# 

Masters in Applied Statistics and Data Science (MASDS)

Department of Statistics

Jahangirnagar University

Savar, Dhaka-1342, Bangladesh.

An Assignment on

*Regression Analysis Problem*

**Introduction to Data Science with Python**

**WM-ASDS04**

Submitted to

**Farhana Afrin Duty**

Submitted by

**Md. Tawfiq-Uz-Zaman**

Roll-20229025

MASDS 9th Batch, Section-A

# Libraries Imported

The very first step which has been carried out is importing libraries like numpy, pandas, seaborn, and matplotlib. Decision Tree, Random Forest Regressor are also imported for regression purposes.

## Dataset

Our dataset, in the form of a csv file, consists of some car features. The name of the file is CarPrice\_Assignment.csv. Here, ‘price’ is the target column.

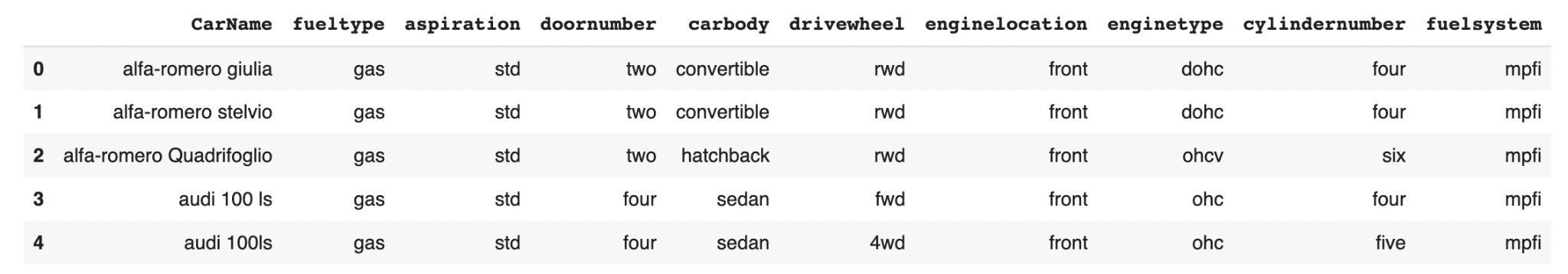
Some observations regarding the dataset include:

1. From the output it can be said that there are 205 rows/records and 26 columns/Features present in our dataset.
2. From the output it can be observed that there are 10 categorical & 16 numerical Attributes.
3. All the features are having correct data-types. So we don't have to do any changes.

## Descriptive Statistics Analysis

Firstly, we checked some basic properties of the dataset such as mean, standard deviation, minimum & maximum values, and 1st, 2nd, & 3rd quartiles. There are no missing or duplicate values found after checking for NaN and duplicate values respectively.

The following image shows only the categorical features.

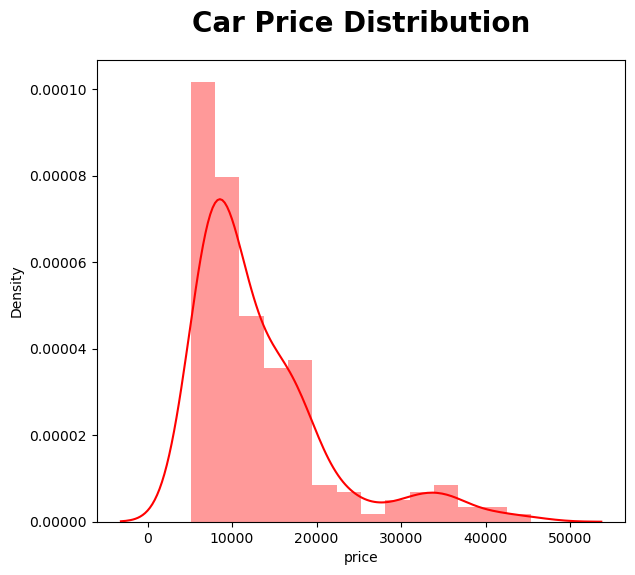


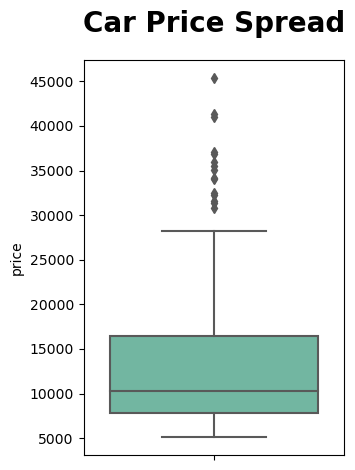
In the CarName Feature we can observe that the values are storing both the car's company name and the Car name. So we have to clean that Feature. We can separate the car company names from Carname Feature.

There were spelling errors in some of the car company names. We fix those up and get a list of the unique company names.

## EDA

At first, we visualized the Target Features using PyPlot.

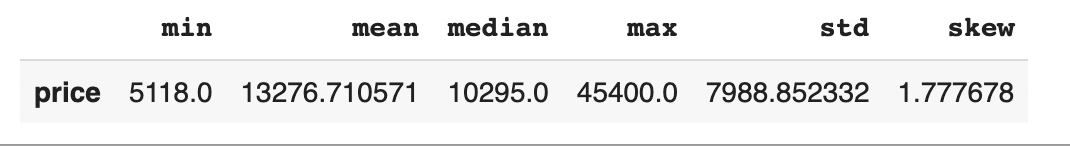




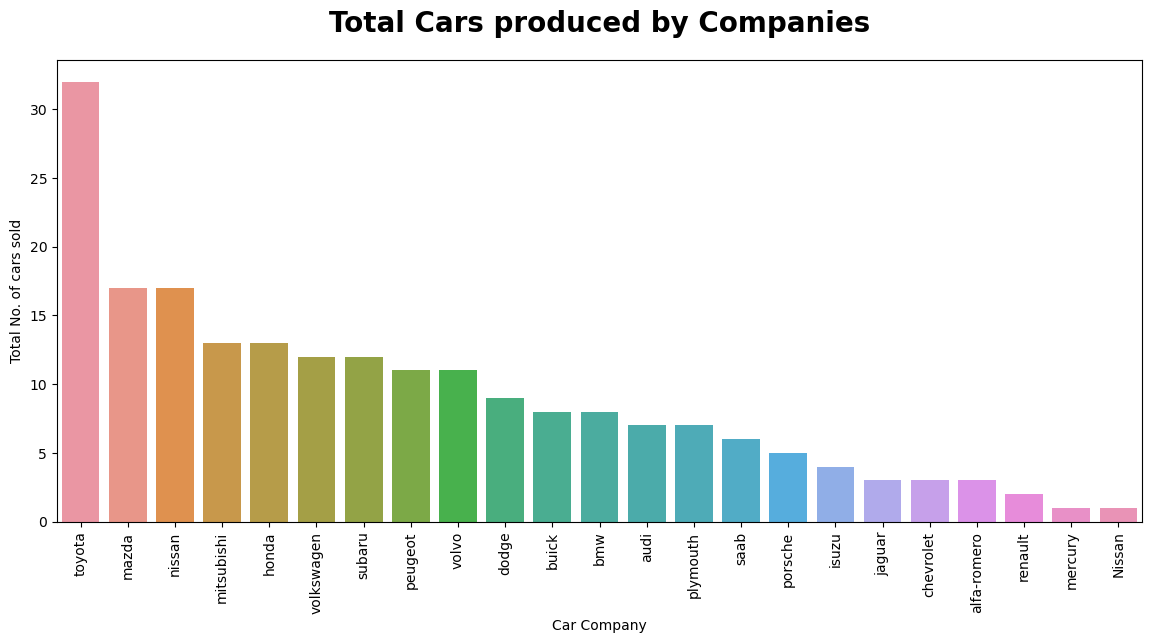
Here are a few observations from the above figures:

* We can clearly observe that our Car Price Feature is Right Skewed.
* We can clearly observe that there is a significant difference between mean & median value.
* We can also make an insight that most of the car's price is below 14000.
* We can also that the skewness of the car price is above 1.5 which means that the data points are highly spread

The descriptive statistics is provided below:



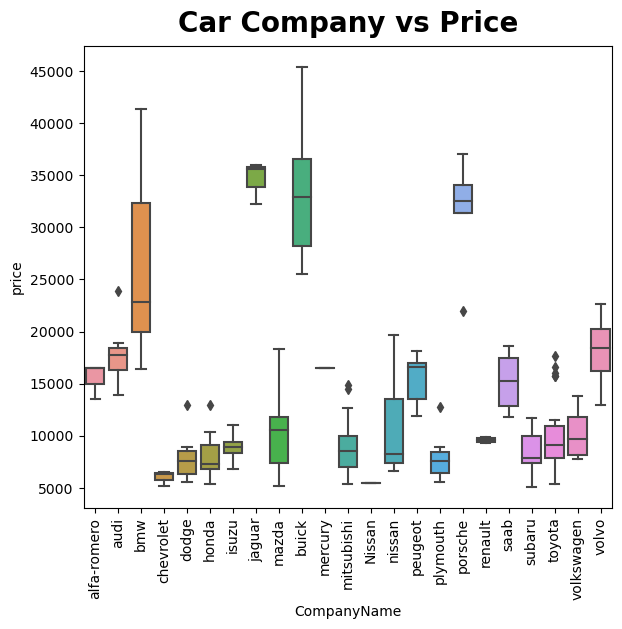
The total number of cars produced by companies is as follows:

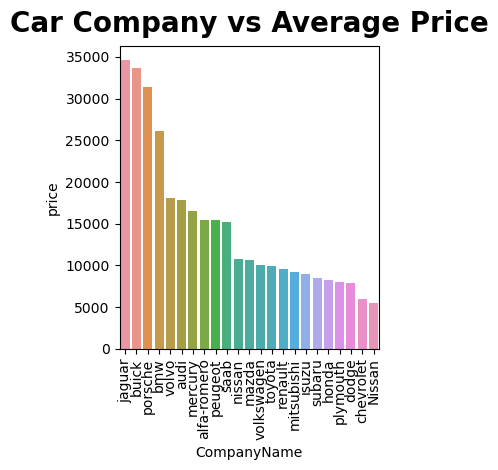


Here, we can observe the following:

* Toyota has sold the highest number of cars.
* So we can say that Toyota is kind of a customer's most favored company.
* Nisaan ,Mercury or Renault are having very low data-points. So we can't make any inference of least sold car companies.

In the following step, we distinguished between Car Company vs Price and Car Company vs Average Price





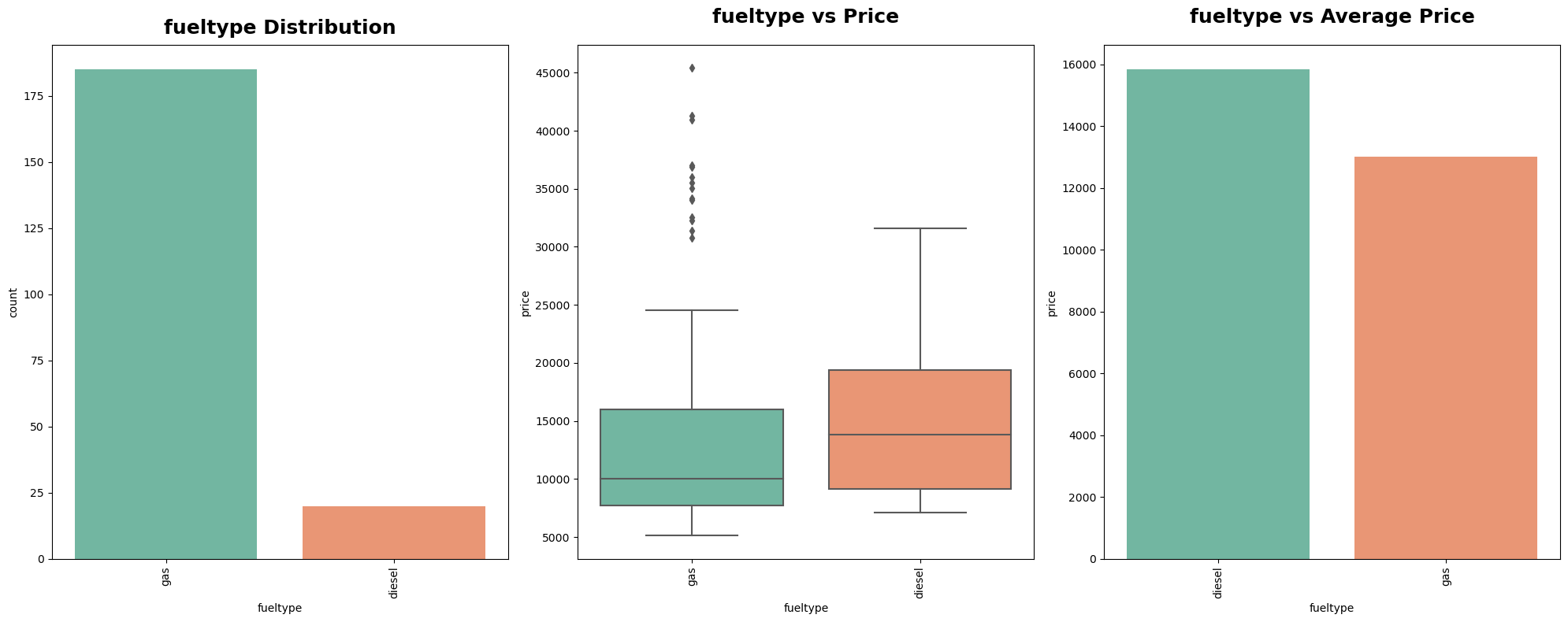
The following things can be observed from the above comparison:

* Jaguar & Buick seem to have the highest price range cars.
* Car companies like Nisaan,Renault & Mercury are having only one to two data points.
* So we can't make any inference related to lowest price range car companies.

Since there are too many categories in car company features. So we derive a new feature Company Price Range which will show the price range as Low Range, Medium Range, High Range.

We have visualized the following plots related to all the features in our dataset.

# Checking Car Fuel Type Feature



Here, we can clearly see that Cars that have a gas fuel system are mostly sold.

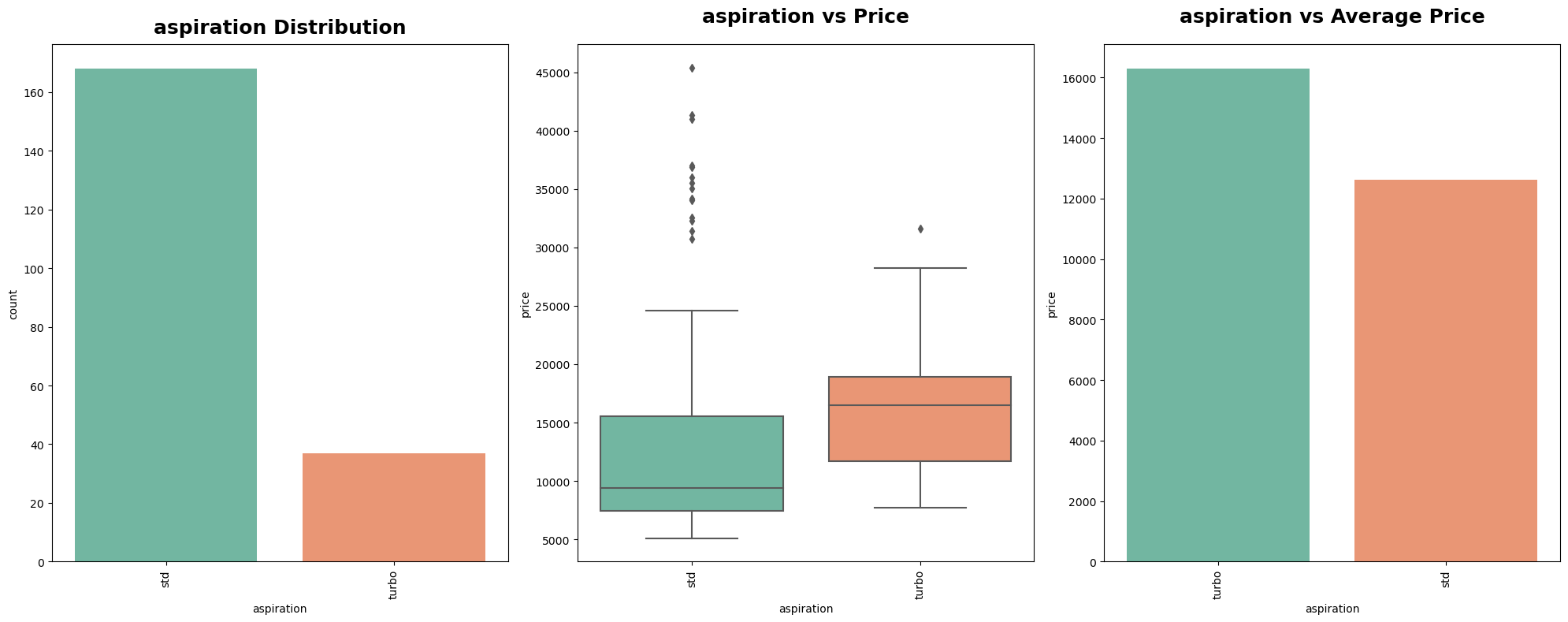
From the second plot we can make an insight that Gas Fuel System cars are available within every price range.

Both types of fuel cars are having most of the cars price near to average price of the cars. But the gas Fuel Type cars are having cars with high prices also.

From the third plot we can make an insight that the Average price of gas fuel type cars are less than diesel fuel type cars.

So we can say that customers prefer those cars which consume less price in fuel.

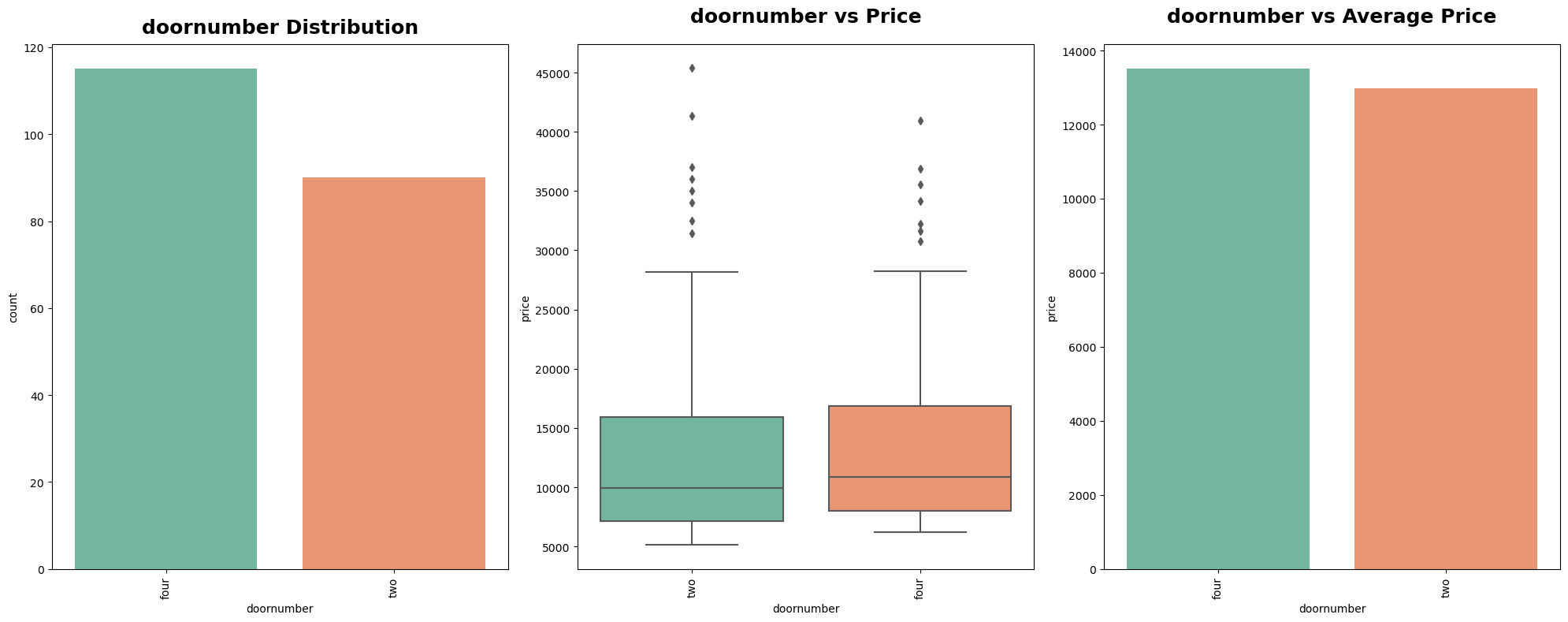
# Checking the Aspiration Feature



Here are some observations for the above plots:

* Cars having STD Aspiration are highly sold when compared with Turbo Aspiration.
* Cars having Turbo Aspiration have a higher price range than the STD Aspiration. Insights
* Outliers present in STD Aspiration states that some of the cars having std aspiration are expensive too.

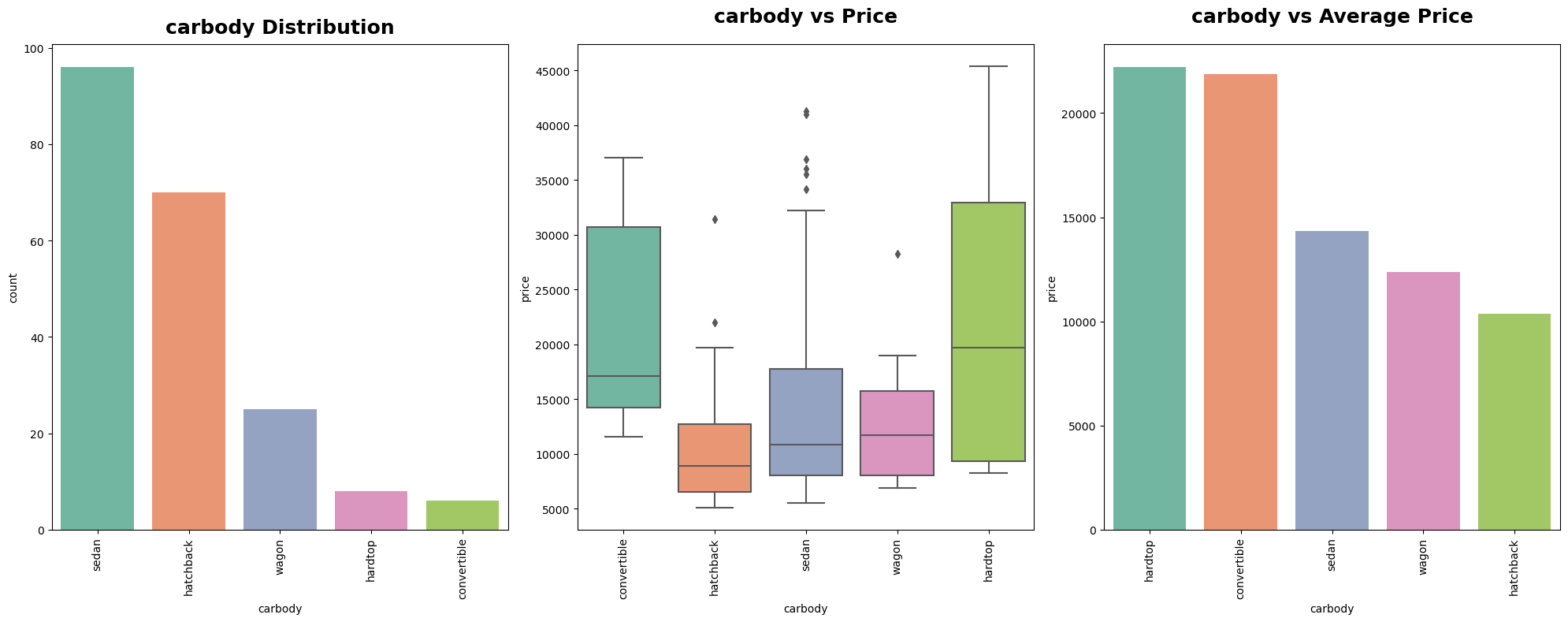
# Checking the Doornumber Feature



Here we observe that:

* Cars having Four Doors are mostly sold when compared with cars having Two Doors.
* Cars having Fours Doors are a little more expensive when we compare the average price of cars having two doors or four doors. Insights
* Cars having Four Doors are mostly sold. But there's not a big difference between the sales of cars having four doors & two doors.
* Outliers present in two doors cars state that some of the cars are more expensive than the cars having four doors Door numbers category is not affecting the price that much.

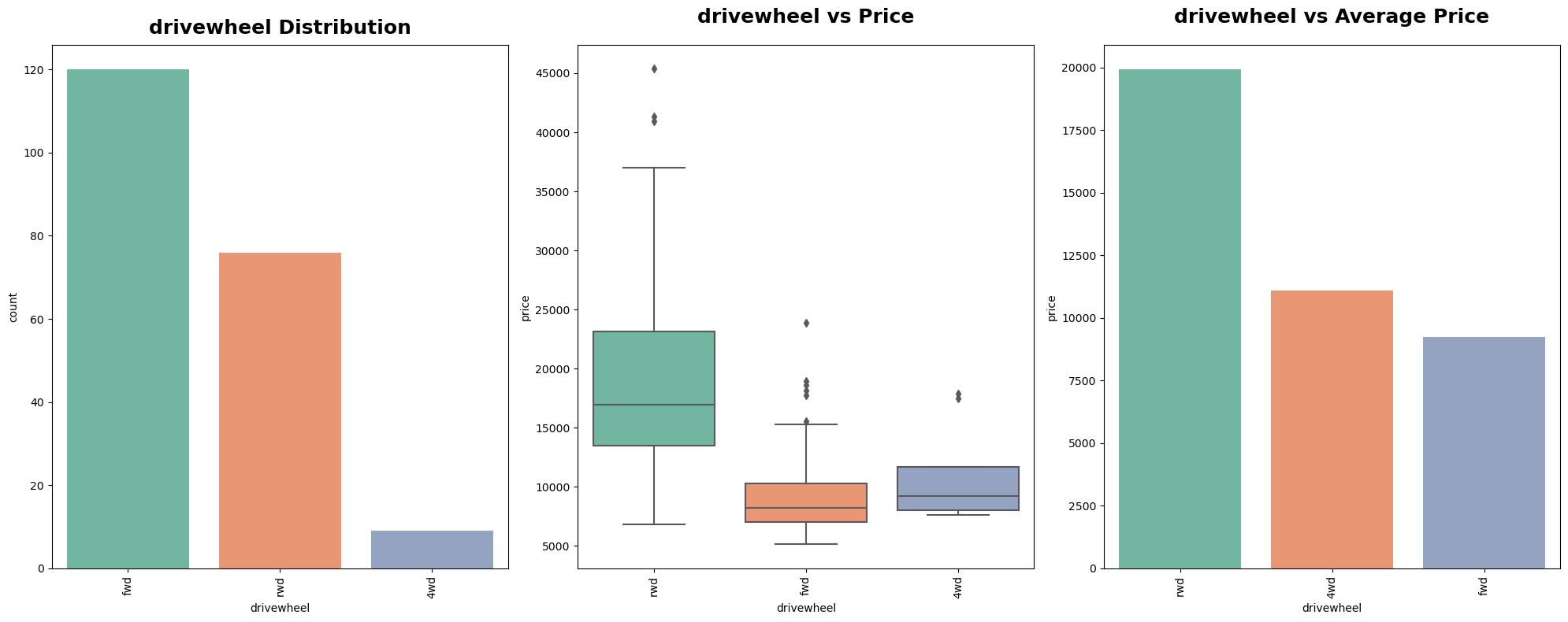
# Checking the Carbody Feature



What we learn from the above graphs are as follows:

* Cars having sedan body-type are mostly sold followed by hatchback.
* Cars having convertible or hardtop body-type are less sold.
* Cars having Hardtop body-type are the most expensive cars followed by convertibles.
* Insights
* Convertible & Hardtop body-types are less sold because they are very expensive cars. So most of the customers couldn't prefer it.
* Cars having Sedan body-type are the third most expensive car. But still it has the most number of car sales. So we can say customers prefer medium price range cars

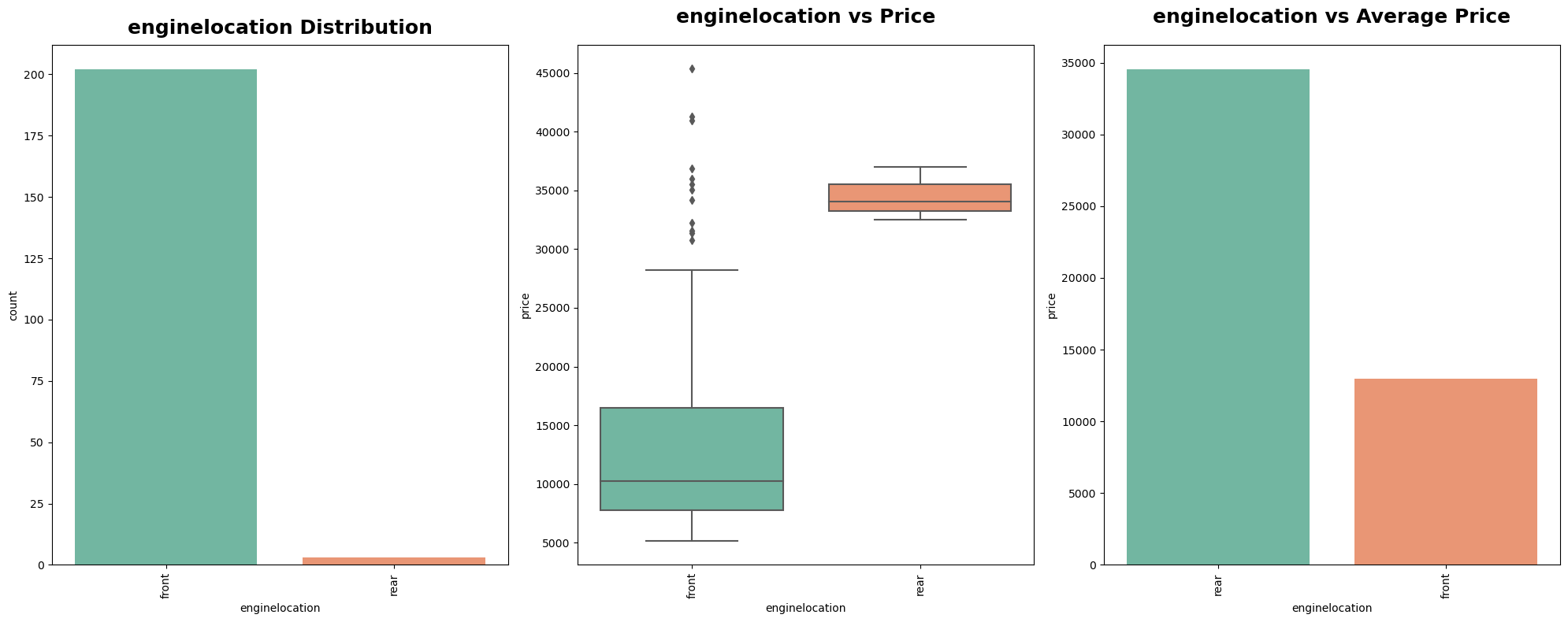
# Checking the DriveWheel Feature



Observations for the above plots

* Most of the cars which are sold have Front Wheel Drive (FWD) followed by Rear Wheel Drive (RWD).
* Cars having Rear Wheel Drive are mostly expensive cars. But cars having Front Wheel Drive are the cheapest.
* Insights
* We can make an insight that the high rated cars must be using Rear Wheel Drive. Cars with low prices must be using Forward Wheel Drive

# Checking the Engine Location Feature

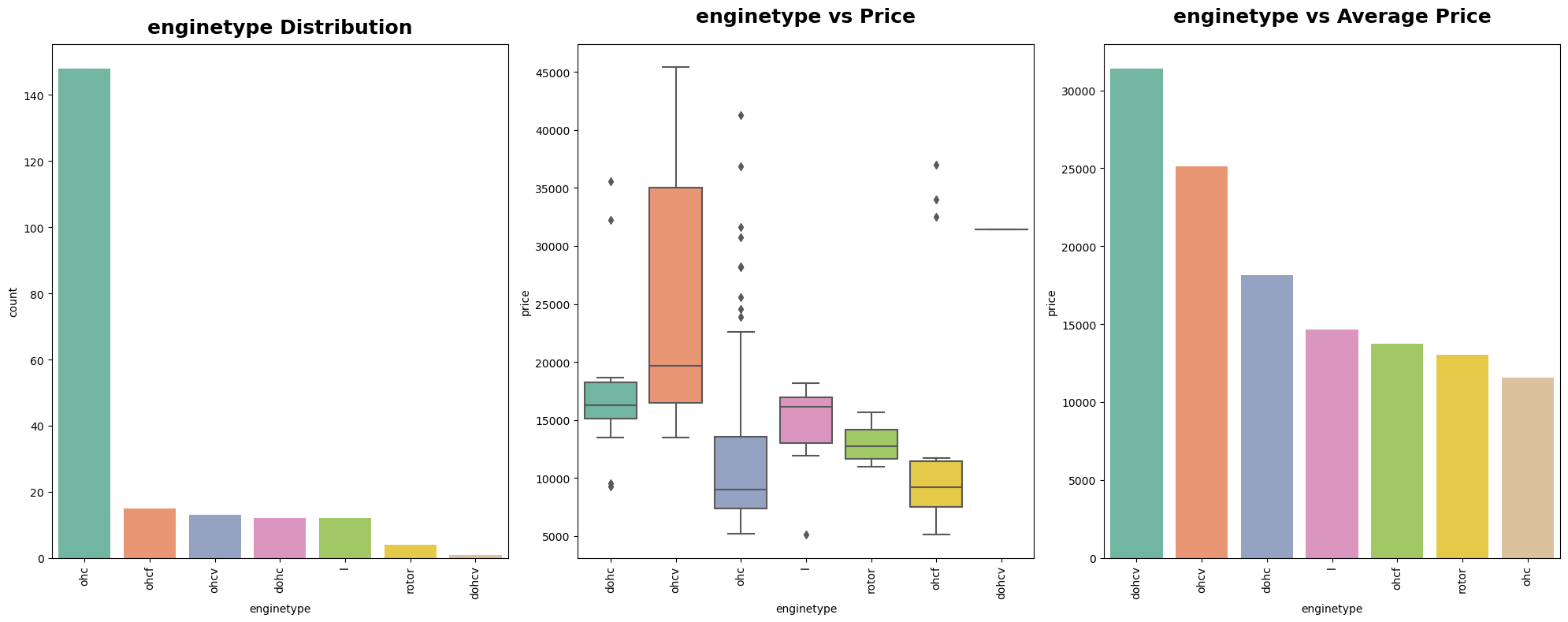


Here we find that:

* Most of the cars have engine location in front.
* There are only 3 data-points for the rear category.

So we can't make any inference of car price when compared with engine location. If we want we can drop this feature before training as it may lead to overfitting.

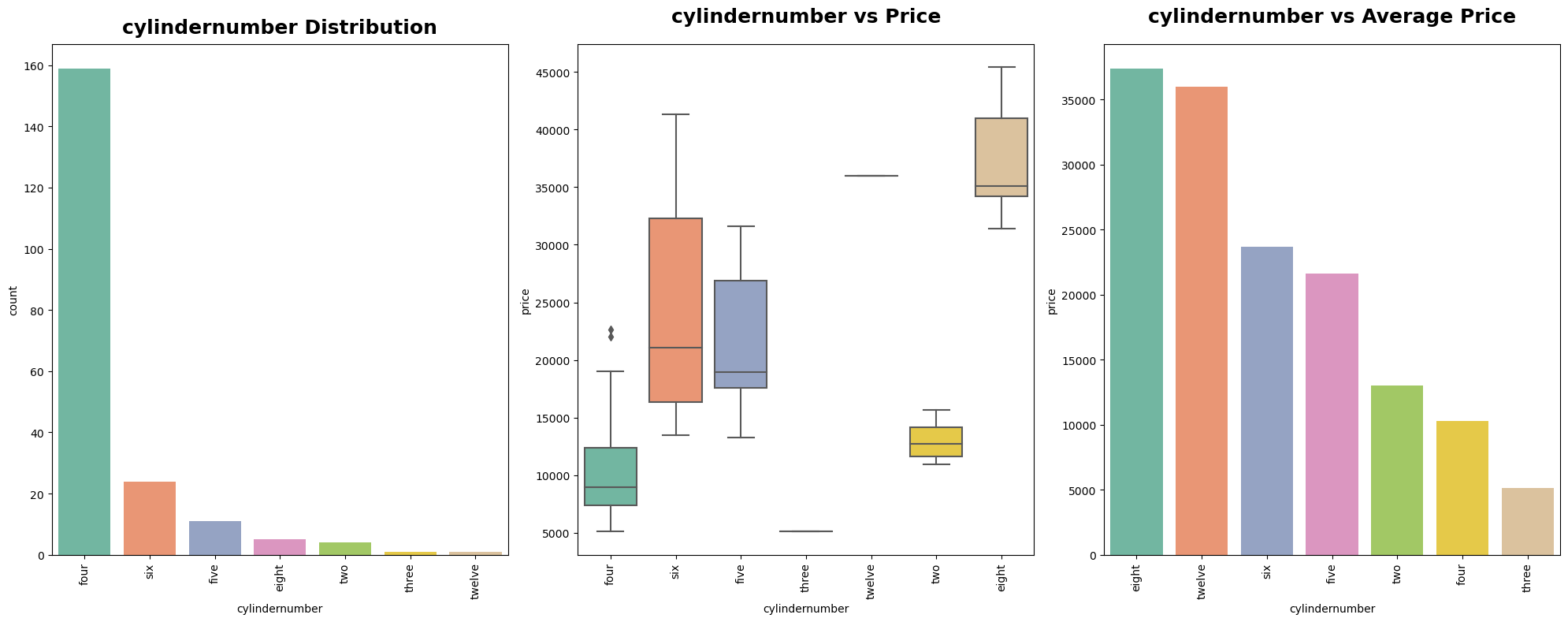
# Checking the Engine type Feature



Observations for the above plots:

* Cars having Overhead Camshaft (OHC) engines are mostly sold.
* Only one car has been sold having engine type dohcv.
* There are very few data-points of engine type dohcv & rotor. So we can say that cars having ohcv engine types are mostly expensive.
* Cars having Overhead Camshaft (OHC) engines are least expensive cars

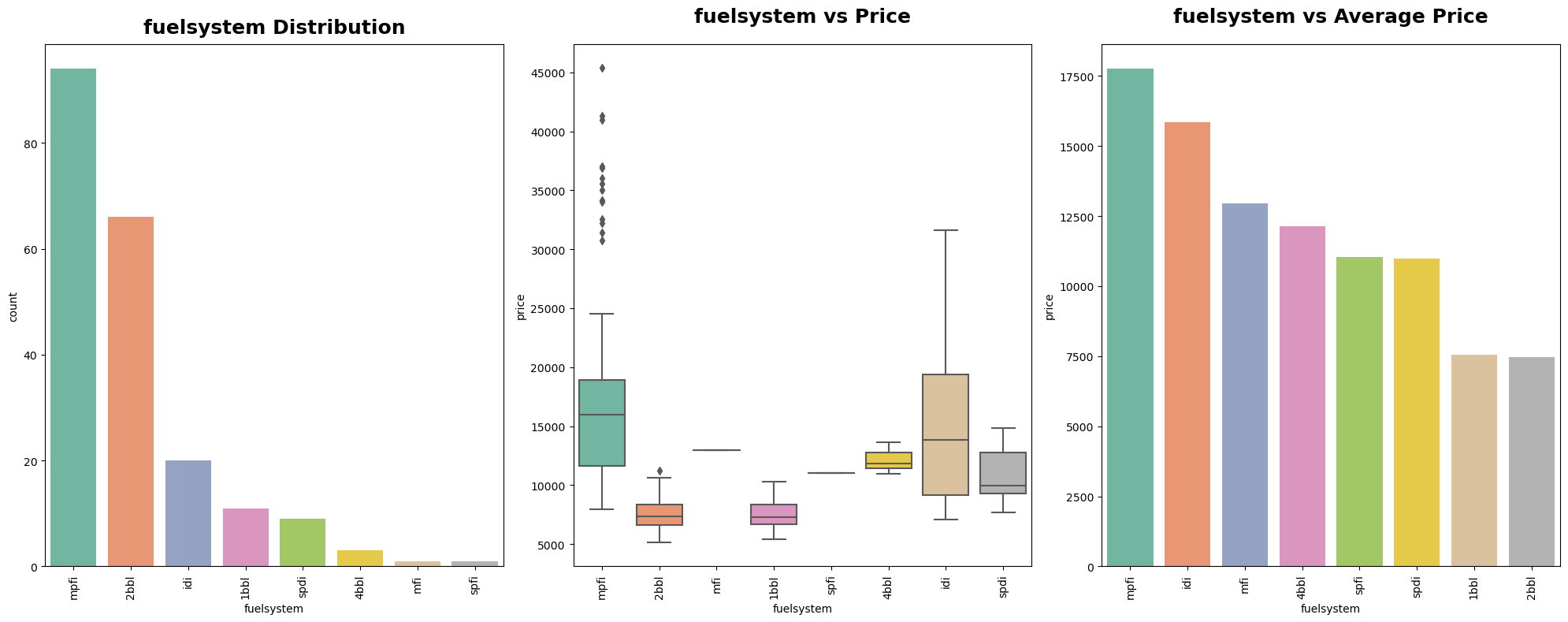
# Checking the Cylinder Number Feature



Observations for the above plots:

* Most of the cars are having Four cyclinders followed by cars having six cyclinders.
* There are only one data-point each for car having Three & Twelve data-points.
* Cars having eight cyclinders are most expensive cars followed by six cyclinders.

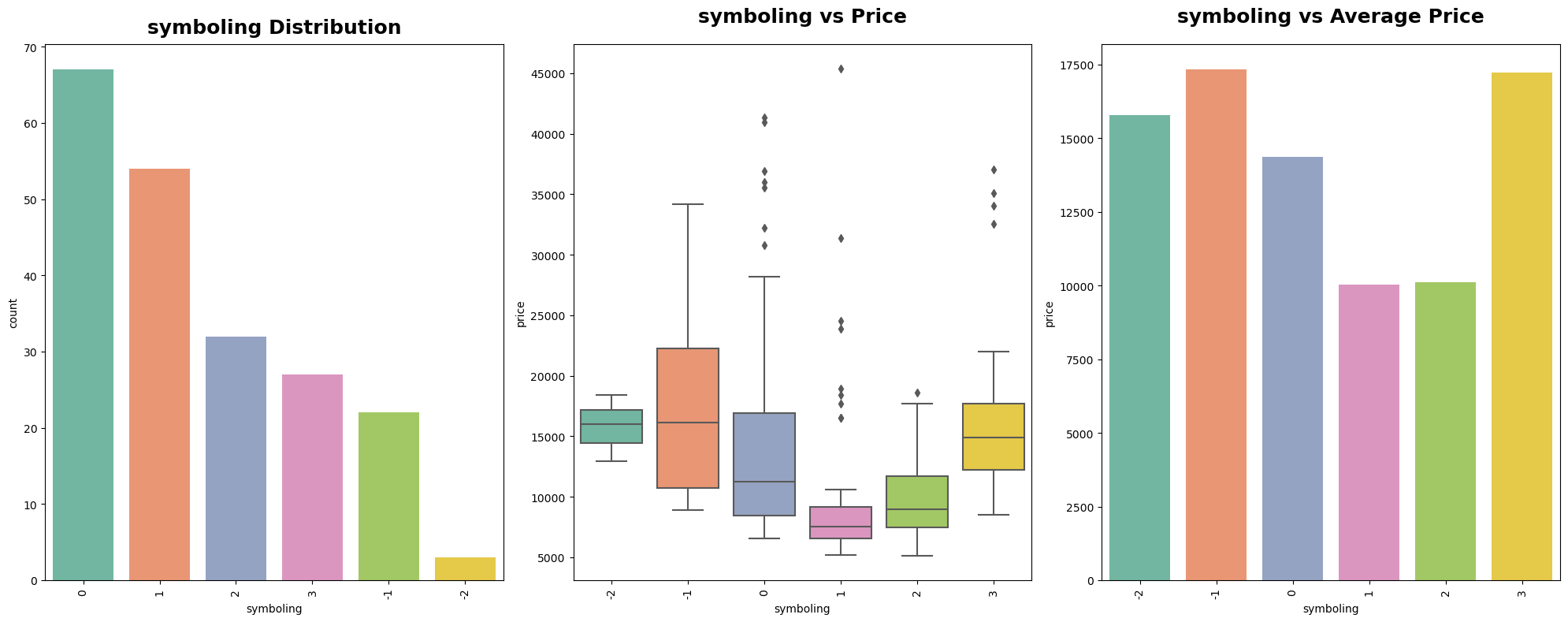
# Checking the Fuel System Feature



Looking at these graphs we can see:

* Most of the cars are having mpfi & 2bbl fuelsystems.
* Cars having mpfi fuel systems are the expensive cars followed by idi fuel system cars.
* There is only one data-point for each mfi and spfi fuelsystem cars. So we can make any further inference.

# Checking the Symboling Feature

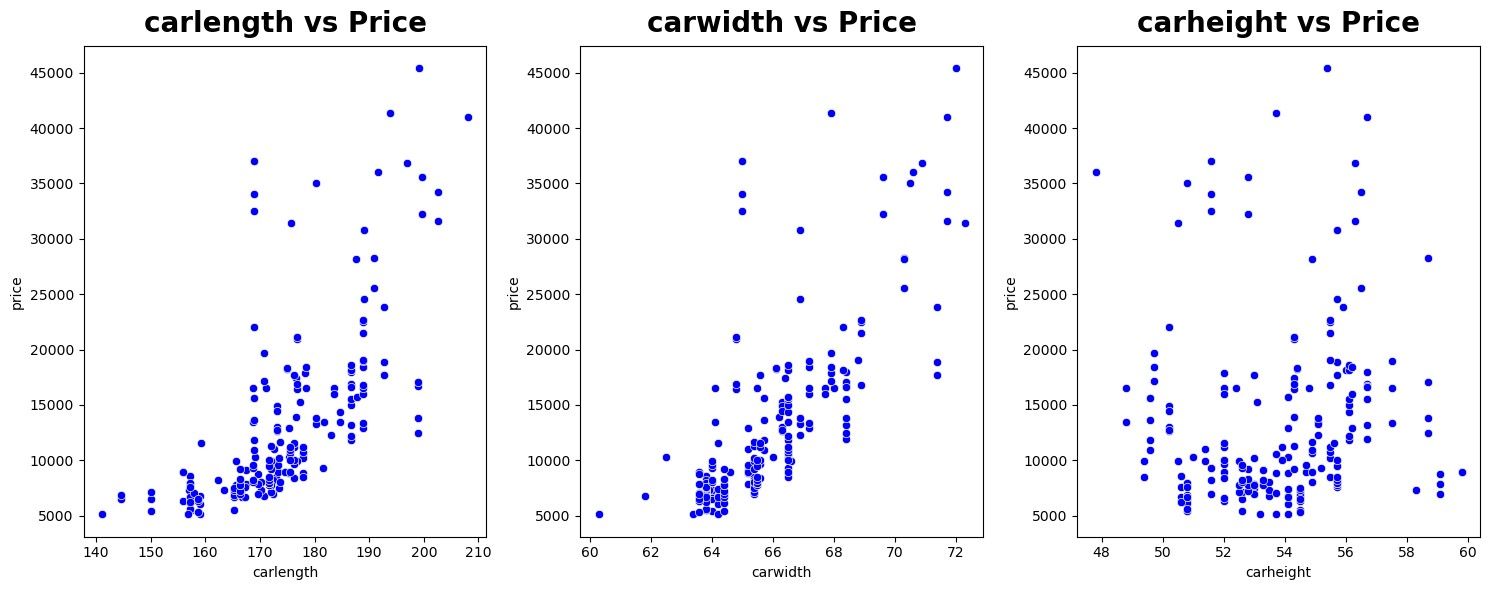


Observations for the above plots:

We can clearly observe cars having symboling 0 or 1 are mostly preferred.

We can also observe that symboling -1, 0, 3 are expensive.

In the next step, we have visualized some scatter plots whose figures and their following observations are provided below:



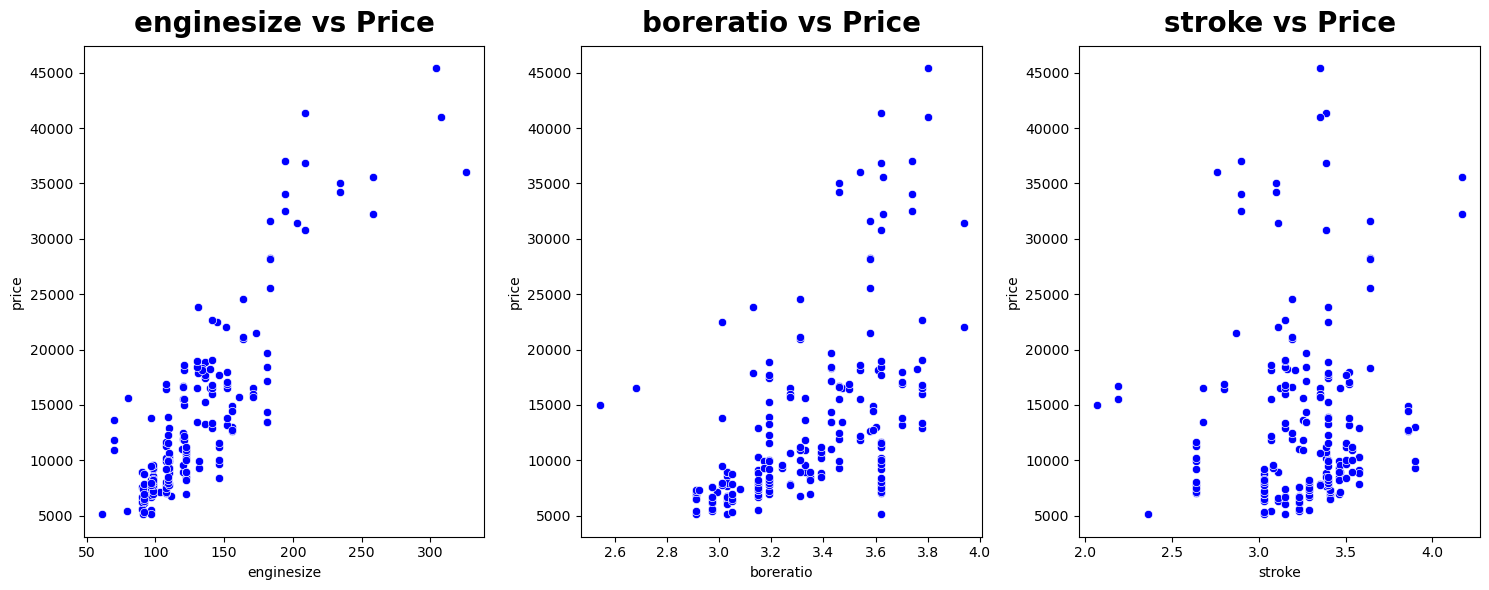
Insights

We can clearly observe that carlength & carwidth features are highly correlated with the price feature.

So we can make an insight that with increment in length & width of the car there is an increment in price too.

From carlength vs price we can't make any inference as the data-points are too scattered.

Since CarHeight is not affecting Price We can drop this feature.



### Insights

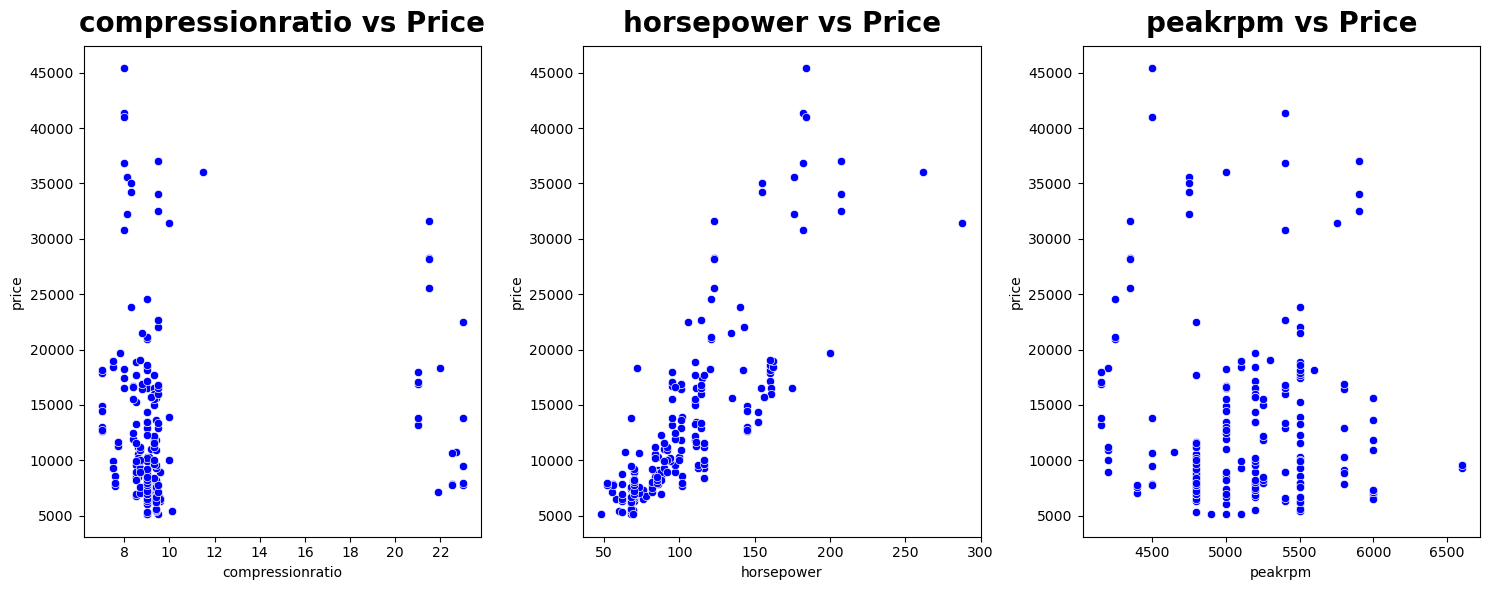
We can clearly observe that Enginesize is highly correlated with price feature.

So we can say with the increment in enginesize the price of the cars increases.

From Boreratio vs Price we can observe that the feature is not highly correlated but still there's a correlation between the features. So we can say with increment in boreratio the price of cars increases.

From Stroke vs Price we can't make any inference as the data-points are too scattered.

Since Stroke is not affecting Price that much. We can drop this feature

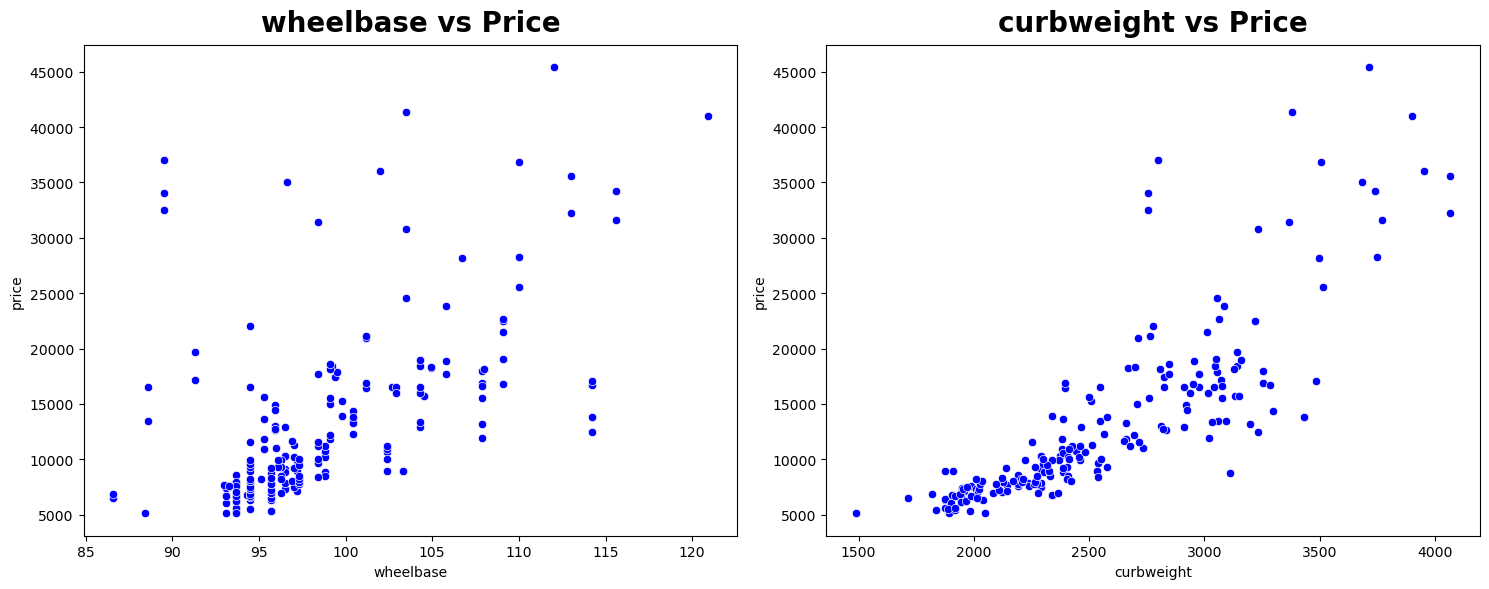


### Insights

We can clearly observe that Horsepower is highly correlated with Price. So we can say with the increment in Horsepower the price of cars also increases.

From Compressionratio vs Price & Peakrpm vs Price visuals we can't make any inference as the data-points are too scattered.

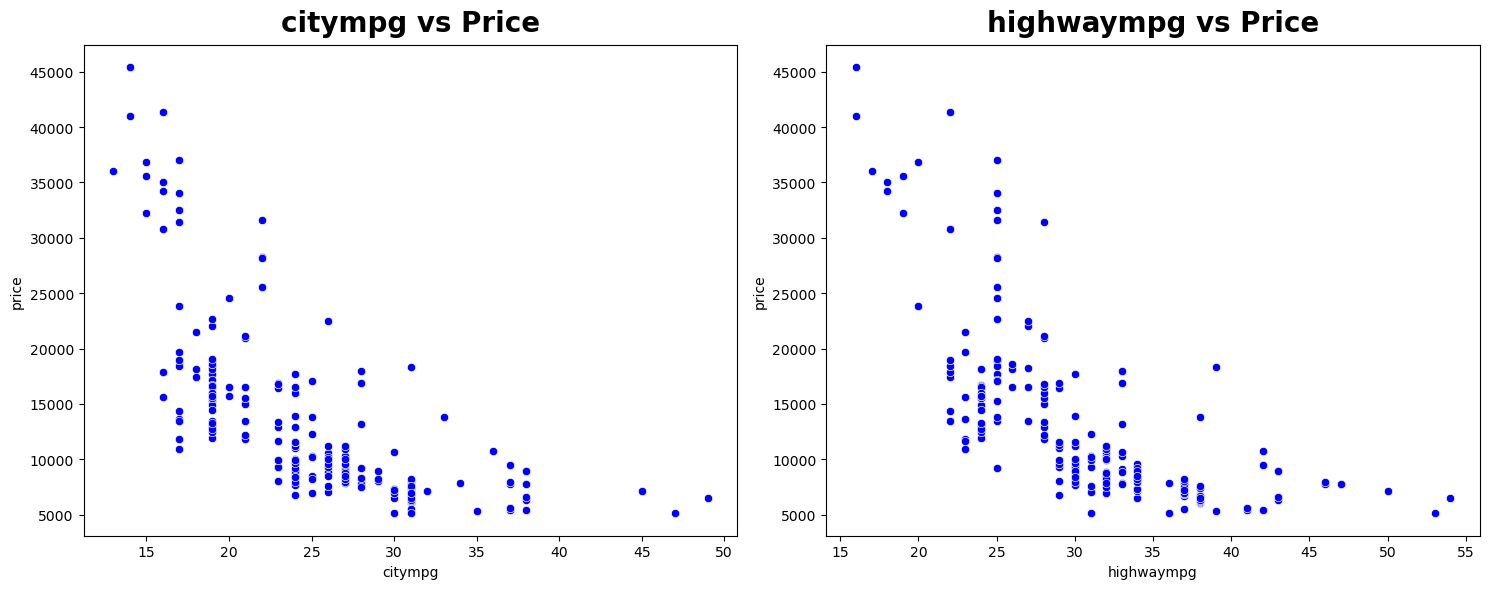
Since Compressionratio & Peakrpm is not affecting price. So we can drop this features.



### Insights

We can clearly observe that Curbweight is highly correlated with Price. So we can say with the increase in curbweight the price of cars increases.

From wheelbase vs price we can say that they are not highly correlated but still there's a correlation. So with wheel base the price of the cars also increases.



Here, we can clearly observe that Citympg & Highwaympg are having negative correlation with the price. So we can say that with the increment in citympg & Highwaympg the price of the cars decreases.

So our both Citympg & Highwaympg are useful features for price prediction.

So far, here are the useful categorical features:

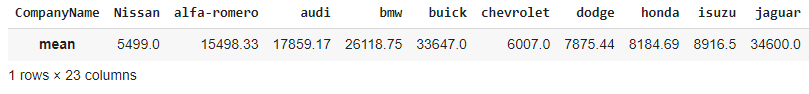
1. CompanyName
2. Fuel Type
3. Aspiration
4. Door Number
5. Car Body
6. Drive Wheel
7. Engine Type
8. Cyclinder Number
9. Fuel System

And the list of useful numerical features:

1. Wheelbase
2. Carlength
3. Carwidth
4. Curbeweight
5. Enginesize
6. Boreratio
7. Horsepower
8. citympg
9. Highwaympg
10. Price

## Feature Engineering

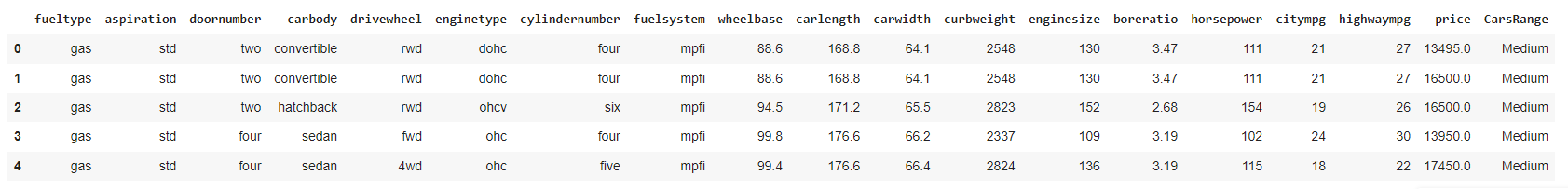
As we made an insight above that we can split the car company name into different price ranges.Like Low Range, Medium Range, High Range cars.



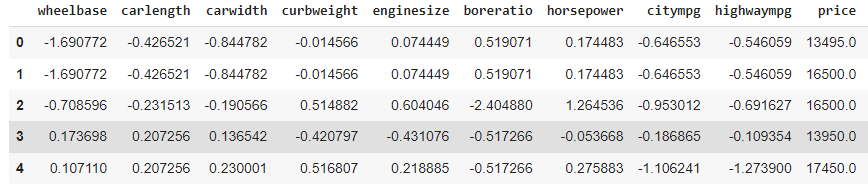
In above output we fetched the avergae price of each individual car companies

Now we have to add this average values a new column in our datset.

## Data Pre Processing

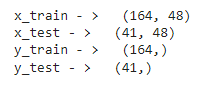


After creating a new dataframe with the useful features, we perform StandardScaling on all of the features.



## Training the Model

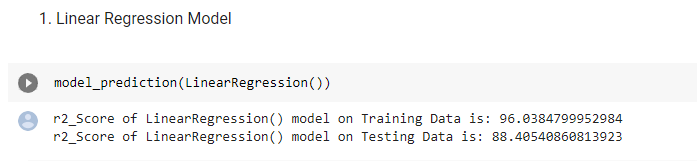
After all of this was done, we were ready to start training our model. We performed a train\_test\_split with 20% of the dataset as test set.



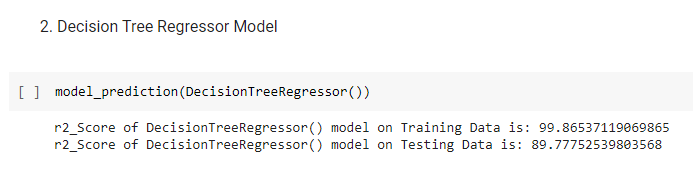
We trained the model on three different models. We used Linear Regression, Random Forest Regressor and Decision Tree Regressor.

Here are the scores for each of the models.

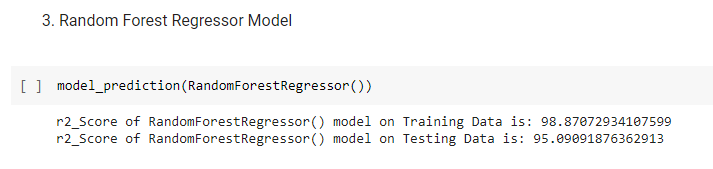
### Linear Regression Model



### Decision Tree Regressor Model



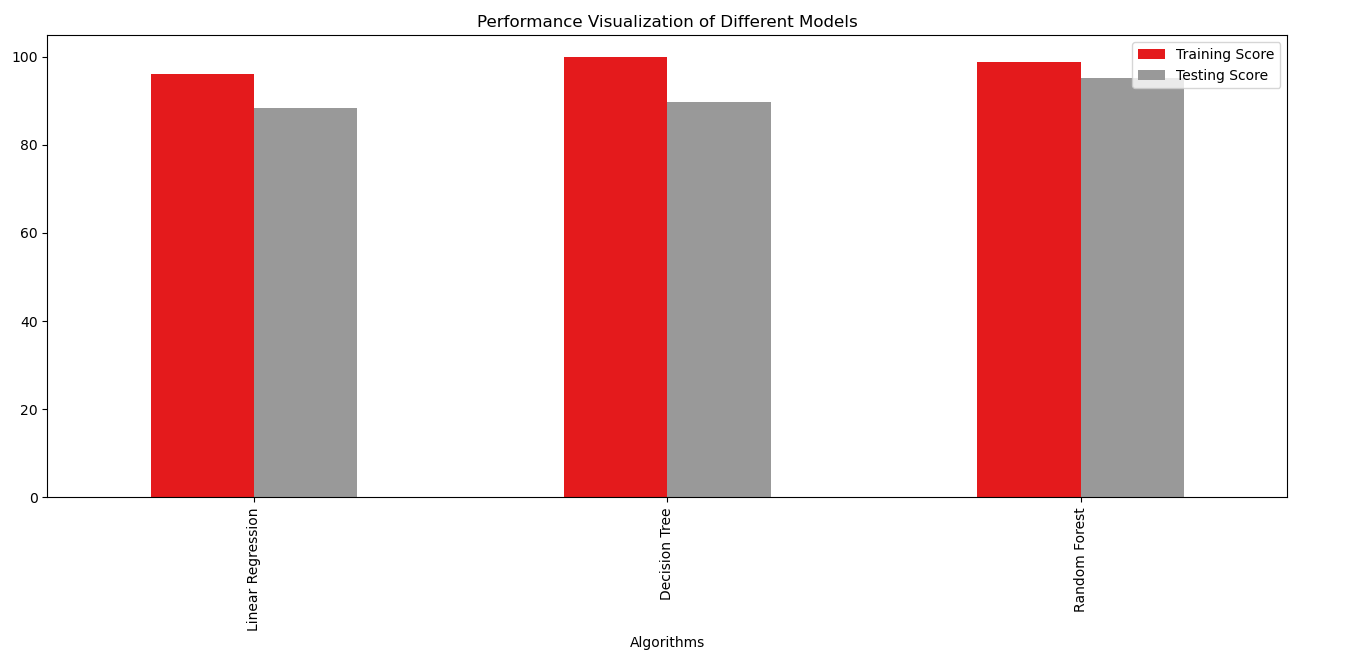
### Random Forest Regressor Model



To make things easier, we compared the three models altogether.



Here’s a visual representation of the scores in a bar graph.



## Conclusion

From the above discussion and comparison, we can see that Random Forest Regressor had the best performance with an accuracy of 95%. So, if we want to solve the regression problem of predicting the price of cars, then we can use the Random Forest Regressor Model.