Faculty of Engineering and Technology

Department of Electrical and Computer Engineering

ENCS5210

Machine Learning

Assignment #1 – Churn Analysis

Prepared by: Abdalraheem Shuaibi 1220148
Heba Mustafa 1221916

Instructor: Dr. Yazan AbuFarha

Date: 10/5/2025

Abstract

In this assignment we will be analyzing a synthetic dataset to determine why customers churn. This includes preprocessing the data including filling in missing data, and handling outliers to prepare the data for proper analysis. We aim to find correlations and identify key insights that help realize the relation between the features and the churn.

Table of Contents

Table of Figures

Table of Tables

Introduction

Data pre-processing and exploratory data analysis **(EDA)** are essential steps in any machine learning project, before start building the predictive module, the data must be cleaned, standardized and explored. Pre-processing take care of inappropriate data such as missing values and outliers then scale the data to make sure all features contribute fairly and consistently. EDA help understanding patterns, trends and relationships through statistical tools and visualizations.

You May view the Graphs and justifications from our website:

Customer Churn Analysis

Dataset Description

To do any analysis, the dataset must be explored and observed first, determining the key concepts and understanding the dataset structure

The dataset analysed in this project represents a synthetic customer database designed to simulate real-world business data related to customer behaviour and churn prediction.

Dataset Features Summary

Table 1 outlines all the features included in the dataset along with their data types and descriptions, help understanding the role of each feature for doing appropriate preprocessing and analysis methods.

Table 2: Features Summary

Feature Name	Type	Description
CustomerID	Object	Unique identifier assigned to each customer.
Age	Float	Represents the customer's age in years.
Gender	Int	Indicates the customer gender Hot-Encoded as 0 = Male, 1 = Female.
Income	Float	Annual income of the customer (in USD).
Tenure	Float	Number of years the customer has been with the company.
ProductType	Int	Indicates the subscription type, Hot-Encoded as 0 = Basic, 1 = Premium.
SupportCalls	Float	Number of support calls made by the customer in the last year.
ChurnStatus	Int	Target variable: 1 = Churned, 0 = Stayed.

Data Set Information

Both **Table 2** and **Table 3** shows the structural and technical details of the dataset. **Table 2** summarizes the overall dataset properties, including the total number of records, number of features, memory usage, data type in general. **Table 3** provides an overview for each feature that lists the non-null count and data type for each attribute.

Table 3: Dataset information - General

Property	Details
Entries	3500 rows
Columns	8 total
Memory Usage	218.9 KB
Data Types	4 float64, 3 int64, 1 object

Table 4: Dataset Information - Features

Column	Non-Null Count	Dtype
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 4 summarizes the entry count and the main statistical measures of each numerical attribute, including the mean, standard deviation, and quartile values. These statistics provide an initial understanding of how each feature behaves and give a general overview of the data distribution before going deeper in each.

Table 5: Dataset Statistics Summary

Statistic	Age	Gender	Income	Tenure	ProductType	SupportCalls	ChurnStatus
Count	3325	3500	3328	3325	3500	3329	3500
Mean	43.61	0.50	140686	5.04	0.30	10.08	0.04
Std	14.93	0.50	433327	2.57	0.46	21.74	0.21
Min	14	0	25037	0	0	1	0
25%	31	0	56530	3	0	3	0
50%	43	0	89533	5	0	7	0
75%	56	1	121502	7	1	11	0
Max	69	1	5.00×10^{6}	9	1	200	1

Features Distribution

While Section 2.3 provided numerical observations through descriptive statistics, this section complements it by visualizing how each feature is distributed across the dataset. Graphical representations help confirm the numerical findings from the previous section and reveal additional insights such as skewness, the presence of outliers, or uneven class distributions.

1- Distribution of Numerical Features

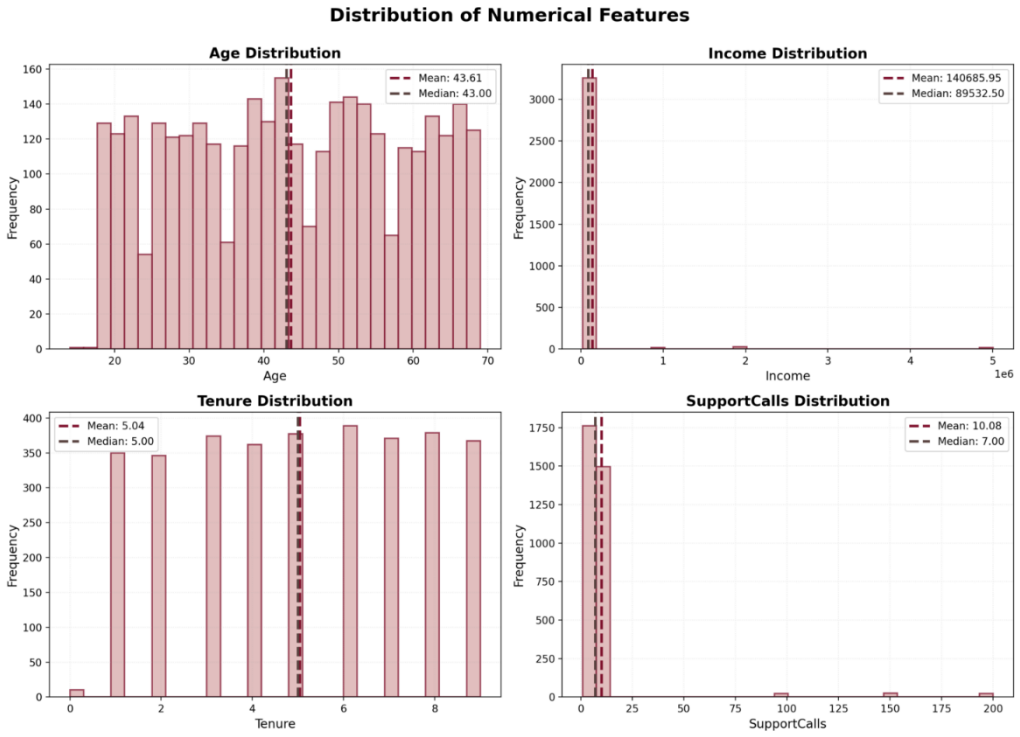

Figure 1: Distribution of Numerical Features

2-

3- Distribution of Categorical Features

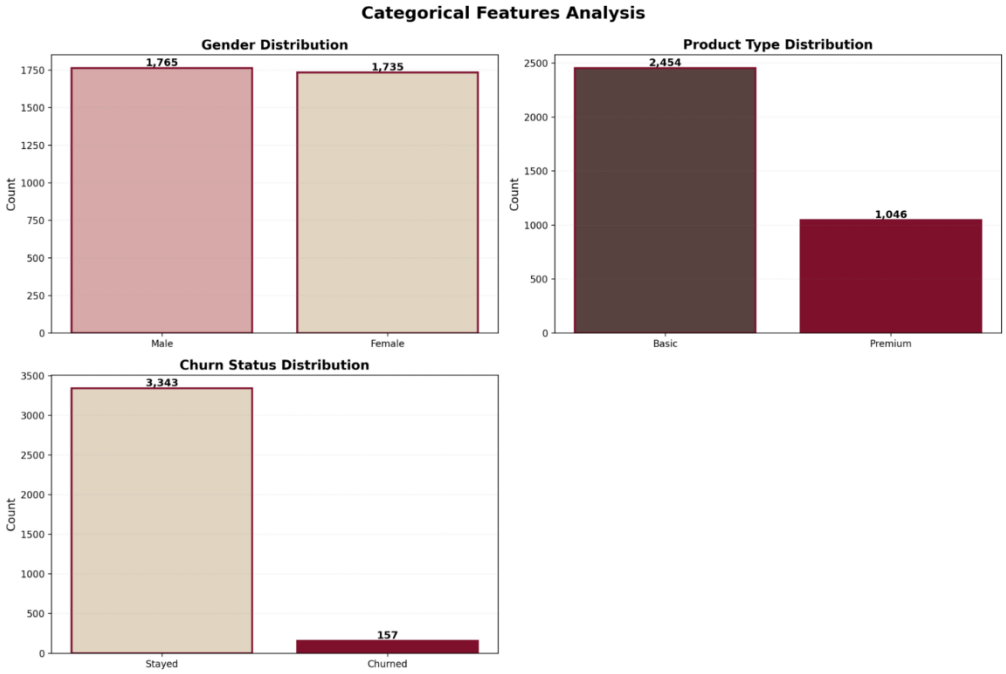

Figure 2: Distribution of Categorical Features

Observations from Data Statistics

Table 5 summarizes the key observations derived from Sections 2.1 through 2.4, compiling the main characteristics and patterns identified for each feature.

Those observations provide an insight for the dataset, highlighting possible data quality issues, feature relevance, and relationships that will guide the preprocessing and analysis steps in the next chapter.

Table 6: Observations from Data Statistics

Feature Name	Observation
CustomerID	Unique identifier, *irrelevant*
Age	Customer ages vary widely; most are middle-aged to older.
Gender	Hot Encoded, Both genders churn almost equally likely *irrelevant*.
Income	Customer incomes vary greatly; the standard deviation is much larger than the mean, indicating outliers.
Tenure	Tenure seems reasonable, the range is from (0-9), and it's distributed about evenly..
ProductType	Hot Encoded, High Majority of customers are subscribed to the basic type.
SupportCalls	Data seems consistent up to the 75% mark, but the max indicates wrongly entered data.
ChurnStatus	Most customers **stayed**.

Data Pre-processing

After gaining an overall understanding of the dataset's structure and basic statistical behaviour in the previous sections, this stage focuses on examining each feature individually. Where every attribute is analysed through its descriptive statistics and visual distribution to identify data quality issues like missing values and outliers. Based on these insights, proper handling methods are applied per feature.

Handle Missing Data & Outliers

4- Age Pre-processing

Missing Values: filled using the **mean**

The missing values were filled using the **mean**, since both **Table 7** and **Figure 3** show that the data distribution is nearly symmetric. The mean (43.6) and median (43) are very close, indicating that there's barley skewness. Therefore, using the mean or median won't make a large difference, hence the mean which is 44 was used.

Outliers: Dropped

Outliers were identified based on domain knowledge rather than purely statistics. Any record with an age value below **18** was considered invalid. Since there were only two entries, remove them won't make such a difference to the dataset as you can see **in table 8 – outlier percentage**.

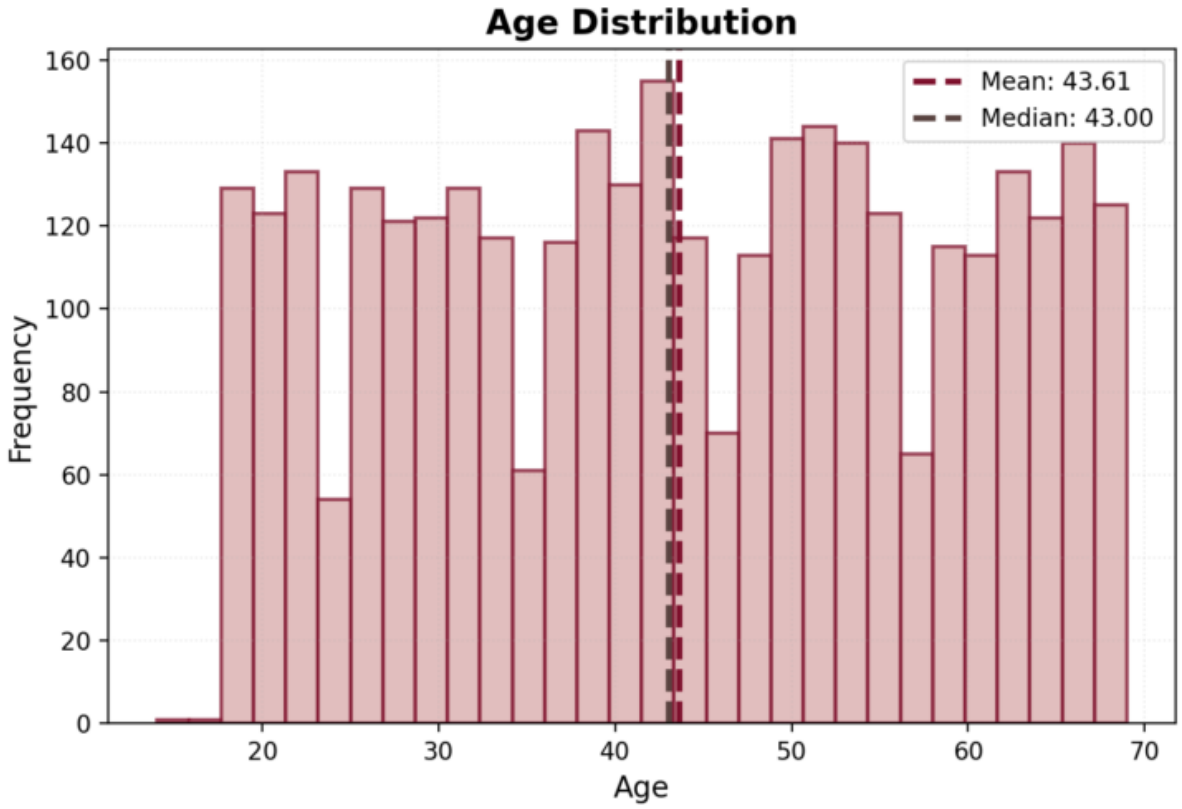

Figure 3: Age Distribution

Table 7: Age Statistics

Statistic	Age
Count	3325
Mean	43.61
Std	14.93
Min	14
25%	31
50%	43
75%	56
Max	69

Table 8: Age Summary

Statistic	Value
Missing Values	175
Missing Values Percentage	5%
Outliers Found	2
Outliers Percentage	0.06%
Skewness	9.6436

5- Income Pre-processing

Missing Values: filled using the **median**

The missing values were filled using the **median**. The skewness value as in **Table 10** appears relatively high, **Figure 4** shows that most income values are concentrated around the median. This indicates that the median represents the data distribution more accurately than the mean, making it the appropriate choice for handling missing values in this feature.

Outliers: Capping at **upper whisker**

Outliers were detected using both the **Interquartile Range (IQR)** method and the **Z-score approach** with a threshold of ±3 standard deviations as justified in **Table 11** and **Figure 5**. Both methods gave the same results, confirming the same extreme income values as in **Table 9**. These outliers were not treated as errors but rather as valid high income entries that reflect real world situation. Therefore, instead of deleting or replacing them with mean or median values—which would reduce data richness—the outliers were **capped at the upper whisker**. This approach preserves their influence while preventing them from disproportionately affecting the analysis.

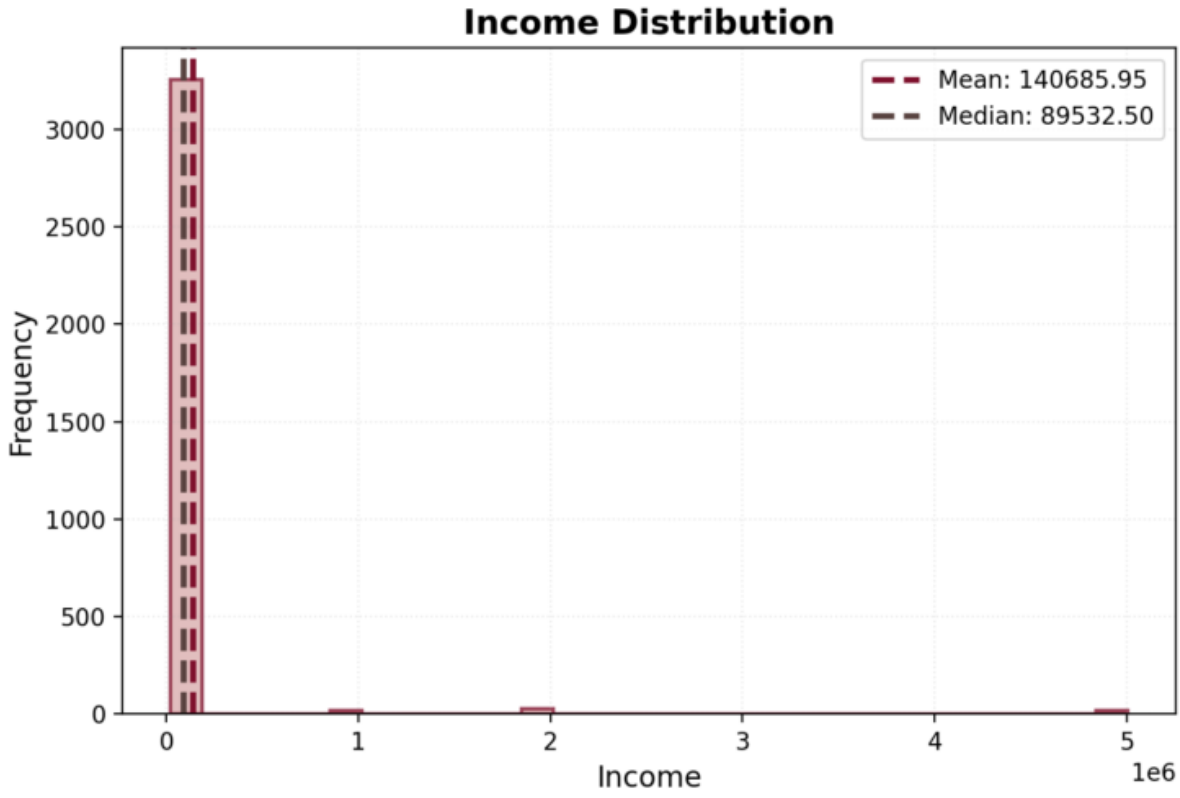

Figure 4: Income Distribution

Table 9: Income Statistics

Statistic	Income
Count	3328
Mean	140686
Std	433327
Min	25037
25%	56530
50%	89533
75%	121502
Max	5.00×10^{6}

Table 10: Income Summary

Statistic	Value
Missing Values	172
Missing Values Percentage	4.9%
Outliers Found	70
Outliers Percentage	2%
Skewness	9.6436

Table 11: Income Box Plot Summary

Statistics	Value
Q1	58,349.50
Q3	119,919.25
IQR	61,569.75
Upper Whisker	212,273.88
Lower Whisker	-34,005.12

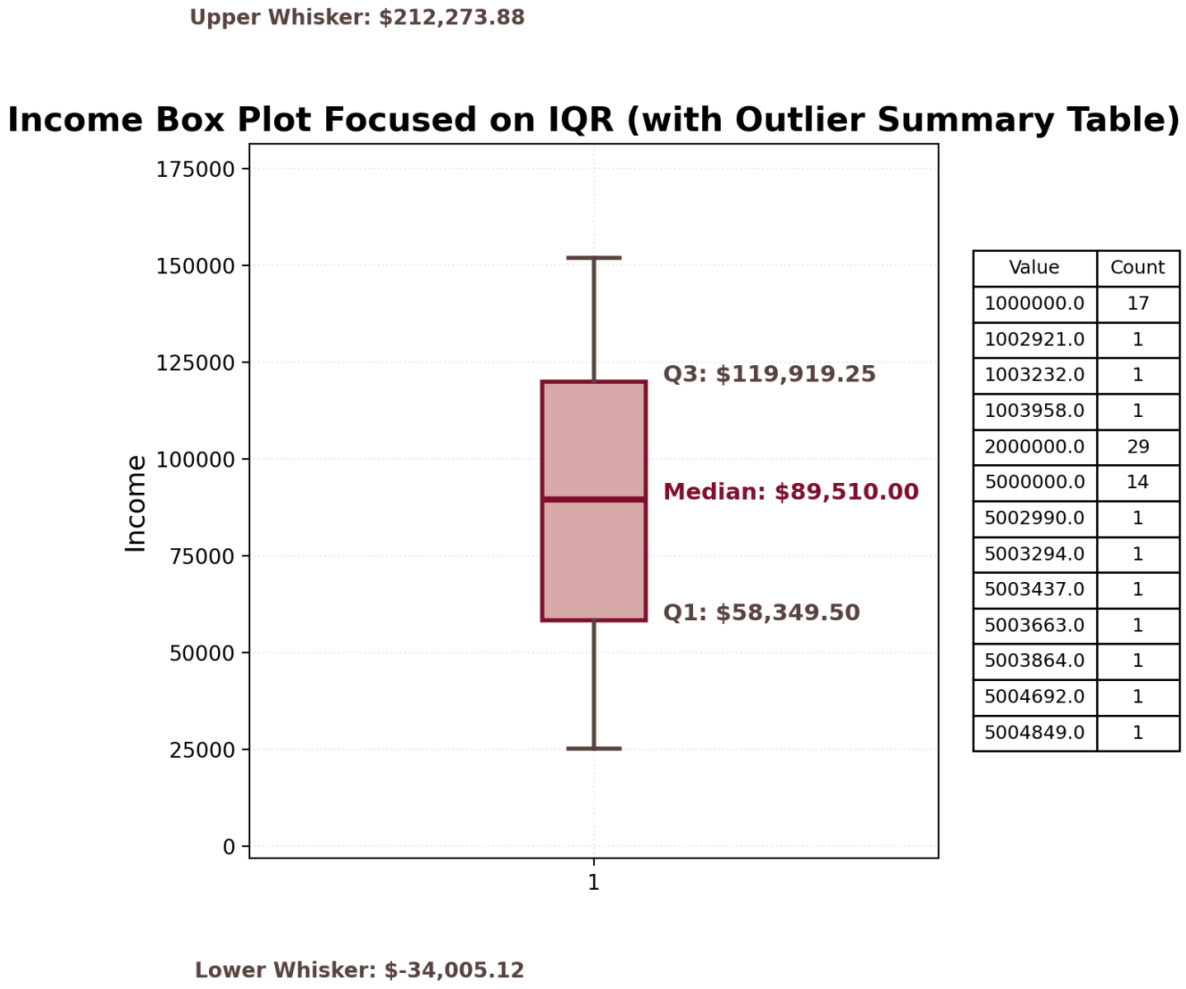

Figure 5: Income Box Plot

Outliers Treatment Options:

~~**- Dropping** — No, we would lose too much data.~~

~~**- Smoothing** — Data is too high; smoothing won't help.~~

- Capping at upper whisker — Keeps the data rows and helps handle outliers ✓

6- Tenure Pre-processing

Missing Values: filled using the **Median**

The missing values were filled using the **median**. As shown in **Figure 6** and **Table 12**, the tenure data follows a roughly uniform distribution —except some with zero values— without significant skewness. In such cases, both mean and median effectively represents the central value of the feature, with roughly same value for each, hence median were used.

Outliers: No significant outliers detected in Tenure data.

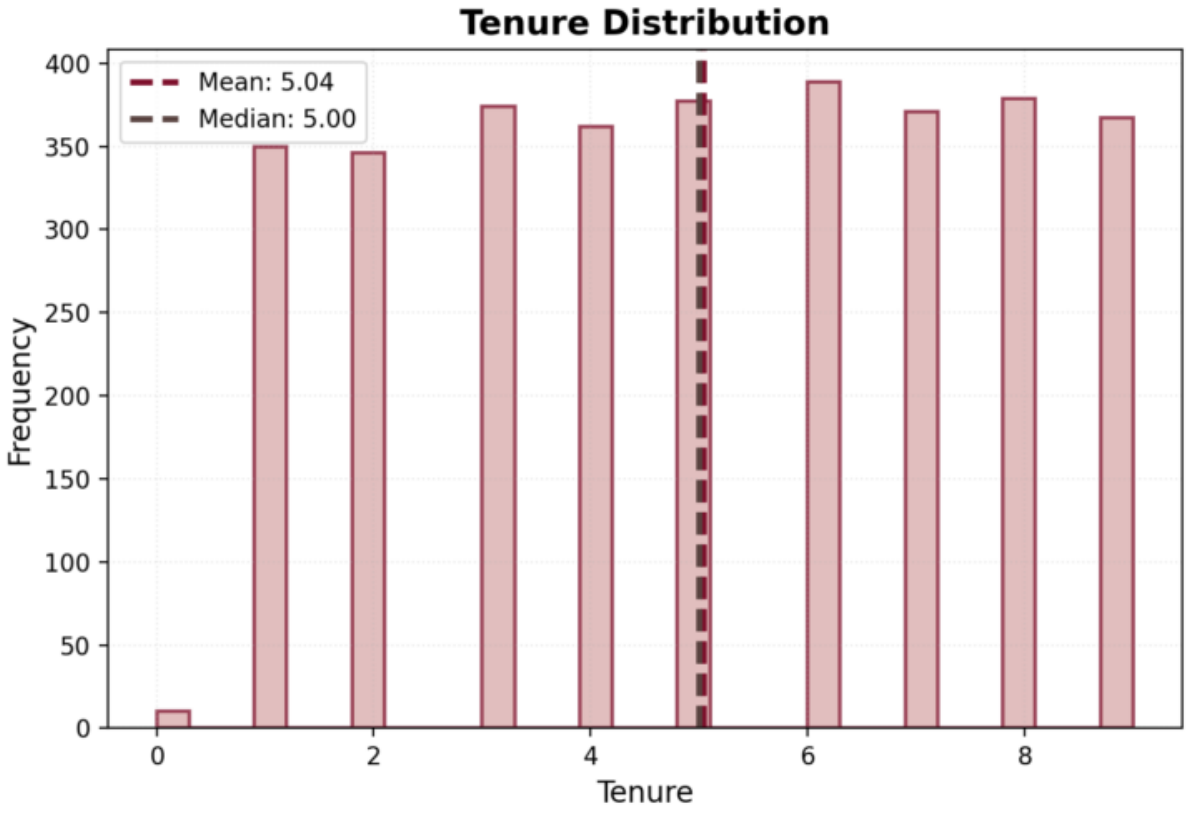

Figure 6: Tenure Distribution

Table 12: Tenure Statistics

Statistic	Tenure
Count	3325
Mean	5.04
Std	2.57
Min	0
25%	3
50%	5
75%	7
Max	9

Table 13: Tenure Summary

Statistic	Value
Missing Values	175
Missing Values Percentage	5%
Outliers Found	0
Outliers Percentage	0.0%
Skewness	-0.0357

7- SupportCalls Pre-processing

Missing Values: Filled with the **Median**

The missing values were filled using the **median**, since the distribution of support calls is **highly skewed**, as shown in **Table 15**. In such cases, the median provides a more reliable estimate of central tendency than the mean, as it is less influenced by extreme high values. This choice is further supported by **Figure 7**, where most data points are crowded around the median, making it the most representative value for imputing the missing entries.

Outliers: Capping to the **Median**

Outliers were detected using the **Interquartile Range (IQR)** method, as presented in **Table 14**. The outliers here are unrealistic and therefore treated as an invalid or erroneous entry. Unlike the **Income** feature, where extreme values represented genuine high-income customers, the unusually large support call realistic. These outliers were handled by **capping them at the median**, effectively treating them as missing or misrecorded values while keeping the general shape and consistency of the distribution.

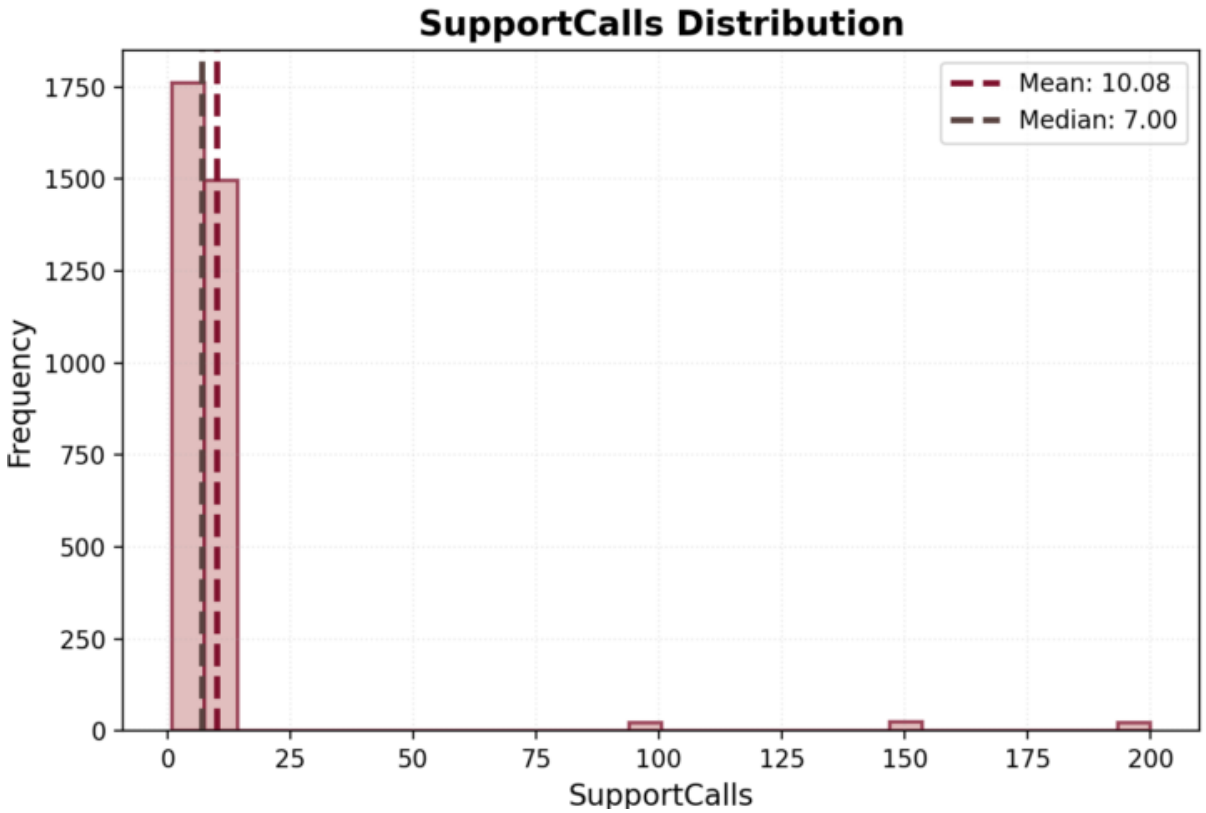

Figure 7: SupportCalls Distribution

Table 14: SupportCalls Statistics

Statistic	Tenure
Count	3329
Mean	10.08
Std	21.74
Min	1
25%	3
50%	7
75%	11
Max	200

Table 15: SupportCalls Summary

Statistic	Value
Missing Values	171
Missing Values Percentage	4.8%
Outliers Found	70
Outliers Percentage	2.1%
Skewness	7.0

8- Gender Pre-processing:

Missing Values: No missing values were found in the **Gender** data.

Outliers: No significant outliers detected in **Gender** data.

Therefore, **no preprocessing was required**.

9- ProductType Pre-processing

Missing Values: No missing values were found in the **ProductType** data**.**

Outliers: No significant outliers detected in **ProductType** data.

Therefore, **no preprocessing was required**.

After Pre-processing

After handling missing values and outliers for each feature, the dataset was re-evaluated to confirm the effect of the preprocessing steps. As shown in the updated **statistics and distributions** in **Table 16**, the data became more consistent and realistic, with reduced variance in highly skewed attributes such as ***Income*** and ***SupportCalls***. At the same time, the overall structure and natural relationships among features were preserved.

Figures **8** and **9** illustrate how the distributions of both numerical and categorical features appear smoother and more balanced after cleaning.

Table 16: Dataset Statistics Summary After Preprocessing

Statistic	Age	Gender	Income	Tenure	ProductType	SupportCalls	ChurnStatus
Count	3498	3498	3498	3498	3498	3498	3498
Mean	43.64	0.50	90115.76	5.04	0.30	7.07	0.04
Std	14.54	0.50	39318.82	2.51	0.46	4.05	0.21
Min	18	0	25037.00	0	0	1	0
25%	32	0	58349.50	3	0	4	0
50% (Median)	44	0	89510.00	5	0	7	0
75%	56	1	119919.25	7	1	10	0
Max	69	1	212273.88	9	1	14	1

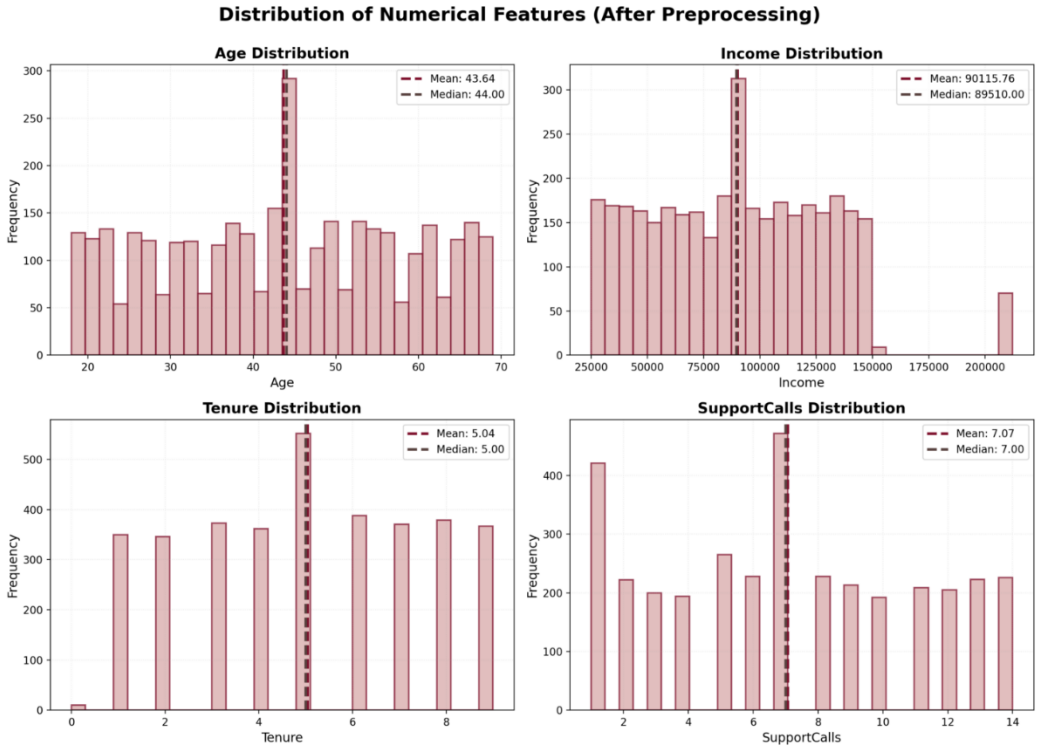

Figure 8: Distribution of Numerical Features After Preprocessing

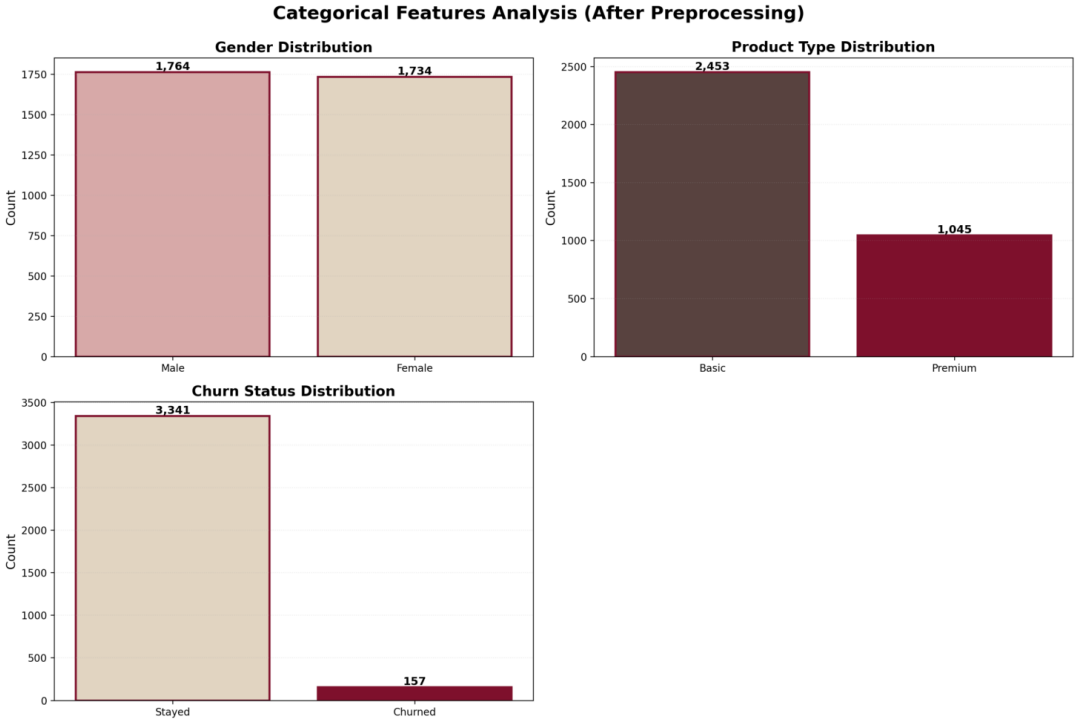

Figure 9: Distribution of Categorical Features After Preprocessing

Standardization

After completing the preprocessing stage, the next step was to **standardize the numerical features** to ensure that all variables contribute equally during analysis. Standardization transforms features to a common scale without distorting differences in their ranges. This step is essential because attributes like *Income* and *SupportCalls* have much larger numeric values than *Age* or *Tenure*, which could otherwise dominate statistical relationships or model training.

To achieve this, the **Z-score standardization** method was applied. In this approach, each feature value is transformed based on the following equation:

$$z = \frac{(x - \mu)}{\sigma}$$

Where x is the original value, μ is the feature's mean, and σ is its standard deviation. This process centers the data around zero and scales it so that each feature has a standard deviation of one. As a result, all standardized features become comparable in scale, improving both the interpretability and performance of subsequent analysis steps. **Figure 10** shows the effect of this transformation, where the categorical features and numerical distributions appear normalized and well-aligned for further **exploratory data analysis (EDA)**.

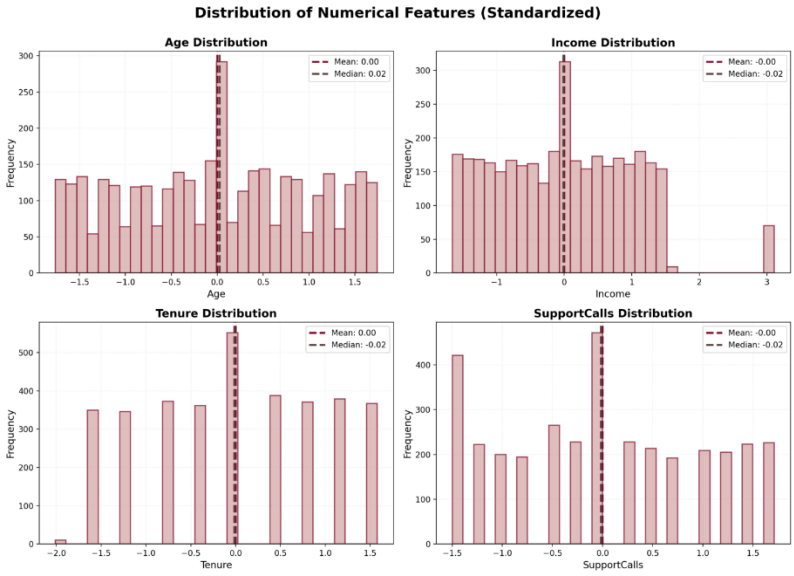

Figure 10: Distribution of Categorical Features After Standarization

Exploratory Data Analysis

After completing the preprocessing and standardization steps, the dataset was ready for **Exploratory Data Analysis (EDA)**. The purpose of this stage is to gain a deeper understanding of the relationships, patterns, and behaviors within the data. Through visualization and statistical examination, EDA helps identify key factors that may influence customer churn and guides future predictive modeling.

In this section, various analyses and plots were conducted to interpret both **numerical** and **categorical** variables with respect to the target variable (*ChurnStatus*):

- **Box plots** were applied to assess distribution ranges and spot variations between churn categories.
- **Bar charts** were used for categorical variables (*Gender* and *ProductType*) to observe churn proportions within each class.
- **Scatter plots** were used to visualize how numerical features such as *Age*, *Income*, *Tenure*, and *SupportCalls* differ between customers who stayed and those who churned.
- **Statistical Compression** summarized the numerical features by comparing mean and median values between churned and non-churned customers, helping to highlight which attributes show the most noticeable behavioral differences.
- **Correlation Analysis** quantified how strongly each feature relates to churn, identifying the variables with the highest predictive potential.

Overall, these combined analyses provide a comprehensive understanding of customer characteristics and the main factors contributing to churn, serving as a foundation for later modeling and decision-making.

Churn Analysis

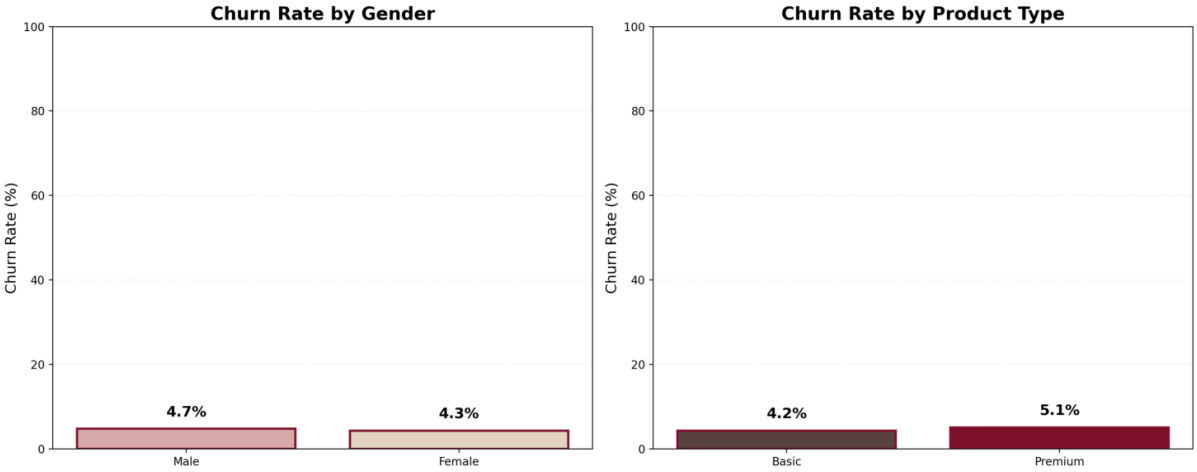

Figure 11: Churn Rate by Categorical

Box Plots

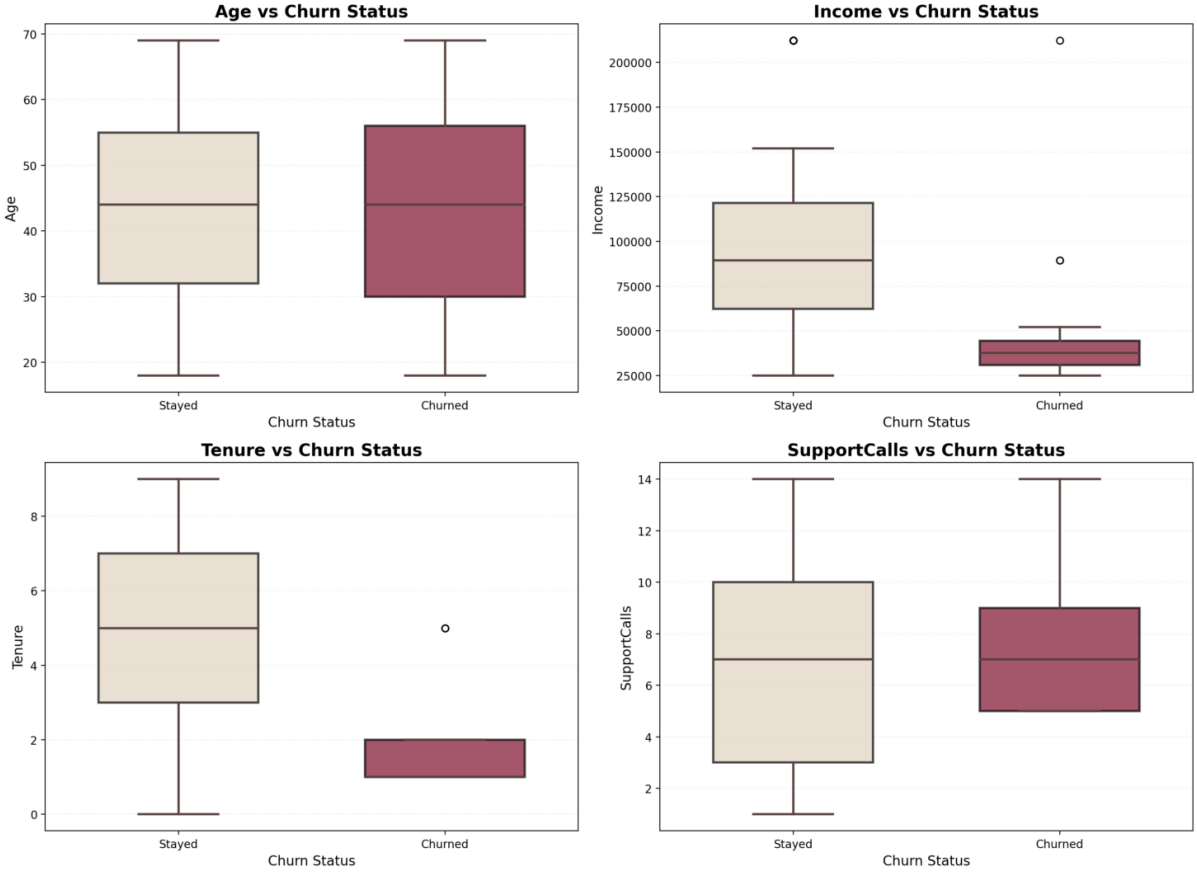

Figure 12: Numerical Box Plot

Scatter Plots

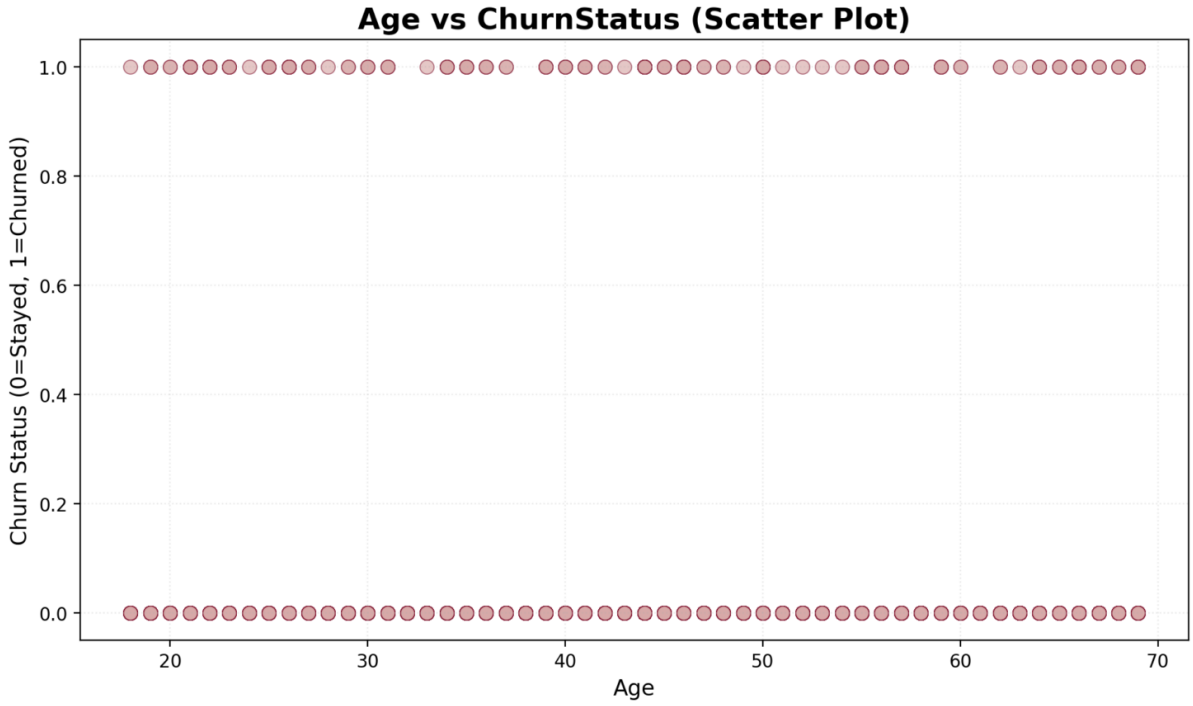

Figure 13: Age vs ChurnStatus

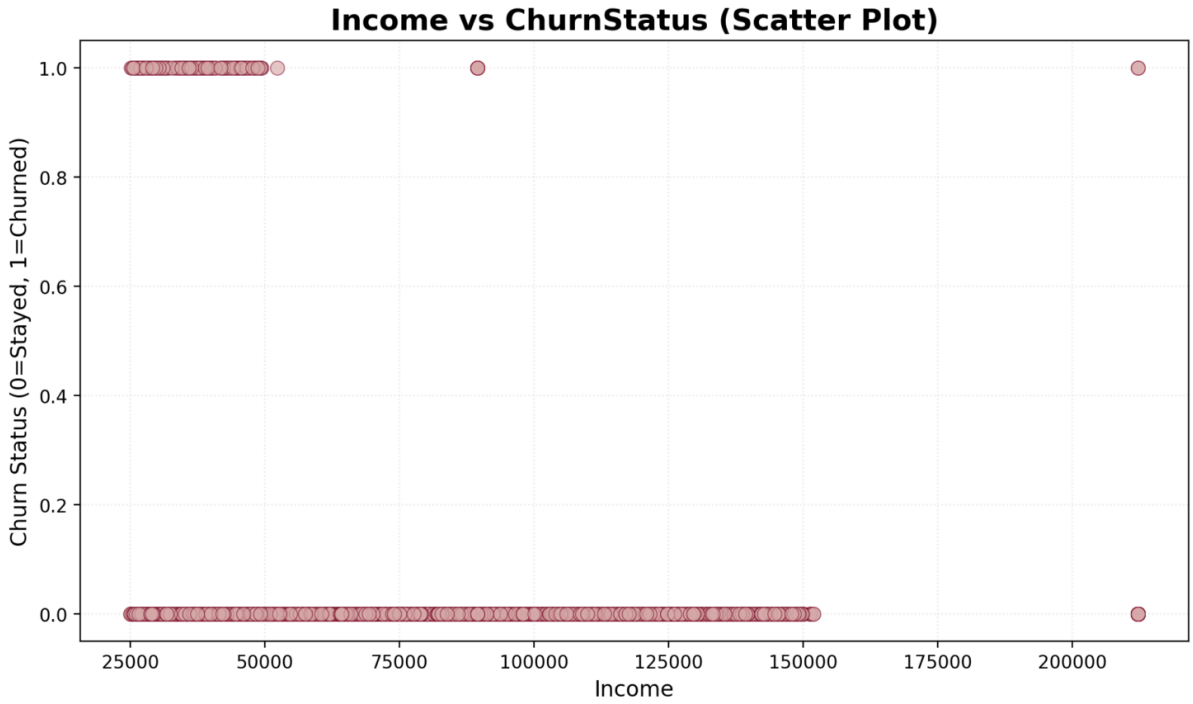

Figure 14: Income vs ChurnStatus

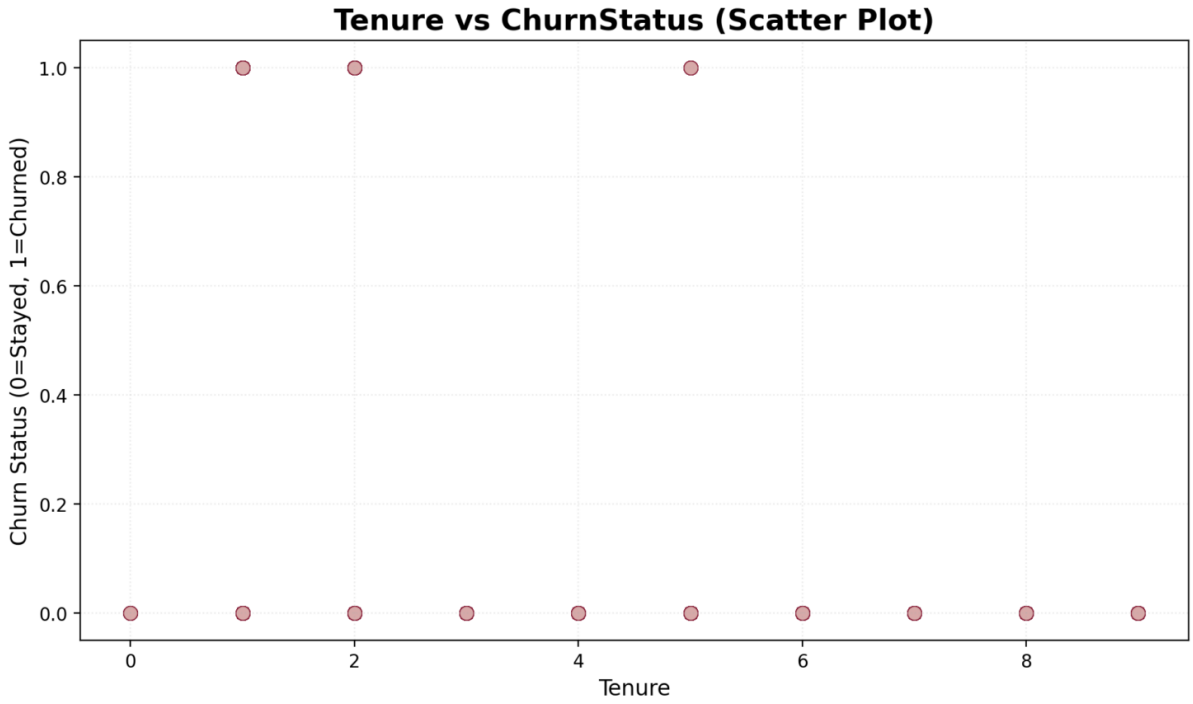

Figure 15: Tenure vs ChurnStatus

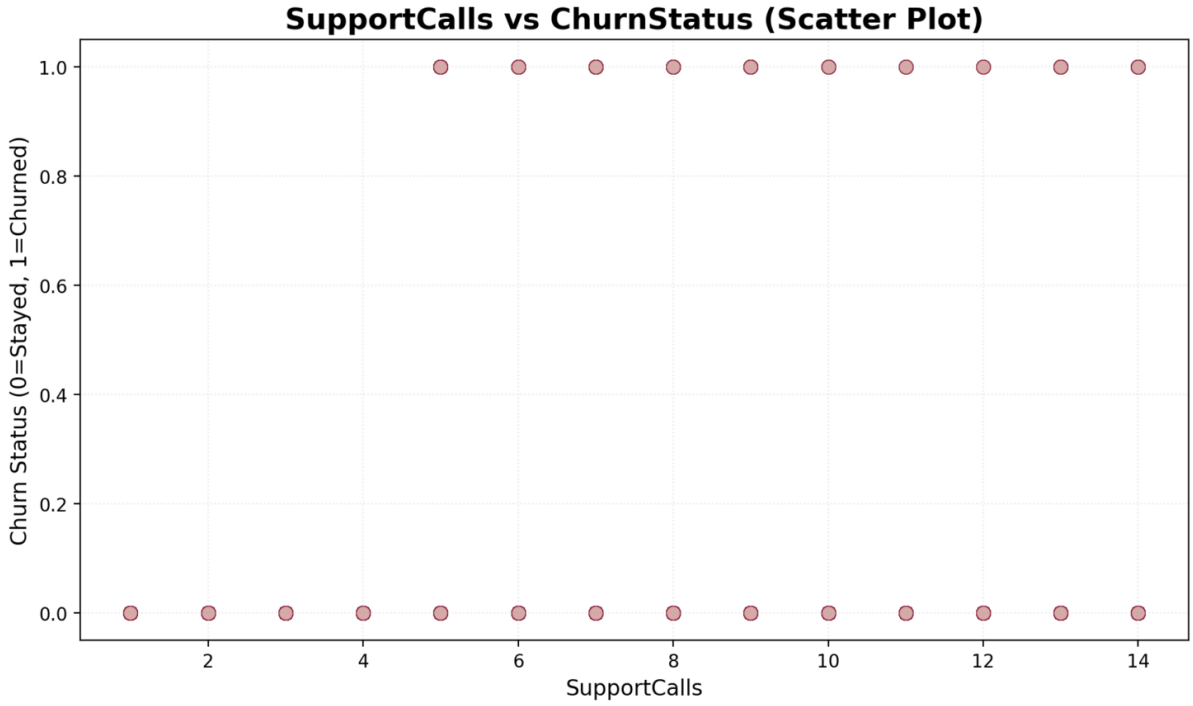

Figure 16: SupportCalls vs ChurnStatus

Statistical Compression

Table 17: Statistical Compression

Feature	Stayed (Mean)	Churned (Mean)	Difference	Stayed (Median)	Churned (Median)
Age	43.65	43.52	0.13	44	44
Income	92,423.11	41,014.77	51,408.34	89,510	37,676
Tenure	5.20	1.62	3.58	5	1
SupportCalls	7.04	7.59	0.54	7	7

Correlation Analysis

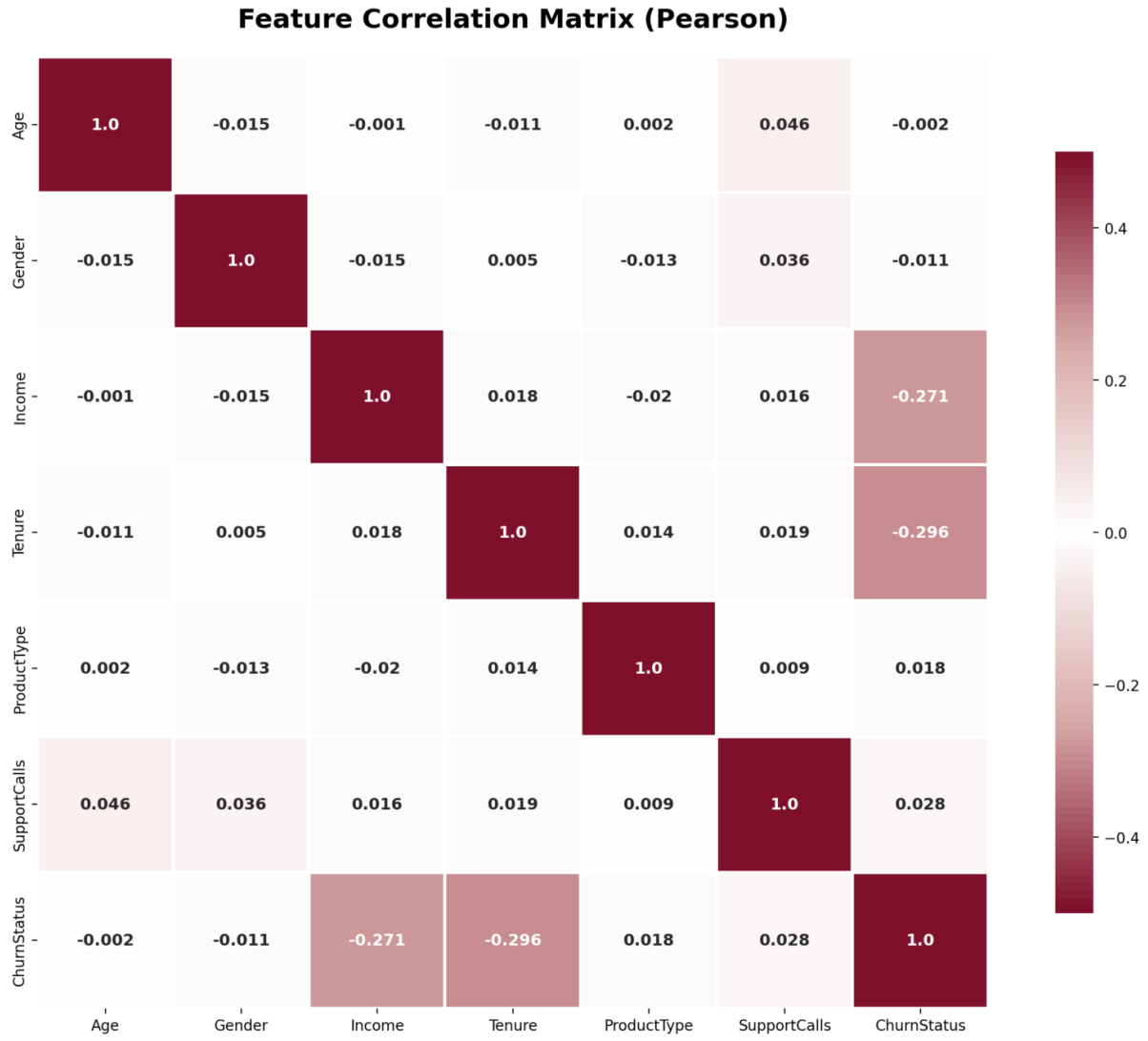

Figure 17: Feature Correlation Matrix

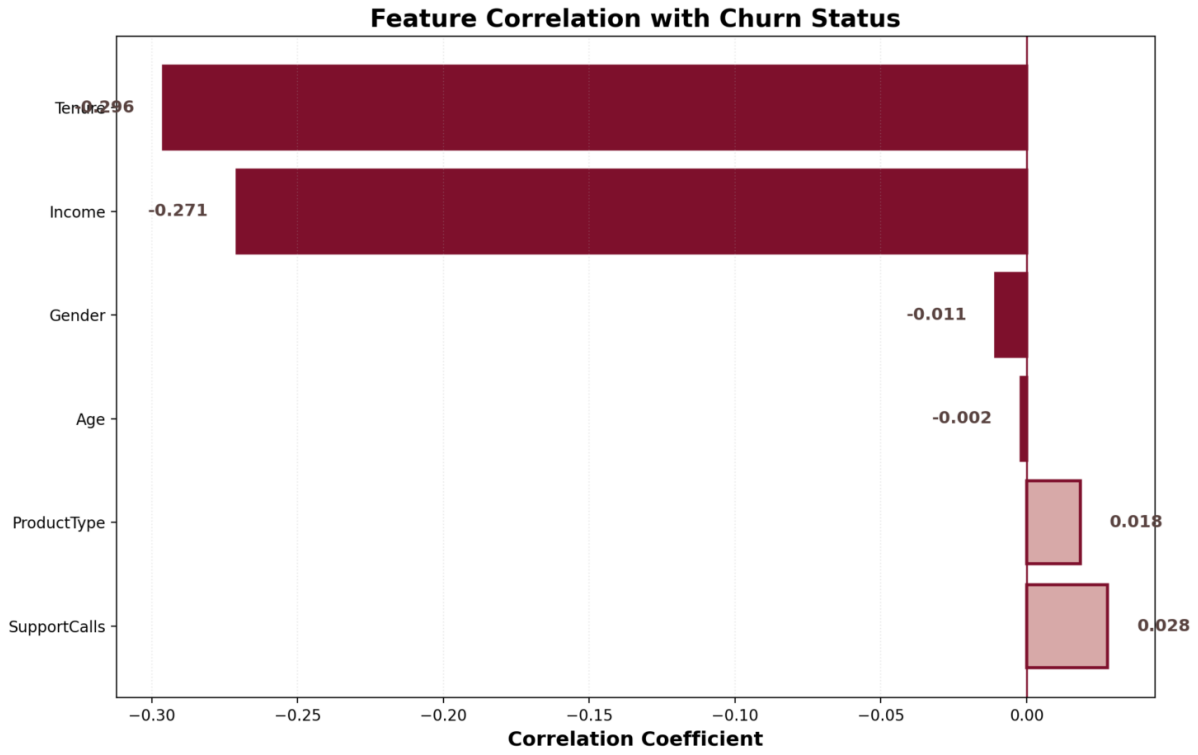

Figure 18: Feature Correlation With Churn States

Table 18: Feature Correlation Summary

Feature	Correlation with Churn	Strength
Tenure	-0.296	Weak
Income	-0.271	Weak
SupportCalls	0.028	Weak
ProductType	0.018	Weak
Gender	-0.011	Weak
Age	-0.002	Weak

Most Predictive Features for Churn:

- **1. Tenure:** Correlation = -0.296
- **2. Income:** Correlation = -0.271
- **3. SupportCalls:** Correlation = 0.028
- These features should be prioritized in predictive modeling.

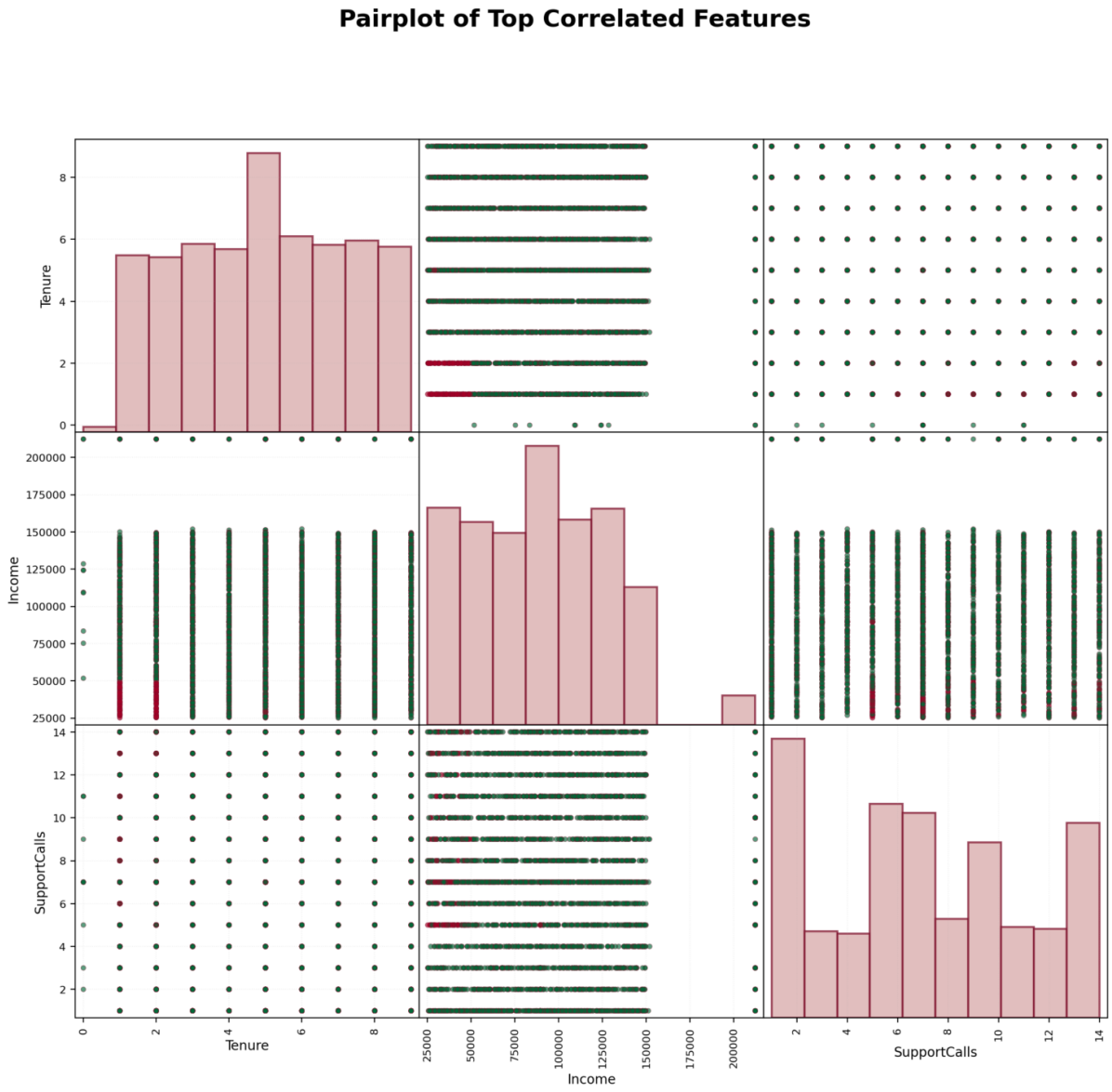

Figure 19: Top Correlation Features

Key Insights from Feature Pair Interactions:

- **Income X Tenure:** If both are low the risk of churn is higher as there is a clear red box that outlines this relation.

Conclusion

In conclusion, the Tenure is the top feature that determines whether a customer will churn. From the data the churn rate in general wasn't too high to begin with, which means that there isn't a high risk or a need to take urgent action. However, enhanced onboarding and engagement programs can help decrease this further as the customers with the low tenure are the ones that churn.

Visit our website:

Customer Churn Analysis