



ETL with R

智庫驅動

Ben Chen



今日重點!!!

整理資料

Extraction, Transform and Load

需要的套件

```
library(ggplot2)  
library(data.table)  
library(dplyr)  
library(reshape2)
```

確定自己的路徑

```
getwd( )
```

```
## [1] "/Users/benjamin/Ben"
```

```
setwd('路徑')
```

小技巧

- **tab**補齊指令
- **command(or control)+ enter**執行source指令
- **command(or control)+ shift+ c**註解或解除註解
- **command+ 1** 滑鼠游標移至source
- **command+ 2** 滑鼠游標移至Console

- **command+ L** 清除Console

先讀檔

讀檔之前，觀察檔案

編碼通常都是**UTF8**或**BIG5**

```
raw <- readLines('檔案路徑', n = 10, encoding = "BIG-5")
```

轉換編碼並存檔

```
raw2 <- iconv(raw, from = "BIG-5", to = "UTF-8")  
# 從big5轉utf8  
write(raw2, "ubikeweatherutf8.csv")  
# 存檔囉~~
```


讀取ubike資料

```
ubike = read.csv('檔案路徑',  
               colClasses = c("factor", "integer", "integer", "factor", "factor",  
                              "numeric", "numeric", "integer", "numeric", "integer",  
                              "integer", "numeric", "numeric", "integer", "integer",  
                              "numeric", "numeric", "numeric", "numeric", "numeric",  
                              "numeric"), fileEncoding = 'utf8')  
# 以colClasses控制每個欄位的class，這可使讀檔加速  
# 以fileEncoding定義檔案編碼
```

```
##  
Read 83.8% of 656711 rows  
Read 656711 rows and 21 (of 21) columns from 0.108 GB file in 00:00:03
```

進階讀取

fread是data.table裡的function

```
ubike = fread('檔案路徑',  
  data.table = FALSE,  
  colClasses = c("factor", "integer", "integer", "factor",  
    "factor", "numeric", "numeric", "integer",  
    "numeric", "integer", "integer", "numeric",  
    "numeric", "integer", "integer", "numeric",  
    "numeric", "numeric", "numeric", "numeric",  
    "numeric"))
```

把欄位中文英文對照

X1	X2	X3	X4
日期	date	車輛數標準差	std.sbi
時間	hour	平均空位數	avg.bemp
場站代號	sno	最大空位數	max.bemp
場站區域	sarea	最小空位數	min.bemp
場站名稱	sna	空位數標準差	std.bemp
緯度	lat	平均氣溫	temp
經度	lng	溼度	humidity
總停車格	tot	氣壓	pressure
平均車輛數	avg.sbi	最大風速	max.anemo
最大車輛數	max.sbi	降雨量	rainfall
最小車輛數	min.sbi		

把欄位換成中文

```
colnames(ubike) <-  
  c("日期", "時間", "場站代號", "場站區域", "場站名稱",  
    "緯度", "經度", "總停車格", "平均車輛數", "最大車輛數",  
    "最小車輛數", "車輛數標準差", "平均空位數", "最大空位數",  
    "最小空位數", "空位數標準差", "平均氣溫", "溼度",  
    "氣壓", "最大風速", "降雨量")
```

讀完檔就是要取值啊～～不然要幹嘛

取值- 座標

```
# head可以取出前幾列
head(ubike)
# tail可以取最後幾列
tail(ubike)
# 利用座標來取值，第一個數表示列位，第二個數表示欄位
ubike[3,2]
# 可一次選擇多列多欄
ubike[c(3:4),c(2:5,7)]
# 加上負號可剔除列位欄位
ubike[c(3:4),-c(6:21)]
```

取值- 指定欄位

```
ubike[,4]  
ubike[, "sna"]  
ubike[["sna"]]  
ubike$場站名稱
```

dplyr

magrittr

- 壓縮的程式碼不好讀
- 展開的程式碼會產生很多暫存變數
- 套件`magrittr`部份解決了這個問題

```
ans1 <- ubike$sna
ans1.1 <- unique(ans1) # unique可列出所有不重複的項目

unique(ubike$sna)

library(magrittr)
ubike$sna %>%
  unique
```

dplyr

- 讓R 使用者可以用更有彈性的方式來處理資料
- 針對`data.frame`做設計（名稱中的d）
- 設計理念
 - 導入資料整理最重要的動作（非常類似SQL）
 - 快
 - 支援異質資料源（`data.frame`或資料庫中的表格）


學習dplyr的官方方式：**vignette**

```
vignette(all = TRUE, package = "dplyr")  
vignette("introduction", package = "dplyr")
```

- 更詳細的dplyr介紹可以閱讀dplyr的小論文
- R 的開發者會針對一個主題撰寫小論文做介紹

dplyr簡介

- **filter** 對列做篩選
- **select** 對欄做篩選
- **mutate** 更改欄或新增欄
- **arrange** 排列
- **group_by+summarise** 分類
- 合併欄位

出處： [資料科學愛好者年會資料分析上手課程：ETL1](#)

小明想在永和找到新房子，希望以後上下班都靠Ubike通勤，希望早上可以輕鬆租到車，下班後也可以輕鬆還車，請幫他找出中和區早上七點腳踏車最多的場站。

中和區腳踏車

SNA	AVG_RATE
中和公園	13.56075
捷運永安市場站	17.51402
中和國民運動中心	27.63830
秀山國小	29.25234

select

X1	X2	X3	X4	X5		X1	X3	X4
V		V	V		⇒	V	V	V
V		V	V			V	V	V
V		V	V			V	V	V
V		V	V			V	V	V
V		V	V			V	V	V
V		V	V			V	V	V

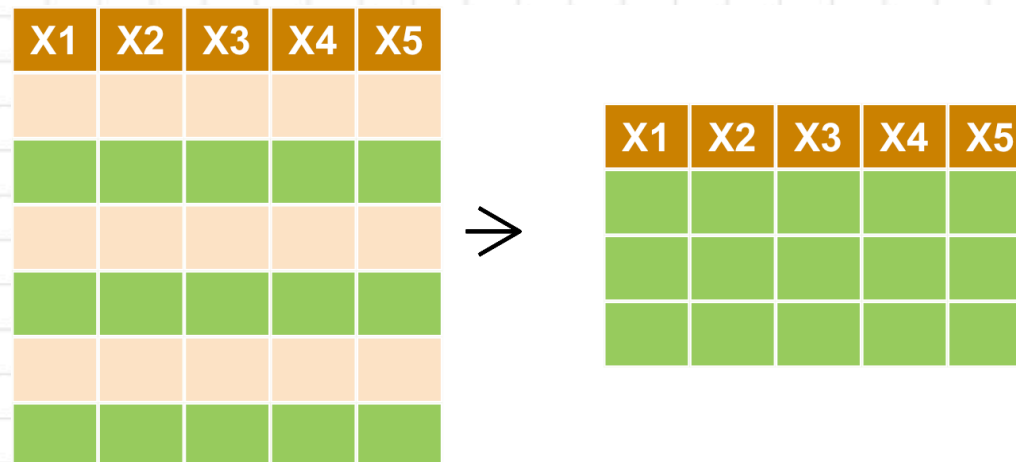
select

- 選擇時間、場站區域、場站名稱平均車輛數

```
ubike1<- select(ubike, hour, sarea, sna, avg.sbi)
```

HOUR	SAREA	SNA	AVG.SBI
15	信義區	捷運市政府站(3號出口)	96.500
15	大安區	捷運國父紀念館站(2號出口)	24.000
15	信義區	台北市政府	10.333
15	信義區	市民廣場	39.333
15	信義區	興雅國中	34.167
15	信義區	世貿二館	31.333

filter



filter

- 過濾出中和區的資料

```
ubike1<- select(ubike, hour, sarea, sna, avg.sbi) %>%  
  filter(sarea=='中和區' & hour==7)
```

HOUR	SAREA	SNA	AVG.SBI
7	中和區	秀山國小	0.000
7	中和區	捷運永安市場站	1.733
7	中和區	中和公園	0.267
7	中和區	秀山國小	33.800
7	中和區	捷運永安市場站	6.467
7	中和區	中和公園	4.600
7	中和區	秀山國小	37.400
7	中和區	捷運永安市場站	12.867
7	中和區	中和公園	1.467
7	中和區	秀山國小	31.867

mutate

- 新增欄位計算有車率

```
ubike1<- select(ubike, hour, sarea, sna, avg.sbi) %>%
  filter(sarea=='中和區' & hour==7) %>%
  mutate(avg.sbi=floor(avg.sbi))
```

HOUR	SAREA	SNA	AVG.SBI
7	中和區	秀山國小	0
7	中和區	捷運永安市場站	1
7	中和區	中和公園	0
7	中和區	秀山國小	33
7	中和區	捷運永安市場站	6
7	中和區	中和公園	4

--- &vcenter .largecontent

group_by

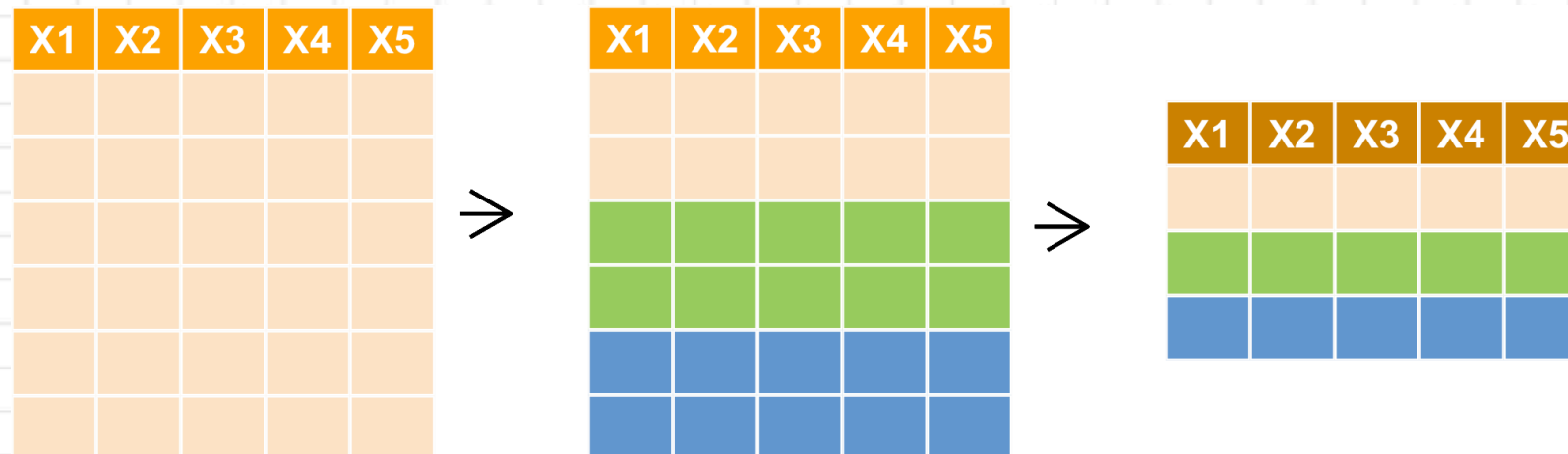
group_by

- 將各站的數據群聚起來

```
ubike1<- select(ubike, hour, sarea, sna, avg.sbi) %>%  
  filter(sarea=='中和區' & hour==7) %>%  
  mutate(avg.sbi=floor(avg.sbi)) %>%  
  group_by(sna)  
ubike1
```

```
## Source: local data frame [368 x 4]  
## Groups: sna [4]  
##  
##      hour  sarea      sna avg.sbi  
##      (int) (chr)    (chr)  (dbl)  
## 1      7 中和區    秀山國小      0  
## 2      7 中和區 捷運永安市場站      1  
## 3      7 中和區    中和公園      0  
## 4      7 中和區    秀山國小     33  
## 5      7 中和區 捷運永安市場站      6  
## 6      7 中和區    中和公園      4
```

summarise



summarise

```
ubike1<- select(ubike, hour, sarea, sna, avg.sbi) %>%  
  filter(sarea=='中和區' & hour==7) %>%  
  mutate(avg.sbi=floor(avg.sbi)) %>%  
  group_by(sna) %>%  
  summarise(avg_rate=mean(avg.sbi))
```

SNA	AVG_RATE
捷運永安市場站	17.51402
秀山國小	29.25234
中和公園	13.56075
中和國民運動中心	27.63830

arrange

X1	X2		X1	X2
3	Ben		1	Rafe
4	Johnson	➤	2	Ning
1	Rafe		3	Ben
2	Ning		4	Johnson

arrange

```
ubike1<- select(ubike, hour, sarea, sna, avg.sbi) %>%  
  filter(sarea=='中和區' & hour==7) %>%  
  mutate(avg.sbi=floor(avg.sbi)) %>%  
  group_by(sna) %>%  
  summarise(avg_rate=mean(avg.sbi)) %>%  
  arrange(avg_rate)
```

SNA	AVG_RATE
中和公園	13.56075
捷運永安市場站	17.51402
中和國民運動中心	27.63830
秀山國小	29.25234

練習一下

小明發現住板橋的話，八點騎腳踏車就可以準時上班，還可以順便吃早餐，請幫忙找出板橋區各車站八點車子最多的站

練習一下

小明發現住板橋的話，八點騎腳踏車就可以準時上班，還可以順便吃早餐，請幫忙找出板橋區各車站八點車子最多的站

SNA	AVG_RATE
永安公園	15.00000
捷運江子翠站(3號出口)	23.33333
音樂公園	23.66667
板橋國民運動中心	24.33333

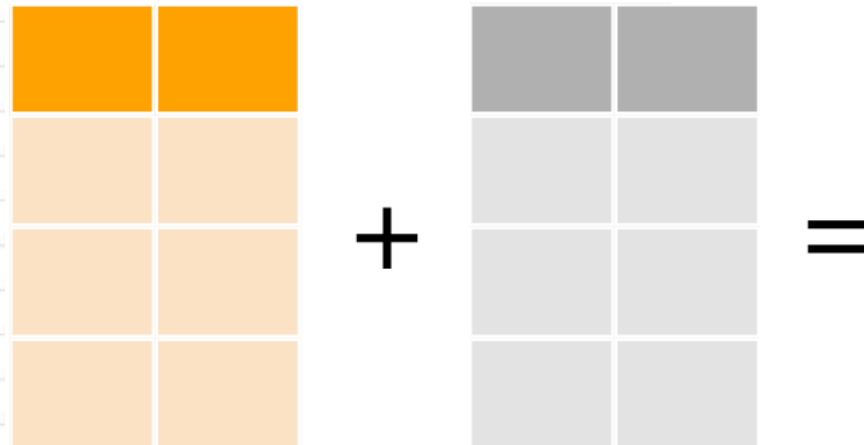
練習一下

小明喜歡玩遙控飛機，在中和希望找一個下午三點風比較小的地點吧

SNA	AVG_ANEMO
捷運永安市場站	2.609531
秀山國小	3.033534
中和公園	2.698973
中和國民運動中心	2.177059

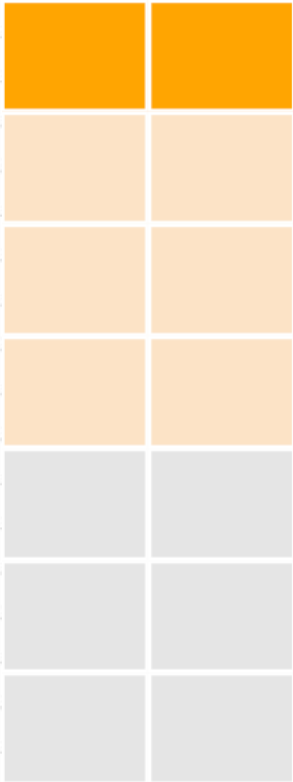
bind

```
bind_rows(a,b)  
bind_cols(a,b)
```

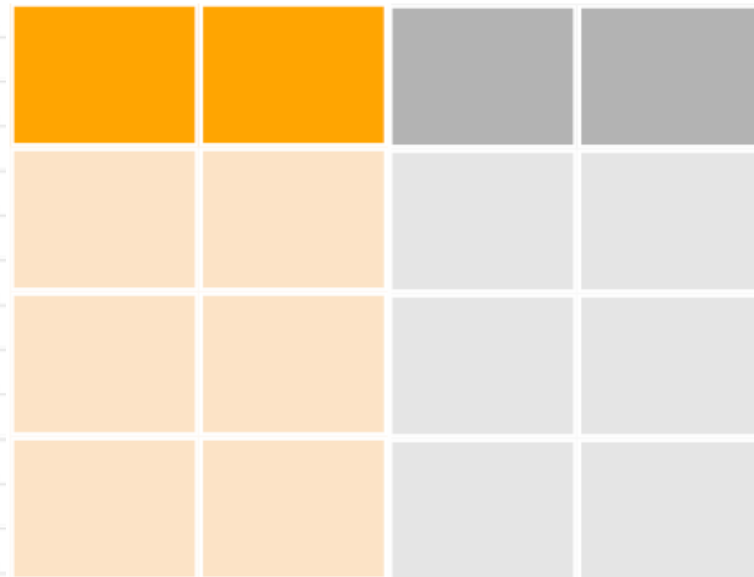


bind

```
bind_rows(a,b)
```



```
bind_cols(a,b)
```



bind

```
bind_rows(ubike1,ubike2)
```

```
## Source: local data frame [8 x 2]
##
##           sna avg_rate
##           (chr)      (dbl)
## 1      中和公園 13.56075
## 2 捷運永安市場站 17.51402
## 3 中和國民運動中心 27.63830
## 4      秀山國小 29.25234
## 5      永安公園 15.00000
## 6 捷運江子翠站(3號出口) 23.33333
## 7      音樂公園 23.66667
## 8 板橋國民運動中心 24.33333
```

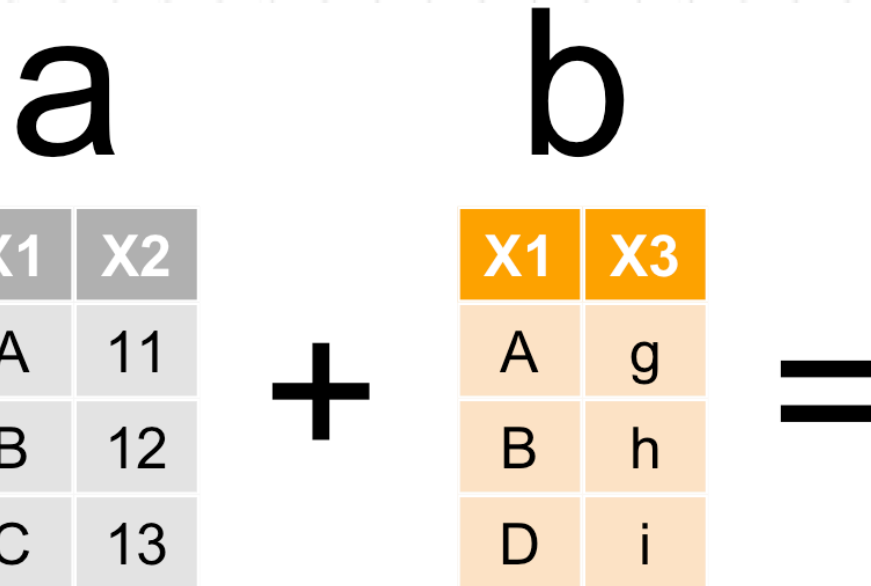
bind

```
bind_cols(ubike1,ubike3)
```

```
## Source: local data frame [4 x 4]
##
##           sna avg_rate
##           (chr)      (dbl)
## 1      中和公園 13.56075
## 2 捷運永安市場站 17.51402
## 3 中和國民運動中心 27.63830
## 4      秀山國小 29.25234
##           sna
##           (chr)
## 1 捷運永安市場站
## 2      秀山國小
## 3      中和公園
## 4 中和國民運動中心
## Variables not shown: avg_anemo (dbl)
```

join

```
left_join(a,b,by=X1)  
right_join(a,b,by=X1)  
inner_join(a,b,by=X1)  
full_join(a,b,by=X1)
```



left_join

```
left_join(a,b,by=X1)
```

a

X1	X2
A	11
B	12
C	13

+

b

X1	X3
A	g
B	h
D	i

=

X1	X2	X3
A	11	g
B	12	h
C	13	NA

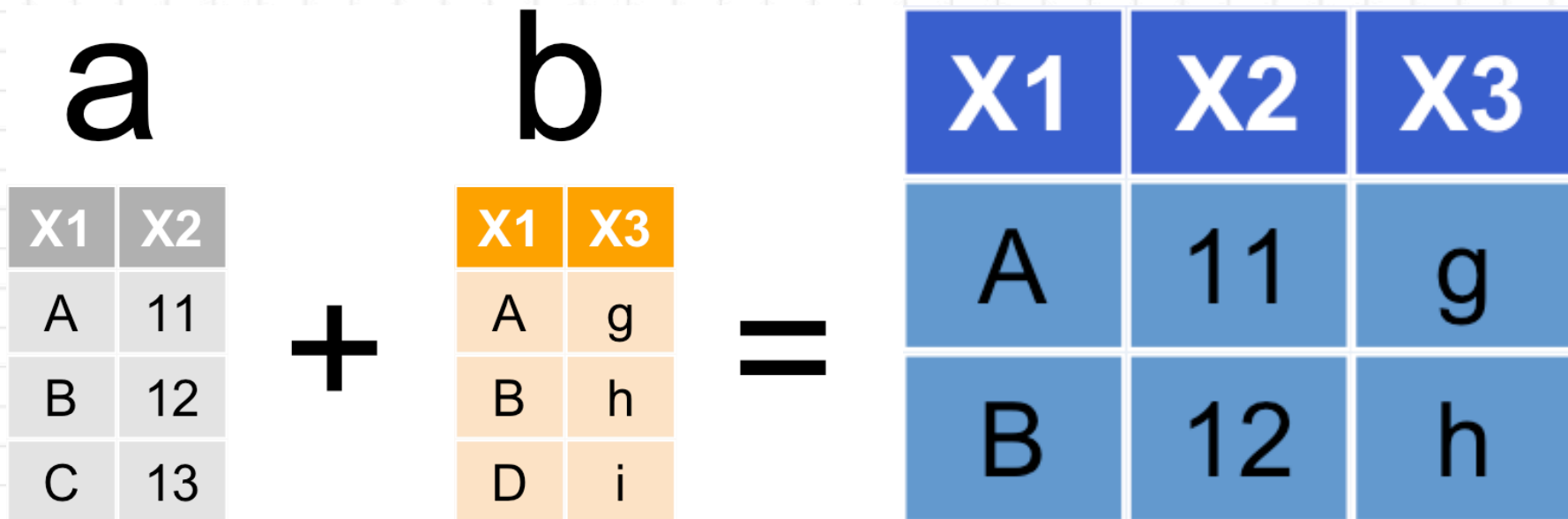
right_join

```
right_join(a,b,by=X1)
```

a		b						
X1	X2		X1	X3				
A	11	+	A	g	=			
B	12		B	h				
C	13		D	i				
X1	X2		X1	X3		X1	X3	X2
A	11		A	g		A	g	11
B	12		B	h		B	h	12
C	13		D	i		D	i	NA

inner_join

```
inner_join(a,b,by=X1)
```



full_join

```
full_join(a,b,by=X1)
```

a

X1	X2
A	11
B	12
C	13

+

b

X1	X3
A	g
B	h
D	i

=

X1	X2	X3
A	11	g
B	12	h
C	13	NA
D	NA	i

練習一下

```
left_join(ubike1,ubike3)
```

```
## Joining by: "sna"
```

```
## Source: local data frame [4 x 3]  
##  
##           sna avg_rate avg_anemo  
##           (chr)   (dbl)   (dbl)  
## 1      中和公園 13.56075  2.698973  
## 2 捷運永安市場站 17.51402  2.609531  
## 3 中和國民運動中心 27.63830  2.177059  
## 4      秀山國小 29.25234  3.033534
```

交集

```
intersect(a,b)
```

a		b			
X1	X2	X1	X2		
A	11	B	12		
B	12	C	13		
C	13	D	14		

+

X1	X2
B	12
C	13

=

X1	X2
B	12
C	13

聯集

```
union(a,b)
```

a		b			
X1	X2		X1	X2	
A	11	+	B	12	=
B	12		C	13	
C	13		D	14	
X1	X2		X1	X2	
A	11				
B	12				
C	13				
D	14				

差集

```
setdiff(a,b)
```

a

X1	X2
A	11
B	12
C	13

+

b

X1	X2
B	12
C	13
D	14

=

X1	X2
A	11

reshape2

reshape2

- melt
 - wide format -> long format
- cast
 - long format -> wide format
 - **dcast** for data.frame
 - **acast** for vector, matrix and array

melt

```
WP.melt=as.data.frame(WorldPhones)  
WP.melt$year <- rownames(WP.melt)  
WP.melt=melt(WP.melt,id='year')  
kable(head(WP.melt))
```

YEAR	VARIABLE	VALUE
1951	N.Amer	45939
1956	N.Amer	60423
1957	N.Amer	64721
1958	N.Amer	68484
1959	N.Amer	71799
1960	N.Amer	76036

cast

```
WP.cast=dcast(WP.melt,year~variable,value.var="value")  
kable(WP.cast)
```

YEAR	N.AMER	EUROPE	ASIA	S.AMER	OCEANIA	AFRICA	MID.AMER
1951	45939	21574	2876	1815	1646	89	555
1956	60423	29990	4708	2568	2366	1411	733
1957	64721	32510	5230	2695	2526	1546	773
1958	68484	35218	6662	2845	2691	1663	836
1959	71799	37598	6856	3000	2868	1769	911
1960	76036	40341	8220	3145	3054	1905	1008
1961	79831	43173	9053	3338	3224	2005	1076

練習一下

小明想知道中和區的腳踏車站晴天和雨天的使用率有何差別

- 提示
 - `filter`、`mutate`、`select`、`group_by`、`summarise`
 - `dcast`

SNA	晴天	雨天
捷運永安市場站	0.6671052	0.6483044
秀山國小	0.4966519	0.4436588
中和公園	0.6363115	0.5917228
中和國民運動中心	0.4571795	0.4603829

學習資源

- [Data Wrangling](#)
- [Introduction to dplyr](#)

Team Project & Lunch time