**FLY THE NEST**

Project Report

On

“Exploratory Data Analysis on the Banking Customer Churn Prediction”

**Submitted by:** Riddhi Limani

**Under The Guidance of** Chintan Patel

**Affiliated With:** Fly the Nest Instituted.

**PROJECT REPORT**

Table of Contents

[Problem Statement: 4](#_Toc197778393)

[Objectives: 4](#_Toc197778394)

[Introduction: 5](#_Toc197778395)

[Exploratory Data Analysis (EDA) 5](#_Toc197778396)

[EDA Pipeline: 5](#_Toc197778397)

[Overview: 6](#_Toc197778398)

[Data Loading / Reading: 6](#_Toc197778399)

[Familiarize with Data & Identifying the Target Variable 7](#_Toc197778400)

[Data Preparation & Transformation: 8](#_Toc197778401)

[Data analysis & visualization 9](#_Toc197778402)

[Overview of Data Before Analysis 9](#_Toc197778403)

[Analysis 9](#_Toc197778404)

[Summary and Suggestions 21](#_Toc197778405)

[Github Link: 26](#_Toc197778406)

# 

# Problem Statement:

Predicting Customer Churn for a Bank the objective is to build a predictive model to identify customers who are likely to leave the bank. By analyzing historical customer data, including demographics, account information, and transaction history, we aim to develop insights into customer behavior and implement strategies to improve customer retention. Understanding the characteristics of customers who churn. Identify Key drivers influencing churn, build a predict customer churn and evaluate the model’s performance and suggest actionable recommendations.

# Objectives:

Understand Customer Demographics and Behavior – Analyze key demographic features such as age, gender, geography, and credit score to identify patterns related to churn.

Explore Account and Banking Features – Investigate features such as account balance, number of products, tenure, and credit card ownership to understand their influence on churn.

Identify Key Drivers of Churn – Use statistical and visual methods to determine which features most strongly correlate with customer churn.

Build a Predictive Model – Develop a machine learning to predict the likelihood of a customer churning using features in dataset.

Evaluate Model Performance – Measure model performance using appropriate metrics (accuracy, precision, recall, F1\_score, ROC-AUC).

Generate Actionable Insights – Provide strategic recommendations for customer retention based on analysis and model findings.

# Introduction:

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize the data, using statistical analysis and data visualization methods. It is a crucial step in any data analysis process, enabling data analysts to uncover patterns, spot anomalies and gain insights to take actions.

## EDA Pipeline:

1. Data Acquisition and Objective
2. Obtain Banking Customer Churn Prediction data (CSV, Excel)
3. Get problem statement from Euromart
4. Choose tools/environment & programming Language
5. Data Loading / Reading
6. Load data in Jupyter Notebook/ VS code (To perform further analysis)
7. Preview first few rows.
8. Check dimensions, data types, and missing values.
9. Familiarize with Data & Identify Target Variable
10. Explore data (column names, data types)
11. Identify target variable based on objective
12. Data Preparation & Transformation
13. Data Cleaning
14. Handle missing values (if required)
15. Removal of unwanted data (if present)
16. Format data types (numerical & categorical)
17. Feature Engineering (Create new features)
18. Data Analysis & Visualization
19. Univariate Analysis

* Numerical variables (mean, median)
* Categorical variable (distribution)

1. Bivariate & Multivariate Analysis (Identify patterns)

* Visualization (Chats: pie, boxplot, histogram, heatmap)

1. Summary and Suggestions

## Overview:

* The Banking Customer Churn Prediction Project aims to predict whether a customer will churn based on historical data such as Customer demographics, Account activity, Transaction Patterns, Credit behavior. This is a classification problem, typically solved using algorithms.
* Customer churn refers to the loss of clients or customer. When a customer stop doing business with a company. In banking, churn means a customer closes their account or discontinues using the bank’s services.
* Retaining a customer is cheaper than acquiring a new one. Churn prediction helps banks focus their marketing and retention strategies. It enables personalized offers and better customer relationship management (CRM).

# Data Loading / Reading:

Import Necessary Library

* NumPy (np): Provides efficient numerical computation tools
* Pandas (pd): Offers data manipulation and analysis structures (Data Frames, Series)
* Seaborn (sns): Creates informative statistical data visualizations based on Matplotlib
* Matplotlib.pypolt (plt): Enables various plot creations for data visualization
* %matplotlib inline (Jupyter Notebook specific): Displays plots within the notebook
* Warnings (with warnings.filterwarnings(“ignore”)): Suppresses warning
* Ploty.express: Simplifies the creation of interactive and visually appealing charts with minimal code.

Load Data in Jupyter Notebook

To begin the analysis, we first load the dataset into a pandas DataFrame using the read\_csv() function. This allows us to explore and manipulate the data efficiently in Jupyter Notebook.

Df = pd.read\_csv(“Churn\_Modelling.csv”)

# Familiarize with Data & Identifying the Target Variable

Explore the provided data (column names, data types)

* We need to understand the data before cleaning the data and cross verify if all the required data are provided by Company

Overview of data

* df.head(): Let’s see the data by displaying the first 5 rows
* df.tail(): Let’s see the last 5 rows
* df.shape is used to get the dimensions (number of rows and columns) of data
* df.size is used to get the total number of elements in a pandas
* df.info(): used to display concise information about

Interpretation:

* Structured Data: The data is provided in a structured table format.
* Dimensions: The dataset contains 14 columns and 10,001 rows, totaling 140,000 elements.
* Column Data Types: Observed a mix of data types:
* Categorical (Object): 6 columns
* Numerical (Float64): 4 columns
* Categorical Variables: All categorical / qualitative variables are Nominal in nature.
* Non-Null Counts: No null values were found in the dataset. Each column has 10,001 non-null values.
* Memory Usage: The dataset consumes approximately 4.59 MB of memory.

(Further optimization may be performed by modifying data types where applicable.)

# Data Preparation & Transformation:

**Data Cleaning**

We need to perform steps mentioned below to clean data:

* Steps involved in handling missing values (imputation, deletion)
* We accept missing values if data is small in dimension
* We delete missing values id:
* When more than 80% of data is missing/null values
* When the percentage of missing values are very small, deleting with have minimal effect on analysis
* Replacing the missing values by imputation
* Imputation: We replace the missing values be Mean, Median or Mode of the variable or perform fill null values with the desired value
* Data Reduction: Remove unwanted data (if present) which are not required for analysis
* Delete unwanted columns
* Delete duplicate rows
* Format data types (numerical & categorical variables)
* Outlier detection and handling (we ignore this step because outliers are valid in our case)
* When data has extreme values that could affect our analysis, we either replace them with Mean or Median or Mode or we accept the outliers
* We identify the outliers by plotting the Box plot

Handling Missing Values:

* df.isnull().sum()- Gives sum of all null values in each column
* df.notnull().sum()- Gives sum of all not null unique values in each column
* Interpretation:
* Data has no null values, so no need to perform process to handle missing values
* But In our data there is no null values or any other Duplicate values means the data provided is properly cleaned and well managed.

Format data types (numerical & categorical)

We need to format columns, that will ease data analysis, below steps are performed to the required format for analysis.

* Rename of column – To keep columns descriptive as well as simple
* Change data types – We changes data type to keep consistency and also for memory optimization

# Data analysis & visualization

## Overview of Data Before Analysis

After Data Wrangling, we check the dataset structure before proceeding with analysis.

Columns in DataFrame: [‘Row Number’, ‘Customer Id’, ‘Surname’, ‘Credit score’, ‘Geography’, ‘Gender’, ‘Age’, ‘Tenure’, ‘Balance’, ‘HascrCard’, ‘Is Active Member’, ‘Estimated Salary’, ‘Exited’]

**Variables/Columns Description**

Row Number - The sequential number of the record in the dataset

Customer Id - A unique identifier for each customer

Surname - The last name of the customer

Credit Score - A numerical score representing the customer’s creditworthiness.

Geography - The country or region where the customer resides.

Gender - The Customer’s Gender

Age - The customer’s age.

Tenure - Number of years the customer has been with the bank.

Balance- The amount of money the customer currently holds in their account.

HasCrCard - Indicates whether the customer has a credit card(1=Yes, 0=No).

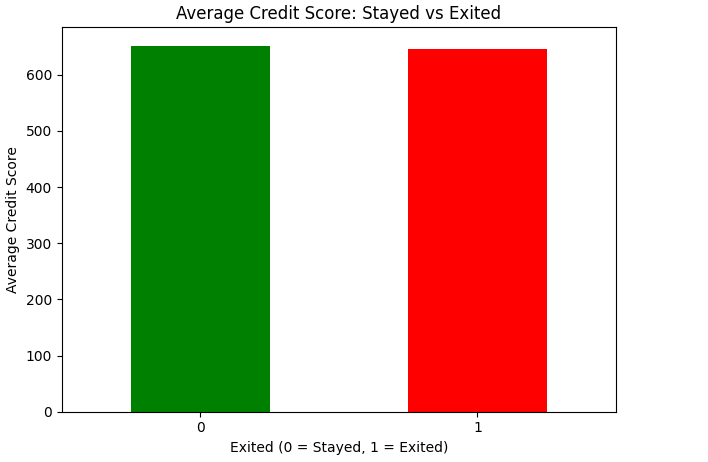
Is Active Member - Indicates whether the customer is considered an active member.

Estimated Salary - An estimation of the customer’s annual salary.

Exited - Indicates whether the customer has left the bank (1 = Yes, 0 = No).

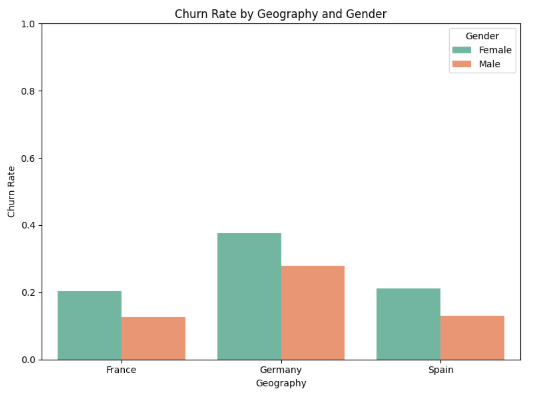
# Analysis

* Average credit score of customer who exited vs stayed?



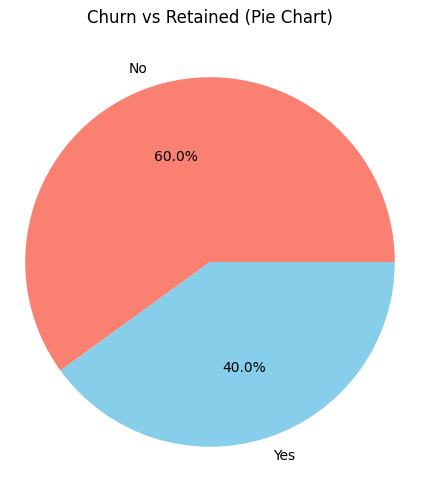
* The bar charts illustrates the “Average Credit Score: Stayed vs Exited” compares the average credit scores of customers who stayed (Exited = 0) versus those who exited (Exited = 1).
* Stayed customers represented by the green bar and average credit score is slightly above 650.
* Exited customers represented by the red bar and average credit score is slightly below 650.
* While the difference is not very large, it suggests that customers with higher credit scores may be slightly more loyal of financially stable, reducing their likelihood of exiting.

Churn rate vary by geography and gender



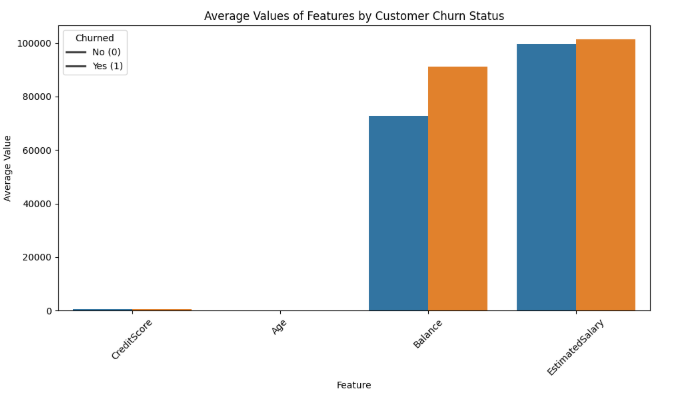
* The bar charts illustrates “Churn Rate by Geography and Gender” presents churn rates in three countries (France, Germany, and Spain) separated by gender (Female in green and Male in orange).
* Females have a higher churn rate than male across all three countries.
* Germany has the highest churn rates for both genders.
* France and Spain show relatively lower and more comparable churn rates.

Churn vs Retained



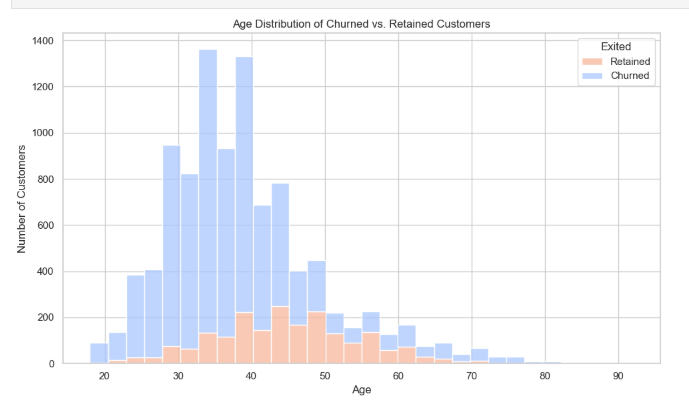
* This pie chart displays the proportion of customers who have churned (“Yes”) versus those who have remained (“No”).
* 40% of customers have churned “Yes” and 60% of customers have been retained “No”.
* This suggests that the majority of customers are still active, but a 40% churn rate is significant and may indicate underlying issues (e.g., dissatisfaction, better alternatives, pricing concerns).

Average values feature by customer churn status



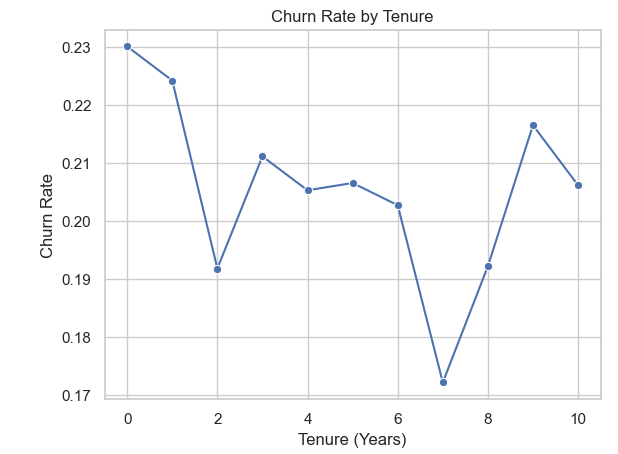
* This bar charts compares average values of four features between two groups of customers: those who churned (Yes/1) and those who did not churn (No/0).
* The suggests customers with lower credit scores may be more likely to churn.
* Age is visually similar between both groups, possibly slightly higher for churned customers.
* Churned customers have a significantly higher average balance. This is noteworthy: customers who churn tend to leave with more money in their accounts.
* It could imply dissatisfaction despite holding substantial balances, or possibly a lack of engagement.
* Estimated salary very similar for both groups. Salary level does not appear to be a strong predictor of churn in this dataset.

Age distribution of Churned vs. Retained Customers



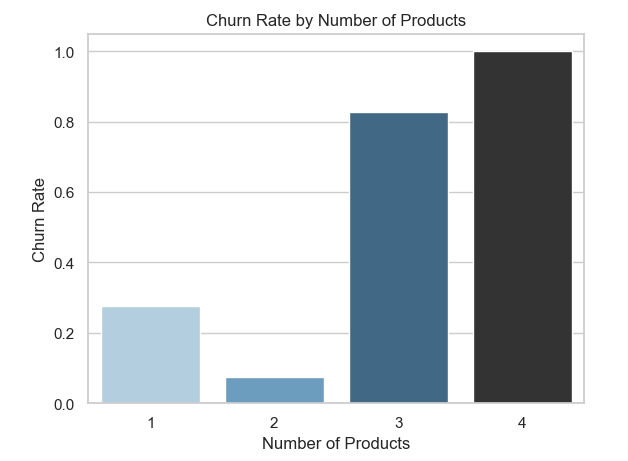
* This histogram shows how customer churn varies with age, comparing the number of churned vs retained customers across different age groups.
* Younger customers (Under 35) is largest population segment. Most of these customers are retained and churn is relatively low in this group.
* Middle-Aged customers (35-50) is churn starts to increase significantly. Around age 40-45, churn becomes more prominent relative to retention.
* This is a critical age range where intervention could reduce churn.
* The older Customer (50+) fewer total customers, but the churn rate is noticeably higher. Churn becomes dominant in many of these age bins, suggesting older customers are more likely to exit.

Churn Rate by Tenure



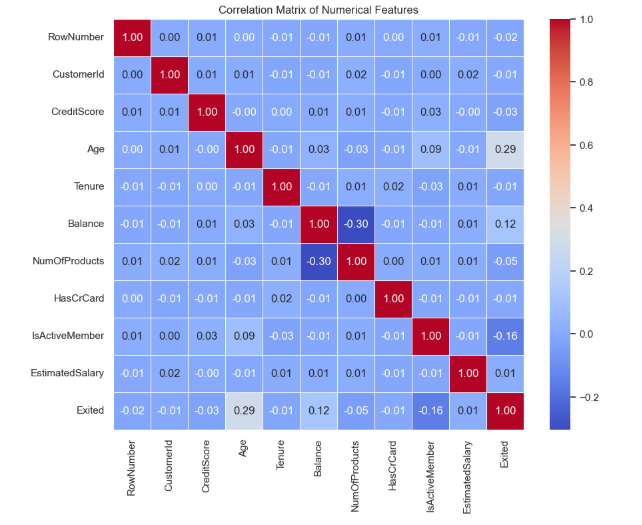
* This chart illustrates how the churn rate varies based on customer tenure the number of years they’ve been with the company.
* Highest Churn at the start customers in their first and second year have the highest churn rates. Indicates that early-stage retention is a critical problem.
* Churn rate significantly drops around tenure = 2 and again at 7. The suggests customers who make it past year 2 more likely to stay longer.
* Mid-Tenure Stability churn rates remain relatively stable between 20-21%. The slight fluctuations, but no sharp spikes or drops.
* The noticeable up stick in churn at 9 years, followed by a slight drop again at 10.
* Could be related to end-of-product cycles, changes in services, or customer fatigue.

Churn Rate by Number of Product



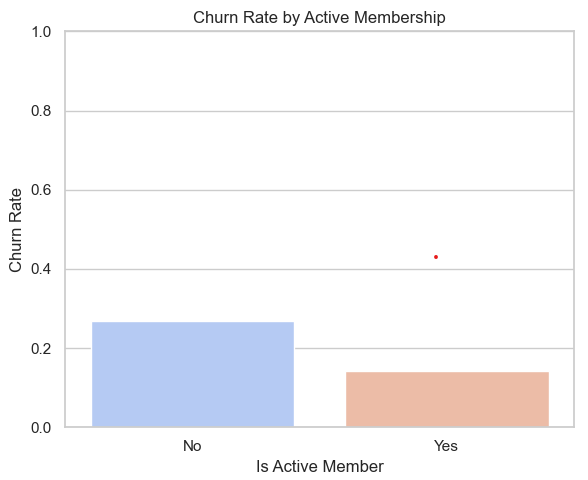
* This bar charts displays how customer churn rates vary based on the number of products they hold.
* Customers with exactly 2 products have the lowest churn rate. That offering a second product may significantly increase retention.
* Churn increases 3 products churn rate jumps to over 80% and 4 products churn rate hits 100% meaning all customers with 4 products churned.
* Moderate churn rate for customers with only 1 product. Indicates limited engagement or value derived from the relationship.

Correlation Matrix of Numerical Feature



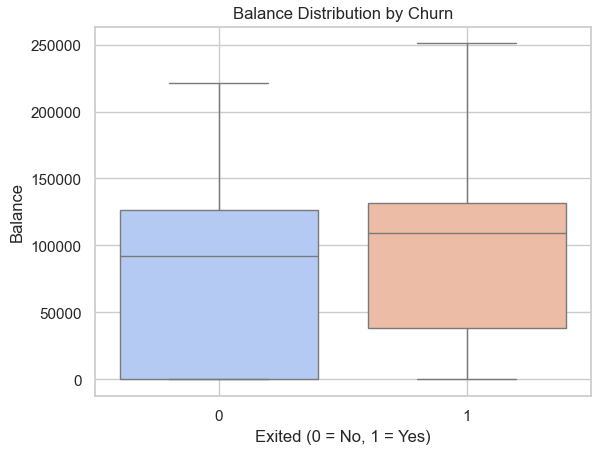
* This heatmap shows a correlation matrix of numerical features from a dataset likely from a bank or customer churn dataset.
* Diagonal is 1.00 as expected, every feature is perfectly correlated with itself.
* Low overall correlation most variables have very low correlation with each other shown by the blue color, indicating weak linear relationships.
* Raw number and customer id these are likely identifiers and negligible correlation with other variable they should probably be excluded from modeling.
* The heatmap helps identify which features might be more relevant when predicting customer behavior. Age and activity level seem to show some predictive potential, while many others like customer id or credit score have little to no linear relationship with the target.

Churn Rate by Active Member



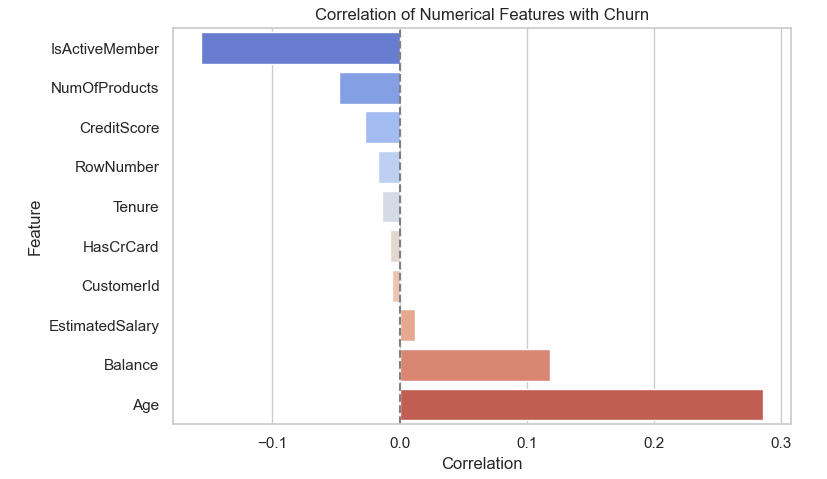
* This bar chart compares the churn rates between customers who are **active members** and those who are not active.
* Inactive Member (No) have a higher churn rate (-27%) and Active Member (Yes) have a lower churn rate (-14%).
* A red dot potentially an outlier or mean indicator appears over the “Yes” group, which might suggest a statistical comparison or highlight.
* Encouraging customers to remain or become active member could be an effective churn-reduction strategy.

Balance Distribution by Churn



* This box plot shows the distribution of customer balances for two groups: 0 (Not Exited / Retained), 1 (Exited / Churned).
* The median balance central line in the box is noticeably higher for churned customers than for retained ones. This suggests customers with higher balance are more likely to churn.
* Churned customer has a broader range of balance, from near 0 up to 2,50,000.
* Retained customers have more lower balances with a tighter interquartile range.
* Both groups have some extreme values. But they’re especially pronounced among retained customers.
* Customers with higher balances might be feeling underserved or under-engaged, leading them to leave despite having significant funds in their accounts.

Correlation of Numerical Features with churn



* This bar charts visualizes the correlation between numerical features and customer churn.
* Age has the highest positive correlation with churn. This means older customers are more likely to churn.
* Balance and Estimated salary both show a positive but weaker correlation with churn. Customer with higher balance salaries may be slightly more likely to churn, but the relationship is weak.
* Active member are less likely to churn, which makes sense logically. Shows the strongest negative correlation.
* Number of products and credit score have slightly negative correlations. That more products or better credit scores slightly reduce the likelihood of churn.
* Other features very low or no correlation with churn. These features are likely not good predictors of churn on their own.

# Summary and Suggestions

Based on the analysis, here are my key summary and suggestions for improving business performance:

Average credit score of customer who exited vs stayed

Summary:

* Stayed customers represented by the green bar and average credit score is slightly above 650.
* Exited customers represented by the red bar and average credit score is slightly below 650.
* While the difference is not very large, it suggests that customers with higher credit scores may be slightly more loyal of financially stable, reducing their likelihood of exiting.

Suggestion:

* Add Numerical Labels: Display the exact average credit scores on top of each bar for clarity.
* Add Legend or Labels: Clarify the meaning of the colors for better accessibility,
* Provide context: Mention the sample size or total number of customers in each category.

Churn rate vary by geography and gender

Summary:

* Germany has the highest churn rate overall, especially among females.
* France and Spain show a lower churn rate, with females still having a slightly higher churn rate than males.
* In all three countries, female customers churn more than male customers.

Suggestion:

* Clarify Axis Titles: Include units or explain if churn rate is a ration or percentage.
* Add Data Labels: Display exact churn rate values above each bar for precision.
* Improve Title: Consider a clearer, more specific title like “customer churn Rate by country and gender”.
* Add gridlines or Annotations: Subtle gridlines or country-level average churn annotations can enhance interpretation.

Customer Vs. Retained

Summary:

* The churn rate is 40%, which is relatively high. Ideally, most business aim for much lower churn, especially in subscription-based or long- term relationship models.
* The majority of customers (60%) are still active, which is good sign, but the substantial churn suggests room for improvement in customer satisfaction, engagement, or value delivery.

Suggestion:

* Customer Segmentation Analysis: Identify which customer groups (e.g., by age, tenure, contract type) have the highest churn, Target interventions accordingly.
* Improve Customer Onboarding & Support: Ensure customer understand how to get value quickly. Provide proactive support or check-ins during early stages of the customer lifecycle.
* Gather Feedback from Churned Customers: Use exit surveys or interviews to learn why they left. Look for common patterns like price concerns, poor service, or better competitor offerings.
* Retention Strategies: Offer loyalty programs, discounts, or incentives for long-term customers. Improve product features or customer experience based on feedback.

Age Distribution of Churn Vs. Retained Customers

Summary:

* Most customers are concentrated between 30 and 50 years old.
* The number of churned customers increases with age, especially noticeable from 40 onwards.
* Younger customers have very low churn rates.
* The churn rate seems relatively higher in the 40-60 age range.

Suggestions:

* Overlay Line Plot for Churn Rate by Age: In addition to counts, plot the percentage churn rate by age group to get a clear sense of churn risk across ages.
* Bin Width Adjustment: Consider finer bins for better granularity, especially if dataset size allows.
* Use Percentage Stacked Bars: Instead of absolute counts, a percentage stacked bar plot would highlight the proportion of churn vs. retention in each age group.
* Highlight Key Age Bands: Add shaded regions or annotations to emphasize age groups with notably higher churn.
* Statistical Summary or Trendline: Fit and overly a smoothed curve to visualize churn trend by age.

Churn rate by tenure

Summary:

* New customers (0-1 year) are most vulnerable to churn.
* A significant drop in churn at 2 years suggests a key retention milestone.
* Customers with 7+ years also show renewed loyalty, except for a spike at 9 years.

Suggestion:

* Focus Heavily on Onboarding and First-Year Experience: The first 12 months are critical. Implement personalized onboarding, frequent check-ins, and early incentives.
* Milestone-Based Engagement: Recognize and reward customers at key points to reinforce loyalty.
* Investigate Year 9 Churn Spike: Understand if policy, pricing, or life-stage transitions drive this spike. Consider special retention offers at this stage.

Balance Distribution by churn

Summary:

* Median Balance: Customers who churned tend to have a slightly higher median balance than those who stayed.
* Distribution Spread: The interquartile range and spread wider for churned customers, indicating more variability in their balances.
* Outliers: Both groups have outliers, especially on the higher and, suggesting some customers have exceptionally large balances.

Suggestion:

* Statistical Testing: Use a t-test or Mann-Whitney U test to determine if the balance difference between churned and non-churned customers is statistically significant.
* Correlation Matrix: Create a correlation matrix including all numerical features and the churn flag to assess liner relationship.
* Multivariate Analysis: Incorporate other feature into a logistics regression or decision tree to evaluate churn drives more comprehensively.
* Segmentation: Segment by customer types and replot to uncover hidden patterns in balance-churn relationships.

Correlation Matrix of Numerical Features

Summary:

* Key correlations with Exited:
* Age: +0.29 – Moderate positive correlation. Older customers are more likely to churn.
* IsActiveMember: -0.16 – Negative correlation. Active members are less likely to churn.
* Balance: +0.12 – Slight positive correlation with churn.
* NumOfProducts: -0.05 – Very weak negative correlation.
* CreditScore, Tenure, EsitimatedSalary: near- zero correlation.
* CustomerId, RowNumber, HasCrCard: Essentially no correlation.
* Other Notable Realtionships:
* Balance vs. NumOfProducts: -0.30 indicating customers with more products may keep less balance.
* Most variables are not strongly correlated with each other a good sign for modeling.

Suggestion:

* Feature Selection:
* Focus on Age, IsActiveMember, and Balance – these show the highest relevance to churn.
* Consider removing or deprioritizing CustomerId, RowNumber, HasCrCard, Tenure and EstimatedSalary.
* Modeling Tips:
* Tree-based models handle weakly correlated features well and may uncover non-linear patterns.
* Standard models like Logistic Regression may benefit from reducing input to only top-correlated features.
* Validate Insights:
* Use SHAP of feature importance plot to confirm which features truly drive churn predictions in your models.

Correlation Of Number Features with Churn

Summary:

* Top Features Positively Correlated with Churn:
* Age: older customers are more likely to churn.
* Balance: Customers with higher balances tend to churn slightly more.
* Estimated Salary: Very weak positive correlation.
* Negatively Correlated with Churn:
* IsActiveMember: Active members are less likely to churn.
* NumOfProducts: Customers with more products tend to stay.
* CreditScore and Tenure: Weak negative correlation.
* Near-Zero Correlation:
* Features like CustomerId, RowNumber, and HasCrCard have nearly zero correlation and are likely irrelevant for churn prediction.

Suggestion:

* Drop Low-Impact Features: Consider removing or deprioritizing CustomerId, RowNumber and possibly HasCrCard in modeling, as they carry little predictive power.
* Feature Engineering: Age and Balance are influential consider binning age group or deriving flags.
* Explore Interactions: Investigate if combinations give deeper insights.
* Model Input Selection: Use top correlated variables as inputs for logistic regression, decision trees, or ensemble models.
* Visualize No-Linear Trends: Since correlation only captures linear relationships, visualize churn by age buckets, balance bands, or product count for deeper insights.

GitHub link: <https://github.com/riddhilimani?tab=repositories>