EDA with R

智庫驅動

Ben Chen

dsp.im

今日重點!!!

畫圖~

ggplot2

需要的套件

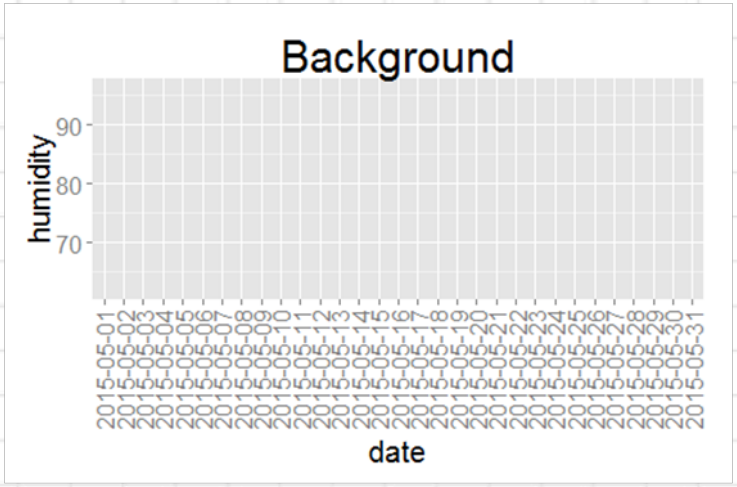
```
library(ggplot2)
library(data.table)
library(dplyr)
library(reshape2)
```

ggplot2 簡介

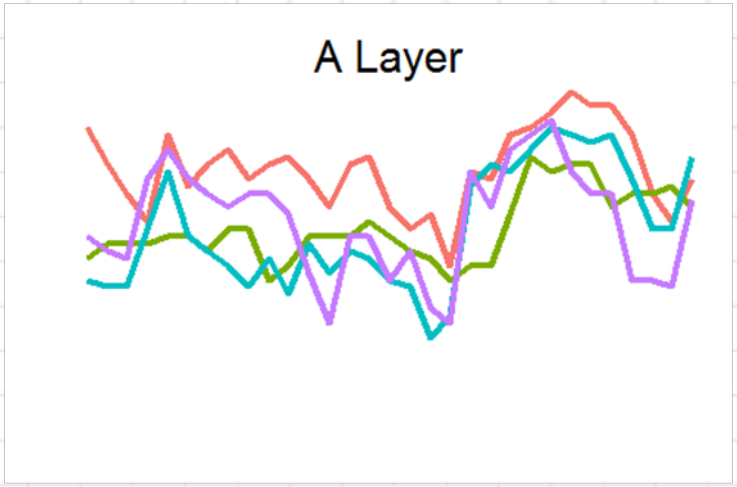
ggplot2簡介

- 2015年，最受歡迎的R套件之一
- R環境下的繪圖套件
- 取自 “The Grammar of Graphics” (Leland Wilkinson, 2005)
- 設計理念
  - 採用圖層系統
  - 用抽象的概念來控制圖形，避免細節繁瑣
  - 圖形美觀

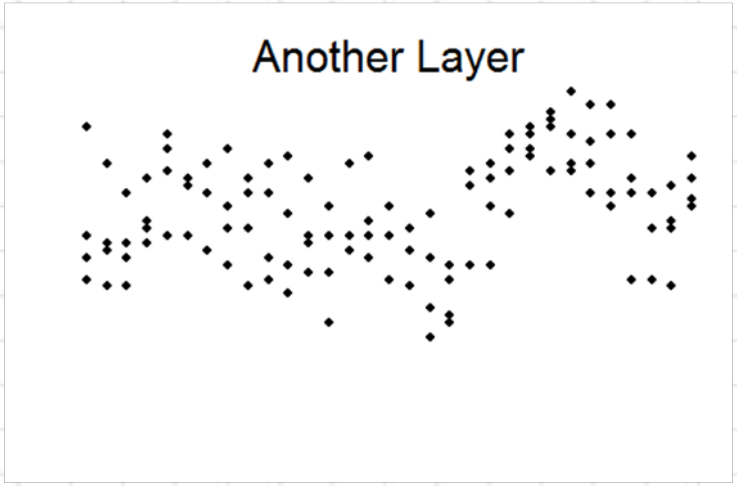
The Anatomy of a Plot



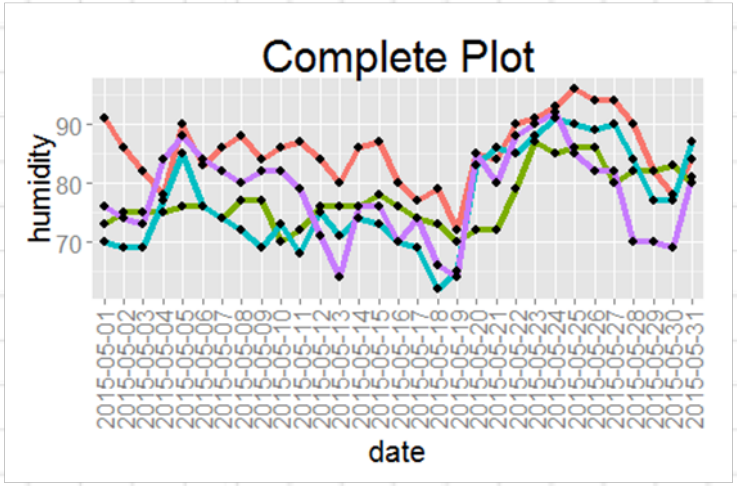
+



+



=



ggplot2核心

- 注意事項
  - 使用 data.frame 儲存資料 (不可以丟 matrix 物件)
  - 使用 long format (利用reshape2將資料轉換成 1 row = 1 observation)
- 基本語法
  - ggplot 描述 data 從哪來
  - aes 描述圖上的元素跟 data 之類的對應關係
  - geom_xxx 描述要畫圖的類型及相關調整的參數
  - 常用的類型諸如：geom_bar, geom_points, geom_line, geom_polygon

一切從讀檔開始 (CSV)

```
# 讀檔起手式
ubike = read.csv('ubikeweatherutf8.csv') #請輸入正確的檔案路徑
# 讀檔進階招式
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"))
# 讀檔大絕招
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"))
```

請輸入正確的檔案路徑

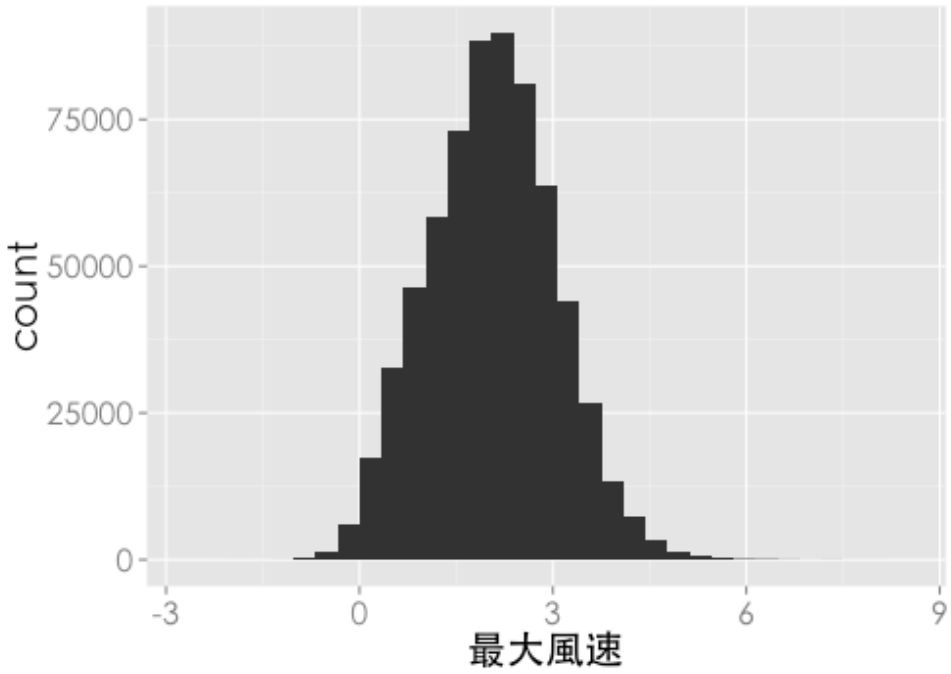
將欄位名稱換成中文

```
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")
```

單一數值：Histogram

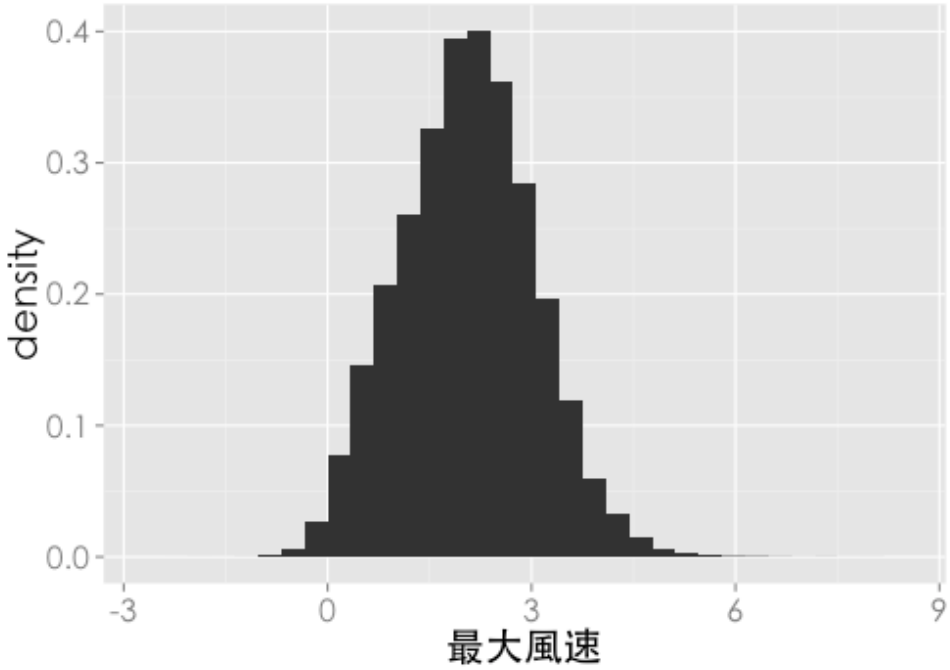
Histogram

```
thm <- theme(text=element_text(size=20,family="STHeiti")) # 控制字體與大小
# STHeiti是只有Mac才有的字體
ggplot(ubike) +
  geom_histogram(aes(x = 最大風速, y=..count..))+thm
```



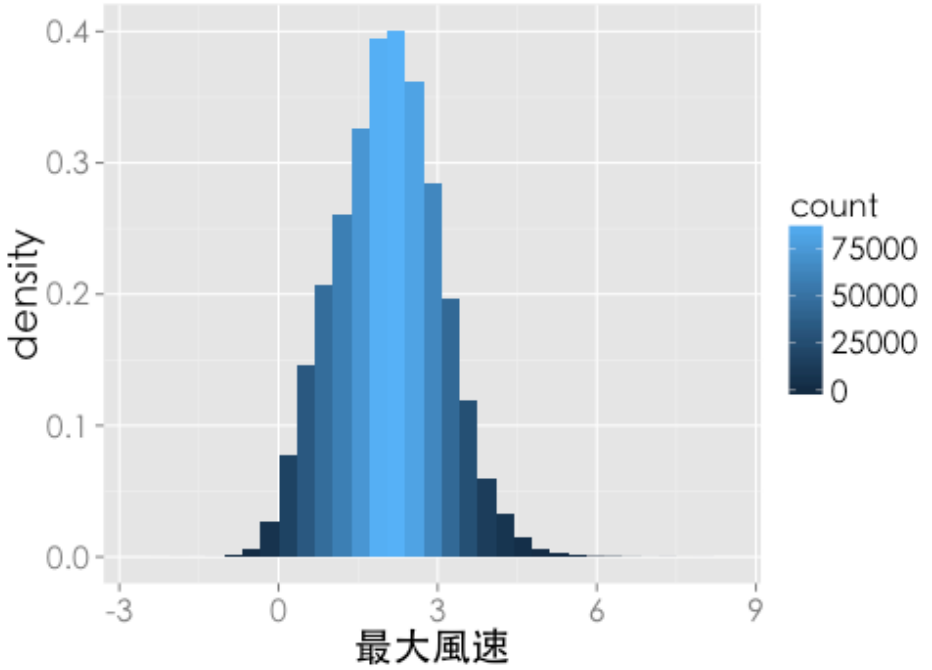
Histogram

```
ggplot(ubike) +
  geom_histogram(aes(x = 最大風速, y=..density..))+thm
```



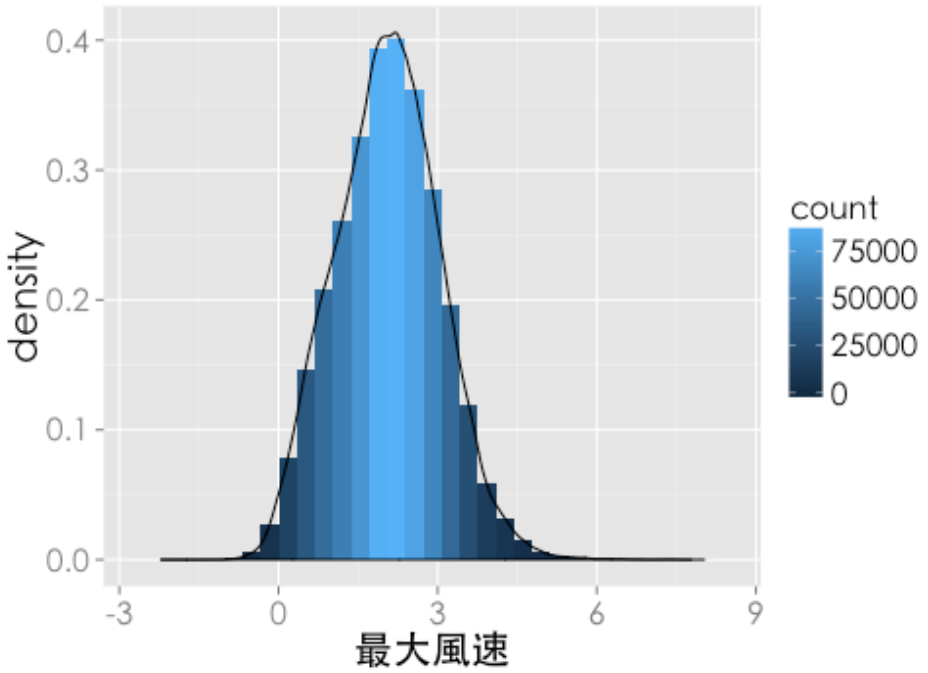
Histogram

```
ggplot(ubike) +
  geom_histogram(aes(x = 最大風速, y=..density..,fill=..count..))+thm
```



Histogram + Density

```
ggplot(ubike,aes(x = 最大風速)) +
  geom_histogram(aes(y=..density..,fill=..count..))+
  geom_density()+thm
```



量化 v.s. 量化：Scatter Plot

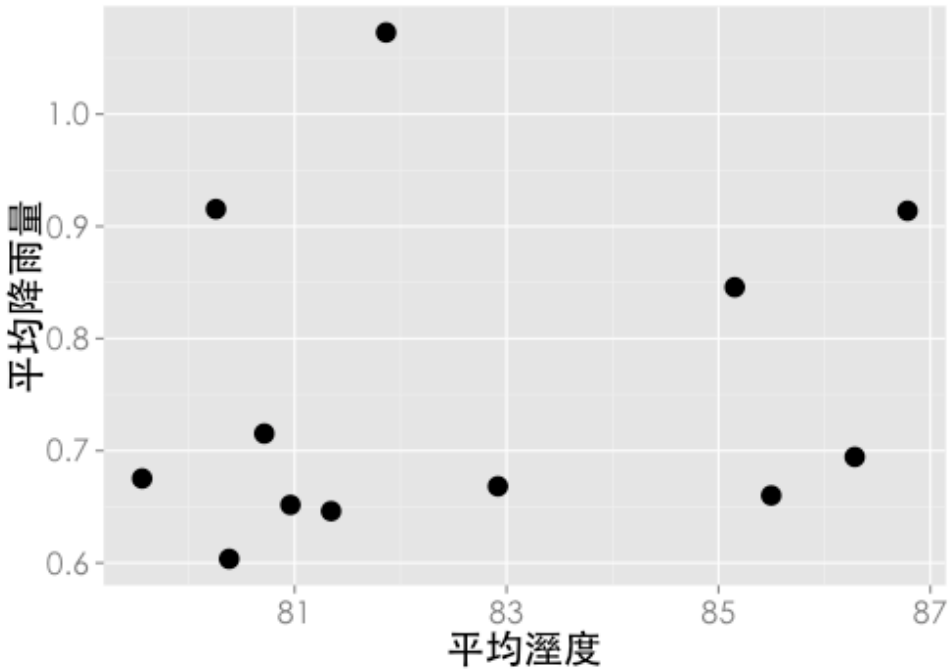
繪圖之前的整理資料

文山區各站點在"2015-02"的平均溼度 vs. 平均雨量

```
x3 <- filter(ubike, grepl("2015-02", 日期, fixed = TRUE), 場站區域 == "文山區") %>%
  group_by(場站名稱) %>%
  summarise(平均降雨量 = mean(降雨量), 平均溼度 = mean(溼度))
```

Scatter Plot

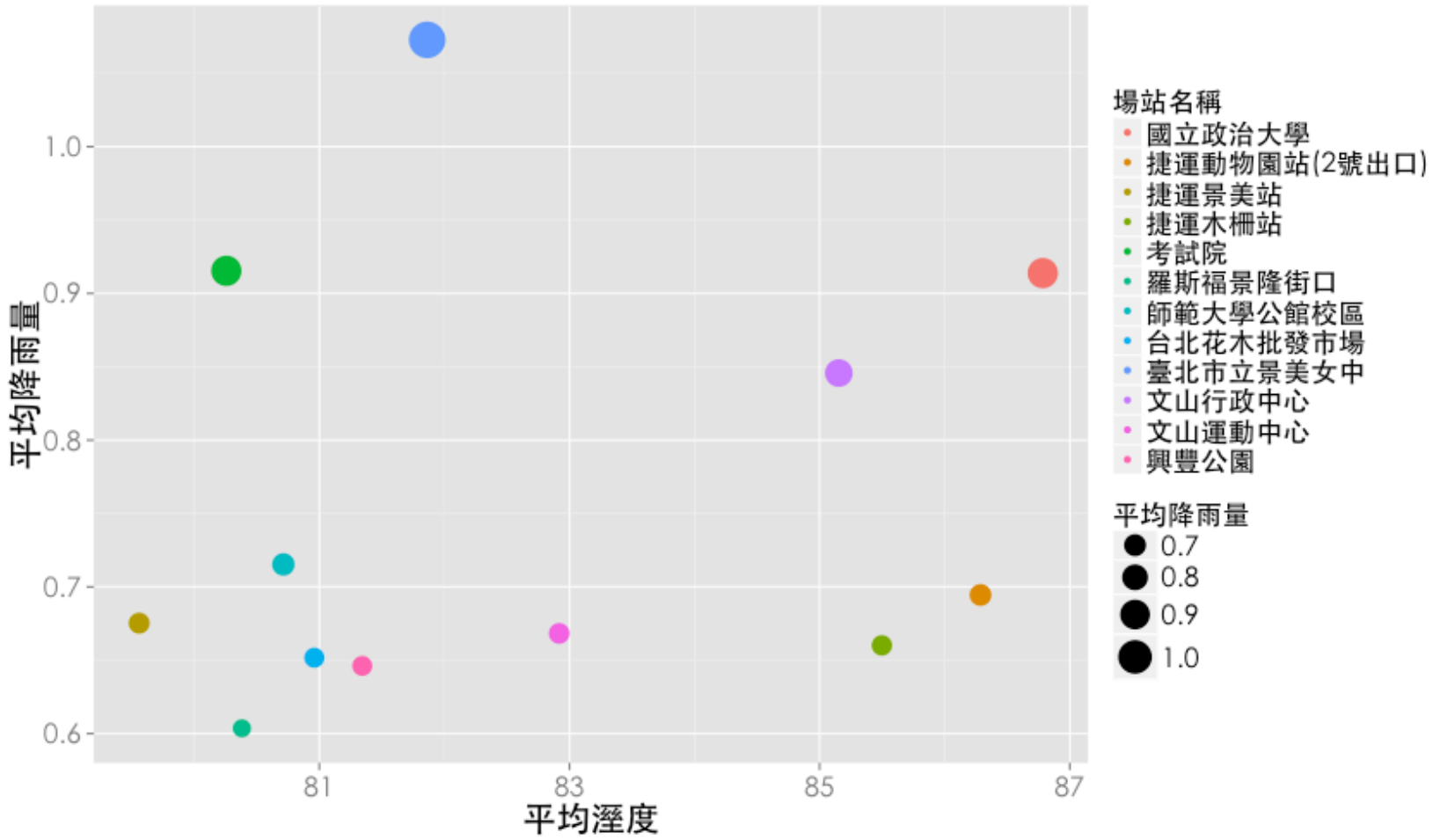
```
ggplot(x3) +
  geom_point(aes(x = 平均溼度, y = 平均降雨量),size=5) + #size控制點的大小
  thm
```



Grouped Scatter Plot

```
ggplot(x3) +
  # 放在aes裡的colour和size可依資料調整顏色和大小
  geom_point(aes(x = 平均溼度, y = 平均降雨量, colour = 場站名稱,size=平均降雨量))+
  # 限制大小
  scale_size(range=c(5,10)) +
  thm
```

Grouped Scatter Plot



量化 v.s. 量化：Line Chart

WorldPhones

```
##      N.Amer Europe Asia S.Amer Oceania
## 1951  45939  21574 2876   1815    1646
## 1956  60423  29990 4708   2568    2366
## 1957  64721  32510 5230   2695    2526
## 1958  68484  35218 6662   2845    2691
## 1959  71799  37598 6856   3000    2868
## 1960  76036  40341 8220   3145    3054
## 1961  79831  43173 9053   3338    3224
##      Africa Mid.Amer
## 1951     89      555
## 1956   1411      733
## 1957   1546      773
## 1958   1663      836
## 1959   1769      911
## 1960   1905     1008
## 1961   2005     1076
```

每年亞洲的電話數量

```
ggplot(WorldPhones,aes(x=?????,y=Asia))......
```

哪裏不對？

```
class(WorldPhones)
```

```
## [1] "matrix"
```

data.frame

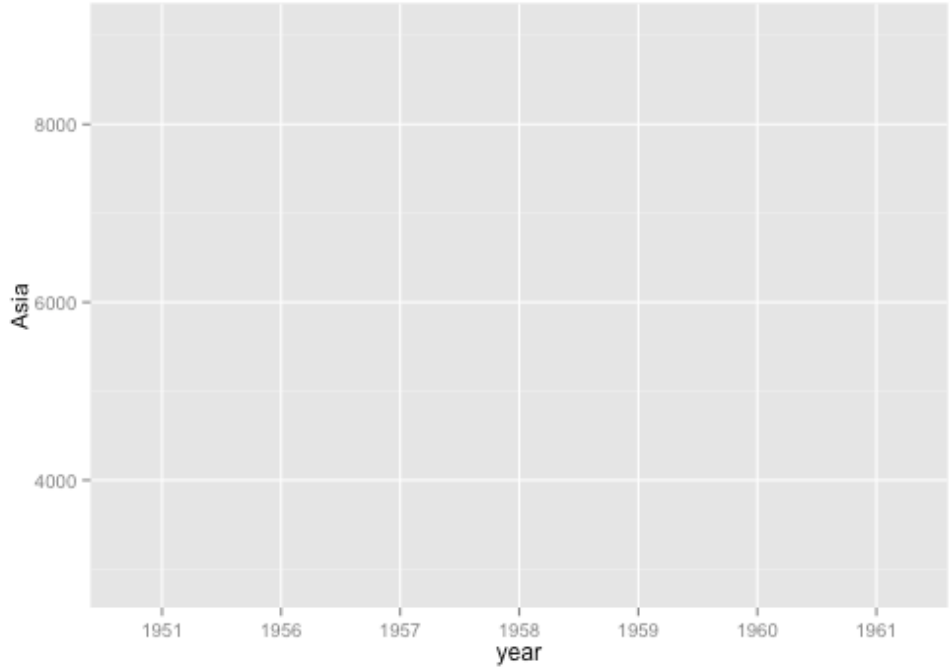
```
WP.df=as.data.frame(WorldPhones)
WP.df$year <- rownames(WP.df)
class(WP.df)
```

```
## [1] "data.frame"
```

Line Chart???

```
ggplot(WP.df,aes(x=year,y=Asia))+geom_line()
```

```
## geom_path: Each group consist of only one observation. Do you need to adjust the group aest
```



Should be Number

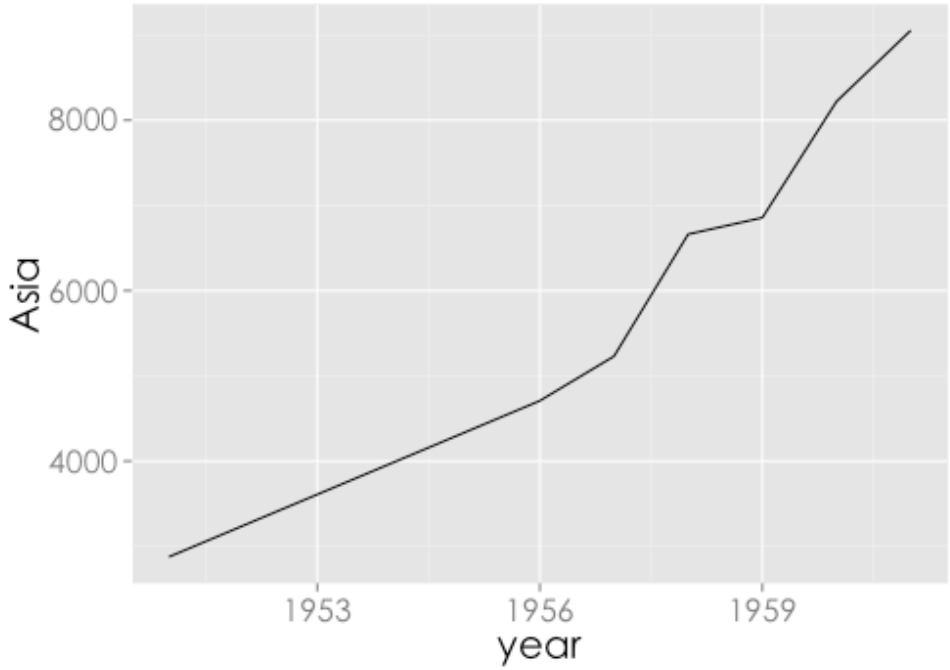
```
str(WP.df)
```

```
## 'data.frame':    7 obs. of  8 variables:
##  $ N.Amer  : num  45939 60423 64721 68484 71799 ...
##  $ Europe  : num  21574 29990 32510 35218 37598 ...
##  $ Asia    : num  2876 4708 5230 6662 6856 ...
##  $ S.Amer  : num  1815 2568 2695 2845 3000 ...
##  $ Oceania : num  1646 2366 2526 2691 2868 ...
##  $ Africa  : num  89 1411 1546 1663 1769 ...
##  $ Mid.Amer: num  555 733 773 836 911 ...
##  $ year    : chr  "1951" "1956" "1957" "1958" ...
```

```
WP.df$year=as.numeric(WP.df$year)
```

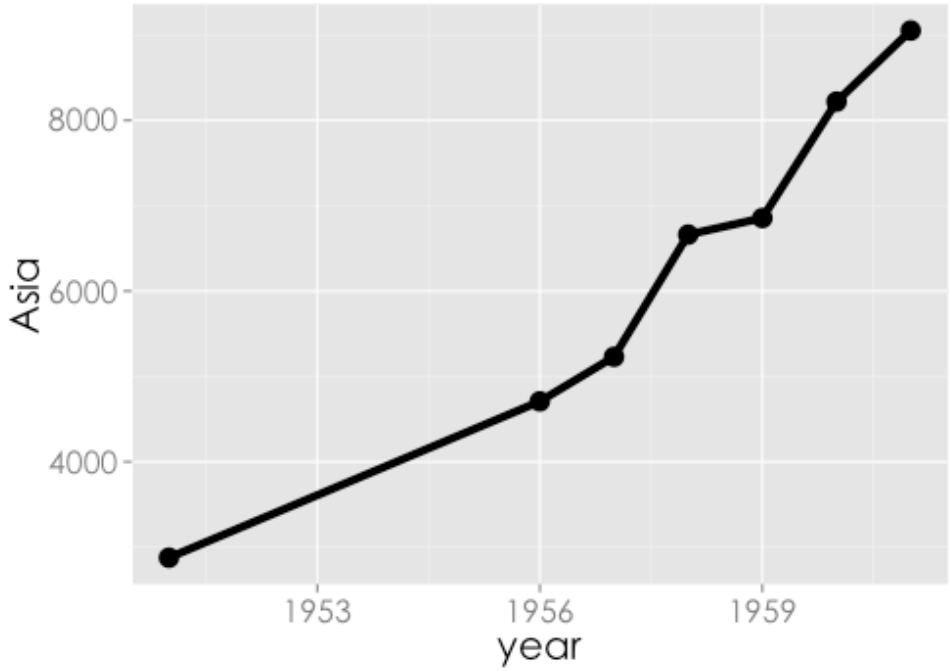
Line Chart!!!

```
ggplot(WP.df,aes(x=year,y=Asia))+
  geom_line()+thm
```



Line Chart and Scatter Ploet

```
ggplot(WP.df,aes(x=year,y=Asia))+
  geom_line(size=2)+ #size控制線的寬度或點的大小
  geom_point(size=5)+thm
```



How to plot multiple line?

Wide format

	N.AMER	EUROPE	ASIA	S.AMER	OCEANIA	AFRICA	MID.AMER	YEAR
1951	45939.00	21574.00	2876.00	1815.00	1646.00	89.00	555.00	1951.00
1956	60423.00	29990.00	4708.00	2568.00	2366.00	1411.00	733.00	1956.00
1957	64721.00	32510.00	5230.00	2695.00	2526.00	1546.00	773.00	1957.00
1958	68484.00	35218.00	6662.00	2845.00	2691.00	1663.00	836.00	1958.00
1959	71799.00	37598.00	6856.00	3000.00	2868.00	1769.00	911.00	1959.00
1960	76036.00	40341.00	8220.00	3145.00	3054.00	1905.00	1008.00	1960.00
1961	79831.00	43173.00	9053.00	3338.00	3224.00	2005.00	1076.00	1961.00

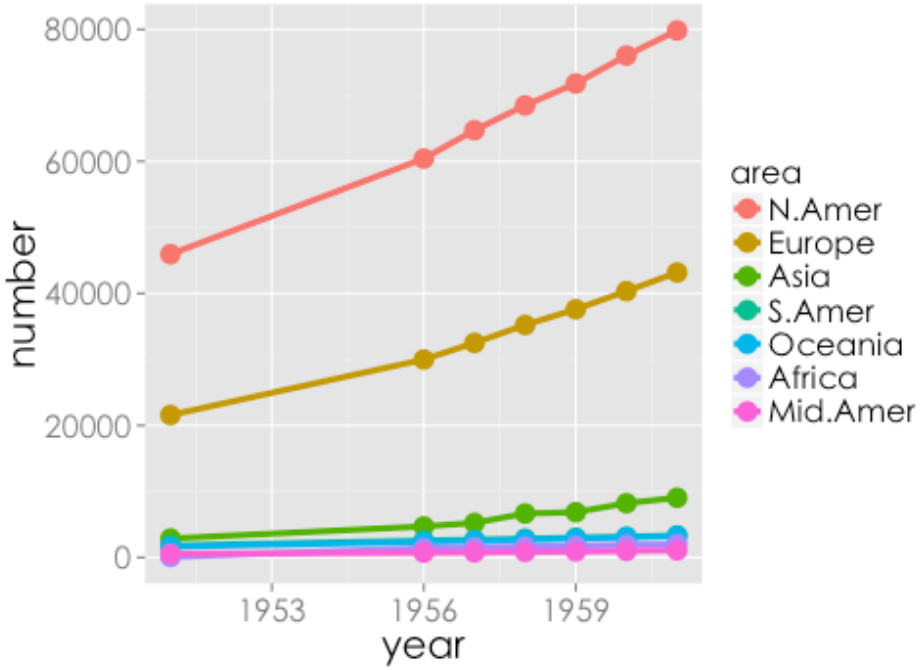
Long format

```
library(reshape2)
WP.long=melt(WP.df,id='year') #id是將保留的欄位名稱
colnames(WP.long)=c('year','area','number')
```

	YEAR	AREA	NUMBER
1	1951.00	N.Amer	45939.00
2	1956.00	N.Amer	60423.00
3	1957.00	N.Amer	64721.00
4	1958.00	N.Amer	68484.00
5	1959.00	N.Amer	71799.00
6	1960.00	N.Amer	76036.00
7	1961.00	N.Amer	79831.00
8	1951.00	Europe	21574.00
9	1956.00	Europe	29990.00
10	1957.00	Europe	32510.00
11	1958.00	Europe	35218.00
12	1959.00	Europe	37598.00

Multiple Line

```
ggplot(WP.long,aes(x=year,y=number,group=area,color=area))+ # gruop按照不同區域劃線
  geom_line(size=1.5)+
  geom_point(size=5)+thm
```



質化 v.s. 量化：Bar Chart

讀取檔案

```
pixnet=read.csv('train.csv',stringsAsFactors = FALSE)
```

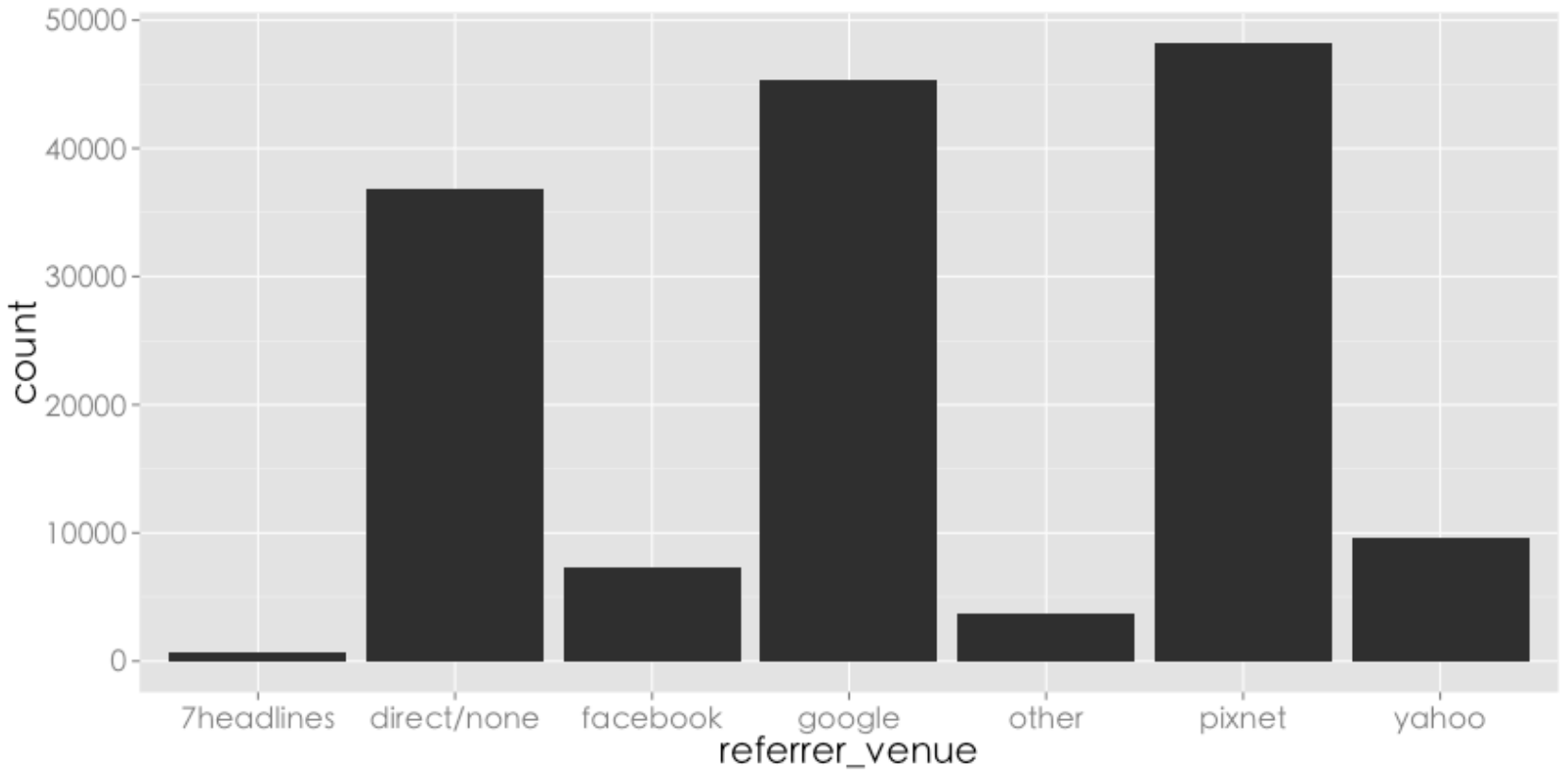
- 2014-11-01 至 2014-11-30 期間，10000 筆隨機取樣的台灣地區網站訪客的瀏覽紀錄

欄位說明

- url_hash - 去識別後的部落格文章 url
- resolution - 瀏覽裝置的螢幕解析度
- browser - 瀏覽裝置的瀏覽器
- os - 瀏覽裝置的作業系統
- device_marketing - 瀏覽裝置的產品型號
- device_brand - 瀏覽裝置的品牌名稱
- cookie_pta - 去識別化的瀏覽者代碼
- date - 瀏覽日期
- author_id - 文章作者 ID 去識別碼
- category_id - 文章分類
- referrer_venue - 訪客來源（網域）

Bar Chart

```
ggplot(pixnet,aes(x=referrer_venue))+
  geom_bar(stat='bin')+thm # stat='bin'算個數
```



兩種類別

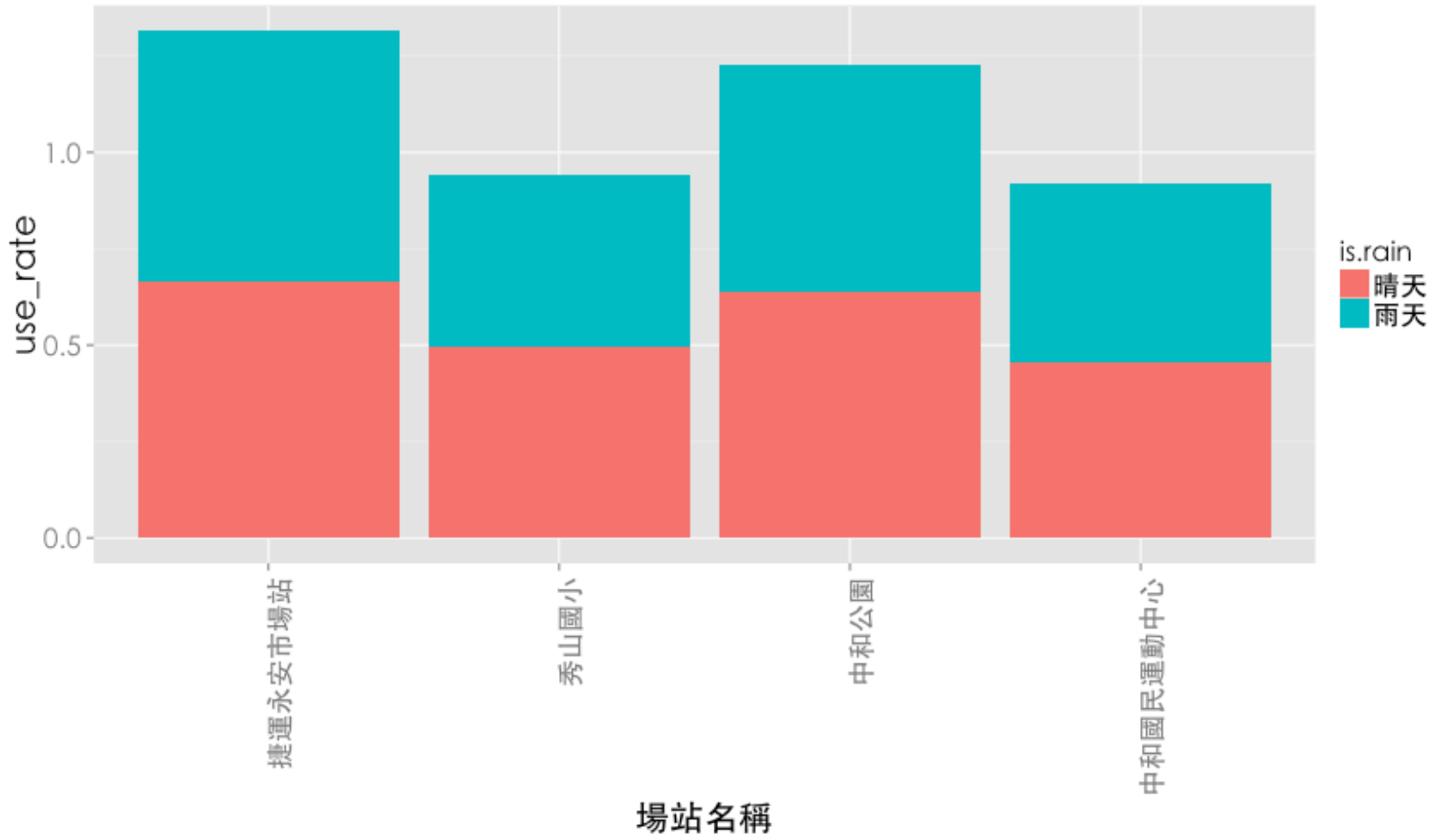
```
ub2=filter(ubike, 場站區域=='中和區',時間==8) %>%
  mutate(is.rain=降雨量>1) %>%
  mutate(is.rain=factor(is.rain, levels=c(FALSE, TRUE),
                        labels = c("晴天","雨天"))) %>%
  select(日期,  平均空位數, 場站名稱, is.rain,總停車格) %>%
  group_by(場站名稱,  is.rain) %>%
  summarise(use_rate=mean(平均空位數/總停車格))
head(ub2)
```

```
## Source: local data frame [6 x 3]
## Groups: 場站名稱 [3]
##
##          場站名稱 is.rain  use_rate
##            (fctr)  (fctr)     (dbl)
## 1 捷運永安市場站    晴天 0.6671052
## 2 捷運永安市場站    雨天 0.6483044
## 3        秀山國小    晴天 0.4966519
## 4        秀山國小    雨天 0.4436588
## 5        中和公園    晴天 0.6363115
## 6        中和公園    雨天 0.5917228
```

兩種類別

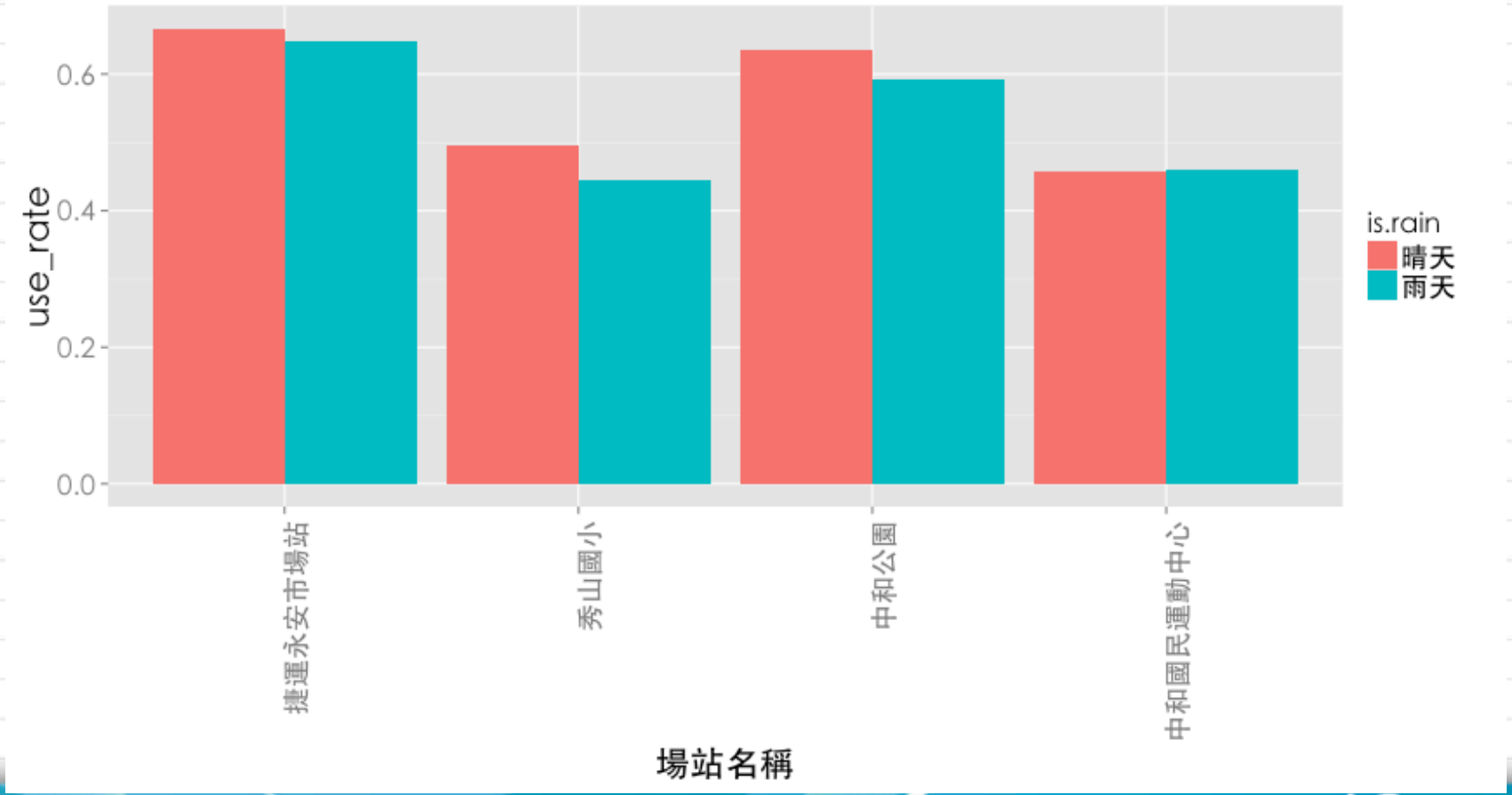
```
las2 <- theme(axis.text.x = element_text(angle = 90, hjust = 1),
              text=element_text(size=20,family="STHeiti")) #控制字的方向
ggplot(ub2,aes(x=場站名稱,y=use_rate,fill=is.rain))+
  geom_bar(stat='identity')+
  las2 # stat='identity'以表格的值做為bar的高度
```

兩種類別: stack


use_rate
1.0
0.5
0.0
捷運永安市場站
秀山國小
中和公園
中和國民運動中心
場站名稱
is.rain
晴天
雨天

兩種類別: dodge

```
ggplot(ub2,aes(x=場站名稱,y=use_rate,fill=is.rain))+
  geom_bar(stat='identity',position = 'dodge')+las2 #dodge類別並排
```

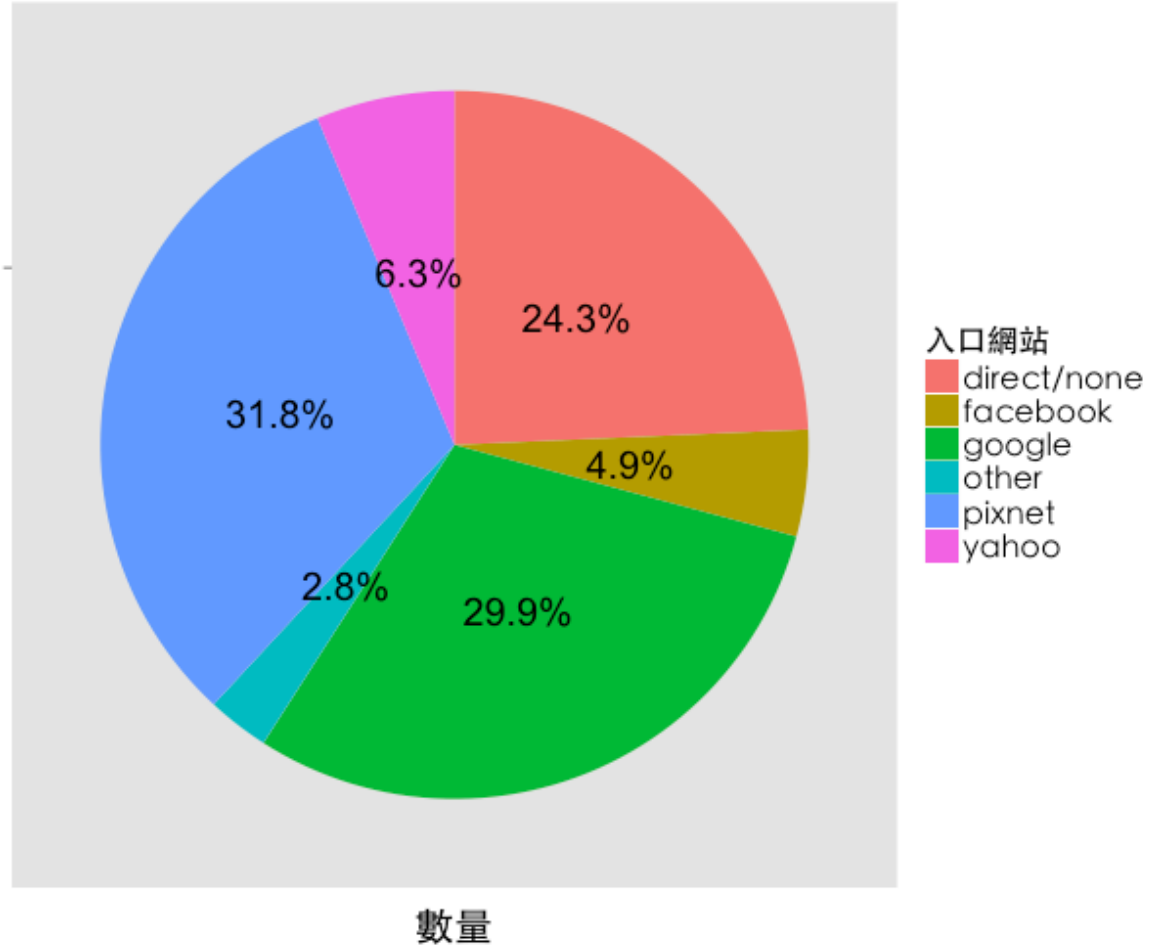


Pie Chart: Bar Chart變形

整理資料

```
pix=data.frame(table(pixnet$referrer_venue)) #table可以算個類別個數
colnames(pix)=c('入口網站','數量')
pix[5,2]=pix[5,2]+pix[1,2]
pix=pix[-1,]
```

Pie Chart: Bar Chart變形



Pie Chart: Bar Chart變形

```
ggplot(pix,aes(x="",y=數量,fill=入口網站))+
  geom_bar(stat='identity',width=1)+
  coord_polar('y')+
  geom_text(aes(y = 數量*0.5+ c(0, cumsum(數量)[-length(數量)]),
                label = paste(round(數量/sum(數量),3)*100,'%',sep="")),
            size=7)+
  theme(axis.title.y = element_blank(),
        axis.text.x=element_blank(),
        panel.grid=element_blank(),
        text=element_text(size=20,family="STHeiti"))
```

The Grammer of Graphics

ggplot2基本架構

- 資料 (data) 和映射 (mapping)
- 幾何對象 (geometric)
- 座標尺度 (scale)
- 統計轉換 (statistics)
- 座標系統 (coordinante)
- 圖層 (layer)
- 刻面 (facet)
- 主題 (theme)

Data and Mapping

```
ggplot(data=WP.df)+geom_line(aes(x=year,y=Asia))
```

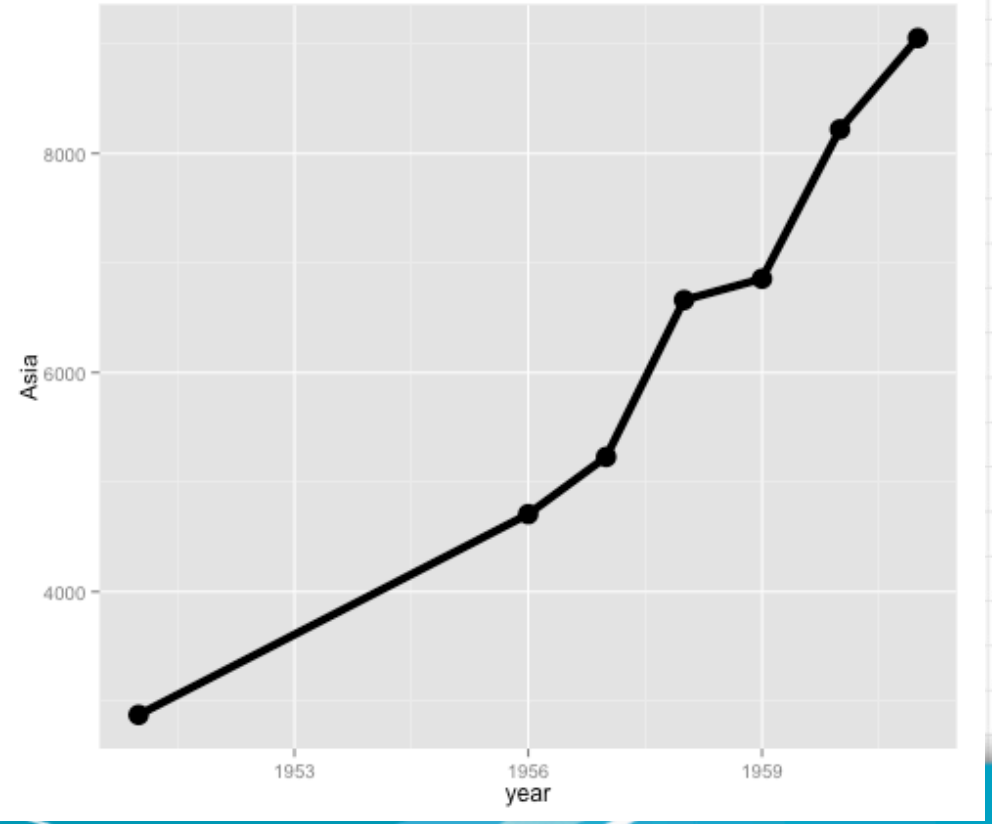
Data is Data

mapping: aes(x=...,y=...)

geometric

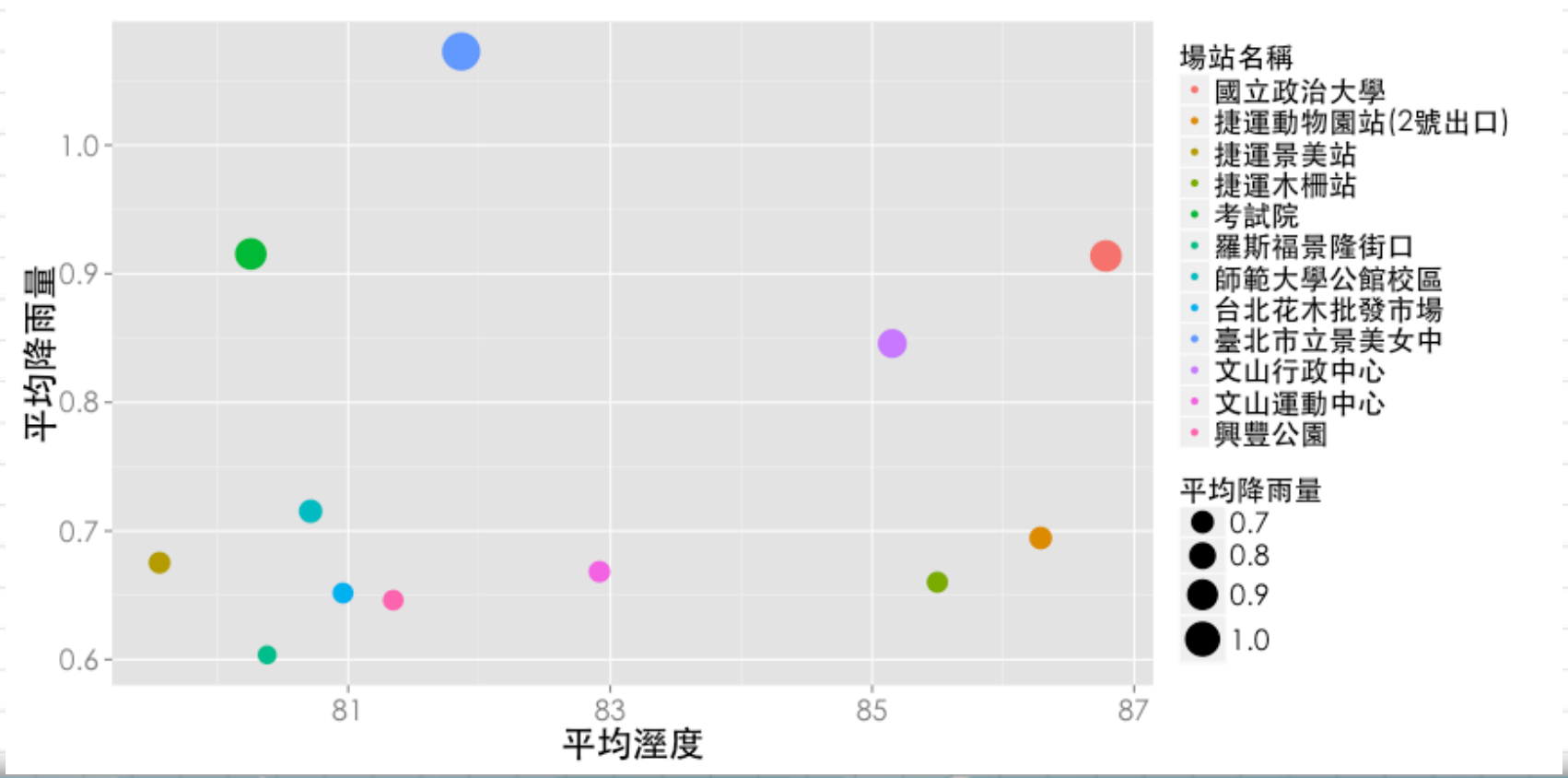
geom_line and geom_point

```
ggplot(WP.df,aes(x=year,y=Asia))+
  geom_line(size=2)+geom_point(size=5)
```



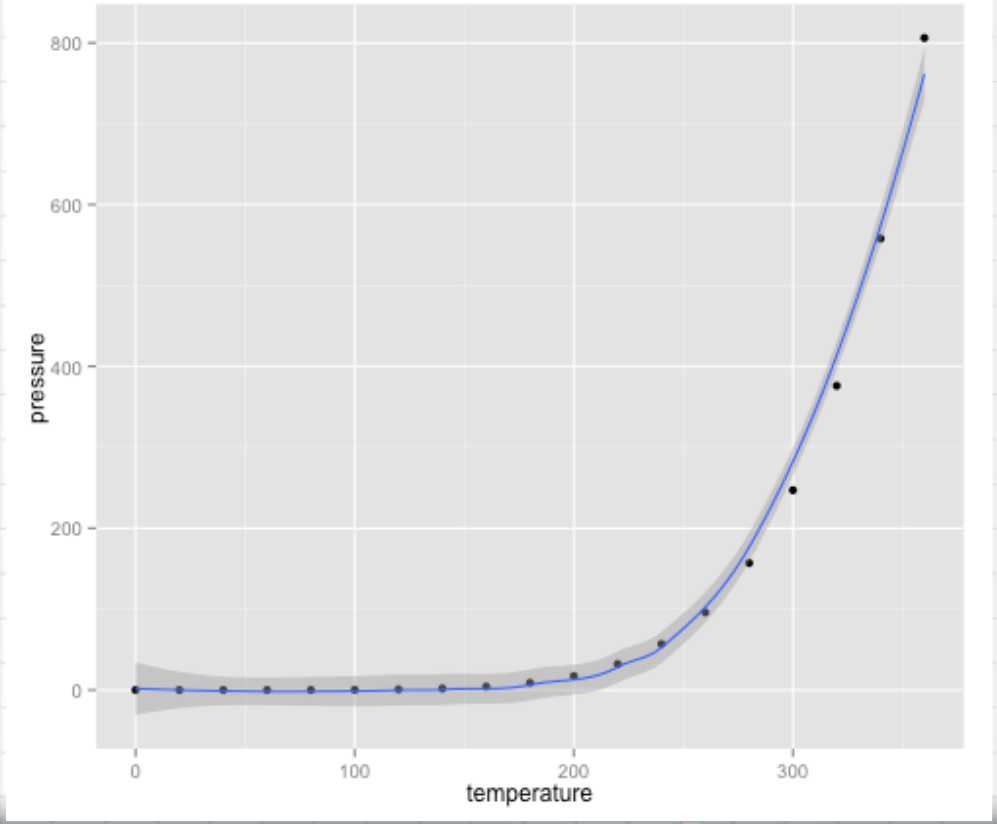
scale

```
ggplot(x3) +
  geom_point(aes(x =平均溼度, y=平均降雨量,colour=場站名稱,size=平均降雨量))+
  scale_size(range=c(5,10)) +thm
```



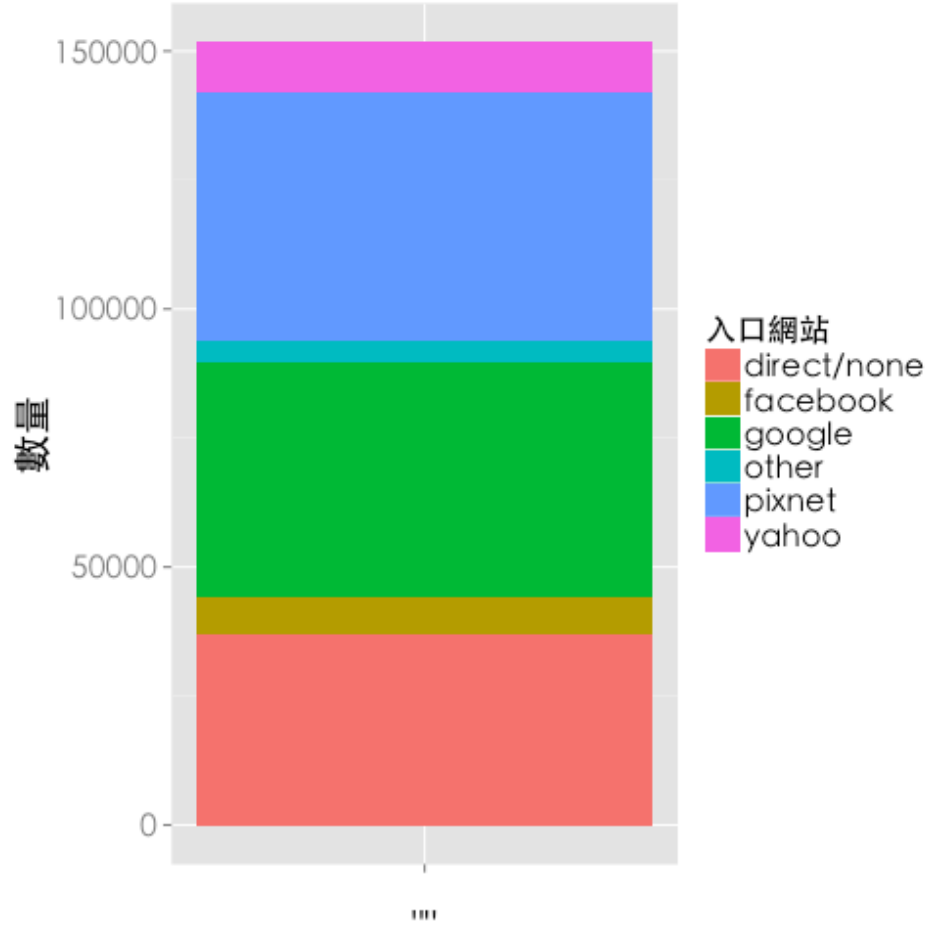
statistics

```
ggplot(pressure,aes(x=temperature,y=pressure))+
 geom_point()+
 stat_smooth()
```



coordinante

```
ggplot(pix,aes(x="",y=數量,fill=入口網站))+
  geom_bar(stat='identity')+thm
```




```
ggplot(pix,aes(x="",y=數量,fill=入口網站))+
  geom_bar(stat='identity',width=1)+
  coord_polar('y')+thm
```

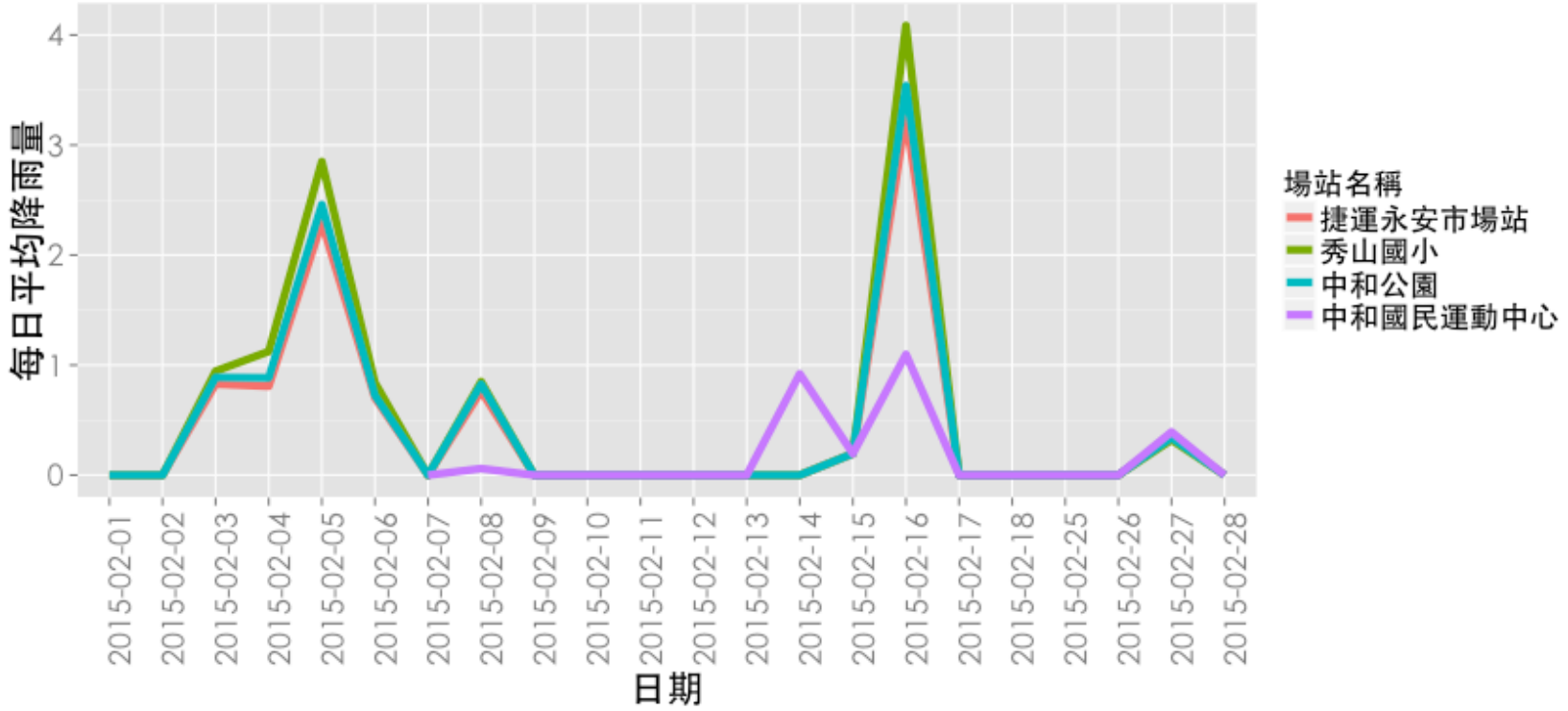
facet

```
rain <- filter(ubike, grepl("2015-02", 日期, fixed = TRUE), 場站區域 == "中和區") %>%
  group_by(日期,場站名稱) %>%
  summarise(每日平均降雨量 = mean(降雨量))
```

facet

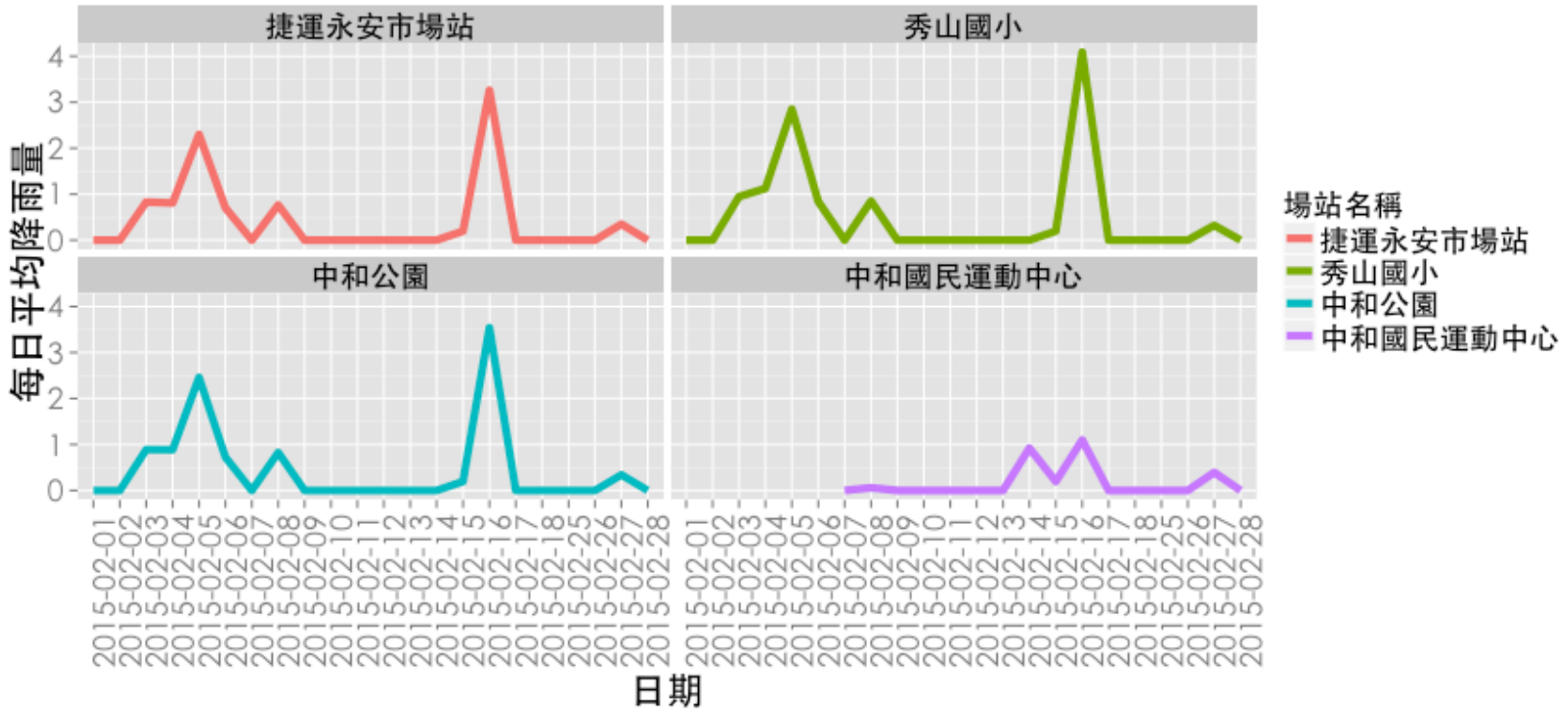
Line Chart

```
ggplot(rain) + thm+las2+
  geom_line(aes(x = 日期, y = 每日平均降雨量,group=場站名稱,colour=場站名稱),size=2)
```



Line Chart in Facets

```
ggplot(rain) +thm+las2+facet_wrap(~場站名稱,nrow=2)+ # facet_wrap將各站的情況分開畫
  geom_line(aes(x = 日期, y = 每日平均降雨量,group=場站名稱,colour=場站名稱),size=2)
```



可以存檔嗎？

存檔

```
# 畫完圖之後，再存檔~~
ggsave('檔案名稱')
```

學習資源

- ggplot2 cheat sheet from RStudio Inc.
- ggplot2 官方文件

本週目標

環境設定

- 建立可以使用R 的環境
- 了解R 的使用界面

學習R 語言

- 透過實際的範例學習R 語言
  - 讀取資料
  - 選取資料
  - 敘述統計量與視覺化

掌握心法後，如何自行利用R 解決問題

- 了解自己的需求
- 詢問關鍵字與函數
  - 歡迎來信 benjamin0901@gmail.com 或其他教師
  - 多多交流
    - Taiwan R User Group ，mailing list: Taiwan-useR-Group-list@meetup.com
    - ptt R_Language版
    - R軟體使用者論壇
  - sos套件，請見Demo

Team Project