

# 天氣與人工智慧 機器學習實作

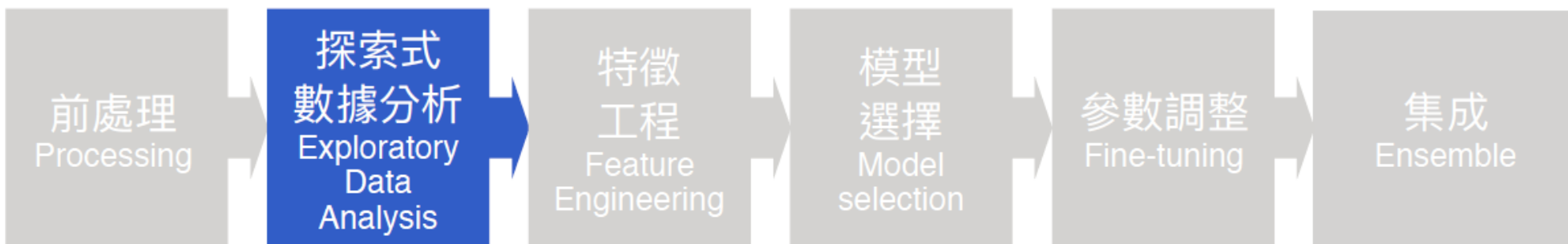
資料清理數據前處理 - 4

2020/11/25

# 知識地圖 探索式數據分析 相關係數的EDA

## 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

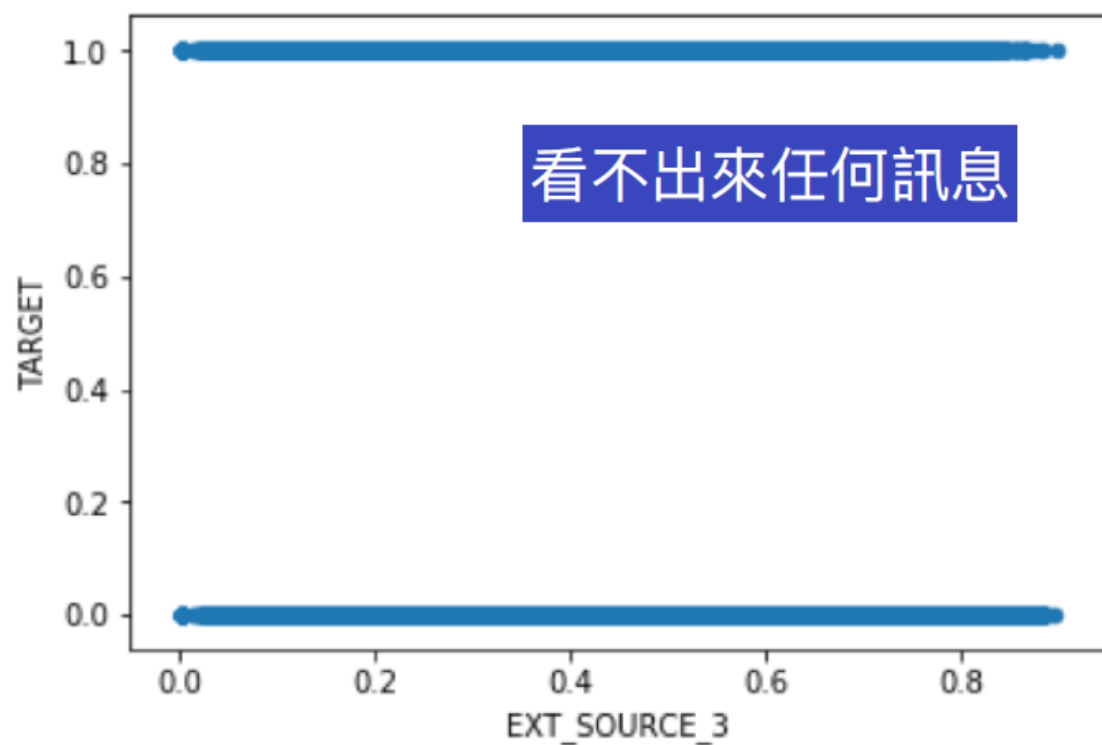
- 可以用相關係數來迅速找到和預測目標最有線性關係的變數
- 相關係數通常搭配散布圖來一起瞭解預測目標和變數的關係

# 相關係數實作

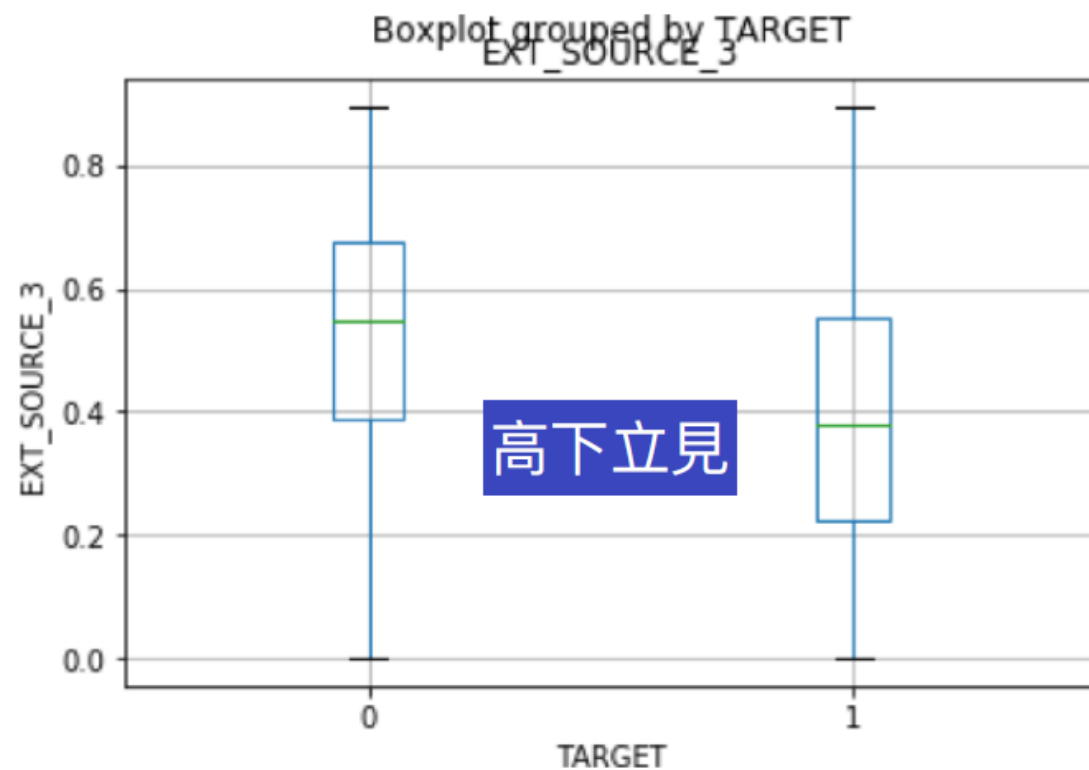
- 列出目標與所有欄位之間相關係數，數值最大及最小各 15 個
- 利用相關係數的結果觀察有興趣的欄位與目標或其他欄位的相關係數，並嘗試找出有趣的訊息
  - 最好的方式當然是畫圖，舉例來說，我們知道 EXT\_SOURCE\_3 這個欄位和 TARGET 之間的相關係數是 -0.178919 (在這個資料集已經是最負的)，那我們可以以 EXT\_SOURCE\_3 為 X 軸，TARGET 為 Y 軸，把資料畫出來。

# Tips: 遇到 $y$ 的本質不是連續數值時

直接以原始數值繪圖

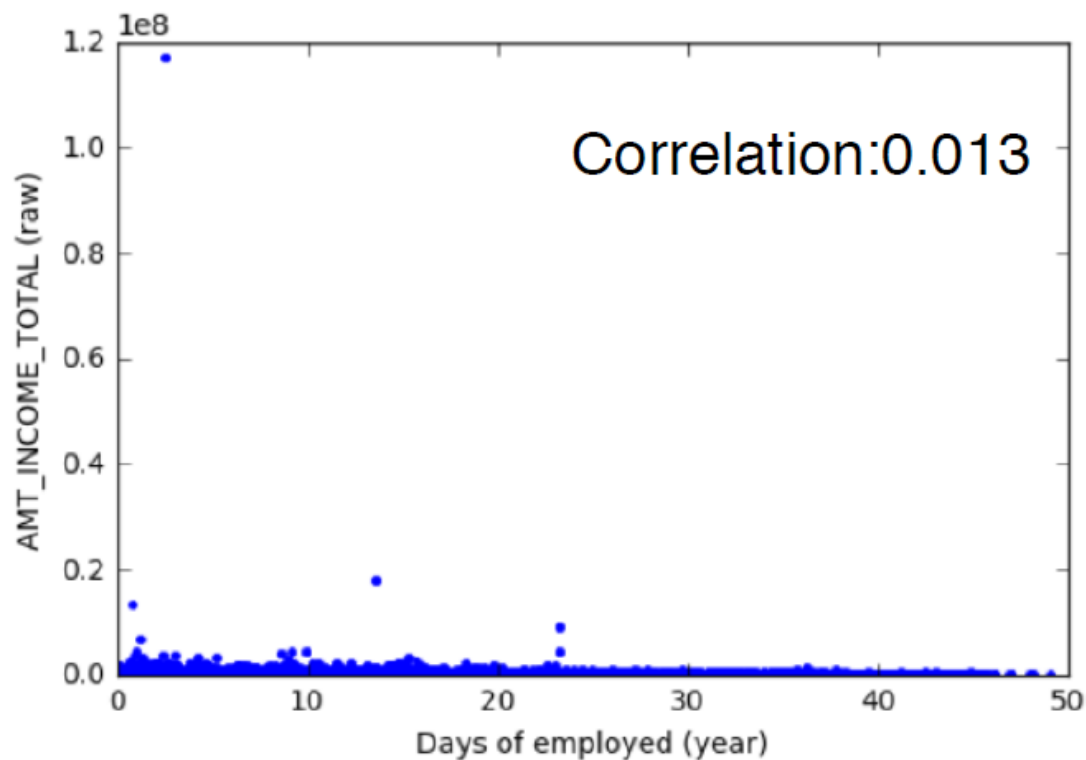


可以換一個角度來看

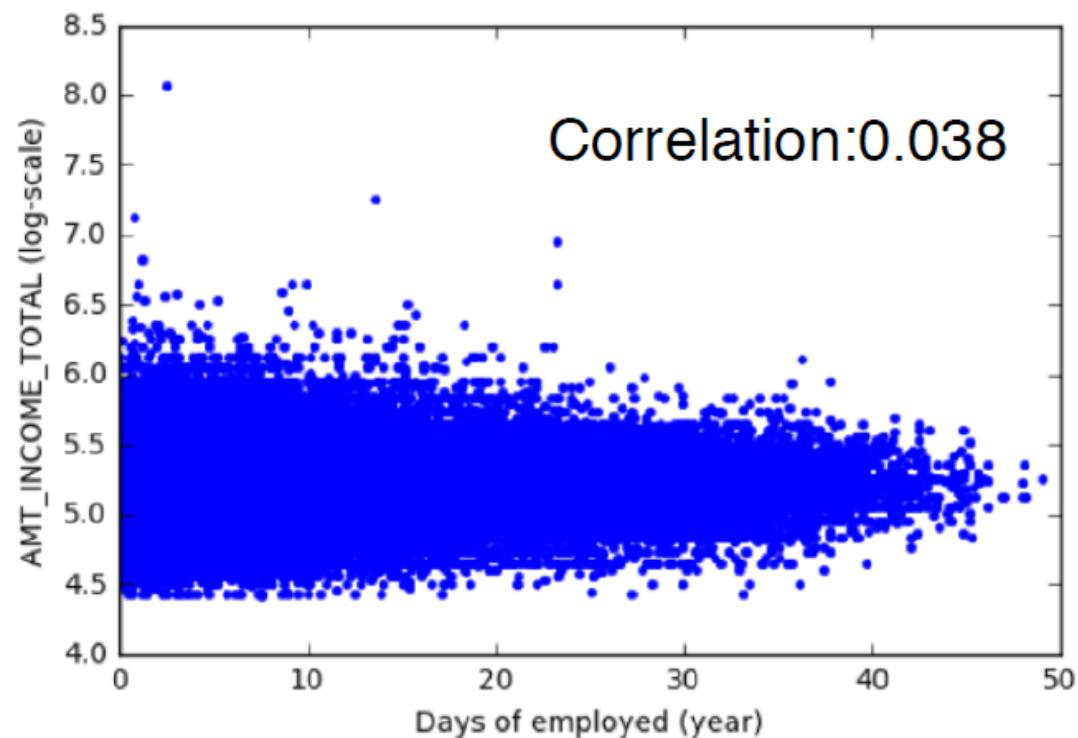


# Tips: 檢視不同數值範圍的變數

直接以原始數值繪圖



將 Y 軸轉換 (log-scale)





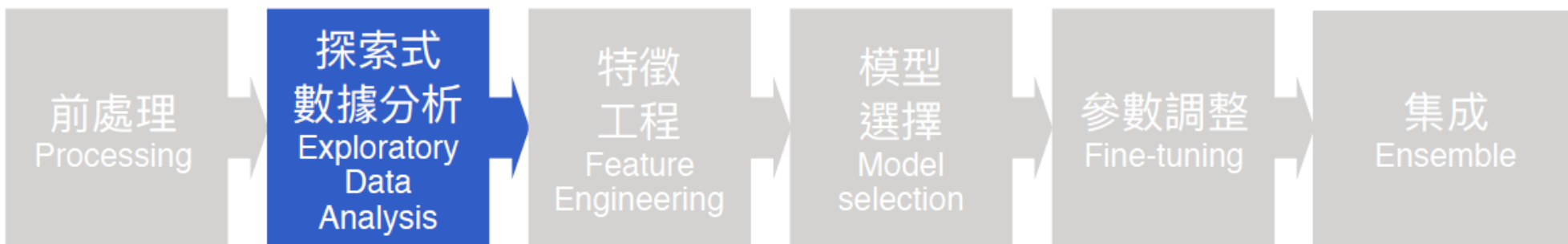
解題時間

It's Your Turn

# 知識地圖 探索式數據分析 核密度函數

## 機器學習概論 Introduction of Machine Learning

### 監督式學習 Supervised Learning

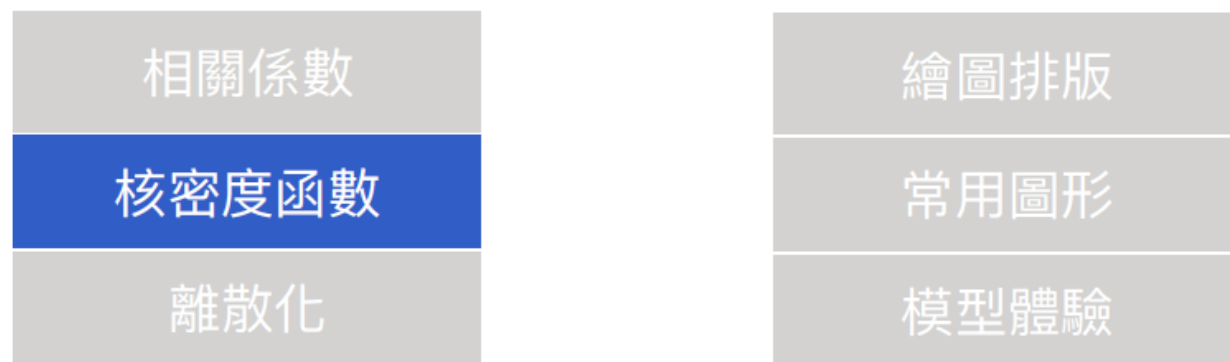


### 非監督式學習 Unsupervised Learning



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

### 統計值的視覺化



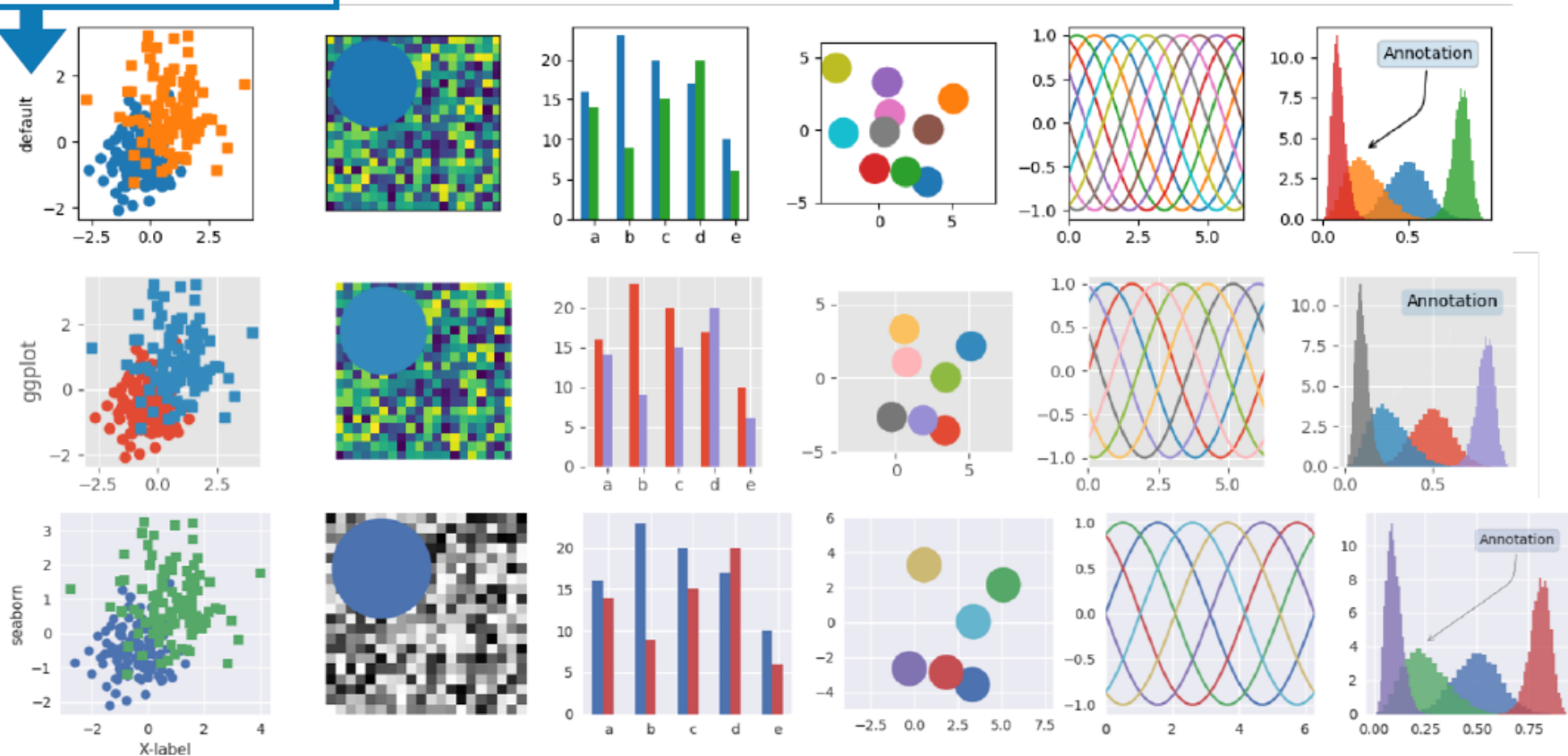


# 繪圖風格

```
plt.style.use('default') # 不需設定就會使用預設  
plt.style.use('ggplot')  
plt.style.use('seaborn') # 或採用 seaborn 套件繪圖
```

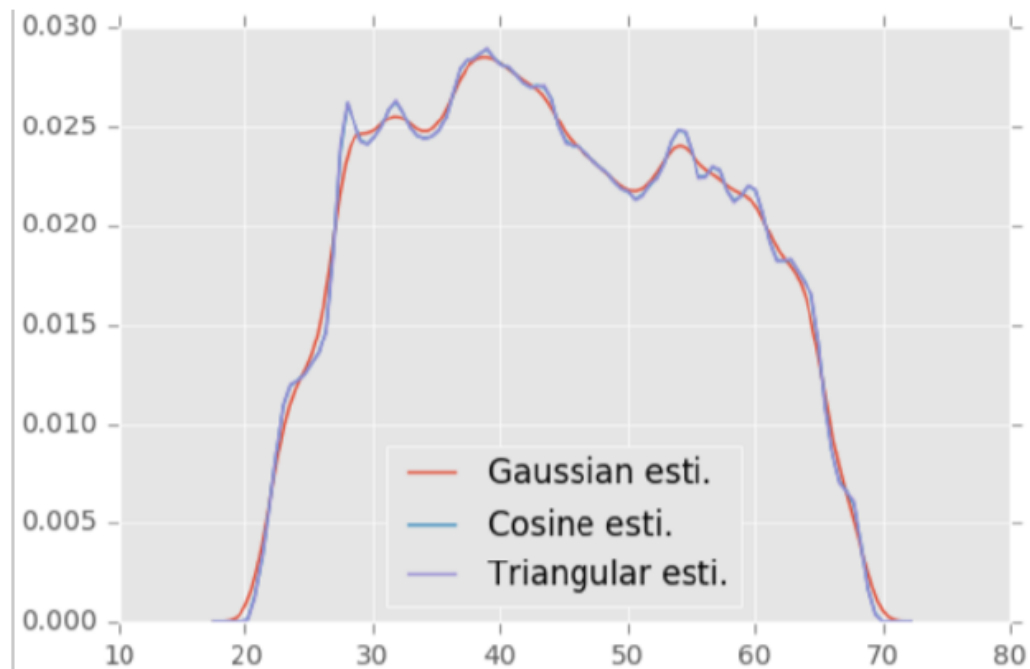
## 轉變繪圖風格的目的

用已經被設計過的風格，  
讓觀看者更清楚明瞭，  
包含色彩選擇、線條、  
樣式等。



# Kernel Density Estimation (KDE)

不同 kernel function 的結果



1

採用無母數方法畫出一個觀察變數的機率密度函數  
某個  $X$  出現的機率為何

2

**Density plot 的特性**

- 歸一：線下面積和為 1
- 對稱： $K(-u) = K(u)$

3

**常用的 Kernel function**

- Gaussian (Normal dist)
- Cosine



解題時間

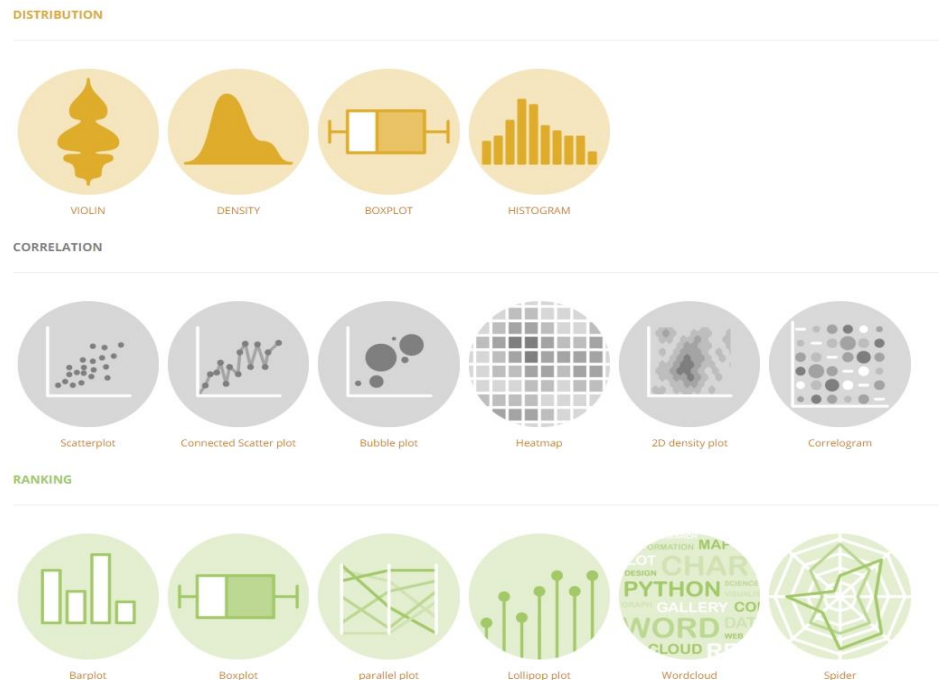
It's Your Turn

# 參考資料

- 繪圖靈感資源參考

- 1. Python Graph Gallery (圖表參考)

這裡整合了 Python 許多繪圖函數的寫法, 同學可以依據自己的喜好與資料形式, 挑選適合的圖形寫作, 並不需要全部看懂, 只需要當成查詢用的工具手冊即可 [網頁連結](#)



# 參考資料

- 繪圖靈感資源參考

- 2. R Graph Gallery

這裡整合了 R 許多繪圖函數的寫法, 與上面的網站是相關網站, 如果較擅長使用 R 做資料科學, 可以先從這邊參考 [網頁連結](#)

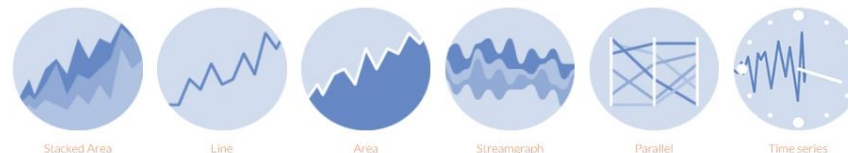
## Rankings



## Part of a whole



## Evolution



# 參考資料

- 繪圖靈感資源參考

- 2. R Graph Gallery

這裡整合了 R 許多繪圖函數的寫法, 與上面的網站是相關網站, 如果較擅長使用 R 做資料科學, 可以先從這邊參考 [網頁連結](#)

## Rankings



Barplot



Spider / Radar



Wordcloud



Parallel



Lollipop / Stem



Circular Barplot

## Part of a whole



Treemap



Dendrogram



Venn Diagram



Stacked Bar



Pie Chart



Doughnut



Circular Packing

## Evolution



Stacked Area



Line



Area



Streamgraph



Parallel



Time series

# 參考資料

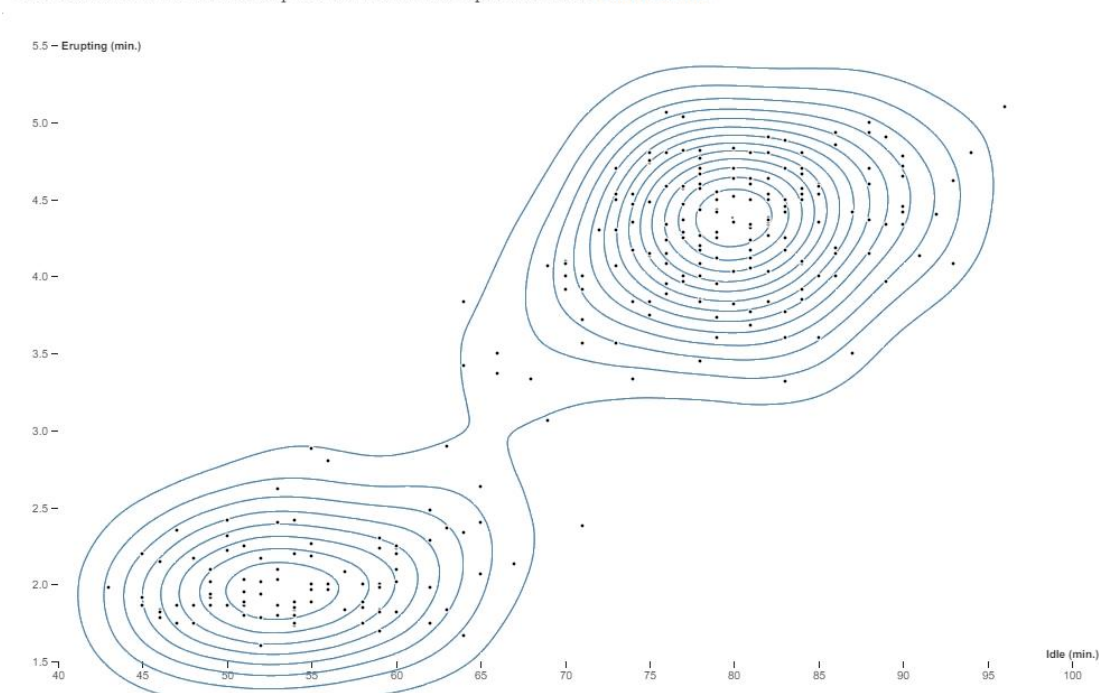
- 繪圖靈感資源參考

- 3. R Graph Gallery (Interactive plot, 互動圖)

可以由 R 語言繪製出的互動圖表, 也是提供同學查詢之用 [網頁連結](#)

## Density Contours

This chart shows the relationship between idle and eruption times for [Old Faithful](#).





## 參考資料

## 繪圖靈感資源參考

- **4. D3.js**

D3.js 是知名的 Javascript 網頁繪圖套件, 如果您是前端工程師, 熟練D3.js 將可使您的網頁圖表豐富起來[網頁連結](#)





# 補充資料

- 核密度估計基礎 - 1 [網頁連結](#)

- 核密度估計基礎 - 2 [網頁連結](#)

- 如果您是對核密度估計函數 ( Kernel Density Estimation, KDE ) 理論有更多的求知慾, 歡迎來到上述兩個網站, 裡面詳盡的解說可提供您查閱

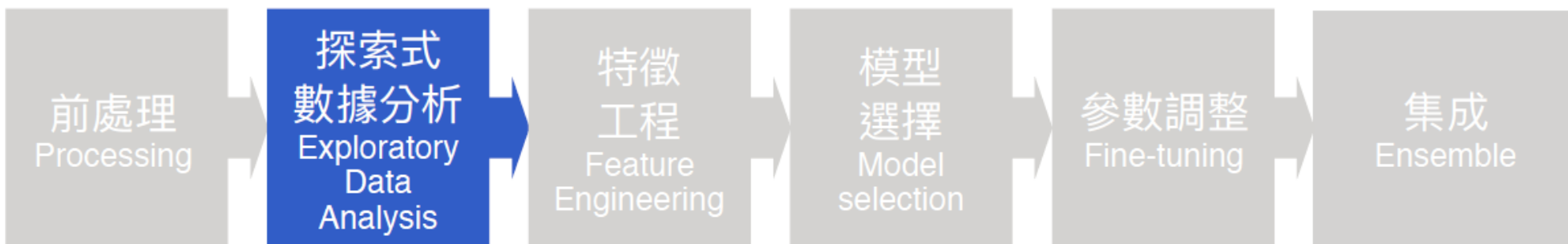
- Seaborn 套件如果發生錯誤的解決辦法 [網頁連結](#)

使用 Seaborn 時, 萬一出現問題 DLL load failed 怎麼辦? 這段討論提供您解決之道

# 知識地圖 探索式數據分析 離散化與EDA

## 機器學習概論 Introduction of Machine Learning

### 監督式學習 Supervised Learning

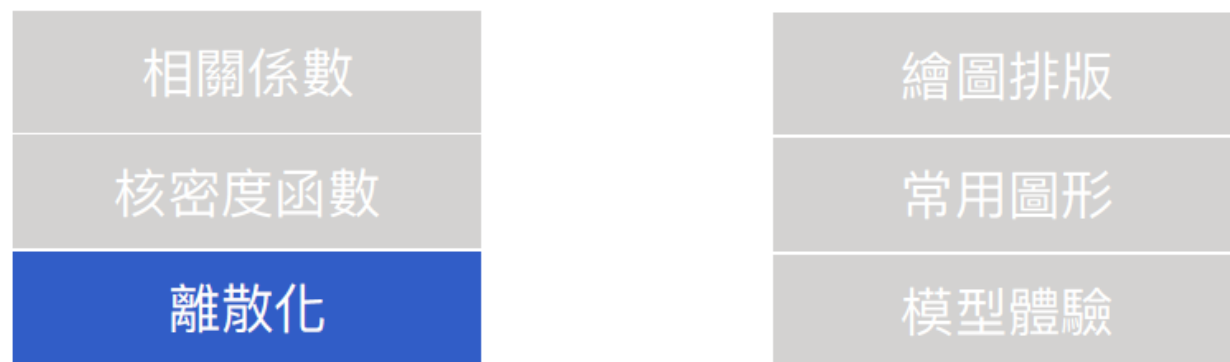


### 非監督式學習 Unsupervised Learning



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

### 統計值的視覺化



# 本節重點

- 了解離散化連續數值的意義以及方法

# 連續型變數離散化

---

## Goal

- 變得更簡單 (可能性變少了)
  - 假設年齡 0-99 (100 種可能性) >> 每 10 歲一組 (10 種可能性)
- 離散化的變數較穩定，假設年齡 > 30 是 1，否則 0。  
如果沒有離散化，outlier 「年齡 300 歲」 會給模型帶來很大的干擾。

## 關鍵點



- 組的數量
  - 一樣以年齡為例子，每 10 歲一組就會有 10 組
- 組的寬度
  - 一組的寬度是 10 歲

# 連續型變數離散化

---

## 主要的方法

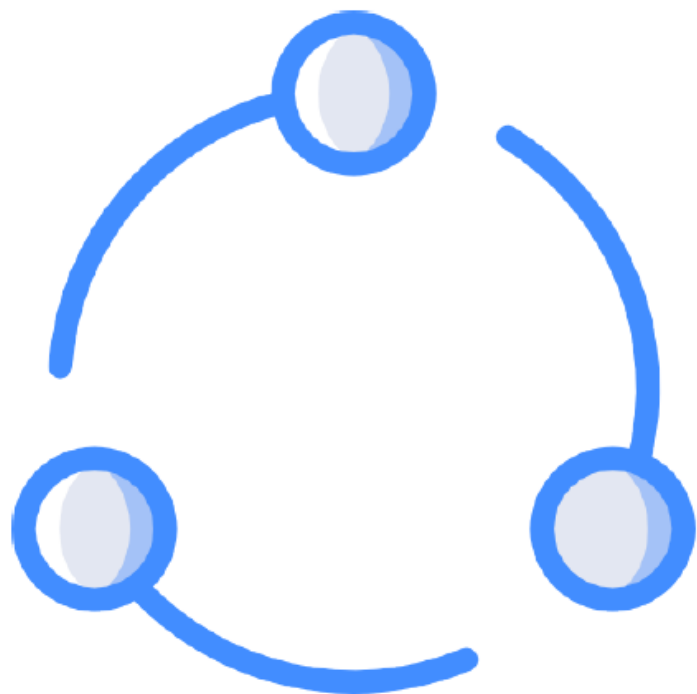
- **等寬劃分**：按照相同寬度將資料分成幾等份。缺點是受到異常值的影響比較大。
- **等頻劃分**：將資料分成幾等份，每等份資料裡面的個數是一樣的。
- **聚類劃分**：使用聚類演算法將資料聚成幾類，每一個類為一個劃分。



除了以上的主要方法，也會因需求而需要自己定義離散化的方式，如何離散化是一門學問！

# 重要知識點複習

---



- 離散化的目的是讓事情變簡單、減少 outlier 對分析以及訓練模型的影響
- 主要的方法是等寬劃分 (對應 pandas 中的 cut) 以及等頻劃分 (對應 pandas 中的 qcut)
- 可以依實際需求來自己定義離散化的方式



解題時間

It's Your Turn

# 連續特徵的離散化：在什麼情況下可以獲得更好的效果(知乎)

- 這個網頁是個討論串，經由幾個網友的討論與補充，很好地說明了離散化的理由：儲存空間小，計算快，降低異常干擾與過擬合(overfitting)的風險，主要想請同學參考**第1位**的回答，至於其他的討論則請同學參考即可。
- [網頁連結](#)