巨匠線上真人

Python 資料科學應用開發

# 第八堂：資料探索（Data Exploration）

# 同學，歡迎你參加本課程

- 請關閉你的FB 、Line等溝通工具，以免影響你上課。
- 考量頻寬、雜音，請預設關閉攝影機、麥克風，若有需要再打開。
- 隨時準備好，老師會呼叫你的名字進行互動，鼓勵用麥克風提問。
- 如果有緊急事情，你必需離開線上教室，請用聊天室私訊給老師，以免老師癡癡呼喚你的名字。
- 軟體安裝請在上課前安裝完成，未完成的同學，請盡快進行安裝。

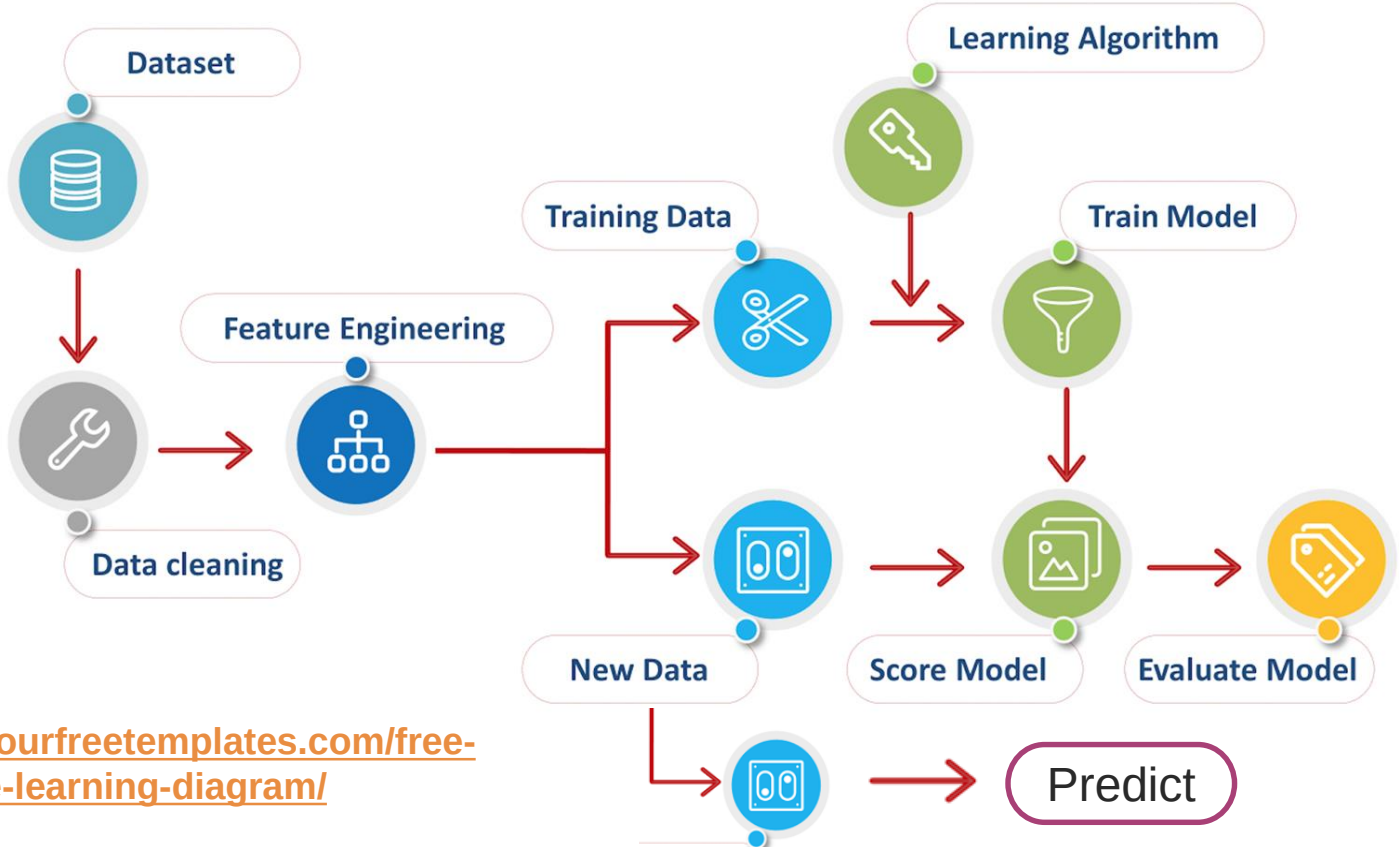
# 課程檔案下載


巨匠電腦線上真人
開課查詢
免費體驗專區
課程總覽
專業師資
學員專區
講師專區
最新消息
您好!
登出
點數卡產品兌換
APCS檢測專區
公告專區
我的課表
IT真人課程劃位
電腦分校課程劃位
外語真人課程劃
美語分校課程劃
取消劃位
課程檔案下載
上課權益查詢
教學平台測試
學習諮詢
常見問題
個資維護
忘記密碼
登出
課程檔案下載
程式語言好難學?
那是因為
你還沒學過Python!
線上老師LIVE直播教學．搶先看
巨匠電腦真人課程

# ZOOM 學員操作說明


www.pcschool.com.tw
您正在觀看蘇世宗的螢幕
查看選項
原始大小
請求遠端控制
共同註記
退出全螢幕
5 查看選項/共同註記/筆（連連看）
滑鼠
文字
筆
箭頭
擦除
格式
撤銷
重做
清除
與會者 (15)
查找參會者
張齡月 (我)
婷婷
解除靜音
舉手
2 共享螢幕（指導演練；點評作品）
老師須先停止共享螢幕
才能請學生共享螢幕
1 聊天
解除靜音
啟動視訊
邀請
與會者
共享螢幕
聊天
錄影
4 解除靜音
3 與會者/舉手

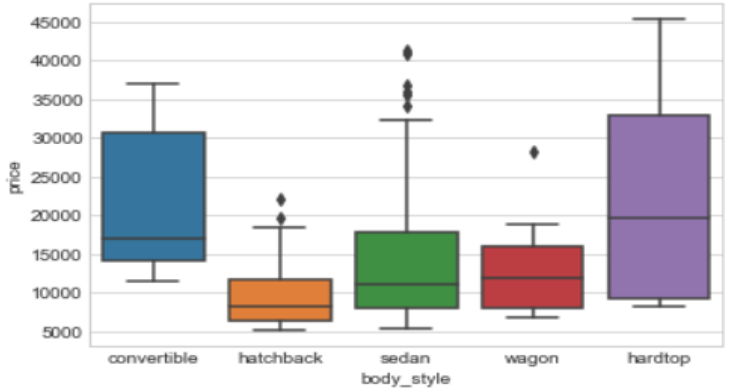
# 機器學習流程




https://yourfreetemplates.com/free-machine-learning-diagram/

# 資料探索

- 資料預處理（Pre-processing）
  - 資料探索與分析（Exploratory Data Analysis；EDA）：分析資料集，以了解資料的主要特性，通常採視覺化判斷
  - Data Clean：資料格式統一、代碼編製、遺漏值（Missing Value）、重複記錄處理



# 課程內容

## 迴歸之資料探索與分析

- 概念介紹
- 實作

## 分類之資料探索與分析

- 概念介紹
- 實作

# 課程內容

## 迴歸之資料探索與分析

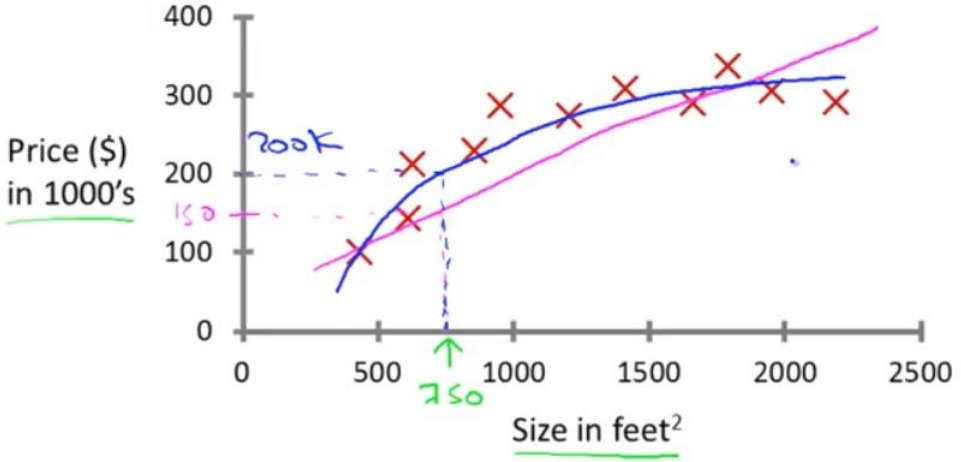
- 概念介紹
- 實作

## 分類之資料探索與分析

- 概念介紹
- 實作

# 迴歸（Regression）

Housing price prediction.




Supervised Learning

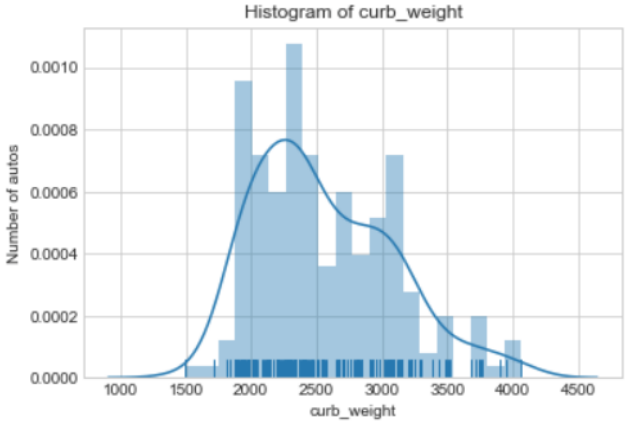
"right answers" given

# 目標

- 探索 automobile pricing 的影響因素
- Lab (DAT275x)
  - Module2-275 / VisualizingDataForRegression.ipynb
- 流程見下頁

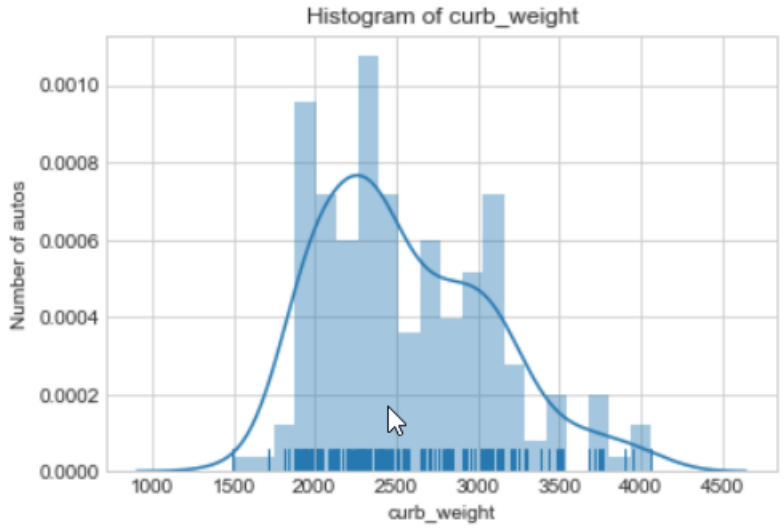
# 處理流程

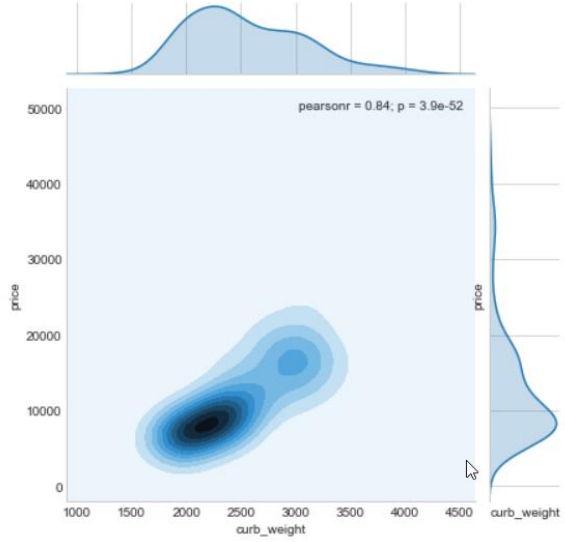
- 資料彙總與處理（Summarizing and manipulating data）
- Developing multiple views of complex data
  - Kernel density plot（KDE）：平滑板的直方圖（Histogram）
  - Combine histograms and kdes
- Matplotlib, Pandas plotting and Seaborn
- 單一變數繪圖（univariate plot）
- 雙變數繪圖（two dimensional plot）
- 美化（Aesthetics）
- Facetted plotting：visualize higher dimensional data



# Seaborn 圖形種類 (1)

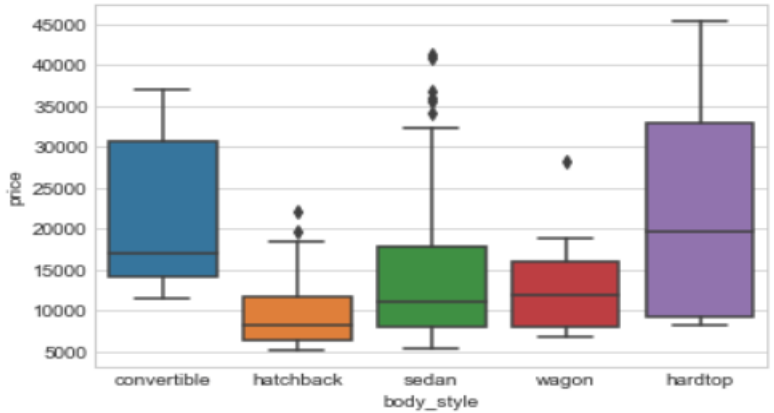
- Kernel density plot：sns.distplot
- Countour plot + Distribution plot：sns.jointplot

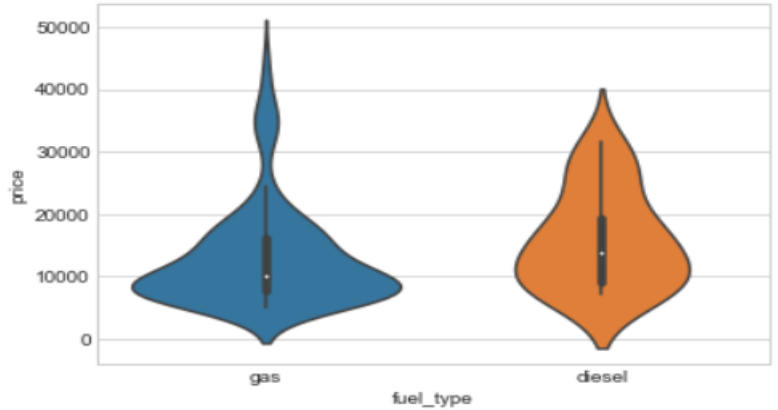


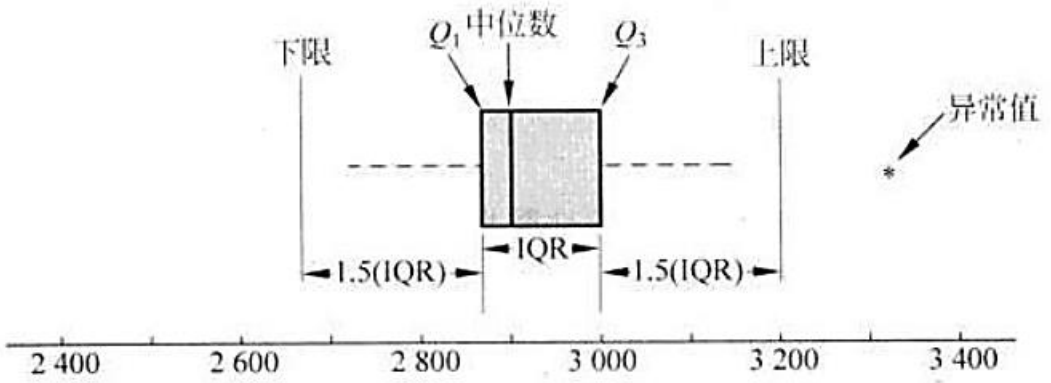
# Seaborn 圖形種類 (2)

- Box plot：sns.boxplot，highlight the quartiles of a distribution.
- Violine plot：sns.violinplot，two back to back KDE curves are used to show the density estimate.



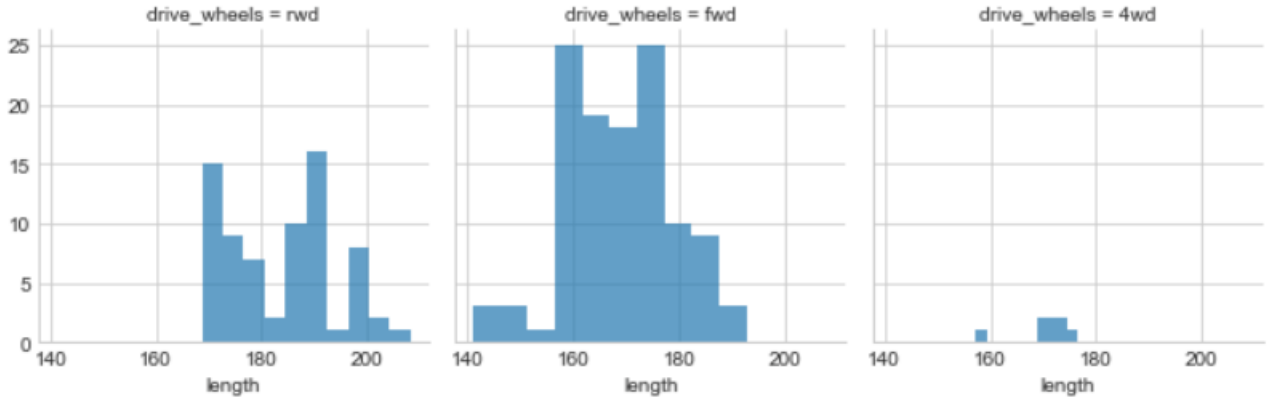



# 箱線圖定義


下限
$Q_1$ 中位數
$Q_3$
上限
異常值
1.5(IQR)
IQR
1.5(IQR)
2 400
2 600
2 800
3 000
3 200
3 400

# Seaborn 圖形種類 (3)

- Conditioned plot：繪製各種類別對應的直方圖
  - FacetGrid：sns.FacetGrid
  - map：sns.map



# 問題(1)

- 直方圖透露甚麼訊息？

1) There are more cars with high miles per gallon than low miles per gallon
2) Most of the cars have larger than 3000 curb weight
3) There are only a few cars with engine size lower than 150
4) Most of the cars cost less than 30,000

# 問題(2)

- 從 kde 圖觀察，哪一個特徵與價格的分配函數最相似？
  1) curb_weight
  2) engine_size
  3) city_mpg

# 問題(3)

- 從散佈圖觀察，哪三個敘述是對的？

1) Both gas and diesel turbo cars are generally more expensive than standard cars.
2) Turbo cars appear to have worse city_mpg at a given price point than standard cars.
3) Standard cars generally has diesel engine.
4) Turbo cars have greater horsepower at a given price point.

# 問題(4)

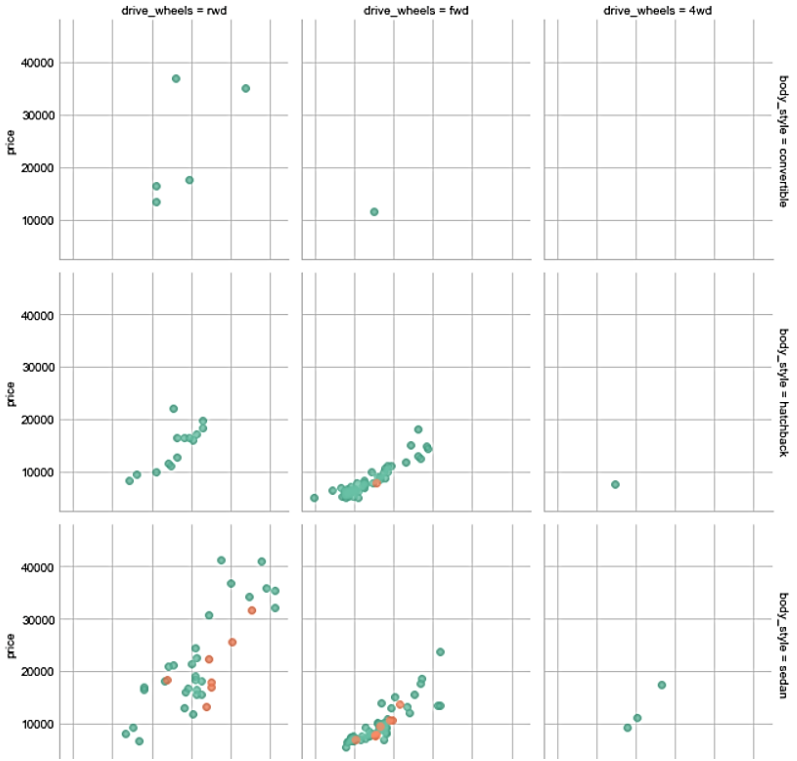
- 從 conditioned plots 觀察，哪三個關聯是對的？

1) The distribution of the values generally increases for length and curb_weight for real wheel drive (rwd) cars, with the values for 4 wheel drive (4wd) and real wheel drive (rwd) overlapping.
2) Generally, 4wd cars have the highest engine size.
3) Cars with fwd have the highest city_mpg, whereas, 4wd and rwd in a similar range.
4) Generally, 4wd cars have the lowest price, with rwd cars having the widest range.

# 問題(5)

- 哪一個組合會產生空白圖？

1) fwd and convertible
2) 4wd and hatchback
3) fwd and hardtop
4) 4wd and convertible



# 課程內容

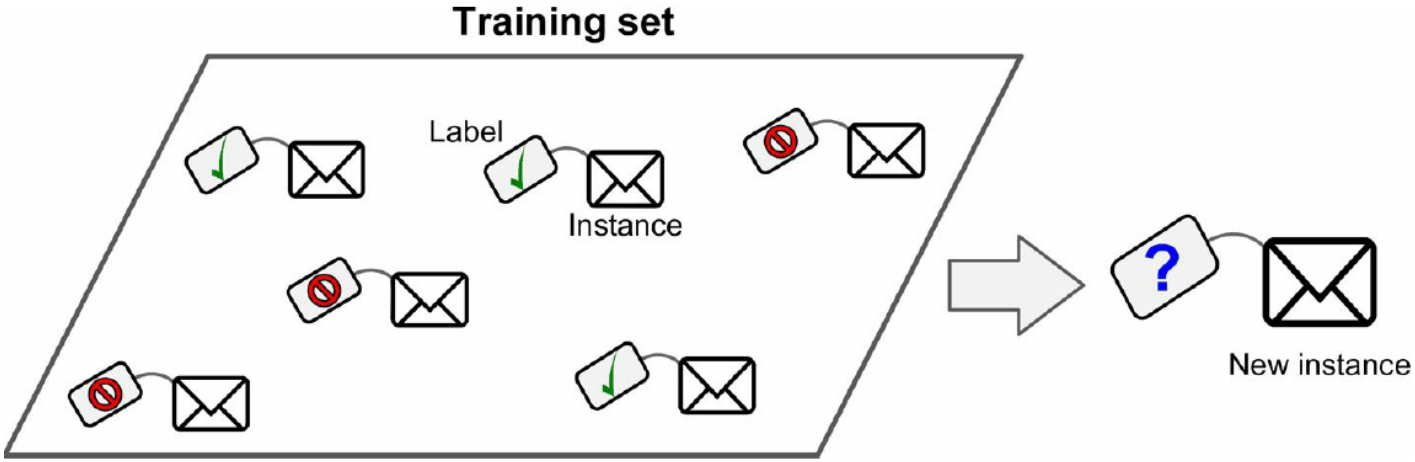
## 迴歸之資料探索與分析

- 概念介紹
- 實作

## 分類之資料探索與分析

- 概念介紹
- 實作

# 分類（Classification）



# 目標

- 探索 German bank credit 的影響因素
- Lab (DAT275x)
  - Module2-275 / VisualizingDataForClassification.ipynb
- 流程見下頁

# 處理流程

- Load and prepare the data set
- Examine classes and class imbalance
- Visualize class separation by numeric features
- Visualize class separation by categorical features

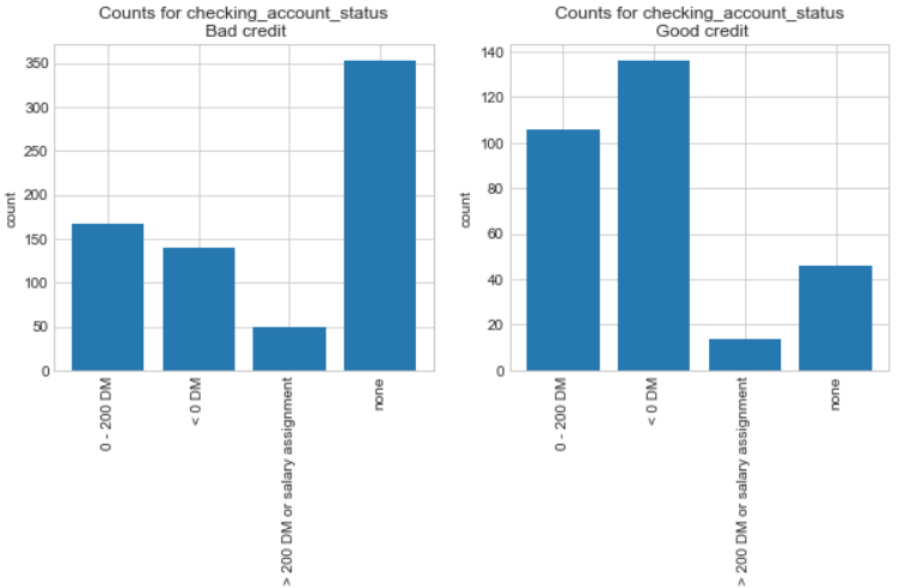
# 問題(1)

- 從圖形觀察，哪兩個特徵可區分信貸的好壞？
  1) payment_pcnt_income
  2) number_loans
  3) age
  4) loan_amount

# 問題(2)

- 從圖形觀察，哪一個特徵可區分信貸的好壞？

1) foreign_worker
2) telephone
3) job_category
4) checking_account_status



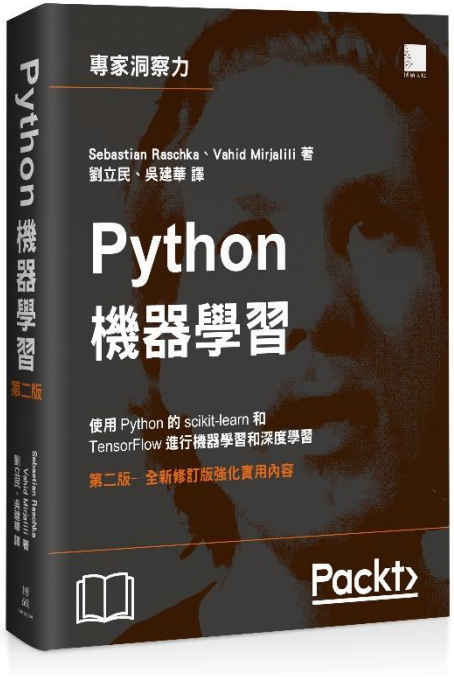
# 作品：分析鐵達尼資料集

- 資料集：鐵達尼（Titanic）
- import seaborn as sns
- titanic = sns.load_dataset("titanic")
- 參考

  Analyzing a dataset — the Step1 of Machine Learning which often gets overlooked

展示

# 參考用書




- 書名：Python機器學習（第二版）

  http://www.drmaster.com.tw/bookinfo.asp?BookID=MP11804
- 作者：Sebastian Raschka, Vahid Mirjalili ISBN
- 譯者：劉立民、吳建華
- 出版社：博碩

# 問卷

http://www.pcschoolonline.com.tw


www.pcschoolonline.com.tw