Επιχειρησιακή Έρευνα και Επιχειρηματική Ευφυΐα

1η Υποχρεωτική Εργασία

Στέφανος Καραμπέρας - ΑΕΜ 2910

Εισαγωγή

Σκοπός της παρούσας εργασίας είναι η εξόρυξη γνώσης από δεδομένα συναλλαγών ενός καταστήματος λιανικής.

Συγκεκριμένα, ως δεδομένο προσφέρεται η ανάλυση 7537 συναλλαγών (καλάθια) που πραγματοποιήθηκαν στην επιχείρηση μέσα σε μια περίοδο 75 ημερών και αφορούν 170 κωδικούς προϊόντων.

Τα ερωτήματα ανάλυσης που καλούμαστε να απαντήσουμε αφορούν ποιοτικά και ποσοτικά χαρακτηριστικά των αγορών, κατηγοριοποίηση των πελατών σύμφωνα με τα μοτίβα συμπεριφοράς που μπορούν να εντοπιστούν από τις αγορές τους αλλά και προσδιορισμό κανόνων συσχετίσεως που παρουσιάζουν ιδιαίτερο ενδιαφέρον για τις μελλοντικές αποφάσεις της επιχείρησης.

Για τον σκοπό της ανάλυσης των δεδομένων θα χρησιμοποιηθεί η γλώσσα προγραμματισμού R.

Δομή κώδικα

Ο κώδικας που έχει συγγραφεί για τους σκοπούς της εξόρυξης γνώσης βρίσκεται στο αρχείο «**main.R**».

Για ευκολία ανάγνωσης, κατανόησης, συντήρησης αλλά και ευκολία αξιοποίησης του κώδικα στα διαφορετικά σενάρια χρήσης που προκύπτουν από το δοθέν καθήκον, έγινε εκτενής διαχωρισμός του υλοποιημένου κώδικα σε συναρτήσεις (functions), ενώ έχει προστεθεί επαρκής σχολιασμός σε όλη την έκτασή του.

Οι συναρτήσεις που έχουν δημιουργηθεί παρουσιάζονται συνοπτικά παρακάτω, σύμφωνα με τα ερωτήματα της εργασίας στα οποία αντιστοιχούν. Αξίζει να σημειωθεί ότι ορισμένες συναρτήσεις είναι αναγκαίες σε περισσότερα από ένα ερωτήματα, συνεπώς θα κατηγοριοποιηθούν σύμφωνα με το ερώτημα βασικής αξιοποίησής τους.

Συναρτήσεις:

- Γενικού σκοπού:
  - execute()
  - binarize(dataColumns, extraColumns=NULL)

- Άσκηση 1:
  - prepareData()

- Άσκηση 2:
  - testAssociationRules(groceriesDiscrete)
  - generateAssociationRulesByConfidence(groceriesDiscrete)

- Άσκηση 3:
  - filterNormalizeCostRecency(groceriesDiscrete)
  - performClustering(normalizedCostAndRecency)
  - printClusteringCharts(groceriesDiscrete)
  - generateGroceriesWithBinaryClusterData(groceriesDiscrete, kmeansFit)

- Άσκηση 4:
  - clusterProductProfile(groceriesWithClusters)

Αναλυτική περιγραφή επίλυσης των ερωτημάτων

Άσκηση 1

1. Μετασχηματισμός των πρωτογενών δεδομένων σε δυαδική μορφή συναλλαγών

Αρχικά, γίνεται ανάγνωση των δεδομένων από το δοθέν αρχείο (GroceriesInitial.csv) με χρήση της εντολής:

```
groceries <- read.csv("GroceriesInitial.csv",header=TRUE,sep=",", stringsAsFactors=TRUE)
```

Στη συνέχεια, πραγματοποιείται μετατροπή των πρωτογενών δεδομένων συναλλαγών σε δυαδική μορφή συναλλαγών (οι ονομασίες των διακριτών προϊόντων γίνονται ιδιότητες (στήλες) και κάθε καταχώρηση (σειρά) μπορεί να έχει είτε **«TRUE»** είτε **«FALSE»** κάτω από κάθε στήλη προϊόντος, ανάλογα με το αν το προϊόν περιλαμβάνεται ή όχι στην συναλλαγή).

Για τον σκοπό αυτό, αξιοποιούμε τη συνάρτηση γενικού σκοπού **binarize(dataColumns, extraColumns=NULL)** που έχουμε δημιουργήσει, με ορίσματα **as.data.frame(groceriesDiscrete$basket_value_dis)** (η στήλη **basket_value_dis** της **groceriesDiscrete** μετετρεμμένη σε data frame) και **groceriesDiscrete** αντίστοιχα.

Η συνάρτηση ξεκινάει εκτελώντας την εντολή :

```
  columnNames <- levels(unlist(dataColumns))
```

με την οποία συλλέγονται τα διακριτά ονόματα του περιεχομένου των στηλών από το σύνολο δεδομένων **dataColumns**. Αυτό επιτυγχάνεται με την ενοποίηση των περιεχομένων των στηλών (όλες οι σειρές) σε μία ενιαία λίστα (χρήση ενσωματωμένης συνάρτησης **unlist(dataColumns)** και στη συνέχεια με την επιλογή των διακριτών ονομάτων του περιεχομένου των στηλών από την προαναφερθείσα ενοποιημένη λίστα (χρήση ενσωματωμένης συνάρτησης **levels()**).

Με δεδομένο ότι στο σύνολο δεδομένων **dataColumns** είναι πιθανό να υπάρχουν στήλες που είναι κενές, είναι προφανές ότι μεταξύ των διακριτών ονομάτων που έχουν εντοπιστεί δύναται ναι συμπεριλαμβάνεται και η κενή συμβολοσειρά (**“”**). Για τον εντοπισμό και την αφαίρεση της κενής συμβολοσειράς από τα διακριτά ονόματα των περιεχομένων των στηλών, γίνεται χρήση των εντολών:

```
# Remove the "" element from (if it exists) from column names
blank <- which(columnNames == "")
if (length(blank) !=0)
  columnNames <- columnNames[-c(blank)]
```

Μετά από την παραπάνω προεπεξεργασία, η συνάρτηση προχωράει στη διαδικασία μετατροπής των δεδομένων σε δυαδική μορφή στις παρακάτω γραμμές κώδικα:

```
binaryResult <- as.data.frame(t(apply(dataColumns, 1
  , function(x) columnNames %in% as.character(unlist(x)))))
```

Πιο αναλυτικά:

- Εφαρμόζεται επαναληπτικά η ανώνυμη συνάρτηση **function(x)** για τις ιδιότητες των προϊόντων (**dataColumns**) κάθε καταχώρησης (σειράς) με χρήση της έτοιμης συνάρτησης **apply()**.

- Μέσω της **function(x)**, για κάθε καταχώρηση (σειρά) συνδυάζονται τα δεδομένα όλων των ιδιοτήτων (στηλών) σε μία ενιαία λίστα με χρήση της εντολής **unlist()**.
  Στη συνέχεια, για κάθε όνομα της **columnNames**, ελέγχεται αν το όνομα εμφανίζεται μέσα στην προαναφερθείσα ενιαία λίστα, με αποτέλεσμα την επιστροφή της τιμής **«TRUE»** αν υπάρχει εμφάνιση ή της τιμής **«FALSE»** αν δεν υπάρχει.

- Ως αποτέλεσμα, επιστρέφεται μια λίστα που περιέχει τις διακριτές τιμές **«TRUE»** και **«FALSE»** με σειρά ανάλογη της σειράς των ονομάτων της **columnNames**. Το αποτέλεσμα αποθηκεύεται στη μεταβλητή **binaryResult**.

- Η έτοιμη συνάρτηση **t()** καλείται για την εφαρμογή ενέργειας αντιμετάθεσης των γραμμών με τις στήλες του πίνακα που προκύπτει παραπάνω (transpose of matrix) προκειμένου να καταλήξουμε στην επιθυμητή μορφή των δεδομένων αποτελέσματος.

2. Φιλτράρισμα των 13 προϊόντων ενδιαφέροντος από το σύνολο δεδομένων

Προκειμένου να είμαστε σε θέση να επιλέξουμε τα 13/170 προϊόντα που μας έχουν γνωστοποιηθεί ως προϊόντα ενδιαφέροντος, ξεκινάμε αντιστοιχίζοντας στη μεταβλητή **productsBinary** τις ονομασίες των προϊόντων με αντιστοίχιση ένα προς ένα. Τα ονόματα των προϊόντων αποθηκεύονται ως τίτλοι των στηλών της **productsBinary**:

```
    names(productsBinary) <- productNames
```

Αμέσως μετά, με την εντολή:

```
filteredProductsBinary <- productsBinary[, c("citrus fruit", "tropical fruit", "whole milk",
"other vegetables","rolls/buns", "chocolate", "bottled water", "yogurt", "sausage", "root
vegetables", "pastry", "soda", "cream")]
```

πραγματοποιείται «φιλτράρισμα» της **productsBinary**, διατηρώντας μόνο τις στήλες με τις ονομασίες που καθορίζονται εντός της λίστας που δημιουργείται από την έτοιμη συνάρτηση **c().**

Ακολούθως, προκειμένου να ενοποιήσουμε τη διακριτοποιημένη λίστα ονομάτων προιόντων αξίας με τις 3 πρώτες στήλες (id, basket_value, recency_days) της αρχικής συλλογής δεδομένων, εκτελούμε την εντολή:

```
groceriesBinary <- cbind(groceries[,1:3], filteredProductsBinary)
```

3. Διακριτοποίηση ιδιότητας αξίας συναλλαγής

Σε αυτό το σημείο πραγματοποιείται διακριτοποίηση της ιδιότητας αξίας συναλλαγής (basket_value) των αρχικών δεδομένων σε τρεις (περίπου) ισοπληθείς κατηγορίες **low_value_basket**, **medium_value_basket**, **high_value_basket**.

Για τον σκοπό αυτό, χρησιμοποιούνται οι εντολές:

```
groceriesDiscrete <- groceriesBinary
cutPoints <- quantile(groceriesDiscrete$basket_value, probs = seq(0, 1, 1/3), na.rm = TRUE,
names = FALSE)
```

ώστε σύμφωνα με τις τιμές της ιδιότητας basket_value να οριστούν τα 3 περίπου ισοπληθή διαστήματα.

Στη συνέχεια, τα διαστήματα αυτά τροφοδοτούνται στην εντολή:

```
groceriesDiscrete$basket_value_dis <- cut(groceriesDiscrete$basket_value, breaks =
cutPoints,
                                labels=c("Low","Medium","High"), include.lowest = TRUE)
```

υπό τη μορφή του ορίσματος σημείων διαχωρισμού (**breaks**). Η έτοιμη συνάρτηση **cut()** κατηγοριοποιεί για κάθε καταχώρηση την τιμή της ιδιότητας «**basket_value**» ανάλογα με το διάστημα στο οποίο ανήκει ως «**Low**», «**Medium**» ή «**High**». Το αποτέλεσμα της κατηγοριοποίησης αποθηκεύεται στη νέα ιδιότητα (στήλη) «**basket_value_dis**» της **groceriesDiscrete.**

Η μετατροπή των τιμών της «**basket_value_dis**» σε δυαδική μορφή γίνεται με κλήση τη συνάρτησης **binarize(dataColumns, extraColumns=NULL)** που έχει δημιουργηθεί και ορίσματα **as.data.frame(groceriesDiscrete$basket_value_dis)** και **groceriesDiscrete** αντίστοιχα. Η λειτουργία της συνάρτησης έχει ήδη περιγραφτεί για τη διαδικασία εξαγωγής των δυαδικών τιμών ονομασιών προϊόντων πιο πάνω. Από το αποτέλεσμα της συνάρτησης, αφαιρείται τελικά η στήλη «**basket_value_dis**», η οποία μετά την εξαγωγή των δυαδικών τιμών κατηγορίας τιμής καλαθιού είναι περιττή.

Τέλος, η συνάρτηση **prepareData()** επιστρέφει ως αποτέλεσμα τη **groceriesDiscrete**, η οποία περιέχει τις αρχικές ιδιότητες **id**, **basket_value**, **recency_days**, τις **διακριτοποιημένες** ιδιότητες με τις **ονομασίες των 13 προϊόντων ενδιαφέροντος**, καθώς και την **διακριτοποιημένη** ιδιότητα κατηγοριοποίησης η οποία αποτελείται από 3 στήλες «**Low**», «**Medium**» και «**High**» των οποίων οι σειρές συμπληρώνονται με τις τιμές «**TRUE**» ή «**FALSE**», υποδεικνύοντας την κατηγορία τιμής κάθε συναλλαγής (σειράς).

Άσκηση 2

Στόχος της άσκησης είναι η μάθηση κανόνων συσχέτισης μέσα από το επεξεργασμένο σύνολο των αρχικών δεδομένων (**groceriesDiscrete**), αποκλειστικά για τα χαρακτηριστικά των προϊόντων και τη διακριτοποιημένη αξία καλαθιού.

1. Πειραματισμός με διαφορετικές τιμές για το ελάχιστο Support

Στο πλαίσιο του πειραματισμού, πραγματοποιούμε αρχικά δοκιμαστικές εκτελέσεις της μεθόδου **Apriori** για την εξαγωγή κανόνων συσχέτισης, χρησιμοποιώντας διαφορετικές τιμές στο όρισμα του **ελάχιστου Support**.

Για την εκτέλεση των δοκιμών, δημιουργήθηκε η συνάρτηση **testAssociationRules(groceriesDiscrete)**, εντός της οποίας πραγματοποιούνται 4 συνολικά εκτελέσεις της **apriori()**, με ορίσματα για το ελάχιστο Support τις τιμές **0.001**, **0.02**, **0.03** και **0.04**.

Τα αποτελέσματα που λαμβάνουμε για κάθε μία από τις τιμές παρατίθενται παρακάτω (για λόγους ευκολίας απεικόνισης και μελέτης, σε κάθε εκτέλεση εμφανίζονται τα πρώτα 20 αποτελέσματα των κανόνων συσχέτισης που προκύπτουν):

I. Για **supp = 0.001**:

```
[1] "rules test 1: "
     lhs                            rhs                      support     confidence lift     count
[1]  {cream}                     => {basket_value_dis=High} 0.001459660 0.8461538  2.545555  11
[2]  {sausage}                   => {basket_value_dis=High} 0.107351380 0.8755411  2.633963 809
[3]  {citrus fruit,chocolate}    => {basket_value_dis=High} 0.007696391 0.9206349  2.769623  58
[4]  {chocolate,pastry}          => {basket_value_dis=High} 0.010483015 1.0000000  3.008383  79
[5]  {chocolate,sausage}         => {basket_value_dis=High} 0.008625265 1.0000000  3.008383  65
[6]  {chocolate,bottled water}   => {basket_value_dis=High} 0.006502123 0.8596491  2.586154  49
[7]  {tropical fruit,chocolate}  => {basket_value_dis=High} 0.010084926 0.9500000  2.857964  76
[8]  {chocolate,root vegetables} => {basket_value_dis=High} 0.007696391 0.9206349  2.769623  58
[9]  {chocolate,yogurt}          => {basket_value_dis=High} 0.010748408 0.8901099  2.677792  81
[10] {chocolate,soda}            => {basket_value_dis=High} 0.014463907 0.8195489  2.465517 109
[11] {rolls/buns,chocolate}      => {basket_value_dis=High} 0.014198514 0.9224138  2.774974 107
[12] {other vegetables,chocolate} => {basket_value_dis=High} 0.014331210 0.8640000  2.599243 108
[13] {whole milk,chocolate}      => {basket_value_dis=High} 0.018842887 0.8658537  2.604820 142
[14] {citrus fruit,pastry}       => {basket_value_dis=High} 0.011279193 0.8854167  2.663673  85
[15] {citrus fruit,sausage}      => {basket_value_dis=High} 0.014729299 1.0000000  3.008383 111
[16] {sausage,pastry}            => {basket_value_dis=High} 0.016321656 1.0000000  3.008383 123
[17] {bottled water,pastry}      => {basket_value_dis=High} 0.010350318 0.8863636  2.666522  78
[18] {tropical fruit,pastry}     => {basket_value_dis=High} 0.017250531 1.0000000  3.008383 130
[19] {root vegetables,pastry}    => {basket_value_dis=High} 0.013402335 0.9351852  2.813395 101
[20] {yogurt,pastry}             => {basket_value_dis=High} 0.021762208 0.9425287  2.835488 164
```

Αξίζει να σημειωθεί ότι για την συγκεκριμένη ελάχιστη τιμή του support, λαμβάνονται συνολικά 750 κανόνες συσχέτισης, από τους οποίους, όπως αναφέρθηκε και πριν, εξετάζονται οι 20.

II. Για **supp = 0.02**:

```
[1] "rules test 2: "
     lhs                                          rhs                    support    confidence lift     count
[1]  {sausage}                                 => {basket_value_dis=High} 0.10735138 0.8755411  2.633963 809
[2]  {yogurt,pastry}                           => {basket_value_dis=High} 0.02176221 0.9425287  2.835488 164
[3]  {rolls/buns,pastry}                       => {basket_value_dis=High} 0.02733546 1.0000000  3.008383 206
[4]  {other vegetables,pastry}                 => {basket_value_dis=High} 0.02720276 0.9234234  2.778012 205
[5]  {whole milk,pastry}                       => {basket_value_dis=High} 0.03662420 0.8440367  2.539186 276
[6]  {yogurt,sausage}                          => {basket_value_dis=High} 0.02561040 1.0000000  3.008383 193
[7]  {sausage,soda}                            => {basket_value_dis=High} 0.03171444 1.0000000  3.008383 239
[8]  {rolls/buns,sausage}                      => {basket_value_dis=High} 0.03994161 1.0000000  3.008383 301
[9]  {other vegetables,sausage}                => {basket_value_dis=High} 0.03516454 1.0000000  3.008383 265
[10] {whole milk,sausage}                      => {basket_value_dis=High} 0.03901274 1.0000000  3.008383 294
[11] {tropical fruit,root vegetables}          => {basket_value_dis=High} 0.02375265 0.8647343  2.601452 179
[12] {tropical fruit,rolls/buns}               => {basket_value_dis=High} 0.02574310 0.8016529  2.411679 194
[13] {rolls/buns,root vegetables}              => {basket_value_dis=High} 0.02547771 0.8033473  2.416776 192
[14] {tropical fruit,whole milk,other vegetables} => {basket_value_dis=High} 0.02016985 0.9047619  2.721871 152
```

Παρατηρούμε ότι για την συγκεκριμένη ελάχιστη τιμή του support, εξάγονται μόλις 14 κανόνες συσχέτισης.

III. Για **supp = 0.03**:

```
[1] "rules test 3: "
    lhs                          rhs                    support    confidence lift     count
[1] {sausage}                 => {basket_value_dis=High} 0.10735138 0.8755411  2.633963 809
[2] {whole milk,pastry}       => {basket_value_dis=High} 0.03662420 0.8440367  2.539186 276
[3] {sausage,soda}            => {basket_value_dis=High} 0.03171444 1.0000000  3.008383 239
[4] {rolls/buns,sausage}      => {basket_value_dis=High} 0.03994161 1.0000000  3.008383 301
[5] {other vegetables,sausage} => {basket_value_dis=High} 0.03516454 1.0000000  3.008383 265
[6] {whole milk,sausage}      => {basket_value_dis=High} 0.03901274 1.0000000  3.008383 294
```

Παρατηρούμε ότι για την συγκεκριμένη ελάχιστη τιμή του support, εξάγονται μόλις 6 κανόνες συσχέτισης.

IV. Για **supp = 0.04**:

```
[1] "rules test 4: "
    lhs          rhs                    support   confidence lift     count
[1] {sausage} => {basket_value_dis=High} 0.1073514 0.8755411  2.633963 809
```

Παρατηρούμε ότι για την συγκεκριμένη ελάχιστη τιμή του support, εξάγεται μόλις 1 κανόνας συσχέτισης.

2. Κανόνες με το υψηλότερο confidence αποκλειστικά για τα προϊόντα

Σε συνέχεια της διεκπεραίωσης των καθηκόντων της άσκησης, καλούμαστε να βρούμε τους 20 κανόνες με το υψηλότερο confidence αποκλειστικά για τα προϊόντα. Για τον σκοπό αυτό, έχει υλοποιηθεί η συνάρτηση **generateAssociationRulesByConfidence(groceriesDiscrete)**.

Θα εφαρμοστεί η ίδια λογική με την διαδικασία πειραματισμού που πραγματοποιήθηκε στην υποενότητα 1 της παρούσας άσκησης, ωστόσο αυτή τη φορά η επιλογή της τιμής του ελάχιστου Support παρουσιάζει αυξημένο ενδιαφέρον: Σκοπός μας είναι η επιλογή της μέγιστης δυνατής τιμής του **Support**, για την οποία λαμβάνουμε ως αποτέλεσμα τουλάχιστον 20 κανόνες συσχέτισης προϊόντων (λόγω των ζητούμενων της άσκησης). Ο λόγος πίσω από την επιλογή της μέγιστης δυνατής τιμής **support** για το όρισμα του **ελάχιστου support** έχει να κάνει με το ότι οι κανόνες συσχέτισης με **υψηλότερο support** αφορούν εξ ορισμού προϊόντα που εμφανίζονται συχνότερα στο σύνολο δεδομένων που έχουμε, συνεπώς αποτελούν σημεία υψηλού ενδιαφέροντος για την ανάλυσή μας.

Δοκιμάζοντας διαφορετικές τιμές ελάχιστου **support**, καταλήγουμε στο συμπέρασμα ότι μία πολύ καλή τιμή για το όρισμα είναι η **supp = 0.001**, μιας και επιστρέφει ακριβώς 20 κανόνες συσχέτισης ως αποτέλεσμα, οι οποίοι εμφανίζουν επαρκή διαφοροποίηση μεταξύ τους στην τιμή του **confidence**.

- Για τιμές μικρότερες του **0.001**, λαμβάνουμε περισσότερους από 20 κανόνες συσχέτισης, οι οποίοι ωστόσο παρουσιάζουν μειωμένη διαφοροποίηση μεταξύ τους ως προς την τιμή του **confidence**.

  Παράλληλα, το γεγονός ότι οι κανόνες συσχέτισης προέκυψαν από χαμηλότερο ελάχιστο **support** υποδηλώνει ότι στο μεταξύ των αποτελεσμάτων υπάρχουν προϊόντα/τιμές που εμφανίζονται σπανιότερα στο σύνολο δεδομένων που επεξεργαζόμαστε. Αυτό μπορεί να αποσπάσει την προσοχή μας από δεδομένα που παρουσιάζουν μεγαλύτερο ενδιαφέρον ανάλυσης λόγω υψηλότερης συχνότητας εμφάνισης στο σύνολο δεδομένων (**υψηλότερο support**).

- Για τιμές μεγαλύτερες του **0.001**, λαμβάνουμε λιγότερους από 20 κανόνες συσχέτισης ως αποτέλεσμα. Είναι προφανές λοιπόν πως λόγω των περιορισμών που δίνονται από την άσκηση, τιμές μεγαλύτερες του **0.001** στο όρισμα του **ελάχιστου support** αποκλείονται.

Τελικά, το αποτέλεσμα της εκτέλεσης είναι το εξής:

```
[1] "Top 20 product rules by Confidence: "
     lhs                                                          rhs                support     confidence lift     count
[1]  {tropical fruit,rolls/buns,sausage,root vegetables}          => {whole milk}       0.001326964 1.0000000  2.998806 10
[2]  {tropical fruit,rolls/buns,bottled water,yogurt,root vegetables} => {whole milk}   0.001061571 1.0000000  2.998806  8
[3]  {tropical fruit,yogurt,sausage,root vegetables}              => {whole milk}       0.001990446 0.9375000  2.811381 15
[4]  {citrus fruit,tropical fruit,whole milk,yogurt,root vegetables} => {other vegetables} 0.001857749 0.9333333  3.696059 14
[5]  {tropical fruit,rolls/buns,bottled water,root vegetables}    => {whole milk}       0.001459660 0.9166667  2.748906 11
[6]  {tropical fruit,yogurt,root vegetables,pastry}               => {whole milk}       0.001326964 0.9090909  2.726187 10
[7]  {citrus fruit,tropical fruit,whole milk,rolls/buns,root vegetables} => {other vegetables} 0.001061571 0.8888889  3.520056  8
[8]  {whole milk,rolls/buns,bottled water,yogurt,root vegetables} => {tropical fruit}   0.001061571 0.8888889  6.490956  8
[9]  {citrus fruit,tropical fruit,whole milk,root vegetables}     => {other vegetables} 0.004113588 0.8857143  3.507484 31
[10] {citrus fruit,root vegetables,pastry}                        => {other vegetables} 0.001990446 0.8823529  3.494173 15
[11] {citrus fruit,tropical fruit,other vegetables,bottled water} => {whole milk}       0.001725053 0.8666667  2.598965 13
[12] {citrus fruit,tropical fruit,rolls/buns,root vegetables}     => {other vegetables} 0.001459660 0.8461538  3.350823 11
[13] {citrus fruit,tropical fruit,other vegetables,yogurt,root vegetables} => {whole milk} 0.001857749 0.8235294  2.469605 14
[14] {citrus fruit,other vegetables,yogurt,root vegetables}       => {whole milk}       0.003052017 0.8214286  2.463305 23
[15] {chocolate,root vegetables,pastry}                           => {other vegetables} 0.001194268 0.8181818  3.240052  9
[16] {citrus fruit,whole milk,root vegetables,pastry}             => {other vegetables} 0.001194268 0.8181818  3.240052  9
[17] {citrus fruit,other vegetables,yogurt,pastry}                => {whole milk}       0.001194268 0.8181818  2.453569  9
[18] {other vegetables,sausage,root vegetables,pastry}            => {yogurt}           0.001194268 0.8181818  4.494037  9
[19] {tropical fruit,rolls/buns,yogurt,root vegetables}           => {whole milk}       0.002919321 0.8148148  2.443472 22
[20] {citrus fruit,tropical fruit,yogurt,root vegetables}         => {other vegetables} 0.002255839 0.8095238  3.205765 17
```

Διερμηνεύοντας το αποτέλεσμα, εντύπωση προκαλούν οι κανόνες συσχέτισης [1] και [2]:

- [1]: **tropical fruit**, **rolls/buns**, **sausage**, **root vegetables** => **whole milk**
- [2]: **tropical fruit**, **rolls/buns**, **bottled water**, **yogurt**, **root vegetable**s => **whole milk**

Παρατηρούμε δηλαδή ένα μοτίβο καταναλωτικής συμπεριφοράς, που υποδεικνύει ότι το προϊόν «**whole milk**» που βρίσκεται στο δεξί μέρος των 2 κανόνων συσχέτισης αγοράζεται με απόλυτη βεβαιότητα (**confidence = 1**) από καταναλωτές που συνδυάζουν την αγορά των προϊόντων «**tropical fruit**», «**rolls/buns**» και «**root vegetables**» με το προϊόν «**sausage**» ή το ζεύγος προϊόντων «**bottled water**» και «**yogurt**».

3. Κανόνες με το υψηλότερο confidence για τα προϊόντα και την διακριτοποιημένη αξία καλαθιού

Σε αυτό το στάδιο της ανάλυσης μας, καλούμαστε να βρούμε τους 20 κανόνες με το υψηλότερο confidence για τα προϊόντα **και** την διακριτοποιημένη αξία καλαθιού.

Θα εφαρμοστεί η ίδια λογική με την διαδικασία πειραματισμού που πραγματοποιήθηκε στην υποενότητα 1 της παρούσας άσκησης, ωστόσο αυτή τη φορά η επιλογή της τιμής του ελάχιστου Support παρουσιάζει αυξημένο ενδιαφέρον: Σκοπός μας είναι η επιλογή της μέγιστης δυνατής τιμής του **Support**, για την οποία λαμβάνουμε ως αποτέλεσμα τουλάχιστον 20 κανόνες συσχέτισης προϊόντων (λόγω των ζητούμενων της άσκησης). Ο λόγος πίσω από την επιλογή της μέγιστης δυνατής τιμής **support** για το όρισμα του **ελάχιστου support** έχει να κάνει με το ότι οι κανόνες συσχέτισης με **υψηλότερο support** αφορούν εξ ορισμού προϊόντα/τιμές που εμφανίζονται συχνότερα στο σύνολο δεδομένων που έχουμε, συνεπώς αποτελούν σημεία υψηλού ενδιαφέροντος για την ανάλυσή μας.

Δοκιμάζοντας διαφορετικές τιμές ελάχιστου **support**, καταλήγουμε στο συμπέρασμα ότι μία πολύ καλή τιμή για το όρισμα είναι η **supp = 0.018**, μιας και επιστρέφει ακριβώς 20 κανόνες συσχέτισης ως αποτέλεσμα, οι οποίοι εμφανίζουν επαρκή διαφοροποίηση μεταξύ τους στην τιμή του **confidence**.

- Για τιμές μικρότερες του **0.018**, λαμβάνουμε περισσότερους από 20 κανόνες συσχέτισης, οι οποίοι ωστόσο παρουσιάζουν μειωμένη διαφοροποίηση μεταξύ τους ως προς την τιμή του **confidence**.

  Παράλληλα, το γεγονός ότι οι κανόνες συσχέτισης προέκυψαν από χαμηλότερο ελάχιστο **support** υποδηλώνει ότι στο μεταξύ των αποτελεσμάτων υπάρχουν προϊόντα/τιμές που εμφανίζονται σπανιότερα στο σύνολο δεδομένων που επεξεργαζόμαστε. Αυτό μπορεί να αποσπάσει την προσοχή μας από δεδομένα που παρουσιάζουν μεγαλύτερο ενδιαφέρον ανάλυσης λόγω υψηλότερης συχνότητας εμφάνισης στο σύνολο δεδομένων (**υψηλότερο support**).

- Για τιμές μεγαλύτερες του **0.018**, λαμβάνουμε λιγότερους από 20 κανόνες συσχέτισης ως αποτέλεσμα. Είναι προφανές λοιπόν πως λόγω των περιορισμών που δίνονται από την άσκηση, τιμές μεγαλύτερες του **0.018** στο όρισμα του **ελάχιστου support** αποκλείονται.

Τελικά, το αποτέλεσμα της εκτέλεσης είναι το εξής:

```
[1] "Top 20 product and value category rules by Confidence: "
     lhs                                                rhs                support    confidence lift     count
[1]  {rolls/buns,pastry}                             => {high_value_basket} 0.02733546 1.0000000  3.008383 206
[2]  {tropical fruit,sausage}                        => {high_value_basket} 0.01817941 1.0000000  3.008383 137
[3]  {sausage,root vegetables}                       => {high_value_basket} 0.01950637 1.0000000  3.008383 147
[4]  {yogurt,sausage}                                => {high_value_basket} 0.02561040 1.0000000  3.008383 193
[5]  {sausage,soda}                                  => {high_value_basket} 0.03171444 1.0000000  3.008383 239
[6]  {rolls/buns,sausage}                            => {high_value_basket} 0.03994161 1.0000000  3.008383 301
[7]  {other vegetables,sausage}                      => {high_value_basket} 0.03516454 1.0000000  3.008383 265
[8]  {whole milk,sausage}                            => {high_value_basket} 0.03901274 1.0000000  3.008383 294
[9]  {tropical fruit,whole milk,yogurt}              => {high_value_basket} 0.01871019 0.9463087  2.846859 141
[10] {yogurt,pastry}                                 => {high_value_basket} 0.02176221 0.9425287  2.835488 164
[11] {other vegetables,pastry}                       => {high_value_basket} 0.02720276 0.9234234  2.778012 205
[12] {tropical fruit,whole milk,other vegetables}    => {high_value_basket} 0.02016985 0.9047619  2.721871 152
[13] {whole milk,rolls/buns,yogurt}                  => {high_value_basket} 0.01804671 0.8888889  2.674118 136
[14] {sausage}                                       => {high_value_basket} 0.10735138 0.8755411  2.633963 809
[15] {whole milk,chocolate}                          => {high_value_basket} 0.01884289 0.8658537  2.604820 142
[16] {tropical fruit,root vegetables}                => {high_value_basket} 0.02375265 0.8647343  2.601452 179
[17] {whole milk,pastry}                             => {high_value_basket} 0.03662420 0.8440367  2.539186 276
[18] {whole milk,other vegetables,rolls/buns}        => {high_value_basket} 0.01924098 0.8238636  2.478498 145
[19] {rolls/buns,root vegetables}                    => {high_value_basket} 0.02547771 0.8033473  2.416776 192
[20] {tropical fruit,rolls/buns}                     => {high_value_basket} 0.02574310 0.8016529  2.411679 194
```

Διερμηνεύοντας το αποτέλεσμα, παρατηρούμε αρχικά ότι στο δεξί μέρος των 20 κανόνων συσχέτισης με το υψηλότερο confidence έχουμε πάντα την τιμή «**high_value_basket**», πράγμα που υποδεικνύει ότι όλες οι προβαλλόμενες συναλλαγές ανήκουν στην κατηγορία υψηλής αξίας.

Με πιο προσεκτική μελέτη των περιεχομένων του αριστερού μέρους των κανόνων συσχέτισης [2] έως [8], παρατηρούμε ότι οι συγκεκριμένοι κανόνες συσχέτισης έχουν confidence = 1, συνεπώς το δεξί μέρος του κανόνα, δηλαδή η υψηλή διακριτοποιημένη αξία του καλαθιού, είναι απόλυτη (βέβαια) συνέπεια για τα περιεχόμενα καλαθιού που παρουσιάζονται στο αριστερό μέρος των κανόνων [2] έως [8].

Προχωρώντας σε ακόμα μεγαλύτερο επίπεδο ανάλυσης, παρατηρούμε ότι στους κανόνες [2] έως [8] το προϊόν «**sausage**» συμπεριλαμβάνεται σε κάθε περίπτωση μεταξύ των προϊόντων του αριστερού μέρους των κανόνων συσχέτισης.

Αυτό μας οδηγεί στη διαπίστωση ότι με απόλυτη βεβαιότητα (**confidence = 1**), οι συναλλαγές στις οποίες συμπεριλαμβάνεται το προϊόν «**sausage**» παρουσιάζουν υψηλή διακριτοποιημένη αξία καλαθιού (**high_value_basket**).

Συνεπώς, λαμβάνοντας υπόψιν τα παραπάνω, είναι πιθανόν ότι το ακριβότερο προϊόν είναι το «**sausage**».

Άσκηση 3

1. Εφαρμογή της μεθόδου ομαδοποίησης k-means στα συνεχή χαρακτηριστικά basket_value και recency_days.

Σε αυτό το στάδιο της επεξεργασίας του συνόλου δεδομένων μας καλούμαστε να εφαρμόσουμε την μέθοδο ομαδοποίησης **k-means** στα 2 συνεχή χαρακτηριστικά **basket_value** και **recency_days** του επεξεργασμένου συνόλου δεδομένων (**groceriesDiscrete**), με σκοπό την εξαγωγή **5 ομάδων (clusters)** συναλλαγών.

Λόγω της εξάρτησης βημάτων ανάλυσης δεδομένων που θα εκτελεστούν αργότερα στην παρούσα εργασία από το φιλτράρισμα, την κανονικοποίηση και την ομαδοποίηση δεδομένων του τρέχοντος σταδίου, δημιουργήθηκαν 2 βοηθητικές συναρτήσεις **filterNormalizeCostRecency(groceriesDiscrete)** και **performClustering(normalizedCostAndRecency)** για την πραγματοποίηση του φιλτραρίσματος/κανονικοποίησης και της ομαδοποίησης αντίστοιχα. Με την χρήση αυτών των συναρτήσεων καθίσταται δυνατή η επαναχρησιμοποίηση του κώδικα σε μεταγενέστερα σημεία της εργασίας, γλιτώνοντας έτσι περιττές επαναλήψεις ενεργειών.

Ακολουθεί συνοπτική περιγραφή της διαδικασίας που ακολουθήθηκε στην R.

Αρχικά απομονώνουμε τις επιθυμητές ιδιότητες (στήλες) από τη **groceriesDiscrete** και τις αποθηκεύουμε στη **normalizedCostAndRecency** με την παρακάτω εντολή:

```
costAndRecency <- groceriesDiscrete[,c("basket_value", "recency_days")]
```

Πριν την εκτέλεση της μεθόδου k-means, είναι σημαντικό για την εγκυρότητα των αποτελεσμάτων να προχωρήσουμε σε κανονικοποίηση των τιμών των ιδιοτήτων «**basket_value**» και «**recency_days**». Συνεπώς, προχωράμε σε κλήση της συνάρτησης **filterNormalizeCostRecency(groceriesDiscrete)** η οποία πραγματοποιεί την κανονικοποίηση με την εκτέλεση της εντολής:

```
normalizedCostAndRecency <- scale(costAndRecency)
```

Τελικά, η συνάρτηση **filterNormalizeCostRecency(groceriesDiscrete)** επιστρέφει το αποτέλεσμα στο σημείο κλήσης της.

Στη συνέχεια, καλούμε τη συνάρτηση **performClustering(normalizedCostAndRecency)** ώστε να πραγματοποιήσουμε ομαδοποίηση των δεδομένων με χρήστη του αλγορίθμου k-means.
Αρχικά, θέτουμε το seed της γεννήτριας τυχαίων αριθμών της R σε προκαθορισμένη τιμή ώστε τα αποτελέσματα της ομαδοποίησης να έχουν δυνατότητα αναπαραγωγής σε μελλοντικό χρόνο. Το παραπάνω επιτυγχάνεται με την εντολή:

```
set.seed(1234)
```

Έπειτα προχωρούμε σε εκτέλεση της ενσωματωμένης συνάρτησης kmeans() της R, δίνοντας ως ορίσματα:

- **normalizedCostAndRecency**: Το σύνολο δεδομένων στο οποίο θέλουμε να εφαρμοστεί ο αλγόριθμος ομαδοποίησης k-means.

- **centers** = 5: Προσδιορίζουμε ότι επιθυμούμε την ομαδοποίηση των δεδομένων σε 5 συστάδες.

- **nstart** = 1000: Προσδιορίζουμε ότι επιθυμούμε 1000 δοκιμαστικές αρχικοποιήσεις του αλγορίθμου (διαδικασία επιλογής τυχαίων κέντρων), από τις οποίες θα διατηρηθεί η καλύτερη.

- **iter.max** = 1000: Προσδιορίζουμε ότι επιθυμούμε την εκτέλεση το πολύ 1000 επαναλήψεων πριν τη διακοπή του αλγορίθμου k-means.

Τελικά, το αποτέλεσμα της ομαδοποίησης επιστρέφεται από τη συνάρτηση **performClustering(normalizedCostAndRecency)** και αποθηκεύεται στην μεταβλητή «**kmeansFit**».

Το αποτέλεσμα, στην αρχική του μορφή, είναι το παρακάτω:

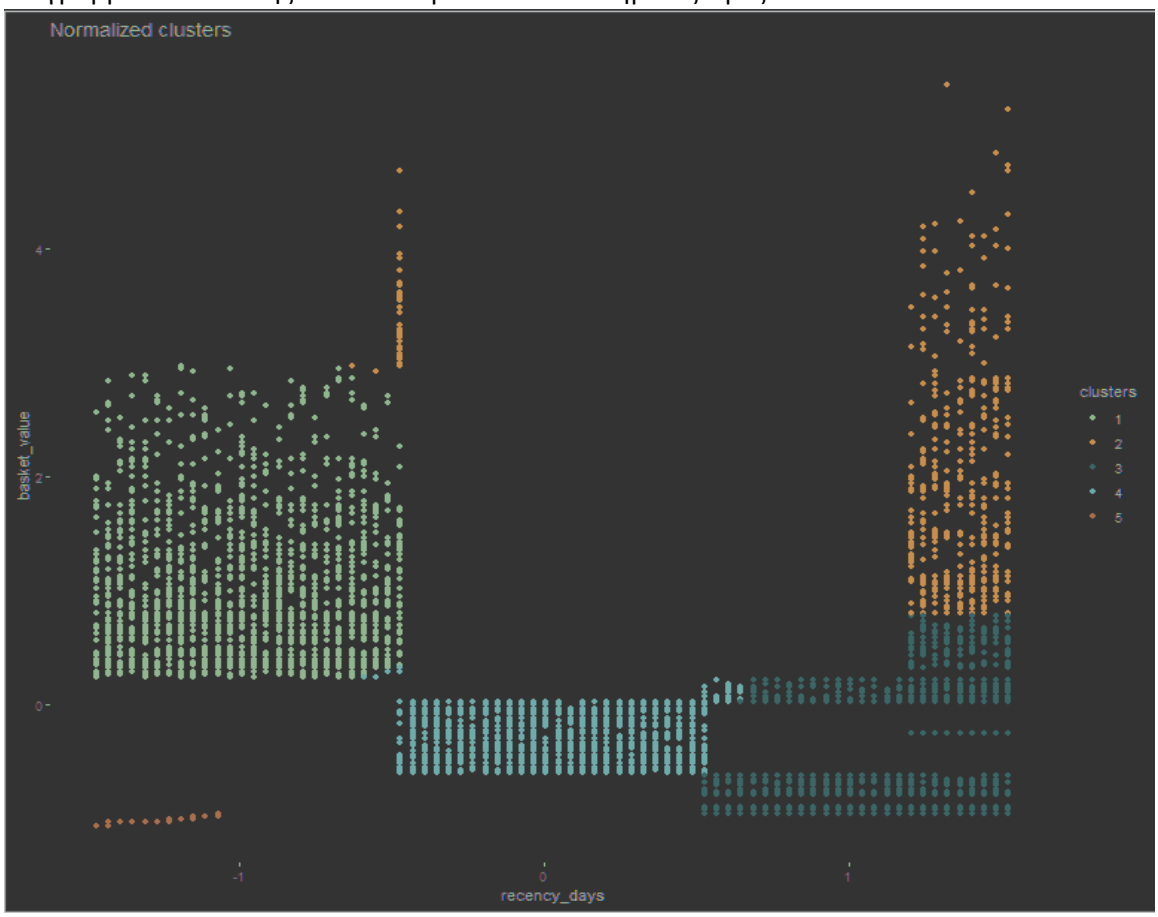
```
[1] "k-means raw result: "
List of 9
 $ cluster     : int [1:7536] 5 4 5 1 3 3 3 4 5 1 ...
 $ centers     : num [1:5, 1:2] 0.989 2.102 -0.508 -0.315 -1.022 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:5] "1" "2" "3" "4" ...
  .. ..$ : chr [1:2] "basket_value" "recency_days"
 $ totss       : num 15070
 $ withinss    : num [1:5] 883 621.8 711.2 250.6 15.6
 $ tot.withinss: num 2482
 $ betweenss   : num 12588
 $ size        : int [1:5] 1867 463 2300 1868 1038
 $ iter        : int 3
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
NULL
```

Για την καλύτερη συνολική απεικόνιση του αποτελέσματος και την εξαγωγή συμπερασμάτων, είναι αναγκαίο σε αυτή τη φάση να προχωρήσουμε στη δημιουργία διαγραμμάτων. Για αυτό τον σκοπό αξιοποιήθηκαν οι δυνατότητες της έτοιμης βιβλιοθήκης «**ggplot2**»:

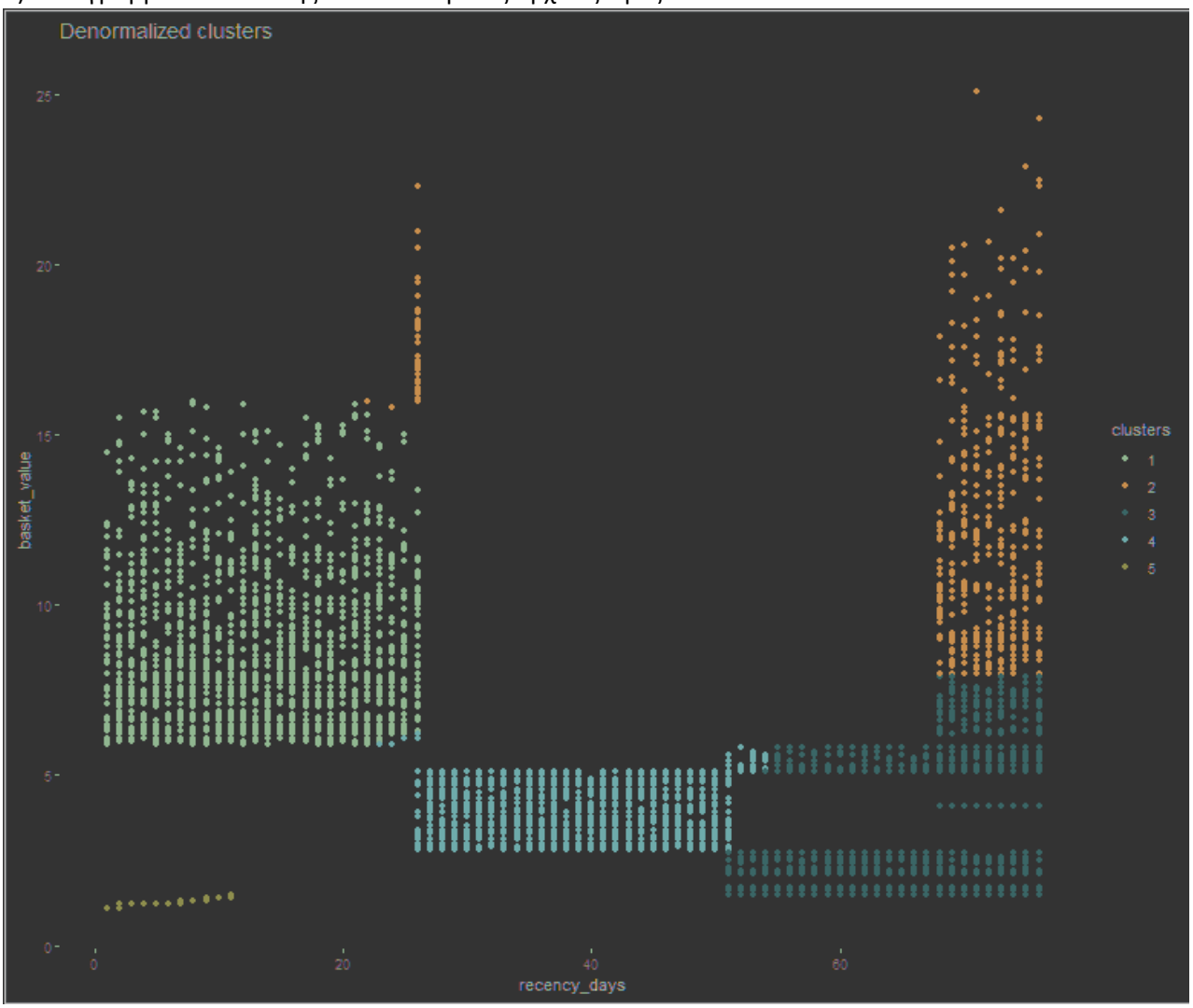
```
# Visualize
#library("factoextra")
library(ggplot2)

# Plot the clustered data returned from kmeans (using normalized axis values)
print(ggplot() + ggtitle( label: "Normalized clusters") +
  geom_point(data = as.data.frame(normalizedCostAndRecency), mapping = aes(x=recency_days, y=basket_value,
  colour = kmeansFit$cluster)) + scale_color_gradient(low="blue", high="red"))

# Plot the clustered data using denormalized axis values
print(ggplot() + ggtitle( label: "Denormalized clusters") +
  geom_point(data = costAndRecency, mapping = aes(x=recency_days, y=basket_value,
  colour = kmeansFit$cluster)) + scale_color_gradient(low="blue", high="red"))
```

Από τον παραπάνω κώδικα, προκύπτουν τα εξής διαγράμματα:

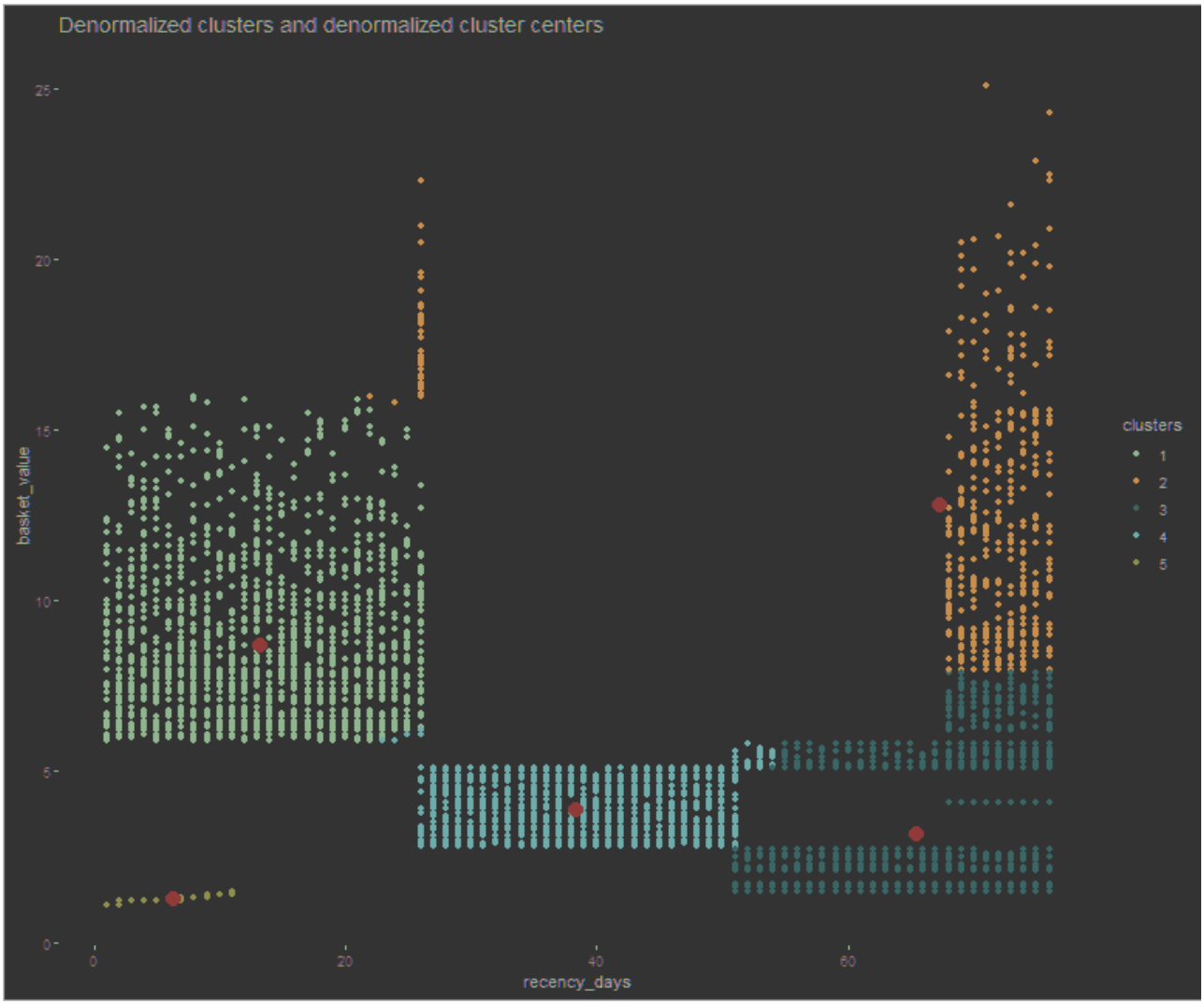
a) Διάγραμμα απεικόνισης συστάδων με κανονικοποιημένες τιμές



b) Διάγραμμα απεικόνισης συστάδων με τις αρχικές τιμές



c) Διάγραμμα απεικόνισης συστάδων και των κέντρων τους σύμφωνα με τις αρχικές (μη κανονικοποιημένες) τιμές



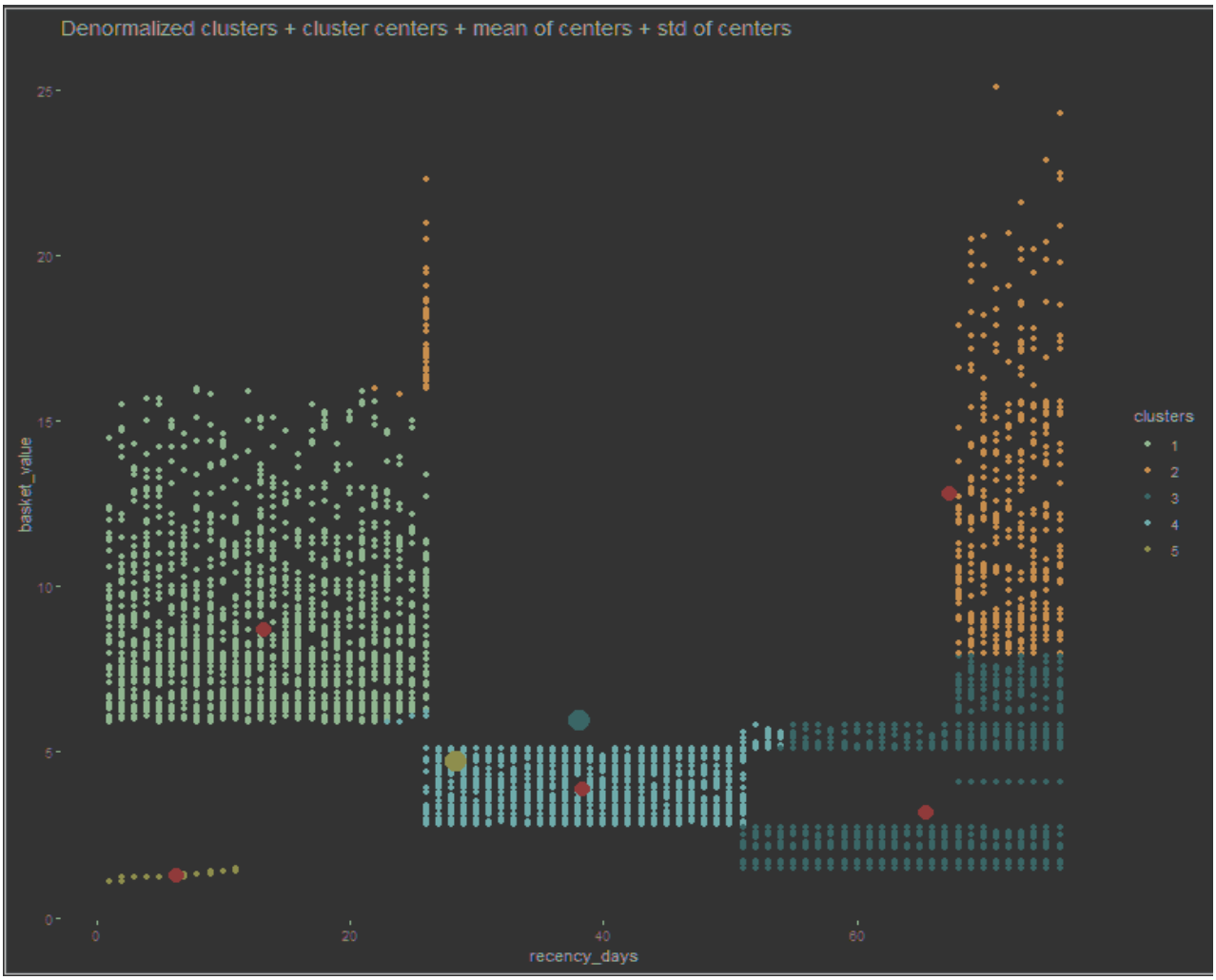
Στο παραπάνω διάγραμμα, οι **κόκκινες** βούλες συμβολίζουν τα κέντρα των συστάδων στην τελική τους μορφή.

2. Αναλυτική εξέταση του αποτελέσματος

Ξεκινάμε αναφέροντας την μέση τιμή των συστάδων που προέκυψαν από τη διαδικασία της ομαδοποίησης, καθώς και την τυπική απόκλισή τους:

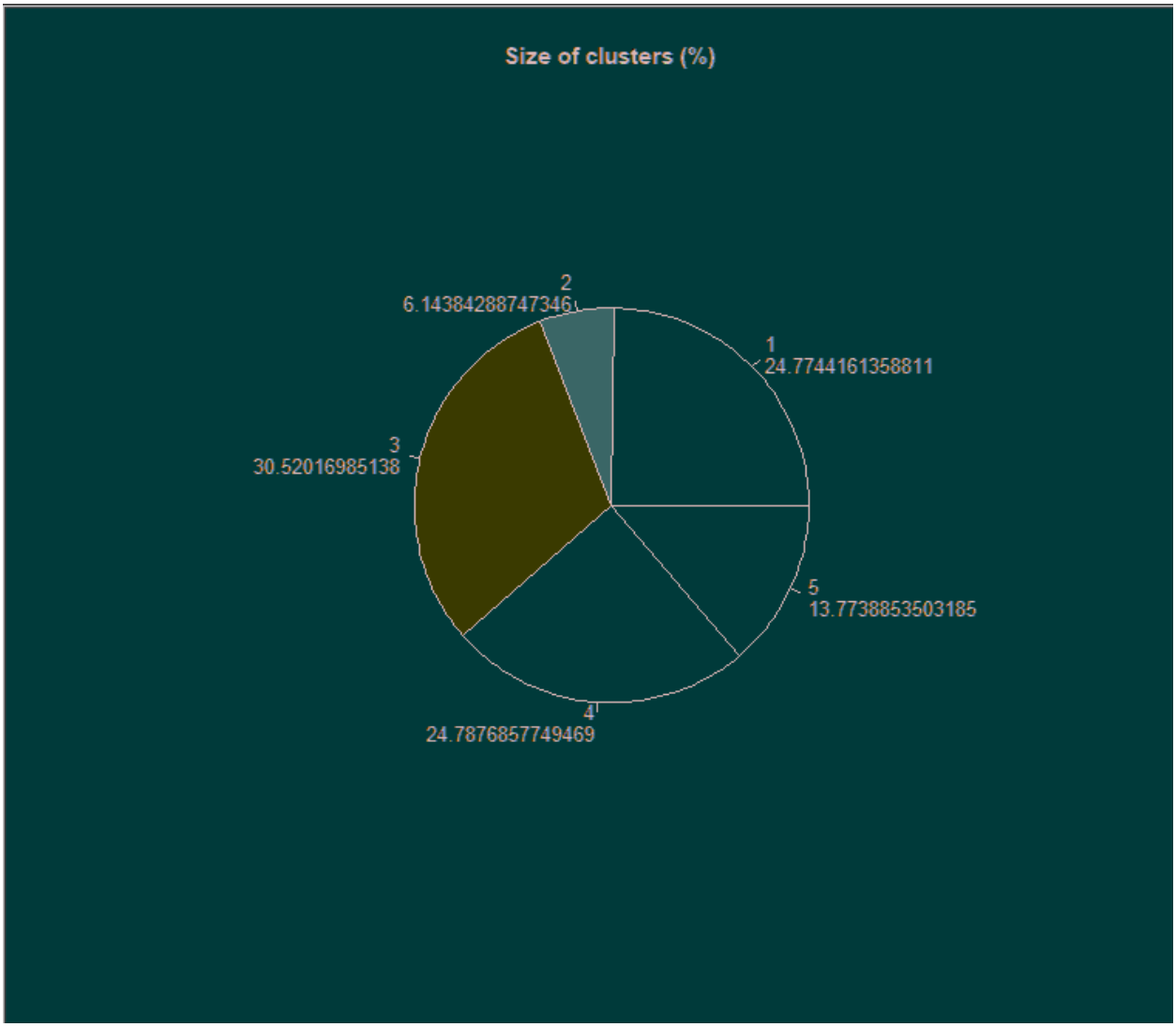
```
[1] "The mean of the denormalized centers is: "
         [,1]     [,2]
[1,] 38.15865 5.955635
[1] "The standard deviation of the denormalized centers is: "
         [,1]     [,2]
[1,] 28.41399 4.701825
```

Για την καλύτερη κατανόηση του αποτελέσματος, προχωράμε και πάλι στη δημιουργία διαγράμματος απεικόνισης συστάδων που περιλαμβάνει τα νέα στοιχεία. Για την καλύτερη συσχέτιση του διαγράμματος με τα αρχικά δεδομένα, η κλίμακα τιμών του διαγράμματος σχηματίστηκε σύμφωνα με τις αρχικές (μη κανονικοποιημένες τιμές):



Στο παραπάνω διάγραμμα, η **κίτρινη** βούλα αντιπροσωπεύει την **τυπική απόκλιση** των συστάδων, ενώ η **πράσινη** βούλα αντιπροσωπεύει την **μέση τιμή** των συστάδων.

Επιπρόσθετα, το μέγεθος κάθε ομάδας συναλλαγών επί του συνόλου των καταγεγραμμένων συναλλαγών αποτυπώνεται στο παρακάτω διάγραμμα:



Σε αυτό το σημείο προχωράμε στην ερμηνεία του **προφίλ** της κάθε ομάδας, χρησιμοποιώντας τα παραπάνω διαγράμματα και λαμβάνοντας υπόψιν τη **μέση τιμή** και την **τυπική απόκλιση** των δεδομένων:

- Ομάδα 1: Πρόκειται για ομάδα δεδομένων που αφορά **πρόσφατες συναλλαγές,** με **αρκετά υψηλή αξία**, οι οποίες αποτελούν περίπου το **24.78%** του συνόλου των συναλλαγών.

- Ομάδα 2: Πρόκειται για ομάδα δεδομένων που αφορά **παλαιές, μη πρόσφατες συναλλαγές**, με **υψηλή αξία**. Μάλιστα, σε αυτή την ομάδα ανήκουν οι συναλλαγές που έχουν τις υψηλότερες μέγιστες αξίες σε σχέση με τις μέγιστες αξίες των συναλλαγών των υπόλοιπων ομάδων. Οι συναλλαγές αυτής της ομάδας αποτελούν περίπου το **6.14%** του συνόλου των συναλλαγών.

- Ομάδα 3: Πρόκειται για ομάδα δεδομένων που αφορά **παλαιές, μη πρόσφατες συναλλαγές**, με **μικρή έως μεσαία αξία**. Οι συναλλαγές αυτής της ομάδας αποτελούν το **30.52%** του συνόλου των συναλλαγών.

- Ομάδα 4: Πρόκειται για ομάδα δεδομένων που αφορά συναλλαγές με χρονική εγγύτητα διεκπεραίωσης κοντά στον μέσο όρο του συνόλου συναλλαγών (**ούτε πολύ παλιές, ούτε πρόσφατες**). Παράλληλα, η αξία των συναλλαγών αυτής της ομάδας είναι **σχετικά χαμηλή**. Οι συναλλαγές αυτής της ομάδας αποτελούν περίπου το **24.79%** του συνόλου των συναλλαγών.

- Ομάδα 5: Πρόκειται για ομάδα δεδομένων που αφορά **πολύ πρόσφατες** συναλλαγές. Μάλιστα, σε αυτή την ομάδα ανήκουν οι συναλλαγές που έχουν τις πιο μικρότερες ελάχιστες τιμές recency_days σε σχέση με τις ελάχιστες τιμές recency_days των συναλλαγών των υπόλοιπων ομάδων. Παράλληλα, η αξία των συναλλαγών αυτής της ομάδας είναι **πολύ χαμηλή**. Μάλιστα, σε αυτή την ομάδα ανήκουν οι συναλλαγές που παρουσιάζουν την μικρότερη ελάχιστη αξία σε σχέση με την ελάχιστη αξία των συναλλαγών των υπόλοιπων ομάδων. Οι συναλλαγές αυτής της ομάδας αποτελούν περίπου το **13.77%** του συνόλου των συναλλαγών.

Διερμηνεύοντας τα προφίλ των ομάδων στο πλαίσιο της αποτίμησης κινδύνου, άξια προσοχής για την ομάδα Marketing της εταιρίας εμφανίζεται η **ομάδα 2**.

Το σύνολο των συναλλαγών που την αποτελούν έχουν αξία εμφανώς υψηλότερη του μέσου όρου του συνόλου συναλλαγών των δεδομένων. Αυτό καθιστά τους καταναλωτές αυτής της ομάδας ιδιαίτερα επικερδείς για την εταιρία.

Ωστόσο, το μεγάλο χρονικό διάστημα που έχει μεσολαβήσει από τις συναλλαγές αυτής της ομάδας καταναλωτών, φανερώνει την ανάγκη για περαιτέρω διερεύνηση των συνθηκών κάτω από τις οποίες πραγματοποιήθηκαν (π.χ. περίοδος εορτών, τουριστική περίοδος κ.λπ.), προκειμένου να καθοριστεί με ασφάλεια από το τμήμα Marketing αν οι ιδιαίτερες συνθήκες που προκάλεσαν την εκδήλωσή τους αποτελούν φυσικά αίτια ή συνέπεια ενός πετυχημένου επιχειρηματικού πλάνου που εφαρμόστηκε από την εταιρεία. Στην τελευταία περίπτωση, το συγκεκριμένο πλάνο πρέπει να εξεταστεί αναλυτικά, να αναπροσαρμοστεί στις τρέχουσες συνθήκες της αγοράς και να τεθεί το συντομότερο ξανά σε εφαρμογή.

2. Εξαγωγή στοιχείων ανάθεσης συναλλαγών σε ομάδες και παραγωγή τελικής μορφής δεδομένων

Η προσθήκη της πληροφορίας ανάθεσης σε ομάδα στα υπάρχοντα δεδομένα συναλλαγών γίνεται εντός της συνάρτησης **generateGroceriesWithBinaryClusterData(groceriesDiscrete, kmeansFit)**.

Εντός της συνάρτησης, ακολουθείται διαδικασία παραγωγής δυαδικής αναπαράστασης της πληροφορίας που αφορά την ανάθεση της κάθε συναλλαγής σε ομάδα. Η διαδικασία είναι όμοια με αυτή που έχει ήδη περιγραφεί αναλυτικά κατά την αρχική επεξεργασία των δεδομένων συναλλαγών για την μετατροπή της απεικόνισής τους σε δυαδική μορφή, συνεπώς για λόγους συντομίας παραλείπεται η αναλυτική εξήγησή της.

Τελικά, η συνάρτηση **generateGroceriesWithBinaryClusterData()** επιστρέφει τη μεταβλητή **groceriesWithClusters**, η οποία περιέχει τα δεδομένα συναλλαγών στη μορφή που βρίσκονταν εντός της **groceriesDiscrete**, με την προσθήκη 5 ιδιοτήτων (στηλών): **«cluster1»**, **«cluster2»**, **«cluster3»**, **«cluster4»**, **«cluster5»**.

Ανάλογα με την ομάδα στην οποία ανήκει κάθε συναλλαγή, η λογική τιμή **«TRUE»** καταγράφεται σε 1 από τις 5 στήλες, με την τιμή **«FALSE»** να καταγράφεται στις υπόλοιπες.

Άσκηση 4

Σε αυτό το στάδιο ζητείται να πραγματοποιήσουμε καταγραφή των 20 κανόνων με το υψηλότερο confidence **αποκλειστικά** για τα **προϊόντα** και τις **ομάδες συναλλαγών**.

Για τον σκοπό αυτό έχει δημιουργηθεί η συνάρτηση clusterProductProfile(groceriesWithClusters) εντός της οποίας πραγματοποιείται η εξαγωγή των κανόνων συσχετίσεων με χρήστη της ενσωματωμένης συνάρτησης **apriori()** καθώς και η τύπωση των αποτελεσμάτων.

Η λογική πίσω από την επιλογή της τιμής του ελάχιστου support στηρίζεται στο γεγονός ότι θέλουμε ιδανικά την υψηλότερη δυνατή τιμή του ορίου ελαχίστου support, έτσι ώστε οι παραγόμενοι κανόνες συσχέτισης να αφορούν όσο το δυνατόν μεγαλύτερους (και άρα περισσότερο αντιπροσωπευτικούς) πληθυσμούς του δείγματος. Παράλληλα, μας ενδιαφέρει η εμπιστοσύνη των εξαγόμενων κανόνων, η οποία θέλουμε να είναι όσο το δυνατόν μεγαλύτερη έτσι ώστε μπορούμε να επικαλεστούμε τους εξαγόμενους κανόνες ως βάση συσχέτισης της συμπεριφοράς των καταναλωτών που αποτυπώνεται στο σύνολο δεδομένων μας. Με δεδομένο ωστόσο πως η αύξηση του support τείνει να οδηγεί σε μείωση του confidence και το αντίστροφο, είναι αναγκαίο να εκτελέσουμε δοκιμές με διάφορες τιμές παραμέτρων για το ελάχιστο support και το ελάχιστο confidence, αξιολογώντας σε κάθε εκτέλεση τις ποιοτικές μεταβολές των παραγόμενων κανόνων συσχέτισης και καθορίζοντας έτσι τη «χρυσή τομή» μεταξύ των 2 αυτών παραμέτρων.

Στην περίπτωση μας, καταλήγουμε πειραματικά στην επιλογή της τιμής **0.01** για το **ελάχιστο support** και της τιμής **0.4** για το **ελάχιστο confidence.**

Επιπρόσθετα, δίνουμε ως παράμετρο για το όρισμα «**appearance**» της **apriori()** το:

```
appearance = list (default="lhs",rhs=c("cluster1", "cluster2", "cluster3", "cluster4", "cluster5"))
```

Με την συγκεκριμένη παράμετρο ορίζουμε ότι θέλουμε να λάβουμε μόνο τους κανόνες συσχέτισης που έχουν στο δεξί μέρος του κανόνα (right hand side – rhs) τις τιμές «cluster1», «cluster2», «cluster3», «cluster4», «cluster5». Με αυτόν τον τρόπο, «φιλτράρουμε» το τελικό αποτέλεσμα από κανόνες συσχέτισης που μπορεί να έχουν τιμές πέρα από αυτές που αφορούν την ομάδα συναλλαγών στο δεξί τους μέρος και που ως αποτέλεσμα δεν παρουσιάζουν ενδιαφέρον για την ανάλυσή μας.

Το αποτέλεσμα εκτέλεσης της εντολής **apriori()** είναι το παρακάτω:

```
[1] "Top 20 product/cluster rules by Confidence: "
     lhs                                                  rhs        support    confidence lift      coun
[1]  {sausage,pastry}                                  => {cluster2} 0.01632166 1.0000000  16.276458 123
[2]  {whole milk,yogurt,pastry}                        => {cluster2} 0.01074841 0.9000000  14.648812  81
[3]  {whole milk,other vegetables,pastry}              => {cluster2} 0.01180998 0.8557692  13.928892  89
[4]  {tropical fruit,pastry}                           => {cluster2} 0.01472930 0.8538462  13.897591 111
[5]  {rolls/buns,sausage}                              => {cluster1} 0.03304140 0.8272425   3.339100 249
[6]  {rolls/buns,sausage,soda}                         => {cluster1} 0.01035032 0.8210526   3.314115  78
[7]  {yogurt,pastry}                                   => {cluster2} 0.01831210 0.7931034  12.908915 138
[8]  {sausage,soda}                                    => {cluster1} 0.02468153 0.7782427   3.141316 186
[9]  {root vegetables,pastry}                          => {cluster2} 0.01114650 0.7777778  12.659467  84
[10] {bottled water,sausage}                           => {cluster1} 0.01207537 0.7711864   3.112834  91
[11] {other vegetables,sausage}                        => {cluster1} 0.02693737 0.7660377   3.092052 203
[12] {citrus fruit,sausage}                            => {cluster1} 0.01127919 0.7657658   3.090954  85
[13] {tropical fruit,other vegetables,root vegetables} => {cluster1} 0.01194268 0.7438017   3.002297  90
[14] {other vegetables,pastry}                         => {cluster2} 0.02176221 0.7387387  12.024050 164
[15] {rolls/buns,pastry}                               => {cluster2} 0.02016985 0.7378641  12.009814 152
[16] {sausage,root vegetables}                         => {cluster1} 0.01433121 0.7346939   2.965535 108
[17] {whole milk,sausage}                              => {cluster1} 0.02773355 0.7108844   2.869429 209
[18] {sausage}                                         => {cluster1} 0.08598726 0.7012987   2.830738 648
[19] {other vegetables,rolls/buns,root vegetables}     => {cluster1} 0.01114650 0.7000000   2.825495  84
[20] {tropical fruit,whole milk,root vegetables}       => {cluster1} 0.01074841 0.6864407   2.770764  81
```

Ερμηνεύοντας το παραπάνω αποτέλεσμα, αξίζει να αναφερθούμε συνοπτικά στα εξής:

- Οι πρώτοι 20 κανόνες σύμφωνα με το confidence αφορούν αποκλειστικά τις **ομάδες συναλλαγών 1 και 2** (cluster1, cluster2), πράγμα από το οποίο προκύπτει ότι για αυτές τις ομάδες συναλλαγών έχουμε πολύ **υψηλή εμπιστοσύνη** αναφορικά με τις συσχετίσεις προϊόντων που τις χαρακτηρίζουν

- Παράλληλα, η απουσία κανόνων συσχέτισης που αφορούν τις ομάδες συναλλαγών 3, 4 και 5 από τη λίστα των 20 κορυφαίων κανόνων σύμφωνα με το confidence (ελάχιστο confidence = 0.7), καθιστά φανερό πως οι κανόνες συσχέτισης των συγκεκριμένων ομάδων συναλλαγών παρουσιάζουν χαμηλή εμπιστοσύνη και συνεπώς προσφέρουν περιορισμένη δυνατότητα εξαγωγής γενικευμένων συμπερασμάτων ως προς το προφίλ συναλλαγών.

- Από τον κανόνα συσχέτισης [1], προκύπτει πως αν τα προϊόντα «**sausage**» και «**pastry**» αγοραστούν μαζί, τότε με απόλυτη βεβαιότητα η συναλλαγή θα ανήκει στην **ομάδα 2** (cluster2).

Στη συνέχεια, μας ζητείται να εντοπίσουμε ποια προϊόντα/συνδυασμοί προϊόντων αγοράζονται συχνότερα από την κάθε ομάδα. Η διαδικασία εξαγωγής των εστιασμένων κανόνων συσχέτισης για το παραπάνω ζητούμενο εκτελείται και πάλι από την συνάρτηση **clusterProductProfile(groceriesWithClusters)**.

Ακολουθεί ανάλυση των 20 κανόνων συσχέτισης σύμφωνα με το confidence που εξάγονται πλέον σε επίπεδο μεμονωμένων ομάδων συναλλαγών (cluster). Η τιμή της παραμέτρου ελάχιστου support επιλέχθηκε πειραματικά, σύμφωνα με τη λογική που έχει ήδη περιγράφει σε προηγούμενες εκτελέσεις του αλγορίθμου apriori.

Για την **ομάδα συναλλαγών 1** (cluster1, supp=0.001):

```
[1] "Top 20 rules by Confidence for *** Cluster 1 ***: "
     lhs                                                                  rhs        support     confidence lift     count
[1]  {citrus fruit,rolls/buns,bottled water,yogurt}                       => {cluster1} 0.001592357 0.9230769  3.725928  12
[2]  {citrus fruit,tropical fruit,bottled water,root vegetables}          => {cluster1} 0.001459660 0.9166667  3.700054  11
[3]  {citrus fruit,other vegetables,rolls/buns,bottled water}             => {cluster1} 0.001194268 0.9000000  3.632780   9
[4]  {citrus fruit,tropical fruit,bottled water,soda}                     => {cluster1} 0.001061571 0.8888889  3.587931   8
[5]  {citrus fruit,bottled water,yogurt,root vegetables}                  => {cluster1} 0.001061571 0.8888889  3.587931   8
[6]  {citrus fruit,whole milk,bottled water,yogurt}                       => {cluster1} 0.002123142 0.8888889  3.587931  16
[7]  {citrus fruit,tropical fruit,whole milk,bottled water,yogurt}        => {cluster1} 0.001061571 0.8888889  3.587931   8
[8]  {citrus fruit,bottled water,yogurt}                                  => {cluster1} 0.004644374 0.8750000  3.531869  35
[9]  {citrus fruit,other vegetables,bottled water,yogurt}                 => {cluster1} 0.001857749 0.8750000  3.531869  14
[10] {citrus fruit,whole milk,other vegetables,bottled water,root vegetables} => {cluster1} 0.001459660 0.8461538  3.415434  11
[11] {rolls/buns,sausage}                                                 => {cluster1} 0.033041401 0.8272425  3.339100 249
[12] {tropical fruit,bottled water,root vegetables}                       => {cluster1} 0.004909766 0.8222222  3.318836  37
[13] {rolls/buns,sausage,soda}                                            => {cluster1} 0.010350318 0.8210526  3.314115  78
[14] {citrus fruit,tropical fruit,bottled water,yogurt}                   => {cluster1} 0.001725053 0.8125000  3.279593  13
[15] {citrus fruit,rolls/buns,bottled water}                              => {cluster1} 0.003317410 0.8064516  3.255179  25
[16] {chocolate,bottled water,root vegetables}                            => {cluster1} 0.001061571 0.8000000  3.229138   8
[17] {citrus fruit,bottled water,sausage}                                 => {cluster1} 0.002123142 0.8000000  3.229138  16
[18] {citrus fruit,other vegetables,bottled water,sausage}                => {cluster1} 0.001061571 0.8000000  3.229138   8
[19] {citrus fruit,whole milk,bottled water,root vegetables}              => {cluster1} 0.002123142 0.8000000  3.229138  16
[20] {tropical fruit,whole milk,bottled water,root vegetables}            => {cluster1} 0.002653928 0.8000000  3.229138  20
```

Παρατηρούμε πολύ υψηλό confidence (~0.92) στον κανόνα συσχέτισης **[1]** σύμφωνα με τον οποίο οι πελάτες που αγοράζουν τον συνδυασμό προϊόντων: **citrus fruit, rolls/buns, bottled water, yogurt** ανήκουν στην **ομάδα 1** στο 92% των περιπτώσεων. Αν και το support του συγκεκριμένου κανόνα συσχέτισης δεν είναι πολύ υψηλό (~0.0016), συνεπώς η συχνότητα αυτής της καταναλωτικής συμπεριφοράς δεν είναι αρκετά υψηλή, η βεβαιότητα εμφάνισής της στο πλαίσιο του συνδυασμού των προαναφερθέντων προϊόντων είναι πολύ μεγάλη.

Επιπρόσθετα, παρατηρούμε και τον κανόνα συσχέτισης **[11]**, ο οποίος εμφανίζει ένα αρκετά υψηλό confidence (~0.83), που αφορά την αγορά του συνδυασμού προϊόντων **rolls/buns, sausage** από την **ομάδα 1** στο 83% των περιπτώσεων.

Σε αυτή την περίπτωση, αν και η βεβαιότητα εμφάνισης αυτής της καταναλωτικής συμπεριφοράς δεν είναι το ίδιο υψηλή με αυτή του κανόνα **[1]**, η υψηλή τιμή του support (~ 0.033) σε σχέση με τους υπόλοιπους κανόνες και συνεπώς η μεγάλη συχνότητα εμφάνισής της συγκεκριμένης καταναλωτικής συμπεριφοράς την καθιστά άξια προσοχής στην εξαγωγή συμπερασμάτων για την αγοραστική τάση της ομάδας 1.

Για την **ομάδα συναλλαγών 2** (cluster2, supp=0.0055):

```
[1] "Top 20 product and value category rules by Confidence for *** Cluster 2 ***: "
     lhs                                        rhs        support     confidence lift     count
[1]  {sausage,pastry}                        => {cluster2} 0.016321656 1.0000000  16.27646 123
[2]  {yogurt,sausage,pastry}                 => {cluster2} 0.005573248 1.0000000  16.27646  42
[3]  {whole milk,sausage,pastry}             => {cluster2} 0.007430998 1.0000000  16.27646  56
[4]  {rolls/buns,yogurt,pastry}              => {cluster2} 0.007563694 1.0000000  16.27646  57
[5]  {rolls/buns,pastry,soda}                => {cluster2} 0.007032909 1.0000000  16.27646  53
[6]  {tropical fruit,other vegetables,pastry} => {cluster2} 0.006369427 0.9600000  15.62540  48
[7]  {tropical fruit,yogurt,pastry}          => {cluster2} 0.005838641 0.9565217  15.56879  44
[8]  {whole milk,root vegetables,pastry}     => {cluster2} 0.007032909 0.9464286  15.40450  53
[9]  {other vegetables,yogurt,pastry}        => {cluster2} 0.008094480 0.9384615  15.27483  61
[10] {other vegetables,pastry,soda}          => {cluster2} 0.006634820 0.9259259  15.07079  50
[11] {yogurt,pastry,soda}                    => {cluster2} 0.005971338 0.9183673  14.94777  45
[12] {tropical fruit,whole milk,pastry}      => {cluster2} 0.007961783 0.9090909  14.79678  60
[13] {whole milk,yogurt,pastry}              => {cluster2} 0.010748408 0.9000000  14.64881  81
[14] {other vegetables,rolls/buns,pastry}    => {cluster2} 0.007032909 0.8833333  14.37754  53
[15] {other vegetables,root vegetables,pastry} => {cluster2} 0.006767516 0.8793103  14.31206  51
[16] {chocolate,pastry}                      => {cluster2} 0.009156051 0.8734177  14.21615  69
[17] {whole milk,rolls/buns,pastry}          => {cluster2} 0.009554140 0.8571429  13.95125  72
[18] {whole milk,other vegetables,pastry}    => {cluster2} 0.011809979 0.8557692  13.92889  89
[19] {tropical fruit,pastry}                 => {cluster2} 0.014729299 0.8538462  13.89759 111
[20] {whole milk,pastry,soda}                => {cluster2} 0.009023355 0.8395062  13.66419  68
```

Παρατηρούμε απόλυτο confidence (=1) στον κανόνα συσχέτισης [1] σύμφωνα με τον οποίο ο συνδυασμός προϊόντων **sausage,pastry** αγοράζεται πάντα από πελάτες που ανήκουν στην **ομάδα 2**. Μάλιστα, το υψηλό support (~ 0.016) του κανόνα σε σχέση με τους υπόλοιπους κανόνες του αποτελέσματος φανερώνει μια αρκετά συχνή καταναλωτική συμπεριφορά, για την οποία μάλιστα έχουμε απόλυτο μέτρο συσχέτισης με το είδος καταναλωτών που την πραγματοποιούν. Συνεπώς, τα συμπεράσματα που εξάγουμε από αυτό τον κανόνα συσχέτισης είναι ιδιαίτερα ενδιαφέροντα.

Στο ίδιο πλαίσιο, οι κανόνες **[2]**, **[3]**, **[4]** και **[5]** εμφανίζουν απόλυτο confidence (=1) αναφορικά με την αγορά των συνδυασμών προϊόντων τους από την **ομάδα 2**. Αν και το support των προαναφερθέντων κανόνων δεν είναι όσο υψηλό είναι το support του κανόνα **[1]**, φανερώνουν παρόλα αυτά ενδιαφέροντα μοτίβα καταναλωτικής συμπεριφοράς για την **ομάδα 2**, και μάλιστα σε απόλυτη εξάρτηση από αυτή.

Γενικά, και οι 20 κορυφαίοι κανόνες συσχέτισης συνδυασμών προιόντων αναφορικά με τις αγορές της ομάδας 2 εμφανίζουν υψηλό confidence (> 0.83) και ικανοποιητικό support (δεν έχουμε δηλαδή υπερβολικά χαμηλές τιμές support που φανερώνουν μεμονωμένες περιπτώσεις). Συνεπώς, αν και για λόγους συντομίας στην παρούσα ανάλυση εστιάσαμε στους κανόνες **[1] – [5]**, μπορούν να αξιοποιηθούν στο σύνολο τους για την εξαγωγή χρήσιμων συμπερασμάτων αναφορικά με την καταναλωτική συμπεριφορά της ομάδας 2.

Για την **ομάδα συναλλαγών 3** (cluster3, supp=0.005, conf=0.1):

```
[1] "Top 20 rules by Confidence for *** Cluster 3 ***: "
     lhs                                rhs        support     confidence lift      count
[1]  {pastry}                        => {cluster3} 0.061173036 0.5268571  1.7262589 461
[2]  {whole milk,pastry}             => {cluster3} 0.015525478 0.3577982  1.1723335 117
[3]  {soda}                          => {cluster3} 0.078290870 0.3440233  1.1271999 590
[4]  {pastry,soda}                   => {cluster3} 0.009023355 0.3285024  1.0763453  68
[5]  {rolls/buns}                    => {cluster3} 0.067940552 0.2830293  0.9273516 512
[6]  {rolls/buns,pastry}             => {cluster3} 0.007165605 0.2621359  0.8588940  54
[7]  {other vegetables,pastry}       => {cluster3} 0.007696391 0.2612613  0.8560282  58
[8]  {whole milk,bottled water}      => {cluster3} 0.010881104 0.2426036  0.7948958  82
[9]  {yogurt}                        => {cluster3} 0.042595541 0.2339650  0.7665915 321
[10] {whole milk,other vegetables}   => {cluster3} 0.020302548 0.2078804  0.6811248 153
[11] {citrus fruit,other vegetables} => {cluster3} 0.007165605 0.1901408  0.6230006  54
[12] {citrus fruit}                  => {cluster3} 0.020435244 0.1891892  0.6198825 154
[13] {root vegetables}               => {cluster3} 0.025875796 0.1819030  0.5960091 195
[14] {tropical fruit}                => {cluster3} 0.024814225 0.1812016  0.5937108 187
[15] {citrus fruit,whole milk}       => {cluster3} 0.007165605 0.1800000  0.5897739  54
[16] {whole milk}                    => {cluster3} 0.055997877 0.1679268  0.5502157 422
[17] {other vegetables,bottled water} => {cluster3} 0.005307856 0.1639344  0.5371347  40
[18] {whole milk,yogurt}             => {cluster3} 0.011677282 0.1597096  0.5232920  88
[19] {other vegetables}              => {cluster3} 0.036624204 0.1450342  0.4752076 276
[20] {bottled water}                 => {cluster3} 0.020833333 0.1444342  0.4732419 157
```

Παρατηρούμε ότι για το προϊόν «**pastry**», υπάρχει **~ 53%** πιθανότητα (confidence ~ 0.526) να αγοραστεί από πελάτη που ανήκει στην **ομάδα 3**.

Με την ίδια λογική, προκύπτουν τα εξής συμπεράσματα:

- Ο συνδυασμός προϊόντων «**whole milk, pastry**» έχει πιθανότητα **~ 36%** να αγοραστεί από πελάτες που ανήκουν στην ομάδα 3 (κανόνας **[2]**).

- Το προϊόν «**soda**» έχει πιθανότητα **~ 34%** να αγοραστεί από πελάτες που ανήκουν στην ομάδα 3 (κανόνας **[3]**).

- Ο συνδυασμός προϊόντων «pastry, soda» έχει πιθανότητα **~ 33%** να αγοραστεί από πελάτες που ανήκουν στην ομάδα 3 (κανόνας **[4]**).

Όσον αφορά τους κανόνες [5] – [20], παρατηρώντας το **lift** τους το οποίο κυμαίνεται σε τιμές **< 1**, καταλήγουμε στο συμπέρασμα ότι η παρουσία της ομάδας 3 (**cluster3**) στο δεξί μέρος του κανόνα συσχέτισης δεν αυξάνει την πιθανότητα εμφάνισης των συναλλαγών που ορίζονται από τους συνδυασμούς προϊόντων του αριστερού μέρους των κανόνων αυτών. Συνεπώς, η ισχύς της συσχέτισης αριστερού και δεξιού μέρους για αυτούς τους κανόνες είναι πολύ χαμηλή για να εξάγουμε συμπεράσματα από τους τελευταίους στην ανάλυσή μας.

Για την **ομάδα συναλλαγών 4** (cluster4, supp=0.01, conf=0.2):

```
[1] "Top 20 rules by Confidence for *** Cluster 4 ***: "
     lhs                                rhs        support    confidence lift      count
[1]  {bottled water,soda}               => {cluster4} 0.01366773 0.3614035  1.4579962 103
[2]  {chocolate}                        => {cluster4} 0.02162951 0.3340164  1.3475094 163
[3]  {whole milk,rolls/buns}            => {cluster4} 0.02335456 0.3159785  1.2747396 176
[4]  {whole milk,root vegetables}       => {cluster4} 0.01857749 0.2910603  1.1742133 140
[5]  {root vegetables}                  => {cluster4} 0.04087049 0.2873134  1.1590974 308
[6]  {whole milk,soda}                  => {cluster4} 0.01459660 0.2791878  1.1263166 110
[7]  {soda}                             => {cluster4} 0.06050955 0.2658892  1.0726665 456
[8]  {bottled water}                    => {cluster4} 0.03834926 0.2658694  1.0725865 289
[9]  {rolls/buns}                       => {cluster4} 0.06157113 0.2564953  1.0347691 464
[10] {other vegetables,root vegetables} => {cluster4} 0.01579087 0.2553648  1.0302083 119
[11] {rolls/buns,soda}                  => {cluster4} 0.01260616 0.2519894  1.0165910  95
[12] {tropical fruit}                   => {cluster4} 0.03436837 0.2509690  1.0124745 259
[13] {other vegetables}                 => {cluster4} 0.06024416 0.2385707  0.9624564 454
[14] {other vegetables,yogurt}          => {cluster4} 0.01340234 0.2365340  0.9542398 101
[15] {yogurt}                           => {cluster4} 0.04219745 0.2317784  0.9350547 318
[16] {whole milk}                       => {cluster4} 0.07722930 0.2315957  0.9343176 582
[17] {other vegetables,rolls/buns}      => {cluster4} 0.01273885 0.2291169  0.9243176  96
[18] {tropical fruit,other vegetables}  => {cluster4} 0.01008493 0.2152975  0.8685662  76
[19] {tropical fruit,whole milk}        => {cluster4} 0.01154459 0.2091346  0.8437037  87
[20] {citrus fruit}                     => {cluster4} 0.02162951 0.2002457  0.8078435 163
```

Λαμβάνοντας υπόψιν την εμπιστοσύνη (confidence) των αποτελεσμάτων, για τους κανόνες **[1] – [12]** παρατηρούμε ότι:

- Ο συνδυασμός προϊόντων «**bottled water, soda**» έχει πιθανότητα **~ 36%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[1]**).

- Το προϊόν «**chocolate**» έχει πιθανότητα **~ 33%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[2]**).

- Ο συνδυασμός προϊόντων «**whole milk, rolls/buns**» έχει πιθανότητα **~ 32%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[3]**).

- Ο συνδυασμός προϊόντων «**whole milk, root vegetables**» έχει πιθανότητα **~ 29%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[4]**).

- Το προϊόν «**root vegetables**» έχει πιθανότητα **~ 29%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[5]**).

- Ο συνδυασμός προϊόντων «**whole milk, soda**» έχει πιθανότητα **~ 28%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[6]**).

- Το προϊόν «**soda**» έχει πιθανότητα **~ 27%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[7]**).

- Το προϊόν «**bottled water**» έχει πιθανότητα **~ 27%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[8]**).

- Το προϊόν «**rolls/buns**» έχει πιθανότητα **~ 26%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[9]**).

- Ο συνδυασμός προϊόντων «**other vegetables, root vegetables**» έχει πιθανότητα **~ 26%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[10]**).

- Ο συνδυασμός προϊόντων «**rolls/buns, soda**» έχει πιθανότητα **~ 25%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[11]**).

- Το προϊόν «**tropical fruit**» έχει πιθανότητα **~ 25%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 4** (κανόνας **[12]**).

Όσον αφορά τους κανόνες συσχέτισης [13] – [20], παρατηρούμε ότι η τιμή του **lift** για κάθε έναν από αυτούς είναι **< 1**. Αυτό σημαίνει ότι η παρουσία της ομάδας 4 (**cluster4**) στο δεξί μέρος του κανόνα συσχέτισης δεν αυξάνει την πιθανότητα εμφάνισης των συναλλαγών που ορίζονται από τους συνδυασμούς προϊόντων του αριστερού μέρους των κανόνων αυτών. Συνεπώς, η ισχύς της συσχέτισης αριστερού και δεξιού μέρους για αυτούς τους κανόνες είναι πολύ χαμηλή για να εξάγουμε συμπεράσματα από τους τελευταίους στην ανάλυσή μας.

Για την **ομάδα συναλλαγών 5** (cluster5, supp=0.001, conf=0.1):

```
[1] "Top 20 rules by Confidence for *** Cluster 5 ***: "
    lhs                   rhs        support    confidence lift      count
[1] {bottled water}    => {cluster5} 0.02720276 0.1885925  1.3692030 205
[2] {whole milk}       => {cluster5} 0.05878450 0.1762833  1.2798373 443
[3] {other vegetables} => {cluster5} 0.03529724 0.1397793  1.0148138 266
[4] {citrus fruit}     => {cluster5} 0.01326964 0.1228501  0.8919061 100
```

Λαμβάνοντας υπόψιν την εμπιστοσύνη (confidence) των αποτελεσμάτων, για τους κανόνες **[1] – [3]** παρατηρούμε ότι:

- Το προϊόν «**bottled water**» έχει πιθανότητα **~ 19%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 5** (κανόνας **[1]**).

- Το προϊόν «**whole milk**» έχει πιθανότητα **~ 18%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 5** (κανόνας **[2]**).

- Το προϊόν «**other vegetables**» έχει πιθανότητα **~ 14%** να αγοραστεί από πελάτες που ανήκουν στην **ομάδα 5** (κανόνας **[3]**).

Όσον αφορά τον κανόνα συσχέτισης **[4]**, παρατηρούμε ότι η τιμή του **lift** του είναι **< 1** (~ 0.89). Αυτό σημαίνει ότι η παρουσία της ομάδας 5 (**cluster5**) στο δεξί μέρος του κανόνα συσχέτισης δεν αυξάνει την πιθανότητα εμφάνισης της συναλλαγής που ορίζεται από το προϊόν «**citrus fruit**» του αριστερού μέρους του κανόνα. Συνεπώς, η ισχύς της συσχέτισης αριστερού και δεξιού μέρους για αυτόν τον κανόνα είναι πολύ χαμηλή για να εξάγουμε συμπεράσματα που αφορούν την ανάλυση μας από αυτόν.

Στην ανάλυση που προηγήθηκε για την Άσκηση 3, καταλήξαμε στον προσδιορισμό της **ομάδας 2** ως «ανησυχητικής». Οδηγηθήκαμε σε αυτή την κατηγοριοποίηση λόγω του ότι αυτή η ομάδα συναλλαγών περιέχει μεν συναλλαγές υψηλότερης αξίας σε σχέση με τον μέσο όρο αξίας συναλλαγών των δεδομένων μας, ωστόσο έχουν παρέλθει περισσότερες μέρες από την πραγματοποίησή τους (**recency**) σε σχέση με τον μέσο όρο ημερών που έχουν παρέλθει από την πραγματοποίηση του συνόλου των συναλλαγών, όπως αυτός προκύπτει από το σύνολο των δεδομένων μας.

Όπως αναφέρθηκε λοιπόν ως συμπέρασμα και στην Άσκηση 3, είναι χρήσιμο και επιθυμητό να προσδιοριστεί τι προκαλεί αυτή την χρονική καθυστέρηση στην πραγματοποίηση των συναλλαγών της **ομάδας 2**.

Κατά την μελέτη των κανόνων συσχετίσεων που προέκυψαν από την εφαρμογή του αλγορίθμου **apriori** για τους συνδυασμούς προϊόντων και την ομάδα συναλλαγών 2 (άσκηση 3), ιδιαίτερο ενδιαφέρον παρουσίασε ο κανόνας **[1]**:

{sausage, pastry} => {cluster2}

Ο συγκεκριμένος κανόνας, πέρα από το απόλυτο **Confidence** (**= 1**) που παρουσιάζει και το οποίο εγγυάται με απόλυτη βεβαιότητα ότι ο συνδυασμός προϊόντων «**sausage, pastry**» αγοράζεται στο **100%** των περιπτώσεων από καταναλωτές που ανήκουν στην ομάδα συναλλαγών 2, έχει και το υψηλότερο Support (**~ 0.016**) μεταξύ των 20 κορυφαίων κανόνων κατά Confidence που μελετήσαμε για την ομάδα 2 στην άσκηση 3.

Λαμβάνοντας υπόψιν το είδος των προϊόντων στα οποία αναφερόμαστε, η υψηλή καθυστέρηση επανάληψης συναλλαγών για την ομάδα 2 μπορεί να αιτιολογηθεί. Συγκεκριμένα, τόσο τα λουκάνικα (**sausage**) όσο και η ζύμη (**pastry**) είναι προϊόντα μακράς διαρκείας, υπό την έννοια ότι πολύ συχνά αποθηκεύονται σε μονάδες μακράς διαρκείας (π.χ. ψυγεία) και αξιοποιούνται μετά από αρκετό διάστημα. Συνεπώς, οι καταναλωτές της **ομάδας 2**, οι οποίοι φανερώνουν απόλυτο **Confidence** και το υψηλότερο **Support** (μεταξύ των 20 κανόνων που μελετήσαμε για την ομάδα 2) για τον συνδυασμό προϊόντων «**sausage, pastry**», είναι αναμενόμενο να διατηρούν τα προϊόντα που προμηθεύονται για μεγάλο διάστημα σε αποθήκευση (μονάδες μακράς διαρκείας), και συνεπώς να καθυστερούν τα επισκεφθούν ξανά το κατάστημα (υψηλή τιμή recency_days) για την αναπλήρωση των αποθεμάτων τους όσον αφορά αυτά τα προϊόντα.

Συμπερασματικά λοιπόν, μπορούμε πλέον να απαντήσουμε στο ερώτημα που θέσαμε στην άσκηση 3 αναφορικά με το αν τα μεγάλα διαστήματα που μεσολαβούν μεταξύ των συναλλαγών της ομάδας 2

οφείλονται σε φυσικά αίτια ή κάποια μεταβολή του επιχειρηματικού πλάνου της επιχείρησης (π.χ. αύξηση τιμών). Στη συγκεκριμένη περίπτωση, το επικρατέστερο σενάριο είναι ότι η φύση των προϊόντων με τα οποία έχει την ισχυρότερη συσχέτιση η ομάδα 2, δηλαδή τα **λουκάνικα** και η **ζύμη**, συνεπάγεται σε πολλές περιπτώσεις την μακρά αποθήκευση τους από τους καταναλωτές και άρα την σπανιότερη προμήθειά τους. Έτσι, το μεγάλο χρονικό διάστημα που μεσολαβεί μεταξύ των συναλλαγών είναι αναμενόμενο και προερχόμενο από φυσική αιτία.