Επιχειρηματική Ευφυΐα και Επιχειρησιακή Έρευνα

Εργασία 1η : Εξόρυξη γνώσης από δεδομένα συναλλαγών καταστήματος λιανικής

Λυσσούδη Πασχαλίνα – ΑΕΜ 3116

Άσκηση 1. Μετασχηματισμός πρωτογενών δεδομένων

Πρώτο βήμα πριν ξεκινήσει η ενασχόληση με τα ζητούμενα της άσκησης, ήταν η λήψη του αρχείου GroceriesInitial.csv . Το αρχείο κατεβαίνει στον φάκελο Downloads του σκληρού δίσκου, οπότε για να μπορεί να ανοίξει μέσω του αρχείου .r που περιέχει τον κώδικα, τοποθετήθηκε μέσα στον εν λόγω φάκελο. Από το αρχείο της .r οι πρώτες εντολές χρησιμεύουν στον προσδιορισμό του φακέλου στον οποίο βρισκόμαστε αρχικά, και την μετακίνηση στον φάκελο που μας ενδιαφέρει.

Σημείωση: Το path που αναφέρεται ανταποκρίνεται στον υπολογιστή όπου εκπονήθηκε η εργασία, για χρήση σε άλλο υπολογιστή θα πρέπει να αλλάξει στο path για τον φάκελο που περιέχει τα αρχεία της εργασίας.

Αφού ανοίξουμε και φορτώσουμε το αρχείο GroceriesInitial.csv στο dataframe original, δημιουργούμε 3 καινούριες στήλες για τον διαχωρισμό της τιμής του καλαθιού σε low, medium και high basket value. Τα χαρακτηριστικά αυτά είναι μορφής logical.

Με την συνάρτηση summary βλέπουμε στοιχεία για τις τιμές του basket value. Παρατηρούμε ότι κυριαρχούν μικρές τιμές, με μέση τιμή το 4.1 και ανώτατη το 25.1 . Με αυτά τα δεδομένα, αποφασίζουμε να ορίσουμε ως low basket value τιμές μέχρι 4, medium τιμές μέχρι 10, και τις διψήφιες τιμές τις θεωρούμε high basket value.

Στη συνέχεια με μια for διατρέχουμε όλες τις εγγραφές, και για κάθε εγγραφή, ανάλογα με το basket value της, κάνουμε True το αντίστοιχο χαρακτηρισμό, σύμφωνα με την περιγραφή της προηγούμενης παραγράφου.

Ακολούθως, μετατρέπουμε σε δυαδικά τα δεδομένα των προϊόντων που μας ενδιαφέρουν. Ο κώδικας προέρχεται από το αρχείο με τις σημειώσεις την R με μικρές αλλαγές, ωστόσο έχει κατανοηθεί πλήρως. Αρχικά δημιουργείται μια λίστα με τα ονόματα των προϊόντων που μας ενδιαφέρουν. Με βάση αυτή την λίστα και τις στήλες του dataframe original που περιέχουν τα items1-32 maximum που μπορεί να περιέχονται σε ένα καλάθι, δημιουργείται ένα νέο dataframe με μια στήλη για κάθε προϊόν της λίστας. Για κάθε εγγραφή, στην αντίστοιχη στήλη μπαίνει τιμή False αν το προϊόν δεν υπάρχει στο αντίστοιχο καλάθι, True διαφορετικά. Οι στήλες αυτές παίρνουν το όνομα του αντίστοιχου προϊόντος για το οποίο περιέχουν δεδομένα.

Τέλος, με την εντολή cbind ενώνονται το dataframe με τις στήλες που παρήξαμε, με τις υπόλοιπες στήλες του original που μας ενδιαφέρουν, δηλαδή τις id, recency_days, basket_value και τις 3 στήλες low/medium/high_value_basket που δημιουργήθηκαν νωρίτερα. Το dataframe Που παράγεται ονομάζεται mydata και είναι το ζητούμενο της άσκησης 1.

Σημείωση: Για την μετατροπή των προϊόντων που μας ενδιαφέρουν σε δυαδική μορφή, έγινε δοκιμή χρήσης πιο hardcoded κώδικα, που τελικά δεν χρησιμοποιήθηκε γιατί τα δεδομένα δεν έμεναν στην

επιθυμητή μορφή (logical) αλλά μετατρεπόταν σε τύπο character παρά τις αρκετές δοκιμές για αλλαγή τους.

Ενδεικτικά, ο κώδικας δημιουργούσε μια στήλη για κάθε προϊόν ξεχωριστά, ομοίως με την διαδικασία για της διακριτοποίηση basket_value, και με χρήση for διέτρεχε όλες τις εγγραφές. Για κάθε εγγραφή, αν έβρισκε κάποιο επιθυμητό προϊόν, έκανε True την τιμή της αντίστοιχης στήλης.

Η διαδικασία απαιτούσε επιπλέον προεπεξεργασία δεδομένων, όπως αφαίρεση κενών και χαρακτήρα /, και τελικά απορρίφθηκε γιατί τα παραγόμενα δεδομένα δεν μπορούσαν, λόγω του τύπου τους, να χρησιμοποιηθούν στον apriori.

Παρακάτω ενδεικτικά ο κώδικας που περιεγράφηκε αλλά δεν χρησιμοποιήθηκε:

```
original$citrusfruit <- as.logical(FALSE)
original$tropicalfruit <-as.logical(FALSE)
original$wholemilk <-as.logical(FALSE)
original$othervegetables <-as.logical(FALSE)
original$rollsbuns <-as.logical(FALSE)
original$chocolate <-as.logical(FALSE)
original$bottledwater <-as.logical(FALSE)
original$yogurt <-as.logical(FALSE)
original$sausage <-as.logical(FALSE)
original$rootvegetables <-as.logical(FALSE)
original$pastry <-as.logical(FALSE)
original$soda <-as.logical(FALSE)
original$cream <-as.logical(FALSE)

original <- as.data.frame(apply(original,2,function(x)gsub('\\s+', '',x)))
original <- as.data.frame(apply(original,2,function(x)gsub('rolls/buns', 'rollsbuns',x)))
```

```
for (row in 1:nrow(original)) {
  vec <- as.vector(original[row, ])

  if(any(vec == "citrusfruit")) {
    original$citrusfruit[row] <- as.logical(TRUE)
  }
  if(any(vec == "tropicalfruit")) {
    original$tropicalfruit[row] <- as.logical(TRUE)
  }
  if(any(vec == "wholemilk")) {
    original$wholemilk[row] <- as.logical(TRUE)
  }
  if(any(vec == "othervegetables")) {
    original$othervegetables[row] <- as.logical(TRUE)
  }
  if(any(vec == "rollsbuns")) {
    original$rollsbuns[row] <- as.logical(TRUE)
  }
  if(any(vec == "chocolate")) {
    original$chocolate[row] <- as.logical(TRUE)
  }
  if(any(vec == "bottledwater")) {
    original$bottledwater[row] <- as.logical(TRUE)
  }
  if(any(vec == "yogurt")) {
    original$yogurt[row] <- as.logical(TRUE)
  }
  if(any(vec == "sausage")) {
    original$sausage[row] <- as.logical(TRUE)
  }
  if(any(vec == "rootvegetables")) {
    original$rootvegetables[row] <- as.logical(TRUE)
  }
  if(any(vec == "pastry")) {
    original$pastry[row] <- as.logical(TRUE)
  }
  if(any(vec == "soda")) {
    original$soda[row] <- as.logical(TRUE)
  }
  if(any(vec == "cream")) {
    original$cream[row] <- as.logical(TRUE)
  }
}
```

```
mydata <- original[, c("id", "recency_days","low_value_basket","medium_value_basket", "high_value_basket", "citrusfruit",
"tropicalfruit", "wholemilk", "othervegetables", "rollsbuns", "chocolate", "bottledwater", "yogurt",
                       "sausage", "rootvegetables", "pastry", "soda", "cream")]
```

Άσκηση 2. Μάθηση κανόνων συσχέτισης με την μέθοδο Apriori

Για την άσκηση 2 είναι απαραίτητη η βιβλιοθήκη arules.

Α) Έγινε δοκιμή της μεθόδου apriori στα δεδομένα των προϊόντων μόνο (στήλες 7 μέχρι και 19). Χρησιμοποιήθηκαν παράμετροι για support, confidence, έγιναν πειράματα με min και max length των κανόνων. Παρατηρήθηκε, εύλογα, ότι όσο περισσότεροι περιορισμού χρησιμοποιήθηκαν, τόσο λιγόστευαν οι κανόνες που παρήγαγε η apriori. Ένα ελάχιστο support που χρησιμοποιήθηκε σε πολλούς πειραματισμούς και έβγαζε πάνω από 250 κανόνες με confidence άνω του 0,5 , ήταν το 0,001 .

Με το ίδιο αυτό support, 0.001, παρατηρήθηκαν 2 κανόνες με confidence 1, όταν οι κανόνες χρησιμοποιούσαν μόνο προϊόντα, και 198 κανόνες με confidence 1, όταν χρησιμοποιήθηκε και η αξία καλαθιού.

Ωστόσο, παρατηρήθηκε ότι όσο μεγάλωνε το support, τόσο μίκραινε το confidence, και οι κανόνες περιοριζόταν στα πιο συχνά εμφανιζόμενα προϊόντα, όπως vegetables (root ή other) και whole milk. Επίσης, ι κανόνες με υψηλό support περιείχαν συνήθως ένα προϊόν στο lhs και ένα στο rhs.

Β) Για τους κανόνες με υψηλότερο confidence μόνο για προϊόντα, χρησιμοποιήθηκαν μόνο οι παράμετροι support = 0.001 και confidence = 0.8

Τα 20 κορυφαία αποτελέσματα ως προς confidence είναι:

```
> inspect(sort(apr, by = 'confidence')[1:20])
     lhs                                                                rhs                support     confidence coverage    lift     count
[1]  {tropical fruit,rolls/buns,sausage,root vegetables}                => {whole milk}       0.001326964 1.0000000  0.001326964 2.998806 10
[2]  {tropical fruit,rolls/buns,bottled water,yogurt,root vegetables}   => {whole milk}       0.001061571 1.0000000  0.001061571 2.998806  8
[3]  {tropical fruit,yogurt,sausage,root vegetables}                    => {whole milk}       0.001990446 0.9375000  0.002123142 2.811381 15
[4]  {citrus fruit,tropical fruit,whole milk,yogurt,root vegetables}    => {other vegetables} 0.001857749 0.9333333  0.001990446 3.696059 14
[5]  {tropical fruit,rolls/buns,bottled water,root vegetables}          => {whole milk}       0.001459660 0.9166667  0.001592357 2.748906 11
[6]  {tropical fruit,yogurt,root vegetables,pastry}                     => {whole milk}       0.001326964 0.9090909  0.001459660 2.726187 10
[7]  {citrus fruit,tropical fruit,whole milk,rolls/buns,root vegetables}=> {other vegetables} 0.001061571 0.8888889  0.001194268 3.520056  8
[8]  {whole milk,rolls/buns,bottled water,yogurt,root vegetables}       => {tropical fruit}   0.001061571 0.8888889  0.001194268 6.490956  8
[9]  {citrus fruit,tropical fruit,whole milk,root vegetables}           => {other vegetables} 0.004113588 0.8857143  0.004644374 3.507484 31
[10] {citrus fruit,root vegetables,pastry}                              => {other vegetables} 0.001990446 0.8823529  0.002255839 3.494173 15
[11] {citrus fruit,tropical fruit,other vegetables,bottled water}       => {whole milk}       0.001725053 0.8666667  0.001990446 2.598965 13
[12] {citrus fruit,tropical fruit,rolls/buns,root vegetables}           => {other vegetables} 0.001459660 0.8461538  0.001725053 3.350823 11
[13] {citrus fruit,tropical fruit,other vegetables,yogurt,root vegetables} => {whole milk}    0.001857749 0.8235294  0.002255839 2.469605 14
[14] {citrus fruit,other vegetables,yogurt,root vegetables}             => {whole milk}       0.003052017 0.8214286  0.003715499 2.463305 23
[15] {chocolate,root vegetables,pastry}                                 => {other vegetables} 0.001194268 0.8181818  0.001459660 3.240052  9
[16] {citrus fruit,whole milk,root vegetables,pastry}                   => {other vegetables} 0.001194268 0.8181818  0.001459660 3.240052  9
[17] {citrus fruit,other vegetables,yogurt,pastry}                      => {whole milk}       0.001194268 0.8181818  0.001459660 2.453569  9
[18] {other vegetables,sausage,root vegetables,pastry}                  => {yogurt}           0.001194268 0.8181818  0.001459660 4.494037  9
[19] {tropical fruit,rolls/buns,yogurt,root vegetables}                 => {whole milk}       0.002919321 0.8148148  0.003582803 2.443472 22
[20] {citrus fruit,tropical fruit,yogurt,root vegetables}               => {other vegetables} 0.002255839 0.8095238  0.002786624 3.205765 17
```

Όπως αναφέρθηκε και νωρίτερα, παρατηρούμε ότι οι περισσότεροι κανόνες περιέχουν προϊόντα καθημερινής χρήσης όπως vegetables και whole milk. Εντύπωση κάνει το γεγονός ότι κανόνες που περιέχουν τροπικά φρούτα εμφανίζονται πιο συχνά από τα εσπεριδοειδή, ενώ δεν εμφανίζεται όσο συχνά θα περιμέναμε το εμφιαλωμένο νερό.

Εντύπωση κάνει ο κανόνας 1, ο οποίος περιέχει προϊόντα από διαφορετικές τροφικές ομάδες και όχι απαραίτητα καθημερινής κατανάλωσης, όπως rolls/buns, λουκάνικο και τροπικά φρούτα, σε αντίθεση με άλλους κανόνες που φαίνονται να συνδυάζουν διάφορα λαχανικά ή διάφορα φρούτα μεταξύ τους, και να περιέχουν νερό ή γάλα.

Γ) Εισάγοντας την αξία καλαθιού στα δεδομένα που χρησιμοποιούν οι κανόνες, παρατηρούμε μεγάλη αλλαγή. Οι περισσότεροι κανόνες πλέον ασχολούνται με την αξία, και αυξάνεται σημαντικά το confidence σε ανάλογο support. Όπως αναφέρθηκε νωρίτερα, έχουμε σχεδόν 200 κανόνες με confidence 1, για support 0,001.

Αυξάνοντας το support σε 0,002 έχουμε πάλι αρκετούς κανόνες με confidence 1. Ενδεικτικά οι 20 πρώτοι είναι:

```
> inspect(sort(apr, by = 'confidence')[1:20])
     lhs                                                  rhs                          support     confidence coverage    lift     count
[1]  {rolls/buns,chocolate,pastry}                     => {original$high_value_basket} 0.003184713 1          0.003184713 9.246626 24
[2]  {tropical fruit,chocolate,sausage}                => {original$high_value_basket} 0.002255839 1          0.002255839 9.246626 17
[3]  {chocolate,yogurt,sausage}                        => {original$high_value_basket} 0.002388535 1          0.002388535 9.246626 18
[4]  {rolls/buns,chocolate,sausage}                    => {original$high_value_basket} 0.003052017 1          0.003052017 9.246626 23
[5]  {citrus fruit,sausage,pastry}                     => {original$high_value_basket} 0.002521231 1          0.002521231 9.246626 19
[6]  {bottled water,sausage,pastry}                    => {original$high_value_basket} 0.002255839 1          0.002255839 9.246626 17
[7]  {tropical fruit,sausage,pastry}                   => {original$high_value_basket} 0.003848195 1          0.003848195 9.246626 29
[8]  {sausage,root vegetables,pastry}                  => {original$high_value_basket} 0.003052017 1          0.003052017 9.246626 23
[9]  {yogurt,sausage,pastry}                           => {original$high_value_basket} 0.005573248 1          0.005573248 9.246626 42
[10] {sausage,pastry,soda}                             => {original$high_value_basket} 0.004511677 1          0.004511677 9.246626 34
[11] {rolls/buns,sausage,pastry}                       => {original$high_value_basket} 0.005042463 1          0.005042463 9.246626 38
[12] {other vegetables,sausage,pastry}                 => {original$high_value_basket} 0.005175159 1          0.005175159 9.246626 39
[13] {whole milk,sausage,pastry}                       => {original$high_value_basket} 0.007430998 1          0.007430998 9.246626 56
[14] {bottled water,sausage,root vegetables}           => {original$high_value_basket} 0.002521231 1          0.002521231 9.246626 19
[15] {citrus fruit,other vegetables,yogurt,sausage}    => {original$high_value_basket} 0.002123142 1          0.002123142 9.246626 16
[16] {citrus fruit,whole milk,other vegetables,sausage} => {original$high_value_basket} 0.003052017 1          0.003052017 9.246626 23
[17] {yogurt,sausage,root vegetables,pastry}           => {original$high_value_basket} 0.002123142 1          0.002123142 9.246626 16
[18] {whole milk,sausage,root vegetables,pastry}       => {original$high_value_basket} 0.002123142 1          0.002123142 9.246626 16
[19] {rolls/buns,yogurt,sausage,pastry}                => {original$high_value_basket} 0.002388535 1          0.002388535 9.246626 18
[20] {whole milk,yogurt,sausage,pastry}                => {original$high_value_basket} 0.003052017 1          0.003052017 9.246626 23
```

Εντύπωση κάνει το γεγονός ότι οι κορυφαίοι κανόνες ασχολούνται μόνο με καλάθια υψηλής αξίας, τα οποία είναι τα πιο σπάνια. Αυτό μπορεί να βοηθήσει στην πρόβλεψη των πιο ακριβών προϊόντων, τα οποία όταν αγοράζονται μαζί, παράγουν σίγουρα ακριβά καλάθια.

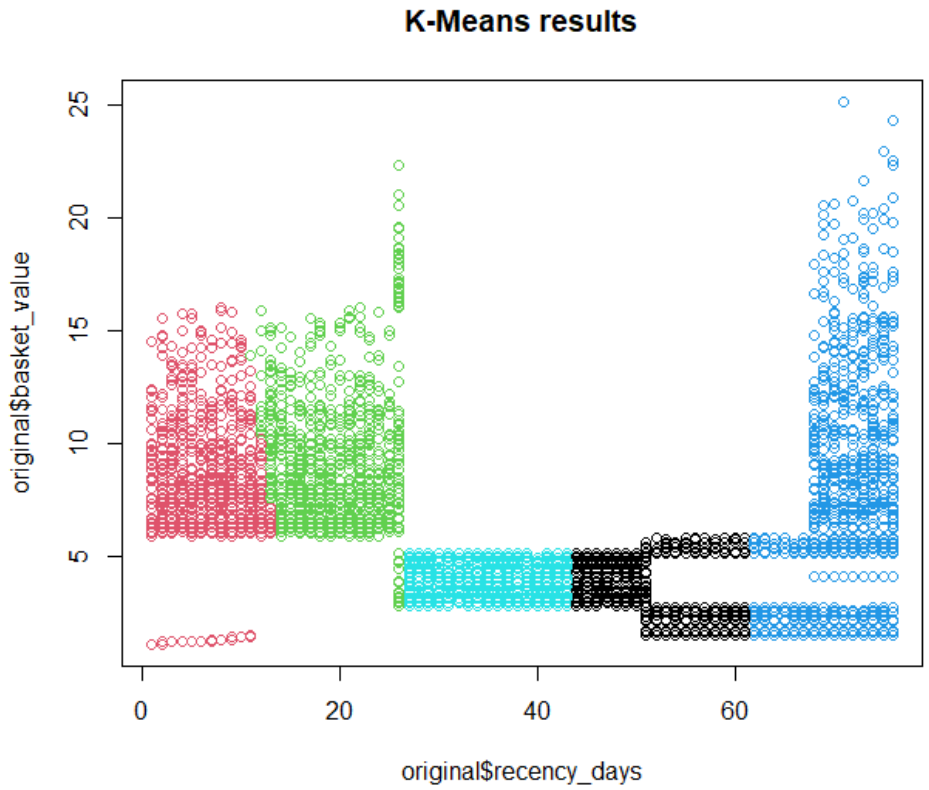
Πιο υψηλής αξίας φαίνεται να είναι τα ζυμαρικά και το λουκάνικο. Κατά πάσα πιθανότητα, το ακριβότερο εκ των 2 είναι το λουκάνικο, αφού εμφανίζεται σε 19 από τους 20 κανόνες, με τα ζυμαρικά να εμφανίζονται σε 13 κανόνες. Πρέπει να σημειωθεί ότι τα 2 αυτά προϊόντα εμφανιζόταν σπάνια πριν την εισαγωγή της αξίας καλαθιού σε κανόνες, και ειδικά σε κανόνες με μεγάλο support. Αυτό σημαίνει ότι αγοράζονται πιο σπάνια, και συνδυάζονται κυρίως με επίσης ακριβά προϊόντα.

Άσκηση 3. Ομαδοποίηση συναλλαγών με χρήση μεθόδου k-means

Α) Χρησιμοποιούμε την μέθοδο kmeans της R, δίνοντας σαν δεδομένα τις στήλες recency days και basket value (με τις συνεχείς και όχι τις διακριτοποιημένες τιμές), και ορίζουμε 5 κέντρα.

Εμφανίζοντας τα δεδομένα, παίρνουμε τον αριθμό εγγραφών ανά cluster, τα 5 κέντρα, το cluster που ανήκε κάθε εγγραφή και άλλες χρήσιμες πληροφορίες.

Φορτώνοντας τις βιβλιοθήκες οπτικοποίησης factoextra και ggplot2, εμφανίζουμε σχηματικά τα clusters που δημιουργούνται. Παρατηρούμε ότι χωρίζονται σε ομάδες σύμφωνα με το χαρακτηριστικό recency days, με τις 5 ομάδες να περιέχουν φαινομενικά σχεδόν τον ίδιο αριθμό ημερών.



Σημείωση: Όσες φορές εκτελέστηκε ο kmeans, ακόμα και με ορισμένα αρχικά κέντρα, το αποτέλεσμα δεν διέφερε πολύ από αυτό της φωτογραφίας.

Β) Παρακάτω βλέπουμε πληροφορίες για το αποτέλεσμα του kmeans:

```
> kmeans_res
K-means clustering with 5 clusters of sizes 1285, 1926, 1072, 1986, 1267

Cluster means:
  original$recency_days original$basket_value
1             52.633463              3.061245
2              6.490135              4.615421
3             19.823694              9.024440
4             70.232125              5.325478
5             34.808208              3.853275

Clustering vector:
   [1] 2 5 2 2 4 1 4 1 2 3 3 1 4 1 2 4 2 2 3 2 4 2 2 1 5 5 1 2 4 1 2 2 4 4 2 4 4 2 2 4 5 5 4 2 3 4 4 4 2 5 4 2 2 3 1 4 5 5 4 2 5 5 1 1 2 3 2 4 1 3 3
  [72] 2 2 4 1 4 4 1 3 4 4 1 2 1 2 2 3 5 2 5 1 5 3 4 5 4 2 3 2 4 1 1 4 4 2 2 2 2 2 4 5 1 5 4 3 4 3 1 2 1 5 4 5 2 1 3 4 2 5 2 1 2 2 5 1 1 1 3 1 1 2 4
 [143] 5 3 2 4 4 5 2 5 5 4 4 4 5 4 4 5 2 5 3 2 2 2 3 4 2 2 3 4 2 3 5 4 2 3 3 4 1 5 5 5 5 3 3 2 5 2 1 2 1 1 1 3 5 4 5 2 3 2 1 1 5 1 2 2 4 3 3 2 4 5 5
 [214] 2 4 2 5 1 5 3 3 2 1 5 2 1 5 3 2 4 3 2 4 5 5 2 1 3 4 4 5 1 3 3 2 1 1 4 4 2 5 5 4 3 1 4 5 2 2 5 5 1 1 5 1 1 2 2 4 4 5 3 4 1 4 4 4 1 4 1 4 2 1 2
 [285] 1 1 4 4 2 2 4 1 2 1 4 2 3 2 5 5 5 2 5 1 2 2 5 3 4 3 4 4 3 2 5 4 2 1 4 4 3 2 4 5 1 2 5 3 2 4 4 5 5 3 4 5 5 4 3 5 4 5 3 2 2 3 2 5 5 4 2 4 2 3 1
 [356] 5 2 2 4 2 3 2 4 2 5 4 3 4 3 2 5 4 5 1 2 4 5 2 2 4 4 5 3 4 2 3 5 1 2 3 5 5 2 4 3 2 4 1 1 5 4 5 1 5 5 2 2 1 4 5 2 4 3 2 4 2 5 5 5 5 2 2 2 2 3 1
 [427] 1 1 1 1 4 2 4 2 1 5 2 5 4 5 4 4 5 2 2 4 4 4 2 1 2 1 3 4 2 2 5 1 5 1 4 4 4 4 3 2 1 3 5 4 1 2 2 2 4 5 5 5 2 1 4 1 2 3 2 3 4 4 5 3 1 2 5 2 2 4 5
 [498] 1 5 2 2 3 5 4 5 3 4 2 2 4 2 4 2 4 5 2 3 4 1 2 5 4 1 2 4 4 4 5 5 5 5 4 2 2 5 4 4 4 5 3 3 4 4 4 5 4 1 1 4 1 2 1 4 4 5 5 3 1 5 2 5 4 5 3 3 5 2 1
 [569] 4 4 4 4 4 5 3 2 2 3 5 3 1 4 2 1 2 2 4 2 3 3 2 4 3 1 1 2 2 5 1 2 4 4 2 5 4 1 2 3 3 2 2 5 3 1 1 1 1 4 4 3 2 4 3 4 3 5 1 2 4 3 2 5 3 1 5 1 2 2 1
 [640] 1 4 1 2 3 4 1 5 3 1 1 3 2 1 4 3 2 2 3 3 3 5 3 1 2 3 5 2 5 2 4 5 2 4 3 5 4 1 2 3 2 4 2 1 3 2 2 3 2 4 3 3 1 5 5 4 4 4 1 4 2 3 4 2 1 2 4 5 5 2 3
 [711] 2 2 4 1 4 2 4 5 5 1 1 5 4 5 2 4 4 1 5 4 3 2 3 3 4 5 5 2 5 5 3 5 1 5 1 4 2 4 3 5 4 1 1 5 4 4 4 1 1 2 2 5 3 3 4 1 5 2 3 2 1 2 3 4 1 3 3 5 1 5 2
 [782] 2 4 4 5 3 4 4 2 1 2 5 2 3 4 2 2 5 5 1 1 3 5 5 2 3 5 2 5 5 2 4 4 5 1 4 1 1 1 4 2 5 4 2 2 1 4 1 4 2 3 4 1 4 3 1 5 5 4 4 4 4 4 5 5 2 2 5 4 1 1 5
 [853] 1 1 4 2 2 4 3 1 4 2 3 4 2 4 3 1 5 5 3 3 3 4 2 2 4 2 4 5 5 5 5 2 1 2 3 2 4 3 4 5 4 1 2 5 5 4 1 5 4 5 2 2 5 2 3 2 2 5 4 5 2 4 3 4 5 2 5 2 1 3 2
 [924] 2 4 1 5 1 4 2 3 4 5 2 3 1 5 3 1 2 4 3 3 3 2 2 3 3 4 2 3 5 2 5 4 2 3 5 5 5 2 4 2 1 5 2 2 2 1 4 2 1 2 4 5 2 3 4 1 5 1 3 4 2 3 4 1 1 1 1 5 1 3 4
 [995] 5 5 5 2 3 4
 [ reached getOption("max.print") -- omitted 6536 entries ]

Within cluster sum of squares by cluster:
[1] 36227.13 50010.62 29475.66 65654.90 31603.39
 (between_SS / total_SS =  95.6 %)
```

Τα 5 clusters δεν διαφέρουν πολύ σε εγγραφές που περιέχουν, και κρίνοντας από τα κέντρα των clusters επιβεβαιώνεται ότι τα clusters έχουν διαχωριστεί με βάση των αριθμό ημερών που περιέχουν.

Όπως αναφέραμε νωρίτερα, η μέση αξία των καλαθιών όλων των εγγραφών κυμαίνεται γύρω στο 5.

```
> summary(original$basket_value)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.100   2.200   4.100   5.037   7.000  25.100
```

Οι μέσες τιμές των clusters αποδεικνύουν ακριβώς αυτό. Μοναδική εξαίρεση αποτελεί το cluster 3 (στο plot αυτό με το πράσινο χρώμα) που αναφέρεται σε αγορές γύρω στις 19 μέρες νωρίτερα. Τα καλάθια αυτής την κατηγορίας έχουν μεγαλύτερες τιμές, αφού το κέντρο έχει τιμή 9.02, αρκετά μεγαλύτερη από τα περισσότερα καλάθια. Αυτό σημαίνει ότι εκείνη την περίοδο οι καταναλωτές έκαναν πιο ακριβές αγορές.

Θα ήταν χρήσιμο για το τμήμα Marketing να ασχοληθεί με τις αγορές εκείνων των ημερών, ώστε να ανακαλύψει τι προκάλεσε την αύξηση του κέρδους (ίσως κάποια καμπάνια, εκπτώσεις που οδήγησαν σε ακριβότερες αγορές, κάποια αργία ή διακοπές κ.α.), ώστε αν είναι δυνατόν να επαναλάβουν κινήσεις από πλευράς του καταστήματος που έγιναν εκείνη την περίοδο.

Ωστόσο, κρίνοντας από το plot, παρατηρούμε ότι και το μπλέ cluster ξεχωρίσει, καθώς πριν από 70+ μέρες, παρατηρούνται οι ακριβότερες αγορές που υπάρχουν, παρά τον μέσο όρο των καλαθιών που δεν παρουσιάζει αυτή την πληροφορία. Θα ήταν χρήσιμη η ανάλυση των συνθηκών και αυτής της περιόδου.

Γενικότερα, τα clusters θα μπορούσαν να περιγραφούν ως εξής:

Cluster1 (μαύρο) : Ομάδα αγορών γύγω στις 54 μέρες πριν (44-61), με καλάθια αξίας περί των 4 ευρώ στις πιο πρόσφατες μέρες του cluster, και ακόμα φθηνότερα είτε ελάχιστα ακριβότερα καλάθια για περισσότερες από 50 μέρες πριν.

Cluster2 (κόκκινο) : Αγορές των τελευταίων 12 ημερών, με μέσο όρο αξίας 4.6, λίγο κάτω από τον γενικό μέσο όρο, με τιμές καλαθιών από τις χαμηλότερες που υπάρχουν, έως και οριακά άνω των 15 ευρώ.

Cluster3 (πράσινο) : Αγορές μεταξύ 13 και 26 ημερών πριν, με τον μεγαλύτερο μέσο όρο τιμής καλαθιών 9,02 και ομοιόμορφες σε τιμές πωλήσεις κάθε μέρα.

Cluster4 (μπλε) : Αντιπροσωπεύει τις παλιότερες αγορές, με το cluster να παρουσιάζει κατά το ήμισι χαμηλότερες τιμές που ρίχνουν τον μέσο όρο του, ενώ στις παλιότερες αγορές συμπεριλαμβάνονται οι ακριβότερες που έχουν γίνει ποτέ στο κατάστημα. Με δεδομένο ότι είναι τα παλιότερα δεδομένα, ίσως ανταποκρίνονται στα εγκαίνια του καταστήματος όπου έγινε κάποια προωθητική ενέργεια.

Cluster5 (γαλάζιο) : Το μεγαλύτερο σε έκταση ημερών cluster, από 27 μέχρι 43 μέρες πριν, με τις λιγότερες αγορές ανά μέρα (λαμβάνοντας υπόψη ότι περιέχει τις περισσότερες μέρες) και τον πιο συμπαγή χαμηλό μέσο όρο τιμών. Οι καταναλωτές ψώνιζαν μόνο τα απολύτως απαραίτητα και φθηνά προϊόντα.

Γ) Με τρόπο παρόμοιο με αυτό της 1ης άσκησης για την αξία καλαθιού, πραγματοποιείται και το 3ο ερώτημα εδώ. Πιο αναλυτικά, δημιουργείται ένα χαρακτηριστικό mydata$cluster, όπου αποθηκεύεται το cluster Την κάθε εγγραφής, όπως προέκυψε από τον kmeans. Ακολούθως δημιουργούνται 5 στήλες, Cluster1 έως Cluster5, που αρχικοποιούνται ως False. Με ένα for loop προσπελαύνονται όλες οι εγγραφές, και για κάθε εγγραφή ανάλογα με την τιμή του cluster, γίνεται True η τιμή του αντίστοιχου Cluster-αριθμού .

Τέλος, αφαιρούμε την στήλη cluster που δεν χρειαζόμαστε πλέον, ώστε να μην δημιουργεί πρόβλημα στον επόμενο apriori.

Άσκηση 4. Συνδυαστική αξιοποίηση μεθόδων: περιγραφή προιοντικού προφίλ ομάδων με χρήση κανόνων συσχέτισης

Για την συγκεκριμένη άσκηση, μας ενδιαφέρει να υπάρχει για κάθε cluster, το αντίστοιχο cluster είτε στο lhs είτε στο rhs. Επειδή δεν υπάρχει κάποια εντολή που να το εγγυάται αυτό, για κάθε cluster τρέχουμε τον apriori 2 φορές, όπου την μια δίνουμε παράμετρο appearance = list(lhs = "όνομα cluster") και την άλλη appearance = list(rhs = "όνομα cluster") και κρατάμε τα 20 καλύτερα confidence.

Πρέπει να σημειωθεί ότι λόγω των πολύ ειδικών κανόνων που αναζητούμε, τόσο το support όσο και το confidence είναι αρκετά χαμηλά, ενώ συναντάμε και αρκετούς αντεστραμμένους κανόνες.

Από το κάθε cluster στο lhs βλέπουμε τα πιο συχνά εμφανιζόμετα προϊόντα μεμονωμένα, ενώ όταν βρίσκεται στο rhs παίρνουμε κανόνες με προϊόντα που εμφανίζονται συχνά μαζί μέσα στα καλάθια της ομάδας.

Cluster1:

```
> inspect(sort(apr_clust1, by = 'confidence'))
     lhs             rhs                support     confidence coverage  lift      count
[1]  {Cluster1} => {soda}              0.044187898 0.25914397 0.1705149 1.1387224 333
[2]  {Cluster1} => {rolls/buns}        0.042728238 0.25058366 0.1705149 1.0438908 322
[3]  {Cluster1} => {whole milk}        0.038747346 0.22723735 0.1705149 0.6814408 292
[4]  {Cluster1} => {other vegetables} 0.028927813 0.16964981 0.1705149 0.6718239 218
[5]  {Cluster1} => {yogurt}            0.028795117 0.16887160 0.1705149 0.9275629 217
[6]  {Cluster1} => {root vegetables}   0.022823779 0.13385214 0.1705149 0.9409606 172
```

```
> inspect(sort(apr_clust1, by = 'confidence'))
     lhs                                    rhs        support     confidence coverage    lift      count
[1]  {soda}                                 => {Cluster1} 0.044187898 0.19416910 0.227574310 1.1387224 333
[2]  {rolls/buns}                           => {Cluster1} 0.042728238 0.17799889 0.240047771 1.0438908 322
[3]  {whole milk,bottled water}             => {Cluster1} 0.007563694 0.16863905 0.044851380 0.9889991  57
[4]  {root vegetables}                      => {Cluster1} 0.022823779 0.16044776 0.142250531 0.9409606 172
[5]  {yogurt}                               => {Cluster1} 0.028795117 0.15816327 0.182059448 0.9275629 217
[6]  {bottled water,soda}                   => {Cluster1} 0.005705945 0.15087719 0.037818471 0.8848331  43
[7]  {tropical fruit}                       => {Cluster1} 0.018444798 0.13468992 0.136942675 0.7899014 139
[8]  {other vegetables,root vegetables}     => {Cluster1} 0.007829087 0.12660944 0.061836518 0.7425127  59
[9]  {whole milk,other vegetables}          => {Cluster1} 0.012340764 0.12635870 0.097664544 0.7410421  93
[10] {bottled water}                        => {Cluster1} 0.017781316 0.12327507 0.144240977 0.7229579 134
[11] {citrus fruit,other vegetables}        => {Cluster1} 0.004644374 0.12323944 0.037685775 0.7227489  35
[12] {whole milk,soda}                      => {Cluster1} 0.006369427 0.12182741 0.052282378 0.7144680  48
[13] {whole milk,bottled water,soda}        => {Cluster1} 0.001194268 0.12162162 0.009819533 0.7132611   9
[14] {chocolate}                            => {Cluster1} 0.007563694 0.11680328 0.064755839 0.6850035  57
```

Στο cluster αυτό κυριαρχούν τα ποτά (νερό, σόδα, γάλα), και εμφανίζονται κυρίως υγιεινά τρόφιμα, όπως γιαούρτι και λαχανικά.

Από το cluster αυτό δημιουργούνται ελάχιστοι κανόνες, γεγονός που δηλώνει ασυνέπεια στα καλάθια των καταναλωτών.

Cluster2:

```
> inspect(sort(apr_clust2, by = 'confidence'))
    lhs                rhs                  support    confidence coverage  lift      count
[1] {Cluster2} => {whole milk}       0.11106688 0.4345794  0.2555732 1.3032195 837
[2] {Cluster2} => {other vegetables} 0.08160828 0.3193146  0.2555732 1.2645061 615
[3] {Cluster2} => {bottled water}    0.04949575 0.1936656  0.2555732 1.3426533 373
[4] {Cluster2} => {rolls/buns}       0.04179936 0.1635514  0.2555732 0.6813286 315
[5] {Cluster2} => {sausage}          0.03808386 0.1490135  0.2555732 1.2153309 287
[6] {Cluster2} => {yogurt}           0.03596072 0.1407061  0.2555732 0.7728581 271
[7] {Cluster2} => {citrus fruit}     0.03224522 0.1261682  0.2555732 1.1680636 243
[8] {Cluster2} => {soda}             0.03131635 0.1225337  0.2555732 0.5384340 236
[9] {Cluster2} => {root vegetables}  0.02959130 0.1157840  0.2555732 0.8139443 223
```

```
> inspect(sort(apr_clust2, by = 'confidence')[1:20])
    lhs                                                                rhs        support     confidence coverage    lift     count
[1]  {tropical fruit,whole milk,other vegetables,bottled water,root vegetables} => {Cluster2} 0.001061571 0.4705882  0.002255839 1.841305   8
[2]  {tropical fruit,other vegetables,bottled water,root vegetables}            => {Cluster2} 0.001592357 0.4615385  0.003450106 1.805895  12
[3]  {tropical fruit,whole milk,other vegetables,bottled water}                 => {Cluster2} 0.002123142 0.4571429  0.004644374 1.788696  16
[4]  {citrus fruit,chocolate,yogurt}                                            => {Cluster2} 0.001194268 0.4500000  0.002653928 1.760748   9
[5]  {other vegetables,rolls/buns,sausage,root vegetables}                      => {Cluster2} 0.001326964 0.4347826  0.003052017 1.701205  10
[6]  {tropical fruit,whole milk,rolls/buns,bottled water}                       => {Cluster2} 0.001592357 0.4285714  0.003715499 1.676903  12
[7]  {tropical fruit,bottled water,root vegetables}                             => {Cluster2} 0.002521231 0.4222222  0.005971338 1.652060  19
[8]  {tropical fruit,whole milk,bottled water}                                  => {Cluster2} 0.004378981 0.4177215  0.010483015 1.634449  33
[9]  {tropical fruit,whole milk,bottled water,root vegetables}                  => {Cluster2} 0.001326964 0.4000000  0.003317410 1.565109  10
[10] {other vegetables,bottled water,yogurt,root vegetables}                    => {Cluster2} 0.001061571 0.4000000  0.002653928 1.565109   8
[11] {tropical fruit,other vegetables,yogurt,soda}                              => {Cluster2} 0.001061571 0.4000000  0.002653928 1.565109   8
[12] {whole milk,other vegetables,rolls/buns,soda}                              => {Cluster2} 0.002255839 0.3953488  0.005705945 1.546910  17
[13] {bottled water,yogurt,root vegetables}                                     => {Cluster2} 0.001990446 0.3947368  0.005042463 1.544515  15
[14] {rolls/buns,sausage}                                                       => {Cluster2} 0.015525478 0.3887043  0.039941614 1.520912 117
[15] {other vegetables,rolls/buns,root vegetables}                              => {Cluster2} 0.006104034 0.3833333  0.015923567 1.499896  46
[16] {other vegetables,bottled water,yogurt,soda}                               => {Cluster2} 0.001061571 0.3809524  0.002786624 1.490580   8
[17] {other vegetables,rolls/buns,sausage}                                      => {Cluster2} 0.004378981 0.3793103  0.011544586 1.484155  33
[18] {whole milk,other vegetables,rolls/buns,root vegetables}                   => {Cluster2} 0.003052017 0.3770492  0.008094480 1.475308  23
[19] {tropical fruit,bottled water,yogurt,soda}                                 => {Cluster2} 0.001194268 0.3750000  0.003184713 1.467290   9
[20] {tropical fruit,bottled water,soda}                                        => {Cluster2} 0.002521231 0.3725490  0.006767516 1.457700  19
```

Το cluster2 είχε πολλούς παραγόμενους κανόνες για rhs. Αφήνουμε τα 29 καλύτερα, 9 για lhs και 20 για rhs, καθώς αποτυπώνεται η διαφορά στο πλήθος εμφανίσεων των κανόνων στα δεδομένα. Παρατηρούμε πολύ μεγάλους κανόνες, ειδικά στην δεύτερη εικόνα, με τα καλάθια να αποτυπώνουν έναν υγιεινό τρόπο ζωής με πολλά τροπικά φρούτα, και ακολούθως λαχανικά ως κυρίαρχα προϊόντα για τους κανόνες με πολλά προϊόντα, ενώ συχνές εμφανίσεις έχουμε κυρίως σε γάλα και λαχανικά.

Cluster3:

```
> inspect(sort(apr_clust3, by = 'confidence'))
     lhs           rhs                  support    confidence coverage  lift     count
[1]  {Cluster3} => {whole milk}       0.06329618 0.4445480  0.1423832 1.333113 477
[2]  {Cluster3} => {other vegetables} 0.05400743 0.3793103  0.1423832 1.502093 407
[3]  {Cluster3} => {sausage}          0.05294586 0.3718546  0.1423832 3.032788 399
[4]  {Cluster3} => {rolls/buns}       0.04949575 0.3476235  0.1423832 1.448143 373
[5]  {Cluster3} => {yogurt}           0.04392251 0.3084809  0.1423832 1.694397 331
[6]  {Cluster3} => {soda}             0.04047240 0.2842498  0.1423832 1.249042 305
[7]  {Cluster3} => {tropical fruit}   0.03688960 0.2590867  0.1423832 1.891935 278
[8]  {Cluster3} => {root vegetables}  0.03609342 0.2534949  0.1423832 1.782031 272
[9]  {Cluster3} => {bottled water}    0.02760085 0.1938490  0.1423832 1.343925 208
[10] {Cluster3} => {citrus fruit}     0.02507962 0.1761417  0.1423832 1.630717 189
[11] {Cluster3} => {chocolate}        0.01579087 0.1109040  0.1423832 1.712649 119
```

```
> inspect(sort(apr_clust3, by = 'confidence')[1:20])
     lhs                                                                rhs        support     confidence coverage    lift     count
[1]  {citrus fruit,bottled water,yogurt,soda}                        => {Cluster3} 0.001061571 1.0000000  0.001061571 7.023299  8
[2]  {whole milk,chocolate,sausage,root vegetables}                  => {Cluster3} 0.001194268 0.9000000  0.001326964 6.320969  9
[3]  {other vegetables,chocolate,yogurt,soda}                        => {Cluster3} 0.001459660 0.8461538  0.001725053 5.942792 11
[4]  {tropical fruit,whole milk,rolls/buns,chocolate}                => {Cluster3} 0.001326964 0.8333333  0.001592357 5.852749 10
[5]  {citrus fruit,tropical fruit,other vegetables,soda}             => {Cluster3} 0.001194268 0.8181818  0.001459660 5.746336  9
[6]  {citrus fruit,tropical fruit,root vegetables,soda}              => {Cluster3} 0.001061571 0.8000000  0.001326964 5.618639  8
[7]  {citrus fruit,whole milk,other vegetables,bottled water,yogurt} => {Cluster3} 0.001061571 0.8000000  0.001326964 5.618639  8
[8]  {citrus fruit,whole milk,bottled water,yogurt}                  => {Cluster3} 0.001857749 0.7777778  0.002388535 5.462566 14
[9]  {whole milk,chocolate,sausage,soda}                             => {Cluster3} 0.001326964 0.7692308  0.001725053 5.402538 10
[10] {rolls/buns,yogurt,root vegetables,soda}                        => {Cluster3} 0.001326964 0.7692308  0.001725053 5.402538 10
[11] {tropical fruit,whole milk,rolls/buns,bottled water,yogurt}     => {Cluster3} 0.001326964 0.7692308  0.001725053 5.402538 10
[12] {tropical fruit,other vegetables,sausage,root vegetables}       => {Cluster3} 0.001725053 0.7647059  0.002255839 5.370758 13
[13] {citrus fruit,whole milk,sausage,soda}                          => {Cluster3} 0.001194268 0.7500000  0.001592357 5.267474  9
[14] {tropical fruit,yogurt,root vegetables,soda}                    => {Cluster3} 0.001194268 0.7500000  0.001592357 5.267474  9
[15] {bottled water,sausage,root vegetables}                         => {Cluster3} 0.001857749 0.7368421  0.002521231 5.175063 14
[16] {chocolate,sausage,root vegetables}                             => {Cluster3} 0.001459660 0.7333333  0.001990446 5.150419 11
[17] {citrus fruit,yogurt,root vegetables,soda}                      => {Cluster3} 0.001061571 0.7272727  0.001459660 5.107854  8
[18] {citrus fruit,rolls/buns,yogurt,soda}                           => {Cluster3} 0.001061571 0.7272727  0.001459660 5.107854  8
[19] {other vegetables,bottled water,sausage,root vegetables}        => {Cluster3} 0.001061571 0.7272727  0.001459660 5.107854  8
[20] {whole milk,rolls/buns,sausage,soda}                            => {Cluster3} 0.002123142 0.7272727  0.002919321 5.107854 16
```

Στο 3ο cluster έχουμε επίσης γάλα, φρούτα και λαχανικά ως τα πιο συχνά προϊόντα, και καλάθια με πιο ακριβά και ανθυγιεινά προϊόντα, όπως λουκάνικα, σόδα και σοκολάτα.

Το cluster αυτό παρουσιάζει τον υψηλότερο μέσο όρο τιμών, που ίσως συνδυάζονται με την συχνή εμφάνιση του λουκάνικου στα καλάθια. Νωρίτερα εκτιμήσαμε ότι πιθανότατα είναι το ακριβότερο προϊόν.

Cluster4:

```
> inspect(sort(apr_clust4, by = 'confidence'))
     lhs           rhs                  support    confidence coverage  lift     count
[1]  {Cluster4} => {pastry}           0.11610934 0.4405841  0.263535 3.7945619 875
[2]  {Cluster4} => {soda}             0.07192144 0.2729104  0.263535 1.1992143 542
[3]  {Cluster4} => {whole milk}       0.06674628 0.2532729  0.263535 0.7595164 503
[4]  {Cluster4} => {rolls/buns}       0.06303079 0.2391742  0.263535 0.9963609 475
[5]  {Cluster4} => {other vegetables} 0.04777070 0.1812689  0.263535 0.7178362 360
[6]  {Cluster4} => {yogurt}           0.04498408 0.1706949  0.263535 0.9375776 339
[7]  {Cluster4} => {tropical fruit}   0.03025478 0.1148036  0.263535 0.8383335 228
```

```
> inspect(sort(apr_clust4, by = 'confidence')[1:20])
     lhs                                  rhs         support     confidence coverage    lift     count
[1]  {pastry}                          => {Cluster4} 0.116109342 1          0.116109342 3.794562 875
[2]  {chocolate,pastry}                => {Cluster4} 0.010483015 1          0.010483015 3.794562  79
[3]  {citrus fruit,pastry}             => {Cluster4} 0.012738854 1          0.012738854 3.794562  96
[4]  {sausage,pastry}                  => {Cluster4} 0.016321656 1          0.016321656 3.794562 123
[5]  {bottled water,pastry}            => {Cluster4} 0.011677282 1          0.011677282 3.794562  88
[6]  {tropical fruit,pastry}           => {Cluster4} 0.017250531 1          0.017250531 3.794562 130
[7]  {root vegetables,pastry}          => {Cluster4} 0.014331210 1          0.014331210 3.794562 108
[8]  {yogurt,pastry}                   => {Cluster4} 0.023089172 1          0.023089172 3.794562 174
[9]  {pastry,soda}                     => {Cluster4} 0.027468153 1          0.027468153 3.794562 207
[10] {rolls/buns,pastry}               => {Cluster4} 0.027335456 1          0.027335456 3.794562 206
[11] {other vegetables,pastry}         => {Cluster4} 0.029458599 1          0.029458599 3.794562 222
[12] {whole milk,pastry}               => {Cluster4} 0.043391720 1          0.043391720 3.794562 327
[13] {citrus fruit,chocolate,pastry}   => {Cluster4} 0.002123142 1          0.002123142 3.794562  16
[14] {chocolate,sausage,pastry}        => {Cluster4} 0.001592357 1          0.001592357 3.794562  12
[15] {chocolate,bottled water,pastry}  => {Cluster4} 0.001194268 1          0.001194268 3.794562   9
[16] {tropical fruit,chocolate,pastry} => {Cluster4} 0.002919321 1          0.002919321 3.794562  22
[17] {chocolate,root vegetables,pastry} => {Cluster4} 0.001459660 1          0.001459660 3.794562  11
[18] {chocolate,yogurt,pastry}         => {Cluster4} 0.002255839 1          0.002255839 3.794562  17
[19] {chocolate,pastry,soda}           => {Cluster4} 0.002919321 1          0.002919321 3.794562  22
[20] {rolls/buns,chocolate,pastry}     => {Cluster4} 0.003184713 1          0.003184713 3.794562  24
```

Στο cluster 4 συναντάμε ως κυρίαρχο προϊόν τα ζυμαρικά, ενώ στα καλάθια έχουμε πιο μικρούς κανόνες. Αξιοσημείωτο είναι το confidence που στους 20 πρώτους κανόνες είναι σταθερά 1, γεγονός που δηλώνει μεγάλη ομοιομορφία μεταξύ των καλαθιών.

Cluster5:

```
> inspect(sort(apr_clust5, by = 'confidence'))
     lhs             rhs                  support    confidence coverage  lift      count
[1]  {Cluster5} => {whole milk}       0.05360934 0.3188635  0.1681263 0.9562097 404
[2]  {Cluster5} => {rolls/buns}       0.04299363 0.2557222  0.1681263 1.0652970 324
[3]  {Cluster5} => {other vegetables} 0.04033970 0.2399369  0.1681263 0.9501651 304
[4]  {Cluster5} => {soda}             0.03967622 0.2359905  0.1681263 1.0369823 299
[5]  {Cluster5} => {yogurt}           0.02839703 0.1689029  0.1681263 0.9277350 214
[6]  {Cluster5} => {bottled water}    0.02733546 0.1625888  0.1681263 1.1272025 206
[7]  {Cluster5} => {root vegetables}  0.02614119 0.1554854  0.1681263 1.0930391 197
[8]  {Cluster5} => {tropical fruit}   0.02361996 0.1404893  0.1681263 1.0258989 178
```

```
> inspect(sort(apr_clust5, by = 'confidence')[1:20])
     lhs                                 rhs         support     confidence coverage   lift      count
[1]  {chocolate}                      => {Cluster5} 0.014596603 0.2254098  0.06475584 1.3407171 110
[2]  {bottled water,soda}             => {Cluster5} 0.008492569 0.2245614  0.03781847 1.3356707  64
[3]  {whole milk,rolls/buns}          => {Cluster5} 0.016188960 0.2190305  0.07391189 1.3027735 122
[4]  {whole milk,root vegetables}     => {Cluster5} 0.013269639 0.2079002  0.06382696 1.2365714 100
[5]  {rolls/buns,soda}                => {Cluster5} 0.009686837 0.1936340  0.05002654 1.1517170  73
[6]  {bottled water}                  => {Cluster5} 0.027335456 0.1895124  0.14424098 1.1272025 206
[7]  {bottled water,yogurt}           => {Cluster5} 0.005573248 0.1858407  0.02998938 1.1053635  42
[8]  {rolls/buns,bottled water}       => {Cluster5} 0.005838641 0.1848739  0.03158174 1.0996133  44
[9]  {root vegetables}                => {Cluster5} 0.026141189 0.1837687  0.14225053 1.0930391 197
[10] {rolls/buns}                     => {Cluster5} 0.042993631 0.1791045  0.24004777 1.0652970 324
[11] {whole milk,soda}                => {Cluster5} 0.009156051 0.1751269  0.05228238 1.0416388  69
[12] {soda}                           => {Cluster5} 0.039676221 0.1743440  0.22757431 1.0369823 299
[13] {tropical fruit}                 => {Cluster5} 0.023619958 0.1724806  0.13694268 1.0258989 178
[14] {tropical fruit,other vegetables} => {Cluster5} 0.007563694 0.1614731  0.04684183 0.9604271  57
[15] {whole milk}                     => {Cluster5} 0.053609342 0.1607640  0.33346603 0.9562097 404
[16] {other vegetables,rolls/buns}    => {Cluster5} 0.008890658 0.1599045  0.05559979 0.9510975  67
[17] {other vegetables}               => {Cluster5} 0.040339703 0.1597478  0.25252123 0.9501651 304
[18] {other vegetables,yogurt}        => {Cluster5} 0.009023355 0.1592506  0.05666136 0.9472079  68
[19] {yogurt}                         => {Cluster5} 0.028397028 0.1559767  0.18205945 0.9277350 214
[20] {tropical fruit,whole milk}      => {Cluster5} 0.008492569 0.1538462  0.05520170 0.9150628  64
```

Σε αυτό το cluster συναντάμε κυρίως προϊόντα πρωινού (γάλα και rolls/buns), ενώ τα καλάθια είναι μικρά και συνδυάζουν κυρίως ποτά και λαχανικά.

Άσκηση 5. Συνδυαστική εφευρετικότητα: εφαρμογή μεθόδων Γραμμικού Προγραμματισμού σε αποτελέσματα ανάλυσης δεδομένων

Μια ιδέα για προωθητική ενέργεια θα ήταν η εξής:

Αρχικά θα πρέπει να βρούμε τα πιο ακριβά προϊόντα, με apriori που στο lhs θα έχει υποχρεωτικά το high_value_basket. Έτσι θα βρούμε τα προϊόντα που περιέχονται συχνότερα στα ακριβά καλάθια. Από τα πιο ακριβά, έστω 5, θα ορίσουμε την προσφορά μας. Για κάθε καλάθι, ανάλογα με το πόσα ακριβά προϊόντα περιέχει, και σε ποια θέση του πίνακα των ακριβότερων βρίσκονται, θα έχει ένα ποσοστό έκπτωσης στο σύνολο των υπόλοιπων προϊόντων. Πχ για το ακριβότερο προϊόν έκπτωση 10%, για το δεύτερο ακριβότερο 8%, για το 3^{ο} 6%, για το 4^{ο} 4% και για το 5^{ο} 2%. Το ποσοστό της έκπτωσης θα αθροίζει, ανάλογα με τα ακριβά προϊόντα του καλαθιού, αλλά δεν θα ξεπερνάει το 16%, εκτός αν περιέχει και τα 5 προϊόντα, οπότε θα γίνεται έκπτωση 25%. Η έκπτωση αφορά τα προϊόντα που βρίσκονται στο καλάθι και δεν ανήκουν στα πιο ακριβά.

Έτσι ο κόσμος θα αγοράζει πιο εύκολα τα ακριβά για να κερδίσει την έκπτωση, η οποία θα γίνεται μόνο στα πιο φθηνά, οπότε δεν θα ζημιώνεται το κατάστημα.