**Assignment-2**

**Abhinav Reddy Ayyadapu**

**16354733**

**Data\_Clean**

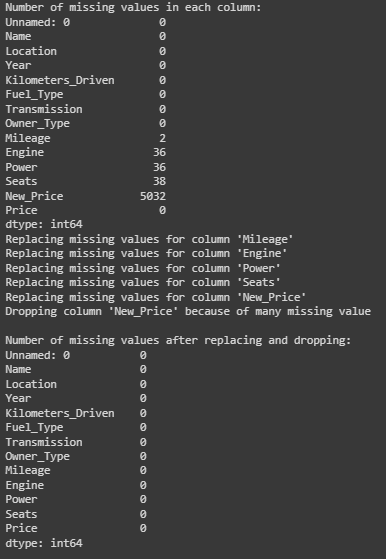
**Given Dataset:**

**A screenshot of a computer

Description automatically generated**

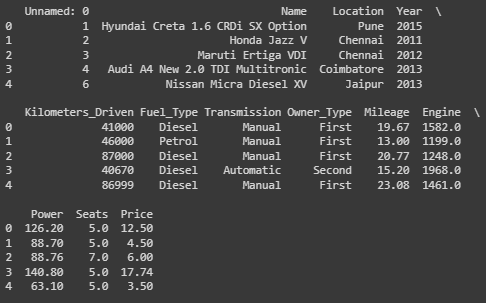
There are different attributes present in the dataset.

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, display the number of missing values in the given dataset in each column
* After that defined a threshold for missing values and dropping values more than the threshold
* Then replaced missing values that are less than the threshold and dropped the column with many empty values
* There are missing values in the columns like Engine, power, Seats. To fill those NA values, we can use mean, mode and median.
* Dropped column new \_price because of too many empty values

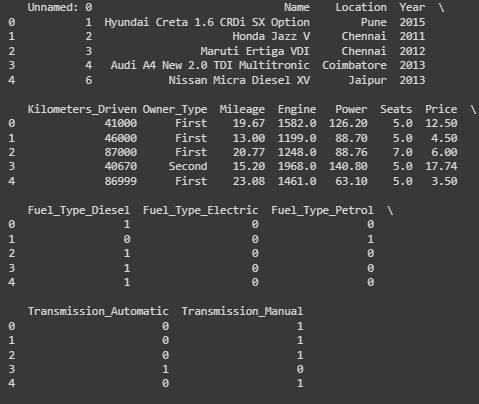
1. **Remove the units from some of the attributes and only keep the numerical values (for example remove kmpl from “Mileage”, CC from “Engine”, bhp from “Power”, and lakh from “New\_price”).**

* Removed the text values and taking only numerical values from the given data set by splitting the values based on space between the numerical and text values
* Removed the units from the given columns as well
* Below is the output of dataset after performing the above mentioned operations



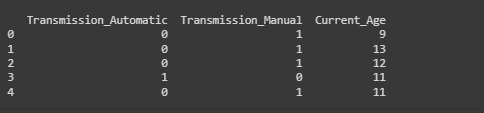
**c) Change the categorical variables (“Fuel\_Type” and “Transmission”) into numerical one hot encoded value.**

* The one hot encoded values are obtained by converting the Fuel\_Type and Transmission category variables. It is evident that every distinct category in "Fuel\_Type" and "Transmission" has been transformed into a distinct binary column, either 1 or 0.
* Below is the output of dataset after performing the above mentioned operations



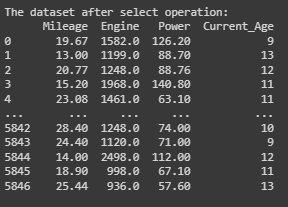
1. **Create one more feature and add this column to the dataset (you can use mutate function in R for this). For example, you can calculate the current age of the car by subtracting “Year” value from the current year.**

* Added new column Current\_Age by calculating the age of car from the columns year
* Below is the output of new column current age

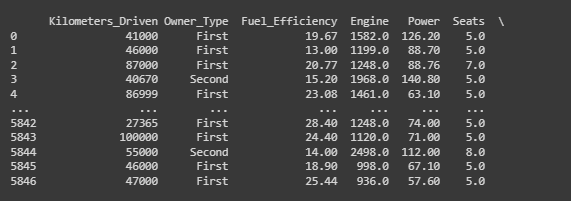
****

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

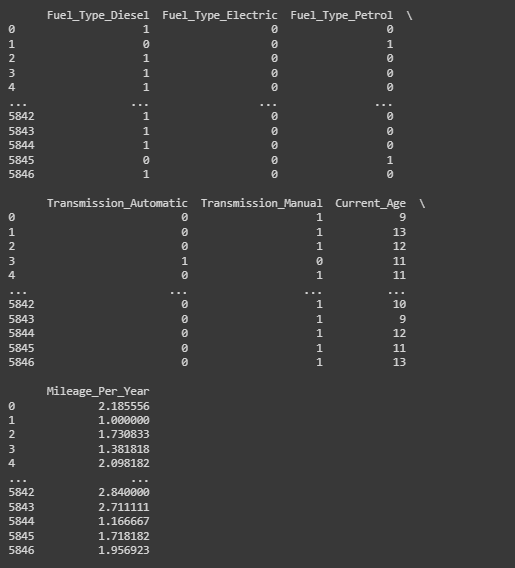
* **Returning the selecting data.**

****

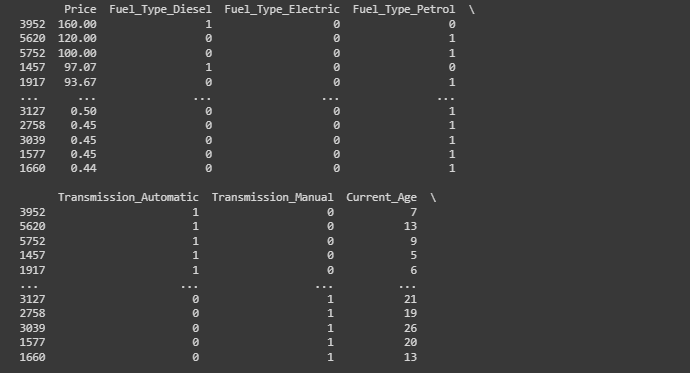
* **Returning the filtered data and Renaming the columns**

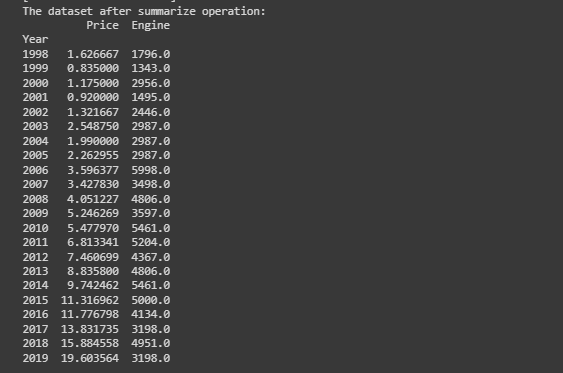
****

* **Add or modifying the columns**

****

* **Arrange and group by operations**

****

****