



# Pandas合併、分組

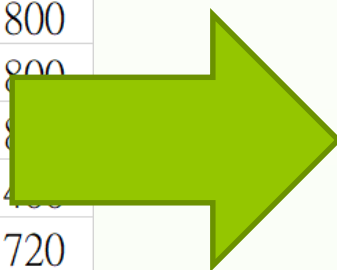
製作人：黃宥輔

日期	客戶編號	貨品編號	數量	單價
2020-01-02 00:00:00	C0001	B	1	720
2020-01-02 00:00:00	C0002	C	1	1400
2020-01-02 00:00:00	C0003	A	4	550
2020-01-02 00:00:00	C0004	B	1	750
2020-01-02 00:00:00	C0005	B	1	800
2020-01-02 00:00:00	C0006	B	1	800
2020-01-02 00:00:00	C0007	B	1	800
2020-01-02 00:00:00	C0008	B	1	800
2020-01-02 00:00:00	C0009	A	2	400
2020-01-02 00:00:00	C0001	B	1	720
2020-01-02 00:00:00	C0009	C	1	970
2020-01-02 00:00:00	C0009	B	1	670

日期	客戶編號	貨品編號	數量	單價
2020-01-02 00:00:00	C0003	A	4	550
2020-01-02 00:00:00	C0009	A	2	400
2020-01-02 00:00:00	C0009	A	1	400
2020-01-03 00:00:00	C0009	A	1	400

日期	客戶編號	貨品編號	數量	單價
2020-01-02 00:00:00	C0001	B	1	720
2020-01-02 00:00:00	C0004	B	1	750
2020-01-02 00:00:00	C0005	B	1	800
2020-01-02 00:00:00	C0006	B	1	800
2020-01-02 00:00:00	C0007	B	1	800

日期	客戶編號	貨品編號	數量	單價
2020-01-02 00:00:00	C0002	C	1	1400
2020-01-02 00:00:00	C0009	C	1	970
2020-01-02 00:00:00	C0011	C	1	1200
2020-01-03 00:00:00	C0012	C	3	1050
2020-01-07 00:00:00	C0022	C	1	1400
2020-01-10 00:00:00	C0029	C	1	1100
2020-01-13 00:00:00	C0011	C	1	1200
2020-01-13 00:00:00	C0005	C	1	1050



項目	程式碼
1	# -*- coding: utf-8 -*-
2	import pandas as pd
3	
4	df = pd.read_excel("資料/銷售資料.xlsx")
5	
6	df_group = df.groupby(by = ["貨品編號"]) # 以貨品編號將資料分組
7	
8	group_keys = df_group.groups.keys() # 分組後每一組的key
9	print(group_keys)
10	print("-----")

```
dict_keys(['A', 'B', 'C'])
```

## 資料分組

項目	程式碼
12	for k in group_keys:
13	df_goods = df_group.get_group(k)
14	print(df_goods) # 取出每一組的df
15	print("-----")

	日期	客戶編號	貨品編號	數量	單價
0	2020-01-02	C0001	B 1	720	
3	2020-01-02	C0004	B 1	750	
4	2020-01-02	C0005	B 1	800	
5	2020-01-02	C0006	B 1	800	
6	2020-01-02	C0007	B 1	800	
...	...	...	...	..	...
990	2020-08-17	C0010	B 1	700	
991	2020-08-18	C0006	B 1	700	
992	2020-08-18	C0021	B 1	700	
993	2020-08-18	C0009	B 1	600	
1001	2020-08-26	C0118	B 1	700	
[520 rows x 5 columns]					

	日期	客戶編號	貨品編號	數量	單價
2	2020-01-02	C0003	A 4	550	
8	2020-01-02	C0009	A 2	400	
14	2020-01-02	C0009	A 1	400	
21	2020-01-03	C0009	A 1	400	
22	2020-01-06	C0014	A 1	530	
...	...	...	...	..	...
1066	2020-12-02	C0009	A 1	480	
1067	2020-12-02	C0001	A 2	350	
1068	2020-12-03	C0009	A 3	480	
1069	2020-12-03	C0009	A 1	480	
1070	2020-12-07	C0039	A 1	550	
[196 rows x 5 columns]					

	日期	客戶編號	貨品編號	數量	單價
1	2020-01-02	C0002	C 1	1400	
10	2020-01-02	C0009	C 1	970	
13	2020-01-02	C0011	C 1	1200	
15	2020-01-03	C0012	C 3	1050	
35	2020-01-07	C0022	C 1	1400	
...	...	...	...	..	...
976	2020-08-12	C0124	C 1	850	
978	2020-08-13	C0083	C 40	750	
982	2020-08-14	C0067	C 1	950	
1002	2020-08-27	C0083	C 60	750	
1003	2020-08-27	C0045	C 1	950	
[355 rows x 5 columns]					

## 資料分組運算



AI.FINTECH  
A I 金 融 科 技 中 心

項目	程式碼
1	# -*- coding: utf-8 -*-
2	import pandas as pd
3	
4	df = pd.read_excel("資料/銷售資料.xlsx")
5	
6	df_group = df.groupby(by = ["貨品編號"]) # 將資料以貨品編號分組
7	print(df_group[["單價", "數量"]].mean())
8	print("-----")

	單價	數量
貨品編號		
A	507.091837	2.285714
B	727.269231	1.709615
C	981.436620	1.915493



## 資料分組運算



AI.FINTECH

AI 金融科技中心

項目	程式碼
10	def cost(data):
11	# 傳入的data格式為Series或DataFrame
12	return data.mean() * 0.5 # 假設成本價是50%，計算成本
13	
14	apply_result = df_group["單價"].apply(cost) # 帶入單價來計算成本
15	print(apply_result)
16	print("-----")

貨品編號

A 253.545918

B 363.634615

C 490.718310

Name: 單價, dtype: float64

項目	程式碼
18	agg_dict = {"單價" : ['mean'],
19	"數量" : ['sum','mean']
20	}
21	agg_result = df_group.agg(agg_dict)
22	print(agg_result)
23	print("-----")

	單價		數量
	mean	sum	mean
貨品編號			
A	507.091837	448	2.285714
B	727.269231	889	1.709615
C	981.436620	680	1.915493

項目	程式碼
1	# -*- coding: utf-8 -*-
2	import pandas as pd
3	
4	df_base = pd.read_excel("資料/銷售資料.xlsx")
5	df_ex = pd.read_excel("資料/擴充銷售資料.xlsx")
6	print(df_base.shape)
7	print(df_ex.shape)
8	
9	df = pd.concat([df_base, df_ex], axis = 0) # 0預設為縱向合併
10	print(df.shape)

(1071, 5)  
(28022, 6)  
(29093, 6)



項目	程式碼
12	<code>df["銷售額"] = df["單價"] * df["數量"] # 計算銷售額</code>
13	<code>df["年"] = df["日期"].dt.year # 取得年</code>
14	<code>df["月"] = df["日期"].dt.month # 取得月</code>
15	
16	<code># 以年、月貨品編號將資料分組</code>
17	<code>df_group = df.groupby(by = ["年", "月", "貨品編號"])</code>
18	
19	<code>df_sales = df_group["銷售額"].mean()</code>
20	<code>print(df_sales)</code>

```

年    月    貨品編號
2020  1    A      1080.000000
        B      1378.297872
        C      1327.222222
        D      1837.648472
        E      2699.744681
        ...
2023  4    E      2344.342105
        F      1701.851852
        5    D      4723.377892
        E      2469.938462
        F      2100.454545
Name: 銷售額, Length: 148, dtype: float64

```

項目	程式碼
1	# -*- coding: utf-8 -*-
2	import pandas as pd
3	
4	file_path = "資料/"
5	
6	df_sell = pd.read_excel(f"{file_path}銷售資料.xlsx")
7	df_goods = pd.read_excel(f"{file_path}商品資料.xlsx")
8	print(df_sell)
9	print(df_goods)

	日期	客戶編號	貨品編號	數量	單價
0	2020-01-02	C0001	B	1	720
1	2020-01-02	C0002	C	1	1400
2	2020-01-02	C0003	A	4	550
3	2020-01-02	C0004	B	1	750
4	2020-01-02	C0005	B	1	800
...	...	...	...	..	...
1066	2020-12-02	C0009	A	1	480
1067	2020-12-02	C0001	A	2	350
1068	2020-12-03	C0009	A	3	480
1069	2020-12-03	C0009	A	1	480
1070	2020-12-07	C0039	A	1	550

	貨品編號	貨品名稱
0	A	商用咖啡豆
1	B	莊園經典咖啡豆
2	C	小農藍山咖啡豆

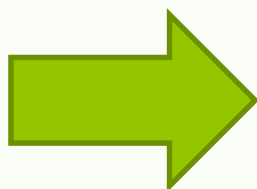
項目	程式碼
11	# 合併資料
12	df_merge = pd.merge(left = df_sell,
13	right = df_goods,
14	left_on = "貨品編號",
15	right_on = "貨品編號"
16	)
17	df_merge = df_merge.sort_values(by = ["日期"])
18	print(df_merge)

	日期	客戶編號	貨品編號	數量	單價	貨品名稱
0	2020-01-02	C0001	B	1	720	莊園經典咖啡豆
522	2020-01-02	C0011	C	1	1200	小農藍山咖啡豆
875	2020-01-02	C0003	A	4	550	商用咖啡豆
876	2020-01-02	C0009	A	2	400	商用咖啡豆
877	2020-01-02	C0009	A	1	400	商用咖啡豆
...	...	...	...	...	...	...
1066	2020-12-02	C0009	A	1	480	商用咖啡豆
1067	2020-12-02	C0001	A	2	350	商用咖啡豆
1068	2020-12-03	C0009	A	3	480	商用咖啡豆
1069	2020-12-03	C0009	A	1	480	商用咖啡豆
1070	2020-12-07	C0039	A	1	550	商用咖啡豆

日期	客戶編號	貨品編號	數量	單價
2020-01-02 00:00:00	C0001	B	1	720
2020-01-02 00:00:00	C0002	C	1	1400
2020-01-02 00:00:00	C0003	A	4	550
2020-01-02 00:00:00	C0004	B	1	750
2020-01-02 00:00:00	C0005	B	1	800
2020-01-02 00:00:00	C0006	B	1	800
2020-01-02 00:00:00	C0007	B	1	800
2020-01-02 00:00:00	C0008	B	1	800



貨品編號	貨品名稱
A	商用咖啡豆
B	莊園經典咖啡豆
C	小農藍山咖啡豆



日期	客戶編號	貨品編號	數量	單價	貨品名稱
2020-01-02 00:00:00	C0001	B	1	720	莊園經典咖啡豆
2020-01-02 00:00:00	C0011	C	1	1200	小農藍山咖啡豆
2020-01-02 00:00:00	C0003	A	4	550	商用咖啡豆
2020-01-02 00:00:00	C0009	A	2	400	商用咖啡豆
2020-01-02 00:00:00	C0009	A	1	400	商用咖啡豆
2020-01-02 00:00:00	C0009	C	1	970	小農藍山咖啡豆
2020-01-02 00:00:00	C0010	B	1	850	莊園經典咖啡豆
2020-01-02 00:00:00	C0002	C	1	1400	小農藍山咖啡豆
2020-01-02 00:00:00	C0001	B	1	720	莊園經典咖啡豆



**章節到此結束，有任何問題歡迎提出來討論！**