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**

A|B|C
1|2|3
4|5|NA

A	B	C
1	2	3
4	5	NA

read_delim("file.txt", delim = "|"**)** Εισαγωγή αρχείων με οποιονδήποτε οριοθέτη (delimiter), ο οποίος εάν δεν έχει καθοριστεί η εντολή θα ορίσει έναν αυτόματα.
Για να δημιουργήσετε ένα αρχείο file.txt μπορείτε να εκτελέσετε:
write_file("A|B|C\n1|2|3\n4|5|NA", file = "file.txt")

A,B,C
1,2,3
4,5,NA

A	B	C
1	2	3
4	5	NA

read_csv("file.csv"**)** Εισαγωγή αρχείου οριοθετημένου με κόμμα με δεκαδικά σημεία τελείας.
write_file("A,B,C\n1,2,3\n4,5,NA", file = "file.csv")

A;B;C
1,5;2;3
4,5;5;NA

A	B	C
1.5	2	3
4.5	5	NA

read_csv2("file2.csv"**)** Εισαγωγή αρχείου οριοθετημένου με ερωτηματικό με δεκαδικά σημεία τελείας.
write_file("A;B;C\n1,5;2;3\n4,5;5;NA", file = "file2.csv")

A B C
1 2 3
4 5 NA

A	B	C
1	2	3
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

A	B	C
1	2	3
4	5	NA

Χωρίς κεφαλίδα (ονόματα στηλών)
read_csv("file.csv", col_names = FALSE)

x	y	z
A	B	C
1	2	3
4	5	NA

Ορισμός κεφαλίδων (ονομάτων στηλών)
read_csv("file.csv",
col_names = c("x", "y", "z"))

Εισαγωγή πολλών αρχείων σε έναν μόνο πίνακα
read_csv(c(“f1.csv”, “f2.csv”, “f3.csv"),
id = "origin_file")

1	2	3
4	5	NA

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

A	B	C
1	2	3

Εισαγωγή ενός υποσυνόλου γραμμών
read_csv("file.csv", n_max = 1)

A	B	C
NA	2	3
4	5	NA

Εισαγωγή με ορισμό τιμών ως τιμές που λείπουν
read_csv("file.csv", na = c("1"))

A;B;C
1,5;2;3,0

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**)**

A	B	C
1	2	3
4	5	NA

A,B,C
1,2,3
4,5,NA

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 εισαγωγή 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
- **jsonlite** - json
- **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

```
spec(x)
# cols(
#   age = col_integer(),
#   sex = col_character(),
#   earn = col_double()
# )
```

Η age είναι ακέραια μεταβλητή
Η sex είναι χαρακτήρας
Η earn είναι αριθμητική μεταβλητή

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 ρεντολή 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(
  file,
  col_type = list(.default = col_double())
)
```

Ορισμός μέσω συνάρτησης ή συντομογραφίας

```
read_csv(
  file,
  col_type = list(x = col_double(), y = "l", z = "_")
)
```

Ορισμός μέσω χαρακτήρων συντομογραφίας

```
# col types: skip, guess, integer, logical, character
read_csv(
  file,
  col_type = "_?ilc"
)
```


• readxl.tidyverse.org and googlesheets4.tidyverse.org • readr 2.0.0 • readxl 1.3.1 • googlesheets4 1.0.0 • Updated: 2021-08

Μετάφραση: Νικόλαος Κουπίδης • koupnikos@gmail.com
Κλεάνθης Κουπίδης • kleanthis.koupidis@gmail.com

Εισαγωγή Υπολογιστικών Φύλλων με το πακέτο readxl

Εισαγωγή Αρχείων Excel

	A	B	C	D	E
1	x1	x2	x3	x4	x5
2	x		z	8	
3	y	7		9	10
	s1				

→

x1	x2	x3	x4	x5
x	NA	z	8	NA
y	7	NA	9	10

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

A	B	C	D	E
s1	s2	s3		

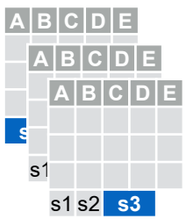
read_excel(path, **sheet = NULL)** Ορισμός φύλλου μέσω του ονόματος ή της θέσης του φύλλου.

read_excel(path, sheet = 1)
read_excel(path, sheet = "s1")

s1	s2	s3

excel_sheets(path**)** Διάνυσμα ονομάτων φύλλων.

excel_sheets("excel_file.xlsx")



Για την εισαγωγή **πολλαπλών φύλλων:**

1. Ορίστε ένα διάνυσμα με τη διαδρομή του αρχείου που περιέχει τα φύλλα.
2. Ορίστε ως όνομα φύλλου το διάνυσμα του προηγούμενου βήματος.
3. 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 **guess_max**.

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(
  path,
  col_types = c("text", "guess", "guess","numeric")
)
```

Τύποι Στηλών

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

- skip
- guess
- logical
- numeric
- text
- date
- list

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

Εισαγωγή υπολογιστικών φύλλων

	A	B	C	D	E
1	x1	x2	x3	x4	x5
2	x		z	8	
3	y	7		9	10
	s1				

→

x1	x2	x3	x4	x5
x	NA	z	8	NA
y	7	NA	9	10

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 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()**.

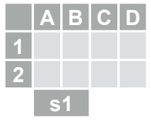
Αποθήκευση υπολογιστικών φύλλων

1	x	4
2	y	5
3	z	6

→

	A	B	C
1	1	x	4
2	2	y	5
3	3	z	6
	s1		

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.

x1	x2	x3
2	y	5
3	z	6

→

	A	B	C
1	x1	x2	x3
2	1	x	4
3	2	y	5
4	3	z	6
	s1		

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

Προδιαγραφές στηλών με το πακέτο googlesheets4

Με το πακέτο googlesheets4 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_sheet()/ range_read()** ορίζουμε τον τύπο δεδομένων των στηλών.

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

- skip - "_" or "-"
- guess - "?"
- logical - "l"
- integer - "i"
- double - "d"
- numeric - "n"
- date - "D"
- datetime - "T"
- character - "c"
- list-column - "L"
- cell - "C" Returns 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.

Άλλα χρήσιμα πακέτα Excel

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**

Προδιαγραφές κελιού με τα πακέτα readxl και googlesheets4

	A	B	C	D	E
1	1	2	3	4	5
2	x		y	z	
3	6	7		9	10
	s1				

→

2	3	4
NA	y	z

Με την παράμετρο **range** των εντολών **readxl::read_excel()** ή **googlesheets4::read_sheet()** γίνεται η εισαγωγή ενός υποσυνόλου κελιών από ένα υπολογιστικό φύλλο.

read_excel(path, range = "Sheet1!B1:D2")
read_sheet(ss, range = "B1:D2")

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

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**.

• readxl.tidyverse.org and googlesheets4.tidyverse.org • readr 2.0.0 • readxl 1.3.1 • googlesheets4 1.0.0 • Updated: 2021-08

Μετάφραση: Νικόλαος Κουπίδης • koupnikos@gmail.com
Κλεάνθης Κουπίδης • kleanthis.koupidis@gmail.com