

Faculty of Engineering & Technology

Department of Electrical and Computer Engineering

Machine Learning and Data Science

ENCS5341

Assignment 1

Data Preprocessing & Exploratory Data Analysis (EDA)

Prepared by:

Hamed Musleh

ID: 1221036

Mohammad Abu Hijleh

ID: 1221350

Instructor: Dr. Yazan Abu Farha

Section: 1

Date: 23/10/2025

Abstract:

This project focuses on exploring customer data to understand the key factors influencing customer churn. The dataset was first cleaned by handling missing values, detecting and treating outliers, and applying feature scaling to ensure consistency. Exploratory Data Analysis (EDA) was then performed to uncover patterns within both numerical and categorical features. Various visualization techniques, including histograms, boxplots, scatter plots, and heatmaps, were used to illustrate relationships and trends in the data.

The analysis revealed that income and tenure are the most influential factors associated with churn customers with lower income levels and shorter tenure are more likely to leave the company. Product type also showed a slight impact, while age and gender had minimal influence. The correlation analysis supported these findings, highlighting negative relationships between churn and both tenure and income.

Overall, this study provides actionable insights that can help businesses design targeted retention strategies, particularly focusing on new or low-income customers to reduce churn and enhance long-term customer loyalty.

Table of Contents

Abstract:	I
Table of Figure:	III
List of Table:	IV
Introduction:	1
Dataset Overview:	1
Data Set Information:	1
Dataset Statistics:	2
Cleaning Strategy:	3
Documenting Missing Values:	3
Data Pre-processing	4
Handling Missing Values:	4
Outlier Detection and Handling:	4
Standardization:	8
Data Visualization:	8
Univariate Analysis – Distribution of Numerical Features:	8
Distribution of Categorical Features	9
Bivariate Analysis – Numerical Features vs ChurnStatus:	9
Bivariate Analysis – Age vs Income by ChurnStatus:	10
Bivariate Analysis - Churn Rate by Gender:	11
Bivariate Analysis Churn Rate by ProductType:	11
Correlation Analysis – Heatmap of Numerical Features:	12
Conclusion	13

Table of Figure:

Figure 1 : Missing Values	3
Figure 2 : Age Distribution	4
Figure 3 : Income Distribution.....	5
Figure 4 : SupportCalls distribution.....	5
Figure 5 : Tenure Distribution	6
Figure 6: Outlier Before Handling.....	7
Figure 7: Outlier After Handling	7
Figure 8 : Distribution of numerical features (Age, Income, Tenure, and SupportCalls) showing the spread and frequency of customer data after cleaning.	8
Figure 9 : Distribution of categorical features.	9
Figure 10 : Relationship between numerical features and ChurnStatus..	9
Figure 11 : Age vs Income by ChurnStatus.	10
Figure 12 : Average churn rate comparison between genders	11
Figure 13 : Average churn rate by product type	11
Figure 14 : Correlation heatmap of numerical features and ChurnStatus.....	12

List of Table:

Table 1 : Dataset Information - Features 2

Table 2 : Summary Statistics 2

Introduction:

Data preprocessing and exploratory data analysis (EDA) are crucial steps in any data science or machine learning project, ensuring that raw data is clean, consistent, and ready for accurate analysis. Since datasets often contain missing values, outliers, and inconsistencies, preprocessing techniques such as data cleaning, imputation, outlier detection, and feature scaling are applied to enhance data quality and model performance.

In this assignment, a customer dataset is prepared and analyzed to identify factors influencing customer churn. After handling missing values and normalizing data, various visualization techniques including histograms, boxplots, scatter plots, and heatmaps are used to explore distributions, correlations, and relationships between demographic and behavioral variables. This systematic approach highlights how proper preprocessing and EDA enable reliable insights and support data-driven decision-making.

Dataset Overview:

Data Set Information:

The dataset contains **3,500 records** and **8 variables** describing customer demographics, behavior, and churn status. The features include CustomerID, Age, Gender, Income, Tenure, ProductType, SupportCalls, and ChurnStatus. The data consists of both numerical and categorical attributes, with a total of **693 missing values** across different columns. Overall, this dataset provides a balanced foundation for analyzing customer characteristics and identifying key factors related to churn behavior.

Table 1 : Dataset Information - Features

Column Name	Non-Null Count	Data Type
CustomerID	3500	object
Age	3325	float64
Gender	3500	int64
Income	3328	float64
Tenure	3325	float64
ProductType	3500	int64
SupportCalls	3329	float64
ChurnStatus	3500	int64

Dataset Statistics:

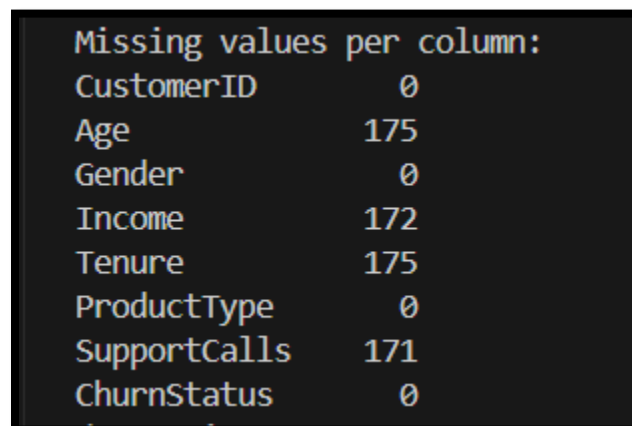
Table 2 : Summary Statistics

Statistic	Age	Gender	Income	Tenure	ProductType	Support Calls	ChurnStatus
count	3325	3500	3328	3325	3500	3329	3500
mean	43.61	0.5	140,686	5.04	0.30	10.08	0.04
std	14.93	0.5	433,327.1	2.57	0.46	21.74	0.21
min	14	0	25,037	0	0	1	0
25%	31	0	56,530.25	3	0	3	0
50% (Median)	43	0	89,532.5	5	0	7	0
75%	56	1	121,502.5	7	1	11	0
max	69	1	5,004,849	9	1	200	1

Cleaning Strategy:

- Missing values were handled through imputation instead of removing rows.
- Age and Tenure were filled with their mean, while Income and SupportCalls used the median.
- Categorical columns were not imputed since mode imputation was not applied.
- Outliers were detected using Z-score (threshold = 3) and handled for Income and SupportCalls using IQR capping/flooring.
- Feature scaling was done using Standardization (Z-score) on numeric features.
- Missing rows were kept to preserve data integrity and maintain dataset size.

Documenting Missing Values:



Missing values per column:	
CustomerID	0
Age	175
Gender	0
Income	172
Tenure	175
ProductType	0
SupportCalls	171
ChurnStatus	0

Figure 1 : Missing Values

The results show that a few numerical features **Age**, **Income**, **Tenure**, and **SupportCalls** contain missing values, each accounting for roughly 5% of the dataset, while the remaining columns are complete with no missing data.

Data Pre-processing

Handling Missing Values:

Missing values were handled carefully to maintain data quality and completeness. Instead of deleting rows, imputation was applied to fill the missing entries. Numerical columns such as Age and Tenure were replaced with their mean values, while Income and SupportCalls were filled using their median values to reduce the impact of outliers. This approach ensured that no important data was lost, keeping the dataset consistent and reliable for further analysis.

Outlier Detection and Handling:

Age Pre-processing

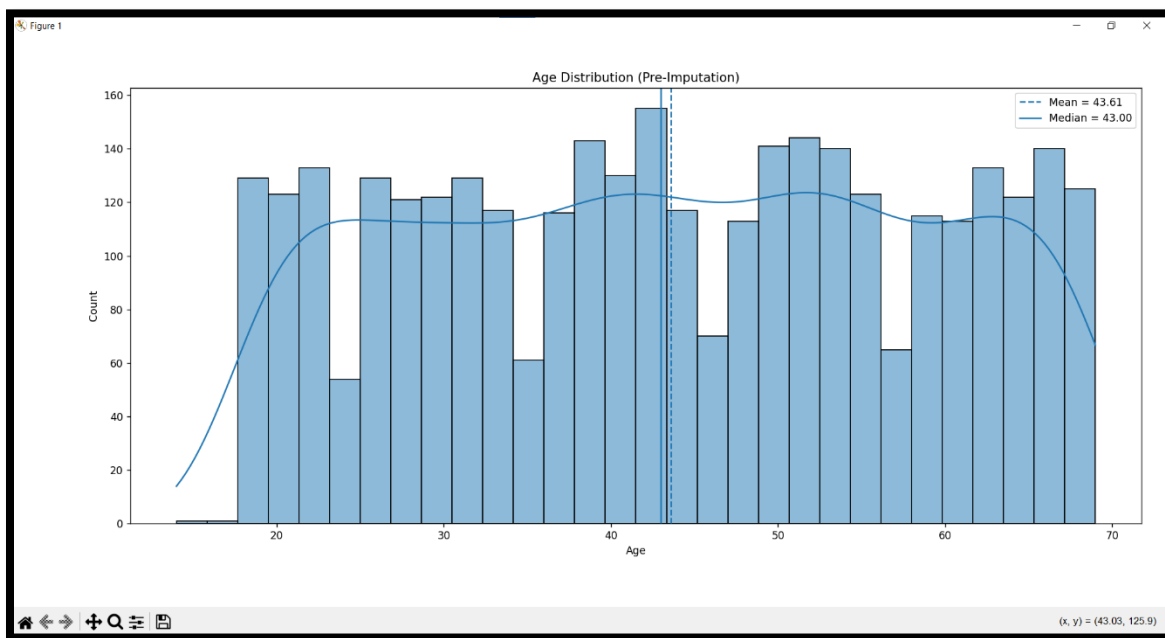


Figure 2 : Age Distribution

This histogram shows the distribution of customer ages before imputation. The **mean** (≈ 43.61) and **median** (≈ 43.00) are almost equal, and the distribution appears roughly **symmetric** with no significant skewness or extreme outliers. This indicates that the **age variable follows an approximately normal distribution**, so using the **mean** for imputing missing values is appropriate and reliable. The mean accurately represents the central tendency of the data without being affected by outliers.

Income Pre-processing

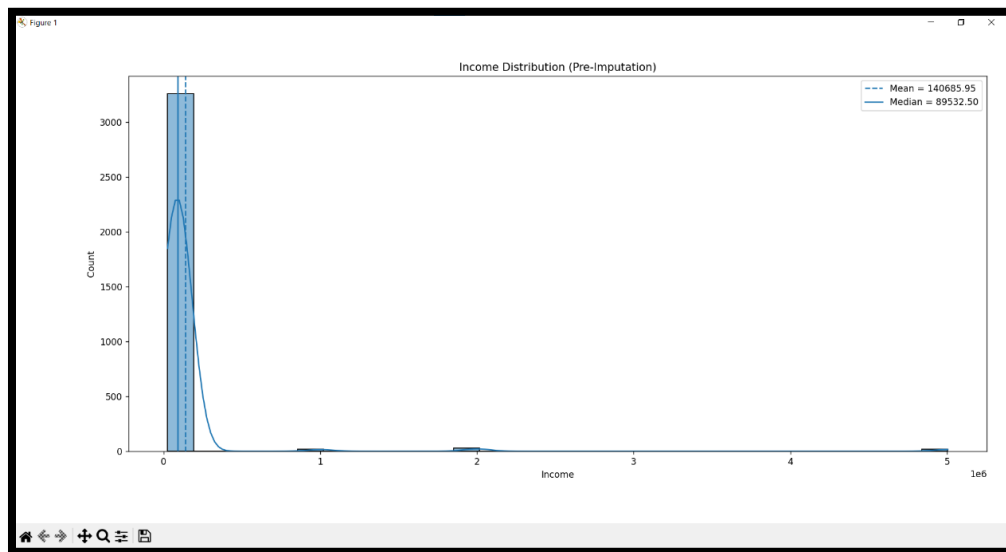


Figure 3 : Income Distribution

The income data is right-skewed, with most values concentrated at lower levels and a few extremely high outliers. The mean ($\approx 140,686$) is pulled to the right by these outliers, while the median ($\approx 89,533$) stays near the center of the main data. Therefore, the median is a better choice for imputation because it is less affected by extreme values and represents the true central tendency more accurately.

SupportCalls Pre-processing

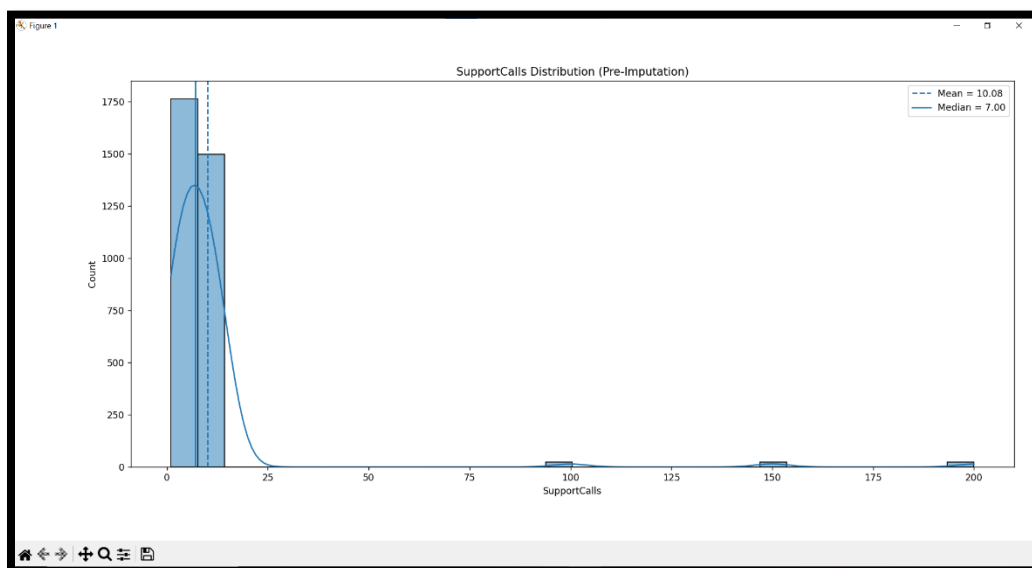


Figure 4 : SupportCalls distribution

The distribution of SupportCalls is **right-skewed**, with most customers making few support calls and a small number making extremely many. The **mean** (≈ 10.08) is pulled to the right by these high outlier values, while the **median** (≈ 7) better represents the typical customer. Thus, the **median is preferred for imputation**, as it is more robust to outliers and reflects the central tendency of the data more accurately.

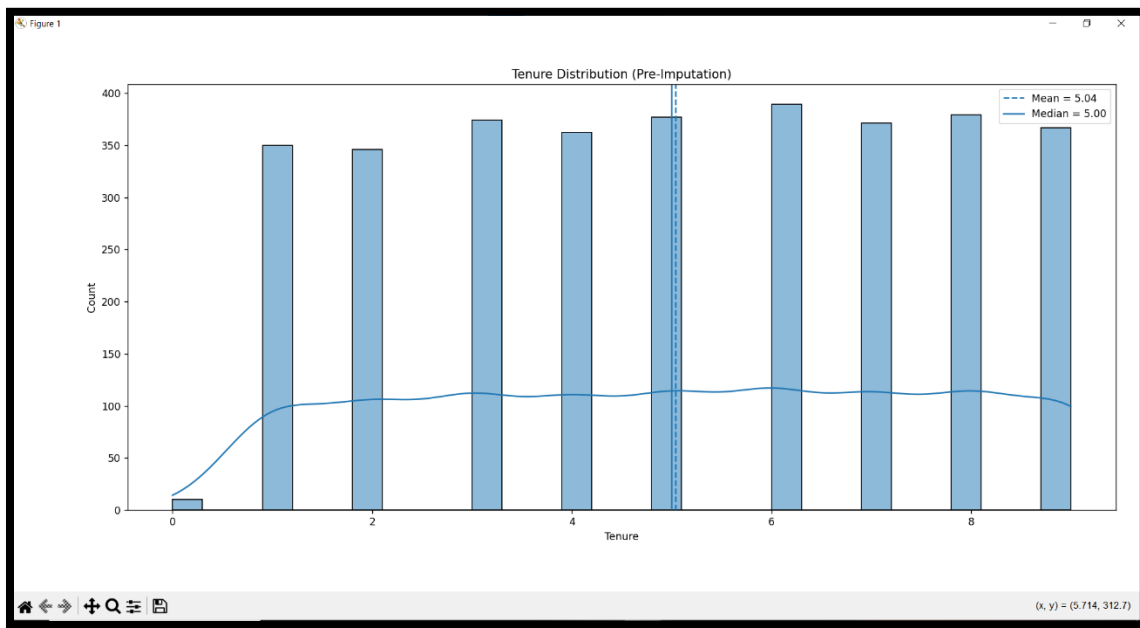


Figure 5 : Tenure Distribution

This histogram shows the distribution of customer tenure before imputation. The **mean** (≈ 5.04) and **median** (≈ 5.00) are almost identical, and the data appears **fairly symmetric** without any significant skewness or outliers. This indicates that the **Tenure** variable follows a nearly normal or uniform distribution, where the mean accurately represents the central tendency. Therefore, using the **mean for imputation** is appropriate and ensures consistency with the data's natural distribution.

The boxplots below show the data **before and after handling outliers**.

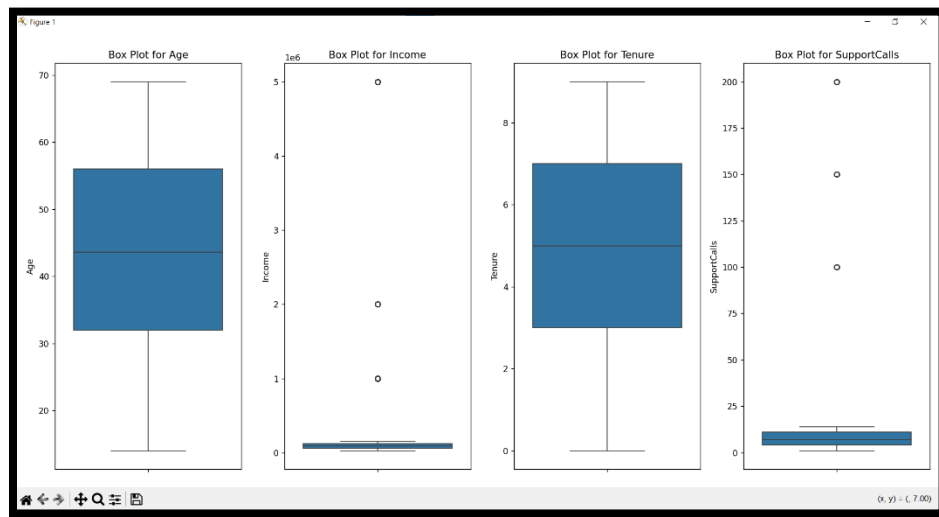


Figure 6: Outlier Before Handling

Before treatment, several extreme values were visible in **Income** and **SupportCalls**, as shown by the points outside the whiskers in the first figure. These outliers could distort statistical analysis and affect the accuracy of future models.

To address this, **IQR-based capping and flooring** was applied values below the lower bound were replaced with the lower limit, and those above the upper bound were capped. This method preserved the overall data structure while reducing the influence of extreme values.

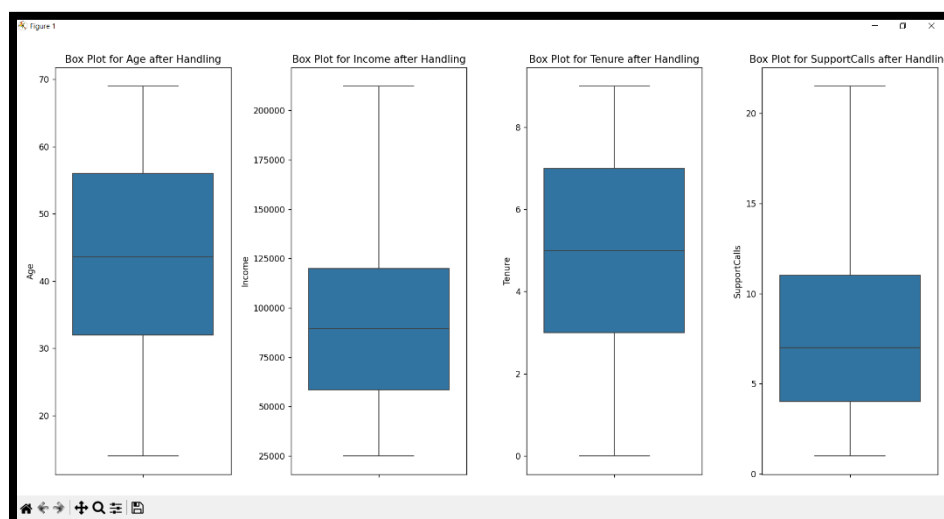


Figure 7: Outlier After Handling

After handling, the second figure shows that the distributions of **Income** and **SupportCalls** became more compact and realistic, with the outliers effectively removed while maintaining the integrity of the dataset.

Standardization:

Feature scaling (standardization) was applied to ensure that all numerical features contribute equally during analysis and future modeling. Since the dataset contains variables measured on different scales such as Age, Income, Tenure, and SupportCalls standardization helps eliminate bias caused by scale differences.

The Z-score standardization method was used, transforming each feature so that it has a mean of 0 and a standard deviation of 1. This approach is particularly useful for algorithms sensitive to feature magnitude, such as distance-based models. After scaling, the numerical features are on a uniform scale, making comparisons and statistical interpretations more accurate and reliable.

Data Visualization:

Univariate Analysis – Distribution of Numerical Features:

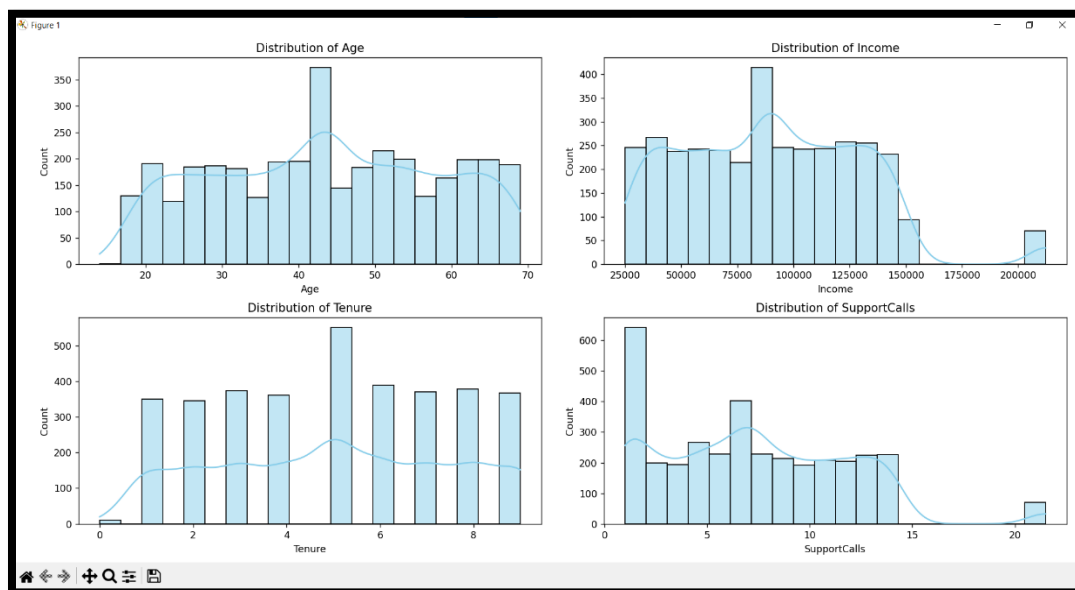


Figure 8 : Distribution of numerical features (Age, Income, Tenure, and SupportCalls) showing the spread and frequency of customer data after cleaning.

The histograms above display the distribution of numerical features in the dataset.

- **Age** and **Income** show relatively even distributions with no strong skewness after cleaning.
- **Tenure** is fairly uniform, indicating customers are spread across different durations.
- **SupportCalls** is slightly right-skewed, showing that most customers made few calls, while a smaller number made many.

Distribution of Categorical Features

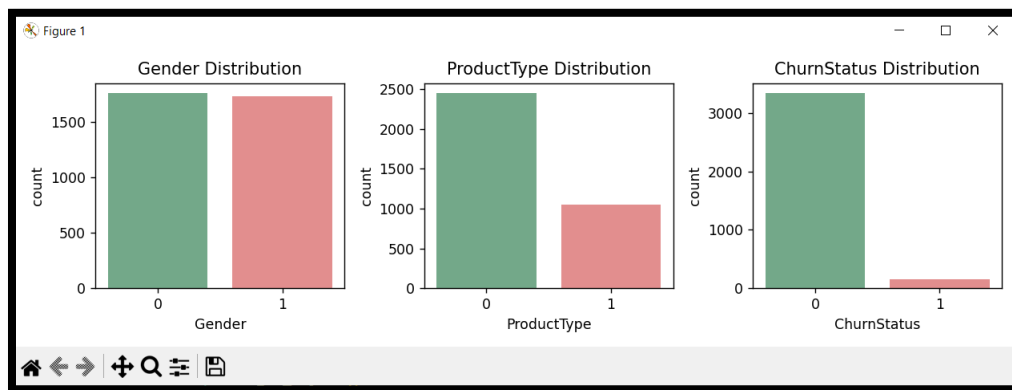


Figure 9 : Distribution of categorical features.

The bar charts display the frequency of each category. Gender is nearly balanced, most customers belong to one main product type, and the dataset shows a strong imbalance in **ChurnStatus**, with far more non-churned customers.

Bivariate Analysis – Numerical Features vs ChurnStatus:

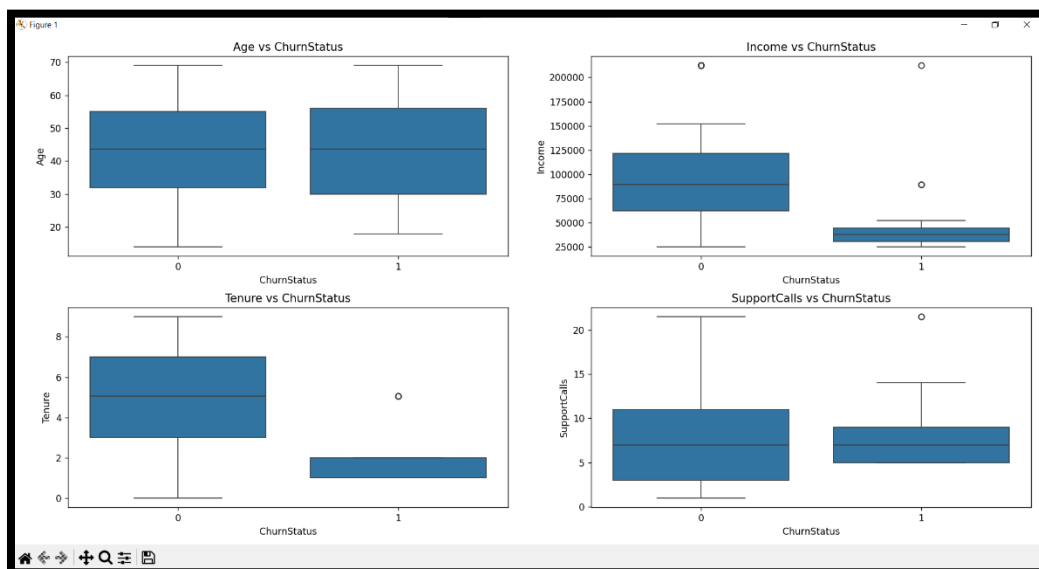


Figure 10 : Relationship between numerical features and ChurnStatus..

The boxplots compare customer attributes across churned (1) and non-churned (0) groups. Churned customers generally show **lower income and tenure**, while **SupportCalls** tend to be slightly higher among them, indicating possible dissatisfaction.

Bivariate Analysis – Age vs Income by ChurnStatus:

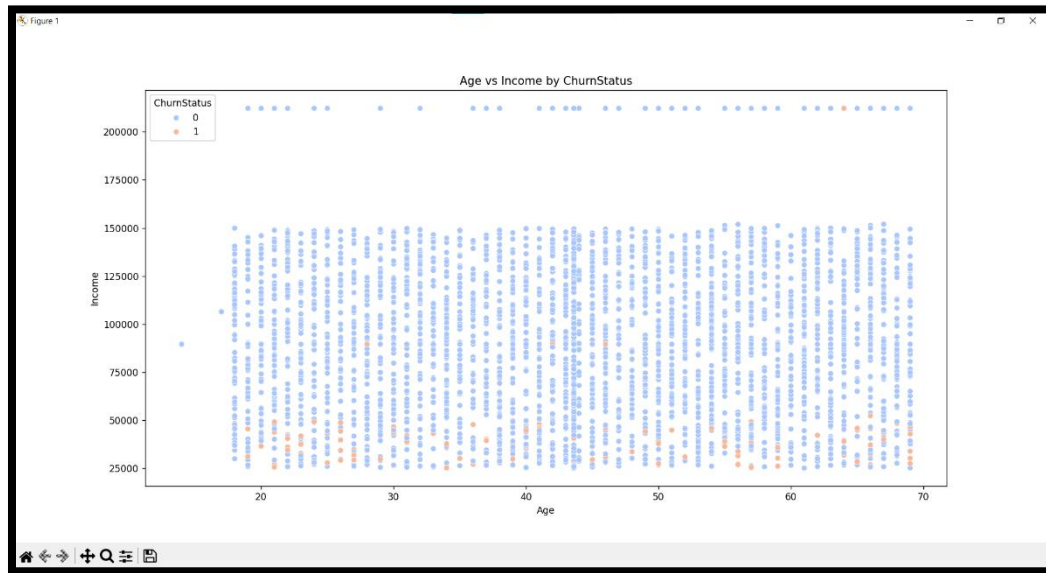


Figure 11 : Age vs Income by ChurnStatus.

The scatter plot shows the relationship between **Age** and **Income**, colored by churn status. Churned customers (in orange) appear more frequently among **lower-income groups**, while **non-churned** customers are spread across all income levels.

Bivariate Analysis - Churn Rate by Gender:

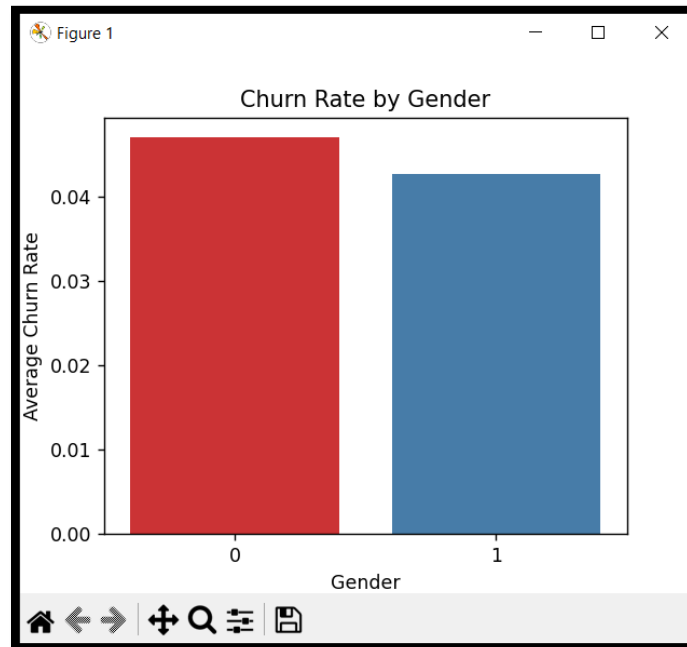


Figure 12 : Average churn rate comparison between genders

The chart shows a slightly higher churn rate for **Gender = 0**, indicating a small difference in customer retention between the two groups.

Bivariate Analysis Churn Rate by ProductType:

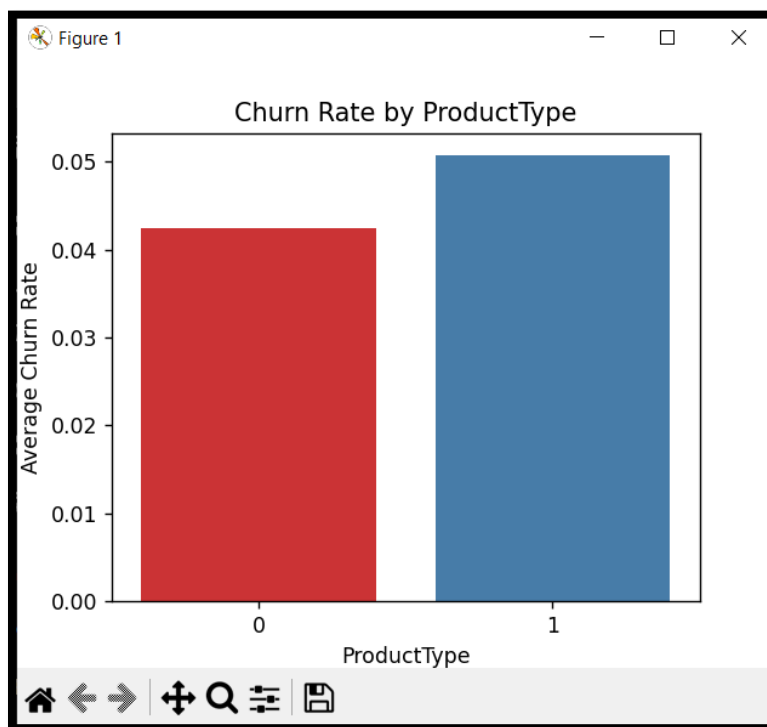


Figure 13 : Average churn rate by product type

Customers with **ProductType = 1** exhibit a higher churn rate, suggesting that this product category may be more associated with customer loss.

Correlation Analysis – Heatmap of Numerical Features:

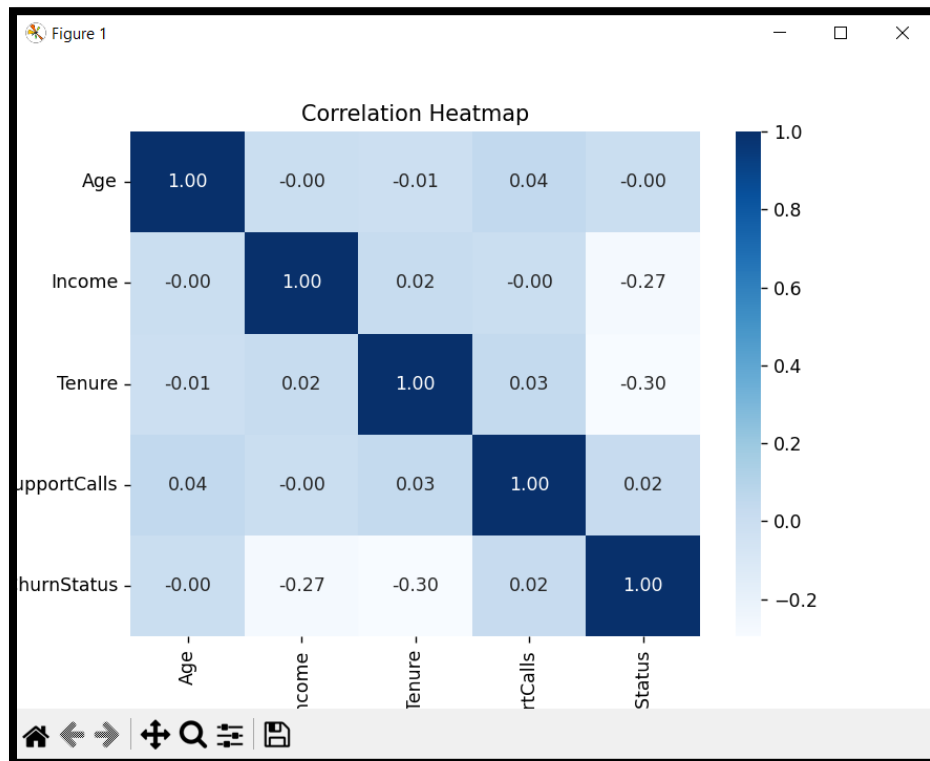


Figure 14 : Correlation heatmap of numerical features and ChurnStatus.

The heatmap indicates weak correlations between most variables. A slight negative correlation exists between **Income**, **Tenure**, and **ChurnStatus**, suggesting that customers with higher income or longer tenure are less likely to churn.

Conclusion

This assignment demonstrated the importance of data preprocessing and exploratory data analysis (EDA) in understanding customer behavior and identifying key drivers of churn. Through systematic data cleaning, missing value imputation, outlier treatment, and feature standardization, the dataset was transformed into a consistent and reliable form suitable for analysis.

The EDA revealed that **income** and **tenure** are the most influential factors affecting customer churn. Customers with lower income levels and shorter tenure are more likely to leave the company. Additionally, **product type** showed a moderate effect on churn, while **age** and **gender** had minimal impact. Correlation analysis supported these insights, showing weak but meaningful negative correlations between churn, income, and tenure.

Overall, the findings suggest that businesses can improve customer retention by focusing on **new and low-income customers**, offering incentives, or personalized services to strengthen loyalty. This analysis highlights how effective data preprocessing and visualization techniques can uncover actionable insights, providing a foundation for predictive modeling and data-driven decision-making in future work.