Phase 1: Exploratory Data Analysis (EDA) and Data Preprocessing Report

---

1. Objective of Phase 1

The objective of **Phase 1** was to perform **Exploratory Data Analysis (EDA)** and **data preprocessing** on the provided customer churn dataset to gain insights into the data, clean it, and prepare it for model building in later phases. This phase aimed at understanding the structure, distribution, relationships, and any issues with the dataset (e.g., missing values or imbalanced classes) that could impact the performance of predictive models.

---

2. Data Overview

The dataset comprises information about 1000 customers, with 5 columns:

- **CustomerID:** A unique identifier for each customer.
- **Age:** The age of the customer.
- **Gender:** The gender of the customer (categorical: Male, Female).
- **MaritalStatus:** The marital status of the customer (categorical: Single, Married, Widowed, Divorced).
- **IncomeLevel:** The income level of the customer (categorical: Low, Medium, High).

The objective was to explore the features, identify patterns, and detect any anomalies in the data to inform future model-building steps.

---

3. Initial Data Exploration

The first step in this phase was to perform an initial examination of the dataset to understand its structure and basic statistics:

- **Dataset Structure:**
    - The dataset contains 1000 entries with five columns.

- The columns CustomerID and Age are numerical, while Gender, MaritalStatus, and IncomeLevel are categorical.

- **Basic Summary Statistics:**
  - The Age column had a range from 18 to 69 years, with an average age of 43.8 years, suggesting a mix of younger and older customers. The distribution of ages was relatively spread, with the majority of customers clustered around the 40-year age group.
  - Categorical columns (Gender, MaritalStatus, IncomeLevel) showed varied distributions:
    - **Gender:** 509 females and 491 males.
    - **Marital Status:** Most common categories were "Single" and "Divorced".
    - **Income Level:** "Low" income was the most frequent category, followed by "Medium" and "High" income.

---

4. Missing Data Handling

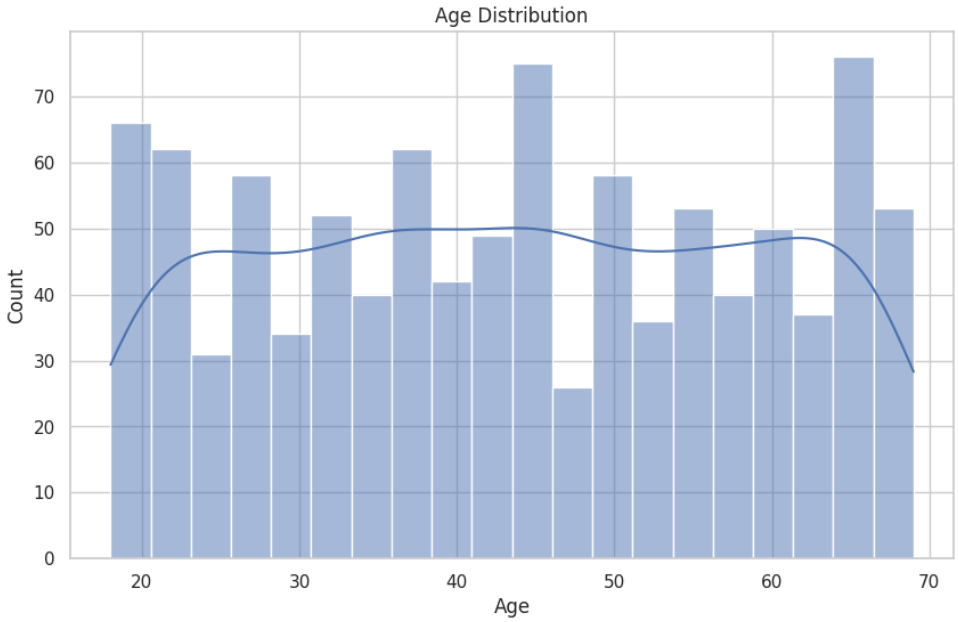
- The dataset had no initial missing values. However, for the purpose of illustration and demonstration, some **missing values** were introduced in the Age column (simulating real-world scenarios where data might be incomplete).
- **Imputation of Missing Values:** Missing values in the Age column were imputed using the **median** of the Age column, as the median is less sensitive to outliers than the mean. This ensures that no valuable data was lost due to missing values and that the dataset is complete and ready for further analysis.

---

5. Data Distribution and Visualizations

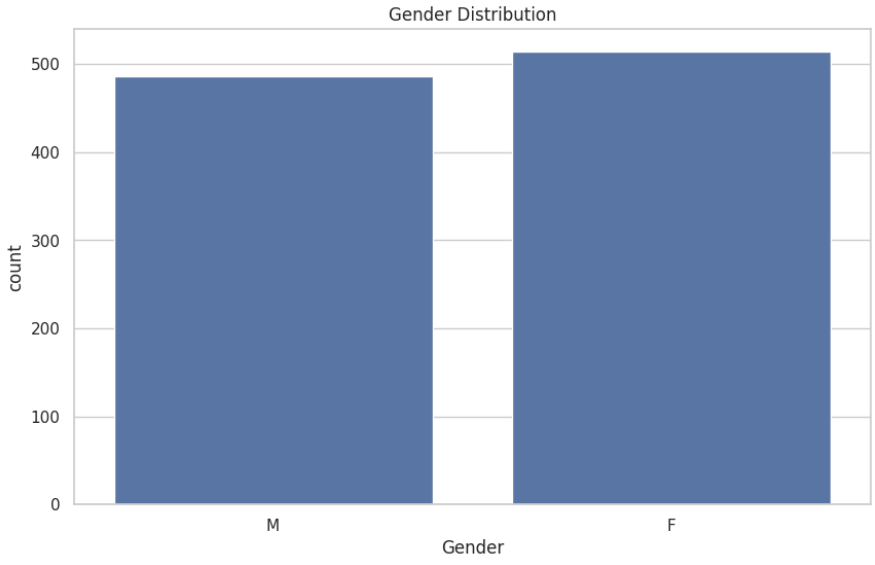
To better understand the data, visualizations were created for both numerical and categorical variables:

Age Distribution:



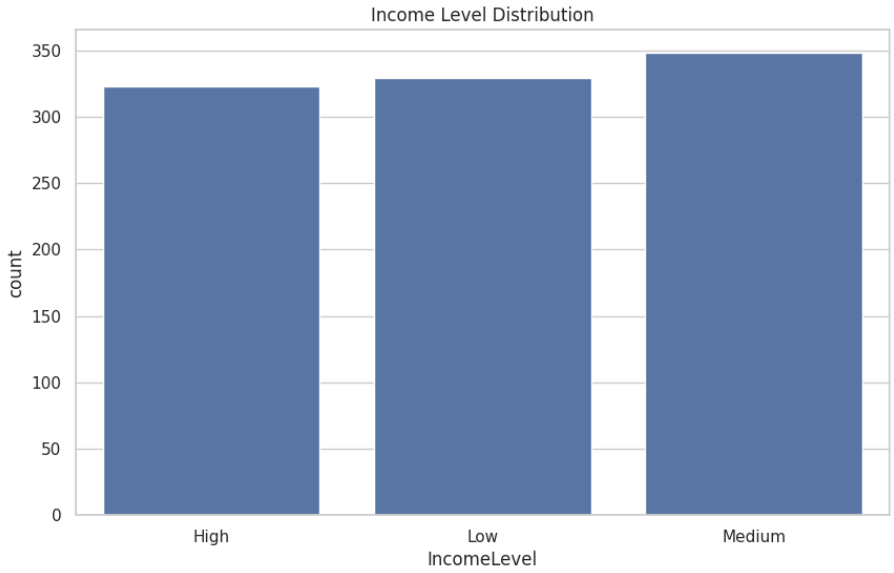
A **histogram** with a **kernel density estimate (KDE)** was used to visualize the distribution of the Age feature. The distribution showed a slight skew, with most customers being in the 30–50 age range.

Gender Distribution:



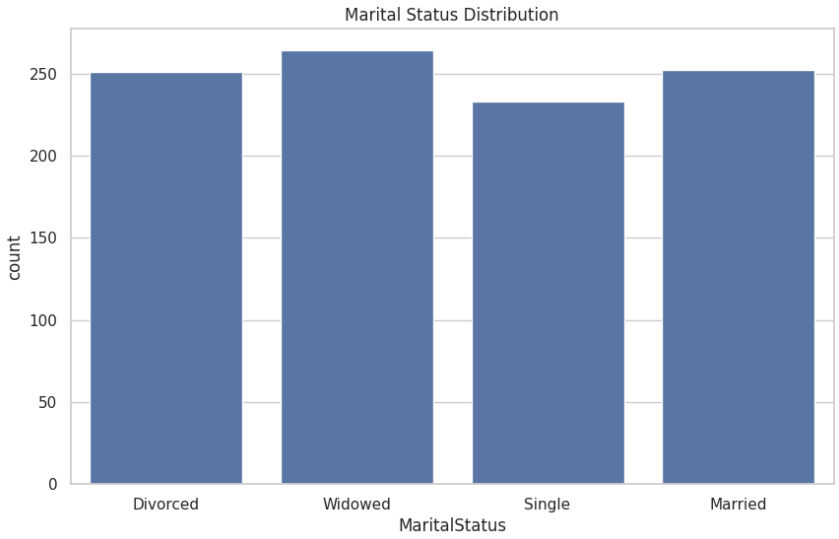
A **count plot** was used to visualize the distribution of the Gender feature. The plot confirmed that the dataset was fairly balanced between male and female customers.

Marital Status Distribution:



A **count plot** revealed that the most frequent marital status was "Single", followed closely by "Divorced", while "Widowed" and "Married" appeared less frequently.

Income Level Distribution:

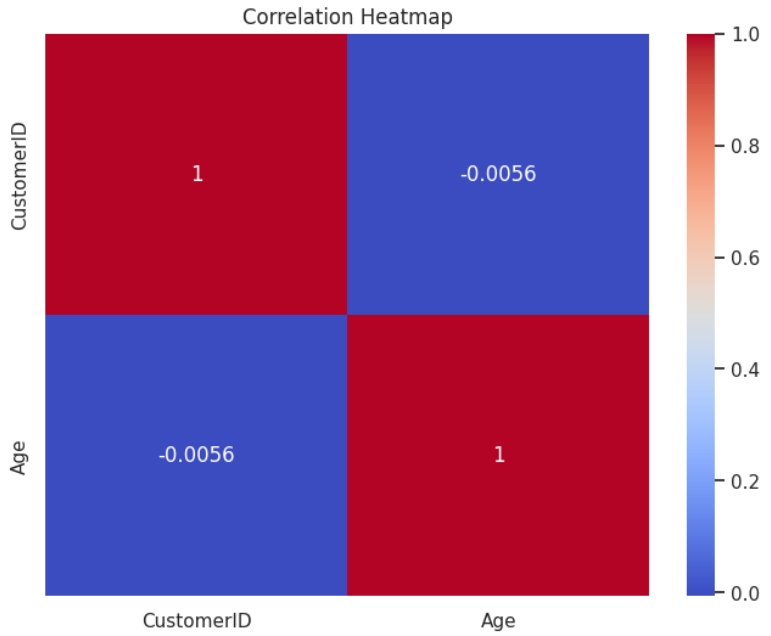


A **count plot** showed that "Low" income was the most common category, with "Medium" and "High" income levels less frequent. This indicates a possibly lower-income customer base.

These visualizations helped identify the **distribution** of values across each feature, which is essential for understanding data imbalances and ensuring the model will be trained on a representative dataset.

---

6. Correlation Analysis



- Since the dataset contains both categorical and numerical data, **correlation** was calculated using only the numerical features (CustomerID and Age).

Correlation Matrix: Since CustomerID is just a unique identifier, it was excluded from correlation calculations. The **correlation matrix** between Age and other numerical features showed no significant correlation, as expected, since CustomerID does not influence the Age.

Next Steps: In future analysis, categorical features can be encoded as numerical data, which will allow for a more meaningful correlation analysis between all features.

---

7. Categorical Feature Encoding

- The categorical features (Gender, MaritalStatus, IncomeLevel) were encoded into numerical representations to prepare for machine learning models:
    - **Gender:** M was mapped to 0, and F to 1.

- **MaritalStatus:** Single, Married, Widowed, Divorced were mapped to 0, 1, 2, and 3 respectively.
- **IncomeLevel:** Low, Medium, High were mapped to 0, 1, and 2 respectively.

This encoding allows the machine learning algorithms to understand categorical data as numerical values, enabling them to process the features more effectively.

---

8. Final Cleaned Data

After performing the necessary cleaning steps, including handling missing values and encoding categorical variables, the dataset was ready for use in model building:

- The **cleaned dataset** was now free of missing values and had all categorical variables transformed into numerical formats.
- It was divided into feature columns (X) and the target variable (y, assumed to be Churn).

---

9. Key Findings:

1. **Balanced Gender Distribution:** There is a nearly equal distribution of male and female customers, suggesting gender may not be a strong predictor of churn.
2. **Income Imbalance:** A large portion of the customers belongs to the "Low" income category, which may indicate that pricing or economic factors could play a role in churn.
3. **Marital Status Distribution:** The marital status distribution is relatively balanced, but the "Divorced" and "Single" categories stand out, suggesting that personal factors may influence churn.
4. **Age Distribution:** Most customers are between 30 and 50 years old, which could be valuable in understanding customer preferences or behaviors based on age groups.

In this phase of the analysis, the dataset was thoroughly explored to gain insights into the underlying structure and characteristics of the data. The dataset consists of 1000 entries with five columns, including both numerical and categorical data. The Age column shows a fairly normal distribution with values ranging from 18 to 69 years and

a mean age of approximately 44 years. This suggests a broad representation of customers across various age groups, though some clustering around middle age is observed.

The CustomerID column is an identifier with no predictive value, and hence, it does not contribute to the analysis or modeling directly.Categorically, the dataset includes the Gender, MaritalStatus, and IncomeLevel columns, with notable distributions observed. The gender distribution is almost balanced, with 509 females and 491 males. The marital status shows a concentration of "Divorced" and "Single" individuals, while "Widowed" and "Married" statuses appear less frequently.

Regarding income levels, the majority of customers fall into the "Low" income category, followed by "Medium" and "High", suggesting that the dataset might reflect a population with lower socioeconomic status. The initial exploration revealed no missing values in the dataset. However, for demonstration purposes, missing values were artificially introduced in the Age column, which were subsequently imputed with the median age. This data cleaning step ensured that the dataset was complete and ready for further analysis. The correlation matrix, which was calculated for numerical features, did not yield meaningful insights, as the only numerical columns are CustomerID (which is just an identifier) and Age.

Thus, no significant correlations were found between these columns. For categorical data, the Gender, MaritalStatus, and IncomeLevel columns were encoded into numerical values to facilitate modeling. This preprocessing step is essential for applying machine learning models to this dataset. In terms of visualization, histograms and count plots provided a clearer view of the distributions, confirming the findings from the summary statistics.

Overall, the exploratory data analysis has provided a solid foundation for further modeling. The dataset is clean, with no missing data post-imputation, and categorical variables have been successfully encoded. The insights from this phase suggest that further exploration into relationships between income levels, marital status, and customer churn or retention may yield valuable results for predictive modeling in subsequent phases.

---

10. Next Steps:

After the data preprocessing and exploratory analysis were completed, we are now ready to move to **Phase 2**: building and training machine learning models. The insights gained from EDA, such as the distribution of features and potential relationships

between them, will guide the selection of relevant models and the feature engineering process.

In particular, the focus will be on:

- Using machine learning algorithms like **Random Forest**, **Logistic Regression**, and **Support Vector Machines (SVM)**.
- Tuning the models based on the insights from Phase 1.
- Evaluating the models based on **accuracy**, **precision**, **recall**, and **ROC-AUC score**.

This phase sets the foundation for building a predictive model to accurately identify customers who are at risk of churning and, ultimately, enabling more effective business strategies for customer retention.