

Chapter 10

BASIC DATA PROCESSING (2)

長寬表格轉換 (tidyverse)

```
1 library(tidyverse)
2 library(tidyr)
3
4 Player <- c("Stephen Curry", "Klay Thompson")
5 Pts <- c(30.1, 22.1)
6 T3p <- c(402, 276)
7
8 collec <- data.frame(Player, Pts, T3p,
9                       stringsAsFactors = FALSE)
10
11 gather(collec, key = stat, value = value, Pts, T3p)
12 View(collec)
```

1.R x collec x			
Filter			
	Player	Pts	T3p
1	Stephen Curry	30.1	402
2	Klay Thompson	22.1	276

```
> gather(collec, key = stat, value = value, Pts, T3p)
  Player stat value
1 Stephen Curry Pts  30.1
2 Klay Thompson Pts  22.1
3 Stephen Curry T3p 402.0
4 Klay Thompson T3p 276.0
```

Install Packages

Install from: [? Configuring Repositories](#)

Repository (CRAN, CRANextra) ▼

Packages (separate multiple with space or comma):

tidyverse

結構化查詢 (tidyverse)

```
1 library(tidyverse)
2
3 Player <- c("Stephen Curry", "Klay Thompson")
4 Pts <- c(30.1, 22.1)
5 T3p <- c(402, 276)
6 Tp <- c(0.454, 0.425)
7 Season <- c("2015-2016", "2015-2016")
8 Shoes <- c("UA", "Anta")
9
10 collec <- data.frame(Player, Pts, T3p, Tp, Season, Shoes,
11                      stringsAsFactors = FALSE)
12 KD <- c("Kevin Durant", 28.2, 186, 0.387,
13         "2015-2016", "Nike")
14 collec <- rbind(collec, KD)
15
16 filter(collec, T3p >= 200)
17
18 filter(collec, T3p >= 150 & Tp > 0.45)|
```

Function	Meaning
filter()	篩選 (過濾)
select()	選擇
mutate()	新增
arrange()	排序
summarise()	聚合函數
group_by()	分組

```
> filter(collec, T3p >= 200)
      Player  Pts T3p   Tp   Season Shoes
1 Stephen Curry 30.1 402 0.454 2015-2016   UA
2 Klay Thompson 22.1 276 0.425 2015-2016  Anta
>
> filter(collec, T3p >= 150 & Tp > 0.45)
      Player  Pts T3p   Tp   Season Shoes
1 Stephen Curry 30.1 402 0.454 2015-2016   UA
```

結構化查詢 (tidyverse)

```
1 library(tidyverse)
2
3 Player <- c("Stephen Curry", "Klay Thompson")
4 Pts <- c(30.1, 22.1)
5 T3p <- c(402, 276)
6 Tp <- c(0.454, 0.425)
7 Season <- c("2015-2016", "2015-2016")
8 Shoes <- c("UA", "Anta")
9
10 collec <- data.frame(Player, Pts, T3p, Tp, Season, Shoes,
11                      stringsAsFactors = FALSE)
12 KD <- c("Kevin Durant", 28.2, 186, 0.387,
13        "2015-2016", "Nike")
14 collec <- rbind(collec, KD)
15
16 select(collec, Player)
17
18 select(collec, Name = Player)
```

Function	Meaning
filter()	篩選 (過濾)
select()	選擇
mutate()	新增
arrange()	排序
summarise()	聚合函數
group_by()	分組

```
> select(collec, Player)
      Player
1 Stephen Curry
2 Klay Thompson
3 Kevin Durant
>
> select(collec, Name = Player)
      Name
1 Stephen Curry
2 Klay Thompson
3 Kevin Durant
```

結構化查詢 (tidyverse)

```
1 library(tidyverse)
2
3 Player <- c("Stephen Curry", "Klay Thompson", "Kevin Durant")
4 Pts <- c(30.1, 22.1, 28.2)
5 T3p <- c(402, 276, 186)
6 T3n <- c(886, 650, 481)
7
8 collec <- data.frame(Player, Pts, T3p, T3n,
9                       stringsAsFactors = FALSE)
10
11 mutate(collec, Tp = T3p/T3n)
12
```

Function	Meaning
filter()	篩選 (過濾)
select()	選擇
mutate()	新增
arrange()	排序
summarise()	聚合函數
group_by()	分組

```
> mutate(collec, Tp = T3p/T3n)
  Player Pts T3p T3n      Tp
1 Stephen Curry 30.1 402 886 0.4537246
2 Klay Thompson 22.1 276 650 0.4246154
3 Kevin Durant 28.2 186 481 0.3866944
```

結構化查詢 (tidyverse)

```
1 library(tidyverse)
2
3 Player <- c("Stephen Curry", "Klay Thompson", "Kevin Durant")
4 Pts <- c(30.1, 22.1, 28.2)
5 T3p <- c(402, 276, 186)
6 T3n <- c(886, 650, 481)
7
8 collec <- data.frame(Player, Pts, T3p, T3n,
9                       stringsAsFactors = FALSE)
10
11 collec <- mutate(collec, Tp = T3p/T3n)
12
13 arrange(collec, desc(Pts))
14
15 arrange(collec, Tp)|
```

Function	Meaning
filter()	篩選 (過濾)
select()	選擇
mutate()	新增
arrange()	排序
summarise()	聚合函數
group_by()	分組

```
> arrange(collec, desc(Pts))
      Player  Pts T3p T3n      Tp
1 Stephen Curry 30.1 402 886 0.4537246
2 Kevin Durant 28.2 186 481 0.3866944
3 Klay Thompson 22.1 276 650 0.4246154
>
> arrange(collec, Tp)
      Player  Pts T3p T3n      Tp
1 Kevin Durant 28.2 186 481 0.3866944
2 Klay Thompson 22.1 276 650 0.4246154
3 Stephen Curry 30.1 402 886 0.4537246
```

結構化查詢 (tidyverse)

```
1 library(tidyverse)
2
3 Player <- c("Stephen Curry", "Klay Thompson", "Kevin Durant")
4 Pts <- c(30.1, 22.1, 28.2)
5 T3p <- c(402, 276, 186)
6 T3n <- c(886, 650, 481)
7
8 collec <- data.frame(Player, Pts, T3p, T3n,
9                       stringsAsFactors = FALSE)
10
11 collec <- mutate(collec, Tp = T3p/T3n)
12
13 summarise(collec, mean(Pts))
14
15 summarise(collec, mean(Tp))|
```

Function	Meaning
filter()	篩選 (過濾)
select()	選擇
mutate()	新增
arrange()	排序
summarise()	聚合函數
group_by()	分組

```
> summarise(collec, mean(Pts))
  mean(Pts)
1      26.8
>
> summarise(collec, mean(Tp))
  mean(Tp)
1 0.4216781
> |
```

結構化查詢 (tidyverse)

```
1 library(tidyverse)
2
3 Player <- c("Stephen Curry", "Klay Thompson",
4             "Kevin Durant", "Russell Westbrook")
5 Pts <- c(30.1, 22.1, 28.2, 23.5)
6 T3p <- c(402, 276, 186, 101)
7 T3n <- c(886, 650, 481, 341)
8 team <- c("GSW", "GSW", "OKC", "OKC")
9 collec <- data.frame(Player, Pts, T3p, T3n, team,
10                      stringsAsFactors = FALSE)
11
12 collec <- mutate(collec, Tp = T3p/T3n)
13
14 a <- group_by(collec, team)
15 b <- summarise(a, mean(Pts))
16 c <- as.data.frame(b)
17
18 print(c)
```

Function	Meaning
filter()	篩選 (過濾)
select()	選擇
mutate()	新增
arrange()	排序
summarise()	聚合函數
group_by()	分組

```
> print(c)
  team mean(Pts)
1 GSW      26.10
2 OKC      25.85
```


結構化查詢 (%>%)

```
1 library(tidyverse)
2 Player <- c("Stephen Curry", "Klay Thompson",
3             "Kevin Durant", "Russell Westbrook")
4 Pts <- c(30.1, 22.1, 28.2, 23.5)
5 T3p <- c(402, 276, 186, 101)
6 T3n <- c(886, 650, 481, 341)
7 team <- c("GSW", "GSW", "OKC", "OKC")
8 collec <- data.frame(Player, Pts, T3p, T3n, team,
9                       stringsAsFactors = FALSE)
10 collec <- mutate(collec, Tp = T3p/T3n)
11
12 #a <- group_by(collec, team)
13 #b <- summarise(a, mean(Pts))
14 #c <- as.data.frame(b)
15
16 group_by(collec, team) %>%
17   summarise(mean(Pts)) %>%
18   as.data.frame()
```

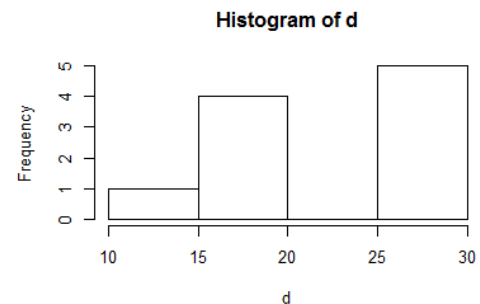
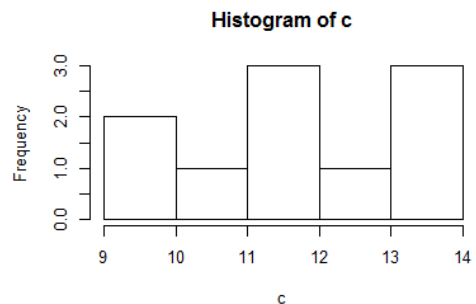
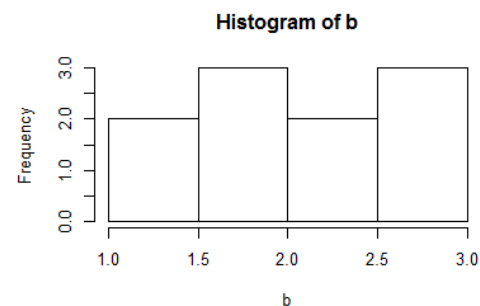
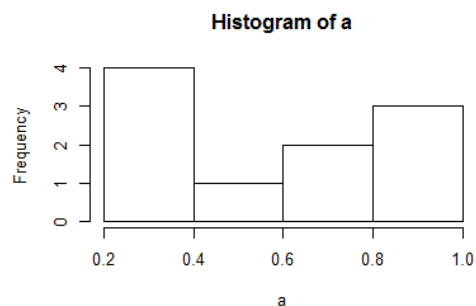
Function	Meaning
filter()	篩選 (過濾)
select()	選擇
mutate()	新增
arrange()	排序
summarise()	聚合函數
group_by()	分組

```
> group_by(collec, team) %>%
+   summarise(mean(Pts)) %>%
+   as.data.frame()
  team mean(Pts)
1  GSW    26.10
2  OKC    25.85
```

隨機分佈亂數 (runif)

```
1 a <- runif(10)
2 print(a)
3
4 b <- runif(10)*3
5 print(b)
6
7 c <- runif(10)*10 + 5
8 print(c)
9
10 d <- runif(10)*20 + 10
11 print(d)
12
13 par(mfrow = c(2,2))
14 hist(a)
15 hist(b)
16 hist(c)
17 hist(d)
```

```
> print(a)
[1] 0.3449741 0.5854575 0.3694249 0.8964922 0.6516345 0.3966556 0.9982142 0.9306506 0.2162981 0.6192556
>
> b <- runif(10)*3
> print(b)
[1] 1.412705 1.546875 2.571578 2.071702 2.991086 1.934721 2.618626 1.721352 2.319308 1.019049
>
> c <- runif(10)*10 + 5
> print(c)
[1] 13.731548 10.113980 13.593970 13.766539 9.534819 12.066363 11.541722 11.728166 9.208419 11.025521
>
> d <- runif(10)*20 + 10
> print(d)
[1] 16.30397 10.19893 28.10036 25.08971 17.61423 16.19506 25.90539 28.48024 27.58171 15.67576
>
```



利用 apply() 取代迴圈

apply(data, MARGIN, FUN)

MARGIN

1: 列 2: 行

FUN

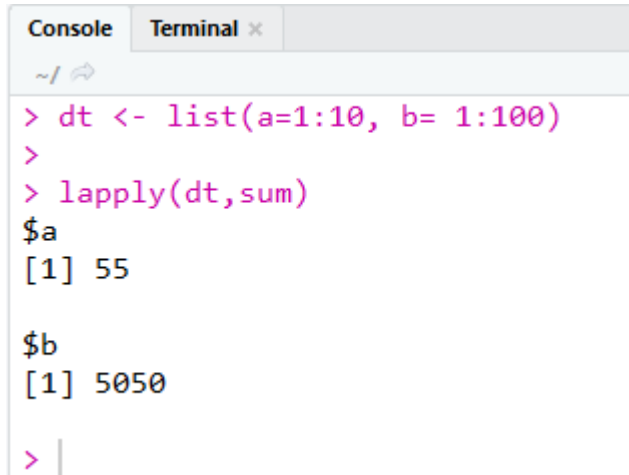
內建 / 自訂函數

```
1
2 dt <- array(1:9, dim= c(3,3))
3 print(dt)
4
5 apply(dt,1,sum)
6 apply(dt,2,sum)
7
8
```

```
Console Terminal x
~/
> dt <- array(1:9, dim= c(3,3))
> print(dt)
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
>
> apply(dt,1,sum)
[1] 12 15 18
> apply(dt,2,sum)
[1]  6 15 24
.
```

利用 lapply() 取代迴圈

```
1 dt <- list(a=1:10, b= 1:100)
2
3 lapply(dt,sum)
```



The screenshot shows a R console window with two tabs: 'Console' and 'Terminal x'. The 'Console' tab is active. The prompt is '~/' with a refresh icon. The following commands and output are shown:

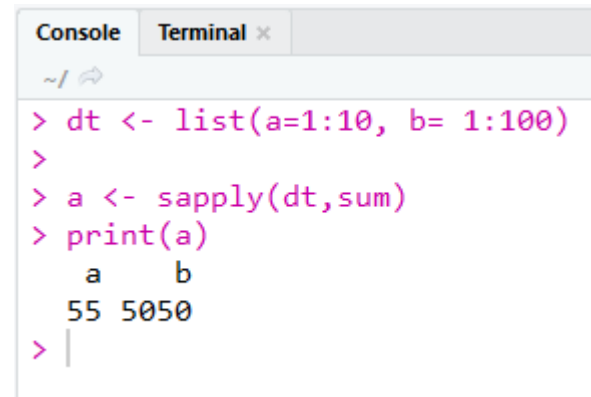
```
> dt <- list(a=1:10, b= 1:100)
>
> lapply(dt,sum)
$a
[1] 55

$b
[1] 5050

> |
```

利用 `sapply()` 取代迴圈

```
1 dt <- list(a=1:10, b= 1:100)
2
3 a <- sapply(dt,sum)
4 print(a)
5
```



The screenshot shows a R console window with two tabs: 'Console' and 'Terminal x'. The 'Console' tab is active. The code entered is: `> dt <- list(a=1:10, b= 1:100)`, `>`, `> a <- sapply(dt,sum)`, and `> print(a)`. The output of `print(a)` is displayed as a matrix with two columns, 'a' and 'b', and one row of values: 55 and 5050. The prompt `>` is shown at the bottom of the console.

```
Console Terminal x
~/
> dt <- list(a=1:10, b= 1:100)
>
> a <- sapply(dt,sum)
> print(a)
      a      b
55 5050
> |
```

隨堂練習 1

1. 隨機產生 50 個人的 3 分球投進與沒投進的次數，並加總投籃次數

```
> print(df)      > print(df2)
   fg  fm      fg  fm  fa
1  51  56    1  51  56 107
2  62  47    2  62  47 109
3  93  11    3  93  11 104
4  25  94    4  25  94 119
5  57  61    5  57  61 118
```

2. 使用 `apply()` / `mapply()` 來計算命中率

```
      x
1 0.47663551
2 0.56880734
3 0.89423077
4 0.21008403
5 0.48305085
```

Any Questions !?