



ΕΙΣΑΓΩΓΗ ΣΤΟΥΣ Η/Υ & ΕΦΑΡΜΟΓΕΣ

Σημειώσεις Εργαστηρίου: Εργαστηριακή Άσκηση 7
-R πλαίσια δεδομένων (dataframes)

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

Περιεχόμενα

Στόχος	2
Ο τύπος factor για ποιοτικές μεταβλητές	2
Ορισμός ποιοτικής μη-διατάξιμης μεταβλητής	2
Ορισμός ποιοτικής διατάξιμης μεταβλητής.....	2
Πίνακας συχνοτήτων ποιοτικών μεταβλητών	3
Ο τύπος dataframe (πλαίσιο δεδομένων).....	3
Τι είναι.....	3
Δημιουργία.....	3
Δημιουργία πλαισίου δεδομένων από διανύσματα.....	4
Δημιουργία πλαισίου δεδομένων από αρχείο CSV	5
Δημιουργία δοκιμαστικού αρχείου CSV.....	6
Ανάκτηση τμημάτων πλαισίου δεδομένων.....	10
Αναφορά σε στοιχείο	10
Αναφορά σε τμήμα πλαισίου δεδομένων.....	10
Αναφορά σε στήλες.....	10
Αναφορά σε γραμμές.....	14
Τεμαχισμός (slicing).....	14
Ανάκτηση τεμαχίου	14
Ανάκτηση τεμαχίου με λογικές συνθήκες (κριτήρια).....	15
Δημιουργία νέων στηλών σε πλαίσιο δεδομένων	16
Βασική στατιστική ανάλυση.....	17

Στόχος

Ο στόχος της άσκησης **ΕΡΓΑΣΤΗΡΙΟ 7: R-πλαίσια δεδομένων (dataframe)** είναι η εξοικείωση με τον τύπο παράγοντα (factor) για την αναπαράσταση ποιοτικών μεταβλητών και τον τύπο πλαίσια δεδομένων (dataframe) και τις σχετικές συναρτήσεις στο προγραμματιστικό περιβάλλον R για στατιστική επεξεργασία δεδομένων.

Μελετήστε τα ακόλουθα και εκτελέστε την με τη χρήση εντολών στο RStudio!

Ο τύπος factor για ποιοτικές μεταβλητές

Στο πλαίσιο της Οικονομικής Επιστήμης, εκτός από τις **ποσοτικές μεταβλητές** για την περιγραφή των οικονομικών φαινομένων (για παράδειγμα το ΑΕΠ, ο πληθωρισμός, κλπ.), χρησιμοποιούμε και **ποιοτικές** ή **κατηγορικές** μεταβλητές.

Οι ποιοτικές μεταβλητές είναι αυτές που δείχνουν ότι οι διάφοροι παράγοντες μεταβάλλονται κατά είδος. Διακρίνονται σε **διατάξιμες**, με τιμές που ιεραρχούνται (π.χ. επίπεδα θερμοκρασίας ως χαμηλή, μεσαία, υψηλή) και **μη-διατάξιμες ή κατηγορικές** (π.χ. φύλλο, επάγγελμα, κλπ.).

Στη γλώσσα R για την αναπαράσταση ποιοτικών μεταβλητών χρησιμοποιείται ο τύπος μεταβλητών **factor** (παράγοντας).

Ορισμός ποιοτικής μη-διατάξιμης μεταβλητής

Για να ορίσουμε μία ποιοτική μεταβλητή μη-διατάξιμη χρησιμοποιούμε τη συνάρτηση `factor()` με ορίσματα ένα διάνυσμα `c()` με τις τιμές που παίρνει η μεταβλητή. Η συνάρτηση δημιουργεί μία ποιοτική μεταβλητή αντιλαμβανόμενη τις μοναδικά διαφορετικές τιμές ως βαθμίδες (levels).

```
> # ποιοτική μη-διατάξιμη
> nationality<-factor(c('Italian','French','French','Greek','Italian'))
> nationality
[1] Italian French  French  Greek   Italian
Levels: French Greek Italian
```

myweight	80
nationality	Factor w/ 3 levels "French","Greek",...: 3 1 1 2 3
z	3+9i

Εικόνα 1: Η εμφάνιση μίας ποιοτικής μεταβλητής μη-διατάξιμης (τύπου factor) στο Περιβάλλον

Ορισμός ποιοτικής διατάξιμης μεταβλητής

Στην περίπτωση της διατάξιμης μεταβλητής, ο ορισμός με τη συνάρτηση `factor()` περιέχει 3 ορίσματα: τις τιμές, το όρισμα `orderd=TRUE` που σημαίνει ότι είναι διατάξιμη και τις βαθμίδες με προκαθορισμένη διάταξη ως διάνυσμα, για παράδειγμα `levels=c('good','very good','excellent')`, ώστε να καθοριστεί η διάταξη μεταξύ των βαθμίδων ως `good < very good < excellent`.

```

> replies=c('excellent','excellent','good','very good','good','good')
> results<-factor(replies,ordered=TRUE,levels=c('good','very
good','excellent'))
> results
[1] excellent excellent good very good good good
Levels: good < very good < excellent

```

replies	chr [1:6]	"excellent"	"excellent"	"good"	"very good"	"good"	"good"
results	ord.factor w/ 3 levels	"good"<	"very good"<	..:	3	3	1 2 1 1

Εικόνα 2: Η εμφάνιση μίας ποιοτικής μεταβλητής διατάξιμης (τύπου factor) στο Περιβάλλον

Πίνακας συχνοτήτων ποιοτικών μεταβλητών

Με τη χρήση της συνάρτησης `table()` και όρισμα το όνομα της ποιοτικής μεταβλητής μπορούμε να δημιουργήσουμε τον αντίστοιχο πίνακα συχνοτήτων. Για παράδειγμα:

```

> table(nationality)
nationality
French   Greek Italian
      2      1      2
> table(results)
results
      good very good excellent
      3      1      2

```

Ο τύπος dataframe (πλαίσιο δεδομένων)

Τι είναι

Οι μεταβλητές τύπου `dataframe` χρησιμοποιούνται για την **δισδιάστατη** αναπαράσταση δεδομένων όπου η **κάθε στήλη μπορεί να είναι διαφορετικού τύπου**.

Δημιουργία

Τα δεδομένα για τη δημιουργία ενός πλαισίου δεδομένων μπορεί να προέρχονται από:

1. **Διανύσματα**, όπου για τη δημιουργία τους χρησιμοποιούμε τη συνάρτηση `data.frame()`
2. **Αρχεία δεδομένων**, όπως κειμένου μορφής CSV (Comma Separated Values), όπου το πλαίσιο δεδομένων είναι η επιστρεφόμενη τιμή μίας συνάρτησης διαβάσματος του αρχείου όπως η `read.csv()`.

Δημιουργία πλαισίου δεδομένων από διανύσματα

Ας υποθέσουμε ότι έχουμε τα στοιχεία 4 φοιτητών σε τρία διανύσματα ως εξής:

```
names<-c("Maria","Nikos","Thanos","Niki") # διάνυσμα ονομάτων  
grades<-c(10,8,9,7) # διάνυσμα βαθμών  
gender<-factor(c("female","male","male","female")) # ποιοτική γένος
```

Για τη δημιουργία του πλαισίου δεδομένων θα έχουμε:

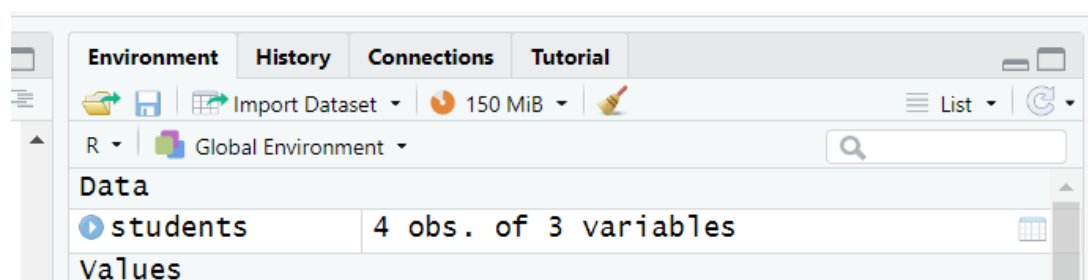
```
# δημιουργία πλαισίου δεδομένων  
students<-data.frame(names,grades,gender)
```

Για να δούμε το αποτέλεσμα της δημιουργίας μπορούμε στην *Κονσόλα* να γράψουμε:

```
> students  
  names grades gender  
1 Maria     10 female  
2 Nikos      8  male  
3 Thanos     9  male  
4 Niki       7 female
```

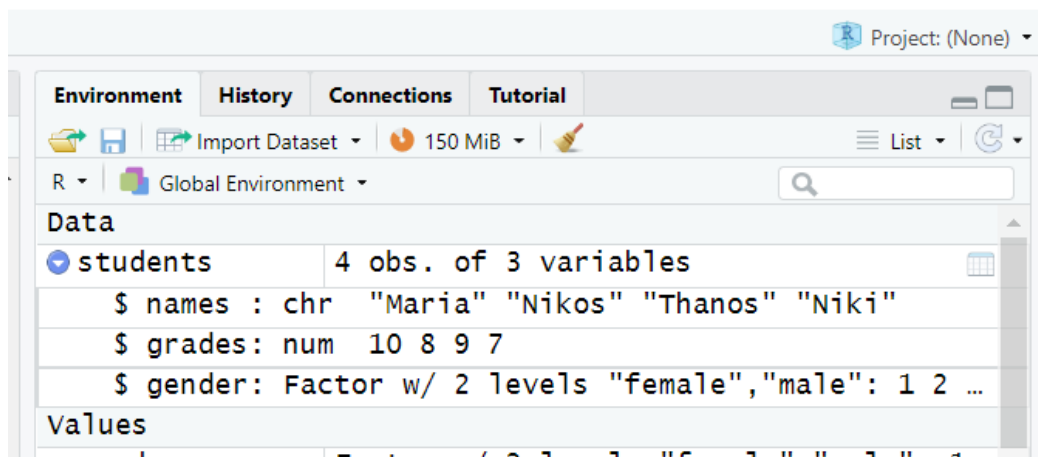
Παρατηρούμε ότι κάθε διάνυσμα αποτελεί μία στήλη του πλαισίου δεδομένων, και ότι το όνομα του διανύσματος λειτουργεί ως όνομα της στήλης.

Στο παράθυρο του *Περιβάλλοντος* παρατηρούμε επίσης, ότι το πλαίσιο δεδομένων έχει δημιουργηθεί στην περιοχή *Data* και οι γραμμές ονομάζονται *obs*, δηλαδή *objects*, και οι στήλες ονομάζονται *variables*, μεταβλητές.



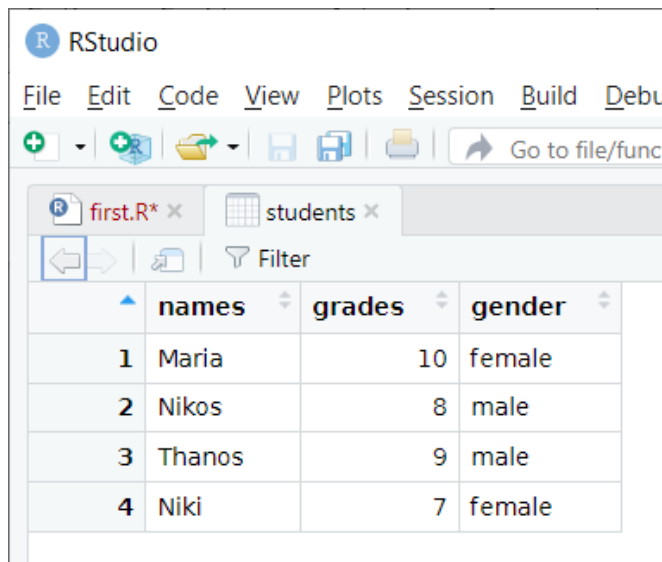
Εικόνα 3: Η εμφάνιση μίας μεταβλητής τύπου *dataframe* στο Περιβάλλον

Όταν πατάμε πάνω στο μπλε βελάκι που εμφανίζεται πριν από το όνομα το πλαισίου μπορούμε να δούμε αναλυτικά τις στήλες του: το όνομα, τον τύπο και τις πρώτες τιμές.



Εικόνα 4: Η εμφάνιση των στηλών μίας μεταβλητής τύπου dataframe στο Περιβάλλον

Αν πατήσουμε διπλό κλικ πάνω στο όνομα, εκτελείται η εντολή **View(students)** και παρουσιάζονται τα περιεχόμενα του πλαισίου δεδομένων σε ξεχωριστό tab στην περιοχή του *Κειμενογράφου*.



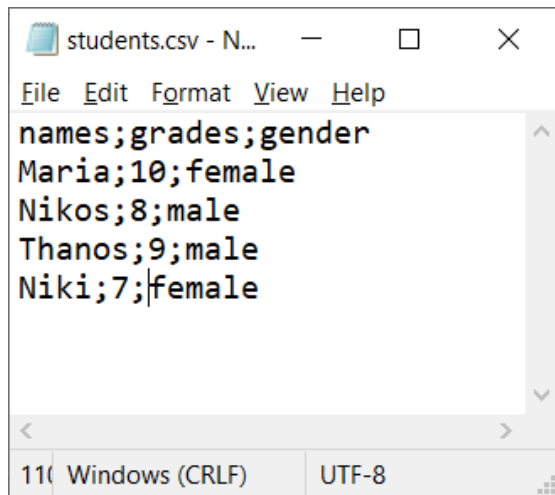
Εικόνα 5: Το αποτέλεσμα της εντολής View(students) σε ξεχωριστό tab στην περιοχή του Κειμενογράφου

Δημιουργία πλαισίου δεδομένων από αρχείο CSV

Τα αρχεία CSV (Comma Separated Values) είναι αρχεία κειμένου που προσομοιάζουν τη μορφή πίνακα. Οι στήλες χωρίζονται με ένα ειδικό **διαχωριστικό χαρακτήρα** (separator ή delimiter) που μπορεί να είναι το κόμμα (,), το ελληνικό ερωτηματικό (;), το tab, κλπ. Συνήθως η πρώτη γραμμή περιέχει μία **επικεφαλίδα** (header). Οι τιμές των αλφαριθμητικών συνήθως περιέχονται σε double quotes ("), ώστε όποιοι ειδικοί χαρακτήρες να μην λαμβάνονται ως διαχωριστικός χαρακτήρας.

Τα αρχεία CSV επεξεργάζονται από λογισμικό επεξεργασίας κειμένου, όπως το Notepad (Σημειωματάριο) και όχι από το λογισμικό λογιστικών φύλων όπως το MS Excel.

Ένα παράδειγμα αρχείου CSV με όνομα students.csv που έχει ανοιχθεί με το Σημειωματάριο είναι το ακόλουθο:



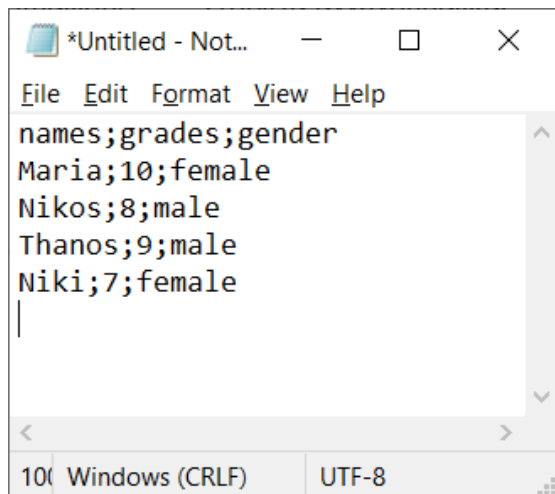
```
students.csv - N...
File Edit Format View Help
names;grades;gender
Maria;10;female
Nikos;8;male
Thanos;9;male
Niki;7;female
11( Windows (CRLF) UTF-8
```

Εικόνα 6: Τα περιεχόμενα του αρχείου students.csv

Δημιουργία δοκιμαστικού αρχείου CSV

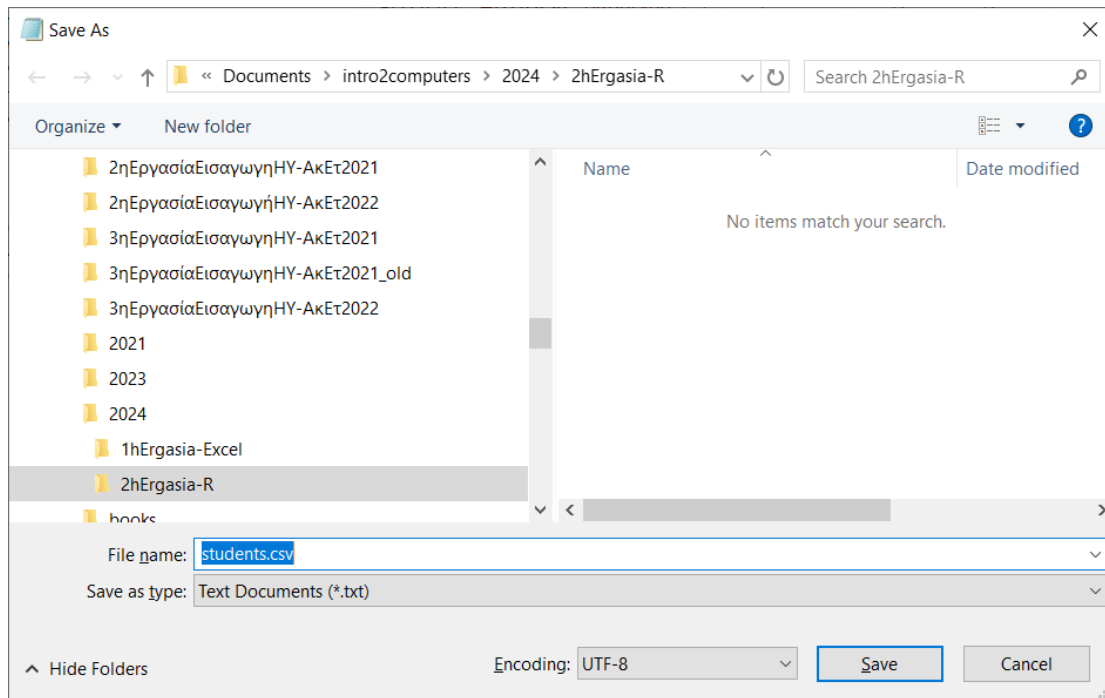
1. Ανοίξτε το Σημειωματάριο (Notepad) και αντιγράψτε τα ακόλουθα περιεχόμενα:

```
names;grades;gender
Maria;10;female
Nikos;8;male
Thanos;9;male
Niki;7;female
```



```
*Untitled - Not...
File Edit Format View Help
names;grades;gender
Maria;10;female
Nikos;8;male
Thanos;9;male
Niki;7;female
|
10( Windows (CRLF) UTF-8
```

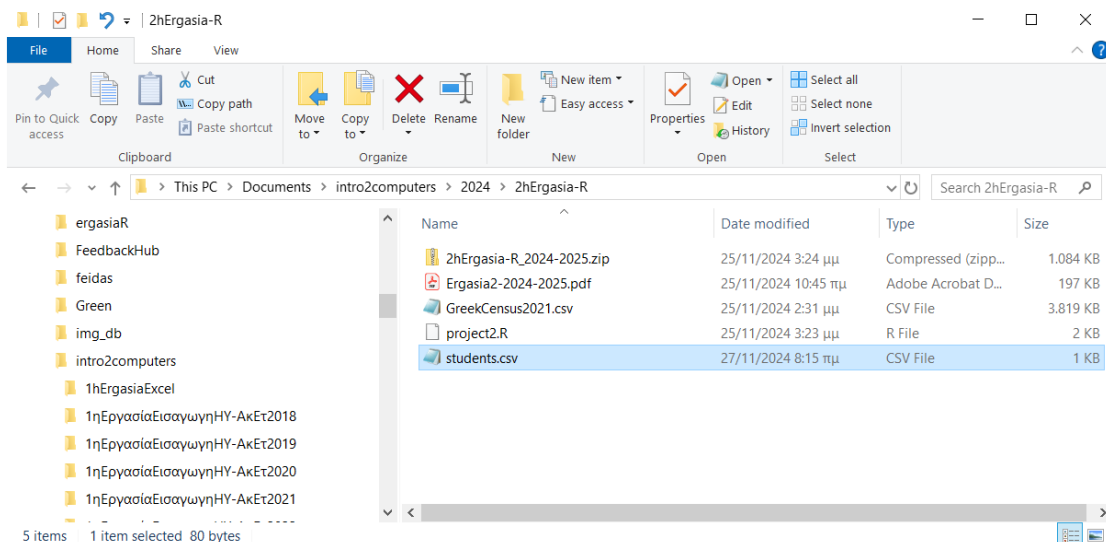
2. Αποθηκεύστε το αρχείο με το όνομα students.csv σε ένα συγκεκριμένο Φάκελο (να γνωρίζετε που το αποθηκεύσατε!). Στην αποθήκευση να σιγουρέψετε ότι η επιλογή Encoding αναφέρει UTF-8.



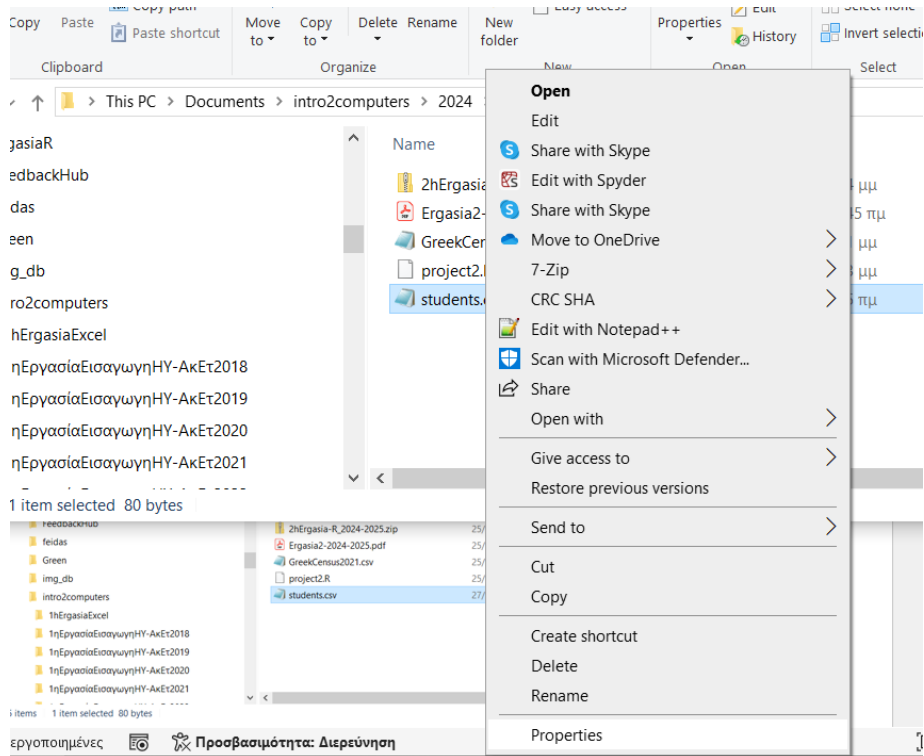
Για να διαβάσουμε ένα αρχείο δεδομένων θα πρέπει η γλώσσα R να γνωρίζει σε ποιο φάκελο βρίσκεται. Αυτό το κάνει η εντολή `setwd(τοποθεσία_στο_δίσκο)` (set working directory).

Για να βρούμε το φάκελο που βρίσκεται ένα αρχείο **CSV (αποσυμπιεσμένο)** κάνουμε τα εξής:

1. Ανοίγουμε τον File Explorer και πηγαίνουμε στο **Φάκελο** που βρίσκεται το αρχείο CSV

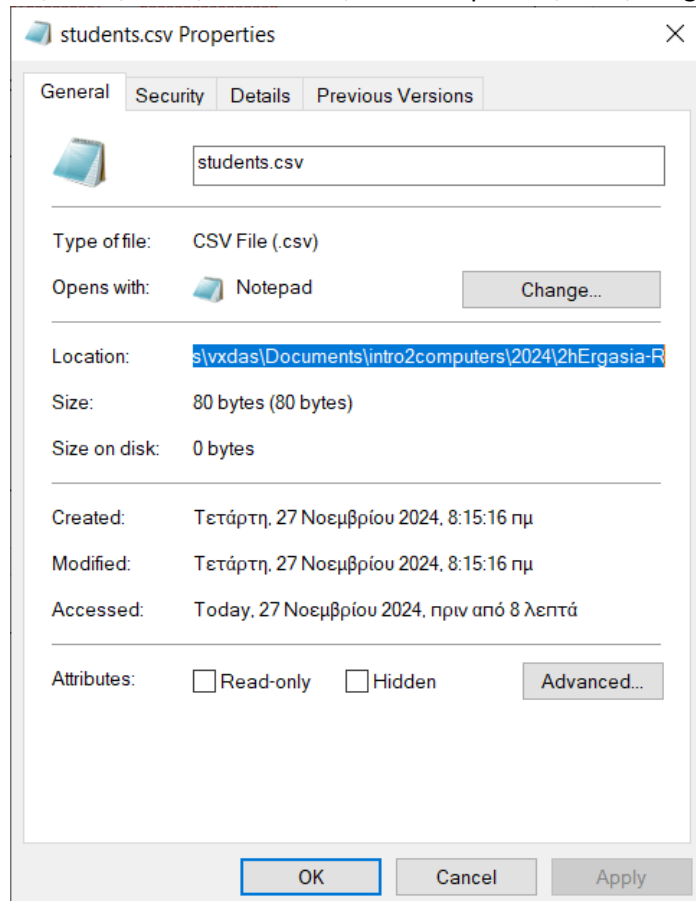


2. Κάνουμε δεξί κλικ πάνω στο αρχείο CSV και επιλέγουμε **Properties (Ιδιότητες)**



3. Στο παράθυρο που εμφανίζεται επιλέγουμε ότι αναφέρεται στο **Location (Τοποθεσία)** και το αντιγράφουμε (Ctrl+C). Στο συγκεκριμένο παράδειγμα στο Πρόχειρο του υπολογιστή μας θα υπάρχει το κείμενο:

C:\Users\vxdas\Documents\intro2computers\2024\2hErgasia-R



4. Στη συνέχεια επικολλούμε την τοποθεσία σε ένα R script και μέσα σε μία εντολή `setwd("τοποθεσία_στο_δίσκο")` αντικαθιστώντας παντού τους χαρακτήρες **back slash (\) με slash (/)**. Ο χαρακτήρας back slash (\) είναι ειδικός για την R και πρέπει να αντικατασταθεί.
5. Στη συνέχεια θα διαβάσουμε το αρχείο στη μνήμη σε μία μεταβλητή data με την εντολή `read.csv()`

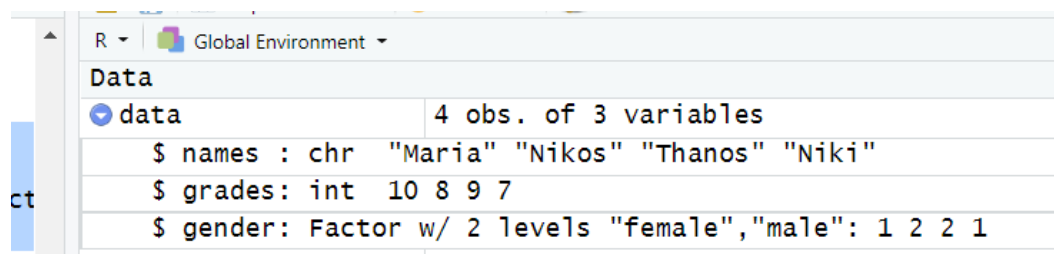
Ο κώδικας είναι ο ακόλουθος. Μπορείτε να τον αντιγράψετε σε ένα R script και να τον εκτελέσετε.

```
# αντικαθιστώ τα back slash \ με slash
setwd("C:/Users/vxdas/Documents/intro2computers/2024/2hErgasia-R")
data=read.csv('students.csv',header=TRUE,sep=';',fileEncoding =
'utf-8',stringsAsFactors = FALSE)
data$gender<-as.factor(data$gender) # μετατροπή σε factor
```

Τα ορίσματα της εντολής `read.csv()` είναι τα ακόλουθα:

<code>'students.csv'</code>	το όνομα του αρχείου σε quotes ή double quotes
<code>header=TRUE</code>	ορίζουμε αν η πρώτη γραμμή έχει τα ονόματα των στηλών, δηλαδή αν το αρχείο έχει header
<code>sep=';'</code>	ορίζουμε ποιός χαρακτήρας που διαχωρίζει τις στήλες. Στο συγκεκριμένο αρχείο είναι ο <code>';'</code>
<code>fileEncoding = 'utf-8'</code>	ορίζουμε ότι η κωδικοποίηση των χαρακτήρων του αρχείου είναι η UTF-8, ώστε να υποστηρίζονται γλώσσες όπως τα ελληνικά
<code>stringsAsFactors = FALSE</code>	η προκαθορισμένη τιμή του ορίσματος κάνει τη συνάρτηση να μετατρέπει τις στήλες με αλφαριθμητικά σε τύπου factor. Στη δική μας περίπτωση αυτό δεν ισχύει για το όνομα (στήλη names). Στη συνέχεια εμείς μέσω της συνάρτησης <code>as.factor()</code> θα αλλάξουμε τον τύπο στις στήλες που επιθυμούμε.

Στη μνήμη θα δημιουργηθεί μία μεταβλητή τύπου dataframe με όνομα data ως εξής:



Εικόνα 7: Η μεταβλητή data στο Περιβάλλον

Ανάκτηση τμημάτων πλαισίου δεδομένων

Αναφορά σε στοιχείο

Έχουμε πρόσβαση σε ένα **μεμονωμένο στοιχείο** του πλαισίου δεδομένων με τη χρήση του **ονόματος** του πλαισίου και τον τελεστή `[]`, και προσδιορίζοντας τη **θέση της γραμμής και το όνομα της στήλης μέσα σε quotes** (*single quotes ' ή double quotes "*), **χωριζόμενες με ένα κόμμα**.

Για παράδειγμα, για να ανακτήσουμε το *βαθμό* στην *1^η γραμμή* (όπου είναι δυνατόν σημειώνεται με **κίτρινο** το ζητούμενο), ισχύει ότι έχουμε μάθει για τα μητρώα, αλλά μπορούμε επίσης να πούμε:

```
> students
  names grades gender
1 Maria      10 female
2 Nikos       8  male
3 Thanos     9  male
4 Niki       7 female
> students[1,2]
[1] 10
> students[1, 'grades']
[1] 10
> students[1, "grades"]
[1] 10
```

Αναφορά σε τμήμα πλαισίου δεδομένων

Αναφορά σε στήλες

Για την αναφορά σε στήλη ενός πλαισίου δεδομένων έχουμε 3 τρόπους:

1. Με χρήση του τελεστή `$` και του **ονόματος της στήλης (χωρίς quotes)**
2. Με χρήση του τελεστή `[]` και του **ονόματος της στήλης με quotes**
3. Με χρήση του τελεστή `[]` και του **αριθμού της στήλης**

Με χρήση τελεστή `$`

Η αναφορά σε **στήλη** (μεταβλητή) γίνεται με το όνομα του πλαισίου δεδομένων, τον **τελεστή `$`** (δολάριο) και το **όνομα της στήλης (χωρίς quotes)**.

Όταν βρισκόμαστε μέσα στον κειμενογράφο του RStudio, και γράφουμε το όνομα ενός dataframe βάζοντας δίπλα τον τελεστή `$`, ήδη μας βοηθά παρουσιάζοντας τα ονόματα των στηλών και τον τύπο και τις πρώτες τιμές της στήλης.

```

47
48 names<-c("Maria","Nikos","Thanos","Niki") # διάνυσμα ονομάτων
49 grades<-c(10,8,9,7) # διάνυσμα βαθμών
50 gender<-factor(c("female","male","male","female")) # ποιοτική
51 students<-data.frame(names,grades,gender)
52
53 snames<-students$
54 students[1,] # στήλη
55 students[1:2,] # στήλες
56

```

names	grades	gender
Maria	10	female
Nikos	8	male
Thanos	9	male
Niki	7	female

```

$ names : chr "Ma
$ grades: num 10
$ gender: Factor v
Values
gender      Facto
grades      num [

```

Εικόνα 8: Με τη χρήση του τελεστή \$ στον κειμενογράφο δίπλα στο όνομα ενός dataframe εμφανίζονται τα ονόματα των στηλών (σε μπλέ πλαίσιο), ο τύπος τους και οι πρώτες τιμές.

```

> snames<-students$names # στήλη names
> snames
[1] "Maria" "Nikos" "Thanos" "Niki"

```

Με χρήση του τελεστή [] και του ονόματος της στήλης με quotes

Η αναφορά σε **στήλη** (μεταβλητή) γίνεται με το **όνομα του πλαισίου δεδομένων**, τον **τελεστή []** και το **όνομα της στήλης με quotes**.

Αν χρησιμοποιήσουμε κόμμα (,) για να πάρουμε ολόκληρη τη στήλη, δεν καθορίζουμε γραμμή που σημαίνει όλες οι γραμμές.

```

> students
  names grades gender
1  Maria     10 female
2  Nikos      8  male
3  Thanos     9  male
4   Niki      7 female
> students['names'] # χωρίς κόμμα, επιστρέφει dataframe
  names
1 Maria
2 Nikos
3 Thanos
4  Niki
> students[, 'names'] # με κόμμα, επιστρέφει vector
[1] "Maria" "Nikos" "Thanos" "Niki"
> students[,"names"] # με κόμμα, επιστρέφει vector
[1] "Maria" "Nikos" "Thanos" "Niki"

```

Στην περίπτωση αυτή μπορούμε να ανακτήσουμε **περισσότερες στήλες** θέτοντας τα **ονόματά τους σε διάνυσμα** με τη χρήση της συνάρτησης **c()** ως ακολούθως:

```
> students
  names grades gender
1 Maria     10 female
2 Nikos     8  male
3 Thanos    9  male
4 Niki      7 female
> students[c('names','grades')] # χωρίς κόμμα για γραμμή
  names grades
1 Maria     10
2 Nikos     8
3 Thanos    9
4 Niki      7
> students[,c('names','grades')] # με κόμμα για γραμμή
  names grades
1 Maria     10
2 Nikos     8
3 Thanos    9
4 Niki      7
```

Με χρήση του τελεστή **[]** και του αριθμού της στήλης

Όταν χρησιμοποιούμε τον τελεστή **[]** και τον αριθμό της στήλης **ισχύουν ακριβώς ότι μάθαμε για τα μητρώα**, δηλαδή:

Μία στήλη

```
> students
  names grades gender
1 Maria     10 female
2 Nikos     8  male
3 Thanos    9  male
4 Niki      7 female
> students[1] # στήλη στη θέση 1, χωρίς κόμμα επιστρέφει dataframe
```

```

names
1 Maria
2 Nikos
3 Thanos
4 Niki
> students[,1] # στήλη στη θέση 1, με κόμμα επιστρέφει vector
[1] "Maria" "Nikos" "Thanos" "Niki"

```

Συνεχόμενες στήλες

```

> students
  names grades gender
1 Maria    10 female
2 Nikos     8  male
3 Thanos    9  male
4 Niki     7 female
> students[2:3] # συνεχόμενες στήλες, θέσεις από 2 έως 3
  grades gender
1    10 female
2     8  male
3     9  male
4     7 female

```

Μεμονωμένες στήλες

```

> students
  names grades gender
1 Maria    10 female
2 Nikos     8  male
3 Thanos    9  male
4 Niki     7 female
> students[c(1,3)] # μη συνεχόμενες στήλες, θέσεις 1 και 3
  names gender

```

```
1 Maria female
2 Nikos male
3 Thanos male
4 Niki female
```

Αναφορά σε γραμμές

Η αναφορά σε **γραμμή** ενός πλαισίου δεδομένων γίνεται με τον προσδιορισμό του **αριθμού της γραμμής**. Μπορούμε να αναφερθούμε σε **συγκεκριμένη γραμμή**, σε **συνεχόμενο τμήμα γραμμών** (τελεστής :) ή σε **μεμονωμένες γραμμές** (χρήση διανύσματος c()) όπως ακριβώς μάθαμε στα μητρώα. Ακολουθούν παραδείγματα:

```
> students
  names grades gender
1 Maria    10 female
2 Nikos     8  male
3 Thanos    9  male
4  Niki     7 female

> students[1,] # συγκεκριμένη γραμμή 1
  names grades gender
1 Maria    10 female

> students[1:2,] # συνεχόμενες γραμμές 1 έως 2
  names grades gender
1 Maria    10 female
2 Nikos     8  male

> students[c(1,4),] # μη συνεχόμενες γραμμές 1 και 4
  names grades gender
1 Maria    10 female
4  Niki     7 female
```

Τεμαχισμός (slicing)

Ανάκτηση τεμαχίου

Για την ανάκτηση ενός συγκεκριμένου τμήματος αρκεί να χρησιμοποιήσουμε το όνομα του πλαισίου δεδομένων και να ορίσουμε τις γραμμές και τις στήλες σε οποιοδήποτε παραπάνω συνδυασμό με τη χρήση του τελεστή [] και να ορίσουμε:

```
όνομα_πλαισίου[γραμμές, στήλες]
```

Για παράδειγμα για να πάρουμε το όνομα και τους βαθμούς (συνεχόμενες στήλες 1 έως 2) των φοιτητριών (γραμμές 1 και 4) θα γράψουμε:

```
> students
  names grades gender
1 Maria    10 female
2 Nikos     8  male
3 Thanos    9  male
4 Niki     7 female
> students[c(1,4),c('names','grades')]
  names grades
1 Maria    10
4 Niki     7
> students[c(1,4),1:2]
  names grades
1 Maria    10
4 Niki     7
```

Ανάκτηση τεμαχίου με λογικές συνθήκες (κριτήρια)

Για να πάρουμε ένα τμήμα ενός πλαισίου δεδομένων που πληροί συγκεκριμένα κριτήρια, για παράδειγμα «τα στοιχεία των φοιτητριών» (στη στήλη gender να υπάρχει η τιμή "female"), «τα στοιχεία των φοιτητών/φοιτητριών με βαθμό ≥ 8.5 (στη στήλη βαθμός η τιμή να είναι ≥ 8.5)», κλπ., δημιουργούμε λογικές συνθήκες. Στη συνέχεια για να πάρουμε το τμήμα που ισχύουν τα κριτήρια, χρησιμοποιούμε τον τελεστή [] και **θέτουμε τα λογικά κριτήρια στη θέση των γραμμών**:

όνομα_πλαισίου[λογική_συνθήκη, [στήλες]]

Λογική συνθήκη

Για να σχηματίσουμε ένα απλό κριτήριο χρησιμοποιούμε το **όνομα μίας στήλης** και έναν από τους **τελεστές σύγκρισης**. Το αποτέλεσμα είναι ένα **διάνυσμα με λογικές τιμές** (TRUE, FALSE) που προκύπτει από τη σύγκριση κάθε στοιχείου της στήλης.

Τελεστής	Ερμηνεία	Παράδειγμα συνθήκης	Αποτέλεσμα
==	Ίσο	students\$gender=="female"	[1] TRUE FALSE FALSE TRUE
>	Μεγαλύτερο	students\$grade>8.5	[1] TRUE FALSE TRUE FALSE
<	Μικρότερο	students\$grade>5	[1] TRUE TRUE TRUE TRUE
>=	Μεγαλύτερο ή ίσο	students\$grade>=8.5	[1] TRUE FALSE TRUE FALSE
<=	Μικρότερο ή ίσο	students\$grade<=5	[1] FALSE FALSE FALSE FALSE
!=	Διάφορο	students\$name!="Thanos"	[1] TRUE TRUE FALSE TRUE

Οι λογικές συνθήκες ενώνονται με λογικούς τελεστές.

Λογικοί τελεστές

Στην R οι λογικοί τελεστές είναι:

Τελεστής R	Ερμηνεία	Παράδειγμα συνθήκης	Αποτέλεσμα
&	ΚΑΙ (TRUE αν όλες οι συνθήκες TRUE)	<code>students\$gender=="female" & students\$grades>8</code>	[1] TRUE FALSE FALSE FALSE
	Ή (TRUE αν έστω μία συνθήκη TRUE)	<code>students\$gender=="female" students\$grades>8</code>	[1] TRUE FALSE TRUE TRUE
!	ΌΧΙ (αντίθετο)	<code>!(students\$gender=="female")</code>	[1] FALSE TRUE TRUE FALSE

Για να πάρουμε το τεμάχιο που ισχύει η συνθήκη, χρησιμοποιούμε το **όνομα του dataframe**, τον τελεστή `[]` και στη **θέση του αριθμού γραμμών θέτουμε τη συνθήκη**. Μετά το **κόμμα (,)** ορίζουμε τη **στήλη** που επιθυμούμε. Ακολουθούν ορισμένα παραδείγματα:

Ερώτημα	R κώδικας	Αποτέλεσμα
Τα στοιχεία των φοιτητριών	<code>students[students\$gender=='female',]</code>	<pre>names grades gender 1 Maria 10 female 4 Niki 7 female</pre>
Το όνομα των φοιτητριών με βαθμό ≥ 8.5	<code>students[students\$gender=='female' & students\$grades>=8.5, 'names']</code>	[1] "Maria"
Το όνομα και το γένος των φοιτητών και φοιτητριών με τον υψηλότερο βαθμό	<code>students[students\$grades == max(students\$grades), c('names', 'gender')]</code>	<pre>names gender 1 Maria female</pre>

Για να βρούμε πόσες γραμμές πληρούν μία λογική συνθήκη που περιλαμβάνει στήλη ενός πλαισίου δεδομένων έχουμε δύο δυνατότητες:

- Χρησιμοποιούμε τη συνάρτηση `sum(συνθήκη)`, καθώς η R αναπαριστά όλα τα TRUE ως 1 και τα FALSE ως 0
- Χρησιμοποιούμε τη συνάρτηση `which(συνθήκη)`, που επιστρέφει ένα διάνυσμα με τους αριθμούς των γραμμών που πληρούν τη συνθήκη και σε αυτό χρησιμοποιούμε τη συνάρτηση για το μήκος `length()`, δηλαδή `length(which(συνθήκη))`

Δοκιμάστε τα!

Δημιουργία νέων στηλών σε πλαίσιο δεδομένων

Ένας τρόπος για να ορίσουμε μία νέα γραμμή σε ένα πλαίσιο δεδομένων είναι να ορίσουμε το όνομά της και συνήθως και την τιμή που θέλουμε να περιέχει. Η τιμή της νέας στήλης

μπορεί να είναι μία σταθερά, μία έκφραση από άλλη στήλη, το άθροισμα άλλων στηλών, κλπ.

Για να δημιουργήσουμε μία στήλη ως άθροισμα άλλων στηλών χρησιμοποιούμε τη συνάρτηση `rowSums()` που αθροίζει στήλες ενός πλαισίου δεδομένων.

Τα παραπάνω σε παραδείγματα:

```
> students
  names grades gender
1 Maria     10 female
2 Nikos      8  male
3 Thanos     9  male
4 Niki       7 female
> students$vathmos<-2 # νέα στήλη με τιμές σε όλες τις γραμμές το 2
> students$vathmos
[1] 2 2 2 2
> students$vathmos<-students$grades/2 # νέα στήλη με τιμές σε κάθε γραμμή grades/2
> students$vathmos
[1] 5.0 4.0 4.5 3.5
> students$final<-rowSums(students[c('grades','vathmos')]) #νέα στήλη άθροισμα στηλών
> students$final
[1] 15.0 12.0 13.5 10.5
```

Βασική στατιστική ανάλυση

Η συνάρτηση `summary(όνομα_πλαισίου)` μας δίνει τις βασικές στατιστικές μετρικές ενός πλαισίου. Για παράδειγμα:

```
> students
  names grades gender vathmos final
1 Maria     10 female    5.0  15.0
2 Nikos      8  male     4.0  12.0
3 Thanos     9  male     4.5  13.5
4 Niki       7 female    3.5  10.5
> summary(students)
  names          grades          gender          vathmos          final
Length:4      Min.   : 7.00  female:2      Min.   :3.500      Min.   :10.50
```

Class :character	1st Qu.: 7.75	male :2	1st Qu.:3.875	1st Qu.:11.62
Mode :character	Median : 8.50		Median :4.250	Median :12.75
	Mean : 8.50		Mean :4.250	Mean :12.75
	3rd Qu.: 9.25		3rd Qu.:4.625	3rd Qu.:13.88
	Max. :10.00		Max. :5.000	Max. :15.00

Η συνάρτηση `summary()` σε αριθμητικές μεταβλητές παρουσιάζει τα τεταρτημόρια και τον αριθμητικό μέσο ως εξής:

- Ελάχιστο (Min., 0%)
- 1ο Τεταρτημόριο (1 st Qu., 25%)
- Διάμεσο (Median) (= p50=Q2, 50%)
- Αριθμητικό μέσο (Mean)
- 3ο Τεταρτημόριο (3 rd Qu., 75%)
- Μέγιστο (Max.,100%)

Για τις ποιοτικές μεταβλητές παρουσιάζει τον πίνακα συχνοτήτων (εδώ μεταβλητή `gender`).

Για την στατιστική επεξεργασία μίας στήλης ενός πλαισίου δεδομένων έχουμε τις ακόλουθες επίσης συναρτήσεις:

Ελάχιστο: `min(στήλη_πλασίου_δεδομένων)`

Μέγιστο: `max(στήλη_πλασίου_δεδομένων)`

Μέσος όρος: `mean(στήλη_πλασίου_δεδομένων, na.rm = TRUE)`

Διάμεσος: `median(στήλη_πλασίου_δεδομένων)`

Τυπική απόκλιση δείγματος: `sd(στήλη_πλασίου_δεδομένων)`

Διακύμανση δείγματος: `var(στήλη_πλασίου_δεδομένων)`

Μπορούμε να θέσουμε το όρισμα `na.rm = TRUE` όταν ο υπολογισμός πρέπει να αγνοήσει τις κενές τιμές. Για παράδειγμα ο μέσος όρος γίνεται:

`mean(students$grades, na.rm = TRUE)`

Δοκιμάστε τις στατιστικές συναρτήσεις!

Χρειάζεται προσοχή όταν ζητείται τυπική **απόκλιση** ή **διακύμανση πληθυσμού**.

Δημιουργήστε μόνοι ή αναζητήστε στο Διαδίκτυο τον τύπο για τον πληθυσμό όταν γνωρίζουμε την τυπική απόκλιση ή διακύμανση δείγματος.