R-Ladies Taipei

R-basic Lesson 8

Descriptive Statistics
敘述統計

Ricci Chen , 黃舒瑜 , 蔡旻均

2017.8.28

R-Ladies Taipei

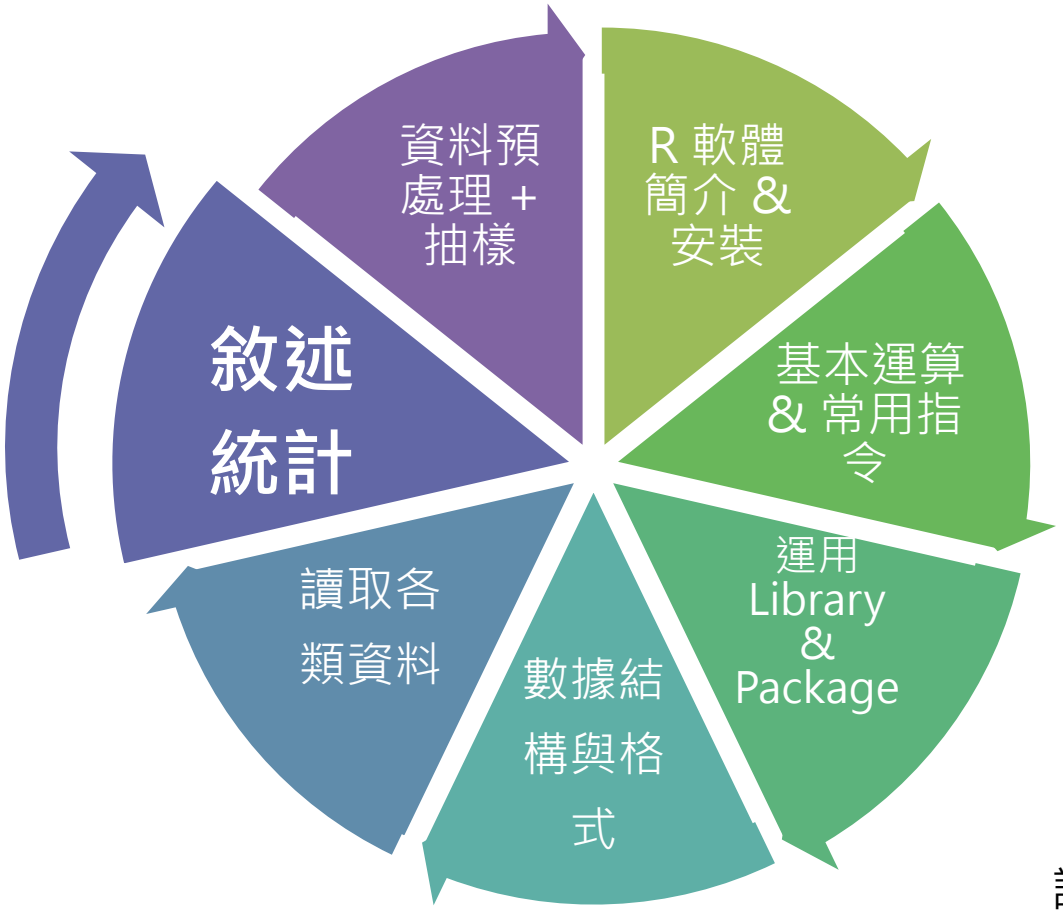
Quick Pickup

新朋友：

中途來聽可以嗎？

R-Ladies Taipei

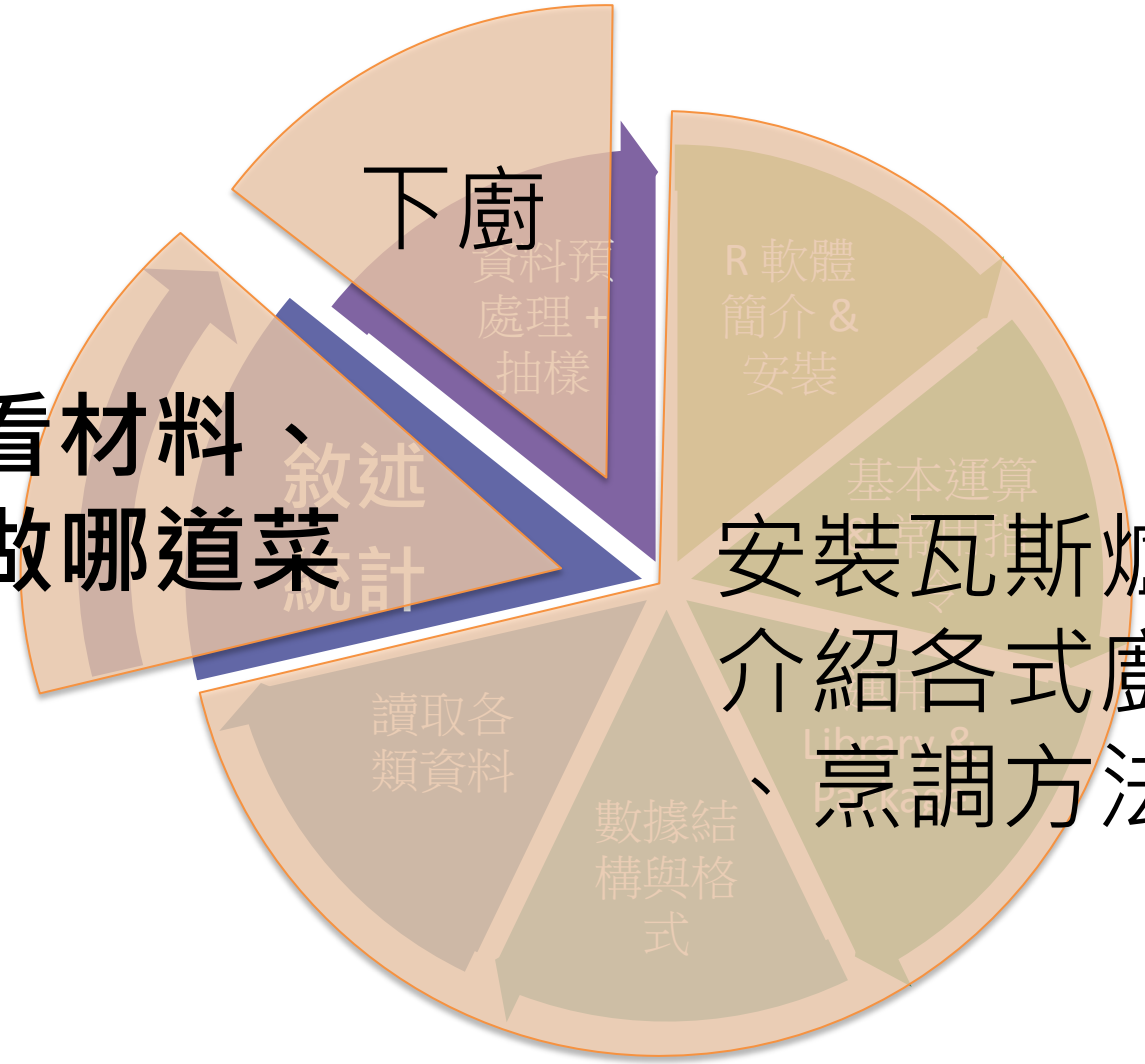
2017 R-basic 提供了下列課程



課程簡報在這裡

https://rladiestaipei.github.io/R-Ladies-Taipei/

如果用做菜來比喻



看看材料、要做哪道菜

安裝瓦斯爐，介紹各式廚具、烹調方法

R-Ladies Taipei

課程內容深入淺出

如何在Windows和 Linux環境安裝 R ，設定 R Studio

如何利用索引值來選取資料、修改資料 ；使用DPLYR 整理結構化資料

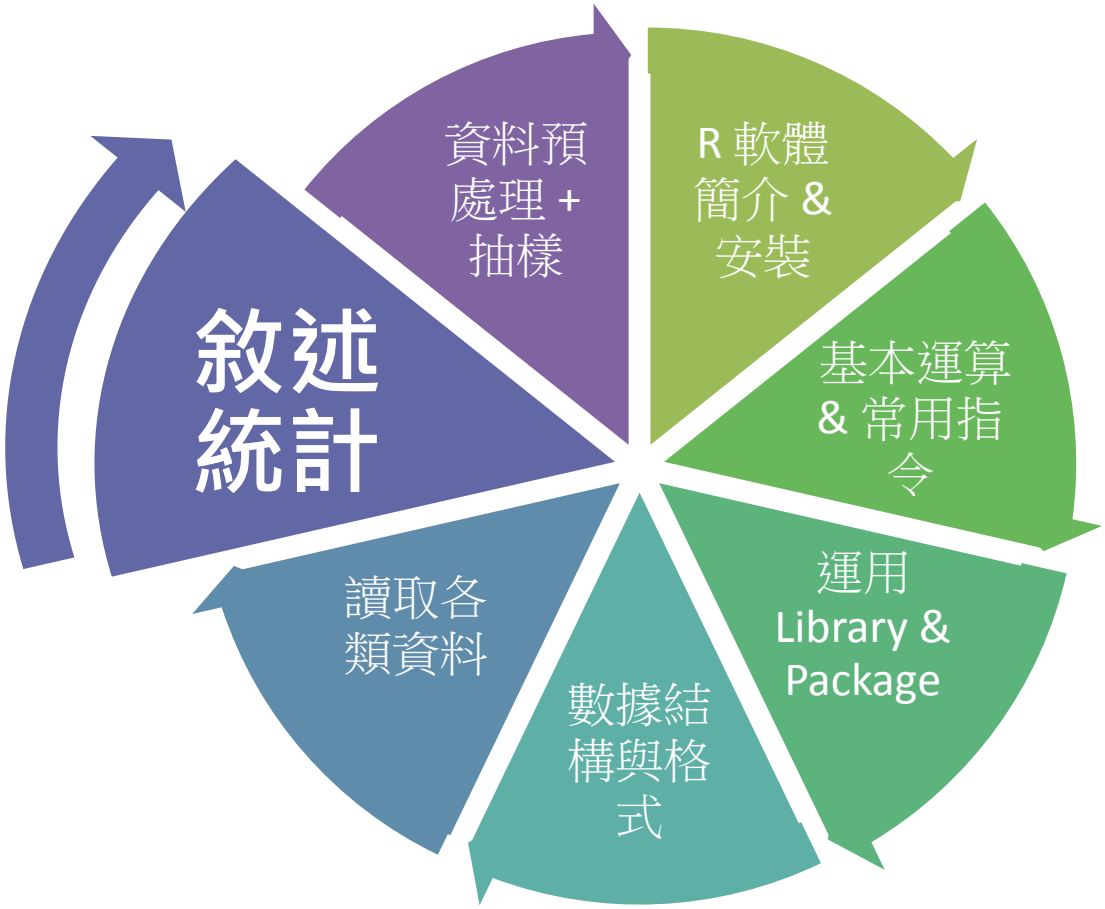
如何安裝、使用、建立package。超人氣package: ggplot2, reshape, stringr

如何有效率地讀存又多又大的檔案；設定編碼環境；特殊檔案讀取 Shape / JSON / MongoDB

介紹資料格式(data type)，如何查詢、強制轉換。介紹資料結構(data structure)，以及如何查詢、抽取、合併。

R-Ladies Taipei

今晚課程



Descriptive Statistics 敘述統計

- What is Descriptive Statistics for ?
  敘述統計在敘述甚麼？
- Measurement　認識手上的材料
- Summary　資料概述
- The squeezed M curve　統計迷思?
- Statistic Plot　常見的統計圖表
- Scatter Plot　散布圖的秘密

R-Ladies Taipei

What is Descriptive Statistics for ?

敘述統計在敘述甚麼？

R-Ladies Taipei

Wikipedia

- 統計學是在資料分析的基礎上，研究如何測定、收集、整理、歸納和分析反映資料資料，以便給出正確訊息的科學。

- 隨著巨量資料（Big Data）時代來臨，與資訊、計算等領域密切結合，是資料科學（Data Science）中的重要主軸之一。

R-Ladies Taipei

Wikipedia

- 自一組資料中，可以摘要並描述這份資料的集中和離散情形，這個用法稱作為描述統計學。

- 觀察者以資料的形態，建立出一個用以解釋其隨機性和不確定性的數學模型，以之來推論研究中的步驟及母體，這種用法被稱做推論統計學。

R-Ladies Taipei

給定一片森林，
如何精確地描述一枚不確定的樹葉？

R-Ladies Taipei

描述特徵

- 以圖表直觀了解整體資料分佈
  - 機率密度函數圖、直方圖、散布圖、盒鬚圖...等
- 以數據觀察集中和離散的情形
  - 集中：

    平均數、眾數、中位數/四分位數/百分位數...等
  - 分散：

    全距(最大值 - 最小值)、標準差/變異數、四分位差、平均絕對離差(MAD)...等

R-Ladies Taipei

初步資料探索

- 各個變數特性/樣本分布情形
- 是否需要變數轉換
- 無效值、異常值處理決策
- 迅速了解變數之間是否線性相關，可作為降維(變數刪減)參考

R-Ladies Taipei

Measurement

認識手上的材料

R-Ladies Taipei

Measurement 衡量尺度

資料衡量尺度	變數形態	特性
名目資料(nomial)	質化	類別
順序資料(ordinal)	質化	優先順序
區間資料(interval)	量化	大小距離
比例資料(ratio)	量化	比值

認識手上的材料

處理過的資料：分類/分組後的平均或加總

ex.氣象資料

項目	溫度(℃)			雨量	風速(公尺/秒)/風向(360°)/日期		相對溼度(%)		測站氣壓	降水日數>=0.1毫米	日照時數
測站	平均	最高/日期	最低/日期	(毫米)	最大十分鐘風	最大瞬間風	平均	最小/日期	(百帕)	(天)	(小時)
阿里山	15.0	22.0/30	10.3/14	940.3	6.3/170.0/31	16.3/180.0/31	93	49/19	764.6	24	112.7
鞍部	23.6	30.7/5	19.1/7	160.0	15.4/190.0/29	38.9/120.0/29	85	56/20	918.3	11	158.8
板橋	29.7	37.5/19	24.3/2	212.6	11.9/70.0/29	28.4/80.0/29	72	41/19	1006.8	17	205.2
成功	28.1	33.2/28	23.2/16	205.1	12.3/210.0/29	27.2/200.0/29	79	55/29	1004.4	10	255.2

R-Ladies Taipei

認識手上的材料

處理過的資料：分類/分組後的平均或加總

ex.財務報表

台積電簡明財報

2017年1~6月

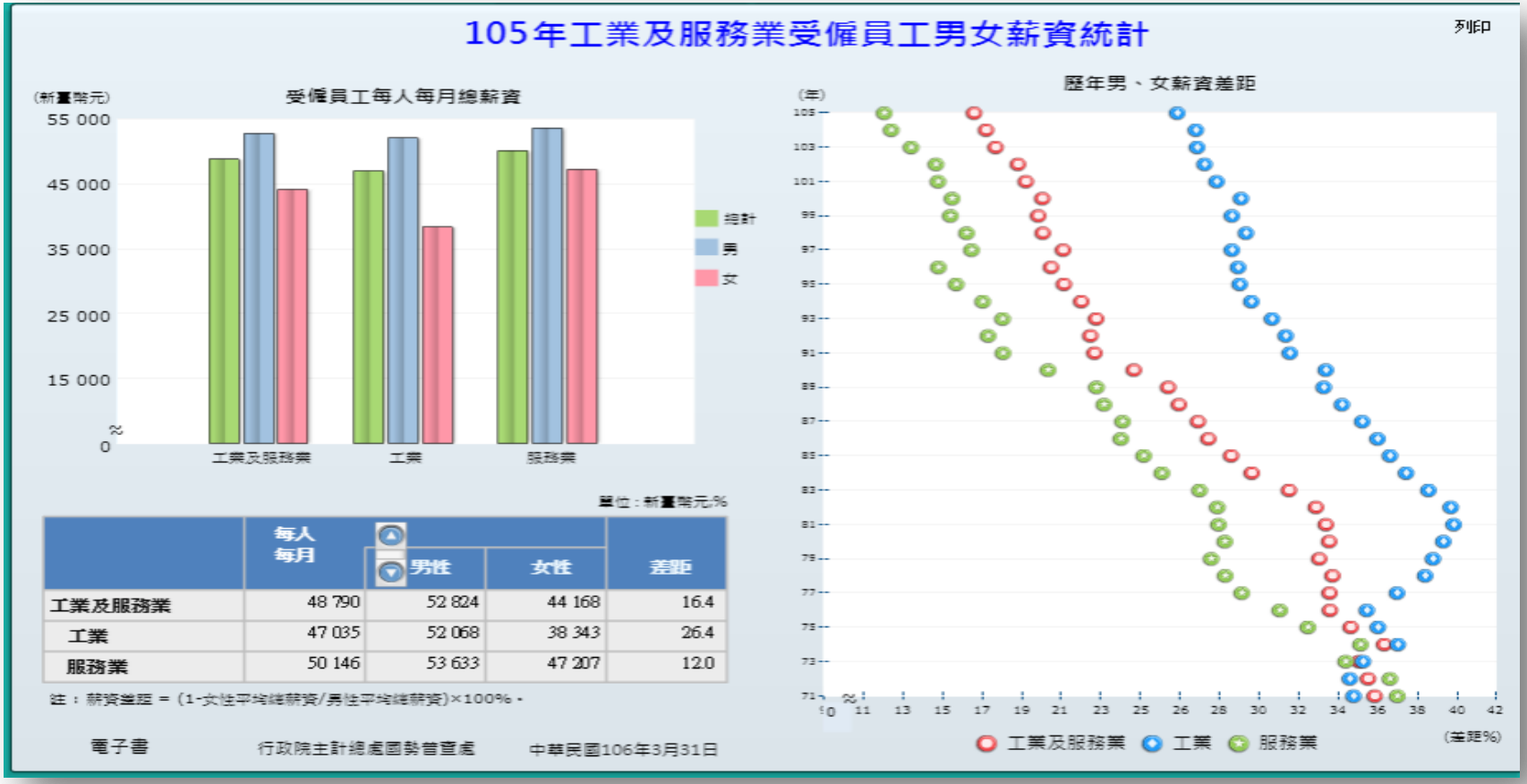
金額單位：新台幣仟元

	科目	2017Q2	2017Q1	2016Q4	2016Q3	2016Q2
股東權益	股本年合併	259,303,805.00	259,303,805.00	259,303,805.00	259,303,805.00	259,303,805.00
	資本公積	56,282,780.00	56,282,118.00	56,272,304.00	56,269,958.00	56,263,141.00
	保留盈餘	1,044,395,423.00	1,159,637,067.00	1,072,008,169.00	972,758,173.00	875,999,117.00
	庫藏股					
	權益調整					
	股東權益	1,342,365,645.00	1,456,310,025.00	1,390,051,126.00	1,283,892,649.00	1,197,330,978.00
損益	營業收入	447,769,610.00	233,914,400.00	947,938,340.00	685,711,090.00	425,305,210.00
	營業毛利	230,238,909.00	121,485,666.00	812,611,955.00	337,750,784.00	205,711,712.00
	營業損益	178,607,791.00	95,352,390.00	646,008,215.00	268,050,437.00	161,788,315.00
	業外收入	5,332,175.00	2,470,109.00	13,863,070.00	5,861,468.00	3,873,975.00
	業外支出					
	稅前盈餘	183,939,966.00	97,822,499.00	659,871,285.00	273,911,905.00	165,662,290.00
	稅後盈餘	153,891,560.00	87,620,908.00	568,448,225.00	234,109,989.00	137,320,876.00

認識手上的材料

處理過的資料：分類/分組後的平均或加總

ex.主計處資料

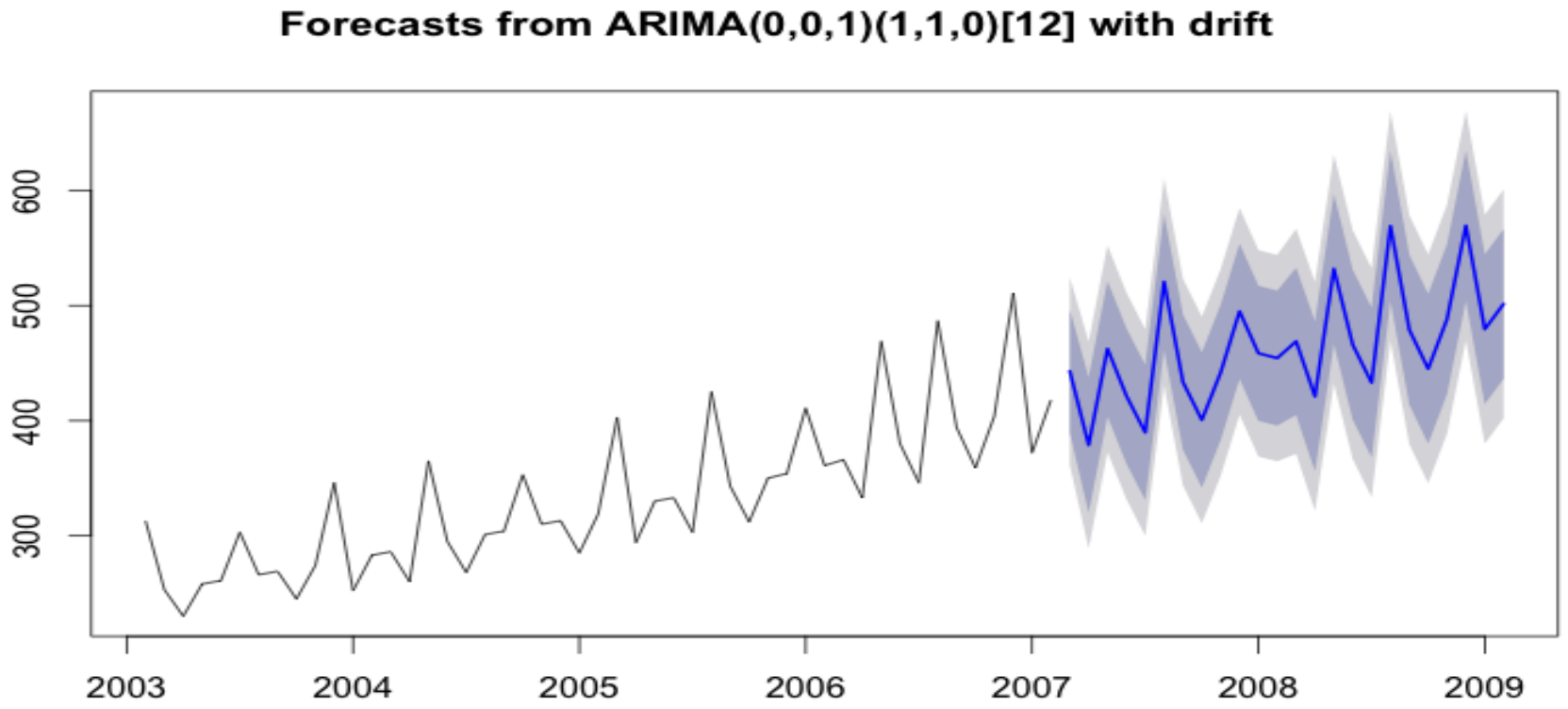


R-Ladies Taipei

認識手上的材料

時間資料：處理與時間相關的波動和趨勢

ex.時間序列預測模型ARIMA



R-Ladies Taipei

認識手上的材料

長字串，

ex. 記者的文筆和捐款有沒有關係？

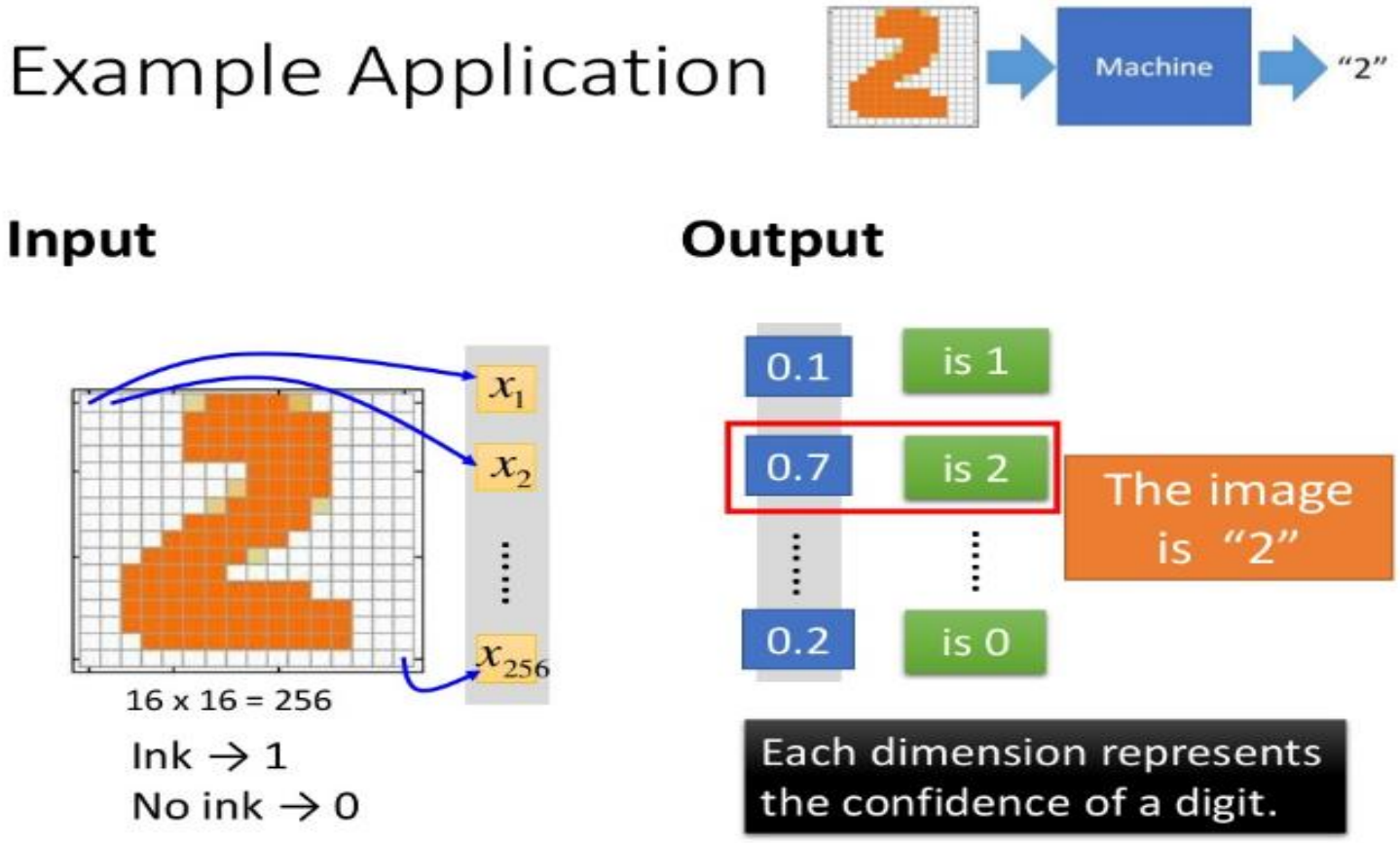
透過 Text Mining 了解文章的遣詞用字
如何影響人們的捐款行為

R-Ladies Taipei

認識手上的材料

圖像辨識

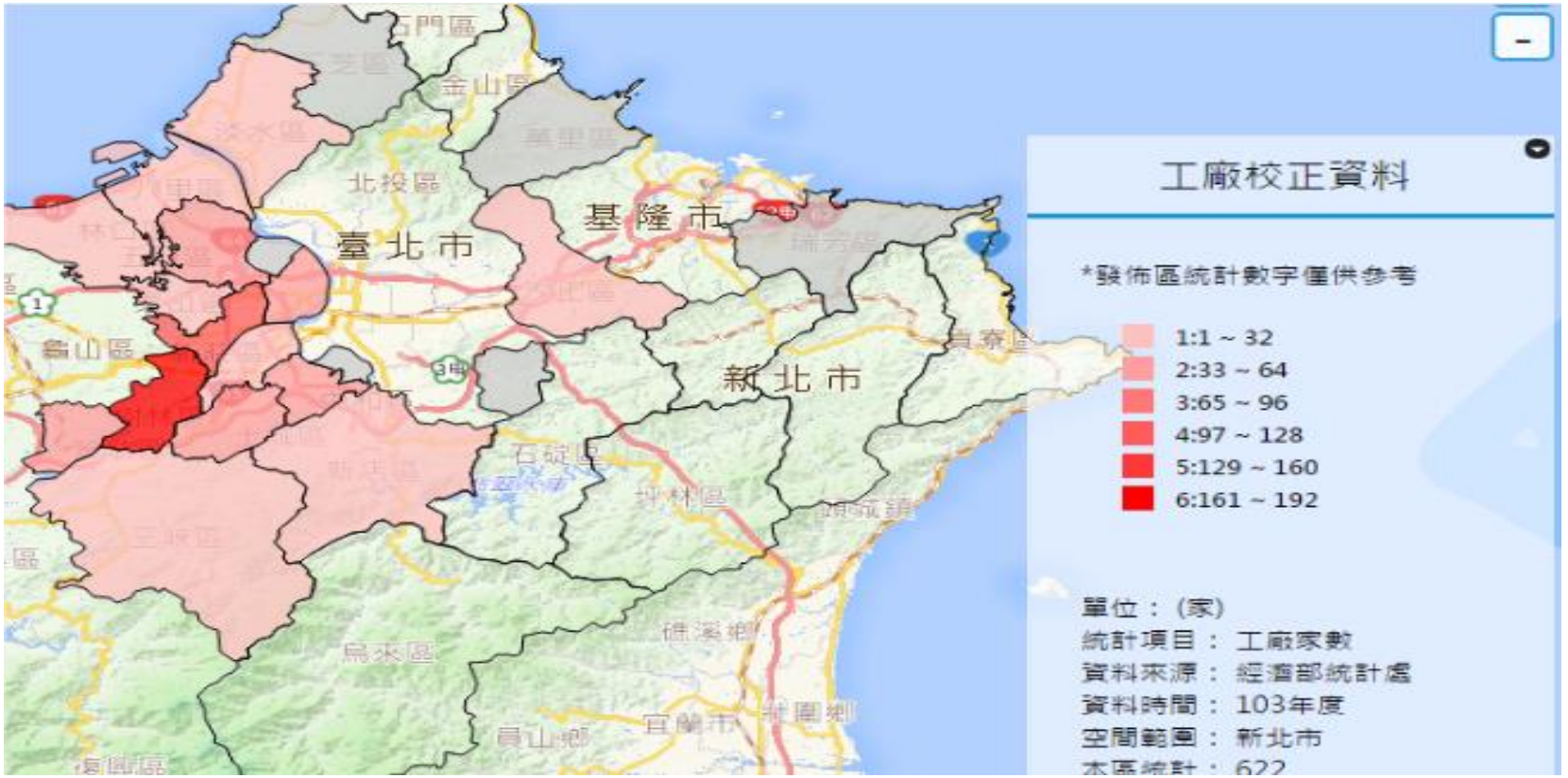
ex.手寫數字圖片之像素矩陣



認識手上的材料

地理資料：

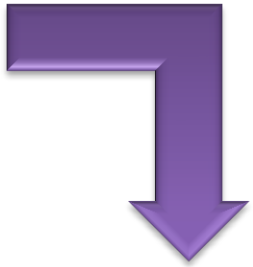
ex.地址、IP位址、地籍圖、經緯度



R-Ladies Taipei

衡量尺度 v.s. 資料格式

資料衡量尺度	變數形態	特性
名目資料(nomial)	質化	類別
順序資料(ordinal)	質化	優先順序
區間資料(interval)	量化	大小距離
比例資料(ratio)	量化	比值



資料格式	特徵
字元(character)	文字字串，用雙引號包起來。
數值(numeric)	雙倍精準度數值，就是double。
整數(integer)	沒有小數位的數值。
複數(complex)	虛數，實際上少用。
邏輯(logical)	TRUE /FALSE，縮寫成T/F。
**日期時間	POSIXct、POSIXt

R-Ladies Taipei

Summary

資料概述

R-Ladies Taipei

IRIS資料集

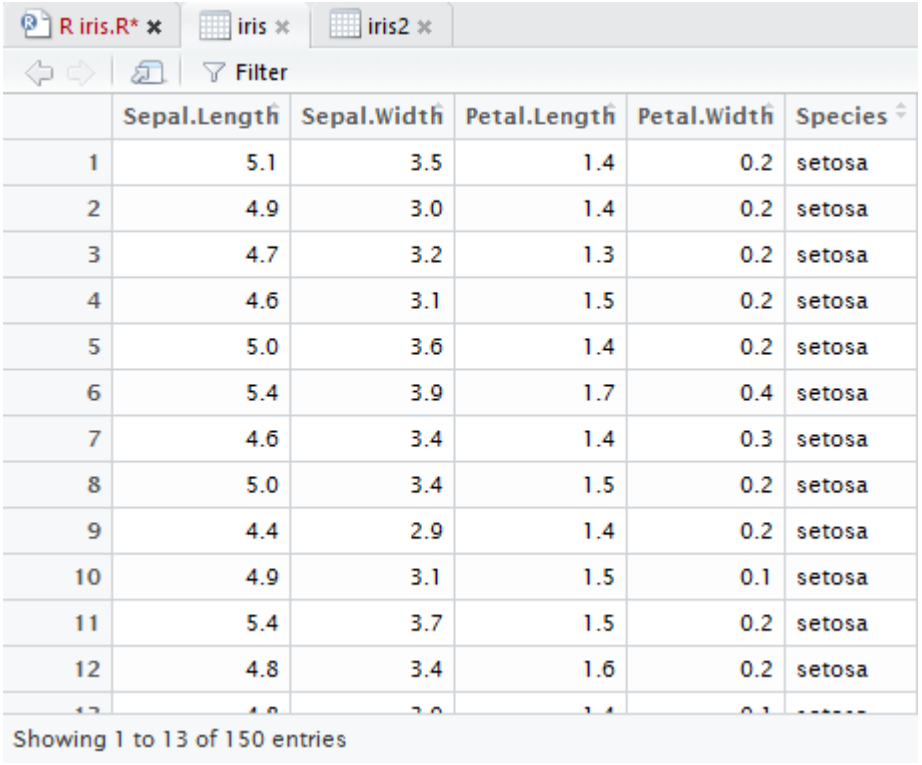
- 網址: http://archive.ics.uci.edu/ml/datasets/Iris
- 資料為150筆，共有五個欄位：

1. 花萼長度(Sepal Length)
2. 花萼寬度(Sepal Width)
3. 花瓣長度(Petal Length)
4. 花瓣寬度(Petal Width)
5. 類別(Species)：可分為Setosa，Versicolor和Virginica三個品種。

R-Ladies Taipei

IRIS資料集

- > View(iris)



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa

Showing 1 to 13 of 150 entries

看看IRIS資料集的頭尾

- > head(iris,5)

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
```

- > tail(iris,5)

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
146          6.7         3.0          5.2         2.3 virginica
147          6.3         2.5          5.0         1.9 virginica
148          6.5         3.0          5.2         2.0 virginica
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

R-Ladies Taipei

極值指令

- 極小值
  > min(dataframe $ column)
- 中位數
  > median(dataframe $ column)
- 平均數
  > mean(dataframe $ column)
- 極大值
  > max(dataframe $ column)
- 四分位距
  > quantile(dataframe $ column)
- 全距
  > range(dataframe $ column)

R-Ladies Taipei

極值指令

```
> min(iris$Sepal.Length)
[1] 4.3
> median(iris$Sepal.Length)
[1] 5.8
> mean(iris$Sepal.Length)
[1] 5.843333
> max(iris$Sepal.Length)
[1] 7.9
> quantile(iris$Sepal.Length)
  0%  25%  50%  75% 100%
 4.3  5.1  5.8  6.4  7.9
> range(iris$Sepal.Length)
[1] 4.3 7.9
```

Note : 注意變數種類

類別資料算平均，得到NA不是夢

```
> mean(iris$Species)
[1] NA
Warning message:
In mean.default(iris$Species) :
  argument is not numeric or logical: returning NA
```

Species
setosa
setosa
setosa
setosa
setosa
setosa
setosa
setosa
setosa
setosa
setosa
setosa

Parameters一個個打好累... ...

來個All in one吧!

R-Ladies Taipei

summary指令

- > summary(dataframe)

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

Note : summary指令沒有的...

- 標準差

  > sd(dataframe $ column)

- 眾數

  > n=table(iris$Sepal.Length)

  > as.numeric(names(n))[which.max(n)]

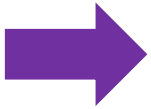
Note : summary指令沒有的...

```
> sd(iris$Sepal.Length)
[1] 0.8280661
> n=table(iris$Sepal.Length)
> as.numeric(names(n))[which.max(n)]
[1] 5
```

Note:遇到有雷的資料集

- >x[is.na(x)] <- 不違反假設的數字
  舉例此數為0

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	4	3	NA	3	7	6	6	10	6	5
2	9	8	9	5	10	NA	2	1	7	2
3	1	1	6	3	6	NA	1	4	1	6
4	NA	4	NA	7	10	2	NA	4	1	8
5	1	2	4	NA	2	6	2	6	7	4
6	NA	3	NA	NA	10	2	1	10	8	4
7	4	4	9	10	9	8	9	4	10	NA
8	5	8	3	2	1	4	5	9	4	7
9	3	9	10	1	9	9	10	5	3	3
10	4	2	2	5	NA	9	7	2	5	5



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	4	3	0	3	7	6	6	10	6	5
2	9	8	9	5	10	0	2	1	7	2
3	1	1	6	3	6	0	1	4	1	6
4	0	4	0	7	10	2	0	4	1	8
5	1	2	4	0	2	6	2	6	7	4
6	0	3	0	0	10	2	1	10	8	4
7	4	4	9	10	9	8	9	4	10	0
8	5	8	3	2	1	4	5	9	4	7
9	3	9	10	1	9	9	10	5	3	3
10	4	2	2	5	0	9	7	2	5	5

R-Ladies Taipei

The squeezed M curve

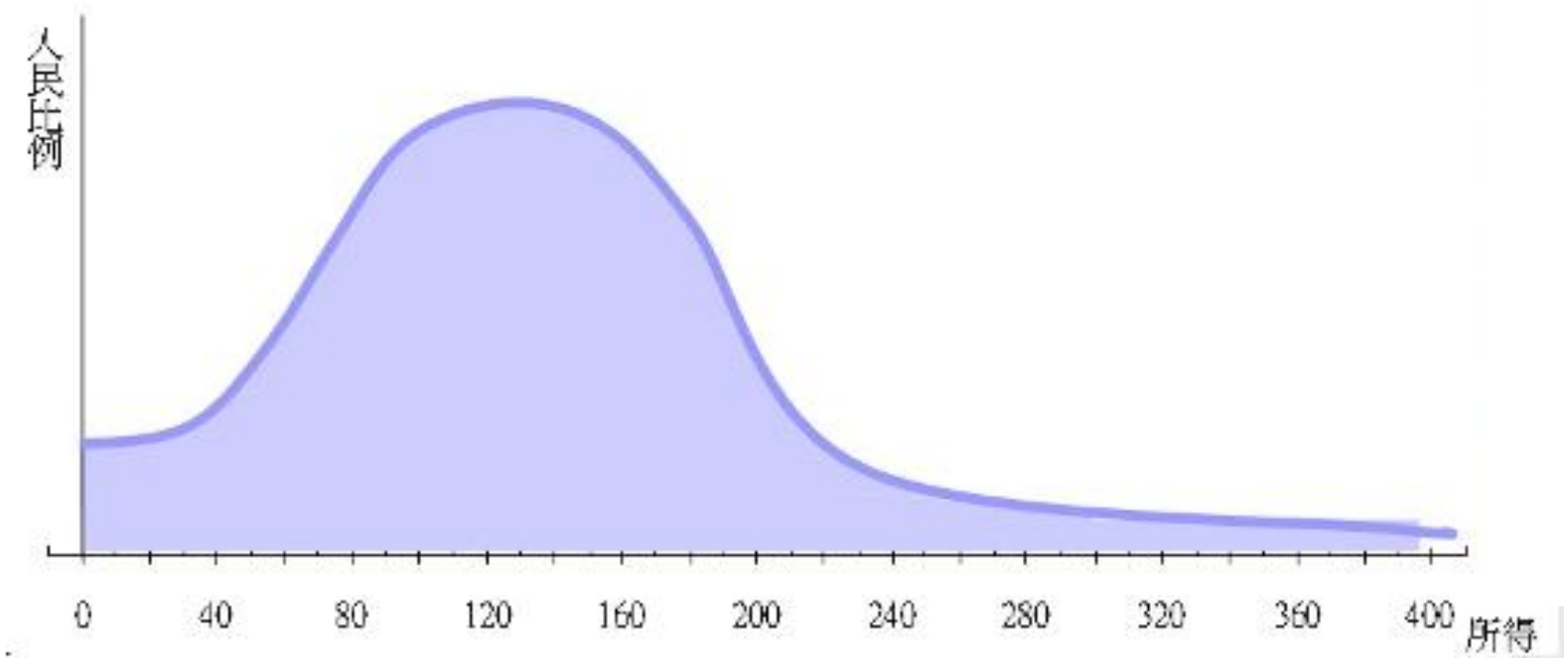
統計迷思?

在你洋洋灑灑寫Code之前... ...

請小心統計資料的**合理性**!

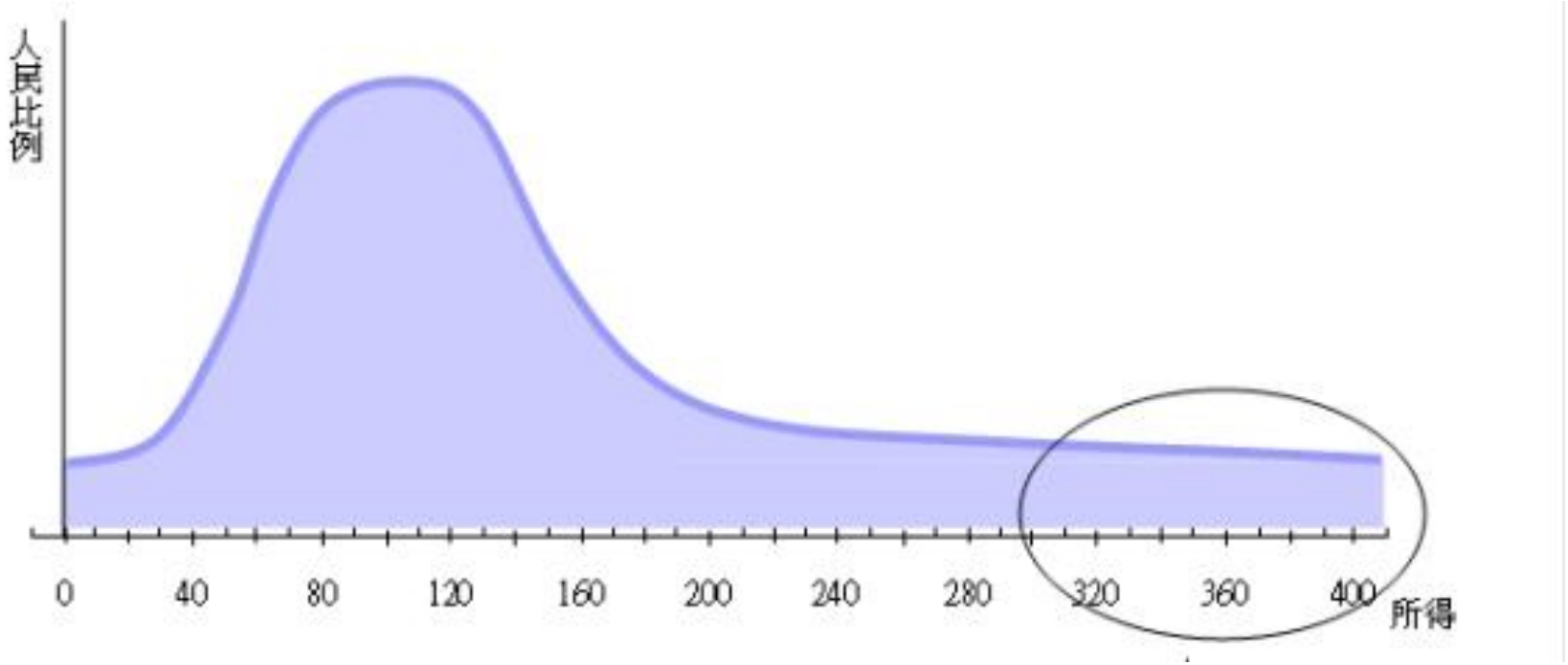
統計迷思:M型社會

- 一般的左高右長型分配



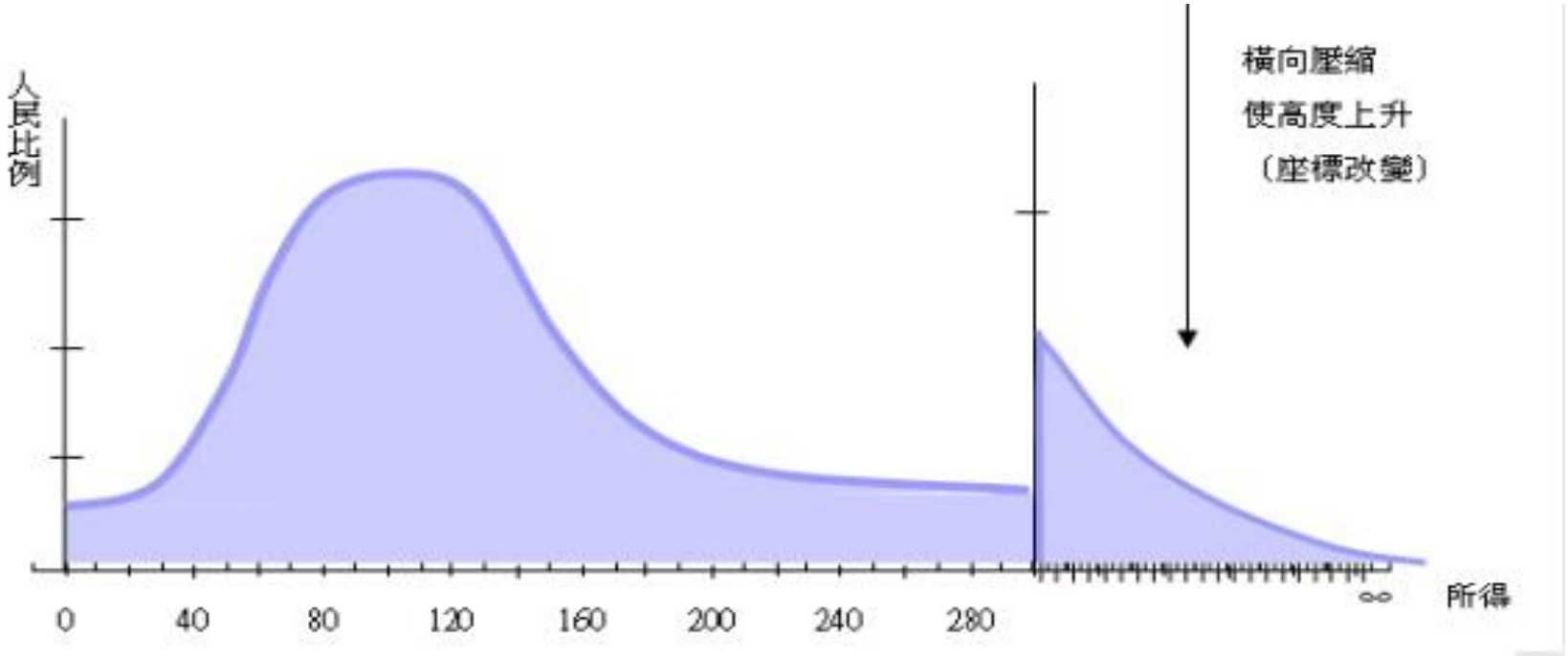
統計迷思:M型社會

- 全球化使高所得國家很多中低技術人員的工作被開發中國家取代，因此這類人員的所得降低，所得分配的高峰往左移，也就是相對較窮的人增加而所得分配惡化



統計迷思:M型社會

- 大談M型社會的大前研一乃是**把右端好多個所得層級加總在一起(橫軸壓縮成較短)**使同樣長度包含更廣的所得範圍，因此每單位長度的人數比例當然變高



統計迷思:M型社會

- 結論:

1.所得分配這樣惡化的結果，分配曲線看起來更像是L型而非M型。

2.事實上所得分配雖然確實惡化，但並不是變成兩極化而很少中間的M型

參考資料:

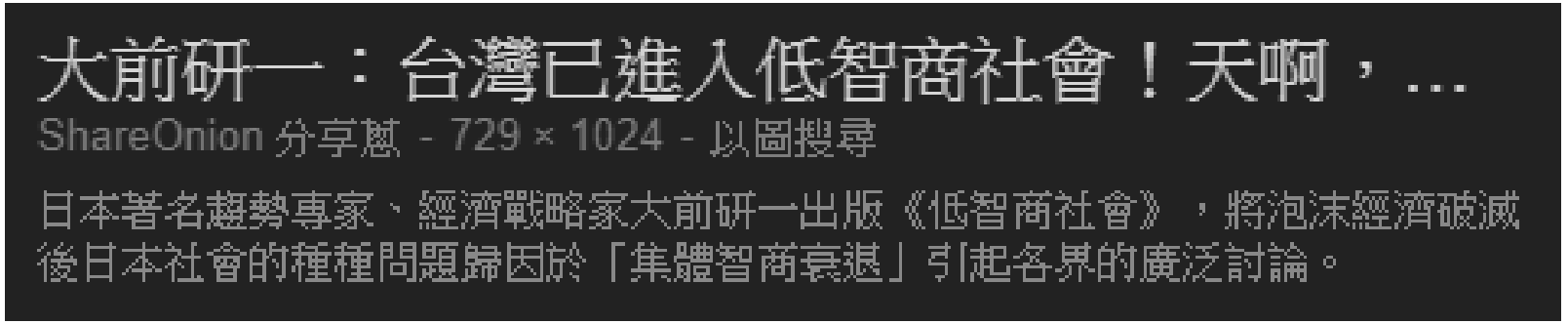
http://www.taiwanthinktank.org/chinese/page/9/26/1041/0

http://valerienliu.spaces.live.com/

R-Ladies Taipei

結果還Google到這本書

最狂的是這本書的註解



台灣在各個領域確實有很多能改進的空間，
從辨別資訊真偽著手應該也不錯

統計迷思:不存在的平均數

- 2016-02-02 18:47:45 中央社 中央社記者張建中　(部分新聞稿)

- 新竹2日電 晶圓代工廠台積電員工現金獎金與現金酬勞**總計達新台幣411億1378萬元，以台積電台灣員工4萬人計，平均1位員工將可拿到102.78**。

統計迷思:不存在的平均數

噓 Biboy: 又是平均...一堆2X的被這種新聞搞的在家裡兩面不是人
推 a58747912: 去頭去尾取中數最準啦
噓 DickMartin: 平均。

natek: 除了GG之外
還有哪些公司是基層的工程師可以拿到的平均值的呢？

R-Ladies Taipei

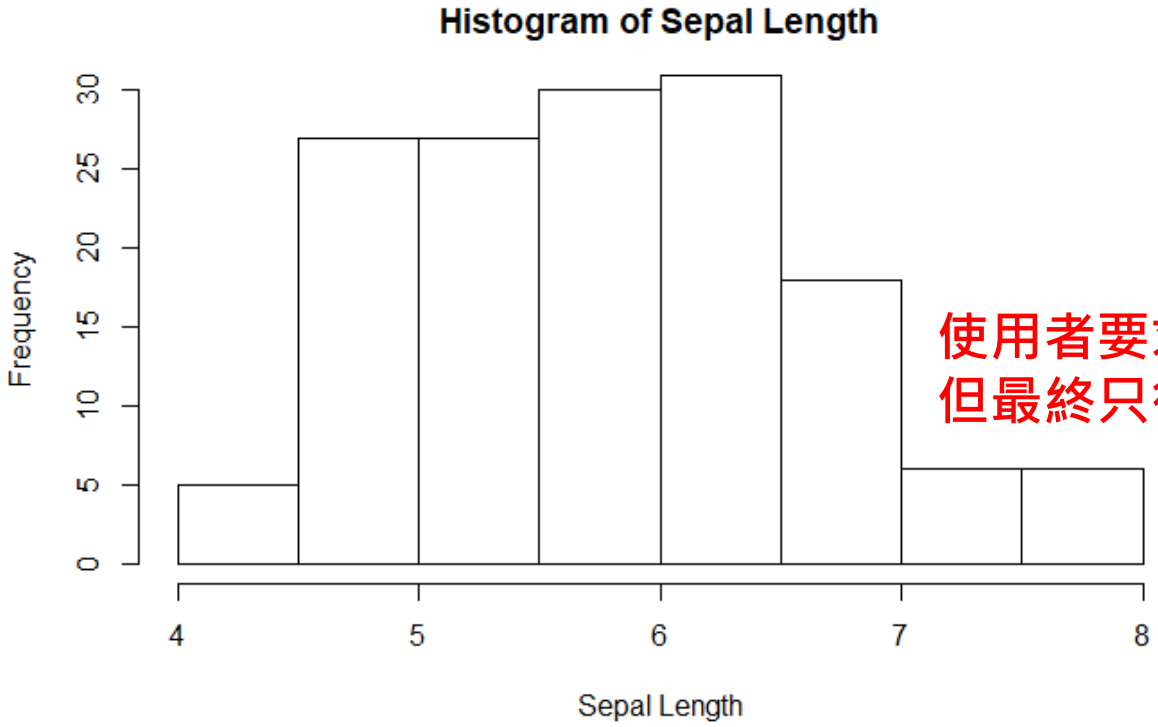
Statistic Plot

常見的統計圖表

直方圖Histogram

```
> h1<-hist(iris$Sepal.Length, breaks=10, xlab="Sepal Length", main="Histogram of Sepal Length")
```

breaks代表組數



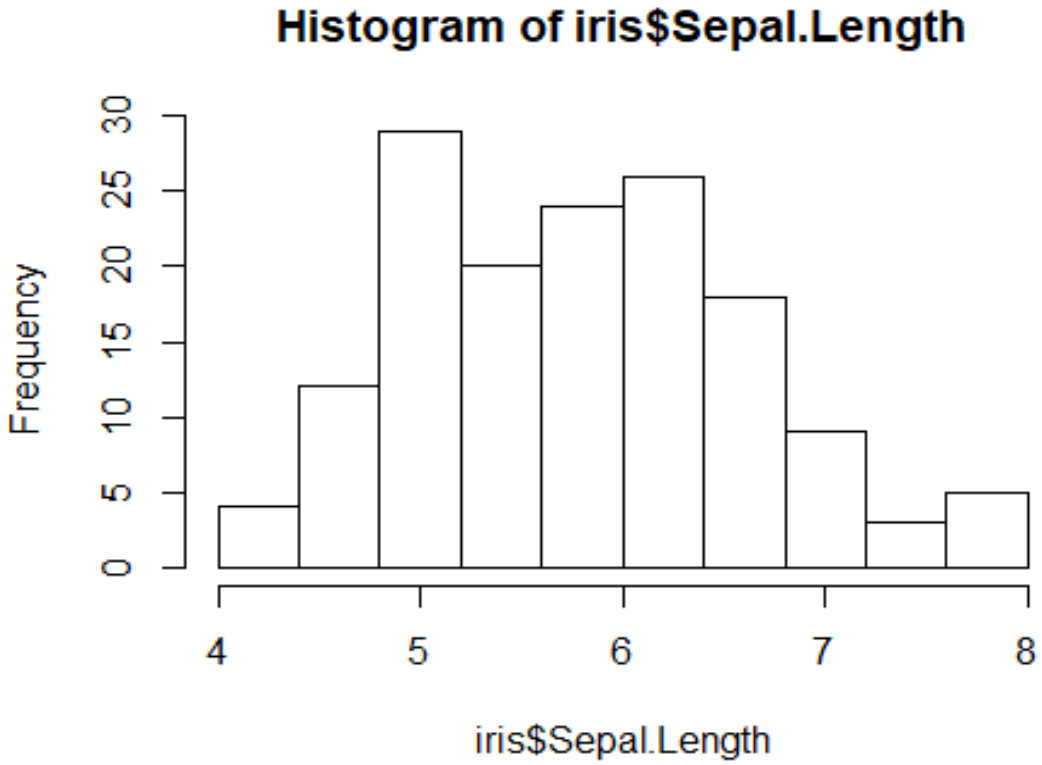
使用者要求切成10組，
但最終只得到8組？

```
> h1$breaks
## [1] 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
```

R-Ladies Taipei

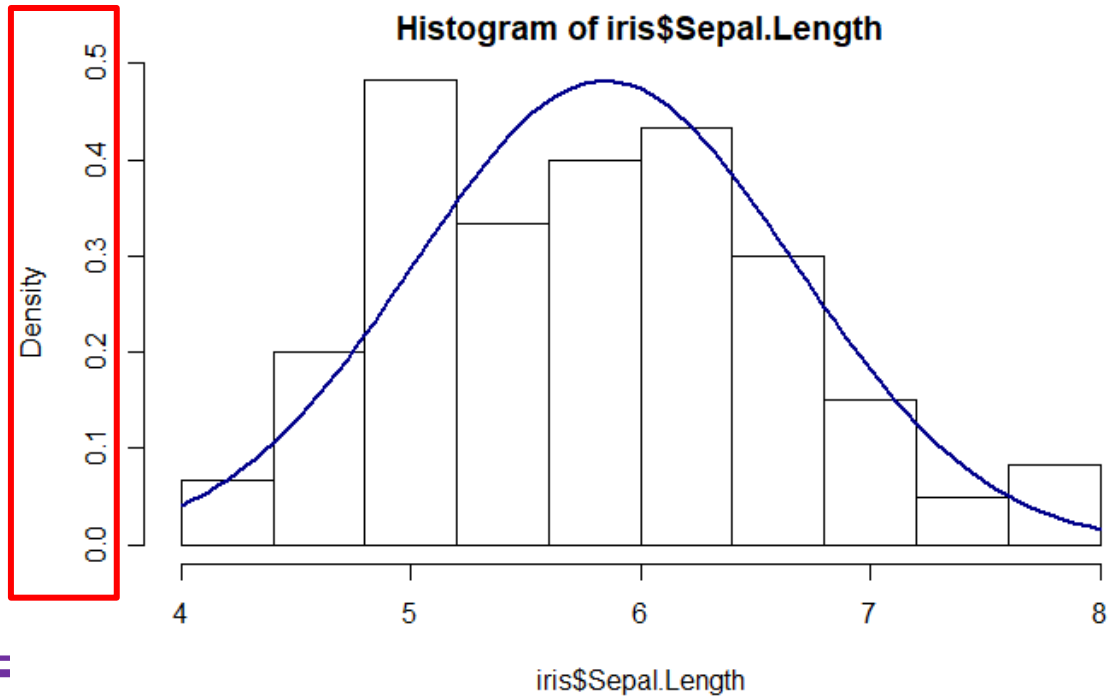
直方圖-組距數

```
> seq(4,8,by=0.4)
##  [1] 4.0 4.4 4.8 5.2 5.6 6.0 6.4 6.8 7.2 7.6 8.0
>h2<-hist(iris$Sepal.Length, breaks=seq(4,8,by=0.4))
```



直方圖-常態分配

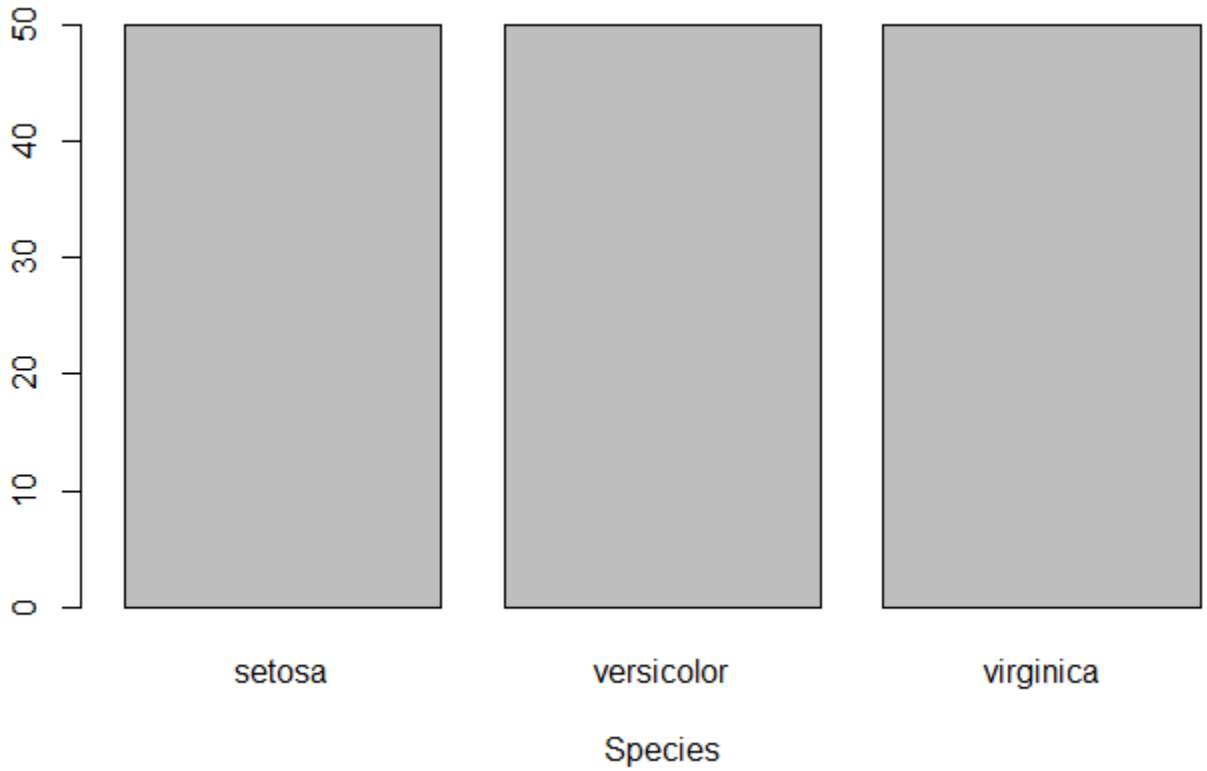
```
> h3<-hist(iris$Sepal.Length, breaks=seq(4,8,by=0.4), prob=TRUE)
> m = mean(iris$Sepal.Length)
> std = sqrt(var(iris$Sepal.Length))
> curve(dnorm(x, mean=m, sd=std), col="darkblue", lwd=2, add=TRUE)
```



R-Ladies Taipei

長條圖Bar Chart

```
> plot(iris$Species, xlab="Species")
```

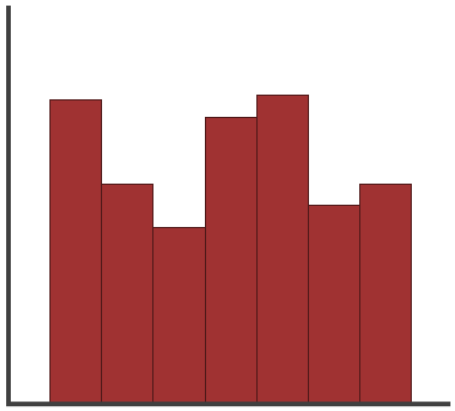


R-Ladies Taipei

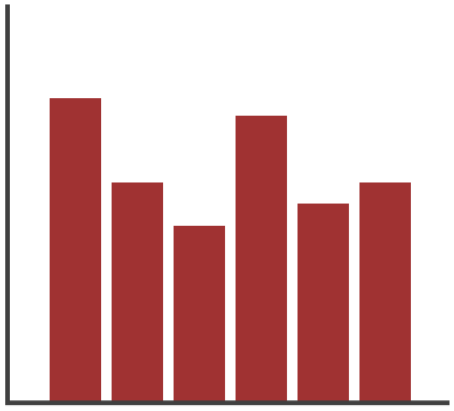
直方圖 vs 長條圖

- 都用來表示次數分配
- 長條圖的每一根柱子之間是分開的，而直方圖則是相連
- 長條圖的X軸用來描述離散變數，直方圖則描述連續變數
  - nominal variable或ordinal variable適用長條圖
  - Interval variable或ratio variable則是用直方圖

直方圖
Histogram



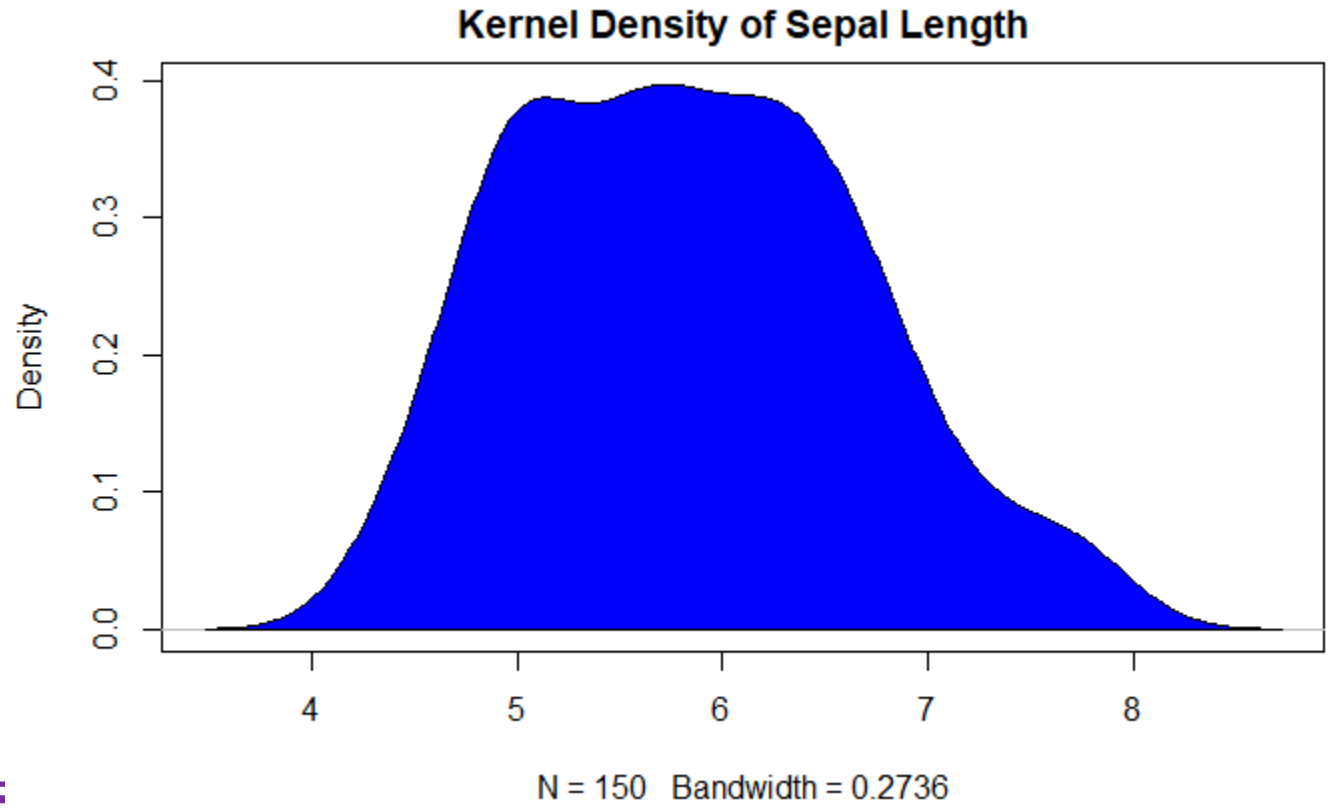
長條圖
Bar chart



R-Ladies Taipei

密度圖Density Plot

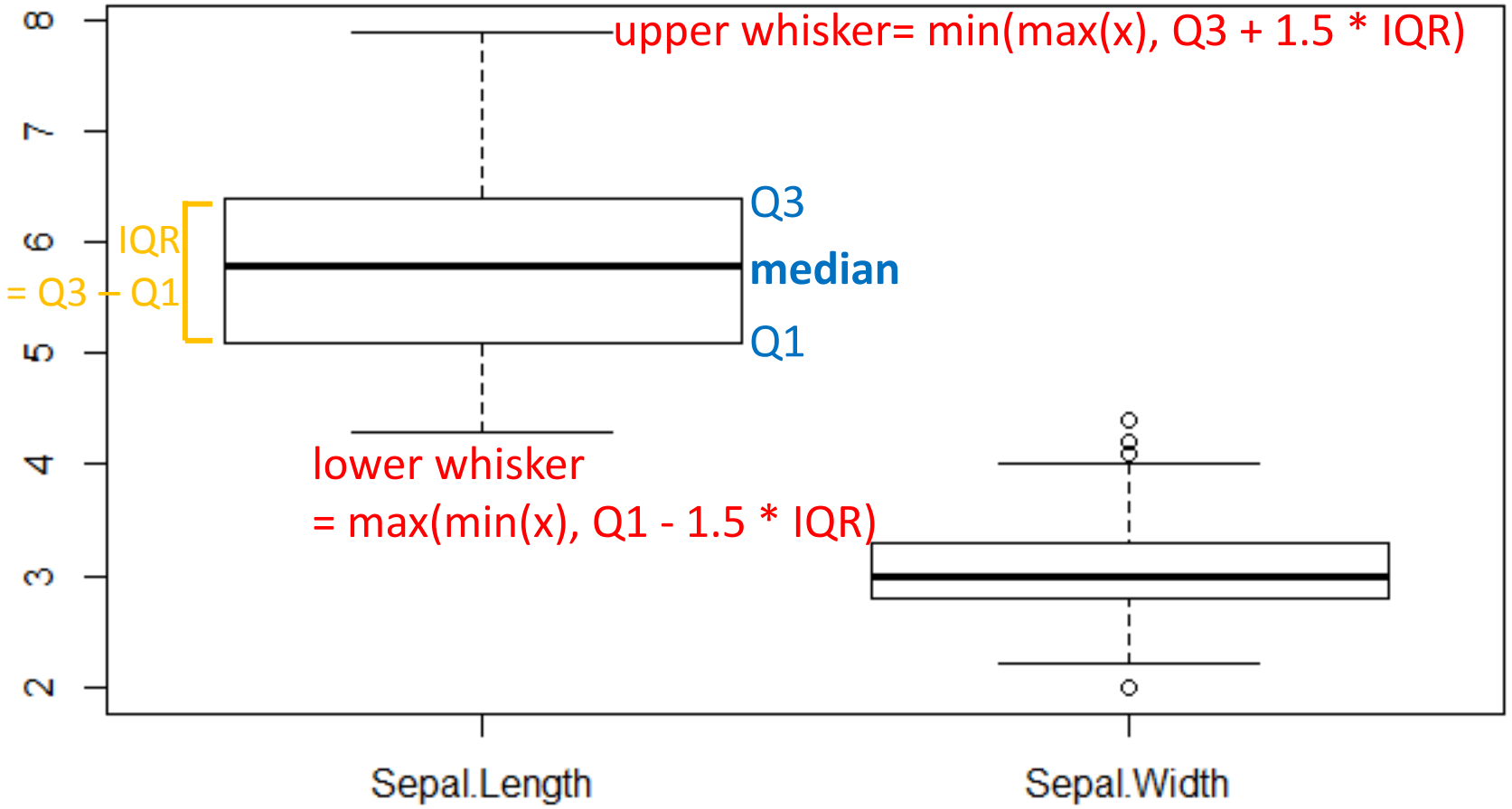
```
> plot(density(iris$Sepal.Length), main="Kernel Density of Sepal Length")
> polygon(density(iris$Sepal.Length), col="blue")
```



R-Ladies Taipei

盒鬚圖Box Plot

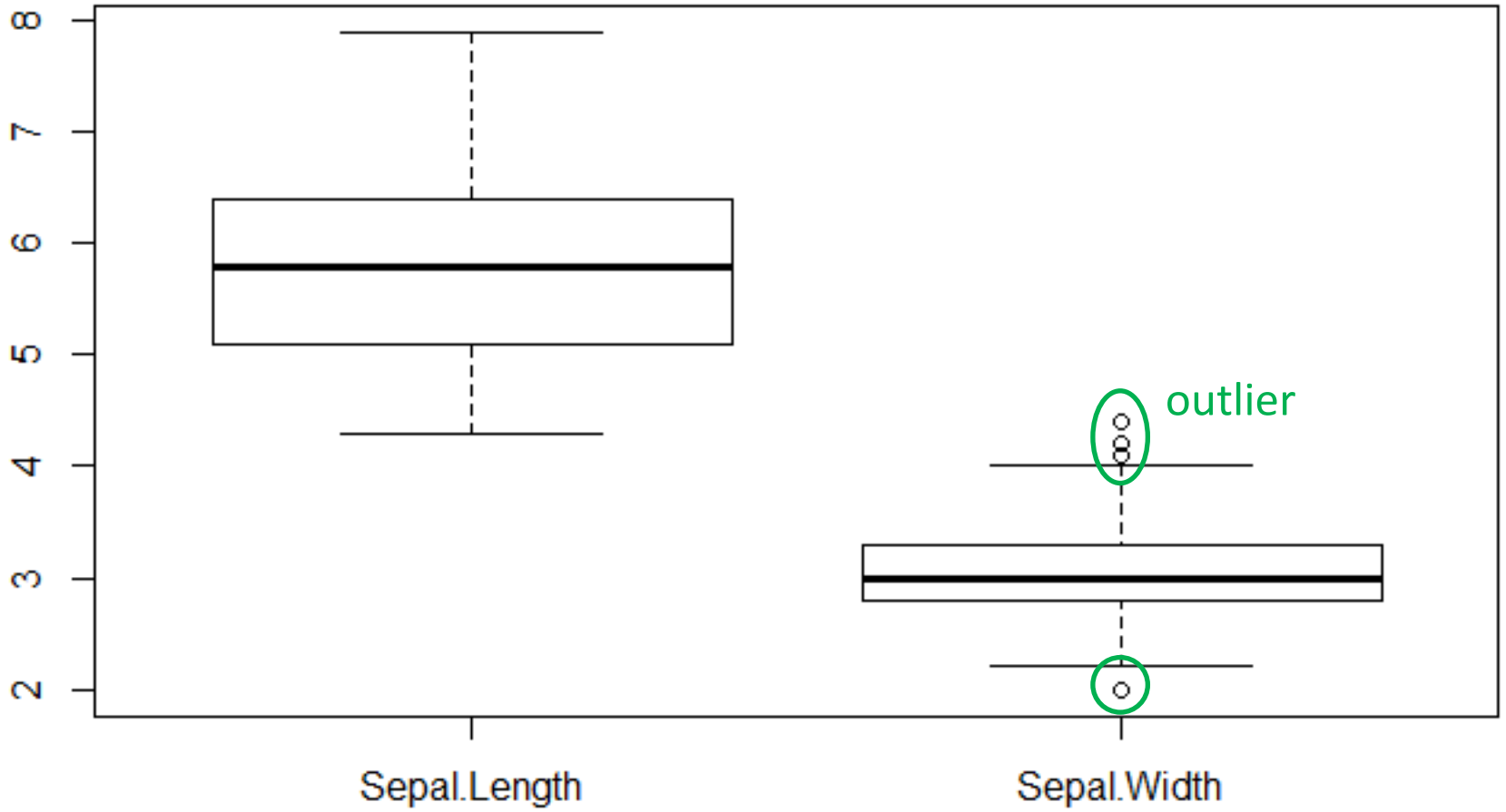
```
> boxplot(iris[,1:2])
```



R-Ladies Taipei

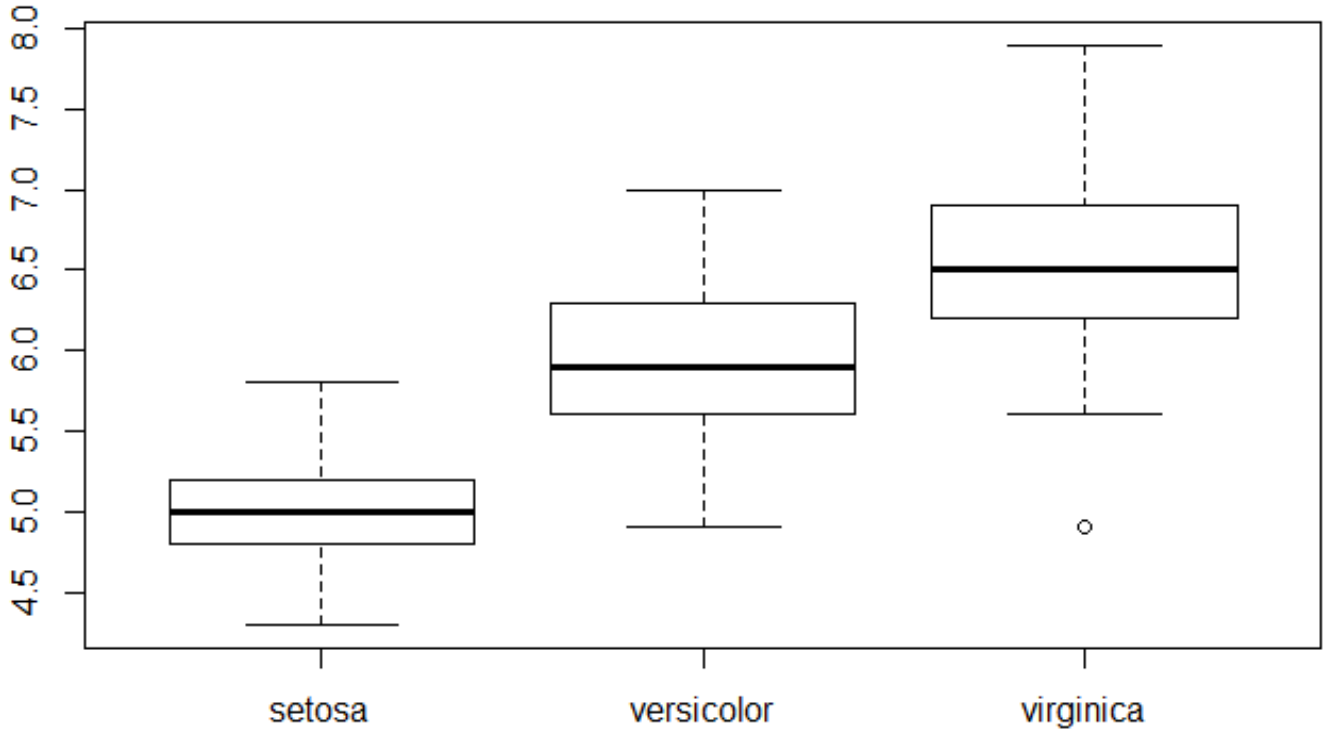
盒鬚圖Box Plot

```
> boxplot(iris[,1:2])
```



盒鬚圖-兩個變數

```
> boxplot(iris$Sepal.Length~iris$Species)
```



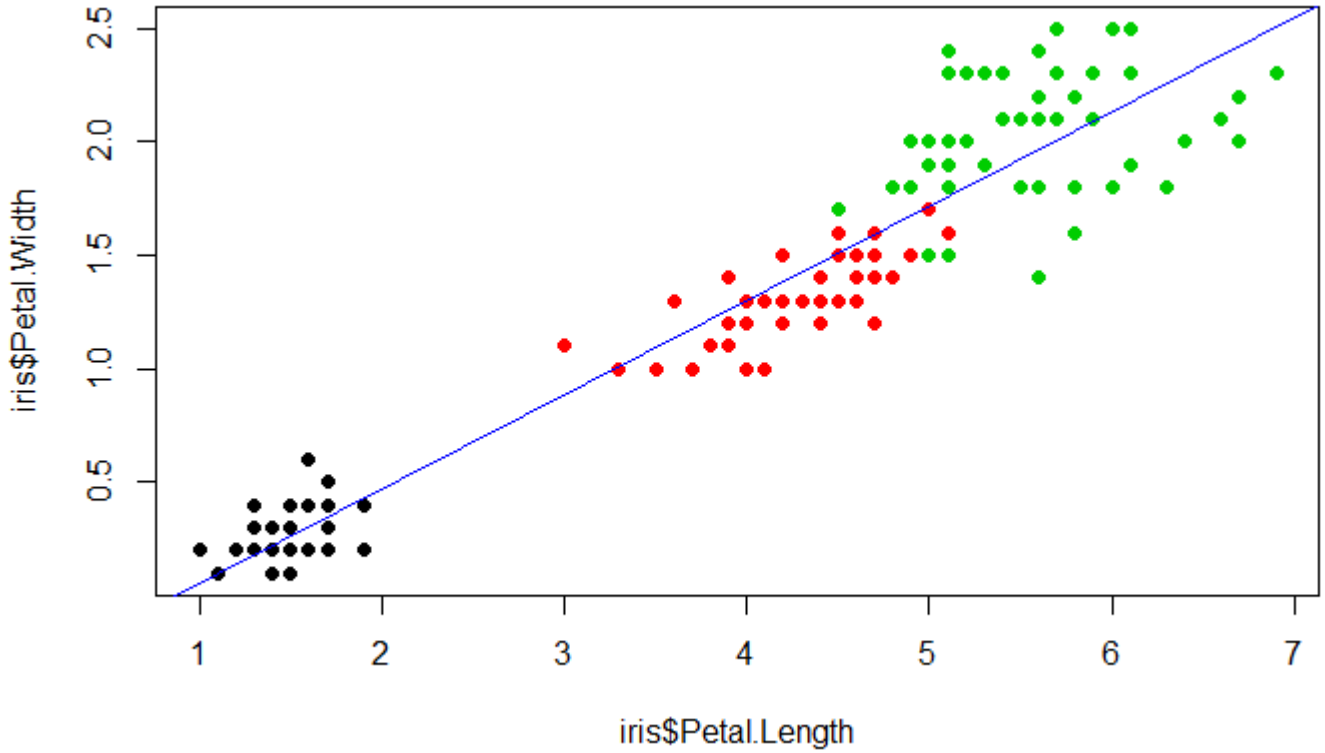
R-Ladies Taipei

Scatter Plot

散布圖的秘密

散布圖Scatter Plot

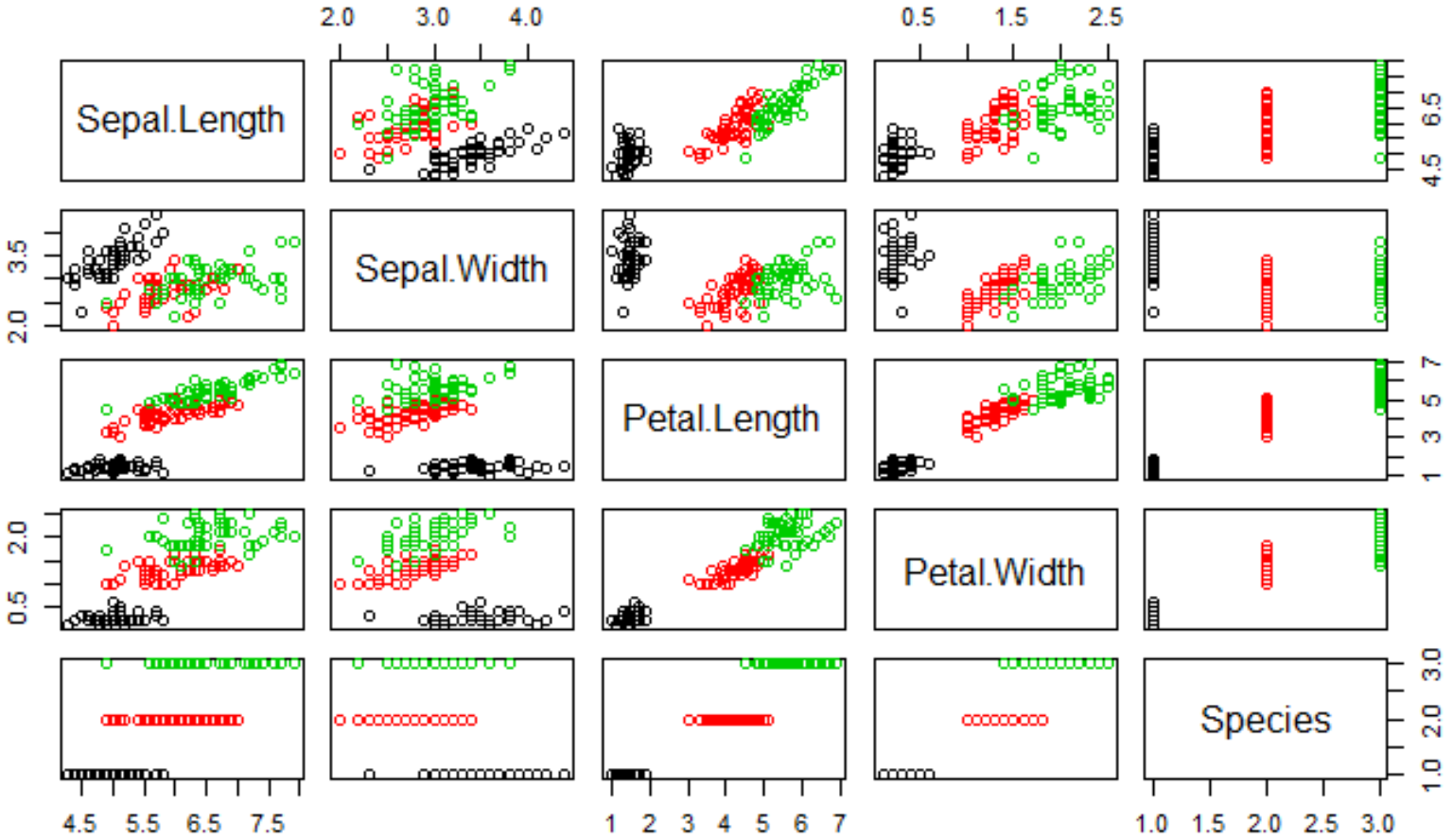
```
> plot(iris$Petal.Length, iris$Petal.Width, pch=19, col=iris$Species)
> abline(lm(iris$Petal.Width~iris$Petal.Length), col="blue")
```



R-Ladies Taipei

散布矩陣圖

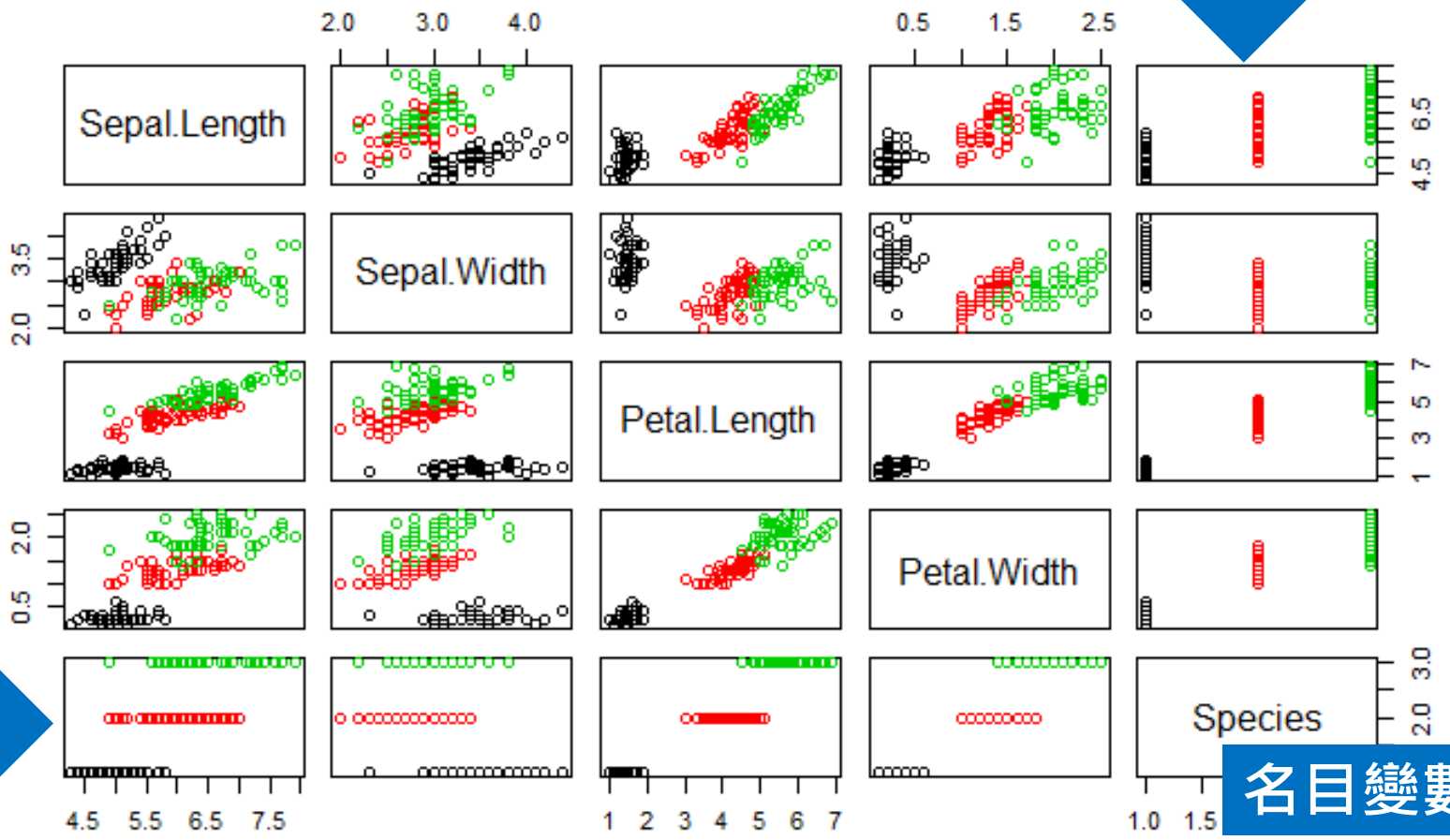
```
> with(iris, plot(iris, col = Species))
```



R-Ladies Taipei

散布矩陣圖

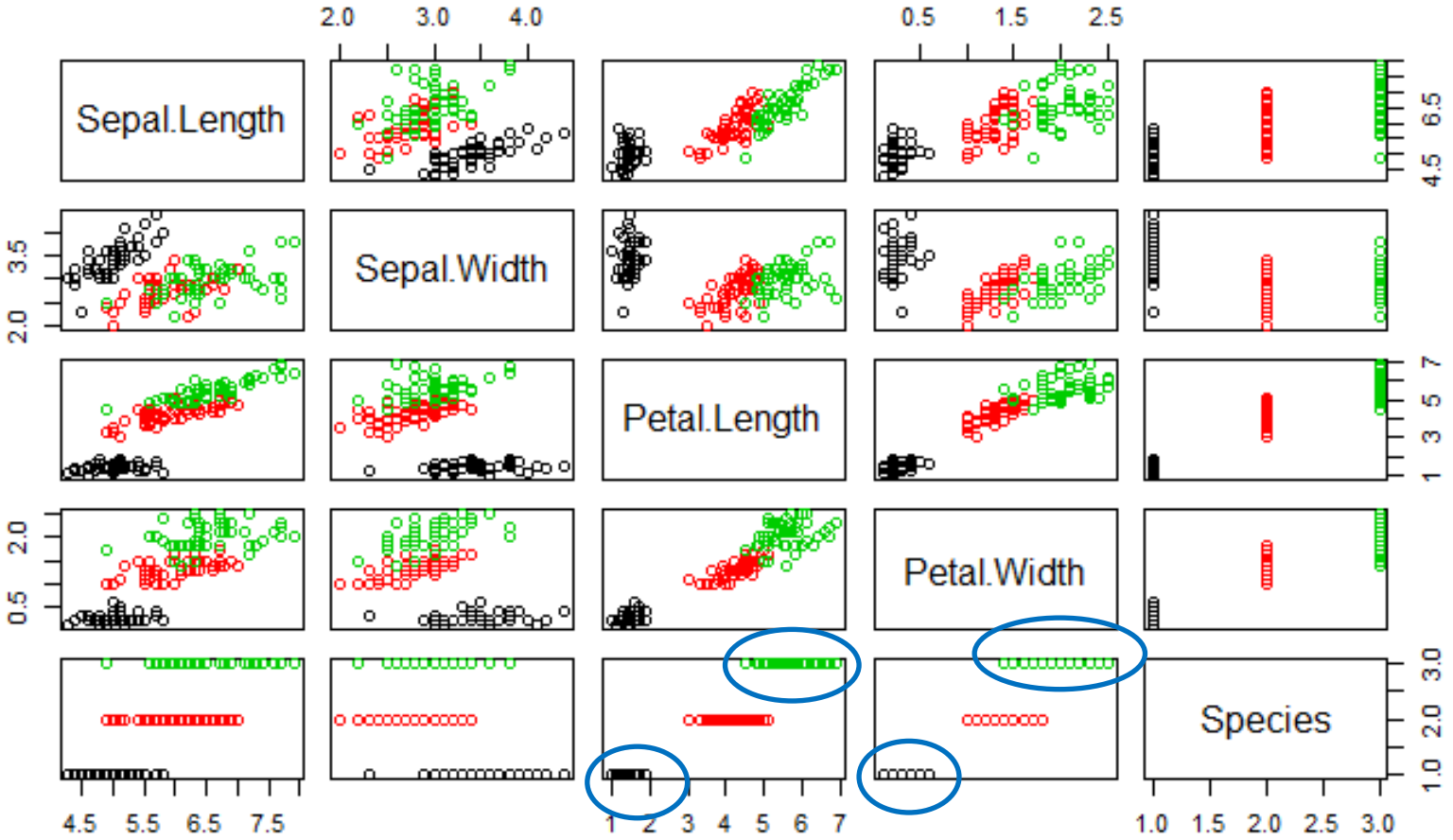
```
> with(iris, plot(iris, col = Species))
```



R-Ladies Taipei

散布矩陣圖

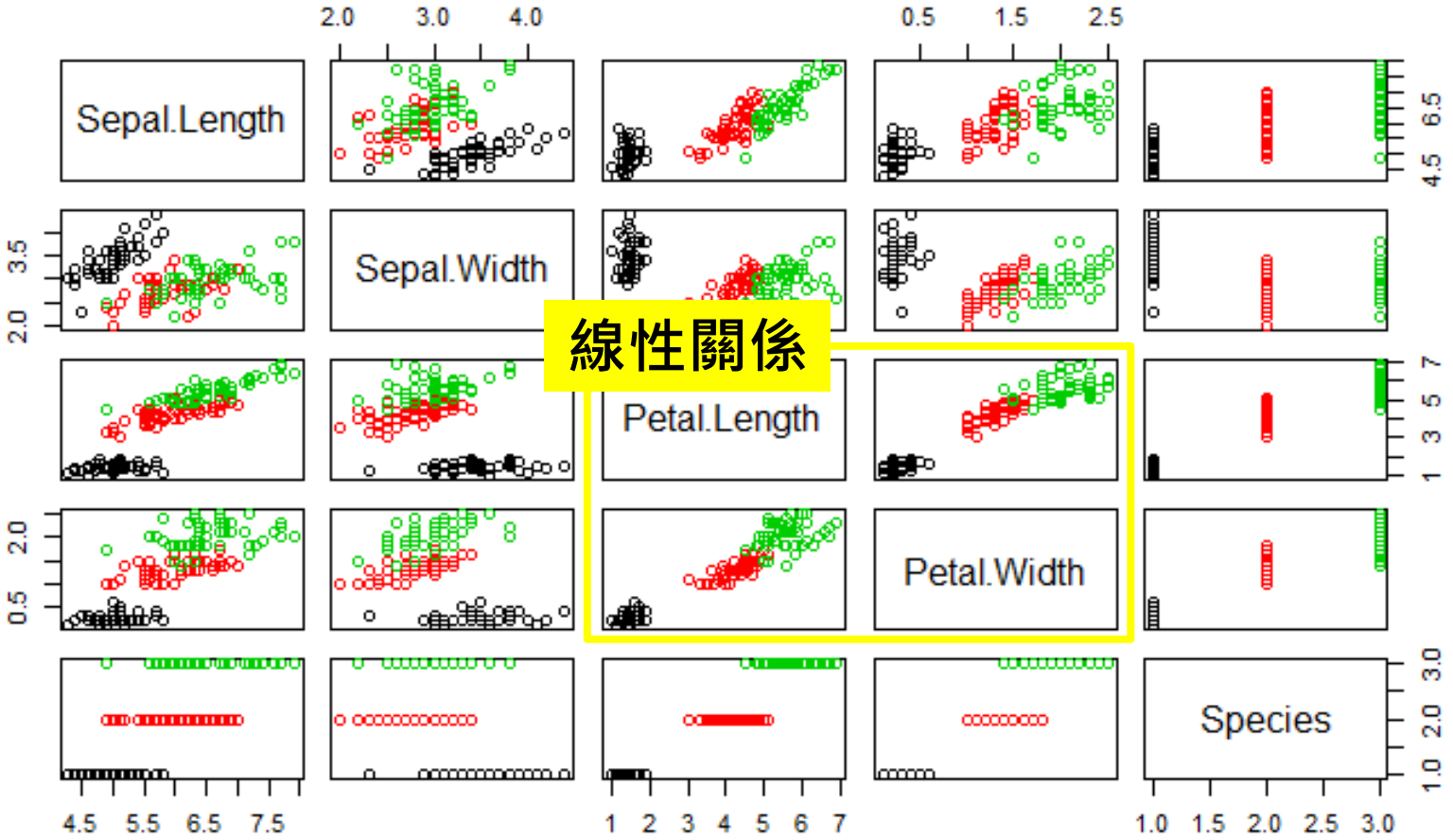
```
> with(iris, plot(iris, col = Species))
```



R-Ladies Taipei

散布矩陣圖

```
> with(iris, plot(iris, col = Species))
```



R-Ladies Taipei

散布矩陣圖-用意

- 觀察變數間的相關性
- 變數的篩選：
  - 相關係數(Correlation Coefficient)

  *補充：
  - 主成分分析(Principal Component Analysis)
  - Step-wise forward selection
  - Step-wise backward selection

R-Ladies Taipei

繪圖套件

- 以上僅介紹使用內建的函數繪圖
- 常用繪圖套件：ggplot2
  - ggplot2.tidyverse.org/reference/

R-Ladies Taipei

延伸閱讀

- M型社會討論來自：台灣智庫
- 陳鍾誠：用十分鐘瞭解 機率、統計、還有R軟體
- 台灣資料科學年會：給工程師的統計學及資料分析 123、手把手教你R語言資料分析實務
- **台大計算機中心：R統計分析與資料探勘入門—以鳶尾花資料集為例**
- R-bloggers : Data Science for Doctors – Part 2 : Descriptive Statistics