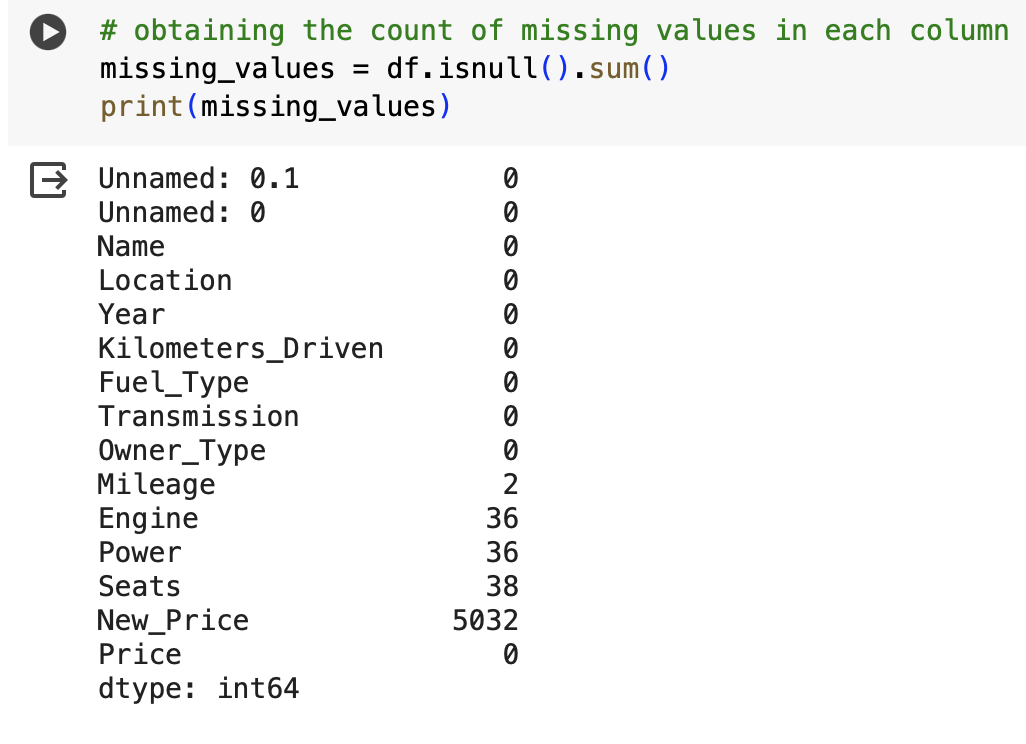
**Principles of Data Science**

**Assignment-2**

**Name: Sreevardhan Reddy Soma ID:16352646**

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.**

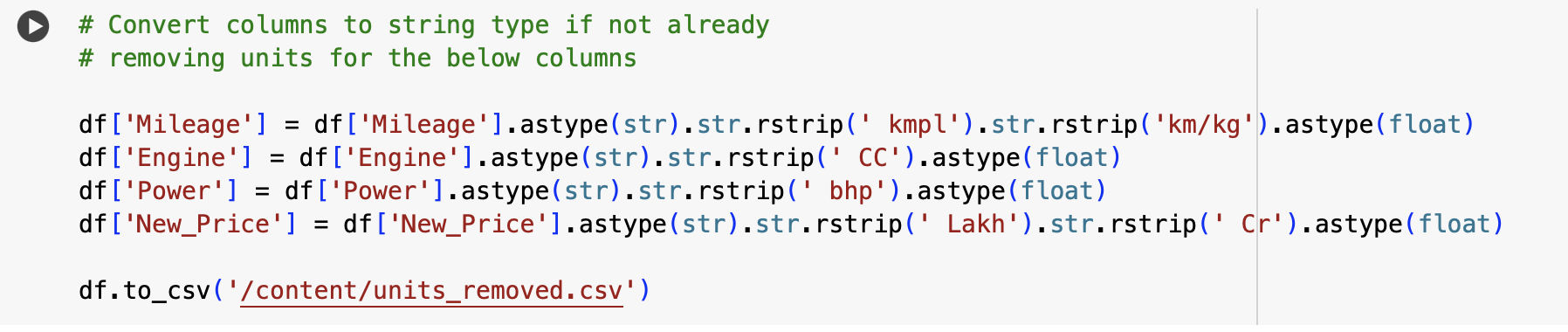
At first, I loaded the data set and identified the missing values using isnull( ) function by adding sum( ) to it.



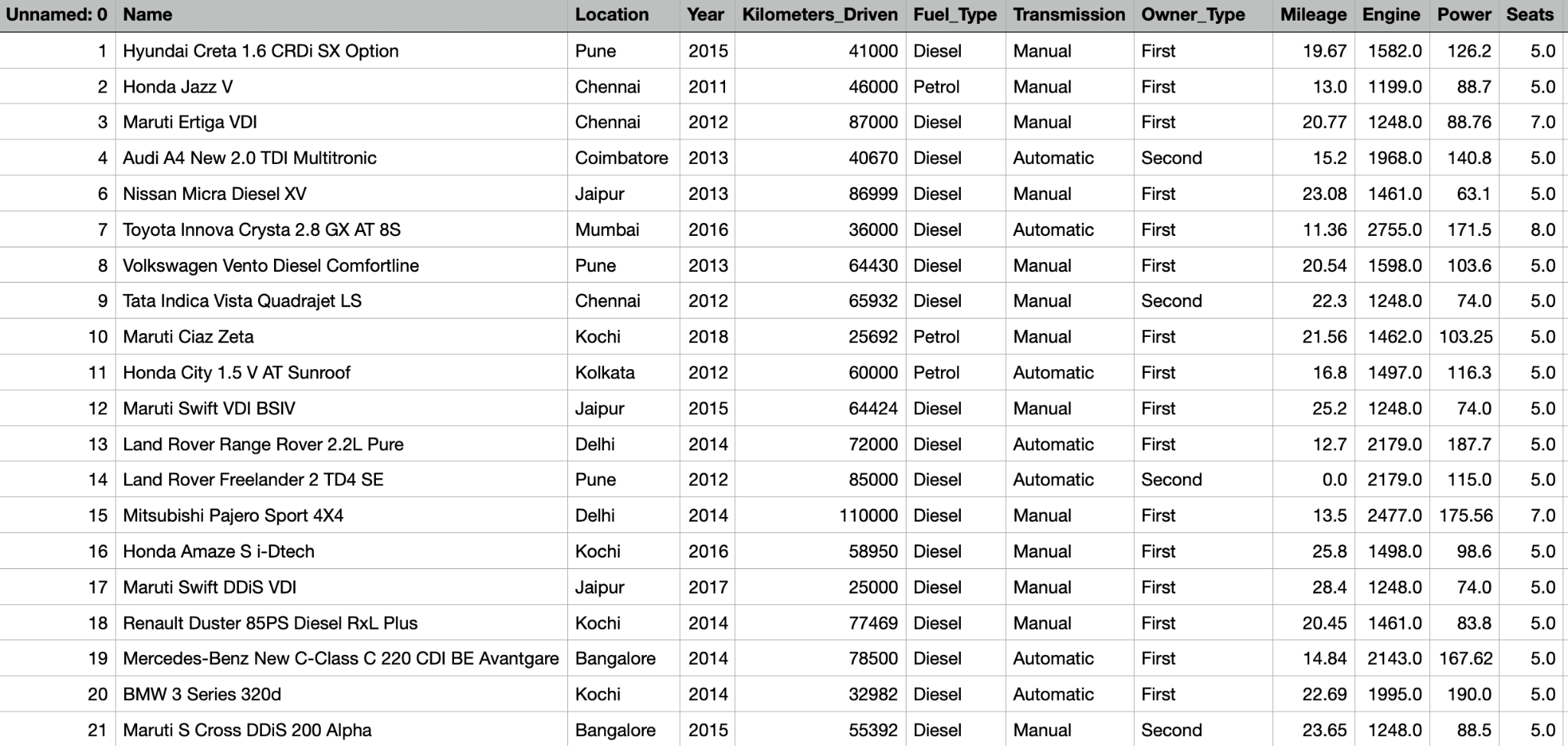
Here, we can identify that Mileage, Engine, Power, Seats and New\_price are having missing values in the dataset. Since I want to try different approaches to fill in the missing data, I have used mean and median to impute the missing values and removed the column “New\_Price” since they have a large number of missing values. Since “Mileage” and “Power” have only few missing values, I tried using median to fill up the missing values because the median almost represents the mid value of the entire data present and imputing those will not have much variance and also a better choice to get rid of sensitive outliers. This also helps to avoid bias in the dataset. To try another approach, I used the mean to fill up “Engine” and “Seats”. This also performs similar to the median and by doing this maintains the data integrity and reduces bias when compared to leaving those empty cells as it is. Moreover, the column “New\_Price” has more percentage of missing values, it will not have any impact even if we replace them with some values. Also, since the data is revolving around used cars, New\_Price of the car is irrelevant to compare while purchasing used ones. So I thought dropping this would be good as it is not crucial for data analysis.



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”).**

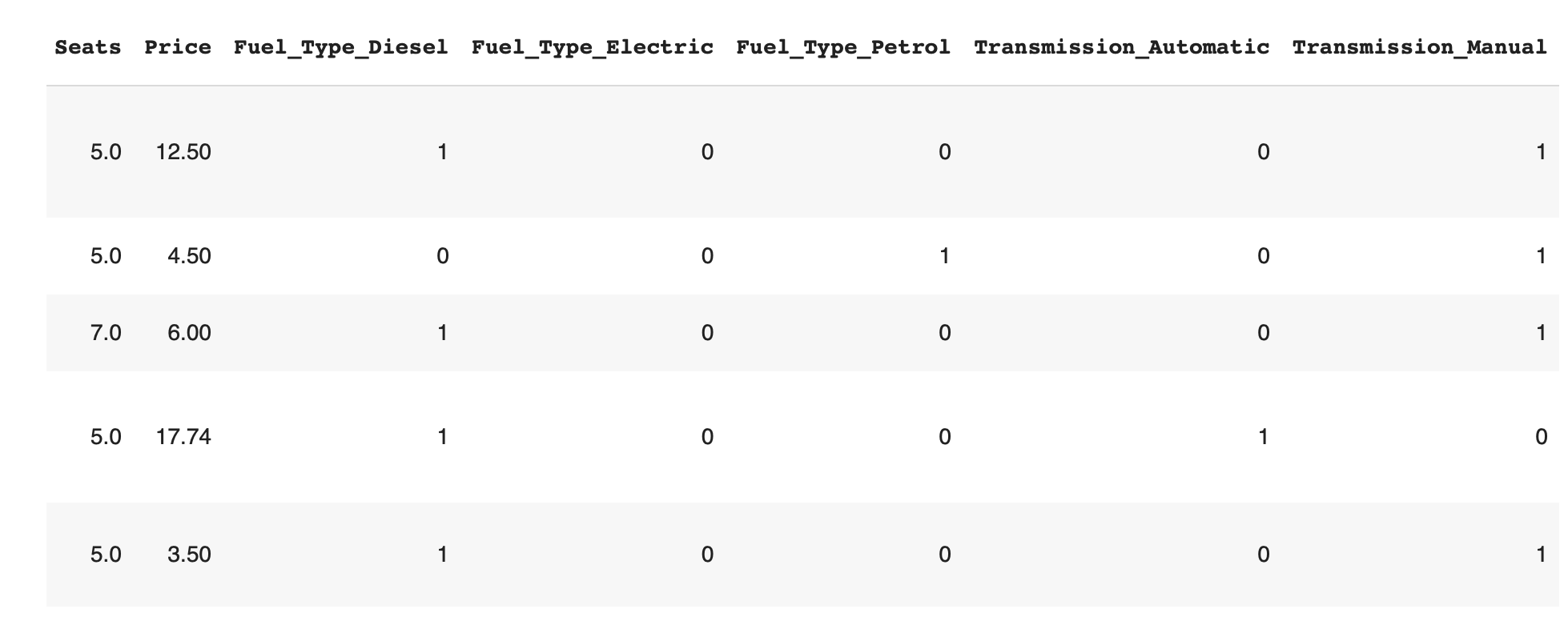


Here, we need to remove any unit present in the Mileage, Engine, Power and New\_price. Mileage has kmpl and km/kg as its unit, Engine has CC, Power has bhp and New\_price has Lakh and Cr as their units.



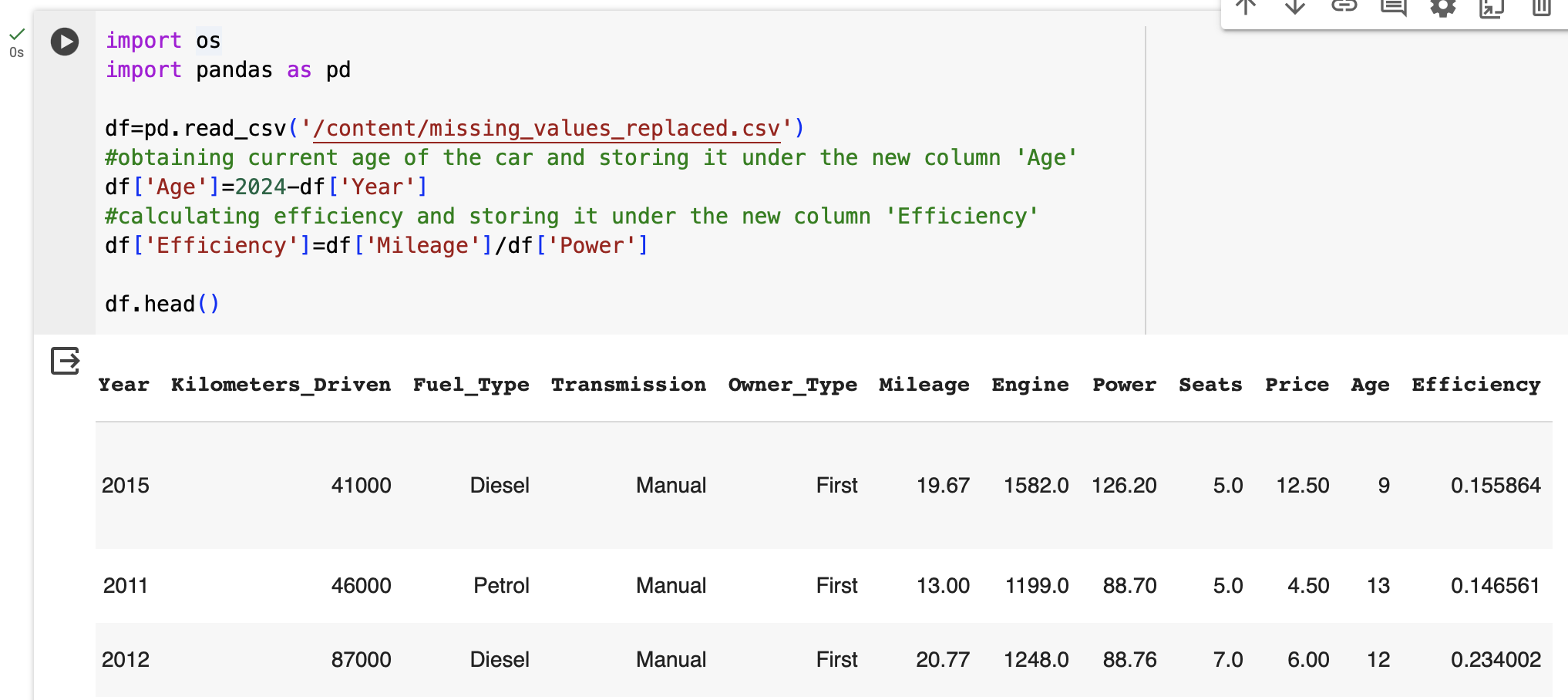
As we can observe, the dataset is free from the units in their respective fields.

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



In this task, we have used get\_dummies function of pandas to change categorical variables into one hot encoded value. Since this transforms categorical data into numeric data, it will be helpful in reducing the bias if we are performing model training.

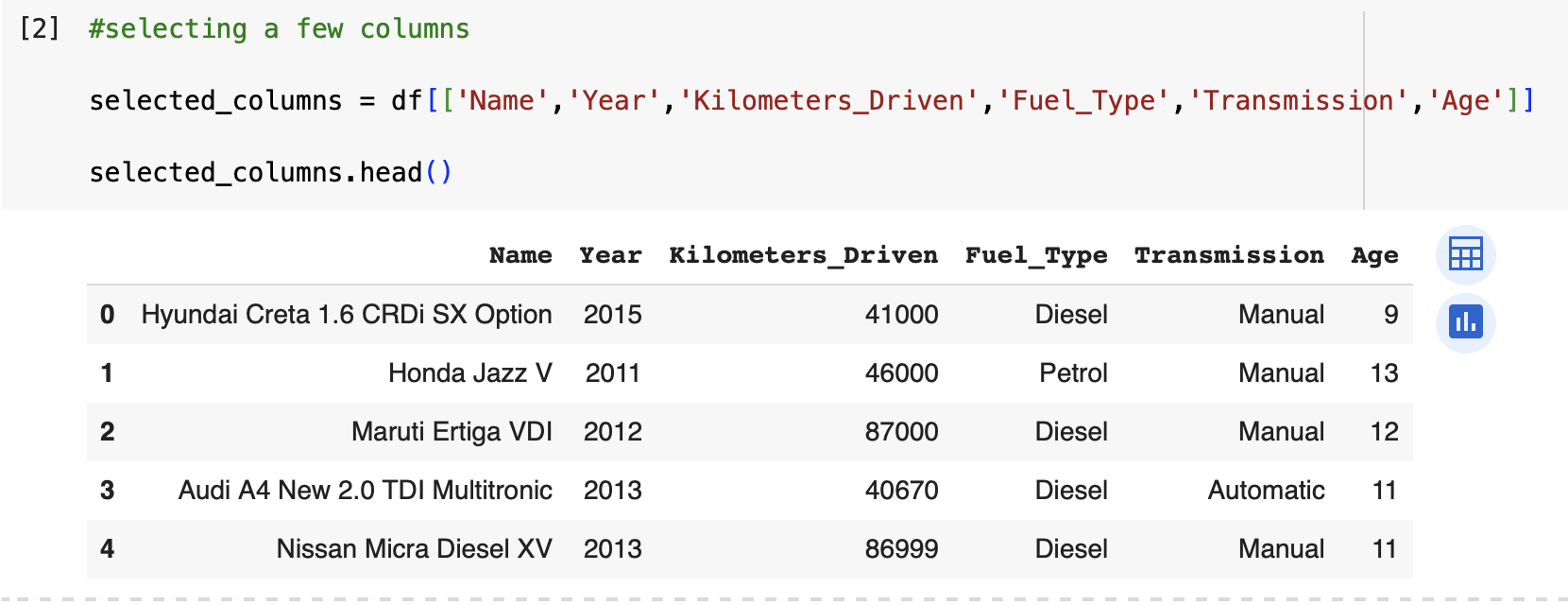
1. **Create one more feature and add this column to the dataset (you can use the mutate function in R for this). For example, you can calculate the current age of the car by subtracting the “Year” value from the current year.**



Here, I have created a new column named “Age” which calculates the current age of the car. It has been calculated by subtracting the car manufactured year from the current year which is 2024. This gives us information about the age of the car since it is manufactured. Along with this, I tried to incorporate the efficiency of the car by dividing the Mileage with power. This data helps the customers to go towards the cars with high efficiency.

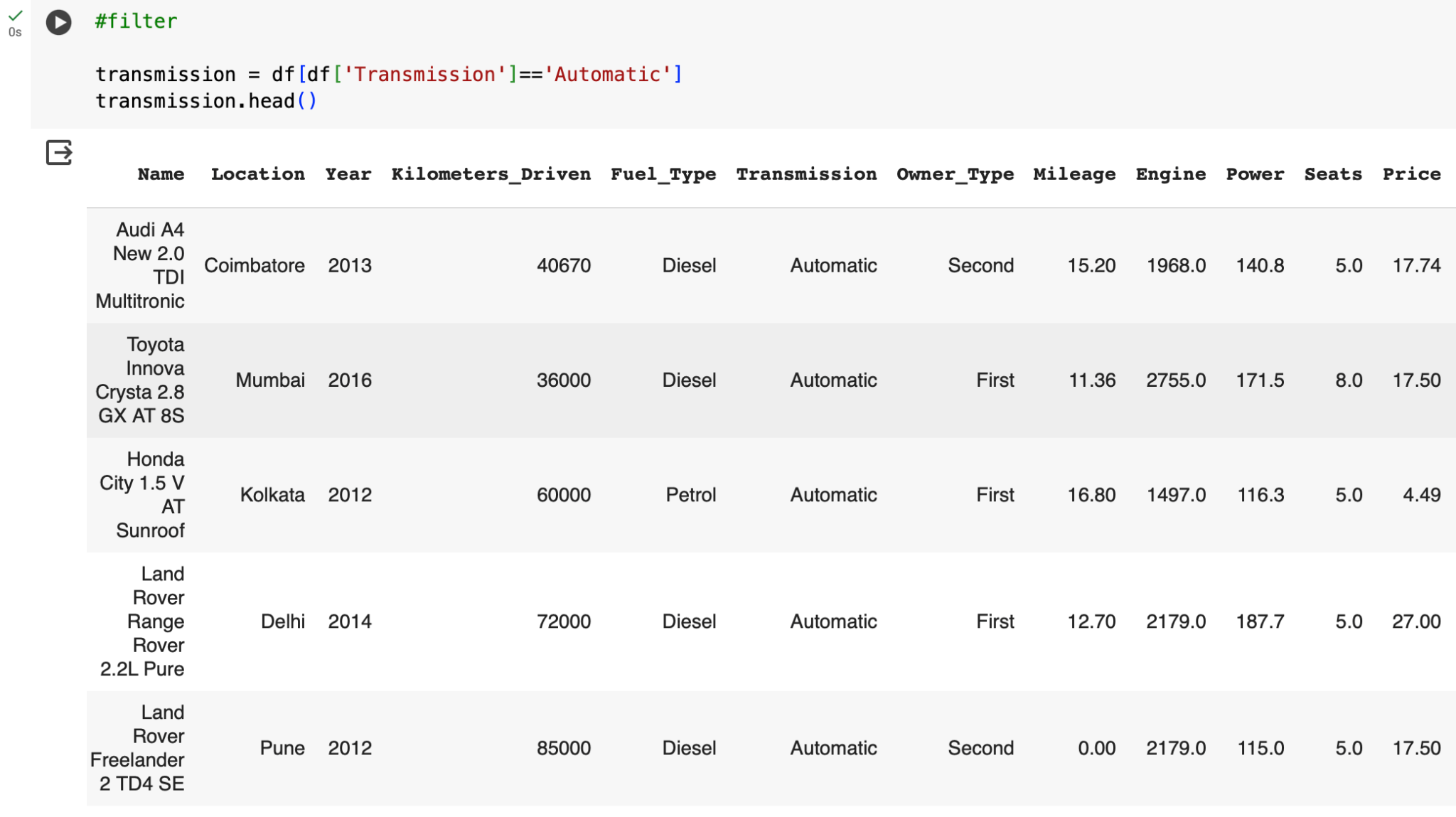
1. **Perform select, filter, rename, mutate, arrange and summarize with group by operations (or their equivalent operations in python) on this dataset.**

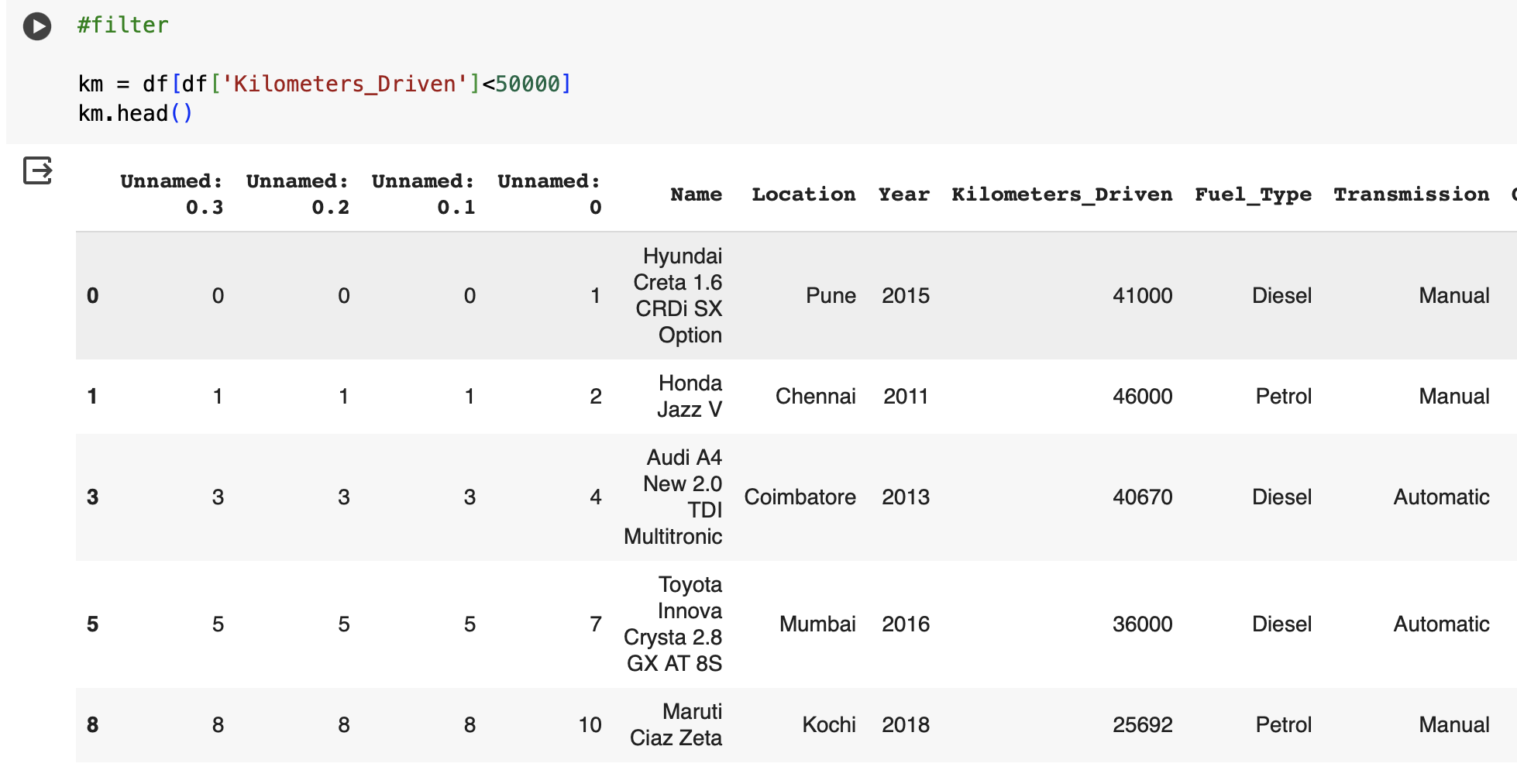
**i) select**

****

I have selected a few of the columns and displayed them as above. This snippet gives an idea about the Car’s name along with the manufactured year, kilometers driven, its fuel type and transmission mode and also the age of the car.

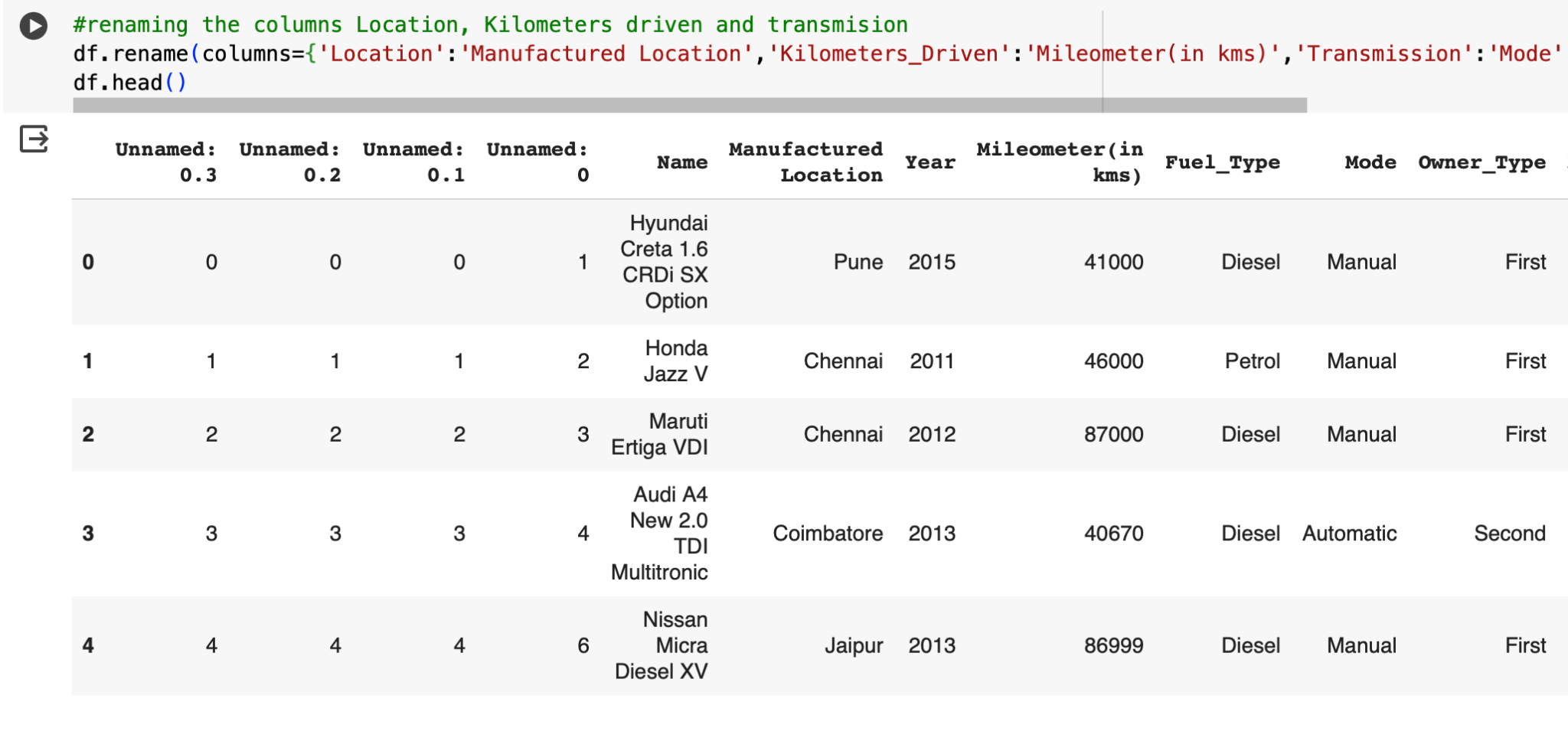
**ii) filter**





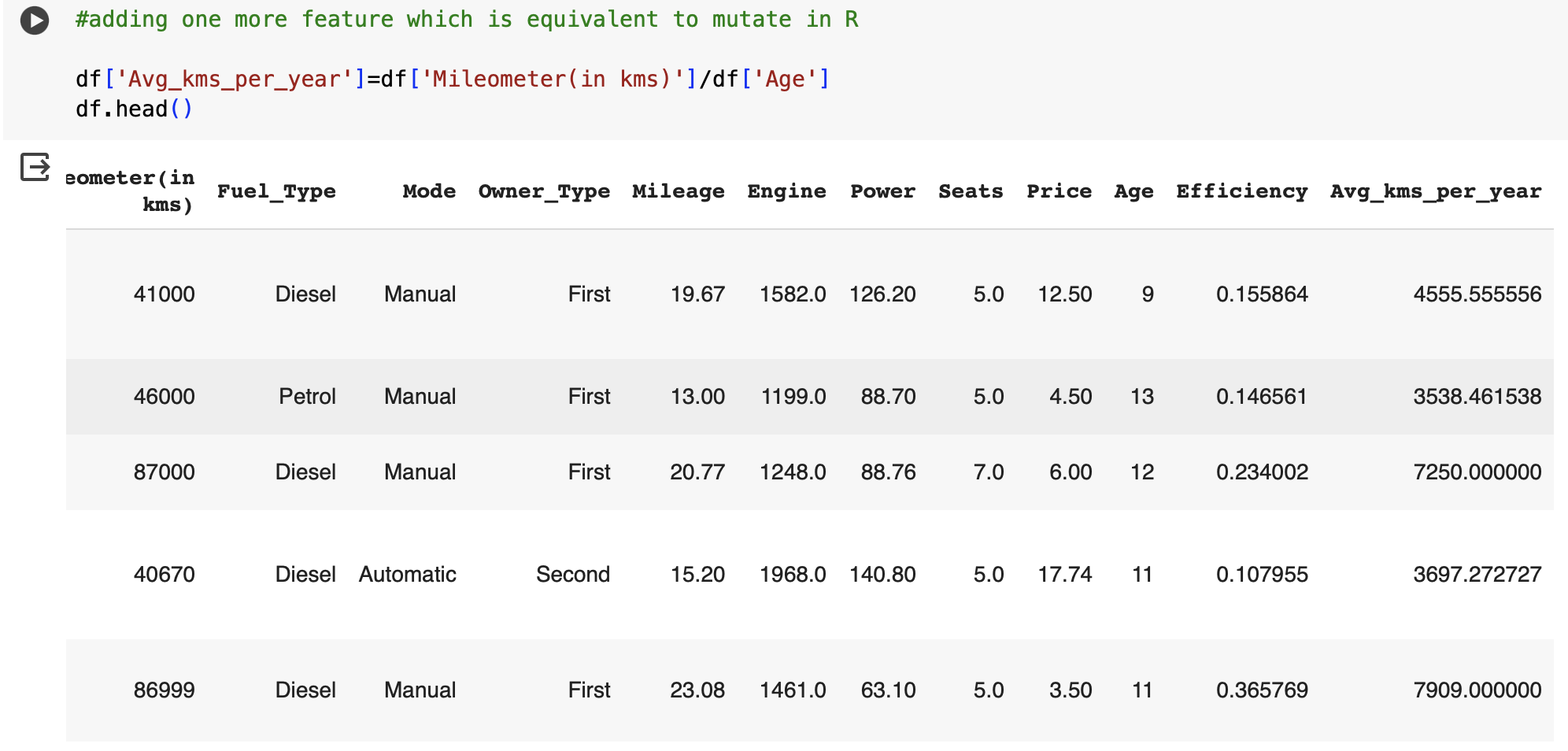
I have filtered all the cars which are Automatic and also the cars which have driven less than 50000. This type of operation helps in getting only the desired data as per the conditions which we have given.

**iii) rename**

****

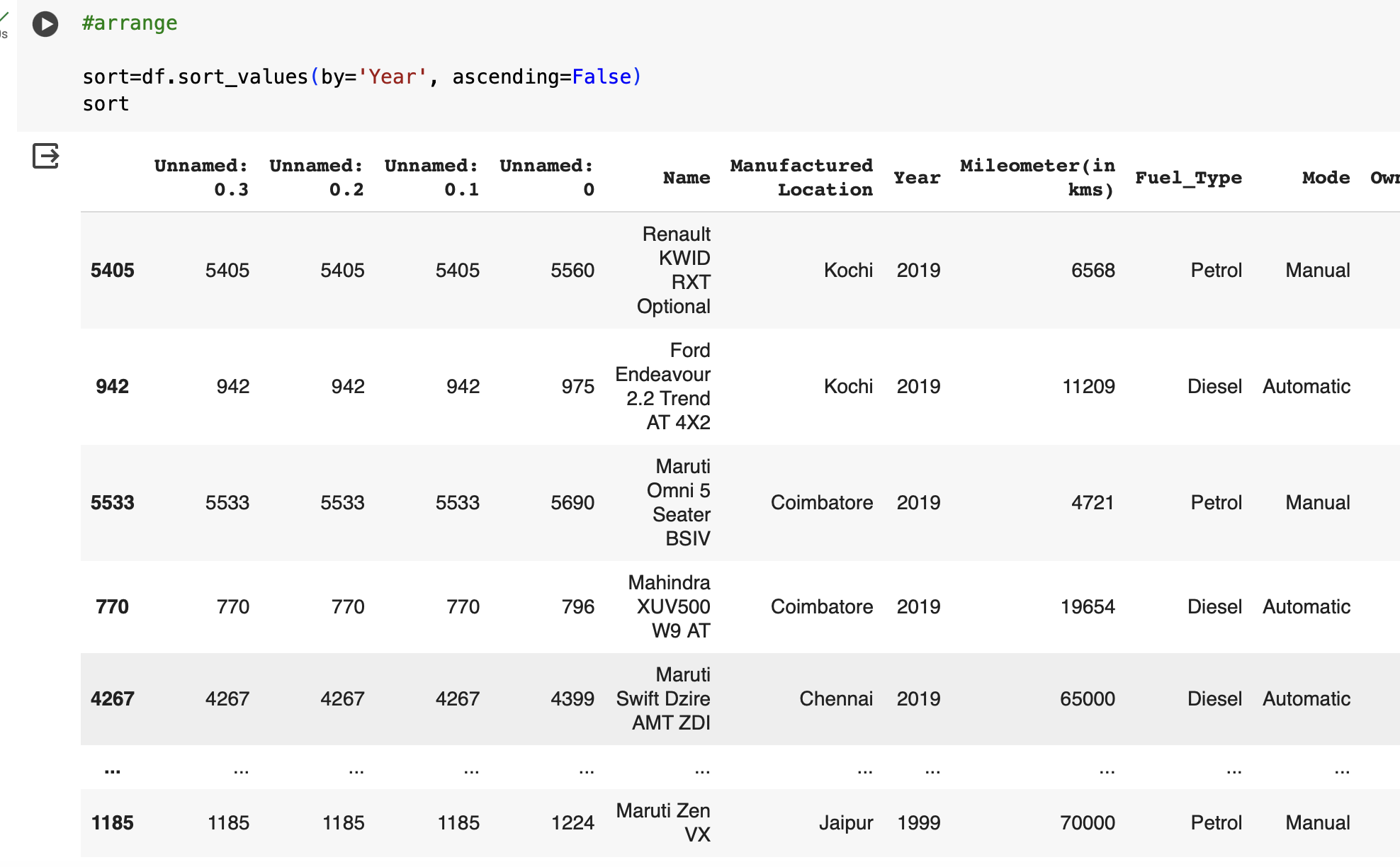
Renaming the certain columns helps the users to rename the most complicated terms in the dataset to a simplest form as per our convenience. Here I have changed the names of three columns namely Location, Kilometers\_Driven and Transmission to Manufactured Location, Mileometer(in kms) and Mode. We will perform this type of operation to change the names of attributes as per our convenience.

**iv) mutate**

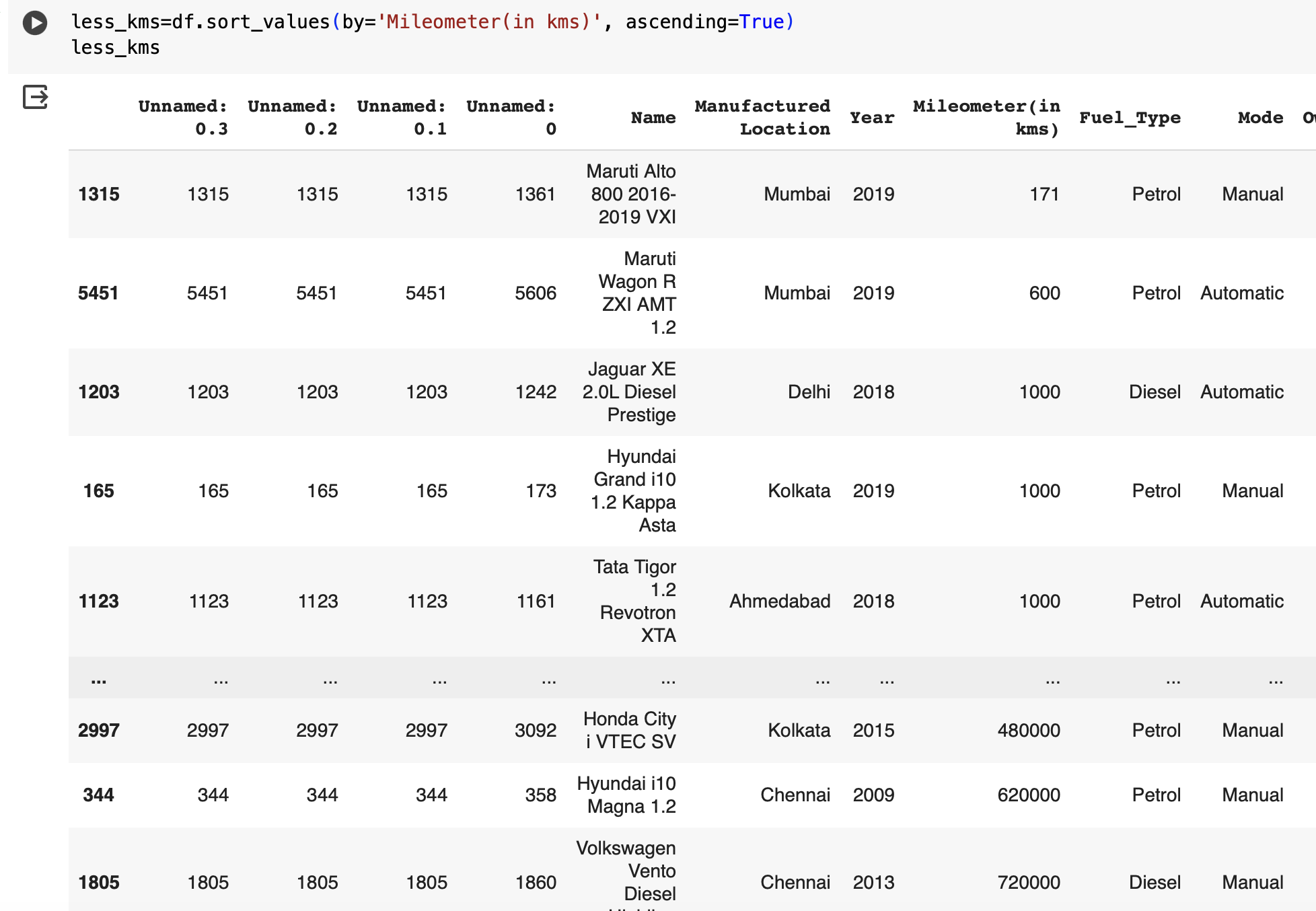
****

To perform a mutate operation which is nothing but creating a new attribute by performing some mathematical operations, I have chosen to divide the kilometers the car has driven in its lifetime until now by the age of the car. This gives us an idea about how many kms a particular car has been used per year. This attribute might become an important feature in purchasing the used car.

**v) arrange**

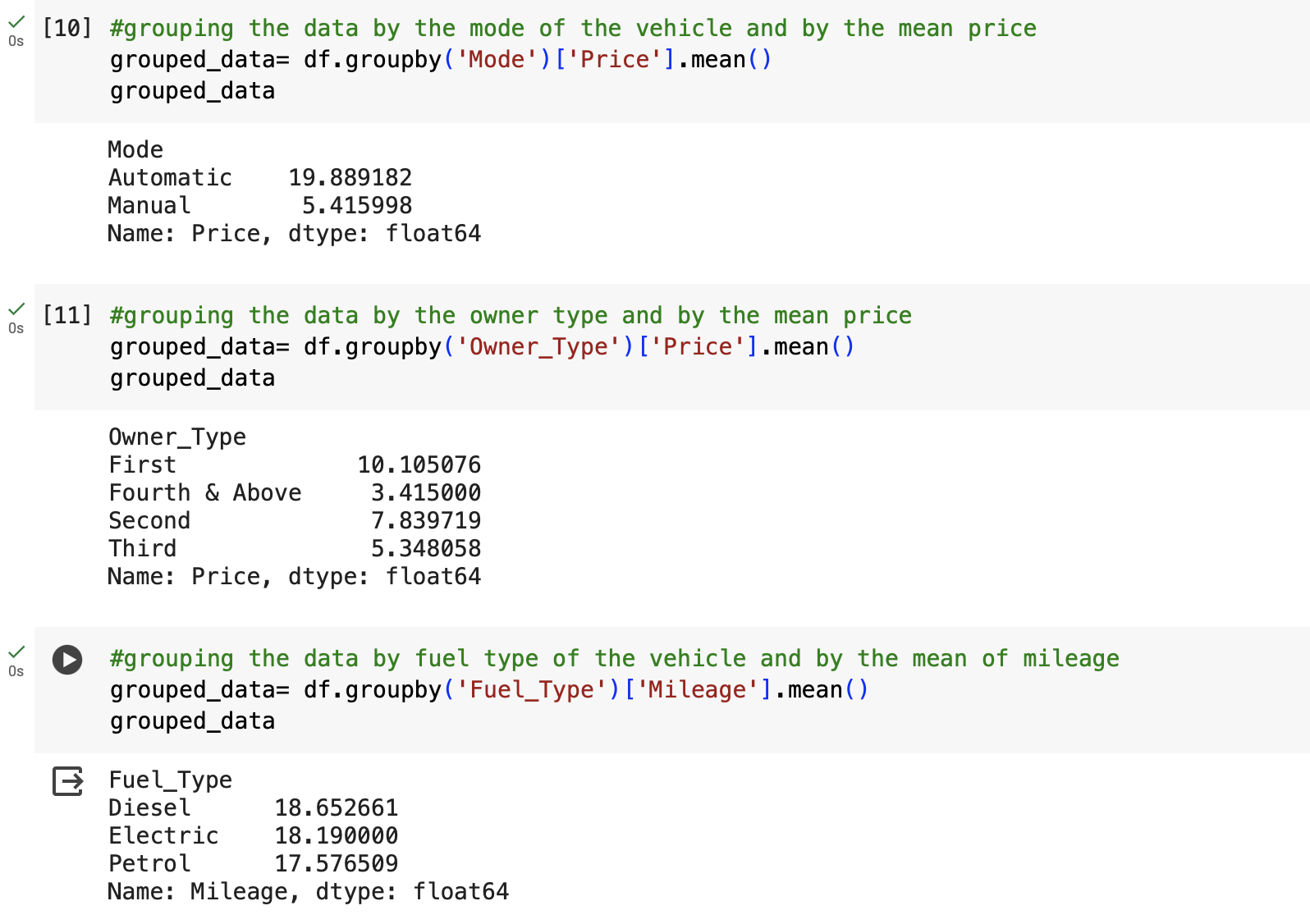
****

Here I arranged the data of the cars based on the year in the descending order which gives us the newest among the used cars.



And also the list of cars as per the ascending order of the kms driven. This list infers about the cars which are driven less since its manufacture. These are the cars most likely people are willing to buy.

**vi) group by**



In group by operations, I tried with different combinations. Firstly, I have grouped by the Mode of the vehicle with its mean price. From that, we can infer that the price of the Automatic cars are much higher than that of the manual cars. Secondly, I have grouped by the type of owner with the mean price. From this data, we can see that the First owner type cars are the highest and the cars with the owner\_type fourth and above are much cheaper than other cars. Lastly, I grouped by the data with the fuel type and its mean mileage. Diesel and electric vehicles are giving almost the same mileage and petrol cars are giving less mileage among the cars. All these operations have a certain impact in making a decision while buying used cars.