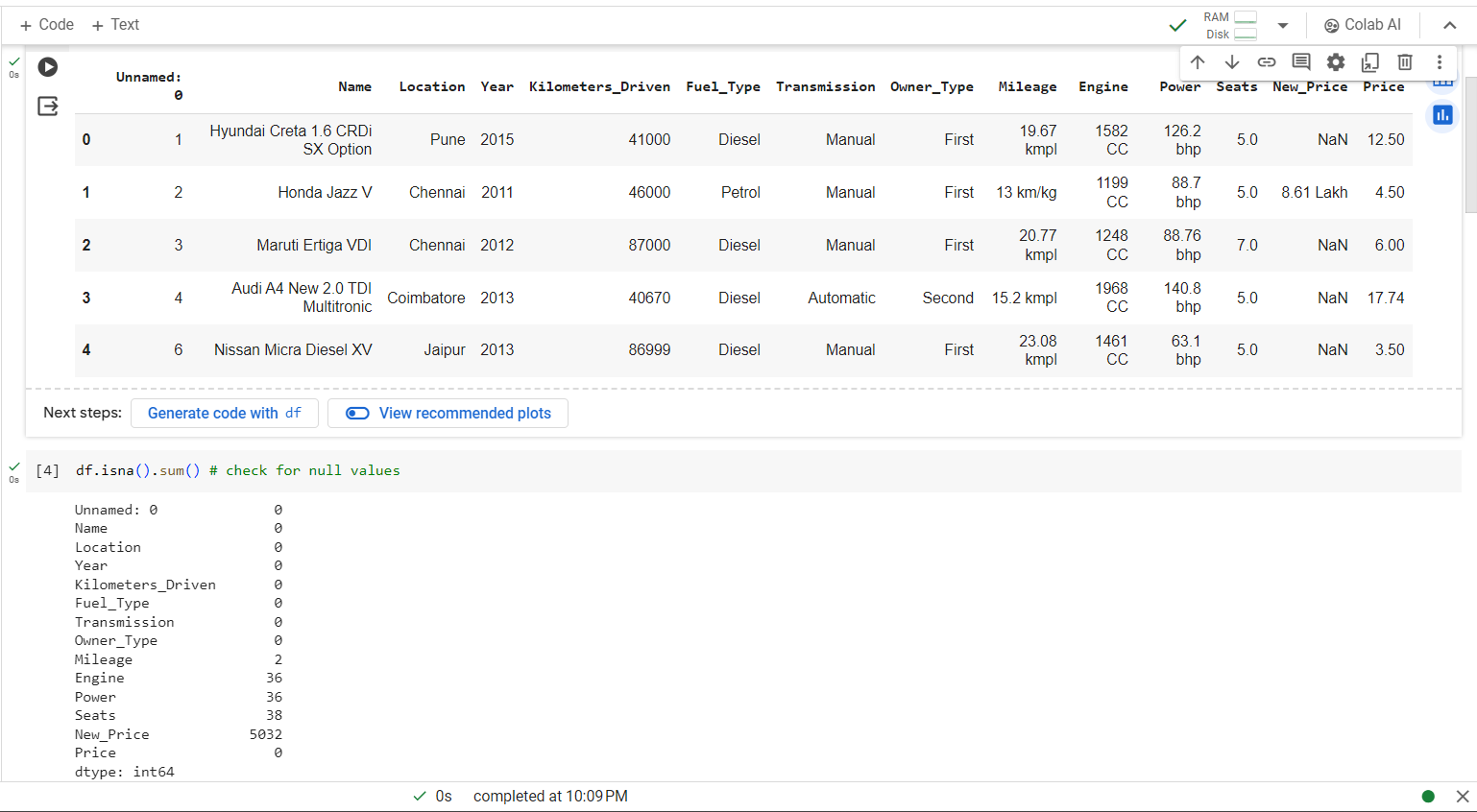
Name: Krishna Uttej Kondisetty

Student ID- 16352970

**Assignment-2**

a)Looking for the missing values in all the columns and either imputing them (replace with mean, median, or mode) or dropping them.



Replacing missing values in Mileage, Engine, and Power with the median is chosen because it provides a robust representation of the typical value in the dataset. This method is preferred due to its ability to handle outliers more effectively and maintain the variability of the data.

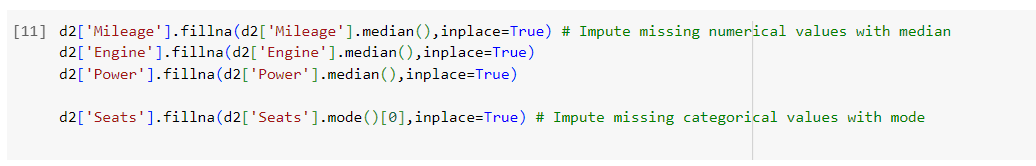
Using the mode to fill missing values in the Seat column is appropriate as it indicates the most commonly occurring whole number of seats. This approach ensures that the imputed values align with the prevailing pattern in the dataset.

Dropping the New\_Price column is warranted because it contains a substantial proportion of missing values, exceeding 50%. Attempting to fill such a large number of missing values could compromise the integrity of the data and subsequent analyses. Therefore, removing this column is a prudent decision to maintain the overall quality of the dataset.



Saving the results in a new file called “preprocessed\_data.csv”

b) 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”).

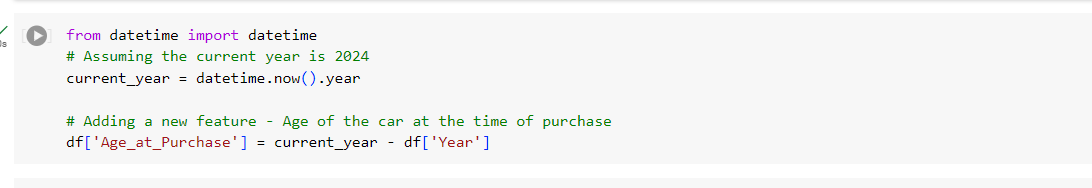


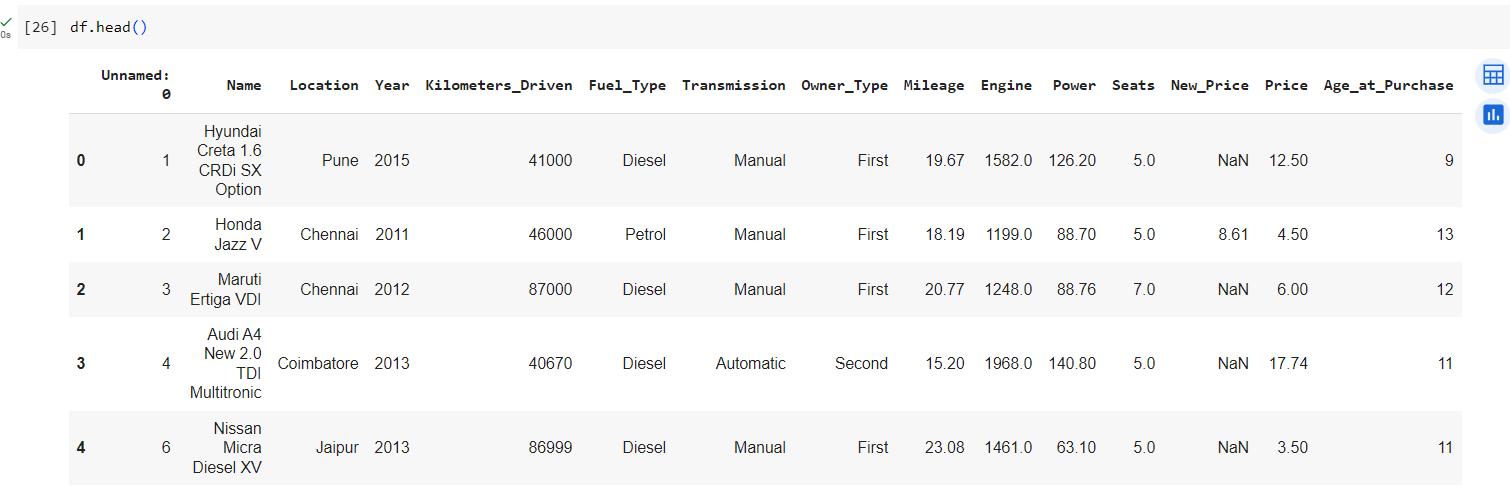
c) Change the categorical variables (“Fuel\_Type” , “Transmission” , “owner\_type”) into numerical one hot encoded value



Saving the final results in a new file called “encoded\_data.csv”

d) Create one more feature and add this column to the dataset





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



