Spaceship Titanic

DM23-Team09
112753203 資智一 張芮瑀
111354014 統碩二 劉貞莉
111354020 統碩二 黃詩涵

TABLE OF CONTENTS

01
資料前處理
Data Preprocessing

02
探索性資料分析
EDA

03
遺失值填補
Imputation

04
特徵創建與選取
Feature Engineering

05
模型配適
Model Fitting

06
模型成果
Result

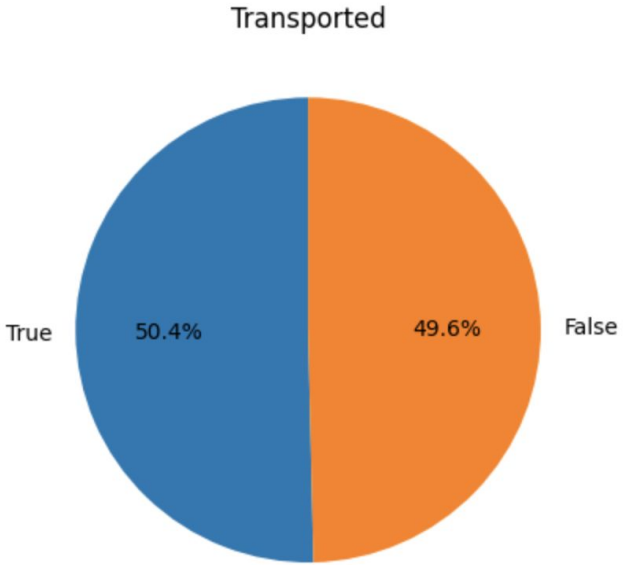
01

Data Preprocessing

資料簡介｜切分欄位資料

資料簡介

- Training set
  - 14個變數
  - 8693筆資料
  - Transported=True的比例為50.4%
- Testing set
  - 13個變數
  - 4277筆資料



切分欄位資料

- **PassengerId**:

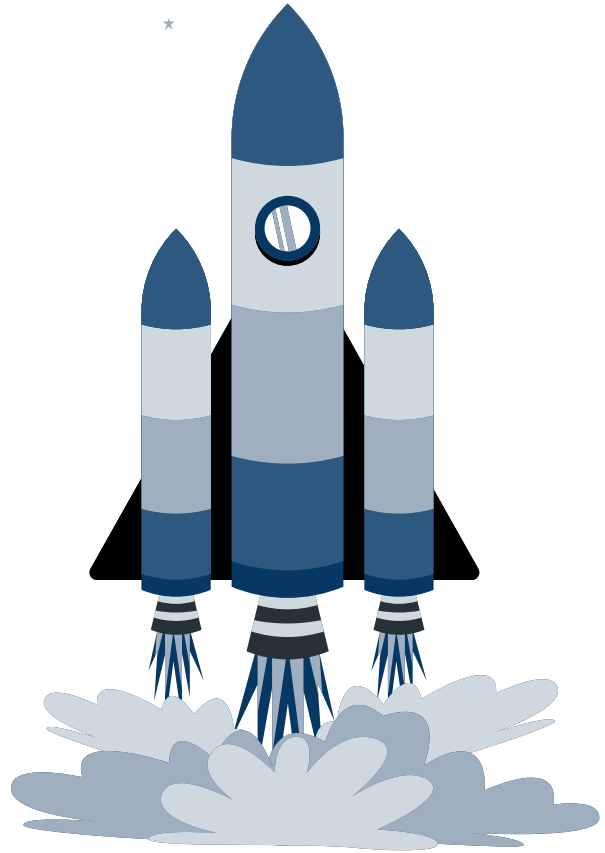
  「gggg_pp」，切分出「gggg」後統計每組人數，新增為”Group”變數

- **Cabin**:

  「deck/num/side」，以/切分出”Deck”、“Number”、”Side”三個變數

PassengerID	Group
0001_01	3
0001_02	3
0001_03	3
0002_01	2
0002_02	2

Cabin	Deck	Number	Side
B/0/P	B	0	P
F/0/S	F	0	S
F/1/S	F	1	S



02

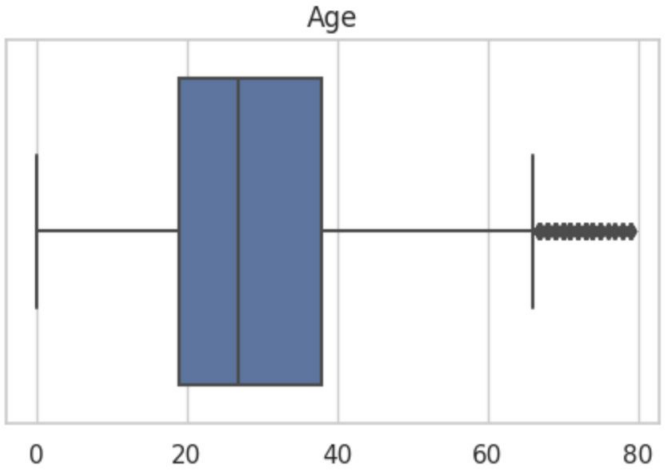
EDA

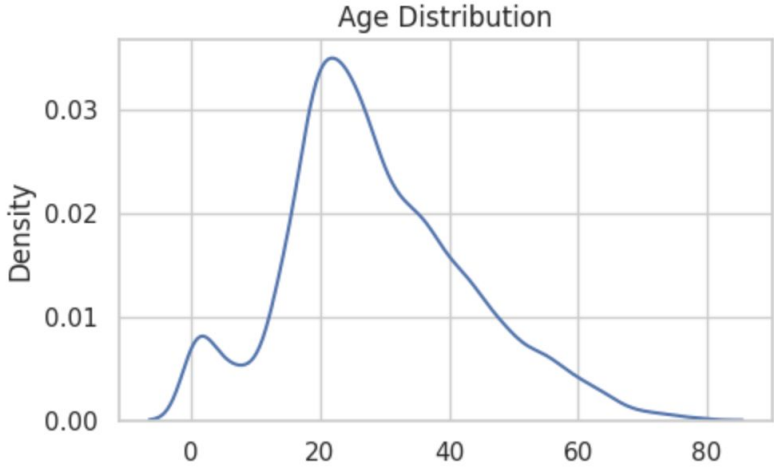
單變數｜雙變數｜
與Transported的關係

單變數EDA-連續型變數

- **連續型變數:**

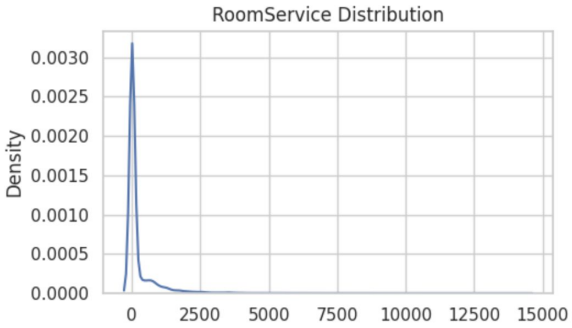
  Age, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck

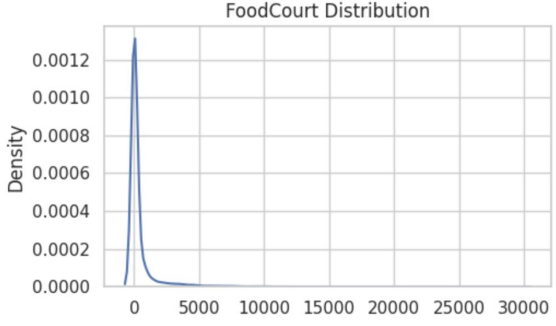


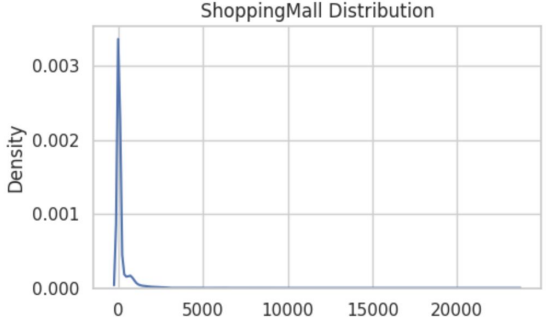


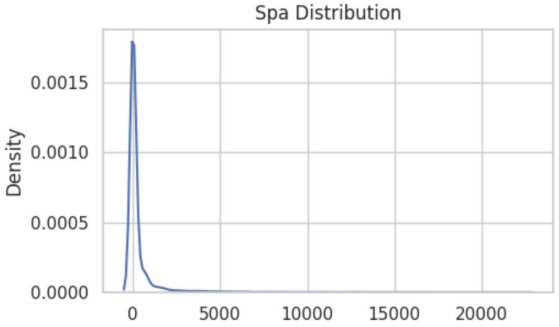
單變數EDA-連續型變數

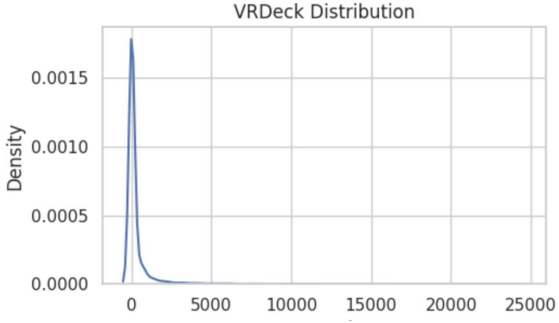
- **觀察到連續型變數皆有很多離群值，且有很嚴重的右偏**





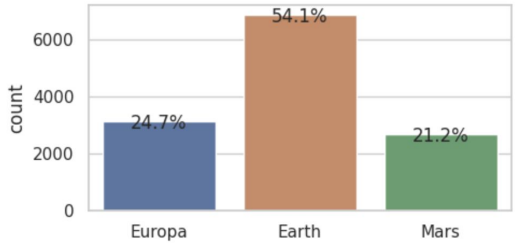




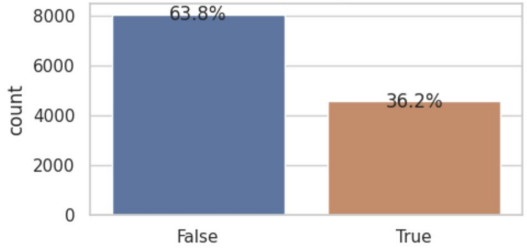


單變數EDA-離散型變數

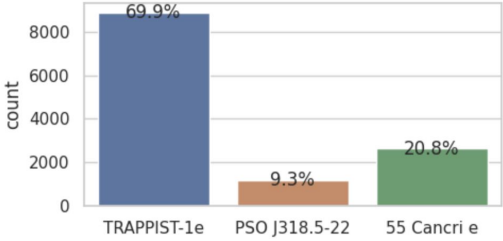
HomePlanet



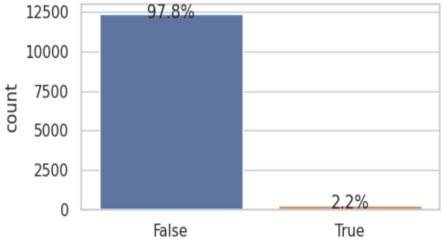
CryoSleep



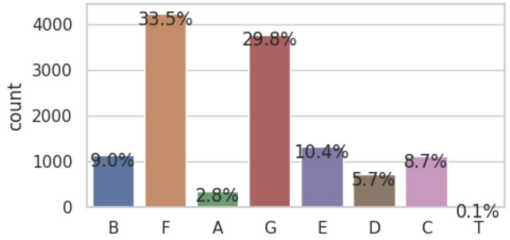
Destination



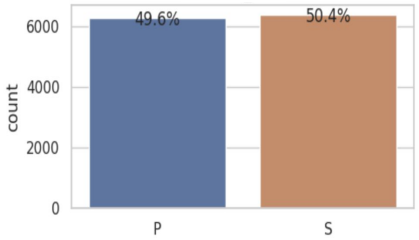
VIP



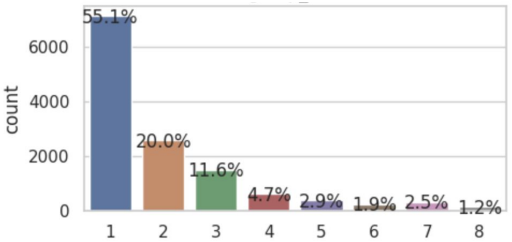
Deck



Side



Group

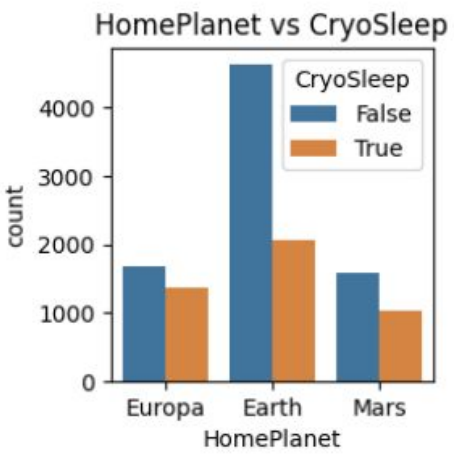


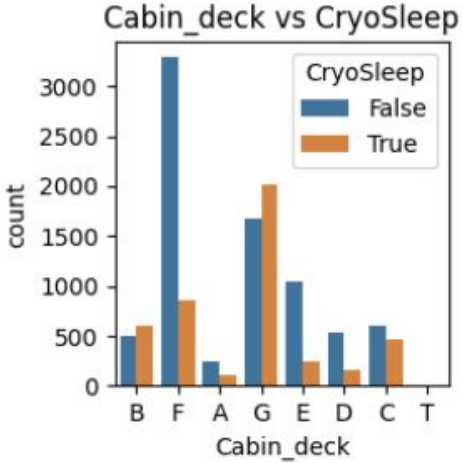
雙變數EDA-連續vs連續

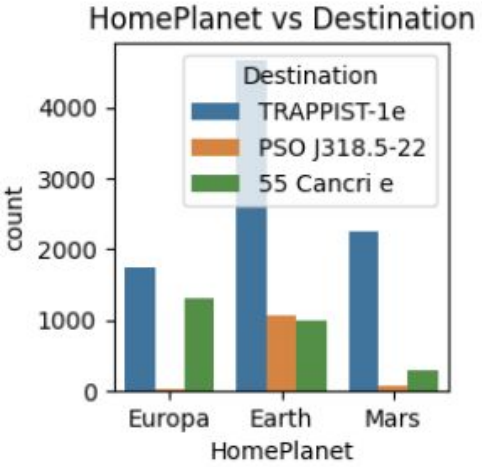
- 計算連續型變數的pearson相關係數矩陣，發現兩兩變數之間的線性關係都較弱，其中相關程度最高的為**FoodCourt & VRDeck**

雙變數EDA-離散vs離散

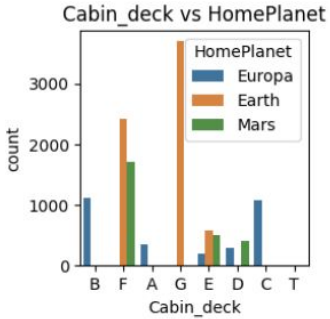
- HomePlanet為Earth的乘客有較高比例不會選擇冷凍睡眠
- Deck為F, E, D的乘客有較高比例不會選擇冷凍睡眠
- Destination為PSO的旅客，有很大的機率HomePlanet為Earth



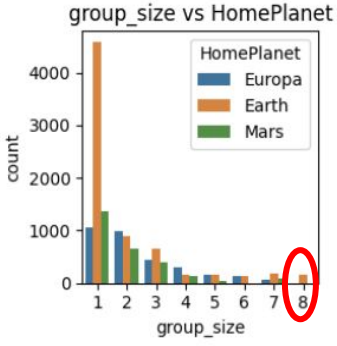




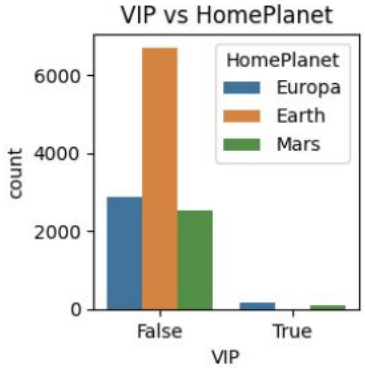
雙變數EDA-離散vs離散



- Deck為A,B,C,T的乘客，HomePlanet皆為Europa
- Deck為G的乘客，HomePlanet皆為Earth



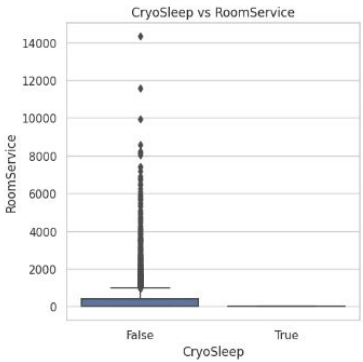
- Group為8的乘客，HomePlanet皆為Earth

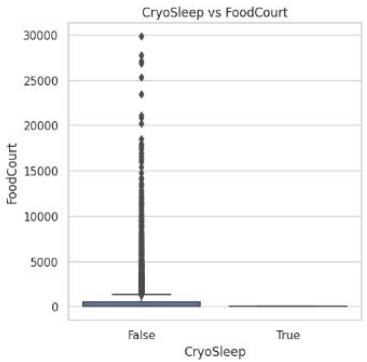


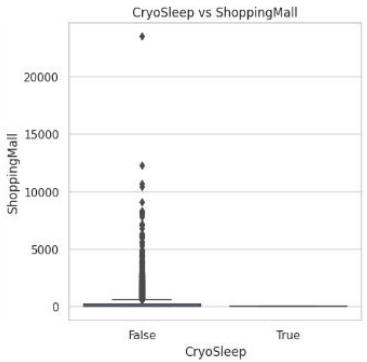
- HomePlanet為Earth的乘客，VIP皆為False

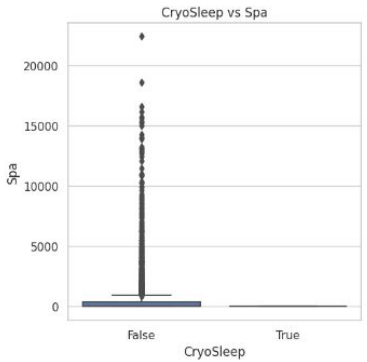
雙變數EDA-離散vs連續

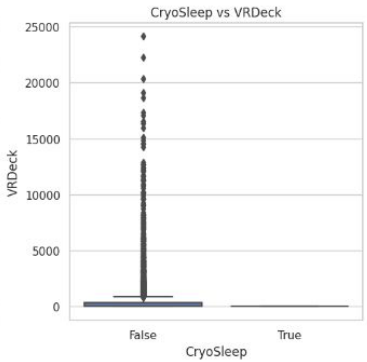
- Cryosleep與5個消費變數的boxplot,
  發現CryoSleep = True的乘客不會有任何消費





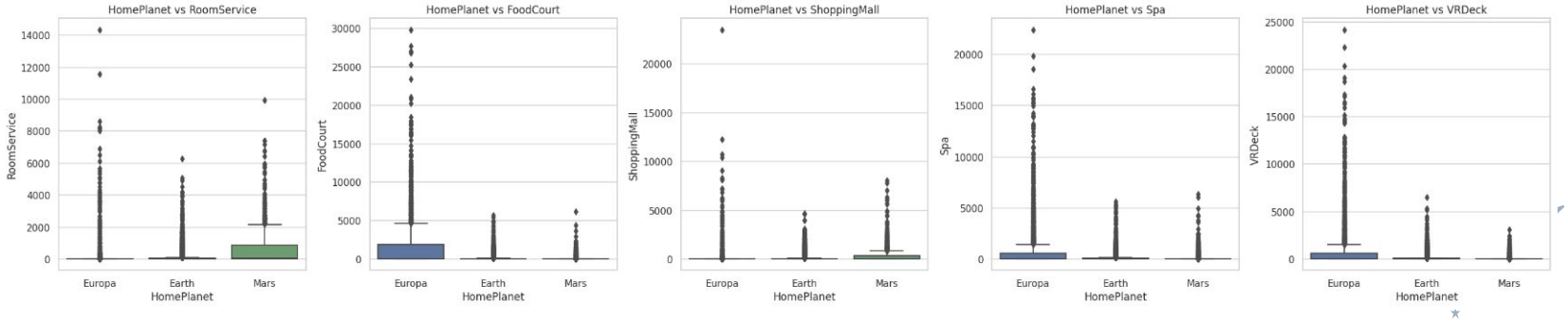




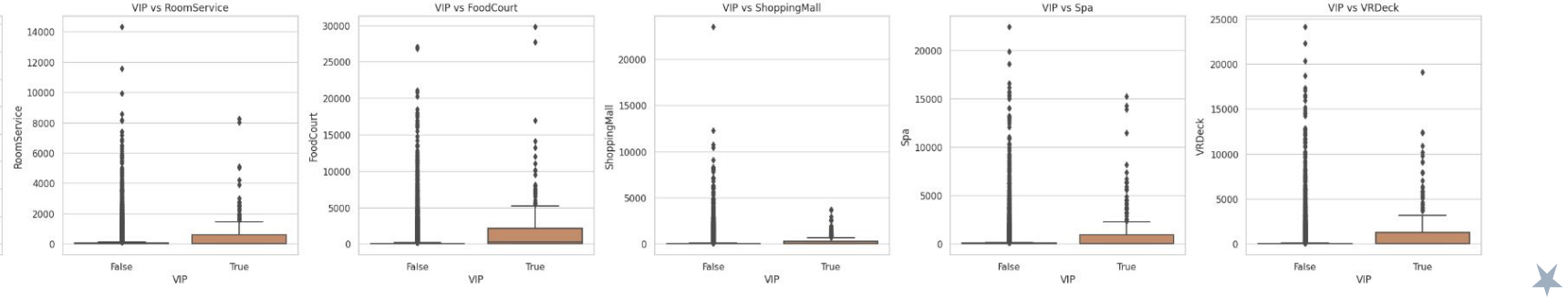


雙變數EDA-離散vs連續

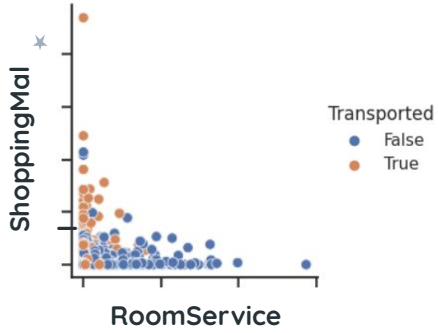
- **HomePlanet = Europa 的乘客消費金額普遍較高**



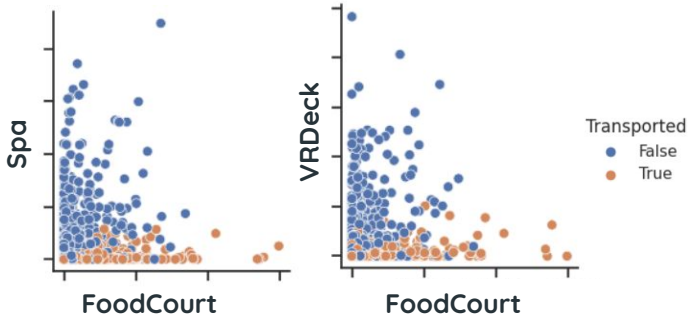
- **VIP = True 的乘客普遍消費金額較高**



雙變數EDA-各變數vs “Transported”

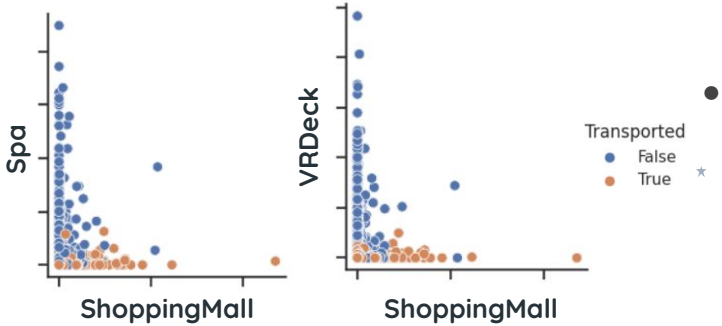


- ShoppingMall金額較高的旅客通常RoomService金額較低，且大多Transprted=True

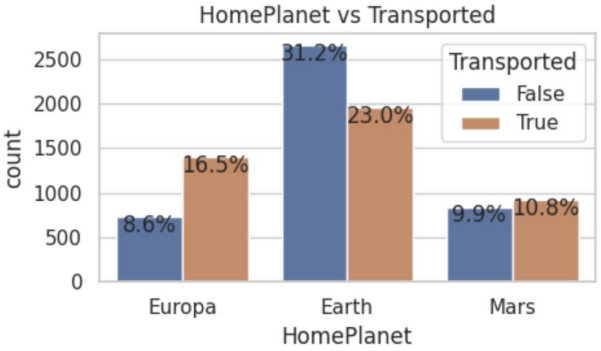


- Spa及VRDeck金額較高的旅客，通常Transported=False
- FoodCourt消費金額若較高，則Spa及VRDeck消費金額會較低，且大多Transprted=True

雙變數EDA-各變數vs “Transported”



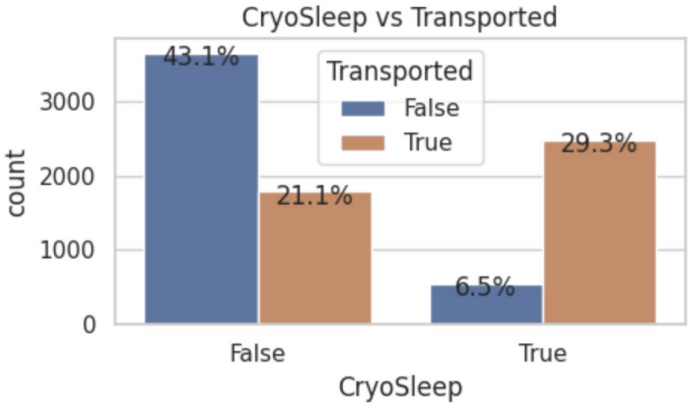
- 若Spa或VRDeck的消費金額較高，則ShoppingMall的消費金額會較低，且Transported=False

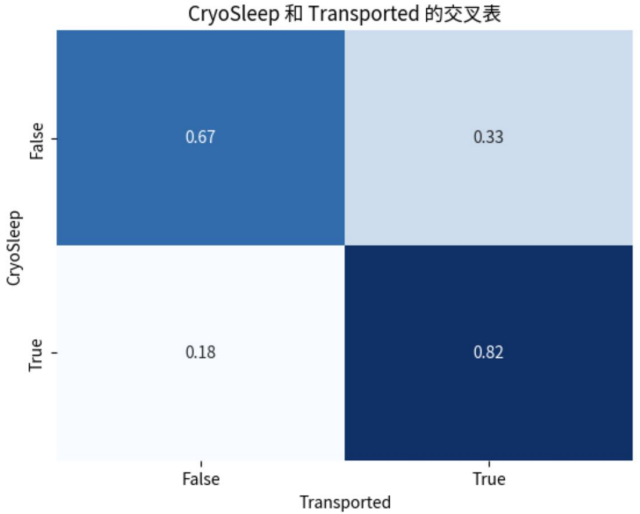


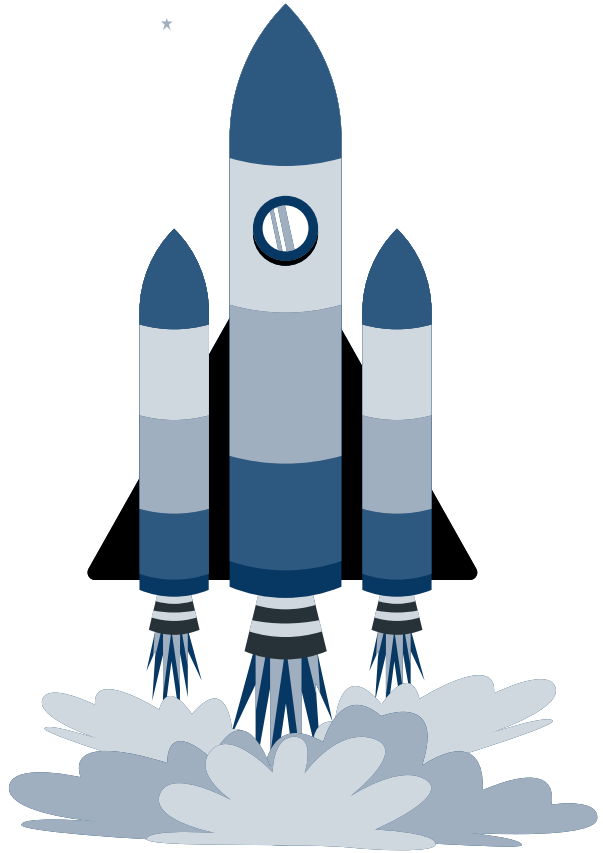
- HomePlanet為Europa及Mars的乘客，被傳送的比例較高；Earth的乘客則是沒有被傳送的比例較高

雙變數EDA-各變數vs “Transported”

- CryoSleep是對Transported較有影響的變數，
  在選擇冷凍睡眠的條件下，乘客被傳送的比例較高(82%)；
  在沒有選擇冷凍睡眠的條件下，乘客沒有被傳送的比例較高(67%)。







03

Imputation

EDA觀察｜平均數眾數

Method 1. 根據EDA觀察

觀察	現象	填補	個數
Deck A, B, C, T	HomePlanet 皆是 Europa	HomePlanet	129
Deck G, Group 8	HomePlanet 皆是 Earth		
HomePlanet Earth, Deck T, Group 8, Age < 18	VIP 皆是 False	VIP	173
CryoSleep True	不會有任何消費	RoomService, FoodCourt, ShoppingMall, Spa, VRDeck	598
有消費, Deck T	CryoSleep 皆是 False	CryoSleep	174

Method 2. 用眾數與平均數補值

- **套件**

  sklearn.impute SimpleImputer

- **眾數 most_frequent**

  針對類別變數

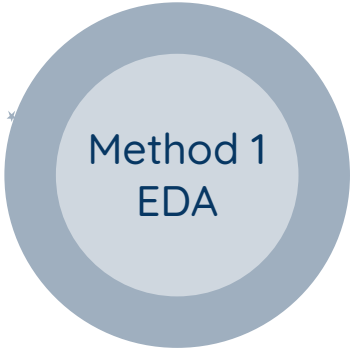
  CryoSleep, Deck, Side, VIP, HomePlanet, Destination, Group

- **平均數 mean**

  針對數值變數

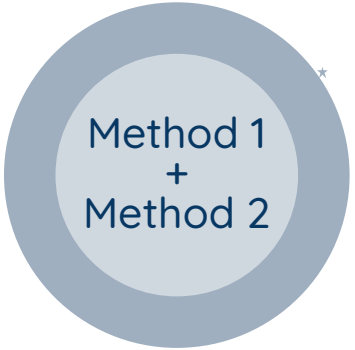
  ShoppingMall, FoodCourt, RoomService, Spa, VRDeck, Age

補値方法比較



0.80687

0.80827



0.80734

04

Feature Engineering

新增變數｜刪減變數

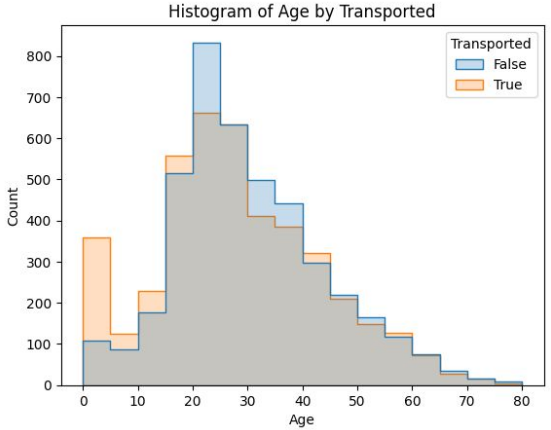
新增變數

- **年齡分組 Age_group**
- **花費總和 Expenses**

  RoomService + FoodCourt + Spa + VRDeck + ShoppingMall
- **One-Hot Encoding**

  Deck, HomePlanet, Destination, Group, Age_group
- **Data Transformation**

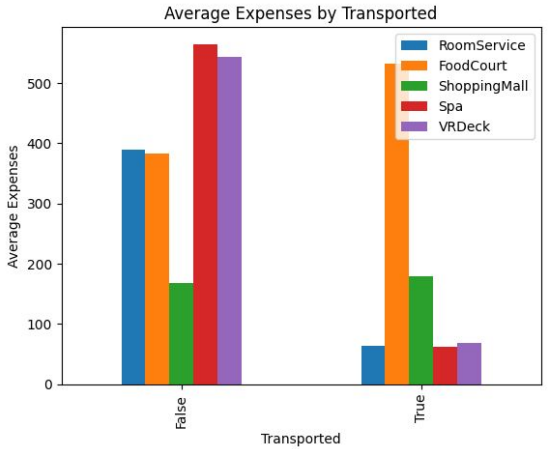
  log, minmax, standardize
- **Clustering**



Age_group	Age
1	0~5
2	6~10
3	11~20
4	21~30
5	31~50
6	51~60
7	61~70
8	>70

刪減變數

- **已建立新變數**: 移除 Age
- **對 Transported 影響不顯著**: 移除 ShoppingMall
- **相關性高的變數**: 移除 Destination_55 Cancri e, FoodCourt, HomePlanet_Earth

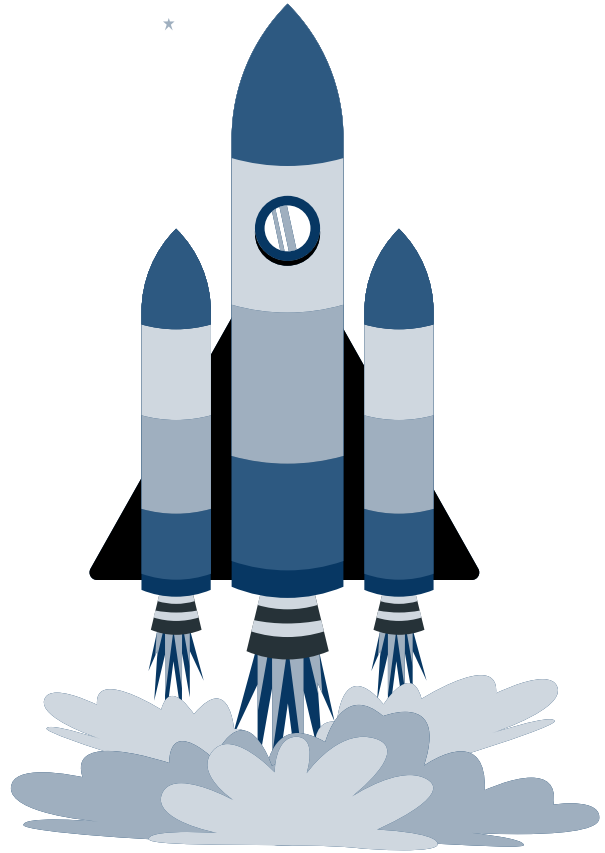


變數	t-test 統計量	p-value	是否顯著
RoomService	-23.4032	< 0.0001	V
FoodCourt	4.1192	< 0.0001	V
ShoppingMall	0.7166	0.4736	
Spa	-21.0460	< 0.0001	V
VRDeck	-19.6559	< 0.0001	V

刪減變數

- **已建立新變數**: 移除 Age
- **對 Transported 影響不顯著**: 移除 ShoppingMall
- **相關性高的變數**: 移除 Destination_55 Cancri e, FoodCourt, HomePlanet_Earth

變數一	變數二	相關係數
Destination_55 Cancri e	Destination_TRAPPIST-1e	0.7831
FoodCourt	Expenses	0.7421
HomePlanet_Earth	HomePlanet_Europa	0.6332
Spa	Expenses	0.5924
...	...	...



05

Model Fitting

Model Selection |
Feature Importance |
Hyperparameter Tuning

Model Selection

- 使用 train_test_split() 以8 : 2的比例分為 :

1) 訓練集: 6954 筆
2) 測試集: 1739 筆

- 評估模型的效能 :

➢ StratifiedKFold 交叉驗證

➢ 最終選擇 XGBoost classifier

	Algorithm	CrossValMeans	CrossValerrors
0	LogisticRegression	0.788569	0.011812
1	SVC	0.788568	0.014188
2	RandomForest	0.787992	0.009447
3	GradientBoosting	0.798118	0.014444
4	KNeighboors	0.760497	0.012738
5	XGBClassifier	0.803525	0.012524

Feature Importance

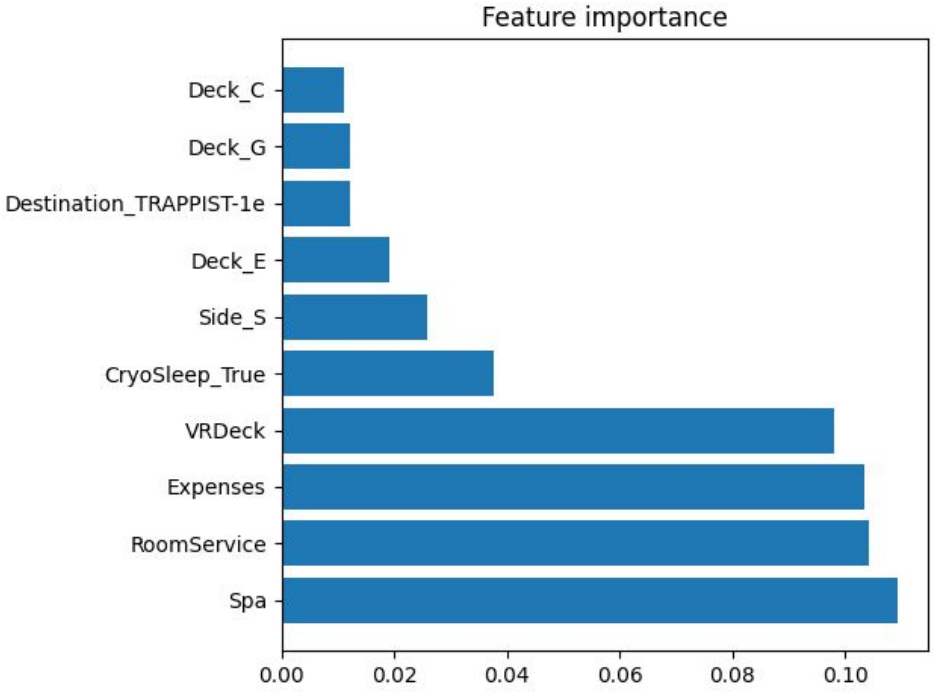
- **Permutation Importance**

➢ **訓練一個模型，可以得到一個基準的評估指標，例如準確率或 R2 等。**

➢ **隨機打散資料集的特徵。**

➢ **計算每個特徵重要性：**

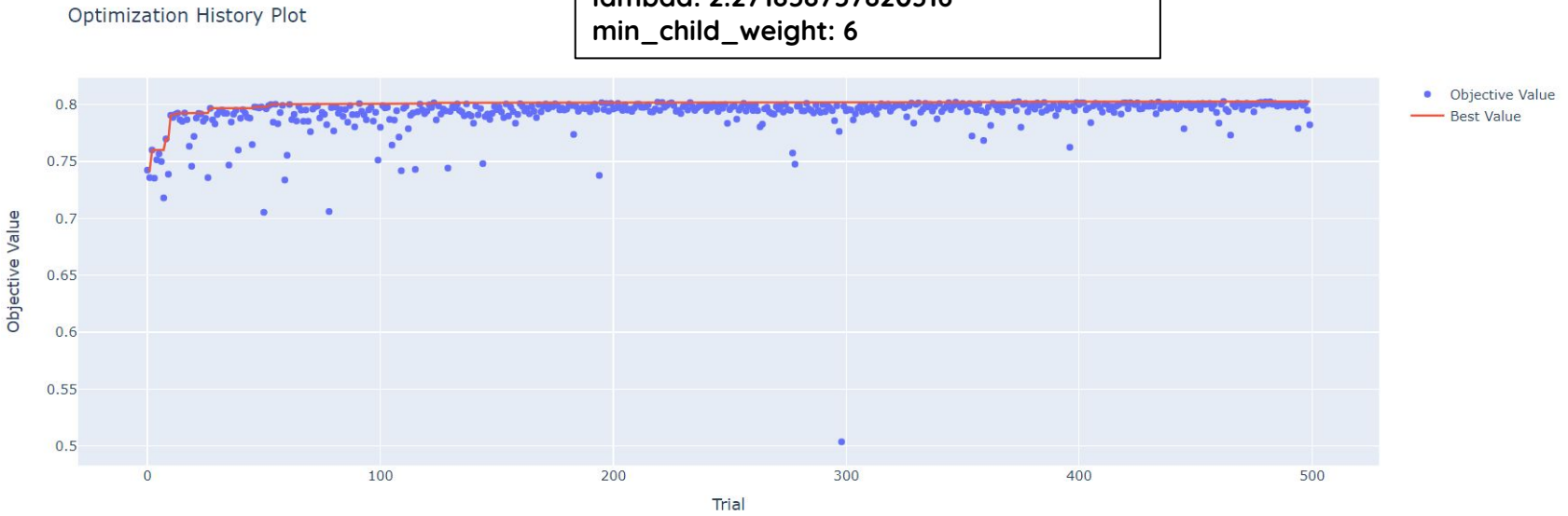
打散資料集的 error - 原始資料集的 error



Hyperparameter Tuning

- **OPTUNA**
- **Best hyperparameters :**

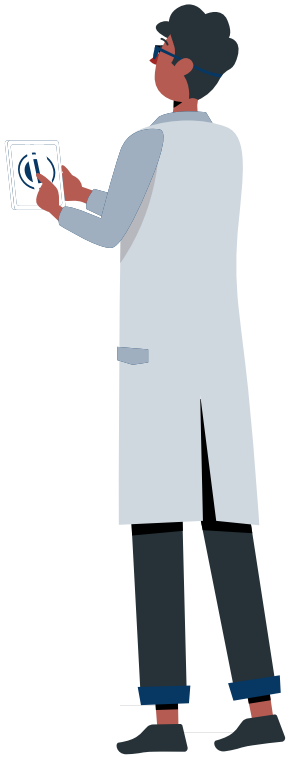
n_estimators: 918
max_depth: 9
learning_rate: 0.0827280846016892
subsample: 0.978321164193843
colsample_bytree: 0.5330631152248969
alpha: 4.370034992967263
lambda: 2.271858757820316
min_child_weight: 6



06

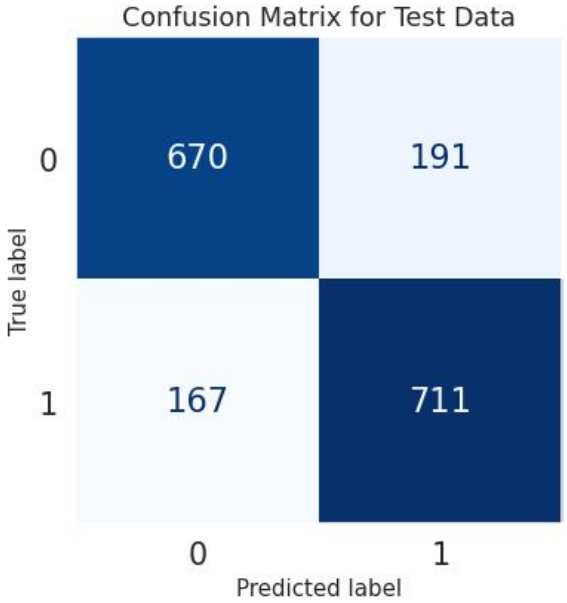
Result

Result | Leaderboard



Result

- **Test Dataset**

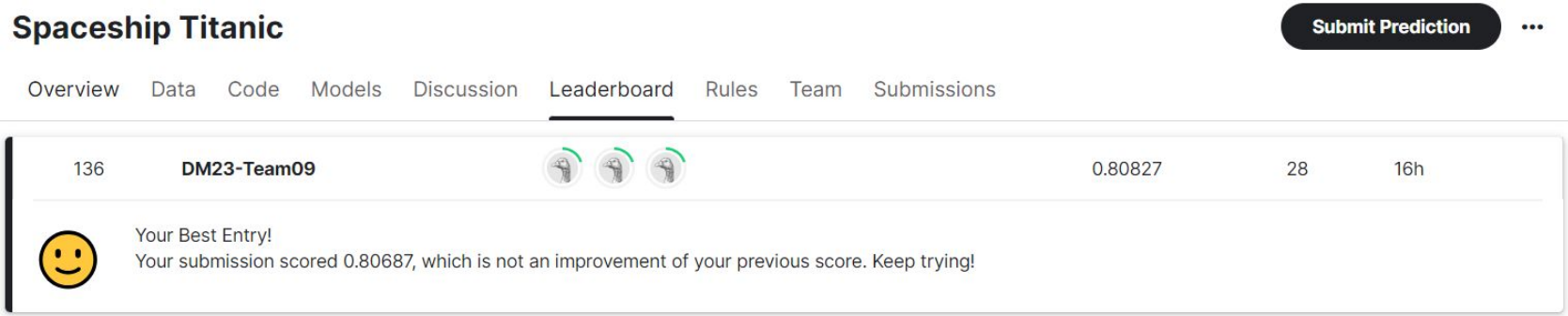


XGBoost Performance Summary on Test Data

	XGBoost
Accuracy	79.41%
Macro Precision	79.44%
Macro Recall	79.4%
Macro F1-score	79.4%
Macro AUC	88.25%

Leaderboard

- **Rank : 136**
- **Score : 0.80827**



THANKS