# Εισαγωγή Δεδομένων με το tidyverse:: Σύντομος Οδηγός

## Εισαγωγή Δεδομένων σε μορφή πίνακα με το πακέτο readr

read \*(file, col\_names = TRUE, col\_types = NULL, col\_select = NULL, id = NULL, locale, n\_max = Inf, skip = 0, na = c("", "NA"), guess\_max = min(1000, n\_max), show col types = TRUE)Δείτε?read delim



read\_delim("file.txt", delim = "|") Εισαγωγή αρχείων με οποιονδήποτε οριοθέτη (delimiter), ο οποίος εάν δεν έχει καθοριστεί η εντολή θα ορίσει

Για να δημιουργήσετε ένα αρχείο file.txt μπορείτε να εκτελέσετε:

read\_csv("file.csv") Εισαγωγή αρχείου οριοθετημένου με κόμμα με

read\_csv2("file2.csv") Εισαγωγή αρχείου οριοθετημένου με ερωτηματικό

ABC123 4 5 NA

read\_tsv("file.tsv") Εισαγωγή αρχείου οριοθετημένου με tab. Το ίδιο γίνεται και με την εντολή read\_table().

read\_fwf("file.tsv", fwf\_widths(c(2, 2, NA))) Εισαγωγή αρχείων σταθερού πλάτους. write file("A\tB\tC\n1\t2\t3\n4\t5\tNA\n", file = "file.tsv")

Χρήσιμες Παράμετροι Εισαγωγής Δεδομένων

id = "origin\_file")

Α	В	С	Χωρίς κεφαλίδα (ονόματα στηλών)
1	2	3	read_csv("file.csv", <b>col_names = FALSE</b> )
4	5	NA	
_		_	Ορισμός κεφαλίδων (ονομάτων στηλών)
X	У	Z	
Α	В	С	read_csv("file.csv",
1	2	3	col_names = c("x", "y", "z"))
4	5	NA	
	<b>)</b>		Εισαγωγή πολλών αρχείων σε έναν μόνο πίνακα
4			read_csv(c("f1.csv", "f2.csv", "f3.csv"),



#### Παράλειψη γραμμών read csv("file.csv", skip = 1)

Εισαγωγή ενός υποσυνόλου γραμμών read\_csv("file.csv", n\_max = 1)



1 2 3

Εισαγωγή με ορισμό τιμών ως τιμές που λείπουν



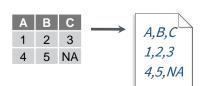
NA 2 3 read\_csv("file.csv", na = c("1"))



Ορισμός δεκαδικών ψηφίων read\_delim("file2.csv", locale = locale(decimal\_mark = ","))

## Αποθήκευση Δεδομένων με το πακέτο readr

write\_\*(x, file, na = "NA", append, col\_names, quote, escape, eol, num\_threads, progress)



write delim(x, file, delim = " ") Αποθηκεύστε αρχεία με οποιοδήποτε οριοθέτη

write csv(x, file) Αποθηκεύστε ένα αρχείο οριοθετημένο με κόμμα.

write\_csv2(x, file) Αποθηκεύστε ένα αρχείο οριοθετημένο με ερωτηματικό.

write tsv(x, file) Αποθηκεύστε ένα αρχείο οριοθετημένο με tab.

Ένα από τα πρώτα βήματα της ανάλυσης δεδομένων είναι η εισανωγή δεδομένων από διάφορες πηγές στην R. Τα δεδομένα αποθηκεύονται συχνά σε μορφές πίνακα, όπως σε αρχεία csv ή σε υπολογιστικά φύλλα.



Η πρώτη σελίδα του οδηγού δίνει σύντομες οδηγίες για την εισαγωγή και αποθήκευση δεδομένων στην R, χρησιμοποιώντας το πακέτο readr.



Η επόμενη σελίδα παρέχει σύντομες οδηγίες για την εισαγωγή υπολογιστικών φύλλων από αρχείαExcel χρησιμοποιώντας το πακέτο readxl ή υπολογιστικά φύλλα Google με το πακέτο googlesheets4.

Άλλοι τύποι δεδομένων

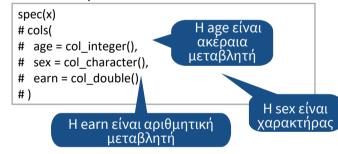
Δοκιμάστε ένα από τα παρακάτω πακέτα για να εισαγάγετε άλλους τύπους αρχείων:

- haven αρχεία SPSS, Stata, και SAS
- DBI βάσεις δεδομένων
- isonlite ison
- xml2 XML
- httr Web APIs
- rvest HTML (Web Scraping)
- readr::read lines() αρχεία κειμένου

## Προδιαγραφές Στηλών με το πακέτο readr

Με το πακέτο readr υπάρχει η δυνατότητα να καθορίσουμε τον τύπο δεδομένων της κάθε στήλης κατά την εισαγωγή δεδομένων στην R. Αν δεν οριστεί από τον χρήστη το πακέτο θα ορίσει αυτομάτως κάποιες προδιαγραφές για τις στήλες.

spec(x) Εξαγωγή προδιαγραφών στηλών για ένα πλαίσιο δεδομένων



#### Τύποι Στηλών

Για κάθε τύπος στήλης αντιστοιχεί μια συνάρτηση και μια αντίστοιχη συντομογραφία.

- col\_logical() "l"
- col integer() "i"
- col double() "d"
- col number() "n"
- col character() "c"
- col\_factor(levels, ordered = FALSE) "f"
- col datetime(format = "") "T"
- col date(format = "") "D"
- col time(format = "") "t"
- col\_skip() "-", "\_"
- col\_guess() "?"

Χρήσιμες παράμετροι για τον ορισμό στηλών Απόκρυψη μηνύματος προδιαγραφών στηλών read \*(file, show col types = FALSE)

Εισαγωγή επιλεγμένων στηλών Η επιλογή γίνεται μέσω ονομάτων ή θέσης στηλών read \*(file, col select = c(age, earn))

Μαντέψτε τους τύπους στηλών

Για να μαντέψετε τον τύπο δεδομένων μιας στήλης, η rεντολή ead\_ \*() σας δίνει πρόσβαση στις πρώτες 1000 γραμμές των δεδομένων, κάτι που μπορεί να αλλάξει από την παράμετρο guess max. read\_\*(file, guess\_max = Inf)

Ορισμός προδιαγραφών στήλης Ορισμός προεπιλεγμένου τύπου

read csv( col type = list(.default = col double())

Ορισμός μέσω συνάρτησης ή συντομογραφίας

col type = list(x = col double(), v = "l", z = "")

Ορισμός μέσω χαρακτήρων συντομογραφίας

# col types: skip, guess, integer, logical, character read\_csv( file. col type = "?ilc"

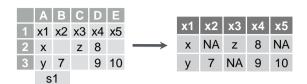




### Εισαγωγή Υπολογιστικών Φύλλων

### με το πακέτο readxl

#### Εισανωνή Αρχείων Excel



read excel(path, sheet = NULL, range = NULL) Εισαγωγή ενός αρχείου .xls ή .xlsx. Δείτε την πρώτη σελίδα για περισσότερες λεπτομέρειες για την παραμετροποίηση της εντολής. Παρόμοιες εντολές: read xls() and read xlsx().

read\_excel("excel\_file.xlsx")

#### Εισαγωγή Δεδομένων από Υπολογιστικά Φύλλα



read\_excel(path, sheet = NULL) Ορισμός φύλλου μέσω του ονόματος ή της θέσης του φύλλου.

read excel(path, sheet = 1) read excel(path, sheet = "s1")



ABCDE

ABCDE

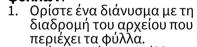
s1 s2 s3

ABCDE

excel\_sheets(path) Διάνυσμα ονομάτων φύλλων.

excel sheets("excel file.xlsx")

#### Για την εισαγωγή πολλαπλών φύλλων:



2. Ορίστε ως όνομα φύλλου το διάνυσμα του προηγούμενου βήματος.

Χρησιμοποιείστε την purrr::map\_dfr() για να εισάγετε πολλαπλά αρχεία σε ένα πλαίσιο δεδομένων

path <- "your\_file\_path.xlsx" path %>% excel sheets() %>% set names() %>% map\_dfr(read\_excel, path = path)

#### Προδιαγραφές στηλών με το πακέτο readxl

Με το πακέτο readxl υπάρχει η δυνατότητα να καθορίσουμε τον τύπο δεδομένων της κάθε στήλης κατά την εισαγωγή δεδομένων στην R.

Με την παράμετρο col\_types της εντολής read\_excel() ορίζουμε τον τύπο δεδομένων των στηλών.

Μαντέψτε τους τύπους δεδομένων των στηλών Μαντέψτε τους τύπους στηλών με την εντολή read excel(), η οποία διαβάζει τις πρώτες 1000 γραμμές των δεδομένων. Αυτό αλλάζει με την παράμετρο

read excel(path, guess\_max = Inf)

Ορίστε τον ίδιο τύπο στηλών σε όλες τις στήλες, π.χ. χαρακτήρας

read\_excel(path, col\_types = "text")

#### Ορίστε τον τύπο κάθε στήλης ξεχωριστά

read excel( col\_types = c("text", "guess", "guess", "numeric")

#### Τύποι Στηλών

	αριθμός numeric		ημερομηνία date	λίστα list
TRUE	2	hello	1947-01-08	hello
FALSE	3.45	world	1956-10-21	1

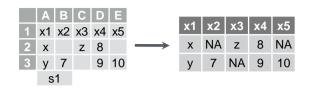
- skip
- logical
- date

- guess
- numeric list
  - text

Χρησιμοποιείστε την λίστα (list) για στήλες που περιέχουν πολλαπλούς τύπους δεδομένων. Δείτε τα πακέτα tidyr και purrr για δεδομένα λίστας σε στήλες.

### με το πακέτο googlesheets4

#### Εισανωνή υπολονιστικών φύλλων



read\_sheet(ss, sheet = NULL, range = NULL) Εισαγωγή ενός υπολογιστικού φύλλου από ένα URL, ή ID, ή από ένα dribble από το πακέτο googledrive. Δείτε την πρώτη Με την παράμετρο col\_types της εντολής σελίδα για περισσότερες λεπτομέρειες για την παραμετροποίηση της εντολής. Παρομοίως range\_read().

#### Μεταδεδομένα Φύλλων

URLs της μορφής:

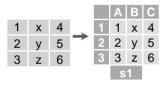
https://docs.google.com/spreadsheets/d/ SPREADSHEET ID/edit#gid=SHEET ID

gs4\_get(ss) Μεταδεδομένα υπολογιστικού φύλλου.

gs4\_find(...) Μεταδεδομένα για όλα τα αρχεία υπολογιστικών φύλλων.

sheet\_properties(ss) Πλειάδα (tibble) ιδιοτήτων για κάθε φύλλο εργασίας. Παρομοίως sheet\_names().

#### Αποθήκευση υπολογιστικών φύλλων



ABCD

write sheet(data, ss = NULL, sheet = NULL) Αποθήκευση ενός πλαισίου δεδομένων σε ένα νέο ή υπάρχον Φύλλο.

gs4\_create(name, ..., sheets = NULL) Δημιουργία ενός νέου Φύλλου με ένα διάνυσμα ονομάτων, ένα πλαίσιο δεδομένων ή μια λίστα πλαισίων δεδομένων.

sheet\_append(ss, data, sheet = 1) Προσθήκη σειρών στο τέλος ενός φύλλου εργασίας.

### Προδιαγραφές στηλών με το πακέτο googlesheets4

googlesheets

Με το πακέτο googlesheets4 υπάρχει η δυνατότητα να καθορίσουμε τον τύπο δεδομένων της κάθε στήλης κατά την εισαγωγή δεδομένων στην R.

read\_sheet()/ range\_read() ορίζουμε τον τύπο δεδομένων των

Μαντέψτε τους τύπους δεδομένων των στηλών Μαντέψτε τους τύπους στηλών με την εντολή read\_sheet()/range\_read() , η οποία διαβάζει τις πρώτες 1000 γραμμές των δεδομένων. Αυτό αλλάζει με την παράμετρο guess\_max. read\_sheet(path, guess\_max = Inf)

Ορίστε τον ίδιο τύπο στηλών σε όλες τις στήλες, π.χ. χαρακτήρας

read\_sheet(path, col\_types = "c")

Ορίστε τον τύπο κάθε στήλης ξεχωριστά

# τύποι στηλών: skip, guess, integer, logical, character read sheets(ss, col types = "?ilc")

#### Τύποι Στηλών

λογικοί	αριθμός	κείμενο	ημερομηνία	λίστα		
logical	numeric	text	date	list		
TRUE	2	hello	1947-01-08	hello		
FALSE	3.45	world	1956-10-21	1		
<ul> <li>skip - "_" or "-"</li> <li>guess - "?"</li> <li>logical - "l"</li> <li>integer - "i"</li> <li>datetime - "</li> <li>character - "</li> <li>list-column</li> <li>cell - "C" Re</li> </ul>						
<ul><li>doubl</li><li>nume</li></ul>	le - "d"	list of raw cell data.				

Χρησιμοποιείστε την λίστα (list) για στήλες που περιέχουν πολλαπλούς τύπους δεδομένων. Δείτε τα πακέτα tidyr και purrr για δεδομένα λίστας σε στήλες.

### Προδιαγραφές κελιού με τα πακέτα readxl και googlesheets4

2 3 4

NA y z

Για συναρτήσεις αποθήκευσης δεδομένων σε αρχεία Excel, δείτε τα πακέτα:

- openxlsx
- writexl

Για δεδομένα τα οποία δεν είναι σε μορφή Excel πίνακα, δείτετα πακέτα:

tidyxl

Με την παράμετρο range των εντολών readxl::read excel() ή googlesheets4::read\_sheet() γίνεται η εισαγωγή ενός υποσυνόλου κελιών από ένα υπολονιστικό φύλλο.

АВС

1 x1 x2 x3

→ 2 1 x 4

read excel(path, range = "Sheet1!B1:D2") read\_sheet(ss, range = "B1:D2")

2 y 5

3 z 6

Επίσης με την παράμετρο **range** μπορούμε να χρησιμοποιήσουμε περισσότερες επιλογές για το κελί με τις εντολές cell limits(), cell rows(), cell cols(), and anchored().

#### Λειτουργίες σε αρχεία

date - "D"

Το πακέτο googlesheets4 προσφέρει επιπλέον λειτουργίες διαχείρισης υπολογιστικών φύλλων (π.χ. ορισμός πλάτος στήλης). Περισσότερες λεπτομέρειες στο googlesheets4.tidyverse.org.

Για πλήρεις διαδικασίες διαχείρισης αρχεχίων (π.χ. μετονομασία, κοινή χρήση, μετακίνηση), δείτε περισσότερες λεπτομέρειες στο πακέτο googledrive στο googledrive.tidyverse.org.



Άλλα χρήσιμα πακέτα Excel





3 6 7 9 10

s1