**Principles of Data Science (5530)-Assignment 2**

**Name: Praveen Kumar Reddy Kadapala**

**ID: 16355737**

1. 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.

First Loaded the train dataset provided and observed all the columns and found one column named ‘unnamed:’ which is having serial numbers, I dropped this column since the serial numbers are not in the order.

Then identified all the missing values in all columns and started replacing them. I replaced the missing values of categorical columns with mode, substituting missing values with the mode maintains data distribution and sample size, acting as a simple imputation method it can preserve relationships and reduce bias.

Replaced the missing values of numerical columns with the mean, substituting missing values with the mean maintains data integrity, sample size, and reduces the likelihood of introducing bias.

A screenshot of a computer

Description automatically generated

1. 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”)

In this question, I removed the units from the attributes of Mileage, Engine, Power, New\_price columns and made them the numerical values.

Mileage: Units ‘kmpl’ and ‘km/kg’ are removed.

Engine: Units ‘CC’ are removed.

Power: Units ‘bhp’ are removed.

New\_Price: Units ‘Lakh’ and ‘Cr’ are removed.

A screenshot of a computer

Description automatically generated

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

In this question encoded the categorical variables “Fuel\_Type” and “Transmission” into numerical values using pd.get\_dummies() function.

A screenshot of a computer

Description automatically generated

1. 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.

In this question, first I calculated the current year using the datetime library and then subtracted the year of each car from the current year to obtain car’s Current\_Age. After calculating the Current\_Age it is added as a new column in the dataset.

A screenshot of a computer

Description automatically generated

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

In this question, I performed all the mentioned operations select, filter, rename, mutate, arrange, and summarize with group by on the provided dataset.

1. **Select Operation**

A screenshot of a computer

Description automatically generated

1. **Filter Operation**

A screenshot of a computer

Description automatically generated

1. **Rename Operation**

A screenshot of a computer

Description automatically generated

1. **Mutate Operation**

A screenshot of a computer

Description automatically generated

1. **Arrange Operation**

A screenshot of a computer

Description automatically generated

1. **Group By**

A screenshot of a computer

Description automatically generated