Report on Data Preprocessing and Exploratory Data Analysis (EDA)

1. Introduction

This report summarizes the data preprocessing procedures and exploratory data analysis (EDA) methods performed on the dataset. The major aim is to clean and prepare the data for subsequent analysis by managing missing values, dropping duplicates, dealing with outliers, and performing statistical and visual explorations.

2. Data Preprocessing

Handling Missing Values:

The dataset was read using pandas.read\_csv().

Missing values were detected using df.isnull().sum().

Rows with missing values were dropped using df.dropna(inplace=True).

Columns with over 30% missing values were also dropped using df.dropna(axis=1, thresh=len(df) \* 0.7, inplace=True).

Removing Duplicate Records:

Duplicate records were detected using df.duplicated().sum().

Duplicate rows were dropped using df.drop\_duplicates(inplace=True).

The dropping was checked using df.duplicated().sum().

Outlier Treatment:

Outliers were plotted using box plots through sns.boxplot().

Outliers were dropped using IQR (Interquartile Range) filtering:

Q1 = df.quantile(0.25)

Q3 = df.quantile(0.75)

IQR = Q3 - Q1

df = df[~((df < (Q1 - 1.5 \* IQR)) | (df > (Q3 + 1.5 \* IQR))).any(axis=1)]

Standardizing Categorical Values:

Categorical values were standardized by using df['column\_name'] = df['column\_name'].str.lower().str.strip().

One-hot encoding was used by pd.get\_dummies().

3. Exploratory Data Analysis (EDA)

Univariate Analysis (Single-Variable Exploration)

Summary statistics such as mean, median, mode, variance, and skewness were obtained by df.describe().

Frequency distributions were examined for categorical variables by using df['column\_name'].value\_counts().

Histograms and box plots were plotted by sns.histplot() and sns.boxplot().

Bivariate Analysis (Two-Variable Exploration)

A correlation matrix was calculated by df.corr() and plotted with sns.heatmap().

Scatter plots were created with sns.scatterplot() to visualize relationships between numerical variables.

Bar plots and violin plots were produced with sns.barplot() and sns.violinplot() to compare categorical and numerical variables.

Multivariate Analysis (Multiple-Variable Exploration)

Pair plots were created using sns.pairplot(df) to explore multiple relationships at once.

Heatmaps were employed to represent correlations between multiple variables with sns.heatmap(df.corr(), annot=True).

Grouped comparisons were done using df.groupby(['category\_column']).mean().

4. Conclusion

The preprocessing of data and EDA steps have served to clean and organize the dataset for subsequent analysis. The findings obtained from univariate, bivariate, and multivariate explorations will be informative in informing future modeling and decision-making activities.