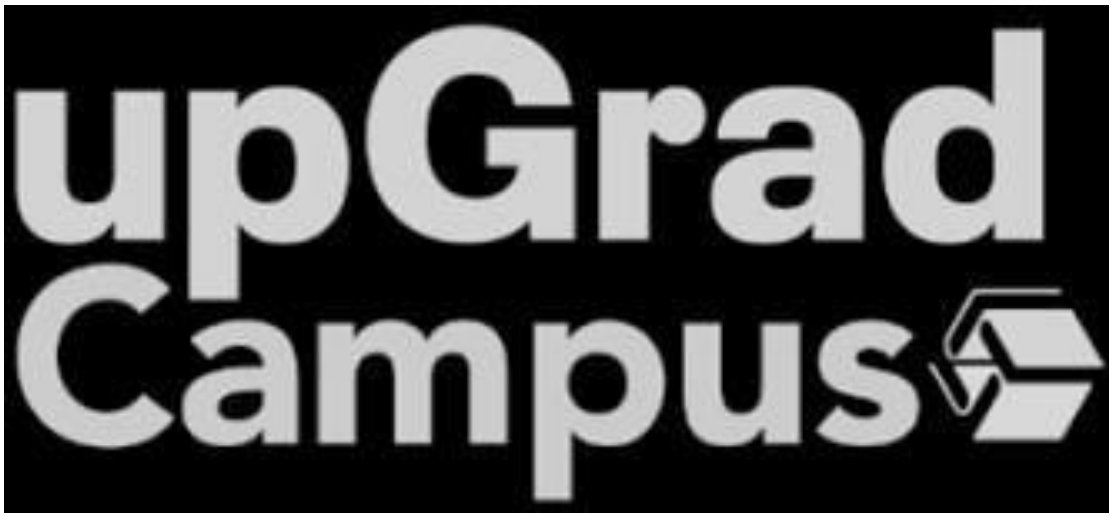
LOVELY PROFESSIONAL UNIVERSITY

NAME : Karri John Pradeep Reddy.

SECTION: K21UT
ROLL NO: RK21UTA15
Course Name: EDA PROJECT
Course Code: INT-353
Project Report of EDA Project on car sales of Quickr
Submitted to : Shivangini Gupta



Domain/Topic knowledge:

My dataset is cars selling prices and the features of the car.

Form this dataset we can get to know about the prices of the cars of different categories and there is 2059 different car data.

Also every datapoint is about the sale of the car in India. So if any company want to know about the data of sales in india this dataset gives an understanding of the car industry in india.

Selling Price:

We can be familiarize with the selling prices of the car and the factors which affect the price of the car.

performance and features:

For comparing the performance and features of different car models based on their mileage, engine power, torque, etc.

Trends in the used car market:

The dataset includes cars from different years, which can be useful for analysing trends in the used car market.

Fuel Efficient cars:

The dataset provides information about the fuel type, which can be useful for analysing the popularity of different types of fuel-efficient cars.

Knowledge on Indian car industry:

Many companies want start their production in India specifically the electric cars. This data can be very useful for those companies to start their sales with the features that these cars have.

Data Understanding:

1. **Make**: Company of the car. It is categorical data which basically gives the company name
2. **Model**: Name of the car. It is categorical data which tells the car's name.
3. **Price**: Selling price of the car in INR. It is Numeric data but all are unique numerical values, it basically gives selling price of the car.
4. **Year**: Manufacturing Year of the car. It is numeric data which gives year of manufacturing.
5. **Kilometer**: Total kilometres Driven. It is numeric data.
6. **Fuel_Type**: Fuel type of the car. It is ordinal data there are only three types of fuel 'Petrol', 'Diesel', 'CNG', 'LPG', 'Electric', 'CNG + CNG','Hybrid', 'Petrol + CNG', 'Petrol + LPG'.
7. **Transmission**: Gear transmission of the car. It is ordinal categorical data. It has only two type of entries Manual, Automatic
8. **Location**: City in which car is being sold. It is categorical data.
9. **Color**: Color of the car. It is categorical data.
10. **Owner**: Number of previous owners. It is categorical data, It has 'First', 'Second', 'Third', 'Fourth', 'Unregistered Car', '4 or More'.
11. **Seller_Type**: it is a categorical value which have values like 'Corporate', 'Individual', 'Commercial Registration'.
12. **Engine**: it is a numerical value which gives the horse power of the car.
13. **Max_Power**: it gives the maximum power of the engine and is a numerical data.

14. **Max_Torque**: it gives information about maximum torque of the engine and it is numeric data.
15. **Drivetrain**: it basically gives information of the cars drive train like 'FWD', 'RWD', 'AWD'.
16. **Length**: it tells about the total length of the car horizontally. It is a numeric value.
17. **Width**: it gives the information about width of the car horizontally it is a numeric data.
18. **Height**: it tells information about height of the car from ground vertically and it is numeric data.
19. **Seating-Capacity**: it gives the information about how many members can a car have at max. it is a numeric value.
20. **Fuel_Tank_Capacity**: it gives information about total fuel tank capacity of the car it will be used for the maximum milage of the car. It is a numeric data.

By seeing the above information we can get the all the information about the dataset I have chosen.

Reasons:

a car details dataset can serve a wide range of purposes, from helping individuals make informed car-buying decisions to supporting research, industry analysis, and regulatory efforts in the automotive domain.

Car enthusiasts: Anyone can access car details datasets for various purposes, including building custom cars, tracking the value of collectible cars, or participating in motorsports.

Educational Purposes: I am interested in data science, analytics, and statistical analysis so I choose this dataset for educational projects, classroom exercises, or research assignments.

Curiosity on car sales: I have a great curiosity on cars and their features so I choose this dataset to basically understand about the car market in India and to know whether used cars can be used or the new ones.

These are some reasons why I have chosen this dataset.

Libraries Used :

Numpy:

"""numpy is used for numerical calculations in python""" EX: cars['Engine']=np.where(cars["Engine"].isna(),cars["Engine "].median(),cars["Engine"]) Pandas:

"""pandas is used for data preprocessing, data manipulation, specifically for spreadsheat data and also used to work with data frames, series""" Pandas provides two primary data structures - DataFrames and Series, which are highly versatile and allow for the representation and manipulation of structured data.

EX: cars["Age"]=cars["Current_year"]-

cars["Year"] cars.head(3) data manipulation:

cars["Current_year"]=2023

Data Import/Export: Pandas supports various file formats, such as CSV, Excel, SQL databases, and more, making it easy to read and write data from different sources.

df=pd.read_csv("BlackFriday.csv")# Load data from a CSV file

4. Data Manipulation: You can perform operations like filtering, sorting, merging, and reshaping data effortlessly, making it a crucial library for data wrangling tasks. # Extract numerical power values and RPM values cars['Power_bhp'] = cars['Max_Power'].str.extract(r'(\d+) bhp').astype(float) cars['Engine_RPM'] = cars['Max_Power'].str.extract(r'@ (\d+) rpm').astype(float) **Matplotlib:**

"""matplotlib is used for plotting the graphs this library is specifically for data visualization purposes only"""

Seaborn:

"""Seaborn is based on matplotlib which does higher level visualizations which works on top of matplotlib to give plots more interactive and particularly used for statistical and exploratory data analysis also it is good in integration with pandas."""

Seaborn is a visualization library that complements pandas and makes it easier to create aesthetically pleasing statistical graphics. It offers a high-level interface for drawing attractive and informative statistical graphics, making data visualization more accessible and allowing you to quickly explore and understand data patterns.

1.Data Visualization: Seaborn specializes in statistical data visualization, offering a high-level interface for creating attractive and informative plots.
sns.countplot(cars.Drivetrain) plt.title("Drive train Count") plt.xticks(rotation=100) plt.show()

3. Statistical Plots: Seaborn provides a wide range of statistical plots, including scatter plots, bar plots, histograms, and regression plots, which are particularly useful for data exploration and analysis.

categorical_columns = ["Make", "Fuel Type", "Transmission", "Location", "Color", "Owner", "Seller Type", "Drivetrain"] for col in categorical_columns: sns.countplot(data=cars, x=col) plt.figure(figsize=(10, 5)) plt.title(f"{col} Count") plt.xticks(rotation=0) plt.show()

4. Categorical Data: It excels at handling categorical data, making it straightforward to create categorical plots like bar plots, box plots, violin plots, and swarm plots.

5. Multivariate Analysis: It supports complex multivariate visualizations, such as pair plots and heatmap correlations, which are essential for understanding relationships between multiple variables.

```
plt.figure(figsize=(10, 6)) correlation_matrix
= cars.corr()
sns.heatmap(correlation_matrix,
annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap') plt.show()
```

Approaches followed:

Data Cleaning:

Check for missing values in each column and decide how to handle them (imputation or removal).

Remove duplicates, if any, in the dataset.

Ensure data consistency and accuracy.

Data Summary:

Calculate basic summary statistics (mean, median, standard deviation, etc.) for numerical columns like Price, Year, Kilometer, Engine, etc.

For categorical columns like Make, Fuel_Type, Transmission, Location, Color, Owner, Seller_Type, and Drivetrain, summarize the frequency distribution of categories.

Visualize data distributions and relationships.

Data Visualization:

Creating visualizations to understand the distribution of Price, Kilometer, Engine, and other numerical features. Histograms, box plots, and scatter plots are useful for this.

Explore the relationships between categorical variables and the target variable (e.g., Price) using bar charts or violin plots.

Analyze trends over time by visualizing Year against Price or other relevant attributes.

Using pair plots or correlation matrices to identify relationships between numerical variables. **Outlier Detection and Handling:**

Identify outliers in numerical columns using box plots or scatter plots.

Decide whether to remove or transform outliers based on domain knowledge.

Approaches for specific columns for data pre-processing:

- For price column I removed the punctuation so that it will be integer data type so that I can do numeric analysis.
- For Engine column I removed the “cc” after every entry again to make it as int data type column.
- For maximum power and torque I divided the columns for better Understanding and I sorted the cars based on their power in bhp
- If power is equal for two cars then I checked the rpm if car1 is greater then it should come first.

Data Pre-Processing

```
In [51]: 1 cars["Engine"]=cars["Engine"].str.split(" ").str.get(0)

In [52]: 1 cars['Engine']=np.where(cars["Engine"].isna(),cars["Engine"].median(),cars["Engine"])

In [53]: 1 cars["Engine"]=cars["Engine"].astype(int)

In [54]: 1 cars.head()
```

- For removing null values first I tried capping the values with median but that didn’t work as in the same rows other values are also missing so I tried to analyse the missing values Then I came to know that they are electric cars as these cannot be in this dataset as this dataset contains only fuel based cars in more proportion.
- I dropped all the electric vehicles they are approximately 25 cars in the dataset.
- After I was searching for the null values in fuel capacity column but a mysterious fact has been unveiled that 29 BMW cars are not having tank capacity information so I deleted all the 29 rows ass they can hamper the analysis.

- The Engine column is comprised of maximum power and its rpm but this is stored in object data type so by using string slicing I extracted the power in bhp and stored it as “Power_bhp” and its rpm is stored as “Engine_Rpm”.
- Same process I have followed for the max_torque column and divided it into “Torque_Nm”,”RPM”.
- Deleted the “Max_Torque ” and “Max_Power” columns because they are extracted as 4 different columns.
- **Null Values Treatment:**

```
In [21]: cars = cars[~((cars['Engine'].isnull()) & (cars['Fuel_Type'].eq('Electric')))]

In [22]: cars['Engine'].isnull().sum()
Out[22]: 73

In [24]: same_null_rows = cars['Engine'].isnull().equals(cars['Seating_Capacity'].isnull())

In [25]: same_null_rows
Out[25]: False

In [26]: cars=cars[cars['Engine'].notnull()]

In [27]: cars['Engine'].isnull().sum()
Out[27]: 0
```

- I haven’t removed all the null it leads to data loss
- For treating the null values I followed both imputation, removal of the null values.

```
In [31]: bmw=cars[(cars['Make']=='BMW')]

In [32]: bmw['Fuel_Tank_Capacity'].mode()
Out[32]: 0    65.0
         Name: Fuel_Tank_Capacity, dtype: float64

In [33]: # filling the null values of a column with specific condition
         cars.loc[cars['Make'] == 'BMW', 'Fuel_Tank_Capacity'] = cars.loc[cars['Make'] == 'BMW', 'Fuel_Tank_Capacity'].fillna(65.0)
```

- For this I fllowed Z-Score method to consider the outliers if they are out of specific range.
- Also I have plotted boxplots for getting visual representation of the data distribution.
- I haven’t cleared all the values because it will cause a data loss.

- This Fuel Tank capacity column is after the clearing of the outliers although it may have some outliers I didn't consider them as outliers as they are continuous.
- I have spotted electric cars in the dataset but in less quantity as they are new in the market I removed them as they don't have engine specifications. They can hamper the analysis of whole data.

```
In [38]: 1 cars=cars[cars['Fuel_Tank_Capacity'].notnull()]

In [39]: 1 cars['Fuel_Tank_Capacity'].isnull().sum()

Out[39]: 0

In [40]: 1 cars.isnull().sum()

Out[40]: Make                0
         Model               0
         Price               0
         Year                0
         Kilometer           0
         Fuel_Type           0
         Transmission        0
         Location            0
         Color               0
         Owner               0
         Seller_Type         0
         Engine              0
         Max_Power           0
         Max_Torque          0
         Drivetrain          0
         Length              0
         Width               0
         Height              0
         Seating_Capacity    0
         Fuel_Tank_Capacity  0
         dtype: int64
```

- **Now the dataset is null values free**

Duplicate Values:

```
In [41]: 1 cars[cars.duplicated()]
```

Out[41]:

Make	Model	Price	Year	Kilometer	Fuel_Type	Transmission	Location	Color	Owner	Seller_Type	Engine	Max_Power	Max_Torque	Drivetrain	Length	Widt

So there are no duplicate values in the dataset

There are null values but they are removed when removing the null values.

Outliers Treatment:

In [57]:

```
from scipy import stats
#removing the outliers
column_name = 'Price'

# Calculate z-scores
z_scores = np.abs(stats.zscore(cars[column_name]))

# Define a threshold
threshold = 5

# Create a mask to identify outliers
outlier_mask = (z_scores > threshold)

# Extract the rows that contain outliers
outliers_cars = cars[outlier_mask]
no_outlier_mask = (z_scores <= threshold)

```

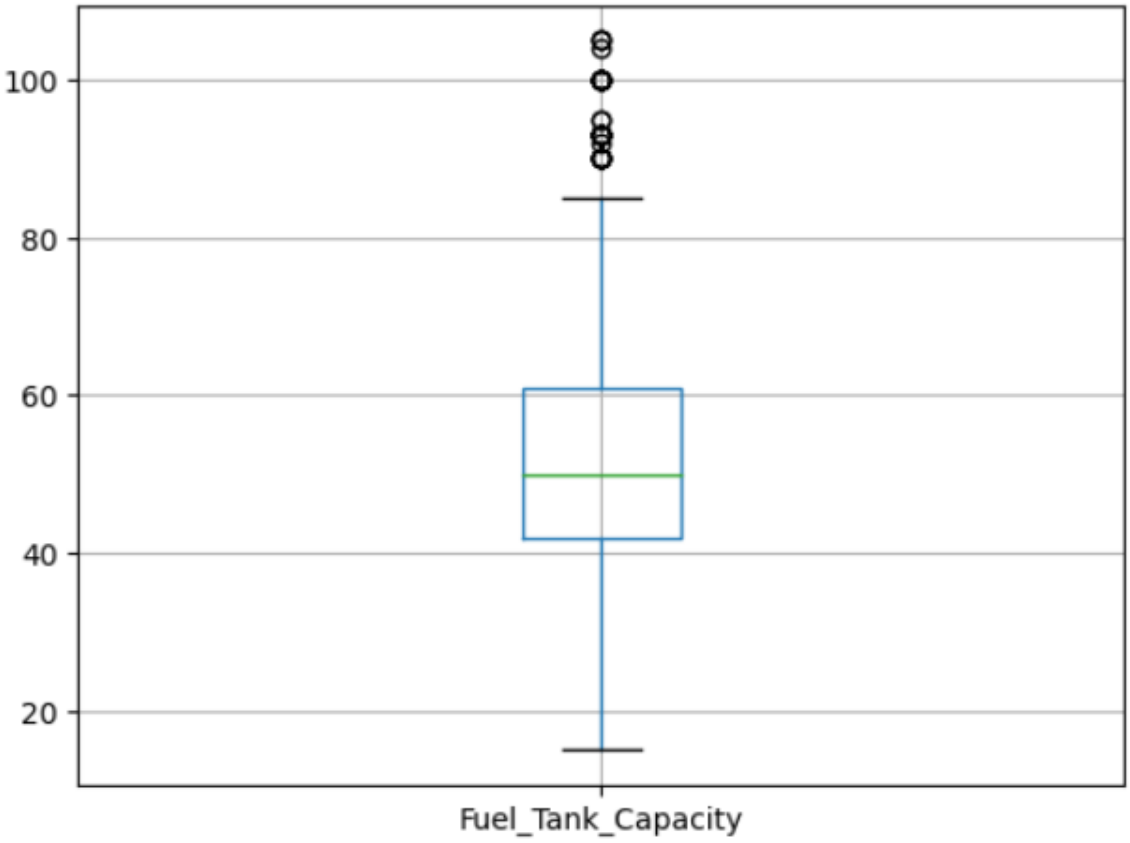
```
In [58]: 1 outliers_cars
```

Out[58]:

	Make	Model	Price	Year	Kilometer	Fuel_Type	Transmission	Location	Color	Owner	Seller_Type	Engine	Max_Power
95	Porsche	Cayenne Coupe Platinum Edition	16200000	2022	2766	Petrol	Automatic	Mumbai	Blue	First	Individual	2995	335 bhp @ 5300 rpm
442	Mercedes-Benz	S-Class Maybach S 560	18500000	2021	21000	Petrol	Automatic	Gurgaon	Black	First	Individual	3982	463 bhp @ 5250 rpm
483	Ferrari	488 GTB	35000000	2018	9500	Petrol	Automatic	Delhi	Black	First	Individual	3902	660 bhp @ 8000 rpm
582	Land Rover	Range Rover 3.0 V6 Diesel Vogue	22000000	2019	35000	Diesel	Automatic	Pune	Blue	First	Individual	2993	244 bhp @ 4000 rpm
977	Rolls-Royce	Ghost 6.5	18000000	2011	60000	Petrol	Automatic	Mumbai	Maroon	Second	Corporate	6592	570 bhp @ 5250 rpm
1246	Rolls-Royce	Ghost Extended Wheelbase	20000000	2011	27000	Petrol	Automatic	Delhi	Blue	Third	Individual	6592	570 bhp @ 5250 rpm

```
In [60]:  1 pd.Series(z_scores).describe()

Out[60]: count    1976.000000
         mean        0.596487
         std         0.802826
         min         0.000701
         25%         0.313586
         50%         0.466844
         75%         0.577631
         max        13.920858
         Name: Price, dtype: float64
```



The Above image is after cleaning the outliers I left some values because they are continuous.

Exploratory Data Anslysis:

EDA stands for Exploratory Data Analysis. It is a critical step in the data analysis process that involves examining and visualizing data to summarize its main characteristics, often with the help of statistical graphics and other data visualization techniques. The primary goal of EDA is to gain insights into the data, discover patterns, detect anomalies, and formulate hypotheses for further analysis.

Univariate Analysis:

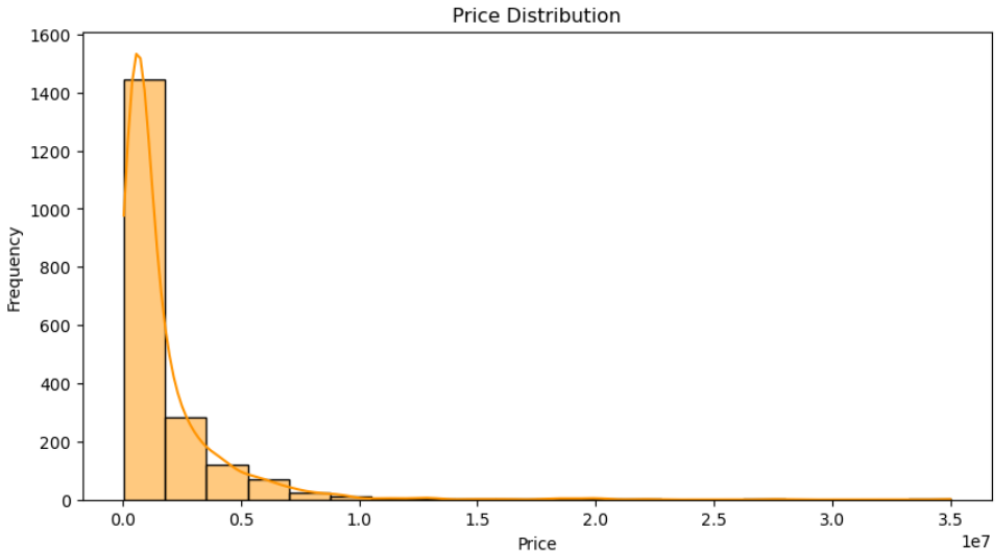
Firstly Analysing the statistical description of all the numerical values.

	Price	Year	Kilometer	Engine	Length	Width	Height	Seating_Capacity	Fuel_Tank_Capacity	Age
count	1.874000e+03	1874.000000	1.874000e+03	1874.000000	1874.000000	1874.000000	1874.000000	1874.000000	1874.000000	1874.000000
mean	1.718279e+06	2016.713447	5.317814e+04	1682.829242	4281.512807	1767.886339	1588.967983	5.295091	52.217343	6.286553
std	2.426090e+06	3.138477	5.878833e+04	633.055721	436.220747	131.344883	134.627659	0.807008	15.167250	3.138477
min	4.900000e+04	1988.000000	0.000000e+00	624.000000	3099.000000	1475.000000	1213.000000	2.000000	15.000000	1.000000
25%	5.000000e+05	2015.000000	2.801975e+04	1197.000000	3985.000000	1695.000000	1485.000000	5.000000	42.000000	4.000000
50%	8.424995e+05	2017.000000	4.879750e+04	1497.000000	4360.000000	1770.000000	1544.000000	5.000000	50.000000	6.000000
75%	1.908250e+06	2019.000000	7.100000e+04	1995.000000	4620.000000	1831.000000	1670.750000	5.000000	60.000000	8.000000
max	3.500000e+07	2022.000000	2.000000e+06	6592.000000	5569.000000	2220.000000	1995.000000	8.000000	105.000000	35.000000

Power_bhp	Engine_RPM	Torque (Nm)	RPM	Torque_Nm
1804.000000	1802.000000	1804.000000	1804.000000	1804.000000
129.094235	4823.091010	245.292221	2616.895787	245.292221
63.931779	1098.272917	138.449752	1206.765311	138.449752
7.000000	2910.000000	54.000000	150.000000	54.000000
83.000000	4000.000000	115.000000	1600.000000	115.000000
115.500000	4200.000000	200.000000	1800.000000	200.000000
170.000000	6000.000000	350.000000	4000.000000	350.000000
660.000000	8250.000000	780.000000	5600.000000	780.000000

From the describe method I got the above description of the columns but in the price column there is a max value which is way larger than the minimum value so this is a kind of outlier, So this might hamper tha analysis of the whole price column introducing skewness in the price column so these entries should be removed and it might be a sport car

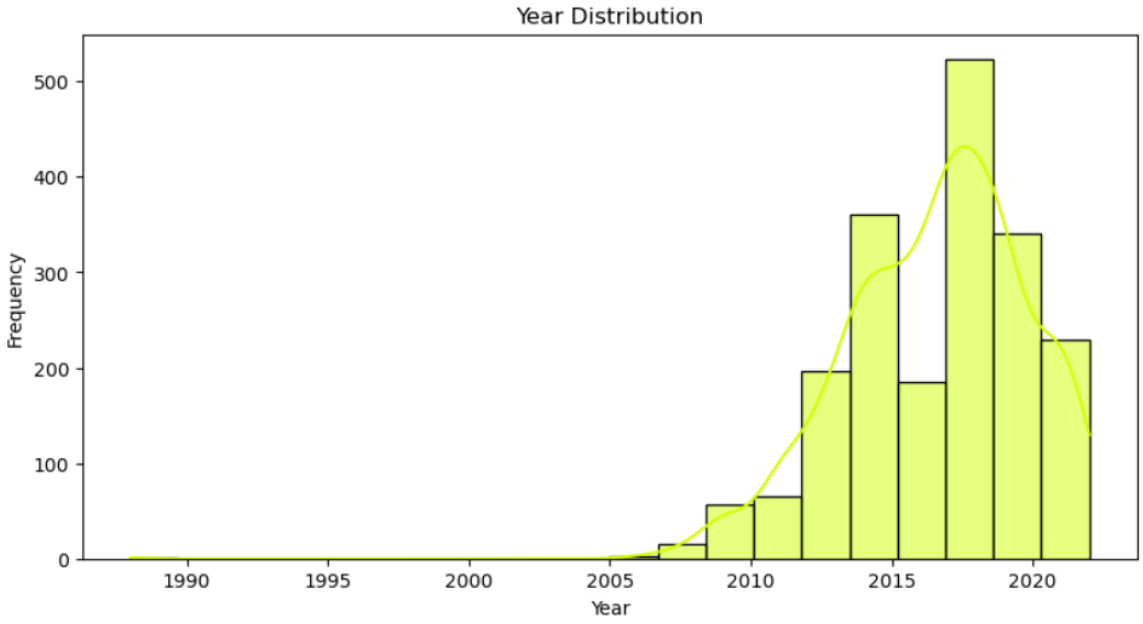
so all the engine secifications will be way higher than all the cars so removing this is the best choice.




```
Percentage of cars in each price range:
 (400000, 700000]      30.764636
(800000, 1100000]     13.201912
(100000, 300000]      10.454002
(300000, 400000]       9.737157
(1200000, 1500000]     7.108722
(700000, 800000]       6.750299
(1500000, 1800000]     6.212664
(1800000, 2100000]     4.241338
(2100000, 2400000]     3.345281
(2700000, 3000000]     3.106332
(2400000, 2700000]     2.628435
(1100000, 1200000]     2.449223
```

Now we can see all the price categories clearly

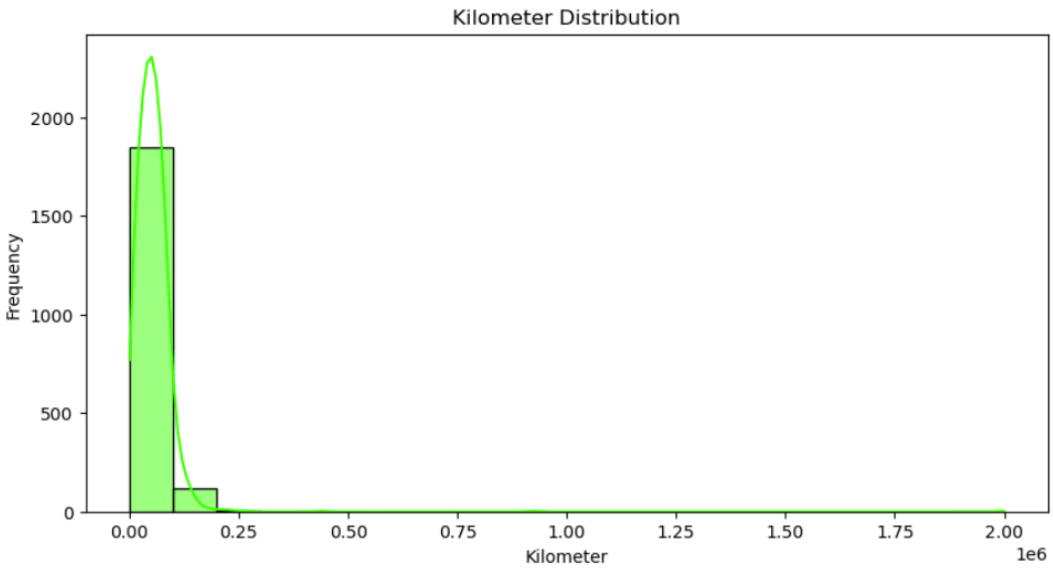
1.) We can that in 4 to 7 lakh Price range exactly 30% of cars exist and these are considered as average cars which comes average price segment in which most of the cars are sold. 2.) And in 8-11 lakh over 13% of the cars exist. 3.) In 1-3 lakh price range approximately 11% data exist and these are considered as cheapest cars.




```
Percentage of cars in each price range:
 (2016, 2018]    26.507856
(2014, 2016]    18.297010
(2018, 2020]    17.232641
(2012, 2014]    15.458692
(2020, 2022]    11.606690
(2010, 2012]     7.197162
(2006, 2010]     3.699949
(2022, 2023]     0.000000
```

Inference from this distribution of the year column

1.) We can see that more cars are sold in 2016-2018 year range approximately 25% of the data lies in this range 2.)In overall conclusion from this we can say that more cars are sold from 2012 and 2020 by slowly increasing thir sales and after 2020 the sales are slowly decreasing.

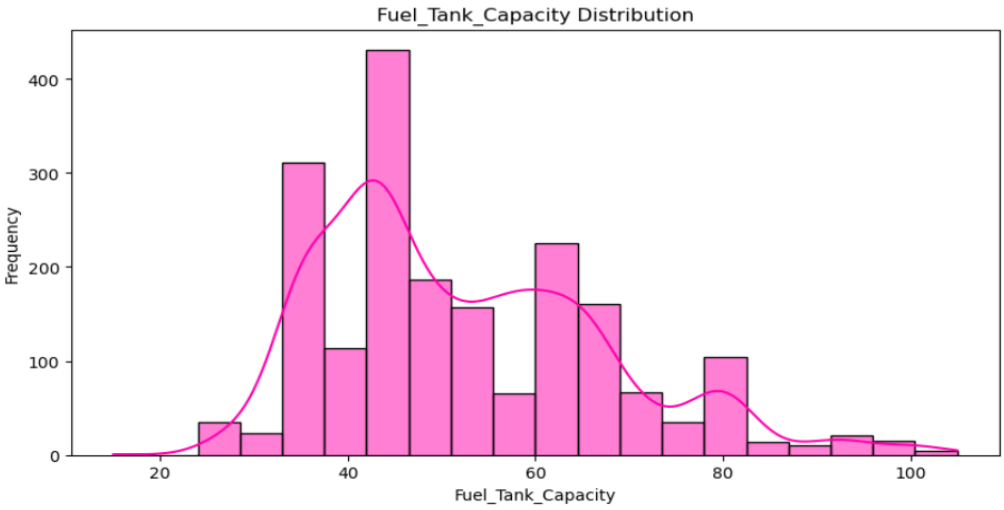


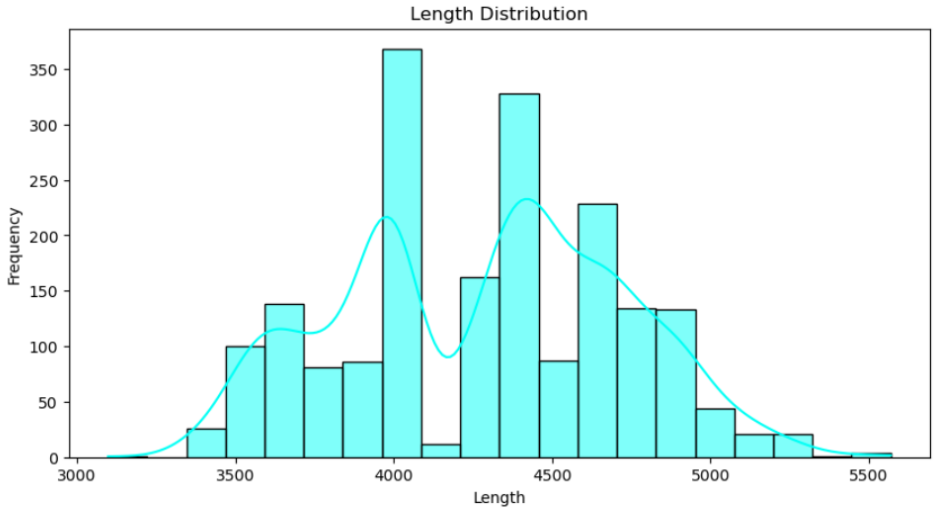

```
Percentage of cars in each price range:
 (50000, 90000]      39.622642
(30000, 50000]      24.732279
(10000, 30000]      19.887812
(0, 10000]           6.833248
(90000, 110000]      5.048445
(110000, 130000]     2.549720
(130000, 150000]     1.070882
(150000, 160000]     0.254972
```

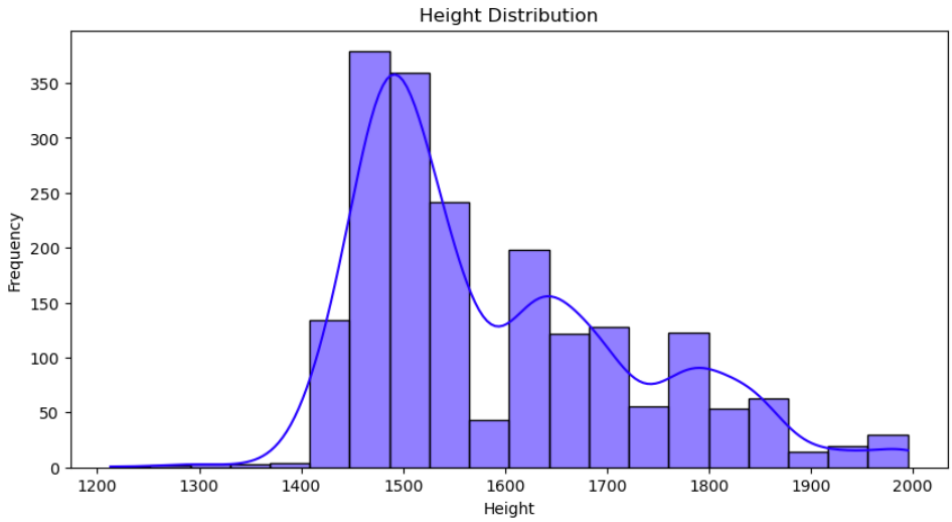
Inferences:

1.)Now we can see that many of that many cars have travelled 30k to 50k kilometers before they sold to others approximately 25% of the data in this segment.

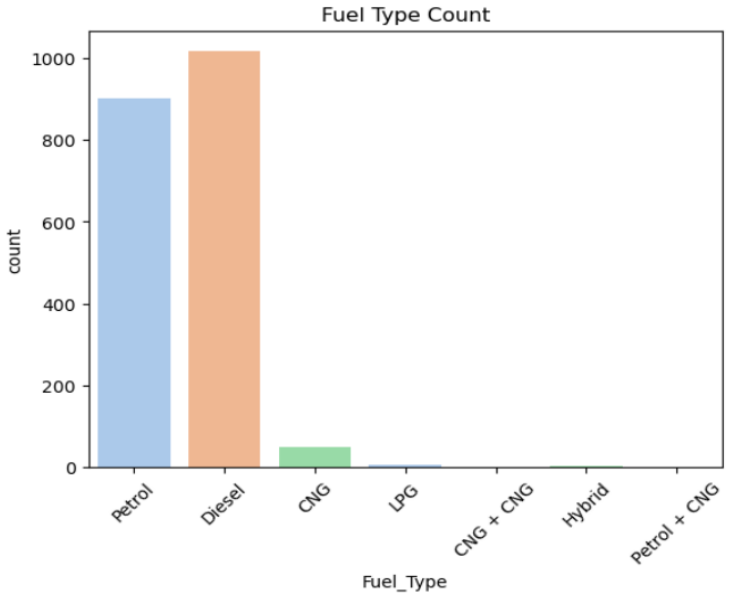
2.)also by seeing carefully we can see that from 50k to 90k almost 39% of the data exists.

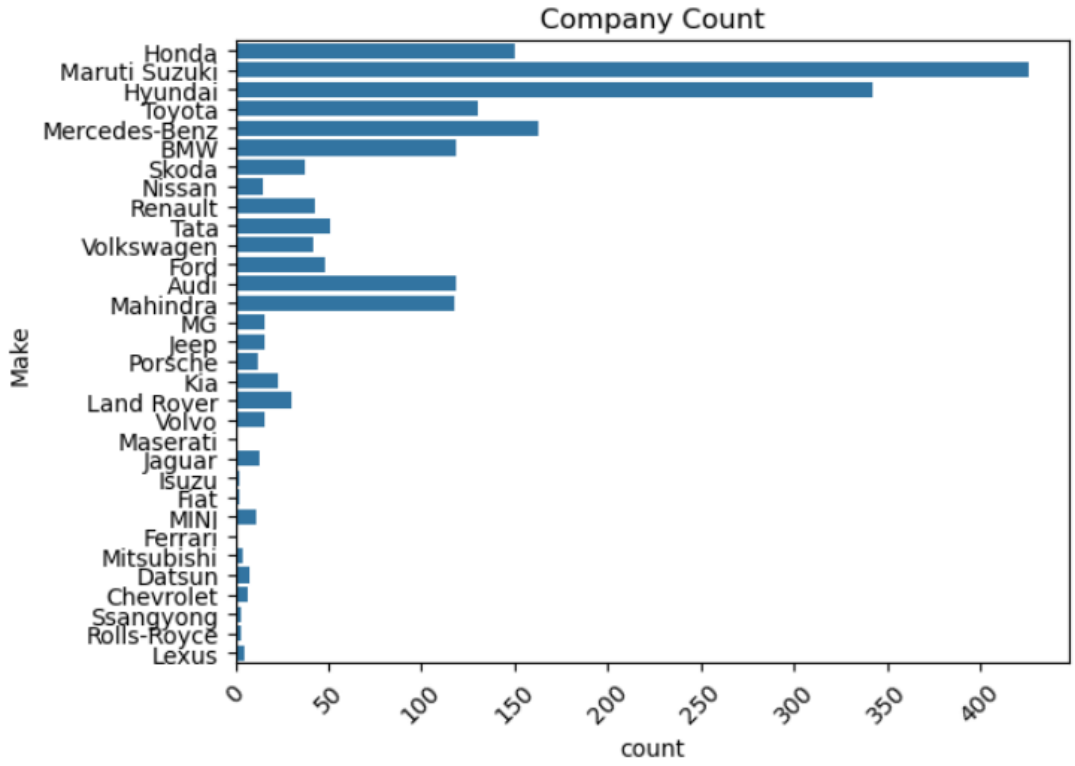




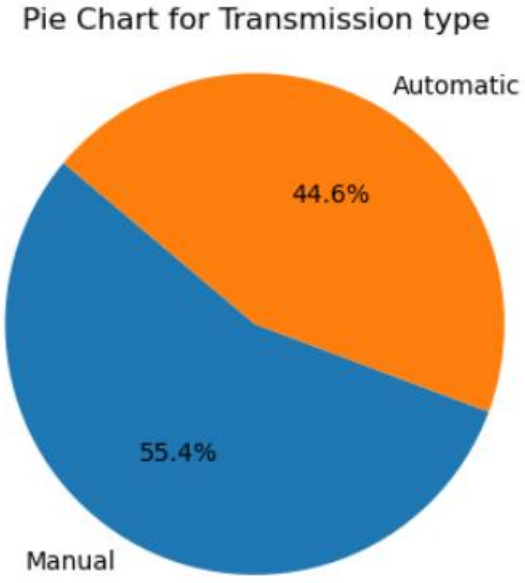


Categorical column analysis:

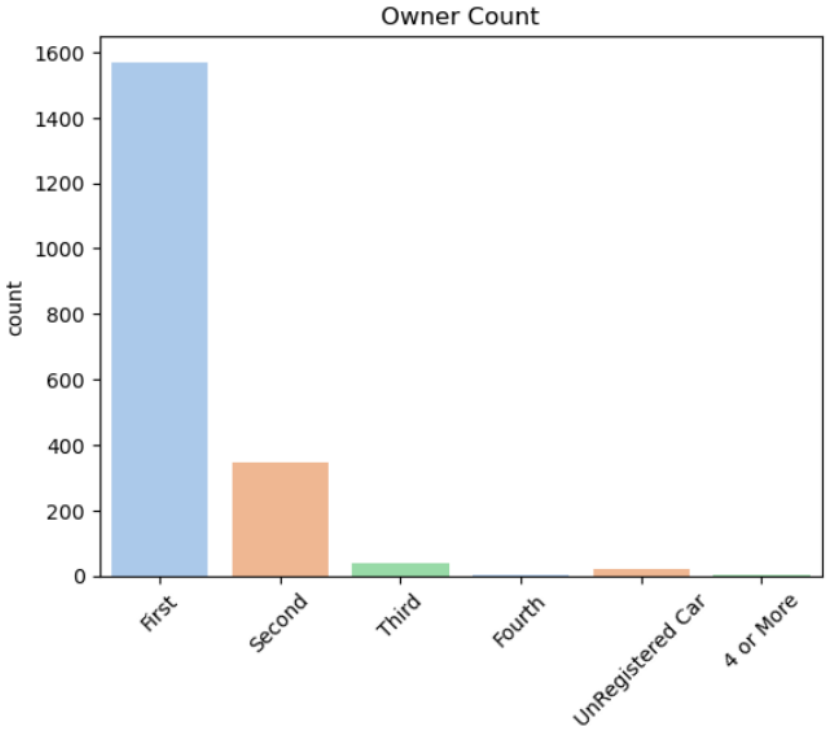




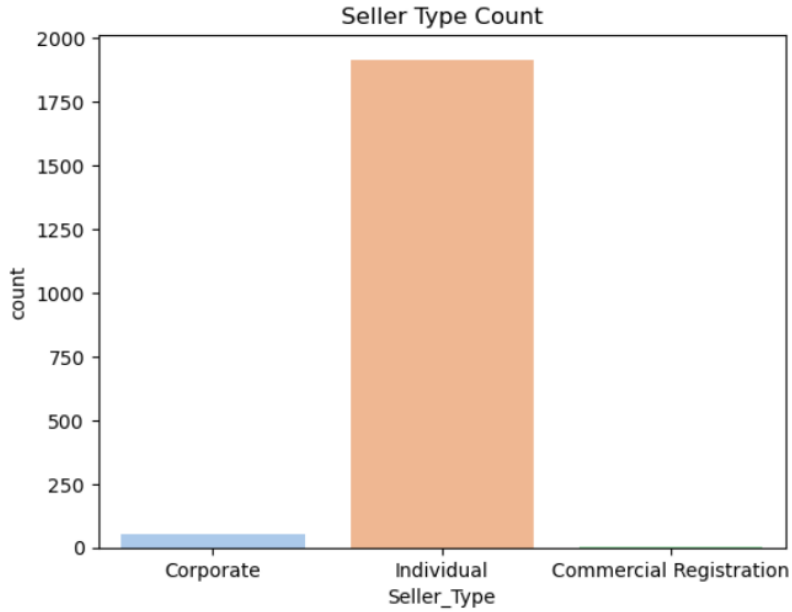
As expected the Maruti Suzuki cars are the highest sales in this data because this company targets budget cars for the people.



It turned out that there are more manual cars over 1100 are from manual transmission gear system as most of the budget cars come in this sector.

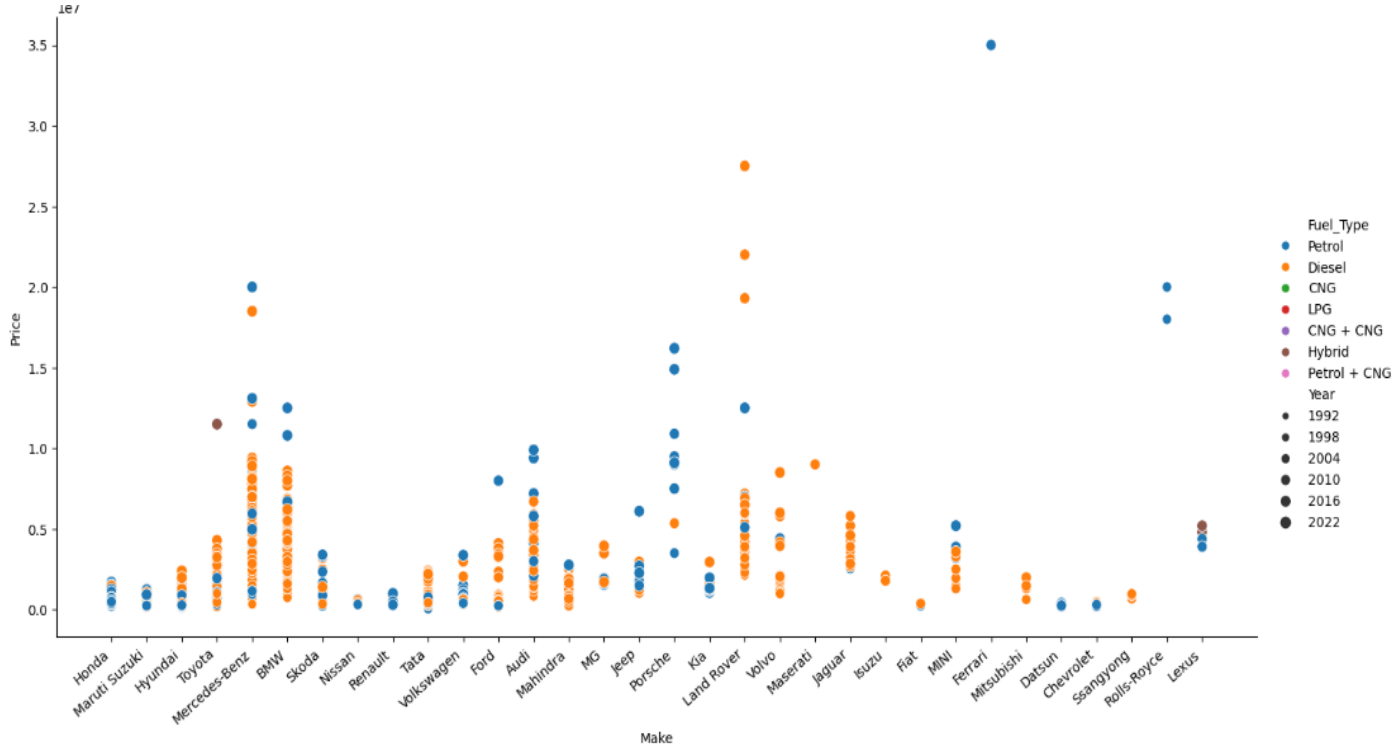


It tells that there are more people who are interested in First owner cars because they are buying the cars which are used and at any cost people try to buy the best one right.



Again the same story is going on as because car dekho company is buying more from the individual customers but it seems like it has less partnership with the commercial organizations.

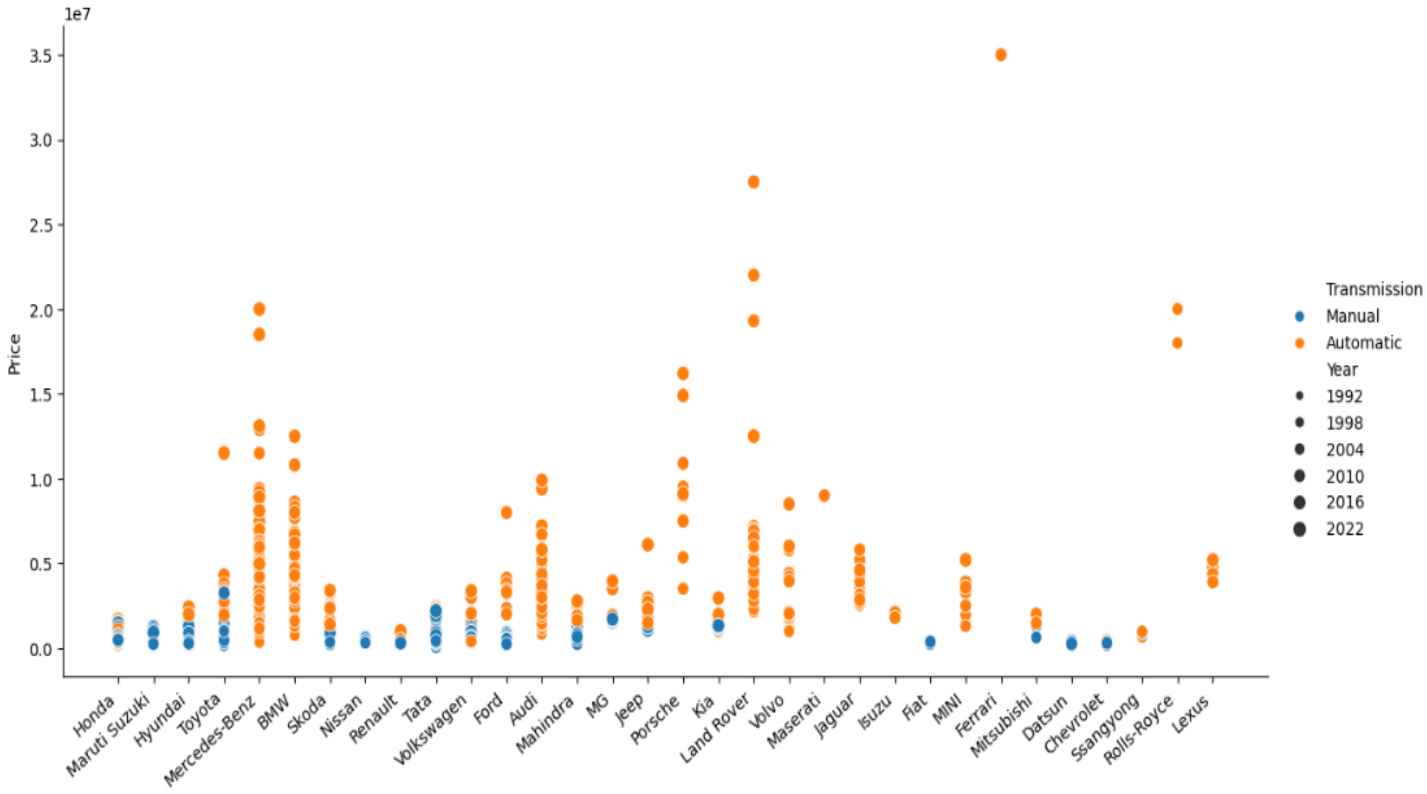
Bivariate Analysis:



Turns out that the Maruti suzuki resale value is low as always these company cars are budget cars.

Exceptionally there are some brands which are less popular in india but are doing good in the resale market like Land rover.

Also the low budget cars are almost petrol from Suzuki and the honda company.



Coming to this relplot It is telling that the companie's which are very popular for budget cars are using manual transmission but the surprising thing is All the Automatic gear cars are on high priced side.

Also they are latest too.

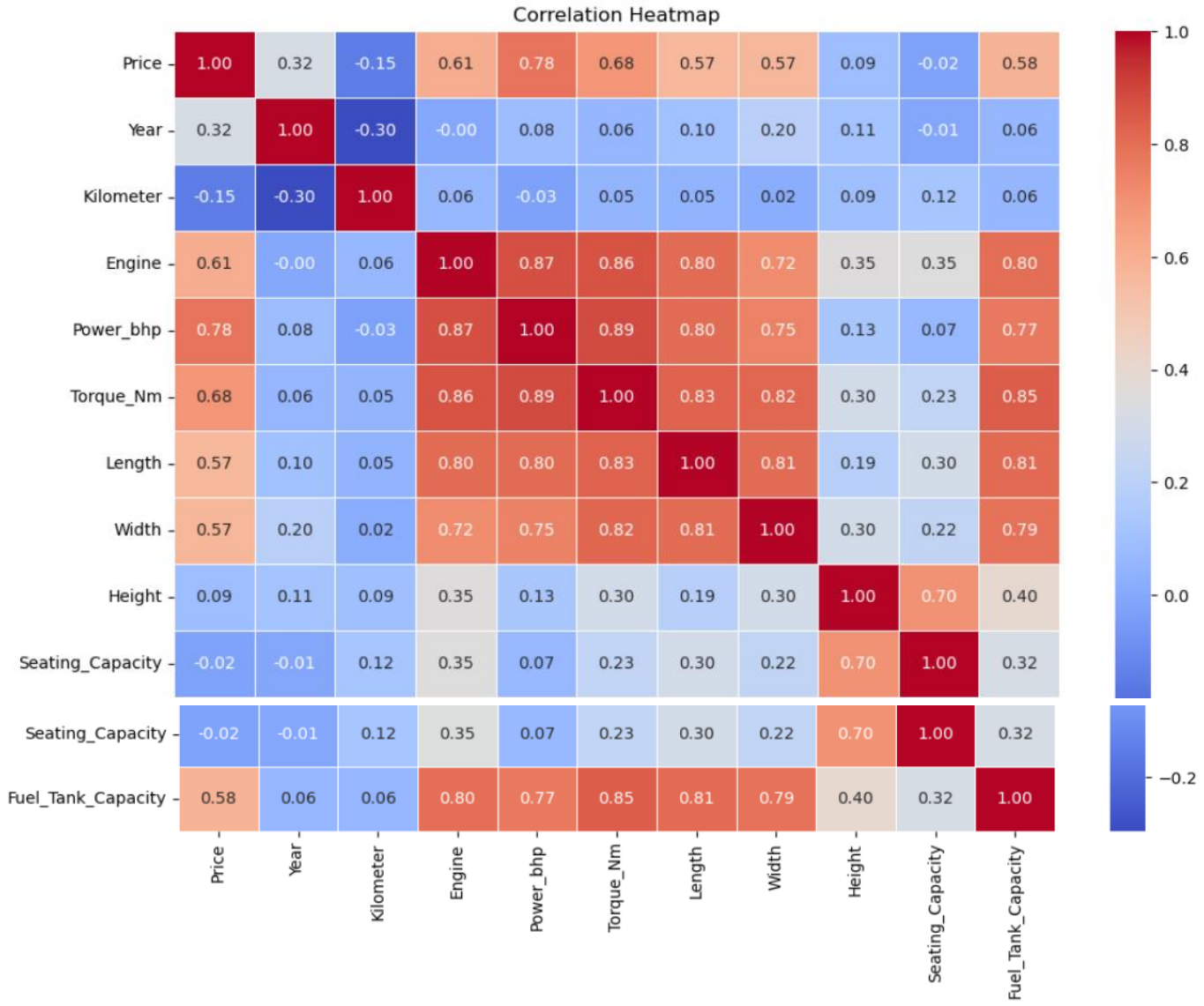
Every foreign company is using automatic Transmission system.

Multi-Variate Analysis:

I visualized a heatmap for getting all the information upon seeing the plot.

I plotted heatmap for correlation and co-variance

Although co-variance is not a standardised method for analysing the data but for getting an understanding about the data we should use it.



From this heat map we set our target to the price column because

Outcomes of CAR Sales DATA EDA:

To get more insights from the data and uncover the underlying patterns in the data like:

Finding relationships between the car sales area wise, the factors which effect the car sale price e.t.c

Quality of this Dataset:

1.)Model of the car should be 3 words so for the correct prediction it is important

2.)Engine, Max_Power, Max_Torque, Drivetrain, Length, Width, Height, Seating_Capacity, Fuel_Tank_Capacity these columns has null values in the dataset

3.)There are some cars which have null values in specified columns above so these can be removed.

4.) in engine column the entries are with cc but for analysis it is not required and calculations get wrong as it is string type.

So I the data quality in this dataset is not upto the mark for data analysis and EDA.

4.) Some Questions that I got and their answers.

These answers might help in getting the insights from the data.

1.What is the average selling price of the cars in the dataset?
cars.Price.mean()

1718279.036819637

2.What is the most common fuel type used in the cars?

Diesel=954

Percentage of most common fuel type is: 50.907150480256135 %

3.What is the average fuel tank capacity of the cars in the dataset?
Average fuel tank capacity in litres of the cars in the dataset is:
52.21734258271077

4.What is the average engine capacity of the cars in the dataset?
average engine capacity of the cars in the dataset is : 1682.82924226254

5.What is the average number of seats in the cars?
5.295090715048025

6.What is the most common seller type in the dataset?
array(['Corporate', 'Individual', 'Commercial Registration'], dtype=object)

7. What is the most common transmission type in the cars?

Manual 1037

Automatic 837

8.What is the average age of the cars in the dataset?
For this question there is no direct answer beacause there is no direct column representing age of the car in the

dataset so we need to create age column, but what is the year this dataset is updated? for this we need the maximum year in the year column so we will apply max function on that and make it as the current year and we subtract the year from the current year so that we can get the age

6.286552828175027

9. What is the most common brand of cars in the dataset
It is obvious from the figure that "Maruthi Suzuki" is the common among all the companies it is almost 20.23% among all the values

10. What is the average distance covered by the cars in the dataset?
The average kilometers driven for the cars in this dataset is "53178.13980789755"

11. What is the most common owner type in the cars?
First 1504

Second 322

Third 30

UnRegistered Car 18

Name: Owner, dtype: int64

from the figure we can say that first owner cars are more in the dataset and it is obvious only first

owner cars sell more than any other type of cars

12. What is the average maximum power of the cars in the dataset? So from

this data the average maximum power of cars is "129.09423503325942"

13. What is the average torque of the cars in the dataset?

69.0

14. What is the most expensive car in the dataset? 35000000

It is obvious from the dataset "ferrari 488 GTB" is the highest selling price in this dataset

15. What is the cheapest car in the dataset?

In the dataset the minimum selling price price of all cars is 49000 and it is Tata Nano Base

18. What is the most powerful car in the dataset?

the ferrari 488 GTB is the most powerful car in this dataset

19. What is the least powerful car in the dataset?

```
Make                          Toyota
Model                 Fortuner 3.0 MT
Price                        1000000
Year                            2010
Kilometer                     127000
Fuel_Type                     Diesel
Transmission                  Manual
Location                    Guwahati
Color                          White
Owner                          Third
Seller_Type                Individual
Engine                           2982
Max_Power                   171@3600
Max_Torque                  343@1400
Drivetrain                       AWD
Length                        4695.0
Width                         1840.0
Height                        1850.0
Seating_Capacity                 7.0
Fuel_Tank_Capacity              80.0
Current_year                    2023
Age                               13
Power(bhp)                       NaN
Engine(RPM)                      NaN
Torque (Nm)                      NaN
RPM                              NaN
Power_bhp                        NaN
Engine_RPM                       NaN
Name: 2036, dtype: object
```

20. What is the most common model of cars in the dataset? X1 sDrive20d xLine
    15

21. Which columns in the dataset directly impact on the resale of the cars?
    From the heat map we can see that power(bhp) is the column that has highest impact on selling price of the car

22. What is the most common fuel type used in the cars?
    from the statistics we can see that Diesel is the most common fuel type in the dataset that is almost 954 out of 1874 are of diesel and it is almost 50.9% of all the cars in the dataset.

Hypothesis Testing:

For this I have assumed one hypothesis and have done all the tests that are T-Test, F-Test, Z-Test, Chi- Squared test.

Hypothesis 1:

The average price of diesel cars is significantly different from the average price of petrol cars.

These are the results from those tests that are evident that my hypothesis is true.

T-test

```
T-Statistic: 6.843165463479428
P-Value: 1.0699683749158076e-11
The difference in average prices is statistically significant.
```

The results indicate that there is a statistically significant difference in the average prices between diesel and petrol cars. The low p-value (close to zero) suggests strong evidence against the null hypothesis¶

F-test:

```
F-Statistic: 47.372131553562454
P-Value: 8.052339105325683e-12
There is a significant difference in average prices between Diesel and Pet
rol cars.
```

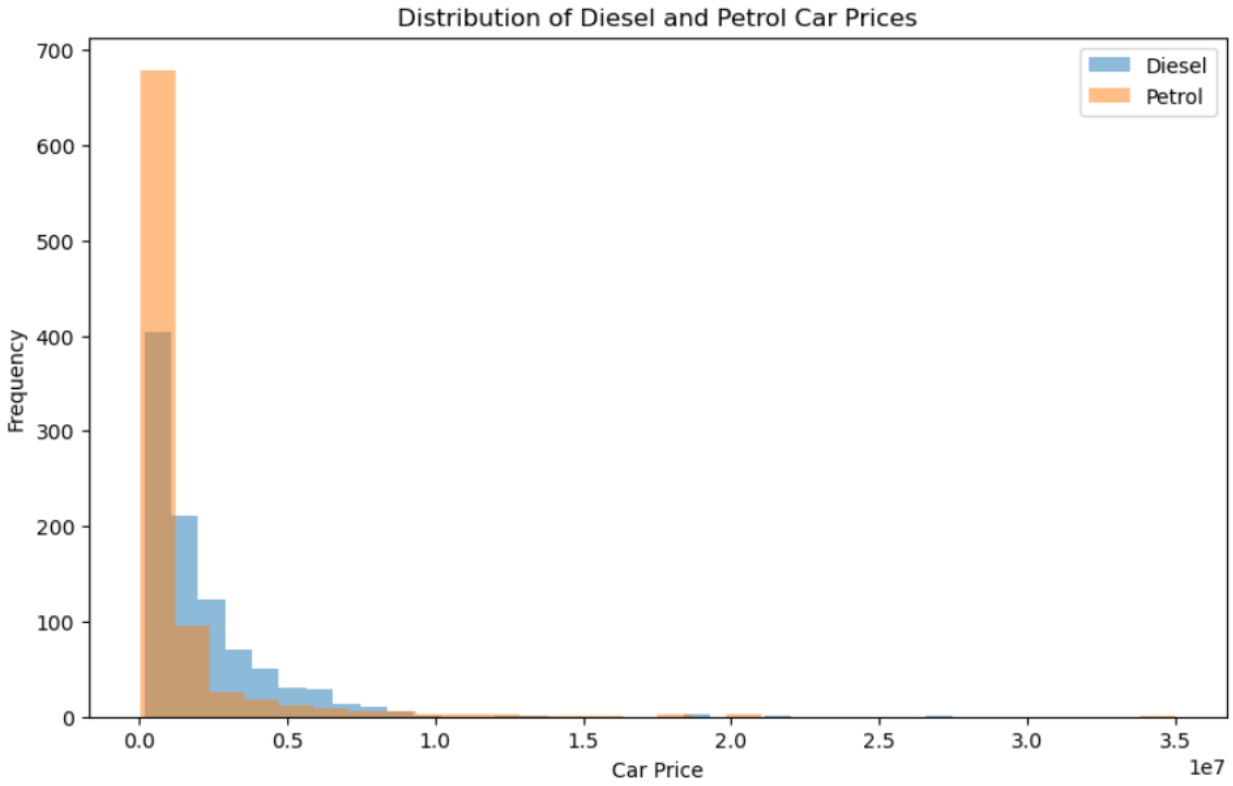
Z-test:

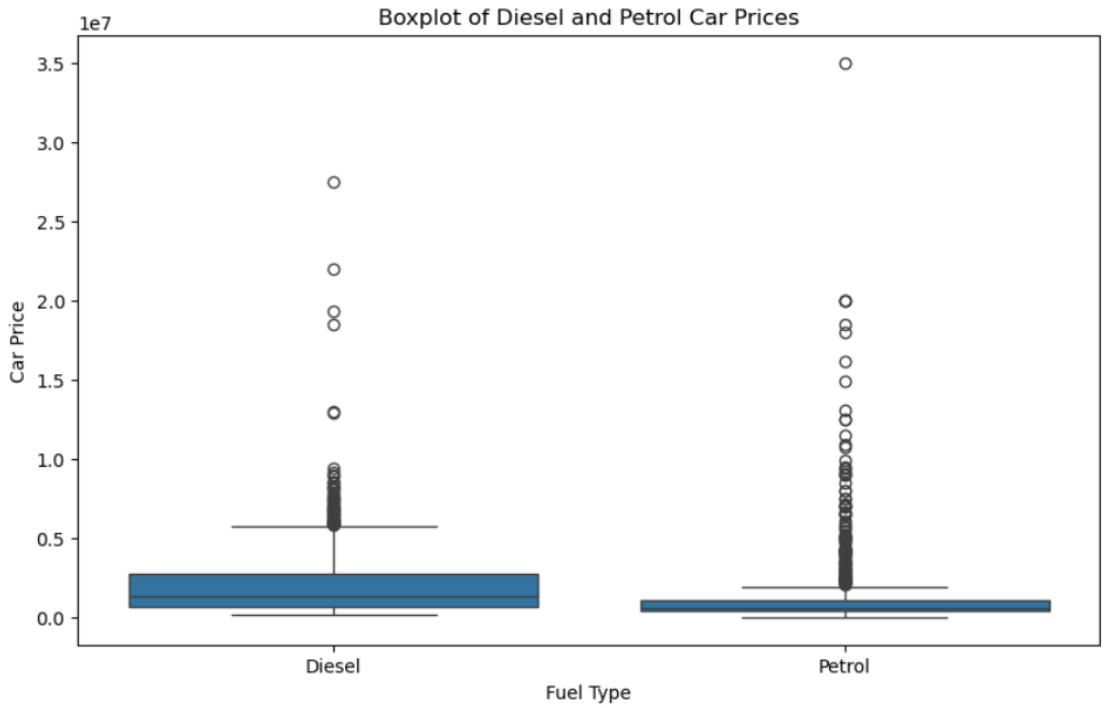
```
Z-Statistic: 6.882741572481309
P-Value: 5.871148223101824e-12
There is a significant difference in average prices between Diesel and Petrol
 cars.
```

The Z-test assumes a known population standard deviation, which is often not the case in practice. The t-test is more commonly used when the population standard deviation is unknown.

Chi- Squared Test:

```
Chi-Square Statistic: 68.9082187336131
P-Value: 6.845717174801284e-13
There is a significant association between Fuel Type and Transmission.
```





There are some visualizations for this hypothesis:

This indicates that there is difference in price of the cars in every price range in diesel and petrol cars.

This image suggests that there are more outliers in the petrol cars.

And more petrol cars are in the lower price segment

These all tests have their P-values less than 0.05 so these tests prove that there is significant difference in prices of the diesel and petrol cars.

Hypothesis 2:

There will be a significant difference in the price location wise for the same model of the car

These are the test results that when I haven't considered the same model

```
F-Statistic: 2.353395921334929
P-Value: 1.386742200085372e-09
```

```
There are significant differences in car prices across different locations
.
```

These are the test results when I considered the model of the car for testing

```
           sum_sq      df           F        PR(>F)
Location  5.044747e+13    75.0   2.868396  2.098638e-13
Model     1.013473e+16  1001.0  43.175668  0.000000e+00
Residual  2.129241e+14   908.0         NaN           NaN
```

Then I got mixed answers

→The F-statistic is 2.87, and the p-value is very close to zero (2.10e-13). This suggests that there are significant differences in car prices across different locations. The null hypothesis (no difference in prices between locations) is rejected.

→The F-statistic is 43.18, and the p-value is zero. This indicates that there are significant differences in car prices based on different car models. The null hypothesis (no difference in prices between models) is strongly rejected.

→The residual sum of squares represents the unexplained variance in car prices after considering the effects of both location and model. The degrees of freedom for residuals are 908.

→Both 'Location' and 'Model' have a significant impact on car prices, as indicated by their low p-values. The null hypotheses for both factors are rejected.The residual sum of squares represents unexplained variance in prices that is not accounted for by 'Location' and 'Model.'

Conclusion:

All conducted tests consistently suggest that there is a significant difference in average prices between Diesel and Petrol cars. Additionally, there is a significant association between Fuel Type and Transmission. These findings provide statistical evidence to support the initial hypothesis that there are significant variations in prices and associations with transmission types between Diesel and Petrol cars.

The t-test and ANOVA were both suitable and produced consistent results.

The Chi-Square test was appropriate for assessing the association between Fuel Type and Transmission.

Considering the consistency of results and the nature of your analysis, either the t-test or ANOVA is a suitable choice for comparing average prices between Diesel and Petrol cars.

Resources:

https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho

google.com

learn.upgrad.com