ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΨΗΦΙΑΚΟ ΜΑΡΚΕΤΙΝΓΚ

(DIGITAL MARKETING)

Μάθημα: Ανάλυση δεδομένων μεγάλου όγκου και μέθοδοι έρευνας (ARM)

Εξάμηνο: Β' ΕΞΑΜΗΝΟ

Διδάσκοντες:
Δρ. Κούγκα Γεωργία
Δρ. Στυλιανός Κρηνίδης

ΔΙΚΤΥΑ CHORD & ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΤΗ ΓΛΩΣΣΑ R

Των Μεταπτυχιακών Φοιτητριών:

1. Άννα Μαρία Βαρσάμη – ΑΕΜ: [303]

2. Μυρσίνη Γκίνη – ΑΕΜ: [296]

Σημείωση: Για την υλοποίηση της εργασίας μας χρησιμοποιήθηκαν τα ΑΕΜ μας 303 και 296.

Περιεχόμενα

ΘΕΜΑ 1ο: Ανάλυση Δικτύου CHORD

1. Υπολογισμός Πλήθους Ενεργών Κόμβων

Για την εκκίνηση της μελέτης του CHORD δικτύου, βασιστήκαμε στα Ακαδημαϊκά Μητρώα της ομάδας μας (ΑΜ1 = 296 και ΑΜ2 = 303). Σύμφωνα με τις προδιαγραφές της εκφώνησης, το συνολικό πλήθος των ενεργών κόμβων που απάρτισαν το peer-to-peer σύστημά μας προέκυψε από τη μαθηματική σχέση:

Πλήθος Κόμβων = [(ΑΜ1 + ΑΜ2) (mod 8)] + 8

Αρχικά, υπολογίσαμε το άθροισμα των δύο μητρώων, το οποίο ισούταν με 599 (296 + 303 = 599). Στη συνέχεια, βρήκαμε το υπόλοιπο της διαίρεσης αυτού του αριθμού με το 8, κάνοντας την πράξη 599 (mod 8). Επειδή το 8 χωράει 74 ολόκληρες φορές μέσα στο 599 (αφού 74 x 8 = 592), μας περίσσεψε ένα αριθμητικό υπόλοιπο ίσο με 7 (καθώς 599 - 592 = 7).

Αντικαθιστώντας αυτή την τιμή στον αρχικό τύπο και προσθέτοντας τη σταθερά 8, καταλήξαμε στο παρακάτω αποτέλεσμα:

Πλήθος Κόμβων = 7 + 8 = 15

Το παραπάνω θεωρητικό αποτέλεσμα διασταυρώθηκε πλήρως με τη χρήση του αρχείου Excel «Κόμβοι». Αφού καταχωρίσαμε τα ΑΜ μας στα αντίστοιχα κελιά, το υπολογιστικό φύλλο εκτέλεσε αυτόματα τις πράξεις και μας επέστρεψε την τιμή 15 στον αριθμό των κόμβων. Παράλληλα, με την ενημέρωση αυτής της τιμής, ανανεώθηκαν αυτόματα και όλες οι υπόλοιπες στήλες του αρχείου με βάση τα δικά μας δεδομένα, επιβεβαιώνοντας έτσι τους υπολογισμούς μας.

2. Εντοπισμός των Θέσεων των Κόμβων στον Δακτύλιο

Το CHORD δίκτυο που μελετήσαμε βασίστηκε σε έναν νοητό κύκλο 64 πιθανών θέσεων. Στα συστήματα αυτά, η αρίθμηση των θέσεων ξεκινάει συμβατικά από το μηδέν, με αποτέλεσμα οι διαθέσιμες θέσεις να εκτείνονται από το 0 έως το 63.

Οι πραγματικές θέσεις των κόμβων προέκυψαν αυτόματα στο φύλλο εργασίας μέσω της συνάρτησης κατακερματισμού που έχει ενσωματωθεί στο αρχείο Excel. Η συνάρτηση αυτή, λαμβάνοντας ως παραμέτρους εισόδου τα Ακαδημαϊκά Μητρώα της ομάδας μας, εξέτασε κάθε μία από τις 64 πιθανές θέσεις του κύκλου και προσδιόρισε ποιες από αυτές θα παραμείνουν ενεργές, αποκτώντας τη δυαδική τιμή 1, και ποιες θα θεωρηθούν κενές, παίρνοντας την τιμή 0.

Με την εισαγωγή των Ακαδημαϊκών μας Μητρώων, το υπολογιστικό φύλλο ενεργοποίησε αυτόματα 15 συγκεκριμένους κόμβους. Για την καταγραφή τους, ανατρέξαμε στον πίνακα του Βήματος 3 και απομονώσαμε τους αριθμούς από τη στήλη «Θέση Κόμβων» για τις 15 πορτοκαλί γραμμές.

Όπως παρατηρήσαμε, ο πρώτος ενεργός κόμβος του συστήματός μας τοποθετήθηκε ακριβώς στην αρχή του δακτυλίου, δηλαδή στη θέση 0. Συνολικά, οι 15 πραγματικές θέσεις των υπολογιστών της ομάδας μας διαμορφώθηκαν ως εξής:

N0, N3, N7, N10, N14, N17, N21, N24, N28, N31, N35, N38, N42, N45, N49

Καθώς η εκφώνηση όρισε ότι η διαδικασία αναζήτησης πρέπει να εκκινήσει από τον κόμβο N0, και με δεδομένο ότι ο κόμβος αυτός ήταν υπαρκτός και ενεργός στο δίκτυό μας, η εκκίνηση έγινε κανονικά από αυτόν.

3. Υπολογισμός των Κλειδιών Αναζήτησης (ΚΧ και ΚΨ)

Πριν προχωρήσουμε στη διαδικασία της δρομολόγησης των μηνυμάτων μέσα στον δακτύλιο, προσδιορίσαμε τους αριθμούς των δύο κλειδιών - αρχείων που έπρεπε να εντοπιστούν στο δίκτυο. Για τον σκοπό αυτό, εφαρμόσαμε τη συνάρτηση Mod64 στα δύο Ακαδημαϊκά Μητρώα της ομάδας μας, σύμφωνα με τις οδηγίες της εκφώνησης:

- Πρώτο Κλειδί (ΚΧ): Υπολογίσαμε την παράμετρο Χ μέσω της πράξης ΑΜ1 mod 64, δηλαδή 296 mod 64. Καθώς ο αριθμός 64 χώρεσε 4 ολόκληρες φορές μέσα στο 296 (4 * 64 = 256), προέκυψε αριθμητικό υπόλοιπο ίσο με 40 (296 - 256 = 40). Συνεπώς, το πρώτο αναζητούμενο αρχείο αντιστοιχούσε στο κλειδί Κ40.
- Δεύτερο Κλειδί (ΚΨ): Αντίστοιχα, υπολογίσαμε την παράμετρο Ψ μέσω της πράξης ΑΜ2 mod 64, δηλαδή 303 mod 64. Ο αριθμός 64 χώρεσε επίσης 4 φορές μέσα στο 303 (4 * 64 = 256), αφήνοντας αριθμητικό υπόλοιπο ίσο με 47 (303 - 256 = 47). Συνεπώς, το δεύτερο αναζητούμενο αρχείο αντιστοιχούσε στο κλειδί Κ47.

4. Κατασκευή του Πίνακα Δακτύλων (Finger Table) για τον Κόμβο N0

Για τη διεκπεραίωση της δρομολόγησης των μηνυμάτων, κάθε ενεργός σταθμός του δικτύου CHORD οφείλει να διατηρεί έναν Πίνακα Δακτύλων (Finger Table) μεγέθους m=6. Για τον αρχικό κόμβο N0, ο οποίος αποτέλεσε την αφετηρία των αναζητήσεών μας, οι θεωρητικές θέσεις-στόχοι προσδιορίστηκαν βάσει του τύπου (N + 2i-1) (mod 64).

Για κάθε μία από τις 6 γραμμές του πίνακα, αναζητήσαμε τον πρώτο πραγματικά ενεργό κόμβο του δικτύου μας που συναντάται αν κινηθούμε με δεξιόστροφη φορά στον δακτύλιο, ξεκινώντας από τη θέση-στόχο. Με βάση τη διαμορφωμένη λίστα των 15 κόμβων μας, οι υπολογισμοί εξελίχθηκαν ως εξής:

- Γραμμή 1 (i=1): Η θεωρητική θέση-στόχος ήταν $0 + 2^0 = 1$. Ο πλησιέστερος ενεργός κόμβος που εντοπίστηκε μετά τη θέση 1 ήταν ο N3.
- Γραμμή 2 (i=2): Η θεωρητική θέση-στόχος ήταν $0 + 2^1 = 2$. Ο πλησιέστερος ενεργός κόμβος παρέμεινε ο N3.
- Γραμμή 3 (i=3): Η θεωρητική θέση-στόχος ήταν $0 + 2^2 = 4$. Καθώς οι θέσεις 4, 5 και 6 ήταν κενές, ο πρώτος υπαρκτός κόμβος που συναντήσαμε ήταν ο N7.
- Γραμμή 4 (i=4): Η θεωρητική θέση-στόχος ήταν $0 + 2^3 = 8$. Ο πρώτος ενεργός κόμβος που εντοπίστηκε μετά τη θέση 8 ήταν ο N10.
- Γραμμή 5 (i=5): Η θεωρητική θέση-στόχος ήταν $0 + 2^4 = 16$. Ο πρώτος ενεργός κόμβος που βρέθηκε στον δακτύλιο μετά τη θέση 16 ήταν ο N17.

- Γραμμή 6 (i=6): Η θεωρητική θέση-στόχος ήταν $0 + 2^5 = 32$. Ο πλησιέστερος ενεργός κόμβος μετά τη θέση 32 ήταν ο N35.

4.1 Αναλυτικό Παράδειγμα Υπολογισμού για τον Κόμβο N0

Για την κατανόηση της μεθοδολογίας, παρουσιάζεται αρχικά ο αναλυτικός υπολογισμός των εγγραφών δρομολόγησης για τον κεντρικό κόμβο N0. Χρησιμοποιώντας τη μαθηματική σχέση $(N + 2^{i-1}) \pmod{64}$ για τις 6 γραμμές (m=6), προσδιορίζεται η θέση στόχος και στη συνέχεια εντοπίζεται ο πλησιέστερος ενεργός κόμβος που είναι υπεύθυνος για αυτήν τη θέση στον δακτύλιο CHORD:

Δείκτης (i)	**Θέση Στόχος**	**Πραγματικός Κόμβος**
1	1	N3
2	2	N3
3	4	N7
4	8	N10
5	16	N17
6	32	N35

4.2 Συγκεντρωτικός Πίνακας Δακτύλων (Finger Table) για Όλους τους Κόμβους

Εφαρμόζοντας την ίδια ακριβώς μαθηματική διαδικασία και για τους 15 ενεργούς κόμβους του δικτύου μας, προκύπτουν οι ακόλουθοι δάκτυλοι δρομολόγησης:

Ενεργός Κόμβος	**Δάκτυλος 1 (i=1)**	**Δάκτυλος 2 (i=2)**	**Δάκτυλος 3 (i=3)**	**Δάκτυλος 4 (i=4)**	**Δάκτυλος 5 (i=5)**	**Δάκτυλος 6 (i=6)**
N0	N3	N3	N7	N10	N17	N35
N3	N7	N7	N7	N14	N21	N35
N7	N10	N10	N14	N17	N24	N42
N10	N14	N14	N14	N21	N28	N42
N14	N17	N17	N21	N24	N31	N49
N17	N21	N21	N21	N28	N35	N49
N21	N24	N24	N28	N31	N38	N0
N24	N28	N28	N28	N35	N42	N0
N28	N31	N31	N35	N38	N45	N0
N31	N35	N35	N35	N42	N49	N0
N35	N38	N38	N42	N45	N0	N3
N38	N42	N42	N42	N49	N0	N7
N42	N45	N45	N49	N0	N0	N10
N45	N49	N49	N49	N0	N0	N14
N49	N0	N0	N0	N0	N3	N17

5. Διαδρομές Αναζήτησης Κλειδιών (K40 και K47)

Αφού ολοκληρώσαμε την κατασκευή του πίνακα δακτύλων για τον κόμβο εκκίνησης N0, προχωρήσαμε στη διαδικασία εντοπισμού των δύο κλειδιών, K40 και K47, μέσα στον δακτύλιο. Η αναζήτηση εκτελέστηκε με βάση τη λογική της απλής δρομολόγησης του CHORD, όπου κάθε κόμβος προωθεί το αίτημα στον πλησιέστερο γνωστό του κόμβο που δεν ξεπερνά την τιμή του αναζητούμενου κλειδιού.

Α. Διαδρομή Αναζήτησης για το Κλειδί KX (K40)

Ο κόμβος N0 εκκίνησε την αναζήτηση για το κλειδί 40. Καθώς στο δίκτυο δεν υπήρχε ενεργός σταθμός στη θέση 40, το κλειδί αυτό όφειλε να αποθηκευτεί στον πρώτο ενεργό κόμβο που το διαδέχεται στον δακτύλιο, ο οποίος με βάση τα δεδομένα μας ήταν ο N42. Η δρομολόγηση εξελίχθηκε σε 3 βήματα ως εξής:

- Βήμα 1 (N0 → N35): Ο κόμβος N0 εξέτασε τον πίνακα δακτύλων του και εντόπισε τον πλησιέστερο γνωστό του κόμβο που προηγείται του στόχου χωρίς να τον ξεπερνά. Αυτός ήταν ο κόμβος N35 (θέση 32 στον πίνακα). Συνεπώς, το αίτημα προωθήθηκε στον N35.
- Βήμα 2 (N35 → N38): Ο κόμβος N35 έλαβε το αίτημα και συμβουλεύτηκε τις δικές του εγγραφές. Ο πλησιέστερος ενεργός κόμβος που προηγείται αυστηρά του 40 ήταν ο N38, οπότε το αίτημα μεταβιβάστηκε σε αυτόν.
- Βήμα 3 (N38 → N42): Ο κόμβος N38, έχοντας ως άμεσο διάδοχο στον δακτύλιο τον κόμβο N42, αναγνώρισε ότι η θέση-στόχος καλύπτεται από αυτόν. Το αίτημα παραδόθηκε στον N42, ο οποίος επιβεβαίωσε την κατοχή του κλειδιού K40 και η αναζήτηση τερματίστηκε.

Β. Διαδρομή Αναζήτησης για το Κλειδί ΚΨ (Κ47)

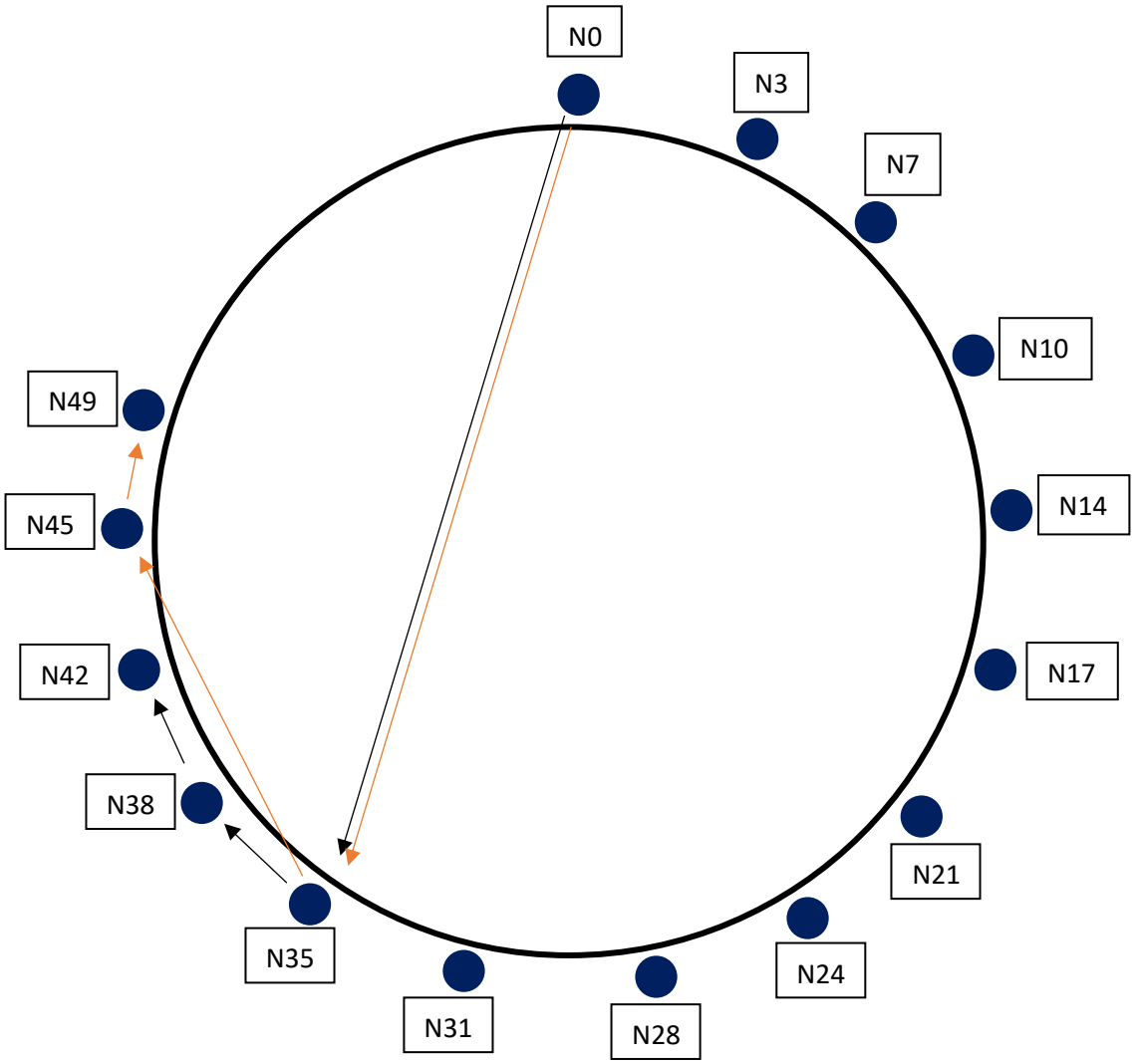
Ο κόμβος Ν0 εκκίνησε τη δεύτερη αναζήτηση για το κλειδί 47. Καθώς δεν υπήρχε ενεργός σταθμός στις θέσεις 47 ή 48, το κλειδί αυτό έπρεπε να καταλήξει στον αμέσως επόμενο διαθέσιμο κόμβο του δακτυλίου, δηλαδή στον Ν49. Η ροή της δρομολόγησης πραγματοποιήθηκε σε 3 βήματα ως εξής:

- Βήμα 1 (Ν0 → Ν35): Ο κόμβος Ν0 ανέτρεξε στον πίνακα δακτύλων του. Ο μεγαλύτερος ενεργός κόμβος που δεν ξεπερνούσε την τιμή 47 ήταν και πάλι ο Ν35, προς τον οποίο και προωθήθηκε το μήνυμα.
- Βήμα 2 (Ν35 → Ν45): Ο κόμβος Ν35 έλαβε το αίτημα και εξέτασε τον δικό του πίνακα. Ο πλησιέστερος γνωστός του σταθμός που προηγούνταν του 47 ήταν ο Ν45 (καθώς ο επόμενος δικός του δείκτης θα οδηγούσε στον Ν49, ο οποίος ξεπερνά την τιμή του κλειδιού). Το αίτημα στάλθηκε στον Ν45.
- Βήμα 3 (Ν45 → Ν49): Ο κόμβος Ν45 διαπίστωσε ότι ο άμεσος διάδοχός του στον δακτύλιο είναι ο κόμβος Ν49. Εφόσον η τιμή του κλειδιού (47) ήταν μικρότερη από το αναγνωριστικό του Ν49, το αίτημα προωθήθηκε στον τελικό αυτό κόμβο, ο οποίος ανέλαβε την αποθήκευση του κλειδιού Κ47.

6. Γραφική Αναπαράσταση του Δικτύου Chord και των Διαδρομών Αναζήτησης

Για την ολοκληρωμένη παρουσίαση του συστήματος, σχεδιάστηκε ο δακτύλιος του δικτύου CHORD για m=6 (64 συνολικές θέσεις αναγνωριστικών, από 0 έως 63). Στο διάγραμμα αποτυπώνονται οι 15 πραγματικά ενεργοί κόμβοι της ομάδας μας στις ακριβείς τους θέσεις, όπως αυτοί προέκυψαν από τους υπολογισμούς με βάση τα Ακαδημαϊκά Μητρώα.

Επιπλέον, εσωτερικά του δακτυλίου έχουν σχεδιαστεί οι δύο διαδρομές δρομολόγησης για τα κλειδιά K40 και K47 με αφετηρία τον κόμβο N0, αναδεικνύοντας γραφικά τα άλματα που εκτέλεσε το πρωτόκολλο μέχρι τον τελικό εντοπισμό των κόμβων διαδόχων(N42 και N49 αντίστοιχα).



Υπόμνημα Διαγράμματος:

- **Μαύρος Κύκλος:** Ο δακτύλιος αναγνωριστικών του CHORD (64 θέσεις).
- **Μπλε Κόμβοι:** Οι 15 ενεργοί σταθμοί του δικτύου (N0 έως N49).
- **Μαύρα Βέλη:** Η διαδρομή αναζήτησης για το κλειδί K40 (N0 → N35 → N38 → N42).
- **Πορτοκαλί Βέλη:** Η διαδρομή αναζήτησης για το κλειδί K47 (N0 → N35 → N45 →N49).

7. Συμπεράσματα & Αξιοσημείωτες Παρατηρήσεις

Με την ολοκλήρωση της προσομοίωσης του δικτύου CHORD για τους 15 ενεργούς κόμβους της ομάδας μας, προκύπτουν ορισμένα εξαιρετικά ενδιαφέροντα και αξιοσημείωτα συμπεράσματα σχετικά με τη συμπεριφορά του συστήματος:

- Η Δύναμη των Συντομεύσεων (Finger Tables): Το πιο εντυπωσιακό στοιχείο φάνηκε στο πρώτο κιόλας βήμα της αναζήτησης και για τα δύο κλειδιά (K40 και K47). Ξεκινώντας από τον κόμβο N0, το σύστημα δεν χρειάστηκε να ψάξει σειριακά τους κόμβους έναν προς έναν. Αντίθετα, χάρη στον 6ο δάκτυλο, έκανε ένα τεράστιο «άλμα» απευθείας στον N35. Με αυτό το ένα και μοναδικό βήμα, προσπεράστηκαν ακαριαία 9 ενεργοί κόμβοι (N3, N7, N10, N14, N17, N21, N24, N28, N31), αποδεικνύοντας έμπρακτα πώς το CHORD εκμηδενίζει τις αποστάσεις και τον χρόνο.
- Επαλήθευση της Λογαριθμικής Πολυπλοκότητας O: Παρά το γεγονός ότι ο αναγνωριστικός χώρος του δικτύου είναι σχετικά μεγάλος (26 = 64 πιθανές θέσεις), η αναζήτηση και για τα δύο κλειδιά απαιτούσε ακριβώς 3 βήματα μέχρι τον τελικό προορισμό. Αυτό αποτελεί την τέλεια πρακτική επιβεβαίωση της θεωρίας των Κατανεμημένων Πινάκων Κατακερματισμού, οι οποίοι εγγυώνται εντοπισμό δεδομένων σε λογαριθμικό χρόνο, καθιστώντας το δίκτυο εξαιρετικά αποδοτικό.
- Σύγκλιση Διαδρομών: Παρατηρήθηκε ότι για δύο διαφορετικά κλειδιά (K40 και K47), η αρχική διαδρομή που ακολουθήθηκε ήταν ακριβώς η ίδια (N0 → N35). Αυτό συμβαίνει επειδή το CHORD καθοδηγεί τα μηνύματα προσεγγιστικά προς την περιοχή του στόχου. Μόλις το μήνυμα φτάσει στην ευρύτερη γειτονιά του κλειδιού (στον N35), τότε η διαδρομή διακλαδίζεται (ο N35 στέλνει το K40 στον N38, ενώ το K47 στον N45), κάνοντας πλέον μικρά και ακριβή βήματα για τον τελικό εντοπισμό.
- Διαχείριση της Αραιότητας του Δικτύου: Από τους 64 πιθανούς κόμβους, στο δίκτυό μας είναι ενεργοί μόνο οι 15 και ότι το δίκτυο μας είναι αρκετά αραιό. Η προσομοίωση έδειξε ότι ο κανόνας του επόμενου υπαρκτού σταθμού λειτουργεί άψογα. Ακόμα κι αν οι μαθηματικοί στόχοι έπεφταν σε άδεια σημεία (π.χ. θέση 32), το πρωτόκολλο μετέφερε την ευθύνη στον αμέσως επόμενο ενεργό κόμβο (N35) χωρίς να προκληθεί καμία ασυνέχεια ή σφάλμα στην αναζήτηση.

ΘΕΜΑ 2ο : Ανάλυση δεδομένων με τη χρήση του R Studio

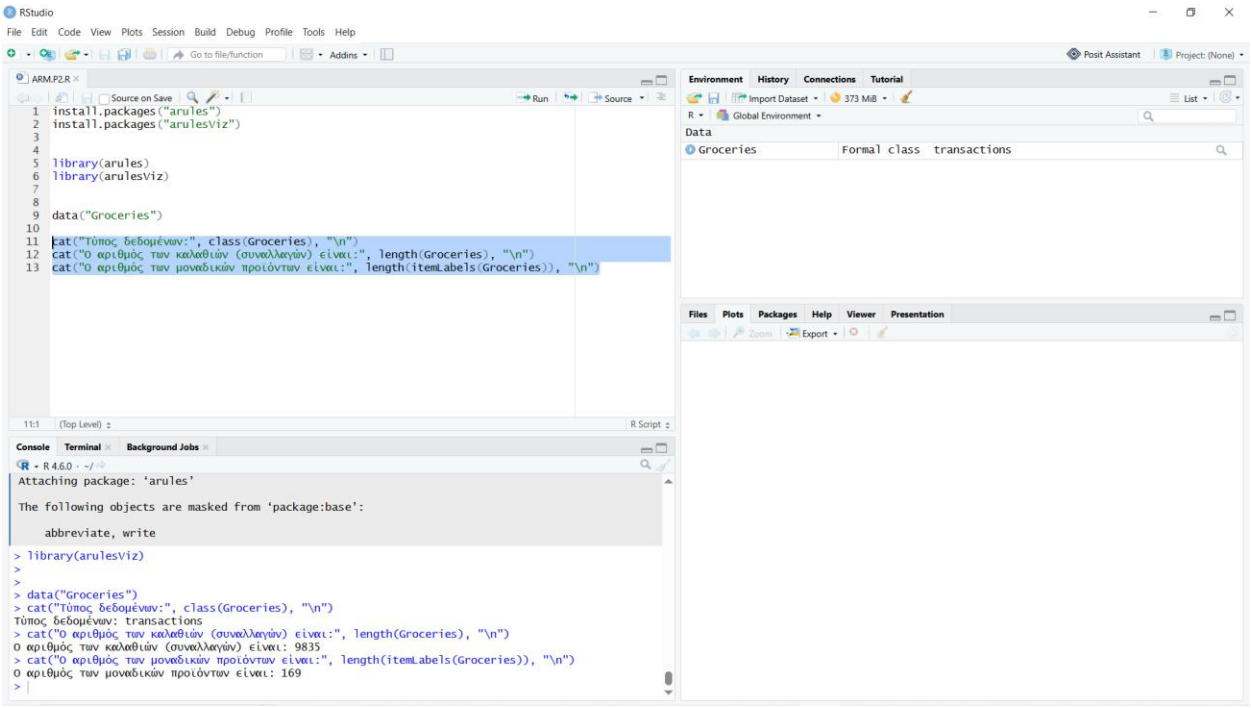
-Ερώτημα: Φόρτωση, μελέτη των δεδομένων "Groceries" και καταγραφή των χαρακτηριστικών τους (τύπος δεδομένων, αριθμός συναλλαγών, κατηγορίες προϊόντων).

1. Κώδικας στην R

Για τη φόρτωση του πακέτου arules, την εισαγωγή του dataset και την εξαγωγή των βασικών χαρακτηριστικών του με τη χρήση της συνάρτησης cat(), χρησιμοποιήθηκε ο ακόλουθος κώδικας:

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

library(arules)
library(arulesViz)

# Φόρτωση του ενσωματωμένου dataset 'Groceries'
data("Groceries")

#Προβολή βασικών πληροφοριών με τη συνάρτηση cat()
cat("Τύπος δεδομένων (Class):", class(Groceries), "\n")
cat("Ο αριθμός των καλαθιών (συναλλαγών) είναι:", length(Groceries), "\n")
cat("Ο αριθμός των μοναδικών προϊόντων είναι:", length(itemLabels(Groceries)), "\n")
```

2. Αποτέλεσμα



3. Επεξήγηση & Συμπεράσματα

- **Τύπος Δεδομένων (transactions):** Το dataset δεν είναι ένας απλός πίνακας (Data Frame), αλλά ανήκει στην ειδική κλάση transactions του πακέτου arules. Η δομή αυτή είναι ένας αραιός πίνακας (sparse matrix), ο οποίος είναι βελτιστοποιημένος για την αποθήκευση δεδομένων αγορών, καθώς εξοικονομεί μνήμη καταγράφοντας μόνο τα προϊόντα που όντως αγοράστηκαν σε κάθε συναλλαγή.
- **Αριθμός Συναλλαγών:** Το σύνολο δεδομένων περιλαμβάνει συνολικά **9.835 καλάθια αγορών** (αγορές διαφορετικών καταναλωτών). Το μέγεθος του δείγματος είναι επαρκώς μεγάλο για την εξαγωγή στατιστικά ασφαλών συμπερασμάτων.
- **Μοναδικά Προϊόντα:** Στο κατάστημα υπάρχουν συνολικά **169 διαφορετικές κατηγορίες προϊόντων** (item labels) που μπορούν να επιλέξουν οι καταναλωτές.

-Ερώτημα: Πυκνότητα ή αραιότητα των δεδομένων.

1. Κώδικας στην R

summary(Groceries)

2. Αποτέλεσμα

```
> data("Groceries")
> cat("Τύπος δεδομένων:", class(Groceries), "\n")
Τύπος δεδομένων: transactions
> cat("Ο αριθμός των καλαθιών (συναλλαγών) είναι:", length(Grocer
Ο αριθμός των καλαθιών (συναλλαγών) είναι: 9835
> cat("Ο αριθμός των μοναδικών προϊόντων είναι:", length(itemLabe
Ο αριθμός των μοναδικών προϊόντων είναι: 169
> summary(Groceries)
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146
```

density : 0.02609146

3. Επεξήγηση & Συμπεράσματα

- **Πυκνότητα (Density = 0,02609146 ή 2,61%):** Το ποσοστό αυτό μας δείχνει τον αριθμό των μη μηδενικών κελιών στον πίνακα «Συναλλαγές Χ Προϊόντα». Σημαίνει ότι από όλους τους πιθανούς συνδυασμούς θέσεων, μόνο το 2,61% περιέχει κάποιο προϊόν.
- **Αραιότητα (Sparsity = 1 - 0,02609146 = 0,97390854 ή 97,39%):** Το υπόλοιπο 97,39% του πίνακα είναι εντελώς κενό (μηδενικά στοιχεία).

Συμπέρασμα: Η ανάλυση δείχνει ότι το dataset είναι εξαιρετικά **αραιό (sparse)**. Αυτό είναι ένα κλασικό χαρακτηριστικό των δεδομένων αγορών (Market Basket Data), καθώς ο μέσος καταναλωτής αγοράζει ελάχιστα προϊόντα (συνήθως 3 με 5 είδη) σε κάθε του επίσκεψη, σε σύγκριση με το σύνολο των 169 διαθέσιμων προϊόντων του καταστήματος.

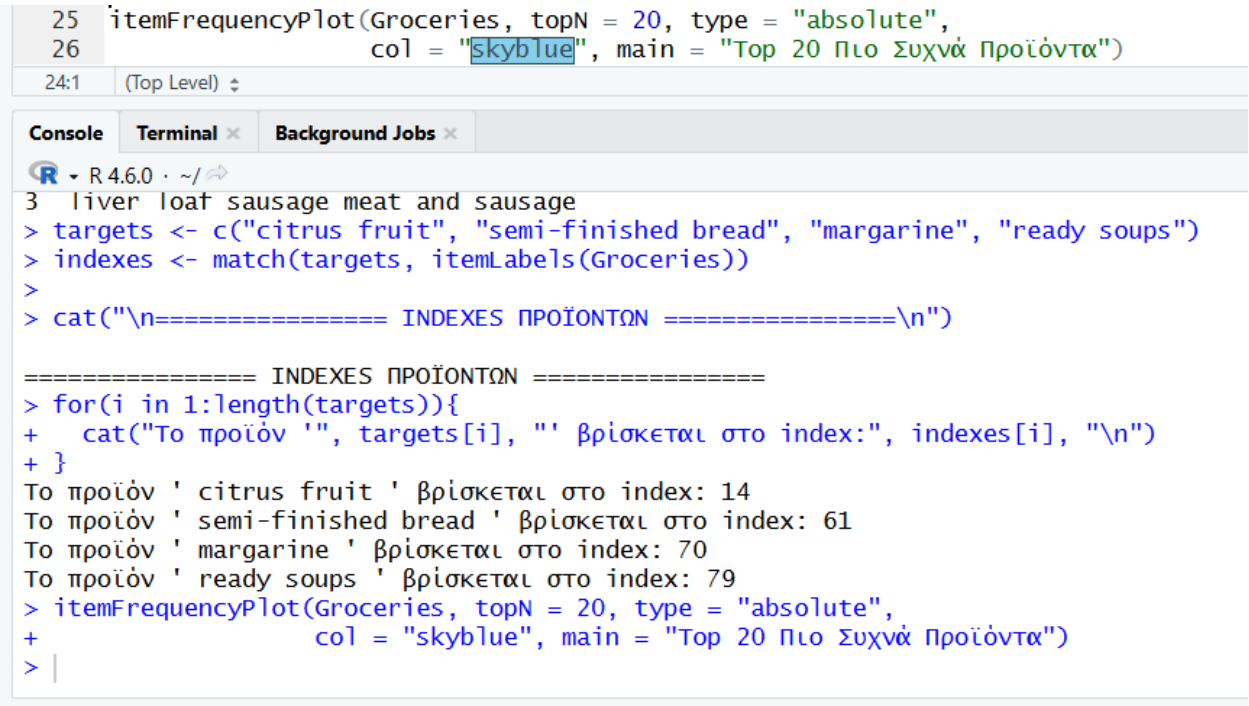
-Ερώτημα: Εντοπισμός του index των προϊόντων citrus fruit, semi-finished bread, margarine, ready soups.

1. Κώδικας στην R

Για να βρούμε την ακριβή θέση (index) των συγκεκριμένων τεσσάρων προϊόντων μέσα στη λίστα των 169 προϊόντων του dataset, χρησιμοποιήθηκε η συνάρτηση match():

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

indexes <- match(targets, itemLabels(Groceries))


cat("\n================ INDEXES ΠΡΟΪΟΝΤΩΝ ================\n")

for(i in 1:length(targets)){

  cat("Το προϊόν '", targets[i], "' βρίσκεται στο index:", indexes[i], "\n")

}
```

2. Αποτέλεσμα



3. Επεξήγηση & Συμπεράσματα

Η συνάρτηση match() επέστρεψε τις ακριβείς θέσεις (indices) των προϊόντων στον πίνακα αναγνώρισης του πακέτου arules. Ο εντοπισμός αυτός μας επιτρέπει να γνωρίζουμε πώς είναι κωδικοποιημένα τα προϊόντα στο dataset. Συγκεκριμένα:

- Το **citrus fruit** (εσπεριδοειδή) είναι το 14ο προϊόν.
- Το **semi-finished bread** (μισοψημένο ψωμί) είναι το 61ο προϊόν.
- Η **margarine** (μαργαρίνη) είναι το 70ο προϊόν.
- Οι **ready soups** (έτοιμες σούπες) είναι το 79ο προϊόν.

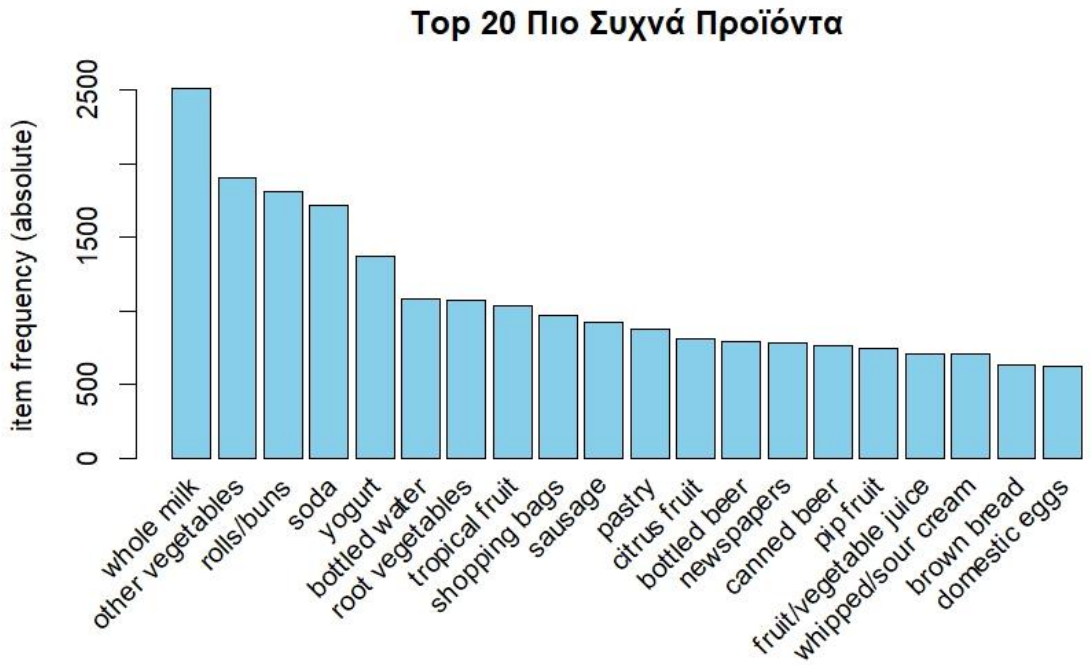
-Ερώτημα: Δημιουργία γραφικής παράστασης με τα προϊόντα που εμφανίζονται πιο συχνά και αναφορά αυτών.

1. Κώδικας στην R

Για την οπτικοποίηση των πιο συχνών προϊόντων στο σύνολο δεδομένων Groceries, χρησιμοποιήθηκε ο ακόλουθος κώδικας:

```
itemFrequencyPlot(Groceries, topN = 20, type = "absolute",
                  col = "skyblue", main = "Top 20 Πιο Συχνά Προϊόντα")
```

2. Αποτέλεσμα (Γράφημα)



3. Επεξήγηση & Συμπεράσματα

Με βάση το παραπάνω ραβδόγραμμα (Barplot), προκύπτει η απόλυτη συχνότητα εμφάνισης των προϊόντων στις 9.835 συναλλαγές του καταστήματος. Τα πέντε (5) πιο συχνά αγοραζόμενα προϊόντα, τα οποία ξεχωρίζουν ξεκάθαρα, είναι τα εξής:

1. **whole milk (Πλήρες Γάλα):** Εμφανίζεται σε περισσότερες από 2.500 συναλλαγές, αποτελώντας το πιο δημοφιλές προϊόν του dataset.
2. **other vegetables (Άλλα Λαχανικά):** Ακολουθεί στη δεύτερη θέση με σχεδόν 2.000 εμφανίσεις.
3. **rolls/buns (Ψωμάκια/Ρολά):** Βρίσκεται στην τρίτη θέση, λίγο κάτω από τις 2.000 εμφανίσεις.
4. **soda (Αναψυκτικά):** Εμφανίζει υψηλή συχνότητα, πλησιάζοντας τις 1.700 συναλλαγές.
5. **yogurt (Γιαούρτι):** Συμπληρώνει την πρώτη πεντάδα με περίπου 1.400 εμφανίσεις.

Συμπέρασμα Ανάλυσης: Η γραφική παράσταση αποκαλύπτει ότι η καταναλωτική συμπεριφορά στο συγκεκριμένο κατάστημα κυριαρχείται από την αγορά **ειδών πρώτης ανάγκης** και προϊόντων **καθημερινής/φρέσκιας διατροφής** (γαλακτοκομικά, λαχανικά, αρτοσκευάσματα), με μοναδική εξαίρεση τα αναψυκτικά (soda).

-Ερώτημα: Να βρείτε ποια είναι τα προϊόντα που αγοράζονται σπάνια.

1. Κώδικας στην R

Για τον εντοπισμό των προϊόντων με τη χαμηλότερη συχνότητα εμφάνισης, υπολογίστηκε η απόλυτη συχνότητα όλων των ειδών και στη συνέχεια ταξινομήθηκαν σε αύξουσα σειρά με τον ακόλουθο κώδικα:

 # Υπολογισμός της απόλυτης συχνότητας για κάθε προϊόν

item_counts <- itemFrequency(Groceries, type = "absolute")

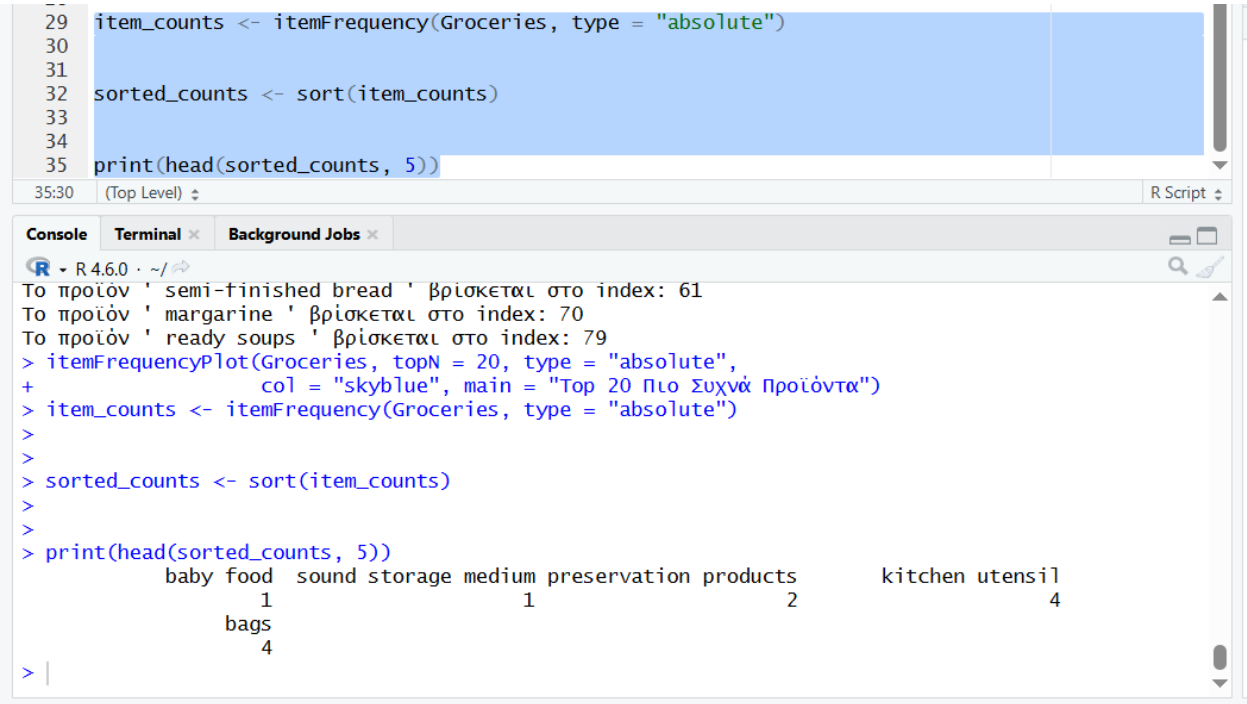
 # Ταξινόμηση από το λιγότερο συχνό στο περισσότερο συχνό

sorted_counts <- sort(item_counts)

 # Εμφάνιση των 5 πιο σπάνιων προϊόντων

print(head(sorted_counts, 5))

2. Αποτέλεσμα



3. Επεξήγηση & Συμπεράσματα

Από την ανάλυση των αποτελεσμάτων της κονσόλας προκύπτει ότι οι καταναλωτές αγοράζουν εξαιρετικά σπάνια ορισμένες κατηγορίες προϊόντων. Τα πέντε (5) πιο σπάνια είδη σε ολόκληρο το dataset (σε σύνολο 9.835 συναλλαγών) είναι:

1. **baby food** (βρεφικές τροφές) – Εμφανίζεται μόλις **1 φορά**.
2. **sound storage medium** (μέσα αποθήκευσης ήχου / CD-DVD) – Εμφανίζεται μόλις **1 φορά**.
3. **preservation products** (προϊόντα συντήρησης) – Εμφανίζεται **2 φορές**.
4. **kitchen utensil** (κουζινικά σκεύη) – Εμφανίζεται **4 φορές**.
5. **bags** (τσάντες/σακούλες) – Εμφανίζεται **4 φορές**.

Συμπέρασμα Ανάλυσης:

Αυτά τα προϊόντα παρουσιάζουν σχεδόν μηδενική ζήτηση στο δείγμα μας. Αυτό σημαίνει ότι η σχετική τους συχνότητα (π.χ. για το baby food είναι 1 / 9835 ~ 0,0001$ ή $0,01%) είναι πολύ χαμηλότερη από το ελάχιστο όριο υποστήριξης (Support = 0.006 ή 0,6%) που ορίζει η άσκηση. Κατά συνέπεια, ο αλγόριθμος Apriori θα απορρίψει αμέσως αυτά τα στοιχεία στο πρώτο κιόλας βήμα του (Pruning phase), καθώς δεν θεωρούνται "συχνά σύνολα στοιχείων" (frequent itemsets).

-Ερώτημα: Εφαρμογή του αλγορίθμου Apriori με support level 0.006, ελάχιστο length 2 και ελάχιστο confidence 0.25

1. Κώδικας στην R

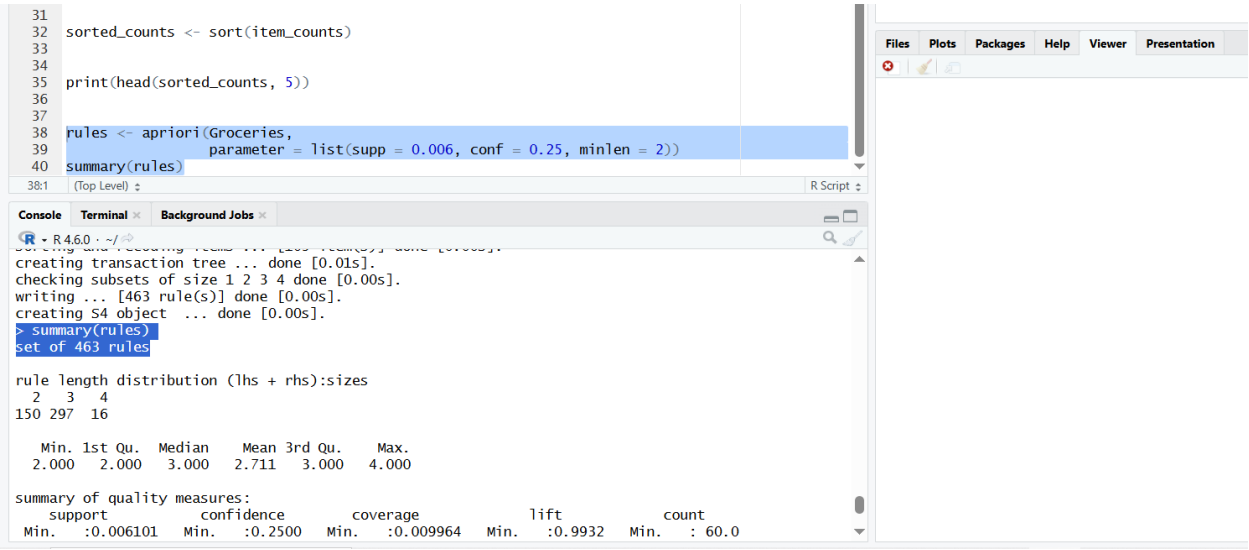
Για τη δημιουργία των κανόνων συσχέτισης, εφαρμόστηκε ο αλγόριθμος Apriori θέτοντας ως ελάχιστη υποστήριξη (LHS & RHS μαζί) το 0.006 (0.6%), ως ελάχιστη εμπιστοσύνη το 0.25 (25%) και ως ελάχιστο μήκος κανόνα τα 2 στοιχεία (ώστε να αποφευχθούν οι κενοί κανόνες):

```
rules <- apriori(Groceries,
                 parameter = list(supp = 0.006, conf = 0.25, minlen = 2))
summary(rules)
```

2. Αποτέλεσμα στην Κονσόλα

Από τα στοιχεία που επιστρέφει το summary, καταγράφουμε το βασικό αποτέλεσμα:

set of 463 rules



3. Επεξήγηση & Συμπεράσματα

Με βάση τα κριτήρια που τέθηκαν, ο αλγόριθμος Apriori παρήγαγε συνολικά **463 έγκυρους κανόνες συσχέτισης**.

Συμπέρασμα Ανάλυσης:

- **Support (0.006):** Σημαίνει ότι για να θεωρηθεί ένας συνδυασμός προϊόντων "συχνός", θα πρέπει να εμφανίζεται σε τουλάχιστον 59 συναλλαγές (9835 X 0.006 = 59.01). Αυτό το σχετικά χαμηλό όριο επιτρέπει στον αλγόριθμο να εντοπίσει αρκετά μοτίβα, χωρίς όμως να κατακλυστεί από τυχαίους συνδυασμούς.
- **Confidence (0.25):** Εξασφαλίζει ότι στους κανόνες που παράγονται, η πιθανότητα να αγοραστεί το προϊόν στο δεξί μέρος (RHS), δεδομένου ότι αγοράστηκε το προϊόν στο αριστερό μέρος (LHS), είναι τουλάχιστον 25%.
- **Minlen (2):** Το φιλτράρισμα αυτό απέκλεισε κανόνες με άδειο αριστερό μέρος (LHS), διασφαλίζοντας ότι κάθε κανόνας εκφράζει μια πραγματική σχέση εξάρτησης μεταξύ τουλάχιστον δύο προϊόντων.

Το πλήθος των 463 κανόνων κρίνεται ικανοποιητικό, καθώς παρέχει ένα πλούσιο σύνολο πληροφοριών για τη μελέτη της καταναλωτικής συμπεριφοράς, τις οποίες θα ταξινομήσουμε και θα φιλτράρουμε στα επόμενα ερωτήματα.

-Ερώτημα: Εντοπισμός των top 5 συνόλων στοιχείων (κανόνων) που έχουν το υψηλότερο επίπεδο confidence.

1. Κώδικας στην R

Για να απομονώσουμε τους 5 κανόνες με τη μεγαλύτερη τιμή εμπιστοσύνης (confidence), ταξινομήσαμε το σύνολο των κανόνων σε φθίνουσα σειρά ως προς τη συγκεκριμένη μετρική και χρησιμοποιήσαμε τη συνάρτηση inspect():

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

inspect(head(rules_sorted_by_conf, 5))

2. Αποτέλεσμα

```
30
31
32  sorted_counts <- sort(item_counts)
33
34
35  print(head(sorted_counts, 5))
36
37
38  rules <- apriori(Groceries,
39                   parameter = list(supp = 0.006, conf = 0.25, minlen = 2))
40  summary(rules)
41
42
43  rules_sorted_by_conf <- sort(rules, by = "confidence", decreasing = TRUE)
44  inspect(head(rules_sorted_by_conf, 5))
```

43:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.6.0 · ~/

```
                                                                    call
 apriori(data = Groceries, parameter = list(supp = 0.006, conf = 0.25, minlen = 2))
> rules_sorted_by_conf <- sort(rules, by = "confidence", decreasing = TRUE)
> inspect(head(rules_sorted_by_conf, 5))
    lhs                               rhs          support     confidence
[1] {butter, whipped/sour cream} => {whole milk} 0.006710727 0.6600000
[2] {butter, yogurt}             => {whole milk} 0.009354347 0.6388889
[3] {root vegetables, butter}    => {whole milk} 0.008235892 0.6377953
[4] {tropical fruit, curd}       => {whole milk} 0.006507372 0.6336634
[5] {tropical fruit, butter}     => {whole milk} 0.006202339 0.6224490
    coverage    lift     count
[1] 0.010167768 2.583008 66
[2] 0.014641586 2.500387 92
[3] 0.012913066 2.496107 81
[4] 0.010269446 2.479936 64
[5] 0.009964413 2.436047 61
>
```

3. Επεξήγηση & Συμπεράσματα

Οι 5 κορυφαίοι κανόνες συσχέτισης με βάση το επίπεδο Confidence (Εμπιστοσύνη) παρουσιάζονται αναλυτικά παρακάτω:

1. **{butter, whipped/sour cream} => {whole milk}**
    - Confidence: **66,00%** | Support: 0,67% | Lift: 2,58
    - Ερμηνεία: Το 66% των πελατών που αγοράζουν βούτυρο και σαντιγί/ξινή κρέμα μαζί, θα αγοράσουν και πλήρες γάλα.
2. **{butter, yogurt} => {whole milk}**
    - Confidence: **63,89%** | Support: 0,94% | Lift: 2,50
    - Ερμηνεία: Το 63,89% όσων αγοράζουν βούτυρο και γιαούρτι, επιλέγουν επίσης πλήρες γάλα.
3. **{root vegetables, butter} => {whole milk}**
    - Confidence: **63,78%** | Support: 0,82% | Lift: 2,50
    - Ερμηνεία: Το 63,78% των καταναλωτών που αγοράζουν βολβούς/λαχανικά ρίζας και βούτυρο, αγοράζουν και πλήρες γάλα.
4. **{tropical fruit, curd} =>{whole milk}**
    - Confidence: **63,37%** | Support: 0,65% | Lift: 2,48
    - Ερμηνεία: Όταν αγοράζονται τροπικά φρούτα και πηγμένο γάλα (curd), υπάρχει 63,37% πιθανότητα να αγοραστεί και πλήρες γάλα.
5. **{tropical fruit, butter} => {whole milk}**
    - Confidence: **62,24%** | Support: 0,62% | Lift: 2,44
    - Ερμηνεία: Το 62,24% όσων αγοράζουν τροπικά φρούτα και βούτυρο, αγοράζουν ταυτόχρονα πλήρες γάλα.

Συμπέρασμα Ανάλυσης:

Το πιο αξιοσημείωτο εύρημα είναι ότι και οι πέντε κορυφαίοι κανόνες έχουν ως δεξί μέρος (rhs) το **whole milk**. Αυτό εξηγείται από το γεγονός ότι το πλήρες γάλα είναι το πιο συχνό προϊόν στο κατάστημα (όπως είδαμε στο πρώτο μας γράφημα), οπότε έχει υψηλή εκ των προτέρων πιθανότητα εμφάνισης.

Επιπλέον, βλέπουμε ότι το **butter (βούτυρο)** εμφανίζεται στο αριστερό μέρος (lhs) σε 4 από τους 5 κανόνες. Αυτό δείχνει μια ισχυρότατη συνδυαστική τάση: οι καταναλωτές που αγοράζουν βούτυρο μαζί με κάποιο άλλο γαλακτοκομικό ή φρέσκο προϊόν (όπως γιαούρτι, κρέμα,

φρούτα/λαχανικά), σχεδόν πάντα ολοκληρώνουν τις αγορές τους προσθέτοντας και γάλα στο καλάθι τους. Οι τιμές του δείκτη **Lift είναι όλες πάνω από 2,4** , πράγμα που σημαίνει ότι αυτοί οι συνδυασμοί αυξάνουν την πιθανότητα αγοράς γάλακτος κατά τουλάχιστον 2,4 φορές σε σχέση με μια τυχαία αγορά.

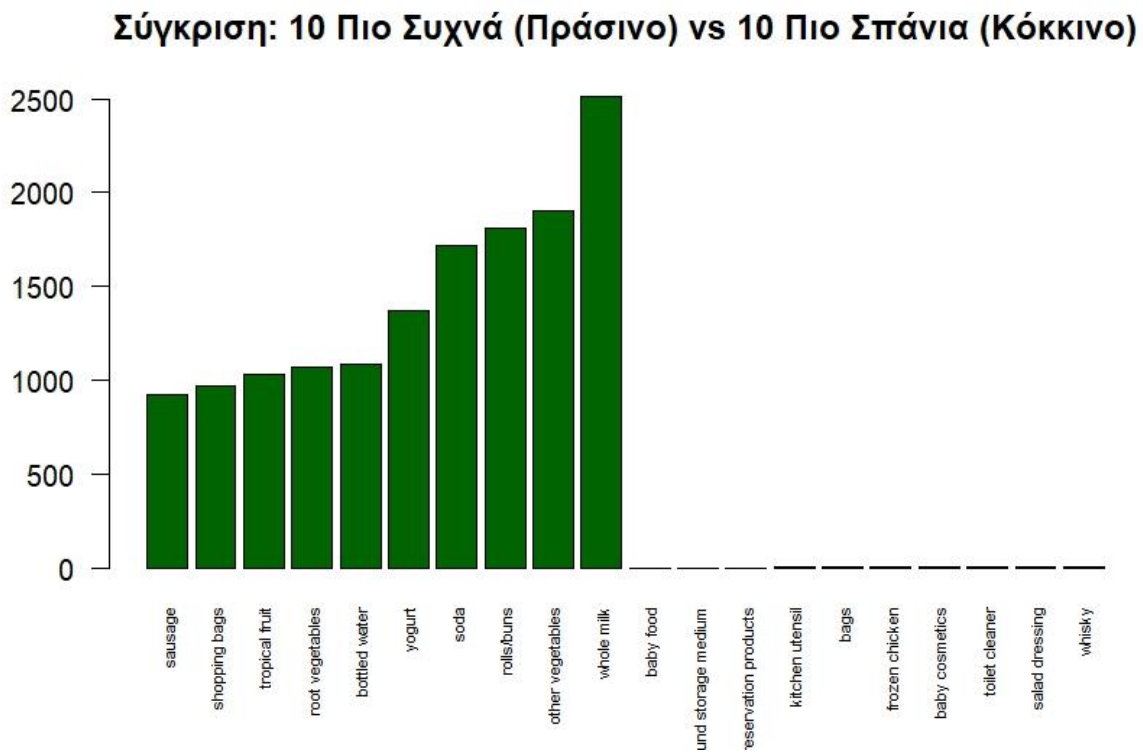
-Ερώτημα: Δημιουργία γραφικής παράστασης που απεικονίζει τη σχέση μεταξύ των προϊόντων που αγοράζονται σπάνια και αυτών που αγοράζονται συχνά.

1. Κώδικας στην R

Για τη γραφική αποτύπωση της σχέσης μεταξύ συχνών και σπάνιων προϊόντων, απομονώθηκαν τα 10 δημοφιλέστερα και τα 10 λιγότερο δημοφιλή προϊόντα του dataset και σχεδιάστηκαν σε ένα κοινό ραβδόγραμμα με τον ακόλουθο κώδικα:

```
item_counts <- itemFrequency(Groceries, type = "absolute")

sorted_counts <- sort(item_counts)


comparison_data <- c(tail(sorted_counts, 10), head(sorted_counts, 10))


barplot(comparison_data, las = 2, col = c(rep("darkgreen", 10), rep("darkred", 10)),

    main = "Σύγκριση: 10 Πιο Συχνά (Πράσινο) vs 10 Πιο Σπάνια (Κόκκινο)", cex.names = 0.55)
```

2. Αποτέλεσμα



3. Επεξήγηση & Συμπεράσματα

Το συγκεκριμένο διάγραμμα παρουσιάζει μια έντονη οπτική αντίθεση, τοποθετώντας δίπλα-δίπλα τα δύο άκρα της καταναλωτικής συμπεριφοράς:

- **Η «Κεφαλή» των Δεδομένων (Πράσινες Μπάρες - Αριστερά):** Περιλαμβάνει τα 10 πιο συχνά αγοραζόμενα προϊόντα (π.χ. *whole milk, other vegetables, rolls/buns* κλπ.). Οι συχνότητές τους είναι εξαιρετικά υψηλές, καθώς το πλήρες γάλα αγγίζει τις 2.513 εμφανίσεις και το δέκατο προϊόν (root vegetables) τις 1.072 εμφανίσεις.
- **Η «Μακρά Ουρά» (Κόκκινες Μπάρες - Δεξιά):** Περιλαμβάνει τα 10 πιο σπάνια προϊόντα (π.χ. *baby food, sound storage medium, preservation products* κλπ.). Οι μπάρες αυτές είναι σχεδόν μη ορατές στο γράφημα, καθώς οι απόλυτες συχνότητές τους κυμαίνονται από μόλις 1 έως 4 εμφανίσεις σε σύνολο 9.835 συναλλαγών.

Συμπέρασμα Ανάλυσης: Η γραφική παράσταση αποδεικνύει την ύπαρξη του φαινομένου της **ασύμμετρης κατανομής (Long Tail Distribution)** στα δεδομένα του σούπερ μάρκετ. Η συντριπτική πλειοψηφία των αγορών συγκεντρώνεται γύρω από έναν πολύ μικρό και σταθερό πυρήνα προϊόντων καθημερινής ανάγκης (πράσινο τμήμα). Αντίθετα, υπάρχει ένας μεγάλος αριθμός εξειδικευμένων προϊόντων (κόκκινο τμήμα) που, αν και είναι διαθέσιμα στο κατάστημα, σπάνια μπαίνουν στο καλάθι κάποιου πελάτη.

Αυτή η τεράστια απόσταση δικαιολογεί απόλυτα τη χρησιμότητα του φιλτραρίσματος Support στον αλγόριθμο Apriori: θέτοντας το ελάχιστο Support στο 0.006, ο αλγόριθμος "κόβει" αυτόματα όλη την κόκκινη ουρά των σπάνιων προϊόντων, επιτρέποντας η ανάλυση να επικεντρωθεί αποκλειστικά στα προϊόντα εκείνα που έχουν εμπορική αξία και δυναμική για τη δημιουργία στρατηγικών marketing (π.χ. τοποθέτηση προϊόντων στα ράφια, προσφορές).

-Ερώτημα: Ποιο είναι το μοτίβο των αγορών του καταναλωτή;

Απάντηση & Συμπεράσματα

Από τη συνολική ανάλυση του dataset Groceries (τόσο από τη διερευνητική ανάλυση συχνοτήτων όσο και από την εφαρμογή του αλγορίθμου Apriori), χαρτογραφείται με σαφήνεια το αγοραστικό μοτίβο του καταναλωτή. Τα βασικά χαρακτηριστικά αυτού του μοτίβου είναι τα εξής:

1. **Κυριαρχία των Ειδών Πρώτης Ανάγκης (Core Grocery Pattern):** Ο καταναλωτής επισκέπτεται το κατάστημα πρωτίστως για να προμηθευτεί βασικά αγαθά καθημερινής διατροφής και άμεσης κατανάλωσης. Το πλήρες γάλα (*whole milk*), τα λαχανικά (*other vegetables*), τα ψωμάκια (*rolls/buns*) και το γιαούρτι (*yogurt*) αποτελούν τη βάση σχεδόν κάθε μεγάλου καλαθιού.

2. **Συνδυαστική Αγορά Συμπληρωματικών Προϊόντων (Cross-Selling / Companion Pattern):** Οι καταναλωτές σπάνια αγοράζουν ένα προϊόν απομονωμένα. Υπάρχει ένα έντονο μοτίβο «συνταγής» ή συμπληρωματικότητας. Για παράδειγμα, όπως είδαμε στους κορυφαίους κανόνες εμπιστοσύνης (Confidence), η αγορά υλικών μαγειρικής ή πρωινού (όπως το *butter*) λειτουργεί ως «μαγνήτης» που οδηγεί με μαθηματική βεβαιότητα (άνω του 62%) στην ταυτόχρονη αγορά φρέσκου γάλακτος.

3. **Μοτίβο "Long Tail" (Μακρά Ουρά):** Οι αγορές χαρακτηρίζονται από ακραία συγκέντρωση. Το μοτίβο δείχνει ότι το super market βασίζεται σε περίπου 15-20 δημοφιλή προϊόντα για τον μεγάλο όγκο των καθημερινών του συναλλαγών, ενώ ένας τεράστιος αριθμός προϊόντων (όπως *baby food, kitchen utensils, preservation products*) αγοράζονται εντελώς σποραδικά και δεν συμμετέχουν στα σταθερά μοτίβα των πελατών.

Σύνοψη Μοτίβου: Το αγοραστικό προφίλ του καταναλωτή στο συγκεκριμένο dataset είναι αυτό του **«οικογενειακού/καθημερινού αγοραστή»**. Οι κανόνες δείχνουν ότι οι καταναλωτές σχεδιάζουν τα καλάθια τους γύρω από γεύματα (π.χ. λαχανικά ρίζας, φρούτα, γαλακτοκομικά) και κάθε φορά που προσθέτουν στο καλάθι τους δύο ή περισσότερα premium/φρέσκα προϊόντα (π.χ. τροπικά φρούτα και βούτυρο), η τάση τους να ολοκληρώσουν τη συναλλαγή με ένα βασικό αγαθό (πλήρες γάλα) είναι ισχυρότατη.

-Ερώτημα: Ποιοι είναι οι συνδυασμοί (κανόνες) που οδηγούν στην αγορά γάλακτος;

1. Κώδικας στην R

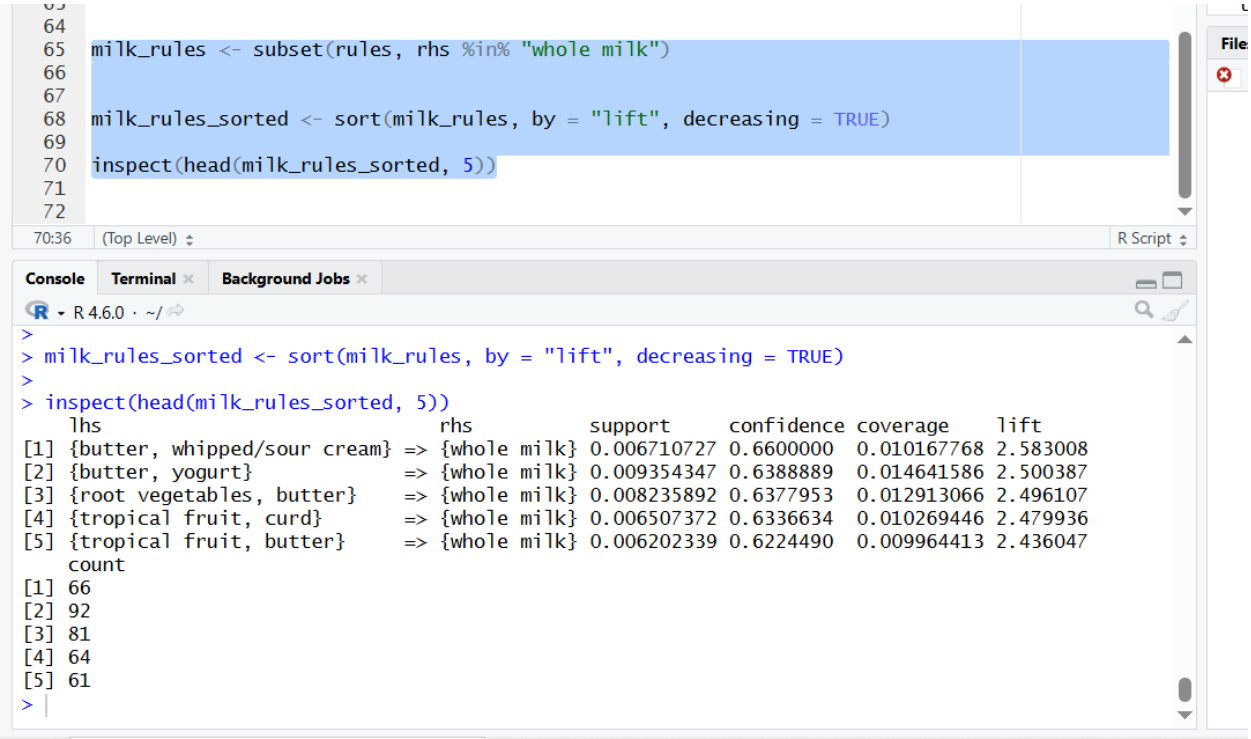
Για να απομονώσουμε αποκλειστικά τους κανόνες που οδηγούν στην αγορά πλήρους γάλακτος, χρησιμοποιήθηκε η συνάρτηση subset() ώστε το whole milk να βρίσκεται στη δεξιά πλευρά (RHS) του κανόνα. Στη συνέχεια, οι κανόνες ταξινομήθηκαν σε φθίνουσα σειρά με βάση τον δείκτη ανασηκώματος (Lift):

milk_rules <- subset(rules, rhs %in% "whole milk")

milk_rules_sorted <- sort(milk_rules, by = "lift", decreasing = TRUE)

inspect(head(milk_rules_sorted, 5))

2. Αποτέλεσμα



3. Επεξήγηση & Συμπεράσματα

Οι 5 ισχυρότεροι συνδυασμοί προϊόντων που ωθούν τον καταναλωτή να αγοράσει πλήρες γάλα (ταξινομημένοι βάσει του δείκτη Lift) είναι:

1. **{butter, whipped/sour cream} =>{whole milk}** (Lift: 2,58)
2. **{butter, yogurt} => {whole milk}** (Lift: 2,50)
3. **{root vegetables, butter} => {whole milk}** (Lift: 2,50)
4. **{tropical fruit, curd} => {whole milk}** (Lift: 2,48)
5. **{tropical fruit, butter} => {whole milk}** (Lift: 2,44)

Συμπέρασμα Ανάλυσης:

- **Ερμηνεία του Lift:** Ο δείκτης Lift και στους 5 κανόνες είναι εξαιρετικά υψηλός (πάνω από 2,4). Αυτό σημαίνει ότι η αγορά αυτών των συνδυασμών (π.χ. βούτυρο μαζί με ξινή κρέμα ή γιαούρτι) αυξάνει την πιθανότητα ο πελάτης να αγοράσει και πλήρες γάλα κατά **2,4 έως 2,58 φορές** σε σχέση με έναν πελάτη που θα αγόραζε τυχαία.

- **Ο ρόλος του Βουτύρου (Butter):** Είναι προφανές ότι το **βούτυρο** αποτελεί τον βασικότερο κοινό παρονομαστή, αφού εμφανίζεται σε 4 από τους 5 κορυφαίους κανόνες. Όταν το βούτυρο συνδυάζεται με άλλα γαλακτοκομικά (whipped/sour cream, yogurt) ή με φρέσκα προϊόντα (root vegetables, tropical fruit), δημιουργεί ένα πολύ ισχυρό αγοραστικό προφίλ που σχετίζεται με τη μαγειρική, τη ζαχαροπλαστική ή την προετοιμασία γευμάτων, οδηγώντας τον καταναλωτή στο να ολοκληρώσει τις αγορές του παίρνοντας και γάλα.

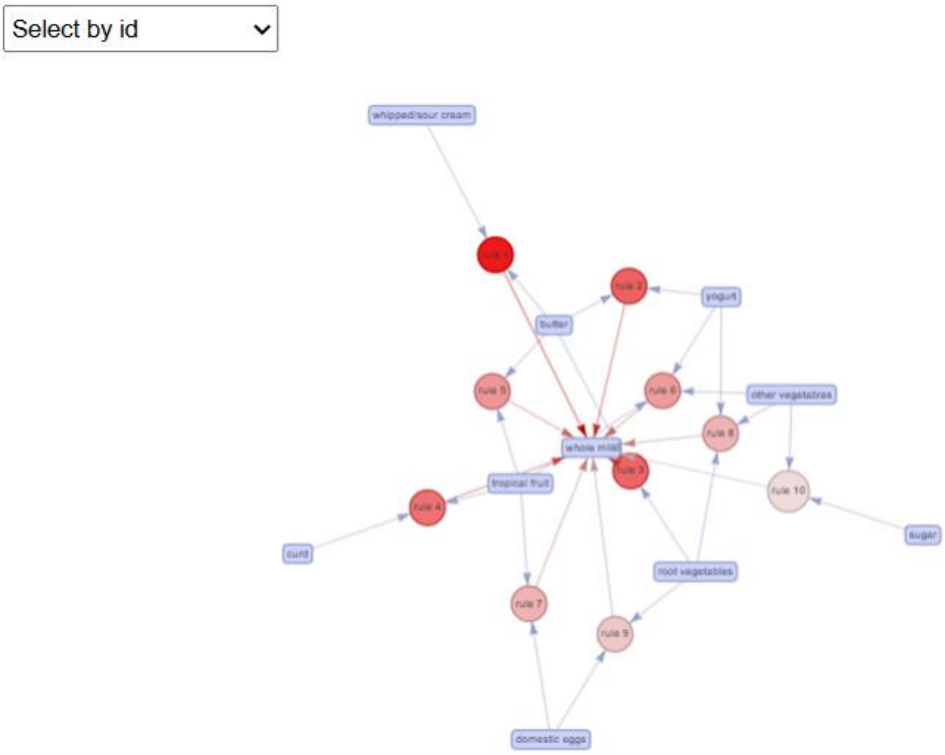
-Ερώτημα: Να δημιουργήσετε τη γραφική απεικόνιση αυτού του συνδυασμού.

1. Κώδικας στην R

Για τη γραφική οπτικοποίηση των 10 ισχυρότερων κανόνων που οδηγούν στην αγορά πλήρους γάλακτος, χρησιμοποιήθηκε η μέθοδος δικτύου (method = "graph") με τη διαδραστική μηχανή σχεδίασης htmlwidget:

plot(head(milk_rules_sorted, 10), method = "graph", engine = "htmlwidget")

2. Αποτέλεσμα



3. Επεξήγηση & Συμπεράσματα

Η γραφική παράσταση αποτυπώνει τις δομικές σχέσεις μεταξύ των προϊόντων σε μορφή **Δικτύου (Network Graph)**. Από την ανάλυση του διαγράμματος προκύπτουν τα εξής βασικά συμπεράσματα:

- **Κεντρικός Στόχος (Whole Milk):** Το whole milk βρίσκεται στην «καρδιά» του δικτύου. Όλα τα βέλη από τους κύκλους των κανόνων (rules) δείχνουν κατευθείαν προς αυτό, επιβεβαιώνοντας οπτικά ότι αποτελεί το τελικό προϊόν αγοράς (Consequent / RHS) για όλους τους κορυφαίους κανόνες.
- **Κόμβοι Προϊόντων (Items):** Τα γαλάζια ορθογώνια αντιπροσωπεύουν τα προϊόντα που ωθούν στην αγορά γάλακτος (LHS). Παρατηρούμε πώς το butter (βούτυρο), τα other vegetables (λαχανικά), το yogurt (γιαούρτι) και τα tropical fruit (τροπικά φρούτα) συνδέονται με πολλαπλά βέλη, δείχνοντας ότι συμμετέχουν σε παραπάνω από έναν ισχυρούς κανόνες.
- **Κύκλοι Κανόνων (Rules):** Οι κυκλικοί κόμβοι (π.χ. rule 1, rule 2, κλπ.) αναπαριστούν τους ίδιους τους κανόνες συσχέτισης:

- **Η ένταση του κόκκινου χρώματος** υποδηλώνει το **Lift**. Ο κύκλος rule 1 (που αντιστοιχεί στον κανόνα {butter, whipped/sour cream} => {whole milk}) έχει το πιο σκούρο και έντονο κόκκινο χρώμα, γεγονός που μαρτυρά ότι διαθέτει τη μεγαλύτερη τιμή ανασηκώματος (Lift = 2,58) και άρα την ισχυρότερη στατιστική εξάρτηση.
- **Το μέγεθος του κύκλου** είναι ανάλογο του **Support**. Κανόνες που περιλαμβάνουν πολύ δημοφιλή προϊόντα, όπως τα λαχανικά (other vegetables), εμφανίζονται ως ελαφρώς μεγαλύτεροι κύκλοι, καθώς έχουν μεγαλύτερη συχνότητα εμφάνισης στο σύνολο των 9.835 συναλλαγών.

Τελικό Επιχειρηματικό Συμπέρασμα: Το δίκτυο αυτό αποτελεί έναν πολύτιμο οδηγό για τη διοίκηση του σούπερ μάρκετ. Αποδεικνύει ότι η αγορά γάλακτος δεν είναι τυχαία, αλλά καθοδηγείται από συγκεκριμένες καταναλωτικές ανάγκες (όπως η προετοιμασία φαγητού/πρωινού). Με βάση αυτό το γράφημα, το κατάστημα μπορεί να εφαρμόσει επιτυχημένες στρατηγικές **Cross-Merchandising**, τοποθετώντας τα προϊόντα που βρίσκονται στην περιφέρεια του δικτύου (π.χ. βούτυρο, κρέμες, γιαούρτια) σε στρατηγικά σημεία που θα διευκολύνουν τον πελάτη να κατευθυνθεί και προς τα ψυγεία με το γάλα, αυξάνοντας έτσι την κερδοφορία ανά επίσκεψη.

Γενική Ανακεφαλαίωση Θέματος 2ου

Κλείνοντας την ανάλυση του Θέματος 2ου, τα συνολικά ευρήματα από το dataset **Groceries** επιβεβαιώνουν ότι η R και το πακέτο *arules* αποτελούν ισχυρά εργαλεία Business Intelligence. Κατάφεραν να μετατρέψουν 9.835 χαοτικές καθημερινές συναλλαγές σε δομημένη γνώση, αποδεικνύοντας ότι:

- **Το Πλήρες Γάλα** λειτουργεί ως ο απόλυτος πυρήνας του καταστήματος, γύρω από τον οποίο οι καταναλωτές χτίζουν το καλάθι τους.
- **Η στρατηγική Support** είναι απαραίτητη για να κόβεται η τεράστια Long Tail ουρά των σπάνιων προϊόντων (όπως βρεφικές τροφές ή κουζινικά), επιτρέποντας στην επιχείρηση να εστιάζει σε συνδυασμούς με πραγματική εμπορική αξία.
- **Οι κανόνες συσχέτισης** (με κυρίαρχο το βούτυρο που εκτοξεύει την αγορά γάλακτος έως και 2,58 φορές) προσφέρουν έτοιμες λύσεις για στοχευμένες προσφορές (Bundling) και αύξηση της κερδοφορίας.