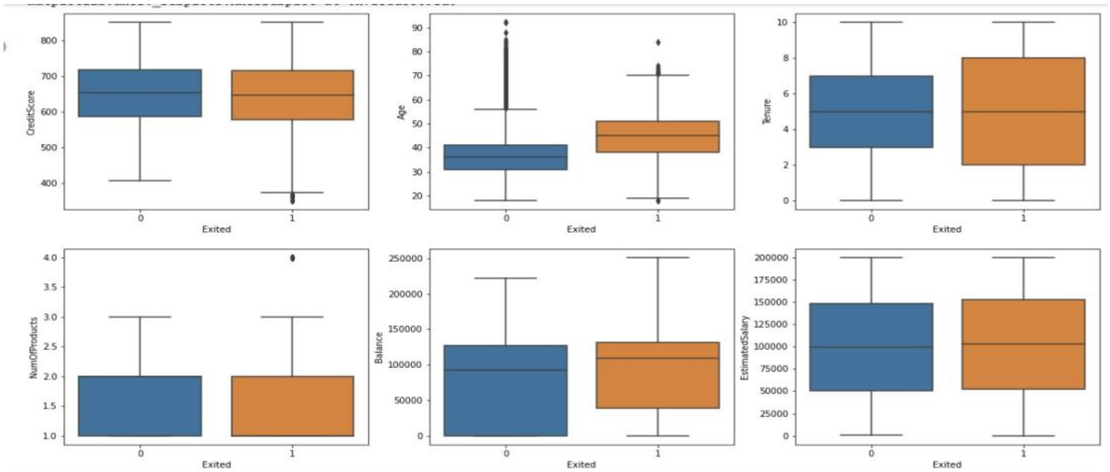
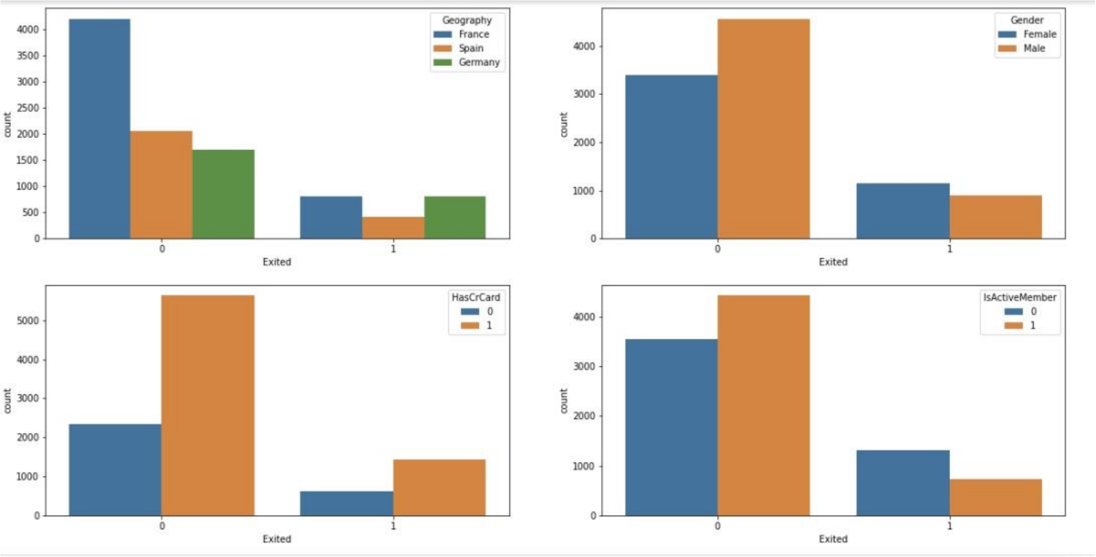
Week 2 – Ingest and Explore the Dataset

The project is banking customer churn prediction. It is a binary classification problem. The dataset is from Kaggle. The target variable was Exited. Exited could be either 1 or 0. When Exited was 1, it represented customers change the bank. When exited was 0, it meant customers stay at the same bank. The predictors could be age, gender, geography, estimated salary, number of products customers own, whether customers have credit card and etc.

The dataset consisted of 10,000 rows and 14 columns. There was no missing value for this dataset. I would drop three unnecessary attributes in the future when building models, which were RowNumber, CustomerId and surname. The rest 10 columns (except target variable) were CreditScore, Gengraphy, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, and EstimatedSalary. Among these 10 attributes, there were 2 categorical variables and the rest were numerical variables. I would convert 2 categorical features geography and gender into numerical variables before building models.

After understanding the dataset, I used matplotlib to do visualization in python. I did boxplots for both numerical and categorical features. From the graph, I found out that for the age attribute, there were lots of outliers for the Exit 0 group. This meant elder people are more willing to stay at the same bank. Also, I found out that Germany were more likely to churn compared to France and Spain.




Geography
🟦 France
🟧 Spain
🟩 Germany
count
4000
3500
3000
2500
2000
1500
1000
500
0
0
1
Exited
Gender
🟦 Female
🟧 Male
count
4000
3000
2000
1000
0
0
1
Exited
HasCrCard
🟦 0
🟧 1
count
5000
4000
3000
2000
1000
0
0
1
Exited
IsActiveMember
🟦 0
🟧 1
count
4000
3000
2000
1000
0
0
1
Exited