Python for Programming

VENKATA SAI TEJA GANGUMALLA| POOJA UDAYANJALI KANNURI

Final Project

Bank Customer Direct Marketing Campaign Analysis

Date: 10/13/2023

1 Project Overview

The Bank Customer Direct Marketing Campaign Analysis project aims to provide data-driven insights to a bank's marketing team in order to optimize their direct marketing campaign. The bank is interested in understanding customer behavior and factors contributing to churn (i.e., customers leaving the bank) in order to design more effective marketing strategies and improve customer retention.

2 Dataset

2.1 Dataset Overview

"Bank Customer Churn Prediction" (linked here) presents a comprehensive collection of customer-related information within the context of a banking institution. It is structured to distinguish customers who have engaged in churn (terminated their relationship with the bank) from those who have exhibited retention (maintained their affiliation). The primary aim of this dataset is to facilitate the development of predictive models that can discern the proclivity of customers to engage in churn, thereby enabling the bank to implement preemptive strategies for customer retention.

2.2 Dataset Contents:

- CustomerId: A unique identifier for each customer.
- Surname: The customer's surname.
- CreditScore: The customer's credit score.
- Geography: The customer's country of residence.
- Gender: The customer's gender.
- Age: The customer's age.
- Tenure: The number of years the customer has been a customer of the bank.
- Balance: The customer's account balance.
- NumOfProducts: The number of products the customer has with the bank.
- HasCrCard: Whether or not the customer has a credit card with the bank.
- IsActiveMember: Whether or not the customer is an active member of the bank.
- EstimatedSalary: The customer's estimated salary.
- Churn: Whether or not the customer has churned

2.3 Dataset Utility:

This dataset serves as a pivotal resource for the development of machine learning models that can effectively predict customer churn. By leveraging the information contained within, the bank can implement targeted retention initiatives. These initiatives, in turn, can lead to a reduction in customer churn, thereby optimizing customer relationships and preserving revenue streams.

3 Comprehensive Data Analysis and Insights

3.1 Data Collection and Preparation

Data Collection and Preparation: Gather data related to customer information, their interactions with the bank, and campaign outcomes. Clean and preprocess the data to ensure its quality and readiness for analysis.

3.1.1 Data Reading and Overview

Reading the data from a CSV file into a Pandas DataFrame. The sep=',' parameter specifies that the data is comma-separated. df.head() displays the first few rows of the DataFrame, and df.describe() provides summary statistics for the numerical columns, such as count, mean, and standard deviation. df.shape shows the number of rows and columns in the dataset.

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedS
0	42	15738148	Clarke	465.0	France	Female	51.0	8.0	122522.32	1.0	0.0	0.0	1812
1	44	15755196	Lavine	834.0	France	Female	49.0	2.0	131394.56	1.0	0.0	0.0	1943
2	53	15683553	O'Brien	788.0	Fr	F	33.0	5.0	0.00	2.0	0.0	0.0	1169
3	81	15706021	Buley	NaN	Fr	F	34.0	1.0	96645.54	2.0	0.0	0.0	1714
4	83	15641732	Mills	NaN	Fr	F	36.0	NaN	0.00	2.0	0.0	0.0	260

3.1.2 Splitting Data into Numerical and Categorical Columns

In this part, we separate the columns of the dataset into numerical and categorical columns for easier analysis.

num_cols = ['CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'EstimatedSalary']

cat_cols = ['Geography', 'Gender', 'HasCrCard', 'IsActiveMember', 'Exited']

3.1.3 Missing Value Handling

In this section, we first replace any 'NULL' values with NaN (missing values). Code then attempts to convert the columns in **num_cols** to the appropriate data types (int or float).

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
count	10000.00000	1.000000e+04	9848.000000	9637.000000	9978.000000	9990.000000	9943.000000	9918.000000	9911.000000	10000.000000
mean	5000.50000	1.569094e+07	650.462530	38.919685	5.015033	76428.280420	1.530725	0.728776	0.515589	100089.748400
std	2886.89568	7.193619e+04	96.593717	10.471967	2.891540	62389.725467	0.581721	0.545999	0.499782	57510.491042
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.000000	0.000000	11.000000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.000000	0.000000	51001.750000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97089.635000	1.000000	1.000000	1.000000	100193.500000
75%	7500.25000	1.575323e+07	717.000000	44.000000	7.750000	127603.687500	2.000000	1.000000	1.000000	149387.750000
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	5.000000	1.000000	199992.000000

3.1.4 Missing Value Analysis

This part defines a function missing_values that calculates and presents the number and percentage of missing values in each column of the dataset. The function returns a DataFrame with this information.

The output of this function provides insights into the extent of missing data in different columns.]

	Variable Name	Number of Missing Values	% of Missing Values
6	Age	363	3.63
3	CreditScore	152	1.52
11	IsActiveMember	89	0.89
10	HasCrCard	82	0.82
9	NumOfProducts	57	0.57
5	Gender	28	0.28
7	Tenure	22	0.22
8	Balance	10	0.10
0	RowNumber	0	0.00
1	CustomerId	0	0.00
2	Surname	0	0.00
4	Geography	0	0.00
12	EstimatedSalary	0	0.00
13	Exited	0	0.00

3.1.5 Missing Value Imputation

Here we address missing value imputation. For categorical variables (e.g., 'Gender', 'HasCrCard', 'IsActiveMember'), missing values are filled with the mode (most common value) of the respective column. For numerical variables (e.g., 'Age', 'CreditScore', 'NumOfProducts', 'Tenure', 'Balance'), missing values are filled with the median of the respective column.

	Variable Name	Number of Missing Values	% of Missing Values
0	RowNumber	0	0.0
1	CustomerId	0	0.0
2	Surname	0	0.0
3	CreditScore	0	0.0
4	Geography	0	0.0
5	Gender	0	0.0
6	Age	0	0.0
7	Tenure	0	0.0
8	Balance	0	0.0
9	NumOfProducts	0	0.0
10	HasCrCard	0	0.0
11	IsActiveMember	0	0.0
12	EstimatedSalary	0	0.0
13	Exited	0	0.0

3.1.6 Data Cleaning and Transformation

This section involves cleaning and transforming the data. It replaces certain values in 'Geography' and 'Gender' columns to make them more interpretable. It also simplifies the 'HasCrCard' column by converting values greater than 0 to 1.

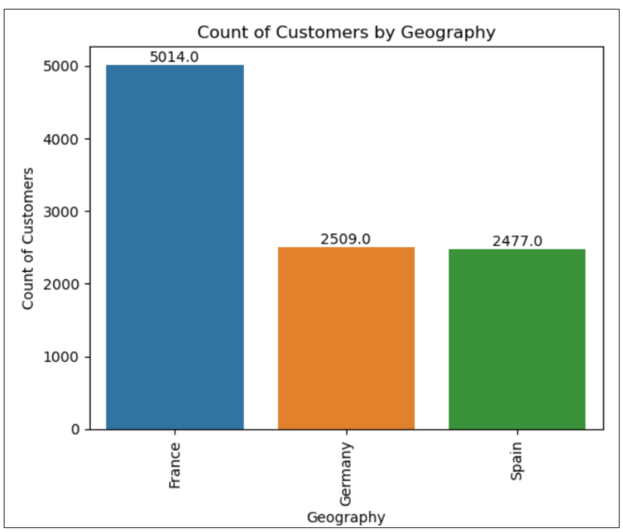
```
Number of unique values of  Geography : 3
unique values of  Geography are : Index(['France', 'Germany', 'Spain'], dtype='object')
Number of unique values of  Gender : 2
unique values of  Gender are : Index(['Male', 'Female'], dtype='object')
Number of unique values of  HasCrCard : 2
unique values of  HasCrCard are : Float64Index([1.0, 0.0], dtype='float64')
Number of unique values of  IsActiveMember : 2
unique values of  IsActiveMember are : Float64Index([1.0, 0.0], dtype='float64')
Number of unique values of  Exited : 2
unique values of  Exited are : Int64Index([0, 1], dtype='int64')
```

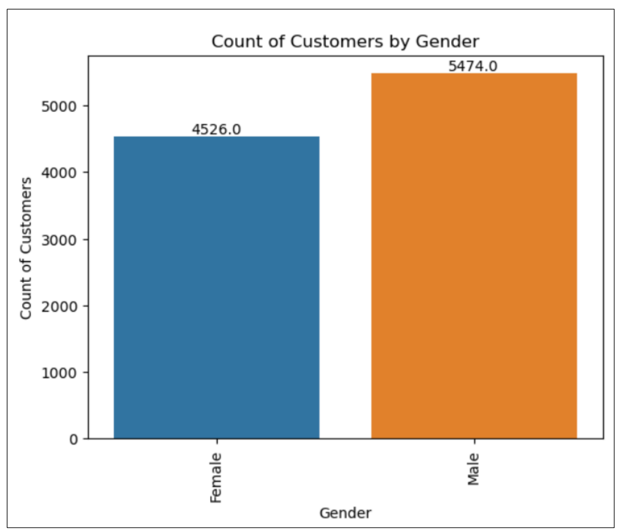
3.2 Exploratory Data Analysis (EDA)

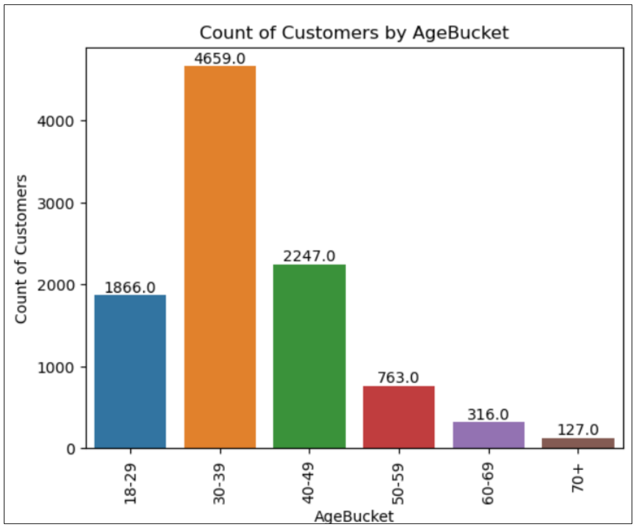
Perform EDA to understand the data's characteristics, identify patterns, and uncover insights. Explore the distribution of various customer attributes, such as age, gender, geography, credit score, and more, and their impact on churn rates.

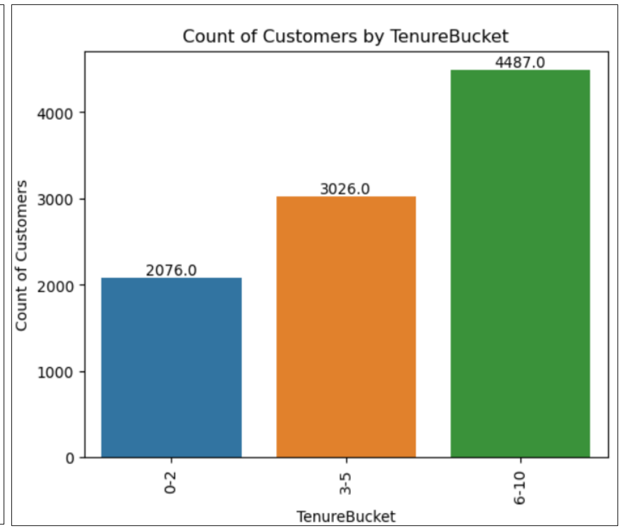
3.2.1 Univariate Analysis

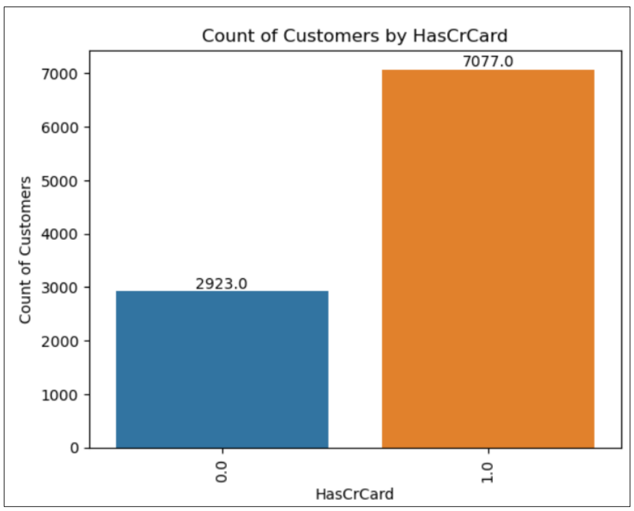
In univariate analysis, individual variables are examined separately. It provides insights into the distribution and characteristics of each variable.

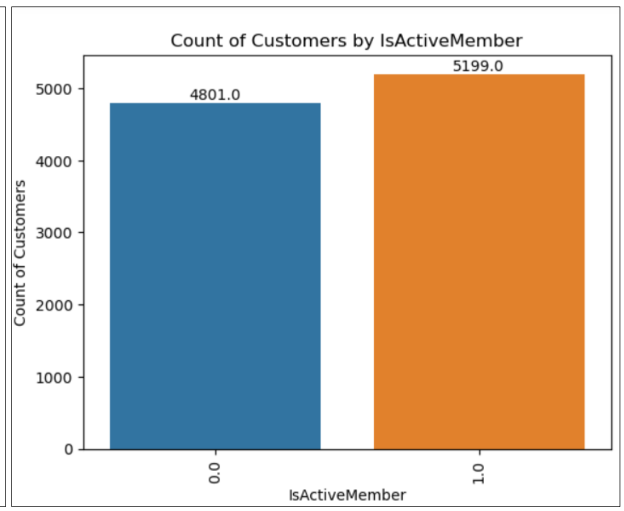










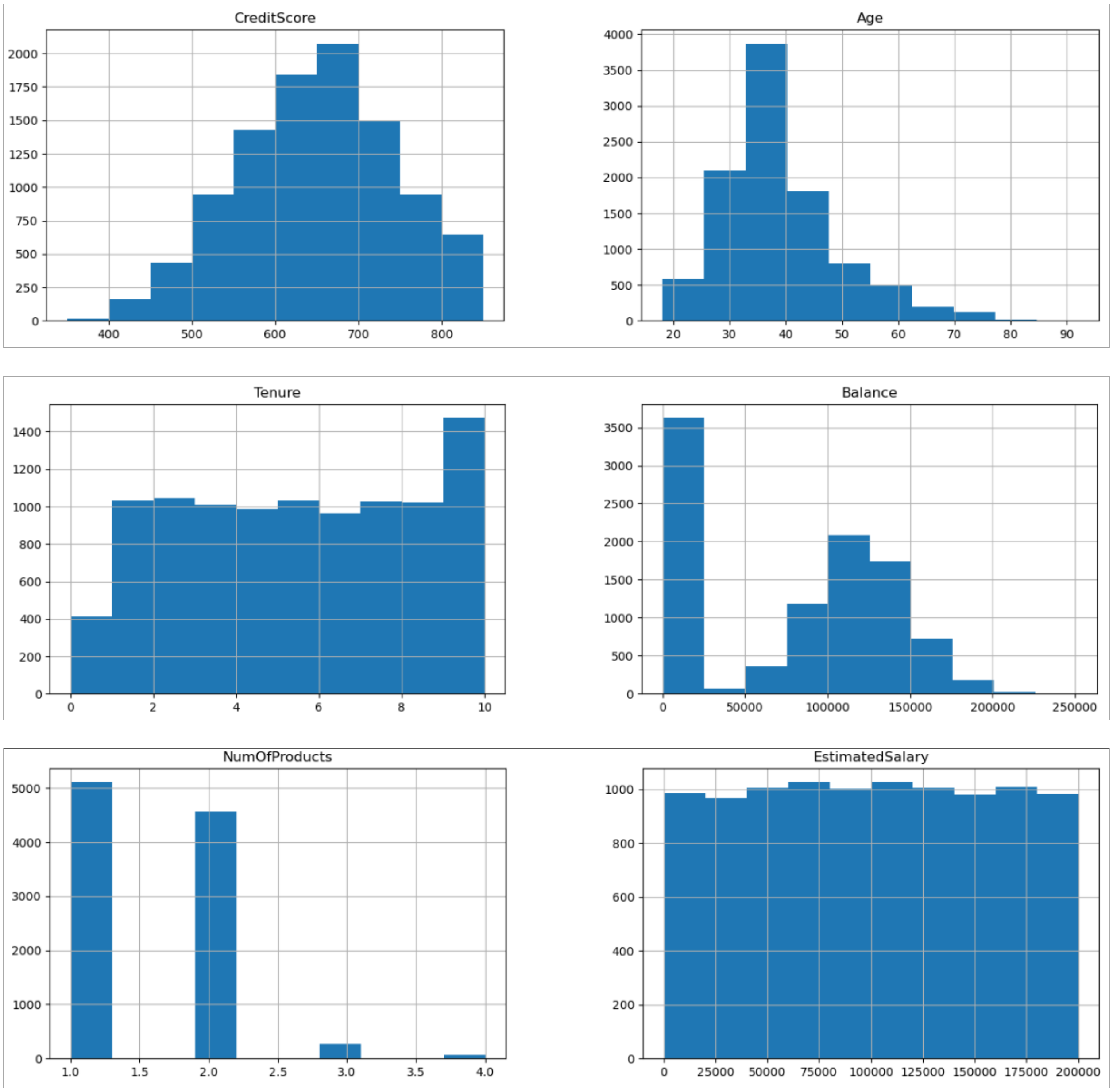


Findings:

- The majority of customers are from France, followed by Germany and Spain.
- More male customers than female customers.
- Most customers fall into the 30-39 age group, followed by 40-49 and 18-29.
- Approximately 70% of customers have a credit card.
- Around 52% are active members.

3.2.2 Univariate Analysis of Continuous Variables

Focuses on exploring the distribution and characteristics of continuous variables in the dataset. In this analysis, continuous variables such as "Age," "CreditScore," "Tenure," "Balance," "NumOfProducts," and "EstimatedSalary" are individually examined to understand their distributions.

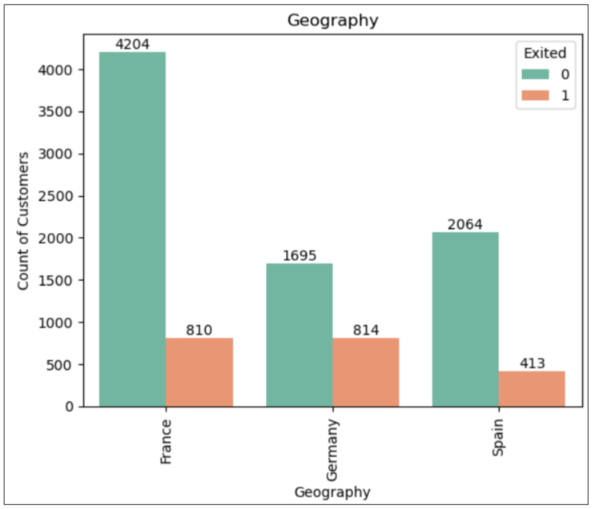


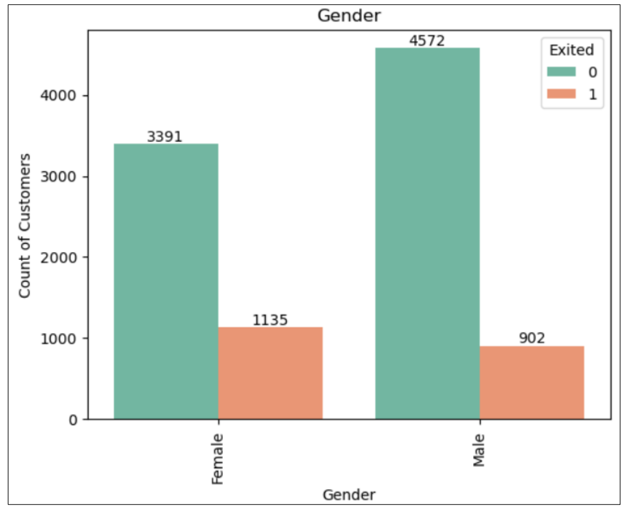
3.2.3 Bivariate Analysis

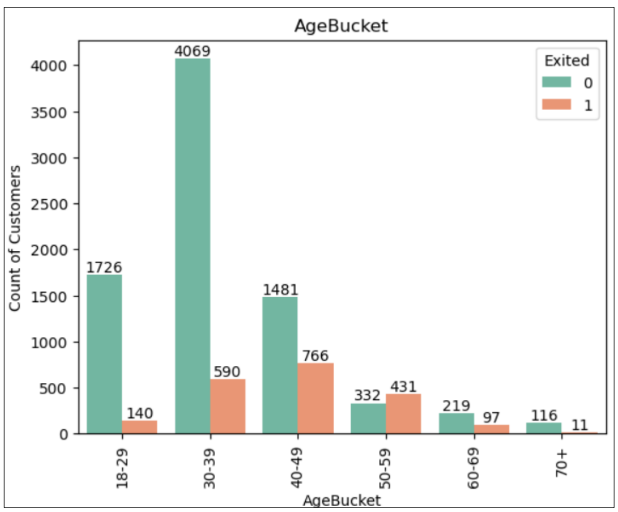
Bivariate analysis explores the relationships between two variables. It helps identify patterns and correlations between different attributes.

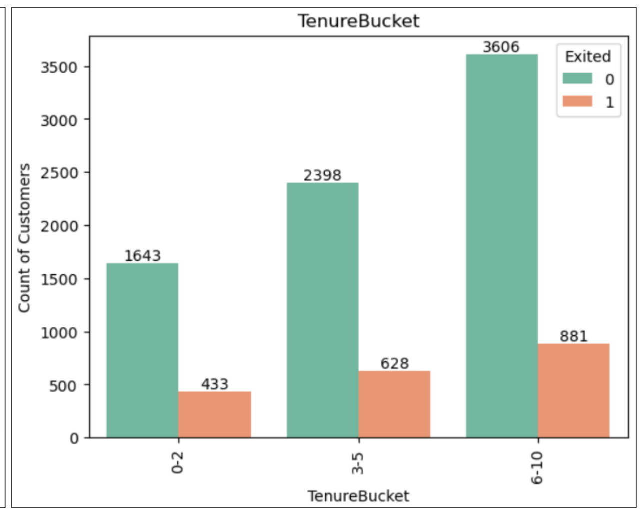
Findings:

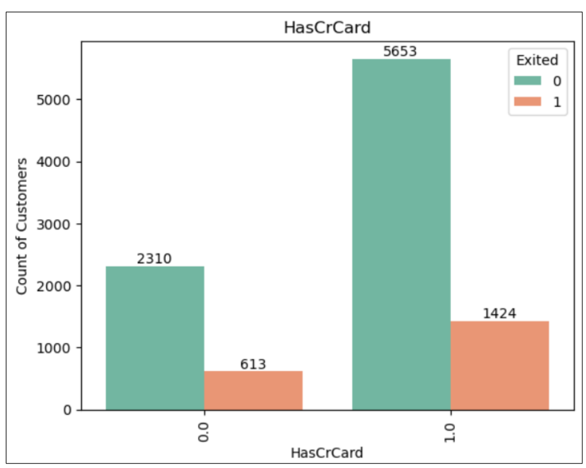
- The churn rate is highest in Germany (32.56%), followed by Spain (20%) and France (16%).
- The 50-59 age group has the highest churn rate (56.48%), followed by the 40-49 age group (34.81%).
- Both customers with and without a credit card have similar churn rates.
- Slightly higher churn rate for females compared to males.
- Churn rate is consistent across tenure groups.





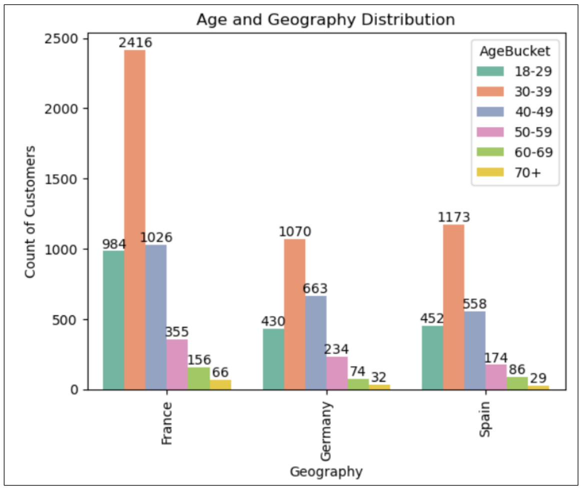
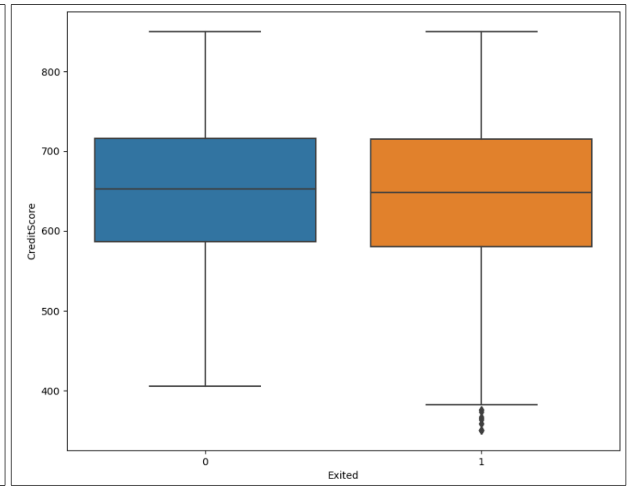
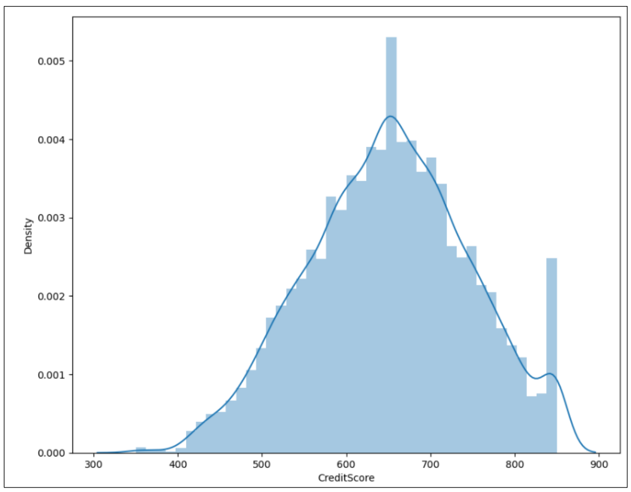
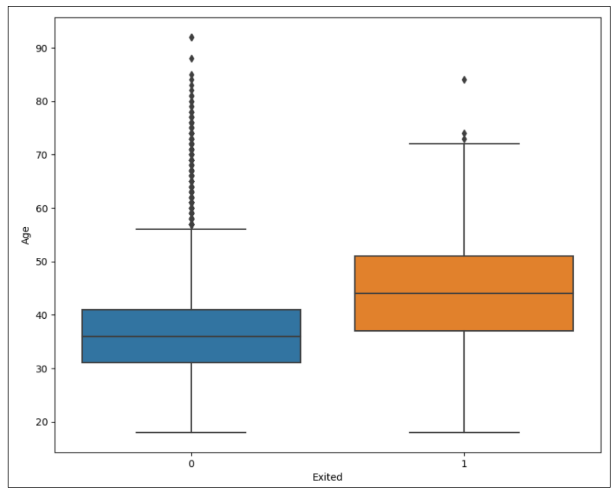
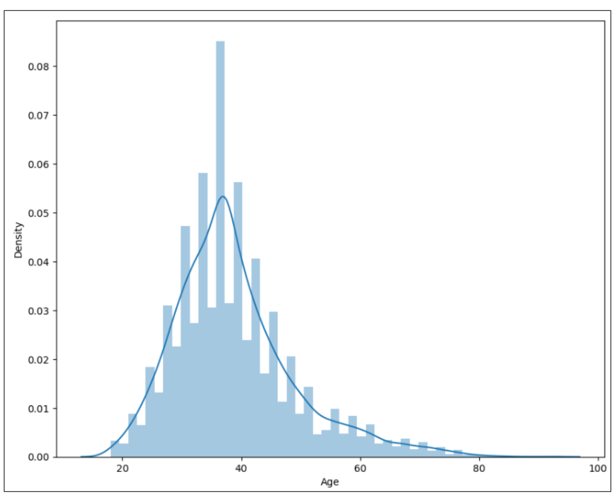
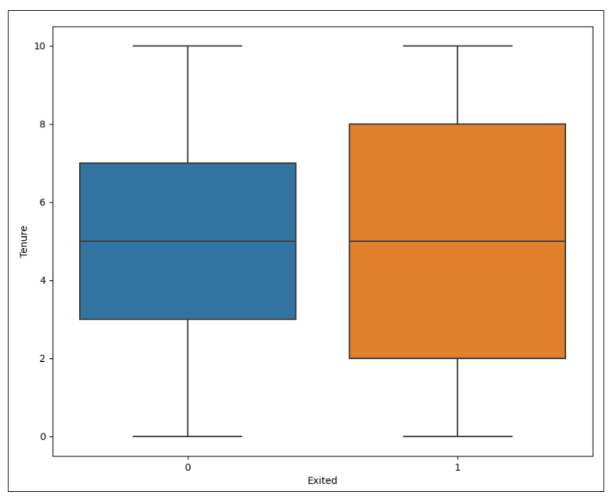


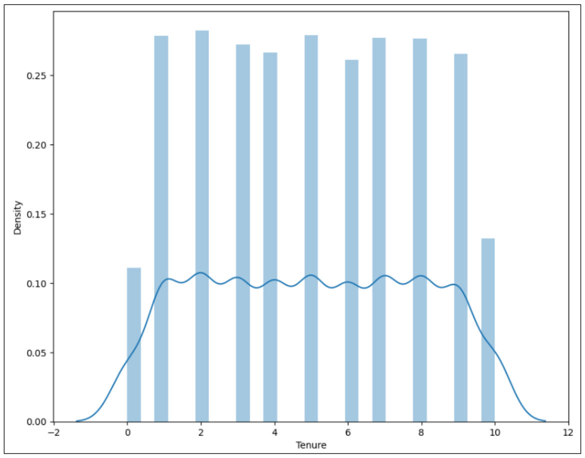


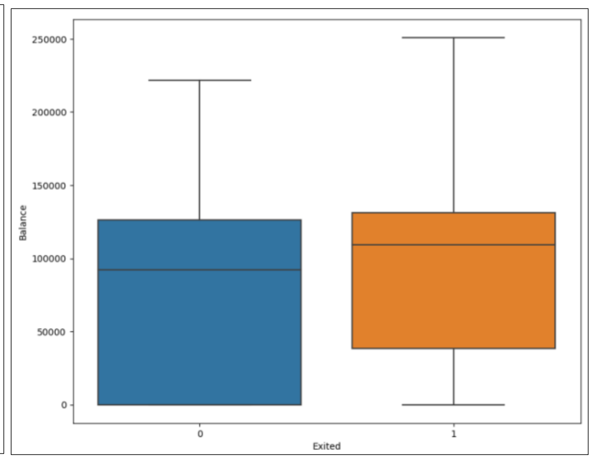


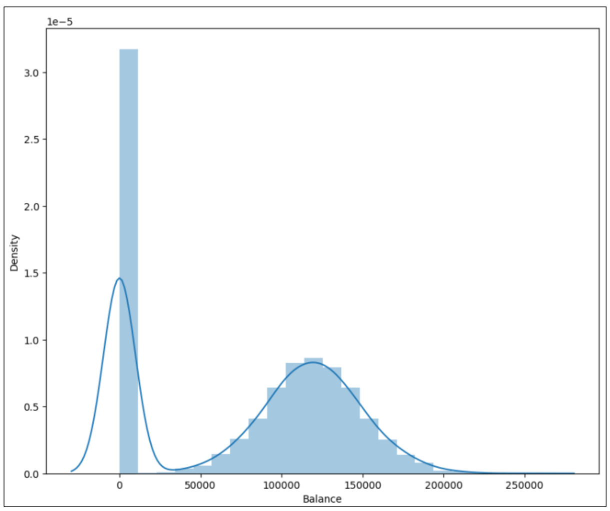
3.3 Churn Rate by Geography, Age, Credit Card, Gender, and Tenure

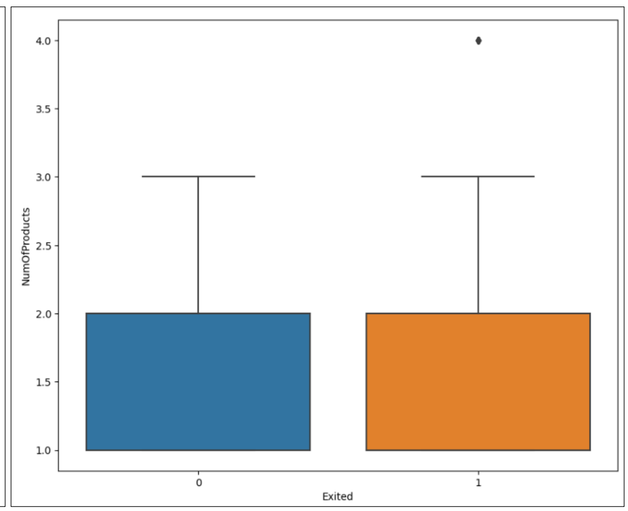
Bivariate analysis in this segment compares the churn rate (customers who have exited) across different categories of geography, age, credit card ownership, gender, and tenure. It identifies which demographic or behavioral factors may be associated with higher churn rates.


Age and Geography Distribution
Count of Customers
Geography
AgeBucket
18-29
30-39
40-49
50-59
60-69
70+
France
Germany
Spain
984
2416
1026
355
156
66
430
1070
663
234
74
32
452
1173
558
174
86
29

CreditScore
Exited
0
1

Density
CreditScore

Age
Exited
0
1

Density
Age

Tenure
Exited
0
1

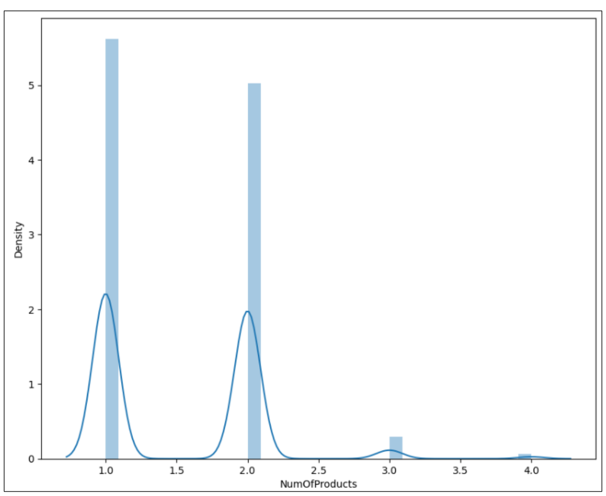

Density
0.00
0.05
0.10
0.15
0.20
0.25
−2 0 2 4 6 8 10 12
Tenure

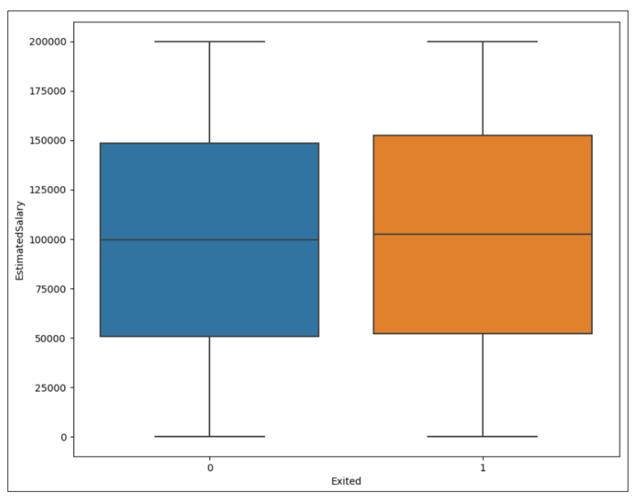

Balance
0
50000
100000
150000
200000
250000
0 1
Exited

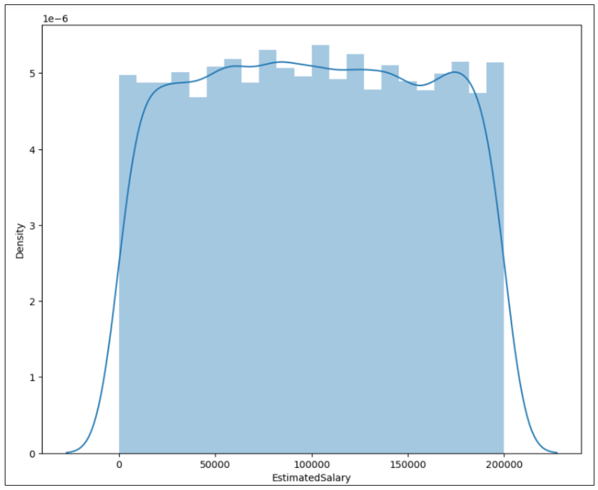

1e−5
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0 50000 100000 150000 200000 250000
Balance


NumOfProducts
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0 1
Exited

3


Density
0
1
2
3
4
5
1.0 1.5 2.0 2.5 3.0 3.5 4.0
NumOfProducts


EstimatedSalary
0
25000
50000
75000
100000
125000
150000
175000
200000
0 1
Exited



Findings:

- The Churn rate is high on Female when compared to Male.
- The people in Germany are more inclined to churn than the remaining 2 Geographical areas.
- Higher churn is observed in age buckets 50-59 followed by 40-49.
- Estimated Salary, Tenure, Has Credit Card do not impact the Churn Rate

4 Recommendations

Based on the inferences drawn from the analysis, here are some recommendations:

1. **Gender-Specific Strategies**: Given that the churn rate is higher among females, consider implementing gender-specific strategies. This might include personalized marketing, products, or services tailored to the needs and preferences of female customers.
2. **Focus on Germany**: Since Germany has a higher churn rate compared to other geographical areas, it's important to understand the specific factors contributing to this trend. Conduct further research or customer surveys in Germany to identify areas for improvement.
3. **Age-Targeted Campaigns**: The age buckets of 50-59 and 40-49 have higher churn rates. Create targeted campaigns or offers for customers in these age groups to retain them. These campaigns could include loyalty programs, discounts, or services tailored to the needs of customers in these age ranges.
4. **Improve Customer Engagement**: Factors like tenure, having a credit card, and estimated salary don't appear to significantly impact churn rate. Instead of focusing on these factors, consider implementing strategies to improve overall customer engagement and satisfaction. This might involve providing better customer support, enhancing the user experience, or offering more value to customers.
5. **Customer Feedback and Surveys**: To gain deeper insights into why customers are churning, consider conducting customer feedback surveys. Understanding the specific reasons for churn will help you develop targeted strategies to address these issues.
6. **Churn Prediction Models**: Implement churn prediction models that can proactively identify customers at risk of churning. This allows you to take preventive actions before customers actually churn. Machine learning models can be effective for this purpose.

7. **Customer Retention Programs**: Develop customer retention programs and loyalty initiatives to incentivize existing customers to stay with your service or product. Offering rewards, discounts, or exclusive benefits can help retain valuable customers.
8. **Competitive Analysis**: Analyze the competitive landscape to understand how your offerings compare to competitors. Identify areas where you can differentiate and improve your products or services to better meet customer needs and reduce churn.