**Data Analysis Project**

**Bank Customer Churn Prediction**

**TABLE OF CONTENT**

1. [INTRODUCTION 3](#_TOC_250029)
   1. [Aim of project 3](#_TOC_250028)
   2. [The Project’s Value 3](#_TOC_250027)
   3. [Motivation 3](#_TOC_250026)
2. [ABOUT THE DATASET 4](#_TOC_250025)
   1. [Data Source 4](#_TOC_250024)
   2. [Data Description 4](#_TOC_250023)
3. [EXPLORATORY DATA ANALYSIS 5](#_TOC_250022)
   1. [Check missing values and Dropping unnecessary columns 5](#_TOC_250021)
   2. [Basic Data Discovery Statistical Overview of Variables 6](#_TOC_250020)
   3. [Distribution and correlation between input variables and the churn rate analysis. 7](#_TOC_250019)
4. [CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING 21](#_TOC_250018)
   1. [Data Preprocessing 21](#_TOC_250017)
   2. [The Elbow Method 21](#_TOC_250016)
   3. [Clusters Characteristics 22](#_TOC_250015)
5. [EXPERIMENTS 27](#_TOC_250014)
   1. [Data Preprocessing 27](#_TOC_250013)
   2. [Building models 28](#_TOC_250012)
   3. [Evaluation 32](#_TOC_250011)
6. [LIMITATIONS 35](#_TOC_250010)
   1. [Data Availability 35](#_TOC_250009)
   2. [Feature Selection and Engineering: 35](#_TOC_250008)
   3. [Imbalanced Dataset: 35](#_TOC_250007)
   4. [Model Limitations: 35](#_TOC_250006)
   5. [Interpretation and Implementation Challenges: 35](#_TOC_250005)
7. [CONCLUSION 36](#_TOC_250004)
   1. [Age is the key Predictors of Churn 36](#_TOC_250003)
   2. [Customer Segmentation: 36](#_TOC_250002)
   3. [Model Performance: 36](#_TOC_250001)
   4. [Retention Strategies Based on Insights: 36](#_TOC_250000)

# INTRODUCTION

# Aim of project

Project’s target is to analyze and anticipate the customer’s turnover rate at the bank based on their behaviors and transaction information. Regarding the customers’ behavior (continuing or discontinuing using bank services), it could be influenced by many factors, for instance, the quality of services or the customers’ financial stability.

In this project, our team will use data analysis methods to evaluate, define the factors affecting the customers’ behavior in order to identify the customers having high risk of stopping using the services.

# The Project’s Value

With the analysis in the project, the banks can have a more comprehensive understanding about the reasons why their customers decided to leave. Therefore, they can optimize or improve the services which are not satisfactory according to the customers’ experience. In addition, banks can develop strategies, campaigns to boost their business efficiency and have more loyal customers for their businesses.

# Motivation

In today’s developing era, there are many banks operating strongly, therefore the financial market is increasingly competitive. Customer retention is becoming a vital factor to support the banks develop more sustainably and lower the rate of leaving customers. By the following research, our team could comprehend the reasons that make the customers stop using the service, thereby the banks could develop more strategies to level up their services, especially customer care services.

# ABOUT THE DATASET

This dataset, titled the “Banking Customer Churn Prediction Dataset,” provides detailed information about bank customers, specifically whether they have exited the bank or remained as active clients. It is suitable for investigating the factors contributing to customer churn in banking institutions and for developing predictive models to identify customers at high risk of churning.

# Data Source

The data was sourced from Kaggle (from the "Bank Customer Churn Prediction" dataset). Kaggle is a platform that offers a wide range of datasets with data exploration and analysis. This dataset is rated as a gold-standard dataset and has received 132 upvotes.

# Data Description

The dataset includes a mix of input variables, which represent the customer's demographics, financial status, and engagement with the bank, as well as a target variable, **Exited**, which indicates customer churn. This variable is essential for predicting which customers are likely to leave the bank, enabling banks to better understand and proactively address factors contributing to customer churn.

The dataset contains 14 columns and 10,000 rows. The variables included in the dataset are as follows:

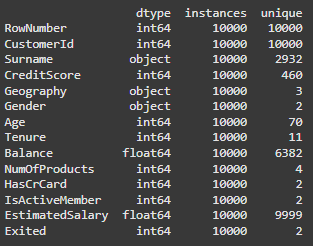
* **RowNumber**: The sequential number assigned to each row in the dataset.
* **CustomerId**: A unique identifier for each customer.
* **Surname**: The customer's surname.
* **CreditScore**: The customer’s credit score.
* **Geography**: The customer's geographical location (e.g., country or region).
* **Gender**: The customer's gender.
* **Age**: The customer’s age.
* **Tenure**: The number of years the customer has been with the bank.
* **Balance**: The customer’s account balance.
* **NumOfProducts**: The number of bank products the customer has.
* **HasCrCard**: Indicates whether the customer has a credit card (binary: yes/no).
* **IsActiveMember**: Indicates whether the customer is an active member (binary: yes/no).
* **EstimatedSalary**: The customer's estimated salary.
* **Exited**: Indicates whether the customer has exited the bank (binary: yes/no), serving as the target variable for churn analysis.

# EXPLORATORY DATA ANALYSIS

**Exploratory Data Analysis (EDA)** is the process of analyzing and summarizing datasets to uncover patterns, relationships, and insights, often as an initial step before building predictive models or performing statistical tests.

# Check missing values and Dropping unnecessary columns

* + 1. **Basic column information**

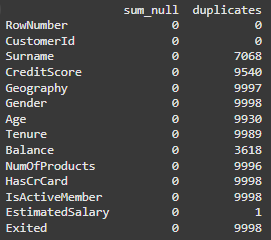
****

The data consists of **11 important variables**:

* + - * **6 Numerical variables:** CreditScore, Age, Tenure, Balance, NumOfProducts, EstimatedSalary.
      * **2 Categorical variables:** Geography, Gender.
      * **3 Binary variables:** HasCrCard, IsActiveMember, Exited.

The remaining **3 unnecessary columns** (RowNumber, CustomerID, Surname) will be dropped.

# Missing and duplicate data information

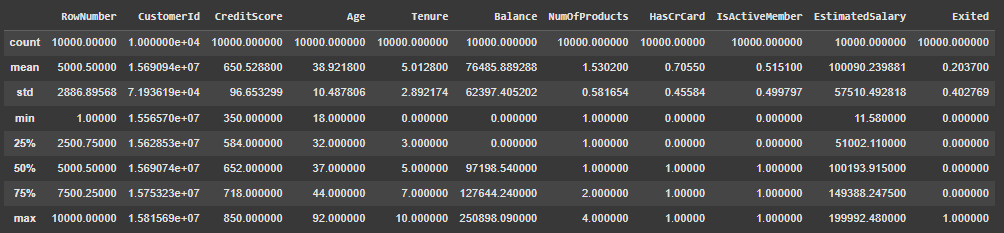
****

The data is cleaned. There are no missing values and no duplicates of rows.

# Basic Data Discovery Statistical Overview of Variables

* + 1. **Categorical Features:**
       - **Geography:** Germany, France, and Spain.
       - **Gender:** Male, Female

# Numerical Features:

****

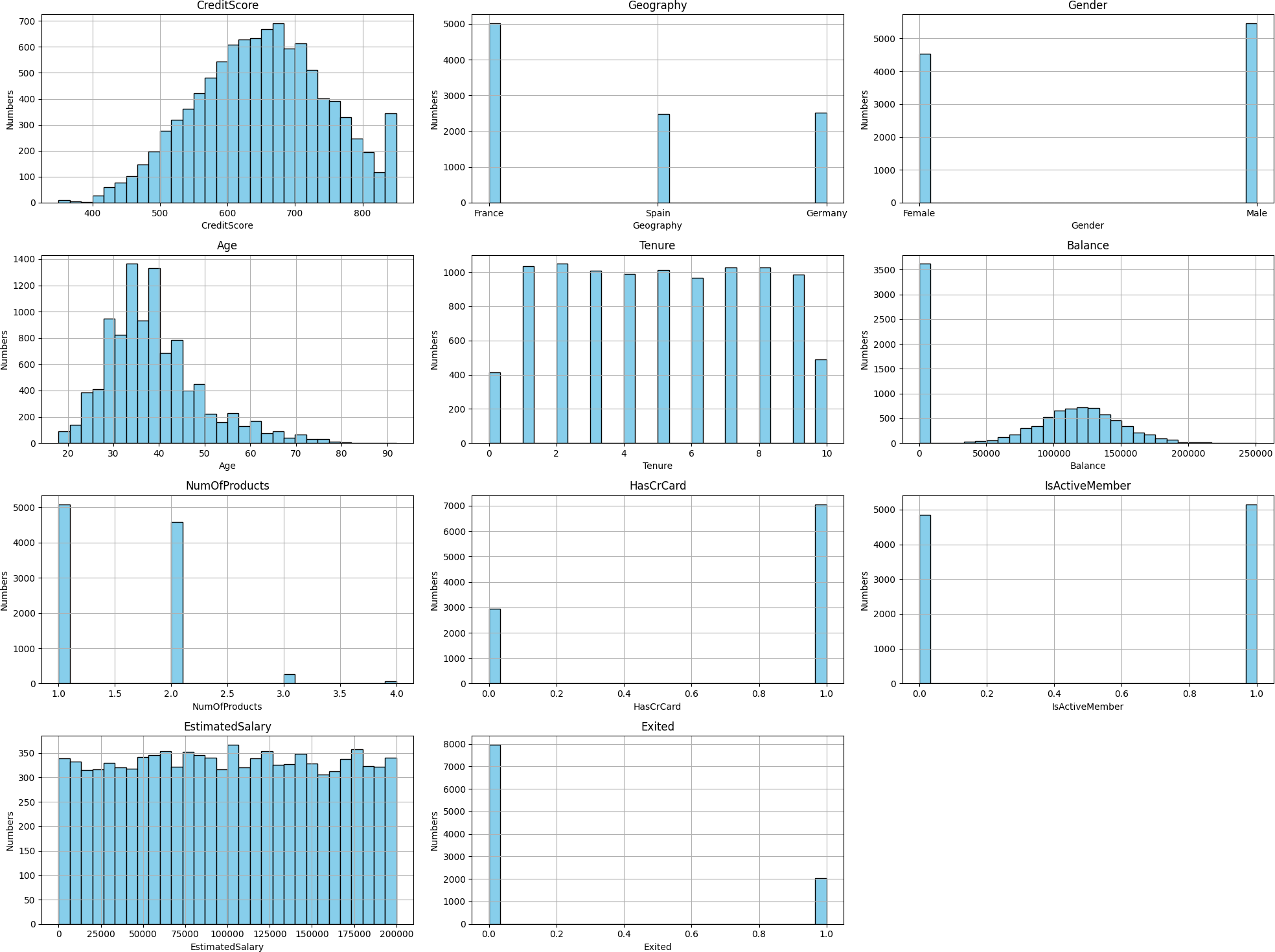
* + - * **CreditScore:** The average credit score is 650.53, with a standard deviation of 96.65. Scores range from a low of 350 to a high of 850, indicating a wide variety of customer credit levels.
      * **Age:** The average age is 38.92, with customers ranging from 18 to 92 years old. The standard deviation is 10.49, showing moderate dispersion.
      * **Tenure:** Average tenure is 5.01 years, ranging from 0 to 10 years. The standard deviation of 2.89 shows diversity in customer tenure.
      * **Balance:** The average balance is 76,485, but the minimum value is 0, indicating many customers have no account balance. The highest balance is 250,898, showing an uneven distribution.
      * **NumOfProducts:** The average number of products used by customers is 1.53, with a range of 1 to 4 products.
      * **EstimatedSalary:** The average customer salary is 100,090, with a high standard deviation of 57,510. Salaries range from 11,580 to 199,992, indicating diverse income levels.

# Binary Features:

* + - * **Exited:** The churn rate is about 20.37%, indicating most customers retain their accounts.
      * **HasCrCard:** 70.55% of customers have a credit card.
      * **IsActiveMember:** around 51.51% of customers being active members. Active customers tend to maintain their accounts longer.

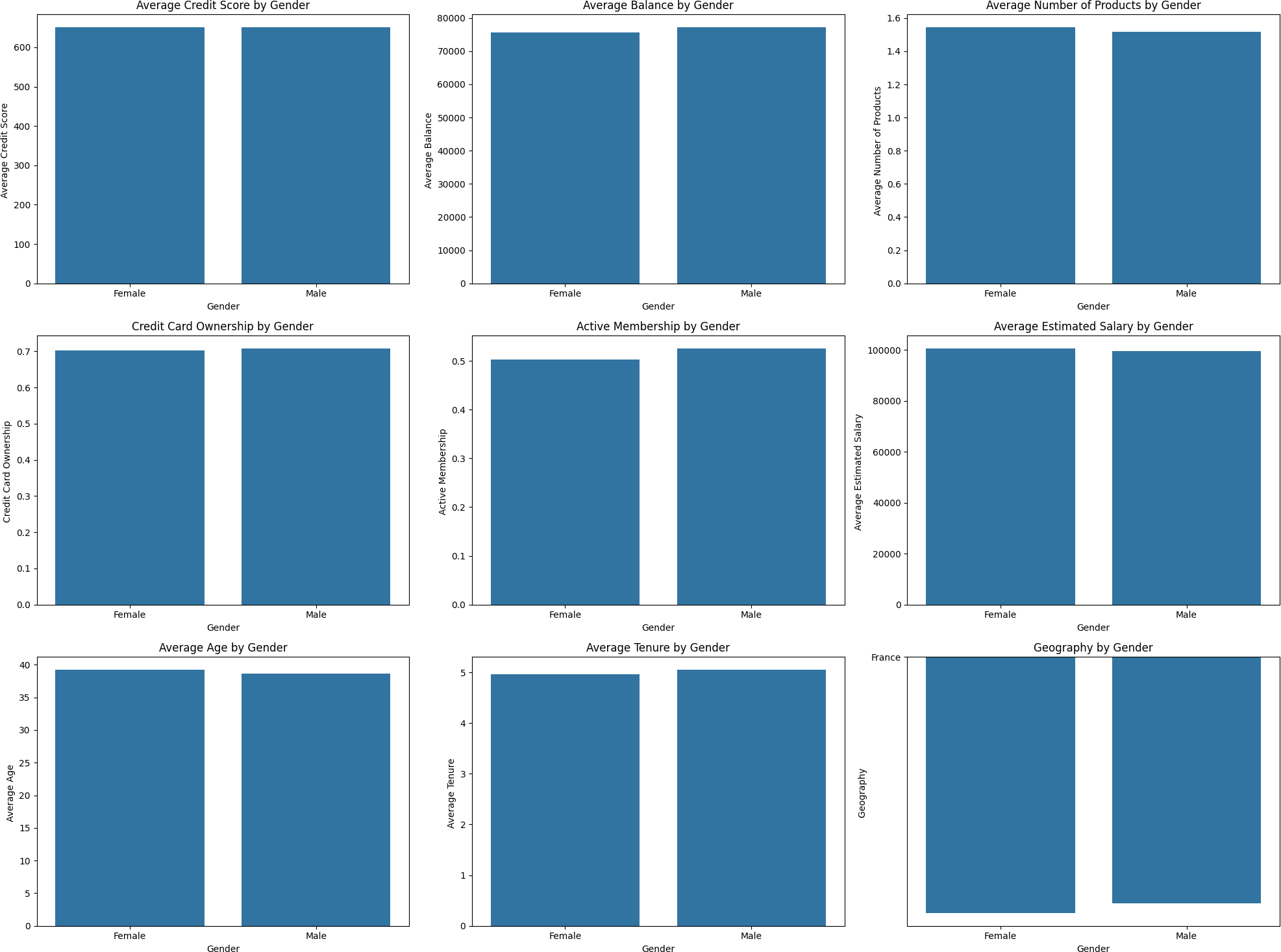
# Distribution and correlation between input variables and the churn rate analysis.

* + 1. **Descriptive input variables.**

****

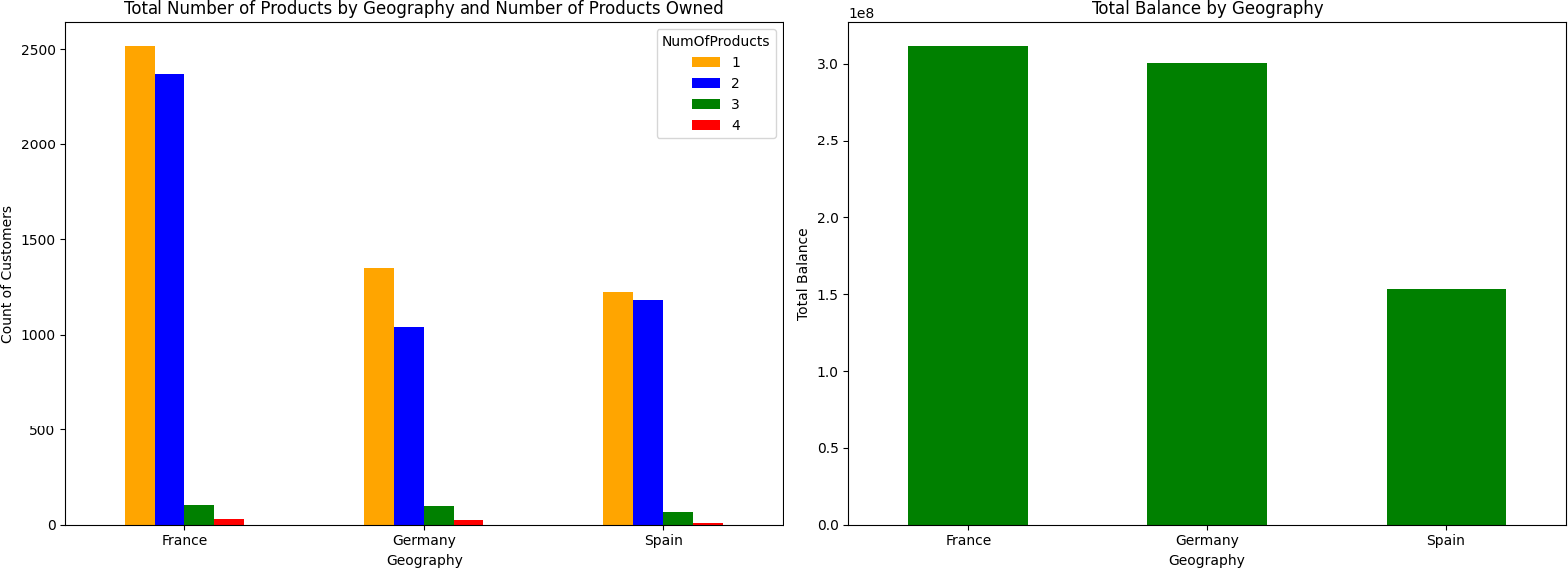
* + - * **CreditScore:** Relatively normal distribution centered around 600-700, suggesting most customers have mid-range credit scores. There are fewer customers at the lower (300-400) and higher (800+) score extremes.
      * **Geography:** The highest number of customers in France, followed by Germany, and then Spain.
      * **Gender:** The distribution between genders is roughly even, with slightly more male customers than female customers.
      * **Age:** The age distribution is slightly skewed towards younger adults, with a concentration of customers in the 30-40 age range. Fewer customers are above 60 or below 20.
      * **Tenure:** This feature shows a uniform distribution, meaning customers have varying lengths of relationships with the bank, from 0 to 10 years, with no specific tenure length dominating.
      * **Balance:** The balance distribution has a peak near zero, indicating many customers have zero or very low balances, while others have a range of balances up to approximately 250,000.
      * **NumOfProducts:** Most customers have only one or two products, with very few having three or four. This suggests limited product engagement beyond the primary products.
      * **HasCrCard:** This binary feature shows that most customers have a credit card, with a much smaller group without one.
      * **IsActiveMember:** The number of active members is slightly larger than inactive ones.
      * **EstimatedSalary:** This feature is evenly distributed across its range, indicating that customers have a wide variety of income levels without any particular concentration.
      * **Excited:** This is the target variable, showing that most customers did not churn , while a smaller proportion did. This imbalance suggests that churn is less common, and may need attention when modeling to avoid bias.

# The average values of key variables between two Gender groups.

****

There are **no major differences** between male and female in customers' behavioral or financial characteristics. However, on average, males tend to use more products, are more active, have higher account balances, and earn more than females.

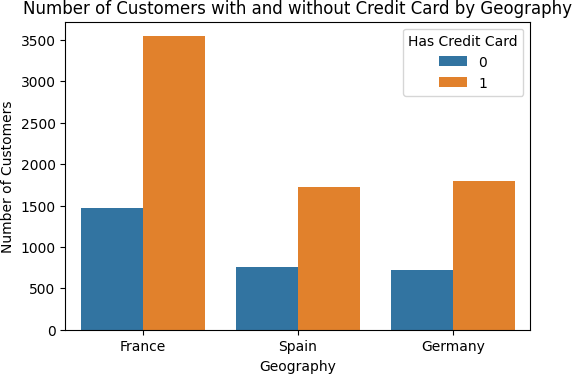
# The distributions and categorical breakdowns by Geography

****

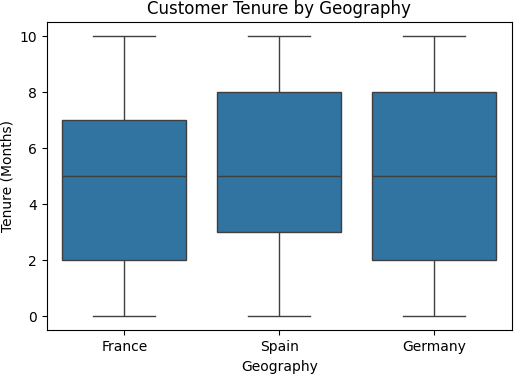
* + - * **Total Number of Products by Geography and Number of Products Owned**
        + **France**, **Germany** and **Spain** have similar patterns, with most customers owning 1 or 2 products.
        + Only a small number of customers in each country own 3 or 4 products, indicating that most customers do not engage extensively with multiple bank products.

# Total Balance by Geography

* + - * + **France** and **Germany** have significantly higher total balances (3x10^8) compared to **Spain** (1.5x10^8).
        + Although the number of customers in Germany is half of France, the 2 countries have nearly the same balance. This indicates that the average customers in Germany are more likely to have a much higher balance in their account than French.
        + **Spain** has a lower total balance.



=> The proportion of customers with a credit card is twice as high as those without one in three countries. Germany and Spain have nearly equal numbers of credit card and non-credit card customers.

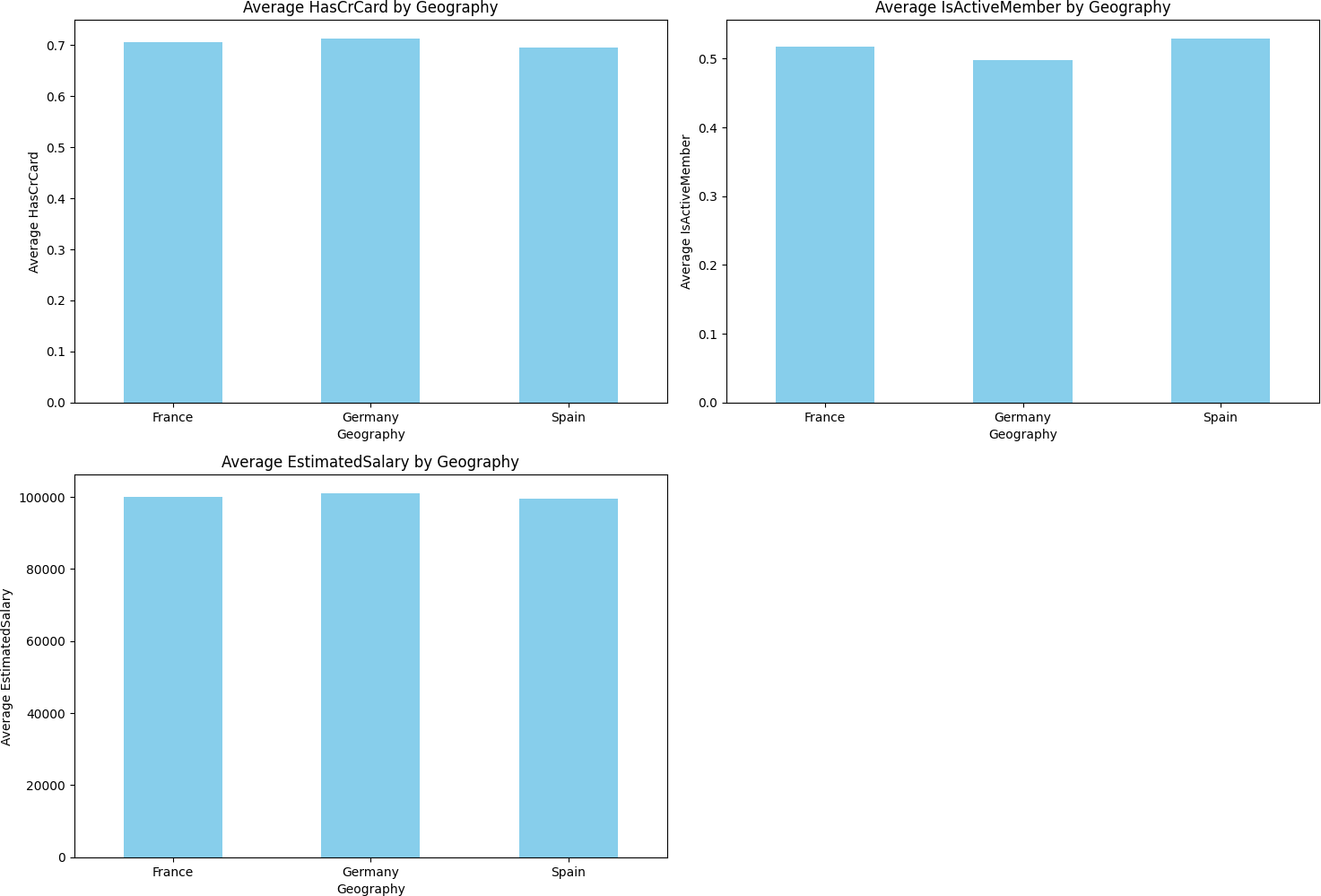


# Overall Tenure Distribution

* + - * + **Median:** The median tenure of customers in all three countries is around 5 years. This indicates that the average tenure of customers is similar across the countries.
        + **Interquartile Range (IQR):** The IQR, ranging from the 25th to the 75th percentile, is relatively similar across all three countries, spanning from about 3 to 7 years. This shows that most customers in these countries have a tenure between 3 and 7 years.

# Upper and Lower Whiskers

* + - * + **Lower Whisker:** The lower whisker in all three countries extends to 0, indicating some customers are new or have a very short tenure in each country.
        + **Upper Whisker:** The upper whisker extends to 10 years, representing customers who have been with the bank for the maximum recorded tenure of 10 years.
        + **Consistency Across Countries:** The length of the upper and lower whiskers is similar across countries, indicating that the diversity in customer tenure is comparable in all three countries.



# Average HasCrCard by Geography:

* + Approximately 70% of customers in each country own a credit card, indicating a similar level of credit card ownership across these geographies.

# Average IsActiveMember by Geography:

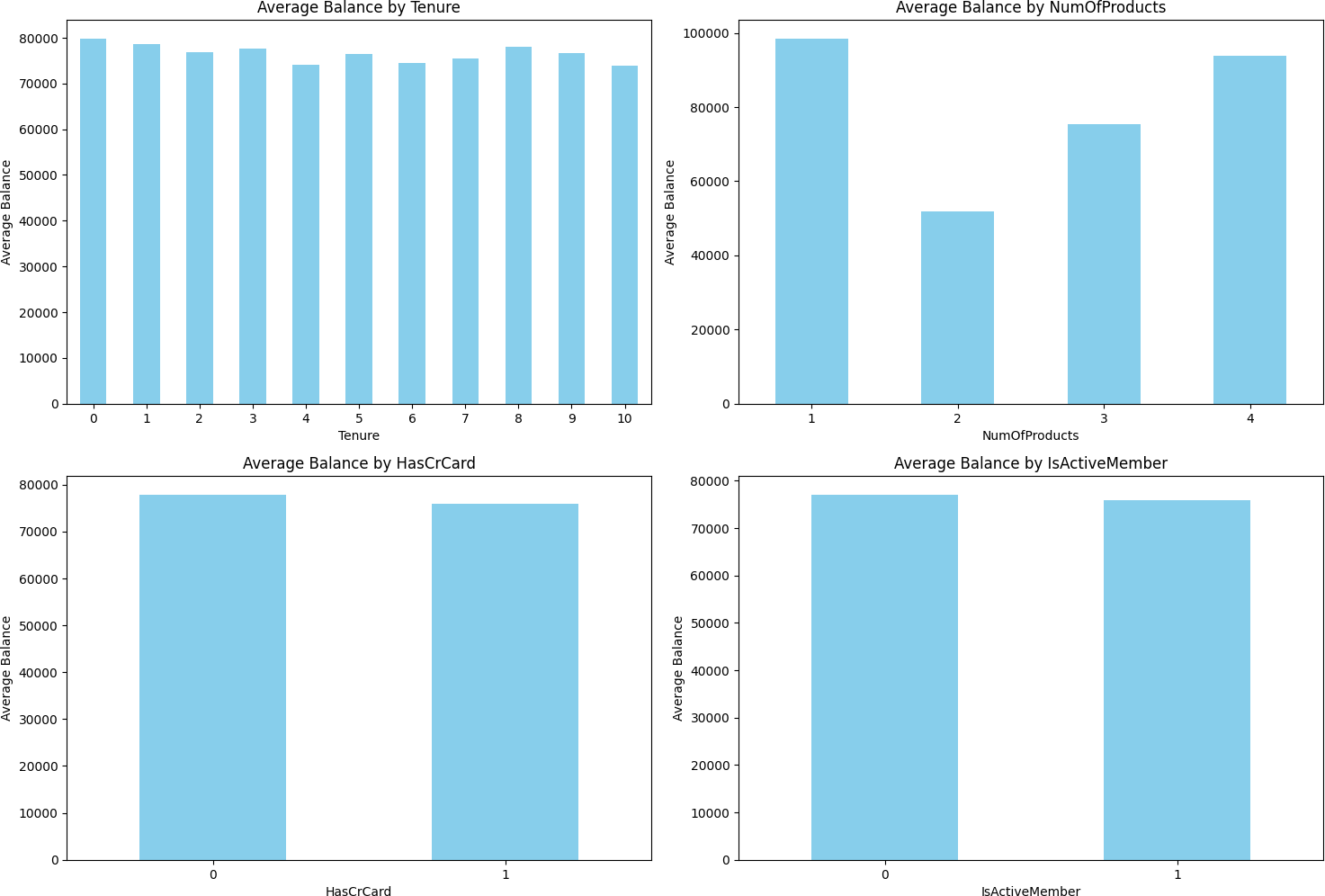
* + The average values for IsActiveMember across the three countries are also quite similar, slightly above 0.5.
  + This suggests that around 50% of customers in each country are actively using the service, with no significant difference in active membership among the countries.

# Average EstimatedSalary by Geography:

* + The average estimated salary for customers is consistent across France, Germany, and Spain, with values close to each other, approximately around 100,000.

There is no strong difference in terms of **Average HasCrCard, Average EstimatedSalary, Average IsActiveMember across 3 countries.**

# The distributions and categorical breakdowns by Balance

****

* + - * **Average Balance by Tenure:**

The average balance remains fairly consistent across all tenure years (from 0 to 10 years), indicating that the length of time a customer has been with the bank does not significantly impact their balance.

# Average Balance by Number of Products:

* + - * + Customers with only 1 or 4 products have the highest average balances.
        + Customers with 2 products have the lowest average balance, while those with 3 products have a moderate balance.
        + This could indicate that customers with fewer products might focus more on savings, or there may be a specific product type associated with higher balances.

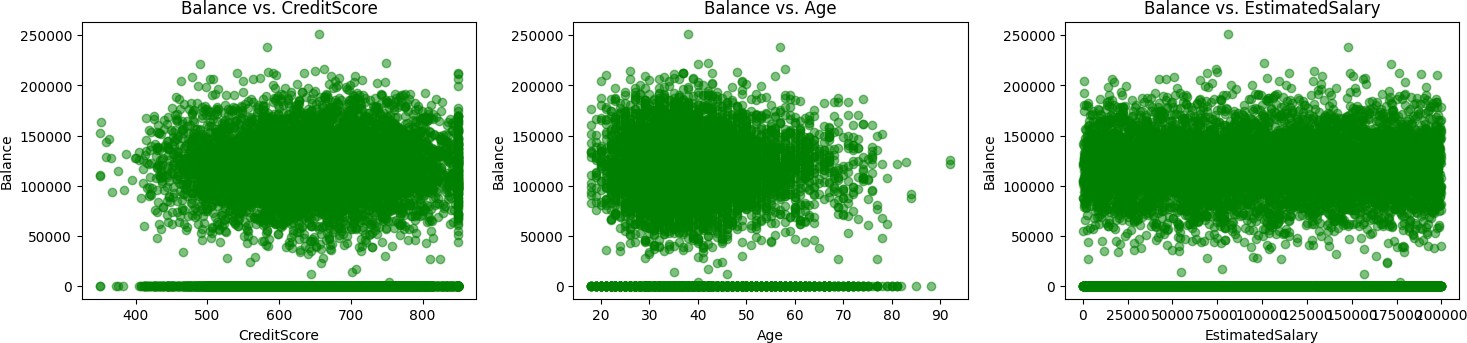
# Average Balance by HasCrCard:

The average balance is similar between customers with a credit card (1) and those without (0), suggesting that having a credit card does not significantly impact the balance.

# Average Balance by IsActiveMember:

The average balance is also consistent between active members (1) and inactive members (0), implying that active status may not have a direct influence on the balance amount.

**=>** These charts highlight that balance does not vary significantly by tenure, credit card ownership, or membership activity status, though there is some variation by number of products owned.



# Balance vs. CreditScore

* + - * + The scatter plot shows no visible relationship between Balance and CreditScore.
        + Customers with both high and low credit scores can have a wide range of balance values, from 0 to over 250,000.

# Balance vs. Age

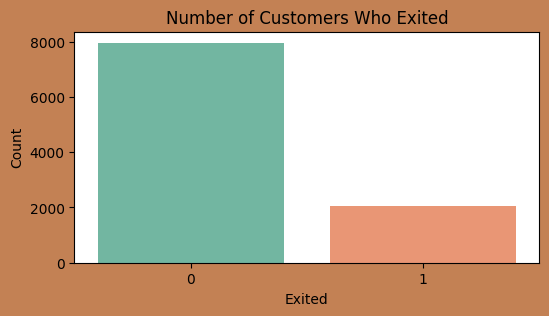
* + - * + There is no clear correlation between Balance and Age in this plot.
        + Customers across all age groups tend to have a wide range of balances.

# Balance vs. EstimatedSalary

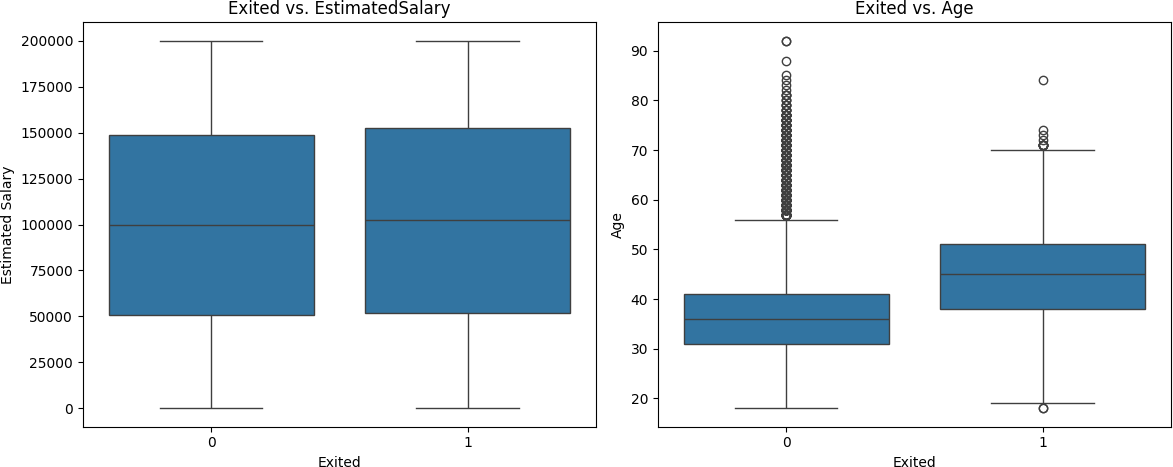
* + - * + This plot also shows no strong relationship between Balance and EstimatedSalary.
        + Customers with a wide range of salaries (from 0 to 200,000) can have balances ranging from 0 to 250,000 and there's no clear trend or correlation.

# => All three scatter plots suggest that there is no clear or strong linear relationship between Balance and the variables CreditScore, Age, and EstimatedSalary.

* + 1. **Churn rate analysis**

****

The number of retained customers is much higher than those who left, at 8,000 compared to 2,000, indicating a high retention rate.



# Exited vs. EstimatedSalary

* + - * + This box plot shows the distribution of EstimatedSalary for customers who exited (Exited = 1) and those who stayed (Exited = 0).
        + Observation: There is no significant difference in the median or overall distribution of estimated salary between the two groups.

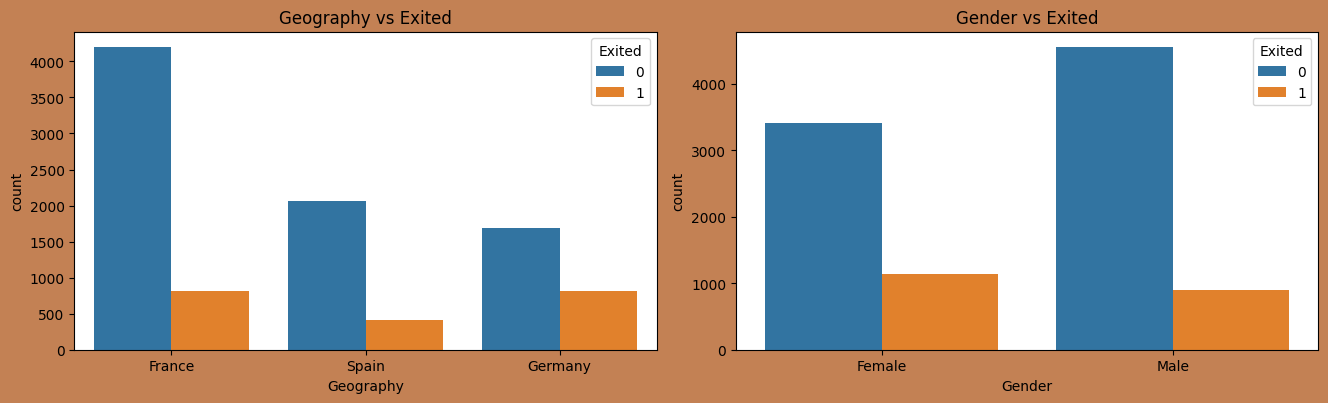
**=>** EstimatedSalary may not have a strong influence on customer churn. Both groups (churned and non-churned) have a similar range and median salary, indicating that salary alone may not be a determining factor for churn in this dataset.

# Exited vs. Age:

* + - * + Customers who did not churn (Exited = 0) have a wider age range, with a large number of outliers extending into older ages, suggesting that some older customers tend to stay with the service.
        + Customers who churned (Exited = 1) have a slightly higher median age compared to those who did not churn. The age distribution is more centered, with fewer outliers, indicating that older customers may be slightly more likely to leave.
        + The median age is higher for customers who churned, which could suggest that

older customers are more likely to exit.

=> Age appears to be a relevant factor in churn. Older customers are more likely to leave the bank, as indicated by the higher median age in the exited group.

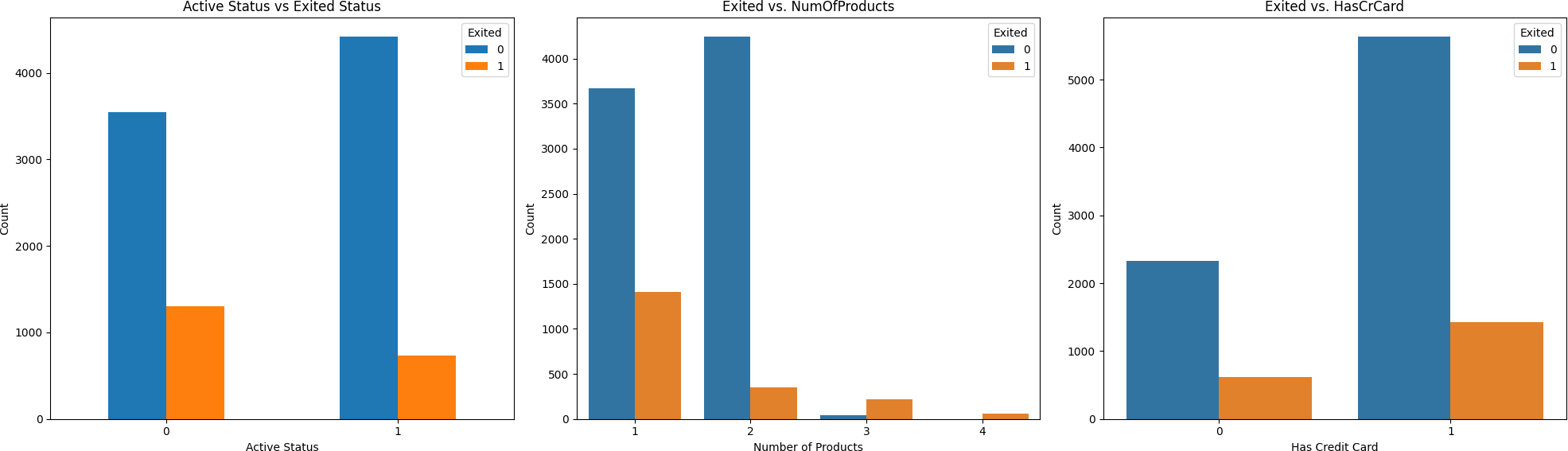


# Geography vs. Exited (Geography and Exit Status Chart):

* + - * + Customers in **France** have the lowest churn rate, with most remaining.
        + **Germany** has the highest churn rate compared to other countries, where the churn rate is nearly half the retention rate.

# Gender vs. Exited (Gender and Exit Status Chart):

* + - * + Male customers have a lower churn rate than female customers, with significantly more men retaining their accounts.
        + Female customers have a higher churn rate than males, with the number of female customers leaving accounting for a bigger proportion.



# Active Status vs. Exited Status

* + - * + **Active Status (1)**: Customers who are actively using the service have a much lower churn rate. Most active customers stay with the service.
        + **Inactive Status (0)**: Inactive customers show a higher likelihood of churning. A high portion of inactive customers have exited.

**=>** Being actively engaged with the service seems to correlate with customer retention, suggesting that encouraging customer activity may reduce churn.

# Exited vs. Number of Products:

* + - * + **1 Product**: The churn rate is quietly low with this group, nearly below ⅓ customers with 1 product exit the bank.
        + **2 Products**: Customers with two products have the lowest churn rate.

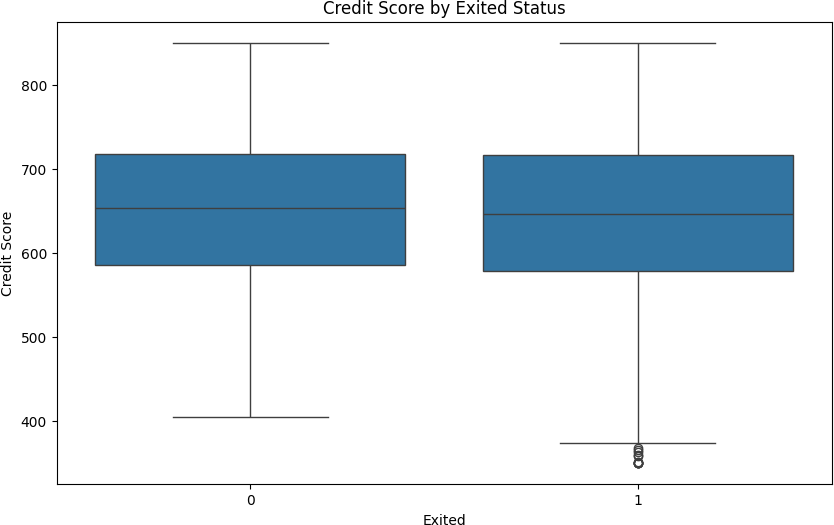
# 3 or More Products:

=> The customers, who own 1 and 2 products tend to remain with the bank, especially with 2 products.

# Exited vs. Has Credit Card:

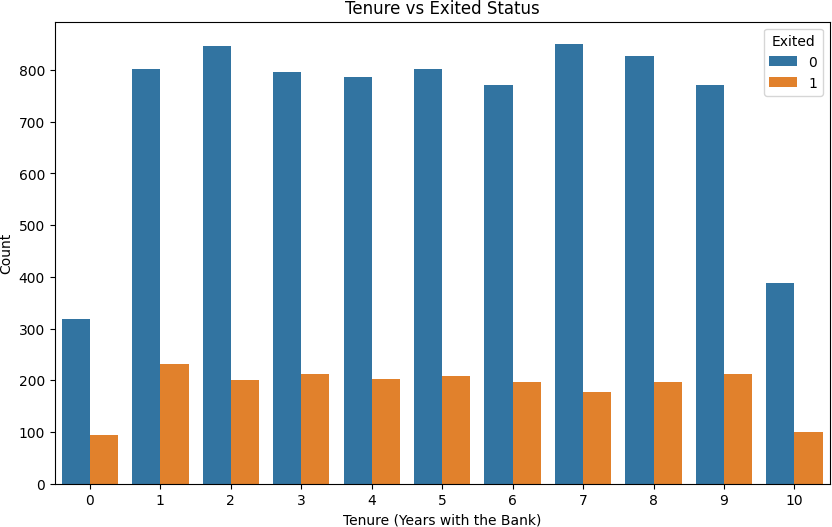
* + - * + **No Credit Card (0)**: Customers without a credit card have a higher churn rate compared to those who have one.
        + **Has Credit Card (1)**: Customers with a credit card are more likely to stay (Exited = 0) than to leave.

**=>** Owning a credit card may indicate a stronger financial relationship with the service, which could contribute to customer retention.



* There is no significant difference in the median credit scores between customers who churned (Exited = 1) and those who did not (Exited = 0).
* The range (from minimum to maximum) and distribution of credit scores are similar for both groups, suggesting that credit score is relatively consistent regardless of churn status.
* There are a few outliers on the lower end for customers who exited, but they do not substantially affect the overall distribution.

# => There is no clear difference in credit scores between customers who left and those who stayed, suggesting that credit score may not be a primary factor influencing customer churn.

****

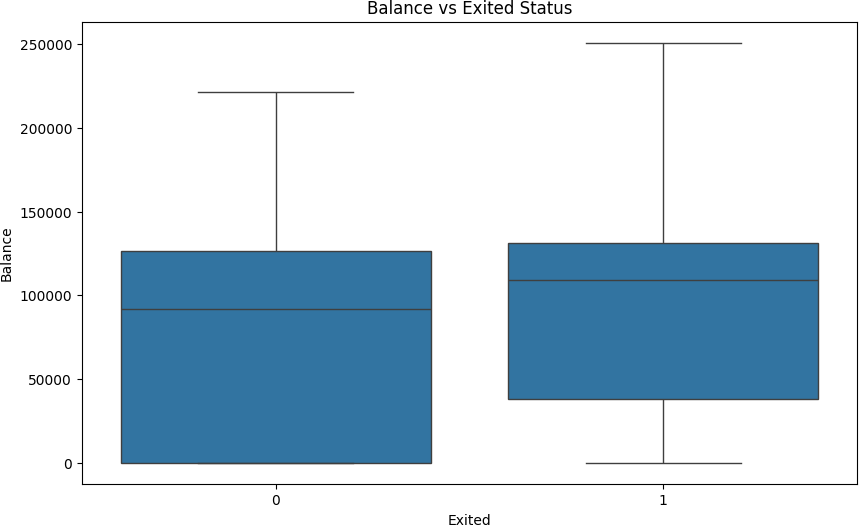
* **Current Customers (Exited = 0):**
  + The number of customers who stayed in each tenure group is stable over the years, ranging from around 800 to 900.
  + Regardless of tenure (from 0 to 10 years), the number of customers remaining with the bank does not vary significantly.

# Customers Who Left (Exited = 1):

* + The churn rate relative to retained customers is relatively low across different tenure groups.
  + The number of customers who left is fairly consistent across tenure years, ranging from around 100 to 300 in each group.

# No Clear Trend Based on Tenure:

* + Based on the chart, there is no clear trend indicating that customers with longer tenures have a higher or lower churn rate.
  + The number of customers who left and those who stayed remains nearly constant regardless of how long they have been with the bank.
* 0 and 10 years have both low numbers of customers in terms of staying and retention.



# Customers Who Stayed with the Bank (Exited = 0):

* + **Median:** The median balance for this group is approximately 100,000.
  + **Interquartile Range (IQR):** The range from the 25th to 75th percentile extends from about 0 to 120000.
  + **Upper and Lower Whiskers:** The upper whisker reaches nearly 220,000, while the lower whisker touches 0.
  + **Outliers:** No clear outliers are present in this group.

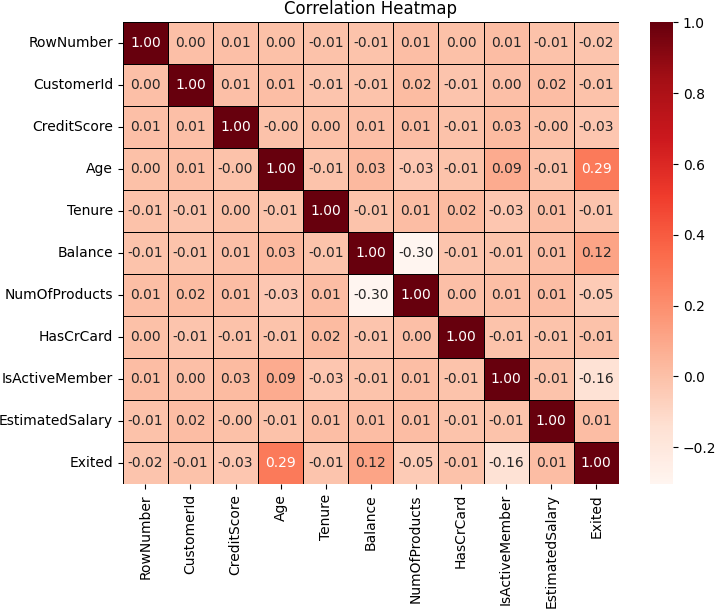
# Customers Who Left (Exited = 1):

* + **Median:** The median balance for customers who left is also approximately 100,000, similar to the customers who stayed.
  + **Interquartile Range (IQR):** The 25th to 75th percentile range is similar to the group that stayed, extending from around 0 to 130,000.
  + **Upper and Lower Whiskers:** The upper whisker extends to around 250,000, while the lower whisker touches 0.

# Comparison Between Groups (Exited = 0 vs. Exited = 1):

* + **Similar Median:** Both groups have a nearly similar median balance, indicating that the average account balance does not significantly differ between customers who left and those who stayed.
  + **Distribution:** Customers Who Left have wider Upper and Lower Whiskers than that of Who Stayed with the Bank.

=> **Balance is not a decisive factor:** Due to the similarity in balance distribution between the two groups, it can be inferred that account balance may not be an important factor affecting customers' decision to leave.



# Exited (Churned Customers):

* + **Correlation with Age (0.29):** This positive correlation indicates that older customers are more likely to leave the bank but it is quite weak.
  + **Negative Correlation with IsActiveMember (-0.16):** Active customers (frequent users of bank services) have a lower likelihood of leaving. This suggests that maintaining customer engagement is an important factor in retention.
  + **Negative Correlation with NumOfProducts (-0.03):** Although really weak,

this negative correlation shows that customers who use more products are less likely to churn.

* + **No Significant Correlation with Other Variables:** Factors such as

CreditScore, Balance, and EstimatedSalary have minimal impact on the likelihood of churn.

# Age:

* + **Negative Correlation with NumOfProducts (-0.30):** Younger customers tend to own more products compared to older customers.
  + **Positive Correlation with Exited (0.29):** Older customers have a higher churn rate.
  + **No Strong Correlation with Other Variables:** There is no notable association between age and other factors like Balance or CreditScore.

# NumOfProducts:

* + **Negative Correlation with Exited (-0.03):** Customers with more products are less likely to leave, although this relationship is weak.
  + **Negative Correlation with Age (-0.30):** Younger customers tend to own more products, indicating diverse service needs in this group.

# No Significant Correlation with Other Variables.

* **IsActiveMember:**
  + **Negative Correlation with Exited (-0.16):** Active customers are less likely to leave.
  + **No Strong Correlation with Other Variables:** This indicates that being an active member is relatively independent of other factors like Balance or EstimatedSalary.

# Balance:

* + **No Significant Correlation with Exited (0.07):** Account balance has minimal influence on customer churn.
  + **No Strong Correlation with Other Variables:** Balance does not have a strong relationship with other features in the dataset.

# EstimatedSalary:

* + **No Significant Correlation with Any Variables:** The estimated salary of customers has little relationship with other factors, suggesting that this may not be an important factor in churn analysis.

# Variables with No Significant Relationships:

* + **CreditScore:** No strong correlation with Exited or other variables, indicating that credit score may not be a key factor influencing customers' decisions to leave.
  + **Tenure:** Tenure has no strong correlation with Exited or other variables, suggesting that how long a customer has stayed with the bank does not have a strong link to whether they will leave or not.

# Recommendations

* + - * **Focus on Older Customer Segments:**
        + Analysis shows that the churn rate increases with age, particularly for customers aged 40 and above.
        + Suggestion: The bank should develop products and services suitable for older customers, such as long-term savings programs, retirement plans, or loan interest rate incentives to retain this group.

# Pay Attention to Customers in Countries with High Churn Rates:

* + - * + Churn rates vary by country, with Germany having the highest rate.
        + Suggestion: The bank should focus on improving service and enhancing customer care in areas with high churn rates like Germany, while also conducting surveys to identify specific reasons for customer departures.

# Consider Retention Policies for Female Customers:

* + - * + The churn rate is higher for females than males, indicating that women are more likely to leave.
        + Suggestion: The bank could implement customer care programs or financial products specifically for women to increase retention within this customer group.

# Increase Engagement of Inactive Customers:

* + - * + Inactive customers have a higher churn rate compared to active customers.
        + Suggestion: The bank should run incentive campaigns to increase interactions with inactive customers, such as providing benefits for frequent transactions or using banking services regularly.

# Conclusion of the Exploratory Data Analysis (EDA)

The EDA of the bank customer dataset reveals several insights into customer behaviors, characteristics, and factors potentially influencing churn. Key findings include:

# Customer Demographics and Behavior:

* + - * + **Age** has a notable positive correlation with churn, especially for customers over 40, who tend to leave at higher rates. This suggests that older customers may require specialized services to retain them.
        + **Geography** plays a role in churn, with **Germany** showing the highest churn rate, indicating regional differences in customer retention.
        + **Gender** analysis shows that **female customers** are more likely to churn than male customers, highlighting a potential need for targeted engagement or services for female clients.

# Product Engagement:

* + - * + Those with more products (3 to 4) show higher churn rates, implying that cross-selling additional products could increase churn rate.
        + **Inactive members** have a higher churn rate than active ones, underscoring the importance of maintaining regular customer interactions.

# Financial Variables:

* + - * + **Balance** does not significantly impact churn; customers with both high and low balances show similar retention patterns.
        + Other financial variables, like **CreditScore** and **EstimatedSalary**, do not show strong correlations with churn, suggesting they may not be primary factors in determining customer loyalty.

# Overall Retention Rate:

* + - * + The majority of customers stay with the bank, with only about 20% churned.

However, targeted strategies could further improve retention in identified at-risk segments (older customers, certain geographic regions, and inactive customers).

# CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING

# Data Preprocessing

* + 1. **Feature Selection**

We choose the most relevant features for clustering. Irrelevant or redundant features can introduce noise, making clusters harder to interpret and reducing clustering quality.

We will select numeric features that are most likely to influence the customer’s behavior, such as **CreditScore**, **Age**, **Tenure, Balance**, **NumOfProducts**, and **EstimatedSalary**. Avoid using features that directly indicate the outcome, like **Exited** in churn analysis, as these would bias the clustering towards that outcome.

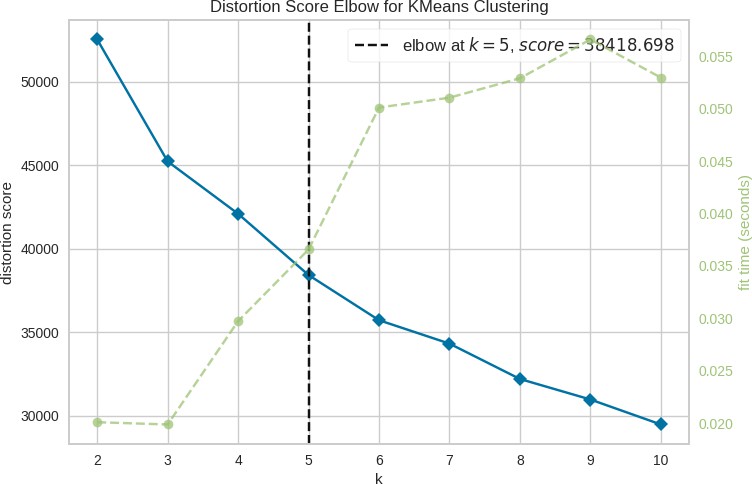
# Scaling data

K-Means is a distance-based algorithm (usually using Euclidean distance), so it is sensitive to feature magnitudes. Features with larger ranges can dominate the clustering outcome. Use **StandardScaler** to scale features so that they have a mean of 0 and a standard deviation of 1. Standardizing ensures that all features contribute equally to the clustering.

# The Elbow Method

The **Elbow Method** is a technique used to determine the optimal number of clusters in KMeans clustering. It involves plotting the distortion score (or within-cluster sum of squares) against the number of clusters, k. As k increases, the distortion score decreases because each cluster becomes smaller, which generally leads to lower intra-cluster variance.

However, after a certain point, the rate of decrease slows down, forming an "elbow" shape in the plot. This elbow point represents the optimal k value, where adding more clusters yields diminishing returns in terms of variance reduction.

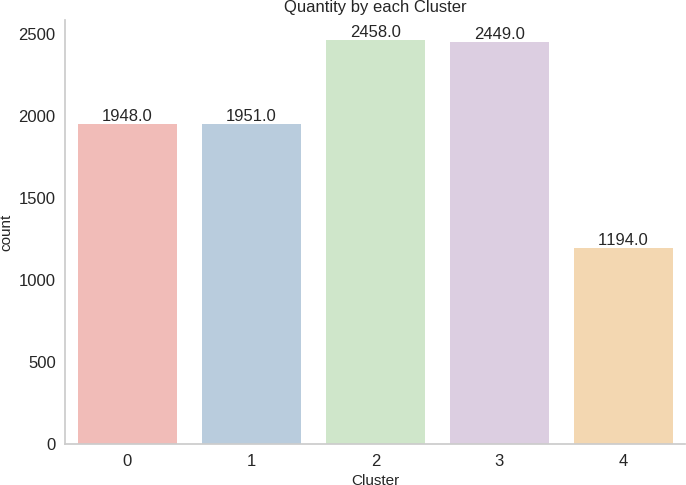


In the plot above, the elbow is observed at **k = 5** with a distortion score of approximately **38418.698**. This suggests that **5 clusters** is a suitable choice, as it balances the compactness of clusters with model simplicity.

# Clusters Characteristics

* + 1. **Quantity by each clusters**

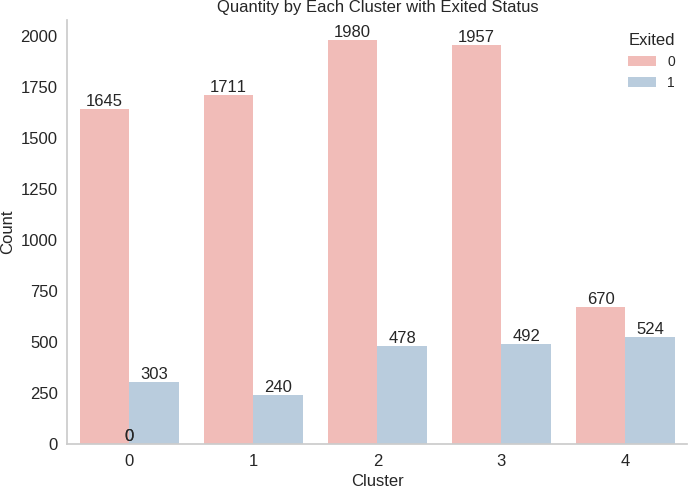
Given k = 5 is the optimal choice, we apply it into the K-Means model



From the graph above, we can see that:

* + - * **Cluster 2** has the highest number of customers, closely followed by **Cluster 3**. These two clusters represent the largest segments of the customer base.
      * **Clusters 0 and 1** have almost equal numbers of customers, indicating moderately sized customer segments.
      * **Cluster 4** has the smallest number of customers, suggesting a more distinct and potentially unique group within the customer base.

# Quantity by each clusters with Exited status:

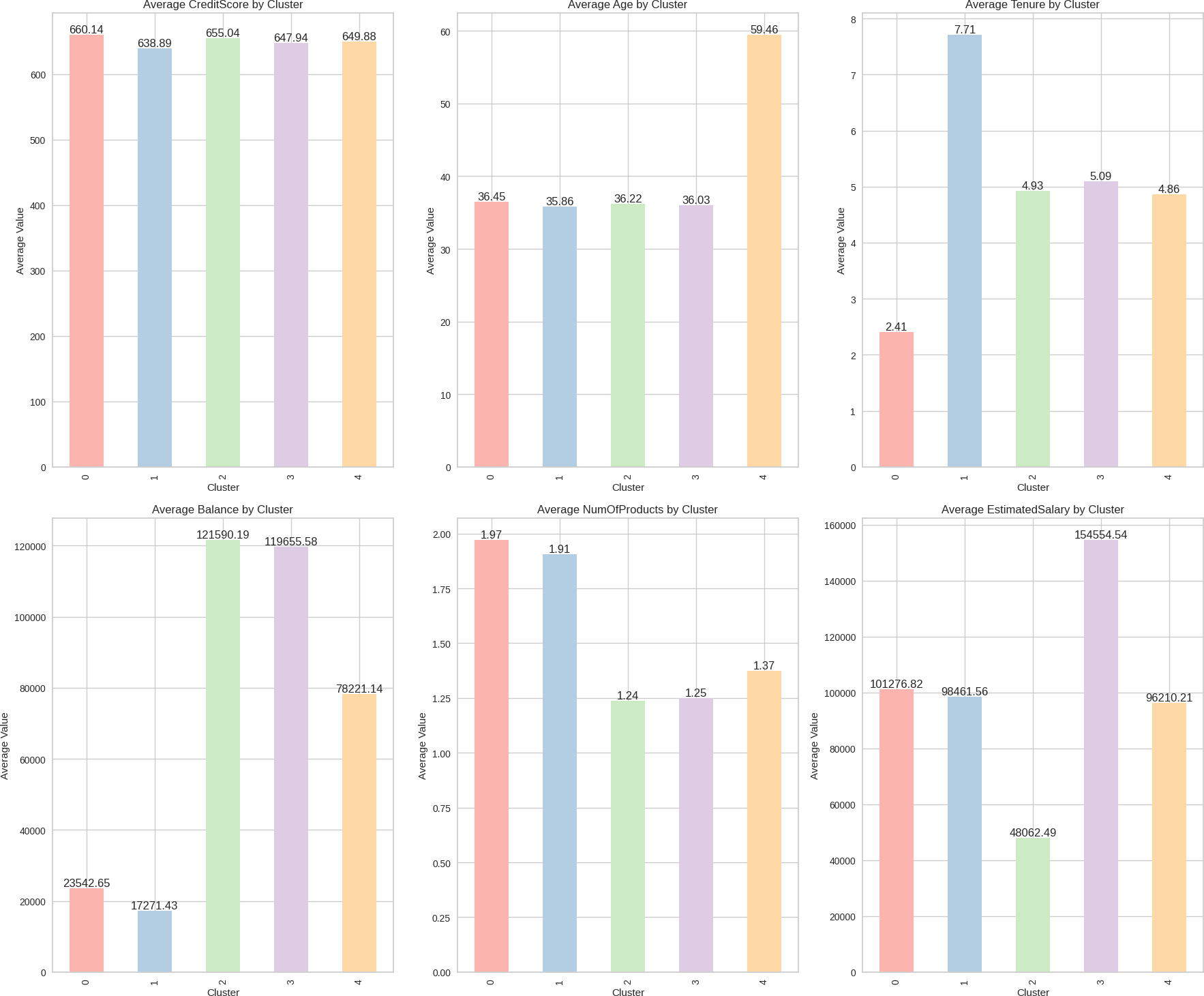
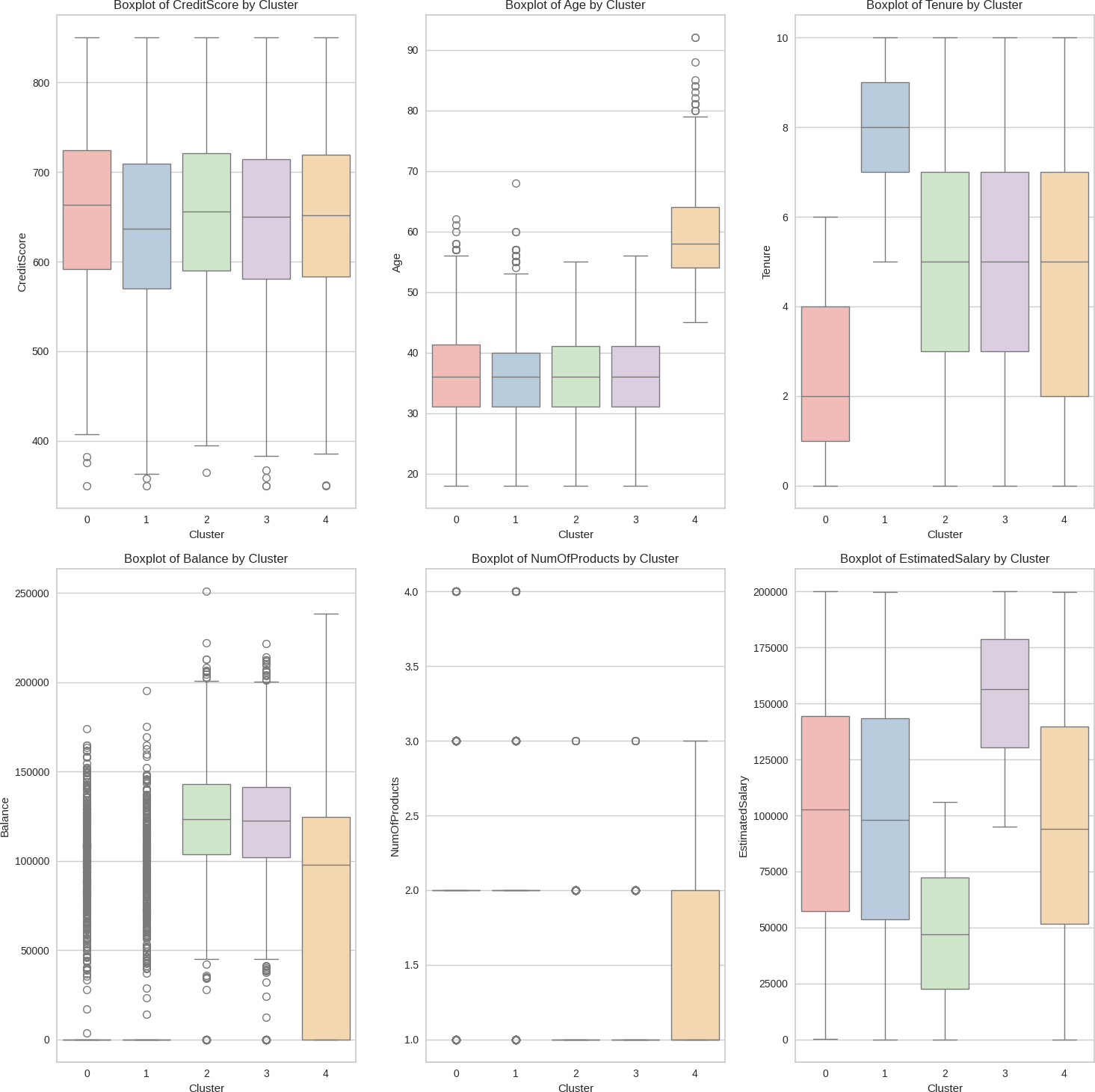
****

This graph highlights that **Cluster 4 has the highest rate of customer churn**, indicating a significant risk of attrition within this segment, while **Cluster 1 shows the lowest churn rate**, suggesting stronger customer retention.

In the next section, we will delve deeper into the characteristics of each cluster to better understand the factors driving these differences and identify targeted strategies for improving retention.

# Characteristics of each clusters and recommendations:

To gain a deeper understanding of the characteristics of each cluster, we generated boxplots and bar charts depicting the average values of key features. This approach provides a clear and detailed view of how each cluster varies, enabling targeted insights and tailored strategies for each segment.



Now, we will analyze the characteristics of each cluster and give recommendations.

# Cluster 0 - Younger, New Customers with Moderate Balances

* + - * **CreditScore**: The highest average credit score (~660), suggesting that these customers are financially responsible or have good credit management habits.
      * **Age**: The average age is ~36, indicating a younger demographic, possibly early in their financial journey.
      * **Tenure**: Average tenure is the lowest among all clusters (~2.4 years), so these customers are relatively new to the bank.
      * **Balance**: Lowest average balance (~23,542), which might indicate that these customers are just starting to build wealth or are not fully financially engaged with the bank.
      * **NumOfProducts**: Highest average (1.97), indicating that these customers are interested in using multiple products despite their low balances.
      * **EstimatedSalary**: Moderate salary (~101,767), suggesting a reasonable earning capacity, though not necessarily wealthy.

# Churn Analysis:

* + - * **Exited Rate**: Moderate churn rate, with 303 churned customers out of a total of 1,948 in the cluster. (15.55%)
      * **Interpretation**: These are likely younger, financially stable customers who have room to increase engagement with the bank. Their lower balances and short tenure may make them more susceptible to churn if they don’t see value quickly.

# Actionable Insights:

* + - * **Engagement Programs**: Introduce financial education programs or personalized recommendations to encourage increased savings or investment.
      * **Onboarding Experience**: As newer customers, focus on improving the onboarding experience to deepen their connection to the bank, perhaps through incentives for increasing product usage.

# Cluster 1 - Loyal, Long-Term Customers with Lower Balances and Salary

* + - * **CreditScore**: Lowest average score among all clusters (~638).
      * **Age**: Average age is similar to Cluster 0 (~35.9), indicating a relatively young demographic, yet they have a long relationship with the bank.
      * **Tenure**: The highest average tenure (~7.7 years), indicating strong loyalty and a long-standing relationship.
      * **Balance**: Lowest balance (~17,271), suggesting these customers have accumulated small funds or are not fully financially engaged with the bank..
      * **NumOfProducts**: Slightly below average (1.91), which may indicate they are selective with the products they use.
      * **EstimatedSalary**: Moderate salary (~98,461), indicative of a quite financially stable group.

# Churn Analysis:

* + - * **Exited Rate**: Lowest churn rate, with only 240 churned customers out of a total of 1,951 in this cluster. (12.53%)
      * **Interpretation**: This cluster represents stable, high-value, and long-term customers. They’re less likely to churn due to their strong relationship with the bank.

# Actionable Insights:

* + - * **Loyalty Rewards**: Implement loyalty rewards or benefits programs to reinforce their commitment.
      * **Exclusive Offers**: Consider offering exclusive products or services tailored to their financial stability, such as premium savings accounts, investment products, or personal financial advisors.

# Cluster 2 - High Balance, Selective Product Use, and Moderate Engagement

* + - * **CreditScore**: High (~655), close to Cluster 0’s level.
      * **Age**: Similar age range (~36 years) as Clusters 0 and 1.
      * **Tenure**: Moderate tenure (~4.9 years), indicating they’ve been with the bank long enough to build a relationship.
      * **Balance**: Highest balance (~121,509) across all clusters, indicating strong financial engagement.
      * **NumOfProducts**: Lowest among all clusters (~1.24), suggesting these customers are very selective with product choices.
      * **EstimatedSalary**: Lowest among all clusters (~48,062), possibly a low to middle-income group.

# Churn Analysis:

* + - * **Exited Rate**: Moderate churn rates, with only 478 churned customers out of a total of 2,458 in this cluster. (19.44%)
      * **Interpretation**: These are high-balance customers who may rely on the bank for specific purposes, such as holding substantial savings but are not interested in additional products.

# Actionable Insights:

* + - * **Cross-Selling Opportunities**: Since they use few products, introduce them to new options that complement their high balances, like investment services or high-interest savings accounts.
      * **Value Proposition**: Emphasize the unique value of each product to encourage further engagement, especially for customers who tend to be selective.

# Cluster 3 - Moderate Tenure, Low Salary, and Balanced Engagement

* + - * **CreditScore**: Moderate among clusters (~647).
      * **Age**: Average age is ~36, similar to other clusters.
      * **Tenure**: Moderate tenure (~5 years), showing some level of established relationship with the bank.
      * **Balance**: Moderate to high balance (~119,656), suggesting high financial stability.
      * **NumOfProducts**: Almost the lowest (1.25), similar to Cluster 2, but the balance of Cluster 3 is not as high as Cluster 2’s .
      * **EstimatedSalary**: Highest among all clusters (~154,554), possibly reflecting a high-income demographic.

# Churn Analysis:

* + - * **Exited Rate**: Moderate churn, with 492 exited customers out of 2,449 in this cluster. (20.08%)
      * **Interpretation**: This group may include customers with limited financial means who have a moderate attachment to the bank but aren’t fully engaged due to their lower salary.

# Actionable Insights:

* + - * **Affordable Products**: Offer products that align with their budget, such as low-fee accounts or credit-building products.
      * **Retention Programs**: Since they have moderate tenure and financial stability, offer incentives for further engagement, like reduced fees or targeted product education.

# Cluster 4 - Older, High-Balance Customers at High Risk of Churn

* + - * **CreditScore**: Slightly lower (~649), but still within acceptable range.
      * **Age**: Significantly older demographic (~59 years on average), indicating customers who are likely closer to retirement.
      * **Tenure**: High tenure (~4.9 years), showing a stable relationship with the bank.
      * **Balance**: Moderate average balance (~78,221), suggesting financial stability.
      * **NumOfProducts**: Low (~1.37), indicating they are not using many bank products.
      * **EstimatedSalary**: Moderate (~96,210), likely reflecting their established financial status.

# Churn Analysis:

* + - * **Exited Rate**: Highest churn rate, with 524 churned customers out of a total of 1,194 in this cluster. (43.88%)
      * **Interpretation**: This older, financially stable group is at high risk of churn. They may no longer need certain banking services or might be considering alternatives as they approach retirement.

# Actionable Insights:

* + - * **Retirement Planning Services**: Offer tailored retirement products, such as annuities or retirement planning consultations, to meet their needs at this life stage.
      * **Retention Incentives**: Provide loyalty-based benefits to encourage them to stay with the bank, possibly through fee reductions or better interest rates on deposits.

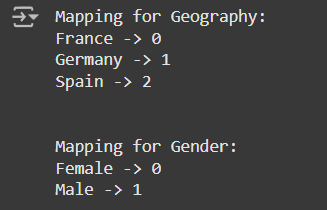
# EXPERIMENTS

# Data Preprocessing

In the Data Preprocessing phase of our churn analysis project, we aim to prepare the raw dataset for machine learning by transforming and refining it for optimal model performance. This phase is crucial as it ensures that the data is in a suitable format and scale for effective model training.

# Data Encoding

In the dataset, certain columns are categorical, and retaining them as strings prevents them from being used in modeling. Therefore, we need to encode these columns, converting text data into numeric format—a process known as encoding.

Following our Exploratory Data Analysis (EDA), we identified that the *Geography* and *Gender* columns are currently in string format. We will convert these columns into numerical format using Label Encoding. The resulting encoded output is shown below.

# Addressing Imbalanced data

From the Exploratory Data Analysis (EDA), we observed a significant class imbalance between customers who exited and those who did not. This imbalance can lead to several issues in model performance, as machine learning algorithms tend to favor the majority class when trained on imbalanced data. Consequently, the model may achieve high accuracy simply by predicting the majority class (non-exit) while failing to effectively identify instances of the minority class (exit), which is crucial for our churn analysis.

To mitigate this, we will apply the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates synthetic examples for the minority class (customers who exited) by interpolating between existing examples, effectively increasing the representation of this class. Over-sampling with SMOTE is preferable here because it allows us to retain all original data points from the majority class while balancing the dataset. This approach aims to improve the model's sensitivity to churn patterns, thereby enhancing its predictive performance and generalization ability.

Here is the output:



# Scaling data

We need to scale the data because different features in the dataset may have different units or ranges. For instance, some columns like *Age*, *Balance*, and *EstimatedSalary* may vary greatly in scale. Machine learning algorithms can be significantly affected by the scale of the features.

When features are on different scales, larger values can dominate the learning process, leading the model to place more weight on certain features over others. Common scaling techniques include:

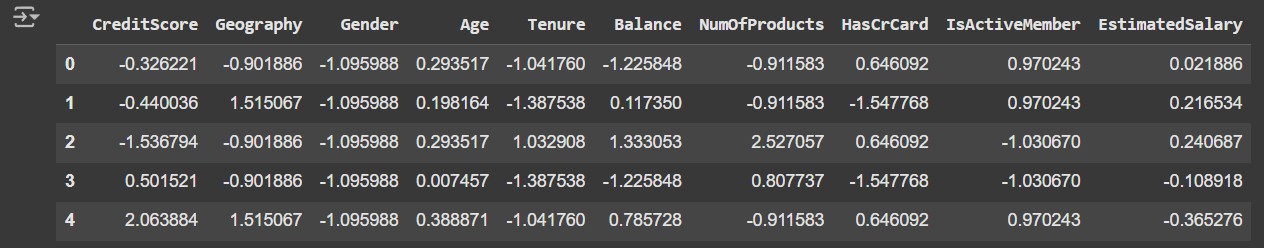
* + - * **Standardization**: Centers the data around zero with a standard deviation of one. This

is useful for data that follows a Gaussian (normal) distribution.

* + - * **Min-Max Scaling**: Scales the data to a fixed range, often [0, 1]. This is particularly helpful when all features are expected to have similar importance in the model.

For our churn analysis project, we will use **StandardScaler** to standardize the features.

Here is the output:



# Building models

In this project, We will develop 18 machine learning models. Then, we will select the top 3 models and perform hyperparameter tuning to find the most suitable model.

1. Logistic Regression
2. K-Nearest Neighbors
3. Linear SVM
4. Gradient Boosting Classifier
5. Decision Tree
6. Random Forest
7. Naive Bayes (Gaussian)
8. Naive Bayes (Bernoulli)
9. AdaBoost Classifier
10. Bagging Classifier
11. Extra Trees Classifier
12. Ridge Classifier
13. Passive Aggressive Classifier
14. Perceptron
15. Linear Discriminant Analysis
16. Quadratic Discriminant Analysis
17. XGBoost Classifier
18. LightGBM Classifier

# Evaluating Models on Training and Test Sets

In this step, we evaluate each model's performance on both the training and test datasets. The goal is to measure how well each algorithm learns from the training data and generalizes to unseen data in the test set. This evaluation process helps us identify any overfitting or underfitting issues.

* + - * **Training Accuracy (train\_score)**: By calculating the accuracy on the training set, we

can assess how well the model has fit the data it was trained on. A high training

accuracy suggests that the model has captured the patterns in the training data effectively. However, if the training accuracy is significantly higher than the test accuracy, it could indicate overfitting.

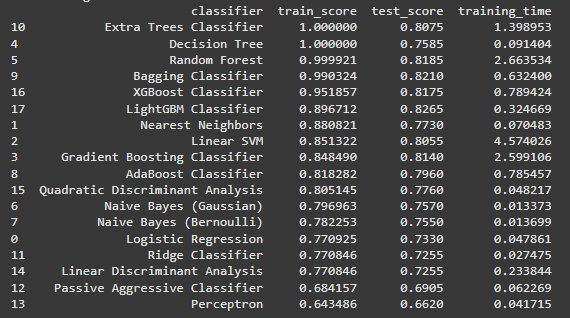
* + - * **Test Accuracy (test\_score)**: Calculating accuracy on the test set gives us an indication

of how well the model generalizes to new, unseen data.

* + - * **Training Time**: Recording the training time of each model is useful for assessing the computational efficiency of each algorithm. This information is particularly valuable when working with large datasets or limited computational resources.

By analyzing the train and test accuracies alongside the training time, we can compare the effectiveness and efficiency of each model. This allows us to select models that not only perform well in terms of accuracy but also train efficiently.

Here is the output:



# Overfitting Indicators:

* + - * + **Random Forest, Extra Trees Classifier, and Decision Tree** models achieve perfect or near-perfect accuracy on the training set (train score = 1.0 for some), but their test scores are significantly lower.
        + This discrepancy indicates **overfitting**, as these models memorize the training data but fail to generalize effectively to unseen data.
        + These models may need further tuning or regularization to mitigate overfitting.

# Balanced Performance:

* + - * + **LightGBM Classifier, Gradient Boosting Classifier, and AdaBoost Classifier** show a good balance between train and test scores, with similar values on both sets.
        + These models generalize well to new data, exhibiting minimal overfitting.
        + This balanced performance makes them promising candidates for further tuning and potential deployment, as they capture patterns in the data without overfitting.

# Underfitting Indicators:

* + - * + **Perceptron and Passive Aggressive Classifier** show relatively low train and test scores, indicating **underfitting**.
        + These models are not able to capture the patterns in the data effectively, resulting in poor performance on both training and test sets.
        + These models may benefit from additional feature engineering, tuning, or potentially using more complex algorithms if better performance is required.

# Training Time:

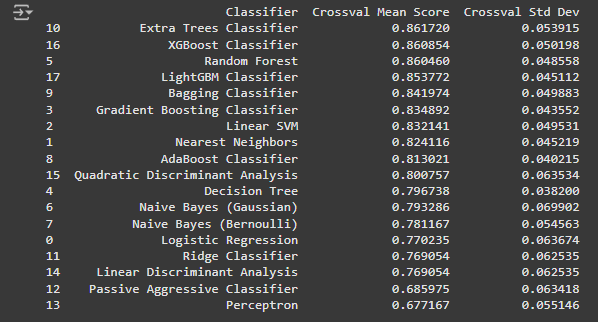
* + - * + The training time varies significantly across models, with **Linear SVM** taking the longest and models like **Naive Bayes (Gaussian)** and **Ridge Classifier** taking very little time.
        + Models like **Random Forest**, **XGBoost**, and **Bagging Classifier** also have relatively longer training times, though their test scores justify the added computation.

# Cross Validation

To better evaluate the models, I will use **cross-validation**. This technique provides a stronger assessment by dividing the dataset into multiple folds and training the model on different subsets of the data.

By implementing cross-validation, we aim to get a more comprehensive understanding of each model's generalization ability, which will help in selecting the top-performing models for further tuning.

Here is the output:



Based on the cross-validation results, we can select the top three models with the highest mean cross-validation scores. These models demonstrate the best performance in terms of consistent accuracy across multiple folds, as indicated by their mean scores and relatively low standard deviations. The top three models are:

1. **Extra Trees Classifier**
2. **XGBoost Classifier**
3. **Random Forest**

These models not only have the highest cross-validation scores but also demonstrate stability, as seen from their low standard deviation values. This stability indicates that these models perform consistently across different subsets of the data, making them strong candidates for further hyperparameter tuning to optimize performance.

# Hyperparameter Tuning

Hyperparameter tuning is the process of optimizing the hyperparameters of a machine learning model to improve its performance. For this project, we will use **Grid Search** to tune the hyperparameters of our selected models. By defining a range of possible values for each hyperparameter, Grid Search will systematically evaluate each combination to find the configuration that yields the highest performance based on our scoring metric. Although Grid Search can be time-intensive, it provides a comprehensive search, ensuring that we thoroughly explore the hyperparameter space and achieve the best possible model configuration.

# Choosing Recall as the Scoring Metric

For our customer churn analysis, we will use **Recall** as the scoring metric during hyperparameter tuning. Here is the reasons:

* + - * **Recall**, also known as sensitivity or true positive rate, measures the proportion of

actual positive cases (churned customers) that the model correctly identifies.

* + - * **Why Recall is Better for Customer Churn**: In a customer churn problem, identifying churned customers accurately is critical, as the business impact of missing a churned customer can be significant. Focusing on recall ensures that we capture as many churn cases as possible, reducing the number of missed cases (false negatives). High recall is essential because:
        + **Identifying Churned Customers**: Missing a churned customer (false

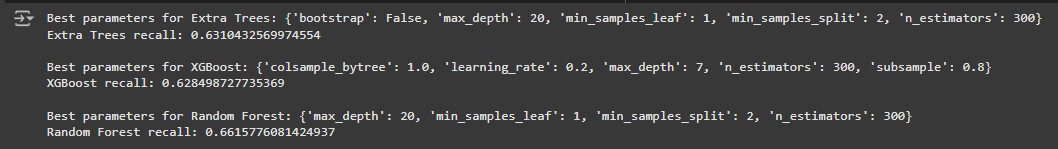
negative) can result in lost revenue, as these customers may leave without any retention efforts.

* + - * + **Prioritizing Retention Efforts**: With high recall, the model will capture most

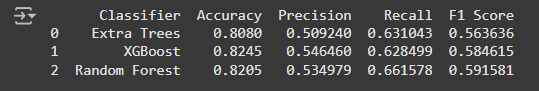
of the customers likely to churn, allowing the business to focus its retention strategies effectively.

While accuracy or precision may be valuable in other contexts, recall is more important in churn prediction because it prioritizes finding as many actual churners as possible, even if some non-churners are mistakenly identified as churners. This approach helps maximize the business's ability to retain valuable customers.

Here is the output:



The output shows the **best hyperparameters** found through Grid Search for each of the top three models, along with their corresponding **recall scores**.

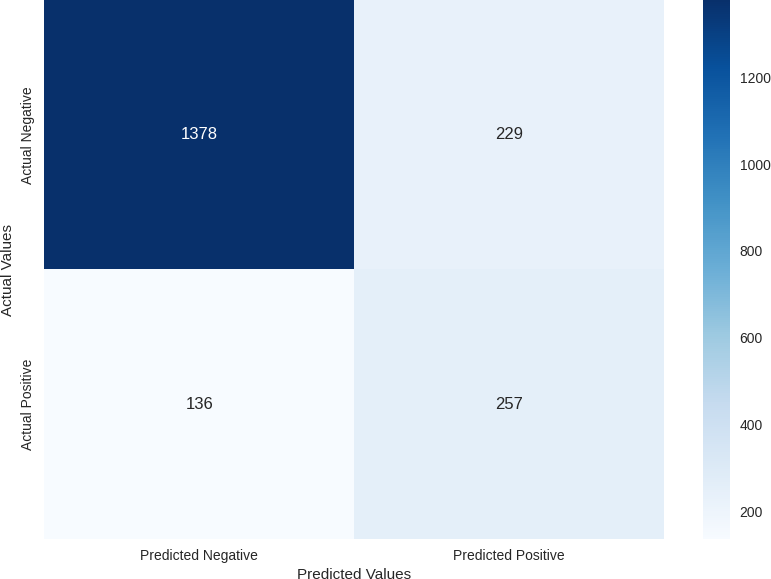


We will choose the **Random Forest** with its best hyperparameters obtained from our Grid Search results with the **highest recall**. Although the Random Forest model has slightly lower accuracy and precision compared to the others, its higher recall makes it the better choice for this application, where capturing as many potential churn cases as possible is prioritized.

# Evaluation

* + 1. **Confusion matrix**

The **confusion matrix** is a performance evaluation tool for classification models that provides insights into how well a model distinguishes between classes. It displays the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), which help assess the model's accuracy and its balance between sensitivity and specificity.



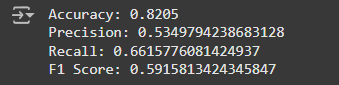
In this confusion matrix:

* + - * **True Negatives (TN)**: 1378 — The model correctly predicted 1378 customers as non-churn (Actual Negative, Predicted Negative).
      * **False Positives (FP)**: 229 — The model incorrectly predicted 229 customers as churn who were actually non-churn (Actual Negative, Predicted Positive).
      * **False Negatives (FN)**: 136 — The model missed 136 actual churn customers, predicting them as non-churn (Actual Positive, Predicted Negative).
      * **True Positives (TP)**: 257 — The model correctly identified 257 customers as churn (Actual Positive, Predicted Positive).

This confusion matrix highlights the model's ability to classify churn vs. non-churn customers. While the model successfully identified many actual churn cases (257), there are still a significant number of misclassifications (229 false positives and 136 false negatives), which suggests there is room for improvement.

We will examine other metrics relating to this confusion matrix, such as accuracy, precision, recall, and F1-score, in the next session to gain a more comprehensive understanding of the model’s performance.

# Accuracy, Precision, Recall, and F1 Score

****

* + - * **Accuracy**: 0.8205 — This metric represents the overall correctness of the model, calculated as the proportion of correct predictions out of total predictions. An accuracy of 82.05% indicates that the model performs fairly well in general. However, accuracy alone may not be the best measure for churn prediction, as it doesn’t specifically address the model's effectiveness in identifying churn cases.
      * **Precision**: 0.5349 — Precision measures how accurately the model identifies positive

cases (churn). In this case, a precision of 53.49% means that when the model predicts a customer will churn, it is correct slightly more than half of the time. While precision helps reduce false positives, it may miss true churn cases, which can be crucial in customer retention efforts.

* + - * **Recall**: 0.6616 — A recall of 66.16% means the model captures about two-thirds of

all churn cases. Recall is particularly important for churn analysis because identifying as many true churners as possible helps proactively target at-risk customers. This relatively high recall score suggests that the model is effective at detecting churn cases, which aligns well with the objective.

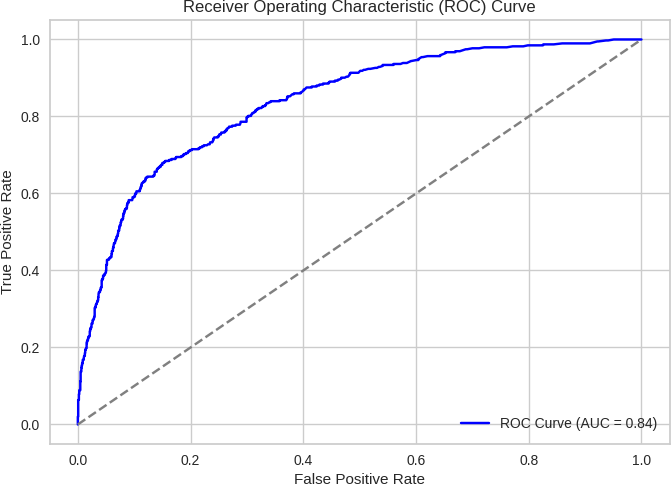
* + - * **F1 Score**: 0.5916 — The F1 Score is the harmonic mean of precision and recall,

balancing the two. An F1 score of 59.16% indicates a moderate trade-off between precision and recall, suggesting that while the model is reasonably effective in identifying churn cases, there is room for improvement, particularly in balancing false positives and false negatives.

Overall, the model's relatively high recall makes it suitable for churn prediction, where identifying true churners is a priority. However, the precision could be improved to reduce false positives, enhancing the model's reliability in predicting churn cases.

# ROC Curve

The ROC Curve plots the **True Positive Rate (Recall)** against the **False Positive Rate** at various thresholds, providing a view of the model's ability to distinguish between positive (churn) and negative (non-churn) cases across all thresholds.



* + - * The **AUC (Area Under the Curve)** score is **0.84**, which measures the model’s ability to distinguish between positive and negative classes (in this case, churn vs.

non-churn). AUC values range from 0 to 1, with values closer to 1 indicating stronger discriminatory power.

* + - * The **ROC curve** itself plots the **True Positive Rate (Recall)** against the **False**

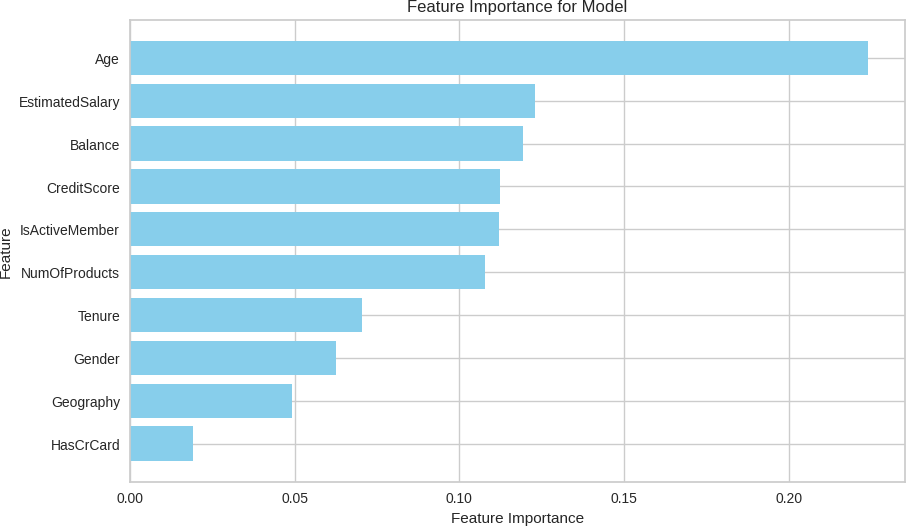
**Positive Rate** at various threshold settings. A steeper curve and higher AUC indicate that the model can achieve a high true positive rate while keeping the false positive rate low.

* + - * An AUC of 0.84 suggests that the model is fairly effective at distinguishing between

customers who are likely to churn and those who are not. This is a good result, as an AUC above 0.8 generally indicates strong performance, especially for customer churn prediction.

# Feature Importance

The **Feature Importance** chart shows the relative contribution of each feature in the predictions made by the **Random Forest** model. In Random Forest, feature importance is calculated based on how much each feature improves the purity (reduces error) of the splits in the trees across the ensemble. Higher feature importance indicates a greater impact on the model’s performance, while lower values suggest lesser influence.



The feature importance analysis highlights that **Age**, **EstimatedSalary** and **Balance** are the most influential factors in predicting customer churn. This information is valuable for developing targeted strategies to retain at-risk customers. Less influential factors, such as **Gender** and **HasCrCard**, might not require as much attention in retention strategies.

# LIMITATIONS

Despite the valuable insights derived from this analysis, there are several limitations to consider:

# Data Availability

Some important factors that might impact customer churn, such as customer satisfaction, service quality, or competitor activity, are not available in the dataset. Including these variables could potentially improve model accuracy and the depth of insights.

# Feature Selection and Engineering

Although the selected features provide meaningful predictions, there may be other, unobserved factors influencing customer churn. For instance, personal interactions with bank representatives, recent complaints, or external economic factors could also be significant predictors but were not included due to data unavailability.

Feature engineering could be further refined. For example, creating interaction terms or time-based features (such as changes in balance or activity trends) might capture more complex customer behaviors and enhance model performance.

# Imbalanced Dataset

Customer churn is an imbalanced problem, as the majority of customers do not churn. While techniques such as SMOTE were applied to address this, the imbalance may still affect the model's ability to generalize, potentially leading to over- or under-prediction of the minority class (churn).

Balancing techniques, while helpful, may not fully eliminate the bias toward the majority class, potentially limiting recall in identifying true churn cases.

# Model Limitations

While the Random Forest performed best among the models tested, it may not be the optimal model for all cases. This model’s effectiveness is constrained by its interpretability and the computational resources required, especially as the dataset scales or if real-time predictions are desired.

# Interpretation and Implementation Challenges:

While the model provides insights into which features are most predictive of churn, understanding the causality behind these factors requires further qualitative analysis. For example, while age and number of products are associated with churn risk, the underlying reasons may vary widely and require further investigation.

Implementing the recommended retention strategies may require organizational changes, resources, and alignment across multiple departments. This could be a complex and

time-consuming process, limiting the immediate impact of the insights gained.

# CONCLUSION

The customer churn analysis project provides a comprehensive understanding of the key factors that drive customer churn at the bank. Through EDA, clustering, and predictive modeling, we have identified significant patterns in customer behavior and built a model to predict potential churners. Here is a summary of the main findings and actionable insights:

# Age is the key Predictors of Churn

Age was identified as the most influential factor in predicting churn. Older customers are more likely to churn, which could be due to differing financial needs, such as retirement planning or reduced engagement with the bank’s products. Targeting older customers with specialized products and personalized service could help in reducing churn.

# Customer Segmentation:

The KMeans clustering analysis revealed 5 distinct customer segments based on behavior and financial characteristics. This segmentation can be used to tailor marketing and retention strategies:

* **Cluster of High Churn Risk**: The analysis identified clusters with high churn rates,

particularly those who are in cluster 4 (Older, High-Balance Customers). This cluster should be prioritized for retention efforts.

* **Loyal and Engaged Cluster**: Some clusters represent customers with a long tenure

and multiple products, showing lower churn risk. For these customers, loyalty programs and value-added services can strengthen their commitment further.

# Model Performance:

The **Random Forest** model emerged as the best choice, with a balanced performance across recall, precision, and F1 score. While its recall was only slightly higher than other models, its AUC score of 0.84 demonstrates strong discriminatory power, making it effective in distinguishing between churners and non-churners. By prioritizing recall, the model aligns with the business goal of capturing as many actual churn cases as possible to help reduce customer attrition.

# Retention Strategies Based on Insights:

* **Enhanced Engagement for High-Risk Segments**: Focus on engaging older customers and those with low product usage by offering personalized consultations, targeted promotions, and bundled products that cater to their specific financial needs.
* **Cross-Selling and Upselling**: Encourage customers with fewer products to explore additional offerings. This could be achieved through tailored marketing campaigns, discounts on bundled products, and incentive programs for multi-product engagement. However, it’s important not to pressure customers excessively, as those with 3 or 4

products already show a high churn rate. Instead, focus on encouraging customers to purchase only the products that genuinely meet their needs.

* **Proactive Approach for High-Balance Customers**: For high-balance customers who

may be looking for better investment opportunities, consider introducing high-yield savings accounts, exclusive investment products, or premium services to retain their loyalty.

* **Focus on Inactive Members**: Increase engagement with inactive members through

targeted outreach, loyalty rewards, or periodic account reviews. These customers may respond positively to re-engagement efforts, reducing their likelihood of churning.

# Overall Impact:

By implementing the strategies derived from this analysis, the bank can proactively address customer needs, reduce churn, and enhance customer satisfaction. Retention efforts focused on high-risk clusters and key predictors will improve the bank's ability to retain valuable customers and foster long-term loyalty.

# Final Recommendation

This data-driven approach provides the bank with actionable insights to retain at-risk customers and strengthens its overall retention strategy. Regularly updating the model and monitoring key predictors such as age, number of products, balance, and salary will ensure that the bank remains responsive to evolving customer behavior. By leveraging the Random Forest and the insights from this analysis, the bank can proactively manage customer churn, improve service delivery, and ultimately achieve a stronger customer relationship.