

天氣與人工智慧 機器學習實作

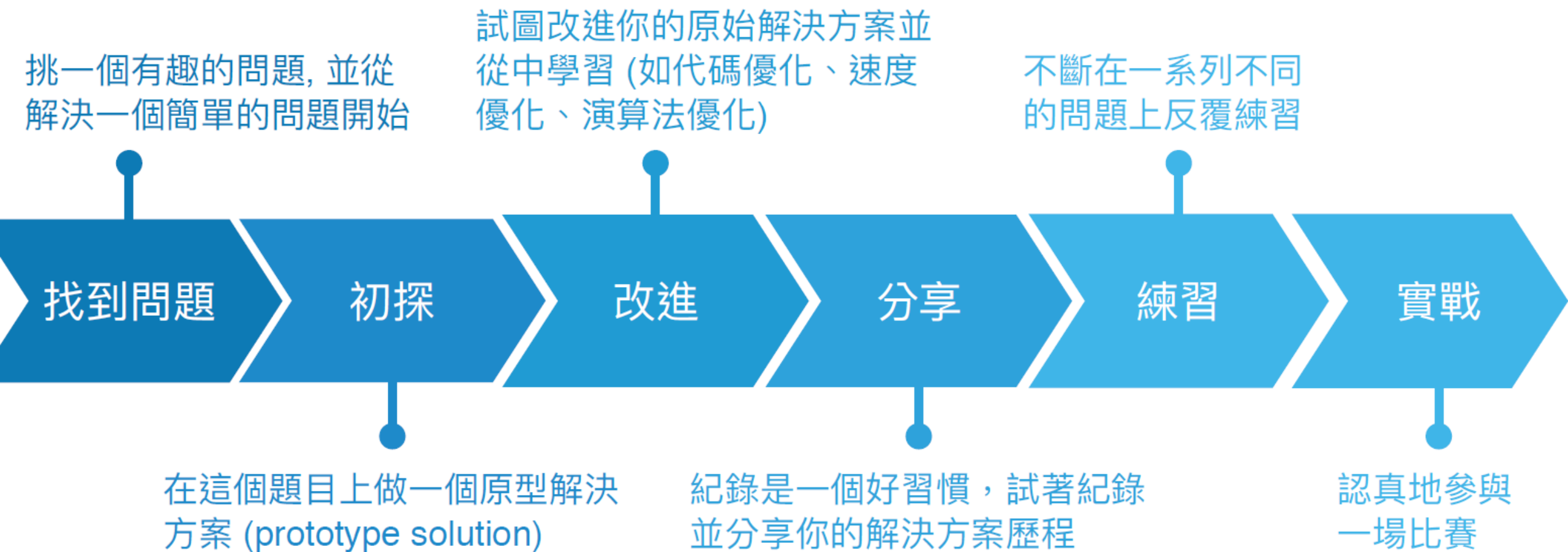
資料清理數據前處理 - 1

2022/11/09

Subject

- 資料介紹與評估指標
- **EDA**/讀取資料
- 如何新建一個 **dataframe** ?

學習路徑



首次面對資料，我們應該思考哪些問題？

Questions	Explanation	Examples
為什麼這個問題重要？ (Why it is important)	A. 好玩 B. 企業的核心問題 C. 公眾利益 / 影響政策方向 D. 對世界很有貢獻	A. 預測生存 (吃雞) 遊戲誰可以活得久 PUBG B. 用戶廣告投放, ADPC C. 停車方針 計程車載客優化 D. 肺炎偵測
資料從何而來？ (Where do data come from)	<ul style="list-style-type: none">來源與品質息息相關根據不同資料源，我們可以合理的推測/懷疑異常資料異常的理由與頻率	資料來源如： 網站流量、購物車紀錄、網路爬蟲、格式化表單、 Crowdsourcing 、紙本轉電子檔
資料的型態是什麼？ (What are they)	A. 結構化資料需要檢視欄位意義以及名稱 B. 非結構化資料需要思考資料轉換與標準化方式	A. 結構化：數值, 表格, ...etc B. 非結構化：圖像、影片、文字、音訊, ... etc
我們可以回答什麼問題？ 問題：指標 (What is our goal)	每個問題都應該要可以被驗證 → 有一個可供衡量的數學評估指標 (Evaluation Metrics)	常見的衡量指標如： 分類問題：正確率, AUC, MAP, ...etc 迴歸問題：MAE, RMSE, ...etc 補充資料： 衡量指標

範例一：我們應該要 / 可以回答什麼問題？

生存 (吃雞) 遊戲

- 玩家排名：平均絕對誤差 (Mean Absolute Error, MAE)
- 怎麼樣的人通常活得久/不久 (如加入遊戲的時間、開始地點、單位時間內取得的資源量, ...) → 玩家在一場遊戲中的存活時間：迴歸 (Mean Squared Error, MSE)



範例二：我們應該要 / 可以回答什麼問題？

廣告投放

- 不同時間點的客群樣貌如何 → 廣告點擊預測 → 預測哪些受眾會點擊或行動：Accuracy / Receiver Operating Curve, ROC
- 哪些素材很好/不好 → 廣告點擊預測 → 預測在版面上的哪個廣告會被點擊：ROC / MAP@N (eg. MAP@5, MAP@12)





解題時間

It's Your Turn

資料簡介：房貸風險預測

資料來源

[描述]

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. **Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.**

資料簡介：房貸風險預測

Questions	Explanation
為什麼這個問題重要？	<p>許多人因為沒有信用歷史，所以沒辦法申請貸款 → 這群人常會轉向風險較高的放款者 → 可能導致這群人的生活狀況更糟 → 如果這群人可以接受正向的幫助，他們將能步入良好正常生活</p> <p>Home Credit 想透過放寬貸款條件，提供給這群人可以有好的借貸經驗 → 但即使放寬貸款條件，公司仍不能接受嚴重呆帳 (未還款) 發生 → 預測還款能力，讓公司可以在放寬貸款條件下，仍不致有貸給無法還債者。</p>
資料從何而來？	信用局 (Credit Bureau) 調閱紀錄、Home Credit 內部紀錄 (如過去借貸狀況、信用卡狀況)
資料的型態是什麼？	<p>[Data]</p> <p>皆為結構化資料：數值、類別資料</p>
我們可以回答什麼問題？ 問題：指標	<p>[Evaluation] 分類問題, 預測各個客戶 ID 是否會還款，以還款機率 (0 ~ 1) 作為最終輸出</p> <p>以 Area Under the ROC curve (ROC) 評估 <small>[註1]</small></p>

註1：在 AUROC, 0.5 代表隨機猜測, 越趨近於 1 則代表模型預測力越好

資料的樣子是什麼？

我們有多少資料

- 多少個資料來源？資料的格式是什麼？資料之間關係是什麼？
- 資料欄位的意義？每一 row 的意義？
- 仔細閱讀 Kaggle 上提供的 [資料說明](#)

你會遇到很多具體的問題

- 怎麼讀資料？在 Python 做資料前處理，我們第一步就是引入常用的套件
 - [pandas](#)：用於讀取以及管理資料
 - [numpy](#)：用於數學函數的運算
- 有多少筆資料？有多少個欄位？
- 有沒有遺失值等等

這些問題的本質其實是在了解資料，我們稱為
「探索式資料分析」(Exploratory Data Analysis)

什麼是EDA？

01

初步透過**視覺化/統計工具**進行分析，達到三個主要目的

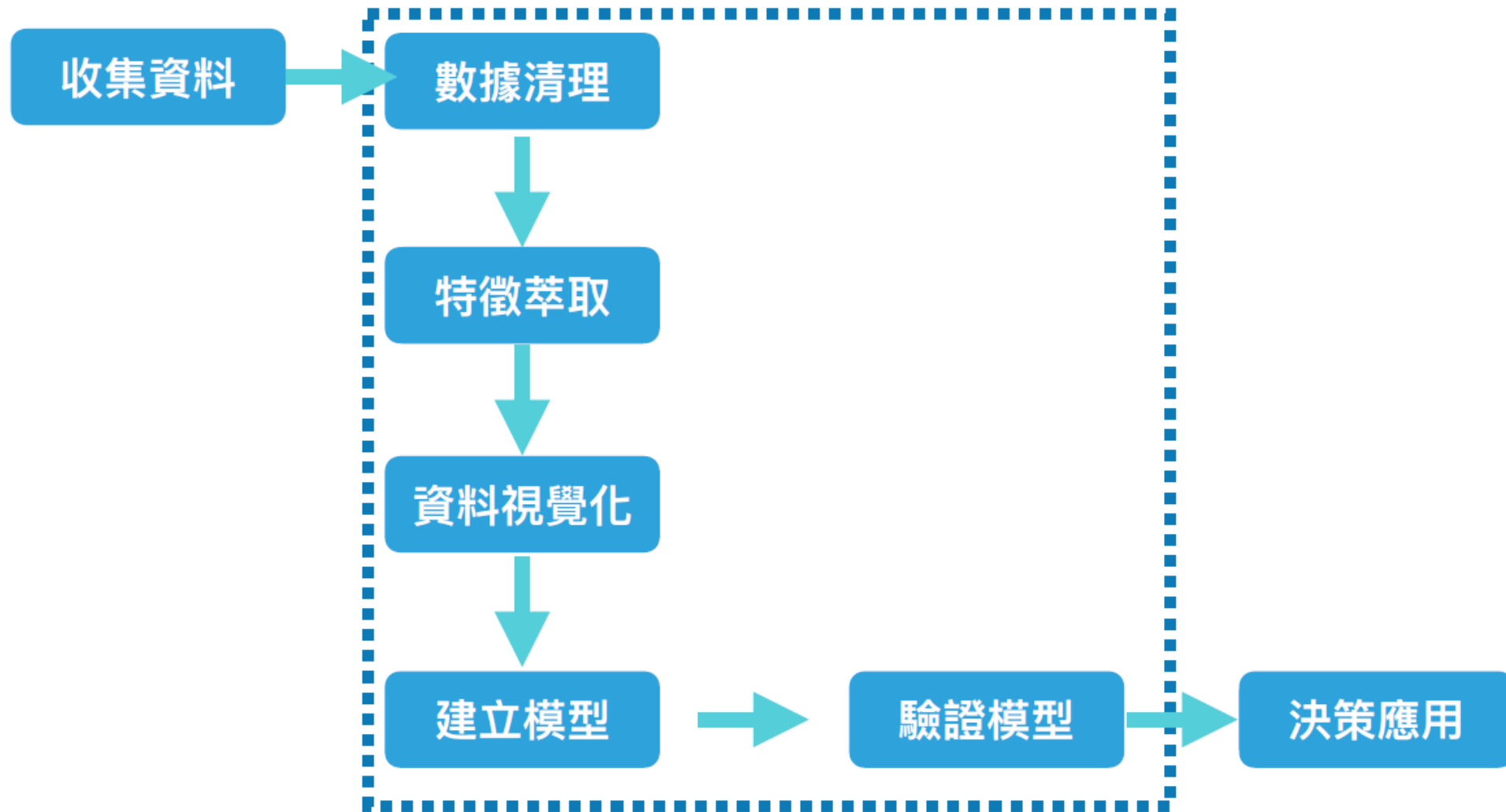
- **了解資料**
 - 獲取資料所包含的資訊、結構和特點
- **發現 outliers 或異常數值**
 - 檢查資料是否有誤
- **分析各變數間的關聯性**
 - 找出重要的變數

從 EDA 的過程中觀察現象，檢查資料是否符合分析前的假設

- 可以在模型建立之前，先發現潛在的錯誤
- 也可以根據 EDA 的結果來調整分析的方向

02

數據分析的流程





解題時間

It's Your Turn

為什麼新建一個 dataframe 重要？



需要把分析過程中所產生的數據或者結果儲存為結構化的資料

- Ex 1: 將每筆交易資料匯總計算平均值、標準差等統計數值
- Ex 2: Kaggle 比賽要上傳的結果



測試程式碼

- 有時候原始資料太大了，有些資料的操作很費時，先在具有同樣結構的資料上測試程式碼是否能夠得到理想中的結果。
- 不確定視覺化程式碼中所需要的資料結構，用新建立的 dataframe 結構來去了解，而不是急著在原始資料上操作。

讀取其他非csv資料格式？

檔案格式

讀取範例

文本 (txt)

```
with open('example.txt', 'r') as f:  
    data = f.readlines()  
print(data)
```

Json

```
import json  
with open('example.json', 'r') as f:  
    data = json.load(f)  
print(data)
```

矩陣檔 (mat)

```
import scipy.io as sio  
data = sio.loadmat('example.mat')
```

讀取其他非csv資料格式？

檔案格式

讀取範例

圖像檔 (PNG / JPG ...)

```
image = cv2.imread(...) # 注意 cv2 會以 BGR 讀入  
image = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
```

```
from PIL import Image  
image = Image.read(...)  
import skimage.io as skio  
image = skio.imread(...)
```

Python npy

```
import numpy as np  
arr = np.load(example.npy)
```

Pickle (pkl)

```
import pickle  
with open('example.pkl', 'rb') as f:  
    arr = pickle.load(f)
```



解題時間

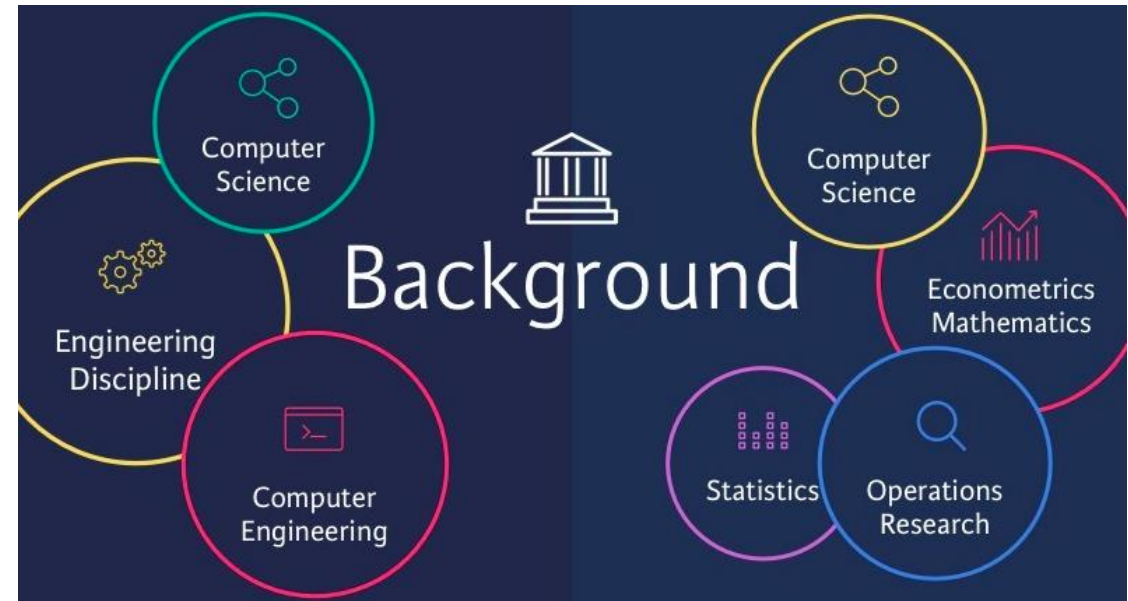
It's Your Turn

參考資料

Data Scientist、Data Analyst、Data Engineer 的區別是什麼？

[原始連結\(英文\)](#) [後續討論\(簡中\)](#)

想必大家在之前多少聽過這些名詞，也帶有不少疑惑，就讓我們看看在業內的專家們怎麼說吧。簡單來說：資料科學家 (Data Scientist) 需要擅長的是數字的敏感度與資料分析工具，訓練偏重統計，也就是本課程想要帶給各位同學的內容。而資料工程師 (Data Engineer) 需要對計算機本身較為熟悉，訓練偏重資料工程，往往需要透過實務的親身經歷來成長，這部分比較難以線上課程的方式提供。



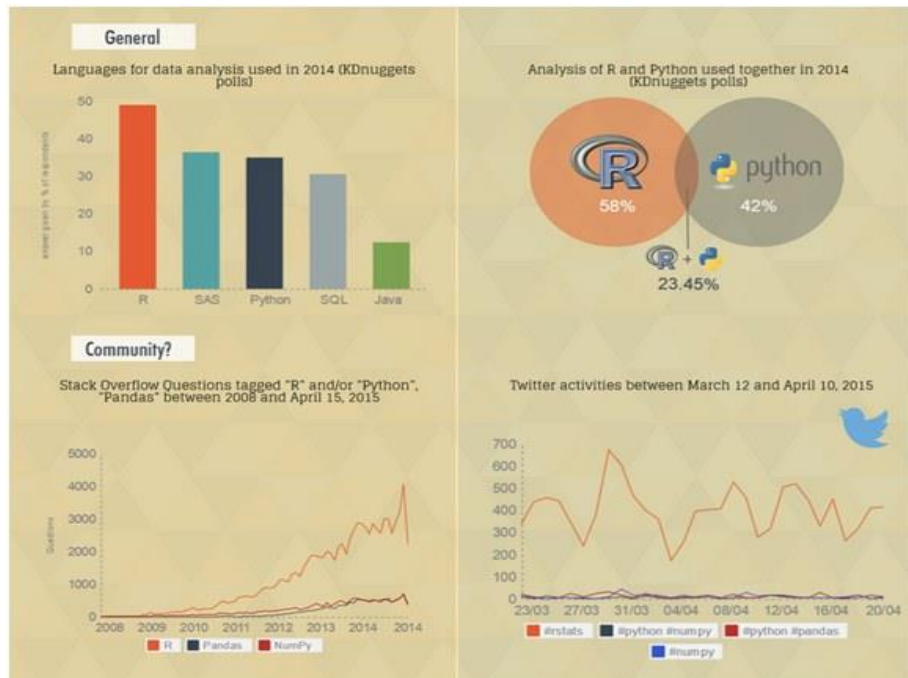
R or Python for Data Science?

[網頁連結](#)

"學 Python 還是 R 語言好?" 想必這個經典問題,也曾是不少同學的煩惱吧?

這個網站的回答雖然也很經典,但是製表的日期已經是 2014 年了,以老師現在 (2019年) 的觀察來說, R 語言雖然在機器學習上比 Python 略為好用,可是在深度學習上, Python 可以說壓著 R 語言打呢,所以還是建議同學先學 Python 比較穩當。

此外, R 語言的另一個好處,是由大量碩博士生貢獻的套件,這個學界霸主的地位已經逐步被 PyTorch 所取代,而業界因為生態系完整的關係,還是以 TensorFlow / Keras 為主,後兩者都是在 Python 上的套件,所以怎麼看,先學 Python 還是比較不虧的。



[其他參考連結]

[Why Data Scientist Must Focus on Developing Product Sense](#)

資料科學家需要目標的領域知識

[Think twice before getting into data science](#) (原文：Why so many data scientist leaving their jobs)

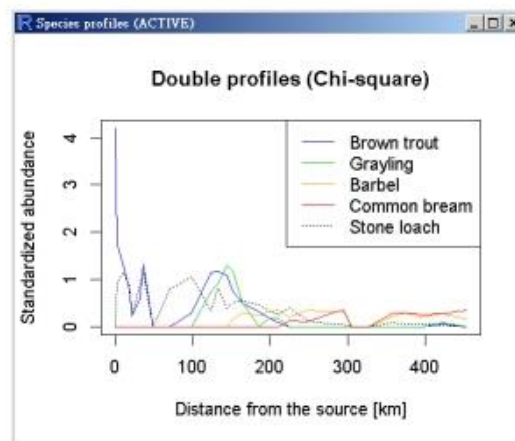
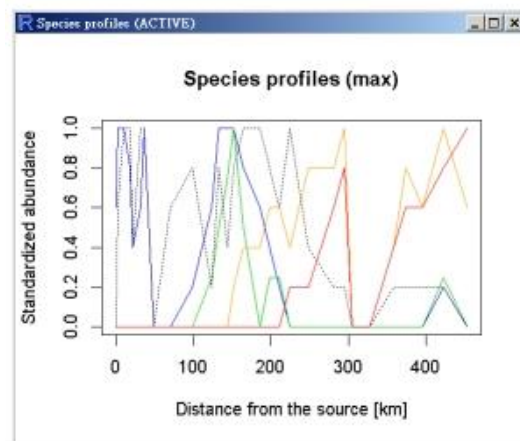
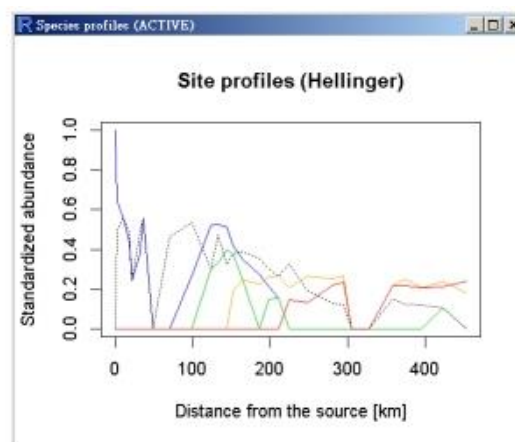
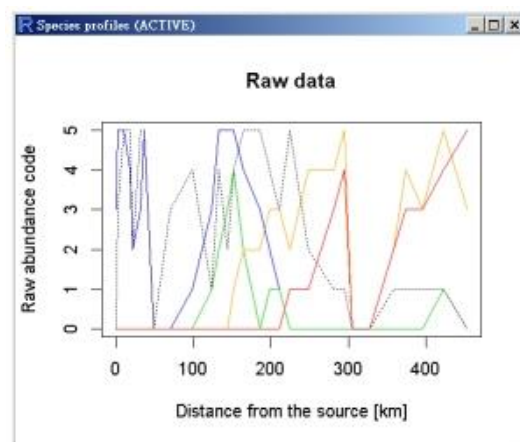
想當資料科學家：三思而後行

探索式資料分析簡介

吳漢銘老師 [網頁連結](#)

這是吳老師講解探索式資料分析 (EDA) 的內容，比較側重於理論部分，同學可以在這裡看到許多豐富的資料，對照後續的課程內容，可以讓您更了解EDA的全貌。

建議同學可以閱讀自己有興趣的部分即可，有必要了解的細節，會在後續課程中提到。

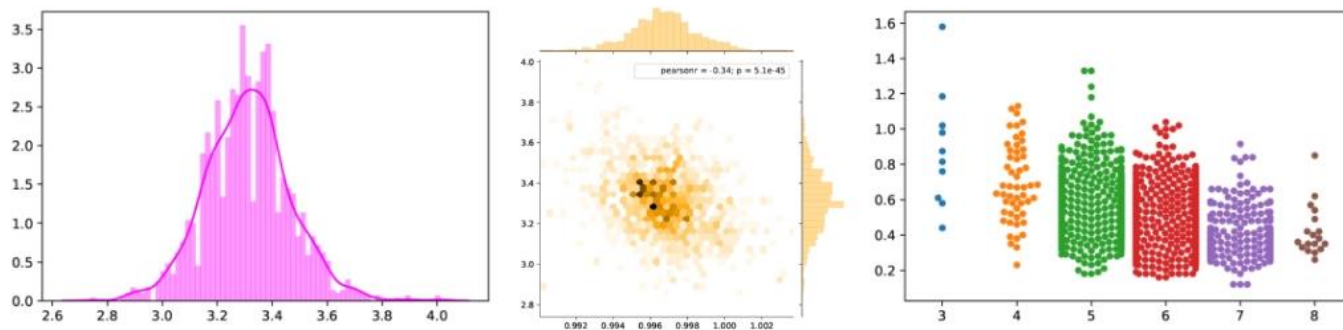


What is Exploratory Data Analysis?

towards data science Prasad Patil

[網頁連結](#)

這是另一份 EDA 的教學，比較側重於圖形與對應的程式碼，比較接近我們後續的課程形式。從這些範例中，我們可以看到有分布圖 (左)，蜂窩聯合圖(中)，分類散佈圖(右)，這些圖形都可以輕鬆藉由 seaborn 套件繪製，所以只要我們在後續課程中學會這些，就可以輕鬆完成資料視覺化。



What is Exploratory Data Analysis?




Prasad Patil [Follow](#)
Mar 24, 2018 · 6 min read


Pandas Foundations : Data ingestion & inspection

Pandas Foundations

[網頁連結](#)

第一個 chapter 是免費的，建議可用來練習 pandas，如果覺得英文聽不懂也沒關係，可以按部就班跟著後面的課程，也可以學到相關的內容。












pandas Foundations 

pandas DataFrames

- Example: DataFrame of Apple Stock data

Date	Open	High	Low	Close	Volume	Adj Close
2014-09-16	99.80	101.26	98.89	100.86	66818200	100.86
2014-09-15	102.81	103.05	101.44	101.63	61216500	101.63
2014-09-12	101.21	102.19	101.08	101.66	62626100	101.66
...

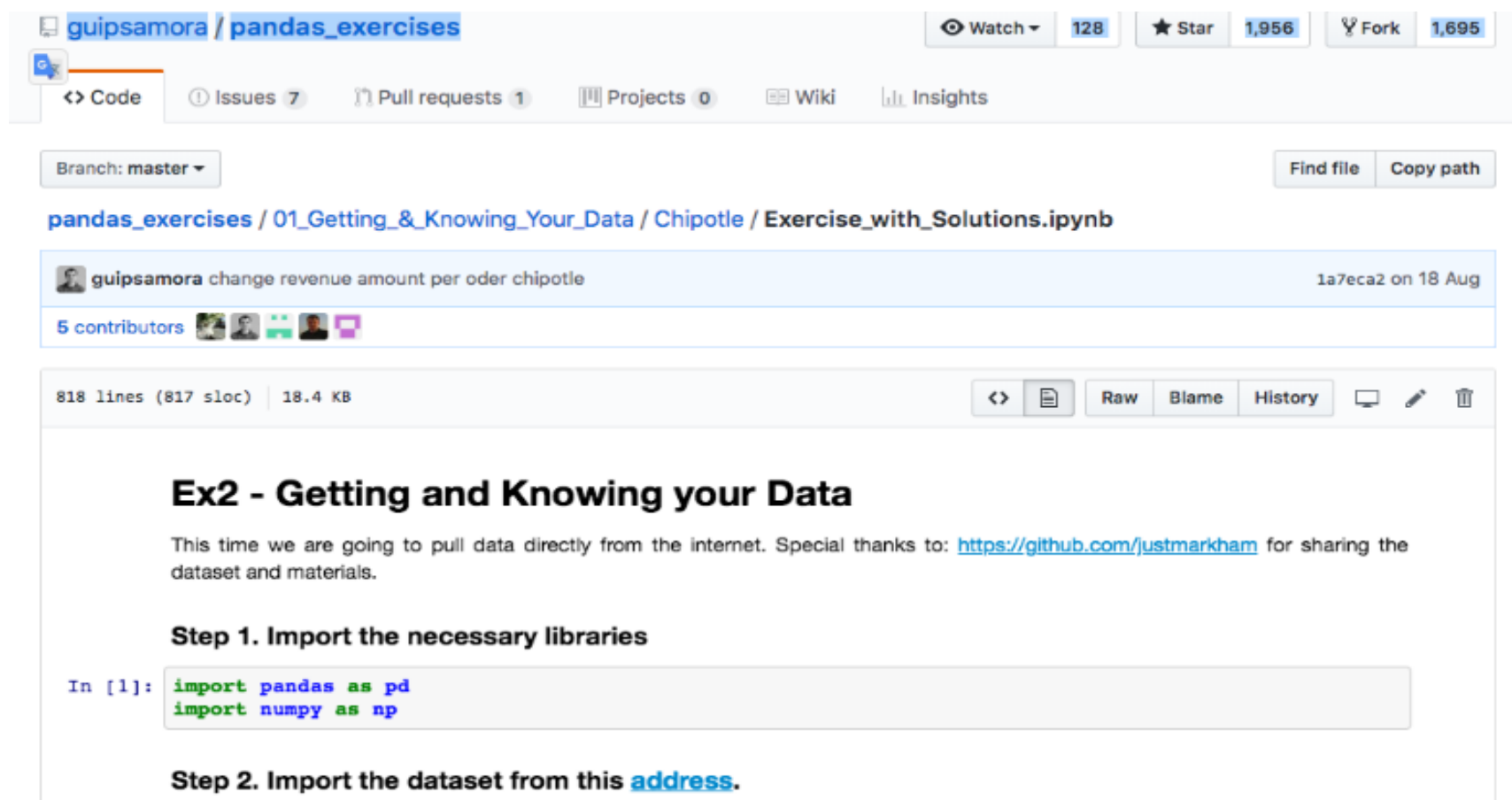




-3:47 1x auto

推薦 github repo

之後實作課程也會有 pandas 操作相關的練習，但若你迫不及待想要更精進自己 pandas 技能，可以到這個 [github repo](#) 挑戰！



The screenshot shows the GitHub interface for the repository 'guipsamora/pandas_exercises'. At the top, it displays the repository name, a 'Watch' button with 128 subscribers, a 'Star' button with 1,956 stars, and a 'Fork' button with 1,695 forks. Below this, there are tabs for 'Code', 'Issues' (7), 'Pull requests' (1), 'Projects' (0), 'Wiki', and 'Insights'. The 'Code' tab is selected, showing the file path 'pandas_exercises / 01_Getting_&_Knowing_Your_Data / Chipotle / Exercise_with_Solutions.ipynb'. A commit by 'guipsamora' is shown with the message 'change revenue amount per oder chipotle' and a commit hash '1a7eca2' from '18 Aug'. Below the commit, it says '5 contributors'. The file details show '818 lines (817 sloc)' and '18.4 KB'. The file content is displayed in a code editor with tabs for '<>', 'Raw', 'Blame', and 'History'. The content starts with the title 'Ex2 - Getting and Knowing your Data', followed by an introduction: 'This time we are going to pull data directly from the internet. Special thanks to: <https://github.com/justmarkham> for sharing the dataset and materials.' Then, 'Step 1. Import the necessary libraries' is shown with a code block:

```
In [1]: import pandas as pd
import numpy as np
```

 Finally, 'Step 2. Import the dataset from this [address](#).'