

读取数据

王敏杰

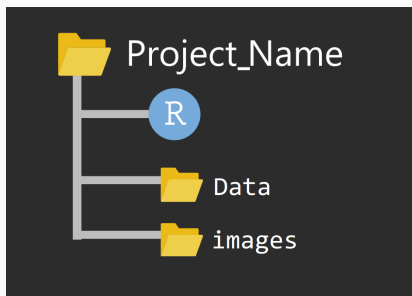
2023 年 3 月 16 日

北清迈大学创新学院

项目管理

项目管理

- 把项目所需的文件（代码、数据、图片等），放在一个文件夹里



- 放在一个没有中文和空格的路径下

文件夹命名

推荐我自己的文件夹命名习惯 (项目名 + 日期), 注意这里不要有中文和空格, 比如下面风格的就比较好

- *homework20220618*
- *project20220618*
- *Emotional_experiment20220618*
- *R_ladies_20220520*
- *R_You_with_Me*

数据读取

读取数据

R 语言提供了很多读取数据的函数。

| 文件格式 | R 函数 |
|----------------------|---|
| .txt | <code>read.table()</code> |
| .csv | <code>read.csv()</code> and <code>readr::read_csv()</code> |
| .xls and .xlsx | <code>readxl::read_excel()</code> and <code>openxlsx::read.xlsx()</code> |
| .sav(SPSS files) | <code>haven::read_sav()</code> and <code>foreign::read.spss()</code> |
| .Rdata or rda | <code>load()</code> |
| .rds | <code>readRDS()</code> and <code>readr::read_rds()</code> |
| .dta | <code>haven::read_dta()</code> and <code>haven::read_stata()</code> |
| .sas7bdat(SAS files) | <code>haven::read_sas()</code> |
| Internet | <code>download.file()</code> |

The tidyverse packages to import your data

readr

- `readr` package:
 - `read_csv()`: comma separated (,)
 - `read_csv2()`: separated (;)
 - `read_tsv()`: tab separated
 - `read_delim()`: general delimited files, auto-guesses delimiter
 - `read_fwf()`: fixed width files
 - `read_table()`: columns separated by white-space(s)

readxl

To import excel files (`.xls` and `.xlsx`):

- `read_excel()`
 - `read_xls()`
 - `read_xlsx()`



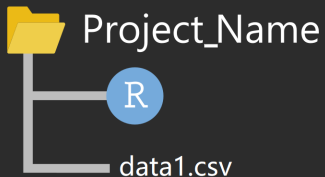
haven

- `read_sas()` for SAS
- `read_sav()` for SPSS
- `read_dta()` for Stata



¹图片来源 <https://rworkshop.uni.lu/>

文件路径，推荐使用相对路径

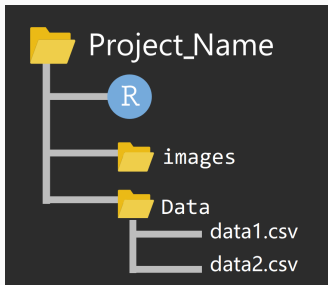


```
d <- read_csv("./data1.csv")
```

or

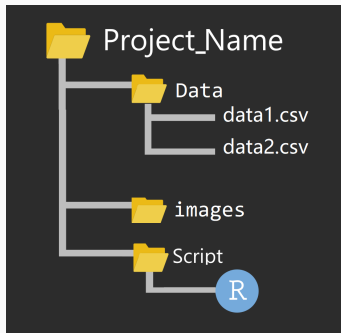
```
d <- read_csv("data1.csv")
```


文件路径



```
d <- read_csv("./Data/data1.csv")
```

文件路径



```
d <- read_csv("../Data/data1.csv")
```

范例

```
library(readr)
wages <- read_csv("./data/wages.csv")
wages
#> # A tibble: 1,379 x 6
#>       earn height sex    race    ed    age
#>   <dbl>   <dbl> <chr>  <chr>  <dbl> <dbl>
#> 1 79571.   73.9 male   white   16    49
#> 2 96397.   66.2 female white   16    62
#> 3 48711.   63.8 female white   16    33
#> 4 80478.   63.2 female other    16    95
#> 5 82089.   63.1 female white   17    43
#> 6 15313.   64.5 female white   15    30
#> 7 47104.   61.5 female white   12    53
#> 8 50960.   73.3 male    white   17    50
```

变量类型

| Variable Type | Long form | Abbreviation |
|-----------------------------|-----------------------------|--------------|
| Logical (TRUE/FALSE) | col_logical() | l |
| Integer | col_integer() | i |
| Double | col_double() | d |
| Character | col_character() | c |
| Factor (nominal or ordinal) | col_factor(levels, ordered) | f |
| Date | col_date(format) | D |
| Time | col_time(format) | t |
| Date & Time | col_datetime(format) | T |
| Number | col_number() | n |
| Don't import | col_skip() | - |
| Default Guessing | col_guess() | ? |

指定类型

```
wages <- read_csv(  
  file = "../data/wages.csv",  
  col_types = list(  
    col_double(),  
    col_double(),  
    col_character(),  
    col_character(),  
    col_character(),  
    col_guess()  
  )  
)
```

```
#> # A tibble: 1,379 x 6  
#>   earn height sex   race   ed   age  
#>   <dbl> <dbl> <chr> <chr> <chr> <dbl>  
#> 1 79571.   73.9 male  white  16    49  
#> 2 96397.   66.2 female white  16    62
```

习题

读取 demo_data 文件夹下 kidiq.RDS 文件

变量含义：

| 变量 | 含义 |
|-----------|----------|
| kid_score | 小孩考试分数 |
| mom_hs | 母亲是否完成高中 |
| mom_iq | 母亲 IQ 值 |
| mom_age | 母亲年龄 |

请说出数据框中每一列的变量类型