



ΕΙΣΑΓΩΓΗ ΣΤΟΥΣ Η/Υ & ΕΦΑΡΜΟΓΕΣ

Σημειώσεις Εργαστηρίου: Εργαστηριακή Άσκηση 8-
R Γραφήματα

Βικτωρία Δασκάλου, Εμμανουήλ Τζαγκαράκης
daskalou@upatras.gr, tzagara@upatras.gr

Περιεχόμενα

Στόχος	2
Διάβασμα και μετασχηματισμός δεδομένων.....	2
Δημιουργία σεναρίου.....	2
Διάβασμα αρχείου δεδομένων	2
Ερώτημα 1: Ραβδόγραμμα ή γράφημα στηλών.....	5
Προετοιμασία δεδομένων: Άθροισμα στηλών	6
Ραβδόγραμμα.....	6
Ερώτημα 2: Γράφημα μπάρας.....	9
Πίνακας συχνοτήτων κατηγορικών μεταβλητών.....	9
Ερώτημα 3	11
Θηκόγραμμα.....	11
Ιστόγραμμα	13

Στόχος

Ο στόχος της Εργαστηριακής Άσκησης 8 είναι η εισαγωγή σε γραφήματα με τη χρήση της βασικής βιβλιοθήκης της R (base).

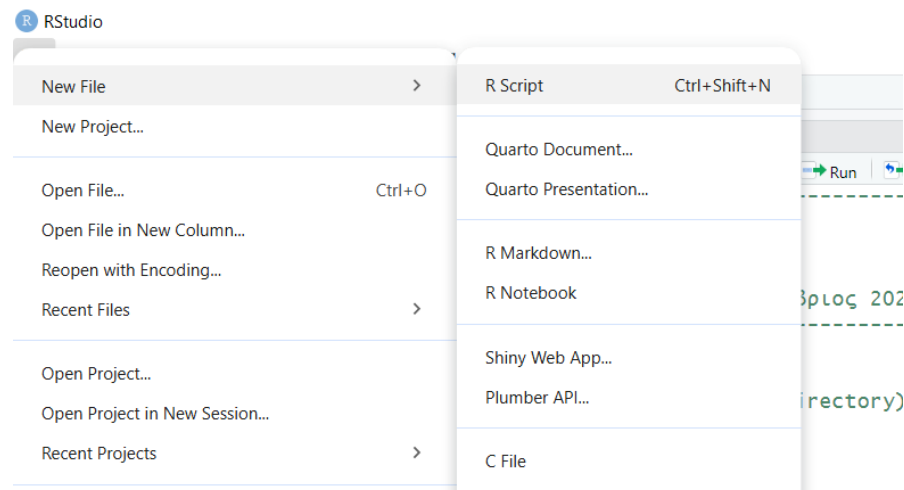
Διάβασμα και μετασχηματισμός δεδομένων

Κατεβάζουμε το αρχείο δεδομένων της Άσκησης με όνομα `greek_cities_census_2021.csv` που περιέχει στοιχεία για τους δήμους της Ελλάδας σύμφωνα με την απογραφή του 2021 στις στήλες:

- `code`: γεωγραφικός κωδικός δήμου
- `name`: όνομα δήμου
- `regional_unit`: περιφερειακή ενότητα που ανήκει ο δήμος
- `region`: περιφέρεια που ανήκει ο δήμος
- `total`: συνολικός πληθυσμός δήμου
- `men`: συνολικός πληθυσμός ανδρών δήμου
- `women`: συνολικός πληθυσμός γυναικών δήμου
- `num_de`: αριθμός δημοτικών ενοτήτων που ανήκουν στο δήμο
- `num_dk`: αριθμός δημοτικών κοινοτήτων που ανήκουν στο δήμο
- `num_oik`: αριθμός οικισμών που ανήκουν στο δήμο

Δημιουργία σεναρίου

Δημιουργούμε ένα νέο σενάριο στο RStudio, επιλέγοντας «**File->New File-> R Script**» (Εικόνα 1). Το προκαθορισμένο όνομα του νέου σεναρίου είναι `Untitled1` και πρέπει να το αλλάξουμε όταν το αποθηκεύσουμε πρώτη φορά.

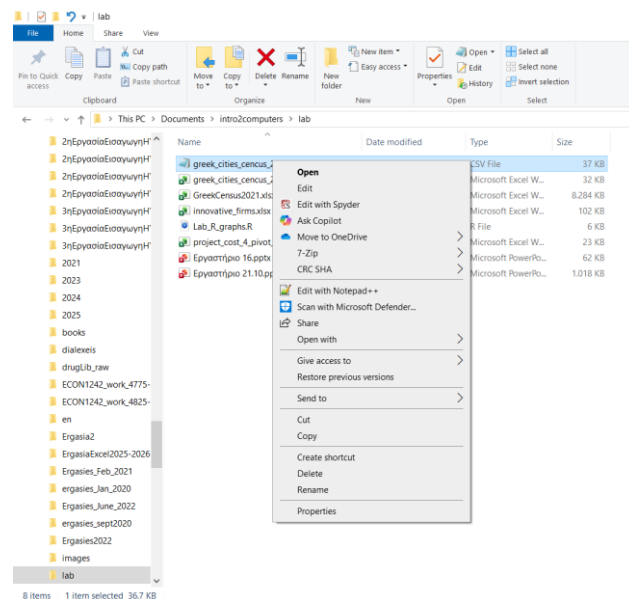


Εικόνα 1: Επιλογές για δημιουργία νέου σεναρίου στο RStudio

Διάβασμα αρχείου δεδομένων

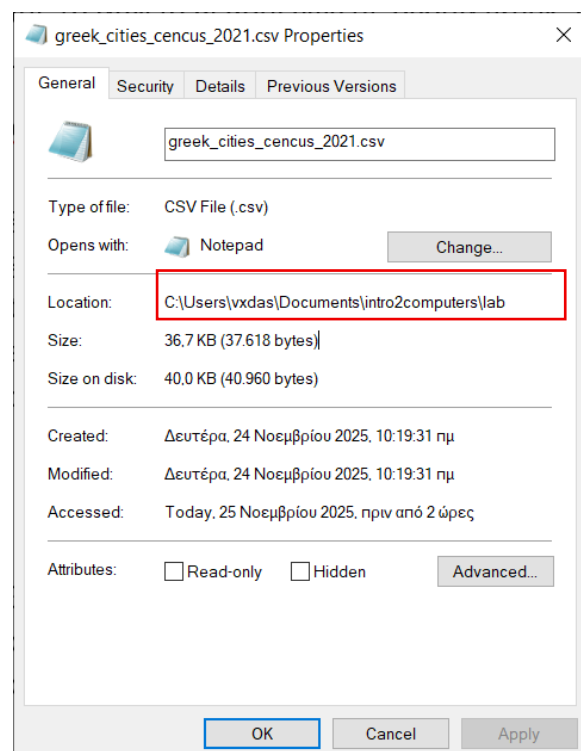
Για να διαβάσουμε το αρχείο δεδομένων θα πρέπει να γνωρίζουμε το φάκελο που έχουμε αποθηκεύσει το αρχείο δεδομένων που δίνει η άσκηση. Θα χρειαστεί να αντιγράψουμε το **μονοπάτι του φακέλου** στο πρόγραμμά μας. Στο πρόγραμμα **File Explorer** των MS Windows

βρίσκουμε το αρχείο και κάνουμε **δεξί κλικ** πάνω στο όνομα του αρχείου ώστε να εμφανιστεί το μενού όπου επιλέγουμε «**Properties (Ιδιότητες)**».



Εικόνα 2: Εύρεση Ιδιοτήτων (Properties) αρχείου στον Εξερευνητή Αρχείων (File Explorer)

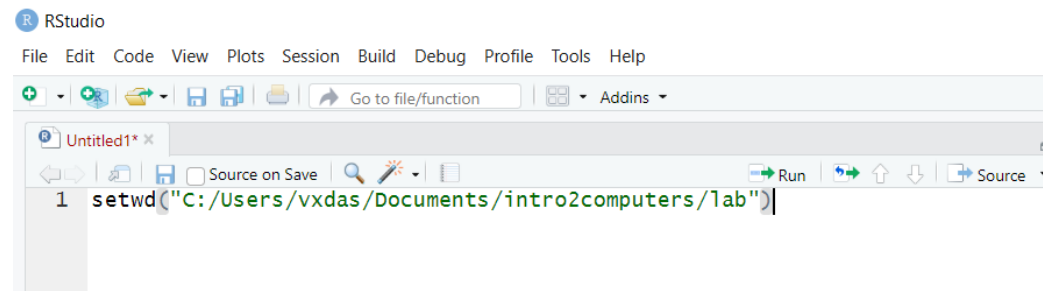
Στο παράθυρο με τις ιδιότητες του αρχείου επιλέγουμε και **αντιγράφουμε** τα περιεχόμενα του πεδίου «**Location**» για να τα χρησιμοποιήσουμε στο πρόγραμμά μας (βλέπε Εικόνα 3).



Εικόνα 3: Από τις Ιδιότητες αντιγράφουμε το μονοπάτι του αρχείου από το πεδίο Location

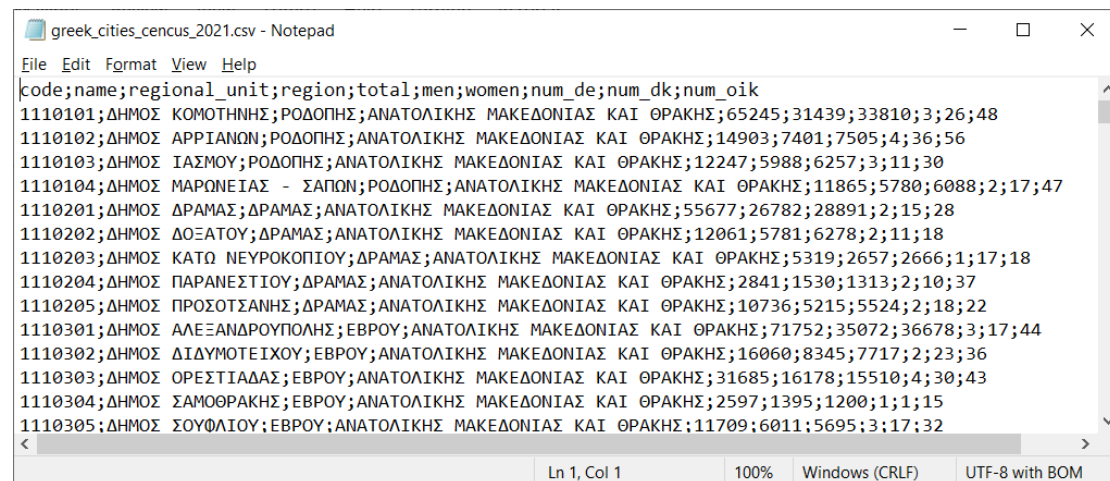
Το location του αρχείου "**C:/Users/vxdas/Documents/intro2computers/lab**" γίνεται όρισμα της συνάρτησης **setwd()** (set working directory), που ορίζει το **τρέχον ευρετήριο**, δηλαδή το φάκελο όπου θα αναζητά η R τα αρχεία προς διάβασμα και εκεί όπου θα γράφει τα αρχεία

εξόδου. **Προσοχή:** οι χαρακτήρες **backslash (\)** στο μονοπάτι θα πρέπει να αλλάξουν σε **slash (/)** (βλέπε Εικόνα 3).



Εικόνα 4: Η εντολή `setwd()` καθορίζει το τρέχον ευρετήριο όπου βρίσκονται τα δεδομένα, προσοχή στην αλλαγή του **backslash (\)** σε **slash (/)**

Το αρχείο εισόδου **`greek_cities_cencus_2021.csv`** είναι CSV, Comma Separated Value, αλλά όπως θα δούμε ανοίγοντας το αρχείο με το Notepad (βλέπε Εικόνα 5), το αρχείο έχει διαχωριστή στηλών το χαρακτήρα **';**'. Επίσης έχει επικεφαλίδα με τα ονόματα των στηλών και περιέχει ελληνικούς χαρακτήρες.



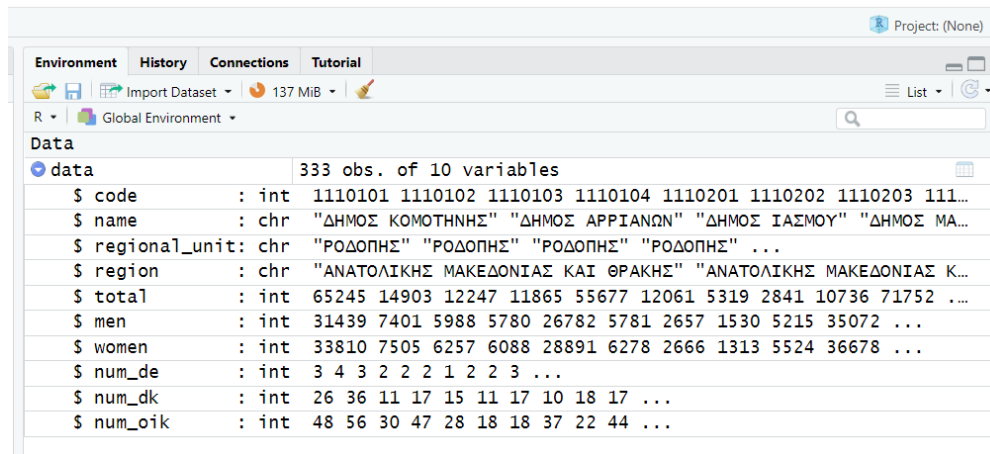
Εικόνα 5: Το αρχείο δεδομένων τύπου CSV

Στη συνέχεια θα χρησιμοποιήσουμε την εντολή **`read.csv()`** για να διαβάσουμε το αρχείο CSV με τα ακόλουθα ορίσματα. Τα σχόλια εξηγούν το κάθε όρισμα.

```
data <- read.csv('greek_cities_cencus_2021.csv',  
                header=TRUE, # έχει επικεφαλίδες  
                sep=';', # διαχωριστής στηλών  
                fileEncoding = 'utf-8', # κωδικοποίηση για ελληνικά  
                stringsAsFactors = FALSE) # όχι μετατροπή στηλών χαρακτήρων σε παράγοντες
```

Το τελευταίο όρισμα, **`stringsAsFactors = FALSE`**, το χρησιμοποιούμε γιατί θέλουμε εμείς συγκεκριμένα να καθορίσουμε ποιες στήλες χαρακτήρων θα γίνουν παράγοντες. Διαφορετικά όλες οι στήλες χαρακτήρων θα μετατρέπονταν σε παράγοντες, και η στήλη των ονομάτων των δήμων που είναι απλά αλφαριθμητικά και όχι κάποια κατηγορική μεταβλητή.

Μετά την εκτέλεση της εντολής η μεταβλητή **data** είναι ένα **πλαίσιο δεδομένων** (dataframe), με 333 παρατηρήσεις (obs από το observations) και 10 μεταβλητές (variables), όπου έχουν διαβαστεί τα περιεχόμενα του αρχείου όπως βλέπουμε στην περιοχή **Environment** του RStudio (Εικόνα 6).



The screenshot shows the RStudio Environment pane. The 'Data' tab is active, displaying a dataframe named 'data'. It has 333 observations and 10 variables. The variables and their data types are: code (integer), name (character), regional_unit (character), region (character), total (integer), men (integer), women (integer), num_de (integer), num_dk (integer), and num_oik (integer). The first few rows of data are visible.

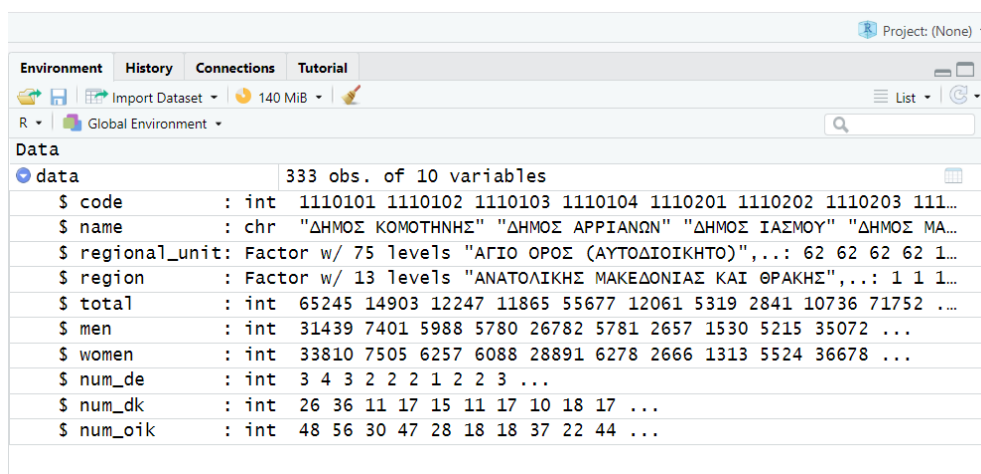
	code	name	regional_unit	region	total	men	women	num_de	num_dk	num_oik
1	1110101	"ΔΗΜΟΣ ΚΟΜΟΤΗΝΗΣ"	"ΡΟΔΟΠΗΣ"	"ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ ΚΑΙ ΘΡΑΚΗΣ"	65245	31439	33810	3	26	48
2	1110102	"ΔΗΜΟΣ ΑΡΡΙΑΝΩΝ"	"ΡΟΔΟΠΗΣ"	"ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ ΚΑΙ ΘΡΑΚΗΣ"	14903	7401	7505	4	36	56
3	1110103	"ΔΗΜΟΣ ΙΑΣΜΟΥ"	"ΡΟΔΟΠΗΣ"	"ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ ΚΑΙ ΘΡΑΚΗΣ"	12247	5988	6257	2	11	30
4	1110104	"ΔΗΜΟΣ ΜΑ..."	"ΡΟΔΟΠΗΣ"	"ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ ΚΑΙ ΘΡΑΚΗΣ"	11865	5780	6088	2	17	47

Εικόνα 6: Το πλαίσιο δεδομένων όπου διαβάστηκαν τα περιεχόμενα του αρχείου δεδομένων στην περιοχή **Environment** του RStudio

Κάνουμε τους μετασχηματισμούς των δύο κατηγορικών στηλών σε παράγοντες ως εξής:

```
data$regional_unit <- as.factor(data$regional_unit)
data$region <- as.factor(data$region)
```

Πλέον στο πλαίσιο δεδομένων οι μεταβλητές **region** και **regional_unit** είναι παράγοντες και βλέπουμε στην Εικόνα 7 ότι υπάρχουν 13 περιφέρειες και 75 περιφερειακές ενότητες.



The screenshot shows the RStudio Environment pane after the conversion. The 'data' dataframe now has 'regional_unit' as a factor with 75 levels and 'region' as a factor with 13 levels. The other variables remain the same.

	code	name	regional_unit	region	total	men	women	num_de	num_dk	num_oik
1	1110101	"ΔΗΜΟΣ ΚΟΜΟΤΗΝΗΣ"	"ΑΓΙΟ ΟΡΟΣ (ΑΥΤΟΔΙΟΙΚΗΤΟ)"	"ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ ΚΑΙ ΘΡΑΚΗΣ"	65245	31439	33810	3	26	48
2	1110102	"ΔΗΜΟΣ ΑΡΡΙΑΝΩΝ"	"ΑΓΙΟ ΟΡΟΣ (ΑΥΤΟΔΙΟΙΚΗΤΟ)"	"ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ ΚΑΙ ΘΡΑΚΗΣ"	14903	7401	7505	4	36	56
3	1110103	"ΔΗΜΟΣ ΙΑΣΜΟΥ"	"ΑΓΙΟ ΟΡΟΣ (ΑΥΤΟΔΙΟΙΚΗΤΟ)"	"ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ ΚΑΙ ΘΡΑΚΗΣ"	12247	5988	6257	2	11	30
4	1110104	"ΔΗΜΟΣ ΜΑ..."	"ΑΓΙΟ ΟΡΟΣ (ΑΥΤΟΔΙΟΙΚΗΤΟ)"	"ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ ΚΑΙ ΘΡΑΚΗΣ"	11865	5780	6088	2	17	47

Εικόνα 7: Το πλαίσιο δεδομένων μετά την μετατροπή των μεταβλητών **region** και **regional_unit** σε παράγοντες

Ερώτημα 1: Ραβδόγραμμα ή γράφημα στηλών

Βρείτε το σύνολο των ανδρών και γυναικών σε όλους τους δήμους και δημιουργήστε ένα ραβδόγραμμα. Πόσοι είναι οι άνδρες;

Προετοιμασία δεδομένων: Άθροισμα στηλών

Για να βρούμε το σύνολο των ανδρών και των γυναικών σε όλους τους δήμους θα πρέπει να βρούμε το **άθροισμα των στηλών men και women** του πλαισίου δεδομένων. Αυτό το κάνει η εντολή **colSums()** ως εξής:

```
total_m_w <- colSums(data[,c("men", "women")])
```

Η μεταβλητή `total_m_w` είναι ένα **ονοματισμένο διάνυσμα (named vector)** ακεραίων με δύο στοιχεία ως το άθροισμα των στηλών `men` και `women`, που το καθένα παίρνει το όνομα της στήλης από την οποία προήρθε και μπορούμε να το ανακτήσουμε με τον τελεστή `[]` και το όνομά του ως εξής:

```
> total_m_w
  men  women
5125968 5356480
> total_m_w["men"]
  men
5125968
```

Ραβδόγραμμα

Για τη σχεδίαση ενός ραβδογράμματος ή γραφήματος στηλών χρησιμοποιούμε τη συνάρτηση **barplot()** με τις βασικές παραμέτρους ως εξής:

```
barplot(height, xlab, ylab, main, names.arg, col)
```

Κύριοι παράμετροι:

- **height:** ένα διάνυσμα με τις τιμές που θα καθορίσουν το ύψος της κάθε στήλης
- **xlab:** ο τίτλος του άξονα X
- **ylab:** ο τίτλος του άξονα Y
- **main:** ο τίτλος του γραφήματος στηλών
- **names.arg:** διάνυσμα με τον τίτλο της κάθε στήλης, αν παραληφθεί και το `height` είναι ονοματισμένο διάνυσμα χρησιμοποιούνται τα ονόματα των στοιχείων του
- **col:** το χρώμα των στηλών

Δείτε την [τεκμηρίωση](#) για όλες τις παραμέτρους.

Εκτελώντας την εντολή

```
b1<-barplot(total_m_w, main="Σύνολο ανδρών-γυναικών δήμων", col=c("red","blue"))
```

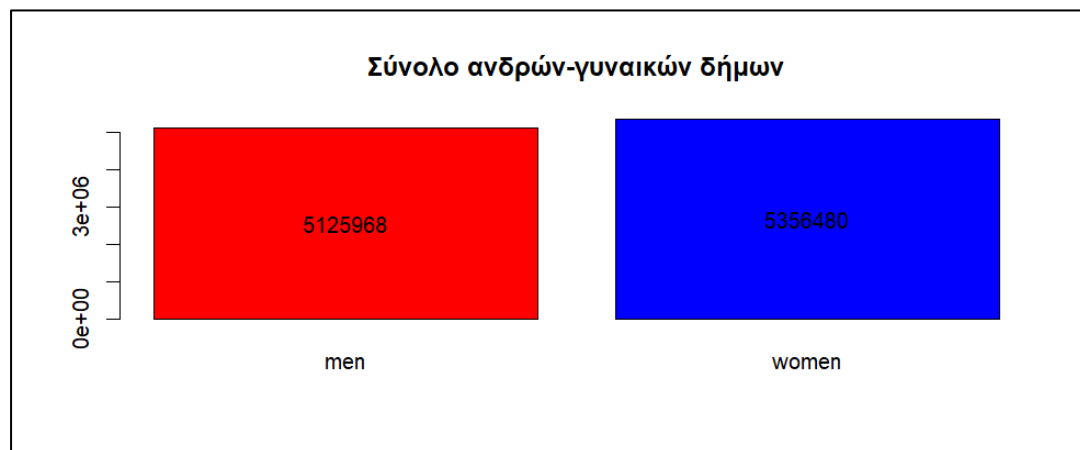
έχουμε το ακόλουθο αποτέλεσμα στο χώρο των **Plots** του RStudio (βλέπε Εικόνα 8).



Εικόνα 8: Στην περιοχή Plots του RStudio εμφανίζονται τα γραφήματα ως αποτελέσματα των σχετικών εντολών

Για να βάλουμε ετικέτες στο γράφημα με τις τιμές θα χρησιμοποιήσουμε την ακόλουθη χρήση της συνάρτησης `text(x, y, labels)`, όπως θα εξηγήσουμε στη συνέχεια:

```
text (b1, total_m_w/2, labels=total_m_w)
```



Εικόνα 9: Για προσθήκη ετικετών στο γράφημα χρησιμοποιούμε τη συνάρτηση `text()`

Κάθε γράφημα στο σύστημα γραφικών της R θεωρείται ως *σχήμα (figure)* και αποτελείται από 2 κύριες περιοχές: την *περιοχή σχεδίασης (plot region)* που ορίζεται από τους άξονες και περιβάλλεται από την περιοχή με τα *περιθώρια (margins)*, τα οποία περιέχουν τους τίτλους και τις ετικέτες των αξόνων. Οι περιοχές παρουσιάζονται στην Εικόνα 10.

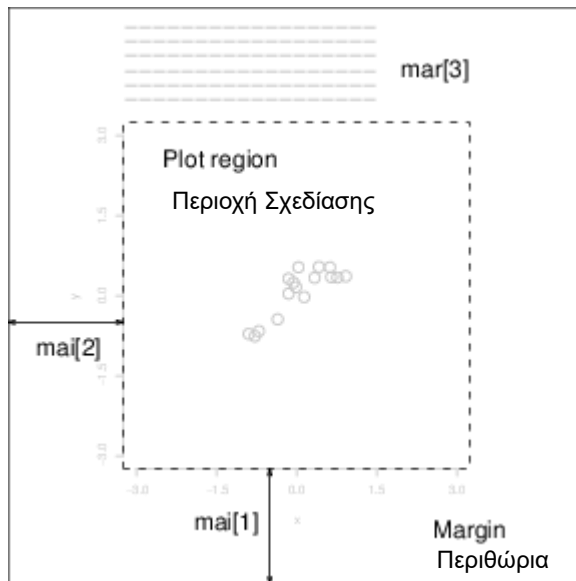


Figure region -Περιοχή σχήματος

Εικόνα 10: Οι περιοχές ενός γραφήματος σε R. Πηγή: <https://rstudio.github.io/r-manuals/r-intro/Graphics.html>

Αναλυτικότερα:

1. Περιοχή σχήματος (Figure region)

- **Ποια είναι:** Συνολική περιοχή που περιλαμβάνει το συνολικό γράφημα με το σχέδιο, άξονες, τίτλους και ετικέτες.
- **Σύστημα συντεταγμένων:** δεν ελέγχεται από το χρήστη

2. Περιθώρια (Margins)

- **Ποια είναι:** Είναι η περιοχή εκτός σχήματος, που ορίζεται από τους άξονες και όπου εμφανίζονται οι τίτλοι και τα σημεία των αξόνων, καθώς και ο κύριος τίτλος του γραφήματος
- **Έλεγχος περιοχής:** Προσαρμόζεται με την εντολή `par(mar = c(bottom, left, top, right))`. Οι τιμές αφορούν γραμμές κειμένου.

3. Περιοχή Σχεδίασης (Plot area)

- **Ποια είναι:** Το κεντρικό ορθογώνιο όπου σχεδιάζονται τα δεδομένα, ανάλογα με το είδος του γραφήματος, σε μορφή ράβδων, σημείων, γραμμών και όπου εμφανίζονται απλά γραφικά με τη χρήση της συνάρτησης `text()`.
- **Σύστημα συντεταγμένων:** Είναι **Συντεταγμένες Χρήστη (user coordinates)** που αφορούν:
 - Τιμές x και y που αναφέρονται στο εύρος των δεδομένων μας, για παράδειγμα το y εδώ θα κυμαίνεται από 0 έως 5356480
 - Η συνάρτηση `barplot()` αυτόματα καθορίζει τα όρια των x και y για να προσαρμόζονται στα δεδομένα μας και **επιστρέφει το μέσο της κάθε στήλης**. Εδώ:

```
> b1
      [,1]
[1,]  0.7
[2,]  1.9
```

Έτσι η συνάρτηση `text(b1, total_m_w/2, labels=total_m_w)` θα εμφανίσει το κείμενο για τις ετικέτες με τις τιμές για τα x και y στο μέσο της κάθε μπάρας.

Ερώτημα 2: Γράφημα μπάρας

Δημιουργήστε πίνακα συχνοτήτων και οριζόντιο ραβδόγραμμα με το πλήθος των δήμων ανά περιφέρεια. Πόσους δήμους έχει η περιφέρεια «ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ»;

Πίνακας συχνοτήτων κατηγορικών μεταβλητών

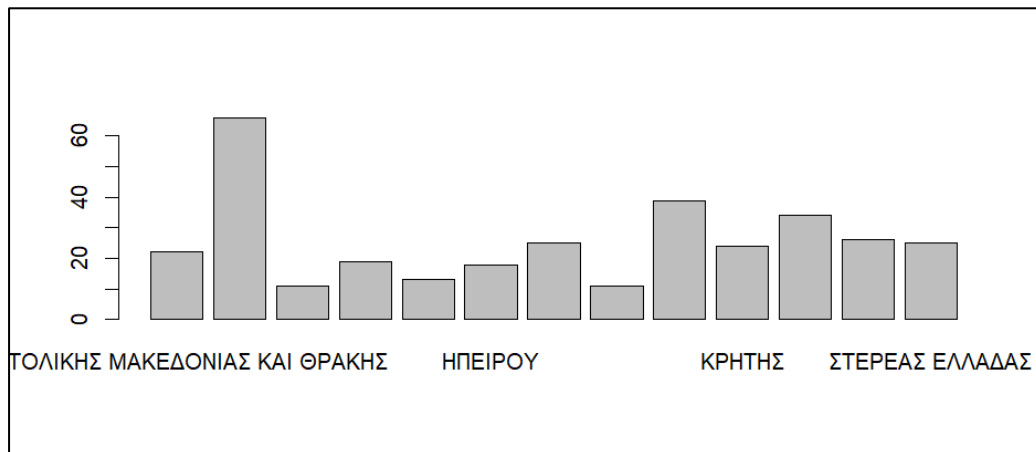
Για να βρούμε το πλήθος των δήμων ανά περιφέρεια μπορούμε να χρησιμοποιήσουμε τη συνάρτηση `table()`, η οποία βρίσκει τη συχνότητα εμφάνισης σε κατηγορικές μεταβλητές. Θα εκτελέσουμε λοιπόν την εντολή:

```
> cities_per_region <- table(data$region)
> cities_per_region
```

ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ ΚΑΙ ΘΡΑΚΗΣ	22	ΑΤΤΙΚΗΣ	66
ΒΟΡΕΙΟΥ ΑΙΓΑΙΟΥ	11	ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ	19
ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ	13	ΗΠΕΙΡΟΥ	18
ΘΕΣΣΑΛΙΑΣ	25	ΙΟΝΙΩΝ ΝΗΣΩΝ	11
ΚΕΝΤΡΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ	39	ΚΡΗΤΗΣ	24
ΝΟΤΙΟΥ ΑΙΓΑΙΟΥ	34	ΠΕΛΟΠΟΝΝΗΣΟΥ	26
ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ	25		

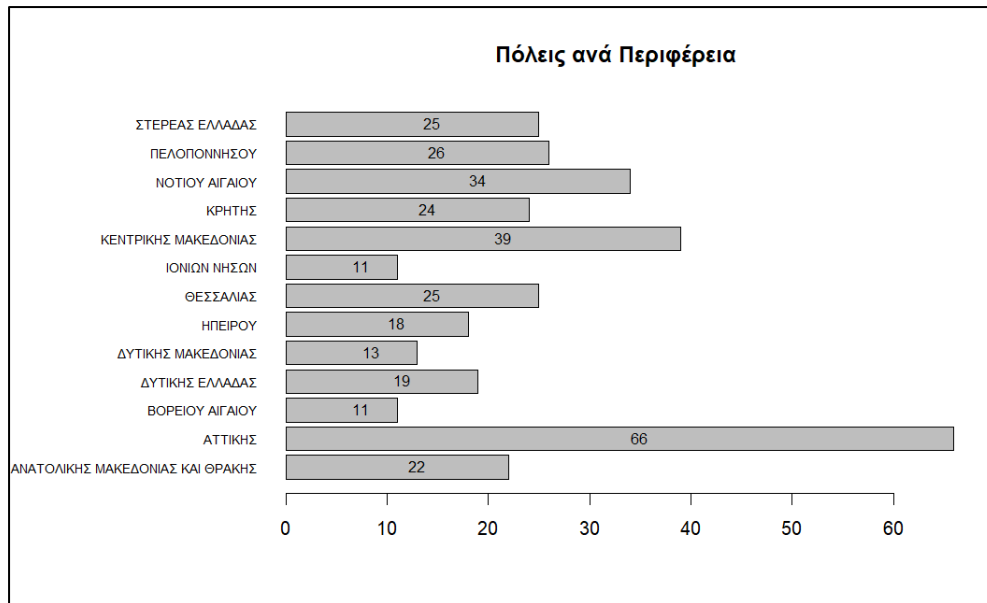
```
>
```

Αν χρησιμοποιήσουμε τη συνάρτηση `barplot(cities_per_region)` όπως προηγουμένως, επειδή τα ονόματα στον άξονα x είναι μεγάλα θα παίρναμε το αποτέλεσμα:



Εικόνα 11: Λόγω του μεγάλου μήκους των ονομάτων στον άξονα x, το διάγραμμα θα πρέπει να γίνει οριζόντιο

Για το λόγο αυτό μία λύση είναι να κάνουμε οριζόντιο γράφημα και να μεγαλώσουμε το αριστερό περιθώριο ώστε να χωρούν τα ονόματα ως ετικέτες, όπως παρουσιάζεται στην Εικόνα 12. Άλλη λύση θα ήταν να βάζαμε κάθετα τα ονόματα και να μεγαλώσουμε το κάτω περιθώριο. Θα προχωρήσουμε με την πρώτη λύση.



Εικόνα 12: Οριζόντιο γράφημα με μπάρες

Για το χειρισμό των περιθωρίων αρχικά θα αποθηκεύσουμε τις τρέχουσες παραμέτρους του γραφικού περιβάλλοντος που ορίζονται από το `par()`, θα μεγαλώσουμε το περιθώριο, και τέλος θα επαναφέρουμε τις αρχικές. Ενδιάμεσα θα φτιάξουμε το διάγραμμα και θα του βάλουμε ετικέτες. Οι εντολές είναι οι ακόλουθες:

```
# Αποθήκευσε τρέχουσες παραμέτρους γραφικού περιβάλλοντος
# αποθήκευσε μόνο τις παραμέτρους που ορίζονται από χρήστη
op <- par(no.readonly = TRUE)

# όρισε τα περιθώρια, margins, με μεγάλο το αριστερό
par(mar = c(5, 12, 4, 2))

# Εμφάνισε οριζόντιο ραβδόγραμμα (ορίσματα horiz = TRUE, las = 1)
b2 <- barplot(cities_per_region, horiz = TRUE, las = 1, cex.names = 0.7,
  main = "Πόλεις ανά Περιφέρεια") # προσθήκη τίτλου
# επανέφερε τις αρχικές παραμέτρους
par(op)
```

Για να εμφανίσουμε οριζόντιο γράφημα μπάρας πρέπει να χρησιμοποιήσουμε τη συνάρτηση `barplot()` με το όρισμα `horiz = TRUE`. Για αλλαγές προσανατολισμό στις ετικέτες και να φαίνονται οριζόντια χρησιμοποιούμε το όρισμα `las=1` (0 : παράλληλα με άξονα. 1 : οριζόντια στον άξονα. 2 : κάθετα στον άξονα).

Η συνάρτηση `barplot()` επιστρέφει στη μεταβλητή `b2`, τα σημεία στη μέση της μπάρας και αφού είναι οριζόντιο γράφημα, αφορούν τα `y`. Αντίστοιχα το `x` θα κυμαίνεται από 0 έως το μέγιστο πλήθος, δηλαδή το 66.

```
> b2
      [,1]
[1,] 0.7
[2,] 1.9
[3,] 3.1
[4,] 4.3
[5,] 5.5
[6,] 6.7
[7,] 7.9
[8,] 9.1
```

```
[9,] 10.3  
[10,] 11.5  
[11,] 12.7  
[12,] 13.9  
[13,] 15.1
```

Με βάση αυτά, για τις ετικέτες, έγινε και η σύνταξη της `text()` ως:

```
text(x = cities_per_region/2, # x δια 2 ώστε να κεντραριστούν  
     y = b2,                  # οι θέσεις y που δημιουργήθηκε το γράφημα  
     labels = cities_per_region, # κείμενο οι συχνότητες που βρήκαμε  
     pos = 4,                 # pos = 4 βάλε ετικέτα αριστερά του x  
     cex = 0.8)              # μέγεθος στο 80%
```

Ενδιαφέροντα παραδείγματα για ραβδογράμματα μπορείτε να βρείτε εδώ:

<https://viz-base.rsquaredacademy.com/bar>

Ερώτημα 3

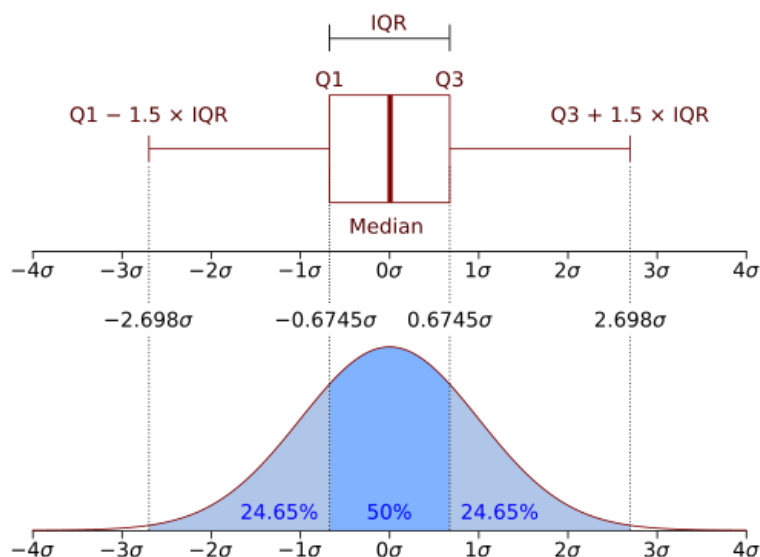
Δημιουργήστε θηκόγραμμα και ιστόγραμμα του αριθμού οικισμού δήμων

Θηκόγραμμα

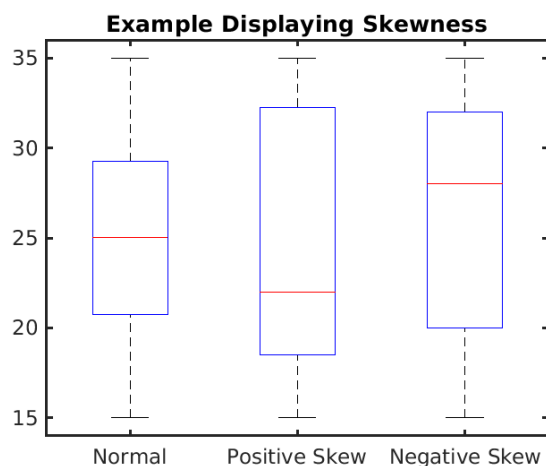
Το θηκόγραμμα (boxplot) είναι ένα γραφικό εργαλείο στην περιγραφική στατιστική που απεικονίζει βασικές τιμές μιας κατανομής δεδομένων και χρησιμοποιείται για την εύκολη σύγκριση συνόλων δεδομένων, τον έλεγχο συμμετρίας και την επισήμανση ακραίων τιμών. Τα μέρη ενός θηκογράμματος είναι:

- **Ορθογώνιο κουτί:** Η κάτω βάση του αντιστοιχεί στο πρώτο τεταρτημόριο (Q1) και η πάνω βάση στο τρίτο τεταρτημόριο (Q3). Το μήκος του είναι το ενδοτεταρτημοριακό εύρος $IQR=Q3-Q1$.
- **Διάμεσος:** Μια γραμμή μέσα στο κουτί, που δείχνει το δεύτερο τεταρτημόριο (Q2), δηλαδή τη διάμεσο των δεδομένων.
- **Κεραίες (Whiskers):** Γραμμές που εκτείνονται από το κουτί προς τα πάνω και προς τα κάτω, για να δείξουν τις οριακές τιμές των δεδομένων.
- **Ακραίες τιμές:** Σημεία εκτός των κεραίων, που συνήθως σημειώνονται με αστερίσκους ή κουκκίδες, για να υποδηλώσουν ακραίες τιμές.

Να θυμηθούμε το θηκόγραμμα σε μία κανονική κατανομή που είναι συμμετρική στην Εικόνα 13, ενώ στην Εικόνα 14 βλέπουμε τη μορφή που παίρνει το θηκόγραμμα στις 3 περιπτώσεις: της συμμετρίας, της θετικής και της αρνητικής συμμετρίας.



Εικόνα 13: Θηκόγραμμα της κανονικής κατανομής, Πηγή: [Jhguch at en.wikipedia](https://en.wikipedia.org/wiki/Normal_distribution#/media:File:Normal_distribution_boxplot.svg), [CC BY-SA 2.5](https://creativecommons.org/licenses/by-sa/2.5/), via Wikimedia Commons

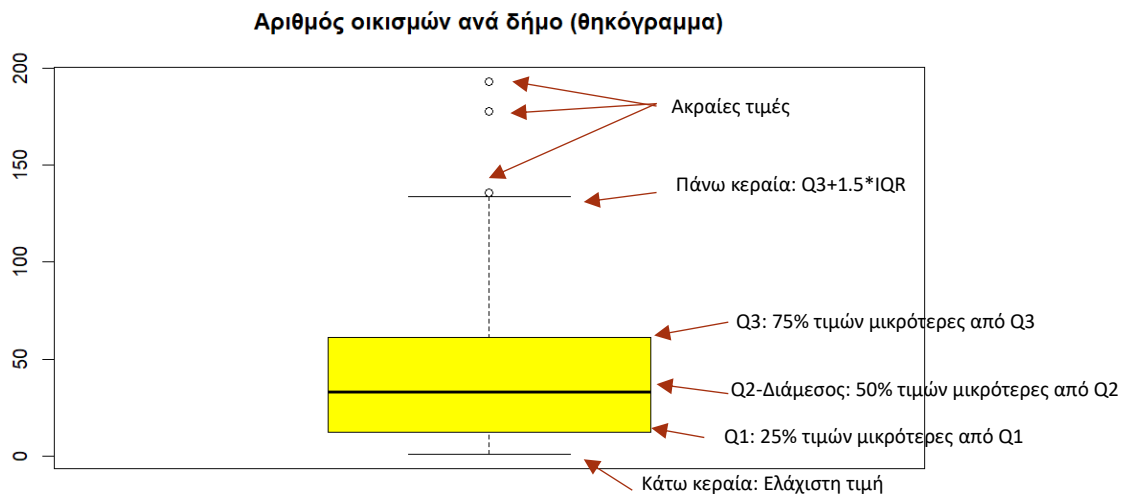


Εικόνα 14: Το θηκόγραμμα όταν έχουμε συμμετρία, θετική και αρνητική ασυμμετρία. Πηγή: [Ever.chae](https://commons.wikimedia.org/wiki/File:Boxplot_skewness.svg), [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/), via Wikimedia Commons

Για να δημιουργήσουμε το θηκόγραμμα στην R χρησιμοποιούμε την εντολή:

```
b4 <- boxplot(data$num_oik, col=c('yellow'), main="Αριθμός οικισμών ανά δήμο (θηκόγραμμα)")
```

Το αποτέλεσμα της εντολής με τις κατάλληλες επεξηγήσεις στα μέρη του θηκογράμματος παρουσιάζεται στην Εικόνα 15. Μελετήστε μόνοι σας τη μεταβλητή b4 που επιστρέφει η συνάρτηση και θα κατανοήσετε καλύτερα και τα συμπεράσματα.



Εικόνα 15: Θηκόγραμμα του αριθμού οικισμών ανά δήμο (με εξηγήσεις για τα σημεία)

Η μεταβλητή **b4** (Εικόνα 1), που επιστρέφει η συνάρτηση `boxplot()`, είναι μία λίστα με χρήσιμα στοιχεία. Το στοιχείο **b4\$stats** μας δίνει τις τιμές για τα 5 στατιστικά στοιχεία (Κάτω κεραία, Q1, Q2, Q3, Πάνω κεραία) και το στοιχείο **b4\$out** τις ακραίες τιμές.

b4	List of 6
\$ stats: num [1:5, 1]	1 12 33 61 134
\$ n : num	333
\$ conf : num [1:2, 1]	28.8 37.2
\$ out : num [1:4]	136 136 178 193
\$ group: num [1:4]	1 1 1 1
\$ names: chr	""

Εικόνα 16: Η μεταβλητή τύπου λίστας που επιστρέφει η συνάρτηση `boxplot()`

Κάποια συμπεράσματα από το θηκόγραμμα:

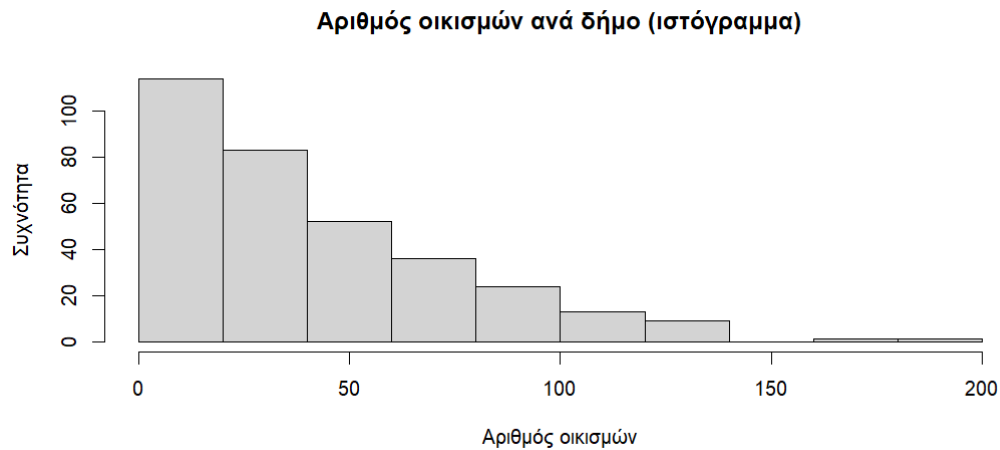
- Το 50% των δήμων έχουν από 12 (Q1) έως 61 (Q3) οικισμούς
- Ένας τυπικός δήμος έχει 33 οικισμούς (Q2, διάμεσος)
- Υπάρχουν ακραίες τιμές με τουλάχιστον 4 δήμους με πάνω από 134 οικισμούς (πάνω κεραία)
- Η κατανομή παρουσιάζει θετική ασυμμετρία. Αυτό σημαίνει ότι οι περισσότεροι δήμοι έχουν μικρότερο αριθμό οικισμών, αλλά μερικοί δήμοι έχουν πολύ υψηλό αριθμό, ανεβάζοντας τον μέσο όρο υψηλότερα από τη διάμεσο.

Ιστόγραμμα

Το ιστόγραμμα χρησιμοποιείται για τη γραφική απεικόνιση της κατανομής συχνοτήτων αριθμητικών μεταβλητών, όπως εδώ έχουμε το πλήθος των οικισμών ανά δήμο. Για τη δημιουργία του χρησιμοποιείται η εντολή:

```
h2 <- hist(data$num_oik, main="Αριθμός οικισμών ανά δήμο (ιστόγραμμα)",  
  xlab="Αριθμός οικισμών", # τίτλος άξονα x  
  ylab="Συχνότητα") # τίτλος άξονα y
```

Το αποτέλεσμα της εντολής με το ιστόγραμμα του αριθμού των οικισμών ανά δήμο, με τίτλους αξόνων και γραφήματος παρουσιάζεται στην Εικόνα 17.



Εικόνα 17: Ιστόγραμμα του αριθμού των οικισμών ανά δήμο

Βλέπουμε ότι η πλειοψηφία των δήμων έχουν μικρό αριθμό δήμων (πρώτες 3 μπάρες), ενώ υπάρχουν και ένας μικρός αριθμός δήμων με μεγάλο αριθμό οικισμών (τελευταίες μπάρες).

Λεπτομέρειες για διαμόρφωση ιστογράμματος μπορείτε να βρείτε και εδώ <https://viz-base.rsquaredacademy.com/hist>