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().

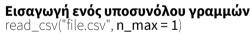
read_fwf("file.tsv", fwf_widths(c(2, 2, NA))) Εισαγωγή αρχείων σταθερού πλάτους. write file("A\tB\tC\n1\t2\t3\n4\t5\tNA\n", file = "file.tsv")

Χρήσιμες Παράμετροι Εισαγωγής Δεδομένων

Α	В	С	Χωρίς κεφαλίδα (ονόματα στηλών)
1	2	3	read_csv("file.csv", col_names = FALSE)
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	
	→		Εισαγωγή πολλών αρχείων σε έναν μόνο πίνακα read_csv(c("f1.csv", "f2.csv", "f3.csv"), id = "origin_file")



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





1 2 3

Εισαγωγή με ορισμό τιμών ως τιμές που λείπουν

4 5 NA

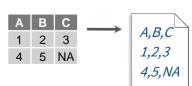
NA 2 3 read_csv("file.csv", na = c("1"))



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 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

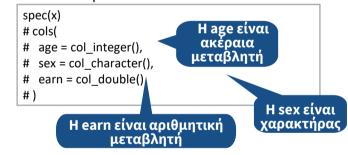
Δοκιμάστε ένα από τα παρακάτω πακέτα για να εισαγάγετε άλλους τύπους αρχείων:

- haven αρχεία SPSS, Stata, και SAS
- **DBI** βάσεις δεδομένων
- 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 προδιανραφών 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

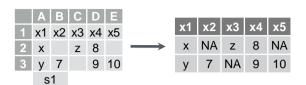
col types: skip, guess, integer, logical, character read_csv(file. col type = "?ilc"



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



s1 s2 s3

ABCDE

ABCDE

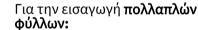
s1 s2 s3

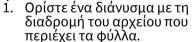
ABCDE

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")







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(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

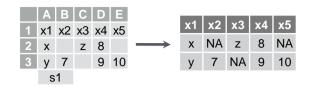
- guess
- numeric list

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:

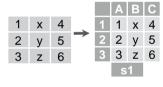
https://docs.google.com/spreadsheets/d/ **SPREADSHEET ID**/edit#gid=**SHEET ID**

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 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

	αριθμός numeric	κείμενο text	ημερομηνία date	λίστα list
TRUE	2	hello	1947-01-08	hello
FALSE	3.45	world	1956-10-21	1
skip -guesslogica	- "T" r - "c" nn - "L"			
 integer 	er - "i"	cell - "C" Returns		
 doubl 	le - "d"	list of raw cell		

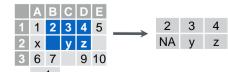
date - "D"

numeric - "n"

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

data.

Προδιαγραφές κελιού με τα πακέτα readxl και googlesheets4



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

ABC

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

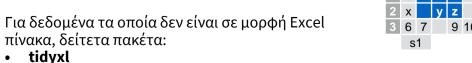
Επίσης με την παράμετρο **range** μπορούμε να χρησιμοποιήσουμε περισσότερες επιλογές για το κελί με τις εντολές cell limits(), cell rows(), cell cols(), and anchored().

Λειτουργίες σε αρχεία

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.





Excel, δείτε τα πακέτα:

openxlsx

writexl