FINTECH PROJECT- Loosing Bank Customers (Churn Analysis)

PROBLEM STATEMENT: In the rapidly evolving banking sector, customer retention has become a critical concern. Banks are increasingly seeking to understand the factors that influence customer decisions to stay with or leave their banking service provider. This project focuses on analyzing a dataset containing various attributes of bank customers to identify key predictors of customer churn. By leveraging data analytics, we aim to uncover patterns and insights that could help devise strategies to enhance customer retention and reduce churn rates.

Objectives:

Identify and Understand Key Factors: Recognize which customer attributes significantly impact the likelihood of churn.

Develop Analytical Skills: Enhance proficiency in using statistical tools and software for analyzing complex datasets.

Generate Insights: Produce actionable insights that could theoretically be used by a bank to improve customer retention strategies.

Present Findings: Effectively communicate findings and recommendations through clear, compelling visualizations and reports.

Importance of Understanding Customer Churn:

Revenue Optimization: By identifying the key drivers of customer churn, banks can optimize their revenue generation strategies. This includes personalized offers, targeted marketing campaigns, and improved customer service.

Customer Experience Enhancement: Understanding churn factors allows banks to tailor their services to meet customer needs effectively. This may involve improving customer support, enhancing digital banking experience, and providing personalized financial advice.

Operational Efficiency: By accurately predicting churn, banks can optimize their operational processes, such as resource allocation, customer service staffing, and marketing spend, leading to cost savings and improved service reliability.

Customer Loyalty and Trust: Banks aim to build long-term relationships with their customers. Reducing churn helps in maintaining customer loyalty and trust, which are crucial for the bank's reputation and growth.

Competitive Advantage: Understanding and mitigating churn gives banks a competitive edge in retaining customers in a highly competitive market.

Data Description:

The dataset includes the following attributes:

RowNumber: Corresponds to the record (row) number and has no effect on the output.

CustomerId: Contains random values and has no effect on customer leaving the bank.

Surname: The surname of a customer has no impact on their decision to leave the bank.

CreditScore: Can affect customer churn, as a customer with a higher credit score is less likely to leave the bank.

Geography: A customer's location can affect their decision to leave the bank.

Gender: It's interesting to explore whether gender plays a role in a customer leaving the bank.

Age: This is relevant since older customers are less likely to leave their bank than younger ones.

Tenure: Refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.

Balance: Also a good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.

NumOfProducts: Refers to the number of products that a customer has purchased through the bank.

HasCrCard: Denotes whether or not a customer has a credit card. This column is relevant, as people with a credit card are less likely to leave the bank.

IsActiveMember: Active customers are less likely to leave the bank.

EstimatedSalary: As with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.

Exited: Whether or not the customer left the bank.

Complain: Customer has a complaint or not.

Satisfaction Score: Score provided by the customer for their complaint resolution.

Card Type: Type of card held by the customer.

Points Earned: The points earned by the customer for using a credit card.

By gaining insights into the factors driving customer churn, banks can make informed decisions that enhance customer retention, optimize revenue, and improve operational efficiency. This not only benefits the bank's financial performance but also contributes to building long-term customer relationships and trust.

ANALYZING BASIC METRICS: Analyzing basic metrics involves examining fundamental characteristics and summary statistics of the dataset to gain an initial understanding of its structure and content. This step is crucial in exploratory data analysis (EDA) as it provides insights into the data's distribution, variability, and potential issues that may need to be addressed before proceeding with further analysis. Here's how we can define and conduct basic metrics analysis:

Analyzing basic metrics involves:

1. **Data Overview:** Let's start by loading the dataset and getting an overview of its structure and content.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Importing dataset
data = pd.read_csv('G:/dsml-scaler/fintech_project/Bank-Records.csv')
data.head()
```

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Bala
0	1	15634602	Hargrave	619	France	Female	42	2	(
1	2	15647311	Hill	608	Spain	Female	41	1	83807
2	3	15619304	Onio	502	France	Female	42	8	15966(
3	4	15701354	Boni	699	France	Female	39	1	(
4	5	15737888	Mitchell	850	Spain	Female	43	2	12551(

```
data.shape
```

```
(10000, 17)
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   CustomerId          10000 non-null  int64  
 1   Surname             10000 non-null  object 
 2   CreditScore         10000 non-null  int64  
 3   Geography           10000 non-null  object 
 4   Gender              10000 non-null  object 
 5   Age                 10000 non-null  int64  
 6   Tenure              10000 non-null  int64  
 7   Balance             10000 non-null  float64
 8   NumOfProducts       10000 non-null  int64  
 9   HasCrCard           10000 non-null  int64  
 10  IsActiveMember      10000 non-null  int64  
 11  EstimatedSalary     10000 non-null  float64
 12  Exited              10000 non-null  int64  
 13  Complain            10000 non-null  int64  
 14  Satisfaction Score  10000 non-null  int64  
 15  Card Type           10000 non-null  object 
 16  Point Earned        10000 non-null  int64  
dtypes: float64(2), int64(11), object(4)
memory usage: 1.3+ MB
```

Insights: From the provided dataset information:

- It contains 10,000 entries and 17 columns.
- There are no missing values (non-null count is the same for all columns).
- The dataset includes a mix of numerical (int64 and float64) and categorical (object) columns.

2. **Descriptive Statistics:** Calculating summary statistics such as mean, median, mode, standard deviation, minimum, maximum, and quartiles for numerical as well as object variables.

```
data.describe()
```

	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProduct
count	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000
mean	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.53020
std	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.58165
min	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.00000
25%	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.00000
50%	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.00000
75%	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.00000
max	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.00000

Insights:

CreditScore:

- The average credit score is approximately 650, with a standard deviation of around 97.
- Credit scores range from a minimum of 350 to a maximum of 850.
- The majority of customers have credit scores between 584 (25th percentile) and 718 (75th percentile).

Age:

- The average age of customers is approximately 39 years old, with a standard deviation of around 10.
- Ages range from a minimum of 18 to a maximum of 92.
- Most customers fall between the ages of 32 (25th percentile) and 44 (75th percentile).

Tenure:

- The average tenure of customers with the bank is approximately 5 years, with a standard deviation of around 2.9.
- Tenure ranges from a minimum of 0 years to a maximum of 10 years.
- The majority of customers have tenures between 3 (25th percentile) and 7 (75th percentile) years.

Balance:

- The average account balance is approximately $76,486, with a standard deviation of around $62,397.

- Balances range from a minimum of $0 to a maximum of $250,898.
- There is a wide variation in account balances, with some customers having zero balance and others having high balances.

NumOfProducts:

- The average number of products held by customers is approximately 1.53, with a standard deviation of around 0.58.
- The number of products ranges from a minimum of 1 to a maximum of 4.
- Most customers have 1 (25th percentile) or 2 (75th percentile) products.

HasCrCard:

Around 70.5% of customers have a credit card (HasCrCard), indicating that credit card ownership is common among customers.

IsActiveMember:

Approximately 51.5% of customers are active members, suggesting that the majority of customers engage with the bank's services.

EstimatedSalary:

- The average estimated salary of customers is approximately $100,090, with a standard deviation of around $57,510.
- Estimated salaries range from a minimum of $11.58 to a maximum of $199,992.

Exited:

Around 20.4% of customers have churned (Exited), indicating that churn is present in the dataset.

Complain:

Approximately 20.4% of customers have filed a complaint, showing that a similar proportion of customers have experienced issues or dissatisfaction.

Satisfaction Score:

- The average satisfaction score provided by customers is approximately 3, with a standard deviation of around 1.41.
- Satisfaction scores range from a minimum of 1 to a maximum of 5.
- Most customers have satisfaction scores between 2 (25th percentile) and 4 (75th percentile).

Point Earned:

- The average number of points earned by customers is approximately 607, with a standard deviation of around 226.
- Points earned range from a minimum of 119 to a maximum of 1000.

- These insights provide a comprehensive understanding of the numerical attributes in the dataset, which can further guide exploratory data analysis and inform subsequent analysis and decision-making processes.

```
data.describe(include = 'object')
```

	Surname	Geography	Gender	Card Type
count	10000	10000	10000	10000
unique	2932	3	2	4
top	Smith	France	Male	DIAMOND
freq	32	5014	5457	2507

Insights:

- The most common surname among customers is "Smith," with a frequency of 32 occurrences.
- The majority of the data pertains to customers from France, with a frequency of 5,014 entries. This indicates that they can focus more on the customers from France.
- Male customers predominate the dataset, with a frequency of 5,457.
- The most frequently used card type is the "Diamond" card, with a frequency of 2,507 customers.

3. Handling Missing Values:

```
data.isna().sum()
```

```
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
Complain              0
Satisfaction Score    0
Card Type             0
Point Earned          0
dtype: int64
```

There are no missing values found. The data is clean.

4. **Non- Graphical Analysis:** First we are going to detect duplicate rows.

```
# Check for duplicate rows
duplicates = data.duplicated()

# Count the number of duplicate rows
num_duplicates = duplicates.sum()
print(f"\nNumber of duplicate rows: {num_duplicates}")
```

```
Number of duplicate rows: 0
```

No duplicate values are found in the dataset.
Now we are going to check for the unique values in dataset.

```
# Categorical columns in the dataset
categorical_columns = ['Geography', 'Gender', 'Card Type']

# Display unique values and their counts for each categorical column
for column in categorical_columns:
    unique_values = data[column].value_counts()
    print(f"Unique values in '{column}':")
    print(unique_values)
    print()
```

```
Unique values in 'Geography':
Geography
France     5014
Germany    2509
Spain      2477
Name: count, dtype: int64

Unique values in 'Gender':
Gender
Male      5457
Female    4543
Name: count, dtype: int64

Unique values in 'Card Type':
Card Type
DIAMOND     2507
GOLD        2502
SILVER      2496
PLATINUM    2495
Name: count, dtype: int64
```

```
#identify numerical columns
numerical_columns = data.select_dtypes(include=['int64', 'float64']).columns

# Count the number of unique values in each numerical column
unique_counts_numerical = data[numerical_columns].nunique()

# Display the counts of unique values
print("Count of unique values in numerical columns:")
print(unique_counts_numerical)

# Optionally, inspect the actual unique values for each numerical column
for column in numerical_columns:
    unique_values = data[column].unique()
    print(f"\nUnique values in '{column}':")
    print(unique_values)
```

```
Count of unique values in numerical columns:
CustomerId            10000
CreditScore             460
Age                      70
Tenure                   11
Balance                6382
NumOfProducts             4
HasCrCard                 2
IsActiveMember            2
EstimatedSalary        9999
Exited                    2
Complain                  2
Satisfaction Score        5
Point Earned            785
```

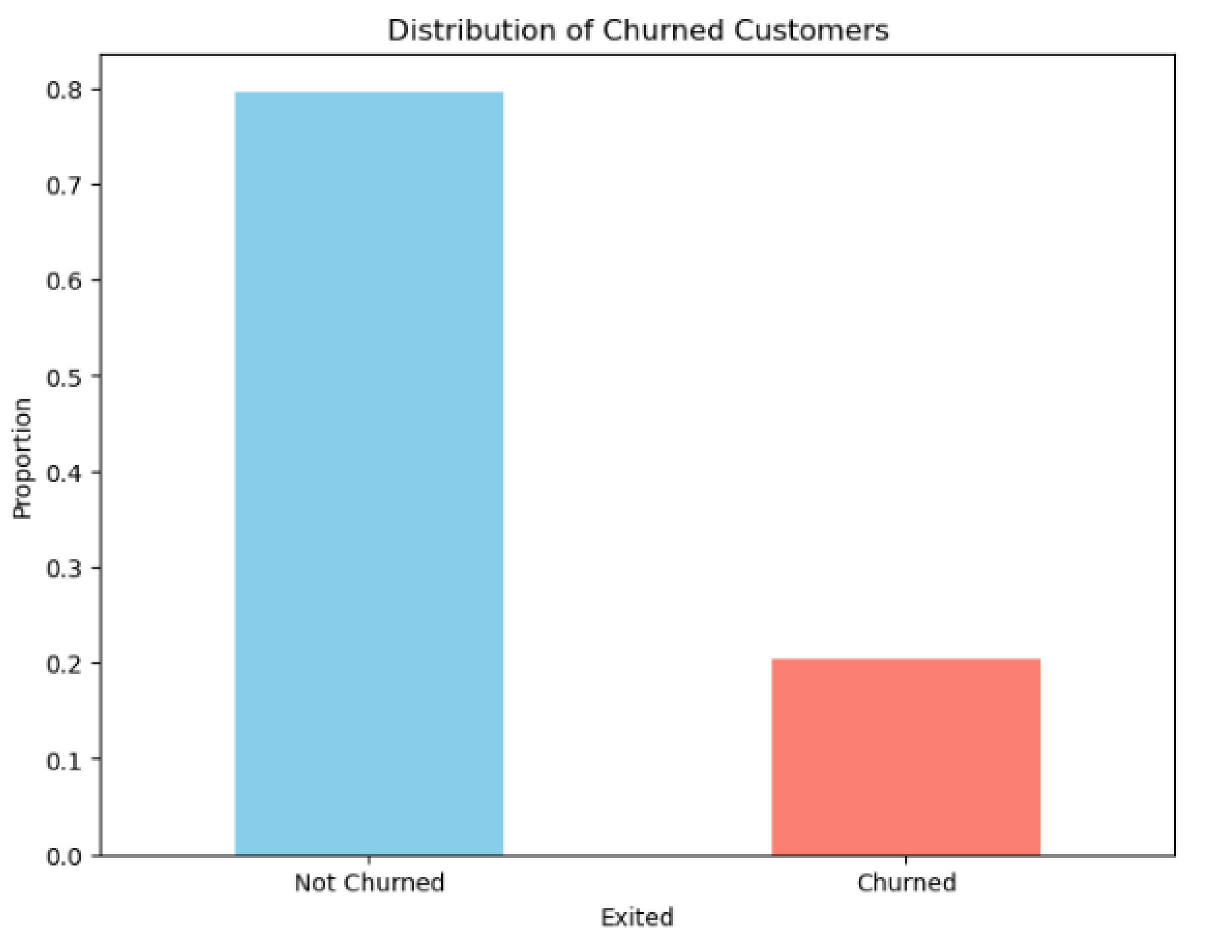
5. **Distribution Analysis:** Examining the distribution of numerical variables through histograms, kernel density estimation (KDE) plots. First we are going to start with univariate analysis.

   a) **Univariate Analysis:** Univariate analysis serves as the foundation of exploratory data analysis, providing valuable insights into individual variables that inform subsequent steps in the data analysis process. It helps us understand the data's structure, identify areas of interest, and lay the groundwork for more in-depth analyses and modeling. Through univariate analysis, we can identify patterns, trends, and anomalies within each variable. This helps in detecting any inconsistencies or unexpected values that may require further investigation or preprocessing before proceeding with more complex analyses.

```
# Distribution of the target variable
churn_counts = data['Exited'].value_counts(normalize=True)
print(churn_counts)
```

```
Exited
0    0.7962
1    0.2038
Name: proportion, dtype: float64
```

```
# Plotting the distribution of the target variable
plt.figure(figsize=(8, 6))
churn_counts.plot(kind='bar', color=['skyblue', 'salmon'])
plt.title('Distribution of Churned Customers')
plt.xlabel('Exited')
plt.ylabel('Proportion')
plt.xticks(ticks=[0, 1], labels=['Not Churned', 'Churned'], rotation=0)
plt.show()
```

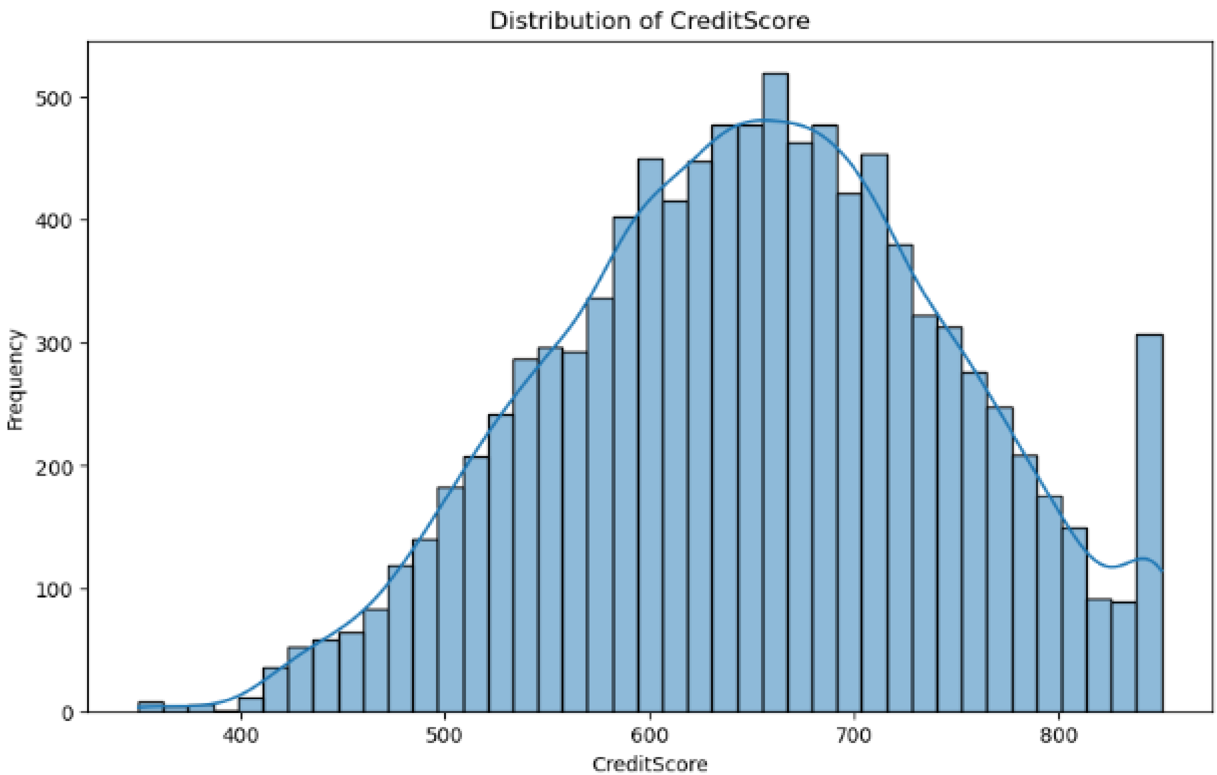


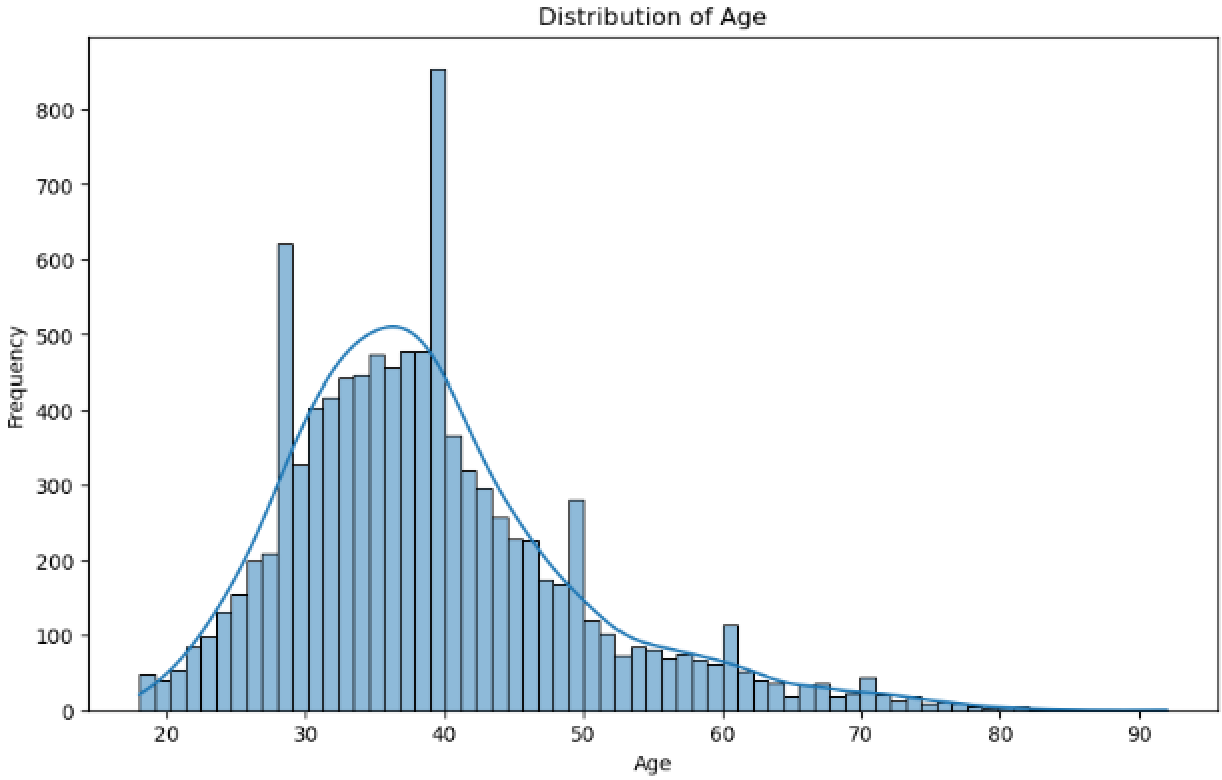
Insights:

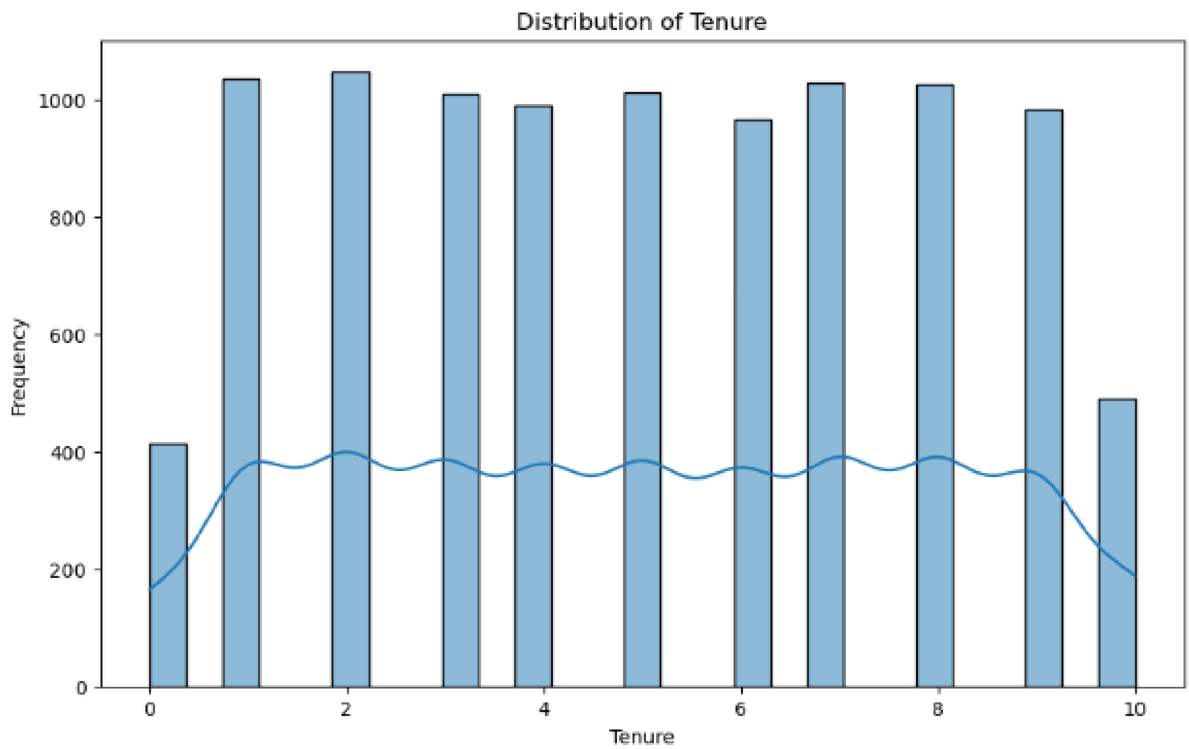
- **Non-Churned Customers (Exited = 0):** The proportion of non-churned customers is approximately 79.62%. This indicates that the majority of customers in the dataset did not churn or leave the bank.
- **Churned Customers (Exited = 1):** The proportion of churned customers is approximately 20.38%. This suggests that a significant minority of customers in the dataset churned or left the bank.
- **Customer Retention Strategies:** With approximately one-fifth of customers churning, it's crucial for the bank to implement effective customer retention strategies. This could include improving customer service, personalized offerings, loyalty programs, and targeted marketing campaigns to reduce churn and increase customer loyalty.

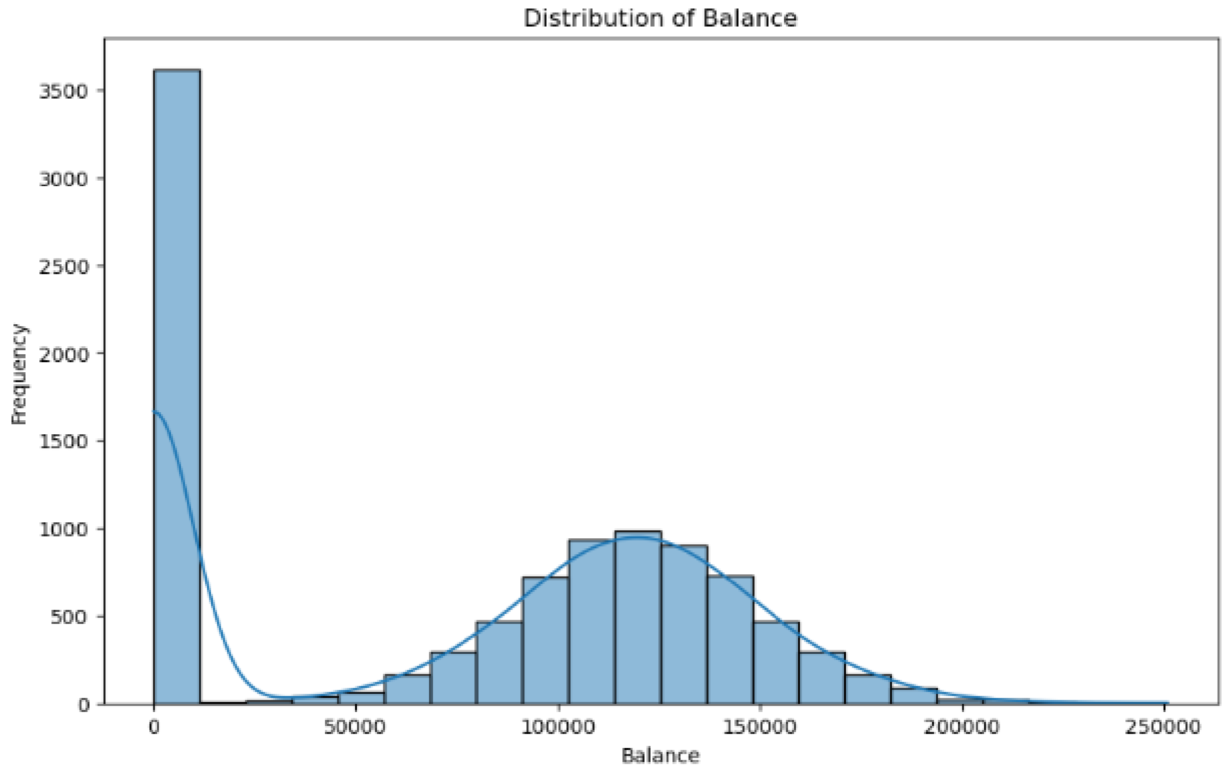
- **Customer Segmentation:** Analyzing churned customers' demographics, behaviors, and preferences can help identify patterns and segments that are more likely to churn. This information can be used to tailor retention strategies and product offerings to specific customer segments, ultimately improving retention rates. Also we can target on the 80% customer and provide them some offers so that they are always satisfied and become our fixed income source.

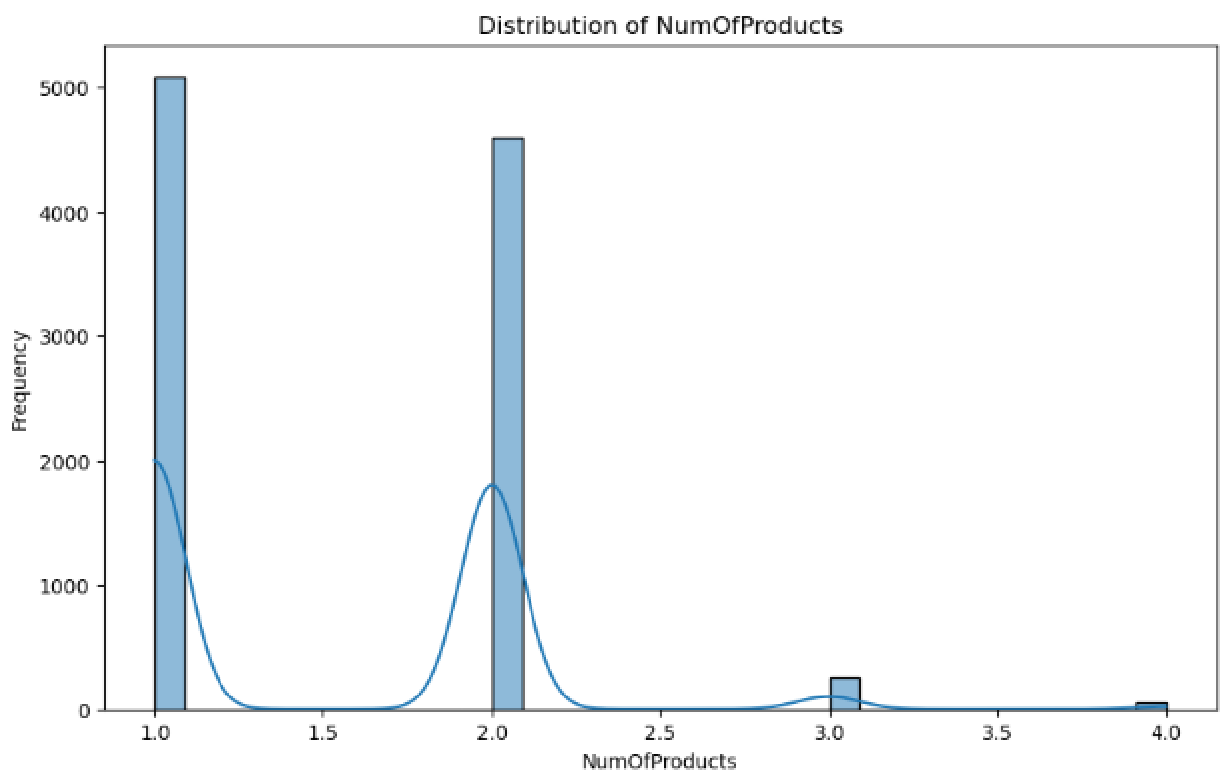
```
# Plot histograms with KDE
for column in numerical_columns:
    plt.figure(figsize=(10, 6))
    sns.histplot(data[column], kde=True)
    plt.title(f'Distribution of {column}')
    plt.xlabel(column)
    plt.ylabel('Frequency')
    plt.show()
```

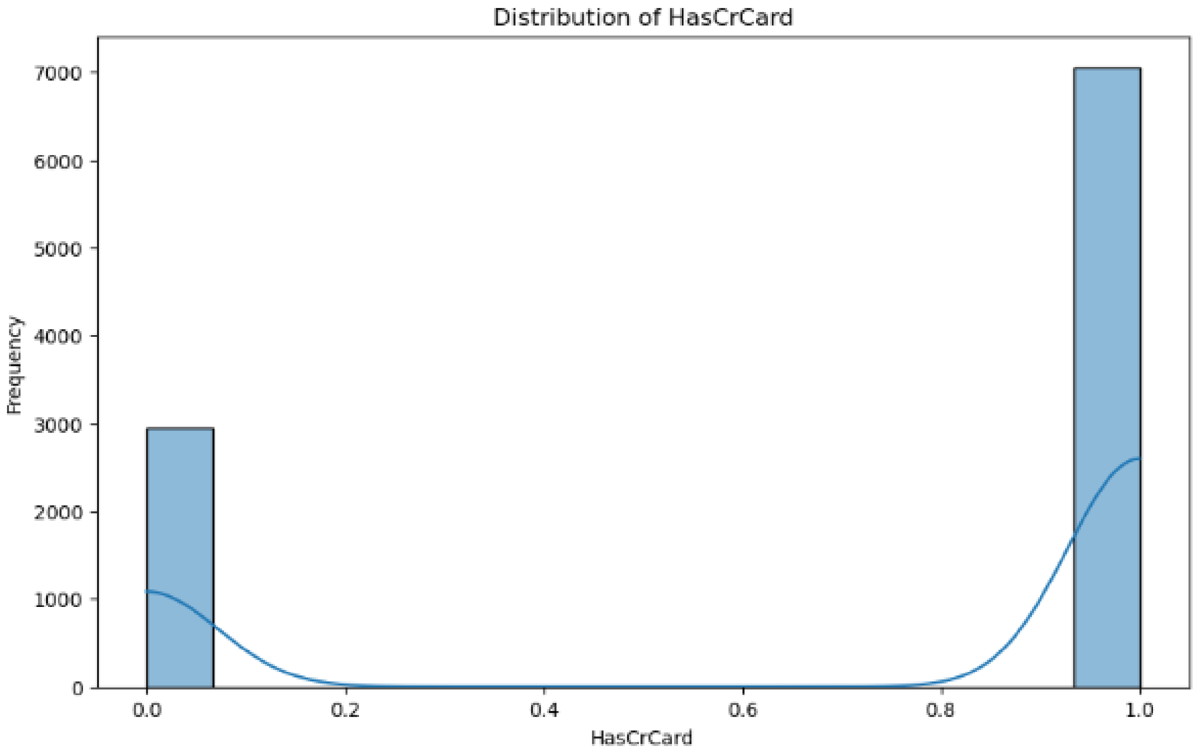


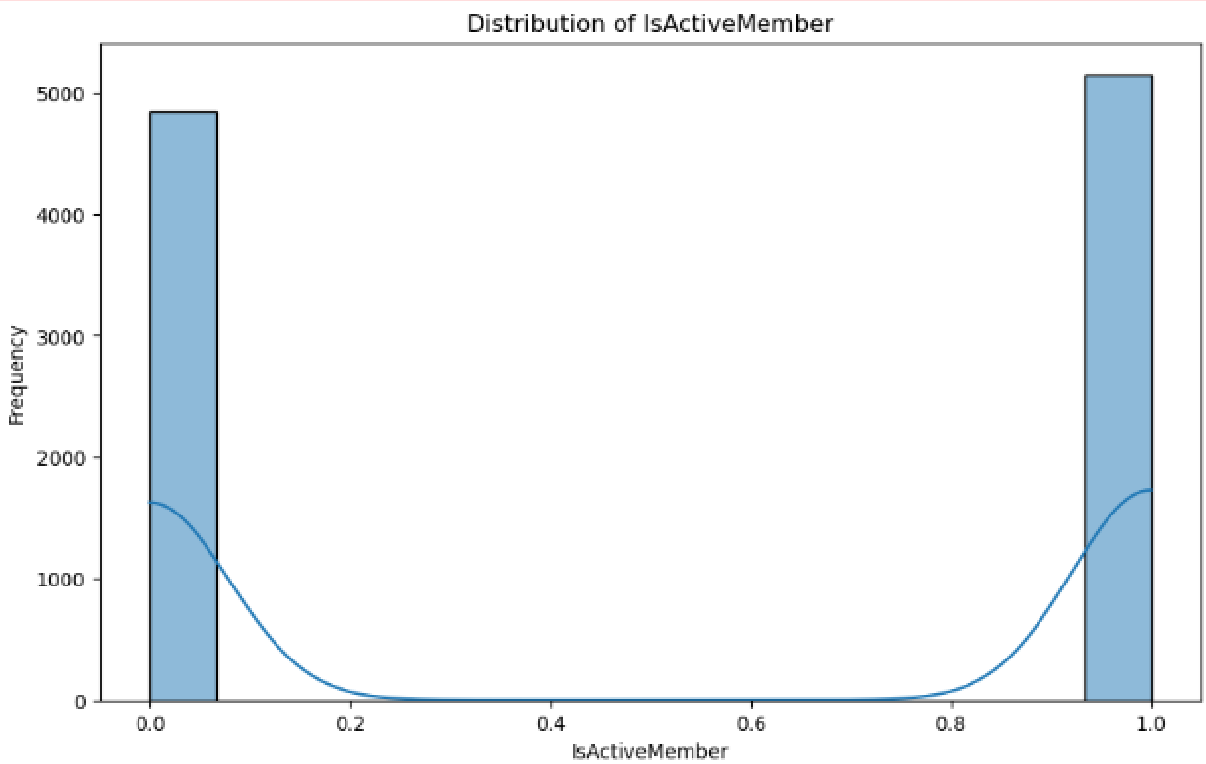

Distribution of Age
Frequency
Age
0
100
200
300
400
500
600
700
800
20
30
40
50
60
70
80
90

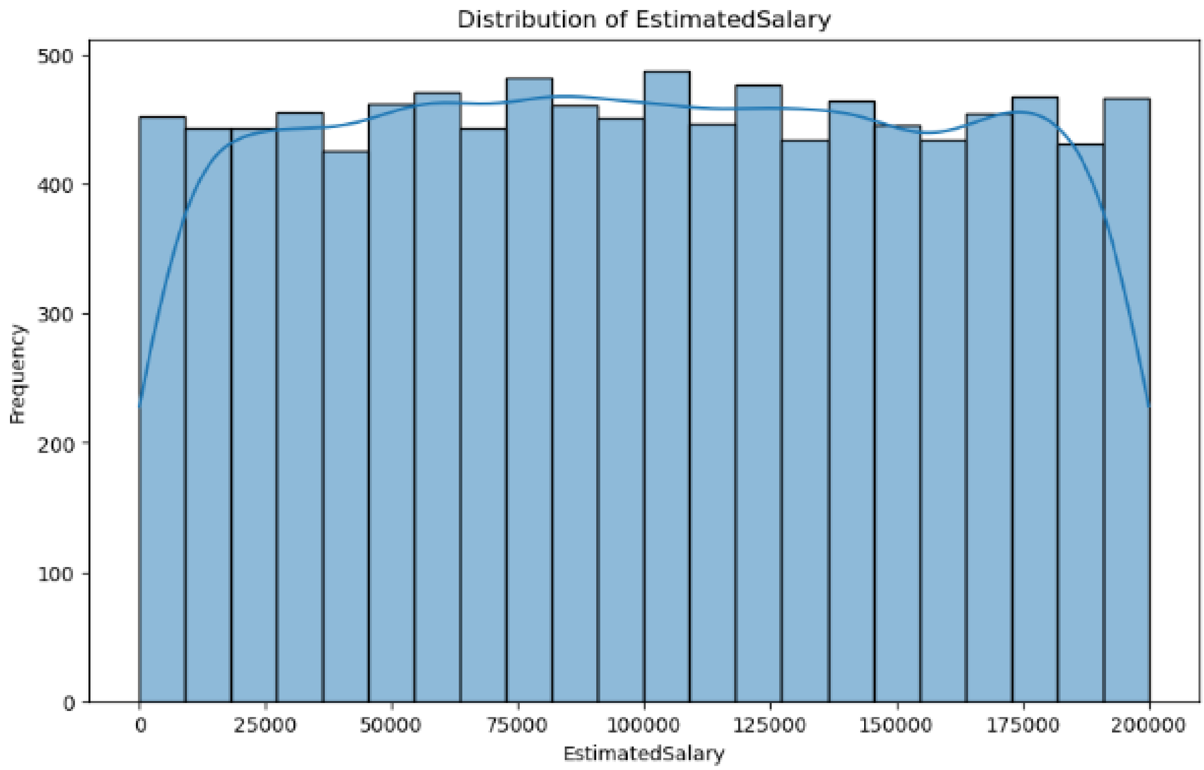

Distribution of Tenure
Frequency
Tenure
0
200
400
600
800
1000
0
2
4
6
8
10

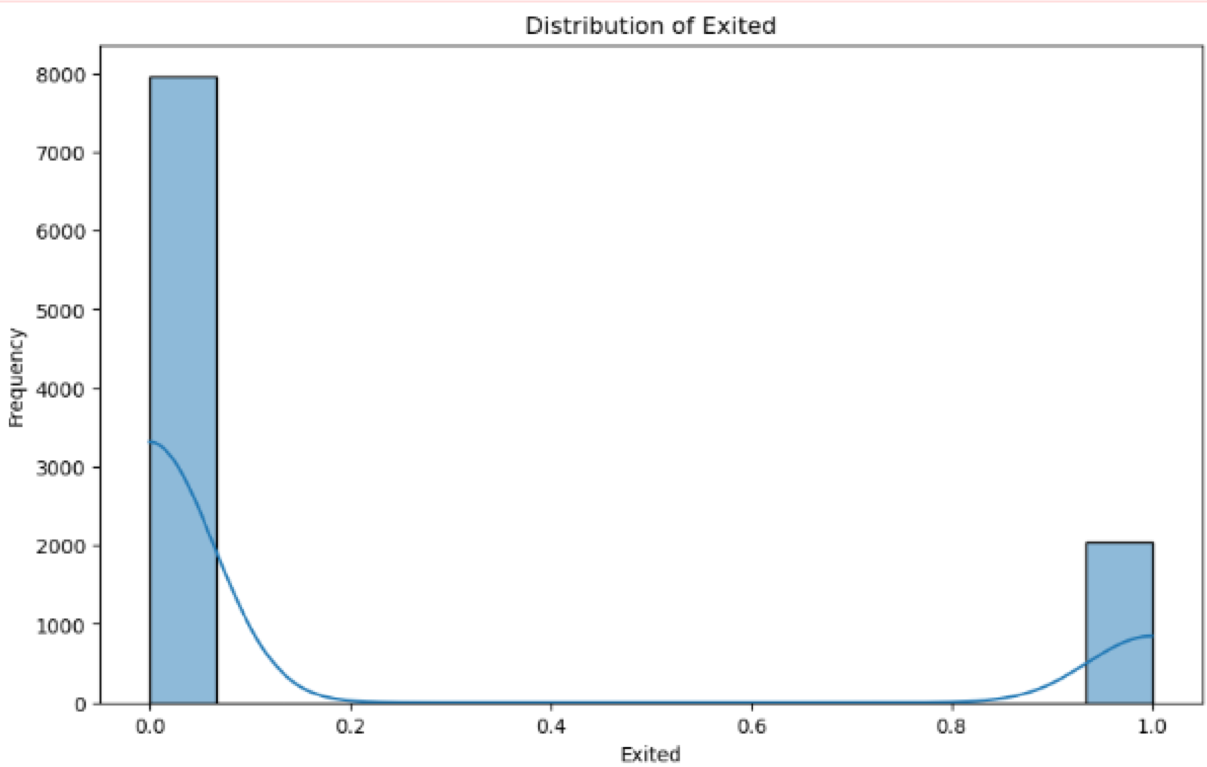

Distribution of Balance
Frequency
Balance
0
500
1000
1500
2000
2500
3000
3500
50000
100000
150000
200000
250000

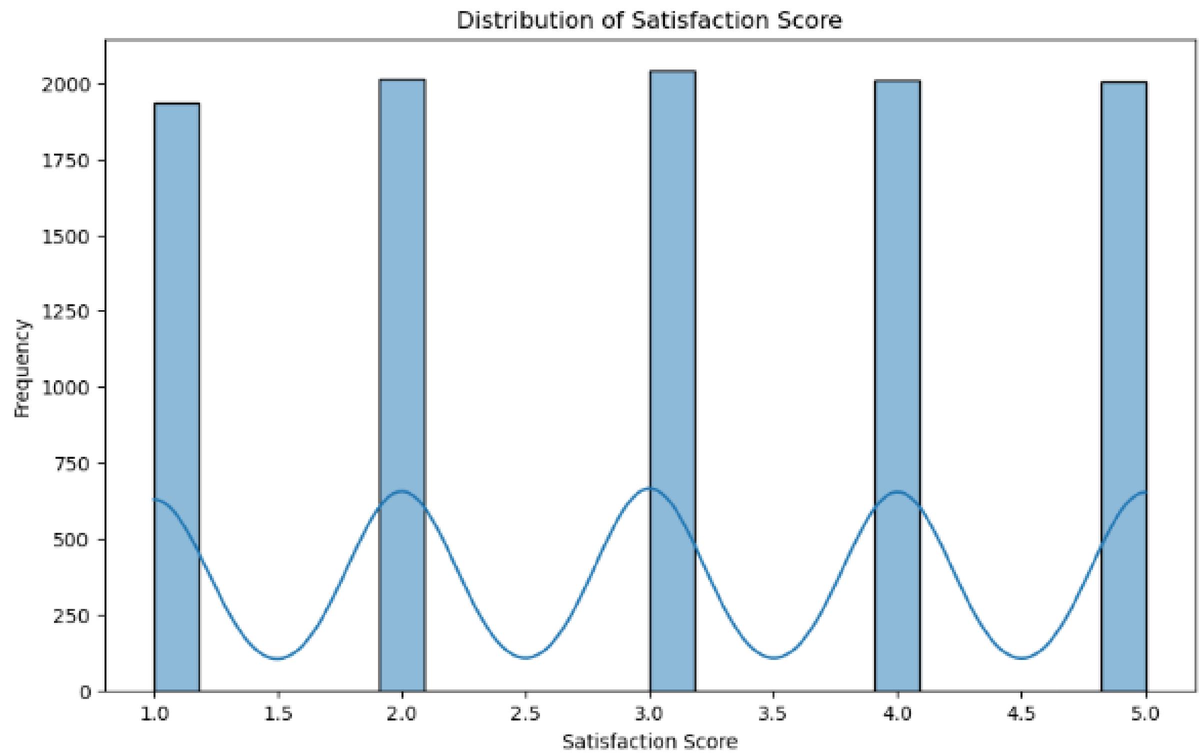

Distribution of NumOfProducts
Frequency
NumOfProducts
0
1000
2000
3000
4000
5000
1.0
1.5
2.0
2.5
3.0
3.5
4.0

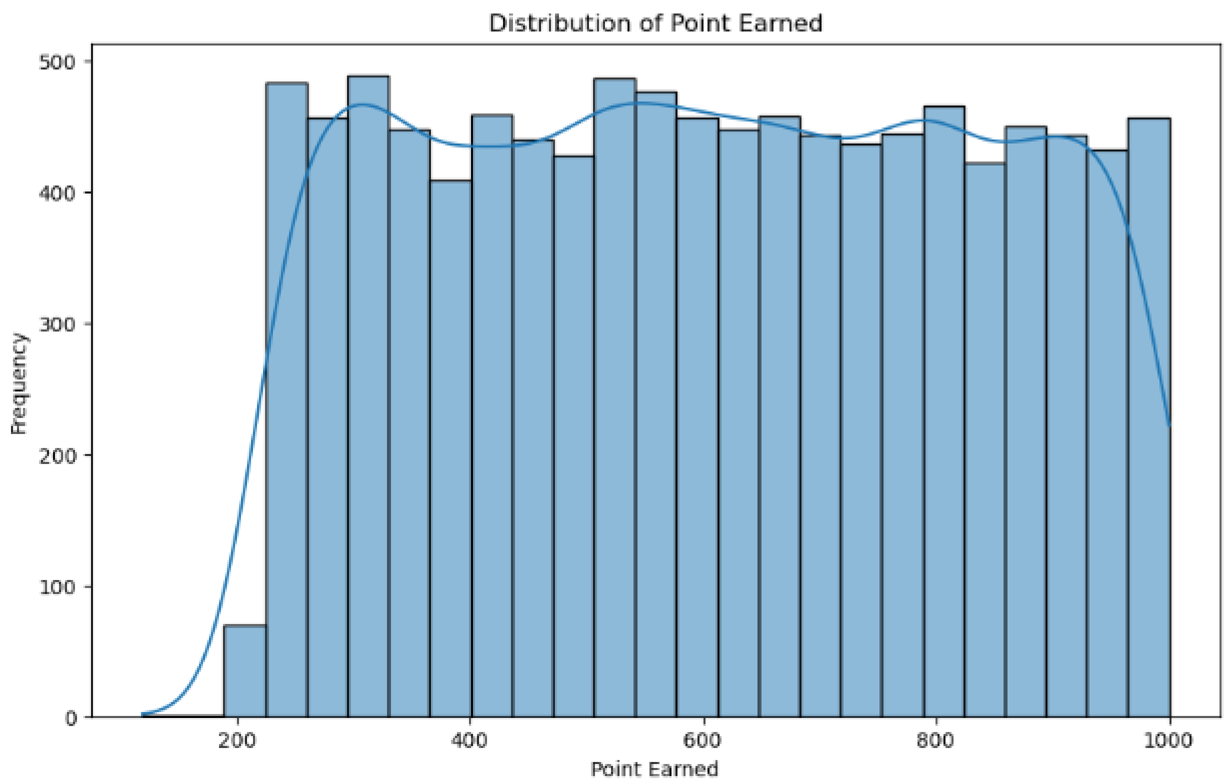

Distribution of HasCrCard
Frequency
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
HasCrCard


Distribution of IsActiveMember
Frequency
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
IsActiveMember


Distribution of EstimatedSalary
500
400
300
200
100
0
Frequency
0
25000
50000
75000
100000
125000
150000
175000
200000
EstimatedSalary


Distribution of Exited
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Exited





Insights:

Credit Score Distribution: The majority of customers have credit scores ranging from 610 to 710, indicating a central tendency in creditworthiness. The distribution appears slightly left-skewed, suggesting that more customers have higher credit scores.

Age Distribution: The primary age range of customers falls between 31 to 41 years, with the maximum age around 41. The right skewness observed in the distribution is expected, as some customers are older, leading to a gradual decrease in the frequency of older age groups.

Tenure Distribution: The distribution of tenure spans across various years, typically ranging from 1 to 19 years. This indicates a relatively consistent distribution of customer tenure, with no significant skewness observed.

Balance Distribution: The distribution of account balances appears to be nearly normal, with a notable number of customers maintaining zero balances. This suggests a diverse range of financial behaviors among customers, with some maintaining significant balances and others keeping minimal balances.

Number of Products Distribution: The majority of customers hold one to two products, as evidenced by the distribution of the number of products they possess.

Credit Card Ownership: A significant portion of customers possesses credit cards, indicating the widespread use and acceptance of this financial instrument among the customer base.

Active Membership: The number of active and non-active customers appears to be relatively balanced, indicating a comparable engagement level among customers with the bank's services.

Salary Distribution: The distribution of estimated salaries among customers shows a similar pattern across different salary levels, suggesting a relatively uniform distribution of income among the customer base.

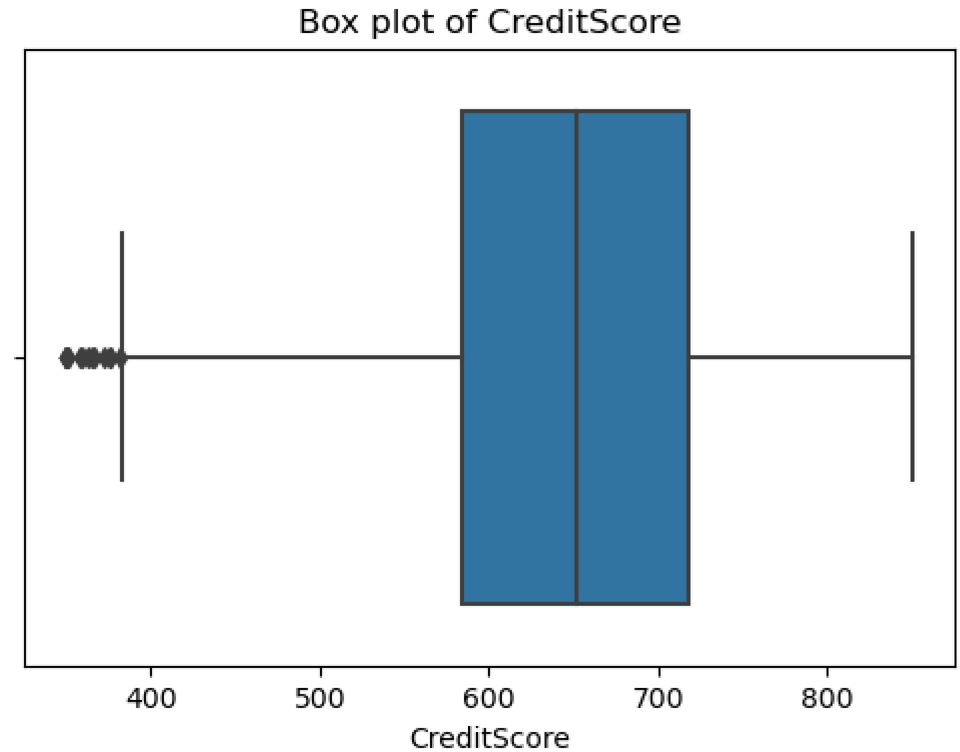
Churn Rate: Approximately 80% of customers did not churn, indicating a high retention rate among the bank's clientele.

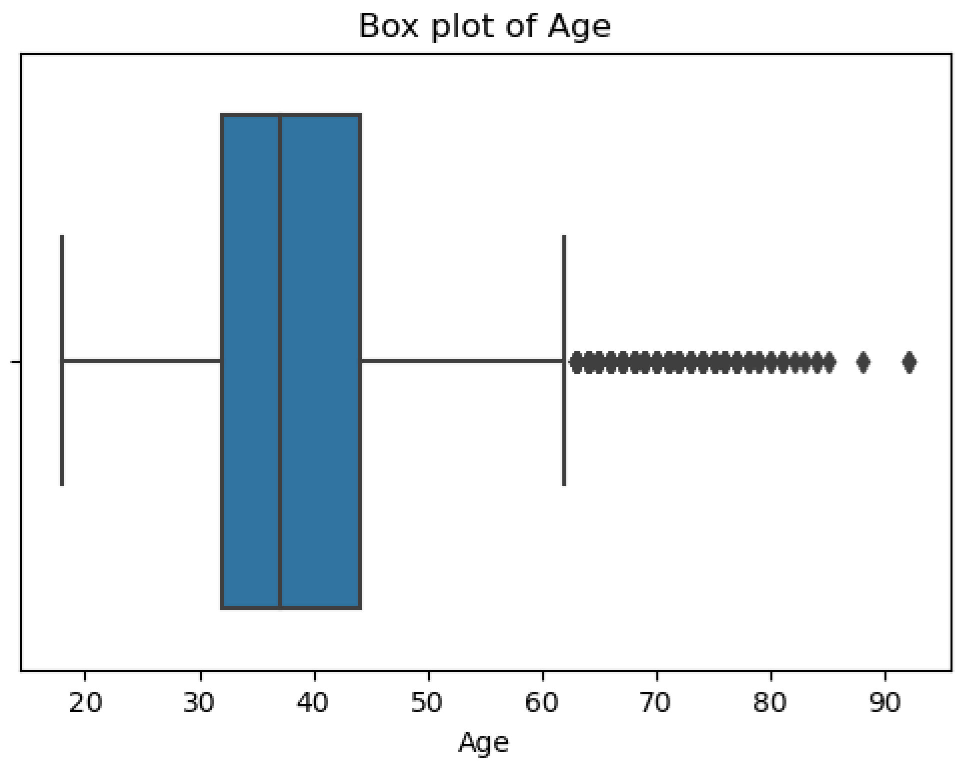
Satisfaction Score Distribution: The distribution of satisfaction scores appears consistent, with no significant deviations observed across different satisfaction levels.

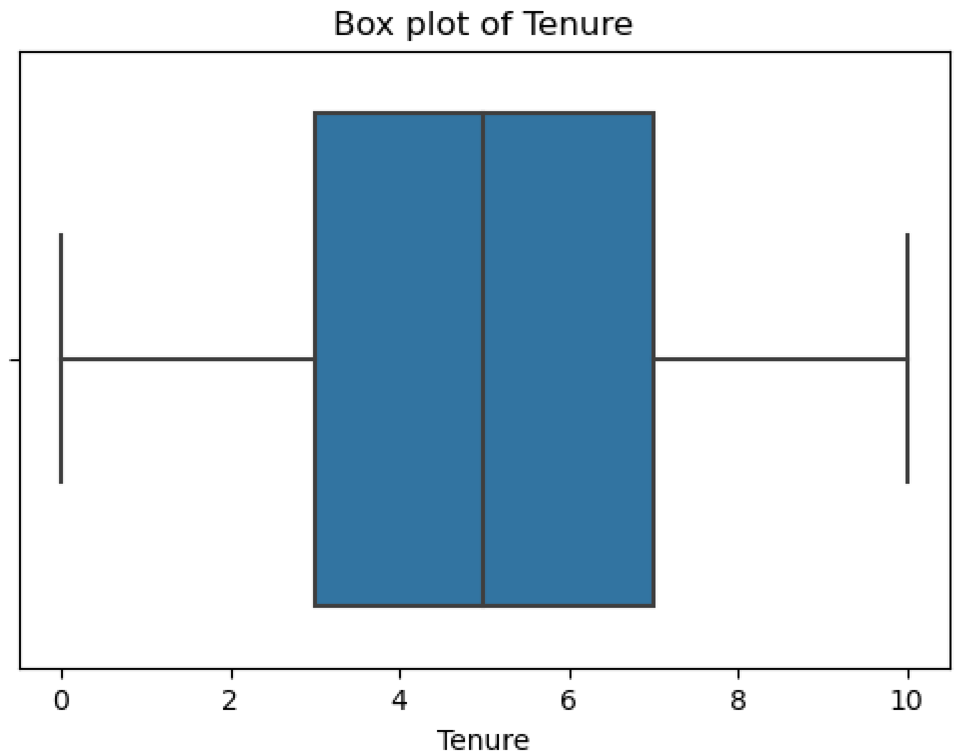
Points Earned Distribution: The distribution of points earned from credit card usage ranges from 300 to 1000, suggesting varying levels of engagement with the bank's loyalty program among customers.

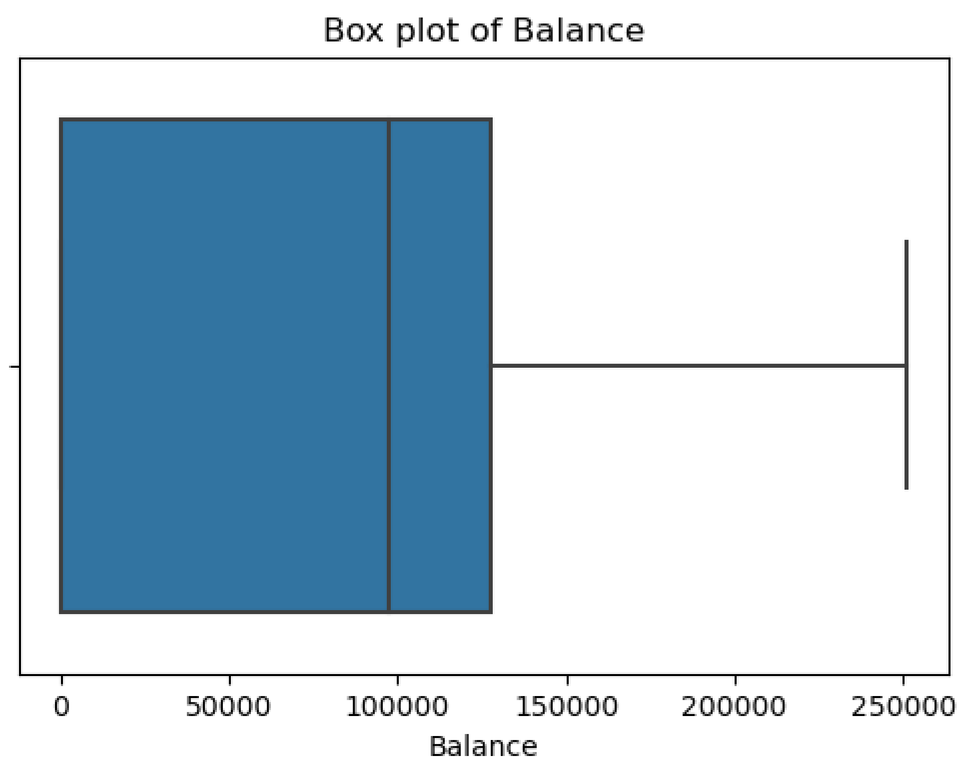
b) **Handling Outliers:** Detecting and handling outliers is a very crucial step in EDA. Now we are going to perform outlier detection which will head our way for further approach.

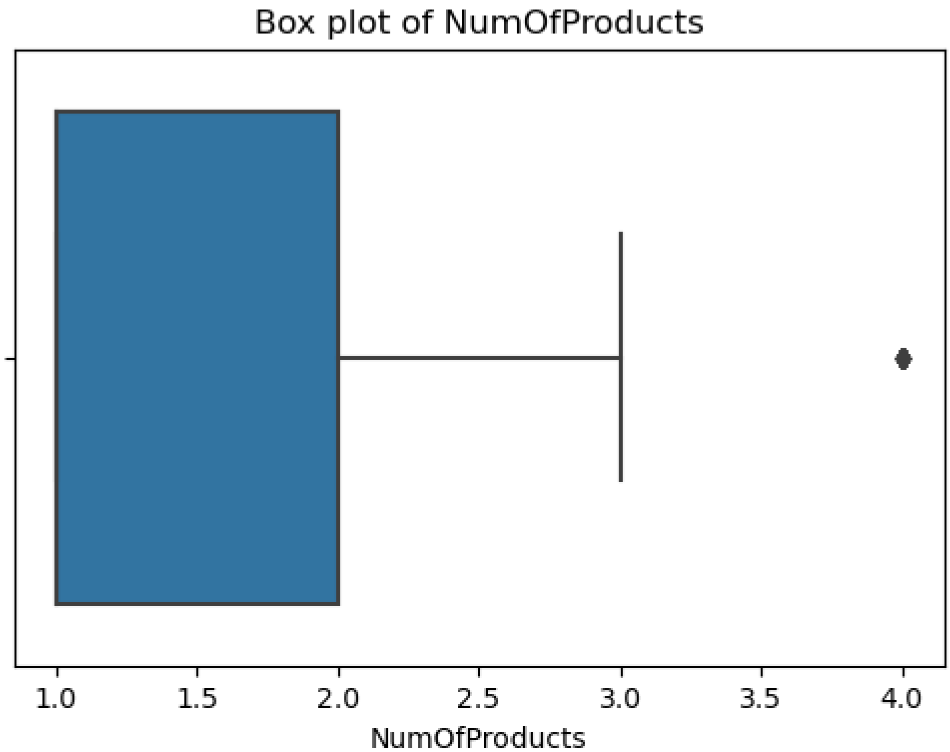
```
# Create box plots for each numerical column
for column in numerical_columns:
    plt.figure(figsize=(6, 4))
    sns.boxplot(x=data[column])
    plt.title(f'Box plot of {column}')
    plt.show()
```

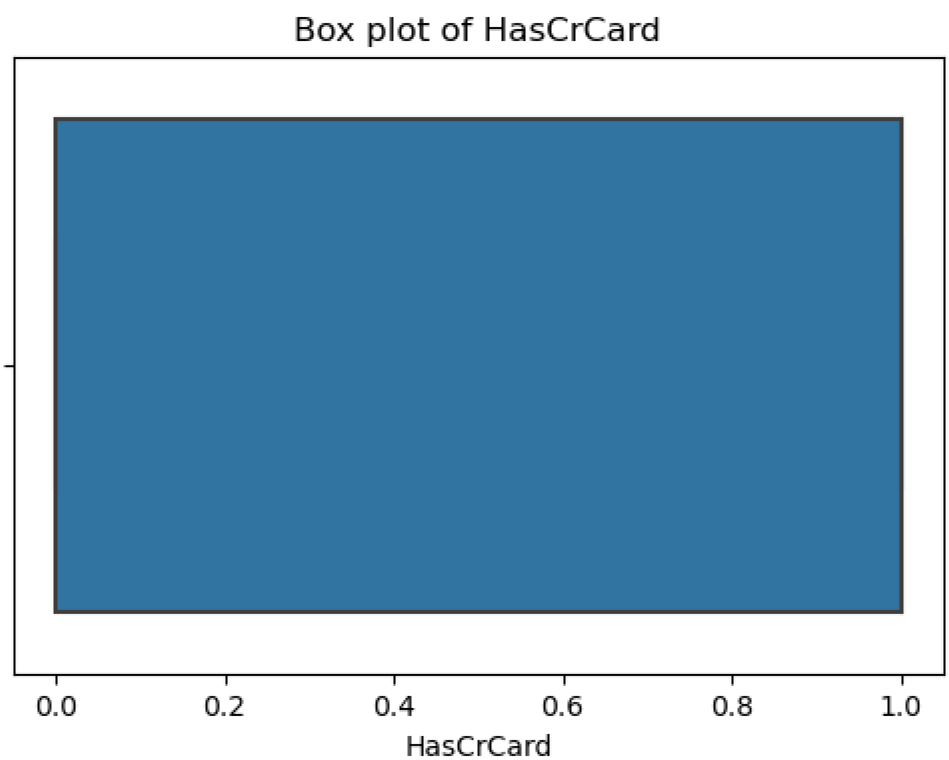

Box plot of CreditScore
400
500
600
700
800
CreditScore


Box plot of Age
20
30
40
50
60
70
80
90
Age

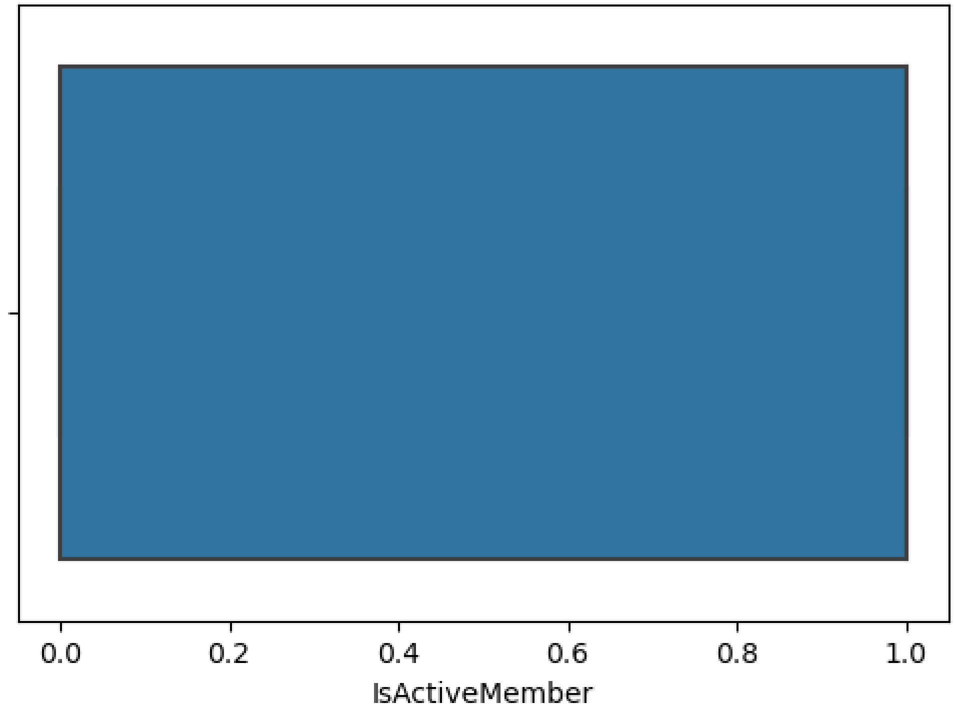

Box plot of Tenure
0
2
4
6
8
10
Tenure


Box plot of Balance
0
50000
100000
150000
200000
250000
Balance

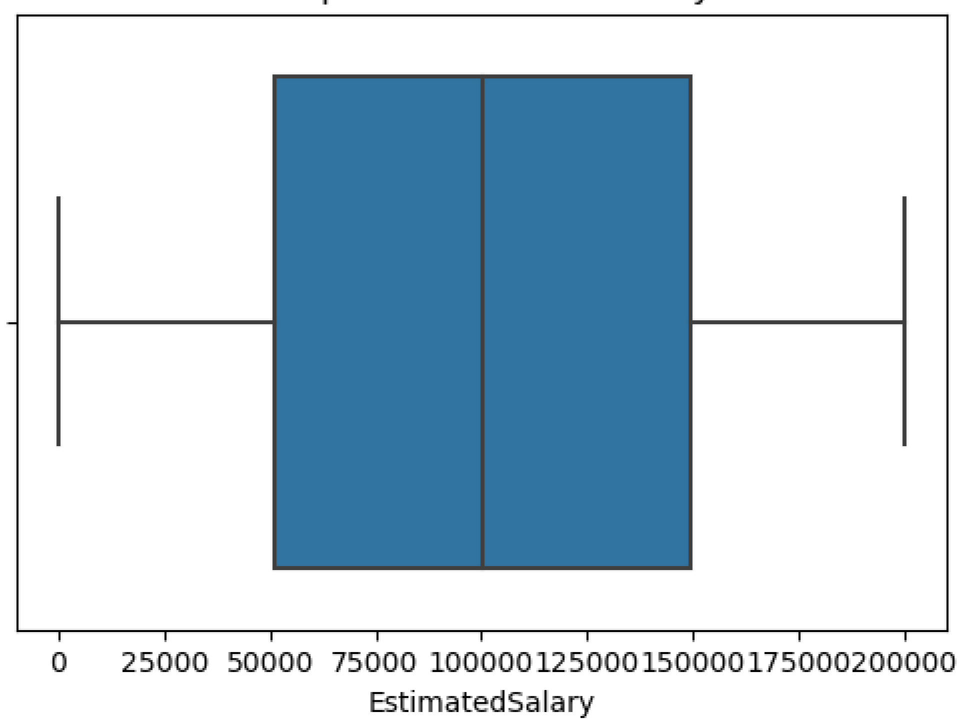

Box plot of NumOfProducts
1.0
1.5
2.0
2.5
3.0
3.5
4.0
NumOfProducts

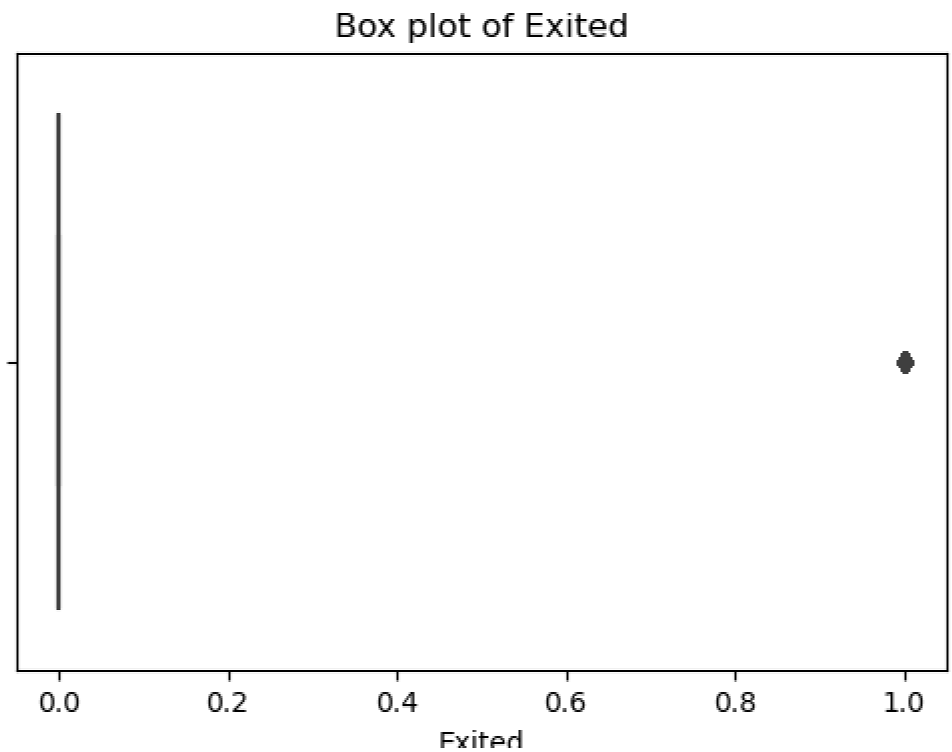

Box plot of HasCrCard
0.0
0.2
0.4
0.6
0.8
1.0
HasCrCard

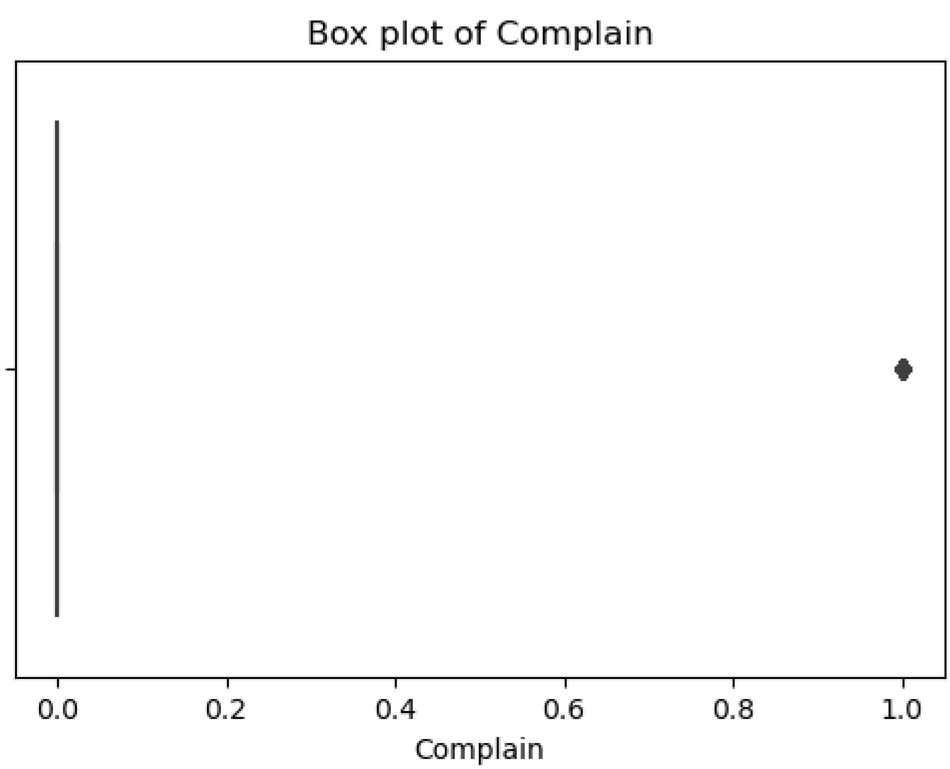
Box plot of IsActiveMember

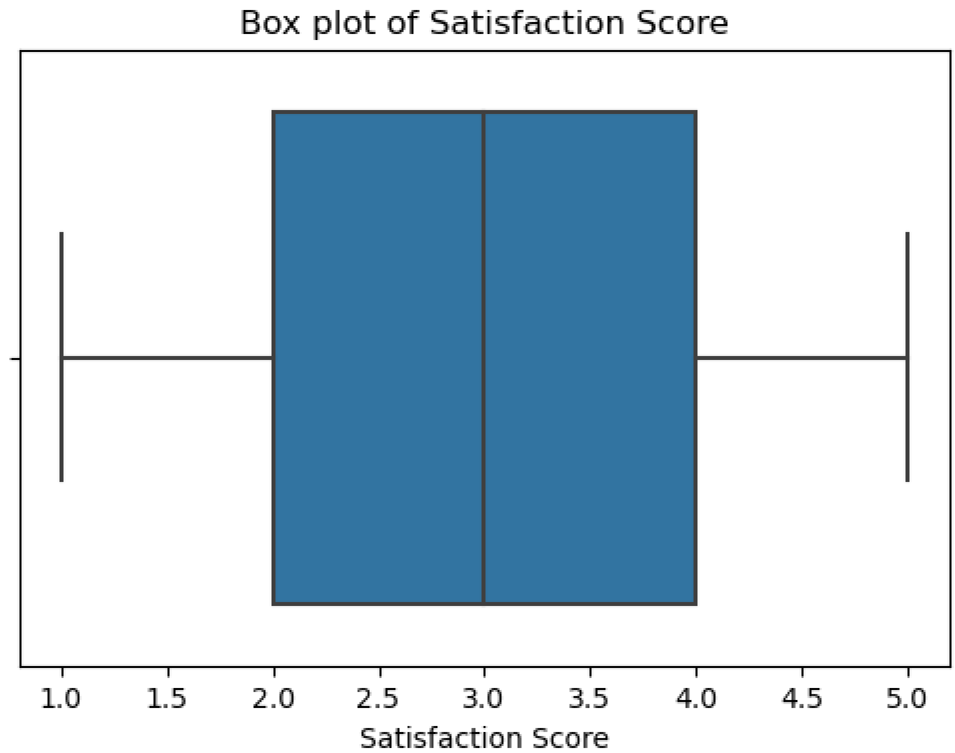


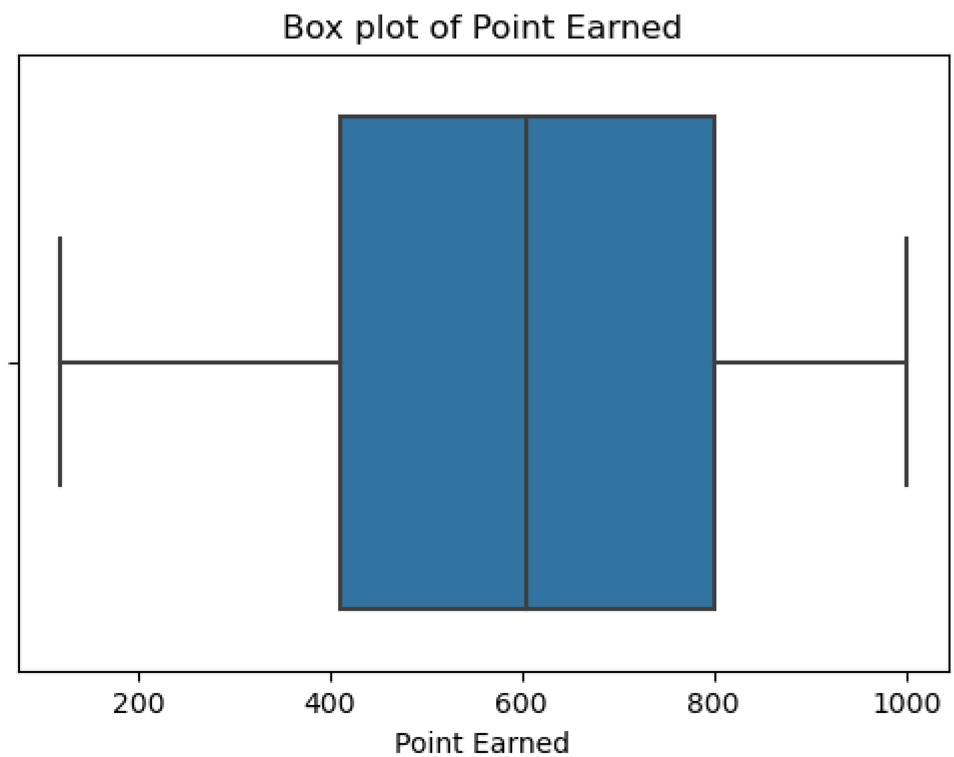
Box plot of EstimatedSalary




Box plot of Exited
0.0
0.2
0.4
0.6
0.8
1.0
Exited


Box plot of Complain
0.0
0.2
0.4
0.6
0.8
1.0
Complain





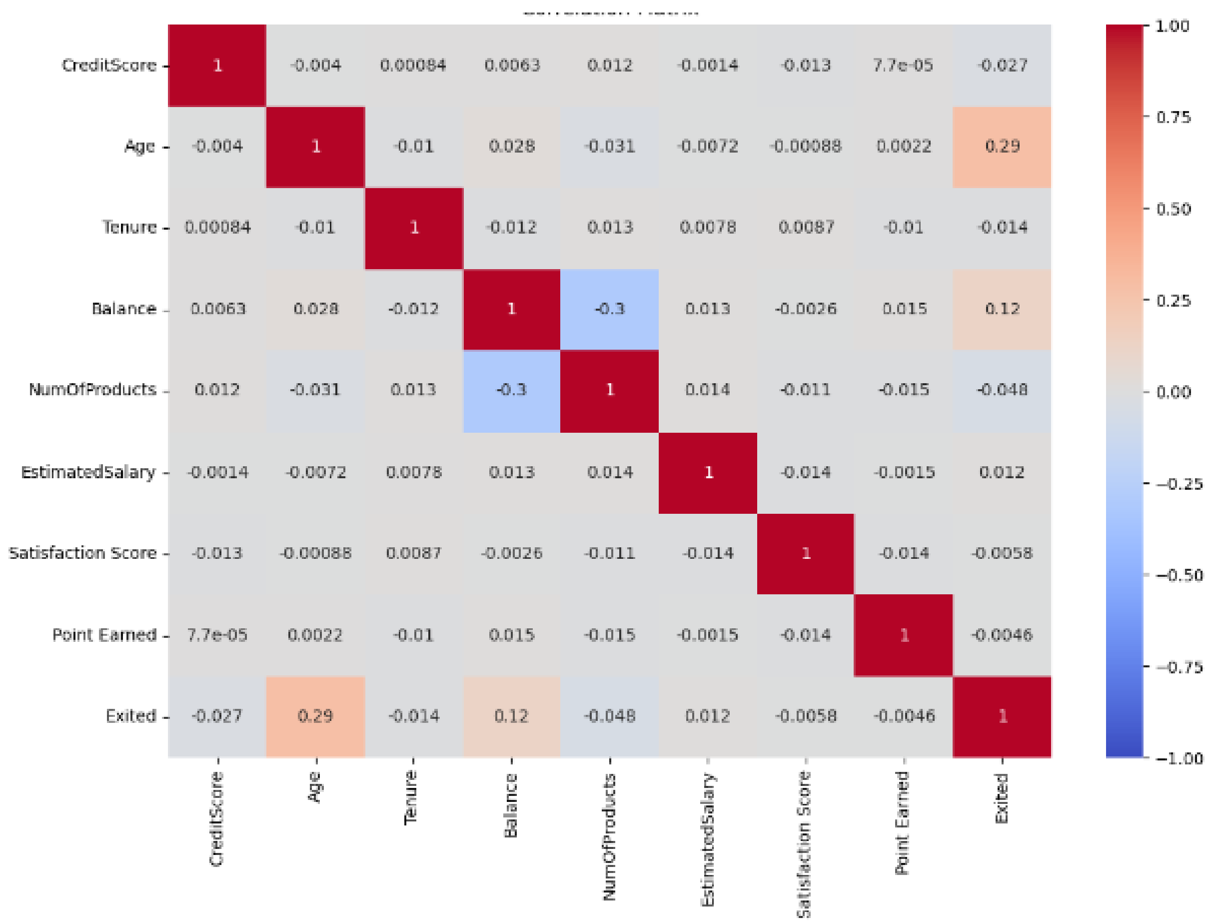
Insights: The outliers present in the dataset are useful for deriving the insights, so we are not going to remove them.

6. **EDA(Exploratory Data Analysis):** Now we are going to perform EDA on our dataset to gain more information.

a) **Correlation Analysis:** By examining the correlation coefficients, you can identify which numerical features are potentially strong predictors of customer churn **(Exited).** This correlation analysis is a crucial step in understanding the relationships between features and the target variable, helping to identify potential predictors of churn.

```
# List of numerical columns to check for correlation
numerical_columns = ['CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', '

# Add the target column 'Exited' to the list
numerical_columns_with_target = numerical_columns + ['Exited']

# Calculate the correlation matrix
correlation_matrix = data[numerical_columns_with_target].corr()

# Display the correlation matrix
print(correlation_matrix)

# Visualize the correlation matrix using a heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)
plt.title('Correlation Matrix')
plt.show()
```

```
                    CreditScore       Age    Tenure   Balance  NumOfProducts  \
CreditScore            1.000000 -0.003965  0.000842  0.006268       0.012238
Age                   -0.003965  1.000000 -0.009997  0.028308      -0.030680
Tenure                 0.000842 -0.009997  1.000000 -0.012254       0.013444
Balance                0.006268  0.028308 -0.012254  1.000000      -0.304180
NumOfProducts          0.012238 -0.030680  0.013444 -0.304180       1.000000
EstimatedSalary       -0.001384 -0.007201  0.007784  0.012797       0.014204
Satisfaction Score    -0.012599 -0.000876  0.008663 -0.002588      -0.011394
Point Earned           0.000077  0.002222 -0.010196  0.014608      -0.015330
Exited                -0.026771  0.285296 -0.013656  0.118577      -0.047611

                    EstimatedSalary  Satisfaction Score  Point Earned  \
CreditScore               -0.001384           -0.012599      0.000077
Age                       -0.007201           -0.000876      0.002222
Tenure                     0.007784            0.008663     -0.010196
Balance                    0.012797           -0.002588      0.014608
NumOfProducts              0.014204           -0.011394     -0.015330
EstimatedSalary            1.000000           -0.013747     -0.001515
Satisfaction Score        -0.013747            1.000000     -0.014400
Point Earned              -0.001515           -0.014400      1.000000
Exited                     0.012490           -0.005849     -0.004628

                      Exited
CreditScore        -0.026771
Age                 0.285296
Tenure             -0.013656
Balance             0.118577
NumOfProducts      -0.047611
EstimatedSalary     0.012490
Satisfaction Score -0.005849
```



Insights:

- **Age** and **Balance** are the most significant predictors of customer churn.
- Other variables show minimal impact on predicting churn.

To improve customer retention, focus on strategies targeting older customers and those with higher balances.

b) **Customer Profile Analysis:** Segmented customers based on key demographics (Age, Geography, Gender) to identify which groups are more likely to churn. Using **bivariate analysis** represented it as visuals.

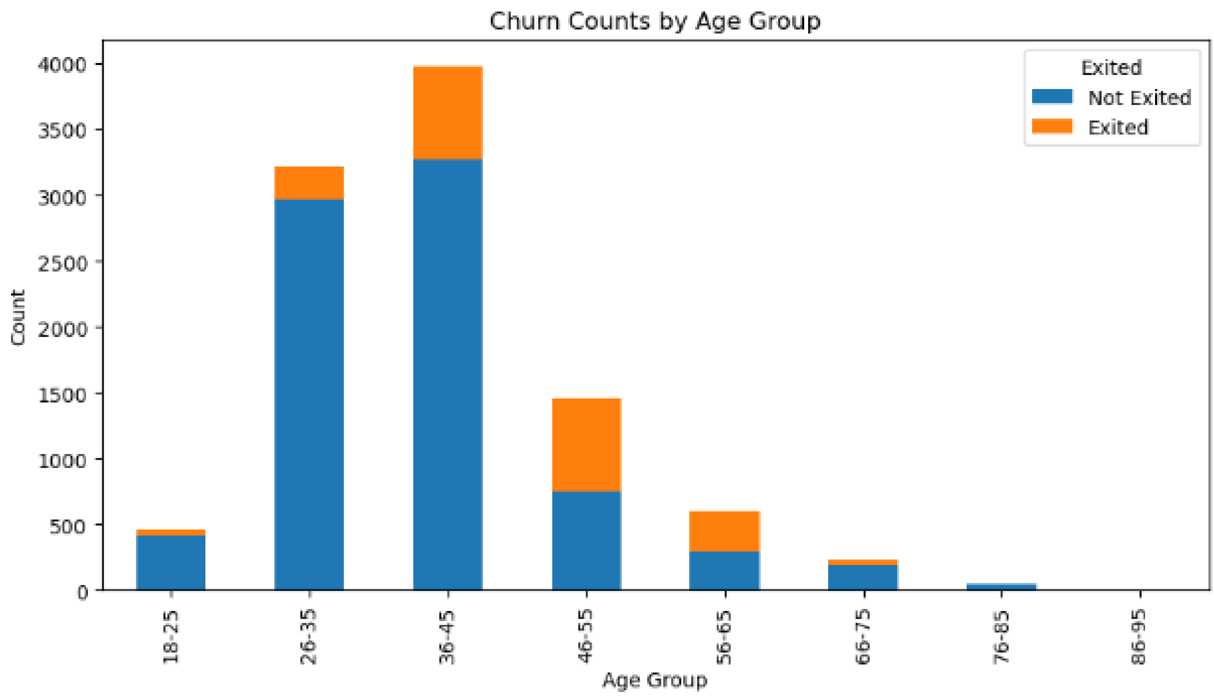
```python
# Define age groups
bins = [18, 25, 35, 45, 55, 65, 75, 85, 95]
labels = ['18-25', '26-35', '36-45', '46-55', '56-65', '66-75', '76-85', '86-95']
data['AgeGroup'] = pd.cut(data['Age'], bins=bins, labels=labels, right=False)

# Calculate counts for each age group
age_group_counts = data.groupby(['AgeGroup', 'Exited']).size().unstack()
print(age_group_counts)
```

```
Exited       0    1
AgeGroup
18-25      417   40
26-35     2972  250
36-45     3277  704
46-55      756  702
56-65      301  299
66-75      186   42
76-85       49    1
86-95        4    0
```

```python
age_group_churn = data.groupby('AgeGroup')['Exited'].mean().sort_values()
print(age_group_churn)
```

```
AgeGroup
86-95    0.000000
76-85    0.020000
26-35    0.077592
18-25    0.087527
36-45    0.176840
66-75    0.184211
46-55    0.481481
56-65    0.498333
Name: Exited, dtype: float64
```

```python
# Bivariate Analysis
# Plot counts by Age Group
age_group_counts.plot(kind='bar', stacked=True, figsize=(10, 5))
plt.title('Churn Counts by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Count')
plt.legend(title='Exited', labels=['Not Exited', 'Exited'])
plt.show()
```

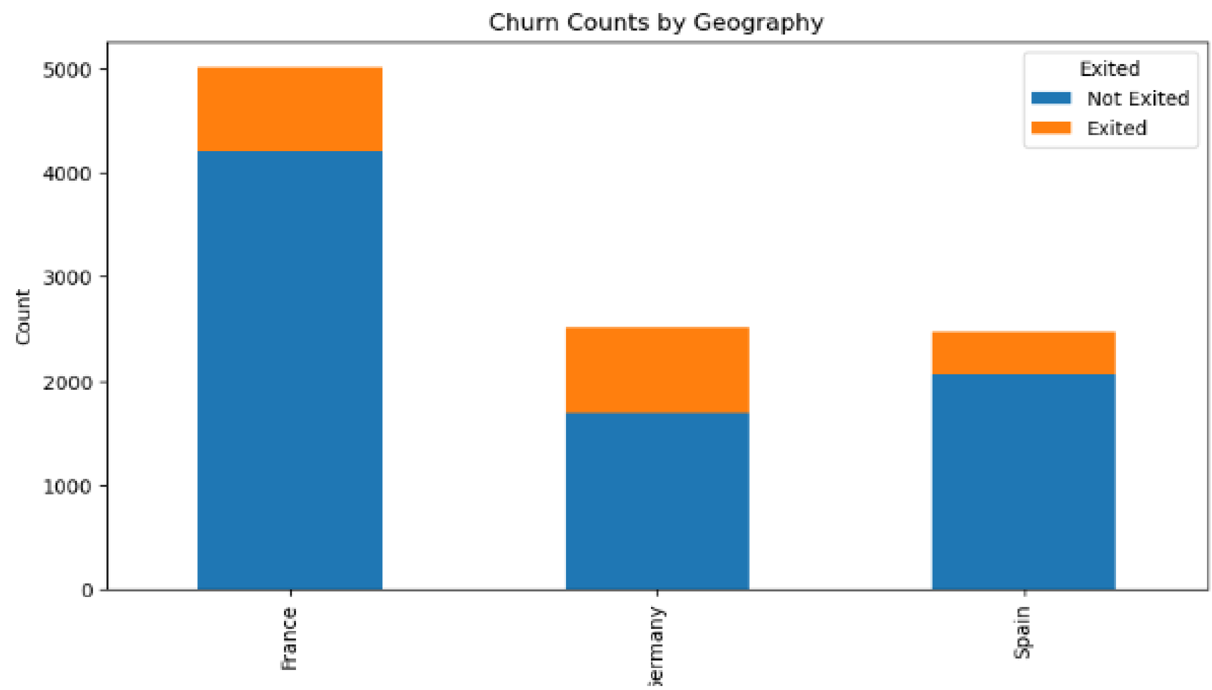



```python
# Calculate counts for each geographical location
geography_counts = data.groupby(['Geography', 'Exited']).size().unstack()
print(geography_counts)
```

```
Exited        0    1
Geography
France     4203  811
Germany    1695  814
Spain      2064  413
```

```python
geo_group_churn = data.groupby('Geography')['Exited'].mean().sort_values()
print(geo_group_churn)
```

```
Geography
France     0.161747
Spain      0.166734
Germany    0.324432
Name: Exited, dtype: float64
```



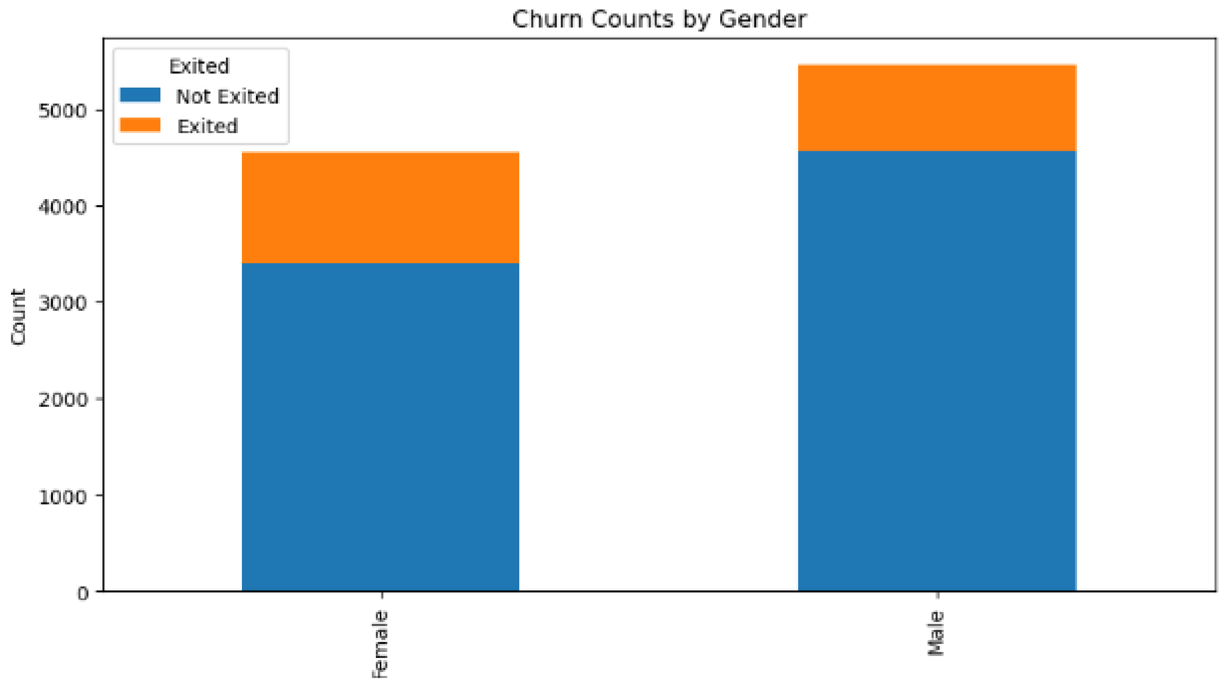

```python
# Calculate counts for each gender
gender_counts = data.groupby(['Gender', 'Exited']).size().unstack()
print(gender_counts)
```

```
Exited     0     1
Gender
Female  3404  1139
Male    4558   899
```

```python
gender_group_churn = data.groupby('Gender')['Exited'].mean().sort_values()
print(gender_group_churn)
```

```
Gender
Male      0.164743
Female    0.250715
Name: Exited, dtype: float64
```

```python
# Plot counts by Gender
gender_counts.plot(kind='bar', stacked=True, figsize=(10, 5))
plt.title('Churn Counts by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.legend(title='Exited', labels=['Not Exited', 'Exited'])
plt.show()
```



Insights:

By Age-

- 46-65 i.e mid- age group show higher exit counts compared to younger and older age groups with around 4.9% of churn percentage.
- Bank should focus on the factors why this age group is churning and provide them better services.

By Geography-

- Germany has the highest churning rate though France has the highest number of customer.
- Germany is the country with second highest customers of bank so Bank should focus on retaining the customers from Germany and give them good services.

By Gender-

- The churn rate of female customer is higher than men.
- Bank should provide some special schemes or offers for female so that they don’t churn.

7. COMPARATIVE ANALYSIS: In Comparative Analysis we are going to statistically compare the churn with geography, gender and age.

a) Churn by Geography: For churn by geography differences in churn analyses using the Chi-square test, we can define the null and alternative hypotheses as follows:

- H0: There is no association between geography and customer churn. In other words, the churn rate is independent of the geographical location.
- H1: There is an association between geography and customer churn. In other words, the churn rate depends on the geographical location.

```
# Assume aplha 0.05.
from scipy.stats import chi2_contingency

# Churn by Geography
contingency_geo = pd.crosstab(data['Geography'], data['Exited'])
chi2_geo, p_geo, dof_geo, expected_geo = chi2_contingency(contingency_geo)
print("Chi-square statistic for churn by geography:", chi2_geo)
print("P-value for churn by geography:", format_p_value(p_geo))

if p_geo < 0.05:
    print("Reject the null hypothesis: There is an association between geography and c
else:
    print("Fail to reject the null hypothesis: There is no association between geograp
```

```
Chi-square statistic for churn by geography: 300.6264011211942
P-value for churn by geography: 0.0
Reject the null hypothesis: There is an association between geography and customer ch
urn.
```

Insights: Since the p-value is less than 0.05, we reject the null hypothesis. There is an association between geography and customer churn.

b) Churn by Gender: For churn by gender differences in churn analyses using the Chi-square test, we can define the null and alternative hypotheses as follows:

- H0: There is no association between gender and customer churn. In other words, the churn rate is independent of gender.
- H1: There is an association between gender and customer churn. In other words, the churn rate depends on gender.

```python
# Gender Differences in Churn
contingency_gender = pd.crosstab(data['Gender'], data['Exited'])
chi2_gender, p_gender, dof_gender, expected_gender = chi2_contingency(contingency_gend
print("Chi-square statistic for gender differences in churn:", chi2_gender)
print("P-value for gender differences in churn:", format_p_value(p_gender))

if p_gender < 0.05:
    print("Reject the null hypothesis: There is an association between gender and cust
else:
    print("Fail to reject the null hypothesis: There is no association between gender
```

```
Chi-square statistic for gender differences in churn: 112.39655374778587
P-value for gender differences in churn: 0.0
Reject the null hypothesis: There is an association between gender and customer chur
n.
```

Insights: Since the p-value is less than 0.05, we reject the null hypothesis. There is an association between gender and customer churn.

c) Churn by Age: For churn by age differences in churn analyses using the Chi-square test, we can define the null and alternative hypotheses as follows:

- H0: There is no association between age and customer churn. In other words, the churn rate is independent of age.
- H1: There is an association between age and customer churn. In other words, the churn rate depends on age.

```python
# Gender Differences in Churn
contingency_age = pd.crosstab(data['AgeGroup'], data['Exited'])
chi2_age, p_age, dof_age, expected_age = chi2_contingency(contingency_age)
print("Chi-square statistic for gender differences in churn:", chi2_age)
print("P-value for gender differences in churn:", format_p_value(p_age))

if p_gender < 0.05:
    print("Reject the null hypothesis: There is an association between age and custome
else:
    print("Fail to reject the null hypothesis: There is no association between age and
```

```
Chi-square statistic for gender differences in churn: 1397.7608390828918
P-value for gender differences in churn: 0.0
Reject the null hypothesis: There is an association between age and customer churn.
```

Insights: Since the p-value is less than 0.05, we reject the null hypothesis. There is an association between gender and customer churn.

8. BEHAVIORAL ANALYSIS:

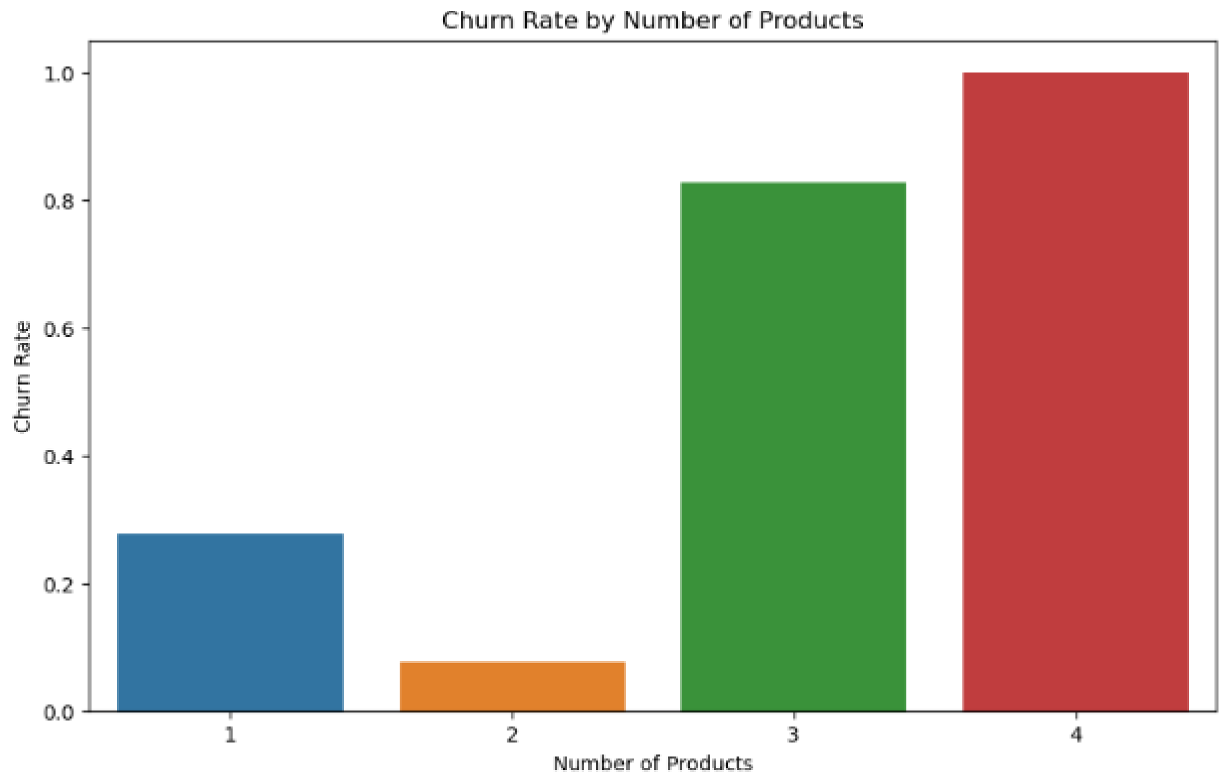
To conduct a behavioral analysis on how the number of products a customer uses and their activity level affects their likelihood to churn, we can use both descriptive statistics and hypothesis testing.

a) Product and services usage: We are going to examine how the number of products (NumOfProducts) a customer uses affects their likelihood to churn.

- H0: There is no association between product and customer churn. In other words, the churn rate is independent of product.
- H1: There is an association between product and customer churn. In other words, the churn rate depends on product.

```
# Descriptive Analysis
product_churn_rate = data.groupby('NumOfProducts')['Exited'].mean()
print("Churn rate by number of products:\n", product_churn_rate)

# Hypothesis Testing - Chi-Square Test
contingency_products = pd.crosstab(data['NumOfProducts'], data['Exited'])
chi2_products, p_products, dof_products, expected_products = chi2_contingency(continge


print("Chi-square statistic for number of products and churn:", chi2_products)
print("P-value for number of products and churn:", format_p_value(p_products))

if p_products < 0.05:
    print("Reject the null hypothesis: There is an association between the number of p
else:
    print("Fail to reject the null hypothesis: There is no association between the num
```

```
Churn rate by number of products:
 NumOfProducts
1    0.277144
2    0.076035
3    0.827068
4    1.000000
Name: Exited, dtype: float64
Chi-square statistic for number of products and churn: 1501.5048306588592
P-value for number of products and churn: 0.0
Reject the null hypothesis: There is an association between the number of products an
d customer churn.
```



Insights:

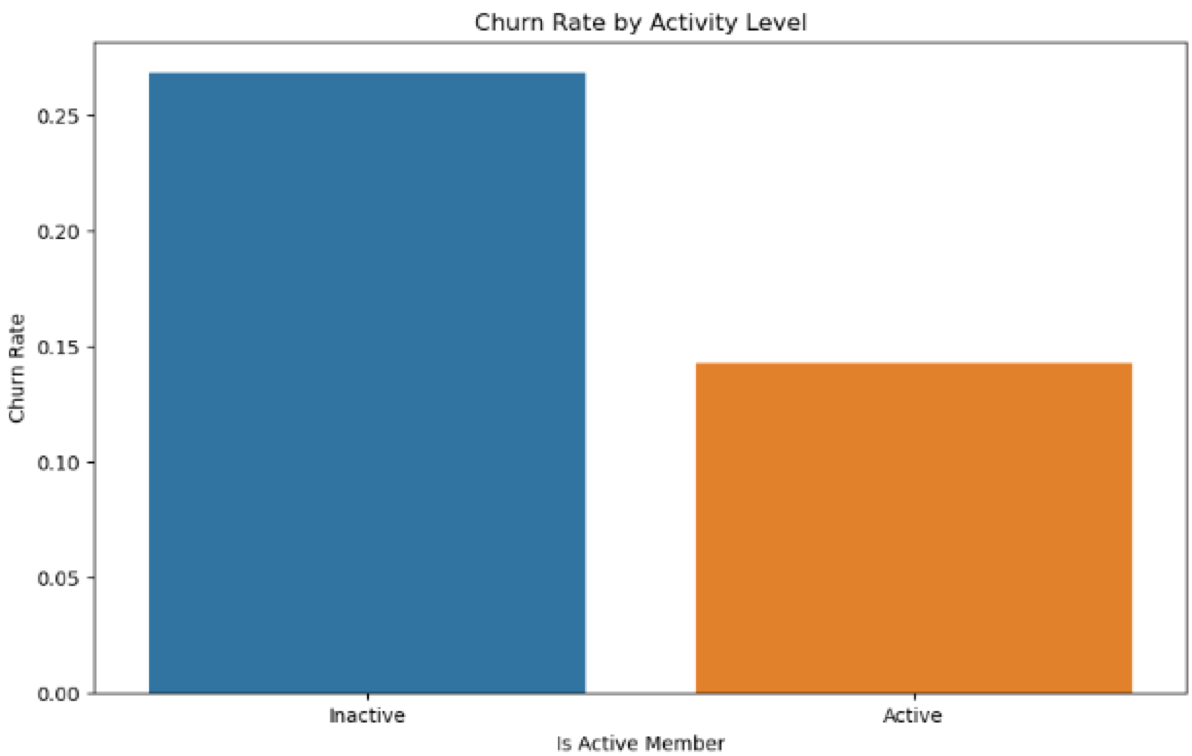
- **S**ince the p-value is less than 0.05, we reject the null hypothesis. There is an association between product services and customer churn.
- Product 4 is more likely to be churned. So bank must either try to avoid giving product 4 or try to improve product 4.

b) Activity Level: We are going to investigate the relationship between being an IsActiveMember and customer churn.

- H0: There is no association between product and customer churn. In other words, the churn rate is independent of product.
- H1: There is an association between product and customer churn. In other words, the churn rate depends on product.

```
# Descriptive Analysis
activity_churn_rate = data.groupby('IsActiveMember')['Exited'].mean()
print("Churn rate by activity level:\n", activity_churn_rate)

# Hypothesis Testing - Chi-Square Test
contingency_activity = pd.crosstab(data['IsActiveMember'], data['Exited'])
chi2_activity, p_activity, dof_activity, expected_activity = chi2_contingency(continge

print("Chi-square statistic for activity level and churn:", chi2_activity)
print("P-value for activity level and churn:", p_activity)

if p_activity < 0.05:
    print("Reject the null hypothesis: There is an association between activity level
else:
    print("Fail to reject the null hypothesis: There is no association between activit
```

```
Churn rate by activity level:
 IsActiveMember
0    0.268715
1    0.142691
Name: Exited, dtype: float64
Chi-square statistic for activity level and churn: 243.6948024819593
P-value for activity level and churn: 6.153167438113408e-55
Reject the null hypothesis: There is an association between activity level and custom
er churn.
```



Insights:

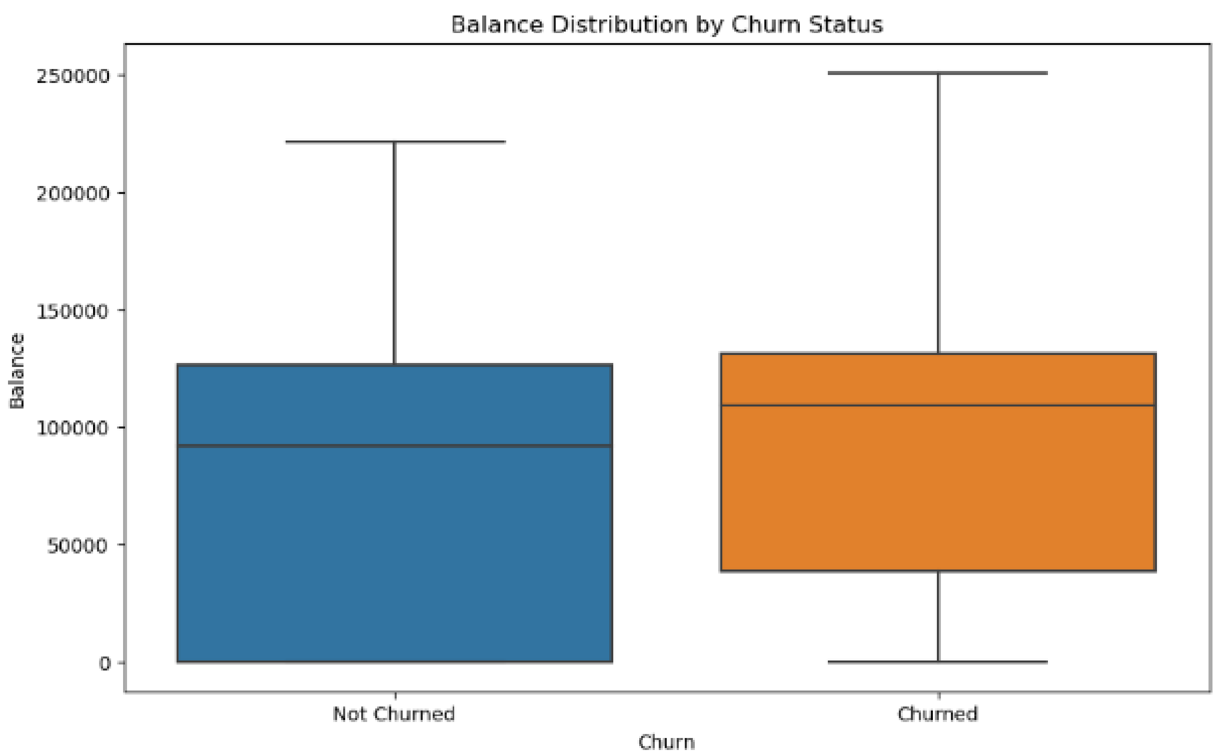
- Since the p-value is less than 0.05, we reject the null hypothesis. There is an association between activity and customer churn.
- Inactive customers are more likely to churn. So bank must either try to keep their customers engaged and active.

9. FINANCIAL ANALYSIS:

a) Balance vs Churn:

To analyze how customer balance levels correlate with churn rates, we can start with descriptive statistics and visualization, followed by a statistical test if needed.

```python
# Visualization
plt.figure(figsize=(10,6))
sns.boxplot(x='Exited', y='Balance', data=data)
plt.title('Balance Distribution by Churn Status')
plt.xlabel('Churn')
plt.ylabel('Balance')
plt.xticks([0, 1], ['Not Churned', 'Churned'])
plt.show()
```



- H0: There is no association between balance and customer churn. In other words, the churn rate is independent of balance.
- H1: There is an association between balance and customer churn. In other words, the churn rate depends on balance.

```python
from scipy.stats import mannwhitneyu

# Separate the balance data into churned and not churned
balance_churned = data[data['Exited'] == 1]['Balance']
balance_not_churned = data[data['Exited'] == 0]['Balance']

# Mann-Whitney U test
stat, p_value = mannwhitneyu(balance_churned, balance_not_churned)
formatted_p_value = f"{p_value:.3f}"

print("Mann-Whitney U test statistic for balance and churn:", stat)
print("P-value for balance and churn:", formatted_p_value)

if p_value < 0.05:
    print("Reject the null hypothesis: There is a significant difference in balances b
else:
    print("Fail to reject the null hypothesis: There is no significant difference in b
```

```
Mann-Whitney U test statistic for balance and churn: 9373909.5
P-value for balance and churn: 0.000
Reject the null hypothesis: There is a significant difference in balances between cus
tomers who churned and those who didn't.
```

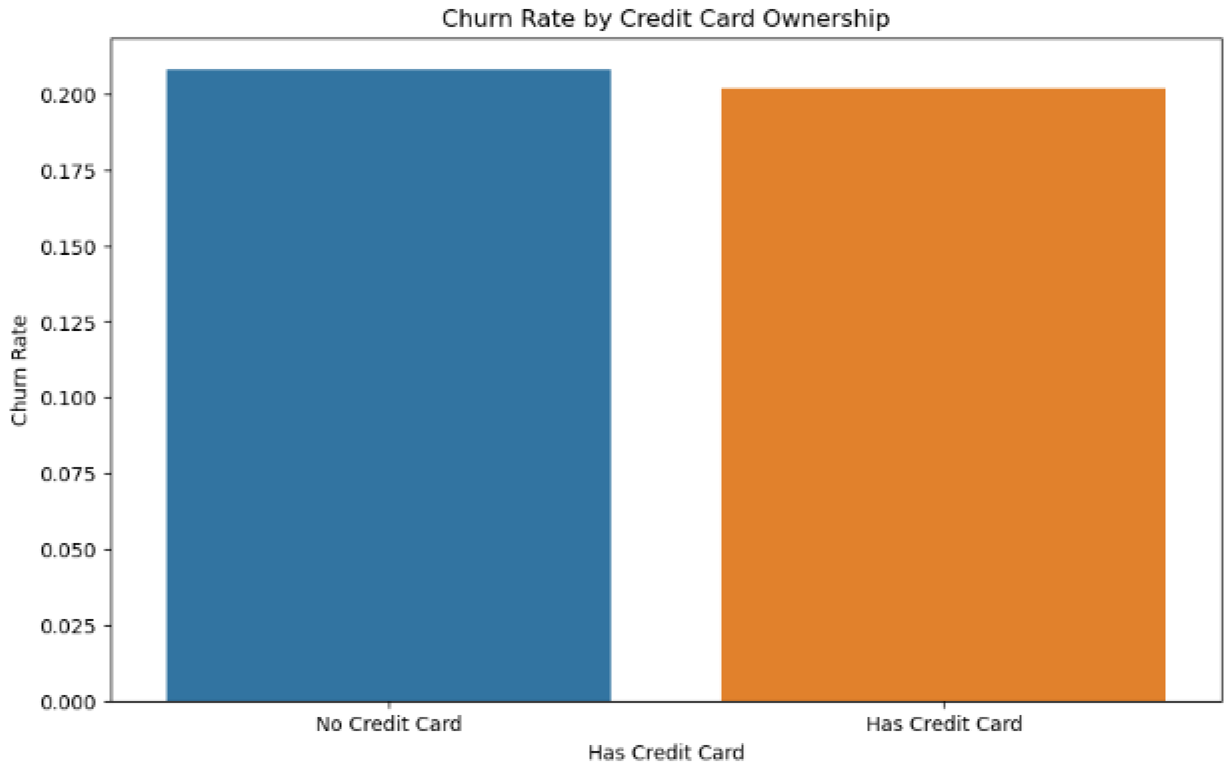
Insights:

- **S**ince the p-value is less than 0.05, we reject the null hypothesis. There is an association between balance and customer churn.
- The people maintaining more balance are more likely to churn. So keeping this balance constraint must be taken care off.

b) Credit Card Ownership: We will determine if owning a credit card (HasCrCard) impacts customer loyalty.

```python
# Descriptive Analysis
credit_card_churn_rate = data.groupby('HasCrCard')['Exited'].mean()
print("Churn rate by credit card ownership:\n", credit_card_churn_rate)

# Visualization
plt.figure(figsize=(10,6))
sns.barplot(x=[0, 1], y=credit_card_churn_rate.values)
plt.title('Churn Rate by Credit Card Ownership')
plt.xlabel('Has Credit Card')
plt.ylabel('Churn Rate')
plt.xticks([0, 1], ['No Credit Card', 'Has Credit Card'])
plt.show()
```

```
Churn rate by credit card ownership:
 HasCrCard
0    0.208149
1    0.201984
Name: Exited, dtype: float64
```



- H0: There is no association between has credit card and customer churn. In other words, the churn rate is independent of has credit card.
- H1: There is an association between has credit card and customer churn. In other words, the churn rate depends on has credit card.

```python
from scipy.stats import chi2_contingency

# Contingency table
contingency_crcard = pd.crosstab(data['HasCrCard'], data['Exited'])

# Chi-Square Test
chi2_crcard, p_crcard, dof_crcard, expected_crcard = chi2_contingency(contingency_crca
formatted_p_crcard = f"{p_crcard:.3f}"

print("Chi-square statistic for credit card ownership and churn:", chi2_crcard)
print("P-value for credit card ownership and churn:", formatted_p_crcard)

if p_crcard < 0.05:
    print("Reject the null hypothesis: There is a significant association between cred
else:
    print("Fail to reject the null hypothesis: There is no significant association bet
```

```
Chi-square statistic for credit card ownership and churn: 0.4494039375253385
P-value for credit card ownership and churn: 0.503
Fail to reject the null hypothesis: There is no significant association between credi
t card ownership and customer churn.
```

Insights:

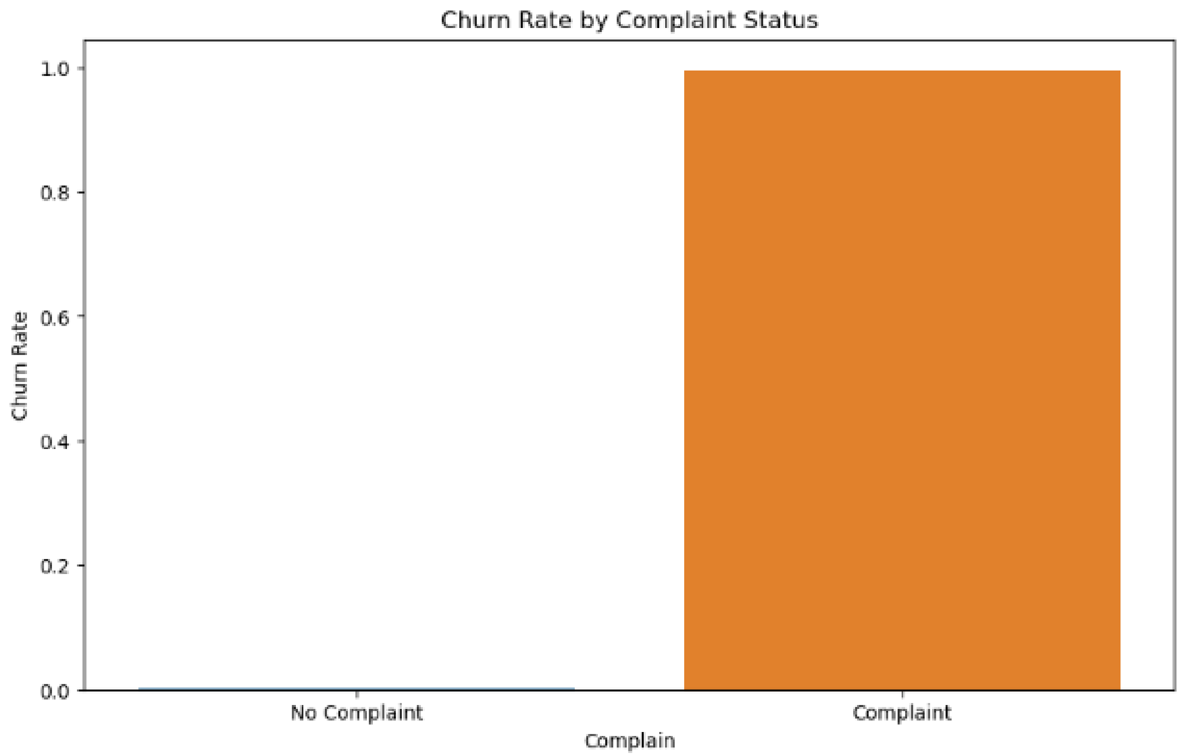
- Since the p-value is more than 0.05, we fail to reject the null hypothesis. There is no association between has credit card and customer churn.
- There is no impact of customers having credit card and not having card on the churn rate. So this can be ignored.

10. CUSTOMER SATISFACTION AND FEEDBACK:

a) Complaint Analysis: We are going to study the impact of having a complaint (Complain) on customer churn.

```python
# Descriptive Analysis
complaint_churn_rate = data.groupby('Complain')['Exited'].mean()
print("Churn rate by complaint status:\n", complaint_churn_rate)

# Visualization
plt.figure(figsize=(10,6))
sns.barplot(x=[0, 1], y=complaint_churn_rate.values)
plt.title('Churn Rate by Complaint Status')
plt.xlabel('Complain')
plt.ylabel('Churn Rate')
plt.xticks([0, 1], ['No Complaint', 'Complaint'])
plt.show()
```

```
Churn rate by complaint status:
 Complain
0    0.000503
1    0.995108
```



- H0: There is no association between complaint and customer churn. In other words, the churn rate is independent of complaint.
- H1: There is an association between complaint and customer churn. In other words, the churn rate depends on complaint.

```
from scipy.stats import chi2_contingency

# Contingency table
contingency_complain = pd.crosstab(data['Complain'], data['Exited'])

# Chi-Square Test
chi2_complain, p_complain, dof_complain, expected_complain = chi2_contingency(conting
formatted_p_complain = f"{p_complain:.3f}"

print("Chi-square statistic for complaint status and churn:", chi2_complain)
print("P-value for complaint status and churn:", formatted_p_complain)

if p_complain < 0.05:
    print("Reject the null hypothesis: There is a significant association between comp
else:
    print("Fail to reject the null hypothesis: There is no significant association bet
```

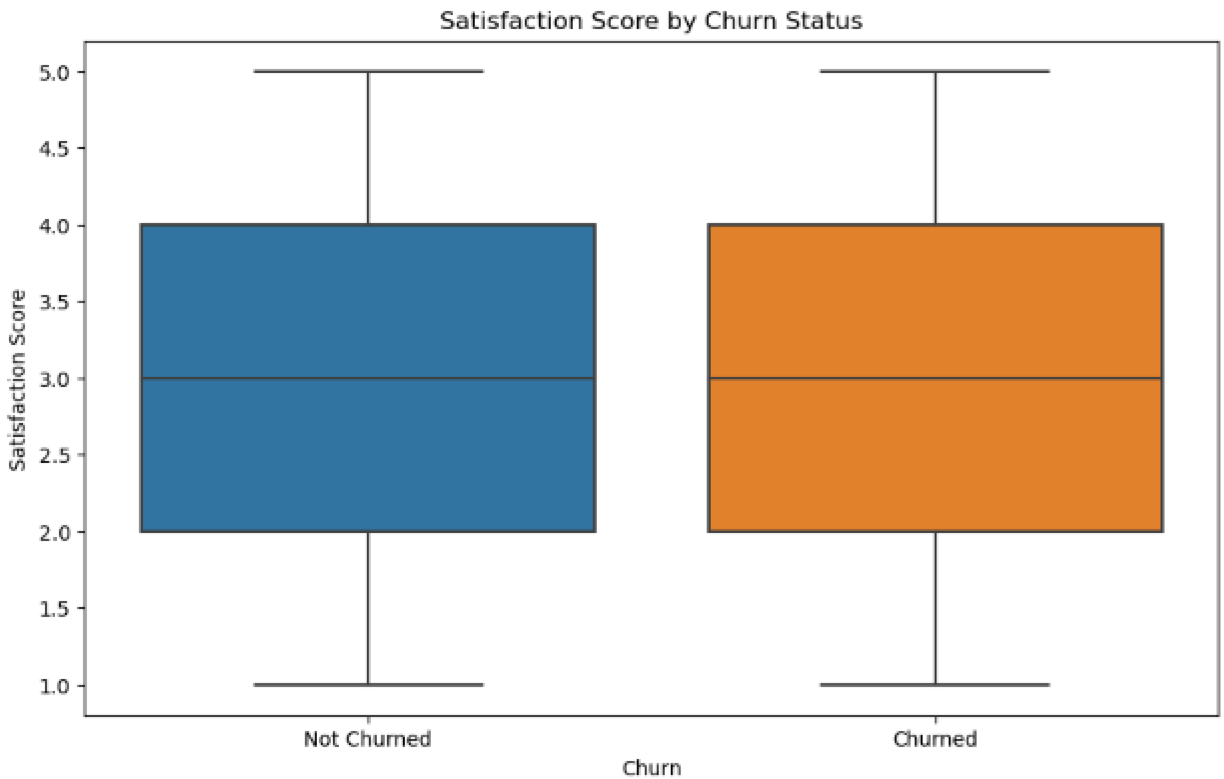
```
Chi-square statistic for complaint status and churn: 9907.907035880155
P-value for complaint status and churn: 0.000
Reject the null hypothesis: There is a significant association between complaints and
customer churn.
```

Insight:

- Since the p-value is less than 0.05, we reject the null hypothesis. There is a significant association between has complaint and customer churn.
- There is a significant impact of complaints on churn rate. Therefore, bank must provide good and satisfactory resolution to the complaints so as to retain the customers

b) Satisfaction and Churn: We are going to explore how the Satisfaction Score relates to churn, especially among those who have filed complaints.

```python
# Visualization
plt.figure(figsize=(10,6))
sns.boxplot(x='Exited', y='Satisfaction Score', data=data)
plt.title('Satisfaction Score by Churn Status')
plt.xlabel('Churn')
plt.ylabel('Satisfaction Score')
plt.xticks([0, 1], ['Not Churned', 'Churned'])
plt.show()
```



- H0: There is no association between satisfaction and customer churn. In other words, the churn rate is independent of satisfaction.
- H1: There is an association between satisfaction and customer churn. In other words, the churn rate depends on satisfaction.

```
# Separate the satisfaction scores into churned and not churned
satisfaction_churned = data[data['Exited'] == 1]['Satisfaction Score']
satisfaction_not_churned = data[data['Exited'] == 0]['Satisfaction Score']

# Mann-Whitney U test
stat_satisfaction, p_value_satisfaction = mannwhitneyu(satisfaction_churned, satisfact
formatted_p_value_satisfaction = f"{p_value_satisfaction:.3f}"

print("Mann-Whitney U test statistic for satisfaction score and churn:", stat_satisfac
print("P-value for satisfaction score and churn:", formatted_p_value_satisfaction)

if p_value_satisfaction < 0.05:
    print("Reject the null hypothesis: There is a significant difference in satisfacti
else:
    print("Fail to reject the null hypothesis: There is no significant difference in s
```

```
Mann-Whitney U test statistic for satisfaction score and churn: 8046002.0
P-value for satisfaction score and churn: 0.555
Fail to reject the null hypothesis: There is no significant difference in satisfactio
n scores between customers who churned and those who didn't.
```

Insights:

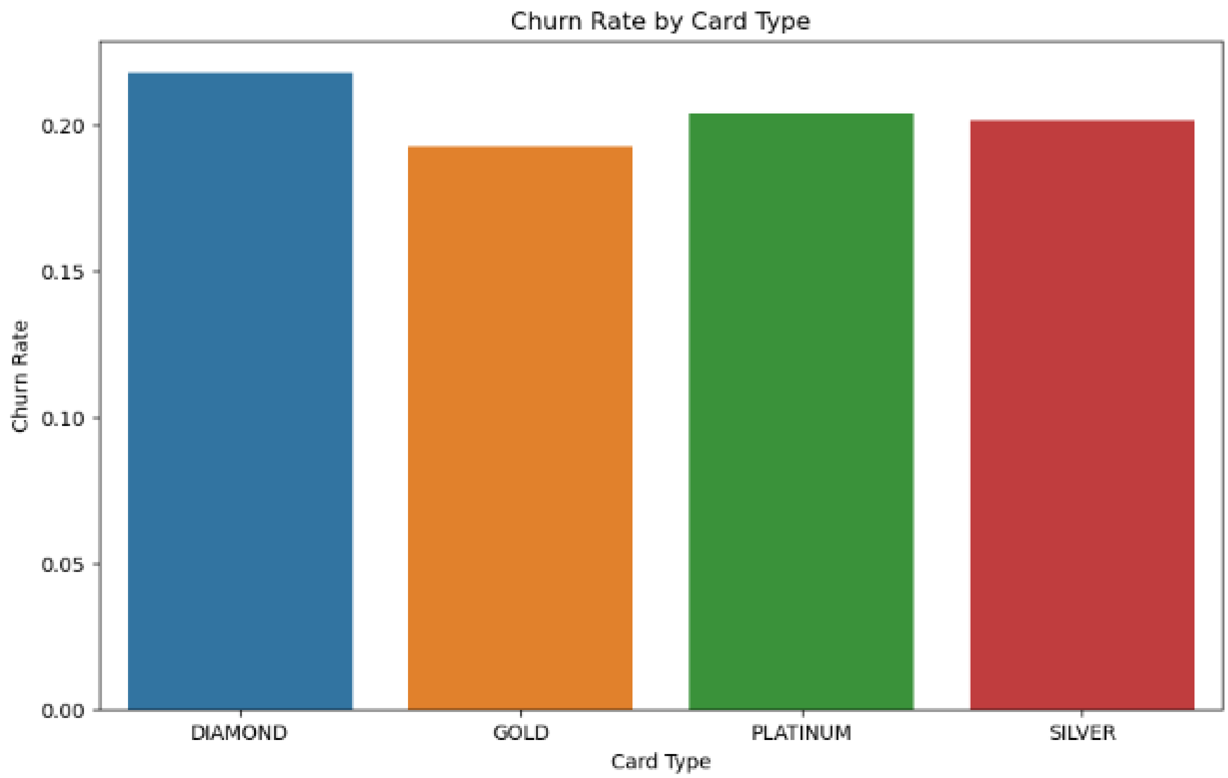
- **S**ince the p-value is more than 0.05, we fail to reject the null hypothesis. There is no association between has satisfaction and customer churn.

11.CARD USAGE ANALYSIS:

a) Impact of Card Type on Churn: We are going to examine if different Card Types have different churn rates.

```
# Calculate churn rates for different card types
card_type_churn_rate = data.groupby('Card Type')['Exited'].mean()
print("Churn rate by card type:\n", card_type_churn_rate)

# Visualization
plt.figure(figsize=(10,6))
sns.barplot(x=card_type_churn_rate.index, y=card_type_churn_rate.values)
plt.title('Churn Rate by Card Type')
plt.xlabel('Card Type')
plt.ylabel('Churn Rate')
plt.show()
```

```
Churn rate by card type:
 Card Type
DIAMOND     0.217790
GOLD        0.192646
PLATINUM    0.203607
SILVER      0.201122
Name: Exited, dtype: float64
```



- H0: There is no association between card type and customer churn. In other words, the churn rate is independent of card type.
- H1: There is an association between card type and customer churn. In other words, the churn rate depends on card type.

```
# Contingency table
contingency_card_type = pd.crosstab(data['Card Type'], data['Exited'])

# Chi-Square Test
chi2_card_type, p_card_type, dof_card_type, ex_card_type = chi2_contingency(contingenc

print(f"Chi-Square test statistic: {chi2_card_type:.4f}")
print(f"P-value: {p_card_type:.4f}")

if p_card_type < 0.05:
    print("Reject the null hypothesis: There is a significant difference in card type
else:
    print("Fail to reject the null hypothesis: There is no significant difference in c
```

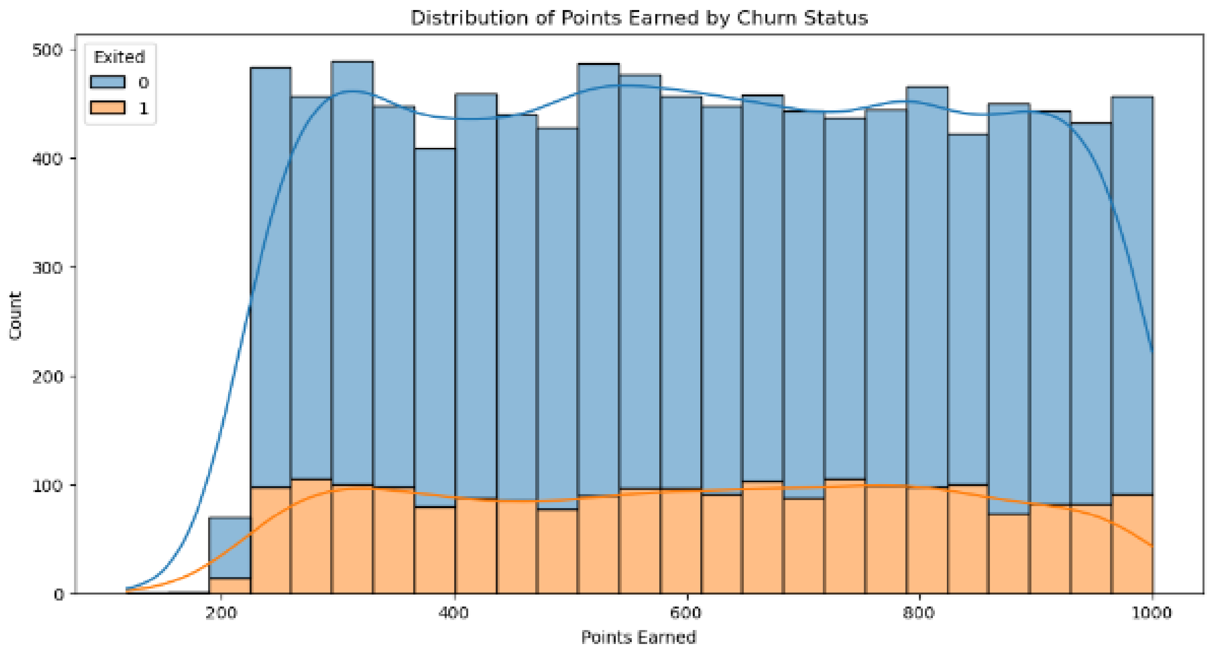
```
Chi-Square test statistic: 5.0532
P-value: 0.1679
Fail to reject the null hypothesis: There is no significant difference in card type b
etween customers who churned and those who didn't.
```

Insights:

- Since the p-value is more than 0.05, we fail to reject the null hypothesis. There is no association between card type and customer churn.

b)Loyalty Points Analysis: Now we will investigate whether points earned from credit card usage influence customer retention

```
# Distribution of points earned by churn status
plt.figure(figsize=(12,6))
sns.histplot(data=data, x='Point Earned', hue='Exited', kde=True, multiple='stack')
plt.title('Distribution of Points Earned by Churn Status')
plt.xlabel('Points Earned')
plt.ylabel('Count')
plt.show()
```



- H0: There is no association between loyalty points earned and customer churn. In other words, the churn rate is independent of loyalty points earned.
- H1: There is an association between loyalty points earned and customer churn. In other words, the churn rate loyalty points earned.

```python
from scipy.stats import ttest_ind
# Points earned by churn status
points_churned = data[data['Exited'] == 1]['Point Earned']
points_not_churned = data[data['Exited'] == 0]['Point Earned']

# T-test
t_stat, p_value = ttest_ind(points_churned, points_not_churned)

print(f"T-test statistic: {t_stat:.4f}")
print(f"P-value: {p_value:.4f}")

if p_value < 0.05:
    print("Reject the null hypothesis: There is a significant difference in points ear
else:
    print("Fail to reject the null hypothesis: There is no significant difference in p
```

```
T-test statistic: -0.4628
P-value: 0.6435
Fail to reject the null hypothesis: There is no significant difference in points earn
ed and between customers who churned and those who didn't.
```

Insights:

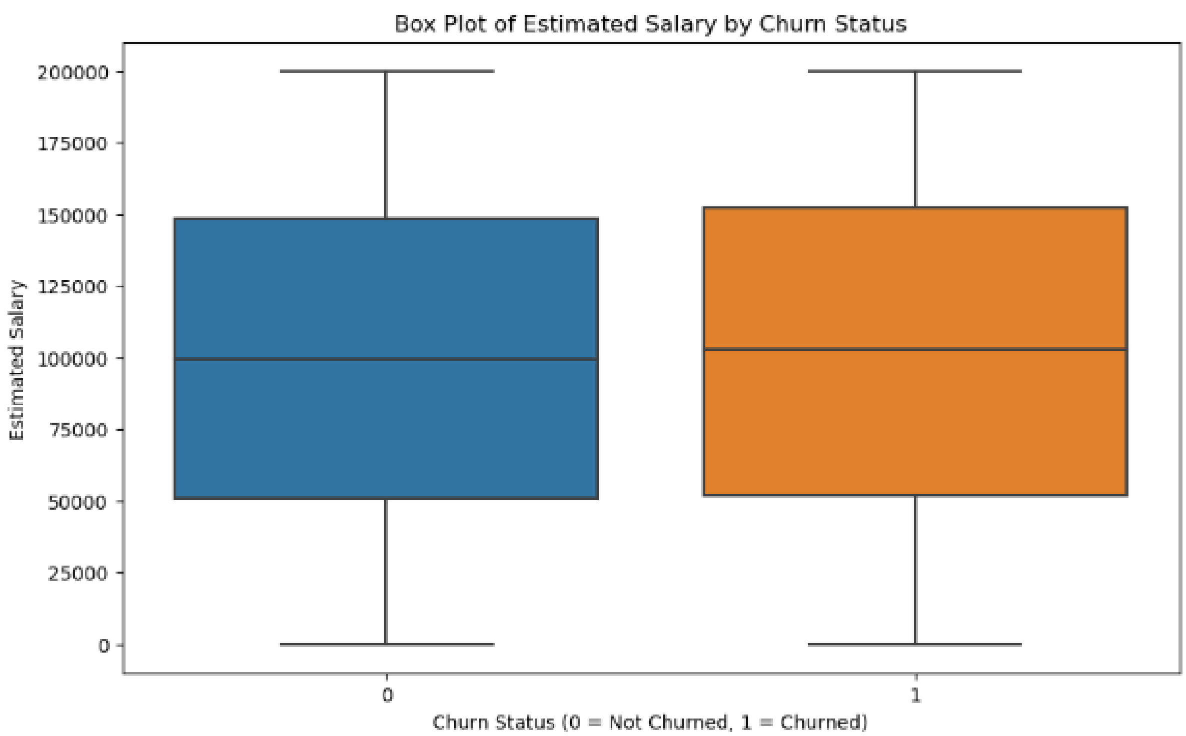
- Since the p-value is more than 0.05, we fail to reject the null hypothesis. There is no association between loyalty points and customer churn.

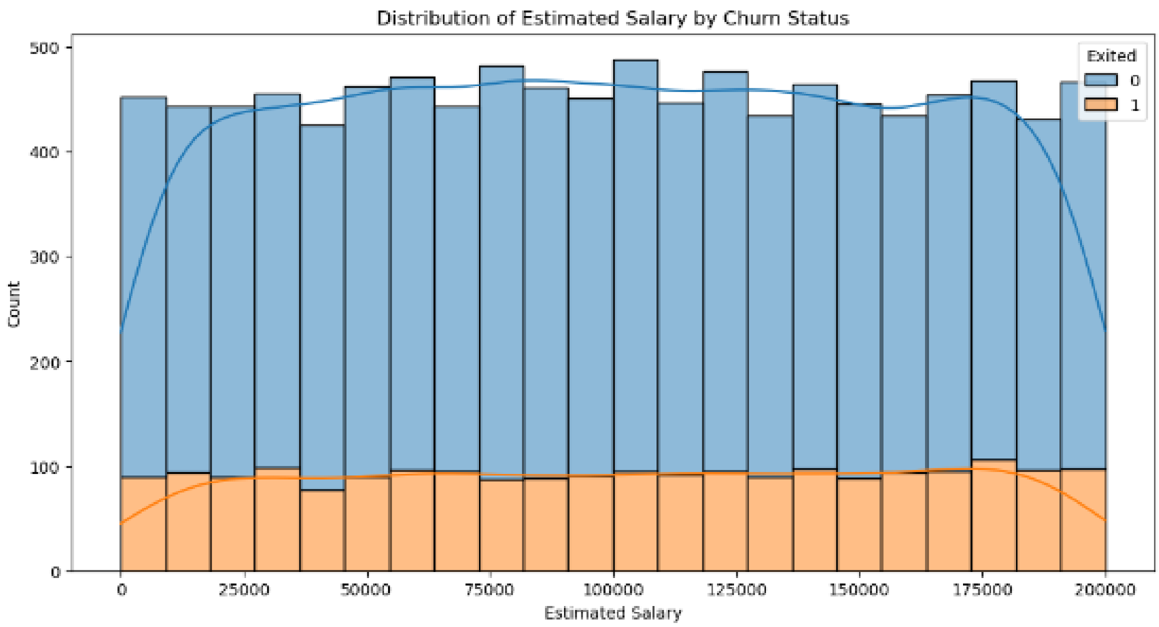
12.SALARY ANALYSIS: We are going to analyze the relationship between EstimatedSalary and customer churn, focusing on how financial well-being might influence churn decisions.

```python
# Summary statistics for estimated salary by churn status
salary_stats = data.groupby('Exited')['EstimatedSalary'].describe()
print(salary_stats)
```

```
         count           mean           std    min        25%         50%  \
Exited
0       7962.0   99726.853141  57399.956717  90.07  50783.295   99620.355
1       2038.0  101509.908783  57932.623392  11.58  51924.020  102489.335

              75%        max
Exited
0       148602.4450  199992.48
1       152443.8575  199808.10
```

```python
# Box plot
plt.figure(figsize=(10,6))
sns.boxplot(x='Exited', y='EstimatedSalary', data=data)
plt.title('Box Plot of Estimated Salary by Churn Status')
plt.xlabel('Churn Status (0 = Not Churned, 1 = Churned)')
plt.ylabel('Estimated Salary')
plt.show()

# Histogram
plt.figure(figsize=(12,6))
sns.histplot(data=data, x='EstimatedSalary', hue='Exited', kde=True, multiple='stack')
plt.title('Distribution of Estimated Salary by Churn Status')
plt.xlabel('Estimated Salary')
plt.ylabel('Count')
plt.show()
```





- H0: There is no association between estimated salary and customer churn. In other words, the churn rate is independent of estimated salary.
- H1: There is an association between estimated salary and customer churn. In other words, the churn rate estimated salary.

```
 for churned and non-churned customers
ata[data['Exited'] == 1]['EstimatedSalary']
 = data[data['Exited'] == 0]['EstimatedSalary']


alue_salary = ttest_ind(salary_churned, salary_not_churned)

tistic: {t_stat_salary:.4f}")
p_value_salary:.4f}")

< 0.05:
the null hypothesis: There is a significant difference in estimated salary and between

 reject the null hypothesis: There is no significant difference in estimated salary and
```

```
T-test statistic: 1.2489
P-value: 0.2117
Fail to reject the null hypothesis: There is no significant difference in estimated s
alary and between customers who churned and those who didn't.
```

Insights:

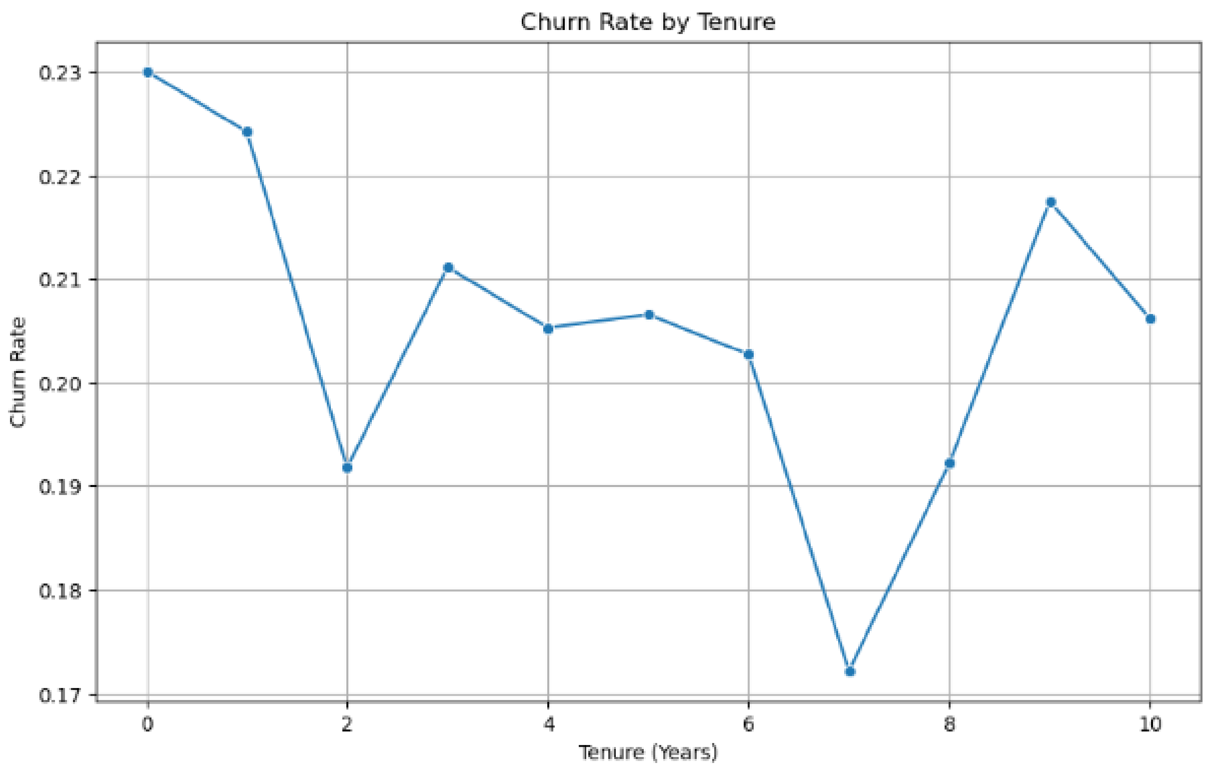
- Since the p-value is more than 0.05, we fail to reject the null hypothesis. There is no association between estimated salary and customer churn.

c) Tenure Analysis: We are going to investigate if there is any particular tenure when customers churn the most or not.

```python
# Calculate churn rate for each tenure level
tenure_churn_rate = data.groupby('Tenure')['Exited'].mean().reset_index()
tenure_churn_rate.columns = ['Tenure', 'ChurnRate']
print(tenure_churn_rate)
```

```
    Tenure  ChurnRate
0        0   0.230024
1        1   0.224155
2        2   0.191794
3        3   0.211100
4        4   0.205258
5        5   0.206522
6        6   0.202689
7        7   0.172179
8        8   0.192195
9        9   0.217480
10      10   0.206122
```

```python
# Line plot of churn rate by tenure
plt.figure(figsize=(10,6))
sns.lineplot(x='Tenure', y='ChurnRate', data=tenure_churn_rate, marker='o')
plt.title('Churn Rate by Tenure')
plt.xlabel('Tenure (Years)')
plt.ylabel('Churn Rate')
plt.grid(True)
plt.show()
```



- H0: There is no association between tenure and customer churn. In other words, the churn rate is independent of tenure.
- H1: There is an association between tenure and customer churn. In other words, the churn rate tenure.

```
from scipy.stats import f_oneway

# ANOVA test
groups = [data['Exited'][data['Tenure'] == tenure] for tenure in data['Tenure'].unique
f_stat, p_value = f_oneway(*groups)

print(f"ANOVA F-statistic: {f_stat:.4f}")
print(f"P-value: {p_value:.4f}")

if p_value < 0.05:
    print("Reject the null hypothesis: There is a significant difference in tenure and
else:
    print("Fail to reject the null hypothesis: There is no significant difference in t
```

```
ANOVA F-statistic: 1.4063
P-value: 0.1703
Fail to reject the null hypothesis: There is no significant difference in tenure and
between customers who churned and those who didn't.
```

- **Insights: S**ince the p-value is more than 0.05, we fail to reject the null hypothesis. There is no association between tenure and customer churn.

FINAL INSIGHTS:

1. **Customer Churn and Complaints:**
    - Around 20.4% of customers have churned (Exited), indicating a significant churn rate in the dataset.
    - Similarly, approximately 20.4% of customers have filed complaints, suggesting that customer dissatisfaction is a key issue.
    - **Actionable Insight:** The bank should prioritize resolving customer complaints effectively to reduce churn, as dissatisfaction directly correlates with churn.
2. **Age Distribution:**
    - The average age of customers is approximately 39 years, with most customers falling between 32 (25th percentile) and 44 (75th percentile) years.
    - **Actionable Insight:** To retain both younger and older customers, the bank can offer tailored schemes and special offers.
3. **Customer Tenure:**
    - The average tenure with the bank is around 5 years, with most customers having tenures between 3 and 7 years.

- **Actionable Insight:** Focus on ensuring a smooth experience during the first three years of the customer journey to prevent early churn.

4. **Account Balance:**
    - The average account balance is approximately $76,486, with many customers not maintaining a balance, which could lead to charges and dissatisfaction.
    - **Actionable Insight:** Consider offering zero-balance accounts or lower minimum balance requirements to reduce the burden on customers and improve retention.
5. **Product Usage:**
    - Most customers use 1 (25th percentile) or 2 (75th percentile) products.
    - **Actionable Insight**: Enhance and promote additional products to engage customers more deeply and reduce churn.
6. **Estimated Salary:**
    - The average estimated salary is around $100,090, with those earning less potentially more likely to churn.
    - **Actionable Insight:** Offer financial products and services that cater to lower-income customers to help retain them.
7. **Satisfaction Score:**
    - The average satisfaction score is around 3, indicating a mixed sentiment among customers.
    - **Actionable Insight:** Improve customer service and provide additional benefits to enhance satisfaction and retention.
8. **Geographical Insights:**
    - The majority of the data pertains to customers from France, with a frequency of 5,014 entries.
    - **Actionable Insight:** Focus retention efforts and personalized marketing strategies on French customers to leverage the largest customer base.
9. **Gender Distribution:**
    - Male customers predominate the dataset with a frequency of 5,457.
    - **Actionable Insight:** Develop targeted offers and discounts to attract and retain more female customers while continuing to provide value to male customers.
10. **Card Type:**
    - The most frequently used card type is the "Diamond" card, with 2,507 customers.
    - **Actionable Insight:** Promote the benefits of the Diamond card to retain existing customers and attract new ones.
11. **Predictors of Customer Churn:**
    - Based on the correlation analysis, Age and Balance are significant predictors of customer churn.
    - **Actionable Insight:** Implement strategies focused on these factors, such as personalized financial advice for older customers and balance maintenance incentives.

By addressing these insights, the bank can develop targeted strategies to enhance customer retention, improve satisfaction, and ultimately reduce churn.

RECOMMENDATIONS:

1. **Addressing Churn Among Mid-Age Customers:**
    - **Insight:** The 46-65 age group shows higher exit counts, accounting for around 4.9% of churn.
    - **Recommendation:** Investigate the specific needs and pain points of this age group. Implement targeted services such as personalized financial advice, retirement planning, and health insurance options to cater to their unique requirements.
2. **Enhancing Customer Experience in Germany:**
    - **Insight:** Germany has the highest churn rate, despite France having the most customers. Germany is also the second-largest customer base for the bank.
    - **Recommendation:** Focus on improving the customer experience in Germany. Conduct surveys to understand their dissatisfaction, provide localized customer service, and introduce special offers or loyalty programs tailored to the German market.
3. **Reducing Female Customer Churn:**
    - **Insight:** Female customers have a higher churn rate compared to male customers.
    - **Recommendation:** Develop and promote special schemes or offers for female customers, such as financial products geared towards women, family-oriented benefits, and discounts on products and services that are popular among female customers.
4. **Improving Product and Service Engagement:**
    - **Insight:** Most customers use only 1 or 2 products, and the number of products used correlates with churn.
    - **Recommendation:** Encourage customers to use more bank products by offering bundled services at discounted rates. Educate customers on the benefits of additional products through targeted marketing campaigns and personalized recommendations.
5. **Activity Level and Retention:**
    - **Insight:** Active members (IsActiveMember) are less likely to churn.
    - **Recommendation:** Increase customer engagement through regular communication, loyalty programs, and exclusive events. Provide incentives for account activity, such as rewards for frequent transactions and bonuses for maintaining a high activity level.
6. **Addressing Zero Balance Accounts:**
    - **Insight:** Many customers maintain a zero balance, which might lead to dissatisfaction and churn.
    - **Recommendation**: Introduce zero-balance account options or reduce minimum balance requirements to alleviate this issue. Additionally, offer financial management tools to help customers manage their finances better and avoid zero balance situations.
7. **Handling Customer Complaints**:
    - **Insight:** Customers who have filed complaints are more likely to churn.
    - **Recommendation:** Enhance the complaint resolution process to ensure timely and satisfactory responses. Implement a follow-up system to check on the customer's

satisfaction after a complaint has been resolved. Use feedback from complaints to improve overall service quality.

8. **Increasing Customer Satisfaction:**
   - **Insight:** The average satisfaction score is around 3, indicating a mixed sentiment among customers.
   - **Recommendation:** Conduct regular customer satisfaction surveys to identify areas for improvement. Invest in training for customer service representatives to enhance their interaction skills. Introduce a customer satisfaction incentive program for bank staff to motivate them to provide excellent service.

By addressing these recommendations, the bank can reduce churn rates, improve customer satisfaction, and enhance overall customer retention, ultimately leading to a more stable and loyal customer base. Given more data we can provide more factors and recommendations.