PRE-OWNED CAR

Gayatri Krishna

21BDA16

USED CAR MARKET IN INDIA

- In India the used car market is segmented by vehicle type (hatchbacks, sedan, and sports utility vehicles), fuel type (petrol, diesel, electric, CNG, LPG).

- The increased sale of used car is mainly found in metro cities and also a rise in online sales platforms, such as CarDekho, Cars24 etc.

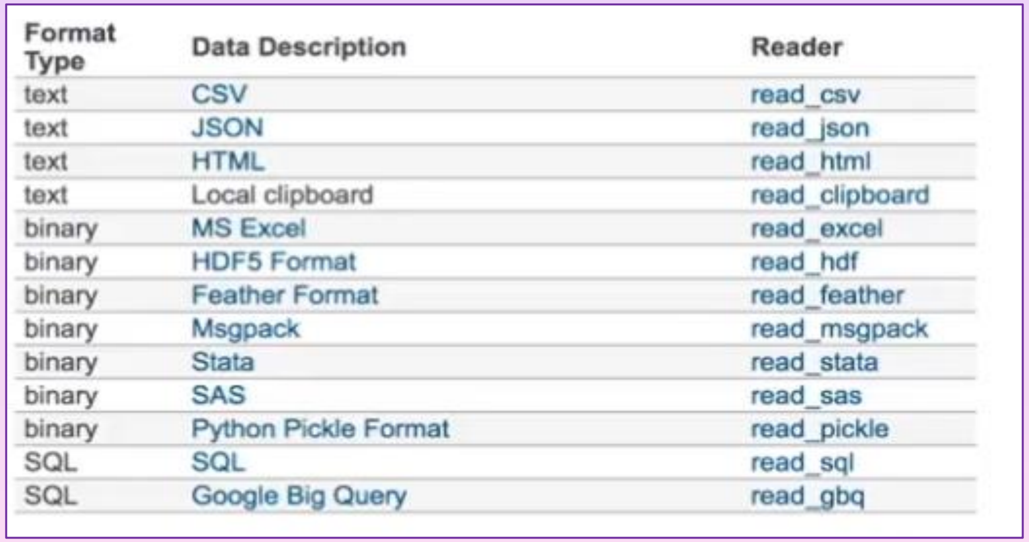
INTRODUCTION TO PANDAS

Pandas is a powerful Python data analysis toolkit for :

1. Reading different varieties of data
2. Functions for filtering, selecting and manipulating the data
3. Plotting data for visualization and exploration purposes

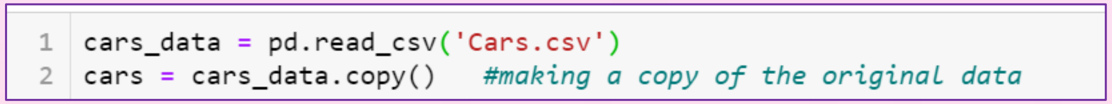
READING A SPREADSHEET FILE

Pandas can help us read data of different types of file.



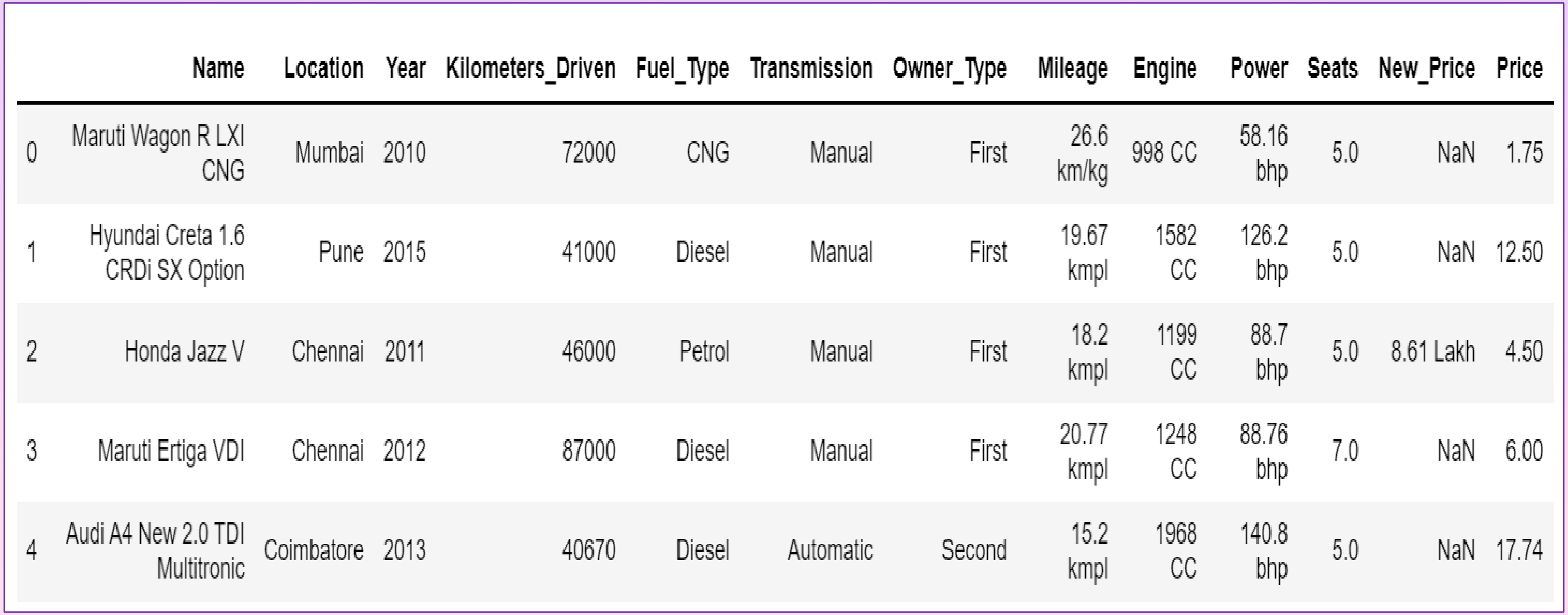
Format Type	Data Description	Reader
text	CSV	read_csv
text	JSON	read_json
text	HTML	read_html
text	Local clipboard	read_clipboard
binary	MS Excel	read_excel
binary	HDF5 Format	read_hdf
binary	Feather Format	read_feather
binary	Msgpack	read_msgpack
binary	Stata	read_stata
binary	SAS	read_sas
binary	Python Pickle Format	read_pickle
SQL	SQL	read_sql
SQL	Google Big Query	read_gbq

Here the dataset is csv



```
cars_data = pd.read_csv('Cars.csv')
cars = cars_data.copy()   #making a copy of the original data
```

DATASET



	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	NaN	1.75
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN	12.50
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN	17.74

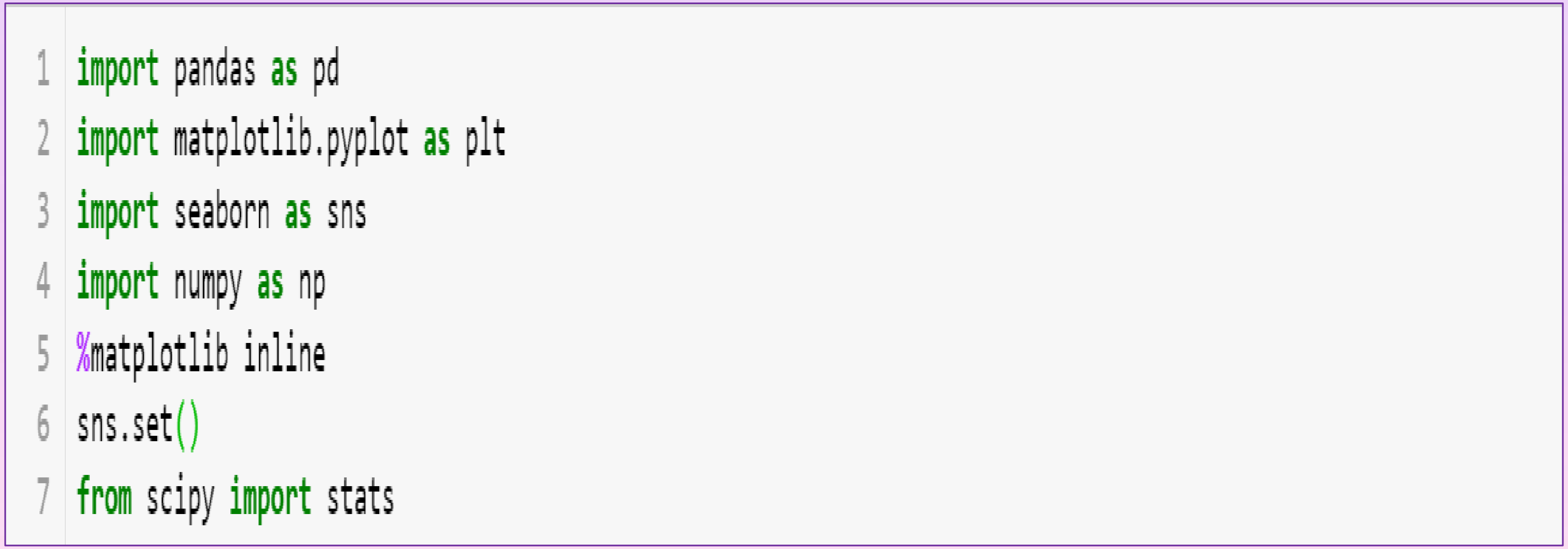
ABOUT THE DATASET

- The dataset is about the pre-owned cars from 1998 to 2019.

- There are 6019 rows and 13 columns in this dataset. The first 5 observations from the dataset is displayed.

- The dataset consist of the pre-owned cars in 11 different states in India.

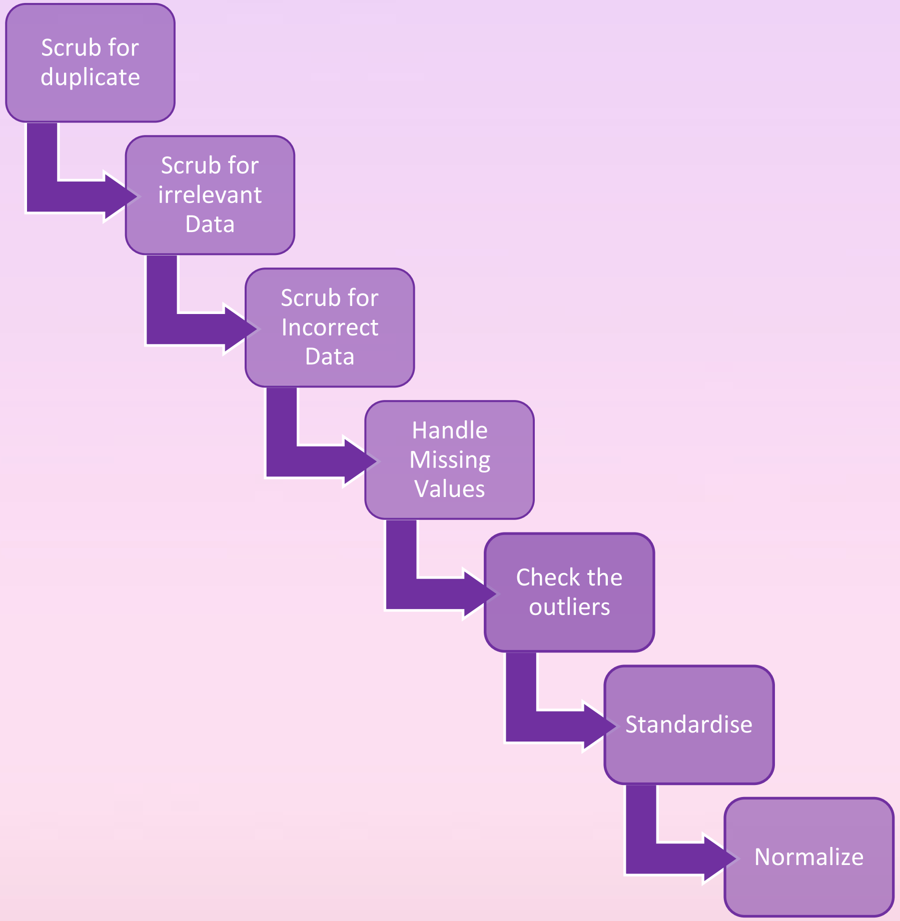
NECESSARY LIBRARIES IN PYTHON



```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
%matplotlib inline
sns.set()
from scipy import stats
```

DATA CLEANING

Data cleaning is the process of identifying, deleting, and/or replacing inconsistent or incorrect information from the database.



Scrub for duplicate

Scrub for irrelevant Data

Scrub for Incorrect Data

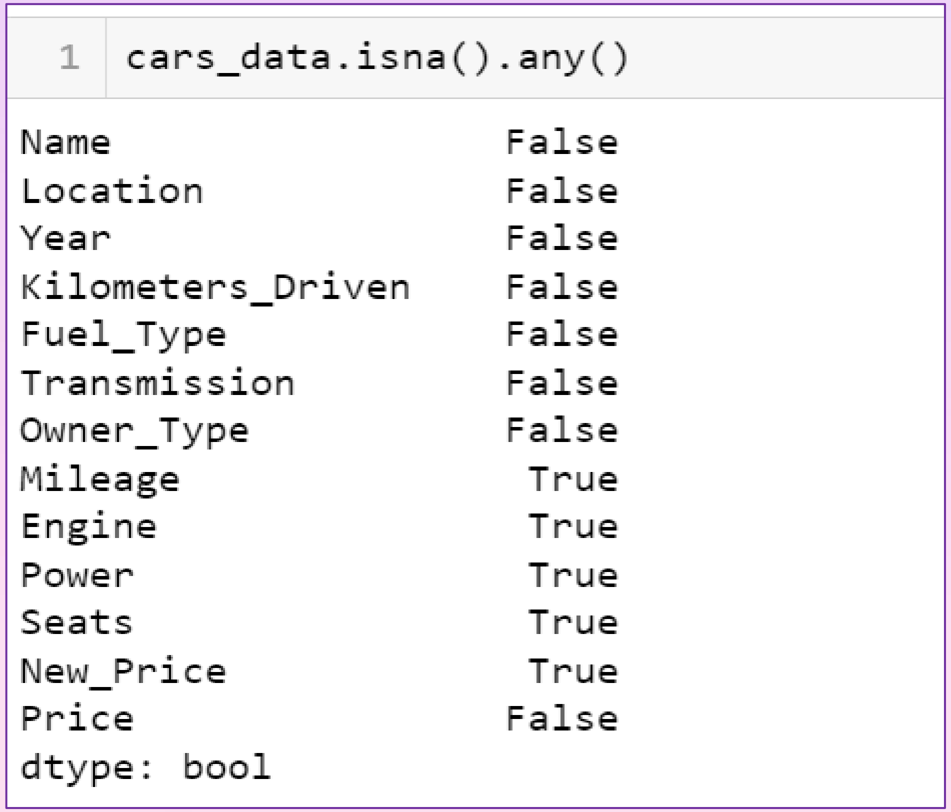
Handle Missing Values

Check the outliers

Standardise

Normalize

1. Find the out how many variables have missing values



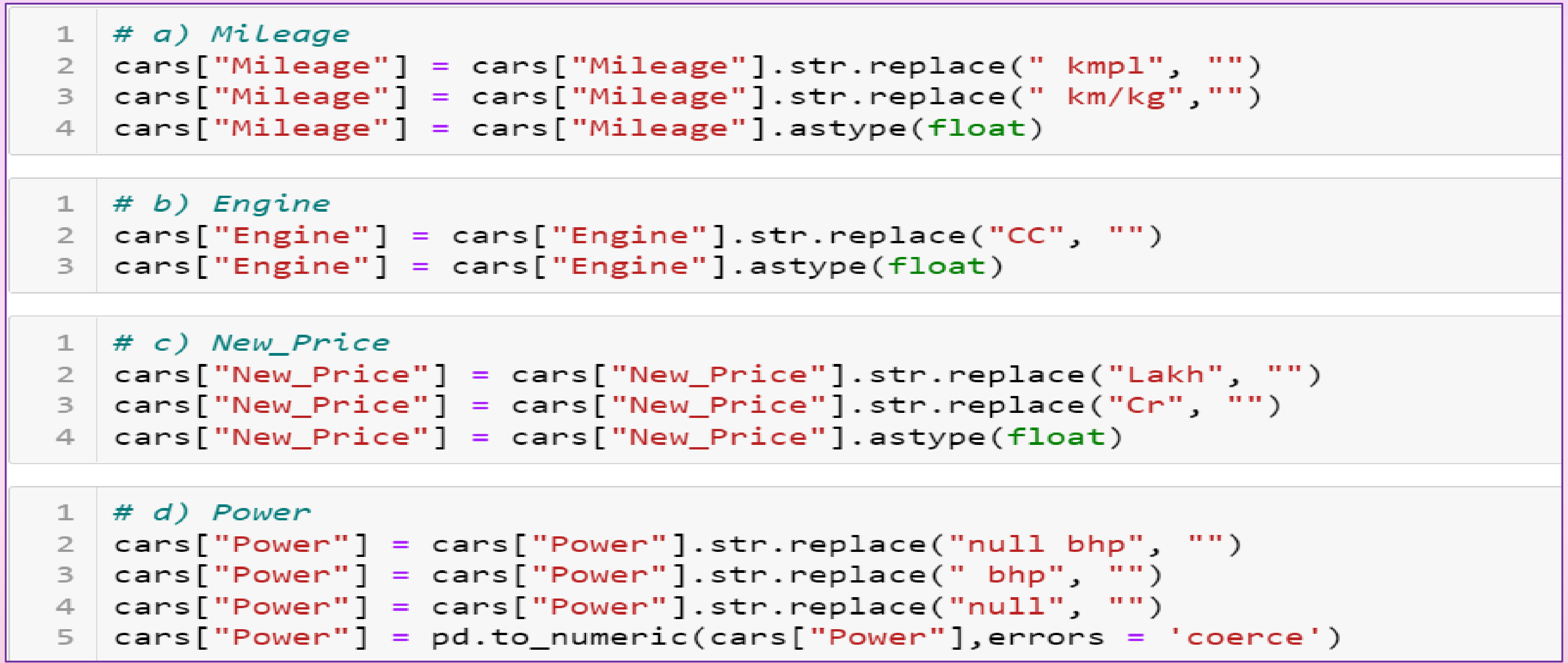
```
cars_data.isna().any()
```

```
Name                 False
Location             False
Year                 False
Kilometers_Driven    False
Fuel_Type            False
Transmission         False
Owner_Type           False
Mileage               True
Engine                True
Power                 True
Seats                 True
New_Price             True
Price                False
dtype: bool
```

So we see that Mileage, Engine, Power, New_Price and Seats have missing values… (Displayed by the boolean **True**). All other columns have complete information.

2. Removing the substring

Substrings are prefix or suffix of any string. Here Mileage, Engine and Power have substrings. So we replaced them and also converted the string type to float type to do statistical operations.



```
# a) Mileage
cars["Mileage"] = cars["Mileage"].str.replace(" kmpl", "")
cars["Mileage"] = cars["Mileage"].str.replace(" km/kg","")
cars["Mileage"] = cars["Mileage"].astype(float)
```

```
# b) Engine
cars["Engine"] = cars["Engine"].str.replace("CC", "")
cars["Engine"] = cars["Engine"].astype(float)
```

```
# c) New_Price
cars["New_Price"] = cars["New_Price"].str.replace("Lakh", "")
cars["New_Price"] = cars["New_Price"].str.replace("Cr", "")
cars["New_Price"] = cars["New_Price"].astype(float)
```

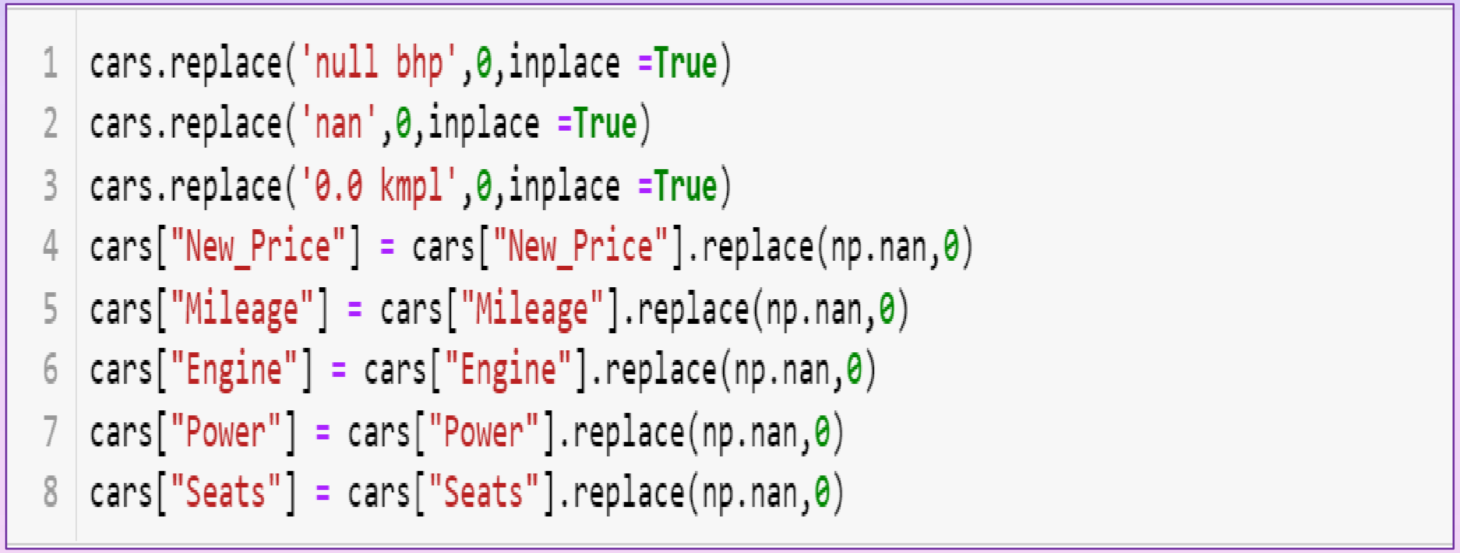
```
# d) Power
cars["Power"] = cars["Power"].str.replace("null bhp", "")
cars["Power"] = cars["Power"].str.replace(" bhp", "")
cars["Power"] = cars["Power"].str.replace("null", "")
cars["Power"] = pd.to_numeric(cars["Power"],errors = 'coerce')
```

3. Replacing the missing values by 0

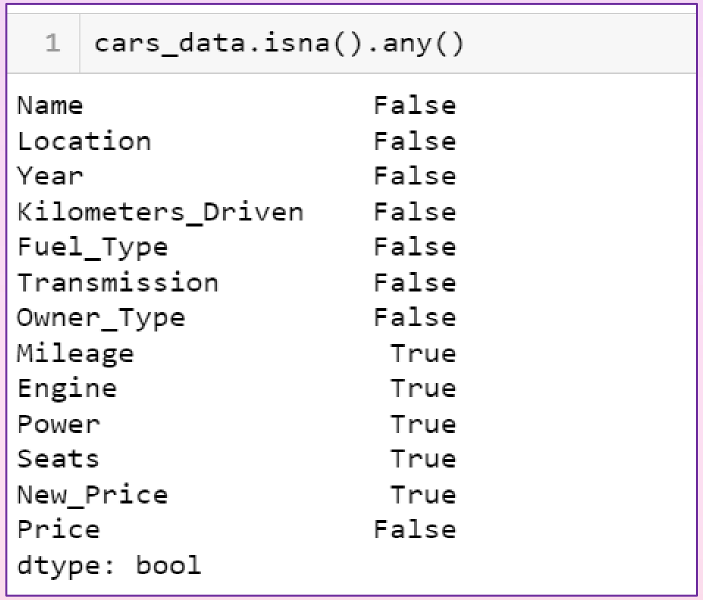
Here we use the replace() function to replace the missing values by 0. The inplace is an argument in pandas. The default value of this attribute is False and it returns the copy of the object.

- null bhp’ is present in the ‘Power’ column.
- ‘nan’ is present in some columns of the dataset.
- ‘0.0 kmpl’ is present in Mileage column.
- np.Nan present in all the columns

all are replaced by 0.

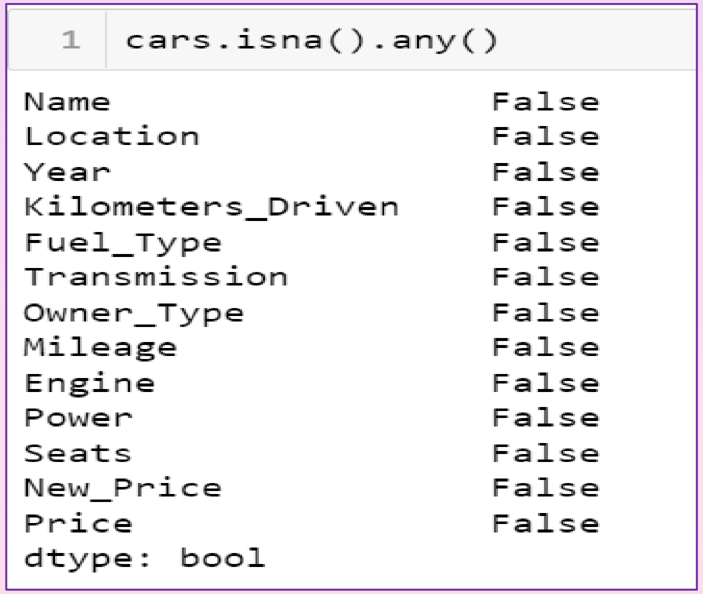


```
cars.replace('null bhp',0,inplace =True)
cars.replace('nan',0,inplace =True)
cars.replace('0.0 kmpl',0,inplace =True)
cars["New_Price"] = cars["New_Price"].replace(np.nan,0)
cars["Mileage"] = cars["Mileage"].replace(np.nan,0)
cars["Engine"] = cars["Engine"].replace(np.nan,0)
cars["Power"] = cars["Power"].replace(np.nan,0)
cars["Seats"] = cars["Seats"].replace(np.nan,0)
```



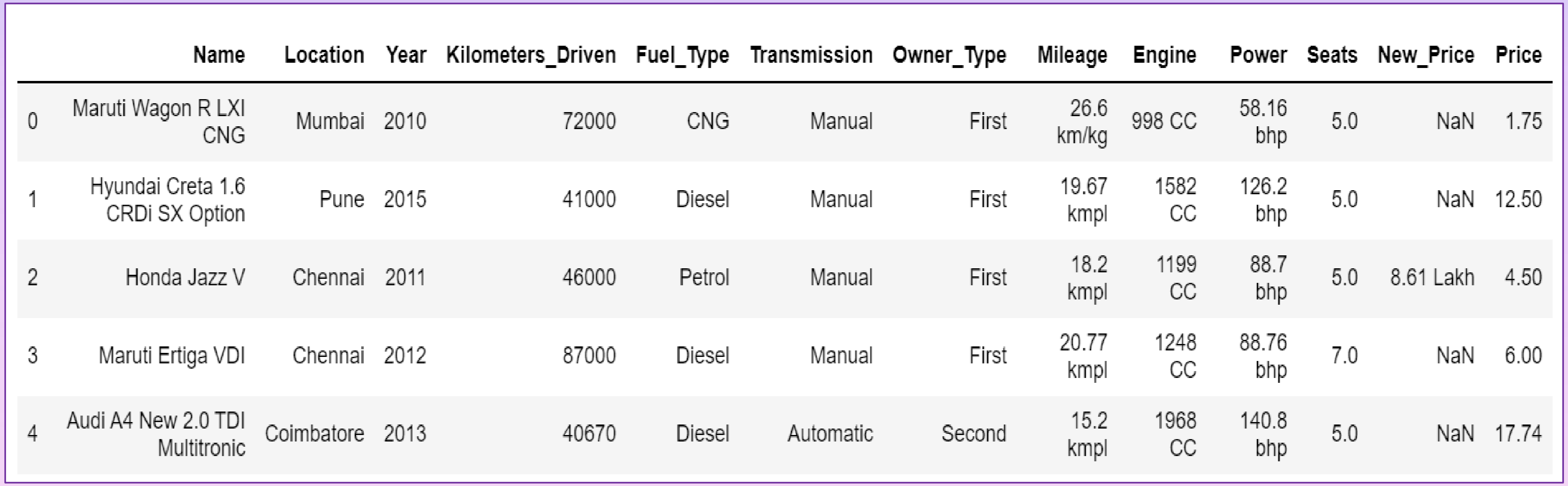
```
cars_data.isna().any()
```

```
Name                 False
Location             False
Year                 False
Kilometers_Driven    False
Fuel_Type            False
Transmission         False
Owner_Type           False
Mileage               True
Engine                True
Power                 True
Seats                 True
New_Price             True
Price                False
dtype: bool
```

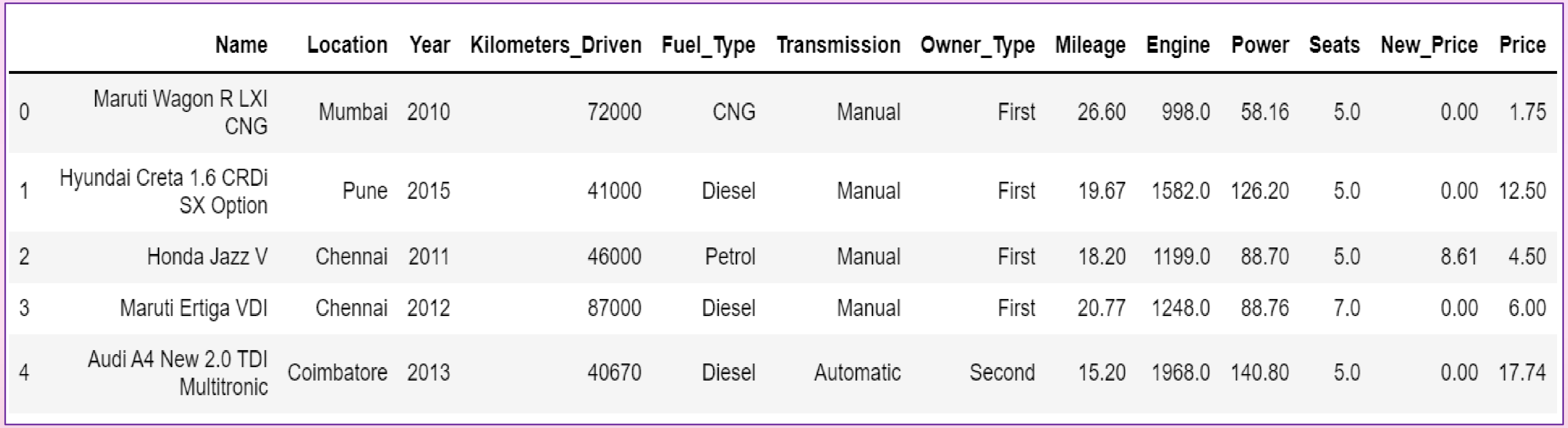


```
cars.isna().any()
```

```
Name                 False
Location             False
Year                 False
Kilometers_Driven    False
Fuel_Type            False
Transmission         False
Owner_Type           False
Mileage              False
Engine               False
Power                False
Seats                False
New_Price            False
Price                False
dtype: bool
```



	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	NaN	1.75
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN	12.50
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN	17.74

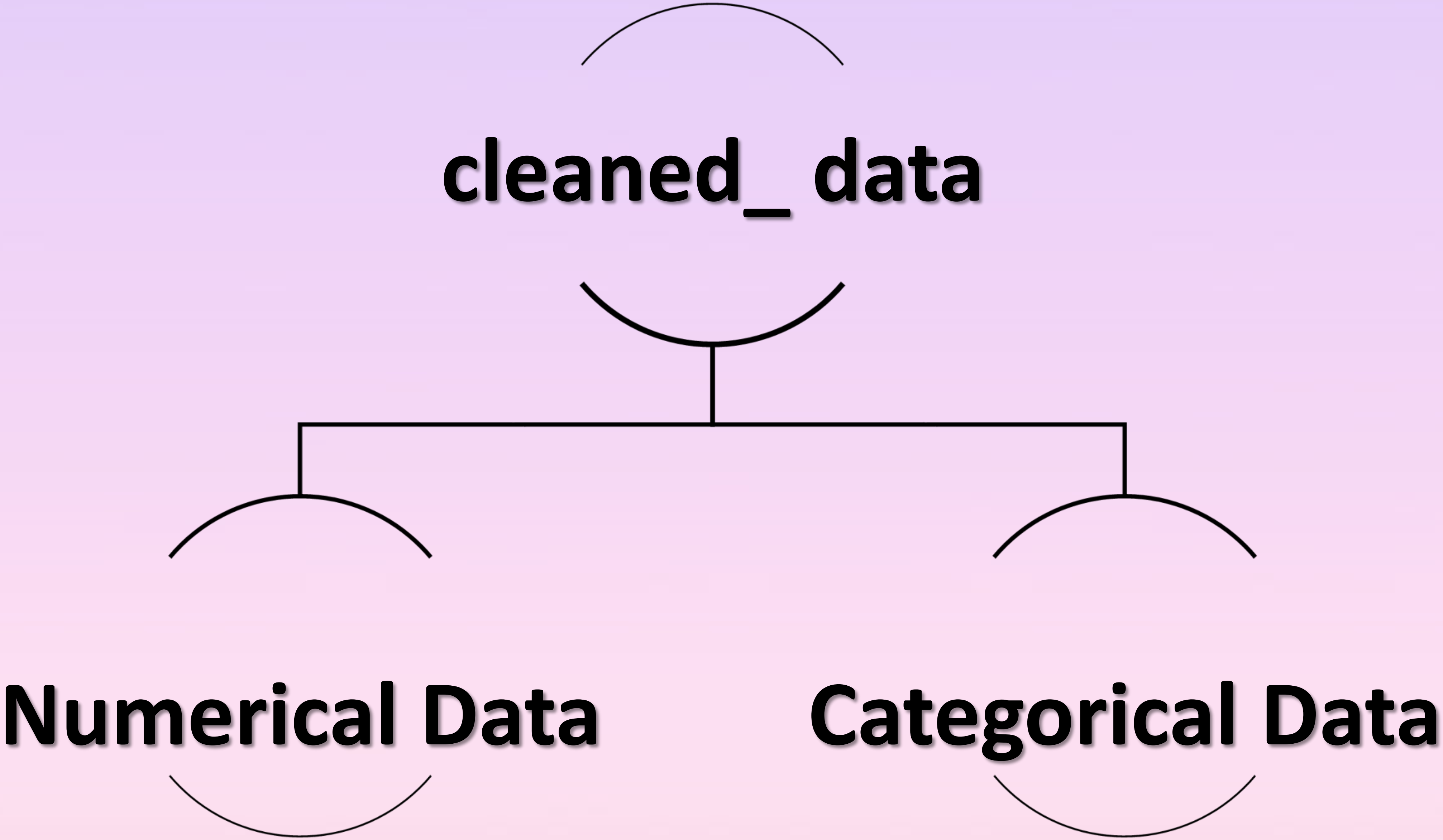


	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.60	998.0	58.16	5.0	0.00	1.75
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67	1582.0	126.20	5.0	0.00	12.50
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.20	1199.0	88.70	5.0	8.61	4.50
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77	1248.0	88.76	7.0	0.00	6.00
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.20	1968.0	140.80	5.0	0.00	17.74

DECRIPTION OF THE DATA

- The data is now cleaned and we have replaced all the null values with 0. Also in the data cleaning process the substring were also removed so as to help in further statistical analysis. The number of observations still remains the same (ie) 6019, it is not reduced.

- The data is copied to the variable name ‘cleaned_data’.



cleaned_ data

Numerical Data

Categorical Data

NUMERICAL DATA

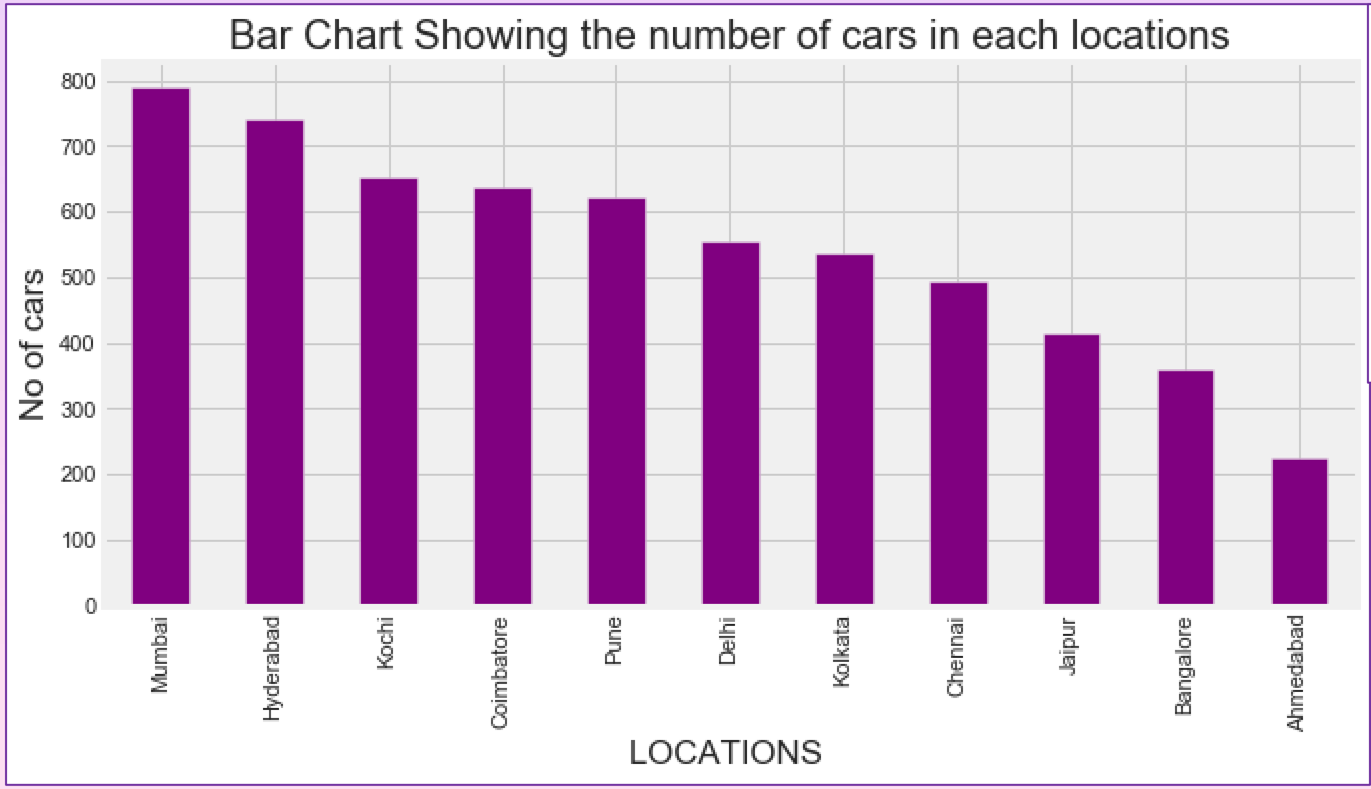
The data that has numerical values is called numerical data or quantitative data.
In the dataset we have Kilometers_Driven, Mileage, Engine, Power, Price as numerical data.

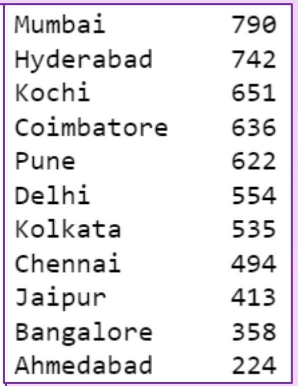
CATEGORICAL DATA

The data that has no numerical values(ie) has attributes is called categorical data or qualitative data.
In the dataset we have Seats, Locations, Year, Fuel_Type, Transmission, Owner_Type as categorical data.

DATA VISUALIZATION

1. Bar Plot showing the sales of cars in each location



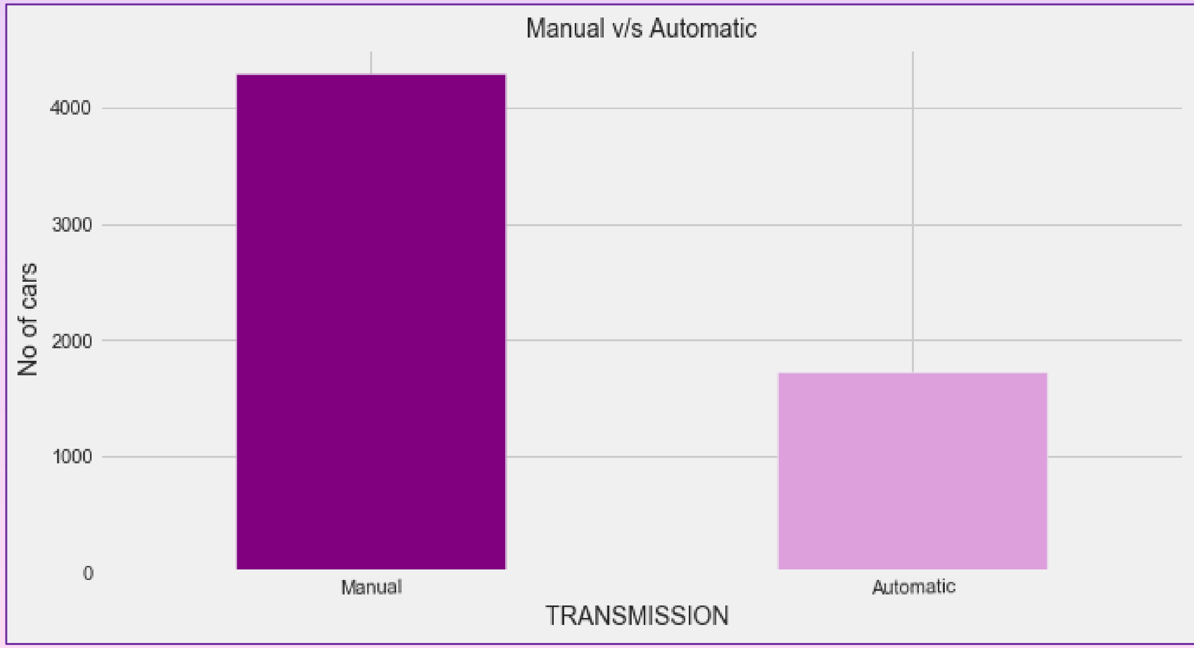


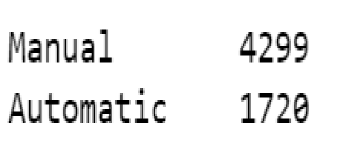
```
Mumbai        790
Hyderabad     742
Kochi         651
Coimbatore    636
Pune          622
Delhi         554
Kolkata       535
Chennai       494
Jaipur        413
Bangalore     358
Ahmedabad     224
```

The bar plot shows the number of cars in different location, we observe that the sale of cars is more in Mumbai followed by Hyderabad and Kochi. The lowest sales of car is Ahmedabad which is 224 units only.

2. Bar plot to show the number of cars in different transmission

- As we know that there are two transmissions (ie) automatic and manual. Clearly from the bar plot we see that the manual transmission is more than automatic transmission.
- The reason is that automatic cars though it came to India but it wasn't that famous. The automatic cars gained popularity from last 3 years.

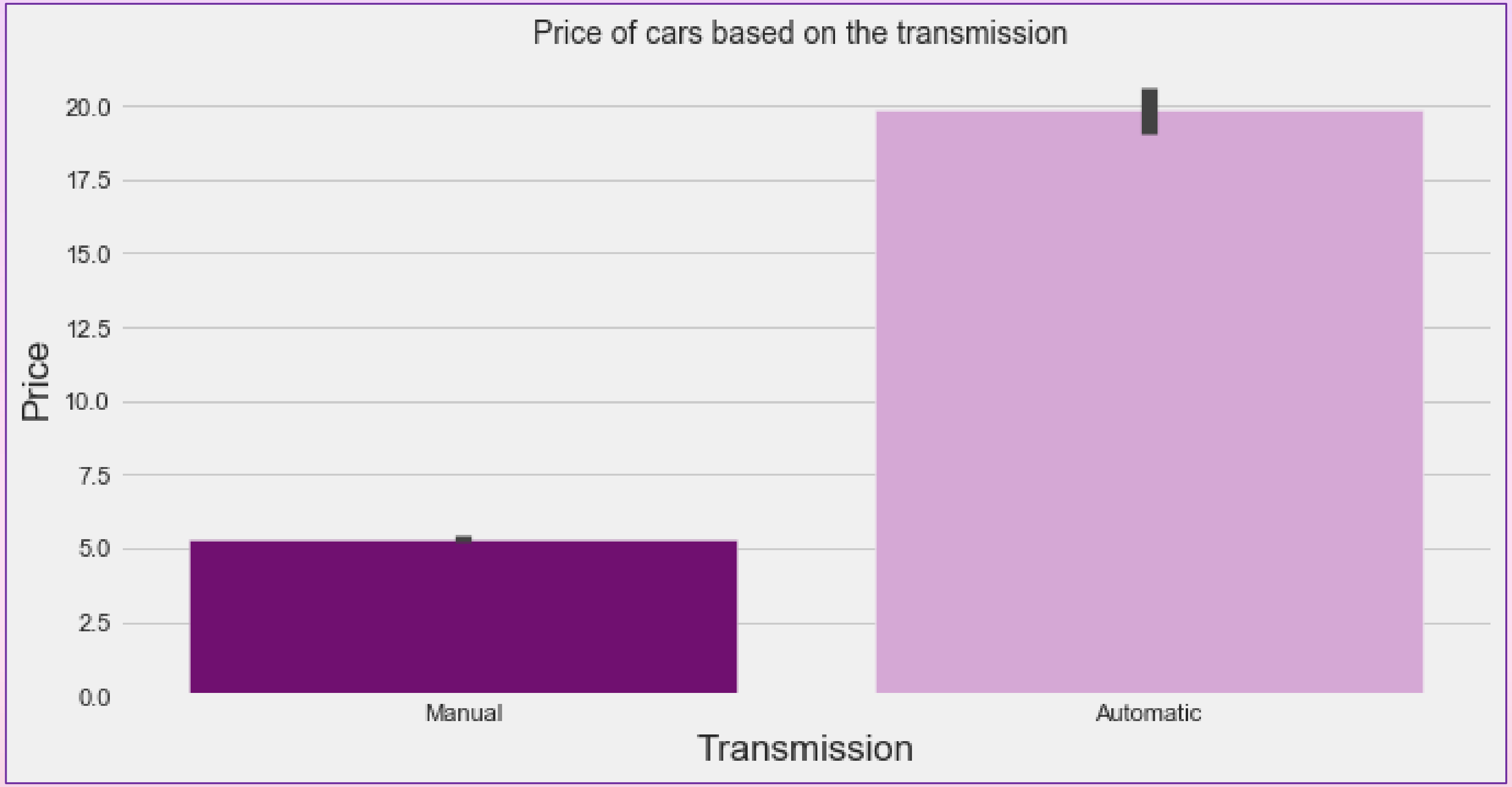




```
Manual       4299
Automatic    1720
```

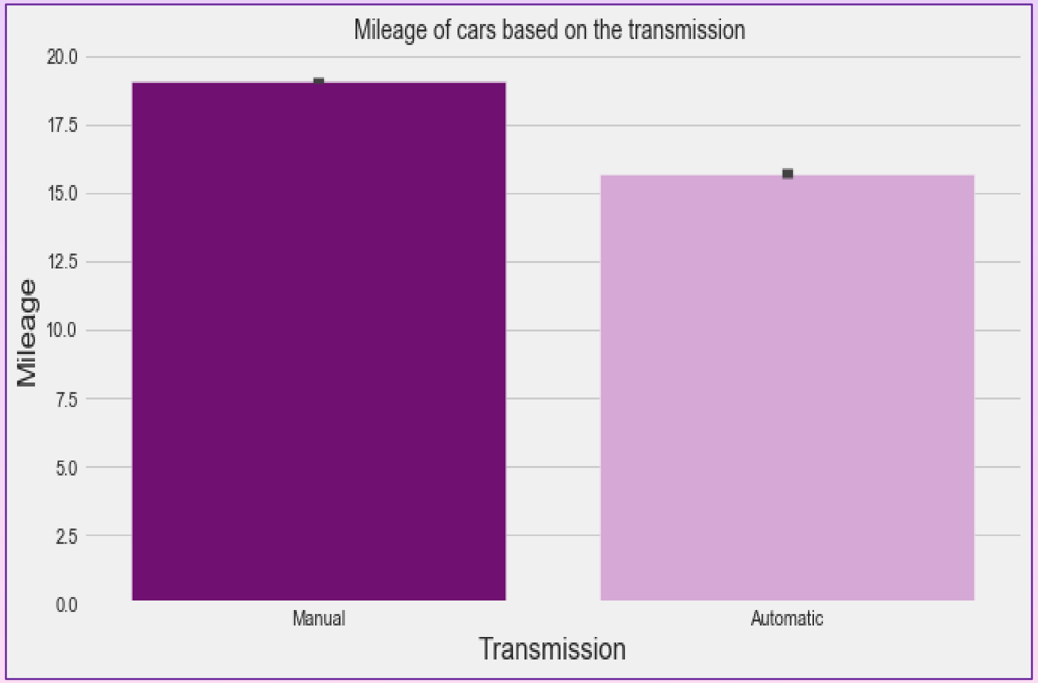
3. Bar Plot of Price vs Transmission

The below bar plot is shows the price of the cars based on the transmission. Clearly manual cars price less compared to automatic cars. So this could be one reason why the manual users were more than automatic users. The automatic cars are more expensive as the AT gearboxes cost carmakers more money as most of them are not made in India unlike the manual versions.

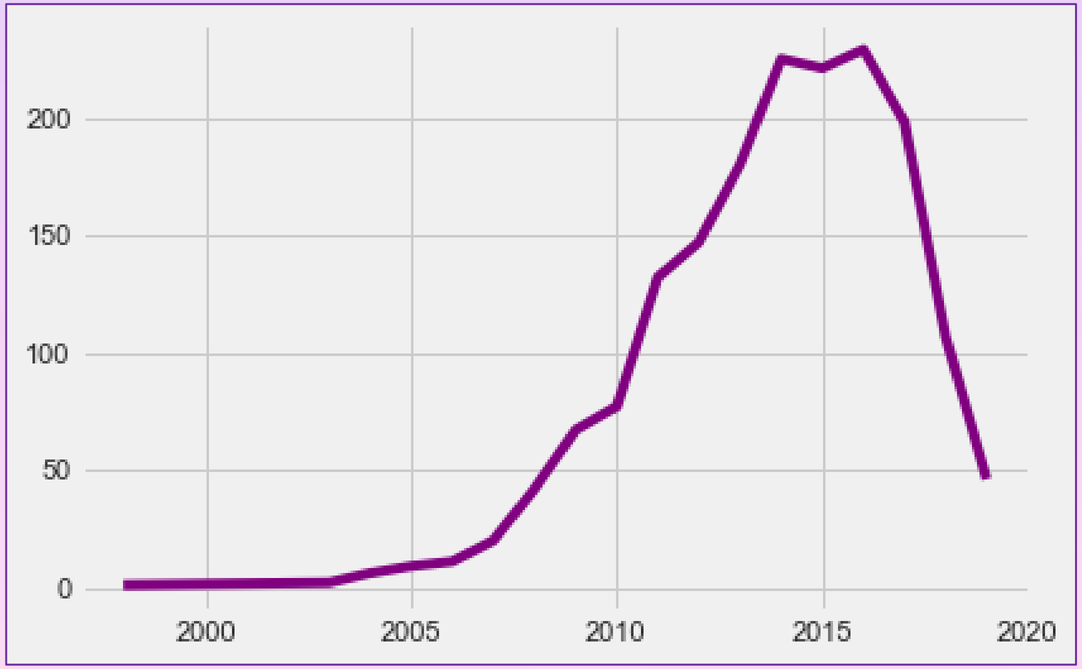


4. Bar Plot of Mileage vs Transmission

- The mileage offered by manual cars are more compared to automatic. But now the automatic cars have gained popularity.
- The buyers don't really bother about fuel consumption as we would imagine.
- We want to enjoy driving in our congested cities, comfort while driving and no headache of shifting gears. Just keep the gear lever on D (drive) mode and relax, accelerate and brake when needed by using only the right foot, while the left foot rests and you can drive with both hands on the steering wheel, eyes focused on the road.

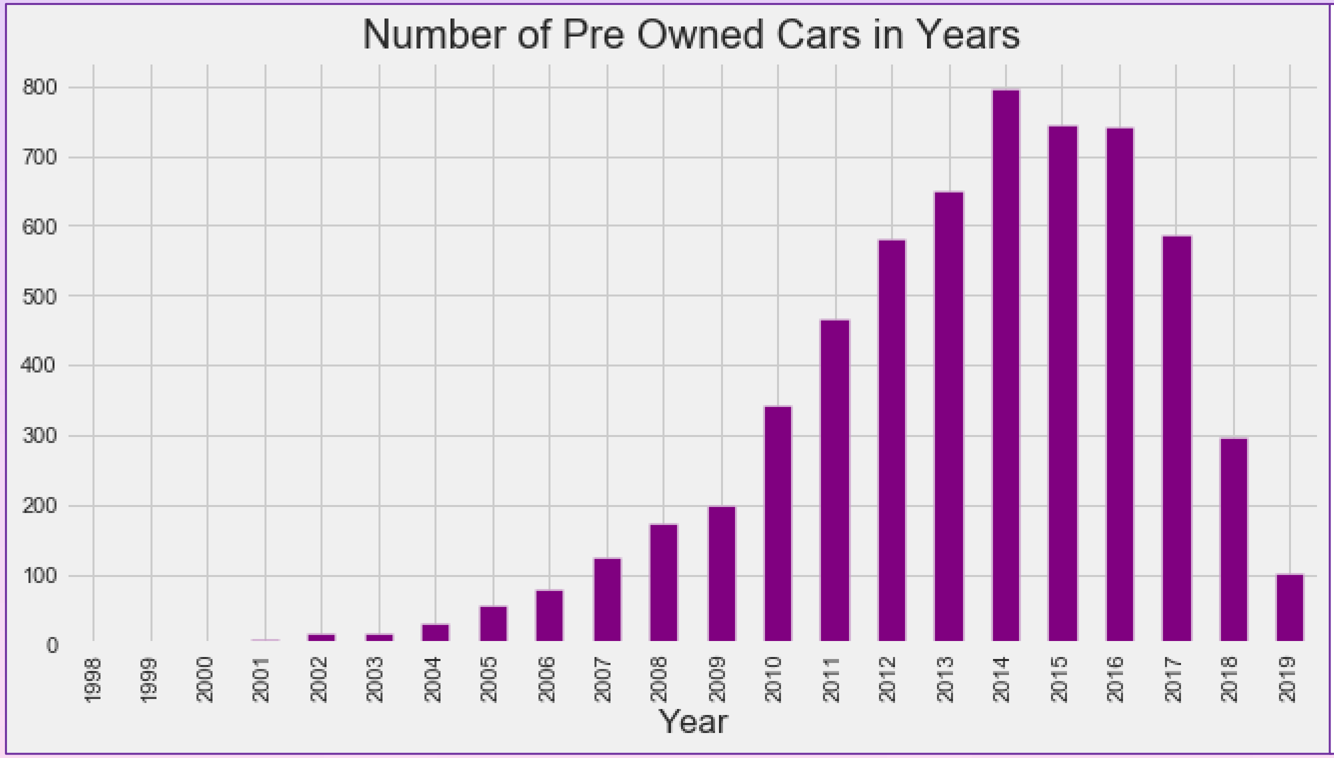


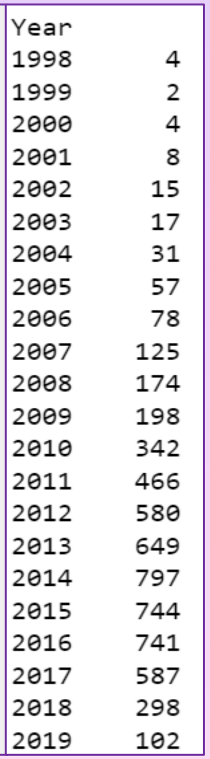
5. Demand Surge for Automatic Cars in India



From the adjacent graph, we see that the automatic cars gained in the year 2015. However we see that there is a decline in the demand. The reasons could be that mileage offered by the automatic cars is very low and also the price of the automatic cars are more than that of manual cars.

6. Number of Pre Owned Cars in Years

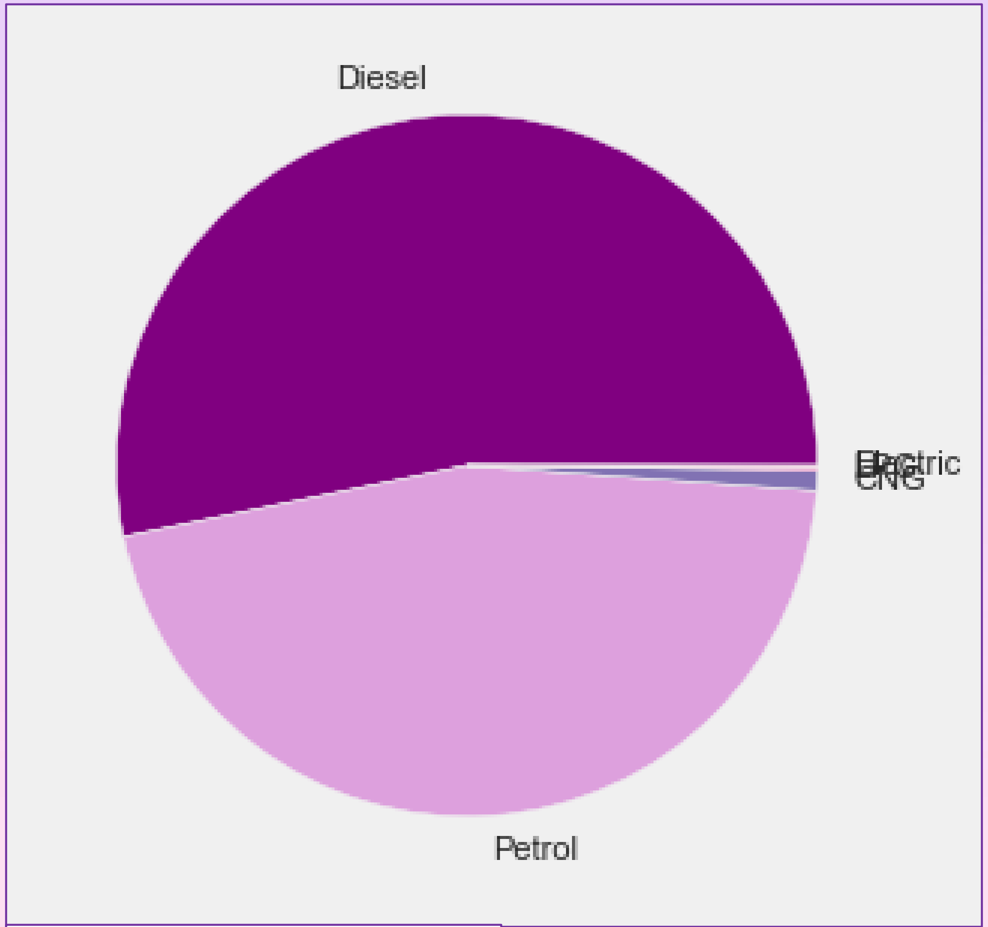


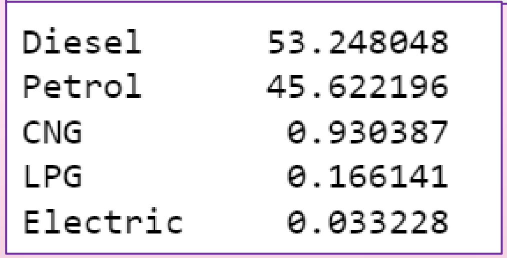


```
Year
1998       4
1999       2
2000       4
2001       8
2002      15
2003      17
2004      31
2005      57
2006      78
2007     125
2008     174
2009     198
2010     342
2011     466
2012     580
2013     649
2014     797
2015     744
2016     741
2017     587
2018     298
2019     102
```

From the above graph we see that most of the cars are in the year 2014 (ie) 797 units. The least is in the year 1999 (ie) 2 units

6. Fuel Type pie chart



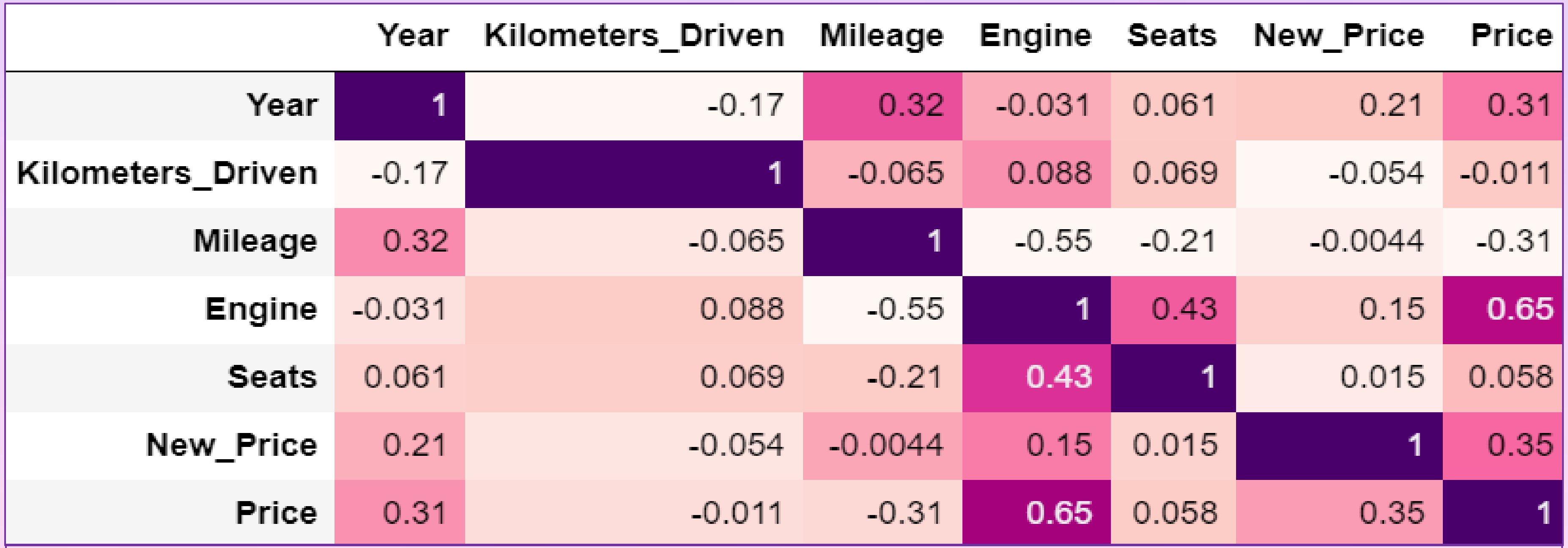


```
Diesel      53.248048
Petrol      45.622196
CNG          0.930387
LPG          0.166141
Electric     0.033228
```

It is seen that the consumer preference is for diesel driven fuel cars which is a bit surprising in the current context ; the price difference between petrol and diesel is marginal.

EV cars are slowly picking up, however availability of charging points is a challenge across cities.

CORRELATION



	Year	Kilometers_Driven	Mileage	Engine	Seats	New_Price	Price
Year	1	-0.17	0.32	-0.031	0.061	0.21	0.31
Kilometers_Driven	-0.17	1	-0.065	0.088	0.069	-0.054	-0.011
Mileage	0.32	-0.065	1	-0.55	-0.21	-0.0044	-0.31
Engine	-0.031	0.088	-0.55	1	0.43	0.15	0.65
Seats	0.061	0.069	-0.21	0.43	1	0.015	0.058
New_Price	0.21	-0.054	-0.0044	0.15	0.015	1	0.35
Price	0.31	-0.011	-0.31	0.65	0.058	0.35	1

From the above corrplot we see that there is strong positive correlation between Engine and Price.

THANK YOU