



Exploratory Data Analysis with R

智庫驅動

Wush Wu

dSp
dsp.im

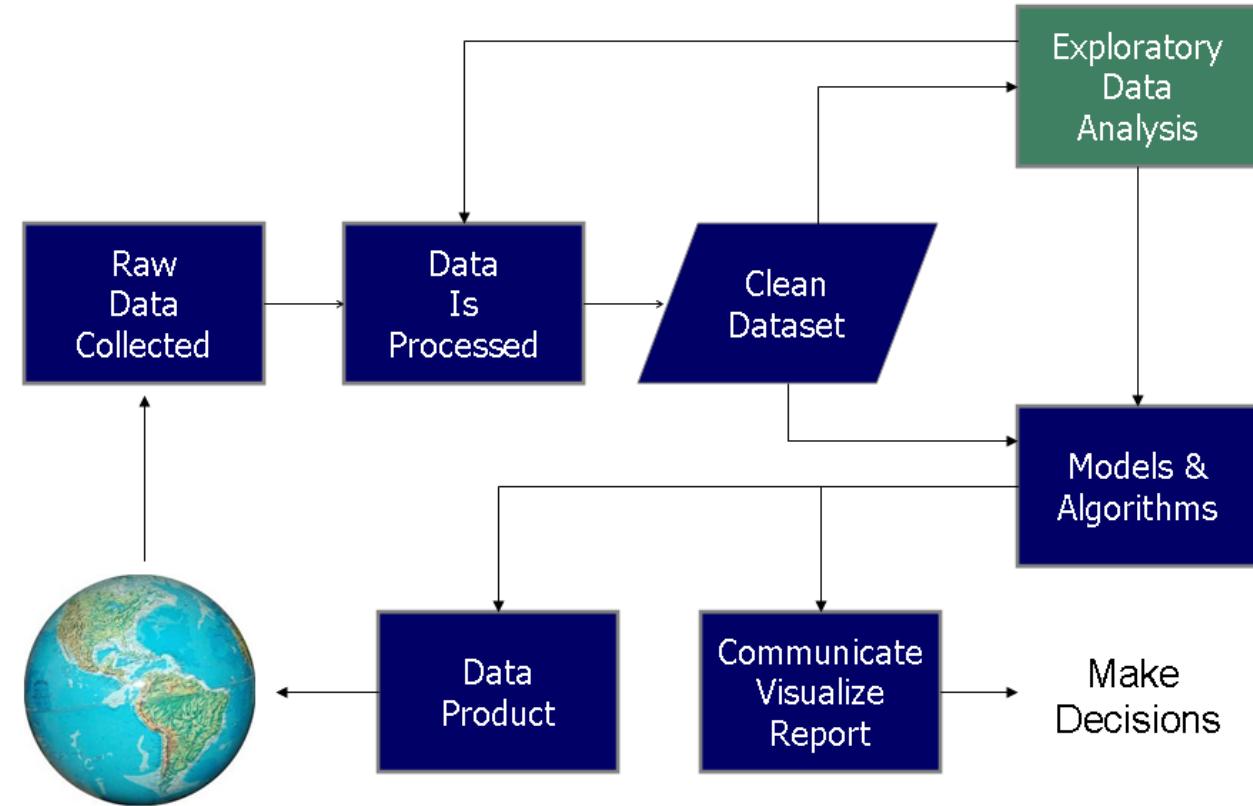
關於講師

Wush Wu

- Taiwan R User Group 共同創辦人
- 以下R 套件的貢獻者：
 - FeatureHashing
 - digest
 - rcppcnpy
 - knitr
- 臺大電機所博士生



Data Science Process



出處：http://en.wikipedia.org/wiki/File:Data_visualization_process_v1.png

大綱

- EDA 的目的
- 如何掌握資料的脈絡
- 觀察數據的方法

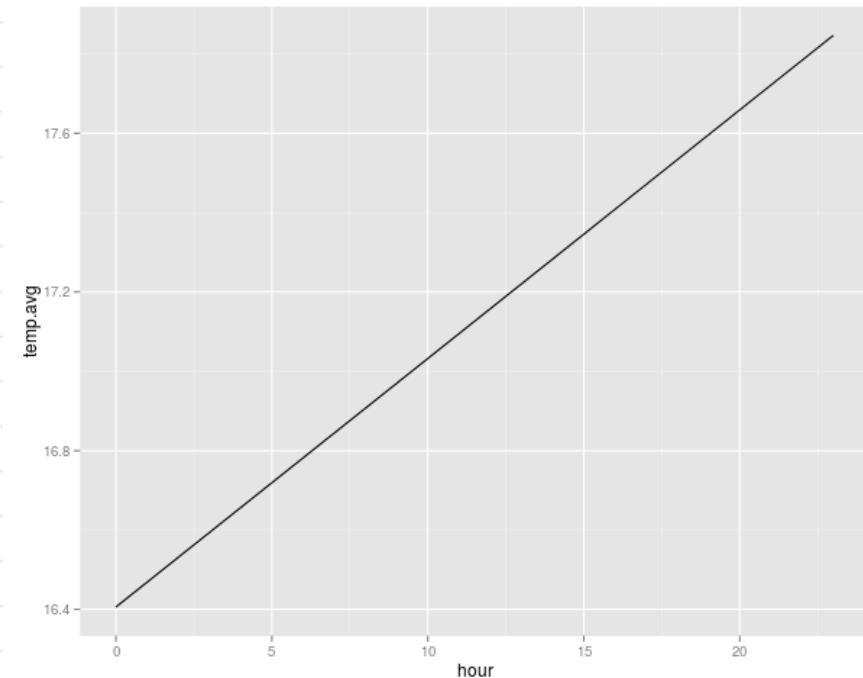
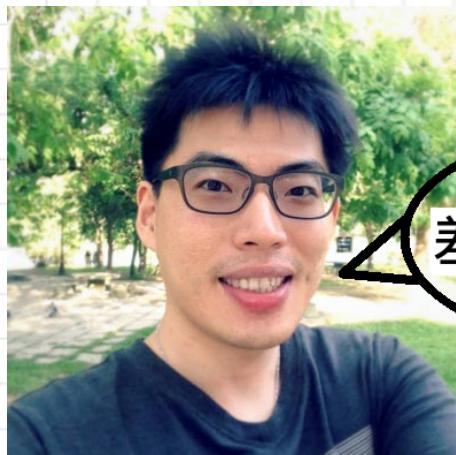
EDA的目的

Tukey, John W. (1977). Exploratory Data Analysis. Pearson. ISBN 978-0201076165.

- 檢查資料的正確性
- 尋找現象中可能的因果關係
- 確認進階分析中的假設是否合理
- 選擇正確的分析工具和技術
- 建議未來的數據收集方向

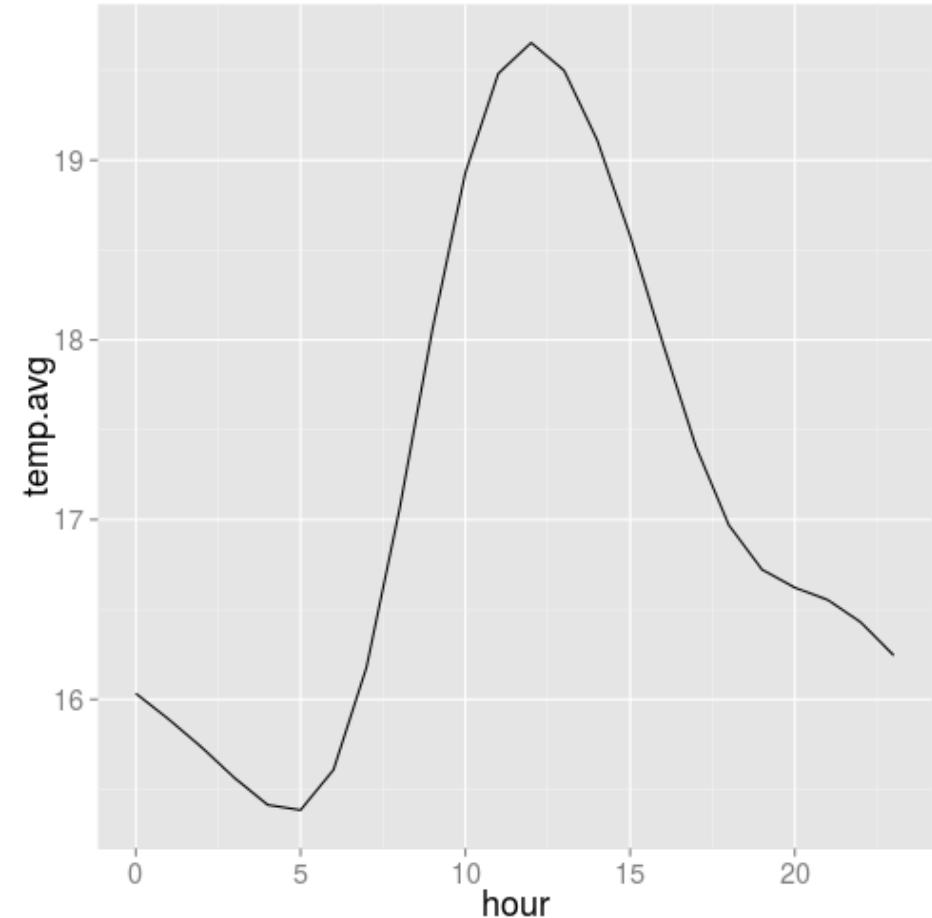
檢查資料的正確性

- 錯誤的資料很常見
 - 輸入錯誤
 - 處理的程式錯誤
 - 對資料的值解釋錯誤



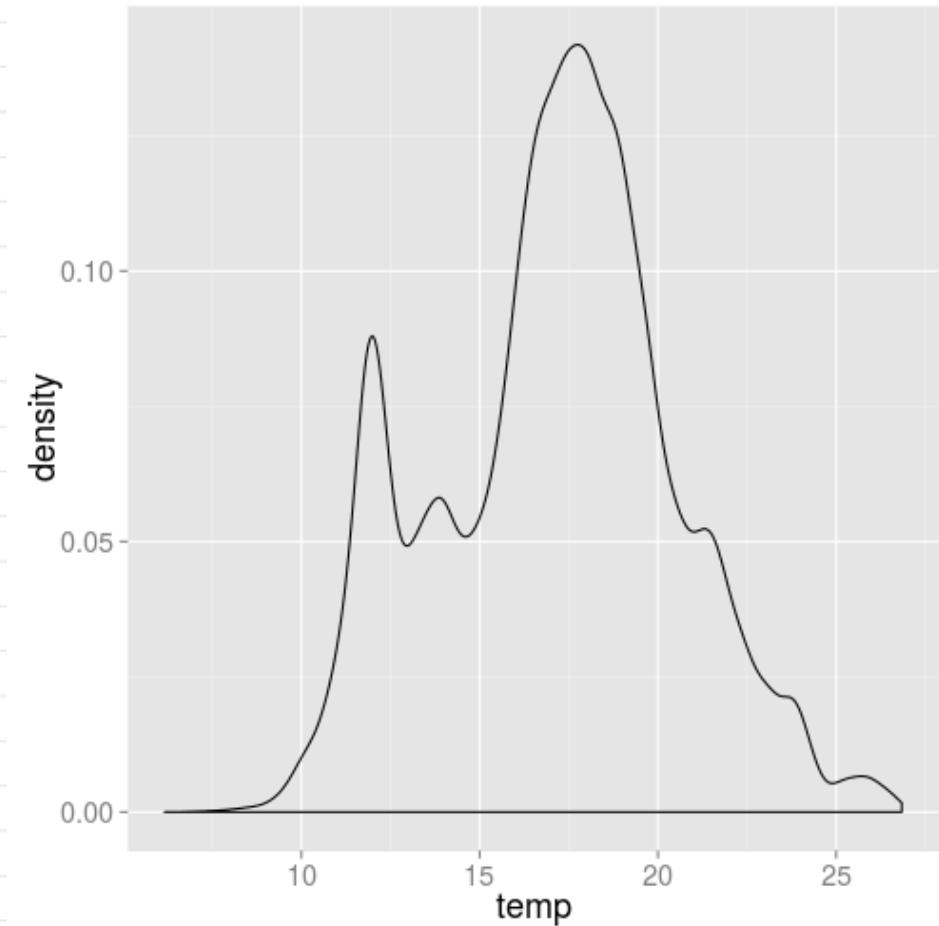
尋找數據中潛在的因果關係

- 因果關係不容易發現
- 需要對問題本質的理解，加上數據的佐證



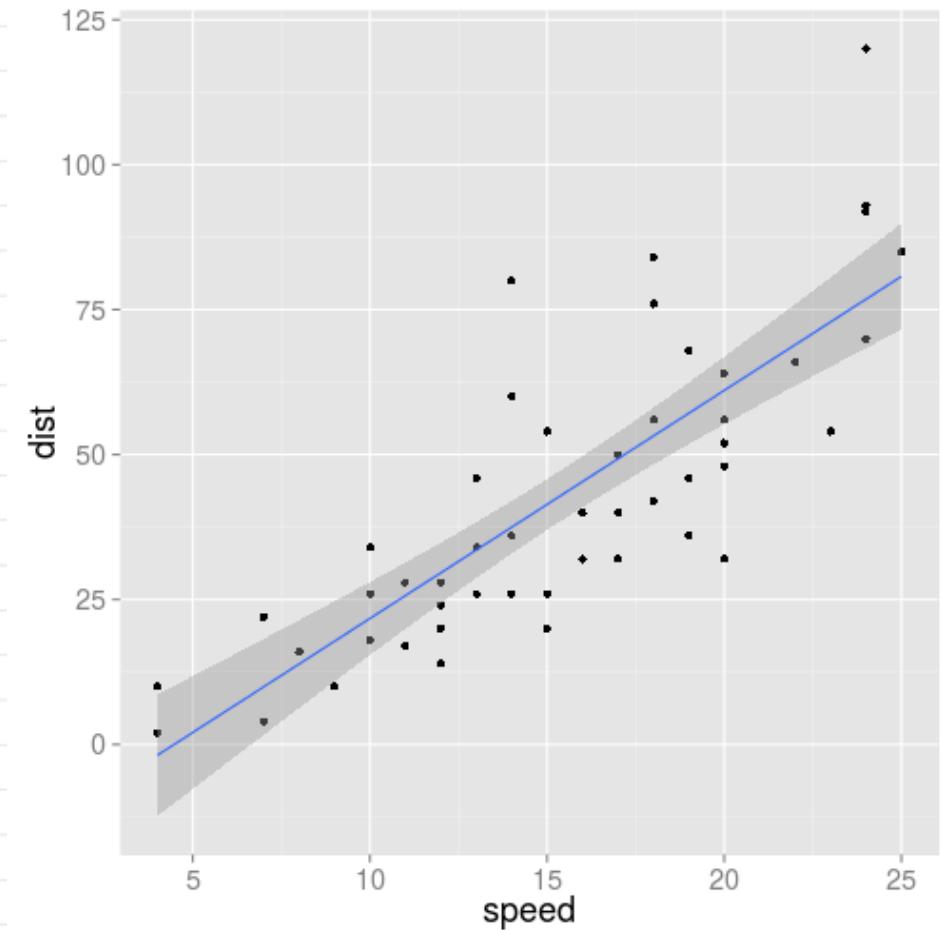
確認進階分析中的假設是否合理

- 可以用常態分佈來描述這個數據嗎？

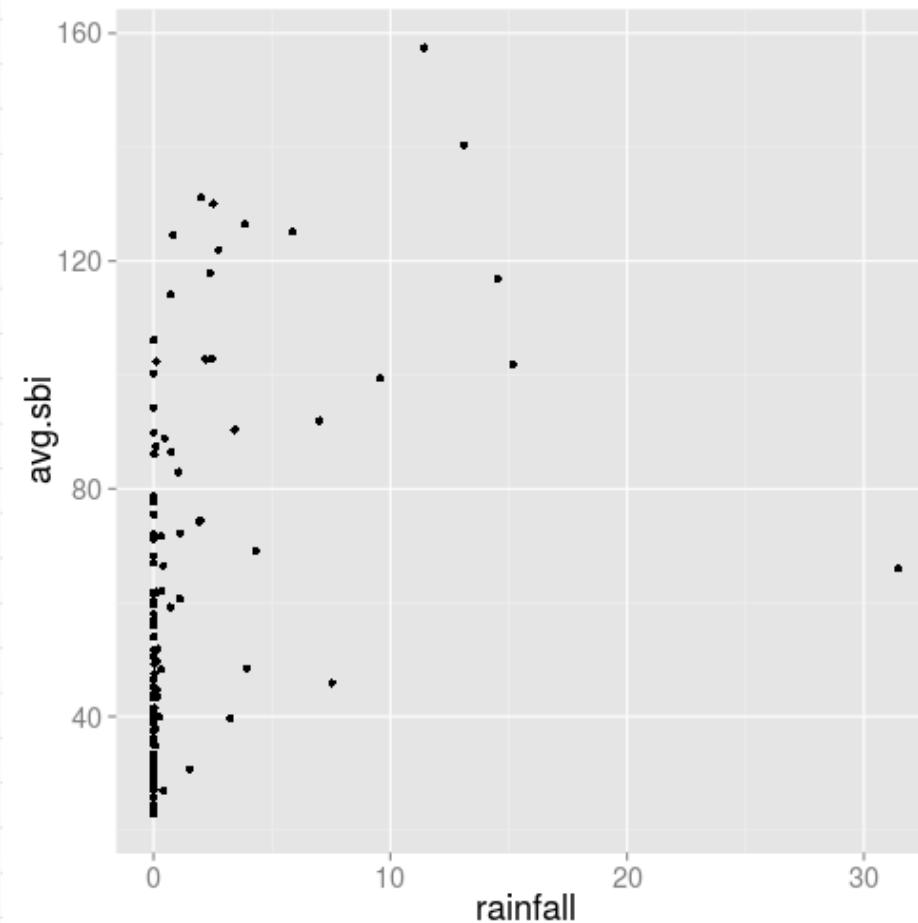


選擇正確的分析工具和技術

- 是線性關係，還是非線性關係呢？



建議未來的數據收集方向和分析方向



如何掌握資料的脈絡

如何掌握資料的脈絡

- 了解資料的格式與物理意義
- 了解資料的欄位與對應的類型
- 檢查和整理資料

了解資料的格式與物理意義

Kaggle

seasons.csv

This file identifies the different seasons included in the historical data, along with certain season-level properties.

- "season" - indicates the year in which the tournament was played
- "dayzero" - tells you the date corresponding to daynum=0 during that season. All game dates have been aligned upon a common scale so that the championship game of the final tournament is on daynum=154. Working backward, the national semifinals are always on daynum=152, the "play-in" games are on days 134/135, Selection Sunday is on day 132, and so on. All game data includes the day number in order to make it easier to perform date calculations. If you really want to know the exact date a game was played on, you can combine the game's "daynum" with the season's "dayzero". For instance, since day zero during the 2011-2012 season was 10/31/2011, if we know that the earliest regular season games that year were played on daynum=7, they were therefore played on 11/07/2011.

了解資料的格式與物理意義

Ubike

DATE	HOUR	SAREA	SNA	LAT	LNG
2014-12-08	15	信義區	捷運市政府站(3號出口)	25.04	121.57
2014-12-08	15	大安區	捷運國父紀念館站(2號出口)	25.04	121.56
2014-12-08	15	信義區	台北市政府	25.04	121.57
2014-12-08	15	信義區	市民廣場	25.04	121.56
2014-12-08	15	信義區	興雅國中	25.04	121.57
2014-12-08	15	信義區	世貿二館	25.03	121.57

了解資料的欄位與對應的類型

- 數值型變數
- 類別型變數
- 標籤型變數

了解資料的欄位與對應的類型 - 數值型變數

- 一定是以數字表示
- 可以做加法和減法的運算
- 一定擁有相對的大小關係
- 範例
 - 氣溫
 - 時間
 - 價格
 - 數量

了解資料的欄位與對應的類型 - 類別型變數

- 可能是以數字表示
- 不能做加法和減法的運算
- 不一定擁有相對的大小關係
- 範例
 - 日期、月份
 - 性別
 - 數值區間，如年齡區間

了解資料的欄位與對應的類型 - 標籤型變數

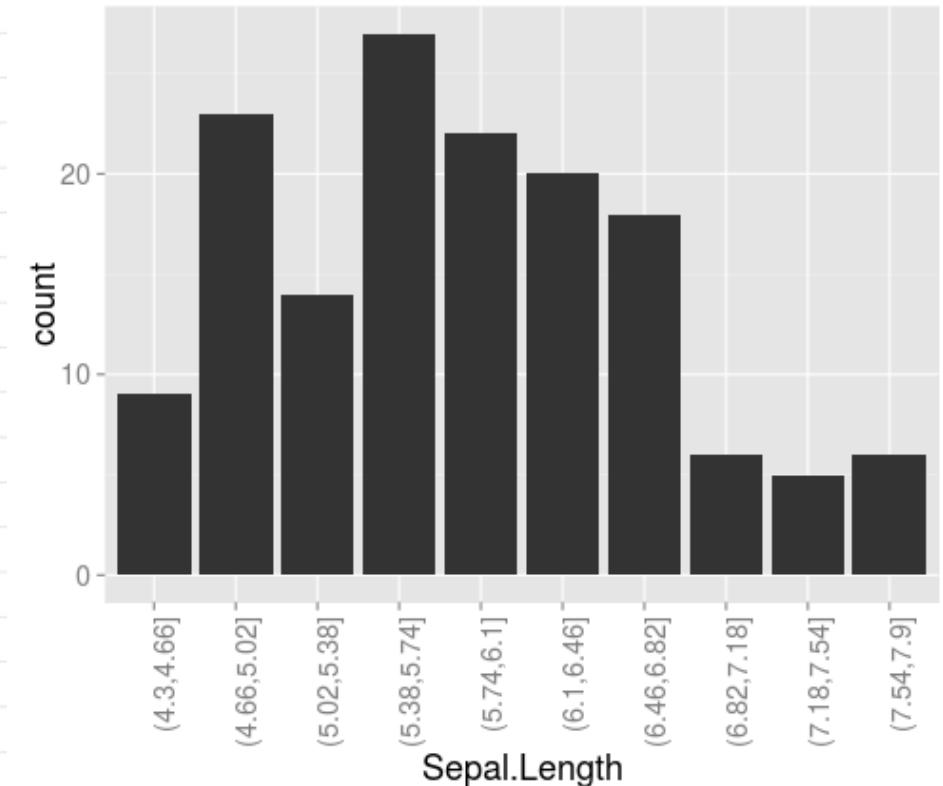
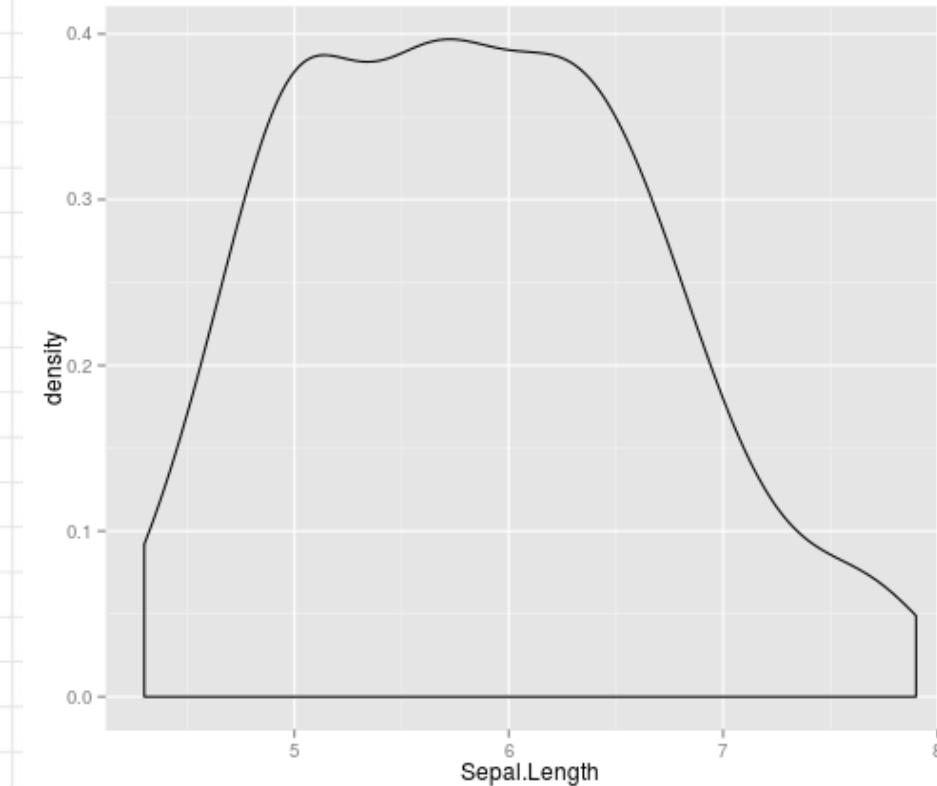
- 通常是類別型或數值型變數的壓縮
- 已經被定義在某些資料庫系統之中
- 轉換為對應的類別型變數與數值型變數
- 範例
 - 文章分類標籤
 - 臉書標籤



出處: <http://www.techbang.com/posts/13673-import-facebook-official-hashtag-label-features-marked-with-a-keywords>

數值型變數可以轉換為類別型變數

- Data Binning



檢查和整理資料

- 檢查資料是否正確
- 檢查資料內容和物理意義的一致性

觀察數據的方法

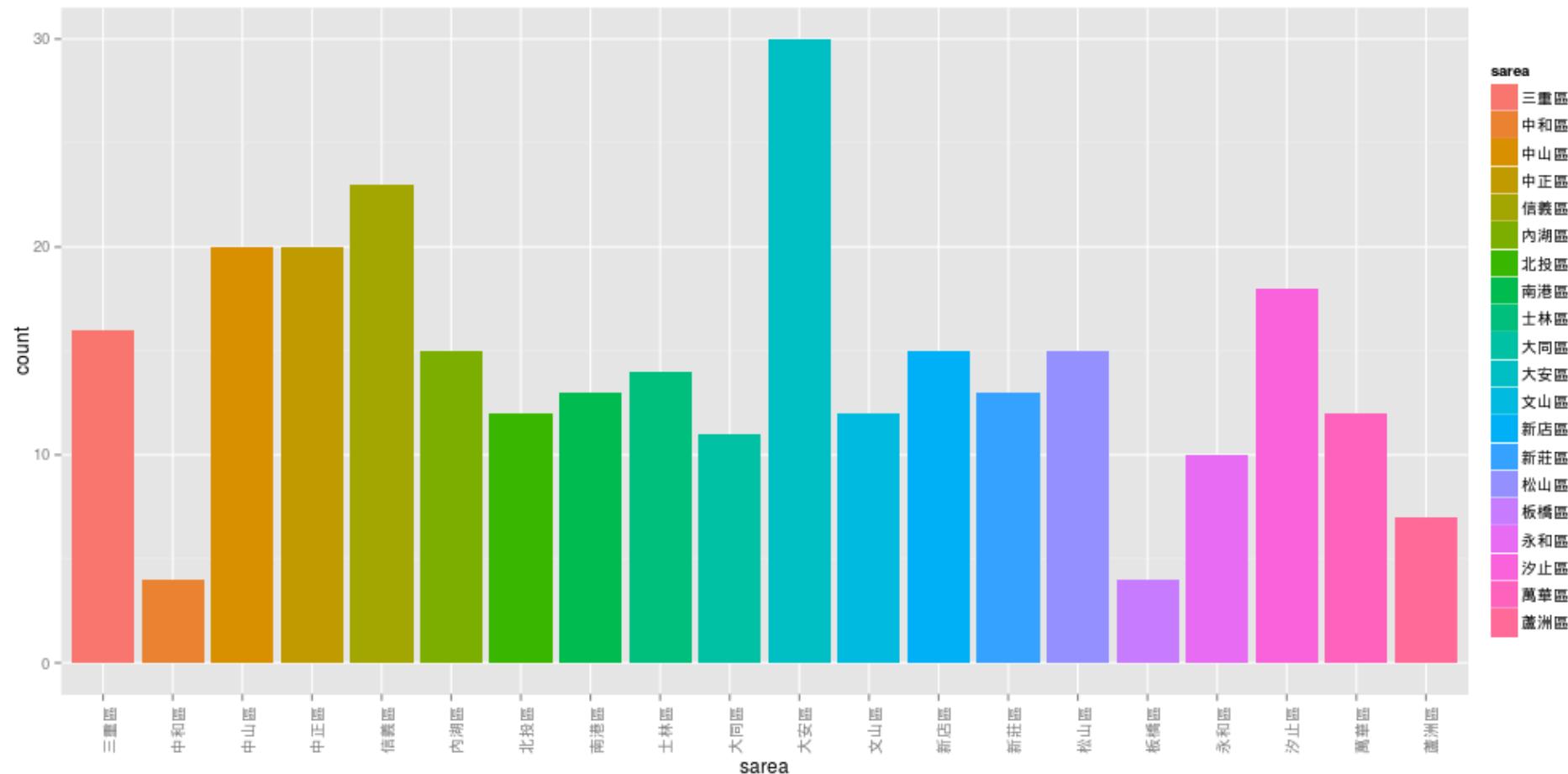
觀察單一欄位的資料

- 數值型變數
- 類別型變數

類別型變數 - 各類別的值與分佈

三重區	16	內湖區	15	大安區	30	板橋區	4
中和區	4	北投區	12	文山區	12	永和區	10
中山區	20	南港區	13	新店區	15	汐止區	18
中正區	20	士林區	14	新莊區	13	萬華區	12
信義區	23	大同區	11	松山區	15	蘆洲區	7

類別型變數 - 各類別的值與分佈

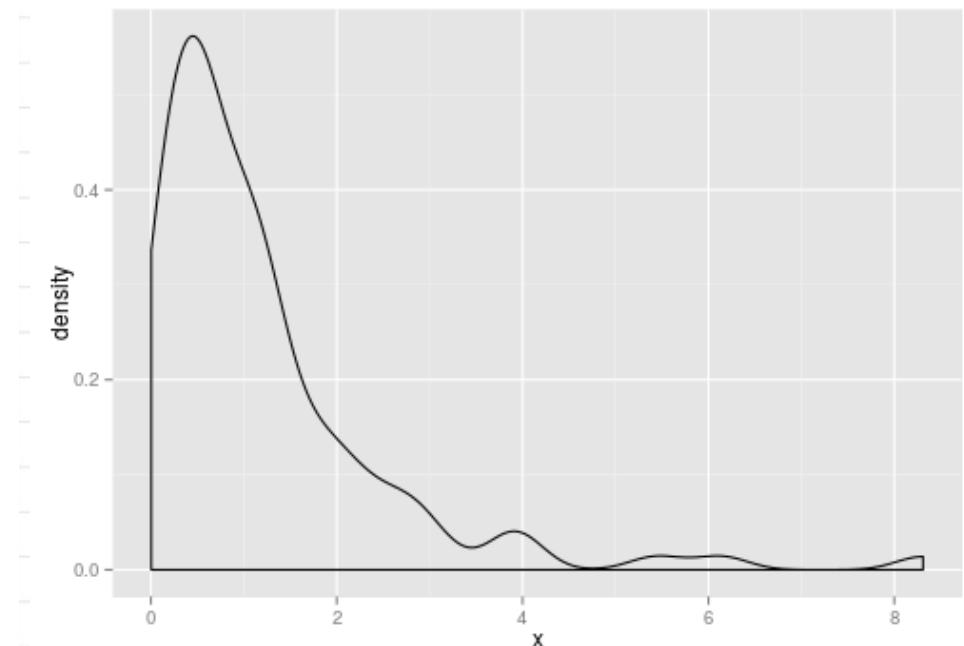
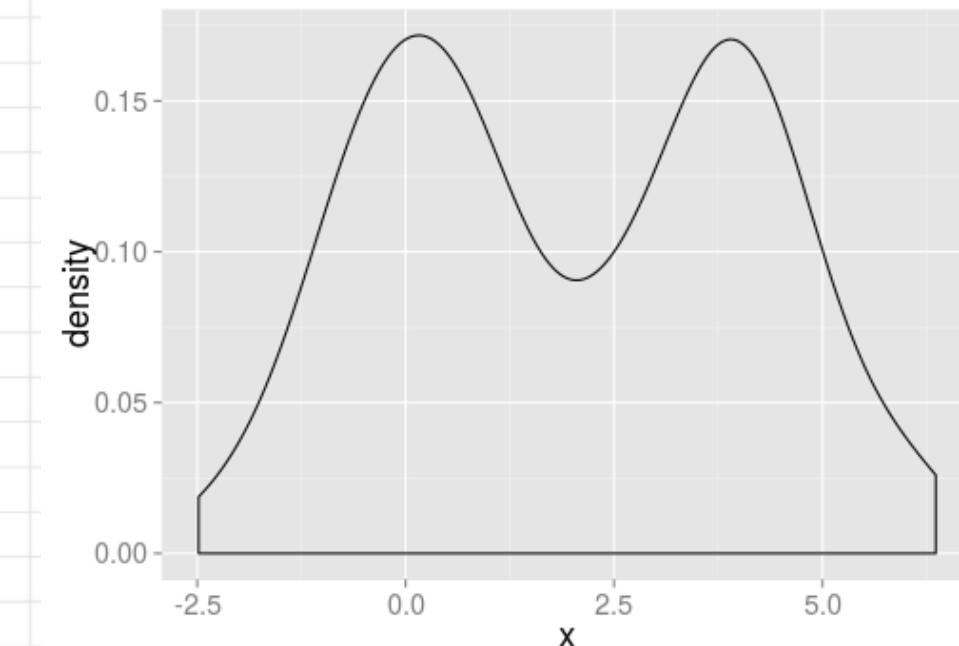


單一欄位的資料 - 數值型變數

- 數值分佈
- 中心位置
- 分散程度

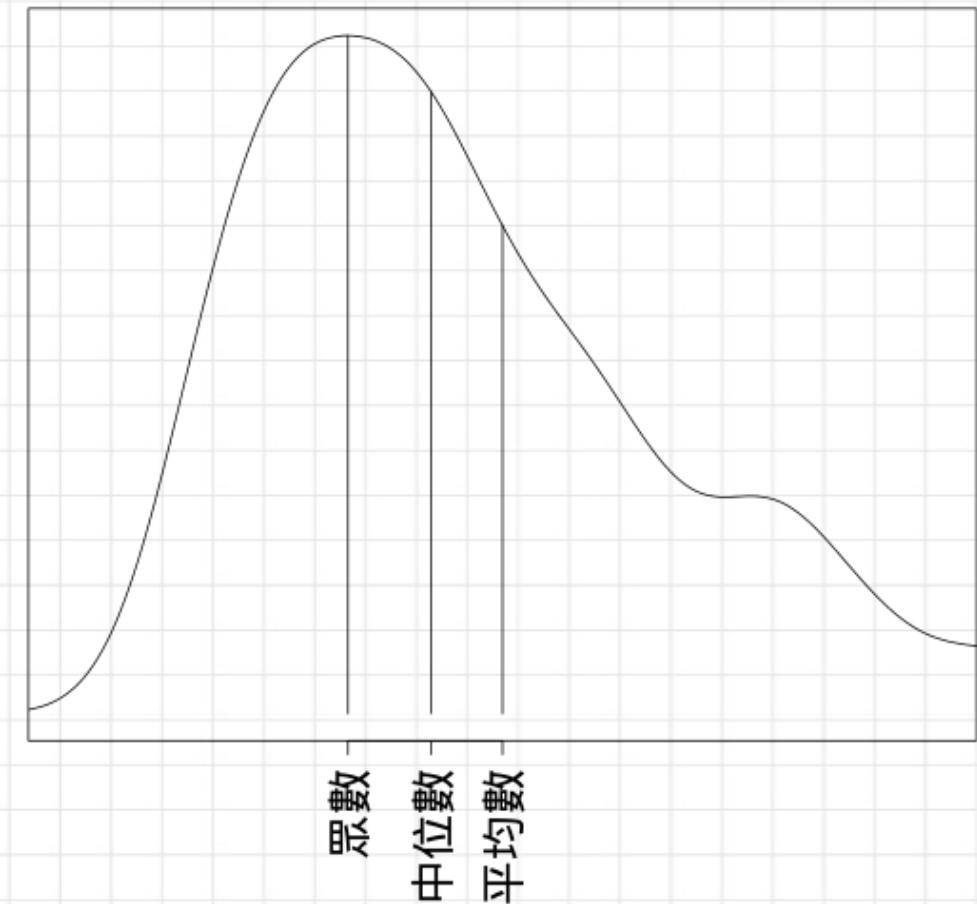
單一欄位的資料 - 數值分佈

- 單峰或雙峰
- 左偏或右偏



單一欄位的資料 - 中心位置

- 平均數 : $\min \sum (x_i - m)^2$
- 中位數 : $\min \sum |x_i - m|$
- 眾數

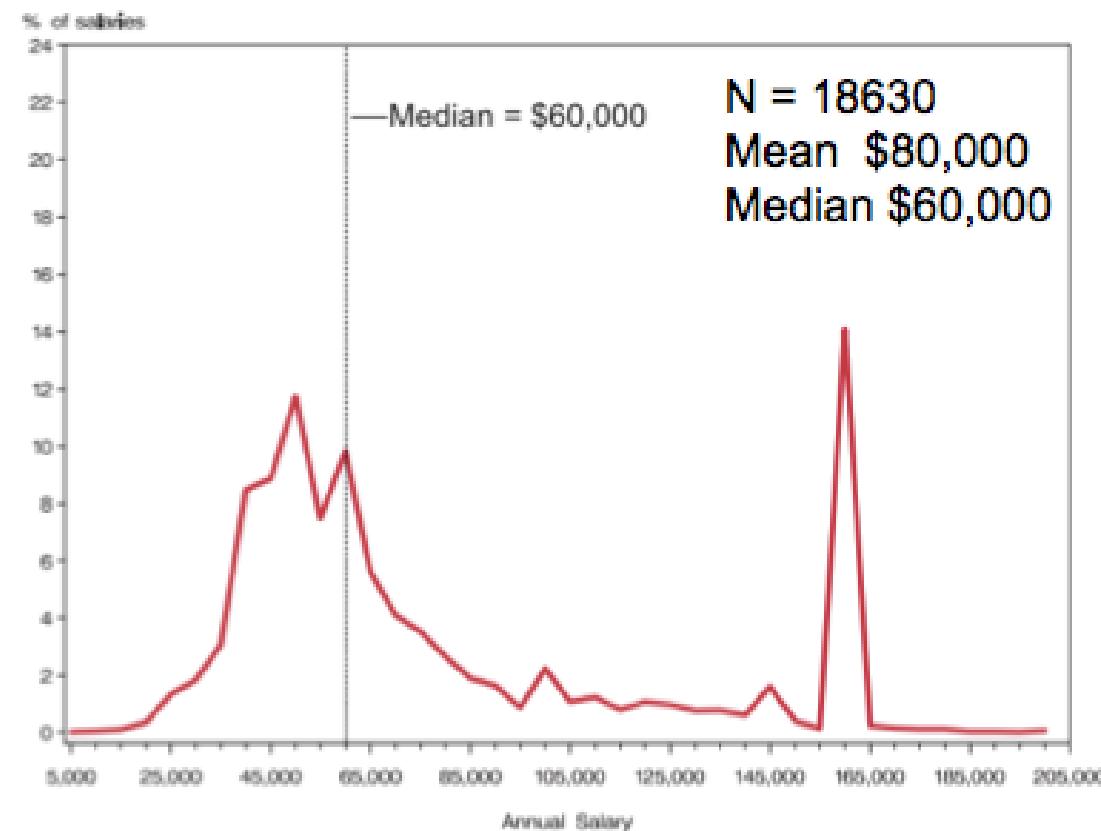


101年中華民國家庭收支統計

- 每戶可支配所得
 - 平均為92.4萬元
 - 中位數為80.8萬元

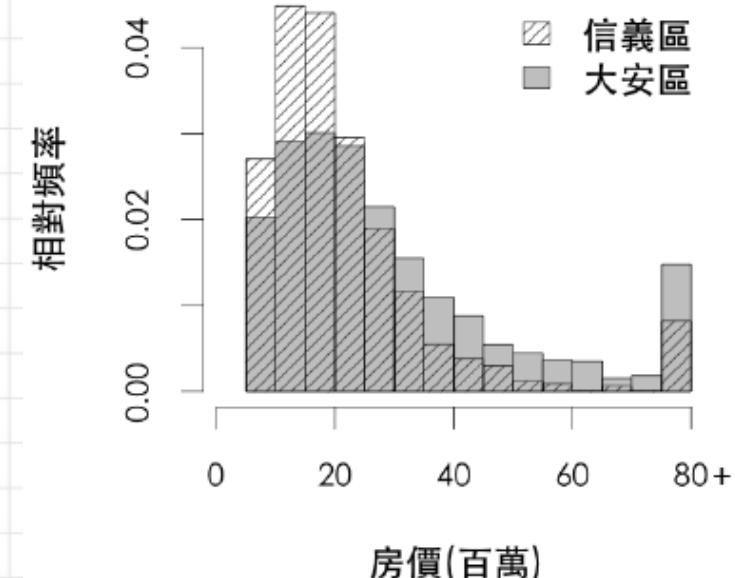
美國新進律師薪水分佈

Distribution of Full-time Salaries — Class of 2011



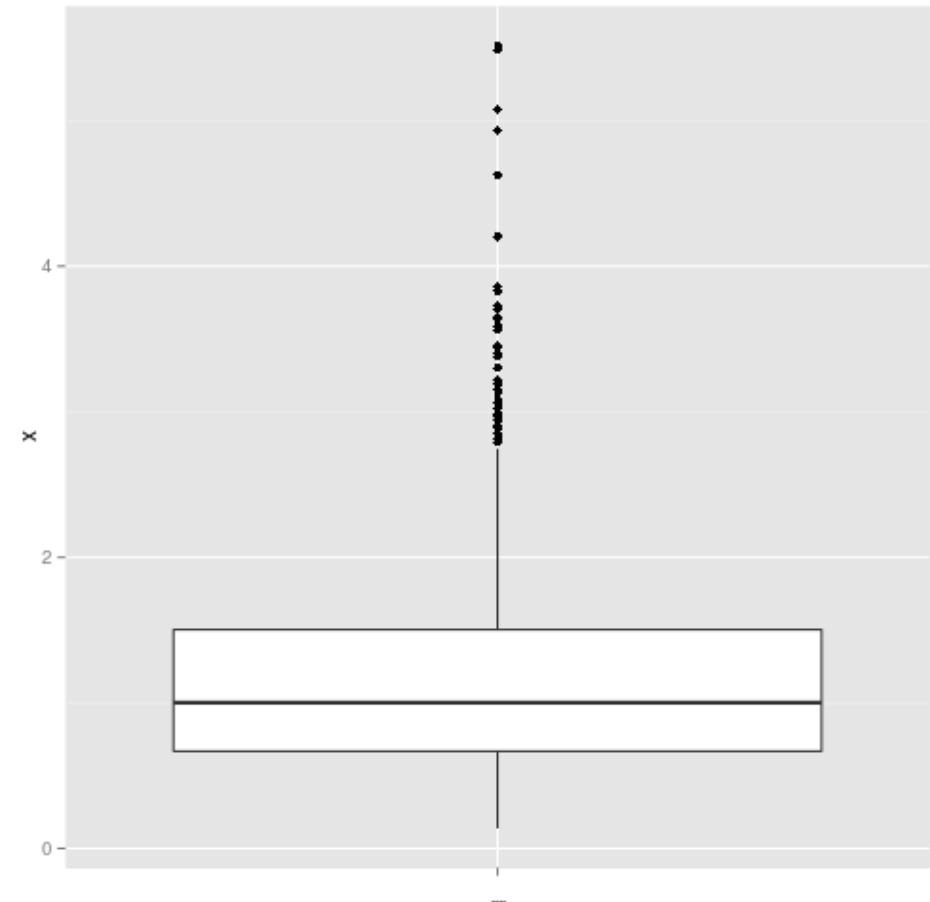
信義區與大安區住宅價格比較

房價	信義區	大安區
1 25%	12.50	14.80
2 50%	18.00	23.40
3 平均	25.20	31.50
4 75%	25.60	37.40



單一欄位的資料 - 分散程度

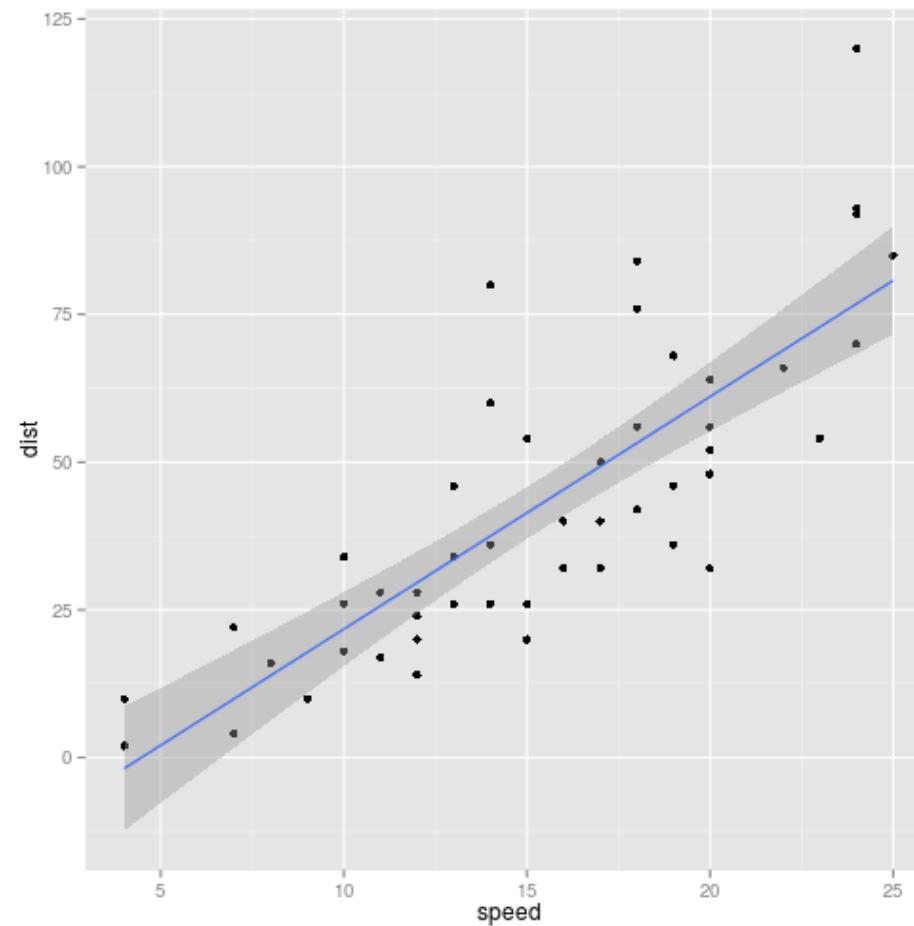
- 標準差 -- 平均數的最小化目標
- 離差 -- 中位數的最小化目標
- 四分位差
- 變異係數



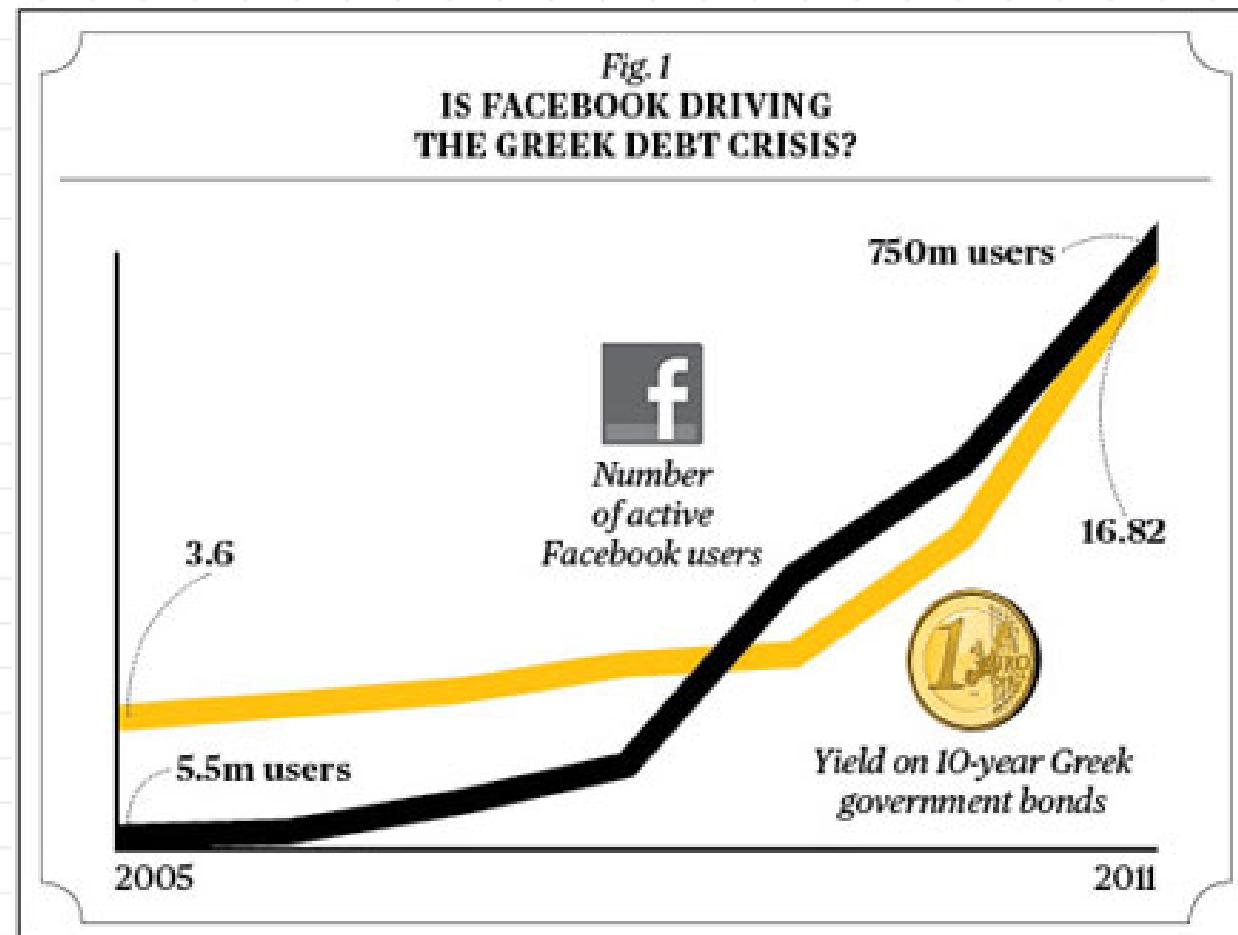
觀察兩個欄位的資料 - 相關性

- 有因果關係一定有相關性
- 相關性不一定代表有因果關係

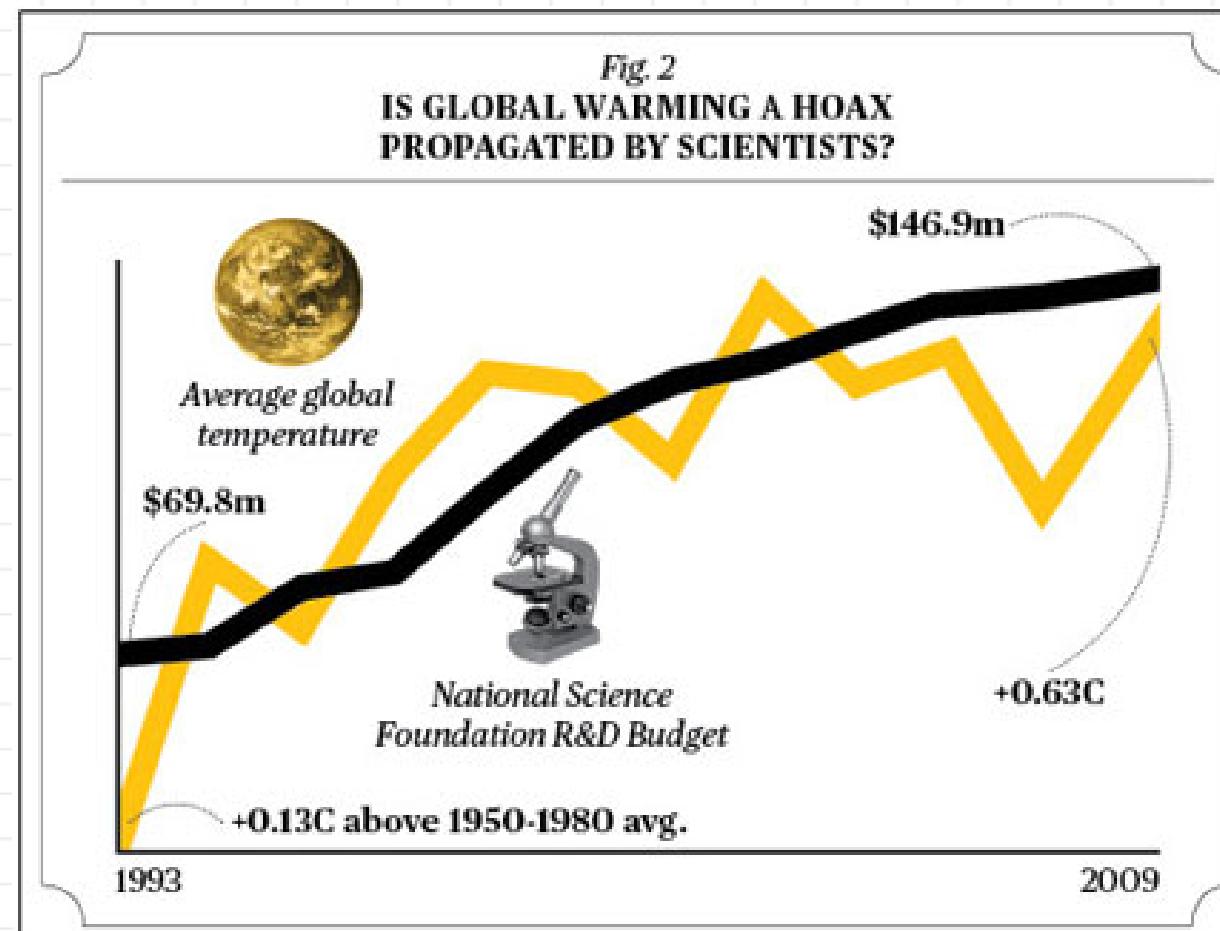
車速和煞車的關係



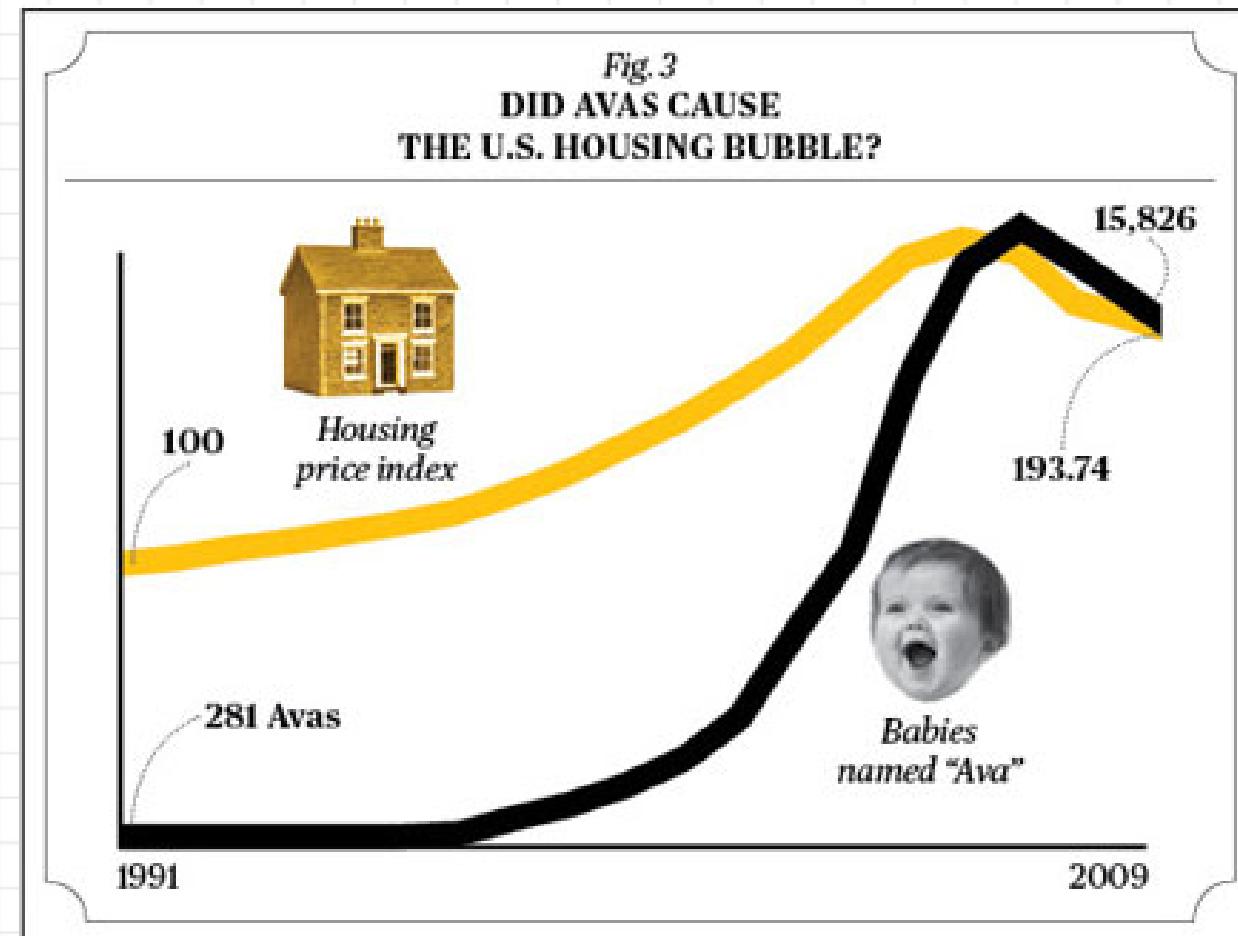
相關性還是因果關係？



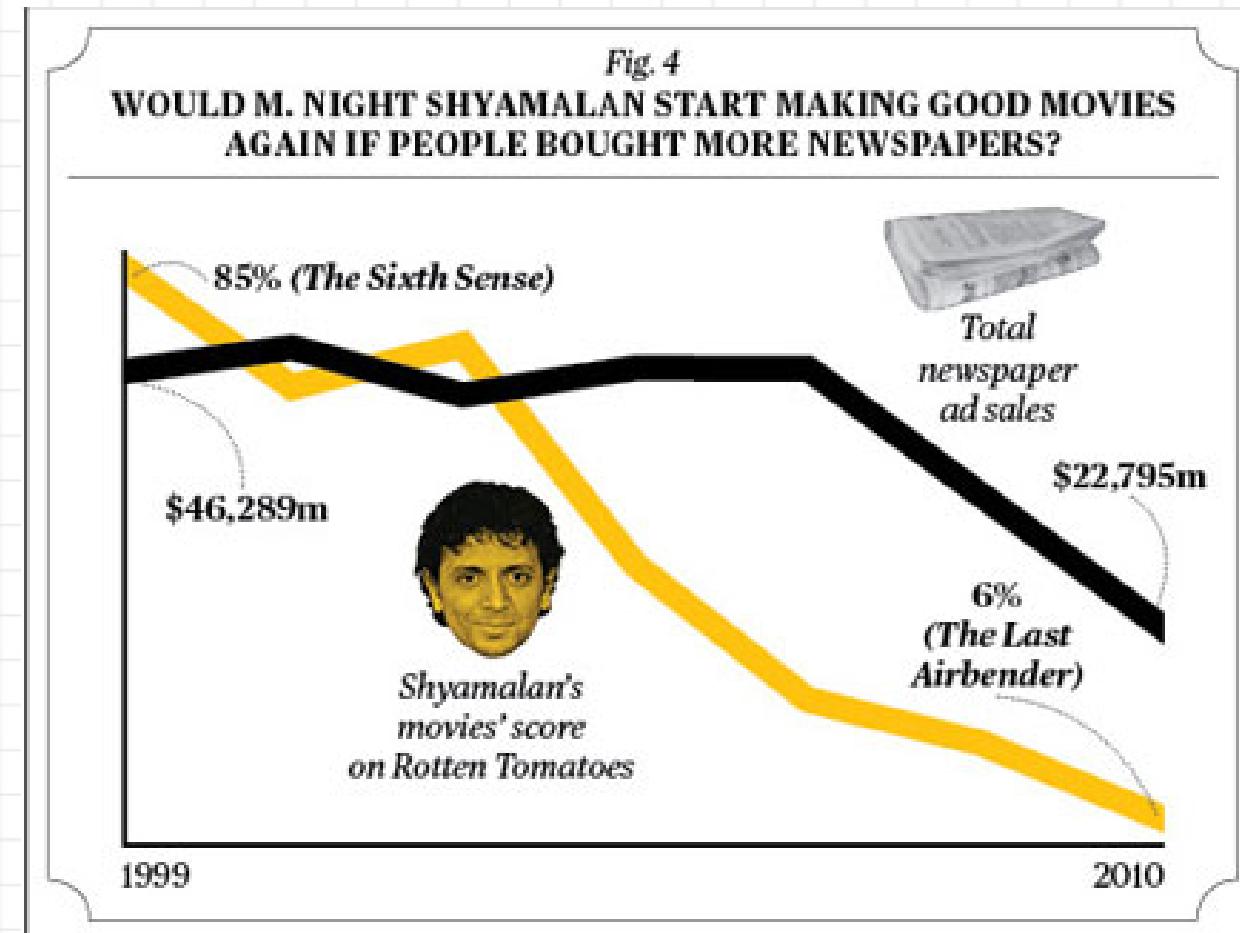
相關性還是因果關係？



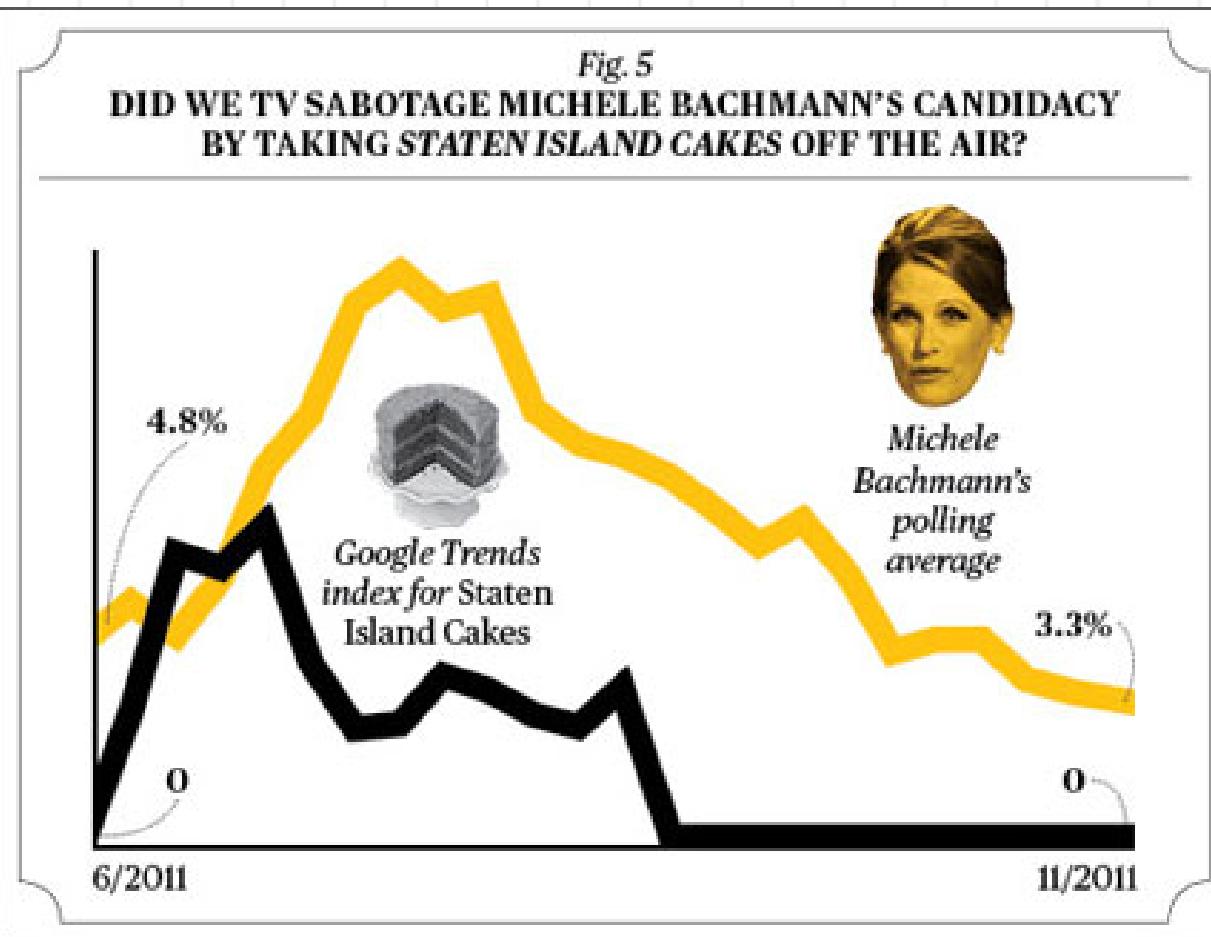
相關性還是因果關係？



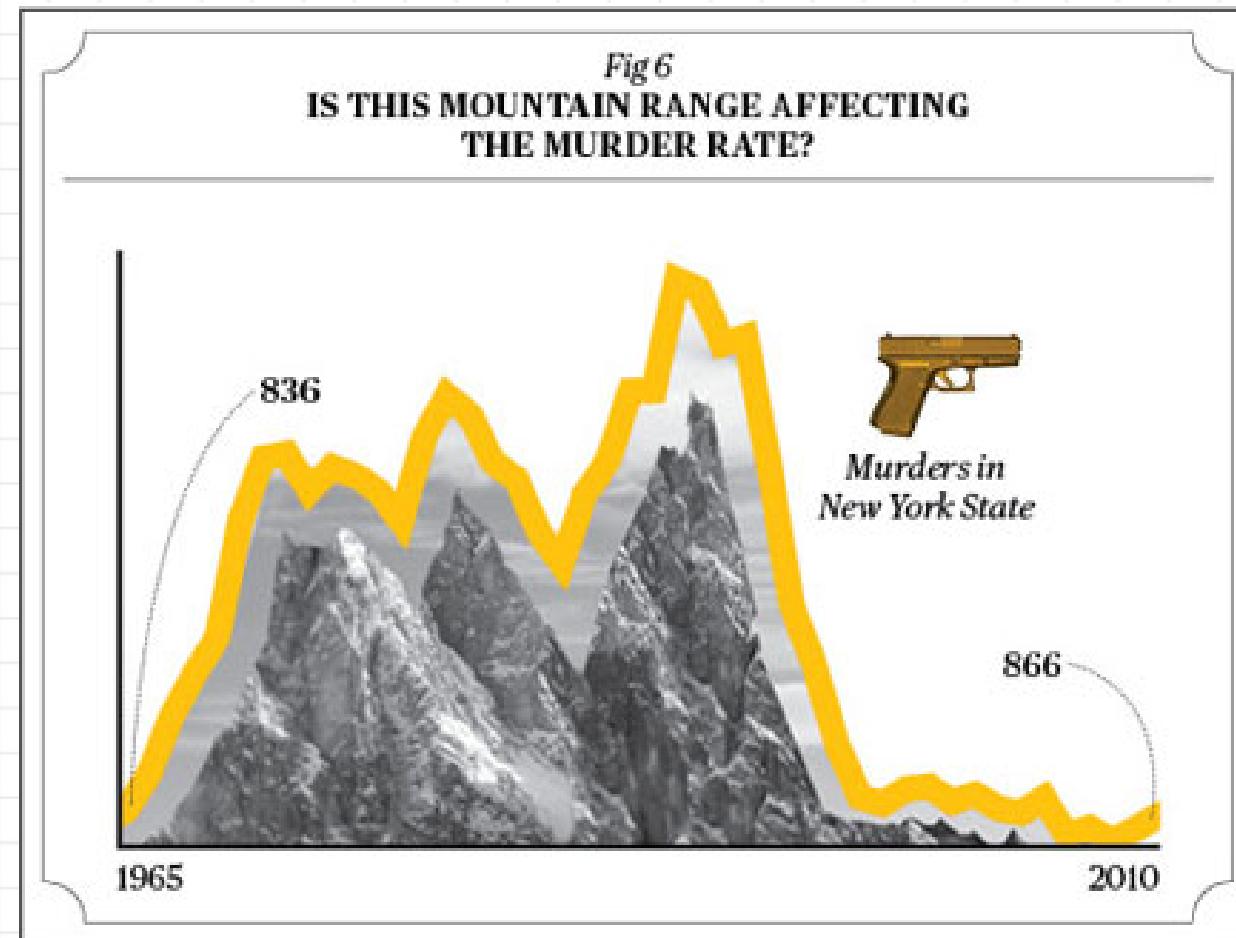
相關性還是因果關係？



相關性還是因果關係？



相關性還是因果關係？



預測和相關性

- 只有單一數據時，基本的預測方式：
 - 類別型變數：猜出現最多次的類別（眾數）
 - 數值型變數：猜中心點
- 只用一月份的平均氣溫猜測二月份的氣溫，平均誤差的平方為：11.667782
- 考慮每小時氣溫的變化之後，平均誤差的平方變成：9.736795

觀察兩個欄位的資料 - 類別 vs 類別

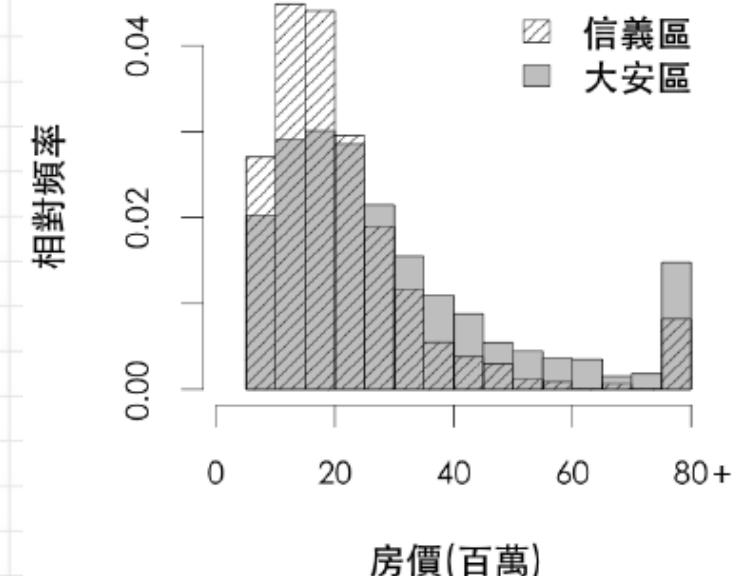
列聯表

SURVIVED	COUNT
No	1490
Yes	711

	"Sex"	"Male"	"Female"
"Survived"			
"No"	1364	126	
"Yes"	367	344	

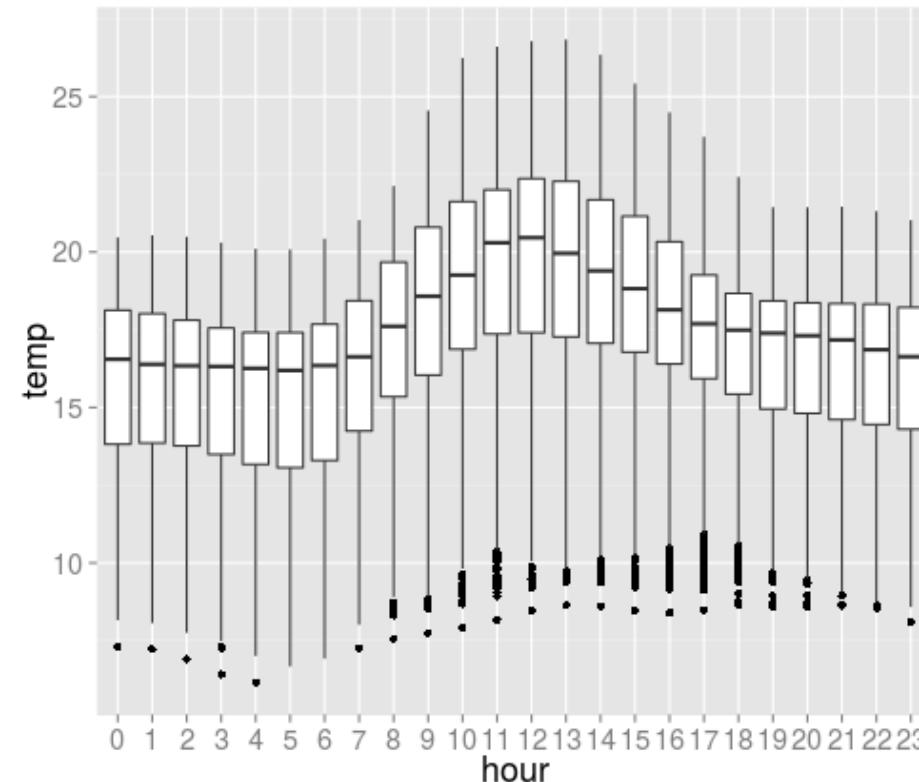
觀察兩個欄位的資料 - 類別 vs 數值

	房價	信義區	大安區
1	25%	12.50	14.80
2	50%	18.00	23.40
3	平均	25.20	31.50
4	75%	25.60	37.40



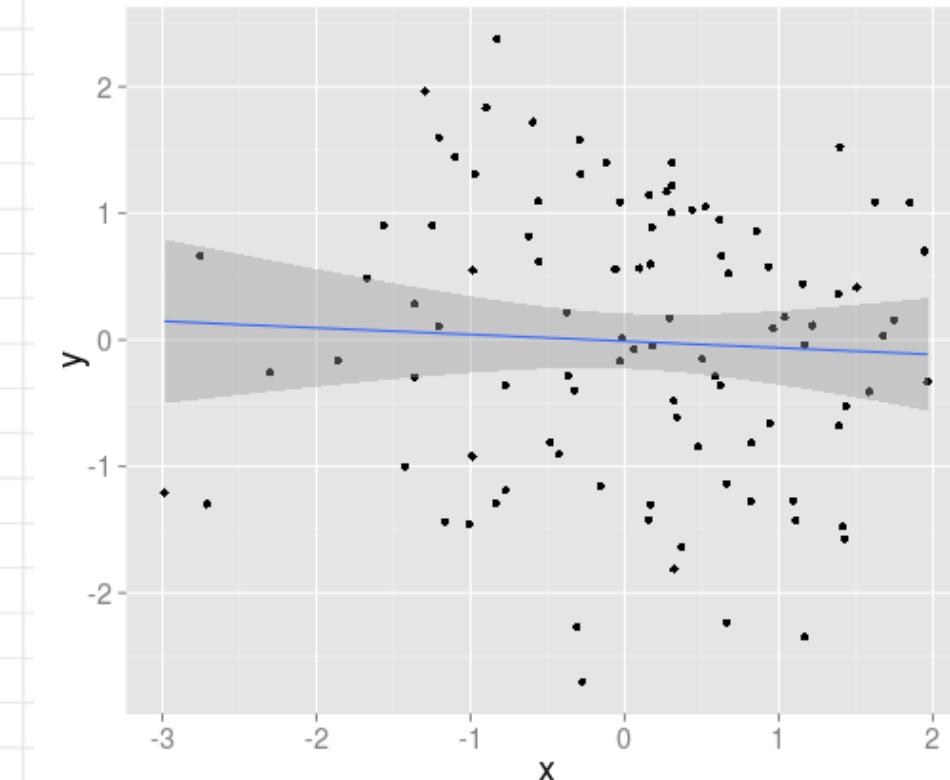
觀察兩個欄位的資料 - 類別 vs 數值

BoxPlot

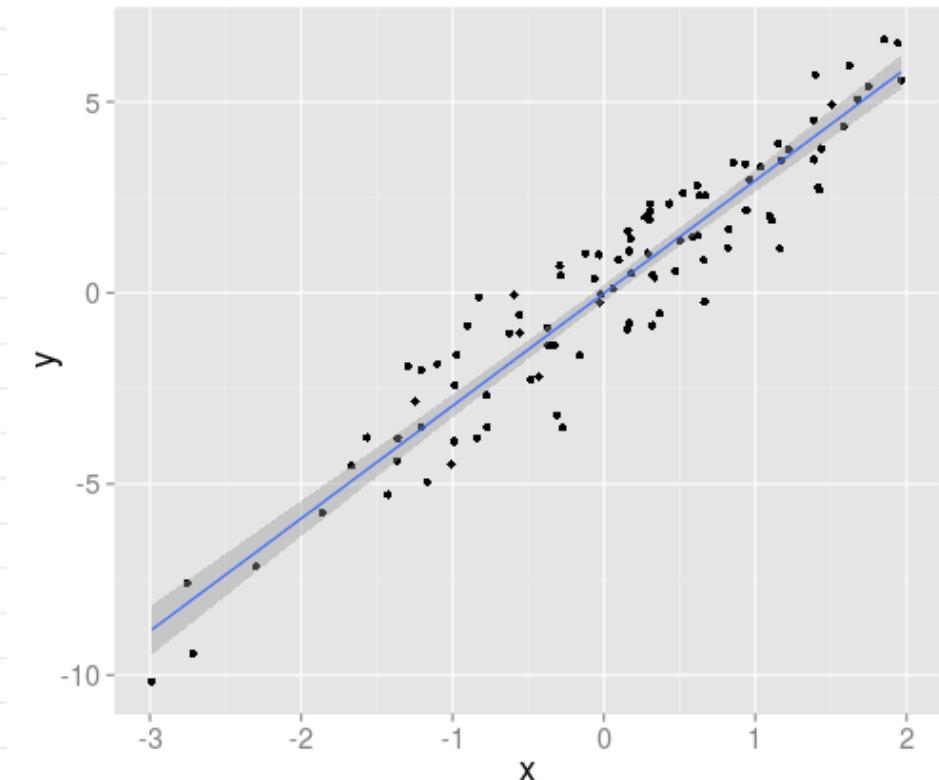


觀察兩個欄位的資料 - 數值 vs 數值

相關係數



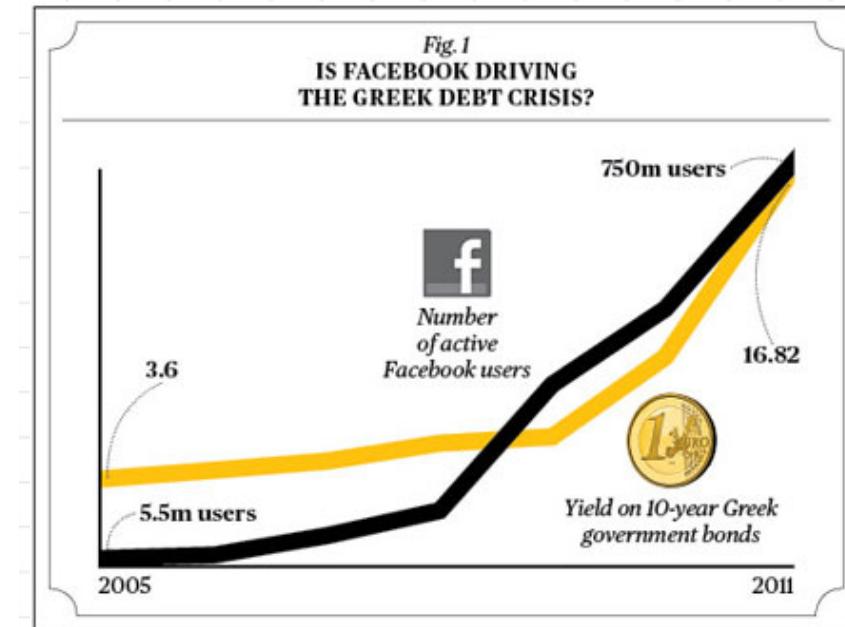
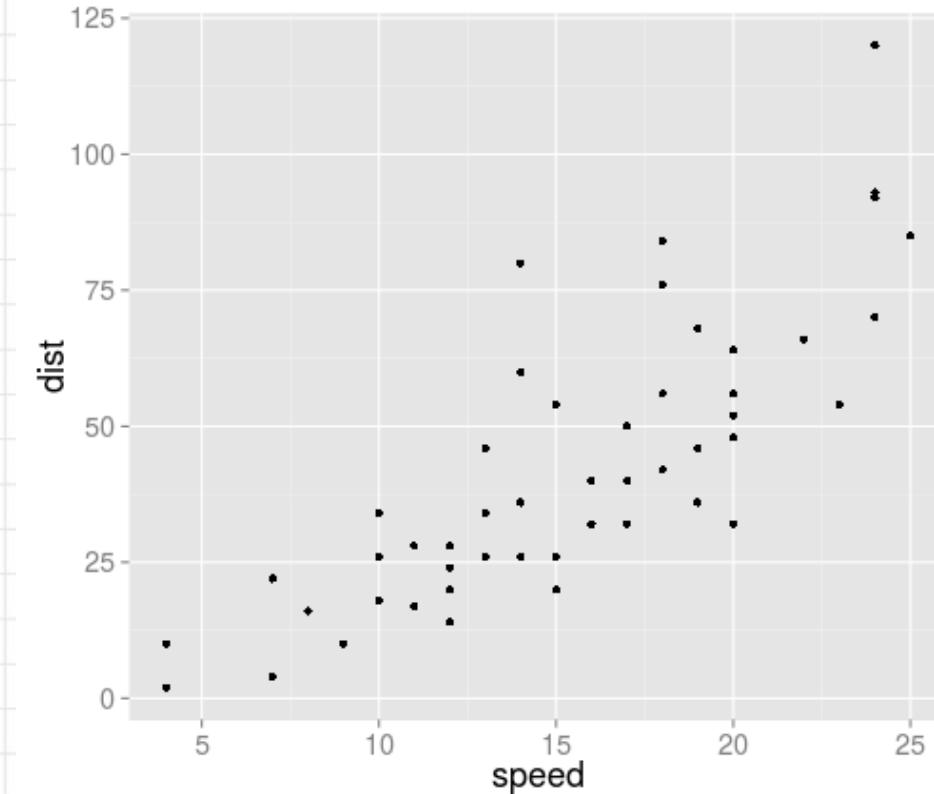
相關係數 : -0.0523



相關係數 : 0.9466

觀察兩個欄位的資料 - 數值 vs 數值

X-Y散佈圖



總結

單一數據

- 類別型數據
 - 類別分佈
 - 表格
 - barplot
- 數值型數據
 - 數值分佈
 - density
 - Boxplot
 - 中心位置
 - 分散程度

總結

雙數據 - 檢視相關性

- 類別 vs 類別
 - 列聯表
- 類別 vs 數值
 - 條件分佈
 - 條件Boxplot
- 數值 vs 數值
 - 相關係數
 - X-Y 散部圖
 - 時間變化圖

觀察數據的工具：R

R來自世界上最專業的統計學家



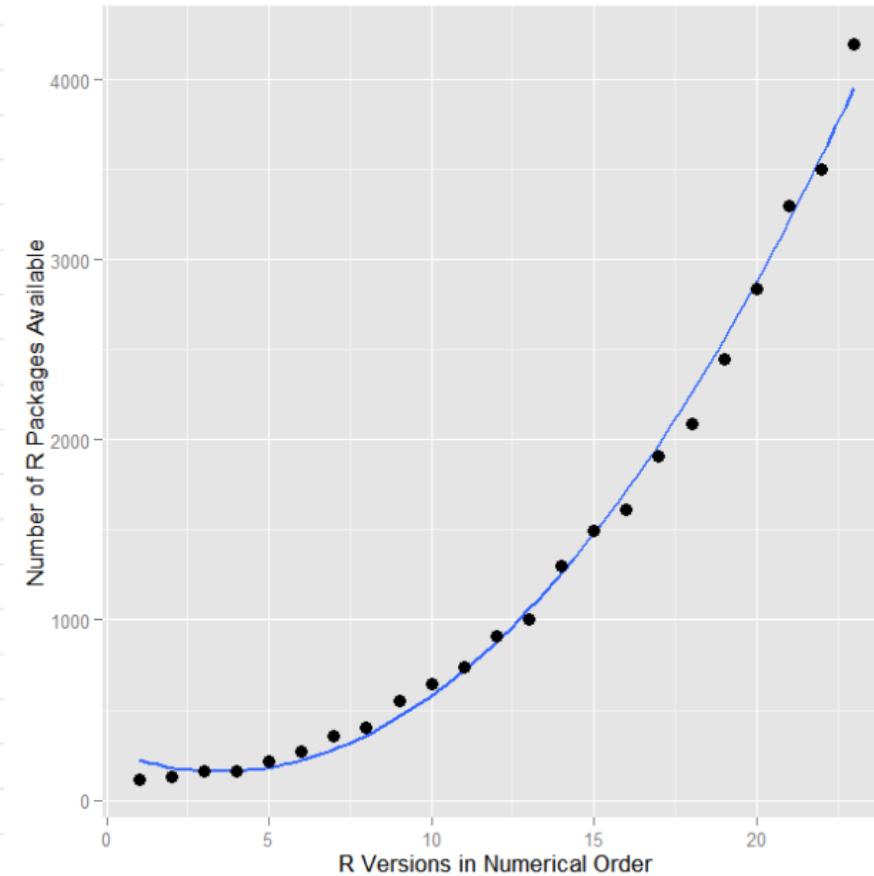
圖片來源：<http://myfootpath.com/careers/engineering-careers/statistician-careers/>

R 可以輸出高品質的視覺化



取自<http://www.r-bloggers.com/mapping-the-worlds-biggest-airlines/>

R 有驚人彈性和潛力



取自 <http://r4stats.com/2013/03/19/r-2012-growth-exceeds-sas-all-time-total/>

R 很容易和其他工具整合

RPostgreSQL
rredis RJDBC
rmongodb
ROpenOffice
RSelenium
Rcpp RODBC
rpy2 rJava
RMySQL
R hadoop

R 很容易擴充和客製化



來源：http://img.diynetwork.com/DIY/2003/09/18/t134_3ca_med.jpg

今天的目標

環境設定

- 建立可以使用R的環境
- 了解R的使用界面