# Code Report

## Introduction

This code focuses on data preprocessing and analysis tasks, including identifying missing values, detecting and handling outliers, and visualizing data distributions. The primary objective is to clean and prepare the dataset for further machine learning or statistical modeling. The implementation employs Python's **pandas** and **matplotlib** libraries, ensuring efficient data manipulation and visualization.

## Methodology

### 1. Handling Missing Values:

- The code checks for missing values in each column using “isnull()” and counts them with “sum()”.
- Missing values are filled based on column data types:
  - Numerical columns: Replaced with the column's mean.
  - Categorical columns: Filled with a placeholder value such as “Unknown”.

### 2. Outlier Detection:

- Outliers are identified using the Interquartile Range (IQR) method:
  - IQR = Q3 - Q1
  - Thresholds for outliers:
    - Lower bound: Q1 - 1.5*IQR
    - Upper bound: Q3 + 1.5*IQR
- Outliers are filtered and counted using the logical condition df[(df[col] < lower) | (df[col] > higher)].

### 3. Visualization:

- Histograms and boxplots are generated for visualizing data distribution and identifying skewness and outliers.

### 4. Experiments

The code performs the following experiments:

- *Skewness Analysis:*

- Skewness is calculated using **df[col].skew()**, indicating whether the data distribution is symmetric or skewed (left or right).

- Histograms visually represent the distribution of numerical columns.

- *Boxplot Analysis:*

- Boxplots are generated for numerical columns grouped by categorical variables to show variability and detect outliers.

- *Handling Missing Data:*

- Columns are processed to replace missing values dynamically based on their type, ensuring no gaps in the dataset.

## Results & Discussion:

*1. Missing Values:*

- Missing values in numerical columns were successfully replaced by their respective means, reducing the risk of bias caused by gaps in data.

- Categorical columns were filled with “Unknown”, providing a consistent placeholder.

*2. Outlier Detection:*

- Several outliers were detected based on IQR thresholds. These can either be removed or handled depending on their impact on the analysis.

*3. Skewness:*

- Skewness values were calculated, and histograms revealed that some features exhibited right-skewed distributions.

*4. Visualizations:*

- Histograms and boxplots effectively highlighted data characteristics, including the spread, central tendency, and presence of outliers.

## Conclusion

This code successfully addresses common data preprocessing challenges like missing values, outliers, and data distribution analysis. The methodology ensures clean and consistent data suitable for advanced analytical tasks. Future improvements could include automating feature transformation for skewed data and dynamic detection of categorical variables for visualization.