**PROJECT 2 REPORT**

(Arun Jiju Joseph Id:1002140653)

Introduction:

The task in the assignment is to follow the instructions in order to clean the data, draw insights, shape the data, and achieve a conclusion. The assignment consists of two sections:

1. Data from our lives
2. Data manipulation/Exploratory Data Analysis
3. Multiple regression Analysis

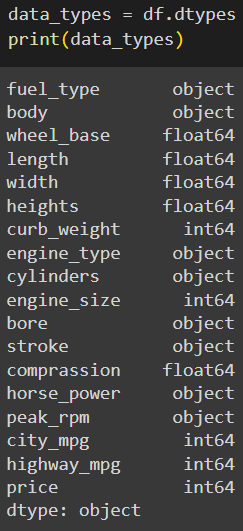
**Part One: Data from our lives:**

Described a real-world problem that uses the regression model and five predictors used in it is also defined. The example which I used was based on car racing, that is, Formula 1 race is going to happen, and wondering who is going to win the race, a regression model can be built to predict that. Some relevant predictors for building the model are Previous race results, driver skill, Weather conditions, Race strategies, Condition of the vehicle and so on.

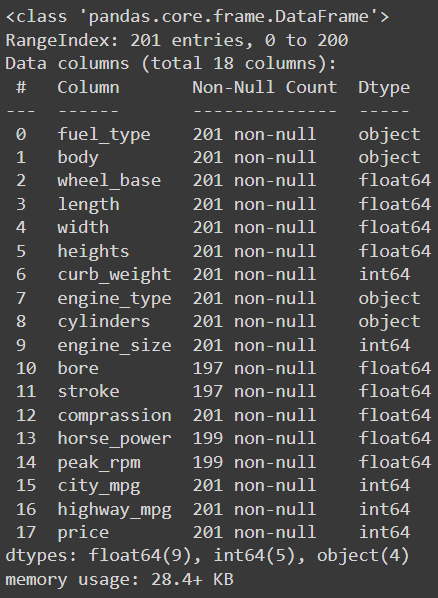
**Part Two:Data manipulation/Exploratory Data Analysis:**

The main objective of this step is to clean the data in order to get it ready for analysis. The dataset we're using for Part two is taken from the 1985 Auto Imports Dataset. This data must be processed because it is raw before the analysis can be done. The null values, duplicate entries, etc. in the data may affect the final results. We clean and organise the data in order to make it understandable so that we can draw any inferences from it. Data cleaning is the process of removing unnecessary or undesirable data from a dataset. Data wrangling is the process of getting this clean data into a format that can be read and used for analysis.

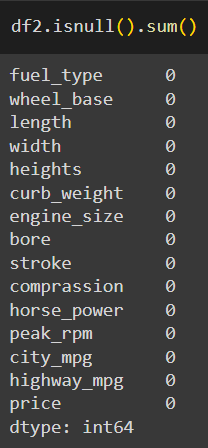
We are using the Auto Imports Dataset, which is a CSV file. We'll outline the procedure we'll use to clean this in the manner that follows. The first step is to read the csv file and import all the relevant libraries, such as NumPy, pandas, and sklearn. We will examine the variable data types. The method is demonstrated in the following:



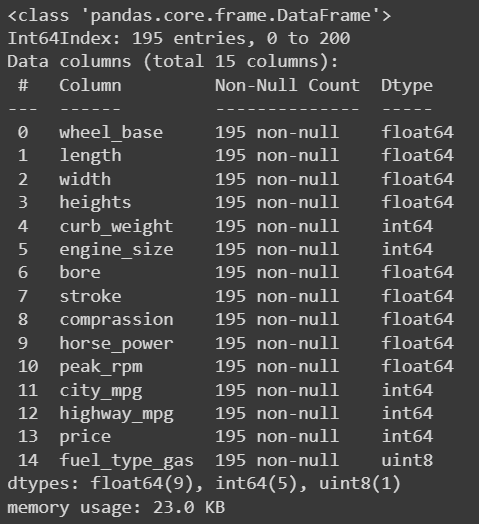
The three main data types are float64, int64, and object. There are 18 different types of this variable data. Next task is to Replace '?' with None and to Change the variables: bore, stroke, horse\_power, peak\_rpm to float64.



The datatypes of the variables bore, stroke, horse\_power, peak\_rpm is changed to float64. The columns that have just null values must also be removed, which is done as shown below.



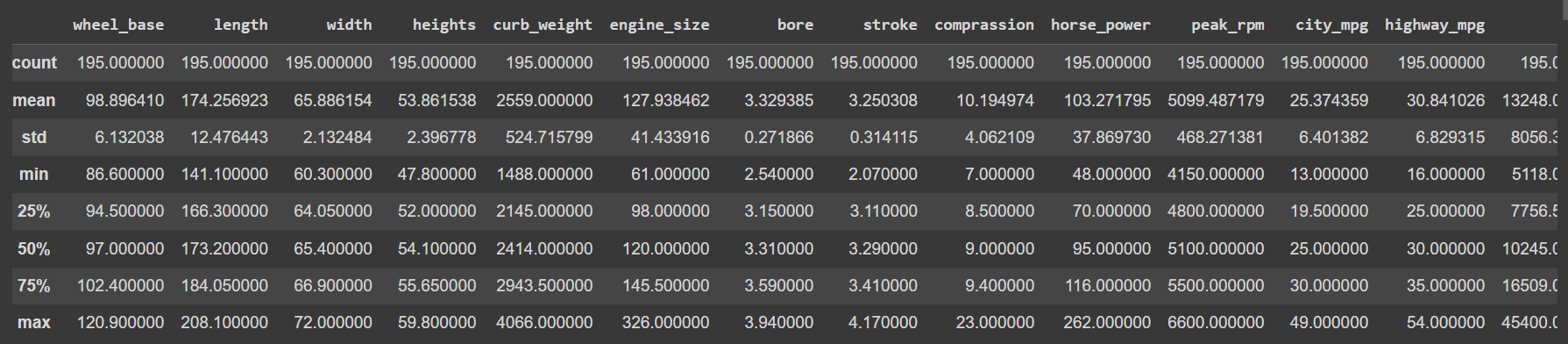
The dummy variables for fuel\_type within is also removed using get dummies method which is shown below.After executing the code, the dummy variables are removed which the total variables number of variables in the code will be 15.



Exploratory Data Analysis:

Step 1: Descriptions and features

The data set that we are using has in total 15 variables with float64, int64 and object as their data type. To obtain the mean, maximum, minimum, and other statistical figures for each column, we use the describe () function. We describe the data in order to highlight central tendencies and the distributional structure of the dataset.

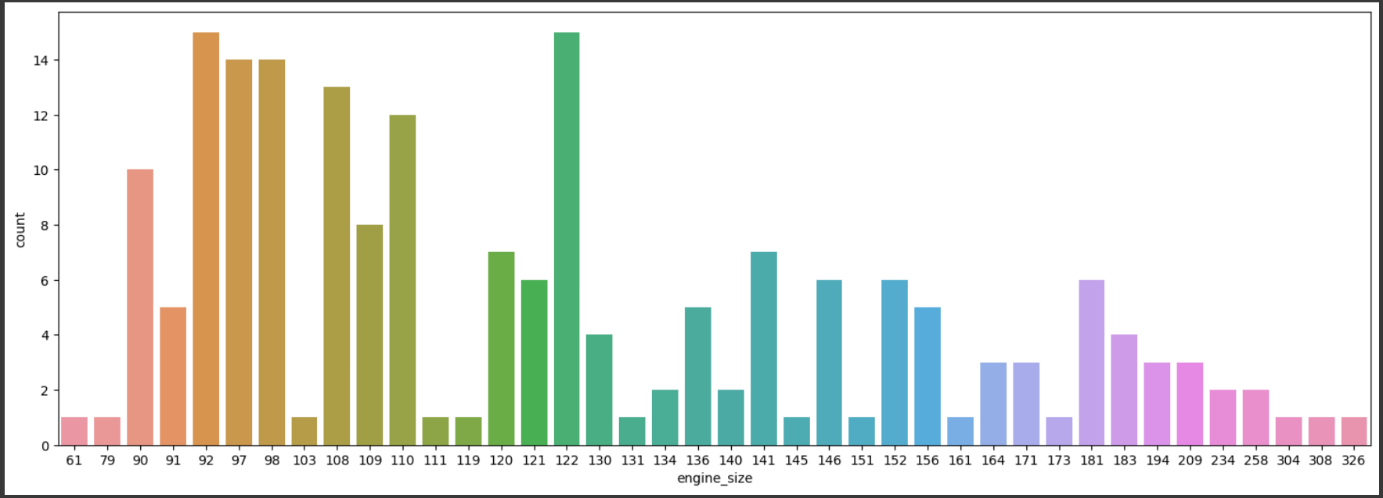
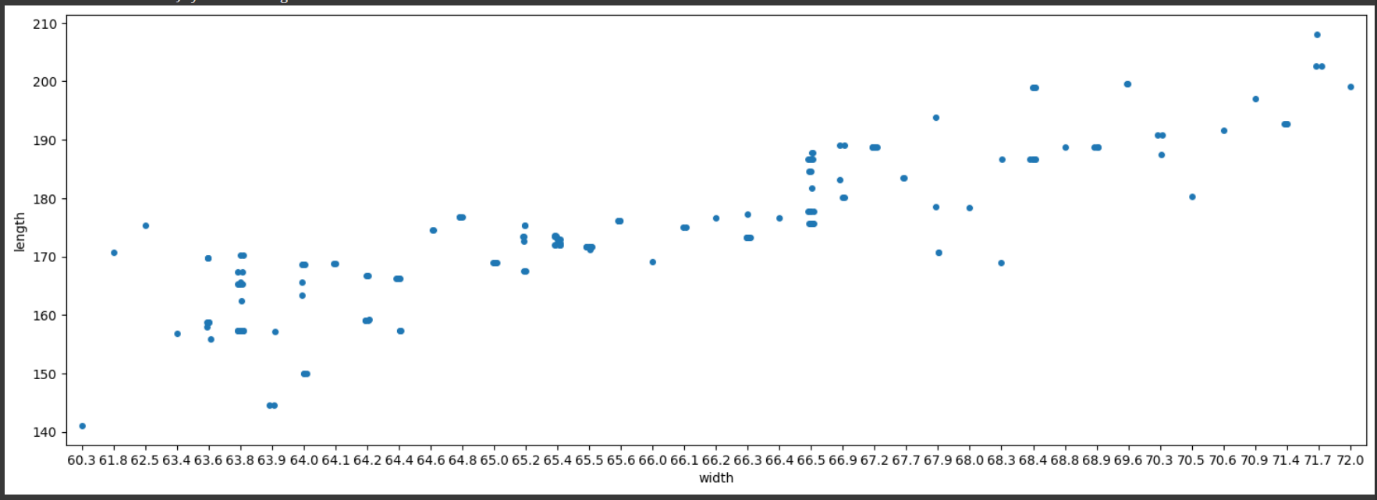
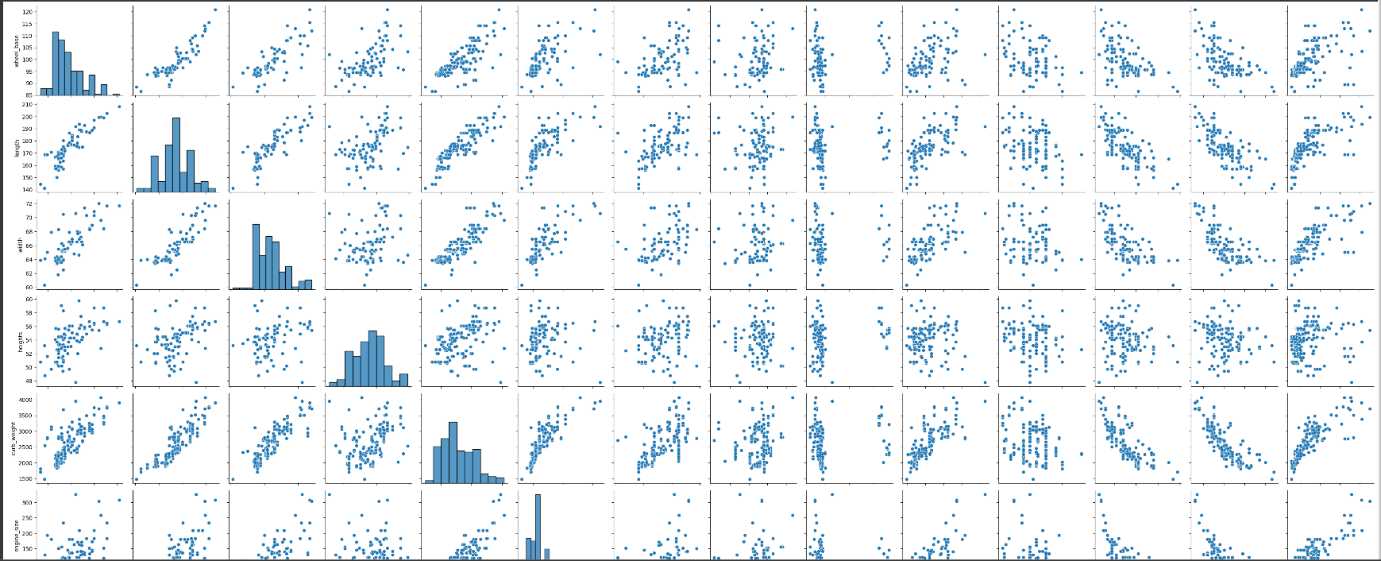


Step 2: Checking Missing value

Looking for null values, duplicate values, and the number of unique values in the columns. The count is used to determine whether the data is balanced.

Step 3: Checking the shape of the data

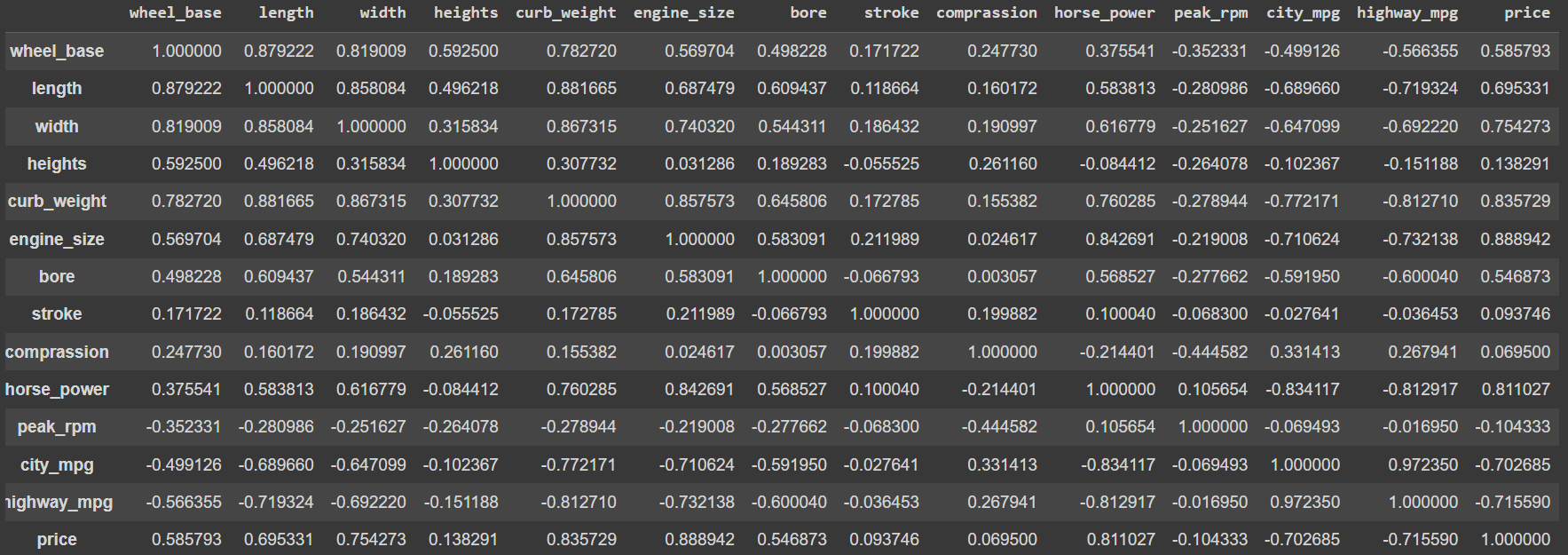
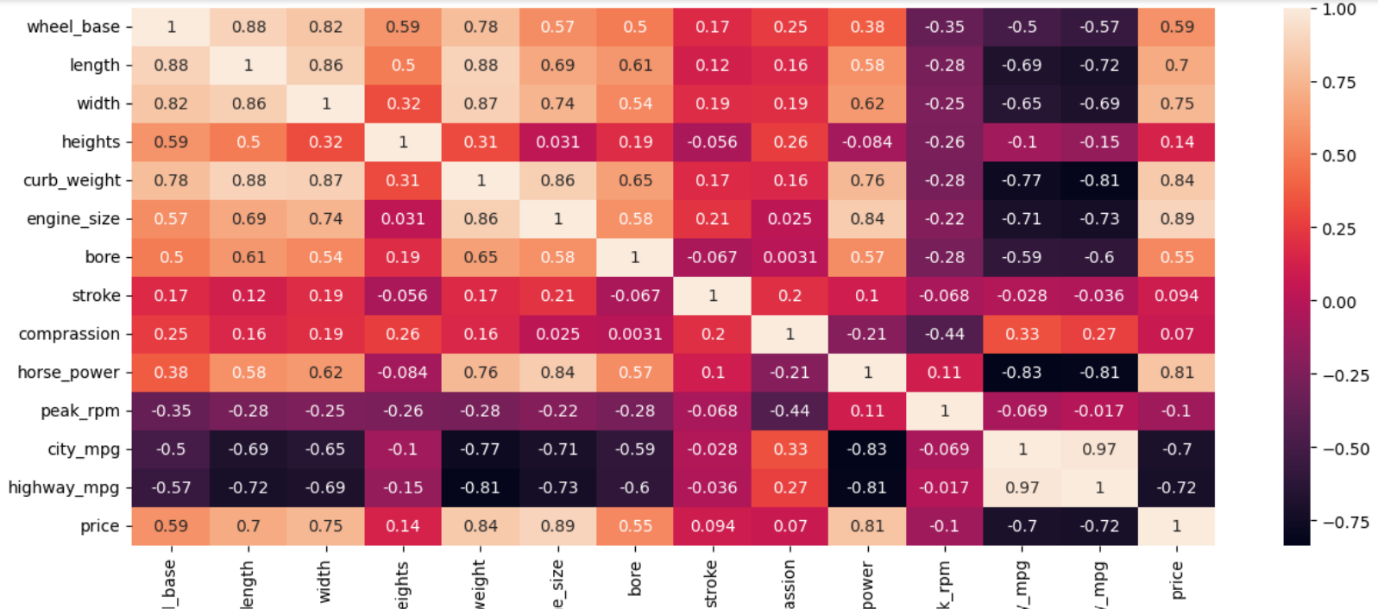
To determine the shape of the data, we produced countplot, stripplot for variables length, width and engine size along with historgram and pairplot.

Step 4: Identifying significant correlations

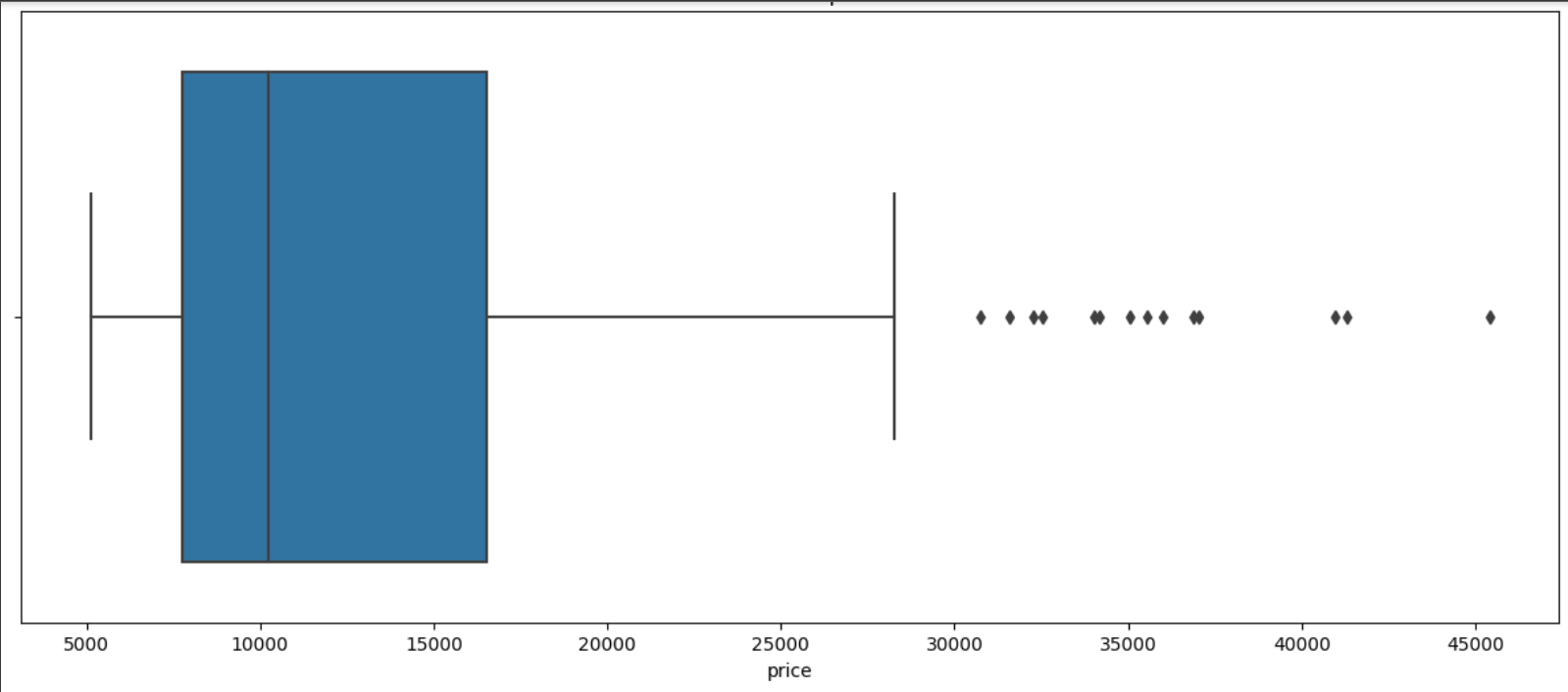
The heatmap, a coloured matrix, displays the correlation between the variables in the data set. All correlations are demonstrated to be positive, and the grid shows how each connection is connected to the others. The variables in the dataset are either positively or directly connected.

Correlation is used to look at how the variables are related. Based on a scale of -1 to 1, correlation is defined as a negative or indirect connection, +1 as a positive or direct association, and 0 as no correlation. We used pearson correlation to determine how much the variables were linearly correlated. We created a heatmap to provide as proof of this.

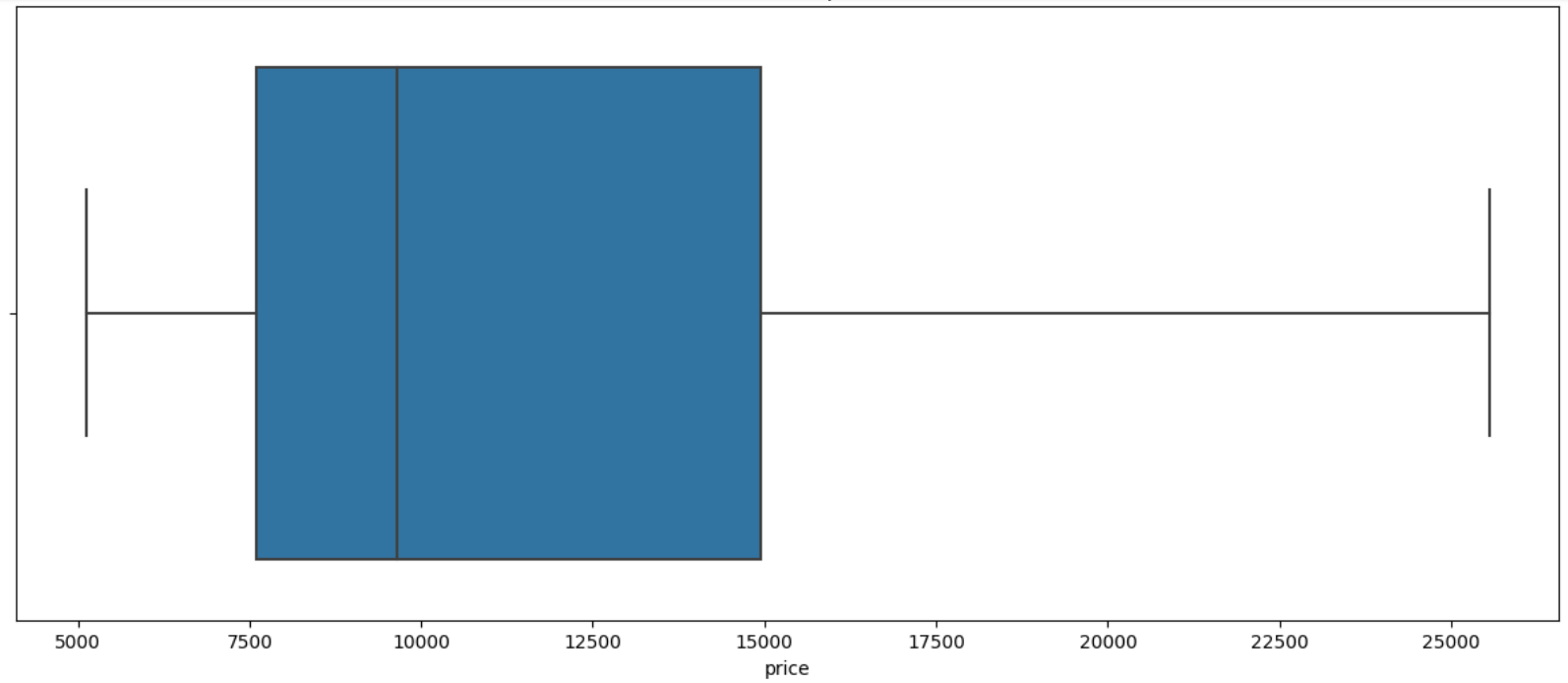
 

Step 5: Detecting and Handling outliers

Since outliers are outside the box plots, they are easy to identify. Because outliers might make it harder to understand the data, we need to be aware of them. We ultimately plotted a box plot by removing the outliers. We are removing outliers for price variable by plotting a boxplot and the removing from it.

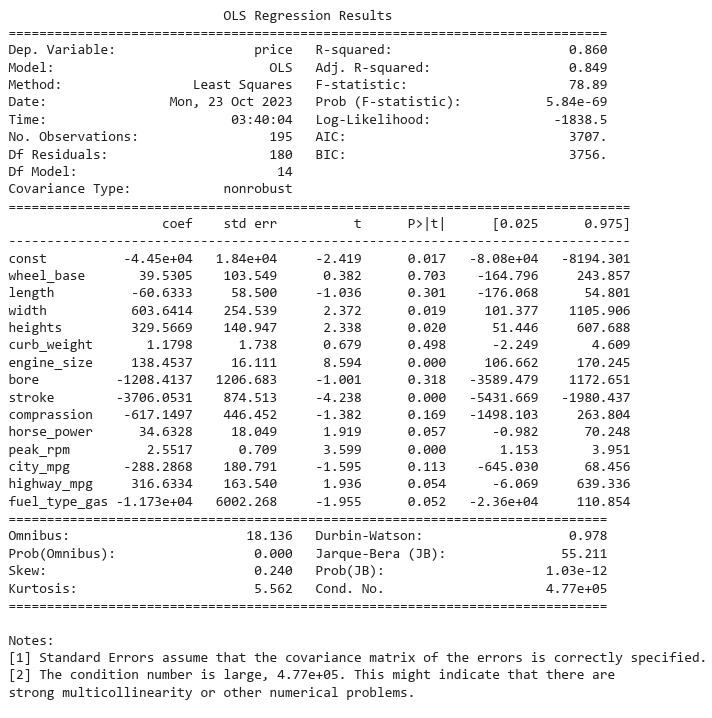


After identifying which rows the outliers are present we will removing it from the data, which we will get as:



**Part Three: Multiple Regression Analysis:**

First, we will be creating summary of the regression model (coefficents, stanrard errors, confidence intervals and other metrics shown in class). The output will be:



1.How do you interpret the intercept?

In a linear regression model, the intercept represents the predicted value of the dependent variable when all independent variables are set to zero. The interpretation of the intercept depends on the specific variables in the given dataset

2.How many variables are statistically significant?

five variables are statistically significant that are width, heights, engine\_size, stroke and peak\_rpm

3.What is the variance of the model?

Variance of the model: 9802534.78

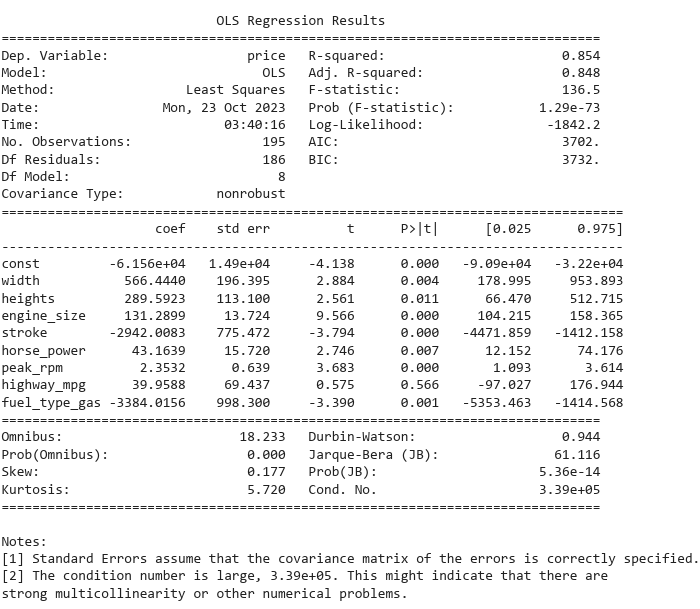
4.What is the coefficeint of determination and how do you interpret it?

Coefficient of Determination (R-squared): 0.86

5. What is the F-statistics used for? How do you interpret it for this model?

F-statistic: 78.89

Second, we will Run another regression model with price as the dependent variable and the rest of the variables as the independent variables.



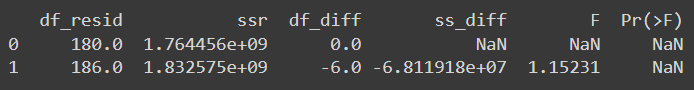
Seven variables are statistically significant that are width, heights, engine\_size, stroke and peak\_rpm, horse\_power and fuel\_type\_gas

Variance of the model: 9852556.11

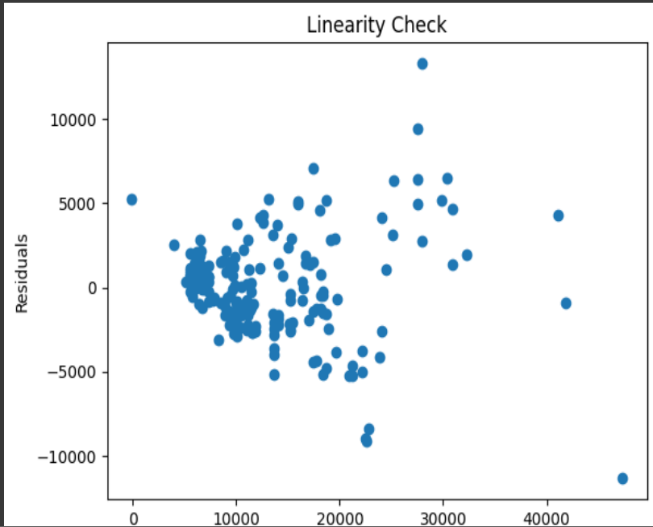
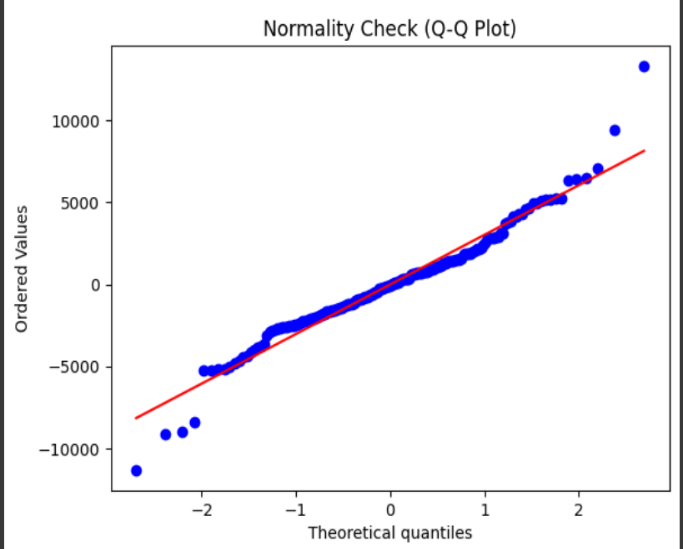
Coefficient of Determination (R-squared): 0.85

F-statistic: 136.5

Third part we will compare with ANOVA models



Fourth will be assumptions.



Fifth we will check for Multicollinearity in the data

