# AUTOMOBILE DATA ANALYSIS

Name
PGP-DSBA Online
January' 21
Date: dd/mm/yyyy

# Table of Contents

## Contents

# List of Figures



# List of Tables

## Executive Summary

A showroom owner deals with different types/models of cars. The dataset consists of various characteristics of the cars based on the models available in the showroom. Based on the different attributes/characteristics the price of the car is defined. In this problem statement we will explore the different attributes of the car and its contribution to the price of the car.

## Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of 127 different cars with 8 unique brands of car. Analyse the different attributes of the car make which can help in analysing the price of the car. This assignment should help the student in exploring the summary statistics, contingency tables, conditional probabilities & hypothesis testing.

## Data Description

1. make: brand of the car (for example: honda, mazda, mitsubishi, Nissan etc.)
2. fuel-type: type of the fuel used in the car (diesel, gas)
3. aspiration: std, turbo.
4. num-of-doors: four, two.
5. body-style: hardtop, wagon, sedan, hatchback, convertible.
6. drive-wheels: 4wd, fwd, rwd.
7. engine-location: front, rear.
8. num-of-cylinders: eight, five, four, six, three, twelve, two.
9. wheel-base: continuous from 86.6 120.9.
10. length: continuous from 141.1 to 208.1.
11. width: continuous from 60.3 to 72.3.
12. height: continuous from 47.8 to 59.8.
13. curb-weight: continuous from 1488 to 4066.
14. engine-size: continuous from 61 to 326.
15. horsepower: continuous from 48 to 288.
16. city-mpg: continuous from 13 to 49.
17. highway-mpg: continuous from 16 to 54.
18. price: continuous from 5118 to 45400.

### Sample of the dataset:

| | make | fuel-type | aspiration | num-of-doors | body-style | drive-wheels | engine-location | num-of-cylinders | wheel-base | length | width | height | curb-weight | engine-size | horsepower | city-mpg | highway-mpg | price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | honda | gas | std | two | hatchback | fwd | front | four | 86.6 | 144.6 | 63.9 | 50.8 | 1713 | 92 | 58 | 49 | 54 | 6479 |
| 1 | honda | gas | std | two | hatchback | fwd | front | four | 86.6 | 144.6 | 63.9 | 50.8 | 1819 | 92 | 76 | 31 | 38 | 6855 |
| 2 | honda | gas | std | two | hatchback | fwd | front | four | 93.7 | 150.0 | 64.0 | 52.6 | 1837 | 79 | 60 | 38 | 42 | 5399 |
| 3 | honda | gas | std | two | hatchback | fwd | front | four | 93.7 | 150.0 | 64.0 | 52.6 | 1940 | 92 | 76 | 30 | 34 | 6529 |
| 4 | honda | gas | std | two | hatchback | fwd | front | four | 93.7 | 150.0 | 64.0 | 52.6 | 1956 | 92 | 76 | 30 | 34 | 7129 |

Table 1. Dataset Sample

Dataset has 18 variables with 8 different types of the car make. Each car make has different sets of attributes. Based on the characteristic price of the car is defined.

# Exploratory Data Analysis

## Let us check the types of variables in the data frame.

```
make               object
fuel-type          object
aspiration         object
num-of-doors       object
body-style         object
drive-wheels       object
engine-location    object
num-of-cylinders   object
wheel-base         float64
length             float64
width              float64
height             float64
curb-weight        int64
engine-size        int64
horsepower         int64
city-mpg           int64
highway-mpg        int64
price              int64
dtype: object
```

There are total 127 rows and 18 columns in the dataset. Out of 18, 8 columns are of object type and rest 10 are of either integer or float data type.

## Check for missing values in the dataset:

```
RangeIndex: 127 entries, 0 to 126
Data columns (total 18 columns):
make               127 non-null object
fuel-type          127 non-null object
aspiration         127 non-null object
num-of-doors       127 non-null object
body-style         127 non-null object
drive-wheels       127 non-null object
engine-location    127 non-null object
num-of-cylinders   127 non-null object
wheel-base         127 non-null float64
length             127 non-null float64
width              127 non-null float64
height             127 non-null float64
curb-weight        127 non-null int64
engine-size        127 non-null int64
horsepower         127 non-null int64
city-mpg           127 non-null int64
highway-mpg        127 non-null int64
price              127 non-null int64
```

From the above results we can see that there is no missing value present in the dataset.

## Correlation Plot

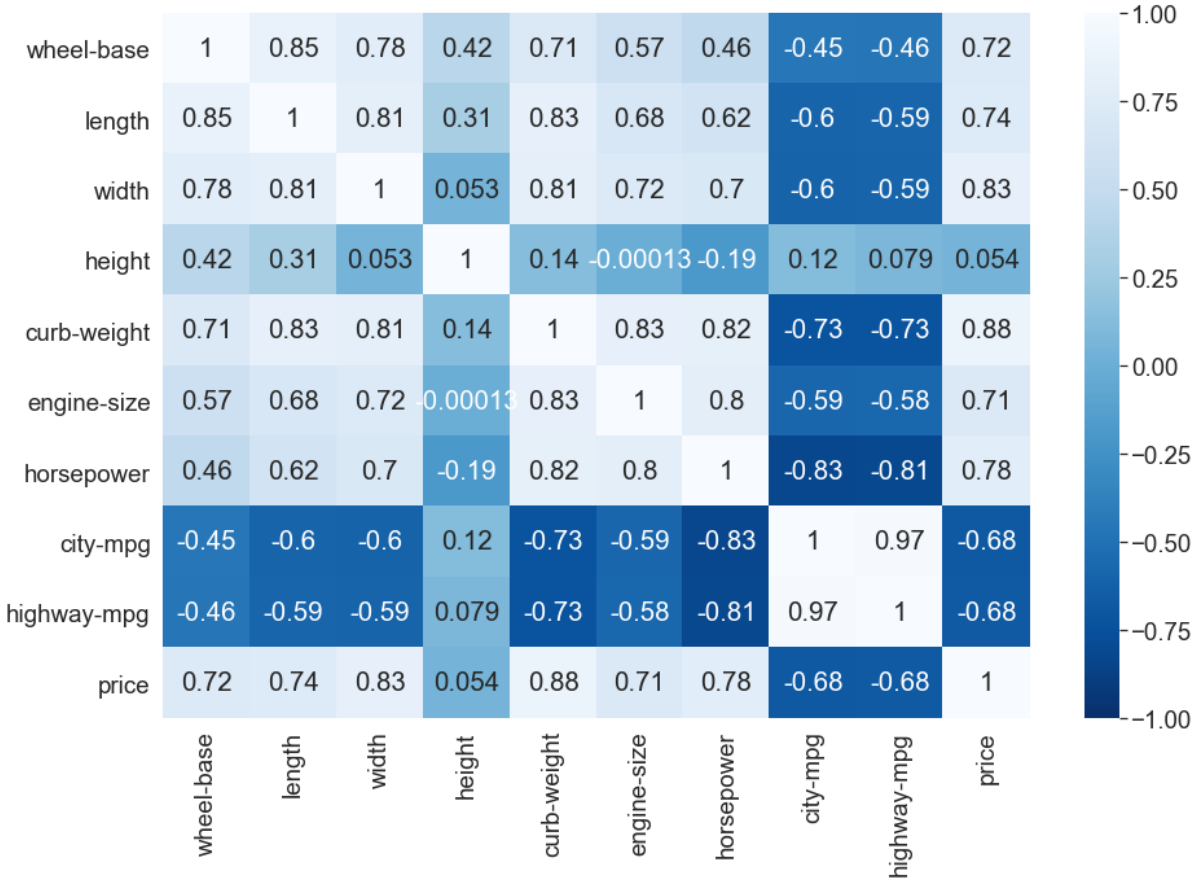


Fig.1 – Correlation Heatmap

From the correlation plot, we can see that various attributes of the car are highly correlated to each other. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

## Pairplot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

From the graph, we can see that there is positive linear relationship between variables like highway-mpg and city-mpg. From the histogram we can see that the price of the whole dataset is right skewed.

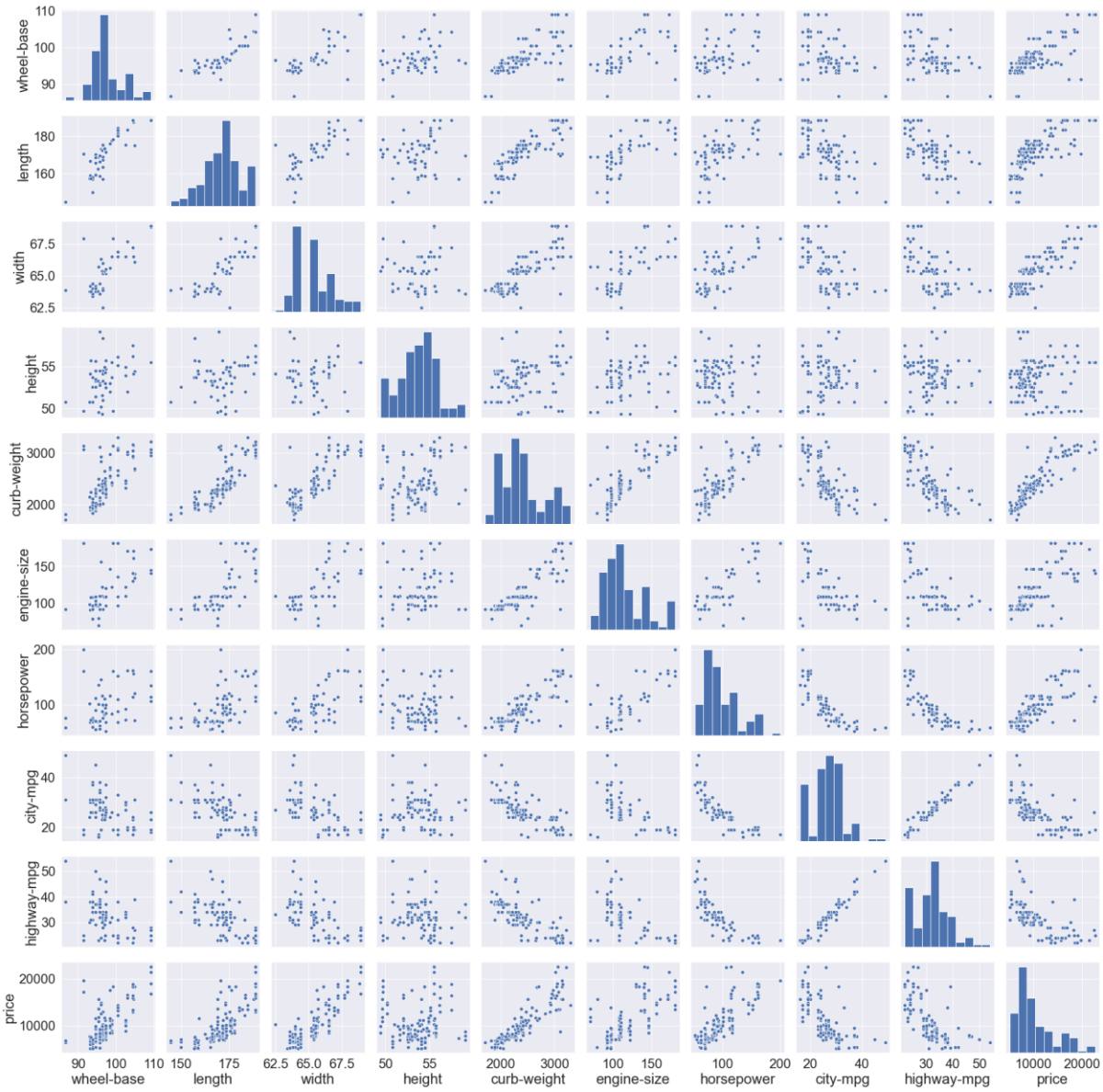


Fig.2 – Pairplot

## Q1: Use methods of descriptive statistics to summarize data. Find out the costliest car and the cheapest car by average price.

Descriptive statistics help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of centre: the mean, median, and mode, which are used at almost all levels of math and statistics.

| | count | unique | top | freq | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **make** | 127 | 8 | toyota | 32 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **fuel-type** | 127 | 2 | gas | 117 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **aspiration** | 127 | 2 | std | 110 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **num-of-doors** | 127 | 2 | four | 71 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **body-style** | 127 | 5 | sedan | 56 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **drive-wheels** | 127 | 3 | fwd | 86 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **engine-location** | 127 | 1 | front | 127 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **num-of-cylinders** | 127 | 4 | four | 110 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **wheel-base** | 127 | NaN | NaN | NaN | 97.4441 | 4.17486 | 86.6 | 94.5 | 96.5 | 98.8 | 109.1 |
| **length** | 127 | NaN | NaN | NaN | 171.287 | 9.65843 | 144.6 | 166.3 | 171.7 | 176.2 | 188.8 |
| **width** | 127 | NaN | NaN | NaN | 65.3756 | 1.40202 | 62.5 | 64 | 65.4 | 66.5 | 68.9 |
| **height** | 127 | NaN | NaN | NaN | 53.6457 | 2.13804 | 49.4 | 52.5 | 54.1 | 55.1 | 59.1 |
| **curb-weight** | 127 | NaN | NaN | NaN | 2405.47 | 388.094 | 1713 | 2114.5 | 2326 | 2614.5 | 3296 |
| **engine-size** | 127 | NaN | NaN | NaN | 115.74 | 26.4249 | 70 | 97 | 109 | 130 | 181 |
| **horsepower** | 127 | NaN | NaN | NaN | 94.622 | 30.8177 | 52 | 69 | 86 | 114 | 200 |
| **city-mpg** | 127 | NaN | NaN | NaN | 26.5984 | 5.84626 | 16 | 23.5 | 26 | 31 | 49 |
| **highway-mpg** | 127 | NaN | NaN | NaN | 32.1654 | 6.18724 | 22 | 28 | 32 | 37 | 54 |
| **price** | 127 | NaN | NaN | NaN | 10415.5 | 4040.54 | 5118 | 7481 | 9233 | 12459.5 | 22625 |

Table 2. Summary of the data

From the descriptive statistics, we can see that there are 8 unique types of make (car) available in the dataset. Toyota is the most frequent car type. Average length of the car is 171.287 with the average width of 65.37. The average price of the overall dataset is 10415 dollars. The range of the price of the car varies from 5118 to 22625. That means the price is spread over a big range.

NaN shows that the values cannot be calculated for that particular variables. Like we can calculate mean for a categorical/object type variable. And in a same way unique value for a numerical variable.

**Calculating the average price of the car based on the make.**

| | make | avg_price |
|---|---|---|
| 0 | honda | 8184.692308 |
| 1 | subaru | 8541.250000 |
| 2 | mitsubishi | 9239.769231 |
| 3 | toyota | 9885.812500 |
| 4 | volkswagen | 10077.500000 |
| 5 | nissan | 10415.666667 |
| 6 | mazda | 10644.000000 |
| 7 | volvo | 18063.181818 |

Table 3. Avg. Price by make of the car

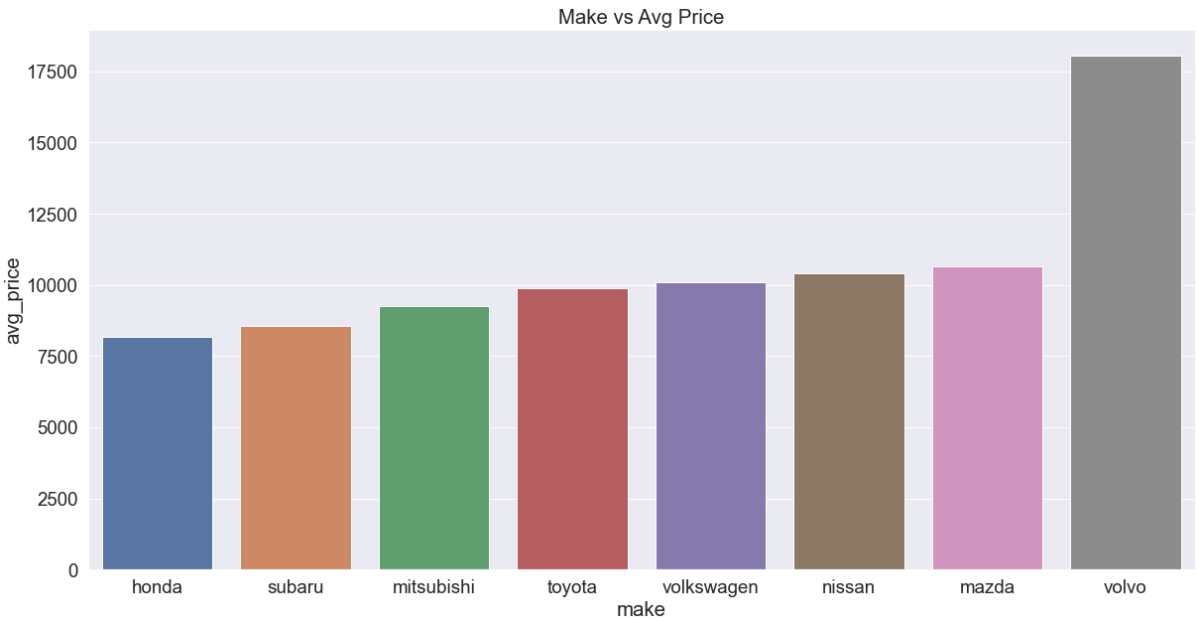


Fig.3 – Make vs Avg. Price bar plot

After calculating the average price of the car based on different make, we found that **Volvo** is the costliest car type and **honda** is the cheapest car.

## Q2: Which fuel type car has the highest average price?

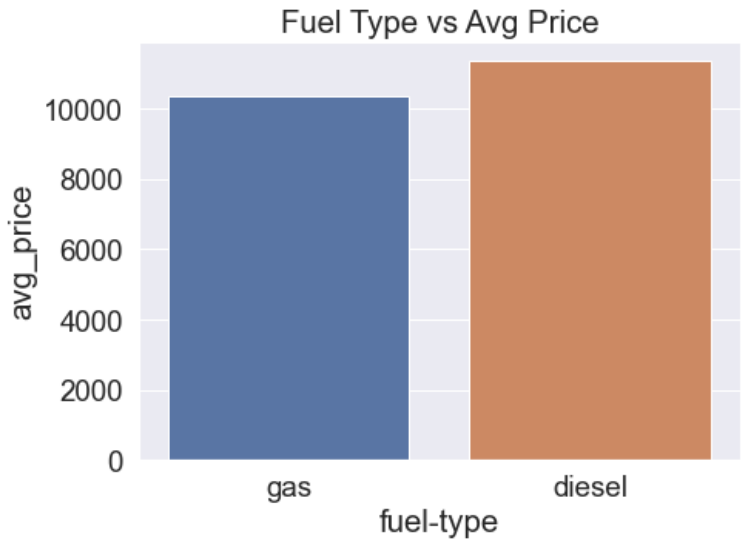


Fig.4 – Fuel Type Vs Avg. Price bar plot

We will calculate the average price of the car based on the fuel-type and plot the graph. From the above plot, we can see that car with **diesel fuel-type** has the highest average price.

## Q3: For this data, construct the following contingency table (Keep make as row variable)

**Contingency Table**: A table showing the distribution of one variable in rows and another in columns, used to study the correlation between the two variables.

### a. Make and body-style

| body-style | convertible | hardtop | hatchback | sedan | wagon |
|---|---|---|---|---|---|
| make | | | | | |
| honda | 0 | 0 | 7 | 5 | 1 |
| mazda | 0 | 0 | 10 | 6 | 0 |
| mitsubishi | 0 | 0 | 9 | 4 | 0 |
| nissan | 0 | 1 | 5 | 9 | 3 |
| subaru | 0 | 0 | 3 | 5 | 4 |
| toyota | 1 | 3 | 14 | 10 | 4 |
| volkswagen | 1 | 0 | 1 | 9 | 1 |
| volvo | 0 | 0 | 0 | 8 | 3 |

Table 4. Contingency Table: Make vs Body-style

## Q4: Assume that the sample is a representative of the population of the cars available at showroom. Based on the data, answer the following questions:

### a. What is the probability that a randomly selected car will be Honda?

Honda car Probability = (Total number of Honda cars)/ (Total number of cars at showroom).

Prob_honda = 13/127 = 0.102362

### b. What is the probability that a randomly selected car will be toyota?

Toyota car Probability = (Total number of Toyota cars)/ (Total number of cars at showroom).

Prob_toyota = 32/127 = 0.251969

## Q5: Form the null and alternate hypothesis to test whether the Price of gas cars is significantly different from that of diesel cars

A **null hypothesis** is a hypothesis that says there is no statistical significance between the two variables in the hypothesis. It is the hypothesis that the researcher is trying to disprove.

An **alternative hypothesis** simply is the inverse, or opposite, of the null hypothesis.

In testing the company would like to show that the average price of gas cars is significantly different from that of diesel cars

**Null hypothesis:** states that the difference in average Price of gas cars & diesel cars is 0

**Alternative hypothesis:** states that the difference in average Price of gas cars & diesel cars is not equal 0 at 95% confidence.

$H_0$: $\mu$ (gas) - $\mu$ (diesel) = 0
$H_A$: $\mu$ (gas) - $\mu$ (diesel)! = 0

## Q6: Conduct the test of hypothesis and find the p-value. Interpret the p-value. Is there evidence at the 0.05 level of significance that average price of gas cars is significantly different from that of diesel cars?

The difference in means between two Normal distributions with unknown variance follows a Student's t-distribution. The t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. The Student t-test is one of the oldest and widely used hypothesis test.

We don't know the variance of the population. A common strategy to assess hypothesis is to conduct a t test. A t test can tell whether two groups have the same mean. A t test can be estimated for:

1) One sample t test

2) Two sample t test (including paired t test)

We assume that the samples are randomly selected, independent and come from a normally distributed population with unknown but equal variances.

### T-Test

Now we will perform the t-test on the log-transformed data to prove the hypothesis.

First decide the level of significance:

The level of significance is defined as the probability of rejecting a null hypothesis by the test when it is really true, which is denoted as α. That is, P (Type I error) = α. Confidence level: The level of significance 0.05 is related to the 95% confidence level.

Level of Significance α = 0.05

From the two sample t- test performed, we got the below results:

Two sample t test

**t statistic: 0.58 p value: 0.574**

At the level of 5% significance, p-value = 0.57. Since p-value > 0.05, that means the difference in average Price of gas cars & diesel cars is 0 at the 5% level of significance

We have no evidence to reject the null hypothesis since p value > Level of significance

## Conclusion & Recommendation

From the data exploration, we can suggest/recommend below suggestions to the car makers:

1. There is skewness in car type availability the terms of fuel type. Many of the brands have only one type of car. The brand companies could try different cars of different types.
2. There is not much difference in the average price of diesel car and gas fuel-type car which again can suggest the makers to explore the car modification with cost effective price based on the fuel-type.
3. Apart from these attributes, there are latest features which could help in attracting customers. So, the companies/manufactures could explore those features as well.

## THE END!