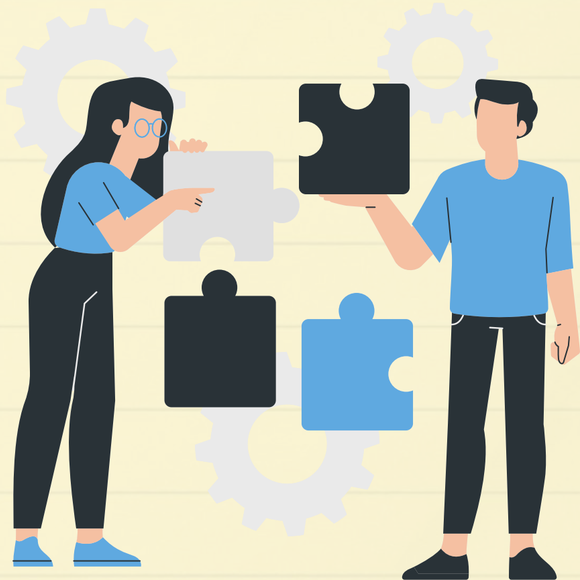
PERSONAL PROJECT #5

BANK CUSTOMER CHURN

Jihan Rana Ayunda Dewana



CONTENTS

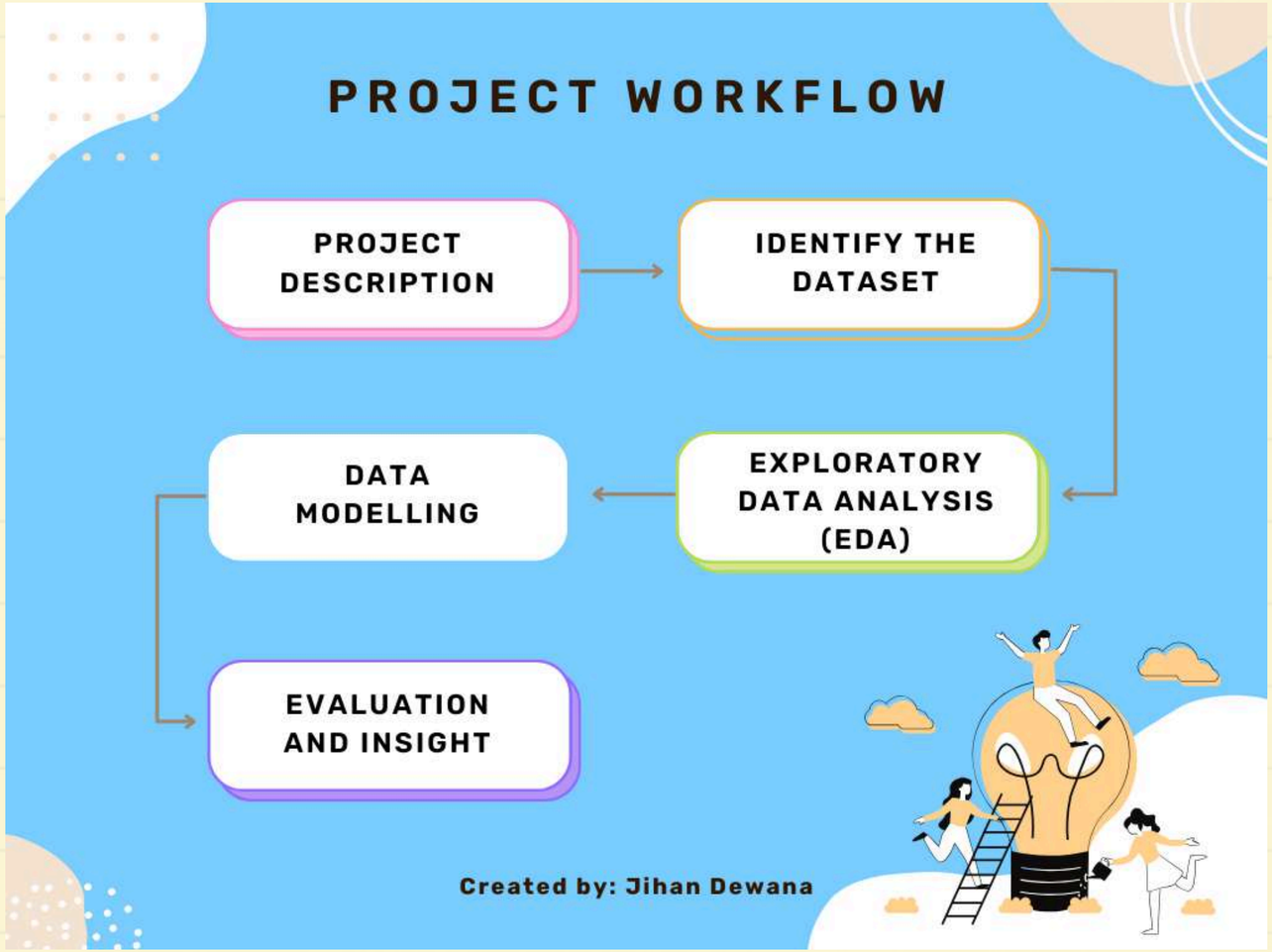
Tools and Language : 

PROJECT DESCRIPTION AND WORKFLOW

Customer churn refers to the number or percentage of customers who cease using a business's product or service within a specified time period. This metric is important because it indicates levels of customer satisfaction and loyalty, as well as potential lost revenue and negative impact on brand image if not handled effectively.

In this project, we have data on 10,000 customers at a European bank, including details on their credit scores, balances, products, and whether they have churned. The primary goal is to identify patterns and factors that lead to customers leaving a bank and to predict which customers are likely to churn in the future.

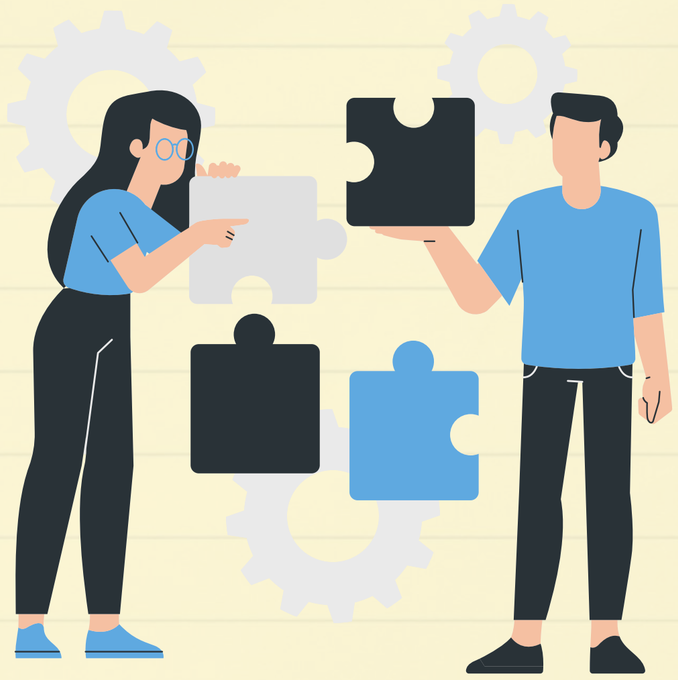
PROJECT DESCRIPTION AND WORKFLOW



IDENTIFY THE DATASET

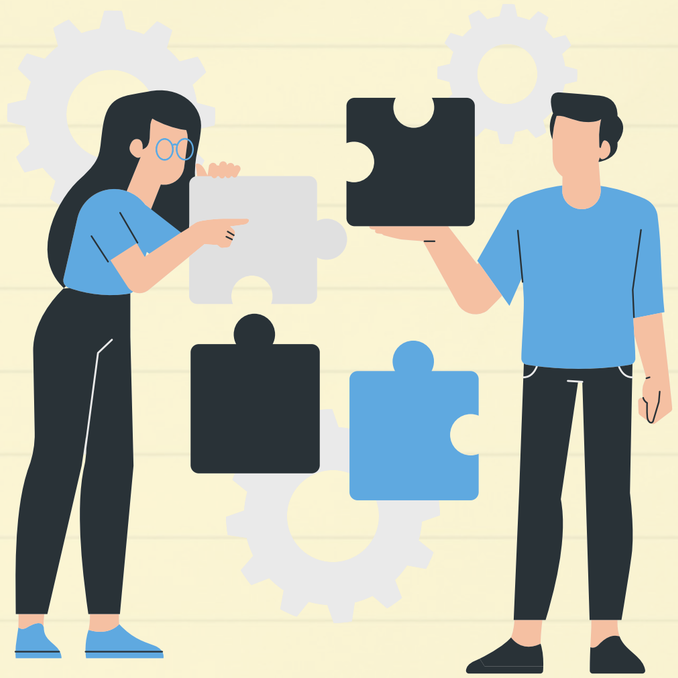
Data source is from Kaggle
There are the details of the dataset we used in this project.

CLASSIFICATION	VARIABLE & DEFINITION
Key Identifiers	• **RowNumber:** A sequential number for each record • **CustomerId:** A unique identity number of each customer • **Surname:** The customer's last name
Financial Indicators	• **CreditScore:** A numerical value that represents your credit worthiness. Based on FICO (Excellent: 800 - 850, Very Good: 740 - 799, Good: 670 - 739, Fair: 580 - 669 and Poor: 300 - 579). A higher credit score indicates a lower likelihood of churn • **Balance:** The amount of the customer's account. Higher balances are less at risk of churn • **EstimatedSalary:** The estimated salaries of customers. A larger income represents financial stability, and vice versa.



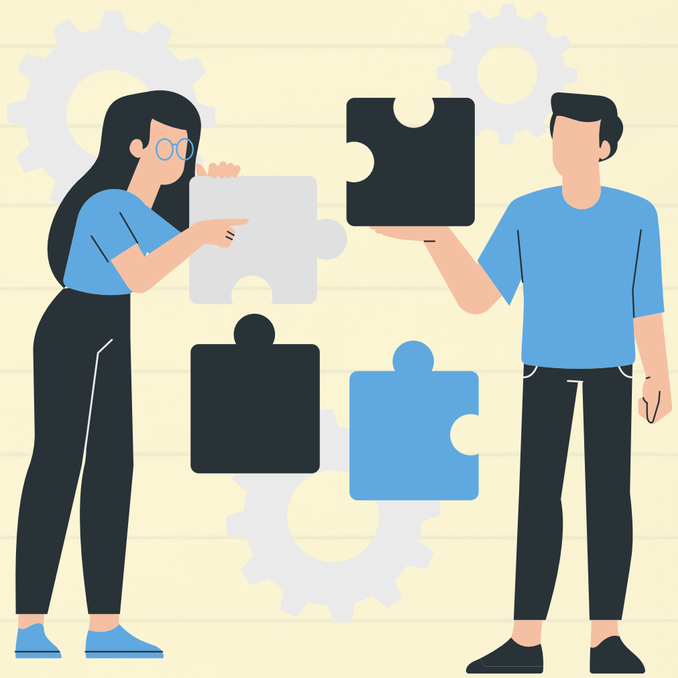
IDENTIFY THE DATASET

CLASSIFICATION	VARIABLE & DEFINITION
Demography Indicators	• **Geography:** The customer's location/region. It will affect the behaviour of the customer to churn or not. • **Gender:** The customer's gender (Female/Male). May shows a pattern of churn. • **Age:** The customer's age. Mostly, the oldest customers are more loyal and less likely to churn.
Customer Indicators	• **Tenure:** The number of years the customer has been a customer of the bank. Longer tenure suggests higher loyalty. • **NumofProducts:** The number of bank products the customer uses. More products may shrink churn likelihood. • **HasCrCard:** Whether the customer has a credit card (0 = No, 1 = Yes). Credit card users are less likely to churn.

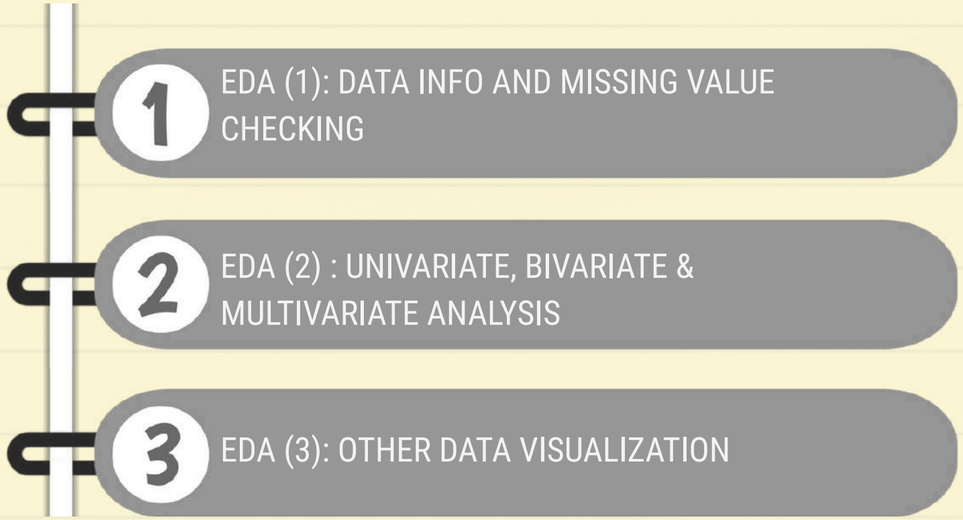


IDENTIFY THE DATASET

CLASSIFICATION	VARIABLE & DEFINITION
Customer Indicators	• **IsActiveMember:** Whether the customer is an active member (0 = No, 1 = Yes). Active users are less likely to churn. • **Card Type:** The type of credit card that customers use. Illustrate the financial stability. • **Points Earned:** The points earned by the customer for using a credit card. Greater points earned indicated the loyal one. • **Complain:** Whether the customer has filled a complain (0 = No, 1 = Yes). Complaints are a strong indicator of potential churn. • **Satisfaction Score:** The rating of the outcome of the complaint resolution. Impacts churn likelihood.
Dependent Variable	• **Exited:** Whether the customer left the bank (0 = Stayed, 1 = Left)



EXPLORATORY DATA ANALYSIS (EDA)



Exploratory Data Analysis (EDA) is the crucial first step in any data science project. It's the process of investigating a dataset to summarize its main characteristics, uncover hidden patterns, identify anomalies, and form hypotheses.

The purpose of EDA is to:

- Understand the data's structure and content
- Identify missing values and data errors
- Uncover trends and relationships
- Detect outliers and anomalies
- Formulate hypotheses

EDA (1): DATA INFO AND MISSING VALUE CHECKING

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 18 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   RowNumber           10000 non-null  int64
 1   CustomerId          10000 non-null  int64
 2   Surname             10000 non-null  object
 3   CreditScore         10000 non-null  int64
 4   Geography           10000 non-null  object
 5   Gender              10000 non-null  object
 6   Age                 10000 non-null  int64
 7   Tenure              10000 non-null  int64
 8   Balance             10000 non-null  float64
 9   NumOfProducts       10000 non-null  int64
 10  HasCrCard           10000 non-null  int64
 11  IsActiveMember      10000 non-null  int64
 12  EstimatedSalary     10000 non-null  float64
 13  Exited              10000 non-null  int64
 14  Complain            10000 non-null  int64
 15  Satisfaction Score  10000 non-null  int64
 16  Card Type           10000 non-null  object
 17  Point Earned        10000 non-null  int64
dtypes: float64(2), int64(12), object(4)
memory usage: 1.4+ MB
```

	0
RowNumber	0
CustomerId	0
Surname	0
CreditScore	0
Geography	0
Gender	0
Age	0
Tenure	0
Balance	0
NumOfProducts	0
HasCrCard	0
IsActiveMember	0
EstimatedSalary	0
Exited	0
Complain	0
Satisfaction Score	0
Card Type	0
Point Earned	0

dtype: int64

```
Percentage of Total Missing Values is  0.0 %
Missing Value Estimation :
```

```
[ ]  ## checking duplicate data
     data.duplicated().sum()

     np.int64(0)
```

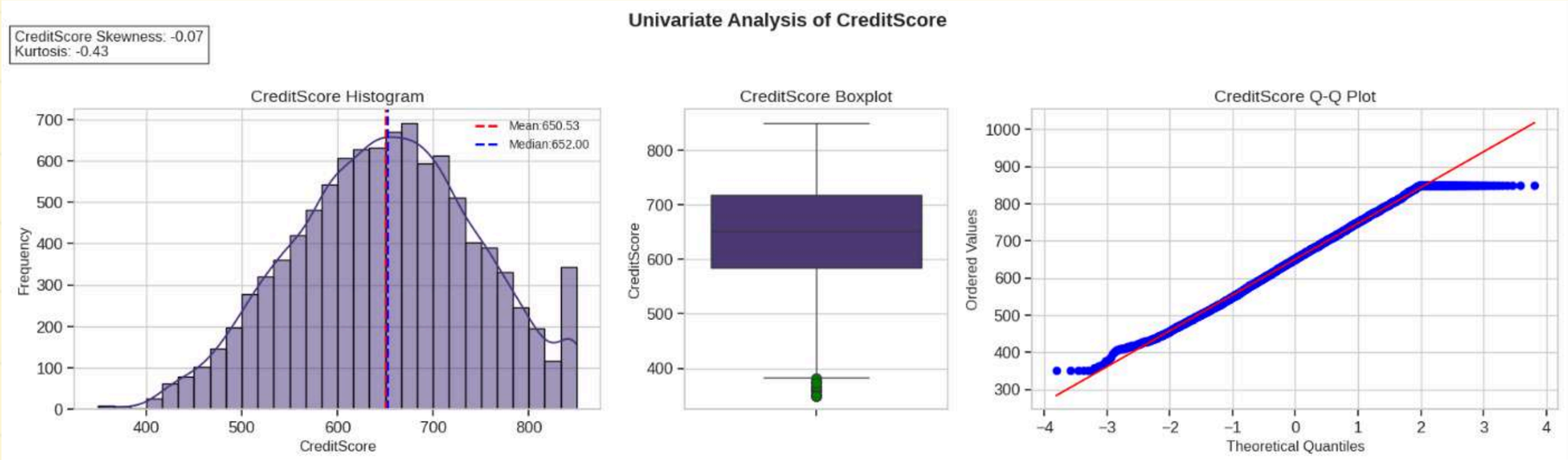
Based on data from BankCustomerChurn.csv, there are 10,000 data entries with 18 variables.

- Two variables with floating are Balance and EstimatedSalary.
- Four variables with object/string are Surname, Geography, Gender, and Card Type.
- Afterwards, the last variable with integer types is RowNumber, CustomerId, CreditScore, Age, Tenure, NumOfProducts, HasCrCard, IsActiveMember, Exited, Complain, Satisfaction Score, and Points Earned.
- There is no null or missing value on data set. Also, there is no duplicate data.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

Method	Method Detail
Univariate	Focusing on a single variable to understand its distribution, central tendency, and spread. We divide the data into two types: numerical and categorical. For the detailed techniques • Numerical: histogram, boxplot, Q-Q plot, skewness & kurtosis • Categorical: pie chart and table visualization
Bivariate	The study of the relationship between two variables. We use Plots and bar charts, which are essential for understanding relationships.
Multivariate	The analysis of three or more variables to understand their complex interactions. We use several methods, starting from the correlation matrix, charts, and plots for visualization. We then applied machine learning, which we'll cover in the next part.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

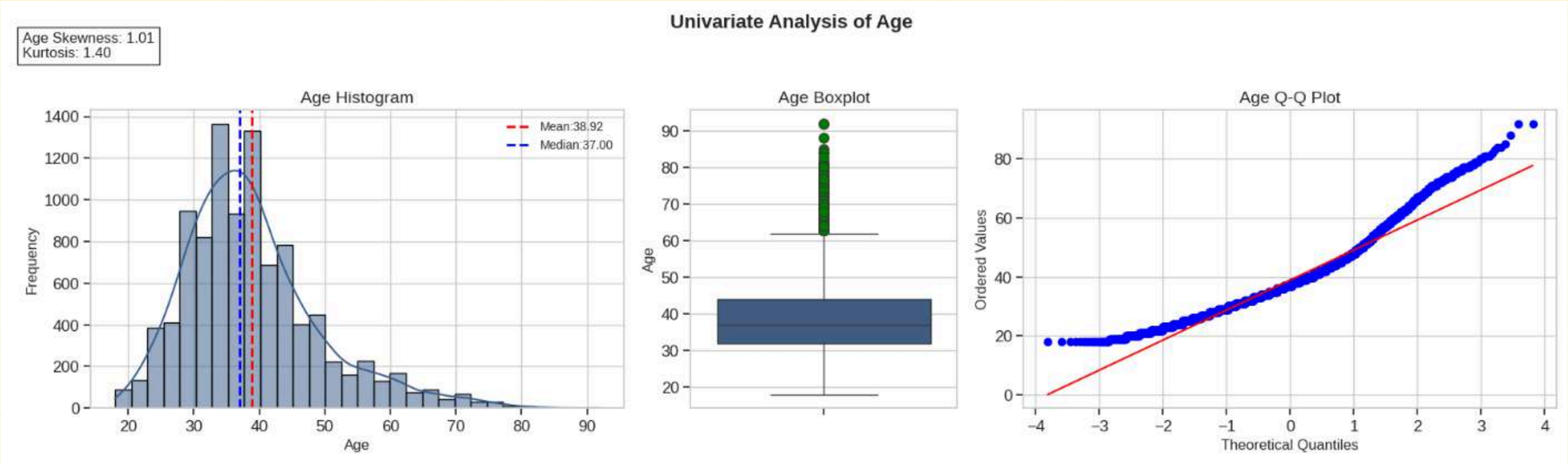


CreditScore (Mean: 650.53, Median: 652.0, Skewness: -0.07, Kurtosis: -0.43, R^2: 0.9941)

- The histogram displays a roughly bell-shaped distribution with a slight left skew (skewness: -0.07), indicating that major customers have higher credit scores.
- The Q-Q plot's high R^2 (0.9941) confirms that the data is nearly normally distributed, with minor deviations at the tails.
- The box plot shows a few outliers below 400, which could represent high-risk customers. This near-normal distribution suggests that the CreditScore is already well-behaved for modeling purposes.

Because the CreditScore near-normal distribution, no transformation is needed. However, the few outliers below 400 may need attention.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

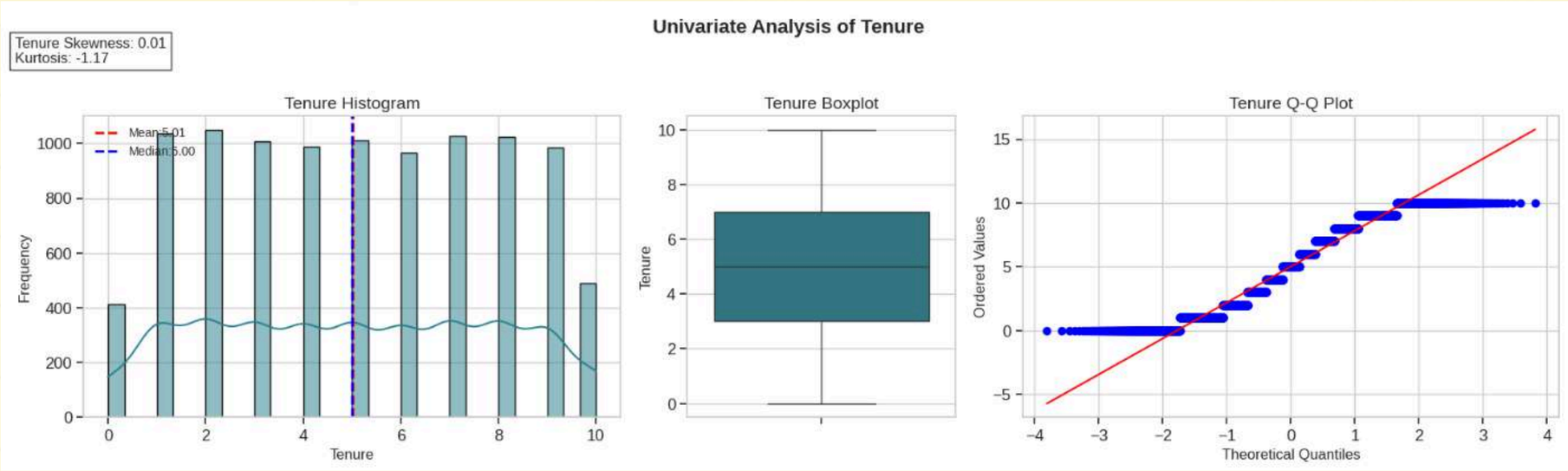


Age (Mean: 38.92, Median: 37.0, Skewness: 1.01, Kurtosis: 1.39, R^2: 0.9441)

- The age distribution is right-skewed (skewness: 1.01), with a peak around 30–40 years and a long tail extending to 90+.
- The Q-Q plot (R^2 = 0.9441) displays deviation from normality, particularly in the upper tail, indicating a non-normal distribution.
- The box plot reveals outliers above 60, likely representing older customers. This skewness suggests younger customers dominate the dataset, which aligns with the dataset description that younger customers are more likely to churn.

Based on the analysis above, we suggest a log transformation to reduce skewness and improve normality

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

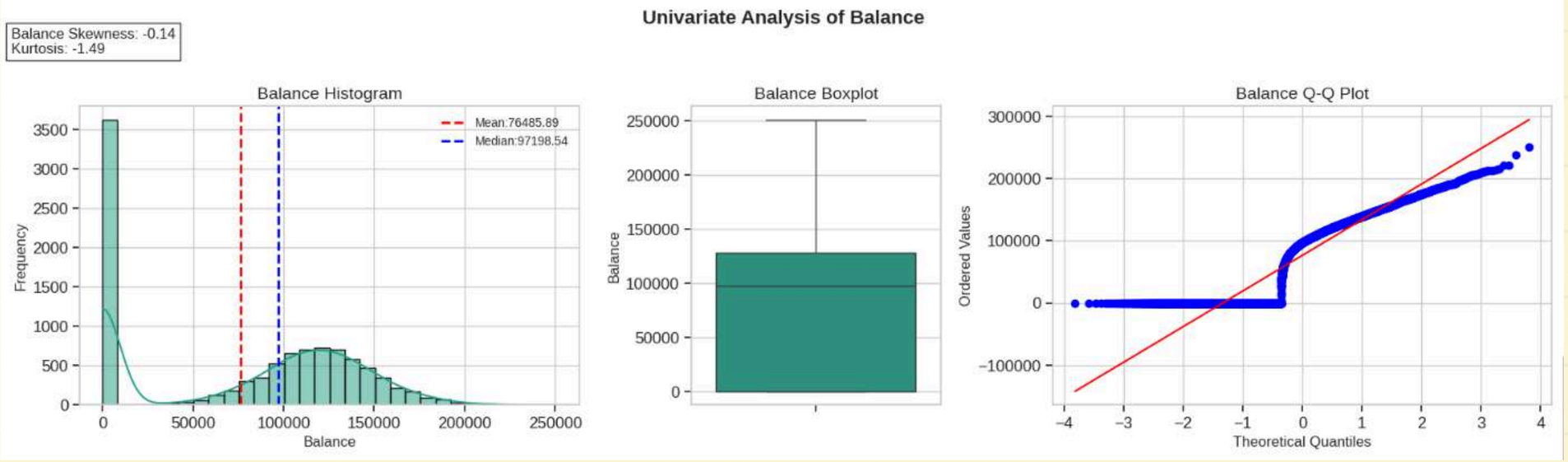


Tenure (Mean: 5.01, Median: 5.0, Skewness: 0.01, Kurtosis: -1.17, R^2: 0.9489)

- Tenure is moderately uniform, with values ranging from 0 to 10 years and a near-zero skewness (0.01). The histogram shows a relatively flat distribution.
- The Q-Q plot (R^2: 0.9489) indicates it's not perfectly normal, as expected for a discrete variable.
- The box plot shows no outliers, confirming the even spread. Customers are equally represented across the various loyalty durations, as shown by the uniform distribution.

The variable is already in a desirable state for modeling. Its discrete, uniform distribution, confirmed by the value of Rsqd, eliminates the need for any transformation.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

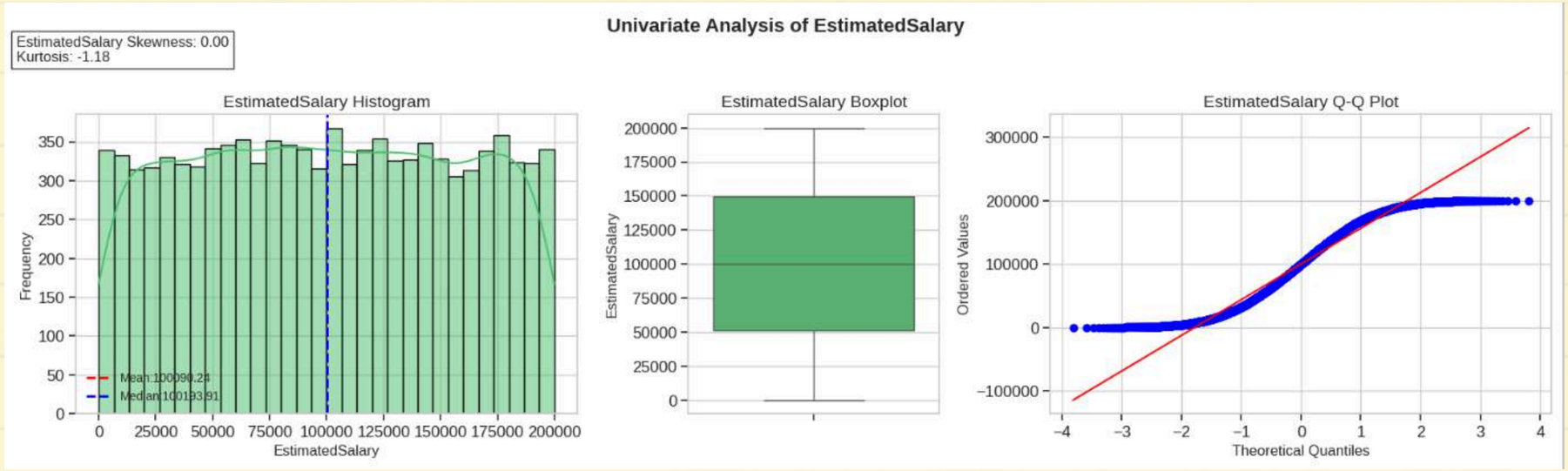


Balance (Mean: 76485.89, Median: 97198.54, Skewness: -0.14, Kurtosis: -1.49, R^2: 0.8458)

- The balance distribution has a unique shape, characterized by a significant peak at 0 (indicating that many customers have no balance) and a right-skewed spread for non-zero balances (skewness: -0.14).
- The Q-Q plot's low R^2 (0.8458) confirms non-normality, with significant deviation due to the zero-inflated nature.
- The box plot shows no significant outliers, but the zero peak is a key characteristic. It suggests a mix of inactive (zero-balance) and active accounts.

To normalize the skewed balance data, we can apply a logarithmic transformation (specifically, log1p to handle zero values). Alternatively, we could create a binary feature called HasBalance to distinguish between customers with and without a balance.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

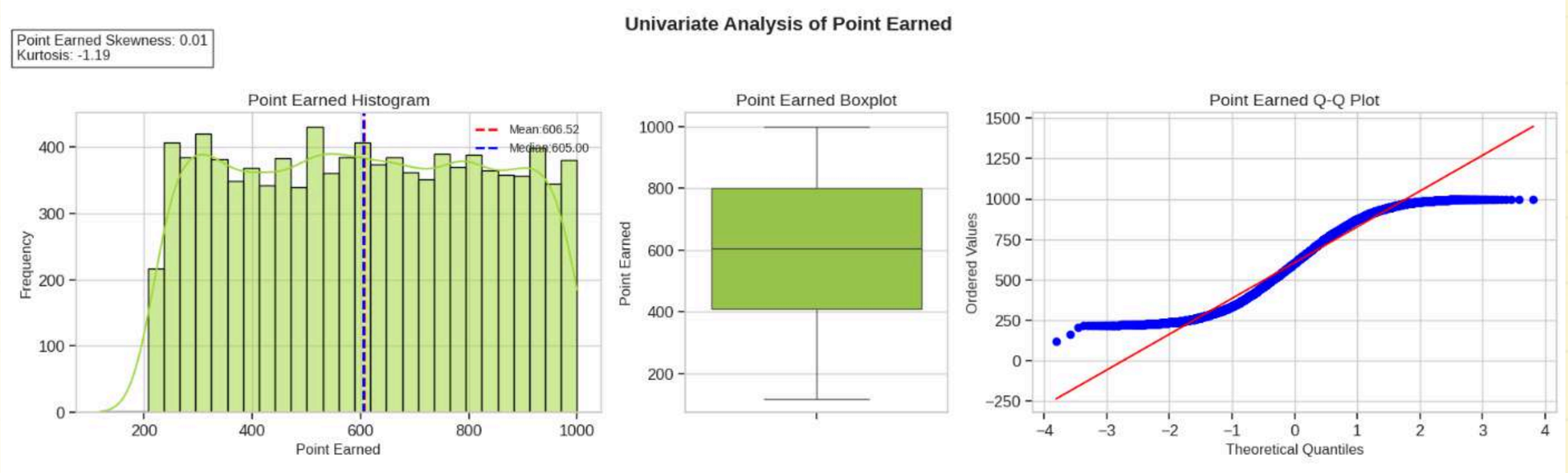


EstimatedSalary (Mean: 100090.24, Median: 100193.92, Skewness: 0.00, Kurtosis: -1.18, R^2: 0.9569)

- The salary distribution is almost uniform across the range (11.58 to 199992.48), with a skewness of 0.00.
- The Q-Q plot (R^2 = 0.9569) shows a slight deviation from normality, which is typical for a uniform distribution.
- The box plot shows no outliers, indicating a balanced spread. This uniformity suggests the bank serves customers across an extensive income spectrum.

The variable's uniform distribution makes it suitable for direct use in models. With an R^2: 0.9569, there's no significant skew, so no transformation is needed.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

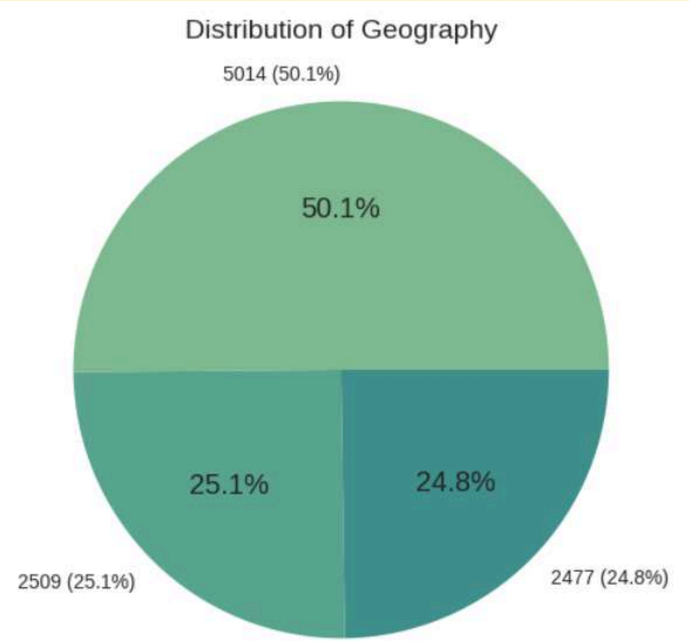


Point Earned (Mean: 606.52, Median: 605.0, Skewness: 0.01, Kurtosis: -1.19, R^2: 0.9555)

- Points earned are relatively uniform between 119 and 1000, with a slight peak around 400–600 (skewness: 0.01).
- With an R^2: 0.9555, the Q-Q plot shows that the distribution is very close to a theoretical ideal. Although there are slight departures from the straight line, the data's variance is well-described by, and the distribution is considered nearly uniform.
- The box plot shows no outliers, confirming an even spread. These suggest consistent credit card usage across customers, with no extreme earners.

The near-uniform distribution (R^2: 0.9555) is fine as is, with no transformation needed.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS



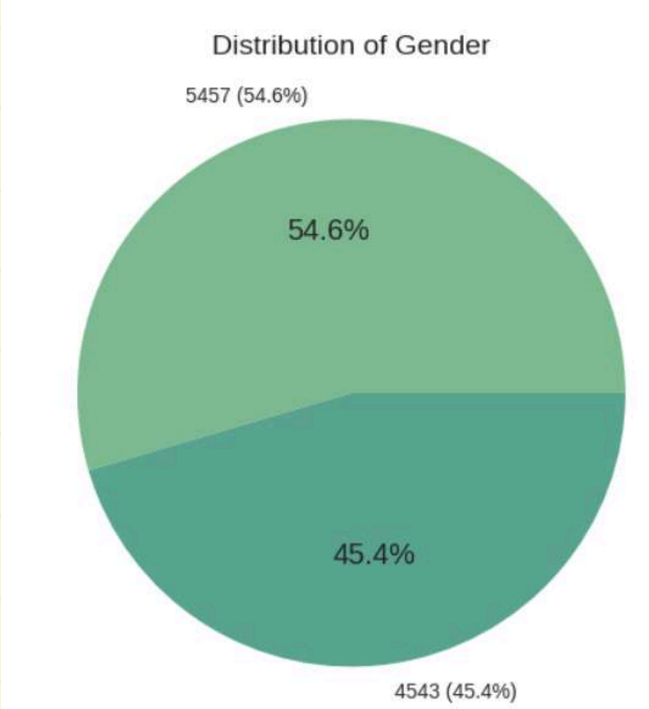
Value	Count	Percentage
France	5014	50.1
Germany	2509	25.1
Spain	2477	24.8

Geography (France: 5014, Germany: 2509, Spain: 2477).

- France dominates with 50.1%, while Germany and Spain each account for 25.1% and 24.8%, respectively. The distribution suggests that churn is not uniform across regions.
- **France, in particular, seems to have a higher churn rate** or distinct retention problems compared to other geographic areas.

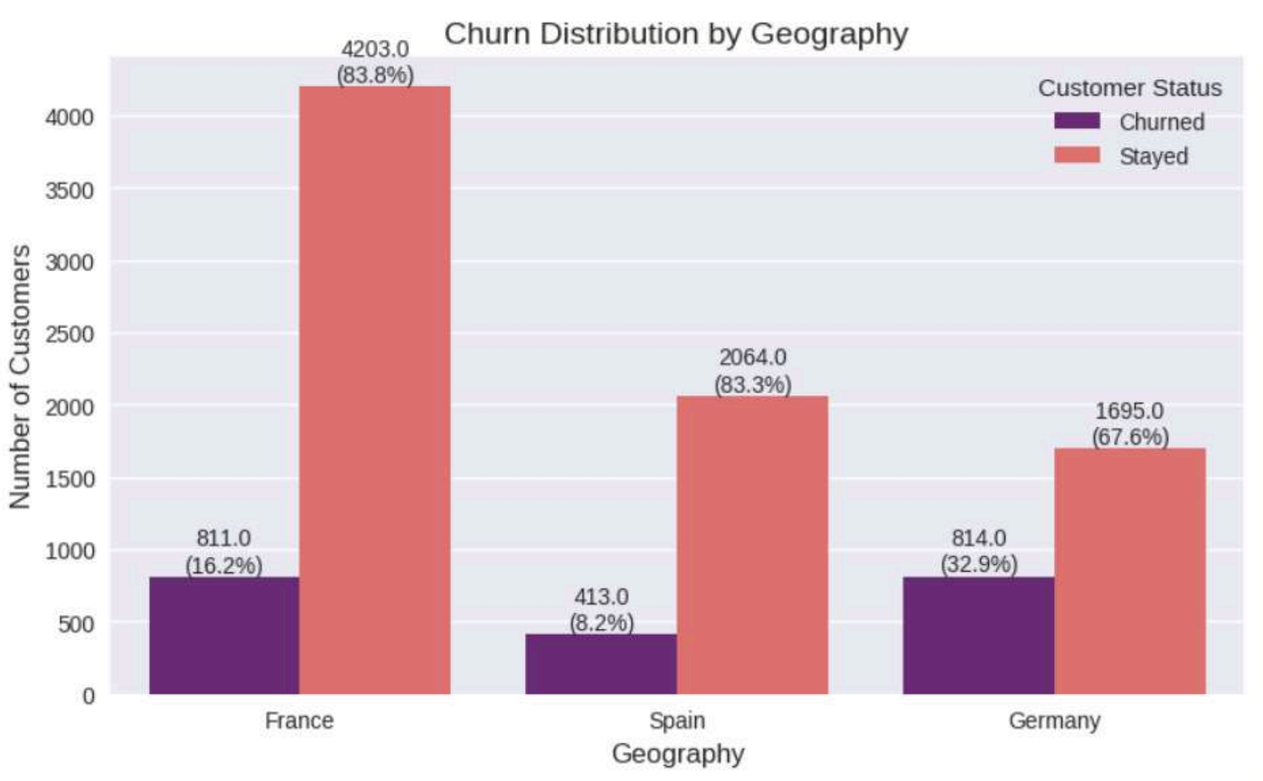
Gender (Male: 5457, Female: 4543)

- Males (54.6%) slightly outnumber females (45.4%), indicating a near-balanced gender split.
- **This minor imbalance could hint at gender-specific churn patterns worth exploring further.**



Value	Count	Percentage
Male	5457	54.6
Female	4543	45.4

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

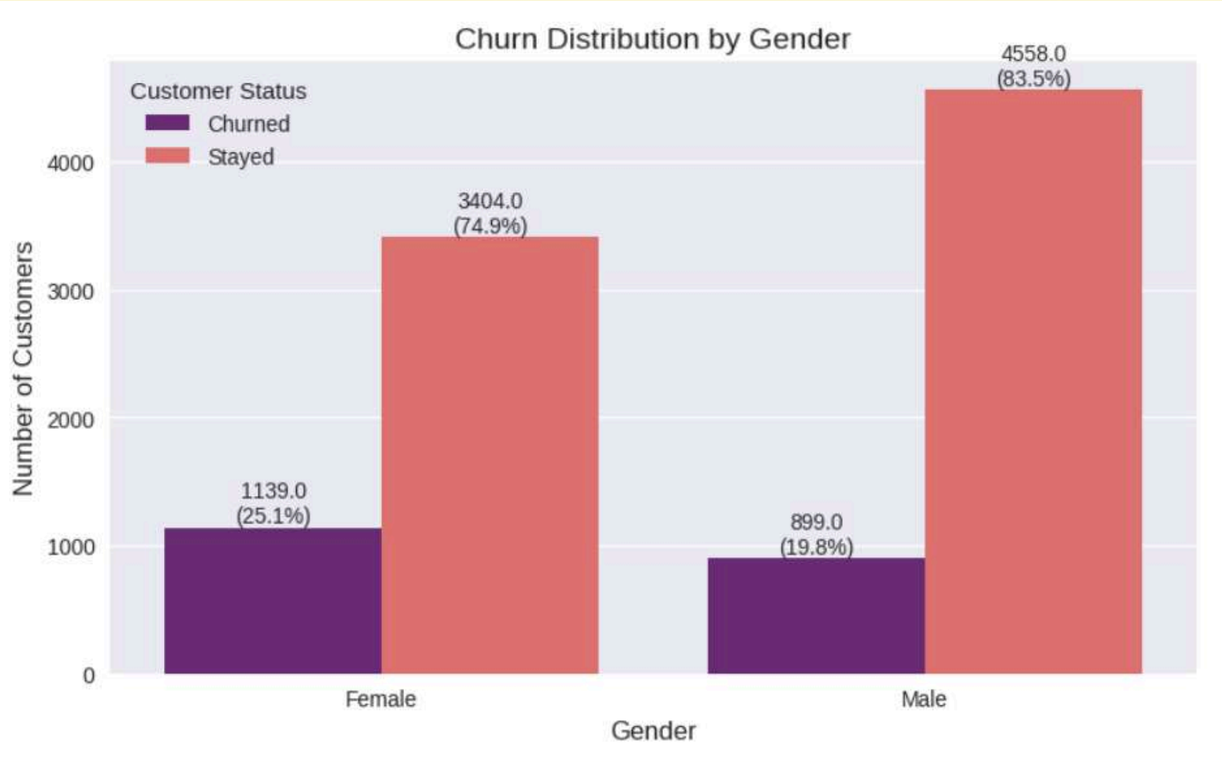


Based on the previous pie chart regarding geography, it is evident that France has the biggest customer base.

Further analysis, comparing the churn rate, shows that 814 (32.9%) customers from Germany are more likely to churn, compared to only 811 (16.2%) from France. In fact, the most loyal (remaining) customers are from France, with 4,203 (83.8%).

These suggest that the previous hypothesis that France has a higher churn rate is not entirely correct.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

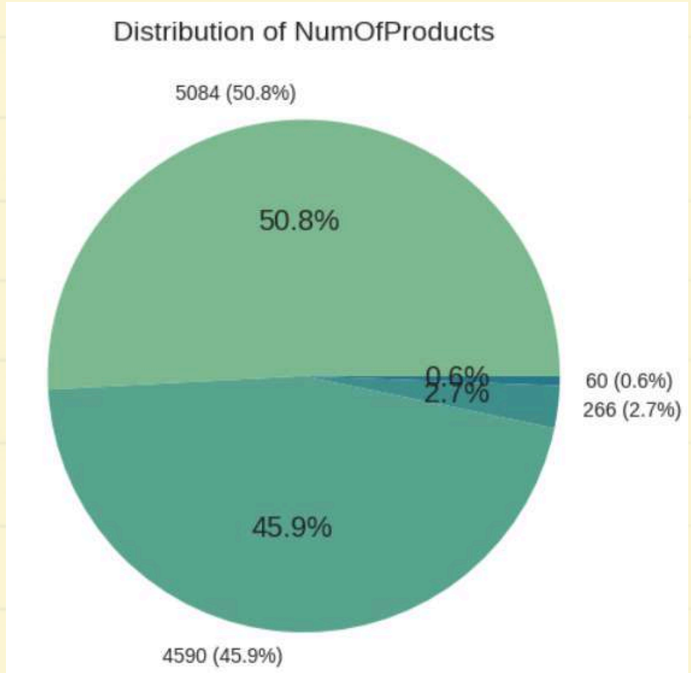


Analyzing the churn rate and the number of customers by gender shows that female are more likely to churn than male.

It can be seen in the churn rate of 1,139 (25.1%) among female customers compared to 899 (19.8%) among male customers, out of the total number of customers per gender. **The data confirms our suspicion that gender plays a role in churn rates.**

However, other external factors, such as credit score, card type used, or other supporting data, need to be included as a basis for analyzing trends using gender.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS



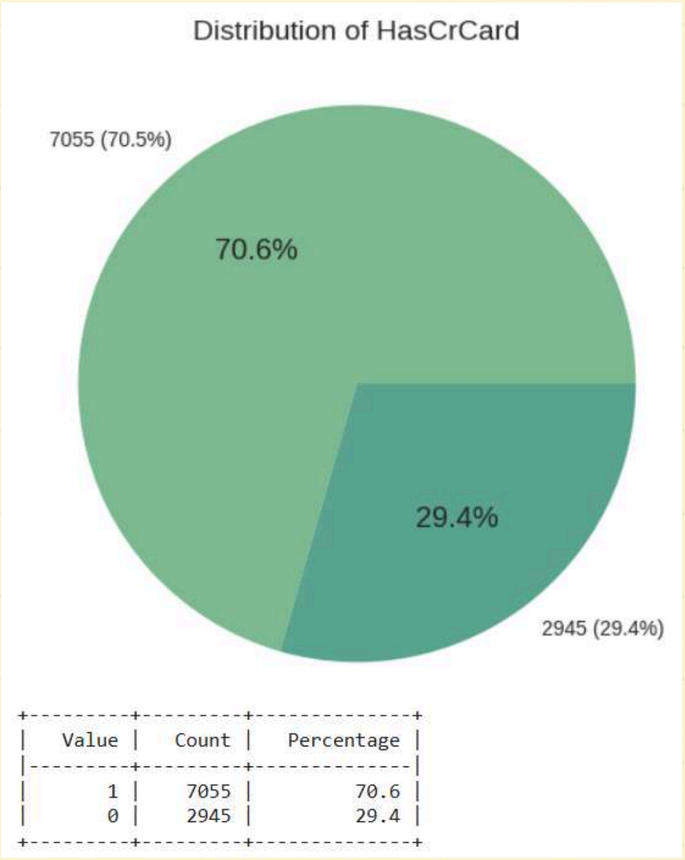
Value	Count	Percentage
1	5084	50.8
2	4590	45.9
3	266	2.7
4	60	0.6

NumOfProducts (1: 5084, 2: 4590, 3: 266, 4: 60)

- Most customers (50.8%) use one product, followed by 45.9% with two, and a small fraction with three (2.7%) or four (0.6%).
- **The dominance of one or two products suggests limited product engagement**, which may correspond with higher churn risk.

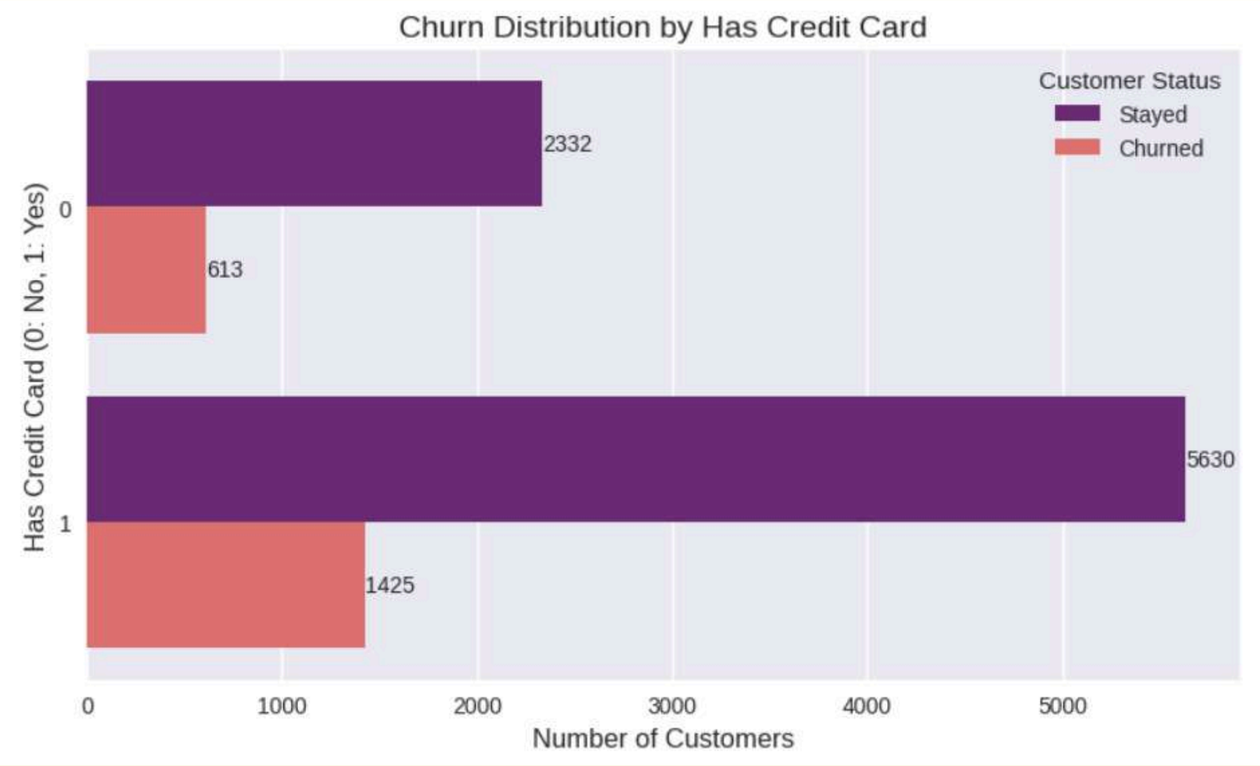
HasCrCard (1: 7055, 0: 2945)

- A majority (70.6%) have a credit card, while 29.4% do not.
- There is a strong indication that credit card ownership is associated with customer loyalty. The data shows a notable split, **with non-credit card holders potentially more likely to leave than those who own a card.**



Value	Count	Percentage
1	7055	70.6
0	2945	29.4

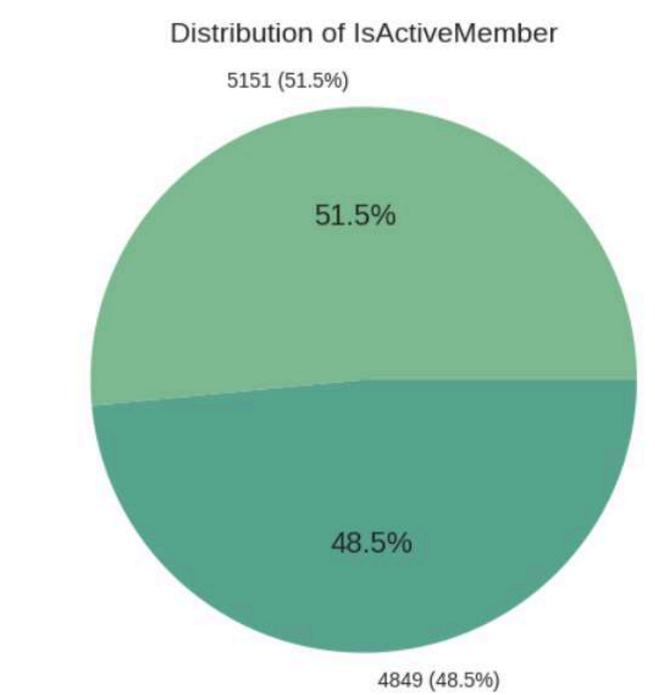
EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS



Based on the previous pie chart analysis, the data indicated a higher churn rate among customers who did not possess a credit card.

When represented in a bar chart, this hypothesis is correct. A total of 1,425 customers without credit cards chose to churn compared to those with credit cards (613). However, numerically, there are still more loyal customers who don't use credit cards than those who do.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS



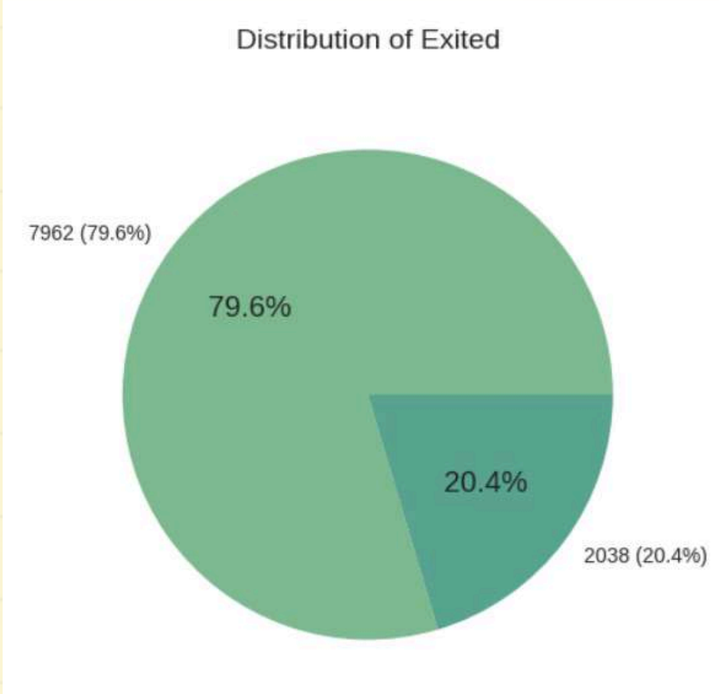
Value	Count	Percentage
1	5151	51.5
0	4849	48.5

IsActiveMember (1: 5151, 0: 4849)

- Active members (51.5%) are nearly equal to inactive ones (48.5%), suggesting a balanced engagement level.
- **Inactive customers might be at higher risk of churn due to disinterest.**

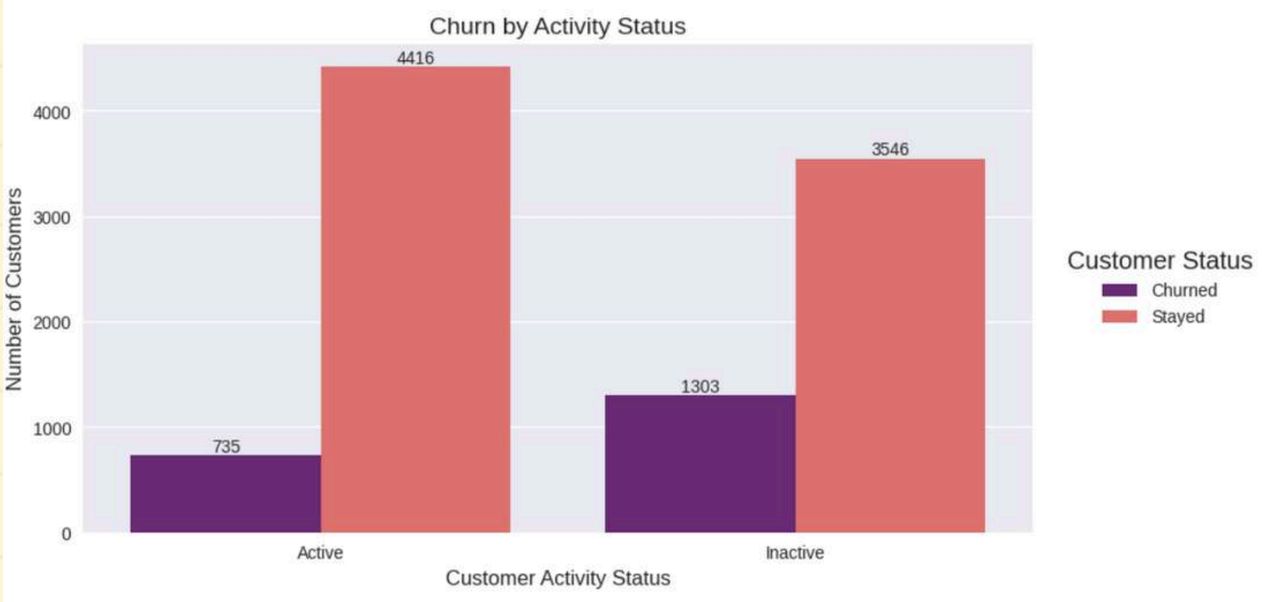
Exited (0: 7962, 1: 2038)

- Most customers (79.6%) stay, while 20.4% churn.
- This imbalance highlights a challenge for modeling, **as the minority class (churned) needs careful handling to avoid bias.**



Value	Count	Percentage
0	7962	79.6
1	2038	20.4

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

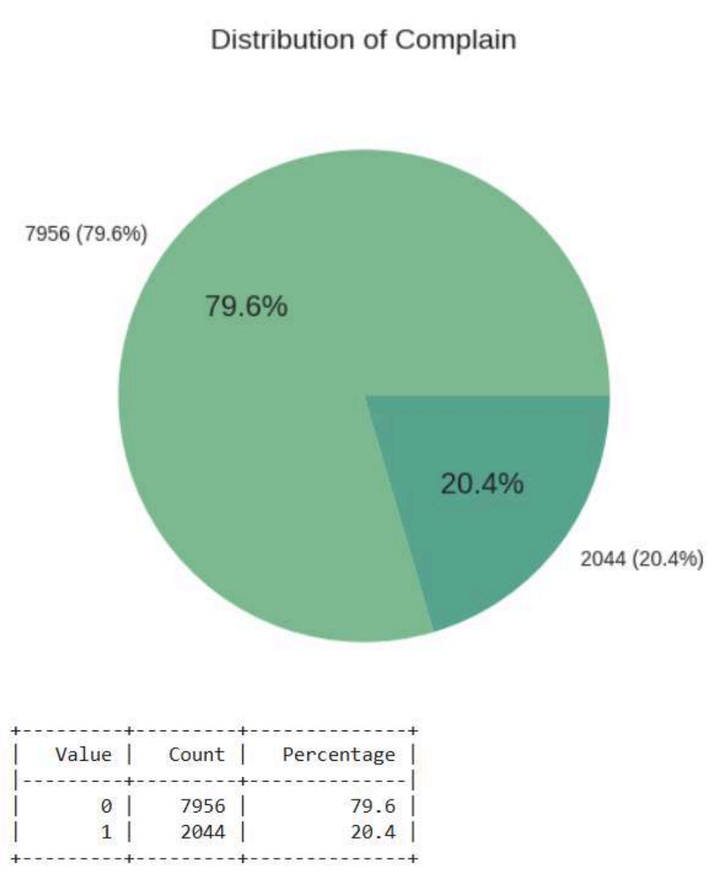


The previous pie chart analysis suggests that inactive customers are more likely to leave the bank.

When illustrated with a bar chart, this hypothesis is correct. A total of 1,303 inactive customers chose to churn compared to 735 active customers.

However, for further analysis, other variables or data could be used as external factors to strengthen the reasons why many customers, especially inactive ones, chose to leave the bank.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS



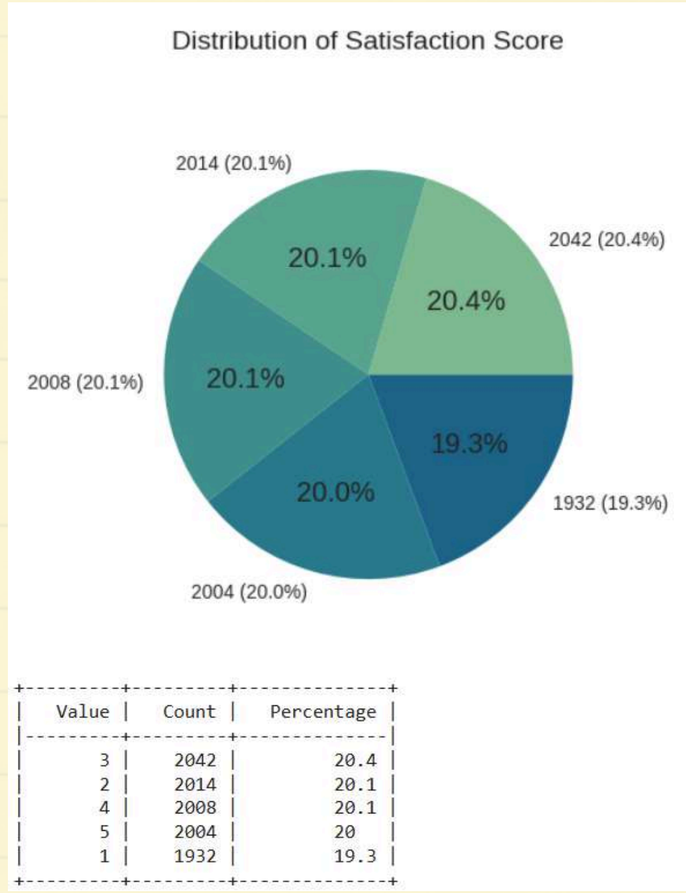
Value	Count	Percentage
0	7956	79.6
1	2044	20.4

Complain (0: 7956, 1: 2044).

- Similar to Exited, 79.6% have no complaints, and 20.4% do.
- **This alignment suggests complaints are a strong predictor of churn**, warranting further investigation.

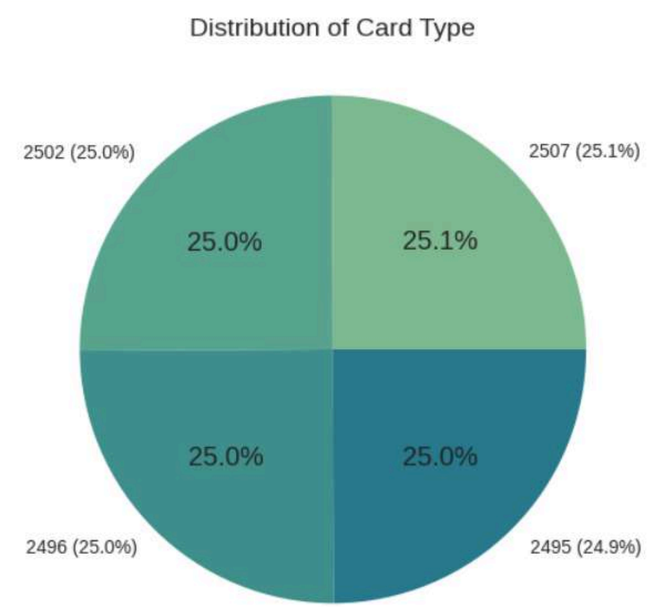
Satisfaction Score (3: 2042, 2: 2014, 4: 2008, 5: 1932)

- Scores are evenly distributed across 2 (20.1%), 3 (20.4%), 4 (20.1%), and 5 (19.3%), indicating consistent feedback with no dominant satisfaction level.
- **This uniformity might dilute its predictive power unless combined with other features.**



Value	Count	Percentage
3	2042	20.4
2	2014	20.1
4	2008	20.1
5	2004	20
1	1932	19.3

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS



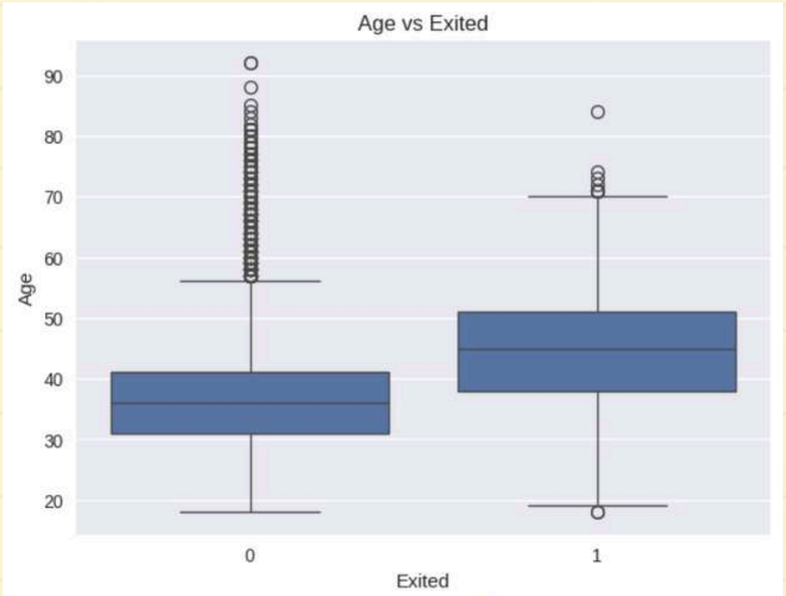
Value	Count	Percentage
DIAMOND	2507	25.1
GOLD	2502	25
SILVER	2496	25
PLATINUM	2495	25

Card Type (DIAMOND: 2507, GOLD: 2502, SILVER: 2496, PLATINUM: 2495)

- Card types are nearly equal (25.1% to 25.0%), reflecting a balanced product offering.
- **This even distribution suggests no single card type drives churn independently.**

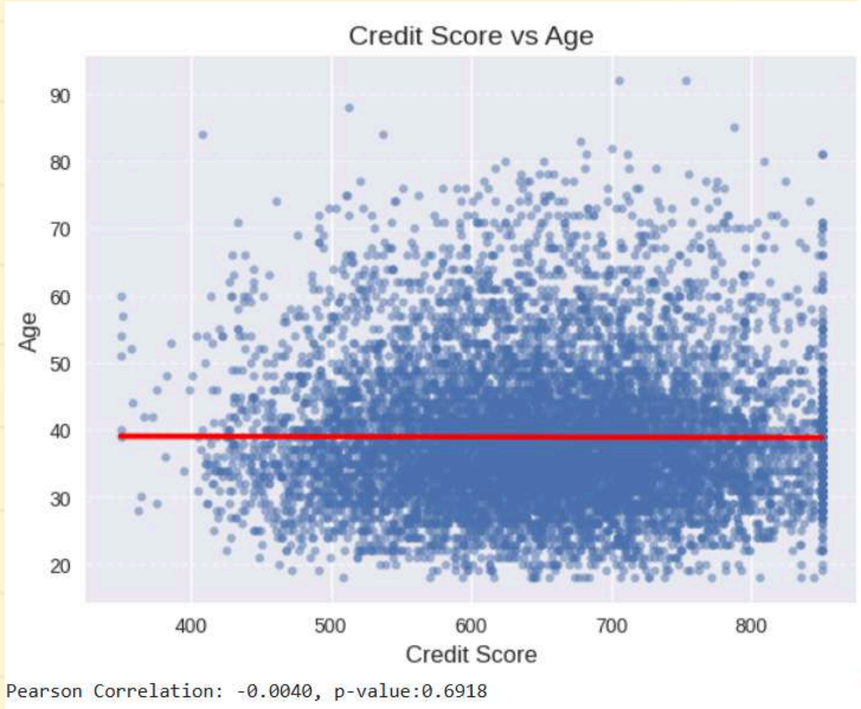
Based on the box plot between age and exit, it shows.

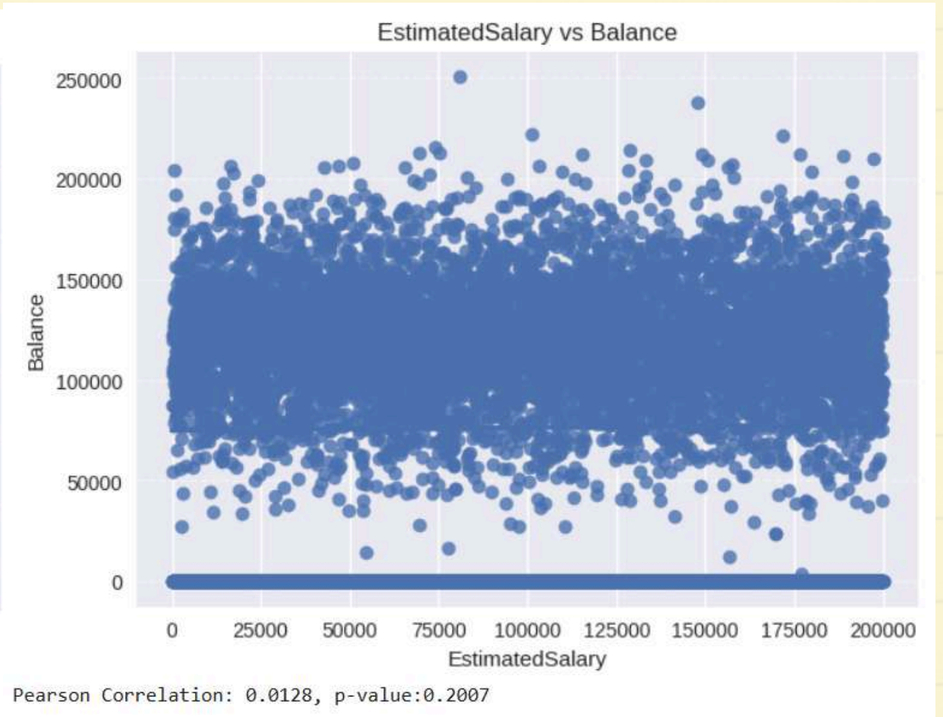
- Customers who left the bank have a notably higher median age than those who stayed. The Mann-Whitney U test result (p-value = 0.000) confirms this observation, providing strong evidence that older age is associated with a higher churn exposure.
- **The result supports the previous univariate analysis of age, which identified a right-skewed distribution and outliers at the older end of the age range.**



Mann-Whitney U Statistic:4347741.0000, p-value:0.0000

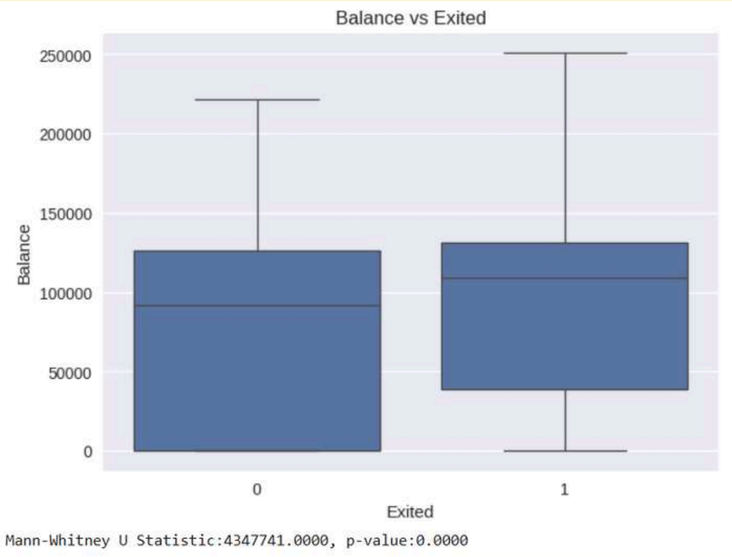
EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS





Based on the two analyses above, there is no correlation between credit score and age, estimated salary, and balance.
The results are not statistically significant, as shown by the p-values (0.6918 and 0.2007), which exceed the 0.05 threshold.

EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

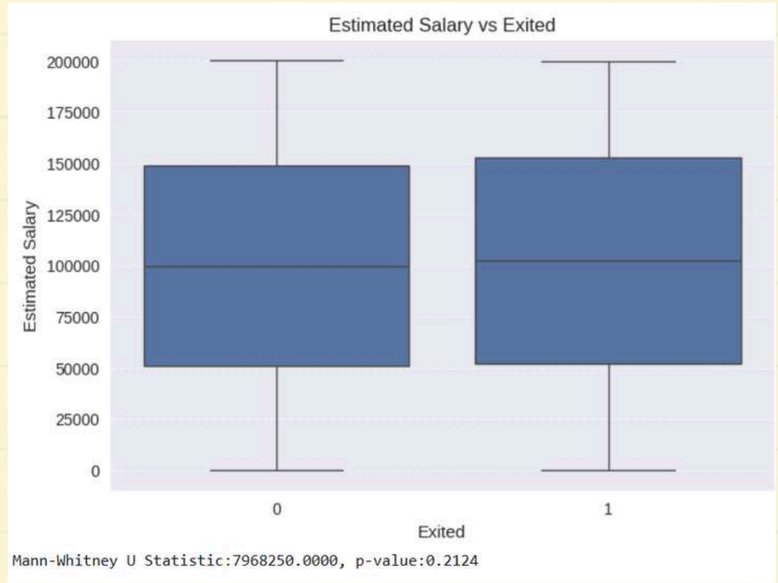


The figure reveals that churned customers hold higher balances than their non-churned counterparts.

- Customers who exited the bank had a significantly higher median balance (~90,000) than those who remained (~70,000), which suggests that a higher balance may increase the risk of churn.
- The Mann-Whitney U test, with a p-value of 0.000, validated the finding, confirming its statistical significance. The outcome could refer to a customer's dissatisfaction or their financial habits.

The figure reveals that both stayed and left customers hold the same estimated salary.

- The data shows that the median estimated salary is essentially the same for both churned and non-churned customers (~100,000), suggesting that salary does not play a role in customer retention.
- The Mann-Whitney U test is non-significant (p-value > 0.05, e.g., 0.2124), suggesting no statistically significant difference in EstimatedSalary between churned and non-churned customers.



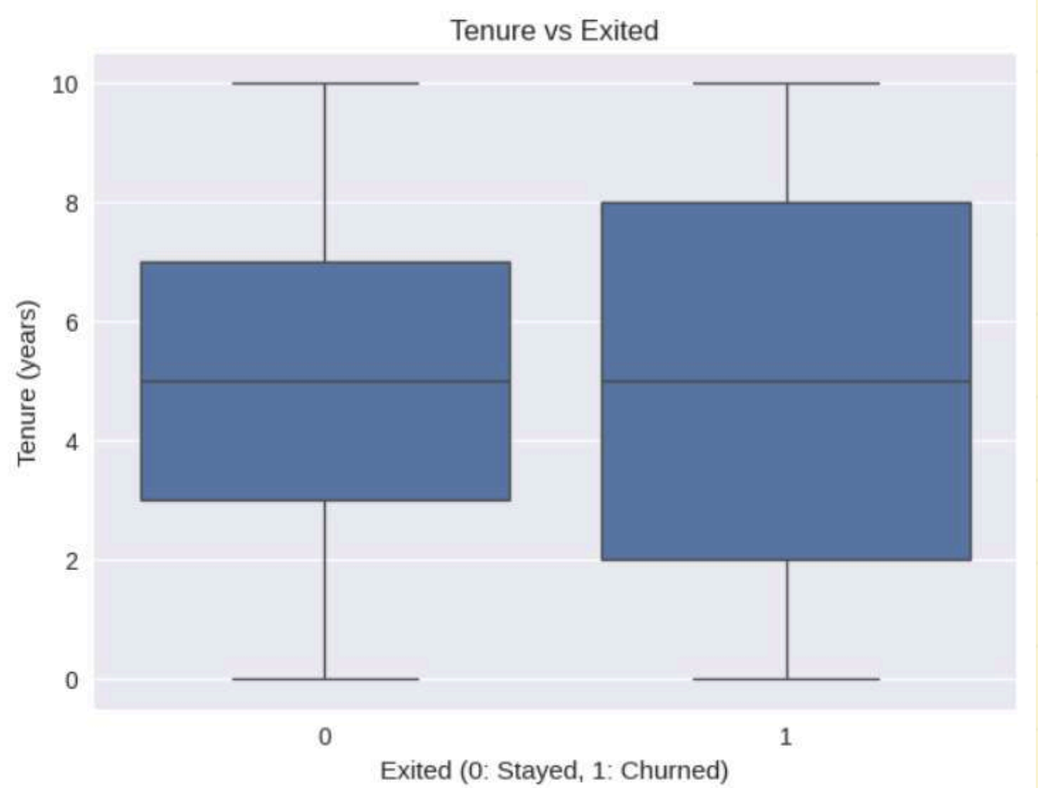
EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS

The figure reveals that both stayed and left customers hold the same tenure period.

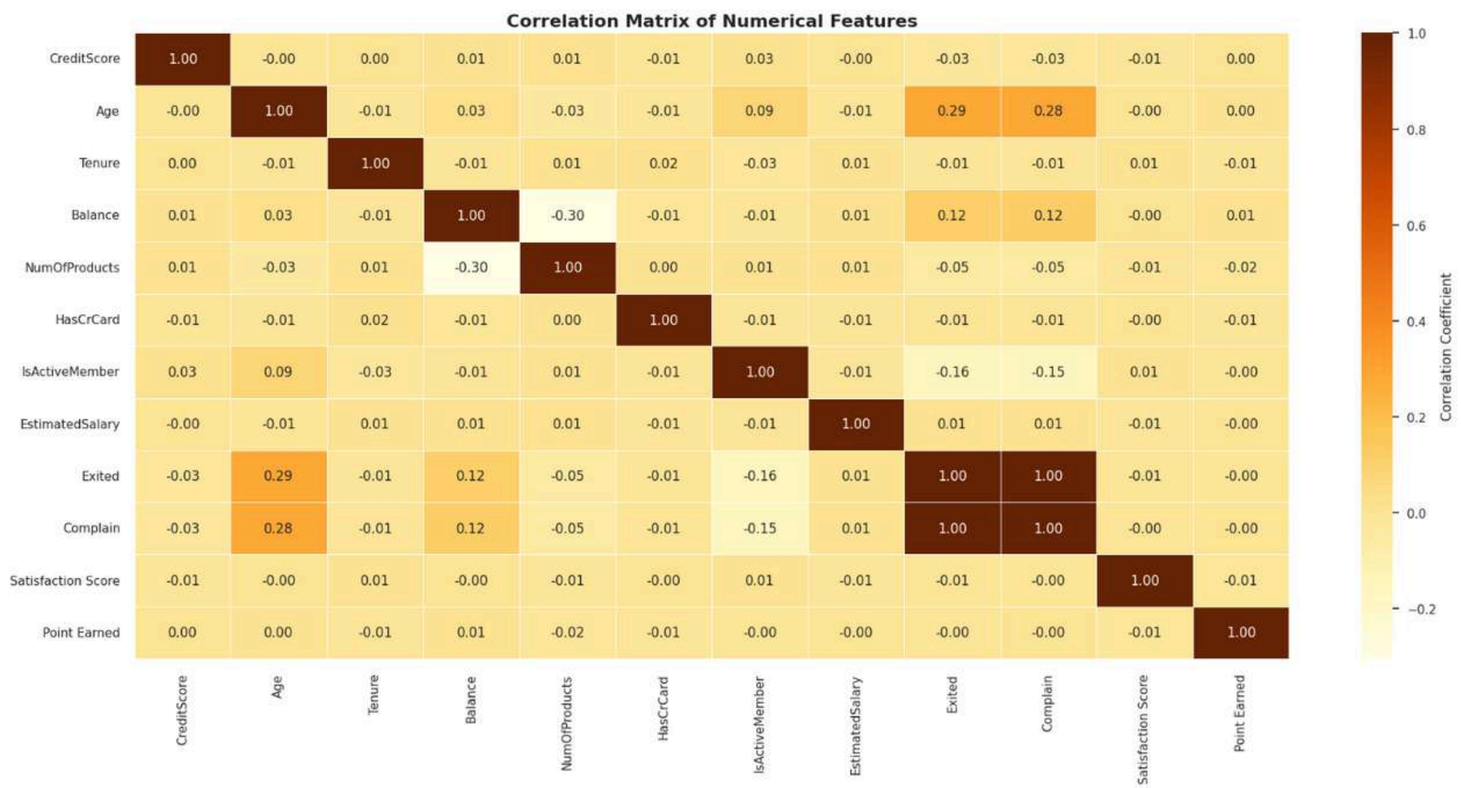
- The data appears that the median estimated salary is essentially the same for both churned and non-churned customers (~5), suggesting that tenure does not play a role in customer retention.
- The Mann-Whitney U test is non-significant (p-value > 0.05, e.g., 0.1730), suggesting no statistically significant difference in Tenure between churned and non-churned customers.

Based on our previous analysis, we preprocessed the dataset to make it suitable for machine learning models.

- Log transform Age, log1p for Balance.
- One-hot encode Geography, Gender, and Card Type.
- Address the existing imbalance with SMOTE or class weighting.
- Scale numerical features using StandardScaler.

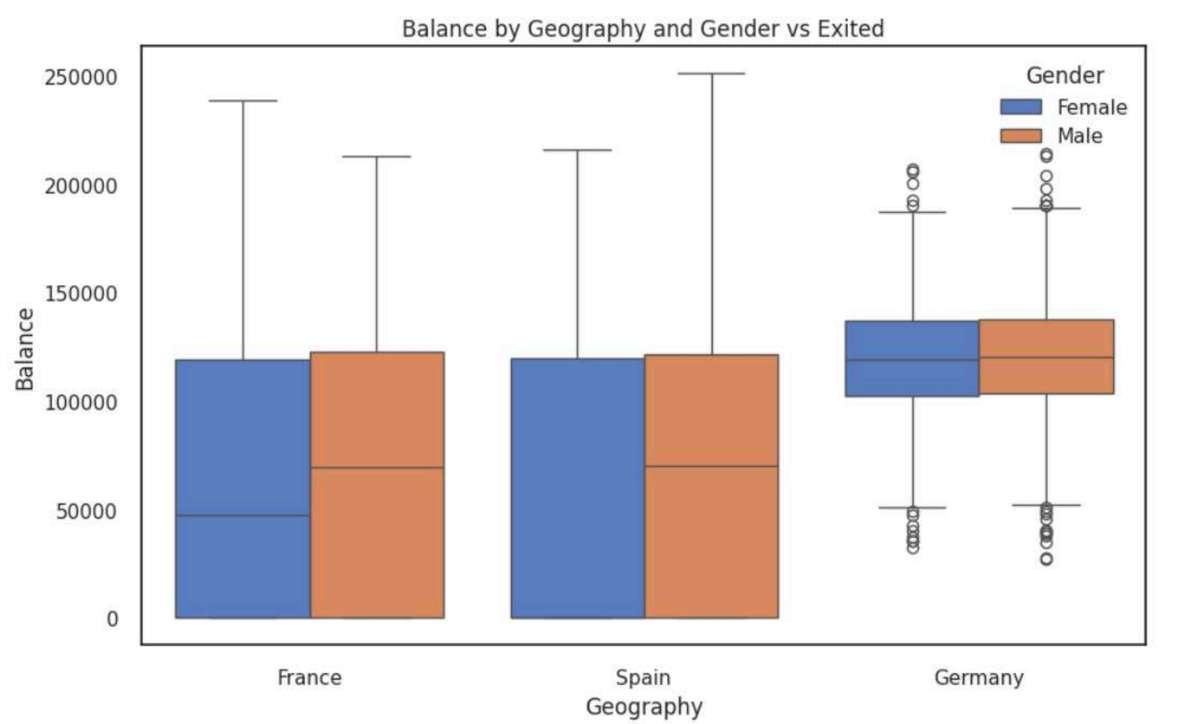


EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS



A heatmap of a correlation matrix provides a visual summary of the linear relationships between every pair of numerical variables. **Based on the result, we spotted two variables that have a stronger positive correlation (1.000). Specifically, there are Exited and Complain.**

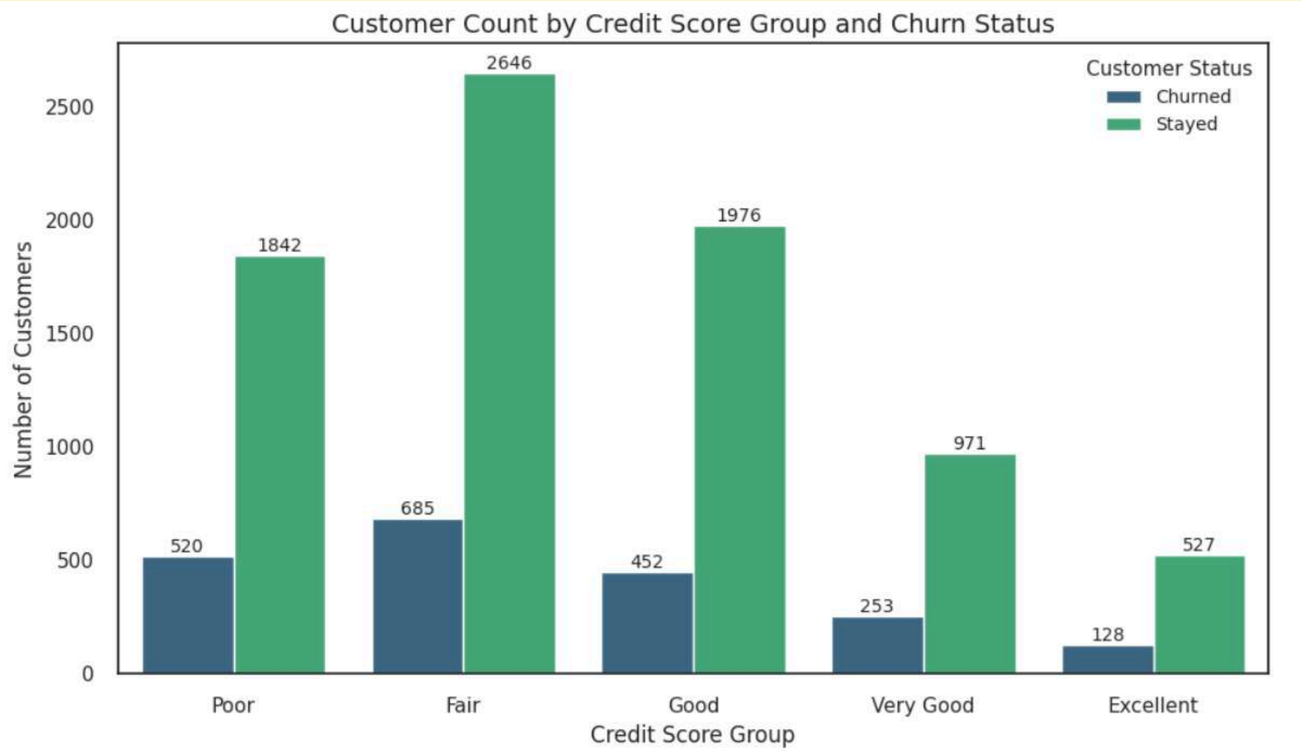
EDA (2) : UNIVARIATE, BIVARIATE & MULTIVARIATE ANALYSIS



Mann-Whitney U Statistic:8271000.0000, p-value:0.1730

The analysis found no significant difference in the median balance between genders within each region (France, Spain, and Germany). **Consequently, it's not possible to conclude that a customer's gender or their balance size is a dominant factor in predicting churn.** It is also confirmed by the p-value of the Mann-Whitney U statistic of 0.1730, indicating no statistically significant effect.

EDA (3) : OTHER DATA VISUALIZATION



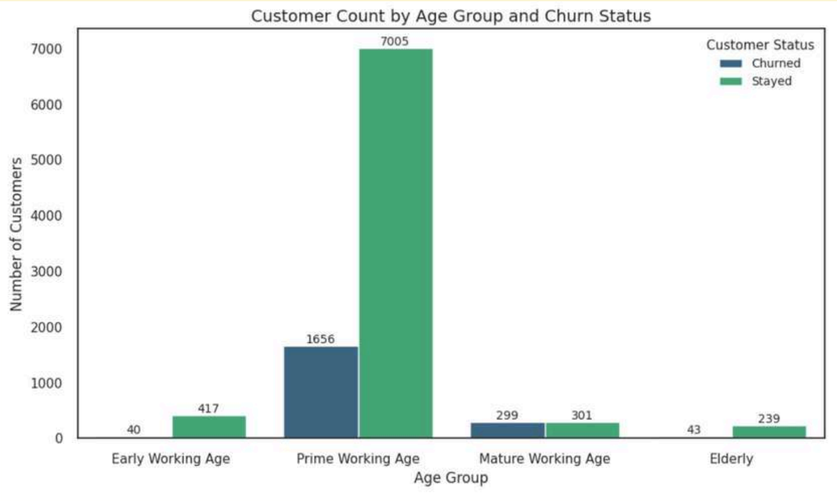
Shifting our focus to another visualization, the graph below will help us analyze whether a customer's credit score is related to their probability of churning.

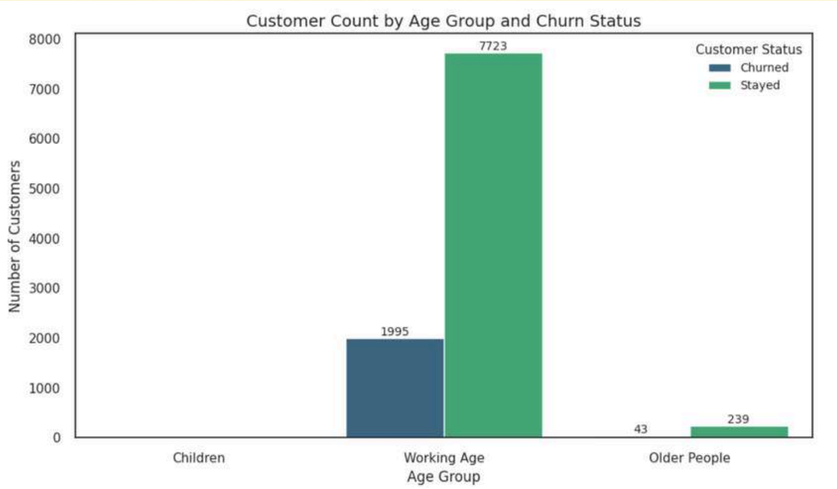
We categorized the credit scores using the FICO scoring model, which segments them into different ranges.

- Exceptional/Excellent: 800 to 850
- Very good: 740 to 799
- Good: 670 to 739
- Fair: 580 to 669
- Poor: 300 to 579

The data revealed the highest churn rate in customers with a fair credit score (685). Later, the highest rates were among customers with a poor credit score (520) and a good credit score (452).

EDA (3) : OTHER DATA VISUALIZATION





Next, we'll use a new visualization to analyze how age categories affect the likelihood of customer churn.
In this analysis, we used two age categories: the USA and Europe

Age Category	Detail Age Category
USA	• Early Working Age (15-24 years) • Prime Working Age (25-54 years) • Mature Working Age (55-64 years) • Elderly (64+ years)
Europe	• Children (0 - 14 years) • Working Age (15 - 64 years) • Older People (65+ years)

Although the majority of customers fall into the working-age category, they are also the most likely to churn.

DATA PROCESSING AND CLEANING

FUTURE SELECTION

Feature selection in machine learning is the process of selecting the most relevant and informative subset of features (input variables) from an existing dataset for training a model.

Why is Feature Selection Important?

- **Improves Model Accuracy:** By removing irrelevant or redundant features, you can reduce noise and improve the model's ability to generalize to new data.
- **Reduces Overfitting:** A model with too many features might "memorize" the training data instead of learning the underlying patterns. Selecting a smaller set of features of the model to focus on the most significant information, which leads to better performance on unseen data.
- **Speeds up Training:** Fewer features mean less data to process, which can significantly reduce the computational cost and training time of a model.
- **Increases Interpretability:** A model with fewer features is easier to understand and explain. It is necessary in fields like finance and healthcare, where model decisions must be transparent.

1. MUTUAL INFORMATION SCORES

Mutual information measures the dependency between two variables. In this case, it quantifies how much information each numerical feature provides about the 'Exited' variable (whether a customer churned or not).

Here's how to interpret the scores:

- **Higher Score:** A higher mutual information score indicates a stronger relationship between the feature and the target variable ('Exited'). It means the feature is more informative for predicting churn.
- **Lower Score:** A lower score suggests a weaker relationship. A score of 0 means the feature and the target are independent.

2. CHI-SQUARE TEST

The Chi-squared test for independence is used to determine if there is a statistically significant relationship between two categorical variables. In this case, we are testing if each categorical feature is independent of the 'Exited' variable (our target).

- **Chi-squared Statistic:** This value, derived from the data, quantifies the difference between the observed and expected frequencies, assuming the two variables are independent. A larger chi-squared statistic indicates a stronger relationship between variables.
- **P-value:** This is the probability of observing a Chi-squared statistic as extreme as, or more extreme than, the one calculated from your data, assuming that the two variables are actually independent (the null hypothesis).

DATA PROCESSING AND CLEANING

Mutual Information Scores:

	Mutual Info
Complain	0.493656
NumOfProducts	0.075658
Age	0.073814
Point Earned	0.010920
Balance	0.008278
Satisfaction Score	0.006186
Tenure	0.005234
IsActiveMember	0.003541
EstimatedSalary	0.002705
CreditScore	0.000395
HasCrCard	0.000000

dtype: float64

Based on the output of mutual information scores:

- Complain has the highest mutual information score (0.497657), indicating it's the most informative numerical feature for predicting churn among the ones considered.
- NumOfProducts and Age also have relatively higher scores (0.072791 and 0.072575, respectively), suggesting they are also somewhat informative.
- Features like HasCrCard, Satisfaction Score, and Point Earned have scores of 0, implying they have very little to no linear relationship with the 'Exited' variable based on this calculation.

Chi-squared Test Results for Categorical Features:

	Chi-squared Statistic	p-value
NumOfProducts	1501.504831	0.000000e+00
Complain	9907.907036	0.000000e+00
Geography	300.626401	5.245736e-66
IsActiveMember	243.694802	6.153167e-55
Gender	112.396554	2.925368e-26
Card Type	5.053223	1.679411e-01
Satisfaction Score	3.802704	4.333650e-01
HasCrCard	0.449404	5.026182e-01

But, if we look at the output of the chi-squared test: Complain and NumOfProducts are the most predictive variables for churn, as indicated by their statistically significant p-values of 0.0000. The other features are less useful for prediction.

DATA MODELLING

Data Split

Data Modelling

Backward Elimination

Machine Learning

Logistic, DT, RF, GB, DNN

DATA MODELLING: DATA SPLIT

First, we perform the data split process. We import several functions, namely:

- **StandardScaler:** This function standardizes the data by transforming it so that it has a mean of zero (0) and a standard deviation of one (1). The core goal is to standardize the scales of numerical variables to prevent one from disproportionately affecting the model.
- **OneHotEncoder:** This function transforms nominal categorical variables (unordered data) into binary vectors suitable for machine learning models.
- **FunctionTransformer:** This function applies a custom transformation to the data, and for now, we're using log.

Second, we create a separation between the independent variables/features (X) and the dependent/target variable (Y), namely 'Exited'. The next step was to determine the data types, identifying whether each variable was categorical or numerical.

Third, we preprocessed the variables as discussed previously, such as:

- Standardize numerical variables using StandardScaler
- Categorize variables using OneHotEncoder
- Log transform the Age & Balance features

Lastly, we divided the data into training and testing sets, with 20% of the data allocated for testing, as seen in the figure.

```
# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

print("Data splitting and preprocessing setup complete.")
print(f"Shape of X_train: {X_train.shape}")
print(f"Shape of X_test: {X_test.shape}")
print(f"Shape of y_train: {y_train.shape}")
print(f"Shape of y_test: {y_test.shape}")
```

```
Data splitting and preprocessing setup complete.
Shape of X_train: (8000, 13)
Shape of X_test: (2000, 13)
Shape of y_train: (8000,)
Shape of y_test: (2000,)
```

DATA MODELLING: BACKWARD ELIMINATION

Backward elimination is a method for selecting features in multiple linear regression (MLR). It works by removing the least significant variables one at a time, which helps improve the model's accuracy and makes the results easier to interpret.

We performed backward elimination using **logistic regression**, a supervised machine learning algorithm that predicts the probability of a categorical outcome. This method was a fitting choice since our target variable, Exited, is binary (0 or 1), making it a binomial classification problem.

Feature	Backward Elimination
Approach	Starts with all variables and removes the least sigificant one iteratively
Initial Model	Begins with all independent variables
Process	Removes variables one by one based on the highest p-value
Stopping Criterion	Stops when all remaining variables have p-value < 0.05
When to Use?	When we want to simplify a full model with all features
Accuracy	Often leads to a more accurate and simplified model

DATA MODELLING: BACKWARD ELIMINATION

```
Optimization terminated successfully.
         Current function value: 0.422867
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                 Exited   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9983
Method:                           MLE   Df Model:                           16
Date:                Sun, 31 Aug 2025   Pseudo R-squ.:                  0.1637
Time:                        05:33:14   Log-Likelihood:                -4228.7
converged:                       True   LL-Null:                       -5056.3
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                -13.1704      0.508    -25.934      0.000     -14.166     -12.175
CreditScore           -0.0006      0.000     -2.286      0.022      -0.001   -9.19e-05
Age                    3.4693      0.120     28.989      0.000       3.235       3.704
Tenure                -0.0153      0.009     -1.627      0.104      -0.034       0.003
Balance                0.0297      0.006      4.885      0.000       0.018       0.042
NumOfProducts         -0.0925      0.048     -1.944      0.052      -0.186       0.001
HasCrCard             -0.0415      0.060     -0.696      0.486      -0.158       0.075
IsActiveMember        -1.0608      0.058    -18.434      0.000      -1.174      -0.948
EstimatedSalary     5.006e-07   4.76e-07      1.051      0.293   -4.33e-07    1.43e-06
Satisfaction Score    -0.0104      0.019     -0.539      0.590      -0.048       0.028
Point Earned          -0.0001      0.000     -1.077      0.281      -0.000       0.000
Geography_Germany      0.7545      0.070     10.825      0.000       0.618       0.891
Geography_Spain        0.0253      0.071      0.357      0.721      -0.114       0.164
Gender_Male           -0.5233      0.055     -9.557      0.000      -0.631      -0.416
Card Type_GOLD        -0.1387      0.077     -1.799      0.072      -0.290       0.012
Card Type_PLATINUM    -0.0767      0.076     -1.004      0.316      -0.226       0.073
Card Type_SILVER      -0.0472      0.076     -0.617      0.537      -0.197       0.103
==============================================================================
```

Let's interpret the results from the statsmodels Logistic Regression summary table from the previous picture:

- The model used is Logistic Regression (**Model: Logit**) with method MLE or Maximum Likelihood Estimation (**Method: MLE**) of 10000 data points (**No.Observation: 10000**) by Exited as target (**Dep.Variable: Exited**)
- According to the results, the model's features explain approximately 16.4% of the variance in the Exited variable, as indicated by the **Pseudo R-squared value of 0.1637**. This was achieved after **seven iterations**. It is a relatively low value, which is common in churn prediction models, but it indicates the features do have some predictive power.
- Take a look at the p-value (**P>|z|**); just a few features have a statistically significant result with 'Exited'
  - **The features CreditScore, Age, Balance, IsActiveMember, Geography_Germany, and Gender_Male have been detected as having a significant relationship with customer churn.**
  - The other Features with high p-values are generally considered not statistically significant at the 0.05 level.

DATA MODELLING: MACHINE LEARNING

We are nearing the end of our process, and the final step is data modeling.

In here, we used some methods of machine learning to estimate the model, such as **Logistic Regression, Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Dense Neural Network (DNN)**

Before training the machine learning model, we first recognized that the dataset was imbalanced. To resolve this, we employed **the Synthetic Minority Oversampling Technique (SMOTE)**, which creates new, synthetic data points for the underrepresented class. The technique creates synthetic data by connecting a minority class instance with a similar instance and generating a new point along that line.

Feature	Logistic Regression	Decision Tree
Type	Linear Model, Supervised Learning	Non-linear Model, Supervised Learning
Method	Uses a sigmoid function to predict the probability of an event.	Creates a flowchart-like structure by splitting data based on feature values.
Intepretation	**High** 🟢. Coefficients show how each feature affects the log-odds of the outcome.	**High** 🟢. Easy to visualize and understand the decision rules.
Performance	**Good** 🟡. Works well for simple, linearly separable data. Can struggle with complex relationships.	**Moderate** 🟡. Prone to overfitting on complex data.
Speed	**Fast** 🟢. Computationally efficient to train.	**Fast** 🟢. Quick to train on small to medium datasets.
Key Limitation	Assumes a linear relationship between features and the log-odds of the outcome.	Prone to overfitting and instability (a small change in data can lead to a very different tree).

DATA MODELLING: MACHINE LEARNING

Feature	Random Forest	Gradient Boosting	Dense Neural Network
Type	Ensemble Method (Bagging), Supervised Learning	Ensemble Method (Boosting), Supervised Learning	Deep Learning, Supervised Learning
Method	Builds multiple, independent decision trees and averages their outputs (for regression) or uses a majority vote (for classification).	Sequentially builds decision trees, with each new tree correcting the errors of the previous ones.	Connects layers of neurons, where each neuron in a layer is connected to all neurons in the next. Learns complex patterns through backpropagation
Intepretation	**Low** 🔴. The "black box" nature of combining many trees makes it hard to interpret.	**Low** 🔴. It's a "black box" as it's a complex combination of many weak models.	**Low** 🔴. The complex, layered structure makes it difficult to understand how predictions are made.
Performance	**High** 🟢. Reduces variance and overfitting, leading to high accuracy.	**Very High** 🟢🟢. Often the top-performing algorithm for structured data due to its focus on correcting errors.	**Very High** 🟢🟢. Can capture highly complex, non-linear patterns in large datasets.
Speed	**Moderate** 🟡. Can be slow because it trains many trees in parallel.	**Slow** 🔴. Slower than Random Forest because it trains trees sequentially.	**Slow** 🔴. Training can be very time-consuming, especially with many layers and large datasets.
Key Limitation	Requires more memory and computational resources than a single decision tree.	Highly susceptible to overfitting if not properly regularized.	Requires a large amount of data and significant computational power. Can be a "black box."

DATA MODELLING: LOGISTIC REGRESSION WITH SMOTE

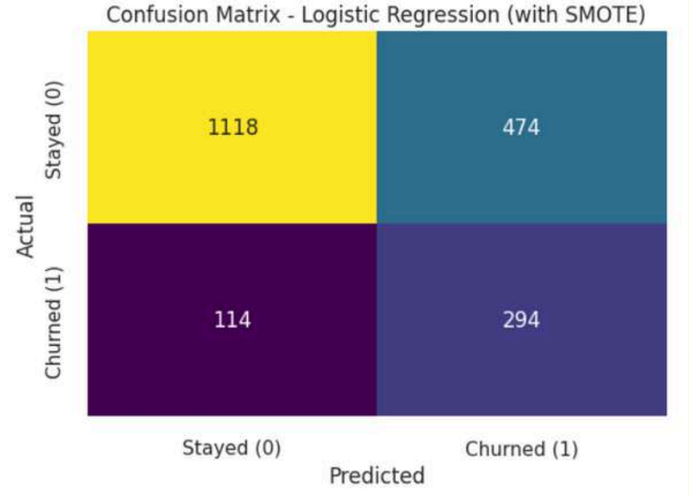
```
--- Training Logistic Regression with SMOTE ---
ROC AUC: 0.7819

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.70      0.79      1592
           1       0.38      0.72      0.50       408

    accuracy                           0.71      2000
   macro avg       0.65      0.71      0.65      2000
weighted avg       0.80      0.71      0.73      2000


Confusion Matrix:
[[1118  474]
 [ 114  294]]
```



First, the **ROC AUC value is 0.7819**, indicating a 78.19% probability that the model will rank a randomly selected churned sample higher than a non-churned sample.

Second, we move on to **the classification report**. For 0 (stayed) and 1 (churn)

0 (stayed)

- **Precision 0.91** indicates that of all the predictions made by the stayed model, only 91% were correct.
- **Recall 0.70** indicates that 70% of all customers did not churn.
- **F1-Score 0.79** is the harmonic mean of precision and recall.
- **Support 1592** indicates the number of customers who stayed.

1 (churn)

- **Precision 0.38** indicates that of all the predictions made by the churn model, only 38% were correct.
- **Recall 0.72** indicates that 72% of all customers churn.
- **F1-Score 0.50** indicates a balance between precision and recall.
- **Support 408** indicates the number of customers who churned.

DATA MODELLING: LOGISTIC REGRESSION WITH SMOTE

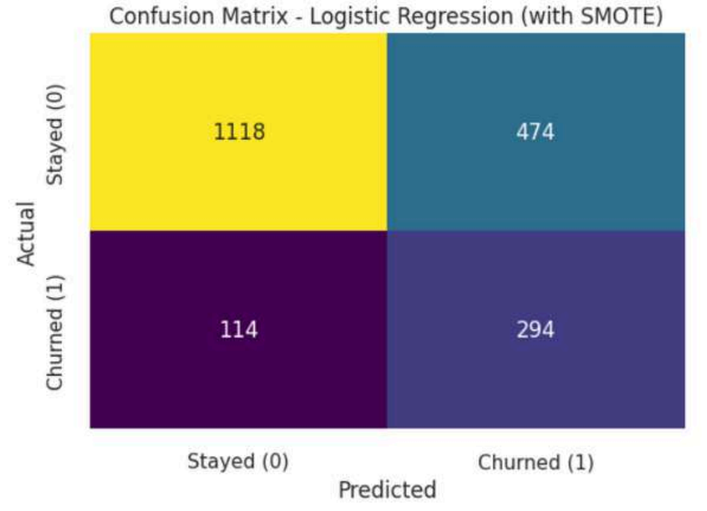
```
--- Training Logistic Regression with SMOTE ---
ROC AUC: 0.7819

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.70      0.79      1592
           1       0.38      0.72      0.50       408

    accuracy                           0.71      2000
   macro avg       0.65      0.71      0.65      2000
weighted avg       0.80      0.71      0.73      2000


Confusion Matrix:
[[1118  474]
 [ 114  294]]
```



Third, the **confusion matrix**. First, we identified:

- **True Negative (TN) = 1118**, or the model correctly predicted that 1118 customers stayed.
- **False Positive (FP) = 474,** or the model incorrectly predicted 474 customers had churned, even though they actually stayed.
- **False Negative (FN) = 114**, or the model incorrectly predicted 114 customers stayed, even though they actually churned.
- **True Positive (TP) = 294**, or the model correctly predicted that 294 customers had churned.

Then, to **evaluate the model, notably since it uses SMOTE to handle data imbalance**,

- **Accuracy** = (294+1118)/(294+1118+474+114) = 1412/2000= **0.706**, or the model's accuracy is approximately 70.6%
- **Precision** = 294/(294+474) = 294/768≈ **0.383**, or of all the "churn" predictions made by the model, only 38.3% actually churned.
- **Recall** = 294/(294+114)=294/408=**0.720,** or the model successfully identified 72% of the total customers who actually churned.

Thus, we can conclude:

- **High recall** 🟢 indicates good results; the model is effective in identifying the majority of customers at risk of churning.
- **Low precision** 🔴, where the model often produces false positives, where customers predicted to churn actually do not. It can have a cost impact if interventions are subject to customers who have not actually churned.

DATA MODELLING: DECISION TREE WITH SMOTE

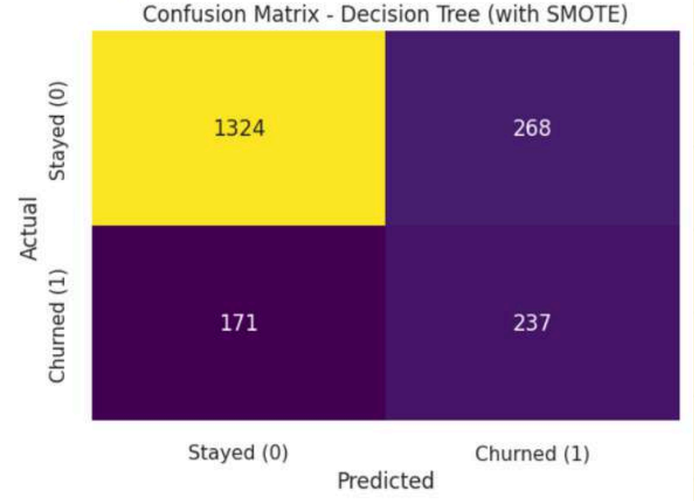
```
--- Training Decision Tree with SMOTE ---
ROC AUC: 0.7063

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.83      0.86      1592
           1       0.47      0.58      0.52       408

    accuracy                           0.78      2000
   macro avg       0.68      0.71      0.69      2000
weighted avg       0.80      0.78      0.79      2000


Confusion Matrix:
[[1324  268]
 [ 171  237]]
```



First, the **ROC AUC value is 0.7063**, indicating a 70.63% probability that the model will rank a randomly selected churned sample higher than a non-churned sample.

Second, we move on to **the classification report**. For 0 (stayed) and 1 (churn)

0 (stayed)

- **Precision 0.89** indicates that of all the predictions made by the stayed model, only 89% were correct.
- **Recall 0.83** indicates that 83% of all customers did not churn.
- **F1-Score 0.86** is the harmonic mean of precision and recall.
- **Support 1592** indicates the number of customers who stayed.

1 (churn)

- **Precision 0.47** indicates that of all the predictions made by the churn model, only 47% were correct.
- **Recall 0.58** indicates that 58% of all customers churn.
- **F1-Score 0.52** indicates a balance between precision and recall.
- **Support 408** indicates the number of customers who churned.

DATA MODELLING: DECISION TREE WITH SMOTE

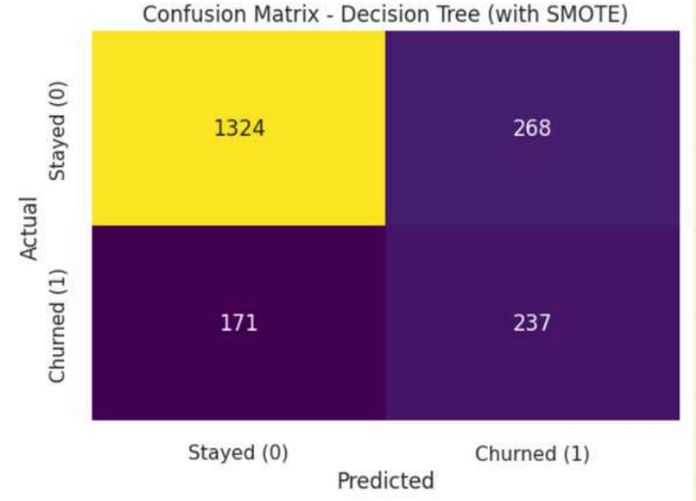
```
--- Training Decision Tree with SMOTE ---
ROC AUC: 0.7063

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.83      0.86      1592
           1       0.47      0.58      0.52       408

    accuracy                           0.78      2000
   macro avg       0.68      0.71      0.69      2000
weighted avg       0.80      0.78      0.79      2000


Confusion Matrix:
[[1324  268]
 [ 171  237]]
```



Based on the confusion **matrix for the Decision Tree model with SMOTE,** here is an interpretation of the results:

- **True Negatives (TN) = 1324:** The model correctly predicted that 1,324 customers would stay (not churn).
- **False Positives (FP) = 268**: The model incorrectly predicted that 268 customers would churn, when in fact they stayed. That is a Type I error or "false alarm."
- **False Negatives (FN) = 171**: The model incorrectly predicted that 171 customers would stay, when in fact they churned. It is a Type II error or "miss."
- T**rue Positives (TP) = 237**: The model correctly predicted that 237 customers would churn.

We can calculate **key performance metrics to better understand the model's effectiveness:**

- **Accuracy**: (237+1324)/(237+1324+268+171)=1561/2000=**0.7805**. The model's overall accuracy is 78.05%.
- **Precision** 237/(237+268)=237/505≈**0.469**. The precision is approximately 46.9%. This means that when the model predicts a customer will churn, it's only correct about half the time.
- **Recall**=237/(237+171)=237/408≈**0.581**. The recall is approximately 58.1%. This indicates that the model successfully identified more than half of the true churn cases.

The Decision Tree model with SMOTE has a reasonably high overall accuracy. However, a deeper look reveals some important details:

- **Moderate Recall 🟡:** The model has a moderate recall of 58.1%, which is good for catching potential churners.
- **Low Precision 🔴**: The precision of 46.9% is relatively low. It means the model makes a significant number of false positive predictions, incorrectly labeling many customers as churn risks.

DATA MODELLING: RANDOM FOREST WITH SMOTE

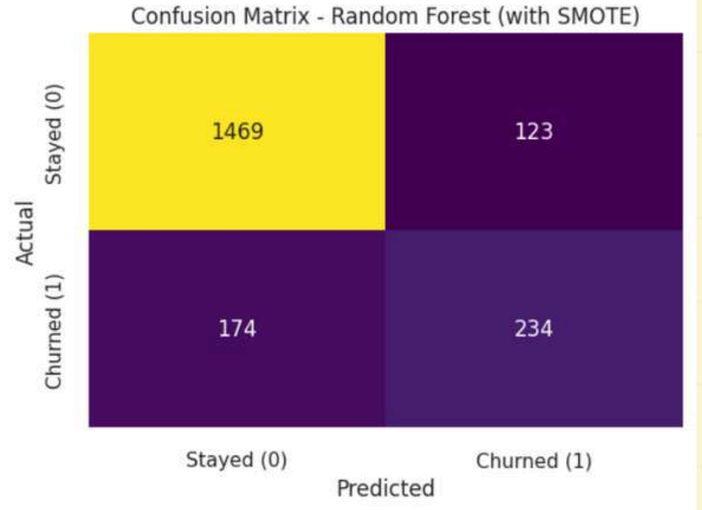
```
--- Training Random Forest with SMOTE ---
ROC AUC: 0.8591

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.91      1592
           1       0.66      0.57      0.61       408

    accuracy                           0.85      2000
   macro avg       0.77      0.75      0.76      2000
weighted avg       0.85      0.85      0.85      2000


Confusion Matrix:
[[1469  123]
 [ 174  234]]
```



First, the **ROC AUC value is 0.8591**, indicating a 85.91% probability that the model will rank a randomly selected churned sample higher than a non-churned sample.

Second, we move on to **the classification report**. For 0 (stayed) and 1 (churn)

0 (stayed)

- **Precision 0.89** indicates that of all the predictions made by the stayed model, only 89% were correct.
- **Recall 0.92** indicates that 92% of all customers did not churn.
- **F1-Score 0.91** is the harmonic mean of precision and recall.
- **Support 1592** indicates the number of customers who stayed.

1 (churn)

- **Precision 0.66** indicates that of all the predictions made by the churn model, only 66% were correct.
- **Recall 0.57** indicates that 57% of all customers churn.
- **F1-Score 0.61** indicates a balance between precision and recall.
- **Support 408** indicates the number of customers who churned.

DATA MODELLING: RANDOM FOREST WITH SMOTE

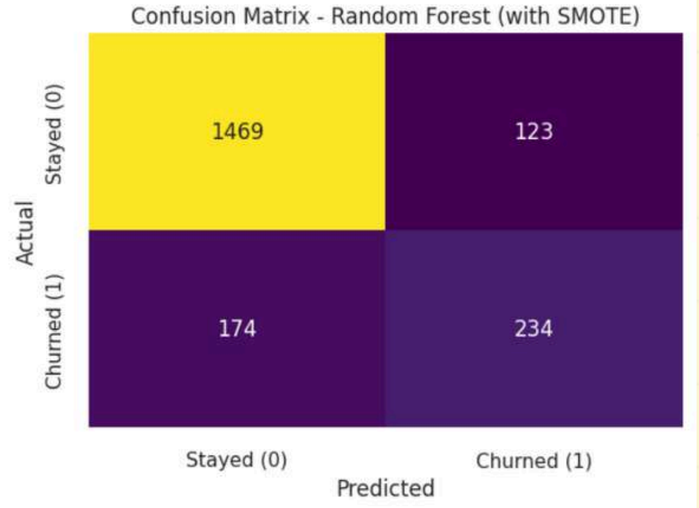
```
--- Training Random Forest with SMOTE ---
ROC AUC: 0.8591

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.91      1592
           1       0.66      0.57      0.61       408

    accuracy                           0.85      2000
   macro avg       0.77      0.75      0.76      2000
weighted avg       0.85      0.85      0.85      2000


Confusion Matrix:
[[1469  123]
 [ 174  234]]
```



Based on the confusion **matrix for the Random Forest model with SMOTE,** here is an interpretation of the results:

- **True Negatives (TN) = 1469:** The model correctly predicted that 1,469 customers would stay (not churn).
- **False Positives (FP) = 123**: The model incorrectly predicted that 123 customers would churn, when in fact they stayed. That is a Type I error or "false alarm."
- **False Negatives (FN) = 174**: The model incorrectly predicted that 174 customers would stay, when in fact they churned. It is a Type II error or "miss."
- T**rue Positives (TP) = 234**: The model correctly predicted that 234 customers would churn.

We can calculate **key performance metrics to better understand the model's effectiveness:**

- **Accuracy**: (234+1469)/(234+1469+123+174)=1703/2000=**0.8515**. The model's overall accuracy is 85.15%.
- **Precision** 234/(234+123)=234/357≈**0.655**. The precision is approximately 65.5%. This means that when the model predicts a customer will churn, it's only correct about half the time.
- **Recall**=234/(234+174)=234/408≈**0.574**. The recall is approximately 57.4%. This indicates that the model successfully identified more than half of the true churn cases.

The Random Forest model with SMOTE has a high overall accuracy, a significant improvement over the Decision Tree model. The most notable result is **the moderate precision 🟡 (65.5%)** compared to the other models. That means it is much better at avoiding false positives.

While **the recall 🟡 (57.4%)** is similar to the Decision Tree model, the increase in precision makes this model more reliable. It correctly identifies a solid portion of actual churners without generating as many "false alarms."

DATA MODELLING: GRADIENT BOOSTING WITH SMOTE

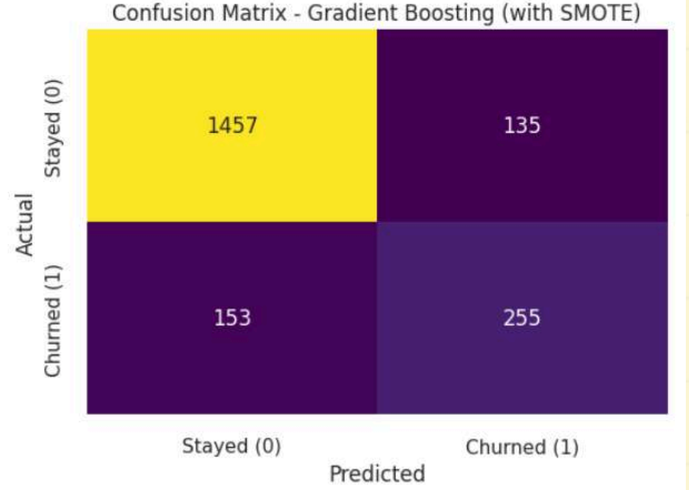
```
--- Training Gradient Boosting with SMOTE ---
ROC AUC: 0.8716

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.92      0.91      1592
           1       0.65      0.62      0.64       408

    accuracy                           0.86      2000
   macro avg       0.78      0.77      0.77      2000
weighted avg       0.85      0.86      0.85      2000


Confusion Matrix:
[[1457  135]
 [ 153  255]]
```



First, the **ROC AUC value is 0.8716**, indicating a 87.16% probability that the model will rank a randomly selected churned sample higher than a non-churned sample.

Second, we move on to **the classification report**. For 0 (stayed) and 1 (churn)

0 (stayed)

- **Precision 0.90** indicates that of all the predictions made by the stayed model, only 90% were correct.
- **Recall 0.92** indicates that 92% of all customers did not churn.
- **F1-Score 0.91** is the harmonic mean of precision and recall.
- **Support 1592** indicates the number of customers who stayed.

1 (churn)

- **Precision 0.65** indicates that of all the predictions made by the churn model, only 65% were correct.
- **Recall 0.62** indicates that 62% of all customers churn.
- **F1-Score 0.64** indicates a balance between precision and recall.
- **Support 408** indicates the number of customers who churned.

DATA MODELLING: GRADIENT BOSSTING WITH SMOTE

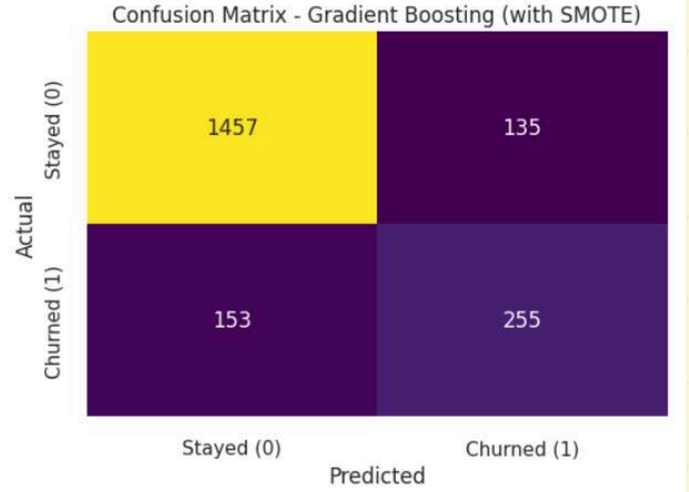
```
--- Training Gradient Boosting with SMOTE ---
ROC AUC: 0.8716

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.92      0.91      1592
           1       0.65      0.62      0.64       408

    accuracy                           0.86      2000
   macro avg       0.78      0.77      0.77      2000
weighted avg       0.85      0.86      0.85      2000


Confusion Matrix:
[[1457  135]
 [ 153  255]]
```



Based on the confusion **matrix for the Random Forest model with SMOTE,** here is an interpretation of the results:

- **True Negatives (TN) = 1457:** The model correctly predicted that 1,457 customers would stay (not churn).
- **False Positives (FP) = 135**: The model incorrectly predicted that 135 customers would churn, when in fact they stayed. That is a Type I error or "false alarm."
- **False Negatives (FN) = 153**: The model incorrectly predicted that 1153 customers would stay, when in fact they churned. It is a Type II error or "miss."
- T**rue Positives (TP) = 255**: The model correctly predicted that 255 customers would churn.

We can calculate **key performance metrics to better understand the model's effectiveness:**

- **Accuracy**: (255+1457)/(255+1457+135+153)=1712/2000=**0.856**. The model's overall accuracy is 85.6%.
- **Precision** 255/(255+135)=255/390≈**0.654**. The precision is approximately 65.4%. This means that when the model predicts a customer will churn, it's only correct about half the time.
- **Recall**=255/(255+153)=255/408≈**0.625**. The recall is approximately 62.5%. This indicates that the model successfully identified more than half of the true churn cases.

The Gradient Boosting model, enhanced with SMOTE, shows a strong overall performance, reflected in its moderate accuracy. A key highlight is its **precision of 65.4%** 🟡, indicating a low rate of false alarms. Furthermore, its solid **recall of 62.5%** 🟡 confirms that the model is quite effective at capturing customers who are actually at risk of churning.

DATA MODELLING: DENSE NEURAL NETWORK WITH SMOTE

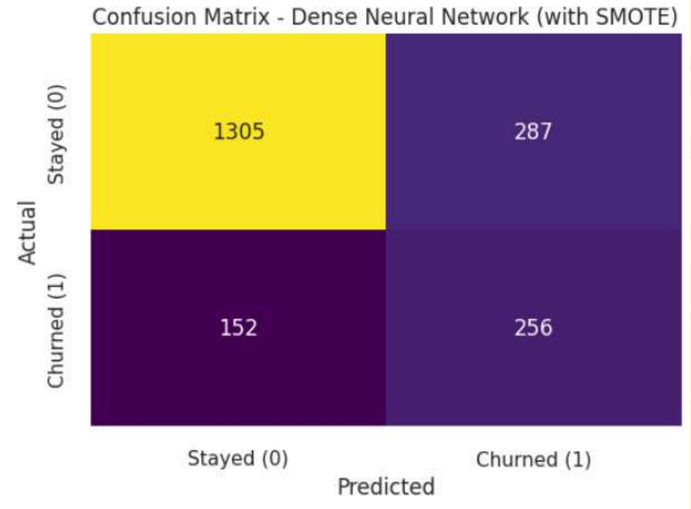
ROC AUC: 0.8190

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.82	0.86	1592
1	0.47	0.63	0.54	408
accuracy			0.78	2000
macro avg	0.68	0.72	0.70	2000
weighted avg	0.81	0.78	0.79	2000

Confusion Matrix:

[[1305 287]
[152 256]]



First, the **ROC AUC value is 0.8190**, indicating a 81.90% probability that the model will rank a randomly selected churned sample higher than a non-churned sample.

Second, we move on to **the classification report**. For 0 (stayed) and 1 (churn)

0 (stayed)

- **Precision 0.90** indicates that of all the predictions made by the stayed model, only 90% were correct.
- **Recall 0.82** indicates that 82% of all customers did not churn.
- **F1-Score 0.86** is the harmonic mean of precision and recall.
- **Support 1592** indicates the number of customers who stayed.

1 (churn)

- **Precision 0.47** indicates that of all the predictions made by the churn model, only 47% were correct.
- **Recall 0.63** indicates that 63% of all customers churn.
- **F1-Score 0.54** indicates a balance between precision and recall.
- **Support 408** indicates the number of customers who churned.

DATA MODELLING: DENSE NEURAL NETWORK WITH SMOTE

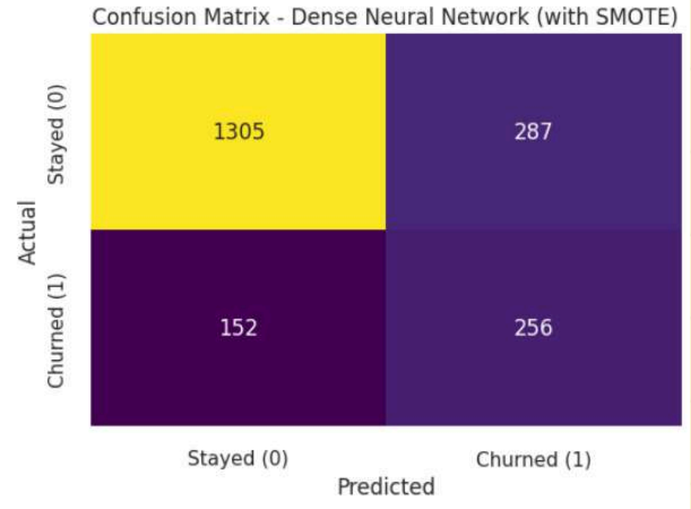
ROC AUC: 0.8190

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.82	0.86	1592
1	0.47	0.63	0.54	408
accuracy			0.78	2000
macro avg	0.68	0.72	0.70	2000
weighted avg	0.81	0.78	0.79	2000

Confusion Matrix:

```
[[1305  287]
 [ 152  256]]
```



Based on the confusion **matrix for the Random Forest model with SMOTE,** here is an interpretation of the results:

- **True Negatives (TN) = 1305:** The model correctly predicted that 1,305 customers would stay (not churn).
- **False Positives (FP) = 287**: The model incorrectly predicted that 287 customers would churn, when in fact they stayed. That is a Type I error or "false alarm."
- **False Negatives (FN) = 152**: The model incorrectly predicted that 152 customers would stay, when in fact they churned. It is a Type II error or "miss."
- T**rue Positives (TP) = 256**: The model correctly predicted that 234 customers would churn.

We can calculate **key performance metrics to better understand the model's effectiveness:**

- **Accuracy**: (256+1305)/(256+1305+287+152)=1561/2000=**0.7805**. The model's overall accuracy is 78.05%.
- **Precision** 256/(256+287)=256/543≈**0.471**. The precision is approximately 47.1%. This means that when the model predicts a customer will churn, it's only correct about half the time.
- **Recall**=256/(256+152)=256/408≈**0.627**. The recall is approximately 62.7%. This indicates that the model successfully identified more than half of the true churn cases.

The Dense Neural Network model, with an overall accuracy of 78.05%, shows a good ability to predict churn. Its recall of 62.7% 🟡 is a key strength, demonstrating that it's effective at capturing real churn cases. However, **a low precision of 47.1%** 🔴 reveals that it often incorrectly flags customers as churn risks.

DATA MODELLING: CONCLUSION

Here we are, in the final part of processing and modeling data using machine learning: evaluating the best model.

The ROC (AUC) and F1-Score metrics are our key indicators for determining which model most accurately predicts customer retention and churn, while balancing precision and recall.

F1 Score is a performance metric used in machine learning to evaluate how well a classification model performs on a dataset, especially when the classes are imbalanced, meaning one class appears much more frequently than another. It is the harmonic mean of precision and recall, combining both metrics into a single value that balances their importance.

Considering both the ROC AUC and F1-score, Gradient Boosting emerges as the most effective model for accurately predicting churn rates.

```
# Create a pandas DataFrame from the results dictionary with SMOTE
results_df_smote = pd.DataFrame(results_smote)
results_df_smote
```

	Model	Balancing (SMOTE)	ROC AUC	F1 Score (Churned)
0	Logistic Regression	Yes	0.781907	0.50
1	Decision Tree	Yes	0.706270	0.52
2	Random Forest	Yes	0.859095	0.61
3	Gradient Boosting	Yes	0.871584	0.64
4	Dense Neural Network	Yes	0.818982	0.54

EVALUATION AND INSIGHT

Based on data analysis, from dataset identification, EDA, to data modeling, we uncovered several interesting insights:

Findings & Caused by:

- **Churn rates are influenced by geographical location. While most customers are from France, the data shows that Germany has the highest churn rate.** This disparity could stem from external factors as well as differing banking regulations, cultural norms, or fluctuations in interest rates across regions.
- **Male gender dominates churn rates** compared to the female gender.
- **Credit card ownership influences churn rates.** It might be a result of customers fully settling their credit card debts before deciding to move to a different bank.
- **The data shows that customers with poor and fair credit scores are the most likely to churn.** This pattern could stem from the bank's initial credit assessment process, which considers factors such as the 5Cs (Character, Capacity, Capital, Collateral, and Condition), or from customers transferring their credit to other financial institutions.

EVALUATION AND INSIGHT

Based on data analysis, from dataset identification, EDA, to data modeling, we uncovered several interesting insights:

Findings & Caused by:

- **Working-age groups churn more often than stayers.** It may be a result of the productive-age group's greater willingness to diversify their credit sources, rather than staying loyal to just one financial institution. Opportunists seeking low interest rates could be a key factor.
- **Inactive customers are more likely to churn**, which could be due to a lack of customer retention.

EVALUATION AND INSIGHT

Based on these findings, the solutions we can offer are:

Solutions:

- **Conduct in-depth research on the products and services to be launched or marketed.** Examples of these considerations include the expenses of product development, the value provided to customers, the market's reaction, and the product's unique selling points against rivals.
- **Increase customer retention, such as enhancing customer relationships** (i.e., maintaining communication with customers, adopting a personalized approach, improving excellent customer service) **and implementing loyalty programs** (e.g., member promotions, tiered rewards, customer loyalty points).
- **Communicate regularly with customers**, including email newsletters, social media, and personalized messages, regarding new products, updates, and customer testimonials about bank services.

EVALUATION AND INSIGHT

If we return to focus on the results of the modeling data:

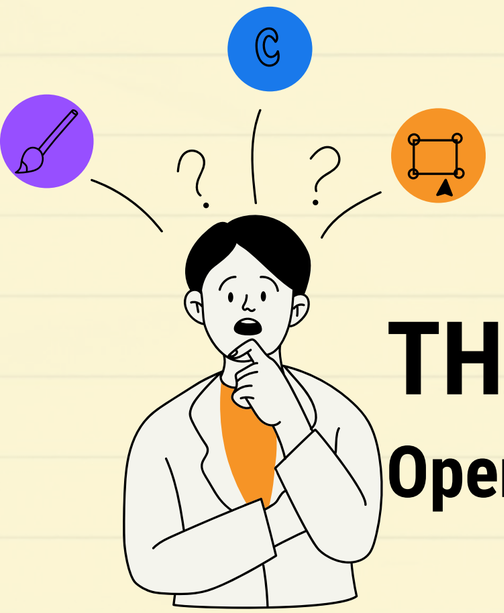
Data Modelling Findings:

Based on the ROC AUC and F1-score results, the Gradient Boosting model is the best performer.

So the suggestion for further research is:

Data Modelling Suggestion:

To further improve the model, we can test other classification algorithms, such as k-means or XGBoost. Additionally, using a different set of features or a different data combination might provide a more robust analysis.



THANK YOU!!

Open Discussion and Feedback:

[Jihan Dewana]

[Jihan Dewana]

[Bank Customer Churn_JihanDewana.ipynb]