数据操纵手册

张晨阳

2018-06-29

献给……

呃，爱谁谁吧

目录

表格

插图

前言

一些对过去工作经验零散凌乱的总结让我有了将其汇总成书的念头。又介于自己并非是什么成功人士、学霸大佬或者文笔绚烂的文人，所以迟迟未“动笔”，总觉得这并非是我该做的事情。

由于 **bookdown** (Xie, 2018) 包给 R 语言用户提供了一种便捷的写作方式，这让我有了尝试的动力。

总的来说这本的书是关于如何操作数据的。由于 **kintr**(Xie, 2015) 强大的多语言支持，可以让我不局限于 R 语言，同样深受欢迎的 Python 和 SQL。

```
sessionInfo()
```

```
## R version 3.5.0 (2018-04-23)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRb
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/lib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets
```

```
## [6] methods   base
##
## other attached packages:
## [1] reticulate_1.7
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.17    bookdown_0.7    lattice_0.20-35
##  [4] digest_0.6.15   rprojroot_1.3-2 grid_3.5.0
##  [7] DBI_1.0.0       jsonlite_1.5    backports_1.1.2
## [10] magrittr_1.5    evaluate_0.10.1 highr_0.6
## [13] stringi_1.2.2   rstudioapi_0.7  Matrix_1.2-14
## [16] rmarkdown_1.9   RMySQL_0.10.15  tools_3.5.0
## [19] stringr_1.3.1   xfun_0.1        yaml_2.1.19
## [22] compiler_3.5.0  htmltools_0.3.6 knitr_1.20
```

致谢

感谢我的父母吧!

张晨阳
上海

作者简介

徜徉在数据的中的一个默默无闻的清道夫。

第一章　牛刀小试

现在我们可以试试 **bookdown** 的一些初级功能了，例如图表。图 1.1 是一幅无趣的散点图，表 1.2 是一份枯燥的数据。

```
head(mtcars, 5)
```

```
##                   mpg cyl disp  hp drat    wt  qsec
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02
##                   vs am gear carb
## Mazda RX4          0  1    4    4
## Mazda RX4 Wag      0  1    4    4
## Datsun 710         1  1    4    1
## Hornet 4 Drive     1  0    3    1
## Hornet Sportabout  0  0    3    2
```

```
import pandas
print(r.mtcars.head())
```

```
##                    mpg  cyl   disp     hp  drat     wt   qsec   vs
## Mazda RX4         21.0  6.0  160.0  110.0  3.90  2.620  16.46  0.0  1
## Mazda RX4 Wag     21.0  6.0  160.0  110.0  3.90  2.875  17.02  0.0  1
## Datsun 710        22.8  4.0  108.0   93.0  3.85  2.320  18.61  1.0  1
```

表 1.1: 5 records

rownames	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2

```
## Hornet 4 Drive     21.4  6.0  258.0  110.0  3.08  3.215  19.44  1.0  0.0
## Hornet Sportabout  18.7  8.0  360.0  175.0  3.15  3.440  17.02  0.0  0.0
##
##                    gear  carb
## Mazda RX4           4.0   4.0
## Mazda RX4 Wag       4.0   4.0
## Datsun 710          4.0   1.0
## Hornet 4 Drive      3.0   1.0
## Hornet Sportabout   3.0   2.0
```

```
select * from mtcars limit 5
```

```
par(mar = c(4, 4, 1, .1))
plot(cars, pch = 19)
```

```
knitr::kable(
  head(iris), caption = ' 雷猴啊，iris 数据！ ',
  booktabs = TRUE
)
```

就这样，你可以一直编下去，直到编不下去。

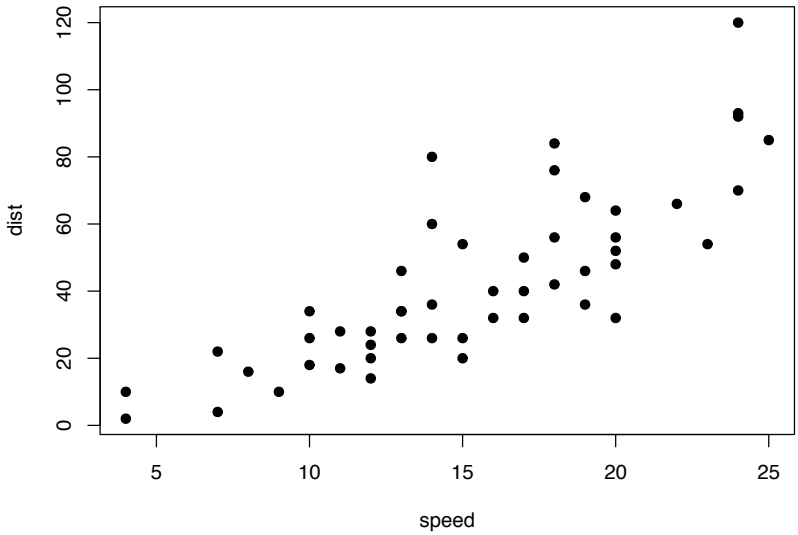

图 1.1: 雷猴啊，散点图！

表 1.2: 雷猴啊，iris 数据！

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

第二章　数据

理解数据类型对操纵数据是十分重要的。同样数据结构也是十分重要的。

2.1　数据类型

R 和 Python 中的数据类型是大致相同的，主要有：

- 整型：Python 中可以处理任意大小的整数。
- 浮点型：浮点数也就是小数，之所以称为浮点数，是因为按照科学记数法表示时，一个浮点数的小数点位置是可变的，浮点数除了数学写法（如 `123.456`）之外还支持科学计数法（如 `1.23456e2`）。
- 字符串型：字符串是以单引号或双引号括起来的任意文本，比如`'hello'` 和`"hello"`, 字符串还有原始字符串表示法、字节字符串表示法、Unicode 字符串表示法，而且可以书写成多行的形式（用三个单引号或三个双引号开头，三个单引号或三个双引号结尾）。
- 布尔型：布尔值只有 `True`、`False` 两种值，要么是 `True`，要么是 `False`，在 Python 中，可以直接用 `True`、`False` 表示布尔值（请注意大小写），也可以通过布尔运算计算出来（例如 `3 < 5` 会产生布尔值 `True`，而 `2 == 1` 会产生布尔值 `False`）。
- 复数型：形如 `3+5j`，跟数学上的复数表示一样，唯一不同的是虚部的 i 换成了 j。

```
a <- TRUE

as.logical(a)
```

```
## [1] TRUE
```

```
as.integer(a)
```

```
## [1] 1
```

```
as.double(a)
```

```
## [1] 1
```

```
as.character(a)
```

```
## [1] "TRUE"
```

```
a = True
print(type(a))
```

```
## <type 'bool'>
```

```
print(int(a))
```

```
## 1
```

```
print(float(a))
```

```
## 1.0
```

```
print(str(a))
```

```
## True
```

2.2 数据结构

貌似大家都喜欢用白萍风这个意境。又如彭玉麟的对联：

> 凭栏看云影波光，最好是红蓼花疏、白苹秋老；
> 把酒对琼楼玉宇，莫孤负天心月到、水面风来。

嘿，玛尼玛尼哄。

附录 A　余音绕梁

呐，到这里朕的书差不多写完了，但还有几句话要交待，所以开个附录，再啰嗦几句，各位客官稍安勿躁、扶稳坐好。

参考文献

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.7.

索引