Day 14 資料清理數據前處理

# 相關係數簡介

出題教練 游為翔 / 杜靖愷

# 知識地圖 探索式數據分析 相關係數簡介

CUPOY

## 機器學習前處理

**監督式學習**
**Supervised Learning**

前處理 Processing → **探索式數據分析 Exploratory Data Analysis** → 特徵工程 Feature Engineering → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

**非監督式學習**
**Unsupervised Learning**

- 分群 Clustering
- 降維 Dimension Reduction

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

**統計值的視覺化**

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

# 本日知識點目標

了解相關係數

# Correlation Coefficient

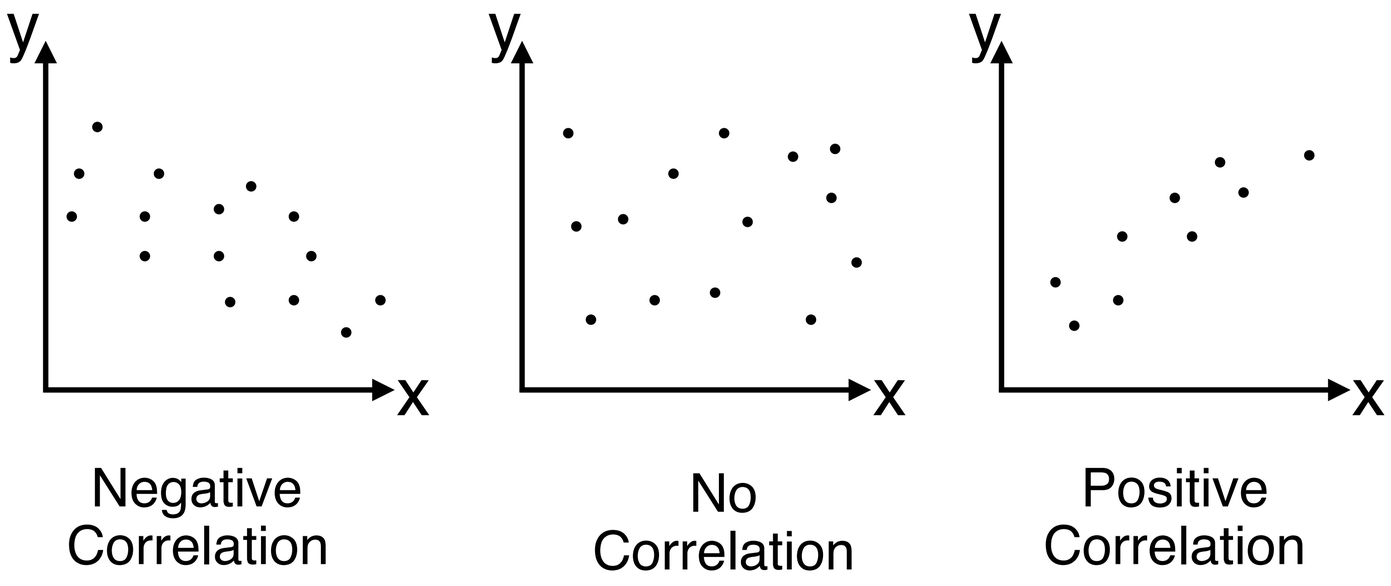
相關係數是其中一個常用來了解各個欄位與我們想要預測的目標之間的關係的指標。相關係數衡量兩個**隨機變量**之間線性關係的強度和方向。雖然不是表示變數之間關係的最好方法，但可以提供我們很直觀的了解。

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

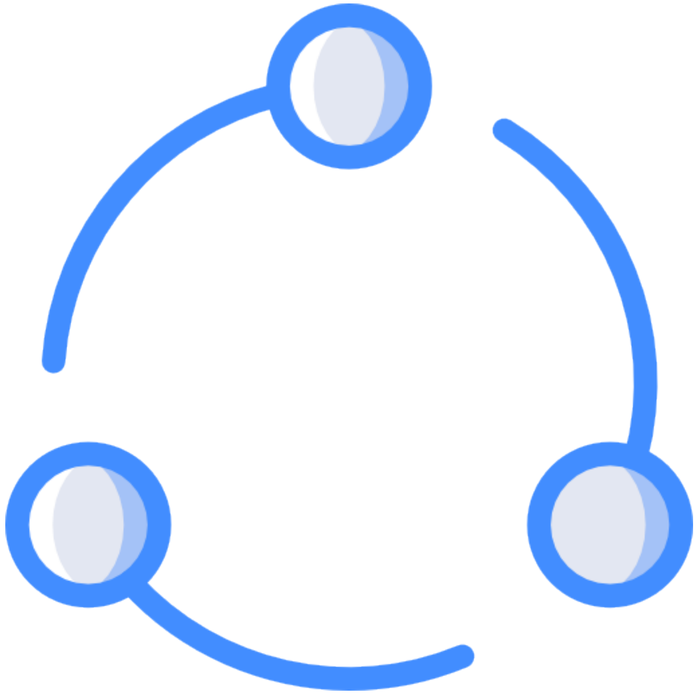
# Correlation Coefficient

**相關係數是一個介於 -1～1 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



# 重要知識點複習



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 -1~1 之間的值來量化兩個變數之間的關係。

# 延伸閱讀

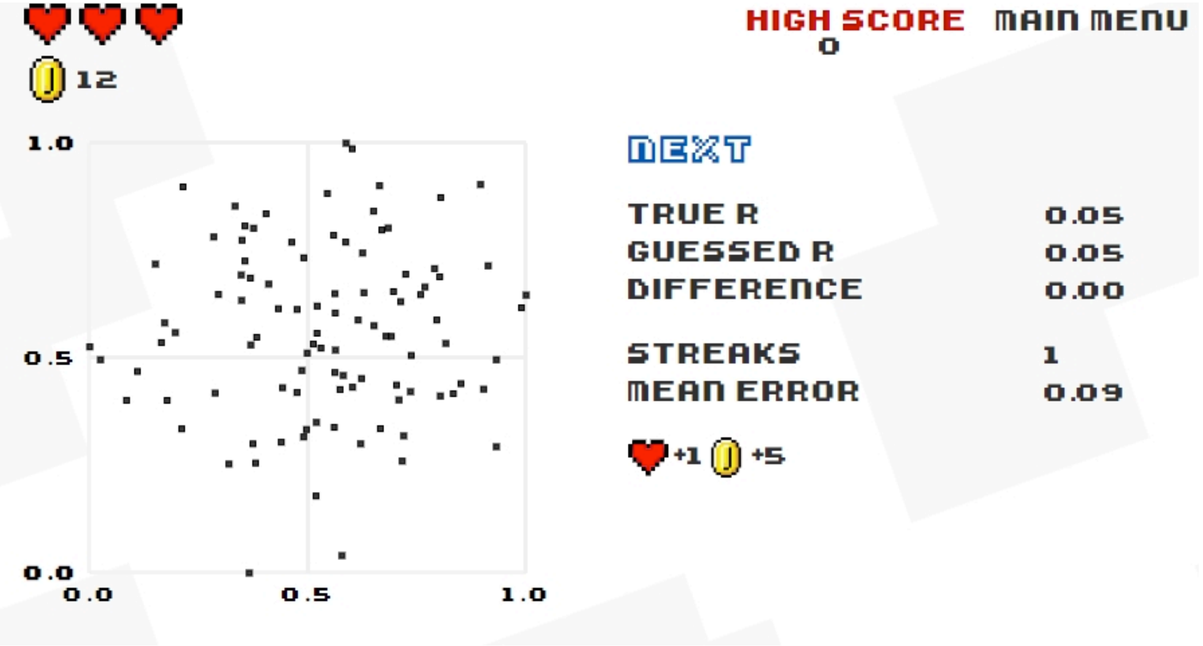
除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度，建議您解完每日題目後，若有多餘時間，可再補充延伸閱讀文章內容。

# 推薦延伸閱讀

Guess the Correlation 相對係數小遊戲 [網頁連結]

什麼!? 覺得自己對相關係數還不夠直覺？ 沒關係...

- 點開上述提供網頁連結，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!!



# 解題時間

# It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題