Assignment-2

Name: Shreyan Reddy Gangwar

St ID: 12599073

A: This data preparation task distinguishes between category and numerical columns in order to handle missing values in a dataset. To preserve the current data distribution and prevent any bias, missing values in categorical columns are imputed using the mode, the most often occurring value in the appropriate category. To maintain the distribution and central tendency of the numerical data, missing values in numerical columns are imputed using the mean, the average value of accessible data points in that column. This strategy increases data integrity, improves the dependability of downstream analysis and modeling efforts, and leads to more accurate and robust data-driven decision-making outcomes.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

B: Standardizing numerical attributes in columns such as "Mileage," "Engine," "Power," and "New\_price" by deleting units or non-numeric characters is part of the data pretreatment stage. This enables uniform analysis while also simplifying processes such as computations, visualizations, and modeling. The cleaned data is stored as "clean\_data.csv" and improves data quality by preserving a structured dataset for meaningful insights and decreasing mistakes caused by inconsistencies in data representations.

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

C: The data preprocessing step involves encoding categorical variables "Fuel\_Type" and "Transmission" into numerical values, making them compatible with machine learning algorithms and data analysis techniques. The modified dataset, saved as "dataset\_with\_one hot \_encoded.csv", allows for comprehensive analysis and more accurate predictions, classifications, and correlations. Encoding these variables ensures the full potential of the dataset in data-driven decision-making and model building, ensuring all relevant information is effectively utilized.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

D: Loading a previously encoded dataset and adding a new column named "Current\_Age" is the data modification task. This column denotes an entity's age at the time of data collection or analysis. The resultant dataset, stored as "data\_with\_currentage\_results.csv," improves the dataset's completeness and usability, allowing for more extensive age-related research and insights. The inclusion of the "Current\_Age" variable allows for more robust data-driven decision-making and modeling, resulting in a more comprehensive and useful dataset.

A screenshot of a computer

Description automatically generated