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)
  + generateAssociationRules(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 προ επεξεργασία, 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) (χρήση έτοιμης συνάρτησης **apply()** ).
* Μέσω της **function(x)**, γιακάθε 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.

1. 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)

1. 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 περιγραφεί 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.

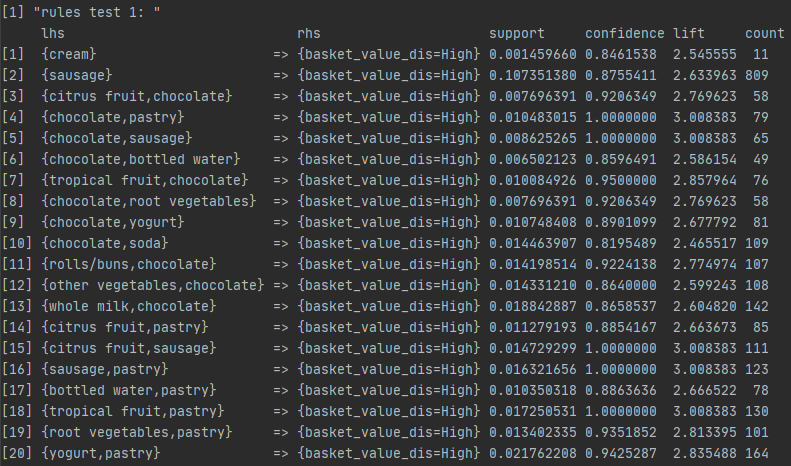
1. Πειραματισμός με διαφορετικές τιμές για το ελάχιστο Support

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):

1. Για **supp = 0.001**:



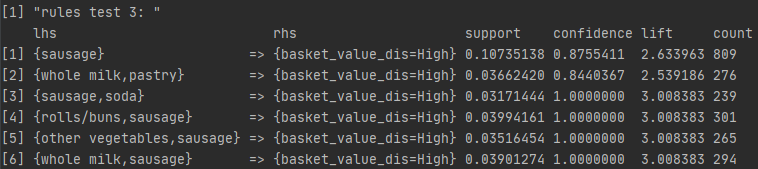
Αξίζει να σημειωθεί ότι για την συγκεκριμένη ελάχιστη τιμή του support, λαμβάνονται συνολικά 750 κανόνες συσχέτισης, από τους οποίους, όπως αναφέρθηκε και πριν, εξετάζονται οι 20.

1. Για **supp = 0.02**:



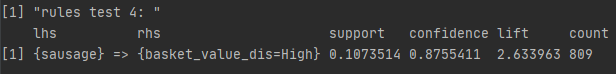
Παρατηρούμε ότι για την συγκεκριμένη ελάχιστη τιμή του support, εξάγονται μόλις 14 κανόνες συσχέτισης.

1. Για **supp = 0.03**:



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

1. Για **supp = 0.04**:



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

1. 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.

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:



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**».

1. 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.

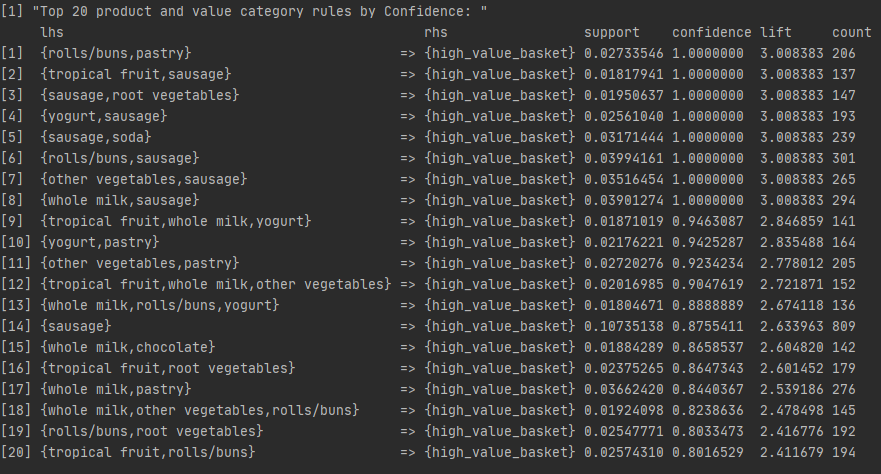
Θα εφαρμοστεί η ίδια λογική με την διαδικασία πειραματισμού που πραγματοποιήθηκε στην υποενότητα 1 της παρούσας άσκησης, ωστόσο αυτή τη φορά η επιλογή της τιμής του ελάχιστου Support παρουσιάζει αυξημένο ενδιαφέρον: Σκοπός μας είναι η επιλογή της μέγιστης δυνατής τιμής του **Support**, για την οποία 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:  


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:

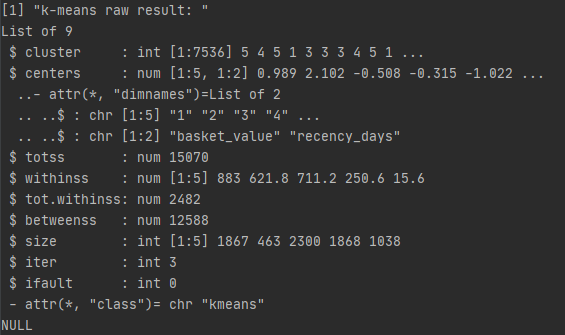
*set.seed*(1234)

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

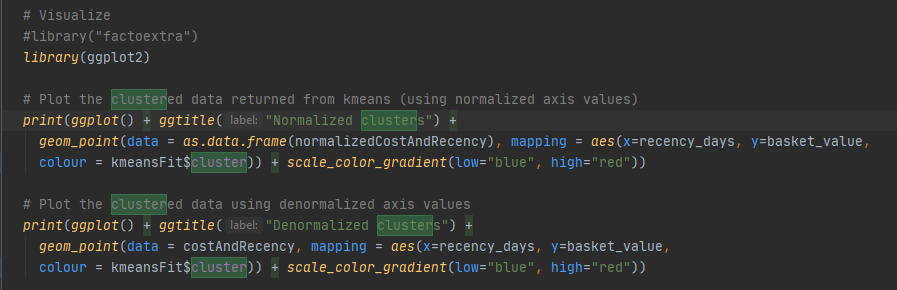
* **normalizedCostAndRecency**: Το σύνολο δεδομένων στο οποίο θέλουμε να εφαρμοστεί ο αλγόριθμος ομαδοποίησης k-means.
* **centers** = 5: Προσδιορίζουμε ότι επιθυμούμε την ομαδοποίηση των δεδομένων σε 5 συστάδες.
* **nstart** = 1000: Προσδιορίζουμε ότι επιθυμούμε 1000 δοκιμαστικές αρχικοποιήσεις του 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:

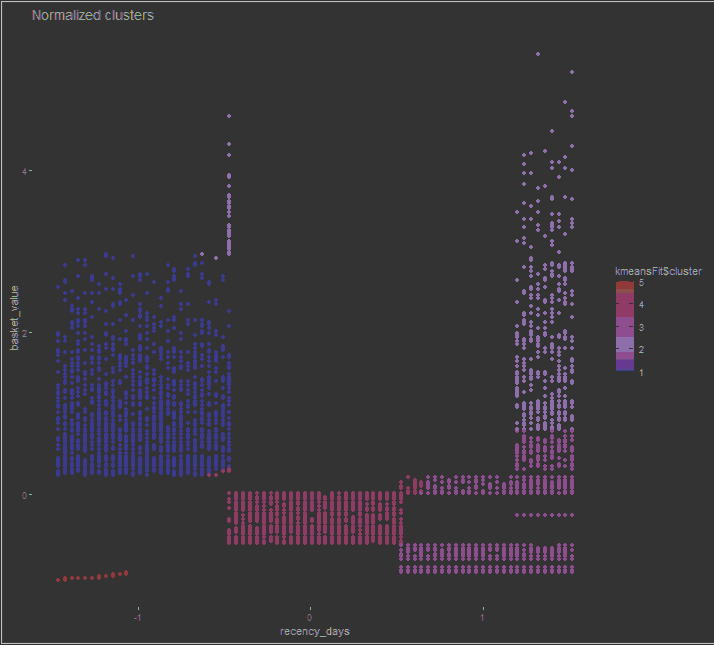


Για την καλύτερη συνολική απεικόνιση του αποτελέσματος και την εξαγωγή συμπερασμάτων, είναι αναγκαίο σε αυτή τη φάση να προχωρήσουμε στη δημιουργία διαγραμμάτων. Για αυτό τον σκοπό αξιοποιήθηκαν οι δυνατότητες της έτοιμης βιβλιοθήκης «**ggplot2**»:

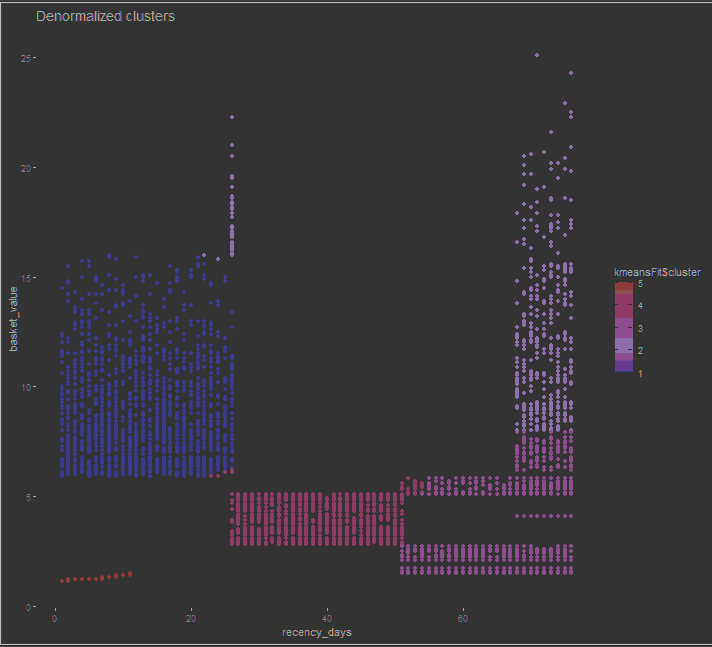


Από τον παραπάνω κώδικα, προκύπτουν τα εξής διαγράμματα:

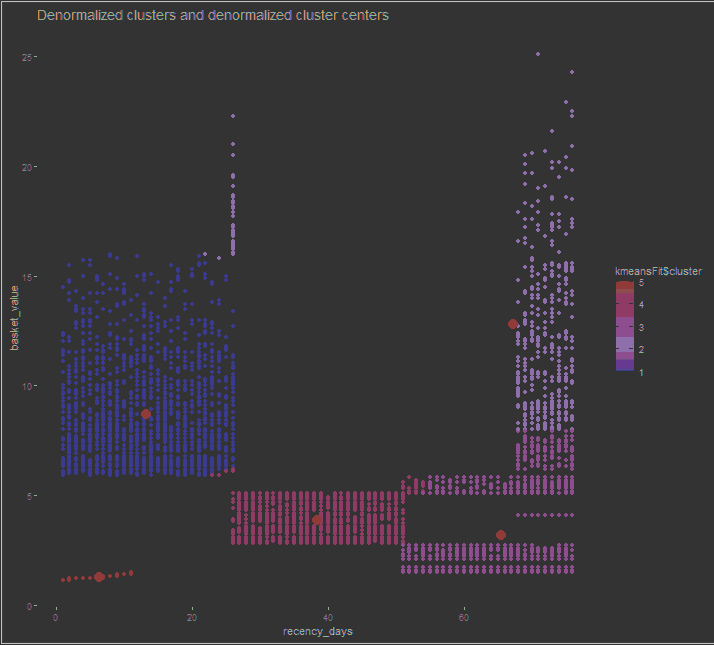
1. Διάγραμμα απεικόνισης συστάδων με κανονικοποιημένες τιμές



1. Διάγραμμα απεικόνισης συστάδων με τις αρχικές τιμές



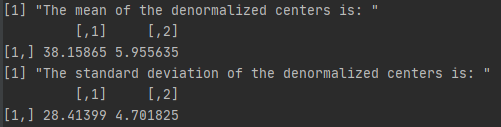
1. Διάγραμμα απεικόνισης συστάδων και των κέντρων τους σύμφωνα με τις αρχικές (μη κανονικοποιημένες) τιμές



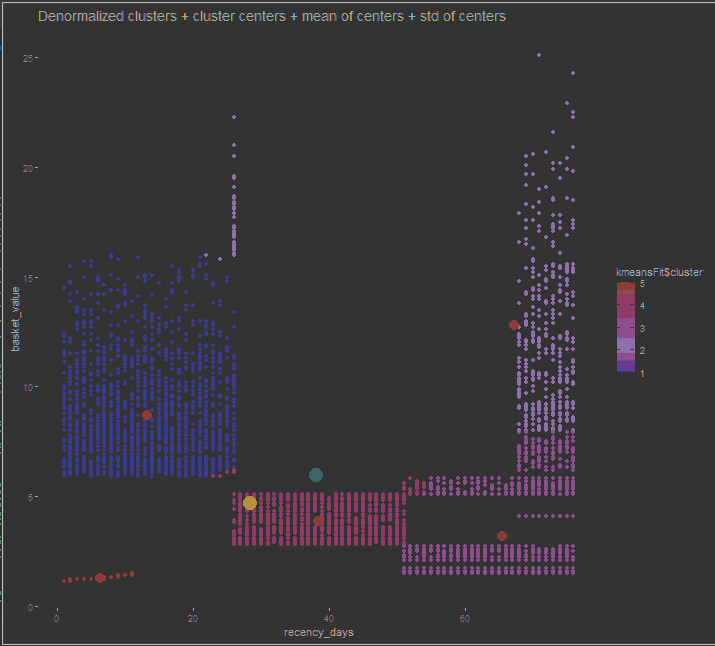
Στο παραπάνω διάγραμμα, οι **κόκκινες** βούλες συμβολίζουν τα κέντρα των συστάδων στην τελική τους μορφή.

1. Αναλυτική εξέταση του αποτελέσματος

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

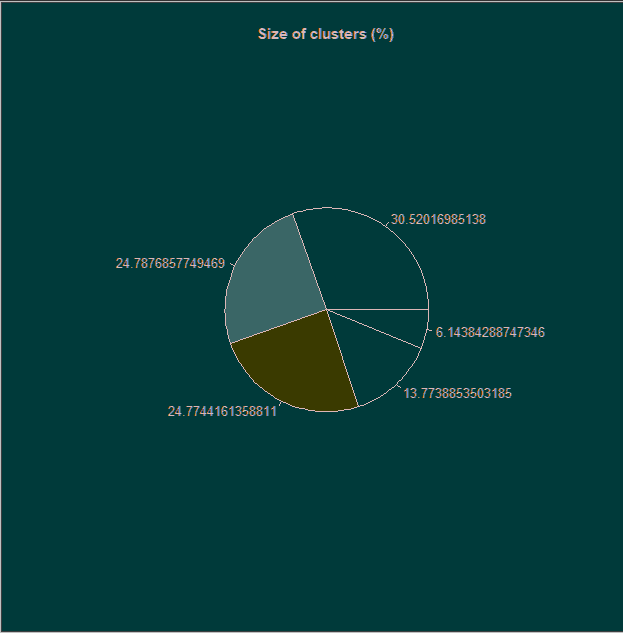


Για την καλύτερη κατανόηση του αποτελέσματος, προχωράμε και πάλι στη δημιουργία διαγράμματος απεικόνισης συστάδων που περιλαμβάνει τα νέα στοιχεία. Για την καλύτερη συσχέτιση του διαγράμματος με τα αρχικά δεδομένα, η κλίμακα τιμών του διαγράμματος σχηματίστηκε σύμφωνα με τις αρχικές (μη κανονικοποιημένες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



Διευκρινίζεται ότι το υπόμνημα των διαγραμμάτων διασποράς (scatter plots) έχει ταξινομηθεί κατά φθίνουσα σειρά μεγέθους ομάδας (Δηλαδή, η ομάδα 1 είναι η μεγαλύτερη σε μέγεθος, ενώ η ομάδα 5 η μικρότερη σε μέγεθος).

Το ίδιο ισχύει και για το pie chart (διάγραμμα πίτας), το οποίο παραθέτει τα μεγέθη των ομάδων σε φθίνουσα σειρά (ξεκινώντας από την ομάδα 1: 30.52% και με φορά αντίθετη του ρολογιού).

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 το **30.52%** 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 το **24.78%** 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 **24.77%** 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 το **13.77%** 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 το **6.14%** 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.

1. 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 **«ΤRUE»** 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.

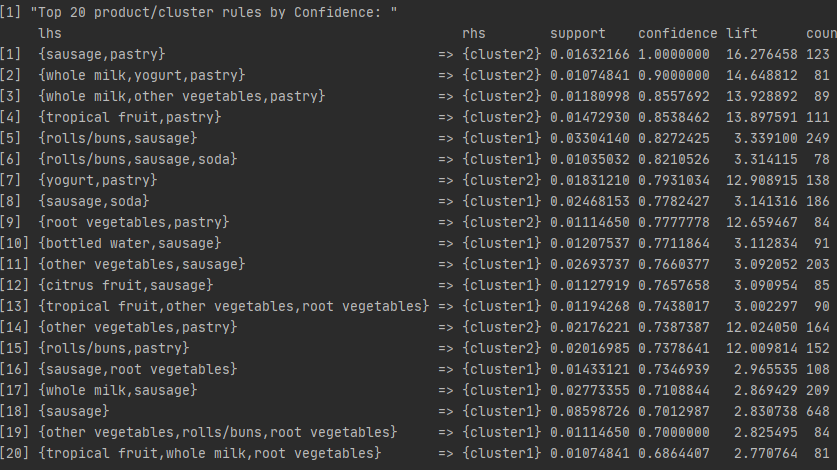
Στην περίπτωση μας, καταλήγουμε πειραματικά στην επιλογή της τιμής **0.01** για το **ελάχιστο support** και της τιμής **0.4** για το **ελάχιστο confidence.**

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:



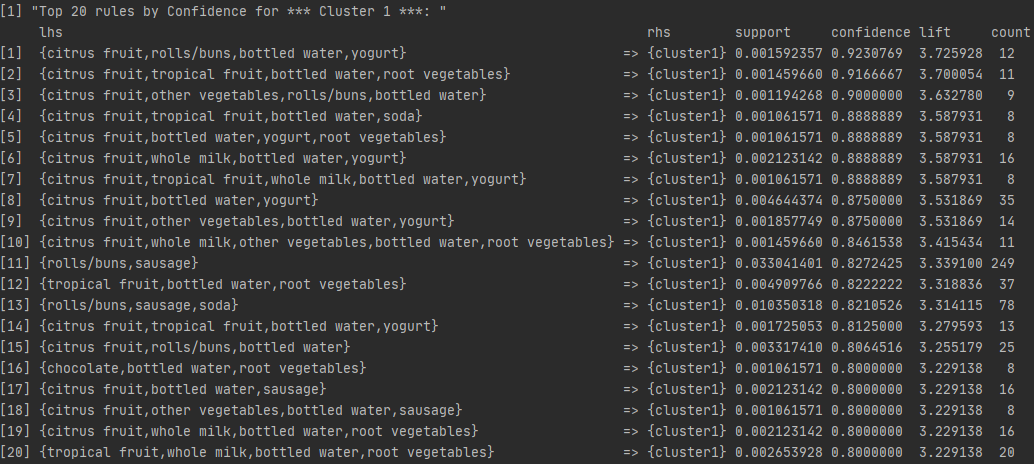
Ερμηνεύοντας το παραπάνω αποτέλεσμα, αξίζει να αναφερθούμε συνοπτικά στα εξής:

* 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).

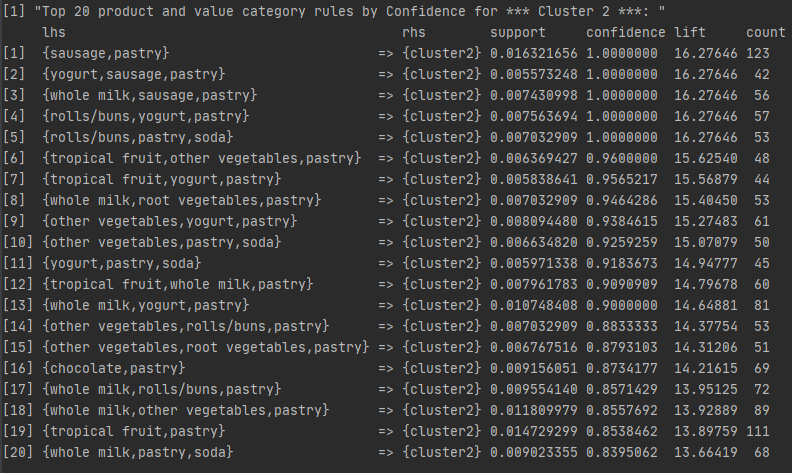
Στη συνέχεια, μας ζητείται να εντοπίσουμε ποια προϊόντα/συνδυασμοί προϊόντων αγοράζονται συχνότερα από την κάθε ομάδα. Η διαδικασία εξαγωγής των εστιασμένων κανόνων συσχέτισης για το παραπάνω ζητούμενο εκτελείται και πάλι από την συνάρτηση **clusterProductProfile(groceriesWithClusters)**.

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.

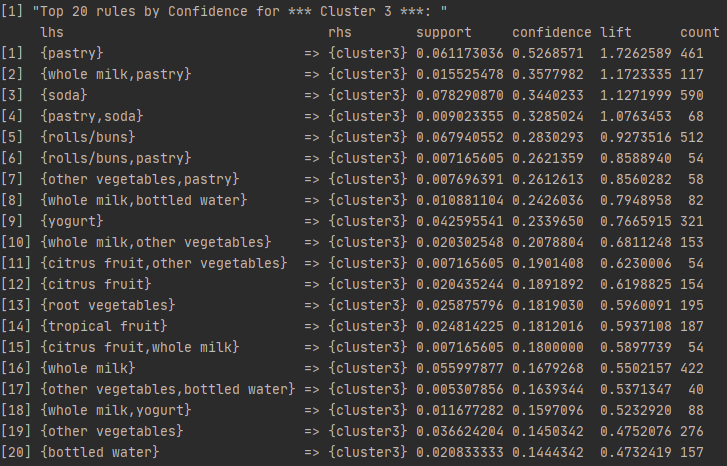
Για την **ομάδα συναλλαγών 1** (cluster1, supp=0.001):



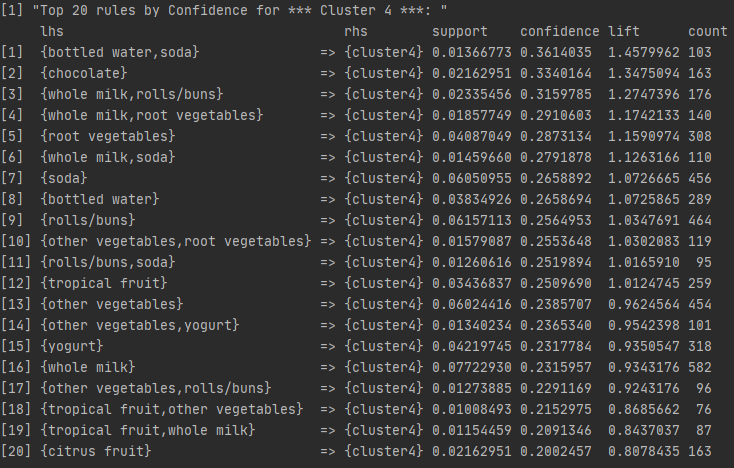
Για την **ομάδα συναλλαγών 2** (cluster2, supp=0.0055):



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



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



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

