



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

The screenshot shows the homepage of the Juei Computer Online Live website. The header is orange with navigation links: 巨匠電腦線上真人 (Juei Computer Online Live), 開課查詢 (Class Inquiry), 免費體驗專區 (Free Experience Area), 課程總覽 (Course Overview), 專業師資 (Professional Faculty), 學員專區 (Student Area), 講師專區 (Instructor Area), and 最新消息 (Latest News). There are also social media icons for 360, Facebook, and YouTube. A user is logged in as '您好!' with a '登出' (Logout) button. The main banner features the text '程式語言好難學?' (Programming Language is so hard to learn?), '那是因為你還沒學過Python!' (That's because you haven't learned Python!), and '(線上老師 LIVE 直播教學 · 搶先看)' (Online Teacher LIVE Streaming Teaching · Preview). A dropdown menu is open from the '學員專區' (Student Area) link, showing options like '點數卡產品兌換' (Points Card Product Exchange), 'APCS檢測專區' (APCS Detection Special Area), '公告專區' (Announcement Special Area), '我的課表' (My Class Schedule), 'IT真人課程劃位' (IT Live Course Seating), '電腦分校課程劃位' (Computer Branch Course Seating), '外語真人課程劃位' (Foreign Language Live Course Seating), '美語分校課程劃位' (American English Branch Course Seating), '取消劃位' (Cancel Seating), '課程檔案下載' (Course Archive Download - highlighted), '上課權益查詢' (Classroom Privilege Inquiry), '教學平台測試' (Teaching Platform Test), '學習諮詢' (Learning Consultation), '常見問題' (Frequently Asked Questions), '個資維護' (Personal Information Maintenance), '忘記密碼' (Forgot Password), and '登出' (Logout). An orange callout bubble points to the '課程檔案下載' option with the text '課程檔案下載'. The background of the banner has a blue and purple digital theme with circuit patterns and a clock face showing 98% and 54%.

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  
外語真人課程劃位  
美語分校課程劃位  
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

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

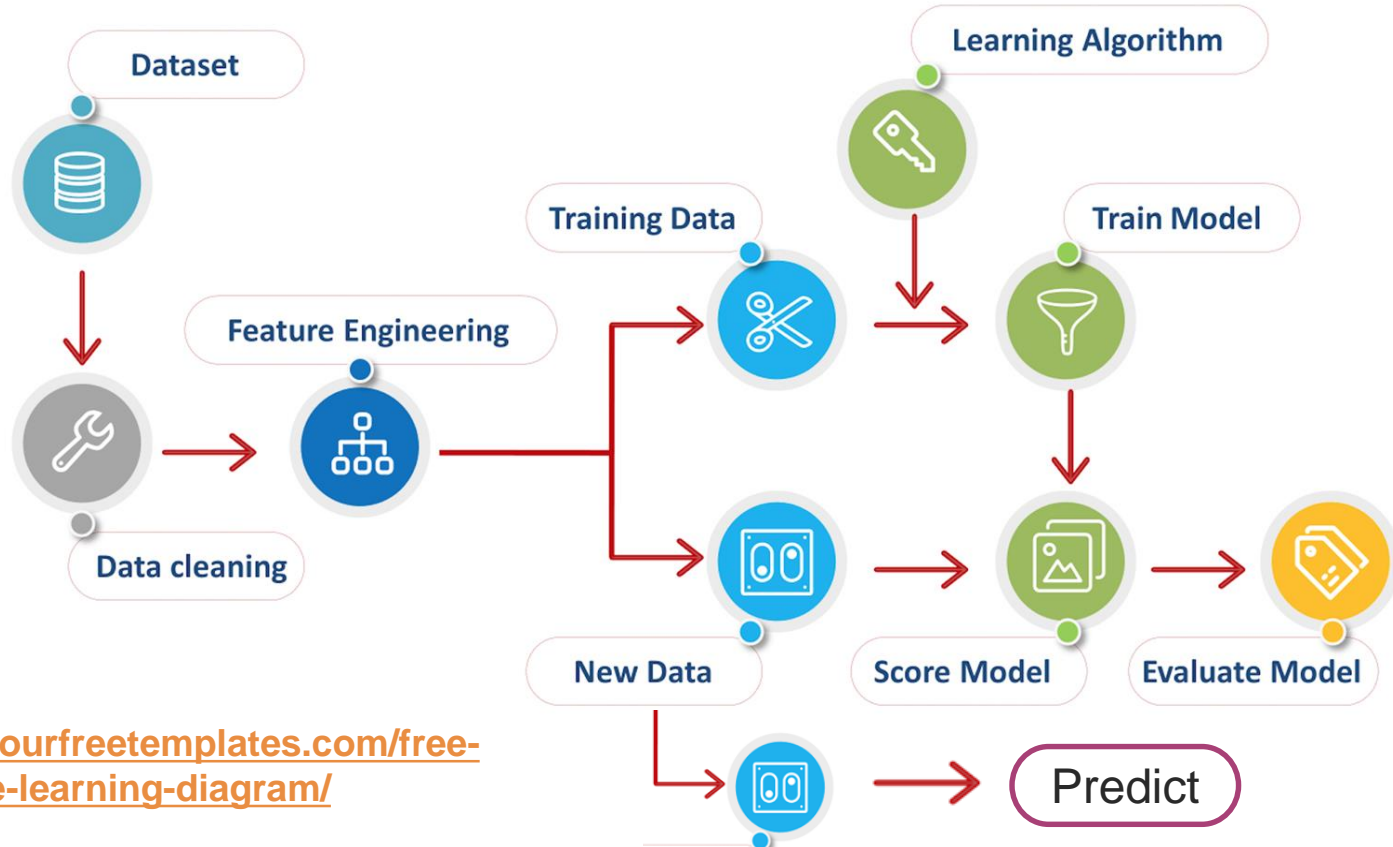
# ZOOM 學員操作說明

The screenshot shows the Zoom interface with several callouts:

- 5 查看選項/共同註記/筆 (連連看)**: Points to the '共同註記' (Co-Annotate) option in the top right menu.
- 2 共享螢幕 (指導演練；點評作品)**: Points to the '共享螢幕' (Share Screen) button in the bottom toolbar. A sub-note says: '老師須先停止共享螢幕才能請學生共享螢幕' (The teacher must first stop sharing the screen before asking the student to share the screen).
- 1 聊天**: Points to the '聊天' (Chat) button in the bottom toolbar.
- 3 與會者/舉手**: Points to the '與會者' (Participants) button in the bottom toolbar.
- 4 解除靜音**: Points to the '解除靜音' (Unmute) button in the bottom toolbar.

Additional interface elements visible include the top bar with 'www.pcschool.com.tw', a toolbar with icons for mouse, text, pen, eraser, format, undo, redo, and delete, and a participants window titled '與會者 (15)' showing a list of users and a '舉手' (Raise Hand) button.

# 機器學習流程



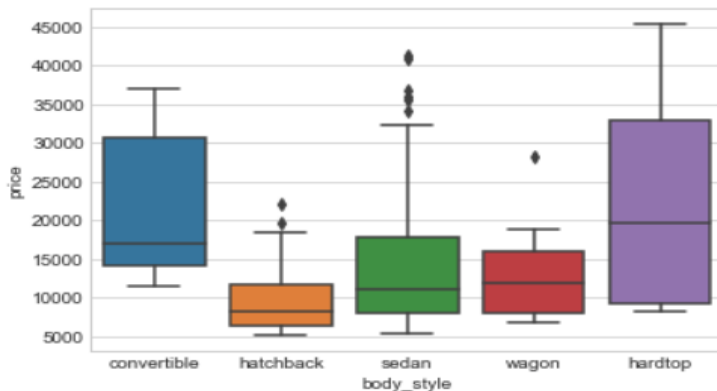
<https://yourfreetemplates.com/free-machine-learning-diagram/>

# 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

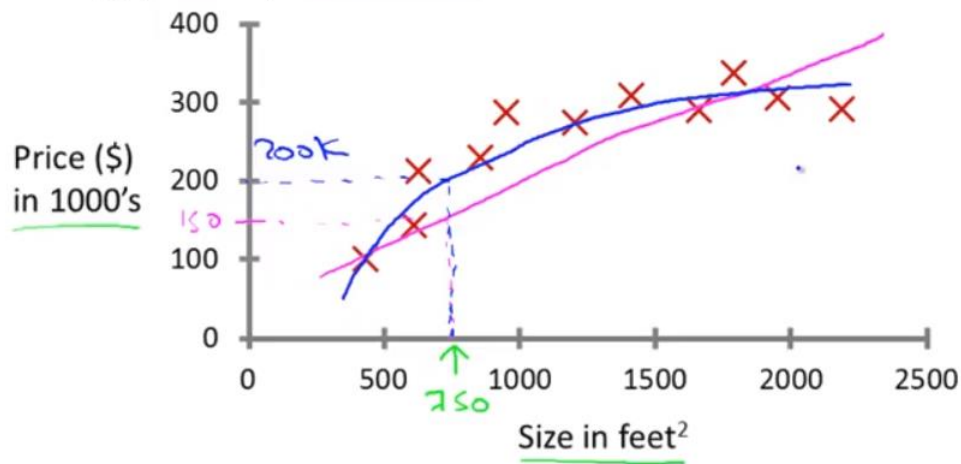
## 分類之資料探索與分析

- 概念介紹
- 實作



# 迴歸 ( Regression )

Housing price prediction.



Supervised Learning

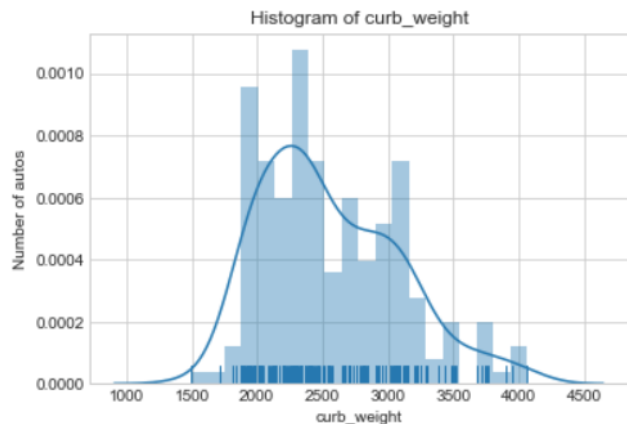
"right answers" given

# 目標

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

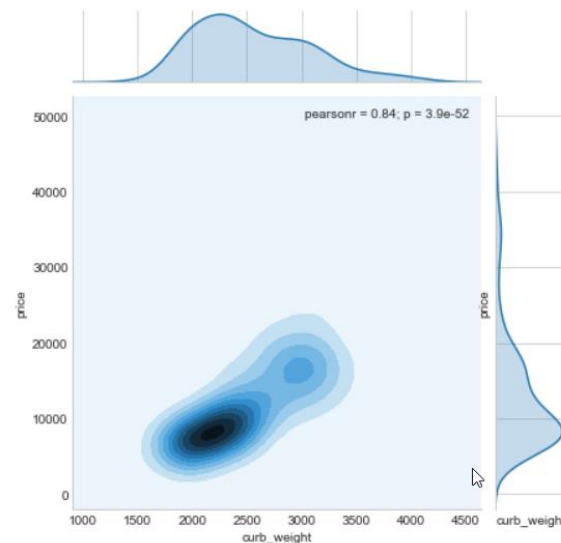
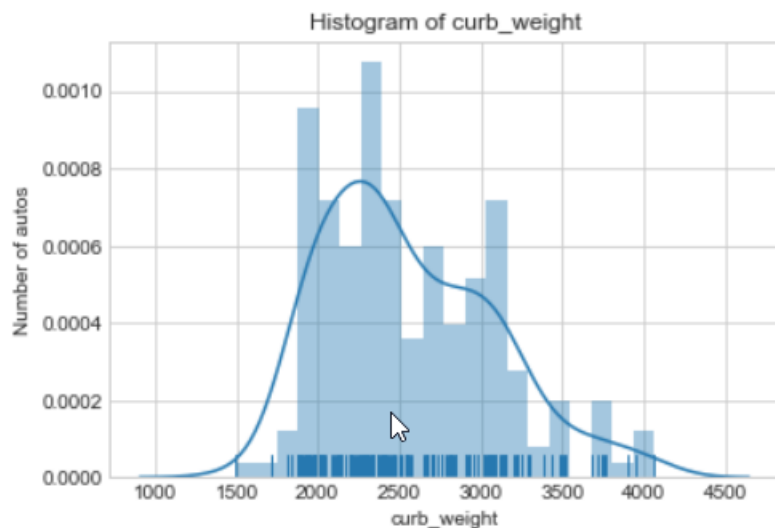
# 處理流程

- ◆ 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



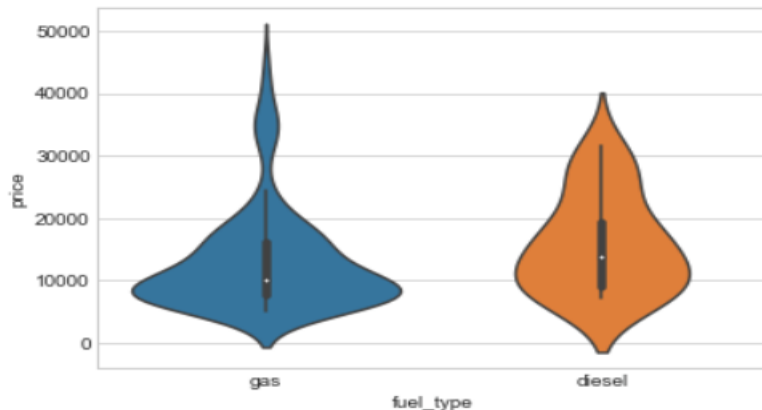
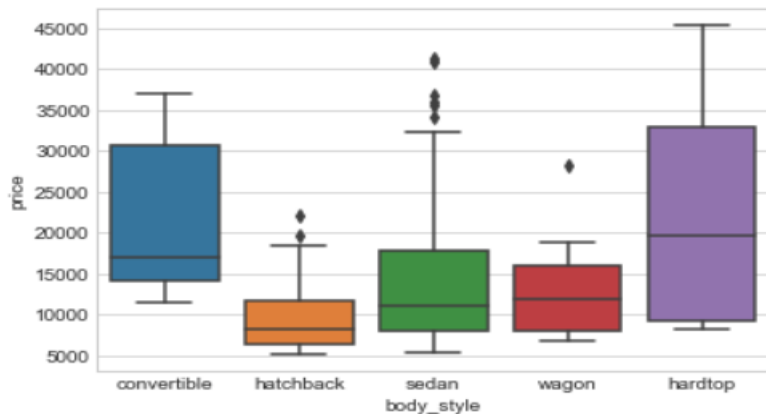
# Seaborn 圖形種類 (1)

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

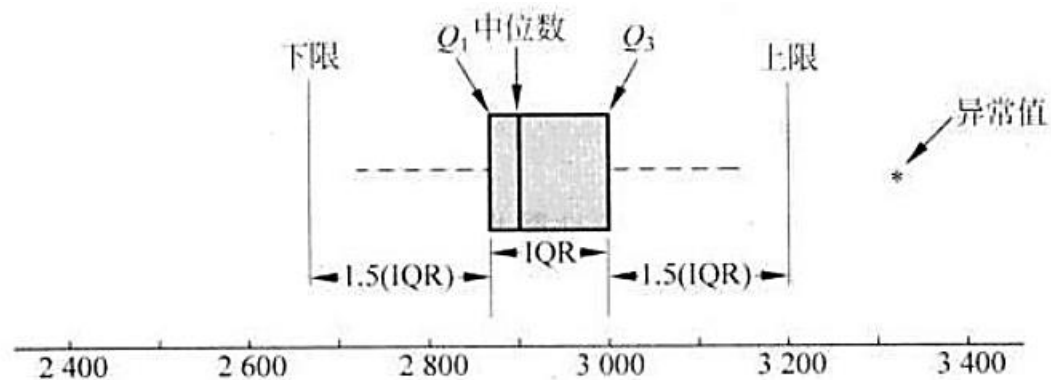


## Seaborn 圖形種類 (2)

- ◆ 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

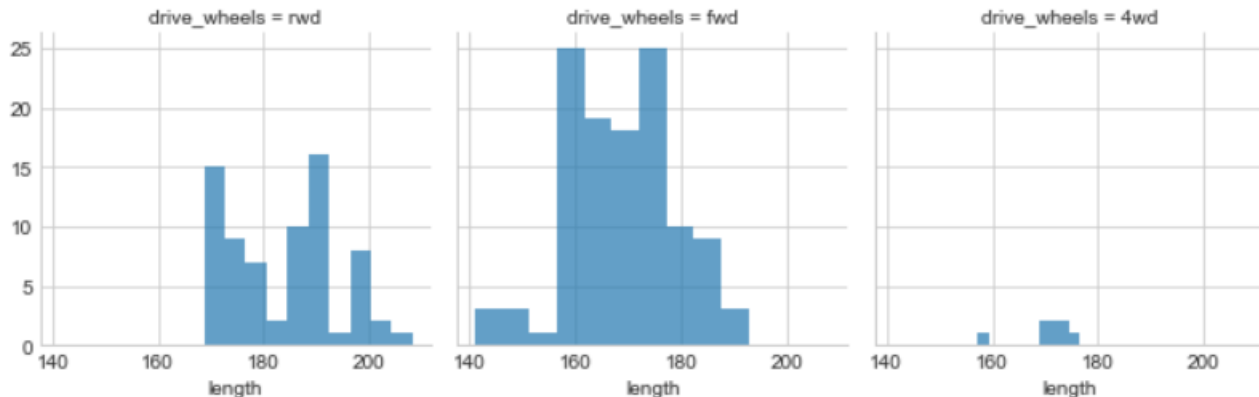


# Seaborn 圖形種類 (3)

◆ 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.

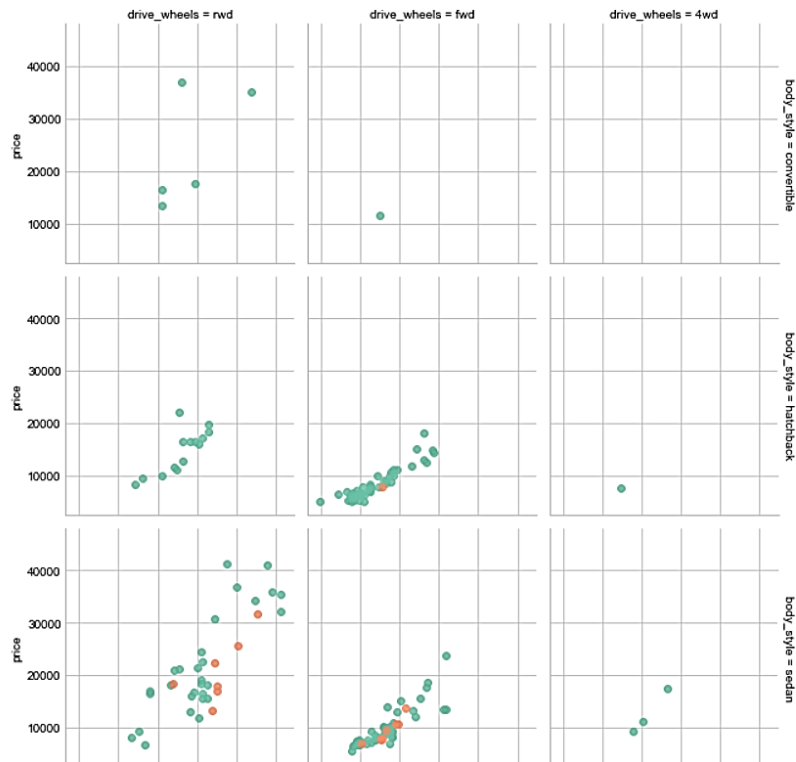
## 問題(4)

- ◆ 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

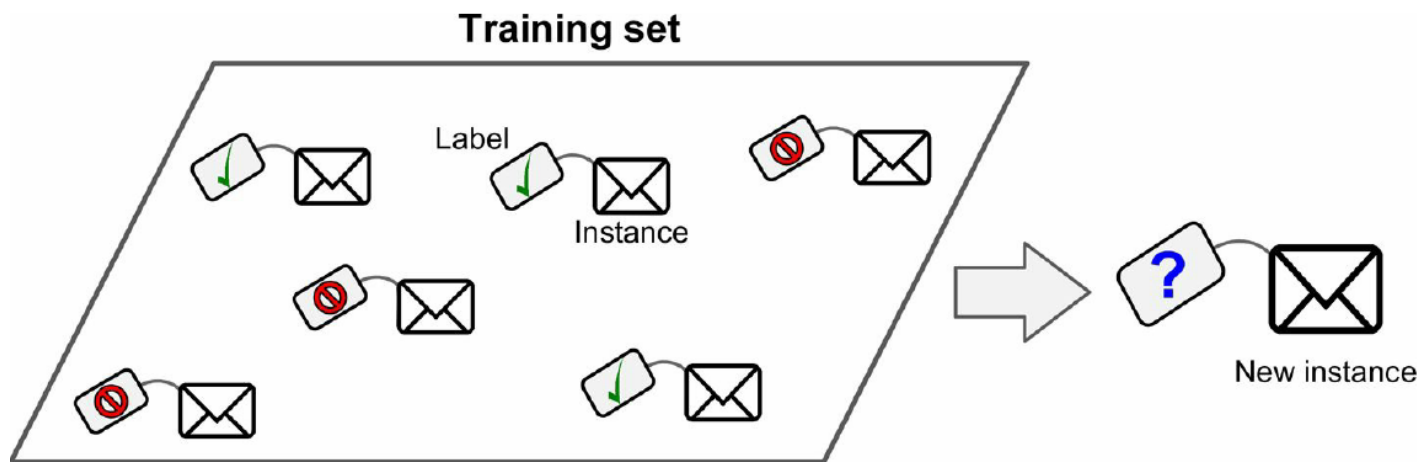
## 迴歸之資料探索與分析

- 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\_pcmt\_income

2) number\_loans

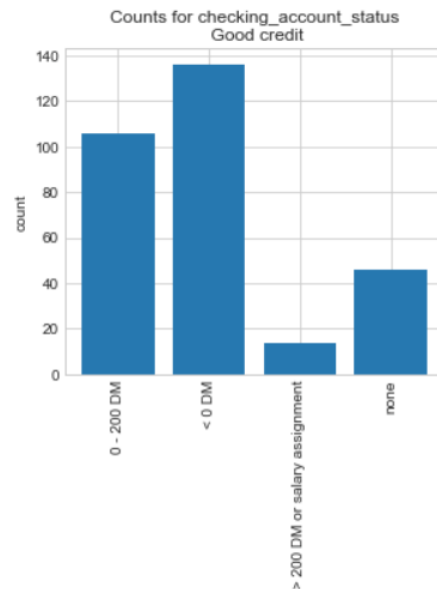
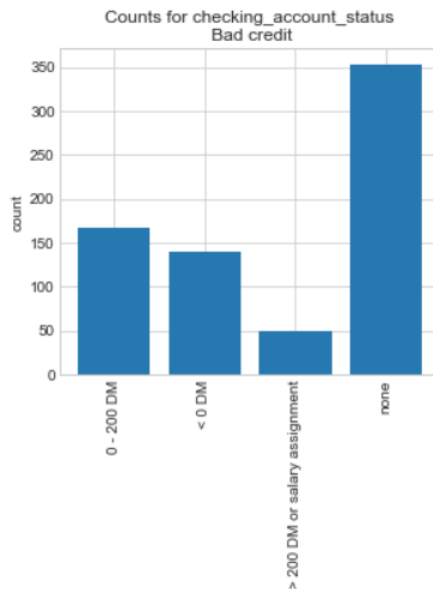
3) age

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

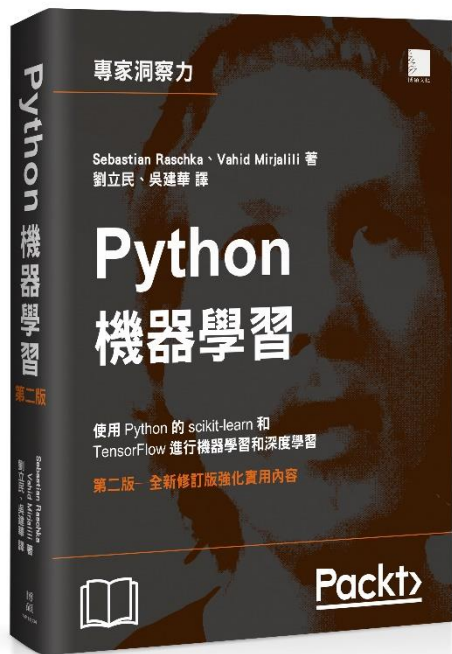
- ◆ 資料集：鐵達尼 ( Titanic )
- ◆ `import seaborn as sns`
- ◆ `titanic = sns.load_dataset("titanic")`
- ◆ 參考

展示



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

# 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>

開課查詢

免費體驗專區

課程總覽

專業師

1

學員專區

講師專區



➤ 課程檔案下載：

學員的「上課教材」，下載檔案為壓縮檔 ([解壓縮操作步驟](#))。  
如無法觀看上課教材，請安裝 [PDF閱讀軟體](#)。

公告專區

我的課表

課程劃位

取消劃位

2

課程檔案下載

自107年1月1日起，課程錄影檔由180天改為365天(含)內無限次觀看 (上課隔日18:00起)。

問  
卷

| 上課日期                   | 課程名稱                 | 課程節次 | 教材下載                 |                     |                      |
|------------------------|----------------------|------|----------------------|---------------------|----------------------|
| 2017/12/27 2000 ~ 2200 | 線上真人-ZBrush 3D動畫造型設計 | 18   | <a href="#">上課教材</a> | <a href="#">錄影檔</a> | <a href="#">課堂問卷</a> |
| 2017/12/20 2000 ~ 2200 | 線上真人-ZBrush 3D動畫造型設計 | 17   | <a href="#">上課教材</a> | <a href="#">錄影檔</a> |                      |
| 2017/12/18 2000 ~ 2200 | 線上真人-ZBrush 3D動畫造型設計 | 16   | <a href="#">上課教材</a> | <a href="#">錄影檔</a> |                      |



巨匠線上真人

[www.pcschoolonline.com.tw](http://www.pcschoolonline.com.tw)