DAC_ship_analysis

匯入資料和查看資料型態

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
ship = pd.read_csv("/Users/shenchingfeng/Downloads/ship_information.csv")
ship.head(10)
ship.info()
```

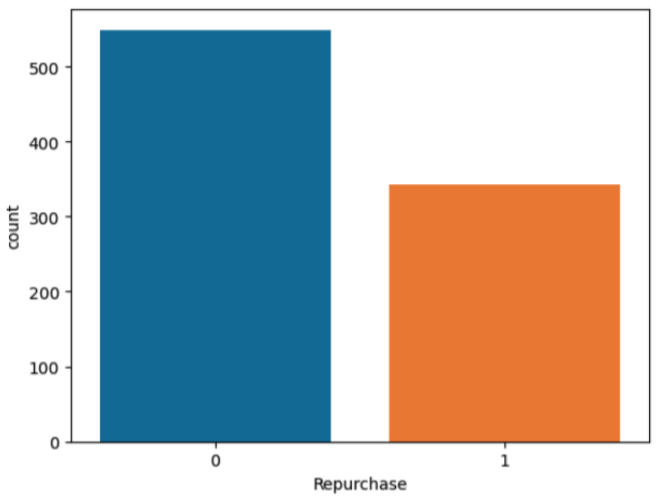
	PassengerId	Repurchase	Pclass	Sex	SibSp	Parch	Fare	Embarked
0	1	0	3	male	1	0	7.2500	S
1	2	1	1	female	1	0	71.2833	C
2	3	1	3	female	0	0	7.9250	S
3	4	1	1	female	1	0	53.1000	S
4	5	0	3	male	0	0	8.0500	S
5	6	0	3	male	0	0	8.4583	Q
6	7	0	1	male	0	0	51.8625	S
7	8	0	3	male	3	1	21.0750	S
8	9	1	3	female	0	2	11.1333	S
9	10	1	2	female	1	0	30.0708	C

資料分析

- 先看一下最重要的Repurchase的分佈

```
# Graph at right
sns.countplot(ship['Repurchase'])

# Outcome present below
pclass = ship.groupby('Pclass').Repurchase.sum() # 各階級回購量
rpc = ship.Repurchase.sum() # 總回購數
print(pclass/rpc)
```



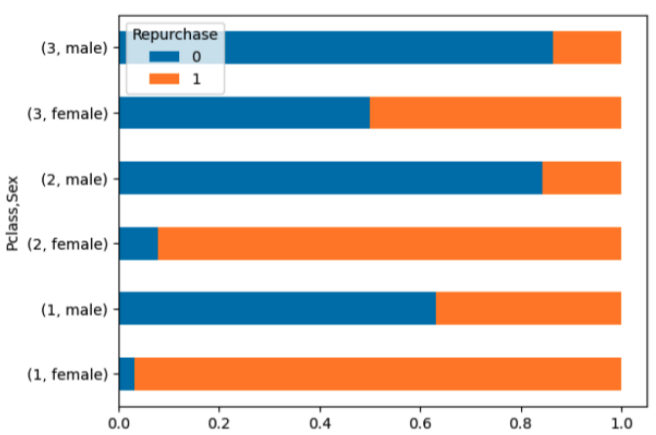
從第一組 code ，結果呈現如右，可以看到大概是 N : Y 回購 大概是 550 : 340

而第二組 code 則說明了回購和階級的關係大致為 0.4 : 0.25 : 0.35 (1 : 2 : 3)

大多數的回購都來自於：class 1: 有本錢花, class3: 拿的票價便宜, 愛花錢 等等

- 因為限定一頁，所以就跳快一點，把 Sex & Pclass 做交叉比對

```
rpc_total = pd.crosstab([ship.Pclass, ship.Sex], ship.Repurchase)
per_class = rpc_total.div(rpc_total.sum(1),axis = 0)
per_class.plot(kind = 'barh', stacked = True)
```

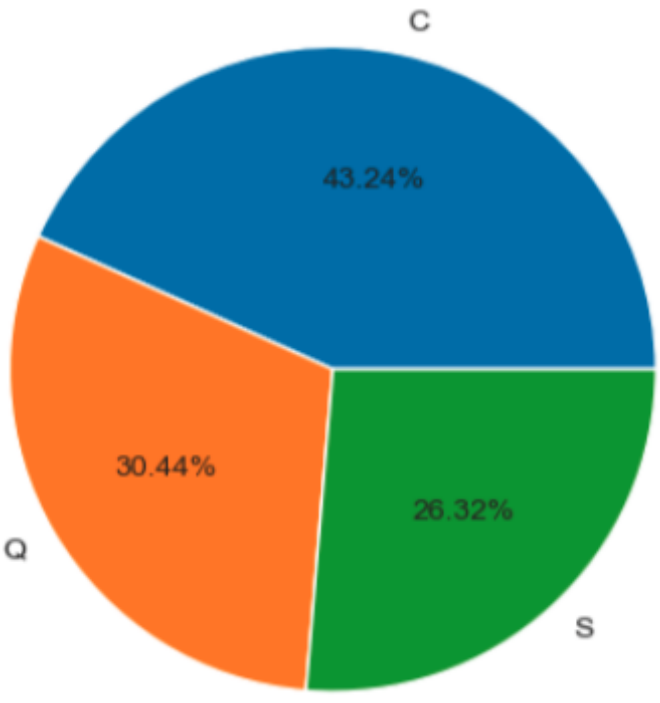


可以看到回購率最高的是來自 **1, 2 級的女性**，甚至 3 級女性回購率都大於 1 級男性

由此可以得到增加 KPI 的第一個方法：多出一些針對**女性**甚至是**貴婦**的方案

- 再來是不同港口間的分析

```
# 各港口回購量
emb_rpc = ship.groupby('Embarked').Repurchase.sum()
# 各港口出發量
emb = ship.groupby('Embarked')
emb_tot = emb.size()
# 各港口回購率
(emb_rpc/emb_tot).plot(kind = 'pie',autopct = '%.2f%%') # 取到小數第二位
```

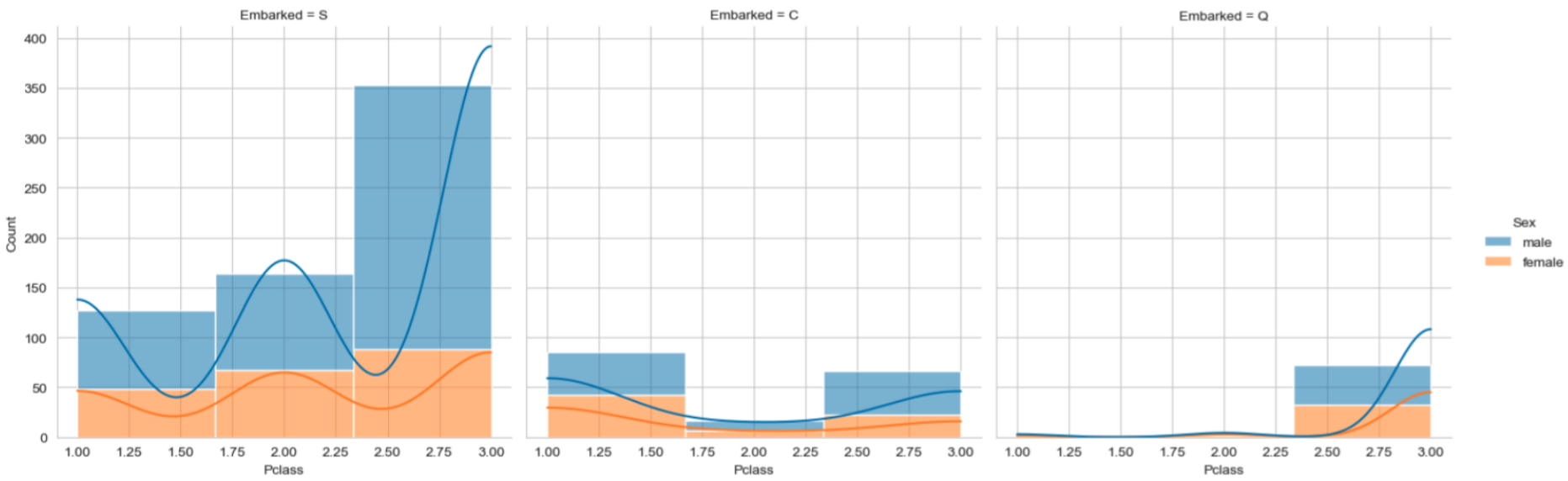


可以看到來自 **C 港** 的回購率是最高的，C > Q > S，那和 **階級** 有關係嗎？

```
sns.displot(x = 'Pclass', bins = 3, hue = 'Sex', kde = True,\
            data = ship, col = 'Embarked')
```

是有的，S 港上船的人數是最多的，但回購率卻最低，建議可以降低在 S 港的投資

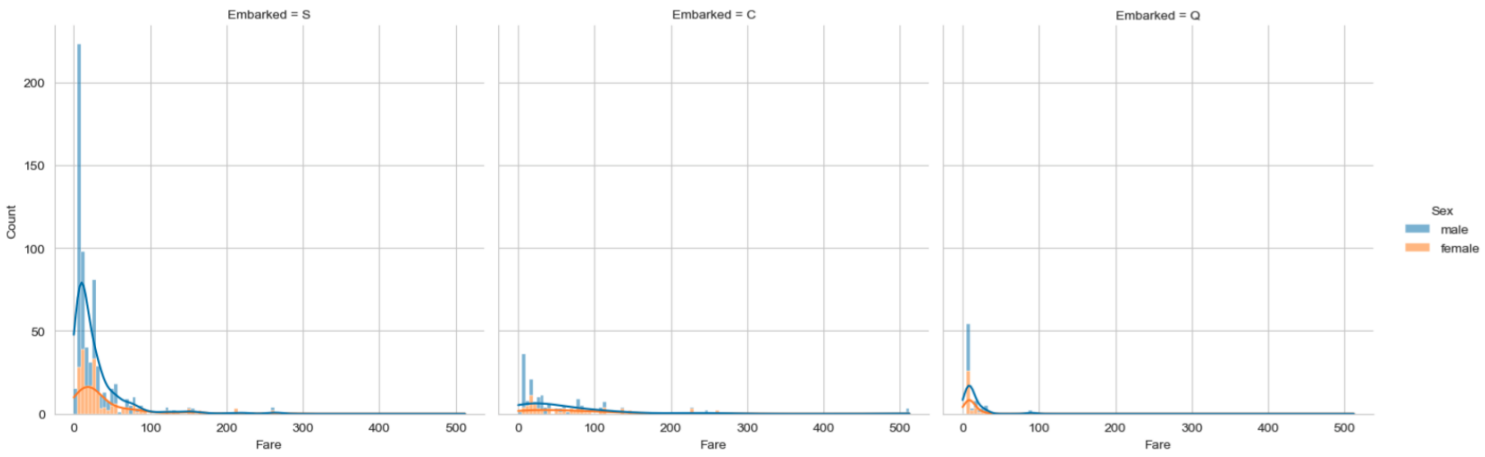
可以從下方的圖表看發現 C 港 的上船人口組成中，回購率最高的 **1, 3 級人類佔了大多數**，故建議可以**加大在 C 港的投資**！



- 票價和港口間的分析

```
emb_total = pd.crosstab([ship.Embarked, ship.Pclass], ship.Repurchase) # 以港口和階級做交叉比對
emb_total
sns.displot(x = 'Fare', kde = True, hue = 'Sex', data = ship, col = 'Embarked', multiple = 'stack')
```

Embarked	Pclass	Repurchase 0	1
C	1	26	59
	2	8	9
	3	41	25
Q	1	1	1
	2	1	2
	3	45	27
S	1	53	74
	2	88	76
	3	286	67

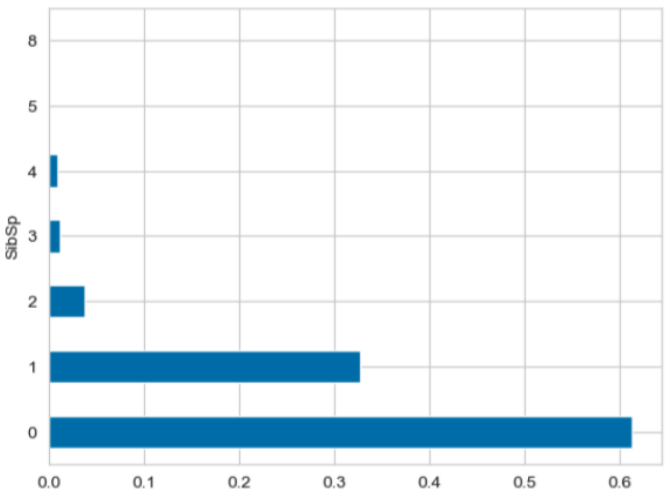


可以觀察到有錢人 (默認 class 1) 佔超過一半的 C港，平均票價是三港中最貴的，但總上船數 (168) 卻才差不多等於 S港 的 class 2 而已，故重複以上，覺得應該大力開發 **C港！**

（有一堆 Fare = 0 的是偷渡的嗎）

- 最後來看到一個很有趣的部分，手足配偶人數和回購率的關係

```
byss = ship.groupby('SibSp').Repurchase.sum()
(byss / rpc).plot(kind = 'barh') # 手足配偶人數回購率
```



可以觀察到，SibSp = 0 (單身), 1 (帶一人) 的回購率是最高的，

故應多增加和**個人**和**配偶**有相關的投資！

總結

1. 增加對**女性**市場的投資
2. 加大在 **C港** 的市場投資，降低在 S港 的支出
3. 把目標客群定調為 **“單身”** 和 **“情侶”**，而非 SciSp ≥ 2 的朋友間出遊