Analysis of Automobile Data

Purpose of Analysis:

The purpose of analysing an automobile dataset is to extract useful information and insights about various aspects of cars, such as their prices, features, and performance. This analysis can help:

- **Understand Trends**: Identify patterns and trends in the automobile market.

- **Make Comparisons**: Compare different car models based on various attributes like price, fuel efficiency, and engine size.

- **Inform Decisions**: Help consumers, manufacturers, and dealers make informed decisions about buying, selling, or manufacturing cars.

- **Identify Relationships**: Discover relationships between different variables, such as how engine size impacts fuel efficiency or how car price varies with features.

- **Predict Future Values**: Use the data to predict future trends.

About the Automobile Dataset:

Context

This dataset consists of data from 1985 Ward's Automotive Yearbook. Here are the sources

Sources:

1)1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.

2) Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038.

3) Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037.

Shape of the dataset:

The dataset consists of 205 rows and 26 columns.

Let us dive deep into the column names and what does it represents:

Symbolling: Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale, this process "symbolling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

Normalised-losses: A numeric value representing the relative average loss payment per insured vehicle year. Higher values indicate higher risk and potential loss for the insurance company.

Make: The manufacturer or brand of the car (e.g., Ford, Chevy, BMW).

Fuel-type: The type of fuel the car uses (e.g., gas, diesel).

Aspiration: aspiration values are ‘std’ for standard/naturally aspirated engines and ‘turbo’ for turbocharged engines.

Num-of-doors: Number of doors present in a car.

Body-style: The style or category of the car body (e.g., sedan, hatchback, convertible etc).

Drive-wheels: This column refers to the type of drivetrain system the vehicle uses to transmit power from the engine to the wheels. The drivetrain system determines which wheels receive power from the engine, affecting the vehicle's handling, performance, and traction.(e.g., FWD , RWD,4WD,AWD).

Engine-location: Location of Engine in the automobile i.e, either in front or rear end of the vehicle.

Wheel-base: the term "wheelbase" refers to the distance between the centre of the front and rear wheels of a vehicle.

Length: length of the vehicle

Width: width of the vehicle

Height: height of the vehicle

Curb-weight: "curb weight" refers to the weight of a vehicle without any passengers or cargo but with all standard equipment and necessary operating consumables such as fuel, oil, and coolant.

Engine-Type: It refers to the classification of the engine based on its design, construction, or operating principles.

Num-of-cylinders: refers to the number of cylinders in the engine of the car. Common configurations include 4-cylinder, 6-cylinder, and 8-cylinder engines.

Engine-size: This measurement is usually expressed in litres(L) or cubic centimetre (cc) and indicates the total volume of all the cylinders in an engine. It essentially measures the space available for the fuel-air mixture to burn and generate power.

Fuel-System: The "fuel system" refers to the system that delivers fuel to the engine.

Bore: It refers to the diameter of the cylinders in the engine of a car.

Stroke: It refers to the engine's stroke, which is a measurement related to the engine's internal geometry.

Compression-ratio: It is a key specification of an engine that impacts its efficiency and performance.

Horsepower: It typically refers to the measure of the engine's power output. It is one of the key variables used to understand and compare the performance capabilities of different vehicles.

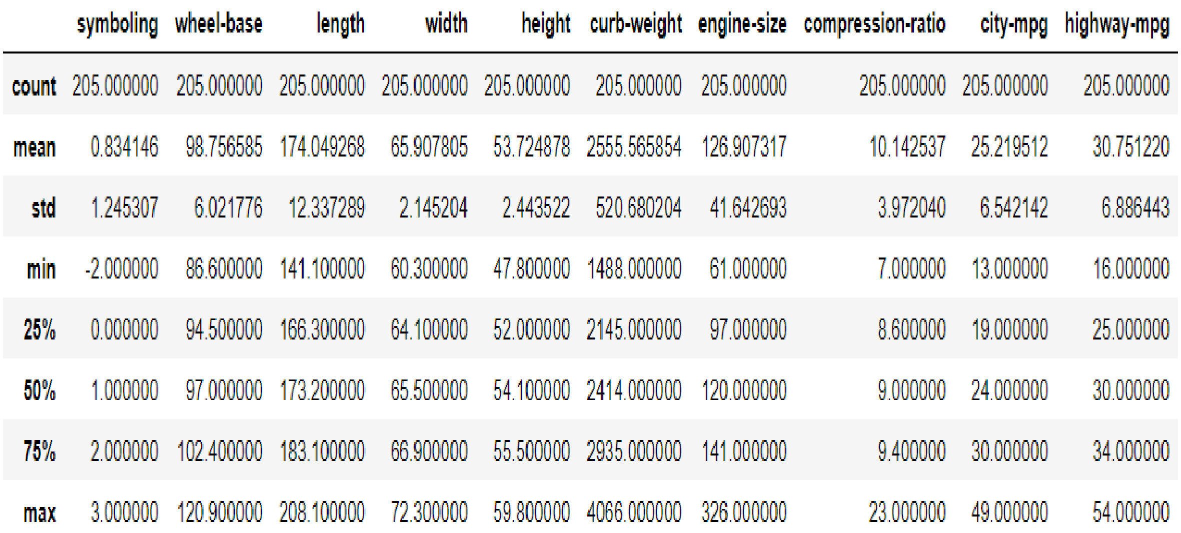
Peak-rpm: refers to the engine speed at which the engine achieves its maximum rotational speed, often measured in revolutions per minute (RPM).

City-mpg: Higher city MPG values indicate that a vehicle can travel more miles on a gallon of fuel in city conditions, which can result in lower fuel costs and reduced environmental impact.

Highway-mpg: Highway MPG is crucial for consumers because it provides an estimate of how fuel-efficient a vehicle is during long-distance or highway travel.

price: The price of the car.

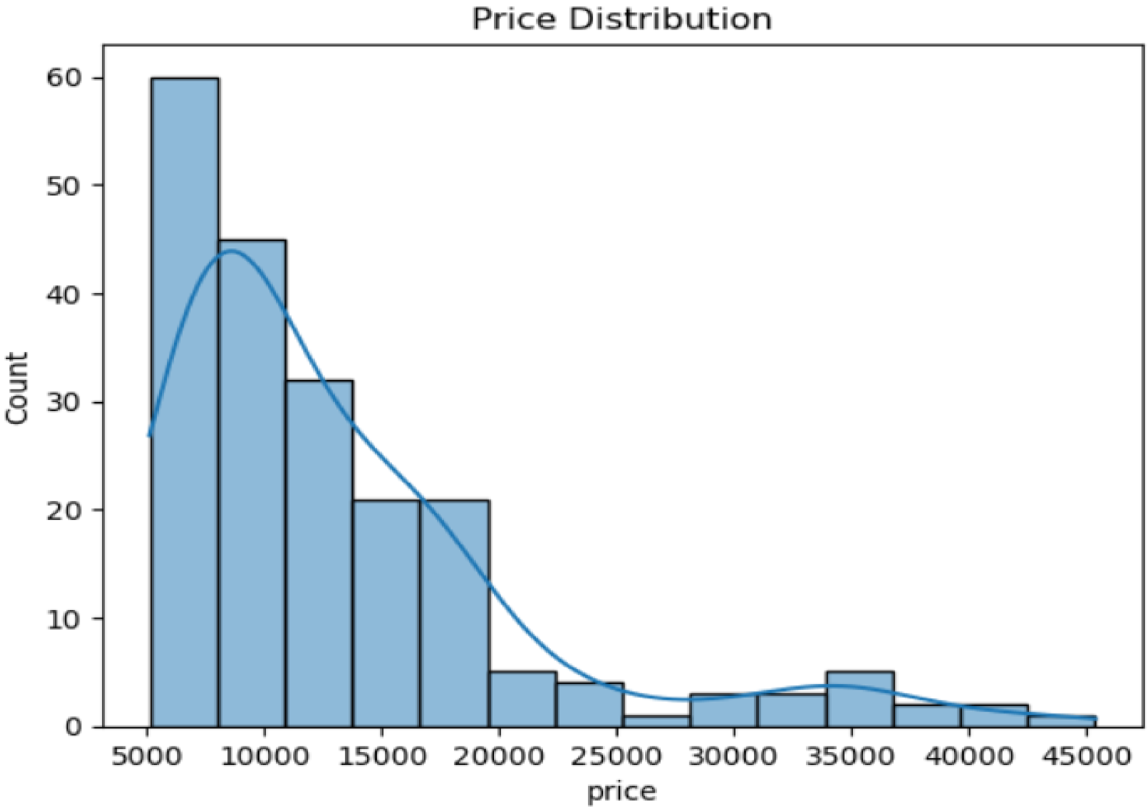
The following table represents the mean, standard deviation, minimum value, 25% or First Quartile ,50% or median or Second Quartile, 75% or Third Quartile and maximum value of each numerical columns or variables in the automobile dataset.



	symboling	wheel-base	length	width	height	curb-weight	engine-size	compression-ratio	city-mpg	highway-mpg
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000
mean	0.834146	98.756585	174.049268	65.907805	53.724878	2555.565854	126.907317	10.142537	25.219512	30.751220
std	1.245307	6.021776	12.337289	2.145204	2.443522	520.680204	41.642693	3.972040	6.542142	6.886443
min	-2.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	7.000000	13.000000	16.000000
25%	0.000000	94.500000	166.300000	64.100000	52.000000	2145.000000	97.000000	8.600000	19.000000	25.000000
50%	1.000000	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000	9.000000	24.000000	30.000000
75%	2.000000	102.400000	183.100000	66.900000	55.500000	2935.000000	141.000000	9.400000	30.000000	34.000000
max	3.000000	120.900000	208.100000	72.300000	59.800000	4066.000000	326.000000	23.000000	49.000000	54.000000

Understanding the Variables through Visualization

Distribution of Price:



Insights

- The distribution of the data in the variable price is skewed to the right or positively skewed
- The maximum value is almost 45000
- The minimum value starts from 5000 approx.
- The graph seems to be a leptokurtic distribution.

From the table we can see that the average price of the automobiles present in the dataset is 13207, the midpoint or the median of the frequency distribution of price is at 10595.

price	
Mean	13207.13
Standard Error	549.5786
Median	10595
Mode	13207.13
Standard Deviation	7868.768
Sample Variance	61917513
Kurtosis	3.354216
Skewness	1.827324
Range	40282
Minimum	5118
Maximum	45400
Sum	2707462
Count	205

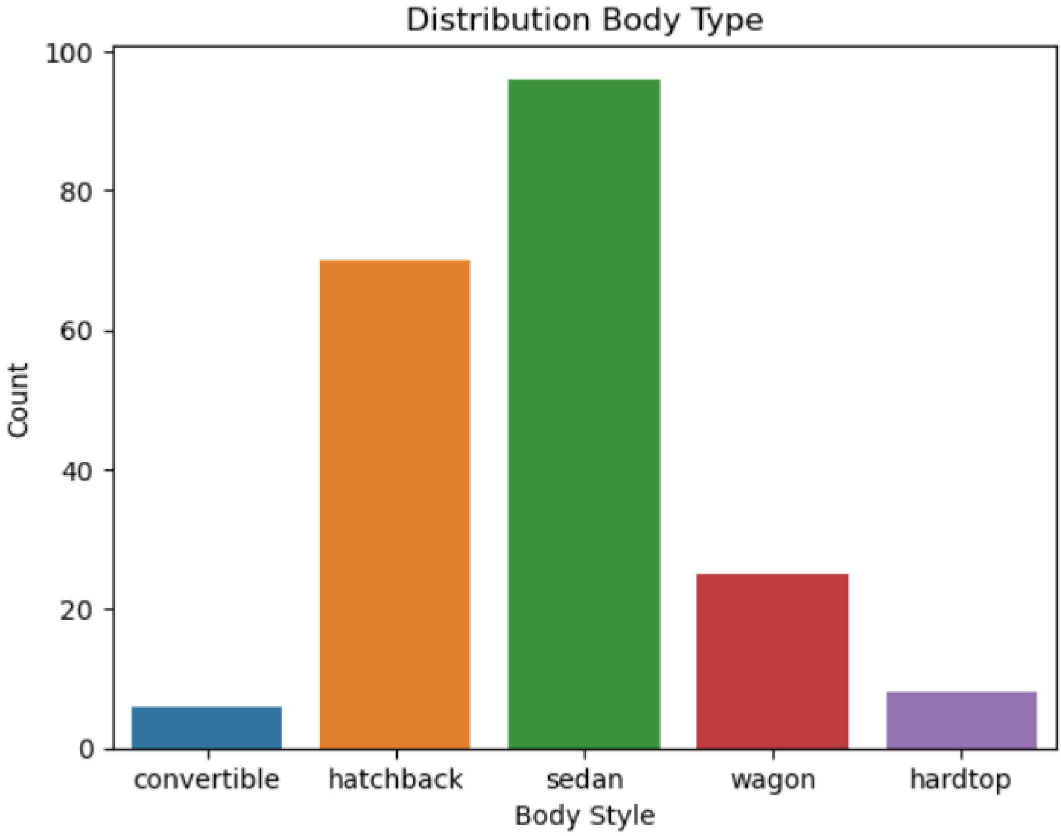
The standard deviation of the price is 7868, it is a measure of the amount of variation or dispersion of automobile prices in the dataset.

The standard error of the price is 549, which tells us how much the average(mean) price of cars in your dataset might differ from the actual average(mean) price of all cars in the population. The highest and the lowest price present in the dataset is 45400 and 5118 respectively, and the difference between them (Range) is 40282.

Kurtosis >3 indicates a distribution with heavier tails (leptokurtic). This means there are more extreme values or outliers present in the dataset than that of in a normal distribution.

A skewness value of 1.82 indicates that the price distribution of cars is positively skewed. In other words, there are more car prices that are lower than the mean price, with a few car prices that are much higher than the mean.

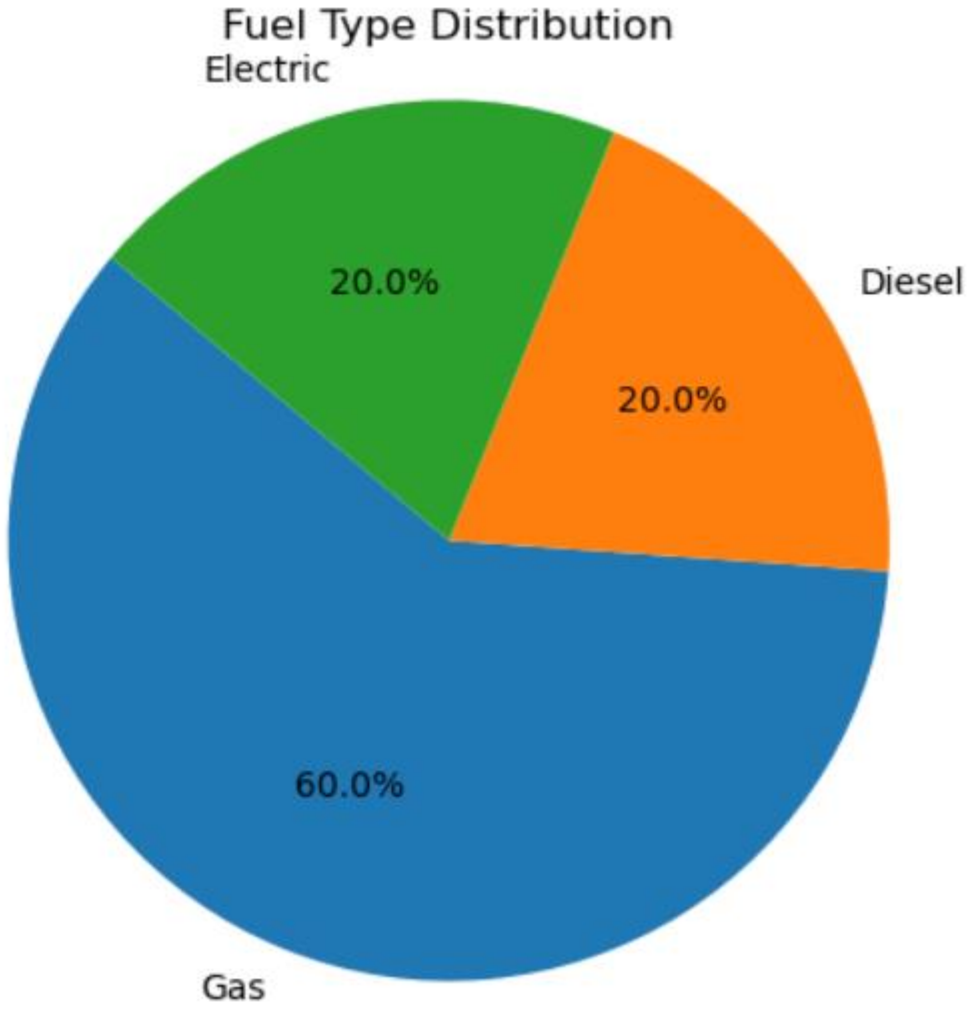
Distribution of Body Type:



Insights

- Sedan and Hatchback are the most common body style present in the dataset.
- Convertible is the least common body style.
- Wagon and Hardtop have some-what a significant percentage in the dataset.

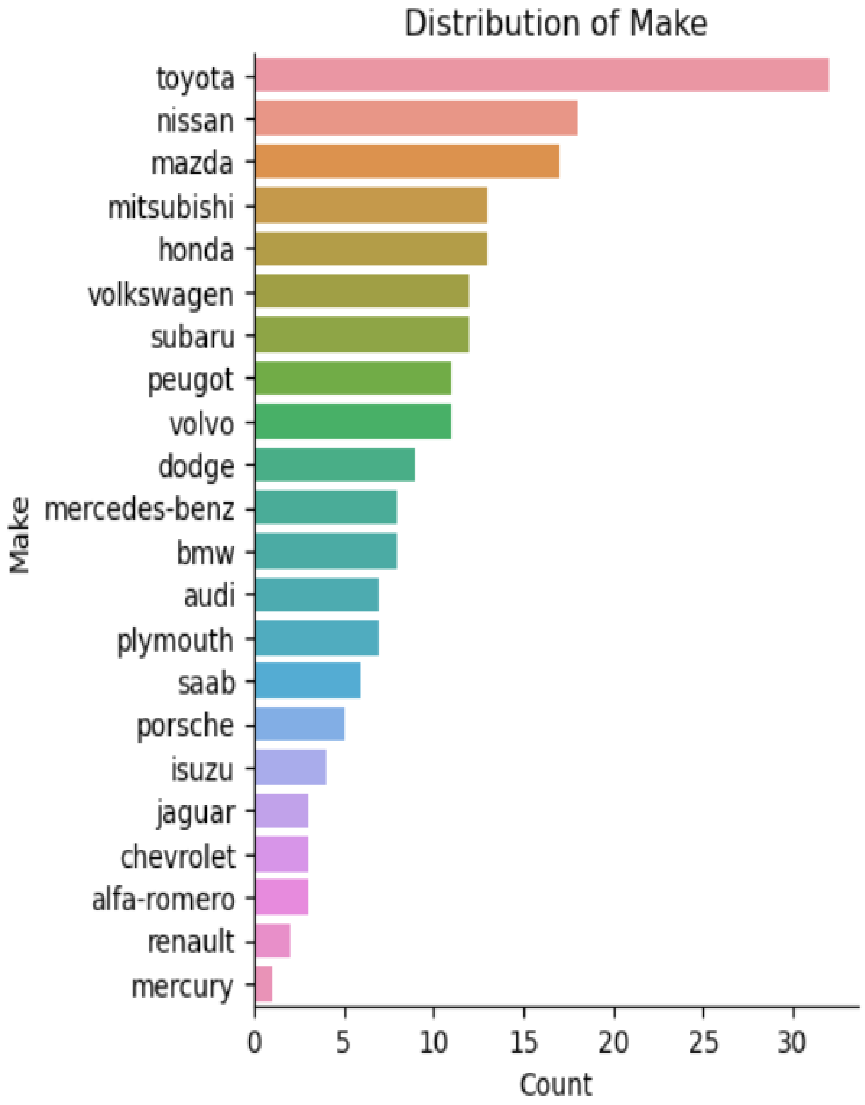
Fuel Type Distribution:



Insights

- Majority of the cars present in the dataset runs on Gas.
- Electric and Disel consists of the rest 40% OF Fuel type distribution

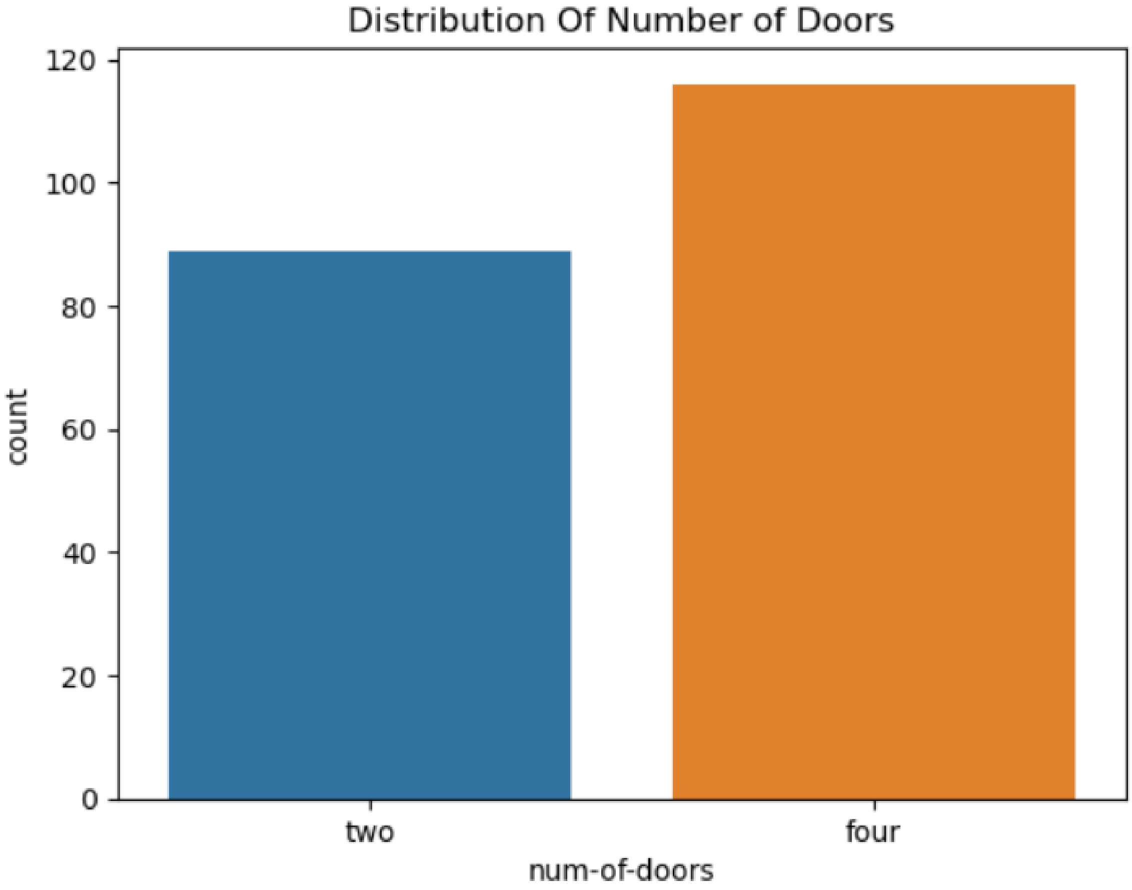
Distribution of Make of Cars



Insights

- **Toyota** has the highest count among all listed makes, making it the most popular in the dataset.
- Other notable makes include **Nissan**, **Mazda**, and **Mitsubishi**.
- **BMW**, **Audi**, and **Mercedes-Benz** also appear in the chart, but with lower counts.

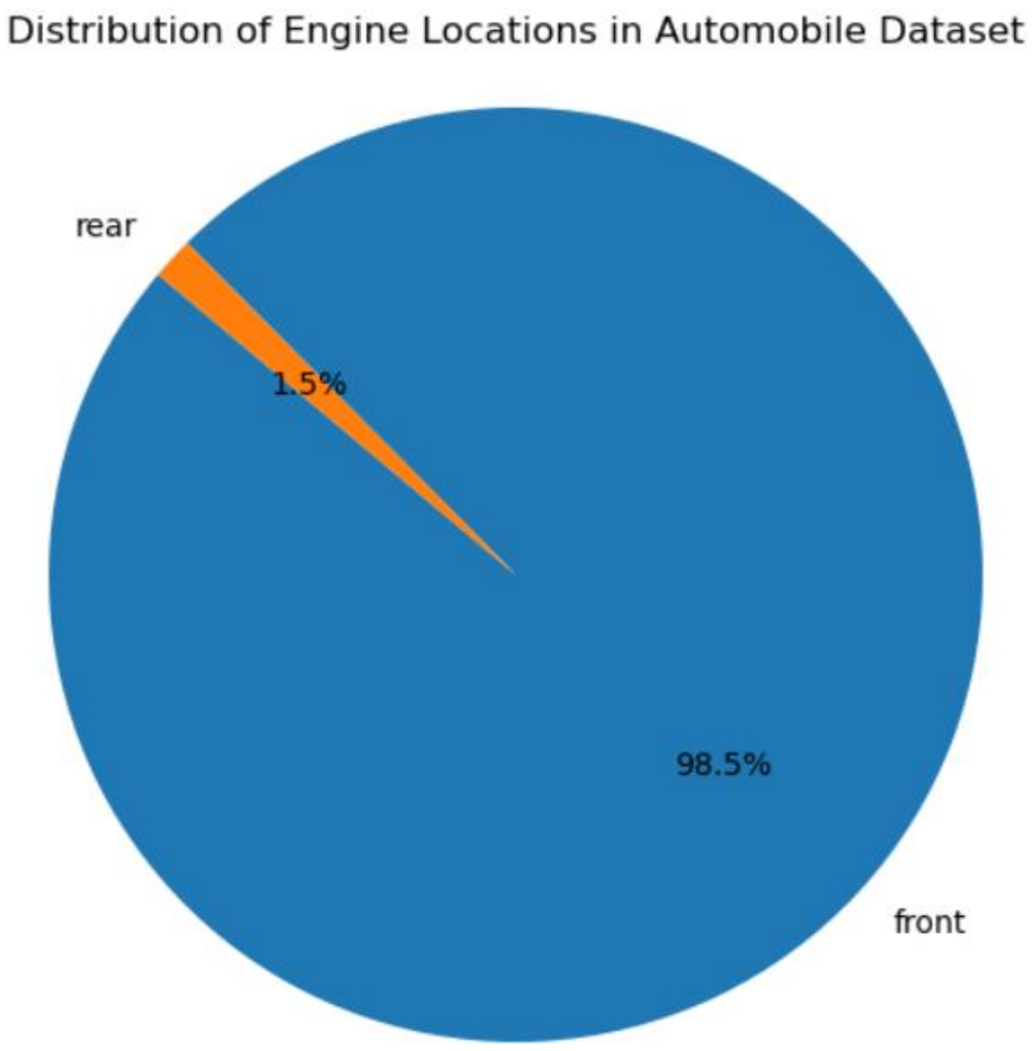
Distribution of Numbers of Doors



Insights:

- The category **“four”** (representing cars with four doors) has a higher count, reaching approximately 120 on the y-axis.
- The category **“two”** (representing cars with two doors) has a lower count, extending up to approximately 80 on the y-axis.

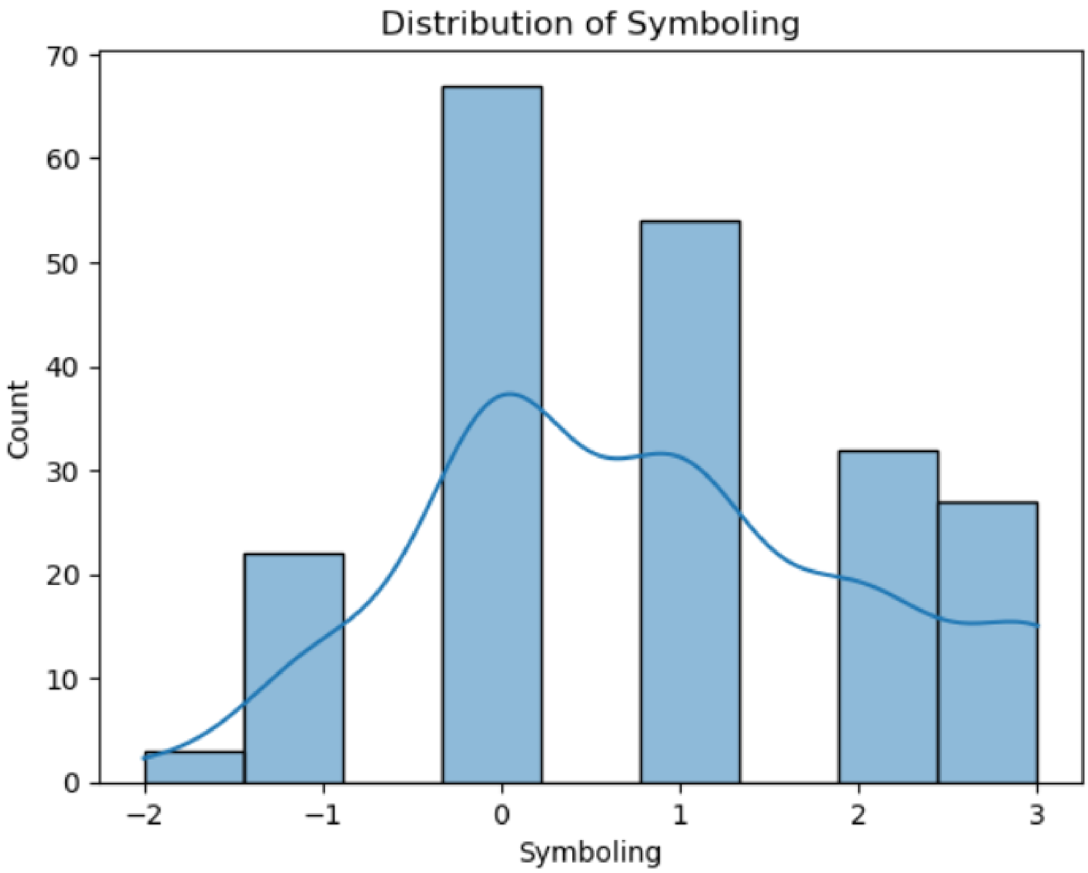
Distribution of Engine Locations



Insights

- 98.5% cars present in the dataset has their engine located in the front
- where-as only 1.5% of them has their engine located in the rear.

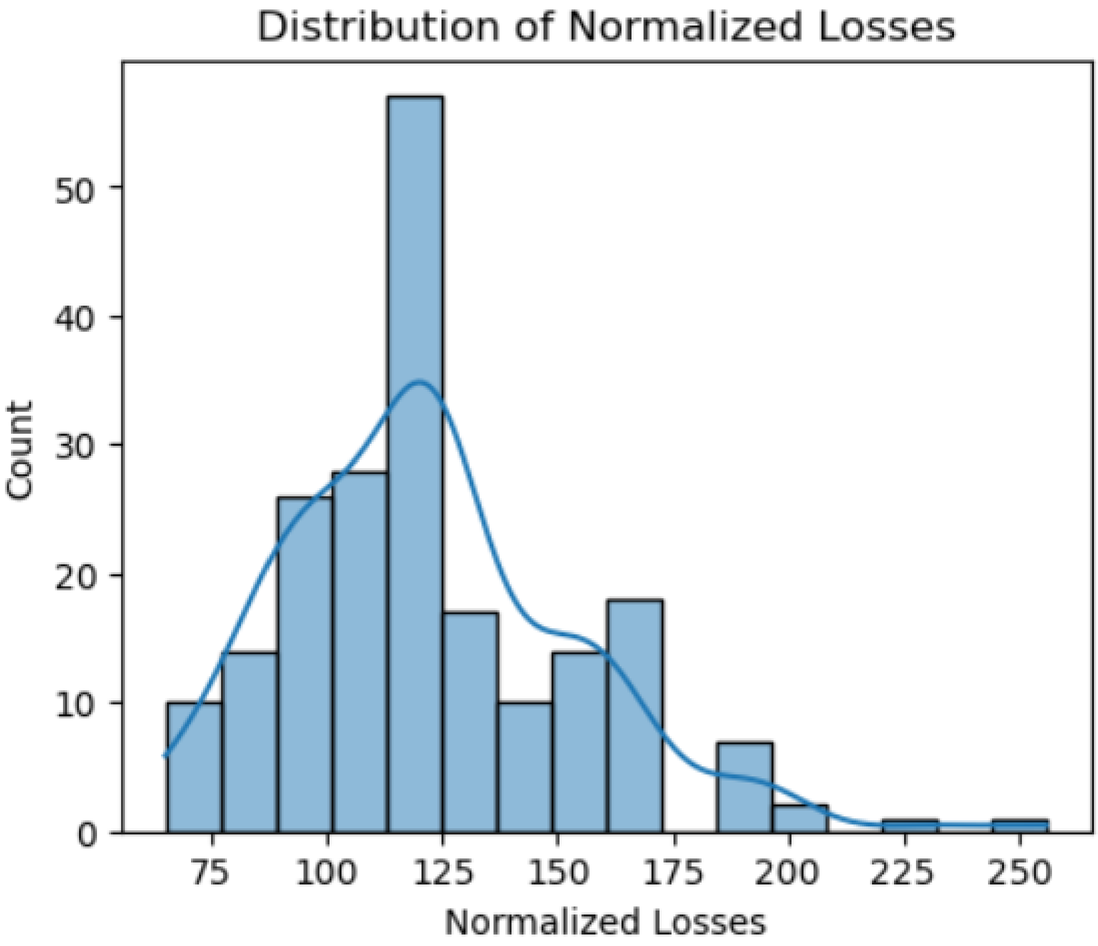
Distribution of Symboling



Insights

- Higher symboling values indicate higher risk (e.g., more prone to accidents or theft).
- The mode for the distribution seems to be at zero which means most of the cars have very less risk.

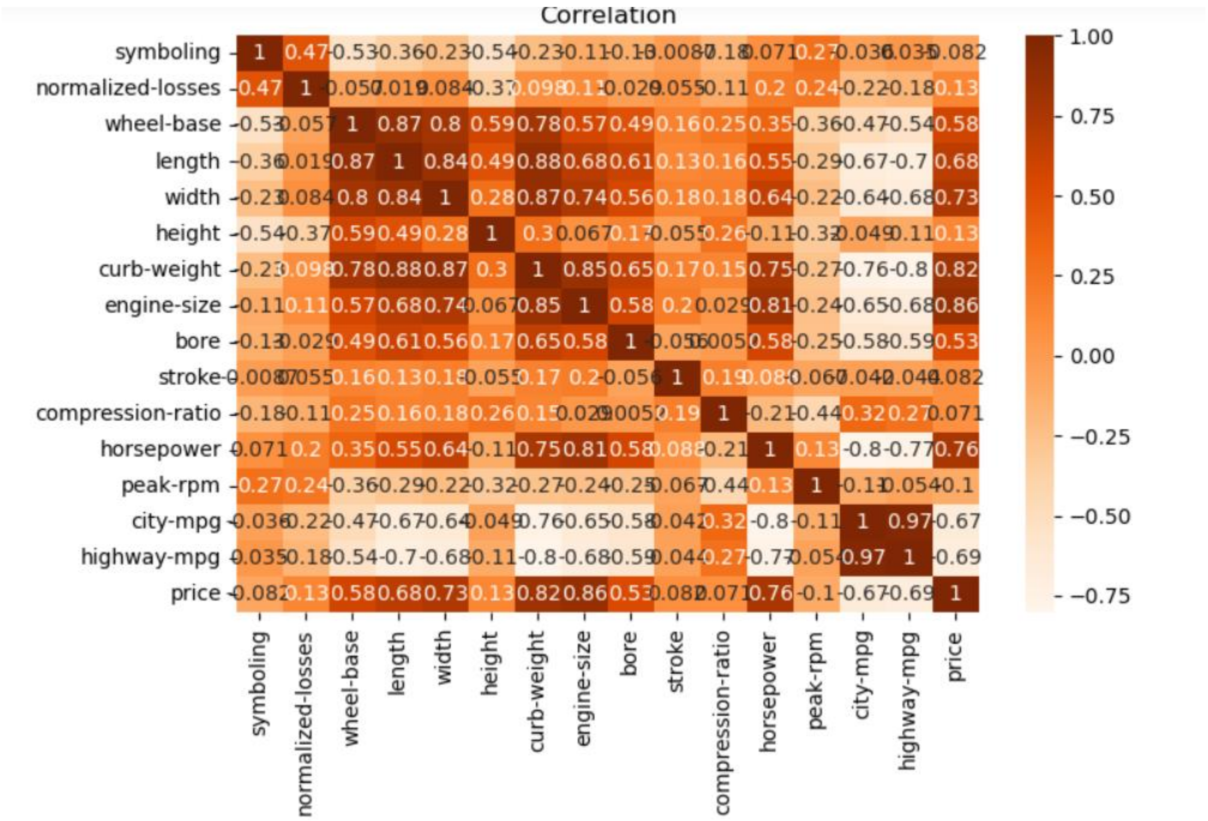
Distribution of Normalised Losses



Insights

- The majority of data points appear to be concentrated around the 100 to 125 range on the x-axis (normalized losses). This suggests that losses are more frequent within this interval.
- The distribution of normalized losses is possibly a normal distribution.
- There seem to be a few outliers with higher normalized losses ,

Analysing Correlation between the variables:



The above heat map shows how the variables are correlated among one another.

The correlation coefficient ranges from -1 to 1.

A coefficient close to 1 indicates a strong positive correlation (as one variable increases, the other tends to increase). Where-as if the value is closer to -1 indicates a strong negative correlation (as one variable increases, the other tends to decrease).

A coefficient around 0 suggests no significant linear relationship between the variables.

TESTING OF HYPOTHESIS

T-test on ‘Price’ and ‘Engine Location’

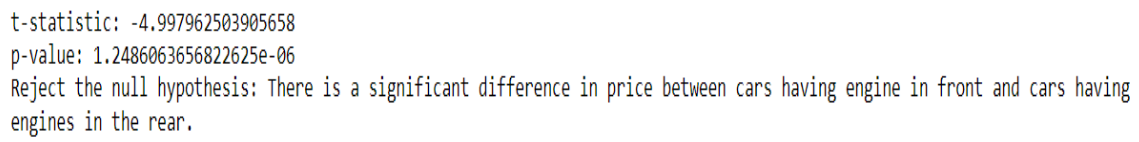
A t-test on Price of the cars and Engine Location has been executed to check if there is any significant difference among price of cars having engine location in the front of the car vs price of cars having engine location in the rear of the car.

Here, we are assuming

H_0 (null hypothesis) = there is no difference between prices of cars based on engine location.

H_1 (alternate hypothesis) = there is a significant difference between prices of cars based on engine location.

The results are as follows:

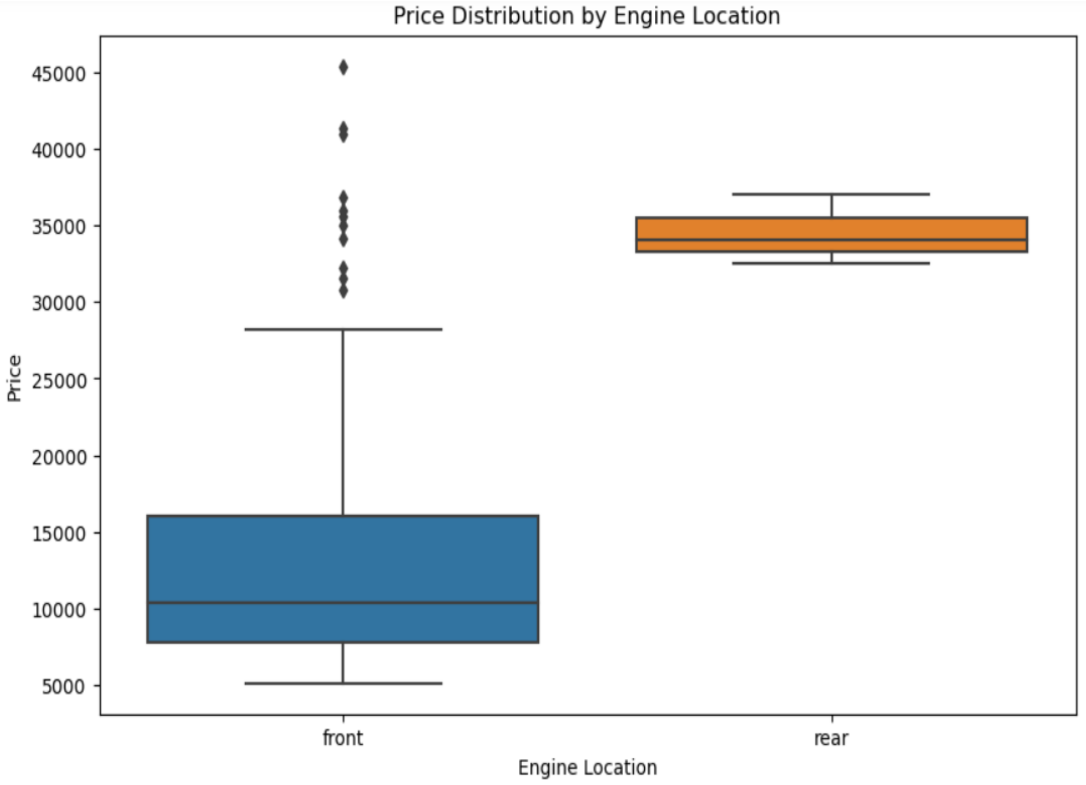


```
t-statistic: -4.997962503905658
p-value: 1.2486063656822625e-06
Reject the null hypothesis: There is a significant difference in price between cars having engine in front and cars having
engines in the rear.
```

We can see that the alternative hypothesis is true.

Let us see the difference with the help of box plots given in the next page;

Distribution of Price based on Engine Location



Insights

- The “front” category has a wider interquartile range (IQR), indicating greater variability in prices. This suggests that vehicles with front engines have a broader range of prices.
- The “rear” category, on the other hand, has a narrower IQR but a higher median price. This implies that vehicles with rear engines tend to be more expensive.
- Both categories have outliers, which are data points significantly different from the rest.

T-test on ‘Highway-milage’ and ‘Fuel-type’

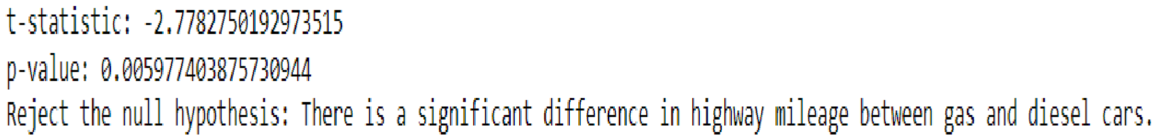
A t-test on highway-milage of the cars and fuel-type has been executed to check if there is any significant difference among highway-milage of cars having gas as a fuel type vs highway-milage of cars having diesel as a fuel type.

Here, we are assuming

H_0 (null hypothesis) = there is no difference between highway-milage of cars based on its fuel type.

H_1 (alternate hypothesis) = there is a significant difference between highway-milage of cars based on its fuel type.

The results are as follows:

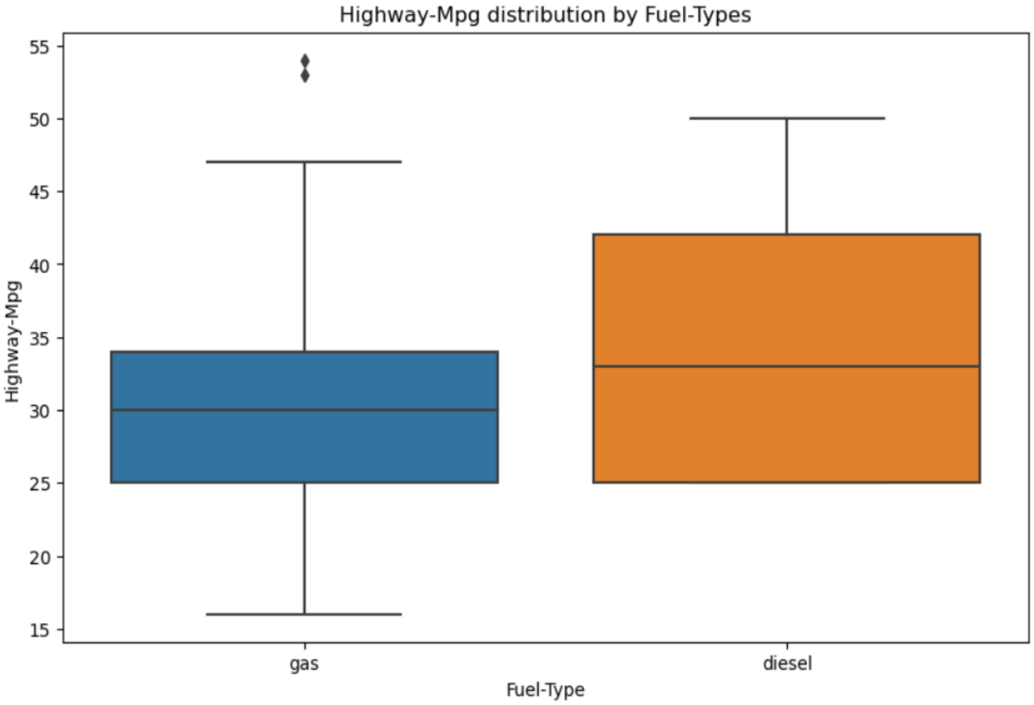


```
t-statistic: -2.7782750192973515
p-value: 0.005977403875730944
Reject the null hypothesis: There is a significant difference in highway mileage between gas and diesel cars.
```

We can see that the alternative hypothesis is true.

Let us see the difference with the help of box plots given in the next page;

Distribution of Highway-Mpg based on Fuel-Types



Insights

- Gas vehicles tend to have a wider range of highway mpg values, with some outliers achieving higher efficiency.
- Diesel vehicles show a narrower range, but their median mpg is generally higher.
- The interquartile range (IQR) for diesel vehicles is smaller, indicating less variability.

Chi-Squared Test on ‘Fuel-Type’ and ‘Aspiration’

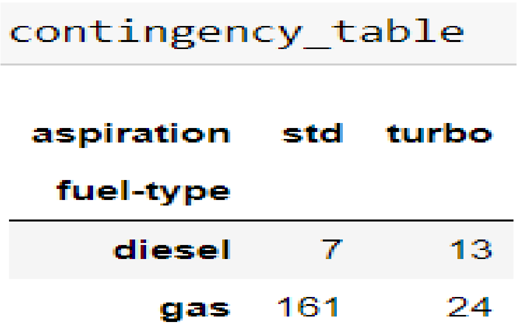
To determine if there is any significant association between ‘fuel-type’ and ‘aspiration’ of the cars, a chi-squared test has been executed.

Here, we are assuming

H_0 (null hypothesis) = there is no association between cars based on its fuel type and aspiration.

H_1 (alternate hypothesis) = there is a significant association between cars based on its fuel type and aspiration.

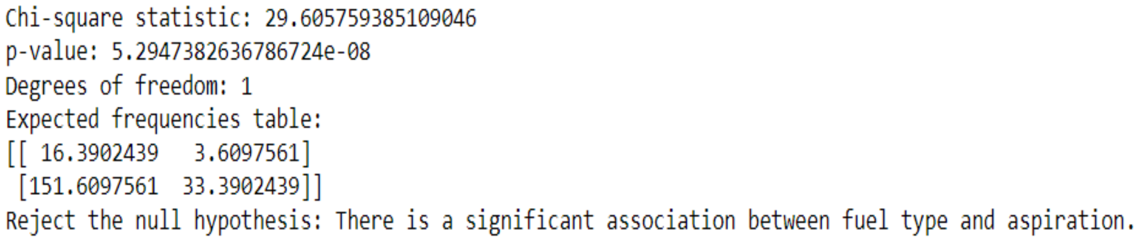
The contingency table and the results are as follows:



```
contingency_table
```

aspiration	std	turbo
fuel-type		
diesel	7	13
gas	161	24

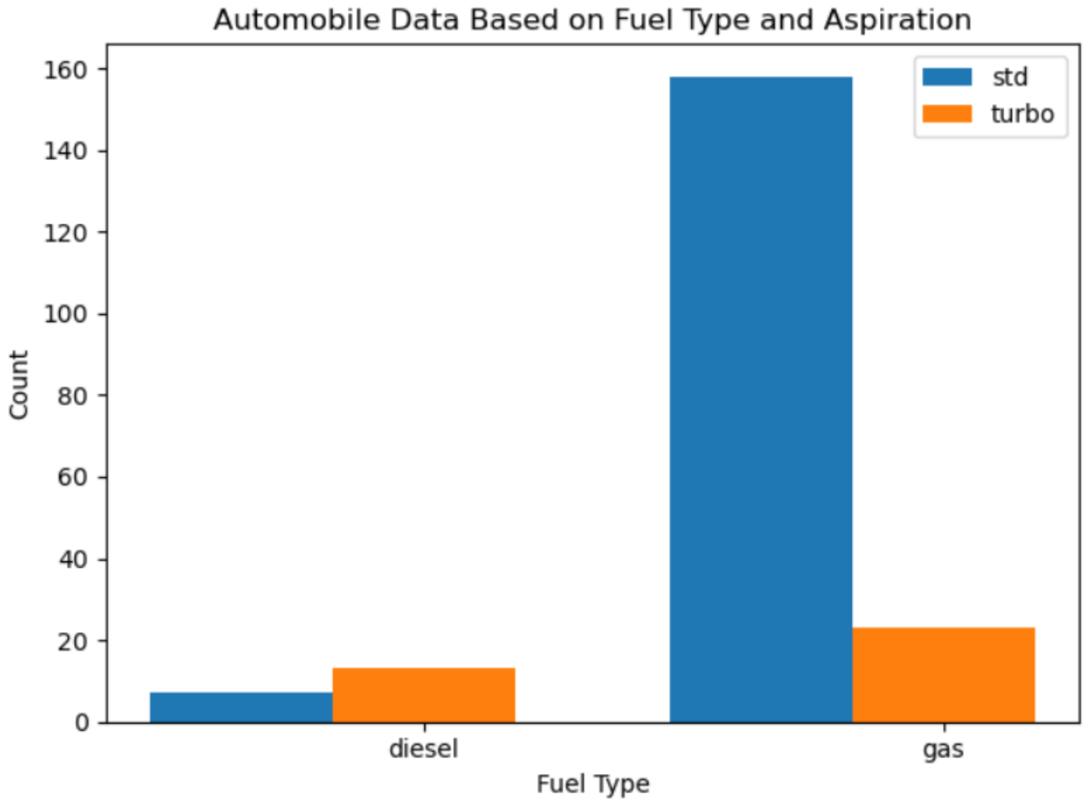
Results



```
Chi-square statistic: 29.605759385109046
p-value: 5.2947382636786724e-08
Degrees of freedom: 1
Expected frequencies table:
[[ 16.3902439   3.6097561]
 [151.6097561  33.3902439]]
Reject the null hypothesis: There is a significant association between fuel type and aspiration.
```

We can see that the alternative hypothesis is true.

Visualization based on Fuel Type and Aspiration



Insights

- Gas-fueled vehicles (both standard and turbocharged) are more common than diesel-fueled ones in the dataset.

- The ‘gas’ category has significantly higher counts for both ‘std’ (standard) and ‘turbo’ (turbocharged) aspiration types.

  Within each fuel type:

- Gas vehicles have more ‘std’ aspiration (standard) than ‘turbo’ aspiration.
- Diesel vehicles also follow a similar trend, but the difference is less pronounced.

Annova Between Fuel Type and Drive Wheels

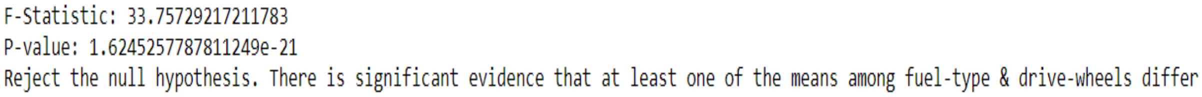
To determine if there is any significant association between ‘fuel-type’, ‘drive-wheels’ and ‘price’ of the cars, an annova test has been executed.

Here, we are assuming

H_0 (null hypothesis) = there is no association between price of cars based on its fuel type and drive-wheels.

H_1 (alternate hypothesis) = there is an association between price of cars based on its fuel type and drive-wheels.

Results derived from the annova test are as follows:

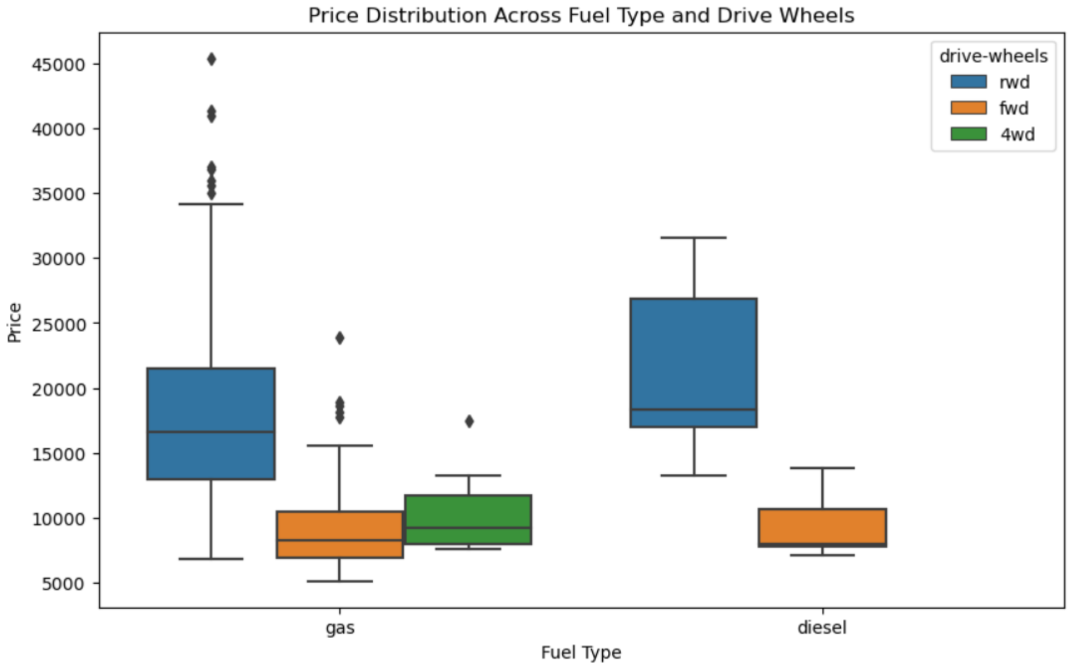


```
F-Statistic: 33.75729217211783
P-value: 1.6245257787811249e-21
Reject the null hypothesis. There is significant evidence that at least one of the means among fuel-type & drive-wheels differ
```

We can see that the alternative hypothesis is true.

To visualize the above relation between fuel-type’, ‘drive-wheels’ and ‘price’ of the cars a box plot graph is presented below.

Price Distribution Across Fuel Type and Drive Wheels



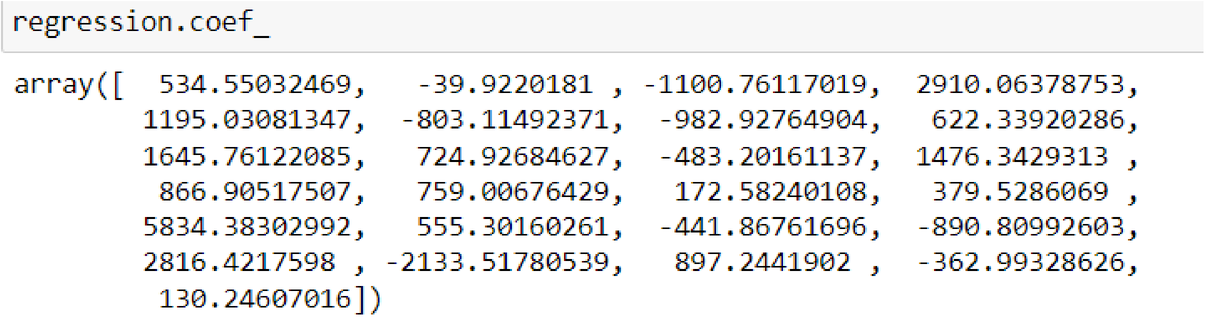
Insights

- We can see that all the three types of drive- wheels are present in the cars having gas as fuel.
- We can observe many outliers for the category gas as fuel type.
- Drive wheel category rwd has the highest price range in diesel cars.

Multiple Linear Regression

To model relationships between a Price and multiple independent variables (like symbolling, normalized losses, make, fuel type, number of doors, aspiration, body -style , drive wheels, engine location, wheel base , number of cylinders, engine size , highway-mpg, city-mpg , peak-rpm etc.) simultaneously a multiple linear regression is executed.

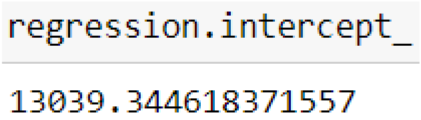
Regression Coefficients are:



```
regression.coef_
```

```
array([  534.55032469,   -39.9220181 , -1100.76117019,  2910.06378753,
        1195.03081347,  -803.11492371,  -982.92764904,   622.33920286,
        1645.76122085,   724.92684627,  -483.20161137,  1476.3429313 ,
         866.90517507,   759.00676429,   172.58240108,   379.5286069 ,
        5834.38302992,   555.30160261,  -441.86761696,  -890.80992603,
        2816.4217598 , -2133.51780539,   897.2441902 ,  -362.99328626,
         130.24607016])
```

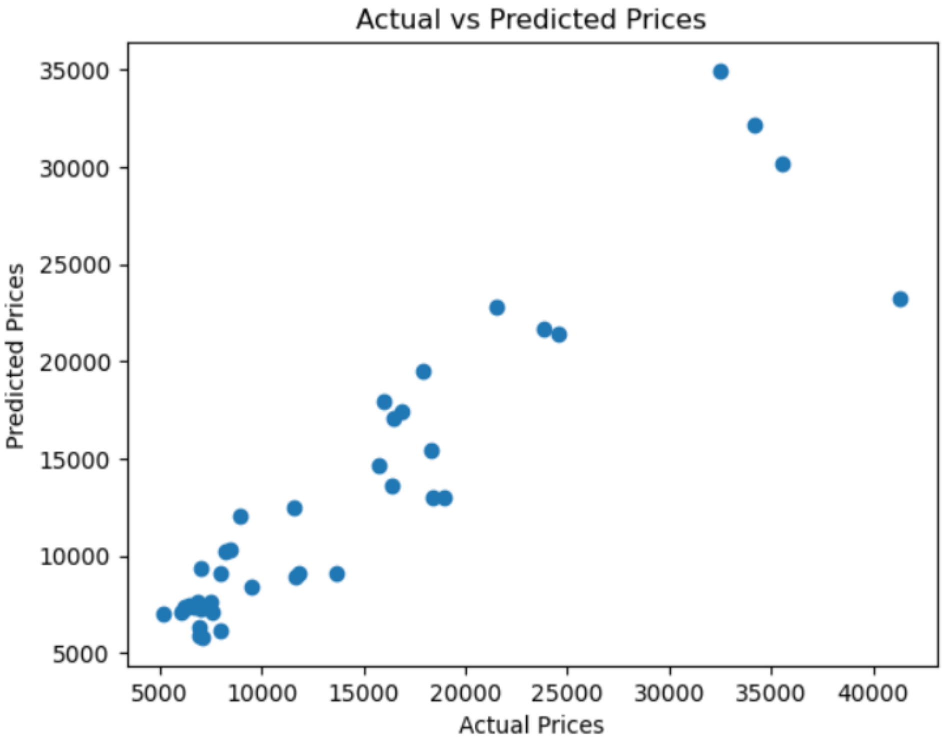
Regression Intercept is at:



```
regression.intercept_
```

```
13039.344618371557
```

Relation among actual and predicted price is shown using a scatter plot in the following page:



- The mean squared error is 13395419.3268793
- The mean absolute error is 2289.069010600281
- The square root of mean squared error is 3659.9753177964603
- The R_2 score is 0.8352309678125501
- The adjusted R_2 score is 0.8122185331494984

List of Findings:

- The price distribution of cars is positively skewed and is Leptokurtic.
- Normalized Losses shows a positively skewed distribution and is platykurtic
- The majority of the cars present in the dataset has their engine located in the front.
- **Toyota** has the highest count among all listed makes, making it the most popular in the dataset.
- The majority of the cars present in the automobile dataset are having four doors.
- Sedan and Hatchback are the most common body style present in the dataset where-as, convertible is the least common body style.
- The mode for the distribution symboling seems to be at zero which means most of the cars have very less risk.
- Gas vehicles tend to have a wider range of highway mpg values, with some outliers achieving higher efficiency.
- We can see that all the three types of drive- wheels are present in the cars having gas as fuel.
- Drive wheel category rwd has the highest price range in diesel cars.
- Gas-fueled vehicles (both standard and turbocharged) are more common than diesel-fueled ones in the dataset.

- The ‘gas’ category has significantly higher counts for both ‘std’ (standard) and ‘turbo’ (turbocharged) aspiration types.

- Gas vehicles have more ‘std’ aspiration (standard) than ‘turbo’ aspiration.
- Diesel vehicles also follow a similar trend, but the difference is less pronounced.

THANK YOU