

Data import with the tidyverse :: CHEAT SHEET

Read Tabular Data with readr

read_*(file, col_names = TRUE, col_types = NULL, col_select = NULL, id = NULL, locale = Inf, skip = 0, na = c("", "NA"), guess_max = min(1000, n_max), show_col_types = TRUE) See ?read_delim

read_delim("file.txt", delim = "|") Read files with any delimiter. If no delimiter is specified, it will automatically guess.
To make file.txt, run: write_file("A|B|C\n1|2|3\n4|5|NA", file = "file.txt")

read_csv("file.csv") Read a comma delimited file with period decimal marks.
write_file("A,B,C\n1,2,3\n4,5,NA", file = "file.csv")

read_csv2("file2.csv") Read semicolon delimited files with comma decimal marks.
write_file("A;B;C\n1;5;2;3\n4;5;NA", file = "file2.csv")

read_tsv("file.tsv") Read a tab delimited file. Also **read_table**().
read_fwf("file.tsv", fwf_widths(c(2, 2, NA))) Read a fixed width file.
write_file("A|B|C\n1|2|3\n4|5|tNA", file = "file.tsv")

USEFUL READ ARGUMENTS

No header
read_csv("file.csv", col_names = FALSE)

Provide header
read_csv("file.csv",
col_names = c("x", "y", "z"))

Read multiple files into a single table
read_csv(c("f1.csv", "f2.csv", "f3.csv"),
id = "origin_file")

Skip lines
read_csv("file.csv", skip = 1)

Read a subset of lines
read_csv("file.csv", n_max = 1)

Read values as missing
read_csv("file.csv", na = c("1"))

Specify decimal marks
read_delim("file2.csv", locale =
locale(decimal_mark = ","))

Save Data with readr

write_*(x, file, na = "NA", append, col_names, quote, escape, eol, num_threads, progress)

write_delim(x, file, delim = "|") Write files with any delimiter.

write_csv(x, file) Write a comma delimited file.

write_csv2(x, file) Write a semicolon delimited file.

write_tsv(x, file) Write a tab delimited file.

One of the first steps of a project is to import outside data into R. Data is often stored in tabular formats, like csv files or spreadsheets.

The front page of this sheet shows how to import and save text files into R using **readr**.

The back page shows how to import spreadsheet data from Excel files using **readxl** or Google Sheets using **googlesheets4**.

Column Specification with readr

Column specifications define what data type each column of a file will be imported as. By default readr will generate a column spec when a file is read and output a summary.

spec(x) Extract the full column specification for the given imported data frame.

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

earn is a double (numeric)

age is an integer

sex is a character

COLUMN TYPES

Each column type has a function and corresponding string abbreviation.

- **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()** - "s", "_"
- **col_guess()** - "g"

OTHER TYPES OF DATA
Try one of the following packages to import other types of files:

- **haven** - SPSS, Stata, and SAS files
- **DBI** - databases
- **jsonlite** - json
- **xml2** - XML
- **httr** - Web APIs
- **rvest** - HTML (Web Scraping)
- **readr::read_lines()** - text data

USEFUL COLUMN ARGUMENTS

Hide col spec message

read_* (file, show_col_types = FALSE)

Select columns to import

Use names, position, or selection helpers.
read_* (file, col_select = c(age, earn))

Guess column types

To guess a column type, read_* () looks at the first 1000 rows of data. Increase with **guess_max**.
read_* (file, guess_max = Inf)

DEFINE COLUMN SPECIFICATION

Set a default type

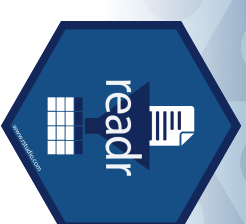
read_csv(
file,
col_type = list(default = col_double()))

Use column type or string abbreviation

read_csv(
file,
col_type = list(x = col_double(), y = "i", z = "s"))

Use a single string of abbreviations

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



Import Spreadsheets

with readxl

READ EXCEL FILES

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

s1

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

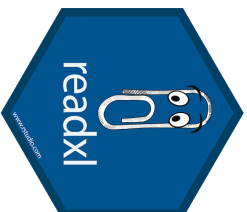
read_excel(path, sheet = NULL, range = NULL)

Read a .xls or .xlsx file based on the file extension.

See front page for more read arguments. Also

read_xls() and **read_xlsx()**.

read_excel("excel_file.xlsx")



with googlesheets4

READ SHEETS

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

s1

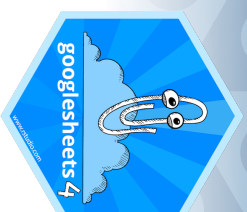
	x1	x2	x3	x4	x5
	x	NA	z	8	NA
y	7	NA	9	10	

read_sheet(ss, sheet = NULL, range = NULL)

Read a sheet from a URL, a Sheet ID, or a dribble

from the googledrive package. See front page for

more read arguments. Same as **range_read()**.



READXL COLUMN SPECIFICATION

Column specifications define what data type each column of a file will be imported as.

Use the **col_types** argument of **read_excel()** to set the column specification.

Guess column types

To guess a column type, **read_excel()** looks at the first 1000 rows of data. Increase with the **guess_max** argument.

read_excel(path, guess_max = Inf)

Set all columns to same type, e.g. character

read_excel(path, col_types = "text")

Set each column individually

read_excel()

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

COLUMN TYPES

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

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

Use **list** for columns that include multiple data types. See **tidyr** and **purrr** for list-column data.

SHEETS METADATA

URLs are in the form:

https://docs.google.com/spreadsheets/d/SPREADSHEET_ID/edit#gid=SHEET_ID

gs4_get(ss) Get spreadsheet meta data.

gs4_find(...) Get data on all spreadsheet files.

sheet_properties(ss) Get a tibble of properties for each worksheet. Also **sheet_names()**.

WRITE SHEETS

write_sheet(data, ss = NULL, sheet = NULL)
Write a data frame into a new or existing Sheet.

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

s1

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

gs4_create(name, ...)
sheets = NULL) Create a new Sheet with a vector of names, a data frame, or a (named) list of data frames.

sheet_append(ss, data, sheet = 1) Add rows to the end of a worksheet.

	x1	x2	x3
1	x1	x2	x3
2	y	5	
3	z	6	

s1

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

OTHER USEFUL EXCEL PACKAGES

For functions to write data to Excel files, see:

- **openxlsx**
- **writexl**

For working with non-tabular Excel data, see:

- **tidyxl**

CELL SPECIFICATION FOR READXL AND 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

Use the **range** argument of **readxl::read_excel()** or **googlesheets4::read_sheet()** to read a subset of cells from a sheet.

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

read_sheet(ss, range = "B1:D2")

Also use the range argument with cell specification functions **cell_limits()**, **cell_rows()**, **cell_cols()**, and **anchored()**.

FILE LEVEL OPERATIONS

googlesheets4 also offers ways to modify other aspects of Sheets (e.g. freeze rows, set column width, manage work/sheets). Go to

googlesheets4.tidyverse.org to read more.

For whole-file operations (e.g. renaming, sharing, placing within a folder), see the tidyverse package **googledrive** at **googledrive.tidyverse.org**.