ΜΕΘΟΔΟΙ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Τμήμα Στατιστικής

Ονοματεπώνυμο: Ελένη Φουρτούνη
Τμήμα Φοίτησης: Πληροφορικής
Αριθμός μητρώου: 3180196
Ακαδημαΐκή Ηλεκτρονική Διεύθυνση: p3180196@aueb.gr
Ακαδημαϊκή περίοδος: Χειμερινό εξάμηνο 2021-2022

Προκειμένου να απαντήσουμε στα ερωτήματα της εκφώνησης, οφείλουμε πρώτα να αναλύσουμε τα δεδομένα μας και να εφαρμόσουμε μία μία τις μεθόδους συσταδοποίησης που μας ενδιαφέρουν να διερευνήσουμε(k-means clustering, hierarchical clustering, model-based clustering). Παρακάτω, παρουσιάζεται βηματικά η μεθοδολογία που ακολουθήθηκε και στο φάκελο της εργασίας μπορούμε να δούμε το script στο οποίο βασίσαμε τα αποτελέσματά μας.

Βήμα 1ο: Τακτοποίηση Αρχείων

- Μέσω της εντολής getwd(), ανακαλύπτουμε το working directory μας και δημιουργούμε σε αυτό τον φάκελο ergasia1-p3180196 με περιεχόμενα τα αρχεία market.csv και assignment1.R
- Εκτελούμε την εντολή: setwd("ergasia1-p3180196") για διευκόλυνσητης εκτέλεσης του script file.

Βήμα 2ο: Προετοιμασία δεδομένων

- Οπτικοποίηση μέρους των δεδομένων(5 πρώτες γραμμές πίνακα):

```
> head(data,5)
  Channel Region Fresh Milk Grocery Frozen Detergents_Paper Delicassen
1       2      3 12669 9656    7561    214             2674       1338
2       2      3  7057 9810    9568   1762             3293       1776
3       2      3  6353 8808    7684   2405             3516       7844
4       1      3 13265 1196    4221   6404              507       1788
5       2      3 22615 5410    7198   3915             1777       5185
```

- Αποφασίζουμε για προληπτικούς λόγους να αφαιρέσουμε της γραμμές στις οποίες εντοπίζονται NA. Ωστόσο, πράγματι, δεν εντοπίζουμε κάποια διαφορά στα δεδομένα μας.

- Αναζητούμε τον τύπο των δεδομένων:

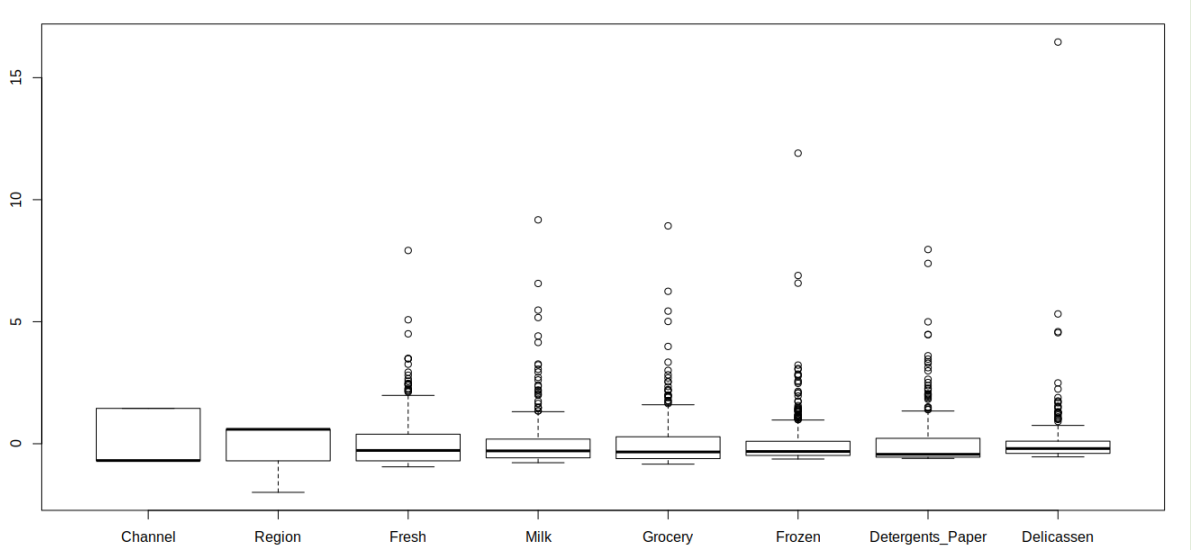
```
> str(data)
'data.frame':   440 obs. of  8 variables:
 $ Channel         : int  2 2 2 1 2 2 2 2 1 2 ...
 $ Region          : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Fresh           : int  12669 7057 6353 13265 22615 9413 12126 7579 5963 6006 ...
 $ Milk            : int  9656 9810 8808 1196 5410 8259 3199 4956 3648 11093 ...
 $ Grocery         : int  7561 9568 7684 4221 7198 5126 6975 9426 6192 18881 ...
 $ Frozen          : int  214 1762 2405 6404 3915 666 480 1669 425 1159 ...
 $ Detergents_Paper: int  2674 3293 3516 507 1777 1795 3140 3321 1716 7425 ...
 $ Delicassen      : int  1338 1776 7844 1788 5185 1451 545 2566 750 2098 ...
```

Συμπέρασμα: όλες οι ιδιότητες έχουν των ίδιο τύπο δεδομένων

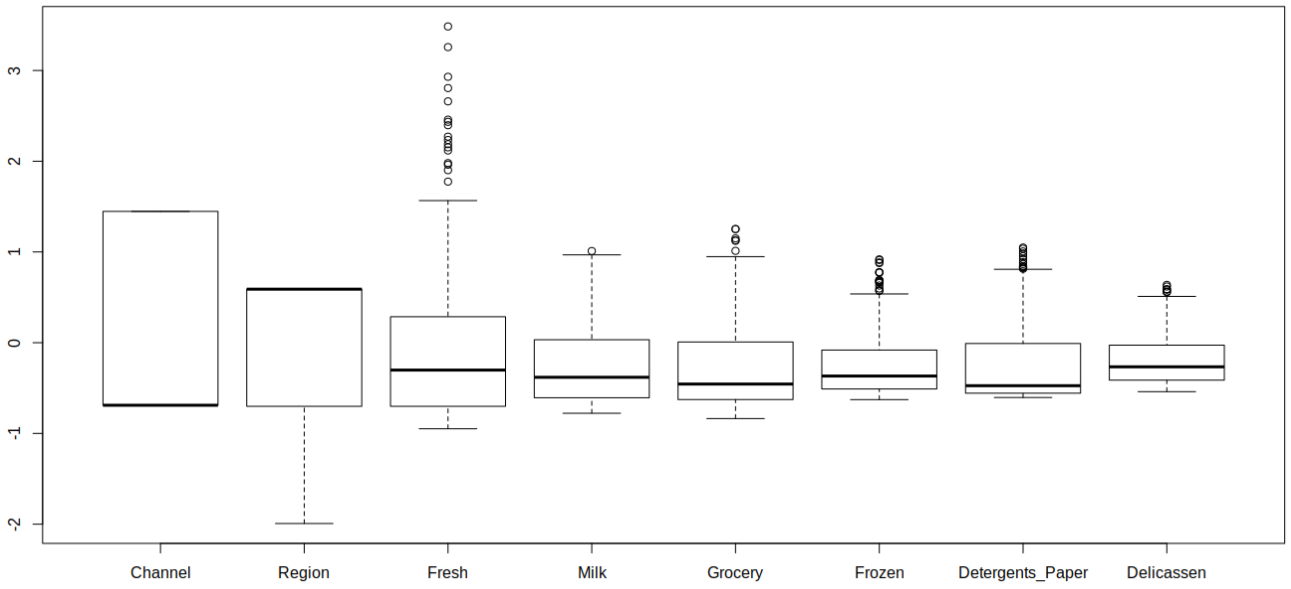
- Μελέτη βασικών μετρικών:

```
> summary(data)
    Channel          Region          Fresh             Milk       
 Min.   :1.000   Min.   :1.000   Min.   :     3   Min.   :   55  
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:  3128   1st Qu.: 1533  
 Median :1.000   Median :3.000   Median :  8504   Median : 3627  
 Mean   :1.323   Mean   :2.543   Mean   : 12000   Mean   : 5796  
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.: 16934   3rd Qu.: 7190  
 Max.   :2.000   Max.   :3.000   Max.   :112151   Max.   :73498  
    Grocery          Frozen        Detergents_Paper    Delicassen     
 Min.   :    3   Min.   :   25.0   Min.   :    3.0   Min.   :    3.0  
 1st Qu.: 2153   1st Qu.:  742.2   1st Qu.:  256.8   1st Qu.:  408.2  
 Median : 4756   Median : 1526.0   Median :  816.5   Median :  965.5  
 Mean   : 7951   Mean   : 3071.9   Mean   : 2881.5   Mean   : 1524.9  
 3rd Qu.:10656   3rd Qu.: 3554.2   3rd Qu.: 3922.0   3rd Qu.: 1820.2  
 Max.   :92780   Max.   :60869.0   Max.   :40827.0   Max.   :47943.0
```

Παρατηρούμε πως ενδέχεται να υπάρχουν outliers, τα οποία οφείλουμε να εξαλείψουμε, προκειμένου να έχουμε πιο “σωστά” αποτελέσματα στις μεθόδους συσταδοποίησης που θα ακολουθήσουμε και να διεξάγουμε πιο ακριβή συμπεράσματα. Πραγματοποιούμε scaling στα δεδομένα μας, προκειμένου να διαχειριζόμαστε μικρότερες τιμές δεδομένων και να είναι έτσι πιο ευκρινή τα διαγράμματά μας και δημιουργούμε boxplots προκειμένου να οπτικοποιήσουμε το πιθανό αυτό πρόβλημα:

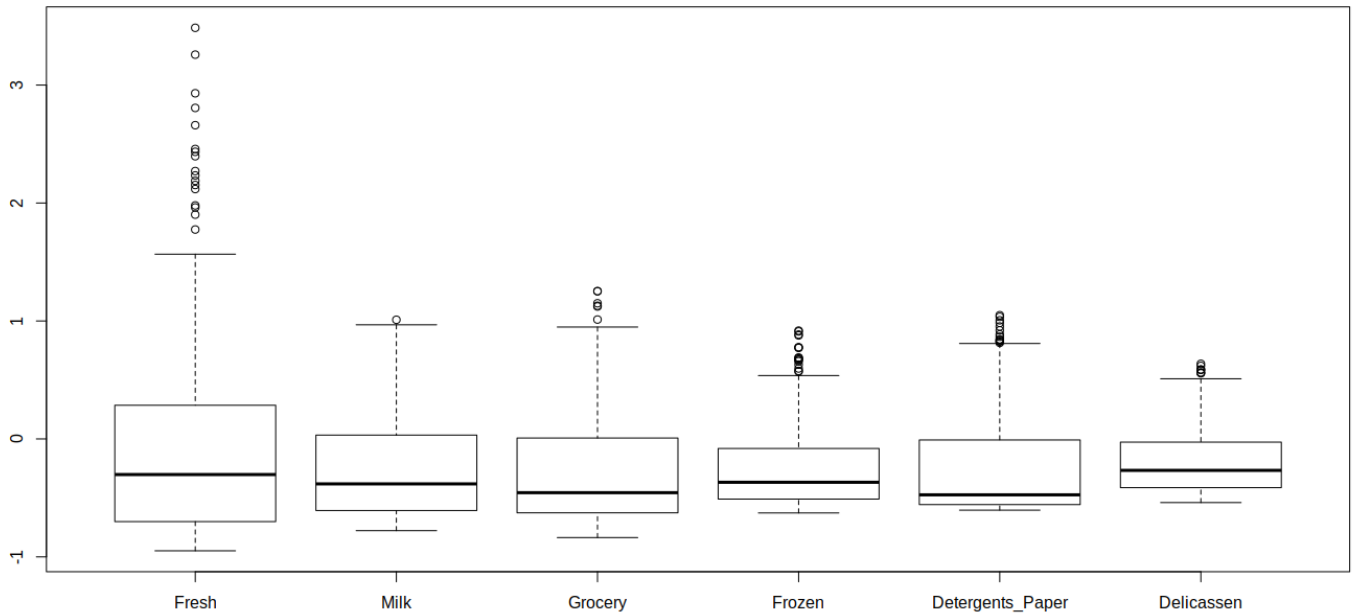


- Εξάλειψη ακραίων ατυπικών τιμών, βασιζόμενοι στην αρχή που λέει πως ένα outlier λαμβάνει τιμή μεγαλύτερη ή μικρότερη από το 150% του IQR



Πλέον, έχουμε ευκρινώς καθαρότερα διαγράμματα

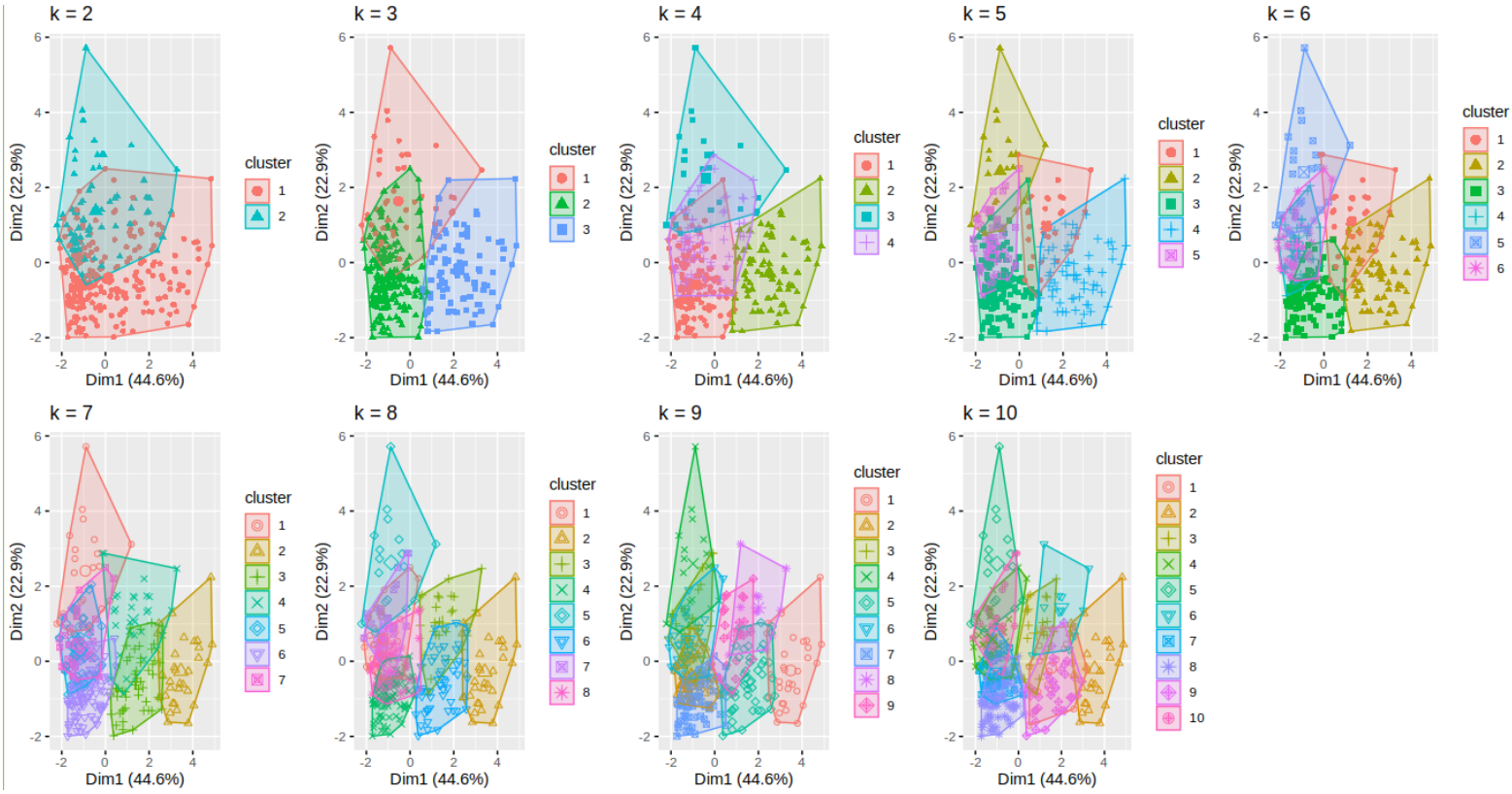
- Αποφασίζουμε να εξαλείψουμε τις στήλες Region, Channel, εφόσον δεν τις χρειαζόμαστε για την απάντηση των 4 πρώτων ερωτημάτων, αλλά επιθυμούμε να διατηρήσουμε την ως τώρα επεξεργασία των δεδομένων μας. Για το λόγο αυτό, αποθηκεύουμε το περιεχόμενο της μεταβλητής data, στη μεταβλητή market.



Είμαστε έτοιμοι να προχωρήσουμε στην εφαρμογή μεθόδων συσταδοποίησης

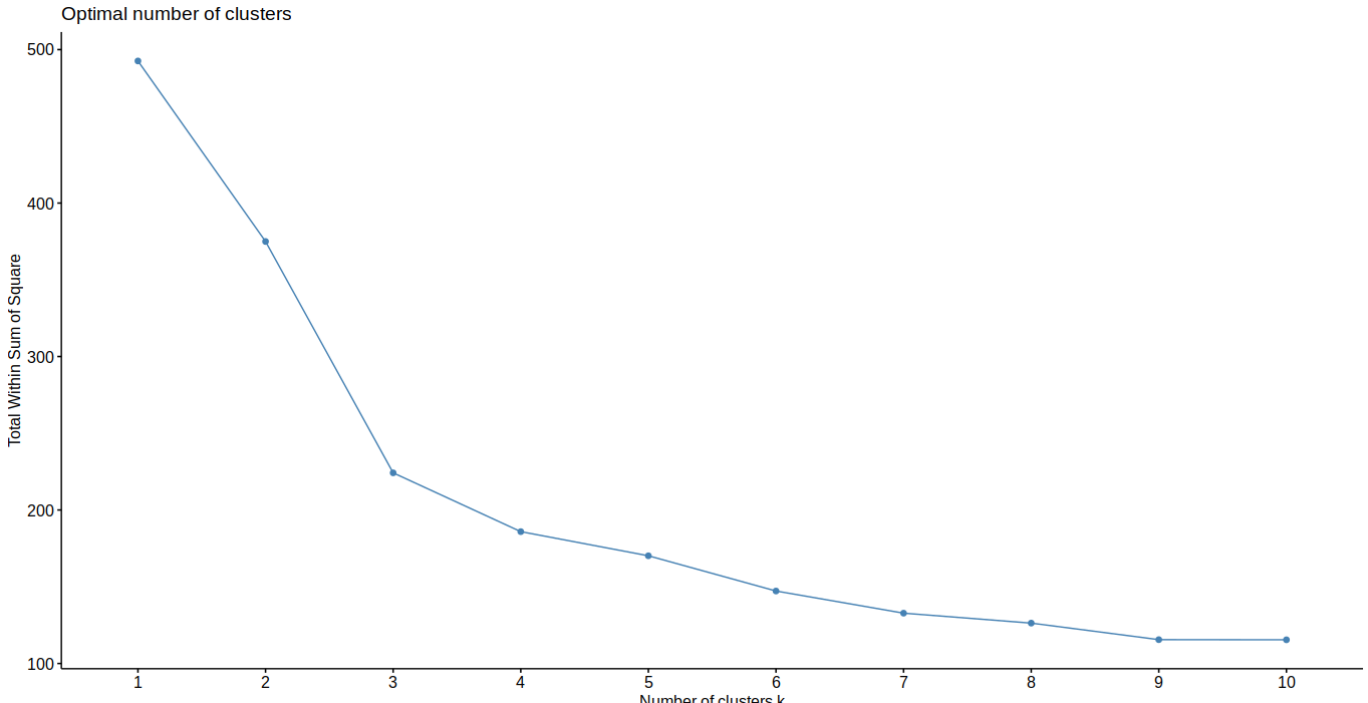
<u>Βήμα 3ο</u>: K-means Clustering

- Οπτικοποίηση του αποτελέσματος της εφαρμογής του αλγορίθμου k-means clustering για k=2,3,4,5,6,7,8,9,10



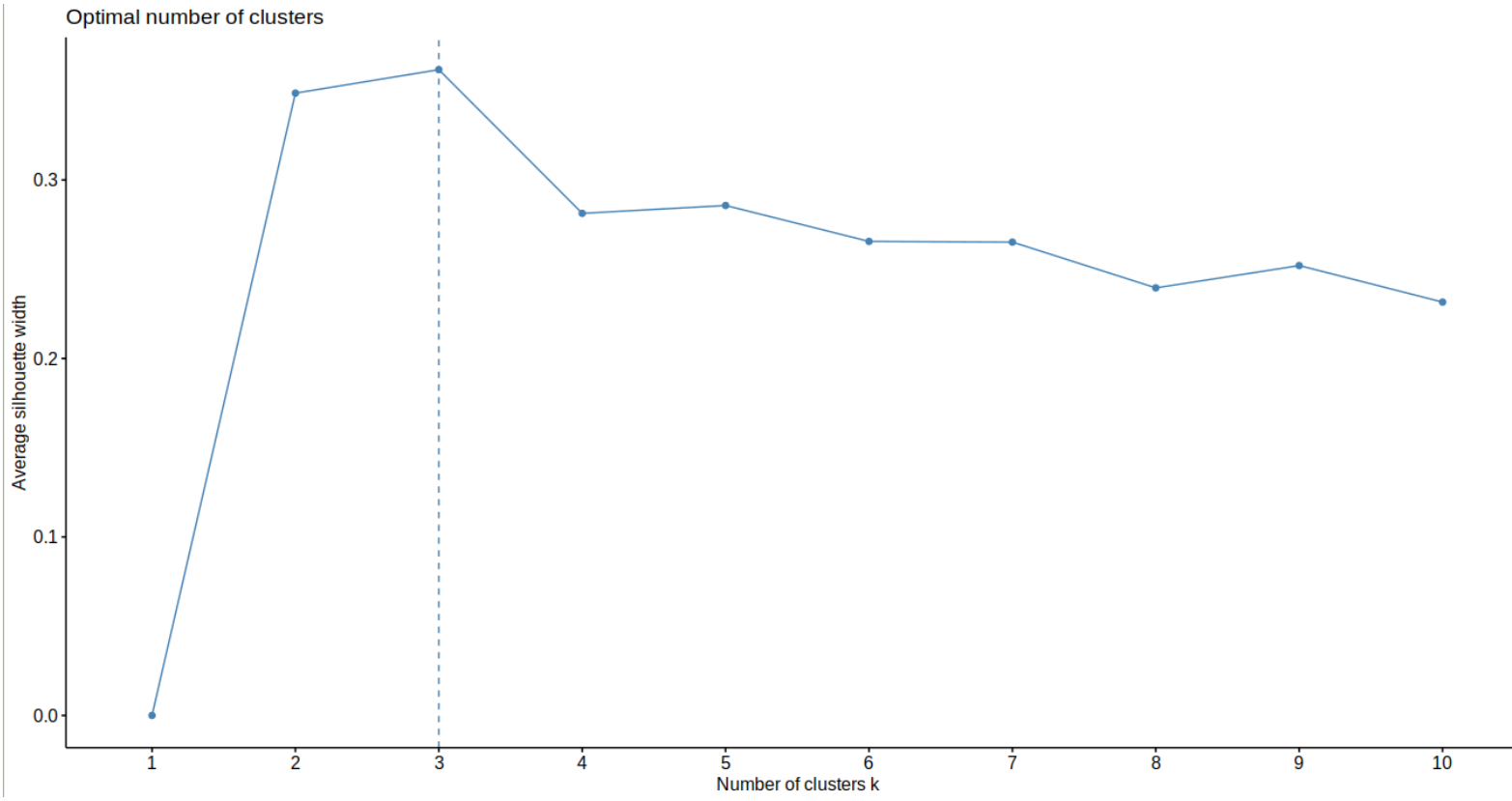
Παρατηρούμε πως για k=3, τα όρια των συστάδων φαίνονται καλύτερα διαχωρισμένα. Πρόκειται λοιπόν, να διερευνήσουμε περαιτέρω αυτή την παρατήρηση

- Μέθοδος Elbow:



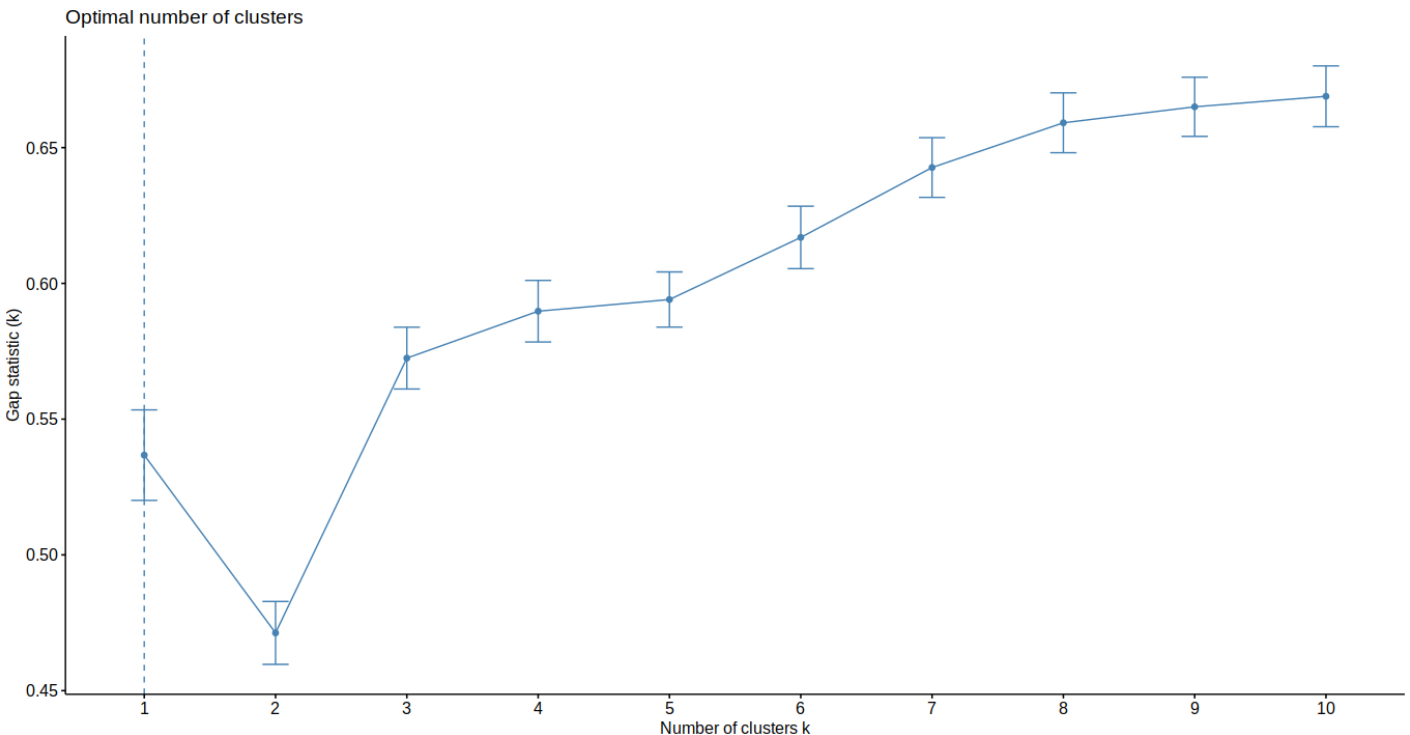
Φαίνεται πως έχουμε τα καλύτερα αποτελέσματα για k=3 ή k=4, όπου διακρίνουμε τη μεγαλύτερη καμπή της καμπύλης.

- Μέθοδος Silhouette:



Εδώ φαίνεται επίσης πως για k=3 έχουμε την καλύτερη συσταδοποίηση των δεδομένων μας, καθώς εκεί βρίσκεται το μέγιστο της γραφικής παράστασης μετρήσεων.

- Μέθοδος Gap Statistics:



```
--> Number of clusters (method 'firstmax'): 1
          logW   E.logW       gap     SE.sim
[1,] 4.849501 5.386252 0.5367509 0.01666661
[2,] 4.675567 5.146829 0.4712615 0.01158581
[3,] 4.480695 5.053175 0.5724801 0.01138252
[4,] 4.386403 4.976160 0.5897574 0.01133789
[5,] 4.327801 4.921855 0.5940539 0.01015942
[6,] 4.263601 4.880527 0.6169269 0.01150337
[7,] 4.209675 4.852356 0.6426812 0.01099592
[8,] 4.168256 4.827409 0.6591532 0.01102484
[9,] 4.139445 4.804483 0.6650377 0.01088659
[10,] 4.114875 4.783780 0.6689043 0.01120784
```

Εδώ δεν είναι ξεκάθαρο το βέλτιστο k, καθώς φαίνεται πως από το ελάχιστο ακρότατό της και μετά, η συνάρτηση αυξάνεται διαρκώς γραμμικά, οπότε δεν μπορούμε να αποφανθούμε για τη μέγιστη τιμή της

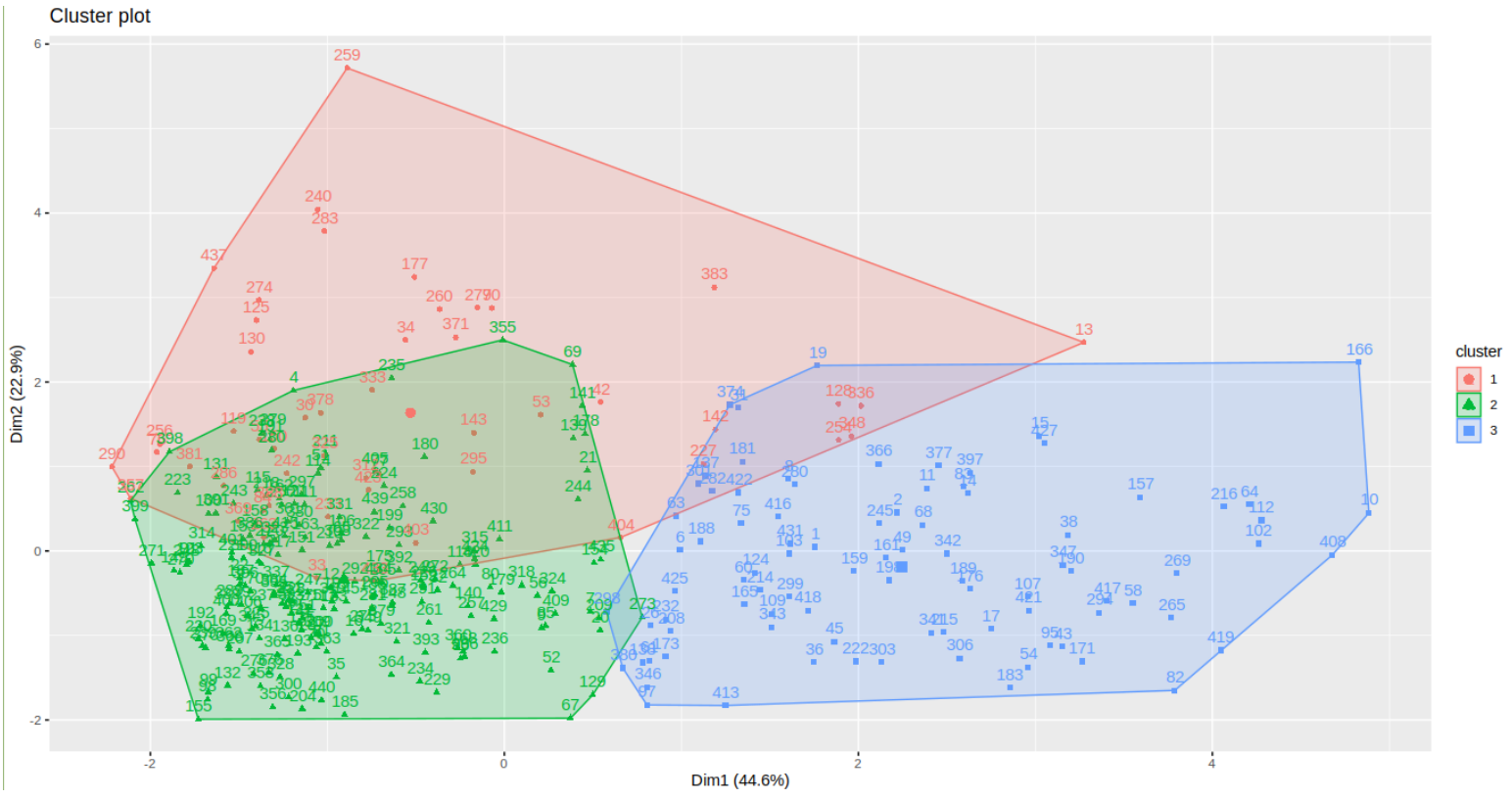
- Επιλογή k

  k=3 είναι ο αριθμός των συστάδων που θα επιλέξουμε, εφόσον αποτελεί κοινή τομή 3 διαφορετικών παρατηρήσεων.

```
> final$tot.withinss
[1] 224.1996
```

```
> final$size
[1]  50 191  89
```

```
Within cluster sum of squares by cluster:
[1] 58.23648 95.87846 70.08464
 (between_SS / total_SS =  54.5 %)
```



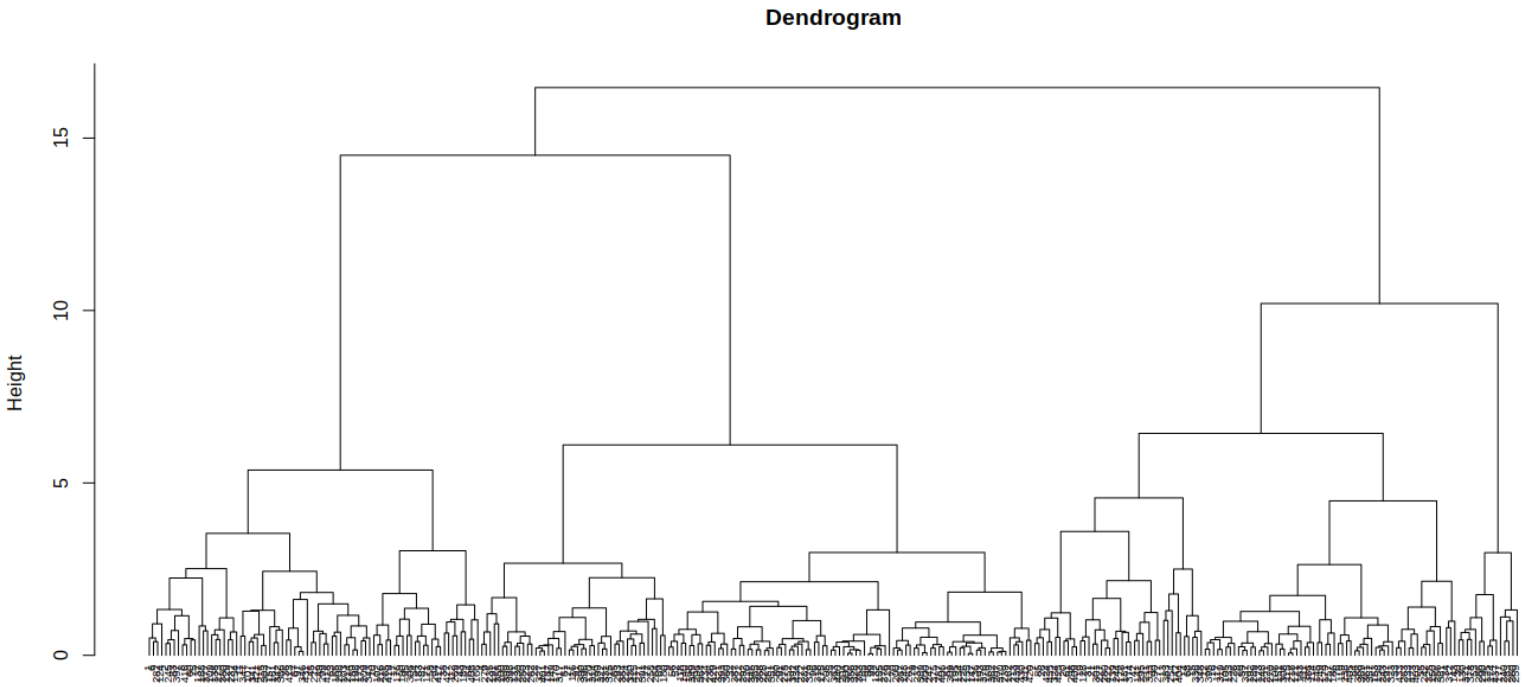
Βήμα 4ο: Hierarchical Clustering

- Επιλογή μεθόδου διασύνδεσης

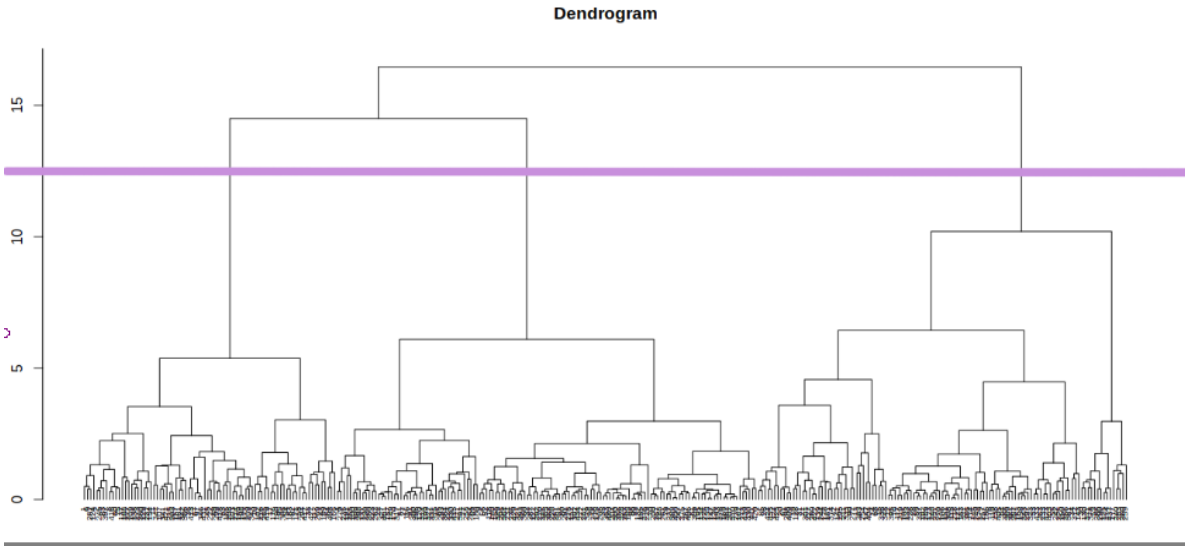
```
> sapply(m, ac)
  average    single  complete      ward
0.8494783 0.6981633 0.9269808 0.9772293
```

Θα προτιμήσουμε τη μέθοδο διασύνδεσης(ward), καθώς είναι πιο αποδοτική, βάσει των παραπάνω τιμών.

- Απεικόνιση δεδομένων, έπειτα από την εφαρμογή της μεθόδου ward



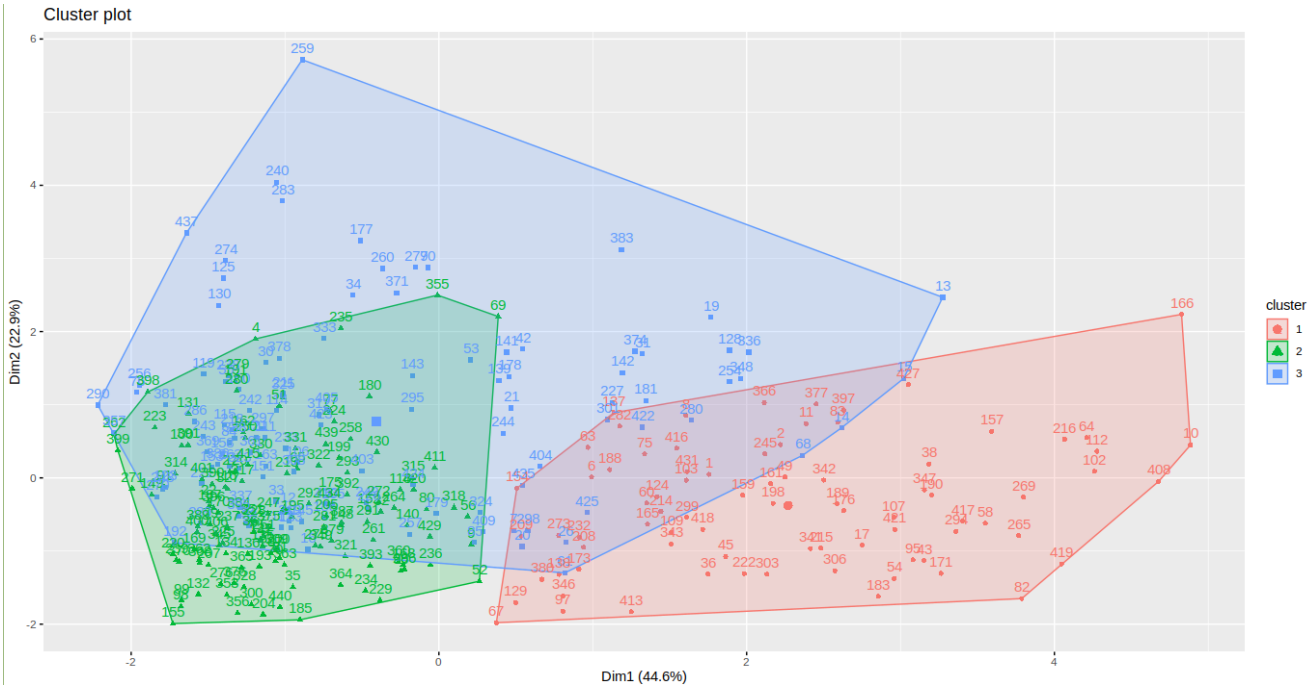
Παρατηρούμε ότι αν τραβήξουμε μια οριζόντια ως προς τον άξονα των x γραμμή,που να περνά από τη μεγαλύτερη κάθετη ευθεία, χωρίς παρεμβολές από οριζόντιες γραμμές, τέμνει το δενδοδιάγραμμά μας σε 3 σημεία(mauve γραμμή του παρακάτω διαγράμματος)



- Εφαρμογή της μεθόδου στα αρχικά μας δεδομένα:

```
> table(groups)
groups
  1   2   3
 80 133 117
```

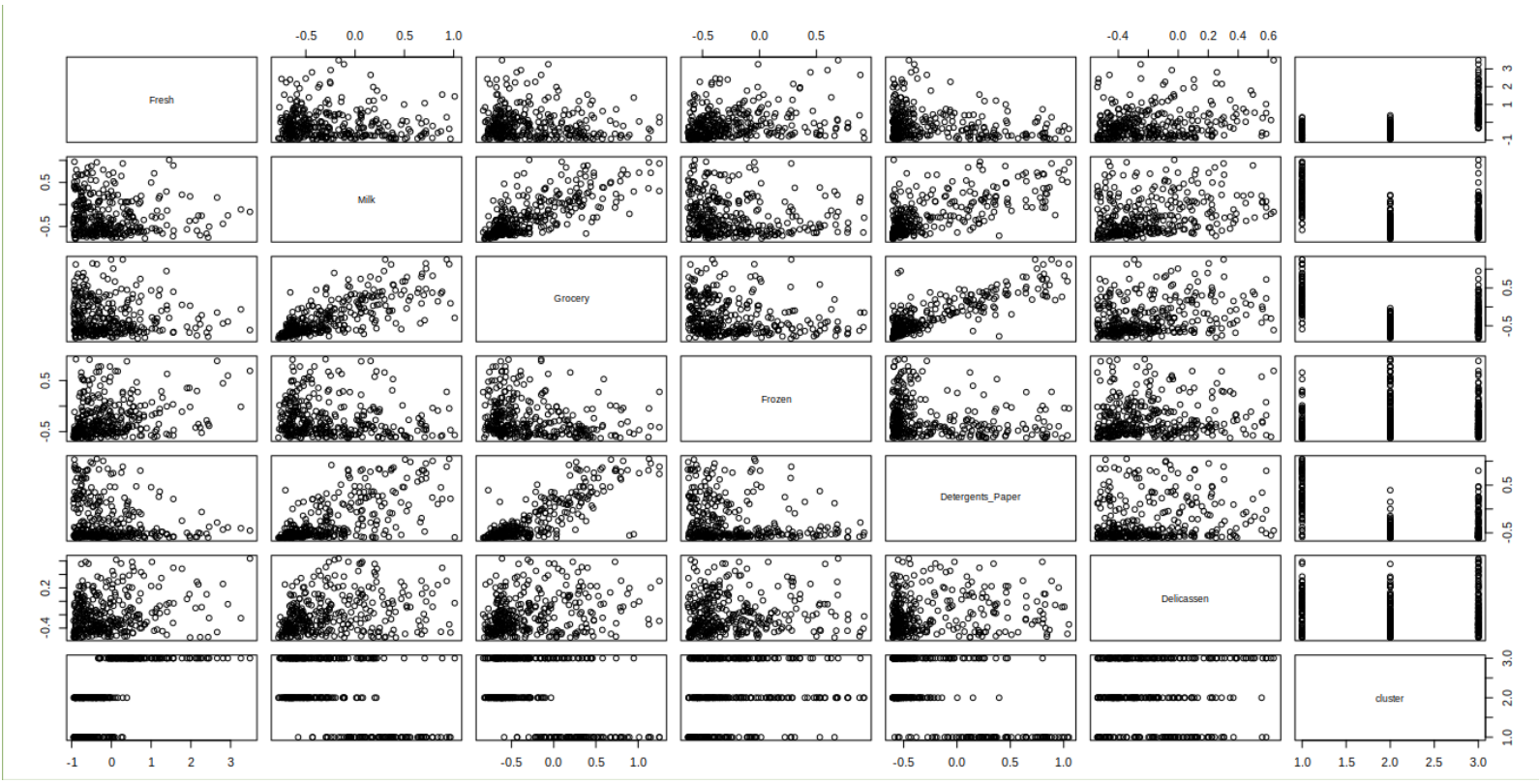
Πλήθος στοιχείων ανά cluster



Γραφική Αναπαράσταση των clusters

```
head(final_data)
       Fresh       Milk     Grocery     Frozen Detergents_Paper  Delicassen
 0.052873004  0.5229725 -0.04106815 -0.5886970      -0.04351919 -0.06626363
-0.390857056  0.5438386  0.17012470 -0.2698290       0.08630859  0.08904969
 0.099997579 -0.6233104 -0.39253008  0.6863630      -0.49802132  0.09330484
-0.204572662  0.3336868 -0.29729863 -0.4955909      -0.22787885 -0.02619421
 0.009939037 -0.3519151 -0.10273183 -0.5339045       0.05421869 -0.34745874
-0.349583519 -0.1138514  0.15518231 -0.2889858       0.09218126  0.36918101
cluster
       1
       1
       2
       1
       3
       1
```

Τα πρώτα 6 περιεχόμενα του πίνακα final_data



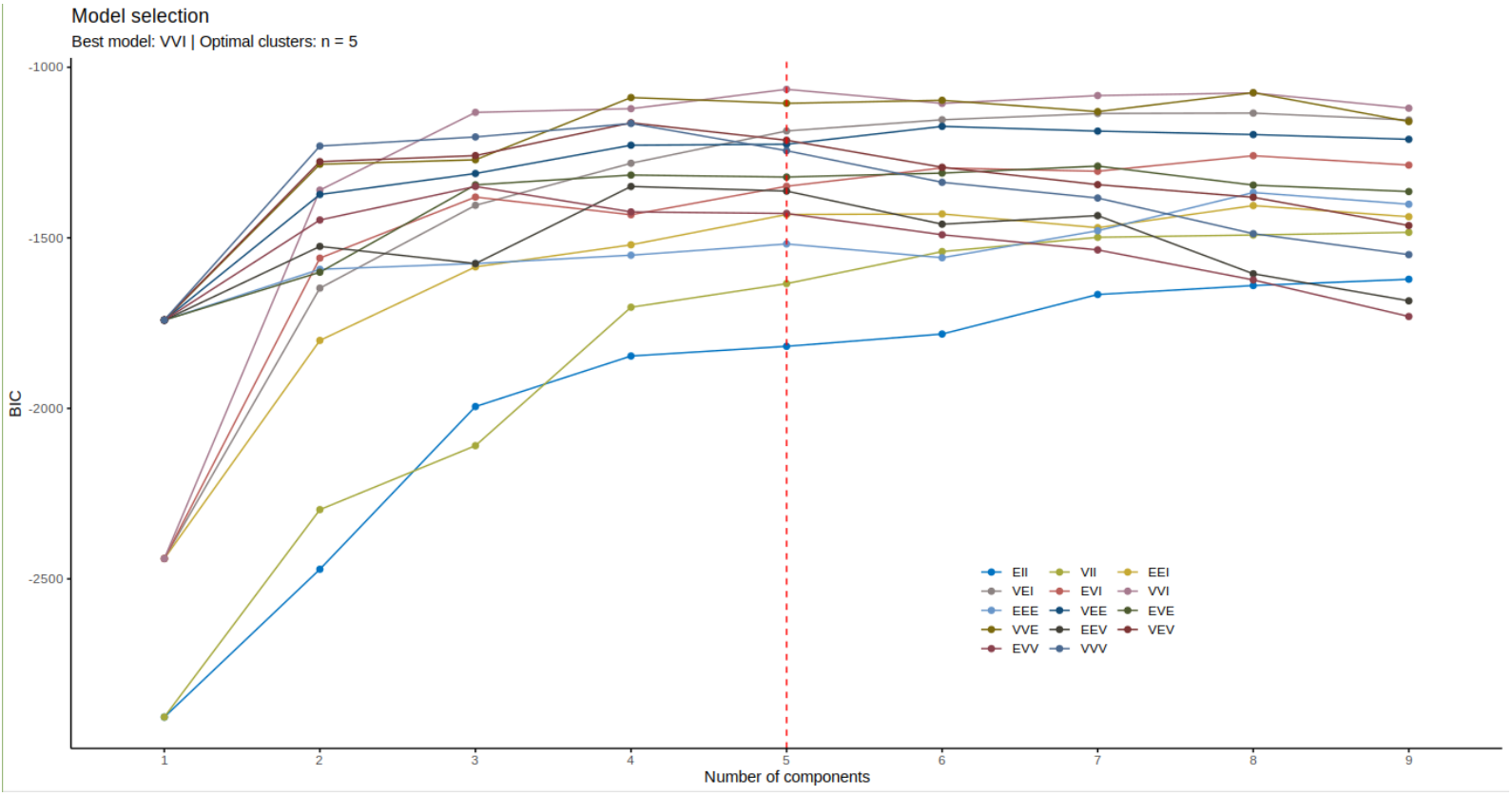
Απεικόνιση των clusters ανά συνδυασμό κατηγοριών

Βήμα 5ο: Model-based Clustering

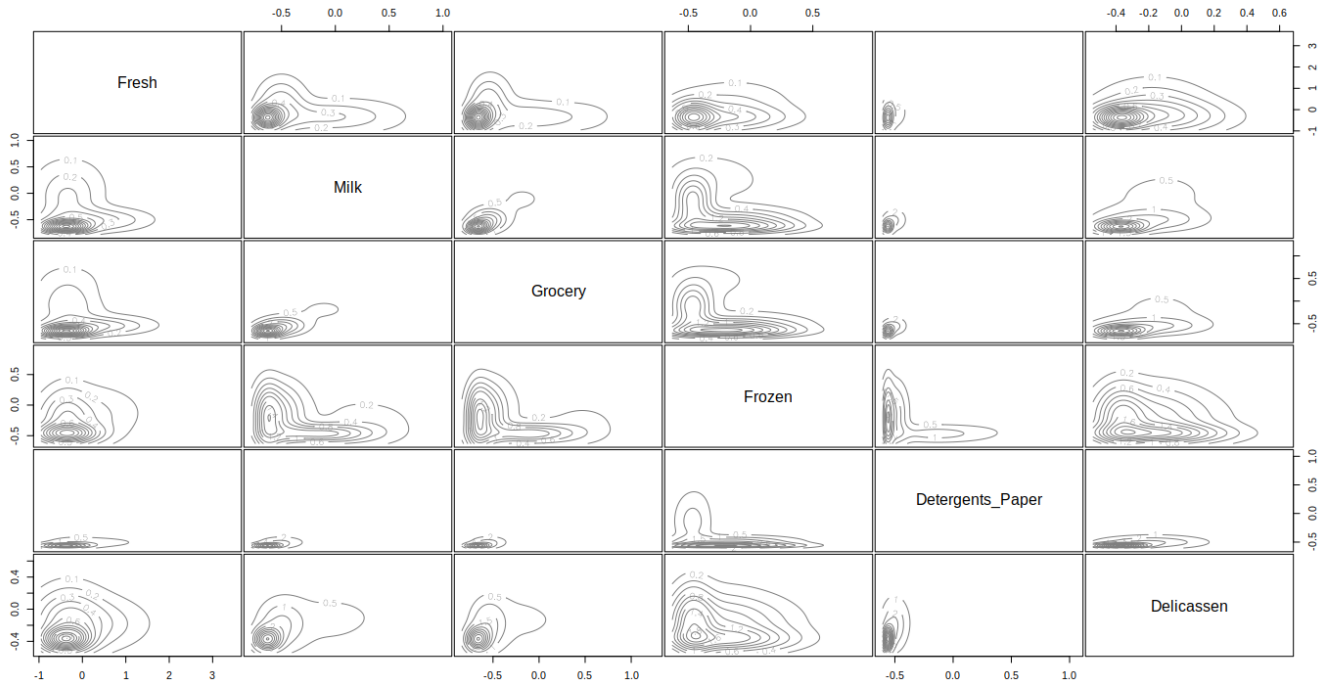
- Επιλογή μεθόδου:

```
> summary(mbc)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVI (diagonal, varying volume and shape) model with 5 components:

 log-likelihood   n df       BIC       ICL
      -346.7021 330 64 -1064.546 -1138.077

Clustering table:
 1  2  3  4  5
77 28 56 98 71
```

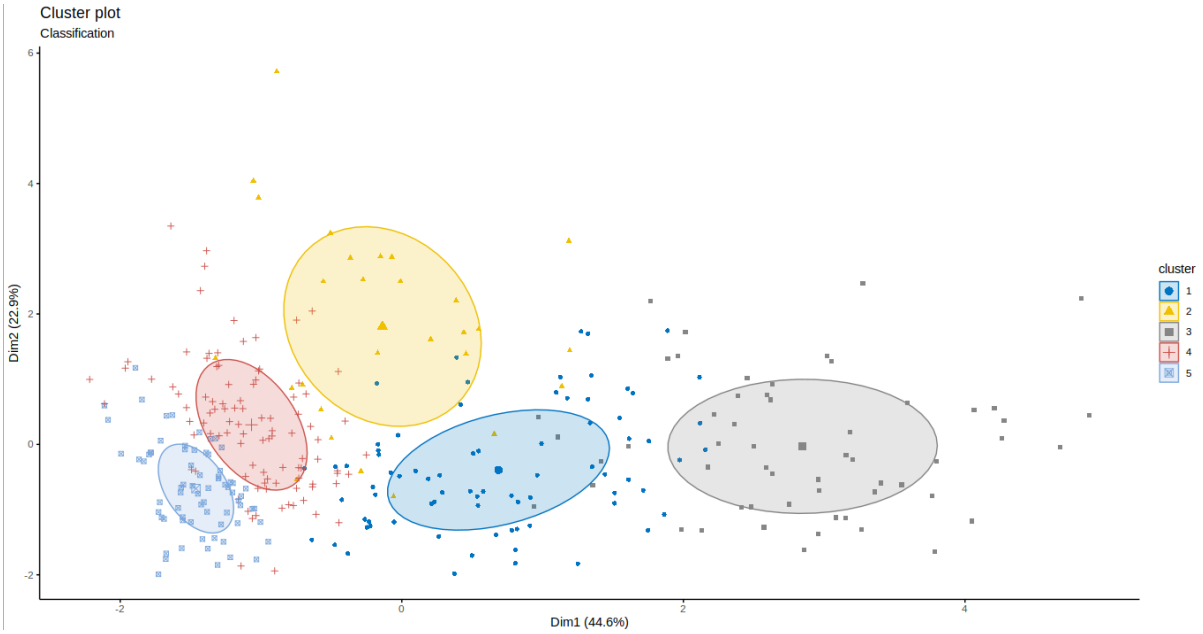
Παρατηρούμε ότι επιλέγεται η μέθοδος VVI με κατηγοριοποίηση σε 5 ομάδες, όπως φαίνεται και στο παρακάτω διάγραμμα:



- Αναπαράσταση της τελικής μορφής των δεδομένων:



<u>Βάσει density ανά συνδυασμό κατηγοριών</u>



Όλα τα clusters σε ένα κοινό διάγραμμα

Βήμα 6ο: Ερώτημα I

Βάσει των τελικών διαγραμμάτων, είναι προτιμότερη η μέθοδος συσταδοποίησης των δεδομένων μας είναι η εφαρμογή της μεθόδου hierarchical clustering με το διαχωρισμό σε 3 clusters, καθώς έτσι, κάθε cluster έχει παρόμοιο μέγεθος (σε αντίθεση με τις ομάδες που χρησιμοποιήθηκαν κατά την εφαρμογή του αλγορίθμου k-means clustering, οι οποίες δεν έχουν τόσο ομοιόμορφο όγκο δεδομένων)

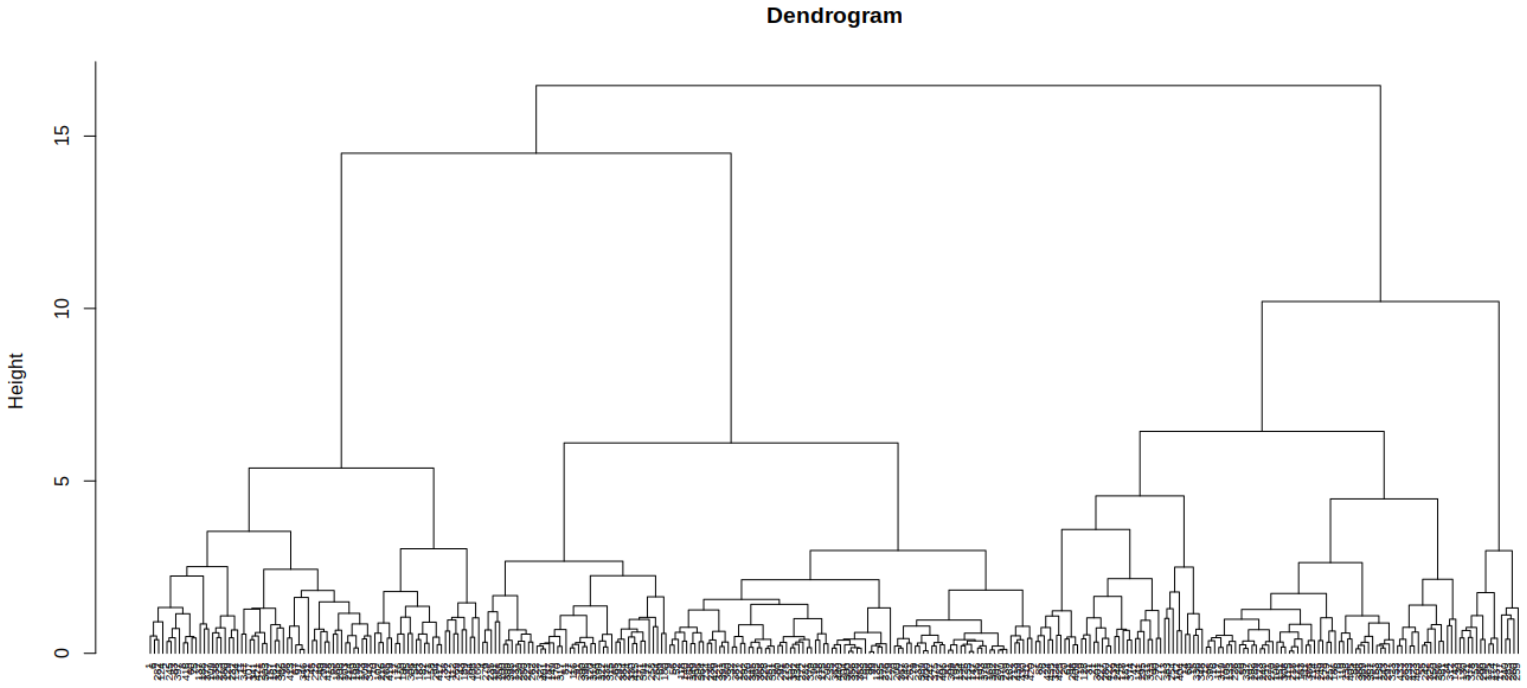
```
> final$size
[1]  50 191  89
```

k-means clustering

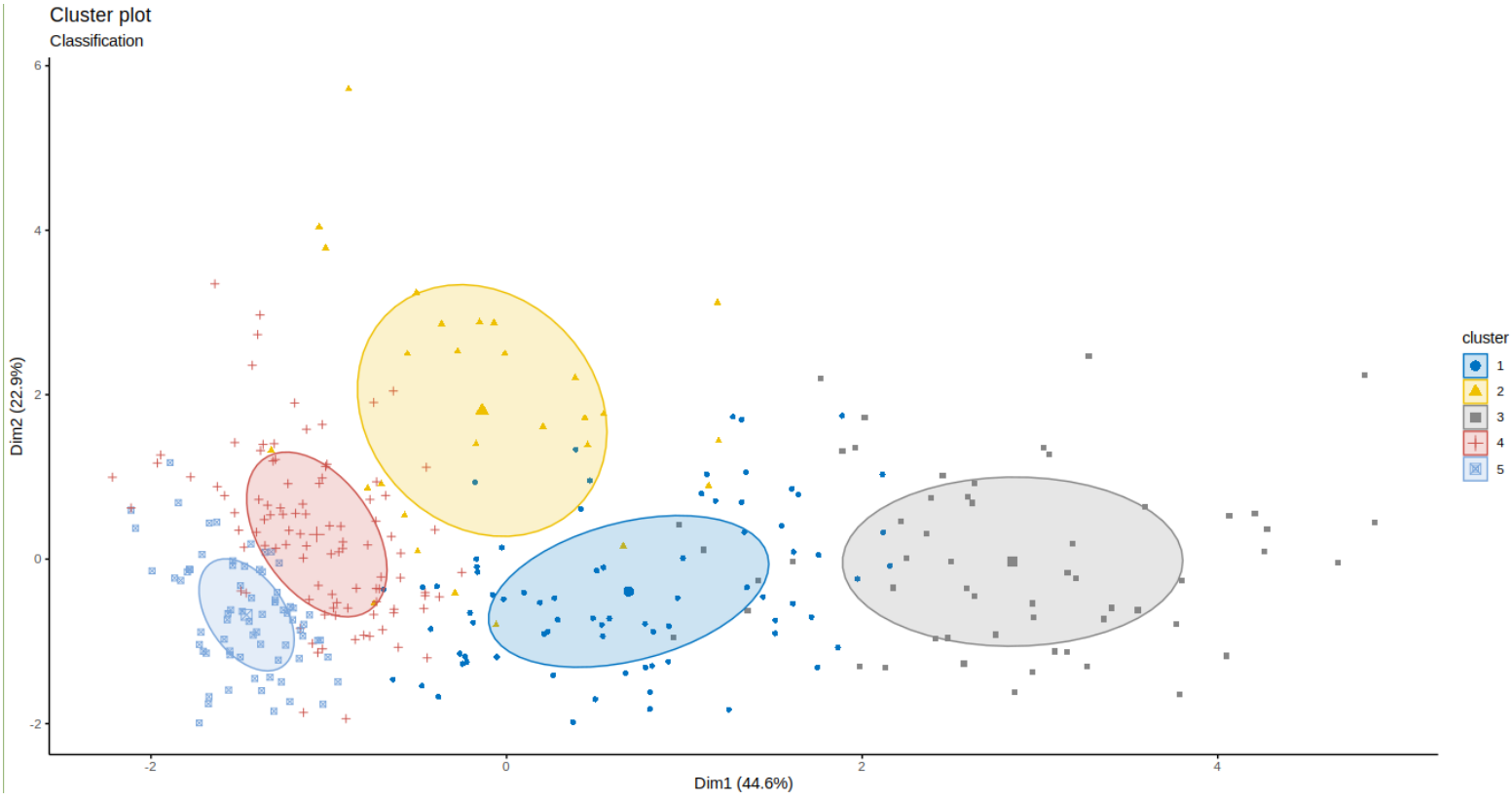
```
> table(groups)
groups
  1   2   3
 80 133 117
```

hierarchical clustering

Επίσης, κάθε περίπτωση λαμβάνει θέση στο εσωτερικό μιας ομάδας(κάτι που βλέπουμε ότι δε συμβαίνει στο τελευταίο διάγραμμα της ανάλυσης του αλγορίθμου model-based clustering).



hierarchical clustering



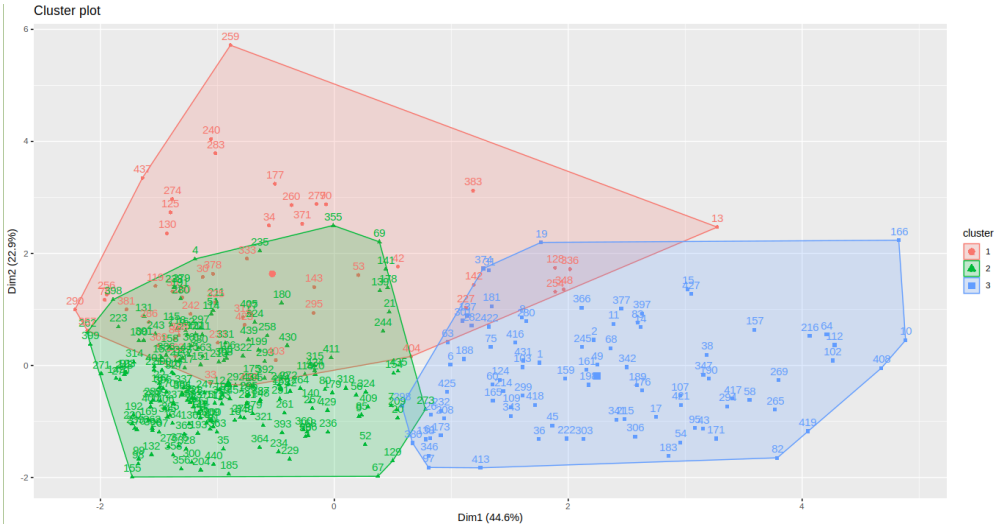
model-based clustering

<u>Βήμα 7ο</u>: Ερώτημα II

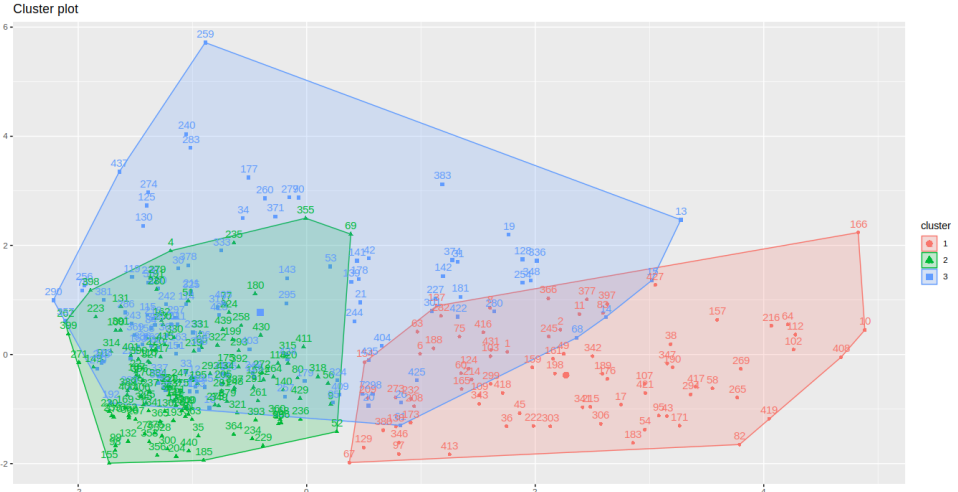
Παρατηρούμε ότι τα δεδομένα μας δεν είναι χωρισμένα σε πλήρως ομοιόμορφα clusters, καθώς βλέπουμε πως όλα τα ιεραρχικά ισάξια clusters δε βρίσκονται στο ίδιο επίπεδο(δεν έχουν την ίδια τιμή στον άξονα των y). Ωστόσο, ικανοποιούν αυτό το χαρακτηριστικό σε ένα μεγάλο βαθμό. Συνεπώς, μπορούμε να συμπεράνουμε πως είναι αρκετά επιτυχημένη η συνολική ομαδοποίηση στην οποία καταλήξαμε, αλλά όχι τέλεια.

<u>Βήμα 8ο</u>: Ερώτημα III

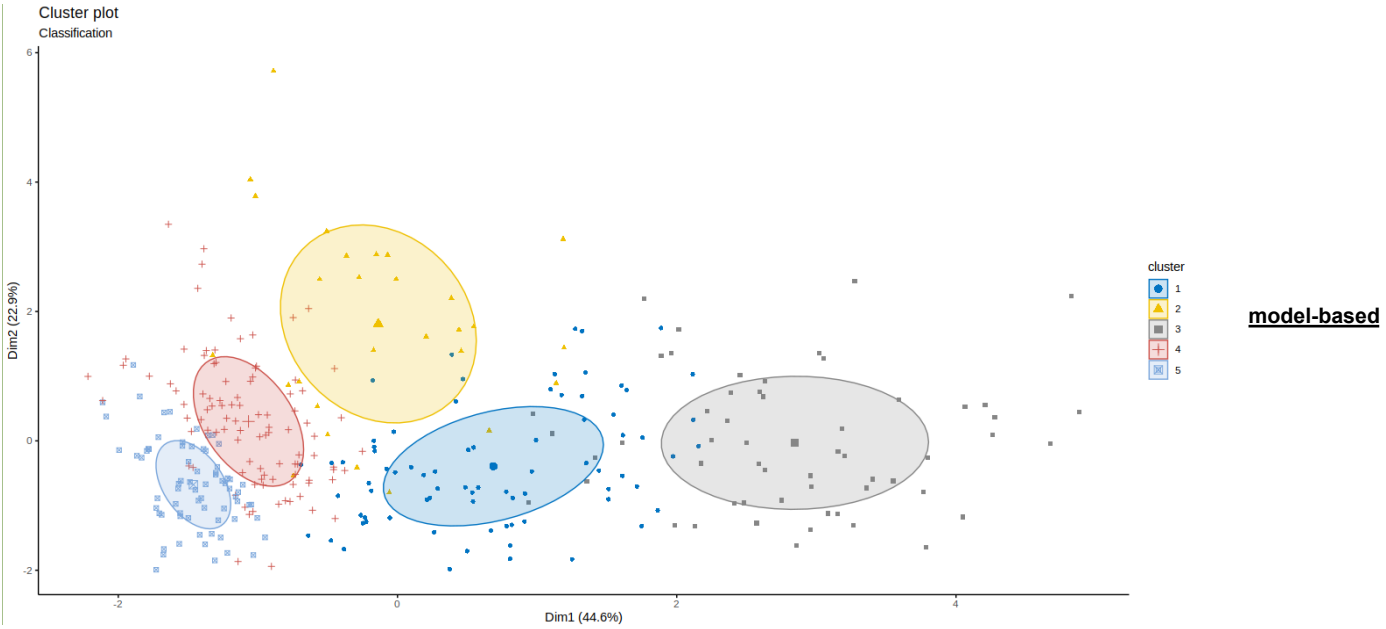
Όπως βλέπουμε και στα παρακάτω διαγράμματα, οι ομαδοποιήσεις των αλγορίθμων k-means και hierarchical clustering μοιάζουν πολύ. Από την άλλη πλευρά, εκείνη που παράγεται από τον αλγόριθμο model-based έχει ως αποτέλεσμα πιο ευδιάκριτα clusters που δε μοιάζουν με εκείνα των δύο προαναφερόμενων μεθόδων.
Συνεπώς, στο ερώτημα αυτό η απάντηση είναι πως δεν υπάρχουν ομάδες που να αναγνωρίζονται με σαφήνεια από όλους τους αλγορίθμους.



<u>k-means</u>



<u>hierarchical</u>



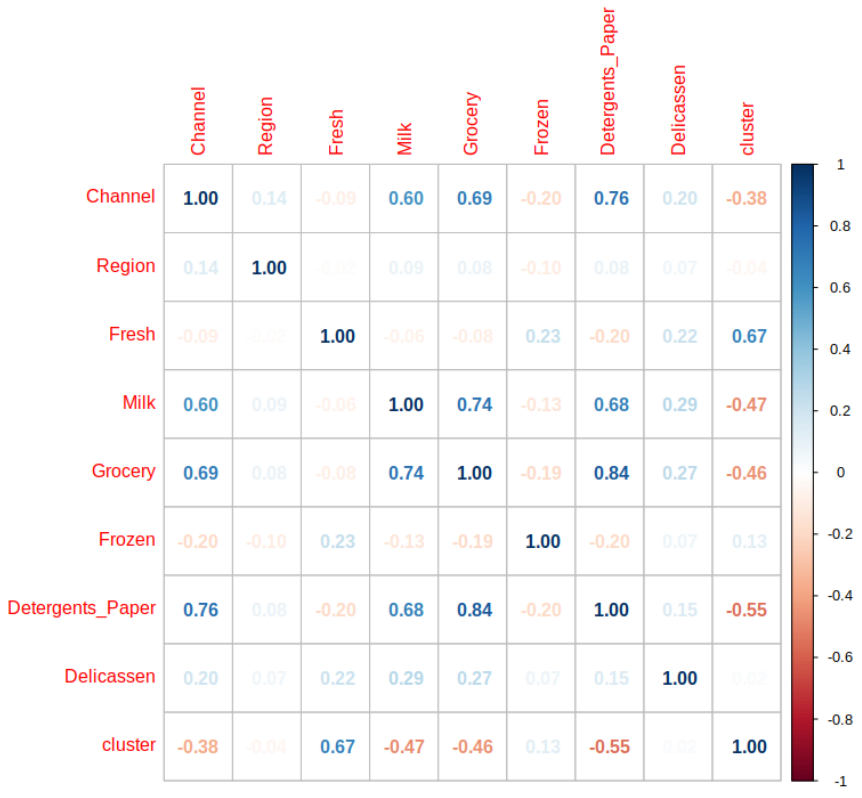
model-based

Βήμα 9ο: Ερώτημα IV

Όπως προκύπτει και από την απάντηση του ερωτήματος III, ξεκάθαρα clusters αντικρίζουμε στον αλγόριθμο model-based. Όσον αφορά τον αλγόριθμο k-means, κανένα cluster δεν είναι ευδιάκριτο, ενώ στον αλγόριθμο hierarchical clustering που επιλέξαμε, διακρίνουμε σαφή διαχωρισμό μεταξύ των clusters 1,2, ενώ τα όρια του cluster 3, δεν είναι τόσο ευδιάκριτα, καθώς το τελευταίο τέμνεται με τα άλλα δύο clusters.

Βήμα 10ο: Ερώτημα V

Θα μελετήσουμε το συντελεστή συσχέτισης μεταξύ των ιδιοτήτων του πίνακα:



Βασμένοι στο παραπάνω διάγραμμα, συμπεραίνουμε πως η μεταβλητή **Channel** εμφανίζει:

1. Ασθενή θετική συσχέτιση με τις μεταβλητές Milk, Grocery, Delicassen
2. Ασθενή και μάλιστα σχεδόν μηδενική συσχέτιση με όλες τις υπόλοιπες μεταβλητές-ιδιότητες(πέραν της ίδιας)

Region εμφανίζει:

1. Ασθενή και μάλιστα σχεδόν μηδενική συσχέτιση με όλες τις μεταβλητές-ιδιότητες(πέραν της ίδιας)

Fresh εμφανίζει:

1. Ασθενή θετική συσχέτιση με τις μεταβλητές Frozen, Delicassen
2. Ασθενή αρνητική συσχέτιση με τη μεταβλητή Detergents_Paper
3. Ασθενή και μάλιστα σχεδόν μηδενική συσχέτιση με όλες τις μεταβλητές-ιδιότητες(πέραν της ίδιας)

Milk εμφανίζει:

1. Μέτρια θετική συσχέτιση με τις μεταβλητές Channel, Grocery, Detergents_Paper
2. Ασθενή συσχέτιση με τη μεταβλητή Delicassen
3. Ασθενή και μάλιστα σχεδόν μηδενική συσχέτιση με όλες τις υπόλοιπες μεταβλητές-ιδιότητες(πέραν της ίδιας)

Grocery εμφανίζει:

1. Ισχυρή θετική συσχέτιση με τη μεταβλητή Detergents_Paper
2. Μέτρια θετική συσχέτιση με τις μεταβλητές Channel, Milk
3. Ασθενή θετική συσχέτιση με τη μεταβλητή Delicassen
4. Ασθενή αρνητική συσχέτιση με τη μεταβλητή Frozen
5. Ασθενή και μάλιστα σχεδόν μηδενική συσχέτιση με όλες τις υπόλοιπες μεταβλητές-ιδιότητες(πέραν της ίδιας)

Frozen εμφανίζει:

1. Ασθενή θετική συσχέτιση με τη μεταβλητή Fresh
2. Ασθενή αρνητική συσχέτιση με τις μεταβλητές Channel, Grocery, Detergents_Paper
3. Ασθενή και μάλιστα σχεδόν μηδενική συσχέτιση με όλες τις υπόλοιπες μεταβλητές-ιδιότητες(πέραν της ίδιας)

Detergents_Paper εμφανίζει:

1. Ισχυρή θετική συσχέτιση με τη μεταβλητή Grocery
2. Μέτρια θετική συσχέτιση με τη μεταβλητή Milk
3. Ασθενή αρνητική συσχέτιση με τις μεταβλητές Frozen, Fresh

4. Ασθενή και μάλιστα σχεδόν μηδενική συσχέτιση με όλες τις υπόλοιπες μεταβλητές-ιδιότητες(πέραν της ίδιας)

Delicassen εμφανίζει:

1. Μέτρια θετική συσχέτιση με τις μεταβλητές Channel, Fresh, Milk, Grocery
2. Ασθενή και μάλιστα σχεδόν μηδενική συσχέτιση με όλες τις υπόλοιπες μεταβλητές-ιδιότητες(πέραν της ίδιας)

Βήμα 11ο: Κατακλείδα

Έπειτα από την επεξεργασία των δεδομένων και την απάντηση των παραπάνω ερωτημάτων, μπορούμε να διεξάγουμε ορισμένα κρίσιμα συμπεράσματα. Αρχικά, αξίζει να αναφέρουμε πως τα αποτελέσματα του κάθε αλγορίθμου unsupervised learning clustering επηρεάζονται από το εύρος τιμών των ιδιοτήτων που εξετάζουμε, τον αλγόριθμο που εφαρμόζουμε, καθώς και τις μετρικές που δίνουμε ως input στον κάθε αλγόριθμο. Έτσι, διακρίνουμε διαφορετικά αποτελέσματα τόσο ως αποτέλεσμα της εφαρμογής διαφορετικού αλγορίθμου, όσο και ως αποτέλεσμα χρήσης διαφορετικών μετρικών στον ίδιο αλγόριθμο.

Επίσης, έπειτα από την ανάλυση της συσχέτισης, μεταξύ των τελικών μας δεδομένων, συμπεραίνουμε πως αιτιολογείται η μορφή και η πληθικότητα των clusters που δημιουργήθηκαν.