ΠΜΣ στο Ψηφιακό Μάρκετινγκ

(MSc. In Digital Marketing)

Ανάλυση Δεδομένων Μεγάλου Όγκου και Μέθοδοι Έρευνας

Εργασία στο Μάθημα

Ανάλυση Δεδομένων Μεγάλου Όγκου και Μέθοδοι Έρευνας

Διδάσκοντες: Επικ. Καθηγητής Στυλιανός Κρηνίδης

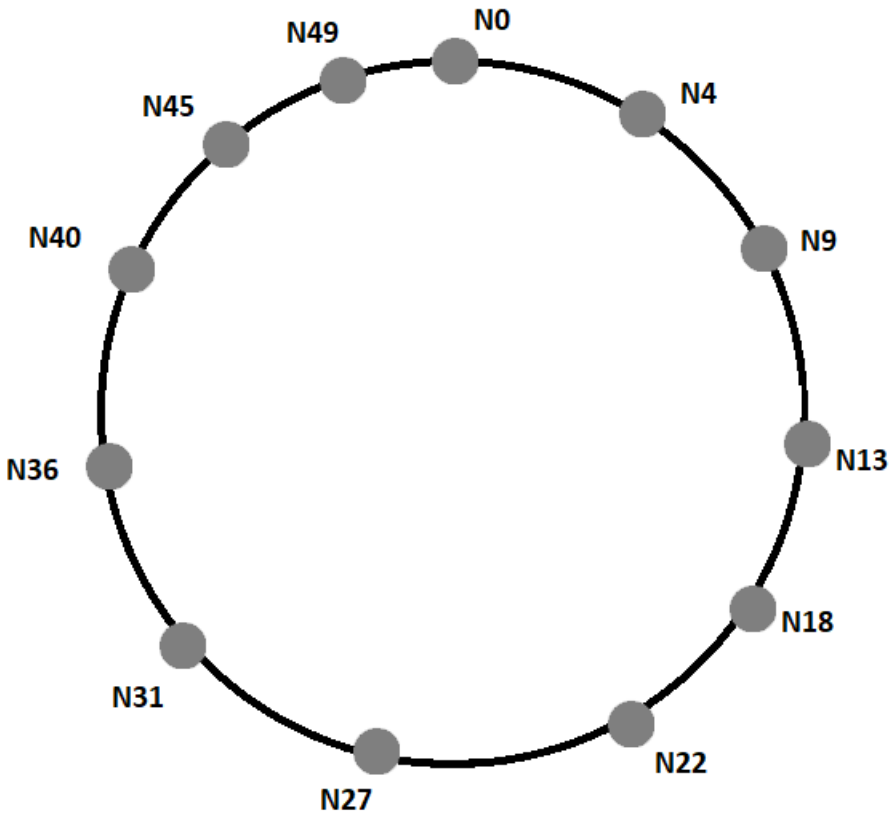
Δρ. Γεωργία Κούγκα

Φοιτητές: Καλαφάτη Ευανθία ΑΜ 140

Φωτοπούλου Ελένη ΑΜ 144

ΘΕΜΑ 1ο

Για το σχεδιασμό του CHORD δικτύου θα αναζητήσουμε τους κόμβους. Οι αριθμοί μητρώου μας είναι 140 και 144. Σύμφωνα με τους υπολογισμούς ο αριθμός των κόμβων είναι 12, οι οποίοι διαφαίνονται στον κύκλο.



Finger tables

Για την εύρεση των finger tables με m=6 για τον κάθε κόμβο υπολογίζουμε με τον τύπο finger[i] = successor (n + 2^(i-1)mod64. Τα finger tables παρουσιάζουν ποιοι κόμβοι συνδέονται μεταξύ τους σε ένα κατανεμημένο σύστημα ώστε να οδηγηθούμε πιο εύκολα και γρήγορα στις αναζητήσεις μας.

ΚΟΜΒΟΣ Ν0		**ΚΟΜΒΟΣ Ν4**		**ΚΟΜΒΟΣ Ν9**		**ΚΟΜΒΟΣ Ν13**		**ΚΟΜΒΟΣ Ν18**	
Ν0+1=1	**Ν4**	**Ν4+1=5**	**Ν9**	**Ν9+1=10**	**Ν13**	**Ν13+1=14**	**Ν18**	**Ν18+1=19**	**Ν22**
Ν0+2=2	**Ν4**	**Ν4+2=6**	**Ν9**	**Ν9+2=11**	**Ν13**	**Ν13+2=15**	**Ν18**	**Ν18+2=20**	**Ν22**
Ν0+4=4	**Ν4**	**Ν4+4=8**	**Ν9**	**Ν9+4=13**	**Ν13**	**Ν13+4=17**	**Ν18**	**Ν18+4=22**	**Ν22**
Ν0+8=8	**Ν9**	**Ν4+8=12**	**Ν13**	**Ν9+8=17**	**Ν18**	**Ν13+8=21**	**Ν22**	**Ν18+8=26**	**Ν27**
Ν0+16=16	**Ν18**	**Ν4+16=20**	**Ν22**	**Ν9+16=25**	**Ν27**	**Ν13+16=29**	**Ν31**	**Ν18+16=34**	**Ν36**
Ν0+32=32	**Ν36**	**Ν4+32=36**	**Ν40**	**Ν9+32=41**	**Ν45**	**Ν13+32=45**	**Ν45**	**Ν18+32=50**	**Ν0**

ΚΟΜΒΟΣ Ν22		**ΚΟΜΒΟΣ Ν27**		**ΚΟΜΒΟΣ Ν31**		**ΚΟΜΒΟΣ Ν36**		**ΚΟΜΒΟΣ Ν40**	
Ν22+1=23	**Ν27**	**Ν27+1=28**	**Ν31**	**Ν31+1=32**	**Ν36**	**Ν36+1=37**	**Ν40**	**Ν40+1=41**	**Ν45**
Ν22+2=24	**Ν27**	**Ν27+2=29**	**Ν31**	**Ν31+2=33**	**Ν36**	**Ν36+2=38**	**Ν40**	**Ν40+2=42**	**Ν45**
Ν22+4=26	**Ν27**	**Ν27+4=31**	**Ν31**	**Ν31+4=35**	**Ν36**	**Ν36+4=40**	**Ν40**	**Ν40+4=44**	**Ν45**
Ν22+8=30	**Ν31**	**Ν27+8=35**	**Ν36**	**Ν31+8=39**	**Ν40**	**Ν36+8=44**	**Ν45**	**Ν40+8=48**	**Ν49**
Ν22+16=38	**Ν40**	**Ν27+16=43**	**Ν45**	**Ν31+16=47**	**Ν49**	**Ν36+16=52**	**Ν0**	**Ν40+16=56**	**Ν0**
Ν22+32=54	**Ν0**	**Ν27+32=59**	**Ν0**	**Ν31+32=63**	**Ν0**	**Ν36+32=68**	**Ν4**	**Ν40+32=72**	**Ν9**

ΚΟΜΒΟΣ Ν45		**ΚΟΜΒΟΣ Ν49**	
Ν45+1=46	**Ν49**	**Ν49+1=50**	**Ν0**
Ν45+2=47	**Ν49**	**Ν49+2=51**	**Ν0**
Ν45+4=49	**Ν49**	**Ν49+4=53**	**Ν0**
Ν45+8=53	**Ν0**	**Ν49+8=57**	**Ν0**
Ν45+16=61	**Ν0**	**Ν49+16=1**	**Ν4**
Ν45+32=77	**Ν13**	**Ν49+32=17**	**Ν18**

Δεδομένου ότι οι ΑΜ1=140 και ΑΜ2=144, βάση των υπολογισμών οι αναζητήσεις θα είναι για Κ12 και Κ16 αντίστοιχα.

Για το Κ12: Ξεκινώντας από τον κόμβο Ν0, μέσω του finger table, οδηγούμαστε απευθείας στο κόμβο Ν9, ο οποίος είναι ο πλησιέστερος στον πιθανό κόμβο που περιέχει το κλειδί. Έτσι ελέγχουμε τον κόμβο Ν9 και τον διπλανό του κόμβο Ν13, όπου περιέχει το κλειδί Κ12.

Για το Κ16: Ξεκινώντας από τον κόμβο Ν0, μέσω του finger table, οδηγούμαστε στον κόμβο Ν9. Ελέγχοντας τον κόμβο Ν9, μέσω του finger table οδηγούμαστε στον κόμβο Ν18 όπου περιέχει το κλειδί Κ16. Διαφορετικά από τον κόμβο Ν0 μπορούμε να πάμε απευθείας στον κόμβο Ν18 όπου πιθανότατα θα περιέχει και το κλειδί που αναζητούμε, το Κ16.

ΘΕΜΑ 2^{ο}

Τα χαρακτηριστικά των δεδομένων είναι:

• Τύπος δεδομένων: cat"Ο τύπος των δεδομένων είναι:", class(Groceries), "\n"

Αποτέλεσμα: Ο τύπος των δεδομένων είναι: transactions

• Αριθμός συναλλαγών: cat("Ο αριθμός των συναλλαγών (καλαθιών) είναι:", length(Groceries), "\n")

Αποτέλεσμα: Ο αριθμός των συναλλαγών (καλαθιών) είναι: 9835

• Κατηγορίες προϊόντων: cat("Οι κατηγορίες των προϊόντων είναι:", itemLabels(Groceries), "\n")

Αποτέλεσμα: Οι κατηγορίες των προϊόντων είναι: frankfurter sausage liver loaf ham meat finished products organic sausage chicken turkey pork beef hamburger meat fish citrus fruit tropical fruit pip fruit grapes berries nuts/prunes root vegetables onions herbs other vegetables packaged fruit/vegetables whole milk butter curd dessert butter milk yogurt whipped/sour cream beverages UHT-milk condensed milk cream soft cheese sliced cheese hard cheese cream cheese processed cheese spread cheese curd cheese specialty cheese mayonnaise salad dressing tidbits frozen vegetables frozen fruits frozen meals frozen fish frozen chicken ice cream frozen dessert frozen potato products domestic eggs rolls/buns white bread brown bread pastry roll products semi-finished bread zwieback potato products flour salt rice pasta vinegar oil margarine specialty fat sugar artif. sweetener honey mustard ketchup spices soups ready soups Instant food products sauces cereals organic products baking powder preservation products pudding powder canned vegetables canned fruit pickled vegetables specialty vegetables jam sweet spreads meat spreads canned fish dog food cat food pet care baby food coffee instant coffee tea cocoa drinks bottled water soda misc. beverages fruit/vegetable juice syrup bottled beer canned beer brandy whisky liquor rum liqueur liquor (appetizer) white wine red/blush wine prosecco sparkling wine salty snack popcorn nut snack snack products long life bakery product waffles cake bar chewing gum chocolate cooking chocolate specialty chocolate specialty bar chocolate marshmallow candy seasonal products detergent softener decalcifier dish cleaner abrasive cleaner cleaner toilet cleaner bathroom cleaner hair spray dental care male cosmetics make up remover skin care female sanitary products baby cosmetics soap rubbing alcohol hygiene articles napkins dishes cookware kitchen utensil cling film/bags kitchen

towels house keeping products candles light bulbs sound storage medium newspapers photo/film pot plants flower soil/fertilizer flower (seeds) shopping bags bags

• Αριθμός μοναδικών προϊόντων: cat("Ο αριθμός των μοναδικών προϊόντων είναι:", length(itemLabels(Groceries)), "\n")

Αποτέλεσμα: Ο αριθμός των μοναδικών προϊόντων είναι: 169

• Μέσος αριθμός αντικειμένων ανά συναλλαγή: average_items_per_transaction <- sum(itemFrequency(Groceries)) / length(Groceries)

cat("Ο μέσος αριθμός αντικειμένων ανά συναλλαγή είναι:", average_items_per_transaction, "\n")

Αποτέλεσμα: Ο μέσος αριθμός αντικειμένων ανά συναλλαγή είναι: 0.0004483433

• Τα 5 συχνότερα προϊόντα που αγοράζονται: item_frequencies <- itemFrequency(Groceries)

top_5_items <- head(sort(item_frequencies, decreasing = TRUE), 5)

cat("Τα 5 πιο συχνά προϊόντα που αγοράζονται είναι:\n")for (i in 1:length(top_5_items)) {cat(i, ": ", names(top_5_items)[i], "\n")}

Αποτέλεσμα:

1 : whole milk

2 : other vegetables

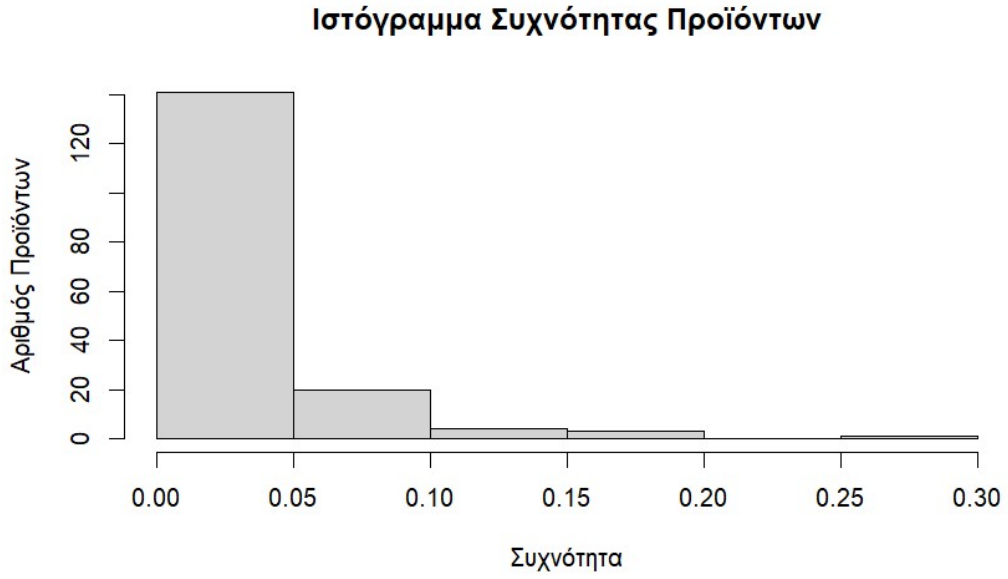
3 : rolls/buns

4 : soda

5 : yogurt

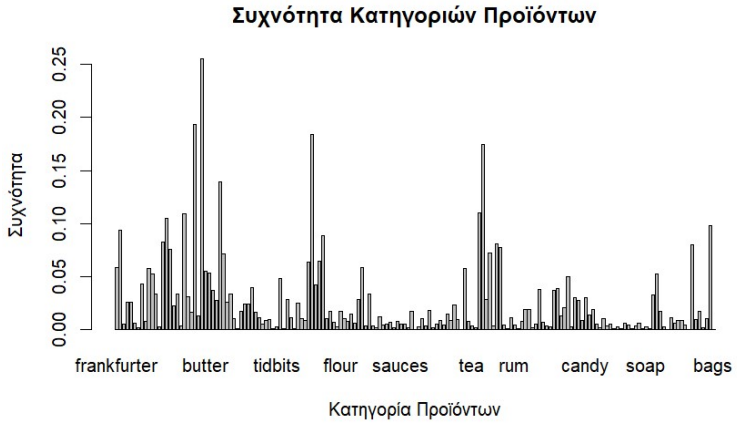
Ιστόγραμμα Συχνότητας Προϊόντων:

```
hist(itemFrequency(Groceries), main = "Ιστόγραμμα Συχνότητας Προϊόντων", xlab =
"Συχνότητα", ylab = "Αριθμός Προϊόντων")
```



Γράφημα ραβδογράμματος (bar plot) που εμφανίζει τη συχνότητα κάθε κατηγορίας προϊόντων:

```
item_frequencies <- itemFrequency(Groceries)

barplot(item_frequencies, main = "Συχνότητα Κατηγοριών Προϊόντων", xlab = "Κατηγορία
Προϊόντων", ylab = "Συχνότητα")
```



ΕΡΩΤΗΣΕΙΣ:

1. Να εντοπίσετε το index των προϊόντων citrus fruit, semi-finished bread, margarine, ready, soups.

Απάντηση: Για τον εντοπισμό του index των προϊόντων :

```
data("Groceries")

items <- itemLabels(Groceries)

products <- c("citrus fruit", "semi-finished bread", "margarine", "ready soups")

indices <- which(items %in% products)

indices
```

Αποτέλεσμα: [1] 14 61 70 79

Οι δείκτες των προϊόντων "citrus fruit", "semi-finished bread", "margarine" και "ready soups" στο dataset 'Groceries' είναι οι εξής:

"citrus fruit" - Δείκτης 14

"semi-finished bread" - Δείκτης 61

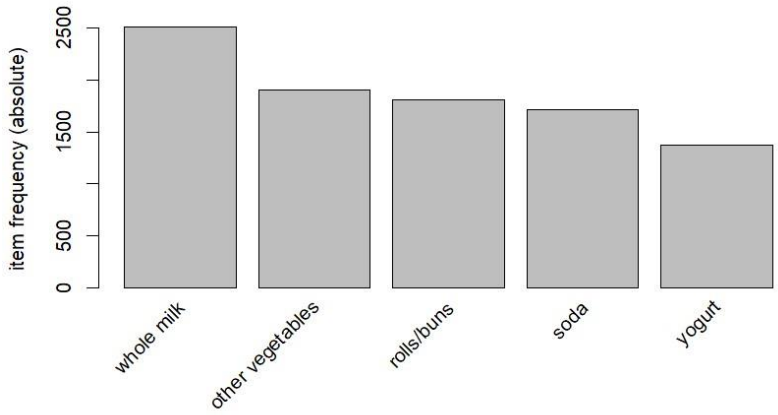
"margarine" - Δείκτης 70

"ready soups" - Δείκτης 79

2. Να δημιουργήσετε τη γραφική παράσταση με τα προϊόντα που εμφανίζονται πιο συχνά σε αυτό το σύνολο δεδομένων και να αναφέρετε ποια είναι αυτά.

```
data("Groceries")

itemFrequencyPlot(Groceries, topN = 5, type = "absolute")
```

Οι πιο συχνές κατηγορίες προϊόντων που θα εμφανιστούν στον άξονα των x και ο αριθμός των συναλλαγών που περιέχουν αυτές τις κατηγορίες θα εμφανιστεί στον άξονα των y. Με βάση τα δεδομένα μας, η γραφική παράσταση θα περιέχει τις 5 πιο συχνές κατηγορίες προϊόντων και τον αριθμό των συναλλαγών που τις περιλαμβάνουν.



3. <u>Να βρείτε ποια είναι αυτά που αγοράζονται σπάνια.</u>

```
data("Groceries")

frequency <- itemFrequency(Groceries)

rare_items <- names(frequency[frequency < 0.01]) # Επιλέγουμε τα προϊόντα που έχουν συχνότητα < 0.01 rare_items
```

<u>Αποτέλεσμα</u>:

[1] "liver loaf" "finished products" "organic sausage"

[4] "turkey" "fish" "nuts/prunes"

[7] "cream" "curd cheese" "specialty cheese"

[10] "mayonnaise" "salad dressing" "tidbits"

[13] "frozen fruits" "frozen chicken" "frozen potato products"

[16] "zwieback" "potato products" "rice"

[19] "vinegar" "specialty fat" "artif. sweetener"

[22] "honey" "ketchup" "spices"

[25] "soups" "ready soups" "Instant food products"

[28] "sauces" "cereals" "organic products"

[31] "preservation products" "pudding powder" "canned fruit"

[34] "specialty vegetables" "jam" "sweet spreads"

[37] "meat spreads" "dog food" "pet care"

[40] "baby food" "instant coffee" "tea"

[43] "cocoa drinks" "syrup" "brandy"

[46] "whisky" "rum" "liqueur"

[49] "liquor (appetizer)" "prosecco" "sparkling wine"

[52] "popcorn" "nut snack" "snack products"

[55] "cooking chocolate" "chocolate marshmallow" "softener"

[58] "decalcifier" "abrasive cleaner" "cleaner"

[61] "toilet cleaner" "bathroom cleaner" "hair spray"

[64] "dental care" "male cosmetics" "make up remover"

[67] "skin care" "female sanitary products" "baby cosmetics"

[70] "soap" "rubbing alcohol" "cookware"

[73] "kitchen utensil" "kitchen towels" "house keeping products"

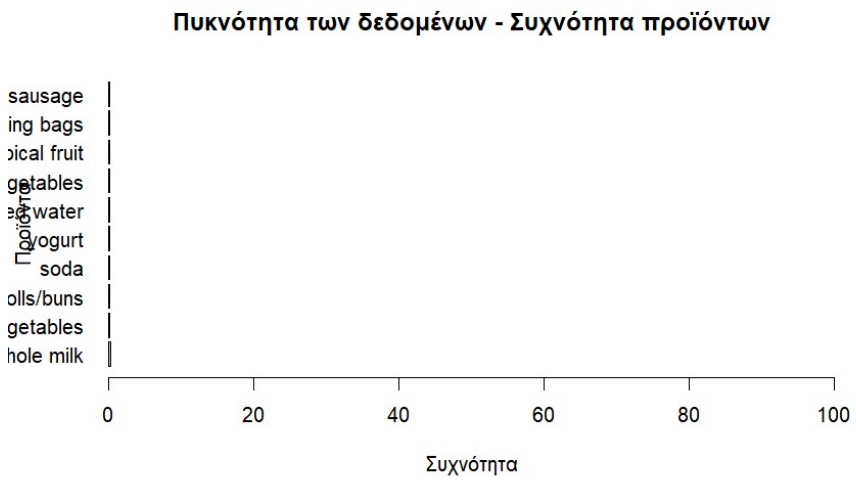
[76] "candles" "light bulbs" "sound storage medium"

[79] "photo/film" "flower soil/fertilizer" "bags"

4. Να δείξετε την πυκνότητα ή αραιότητα των δεδομένων.

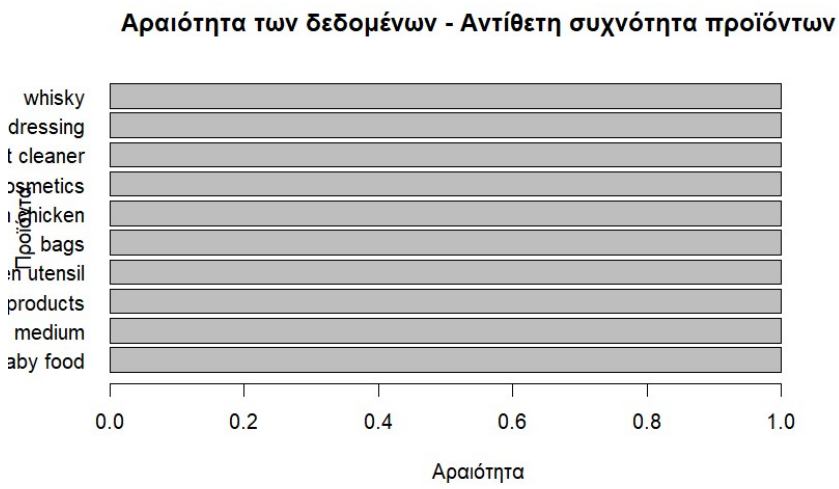
Για την απεικόνιση της **πυκνότητας** των δεδομένων μπορούμε να χρησιμοποιήσουμε τη συνάρτηση itemFrequency() από το πακέτο "arules" για να υπολογίσουμε τη συχνότητα εμφάνισης των προϊόντων. Στη συνέχεια, μπορούμε να δημιουργήσουμε ένα γράφημα ραβδογράμματος για να οπτικοποιήσουμε αυτήν την πληροφορία.

```
# Υπολογισμός της συχνότητας για κάθε προϊόν

product_freq <- itemFrequency(Groceries)


# Ταξινόμηση των προϊόντων βάσει της συχνότητας

sorted_products <- sort(product_freq, decreasing = TRUE)


# Δημιουργία γραφήματος ραβδογράμματος

barplot(sorted_products[1:10], horiz = TRUE, las = 1,

      main = "Πυκνότητα των δεδομένων - Συχνότητα προϊόντων",

      xlab = "Συχνότητα", ylab = "Προϊόντα",

      xlim = c(0, max(sorted_products[1:10]) + 100))
```



Για την απεικόνιση της **αραιότητας** των δεδομένων μπορούμε να ταξινομήσουμε τα προϊόντα βάση της αραιότητας τους και να δημιουργήσουμε ένα γράφημα ραβδογράμματος που απεικονίζει την αραιότητα των 10 πιο αραιών προϊόντων. Θα χρησιμοποιήσουμε την αντίθετη τιμή της συχνότητας, δηλαδή το (1 - συχνότητα), για να μετρήσουμε την αραιότητα. Όσο χαμηλότερη είναι η τιμή, τόσο πιο αραιό είναι το προϊόν.

```
# Υπολογισμός της αραιότητας για κάθε προϊόν

product_sparsity <- 1 - product_freq



# Ταξινόμηση των προϊόντων βάσει της αραιότητας

sorted_products_sparsity <- sort(product_sparsity, decreasing = TRUE)



# Δημιουργία γραφήματος ραβδογράμματος

barplot(sorted_products_sparsity[1:10], horiz = TRUE, las = 1,

      main = "Αραιότητα των δεδομένων - Αντίθετη συχνότητα προϊόντων",

      xlab = "Αραιότητα", ylab = "Προϊόντα",

      xlim = c(0, max(sorted_products_sparsity[1:10]) + 0.1))
```



5. Να εφαρμόσετε τον αλγόριθμο Apriori με support level 0.006, ελάχιστο length 2 και ελάχιστο confidence 0.25.

Εισάγουμε το πακέτο 'arules' και μετατρέπουμε το dataset 'Groceries' σε transactions. Έπειτα, εφαρμόζουμε τον αλγόριθμο Apriori με τις παραμέτρους που ορίστηκαν (support = 0.006, minlen = 2, confidence = 0.25). Η συνάρτηση inspect() χρησιμοποιείται για να εμφανίσει τους κανόνες σε ένα ευανάγνωστο μορφοποιημένο τρόπο.

```
# Εισαγωγή του πακέτου 'arules'

library(arules)



# Μετατροπή του dataset σε transactions

transactions <- as(Groceries, "transactions")



# Εφαρμογή του αλγορίθμου Apriori

rules <- apriori(transactions, parameter = list(support = 0.006, minlen = 2, confidence = 0.25))



# Εκτύπωση των ανακτηθέντων κανόνων

inspect(rules)
```

6. Να εντοπίσετε τα top 5 σύνολα στοιχείων που έχουν το υψηλότερο επίπεδο confidence.

Για να εντοπίσουμε τα top 5 σύνολα στοιχείων με το υψηλότερο επίπεδο confidence από τους ανακτηθέντες κανόνες, μπορούμε να χρησιμοποιήσουμε τη συνάρτηση sort() για να τα ταξινομήσουμε βάσει του επιπέδου confidence και να επιλέξουμε τα πρώτα 5 στοιχεία.

```
# Ταξινόμηση των κανόνων βάσει του επιπέδου confidence

sorted_rules <- sort(rules, by = "confidence", decreasing = TRUE)
```

```
# Επιλογή των πρώτων 5 κανόνων με το υψηλότερο επίπεδο confidence

top_5_confidence <- sorted_rules[1:5]


# Εκτύπωση των top 5 κανόνων

inspect(top_5_confidence)
```

Αποτέλεσμα:

[1] {butter, whipped/sour cream} => {whole milk} 0.006710727 0.6600000 0.010167768 2.583008

[2] {butter, yogurt} => {whole milk} 0.009354347 0.6388889 0.014641586 2.500387

[3] {root vegetables, butter} => {whole milk} 0.008235892 0.6377953 0.012913066 2.496107

[4] {tropical fruit, curd} => {whole milk} 0.006507372 0.6336634 0.010269446 2.479936

[5] {tropical fruit, butter} => {whole milk} 0.006202339 0.6224490 0.009964413 2.436047

Count

[1] 66

[2] 92

[3] 81

[4] 64

[5] 61

7. Ποιο είναι το μοτίβο των αγορών του καταναλωτή;

Το μοτίβο των αγορών του καταναλωτή είναι η τάση αγοράς συγκεκριμένων προϊόντων που συνδέονται μεταξύ τους. Με βάση τα δεδομένα και τα αποτελέσματα που αναφέρθηκαν προηγουμένως, μπορούμε να παρατηρήσουμε ότι ένα από τα δημοφιλέστερα μοτίβα αγορών του καταναλωτή είναι η αγορά του προϊόντος "whole milk" με την συνοδεία των προϊόντων "butter" και "whipped/sour cream". Άλλα δημοφιλή μοτίβα περιλαμβάνουν την αγορά "whole milk" με συνοδευτικά προϊόντα όπως "yogurt", "root vegetables" και "tropical fruit".

8. Ποιοι είναι οι συνδυασμοί που οδηγούν στην αγορά γάλακτος;

Βασισμένοι στα αποτελέσματα του αλγορίθμου Apriori, οι συνδυασμοί που οδηγούν στην αγορά γάλακτος είναι οι εξής:

{butter, whipped/sour cream} => {whole milk}

{butter, yogurt} => {whole milk}

{root vegetables, butter} => {whole milk}

{tropical fruit, curd} => {whole milk}

{tropical fruit, butter} => {whole milk}

Αυτοί οι συνδυασμοί προϊόντων υποδηλώνουν ότι η αγορά των προϊόντων butter, whipped/sour cream, yogurt, root vegetables, και tropical fruit συχνά συνοδεύεται από την αγορά γάλακτος (whole milk).

9. Μπορείτε να δημιουργήσετε τη γραφική απεικόνιση αυτού του συνδυασμού;

Για την γραφική απεικόνιση του συνδυασμού αυτού, αρχικά θα εγκαταστήσουμε το πακέτο "arulesViz".

```
install.packages("arulesViz")
```

Στη συνέχεια θα δημιουργήσουμε ένα γράφημα δέντρου που απεικονίζει τη συσχέτιση μεταξύ των προϊόντων που οδηγούν στην αγορά γάλακτος. Οι κόμβοι του γραφήματος αντιπροσωπεύουν τα προϊόντα, ενώ οι ακμές δείχνουν τις συνδέσεις μεταξύ τους. Όσο πιο κοντά είναι δύο προϊόντα στο γράφημα, τόσο πιο συχνά συνοδεύονται από την αγορά γάλακτος.

```
library(arulesViz)

library(arules)



# Δημιουργία των συνδυασμών που οδηγούν στην αγορά γάλακτος

milk_combinations <- apriori(data = Groceries, parameter = list(supp = 0.006, minlen = 2, conf = 0.25),

                             appearance = list(lhs = c("butter", "whipped/sour cream", "yogurt", "root vegetables", "tropical fruit"),

                                               rhs = "whole milk"))



# Απεικόνιση του γραφήματος δέντρου

plot(milk_combinations, method = "graph")
```