

读取数据

王敏杰

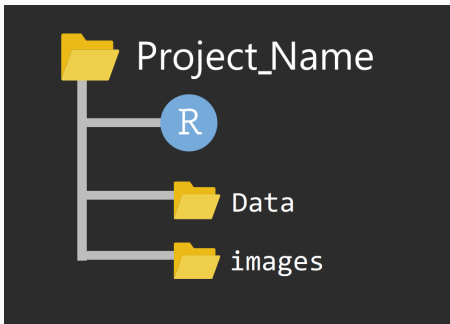
2022 年 4 月 5 日

四川师范大学

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



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

- homework20201014
- project20201014
- Emotional_experiment20201014

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>

Tidyverse 各种宏包

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



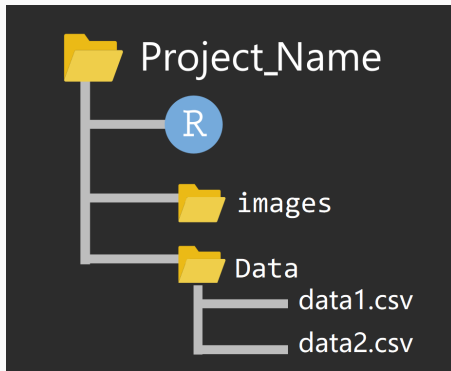
图片来源 <https://rworkshop.uni.lu/>

文件路径

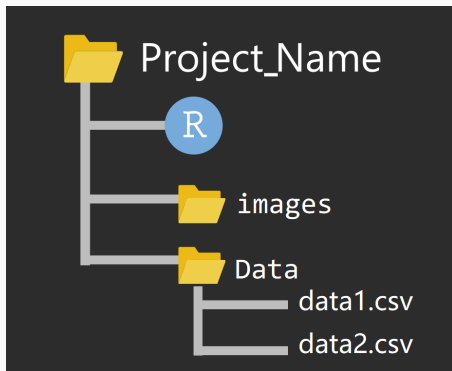
在读取文件时，路径的写法有如下方式（以图像文件“a.jpg”为例）

- 如果以当前项目文件为中心，图片在当前目录，它的路径 “./a.jpg”
- 如果以当前项目文件为中心，图片在下一层目录的 images 文件夹 “./images/a.jpg”
- 如果以当前项目文件为中心，图片在上一层目录下，它的路径 “../a.jpg”
- 如果以当前项目文件为中心，图片在上一层目录的 images 文件夹，它的路径 “../images/a.jpg”
- 从根目录出发，访问 D 盘的 images 文件“b.jpg” 图片，它的路径 “D:/images/b.jpg”

文件路径

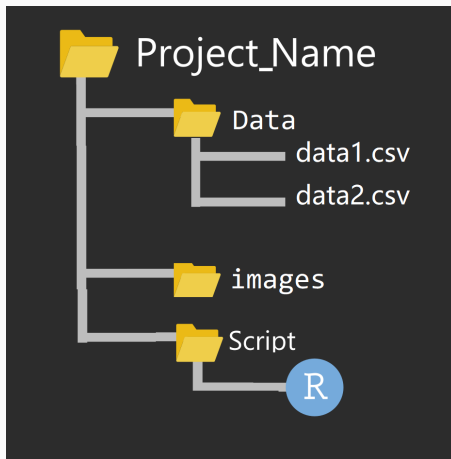


文件路径

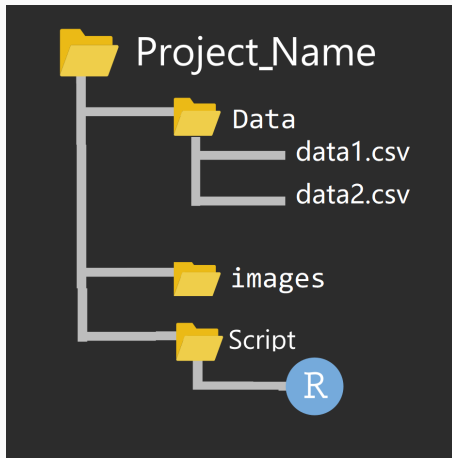


```
d <- read_csv("../Data/data1.csv")
```

文件路径



文件路径



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

范例

```
library(readr)
wages <- read_csv("../data/wages.csv")
head(wages, 6)

#> # A tibble: 6 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
```

范例

```
library(readxl)
d <- read_excel("../demo_data/olympics.xlsx")
tail(d, 6)
#> # A tibble: 6 x 3
#>   Olympic_year Men_score Women_score
#>   <dbl>         <dbl>         <dbl>
#> 1    1984         9.99         11.0
#> 2    1988         9.92         10.5
#> 3    1992         9.96         10.8
#> 4    1996         9.84         10.9
#> 5    2000         9.87         10.8
#> 6    2004         9.85         10.9
```

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