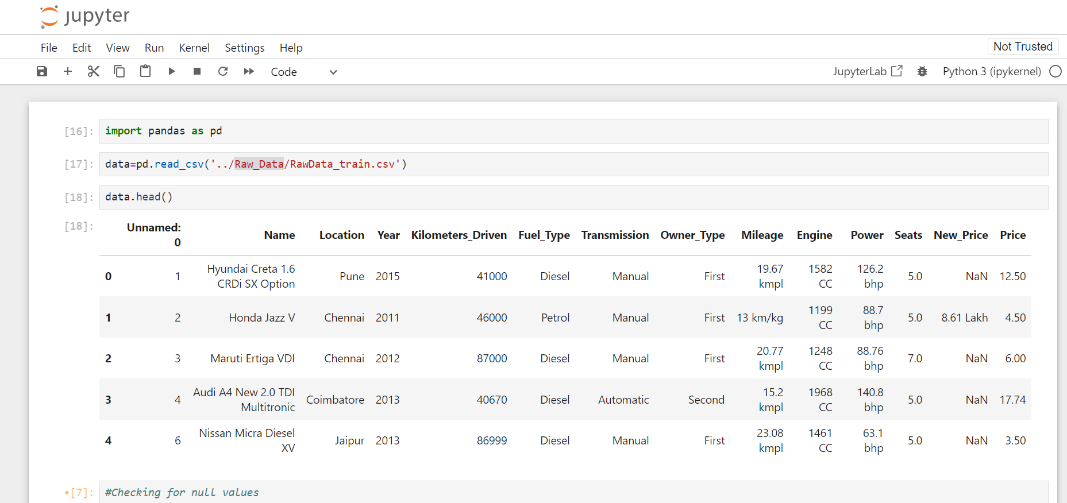
Name: Bhavana Navari

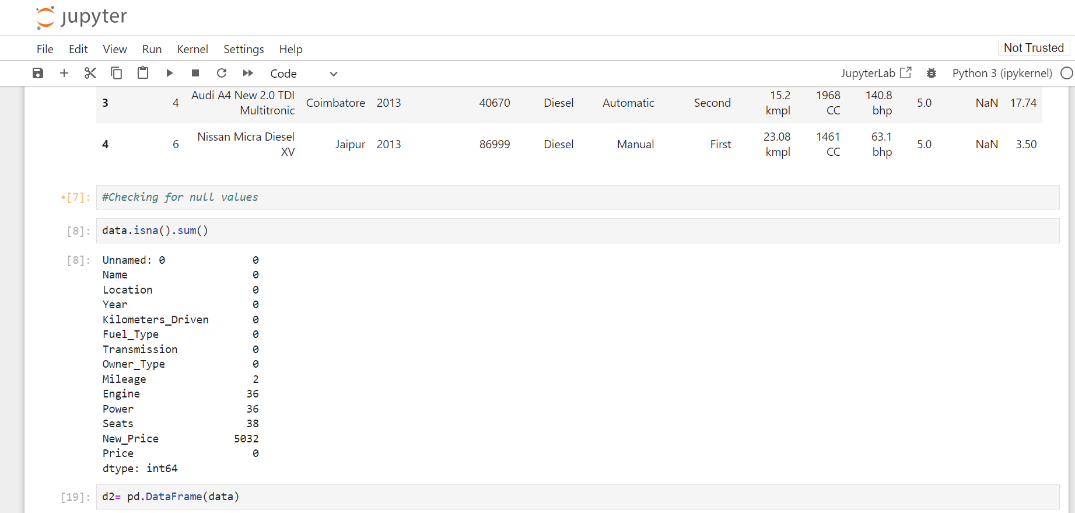
Student ID- 16341885

**Assignment-2**

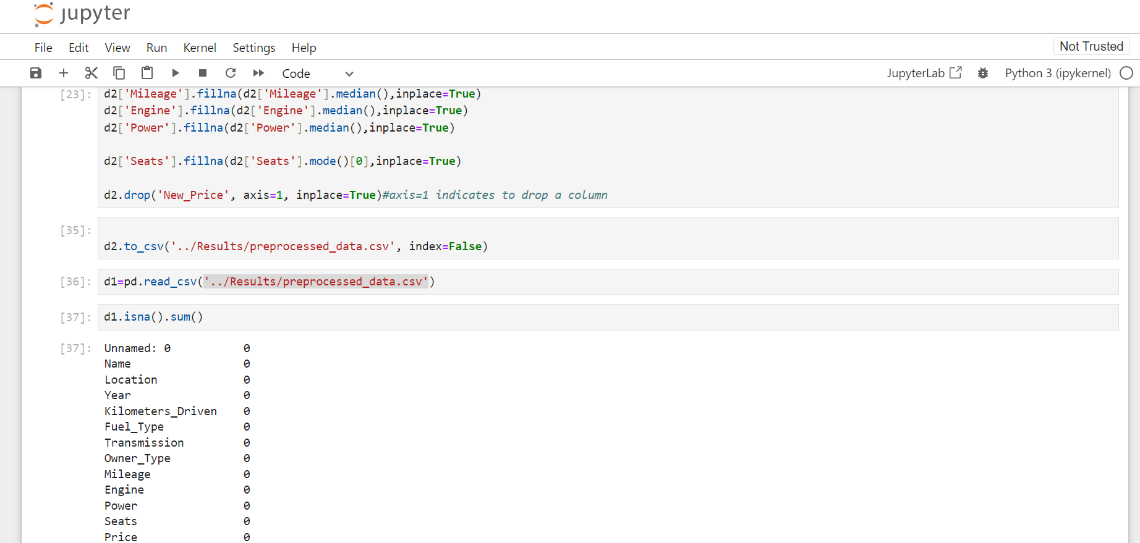
1. Looking for the missing values in all the columns and either imputing them (replace with mean, median, or mode) or dropping them.

* Loading the dataset and making observations. The dataset has 14 variables.



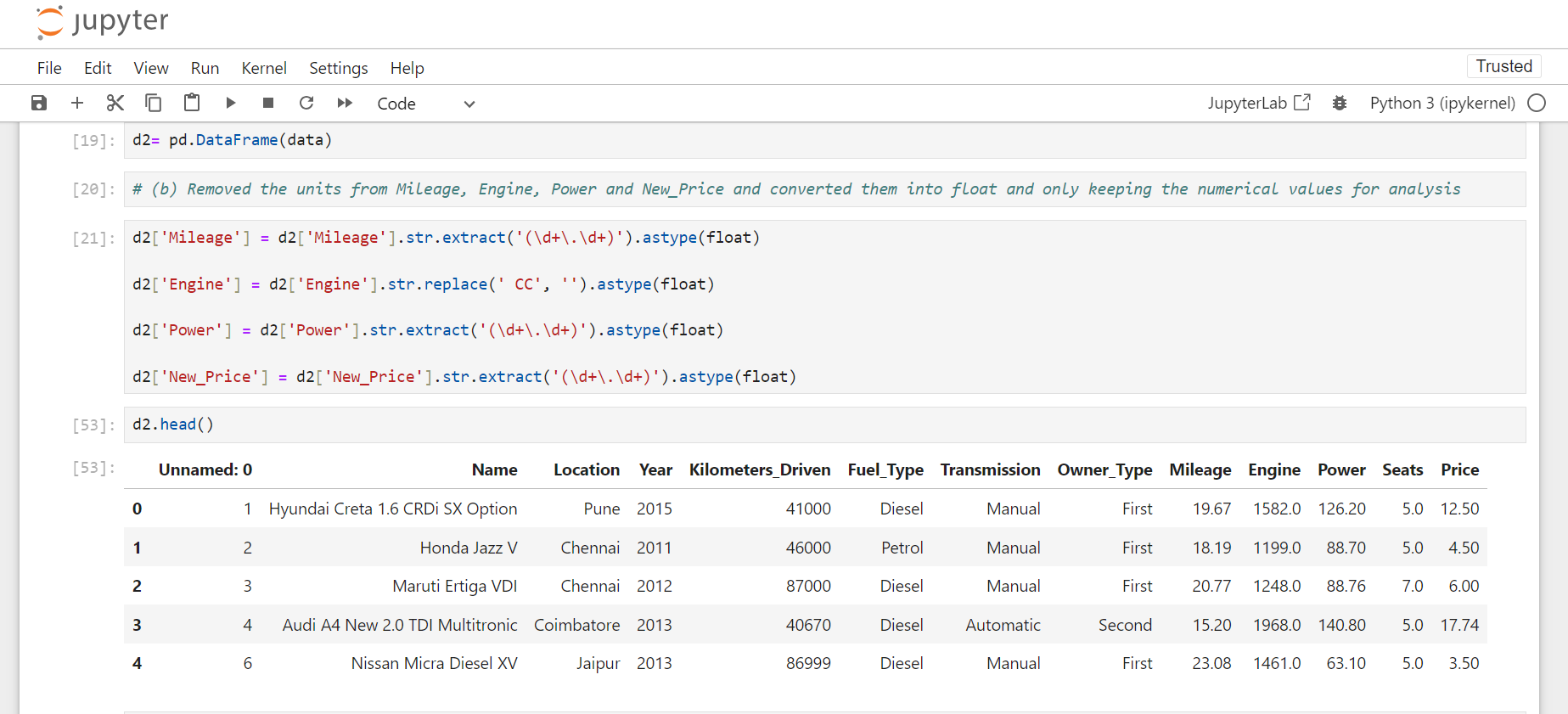


* Replacing the missing values in Mileage, Engine, and Power with the median because it represents the typical value, have more variance and is less affected by outliers.
* Using the mode to replace missing values for the Seat column as it gives the information about the most frequently occurring whole number of seats.
* Dropping the New\_Price column as it contains more than 50% of missing values in it. By filling these missing values which are huge in number could not only impact the data making it faulty as well as its analysis.



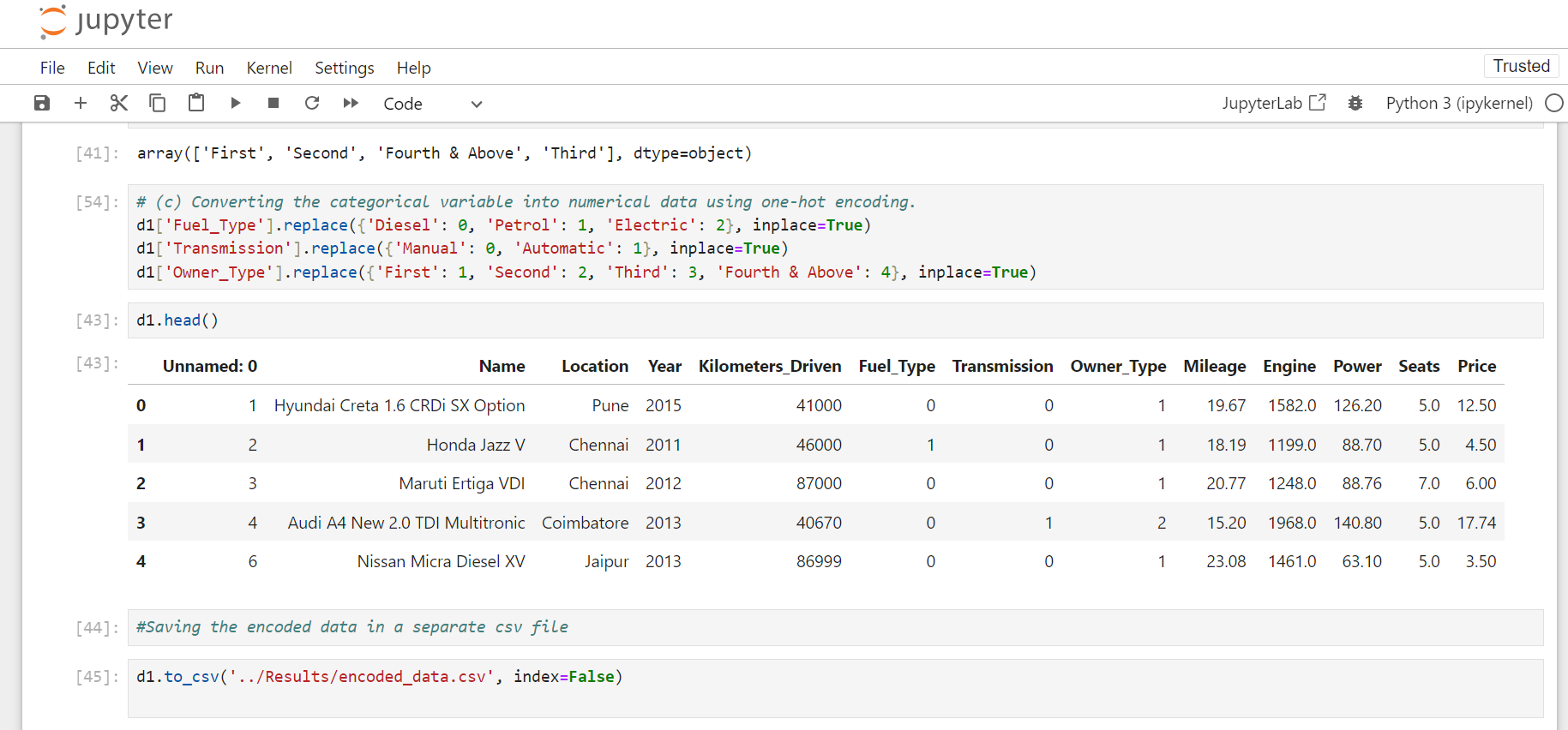
Saving the results in a new file called “preprocessed\_data.csv”

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”).



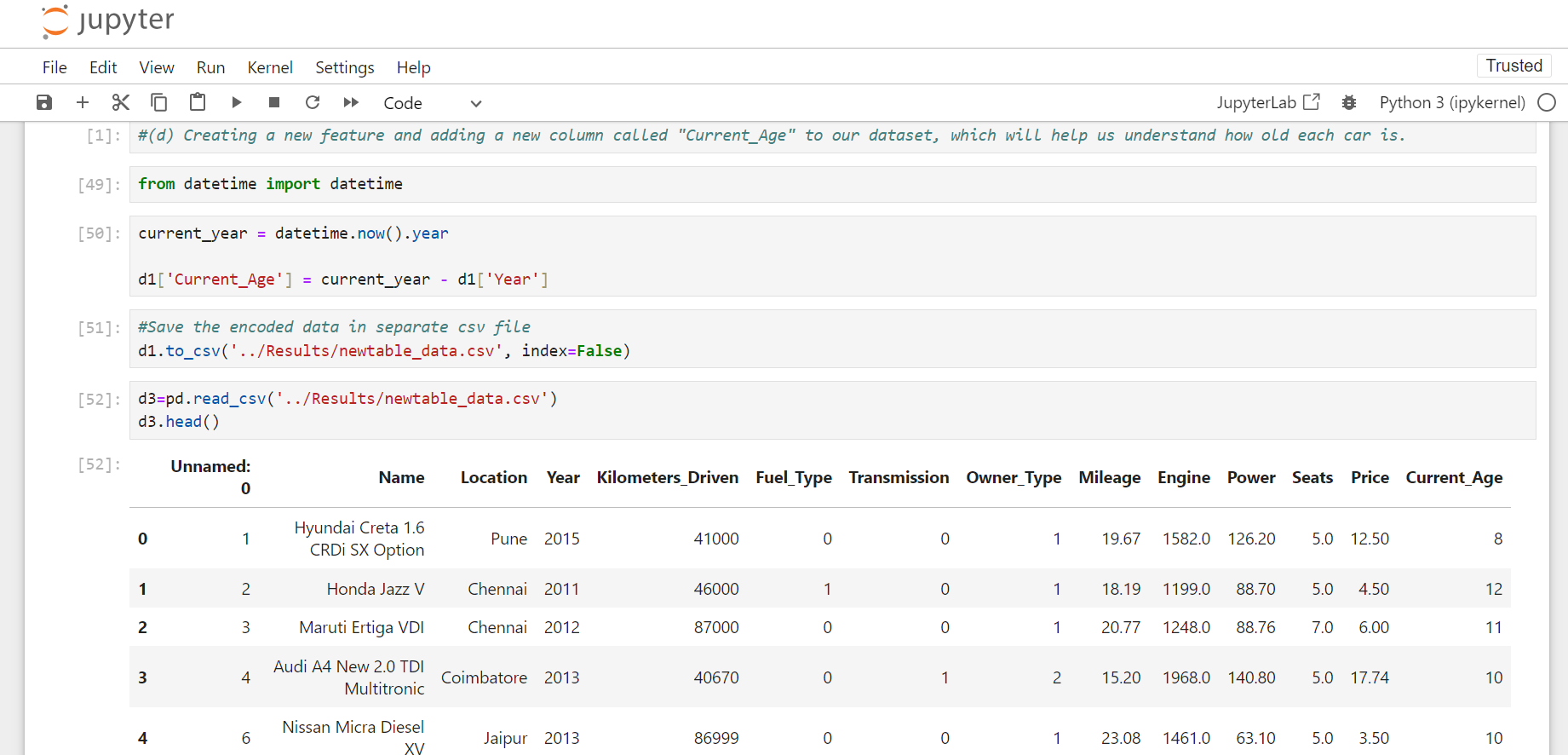
* missing values in Mileage, Engine, and Power with the median because it represents the typical value and is less

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



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

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.

The results are stored in a new file called “newtable\_data.csv”