Day 5

資料清理數據前處理

EDA之資料分布

出題教練

游為翔 / 杜靖愷

知識地圖 探索式數據分析 EDA 資料分布

CUPOY

機器學習概論 Introduction of Machine Learning

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

了解如何通過基本的統計數值以及畫圖來了解資料

EDA - 統計量化的方式？

以單變量分析來說，量化的分析方式可包含

計算集中趨勢

- 平均值 Mean
- 中位數 Median
- 眾數 Mode

計算資料分散程度

- 最小值 Min
- 最大值 Max
- 範圍 Range
- 四分位差 Quartiles
- 變異數 Variance
- 標準差 Standard deviation

基本上使用上述統計特徵就可以讓我們初步了解資料的樣子，並且觀察是否有異樣

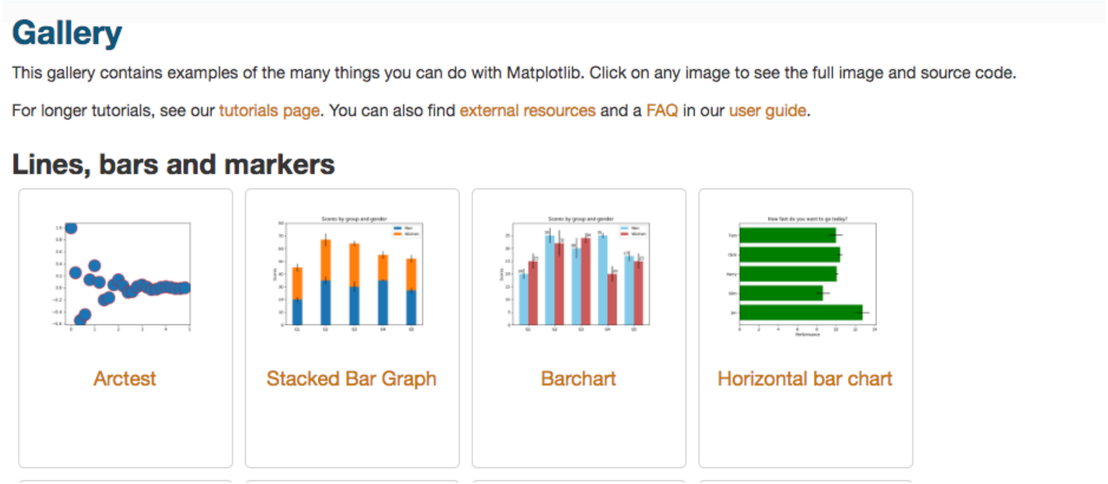
EDA視覺化的方式？

CUPOY

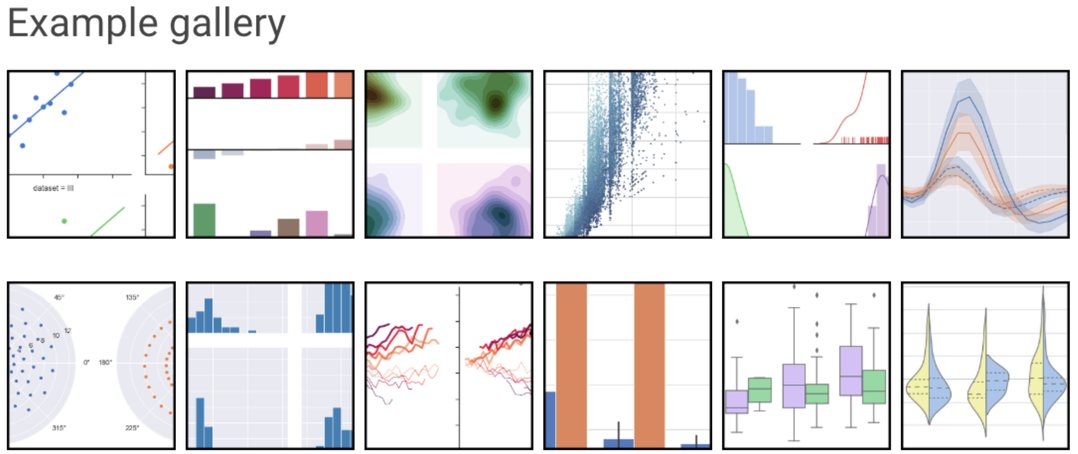
有句話「一畫勝千言」，除了數字，視覺化的方式也是一種很好觀察資料分佈的方式，可參考 python 中常用的視覺化套件

畫圖沒靈感的時候可以到這兩個套件的範例網頁逛逛！

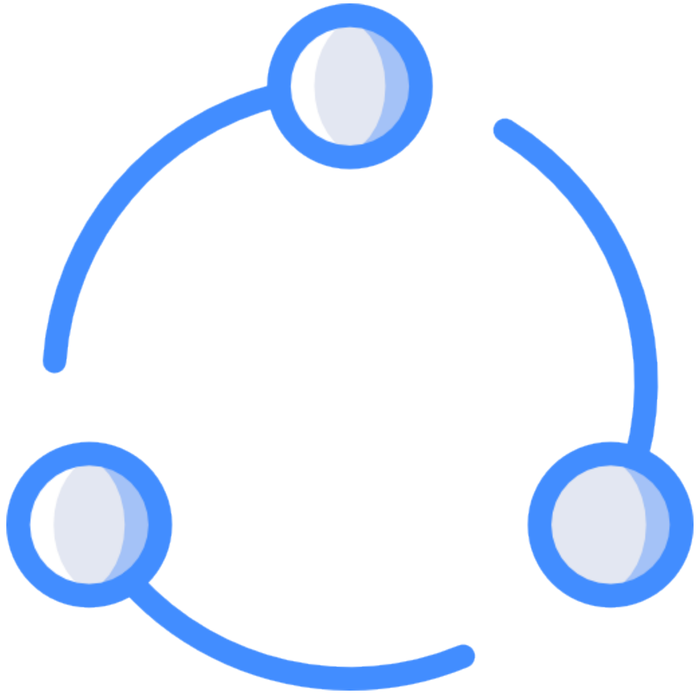
matplotlib



seaborn



重要知識點複習



- 資料大部分時候都是非常多的，我們沒辦法用眼睛一筆一筆都看完，平均值、標準差、最大最小值等統計數值能幫助我們迅速對資料有初步的了解。
- 了解統計數值後，把資料的圖畫出來除了能夠更全面地了解資料，也能幫我們快速觀察到異常的地方
- pandas 有許多已經寫好用來做以上這些觀察的函數，熟悉這些函數的使用能加速觀察資料的過程

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題