Day 13 資料清理數據前處理

常用的 DataFrame 操作

出題教練 游為翔 / 杜靖愷

知識地圖 機器學習前處理 常用的 DataFrame 操作

CUPOY

機器學習前處理

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

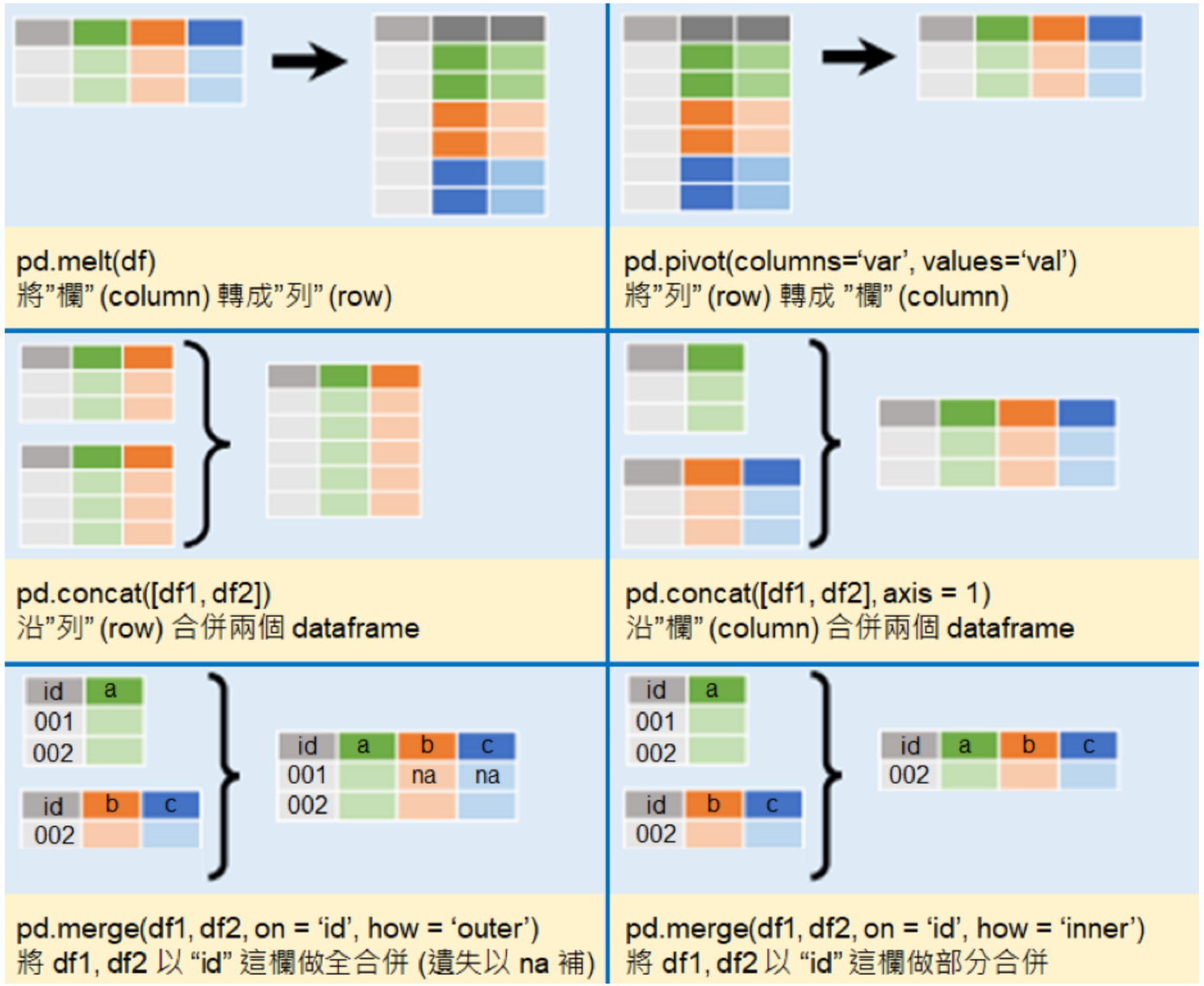
前處理 Processing

資料讀取 → 格式調整 → 填補缺值 → 去離群值 → 特徵縮放

本日知識點目標

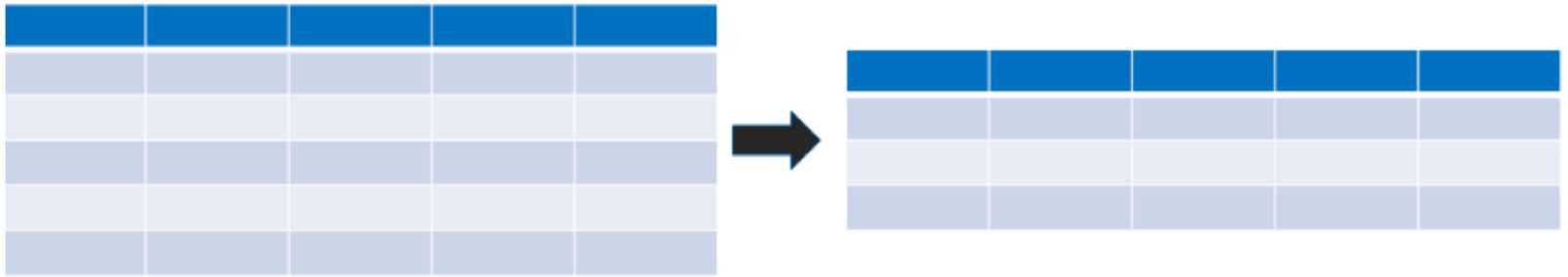
熟悉 python 常用套件 pandas 的操作方式，如排序、合併、分組操作、Indexing 等

轉換與合併 dataframe



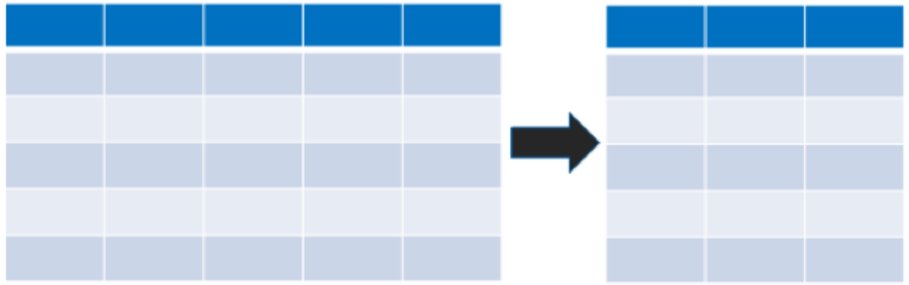
圖片來源: [Pandas Cheat Sheet]

Subset



列篩選 / 縮減	
邏輯操作	sub_df = df[df.age > 20]
移除重複	df = df.drop_duplicates()
前 n 筆	sub_df = df.head(n = 10)
後 n 筆	sub_df = df.tail(n = 10)
隨機抽樣	sub_df = df.sample(frac = 0.5) # 抽 50 %
	sub_df = df.sample(n = 10) # 抽 10 筆
第 n 到 m 筆	sub_df = df.iloc[n : m]

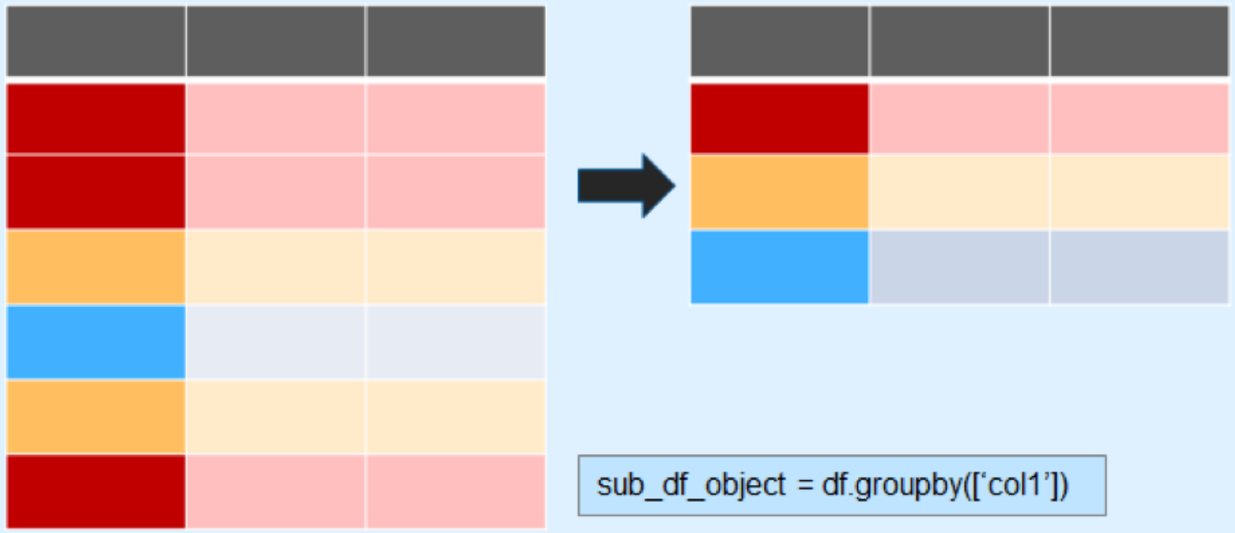
邏輯操作	
大於 / 小於 / 等於	>, <, ==
大於等於 / 小於等於	>=, <=
不等於	!=
&, |, ~, ^	邏輯的 and, or, not, xor
欄位中包含 value	df.column.isin(value)
為 Nan	pd.isnull(obj)
非 Nan	pd.notnull(obj)



欄篩選 / 縮減	
單一欄位	new_df = df['col1'] 或 df.col1
複數欄位	new_df = df[['col1', 'col2', 'col3']]
Regex 篩選	new_df = df.filter(regex = ...)

圖片來源: Pandas Cheat Sheet

Group operations

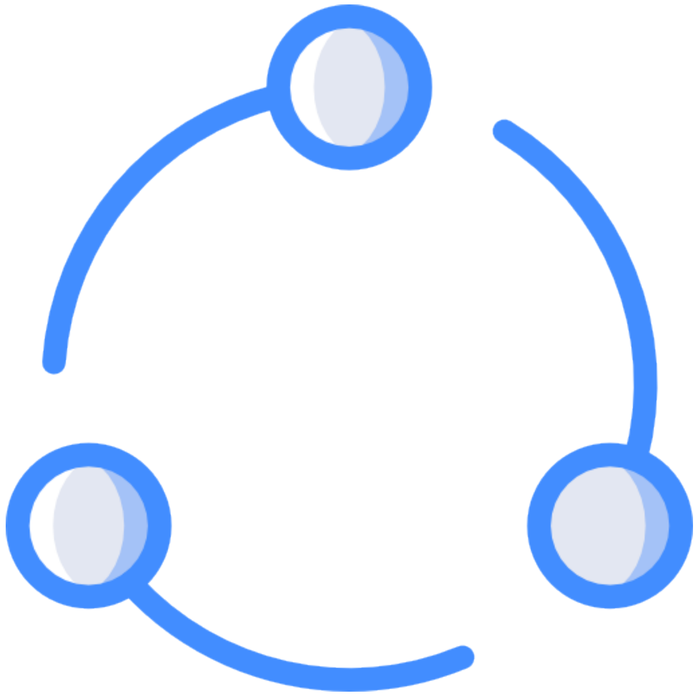


sub_df_object = df.groupby(['col1'])

sub_df_object.size()	計算各組的數量
sub_df_object.describe()	得到各組的基本統計值
sub_df_object['col2'].mean()	根據 col1 分組後，計算 col2 統計值 (以平均值為例, 可用操作如 .max(), .min() …etc)
sub_df_object['col2'].apply()	對依 col1 分組後的 col2 引用操作
sub_df_object['col2'].hist()	對依 col1 分組後的 col2 繪圖 (hist 為例)

圖片來源: Pandas Cheat Sheet

重要知識點複習



- 合併 (concat) 常用於將多個表依照某欄 (key) 結合使用
- 分組 (groupby) 是常用在計算"組"統計值時會用到的功能
- 許多基本操作 (如 >, ==, <, ~) 都是可以在 pandas 作為篩選條件使用

延伸閱讀

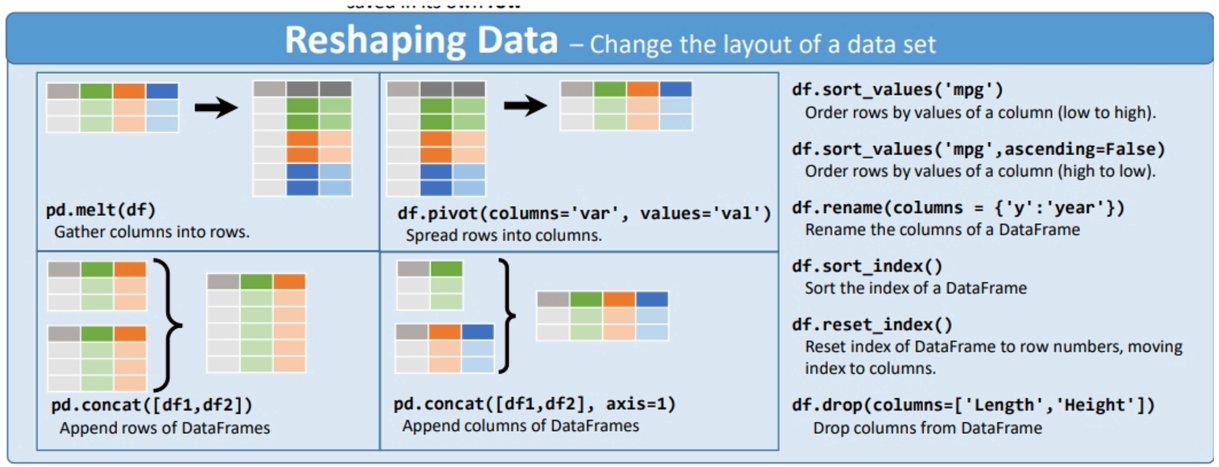
除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度，建議您解完每日題目後，若有多餘時間，可再補充延伸閱讀文章內容。

推薦延伸閱讀

Pandas 官方 Cheat Sheet

Pandas 官方 網頁連結

- 所謂的 Cheat Sheet，中文理解為「懶人包」，各家語言與套件常常有懶人包 (雲端 AI , R 語言....)，目的不外乎是提供使用者便利，以便推廣。不過 Pandas 的懶人包初學時似乎會被滿滿文字困惑住，但學通後回頭看這張，是相當受用的。因此與其他參考資料建議相反：如果未來要將機器學習當作吃飯的工具，強烈建議練習這些範例。

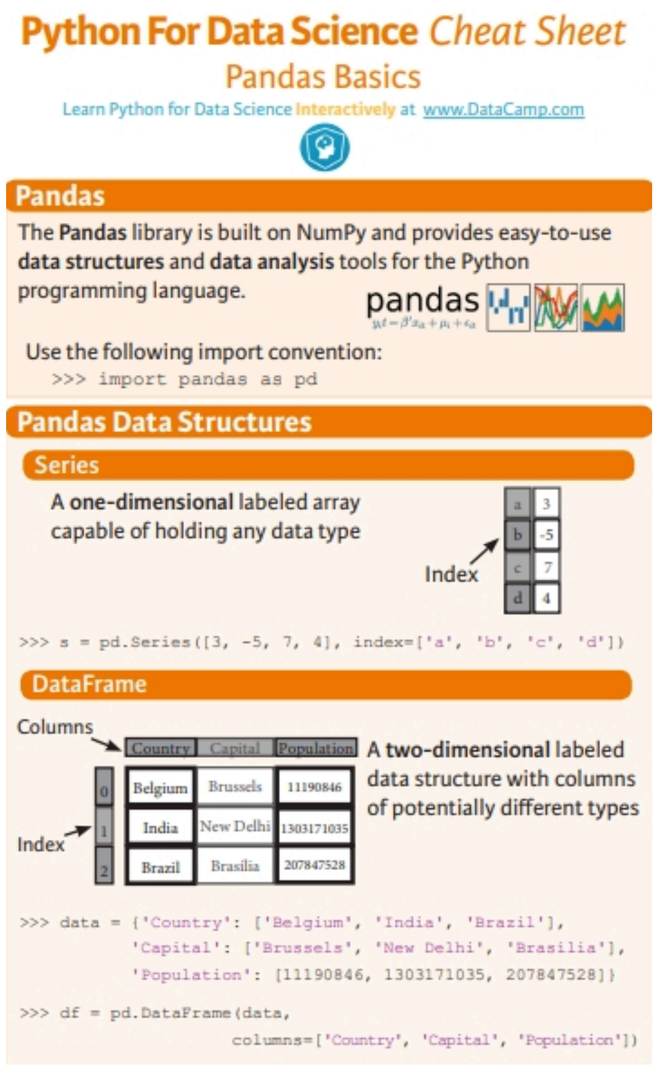


推薦延伸閱讀

Pandas Cheat Sheet

datacamp.com 網頁連結

- 內容上似乎也蠻豐富的，喜歡這種風格的可以嘗試看看，在與官方懶人包之間可以做個選擇，雖然以部分使用者的觀點來說，官方的似乎比較易學易懂一點。



解題時間

It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題