成為初級資料分析師 IR 程式設計與資料科學應用

資料框處理

郭耀仁

Tidy datasets are all alike, but every messy dataset is messy in its own way.

Hadley Wickham

大綱

- 常用檢視資料框的函數
- 基礎資料框處理
- 使用 `dplyr` 處理資料框

常用檢視資料框的函數

常見檢視資料框的函數一覽

- dim()、nrow() 與 ncol() 檢視外觀
- summary() 描述性統計
- str() 詳細資訊
- View()、head() 與 tail() 顯示資料框

In [ ]:
```
csv_url <- "https://s3-ap-northeast-1.amazonaws.com/r-essentials/chicago_bulls_1995_1996.csv"
chicago_bulls <- read.csv(csv_url, stringsAsFactors = FALSE)
```

In [ ]:

```
nrow(chicago_bulls)
ncol(chicago_bulls)
dim(chicago_bulls)
summary(chicago_bulls)
str(chicago_bulls)
head(chicago_bulls)
tail(chicago_bulls)
View(chicago_bulls)
```

基礎資料框處理

基礎資料框處理的技巧

- 解構資料框
    - 選擇
    - 篩選
    - 選擇與篩選
- 排序資料框
- 新增變數
- 摘要
- 分組摘要

解構資料框：選擇

使用 df[, COLUMN_NAME] 或 df$COLUMN_NAME

In [ ]:
```
chicago_bulls[, "Player"]
chicago_bulls$Player
```

解構資料框：篩選

使用 df[EXPR,] 或 df[ROW_INDICES,]

In [ ]:
```
chicago_bulls[chicago_bulls$Player == "Michael Jordan",]
chicago_bull[7, ]
```

解構資料框：選擇與篩選

使用 df[EXPR, COLUMN_NAME] 或 df[ROW_INDICES, COLUMN_NAME]

In [ ]:
```
chicago_bulls[chicago_bulls$Player == "Michael Jordan", "Player"]
chicago_bulls[7, "Player"]
```

隨堂練習： 鐵三角 Michael Jordan, Scottie Pippen 與 Dennis Rodman

In [2]:
```
trio
```

	No.	Player
7	23	Michael Jordan
11	33	Scottie Pippen
12	91	Dennis Rodman

排序資料框

利用 `order()` 函數取得排序後的列索引

In [ ]:
```
ordered_indices <- order(chicago_bulls[, "No."])
chicago_bulls[ordered_indices, ]
```

新增變數

In [ ]:
```
chicago_bulls$Wt_kg <- chicago_bulls$Wt * 0.45359
head(chicago_bulls)
```

隨堂練習： 新增變數 `Ht_cm`

- 1 feet = 30.48 cm
- 1 inch = 2.54 cm

In [4]:
```
head(chicago_bulls)
```

No.	Player	Pos	Ht	Wt	Birth.Date	College	Ht_cm
0	Randy Brown	PG	6-2	190	May 22, 1968	University of Houston, New Mexico State University	187.96
30	Jud Buechler	SF	6-6	220	June 19, 1968	University of Arizona	198.12
35	Jason Caffey	PF	6-8	255	June 12, 1973	University of Alabama	203.20
53	James Edwards	C	7-0	225	November 22, 1955	University of Washington	213.36
54	Jack Haley	C	6-10	240	January 27, 1964	University of California, Los Angeles	208.28
9	Ron Harper	PG	6-6	185	January 20, 1964	Miami University	198.12

摘要

針對欲摘要的變數使用敘述性統計函數

In [ ]:
```
mean(chicago_bulls$Ht_cm)
```

使用 dplyr 處理資料框

使用 dplyr 處理資料框

- 安裝 dplyr 套件
- 載入 dplyr 套件

安裝 dplyr 套件

- 透過 RStudio 的 Packages 功能頁籤
- 透過 install.pacakges() 函數

```
install.pacakges("dplyr")
```

載入 dplyr 套件

- 透過 RStudio 的 Packages 功能頁籤
- 透過 library() 函數

```
library("dplyr")
```

使用 %>% 鏈結函數（chaining functions）

- %>% 運算符來自 magrittr，會隨著 dplyr 一起被安裝
- 讓需要鏈結函數的資料操作可讀性更高
- 在 RStudio 中使用 Ctrl-Shift-M 快捷鍵可以叫出 %>% 運算符

使用 `dplyr` 進行基礎資料框處理

- 解構資料框
    - 選擇與篩選
- 排序資料框
- 新增變數
- 摘要
- 分組摘要

解構資料框：選擇與篩選

- 使用 dplyr::select() 函數選擇
- 使用 dplyr::filter() 函數篩選

In [ ]:
```r
library("dplyr")

# Without %>%
select(chicago_bulls, Player)
filter(select(chicago_bulls, Player), Player == "Michael Jordan")
```

In [ ]:

```
library("dplyr")

# With %>%
chicago_bulls %>%
    select(Player) %>%
    filter(Player == "Michael Jordan")
```

排序資料框

使用 dplyr::arrange() 函數排序

In [ ]:
```
library("dplyr")

chicago_bulls %>%
    arrange(`No.`)
```

新增變數

使用 dplyr::mutate() 函數新增變數

In [ ]:
```
library("dplyr")

chicago_bulls %>%
    mutate(Wt_kg = Wt * 0.45359)
```

摘要

使用 dplyr::summarise() 函數摘要

In [ ]:
```
library("dplyr")

chicago_bulls %>%
    mutate(Wt_kg = Wt * 0.45359) %>%
    summarise(Avg_Wt_kg = mean(Wt_kg))
```

分組摘要

使用 dplyr::group_by() 搭配 dplyr::summarise() 函數分組摘要

In [ ]:
```
library("dplyr")

chicago_bulls %>%
    mutate(Wt_kg = Wt * 0.45359) %>%
    group_by(Pos) %>%
    summarise(Avg_Wt_kg = mean(Wt_kg))
```

隨堂練習：摘要每個鋒衛位置的最高身高

In [6]:
```
ans
```

Pos	max_Ht_cm
C	213.3600
PF	205.1050
PG	192.1933
SF	203.2000
SG	198.1200