
The Dataset is about bank customers churning and can be found on Kaggle:

https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling

Disclaimer: The dataset above is simulated.

Introduction

Banks need to have satisfied customers for it to be successful and this report looks at bank customers and our ability to correctly predict if a bank customer will exit the bank or not. The bank is concerned about potentially losing current customers. They want to be able to predict which customers are at risk of churning and be able to do outreach (in the form of promotions, etc) to convince them to retain their services with the bank.

Data Wrangling

The dataset for the bank customers was obtained from Kaggle and was in a csv format. Before starting any analysis, the shape of the dataset needs to be known which in this case was 10000 rows and 14 columns. Out of the 14 columns in the dataset, 13 are feature columns and the 14th column, **'Exited',** is the response column. The dataset also needed to be checked for null values and fortunately, our dataset did not contain any null values as it was obtained from Kaggle and was already clean. Though, this is not often the case with real world data.

The redundant columns, like **'RowNumber'**, were dropped from the dataframe as they do not provide any useful information and the new shape of the dataset became 10000 rows and 13 columns. Another important step that needs to be carried out is converting the categorical columns into numerical values. For this dataset, there were two such columns **'Geography'** and **'Gender'** that were converted. The 'Geography' column contained values France, Germany and Spain which were converted to 0, 1 and 2 respectively. Similarly, the Gender column contained values Male and Female that were converted to 0 and 1 respectively.

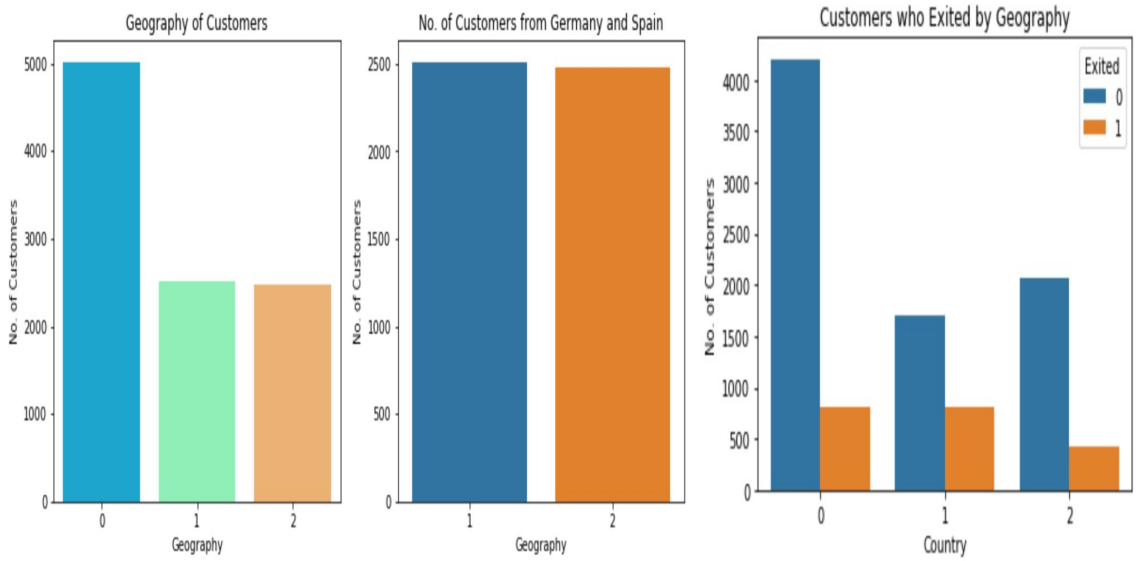
The column 'Exited' was moved to the leftmost side of the table as this allows viewing the data quickly and also makes splitting the dataset into train/test sets easier at a later stage. The data was checked for outliers by looking for any extreme values in the min and max fields of the columns of the dataframe. Fortunately, the data does not contain any outliers.

After performing data wrangling on the dataset, the next step was to investigate, by visualizing the data, to uncover interesting information about the data and learn more about the customers in the bank who are exiting or not. The 'Geography' and 'Gender' categorical columns had been previously converted into numerical values using the following relationship:

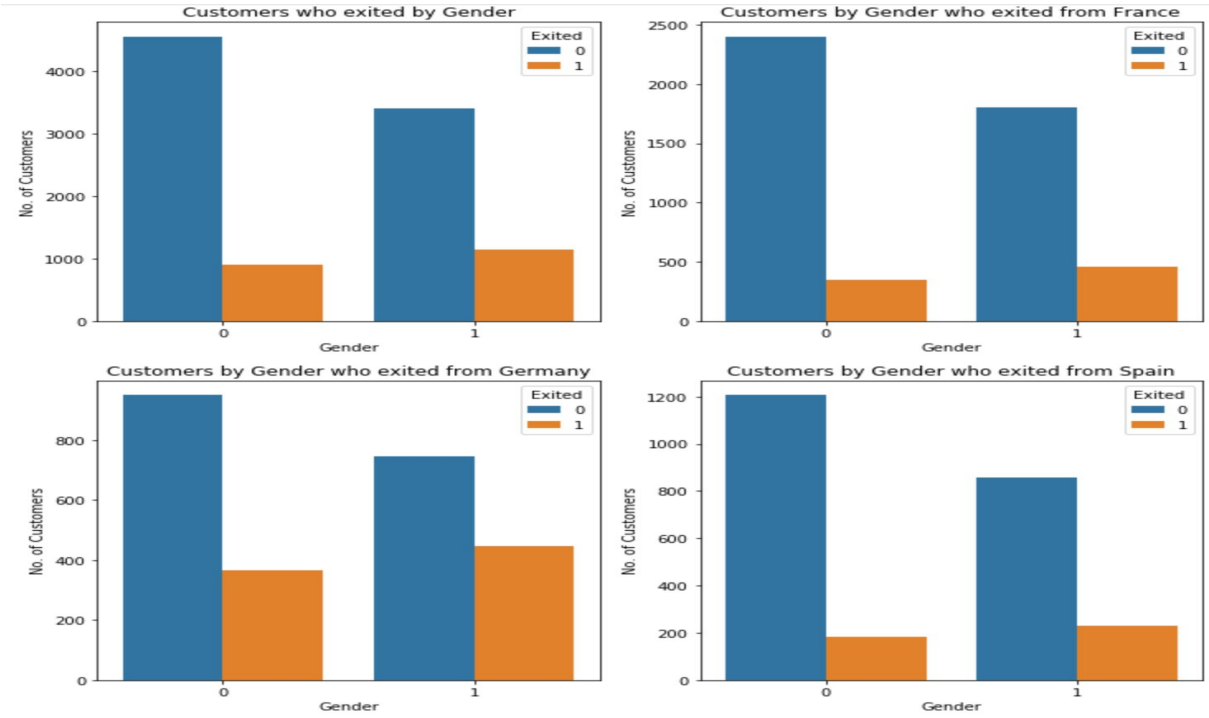
- For the ‘Geography’ column, the values of **France**, **Germany** and **Spain** are represented by **0**, **1** and **2** respectively.
- For the ‘Gender’ column, the values of **Male** and **Female** are represented by **0** and **1** respectively.

Data Visualization

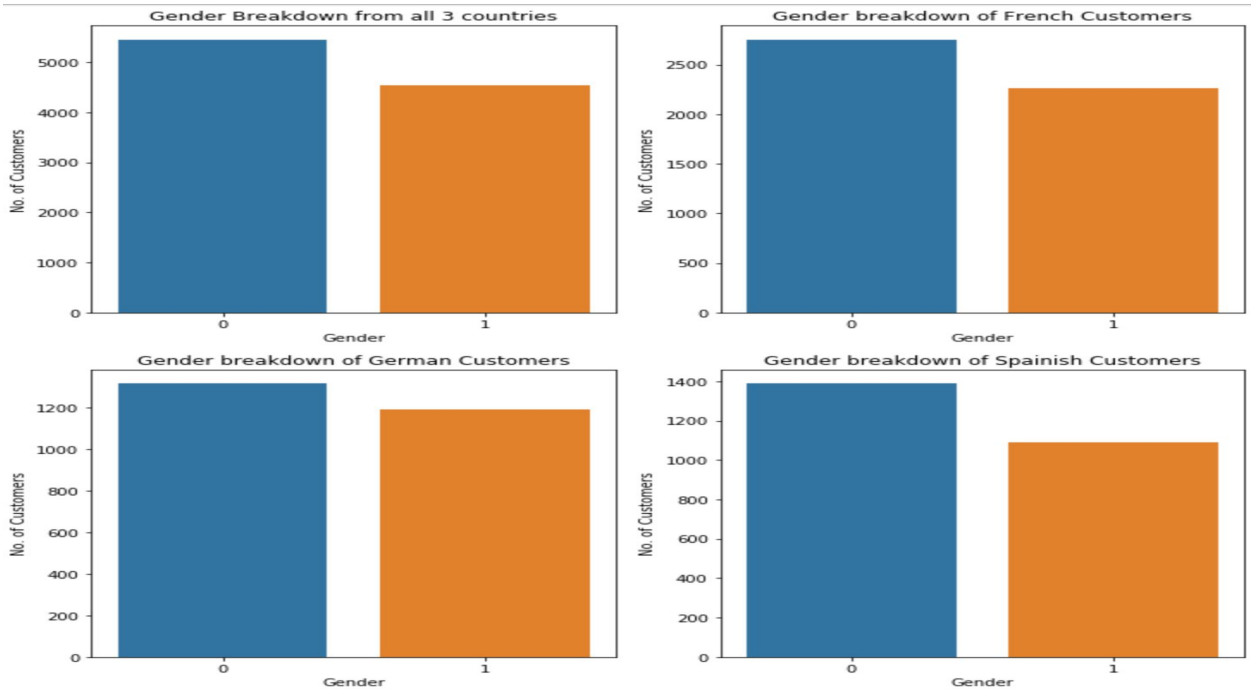
After visualizing where the customers are from, it was discovered that there were twice as many French Customers than from Germany and Spain. Germany also had the highest number of customers who exited the bank followed by France and Spain. These inferences can be seen in the 3 plots below.



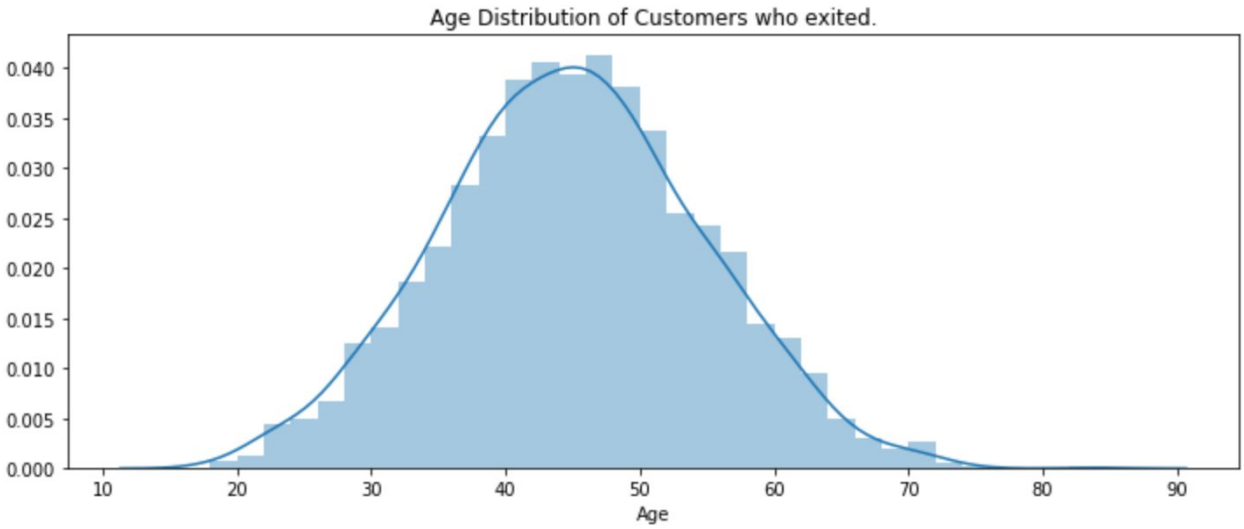
Further visualization of the data showed that female customers left the bank more than male customers, France had the highest number of female customers who exited the bank and Germany had the highest number of customers who left the bank as well as the highest number of males who exited the bank.

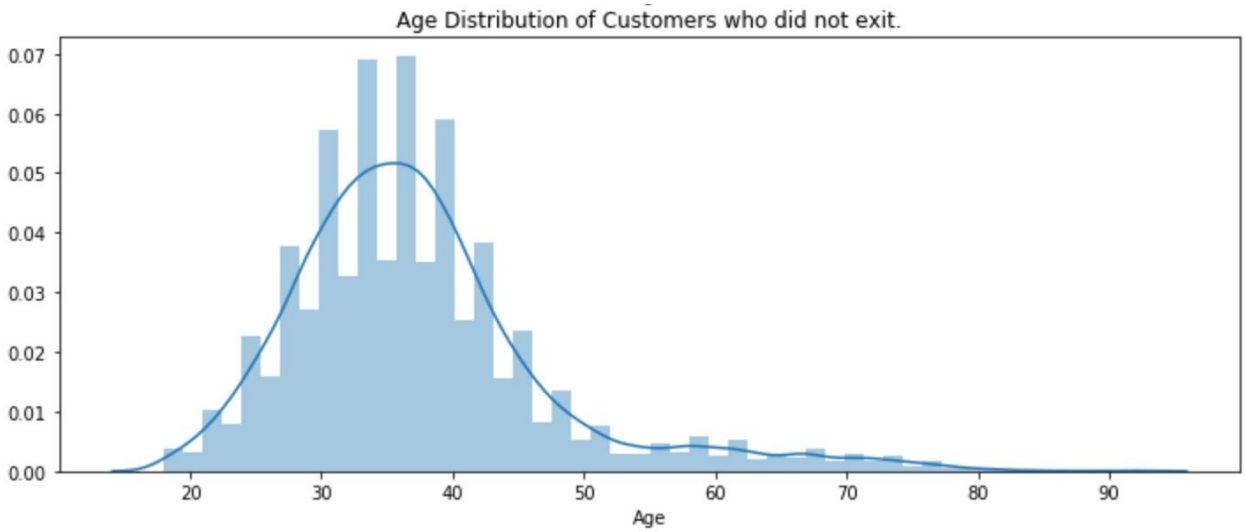


When the gender breakdown of the bank customers was visualized, it showed that France had the highest number of female and male bank customers.

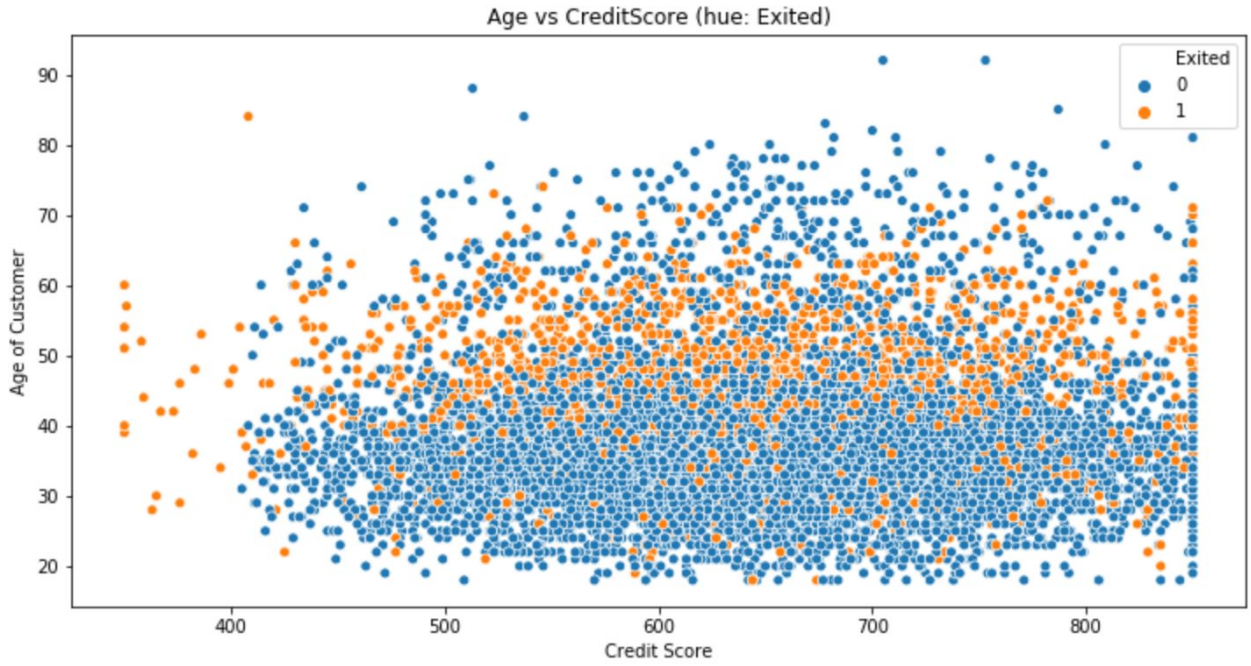


The age distribution of the customers who exited the bank looks similar to a normally distributed distribution.

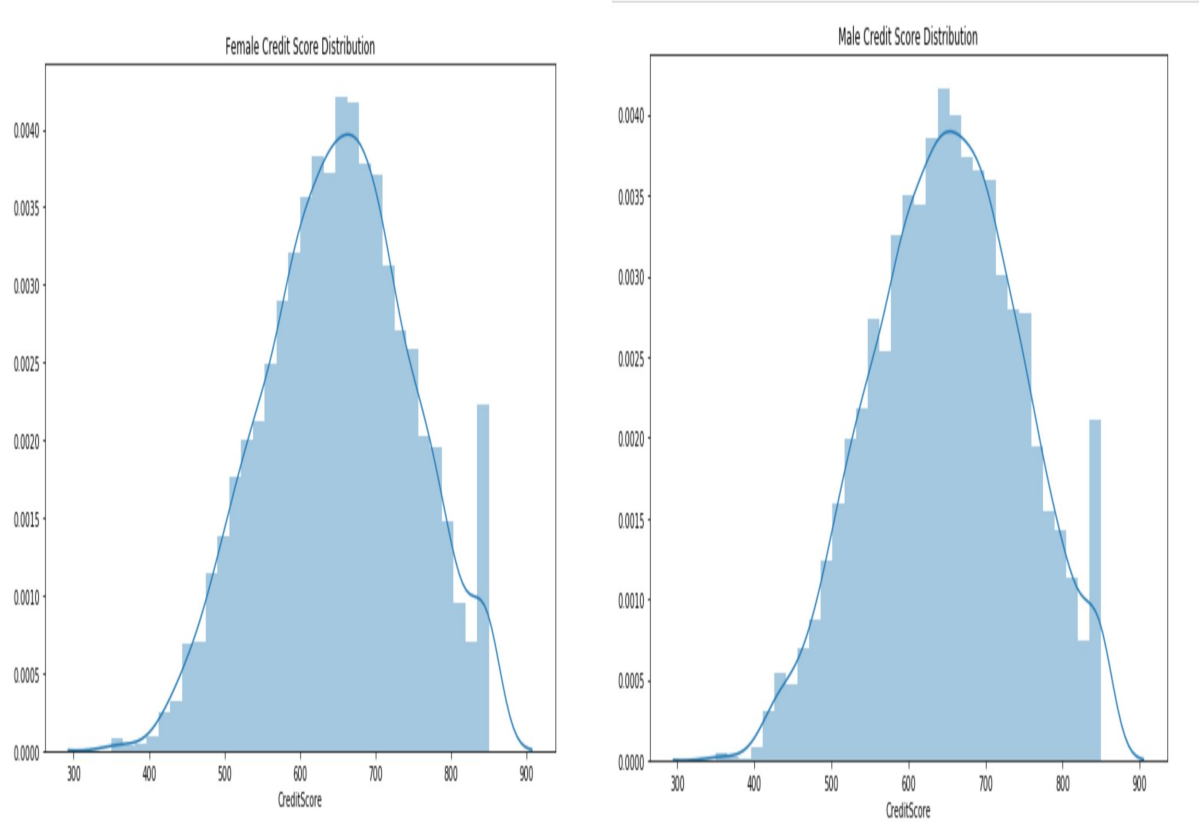




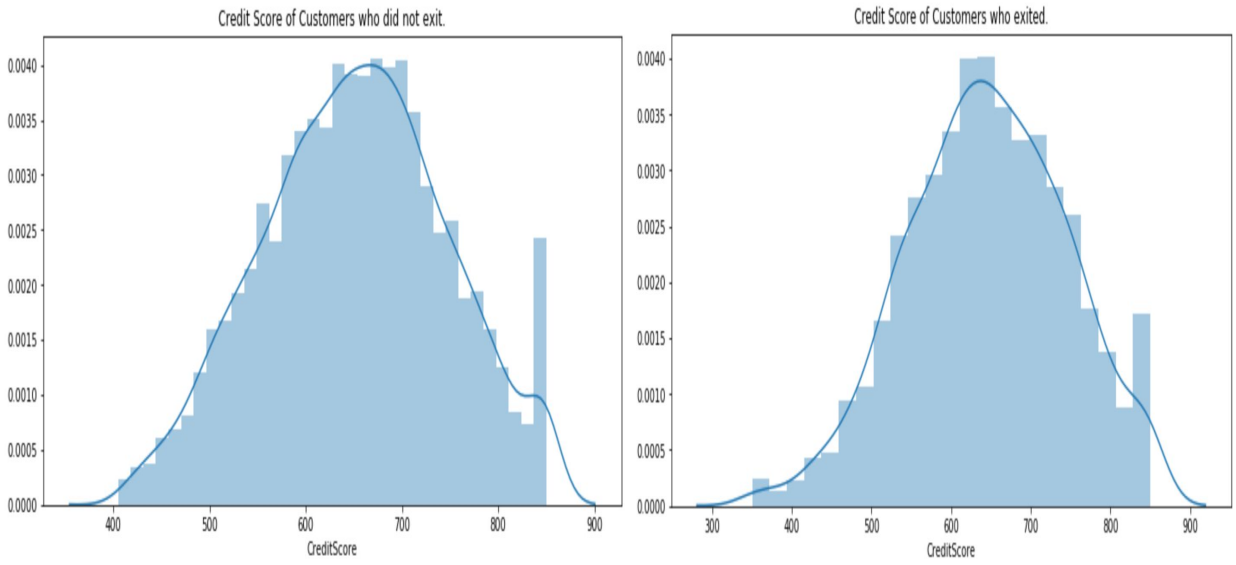
A scatterplot between credit score and age revealed that the majority of the customers who exited were between 40 years and 60 years of age. It would be interesting to see the reason why this particular group of people aged between 40 and 60 years old are exiting the bank? Are they getting better offers from other banks?

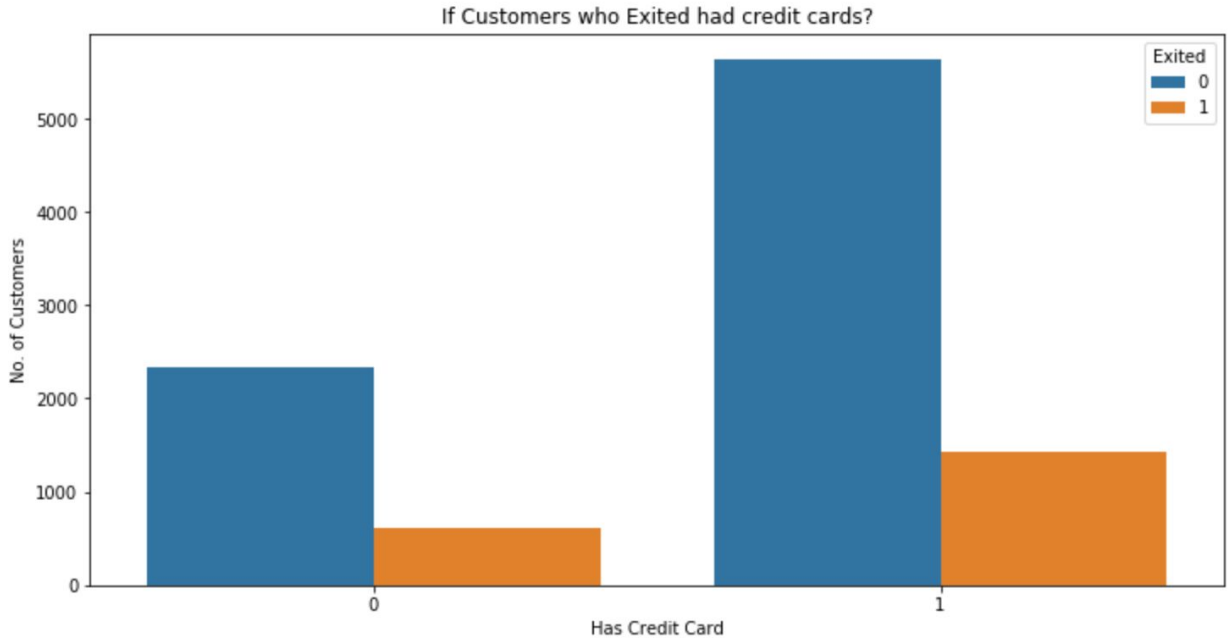


Visualizing the male and female bank customers credit score distributions ,showed no major significant difference in the distribution of both distributions.



It was interesting to find that customers who exited had a roughly lower credit score and customers who had a credit card exited more than customers who didn’t own a credit card.





These are interesting insights about the bank customers and allows us to further ask questions: if the customers could get better offers based on their low credit score from other banks that is causing them to exit? Are other banks offering lower APR on their credit cards to lure customers?

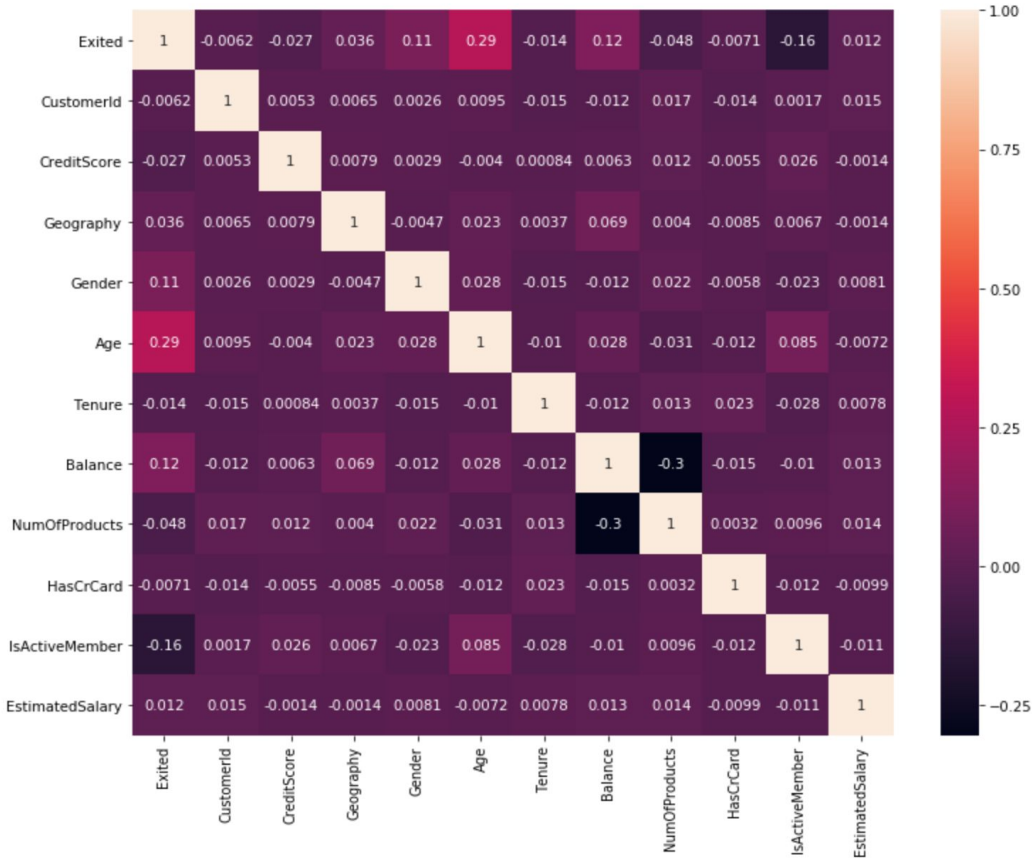
Statistical Analysis

After gaining further information about the dataset by performing exploratory data analysis, the next step is to investigate the dataset through a statistical lens. The main aim is to look for any significant correlations between variables and performing a test to see if there is any statistical

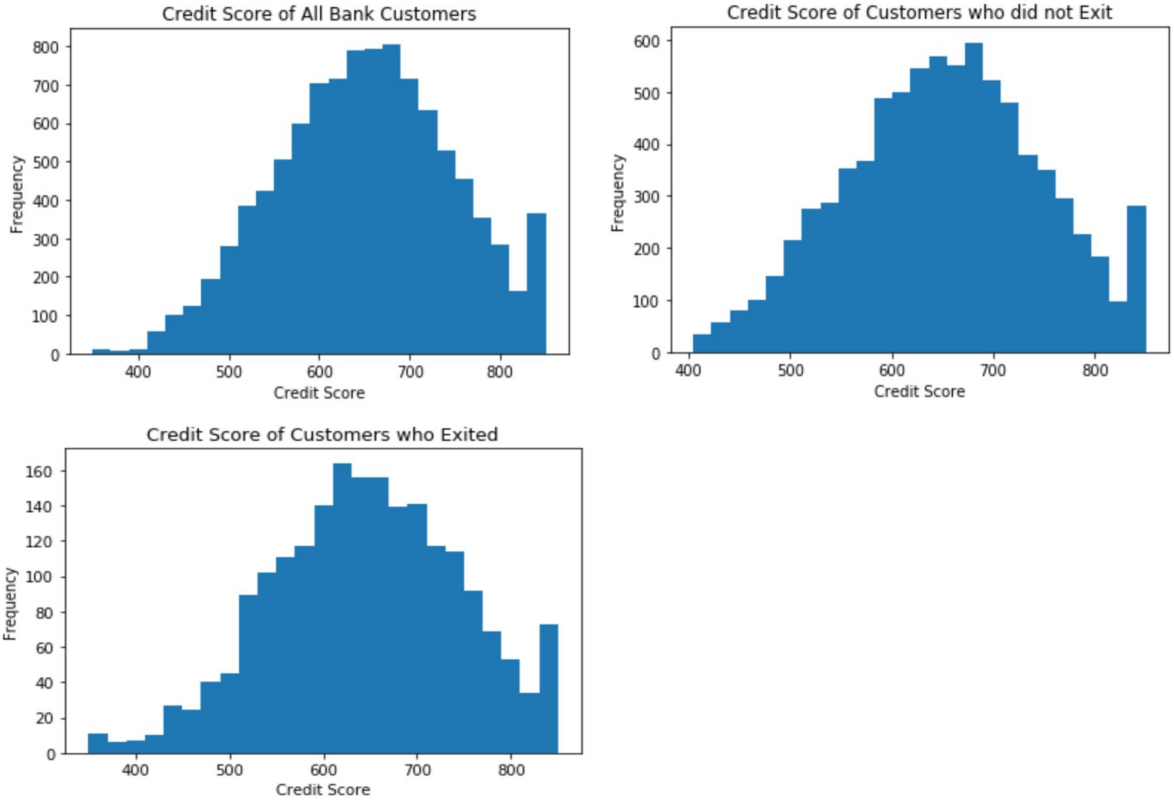
difference in the means of credit score of customers who exited compared to the mean credit score of the entire bank customer population.

The summary statistics of the dataset provide us with some initial information about customers who had stayed with the bank and customers who had exited the bank. They showed that **79.6%** of customers stayed at the bank while **20.4%** of customers exited the bank. The mean credit score of customers who stayed was 651.85 compared to 645.35 for customers who exited the bank.

It is important to check if the variables in the dataset have any meaningful correlations between them. A correlation value close to -1 indicates strong negative correlation, values close to 0 indicate no/very weak correlation and values close 1 indicate a strong positive correlation between variables. Looking at the figure of the heatmap below, the **'Exited'** and **'Age**' variables had the highest correlation value of 0.29, which does not suggest strong correlation between the pairs as the value is close to 0. All the other variables had very weak correlation values as can be seen below in the heatmap.



The figures below show how the credit score distributions looked like for customers who had exited, who had stayed and the entire bank customer population. The credit score distribution of customers who exited appears different than the other two distributions.



Next, a one-sample t-test was performed to test whether a population mean is significantly different from some hypothesized value. In this case, the test was performed to check whether the average credit score of customers that exited differs from the average credit score of the entire bank population. For this test, the Null and Alternate hypothesis were:

Null Hypothesis: The null hypothesis would be that there is no difference in credit score between customers who exited and all bank customers.

Alternate Hypothesis: The alternative hypothesis would be that there is a difference in credit scores between customers who exited and all bank customers.

The one sample t-test was performed at a 95% confidence interval to check if we could correctly reject the null hypothesis based on our result of the test.

The t-distribution left quartile range was -1.96112925575354 and right quartile range was 1.9611292557535396. The test result showed the test statistic 't' is equal to -2.329. T is simply the calculated difference represented in units of standard error and tells us how much the sample mean deviates from the null hypothesis. The null hypothesis can be rejected if the

t-statistic lies outside the quantiles of the t-distribution corresponding to the chosen confidence level and degrees of freedom.

A p-value of 0.019946347165310532 meant that they would expect to see data as extreme as their sample due to chance way less than 5% of the time if the null hypothesis was true. In this case, the p-value is lower than our significance level α (equal to 1-conf.level or 0.05) so the null hypothesis should be rejected.

Based on the statistical analysis of a one sample t-test, there seems to be some significant difference between the mean credit score of bank customers who exited and the entire bank customer population. The low P-value of 0.019946347165310532 at a 5% confidence interval is a good indicator to reject the null hypothesis.

The results of our test showed that there is some significant difference between the mean credit score of bank customers who exited and the entire bank customer population. Further experiments can be conducted or more data can be collected about the bank customers to develop our understanding of the data and potentially get more refined insights.