AUTOMOBILE PRICE PREDICTION

Data Analysis project work

Name: Debanjana Kundu

Christ (Deemed To Be University)

Master of Science Data Science (2022-24)

CONTENTS

1 INTRODUCTION

Data science advances in predictive modeling revolutionize industries like the automotive sector, enabling informed decision-making, optimal pricing strategies, efficient inventory management, and customer satisfaction. This comprehensive dataset encompasses a diverse range of attributes associated with various automobile models, aiding in the creation of robust predictive models. Containing a wealth of information, the dataset covers crucial aspects such as car specifications, technical features, manufacturer details, and market performance indicators. By delving into this dataset, data scientists can embark on a journey to develop and fine-tune machine learning algorithms that predict automobile prices with a high degree of accuracy. In this dataset, you'll find an opportunity to engage with real-world automotive data, enabling you to apply regression techniques and explore feature importance.

2 ORGANIZATION OF THE PROJECT

2.1 The problem

The data in hand is based on 26 attributes and characteristics of automobiles. It has about 205 data points for the analysis. The project focuses on two main problems. We take the price of the automobile as our response variable. The first one is to check every assumption of the multiple linear regression model. The second one is to predict the price using a machine learning multiple linear regression model.

2.2 Objectives

The project has two main objectives. The first one is that normalize and transform data to ensure linear regression assumptions are met. Address heteroscedasticity, minimize multicollinearity, and assess autocorrelation using the Durbin-Watson test to maintain model stability and accuracy. Eliminate correlated variables and ensure model reliability through transformation techniques.

The second one is that develop a robust linear regression model for predicting automobile prices based on a diverse set of attributes. The goal is to create a predictive model that accurately estimates the prices of automobiles, aiding automotive manufacturers, dealerships, and stakeholders in making informed pricing decisions.

3 THE DATA

- **Name:** Automobile
- **Source:** Kaggle
- **Source Link:** https://www.kaggle.com/datasets/shivam2503/diamonds
- **Description:** This classic dataset contains the prices and other attributes of almost 205 automobiles.
- **Response Variable:** Price: Price in New York price
- **Predictors :**
    1. Symboling: Symboling indicates riskiness beyond the price, assigned to vehicles. A value of +3 indicates that the auto is risky, and -3 that it is probably pretty safe.
    2. Normalized losses: Insurance risk assessment
    3. Make: Car company names
    4. Fuel-type: Fuel type of the cars (gas and diesel)
    5. Aspiration: Type of air intake system used in the engine of a vehicle (std and turbo)
    6. Num-of-doors: Number of doors of vehicle (two and four)
    7. Body style: distinctive design and layout of a vehicle's exterior
    8. Drive-wheels: Contains information about the number of driven wheels and whether they are the front wheels, rear wheels, or all wheels (4WD or AWD)
    9. Engine location: Where the engine is situated (front and rear)
    10. Wheel-base: Distance between the centers of the front and rear wheels. A longer wheelbase often results in better ride quality
    11. Length: The length of a vehicle refers to the distance from its front bumper to its rear bumper.
    12. Width: The width of a vehicle is the measurement across its widest point, usually from one side mirror to the other.
    13. Height: Vehicle height refers to the vertical measurement from the ground to the highest point of the vehicle's roof
    14. Curb-weight: Curb width refers to the width of the vehicle from one side to the other
    15. Engine-type: The configuration or layout of the engine
    16. Num-of-cylinders: How many cylinders are present in the engine of each vehicle
    17. Engine-size: Volume of engine size
    18. Fuel-system: The fuel system of an automobile is responsible for delivering fuel from the fuel tank to the engine
    19. Bore: The bore refers to the diameter of the cylinders
    20. Stroke: The stroke of an engine refers to the distance that the piston travels inside the cylinder
    21. Compression-ratio: The compression ratio is a measure of the engine's efficiency and power output
    22. Horsepower: Horsepower (hp) is a unit of measurement for the engine's power output

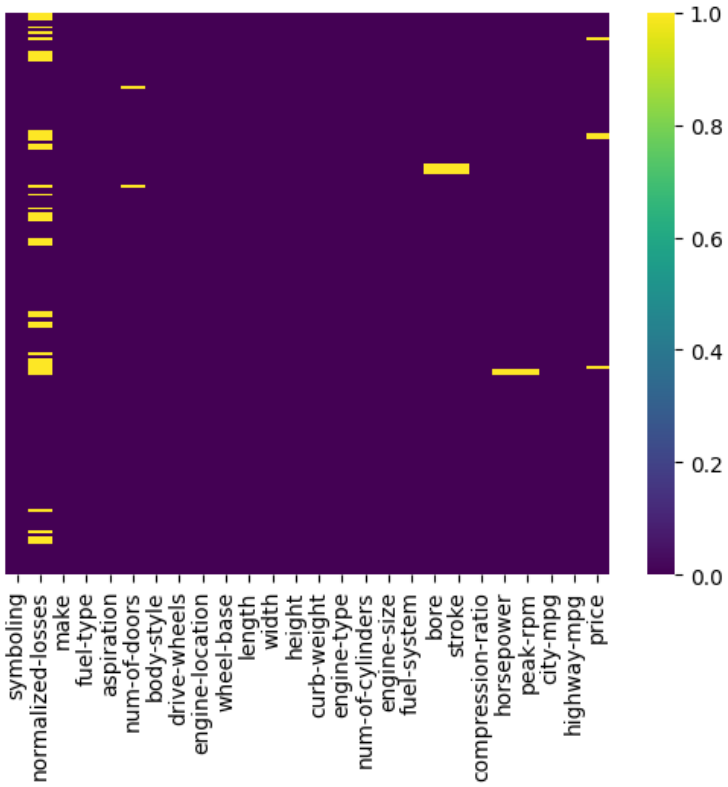
23. Peak-rpm: Peak RPM (Revolutions Per Minute) is the engine speed at which the maximum horsepower is generated
24. City-mpg: City MPG (Miles Per Gallon) indicates how many miles the vehicle can travel per gallon of fuel in city settings.
25. Highway-mpg: Highway MPG represents how many miles the vehicle can travel per gallon of fuel under highway conditions.
26. Price: Car pricing details

4 DATA PRE-PROCESSING

Data preprocessing involves cleaning, transforming, and organizing raw data into a format that is suitable for analysis or for training machine learning models. Proper data processing improves the quality of the data and better analysis.

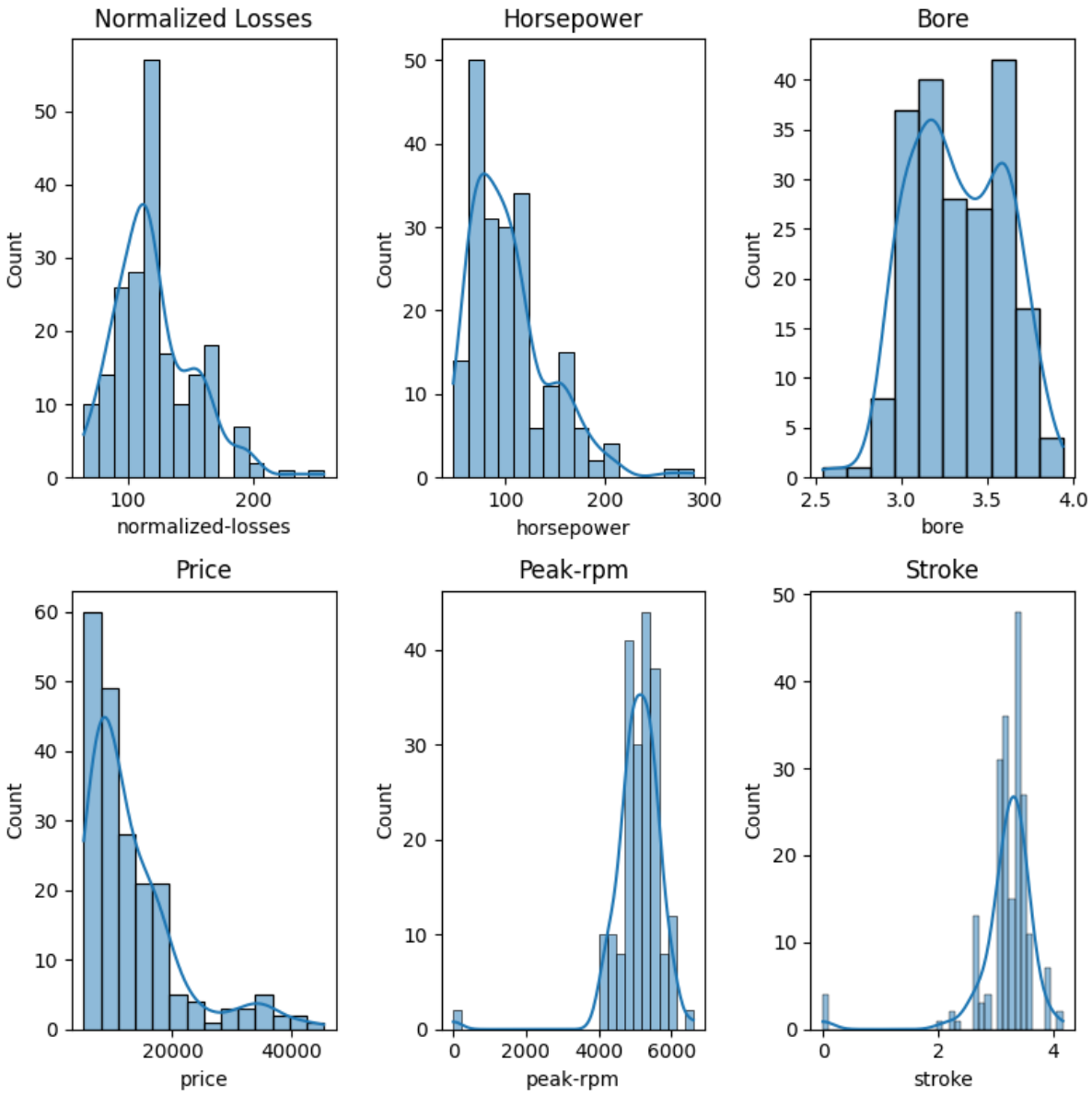
Here are some techniques involved in data processing:

1. **Handling missing value:** after replacing the '?' with nan value find the missing value columns that are['normalized-losses, 'num-of-doors', 'bore', 'stroke', 'horsepower', 'peak-rpm', 'price']



[Fig: 1]

- The above missing value heatmap [Fig:1] shows the missingness occurs completely randomly, and there is no relationship between the missing values and the observed data.
- Fill the categorical variable by mode because it helps maintain the overall distribution and preserves the dominant category, minimizing disruption in the dataset's categorical structure.
- Fill with Mean imputation when the missing values are numerical and the distribution of the variable is approximately normal. And Median imputation is preferred when the distribution is skewed, as the median is less sensitive to outliers than the mean.



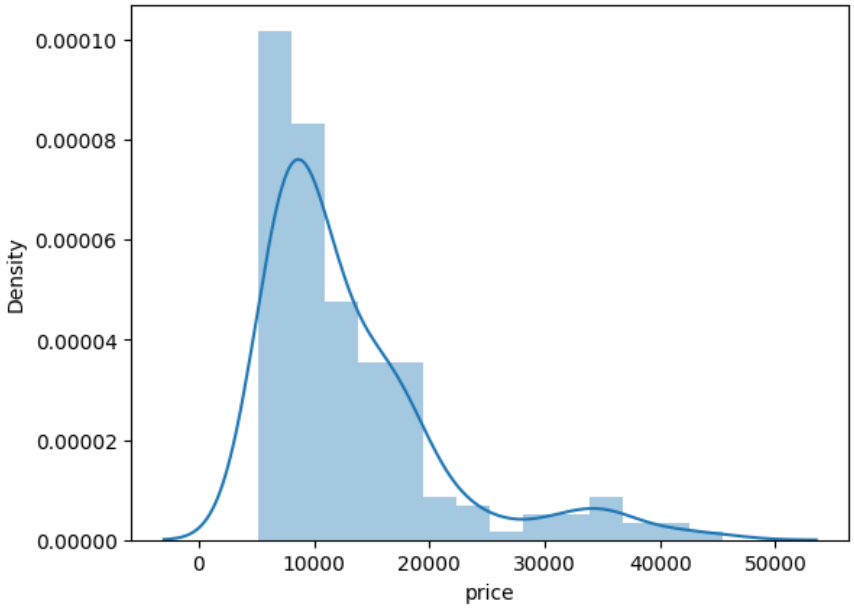
[Fig: 2]

- ['normalized-losses, 'bore', 'stroke', 'horsepower', 'peak-rpm', 'price'] are skewed so replace the missing value by median technique, and 'num-of-doors' is a categorical feature so it replaces by mode technique.

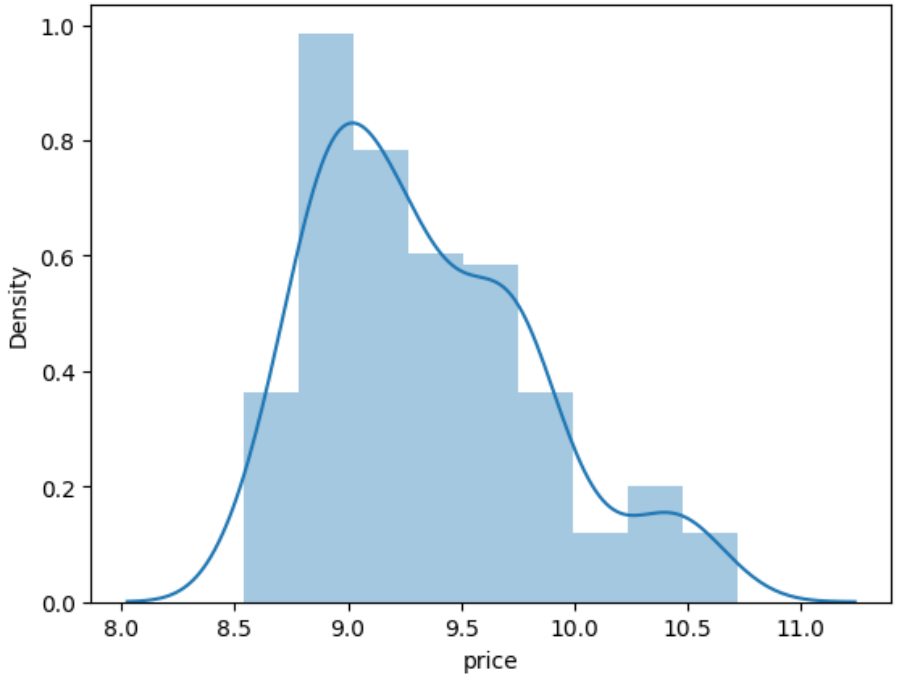
2. **Feature-scaling**: Scale numerical features to a similar range, here used technique is StandardScaler.
3. **Encoding categorical variables**: Convert categorical variables into numerical representations. ['make', 'fuel-type', 'aspiration', 'num-of-doors', 'body-style', 'drive-wheels', 'engine-location', 'engine-type', 'num-of-cylinders', 'fuel-system'] are the categorical variables using LableEncoder technique.

5 EXPLORATORY DATA ANALYSIS

5.1 Nature of response variable



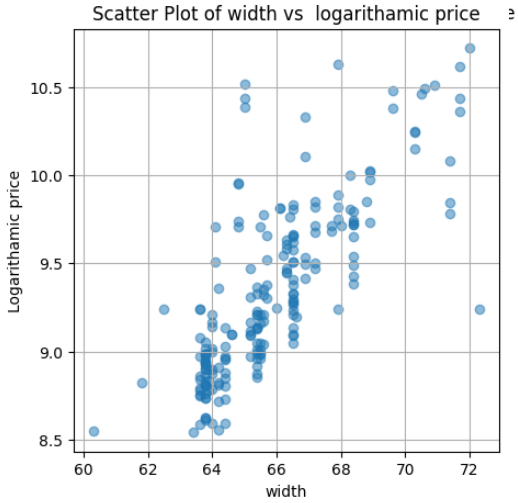
The response variable for the given data is Price. The histogram of the variable price shows that it is positively skewed. So, if we use the data for the purpose of regression, we cannot do multiple linear regression using the method of least squares as it violates the assumption of normality. So, we do a log transformation of the response.

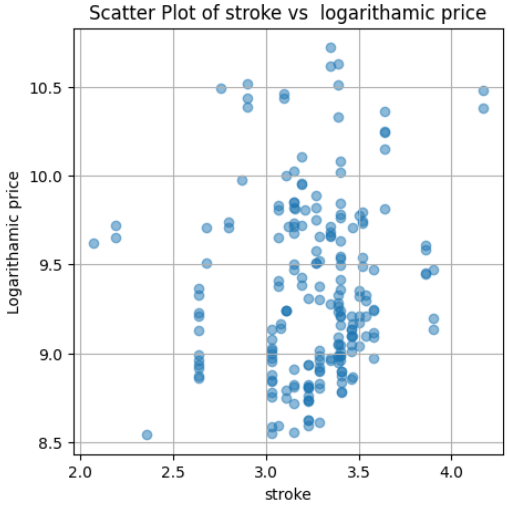


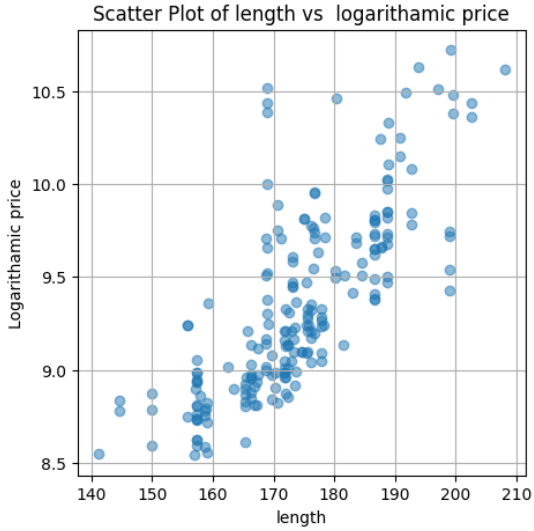
We observe that on the log transformation of the price variable, the histogram is no more positively skewed. It had become almost symmetric roughly. So, applying the method of least squares may not lead to much deviation in accuracy.

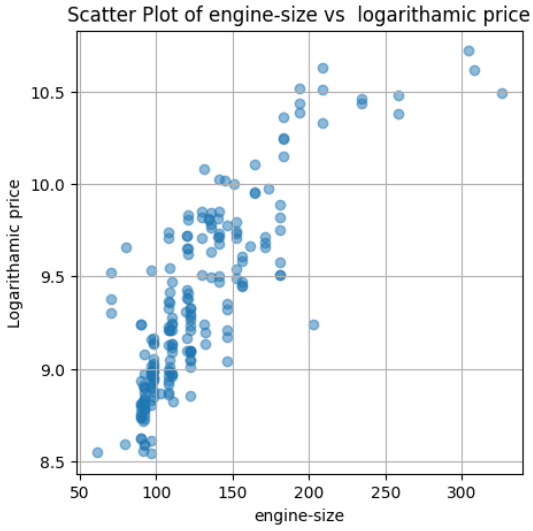
5.2 Plot of Logarithmic of Price vs Continuous Predictors

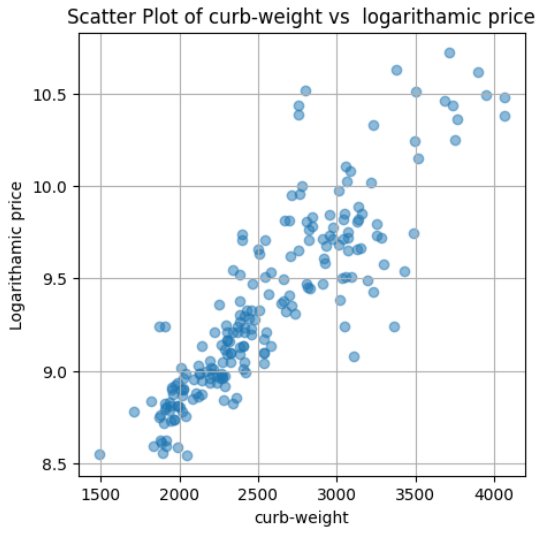
We wish to study whether there is any relationship between the predictors and response variables. Also, how the logarithmic price of the automobile gets influenced by the predictors.

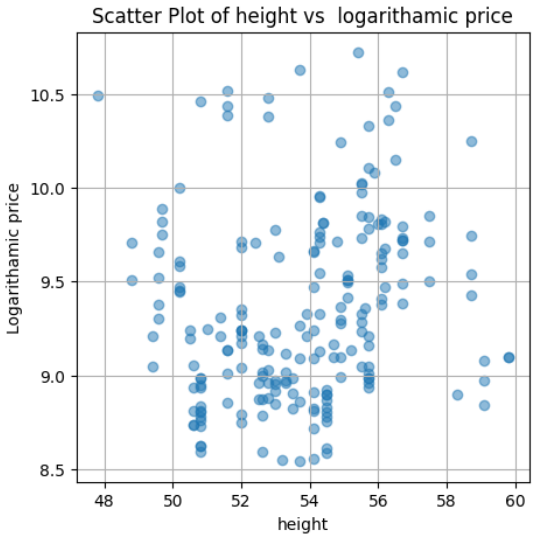


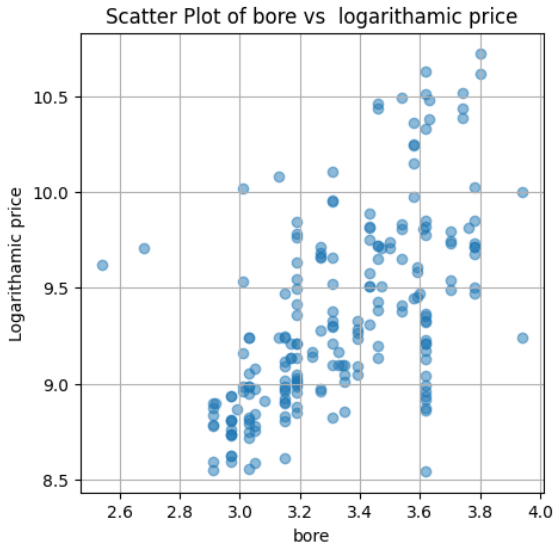


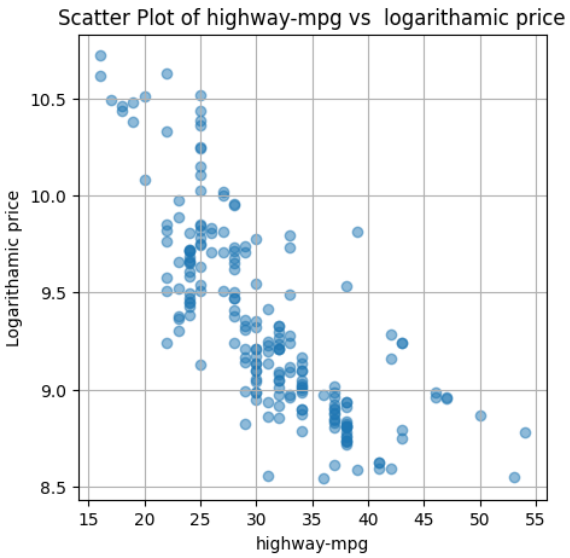


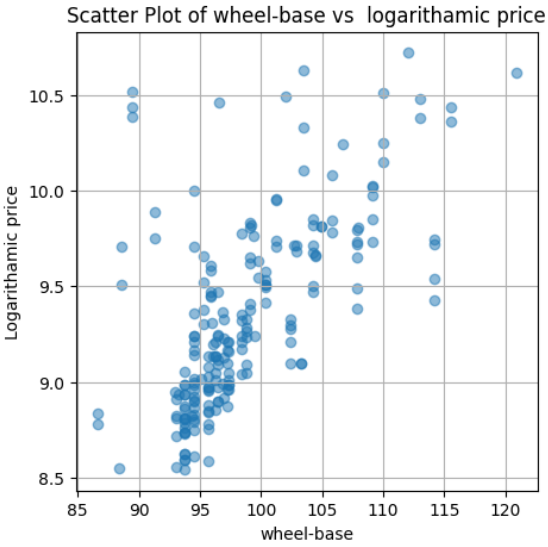


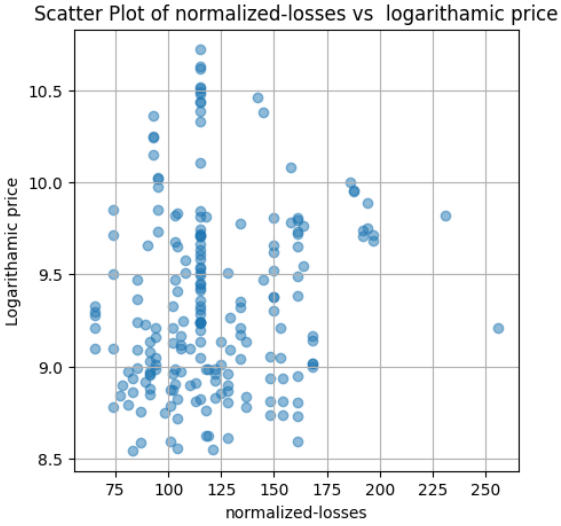


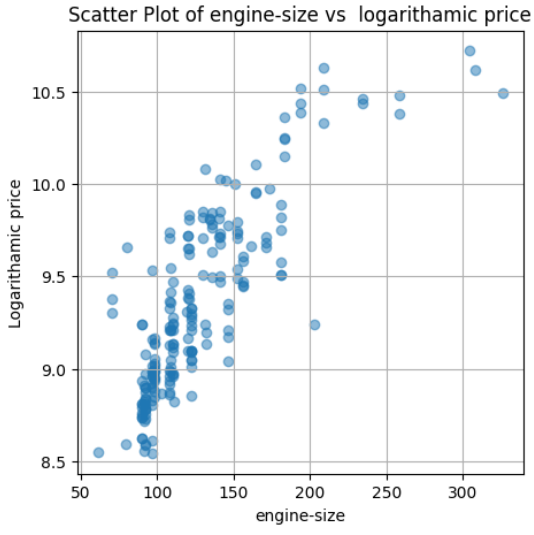


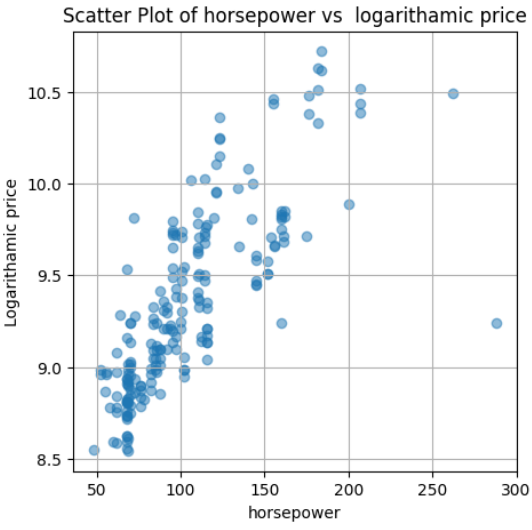

Scatter Plot of bore vs logarithamic price
Logarithamic price
bore
10.5
10.0
9.5
9.0
8.5
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0

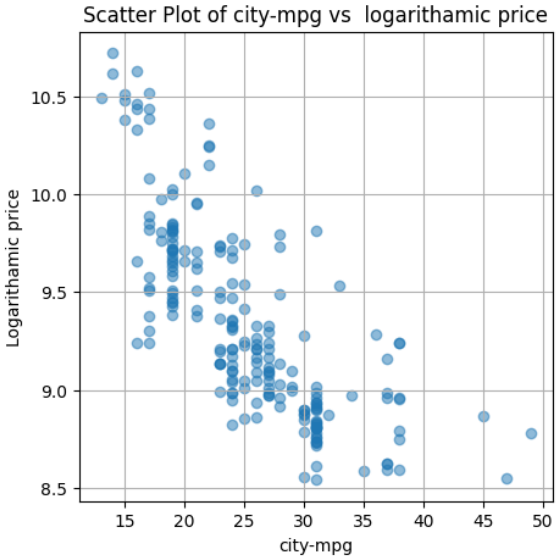

Scatter Plot of highway-mpg vs logarithamic price
Logarithamic price
highway-mpg
10.5
10.0
9.5
9.0
8.5
15
20
25
30
35
40
45
50
55

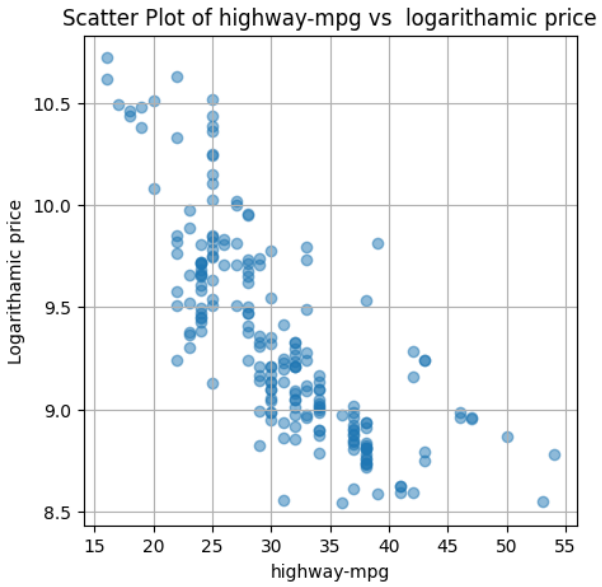

Scatter Plot of wheel-base vs logarithamic price
Logarithamic price
wheel-base
10.5
10.0
9.5
9.0
8.5
85
90
95
100
105
110
115
120


Scatter Plot of normalized-losses vs logarithamic price
Logarithamic price
normalized-losses
10.5
10.0
9.5
9.0
8.5
75
100
125
150
175
200
225
250

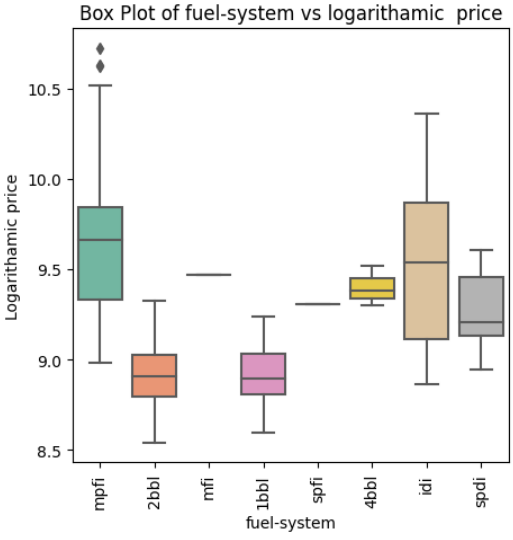

Scatter Plot of engine-size vs logarithamic price
Logarithamic price
engine-size
10.5
10.0
9.5
9.0
8.5
50
100
150
200
250
300

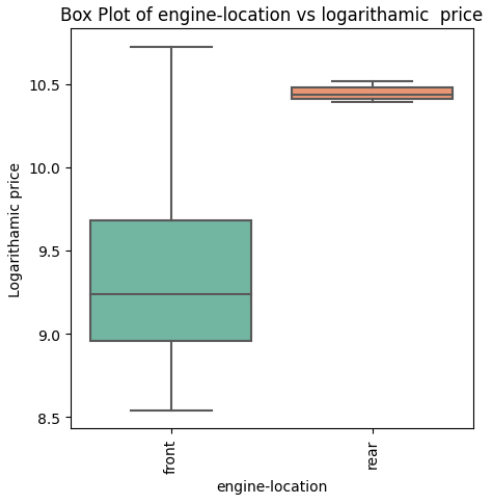

Scatter Plot of horsepower vs logarithamic price
Logarithamic price
horsepower
10.5
10.0
9.5
9.0
8.5
50
100
150
200
250
300

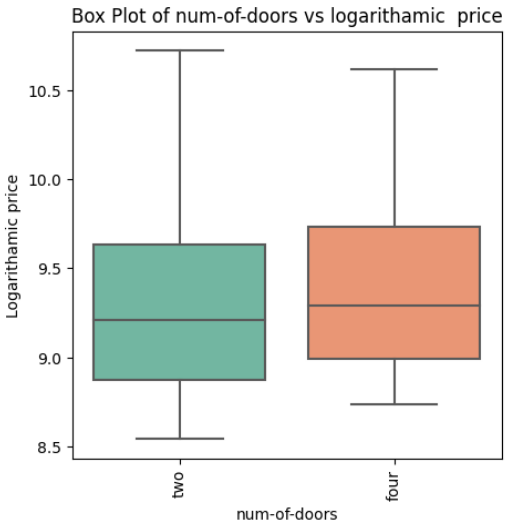


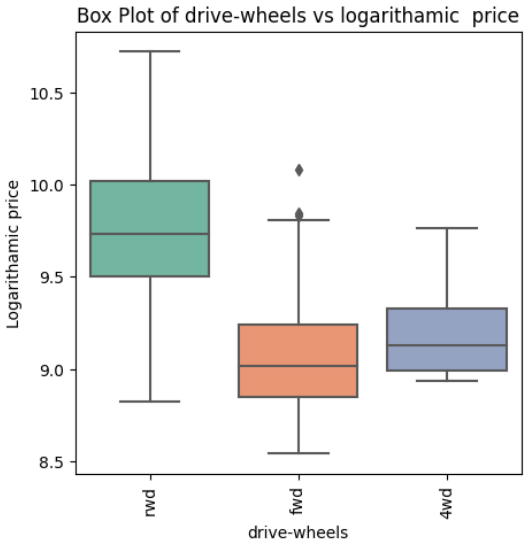


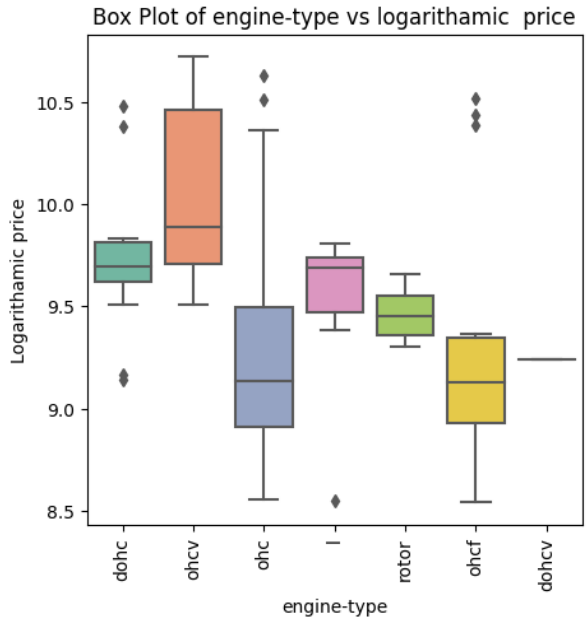
5.3 Plot of Logarithmic of Price vs Categorical Predictors

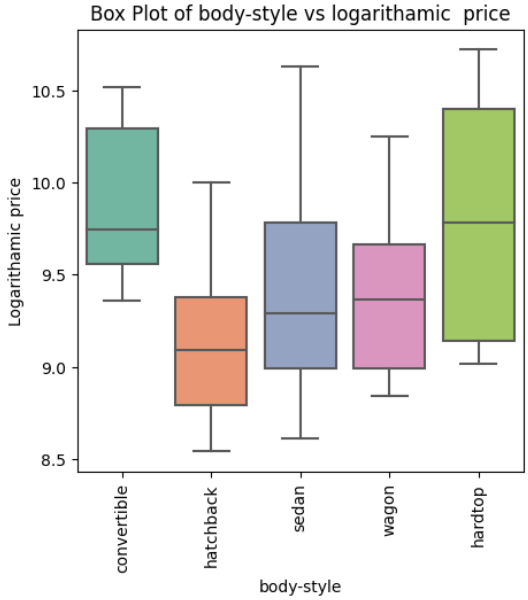


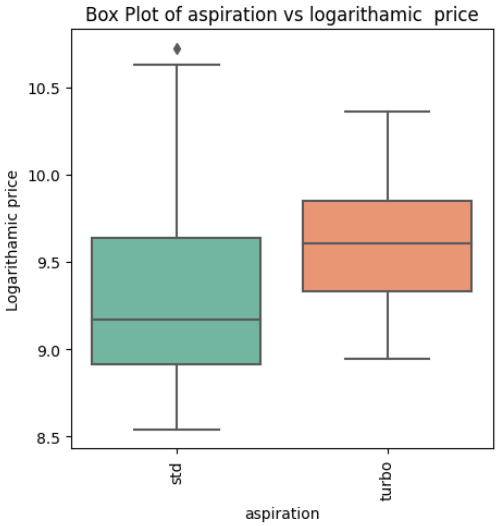


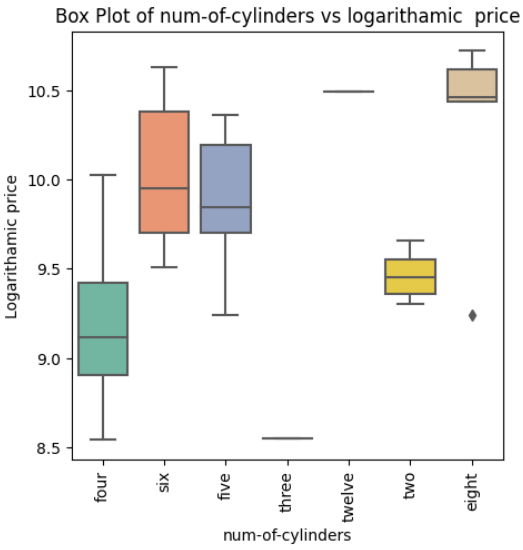

Box Plot of num-of-doors vs logarithamic price
Logarithamic price
10.5
10.0
9.5
9.0
8.5
two
four
num-of-doors

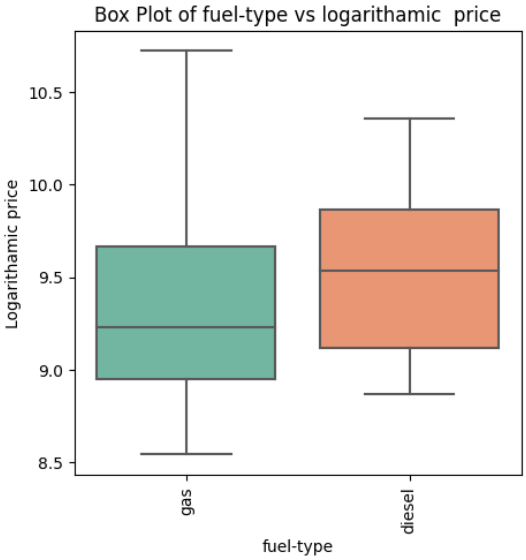

Box Plot of drive-wheels vs logarithamic price
Logarithamic price
10.5
10.0
9.5
9.0
8.5
rwd
fwd
4wd
drive-wheels

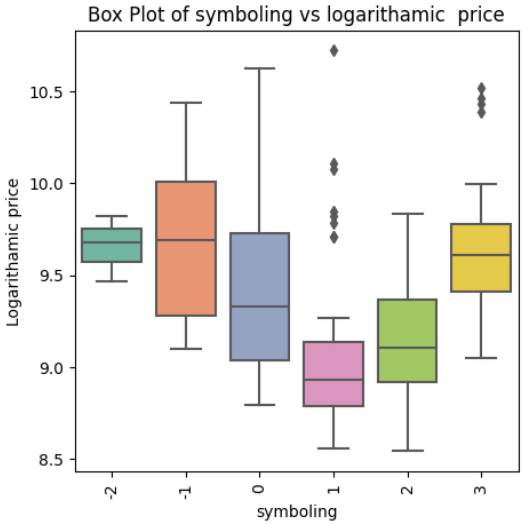

Box Plot of engine-type vs logarithamic price
Logarithamic price
10.5
10.0
9.5
9.0
8.5
dohc
ohcv
ohc
l
rotor
ohcf
dohcv
engine-type


Box Plot of body-style vs logarithamic price
Logarithamic price
10.5
10.0
9.5
9.0
8.5
convertible
hatchback
sedan
wagon
hardtop
body-style


Box Plot of aspiration vs logarithamic price
Logarithamic price
10.5
10.0
9.5
9.0
8.5
std
turbo
aspiration


Box Plot of num-of-cylinders vs logarithamic price
Logarithamic price
10.5
10.0
9.5
9.0
8.5
four
six
five
three
twelve
two
eight
num-of-cylinders


Box Plot of fuel-type vs logarithamic price
Logarithamic price
10.5
10.0
9.5
9.0
8.5
gas
diesel
fuel-type


Box Plot of symboling vs logarithamic price
Logarithamic price
10.5
10.0
9.5
9.0
8.5
-2
-1
0
1
2
3
symboling

6 MULTIPLE LINEAR REGRESSION

Multiple Linear Regression (MLR) is a statistical technique used to model the relationship between two or more independent variables (predictors) and a dependent variable (response). MLR makes certain assumptions to ensure the validity and reliability of the regression analysis. Violation of these assumptions can lead to unreliable results and inaccurate interpretations.

6.1 Assumption

6.1.1 Linearity

This means that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables. Note that this assumption is much less restrictive than it may at first seem. Because the predictor variables are treated as fixed values.

6.1.2 Normality

In the previous section [5.1] we discuss the normality assumption of multiple linear regression. The normality of predictors is crucial in linear regression because violations can affect the reliability of statistical tests, confidence intervals, and coefficient estimates. Non-normal predictors can lead to biased parameter estimates, impacting the model's accuracy and interpretability. Normality assumptions ensure that the sampling distribution of estimates is well-behaved, allowing valid hypothesis testing and confident inferences. Departures from normality might distort significance levels and affect the overall performance and generalization of the regression model.

6.1.3 Multicollinearity

The independent variables should not be highly correlated with each other. High multicollinearity can lead to difficulty in distinguishing the individual effects of variables on the dependent variable. This makes it challenging to interpret the individual effects of predictors on the response variable. Multicollinearity inflates standard errors, reducing the statistical significance of variables. So we have to remove the multicollinearity from the dataset.

One method to detect Multicollinearity is to calculate the **variance inflation factor (VIF)** for each independent variable

The formula for VIF is:

$$VIF = \frac{1}{1-R^2}$$

Where R2= R^2 value is determined to find out how well an independent variable is described by the other independent variables.

VIF equal to 1 = variables are not correlated

VIF between 1 and 5 = variables are moderately correlated

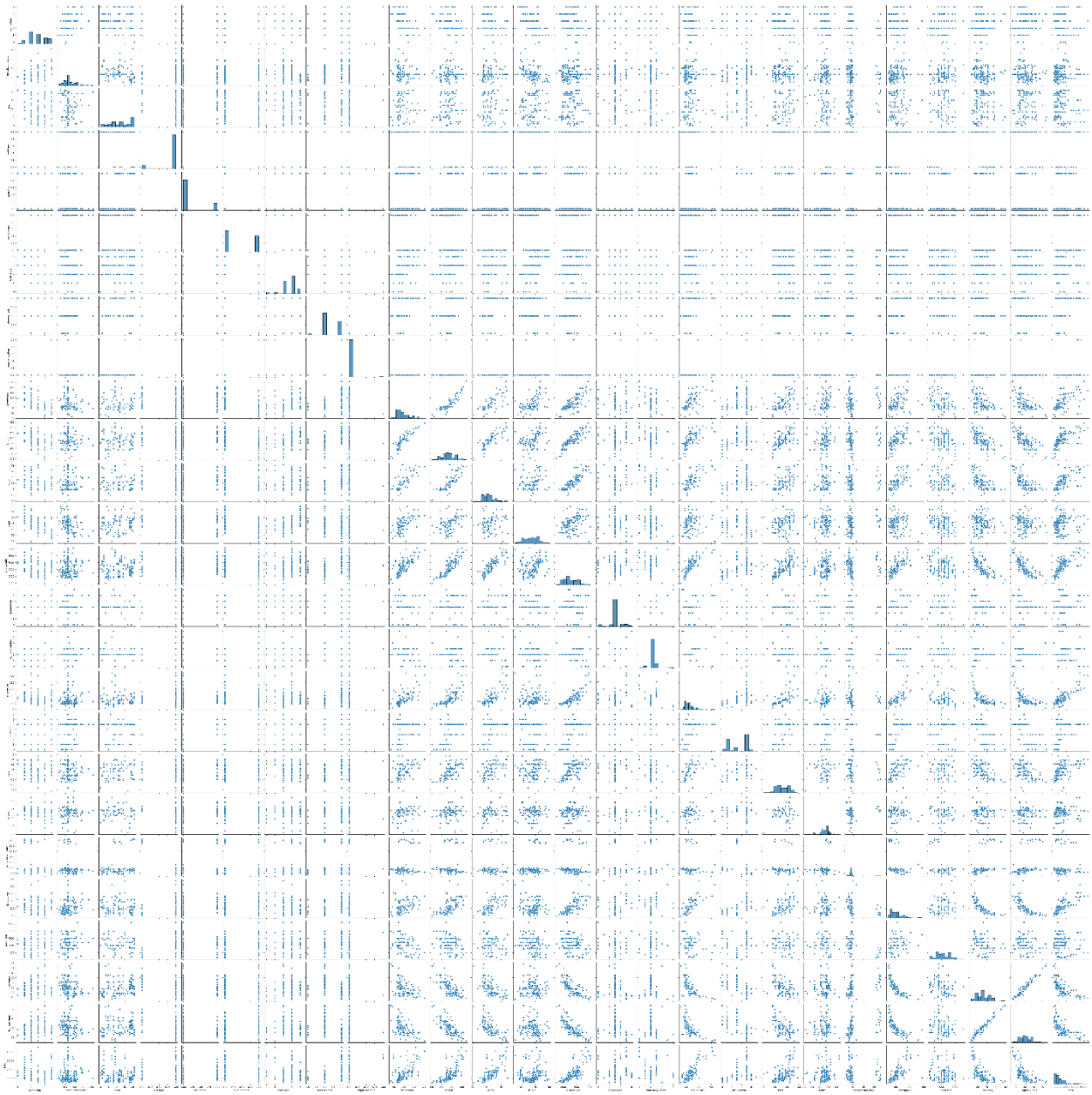
VIF greater than 5 = variables are highly correlated.

	Variable	VIF
0	const	8924.082806
1	symboling	3.391008
2	normalized-losses	1.772721
3	make	1.520081
4	fuel-type	95.078728
5	aspiration	3.508524
6	num-of-doors	2.885799
7	body-style	2.845079
8	drive-wheels	2.764357
9	engine-location	1.674938
10	wheel-base	11.695828
11	length	11.815238
12	width	9.101447
13	height	3.223810
14	curb-weight	18.937729
15	engine-type	1.978516
16	num-of-cylinders	2.501097
17	engine-size	16.291847
18	fuel-system	2.745078
19	bore	2.544093
20	stroke	3.360551
21	compression-ratio	89.160624
22	horsepower	16.551151
23	peak-rpm	1.743734
24	city-mpg	34.309698
25	highway-mpg	29.861781

['fuel-type','wheel-base','length','curb-weight','engine-size','compression-ratio','horsepower','city-mpg','highway-mpg','width'] attributes are highly correlated which is

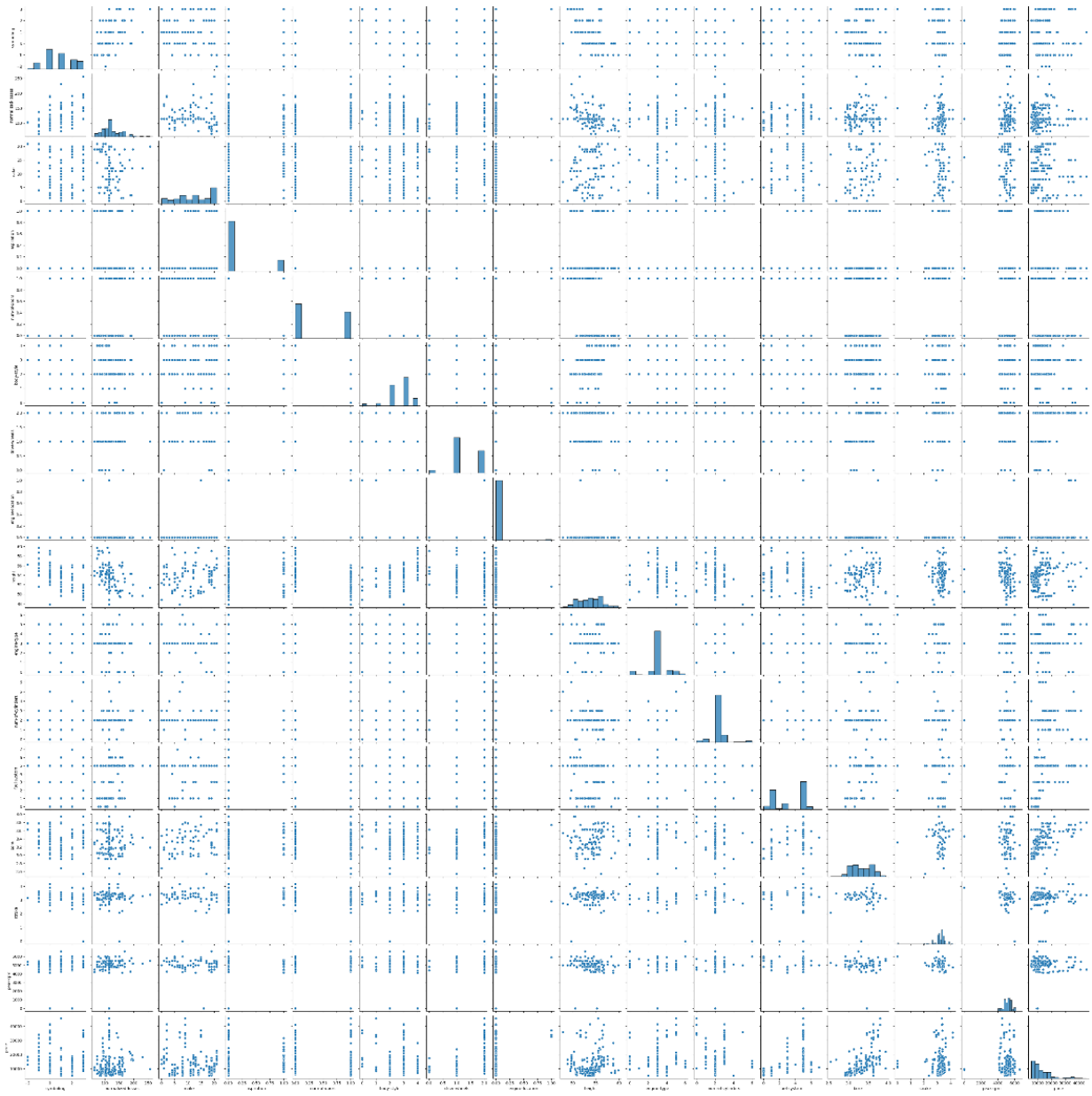
the contradiction of Multiple linear regression. So, for analysis purposes, we remove the variable from the dataset.

Pairplot before removing the attributes:



This visual represents shows there are so many predictors which are highly correlated. So need to remove that attributes because multicollinearity case sensitive in linear regression.

Pairplot after removing the attributes:



This visual represents shows there are no highly correlated attributes present.

6.1.4 Autocorrelation

Autocorrelation sensitivity in linear regression arises from the dependence of response (error) terms on past observations. When autocorrelation is present, the assumption of independent errors is violated, impacting the model's reliability. Detecting and addressing autocorrelation is vital to maintain the integrity of regression results and ensure accurate predictions.

The **Durbin-Watson statistic** is a test statistic to detect autocorrelation in the response (residuals) from a regression analysis.

The assumption of the test is responses are normally with a mean 0.

The hypotheses followed for the Durbin-Watson statistic:

H(0) = First-order autocorrelation does not exist.

H(1) = First-order autocorrelation exists.

The formula is:

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Where, e_t response figure, T is the number of observations.

The Durban-Watson statistic will always assume a value between 0 and 4.

DW = 2 indicates that there is no autocorrelation.

DW <2, indicates a positive autocorrelation

DW>2, indicates negative autocorrelation.

In this analysis got the Durbin-waston test statistics near 1.8 which is approximately no autocorrelations. This is because the response is approximately normally distributed so that got an approx autocorrelation.

6.2 MODEL

Upon initially verifying the fundamental assumptions of linear regression, we proceed with the underlying assumption of homoscedasticity, where we presume that the variance of errors remains uniform across all levels of predictor variables.

A rigorous assessment is conducted after constructing and fitting the regression model to ensure its robustness and practical application integrity.

6.2.1 Test Set and Train Set:

For multiple linear regression, we split the data into a train set and a test set. The train set contains 0.7 proportion of the total observations and the train set contains 0.3 proportion of the total observations. We first regress the price on other covariates in the train set and observe the residual plot. Then we use the obtained equation in the test set and observe how much deviation in the price of the diamond occurs.

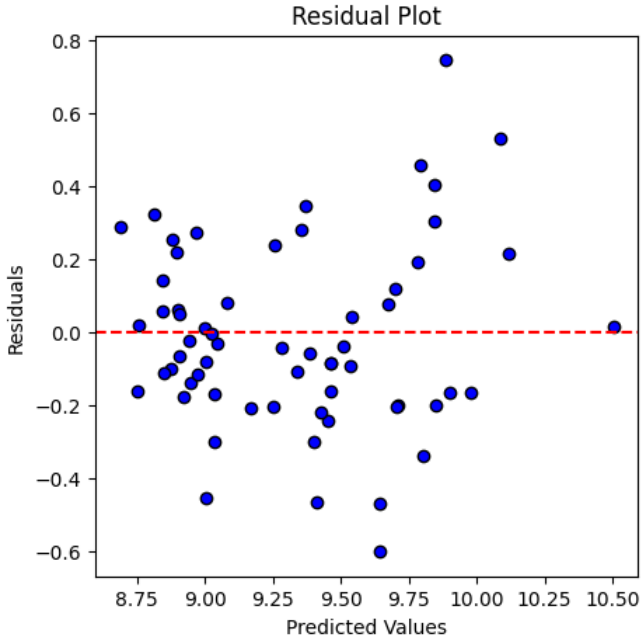
6.2.2 Regression Coefficients and Intercepts:

Here presents regression model coefficients and intercepts in a data frame, sorted by magnitudes in descending order, aiding in understanding features' impact on target variables in regression analysis.

	Features	Coefficients	Intercept
7	engine-location	0.703162	5.70905
12	bore	0.378310	5.70905
6	drive-wheels	0.224502	5.70905
11	fuel-system	0.105162	5.70905
3	aspiration	0.086937	5.70905
10	num-of-cylinders	0.045676	5.70905
9	engine-type	0.041603	5.70905
8	height	0.040601	5.70905
13	stroke	0.022611	5.70905
1	normalized-losses	0.001281	5.70905
14	peak-rpm	-0.000061	5.70905
2	make	-0.019486	5.70905
0	symboling	-0.021392	5.70905
5	body-style	-0.090975	5.70905
4	num-of-doors	-0.150929	5.70905

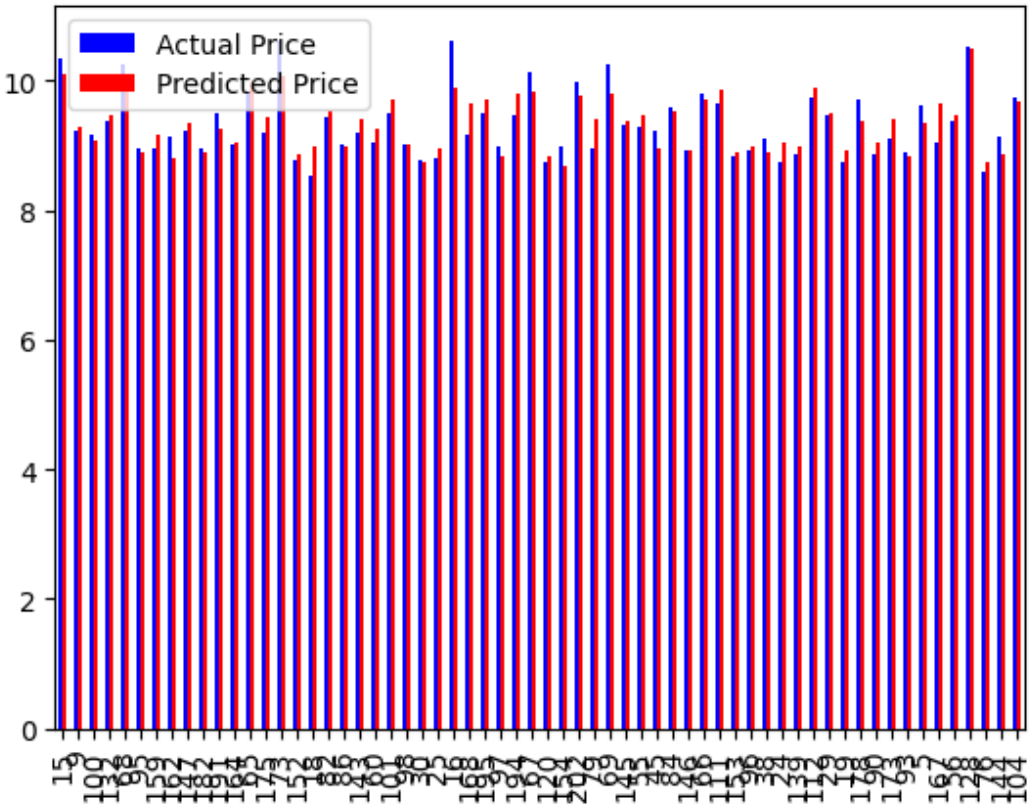
We observe that ‘normalized-losses’, and ‘peak rpm’ are less significant to predict the price.

6.2.3 Residual Plot



In this residual plot, the points are scattered randomly around the residual=0 line. We can conclude that a linear model is appropriate for modeling this data.

6.2.4 Fitting the model in the test set



We observe that the first plot shows each sample's actual (by red) values vs predicted (by blue) difference, and we say that there is less difference.

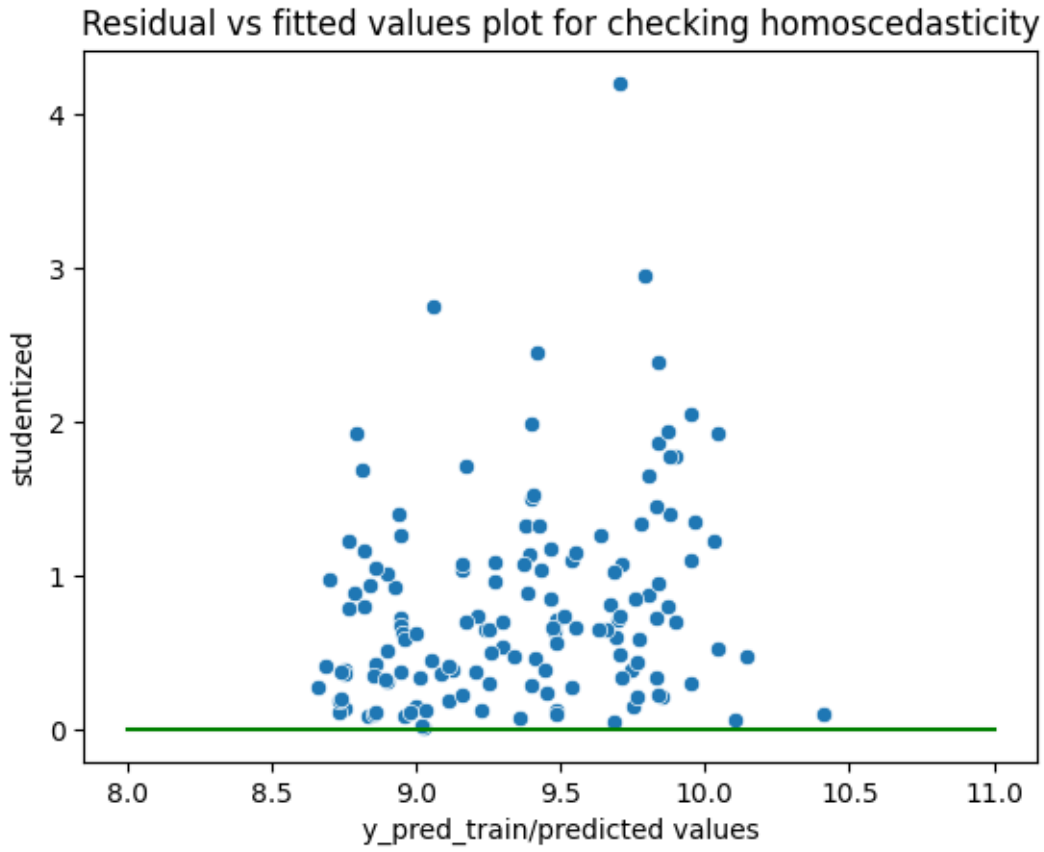
The second plot shows the overall difference between actual target values and predicted values.

6.1.5 Homoscedasticity

Homoscedasticity is sensitive in linear regression because it assumes that the variance of errors remains constant across all levels of predictor variables. When violated, with heteroscedasticity present, the model's assumptions are compromised.

After the model fitting, we show that there is no heteroscedasticity between

Spread Level Curve is a smoothed or fitted line that helps visualize the relationship between the residuals and the predicted values. In the context of homoscedasticity, the spread level curve should ideally be a horizontal line with no discernible pattern.



We observe that the spread level curve is relatively flat and shows no clear pattern, it suggests homoscedasticity, indicating that the assumption of constant variance is met.

7 RESULT

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

Where, y_i is the observed variable value

$\hat{y}$ is the value estimated by the regression line

The value of the RSS is 3.9422876034485173 which is a regression model that is doing a better job of fitting the data. It means that the predicted values are closer to the actual values, and there is less unexplained variance in the data.

Evaluation metric:

```
Multiple regression  report:
Mean Absolute Error: 0.198222450013167
Mean Squared Error: 0.06358528392658899
Root Mean Squared Error: 0.2521612260570387
R-squared: 74.92677 %
```

Observations:

- An R-squared of 0.749 suggests a relatively good fit of the model to the data. It indicates 75% of the variability is accounted for by the regression model. The remaining 25% of the variability is not captured by the model.
- The provided metrics indicate that the model is performing reasonably well. The MAE, MSE, and RMSE values are all relatively low, which suggests that the model is making predictions that are close to the actual values.

8 CONCLUSIONS

In conclusion, the developed regression model exhibits promising predictive potential, as reflected by low Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) values. However, the comparatively low R-squared on both train and test datasets suggests room for further improvement. Given the possibility of outliers impacting the model's performance, applying robust regression techniques such as Lasso or Ridge regression could enhance accuracy and address potential overfitting. Continued refinement using these methods can yield a more robust and reliable model for accurate predictions and insightful analysis.