REPORT OF DAI-101 ASSIGNMENT 1

ANKIT KUMAR

Enrollment: 23116013

Department: ECE

Table of Content:

1. Introduction

In this report, we conduct an in-depth analysis of the dataset, data_ak.csv. The analysis includes data inspection, missing value treatment, outlier handling, univariate, bivariate, and multivariate analysis using statistical and visualization techniques.

---

2. Data Loading and Inspection

Dataset Overview

The dataset is loaded using Pandas, and an initial inspection is conducted using .info() and .describe() methods. This helps us understand:

- The number of entries (rows and columns).
- Data types (numerical vs categorical).
- Presence of missing values.
- Basic statistical properties such as mean, standard deviation, and range of numerical variables.

Handling Missing Values

- **Numerical Columns**: Missing values are replaced with the **mean** of the respective column.
- **Categorical Columns**: Missing values are replaced with the **mode** (most frequent category).

Duplicate Removal

To ensure data integrity, duplicate records are identified and removed using .drop duplicates().

Outlier Treatment

- The **Interquartile Range (IQR) method** is used to detect and replace outliers.
- Values falling outside **1.5 times** the IQR are considered outliers and replaced with the mean of the respective column.

Standardization of Categorical Variables

- Strings are converted to lowercase and stripped of whitespace.

- Common typos in categorical values are corrected.

---

3. Univariate Analysis

Univariate analysis focuses on analyzing individual variables to understand their distribution and properties.

3.1 Statistical Summary

- **Mean, median and mode** are calculated to understand central tendency.
- **Variance and skewness** are analyzed to measure dispersion and asymmetry in numerical variables.

3.2 Visualization

- **Histograms** are plotted to show the frequency distribution of numerical variables.
- **Boxplots** are used to detect outliers and analyze spread.
- **Bar charts** display the frequency of categorical variables.

---

4. Bivariate Analysis

Bivariate analysis examines the relationships between two variables.

4.1 Correlation Matrix

- A heatmap is generated to visualize correlations among numerical variables.
- Strong correlations indicate potential multicollinearity, useful for feature selection in modeling.

4.2 Scatter Plots

- Pairwise scatter plots of numerical variables help identify linear/non-linear relationships.

4.3 Categorical vs. Numerical Analysis

- **Bar plots** compare numerical variables across categorical groups.
- **Violin plots** show distribution shapes within categories.

- **Box plots** highlight variations and outliers.

---

5. Multivariate Analysis

Multivariate analysis extends beyond two variables to explore deeper relationships.

5.1 Pair Plots

- Pairwise scatter plots combined with kernel density estimation (KDE) help visualize multiple variable interactions.

5.2 Advanced Correlation Heatmap

- A refined heatmap further investigates multicollinearity.

5.3 Grouped Comparisons

- **Box plots with hue differentiation** show how multiple categorical features impact numerical variables.

---

6. Conclusion

Key Insights:

- Missing values were successfully handled, ensuring data completeness.
- Outlier treatment was applied to mitigate extreme value influence.
- Univariate analysis highlighted skewness and dominant categories.
- Bivariate analysis revealed strong correlations and category-based variations.
- Multivariate analysis provided deeper insights into variable interactions.