
**Ακαδημαϊκό έτος: 2021-2022**

# Μάθημα: Μέθοδοι Στατιστικής και Μηχανικής Μάθησης

## 1η εργασία

**Ονομ/πώνυμο: Βασίλειος-Ηλίας Δρούζας**

**ΑΜ: p3180051**

Σημείωση: Στην εργασία αυτή πρώτα έχω υπλοποιήσει τους αλγορίθμους και στο τέλος έχω προσθέσει τις απαντήσεις μου στις ερωτήσεις. Ο κώδικας των μεθόδων που εφάρμοσα βρίσκεται στο **.txt** αρχείο και περιέχει όλες τις εντολές κατά σειρά που έτρεξα για να παράξω αυτή τη παρουσίαση.

Πριν αφαιρέσω τις κατηγορικές μεταβλητές Channel,Region μπορώ να εξετάσω τα δεδομένα μου:

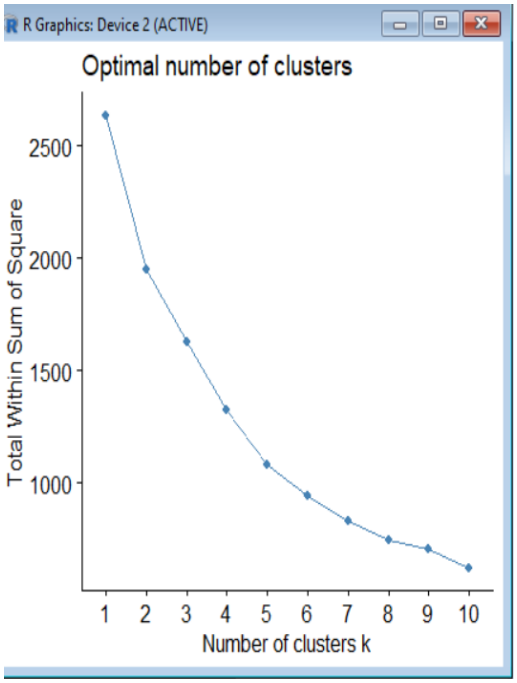
```
> summary(data)
    Channel          Region          Fresh             Milk
 Min.   :1.000   Min.   :1.000   Min.   :     3   Min.   :   55
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:  3128   1st Qu.: 1533
 Median :1.000   Median :3.000   Median :  8504   Median : 3627
 Mean   :1.323   Mean   :2.543   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
```

Ένα γρρήγορο συμπέρασμα που βγάζουμε είναι ότι οι τιμές min και max για όλες τις μεταβλητές πλην των κατηγορικών παρουσιάζουν μεγάλη διαφορά. Αυτό πρακτικά σημαίνει ότι υπάρχουν πελάτες που ξοδεύουν λίγα και πελάτες που ξοδεύουν πολλά.

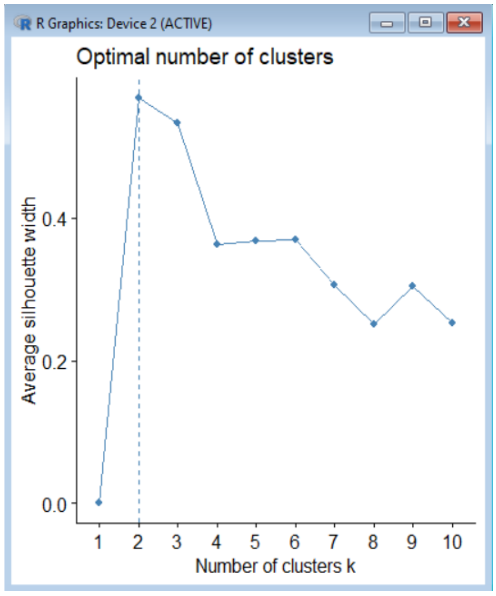
A. K-means

Η πρώτη μας δουλειά στην μέθοδο K-means είναι να διαλέξουμε τον βέλτιστο αριθμό ομάδων για να χωρίσουμε τα δεδομένα μας. Τρεις μεθόδους θα χρησιμοποιήσουμε για να καταλήξουμε: Elbow, Silhouette, Gap statistic.

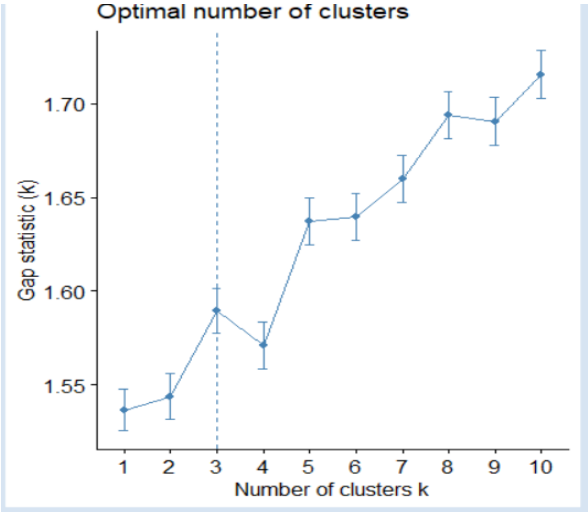
Μέθοδος Elbow->2:




Μέθοδος Silhouette->2:



Μέθοδος Gap statistic->3:




Δεν είμαστε ακόμα τόσο σίγουροι για το αν θα ήταν καλύτερο να χωρίσουμε σε 2 ή 3 ομάδες. Ας δοκιμάσουμε να χωρίσουμε σε 2 και 3 clusters για να δούμε τι ομαδοποίηση πετυχαίνεται:

#Compute K-means clustering with k=2:

Δημιουργούνται 2 clusters με μεγέθη 399,41.

Επίσης

```
Within cluster sum of squares by cluster:
[1] 966.3860 982.9619
 (between_SS / total_SS =  26.0 %)
```

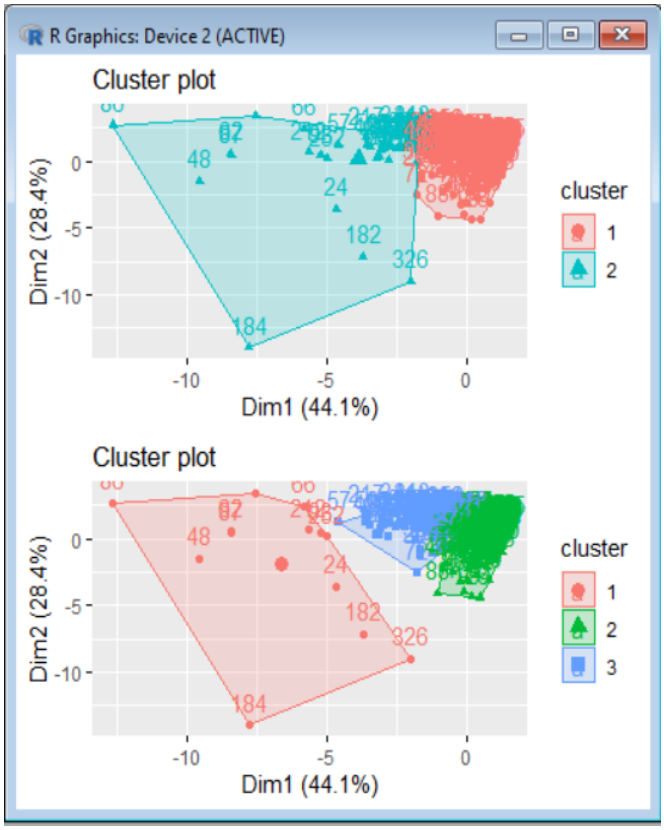
#Compute K-means clustering with k=3:

Δημιουργούνται 2 clusters με μεγέθη 44,3,393.

Επίσης

```
Within cluster sum of squares by cluster:
[1] 441.0021 214.5396 944.8291
 (between_SS / total_SS =  39.2 %)
```

Και μπορούμε να οπτικοποιήσουμε το αποτέλεσμα για 2 και 3 clusters αντίστοιχα:

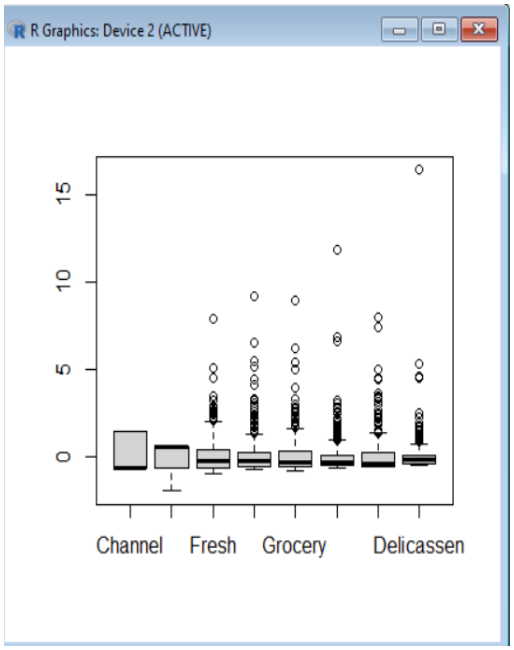



Τελικά θα επιλέξω να χωρίσω σε 3 clusters, γιατί φαίνεται ότι έτσι χωρίζεται η ομάδα 1 για 2 clusters σε δύο επιμέρους. Πιθανόν αυτό το cluster να περιλαμβάνει πελάτες με υψηλή αγοραστική ικανότητα.

Βλέπουμε ότι between_SS/total_SS=39.3%, που είναι πολύ χαμηλό. Βλέποντας και από το σχήμα, ο αλγόριθμος K-means δεν δουλεύει σωστά. Η εκ των προτέρων πιθανότητα για τα 3 clusters διαφέρει σημαντικά, καθώς το πρώτο cluster έχει 44 παρατηρήσεις το δεύτερο μόλις 3 και το τρίτο 393. Επίσης, η μορφή των δεδομένων είναι τέτοια που υπάρχουν και λίγες ατυπικές τιμές (outliers), κάτι που δεν βοηθάει τον αλγόριθμο μας, ο οποίος είναι ευαίσθητος σε αυτές.

Outliers:

Φτιάχνοντας τα boxplots των μεταβλητών φαίνεται ότι υπάρχουν πολλά outliers , τα οποία επηρεάζουν κατα πολύ τον αλγόριθμο k-means από το να αποδόσει οπως θα θέλαμε.




B. Hierarchical clustering

Χρησιμοποιώ την συνάρτηση **agnes** για να κάνω agglomerative ιεραρχική ομαδοποίηση.
Εφαρμόζω στα δεδομένα ιεραρχική ομαδοποίηση με τις ακόλουθες μεθόδους:

1.Κοντινότερου γείτονα

2.Πιο απομακρυσμένου γείτονα

3. Average Link clustering

4. Ward

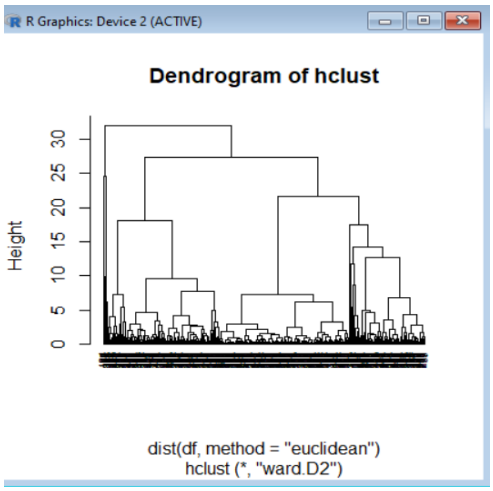
Με τη μέθοδο agnes μπορώ να βρω και το agglomerative coefficient(ac), το οποίο θα μου δώσει μια εικόνα σχετικά με τη δομή του clustering.

```
> map_dbl(m,ac)
  average    single  complete      ward
0.9667024 0.9579012 0.9693598 0.9796740
```

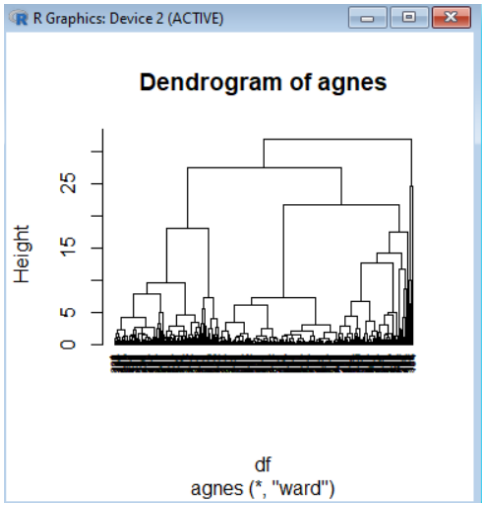
Βλέπουμε λοιπόν ότι η πιο αποδοτική μέθοδος είναι αυτή του Ward. Επομένως, θα χρησιμοποιήσουμε αυτή και στη συνέχεια.

Το δενδρόγραμμα που προκύπτει είναι το εξής:

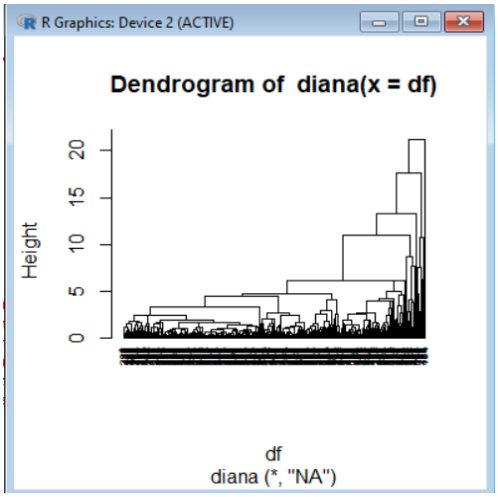
1)Χρησιμοποιώντας hclust:




2)Χρησιμοποιώντας την agnes:

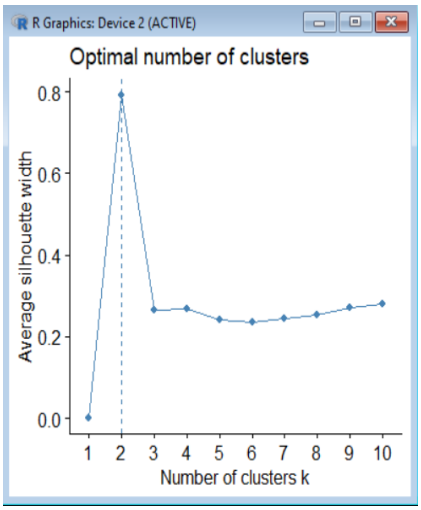



3) Χρησιμοποιώντας την diana (divisive hierarchical clustering):



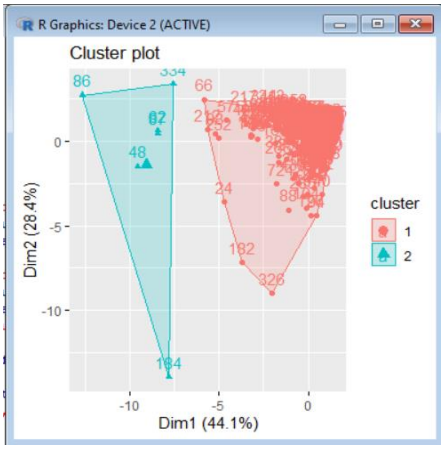

Τώρα πρέπει να βρούμε τον κατάλληλο αριθμό από clusters. Τα σχήματα δε μας επιτρέπουν να εξάγουμε κάποιο ασφαλές συμπέρασμα για το πού μπορούμε να κόψουμε το δέντρο, οπότε θα χρησιμοποιήσουμε εναλλακτικούς τρόπους για να βρούμε τον ιδανικό αριθμό από clusters.

Χρησιμοποιώ την fviz_nbclust() για να βρω τον κατάλληλο αριθμό των ομάδων για τα δεδομένα.Για την Silhouette μέθοδο έχουμε:




, η οποία μας λέει να χρησιμοποιήσουμε 2 clusters.

Για 2 clusters:



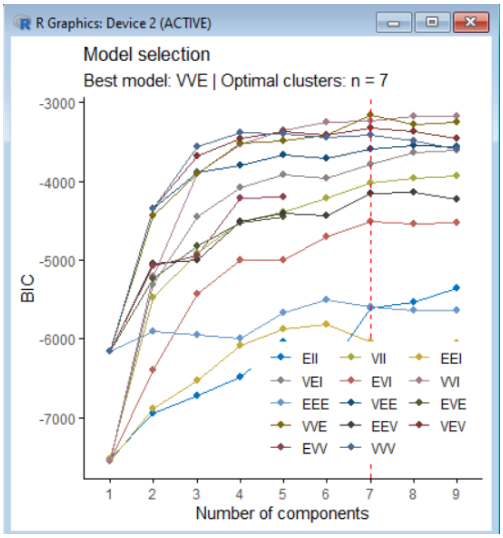

Βλέποντας το διάγραμμα για 2 clusters, και συγκρίνοντας το διάγραμμα που πήραμε για 2 clusters από τον αλγόριμο K-means, βλέπουμε ότι πετυχαίνεται καλύτερη ομαδοποίηση των δεδομένων.

Γ) Model-based clustering

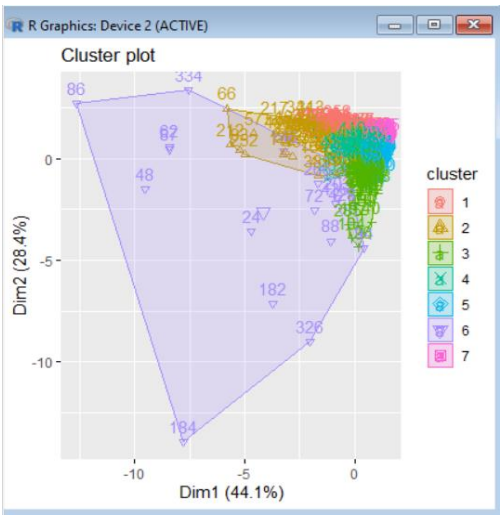
```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVE (ellipsoidal, equal orientation) model with 7 components:

 log-likelihood   n  df       BIC       ICL
      -1262.255 440 105 -3163.621 -3263.066

Clustering table:
 1  2  3  4  5  6  7
69 51 71 88 78 18 65
```

Βλέπουμε λοιπόν ότι για τα δεδομένα μας, ο αλγόριθμος model-based διάλεξε ένα μοντέλο με 7 clusters.Το βέλτιστο μοντέλο είναι το VVE (οι 7 ομάδες έχουν διαφορετικό όγκο, διαφορετικό σχήμα και ίδια κατεύθυνση).

Εξετάζουμε το σχήμα με τα BIC:




Το BIC είναι χαμηλό για το βέλτιστο μοντέλο (7 clusters), κάτι που μας ικανοποιεί. Εμείς όμως θα κρίνουμε και με βάση το οπτικό αποτέλεσμα από το cluster plot:




Το καλό εδώ είναι ότι πετυχαίνουμε μια ομοιομορφία σχετικά με την τοποθέτηση των παρατηρήσεων στα clusters, κάτι που δε συνέβαινε στους προηγούμενους 2 αλγορίθμους, οι οποίοι για 2 clusters πετύχαιναν ανομοιόμορφες κατανομές (ο kmeans πετύχαινε 399-41 κατανομή και ο hierarchical πάλι πετύχαινε λίγες στο ένα cluster).Το πρόβλημα της ανομοιομορφίας όμως συνεχιζόταν και για μεγαλύτερο αριθμό από clusters (για 3 η κατανομή του k-means ήταν 44-393-3 και για 7, αν θέλουμε να συγκρίνουμε με τον model-based, ήταν 5-1-250-2-104-28-50).

Απαντήσεις στα ερωτήματα

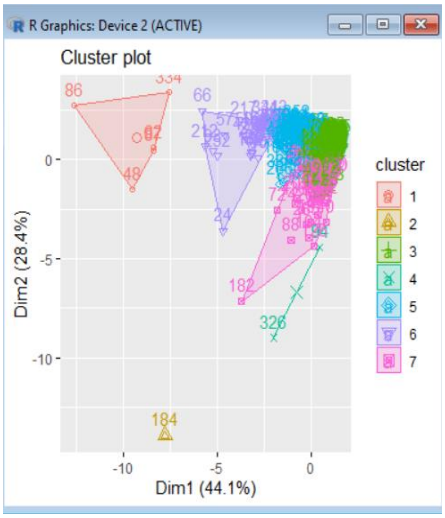
I) Η προτιμότερη μέθοδος φαίνεται να είναι η μέθοδος του model-based clustering καθώς δημιουργεί 7 σχετικά ομοιόμορφες ομάδες. Αντίθετα, η μέθοδος του k-means αποτυγχάνει παταγωδώς στην ομοιομορφία (δημιουργεί χωρισμό 44-3-393) και λόγω των outliers δε μπορεί να λειτουργήσει σωστά. Η μέθοδος hierarchical clustering πάσχει και αυτή από το ίδιο πρόβλημα (της ανοιομορφίας).

II) Εφόσον κατέληξα στο μοντέλο του model-based clustering, μπορώ να κρίνω από τον αριθμό του BIC και από το ποσοστό ομοιομορφίας των δεδομένων (το ιδανικό θα ήταν μέσω του silhouette coefficient αλλά δεν έχω βρει τρόπο να το υπολογίσω στην R). Το BIC έχει καλή τιμή (μικρότερο από -3000) και τα δεδομένα μας έχουν χωριστεί ομοιόμορφα στα clusters, κάτι που παρατηρούμε από το σχήμα.Δυστυχώς στη μέθοδο του model-based δεν έχουμε κάποιο παραπάνω metric για να μας βοηθήσει να βγάλουμε περαιτέρω συμπεράσματα. Παρ'όλα αυτά, όπως είπαμε, η ομαδοποίηση φαίνεται καλή.

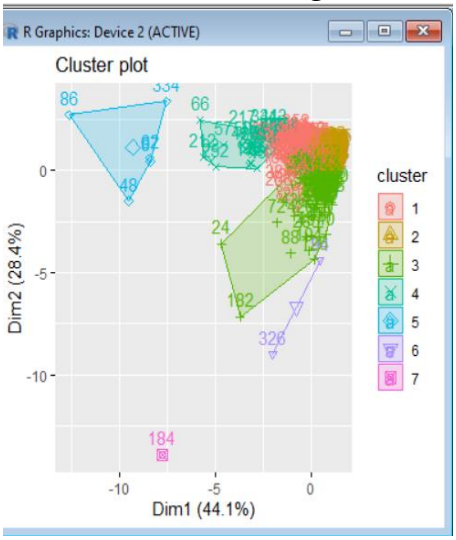
III) Παρατηρούμε ότι το αριστερό κομμάτι, στο οποίο οι τιμές περίπου ισαπέχουν, σε όλες τις περιπτώσεις μπήκε σε ένα cluster (το αριστερό μέρος της ομάδας 6 (μωβ) του model-based με την ομάδα 2 του hierarchical clustering και την ομάδα 1 του k-means για 3 clusters.Σε όλες τις περιπτώσεις η ομάδα αυτή αναγνωρίστηκε από τους αλγορίθμους.

IV) Στο model-based clustering οι ομάδες στο δεξί κομμάτι είναι πολλές και σχετικά ομοιόμορφα χωρισμένες. Στους άλλους δύο αλγορίθμους το δεξί κομμάτι το αποτελούσε μία ομάδα για 2 clusters.
Παραθέτω για 7 ομάδες τα σχήματα των k-means και hierarchical clustering για καλύτερη σύγκριση:

K-means, 7 clusters




Hierarchical clustering, 7 clusters




Οι αλγόριθμοι K-means και Hierarchical Clustering έχουν κάποιες κοινές ομάδες αλλά έχουν και διαφορές στην ομαδοποίηση. Χωρίζουν πάντως διαφορετικά από τον model-based, επομένως δεν μπορούμε να πούμε εδώ ότι υπάρχουν ευδιάκριτες ομάδες.

V) Δεν έχω βρει καταφέρει να βρω τον πίνακα συνάφειας στο model-based στην R, αλλά θα σας παραθέσω τον πίνακα με βάση τους αλγορίθμους K-means και Hierarchical Clustering.

K-means:

```
> data[1:2]%>%
+ mutate(Cluster=k2$cluster)%>%
+ filter(Cluster==1)%>% ungroup() %>% count(Channel,Region)
  Channel Region   n
1       1      1  59
2       1      2  28
3       1      3 210
4       2      1   2
5       2      2   1
6       2      3   5
```

```
> data[1:2]%>%
+ mutate(Cluster=k2$cluster)%>%
+ filter(Cluster==2)%>% ungroup() %>% count(Channel,Region)
  Channel Region   n
1       1      3   1
2       2      1  16
3       2      2  18
4       2      3 100
```

Επομένως αν τα ενώσουμε έχουμε:

| # | Channel | Region | #of obs. in Cluster 1 | #of obs. in Cluster 2 |
|---|---|---|---|---|
| 1 | 1 | 1 | 59 | - |
| 2 | 1 | 2 | 28 | - |
| 3 | 1 | 3 | 210 | 1 |
| 4 | 2 | 1 | 2 | 16 |
| 5 | 2 | 2 | 1 | 18 |
| 6 | 2 | 3 | 5 | 100 |

Hierarchical Clustering:

```
> data[1:2]%>%
+ mutate(Cluster=sub_group)%>%
+ filter(Cluster==1)%>% ungroup() %>% count(Channel,Region)
  Channel Region   n
1       1      3   1
2       2      1  18
3       2      2  19
4       2      3 104
```

```
> data[1:2]%>%
+ mutate(Cluster=sub_group)%>%
+ filter(Cluster==2)%>% ungroup() %>% count(Channel,Region)
  Channel Region   n
1       1      1  59
2       1      2  28
3       1      3 210
4       2      3   1
```

Επομένως αν τα ενώσουμε έχουμε:

| # | Channel | Region | #of obs. in Cluster 1 | #of obs. in Cluster 2 |
|---|---|---|---|---|
| 1 | 1 | 1 | - | 59 |
| 2 | 1 | 2 | - | 28 |
| 3 | 1 | 3 | 1 | 210 |
| 4 | 2 | 1 | 18 | - |
| 5 | 2 | 2 | 19 | - |
| 6 | 2 | 3 | 104 | 1 |

Βλέπουμε ότι και με τους δύο αλγορίθμους η κατανομή είναι πολύ άνιση, δηλαδή έχουμε περιπτώσεις π.χ.210-1, 5-100 που σημαίνει ότι κάθε cluster μπορεί να αποφασίσει το Channel-Region στο οποίο αντιστοιχεί. (διότι δεν έχουμε περιπτώσεις σχεδόν ίσης κατανομής, τύπου 30-35 που δεν δίνει ξεκάθαρο συμπέρασμα)

Τελικά Συμπεράσματα

Στο σετ δεδομένων που μας δόθηκε, εξαιτίας της πληθώρας από ατυπικές τιμές (outliers) ο αλγόριθμος K-means δεν λειτουργεί ορθολογικά, φτιάχοντας ανομοιόμορφες ομάδες, κάτι που ισχύει και για τον αλγόριθμο ιεραρχικής ομαδοποίησης. Ο αλγόριθμος model-based προτείνει 7 ομάδες (αντί 3 και 2 που πρότειναν οι προηγούμενοι) και καταφέρνει να πετύχει μια ομοιόμορφη κατανομή των παρατηρήσεων σε clusters ,γεγονός που κατά την άποψή μου τον καθιστά ως την καλύτερη λύση για τα συγκεκριμένο dataset.