**Assignment 2**

**Task1:**

Loaded the dataset and figured out the missing columns and their count using **isnull().sum()** function. Later handled missing values('Mileage', 'Engine', 'Power', 'Seats') for numerical columns by imputing with mean. Dropped New\_price column as there is a high percentage of missing values.

**Imputing Mean**: Numbers in columns with names like "Mileage," "Engine," "Power," and "Seats" usually represent continuous or semi-continuous data. One typical method for dealing with missing values in these columns is to impute the missing values using the column mean.

**Dropped New\_Price' Column:** When missing values occur randomly and do not inject bias into the dataset, using the mean as an imputation approach is appropriate. It supports the preservation of the data's statistical characteristics, including the central tendency.

I have updated the **Question a** details in **Task1 Folder.**

The output of this task is **Task\_1\_output.csv.**

**Task2:**

I have performed task 2 first here as I was unable to compute mean of the missing columns with the units present in the respective columns.

'Mileage': Units 'kmpl' and 'km/kg' are removed.

'Engine': Units 'CC' are removed.

'Power': Units 'bhp' are removed.

'New\_Price': Units 'Lakh' are removed.

I have updated **Question b** details in **Task2 Folder**

The output of this task is **Task\_2\_output.csv.**

**Task 3:**

To perform One-Hot Encoding, I have used pd.get\_dummies() functionon the specified columns

I have updated **Question c** details in **Task3 Folder.**

The output of this task is **Task\_3\_output.csv.**

**Task 4:**

I have created **“Kilometers\_Per\_Year"** Feature and added this column in the dataset. Calculated the "Age" of each entry in the dataset. It does this by subtracting the "Year" column's values from a reference year, which is set as 2023 . Created a new feature called "Kilometers\_Per\_Year" by dividing the "Kilometers\_Driven" column by the "Age" column. This feature provides an estimate of the number of kilometers driven by the vehicle per year.

I have updated **Question d** details in **Task4 Folder.**

The output of this task is **Task\_4\_output.csv.**