Exploratory Data Analysis (EDA) Report:

1. Introduction:

Objective:

This report presents an exploratory data analysis (EDA) on a dataset obtained from Kaggle. The goal is to perform data cleaning, univariate, bivariate, and multivariate analysis to uncover meaningful insights.

Dataset Description:

The selected dataset contains multiple numerical and categorical variables representing different attributes. The dataset includes missing values, duplicate entries, and potential outliers, which have been handled accordingly.

2. Data Cleaning:

Loading and Inspecting Data

The dataset was loaded using `pandas`, and its structure was examined using `.info()` and `.describe()`.

Handling Missing Values:

Numerical Data: Missing values were imputed using the **mean** or **median**.
Categorical Data: Missing values were filled using the **mode**.

Removing Duplicates

Duplicate records were identified using `.duplicated()` and removed accordingly.

Outlier Detection and Treatment

Outliers in numerical variables were detected using box plotsand the IQR method.
Treatment strategies included removal or transformation (log scaling).

Standardizing Categorical Values

Inconsistent categorical values (e.g., typos) were standardized.

3. Exploratory Data Analysis (EDA)

Univariate Analysis

Summary Statistics

Mean, median, mode, variance, and skewness** were calculated for numerical variables.

Visualizations

Histograms: Distribution of each numerical feature.
Box Plots: Identify potential outliers.
Frequency Plots: Count distributions for categorical variables.

Bivariate Analysis

Correlation Analysis

Pearson correlation matrix to assess relationships between numerical variables.
Scatter plots to examine variable relationships.

Comparisons
Box plots and violin plots to compare numerical variables across categories.
Bar plots to visualize categorical vs. numerical relationships.

Multivariate Analysis
Pair Plots
Examined interactions between multiple numerical variables.

Heatmaps
Used `seaborn.heatmap()` to visualize correlations between multiple numerical variables.

Grouped Comparisons
Analyzed combined effects of multiple categorical and numerical variables.

4. Insights and Conclusions
Key Trends: Significant correlations and distribution patterns were identified.
Outliers: Certain variables exhibited extreme values, affecting overall distribution.
Categorical Analysis: Some categories had dominant values, influencing the dataset structure.
Multivariate Interactions: Identified important relationships between variables.

-