**Customer Churn Analysis: Data Preparation Report**

**1. Dataset Summary & Selection Rationale**

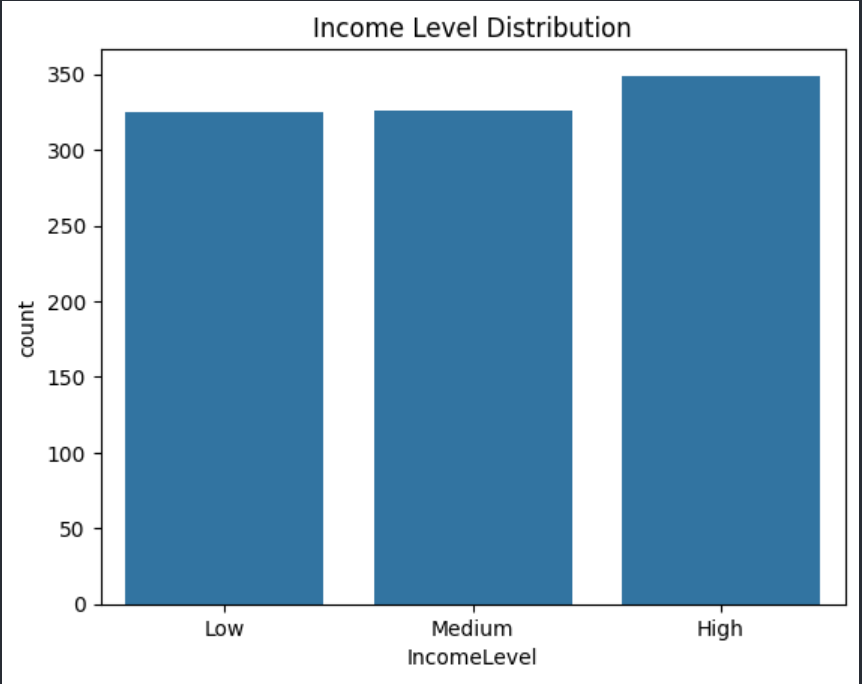
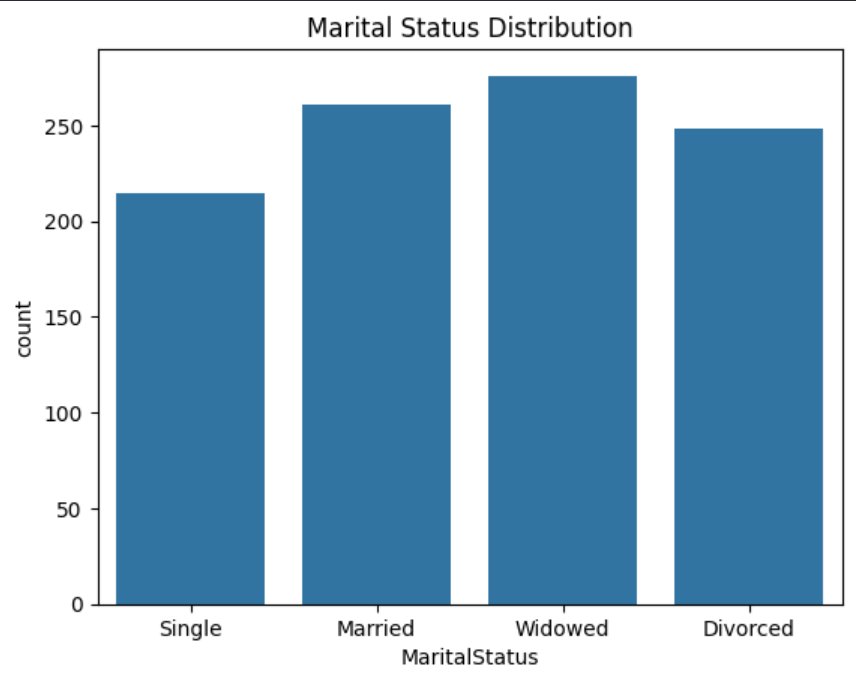
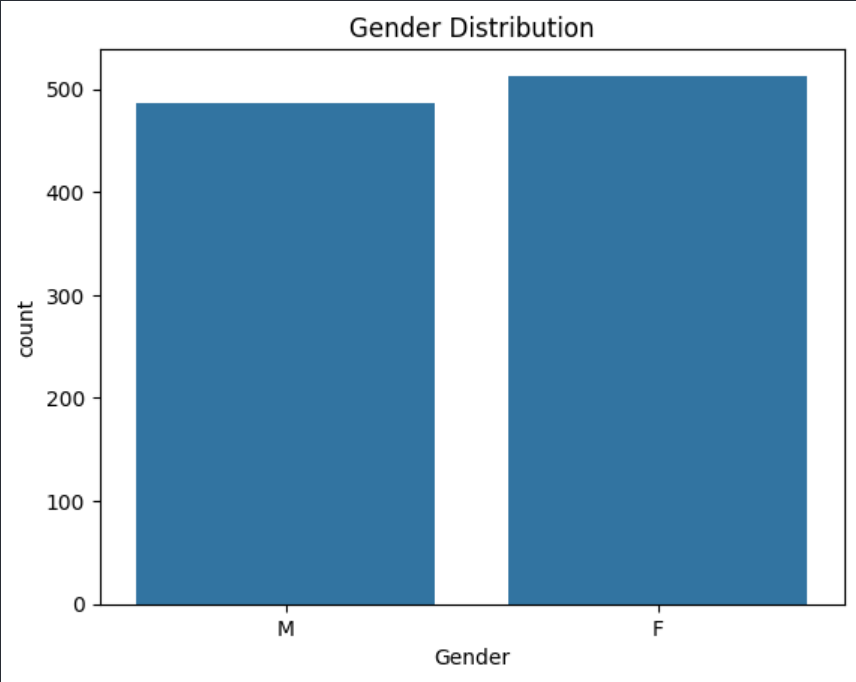
The dataset used in this analysis consists of 1000 customer records. These features were selected because customer demographics often influence churn behaviour. For example, age, marital status, and income can reflect customer preferences, lifestyle, or financial capability—key indicators for predicting churn likelihood.

**2. Exploratory Data Analysis (EDA)**

EDA was conducted to understand variable distributions and relationships with churn. Here are key insights and visualizations:

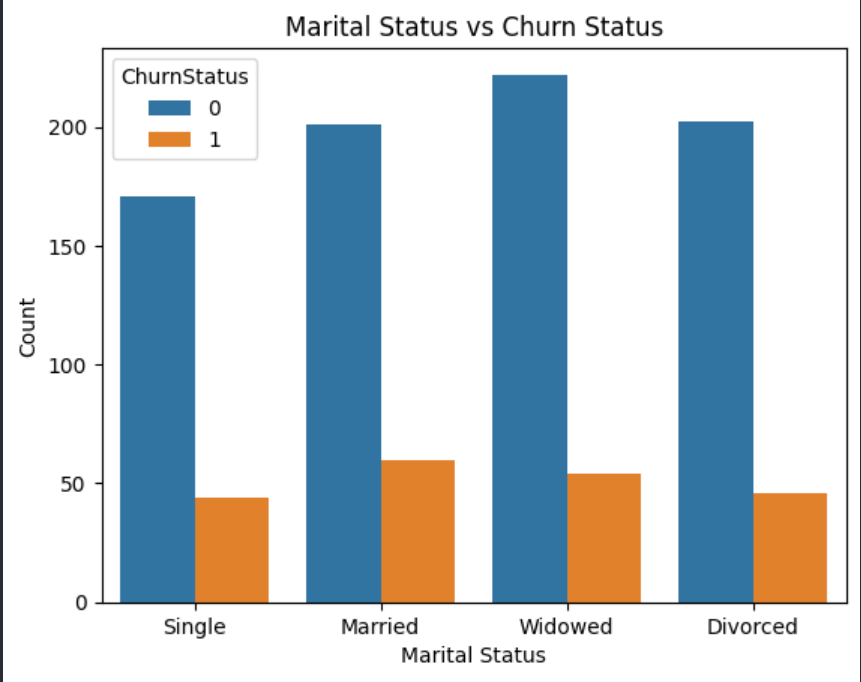
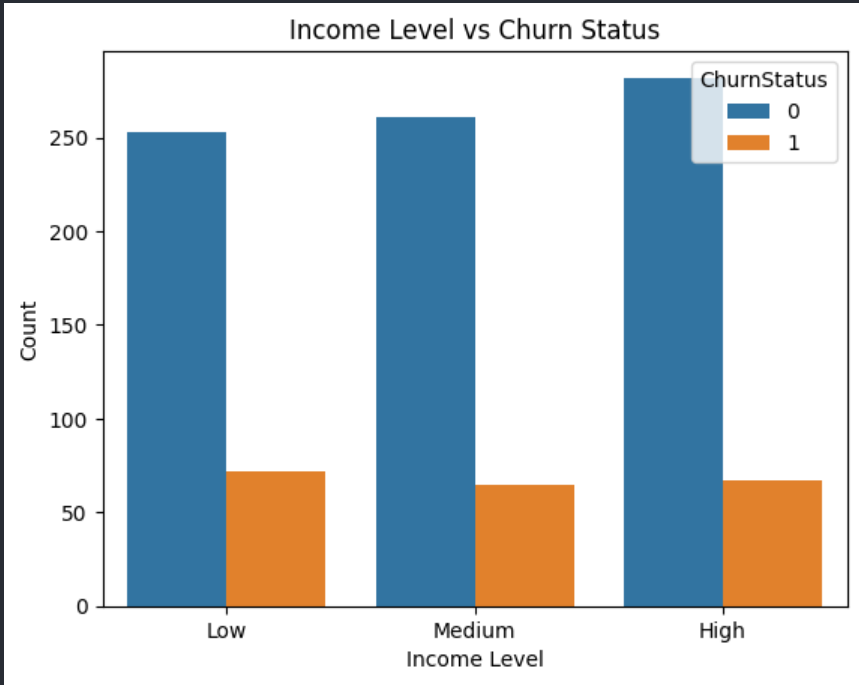
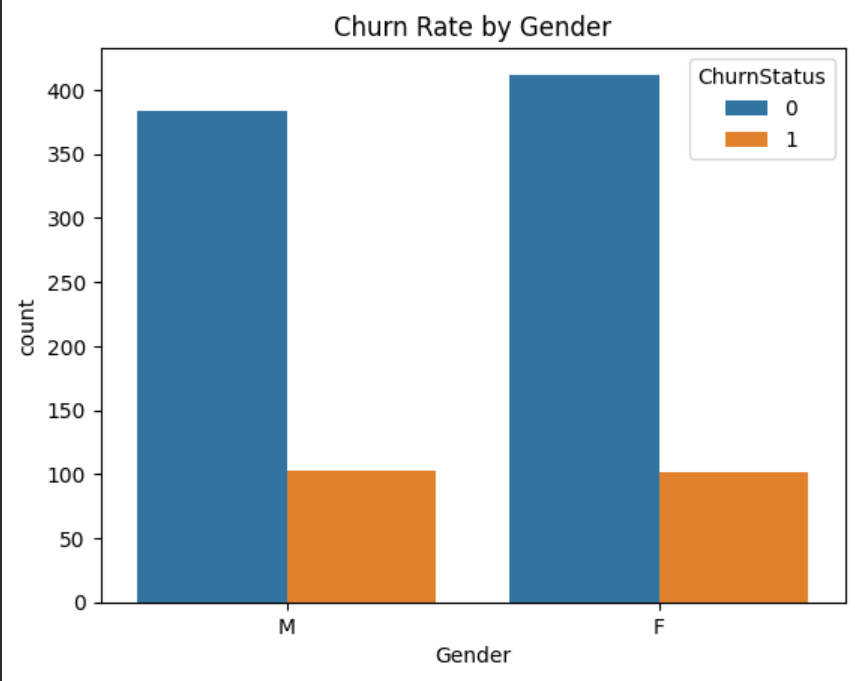
**a) Univariate Analysis**

* **Age**: Ranges from 18 to 69, with a fairly normal distribution.
* **Gender**: Slightly more females than males.
* **Marital Status**: Most customers are *Widowed*, followed by *Married*.
* **Income Level**: Majority fall into *High* income groups.



b) **Bivariate Analysis with “churnstatus”**

* Gender vs Churn – Slightly more females churned than males.
* Marital Status vs Churn - Higher churn observed among *Single* and *Divorced* customers.
* Income Level vs Churn - *Low income* customers show a higher churn rate.
* Age vs Churn (Boxplot) - Younger customers are more likely to churn.



3. **Data Cleaning & Preprocessing Steps**

Missing Values

* No missing values were found in the dataset.

Outliers

* Age values were within expected ranges (18–69), so no major outliers were removed.

Encoding Categorical Variables

* One-hot encoding was applied to convert categorical columns into numerical form

Standardization

* The Age column was standardized using StandardScaler to ensure uniform scale

**4. Final Output: Cleaned Dataset**

* The dataset now contains the target column churnstatus, standardized numerical features, and encoded categorical features.
* Ready for model development.