

Day 8

# 資料清理數據前處理

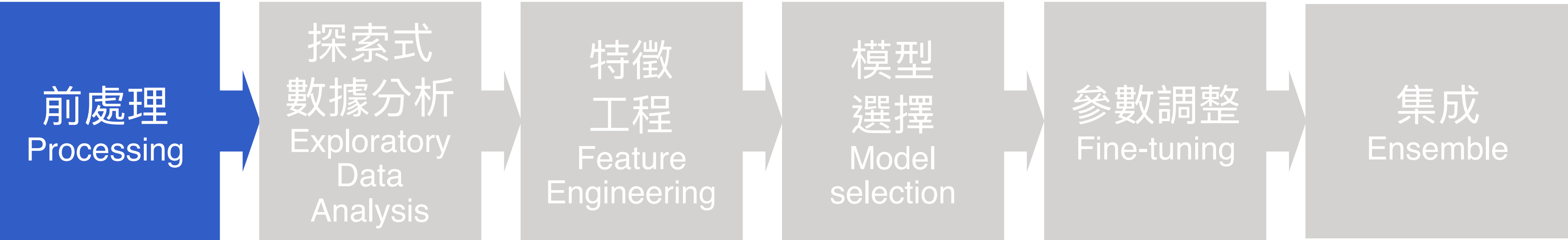
## 常用的 DataFrame 操作



# 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

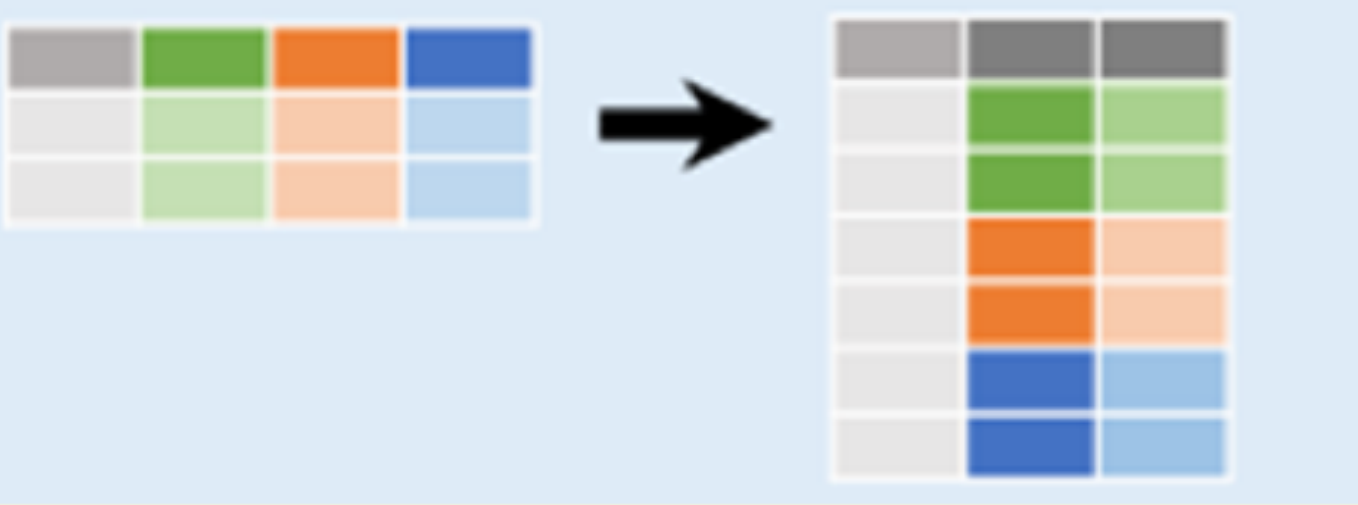


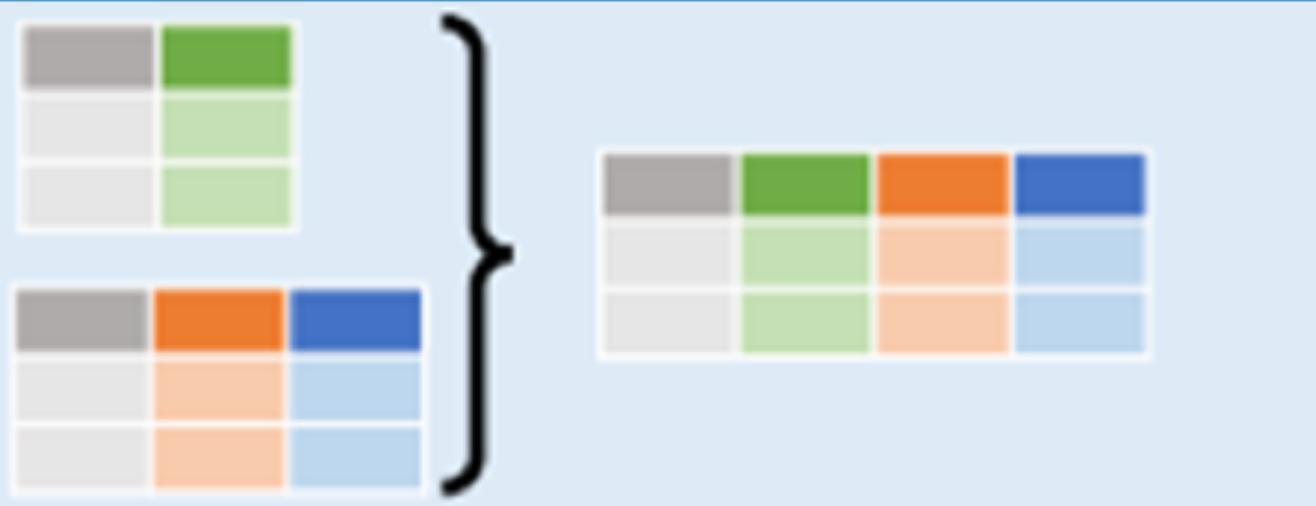
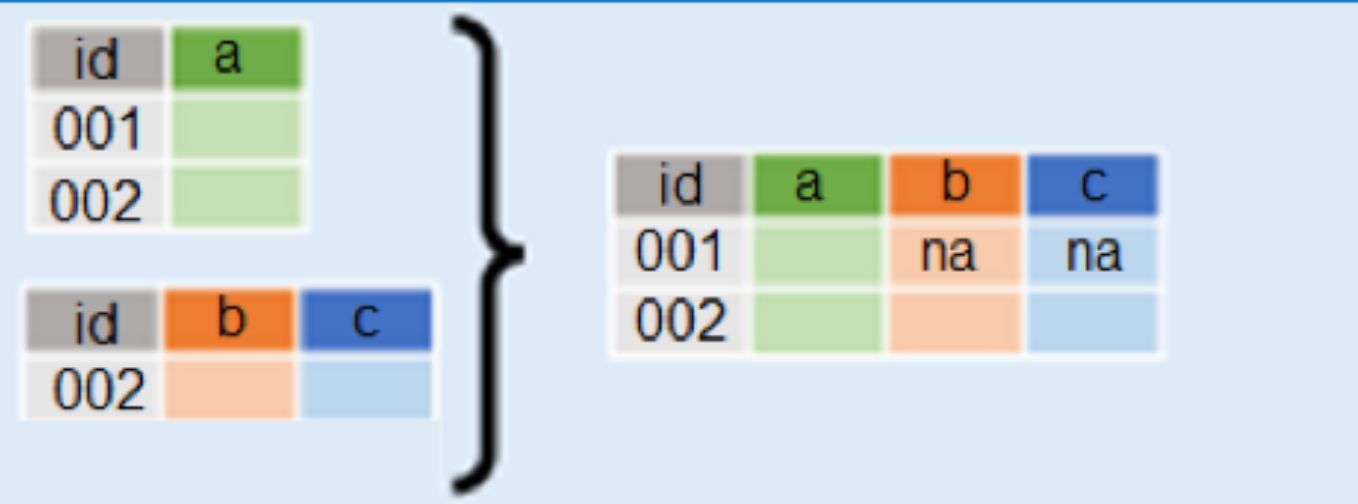
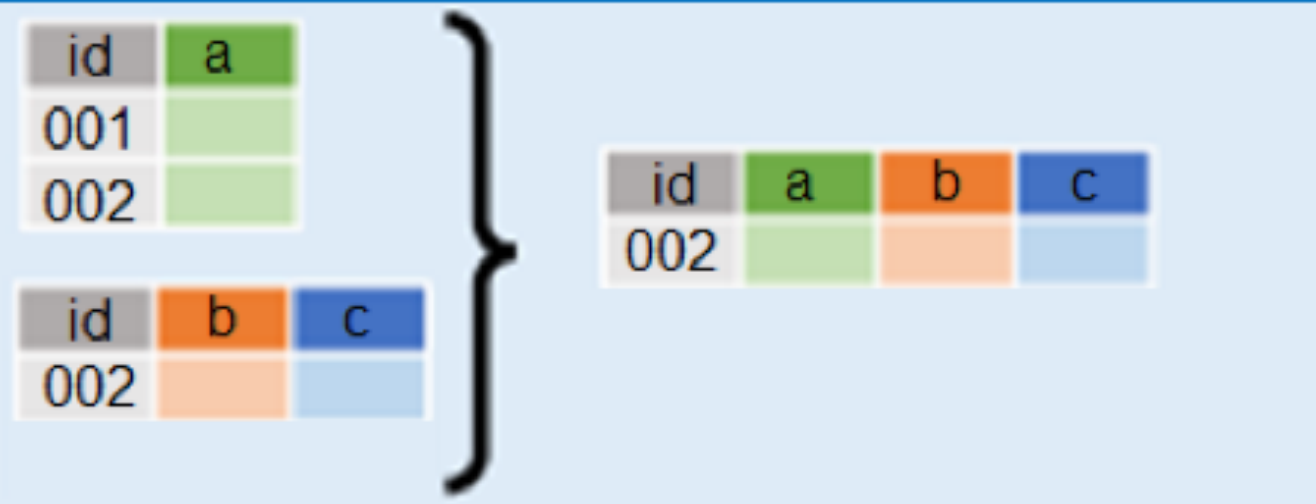




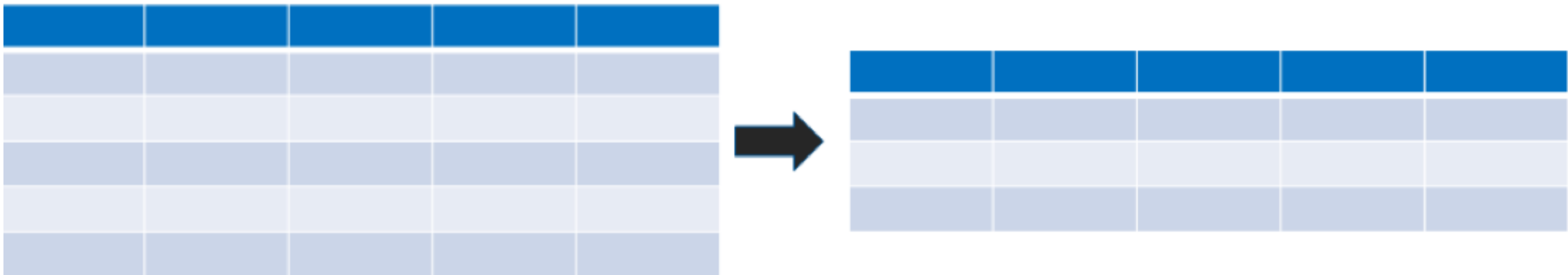
# 本日知識點目標

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

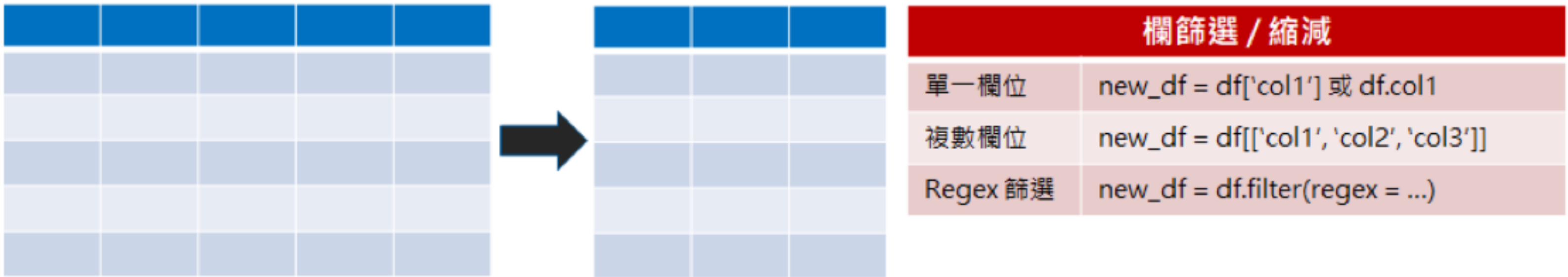
	
<p><code>pd.melt(df)</code> 將"欄" (column) 轉成"列" (row)</p>	<p><code>pd.pivot(columns='var', values='val')</code> 將"列" (row) 轉成 "欄" (column)</p>
	
<p><code>pd.concat([df1, df2])</code> 沿"列" (row) 合併兩個 dataframe</p>	<p><code>pd.concat([df1, df2], axis = 1)</code> 沿"欄" (column) 合併兩個 dataframe</p>
	
<p><code>pd.merge(df1,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

# Subset



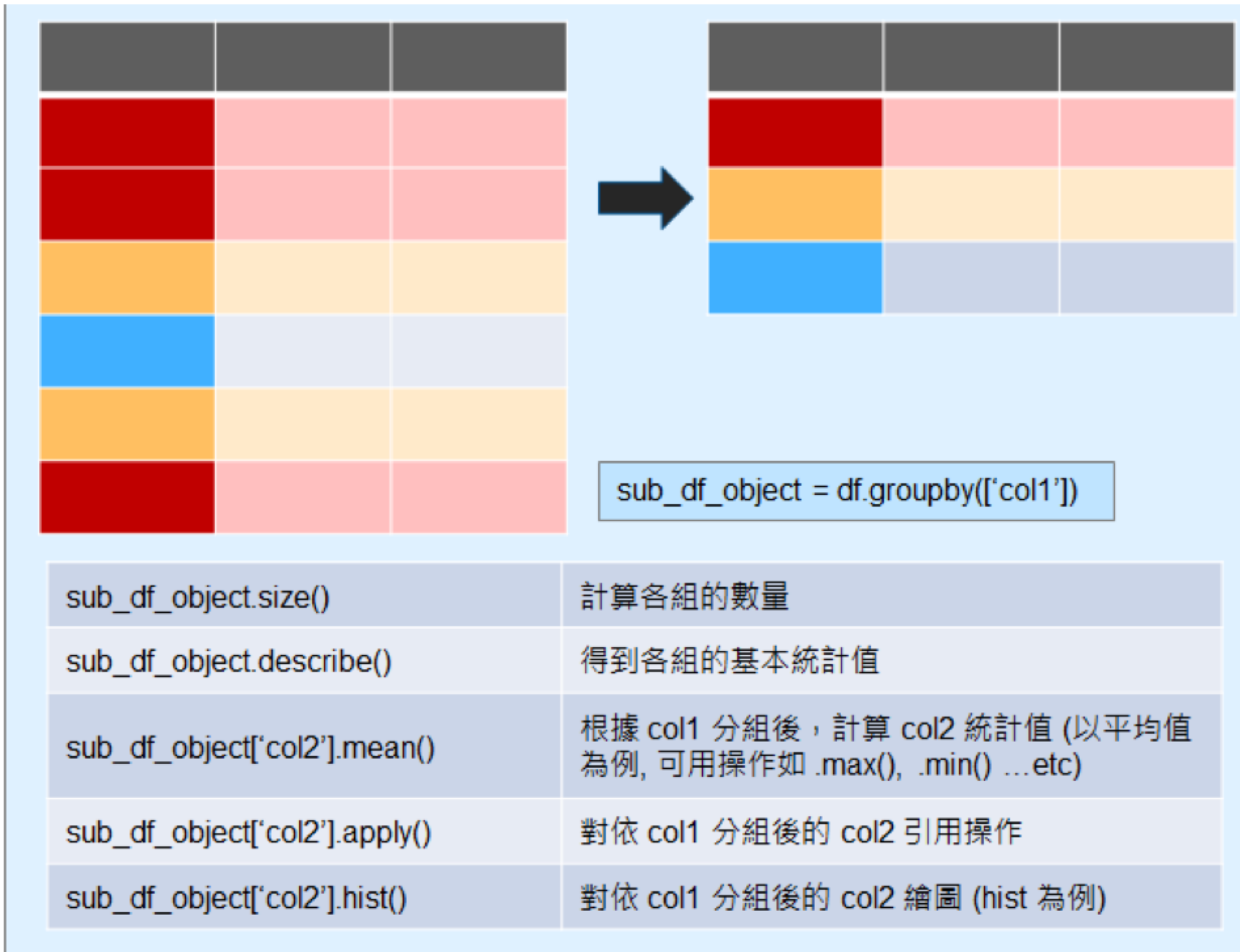
列篩選 / 縮減	
邏輯操作	<code>sub_df = df[df.age &gt; 20]</code>
移除重複	<code>df = df.drop_duplicates()</code>
前 n 筆	<code>sub_df = df.head(n = 10)</code>
後 n 筆	<code>sub_df = df.tail(n = 10)</code>
隨機抽樣	<code>sub_df = df.sample(frac = 0.5) # 抽 50 %</code>
	<code>sub_df = df.sample(n = 10) # 抽 10 筆</code>
第 n 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

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



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

# Group operations

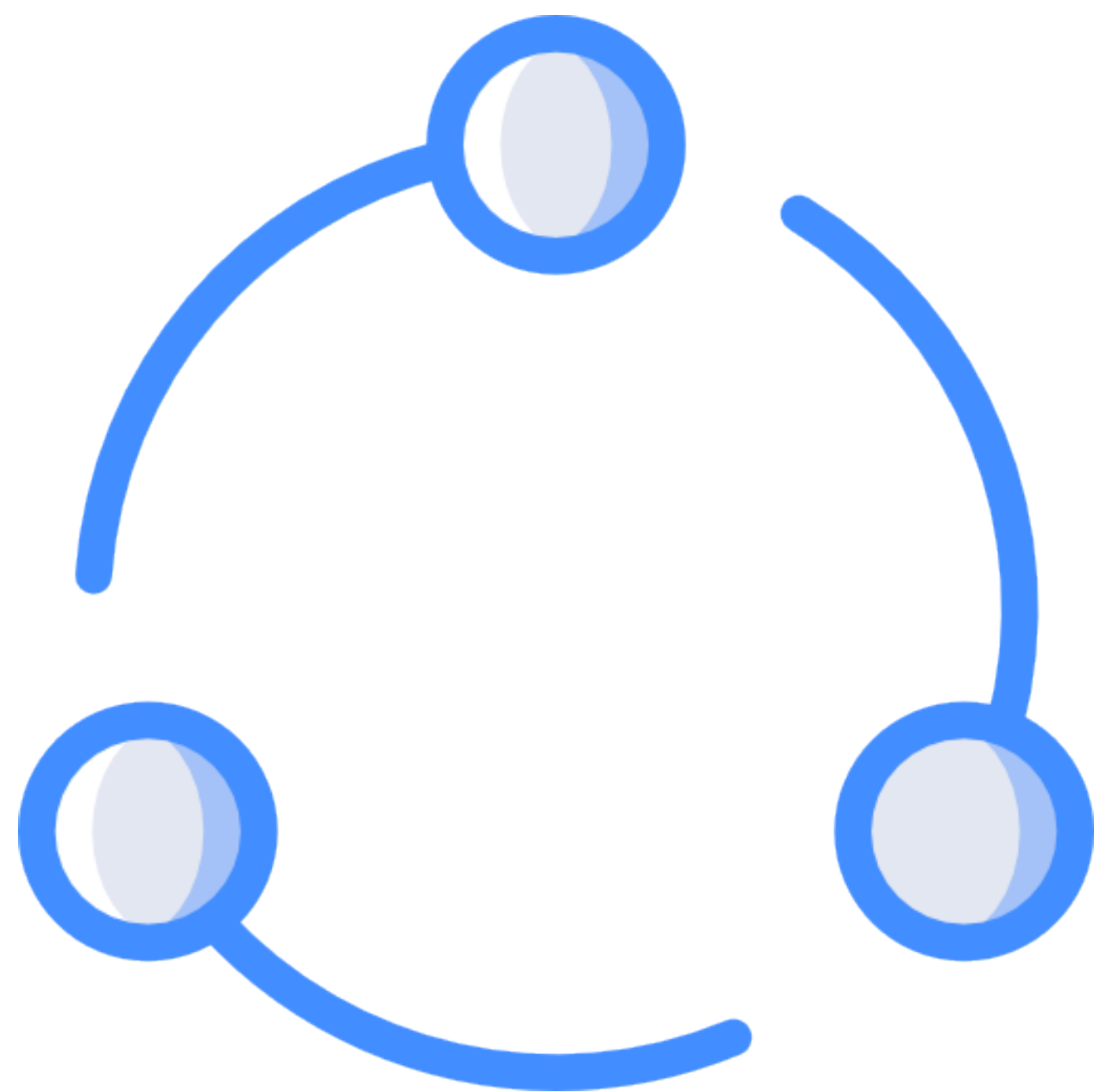


圖片來源: [Pandas Cheat Sheet](#)



# 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  $>$ ,  $==$ ,  $<$ ,  $\sim$ ) 都是可以在 pandas 作為篩選條件使用

# 解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

