Faculty of Engineering & Technology
Electrical & Computer Engineering Department

Machine Learning and Data Science – (ENCS5341)

Assignment 1

Data Preprocessing & Exploratory Data Analysis (EDA)

Prepared by:

Aws Hammad Id : 1221697

Khaled Abu Lebdeh Id : 1220187

Instructor: Dr. Yazan Abu Farha

Section: 2

Date: October 20, 2025

Table of Contents

1. **Data Loading and Preprocessing** 1
   - 1.1 **Data Loading and Initial Inspection** 1
   - 1.2 **Missing Data Handling** 2
   - 1.3 **Outliers Handling** 3
   - 1.4 **Feature Scaling** 4
   - 1.5 **Exploratory Data Analysis (EDA)** 4
     - **A. Univariate Analysis** 4
     - **B. Bivariate Analysis** 8
     - **C. Correlation Analysis** 10
   - 1.6 **Data Visualizations:** 10
2. **Conclusion** 14

List of Figures

Figure 1: Histogram showing the distribution of customer income 5
Figure 2: Histogram displaying the distribution of customer tenure 5
Figure 3: Histogram of the number of support calls made per customer 6
Figure 4: Histogram of customer age values 6
Figure 5: Bar chart showing counts of male (0) and female (1) customers 7
Figure 6: Bar chart comparing counts of Basic (0) vs Premium (1) subscriptions 7
Figure 7: Scatter plot showing the relationship between age and churn 8
Figure 8: Scatter plot of income levels versus churn status 8
Figure 9: Scatter plot showing tenure against churn status 9
Figure 10: Scatter plot showing how number of support calls relates to churn 9
Figure 11: Heatmap of correlations among numerical features and ChurnStatus 10
Figure 12: Pie chart displaying the proportion of churned vs retained customers 11
Figure 13: Heatmap showing churn rates across combined income and tenure quartiles 11
Figure 14: Heatmap showing churn rate variations by support-call frequency and income level 12
Figure 15: Radar plot comparing mean Income, Tenure, and SupportCalls for stayed vs churned customers 12
Figure 16: Pair plot illustrating joint distributions and pairwise relationships between key variables colored by churn status. 13

List of Tables

Table 1-1: Description of Variables in the Customer Dataset 1

1. Data Loading and Preprocessing

1.1 Data Loading and Initial Inspection

The dataset used in this analysis was imported using the pandas library in Python. It contains customer information related to churn behavior. Each row represents one customer, and each column corresponds to an attribute describing that customer (such as demographic or service-related information).

After loading the dataset, an initial examination was carried out to understand its overall structure, size, and key features. The dataset consisted of approximately 3,500 records and 8 attributes, each describing various aspects of customer profiles and behavior. A summary of the variables is presented in Table 1, outlining their type and description.

Feature	Description	Type
Age	Customer's age	Numeric
Income	Annual income value (highly variable and right-skewed)	Numeric
Tenure	Duration of customer relationship with the company	Numeric
SupportCalls	Number of customer-support interactions	Numeric
Gender	Customer gender	Categorical
ProductType	Type of product or service subscribed	Categorical
ChurnStatus	Binary output variable (0 = Stayed, 1 = Churned)	Binary

Table 1-1: Description of Variables in the Customer Dataset

- An initial inspection using **df.info()** and **df.describe()** confirmed that:
    - Most variables were **numeric**, suitable for correlation and visualization.
    - The target variable (ChurnStatus) was **binary**, encoded as 0 and 1.
    - Some columns (i.e. *Income* and *SupportCalls*) showed wide ranges, and high standard deviation, hinting at the presence of **outliers** to be handled in later preprocessing steps.

This step ensured that the dataset was successfully loaded, correctly structured, and ready for cleaning and transformation in the subsequent sections.

1.2 Missing Data Handling

An initial inspection of the dataset revealed that several numerical attributes contained missing entries. Specifically, 175 values were missing in Age, 172 in Income, 175 in Tenure, and 171 in SupportCalls. In contrast, all categorical variables (Gender and ProductType) and the binary target variable (ChurnStatus) were complete, with no missing data.

To maintain the overall integrity of the dataset and avoid a substantial reduction in sample size, the missing values were handled through **imputation** rather than deletion. Because the missing cells were distributed across different records, meaning that removing all rows containing at least one missing value would have eliminated roughly 700 rows (about 20 % of the data). Therefore, imputation was the most appropriate strategy.

Considering the distribution characteristics of each variable, a combination of mean and median imputation was applied:

- For approximately symmetric attributes (i.e. Age and Tenure), missing values were replaced with the **mean** of the respective column.
- For skewed attributes (i.e. Income and SupportCalls), the **median** was used to minimize the impact of extreme values and preserve the central tendency of the data.

After performing these imputations, the dataset was rechecked using **df.isnull().sum()**, confirming that all missing values had been successfully addressed. Thus, the dataset was made both **complete** and **consistent**, ensuring that no valuable information was lost during preprocessing. This step improved the overall data quality and reliability, providing a solid foundation for subsequent tasks such as outlier detection, scaling, and exploratory data analysis.

1.3 Outliers Handling

Outlier detection was performed to identify extreme numerical values that could distort the overall data distribution and affect the accuracy of subsequent analysis. Boxplots and descriptive statistics revealed clear outliers in the features Income and SupportCalls. Both variables displayed right-skewed distributions, with a few extremely high values (for instance, Income reaching several million and SupportCalls exceeding 200).

To address these, the Interquartile Range (IQR) method was first applied to both variables. The IQR was calculated as the difference between the third quartile (Q3) and the first quartile (Q1). Outlier thresholds were defined using the standard rule:

- Lower Bound $= Q1 - 1.5 \times IQR$
- Upper Bound $= Q3 + 1.5 \times IQR$

Observations outside these limits were considered outliers and removed. This step produced clear improvements in both features, with more compact and balanced distributions while preserving the majority of the dataset.

To enhance understanding and further validate the approach, an alternative method based on the mean and standard deviation was tested specifically for the Income variable. The lower and upper bounds were defined as:

- Lower Bound $=$ mean $- 2 \times$ std
- Upper Bound $=$ mean $+ 2 \times$ std

This standard-deviation (Z-score equivalent) method produced similar results and confirmed that the previously detected outliers were indeed extreme values rather than legitimate observations. The final dataset used for subsequent analysis retained the results obtained from this method, as it provided a consistent and statistically sound cutoff for Income while maintaining data integrity.

In summary, both the IQR and standard-deviation approaches effectively identified and removed outliers. The comparison validated that the selected thresholds produced stable,

representative distributions for both Income and SupportCalls, ensuring reliable input for later scaling and modeling steps.

1.4 Feature Scaling

After handling outliers, the numerical attributes were normalized to ensure that all features contributed equally to subsequent analyses and modeling. Because the variables (Age, Tenure, Income, and SupportCalls) had different ranges and units, direct comparison could bias distance-based or gradient-based models.

To standardize the data, Z-score normalization (Standardization) was applied using the formula:

$$Z(scaled\ value) = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

This process rescales each feature to have a mean of 0 and a standard deviation of 1, effectively removing the influence of scale differences among variables.

Z-score normalization was selected over Min–Max scaling because, after outlier removal, the dataset was approximately normally distributed, making standardization the most suitable approach. It preserves the relative variability of each attribute while allowing all numerical features (Age, Tenure, Income, and SupportCalls) to be compared on a common scale.

As a result, the standardized dataset became balanced and consistent, providing a reliable basis for subsequent correlation analysis, visualization, and predictive modeling.

1.5 Exploratory Data Analysis (EDA)

A. Univariate Analysis

To understand the general characteristics of the dataset, the distribution of numerical and categorical variables was explored.

1. Income Distribution:

The histogram of *Income* shows a roughly uniform distribution, indicating that customer

income values are spread relatively evenly across the range. There is a noticeable peak near the center, suggesting a slightly higher frequency of average-income customers.

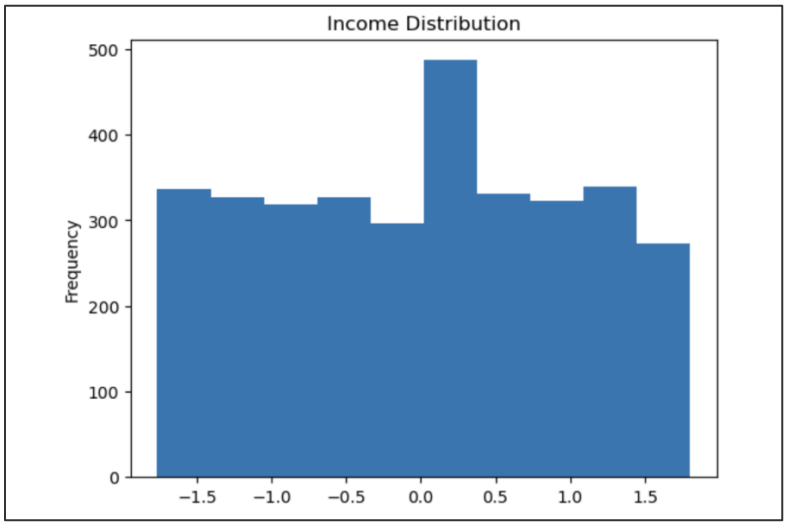

Figure 1: Histogram showing the distribution of customer income

2. Tenure Distribution:

The *Tenure* variable also follows an approximately uniform pattern with a moderate concentration around the center. This suggests that the company has a balanced mix of both new and long-term customers, with no strong skew toward either extreme.

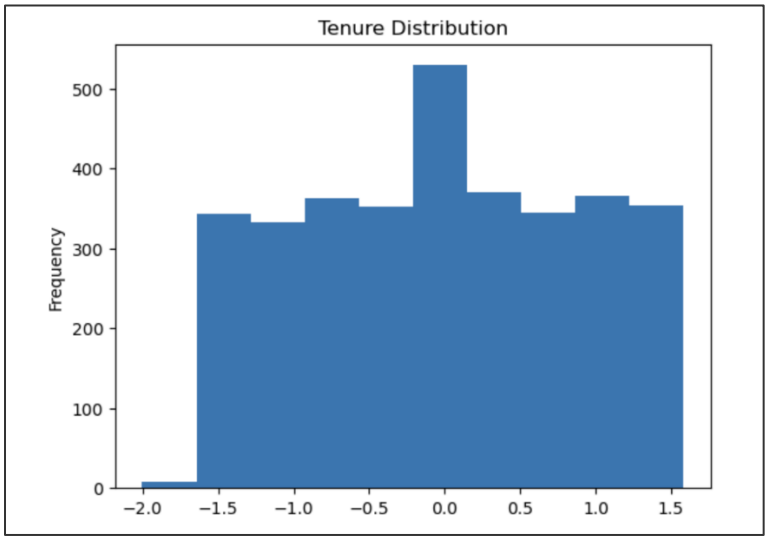

Figure 2: Histogram displaying the distribution of customer tenure

3. Support Calls Distribution:

The *SupportCalls* histogram exhibits visible variation across intervals, with certain bins showing higher frequencies. This implies that while many customers make few support calls, a smaller portion contact support frequently — potentially signaling dissatisfaction among those users.

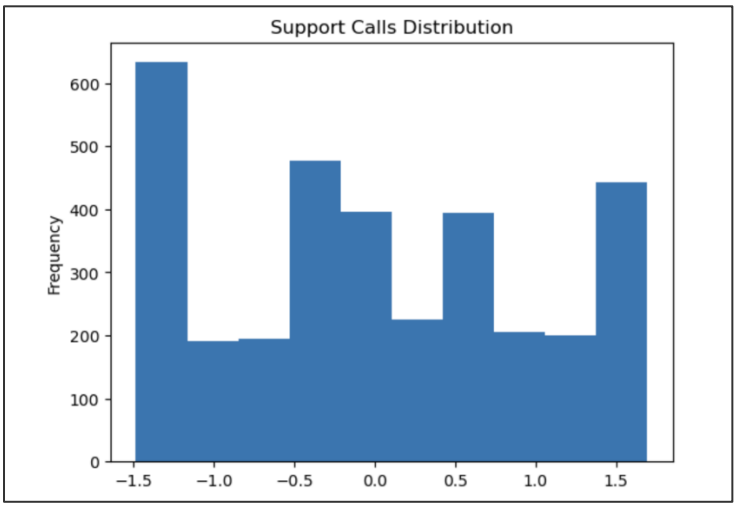

Figure 3: Histogram of the number of support calls made per customer

4. Age Distribution:

The *Age* variable is moderately symmetric with slight central clustering. This indicates that most customers are of middle age, with fewer younger or older individuals represented.

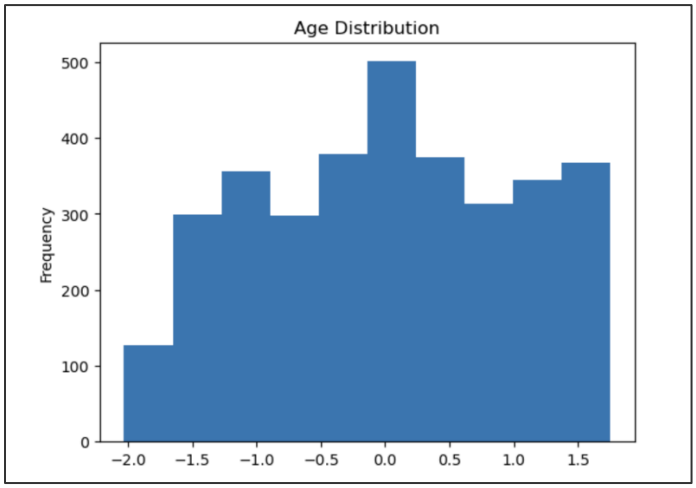

Figure 4: Histogram of customer age values

5. Gender Distribution:

The bar chart shows that both gender categories (0 = Male, 1 = Female) are almost equally represented, confirming that the dataset is gender-balanced.

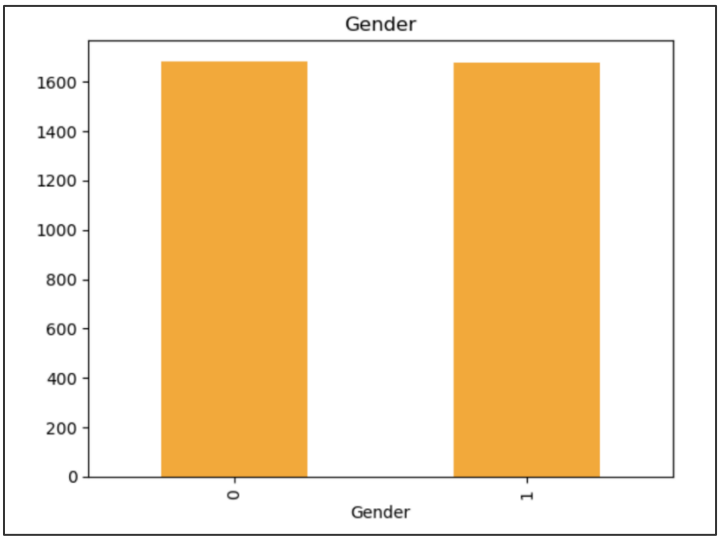

Figure 5: Bar chart showing counts of male (0) and female (1) customers

6. Product Type Distribution:

The *ProductType* bar chart reveals that the *Basic* plan (coded 0) has significantly more customers than the *Premium* plan (coded 1). This suggests that the majority of customers prefer the basic service offering.

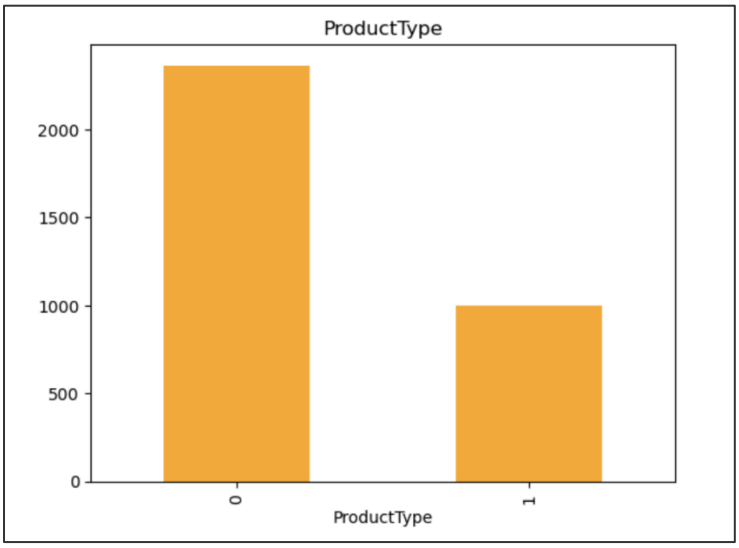

Figure 6: Bar chart comparing counts of Basic (0) vs Premium (1) subscriptions

B. Bivariate Analysis

This step examines how different variables relate to customer churn (ChurnStatus).

1. Age vs. ChurnStatus:

The scatter plot indicates no clear pattern between *Age* and *ChurnStatus*. Customers across all age groups appear in both churned and non-churned categories, implying age alone is not a strong predictor of churn.

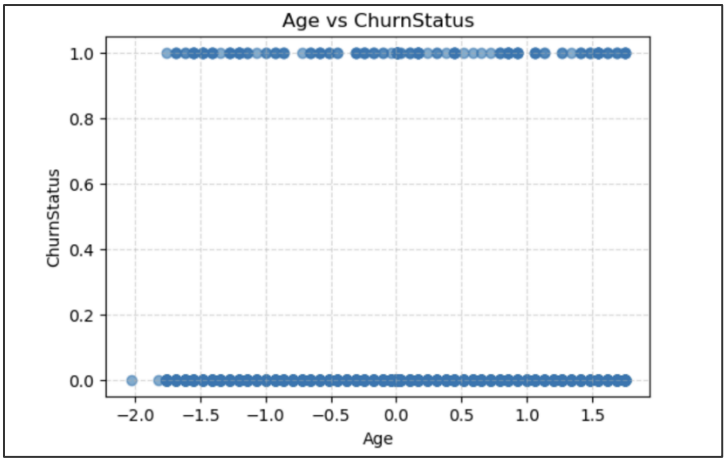

Figure 7: Scatter plot showing the relationship between age and churn

2. Income vs. ChurnStatus:

Similar to age, *Income* shows limited separation between churned and retained customers. Both low- and high-income customers can churn, suggesting income may not directly influence churn probability.

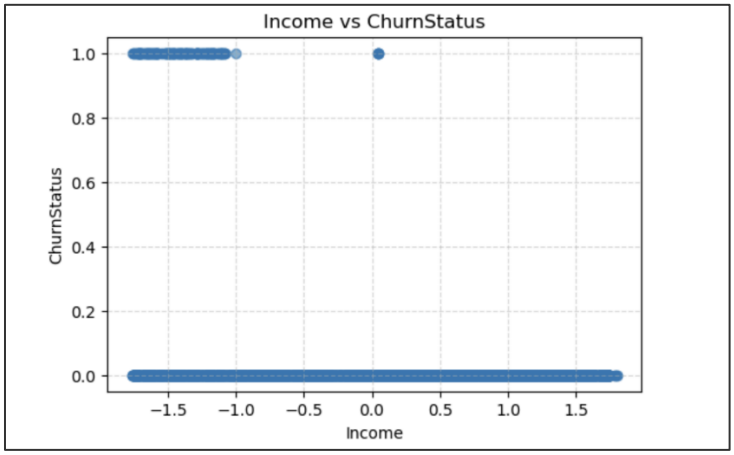

Figure 8: Scatter plot of income levels versus churn status

3. Tenure vs. ChurnStatus:

The *Tenure* scatter shows that churn occurs across different tenure values, but customers with very short tenure appear slightly more likely to churn. This aligns with typical customer-lifecycle behavior, where early disengagement is more common.

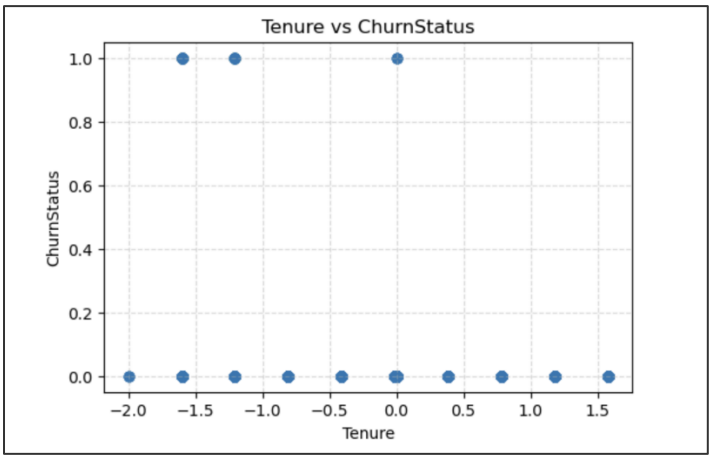

Figure 9: Scatter plot showing tenure against churn status

4. Support Calls vs. ChurnStatus:

A mild upward pattern is visible — customers making more support calls tend to have higher churn likelihood. This relationship suggests that frequent service issues or dissatisfaction could contribute to attrition.

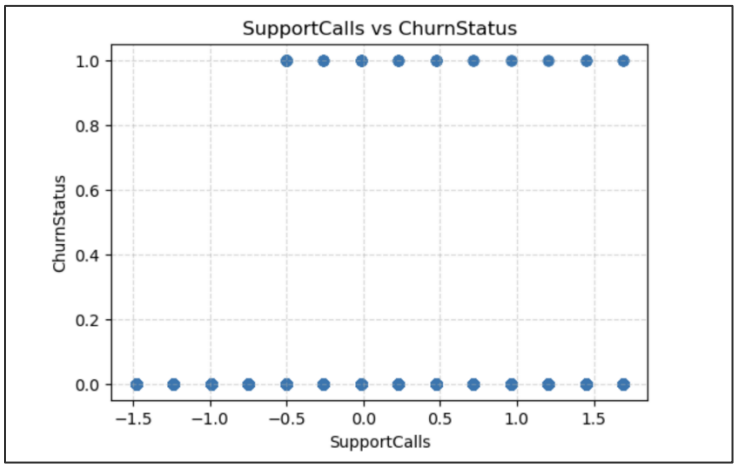

Figure 10: Scatter plot showing how number of support calls relates to churn

C. Correlation Analysis

The correlation heatmap (shown above) provides a compact summary of the relationships between all numerical variables (*Age, Income, Tenure, SupportCalls,* and *ChurnStatus*).

- The diagonal elements indicate perfect self-correlation (1.0).
- Inter-feature correlations among *Age, Income,* and *Tenure* are low, confirming independence between these demographic factors.
- *SupportCalls* shows a slight positive correlation with *ChurnStatus,* supporting the earlier finding that frequent support interactions may signal potential churn.
- Other variables display weak correlations with churn, suggesting that no single attribute drives customer loss on its own.

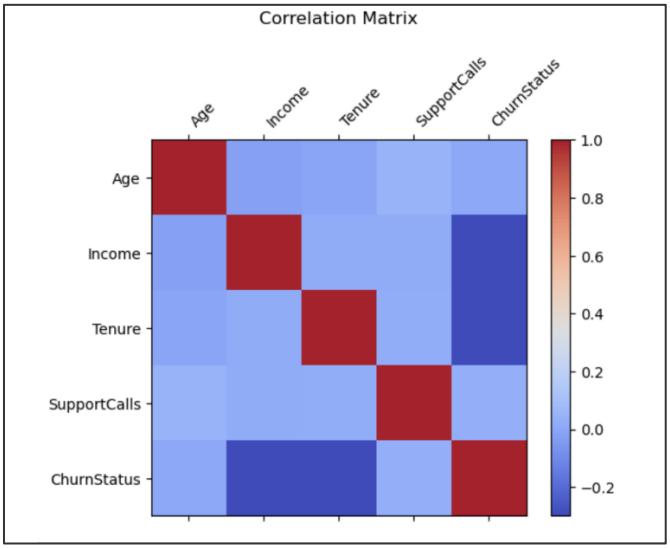

Figure 11: Heatmap of correlations among numerical features and ChurnStatus

1.6 Data Visualizations:

This section presents five concise visualizations that communicate the key insights uncovered during EDA. The plots are well-labeled and easy to interpret, satisfying the assignment requirement to include at least four insight-driven visuals.

1. Output Distribution — Customer Churn:

The pie chart shows a strong class imbalance: ~95.5% of customers stayed (label 0) versus ~4.5% who churned (label 1). This imbalance has important modeling

implications (e.g., using stratified splits, class weights, or sampling methods) to avoid bias toward the majority class.

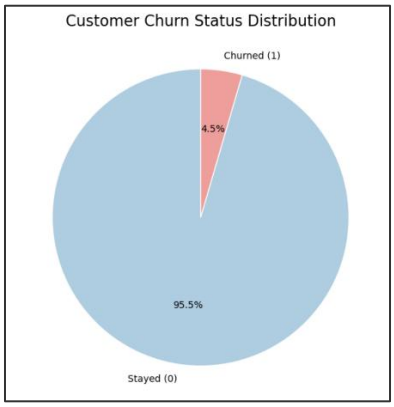

Figure 12: Pie chart displaying the proportion of churned vs retained customers

2. Churn Rates by Key Feature Interactions (Income × Tenure):

The heatmap highlights localized pockets of elevated churn among customers with lower income and lower tenure. While overall churn is low, bins at the Low-Income / Low-Tenure corner exhibit notably higher churn rates than other cells. Because these are quantile bins, note that small cell counts can inflate rates, so this pattern should be validated with raw counts alongside the rates.

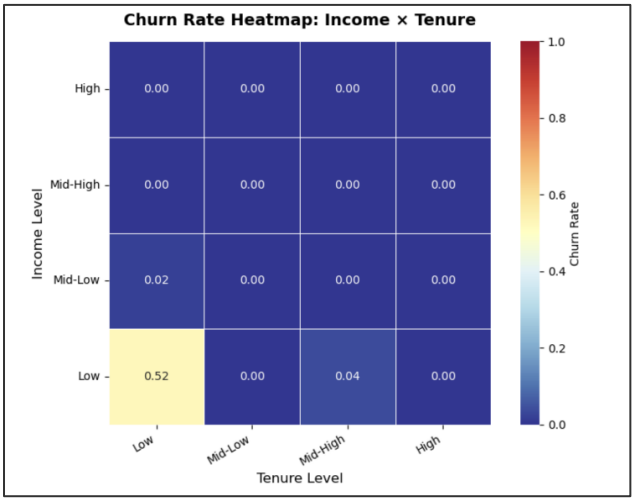

Figure 13: Heatmap showing churn rates across combined income and tenure quartiles

3. Support Calls Distribution (SupportCalls × Income):

Churn rises with more support calls, especially within the lower-income bands. This

aligns with the earlier bivariate results and correlation analysis: heavy support interaction is a practical signal of dissatisfaction and a leading indicator of churn risk.

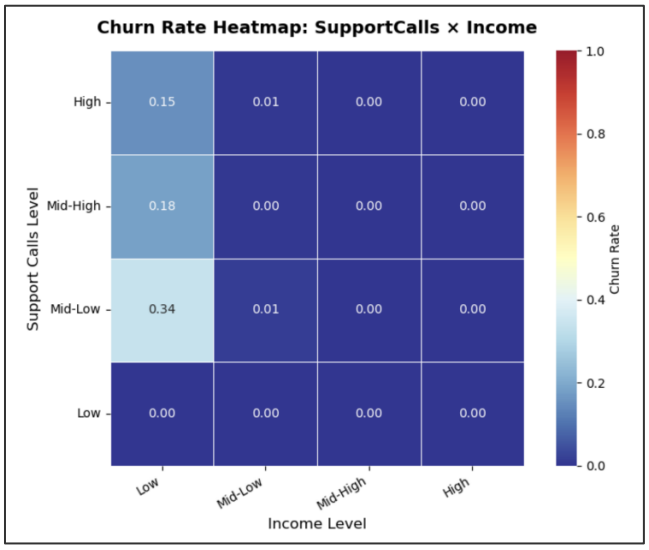

Figure 14: Heatmap showing churn rate variations by support-call frequency and income level

4. Average Feature Levels by Churn Segment:

The radar chart compares mean feature levels for **Stayed (0)** vs **Churned (1)** customers:

- **Churned** customers show **lower average income** and **tenure**, and **higher support calls**, relative to those who **stayed**.
- The shapes for the two segments diverge most strongly on **SupportCalls**, echoing the heatmaps.

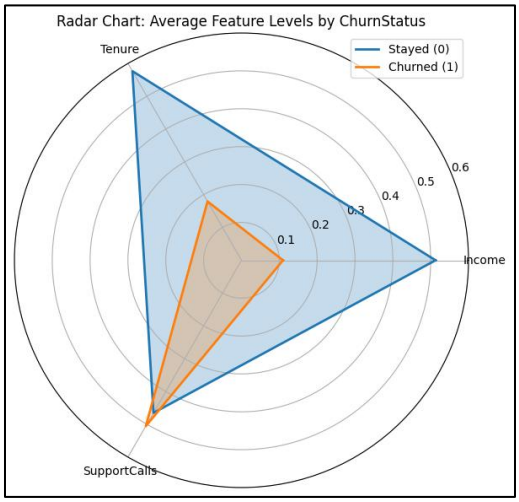

Figure 15: Radar plot comparing mean Income, Tenure, and SupportCalls for stayed vs churned customers

5. Average Feature Levels by Churn Segment:

The pair plot visually corroborates the earlier findings:

- Churn points (label 1) concentrate in regions with **lower income**, **shorter tenure**, and **more support calls**.
- Diagonal density plots show noticeably different distributions for churned vs stayed on **SupportCalls**, and more subtle differences on **Income** and **Tenure**.

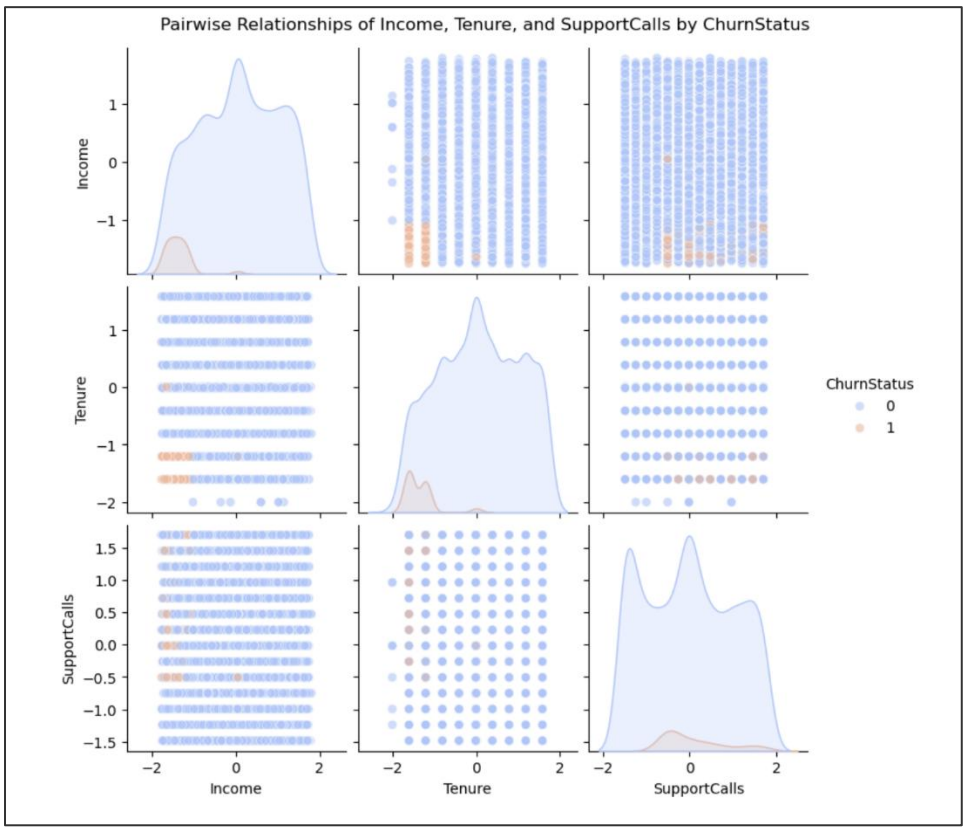

Figure 16: Pair plot illustrating joint distributions and pairwise relationships between key variables colored by churn status.

2. Conclusion

The exploratory data analysis conducted in this assignment provided a comprehensive understanding of customer characteristics, behavioral patterns, and potential churn drivers within the dataset. The preprocessing stages—covering missing-value imputation, outlier treatment, and feature scaling—ensured that the data was both clean and standardized, forming a robust foundation for meaningful exploration.

Through univariate and bivariate analyses, it became clear that most numerical variables (Age, Income, Tenure, SupportCalls) exhibited balanced distributions, indicating a diverse and representative customer base. Categorical attributes, such as Gender and ProductType, revealed an equal gender split and a notable preference for the Basic service plan. These observations confirm the dataset's suitability for modeling both demographic and behavioral influences on churn.

The correlation and visualization analyses highlighted several important insights. Most notably, **SupportCalls** showed a mild but consistent positive relationship with **ChurnStatus**, suggesting that customers who contact support frequently are more likely to discontinue their service. Additionally, heatmap visualizations revealed that **low-income, low-tenure** customers have higher churn rates—an early warning pattern that can guide customer-retention strategies. In contrast, features such as **Age** and **Income** alone showed limited predictive power, indicating that churn is driven by **interacting behavioral and service-related factors** rather than single demographic attributes.

The overall findings suggest that customer churn is a **multifactor phenomenon**, influenced primarily by satisfaction and engagement indicators rather than static demographics. From a business perspective, these results can help organizations prioritize proactive retention strategies, such as improving customer support quality, offering targeted incentives for low-tenure users, and designing personalized upgrade paths for low-income segments.

In conclusion, this study demonstrated the full data-preprocessing and EDA pipeline—from data cleaning to insight generation—highlighting how statistical and visual techniques can uncover actionable business intelligence. The next logical step would be to

apply predictive modeling techniques (e.g., logistic regression, decision trees) to quantify the contribution of each factor and develop a data-driven churn-prediction system.