D15 pandas Split-Apply-Combine Strategy

簡報閱讀 範例與作業 問題討論 學習心得(完成)

- 重要知識點
- 認識 groupby
- 認識 Split-Apply-Combine 策略
- Groupby 針對多個欄位做分析
- Groupby 針對欄位做多個分析
- Groupby 同時針對多個欄位做多個分析
- 知識點回顧
- 參考資料
- 延伸閱讀

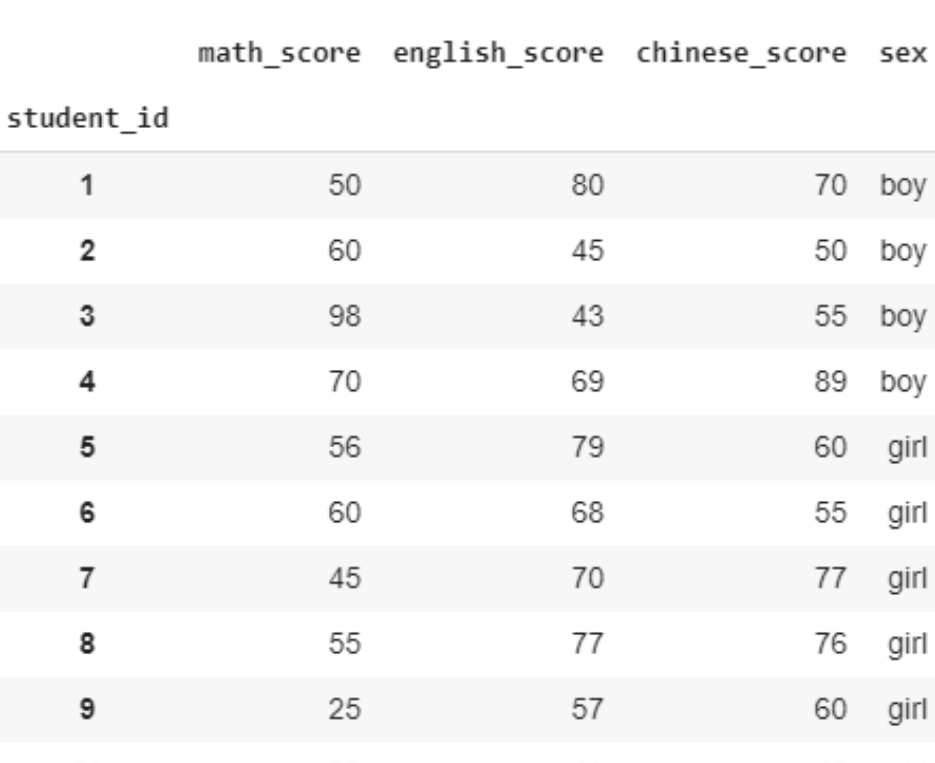
重要知識點

介紹如何透過 pandas groupby 函式實現資料科學的 Split-Apply-Combine 策略

認識 groupby

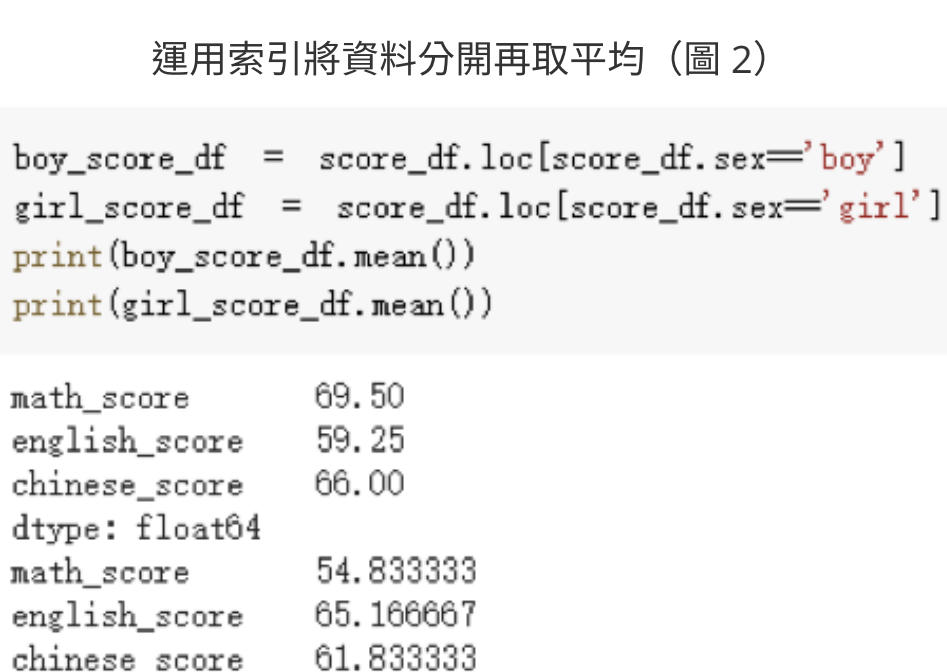
在數據分析中時常會分析不同族群的資料，例如，學生分數資料(如表 1)，你想分析男生與女生的各科差異，前幾天有教到檢索可以將資料分成男生資料與女生資料，在將各資料算平均值(如圖 2)，在這裡有一個函數 goupby 可以一行指令執行以上的邏輯(下圖 3)。

（表 1）



student_id	math_score	english_score	chinese_score	sex
1	50	80	70	boy
2	60	45	50	boy
3	98	43	55	boy
4	70	69	89	boy
5	56	79	60	girl
6	60	68	55	girl
7	45	70	77	girl
8	55	77	76	girl
9	25	57	60	girl
10	88	40	43	girl

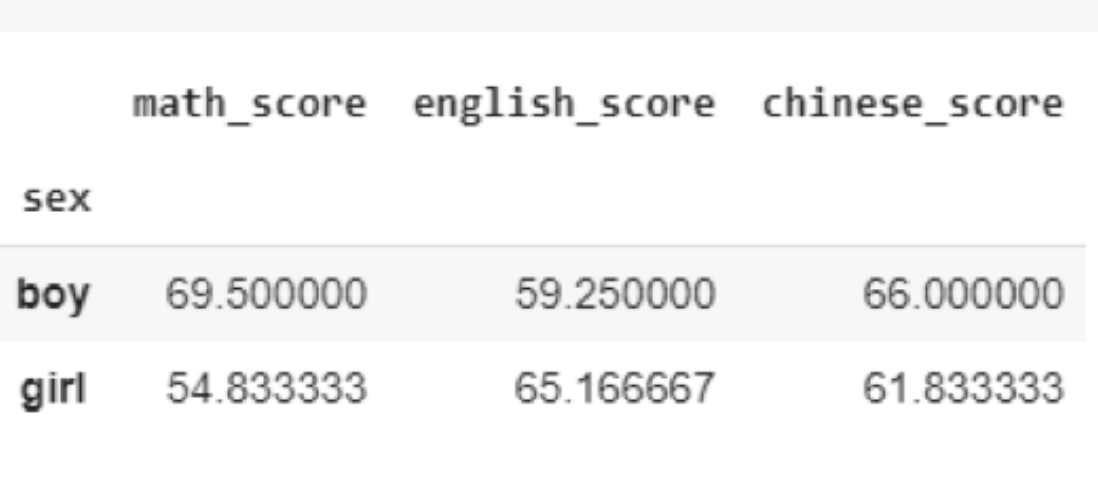
運用索引將資料分開再取平均（圖 2）



```
boy_score_df  =  score_df.loc[score_df.sex=='boy']
girl_score_df  =  score_df.loc[score_df.sex=='girl']
print(boy_score_df.mean())
print(girl_score_df.mean())
```

```
math_score        69.50
english_score     59.25
chinese_score     66.00
dtype: float64
math_score        54.833333
english_score     65.166667
chinese_score     61.833333
dtype: float64
```

運用 groupby 平均（圖 3）



```
score_df.groupby('sex').mean()
```

sex	math_score	english_score	chinese_score
boy	69.500000	59.250000	66.000000
girl	54.833333	65.166667	61.833333

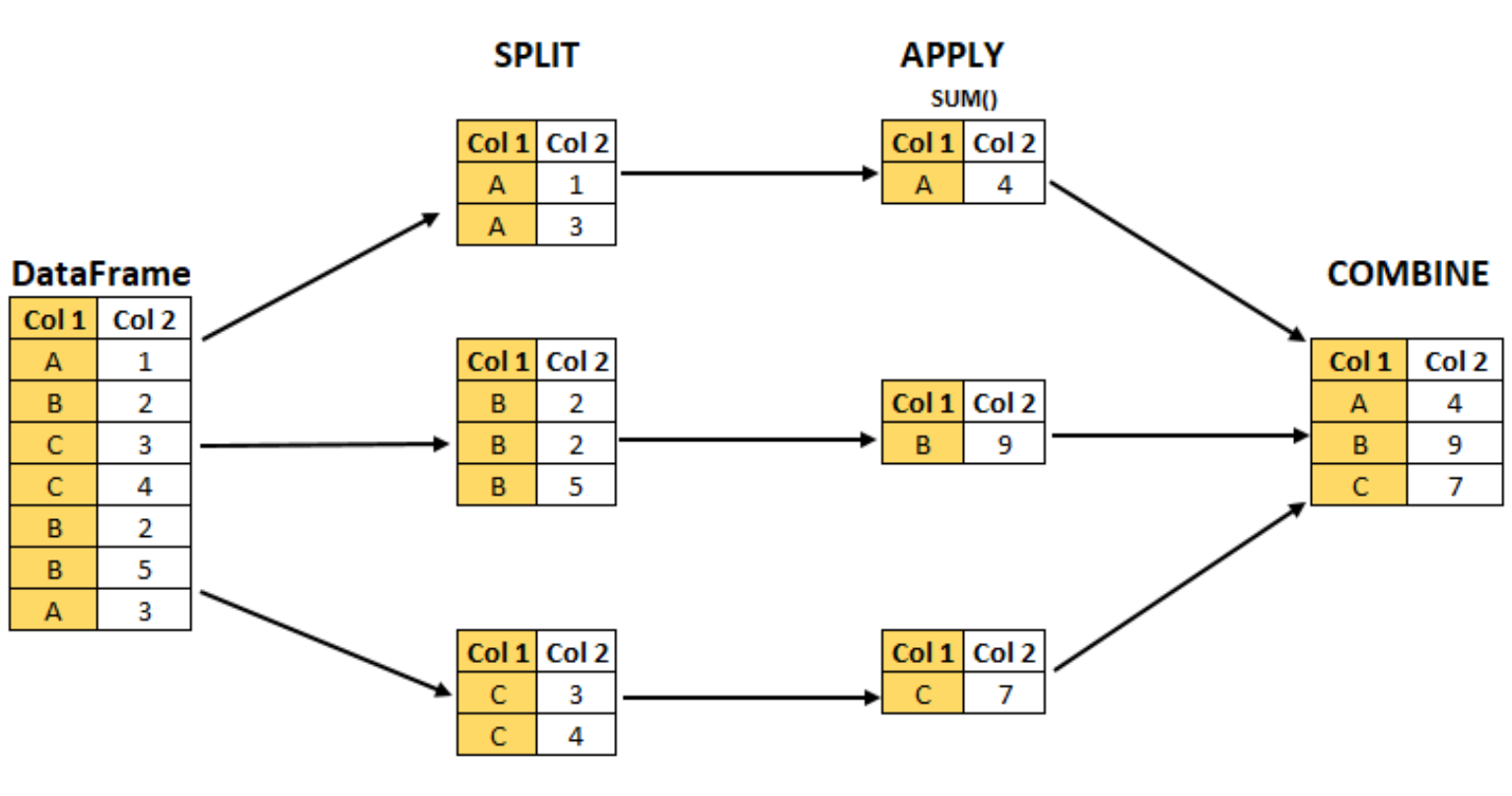
認識 Split-Apply-Combine 策略

以剛剛學生資料來分解一下 groupby 的邏輯過程

- Split：將大的數據集拆成可獨立計算的小數據集(拆成男生、女生資料)
- Apply：獨立計算各個小數據集(成績取平均)
- Combine：將小數據集運算結果合併
-

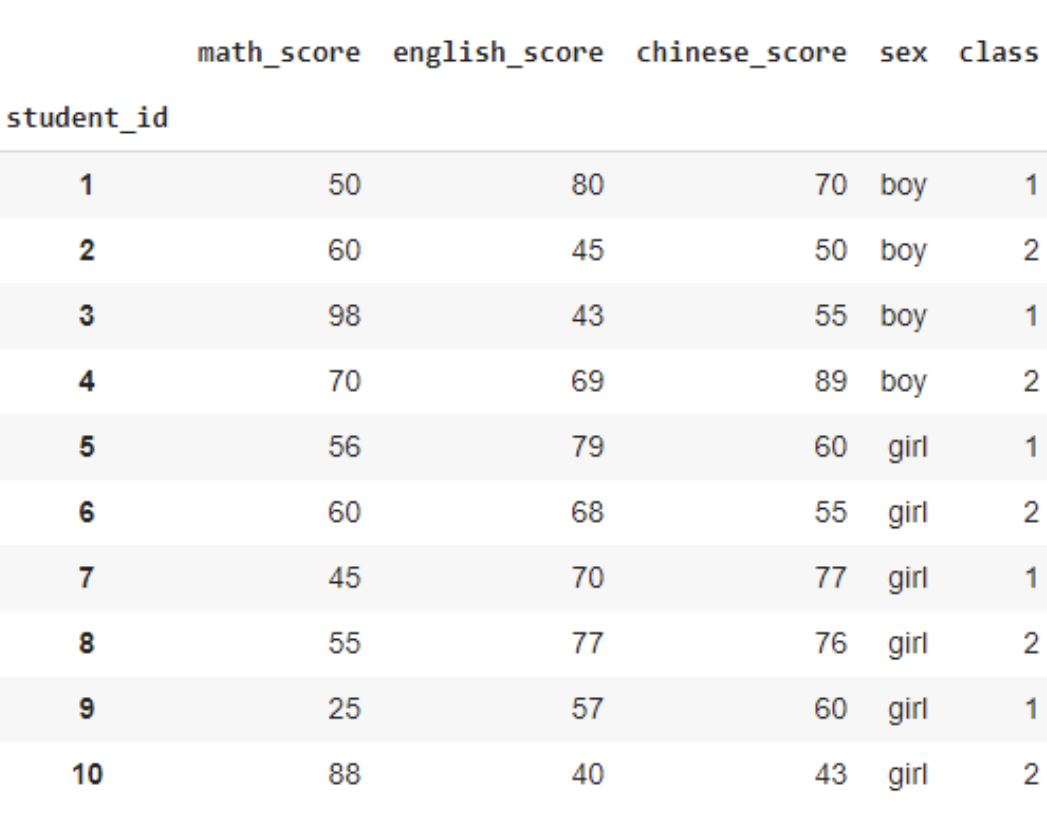
將 DataFrame 依照 A、B、C 拆成三個小數據集[split]，各自計算總合[Apply]，合併結果輸出[Combine]

- 拆分成 A、B、C 小數據集的方法為 groupby

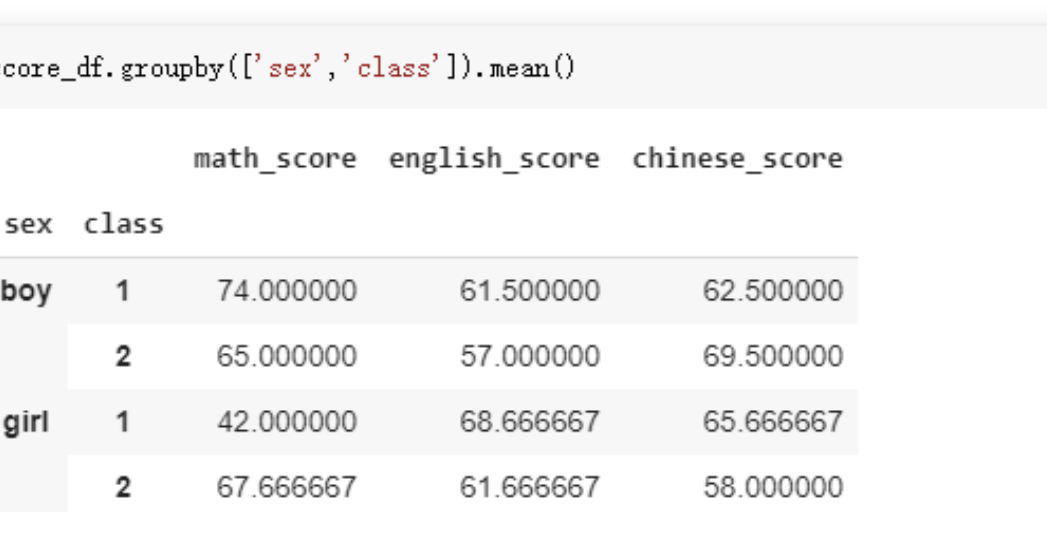


Groupby 針對多個欄位做分析

Groupby 也可以**針對多個欄位做分析**，例如，學生成績資料多一欄位 c 班級(class)，想對班級以及性別做分類，在 groupby 中加入兩個欄位名稱即可(如下圖)，此時 groupby 自動會生成多維度索引(multiple index)



student_id	math_score	english_score	chinese_score	sex	class
1	50	80	70	boy	1
2	60	45	50	boy	2
3	98	43	55	boy	1
4	70	69	89	boy	2
5	56	79	60	girl	1
6	60	68	55	girl	2
7	45	70	77	girl	1
8	55	77	76	girl	2
9	25	57	60	girl	1
10	88	40	43	girl	2

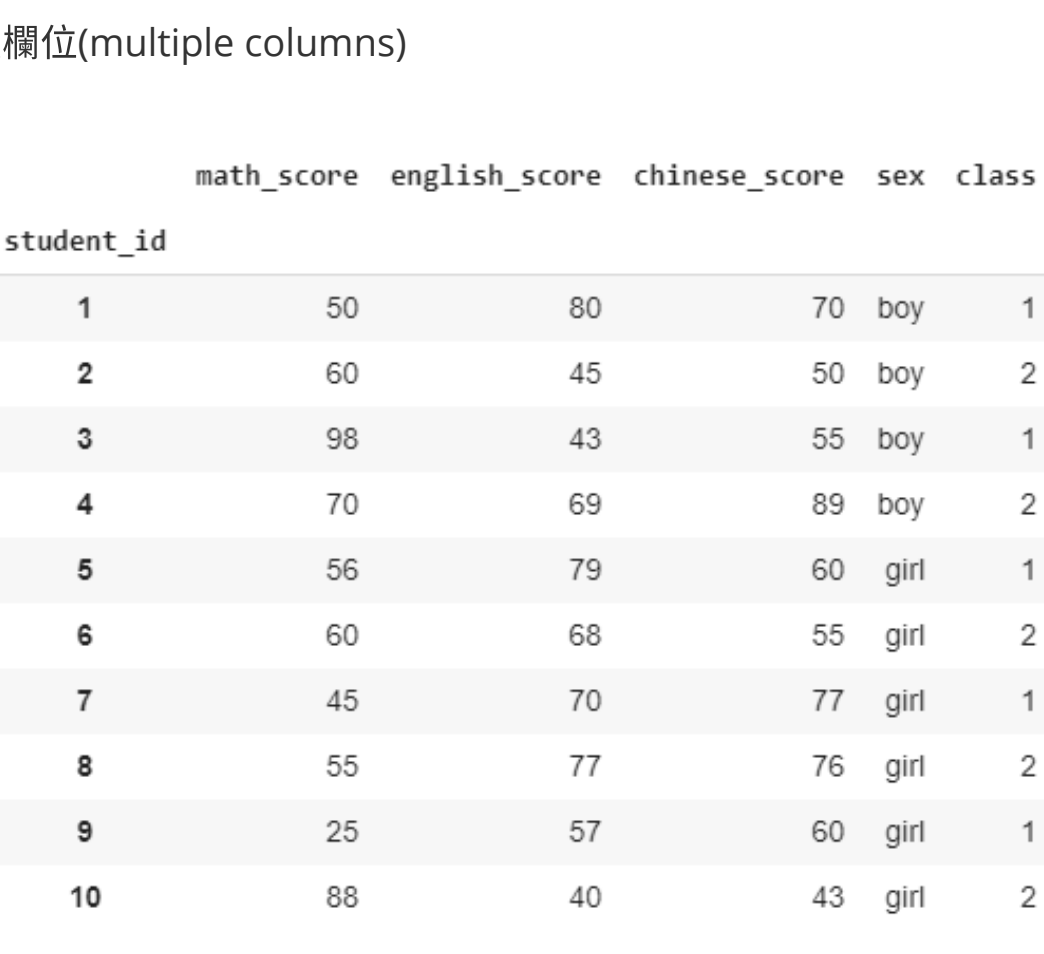


```
score_df.groupby(['sex','class']).mean()
```

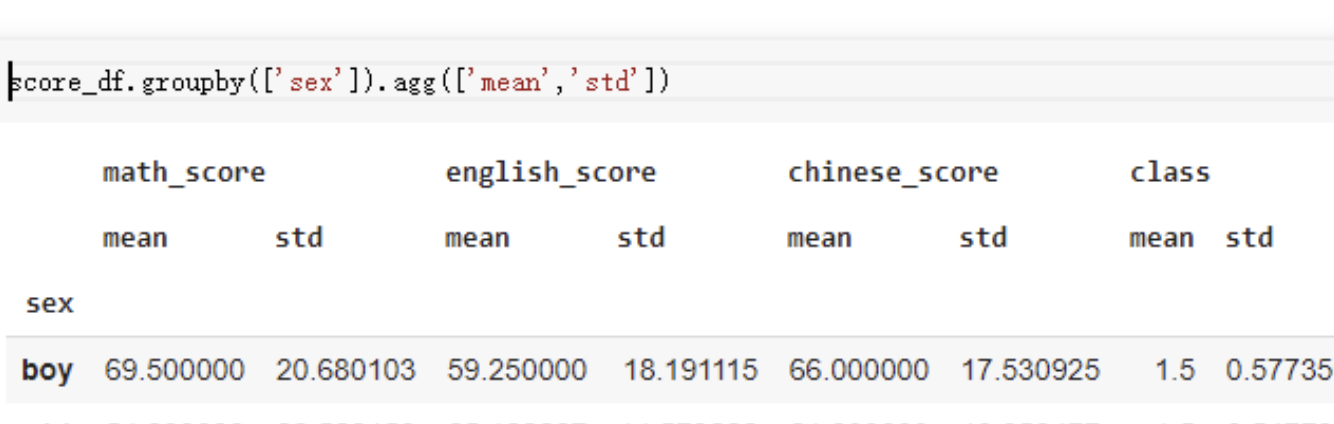
sex	class	math_score	english_score	chinese_score
boy	1	74.000000	61.500000	62.500000
	2	65.000000	57.000000	69.500000
girl	1	42.000000	68.666667	65.666667
	2	67.666667	61.666667	58.000000

Groupby 針對欄位做多個分析

Groupby也可以**針對欄位做多個分析**，例如，學生成績資料，想針對性別做成績平均以及標準差的計算，在 groupby 後加入 agg() (如下圖)，在 agg 中加入計算的邏輯(mean,std)，此時 groupby 自動會生成多維度欄位(multiple columns)



student_id	math_score	english_score	chinese_score	sex	class
1	50	80	70	boy	1
2	60	45	50	boy	2
3	98	43	55	boy	1
4	70	69	89	boy	2
5	56	79	60	girl	1
6	60	68	55	girl	2
7	45	70	77	girl	1
8	55	77	76	girl	2
9	25	57	60	girl	1
10	88	40	43	girl	2

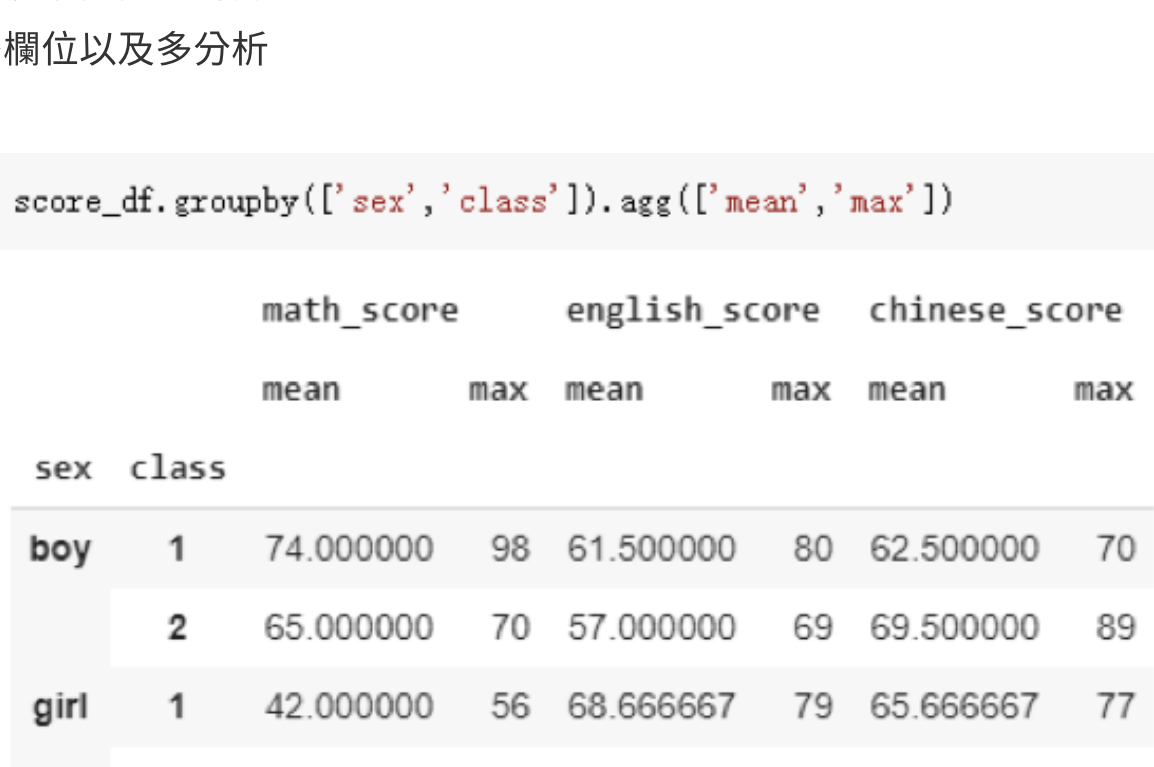


```
score_df.groupby(['sex']).agg(['mean','std'])
```

	math_score		english_score		chinese_score		class	
sex	mean	std	mean	std	mean	std	mean	std
boy	69.500000	20.680103	59.250000	18.191115	66.000000	17.530925	1.5	0.577350
girl	54.833333	20.566153	65.166667	14.579666	61.833333	12.952477	1.5	0.547723

Groupby 同時針對多個欄位做多個分析

- Groupby 也可以**同時針對多個欄位做多個分析**，例如，學生成績資料，想針對性別、班級做成績平均以及最高分的計算
- 合併了多欄位以及多分析



```
score_df.groupby(['sex','class']).agg(['mean','max'])
```

		math_score		english_score		chinese_score	
sex	class	mean	max	mean	max	mean	max
boy	1	74.000000	98	61.500000	80	62.500000	70
	2	65.000000	70	57.000000	69	69.500000	89
girl	1	42.000000	56	68.666667	79	65.666667	77
	2	67.666667	88	61.666667	77	58.000000	76

知識點回顧

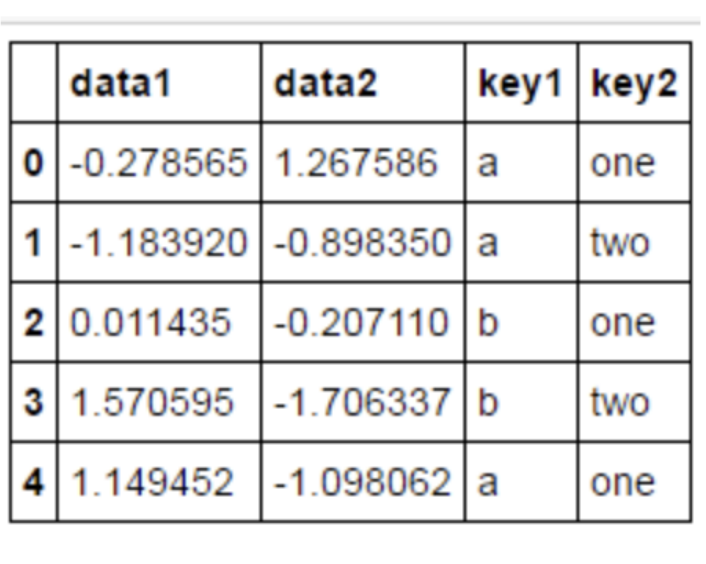
- Groupby 可以拆成
  - Split：將大的數據集拆成可獨立計算的小數據集
  - Apply：獨立計算各個小數據集
  - Combine：將小數據集運算結果合併
- Groupby 可以**同時針對多個欄位做多個分析**

參考資料

groupby
網站：python/pandas數據挖掘（十四）-groupby,聚合，分組級運算

groupby

```
import pandas as pd
df = pd.DataFrame({'key1':list('aabba'),
                   'key2': ['one','two','one','two','one'],
                   'data1': np.random.randn(5),
                   'data2': np.random.randn(5)})
df
```



	data1	data2	key1	key2
0	-0.278565	1.267586	a	one
1	-1.183920	-0.898350	a	two
2	0.011435	-0.207110	b	one
3	1.570595	-1.706337	b	two
4	1.149452	-1.098062	a	one

Split-Apply-Combine Strategy for Data Mining

In a typical exploratory data analysis, we approach the problem by dividing the data set at some granular level and then aggregating the data at that granularity in order to understand the central tendency. Similarly, a famous (must read) paper by, Hadley Wickham, outlines split-apply-combine strategy as one of the most common strategies in data analysis. Be it Marketing Segmentation, or any Behavioral Research, we use this technique at some point during our analysis.

延伸閱讀

Pandas 分組（GroupBy）
網站：易百教程

任何分組（*groupby*）操作都涉及原始對象的以下操作之一。它們是-

- 分割對象
- 應用一個函數
- 結合的結果

在許多情況下，我們將數據分成多個集合，並在每個子集上應用一些函數。在應用函數中，可以執行以下操作-

- *聚合* -計算匯總統計
- *轉換* -執行一些特定於組的操作
- *過濾* -在某些情況下以下數據

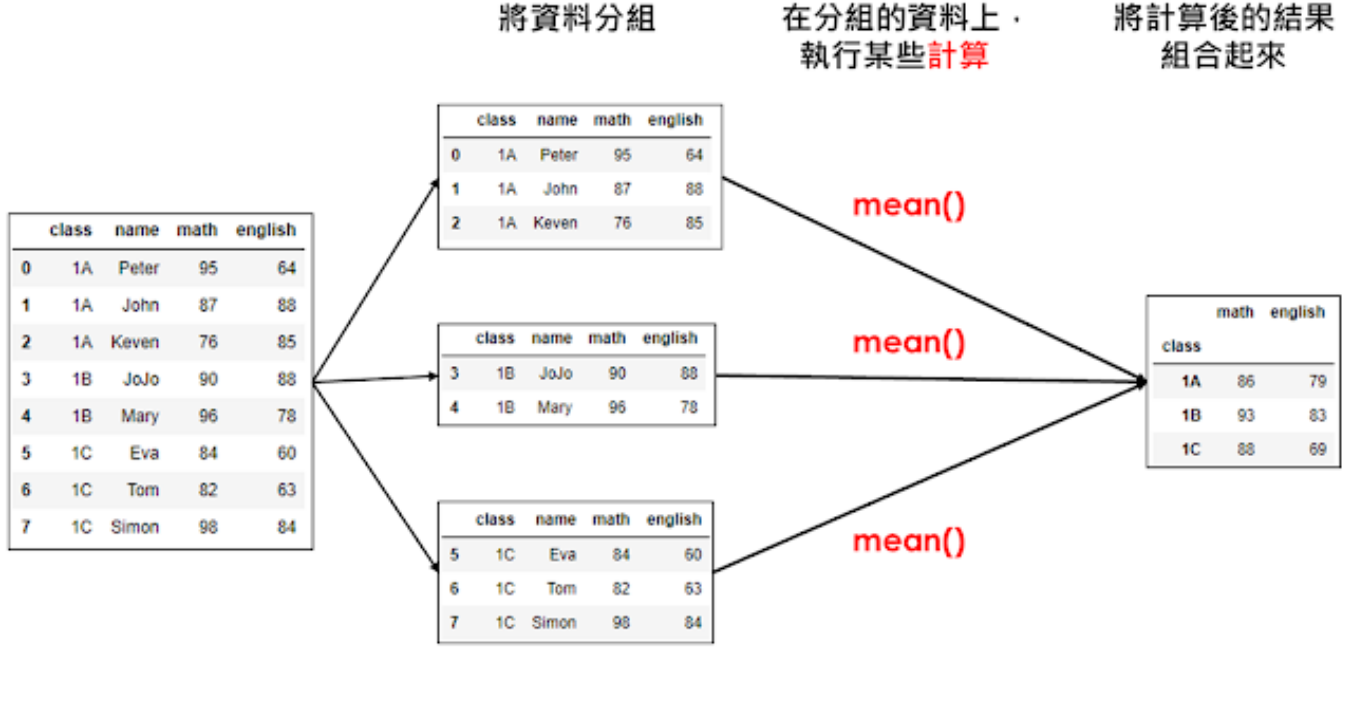
下面來看看創建一個DataFrame對象並進行執行所有操作-

```
import pandas as pd

ipl_data = {'Team': ['Riders', 'Riders', 'Devils', 'Devils', 'Kings',
         'kings', 'Kings', 'Kings', 'Riders', 'Royals', 'Royals', 'Riders'],
         'Rank': [1, 2, 2, 3, 3,4 ,1 ,1,2 , 4,1,2],
         'Year': [2014,2015,2014,2015,2014,2015,2016,2017,2016,2014,2015,2017],
         'Points':[876,789,863,673,741,812,756,788,694,701,804,690]}
df = pd.DataFrame(ipl_data)

print (df)
```

執行上面的示例代碼，得到以下結果-

```
    Points  Rank    Team  Year
0      876     1  Riders  2014
1      789     2  Riders  2015
2      863     2  Devils  2014
3      673     3  Devils  2015
4      741     3   Kings  2014
5      812     4   kings  2015
6      756     1   Kings  2016
7      788     1   Kings  2017
8      694     2  Riders  2016
9      701     4  Royals  2014
10     804     1  Royals  2015
11     690     2  Riders  2017
```

Pandas 的 groupby 語法
網站：Justim的喃喃自語



	class	name	math	english
0	1A	Peter	95	64
1	1A	John	87	88
2	1A	Keven	76	85
3	1B	JoJo	90	88
4	1B	Mary	96	78
5	1C	Eva	84	60
6	1C	Tom	82	63
7	1C	Simon	98	84

class	math	english
1A	86	79
1B	93	83
1C	88	69

下一步：閱讀範例與完成作業