

## Quiz 11

### 題目說明：

下載 Quiz11.ipynb 檔案，將未完成的部分補上

**DataFrame** 請照規定命名，僅需在標示的位置上填空，其餘程式碼無需更動。

### 第一部分：

1. 讀入 'titanic.csv' 成 DataFrame 格式並命名為 df
2. 並把 df 隨機分為 8:2 的兩個 DataFrame，並命名為 train\_df 和 test\_df，train\_df 為 8，test\_df 為 2。( Hint: 此部份可使用 numpy )

#### 輸出範例：

輸出為資料數，df會與範例一樣。train\_df 和 test\_df，因是隨機切分不一定會與範例一模一樣，故只要比例接近 8:2 即可。

```
df: 891
train_df: 711
test_df: 180
```

### 第二部分：

將 def PreprocessData() 中的空白處補上：

1. 刪除名為 'PassengerId', 'Name', 'Ticket', 'Cabin' 的四個 column
2. 'Age', 'Fare' 的空值用其平均值補上
3. 將 'Sex' column 中的 female 轉為 0，male 轉為 1
4. 再將 'Embarked' column 改為 One Hot Encoding 格式，並存成名為 x\_OneHot\_df 的 DataFrame

#### 輸出範例：

輸出為x\_OneHot\_df的資訊，ndarray會是(資料數,10)，Features會是(資料數,9)，Label會是(資料數,)，One Hot Encoding若處理正確，Embarked會被取代成 Embarked\_C, Embarked\_Q, Embarked\_S

```
train_df
ndarray (711, 10)
Features (711, 9)
Label (711,)
  Survived  Pclass  Sex   Age  SibSp  Parch   Fare  Embarked_C  Embarked_Q  \
0         0      3    1  22.0     1     0   7.250         False         False  \
2         1      3    0  26.0     0     0   7.925         False         False

  Embarked_S
0         True
2         True

-----

test_df
ndarray (180, 10)
Features (180, 9)
Label (180,)
  Survived  Pclass  Sex   Age  SibSp  Parch   Fare  Embarked_C  Embarked_Q  \
1         1      1    0  38.0     1     0  71.2833         True         False  \
6         0      1    1  54.0     0     0  51.8625         False         False

  Embarked_S
1         False
6         True

-----
```