**PDS ASSIGNMENT – 2**

**Name:** Vinisha Vimal Kumar Shukla **ID:** 16344280

**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.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**Justification:**

* It ensures data integrity and completeness by retaining all available information in the dataset.
* By keeping observations with missing values, it maintains the sample size, preventing data loss and preserving statistical power.
* Imputation methods like mean and mode utilize existing data distribution, providing representative values for missing entries.
* This approach mitigates the risk of introducing bias into the analysis, as imputed values are derived from the observed data.
* Utilizing available information maximizes the dataset's potential for accurate analysis and modeling.
* It enhances the robustness of the analysis by leveraging the entire dataset, even in the presence of missing values.
* Imputing missing values facilitates downstream tasks like modeling and visualization, as the dataset remains intact and ready for analysis without additional preprocessing steps.

**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”).

A screenshot of a computer program

Description automatically generated

**c).** Change the categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hotencoded value.

A screenshot of a computer

Description automatically generated

A screenshot of a phone

Description automatically generated

**d).** Create one more feature and add this column to the dataset. you can calculate the current age of the car by subtracting “Year” value from the current year.

A screenshot of a computer

Description automatically generated

A screenshot of a phone

Description automatically generated

**e).** Perform select, filter, rename, mutate, arrange and summarize with group by operations (or their equivalent operations in python) on this dataset.

**Select:**

A screenshot of a computer program

Description automatically generated

**Filter:**

A screenshot of a computer

Description automatically generated

**Rename:**

A screenshot of a computer

Description automatically generated

**Mutate:**

A screenshot of a computer

Description automatically generated

**Sort/Arrange:**

A screenshot of a computer

Description automatically generated

**Group By:**

A screenshot of a computer program

Description automatically generated

**Conclusion:**

The dataset consists of attributes of used cars, with the target variable being the price measured in lakhs. Missing values were addressed by imputing them or dropping them based on data integrity. Units were removed from specific attributes, and categorical variables were transformed into numerical one-hot encoded values. Additionally, a new feature indicating the current age of the car was created. Various data manipulation operations were performed for analysis and summary.