**Faculty of Engineering & Technology Electrical & Computer Engineering Department**

** Machine Learning and Data Science ENCS5341**

**Assignment 1**

**Data Preprocessing & Exploratory Data Analysis (EDA)**

**Prepared by:**

Qasim Batrawi 1220204

Ahmad Hamdan 121XXXX

**Instructor:** Dr. Yazan Abu Farha

**Section:** 2

**Date:** 22-10-2025

Table of Contents

[1 Introduction and Dataset Overview 1](#_Toc212134976)

[2 Handling Missing Data 1](#_Toc212134977)

[3 Handling Outliers 3](#_Toc212134978)

[4 Feature Scaling 5](#_Toc212134979)

[5 Exploring Data Analysis (EDA) 6](#_Toc212134980)

[5.1 Univariate Analysis 6](#_Toc212134981)

[5.2 Bivariate Analysis 8](#_Toc212134982)

[5.3 Correlation Analysis 12](#_Toc212134983)

[6 Data Visualizations 13](#_Toc212134984)

[7 Conclusion 13](#_Toc212134985)

Table of Figures

[Figure 1 Number of Missing Values in each Feature. 1](#_Toc212134987)

[Figure 2 ChurnStatus by Tenure Group. 2](#_Toc212134988)

[Figure 3 Boxplot for Numerical values before Resolving Outlier Values. 3](#_Toc212134989)

[Figure 4 Boxplot for Numerical values after Resolving Outlier Values. 5](#_Toc212134990)

[Figure 5 Before Scaling. 6](#_Toc212134991)

[Figure 6 After Scaling. 6](#_Toc212134992)

[Figure 7 Histograms and Boxplots for Numerical Features. 7](#_Toc212134993)

[Figure 8 Bar Plots for Categorical Features. 8](#_Toc212134994)

[Figure 9 Age vs Income colored by ChurnStatus. 9](#_Toc212134995)

[Figure 10 Tenure vs Income colored by ChurnStatus. 10](#_Toc212134996)

[Figure 11 SupportCalls vs Income colored by ChurnStatus. 11](#_Toc212134997)

[Figure 12 Gender and ProductType relate to ChurnStatus. 11](#_Toc212134998)

[Figure 13 Correlation Matrix. 12](#_Toc212134999)

# 1 Introduction and Dataset Overview

Customer churn is an important challenge for businesses. Understanding why customers leave can help companies improve their services and keep more customers. In this assignment, we study the behavior of the customers and identify which customers are more likely to churn. By analyzing the data, we aim to find useful insights that could help to reduce the number of churned customers.

The dataset contains information about the customers including CustomerID, Age, Gender, Income, Tenure, ProductType, SupportCalls, and ChurnStatus. Age, Income, Tenure, and SupportCalls are numerical features, while Gender and ProductType are categorical. ChurnStatus is the target variable, showing whether a customer has churned (1) or stayed (0). This dataset allows us to explore relationships between customer characteristics and churn.

# 2 Handling Missing Data

After inspecting the data, we noticed that some features contain missing values, as shown in the following figure.

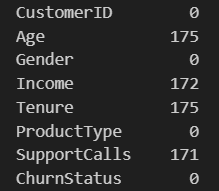


Figure Number of Missing Values in each Feature.

To fix this issue, deleting these rows from the dataset is not a good solution, since we will lose a lot of customers, so each feature have handled independently based on the type of data and how important it is for the analysis.

The Age had 175 missing values. Instead of using the overall mean age, we filled the missing values using the mean age for each gender group separately, since age could vary between male and female. This approach gives more accurate representation of the data.

The Income feature had 172 missing values. To handle this, we grouped customers into age ranges and filled the missing income values using the median income of each age group. This approach assumes that people in similar age are likely to have comparable income. We choose median instead of mean because the income often have extreme values that will affect the mean.

The Tenure feature had 175 missing values. To solve this, we have studied the relation between the ChurnStatus and Tenure resulting with the following plot.

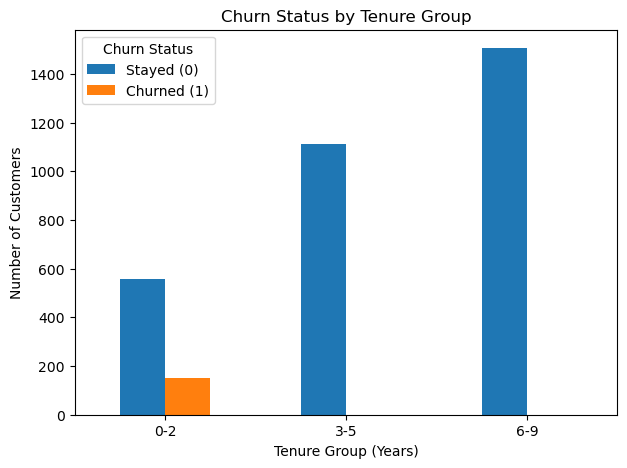


Figure ChurnStatus by Tenure Group.

From the plot, we observed that all customers who churned had been with the company for two years or less. For customers who did not churn, about 17.5% had tenure of 0–2 years, 35% had a tenure of 3–5 years, and 47.5% had a tenure of 6–9 years. Based on this, if the ChurnStatus is 1, we assign the tenure as a random value between 0 and 2 years. Otherwise, we generate a random number between 0 and 1. If this number is less than or equal to 0.175, the tenure is set between 0 and 2 years; if it is less than or equal to 0.475, the tenure is set between 3 and 5 years; otherwise, it falls between 6 and 9 years.

The SupportCalls feature had 171 missing values. Since this information is not significant for the customer, we replaced the missing values with zero. With this, all missing values in the dataset have been resolved without deleting any row, ensuring that the data is complete and reliable.

# 3 Handling Outliers

The following plot shows the boxplot for each numerical feature.

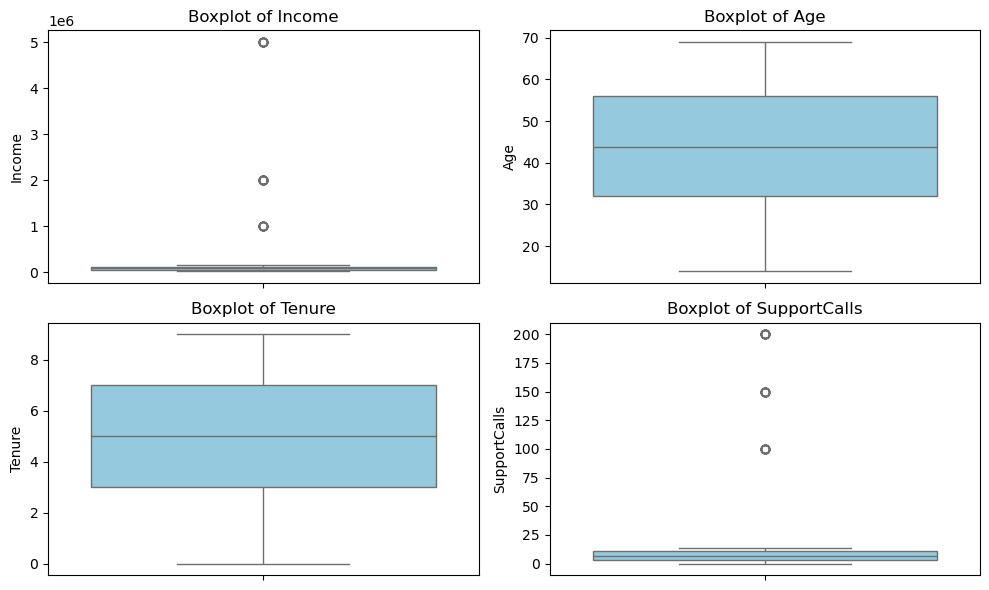


Figure Boxplot for Numerical values before Resolving Outlier Values.

The figure above shows that both Income and SupprotCalls have extreme values marked as outliers. Tenure does not seem to have any outliers. Age also appears that it does not have outliers, but we need to check if there are young customers who may not meet the requirements of the system.

To handle outliers in the Age feature, we considered both Age and Tenure. If the difference between Age and Tenure was less than 15 years (meaning the customer was younger than 15 when registering), we marked the age as an outlier. Using this method, we identified 199 outlier ages. To correct them, we reassigned the age of these customers to a random number between 16 and 20, added to their tenure. This approach ensures that all registered customers are adults, which aligns with the system’s requirement that users must be adults.

For the SupportCalls, we will detect outlier values using the Z-score. We will calculate the Z-score for each support call value using the following equation:

Where:

* is the standardized value,
* is the current value,
* is the mean value, and
* is the standard deviation value.

Any value with a Z-score greater than 3 was considered an outlier. After applying this method, we noticed that there are 70 outlier values. To resolve these outliers, the binning method was used. The data were first sorted, and then grouped into bins of size 500, and each value within a bin was replaced by the median of that bin. The median was used instead of the mean because it is less affected by extreme values and better represents the typical value in each bin.

Same as SupportCalls, outlier values for the Income feature will be detected using the Z-score and using Equation (1). A total of 50 outlier values were identified. These values appeared to be extreme and likely stored in a different unit than the rest of the data. To correct this, the outlier values were divided by 1,000 to match the rest of the data.

The following figure shows the boxplot for the numerical values after resolving the outliers.

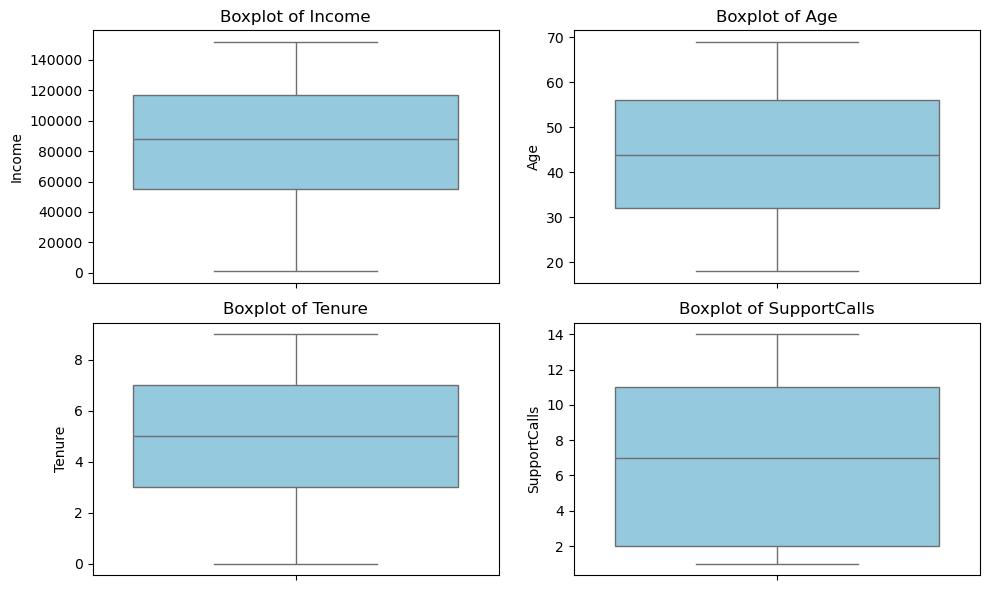


Figure Boxplot for Numerical values after Resolving Outlier Values.

## 4 Feature Scaling

For feature scaling, we chose the Z-score standardization method to scale all numerical features. The Min-Max scaling method is not suitable for our dataset because may not remain bounded over time. In contrast, Z-score standardization centers the data on a mean of 0 with a standard deviation of 1, which preserves the original shape and distribution of each feature. This ensures that all features are equally important for the model training, and helps improve the stability and performance of many machine learning algorithms. The following plots show the numerical features before and after applying Equation (1) to the data.

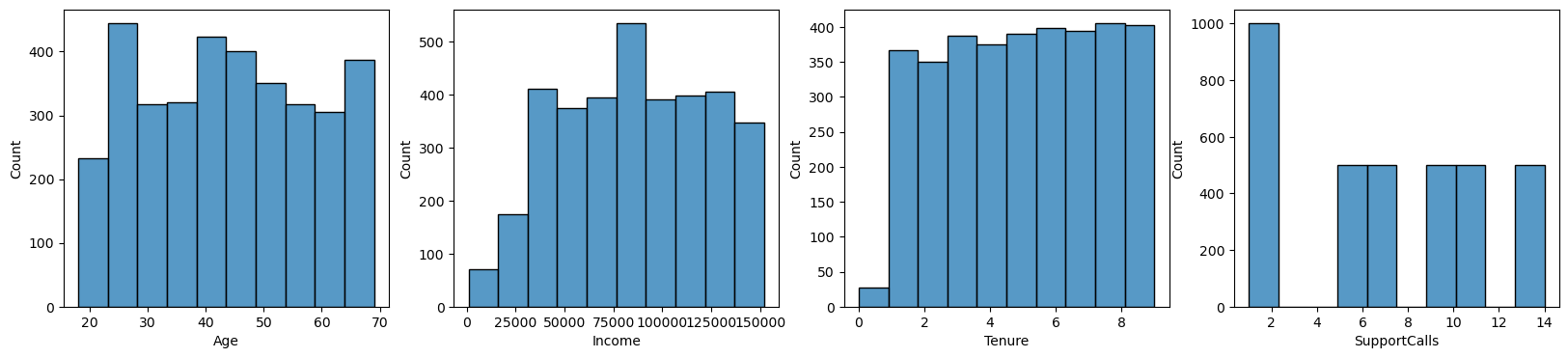


Figure Before Scaling.

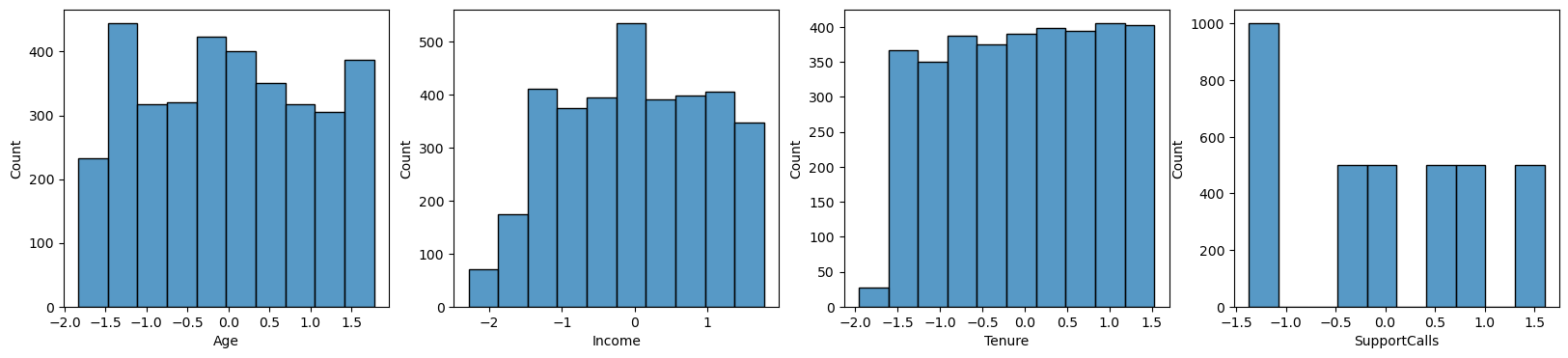


Figure After Scaling.

## 5 Exploring Data Analysis (EDA)

## 5.1 Univariate Analysis

The following figure shows histograms and boxplots for the numerical features.

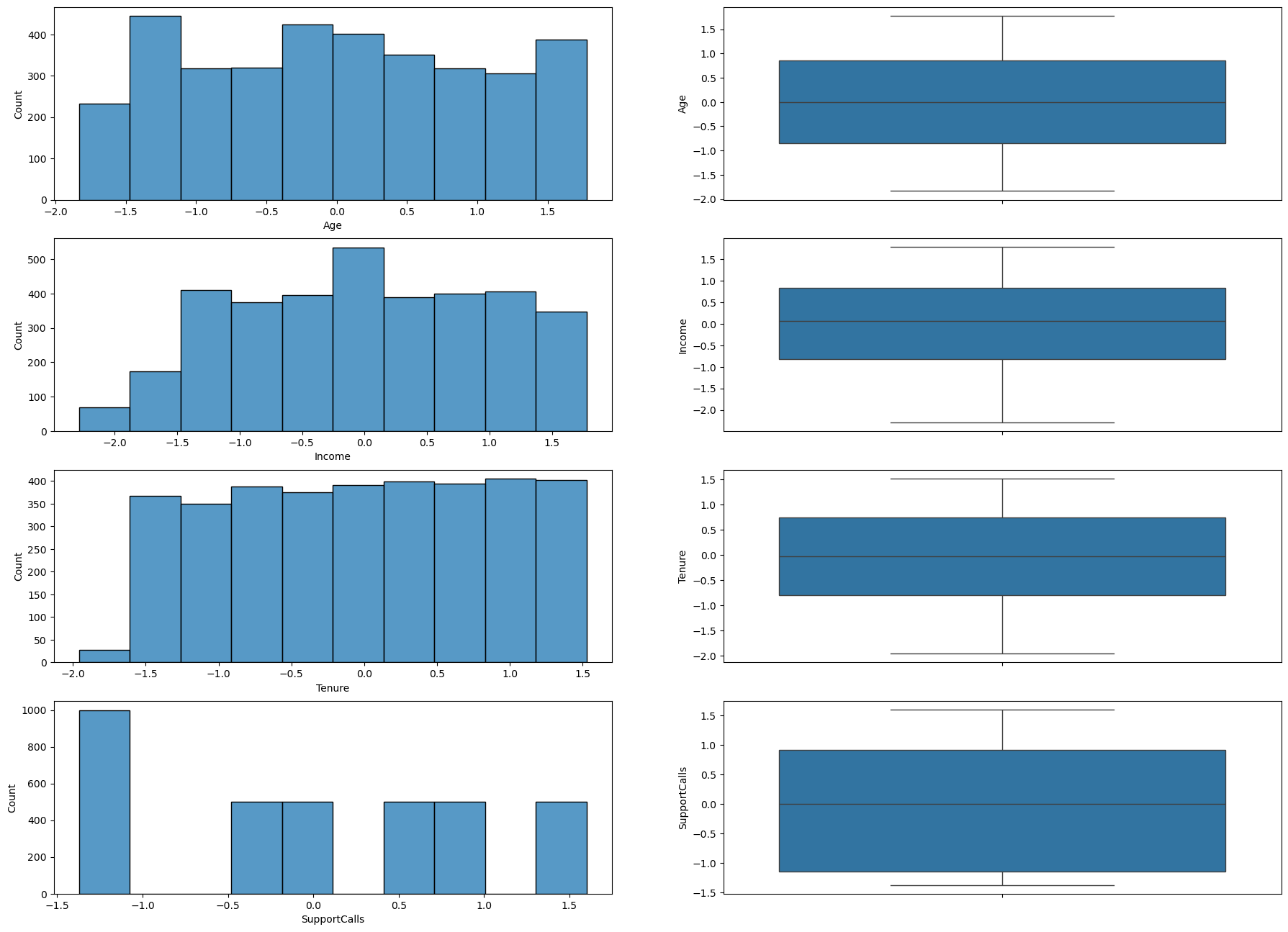


Figure Histograms and Boxplots for Numerical Features.

Based on the plots, we can make the following observations. The Age feature shows that customers are spread uniformly across age distributions with no skew. For Income, after removing outliers, the distribution has become more balanced and easier to interpret. The Tenure feature exhibits values that are evenly distributed across the full range, with no anomalies. Finally, the SupportCalls distribution appears reasonable, with most values concentrated at lower levels after the extreme outliers were removed. The following figure shows the bar plots for the categorical features.

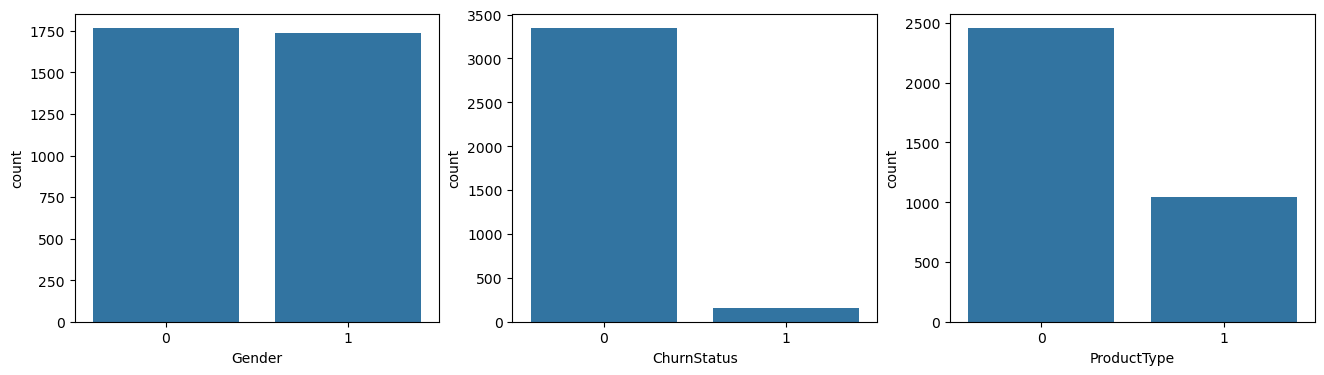


Figure Bar Plots for Categorical Features.

Based on the plots, we can make the following observations. The Gender feature is very balanced, meaning there is no bias in the distribution. In contrast, the Churn Status feature is imbalanced, with relatively few customers churning compared to those who stay. For the Product Type, nearly half of the customers choose the premium product over the basic option.

## 5.2 Bivariate Analysis

The following scatter plot shows the relationship between Age and Income and the points are colored by ChurnStatus.

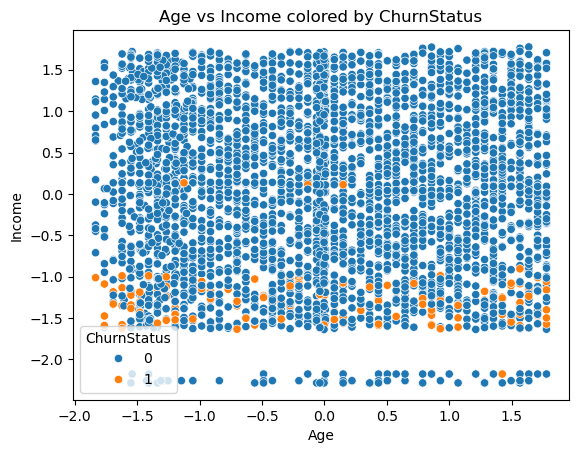


Figure Age vs Income colored by ChurnStatus.

With the outliers already removed, the relationship between the Income and ChurnStatus becomes clear as no customer with an income higher than 50k has churned across all ages. The following plot shows how customers who stayed or churned are distributed across different tenure and income values.

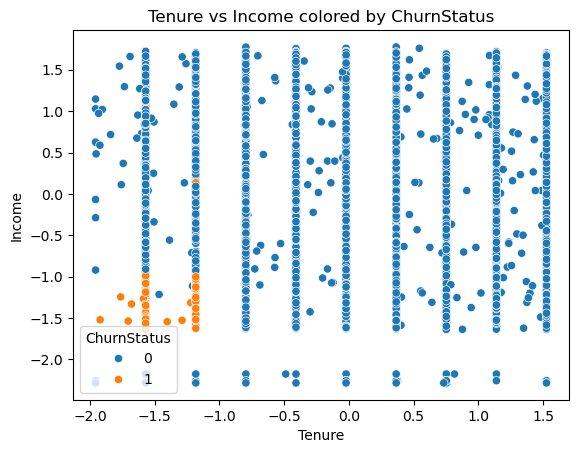


Figure Tenure vs Income colored by ChurnStatus.

The result comes as no surprise as we already learned the relationship between Tenure and Churn, were churned customers having tenure less than 2 years. As well as the relationship between Tenure and Income, were higher customers having higher income. The following plot shows no meaningful pattern between SupportCalls and Income, even after outlier removal, ChurnStatus appears unrelated to SupportCalls.

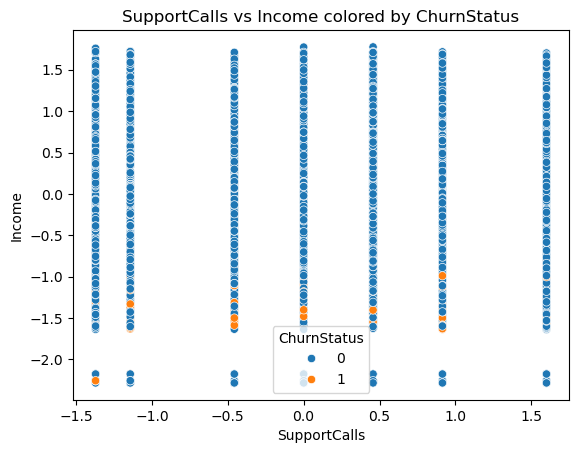


Figure SupportCalls vs Income colored by ChurnStatus.

The following count plots show how Gender and ProductType relate to the ChurnStatus.

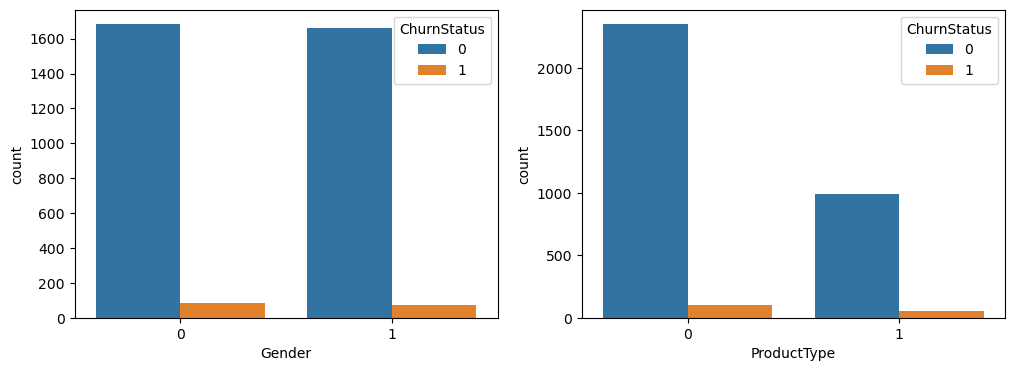


Figure Gender and ProductType relate to ChurnStatus.

Since the churn ratio appears similar for customers who churned or who leaved in both categories, we can conclude that Gender and ProductType play no part in predicting churn status.

## 5.3 Correlation Analysis

The following plot displays the correlation matrix, showing how the numerical features relate to each other as well as to the target variable.

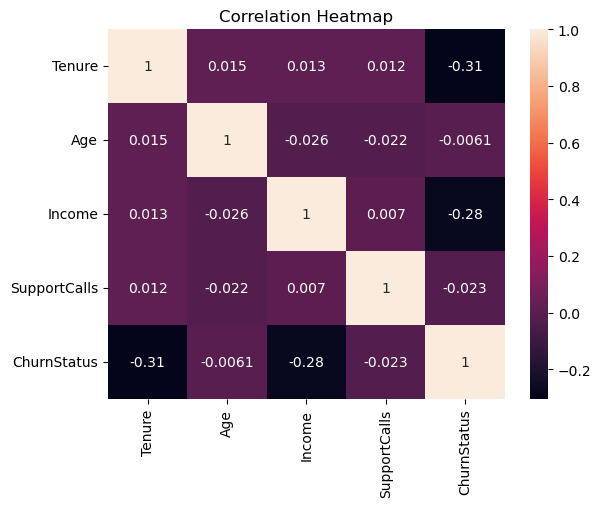


Figure 13 Correlation Matrix.

We can conclude from the matrix that Tenure and Income are the only features with noticeable negative correlation to churn, indicating they are the most relevant predictors for ChurnStatus in the dataset. Other features like Age and SupportCalls are almost independent of churn, since they are more close to zero.­

# 6 Data Visualizations

# 7 Conclusion