

Day 8

資料清理數據前處理

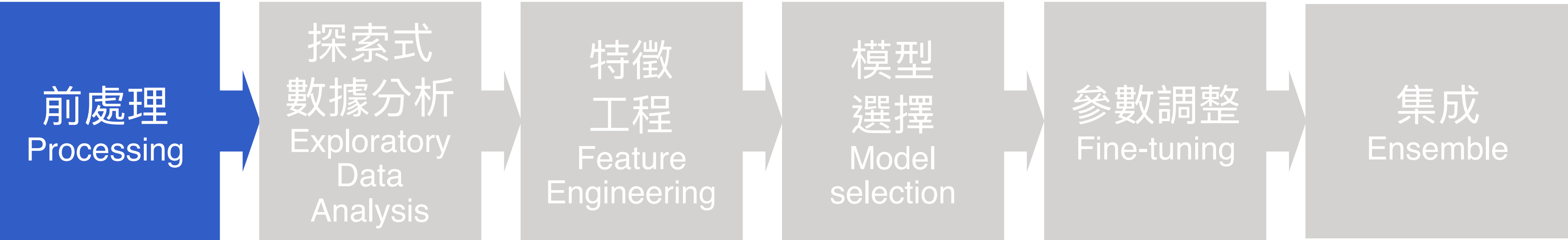
常用的 DataFrame 操作



知識地圖 機器學習前處理 常用的 DataFrame 操作

機器學習概論 Introduction of Machine Learning

監督式學習 Supervised Learning



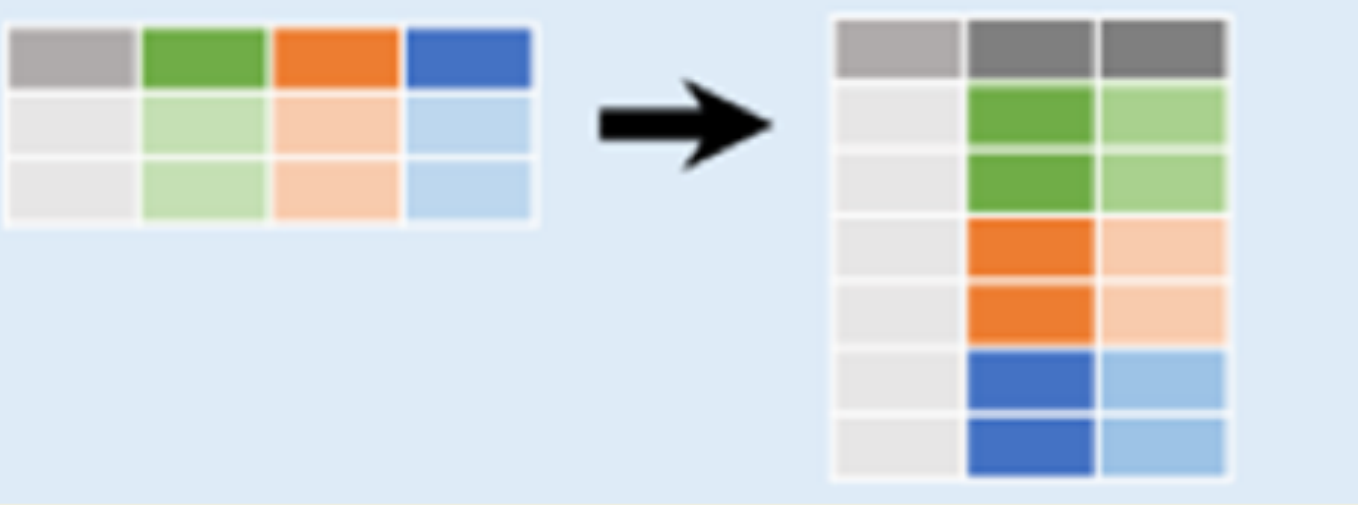


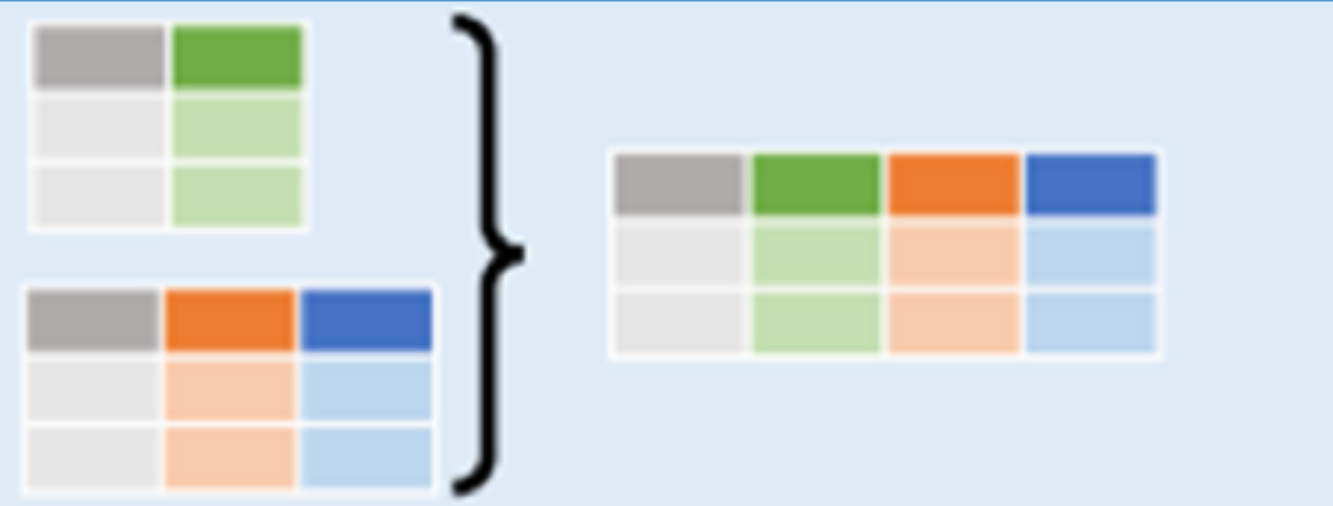
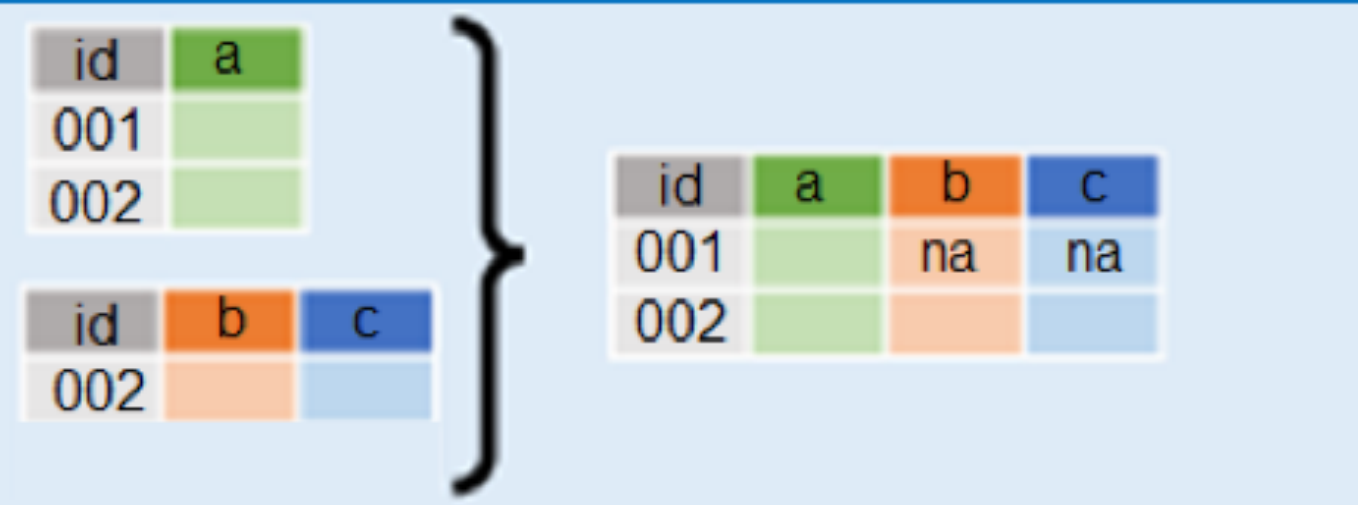
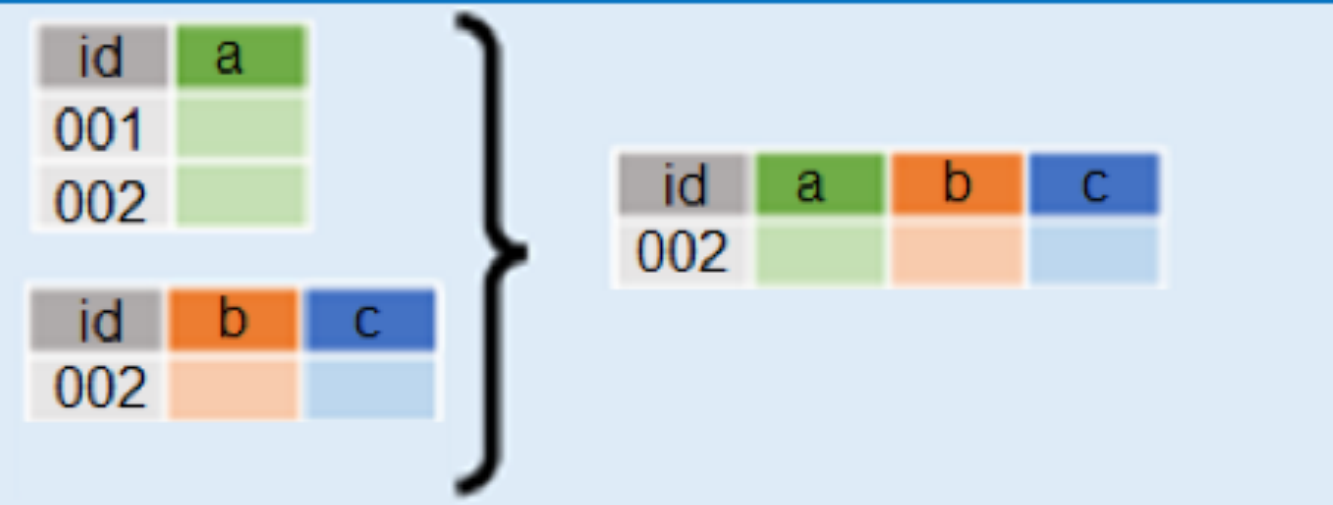
非監督式學習 Unsupervised Learning



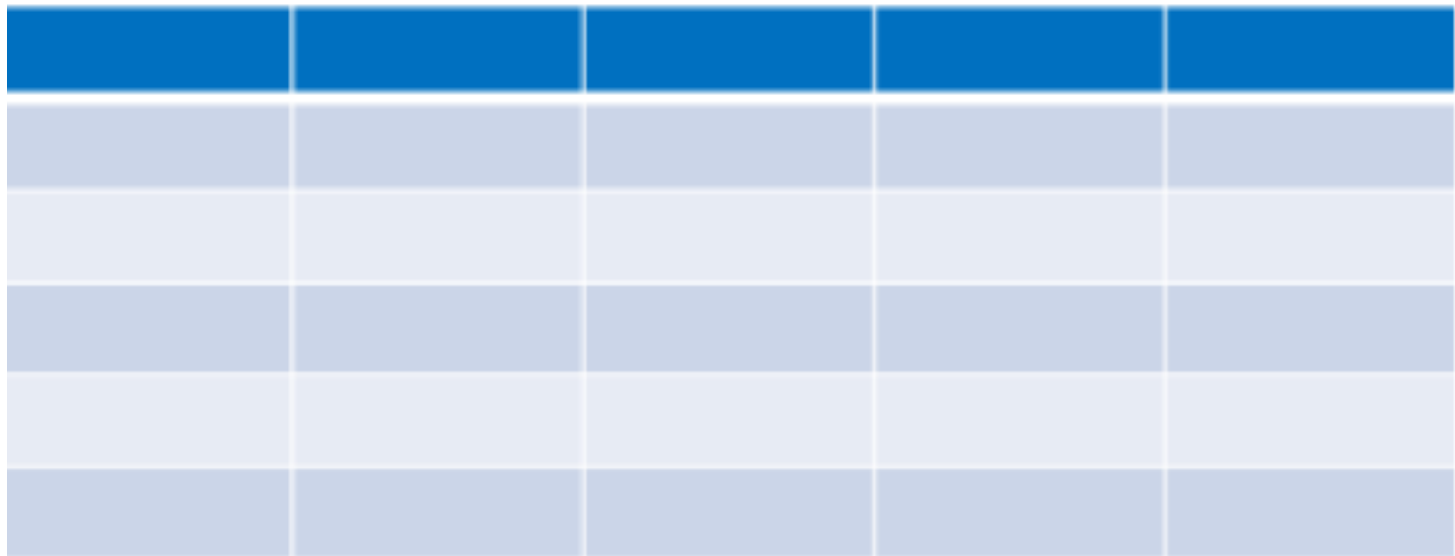
前處理 Processing

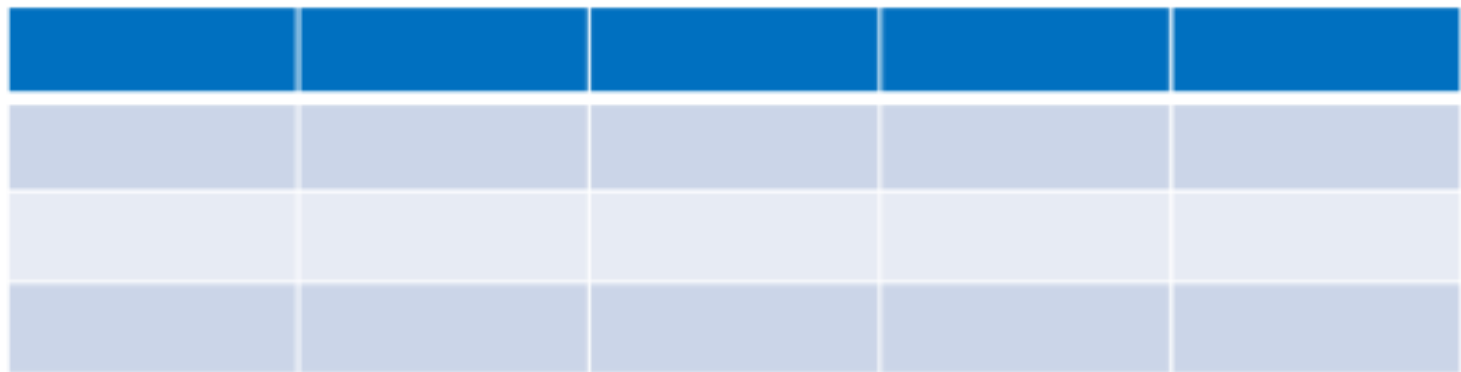


轉換與合併 dataframe

	
<code>pd.melt(df)</code> 將"欄" (column) 轉成"列" (row)	<code>pd.pivot(columns='var', values='val')</code> 將"列" (row) 轉成 "欄" (column)
	
<code>pd.concat([df1, df2])</code> 沿"列" (row) 合併兩個 dataframe	<code>pd.concat([df1, df2], axis = 1)</code> 沿"欄" (column) 合併兩個 dataframe
	
<code>pd.merge(df1, df2, on = 'id', how = 'outer')</code> 將 df1, df2 以 "id" 這欄做全合併 (遺失以 na 補)	<code>pd.merge(df1, df2, on = 'id', how = 'inner')</code> 將 df1, df2 以 "id" 這欄做部分合併

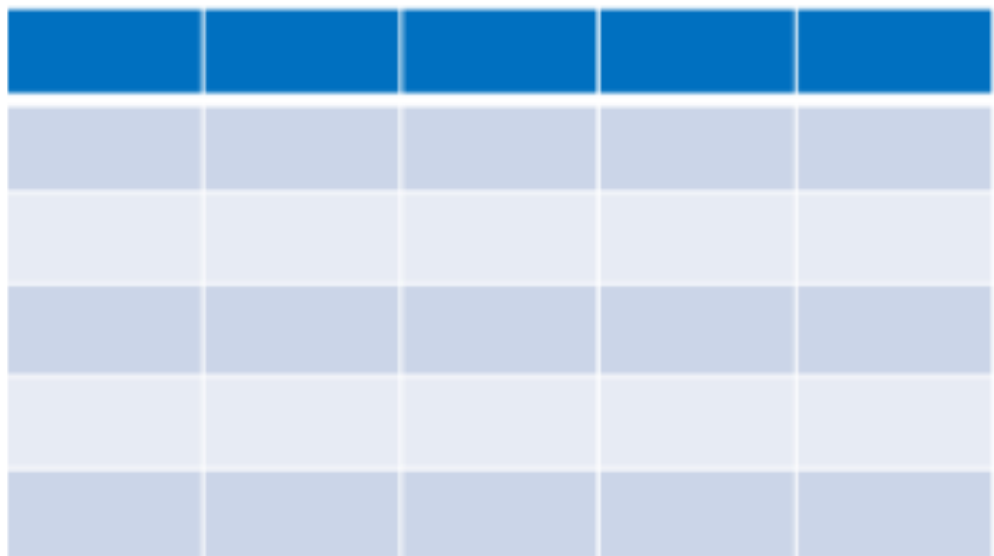
Subset

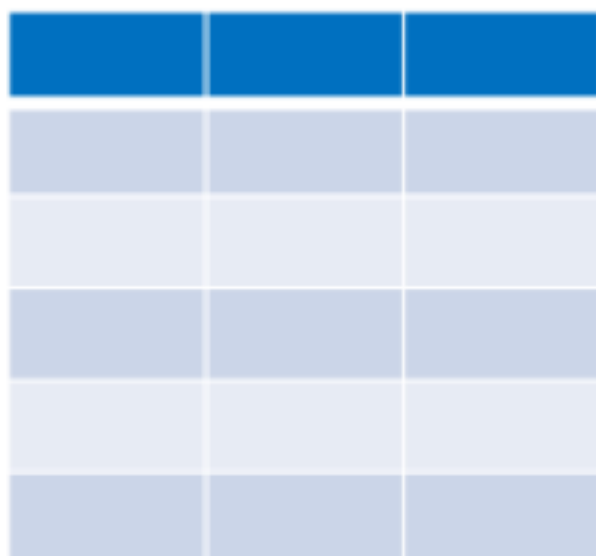




列篩選 / 縮減	
邏輯操作	sub_df = df[df.age > 20]
移除重複	df = df.drop_duplicates()
前 n 筆	sub_df = df.head(n = 10)
後 n 筆	sub_df = df.tail(n = 10)
隨機抽樣	sub_df = df.sample(frac = 0.5) # 抽 50 %
	sub_df = df.sample(n = 10) # 抽 10 筆
第 n 到 m 筆	sub_df = df.iloc[n : m]

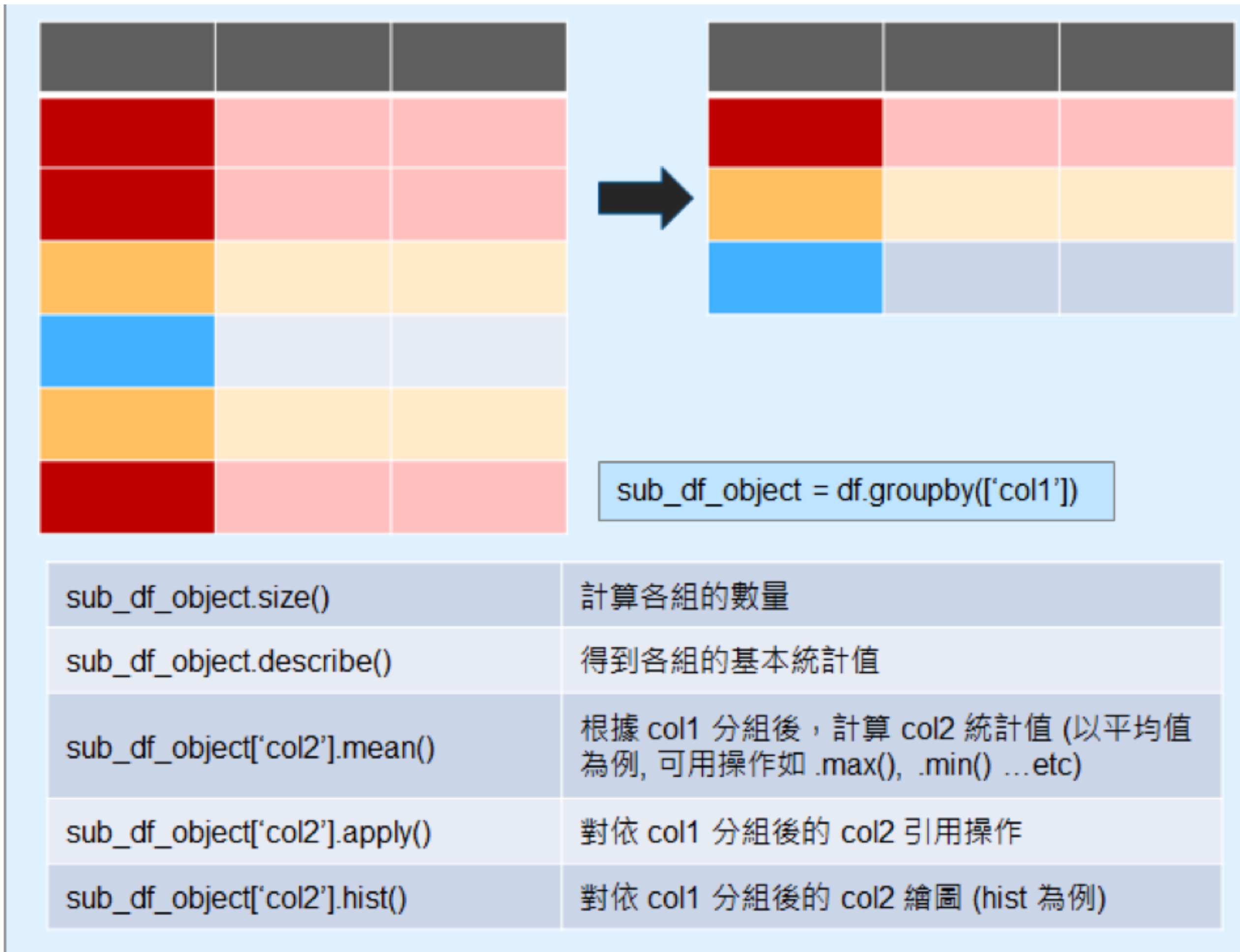
邏輯操作	
大於 / 小於 / 等於	>, <, ==
大於等於 / 小於等於	>=, <=
不等於	!=
&, , ~, ^	邏輯的 and, or, not, xor
欄位中包含 value	df.column.isin(value)
為 Nan	pd.isnull(obj)
非 Nan	pd.notnull(obj)





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 = ...)

Group operations



圖片來源: [Pandas Cheat Sheet](#)

解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

