Importação de dados com tidyverse : : **FOLHA DE REFERÊNCIA**

Ler dados tabulados com 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**)** Ver **?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 = "|") Lê arquivos com qualquer delimitador. Se o delimitador não for especificado, tenta advinhar automaticamente.
Para criar file.txt, execute: 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") Lê arquivo separado com vírgula com ponto como separador decimal.
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") Lê arquivo separado por ponto-e-vírgula com vírgula como separador decimal.
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") Lê arquivo separado com tab. Ver também read_table().
read_fwf("file.tsv", fwf_widths(c(2, 2, NA))) Lê arquivo com largura fixa.
write_file("A\tB\tC\n1\t2\t3\n4\t5\tNA\n", file = "file.tsv")

ARGUMENTOS ÚTEIS DE LEITURA

A	B	C
1	2	3
4	5	NA

Sem cabeçalho
read_csv("file.csv", col_names = FALSE)

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

Fornece cabeçalho
read_csv("file.csv",
col_names = c("x", "y", "z"))

Lê vários arquivos em um única tabela
read_csv(c("f1.csv", "f2.csv", "f3.csv"),
id = "origin_file")

1	2	3
4	5	NA

Pula linhas
read_csv("file.csv", skip = 1)

A	B	C
1	2	3

Lê subconjunto de linhas
read_csv("file.csv", n_max = 1)

A	B	C
NA	2	3
4	5	NA

Lê valor como ausente (NA)
read_csv("file.csv", na = c("1"))

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

Especifica separador decimal
read_delim("file2.csv", locale =
locale(decimal_mark = ","))

Salvar dados com 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 = " ") Grava arquivos com delimitador.

write_csv(x, file) Grava arquivo separado por vírgula.

write_csv2(x, file) Grava arquivo separado por ponto-e-vírgula.

write_tsv(x, file) Grava arquivo separado por tab.

Um dos primeiros passos de um projeto é importar dados externos para o R. Os dados são frequentemente armazenados em formatos tabulados como arquivos .csv ou planilhas.

A página da frente desta folha de resumo mostra como importar e salvar arquivos texto usando o pacote readr.

O verso mostra como importar planilhas do Excel usando readxl ou planilhas do Google Sheets usando o googlesheets4.

OUTROS FORMATOS DE DADOS

Experimente um desses pacotes para importar outros formatos de dados:

- haven – arquivos SPSS, Stata e SAS
- DBI – bando de dados
- jsonlite - json
- xml2 - XML
- httr - Web APIs
- rvest - HTML (Web Scraping)
- readr::read_lines() – dados texto

Especificação de Coluna com readr

Especificação de coluna define qual o tipo de dado cada coluna de um arquivo será importada. Por padrão readr gera uma especificação quando o arquivo é importado e um resumo na saída.

spec(x) Extrai a especificação de coluna completa de um data frame importado.

```
spec(x)
# cols(
#   idade  = col_integer(),
#   sexo = col_character(),
#   salário = col_double()
# )
```

idade é um inteiro

sexo é um caractere

salário é um double (numerico)

TIPOS DE COLUNAS

Cada tipo de coluna tem uma função e uma string de abreviação correspondente.

- 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() - "?"

ARGUMENTOS ÚTEIS PARA COLUNAS

Esconde as mensagem de especificação
read_*(file, show_col_types = FALSE)

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

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

DEFINE ESPECIFICAÇÃO DAS COLUNAS

Define o tipo padrão

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

Usa um tipo ou a string de abreviação

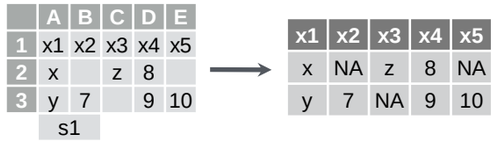
```
read_csv(
  file,
  col_type = list(x = col_double(), y = "l", z = "_")
)
```

Usa um única string de abreviação

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

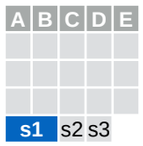
Importando Planilhas com readxl

LER ARQUIVOS EXCEL



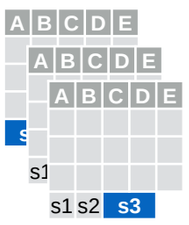
read_excel(path, sheet = NULL, range = NULL) Lê um arquivo .xls ou .xlsx baseado na extensão. Ver primeira página para mais argumentos de leitura. Ver read_xls() e read_xlsx().
read_excel("excel_file.xlsx")

LER PLANILHAS



read_excel(path, sheet = NULL) Especifica qual planilha ler, por nome ou posição.
read_excel(path, sheet = 1)
read_excel(path, sheet = "s1")

excel_sheets(path) Pega os nomes das planilhas em um vetor.
excel_sheets("excel_file.xlsx")



Para ler várias planilhas:
1. Pegue do arquivo um vetor com nomes das planilhas.
2. Defina o vetor de nomes como nomes das planilhas.
3. Use purrr::map_dfr() para ler vários arquivos e gerar um único data frame.

```
path <- "caminho_do_arquivo.xlsx"
path %>% excel_sheets() %>%
  set_names() %>%
  map_dfr(read_excel, path = path)
```

ESPECIFICAÇÃO DE COLUNA - READXL

Especificação de coluna define o tipo de dado que cada coluna do arquivo terá após importada.

Use o argumento col_types da read_excel() para definiar as especificações das colunas.

Adivinhar tipos de colunas
Para adivinhar o tipo da coluna, read_ excel() lê as 100 primeiras linhas. Aumente com o argumento guess_max.
read_excel(path, guess_max = Inf)

Define todas as colunas com o mesmo tipo, ex. caractere
read_excel(path, col_types = "text")

Define cada coluna individualmente

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

TIPOS DE COLUNAS

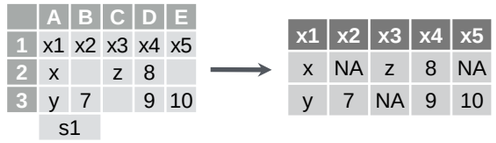
logical	numeric	text	date	list
TRUE	2	hello	1947-01-08	hello
FALSE	3.45	world	1956-10-21	1

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

Use list para colunas que tem múltiplos tipos de dados. Veja sobre colunas de lista em tidyr e purrr.

com googlesheets4

LER PLANILHAS GOOGLE



read_sheet(ss, sheet = NULL, range = NULL)
Lê um arquivo com URL, um ID, ou um objeto drible do pacote googledrive. Veja página da frente para mais argumentos de leitura. Mesmo que range_read().

METADATA DAS PLANILHAS

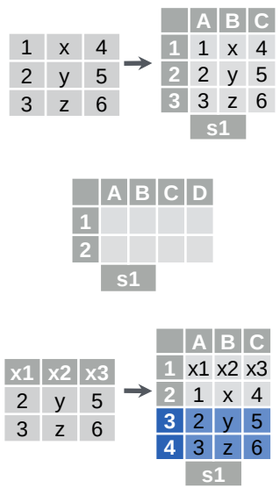
URLs estão na seguinte forma:
https://docs.google.com/spreadsheets/d/ID_ARQUIVO/edit#gid=ID_PLANILHA

gs4_get(ss) Lê o metadado da planilha.

gs4_find(...) Lê dados de todos os arquivos de planilhas.

sheet_properties(ss) Retorna um tibble com propriedades de cada planilha. Veja também sheet_names().

GRAVAR PLANILHAS DO GOOGLE



write_sheet(data, ss = NULL, sheet = NULL) Wgrava um data frame em um planilha nova ou já existente.

gs4_create(name, ..., sheets = NULL) Cria uma nova planilha com um vetor de nomes, um data frame ou uma lista de data frame nomeada.

sheet_append(ss, data, sheet = 1) Adiciona uma linha ao final da planilha.

ESPECIFICAÇÃO DE COLUNAS - GOOGLESHEETS4

Especificação de coluna define o tipo de dado que cada coluna do arquivo terá após importada.

Use o argumento col_types da read_sheet()/range_read() para definir as especificações das colunas.

Adivinhar tipos de colunas
Para adivinhar o tipo da coluna read_sheet()/range_read() lê as peimrias 100 linhas. Aumente com guess_max.
read_sheet(path, guess_max = Inf)

Define todas as colunas com o mesmo tipo, ex. caractere
read_sheet(path, col_types = "c")

Define cada coluna individualmente
# col types: skip, guess, integer, logical, character
read_sheets(ss, col_types = "_?ilc")

TIPOS DE COLUNAS

l	n	c	D	L
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.

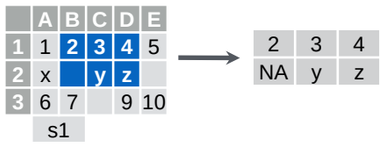
Use list para colunas que tem múltiplos tipos de dados. Veja sobre colunas de lista em tidyr e purrr.

OUTROS PACOTES ÚTEIS PARA

Para funções de gravar dados em arquivos Excel, veja:
- openxlsx
- writexl

Para trabalhar com dados do Excel em formato não-tabular, veja:
- tidyxl

ESPECIFICAÇÃO DE CÉLULAS PARA READXL E GOOGLESHEETS4



Use o argumento range da readxl::read_excel() ou googlesheets4::read_sheet() para ler um subconjunto de células de uma planilha.
read_excel(path, range = "Sheet1!B1:D2")
read_sheet(ss, range = "B1:D2")

Use também o argumento range com funções de especificação de células cell_limits(), cell_rows(), cell_cols() e anchored().

OPERAÇÕES EM NÍVEL DE ARQUIVO

googlesheets4 também oferece várias formas de modificar outros aspectos das planilhas (ex. congelar linhas, definir largura das colunas, gerenciar planilhas, etc). Veja googlesheets4.tidyverse.org para maiores informações.

Para operações de arquivos (ex. renomear, compartilhar, mover para outra pasta, etc), veja mais sobre o pacote googledrive do tidyverse em: googledrive.tidyverse.org.