## **ASSIGNMENT-2**

## **Name: ENUKONDA SAI DEEPTHI STUDENT ID:16343756**

**1) The provided data contains various details and attributes associated with used cars. The target variable, which is the central focus of analysis, is the price of the used cars, and it is measured in lakhs. The data in this dataset is tabular, with rows and columns, where each row represents a specific used car listing, and each column represents a particular attribute or feature of these cars.  Features are Make and model of the car, Location or city of sale, Year of manufacture, Mileage, Odometer (kilometers driven), Fuel type (petrol or diesel), Transmission type (manual or automatic), Number of owners, Engine displacement, Engine horsepower, Number of seats, and Price when the car was new.**

**a)  Look for the missing values in all the columns and either impute them (replace with mean, median, or mode) or drop them. Justify your action for this task.**

In the given data, missing values are systematically identified using the **isnull().sum()** function, offering a meticulous approach to comprehensively recognize gaps in the dataset. This process emphasizes the importance of handling missing values either by imputation using techniques like mean, median, or mode, or by removing them, with the choice justified based on the dataset's context and specific affected columns.

Preserving data integrity is paramount, identifying missing values prevents skewed analyses due to incomplete data, ensuring the dataset's reliability. Imputation maintains statistical properties, enabling informed analysis when missing data is random or limited, without compromising overall dataset integrity. Removing missing values becomes essential when accuracy is significantly impacted, guaranteeing analyses on complete, unbiased data. Emphasizing justification enhances decision-making, becoming the foundation for subsequent analyses and modeling. This deliberate approach ensures the validity and trustworthiness of the findings, reinforcing the importance of a systematic and justified handling of missing data.

**b) Remove the units from some of the attributes and only keep the numerical values (for example remove kmpl from “Mileage”, CC from “Engine”, bhp from “Power”, and lakh from “New\_price”).**

Removing units from numerical attributes is a critical preprocessing step in data analysis, offering several key advantages. First and foremost, it ensures consistency and uniformity by standardizing numerical values, simplifying comparisons and computations across the dataset. This uniformity is essential for accurate statistical analyses and model training. Additionally, eliminating units enhances data interpretability, making visualizations and summaries clearer for both analysts and stakeholders. Moreover, in the context of machine learning, models perform optimally when working with uniform, numeric inputs. Removing units facilitates streamlined data processing, enabling models to identify patterns accurately and make reliable predictions. In summary, this preprocessing step not only enhances data quality and comparability but also significantly impacts the accuracy and reliability of subsequent analyses and predictive modeling, making it a vital practice in data preparation.

**C) Change the categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hot encoded value.**

 Converting categorical variables like "Fuel\_Type" and "Transmission" into numerical one-hot encoded values is a crucial step in data preprocessing, particularly for machine learning tasks. This transformation is justified for several reasons. Firstly, machine learning algorithms operate on numerical inputs, making it essential to encode categorical data into binary vectors. One-hot encoding maintains the integrity of the original categories, ensuring that no unintended ordinal relationships are imposed. It also eliminates potential bias that could arise from assigning arbitrary numerical values to categories, providing an accurate representation of the data. Additionally, one-hot encoding handles situations where categorical variables have multiple categories, allocating a dedicated binary column for each unique category. This approach not only ensures model compatibility but also enhances interpretability, allowing analysts to comprehend and explain the model's predictions in the context of the original categorical variables. Overall, one-hot encoding is a vital technique that preserves meaningful categorical information while enabling effective analysis and modeling.

**d) Create one more feature and add this column to the dataset (you can use mutate function in R for this). For example, you can calculate the current age of the car by subtracting “Year” value from the current year.**

Introducing the "Car\_Age" feature, derived by subtracting the "Year" value from the current year, is pivotal in automotive data analysis. This addition is crucial for assessing market dynamics, as a car's age significantly influences its value, performance, and consumer demand. "Car\_Age" enriches predictive models by capturing intricate patterns related to depreciation and consumer preferences. Additionally, it boosts interpretability, providing stakeholders with clear insights into age-related impacts on outcomes. This feature ensures a comprehensive dataset representation, facilitating holistic analyses of factors affecting target variables. Its presence enhances adaptability for future research, enabling effective exploration of diverse hypotheses. In essence, "Car\_Age" not only deepens analysis but also enhances dataset flexibility, making it a valuable and justified addition to the dataset.

**Submission Link:**

<https://github.com/deepthi978/PDSAssignment2>