R – Μία Στατιστική Γλώσσα Προγραμματισμού

Μανώλης Τζαγκαράκης, Βικτωρία Δασκάλου

When your stats prof starts talking in R commands:

Εισαγωγικά

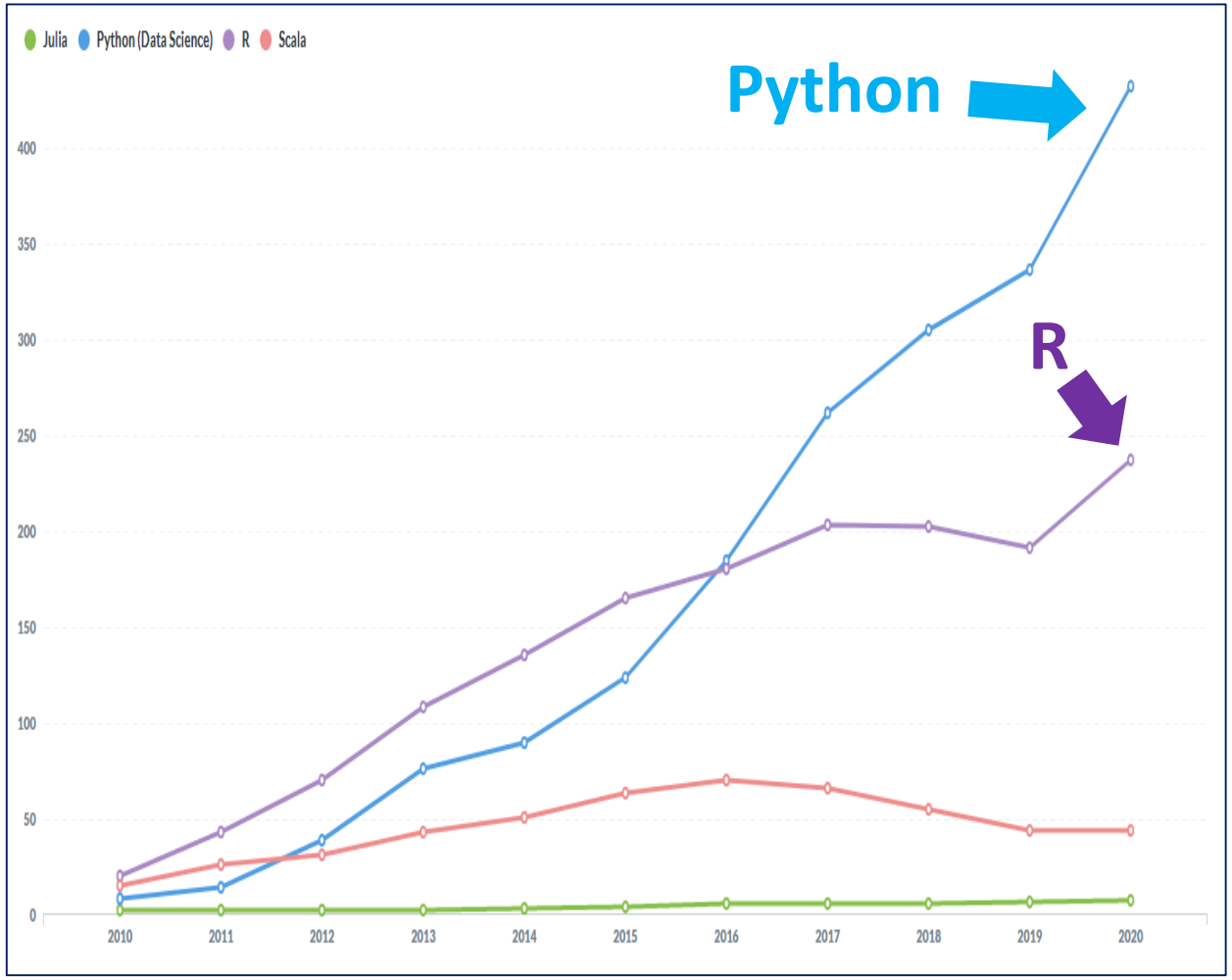
- Τι είναι η R ;
  - Είναι μία γλώσσα προγραμματισμού και περιβάλλον προγραμματισμού, **εξειδικευμένο για υπολογισμούς στατιστικής φύσεως και οπτικοποίησης δεδομένων**.
    - Χαρακτηρίζεται και ως «στατιστική γλώσσα προγραμματισμού»
    - Στατιστική γλώσσα προγραμματισμού; Σημαίνει ότι υποστηρίζει πάρα πολύ καλά όλα τα βήματα στη διαδικασία της στατιστικής ανάλυσης δεδομένων, τα οποία μπορούν να υλοποιηθούν (δλδ να προγραμματιστούν) με ακρίβεια, ευκολία και ταχύτητα.
    - Δηλαδή η R έχει σχεδιαστεί με στόχο να κάνει πολύ εύκολα τα εξής:
      - Την ενσωμάτωση/ανάγνωση δεδομένων
      - Τον καθαρισμό των δεδομένων (προεπεξεργασία) όσο περίπλοκος κι αν είναι αυτός
      - Την εφαρμογή στατιστικών ελέγχων και μεθόδων πάνω στα δεδομένα, και την εξαγωγή συμπερασμάτων
      - Τη δημιουργία και εφαρμογή στατιστικών/οικονομετρικών μοντέλων
      - Την αξιολόγηση των στατιστικών μοντέλων
      - Τη χρήση των στατιστικών μοντέλων για την αντιμετώπιση πραγματικών προβλημάτων

Εισαγωγικά

- Δημοτικότητα της R

Jul 2020	Jul 2019	Change	Programming Language	Ratings	Change
1	2	↑	C	16.45%	+2.24%
2	1	↓	Java	15.10%	+0.04%
3	3		Python	9.09%	-0.17%
4	4		C++	6.21%	-0.49%
5	5		C#	5.25%	+0.88%
6	6		Visual Basic	5.23%	+1.03%
7	7		JavaScript	2.48%	+0.18%
8	20	⇈	R	2.41%	+1.57%
9	8	↓	PHP	1.90%	-0.27%
10	13	↑	Swift	1.43%	+0.31%
11	9	↓	SQL	1.40%	-0.58%
12	16	⇈	Go	1.21%	+0.19%
13	12	↓	Assembly language	0.94%	-0.45%
14	19	⇈	Perl	0.87%	-0.04%
15	14	↓	MATLAB	0.84%	-0.24%
16	11	⇊	Ruby	0.81%	-0.83%
17	30	⇈	Scratch	0.72%	+0.35%
18	33	⇈	Rust	0.70%	+0.36%
19	23	⇈	PL/SQL	0.68%	-0.01%
20	17	↓	Classic Visual Basic	0.66%	-0.35%

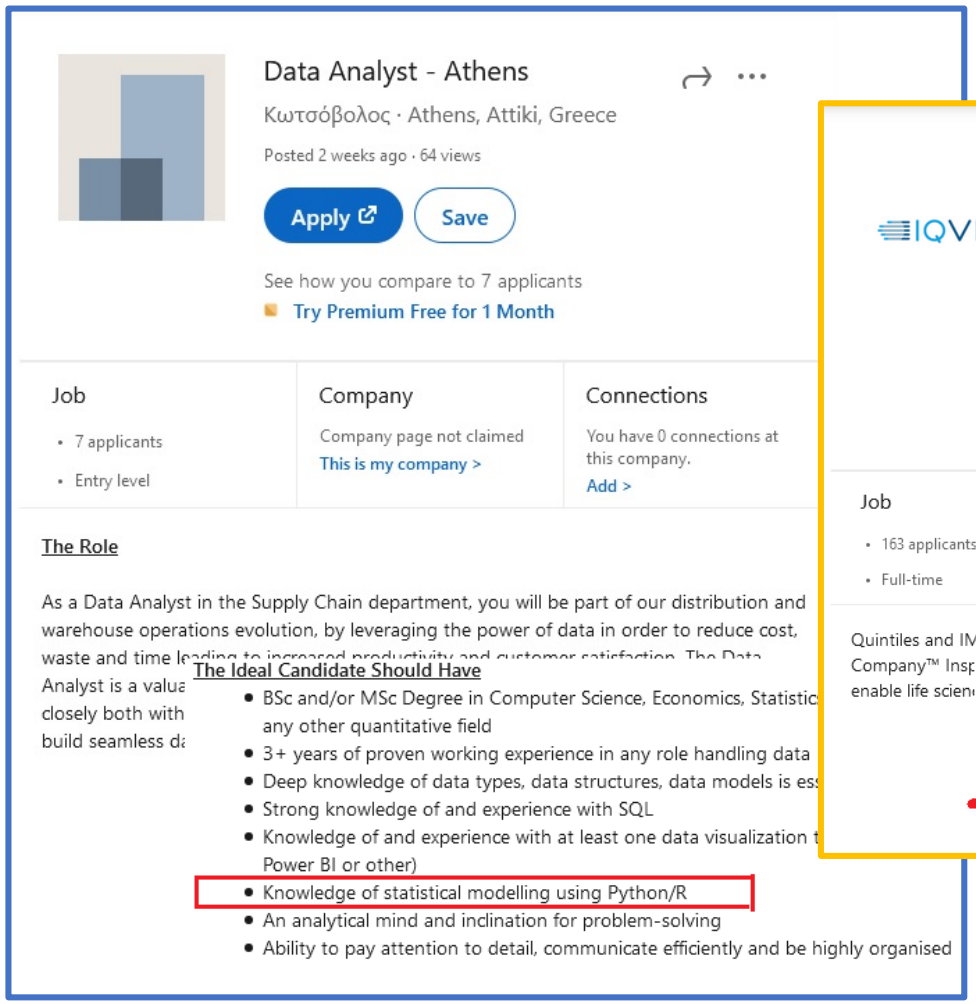
Μεταξύ όλων των γλωσσών προγραμματισμού (2020)

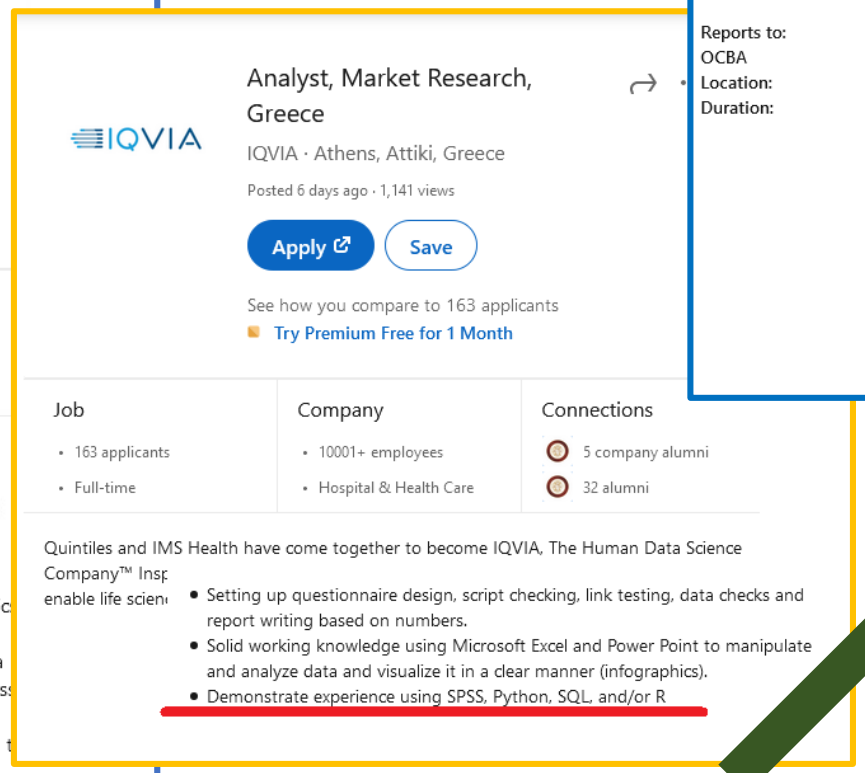


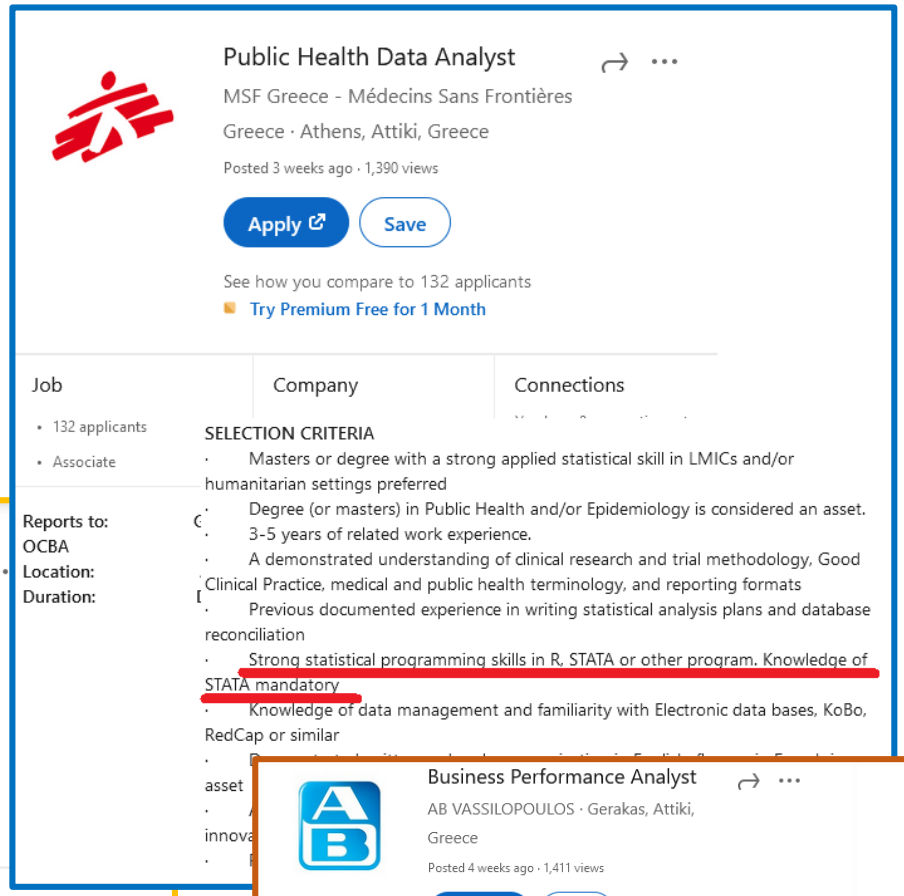
Μεταξύ γλωσσών προγραμματισμού στην επιστήμη δεδομένων(2020)

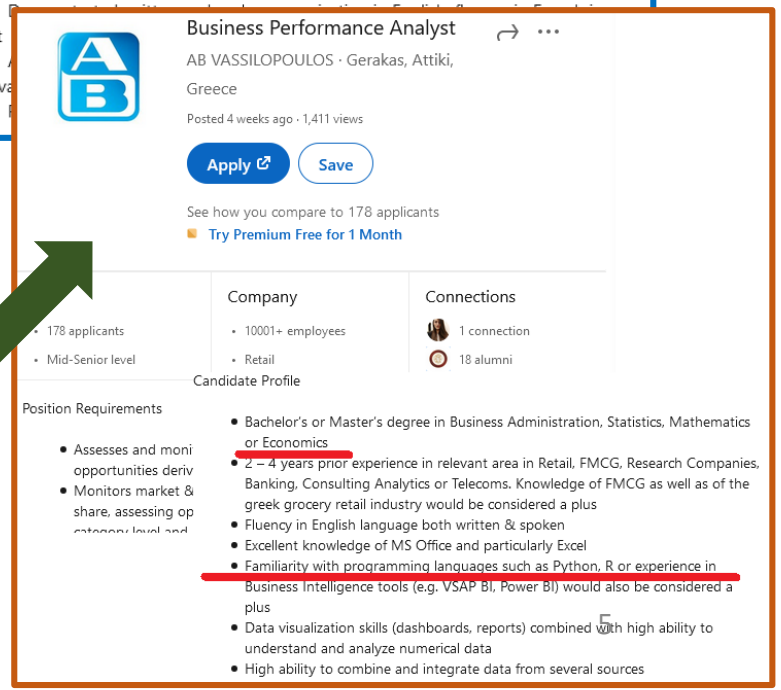
Εισαγωγικά

- Μία πρόχειρη ματιά στην αγορά εργασίας (LinkedIn)









Super Market. W-T-F?

Εισαγωγικά

- Γιατί R ;
  - Η R είναι **ανοικτό λογισμικό και δωρεάν στη χρήση της από οποιονδήποτε** (ιδιώτη ή επιχείρηση). Άλλα τέτοια περιβάλλοντα κοστίζουν χιλιάδες ευρώ.
  - Αποτελεί μία **ολοκληρωμένη και πλήρης στατιστική πλατφόρμα** υποστηρίζοντας όλες τις **τεχνικές και μεθόδους ανάλυσης δεδομένων**. Οποιαδήποτε (οποιαδήποτε) στατιστική ανάλυση δεδομένων μπορεί να γίνει με την R.
  - Δεν υπάρχει **αλγόριθμος/μέθοδος/στατιστικός έλεγχος** που να **μην υπάρχει διαθέσιμος στο περιβάλλον της R**.
  - Επιτρέπει τη **δημιουργία επαγγελματικών και πολύπλοκων γραφημάτων**, που δεν μπορούν να γίνουν με άλλα περιβάλλοντα ή γλώσσες.
  - Η R έχει **σχεδιαστεί με στόχο την υποστήριξη επεξεργασίας, εξερεύνησης και ανάλυσης δεδομένων**.
    - Δεν είναι μία γλώσσα γενικού σκοπού όπου έχει «προστεθεί»/«φορεθεί» από πάνω η υποστήριξη για την ανάλυση δεδομένων.
    - Παρέχει μηχανισμούς για την εύκολη διασύνδεση εντολών επεξεργασίας/ανάλυσης δεδομένων και δημιουργία ροών επεξεργασιών (pipes/σωλήνες)

Εισαγωγικά

- Γιατί R ;
  - Μπορεί με **ευκολία να ενσωματώσει και να χρησιμοποιήσει δεδομένα που υπάρχουν σε εξωτερικές (τρίτες) πηγές** που μπορούν να είναι σε οποιαδήποτε μορφή (όπως αρχεία κειμένου, αρχεία Excel, βάσεις δεδομένων κλπ)
  - Παρέχει ένα **σύστημα βιβλιοθηκών (packages) που επιτρέπει την υλοποίηση και εγκατάσταση νέων συναρτήσεων που υποστηρίζουν νέους τρόπους ανάλυσης δεδομένων**. Μέσω του συστήματος βιβλιοθηκών, νέοι τρόποι ανάλυσης μπορούν εύκολα να υλοποιηθούν, να διανεμηθούν και να εγκατασταθούν.
    - Οποιοσδήποτε μπορεί να αναπτύξει νέα βιβλιοθήκη, η οποία περιέχει συναρτήσεις για την εκτέλεση των στατιστικών μεθόδων.
  - Η R υπάρχει **διαθέσιμη για όλα τα υπολογιστικά περιβάλλοντα και λειτουργικά συστήματα** (Windows, Linux, MacOS κλπ - διαθέσιμη ακόμη και για iPhone).
    - Πρόγραμμα R που συγγράφεται σε ένα υπολογιστικό περιβάλλον, τρέχει σε όλα τα περιβάλλοντα.

Εισαγωγικά

- Python vs R

Python	R
Γλώσσα προγραμματισμού γενικού σκοπού όπου προστέθηκαν δυνατότητες στατιστικής ανάλυσης δεδομένων.	Γλώσσα προγραμματισμού που σχεδιάστηκε με στόχο την στατιστική ανάλυση δεδομένων.
Παρέχει μεγάλο εύρος βιβλιοθηκών για την υποστήριξη επεξεργασίας δεδομένων.	Οι βασικές μέθοδοι στατιστικής επεξεργασίας δεδομένων παρέχονται από τη γλώσσα (πολλοί τρόποι επεξεργασίας δεν απαιτούν τρίτες βιβλιοθήκες).
Βολική εάν οι μέθοδοι στατιστικής ανάλυσης δεδομένων πρέπει να ενσωματωθούν σε άλλα συστήματα/προγράμματα (π.χ. ιστοσελίδες).	Βολική εάν οι μέθοδοι στατιστικής ανάλυσης δεν απαιτείται να ενσωματωθούν με άλλα συστήματα. Standalone computing.
Χρησιμοποιείται στις πολυτεχνικές σχολές και σχολές θετικών επιστημών	Χρησιμοποιείται κυρίως στις ανθρωπιστικές επιστήμες

Εισαγωγικά

- «*Ρε δάσκαλε, τα ίδια μας τα έλεγες για την Python στο 1ο έτος. Γιατί πρέπει να μάθουμε και R τώρα???? Δεν μας λυπάσαι? Όλο αυτή η δουλειά θα γίνεται τώρα???*»
  - Εύλογη η παρατήρηση του συμφοιτητή σας. Τί έχουμε να απαντήσουμε;
    - Δυστυχώς ή ευτυχώς **όλες οι γλώσσες προγραμματισμού ανάλυσης και στατιστικής επεξεργασίας δεδομένων έχουν δυνατά και αδύνατα σημεία**. **Καμία γλώσσα δεν είναι καλή/βολική για όλες τις «δουλειές»**. No silver bullet.
      - Ακόμη και το MS Excel είναι το καλύτερο/πιο βολικό εργαλείο για κάποιες εργασίες. Όχι για όλες.
    - Στόχος των προπτυχιακών σπουδών είναι **να αποκτήσουν οι φοιτητές μία γεύση απ'όλα τα δημοφιλή εργαλεία στατιστικής επεξεργασίας δεδομένων** που χρησιμοποιούνται σήμερα. Αυτό ώστε να μπορούν να αποκτήσουν εμπειρία για τις διαφορετικές προσεγγίσεις που κάνουν χρήση τέτοια εργαλεία.
      - Να αποκτήσουν μία αίσθηση για το πως δουλεύουν τα εργαλεία και να επιλέξουν εκείνο που τους αρέσει περισσότερο αργότερα.

R - Εγκατάσταση

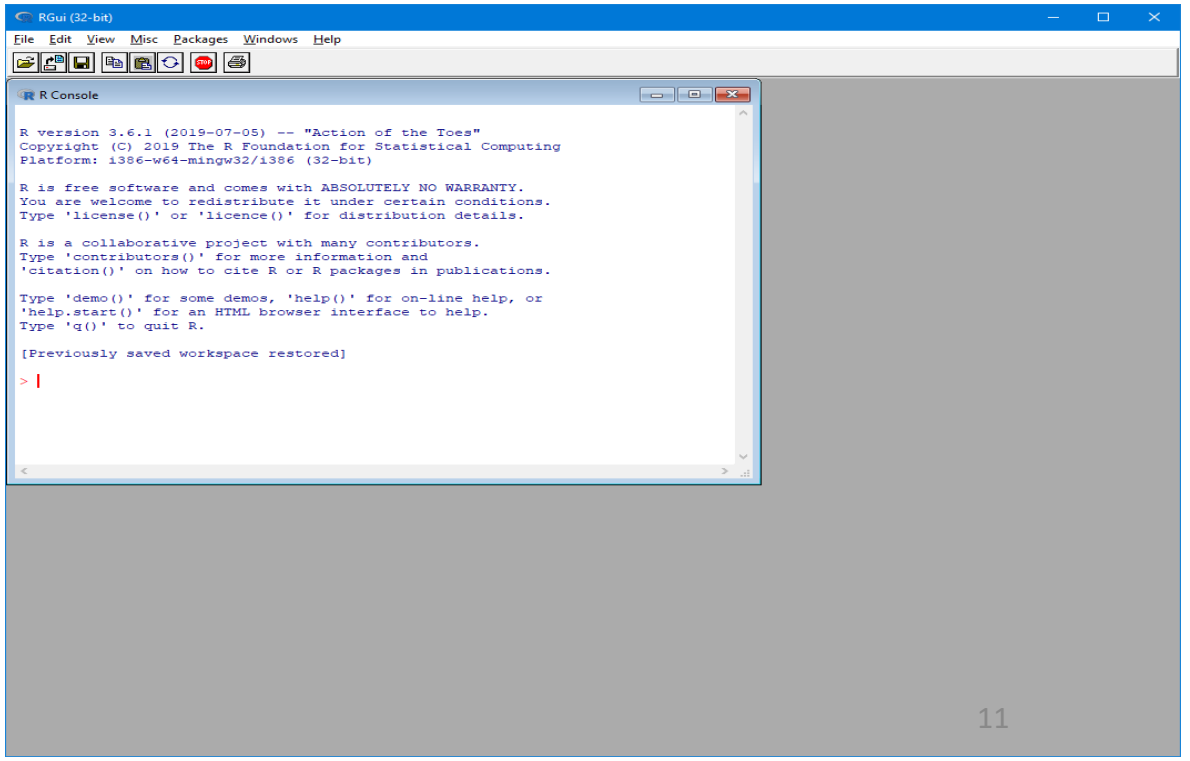
Εγκατάσταση

- Κατέβασμα κατάλληλου αρχείου εγκατάστασης **R** από τις ακόλουθες διευθύνσεις και εκτέλεση του αρχείου:
  - https://cran.r-project.org/bin/windows/base/ (Windows)
  - https://cran.r-project.org/bin/macosx/ (MacOS X)
  - https://cran.r-project.org/bin/linux/ (Linux flavors)

Μετά την επιτυχή εγκατάσταση εμφανίζεται το εικονίδιο,  που εάν εκτελεστεί εμφανίζεται το περιβάλλον της R.

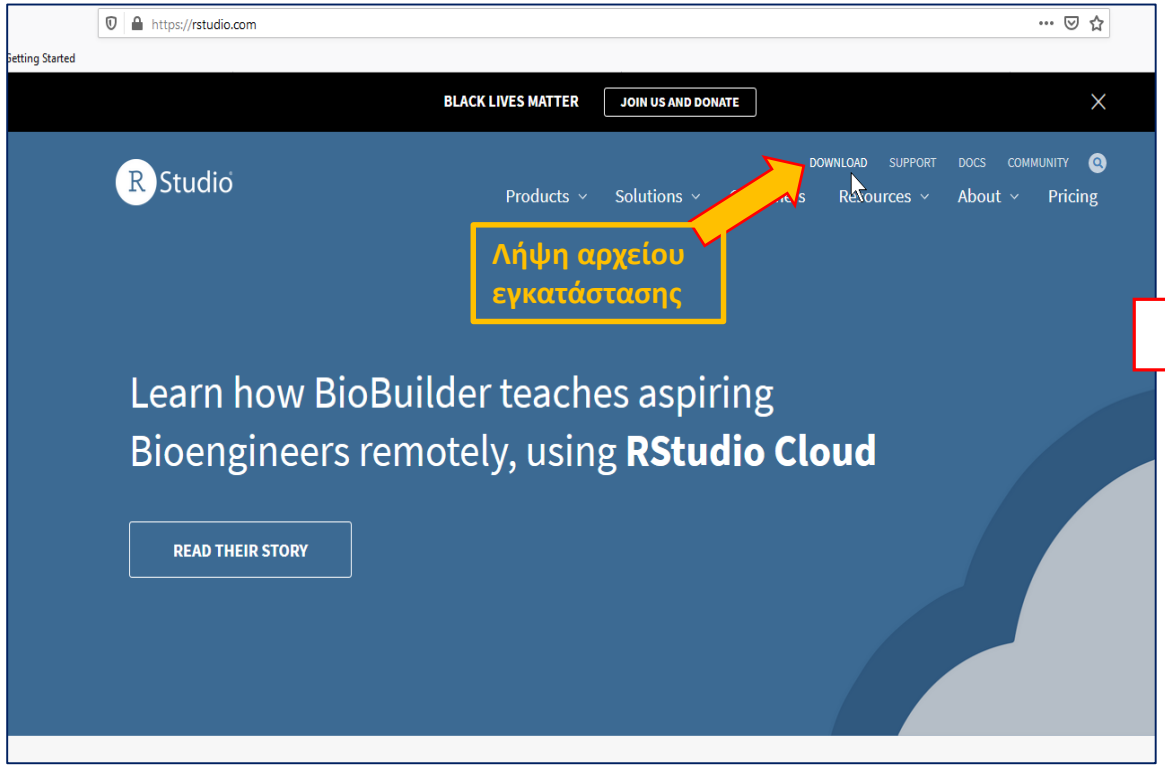
ΠΡΟΣΟΧΗ: Το προκαθορισμένο περιβάλλον της R δεν είναι γραφικό. Παρέχει κονσόλα για τη συγγραφή και εκτέλεση των προγραμμάτων R.

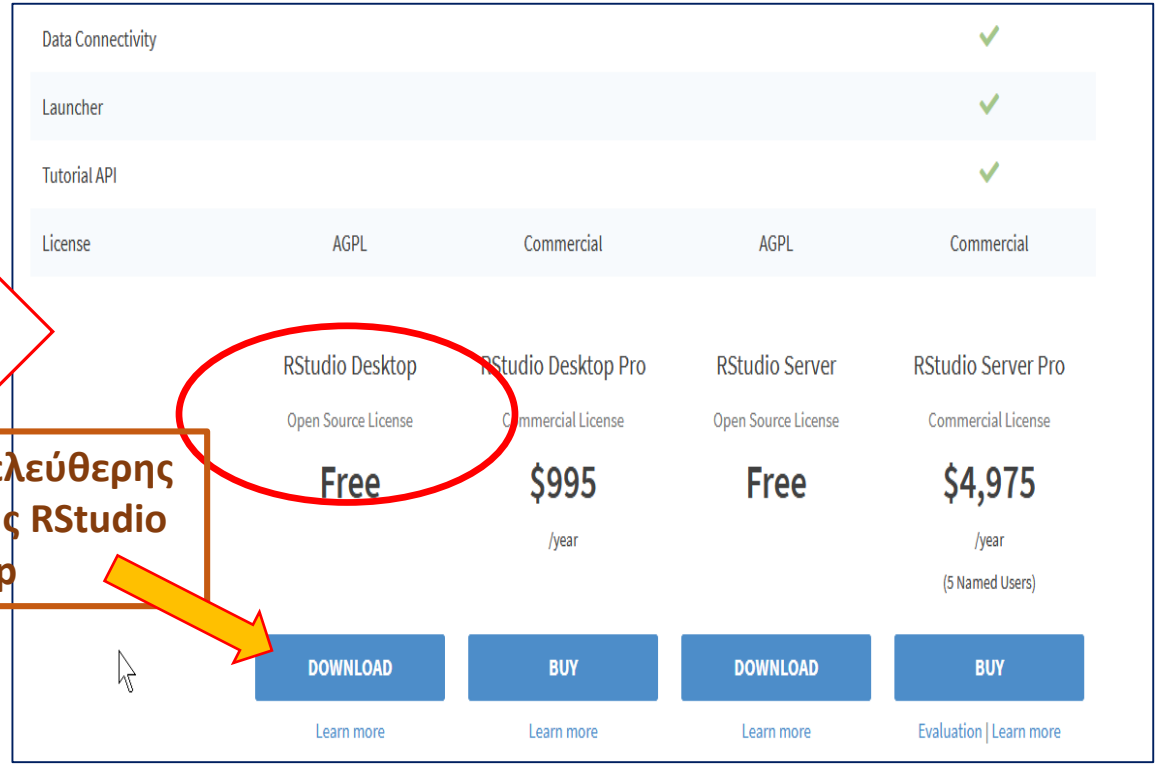
Υπάρχουν άλλα εργαλεία που μπορούν να εγκατασταθούν μετά, για την παροχή μιας πιο φιλικής και βολικής γραφικής διεπαφής αλληλεπίδρασης. Μία τέτοια είναι το **RStudio** .



Εγκατάσταση

- **RStudio**
  - Γραφικό περιβάλλον συγγραφής και εκτέλεσης R προγραμμάτων
  - Εγκατάσταση από http://www.rstudio.com/





Εγκατάσταση

- Εν κατακλείδι, τα βήματα εγκατάστασης και χρήσης (αυστηρά με αυτή τη σειρά!):
  1. Εγκατάσταση R από https://cran.r-project.org/
  2. Εγκατάσταση RStudio από http://www.rstudio.com/
  3. Εκτέλεση RStudio για την εμφάνιση του περιβάλλοντος συγγραφής και εκτέλεσης προγραμμάτων (σεναρίων) R
  4. Συγγραφή και εκτέλεση R προγραμμάτων (ή μάλλον σεναρίων/script) με εκτέλεση του περιβάλλοντος RStudio

Εικονίδιο RStudio στην επιφάνεια εργασίας του υπολογιστή μετά την επιτυχή εγκατάσταση.

Εγκατάσταση

- Μία πρώτη ματιά στο **RStudio**

Εικονίδιο RStudio στην επιφάνεια εργασίας του υπολογιστή μετά την εγκατάσταση. Εκτελέστε αυτό για την έναρξη του περιβάλλοντος.

Κειμενογράφος. Περιοχή συγγραφής προγράμματος (σεναρίου/script) R

Εκτέλεση εντολής (ή εντολών) προγράμματος/σερναρίου R.

Λίστα μεταβλητών που έχουν οριστεί στα προγράμματα και οι τιμές αυτών.

Κονσόλα της R. Επιτρέπει την πληκτρολόγηση και εκτέλεση μεμονωμένων εντολών της R από το προτρεπτικό και την εμφάνιση των αποτελεσμάτων εντολών.

R – Μία Στατιστική Γλώσσα Προγραμματισμού

Μανώλης Τζαγκαράκης, Βικτωρία Δασκάλου

Η γλώσσα R

Η γλώσσα R

- Η R εμφανίστηκε το 1993 και αποτέλεσε μια υλοποίηση της γλώσσας S στην οποία βασίστηκε.
  - Και η S ήταν στατιστική γλώσσα προγραμματισμού (από τα Bell-Labs)
- Η R σχεδιάστηκε και υλοποιήθηκε από τους Ross Ihaka και Robert Gentleman, καθηγητές στατιστικής στη Νέα Ζηλανδία
  - **Σχεδιάστηκε ως στατιστική γλώσσα προγραμματισμού** για την υποστήριξη των μαθημάτων και έρευνας – όχι ως γλώσσα προγραμματισμού γενικού σκοπού

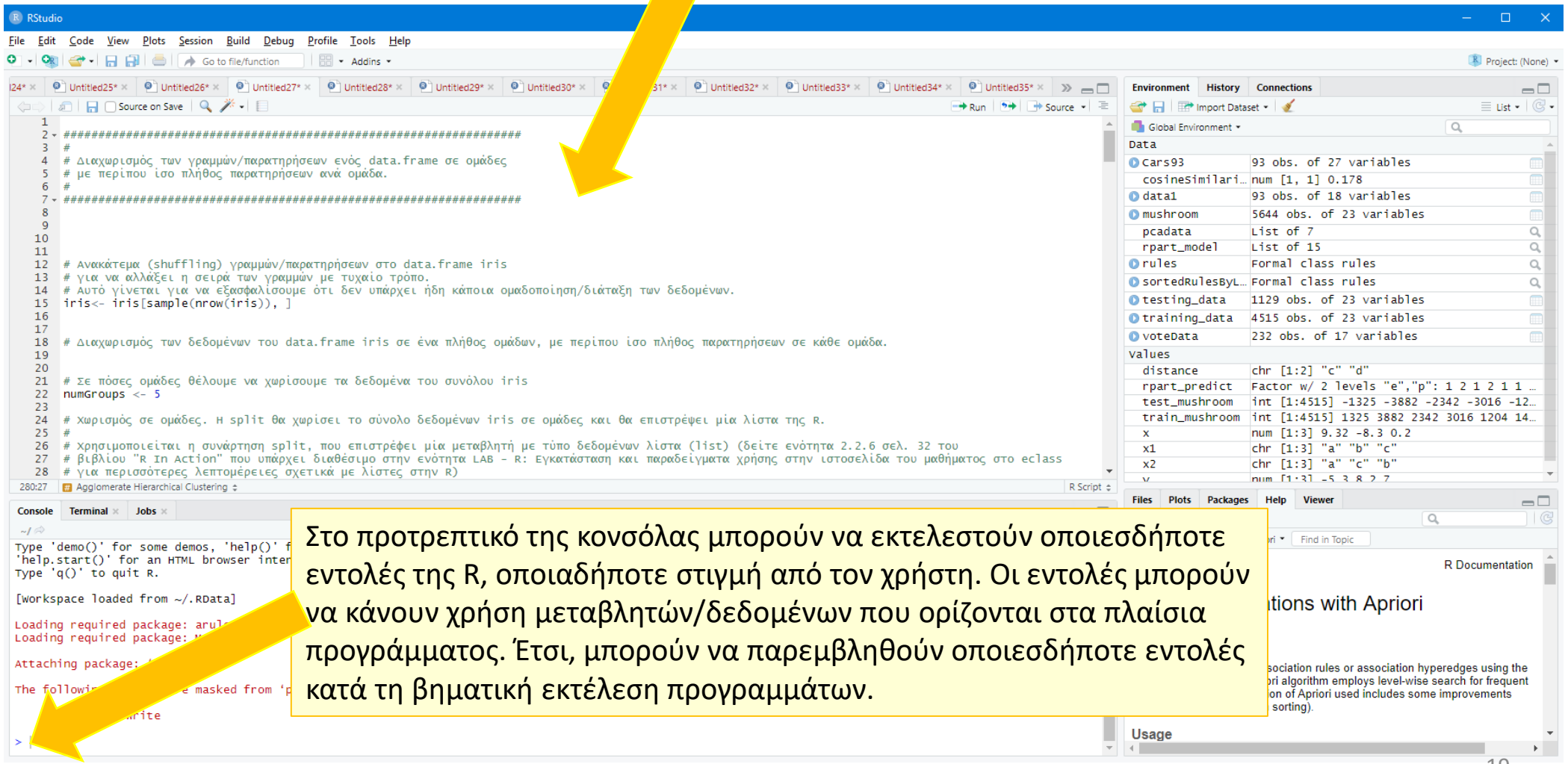
Η γλώσσα R

- Είναι διερμηνευόμενη γλώσσα (interpreted language) – όπως και η Python
  - Οι **εντολές της R** που πληκτρολογεί ο χρήστης (ή υπάρχουν σε ένα πρόγραμμα) εκτελούνται από ένα ειδικό πρόγραμμα, τον διερμηνευτή της R (R Interpreter)
    - Ο διερμηνευτής διαβάζει τις εντολές τις R και τις εκτελεί μία-μία
    - Αυτό σε αντίθεση με μεταγλωττισμένες γλώσσες προγραμματισμού, οι εντολές των οποίων δεν εκτελούνται από διερμηνευτή (εκτελούνται κατ'ευθείαν από τον επεξεργαστή)
  - Επειδή είναι διερμηνευόμενη γλώσσα:
    - οι **εντολές** που πληκτρολογεί ο χρήστης **μπορούν να εκτελεστούν κατ' ευθείαν και δίχως καθυστέρηση** – μπορούν να εκτελεστούν εκτός προγράμματος με δεδομένα που χρησιμοποιεί το πρόγραμμα
    - επιτρέπει την **σταδιακή και βηματική ανάπτυξη και εκτέλεση του προγράμματος σε R** με το ρυθμό που θέλει ο χρήστης.
    - **πολύ ευέλικτη** στην συγγραφή και εκτέλεση προγράμματος
    - …αλλά **πιο αργή η εκτέλεση προγραμμάτων** σε σχέση με μεταγλωττισμένες γλώσσες

Η γλώσσα R

- Ως **διερμηνευόμενη γλώσσα…**

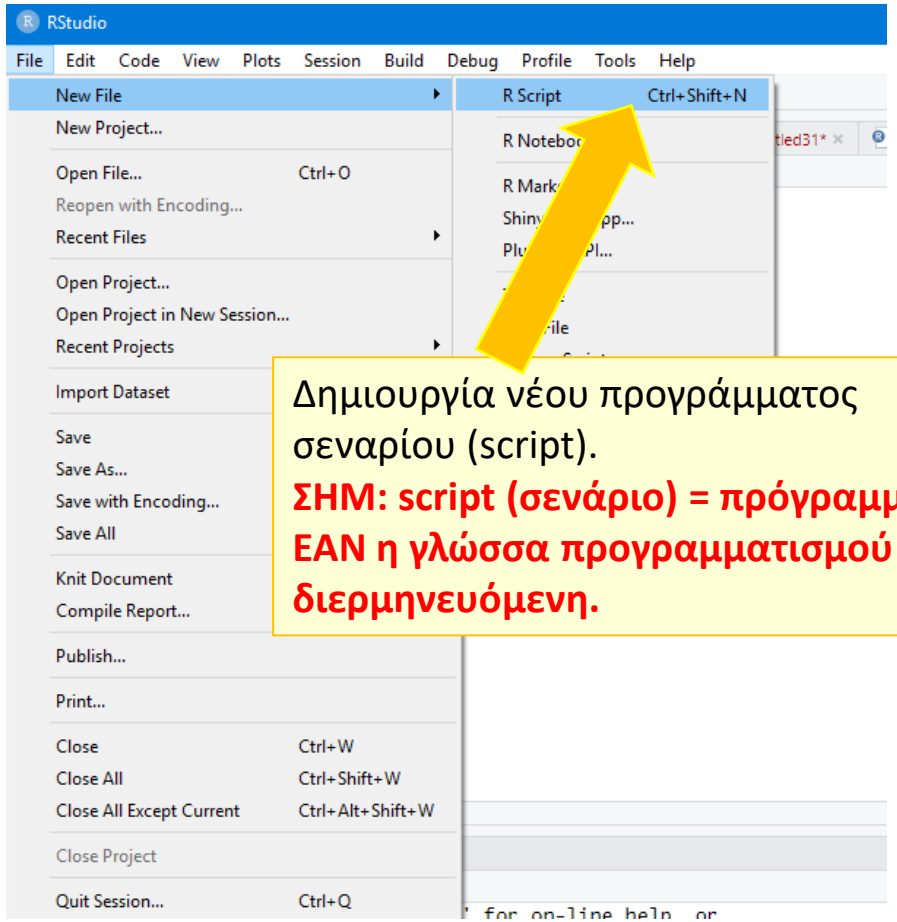
Εντολές προγράμματος (σεναρίου) R. Οι εντολές μπορούν να εκτελεστούν όλες μαζί, ή βηματικά η μία μετά την άλλη, με ρυθμό που επιθυμεί ο χρήστης



Στο προτρεπτικό της κονσόλας μπορούν να εκτελεστούν οποιεσδήποτε εντολές της R, οποιαδήποτε στιγμή από τον χρήστη. Οι εντολές μπορούν να κάνουν χρήση μεταβλητών/δεδομένων που ορίζονται στα πλαίσια προγράμματος. Έτσι, μπορούν να παρεμβληθούν οποιεσδήποτε εντολές κατά τη βηματική εκτέλεση προγραμμάτων.

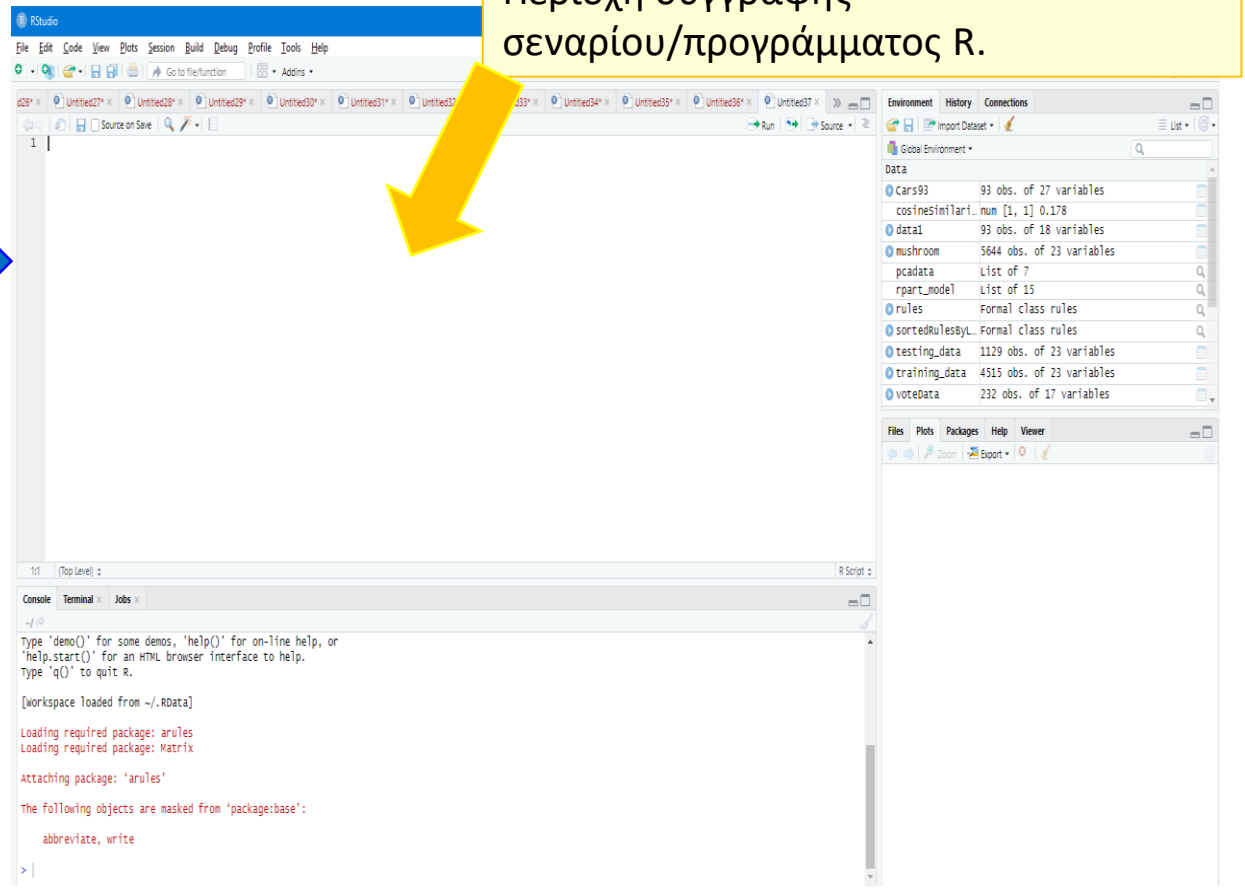
Η γλώσσα R

- Συγγραφή προγράμματος/σεναρίου στο RStudio



Δημιουργία νέου προγράμματος σεναρίου (script).

ΣΗΜ: script (σενάριο) = πρόγραμμα ΕΑΝ η γλώσσα προγραμματισμού είναι διερμηνευόμενη.



Μεταβλητές και Τύποι Δεδομένων στην R

Μεταβλητές και Τύποι Δεδομένων

- Η R παρέχει μεταβλητές, που είναι ένας **τρόπος αποθήκευσης στη μνήμη του υπολογιστή τιμής** η οποία μπορεί να είναι διαφορετικού είδους
  - Είδος τιμών: τύπος δεδομένων (data type)
- Όλες οι μεταβλητές στην R έχουν όνομα που επιλέγεται από τον χρήστη
  - Ονόματα μεταβλητών πρέπει να ακολουθούν κανόνες
    I. Έγκυροι χαρακτήρες σε όνομα μεταβλητής είναι: γράμματα, αριθμοί, τελεία (.) και κάτω παύλα (_) .
    II. Ονόματα δεν μπορούν να ξεκινούν με αριθμό
    III. Μπορούν να ξεκινούν με τελεία, εφόσον ο επόμενος χαρακτήρας δεν είναι αριθμός.

Όνομα μεταβλητής	Έγκυρο/μη έγκυρο
salary	Έγκυρο όνομα μεταβλητής
Employee-salary	Μη έγκυρο όνομα μεταβλητής. Παραβιάζει κανόνα I. αφού περιέχει μη έγκυρο χαρακτήρα - (hyphen)
2bad	Μη έγκυρο όνομα μεταβλητής. Παραβιάζει κανόνα II.
.lmResult	Έγκυρο όνομα μεταβλητής. Όμως το .5lmResult είναι μη έγκυρο αφού παραβιάζει κανόνα III.

Μεταβλητές και Τύποι Δεδομένων

- Ανάθεση τιμής σε μεταβλητή
  - Με τους ειδικούς τελεστές ανάθεσης **<-** ή **=**
    - Οι τελεστές αυτοί ανάθεσης είναι στις περισσότερες περιπτώσεις ισοδύναμοι (όχι πάντα όμως)
    - Η χρήση του τελεστή **<-** πολύ πιο συχνή. Αυτό για λόγους προς τα πίσω συμβατότητας.
      - Κι εδώ θα υιοθετηθεί η χρήση του τελεστή **<-**

Οι εντολές/εκφράσεις που δίνονται ως παραδείγματα μπορούν να πληκτρολογηθούν είτε στην κονσόλα του RStudio και να εκτελεστούν ή μπορεί να είναι μέρος προγράμματος/σεναρίου R.

```
> age <- 19
> result <- 3*5 - 42
> name <- “Jim”
> numberSequence <-  1:42
```

Δημιουργία μεταβλητής με όνομα age και ανάθεση σε αυτήν της τιμής 19 με τον τελεστή <-

Δημιουργία μεταβλητής με όνομα result και ανάθεση σε αυτήν το αποτέλεσμα αριθμητικής πράξης (πράξεων).

Δημιουργία μεταβλητής με όνομα name και ανάθεση σε αυτήν της τιμής Jim.

Δημιουργία μεταβλητής με όνομα numberSequence και ανάθεση σε αυτήν ενός διανύσματος, που περιέχει τιμές από το 1 έως και το 42.

Μεταβλητές και Τύποι Δεδομένων

- Εγγενώς υποστηριζόμενοι τύποι δεδομένων στην R
  - **Τύπος δεδομένων (data type)** ? Περιορισμοί και κανόνες που διέπουν τις τιμές και τη μορφή τους, που αποθηκεύονται σε μία μεταβλητή.
  - Όλες οι μεταβλητές έχουν τύπο δεδομένων. Λέγεται ότι μία "*μεταβλητή έχει τύπο δεδομένων XXX*"
    - **ΣΗΜΑΝΤΙΚΟ!** Ο τύπος δεδομένων μιας μεταβλητής καθορίζει και το είδος των πράξεων που μπορεί να γίνει πάνω σε αυτή.

Βασικοί τύποι δεδομένων (basic data types)

Σύνθετοι τύποι δεδομένων – δομές δεδομένων (data structures)

Εγγενώς υποστηριζόμενοι τύποι δεδομένων στην R	
character (συμβολοσειρά)	**integer** (ακέραιος)
numeric (πραγματικός αριθμός)	**logical** (λογική τιμή, TRUE/FALSE)
complex (μιγαδικός αριθμός)	
vector (διάνυσμα)	**list** (λίστα)
matrix (πίνακας/μήτρα 2 το πολύ διαστάσεων)	**factor** (παράγοντας aka κατηγορική τιμή)
array (πίνακας >2 διαστάσεων)	**data frame** (πλαίσιο δεδομένων)

Σε σύγκριση με τύπους δεδομένων της Python (όπως list, tuples, dictionaries) οι τύποι δεδομένων της R είναι εμπνευσμένοι από τα μαθηματικά/στατιστική.

Μεταβλητές και Τύποι Δεδομένων: Βασικοί τύποι

- character
  - Συμβολοσειρές, αλληλουχία από οποιουσδήποτε αριθμού, οποιονδήποτε χαρακτήρων.
  - Κυριολέκτημα συμβολοσειρά εσωκλείεται μεταξύ “ “ ή ‘ ‘

```
> name <- “James”
> surname <- ‘Bond’
```

- integer
  - Τιμές θετικοί ή αρνητικοί. Απαιτεί ένα **L** στο τέλος της ακέραιας τιμής προκειμένου να μην νοηθεί η τιμή ως πραγματική.

```
> age <- 35L
> temperature <- -7L
```

Ανάθεση αρνητικής ακέραιας τιμής. Προσοχή: ισοδύναμη έκφραση είναι temperature<--7L

- numeric
  - Πραγματικοί αριθμοί

```
> width <- 4.5
> temperature <- -7.4
> age <- 42
```

Επειδή ΔΕΝ ακολουθεί L μετά την τιμή, θα ερμηνευθεί ώς πραγματική τιμή

Μεταβλητές και Τύποι Δεδομένων: Βασικοί τύποι

- logical
  - Λογικές/δυαδικές τιμές True/False

```
> male <- TRUE
> female <- FALSE
```

- complex
  - Μιγαδικοί αριθμοί

```
> z <- 3+5i
> k <- -7+8i
```

Ανάθεση μιγαδικής τιμής.

Μεταβλητές και Τύποι Δεδομένων: Εμφάνιση τύπου

- Εμφάνιση τύπου δεδομένων μιας μεταβλητής
  - Με χρήση των εντολών **class()** ή/και **typeof()** δίνοντας ως όρισμα το όνομα μεταβλητής

```
> name <- “James”
> class( name )
[1] "character"
> age <- 35L
> class(age)
[1] "integer"
> width <- 4.5
> class(width)
[1] "numeric"
> male <- FALSE
> class(male)
[1] "logical"
>z <- 3+9i
> class(z)
[1] “complex"
```

Μεταβλητές και Τύποι Δεδομένων: Εκτύπωση

- Εμφάνιση τιμής μεταβλητής
  - Από την κονσόλα, πληκτρολογώντας το όνομα μεταβλητής

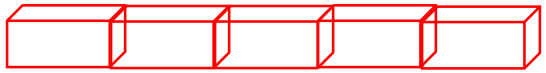
```
> male <- FALSE
> male
[1] FALSE
```

  - Μέσα από πρόγραμμα/σενάριο R, με την εντολή **print()** δίνοντας ως όρισμα το όνομα μεταβλητής

```
1  age <- 42L
2  print(age)
```

```
>print(age)
[1] 42
```

Μεταβλητές και Τύποι Δεδομένων: Vector

- vector (διάνυσμα) 
  - Διατεταγμένο σύνολο τιμών, όπου **όλες οι τιμές πρέπει να έχουν τον ίδιο τύπο δεδομένων**.
    - Διατηρεί πολλές τιμές, διατεταγμένες (σε σειρά)
    - Οι τιμές πρέπει να είναι του ίδιου τύπου δεδομένων όπως character, integer, logical, complex.
  - Δημιουργία με την συνάρτηση **c()** που δημιουργεί νέο διάνυσμα, με στοιχεία τις τιμές που δίνονται ως όρισμα χωρισμένα με κόμμα

```
> aVector <- c(1, 2, 3, -6, 42, 9)
> aVector
[1] 1 2 3 -6 42 9
> anotherVector <- c("Phineas", "Ferb", "Candace", "Doofenschmirtz")
>anotherVector
[1] "Phineas" "Ferb" "Candace" "Doofenschmirtz“
>class(anotherVector)
[1] "character"
```

Δημιουργία μεταβλητής aVector που είναι διάνυσμα το οποίο διάνυσμα έχει 6 τιμές και όλες τους είναι τύπου δεδομένων numeric. Τιμές διαχωρίζονται με κόμμα (,).

Διάνυσμα όπου όλες οι τιμές του είναι συμβολοσειρές.

Χρήση της class() σε μεταβλητή σύνθετου τύπου δεδομένων επιστρέφει τον τύπο δεδομένων των τιμών που περιέχει.

Μεταβλητές και Τύποι Δεδομένων: Vector

- vector(διάνυσμα)
  - Συνάρτηση **length()** με όρισμα μεταβλητή που είναι διάνυσμα για την εμφάνιση του πλήθους των στοιχείων που περιέχει το διάνυσμα
  - Πρόσβαση στα μεμονωμένα στοιχεία του διανύσματος με τον τελεστή **[]** και προσδιορίζοντας τη θέση
    - ΠΡΟΣΟΧΗ! **Στην R η αρίθμηση ξεκινά από το 1 (ένα)!**

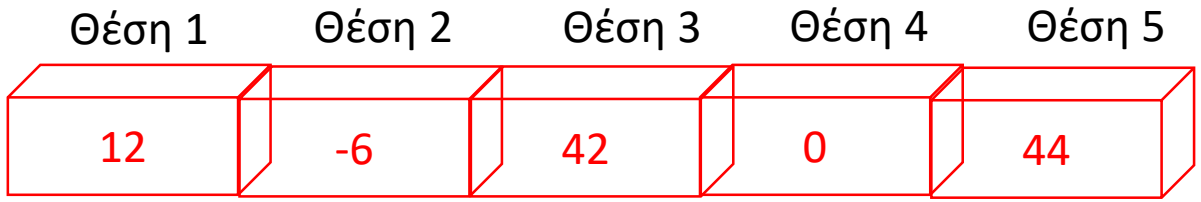
```
> anotherVector <- c("Phineas", "Ferb", "Candace", "Doofenschmirtz")
>length(anotherVector)
[1] 4
>anotherVector[1]
[1] “Phineas"
```

Πλήθος στοιχείων/τιμών διανύσματος anotherVector

Πρώτο στοιχείο διανύσματος anotherVector

Μεταβλητές και Τύποι Δεδομένων: Vector

- vector (διάνυσμα)




```
> intVector <- c(12, -6, 42, 0, 44)
>intVector[3]
[1] 42
>intVector[4] <- -77
>intVector
[1] 12   -6   42   -77   44
```

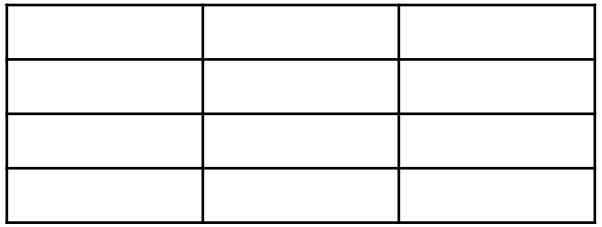
Δημιουργία διανύσματος με 5 ακέραιους αριθμούς

Αναφορά στο στοιχείο στη θέση 3

Τροποποίηση τιμής στη θέση 4, από την τιμή 0 στην τιμή -77

Εμφάνιση διανύσματος intVector

Μεταβλητές και Τύποι Δεδομένων



- matrix (πίνακας δύο διαστάσεων)
  - Δισδιάστατη αποθήκευση τιμών, όπου κι εδώ **όλες οι τιμές πρέπει να έχουν τον ίδιο τύπο δεδομένων**.
    - Οι τιμές που περιέχει ο πίνακας πρέπει να είναι του ίδιου τύπου δεδομένων όπως character, integer, logical, complex.
    - Δημιουργία πίνακα 2 διαστάσεων με χρήση της matrix()

```
> aMatrix <- matrix(nrow = 3, ncol = 3)
> aMatrix
     [,1] [,2] [,3]
[1,]   NA   NA   NA
[2,]   NA   NA   NA
[3,]   NA   NA   NA
> aMatrix <- matrix(1:9, nrow = 3, ncol = 3)
>aMatrix
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

Δημιουργία μεταβλητής aMatrix που είναι μήτρα/πίνακας/matrix με 3 γραμμές και 3 στήλες. Επειδή δεν δίνονται δεδομένα, το περιεχόμενο του πίνακα είναι απροσδιόριστο.

Εμφάνιση πίνακα που έχει δημιουργηθεί. Επειδή ΔΕΝ έχουν δοθεί δεδομένα, το περιεχόμενο του πίνακα εμφανίζει τις τιμές NA, που σημαίνει “Not Available” και σηματοδοτεί απουσία τιμής

Δημιουργία μεταβλητής aMatrix που είναι μήτρα/πίνακας/matrix με 3 γραμμές και 3 στήλες. Ο πίνακας αρχικοποιείται με δεδομένα τις τιμές από 1 έως και 9.

Μεταβλητές και Τύποι Δεδομένων: Matrix

- matrix (πίνακας δύο διαστάσεων)
  - Χρήση του τελεστή **[]** για την προσπέλαση και τροποποίηση στοιχείων του πίνακα με τη μορφή **[<αριθμός γραμμής>, <αριθμός στήλης>]**
    - Αρίθμηση ξεκινά από το 1.

```
> aMatrix <- matrix(1:9, nrow = 3, ncol = 3)
>aMatrix[1,1]
[1] 1
>aMatrix[2,3]
[1] 8
```

Εμφάνιση στοιχείου στην πρώτη γραμμή, πρώτη στήλη

Εμφάνιση στοιχείου στην δεύτερη γραμμή, τρίτη στήλη

Μεταβλητές και Τύποι Δεδομένων: Matrix

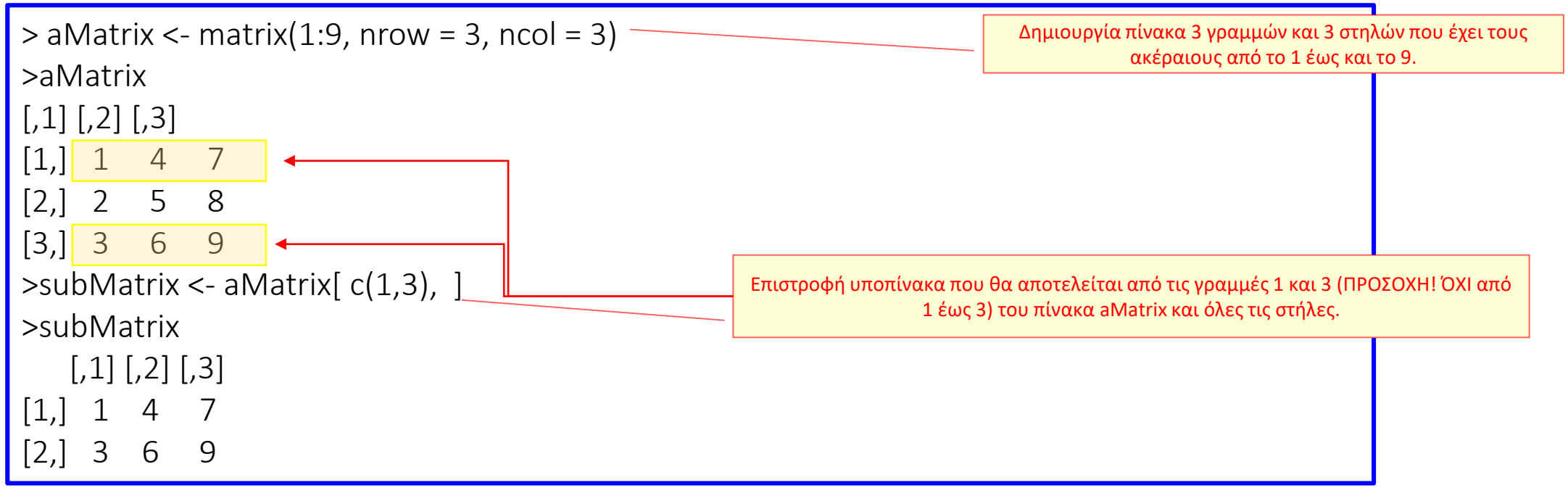
- matrix (πίνακας δύο διαστάσεων)
  - Χρήση του τελεστή **[]** για τον τεμαχισμό πίνακα (slicing) –δλδ τη λήψη συγκεκριμένων γραμμών και στηλών, με προσδιορισμό εύρους γραμμών και στηλών.

```
> aMatrix <- matrix(1:9, nrow = 3, ncol = 3)
>aMatrix
[,1] [,2] [,3]
[1,]  1    4    7
[2,]  2    5    8
[3,]  3    6    9
>subMatrix <- aMatrix[2:3, 1:2]
>subMatrix
     [,1] [,2]
[1,]  2    5
[2,]  3    6
```

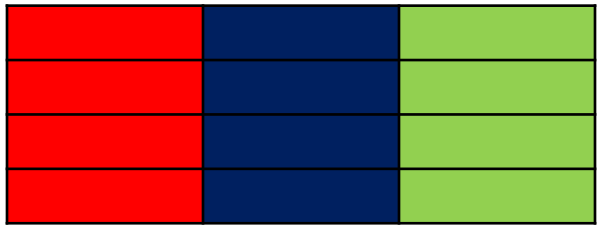
Επιστροφή από τον πίνακα aMatrix του υποπίνακα που αποτελείται από τις γραμμές από 2 έως και 3 και τις στήλες από 1 έως και 2 .

Μεταβλητές και Τύποι Δεδομένων: Matrix

- matrix (πίνακας δύο διαστάσεων)
  - Χρήση του τελεστή **[]** για τον τεμαχισμό πίνακα (slicing) –δλδ τη λήψη συγκεκριμένων γραμμών και στηλών, με προσδιορισμό εύρους γραμμών και στηλών.



Μεταβλητές και Τύποι Δεδομένων



- **Data Frame** (πλαίσιο δεδομένων)
  - Δισδιάστατη αναπαράσταση δεδομένων, **όπου κάθε στήλη μπορεί να είναι οποιουδήποτε τύπου δεδομένων**, όχι υποχρεωτικά η ίδια.
  - Πλαίσια δεδομένων μπορεί να έχουν επικεφαλίδες (ονόματα στηλών) ή όχι.
    - Επικεφαλίδες μπορεί να χρησιμοποιηθούν για την αναφορά ολόκληρης της στήλης ενός πλαισίου δεδομένων
  - Δημιουργία πλαισίων δεδομένων με διάφορους τρόπους. Ένας δημοφιλής είναι με συνένωση διανυσμάτων με την εντολή data.frame().
    - Άλλος, με την ανάγνωση .csv αρχείων.

```
>names <- c("Jim", "Maria", "Karen")
>age<-c(54, 22, 48)
>aDataFrame <- data.frame( names, age)
>aDataFrame
   names age
1    Jim  54
2  Maria  22
3  Karen  48
```

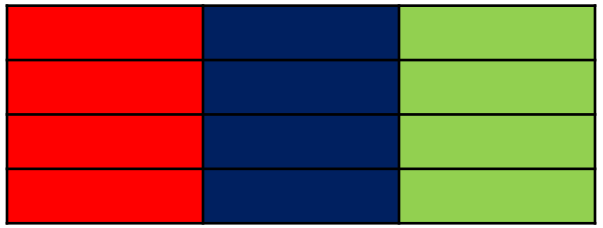
Δημιουργία διανύσματος με τιμές που είναι συμβολοσειρές.

Δημιουργία διανύσματος με τιμές που είναι ακέραιες.

Επικεφαλίδα (ονόματα στηλών) πλαισίου

Δημιουργία νέας μεταβλητής με όνομα aDataFrame που είναι πλαίσιο δεδομένων (data frame) που έχει 2 στήλες και 3 γραμμές με τιμές των διανυσμάτων names και age. Το πλαίσιο δεδομένων θα έχει επικεφαλίδα, με την πρώτη επικεφαλίδα να έχει όνομα names και η δεύτερη age (τα ονόματα των μεταβλητών που χρησιμοποιήθηκαν για τη δημιουργία του πλαισίου δεδομένων).

Μεταβλητές και Τύποι Δεδομένων



- **data.frame()**: Δημιουργία πλαισίου δεδομένων)
  - Χρήση της data.frame() με ορίσματα διανύσματα που περιέχουν τις τιμές των δεδομένων ανά στήλη

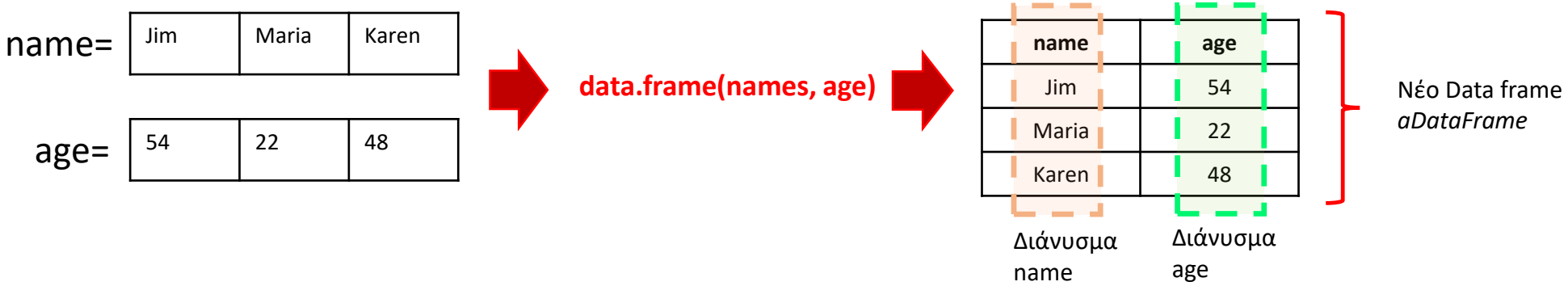
Δημιουργία διανύσματος με τιμές που είναι συμβολοσειρές.

Δημιουργία διανύσματος με τιμές που είναι ακέραιες.

Επικεφαλίδα (ονόματα στηλών) πλαισίου

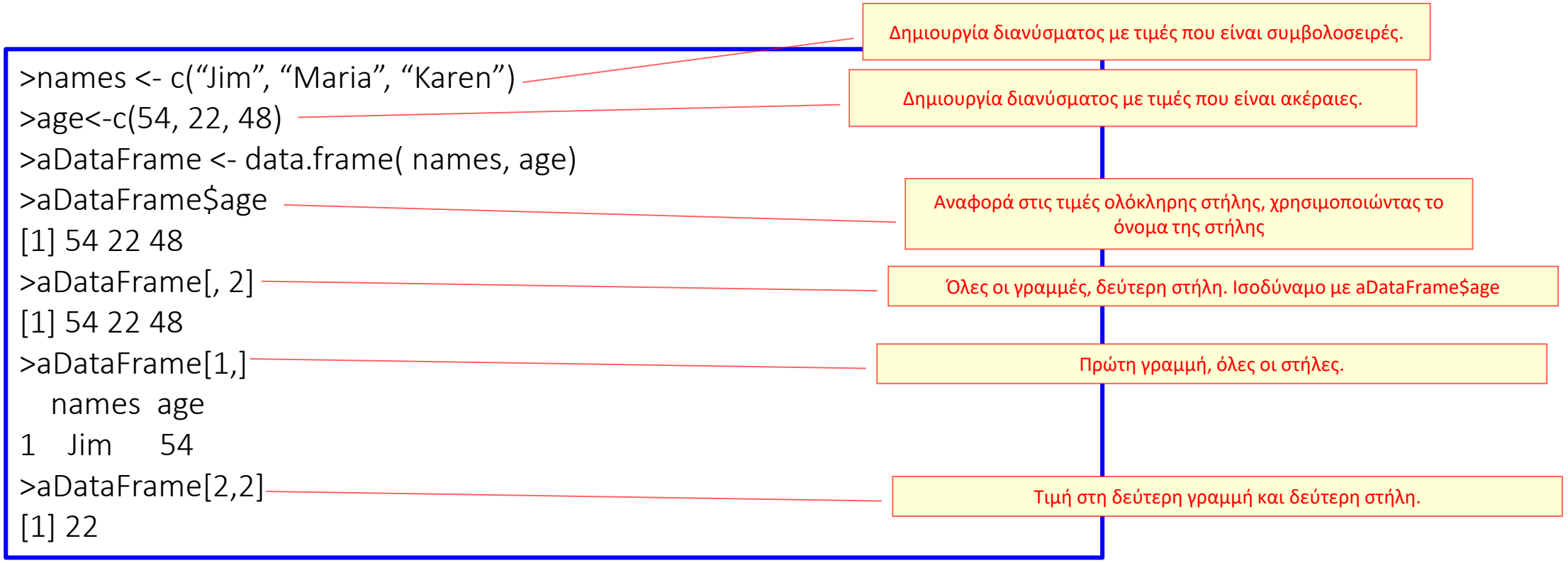
```
>names <- c("Jim", "Maria", "Karen")
>age<-c(54, 22, 48)
>aDataFrame <- data.frame( names, age)
>aDataFrame
   names age
1    Jim  54
2  Maria  22
3  Karen  48
```

Δημιουργία νέας μεταβλητής με όνομα aDataFrame που είναι πλαίσιο δεδομένων (data frame) που έχει 2 στήλες και 3 γραμμές με τιμές των διανυσμάτων names και age. Το πλαίσιο δεδομένων θα έχει επικεφαλίδα, με την πρώτη επικεφαλίδα να έχει όνομα names και η δεύτερη age (τα ονόματα των μεταβλητών που χρησιμοποιήθηκαν για τη δημιουργία του πλαισίου δεδομένων).



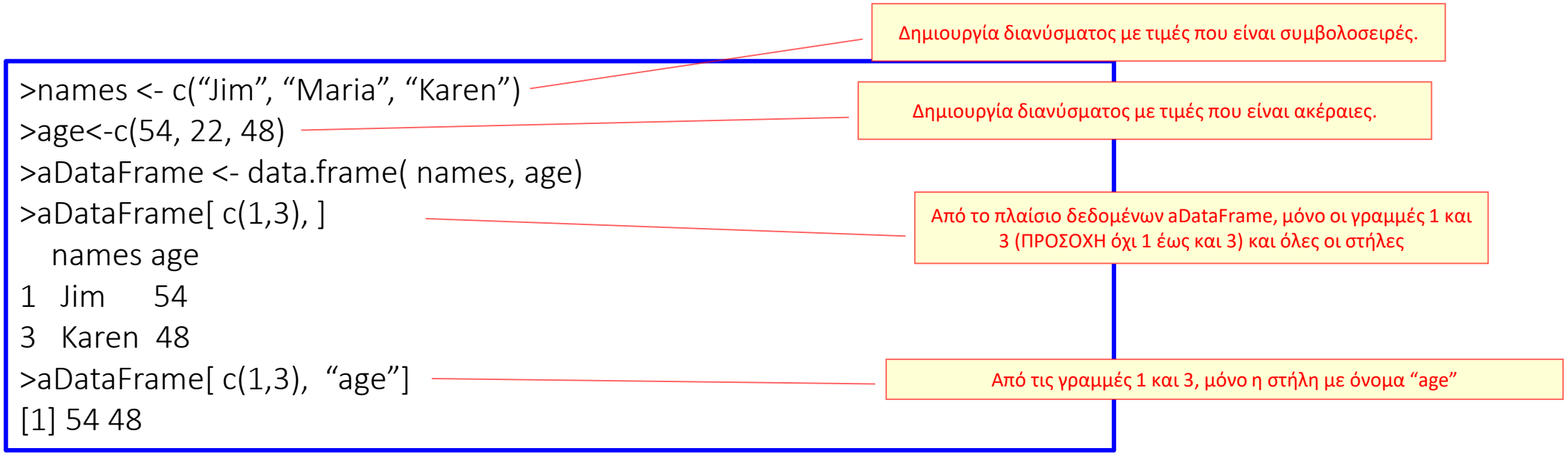
Μεταβλητές και Τύποι Δεδομένων: Data frame

- Data Frame (πλαίσιο δεδομένων)
  - Αναφορά σε στήλες ή γραμμές
    - Με τον τελεστή τεμαχισμού **[]** και αναφορά στον αριθμό γραμμής/στήλης ή στην επικεφαλίδα στήλης εάν υπάρχει. **NOTE:Αρίθμηση ξεκινά από το 1.**
    - Με τον τελεστή **$** και όνομα στήλης (**εάν υπάρχει επικεφαλίδα**)



Μεταβλητές και Τύποι Δεδομένων: Data frame

- Data Frame (πλαίσιο δεδομένων)
  - Αναφορά σε στήλες ή γραμμές
    - Με τον τελεστή τεμαχισμού **[]** και αναφορά στον αριθμό γραμμής/στήλης
    - Με τον τελεστή **$** και όνομα στήλης (**εάν υπάρχει επικεφαλίδα**)



Μεταβλητές και Τύποι Δεδομένων: Data frame

- Τρόποι επιλογής συγκεκριμένων γραμμών και στηλών από ένα πλαίσιο δεδομένων

Τελεστής : (colon) που εκφράζει το *από-έως* (συμπεριλαμβανομένου από και έως) και χρησιμοποιείται για γραμμές και στήλες.

Π.χ. aDataFrame[2:3, 2] που σημαίνει επιστροφή/εμφάνιση των γραμμών 2 και 3 και από αυτές μόνο τη στήλη 2 του πλαισίου δεδομένων aDataFrame (γραμμές και στήλες ξεκινούν την αρίθμησή τους από το 1)

```
>names <- c("Jim", "Maria", "Karen")
>age<-c(54, 22, 48)
>aDataFrame <- data.frame( names, age)
>aDataFrame[ 2:3,  2]
[1] 22 48
>aDataFrame[2:3, "age"]
[1] 22 48
>aDataFrame[:2,]
   names  age
1    Jim     54
>aDataFrame[2,2]
[1] 22
```

Από το πλαίσιο δεδομένων aDataFrame, οι γραμμές από 1 έως και 3 και από αυτές μόνο η στήλη 2.

Από το πλαίσιο δεδομένων aDataFrame, οι γραμμές από έως και 3 και από αυτές μόνο τη στήλη με όνομα/επικεφαλίδα age. (ΣΗΜ: το ίδιο με παραπάνω)

Συντακτικό σφάλμα! Στην R πρέπει να δηλώνονται ρητά τόσο το από όσο και το εώς

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Τρόποι επιλογής συγκεκριμένων γραμμών και στηλών από ένα πλαίσιο δεδομένων

Διάνυσμα συγκεκριμένων τιμών που περιέχει τον αριθμό συγκεκριμένων γραμμών και στηλών που πρέπει να επιστραφούν. Χρήση του τελεστή c() για τη δημιουργία διανύσματος

Π.χ. aDataFrame[c(1, 3), 1] που σημαίνει επιστροφή/εμφάνιση των γραμμών 1 και 3 και από αυτές μόνο τη στήλη 1 του πλαισίου δεδομένων aDataFrame.

```
>names <- c("Jim", "Maria", "Karen")
>age<-c(54, 22, 48)
>aDataFrame <- data.frame( names, age)
>aDataFrame[ c(1,3),  2]
[1] 54 48
```

Από το πλαίσιο δεδομένων aDataFrame, οι γραμμές 1 και 3 και από αυτές μόνο η στήλη 2.

Τέτοιος προσδιορισμός μπορεί να γίνει και για τις στήλες. Οι τιμές μπορεί να είναι είτε αριθμητικές είτε η επικεφαλίδα της στήλης.

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Τρόποι επιλογής συγκεκριμένων γραμμών και στηλών από ένα πλαίσιο δεδομένων

Χρήση κριτηρίων. Π.χ. επιλογή εκείνων των *γραμμών* σε ένα πλαίσιο δεδομένων που έχουν συγκεκριμένη τιμή σε κάποια στήλη.

Χρήση της *which()* η οποία επιστρέφει διάνυσμα με τον αριθμό όλων εκείνων των γραμμών που πληρούν κάποιο κριτήριο σε διάνυσμα που δίνεται ως όρισμα. Π.χ. *which(aDataFrame[, 1] == "Jim")* επιστρέφει διάνυσμα που περιέχει τον αριθμό γραμμής εκείνων των γραμμών του aDataFrame όπου η στήλη 1 έχει τιμή Jim.

```
>names <- c("Jim", "Maria", "Karen")
>age<-c(54, 22, 48)
>aDataFrame <- data.frame( names, age)
>aDataFrame[ which( aDataFrame[,2] < 30),  ]
[1] Maria 22
```

Η which() θα επιστρέψει ένα διάνυσμα με εκείνες τις γραμμές, οι οποίες στη στήλη 2 (age), έχουν τιμή μικρότερη από 30. Το διάνυσμα που προκύπτει θα έχει ως τιμές τους αριθμούς γραμμής που πληρούν το κριτήριο. Το διάνυσμα αυτό δίνεται ως προσδιοριστής γραμμής στον τελεστή [ ] και κατά συνέπεια θα επιστραφούν μόνο αυτές οι γραμμές του πλαισίου δεδομένων aDataFrame

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Περισσότερα για τη *which(…)*:
  - Η which() επιστρέφει διάνυσμα με τον αριθμό θέσης ενός άλλου διανύσματος με **λογικές τιμές** που δίνεται όρισμα στις οποίες θέσεις υπάρχει **τιμή TRUE**.
    - Χρησιμοποιείται κυρίως για τον προσδιορισμό της θέσης μιας τιμής σε έναν πίνακα ή διάνυσμα, η οποία (θέση) πληροί μία συνθήκη.
  - Ορίσματα: διάνυσμα με λογικές (μόνο) τιμές. Μπορεί να είναι και οποιαδήποτε έκφραση που επιστρέφει διάνυσμα με λογικές τιμές.

```
>booleanVector <- c(TRUE, FALSE, TRUE, TRUE, FALSE, TRUE)
>which( booleanVector )
[1] 1  3  4  6
```

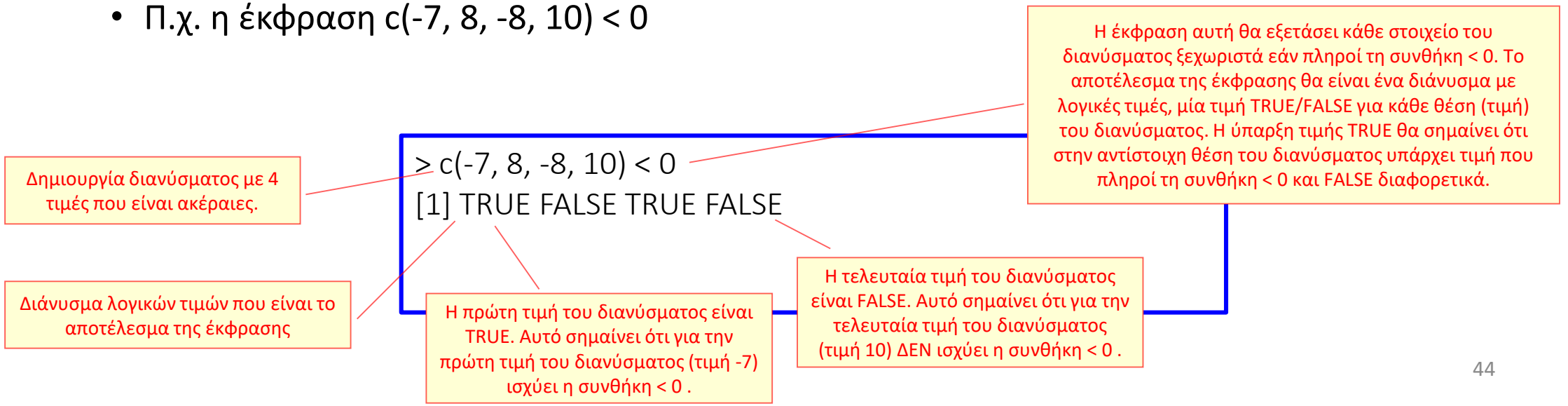
Δημιουργία διανύσματος με 6 (σε πλήθος) λογικών τιμών

Το διάνυσμα με τις 6 λογικές τιμές δίνεται ως όρισμα στη which(). Θα επιστρέψει τον αριθμό θέσεις του διανύσματος booleanVector όπου υπάρχει τιμή TRUE

Αποτέλεσμα της which(): διάνυσμα με αριθμούς θέσεων του διανύσματος booleanVector όπου υπάρχει τιμή TRUE. Εδώ τα 1 3 4 και 6 σημαίνουν ότι στο διάνυσμα λογικών τιμών booleanVector τιμή TRUE υπάρχει στις θέσεις 1, 3,4 και 6.

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Περισσότερα για τη *which(…)*:
  - Οι τελεστές σύγκρισης >, ==, <=, >=, <> μπορούν να εφαρμοστούν και σε διανύσματα όπου συγκρίνεται κάθε στοιχείο ξεχωριστά με την τιμή, *επιστρέφοντας διάνυσμα με λογικές τιμές TRUE/FALSE* που δηλώνει εάν η συνθήκη ισχύει για το στοιχείο στη συγκεκριμένη θέση του διανύσματος ή όχι.
    - Π.χ. η έκφραση c(-7, 8, -8, 10) < 0



Μεταβλητές και Τύποι Δεδομένων: Data frame

- Περισσότερα για τη *which(…)*:
  - Αφού οι τελεστές σύγκρισης επιστρέφουν διάνυσμα με λογικές τιμές TRUE/FALSE, μπορούν να χρησιμοποιηθούν ως ορίσματα στη which()

```
>which( c(-7, 8, -8, 10) < 0 ) )
[1] 1   3
```

Η έκφραση
c(-7, 8, -8, 10) < 0 θα επιστρέψει ως αποτέλεσμα το διάνυσμα
TRUE FALSE TRUE FALSE οποίο δίνεται ως όρισμα στην which() και θα εκτελεστεί η έκφραση
which(**TRUE FALSE TRUE FALSE**)
η οποία θα επιστρέψει τον αριθμό θέσεων όπου υπάρχει η τιμή TRUE στο διάνυσμα που δίνεται ως όρισμα. Έτσι εδώ, το αποτέλεσμα θα είναι
[1] 1 3
που σημαίνει ότι στις θέσεις 1 και 3 του διανύσματος TRUE FALSE TRUE FALSE υπάρχουν οι τιμές TRUE που είναι και αυτές που πληρούν τη συνθήκη < 0 .

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Περισσότερα για τη *which(…)*:
  - Ό,τι ισχύει για τα κριτήρια πάνω σε διανύσματα ισχύει - με την ίδια λογική - για στήλες σε πλαίσια δεδομένων. Έτσι

```
>names <- c(“Jim”, “Maria”, “Karen”)
>age<-c(54, 22, 48)
>aDataFrame <- data.frame( names, age)
>aDataFrame[ , “age” ] < 32
[1] FALSE TRUE FALSE
```

Η έκφραση θα επιστρέψει διάνυσμα με λογικές τιμές, μία για κάθε γραμμή του πλαισίου δεδομένων aDataFrame. Το διάνυσμα θα έχει τιμή TRUE εάν η αντίστοιχη θέση του διανύσματος aDataFrame[, “age”] έχει τιμή < 32 και FALSE διαφορετικά.

Μόνο η δεύτερη τιμή του διανύσματος έχει τιμή TRUE που σημαίνει ότι μόνο η δεύτερη γραμμή του πλαισίου δεδομένων έχει τιμή στη στήλη ηλικία (age) που είναι μικρότερη από 32 (και πληροί τη συνθήκη < 32).

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Περισσότερα για τη *which(…)*:
  - Βασισμένοι στην ιδιότητα αυτή της which, μπορεί να χρησιμοποιηθεί για να γίνει επιλογή εκείνων των γραμμών ενός πλαισίου δεδομένων, που πληρούν κάποιο κριτήριο.
    - Π.χ. *Τα ονόματα των ατόμων με ηλικία μικρότερη από 32 έτη.*

name	age
Jim	54
Maria	22
Karen	48

```
>names <- c("Jim", "Maria", "Karen")
>age<-c(54, 22, 48)
>aDataFrame <- data.frame( names, age)
>aDataFrame[which( aDataFrame[,"age"] < 32),  ]
   names   age
2  Maria     22
```

Η έκφραση θα επιστρέψει εκείνες τις γραμμές από το πλαίσιο δεδομένων aDataFrame, όπου η στήλη ηλικία (age) έχει τιμή που πληροί τη συνθήκη < 32

Πως ερμηνεύεται και εκτελείται η έκφραση: *aDataFrame[which(aDataFrame[,"age"] < 32) ,]*

1. aDataFrame[which(aDataFrame[,"age"] < 32),]

Επιστρέφεται διάνυσμα με τιμές TRUE/FALSE, μία για κάθε γραμμή του πλαισίου aDataFrame ανάλογα εάν ισχύει η συνθήκη ή όχι.

2. aDataFrame[which(c(FALSE, TRUE, FALSE)),]

3. aDataFrame[which(c(FALSE, TRUE, FALSE)),]

Η which() επιστρέφει τις θέσεις στο διάνυσμα που δίνεται ως όρισμα έχουν τιμή TRUE. Εδώ μόνο στη θέση 2.

4. aDataFrame[2,]

Από το πλαίσιο δεδομένων aDataFrame, επεστρεψε τη γραμμή 2, όλες τις στήλες.

	names	age
2	Maria	22

Result

Μεταβλητές και Τύποι Δεδομένων: Data frame

aDataFrame =

name	age
Jim	54
Maria	22
Karen	48

- Παραδείγματα της which() για επιλογές γραμμών σε πλαίσιο δεδομένων

```
aDataFrame[which( aDataFrame[,"age"] > 32),  ]
```

```
  names age
1   Jim  54
3 Karen  48
```

Επιστρέφει εκείνες τις γραμμές (όλες οι στήλες) που έχουν ηλικία > 32

```
aDataFrame[which( aDataFrame[,"age"] > 32), “names”  ]
```

```
Jim Karen
```

Επιστρέφει μόνο την τιμή στην στήλη “names” (όνομα) εκείνων των γραμμών που έχουν ηλικία > 32

```
aDataFrame[which( aDataFrame[,"age"] == max(aDataFrame[,"age"]) ), ]
```

```
  names age
1   Jim  54
```

Επιστρέφει εκείνες τις γραμμές (όλες τις στήλες) που έχουν τη μέγιστη ηλικία.

```
length( which( aDataFrame[,"age"] > 32)  )
```

```
[1] 2
```

Επιστρέφει το πλήθος γραμμών όπου στη στήλη age έχουν τιμή > 32. **Η which() επιστρέφει διάνυσμα** όπου μπορεί να εφαρμοστεί η length().

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Περισσότερα για τη which()
  - Δέχεται και παραπάνω από ένα κριτήριο που μπορούν να συνδυαστούν με λογικές πράξεις. Τελεστές **&** (λογικό ΚΑΙ, σύζευξη), **|** (λογικό Ή, διάζευξη)

Χρήματα που ξοδεύει το άτομο κάθε μήνα ΜΟΝΟ σε τρόφιμα.(Ευρώ)

Χρήματα που ξοδεύει το άτομο κάθε μήνα ΜΟΝΟ σε δραστηριότητες (γυμναστήρια, hobby κλπ) (σε Ευρώ)

aDataFrame =

name	Food expenditure	Leisure expenditure
Jim	400	234
Maria	542	144
Karen	468	399
Jordan	341	198

Εφαρμογή δύο κριτηρίων στις γραμμές του πλαισίου δεδομένων aDataFrame. Η συνθήκη επιστρέφει TRUE για εκείνες τις γραμμές όπου η τιμή της Food expenditure είναι > 400 ΚΑΙ η τιμή της Leisure expenditure είναι < 300 . Η συνθήκη αυτή ισχύει μόνο για τη γραμμή 2 του πλαισίου δεδομένων aDataFrame.

```
>which( aDataFrame[, 2] > 400 & aDataFrame[, 3] < 300 )
[1] 2
>aDataFrame[which( aDataFrame[, 2] > 400 & aDataFrame[, 3] < 300 ), "name" ]
>[1] Maria
```

Μόνο η Maria πληροί και τα δύο κριτήρια.

Το όνομα του ατόμου που ξοδεύει περισσότερα από 400 Ευρώ κάθε μήνα σε τρόφιμα και λιγότερα από 300 Ευρώ σε δραστηριότητες.

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Δυναμική προσθήκη νέων στηλών σε υπάρχον πλαίσιο δεδομένων (Data frame)

Τελεστής $

```
>names <- c("Jim", "Maria", "Karen")
>age<-c(54, 22, 48)
>aDataFrame <- data.frame( names, age)
>aDataFrame$weight <- -1
>aDataFrame
names age  weight
1   Jim  54      -1
2 Maria  22      -1
3 Karen  48      -1
```

Προσθήκη νέας στήλης στο υπάρχον Data frame aDataFrame με όνομα weight με χρήση του τελεστή $. Για κάθε γραμμή του Data frame, η στήλη weight θα λάβει τιμή -1 (τιμή αρχικοποίησης)

Τελεστής []

```
>names <- c("Jim", "Maria", "Karen")
>age<-c(54, 22, 48)
>aDataFrame <- data.frame( names, age)
>aDataFrame["weight"] <- -1
>aDataFrame
names age  weight
1   Jim  54      -1
2 Maria  22      -1
3 Karen  48      -1
```

Προσθήκη νέας στήλης στο υπάρχον Data frame aDataFrame με όνομα weight με χρήση του τελεστή [ ]. Για κάθε γραμμή του Data frame, η στήλη weight θα λάβει τιμή -1 (τιμή αρχικοποίησης)

Χρήση cbind

```
>names <- c("Jim", "Maria", "Karen")
>age<-c(54, 22, 48)
>aDataFrame <- data.frame( names, age)
>newCol <- c(78, 62, 61.5)
>aDataFrame<-cbind(aDataFrame, weight=newCol)
names age  weight
1   Jim  54    78
2 Maria  22    62
3 Karen  48  61.5
```

Προσθήκη νέας στήλης στο υπάρχον Data frame aDataFrame με όνομα weight με χρήση της cbind. Επιτρέπει την προσθήκη συγκεκριμένης τιμής για κάθε γραμμή.

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Εφαρμογή συναρτήσεων περιγραφικής στατιστικής πάνω σε στήλες ενός πλαισίου δεδομένων
  - min, max, mean, sd κλπ

aDataFrame =

name	age
Jim	54
Maria	22
Karen	48

```
>min( aDataFrame[, "age"])
[1] 22
>max( aDataFrame[, "age"])
[1] 54
>var(aDataFrame[, "age"])
[1] 289.3333
>sd(aDataFrame[, "age"])
[1] 17.0098
```

Μικρότερη ηλικία.

Μεγαλύτερη ηλικία.

Διακύμανση ηλικίας.

Τυπική απόκλιση ηλικίας.

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Οι συναρτήσεις περιγραφικής στατιστικής **min, max, var, sd** δέχονται επιπλέον ορίσματα.
  - Από τα πιο σημαντικά το όρισμα **na.rm** το οποίο ελέγχει πως θα χειριστούν οι τιμές που λείπουν (NA): εάν θα πρέπει να αφαιρεθούν πριν τον υπολογισμό της συνάρτησης (***na.rm=TRUE***) ή όχι (***na.rm=FALSE***).

aDataFrame =

name	age
Jim	54
Maria	22
Karen	48
Jordan	NA

```
>min( aDataFrame[, "age"])
[1] NA
>min ( aDataFrame[, "age"], na.rm=TRUE)
[1] 22
```

Επειδή υπάρχει τιμή που λείπει (missing value) στη στήλη age, δεν μπορεί να υπολογιστεί το ελάχιστο.

Με το όρισμα na.rm=TRUE δηλώνεται ότι όλες οι τιμές που λείπουν στη στήλη “age” να μην ληφθούν υπόψιν κατά τον υπολογισμό της ελάχιστης τιμής. Δλδ να αφαιρεθούν πριν τον υπολογισμό της ελάχιστης τιμής. Αυτό θα έχει σαν αποτέλεσμα να υπολογίζεται η ελάχιστη τιμή μεταξύ όλων των υπολοίπων τιμών.

Τιμή που λείπει (missing value) στη στήλη age.

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Άλλες συναρτήσεις που μπορούν να εφαρμοστούν πάνω σε στήλες/γραμμές ενός πλαισίου δεδομένων
  - **rowSums()**: Δέχεται ως όρισμα ένα πλαίσιο δεδομένων και για κάθε γραμμή του, υπολογίζει το άθροισμα συγκεκριμένων στηλών του πλαισίου δεδομένων και επιστρέφει ένα διάνυσμα με τα αθροίσματα αυτά.

Χρήματα που ξοδεύει το άτομο κάθε μήνα ΜΟΝΟ σε τρόφιμα.

Χρήματα που ξοδεύει το άτομο κάθε μήνα ΜΟΝΟ σε δραστηριότητες (γυμναστήρια, hobby κλπ)

aDataFrame =

name	Food expenditure	Leisure expenditure
Jim	400	234
Maria	542	144
Karen	468	399
Jordan	341	198

"Πόσα ξοδεύει κάθε άτομο συνολικά το μήνα (Food expenditure + Leisure expenditure); "
Υπολογισμός με χρήση της rowSums αφού πρέπει για κάθε γραμμή να προστεθούν οι στήλες Food expenditure + Leisure expenditure.

```
>rowSums( aDataFrame[, 2:3] )
[1] 634  686  867  539
```

Για κάθε γραμμή του πλαισίου δεδομένων aDataFrame πρόσθεσε τις τιμές από τις στήλες 2 έως και 3 (2=Food exp, 3=Leisure exp).
Εναλλακτικά:
rowSums(aDataFrame[, c("FoodExpenditure","LeisureExpenditure")])

Διάνυσμα με τα αθροίσματα για κάθε γραμμή. Η πρώτη τιμή είναι το άθροισμα των στηλών 2 και 3 της 1ης γραμμής του πλαισίου δεδομένων aDataFrame (Jim), η δεύτερη το άθροισμα της 2ης κ.ο.κ. Ένα άθροισμα για κάθε γραμμή του πλαισίου aDataFrame.

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Άλλες συναρτήσεις που μπορούν να εφαρμοστούν πάνω σε στήλες/γραμμές ενός πλαισίου δεδομένων
  - **rowSums()**

Χρήματα που ξοδεύει το άτομο κάθε μήνα ΜΟΝΟ σε τρόφιμα.

Χρήματα που ξοδεύει το άτομο κάθε μήνα ΜΟΝΟ σε δραστηριότητες (γυμναστήρια, hobby κλπ)

aDataFrame =

name	Food expenditure	Leisure expenditure
Jim	400	234
Maria	542	144
Karen	468	399
Jordan	341	198

"Τί θα υπολογίσει η παρακάτω έκφραση και ποιο το αποτέλεσμα που θα εμφανιστεί στην οθόνη εάν η μεταβλητή aDataFrame είναι το πλαίσιο δεδομένων που εμφανίζεται στα αριστερά;"

```
aDataFrame[which( rowSums( aDataFrame[, 2:3] ) == max(rowSums( aDataFrame[, 2:3] ))  ), "names"]
```

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Άλλες συναρτήσεις που μπορούν να εφαρμοστούν πάνω σε στήλες/γραμμές ενός πλαισίου δεδομένων
  - **rowSums()**

Χρήματα που ξοδεύει το άτομο κάθε μήνα ΜΟΝΟ σε τρόφιμα.

Χρήματα που ξοδεύει το άτομο κάθε μήνα ΜΟΝΟ σε δραστηριότητες (γυμναστήρια, hobby κλπ)

aDataFrame =

name	Food expenditure	Leisure expenditure
Jim	400	234
Maria	542	144
Karen	468	399
Jordan	341	198

"Τί θα υπολογίσει η παρακάτω έκφραση και ποιο το αποτέλεσμα που θα εμφανιστεί στην οθόνη εάν η μεταβλητή aDataFrame είναι το πλαίσιο δεδομένων που εμφανίζεται στα αριστερά;"

SOL: *Το όνομα εκείνου του ατόμου που ξοδεύει το περισσότερο συνολικό ποσό (Food + Leisure) κάθε μήνα. Το αποτέλεσμα θα είναι*
[1] Karen.

```
aDataFrame[which( rowSums( aDataFrame[, 2:3] ) == max(rowSums( aDataFrame[, 2:3] ))  ), "names"]
```

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Άλλες συναρτήσεις που μπορούν να εφαρμοστούν πάνω σε στήλες/γραμμές ενός πλαισίου δεδομένων
  - **colSums()**: Δέχεται ως όρισμα ένα πλαίσιο δεδομένων και για κάθε στήλη του, υπολογίζει το άθροισμα των τιμών σε αυτές, επιστρέφοντας ένα διάνυσμα με τα αθροίσματα αυτά.

Χρήματα που ξοδεύει το άτομο κάθε μήνα ΜΟΝΟ σε τρόφιμα.

Χρήματα που ξοδεύει το άτομο κάθε μήνα ΜΟΝΟ σε δραστηριότητες (γυμναστήρια, hobby κλπ)

aDataFrame =

name	Food expenditure	Leisure expenditure
Jim	400	234
Maria	542	144
Karen	468	399
Jordan	341	198

"Πόσα ξοδεύει κάθε άτομο συνολικά το μήνα (Food expenditure + Leisure expenditure); "
Υπολογισμός με χρήση της rowSums αφού πρέπει για κάθε γραμμή να προστεθούν οι στήλες Food expenditure + Leisure expenditure.

```
>colSums( aDataFrame[, 2:3] )
FoodExpenditure LeisureExpenditure
           1751                975
```

Άθροισμα στηλών 2 (FoodExpenditure) και 3 (LeisureExpenditure)

Άθροισμα τιμών της στήλης FoodExpenditure (στήλη 2)

Άθροισμα τιμών της στήλης LeisureExpenditure (στήλη 3)

Μεταβλητές και Τύποι Δεδομένων: Data frame

- Άλλες συναρτήσεις που μπορούν να εφαρμοστούν πάνω σε στήλες/γραμμές ενός πλαισίου δεδομένων
  - Οι συναρτήσεις **rowSums()** και **colSums()** δέχονται και όρισμα **na.rm** προκειμένου να χειριστούν τιμές που λείπουν (missing values)
    - Π.χ όρισμα ***na.rm = TRUE*** σημαίνει ότι τιμές που λείπουν δεν θα συμπεριληφθούν στο άθροισμα. Θα αγνοηθούν από τις rowSums() και colSums(). Προκαθορισμένη τιμή είναι FALSE (δλδ ***na.rm = FALSE***)

Ανάγνωση δεδομένων από αρχείο: Τρέχων φάκελος

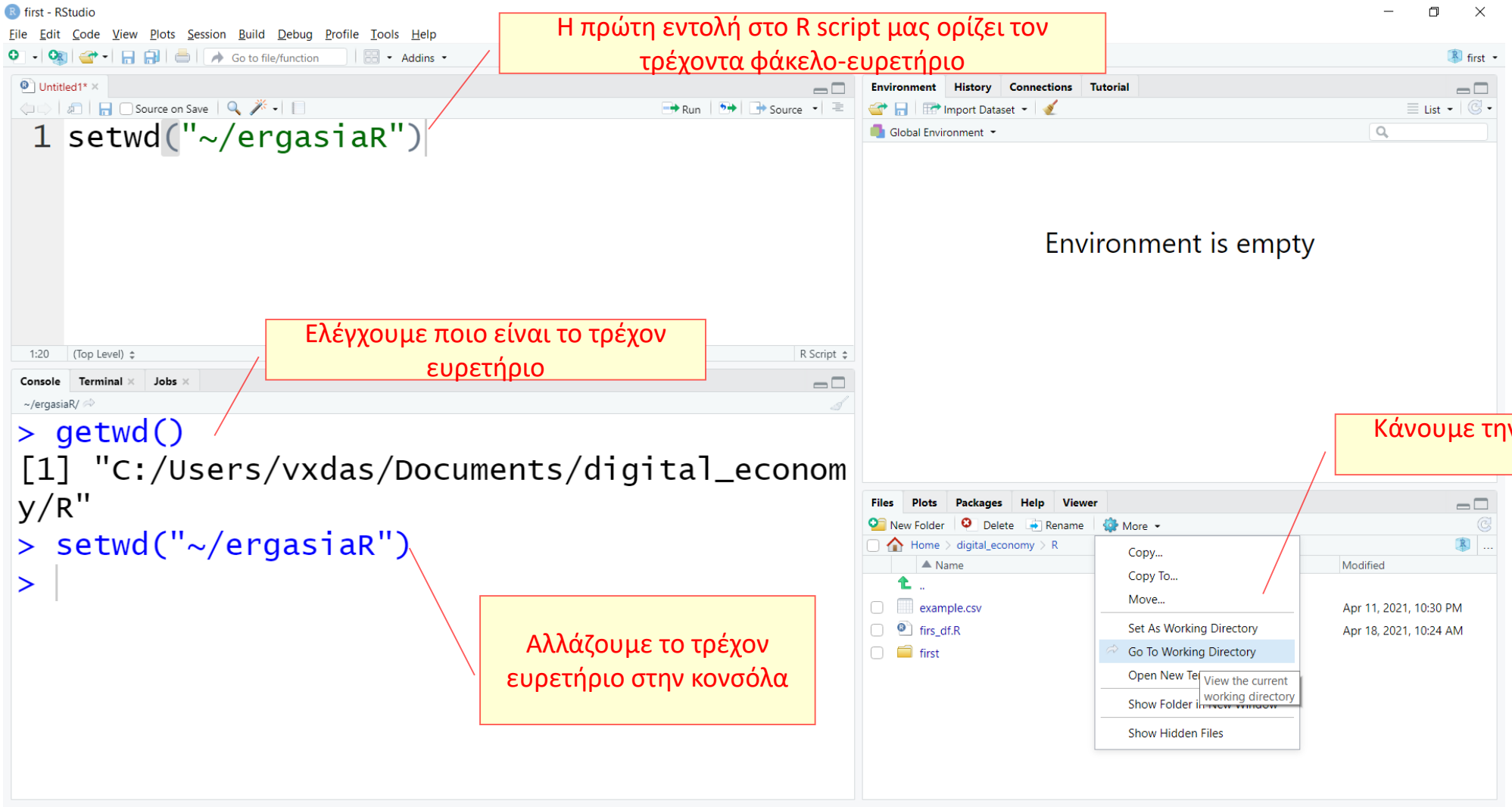
- Το RStudio χρησιμοποιεί το όρο του *global default working directory* (συνολικό προκαθορισμένο φάκελο-ευρετήριο) που είναι ο φάκελος του χρήστη (user home directory) (τυπικά ορίζεται ως ~ στο R)
- Η εντολή dirname("~") δίνει το όνομα του *global default working* directory.
- Θα πρέπει στο R να οριστεί το τρέχον ευρετήριο εργασίας (working directory), δηλαδή εκεί που θα αποθηκεύσετε τα δεδομένα και τον κώδικά σας.
- Η εντολή getwd() επιστρέφει το τρέχον working directory
- Η εντολή setwd() ορίζει ποιο θα είναι το νέο working directory

Ορισμός τρέχοντος φακέλου για την εργασία

- Στην περίπτωση MS Windows στο φάκελο Documents του χρήστη δημιουργήστε τον φάκελο “ergasiaR” με τη χρήση του Windows Explorer. Θα αποθηκεύετε εκεί:
  - Όλα τα παραδείγματα και τα αρχεία δεδομένων του Εργαστηρίου μας
  - Το αρχείο δεδομένων που περιέχει το zip αρχείο της Εργασίας
- Ορίστε το φάκελο αυτό ως το current working directory στην αρχή του κώδικά σας

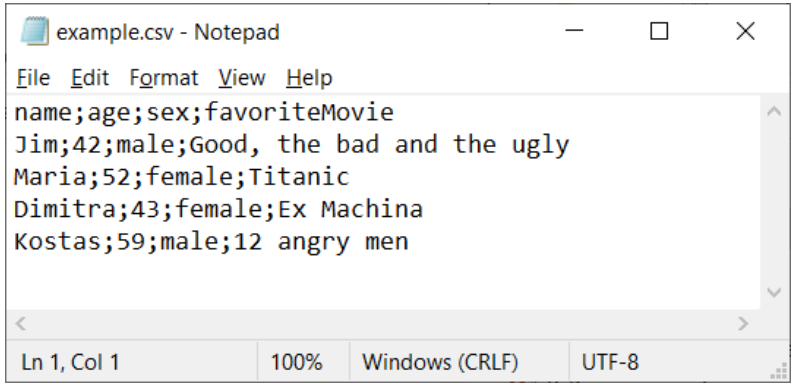
```
> getwd() [1] "C:/Users/vxdas/Documents/digital_economy/R"
> setwd("~/ergasiaR")
>
```

Ορισμός του τρέχοντος φακέλου εργασίας



Το αρχείο CSV

Αρχείο example.csv, έχει επικεφαλίδα (πρώτη γραμμή) και ο διαχωριστής τιμών σε κάθε γραμμή είναι ο χαρακτήρας ;



- Αρχείο CSV:
  - Αρχείο κειμένου (text) με δεδομένα σε γραμμές και στήλες.
  - Κάθε στήλη διαχωρίζεται με ειδικό χαρακτήρα (delimiter) (συνήθως κόμμα ή ελληνικό ερωτηματικό ;)
  - Κάθε γραμμή τελειώνει με τον ειδικό χαρακτήρα αλλαγής γραμμής.
  - Η πρώτη γραμμή συνήθως έχει τη μορφή επικεφαλίδας (header) και ορίζει τα ονόματα των στηλών-μεταβλητών
- Παράδειγμα:
  - Κατεβάστε στο αρχείο example.csv από το eclass στο τρέχον ευρετήριο εργασίας (στο φάκελο Document του χρήστη στο φάκελο ergasiaR)

Ανάγνωση από αρχείο CSV σε Data Frame

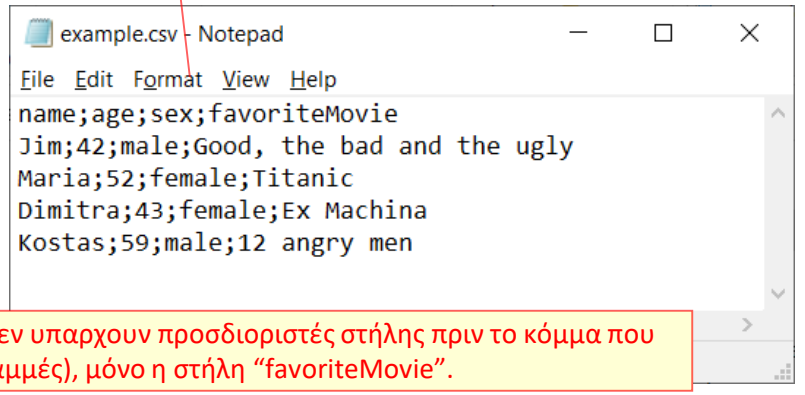
- Χρήση της read.csv() με τα κατάλληλα ορίσματα, τα πιο σημαντικά από τα οποία είναι:
  - Όνομα αρχείου
  - Εάν το αρχείο έχει επικεφαλίδα (όρισμα header)
  - Διαχωριστής (όρισμα sep): Με ποιον τρόπο (χαρακτήρα) διαχωρίζονται οι τιμές μιας γραμμής.
  - Όρισμα stringsAsFactors που δηλώνει εάν οι στήλες που είναι συμβολοσειρές να ερμηνευτούν ως κατηγορικές τιμές –factors- (stringsAsFactors=TRUE) ή όχι (stringsAsFactors=FALSE). **Προκαθορισμένη τιμή** του ορίσματος **stringsAsFactors** (εάν δεν δοθεί) **είναι TRUE**.

Είναι πλαίσιο δεδομένων (Data Frame)

```
> df <- read.csv("example.csv", header=TRUE, sep=";“, stringsAsFactors=FALSE)
> df
    name age gender           favoriteMovie
1    Jim  42   male Good, the bad and the ugly
2  Maria  52 female                Titanic
3 Dimitra  43 female             Ex Machina
4 Kostas  59   male             12 angry men
> df[,"favoriteMovie"]
```

Ανάγνωση αρχείου. Η μεταβλητή df είναι πλαίσιο δεδομένων (data frame)

Αρχείο example.csv που θέλουμε να διαβάσουμε σε πλαίσιο δεδομένων (data frame). Το αρχείο έχει επικεφαλίδα (πρώτη γραμμή) και ο διαχωριστής τιμών σε κάθε γραμμή είναι ο χαρακτήρας ;



Για όλες τις γραμμές (ΠΡΟΣΟΧΗ ότι δεν υπαρχουν προσδιοριστές στήλης πριν το κόμμα που σημαίνει όλες οι γραμμές), μόνο η στήλη “favoriteMovie”.

Δομή πλαισίου δεδομένων και στήλες

- Εμφάνιση δομής πλαισίου δεδομένων με χρήση της **str()**
  - Δομή: στήλες (variables) και τύπος δεδομένων τους σε πλαίσιο δεδομένων

```
>df <- read.csv(“example.csv”, header=TRUE, sep=“;”, stringsAsFactors=FALSE)
>str(df)
 > str(df)
'data.frame':          4 obs. of  4 variables:
 $ name        : chr  "Jim" "Maria" "Dimitra" "Kostas"
 $ age         : int  42 52 43 59
 $ gender      : chr  "male" "female" "female" "male"
 $ favoriteMovie: chr  "Good, the bad and the ugly" "Titanic" "Ex Machina" "12 angry men“
>  str(df$gender)
 chr [1:4] "male" "female" "female" "male"
> str(df[,"gender"])
 chr [1:4] "male" "female" "female" "male"
> str(df["gender"])
'data.frame':          4 obs. of  1 variable:
 $ gender: chr  "male" "female" "female" "male"
```

Δομή πλαισίου δεδομένων df. Εμφανίζονται οι στήλες και ο τύπος δεδομένων των τιμών κάθε στήλης.

Τύπος δεδομένων των τιμών της στήλης age.

Τύπος δεδομένων των τιμών της στήλης gender.

Η στήλη gender ως vector

Η στήλη gender ως dataframe

Κατηγορικές μεταβλητές: Ο τύπος factor

- Κατηγορικές μεταβλητές:
  - Μια μεταβλητή που οι τιμές της κατανέμονται σε κατηγορίες ανάλογα με κάποια ποιοτικά χαρακτηριστικά, όπως το φύλλο, η οικογενειακή κατάσταση, η εθνικότητα κ.α., ονομάζεται *κατηγορική μεταβλητή*. Διακρίνονται σε διατάξιμες (π.χ. επίπεδα θερμοκρασίας) και μη-διατάξιμες (π.χ. φύλλο). Περισσότερα…
  - Στην R οι κατηγορικές μεταβλητές έχουν τον τύπο factor (παράγοντας)

```
> nationality<-factor(c('Italian','French','French','Greek','Italian'))
> nationality
[1] Italian French  French  Greek   Italian
Levels: French Greek Italian
> replies=c('high','high','low','medium','low','low')
> results<-factor(replies,ordered=TRUE,levels=c('low','medium','high'))
> results
[1] high   high   low    medium low    low
Levels: low < medium < high
> table(results)
results
   low     medium   high
     3          1          2
```

nationality είναι κατηγορική μεταβλητή τύπου factor

Έχει 5 τιμές και 3 βαθμίδες (levels): French, Greek, Italian

Το διάνυσμα replies περιέχει απαντήσεις σε ερώτηση για τη θερμοκρασία

Η μεταβλητή results είναι κατηγορική (factor) διατεταγμένη (ordered=TRUE) με δεδομένα τις τιμές που είχε το διάνυσμα replies και 3 διατεταγμένες βαθμίδες (levels)

Δημιουργία πίνακα συχνοτήτων ως προς τη βαθμίδα (level) της κατηγορικής μεταβλητής

Κατηγορικές μεταβλητές σε πλαίσια δεδομένων

- Η συνάρτηση as.factor() μετατρέπει μία στήλη ενός πλαισίου δεδομένων από έναν τύπο δεδομένων (χαρακτήρες ή αριθμητικές τιμές) σε κατηγορική τύπου factor (παράγοντα)

```
> df$gender
[1] "male"   "female" "female" "male“
> df$gender<-as.factor(df$gender)
> str(df)
'data.frame':          4 obs. of  4 variables:
 $ name        : chr  "Jim" "Maria" "Dimitra" "Kostas"
 $ age         : int  42 52 43 59
 $ gender      : Factor w/ 2 levels "female","male": 2 1 1 2
 $ favoriteMovie: chr  "Good, the bad and the ugly" "Titanic" "Ex Machina" "12
angry men"
```

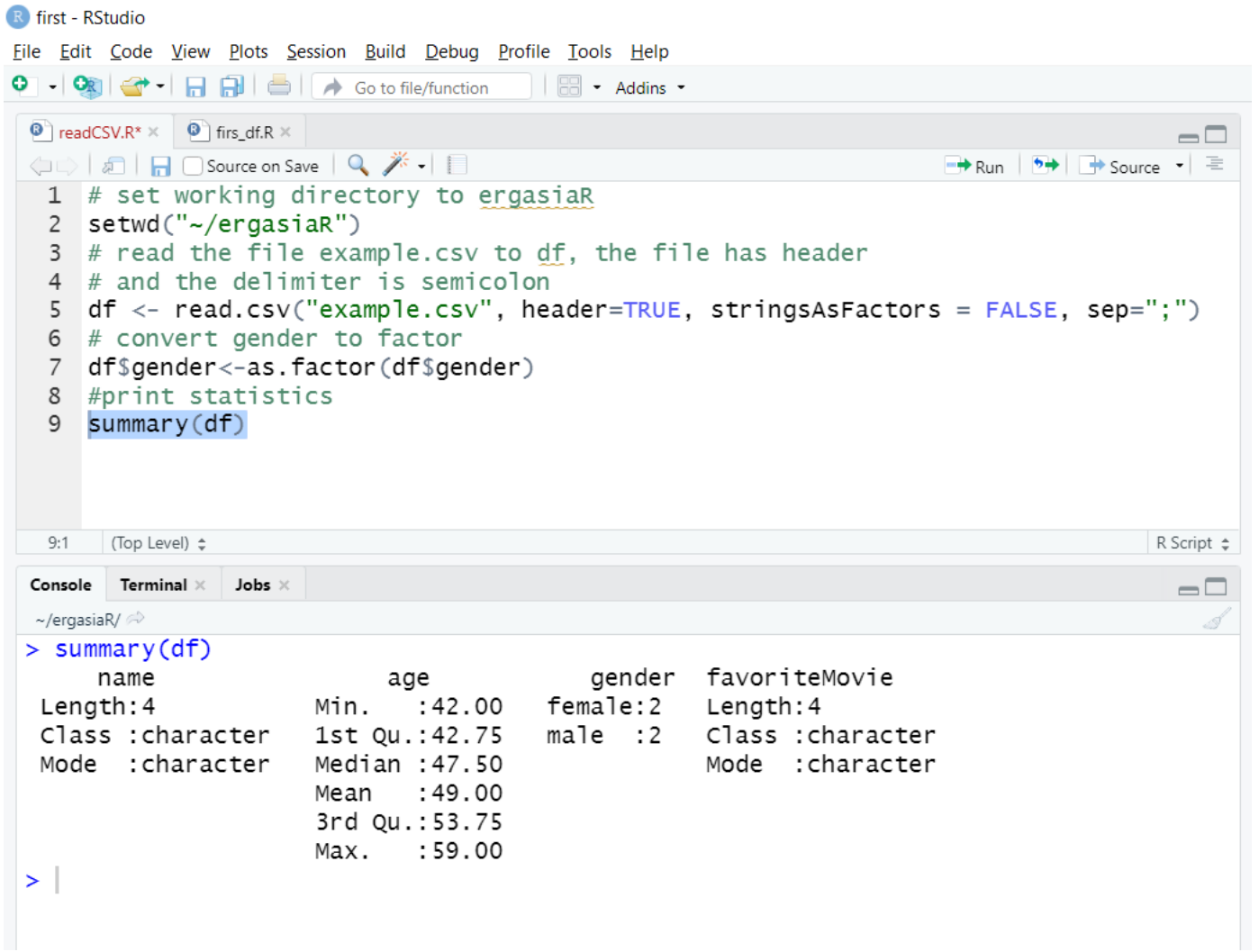
Τα δεδομένα της στήλης gender ως vector συμβολοσειρών

Μετατρέπουμε τα δεδομένα της στήλης gender ως factor

Η στήλη gender είναι τύπου factor με δύο επίπεδα, το female και το male. Η R ταξινομεί τα επίπεδα με λεξικογραφική σειρά, άρα στο επίπεδο “female” δίνει την τιμή 1 και στην τιμή “male” την τιμή 2.

Συγκεντρωτικά στατιστικά μεγέθη σε Data frame

- Η συνάρτηση `summary(object)` χρησιμοποιείται για να παρουσιάσει σύντομα στατιστικά για ένα αντικείμενο
- `summary(df)`:
  - Παρουσιάζει στατιστικά ανά στήλη σε αντικείμενο πλαισίου δεδομένων ανάλογα με τον τύπο της στήλης



Συγκεντρωτικά στατιστικά μεγέθη σε στήλη αριθμητικού τύπου

```
> summary(df$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  42.00   42.75   47.50   49.00   53.75   59.00 
```

- Η συνάρτηση summary() σε αριθμητικά δεδομένα παρουσιάζει:
  - Ελάχιστο (Min.)
  - 1ο Τεταρτημόριο (1st Qu.)
  - Διάμεσο (Median) (= $p_{50}=Q_2$)
  - Αριθμητικό μέσο (Mean)
  - 3ο Τεταρτημόριο (3rd Qu.)
  - Μέγιστο (Max.)

Θυμήσου!
Τα ποσοστημόρια ενός δείγματος συμβολίζονται με p_α. Το ποσοστημόριο p_α είναι η τιμή x, για την οποία ισχύει ότι: το α% των παρατηρήσεων είναι μικρότερες από αυτή και το υπόλοιπο (1-α)% των παρατηρήσεων είναι μεγαλύτερες από αυτή.
Τα ποσοστημόρια διακρίνονται σε:
- *Εκατοστημόρια* (percentiles): p_1, p_2…,p_{99}
- *Δεκατημόρια* (deciles): p_{10}, p_{20},…, p_{90}
- *Τεταρτημόρια* (quartiles): $p_{25}=Q_1$, $p_{50}=Q_2$, $p_{75}=Q_3$

Βασικά στατιστικά μεγέθη σε στήλη αριθμητικού τύπου (Συνολικό R script)

```
# set working directory to ergasiaR
setwd("~/ergasiaR")
# read the file example.csv to df, the file has header
# and the delimiter is semicolon
df <- read.csv("example.csv", header=TRUE, stringsAsFactors = FALSE, sep=";")
# convert gender to factor
df$gender<-as.factor(df$gender)
# summary statistics
print(summary(df))
# statistics of a single numeric column
# min and max of age
cat("Min age:",min(df$age),"Max age:",max(df$age))
# arithmetic mean, ignore Nas if exist
cat("Arithmetic mean of age:",mean(df$age,na.rm = TRUE))
# spread measures
cat("Variance:",var(df$age),"Standard Deviation:",sd(df$age),sqrt(var(df$age)))
```