

R 語言基礎課程 (二)

丘祐瑋
David Chiu

環境資訊頁面

- 所有課程補充資料、投影片皆位於
 - ▣ https://github.com/ywchiu/cdc_course

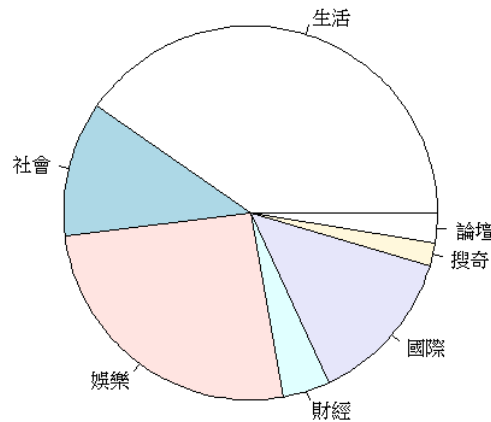


資料探索 (Data Exploration)

敘述性統計

■ 敘述性統計

- 有系統的歸納資料，瞭解資料的輪廓
- 對數據樣本做敘述性陳述，例如：平均數、標準差、計次頻率、百分比
- 對數據資料的圖像化處理，將資料摘要變為圖表



敘述性統計分析

- 多數資料分析，80% 在於如何加總與平均

- 銷售份額
- 客戶數量
- 業績成長量

- 用SQL做敘述性統計

- `select * from tb1 where col1 >= 100 limit 3`

- R有類似的工具嗎？

- `plyr`
- `reshape2`
- `dplyr`



如何操作資料

- 關於操作資料，你需要
 - 可以分割資料(Split)
 - 可以轉換資料(Transformation)
 - 可以聚合資料(Aggregation)
 - 可以探索資料(Exploration)
- 需要如同SQL的語法操作

使用dplyr 做資料分析

為什麼要使用dplyr

- 提供操作資料的基本語法
 - ▣ filter, select, arrange, mutate, summarise, group_by
- 提供資料合併功能(JOIN)
 - ▣ Inner join, left join
- 可以操作資料表(data table) 或資料庫 (Database)的資料

安裝與使用dplyr

■ 安裝dplyr

- `install.packages("dplyr")`

■ 使用dplyr

- `library(dplyr)`

■ 觀看說明頁

- `help(package='dplyr')`

過濾資料

#原先R 提供的過濾功能

```
Dengue[Dengue$感染國家 == "中華民國",]
```

#dplyr 的過濾功能

```
filter(Dengue, 感染國家 == "中華民國")
```

可以使用 AND, OR 與 IN 來過濾資料

#找出病例數超過100的男性案例

```
filter(Dengue, 性別 == "男" & 病例數 > 100)
```

#找出病例數超過100或男性案例

```
filter(Dengue, 性別 == "男" | 病例數 > 100)
```

#找出台北或台南的病例

```
filter(Dengue, 居住縣市 %in% c("台南市", "台北市"))
```


選擇欄位

#原先R 提供的欄位選取

```
Dengue[, c("居住縣市", "病例數")]
```

#dplyr 的欄位選取

```
select(Dengue, 居住縣市, 病例數)
```

但如果我想同時選擇欄位又過濾資料呢？

■ 鏈接(Chaining)

- %>% (Then)

- 來自 magrittr

%>%
magrittr

Ceci n'est pas un pipe.

■ 使用Then (%>%)

Dengue %>%

select(居住縣市, 病例數) %>%

filter(居住縣市 == "台北市")

資料做排序

- 使用Arrange 可以將資料做排序

Dengue %>%

select(居住縣市, 病例數) %>%

filter(居住縣市 == "台北市") %>%

arrange(病例數)

- 由大到小排序 (desc)

Dengue %>%

select(居住縣市, 病例數) %>%

filter(居住縣市 == "台北市") %>%

arrange(desc(病例數))

新增欄位 (mutate)

#計算總和

```
freqsum <- Dengue %>%  
  select(病例數) %>%  
  sum()
```

#使用mutate 新增欄位

```
Dengue %>%  
  select(病例數) %>%  
  mutate(portion= 病例數/freqsum)
```

#儲存新欄位

```
Dengue <- Dengue %>% mutate(portion= 病例數/freqsum)
```

分組計算 (group_by, summarise)

■ 分組計算函式

- group_by: 分組依據
- summarise: 依組別計算結果

■ 統計各性別的總和

Dengue %>%

group_by(性別) %>%

summarise(instance_sum = sum(病例數, na.rm=TRUE))

統計多個欄位

#使用summarise_each 統計portion 與 病例數 的總和

Dengue %>%

group_by(性別) %>%

summarise_each(funs(sum), 病例數, portion)

針對多個欄位做統計

- 可針對不同資料同時做統計

```
Dengue %>%
```

```
  group_by(居住縣市) %>%
```

```
  summarise_each(funs(min(., na.rm=TRUE), max(.,  
na.rm=TRUE))), 病例數)
```

資料計數

■ 一般計數

Dengue %>%

```
select(居住縣市) %>%
```

```
summarise_each(funs(n()))
```

■ 不重複計數

Dengue %>%

```
select(居住縣市) %>%
```

```
summarise_each(funs(n_distinct(居住縣市)))
```



資料視覺化 (Data Visualization)

人是視覺性的動物



70%



30%

這裡面有多少個 9?

3 3 0 3 0 1 8 7 6 8 2 1 4 0 3 8 3 7 7 2 0 5 2 3 2 7 0 2 0
7 1 4 6 0 2 1 3 2 7 6 0 2 5 6 3 2 5 7 6 3 3 0 2 0 3 0 7 2
8 7 5 7 2 8 3 8 7 7 8 2 0 7 7 5 2 3 1 1 5 6 3 8 4 7 8 2 0
0 5 0 5 1 6 1 7 5 6 8 0 4 4 6 7 4 7 1 4 0 0 8 4 4 3 0 3 2
2 4 3 1 3 5 4 9 5 0 7 6 0 7 4 3 1 8 2 7 3 4 6 0 2 4 8 2 3
8 6 2 2 6 5 4 6 7 0 7 6 0 0 3 9 0 2 4 7 1 7 2 3 3 5 8 7 0
0 8 4 5 1 3 1 7 6 4 5 4 1 2 4 5 3 3 5 4 9 6 7 7 6 3 4 2 5
4 7 7 0 2 2 0 1 1 7 7 7 0 2 6 6 4 7 5 8 6 1 4 3 7 8 5 4 6
4 3 6 6 4 6 6 2 8 4 8 5 3 7 8 8 1 3 8 5 4 5 7 4 0 3 2 8 4
5 5 0 3 5 3 5 3 8 3 2 3 8 2 3 1 6 2 7 2 4 6 3 6 4 4 3 2 5
4 4 0 2 1 7 2 4 4 7 4 1 9 2 4 5 2 5 0 4 0 0 5 3 6 3 3 6 7
7 4 6 6 8 7 5 7 9 2 0 2 8 8 8 8 3 2 4 2 6 4 0 4 6 3 7 2 1
0 1 7 1 5 9 1 4 2 8 7 3 7 1 4 5 1 8 7 8 0 5 1 7 0 5 8 8 1
2 8 5 2 1 2 8 7 7 6 2 5 6 2 6 4 1 5 1 6 1 2 1 1 0 5 6 4 0
2 1 1 7 7 2 0 0 1 8 7 0 2 9 0 2 8 5 7 8 4 6 0 6 5 0 7 1 2
0 5 2 4 1 5 3 3 1 5 5 1 4 0 1 6 4 3 3 9 8 8 3 4 6 8 4 8 6
7 3 7 5 2 4 0 2 7 6 3 8 5 5 4 5 8 8 7 5 5 6 5 6 7 9 7 7 4
0 3 2 8 1 4 4 6 0 8 2 3 0 1 3 4 6 2 0 5 7 7 3 6 1 8 7 3 5
4 4 8 3 3 3 5 0 1 0 3 8 6 3 2 0 5 0 6 1 3 3 4 3 6 1 5 8 6
1 0 2 2 7 6 3 3 0 8 8 0 3 1 8 8 1 2 1 7 5 2 9 3 5 8 3 2 5

視覺化的重要性

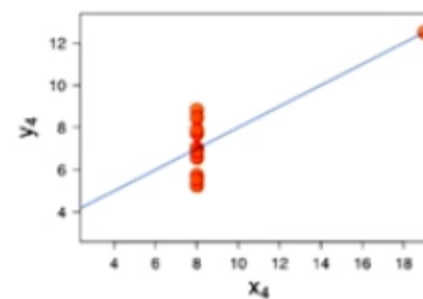
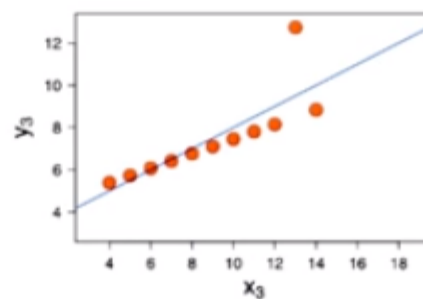
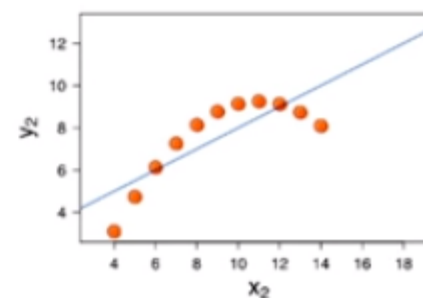
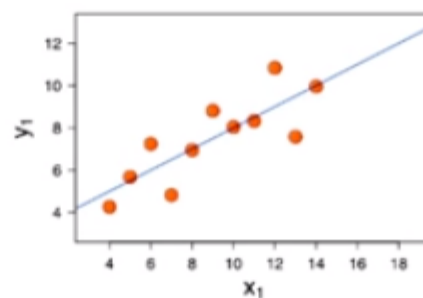
3	3	0	3	0	1	8	7	6	8	2	1	4	0	3	8	3	7	7	2	0	5	2	3	2	7	0	2	0
7	1	4	6	0	2	1	3	2	7	6	0	2	5	6	3	2	5	7	6	3	3	0	2	0	3	0	7	2
8	7	5	7	2	8	3	8	7	7	8	2	0	7	7	5	2	3	1	1	5	6	3	8	4	7	8	2	0
0	5	0	5	1	6	1	7	5	6	8	0	4	4	6	7	4	7	1	4	0	0	8	4	4	3	0	3	2
2	4	3	1	3	5	4	9	5	0	7	6	0	7	4	3	1	8	2	7	3	4	6	0	2	4	8	2	3
8	6	2	2	6	5	4	6	7	0	7	6	0	0	3	9	0	2	4	7	1	7	2	3	3	5	8	7	0
0	8	4	5	1	3	1	7	6	4	5	4	1	2	4	5	3	3	5	4	9	6	7	7	6	3	4	2	5
4	7	7	0	2	2	0	1	1	7	7	7	0	2	6	6	4	7	5	8	6	1	4	3	7	8	5	4	6
4	3	6	6	4	6	6	2	8	4	8	5	3	7	8	8	1	3	8	5	4	5	7	4	0	3	2	8	4
5	5	0	3	5	3	5	3	8	3	2	3	8	2	3	1	6	2	7	2	4	6	3	6	4	4	3	2	5
4	4	0	2	1	7	2	4	4	7	4	1	9	2	4	5	2	5	0	4	0	0	5	3	6	3	3	6	7
7	4	6	6	8	7	5	7	9	2	0	2	8	8	8	8	3	2	4	2	6	4	0	4	6	3	7	2	1
0	1	7	1	5	9	1	4	2	8	7	3	7	1	4	5	1	8	7	8	0	5	1	7	0	5	8	8	1
2	8	5	2	1	2	8	7	7	6	2	5	6	2	6	4	1	5	1	6	1	2	1	1	0	5	6	4	0
2	1	1	7	7	2	0	0	1	8	7	0	2	9	0	2	8	5	7	8	4	6	0	6	5	0	7	1	2
0	5	2	4	1	5	3	3	1	5	5	1	4	0	1	6	4	3	3	9	8	8	3	4	6	8	4	8	6
7	3	7	5	2	4	0	2	7	6	3	8	5	5	4	5	8	8	7	5	5	6	5	6	7	9	7	7	4
0	3	2	8	1	4	4	6	0	8	2	3	0	1	3	4	6	2	0	5	7	7	3	6	1	8	7	3	5
4	4	8	3	3	3	5	0	1	0	3	8	6	3	2	0	5	0	6	1	3	3	4	3	6	1	5	8	6
1	0	2	2	7	6	3	3	0	8	8	0	3	1	8	8	1	2	1	7	5	2	9	3	5	8	3	2	5

告訴我這張表的意含

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

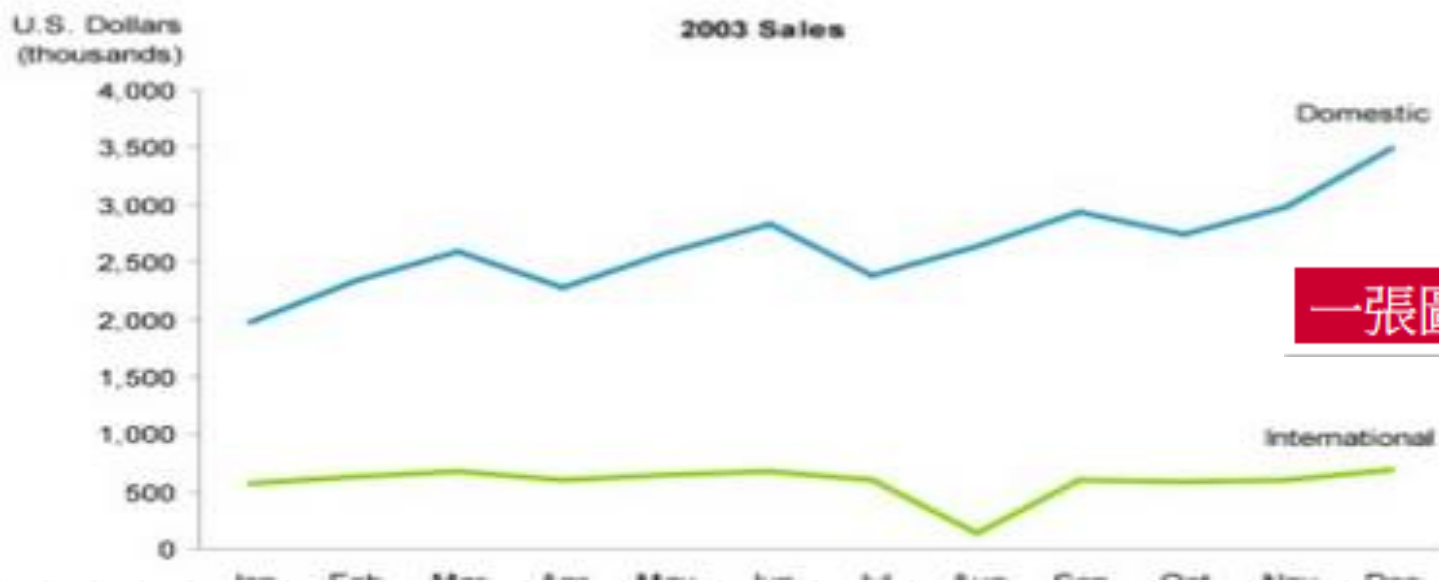


一張圖是否勝過千言萬語？

表格 v.s. 圖表

2003 Sales (U.S. dollars in thousands)

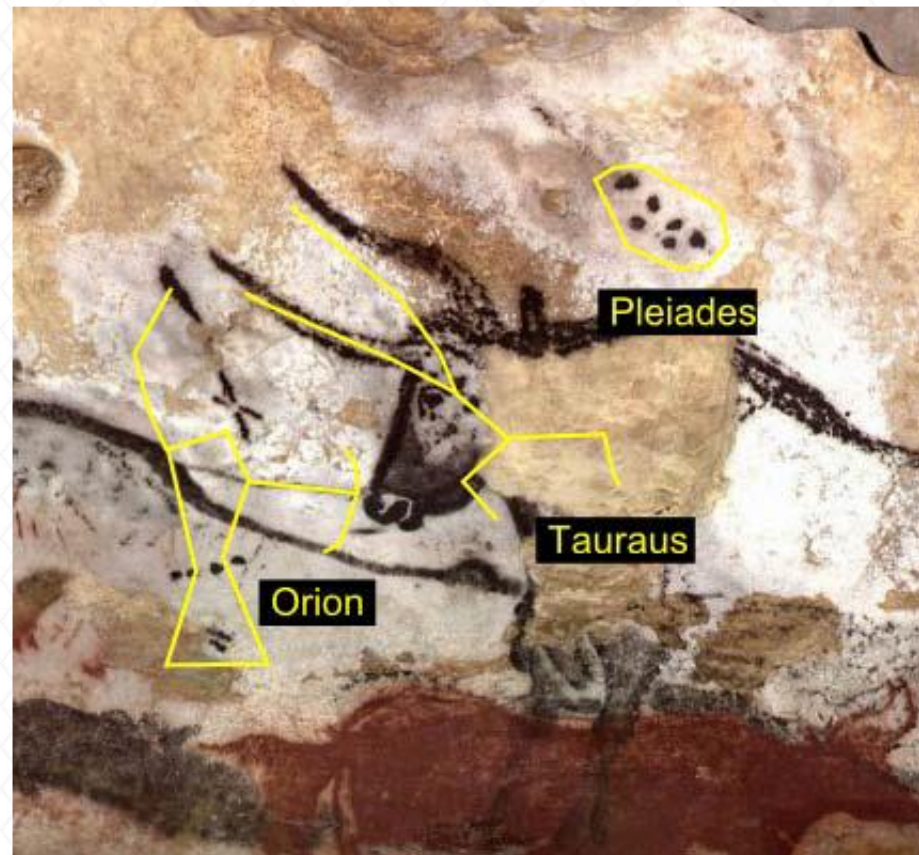
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Domestic	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493
International	574	636	673	593	644	679	593	139	599	583	602	690
	\$2,557	\$2,979	\$3,266	\$2,876	\$3,218	\$3,517	\$2,975	\$2,773	\$3,537	\$3,322	\$3,585	\$4,183



一張圖是否勝過千言萬語？

B.C. 15,000 的視覺化分析

- 法國拉斯考克附近的洞穴牆壁上，獵人畫下了他們所捕獵動物的圖案
- 描述遷移路線和軌跡線條和符木

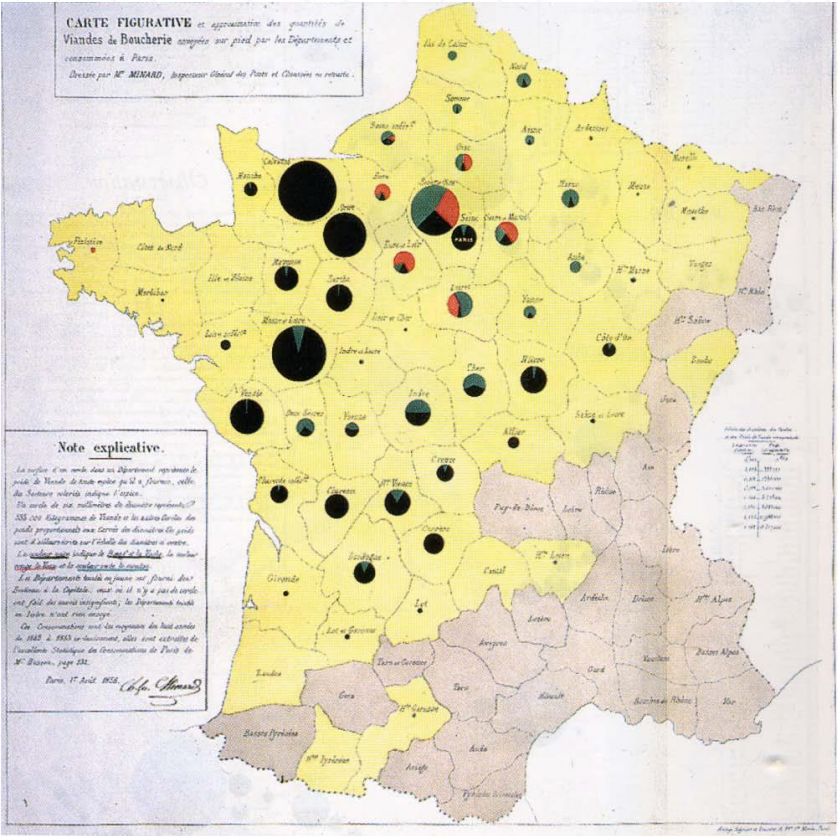


運用視覺化分析找出霍亂疫情元兇

- 1854 年，霍亂疫情爆發，造成十天之內死了五百多人
- Dr. John Snow 將所有病患的住家位置點在地圖上，發現病例聚集在一口井附近

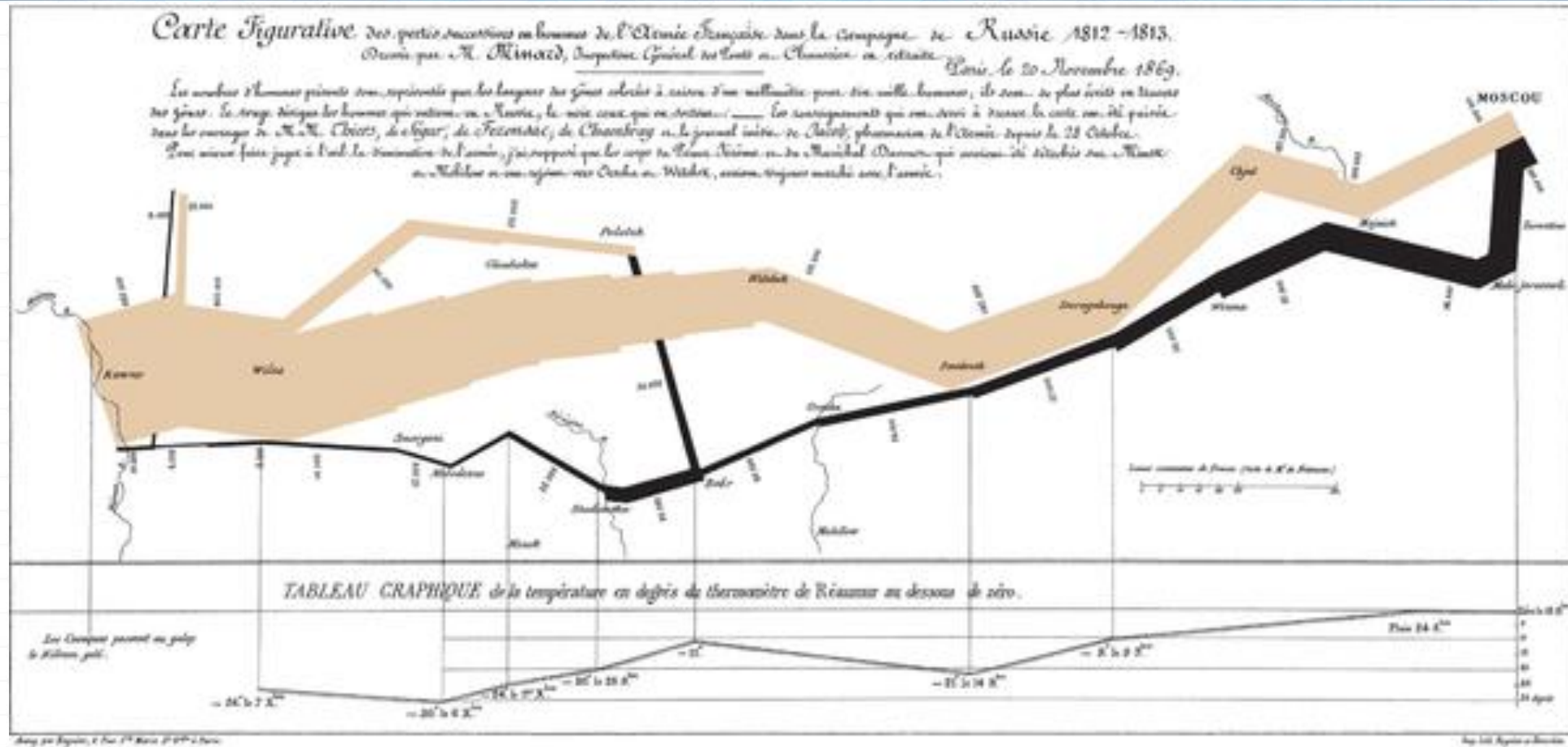


運用不同類型圖表分析地理資訊



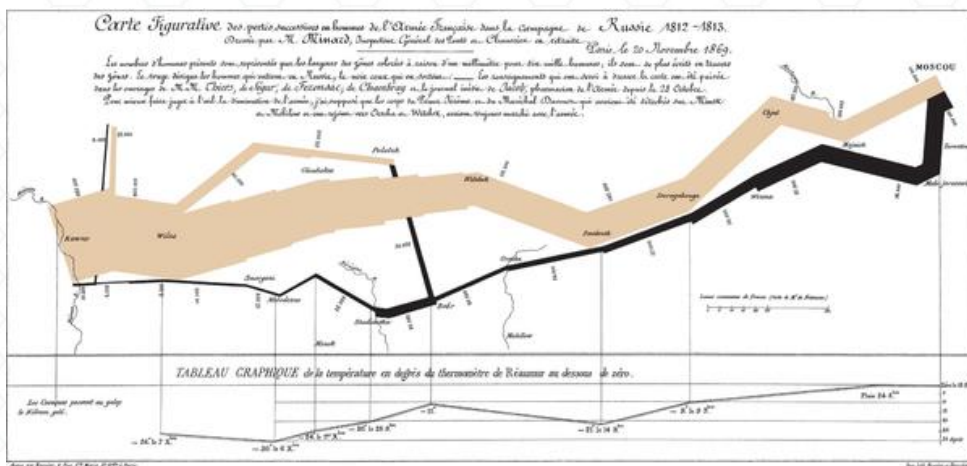
Charles Minard 結合Pie 圖與地圖繪製從法國各地運送到巴黎的牲畜

拿破崙遠征俄國



the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates

資訊圖表的功能



Storytelling 向聽眾溝通已知資訊



Exploration

從資料中發現背後的事實

資訊視覺化

■ 兩種視覺化功能

- Exploration – 發現資料背後蘊含的資訊
- Explanation – 向聽眾說故事

- 用一張圖一致性的呈現大量資料
- 協助使用者了解資料之間的關係
- 不去扭曲資料所要表達的原意
- 繪製圖表需要考量觀眾的期待

資訊視覺化的目標

■ 視覺化的目標

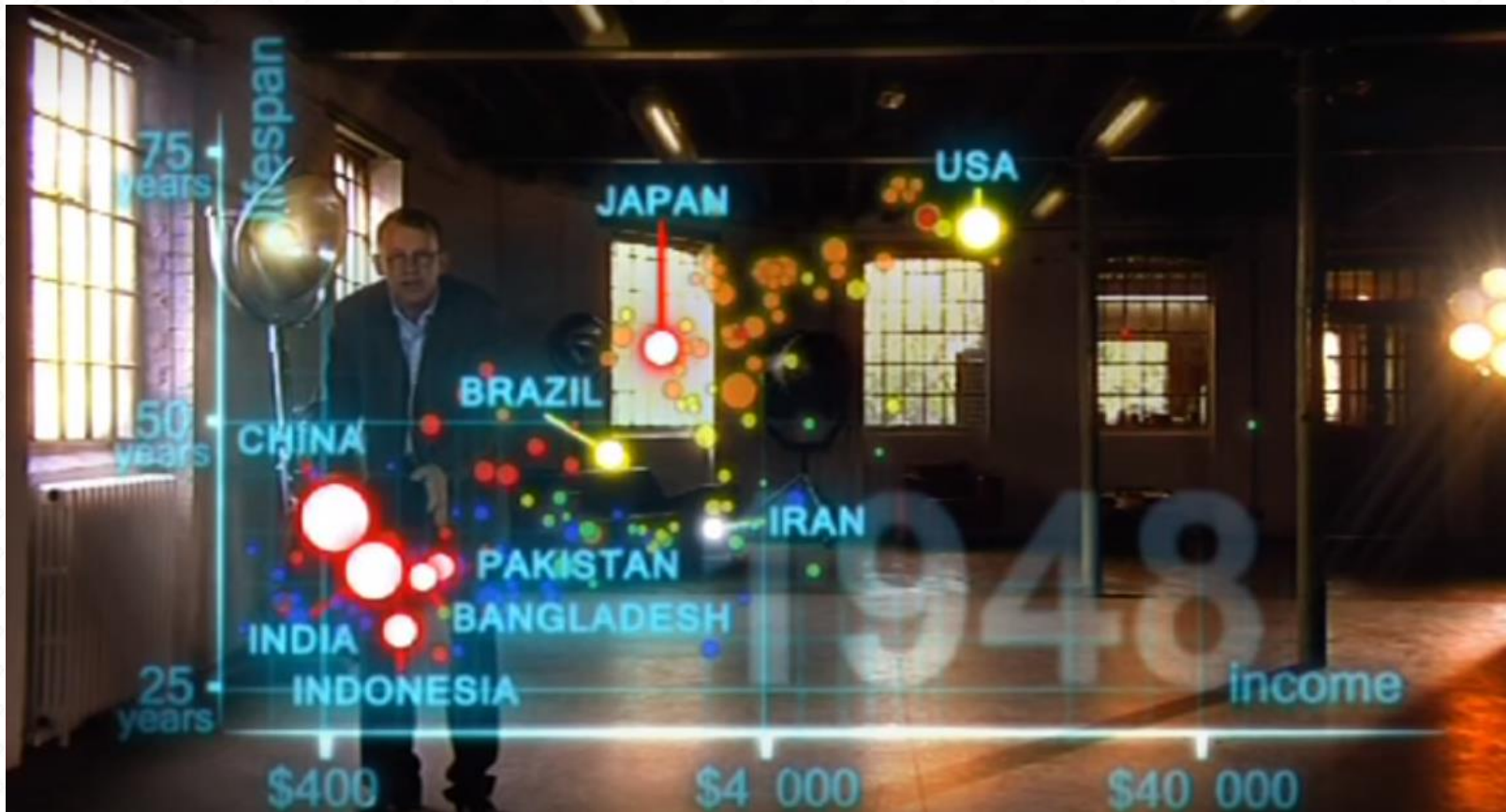
- 有效溝通
- 清楚
- 完整
- 促進參與者的互動

■ 專注在傳達的有效性

■ 視覺化 + 互動 = 成功的視覺化

Hans Rosling's 200 Countries, 200 Years, 4 Minutes

- <https://www.youtube.com/watch?v=jbkSRLYSojo>



資料類型

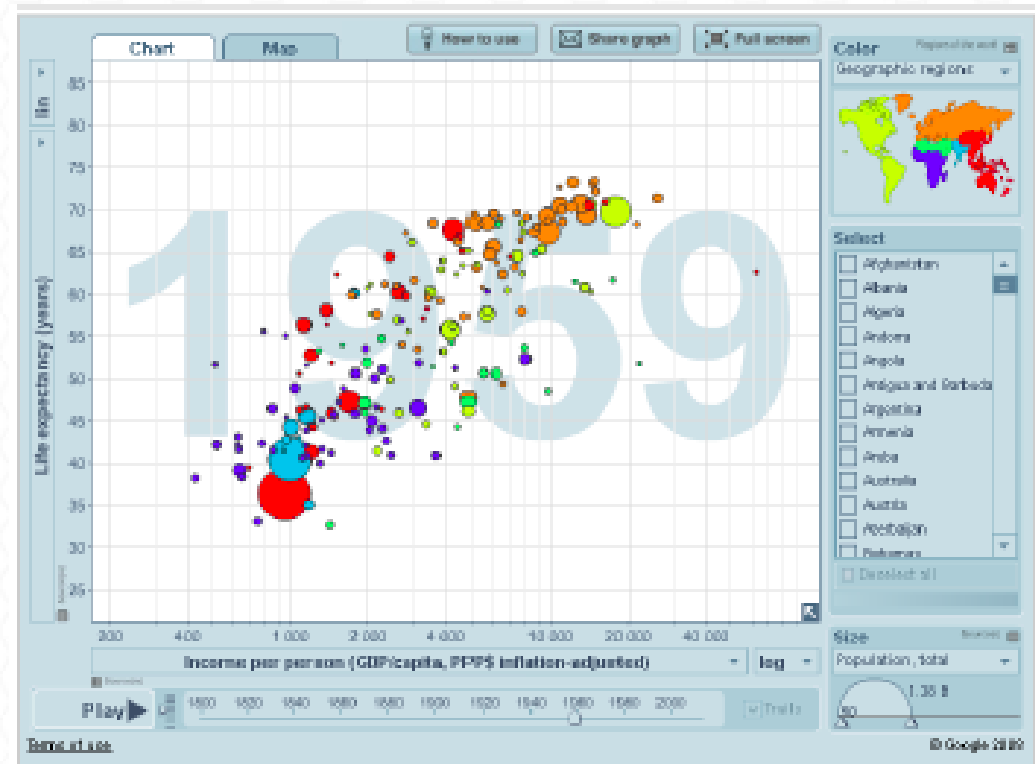
■ 量化資料

- 預期壽命
- 個人收入 (連續型)
- 總人口數 (離散)
- 年

■ 質化資料

- 以名稱列舉
 - 地理區域
- 排序
 - 人口數量範圍 (50 ~ 100萬)
 - 人口分層 (中產階級...)

<https://www.gapminder.org/world/>



基礎繪圖套件

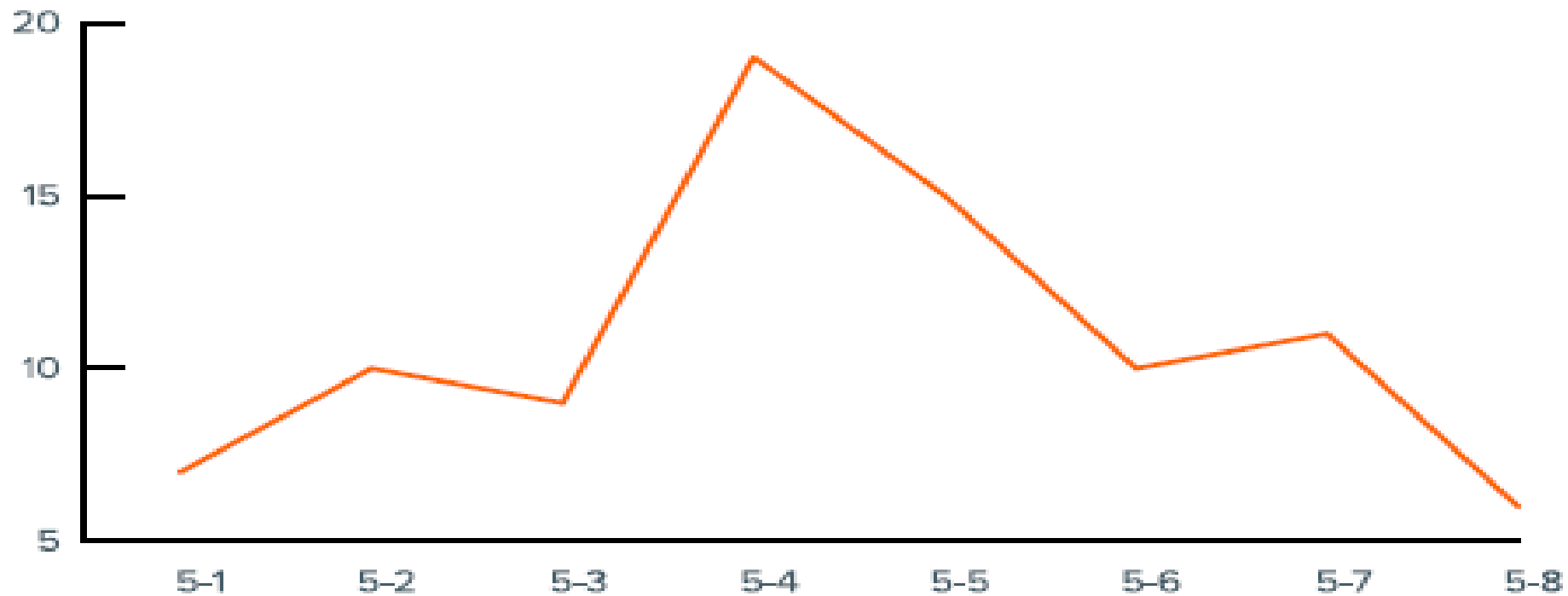
基礎繪圖套件 - graphics

■ 常用統計繪圖

Plot type	function
Line chart	plot() + lines()
bar chart	barplot()
Histogram	hist()
Pie chart	pie()
Mosaic	mosaicplot()
Box plot	boxplot()
Scatter plot	plot() + points()
Stem	stem()
其他補充	demo(image), demo(persp)

Line Chart

DIRECT MARKETING VIEWS, BY DATE



Line Chart

□ 函數：lines(x,y, type=)

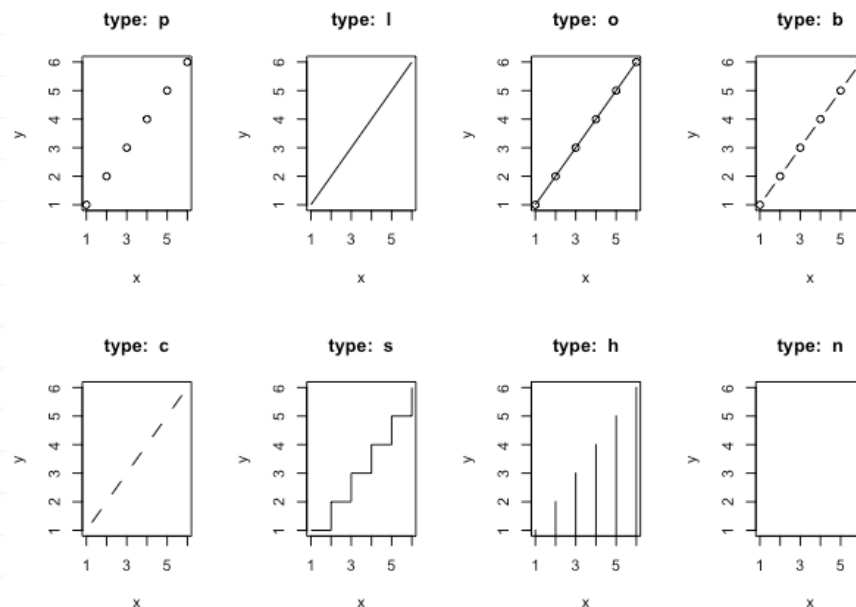
- x,y：數字向量，分別代表x軸與y軸的值。
- type：線的樣式。

□ 參數type的種類：

type	description
p	點。
l	直線。
o	點+直線。(兩者重疊)
b	點+直線。(不重疊)
c	點+直線。(點為空白)
S/s	階梯狀。
h	Histogram狀。
n	空樣式。

Line Chart







```
x <- seq(1,6)
y <- x
par(mfrow=c(2,4))
types = c("p","l","o","b","c","s","h","n")
for(i in 1:length(types)){
  title <- paste("type: ",types[i])
  plot(x, y, type="n", main=title)
  lines(x, y, type=types[i])
}
```



lines() 函數不可單獨使用，需搭配 plot(x,y) 函數一起使用
如果只有一組變數，可直接使用 plot(x,y,type=) 來取代 lines()

graphics各繪圖函數的常用參數整理

lty=

6.'twodash'	
5.'longdash'	
4.'dotdash'	
3.'dotted'	
2.'dashed'	
1.'solid'	
0.'blank'	

pch=

1 = ○	2 = △	3 = +	4 = ×	5 = ◇
6 = ▽	7 = ☒	8 = ※	9 = ◈	10 = ⊕
11 = ☒	12 = ▦	13 = ☒	14 = ☒	15 = ■
16 = ●	17 = ▲	18 = ◆	19 = ●	20 = ●
21 = ●	22 = ■	23 = ◆	24 = ▲	25 = ▼

Line Chart

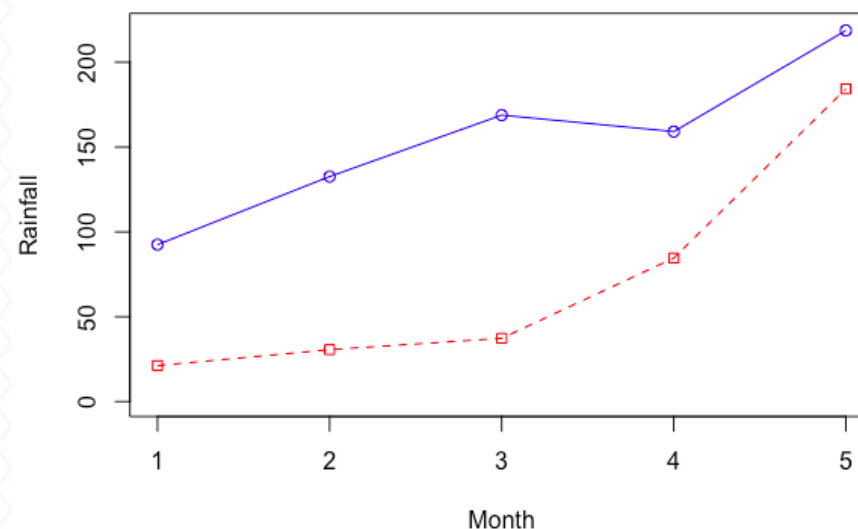
■ 遇到兩組以上的變數，則畫第二條線則要使用`lines()`函數

```
Taipei <- c(92.5, 132.6, 168.8, 159.1, 218.7)
```

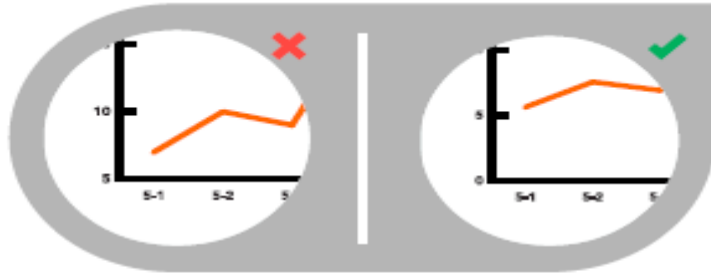
```
Tainan <- c(21.2, 30.6, 37.3, 84.6, 184.3)
```

```
plot(Taipei, type="o", col="blue", ylim=c(0,220),  
      xlab="Month", ylab="Rainfall")
```

```
lines(Tainan, type="o", pch=22, lty=2, col="red")
```

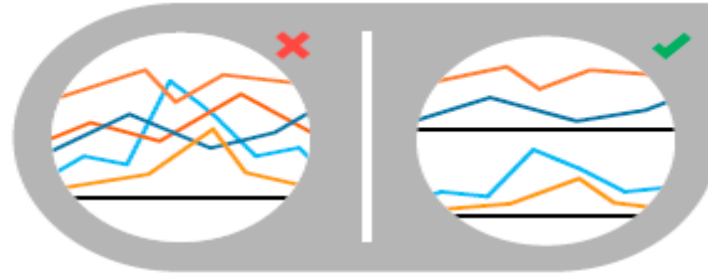


Line Chart 設計原則



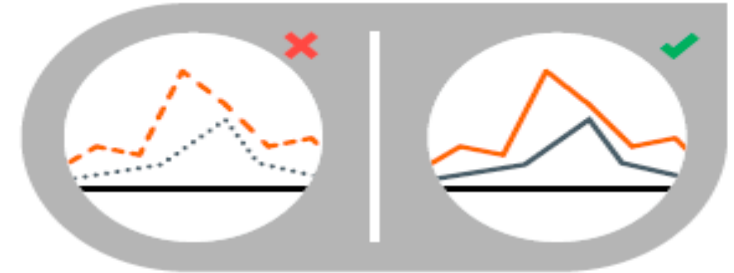
INCLUDE A ZERO BASELINE IF POSSIBLE

Although a line chart does not have to start at a zero baseline, it should be included if possible. If relatively small fluctuations in data are meaningful (e.g., in stock market data), you may truncate the scale to showcase these variances.



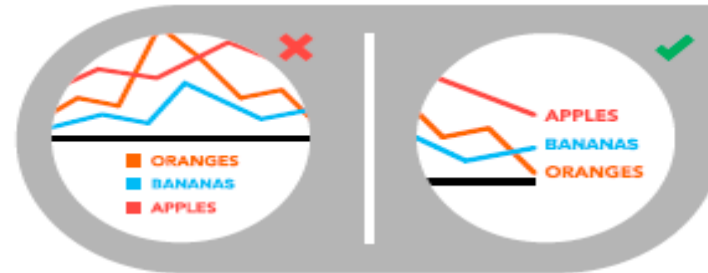
DON'T PLOT MORE THAN 4 LINES

If you need to display more, break them out into separate charts for better comparison.



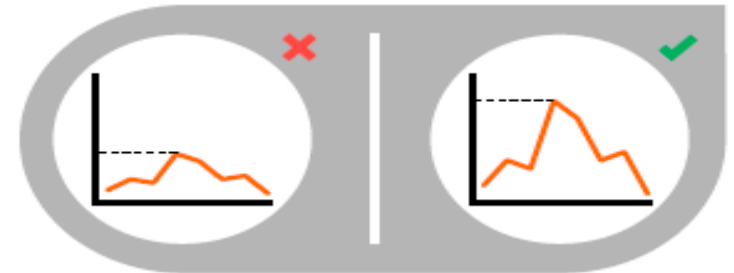
USE SOLID LINES ONLY

Dashed and dotted lines can be distracting.



LABEL THE LINES DIRECTLY

This lets readers quickly identify lines and corresponding labels instead of referencing a legend.

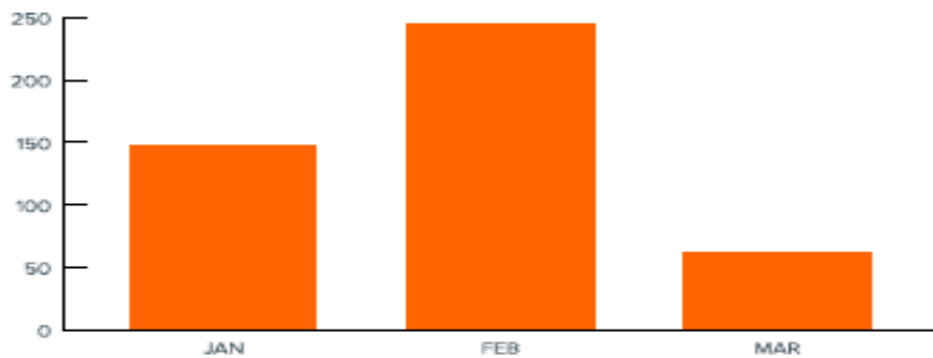


USE THE RIGHT HEIGHT

Plot all data points so that the line chart takes up approximately two-thirds of the y-axis' total scale.

Bar Chart 的種類

PAGE VIEWS, BY MONTH



**VERTICAL
(COLUMN CHART)**

Best used for chronological data (time-series should always run left to right), or when visualizing negative values below the x-axis.

CONTENT PUBLISHED, BY CATEGORY

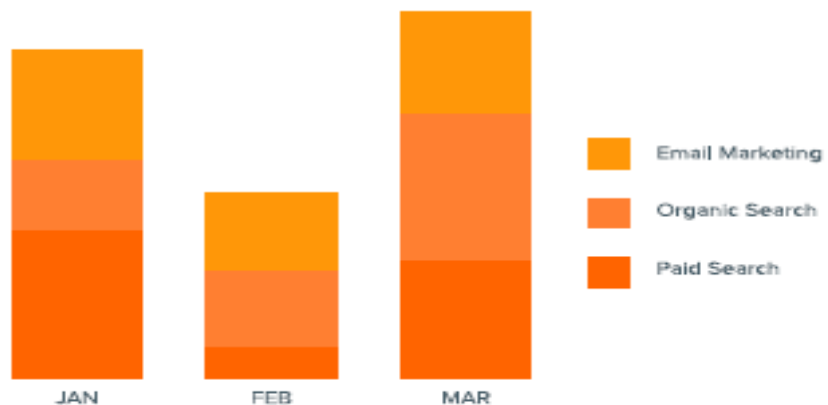


HORIZONTAL

Best used for data with long category labels.

Bar Chart 的種類

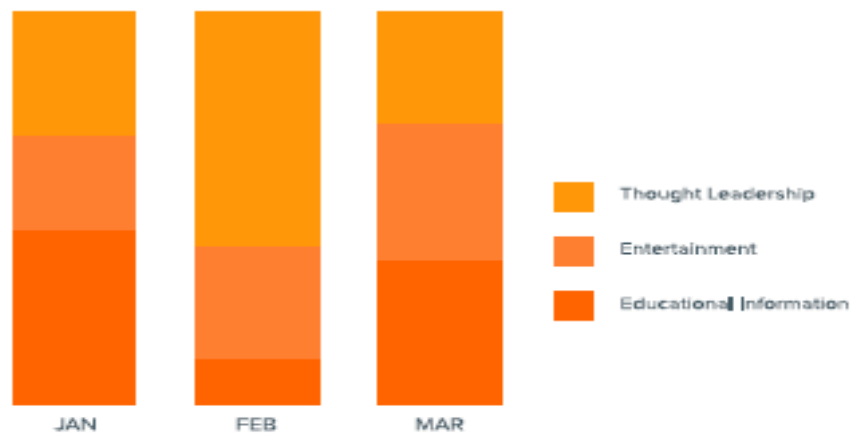
MONTHLY TRAFFIC, BY SOURCE



STACKED

Best used when there is a need to compare multiple part-to-whole relationships. These can use discrete or continuous data, oriented either vertically or horizontally.

PERCENTAGE OF CONTENT PUBLISHED, BY MONTH



100% STACKED

Best used when the total value of each category is unimportant and percentage distribution of subcategories is the primary message.

Bar Chart

■ 函數：barplot()

■ 範例：

```
housePrice <- read.csv('house-prices.csv',header = TRUE)
```

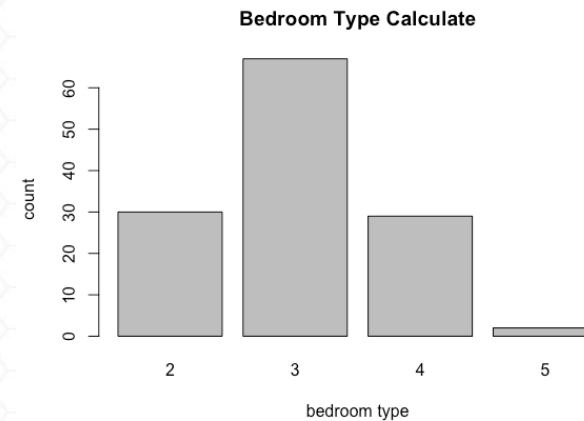
```
bedrooms <- housePrice$Bedrooms
```

```
bedroomsTable <- table(bedrooms)
```

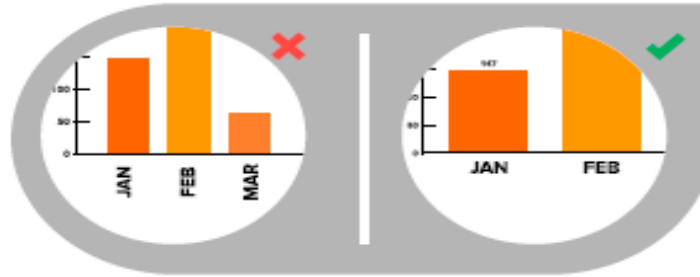
```
barplot(bedroomsTable, main="Bedroom Type Calculate",
```

```
      xlab="bedroom type", ylab="count")
```

	Home	Price	SqFt	Bedrooms	Bathrooms	Offers	Brick	Neighborhood
1	1	114300	1790	2	2	2	No	East
2	2	114200	2030	4	2	3	No	East
3	3	114800	1740	3	2	1	No	East
4	4	94700	1980	3	2	3	No	East
5	5	119800	2130	3	3	3	No	East
6	6	114600	1780	3	2	2	No	North
7	7	151600	1830	3	3	3	Yes	West
8	8	150700	2160	4	2	2	No	West
9	9	119200	2110	4	2	3	No	East
10	10	104000	1730	3	3	3	No	East
11	11	132500	2030	3	2	3	Yes	East
12	12	123000	1870	2	2	2	Yes	East



Bar Chart 設計原則



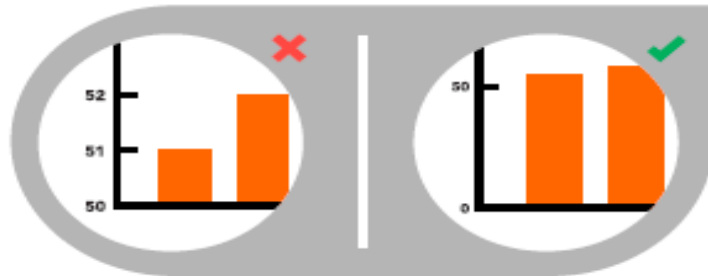
USE HORIZONTAL LABELS

Avoid steep diagonal or vertical type, as it can be difficult to read.



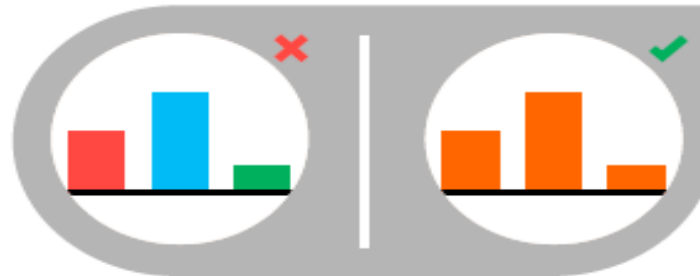
SPACE BARS APPROPRIATELY

Space between bars should be $\frac{1}{2}$ bar width.



START THE Y-AXIS VALUE AT 0

Starting at a value above zero truncates the bars and doesn't accurately reflect the full value.



USE CONSISTENT COLORS

Use one color for bar charts. You may use an accent color to highlight a significant data point.



ORDER DATA APPROPRIATELY

Order categories alphabetically, sequentially, or by value.

Histogram

■ 函數 : hist()

```
load('cdc.Rdata')
```

```
View(cdc)
```

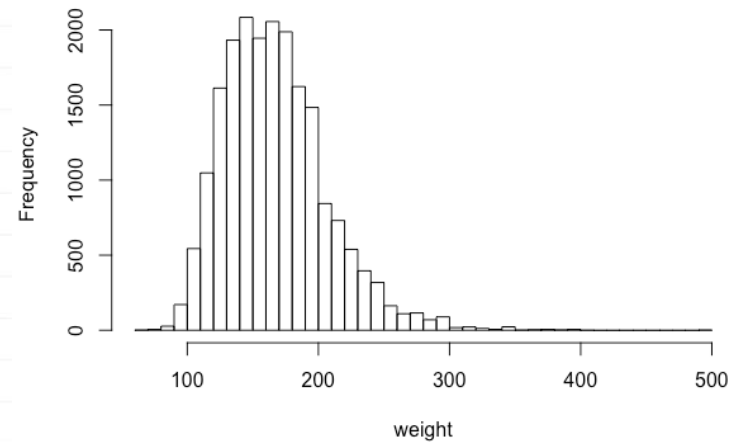
```
weight <- cdc$weight
```

```
hist(weight,breaks=50)
```

	genhlth	exerany	hlthplan	smoke100	height	weight	wtdesired	age	gender
1	good	0	1	0	70	175	175	77	m
2	good	0	1	1	64	125	115	33	f
3	good	1	1	1	60	105	105	49	f
4	good	1	1	0	66	132	124	42	f
5	very good	0	1	0	61	150	130	55	f
6	very good	1	1	0	64	114	114	55	f
7	very good	1	1	0	71	194	185	31	m
8	very good	0	1	0	67	170	160	45	m
9	good	0	1	1	65	150	130	27	f
10	good	1	1	0	70	180	170	44	m
11	excellent	1	1	1	69	186	175	46	m
12	fair	1	1	1	69	168	148	62	m



Histogram of weight



Histogram vs Bar Chart

■ Histogram : y軸數值為x軸各區間資料之頻率

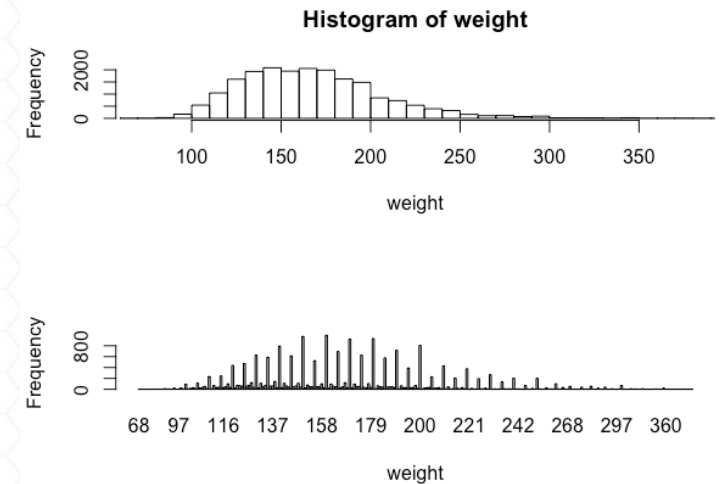
■ Barplot : y軸數值為x軸資料相對應的原始數值

```
par(mfrow=c(2,1))
```

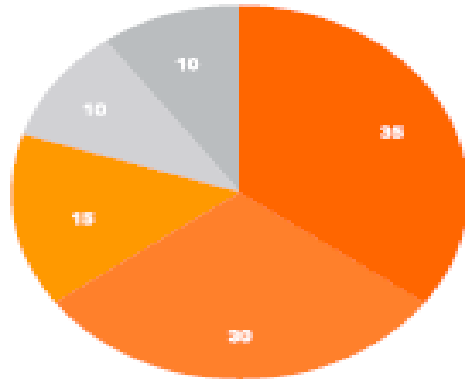
```
hist(weight,breaks=50,xlim=c(70,380))
```

```
barplot(table(cdc$weight),xlab="weight",ylab="Frequency")
```

```
par(mfrow=c(1,1))
```

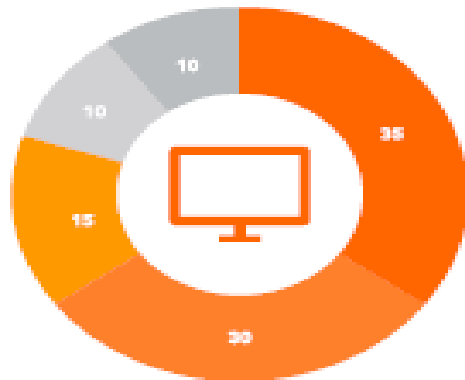


PIE Chart



STANDARD

Used to show part-to-whole relationships.



DONUT

Stylistic variation that enables the inclusion of a total value or design element in the center.

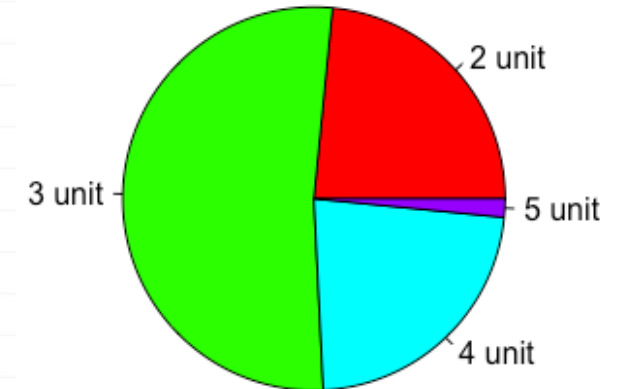
PIE Chart

■ 函數：pie()

■ 範例：

```
housePrice <- read.csv('house-prices.csv',header = TRUE)
bedrooms <- housePrice$Bedrooms
bedroomsTable = table(bedrooms)
Labels <- c("2 unit", "3 unit", "4 unit", "5 unit")
pie(bedroomsTable,labels=labels,
    col=rainbow(length(labels)),
    main="Pie Chart of Bedroom")
```

Pie Chart of Bedroom



PIE Chart 設計原則



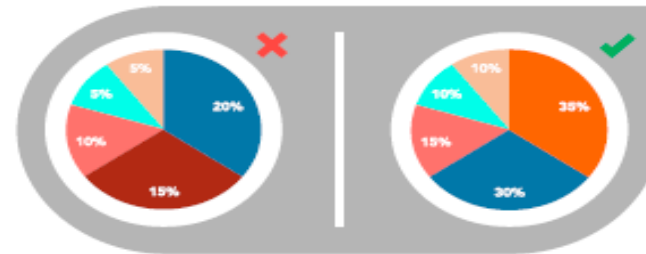
VISUALIZE NO MORE THAN 5 CATEGORIES PER CHART

It is difficult to differentiate between small values; depicting too many slices decreases the impact of the visualization. If needed, you can group smaller values into an “other” or “miscellaneous” category, but make sure it does not hide interesting or significant information.



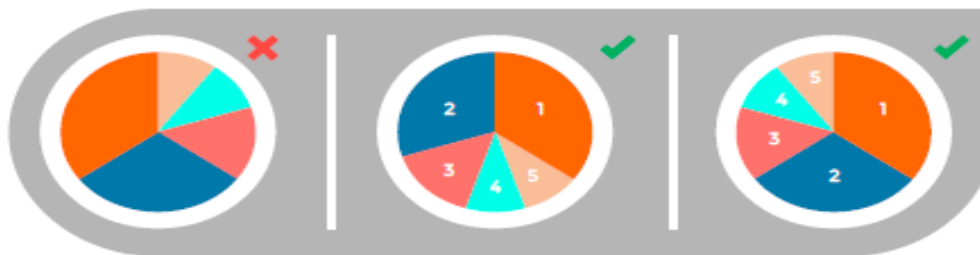
DON'T USE MULTIPLE PIE CHARTS FOR COMPARISON

Slice sizes are very difficult to compare side-by-side. Use a stacked bar chart instead.



MAKE SURE ALL DATA ADDS UP TO 100%

Verify that values total 100% and that pie slices are sized proportionate to their corresponding value.



ORDER SLICES CORRECTLY

There are two ways to order sections, both of which are meant to aid comprehension:

OPTION 1

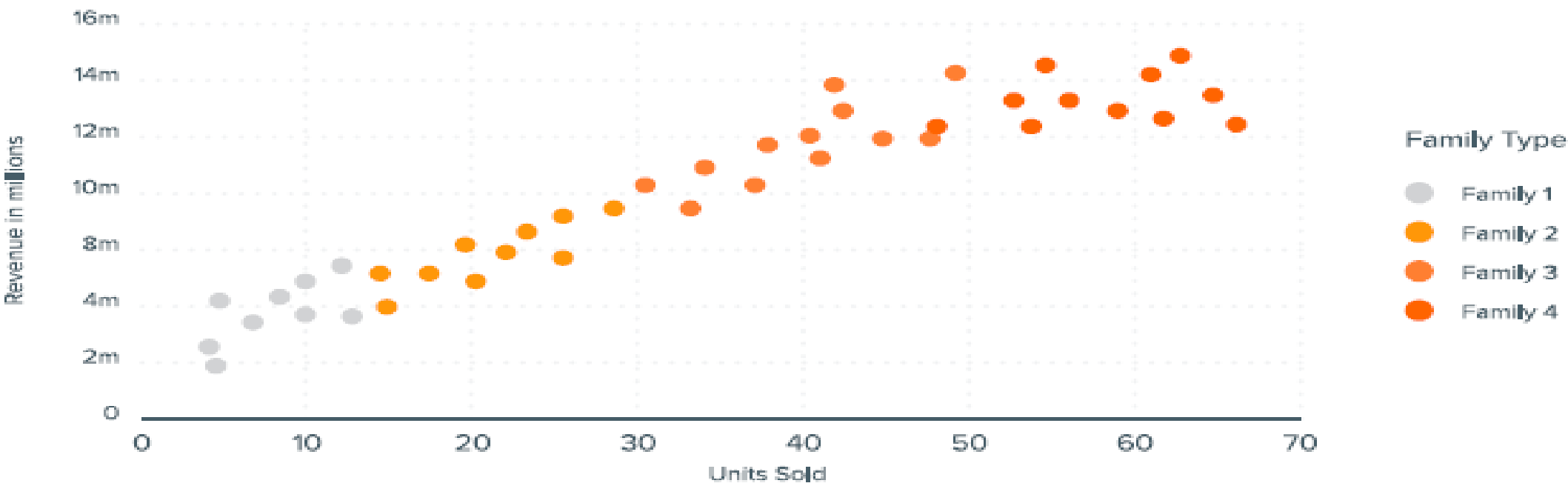
Place the largest section at 12 o'clock, going clockwise. Place the second largest section at 12 o'clock, going counterclockwise. The remaining sections can be placed below, continuing counterclockwise.

OPTION 2

Start the largest section at 12 o'clock, going clockwise. Place remaining sections in descending order, going clockwise.

Scatter Plot

REVENUE, BY PRODUCT FAMILY

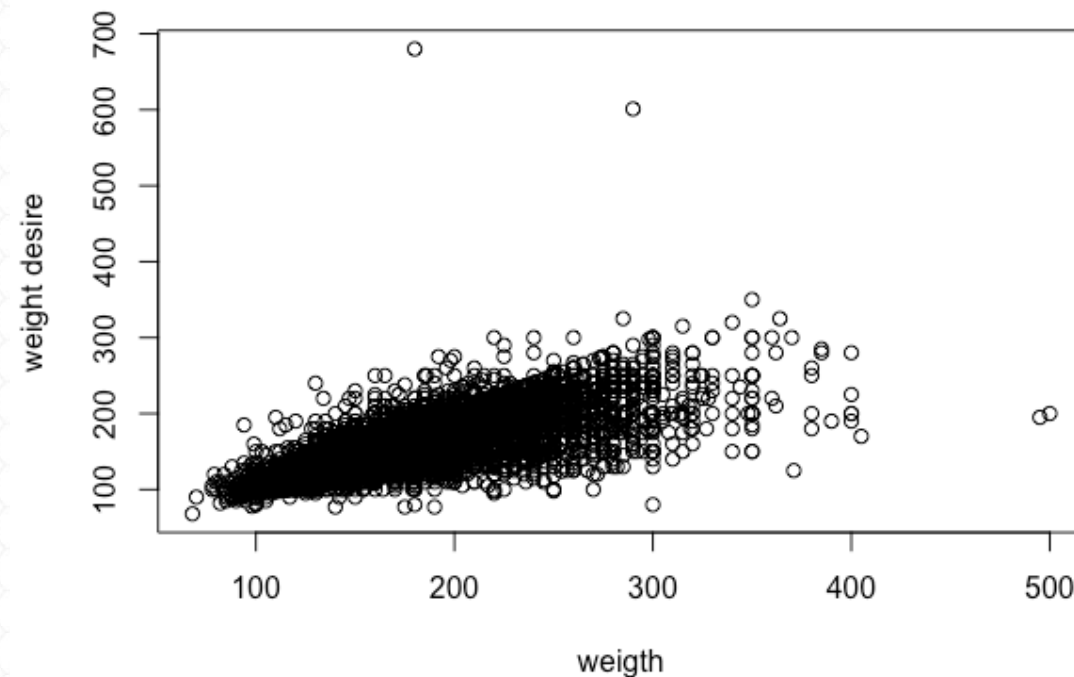


Scatter Plot

■ 函數：plot(x,y)

■ 範例：

`plot(cdc$weight, cdc$wt Desire)`

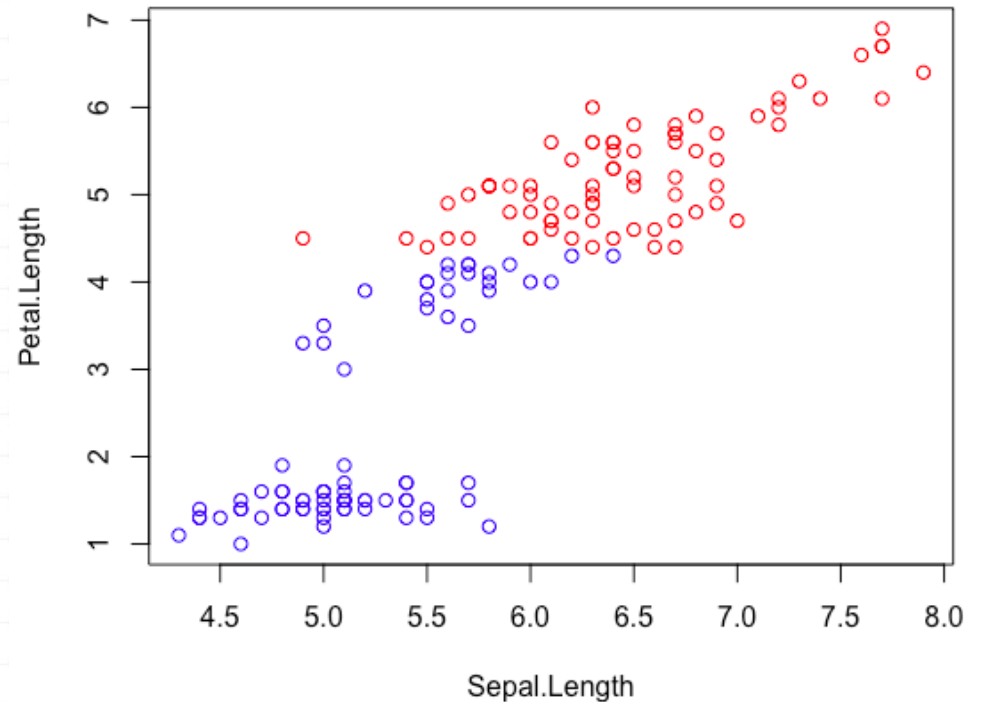


Scatter Plot

■ 函數：plot(x, y) + points(x,y,col=)

■ 範例：

```
data(iris)
xlab = names(iris)[1]
ylab = names(iris)[3]
x = iris[,1]
y = iris[,3]
plot(x, y, xlab=xlab, ylab=ylab,
     col=ifelse(iris[,3] > median(iris[,3]), "red", "blue"))
```



Scatter Chart

■ 函數：plot(x, y) + points(x,y,col=)

■ 範例：

```
data(iris)
```

```
xlab <- names(iris)[1]
```

```
ylab <- names(iris)[3]
```

```
x <- iris[,1]
```

```
y <- iris[,3]
```

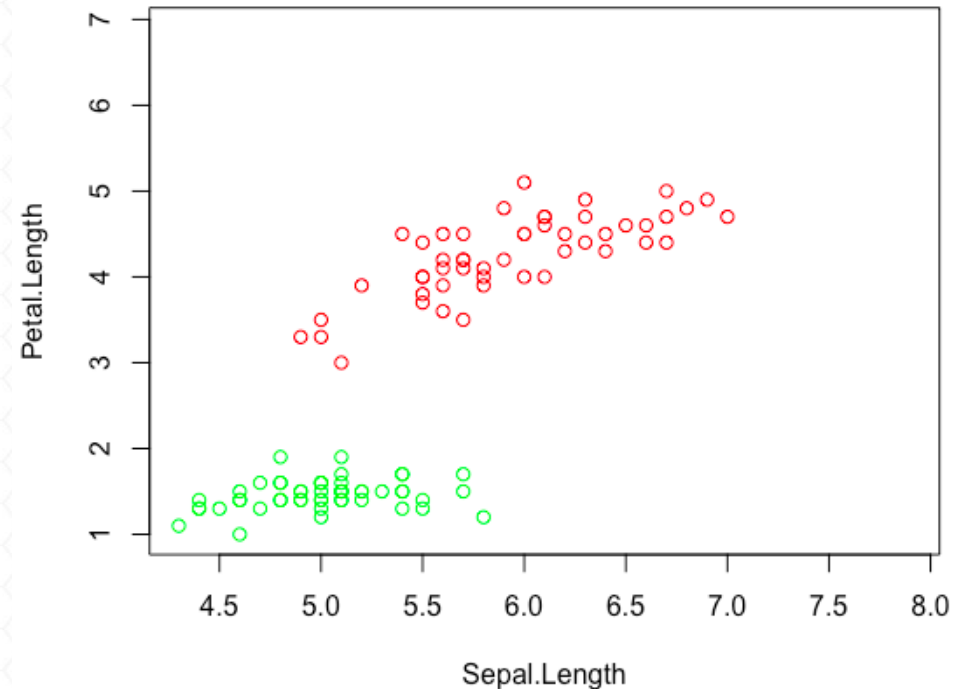
```
plot(x, y, xlab=xlab, ylab=ylab,type="n")
```

```
setosa <- which(iris$Species=="setosa")
```

```
versicolor <- which(iris$Species=="versicolor")
```

```
points(iris[setosa,1],iris[setosa,3],col="green")
```

```
points(iris[versicolor ,1],iris[versicolor,3],col="red")
```

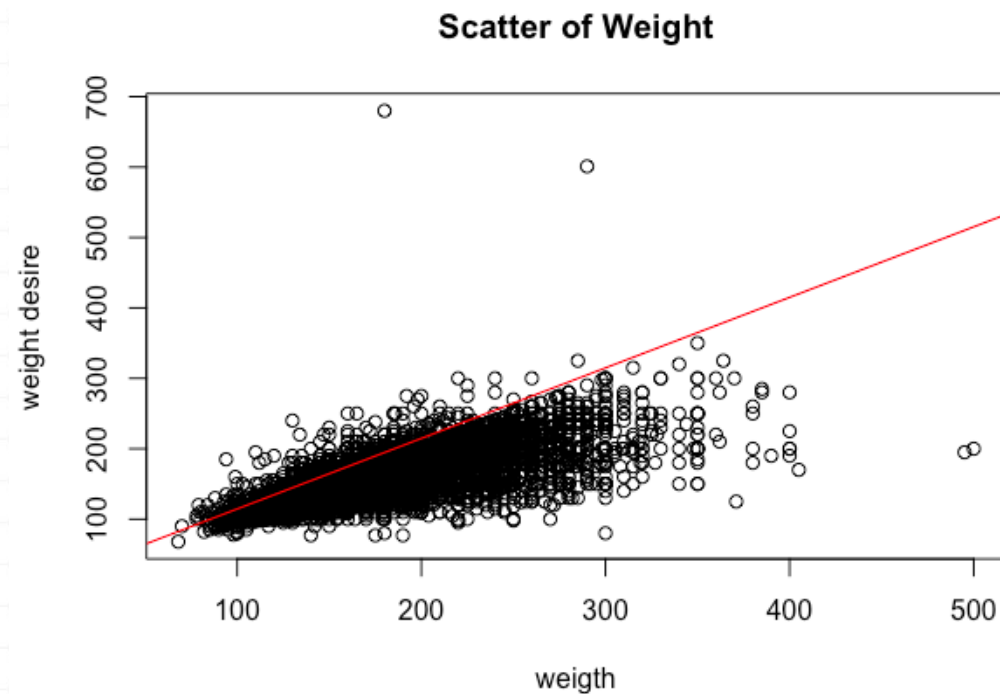


Linear Regression

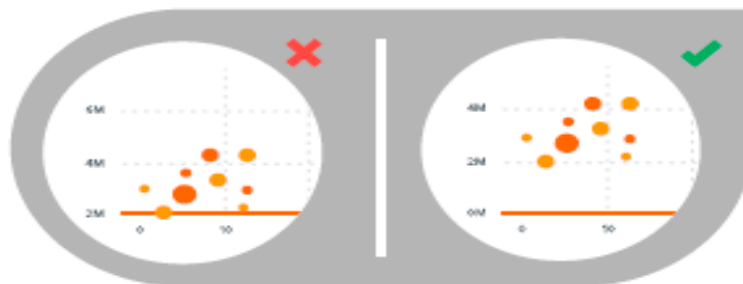
■ 函數：lm(x~y) 搭配 abline(lm(x~y))

■ 範例：

```
plot(cdc$weight, cdc$wt Desire,  
     xlab="weight", ylab="weight desire",  
     main="Scatter of Weight")  
abline(lm(cdc$weight~cdc  
$wt Desire), col="red")
```

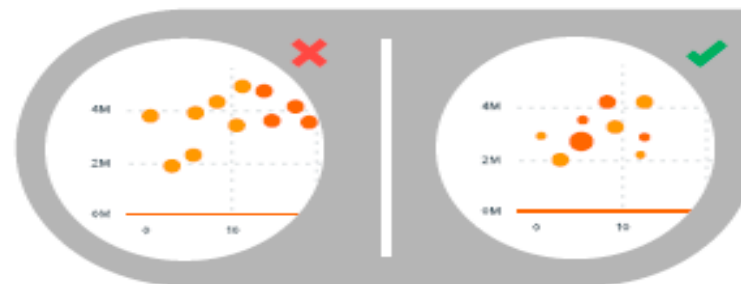


Scatter Plot 設計原則



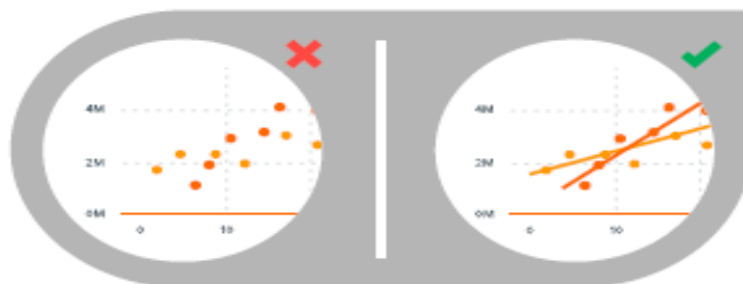
START Y-AXIS VALUE AT 0

Starting the axis above zero truncates the visualization of values.



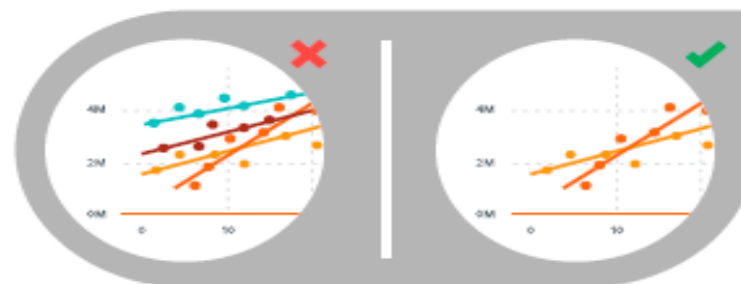
INCLUDE MORE VARIABLES

Use size and dot color to encode additional data variables.



USE TREND LINES

These help draw correlation between the variables to show trends.



DON'T COMPARE MORE THAN 2 TREND LINES

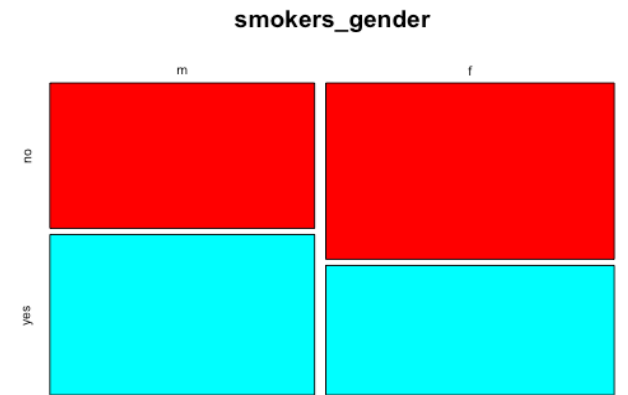
Too many lines make data difficult to interpret.

Mosaic Chart

■ 函數：mosaicplot()

■ 範例：

```
smokers_gender <- table(cdc$gender, cdc$smoke100)
colnames(smokers_gender) <- c("no", "yes")
mosaicplot(smokers_gender
            ,col=rainbow(length(colnames(smokers_gender))))
```

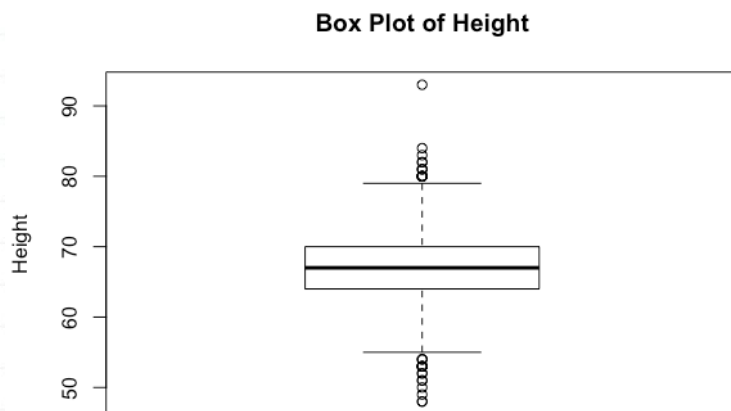


Box Chart

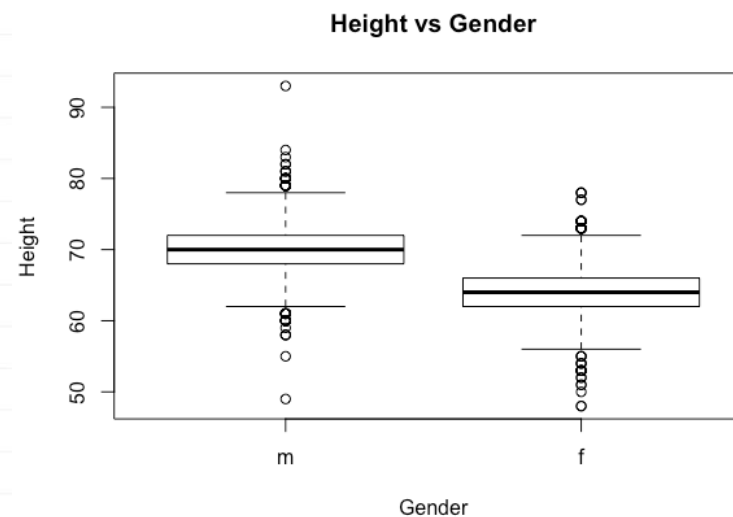
■ 函數：boxplot()

■ 範例：

```
boxplot(cdc$height,  
        ylab="Height",  
        main="Box Plot of Height")
```

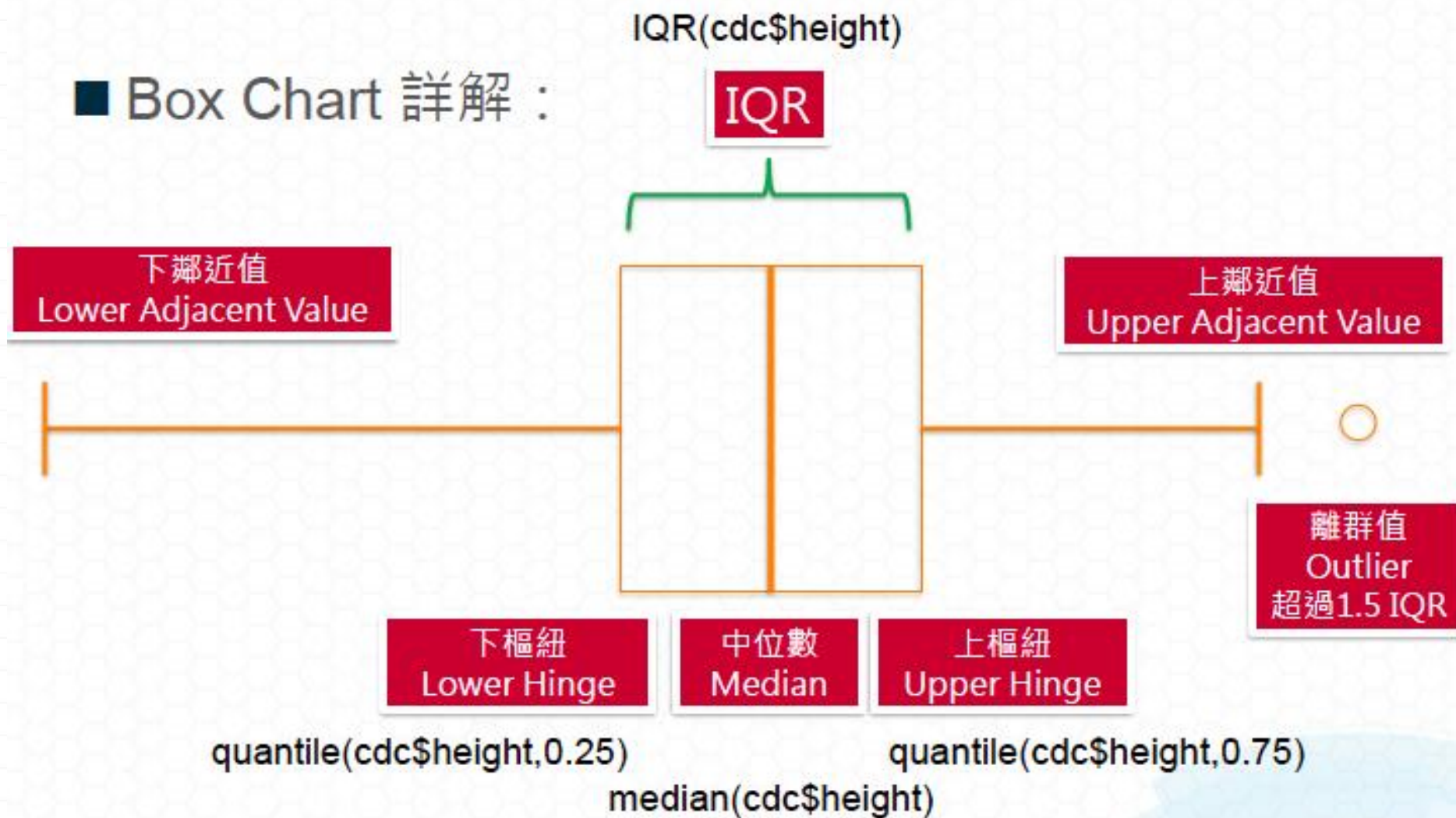


```
boxplot(cdc$height ~ cdc$gender  
        ,ylab="Height",xlab="Gender"  
        ,main="Height vs Gender")
```



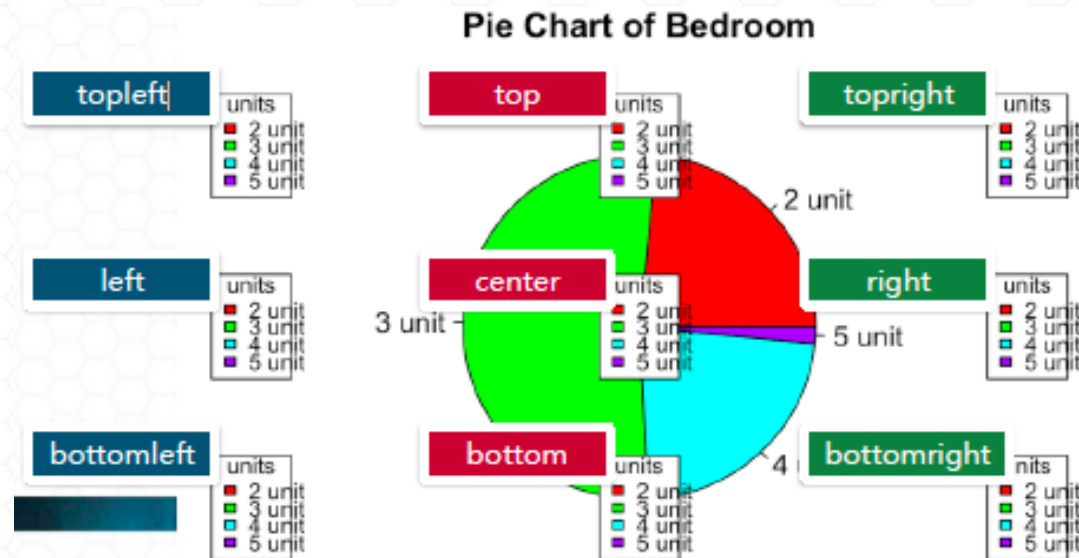
Box Chart

■ Box Chart 詳解：



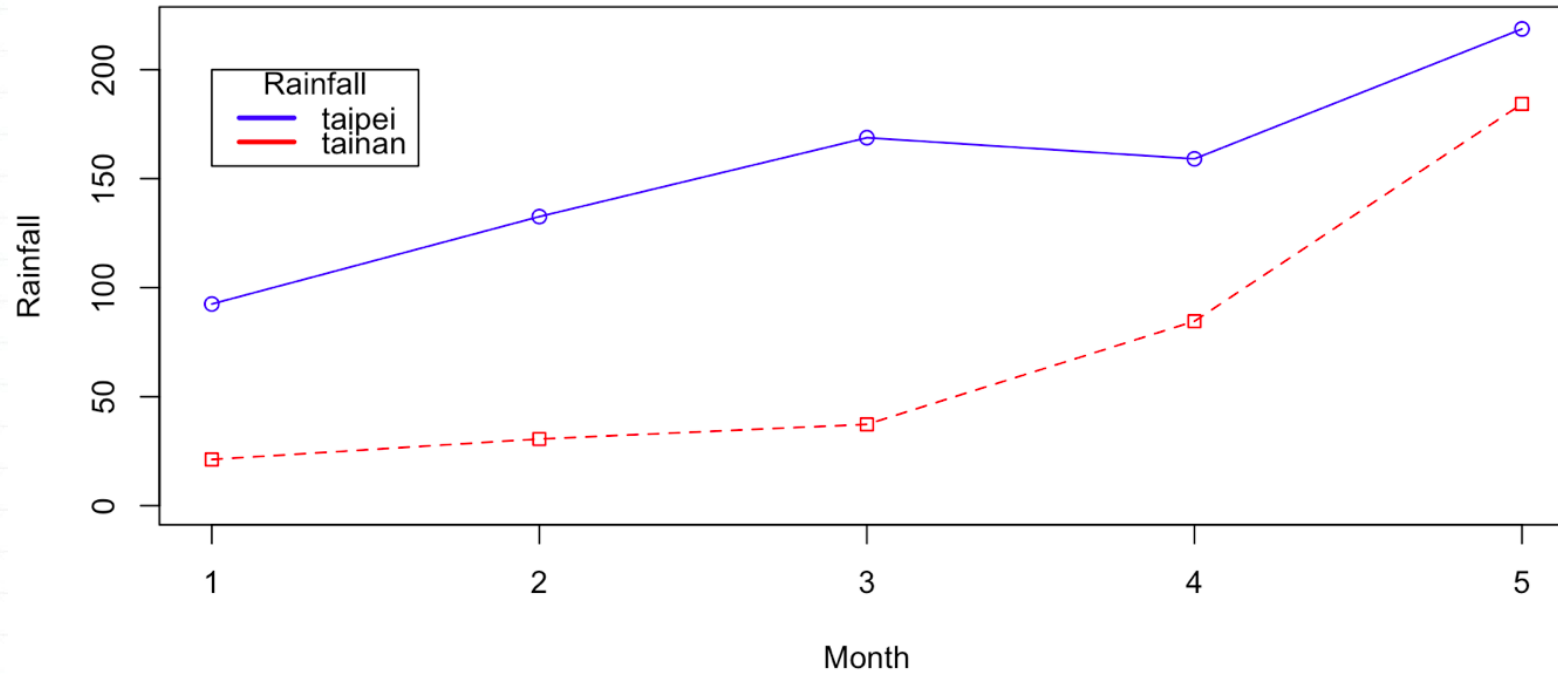
Legend

- 參數：legend(x, y = NULL,...)
- 說明：當圖有x,y軸時(例如 line chart)，x與y參數可指定legend放在座標軸哪個位置；當圖沒有x,y軸時(例如piechart)，可選擇："bottomleft","bottom","bottomright","left","center","right","topleft","top","topright"等位置。



Legend

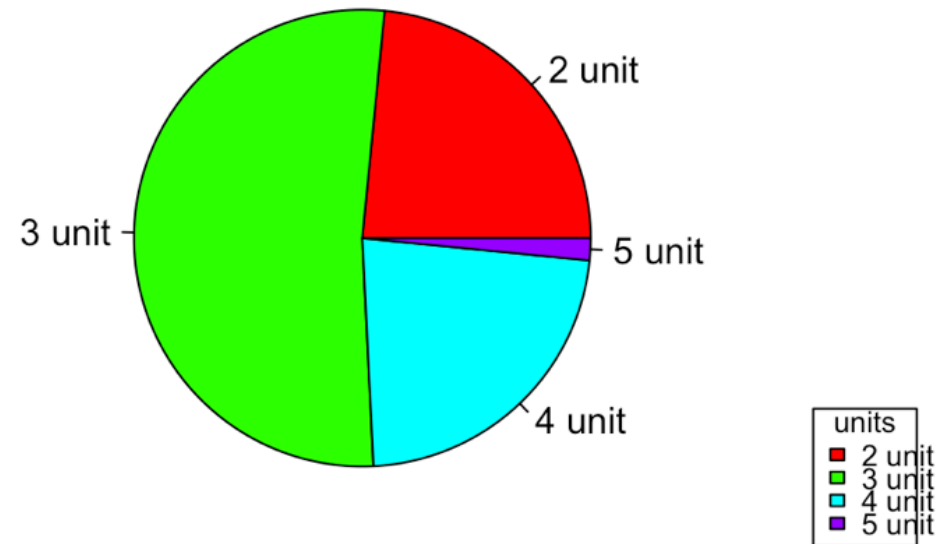
■ 舉例：`legend(1,200, c("taipei","tainan"), lwd=c(2.5,2.5),col=c("blue","red"), title = "Rainfall")`



Legend

- 舉例 : `legend("bottomright", labels,`
`fill=rainbow(length(labels)), title = "units", cex=0.8)`

Pie Chart of Bedroom



graphics各繪圖函數的常用參數整理

繪圖函數參數	說明
type	plot(),lines()的線條呈現樣式。
col/col.axis/col.lab/ col.main/col.sub	顏色。可用rainbow(n), heat.colors(n), terrain.colors(n), topo.colors(n), and cm.colors(n)等函式填充，n代表顏色個數;也可自行輸入顏色代碼。關於顏色代碼請參考： http://research.stowers-institute.org/efg/R/Color/Chart/index.htm
xlim/ylim	x/y軸的上下限。
xaxt/yaxt	x/y軸的顯示，設為“n”則不顯示x/y軸的字。
cex	字體大小。
pch	點的樣式。(預設為1)
lty	線的樣式。(預設為1)
lwd	線的粗細。(預設為1)
main	圖的標題。
xlab/ylab	x/y軸的顯示名稱。
breaks	hist()要大約分成幾組。(套件會自動優化到最佳組數)

全局圖形設定函數：par()

■ 函數：par()

■ 參數：

- `par(mfrow=c(n,m))`：n行 x m欄，由" row"開始填充。
- `par(mfcol=c(n,m))`：n行 x m欄，由" column"開始填充。
- `par(mar=c(bottom,left,top,right))`：圖的margin(單位為number of lines)，需帶入下方、左側、上方、右側四個margin值，預設為`c(5,4,4,2)+0.1`。
- `par(mai=c(bottom,left,top,right))`：同mar，但單位更換為inches。
- 其他詳細資訊可打`?par`查詢。

全局圖形設定函數：par()

```
showLayout = function(n){  
  for(i in 1:n){  
    plot(1,type="n",xaxt="n",yaxt="n",xlab="",ylab="")  
    text(1, 1, labels=i, cex=10)  
  }  
}
```

```
par(mar=c(1,1,1,1),mfrow=c(3,2))  
showLayout(6)
```

```
par(mar=c(3,3,3,3),mfrow=c(3,2))  
showLayout(6)
```

```
par(mar=c(3,3,3,3),mfc col=c(3,2))  
showLayout(6)
```

par(mar=c(1,1,1,1),
mfrow=c(3,2))

1	2
3	4
5	6

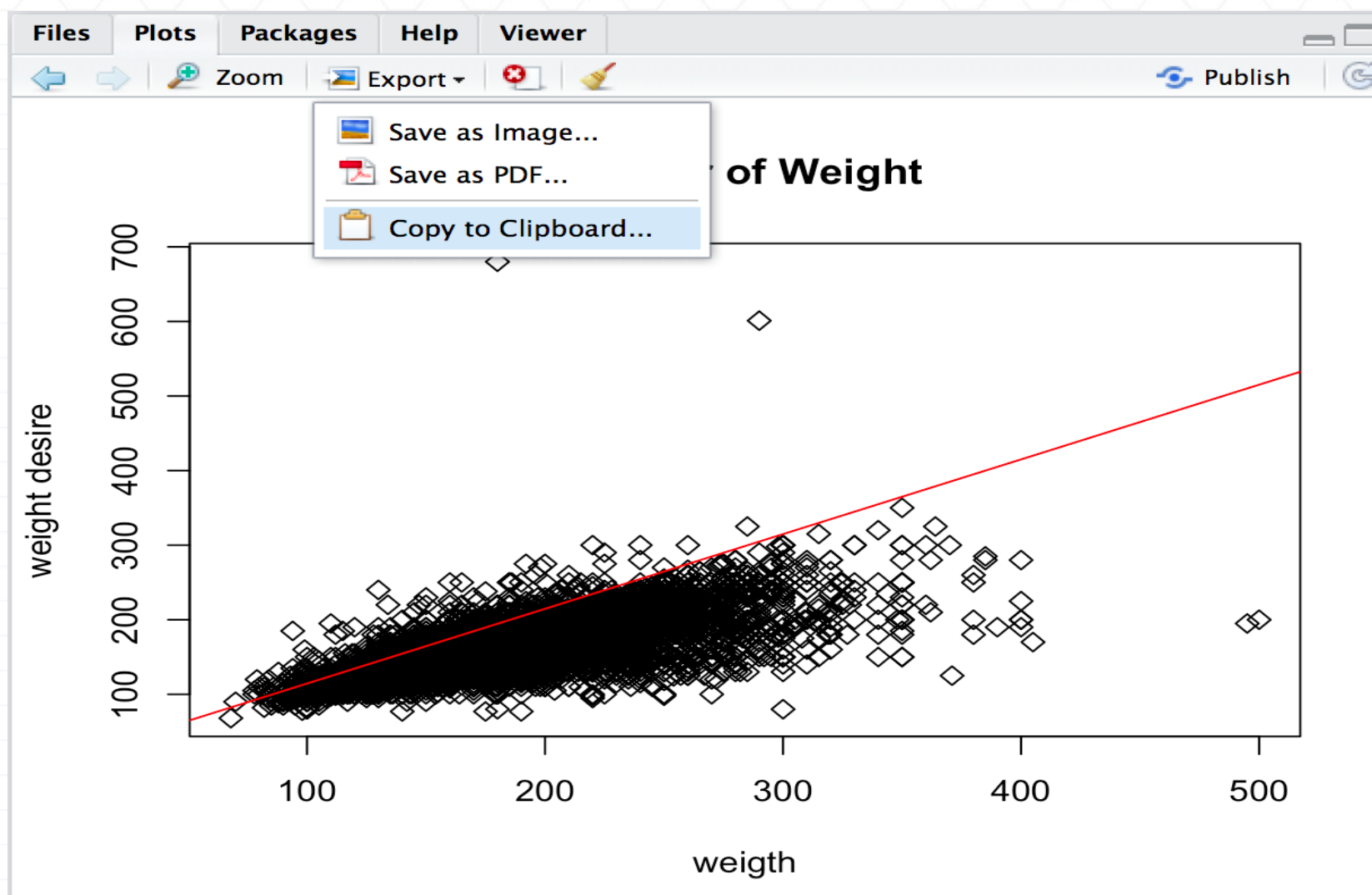
par(mar=c(3,3,3,3),
mfrow=c(3,2))

1	2
3	4
5	6

par(mar=c(3,3,3,3),
mfc col=c(3,2))

1	4
2	5
3	6

圖的輸出



圖的輸出

■ 使用函數：

繪圖函數參數	說明
<code>pdf("xxfile.pdf")</code>	輸出.pdf檔。
<code>png("xxfile.png")</code>	輸出.png檔。
<code>jpeg("xxfile.jpeg")</code>	輸出.jpeg檔。
<code>bmp("xxfile.bmp")</code>	輸出.bmp檔。

■ 範例：

```
jpeg("c:/documents/pieplot.jpg")
```

```
pie(x)
```

```
dev.off() #輸入dev.off()後才會存檔
```

設計遵守原則



1 | DO USE ONE COLOR TO REPRESENT EACH CATEGORY.



2 | DO ORDER DATA SETS USING LOGICAL HEIRARCHY.



3 | DO USE CALLOUTS TO HIGHLIGHT IMPORTANT OR INTERESTING INFORMATION.



4 | DO VISUALIZE DATA IN A WAY THAT IS EASY FOR READERS TO COMPARE VALUES.



5 | DO USE ICONS TO ENHANCE COMPREHENSION AND REDUCE UNNECESSARY LABELING.



6 | DON'T USE HIGH CONTRAST COLOR COMBINATIONS SUCH AS RED/GREEN OR BLUE/YELLOW.



7 | DON'T USE 3D CHARTS. THEY CAN SKEW PERCEPTION OF THE VISUALIZATION.



8 | DON'T ADD CHART JUNK. UNNECESSARY ILLUSTRATIONS, DROP SHADOWS, OR ORNAMENTATIONS DISTRACT FROM THE DATA.



9 | DON'T USE MORE THAN 6 COLORS IN A SINGLE LAYOUT.



10 | DON'T USE DISTRACTING FONTS OR ELEMENTS (SUCH AS BOLD, ITALIC, OR UNDERLINED TEXT).

The background features a light gray hexagonal grid pattern. Overlaid on this is a series of concentric, semi-transparent circles in shades of light blue and white. The circles have a slightly irregular, hand-drawn appearance. A solid dark blue horizontal line runs across the top of the image, and a similar but slightly textured dark blue line runs across the bottom.

THANK YOU