Στατιστική & Μηχανική Μάθηση – 1η Εργασία

Δικαιόπουλος Μιχαήλ - 3180050

Παρουσιάζονται τα αποτελέσματα της ταξινόμησης των πελατών ενός αρχείου συνολικά 440 πελατών σε ομάδες με χρήση διαφορετικών αλγόριθμων ομαδοποίησης (k-means, hierarchical clustering, model based clustering) και η σύγκριση των αποτελεσμάτων.

Περιγραφική ανάλυση των δεδομένων

Το σετ των δεδομένων μας περιέχει 440 πελάτες (παρατηρήσεις) και 8 μεταβλητές, τις Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen. Οι μεταβλητές Channel και Region είναι κατηγορικές (2 και 3 επίπεδα αντίστοιχα), ενώ οι υπόλοιπες αριθμητικές.

Τα πλήθη των παραδειγμάτων για κάθε κατηγορική μεταβλητή, είναι τα εξής :

- Channel : #**1** – 298 #**2** – 142
- Region : #**1** – 77 #**2** – 47 #**3** – 316

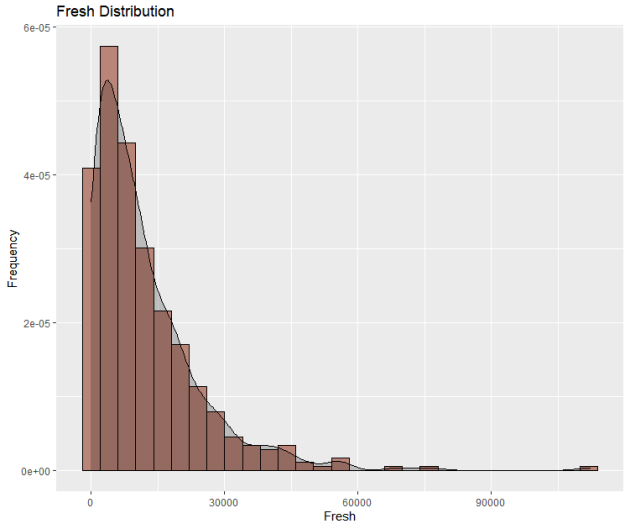
Για κάθε αριθμητική μεταβλητή, παρατίθενται η ελάχιστη και η μέγιστη τιμή της, το άθροισμα των τιμών της, η διάμεσος και η μέση τιμή της, η διασπορά και η τυπική απόκλισή της.

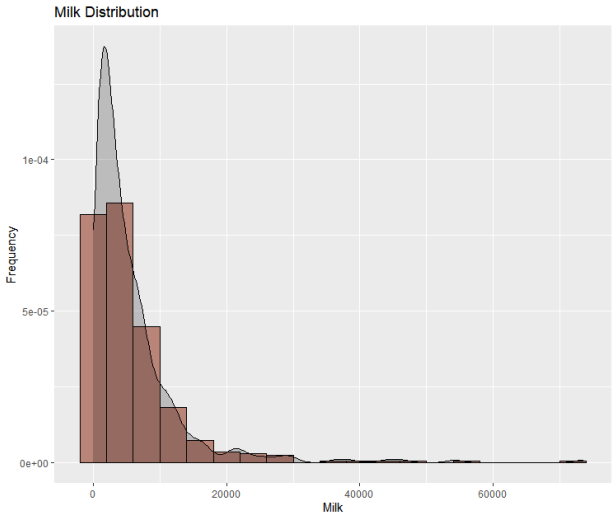
	Fresh	**Milk**	**Grocery**	**Frozen**	**Detergents_Paper**	**Delicassen**
Min	3	55	3	25	3	3
Max	112151	73498	92780	60869	40827	47943
Sum	5280131	2550357	3498562	1351650	1267857	670943
Median	8504	3627	4755.5	1526	816.5	965.5
Mean	12000.3	5796.266	7951.277	3071.932	2881.493	1524.87
Variance	159954927.4	54469967.24	90310103.75	23567853.17	22732436.04	7952997.498
Standard Deviation	12647.33	7380.377	9503.163	4854.673	4767.854	2820.106

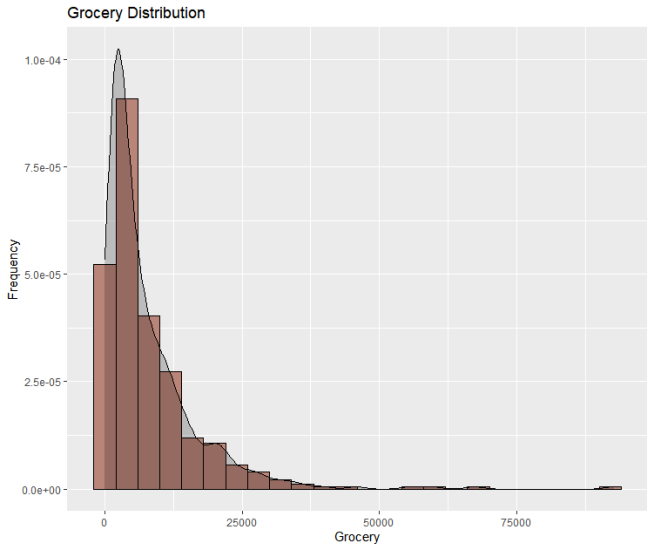
Παρακάτω, παρατίθεται ο πίνακας συσχετίσεων του Pearson για κάθε ζευγάρι μεταβλητών. Με χρήση της τιμής πρακτικής σημαντικότητας 0.7 (Chatfield, Collins – 1992), διακρίνουμε δύο σημαντικές συσχετίσεις, της μεταβλητής **Milk** με τη **Grocery** και της μεταβλητής **Grocery** με τη **Detergents_Paper**.

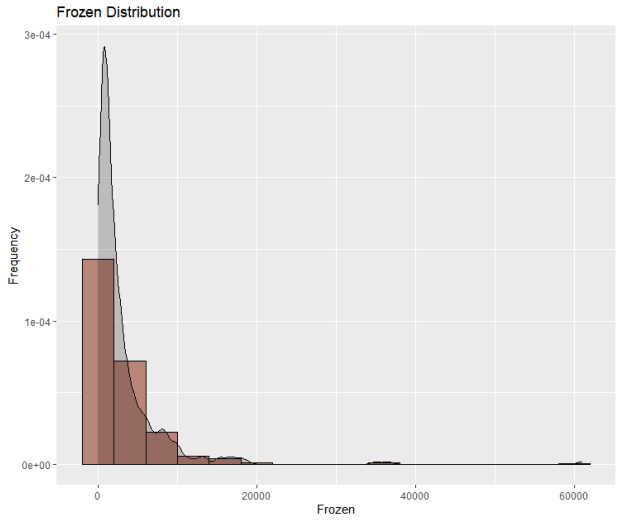
	Fresh	**Milk**	**Grocery**	**Frozen**	**Detergents_Paper**	**Delicassen**
Fresh	1	0.101	-0.012	0.346	-0.102	0.245
Milk	0.101	1	0.728	0.124	0.662	0.406
Grocery	-0.012	0.728	1	-0.04	0.925	0.205
Frozen	0.346	0.124	-0.04	1	-0.132	0.391
Detergents_Paper	-0.102	0.662	0.925	-0.132	1	0.069
Delicassen	0.245	0.406	0.205	0.391	0.069	1

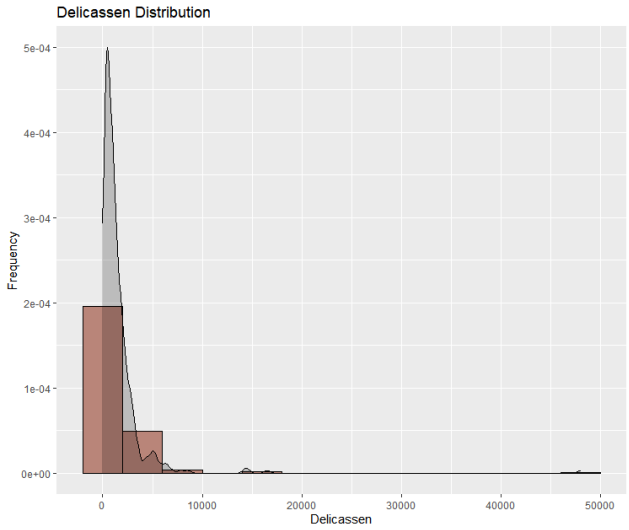
Παρακάτω θα παραθέσουμε τα ιστογράμματα μαζί με τις κατανομές των 6 αριθμητικών μεταβλητών. Γίνεται εύκολα αντιληπτά από τα ιστογράμματα τα διαφορετικά εύρη τιμών (με σταθερο μήκος κάδου) και επομένως η ανάγκη κανονικοποίησης των δεδομένων για πιο αξιόπιστα αποτελέσματα.

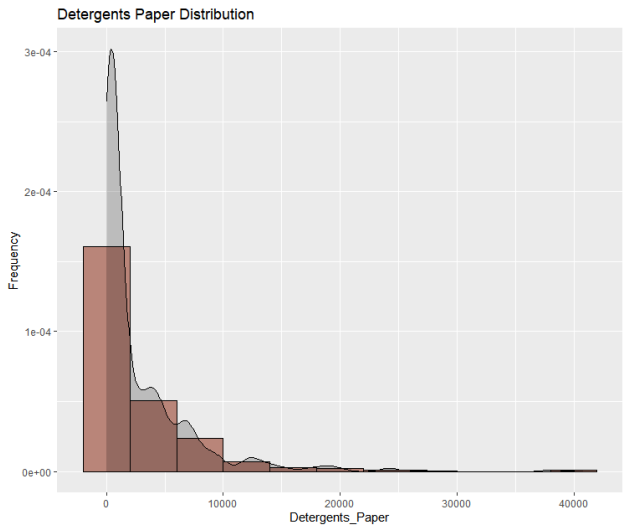












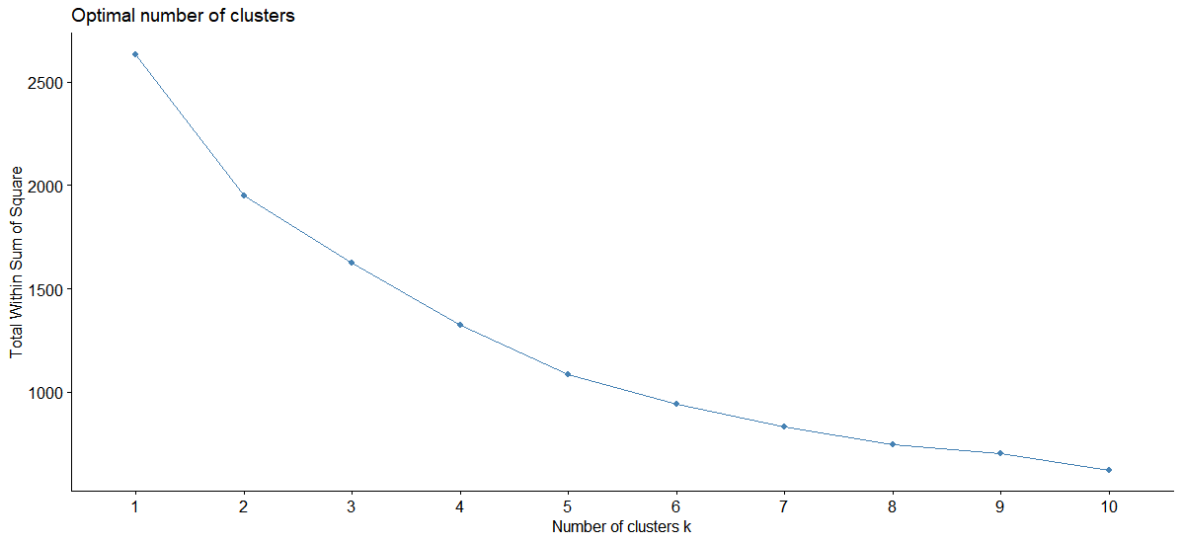
Επίσης, για την ανάλυση των ομαδοποιήσεών μας θα μας είναι χρήσιμο να εντοπίσουμε τις πολύ ακραίες τιμές. Εφόσον το σύνολο των δεδομένων μας είναι πολυμεταβλητό, το καλύτερο μέτρο απόστασης είναι η απόσταση Mahalanobis. Βρίσκουμε 3 πολύ ακραίες τιμές, τις γραμμές 87, 184 και 326, με τη γραμμή 184 μάλιστα να αποτελεί εξαιρετικά ακραία τιμή που ενδεχομένως να επηρεάσει πολύ τις ομαδοποιήσεις μας. Παρατίθεται και το scatterplot με τις αποστάσεις Mahalanobis.

Αλγόριθμος k-means

Θα τρέξουμε τον αλγόριθμο k-means, αρχικά σε όλο το dataset και έπειτα στο dataset δίχως τις ακραίες τιμές που επισημάναμε προηγουμένως. Ο λόγος είναι ότι αναμένουμε πως ο k-means επηρεάζεται πολύ από αυτές τις ακραίες τιμές λόγω της φύσης του (χρήση μέσων τιμών) και αναμένουμε καλύτερη ποιότητα ομαδοποίησης βάσει των αρχικών κατηγοριοποιήσεων με αφαίρεση των ακραίων τιμών.

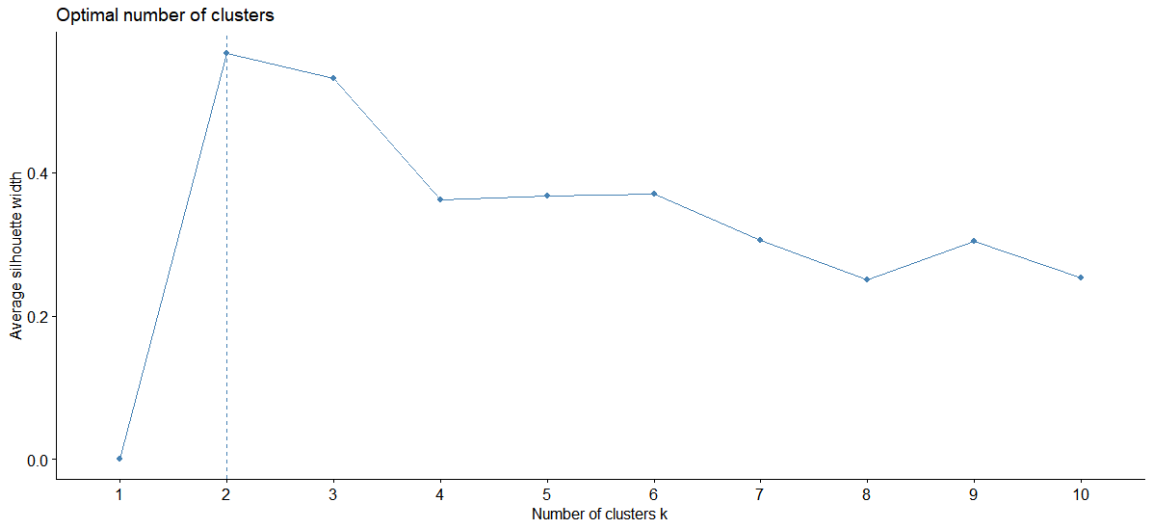
Αφού κάνουμε scale τα δεδομένα μας τρέχουμε τις τρεις μεθόδους εύρεσης του optimal k.

- 1η Μέθοδος : Elbow Method



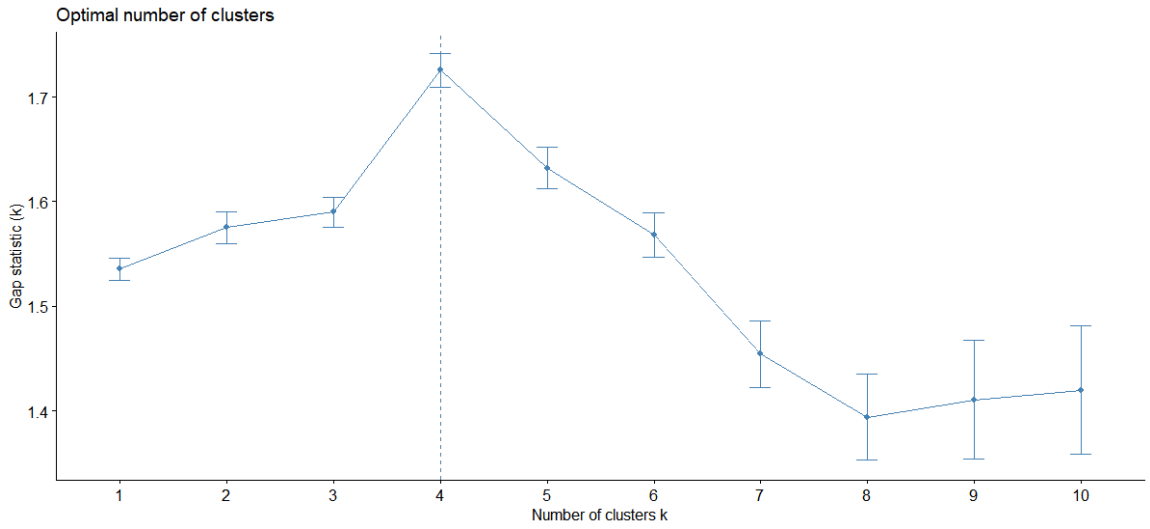
Εδώ το πρόβλημα είναι ότι το γράφημα δεν μοιάζει με αγκώνα, έχει μικρή καμπυλότητα και δεν μπορούμε να ορίσουμε με σιγουριά το k, αλλά θεωρούμε ότι είναι μεταξύ 4 και 5 clusters.

- 2η Μέθοδος : Silhouette Method



Η Silhouette Method ορίζει ως optimal k το k = 2.

- 3η Μέθοδος : Gap Statistic



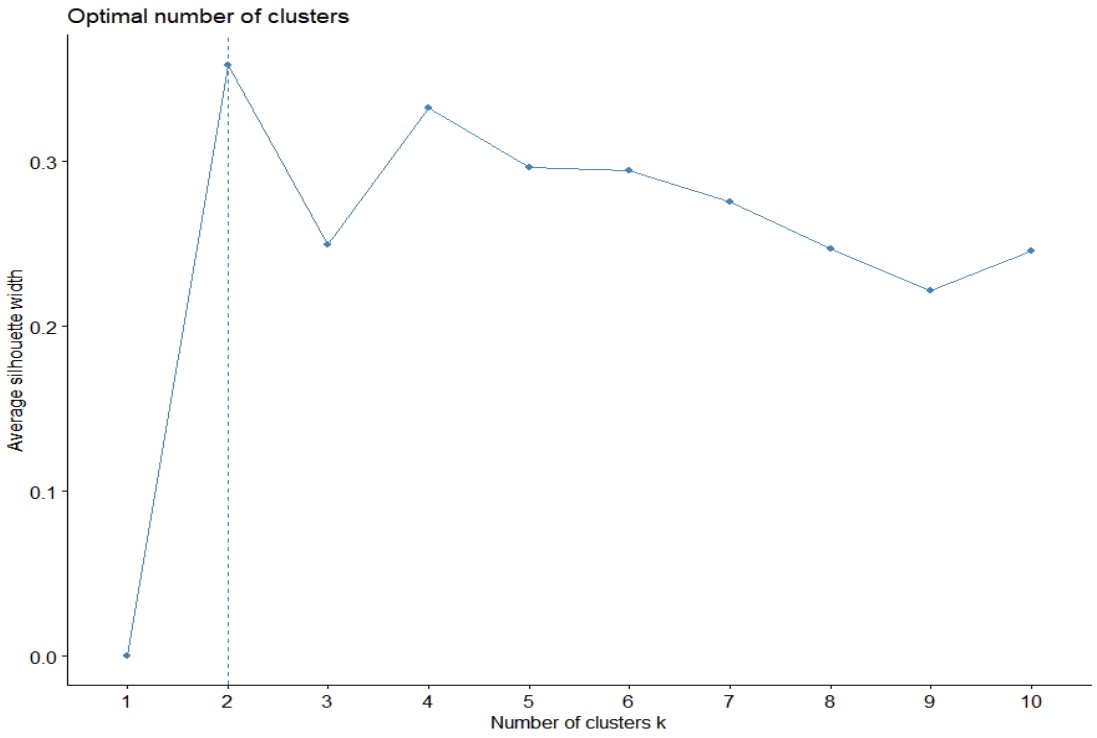
Εδώ, ορίζεται ως βέλτιστο k το k = 4.

Θα ορίσουμε ως βέλτιστο k με βάση τα αποτελέσματα των 3 μεθόδων το k = 3. Έχουμε ήδη από τα δεδομένα μία ισχυρή εκτίμηση ότι μπορούν να ομαδοποιηθούν σε τρία clusters (με βάση το Region) ή σε 2 clusters με βάση το Channel. Επίσης για k = 2 και k = 3 έχουμε το μέγιστο silhouette average coefficient και για k = 3 έχουμε χαμηλότερο wss και υψηλότερο gap statistic απ'ότι για k = 2. Για k = 4 έχουμε αρκετά υψηλό gap statistic, αλλά με βάση το average silhouette coefficient καταλαβαίνουμε ότι πρόκειται για μία ομαδοποίηση με ασθενική δομή (< 0.4)

- Χαρακτηρισμός Ποιότητας Ομαδοποίησης χωρίς Επίβλεψη : Για k = 3, υπολογίζουμε το average silhouette coefficient ως 0.54, ορίζοντας την δομή ως οριακά ενδιαφέρουσα. Ωστόσο επειδή έχουμε και 2 μεταβλητές βάσει των οποίων μπορούμε να κατηγοριοποιήσουμε τα δεδομένα μας σε δύο ή τρεις κατηγορίες, θα εξετάσουμε και την ποιότητα της κατηγοριοποίησης βάσει αυτών των χαρακτηριστικών.

- Χαρακτηρισμός Ποιότητας Ομαδοποίησης με Επίβλεψη : Εξετάζουμε αν υπάρχει ισχυρή γραμμική συσχέτιση μεταξύ των ομάδων που έχει βρει ο k-means αλγόριθμος και της ήδη υπάρχουσας κατηγοριοποίησης των δεδομένων μας. Όσον αφορά την κατηγοριοποίηση με βάση το Region (δηλαδή αν οι πελάτες είναι από τη Λισαβόνα, το Πόρτο ή κάποια άλλη περιοχή), συγκρίνουμε με την ομαδοποίηση για k = 3, εφόσον θέλουμε ο αριθμός των ομάδων να είναι ίδιος με τον αριθμό των κλάσεων. Βρίσκουμε γραμμικό συντελεστή συσχέτισης corr = 0.034, άρα δεν παρατηρούμε κάποια σχέση της ομαδοποίησης μας με την προϋπάρχουσα κατηγοριοποίηση. Αν συγκρίνουμε τώρα την κατηγοριοποίηση των δεδομένων με βάση το Channel και την ομαδοποίηση μας για k = 2, βρίσκουμε γραμμικό συντελεσή συσχέτισης corr = 0.249, που υποδεικνύει καλύτερη μεν, αλλά σχετικά αδύναμη σχέση μεταξύ τους.
  Επομένως, δεν μπορούμε να χαρακτηρίσουμε ως καλή την ποιότητα της ομαδοποίησης ούτε με κριτήριο το average silhouette coefficient ούτε με κριτήριο την προϋπάρχουσα κατηγοριοποίηση.

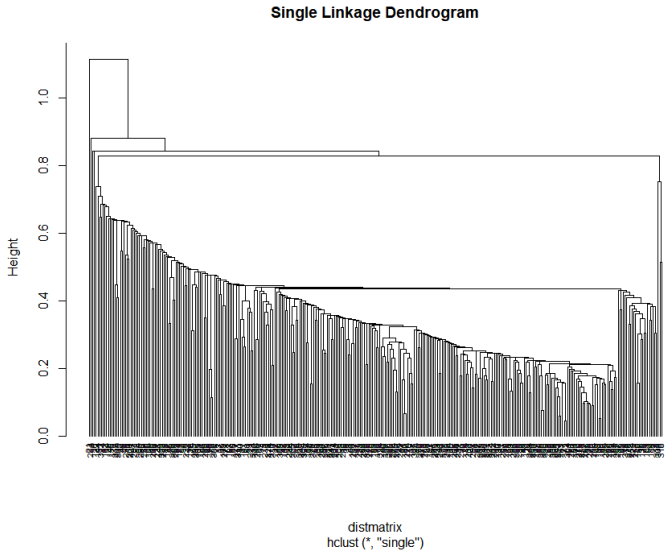
Μετά την αφαίρεση των outliers, ακολουθούμε την ίδια διαδικασία με πριν. Στο γράφημα της Silhouette Method, παρατηρούμε μία ενδιαφέρουσα αλλαγή :

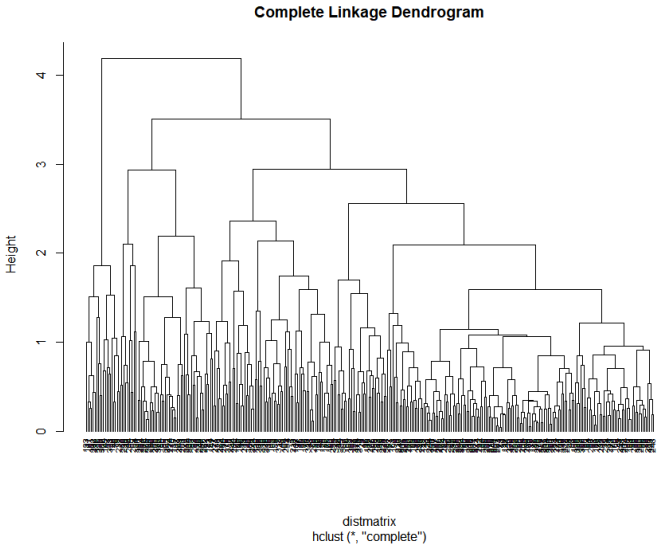


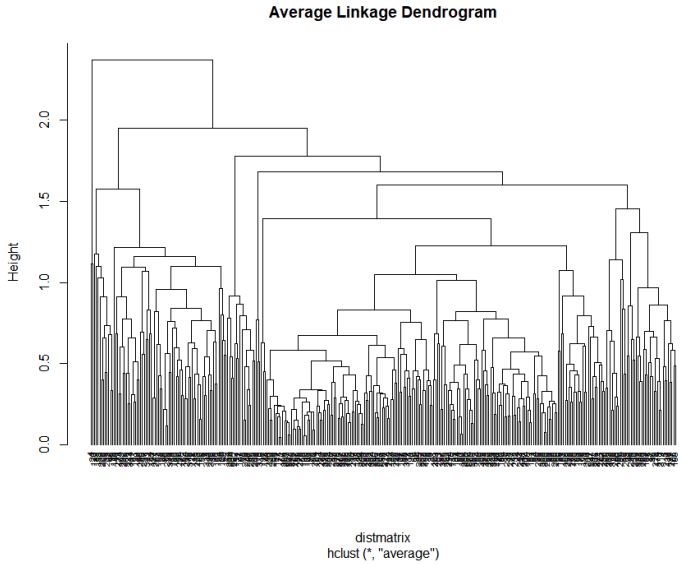
Εδώ το average silhouette coefficient για k = 3 μειώνεται πολύ. Αυτό κρίνουμε ότι συμβαίνει διότι μία εκ των ομάδων πριν αποτελείτο από δεδομένα που είχαν ακραίες τιμές στο σύνολο δεδομένων μας. Πράγματι, αν ελέγξουμε το πλήθος δεδομένων κάθε ομάδας στην προηγούμενη ομαδοποίηση, θα δούμε ότι η μία ομάδα αποτελείτο από τρία δείγματα, εκ των οποίων τα δύο άνηκαν στα outliers που απορρίψαμε. Έχοντας μία αίσθηση για το πώς κατηγοριοποιούνται τα δεδομένα με βάση το Region σε 3 ομάδες και βλέποντας την σχεδόν ανύπαρκτη συσχέτιση που παρατηρήσαμε μεταξύ της προηγούμενης ομαδοποίησης για k = 3 και της κατηγοριοποίησης με βάση το Region(0.034), καταλήγουμε ότι η ομαδοποίηση σε 3 ομάδες δεν είναι η ιδανική και έδινε υψηλό average silhouette coefficient εξαιτίας της μικρής 3ης ομάδας. Τώρα για k = 2, και χωρίς τα outliers, έχουμε ότι το average silhouette coefficient ισούται με 0.366 ενώ ο γραμμικός συντελεστής συσχέτισης με 0.613, παρατηρώντας έτσι μία καλύτερη προσέγγιση της ομαδοποίησης στην κατηγοριοποίηση με βάση το Channel.

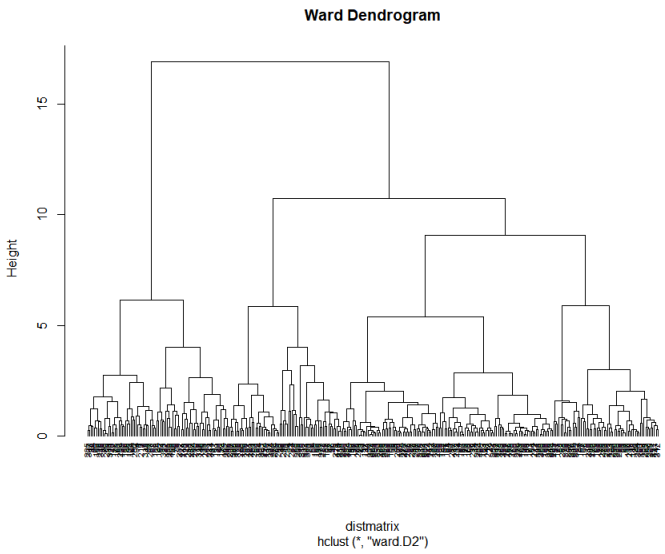
Hierarchical Clustering

Ξεκινάμε όπως και πριν κανονικοποιώντας τα δεδομένα μας. Εκτελούμε ιεραρχική ομαδοποίηση με τις μεθόδους single linkage, complete linkage, average linkage και Ward.



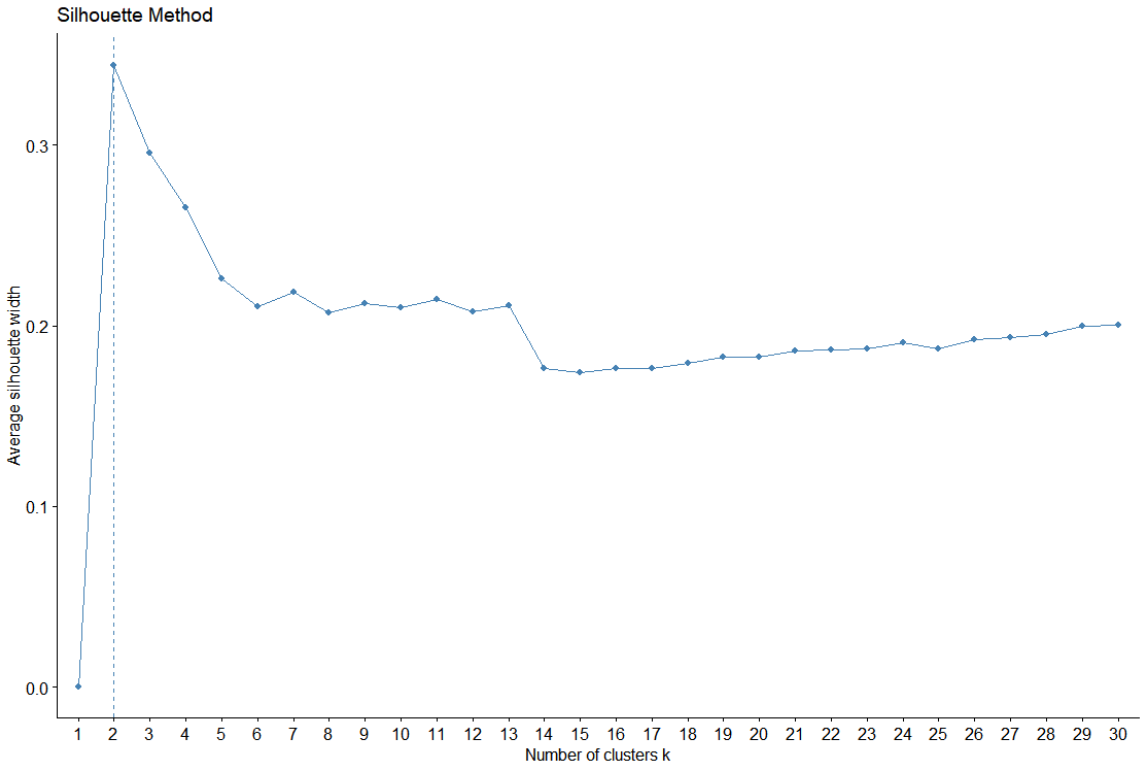






Χρησιμοποιούμε τον συντελεστή agglomerative coefficient για να ελέγξουμε την ισχύ της δομής της ομαδοποίησης για κάθε μέθοδο, με τον αλγόριθμο Ward να παρουσιάζει την ισχυρότερη δομή (coef = 0.979) και την μέθοδο Single Linkage την χαλαρότερη (coef = 0.712)

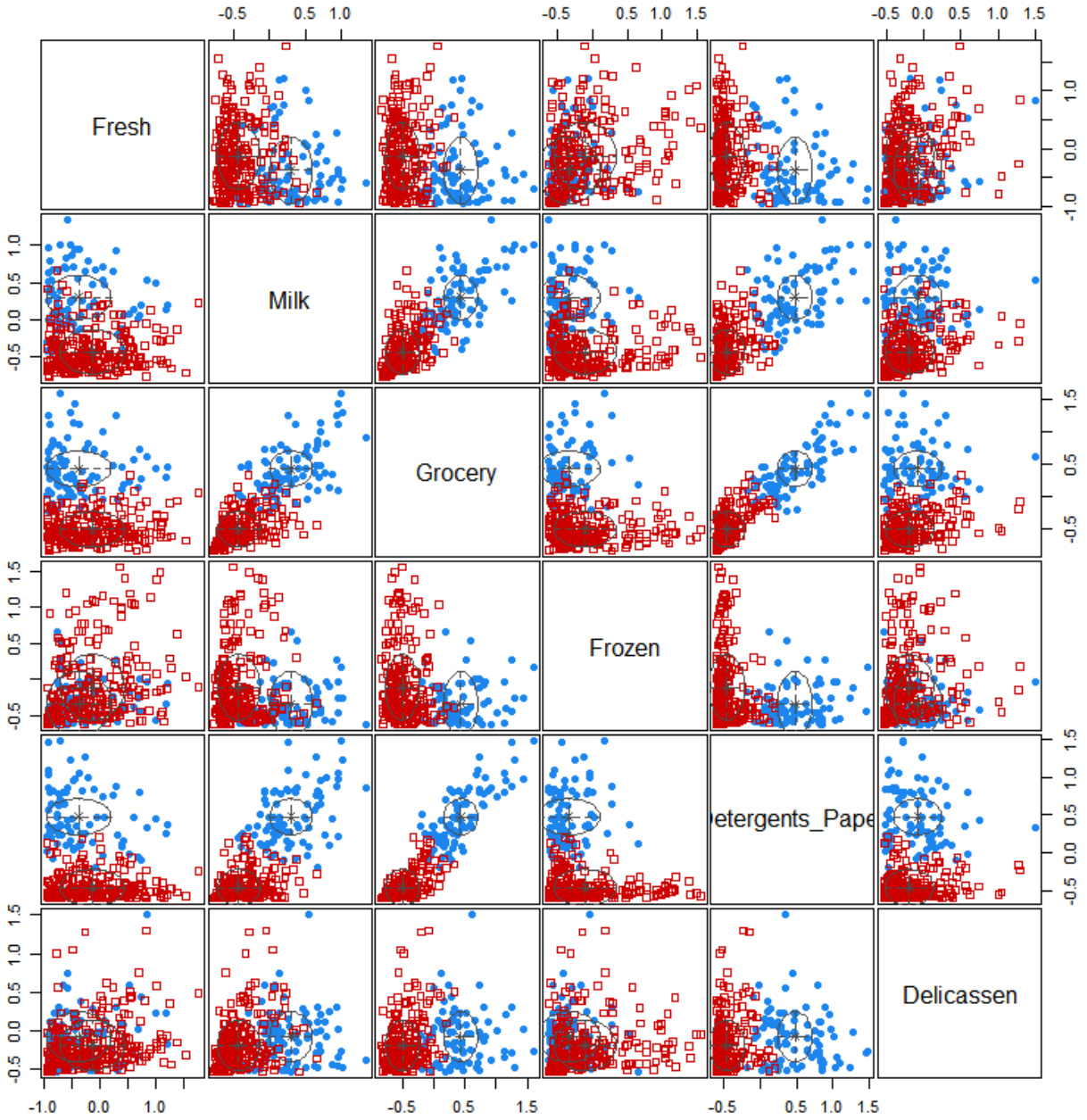
Εξετάζοντας την ποιότητα της ομαδοποίησης, αρχικά χωρίς επίβλεψη, θα δούμε ότι δεν διακρίνουμε κάποια ένδειξη καλής ομαδοποίησης για οποιονδήποτε αριθμό ομάδων. Όπως φαίνεται και στο παρακάτω διάγραμμα με τους average silhouette coefficients, μόνο για k = 2 ο συντελεστής υπερβαίνει το φράγμα του 0.3 (συγκεκριμένα ισούται με 0.344), ορίζοντας μία ασθενική δομή.

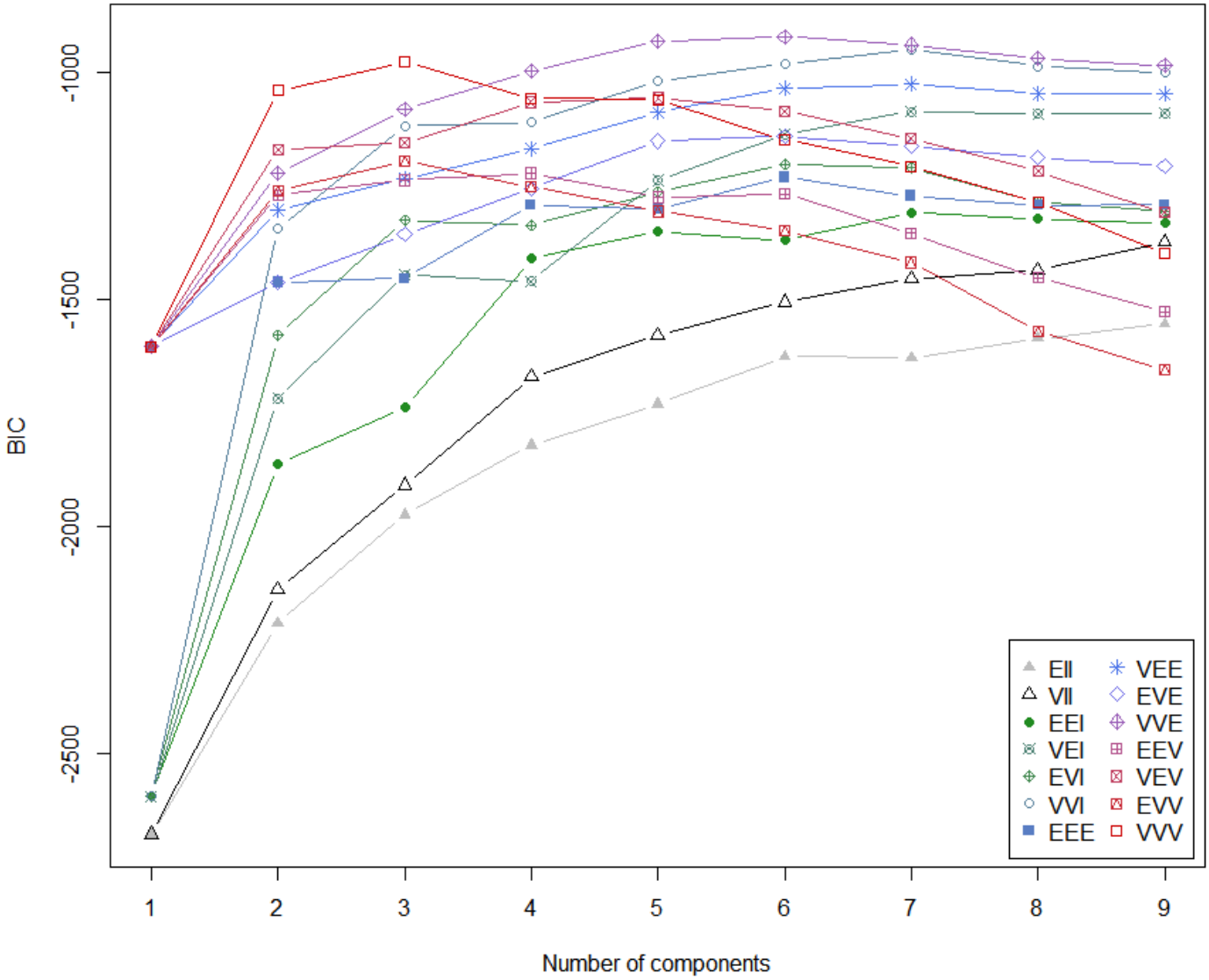


Όσον αφορά την ποιότητα της ομαδοποίησης με βάση την κατηγοριοποίηση με το Channel και το Region, τα αποτελέσματα πάλι είναι χειρότερα από ότι με τον k-means αλγόριθμο. Χρησιμοποιώντας τη μέθοδο του Ward, με την οποία επιτυγχάνουμε τα βέλτιστα αποτελέσματα στο σύνολο των μεθόδων ιεραρχικής ομαδοποίησης, ο γραμμικός συντελεστής συσχέτισης μεταξύ της ομαδοποίησης σε 2 ομάδες και της κατηγοριοποίησης με βάση το Channel ισούται με 0.521, μικρότερο από το 0.613 που βρήκαμε με τον k-means. Ο αντίστοιχος συντελεστής μεταξύ της ομαδοποίησης σε 3 ομάδες και της κατηγοριοποίησης με βάση το Region είναι πολύ κοντά στο 0 και άρα αποκλείει την ύπαρξη γραμμικής σχέσης μεταξύ των δύο.

Model - Based Clustering

Αφού κανονικοποιήσουμε τα δεδομένα μας, πραγματοποιούμε model-based clustering, ελέγχοντας διαφορετικά μοντέλα Γκαουσιανών κατανομών με κριτήριο το average silhouette coefficient του μοντέλου και το γραμμικό συντελεστή συσχέτισης της ομαδοποίησης σε 2 cluster και της κατηγοριοποίησης με βάση το Channel. Με κανένα μοντέλο δεν παρατηρείται γραμμική συσχέτιση μεταξύ της ομαδοποίησης σε 3 cluster και της κατηγοριοποίησης με βάση το Region. Επιλέγουμε ως βέλτιστο μοντέλο το "EEI", δηλαδή οι 2 ομάδες έχουν ίδιο όγκο και σχήμα, καθώς και κατεύθυνση ίδια με τους άξονες των συντεταγμένων. Ο γραμμικός συντελεστής συσχέτισης αυτού του μοντέλου ισούται με 0.652, που είναι και η καλύτερη γραμμική σχέση που έχουμε βρει ως τώρα μεταξύ ομαδοποίησης και κατηγοριοποίησης με βάση το Region. Παρακάτω μπορούμε να δούμε το γράφημα της ομαδοποίησης σε όλα τα δισδιάστατα επίπεδα και το διάγραμμα BIC (χωρίς ωστόσο η επιλογή των ομάδων και του μοντέλου να το λαμβάνει αυστηρά υπόψη, καθώς γίνεται βαρύτητα στην κατηγοριοποίηση που γνωρίζουμε από τα δεδομένα μας).





Τελικά Συμπεράσματα

Το κύριο κριτήριο αξιολόγησης που χρησιμοποιούμε για τους αλγορίθμους που αναφέραμε, είναι η κατηγοριοποίηση που γνωρίζαμε από πριν για τα δεδομένα μας. Όπως παρατηρήσαμε, και οι 3 μέθοδοι σε διαφορετικό βαθμό η κάθε μία, διέκριναν μία ομαδοποίηση σε 2 cluster που προσπαθεί να προσεγγίσει την κατηγοριοποίηση με βάση την κατηγορική μεταβλητή Channel. Αντίθετα, καμία εκ των τριών μεθόδων δεν εντοπίζει την κατηγοριοποίηση με βάση το Region. Τα αποτελέσματα των k-means και model-based clustering είναι παρόμοια, με τις δύο μεθόδους να παρουσιάζουν ίδιας ισχύος δομή και τον model-based clustering να εμφανίζει καλύτερη γραμμική σχέση με την προϋπάρχουσα κατηγοριοποίηση. Με βάση λοιπόν τα παραπάνω στοιχεία, επιλέγουμε ως βέλτιστη μέθοδο ομαδοποίησης το Model-Based Clustering.