
  - έχουν χαμηλό προφίλ δαπανών.
- Ποσοτικές μεταβλητές:
  - Έχουν 3,5 χρόνια επαγγελματικής εμπειρίας.

❖ **3η ομάδα**

- Ποιοτικές μεταβλητές:
  - Η 3η ομάδα χαρακτηρίζεται από δικηγόρους,
  - είναι έγγαμοι και
  - δαπανούν πολλά ως καταναλωτές.
- Ποσοτικές μεταβλητές:
  - Κατά κύριο λόγο είναι περίπου 53 ετών.

❖ **4η ομάδα**

- Ποιοτικές μεταβλητές:
  - Η 4η ομάδα χαρακτηρίζεται από απόφοιτους πανεπιστημίου,
  - είναι καλλιτέχνες και
  - είναι περιστασιακοί καταναλωτές.
- Ποσοτικές μεταβλητές:
  - Κατά κύριο λόγο είναι 74 ετών.

3.4 Μέθοδος Modha – Spangler convex K-means

Η μέθοδος συσταδοποίησης Modha-Spangler υπολογίζει το βέλτιστο βάρος χαρακτηριστικών των ποσοτικών μεταβλητών έναντι των ποιοτικών χρησιμοποιώντας μια προσαρμοστική διαδικασία εύρεσης και επιλογής διαστημάτων στον K-means αλγόριθμο (Modha & Spangler, 2003). Για την εφαρμογή αυτής της μεθόδου έγινε χρήση του πακέτου ***kamila*** της R και πιο συγκεκριμένα της εντολής ***gmsClust()***. Ο προσαρμοσμένος δείκτης Rand έδωσε τιμή 0,0850 για την εφαρμογή αυτής της μεθόδου στα δεδομένα που εξετάζονται.

Πίνακας 3-7: Αποτελέσματα ποιοτικών μεταβλητών Modha-Spangler K-means

```
Description of each cluster by the categories
==============================================
$`1`
                            Cla/Mod    Mod/Cla    Global       p.value     v.test
Profession=Lawyer         94.200000 32.9601120  7.501875 1.438323e-301  37.118028
Ever_Married=Yes          31.237323 86.2141358 59.174794 8.923637e-136  24.800186
Spending_Score=High       49.203187 34.5696291 15.063766 1.566972e-102  21.499700
Var_1=Cat_6               26.608579 83.3449965 67.156789  2.658983e-53  15.368589
Graduated=Yes             24.476347 72.7781666 63.750938  4.136873e-16   8.134487
Segmentation=B            28.180662 31.0006998 23.585896  3.262878e-13   7.283078
Segmentation=A            25.742574 29.1112666 24.246062  1.836273e-06   4.770656
Profession=Executive      26.930693  9.5171449  7.576894  2.272875e-03   3.052046
Var_1=Cat_1               13.461538  0.9797061  1.560390  3.902376e-02  -2.063936
Profession=Engineer       17.869416  7.2778167  8.732183  2.582498e-02  -2.228834
Var_1=Cat_5                8.108108  0.4198740  1.110278  2.367733e-03  -3.039752
Profession=Marketing      12.446352  2.0293912  3.495874  3.363181e-04  -3.585589
Var_1=Cat_7               10.240964  1.1896431  2.490623  1.357792e-04  -3.815733
Profession=Homemaker      10.285714  1.2596221  2.625656  9.418394e-05  -3.905106
Profession=Entertainment  16.069221  9.0972708 12.138035  4.531640e-05  -4.078555
Var_1=Cat_2                9.116022  2.3093072  5.431358  1.630807e-10  -6.392607
Var_1=Cat_3               11.987382  5.3184045  9.512378  1.082920e-10  -6.454898
Profession=Doctor         10.979730  4.5486354  8.882221  3.947409e-12  -6.939052
Graduated=No              16.100993 27.2218334 36.249062  4.136873e-16  -8.134487
Var_1=Cat_4               10.836278  6.4380686 12.738185  8.028930e-18  -8.599178
Segmentation=D            10.984633 13.5059482 26.361590  3.309061e-39 -13.099621
Profession=Healthcare      1.299907  0.9797061 16.159040 2.152350e-100 -21.270010
Spending_Score=Average     4.211793  4.8985304 24.936234 1.836963e-109 -22.228113
Ever_Married=No            7.239985 13.7858642 40.825206 8.923637e-136 -24.800186

$`2`
                            Cla/Mod     Mod/Cla    Global       p.value     v.test
Spending_Score=Average    79.542720  67.5869121 24.936234  0.000000e+00        Inf
Ever_Married=Yes          49.594320 100.0000000 59.174794  0.000000e+00        Inf
Segmentation=C            48.953488  43.0470348 25.806452  1.347356e-90  20.184221
Profession=Artist         43.521898  48.7730061 32.888222  9.956773e-69  17.520714
Graduated=Yes             34.690515  75.3578732 63.750938  2.713936e-38  12.938927
Profession=Executive      53.861386  13.9059305  7.576894  4.119506e-33  11.987748
Gender=Male               34.049497  64.0081800 55.168792  5.358208e-21   9.401909
Segmentation=B            37.340967  30.0102249 23.585896  4.530396e-15   7.839307
Spending_Score=High       37.948207  19.4785276 15.063766  1.949082e-10   6.365300
Profession=Entertainment  34.363412  14.2126789 12.138035  9.646179e-04   3.300648
Var_1=Cat_4               34.040047  14.7750511 12.738185  1.487735e-03   3.177065
Var_1=Cat_7               36.746988   3.1186094  2.490623  3.758900e-02   2.079309
Var_1=Cat_3               32.807571  10.6339468  9.512378  4.604496e-02   1.994981
Profession=Doctor         24.831081   7.5153374  8.882221  1.055863e-02  -2.556977
Var_1=Cat_6               27.792672  63.5991820 67.156789  7.310066e-05  -3.965962
Profession=Homemaker      14.285714   1.2781186  2.625656  2.521116e-06  -4.706415
Profession=Marketing      12.446352   1.4826176  3.495874  5.557065e-10  -6.202507
Segmentation=A            22.648515  18.7116564 24.246062  4.844086e-12  -6.910072
Gender=Female             23.560910  35.9918200 44.831208  5.358208e-21  -9.401909
Graduated=No              19.950331  24.6421268 36.249062  2.713936e-38 -12.938927
Profession=Lawyer          2.400000   0.6134969  7.501875  4.447664e-60 -16.348644
Profession=Healthcare      5.292479   2.9141104 16.159040 1.672373e-101 -21.389529
Segmentation=D             9.163347   8.2310838 26.361590 1.569457e-120 -23.344443
Spending_Score=Low         6.326582  12.9345603 60.000000  0.000000e+00       -Inf
Ever_Married=No            0.000000   0.0000000 40.825206  0.000000e+00       -Inf
```

```
$`3`
                            Cla/Mod     Mod/Cla    Global       p.value     v.test
Spending_Score=Low       49.31232808 99.94931576 60.000000  0.000000e+00        Inf
Ever_Married=No          69.31275266 95.59047136 40.825206  0.000000e+00        Inf
Profession=Healthcare    79.48003714 43.38570705 16.159040 5.269371e-309  37.576163
Segmentation=D           59.02105862 52.55955398 26.361590 1.712256e-205  30.589038
Graduated=No             49.37913907 60.46629498 36.249062 1.770031e-153  26.390313
Profession=Marketing     57.51072961  6.79168779  3.495874  2.322592e-19   8.996863
Var_1=Cat_4              41.69611307 17.94221997 12.738185  8.793786e-16   8.042620
Profession=Doctor        44.08783784 13.22858591  8.882221  5.079149e-15   7.824936
Var_1=Cat_2              48.06629834  8.81905727  5.431358  3.141664e-14   7.592345
Var_1=Cat_3              37.06624606 11.91079574  9.512378  2.166633e-05   4.246988
Gender=Female            31.72690763 48.04865687 44.831208  6.249869e-04   3.420532
Var_1=Cat_5              41.89189189  1.57121135  1.110278  2.421512e-02   2.253699
Segmentation=A           27.04207921 22.14901166 24.246062  9.255529e-03  -2.602469
Gender=Male              27.87598586 51.95134313 55.168792  6.249869e-04  -3.420532
Profession=Homemaker     11.42857143  1.01368474  2.625656  7.667844e-09  -5.775602
Var_1=Cat_6              24.17336908 54.84034465 67.156789  9.963920e-43 -13.701361
Segmentation=B           15.83969466 12.62037506 23.585896  3.219048e-46 -14.273108
Profession=Executive      4.35643564  1.11505322  7.576894  2.185766e-50 -14.927415
Segmentation=C           14.53488372 12.67105930 25.806452  2.859759e-62 -16.653382
Profession=Lawyer         0.80000000  0.20273695  7.501875  1.788392e-72 -18.004739
Profession=Artist        12.91058394 14.34363913 32.888222 3.013696e-106 -21.893232
Graduated=Yes            18.35726053 39.53370502 63.750938 1.770031e-153 -26.390313
Spending_Score=High       0.09960159  0.05068424 15.063766 8.617632e-167 -27.525830
Spending_Score=Average    0.00000000  0.00000000 24.936234 7.860305e-302 -37.134292
Ever_Married=Yes          2.20588235  4.40952864 59.174794  0.000000e+00       -Inf

$`4`
                             Cla/Mod    Mod/Cla    Global       p.value     v.test
Profession=Homemaker     64.00000  8.5692425  2.625656 1.168886e-38  13.003488
Spending_Score=Low       22.73068 69.5485845 60.000000 1.753537e-15   7.957639
Graduated=Yes            22.47588 73.0680949 63.750938 1.916252e-15   7.946648
Ever_Married=No          23.44726 48.8140780 40.825206 7.320165e-11   6.513946
Gender=Female            22.95850 52.4866106 44.831208 6.041497e-10   6.189344
Segmentation=A           24.56683 30.3749044 24.246062 1.535990e-08   5.657494
Var_1=Cat_6              21.42538 73.3741393 67.156789 6.386663e-08   5.407623
Profession=Artist        22.49088 37.7199694 32.888222 3.926462e-05   4.111765
Profession=Entertainment 22.24969 13.7719969 12.138035 4.608385e-02   1.994624
Var_1=Cat_5              10.81081  0.6120888  1.110278 4.623886e-02  -1.993206
Profession=Executive     14.85149  5.7383321  7.576894 4.047980e-03  -2.874398
Spending_Score=Average   16.24549 20.6579954 24.936234 5.243435e-05  -4.044502
Var_1=Cat_4              13.42756  8.7222647 12.738185 4.679416e-07  -5.039012
Profession=Healthcare    13.92758 11.4766641 16.159040 1.185677e-07  -5.295690
Gender=Male              16.88877 47.5133894 55.168792 6.041497e-10  -6.189344
Spending_Score=High      12.74900  9.7934200 15.063766 6.000067e-10  -6.190429
Segmentation=C           14.59302 19.2042846 25.806452 4.778039e-10  -6.226226
Ever_Married=Yes         16.96247 51.1859220 59.174794 7.320165e-11  -6.513946
Graduated=No             14.56954 26.9319051 36.249062 1.916252e-15  -7.946648
Profession=Lawyer         2.60000  0.9946442  7.501875 2.310810e-32 -11.844028
```

Πίνακας 3-8: Αποτελέσματα ποσοτικών μεταβλητών Modha-Spangler K-means

```
Description of each cluster by quantitative variables
======================================================
$`1`
                v.test Mean in category Overall mean sd in category Overall sd       p.value
Age             55.72137        65.124563    43.536084     13.2835215  16.522814  0.000000e+00
Work_Experience -20.94678         0.956613     2.629107      1.4351652   3.405110  2.007441e-97
Family_Size     -30.69155         1.743877     2.841110      0.7216821   1.524628 7.378770e-207

$`2`
                  v.test Mean in category Overall mean sd in category  Overall sd       p.value
Family_Size      18.520660     3.377812e+00 2.841110e+00       1.325732    1.524628  1.407003e-76
Age               8.112639     4.608384e+01 4.353608e+01      10.060203   16.522814  4.953194e-16
ID                2.667102     4.636499e+05 4.635198e+05    2488.938970 2566.239202  7.650856e-03
Work_Experience -21.482365     1.238753e+00 2.629107e+00       1.646323    3.405110 2.276182e-102

$`3`
                v.test Mean in category Overall mean sd in category Overall sd       p.value
Family_Size      21.98840         3.474404     2.841110       1.714249   1.524628 3.719010e-107
Work_Experience -19.66882         1.363913     2.629107       1.982016   3.405110  3.989897e-86
Age             -46.20400        29.114546    43.536084       7.694895  16.522814  0.000000e+00

$`4`
                  v.test Mean in category Overall mean sd in category  Overall sd      p.value
Work_Experience  68.903246     8.448355e+00 2.629107e+00       2.158108    3.405110 0.000000e+00
ID               -2.311614     4.633727e+05 4.635198e+05    2638.830743 2566.239202 2.079896e-02
Age             -13.777829     3.788982e+01 4.353608e+01       9.658204   16.522814 3.465377e-43
Family_Size     -14.797162     2.281561e+00 2.841110e+00       1.245769    1.524628 1.527948e-49
```

- **1η ομάδα**
  - Ποιοτικές μεταβλητές:

- Η 1η ομάδα χαρακτηρίζεται από άτομα που είναι δικηγόροι,
- είναι παντρεμένοι και
- έχουν υψηλό σκορ δαπανών.

➢ Ποσοτικές μεταβλητές:

- Είναι κατά μέσο όρο 65 ετών.

❖ **2η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 2η ομάδα χαρακτηρίζεται από άτομα που έχουν ένα μέσο σκορ δαπανών,
- είναι έγγαμοι και
- είναι καλλιτέχνες.

➢ Ποσοτικές μεταβλητές:

- Ανήκουν σε 3μελείς οικογένειες.

❖ **3η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 3η ομάδα χαρακτηρίζεται από χαμηλό σκορ δαπανών,
- είναι άγαμοι και
- εργάζονται στον τομέα της υγείας.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο ανήκουν σε 3μελείς οικογένειες.

❖ **4η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 4η ομάδα χαρακτηρίζεται από άτομα που ασχολούνται με τα οικιακά,
- έχουν χαμηλό σκορ δαπανών,
- είναι απόφοιτοι πανεπιστημίου και
- είναι άγαμες γυναίκες.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο έχουν εμπειρία 8,5 ετών.

3.5 Μέθοδος FAMD + K-means (Two-step)

Η μέθοδος Ανάλυσης Παραγόντων για Μεικτά Δεδομένα (Factor Analysis of Mixed Data -FAMD) είναι μια μέθοδος ανάλυσης κύριων συνιστώσων για κατηγορικές και συνεχείς μεταβλητές (Pagès, 2004). Η συγκεκριμένη μέθοδος μπορεί να θεωρηθεί ως ένα κράμα της Ανάλυσης σε Κύριες Συνιστώσες (PCA) και της Παραγοντικής Ανάλυσης

των Πολλαπλών Αντιστοιχιών (MCA) (Markos, Moschidis, & Chadjipadelis, Sequential dimension reduction and clustering of mixed-type data, 2020). Για την εφαρμογή αυτής της μεθόδου έγινε χρήση του πακέτου ***FactoMineR*** της **R** και συγκεκριμένα των εντολών ***estim_ncp()*** για τον υπολογισμό του πλήθους των συνιστώσων, της ***FAMD()*** και της ***kmeans()*** του πακέτου ***stats***. Ο προσαρμοσμένος δείκτης Rand δίνει την τιμή 0,1256.

Πίνακας 3-9: Αποτελέσματα ποιοτικών μεταβλητών FAMD & K-means

Description of each cluster by the categories
==========

$`1`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Spending_Score=Average	77.075812	55.002147	24.936234	0.000000e+00	Inf
Ever_Married=Yes	57.581136	97.509661	59.174794	0.000000e+00	Inf
Profession=Artist	59.306569	55.817948	32.888222	2.935246e-184	28.948450
Graduated=Yes	44.128030	80.506655	63.750938	7.815853e-102	21.424986
Segmentation=C	54.186047	40.017175	25.806452	1.680872e-81	19.121233
Segmentation=B	47.964377	32.374410	23.585896	2.800877e-34	12.208476
Var_1=Cat_6	38.047364	73.121511	67.156789	1.832122e-14	7.661893
Gender=Male	38.645635	61.013310	55.168792	1.807982e-12	7.048545
Profession=Executive	44.950495	9.746672	7.576894	1.403293e-06	4.824536
Profession=Entertainment	40.914710	14.212108	12.138035	1.684062e-04	3.762228
Profession=Engineer	30.584192	7.642765	8.732183	2.009689e-02	-2.324534
Profession=Homemaker	25.714286	1.932160	2.625656	8.289508e-03	-2.640045
Var_1=Cat_2	27.348066	4.250751	5.431358	1.555859e-03	-3.164058
Profession=Doctor	26.013514	6.612280	8.882221	1.124100e-06	-4.868568
Segmentation=A	29.393564	20.395019	24.246062	5.831522e-08	-5.423892
Var_1=Cat_4	24.499411	8.930872	12.738185	2.645291e-12	-6.995386
Gender=Female	30.388220	38.986690	44.831208	1.807982e-12	-7.048545
Profession=Marketing	11.158798	1.116359	3.495874	3.477588e-17	-8.429293
Profession=Lawyer	10.200000	2.189781	7.501875	1.042226e-39	-13.187021
Graduated=No	18.791391	19.493345	36.249062	7.815853e-102	-21.424986
Segmentation=D	9.561753	7.213396	26.361590	2.862747e-172	-27.979789
Profession=Healthcare	1.578459	0.729927	16.159040	5.976526e-191	-29.475258
Spending_Score=Low	17.304326	29.712323	60.000000	5.092385e-303	-37.207862
Ever_Married=No	2.131569	2.490339	40.825206	0.000000e+00	-Inf

$`2`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Segmentation=D	58.33807627	66.34304207	26.361590	0.000000e+00	Inf
Spending_Score=Low	38.60965241	99.93527508	60.000000	0.000000e+00	Inf
Profession=Healthcare	86.07242340	60.00000000	16.159040	0.000000e+00	Inf
Ever_Married=No	56.59683940	99.67637540	40.825206	0.000000e+00	Inf
Graduated=No	48.05463576	75.14563107	36.249062	4.642996e-283	35.952398
Profession=Marketing	52.78969957	7.96116505	3.495874	2.178273e-23	9.964564
Var_1=Cat_2	45.30386740	10.61488673	5.431358	1.487982e-21	9.535755
Var_1=Cat_4	35.68904594	19.61165049	12.738185	8.068065e-19	8.859076
Profession=Doctor	32.26351351	12.36245955	8.882221	1.162604e-07	5.299280
Var_1=Cat_3	31.86119874	13.07443366	9.512378	1.379636e-07	5.267939
Gender=Female	25.30120482	48.93203883	44.831208	2.232432e-04	3.691149
Var_1=Cat_5	40.54054054	1.94174757	1.110278	8.635174e-04	3.331589
Var_1=Cat_7	30.12048193	3.23624595	2.490623	3.673008e-02	2.088752
Profession=Engineer	19.93127148	7.50809061	8.732183	4.944118e-02	-1.964767
Gender=Male	21.45771009	51.06796117	55.168792	2.232432e-04	-3.691149
Profession=Entertainment	11.74289246	6.14886731	12.138035	1.941368e-18	-8.760645
Segmentation=A	11.88118812	12.42718447	24.246062	8.670212e-39	-13.026308
Segmentation=C	11.22093023	12.49190939	25.806452	4.402577e-47	-14.411159
Profession=Executive	0.59405941	0.19417476	7.576894	2.641160e-55	-15.664593
Profession=Lawyer	0.20000000	0.06472492	7.501875	2.287705e-58	-16.106740
Var_1=Cat_6	17.09115282	49.51456311	67.156789	4.020378e-61	-16.494475
Segmentation=B	8.58778626	8.73786408	23.585896	2.473122e-64	-16.935248
Spending_Score=High	0.09960159	0.06472492	15.063766	2.355277e-124	-23.717930
Spending_Score=Average	0.00000000	0.00000000	24.936234	7.087706e-225	-32.013446
Profession=Artist	1.82481752	2.58899676	32.888222	1.066740e-241	-33.200275
Graduated=Yes	9.03742057	24.85436893	63.750938	4.642996e-283	-35.952398
Ever_Married=Yes	0.12677485	0.32362460	59.174794	0.000000e+00	-Inf

```
$`3`
                          Cla/Mod    Mod/Cla    Global       p.value     v.test
Spending_Score=High     59.1633466 65.4185022 15.063766  0.000000e+00        Inf
Profession=Lawyer       88.8000000 48.8986784  7.501875  0.000000e+00        Inf
Ever_Married=Yes        22.9462475 99.6696035 59.174794 8.081633e-221  31.720592
Var_1=Cat_6             18.9231457 93.2819383 67.156789  3.820234e-90  20.132648
Profession=Executive    45.3465347 25.2202643  7.576894  7.211585e-75  18.307492
Segmentation=C          22.7325581 43.0616740 25.806452  2.368554e-34  12.222110
Graduated=Yes           16.4038597 76.7621145 63.750938  1.756802e-19   9.027474
Gender=Male             15.6921403 63.5462555 55.168792  3.891589e-08   5.495702
Segmentation=B          15.7124682 27.2026432 23.585896  6.391454e-03   2.726992
Var_1=Cat_1              6.7307692  0.7709251  1.560390  2.889961e-02  -2.184853
Segmentation=A          11.1386139 19.8237885 24.246062  6.703792e-04  -3.401415
Var_1=Cat_5              1.3513514  0.1101322  1.110278  2.546104e-04  -3.657577
Gender=Female           11.0776439 36.4537445 44.831208  3.891589e-08  -5.495702
Var_1=Cat_7              1.2048193  0.2202643  2.490623  8.280875e-09  -5.762639
Profession=Homemaker     1.1428571  0.2202643  2.625656  2.363988e-09  -5.970589
Profession=Marketing     1.2875536  0.3303965  3.495874  8.439237e-12  -6.830885
Var_1=Cat_2              2.4861878  0.9911894  5.431358  7.021927e-14  -7.487460
Profession=Artist        8.9872263 21.6960352 32.888222  1.756891e-15  -7.957402
Spending_Score=Average   7.8219013 14.3171806 24.936234  6.325482e-17  -8.358995
Graduated=No             8.7334437 23.2378855 36.249062  1.756802e-19  -9.027474
Var_1=Cat_3              2.8391167  1.9823789  9.512378  3.291593e-22  -9.691040
Profession=Entertainment 2.9666255  2.6431718 12.138035  1.061809e-27 -10.907461
Var_1=Cat_4              2.8268551  2.6431718 12.738185  4.727379e-30 -11.389306
Profession=Doctor        0.8445946  0.5506608  8.882221  2.889811e-32 -11.825267
Profession=Engineer      0.6872852  0.4405286  8.732183  6.449951e-33 -11.950547
Segmentation=D           5.1223677  9.9118943 26.361590  1.828814e-39 -13.144554
Profession=Healthcare    0.0000000  0.0000000 16.159040  5.578631e-76 -18.446352
Spending_Score=Low       4.6011503 20.2643172 60.000000 4.255367e-153 -26.357101
Ever_Married=No          0.1102536  0.3303965 40.825206 8.081633e-221 -31.720592

$`4`
                          Cla/Mod   Mod/Cla    Global       p.value     v.test
Spending_Score=Low      39.484871 83.855550 60.000000 2.689807e-149  26.023430
Ever_Married=No         41.161338 59.479554 40.825206  2.372713e-83  19.342171
Segmentation=A          47.586634 40.839087 24.246062  1.880966e-82  19.235123
Profession=Engineer     48.797251 15.082315  8.732183  3.876009e-28  10.998719
Profession=Entertainment 44.375773 19.065321 12.138035  1.201909e-25  10.468773
Gender=Female           33.232932 52.734997 44.831208  4.471876e-16   8.125047
Profession=Doctor       40.878378 12.851832  8.882221  4.574477e-12   6.918191
Var_1=Cat_4             36.984688 16.675518 12.738185  3.337481e-09   5.914074
Graduated=Yes           30.430690 68.667021 63.750938  1.354642e-07   5.271295
Profession=Homemaker    45.142857  4.195433  2.625656  1.518810e-06   4.808745
Var_1=Cat_3             32.807571 11.046203  9.512378  8.167925e-03   2.645048
Var_1=Cat_7             36.144578  3.186405  2.490623  2.533649e-02   2.236233
Profession=Marketing    34.763948  4.301646  3.495874  2.735074e-02   2.206475
Profession=Artist       29.881387 34.784918 32.888222  3.912138e-02   2.062908
Graduated=No            24.420530 31.332979 36.249062  1.354642e-07  -5.271295
Var_1=Cat_6             25.938338 61.656930 67.156789  2.622980e-09  -5.953606
Gender=Male             24.204515 47.265003 55.168792  4.471876e-16  -8.125047
Profession=Executive     9.108911  2.442910  7.576894  7.356877e-28 -10.940774
Profession=Healthcare   12.349118  7.063197 16.159040  1.214309e-41 -13.518621
Spending_Score=Average  15.102286 13.329793 24.936234  1.063189e-46 -14.350142
Profession=Lawyer        0.800000  0.212427  7.501875  3.127187e-68 -17.455483
Segmentation=C          11.860465 10.833776 25.806452  3.766833e-77 -18.591474
Ever_Married=Yes        19.345842 40.520446 59.174794  2.372713e-83 -19.342171
Spending_Score=High      5.278884  2.814657 15.063766  9.044259e-88 -19.859927
```

Πίνακας 3-10: Αποτελέσματα ποσοτικών μεταβλητών FAMD & K-means

```
Description of each cluster by quantitative variables
======================================================
$`1`
                   v.test Mean in category Overall mean sd in category Overall sd       p.value
Family_Size      29.574080         3.846602     2.841110       1.635893   1.524628 3.220032e-192
Work_Experience   2.173838         2.794175     2.629107       3.553980   3.405110  2.971730e-02
Age             -46.467239        26.414887    43.536084       5.928105  16.522814  0.000000e+00

$`2`
                   v.test Mean in category Overall mean sd in category Overall sd      p.value
Age              52.34060        70.211454    43.536084     11.6011830  16.522814 0.000000e+00
Family_Size     -12.10109         2.272026     2.841110      0.9628618   1.524628 1.042149e-33
Work_Experience -13.04655         1.258811     2.629107      2.1645368   3.405110 6.648733e-39

$`3`
                   v.test Mean in category Overall mean sd in category  Overall sd      p.value
Work_Experience  13.241324     3.509294e+00 2.629107e+00       3.827992    3.405110 5.064778e-40
ID               -2.455065     4.633969e+05 4.635198e+05    2508.528482 2566.239202 1.408589e-02
Family_Size     -14.967095     2.395645e+00 2.841110e+00       1.508384    1.524628 1.204703e-50
Age             -16.560919     3.819437e+01 4.353608e+01       8.763817   16.522814 1.335391e-61

$`4`
                   v.test Mean in category Overall mean sd in category  Overall sd      p.value
Age             19.106754     4.881280e+01 4.353608e+01      10.981509   16.522814 2.218471e-81
ID               2.190217     4.636138e+05 4.635198e+05    2595.253472 2566.239202 2.850852e-02
Family_Size     -3.335189     2.756119e+00 2.841110e+00       1.308143    1.524628 8.524131e-04
Work_Experience -5.040894     2.342207e+00 2.629107e+00       3.109582    3.405110 4.633629e-07
```

- **1η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 1η ομάδα χαρακτηρίζεται από άτομα με μεσαίο σκορ δαπανών,
    - είναι παντρεμένοι,
    - είναι καλλιτέχνες και
    - είναι απόφοιτοι πανεπιστημίου.
  - Ποσοτικές μεταβλητές:
    - Ανήκουν σε οικογένειες με 4 μέλη και
    - έχουν περίπου 2,8 έτη εργασιακής εμπειρίας.
- **2η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 2η ομάδα χαρακτηρίζεται από άτομα που με χαμηλό σκορ δαπανών,
    - εργάζονται στον χώρο της υγείας,
    - είναι άγαμοι
    - και δεν είναι απόφοιτοι πανεπιστημίου.
  - Ποσοτικές μεταβλητές:
    - Είναι κατά μέσο όρο 70 ετών.
- **3η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 3η ομάδα χαρακτηρίζεται από υψηλό σκορ δαπανών,
    - είναι δικηγόροι και
    - είναι έγγαμοι.
  - Ποσοτικές μεταβλητές:
    - Κατά κύριο λόγο έχουν 3,5 έτη εργασιακής εμπειρίας.
- **4η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 4η ομάδα χαρακτηρίζεται από άτομα με χαμηλό σκορ δαπανών,
    - είναι άγαμοι,
    - είναι μηχανικοί αλλά υπάρχουν και αρκετοί που εργάζονται στον χώρο της διασκέδασης και
    - είναι κατά κύριο λόγο γυναίκες.

3.6 Μέθοδος Mixed Reduced K-means

Η Mixed Reduced K-means είναι μέθοδος που συνδυάζει την μείωση διαστάσεων και τη συσταδοποίηση ποιοτικών, ποσοτικών και μεικτών δεδομένων. Αρχικά ο αλγόριθμος μετατρέπει όλες τις μεταβλητές μεικτού τύπου στην ίδια κλίμακα, από ποιοτικές σε ποσοτικές και αντίστροφα. Έπειτα εξετάζει διάφορες μετρικές απόστασης για μεικτού τύπου μεταβλητές καθώς και αντίστοιχες μεθόδους συσταδοποίησης. Τέλος λαμβάνονται υπόψη αρκετές προσεγγίσεις που συνδυάζουν την μείωση διαστάσεων και την ομαδοποίηση τόσο σειριακά όσο και συνεκτικά (Van de Velden, Iodice D'Enza, & Markos, 2019). Για την εφαρμογή του αλγορίθμου αυτού έγινε χρήση του πακέτου ***clustrd*** της **R** (Markos, Iodice D'Enza, & Van de Velden, 2019) και συγκεκριμένα της εντολής ***cluspcamix()***. Ο προσαρμοσμένος δείκτης Rand δίνει τιμή 0,1261 για την συνεκτικότητα των παρατηρήσεων στην κάθε ομάδα.

Πίνακας 3-11: Αποτελέσματα ποιοτικών μεταβλητών Mixed Reduced K-means

Description of each cluster by the categories

$`1`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Spending_Score=Average	77.978339	55.8620690	24.936234	0.000000e+00	Inf
Ever_Married=Yes	57.682556	98.0603448	59.174794	0.000000e+00	Inf
Profession=Artist	58.667883	55.4310345	32.888222	5.479219e-177	28.364923
Graduated=Yes	43.563191	79.7844828	63.750938	1.482799e-92	20.405876
Segmentation=C	53.372093	39.5689655	25.806452	4.139048e-76	18.462479
Segmentation=B	47.201018	31.9827586	23.585896	2.802002e-31	11.632963
Gender=Male	38.509655	61.0344828	55.168792	1.747742e-12	7.053260
Var_1=Cat_6	37.354781	72.0689655	67.156789	3.387065e-10	6.279939
Profession=Entertainment	43.263288	15.0862069	12.138035	1.084743e-07	5.311923
Profession=Executive	40.396040	8.7931034	7.576894	6.634343e-03	2.714662
Profession=Homemaker	25.714286	1.9396552	2.625656	9.238833e-03	-2.603088
Var_1=Cat_2	28.176796	4.3965517	5.431358	5.825891e-03	-2.757423
Segmentation=A	30.074257	20.9482759	24.246062	3.768525e-06	-4.623756
Profession=Doctor	26.013514	6.6379310	8.882221	1.599992e-06	-4.798323
Var_1=Cat_4	26.030624	9.5258621	12.738185	4.696841e-09	-5.857572
Gender=Female	30.254351	38.9655172	44.831208	1.747742e-12	-7.053260
Profession=Marketing	9.871245	0.9913793	3.495874	4.605364e-19	-8.921375
Profession=Lawyer	11.400000	2.4568966	7.501875	2.825296e-35	-12.393746
Graduated=No	19.412252	20.2155172	36.249062	1.482799e-92	-20.405876
Segmentation=D	9.903244	7.5000000	26.361590	1.251896e-165	-27.428568
Profession=Healthcare	1.764160	0.8189655	16.159040	1.126322e-186	-29.139760
Spending_Score=Low	17.304326	29.8275862	60.000000	8.615578e-299	-36.945457
Ever_Married=No	1.653804	1.9396552	40.825206	0.000000e+00	-Inf

$`2`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Spending_Score=Low	38.709677	85.0549451	60.000000	1.827737e-158	26.821223
Ever_Married=No	40.352811	60.3296703	40.825206	8.229318e-87	19.748710
Segmentation=A	46.349010	41.1538462	24.246062	2.140245e-81	19.108627
Profession=Engineer	47.594502	15.2197802	8.732183	7.112304e-28	10.943839
Profession=Entertainment	42.027194	18.6813187	12.138035	4.520616e-22	9.658586
Gender=Female	32.329317	53.0769231	44.831208	1.249291e-16	8.278316
Profession=Doctor	38.851351	12.6373626	8.882221	1.618882e-10	6.393729
Graduated=Yes	29.771711	69.5054945	63.750938	1.578591e-09	6.036125
Var_1=Cat_4	35.100118	16.3736264	12.738185	9.404215e-08	5.337875
Profession=Homemaker	43.428571	4.1758242	2.625656	3.621579e-06	4.631996
Profession=Artist	29.972628	36.0989011	32.888222	6.659191e-04	3.403239
Var_1=Cat_3	32.176656	11.2087912	9.512378	4.343705e-03	2.852059
Var_1=Cat_7	34.939759	3.1868132	2.490623	2.893788e-02	2.184332
Segmentation=D	25.441093	24.5604396	26.361590	4.017304e-02	-2.051965
Var_1=Cat_6	25.245755	62.0879121	67.156789	8.280822e-08	-5.360900
Graduated=No	22.971854	30.4945055	36.249062	1.578591e-09	-6.036125
Gender=Male	23.225456	46.9230769	55.168792	1.249291e-16	-8.278316
Profession=Executive	7.326733	2.0329670	7.576894	1.021623e-31	-11.718755
Profession=Healthcare	11.234912	6.6483516	16.159040	7.399973e-44	-13.888864
Spending_Score=Average	13.357401	12.1978022	24.936234	3.789766e-54	-15.494315
Profession=Lawyer	1.200000	0.3296703	7.501875	3.751886e-62	-16.637128
Segmentation=C	11.162791	10.5494505	25.806452	6.984498e-77	-18.558328
Spending_Score=High	4.980080	2.7472527	15.063766	2.633687e-85	-19.572880
Ever_Married=Yes	18.306288	39.6703297	59.174794	8.229318e-87	-19.748710

$`3`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Segmentation=D	59.53329539	66.16065781	26.361590	0.000000e+00	Inf
Spending_Score=Low	39.53488372	100.00000000	60.000000	0.000000e+00	Inf
Profession=Healthcare	87.00092851	59.26628716	16.159040	0.000000e+00	Inf
Ever_Married=No	57.99338479	99.81024668	40.825206	0.000000e+00	Inf
Graduated=No	48.34437086	73.87729285	36.249062	9.028921e-273	35.287821
Profession=Marketing	56.65236052	8.34914611	3.495874	5.030951e-28	10.975173
Var_1=Cat_2	45.85635359	10.49968374	5.431358	2.328948e-21	9.489164
Var_1=Cat_4	35.68904594	19.16508539	12.738185	3.356842e-17	8.433428
Profession=Doctor	34.45945946	12.90322581	8.882221	5.990350e-10	6.190685
Var_1=Cat_3	32.17665615	12.90322581	9.512378	3.390747e-07	5.100323
Gender=Female	26.47255689	50.03162555	44.831208	2.027913e-06	4.750623
Var_1=Cat_5	40.54054054	1.89753321	1.110278	1.307397e-03	3.214351
Var_1=Cat_7	30.72289157	3.22580645	2.490623	3.642950e-02	2.092101
Profession=Engineer	20.44673540	7.52688172	8.732183	4.961029e-02	-1.963309
Gender=Male	21.48490617	49.96837445	55.168792	2.027913e-06	-4.750623
Profession=Entertainment	11.24845488	5.75585073	12.138035	1.625958e-21	-9.526551
Segmentation=A	12.25247525	12.52371917	24.246062	3.237546e-39	-13.101279
Segmentation=C	11.56976744	12.58697027	25.806452	1.057443e-47	-14.509333
Var_1=Cat_6	17.80607685	50.41113219	67.156789	5.493687e-57	-15.908939
Profession=Executive	0.39603960	0.12650221	7.576894	1.162410e-58	-16.148562
Profession=Lawyer	0.20000000	0.06325111	7.501875	5.757378e-60	-16.332908
Segmentation=B	8.77862595	8.72865275	23.585896	3.064380e-66	-17.191696
Spending_Score=High	0.00000000	0.00000000	15.063766	2.333205e-130	-24.292845
Spending_Score=Average	0.00000000	0.00000000	24.936234	4.882639e-231	-32.453195
Profession=Artist	1.96167883	2.71979760	32.888222	1.442749e-245	-33.467284
Graduated=Yes	9.71993410	26.12270715	63.750938	9.028921e-273	-35.287821
Ever_Married=Yes	0.07606491	0.18975332	59.174794	0.000000e+00	-Inf

$`4`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Spending_Score=High	61.9521912	65.8898305	15.063766	0.000000e+00	Inf
Profession=Lawyer	87.2000000	46.1864407	7.501875	0.000000e+00	Inf
Ever_Married=Yes	23.9350913	100.0000000	59.174794	8.065370e-239	33.000213
Profession=Executive	51.8811881	27.7542373	7.576894	6.694295e-100	21.216713
Var_1=Cat_6	19.5933870	92.9025424	67.156789	7.106620e-91	20.215812
Segmentation=C	23.8953488	43.5381356	25.806452	9.600367e-38	12.841494
Graduated=Yes	16.9451636	76.2711864	63.750938	7.260987e-19	8.870819
Gender=Male	16.7799837	65.3601695	55.168792	7.231350e-12	6.853007
Segmentation=B	16.5394402	27.5423729	23.585896	2.308865e-03	3.047327
Var_1=Cat_1	6.7307692	0.7415254	1.560390	1.979150e-02	-2.330277
Var_1=Cat_5	1.3513514	0.1059322	1.110278	1.658988e-04	-3.765977
Segmentation=A	11.3242574	19.3855932	24.246062	1.270306e-04	-3.832145
Var_1=Cat_7	1.2048193	0.2118644	2.490623	3.132807e-09	-5.924483
Profession=Homemaker	1.1428571	0.2118644	2.625656	8.440687e-10	-6.136411
Gender=Female	10.9437751	34.6398305	44.831208	7.231350e-12	-6.853007
Var_1=Cat_2	3.0386740	1.1652542	5.431358	3.668658e-13	-7.267255
Profession=Marketing	0.8583691	0.2118644	3.495874	1.556530e-13	-7.382236
Spending_Score=Average	8.6642599	15.2542373	24.936234	9.238812e-15	-7.749317
Profession=Artist	9.3978102	21.8220339	32.888222	8.807659e-16	-8.042427
Graduated=No	9.2715232	23.7288136	36.249062	7.260987e-19	-8.870819
Var_1=Cat_3	2.9968454	2.0127119	9.512378	6.608151e-23	-9.853677
Profession=Entertainment	3.4610630	2.9661017	12.138035	1.314035e-26	-10.676301
Var_1=Cat_4	3.1802120	2.8601695	12.738185	9.859637e-30	-11.325073
Profession=Engineer	0.6872852	0.4237288	8.732183	1.644968e-34	-12.251707
Profession=Doctor	0.6756757	0.4237288	8.882221	3.278078e-35	-12.381824
Segmentation=D	5.1223677	9.5338983	26.361590	3.040931e-43	-13.787260
Profession=Healthcare	0.0000000	0.0000000	16.159040	3.142872e-79	-18.846441
Spending_Score=Low	4.4511128	18.8559322	60.000000	2.733678e-172	-27.981436
Ever_Married=No	0.0000000	0.0000000	40.825206	8.065370e-239	-33.000213

Πίνακας 3-12: Αποτελέσματα ποσοτικών μεταβλητών Mixed Reduced K-means

```
Description of each cluster by quantitative variables
======================================================
$`1`
                   v.test Mean in category Overall mean sd in category  Overall sd      p.value
Age             18.255796     4.859112e+01 4.353608e+01      11.100671   16.522814 1.860784e-74
ID               2.388240     4.636226e+05 4.635198e+05    2599.799895 2566.239202 1.692928e-02
Family_Size     -2.735378     2.771219e+00 2.841110e+00       1.313359    1.524628 6.230876e-03
Work_Experience -4.304762     2.383455e+00 2.629107e+00       3.137295    3.405110 1.671655e-05

$`2`
                   v.test Mean in category Overall mean sd in category  Overall sd      p.value
Work_Experience  12.947542     3.511282e+00 2.629107e+00       3.834641    3.405110 2.425982e-38
ID               -2.906611     4.633706e+05 4.635198e+05    2498.087755 2566.239202 3.653673e-03
Age             -15.324938     3.846946e+01 4.353608e+01       8.956157   16.522814 5.210416e-53
Family_Size     -15.520058     2.367639e+00 2.841110e+00       1.500031    1.524628 2.538212e-54

$`3`
               v.test Mean in category Overall mean sd in category Overall sd      p.value
Family_Size  28.20667         3.784227      2.84111        1.65211   1.524628 4.84386e-175
Age         -46.81908        26.570978     43.53608        5.96278  16.522814  0.00000e+00

$`4`
                   v.test Mean in category Overall mean sd in category Overall sd      p.value
Age              51.84144        69.399151    43.536084      12.156967  16.522814 0.000000e+00
Family_Size     -10.89201         2.339703     2.841110       1.035103   1.524628 1.258354e-27
Work_Experience -12.93356         1.299363     2.629107       2.215459   3.405110 2.910247e-38
```

- **1η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 1η ομάδα χαρακτηρίζεται από άτομα με μεσαίο σκορ δαπανών,
    - είναι παντρεμένοι,
    - είναι καλλιτέχνες και
    - είναι απόφοιτοι πανεπιστημίου.
- **2η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 2η ομάδα χαρακτηρίζεται από άτομα που με χαμηλό σκορ δαπανών,
    - είναι άγαμοι
    - εργάζονται τόσο ως μηχανικοί όσο στον χώρο της διασκέδασης,
    - και είναι γυναίκες.
  - Ποσοτικές μεταβλητές:
    - Έχουν κατά μέσο όρο 3,5 έτη επαγγελματικής εμπειρίας.
- **3η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 3η ομάδα χαρακτηρίζεται από χαμηλό σκορ δαπανών,
    - εργάζονται στον τομέα της υγείας,
    - είναι άγαμοι,
  - Ποσοτικές μεταβλητές:
    - Κατά κύριο λόγο ανήκουν σε 4μελείς οικογένειες.
- **4η ομάδα**

- Ποιοτικές μεταβλητές:
  - Η 4η ομάδα χαρακτηρίζεται από άτομα με υψηλό σκορ δαπανών,
  - είναι δικηγόροι και
  - είναι έγγαμοι.
- Ποσοτικές μεταβλητές:
  - Κατά κύριο λόγο είναι 69 ετών.

3.7 Αλγόριθμος KAMILA (KAy-means for MIxed LArge datasets)

Ο αλγόριθμος KAMILA είναι μια ημιπαραμετρική μέθοδος για τη συσταδοποίηση μεικτού τύπου δεδομένων εμπνευσμένη από την μέθοδο k-means. Αρχικά οι μεταβλητές χρησιμοποιούνται στην αρχική τους κλίμακα ώστε να μην χάνονται πληροφορίες κατά την μετατροπή τους. Έπειτα διασφαλίζεται ισότιμη επιρροή τόσο στις ποιοτικές όσο στις ποσοτικές μεταβλητές. Δεν περιορίζει με αυστηρές παραμετρικές υποθέσεις, γενικεύοντας τον σχηματισμό των ομάδων σε ένα εύρος από ελλιπείς διαμοιρασμούς. Έτσι δεν χρειάζεται από τον χρήστη να ορίσει βάρη στις μεταβλητές, η να χρησιμοποιήσει περίπλοκους κώδικες (Foss, Markatou, Ray, & Heching, 2016). Για την εφαρμογή του αλγορίθμου αυτού έγινε χρήση του πακέτου ***kamila*** της **R** και συγκεκριμένα της εντολής ***kamila()***. Ο προσαρμοσμένος δείκτης Rand, έδωσε την τιμή 0,1170 για την ομοιογένεια των ομάδων.

Πίνακας 3-13: Αποτελέσματα ποιοτικών μεταβλητών KAMILA

```
Description of each cluster by the categories
==============================================
$`1`
                          Cla/Mod    Mod/Cla    Global      p.value     v.test
Profession=Artist        45.392336 46.9561114 32.888222 4.837508e-61  16.483295
Graduated=Yes            38.550247 77.3006135 63.750938 5.757260e-58  16.049558
Spending_Score=Average   43.682310 34.2614441 24.936234 3.404207e-32  11.811503
Segmentation=A           42.141089 32.1378008 24.246062 5.320356e-24  10.103687
Ever_Married=Yes         35.192698 65.5025956 59.174794 5.515443e-13   7.211954
Segmentation=B           39.058524 28.9759320 23.585896 2.717129e-12   6.991629
Profession=Entertainment 39.184178 14.9598867 12.138035 2.112588e-06   4.742345
Profession=Engineer      39.518900 10.8541765  8.732183 3.794555e-05   4.119646
Gender=Female            34.036145 47.9943370 44.831208 3.987672e-04   3.540898
Var_1=Cat_3              36.908517 11.0429448  9.512378 4.033011e-03   2.875568
Segmentation=C           29.767442 24.1623407 25.806452 3.576213e-02  -2.099622
Var_1=Cat_6              30.853441 65.1722511 67.156789 1.880448e-02  -2.349384
Gender=Male              29.970084 52.0056630 55.168792 3.987672e-04  -3.540898
Spending_Score=Low       29.082271 54.8843794 60.000000 6.588089e-09  -5.801101
Spending_Score=High      22.908367 10.8541765 15.063766 1.876094e-11  -6.715355
Ever_Married=No          26.865123 34.4974044 40.825206 5.515443e-13  -7.211954
Segmentation=D           17.757541 14.7239264 26.361590 1.508546e-52 -15.255706
Graduated=No             19.908940 22.6993865 36.249062 5.757260e-58 -16.049558
Profession=Lawyer         1.000000  0.2359604  7.501875 7.762583e-78 -18.675998
Profession=Healthcare     9.377902  4.7663992 16.159040 5.020470e-79 -18.821642

$`2`
                          Cla/Mod    Mod/Cla    Global       p.value     v.test
Segmentation=D           69.379624 58.7469880 26.361590  0.000000e+00        Inf
Spending_Score=Low       47.761940 92.0481928 60.000000  0.000000e+00        Inf
Profession=Healthcare    88.486537 45.9277108 16.159040  0.000000e+00        Inf
Ever_Married=No          65.159868 85.4457831 40.825206  0.000000e+00        Inf
Graduated=No             54.470199 63.4216867 36.249062 3.731449e-208  30.788534
Profession=Doctor        47.466216 13.5421687  8.882221  3.128405e-18   8.706703
Profession=Marketing     53.648069  6.0240964  3.495874  4.558853e-13   7.237840
Var_1=Cat_4              40.518257 16.5783133 12.738185  5.694741e-10   6.198656
Var_1=Cat_2              45.856354  8.0000000  5.431358  1.656147e-09   6.028377
Profession=Homemaker     43.428571  3.6626506  2.625656  5.412369e-04   3.459471
Var_1=Cat_5              50.000000  1.7831325  1.110278  7.247650e-04   3.380035
Var_1=Cat_3              37.066246 11.3253012  9.512378  8.196858e-04   3.346060
Profession=Entertainment 25.957973 10.1204819 12.138035  5.954538e-04  -3.433677
Segmentation=A           25.371287 19.7590361 24.246062  5.955289e-09  -5.818008
Var_1=Cat_6              26.563896 57.3012048 67.156789  4.311815e-30 -11.397322
Profession=Executive      9.504950  2.3132530  7.576894  3.668554e-33 -11.997350
Segmentation=B           13.422392 10.1686747 23.585896  3.963380e-75 -18.340063
Profession=Lawyer         0.800000  0.1927711  7.501875  2.150260e-77 -18.621519
Segmentation=C           13.662791 11.3253012 25.806452  1.162467e-81 -19.140457
Spending_Score=High       5.776892  2.7951807 15.063766  1.468541e-99 -21.179736
Spending_Score=Average    6.438026  5.1566265 24.936234 1.983942e-168 -27.662323
Profession=Artist         9.534672 10.0722892 32.888222 8.018247e-178 -28.432512
Graduated=Yes            17.863027 36.5783133 63.750938 3.731449e-208 -30.788534
Ever_Married=Yes          7.657201 14.5542169 59.174794  0.000000e+00        -Inf
```

```
$`3`
                          Cla/Mod    Mod/Cla    Global       p.value     v.test
Profession=Lawyer        91.4000000 53.1395349  7.501875  0.000000e+00        Inf
Ever_Married=Yes         20.8671400 95.6976744 59.174794 1.373759e-152  26.312662
Spending_Score=High      41.2350598 48.1395349 15.063766 1.259080e-141  25.336713
Var_1=Cat_6              16.5326184 86.0465116 67.156789  2.607512e-41  13.462278
Profession=Executive     21.3861386 12.5581395  7.576894  3.347525e-08   5.522215
Segmentation=B           15.3307888 28.0232558 23.585896  1.230251e-03   3.231772
Segmentation=C           15.0581395 30.1162791 25.806452  2.262445e-03   3.053426
Segmentation=A           14.4183168 27.0930233 24.246062  3.858885e-02   2.068545
Profession=Marketing      6.4377682  1.7441860  3.495874  1.290169e-03  -3.218158
Profession=Homemaker      3.4285714  0.6976744  2.625656  1.601567e-05  -4.314235
Var_1=Cat_7               2.4096386  0.4651163  2.490623  1.797839e-06  -4.774914
Var_1=Cat_2               4.6961326  1.9767442  5.431358  9.533829e-08  -5.335392
Profession=Entertainment  6.9221261  6.5116279 12.138035  7.168536e-09  -5.786928
Var_1=Cat_4               6.8315665  6.7441860 12.738185  1.494738e-09  -6.044931
Profession=Engineer       5.3264605  3.6046512  8.732183  3.429736e-10  -6.277992
Var_1=Cat_3               4.2586751  3.1395349  9.512378  1.795359e-14  -7.664495
Profession=Doctor         3.7162162  2.5581395  8.882221  1.993837e-15  -7.941728
Segmentation=D            7.2282299 14.7674419 26.361590  3.464447e-18  -8.695126
Profession=Artist         7.3905109 18.8372093 32.888222  1.244902e-22  -9.789842
Spending_Score=Average    6.0770156 11.7441860 24.936234  1.264191e-24 -10.243618
Spending_Score=Low        8.6271568 40.1162791 60.000000  2.016075e-36 -12.603645
Profession=Healthcare     0.2785515  0.3488372 16.159040  1.418519e-65 -17.102632
Ever_Married=No           1.3597942  4.3023256 40.825206 1.373759e-152 -26.312662

$`4`
                          Cla/Mod    Mod/Cla    Global       p.value     v.test
Ever_Married=Yes         36.282961 88.8268156 59.174794 4.333002e-193  29.641746
Spending_Score=Average   43.802647 45.1893234 24.936234  1.481803e-95  20.740883
Segmentation=C           41.511628 44.3202980 25.806452  4.996856e-79  18.821891
Profession=Artist        37.682482 51.2725016 32.888222  7.571977e-70  17.666683
Graduated=Yes            30.383620 80.1365611 63.750938  2.438103e-59  16.244629
Segmentation=B           32.188295 31.4090627 23.585896  9.592118e-17   8.309729
Profession=Executive     39.009901 12.2284295  7.576894  1.264159e-14   7.709395
Var_1=Cat_6              26.050045 72.3774053 67.156789  2.303111e-07   5.173044
Spending_Score=High      30.079681 18.7461204 15.063766  3.309543e-06   4.650610
Gender=Male              26.026652 59.4040968 55.168792  8.391793e-05   3.932926
Profession=Entertainment 27.935723 14.0285537 12.138035  8.443793e-03   2.633789
Var_1=Cat_2              19.889503  4.4692737  5.431358  4.737506e-02  -1.982932
Var_1=Cat_1              15.384615  0.9931719  1.560390  2.956762e-02  -2.175835
Profession=Homemaker     14.857143  1.6139044  2.625656  2.362363e-03  -3.040436
Var_1=Cat_4              18.963486  9.9937927 12.738185  1.071055e-04  -3.873904
Gender=Female            21.887550 40.5959032 44.831208  8.391793e-05  -3.932926
Profession=Marketing     13.304721  1.9242706  3.495874  3.079604e-05  -4.167499
Profession=Doctor        16.722973  6.1452514  8.882221  4.492639e-06  -4.587184
Segmentation=A           18.069307 18.1253880 24.246062  1.734099e-11  -6.726821
Profession=Lawyer         6.800000  2.1104904  7.501875  3.927129e-26 -10.574152
Graduated=No             13.245033 19.8634389 36.249062  2.438103e-59 -16.244629
Profession=Healthcare     1.857010  1.2414649 16.159040 1.239167e-109 -22.245781
Spending_Score=Low       14.528632 36.0645562 60.000000 9.265736e-111 -22.361820
Segmentation=D            5.634604  6.1452514 26.361590 4.754887e-121 -23.395447
Ever_Married=No           6.615215 11.1731844 40.825206 4.333002e-193 -29.641746
```

Πίνακας 3-14: Αποτελέσματα ποσοτικών μεταβλητών KAMILA

```
Description of each cluster by quantitative variables
======================================================
$`1`
                  v.test Mean in category Overall mean sd in category Overall sd      p.value
Work_Experience  16.20266         3.645303     2.629107       3.806592   3.405110 4.829104e-59
Family_Size     -12.07736         2.501957     2.841110       1.362589   1.524628 1.391125e-33
Age             -15.13827        38.929061    43.536084       3.744449  16.522814 9.055980e-52

$`2`
                  v.test Mean in category Overall mean sd in category Overall sd       p.value
Age              28.46824        53.260393    43.536084       5.158323   16.52281 2.898041e-178
Work_Experience -10.90517         1.861432     2.629107       2.704109    3.40511  1.088905e-27

$`3`
                   v.test Mean in category Overall mean sd in category Overall sd       p.value
Family_Size      24.989698         3.552894     2.841110       1.650530   1.524628 7.912136e-138
Work_Experience   4.120309         2.891218     2.629107       3.597951   3.405110  3.783649e-05
Age             -56.252363        26.172156    43.536084       4.428947  16.522814  0.000000e+00

$`4`
                  v.test Mean in category Overall mean sd in category Overall sd      p.value
Age              59.79262        74.464407    43.536084       7.331854  16.522814 0.000000e+00
Work_Experience -13.49110         1.190960     2.629107       2.113911   3.405110 1.764500e-41
Family_Size     -16.43867         2.056497     2.841110       1.024654   1.524628 1.011068e-60
```

- **1η ομάδα**
  - Ποιοτικές μεταβλητές:
    - Η 1η ομάδα χαρακτηρίζεται από καλλιτέχνες,

- απόφοιτους πανεπιστημίου και
- έχουν μεσαίο σκορ δαπανών.

➢ Ποσοτικές μεταβλητές:

- Έχουν κατά μέσο όρο 3,6 έτη επαγγελματικής εμπειρίας.

❖ **2η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 2η ομάδα χαρακτηρίζεται από άτομα που με χαμηλό σκορ δαπανών,
- εργάζονται στον τομέα της υγείας,
- είναι άγαμοι και
- δεν έχουν αποφοιτήσει από κάποιο πανεπιστημιακό ίδρυμα.

➢ Ποσοτικές μεταβλητές:

- Είναι κατά μέσο όρο 53 ετών.

❖ **3η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 3η ομάδα χαρακτηρίζεται από δικηγόρους,
- είναι έγγαμοι,
- με υψηλό σκορ δαπανών.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο ανήκουν σε 4μελείς οικογένειες.

❖ **4η ομάδα**

➢ Ποιοτικές μεταβλητές:

- Η 4η ομάδα χαρακτηρίζεται από έγγαμα άτομα,
- με μεσαίο σκορ δαπανών,
- είναι καλλιτέχνες και
- απόφοιτοι πανεπιστημίου.

➢ Ποσοτικές μεταβλητές:

- Κατά κύριο λόγο είναι 74 ετών.

4 Συμπεράσματα

Η τμηματοποίηση πελατών είναι μια σημαντική πρακτική του στρατηγικού σχεδιασμού του μάρκετινγκ και στην ερμηνεία της καταναλωτικής συμπεριφοράς του κοινού. Στηρίζεται στην υπόθεση πως ένα προϊόν δεν μπορεί να είναι αρεστό σε όλους ανεξαιρέτως τους καταναλωτές καθώς κάθε άτομο έχει διαφορετικές προτιμήσεις και τρόπο ζωής. Το μειονέκτημα της εφαρμογής τμηματοποίησης της αγοράς στον στρατηγικό σχεδιασμό μιας επιχείρησης είναι η επένδυση χρόνου και ανθρώπινου δυναμικού για να πραγματοποιηθεί μια σωστά δομημένη και αποτελεσματική έρευνα αγοράς.

Εφαρμόζοντας την προσέγγιση των δέκα βημάτων για την ανάλυση της τμηματοποίησης των πελατών στα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία, παρατηρείται πως η αυτοκινητοβιομηχανία έχει αποφασίσει να υιοθετήσει την πρακτική της τμηματοποίησης της αγοράς (1ο βήμα). Δεν έχει προσδιορίσει όμως τα χαρακτηριστικά αυτής της ομάδας που θεωρεί ιδανικό προφίλ πελάτη (2ο βήμα) καθώς θα ήθελε η έρευνα να προβλέψει αυτό το ιδανικό προφίλ. Έχουν συλλεχθεί τα δεδομένα πιθανών πελατών και η αυτοκινητοβιομηχανία έπειτα από έρευνα αγοράς έχει καταλήξει πως είναι καλύτερο να ταξινομηθούν οι πελάτες σε 4 ομάδες (3ο βήμα). Στην εφαρμογή των μεθόδων ανάλυσης στα πραγματικά δεδομένα έγινε χρήση των εξής μεθόδων για την διερεύνηση των ομάδων (4ο βήμα):

1) Συντελεστής ανομοιότητας του Gower & Partitioning Around Medoids
2) Αλγόριθμος K-prototypes
3) Μέθοδος Mixed K-means
4) Μέθοδος Modha-Spangler convex K-means
5) Μέθοδος FAMD + K-means (Two-step)
6) Μέθοδος Mixed Reduced K-means
7) Αλγόριθμος KAMILA (KAy-means for MIxed LArge datasets)

Μετά την αξιολόγηση των αποτελεσμάτων κάθε μεθόδου βάσει του προσαρμοσμένου δείκτη Rand επιλέχθηκε ως η πιο ικανοποιητική συσταδοποίηση να έχει γίνει από τη μέθοδο Mixed Reduced K-means (ARI = 0,1261) και έπειτα με μικρή διαφορά από το συνδυασμό FAMD + K-means (ARI = 0,1256) (5ο βήμα). Οι 4 ομάδες που σχηματίστηκαν από τη Mixed Reduced K-means έχουν το εξής προφίλ (6ο βήμα):

- Στην 1η ομάδα ανήκουν οι καλλιτέχνες πελάτες που είναι απόφοιτοι πανεπιστημίου και παντρεμένοι και έχουν ένα μεσαίο σκορ δαπανών.

- Στην 2η ομάδα ανήκουν γυναίκες που εργάζονται είτε ως μηχανικοί είτε στο χώρο της διασκέδασης, είναι άγαμες που δεν δαπανούν πολλά και έχουν περίπου 3,5 έτη επαγγελματικής εμπειρίας.
- Στην 3η ομάδα ανήκουν πελάτες που εργάζονται στον τομέα της υγείας, δεν δαπανούν πολλά, είναι άγαμοι και ανήκουν σε 4μελείς οικογένειες.
- Στην 4η ομάδα ανήκουν πελάτες που εργάζονται ως δικηγόροι, είναι παντρεμένοι, έχουν υψηλό σκορ δαπανών και είναι περίπου 69 ετών.

Λόγω της χαμηλής τιμής του προσαρμοσμένου δείκτη Rand, η οποία είναι μικρότερη του 0,6, η επιχείρηση θα μπορούσε να εξετάσει και άλλους τρόπους τμηματοποίησης ή να μαζέψει εκ νέου δεδομένα που να περιγράφουν καλύτερα το καταναλωτικό της κοινό. Γι' αυτό η ενδελεχή περιγραφή των ομάδων παραλείπεται (7ο βήμα) καθώς οι ομάδες θα πρέπει να επαναπροσδιοριστούν. Αφού γίνει η περιγραφή των νέων, πιο ομοιογενών, ομάδων θα μπορούσε να επιλεχθεί η ομάδα που θα στοχεύσει το τμήμα μάρκετινγκ της αυτοκινητοβιομηχανίας (8ο βήμα). Πάνω σε αυτή την ομάδα θα βασιστεί το μείγμα μάρκετινγκ για την προώθηση των αυτοκινήτων της σε νέο καταναλωτικό κοινό (9ο βήμα). Τέλος, η αυτοκινητοβιομηχανία θα χρειαστεί να ορίσει μια ομάδα αναλυτών της για να παρακολουθεί διαρκώς την επιλεγμένη ομάδα στόχευσης για πιθανές αλλαγές στο μέγεθος και στα χαρακτηριστικά τους ώστε το μείγμα μάρκετινγκ που θα χρησιμοποιείται να είναι πάντα επικαιροποιημένο και προσαρμοσμένο στο κοινό του (10ο βήμα).

5 Βιβλιογραφία

Abdalameer, A. K., Alswaitti, M., Alsudani, A. A., & Isa, N. A. (2022, April 1). A new validity clustering index-based on finding new centroid positions using the mean of clustered data to determine the optimum number of clusters. *Expert Systems with Applications*.

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 503-527.

Anand, D. (2020, June 17). *Gower's Distance*. Ανάκτηση από Medium: https://medium.com/analytics-vidhya/gowers-distance-899f9c4bd553

Aprilliant, A. (2021, January 17). *The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical)*. Ανάκτηση από Towards Data Science.

Bai, L., Liang, J., & Cao, F. (2020, September). A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters. *Information Fusion*, σσ. 36-47.

Davies, D., & Bouldin, D. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), σσ. 224-227.

Dolnicar, S., Grün, B., & Leisch, F. (2018). *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful.*

Foss, A., Markatou, M., Ray, B., & Heching, A. (2016). A semiparametric method for clustering mixed data. *Machine Learning*, σσ. 419-458.

Guo, G., Chen, L., Ye, Y., & Jiang, Q. (2017, December). Cluster Validation Method for Determining the Number of Clusters in Categorical Sequences. *IEEE Transactions on Neural Networks and Learning Systems*, σσ. 2936-2948.

Helm, M. (2021, August 20). *A deep dive into partitioning around medoids.* Ανάκτηση από Towards Data Science: https://towardsdatascience.com/a-deep-dive-into-partitioning-around-medoids-a77d9b888881

Huang, J., Sun, H., Kang, J., Qi, J., Deng, H., & Song, Q. (2013, March). ESC: An efficient synchronization-based clustering algorithm. *Knowledge-Based Systems*, σσ. 111-122.

Ikotun, A., Ezugwu, A., Abualigah, L., Abuhaija, B., & Heming, J. (2023, April). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, σσ. 178-210.

Jain, A. K. (2010, June 01). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, σσ. 651-666.

Kaushik, A. (2015). *See, Think, Do Care Winning Combo: Content +Marketing +Measurement!* Ανάκτηση από Kaushik.net:

https://www.kaushik.net/avinash/see-think-do-care-win-content-marketing-measurement

Kolesnikov, A., Trichina, E., & Kauranne, T. (2015, March). Estimating the number of clusters in a numerical data set via quantization error modeling. *Pattern Recognition*, σσ. 941-952.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A Package for Multivariate Analysis.” . *Journal of Statistical Software*, 1-18.

Leich, F. (2006, 11 15). A toolbox for k-centroids cluster analysis. *Computational Statistics & Data Analysis*, *51*(2), 536-544.

Ling, H.-L., Wu, J.-S., Zhou, Y., & Zheng, W.-S. (2016, September 26). How many clusters? A robust PSO-based local density model. *Neurocomputing*, σσ. 264-275.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2022). *cluster: Cluster Analysis Basics and Extensions. R package version 2.1.4 — For new features, see the 'Changelog' file (in the package source)*. Ανάκτηση από https://CRAN.R-project.org/package=cluster

Manochandar, S., Punniyamoorthy, M., & Jeyachitra, R. (2020, March). Development of new seed with modified validity measures for k-means clustering. *Computers and Industrial Engineering*.

Markos, A., Iodice D'Enza, A., & Van de Velden, M. (2019). Beyond Tandem Analysis: Joint Dimension Reduction and Clustering in R. *Journal of Statistical Software*, 1-24.

Markos, A., Moschidis, O., & Chadjipadelis, T. G. (2020). Sequential dimension reduction and clustering of mixed-type data. *International Journal of Data Analysis Techniques and Strategies*, 28-30.

Modha, D. S., & Spangler, W. S. (2003). Feature Weighting in k-Means Clustering. *Machine Learning*, 217-237.

Morgan, N., Whitler, K., Feng, H., & Chari, S. (2019). Research in marketing strategy. *Journal of the Academy of Marketing Science*, σσ. 4-29.

Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée*, 93-111.

Peter, J. P., & Olson, J. C. (2010). *Consumer Behavior & Marketing Strategy*.

Rojas-Thomas, J., Santos, M., & Mora, M. (2017, November 15). New internal index for clustering validation based on graphs. *Expert Systems with Applications*, σσ. 334-349.

Rousseeuw, P. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 53-65.

Rust, R. T., Moorman, C., & Bhalla, G. (2009). Rethinking Marketing. *Harvard Business Review*.

Steinley, D., & Brusco, M. J. (2007). Initializing K-means Batch Clustering: A Critical Evaluation of Several Techniques. *Journal of Classification, 24*, 99-121.

Szepannek, G. (2018). clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *The R Journal*, 200-208.

Van de Velden, M., Iodice D'Enza, A., & Markos, A. (2019, April 10). Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*.

Vollrath, M. D., & Villegas, S. G. (2021). Avoiding digital marketing analytics myopia: revisiting the customer decision journey as a strategic marketing framework. *Journal of Marketing Analytics*.

Zhang, Y., Mańdziuk, J., Hiok Quek, C., & Wooi Goh, B. (2017, November). Curvature-based method for determining the number of clusters. *Information Sciences*, σσ. 414-428.

Zhou, Y., Wu, H., Luo, Q., & Mohamed, A.-B. (2019, January 01). Automatic data clustering using nature-inspired symbiotic organism search algorithm. *Knowledge-Based Systems*, σσ. 546-557.

Παράρτημα Α - Κώδικας R

```
install.packages("MSA")
load("dat.Rdata")

## Apply Clustering to Mixed-Type Data
install.packages("mclust")
require(mclust)
install.packages("cluster")
require(cluster)
install.packages("clustMD")
require(clustMD)
install.packages("clustMixType")
require(clustMixType)
install.packages("kmed")
require(kmed)
install.packages("kamila")
require(kamila)
install.packages("FactoMineR")
require(FactoMineR)
require(fpc)

#################### DISTANCE-BASED METHODS ####################

dat <- dat[complete.cases(dat),]
# 6665 customers left

#1. Gower's distance + PAM (Partitioning Around Medoids)
# Compute Gower distance
gower_dist <- daisy(dat[,-c(1,11)], metric = "gower")
gower_mat <- as.matrix(gower_dist)
# Print most similar objects
```

```
dat[which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]), arr.ind = TRUE)[1, ], ]
# Partitioning Around Medoids
pam_fit <- pam(gower_dist, diss = TRUE, k = 4)
# Average Sihouette Width (ASW)
sil_width <- pam_fit$silinfo$avg.width
sil_width

# ARI = 0.1103
table(pam_fit$clustering)
adjustedRandIndex(pam_fit$clustering,dat$Segmentation)

# Ερμηνεία των ομάδων
catdes(cbind(dat,as.factor(pam_fit$clustering)),12)



# 2. K-prototypes
outk = kproto(dat[,-c(1,11)], 4)

# ARI = 0.1107
table(outk$cluster,dat$Segmentation)
adjustedRandIndex(outk$cluster,dat$Segmentation)

# Ερμηνεία των ομάδων
catdes(cbind(dat,as.factor(outk$cluster)),12)

# 3. Mixed K-means
# Distances for mixed variables data set (gower, wishart, podani, huang, harikumar, ahmad)
mix <- distmix(dat[,-c(1,11)], method = "ahmad", idcat = c(1,2,4,5,7,9), idnum = c(3,6,8))
kmedres <- fastkmed(mix, 4, iterate = 100, init = NULL)
```

```
# ARI = 0.1044
table(kmedres$cluster,dat$Segmentation)
adjustedRandIndex(kmedres$cluster,dat$Segmentation)

# Ερμηνεία των ομάδων
catdes(cbind(dat,as.factor(kmedres$cluster)),12)



# 4. Modha-Spangler K-means
datnew <- dat[,-c(1,11)]
conDf <- data.frame(scale(datnew[,c(3,6,8)]))
catDf <- dummyCodeFactorDf(data.frame(datnew[,c(1,2,4,5,7,9)]))
#Modha-Sprangler
msRes <- gmsClust(conDf, catDf, nclust = 4)

# ARI = 0.0850
adjustedRandIndex(msRes$results$cluster, dat$Segmentation)

# Ερμηνεία των ομάδων
catdes(cbind(dat,as.factor(msRes$results$cluster)),12)



# 5. FAMD + K-means (Two-step)
howmany <- estim_ncp(data.matrix(datnew))$ncp
outpcamix <- FAMD(datnew, ncp = howmany)
outkm <- kmeans(outpcamix$ind$coord, 4, nstart = 100)

# ARI = 0.1256
adjustedRandIndex(outkm$cluster, dat$Segmentation)

# Ερμηνεία των ομάδων
catdes(cbind(dat,as.factor(outkm$cluster)),12)
```

```
# 6. Mixed Reduced K-means
install.packages("clustrd")
require(clustrd)
outc = cluspcamix(datnew, nclus = 4, ndim = 1, nstart = 100)

# ARI = 0.1261
adjustedRandIndex(outc$cluster, dat$Segmentation)

# Ερμηνεία των ομάδων
catdes(cbind(dat,as.factor(outc$cluster)),12)



# 7. KAMILA

outkam <- kamila(datnew[,c(3,6,8)],datnew[,-c(3,6,8)],numClust = 4,numInit =
100)

# ARI = 0.1170
adjustedRandIndex(outkam$finalMemb, dat$Segmentation)

# Ερμηνεία των ομάδων
catdes(cbind(dat,as.factor(outkam$finalMemb)),12)
```