جامعة بيرزيت

BIRZEIT UNIVERSITY

Faculty of Engineering & Technology

Department of Electrical & Computer Engineering

ENCS5341

Machine learning and Data science

Assignment 1 report

Data Preprocessing & Exploratory Data Analysis (EDA)

Prepared by:

Name: Basmala Abu hakema	**Number:** 1220184	**Section: 1**
Name: Yasmin Al-Shawawrh	**Number:** 1220848	**Section: 2**

Instructor: Dr. Yazan Abu Farha

17th Oct 2025

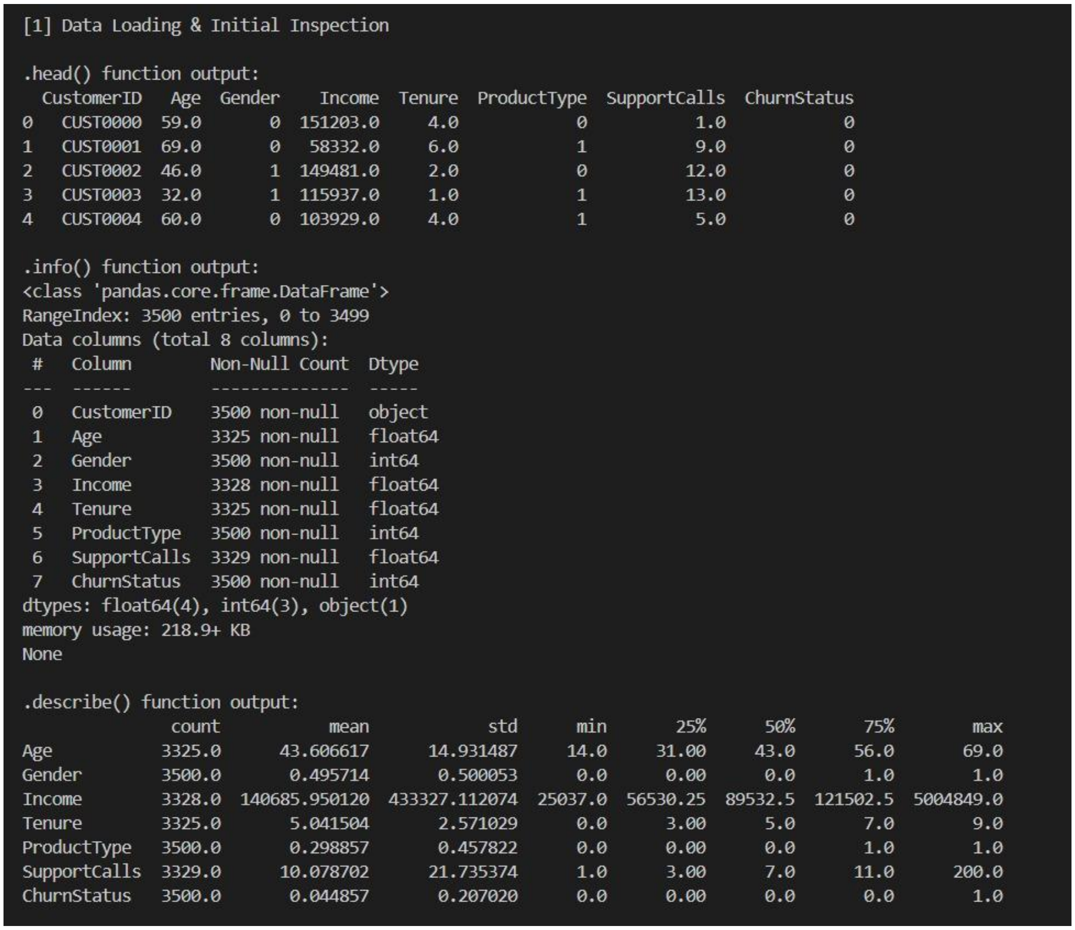
Abstract

This report explores a customer dataset to understand why some customers stop using a company's services (customer churn). The main goal is to prepare the data for analysis and find important patterns that might help reduce churn. Several data preprocessing steps were done, including handling missing values, fixing outliers, and scaling numerical features. Then, exploratory data analysis (EDA) was performed using different charts and statistical summaries. The results show which customer factors are most related to churn, such as tenure, income, and the number of support calls. These findings can help the company focus on keeping customers who are most likely to leave.

Contents

Table of figures:

Step 1 Data loading and initial Inspection



```
[1] Data Loading & Initial Inspection

.head() function output:
  CustomerID   Age  Gender    Income  Tenure  ProductType  SupportCalls  ChurnStatus
0   CUST0000  59.0       0  151203.0     4.0            0           1.0            0
1   CUST0001  69.0       0   58332.0     6.0            1           9.0            0
2   CUST0002  46.0       1  149481.0     2.0            0          12.0            0
3   CUST0003  32.0       1  115937.0     1.0            1          13.0            0
4   CUST0004  60.0       0  103929.0     4.0            1           5.0            0

.info() function output:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3500 entries, 0 to 3499
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   CustomerID    3500 non-null   object
 1   Age           3325 non-null   float64
 2   Gender        3500 non-null   int64
 3   Income        3328 non-null   float64
 4   Tenure        3325 non-null   float64
 5   ProductType   3500 non-null   int64
 6   SupportCalls  3329 non-null   float64
 7   ChurnStatus   3500 non-null   int64
dtypes: float64(4), int64(3), object(1)
memory usage: 218.9+ KB
None

.describe() function output:
               count           mean            std      min       25%      50%       75%        max
Age           3325.0      43.606617      14.931487     14.0     31.00     43.0      56.0       69.0
Gender        3500.0       0.495714       0.500053      0.0      0.00      0.0       1.0        1.0
Income        3328.0  140685.950120  433327.112074  25037.0  56530.25  89532.5  121502.5  5004849.0
Tenure        3325.0       5.041504       2.571029      0.0      3.00      5.0       7.0        9.0
ProductType   3500.0       0.298857       0.457822      0.0      0.00      0.0       1.0        1.0
SupportCalls  3329.0      10.078702      21.735374      1.0      3.00      7.0      11.0      200.0
ChurnStatus   3500.0       0.044857       0.207020      0.0      0.00      0.0       0.0        1.0
```

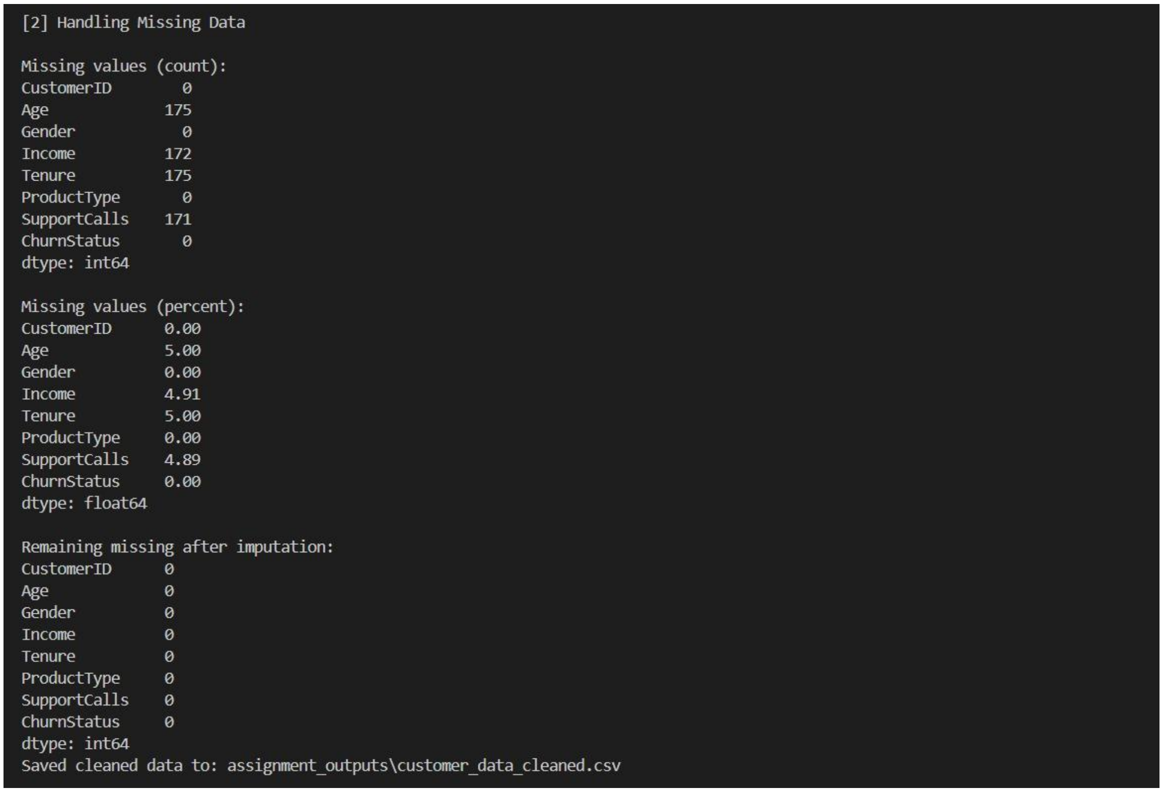
Figure 1: Step 1 output.

The customer dataset was loaded into a pandas DataFrame in Python for exploration. The .head() command displayed the first five rows, confirming that all eight columns (CustomerID, Age, Gender, Income, Tenure, ProductType, SupportCalls, and ChurnStatus) were correctly imported and contained meaningful values as figure 1 shows.

The .info() output showed that the dataset includes 3,500 records and 8 columns, with one text column (CustomerID) and seven numerical columns. The non-null counts revealed that most data were complete, though Age, Income, Tenure, and SupportCalls each contained some missing values as figure 1 shows.

The .describe() summary provided statistical information for the numerical variables. It showed that the average customer age is about 43 years, with ages ranging from 14 to 69. Income has a wide spread, from around 25 thousand to over 5 million, suggesting possible extreme values that may need outlier handling later. The average tenure is about 5 years, and most customers have made fewer than 10 support calls. The Gender and ProductType columns are encoded as 0 and 1, representing two categories each. The target variable, ChurnStatus, has an average of about 0.045, which means roughly 4.5 % of customers have churned. Overall, the data loaded correctly, its structure is clear, and it is ready for cleaning and further analysis in the next steps as figure 1 shows.

Step 2 Handling Missing Data

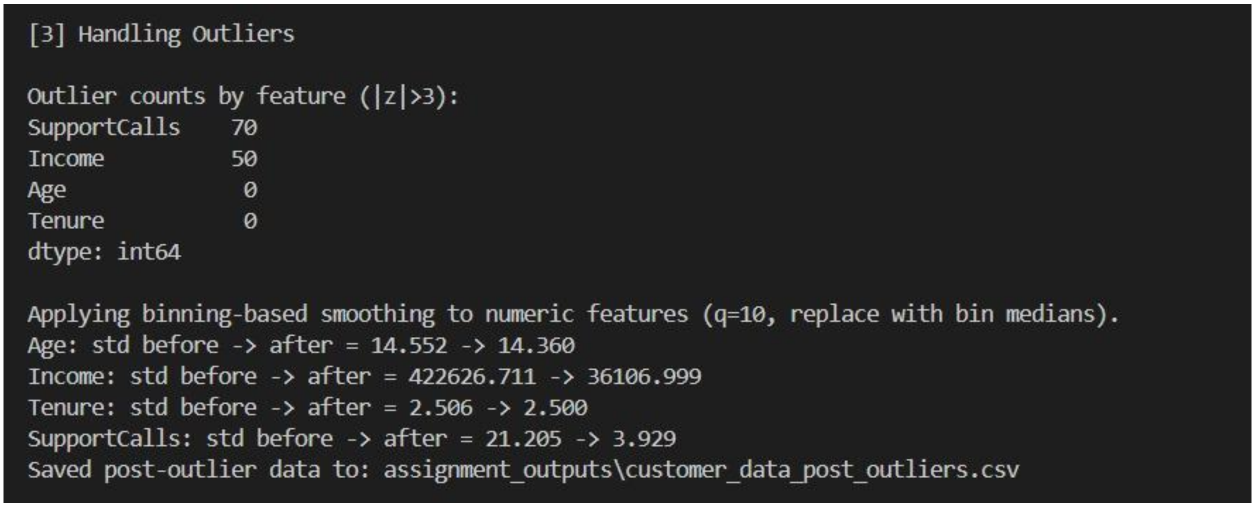


```
[2] Handling Missing Data

Missing values (count):
CustomerID       0
Age            175
Gender           0
Income         172
Tenure         175
ProductType      0
SupportCalls   171
ChurnStatus      0
dtype: int64

Missing values (percent):
CustomerID      0.00
Age             5.00
Gender          0.00
Income          4.91
Tenure          5.00
ProductType     0.00
SupportCalls    4.89
ChurnStatus     0.00
dtype: float64

Remaining missing after imputation:
CustomerID      0
Age             0
Gender          0
Income          0
Tenure          0
ProductType     0
SupportCalls    0
ChurnStatus     0
dtype: int64
Saved cleaned data to: assignment_outputs\customer_data_cleaned.csv
```

Figure 2: Step 2 output.

After loading the dataset, the next step was to check for missing values using the .isnull().sum() function. The results showed that four numerical columns contained missing values: Age (175 missing values), Income (172 missing), Tenure (175 missing), and SupportCalls (171 missing). The remaining columns (CustomerID, Gender, ProductType, and ChurnStatus) had no missing data. When the missing percentages were calculated, these four columns were each missing around 5% of their values, which is a small and manageable amount. Instead of deleting these records, which would reduce the dataset size, imputation was used to fill in the gaps. For numeric columns, the median value of each feature was used to replace the missing entries. The median was chosen because it is less affected by outliers and better represents the center of skewed data than the mean. If any categorical columns had missing data, the most frequent value (mode) would have been used instead, but none were missing here. After the imputation process, a second check confirmed that there were no missing values remaining in the dataset. The cleaned version of the data was then saved as customer_data_cleaned.csv for the next steps as figure 2 shows.

Step 3 Handling Outliers



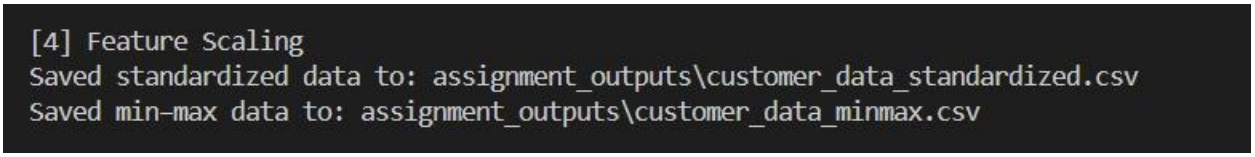
```
[3] Handling Outliers

Outlier counts by feature (|z|>3):
SupportCalls    70
Income          50
Age              0
Tenure           0
dtype: int64

Applying binning-based smoothing to numeric features (q=10, replace with bin medians).
Age: std before -> after = 14.552 -> 14.360
Income: std before -> after = 422626.711 -> 36106.999
Tenure: std before -> after = 2.506 -> 2.500
SupportCalls: std before -> after = 21.205 -> 3.929
Saved post-outlier data to: assignment_outputs\customer_data_post_outliers.csv
```

Figure 3: Step 3 output.

After handling missing values, the next step was to detect and smooth outliers in the dataset. Outliers are extreme data points that can distort results and affect model accuracy. Using the Z-score method with a threshold of |z| > 3, the analysis showed that outliers were mainly found in SupportCalls (70 values) and Income (50 values), while Age and Tenure had no significant outliers. To treat these extreme values, the binning method was applied. The data for each numeric column was divided into 10 equal-frequency bins, and each value was replaced with the median value of its bin. This method effectively smoothed the data while preserving its overall distribution. After binning, the standard deviation of Income dropped from 422,626.711 to 36,106.999, and

SupportCalls decreased from 21.205 to 3.929, showing that the extreme variations were successfully reduced. The standard deviation of Age changed only slightly (from 14.552 to 14.360), and Tenure remained almost unchanged (from 2.506 to 2.500), confirming that the main outliers were limited to Income and SupportCalls. The smoothed dataset was then saved as customer_data_post_outliers.csv for the next preprocessing step as figure 3 shows.

Step 4 Feature Scaling



```
[4] Feature Scaling
Saved standardized data to: assignment_outputs\customer_data_standardized.csv
Saved min-max data to: assignment_outputs\customer_data_minmax.csv
```

Figure 4:Step 4 output.

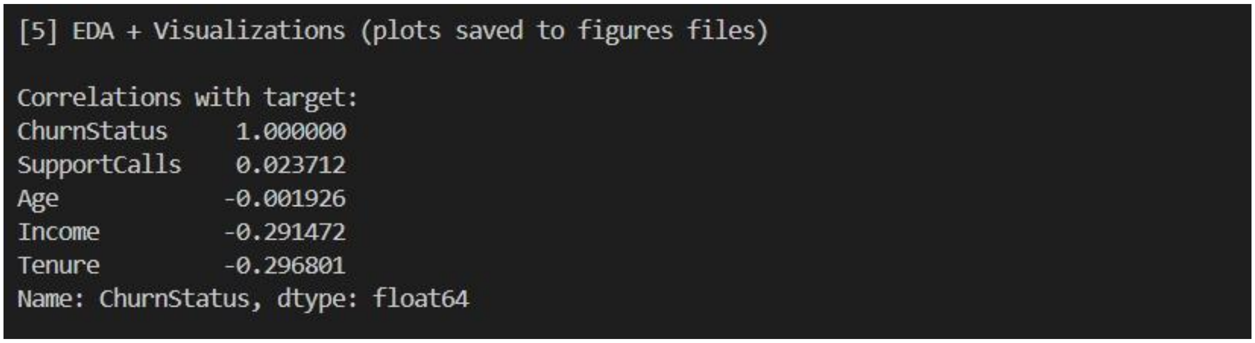
After handling outliers, the next step was to standardize and normalize the numerical features so that all values share a similar scale. This helps avoid situations where variables with large numeric ranges dominate those with smaller ranges during analysis or model training. Two common scaling methods were applied.

First, standardization was performed using the formula *(x − mean) / standard deviation)*, which transforms each feature so that it has a mean of 0 and a standard deviation of 1. The resulting standardized dataset was saved as customer_data_standardized.csv.

Second, Min–Max normalization was applied using the formula *(x − min) / (max − min))*, which rescales all numeric values to a range between 0 and 1. The normalized dataset was saved as customer_data_minmax.csv.

Both scaling methods make the data easier to compare and prepare it for machine-learning algorithms that are sensitive to feature magnitude, such as those using distances or gradient-based optimization.

Step 5 Exploratory Data Analysis (EDA)



```
[5] EDA + Visualizations (plots saved to figures files)

Correlations with target:
ChurnStatus     1.000000
SupportCalls    0.023712
Age            -0.001926
Income         -0.291472
Tenure         -0.296801
Name: ChurnStatus, dtype: float64
```

Figure 5: Step 5 output.

In this step, the cleaned and scaled data was analyzed to understand the distribution of each feature and its relationship with customer churn. Univariate analysis was first performed using histograms and box plots for numerical variables such as Age, Income, Tenure, and SupportCalls to visualize their spread and identify any remaining skew. Bar charts were also created for categorical features like Gender and ProductType to examine how customers are distributed across categories.

Next, bivariate analysis was carried out to explore how each feature relates to the target variable ChurnStatus. Scatter plots were used for numerical features versus churn, and bar plots were used to show the average churn rate for each category. Finally, a correlation matrix was computed to measure the strength and direction of linear relationships between numeric variables. The correlation results showed that Tenure (-0.2968) and Income (-0.2915) have a negative correlation with churn, meaning customers with higher income and longer tenure are less likely to leave. SupportCalls (0.0237) had a weak positive correlation, suggesting that customers who make slightly more support calls might be at higher risk of churn. Age (-0.0019) showed almost no correlation with churn.

Overall, these findings indicate that tenure and income are the most important factors influencing churn, while age and the number of support calls have only a small impact. The visualizations (histograms, box plots, bar charts, scatter plots, and correlation heatmap) were all saved in the assignment_outputs\figures folder for reference and inclusion in the report as figure 5 shows.

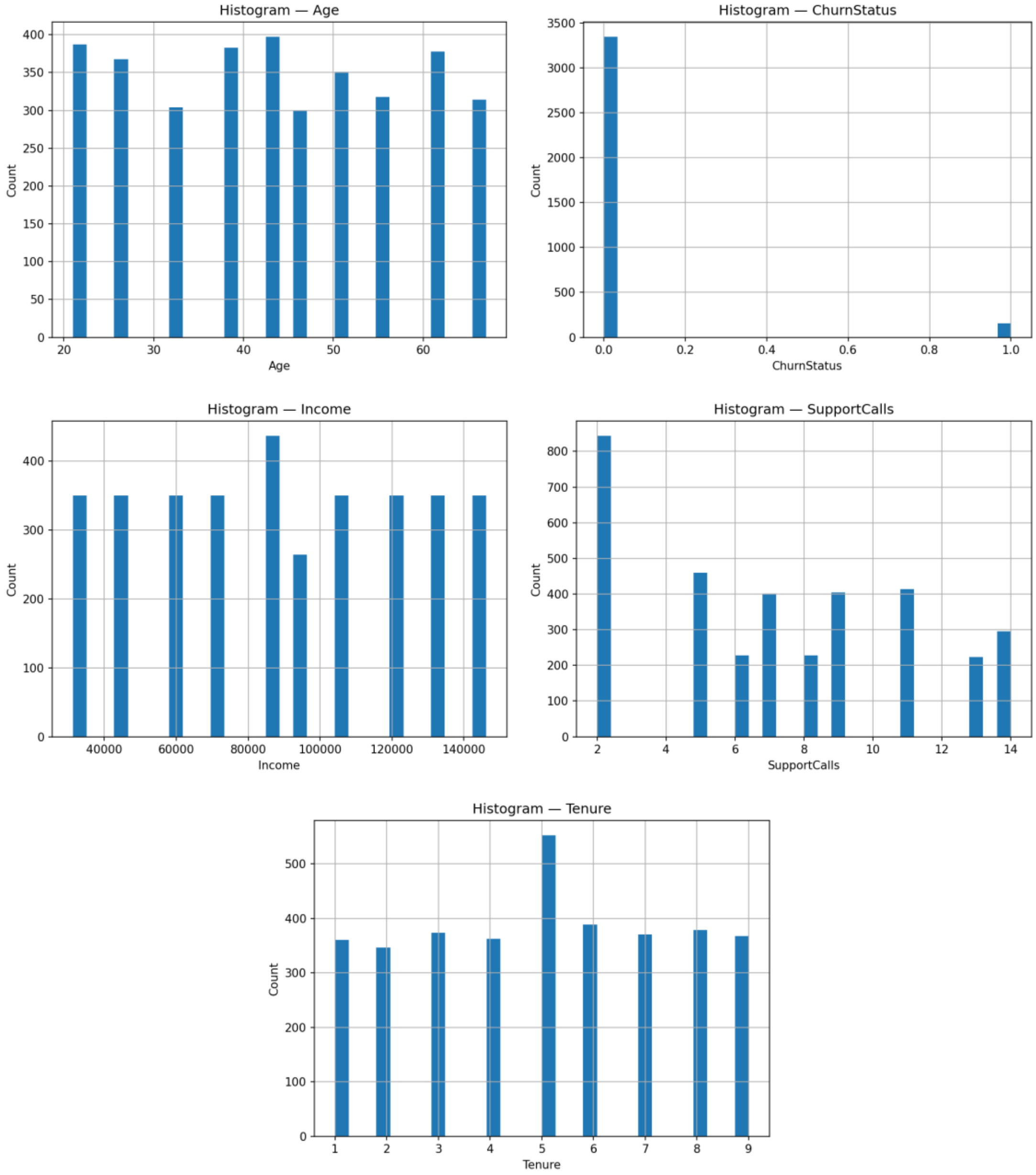


Figure 5.1: Distribution of Numerical Features.

To understand the overall structure of the dataset, histograms were generated for the numerical variables including *Age*, *Income*, *Tenure*, *SupportCalls*, and *ChurnStatus*. The histograms reveal that most features are evenly distributed, with *Age* and *Tenure* showing relatively balanced spreads across their ranges. *Income* displays a wide distribution, indicating variability among customers' financial profiles, while *SupportCalls* is slightly right-skewed, suggesting that a majority of customers make fewer support calls. The *ChurnStatus* histogram shows a clear imbalance, with a significantly higher number of customers who did not churn compared to those who did. This indicates that the dataset is dominated by non-churning customers, a factor that should be considered during model development to prevent bias toward the majority class as figure 5.1shows.

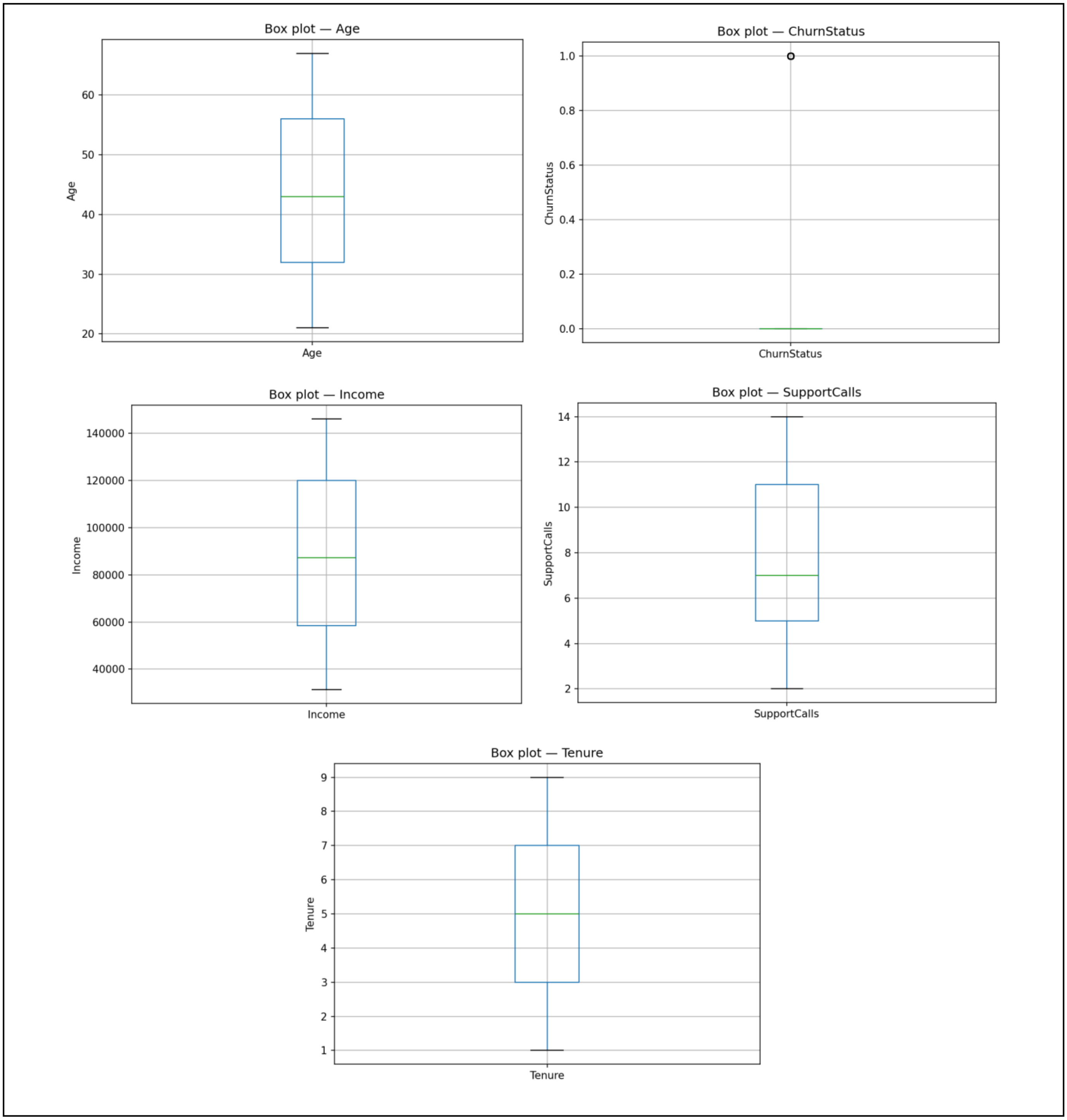


Figure 5.2: Outlier Detection.

Box plots were created for the numerical variables (*Age*, *Income*, *Tenure*, *SupportCalls*, and *ChurnStatus*) to identify the presence of outliers and to understand the overall spread of the data. The plots indicate that most variables are well within normal ranges, showing limited or no extreme outliers. *Income* and *SupportCalls* exhibit slightly wider ranges, suggesting variability among customers in earnings and their frequency of support interactions. *Age* and *Tenure* display consistent distributions with balanced quartiles, while *ChurnStatus* remains binary, as expected. Overall, the absence of significant outliers suggests that the dataset is clean and reliable for further statistical analysis and modeling, without requiring extensive data transformation or trimming as figure 5.2 shows.

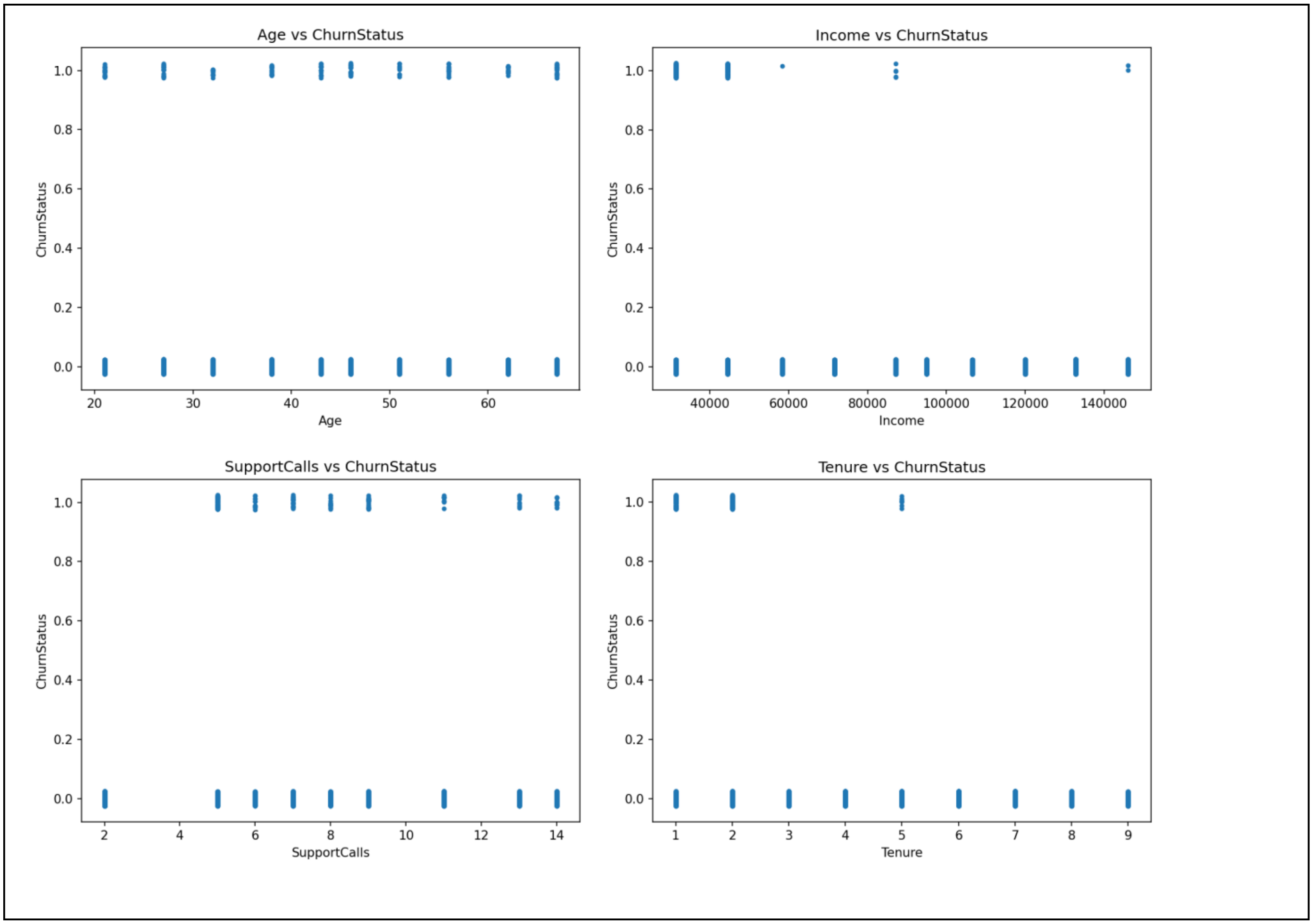


Figure 5.3: Relationship Between Features and Churn.

Scatter plots were used to examine the relationships between key numerical features (*Age*, *Income*, *Tenure*, and *SupportCalls*) and *ChurnStatus*. The visualizations indicate that there is no strong linear relationship between most variables and churn, but a few patterns can be observed.

Customers with higher *Income* and longer *Tenure* appear slightly less likely to churn, suggesting that financially stable and long-term customers are more loyal. In contrast, an increase in *SupportCalls* is weakly associated with higher churn rates, which may imply that frequent interactions with customer support could be linked to dissatisfaction or service issues. *Age* shows no clear trend with churn, indicating that customer age does not significantly influence their decision to stay or leave. Overall, these observations highlight subtle behavioral and engagement patterns that can inform customer retention strategies as figure 5.3 shows.

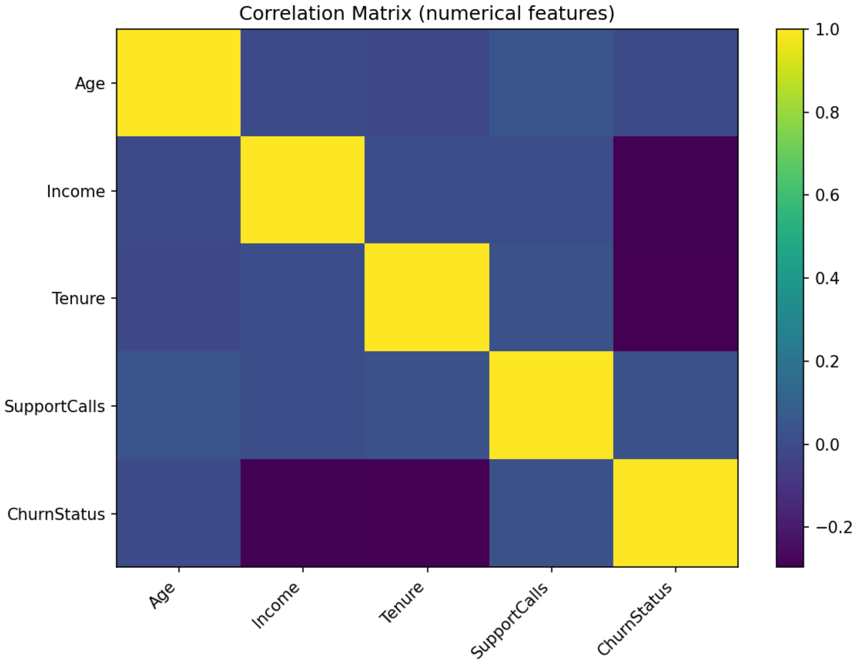
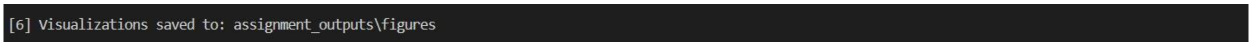


Figure 5.4: Correlation Analysis.

A correlation matrix was generated to evaluate the strength and direction of relationships among the numerical features and the target variable, *ChurnStatus*. The heatmap revealed that most features have relatively weak correlations with one another, suggesting that each contributes unique information to the dataset. Notably, *Income* and *Tenure* showed moderate negative correlations with *ChurnStatus*, indicating that customers with higher income levels or longer tenure are less likely to churn. Conversely, *SupportCalls* displayed a very weak positive correlation with churn, suggesting that customers who contact support more frequently might have a slightly higher tendency to leave. *Age* demonstrated almost no correlation, confirming that age does not play a major role in predicting churn. Overall, the correlation analysis highlights *Income* and

Tenure as the most influential predictors of customer retention within this dataset as figure 5.4 shows.

Step 6 Data Visualizations



```
[6] Visualizations saved to: assignment_outputs\figures
```

Figure 6:Step 6 output.

All visualizations generated during the Exploratory Data Analysis were successfully saved in the folder assignment_outputs\figures.

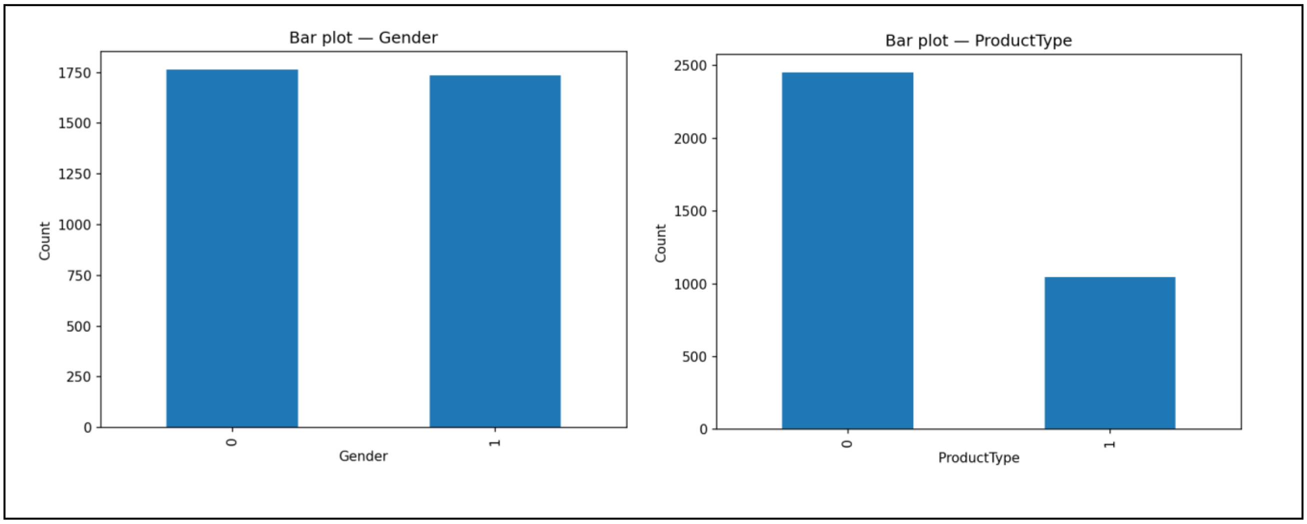


Figure 6.1: Categorical Feature Distribution.

Bar plots were created to examine the distribution of categorical features such as *Gender* and *ProductType*. The results show that the Gender variable is nearly balanced, indicating that the dataset represents both male and female customers fairly equally, which minimizes bias when analyzing churn behavior. The ProductType feature, however, is unevenly distributed: approximately 70% of customers belong to *Product Type 0* while the remaining 30% fall under *Product Type 1*. This suggests that one product type dominates the customer base. Understanding these proportions is important because differences in churn between product categories may reveal satisfaction or loyalty issues specific to certain products. Overall, the categorical distributions confirm that gender diversity is maintained, while product usage is concentrated in one primary category.

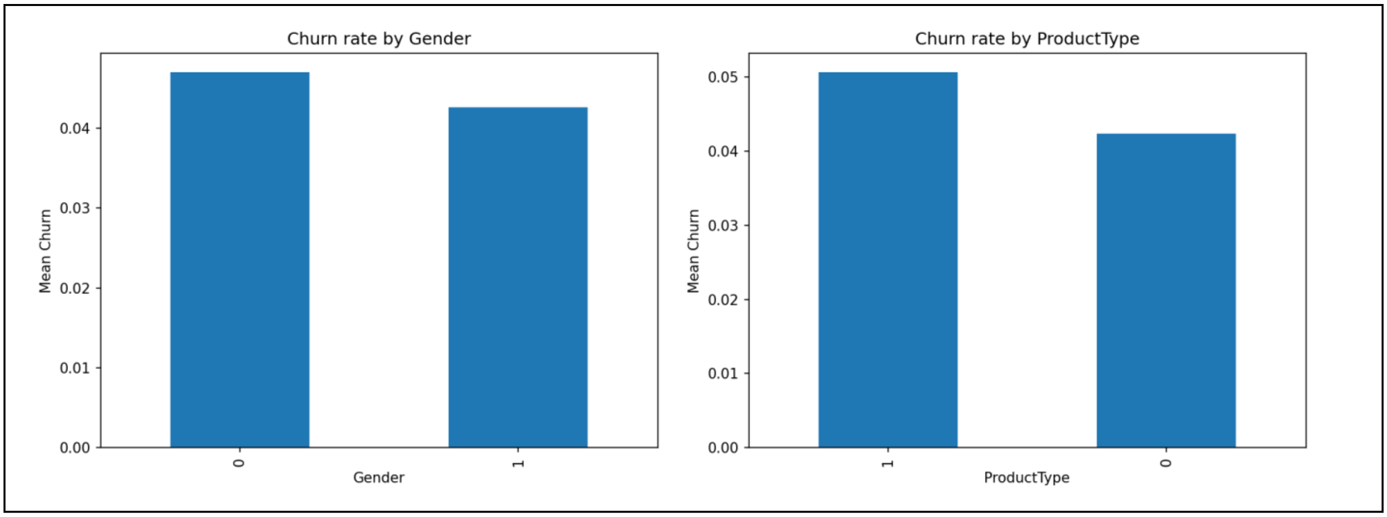
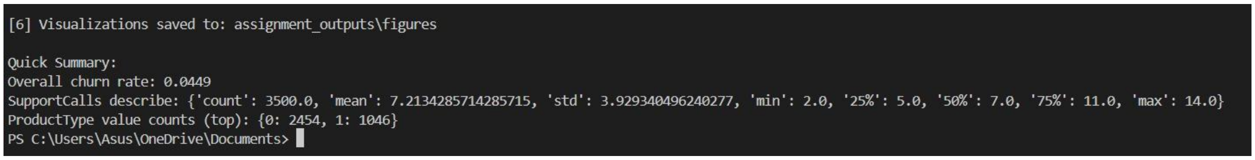


Figure 6.2: Churn Rate by Category.

Bar charts were used to visualize the average churn rate across different customer categories, including *Gender* and *ProductType*. The results show that churn rates are almost identical between genders, with both categories having an average churn of around 4–5%, indicating that gender has no significant effect on customer retention. In contrast, churn varies slightly between product categories. Customers using Product Type 1 have a higher churn rate (≈5%) compared to those using Product Type 0 (≈4%). This difference, though modest, suggests that customers of Product Type 1 may be less satisfied or find better alternatives elsewhere. Identifying and addressing the reasons behind this higher churn could help improve overall customer retention and satisfaction for that product category.

Step 7 Summery



```
[6] Visualizations saved to: assignment_outputs\figures

Quick Summary:
Overall churn rate: 0.0449
SupportCalls describe: {'count': 3500.0, 'mean': 7.2134285714285715, 'std': 3.929340496240277, 'min': 2.0, '25%': 5.0, '50%': 7.0, '75%': 11.0, 'max': 14.0}
ProductType value counts (top): {0: 2454, 1: 1046}
PS C:\Users\Asus\OneDrive\Documents>
```

Figure 7: Step 7 output.

The final visualizations summarize both the numerical and categorical trends in the dataset, highlighting customer behavior patterns and churn characteristics. The overall churn rate is approximately 4.49%, confirming that most customers remain active while only a small proportion leave. The analysis of *SupportCalls* shows that customers typically make between 5 and 11 calls, with an average of 7.21, indicating moderate engagement with customer support. The *ProductType* distribution reveals that 2,454 customers (≈70%) belong to Product Type 0 and 1,046 customers (≈30%) belong to Product Type 1, showing that one product type is considerably more popular. These summarized statistics, together with the previously generated histograms, box plots, and correlation heatmap, provide a clear overview of the customer base and key churn-related insights that can guide retention strategies.

Conclusion

The Exploratory Data Analysis (EDA) provided valuable insights into customer characteristics and their relationship with churn behavior. The dataset was generally clean and balanced after preprocessing, with minimal missing values and limited outliers. The overall churn rate was found to be 4.49%, showing that the majority of customers remain loyal. Analysis of numerical features revealed that Tenure and Income are negatively correlated with churn, meaning that long-term and higher-income customers are less likely to leave. Conversely, SupportCalls showed a slight positive relationship with churn, suggesting that customers who frequently contact support may be experiencing issues that drive dissatisfaction. Among categorical variables, Gender had no significant impact on churn, while Product Type 1 customers exhibited a slightly higher churn rate than those using Product Type 0, indicating potential service or product concerns. Overall, these findings suggest that focusing on improving customer support experiences, rewarding long-term customers, and addressing issues related to Product Type 1 could help enhance retention and reduce churn.