

天氣與人工智慧 機器學習實作

資料清理數據前處理 - 3

2020/11/04

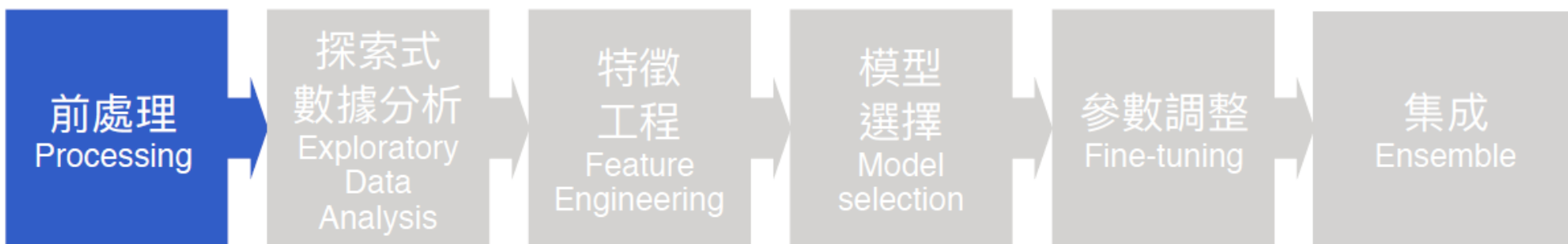
Subject

- 中位數與分位數
- 連續數值標準化
- 常用的 DataFrame 操作
- 相關係數簡介

知識地圖 機器學習前處理 中位數與分位數連續數值標準化

機器學習概論 Introduction of Machine Learning

監督式學習 Supervised Learning



非監督式學習 Unsupervised Learning



前處理 Processing



本節重點

- 如何處理例外值
- 如何進行數據標準化

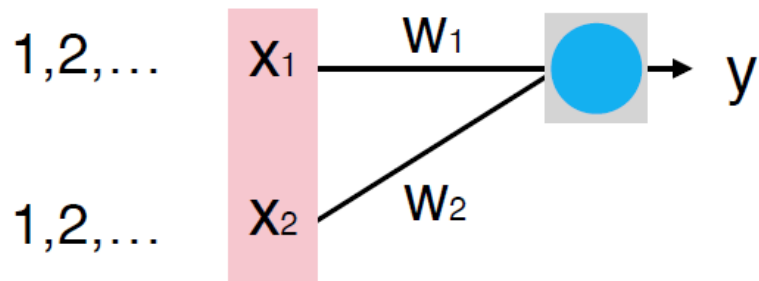
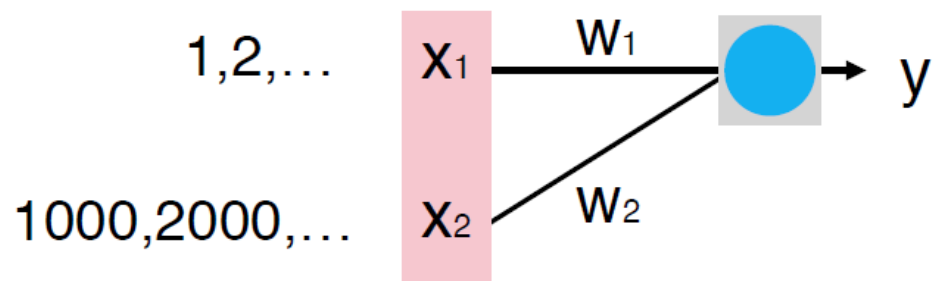
常用以填補缺失值的統計值

| 統計值 | 方法 |
|------------------------|---|
| <u>中位數</u> (median) | <code>np.median(value_arr)</code> |
| <u>分位數</u> (quantiles) | <code>np.quantile(value_arr, q = ...)</code> |
| <u>眾數</u> (mode) | <code>scipy.stats.mode(value_array)</code> 較慢的方法 dictionary method 較快的方法 |
| 平均數 (mean) | <code>np.mean(value_array)</code> |

連續型數值標準化

- 為何要標準化

改變一單位的 x_2 對 y 的影響完全不同



- 是否一定要做標準化 (有沒有做有差嗎)

看使用的模型而定

- Regression model : 有差
- Tree-based model : 沒有太大關係

Requires little data preparation. Other techniques often require data normalization. Since trees can handle qualitative predictors, there is no need to create dummy variables.

連續型數值標準化

常用的標準化方法

公式

Z 轉換

$$\frac{(x - \text{mean}(x))}{\text{std}(x)}$$

空間壓縮

$$Y = 0 \sim 1, \quad \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$Y = -1 \sim 1, \quad \left(\frac{x - \min(x)}{\max(x) - \min(x)} - 0.5 \right) * 2$$

$$Y = 0 \sim 1, \text{ (針對特別影像)}, \quad \frac{x}{255}$$

特殊狀況

有時候我們不會使用 min/max 方法進行標準化，而會採用 Qlow/Qhigh normalization (如將空間壓縮第一例中的 min 改為 q1, max 改為 q99)



解題時間

It's Your Turn

參考資料

- [Is it a good practice to always scale/normalize data for machine learning?](#)

- 閱讀重點：

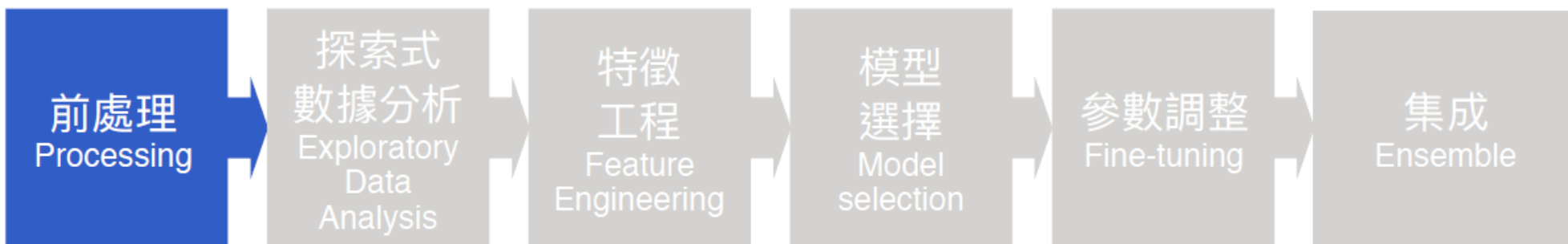
有的時候好，有得時候不好 (但爭議仍在，僅供參考)

- Good
 - 某些演算法 (如 SVM, DL) 等，對權重敏感或對損失函數平滑程度有幫助者
 - 特徵間的量級差異甚大
- Bad
 - 有些指標，如相關不適合在有標準化的空間進行
 - 量的單位在某些特徵上是有意義的

知識地圖 機器學習前處理 常用的 DataFrame 操作

機器學習概論 Introduction of Machine Learning

監督式學習 Supervised Learning



非監督式學習 Unsupervised Learning



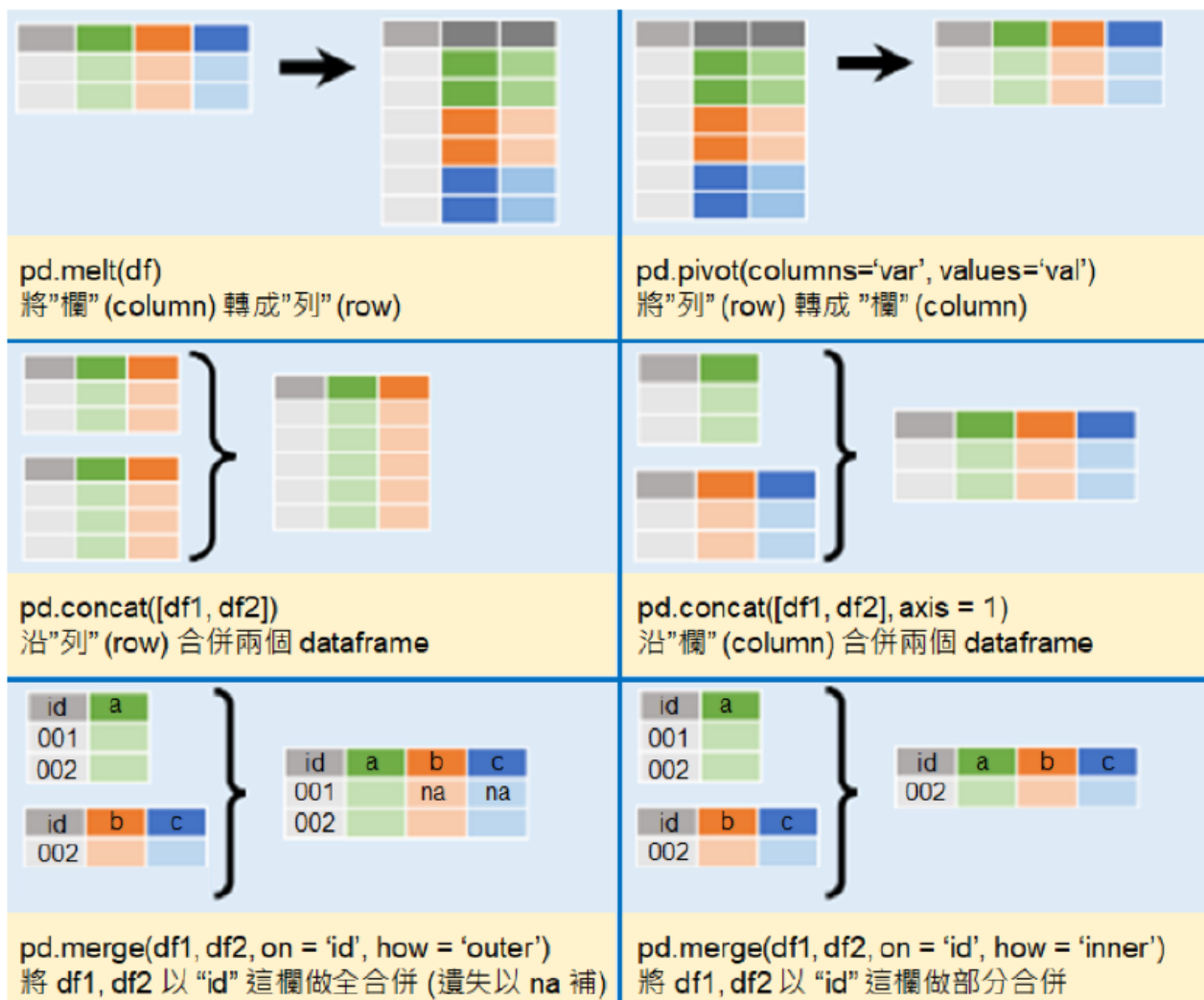
前處理 Processing



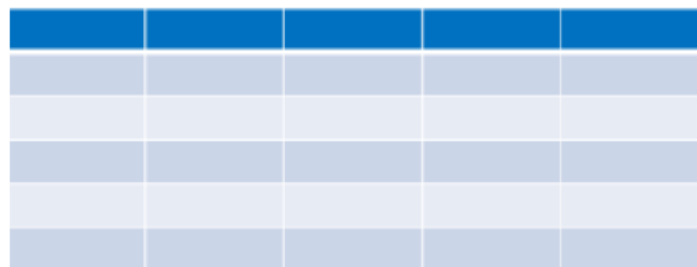
本節重點

- 熟悉 python 常用套件 pandas 的操作方式，如排序、合併、分組操作、indexing 等

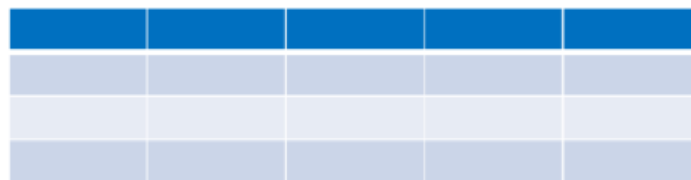
轉換與合併 DataFrame



subset



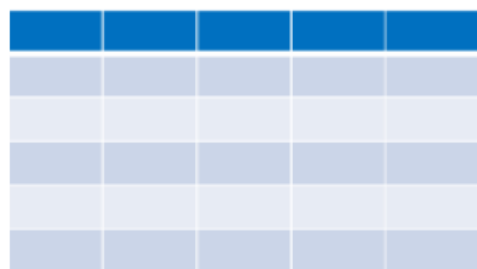
| | | | | |
|--|--|--|--|--|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |



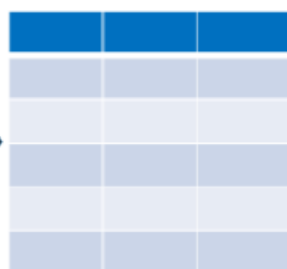
| | | | | |
|--|--|--|--|--|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

| 列篩選 / 縮減 | |
|-----------|--|
| 邏輯操作 | <code>sub_df = df[df.age > 20]</code> |
| 移除重複 | <code>df = df.drop_duplicates()</code> |
| 前 n 筆 | <code>sub_df = df.head(n = 10)</code> |
| 後 n 筆 | <code>sub_df = df.tail(n = 10)</code> |
| 隨機抽樣 | <code>sub_df = df.sample(frac = 0.5) # 抽 50 %</code> |
| | <code>sub_df = df.sample(n = 10) # 抽 10 筆</code> |
| 第 n 到 m 筆 | <code>sub_df = df.iloc[n : m]</code> |

| 邏輯操作 | |
|-----------------------------|------------------------------------|
| 大於 / 小於 / 等於 | <code>>, <, ==</code> |
| 大於等於 / 小於等於 | <code>>=, <=</code> |
| 不等於 | <code>!=</code> |
| <code>&, , ~, ^</code> | 邏輯的 and, or, not, xor |
| 欄位中包含 value | <code>df.column.isin(value)</code> |
| 為 Nan | <code>pd.isnull(obj)</code> |
| 非 Nan | <code>pd.notnull(obj)</code> |



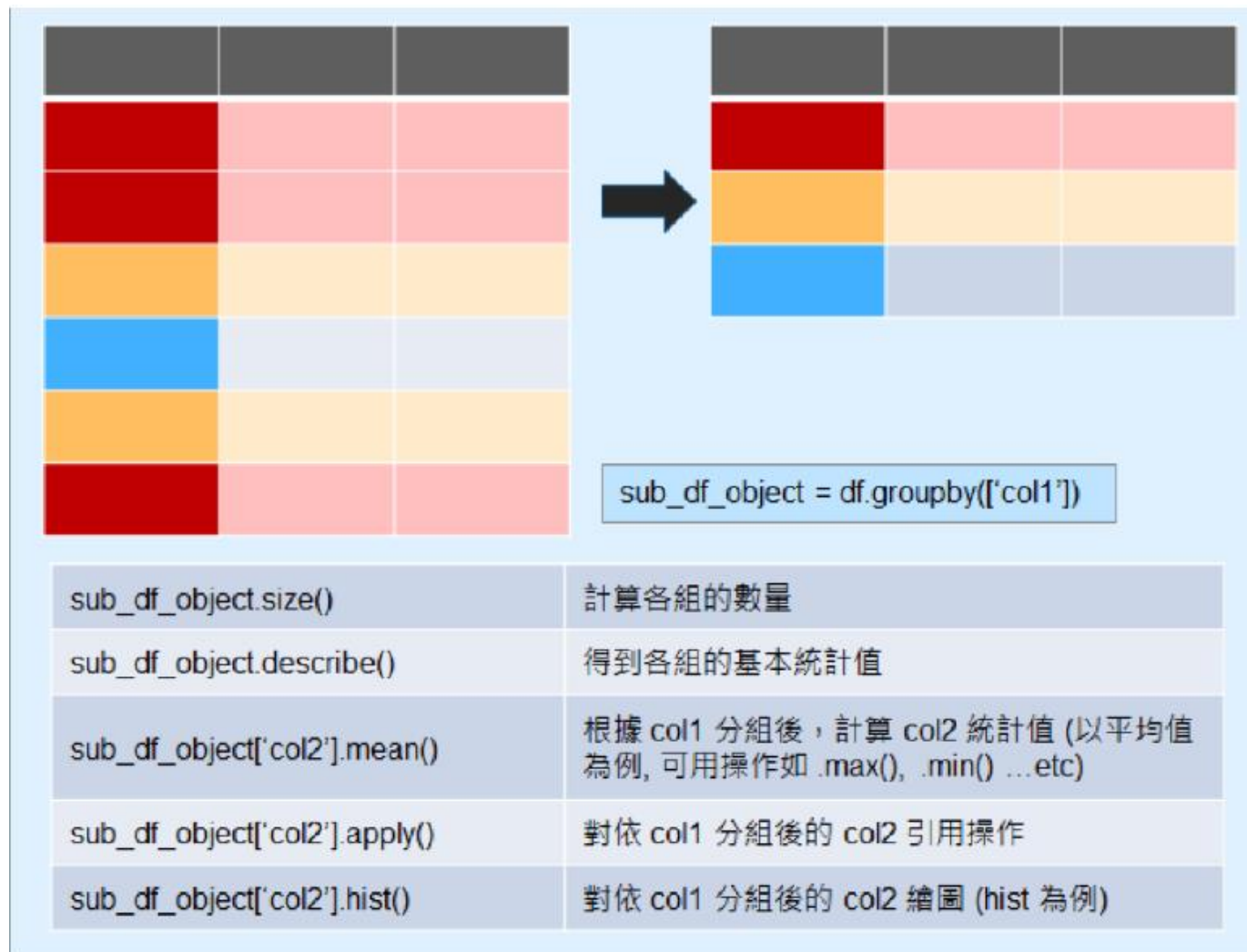
| | | | | |
|--|--|--|--|--|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |



| | | | | |
|--|--|--|--|--|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

| 欄篩選 / 縮減 | |
|----------|---|
| 單一欄位 | <code>new_df = df['col1']</code> 或 <code>df.col1</code> |
| 複數欄位 | <code>new_df = df[['col1', 'col2', 'col3']]</code> |
| Regex 篩選 | <code>new_df = df.filter(regex = ...)</code> |

Group operations



重點複習

- 合併 (concat) 常用於將多個表格依照某欄 (key) 結合使用
- 分組 (groupby) 是常用在計算“組”統計值時會用到的功能
- 許多基本操作 (如： $>$ 、 $<$ 、 $==$ 、 \sim) 都是可以在 pandas 作為篩選條件使用



解題時間

It's Your Turn

Pandas 官方 Cheat Sheet

- Pandas 官方 [網頁連結](#)
- 所謂的 Cheat Sheet, 中文好像該叫做“懶人包”, 各家語言與套件常常有懶人包 (雲端AI, R語言....), 目的不外乎是提供使用者便利, 以便推廣。不過 Pandas 的懶人包初學時似乎會被滿滿文字困惑住, 但學通後回頭看這張, 是相當受用的, 如果未來要將機器學習當作吃飯的工具, 強烈建議練習這些範例。

Pandas Cheat Sheet

- datacamp.com [網頁連結](#)
- 內容上似乎也蠻豐富的，喜歡這種風格的可以嘗試看看，在與官方懶人包之間可以做個選擇，雖然以部分使用者的觀點來說，官方的似乎比較易學易懂一點。

Python For Data Science Cheat Sheet

Pandas Basics

Learn Python for Data Science interactively at www.DataCamp.com



Pandas

The **Pandas** library is built on NumPy and provides easy-to-use data structures and data analysis tools for the Python programming language.



Use the following import convention:

```
>>> import pandas as pd
```

Pandas Data Structures

Series

A one-dimensional labeled array capable of holding any data type

| | |
|---|----|
| a | 3 |
| b | -5 |
| c | 7 |
| d | 4 |

```
>>> s = pd.Series([3, -5, 7, 4], index=['a', 'b', 'c', 'd'])
```

DataFrame

Columns

| | Country | Capital | Population |
|---|---------|-----------|------------|
| 0 | Belgium | Brussels | 11190846 |
| 1 | India | New Delhi | 1303171035 |
| 2 | Brazil | Brasilia | 207847528 |

A two-dimensional labeled data structure with columns of potentially different types

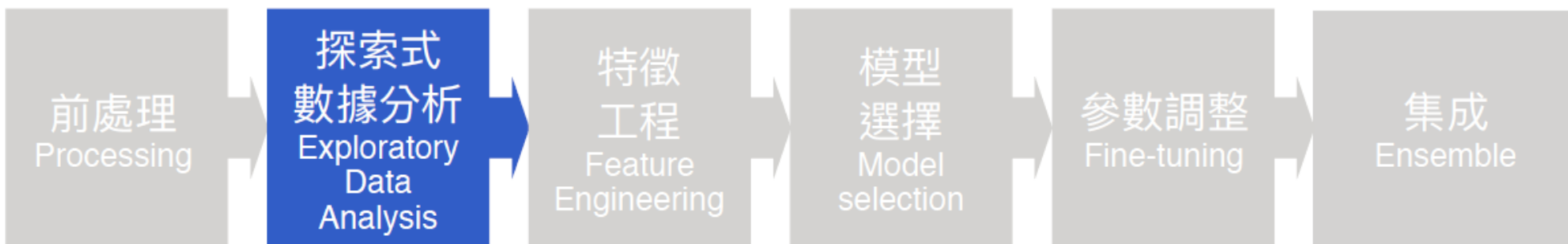
```
>>> data = {'Country': ['Belgium', 'India', 'Brazil'],  
           'Capital': ['Brussels', 'New Delhi', 'Brasilia'],  
           'Population': [11190846, 1303171035, 207847528]}
```

```
>>> df = pd.DataFrame(data,  
                      columns=['Country', 'Capital', 'Population'])
```

知識地圖 探索式數據分析 相關係數簡介

機器學習概論 Introduction of Machine Learning

監督式學習 Supervised Learning



非監督式學習 Unsupervised Learning



探索式數據分析 Exploratory Data Analysis (EDA)

統計值的視覺化

| | |
|-------|------|
| 相關係數 | 繪圖排版 |
| 核密度函數 | 常用圖形 |
| 離散化 | 模型體驗 |

本節重點

- 了解相關係數

Correlation Coefficient

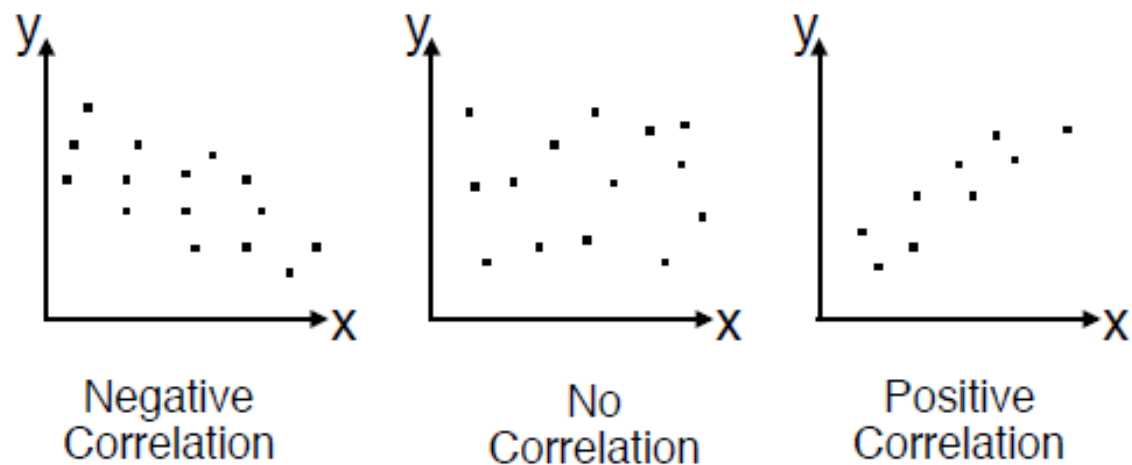
相關係數是其中一個常用來了解各個欄位與我們想要預測的目標之間的關係的指標。相關係數衡量兩個隨機變量之間線性關係的強度和方向。雖然不是表示變數之間關係的最好方法，但可以提供我們很直觀的了解。

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

Correlation Coefficient

相關係數是一個介於 $-1 \sim 1$ 之間的值，負值代表負相關，正值代表正相關，數值的大小代表相關性的強度。

- .00-.19：非常弱相關
- .20-.39：弱相關
- .40-.59：中度相關
- .60-.79：強相關
- .80-1.0：非常強相關



重點複習

基本的統計數值只能了解一個變數，如果我們想要了解兩個變數之間的線性關係時，相關係數是一個還不錯的簡單方法，能給出一個 $-1 \sim 1$ 之間的值來量化兩個變數之間的關係。



解題時間

It's Your Turn

- **Guess the Correlation 相對係數小遊戲** [網頁連結](#)

什麼!?! 還不夠直覺? 沒關係...

點開這個網站, 輸入你所認為的相關係數, 差太多會扣生命值, 命中的話生命值會增加, 挑戰自己看看扣完生命前能賺得多少金幣吧!!

