2021 電子商務技術作業七

資料集 Churn_Modelling.csv 是一份銀行客戶的資料，請對此資料集進行前處理與分析各欄位，並預測潛在的流失客戶

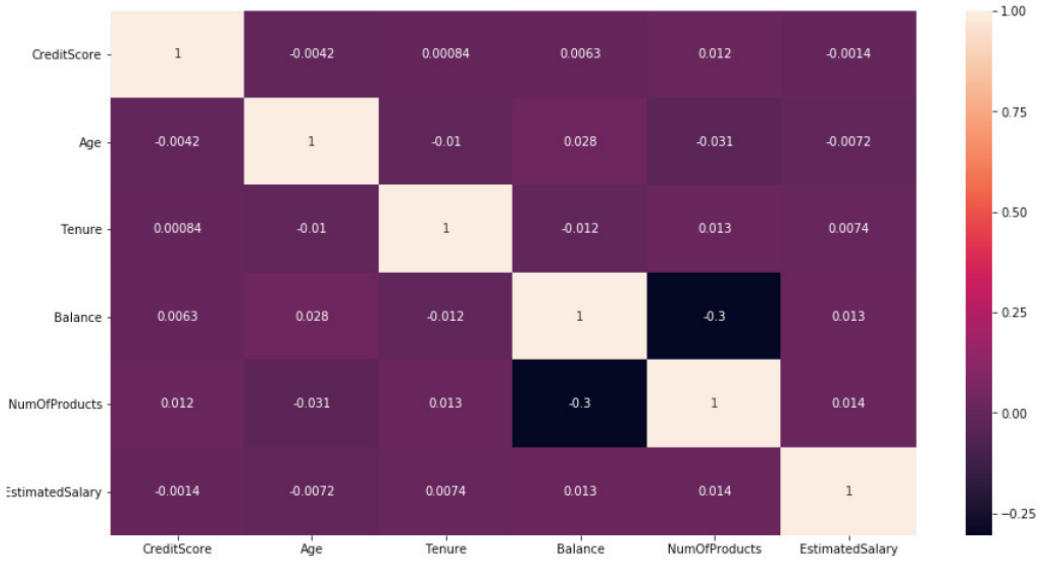
1. 載入 Churn_Modelling.csv 資料集，並印出哪些欄位含有遺漏值(missing value)。 (5%)
2. 以平均值填入 EstimatedSalary 的遺漏值，以眾數填入 Age 與 Gender 的遺漏值。(10%)
3. 修改欄位名稱，將 CredRate 改成 CreditScore、ActMem 改成 IsActiveMember、Prod Number 改成 NumOfProducts、Exited 改成 Churn，以利後續分析資料。 (5%)
4. 去除 CustomerId 欄位，並將 Geography、Gender、HasCrCard、Churn、IsActiveMember 修改資料型態為 category，印出所有欄位的資料型態，並存成新的 CSV 檔 (設定 index=False)。(5%)
5. 對各個欄位進行分析，了解目前銀行客戶的概況：
   (1) 對 HasCrCard 欄位進行分析，說明有多少比例的人持有信用卡，多少比例的人不持有信用卡。(3%)
   (2) 對 Churn 欄位進行分析，說明有多少比例的客戶流失。(3%)
   (3) 對 IsActiveMember 欄位進行分析，說明有多少比例的客戶仍是活躍狀態。(3%)

(4) 對 Churn 進行分析，觀察流失客戶跟未流失客戶的資料平均值。

(Hint: pandas.DataFrame.groupby()) (6%)

(5) 計算屬性間的相關係數，並用 seaborn 繪製出熱力圖 (heatmap)，

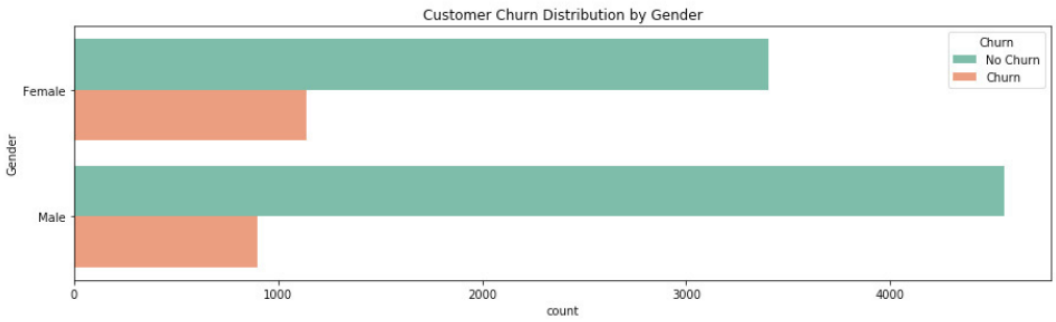
如下圖所示。(Hint: pandas.DataFrame.corr()) (8%)



6. 運用資料視覺化來幫助分析：

(1) 繪出 Gender 與 Churn 的數量關係，分析不同性別於客戶流失的關

係，如下圖所示。 (Hint: seaborn.countplot()) (10%)



(2) 繪出 Geography 與 Churn 的數量關係，分析不同地區於客戶流失

的關係。 (Hint: seaborn.countplot()) (5%)

(3) 繪出 Age 分布與 Churn 的關係，分析不同年齡於客戶流失率的關係，如下圖所示。 (Hint: seaborn.kdeplot()) (10%)

Figure 7.38: Distribution of customer age (churn versus no churn)

(4) 繪出 CreditScore 與 Churn 的關係，分析客戶信用分數於客戶流失率的關係。(Hint: seaborn.kdeplot()) (7%)

7. 載入第四題儲存的新的 CSV 檔，使用 Weka 回答以下題目，截圖並說明操作過程：

   (1) 將 HasCrCard, IsActiveMember, Churn 轉成 Nominal 屬性。 (10%)

   (2) 使用 Attribute Selection，以 CfsSubsetEval 及 BestFirst 來篩選屬性，並說明屬性篩選結果。 (10%)

- **繳交期限：5/19(三) 中午12:00**

- **請繳交 .PDF檔與 .ipynb檔，檔名為 ECT_HW7_學號.pdf/ECT_HW7_學號.ipynb，文件與程式中請適當附上說明文字。**

- **上傳至ee-class作業區，遲交一天扣該次作業5%**