PRINCIPLES OF DATA SCIENCE

**ASSIGNMENT-2**

MANOJ KUMAR SUGGALA

**(16357169)**

**Task-1:**

Initially we are loading our dataset and trying to find out if there are any missing values in the dataset provided. For this we are using isnull(). sum() function.

From the report we got the columns with Mileage, Engine Power and seats as a smaller number of missing values. Those missing values are replaced by mean of those columns because the missing values are very less in number and imputing them with mean doesn’t make much difference in change in data.

Dropping the column named New\_Price because there are huge number of missing values, and we cannot replace them with mean values because we cannot fit those large number of missing values according to the dataset, there might be huge difference in those values.

The question A have been uploaded in task-1 folder.

The output of task-1 is:

A screenshot of a computer

Description automatically generated

**Task-2:**

In this task we are removing the units from the columns like:

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

'Engine': Units 'CC' are removed.

'Power': Units 'bhp' are removed.

A screenshot of a computer

Description automatically generated

From the above output we can say that the units are removed successfully.

**Task-3:**

In this task we need to perform One-Hot Encoding. We are using get\_dummies function on specified columns.

**Question C** Output can be seen in task-3 folder.

From below image we can say that there are Fuel\_Type\_Diesel, Fuel\_Type\_Electric, Fuel\_Type\_petrol, transmission\_automatic and transmission\_Manual columns have been updated with 0 and 1 values as shown below.

A screenshot of a computer

Description automatically generated

**Task-4:**

In this task we are adding more features to our data:

Here we are adding the Car\_age column for this, we are calculating the Car\_Age by subtracting the Year Column from the current year which is 2024.

A screenshot of a computer

Description automatically generated

Next we have added one more feature kilometer\_driven\_ per year by dividing the kilometers driven by the car\_Age.

A screenshot of a computer

Description automatically generated

**Task-5:**

In this task we are performing the select, filter, rename, mutate, arrange and summarize on the provided dataset.

**Select**: Select the columns you want and get them right here:

A screenshot of a computer

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

**Arrange**:

A screenshot of a computer

Description automatically generated

**Summarize with group by:**

A screenshot of a computer

Description automatically generated