**Principles of Data Science**

**Assignment 2**

Partha Nossam(16326619)

**Task 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.**

**Justification:**

For **Mileage** column I have taken the **Mean** value, because the difference in the mileage range for all the cars is very less. It is appropriate to take the mean value.

For **Engine** also I have taken the **Mean** values, as the number of null values is very less, and the range of CC is varying between 1200 to 1700 for most of the cars.

For **Seats** column I have taken the **Mode** value, because the number of seats is either 5 or 7. It is appropriate to take a Mode value for this column.

For **Power** column I have taken the **Mean** value for **Automatic cars** and **Mean** value for **Manual** **cars** separately. I observed a pattern that Automatic cars have more power than Manual cars.

For **New\_Price** column also I have done the same as **Power** column. **Mean** for **Automatic** and **Manual** cars separately. There are almost 45% of the records are null, there is a can that we can neglect it. But, here I am taking the mean values as I am using this column in the upcoming tasks.

**Task2**: **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”).**

For New\_Price column contains some records with crores, I had to convert them into lakhs.

After that, I have converted all the above columns into strings to perform the string operations on it. Using str.replace() I have removed all the units from the above mentioned columns.

**Task3: Change the categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hot encoded value.**

get\_dummies are used to perform the hot encoded technique on Fuel\_type and Transmission.

**Task 4:** **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.**

I have used datetime function in python to find the age of the car.