Customer Churn Predictions

Gursimran Singh

2019

Introduction

Customer churn, also known as customer attrition, occurs when customers stop doing business with a company. The companies are interested in identifying segments of these customers because the price for acquiring a new customer is usually higher than retaining the old one. For example, if Netflix knew a segment of customers who were at risk of churning they could proactively engage them with special offers instead of simply losing them.

Simply put, customer churn occurs when customers or subscribers stop doing business with a company or service. Also known as customer attrition, customer churn is a critical metric because it is much less expensive to retain existing customers than it is to acquire new customers – earning business from new customer's means working leads all the way through the sales funnel, utilizing your marketing and sales resources throughout the process. Customer retention, on the other hand, is generally more cost-effective as you've already earned the trust and loyalty of existing customers.

Customer churn impedes growth, so companies should have a defined method for calculating customer churn in a given period of time. By being aware of and monitoring churn rate, organizations are equipped to determine their customer retention success rates and identify strategies for improvement.

Various organizations calculate customer churn rate in a variety of ways, as churn rate may represent the total number of customers lost, the percentage of customers lost compared to the company's total customer count, the value of recurring business lost, or the per cent of recurring value lost. Other organizations calculate churn rate for a certain period of time, such as quarterly periods or fiscal years. One of the most commonly used methods for calculating customer churn is to divide the total number of clients a company has at the beginning of a specified time period by the number of customers lost during the same period.

In this document, I will create a simple customer churn prediction model using one of the open data set from leading financial institution from USA. This document contains all the necessary steps from Data Cleansing to Visualization and from understanding the data set to creating a machine learning model.

Problem Statement

The objective of this dataset is to identify which customer will churn in the future and predict their count of based on the other parameters. This research will help the financial institution to take decisions in advance to prevent customers from leaving their bank.

Data

Sample from the whole dataset is shown below:-

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age
10001	15798485	Copley	565	France	Male	31
10002	15588959	T'ang	569	France	Male	34
10003	15624896	Ku	669	France	Female	20
10004	15639629	McConnan	694	France	Male	39
10005	15638852	Ts'ui	504	Spain	Male	28

Table 1.1 Sample Data(Columns 1-7)

Tenure	Balance	Products	HasCard	IsActive	EstimatedSalary	Exited
1	0	1	0	1	20443.08	0
4	0	1	0	1	4045.9	0
7	0	2	1	0	128838.7	0
4	173255.5	1	1	1	81293.1	0
10	109291.4	1	1	1	187593.2	0

Table 1.2 Sample Data (Columns 8 - 14)

Descriptions of the attributes are given below:-

Attribute	Description
RowNumber	N/A.
CustomerId	Unique customer Id.
Surname	Last name of the customer.
CreditScore	Credit Score of the customer
Geography	Citizenship of the customer.
Gender	Gender of the customer.
Age	Age of the customer.
Tenure	Tenure of customer associated with the bank.
Balance	Account balance of the customer.
Products	Number of services customer availed from the bank.
HasCard	Has Credit card. 0 :- No 1:- Yes
IsActive	Is active customer or not. 0 :- No 1 :- Yes
EstimatedSalary	Estimated Salary of the customer.
Exited	Customer Left the bank or not. 0 :- No 1 :- Yes

Table 1.3 Attribute Descriptions

Methodology

Any predictive modelling requires looking at the data before start modelling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis (EDA).

Exploratory data analysis (EDA) is a very important step which takes place after feature engineering and acquiring data and it should be done before any modelling. This is because it is very important for a data scientist to be able to understand the nature of the data without making assumptions.

The purpose of EDA is to use summary statistics and visualizations to better understand data, and find clues about the tendencies of the data, its quality and to formulate assumptions and the hypothesis of our analysis. EDA is not about making fancy visualizations or even aesthetically pleasing ones, the goal is to try and answer questions with data. A goal should be to be able to create a figure which someone can look at in a couple of seconds and understand what is going on. If not, the visualization is too complicated (or fancy) and something similar should be used.

EDA is also very iterative since we first make assumptions based on our first exploratory visualizations, and then build some models. We then make visualizations of the model results and tune our models.

Remember the quality of our inputs decide the quality of our output. So, once we have got our business hypothesis ready, it makes sense to spend lot of time and efforts here. Estimating, data exploration, cleaning and preparation can take up to 70% of our total project time.

Variable Identification

Variable identification is the first step in the exploratory data analysis. Identification of the variables in the dataset is totally dependent on the business requirements and need of the client. The main task here is to identify the Target variable on which future decision has to be made. Besides these identifying the independent variables are equally important because end result of the target variable are totally dependent on the independent/predictor variables. Brief introduction of predictor and target variables are given below:-

Predictor Variable

Predictor variables are those variables or attributes in the dataset on which the result of the target variable is totally dependent. These variables are those on which decisions are made by the clients to get the maximum profit from the business.

Target Variable

It is the variable or attribute in the whole dataset on which client is mostly interested. Based on the business requirements category of the target variable is identified by the data sciences experts.

Once the target variable and predictor variables are identified, our next task to identify the data types and categories of the variable. From analysing the dataset of bike renting company, the detailed descriptions of all the variables are given below in the table:-

Attribute	Type of Variable	Data Type	Variable Type
RowNumber	N/A	N/A	N/A
CustomerId	Continuous	Integer	Predictor Variable
Surname	Label	Object	Predictor Variable
CreditScore	Continuous	Integer	Predictor Variable
Geography	Continuous	Object	Predictor Variable
Gender	Categorical	Object	Predictor Variable
Age	Continuous	Integer	Predictor Variable
Tenure	Continuous	Integer	Predictor Variable
Balance	Continuous	Float	Predictor Variable
Products	Categorical	Integer	Predictor Variable
HasCard	Categorical	Integer	Predictor Variable

IsActive	Categorical	Integer	Predictor Variable
EstimatedSalary	Continuous	Float	Predictor Variable
Exited	Categorical	Integer	Target Variable

Table 1.4

In the further stages of exploratory data analysis process, we have to dive deep into the understanding of the each variable present in the dataset. From Business point of view each and every variable is crucial and even a minute mistake here can cause a loss of millions to your client. So to get the detailed summary of all the variables in the dataset on statistical parameters will help to better understanding of data to a data sciences expert. Detailed summary of all the variables are given below:-

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	EstimatedSalary
count	1000.000000	1.000000e+03	1000.000000	1000.000000	1000.00000	1000.000000	1000.00000
mean	10500.500000	1.569274e+07	652.187000	39.220000	4.86000	75160.026950	101500.36066
std	288.819436	7.313593e+04	97.936201	10.764826	2.91082	62975.377861	57860.87521
min	10001.000000	1.556586e+07	366.000000	18.000000	0.00000	0.000000	245.50000
25%	10250.750000	1.562966e+07	582.000000	32.000000	2.00000	0.000000	49099.87250
50%	10500.500000	1.569516e+07	656.000000	38.000000	5.00000	97926.720000	104081.61000
75%	10750.250000	1.575500e+07	719.000000	44.000000	7.00000	128141.972500	151514.41750
max	11000.000000	1.581546e+07	850.000000	91.000000	10.00000	211520.250000	199633.73000

NOTE: I have changed data types of some variables.

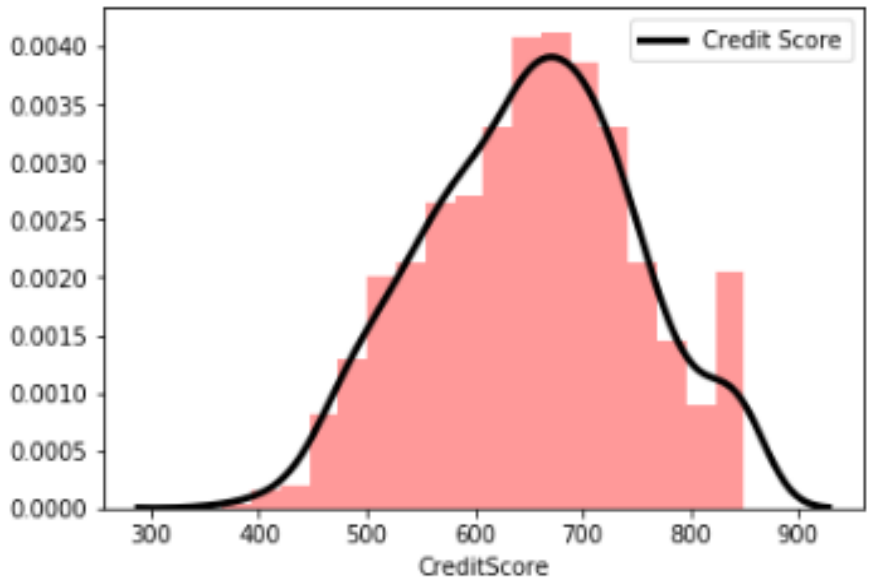
Univariate Analysis

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe. It takes data, summarizes that data and finds patterns in the data.
Method to perform univariate analysis will depend on whether the variable type is categorical or continuous.

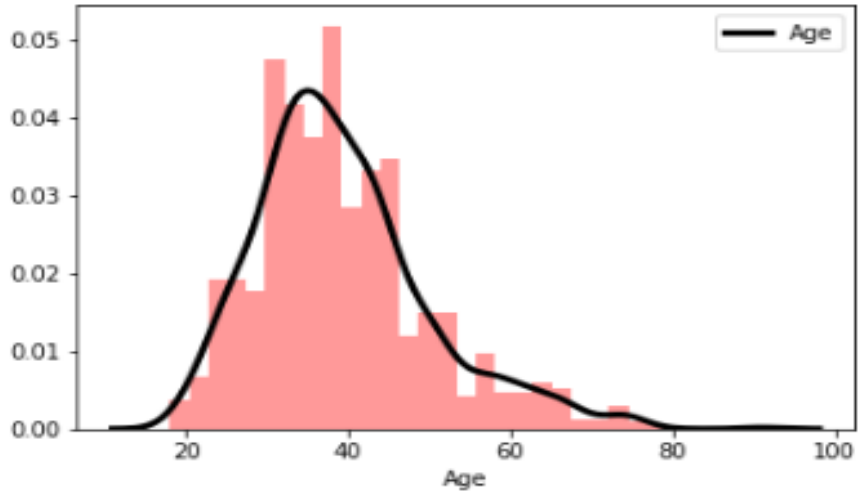
Continuous Variables

A continuous variable is a variable that has an infinite number of possible values. In other words, any value is possible for the variable. All the continuous variables present in our dataset are analysed below:-

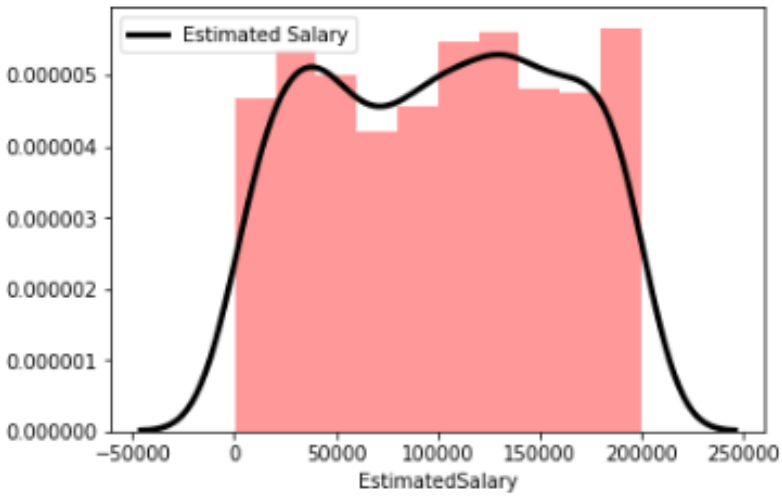
```
count    1000.000000
mean      652.187000
std        97.936201
min       366.000000
25%       582.000000
50%       656.000000
75%       719.000000
max       850.000000
Name: CreditScore, dtype: float64
```



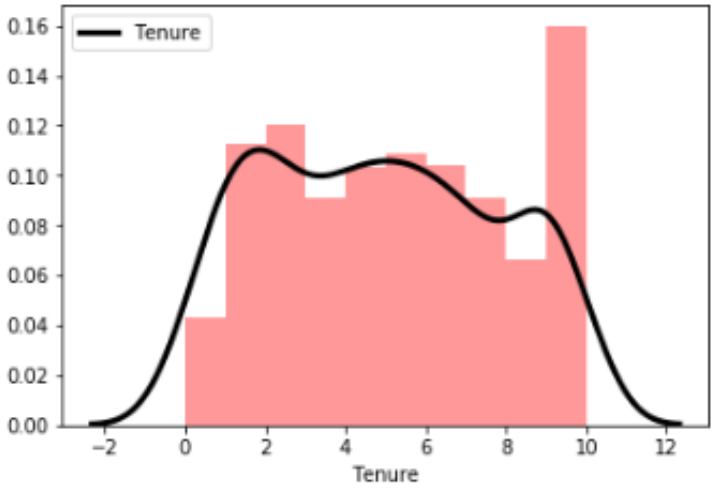
```
count    1000.000000
mean       39.220000
std        10.764826
min        18.000000
25%        32.000000
50%        38.000000
75%        44.000000
max        91.000000
Name: Age, dtype: float64
```



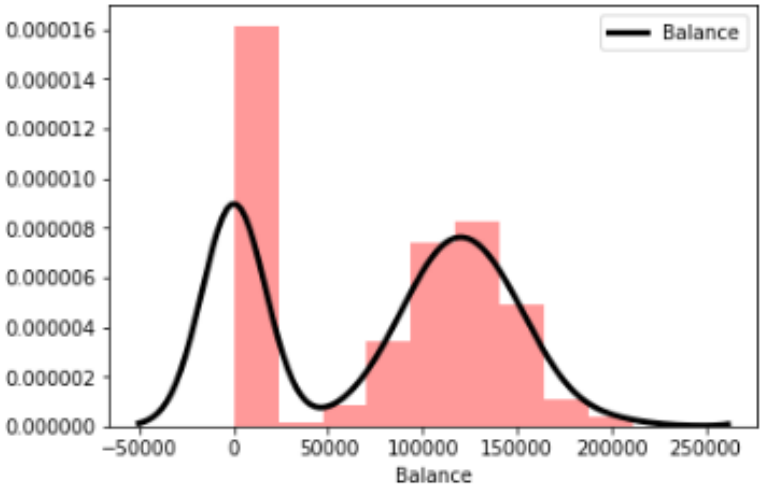

```
count      1000.00000
mean     101500.36066
std       57860.87521
min         245.50000
25%       49099.87250
50%      104081.61000
75%      151514.41750
max      199633.73000
Name: EstimatedSalary, dtype: float64
```



```
count    1000.00000
mean        4.86000
std         2.91082
min         0.00000
25%         2.00000
50%         5.00000
75%         7.00000
max        10.00000
Name: Tenure, dtype: float64
```

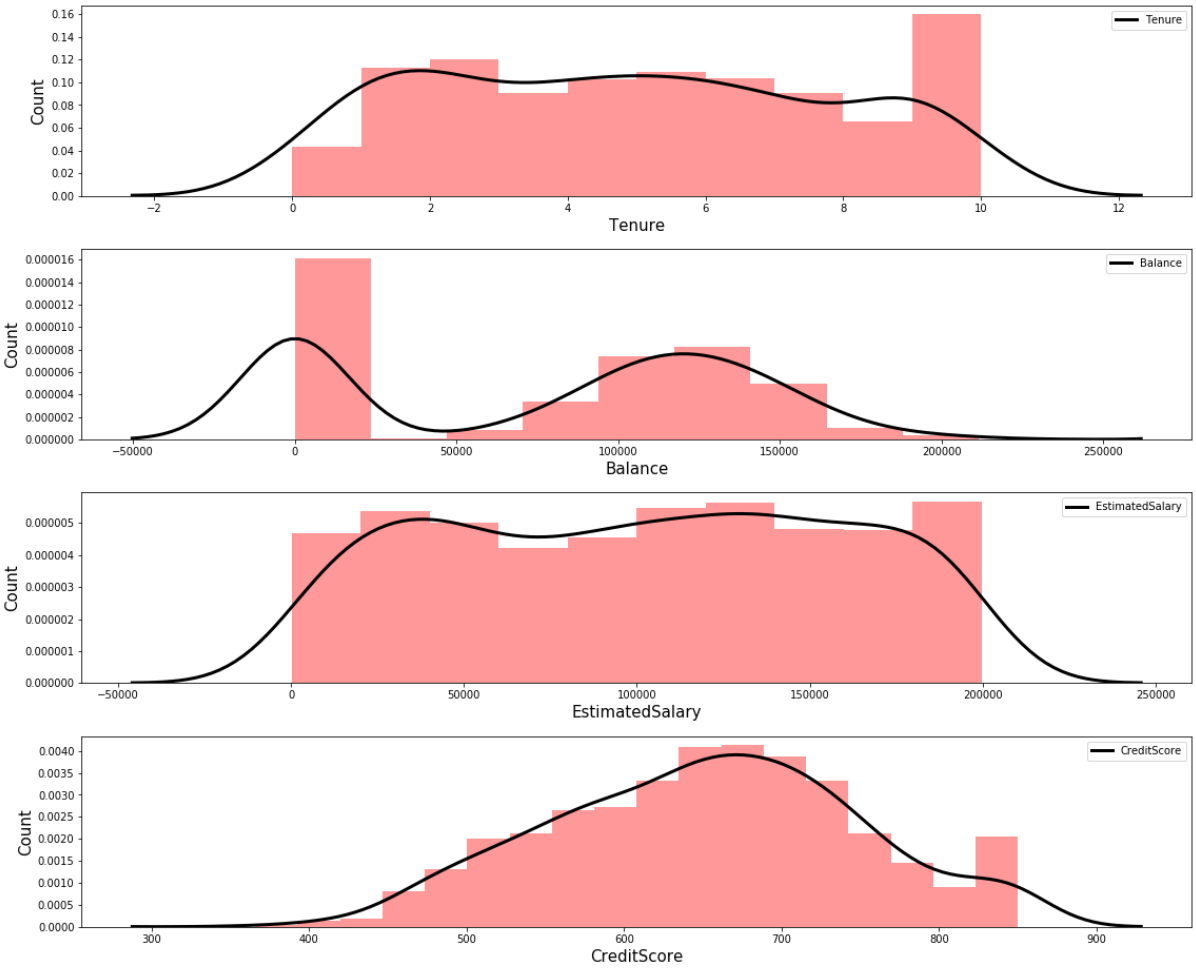



```
count      1000.000000
mean      75160.026950
std       62975.377861
min           0.000000
25%           0.000000
50%       97926.720000
75%      128141.972500
max      211520.250000
Name: Balance, dtype: float64
```



Distribution plot of all numerical Variables

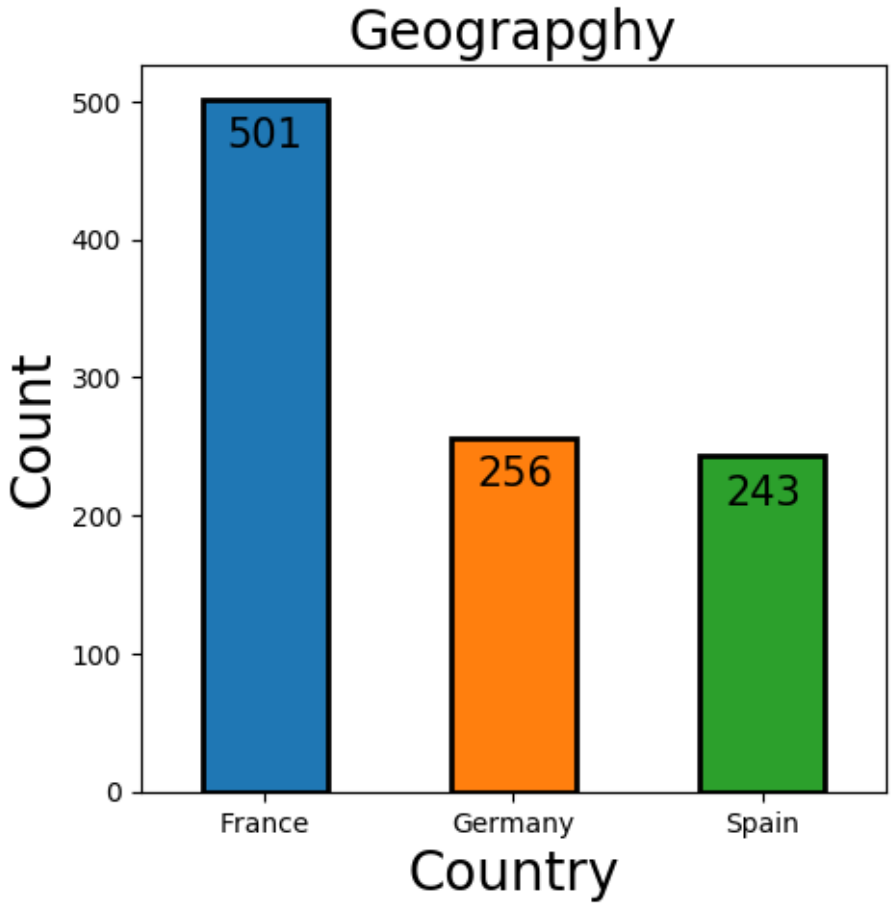
The distribution plot is suitable for comparing range and distribution for groups of numerical data. Distribution of all the numeric variables present in our dataset in shown below:-

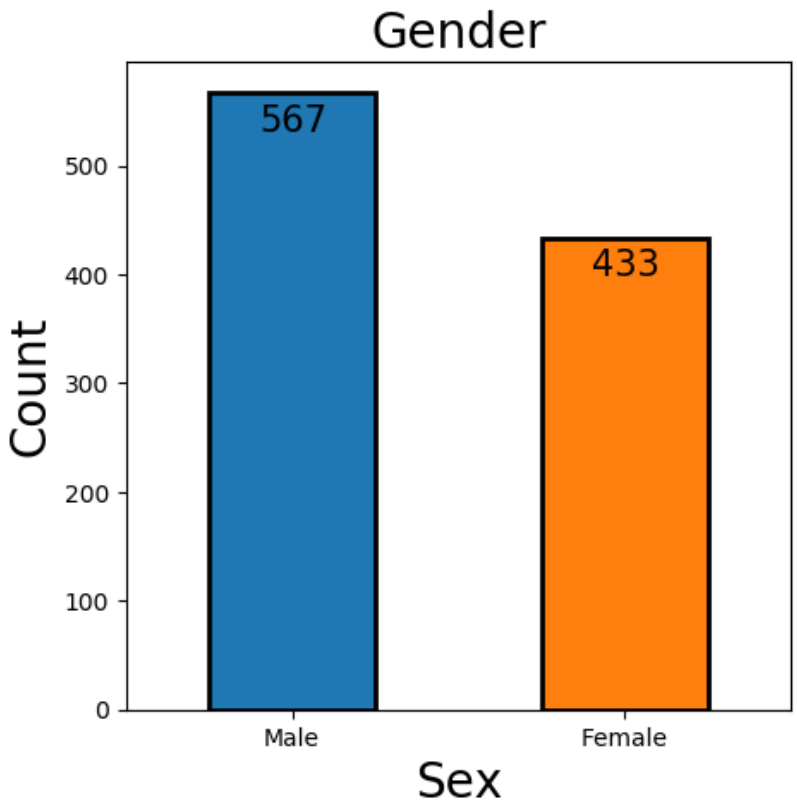


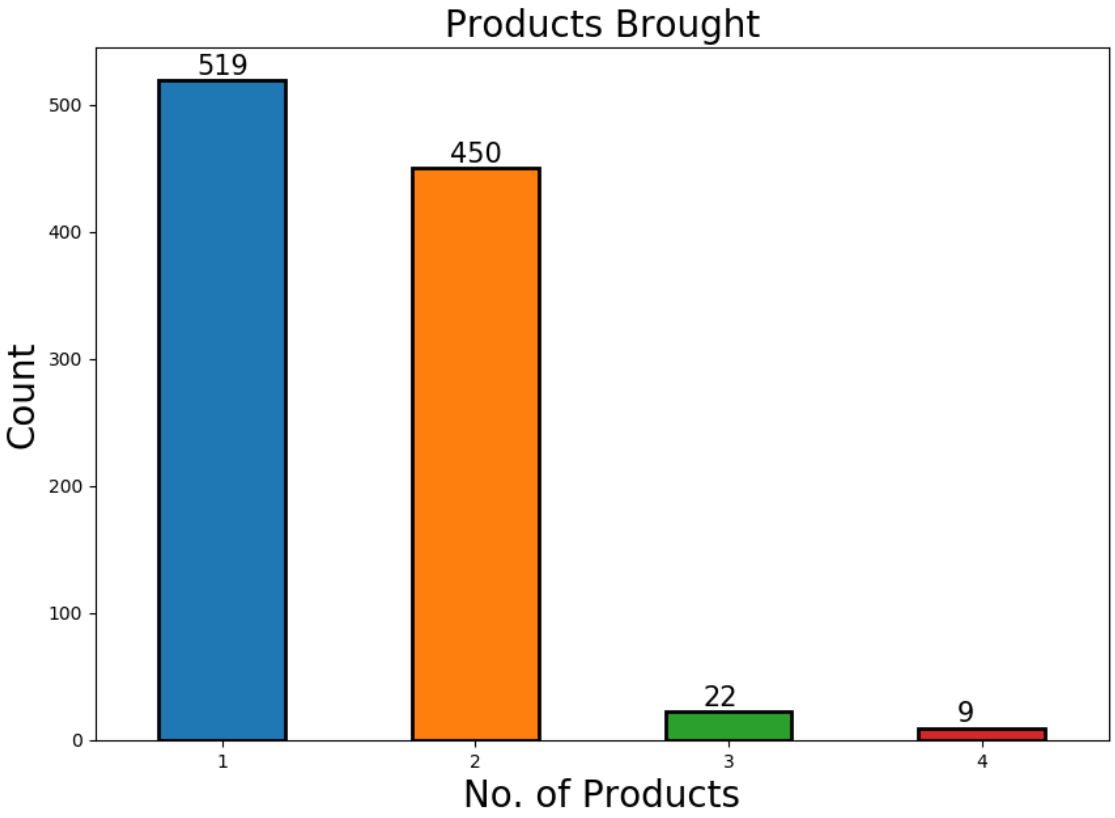
Distribution Plot of all the Numerical Variables

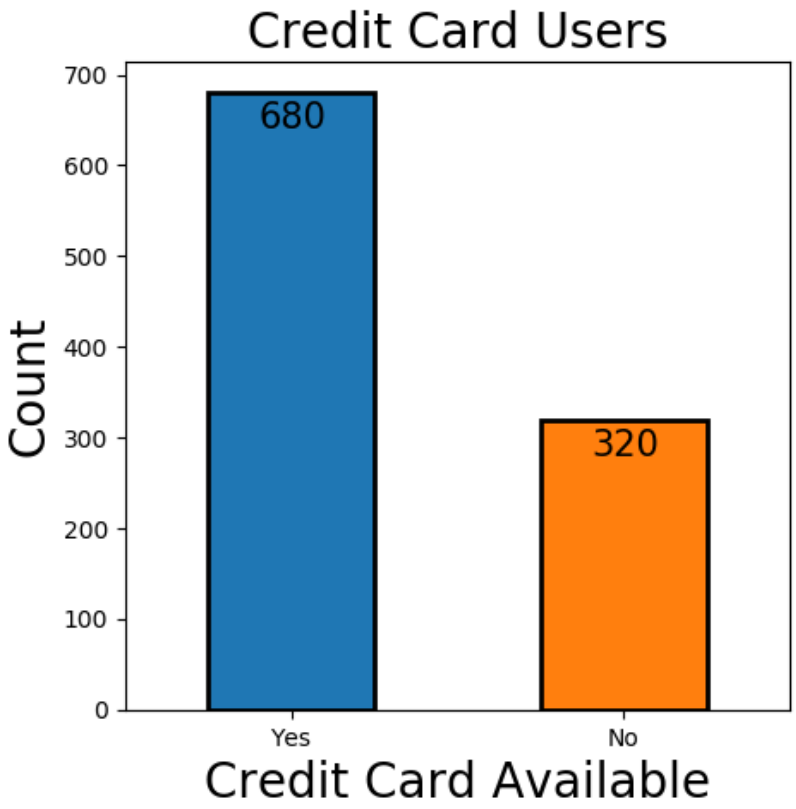
Categorical Variables

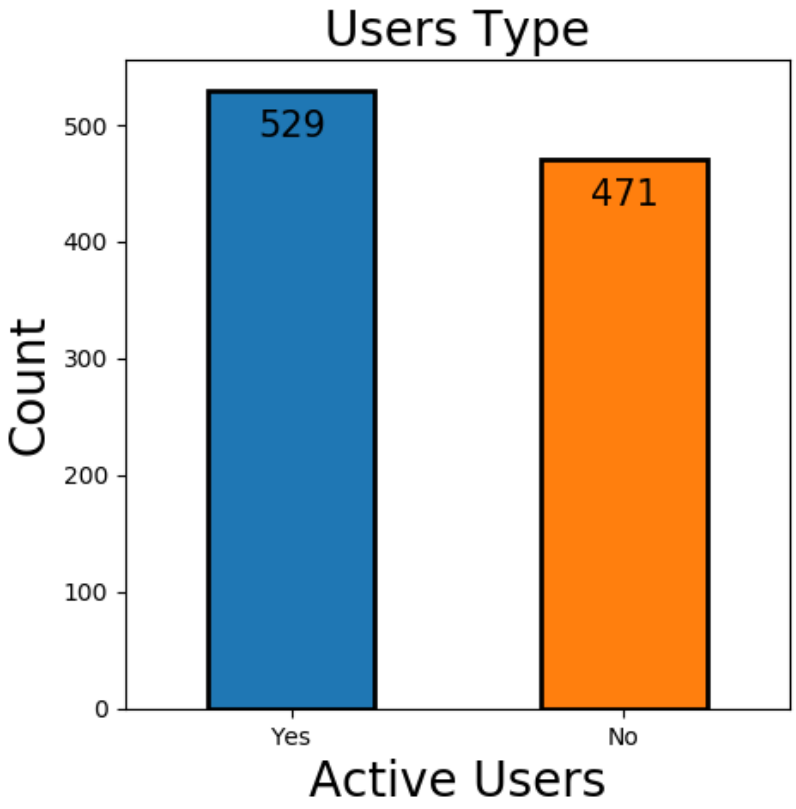
Categorical data are easier to interpret as compared to numerical variables. The best ways to analyze categorical variables are through bar graphs and pie charts. Bar graphs are mostly preferred to visualize the frequency of each category that falls into that variable, whereas pie charts are used to visualize the percentage of each category. Plots for analyzing categorical variables are shown below:-

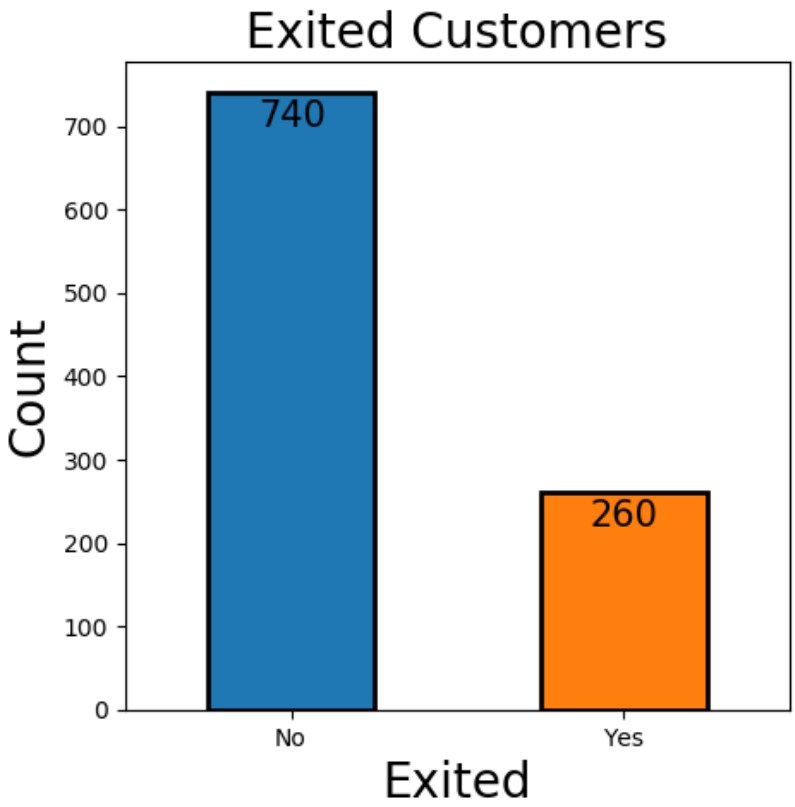



Gender
567
433
500
400
300
200
100
0
Count
Male
Female
Sex


Products Brought
519
450
22
9
500
400
300
200
100
0
Count
1
2
3
4
No. of Products


Credit Card Users
Count
700
600
500
400
300
200
100
0
680
320
Yes
No
Credit Card Available


Users Type
Count
500
400
300
200
100
0
529
471
Yes
No
Active Users

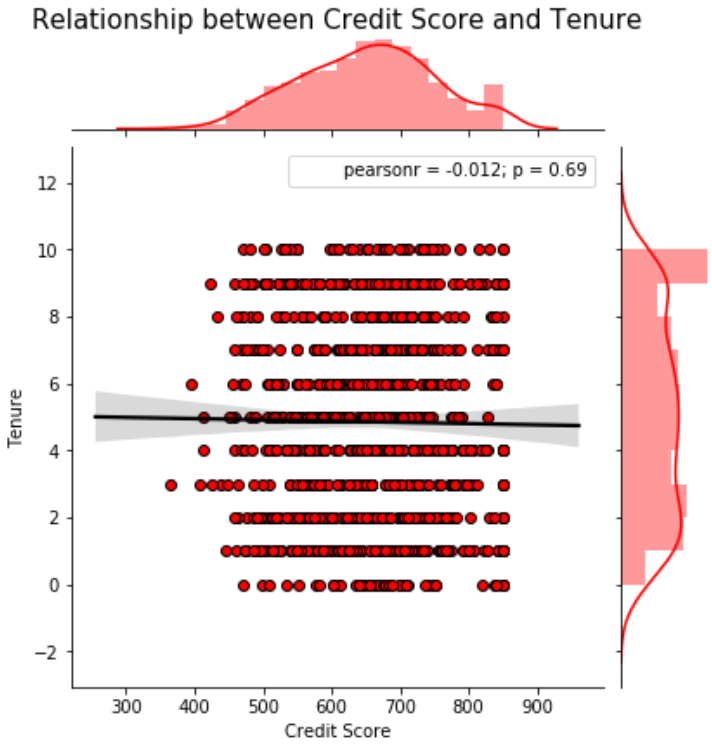

Exited Customers
740
260
700
600
500
400
300
200
100
0
Count
No
Yes
Exited

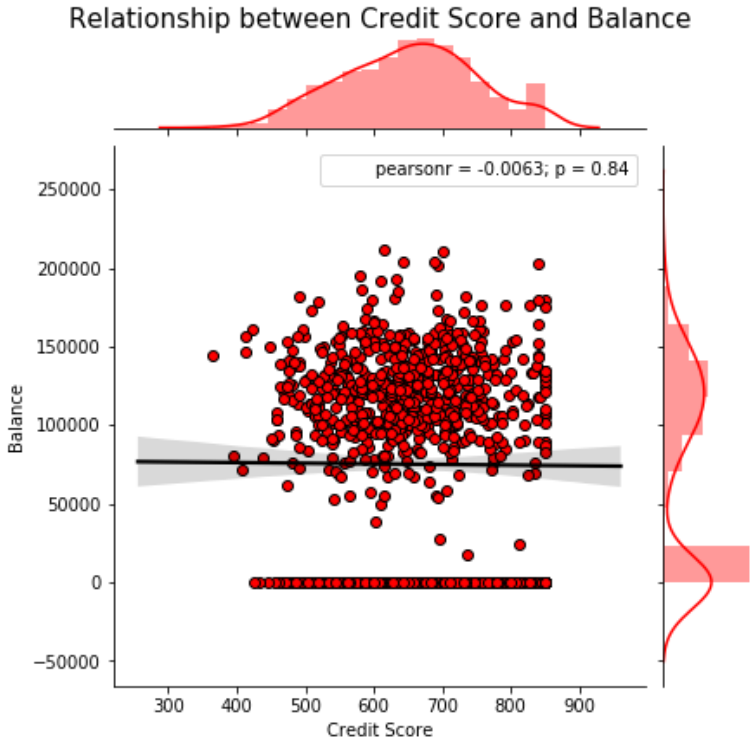
Bivariate Analysis

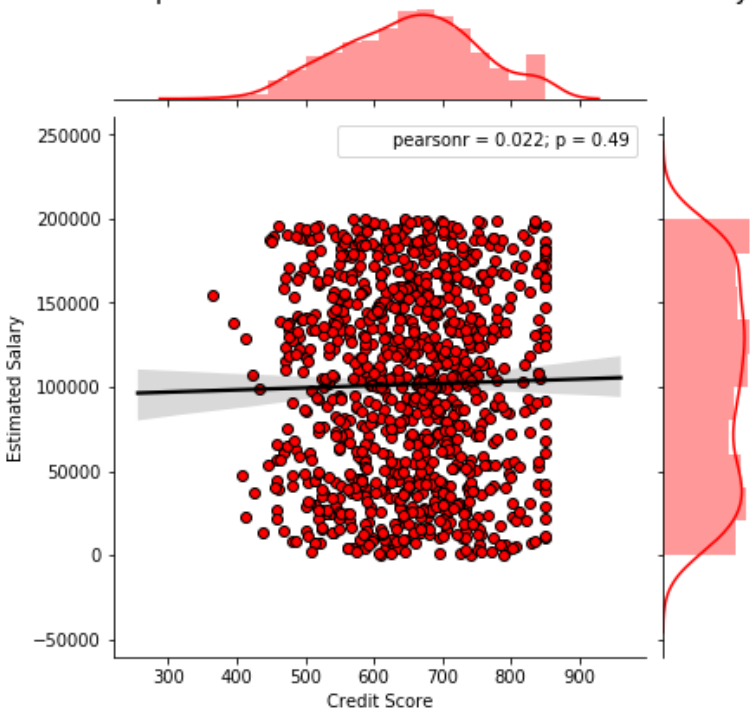
Bivariate analysis is the simultaneous analysis of two variables. It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences.

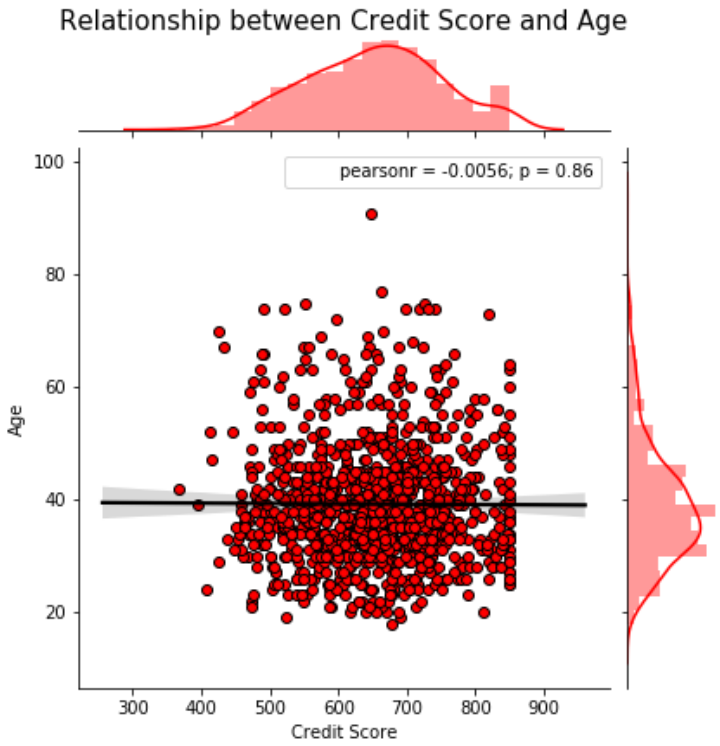
Continuous Variables

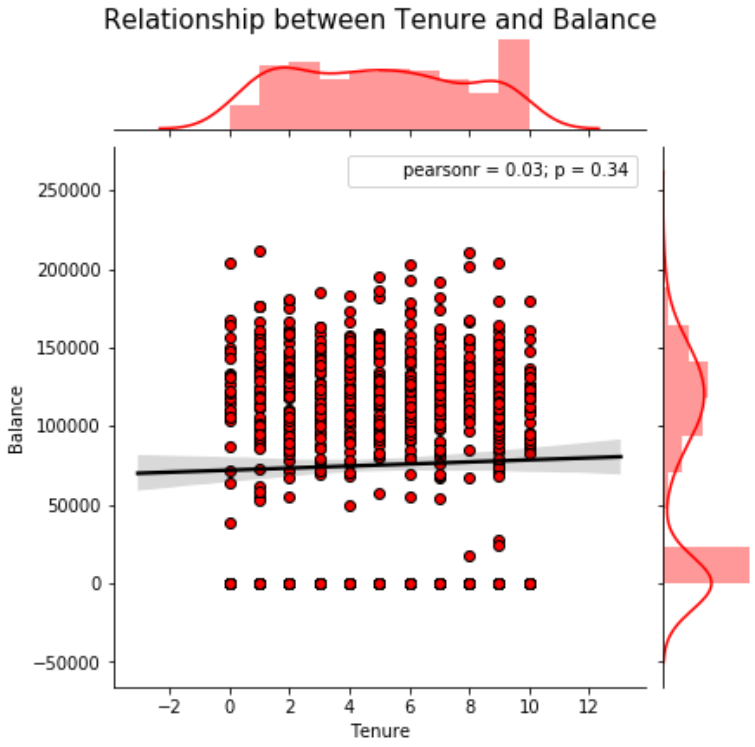
Scatter plot for all the continuous variables are shown below. Correlation value is shown in the legend and it describes the relationship between variables against which graphs are plotted.

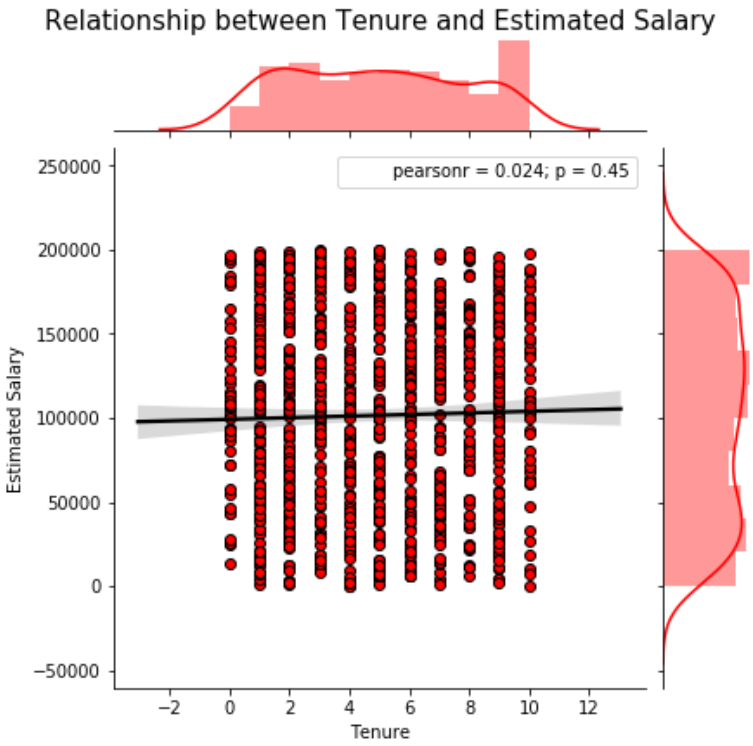


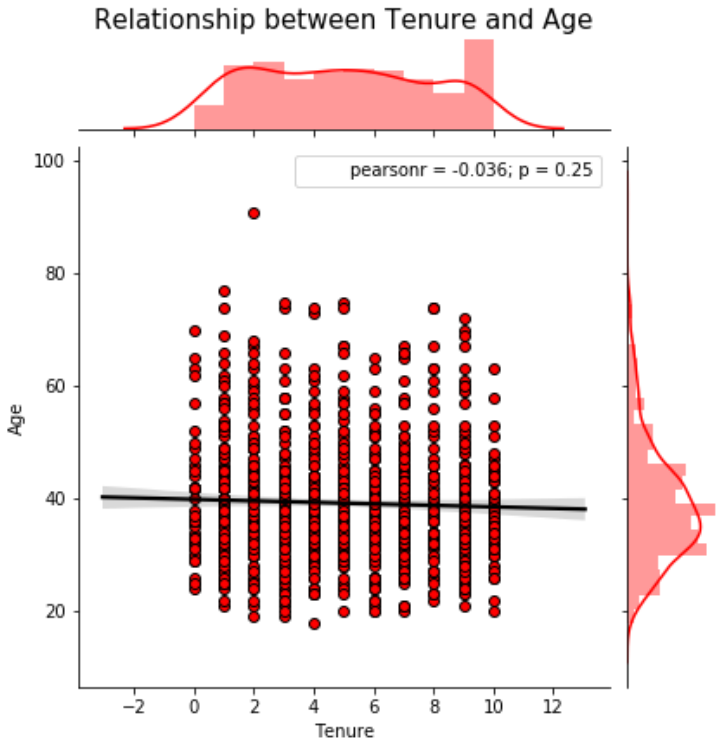

Relationship between Credit Score and Balance
pearsonr = -0.0063; p = 0.84
Balance
Credit Score
250000
200000
150000
100000
50000
0
−50000
300
400
500
600
700
800
900

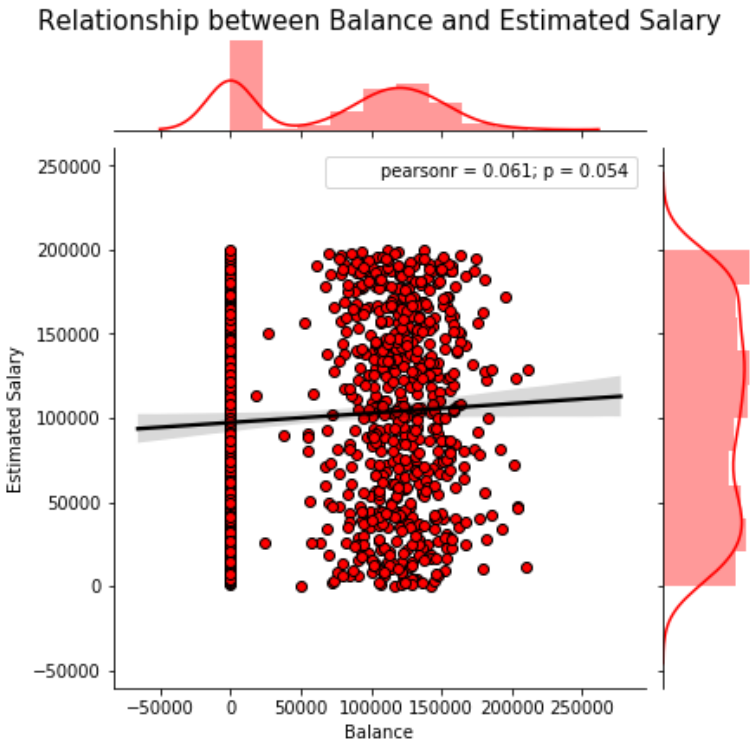

Relationship between Credit Score and Estimated Salary
pearsonr = 0.022; p = 0.49
Estimated Salary
Credit Score
250000
200000
150000
100000
50000
0
−50000
300
400
500
600
700
800
900

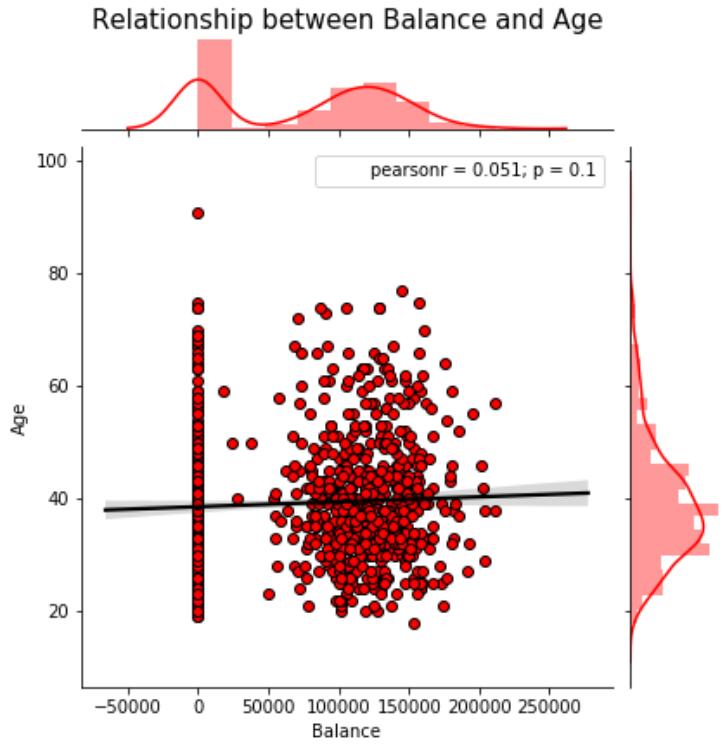

Relationship between Credit Score and Age
pearsonr = -0.0056; p = 0.86
Age
Credit Score
100
80
60
40
20
300
400
500
600
700
800
900

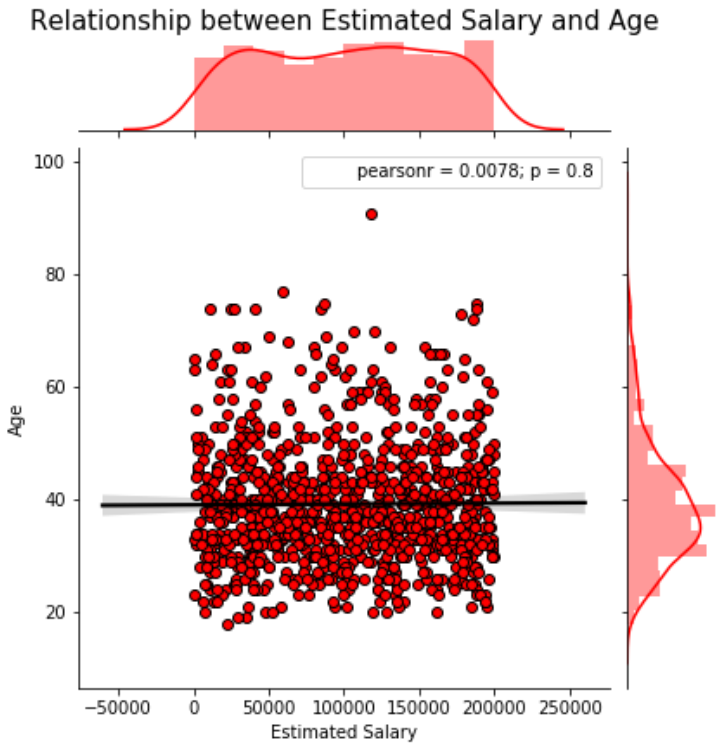

Relationship between Tenure and Balance
pearsonr = 0.03; p = 0.34
Balance
Tenure
250000
200000
150000
100000
50000
0
−50000
−2
0
2
4
6
8
10
12


Relationship between Tenure and Estimated Salary
pearsonr = 0.024; p = 0.45
Estimated Salary
Tenure


Relationship between Tenure and Age
pearsonr = -0.036; p = 0.25
Age
Tenure

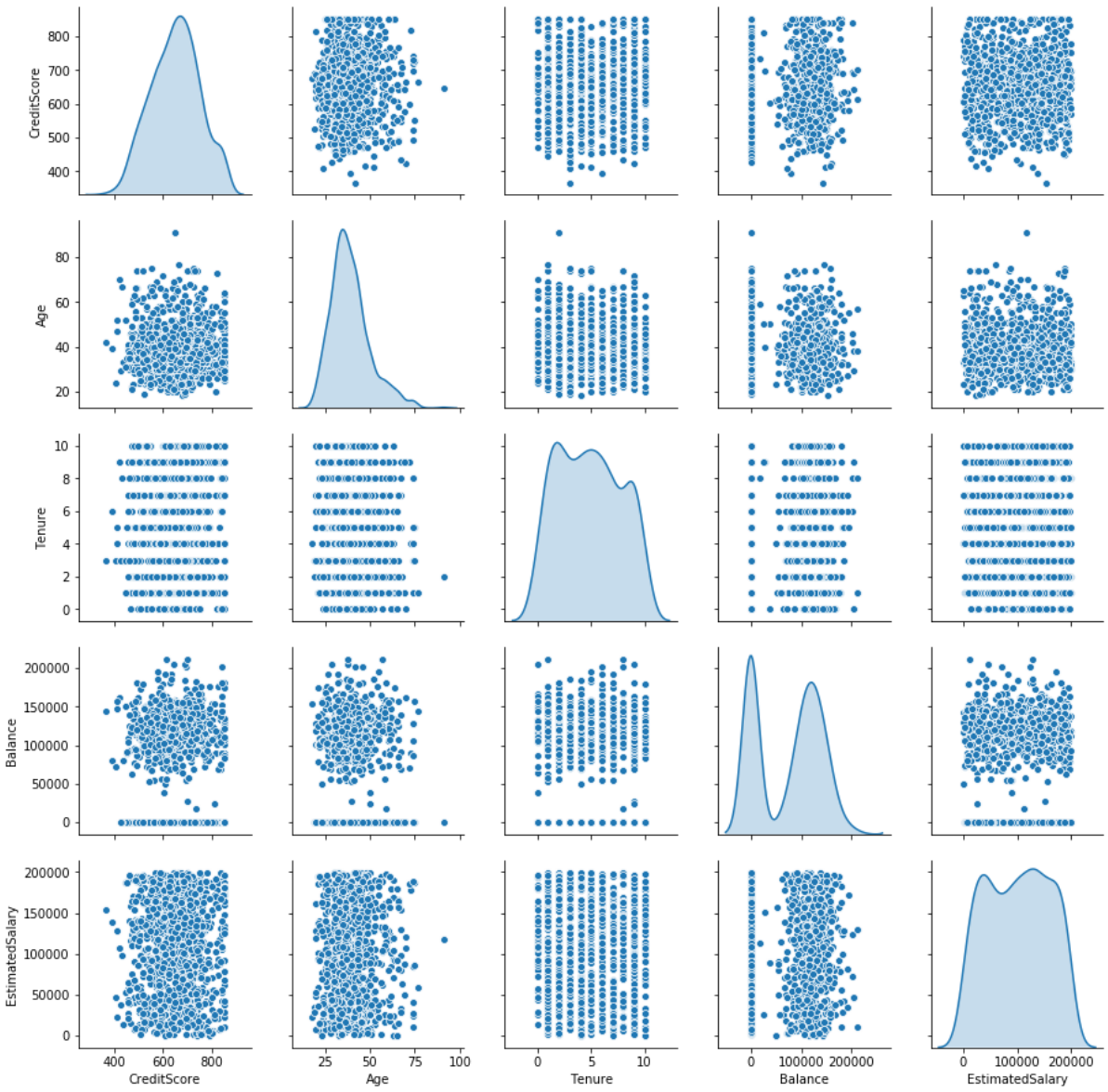

Relationship between Balance and Estimated Salary
pearsonr = 0.061; p = 0.054
Estimated Salary
Balance
250000
200000
150000
100000
50000
0
−50000
−50000
0
50000
100000
150000
200000
250000


Relationship between Balance and Age
pearsonr = 0.051; p = 0.1
Age
Balance
100
80
60
40
20
−50000
0
50000
100000
150000
200000
250000

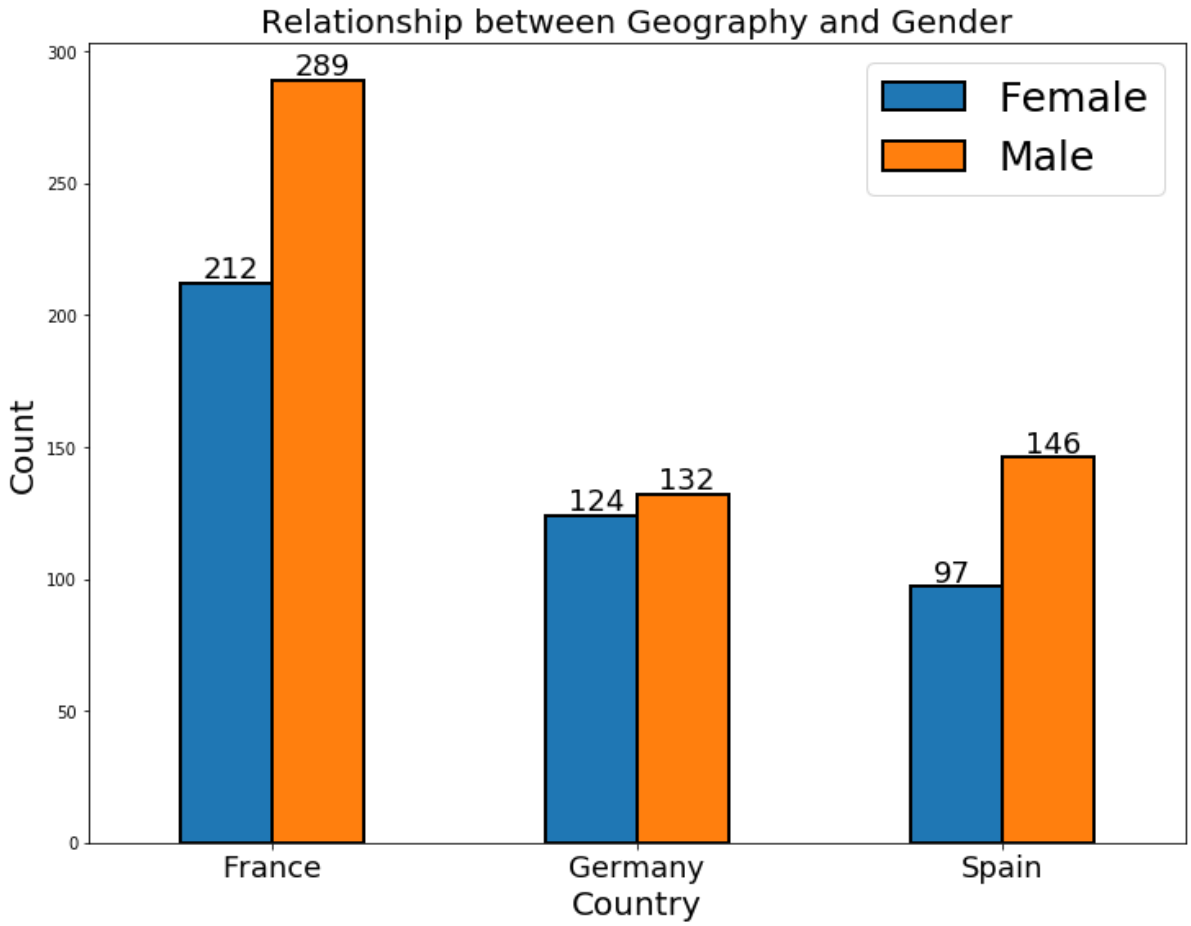

Relationship between Estimated Salary and Age
pearsonr = 0.0078; p = 0.8
Age
Estimated Salary
100
80
60
40
20
−50000
0
50000
100000
150000
200000
250000

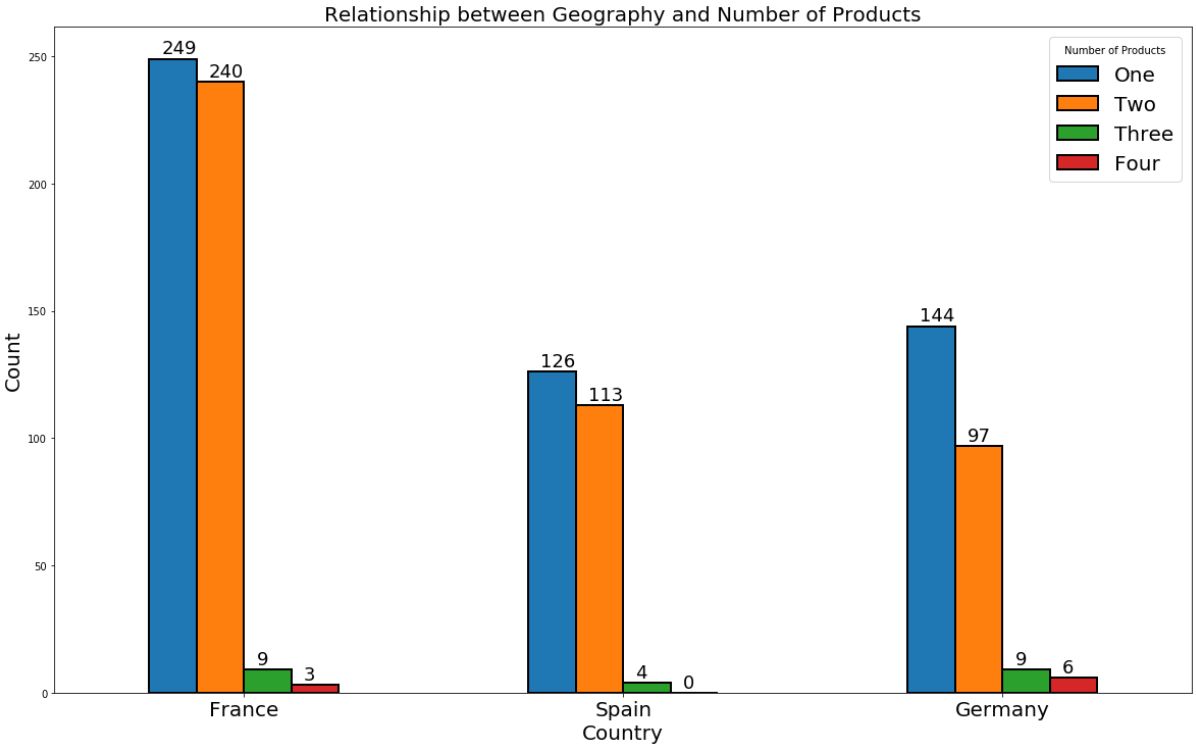
Pair Plot

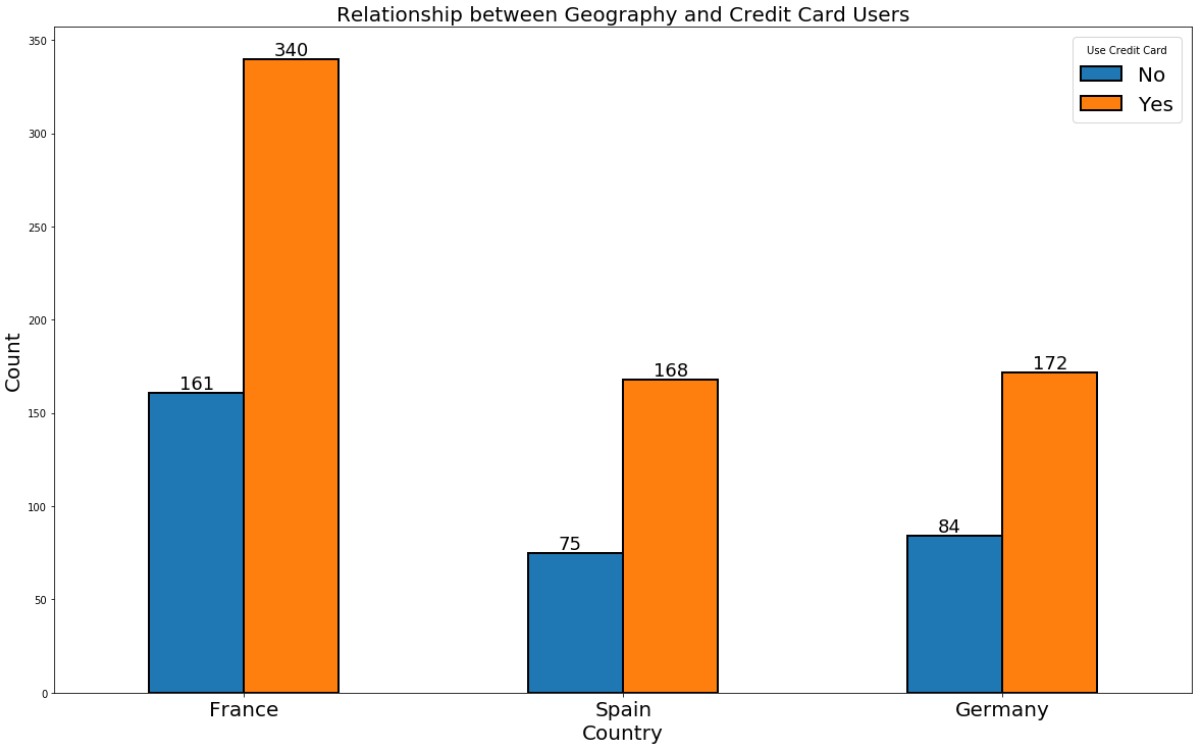
Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our dataset. Pair plots for all our numerical variables are shown below:-

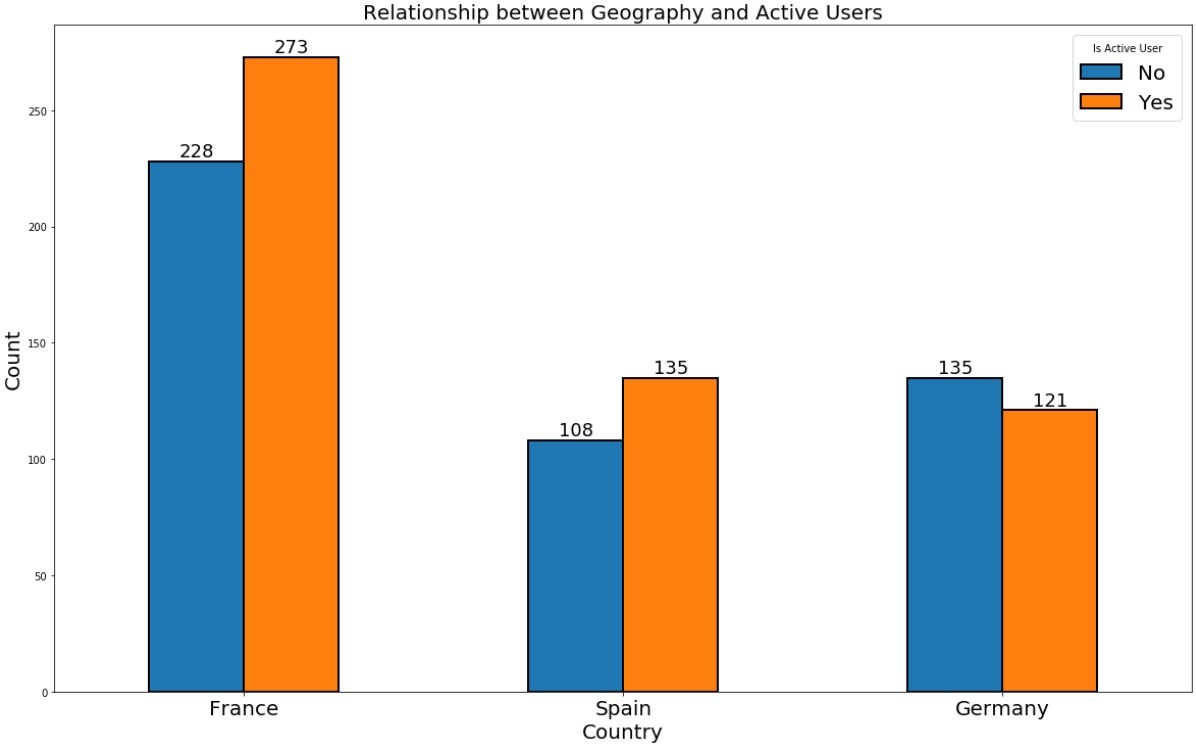


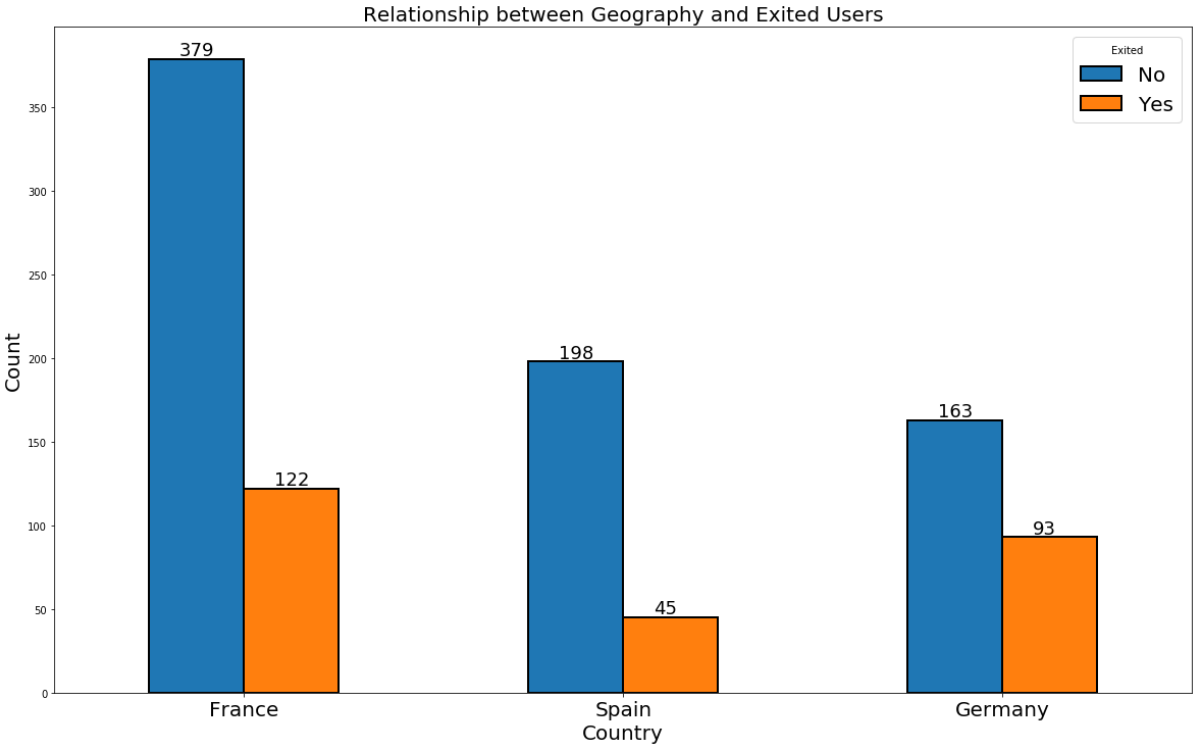
Categorical Variables

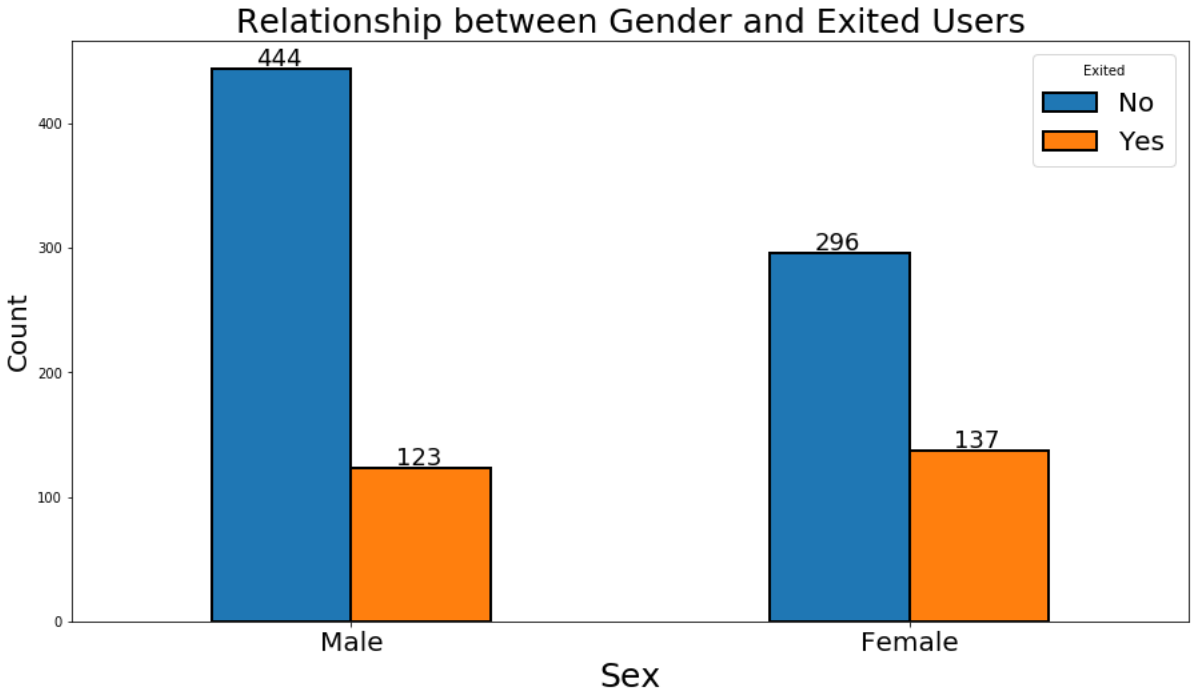


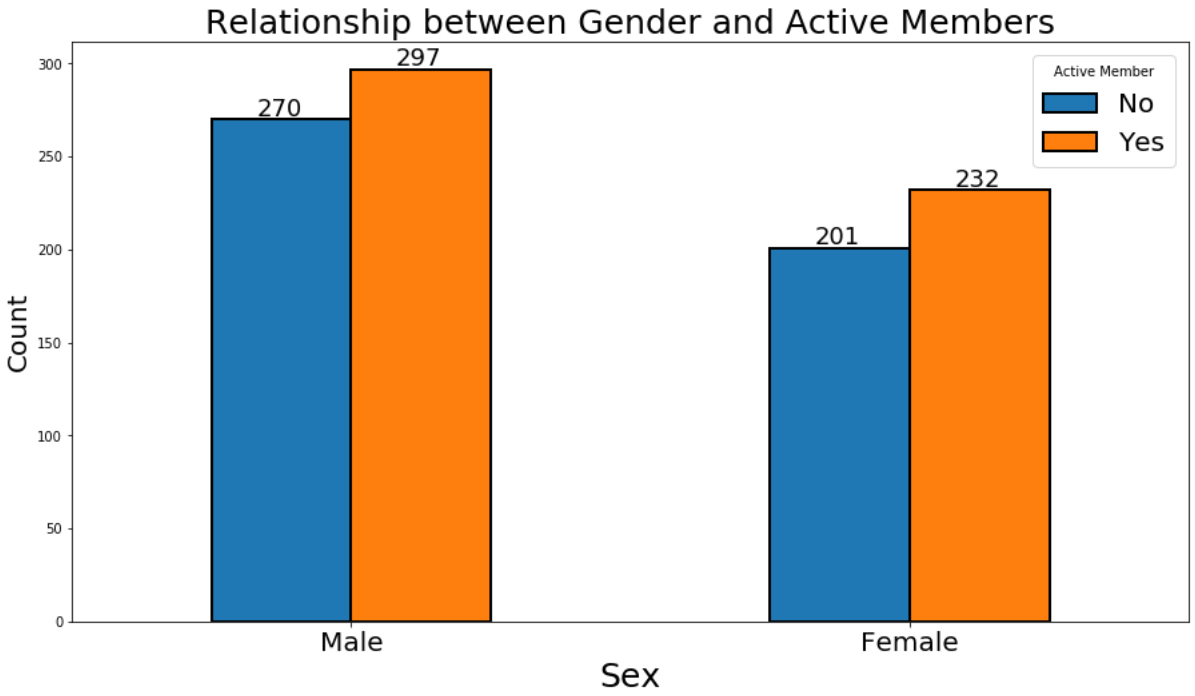


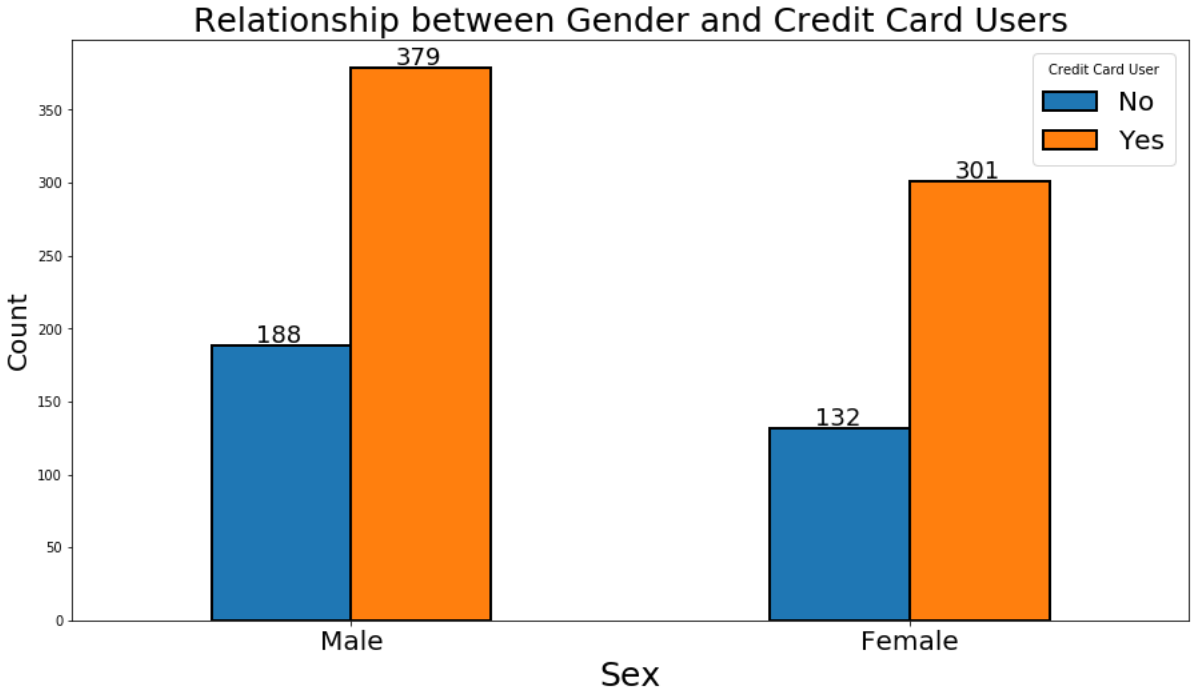

Relationship between Geography and Credit Card Users
Use Credit Card
No
Yes
Count
Country
France
Spain
Germany
161
340
75
168
84
172

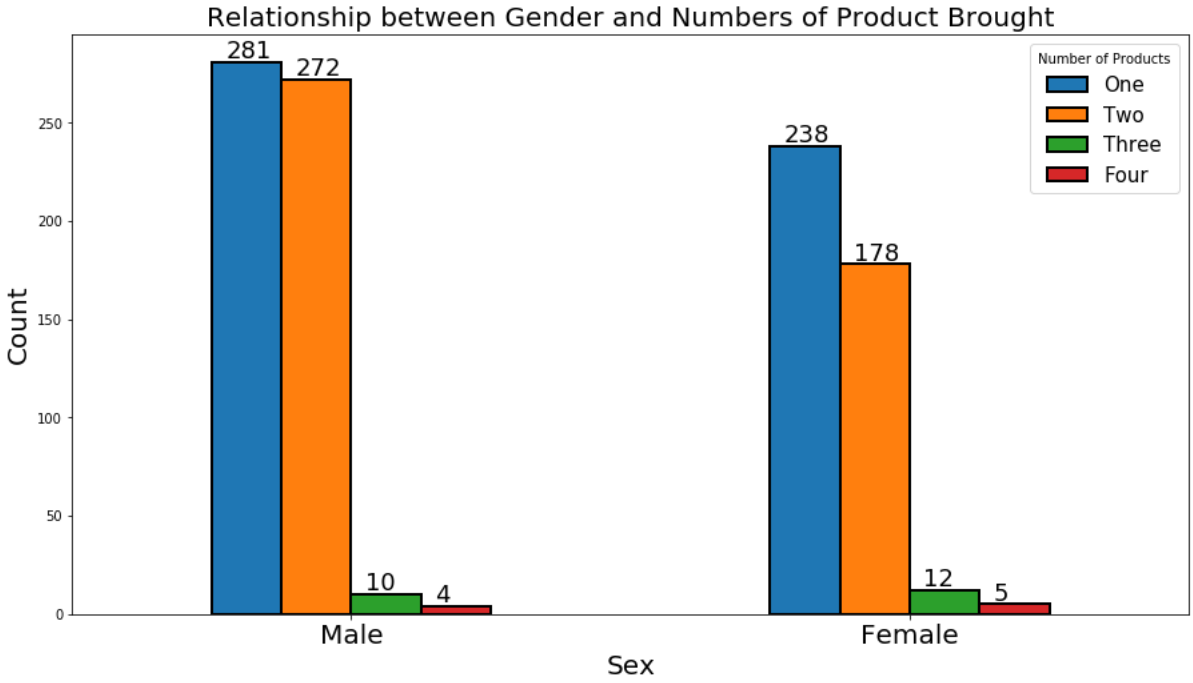

Relationship between Geography and Active Users
Is Active User
No
Yes
Count
Country
France
Spain
Germany
228
273
108
135
135
121

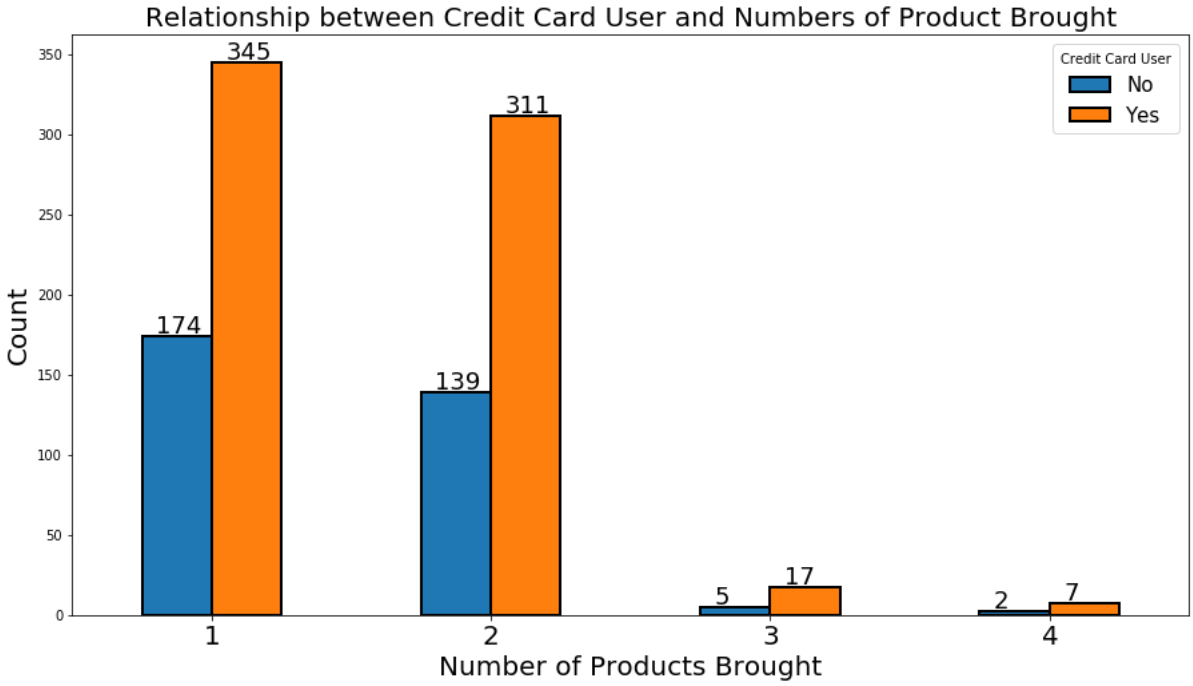

Relationship between Geography and Exited Users
Exited
No
Yes
Count
379
122
198
45
163
93
0
50
100
150
200
250
300
350
France
Spain
Germany
Country

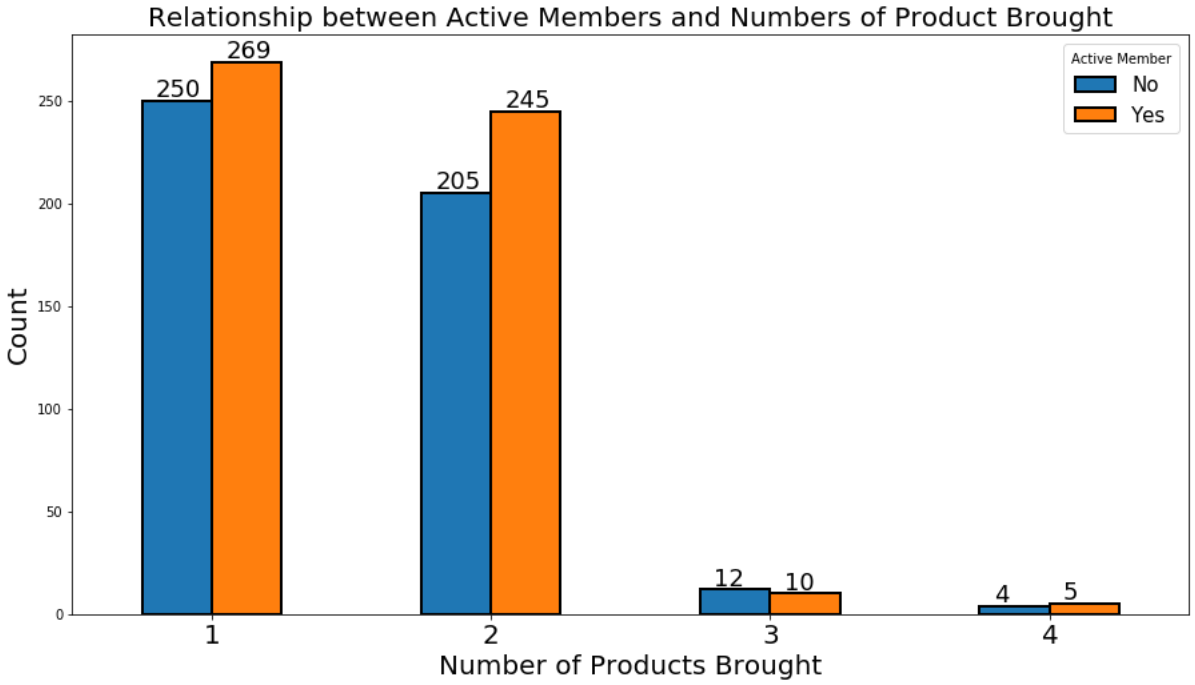

Relationship between Gender and Exited Users
Exited
No
Yes
Count
444
123
296
137
0
100
200
300
400
Male
Female
Sex

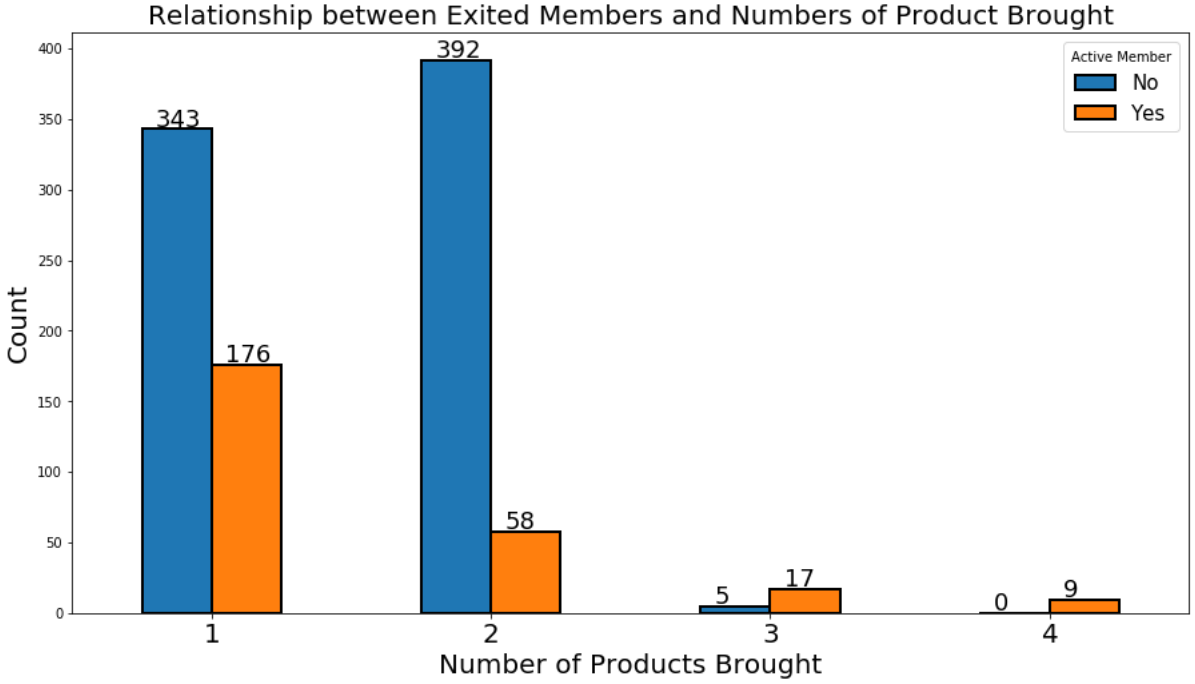

Relationship between Gender and Active Members
Active Member
No
Yes
297
270
232
201
Count
300
250
200
150
100
50
0
Male
Female
Sex

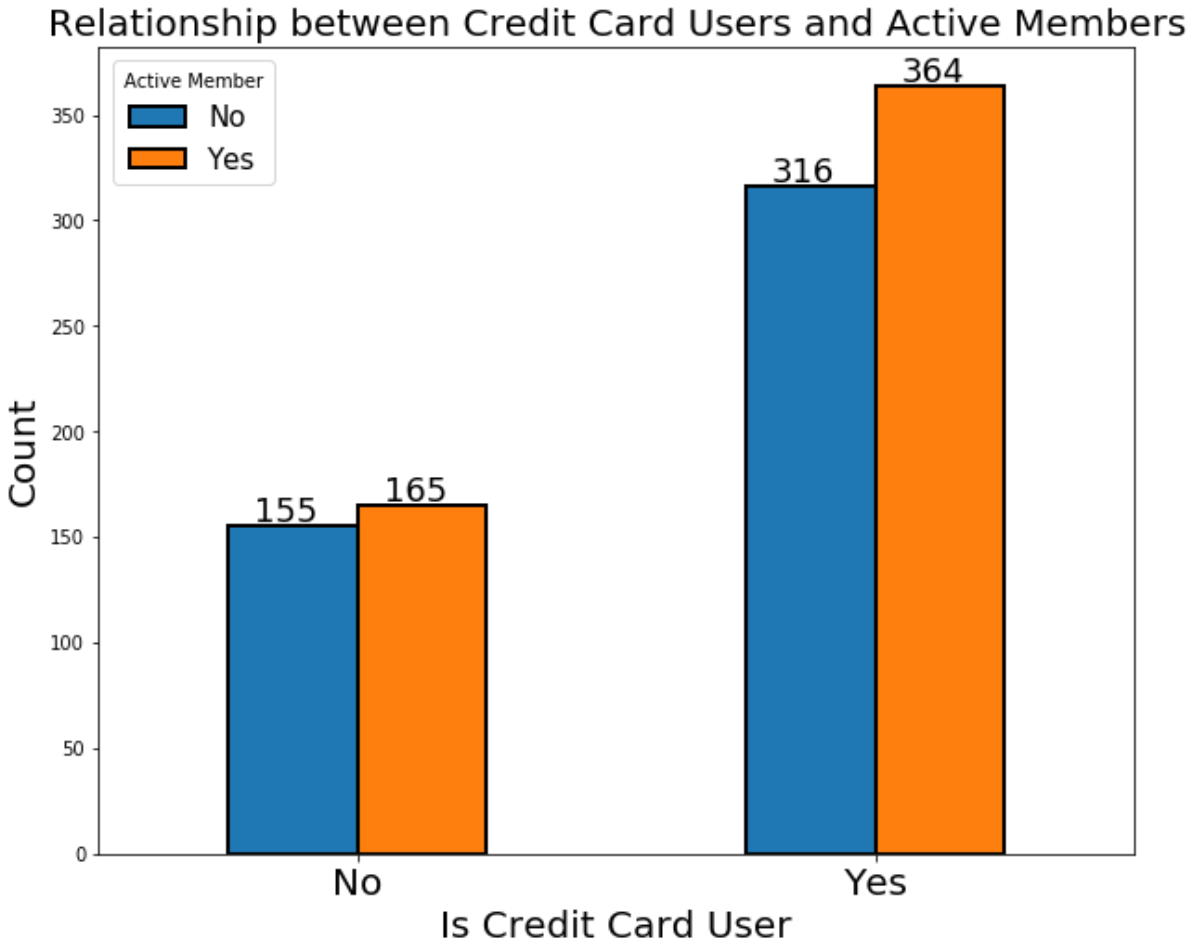

Relationship between Gender and Credit Card Users
Credit Card User
No
Yes
379
301
188
132
Count
350
300
250
200
150
100
50
0
Male
Female
Sex

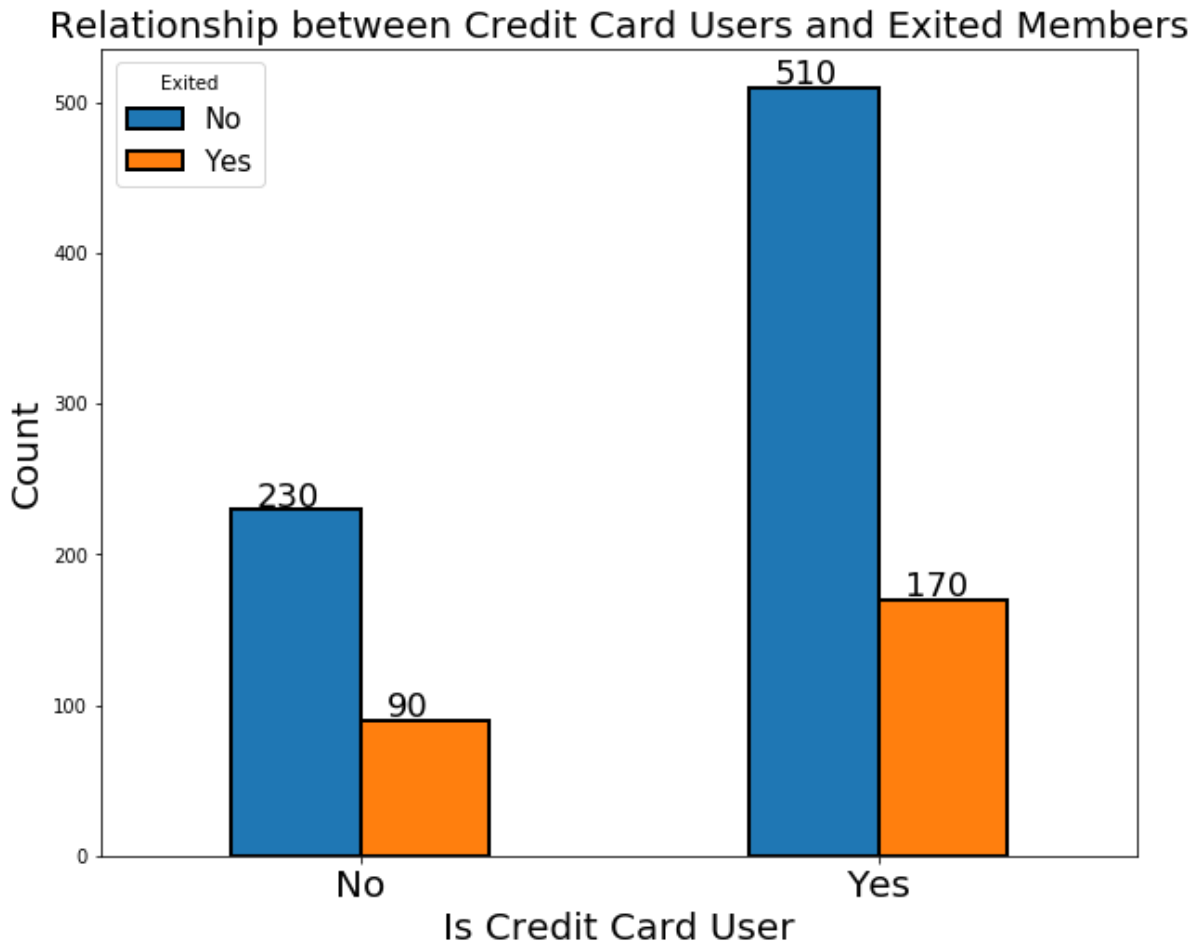

Relationship between Gender and Numbers of Product Brought
Number of Products
One
Two
Three
Four
281
272
10
4
238
178
12
5
Count
Male
Female
Sex

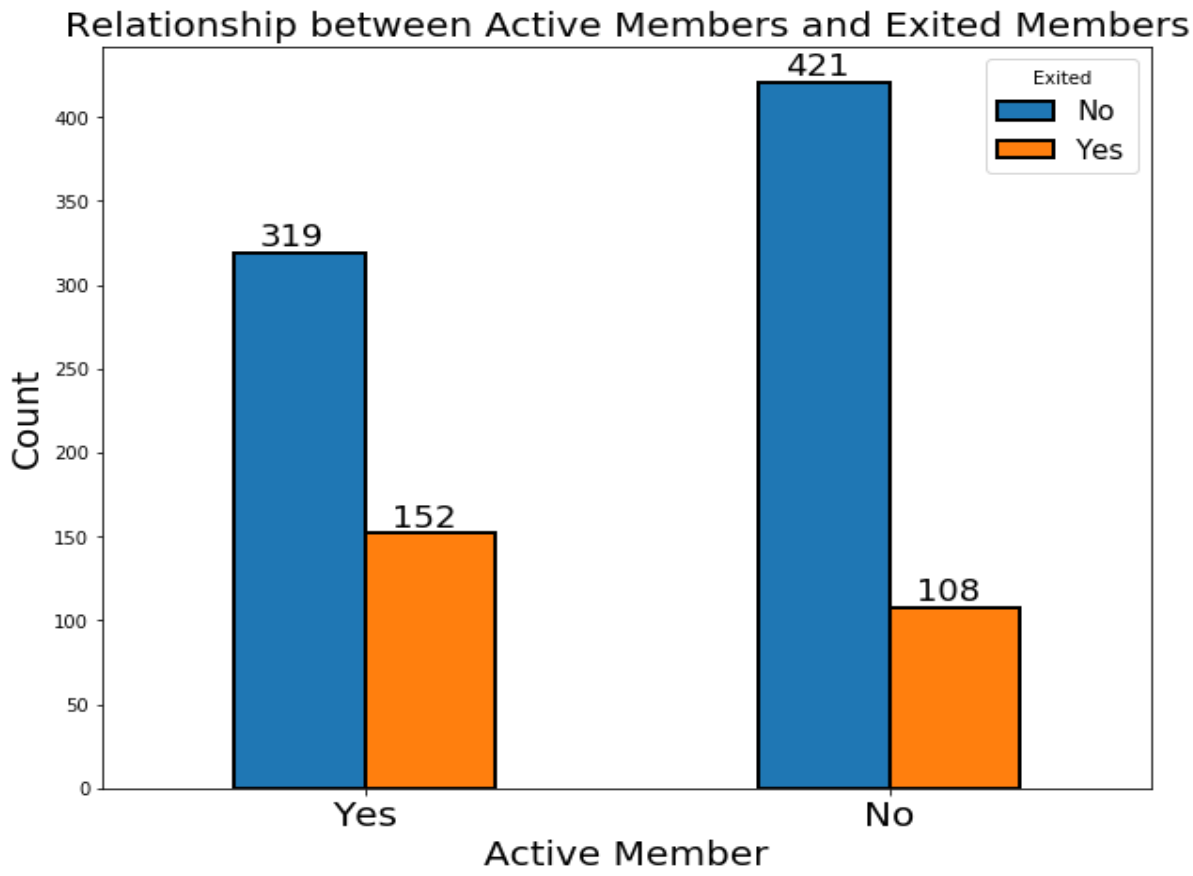

Relationship between Credit Card User and Numbers of Product Brought
Credit Card User
No
Yes
345
174
311
139
17
5
7
2
Count
Number of Products Brought


Relationship between Active Members and Numbers of Product Brought
Active Member
No
Yes
Count
Number of Products Brought
250
269
205
245
12
10
4
5
1
2
3
4

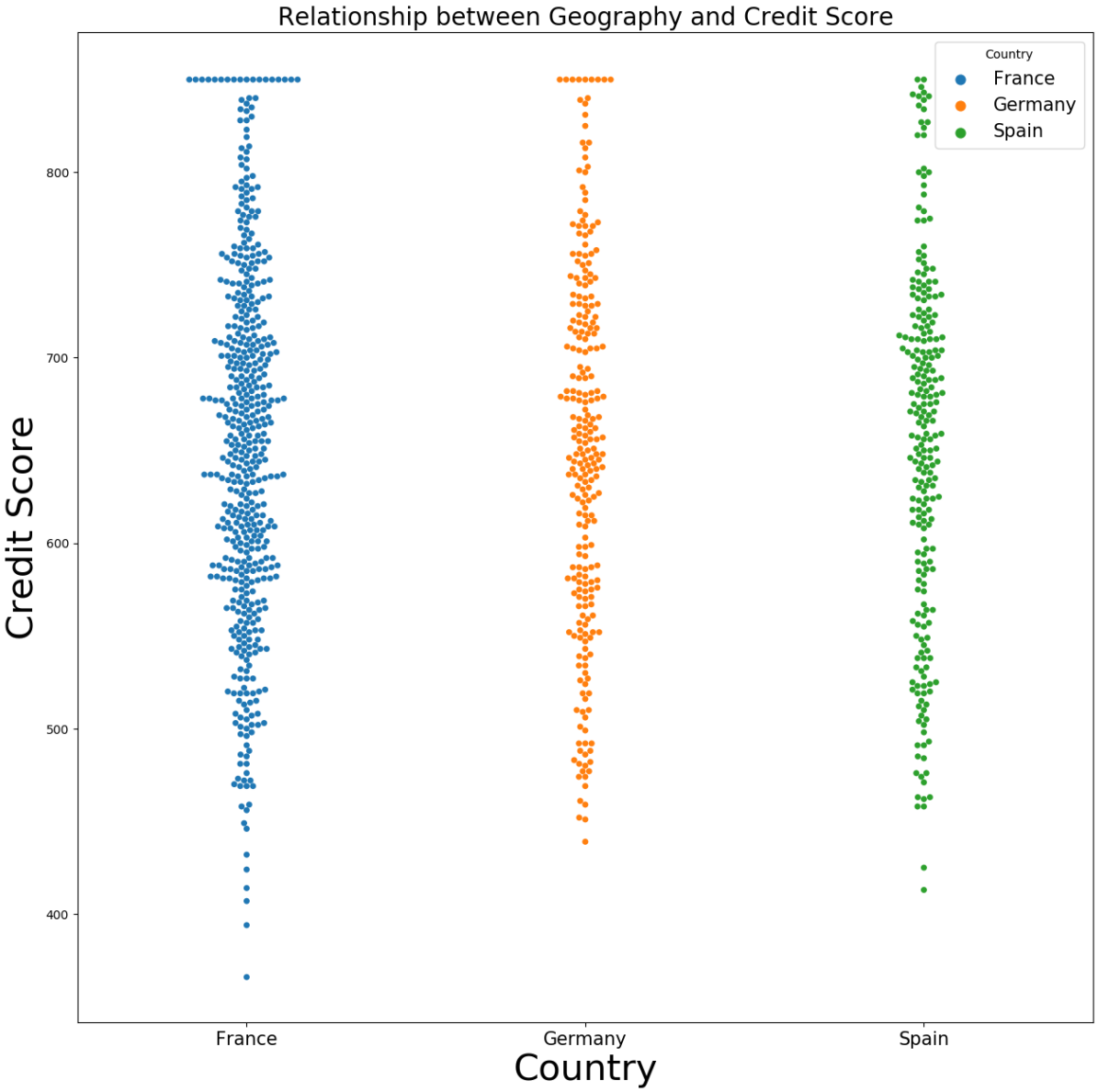

Relationship between Exited Members and Numbers of Product Brought
Active Member
No
Yes
Count
Number of Products Brought
343
176
392
58
5
17
0
9
1
2
3
4

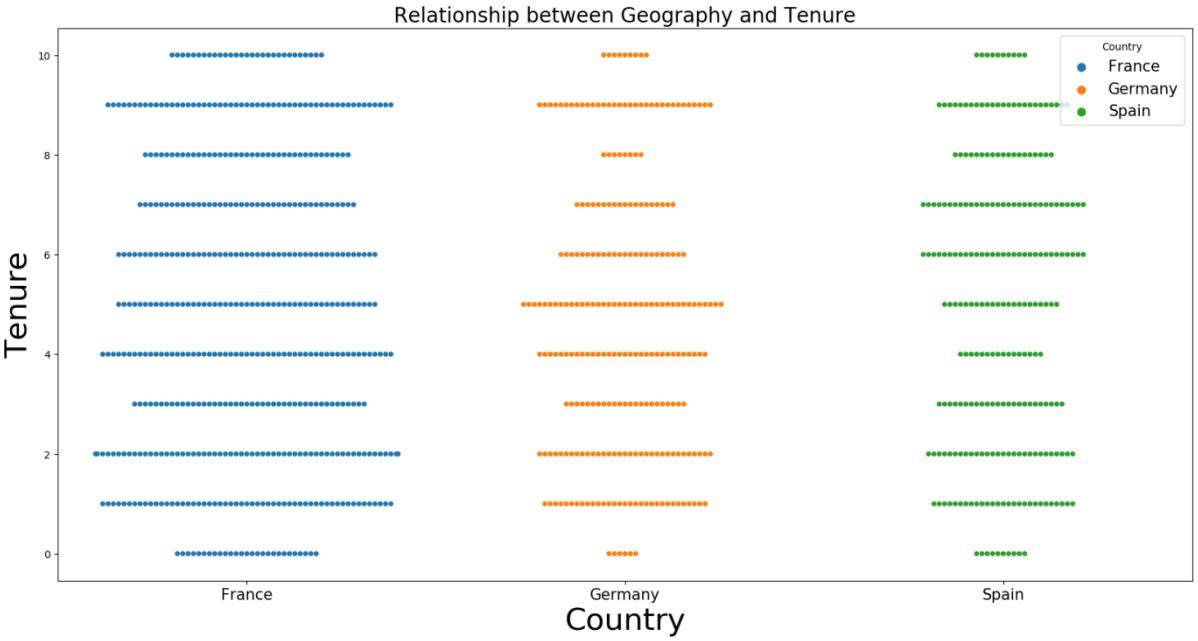

Relationship between Credit Card Users and Active Members
Active Member
No
Yes
Count
0
50
100
150
200
250
300
350
155
165
316
364
No
Yes
Is Credit Card User

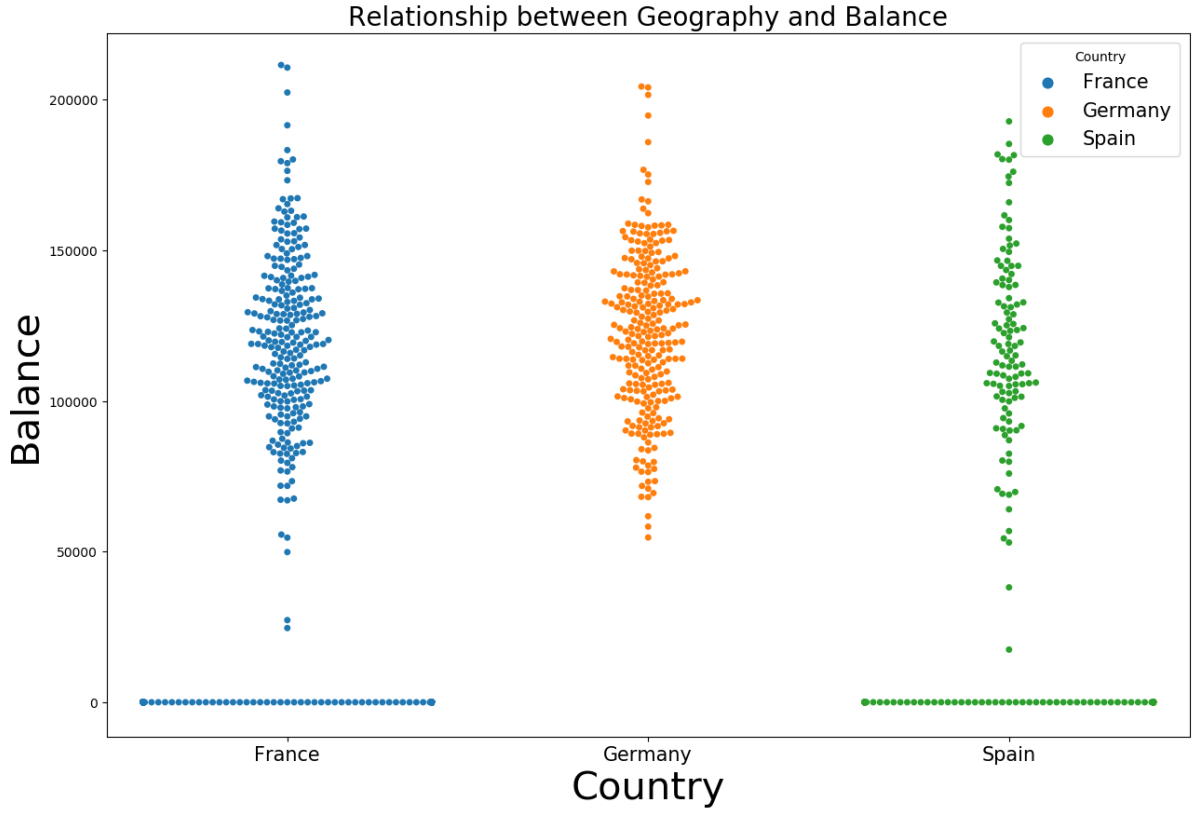

Relationship between Credit Card Users and Exited Members
Exited
No
Yes
Count
0
100
200
300
400
500
230
90
510
170
No
Yes
Is Credit Card User

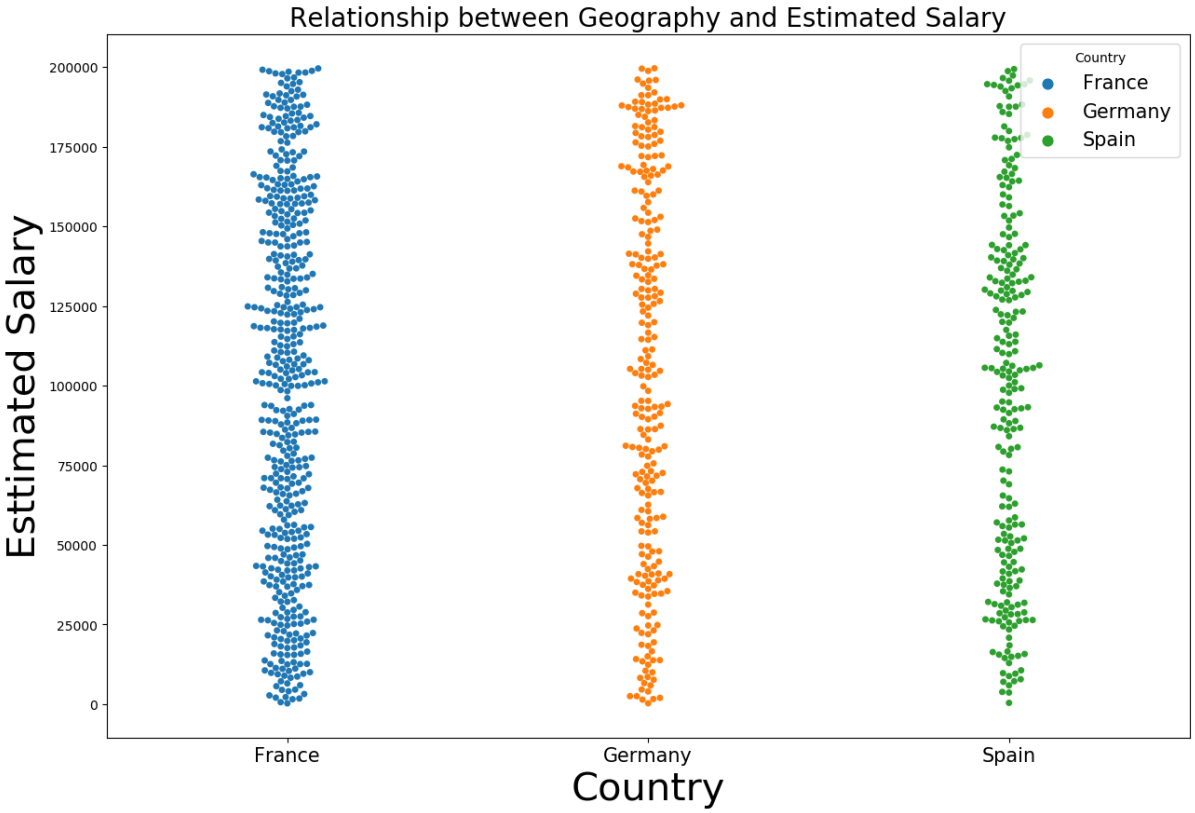


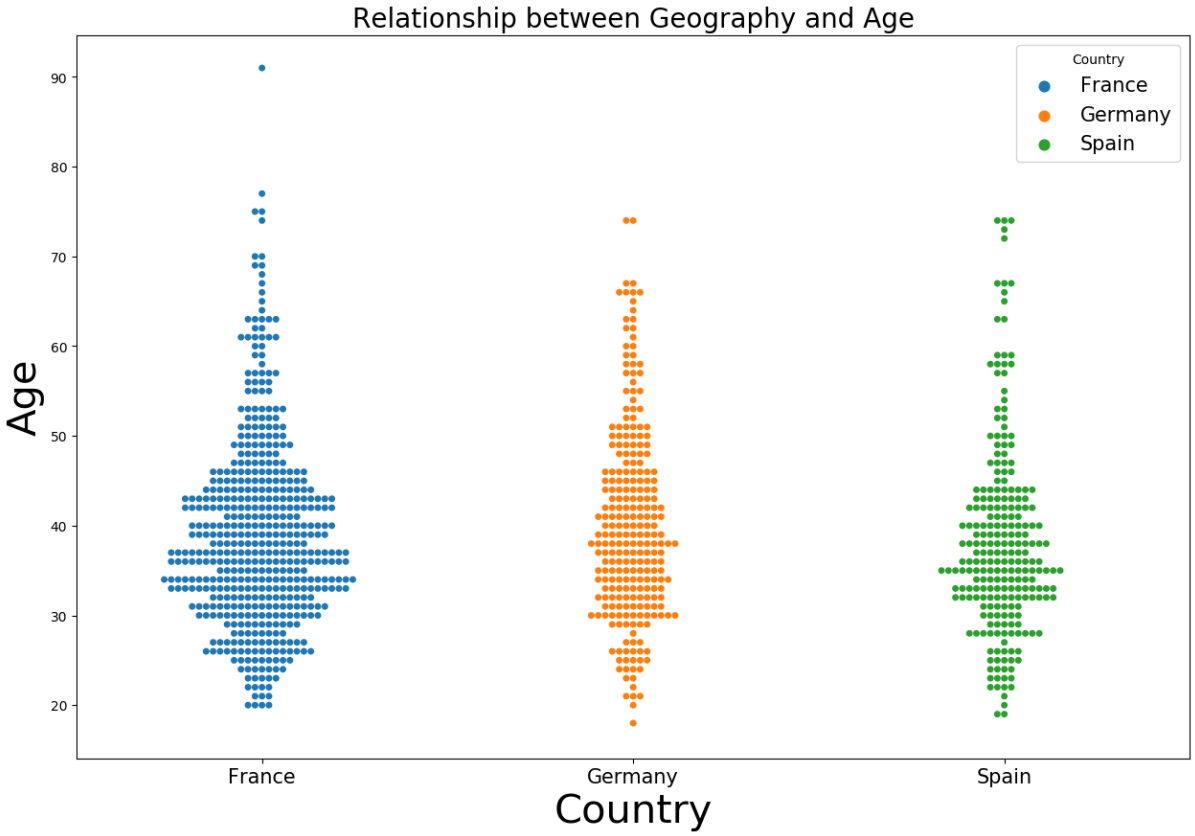
Bivariate Analysis between Categorical and Numerical Variables

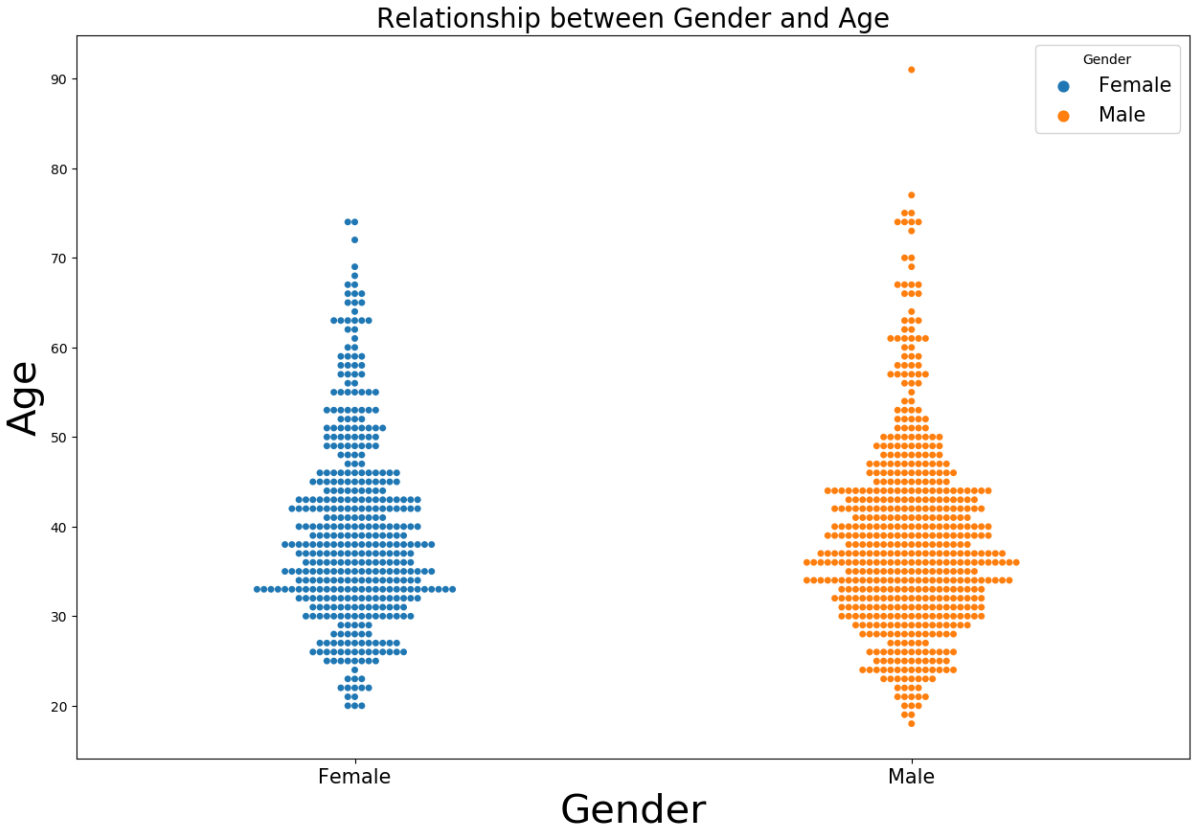

Relationship between Geography and Credit Score
Country
France
Germany
Spain
800
700
600
500
400
Credit Score
France
Germany
Spain
Country


Relationship between Geography and Tenure
Country
France
Germany
Spain
Tenure
0
2
4
6
8
10
France
Germany
Spain
Country

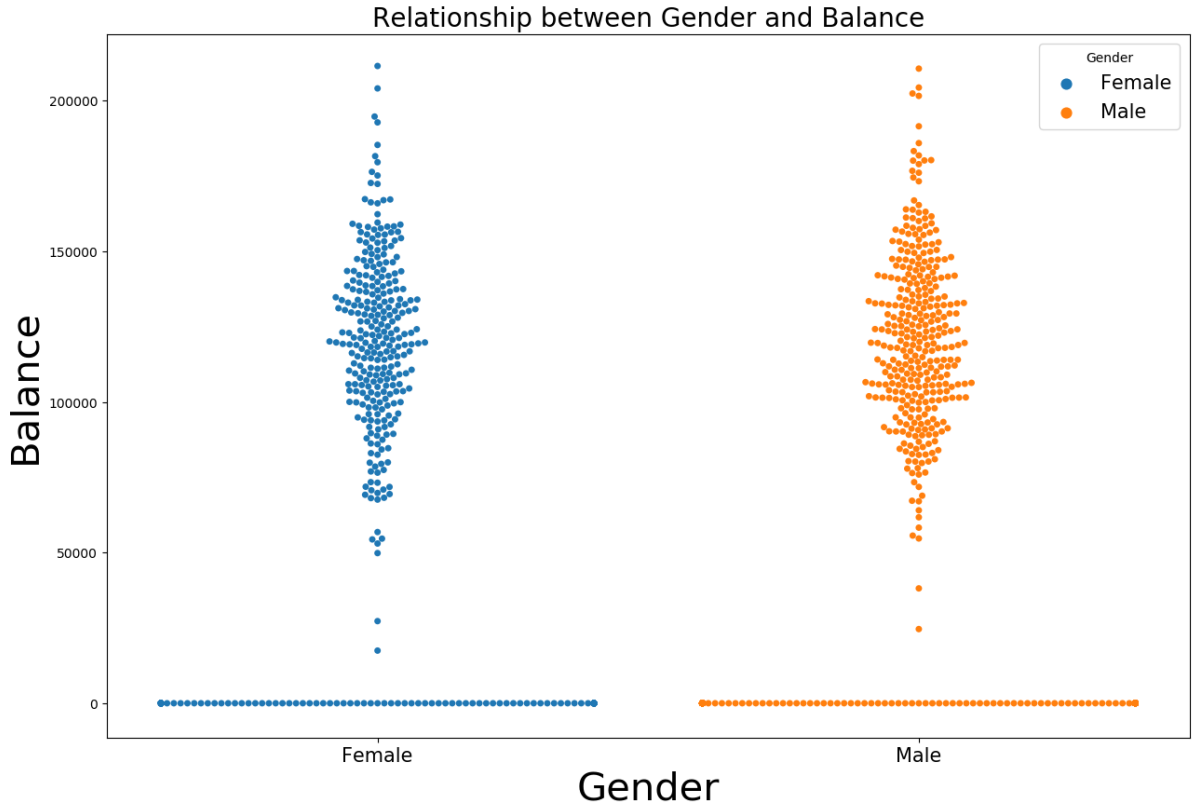

Relationship between Geography and Balance
Country
France
Germany
Spain
Balance
0
50000
100000
150000
200000
France
Germany
Spain
Country

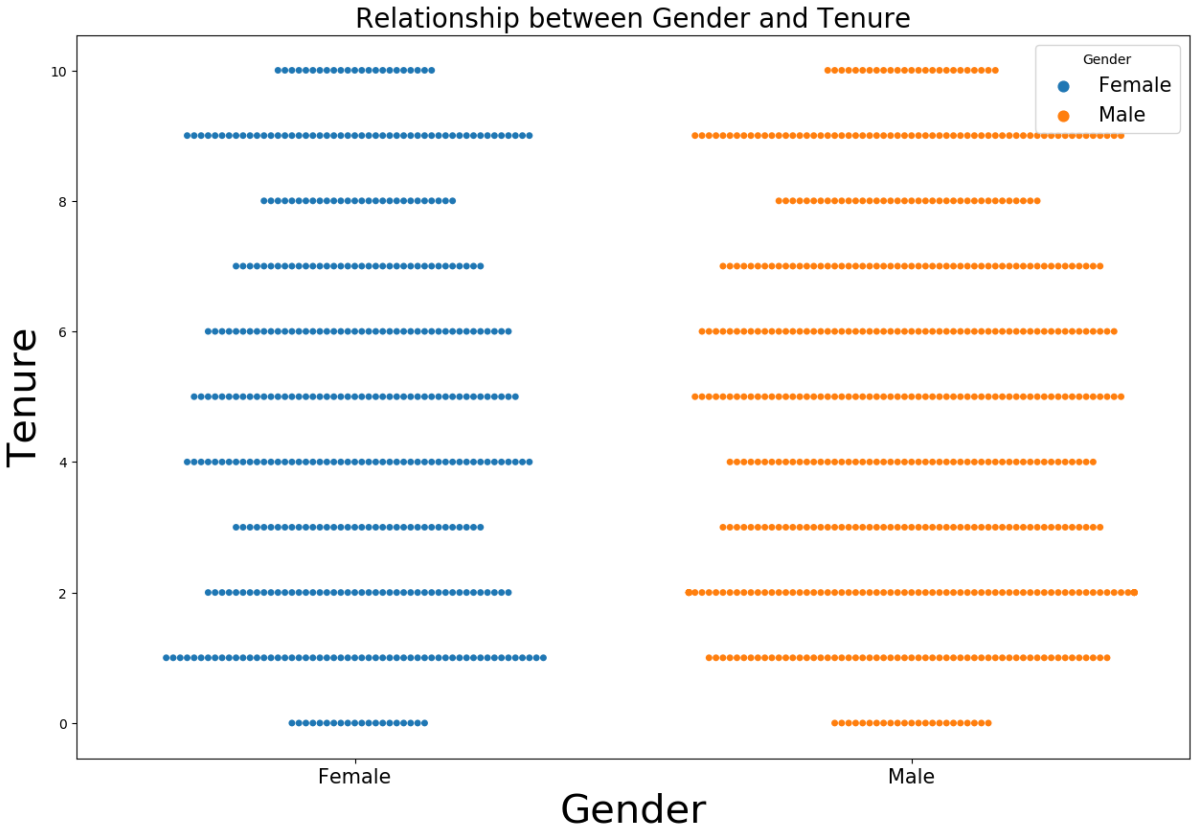

Relationship between Geography and Estimated Salary
Esttimated Salary
200000
175000
150000
125000
100000
75000
50000
25000
0
Country
France
Germany
Spain
France
Germany
Spain
Country

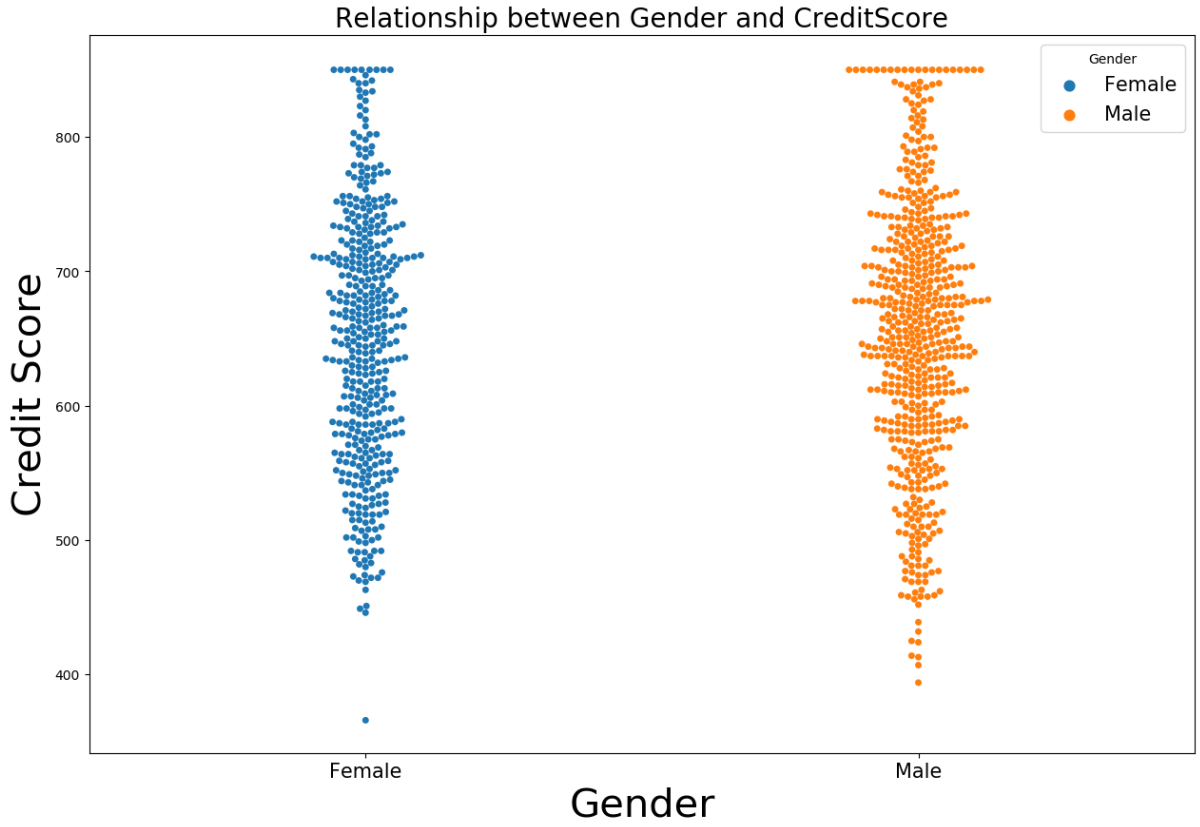

Relationship between Geography and Age
Age
90
80
70
60
50
40
30
20
Country
France
Germany
Spain
France
Germany
Spain
Country

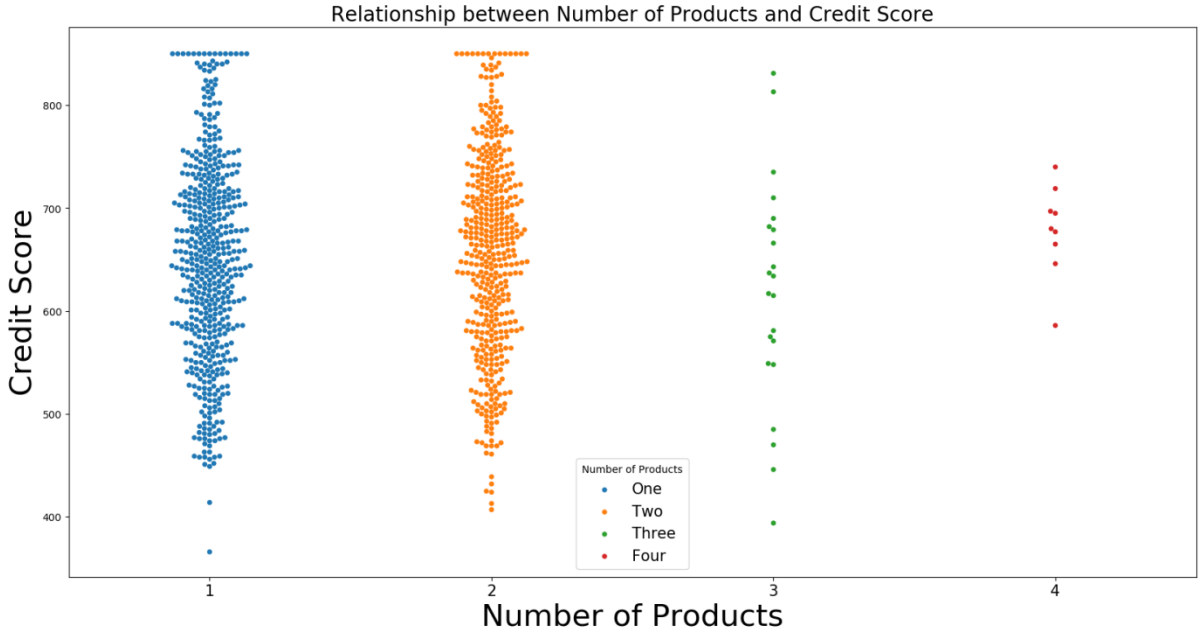

Relationship between Gender and Age
Gender
Female
Male
Age
90
80
70
60
50
40
30
20
Female
Male
Gender

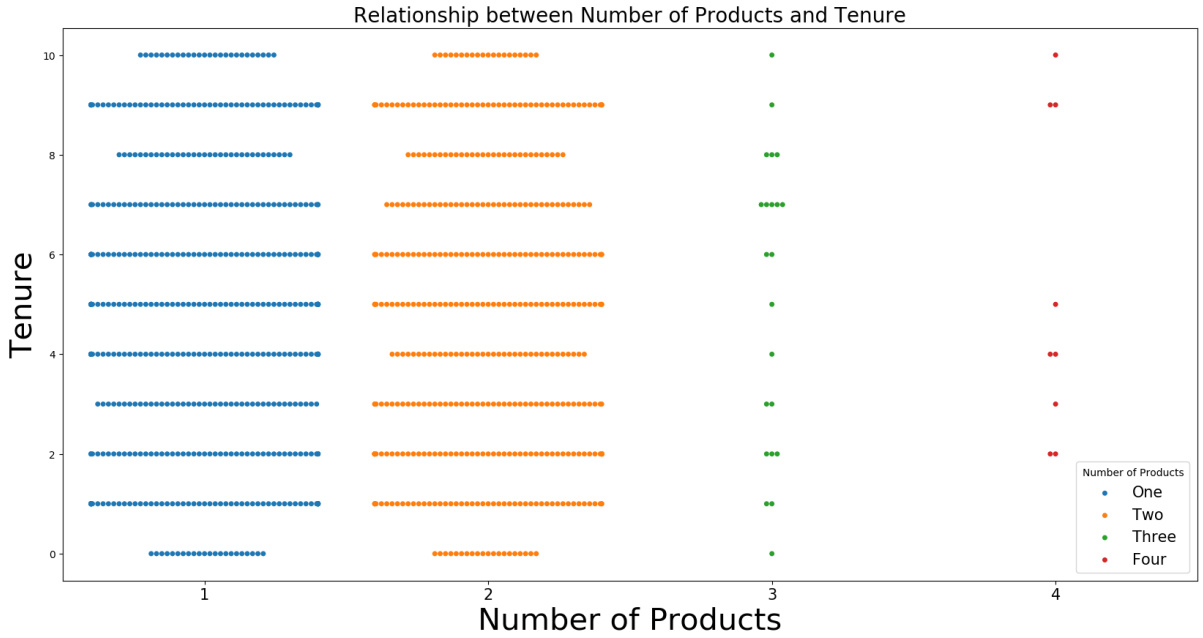

Relationship between Gender and Estimated Salary
Gender
Female
Male
Estimated Salary
200000
175000
150000
125000
100000
75000
50000
25000
0
Female
Male
Gender

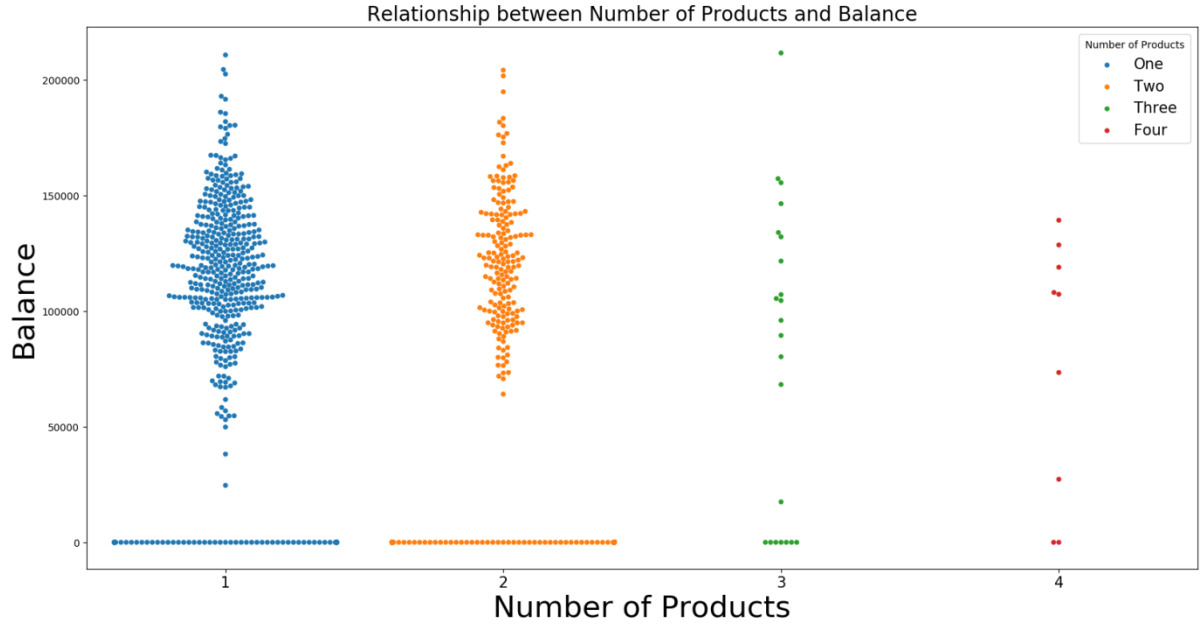

Relationship between Gender and Balance
Gender
Female
Male
200000
150000
100000
50000
0
Balance
Female
Male
Gender

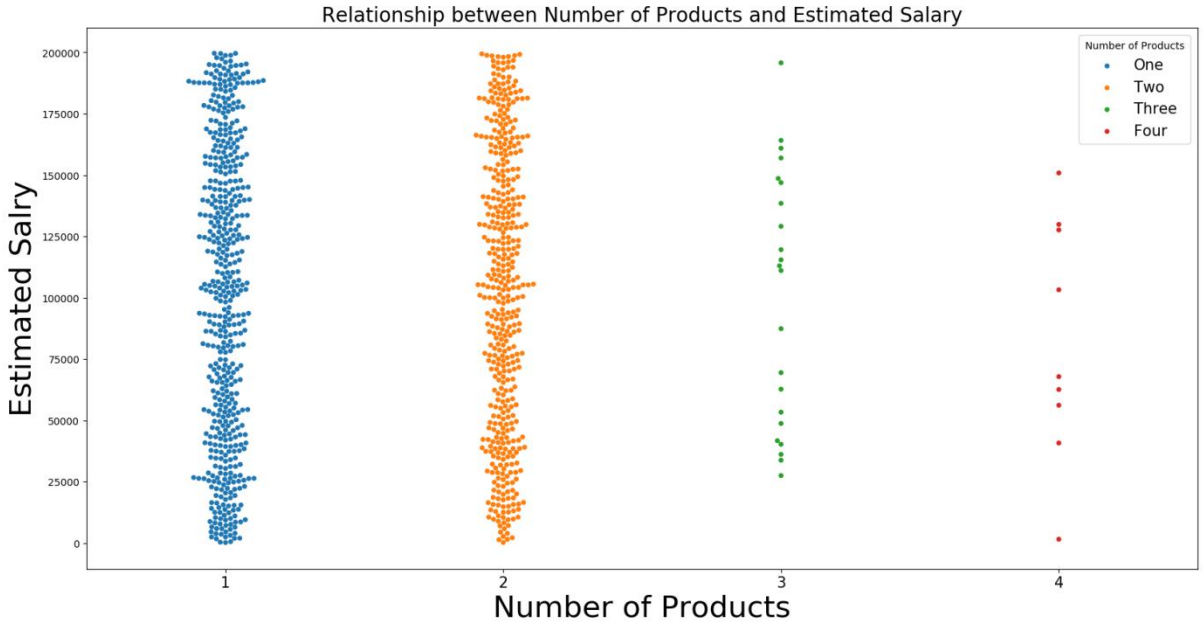

Relationship between Gender and Tenure
Gender
Female
Male
10
8
6
4
2
0
Tenure
Female
Male
Gender


Relationship between Gender and CreditScore
Gender
Female
Male
Credit Score
800
700
600
500
400
Female
Male
Gender

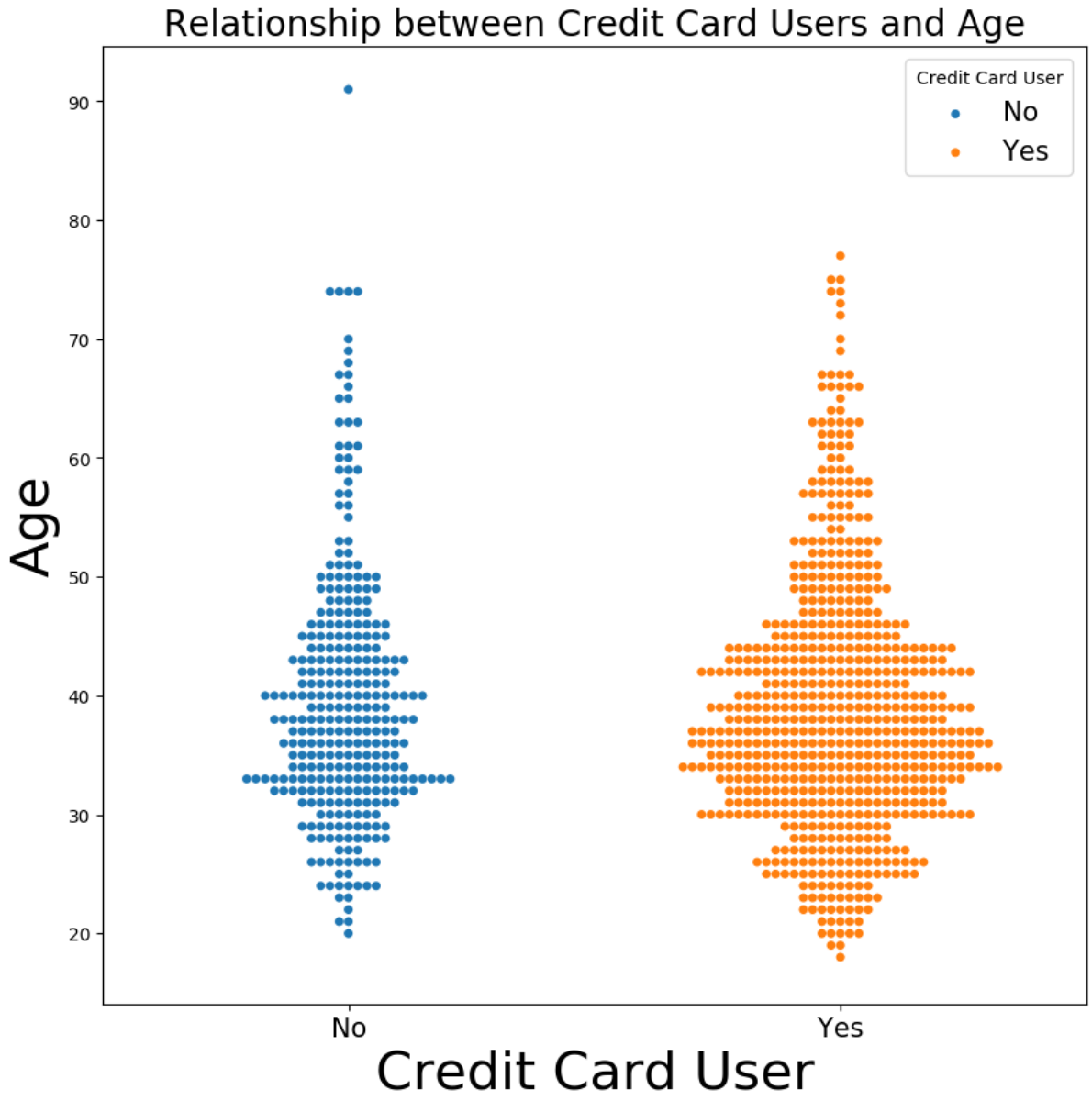

Relationship between Number of Products and Credit Score
Credit Score
800
700
600
500
400
Number of Products
One
Two
Three
Four
1
2
3
4
Number of Products

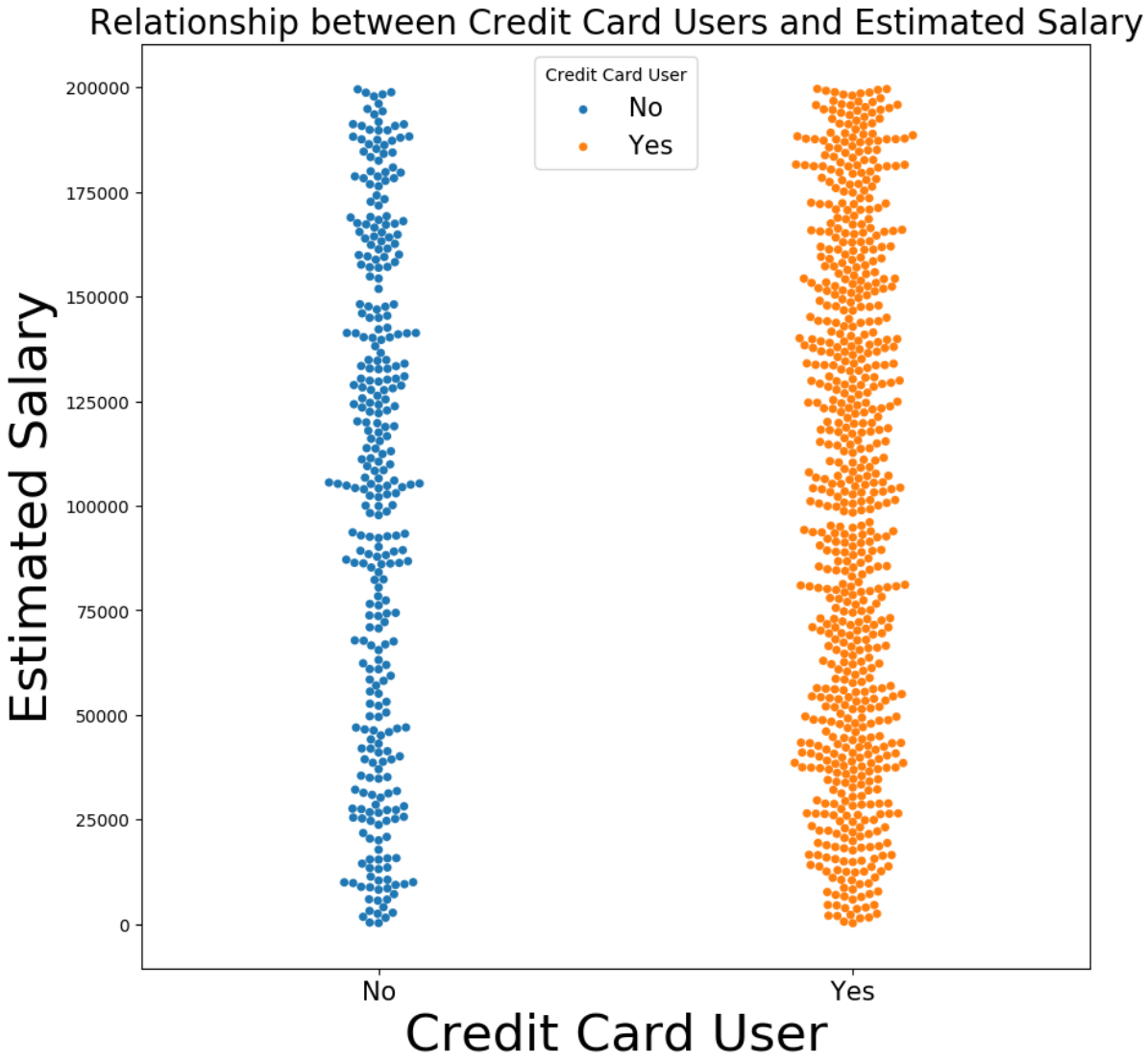

Relationship between Number of Products and Tenure
Tenure
Number of Products
Number of Products
One
Two
Three
Four

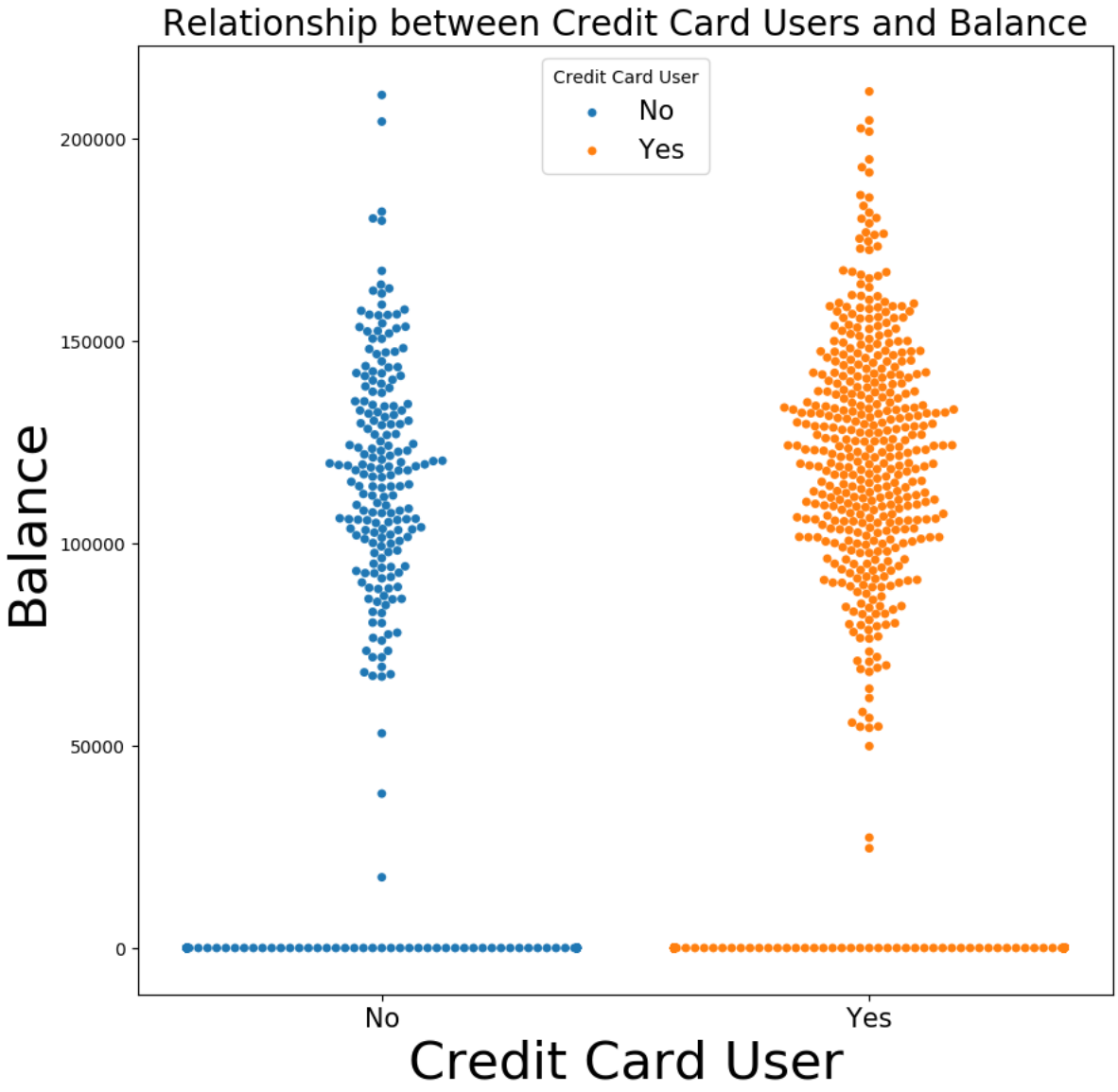

Relationship between Number of Products and Balance
Balance
Number of Products
Number of Products
One
Two
Three
Four

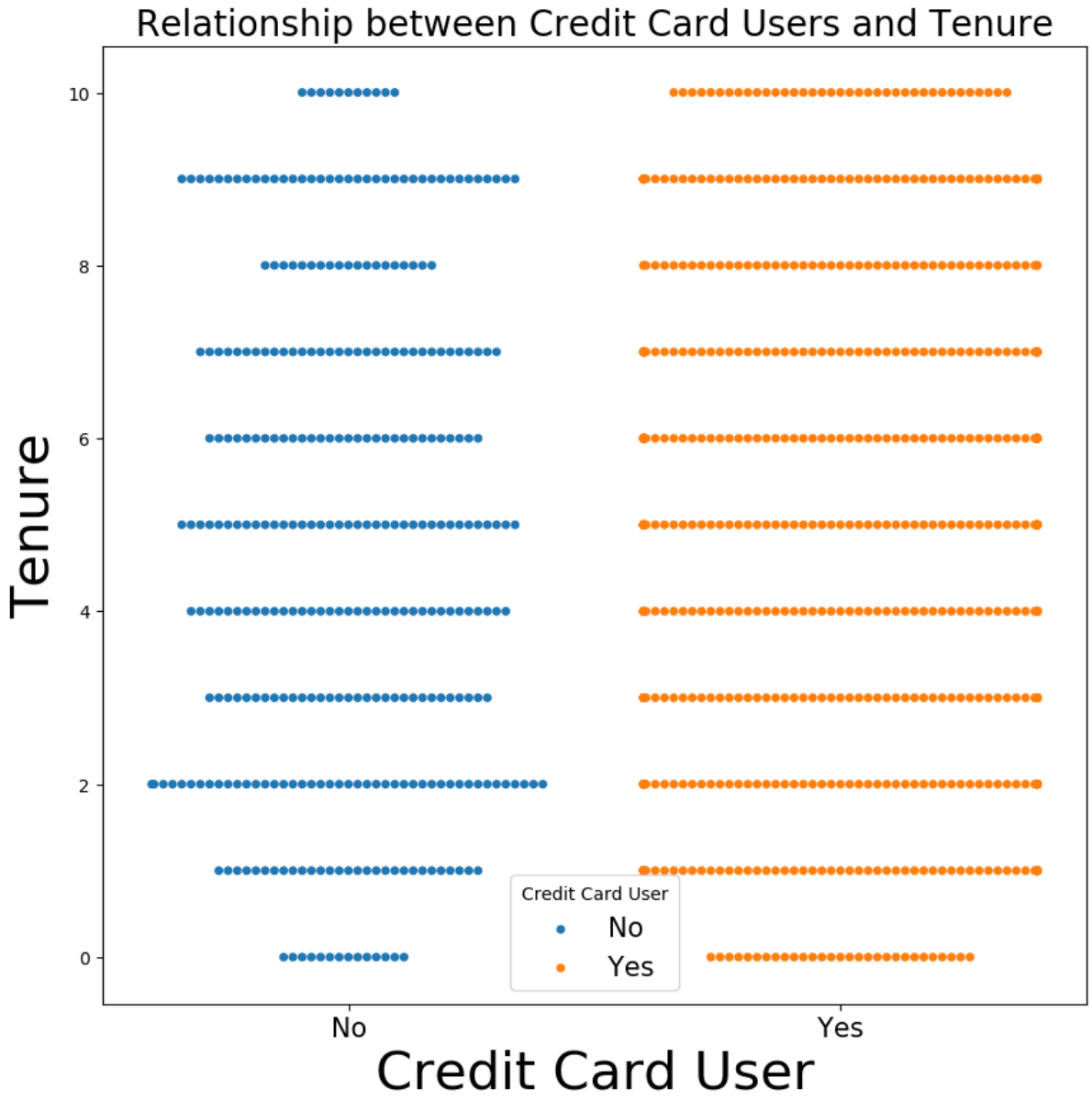

Relationship between Number of Products and Estimated Salary
Estimated Salry
Number of Products
Number of Products
One
Two
Three
Four
200000
175000
150000
125000
100000
75000
50000
25000
0
1
2
3
4

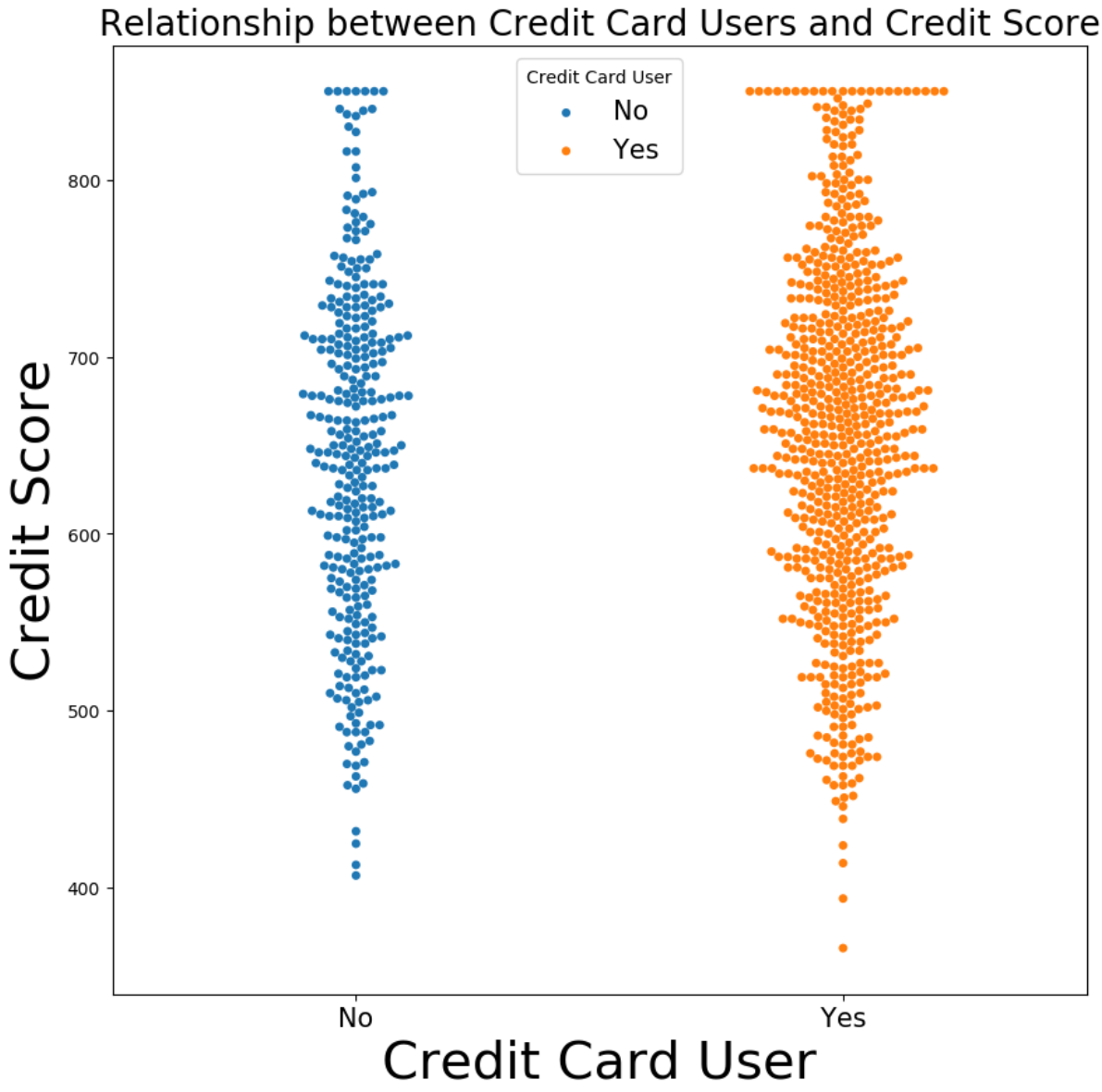

Relationship between Number of Products and Age
Age
Number of Products
Number of Products
One
Two
Three
Four
90
80
70
60
50
40
30
20
1
2
3
4

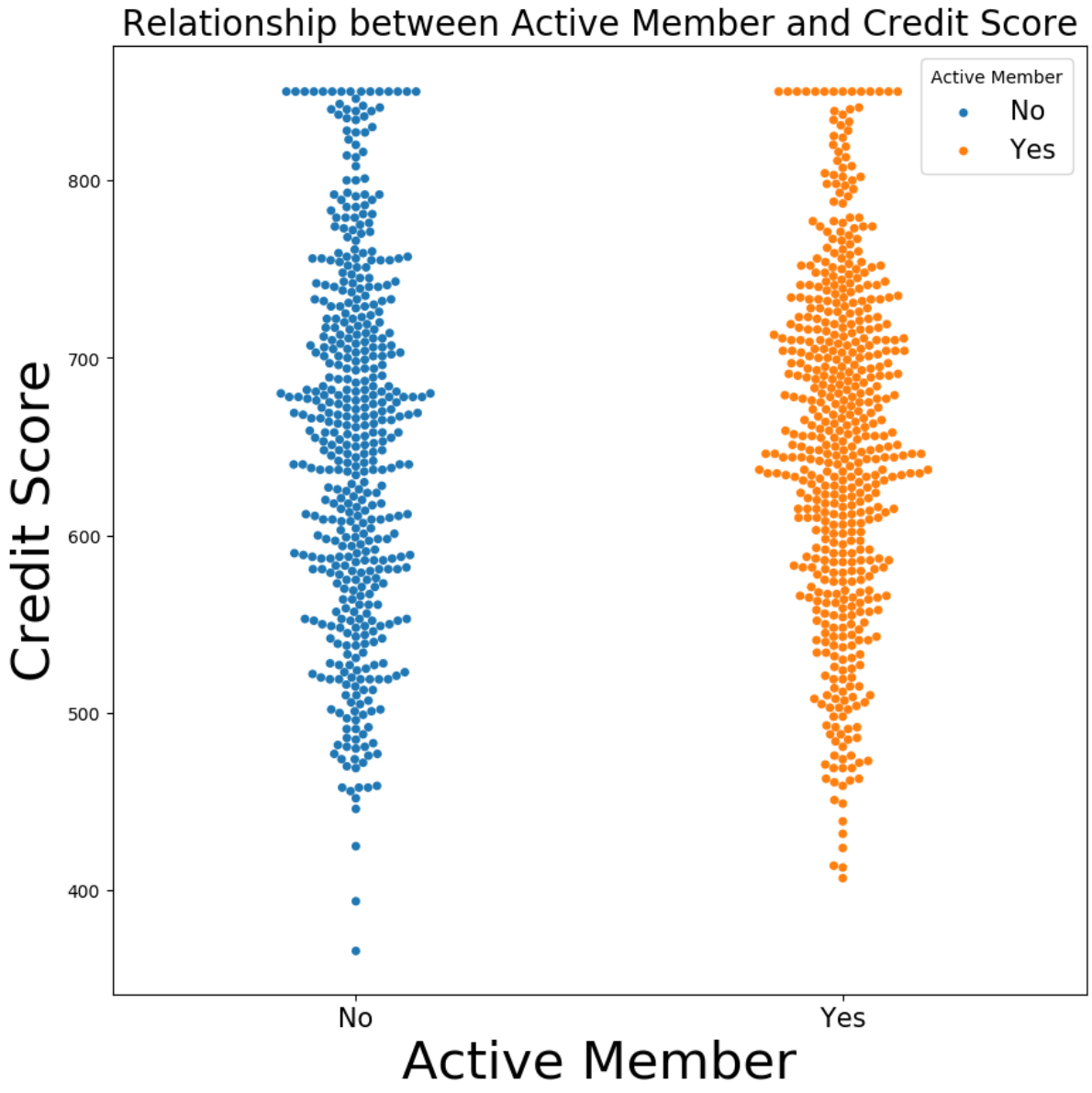

Relationship between Credit Card Users and Age
Credit Card User
No
Yes
Age
90
80
70
60
50
40
30
20
No
Yes
Credit Card User

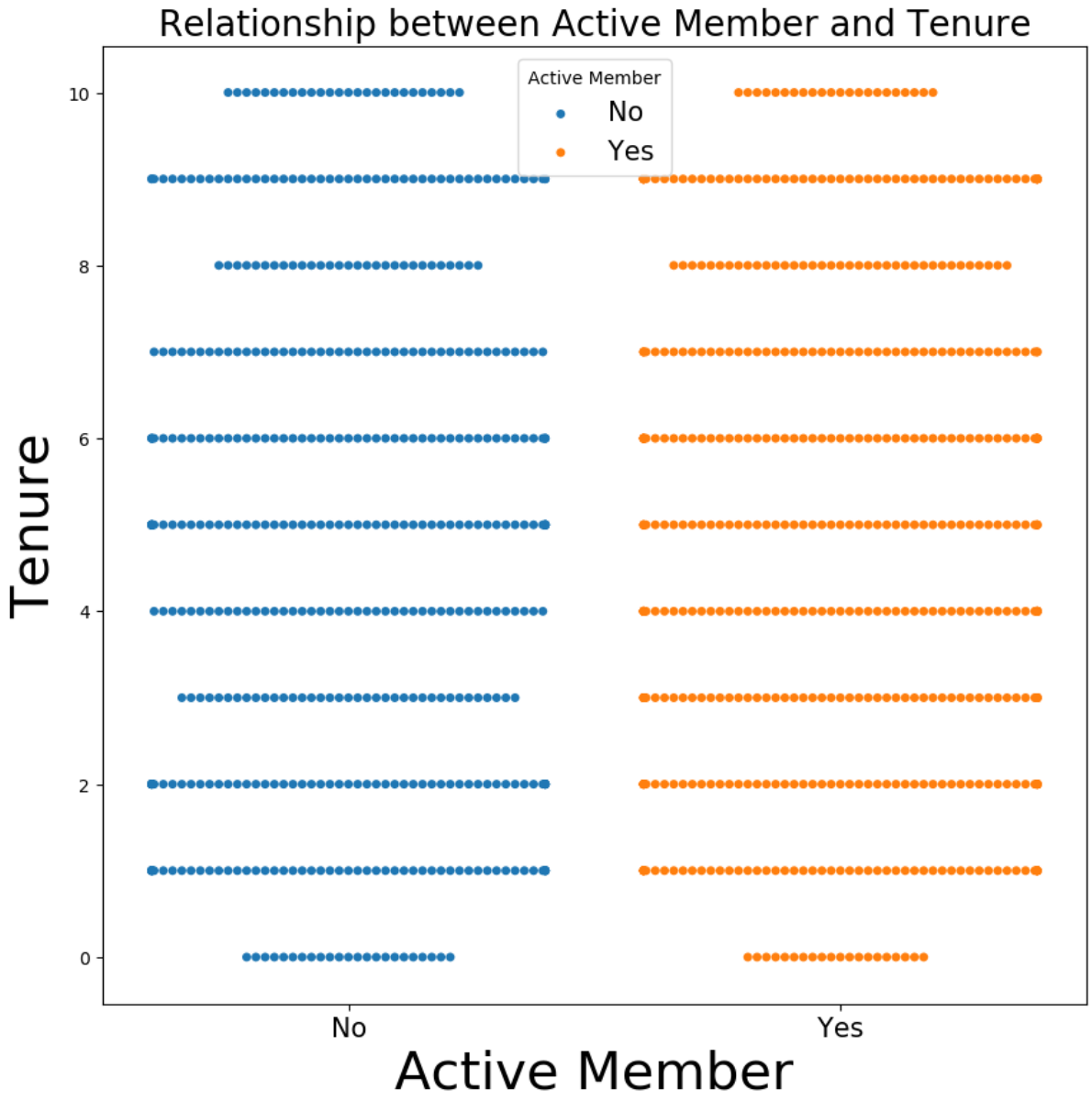

Relationship between Credit Card Users and Estimated Salary
Credit Card User
No
Yes
200000
175000
150000
125000
100000
75000
50000
25000
0
Estimated Salary
No
Yes
Credit Card User


Relationship between Credit Card Users and Balance
Credit Card User
No
Yes
200000
150000
100000
50000
0
Balance
No
Yes
Credit Card User

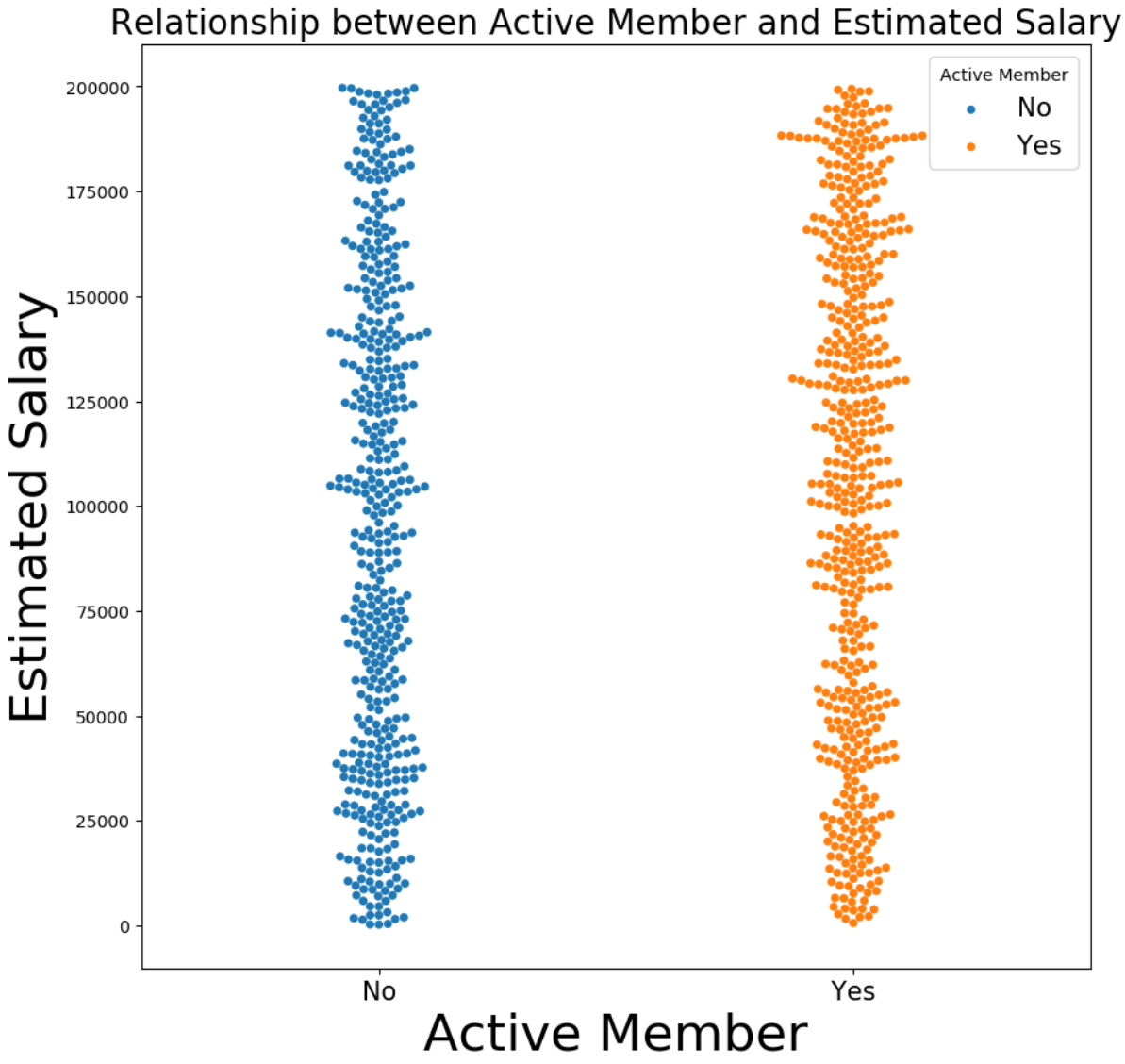

Relationship between Credit Card Users and Tenure
Tenure
10
8
6
4
2
0
No
Yes
Credit Card User
Credit Card User
No
Yes

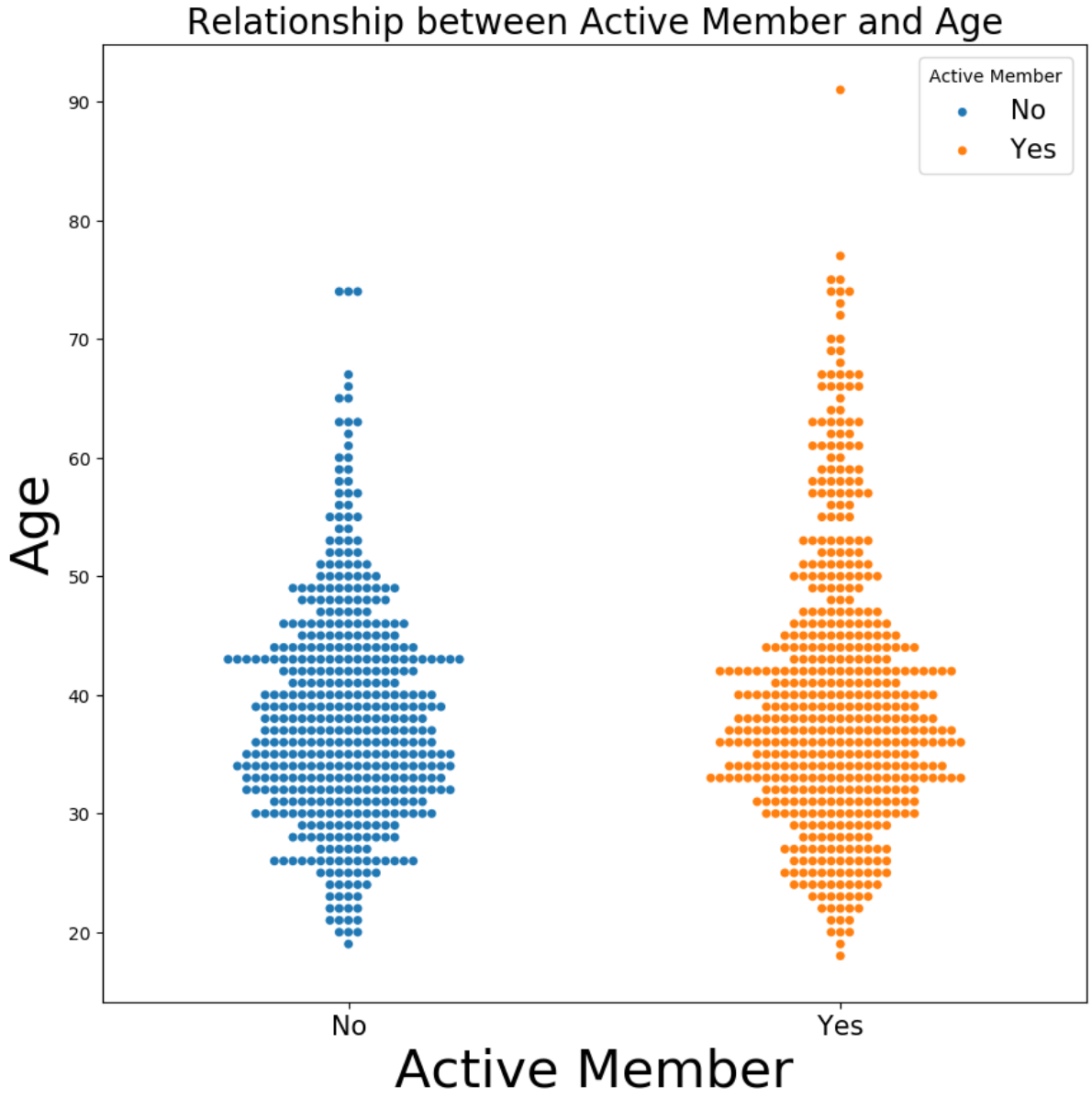

Relationship between Credit Card Users and Credit Score
Credit Card User
No
Yes
800
700
600
500
400
Credit Score
No
Yes
Credit Card User


Relationship between Active Member and Credit Score
Active Member
No
Yes
800
700
600
500
400
Credit Score
No
Yes
Active Member


Relationship between Active Member and Tenure
Active Member
No
Yes
Tenure
10
8
6
4
2
0
No
Yes
Active Member


Relationship between Active Member and Balance
Active Member
No
Yes
Balance
200000
150000
100000
50000
0
No
Yes
Active Member


Relationship between Active Member and Estimated Salary
Active Member
No
Yes
Estimated Salary
200000
175000
150000
125000
100000
75000
50000
25000
0
No
Yes
Active Member


Relationship between Active Member and Age
Active Member
No
Yes
Age
90
80
70
60
50
40
30
20
No
Yes
Active Member

Missing Values

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Missing data can occur because of nonresponse: no information is provided for one or more items or for a whole unit ("subject"). Some items are more likely to generate a nonresponse than others.

In our dataset, there is no missing value present. A table below will throw light on our dataset:-

RowNumber	0
CustomerId	0
Surname	0
CreditScore	0
Geography	0
Gender	0
Age	0
Tenure	0
Balance	0
NumOfProducts	0
HasCrCard	0
IsActiveMember	0
EstimatedSalary	0
Exited	0

Outliers

An Outlier is a rare chance of occurrence within a given data set. In Statistics and Data Science, an Outlier is an observation point that is distant from other observations. An Outlier may be due to variability in the measurement or it may indicate experimental error.

Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

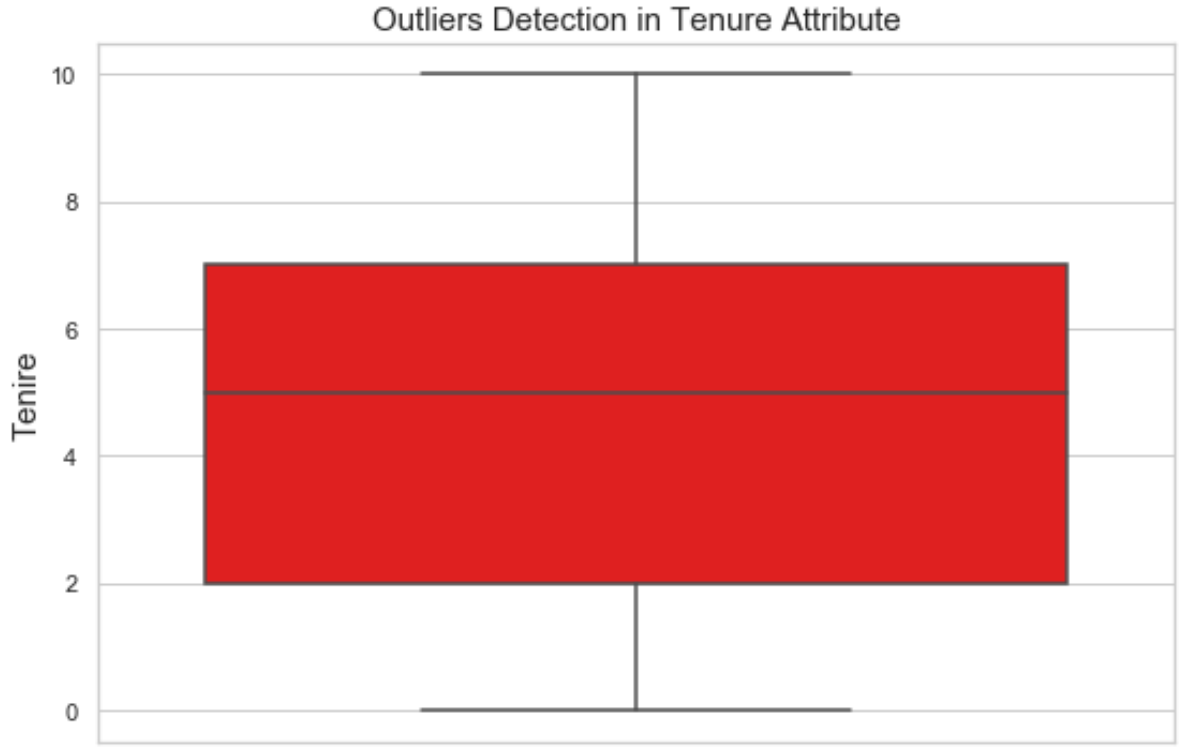
While outliers are attributed to a rare chance and may not necessarily be fully explainable, Outliers in data can distort predictions and affect the accuracy, if you don't detect and handle them.

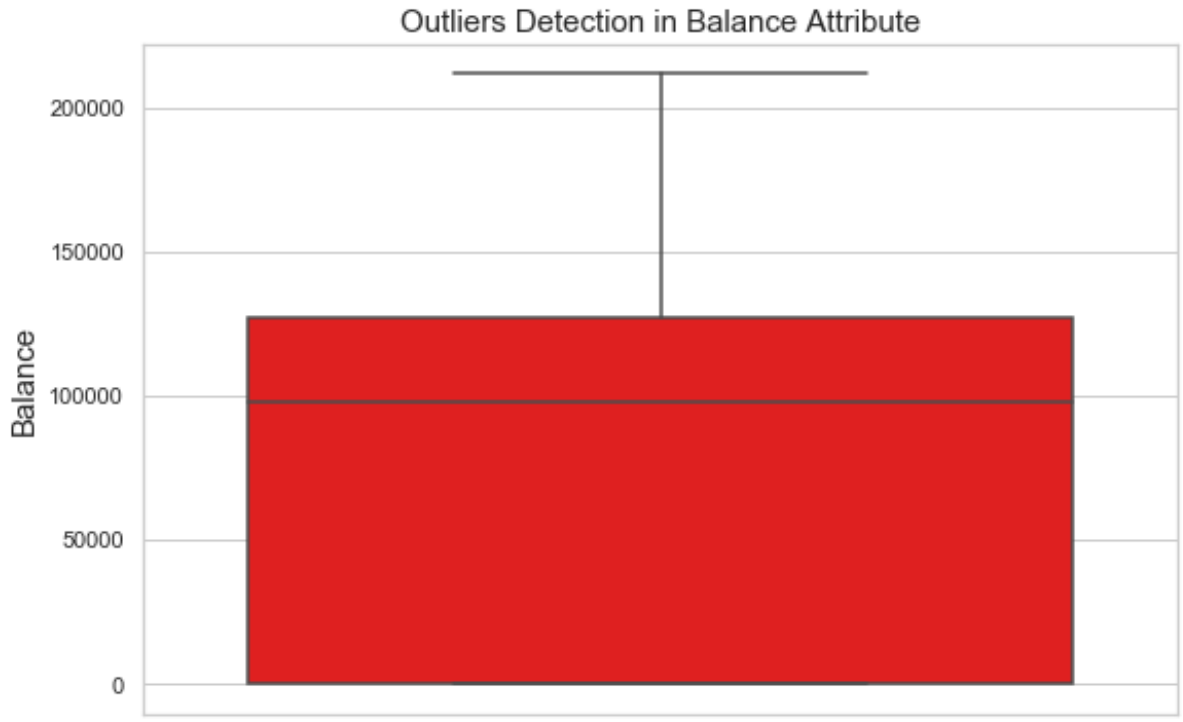
The contentious decision to consider or discard an outlier needs to be taken at the time of building the model. Outliers can drastically bias/change the fit estimates and predictions.

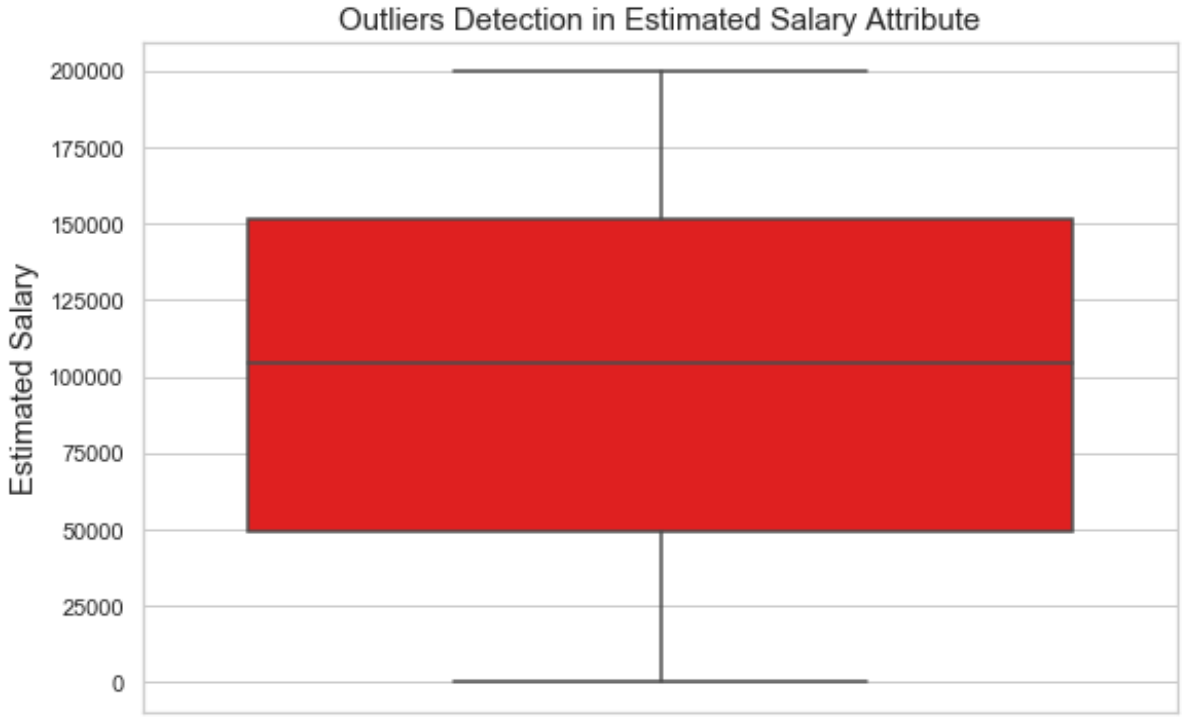
Detecting and Removing Outliers

Mostly outliers are present in the continuous variables and box plot method is best and easy way to detect and remove outliers. Moreover, our dataset contains categorical variables that are encoded so we will perform outlier detections only on continuous variables.


Outliers Detection in Age Attribute
Age
90
80
70
60
50
40
30
20


Outliers Detection in Tenure Attribute
Tenire
10
8
6
4
2
0


Outliers Detection in Balance Attribute
Balance
200000
150000
100000
50000
0


Outliers Detection in Estimated Salary Attribute
Estimated Salary
200000
175000
150000
125000
100000
75000
50000
25000
0

Feature Selection

Machine learning works on a simple rule – if we put garbage in, we will only get garbage to come out.

This becomes even more important when the number of features is very large. We need not use every feature at our disposal for creating an algorithm. We can assist our algorithm by feeding in only those features that are really important. Feature subsets giving better results than complete set of feature for the same algorithm or – "Sometimes, less is better!".

We should consider the selection of feature for model based on below criteria:-

1. The relationship between two independent variable should be less and
2. The relationship between Independent and target variables should be high.

Below figure shows a graphical display of a correlation matrix, called a correlogram. The cells of the matrix are coloured to show the correlation value.

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	EstimatedSalary
RowNumber	1.0	-0.012	0.013	0.081	-0.046	-0.0086	0.071
CustomerId	-0.012	1.0	-0.014	-0.014	0.028	0.051	-0.034
CreditScore	0.013	-0.014	1.0	0.022	-0.01	-0.0017	0.041
Age	0.081	-0.014	0.022	1.0	-0.025	0.054	0.03
Tenure	-0.046	0.028	-0.01	-0.025	1.0	0.029	0.016
Balance	-0.0086	0.051	-0.0017	0.054	0.029	1.0	0.047
EstimatedSalary	0.071	-0.034	0.041	0.03	0.016	0.047	1.0

Feature Scaling

Feature scaling is done to reduce unwanted variation either within or between variables and to bring all of the variables into proportion with one another. I will use Normalization process to perform feature scaling. Formula for Normalization is given below:-

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Attribute before and after normalization are given below:-

CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
0.375000	France	Male	0.317073	0.1	0.000000	1	0	1	0.101298
0.383772	France	Male	0.390244	0.4	0.000000	1	0	1	0.019060
0.603070	France	Female	0.048780	0.7	0.000000	2	1	0	0.644939
0.657895	France	Male	0.512195	0.4	0.819096	1	1	1	0.406481
0.241228	Spain	Male	0.243902	1.0	0.516695	1	1	1	0.939612

Modeling

Model Selection

In the case of this dataset we have to predict whether customer has left the bank or not. The target variable here is a categorical variable and for a categorical variable we can use various Classification models. Trained model having less error rate and more accuracy will be our final model. Different machine learning methods which will be used to train our final model are mentioned below:-

1. Decision Tree Classification Model
2. Random Forest Model
3. Logistic Regression Model
4. KNN Model
5. Naïve Bayes Model

Final model with will be with the higher accuracy which we will able to decide at the end of the modelling process.

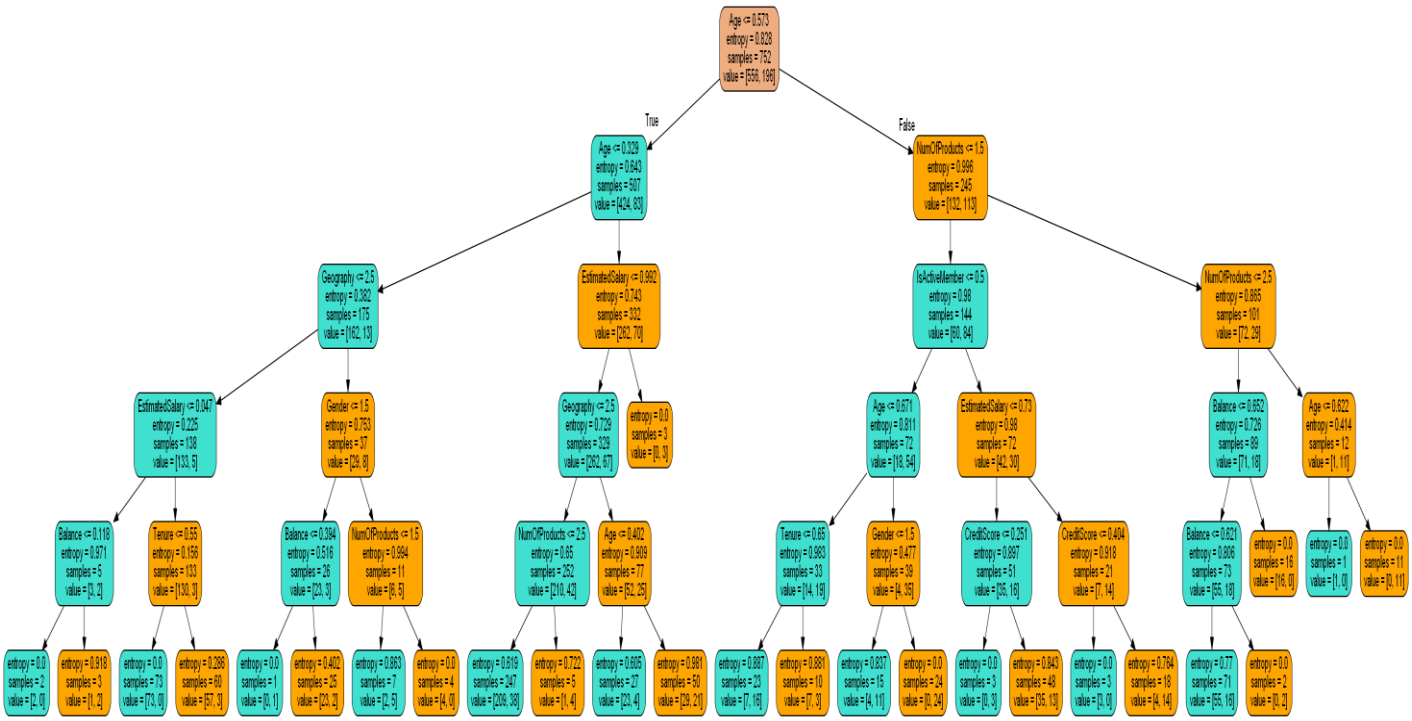
Decision Tree Classification Model

Decision tree builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Trained Decision tree are shown below:-

```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=5,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best')
```

Graphical Visualization of a trained modal is shown below:-



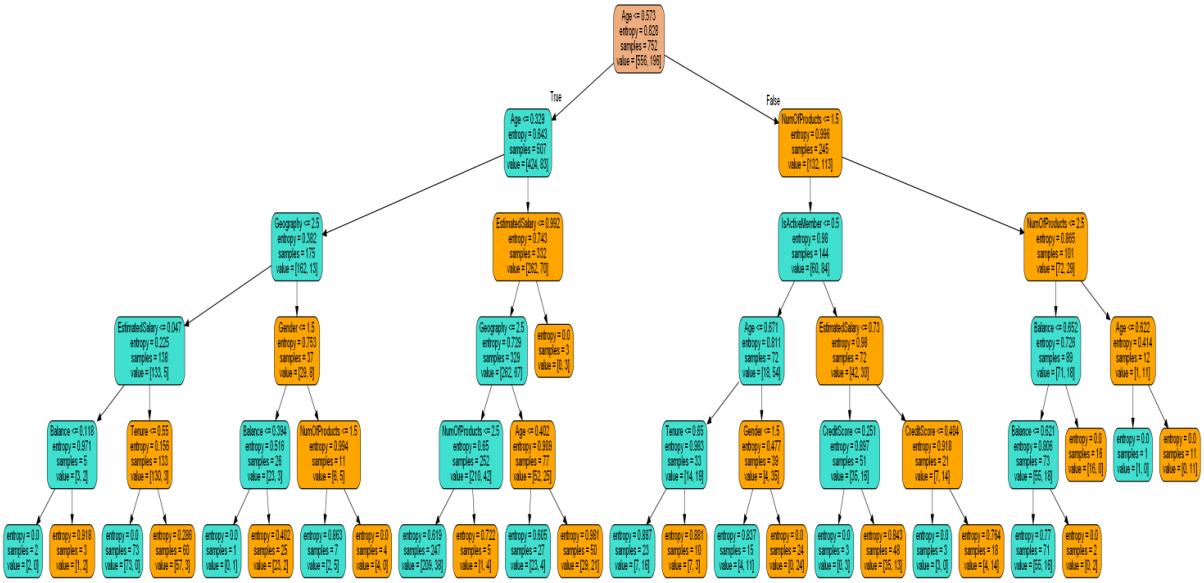
Random Forest Model

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks.

Trained Random forest classification model is shown below:-

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=5, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=1,
            oob_score=False, random_state=0, verbose=0, warm_start=False)
```

Visualization of Trained Random Forest is shown below:-



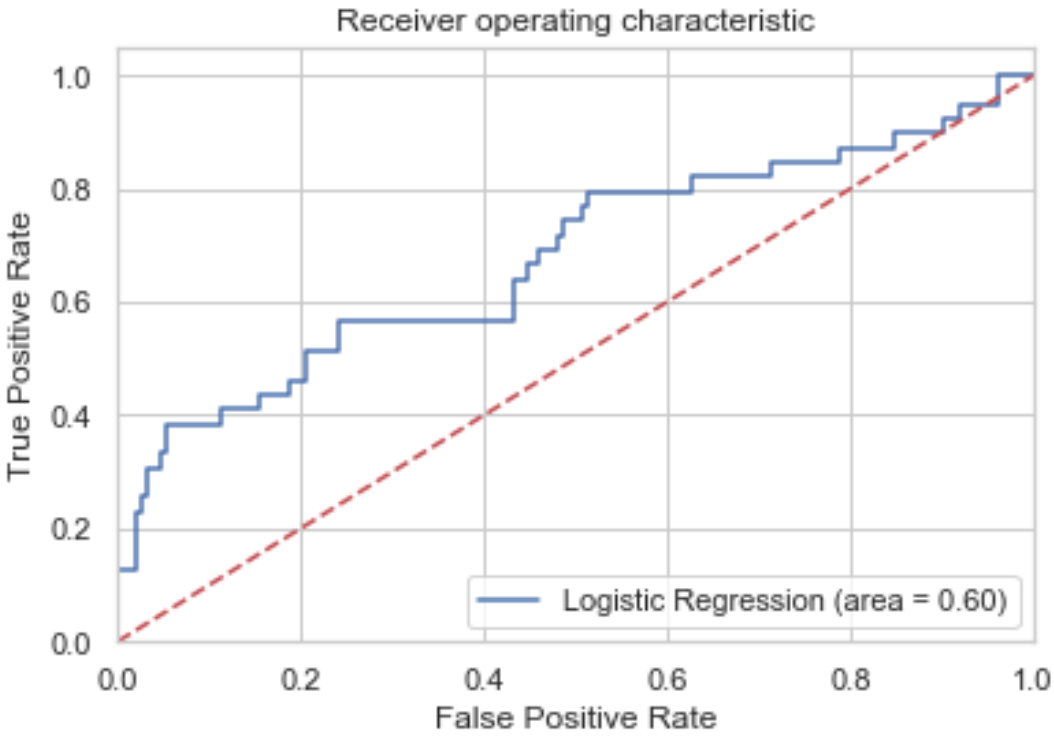
Logistic Regression Model

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

- A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1).

- Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

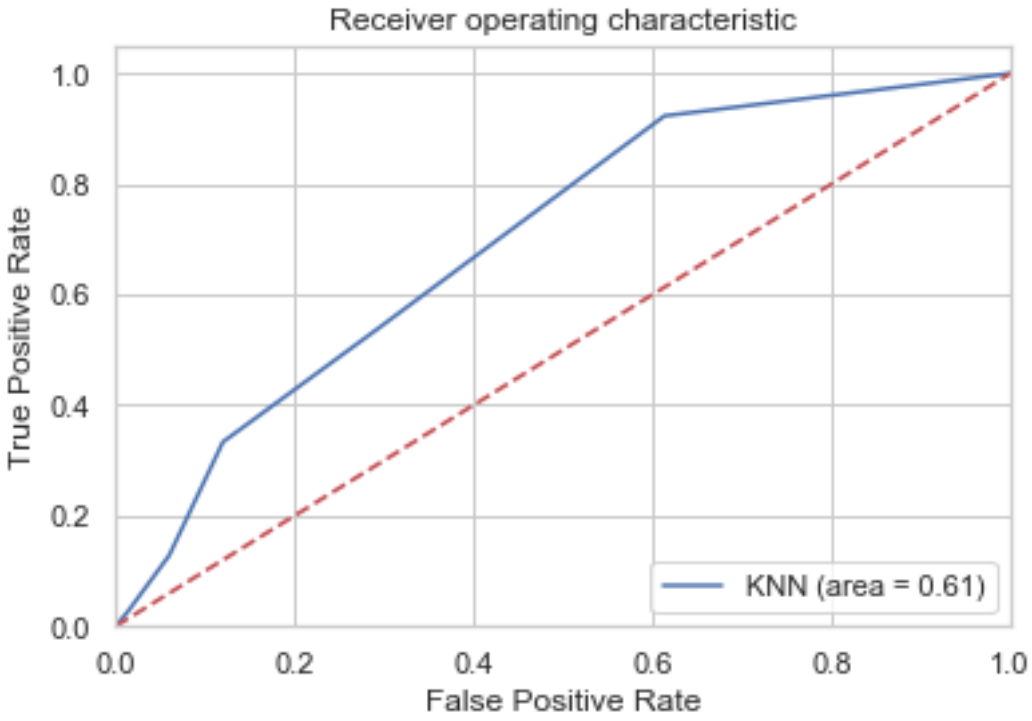
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=0, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```



KNN Model

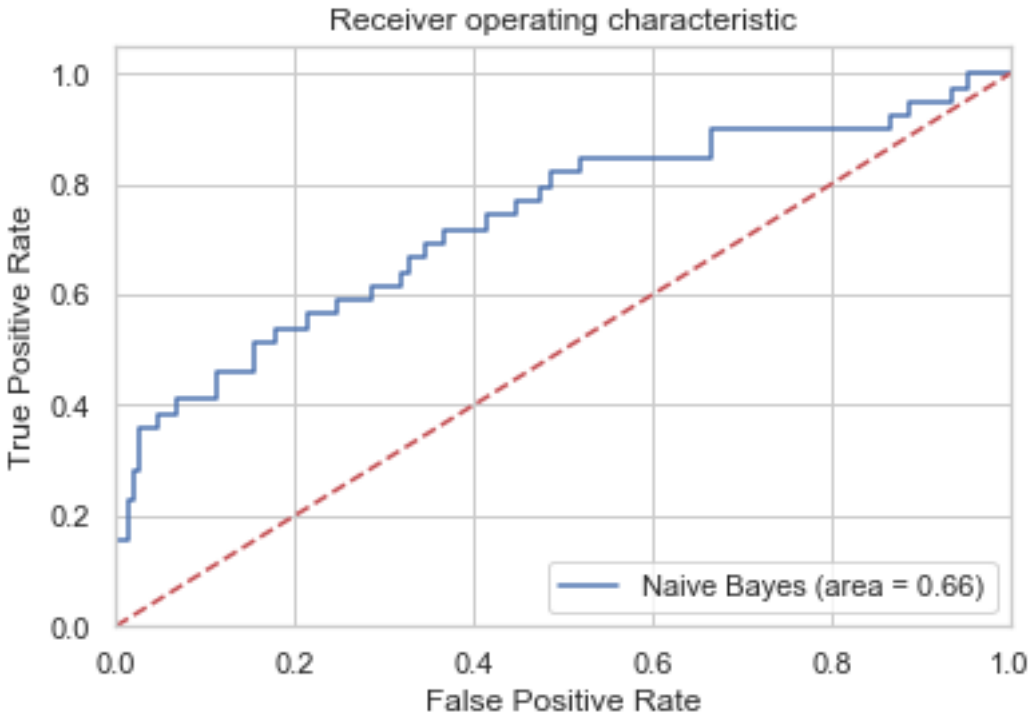
In pattern recognition, the k-nearest neighbour's algorithm is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space.

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=1, n_neighbors=5, p=2,
           weights='uniform')
```



Naive Bayes Model

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.



Accuracy of all the Trained Models

S.No	Model	Accuracy
1	Decision Tree Classification Model	81.48148148148148
2	Random Forest Model	86.24338624338624
3	Logistic Model	82.01058201058201
4	KNN Model	76.71957671957672
5	Naïve Bayes Model	83.06878306878306

Random Forest Model is giving Best Accuracy in this type of Business Problem. Hence it will be our final model.