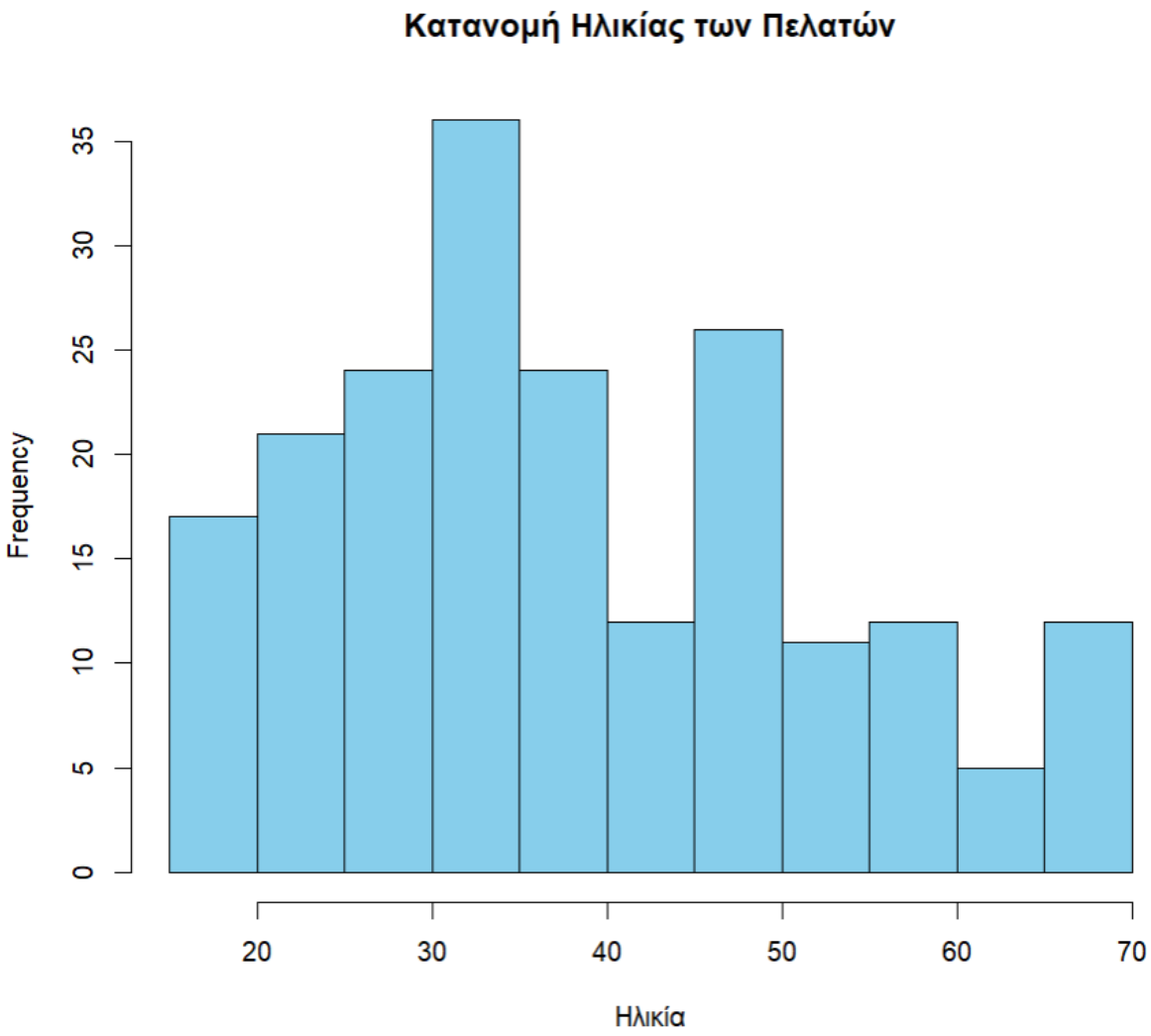
Περιγραφή Project: Μέσα από την πλατφόρμα Kaggle έχω κατεβάσει ένα αρχείο μέσα από το οποίο θα αναλύσουμε δεδομένα πελατών και θα τους κατηγοριοποιήσω σε clusters με βάση τα αγοραστικά τους μοτίβα.

Έχω **200 παρατηρήσεις (γραμμές) και 5 μεταβλητές (στήλες)**.

Οι στήλες είναι:

- CustomerID: Μοναδικός αριθμός πελάτη.
- Genre: Φύλο (Male, Female).
- Age: Ηλικία (18-70).
- Annual.Income..k..: Ετήσιο εισόδημα σε χιλιάδες δολάρια (15-137).
- Spending.Score..1.100.: Δείκτης δαπανών (1-99).

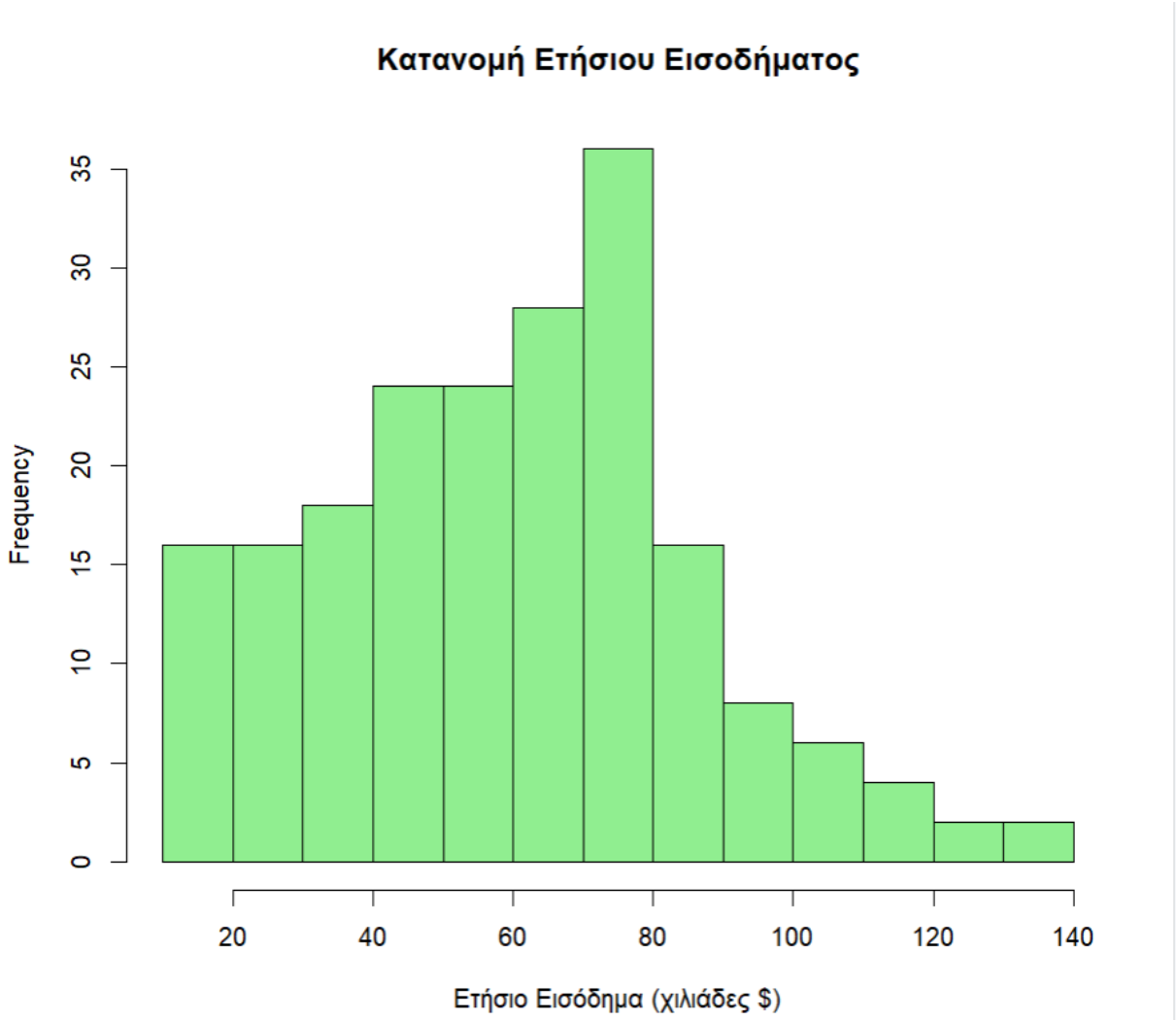


1. Κατανομή Ηλικίας των Πελατών

Τι βλέπουμε:

- Οι περισσότεροι πελάτες είναι μεταξύ **30-50 ετών**.
- Υπάρχουν λίγοι πελάτες κάτω των 25 και άνω των 60.
- Ο πληθυσμός είναι αρκετά **ισοκατανεμημένος**, χωρίς έντονες ανισορροπίες.

Συμπέρασμα:
Η πλειοψηφία των πελατών είναι μεσαίας ηλικίας.

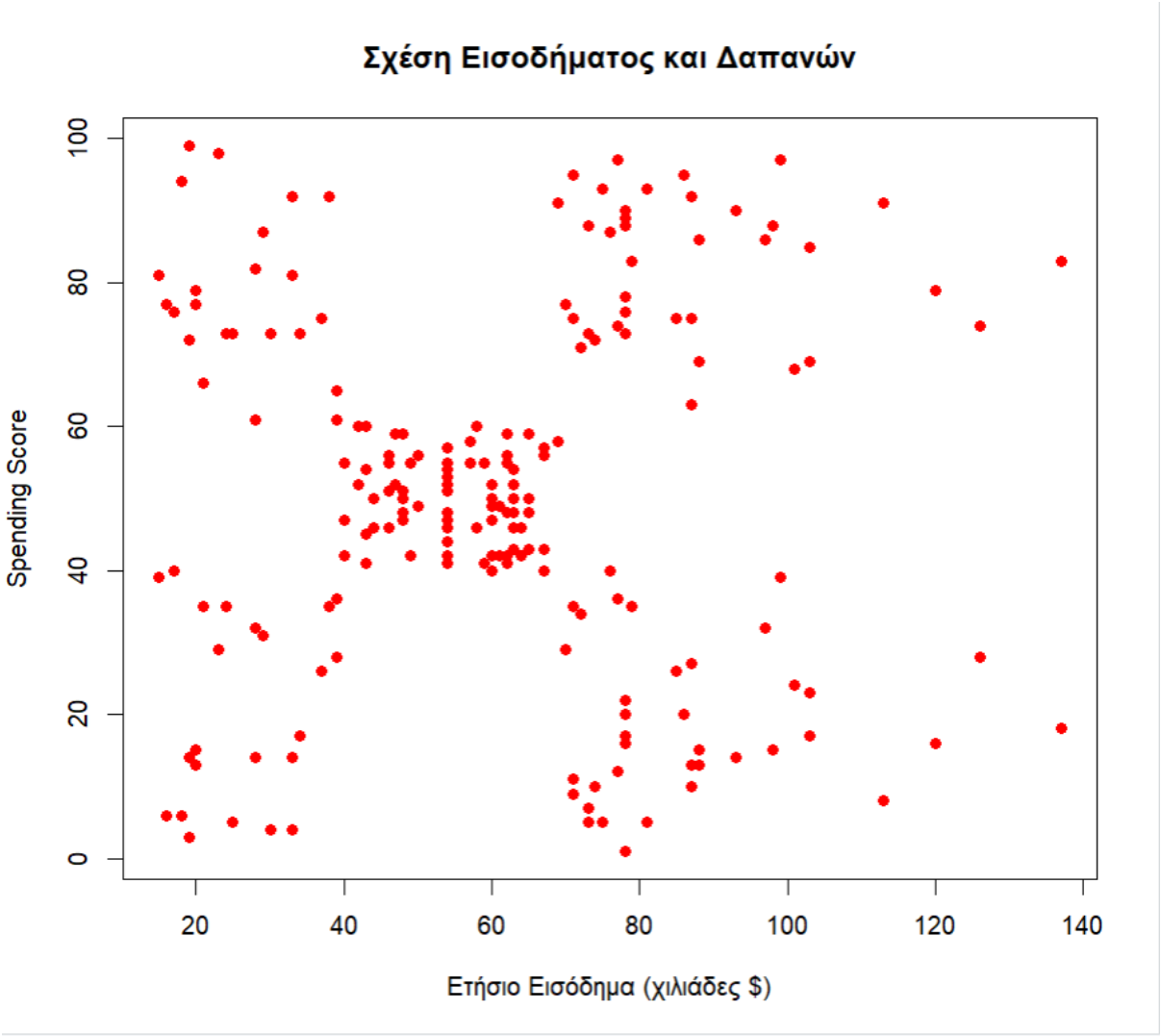


2. Κατανομή Ετήσιου Εισοδήματος

Τι βλέπουμε:

- Η κατανομή είναι **ασύμμετρη** προς τα δεξιά (δεξιά ουρά).
- Οι περισσότεροι πελάτες έχουν **εισόδημα μεταξύ 40k-80k**.
- Υπάρχουν λίγοι πελάτες με εισόδημα πάνω από 100k.

Συμπέρασμα:
Το εισόδημα των πελατών είναι κυρίως μεσαίο.

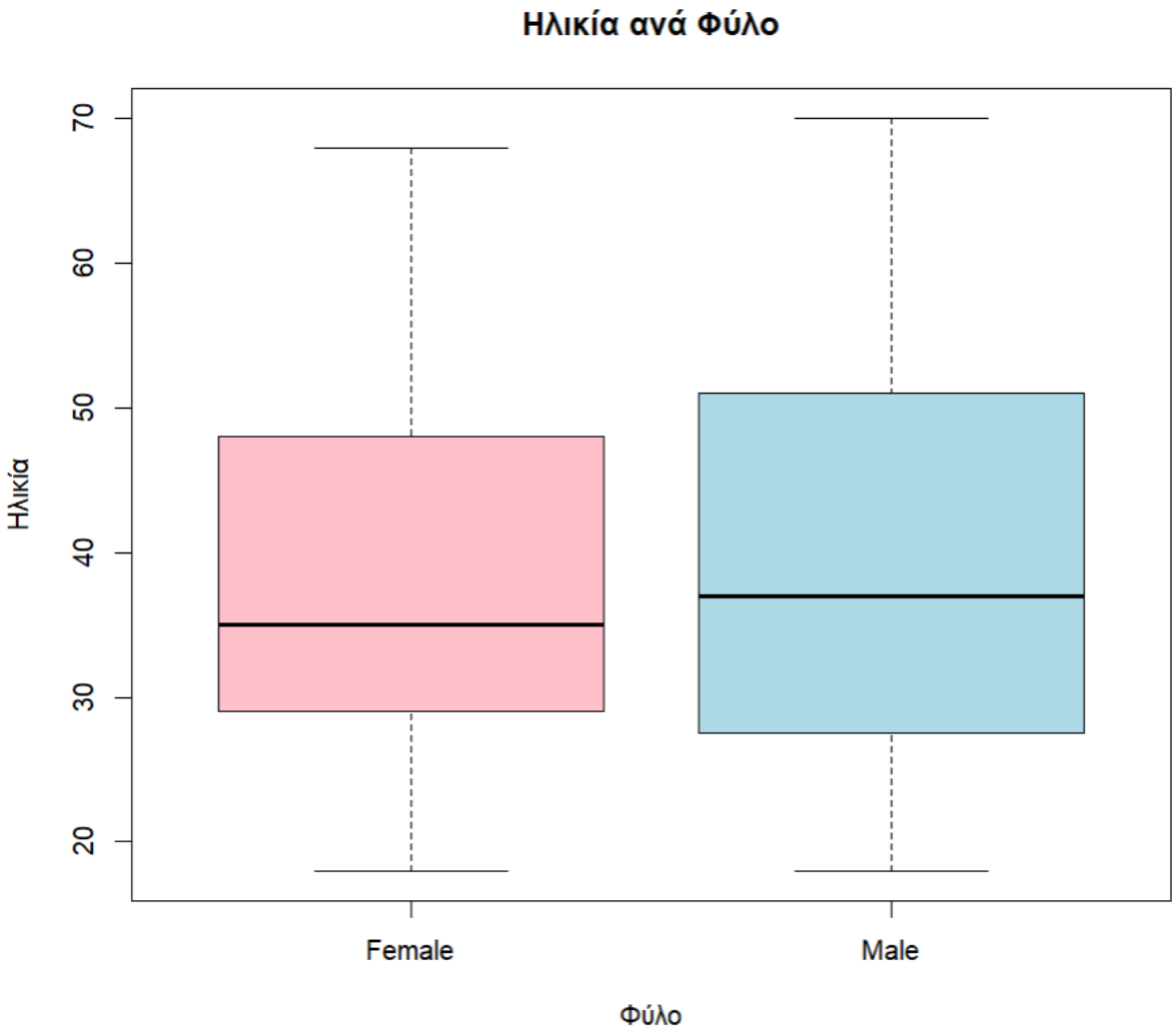


3. Σχέση Εισοδήματος και Δείκτη Δαπανών

Τι βλέπουμε:

- Υπάρχουν **δύο ξεχωριστές ομάδες** πελατών:
  - Μερικοί με **χαμηλό εισόδημα** αλλά **υψηλό spending score**.
  - Μερικοί με **υψηλό εισόδημα** αλλά **χαμηλό spending score**.
- Οι πελάτες με **μεσαίο εισόδημα (40k-80k)** έχουν **μεγαλύτερη διασπορά** στις δαπάνες.

Συμπέρασμα:
Υπάρχουν διαφορετικές καταναλωτικές συμπεριφορές. Μερικοί πελάτες ξοδεύουν πολύ ανεξαρτήτως εισοδήματος, ενώ άλλοι είναι πιο συγκρατημένοι. Αυτό μπορεί να βοηθήσει στην τμηματοποίηση (clustering) των πελατών.



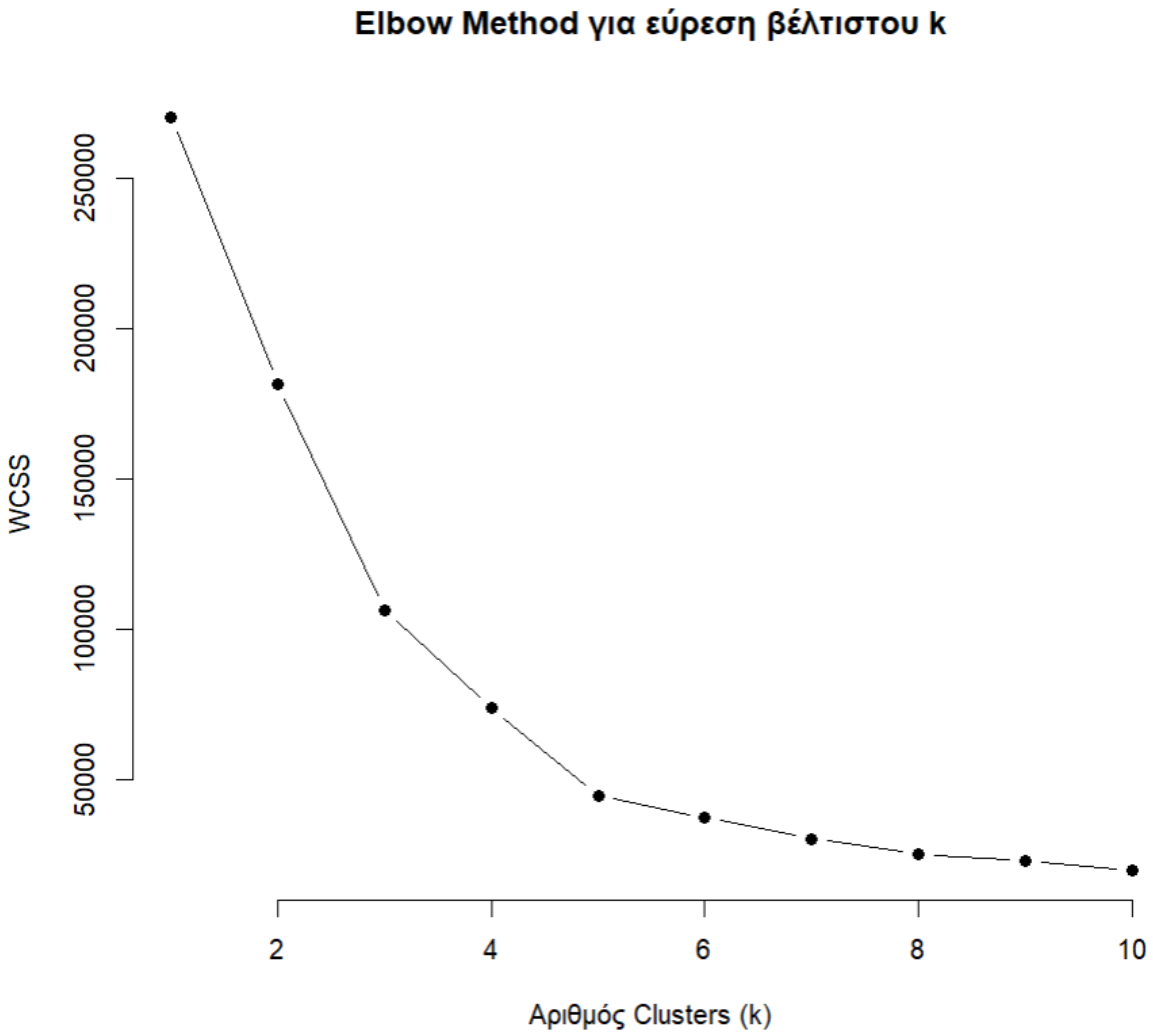
4. Boxplot για Ηλικία ανά Φύλο

Τι βλέπουμε:

- Οι ηλικίες των **γυναικών και ανδρών είναι παρόμοιες**.
- Η **μέση ηλικία** είναι γύρω στα **35-40 έτη** και για τα δύο φύλα.
- Υπάρχουν μερικές **μεγαλύτερες ηλικίες** (πιθανοί outliers), αλλά όχι πολλές.

Συμπέρασμα:
Δεν υπάρχει μεγάλη διαφορά στις ηλικίες ανδρών-γυναικών, οπότε οι καταναλωτικές συνήθειες ίσως να επηρεάζονται περισσότερο από άλλους παράγοντες (εισόδημα, spending score).

k-means clustering: Θα χωρίσουμε σε clusters (τμηματοποίηση πελατών), για να ομαδοποιήσω τους πελάτες με βάση το Ετήσιο Εισόδημα και το Spending Score.



Τι μας δείχνει το Elbow Plot:

-Στον οριζόντιο άξονα (x-axis) έχουμε τον αριθμό των clusters (k).

-Στον κάθετο άξονα (y-axis) έχουμε το WCSS (Within-Cluster Sum of Squares), που δείχνει πόσο καλά σχηματίζονται τα clusters.

- Το **βέλτιστο k** είναι το σημείο που η καμπύλη **"λυγίζει" (elbow)**.
- Πριν από αυτό το σημείο, το WCSS **μειώνεται γρήγορα**.

- Μετά από αυτό, το WCSS **δεν μειώνεται σημαντικά**, άρα δεν υπάρχει λόγος να προσθέσουμε περισσότερα clusters.

Άρα, για το Mall Customers Dataset, το **k = 5** είναι η καλύτερη επιλογή!

```
> kmeans_result
K-means clustering with 5 clusters of sizes 81, 23, 22, 35, 39

Cluster means:
  Annual.Income..k.. Spending.Score..1.100.
1           55.29630                49.51852
2           26.30435                20.91304
3           25.72727                79.36364
4           88.20000                17.11429
5           86.53846                82.12821

Clustering vector:
  [1] 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2
 [43] 2 1 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [85] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 5 4 5 4 5
[127] 1 5 4 5 4 5 1 5 4 5 4 5 4 5 4 5 1 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4
[169] 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4

Within cluster sum of squares by cluster:
[1]  9875.111  5098.696  3519.455 12511.143 13444.051
 (between_SS / total_SS =  83.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
> mall_data$Cluster <- as.factor(kmeans_result$cluster)
> mall_data$Cluster
  [1] 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2
 [43] 2 1 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [85] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 5 4 5 4 5
[127] 1 5 4 5 4 5 1 5 4 5 4 5 4 5 4 5 1 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4
[169] 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4
Levels: 1 2 3 4 5
>
```

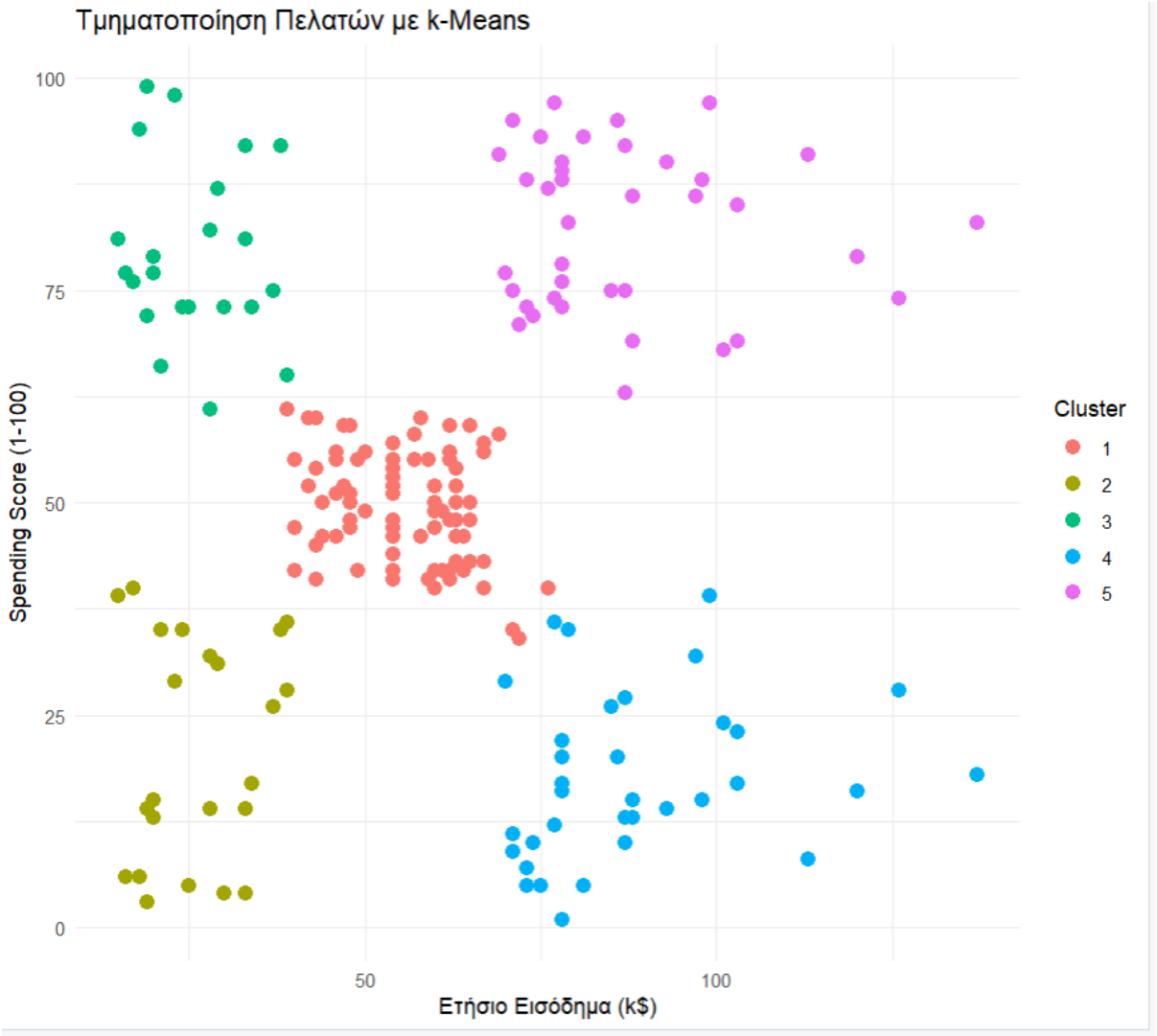
Ανάλυση αποτελεσμάτων

1. **Cluster sizes:**
    - Ο αλγόριθμος δημιούργησε **5 clusters** με αντίστοιχα μεγέθη:
        - **Cluster 1**: 81 πελάτες
        - **Cluster 2**: 23 πελάτες
        - **Cluster 3**: 22 πελάτες
        - **Cluster 4**: 35 πελάτες
        - **Cluster 5**: 39 πελάτες

2. **Cluster Means (Μέσες τιμές κάθε cluster):**

- Δείχνει τον μέσο όρο του **Ετήσιου Εισοδήματος** και του **Spending Score** για κάθε cluster.
- Παρατηρούμε ότι:
    - **Cluster 1** έχει **μεσαίο εισόδημα (55.3k) και μέσο Spending Score (49.5)**.
    - **Cluster 2** έχει **χαμηλό εισόδημα (26.3k) και χαμηλό Spending Score (20.9)**.
    - **Cluster 3** έχει **χαμηλό εισόδημα (25.7k) αλλά πολύ υψηλό Spending Score (79.3)**.
    - **Cluster 4** έχει **υψηλό εισόδημα (88.2k) αλλά χαμηλό Spending Score (17.1)**.
    - **Cluster 5** έχει **υψηλό εισόδημα (86.5k) και υψηλό Spending Score (82.1)**.

Παρατήρηση: Το συνολικό **(between_SS / total_SS = 83.5%)** σημαίνει ότι το **83.5% της διακύμανσης εξηγείται από τα clusters**, που είναι καλό ποσοστό!



Ανάλυση του Διαγράμματος

Τι δείχνει το γράφημα:

- Στον **οριζόντιο άξονα (x-axis)**: **Ετήσιο Εισόδημα (Annual Income in k$)**.
- Στον **κατακόρυφο άξονα (y-axis)**: **Spending Score (1-100)**, δηλαδή πόσο ξοδεύουν οι πελάτες σε σχέση με το εισόδημά τους.
- Κάθε **σημείο** είναι ένας πελάτης.
- Το **χρώμα** αντιπροσωπεύει το cluster στο οποίο ανήκει ο πελάτης.

Ανάλυση των Clusters

1. **Κόκκινο Cluster (1)** → Πελάτες με **μεσαίο εισόδημα** (~40-60k) και **μεσαίο Spending Score**.
2. **Πράσινο Cluster (2)** → Πελάτες με **χαμηλό εισόδημα (~0-40k) και χαμηλό Spending Score (~0-40)**.

3. **Τυρκουάζ Cluster (3)** → Πελάτες με **χαμηλό εισόδημα (~0-40k) αλλά υψηλό Spending Score (~60-100)**.
4. **Μπλε Cluster (4)** → Πελάτες με **υψηλό εισόδημα (~60-100k) αλλά χαμηλό Spending Score (~0-40)**.
5. **Μωβ Cluster (5)** → Πελάτες με **υψηλό εισόδημα (~60-100k) και υψηλό Spending Score (~60-100)**.

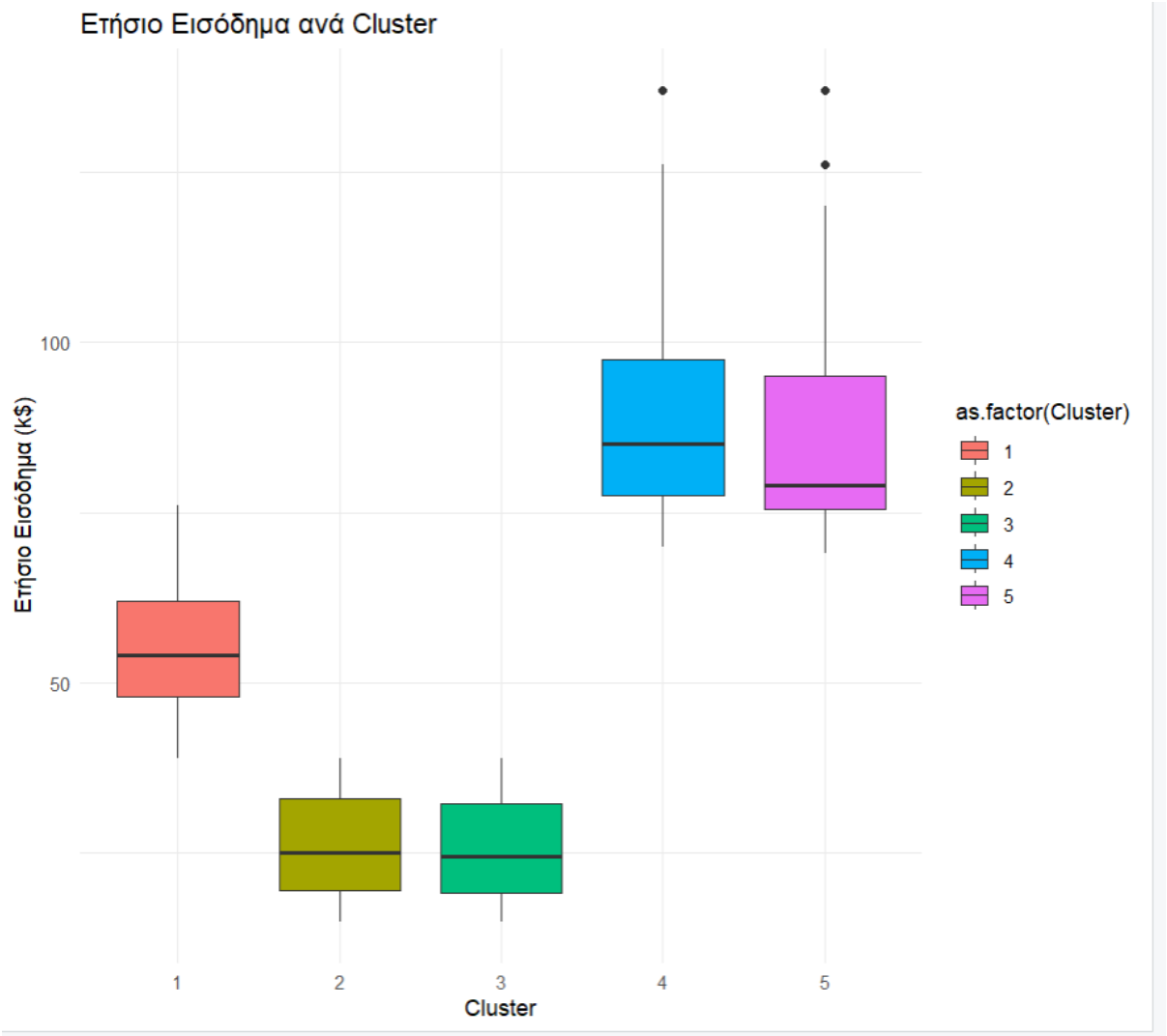
Συμπεράσματα

- Οι πελάτες που **ξοδεύουν πολύ (υψηλό Spending Score)** ανήκουν είτε στο **Cluster 3** (χαμηλό εισόδημα, υψηλή δαπάνη) είτε στο **Cluster 5** (υψηλό εισόδημα, υψηλή δαπάνη).
- Οι πελάτες που **δεν ξοδεύουν πολύ** ανήκουν στα **Cluster 2 και 4**, ανάλογα με το εισόδημά τους.
- Το **Cluster 1** φαίνεται να είναι πιο «ουδέτερο», δηλαδή πελάτες με μέσο εισόδημα και μέσο Spending Score.

	Cluster	Annual.Income..k...mean	Annual.Income..k...median	Annual.Income..k...sd
1	1	55.296296	54.000000	8.988109
2	2	26.304348	25.000000	7.893811
3	3	25.727273	24.500000	7.566731
4	4	88.200000	85.000000	16.399067
5	5	86.538462	79.000000	16.312485

	Spending.Score..1.100..mean	Spending.Score..1.100..median	Spending.Score..1.100..sd
1	49.518519	50.000000	6.530909
2	20.913043	17.000000	13.017167
3	79.363636	77.000000	10.504174
4	17.114286	16.000000	9.952154
5	82.128205	83.000000	9.364489

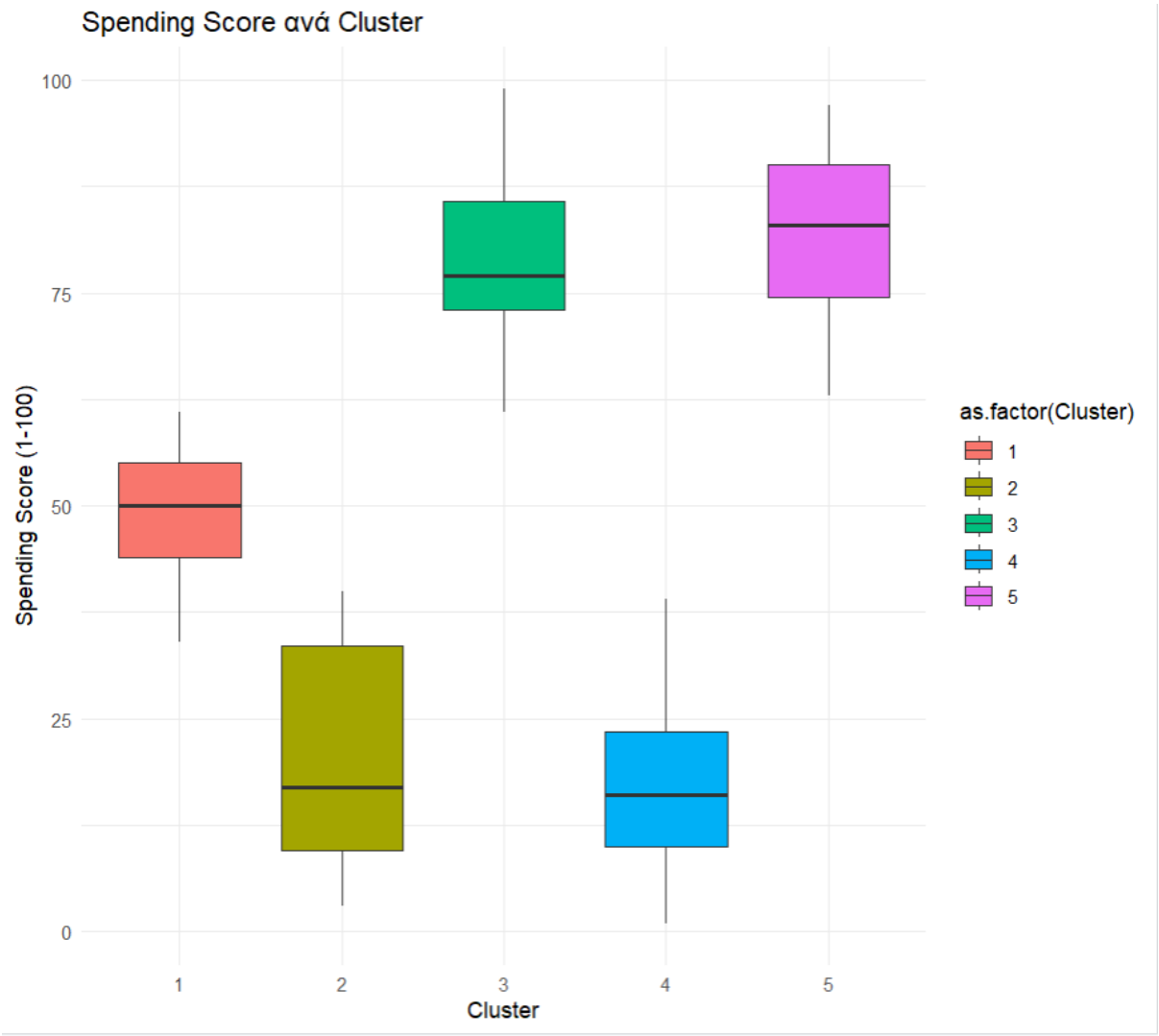
-Υπολογισμός των μεγεθών μέση τιμή, διάμεσο, τυπική απόκλιση για κάθε μεταβλητή, για να δούμε ποιες ομάδες έχουν υψηλή ή καταναλωτική συμπεριφορά.



Η 1η γραφική παράσταση είναι ένα boxplot που αναλύει τα χαρακτηριστικά των clusters ως προς δύο μεταβλητές :

Ετήσιο Εισόδημα ανά Cluster

- Στον πρώτο γράφημα, κάθε cluster απεικονίζεται ως ένα boxplot που δείχνει την κατανομή του ετήσιου εισοδήματος.
- Βλέπουμε ότι:
    - Το Cluster 4 (μπλε) και το Cluster 5 (μωβ) έχουν τα υψηλότερα εισοδήματα.
    - Τα Clusters 2 (λαδί) και 3 (πράσινο) έχουν τα χαμηλότερα εισοδήματα.
    - Το Cluster 1 (κόκκινο) έχει μεσαία εισοδήματα με μικρότερη διασπορά.
    - Υπάρχουν κάποια outliers, ειδικά στα Clusters 4 και 5.



Η 2η γραφική παράσταση είναι ένα boxplot που τα clusters συγκρίνονται ως προς το Spending Score(βαθμός καταναλωτικών δαπανών).

Το Cluster 3 (πράσινο) και το Cluster 5 (μωβ) έχουν τα υψηλότερα Spending Scores, δηλαδή αυτοί οι πελάτες ξοδεύουν περισσότερο.

Το Cluster 4 (μπλε) έχει χαμηλό Spending Score παρά το υψηλό εισόδημα, κάτι που δείχνει ότι αυτοί οι πελάτες είναι πιθανώς πιο συντηρητικοί στις δαπάνες τους.

Το Cluster 2 (λαδί) έχει χαμηλό εισόδημα και χαμηλό Spending Score, άρα πρόκειται για πελάτες με μικρότερη αγοραστική δύναμη.

Το Cluster 1 (κόκκινο) έχει μεσαίες τιμές και στις δύο μεταβλητές.

-Machine Learning: Θα κάνουμε προβλέψεις με Machine Learning, όπου θα χρησιμοποιήσουμε τα δεδομένα που έχουμε ήδη τμηματοποιήσει (clusters) ως labels και εκπαιδεύουμε ένα μοντέλο (Random Forest) που μπορεί να προβλέψει σε ποιο cluster ανήκει ένας νέος πελάτης. Οπότε θα φτιάξω ένα μοντέλο ταξινόμησης (classification model) που θα προβλέπει το Cluster στο οποίο ανήκει ένας πελάτης, με βάση το ετήσιο εισόδημα και το Spending Score. Με την συνάρτηση createDataPartition() χωρίζω τα δεδομένα όπου 80% (training set) για εκπαίδευση και 20% για δοκιμή (testing set).

Τώρα που έχουμε διαχωρίσει τα δεδομένα σε training (train_data) και testing (test_data), θα εκπαιδεύσουμε ένα μοντέλο Machine Learning για να προβλέπει το Cluster.

```
> library(randomForest)
randomForest 4.7-1.2
Type rfNews() to see new features/changes/bug fixes.

Attaching package: 'randomForest'

The following object is masked from 'package:dplyr':

    combine

The following object is masked from 'package:ggplot2':

    margin

Warning message:
package 'randomForest' was built under R version 4.4.3
> library(caret)
> set.seed(123) #για αναπαραγωγιμότητα
> model_rf <- randomForest(Cluster ~ ., data = train_data, ntree = 100)
> model_rf

Call:
 randomForest(formula = Cluster ~ ., data = train_data, ntree = 100)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 1

        OOB estimate of  error rate: 1.85%
Confusion matrix:
   1  2  3  4  5 class.error
1 63  1  0  1  0  0.03076923
2  0 19  0  0  0  0.00000000
3  0  0 18  0  0  0.00000000
4  0  0  0 28  0  0.00000000
5  1  0  0  0 31  0.03125000
>
```

- Εκπαιδεύει ένα **Random Forest Classifier** χρησιμοποιώντας το training set (train_data).
- Το μοντέλο προσπαθεί να προβλέψει το Cluster με βάση τα χαρακτηριστικά Annual.Income..k.. και Spending.Score..1.100.
- Χρησιμοποιεί **100 δέντρα απόφασης** (ntree = 100).

Αποτελέσματα του Μοντέλου:

- **OOB error rate**: 1.85%, που σημαίνει ότι το μοντέλο κάνει πολύ λίγα λάθη.
- **Confusion Matrix**: Δείχνει ότι το μοντέλο ταξινομεί σωστά τις περισσότερες κατηγορίες, με ελάχιστα σφάλματα.

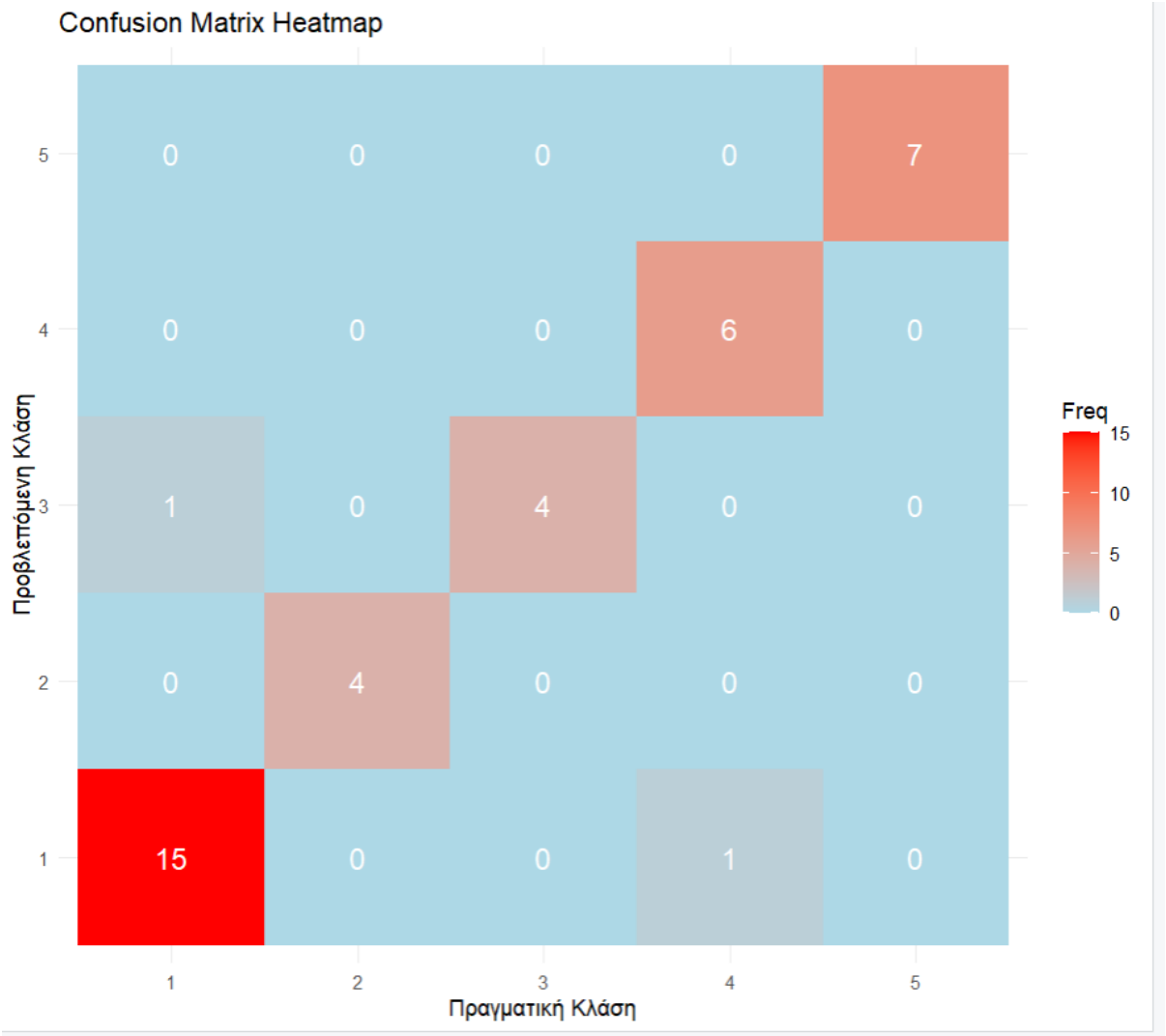
```
> predictions
  1   2  15  26  34  39  40  41  44  47  48  49  67  73  75  80  85  92  93 106 108 109
  2   3   2   3   3   2   3   2   3   1   1   1   1   1   1   1   1   1   1   1   1   1
113 115 125 126 134 145 150 165 168 171 173 178 183 188 191 194
  1   1   1   5   5   4   5   4   5   4   4   5   4   5   4   5
Levels: 1 2 3 4 5
> conf_matrix <- confusionMatrix(predictions, test_data$Cluster)
> conf_matrix
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5
         1 15  0  0  1  0
         2  0  4  0  0  0
         3  1  0  4  0  0
         4  0  0  0  6  0
         5  0  0  0  0  7

Overall Statistics

               Accuracy : 0.9474
                 95% CI : (0.8225, 0.9936)
    No Information Rate : 0.4211
    P-Value [Acc > NIR] : 7.335e-12

                  Kappa : 0.9284

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.9375   1.0000   1.0000   0.8571   1.0000
Specificity            0.9545   1.0000   0.9706   1.0000   1.0000
Pos Pred Value         0.9375   1.0000   0.8000   1.0000   1.0000
Neg Pred Value         0.9545   1.0000   1.0000   0.9688   1.0000
Prevalence             0.4211   0.1053   0.1053   0.1842   0.1842
Detection Rate         0.3947   0.1053   0.1053   0.1579   0.1842
Detection Prevalence   0.4211   0.1053   0.1316   0.1579   0.1842
Balanced Accuracy      0.9460   1.0000   0.9853   0.9286   1.0000
>
```

- **Accuracy**: **94.74%** → Το μοντέλο έχει πολύ υψηλή ακρίβεια!
- **Kappa = 0.9284** → Υποδεικνύει πολύ καλή συμφωνία μεταξύ των πραγματικών και προβλεπόμενων τιμών.
- **Ειδικά Αποτελέσματα ανά Κλάση (Class 1-5):**
  - **Sensitivity** (True Positive Rate): Πάνω από **85% σε όλες τις κατηγορίες**.
  - **Specificity** (True Negative Rate): Κοντά στο **100% στις περισσότερες κατηγορίες**.
  - **Balanced Accuracy**: >92% για όλες τις κλάσεις.

Συμπέρασμα

- Το μοντέλο **Random Forest** έχει πολύ καλή απόδοση με **υψηλή ακρίβεια (94.74%)**.
- Το **OOB error rate** είναι μόλις **1.85%**, που δείχνει ότι το μοντέλο δεν έχει overfitting.
- Ο **πίνακας σύγχυσης (confusion matrix)** δείχνει ότι σχεδόν όλες οι προβλέψεις γίνονται σωστά.

```
> conf_matrix_table
          Reference
Prediction  1  2  3  4  5
         1 15  0  0  1  0
         2  0  4  0  0  0
         3  1  0  4  0  0
         4  0  0  0  6  0
         5  0  0  0  0  7
> df_cm <- as.data.frame(conf_matrix_table)
> df_cm
   Prediction Reference Freq
1           1         1   15
2           2         1    0
3           3         1    1
4           4         1    0
5           5         1    0
6           1         2    0
7           2         2    4
8           3         2    0
9           4         2    0
10          5         2    0
11          1         3    0
12          2         3    0
13          3         3    4
14          4         3    0
15          5         3    0
16          1         4    1
17          2         4    0
18          3         4    0
19          4         4    6
20          5         4    0
21          1         5    0
22          2         5    0
23          3         5    0
24          4         5    0
25          5         5    7
```



Τι δείχνει το γράφημα:

- Οι **γραμμές** αντιπροσωπεύουν τις **προβλεπόμενες κλάσεις** (τι είπε το μοντέλο).
- Οι **στήλες** αντιπροσωπεύουν τις **πραγματικές κλάσεις** (η αλήθεια).
- Τα **χρώματα** δείχνουν τη συχνότητα (πόσες προβλέψεις έγιναν σωστά ή λάθος):
    - **Κόκκινο** = Υψηλή τιμή (πολλές προβλέψεις).
    - **Μπλε** = Χαμηλή τιμή (λίγες ή καθόλου προβλέψεις).

Τέλεια προβλέψεις (διαγώνιος)

- Η διαγώνιος δείχνει τις σωστές προβλέψεις.
- **(1,1) -> 15** σωστές προβλέψεις για την κλάση 1.

- **(2,2) -> 4** σωστές για την κλάση 2.
- **(3,3) -> 4**, **(4,4) -> 6**, **(5,5) -> 7** επίσης σωστές προβλέψεις.

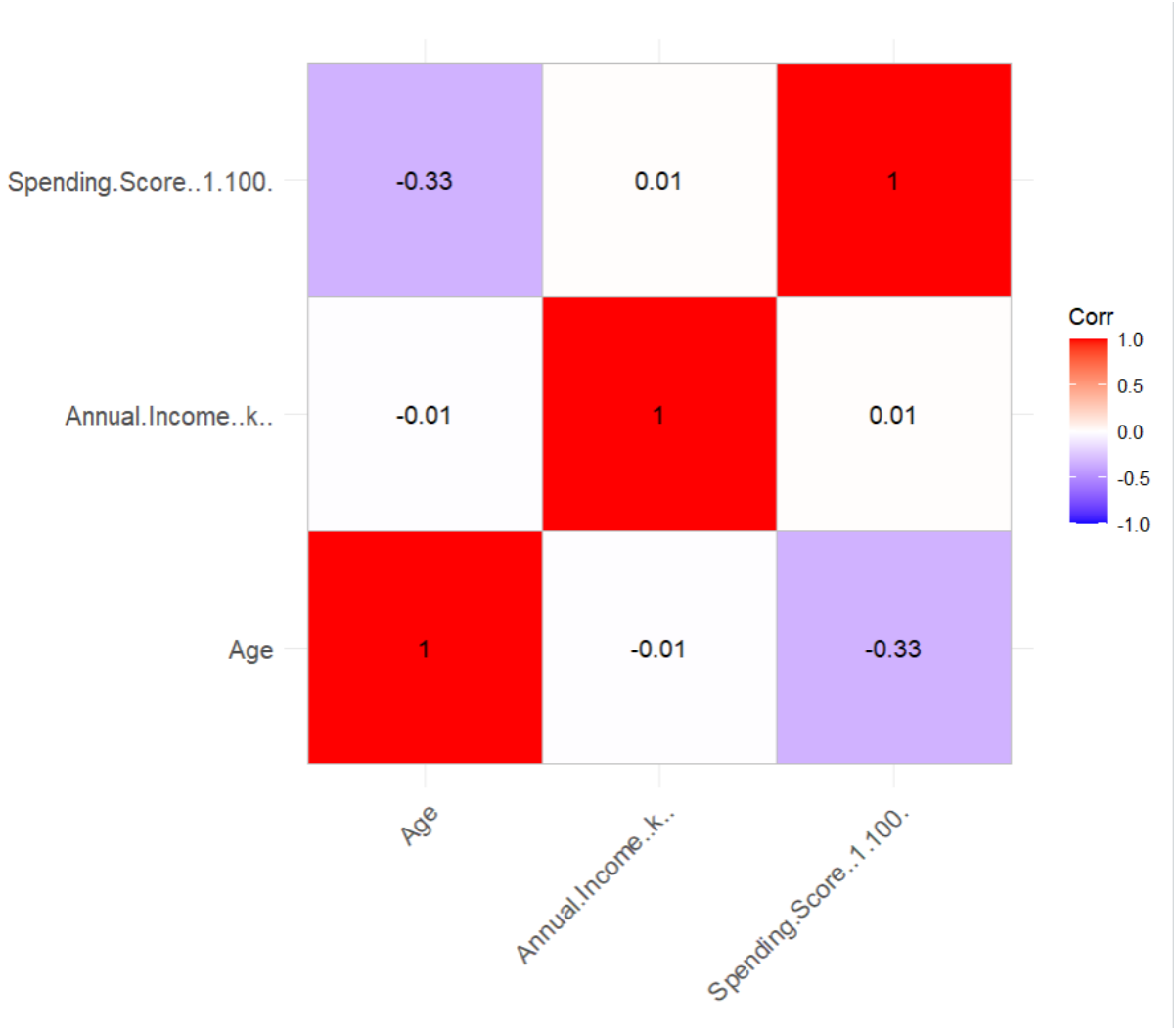
Το μοντέλο είναι αρκετά ακριβές, γιατί οι περισσότερες τιμές είναι στη διαγώνιο.

```
> cor_matrix
                            Age Annual.Income..k.. Spending.Score..1.100.
Age                  1.00000000       -0.012398043           -0.327226846
Annual.Income..k..  -0.01239804        1.000000000            0.009902848
Spending.Score..1.100. -0.32722685    0.009902848            1.000000000
> library(ggcorrplot)
Warning message:
package 'ggcorrplot' was built under R version 4.4.3
> ggcorrplot(cor_matrix, lab = TRUE, colors = c("blue", "white", "red"))
>
```

Υπολογίσαμε την συσχέτιση μεταξύ των βασικών αριθμητικών μεταβλητών (Age, Annual.Income..k.., Spending.Score..1.100.)

Τι μας λέει ο πίνακας:

1. **Ηλικία & Ετήσιο Εισόδημα (-0.012)**
    - Πολύ μικρή αρνητική συσχέτιση (~0), άρα η ηλικία **δεν σχετίζεται** με το εισόδημα.
2. **Ηλικία & Δείκτης Κατανάλωσης (-0.327)**
    - Ελαφρώς **αρνητική συσχέτιση**, δηλαδή όσο μεγαλώνει η ηλικία, οι άνθρωποι **ξοδεύουν λιγότερο**.
3. **Ετήσιο Εισόδημα & Δείκτης Κατανάλωσης (0.009)**
    - Σχεδόν **μηδενική συσχέτιση**, που σημαίνει ότι **το εισόδημα δεν επηρεάζει το πόσο ξοδεύουν οι πελάτες**.



Το γράφημα απεικονίζει τις **συσχετίσεις** μεταξύ των μεταβλητών **Age (Ηλικία)**, **Annual Income (Ετήσιο Εισόδημα)** και **Spending Score (Δείκτης Κατανάλωσης)**.

Τι βλέπουμε στο heatmap:

Μεταβλητές	**Age**	**Annual Income**	**Spending Score**
Age	**1.00**	-0.01	**-0.33**
Annual Income	-0.01	**1.00**	**0.01**
Spending Score	-0.33	0.01	**1.00**

Ερμηνεία των χρωμάτων:

- **Κόκκινο (Θετική συσχέτιση, κοντά στο 1)** → Οι μεταβλητές αυξάνονται μαζί.
- **Μπλε (Αρνητική συσχέτιση, κοντά στο -1)** → Όταν αυξάνεται η μία μεταβλητή, η άλλη μειώνεται.

- **Λευκό (Συσχέτιση κοντά στο 0)** → Δεν υπάρχει ιδιαίτερη σχέση μεταξύ των μεταβλητών.

Συμπεράσματα

1.**Ηλικία & Δείκτης Κατανάλωσης (-0.33, Μπλε-Μωβ)**

- Όσο αυξάνεται η ηλικία, τόσο **μειώνεται η καταναλωτική συμπεριφορά**.
- Νεότεροι πελάτες ξοδεύουν περισσότερο.

2.**Ηλικία & Ετήσιο Εισόδημα (-0.01, Σχεδόν Λευκό)**

- Δεν υπάρχει σχέση μεταξύ ηλικίας και εισοδήματος.

3.**Ετήσιο Εισόδημα & Δείκτης Κατανάλωσης (0.01, Λευκό)**

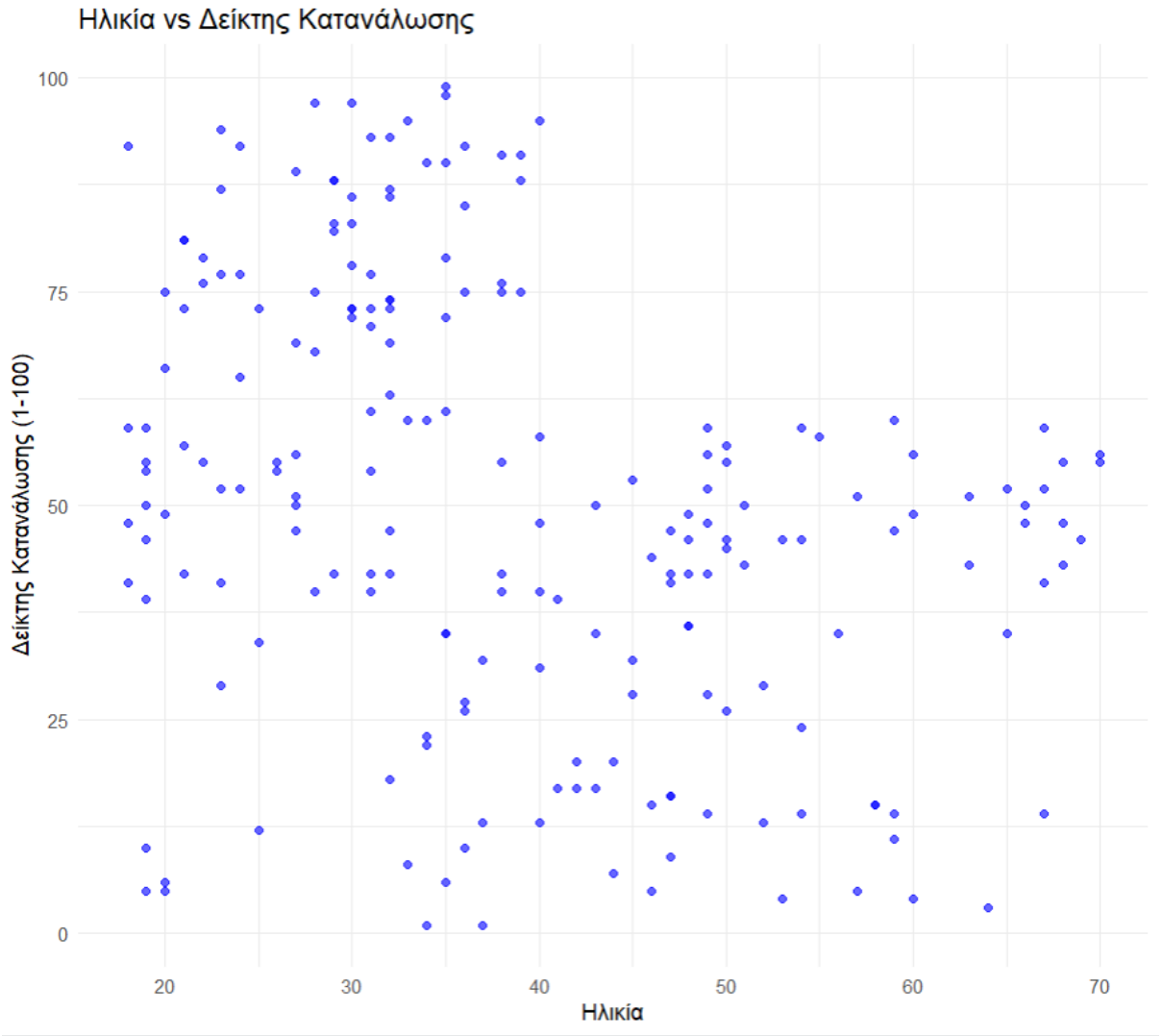
- Το εισόδημα **δεν επηρεάζει άμεσα** το πόσο ξοδεύουν οι πελάτες.

Συνολικά: Η **ηλικία επηρεάζει** το πόσο ξοδεύει κάποιος, αλλά το εισόδημα **όχι ιδιαίτερα**.

-Θα φτιάξουμε 2 scatterplot :

Scatter plot1:

Ηλικία vs Δείκτης Κατανάλωσης (για να δούμε αν οι μεγαλύτεροι καταναλωτές είναι πιο νέοι)

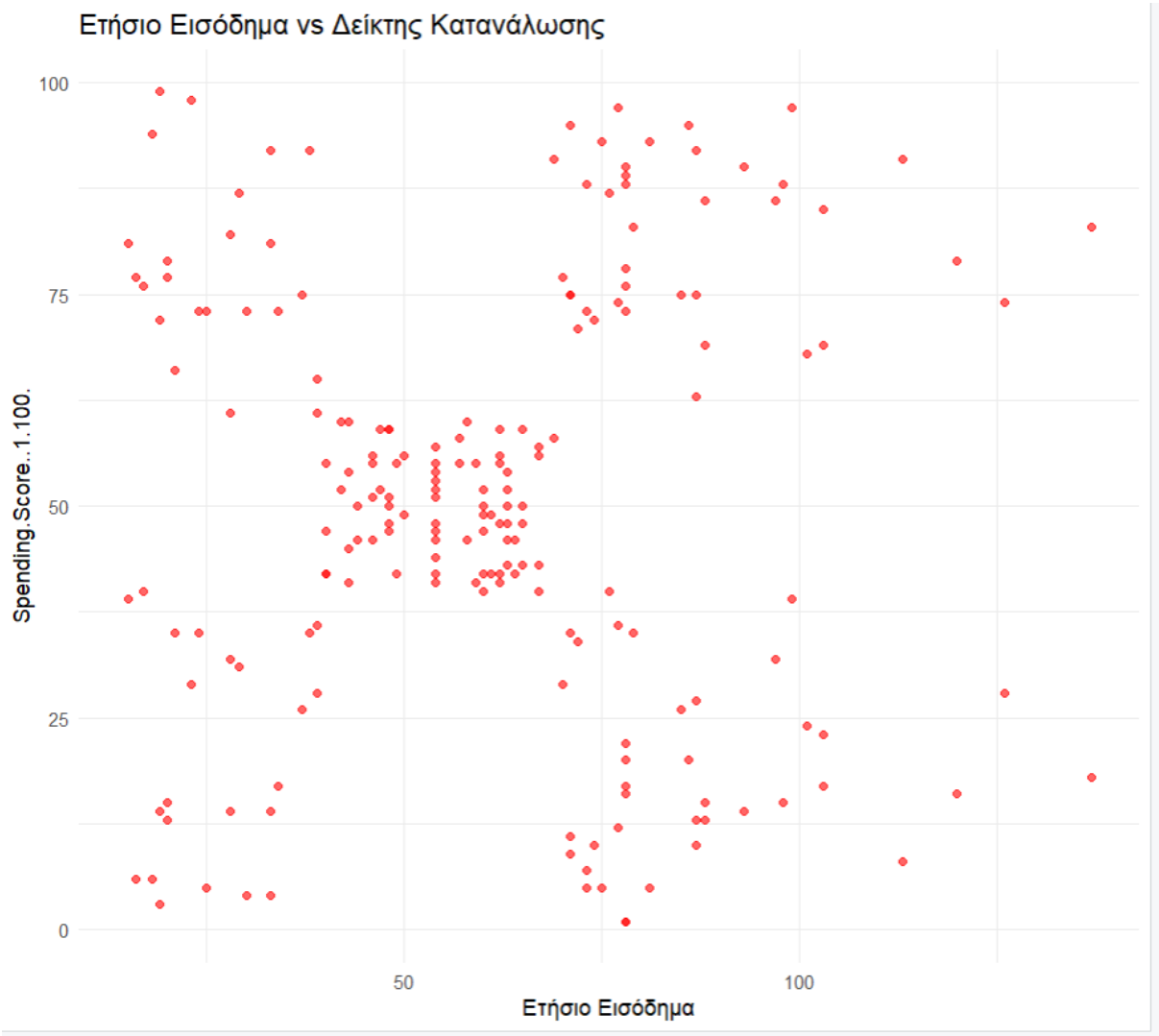


Ηλικία vs Δείκτης Κατανάλωσης (Age vs Spending Score)

- Στον πρώτο γράφημα, παρατηρούμε ότι οι νεότεροι πελάτες (20-40 ετών) τείνουν να έχουν μεγαλύτερη διακύμανση στον δείκτη κατανάλωσης, με αρκετούς να ξοδεύουν πολύ υψηλά ή πολύ χαμηλά.
- Οι μεγαλύτεροι σε ηλικία (50+ ετών) φαίνεται να έχουν πιο σταθερή κατανάλωση, συγκεντρωμένοι κυρίως στα μεσαία επίπεδα δαπανών.
- Η γενική τάση δείχνει μια αρνητική συσχέτιση, δηλαδή όσο αυξάνεται η ηλικία, μειώνεται ο δείκτης κατανάλωσης.

Scatter plot2:

Ετήσιο Εισόδημα vs Δείκτης Κατανάλωσης (για να δούμε αν το εισόδημα επηρεάζει την καταναλωτική συμπεριφορά)



Ετήσιο Εισόδημα vs Δείκτης Κατανάλωσης (Annual Income vs Spending Score)

- Στο δεύτερο γράφημα, υπάρχουν τέσσερις διακριτές ομάδες:
    1. Χαμηλό εισόδημα – Χαμηλή κατανάλωση
    2. Χαμηλό εισόδημα – Υψηλή κατανάλωση
    3. Υψηλό εισόδημα – Χαμηλή κατανάλωση
    4. Υψηλό εισόδημα – Υψηλή κατανάλωση
- Αυτό υποδηλώνει ότι το εισόδημα δεν σχετίζεται γραμμικά με το πόσο ξοδεύουν οι πελάτες. Αντίθετα, υπάρχουν διαφορετικά προφίλ καταναλωτών που ξοδεύουν είτε πολύ είτε λίγο ανεξάρτητα από το εισόδημά τους.

Συμπέρασμα: Οι πελάτες δεν ακολουθούν μια απλή γραμμική σχέση μεταξύ εισοδήματος και δαπανών. Ηλικία και καταναλωτική συμπεριφορά έχουν κάποια συσχέτιση, αλλά δεν είναι απόλυτη.