**Data Analysis and Exploration Report**

Report by Ayush

1. Introduction

This report provides an in-depth analysis of a dataset containing multiple numerical and categorical variables. The objective is to clean the data, handle missing values, remove duplicates, treat outliers, and conduct exploratory data analysis (EDA) using statistical and visualization techniques.

2. Data Cleaning

2.1 Data Loading and Inspection

- The dataset was loaded using Pandas, and its structure was inspected using `.info()` and `.describe()`.

- This step helped in understanding the number of records, data types, missing values, and basic statistical properties.

2.2 Handling Missing Values

- Numerical Variables: Missing values were imputed with the mean of the respective column.

- Categorical Variables: Missing values were replaced with the mode (most frequent value).

2.3 Duplicate Removal

- Duplicate records were identified and removed using `.drop\_duplicates()` to maintain data integrity.

2.4 Outlier Detection and Treatment

- The Interquartile Range (IQR) method was applied to identify and treat outliers.

- Outliers were replaced with the mean to prevent skewed analysis.

2.5 Standardizing Categorical Variables

- Categorical values were converted to lowercase and stripped of whitespace.

- Common typos in categorical labels were corrected.

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

- Summary statistics (mean, median, mode, variance, skewness) were computed for numerical variables.

- Frequency distributions of categorical variables were visualized using \*\*bar plots\*\*.

- \*\*Histograms and box plots\*\* were used to understand the data distribution and detect potential outliers.

3.2 Bivariate Analysis

- Correlation Matrix: A heatmap was generated to visualize relationships between numerical variables.

- Scatter Plots: These were used to identify linear and non-linear relationships between numerical variables.

- Box Plots & Violin Plots: These visualized numerical variable distributions across categorical groups.

3.3 Multivariate Analysis

- Pair Plots: Used to analyze relationships between multiple numerical variables.

- Advanced Heatmaps: Provided deeper insights into multicollinearity among variables.

- Grouped Comparisons: Box plots with hue differentiation were used to explore how categorical variables influence numerical distributions.

4. Findings and Conclusion

- The dataset contained missing values, which were successfully handled using imputation techniques.

- Outliers were treated using the IQR method to ensure accurate analysis.

- Correlation analysis provided valuable insights into feature relationships for potential predictive modeling.

- Univariate, bivariate, and multivariate analysis helped understand the data’s structure and key trends.