HKSSPC Workshop
Data analysis

Jack CM Wong and Dicky CT Law

The University of Hong Kong

5 April 2021

Learning Objectives 學習目標

1. 了解何為是數據
2. 了解數據的各種類型
3. 了解數據分析的基本概念
4. 了解簡單的統計學方法

What is data? 什麼是數據?

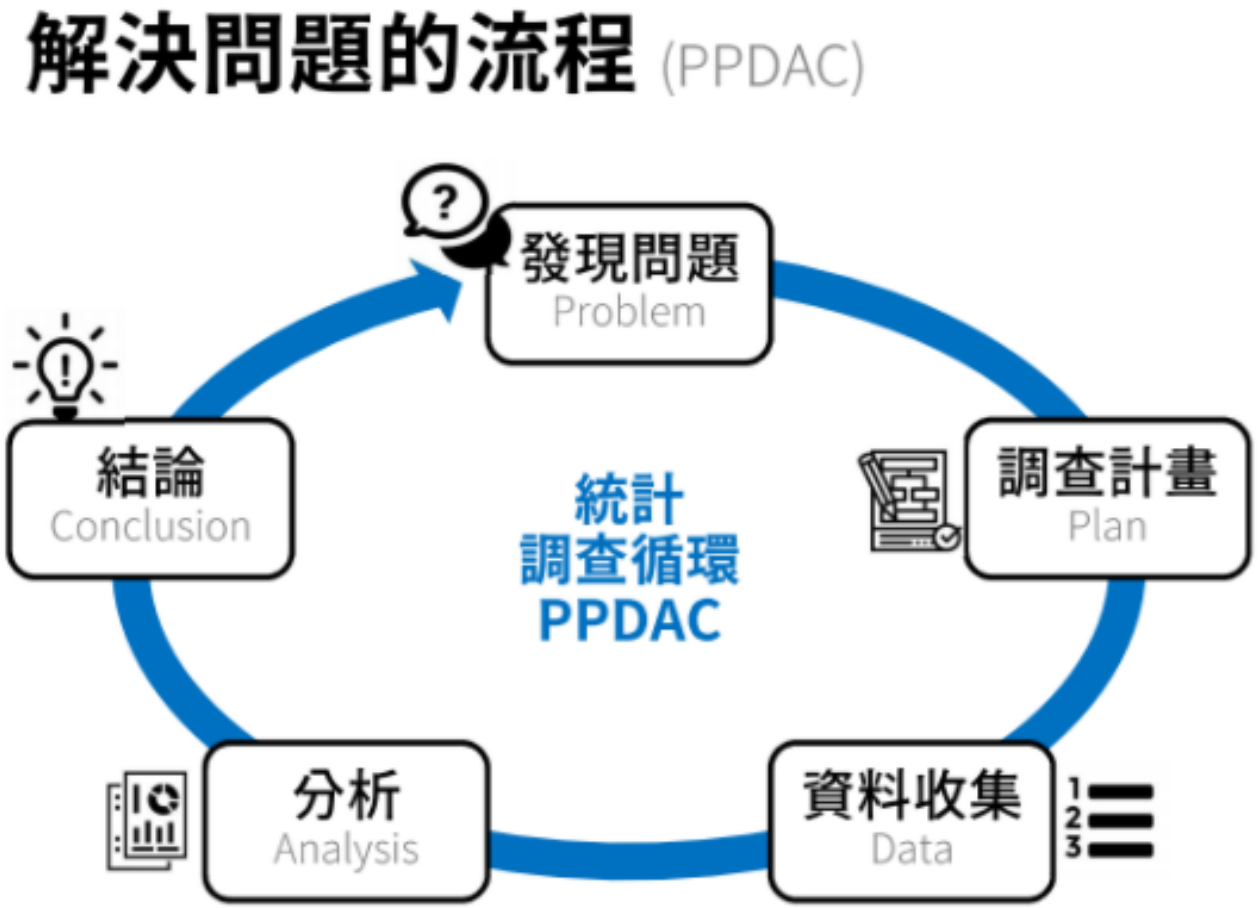
- 數據 (data)是一組關於一個或多個人或物件的定性或定量變數。
- 它可以是一堆雜誌、病歷記錄、民意調查或實驗結果。
- 數據經過測量、收集、報告和分析，可以使用圖表、圖像或者其他分析工具進行視覺化。

Data to knowledge 從數據到知識

- Data: Raw data
- Information: Data organized to convey meaning
- Knowledge: Data organized and processed to convey understanding

- 原始數據 → 信息 → 知識

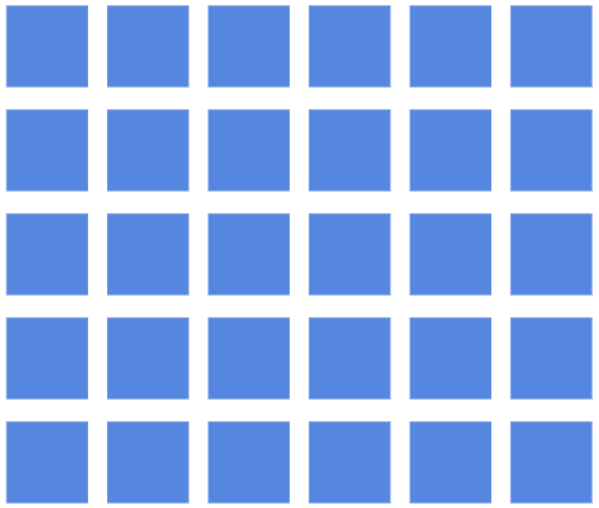
PPDAC research cycle



Structured vs. Unstructured data

結構式資料

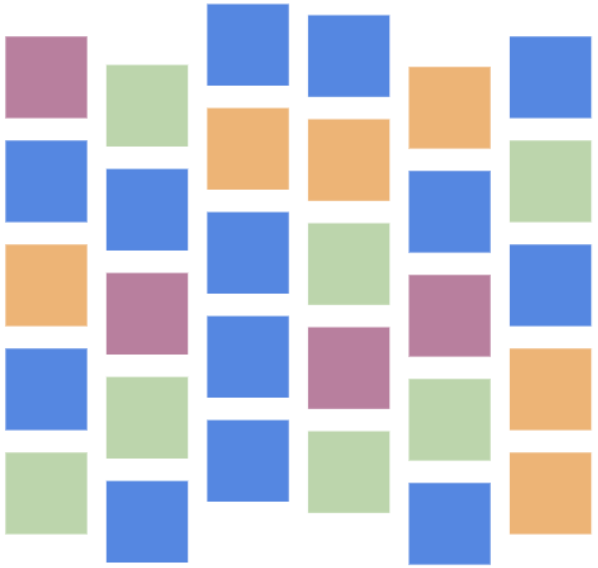
Structured data



Data stored in databases and tables

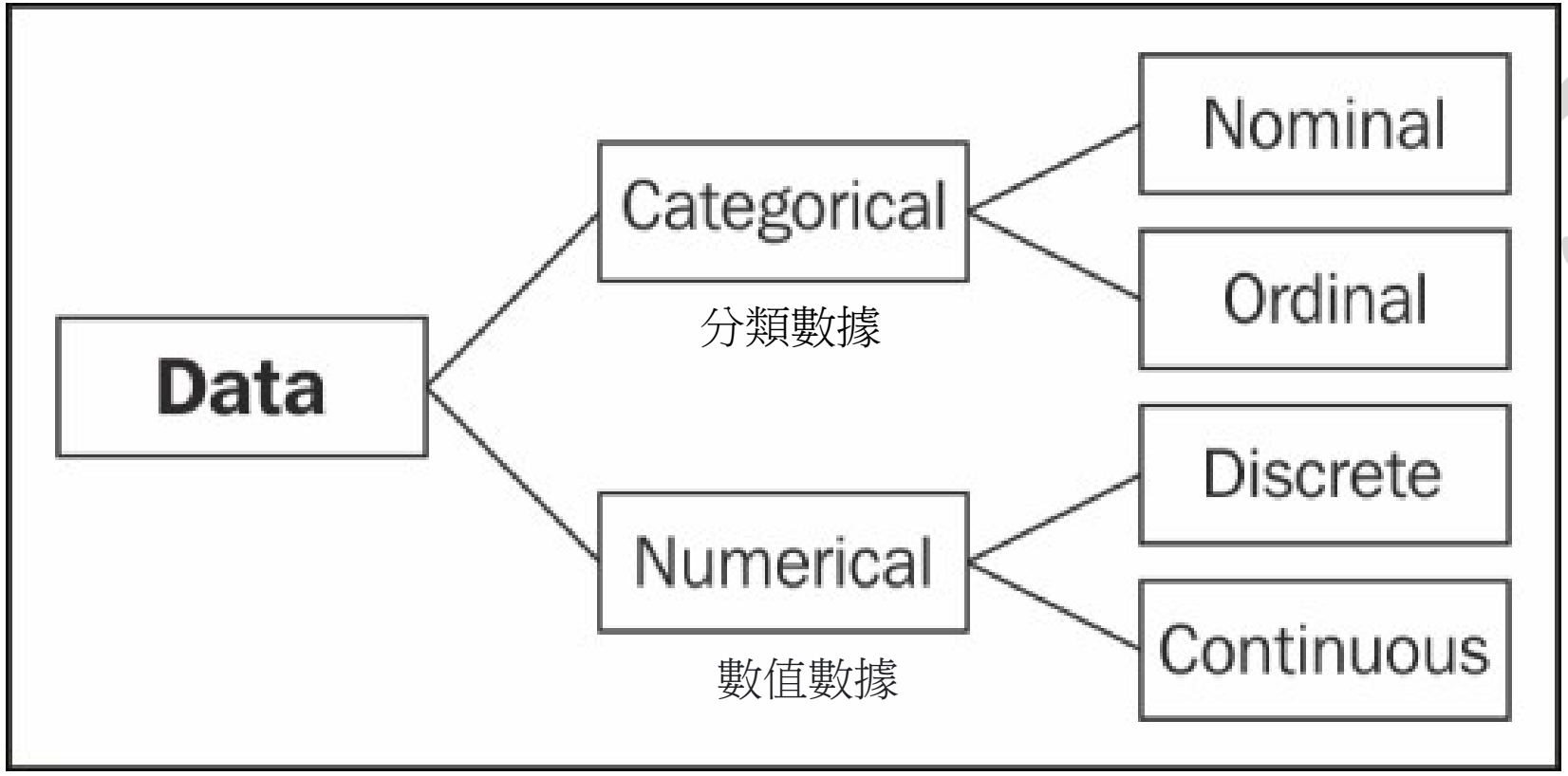
非結構式資料

Unstructured data



Images, text, audio, video, documents

數據類型


Data
Categorical
分類數據
Numerical
數值數據
Nominal
Ordinal
Discrete
Continuous
定類數據
定序數據
離散數據
連續數據

Categorical (qualitative) data 分類數據

- Nominal 定類數據：
  - Unordered categories 無序類別
  - Mutually exclusive 互斥
  - e.g. male/female, smoker/non-smoker

- Ordinal 定序數據
  - Ordered categories 有序類別
  - Mutually exclusive 互斥
  - e.g. Strongly agree, agree, neutral, disagree, strongly disagree

Numerical (quantitative data) 數值數據

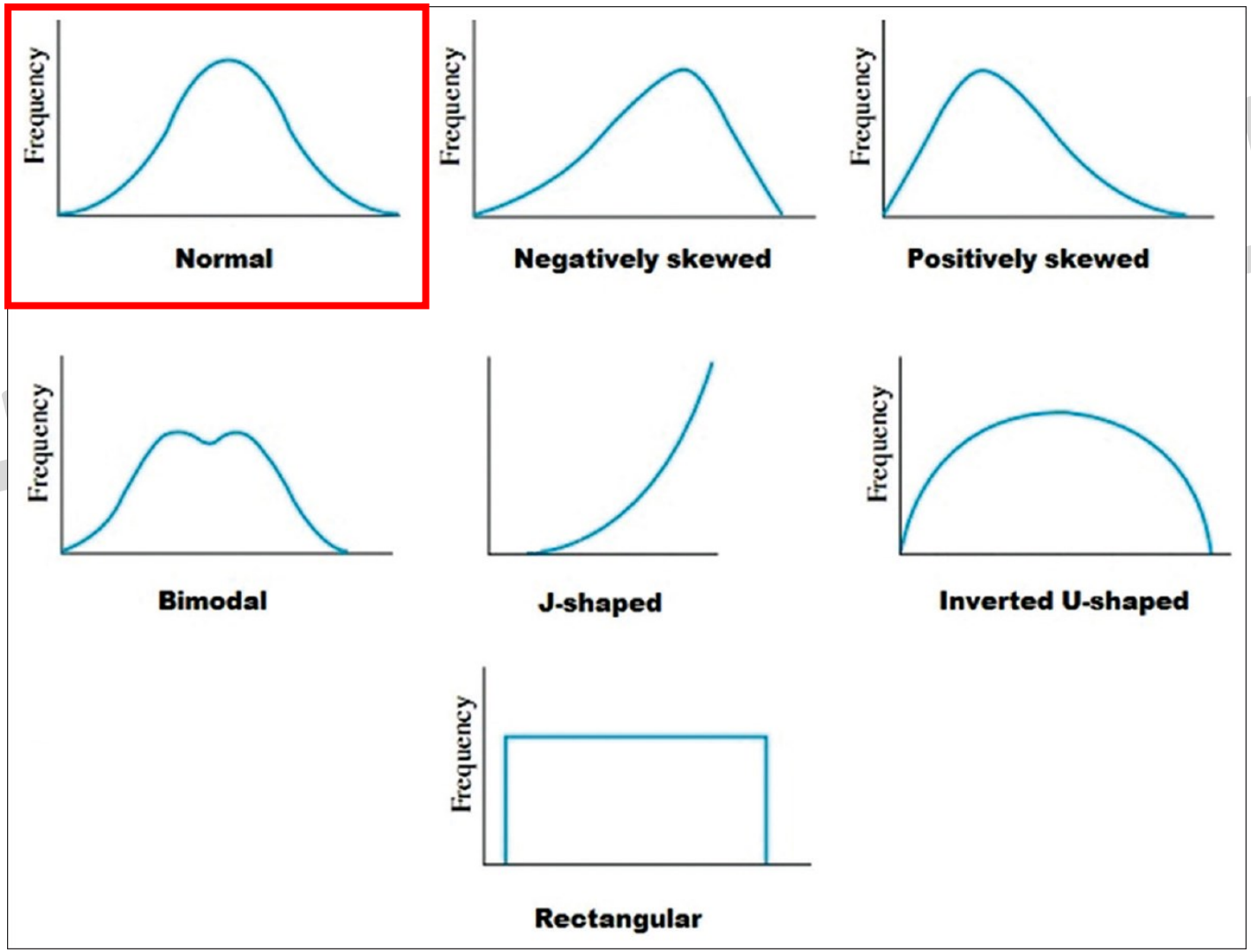
- Discrete 離散數據
  - Whole number value – typically counts 整數值
  - e.g. number of students in the room

- Continuous 連續數據
  - Can take any value within a range 可以非整數值
  - e.g. body weight and height

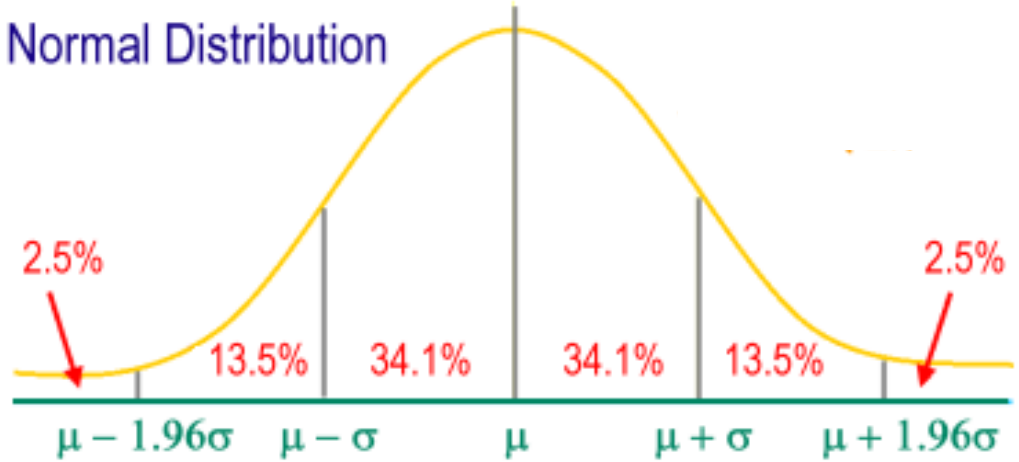
Statistics 統計學

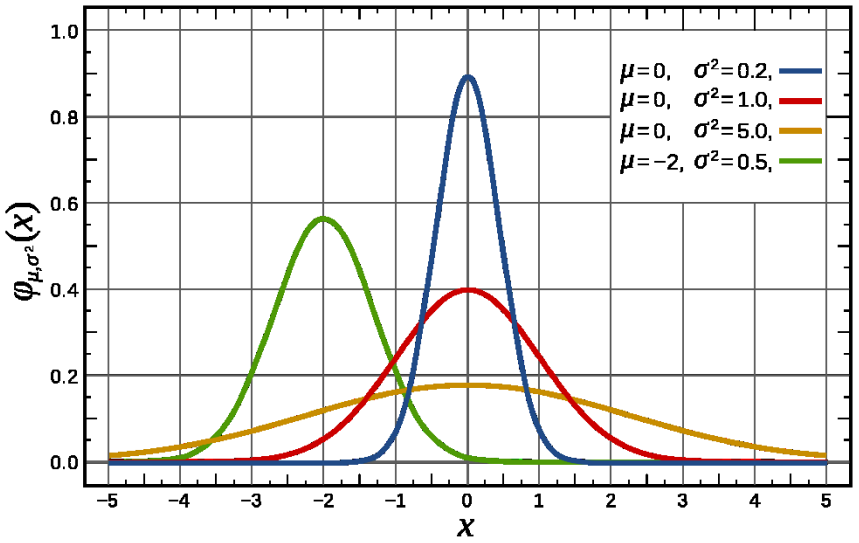
- 統計學是在資料分析的基礎上，研究測定、收集、整理、歸納和分析反映數據資料，以便給出正確訊息的科學。
- 統計研究中的共同目標是分析因果關係，或是從預估數據變化中得出結論。
  - 描述統計(descriptive statistics)是來描繪或總結觀察的基本情況。
  - 推論統計學(inference statistics) 將資料中的數據模型化，計算它的概率並且做出對於母群體的推論

Data distribution 數據分佈



Normal distribution 正態分佈



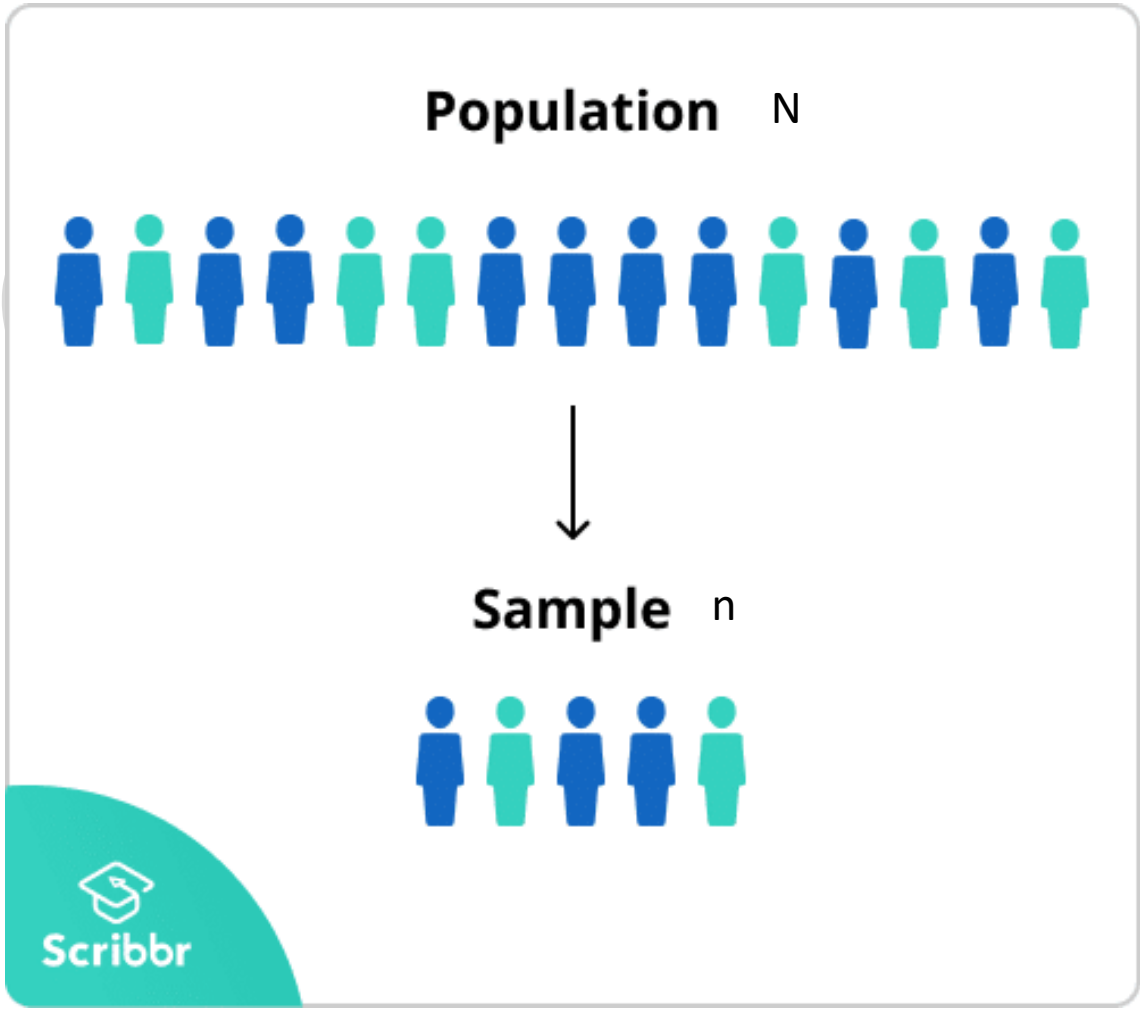


Continuous random variable 連續型隨機變數
例如：體重，身高，時間，溫度等

μ = mean 均值
σ = Standard Deviation 標準差

Random Sampling 隨機抽樣

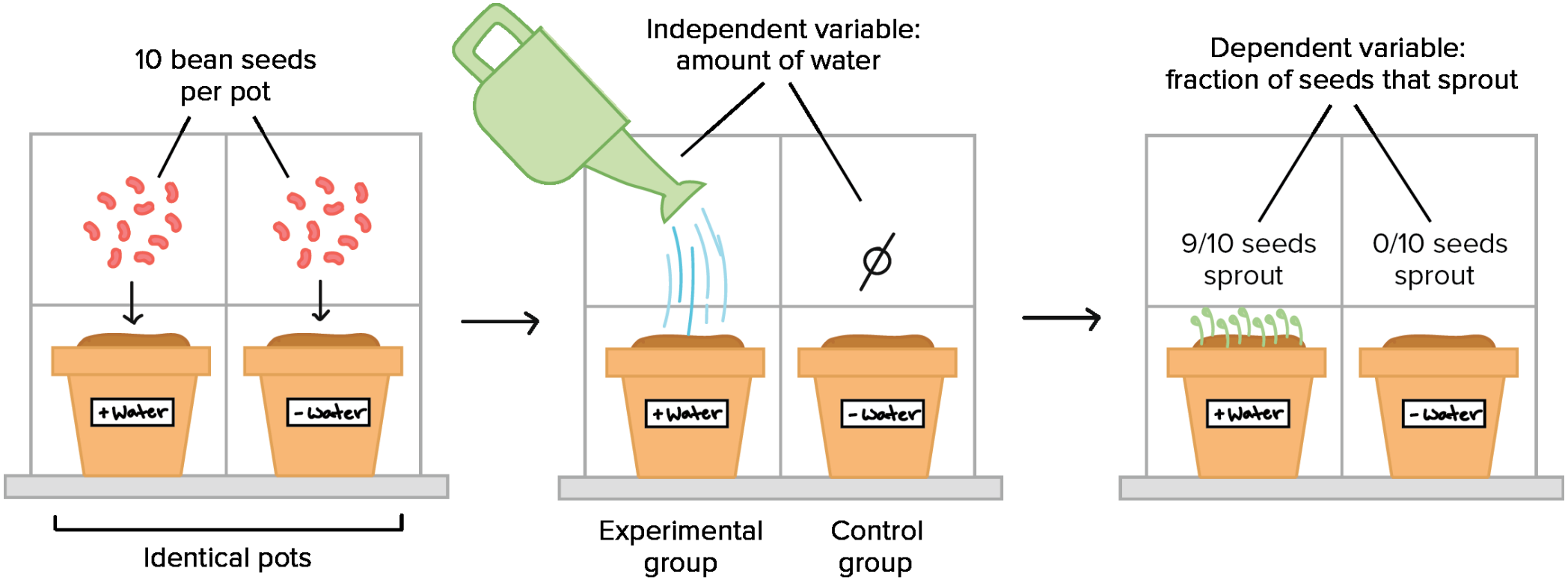
在統計學中，抽樣（Sampling）是一種推論統計方法，它是指從目標母體（Population）中抽取一部分個體作為樣本（Sample），通過觀察樣本的某一或某些屬性，依據所獲得的數據對母體的數量特徵得出具有一定可靠性的估計判斷，從而達到對母體的認識。



隨機抽樣: 從母體N個單位中隨機地抽取n個單位作為樣本，使得每一個容量為樣本都有相同的概率被抽中。

其他抽樣方法:
系統抽樣
分層抽樣

Experimental controls 實驗控制



獨立變數 (Indepdndent variable)：實驗中唯一能改變的因素。
因變數 (Dependent variable): 實驗結果中要測量或比較的項目

Probability 概率

- 又稱或然率、機率、機會率。
- 是一個在0到1之間的實數， 描述隨機事件發生之可能性。
- 事件發生的概率越高，表示這個事件比較可能發生。
- 例如:
  - 太陽從東方升起的概率為 1或 100%
  - 丟銅板時，正面朝上的概率為 0.5 或 50%
  - 太陽從西方升起的概率為 0或 0%

Null hypothesis 虛無假設

- 虛無假設 H0 : 所有因素對變數都不起任何作用。
- 虛無假設H0 認為事件是隨機發生的，而對立假設H1則認為事件有因果關係。
- 當有充足證據拒絕虛無假說時，即可接受對立假說，而若無充足證據證明對立假說為真時，則「不拒絕」虛無假說。

P-value

- P-value 常用於檢定虛無假設。
- P-value是虛無假設為真時獲得特定觀測結果的概率。
- 當p-value 很小時 ， 意味着在虛無假設下觀測特定結果的發生概率很小。
- 當p-value 值低於某個標準（通常為 0.05或 0.01）時，虛無假設將會被拒絕。
- 因此，p-value小於0.05等於統計學上所謂的顯著性差異 (significant differences)

Type I and Type II errors 第一型及第二型錯誤

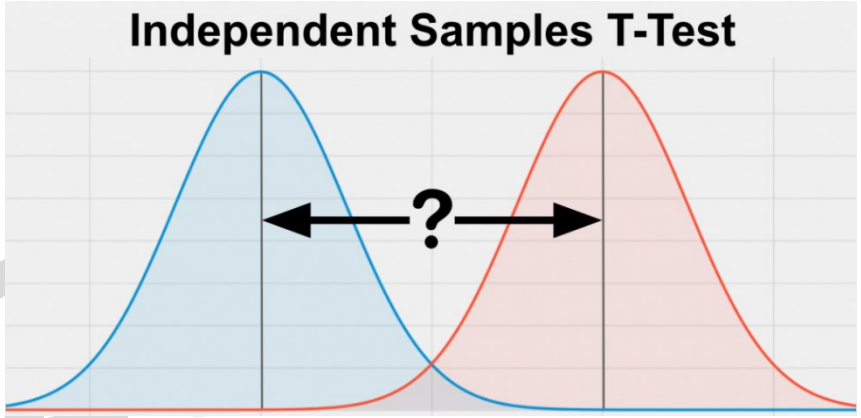
在虛無假設中存在兩種基本誤差：

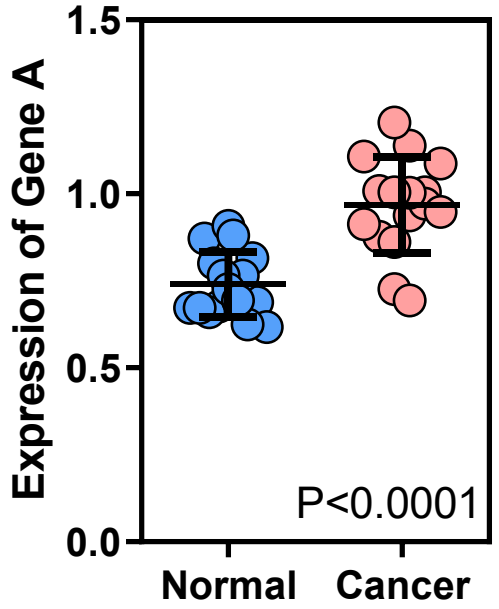
- 第一型錯誤中虛無假設被錯誤地證偽，得出測試結果為「偽陽性」(False positive)。
- 第二型錯誤中虛無假設沒有被及時排除，被錯誤判斷為「偽陰性」(False negative)。

		真實情況	
		H_0（虛無假說）為真	H_α（對立假說）為真
根據研究結果的判斷	拒絕H_0	錯誤判斷 （**偽陽性 false positive**、**型一錯誤 type-1 error**） 發生機率α（**顯著水準**）	正確判斷， 發生機率1-β（**檢定力**）
	不拒絕H_0	正確判斷	錯誤判斷 （**偽陰性 false negative**、**型二錯誤 type-2 error**） 發生機率β

Student's t-test

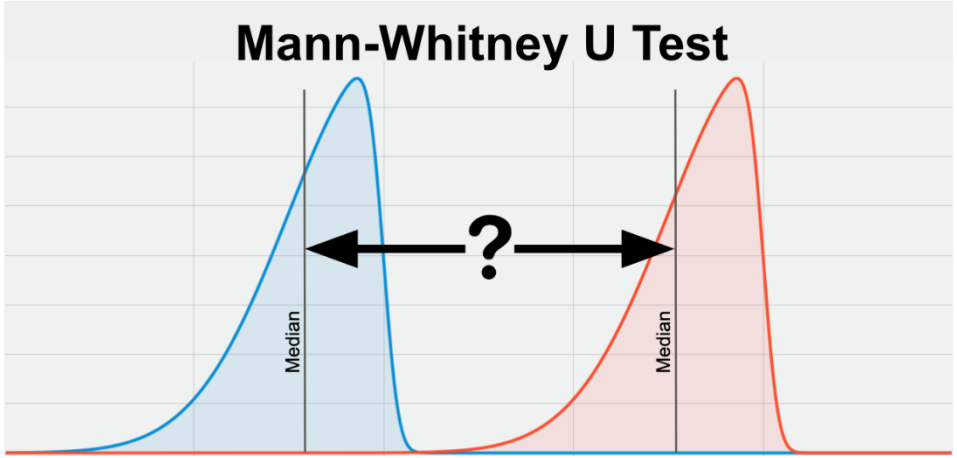
- 比較兩組的均值和標準差
- 連續變量
- 正態分佈
- 例如:
  - **18歲男孩和女孩的體重**
  - 基因**A**在癌症患者和健康人群中的表達





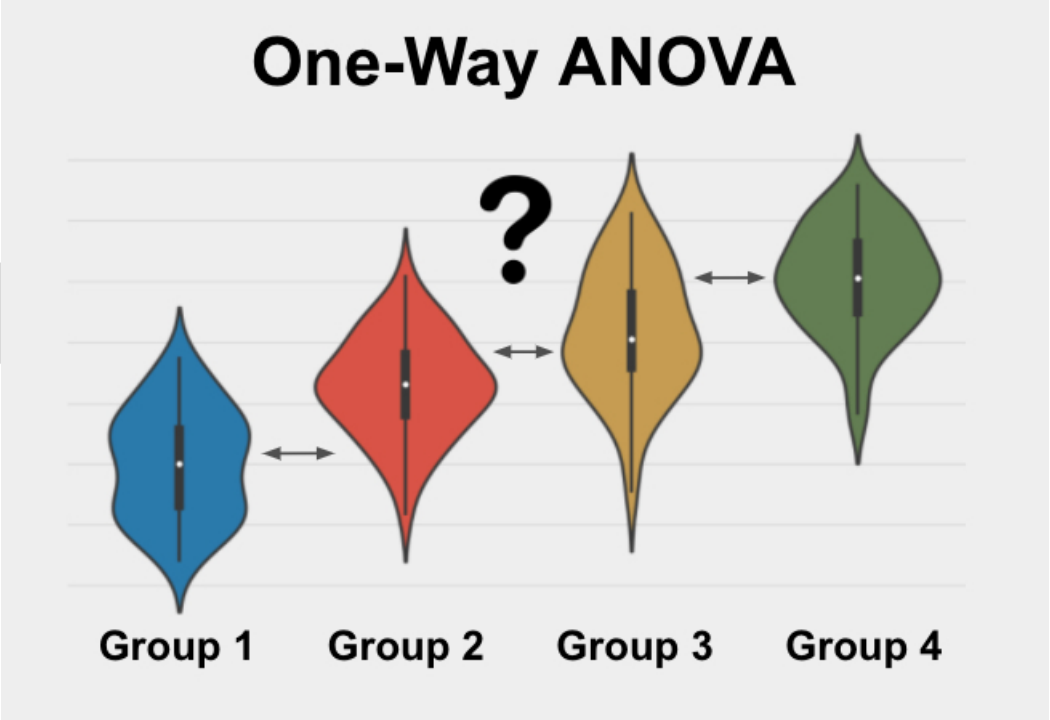
Mann-Whitney U test

- 比較兩組的中位數
- 連續變量
- 不假定正態分佈
- 例如:
  - 香港人口的收入



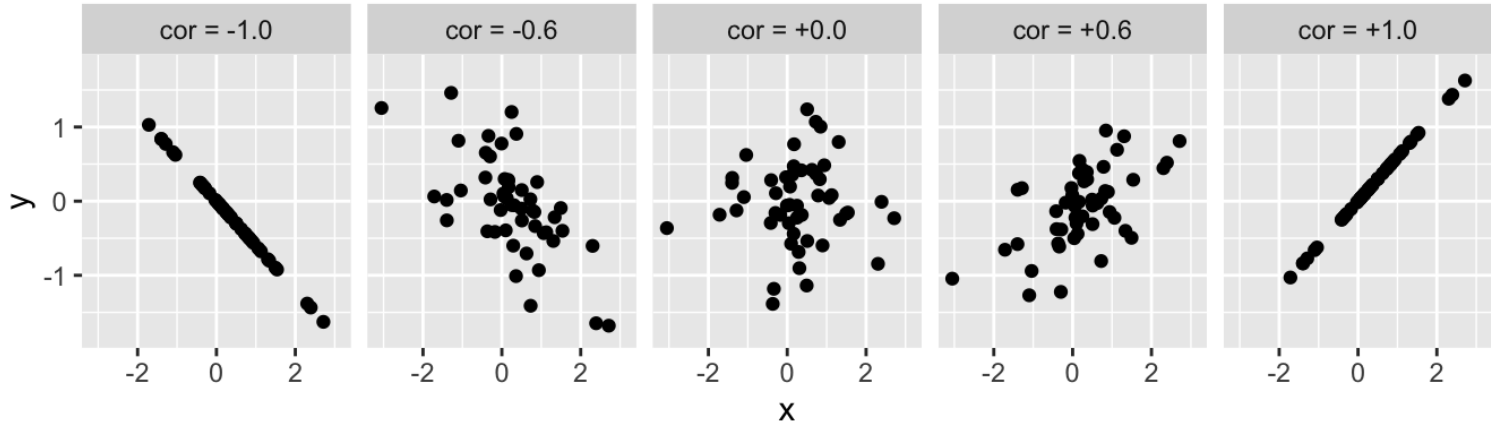
One-Way ANNVA

- 比較多於兩組的均值
- 連續變量
- 正態分佈



Correlation tests

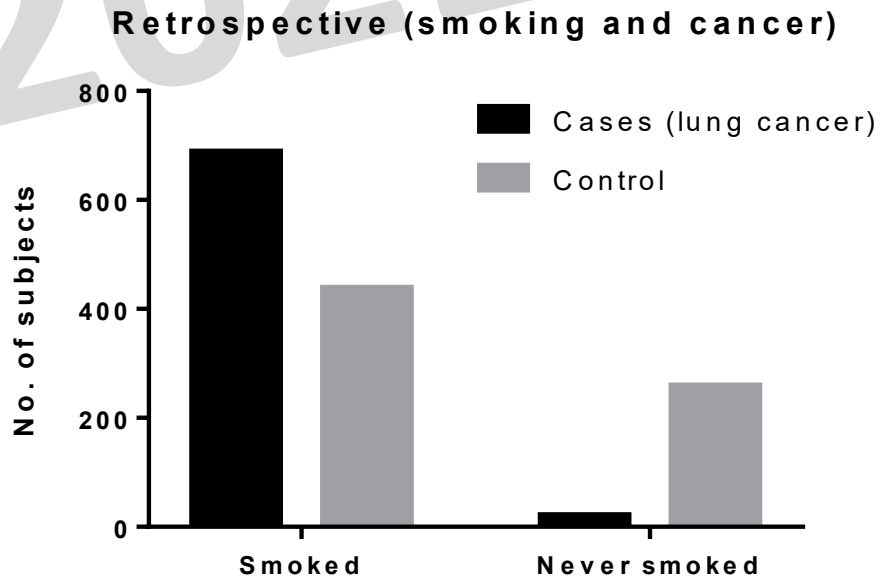
- 評價兩個統計變數的相關性。
- Pearson correlation
  - 連續變量
  - 正態分佈
- Spearman correlation
  - 連續變量
  - 不假定正態分佈



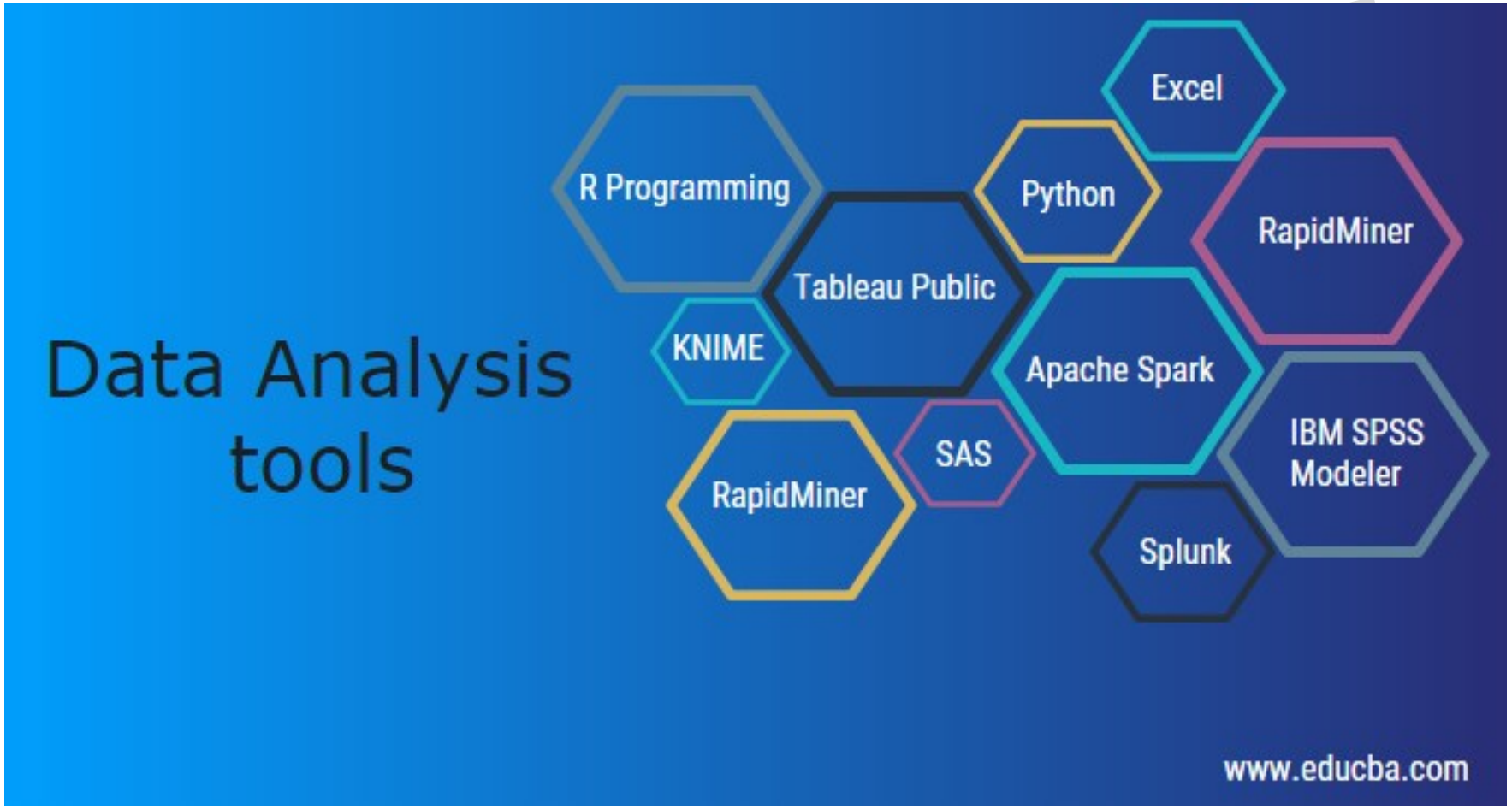
Chi-Square test or Fisher's exact test

- 比較兩組(或多組)分類數據的相關性
- 例如: 肺癌患者吸煙者比例

	Cases (lung cancer)	Control
Smoked	688	438
Never smoked	21	259



Data analysis tools



Data visualization

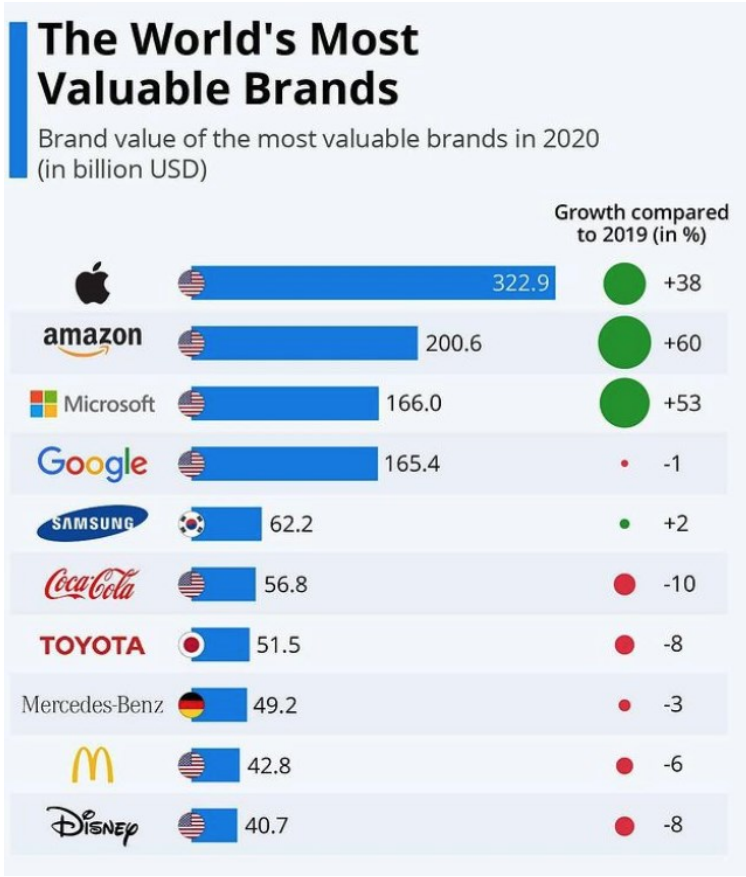
Outline

- Introduction of data visualization
- Basic data visualization
- Analysis and data visualization in Excel
- Practical session

Graph vs Text

According to the newly released Best Global Brands 2020 report from Interbrand, the company which boasts the largest value backing up its brand is currently Apple, with an estimated $322.9 billion - taking it again to the top of this particular list. A touch behind is Amazon,with a value of $200,677 million. Jeff Bezos' company was even able to increase its brand value by 60 % compared to last year. Tech still dominates the top ten, in fact, as do U.S. brands with three geographical exceptions : the South Korean giant Samsung , Japanese car manufacturer Toyota as well as Mercedes Benz (Germany).

Amid a global pandemic, social media and communication brands have fared well in the past 12 months with Instagram (#19), YouTube (#30) and Zoom (#100) entering the rankings for the first time. Tesla has re-entered the rankings at #40 with a brand value of US$12,785m, having last appeared in the Best Global Brands table in 2017.

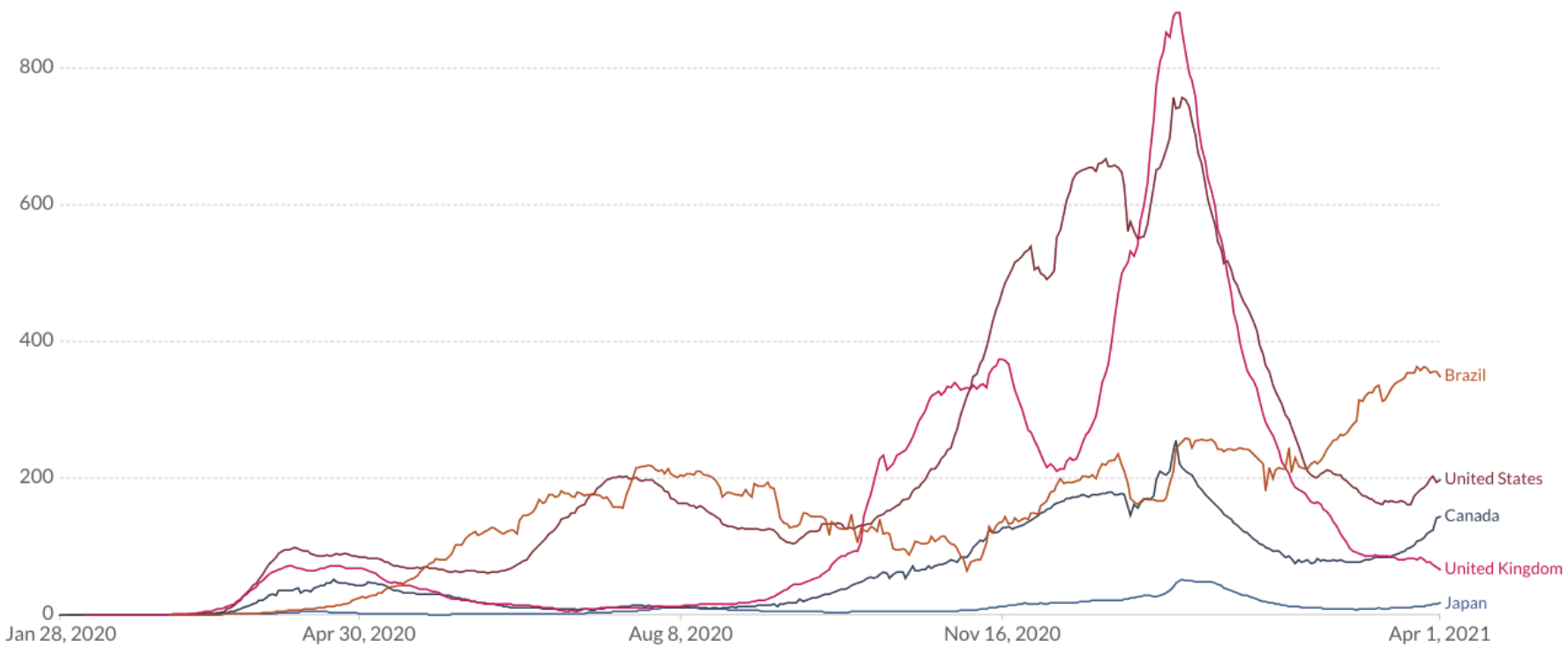


Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

Our World in Data

LINEAR LOG



Data visualization

Aim:

- Graphically present data

- Help audiences to comprehend data
  - Text vs graph

Principle:

- Avoid distorting the data (or torturing data)

- Do not mislead audiences

Popular programs for data visualization

Programming languages:

- Python (Free, general purpose, the fastest growing rate)
- R (Free, specialized in data visualization and statistics)
- MATLAB (Subscription basis, excellent in Math related analysis)

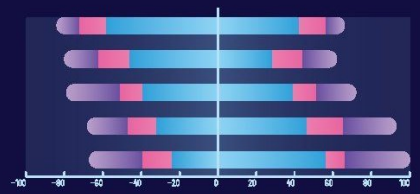
Software:

- GraphPad Prism (Subscription basis)
- Excel (The most accessible)
- LibreOffice (Alternative of Excel, free)

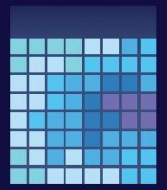
BAR CHART

SURPLUS

DIVERGING CHART



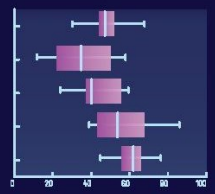
HEATMAP



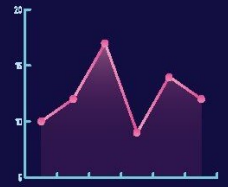
PICTORISL FRACTION CHART

CONTOUR MAP

BOXPLOT

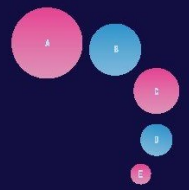


SCATTRPLOT

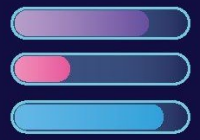


POPULATION PYRAMID

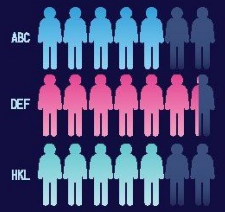
PROPORTION FILTERS



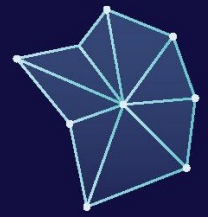
PROGRESS



FRACTION CHART

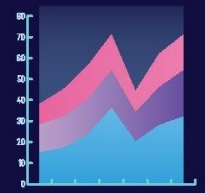


RADAR

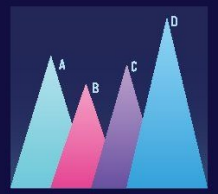


CANDLESTICK

AREA CHART



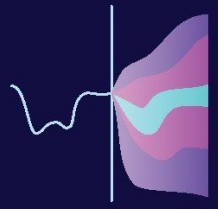
PYRAMID CHART



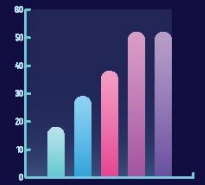
TIMELINE

COMBINATION CHART

FAN CHART

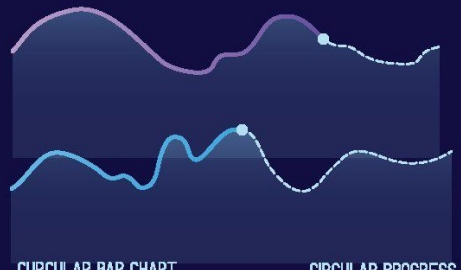


WATERFALL CHART



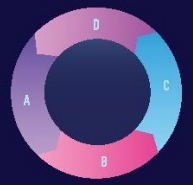
NETWORK SCHEME

TRACKING

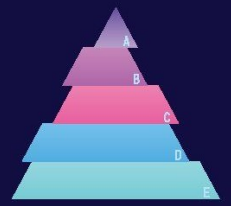


CIRCULAR BAR CHART

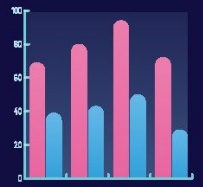
PROCESS CIRCLE



PYRAMID CHART



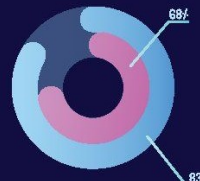
GROUPED BAR CHART



POLAR CHART

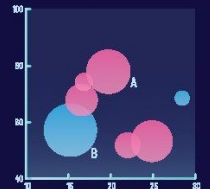
DONAT BAR CHART

CURCULAR BAR CHART

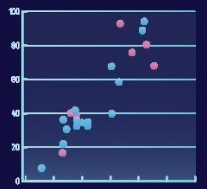


CIRCULAR PROGRESS

SCATTERPLOT



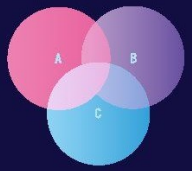
BUBBLE CORRELATION



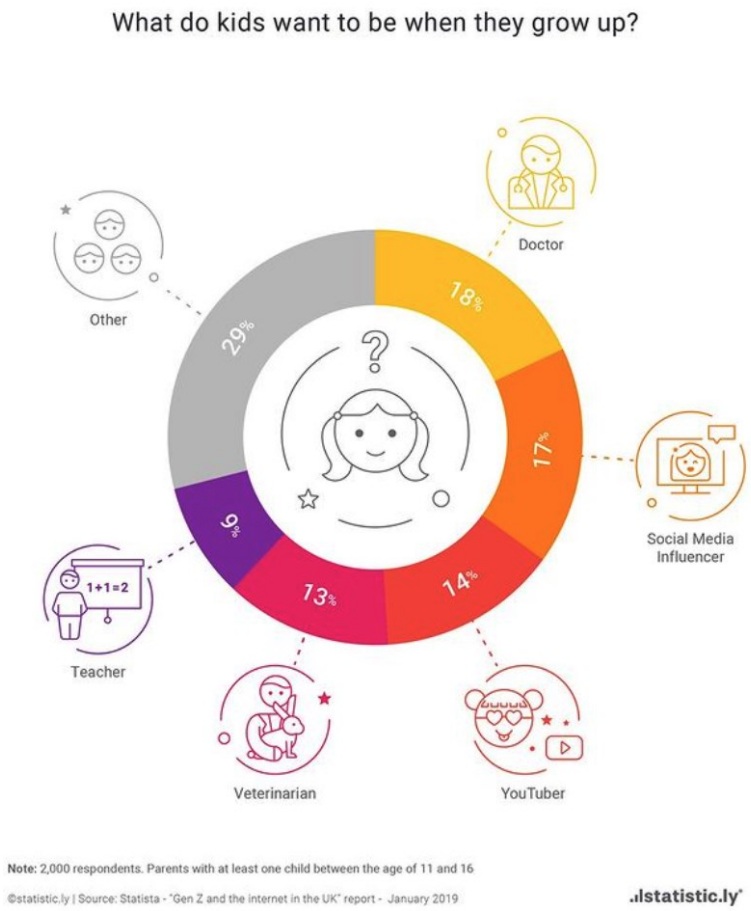
SPLINE GRAPH

FLOW MAP

VENN



Pie chart (圓形圖)



Features:

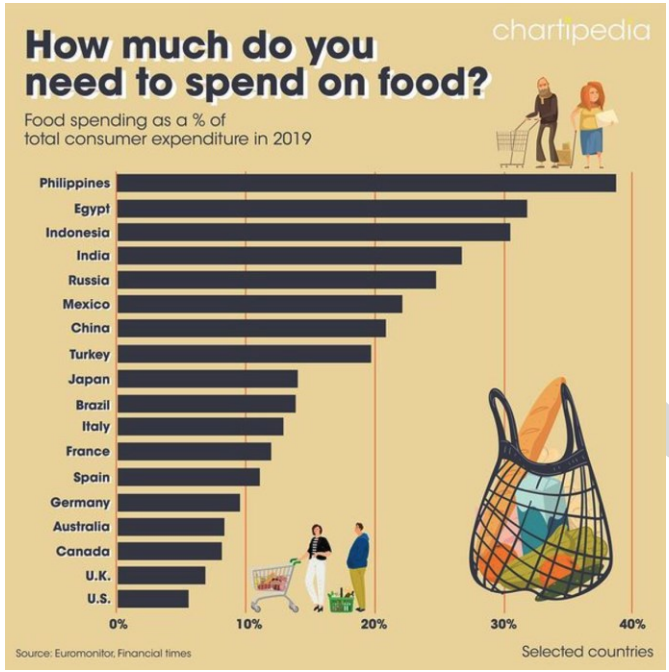
The size of each slice is proportional to the quantity.

Purpose:

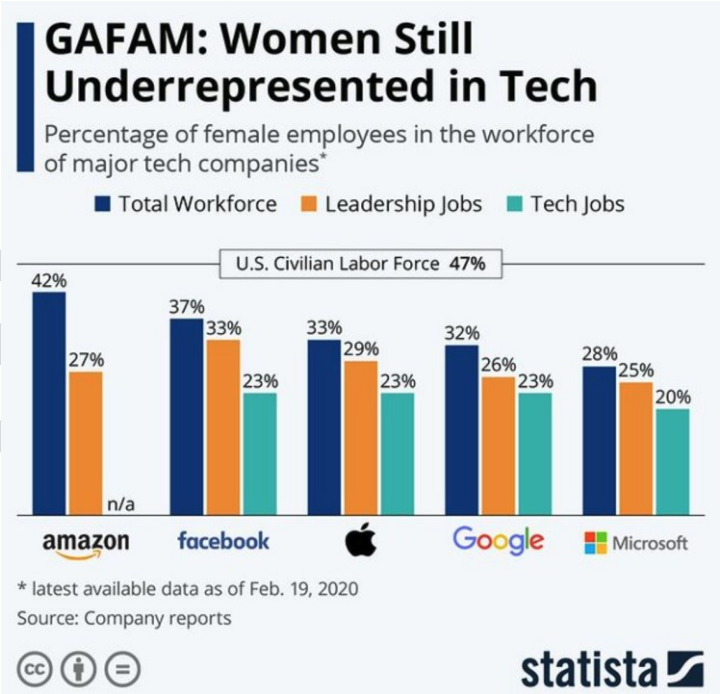
To illustrate and compare proportion of different category in the data. (Categorical data)

Bar chart (棒形圖)

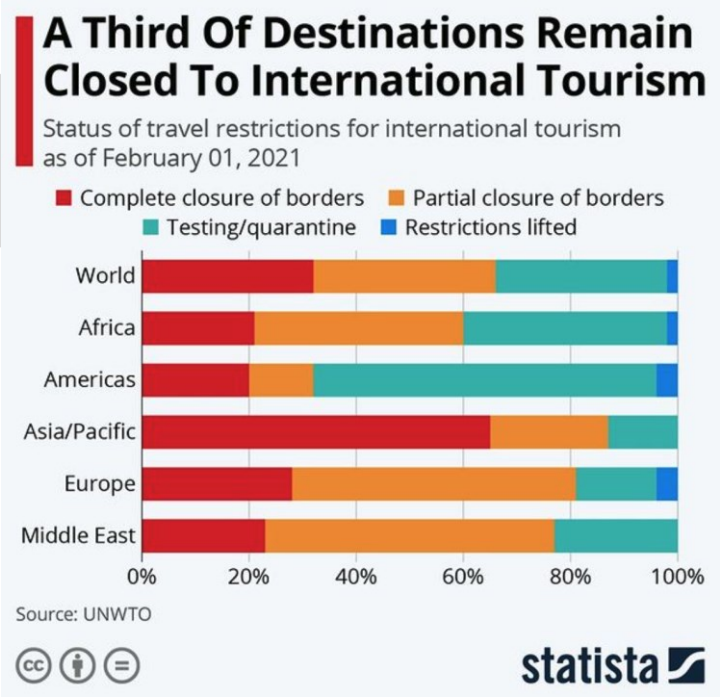
Ordinary



Grouped



Stacked



Feature:

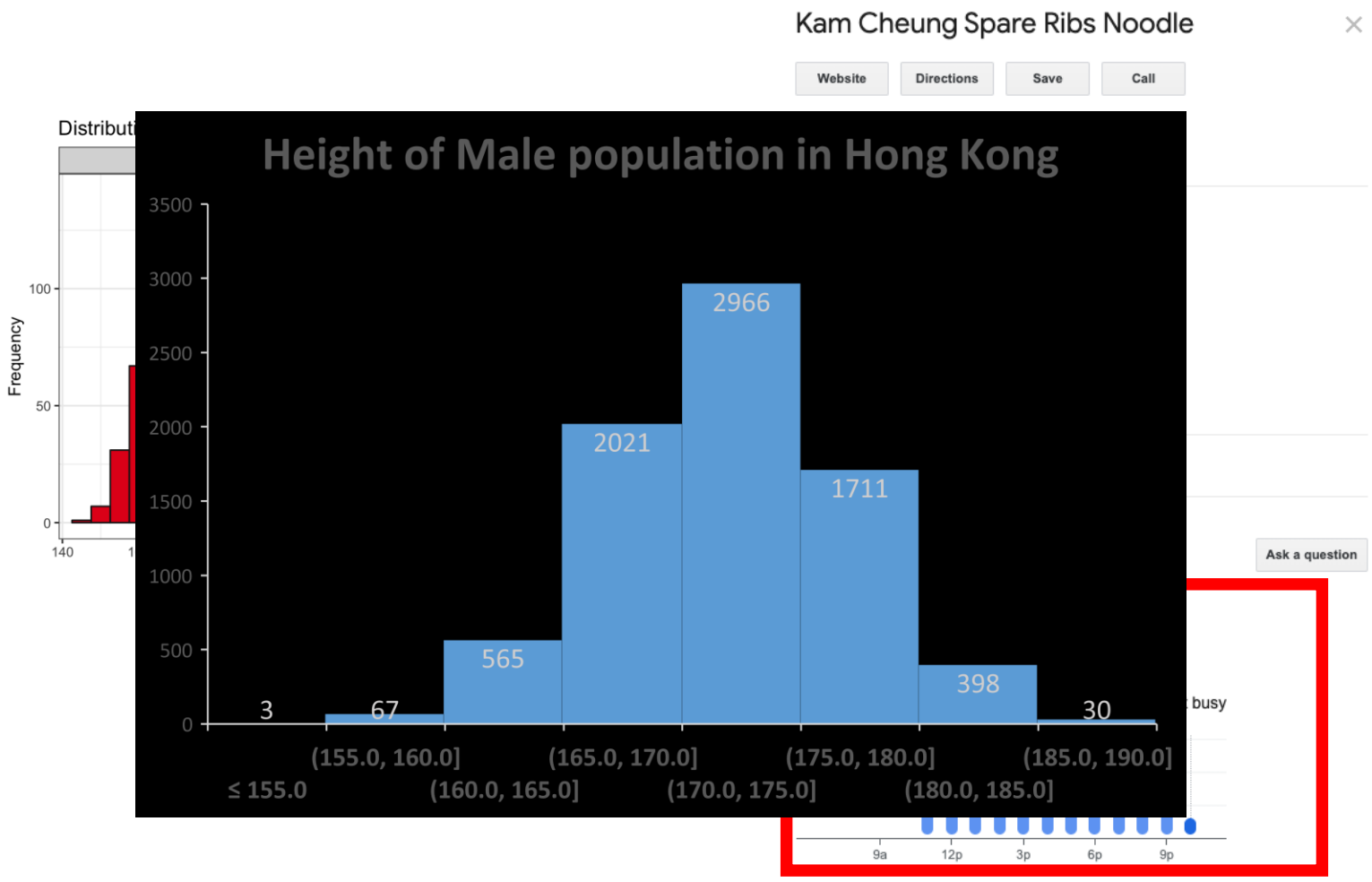
X-axis: Data category

Y-axis: Value

Purpose:

To compare values among different categories

Histogram (直方圖)



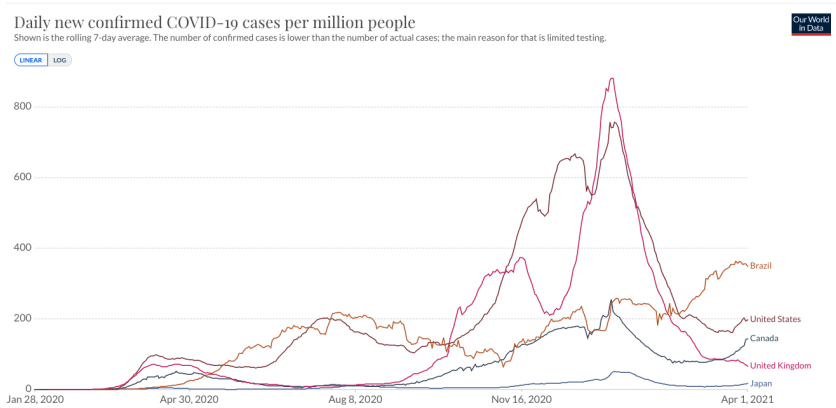
Features:

Separating continuous data into bins and counting the number in each bin (frequency).

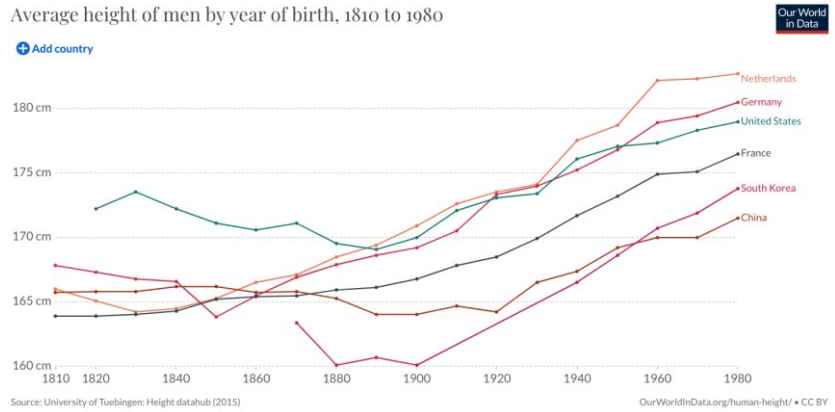
Purpose:

Understand the distribution or nature of the data.

Line graph (線形圖)



Retrieved from http://ourworldindata.org/



Retrieved from http://ourworldindata.org/

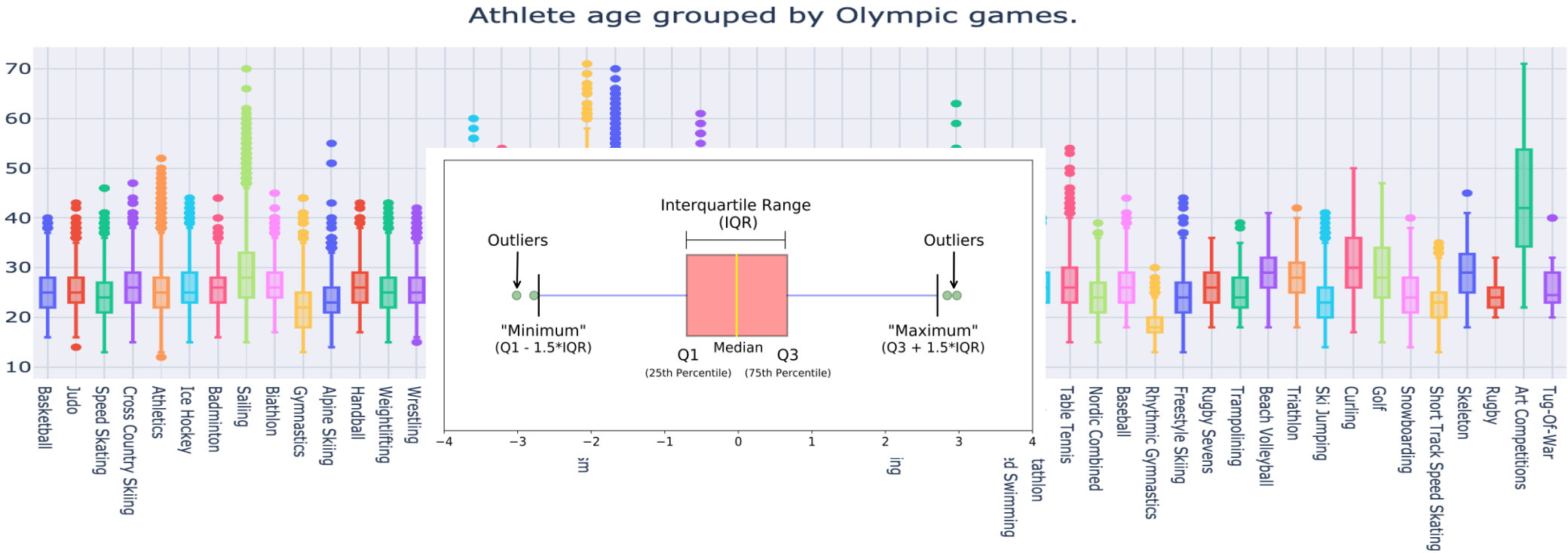
Features:

X – axis: continuous variable (usually, time-series)

Purpose:

To demonstrate and compare the changes of value over a time series.

Box plot (箱形圖)



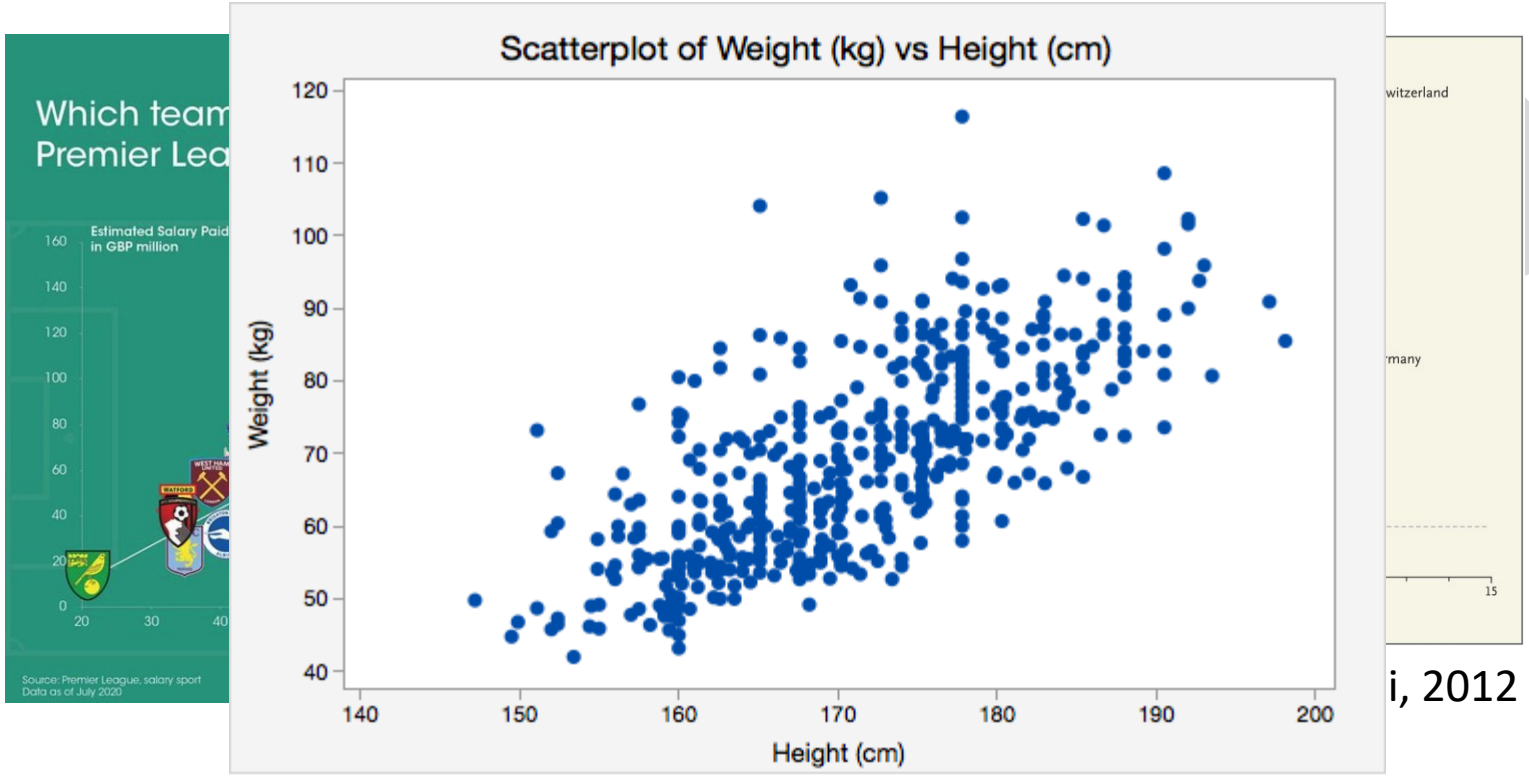
Features:

Always indicating the Quartiles (25, 50, 75%)

Purpose:

To generally demonstrate the data distribution for comparison

Scatter plot (XY plot) (散點圖)

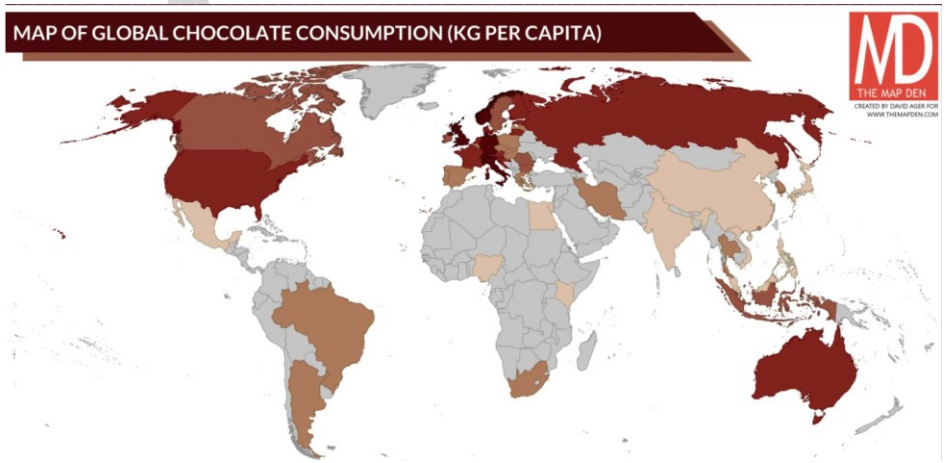


i, 2012

Features:
X – axis: Numeric data
Y – axis: Numeric data

Purpose:
To demonstrate correlation between two variables

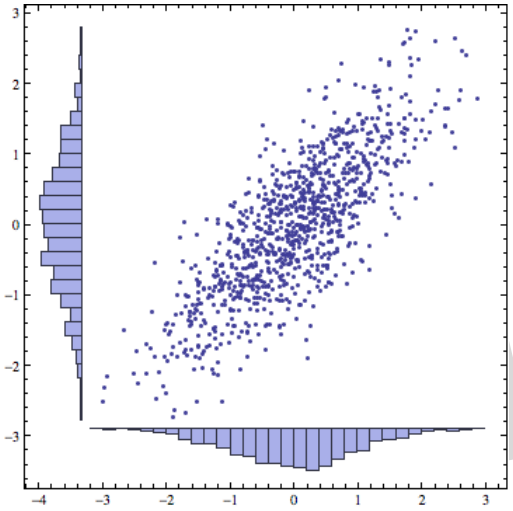
Map Chart

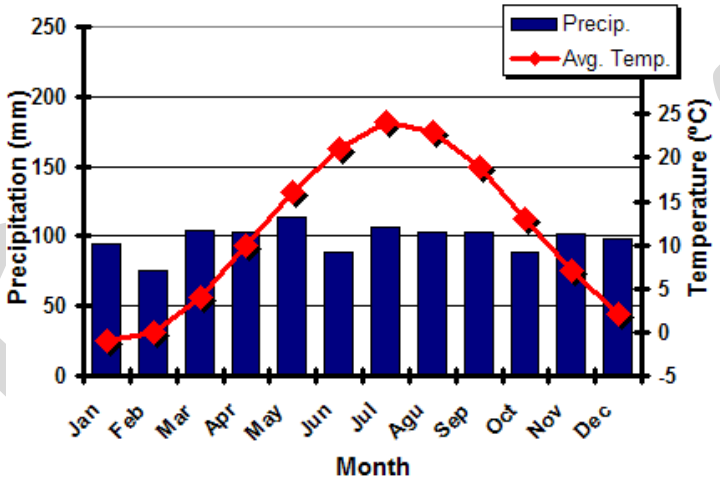


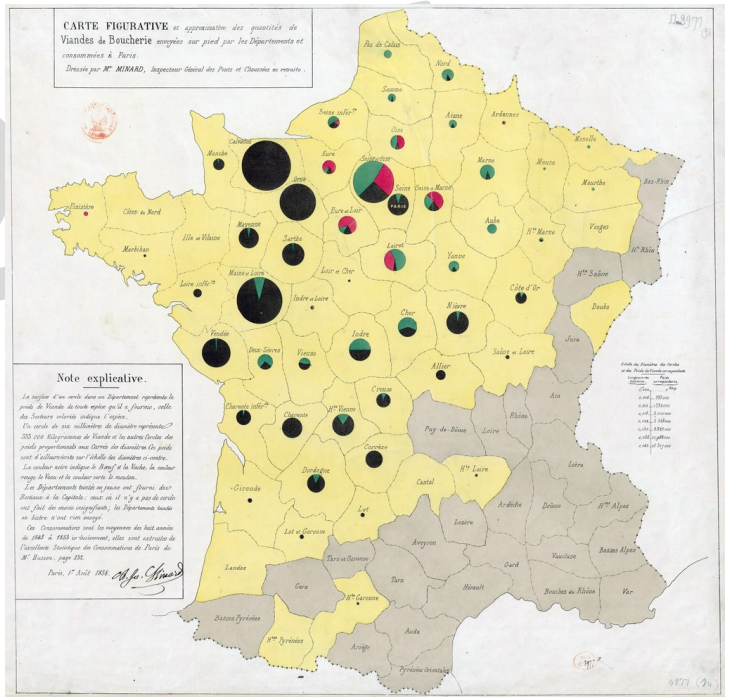
Features:
Displaying values on a map. Usually, values are indicated by colors.

Purpose:
To visualize the geographic relationship.

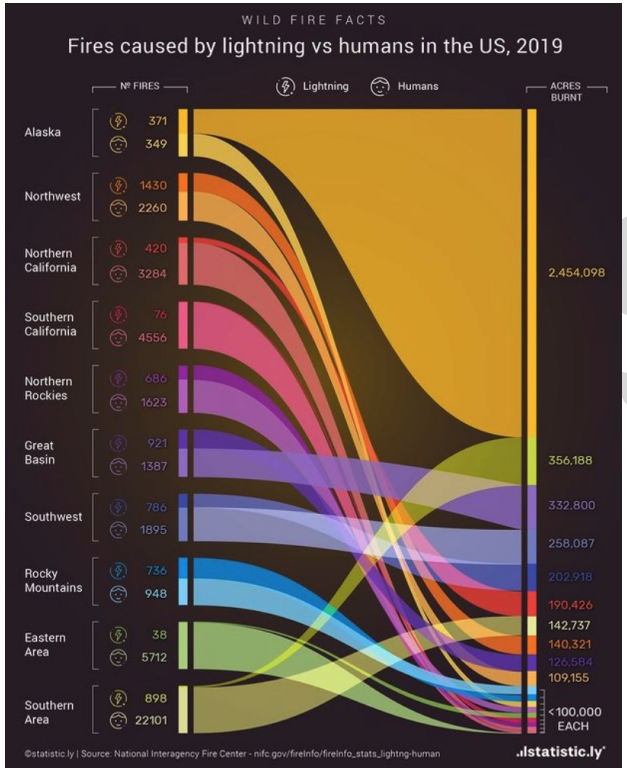
Mixture of graphs

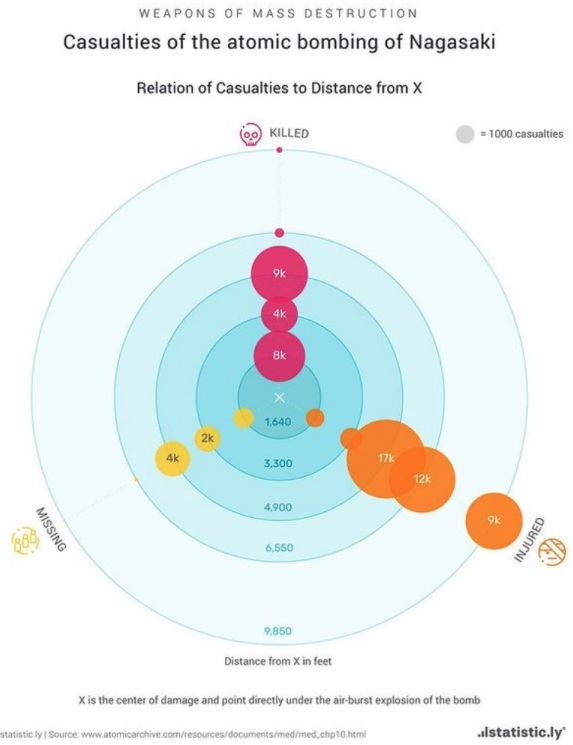


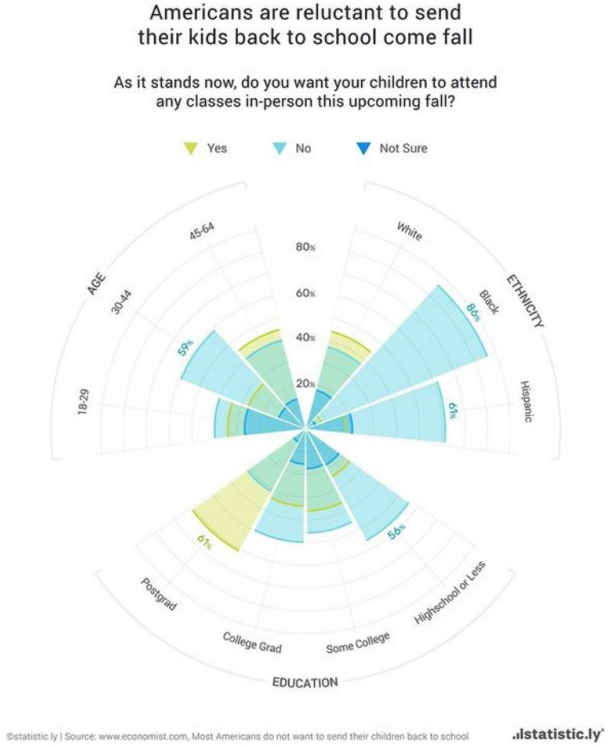




Customized


WILD FIRE FACTS
Fires caused by lightning vs humans in the US, 2019
Nº FIRES
Lightning
Humans
ACRES BURNT
Alaska
371
349
Northwest
1430
2260
Northern California
420
3284
Southern California
76
4556
Northern Rockies
686
1623
Great Basin
921
1387
Southwest
786
1895
Rocky Mountains
736
948
Eastern Area
38
5712
Southern Area
898
22101
2,454,098
356,188
332,800
258,087
202,918
190,426
142,737
140,321
126,584
109,155
<100,000 EACH
©statistic.ly | Source: National Interagency Fire Center - nifc.gov/fireInfo/fireInfo_stats_lightng-human
.ıstatistic.ly


WEAPONS OF MASS DESTRUCTION
Casualties of the atomic bombing of Nagasaki
Relation of Casualties to Distance from X
KILLED
= 1000 casualties
9k
4k
8k
1,640
2k
4k
3,300
17k
12k
4,900
9k
6,550
MISSING
INJURED
9,850
Distance from X in feet
X is the center of damage and point directly under the air-burst explosion of the bomb
©statistic.ly | Source: www.atomicarchive.com/resources/documents/med/med_chp10.html
.ıstatistic.ly


Americans are reluctant to send their kids back to school come fall
As it stands now, do you want your children to attend any classes in-person this upcoming fall?
Yes
No
Not Sure
AGE
45-64
30-44
18-29
59%
80%
60%
40%
20%
ETHNICITY
White
Black
Hispanic
84%
61%
61%
56%
Postgrad
College Grad
Some College
Highschool or Less
EDUCATION
©statistic.ly | Source: www.economist.com, Most Americans do not want to send their children back to school
.ıstatistic.ly

Choice of graphs

- Easy to comprehend

- Standard > Mixed > tailor-made

Practical session - Excel

- Useful operation in Excel
- Perform statistical analysis
  - T-test
- Create statistic graphs
  - Line graph
  - Pie graph
  - Scatter plot
  - Bar charts
  - Map chart

Analysis Demonstration - Excel

- Dataset:
  - Simulated data including the following fields of 55 students
    - Gender
    - Age
    - Height
    - Chinese, English, Math exam score
    - House
    - Place want to go after the pandemic

Analysis

- Analyze and visualize gender proportion
  - Overall (Pie chart)
  - Overall (Ordinary bar chart)
  - In each house (grouped bar chart)
- Visualize height distribution
- Visualize the change of height over time in female
  - Comparing the correlation of height in different age stages
- Comparing the score between gender

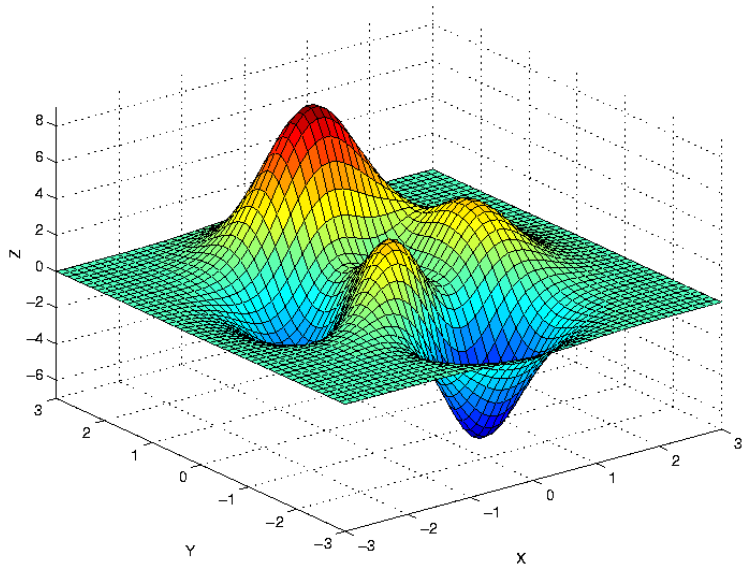
Practice – Nobel laureate

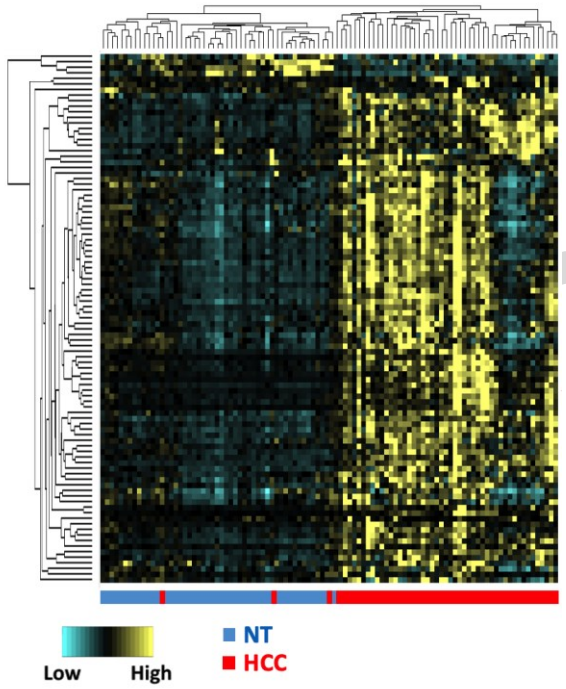
Download from

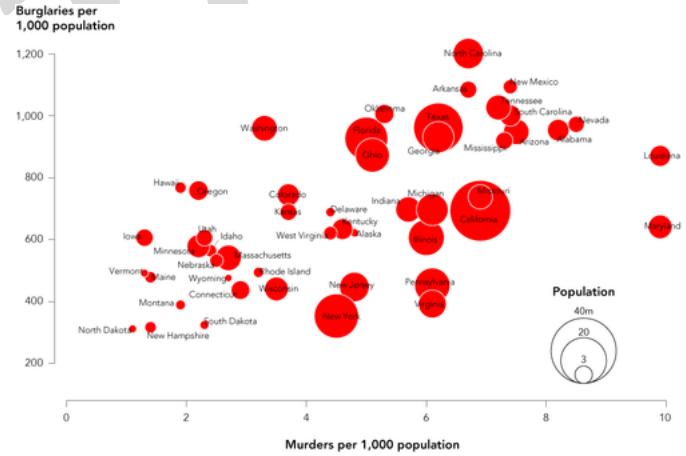
- https://bit.ly/39H7NNY

- Information of each Nobel prize winner from 1901 to 2020
  - Age
  - Awarded year
  - Birthplace
  - Prize Category
  - Gender
  - etc

Graphs in R


Z
Y
X


Low
High
NT
HCC


Burglaries per
1,000 population
Murders per 1,000 population
Population
40m
20
3

Useful online resource

- https://www.edx.org/
- https://www.coursera.org/

Or

- Search R/Python tutorial on Youtube