Introduction

In today's competitive banking landscape, retaining customers has become increasingly important for financial institutions. Customer churn, the loss of customers to a competitor or due to other factors, can have a significant impact on a bank's profitability and long-term growth . Bank churn analysis involves studying the factors influencing customers' decisions to terminate their relationships with a bank. By understanding the reasons behind customer churn, banks can develop targeted strategies to address these issues and reduce the rate of customer attrition

About Dataset

For this data visualization project, we will be using the BankChurners dataset, which includes data on 10,127 customers and their banking profiles. The dataset has 23 columns with various attributes, such as customer demographics, banking behaviors, and product usage. Key variables include customer age, gender, credit limit, transaction amounts, and whether the customer has churned (left the bank).
We chose this dataset because it provides a comprehensive look at customer behavior within the banking sector, making it an ideal resource for identifying trends and patterns. Additionally, with the inclusion of an attrition flag, this dataset allows for churn analysis, which is valuable for understanding factors that influence customer retention. These insights can be essential for data-driven decision-making, particularly in customer service, marketing, and product management within financial institutions.

Data Cleaning

In this process, we tackled the issue of "Unknown" values in the dataset to prepare it for accurate analysis. We used a Label Encoder to convert categorical text into numbers, then employed an Iterative Imputer with a RandomForestClassifier as its core to predict and fill in these "Unknown" values based on patterns found in other data. This machine learning approach allowed us to infer the missing values smartly and preserve the dataset's integrity for more reliable results. Once the imputation was complete, we converted the numbers to their original categories, effectively replacing "Unknown" values with meaningful data.

This is the data as it exists before we perform any data cleaning
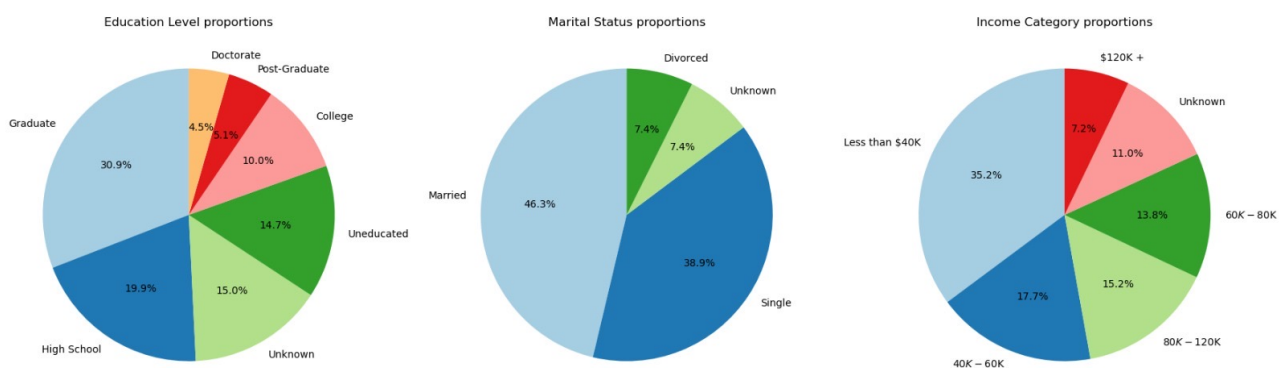
After completing data cleaning, here is the updated data