

Car Price Prediction Model

*A report submitted in partial fulfillment of the requirements for
the award of the degree of*

Master of Computer Applications

by

**Mohammad Salman
(17419MCA020)**



**Department of Computer Science
Institute of Science
Banaras Hindu University, Varanasi – 221005
May 2018**

CANDIDATE'S DECLARATION

I **Mohammad Salman** hereby certify that the work, which is being presented in the project report, entitled **Car Price Prediction Model**, in partial fulfillment of the requirement for the award of the Degree of **Master of Computer Applications** and submitted to the institution is an authentic record of my/our own work carried out during the period **January, 2020 to May, 2020** under the supervision of **Dr Gaurav Baranwal**. I also cited the reference about the text(s) /figure(s) /table(s) /equation(s) from where they have been taken.

The matter presented in this report as not been submitted elsewhere for the award of any other degree or diploma from any Institutions.

Date:

Signature of the Candidate

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge. The Viva-Voce examination of **Mohammad Salman**, M.C.A. Student has been held on _____.

**Signature of
Research Supervisor**

**Signature of
Head of the Department**

ABSTRACT

Nowadays, the prediction of the price is the dynamic trend as well as challenge in the market as every company & organization wants to aid good growth in their business & economy. It also helps to make a sustainable plan for future goals to achieve.

As we know in today's World, the presence & continue manufacturing of vast range & varieties of cars by the automobile companies regarding its quality, specifications, the demand for good brands by public & interest of the public created a new challenge for car price prediction in the market day by day. So in this project, I tried to make a model to predict the price/cost of cars. It also needs a considerable number of distinct attributes to examine the accurate & reliable car price prediction. I implemented four supervised machine learning techniques that are Multiple Linear Regression, Polynomial Regression, Decision Tree Regression Model and Random Forest, which helps to build a model for automobile price prediction. These predictions are mainly based on data collected from UCI Machine Learning Repository. After that, these models are evaluated and compared in order to find those which provide the best performances and accuracy.

Keywords - Car price prediction, Supervised Learning, Machine Learning, Multiple Linear Regression, polynomial Regression, Decision Tree Regression, Random Forest Regression.

TABLE OF CONTENTS

Title	Page No.
ABSTRACT	iii
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF NOTATIONS.....	xi
 CHAPTER 1 INTRODUCTION	
1.1 General Introduction.....	1
 CHAPTER 2 PROPOSED APPROACH	
2.1 Introduction.....	2
2.2 Description of Dataset	2
2.3 Data Science Methodology	5
2.3.1 Business Understanding	6
2.3.2 Analytic Approach	6
2.3.3 Data Requirements.....	7
2.3.4 Data Collection	7
2.3.5 Data Understanding	7
2.3.6 Data Preparation	7
2.3.7 Modeling	7
2.3.8 Evaluation.....	8
2.3.9 Deployments.....	8
2.3.10 Feedback.....	8
2.4 Machine Learning.....	8
2.4.1 The various feature of machine learning.....	9
2.4.2 Application of Machine Learning.....	9
2.4.3 Machine Learning Life Cycle.....	9
2.4.4 Types of Machine Learning.....	10
2.4.5 Supervised machine learning.....	10

2.4.6	Regression	11
2.4.7	Decision Tree.....	12
2.4.8	Random Forest.....	12
2.5	Important Python libraries for data science	12
2.5.1	NumPy.....	13
2.5.2	SciPy	13
2.5.3	Pandas	13
2.5.4	SciKit-Learn	13
2.5.5	Matplotlib	14
2.5.6	Seaborn.....	14
2.5.7	Pyplot	14
2.6	Some Important plots for Visualization.....	15
2.6.1	Scatter plot.....	15
2.6.2	Box plot.....	15
2.6.3	Pie Chart	16
2.6.4	Regression Plot	17
2.6.5	Distribution Plot.....	17
2.6.6	Heatmap	18
2.7	Statistical and Mathematical tools for data science.....	18
2.7.1	Correlation.....	18
2.7.2	P-value.....	19
2.7.3	Mean Square Error (MSE)	19
2.8	Cross-Validation.....	20
2.8.1	K-fold Cross-Validation	20
2.9	Data Normalization	21
2.10	Data Standardization	21

CHAPTER 3 IMPLEMENTATION

3.1	Introduction.....	22
3.2	Data Acquisition.....	22
3.3	Data Cleansing	25
3.3.1	Identify missing data.....	25

3.3.2	Deal with missing data.....	28
3.3.3	Correct data format	29
3.4	Data Analysis and Visualization	30
3.4.1	Analysis of continuous numerical variables	30
3.4.2	Analysis of Categorical Variables	36
3.4.3	An examination of price trend	41
3.4.4	Conclusion: Important variables.....	43
3.5	Data Preprocessing	43
3.5.1	One Hot Encoding	43
3.5.2	Data Standardization.....	44
3.5.3	Data Normalization.....	44
3.6	Model Development	44
3.6.1	Data splitting	45
3.6.2	Model -1: Multiple Linear Regression.....	45
3.6.3	Model -2: Polynomial Regression	46
3.6.4	Model -3: Polynomial Regression	48
3.6.5	Model -4: Random Forest Regression	49
3.7	Model Evaluation	51
3.7.1	General Evaluation.....	51
3.7.2	Model Evaluation by using K-fold Cross-Validation	52
3.8	Save Model	54

CHAPTER 4 RESULTS AND DISCUSSION

4.1	Introduction.....	55
4.2	Testing The model With new Dataset	55
4.3	Visualization of Result	55
4.4	Discussion on Result	57

CHAPTER 5 CONCLUSION AND FUTURE WORK

5.1	Conclusion	58
5.2	Limitations of the Study and Suggestion for Further work	58

REFERENCES	59
PLAGIARISM REPORT	60

LIST OF TABLES

Table No.	Title	Page No.
2.1	Description of the Attributes of the Automobile dataset.....	5
3.1	R-squares and RMSE value for each model	35
3.2	R ² score, mean and standard deviation for each model in cross-validation	53
3.3	RMSE values, mean and standard deviation for each model in cross-validation	53
4.1	R ² score and RMSE value for test dataset	57

LIST OF FIGURES

Figure No.	Title	Page No.
2.1	Building block of the working process in data science methodology	6
2.2	Machine Learning Life Cycle	9
2.3	Scatter Plot	15
2.4	Box Plot	16
2.5	Pie Chart	16
2.6	Regression Plot.....	17
2.7	Distribution Plot	17
2.8	Heatmap	18
2.9	Correlation	19
3.1	Top five rows of dataset	22
3.2	Information about dataset	23
3.3	Describe numerical variables	24
3.4	Describe categorical variables	24
3.5	NaN values in the dataset	25
3.6	Dataset in Boolean form (True or False)	26
3.7	The number of NaN value in each variable	27
3.8	Variables and its correct data types	29
3.9	The correlation between the variables through heat map	31
3.10	P-value of each numerical Variable with respect to price	32
3.11	The positive linear relationship	34
3.12	The negative linear relationship	35
3.13	Weak linear relationship	36
3.14	Pie-chart and box plots for illustrating the percentage of labels in the variables and the relationship between categorical variables and price	40
3.15	Plots for illustrating the price trends through histogram and box plot and also show its percentile, mean and standard deviation.....	42
3.16	The change of categorical variable into numerical variables.....	44

3.17	Distribution plot and Scatter plot between actual value and predicted value of train data for Multiple Linear Regression	45
3.18	Distribution plot and Scatter plot between actual value and predicted value of test data for Multiple Linear Regression	46
3.19	Distribution plot and Scatter plot between actual value and predicted value for train data for Polynomial Regression	47
3.20	Distribution plot and Scatter plot between actual value and predicted value for test data for Polynomial Regression	47
3.21	Distribution plot and Scatter plot between actual value and predicted value for train data for Decision Tree Regression	48
3.22	Distribution plot and Scatter plot between actual value and predicted value for test data for Decision Tree Regression	49
3.23	Distribution plot and Scatter plot between actual value and predicted value for train data and Random Forest Regression.....	50
3.24	Distribution plot and Scatter plot between actual value and predicted value for test data and Random Forest Regression	50
4.1	Distribution plot for the predicted values and the actual values for model testing	55
4.2	Scatter plot of the predicted value with respect to actual value	56

LIST OF NOTATIONS

Notation	Description
$A < B$	A is less than B
$A > B$	A is greater than B
$A \leq B$	A is less than or equal to B
$A \geq B$	B is greater than or equal to B

CHAPTER 1

INTRODUCTION

1.1. General Introduction:

Every human being has got his/her business in this world. The term business comes with certain tagged words such as sales, profit and loss. The success of a business is not only determined by the profit on the product in the sales, and its number of sales also identifies the performance of its product in the industry. Hence in order to improve the sales of the product suitable and customer friendly price is essential for every product. So sales prediction is one of the master trades of business which may open the gateways for obtaining knowledge about the existing market trends and the ways to conquer the market. The price of any product depends on its features, looks, quality and many factors. We have to do an in-depth analysis of those factors and choose the vital factor that acts a significant role to decide the cost of the product.

In this project, we build a model to predict the estimated price of an car by using supervised machine learning techniques. Car price prediction is a challenging task due to the large number of attributes that should be considered for the accurate prediction. The main step in the prediction process is the collection and preprocessing of the dataset. We used an automobile dataset in this project. The dataset was downloaded from UCI Machine Learning Repository. Before building a model, we analyze all the factors and attributes that make a significant effect on the price. The most important ones are its make (and model), the origin of the car (the original country of the manufacturer), its mileage and its horsepower. Other factors such as the type of fuel it uses, the interior style, the braking system, acceleration, the volume of its cylinders (measured in cc), the number of cylinders safety index, aspiration, body-style, its size, number of doors, the weight of the car, length and height of the car, drive-wheel, engine-location and so forth. After analysis of the attributes, we build four models by using supervised machine learning techniques (Multiple Linear Regression, Polynomial Regression, Decision Tree Regression Model and Random Forest) and then select the best model for prediction based on their R-square score, mean square error and cross-validation.

CHAPTER 2

PROPOSED APPROACH

2.1. Introduction:

In this section, we will introduce our dataset, all data science methodology and approaches that we applied in this project. We also discuss some essential tools and python libraries, which play a vital role in building our model.

2.2. Description of Dataset:

The data used in this project was downloaded from UCI Machine Learning Repository. It was uploaded donated by Jeffrey C Schlimmer in 1987. This is a multivariate dataset. It has 26 attributes and 205 instances. The description of the attribute value is given in the table below –

S.No	Attributes	Type	Description
1.	Symboling	int	Symboling corresponds to the degree to which the car is more risky than its price indicates. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. A value of +3 indicates that the auto is risky, -2 that it is probably pretty safe. The values ranges from -3 to +3.
2.	Normalized-losses	int	It is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door, small, station wagons, sports/specialty, etc...). The values range from 65 to 256.

3.	make	object	It determines the company name of the car(alfa-romero, audi, bmw etc..)
4.	Fuel-type	object	It determines types of fuels cars uses (gas or diesel)
5.	Aspiration	object	It is an internal combustion engine in which air intake depends solely on atmospheric pressure and which does not rely on forced induction through a turbocharger or a supercharger (turbo or std).
6.	Num-of-doors	object	It determines the number of doors in car (two or four).
7.	Body-style	object	It determines body style of the car (hardtop, wagon, sedan, hatchback, convertible)
8.	Drive-wheels	object	A drive wheel is a wheel of a motor vehicle that transmits force, transforming torque into tractive force from the tires to the road, causing the vehicle to move. These are three type that are front wheel drive (FWD), rear wheel drive (RWD), and 4WD (4 wheel drive).
9.	Engine-location	object	It determines the location of the engine in the car(front and rear)
10.	Wheel-base	float	The wheelbase is the distance between a car's front and rear wheels.
11.	Length	float	It determines the length of the cars.

12.	Width	float	It determines the width of the cars.
13.	Height	float	It determines the height of the cars.
14.	Curb-weight	int	Curb weight is the actual weight of a vehicle assigned by its manufacturer.
15.	Engine-type	object	It states how the engine is assembled or design in terms of operations of valves and cylinders. In this dataset we have seven type of engine type, dohc(Dual Overhead Cam), dohcv(Dual Overhead Cam and valve), l(L - engine), ohc, ohcf, ohcv and rotor.
16.	Num-of-cylinders	int	It determine the number of cylinders used in the cars.(2 to 12)
17.	Engine-size	int	It determine the volume of engine (in cc) used in the cars.
18.	Fuel-system	object	Fuel system is the fuel supplied in the engine with the correct amount of the fuel, for all operating circumstances(1bbl, 2bbl, 4bbl, idi, mfi, mpfi, scdi, stfi)
19.	Bore	float	Bore determine the diameter of each cylinder.
20.	Stroke	float	The stroke is the depth of the hole of the piston of the cylinder used in the car.

21.	Compression-ratio	float	It is the ratio of the volume of the cylinder and the combustion chamber when the piston at the bottom, and the volume of the combustion chamber when the piston is at the top.
22.	Horsepower	object	The horse power is the maximum power that the engine can put out.
23.	Peak-rpm	float	It determines the peak revolution per minute.
24.	City-mpg	int	It determines the average mile per gallon for the cars in the city.
25.	Highway-mpg	int	It determines the average mile per gallon for the cars on the highway.
26.	Price	float	It shows the price of each car.

Table 2.1: Description of the Attributes of the Automobile dataset

2.3. Data Science Methodology:

Data Science Methodology indicates the routine for finding solutions to a specific problem. The aim of this methodology to answer the following 10 questions in this prescribed sequence:

1. What is the problem that you are trying to solve?
2. How can you use data to answer the question?
3. What data do you need to answer the question?
4. Where is the data coming from (identify all sources) and how will you get it?
5. Is the data that you collected representative of the problem to be solved?
6. What additional work is required to manipulate and work with the data?
7. In what way can the data be visualized to get the answer that is required?

8. Does the model used really answer the initial question or does it need to be adjusted?
9. Can you put the model into practice?
10. Can you get constructive feedback into answering the question?

We answer all the question in the same manner that they asked for. Now we explained all the working process in data science methodology through the following building block-

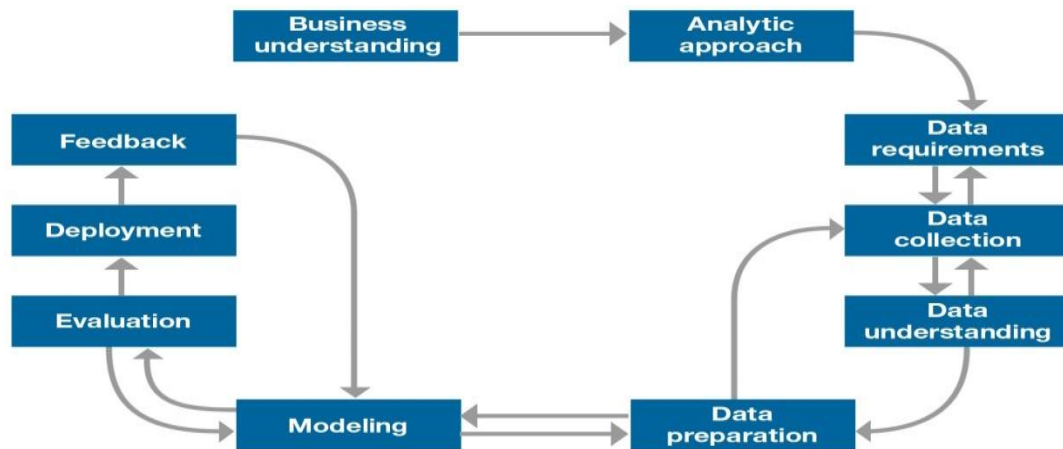


Figure 2.1: Building block of the working process in data science methodology

2.3.1. Business Understanding:

This stage describes “What is the problem that you are trying to solve?” We should have the clarity of what is the exact problem we are going to solve. Business understanding forms a concrete base, which further leads to easy resolution of queries.

2.3.2. Analytic Approach:

This stage is important because it helps to identify what type of patterns will be needed to address the question most effectively. “How can you use data to answer the question?” The approaches can be of 4 types: Descriptive approach (current status and information provided), Diagnostic approach (statistical analysis, what is happening and why it is happening), Predictive approach(it forecasts on the trends or future events probability) and Prescriptive approach(how the problem should be solved actually).

2.3.3. Data Requirements:

This stage is performed to identify When, Where, How, Who, Why and What are the data requirements to solve the problem.

2.3.4. Data Collection:

Collecting appropriate data is essential to provide required solution. Data collected can be obtained in any random format. So, according to the approach chosen and the output to be obtained, the data collected should be validated. Thus, if required one can gather more data or discard the irrelevant data.

2.3.5. Data Understanding:

Data understanding answers the question “Is the data collected representative of the problem to be solved?”. Descriptive statistics calculates the measures applied over data to access the content and quality of matter. This step may lead to reverting the back to the previous step for correction.

2.3.6. Data Preparation:

This stage describes, what are the possible ways can data be prepared. Here noise removal is done. Taking only significant items in the dataset, if we don't need specific data then we should not consider it for further process. This whole process includes transformation, normalization etc.

2.3.7. Modeling:

This stage is all about data visualization. In what way can the data be visualized to get the solution that is required. Modeling decides whether the data prepared for processing is appropriate or requires more finishing and seasoning. This phase focuses on the building of predictive/descriptive models.

2.3.8. Evaluation:

This stage checks the model which is used to really answer the initial question or does it need to be adjusted. It undergoes diagnostic measure phase (the model works as intended and where are modifications required) and statistical significance testing phase (ensures about proper data handling and interpretation).

2.3.9. Deployment:

As the model is effectively evaluated it is made ready for deployment in the business market. Deployment phase checks how much the model can withstand in the external environment and perform superiorly as compared to others.

2.3.10. Feedback:

Feedback is the necessary purpose which helps in refining the model and accessing its performance and impact. Steps involved in feedback define the review process, track the record, measure effectiveness and review with refining.

2.4. Machine Learning:

Learning is the method by which anyone can perform work more efficiently. Learning makes useful changes in our mind so that work can be done better than previous.

Definition: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .

Machine learning is a class of algorithm which is data-driven. This is not like the normal algorithm. It is the data that suggest what would be the good answer. It provides the computer to work efficiently without any explicit programmed. It provides ability to make decision to the computer. It can easily find the relation between the data and hidden pattern of the data.

2.4.1. The various feature of machine learning algorithm some of them are :

- It uses the data set to find the pattern or relation and adjust the action of programmed according to the work.
- It focuses on the development of such type of computer programs that has ability to teach them to grow and make changes when provide new data.
- It enables system to discover hidden things using iterative algorithms without explicitly programming.
- It is useful method of data analysis.

2.4.2. Application of Machine Learning:

- Used in Google map
- Used in facebook face recognition
- In Uber
- In Dynamic Pricing
- In Netflix suggestion
- In Amazon

2.4.3. Machine Learning Life Cycle:

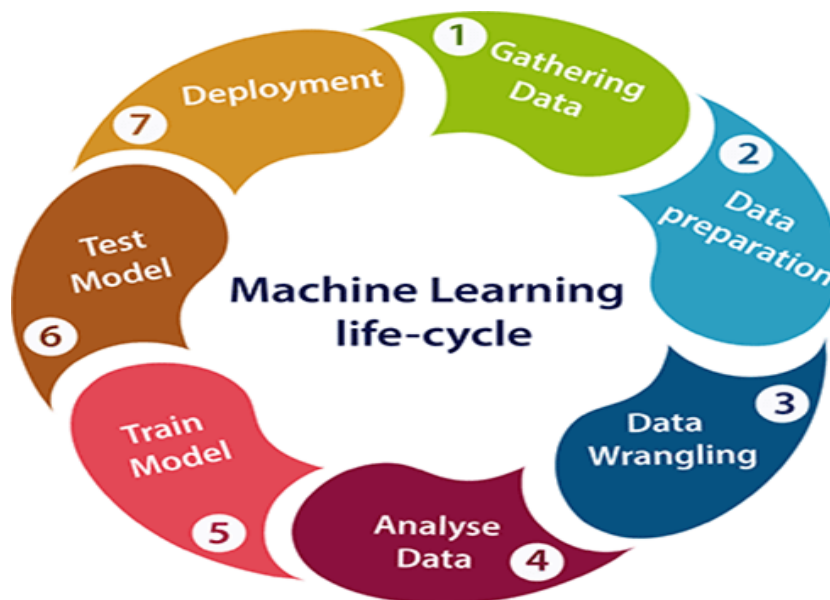


Figure 2.2: Machine Learning Life Cycle

- I. **Gathering Data:** Data can be obtained from various sources - pdf, excel, document, YouTube, databases, twitter.
- II. **Data Preprocessing:** Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.
- III. **Data wrangling:** This can be done in various steps. First of all acquire the data from the sources. Then do the data filtration process. After this clean the data.
- IV. **Analysis Data:** In this step data is converted into data model.
- V. **Train Model:** Train the algorithm so that algorithm may understand the pattern and role of the data.
- VI. **Test Model:** In this part test the accuracy of the data model.
- VII. **Deployment:** If model is feasible then it will be deployed in the real system.

2.4.4. Types of Machine Learning:

There are mainly three types of machine learning

- Supervised Machine learning
- Unsupervised Machine learning
- Reinforcement Learning

Note: In this project, we use only supervised machine learning. So, we will discuss supervised machine learning only.

2.4.5. Supervised machine learning:

Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. In other words, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. Supervised learning is the type of machine learning where you know the input variables(x) and output variable (y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

There are mainly two types of supervised learning –

- A. Classification
- B. Regression

Note: In this project, we use only regression of supervised machine learning. So, we will discuss regression only.

2.4.6. Regression:

Regression is a statistical process for estimating the relationships between the dependent variables or criterion variables and one or more independent variables or predictors. Regression analysis explains the changes in criteria in relation to changes in select predictors. The conditional expectation of the criteria based on predictors where the average value of the dependent variables is given when the independent variables are changed. Three major uses for regression analysis are determining the strength of predictors, forecasting an effect, and trend forecasting.

Now, we will discuss some types of regression that we will use in the modeling process.

2.4.6.1. Linear regression:

linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression

- **Simple Linear Regression:**

$$Y = mX + c$$

Where m is a slope and c is a intercept.

- **Multiple Linear Regression:**

$$Y = a + m_1 * x_1 + m_2 * x_2 + m_3 * x_3 + \dots$$

2.4.6.2. Polynomial Regression:

Polynomial regression is used for curvilinear data. Polynomial regression is fit with the method of least squares. The goal of regression analysis to model the expected value of a dependent variable y in regards to the independent variable x . The variable x has degree more than one. This is also two types, one is simple polynomial which have only one independent variable and other is multiple polynomial regression that

have more than one independent variable. The equation for polynomial regression for one independent is:

$$y = w_1x_1 + w_2x_2^2 + 3$$

Two other method we used in the model are -

- **Decision tree**
- **Random forest**

2.4.7. Decision tree:

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. . The final result is a tree with **decision nodes** and **leaf nodes**. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

2.4.8. Random Forest:

Random Forest technique is capable of performing both regression and classification task more efficiently than any other model. Random Forest uses an ensemble of decision trees to perform regression tasks. In ensemble learning different algorithm together improve the prediction of a model. This uses different model implementation and then chooses the maximum information gain technique. In this technique maximum depth should also be define that is pruning should be made to avoid the overfitting.

2.5. Important Python libraries for data science:

Python is one of the most popular languages used by data scientists and software developers alike for data science tasks. The best part of the python is there are some open-source libraries that make Python data tasks much, much easier. Here's a line-up of the most important Python

libraries for data science tasks, covering areas such as data processing, modeling, and visualization.

2.5.1. NumPy:

NumPy (Numerical Python) is a perfect tool for scientific computing and performing basic and advanced array operations. The library offers many handy features performing operations on n-arrays and matrices in Python. It helps to process arrays that store values of the same data type and makes performing math operations on arrays (and their vectorization) easier. In fact, the vectorization of mathematical operations on the NumPy array type increases performance and accelerates the execution time.

2.5.2. SciPy:

This useful library includes modules for linear algebra, integration, optimization, and statistics. Its main functionality was built upon NumPy, so its arrays make use of this library. SciPy works great for all kinds of scientific programming projects (science, mathematics, and engineering). It offers efficient numerical routines such as numerical optimization, integration, and others in sub modules. The extensive documentation makes working with this library really easy.

2.5.3. Pnadas:

Pandas is a library created to help developers work with "labeled" and "relational" data intuitively. It's based on two main data structures: "Series" (one-dimensional, like a list of items) and "Data Frames" (two-dimensional, like a table with multiple columns). Pandas allows converting data structures to DataFrame objects, handling missing data, and adding/deleting columns from DataFrame, imputing missing files, and plotting data with histogram or plot box. It's a must-have for data wrangling, manipulation, and visualization.

2.5.4. SciKit-Learn:

This is an industry-standard for data science projects based in Python. Scikits is a group of packages in the SciPy Stack that were created for specific functionalities – for example, image

processing. Scikit-learn uses the math operations of SciPy to expose a concise interface to the most common machine learning algorithms. Data scientists use it for handling standard machine learning and data mining tasks such as clustering, regression, model selection, dimensionality reduction, and classification. Another advantage? It comes with quality documentation and offers high performance.

2.5.5. Matplotlib:

This is a standard data science library that helps to generate data visualizations such as two-dimensional diagrams and graphs (histograms, scatterplots, non-Cartesian coordinates graphs). Matplotlib is one of those plotting libraries that are really useful in data science projects — it provides an object-oriented API for embedding plots into applications. It's thanks to this library that Python can compete with scientific tools like MatLab or Mathematica. However, developers need to write more code than usual while using this library for generating advanced visualizations. Note that popular plotting libraries work seamlessly with Matplotlib.

2.5.6. Seaborn:

Seaborn is based on Matplotlib and serves as a useful Python machine learning tool for visualizing statistical models – heatmaps and other types of visualizations that summarize data and depict the overall distributions. When using this library, you get to benefit from an extensive gallery of visualizations (including complex ones like time series, joint plots, and violin diagrams).

2.5.7. Pydot:

This library helps to generate oriented and non-oriented graphs. It serves as an interface to Graphviz (written in pure Python). You can easily show the structure of graphs with the help of this library. That comes in handy when you're developing algorithms based on neural networks and decision trees.

2.6. Some Important plots for Visualization :

2.6.1. Scatter Plot:

A scatter plot is a diagram where each value in the data set is represented by a dot. The Matplotlib and Seaborn module have a method for drawing scatter plots, it needs two arrays of the same length, one for the values of the x-axis, and one for the values of the y-axis.

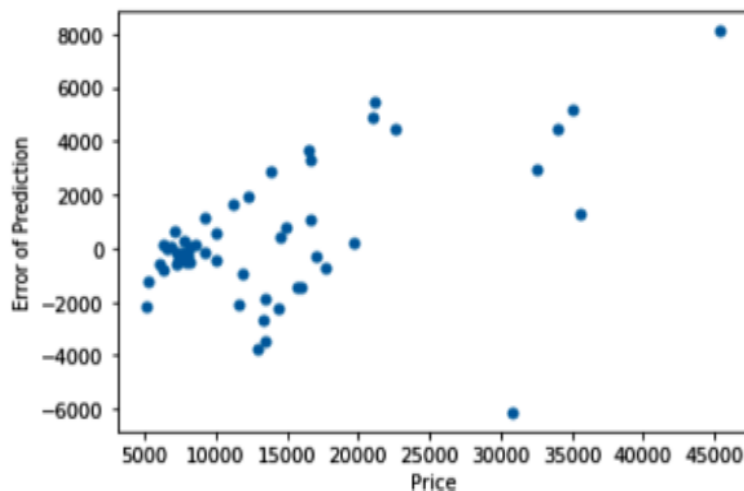


Figure 2.3: Scatter plot

2.6.2. Box plot:

Box Plot is the visual representation of the depicting groups of numerical data through their quartiles. Boxplot is also used for detect the outlier in data set. It captures the summary of the data efficiently with a simple box and whiskers and allows us to compare easily across groups. Boxplot summarizes a sample data using 25th, 50th and 75th percentiles. These percentiles are also known as the lower quartile, median and upper quartile.

A box plot consists of 5 things.

- Minimum
- First Quartile or 25%
- Median (Second Quartile) or 50%
- Third Quartile or 75%
- Maximum

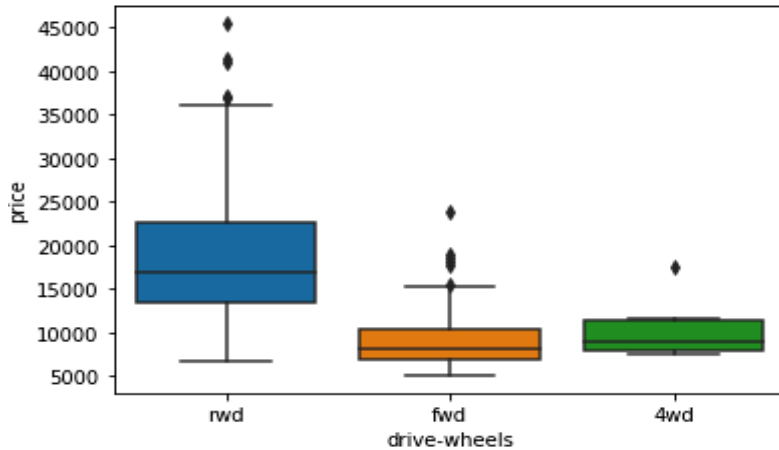


Figure 2.4: Box plot

2.6.3. Pie Chart:

A Pie Chart is a circular statistical plot that can display only one series of data. The area of the chart is the total percentage of the given data. The area of slices of the pie represents the percentage of the parts of the data. The slices of pie are called wedges. The area of the wedge is determined by the length of the arc of the wedge. Pie charts are commonly used in business presentations like sales, operations, survey results, resources, etc as they provide a quick summary.

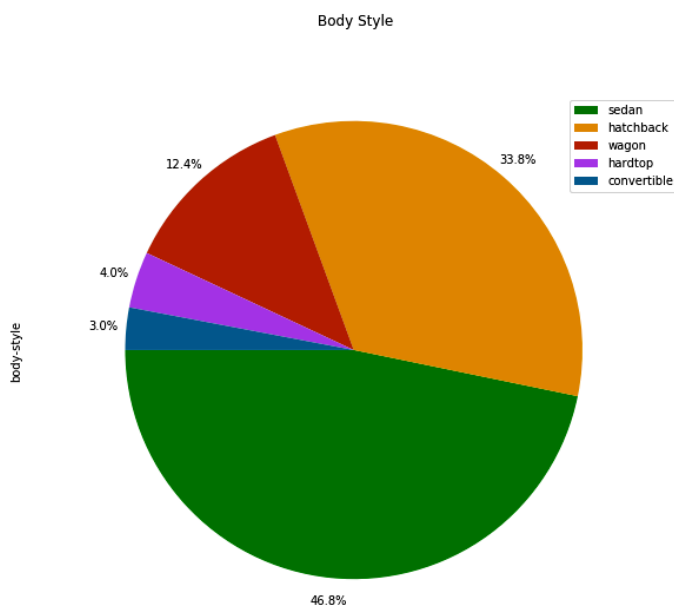


Figure 2.5: Pie Chart

2.6.4. Regression plots:

The regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyses. Regression plots as the name suggests creates a regression line between 2 parameters and helps to visualize their linear relationships.

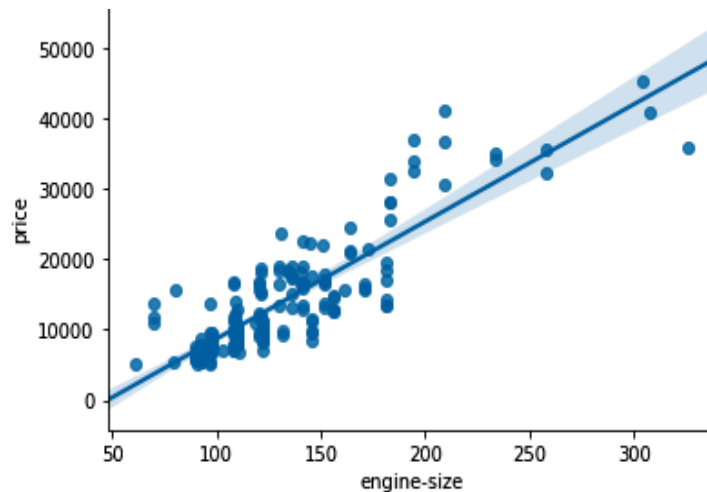


Figure 2.6: Regression plot

2.6.5. Distribution plot:

The distribution plots is used for examining univariate and bivariate distributions.

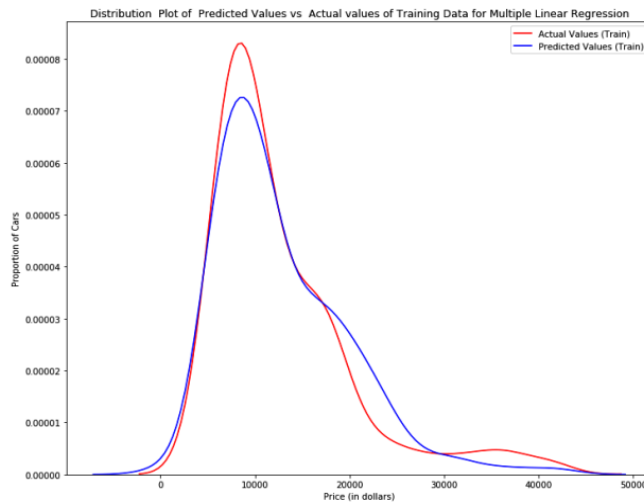


Figure 2.7: Distribution Plot

2.6.6. Heatmap:

Heatmap is defined as a graphical representation of data using colors to visualize the value of the matrix. In this to represent more common values or higher activities brighter colors basically reddish colors are used and to less common or activity values darker colors are preferred. Heatmap is also defined by the name of the shading matrix.

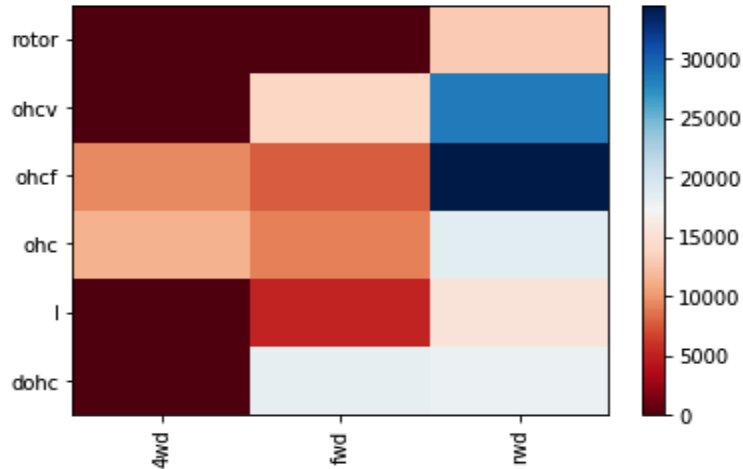


Figure 2.8: Heatmap

2.7. Statistical and Mathematical tools for data science:

2.7.1. Correlation:

The correlation show whether and how strongly pairs of variables are related to each other. It gives the direction and strength of relationship between variables. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.

Formula for correlation:

$$Corr(x, y) = \frac{\sum_{i=1}^n (x_i - x') (y_i - y')}{\sqrt{\sum_{i=1}^n (x_i - x')^2 \sum_{i=1}^n (y_i - y')^2}}$$

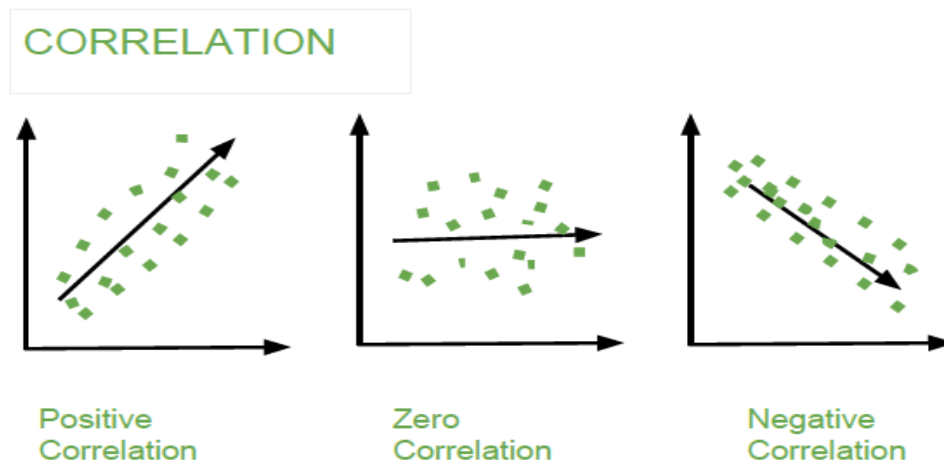


Figure 2.9: Correlation

2.7.2. P-Value:

P-value helps us determine how likely it is to get a particular result when the null hypothesis is assumed to be true. It is the probability of getting a sample like ours or more extreme than ours if the null hypothesis is correct. Therefore, if the null hypothesis is assumed to be true, the p-value gives us an estimate of how “strange” our sample is. If the p-value is very small (<0.05 is considered generally), then our sample is “strange,” and this means that our assumption that the null hypothesis is correct is most likely to be false. Thus, we reject it.

2.7.3. Mean Square Error (MSE):

The Mean Square Error of an estimator measures the average of error squares i.e. the average squared difference between the estimated values and true value. It is a risk function, corresponding to the expected value of the squared error loss. It is always non – negative and values close to zero are better.

Formula For MSE:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

2.7.4. R-squared Score:

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

Formula for R-Squared:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Where, SS_{RES} = Sum squared regression error

SS_{TOT} = Sum Squared total error

2.8. Cross-Validation:

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows :

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

2.8.1. K-fold Cross-Validation:

It is one of the methods of cross-validation. In this method, we split the data-set into k number of subsets (known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

2.9. Data Normalization:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Formula for normalization-

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

2.10. Data Standardization:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Formula for standardization-

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

CHAPTER 3

IMPLEMENTATION

3.1. Introduction:

Now, we will implement our automobile price prediction model on Jupyter Notebook. There are five main step in the working of the project.

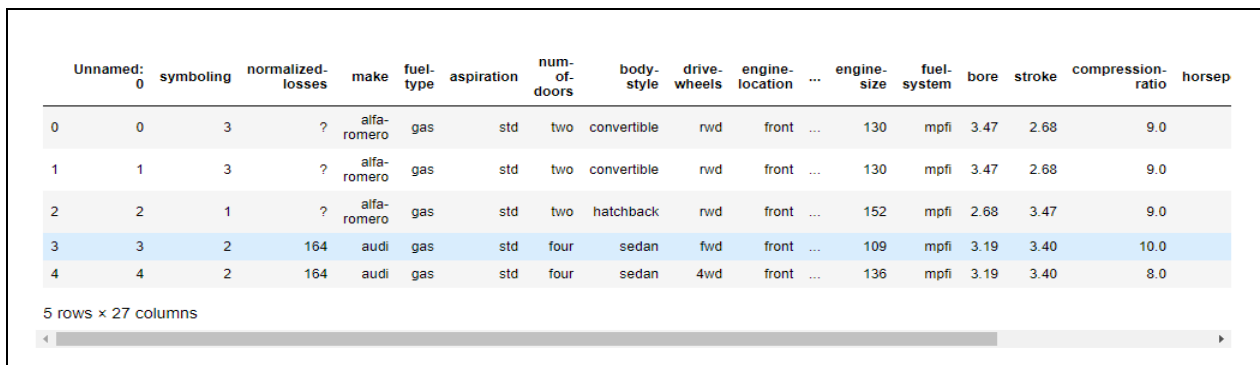
1. Data Acquisition
2. Data Cleansing
3. Data Visualization and Analysis
4. Data Preprocessing
5. Model development
6. Model Evaluation
7. Save Model

3.2. Data Acquisition:

In this section, we load the main dataset (automobile dataset) into our Jupyter Notebook by using pandas library of the python. Here the pandas library is used for reading the CSV file by read_csv function.

Basic Insight of Dataset:

- By using the head function, we find the top five values of the data in which all attribute values of the corresponding data are given.



Unnamed: 0	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower
0	0	3	? alfa-romero	gas	std	two	convertible	rwd	front	...	130	mpfi	3.47	2.68	9.0	
1	1	3	? alfa-romero	gas	std	two	convertible	rwd	front	...	130	mpfi	3.47	2.68	9.0	
2	2	1	? alfa-romero	gas	std	two	hatchback	rwd	front	...	152	mpfi	2.68	3.47	9.0	
3	3	2	164 audi	gas	std	four	sedan	fwd	front	...	109	mpfi	3.19	3.40	10.0	
4	4	2	164 audi	gas	std	four	sedan	4wd	front	...	136	mpfi	3.19	3.40	8.0	

5 rows x 27 columns

Figure 3.1- Top five rows of dataset

- By using the info function, we get information of the class type, the number of value in the dataset and data type of each attributes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
symboling                205 non-null int64
normalized-losses        205 non-null object
make                     205 non-null object
fuel-type                 205 non-null object
aspiration                205 non-null object
num-of-doors              205 non-null object
body-style                205 non-null object
drive-wheels              205 non-null object
engine-location           205 non-null object
wheel-base               205 non-null float64
length                   205 non-null float64
width                    205 non-null float64
height                   205 non-null float64
curb-weight               205 non-null int64
engine-type               205 non-null object
num-of-cylinders          205 non-null object
engine-size               205 non-null int64
fuel-system               205 non-null object
bore                      205 non-null object
stroke                   205 non-null object
compression-ratio         205 non-null float64
horsepower                205 non-null object
peak-rpm                  205 non-null object
city-mpg                  205 non-null int64
highway-mpg               205 non-null int64
price                     205 non-null object
dtypes: float64(5), int64(5), object(16)
memory usage: 41.8+ KB
```

Figure 3.2 - Information about dataset

- By using describe function, we represent mean, standard deviation, minimum value, the maximum value count number of data and quartile value for each numerical variable and represent number of unique values and frequency of most frequent value for categorical variables.

	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

Figure 3.3- Describe numerical variables

	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	engine-type	num-of-cylinders	fuel-system	bore	stroke	horsepower	peak-rpm	price
count	205	205	205	205	205	205	205	205	205	205	205	205	205	205	205	205
unique	52	22	2	2	3	5	3	2	7	7	8	39	37	60	24	187
top	?	toyota	gas	std	four	sedan	fwd	front	ohc	four	mpfi	3.62	3.40	68	5500	?
freq	41	32	185	168	114	96	120	202	148	159	94	23	20	19	37	4

Figure 3.4- Describe categorical variables

3.3. Data Cleansing:

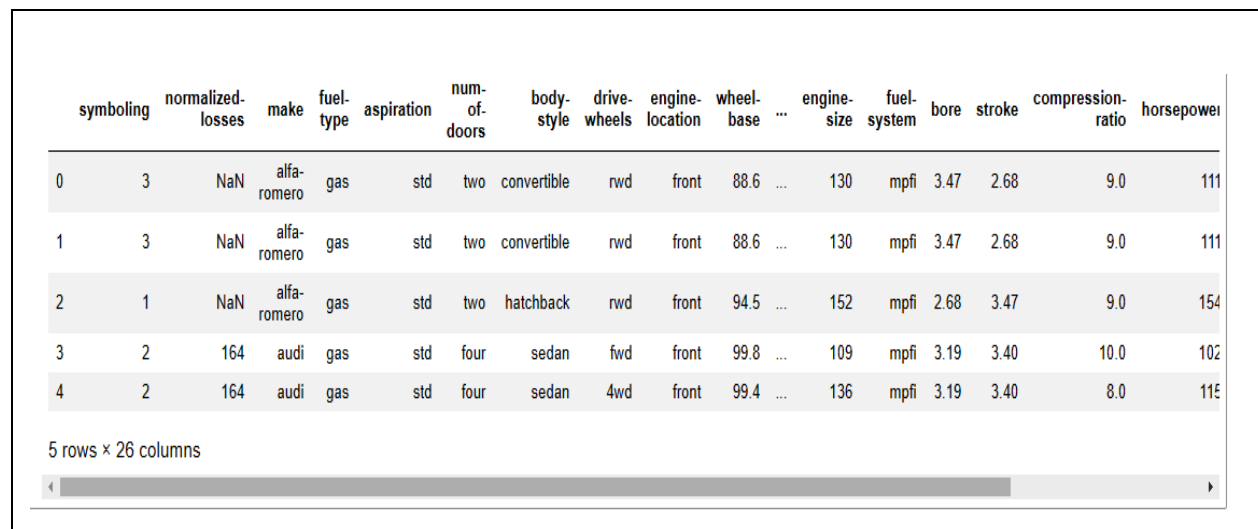
In this section, we will do the process of detecting and correcting corrupt or inaccurate records from the dataset and transform this raw data into a useful one because there are many problems with the datasets and it has a lot of missing values. We identify and remove all the problem one by one.

As we can see, there are several question marks appeared in the data frame; those are missing values which may hinder our further analysis. So, we identify all those missing values and deal with them. We apply the following three steps for working with missing data:

1. Identify missing data
2. Deal with missing data
3. Correct data format

3.3.1. Identify missing data:

In the car dataset, missing data comes with the question mark "?". We replace "?" with NaN (Not a Number)



	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower
0	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111
1	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111
2	1	NaN	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	154
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0	102
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0	115

5 rows x 26 columns

Figure 3.5: NaN values in the dataset

After that, we convert all the data in the form of Boolean values, in which, "True" stands for missing value, while "False" stands for not missing value.

	symboling	normalized- losses	make	fuel- type	aspiration	num- of- doors	body- style	drive- wheels	engine- location	wheel- base	...	engine- size	fuel- system	bore	stroke	compression- ratio	horsepower	pe
0	False	True	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	F
1	False	True	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	F
2	False	True	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	F
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	F
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	F

5 rows x 26 columns

Figure 3.6: Dataset in Boolean form (True or False)

Using for loop in Python, we can figure out the number of missing values in each column. As mentioned above, "True" represents a missing value, "False" means the value is present in the dataset. In the body of the for loop the method ".value_counts()" counts the number of "True" values.

symboling False 205 Name: symboling, dtype: int64 normalized-losses False 164 True 41 Name: normalized-losses, dtype: int64 make False 205 Name: make, dtype: int64 fuel-type False 205 Name: fuel-type, dtype: int64 aspiration False 205 Name: aspiration, dtype: int64 num-of-doors False 203 True 2 Name: num-of-doors, dtype: int64 body-style False 205 Name: body-style, dtype: int64 drive-wheels False 205 Name: drive-wheels, dtype: int64 engine-location False 205 Name: engine-location, dtype: int64	wheel-base False 205 Name: wheel-base, dtype: int64 length False 205 Name: length, dtype: int64 width False 205 Name: width, dtype: int64 height False 205 Name: height, dtype: int64 curb-weight False 205 Name: curb-weight, dtype: int64 engine-type False 205 Name: engine-type, dtype: int64 num-of-cylinders False 205 Name: num-of-cylinders, dtype: int64 engine-size False 205 Name: engine-size, dtype: int64 fuel-system False 205 Name: fuel-system, dtype: int64
bore False 201 True 4 Name: bore, dtype: int64 stroke False 201 True 4 Name: stroke, dtype: int64 compression-ratio False 205 Name: compression-ratio, dtype: int64 horsepower False 203 True 2 Name: horsepower, dtype: int64	peak-rpm False 203 True 2 Name: peak-rpm, dtype: int64 city-mpg False 205 Name: city-mpg, dtype: int64 highway-mpg False 205 Name: highway-mpg, dtype: int64 price False 201 True 4 Name: price, dtype: int64

Figure 3.7: The number of NaN values in each variable

Based on the summary above, all column has 205 rows of data, out of those there is 19 columns have 205 false value and 0 true value, but there are seven such columns which have less than 205 false values and some true values. Those columns which have some true value show missing value columns. Those columns and there missing values are-

- a. *'normalized-losses'*: 41 missing data
- b. *'num-of-doors'*: 2 missing data
- c. *'bore'*: 4 missing data
- d. *'stroke'* : 4 missing data
- e. *'horsepower'*: 2 missing data
- f. *'peak-rpm'*: 2 missing data
- g. *'price'*: 4 missing data

3.3.2. Deal with missing data:

After identify missing values, now we start dealing with missing value. We do this by one of the two following method-

A. Replace data:

It can be done by one of the three following method according to situation:

- a. Replace it by mean (for numeric data)
- b. Replace it by frequency (for categorical data)
- c. Replace it based on other functions (for both)

B. Drop data:

It can be done by one of the three following method according to situation:

- a. drop the whole row
- b. drop the whole column

The missing value of the numeric attributes (*'normalized-loss'*, *'bore'*, *'stroke'*, *'horsepower'*, and *'peak-rpm'*) replace by their mean, and the missing value of categorical attribute (*'num-of door'*) replace by *'four'* because 84% sedans is four doors.

‘price’ has 4 missing data, simply we delete the hole row because price is what we want to predict. Any data entry without price data cannot be used for prediction; therefore any row now without price data is not useful to us.

3.3.3. Correct data format:

As we can see above, some columns are not of the correct data type. Numerical variables should have type ‘float’ or ‘int’, and variables with strings such as categories should have type ‘object’. In our dataset, ‘price’, ‘bore’, ‘stroke’ and ‘peak-rpm’ variables are numerical values, but these are in object format and ‘normalized-loss’ is an integer value, but is in float format. So we have to convert them into proper format by using the "astype()" method.

symboling	int64
normalized-losses	int32
make	object
fuel-type	object
aspiration	object
num-of-doors	object
body-style	object
drive-wheels	object
engine-location	object
wheel-base	float64
length	float64
width	float64
height	float64
curb-weight	int64
engine-type	object
num-of-cylinders	object
engine-size	int64
fuel-system	object
bore	float64
stroke	float64
compression-ratio	float64
horsepower	object
peak-rpm	float64
city-mpg	int64
highway-mpg	int64
price	float64

Figure 3.8- Variables and its correct data types

3.4. Data Visualization and Analysis:

In this section, we analyze the data by applying a variety of mathematical and statistical tools and techniques to explore the data and find a pattern in it and extract useful information from data. This information is beneficial for decision-making. We also use data visualization to examine the data in a graphical format to obtain additional insight regarding the messages within the data.

In our dataset, there are two types of variables present, first one is numerical variable and other is categorical variables. We will apply different tools and techniques for these two type of variables for analysis the data.

3.4.1. Analysis of continuous numerical variables:

For analysis of continuous numerical variables, first we have to find correlation and P-value between the variables.

A. Correlation:

It is a measure of the extent of interdependence between variables. The resulting coefficient is a value between -1 and 1 inclusive, where:

- **1:** Total positive linear correlation.
- **0:** No linear correlation, the two variables most likely do not affect each other.
- **-1:** Total negative linear correlation

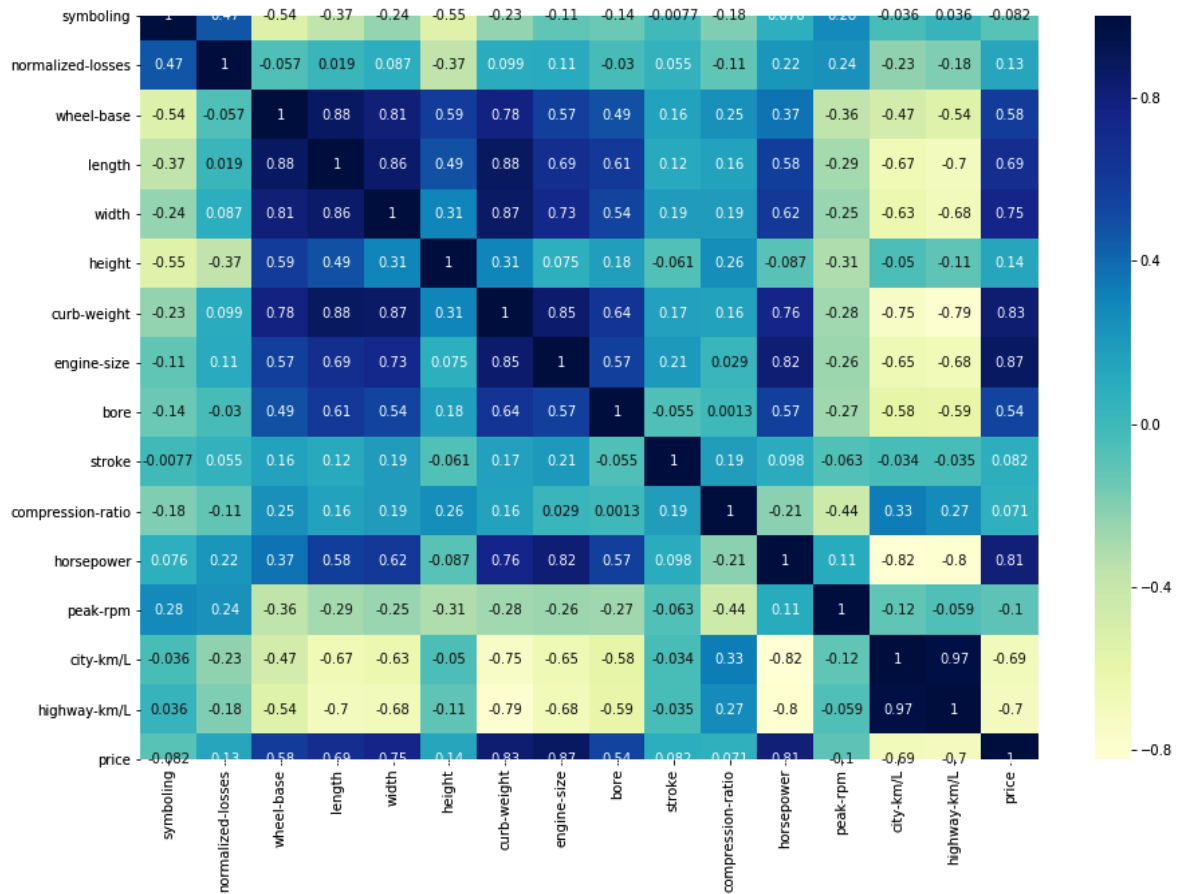


Figure 3.9: The correlation between the variables through heat map

As we saw the above figure, we found that correlation for **horsepower, engine-size, curb-weight, width and length of the car with the price is highly positive**. Their relationship is statistically significant, is shown in dark blue color or nearly dark blue color in the figure. And the other side we found that the correlation **for city-km/L and highway-km/L** with the price is highly negative and their relationship is also statistically significant, is shown in light cream color in the figure. But some attributes show correlation value close to zero, and these attributes are not significant for modeling, is shown in sky blue color in the figure.

B. P- Value:

The P-value is the probability value that the correlation between these two variables is statistically significant. Normally, we choose a significance level of 0.05, which means that we are 95% confident that the correlation between the variables is significant.

By convention, when the

- p-value is $\ll 0.001$: we say there is strong evidence that the correlation is significant.
- p-value is $\ll 0.05$: there is moderate evidence that the correlation is significant.
- p-value is $\ll 0.1$: there is weak evidence that the correlation is significant.
- p-value is $\gg 0.1$: there is no evidence that the correlation is significant.

The table shown on the next page, describe the P-value of each numerical variable. As we have seen that **wheel-base, width, length, curb-weight, bore, engine-size, horsepower, city-km/L, and highway-km/L have very low P-value($\ll 0.001$)**. So, we can say these attributes have powerful evidence that the correlation is significant, and these are helpful for the development of the model. On the other hand, remaining numerical attributes have P-value nearly equal to 0.1. So, we can say that these attributes have no evidence that the correlation is significant.

Atributes	P-valu
wheel-base	8.07 e-20
length	8.01 e-30
width	9.20 e-38
curb-weight	2.19 e-53
bore	8.05 e-17
Engine-size	9.26 e-64
Horsepower	6.27 e-48
city-km/L	2.26 e-29
highway-km/L	1.74 e-31
symboling	0.240
normalized-losses	0.057
height	0.055
stroke	0.245
compression-ratio	0.315
peak_rpm	0.151

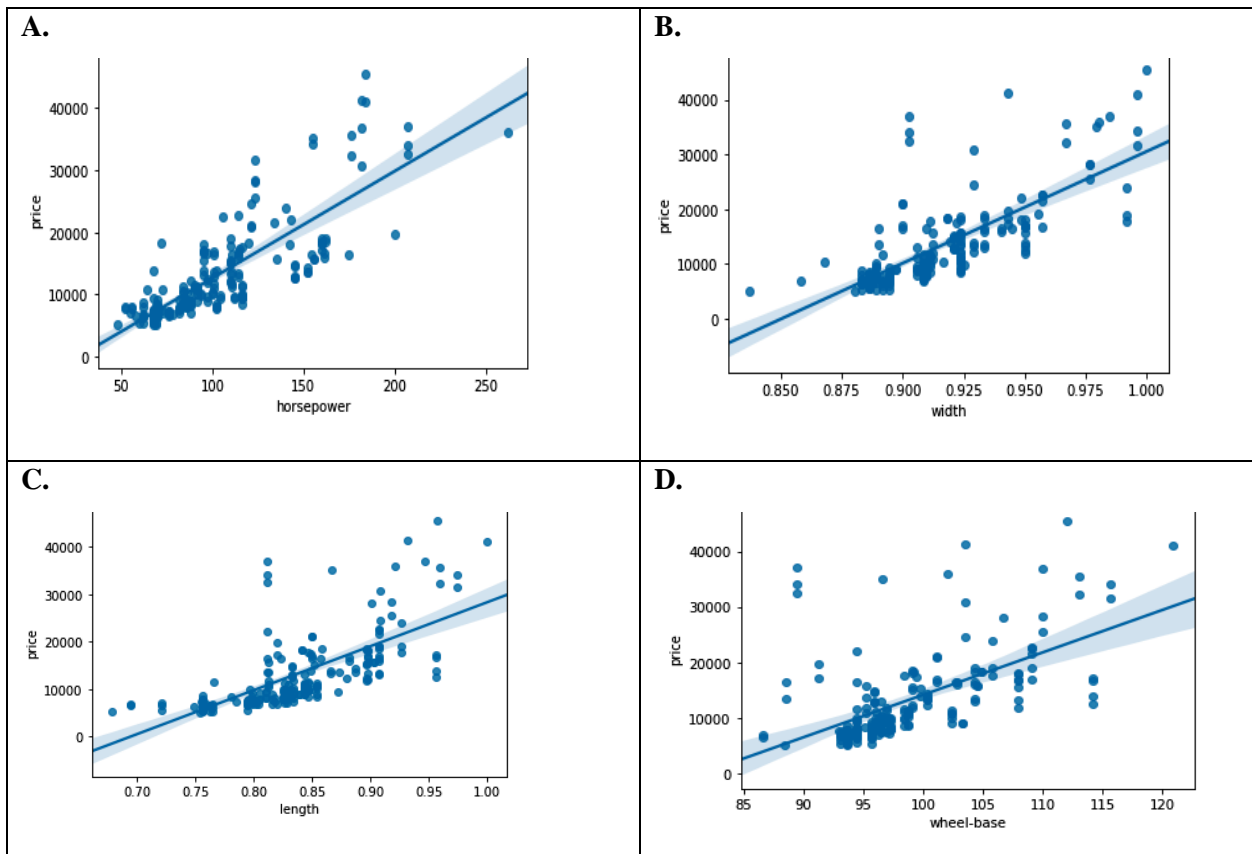
Figure 3.10- P-value of each numerical Variable with respect to price

C. Linear Relationship:

Here we examine the linear relationship between an individual variable and the price. It will help to understand how our continuous variables depend on the price of the car. It will show one of the three type of relationship positive linear relationship (increasing order), negative linear relationship (decreasing order) or weak linear relationship(horizontal). In order to start understanding the (linear) relationship, we plot the scatter plot by using "regplot". It plots the scatter plot plus the fitted regression line for the data. The following plots show that relationship -

- **Positive Linear Relationship:**

If we look at the plots shown below, we observe that, the values of the attributes go up; the price goes up; this indicates a positive direct correlation between these two variables. So, these attributes seem like a pretty good predictor of price since the regression line is almost a perfect diagonal line.



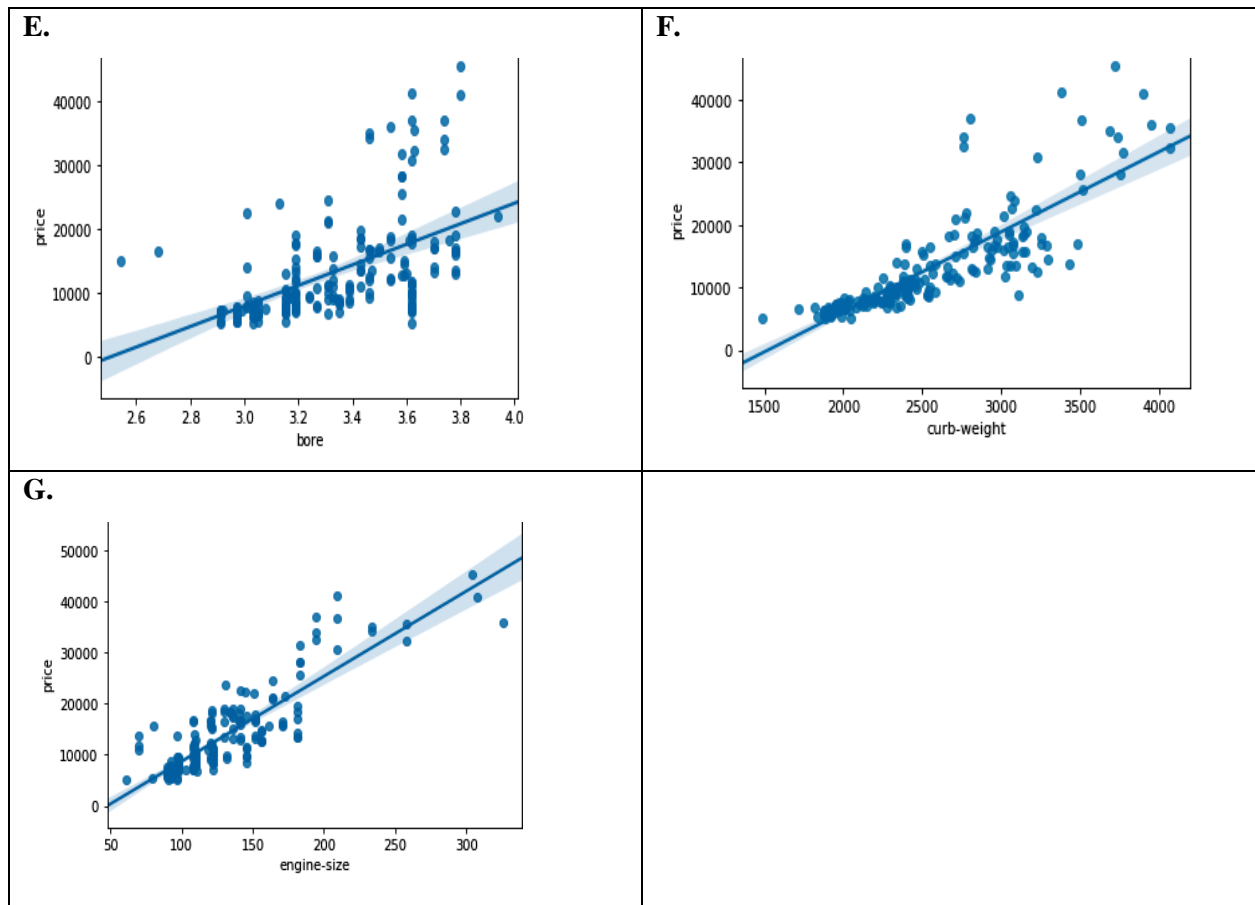


Figure 3.11-The positive linear relationship

- **Negative Linear Relationship:**

If we look at the plots shown below, we observe that the values of highway-km/L and city-km/L increase, the price goes down, this indicates an inverse/negative relationship between these two variables. So, these attributes also seem like a pretty good predictor of price.

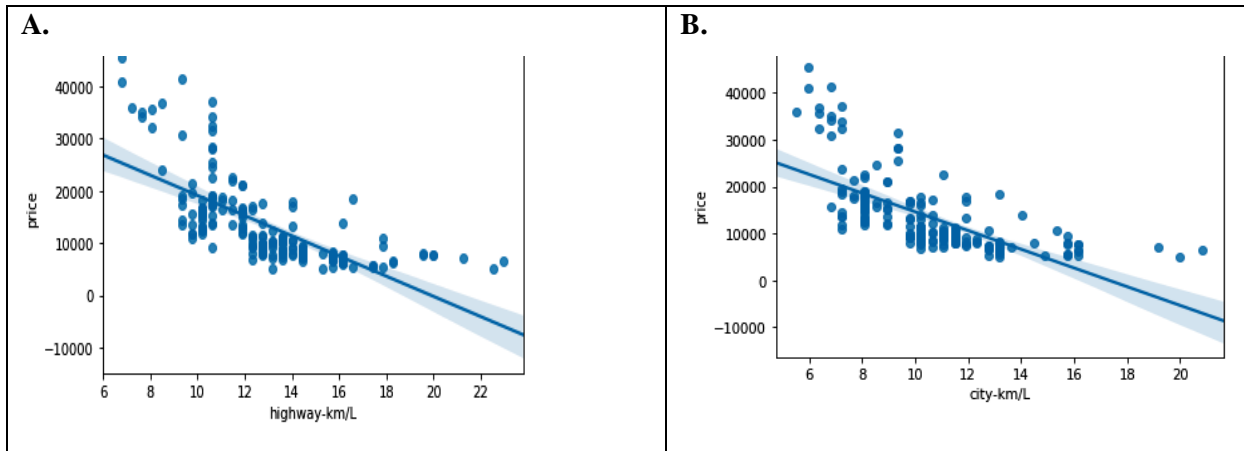
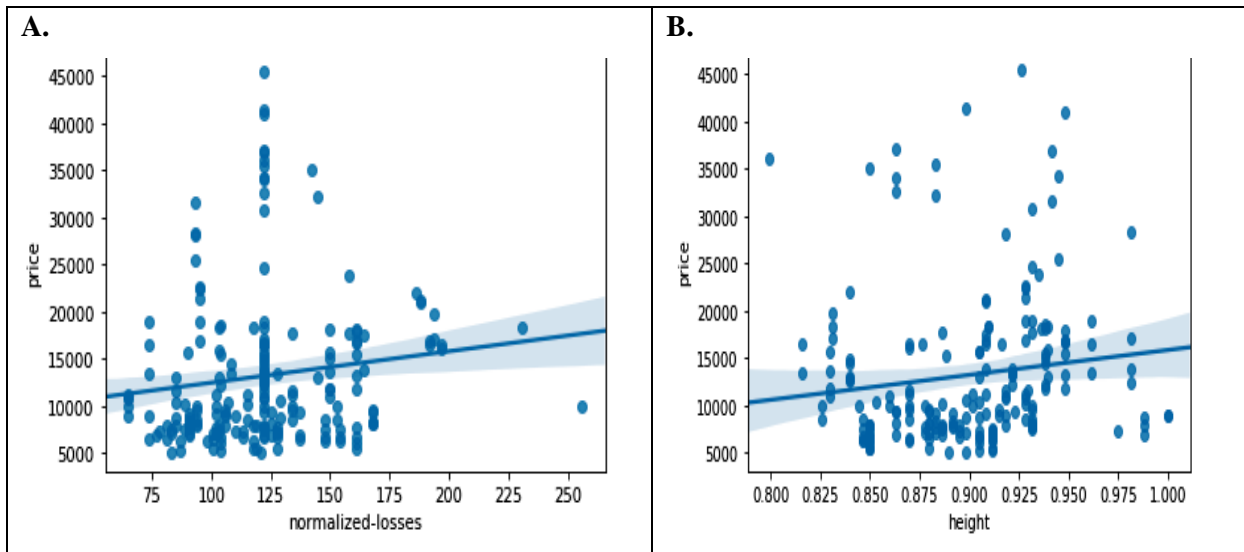


Figure 3.12-The negative linear relationship

- **Weak Linear Relationship:**

If we look at the plots shown below, we observe that the attributes do not seem like a good predictor of the price at all since the regression line is close to horizontal. Also, the data points are very scattered and far from the fitted line, showing lots of variability. Therefore these are not a reliable variable.



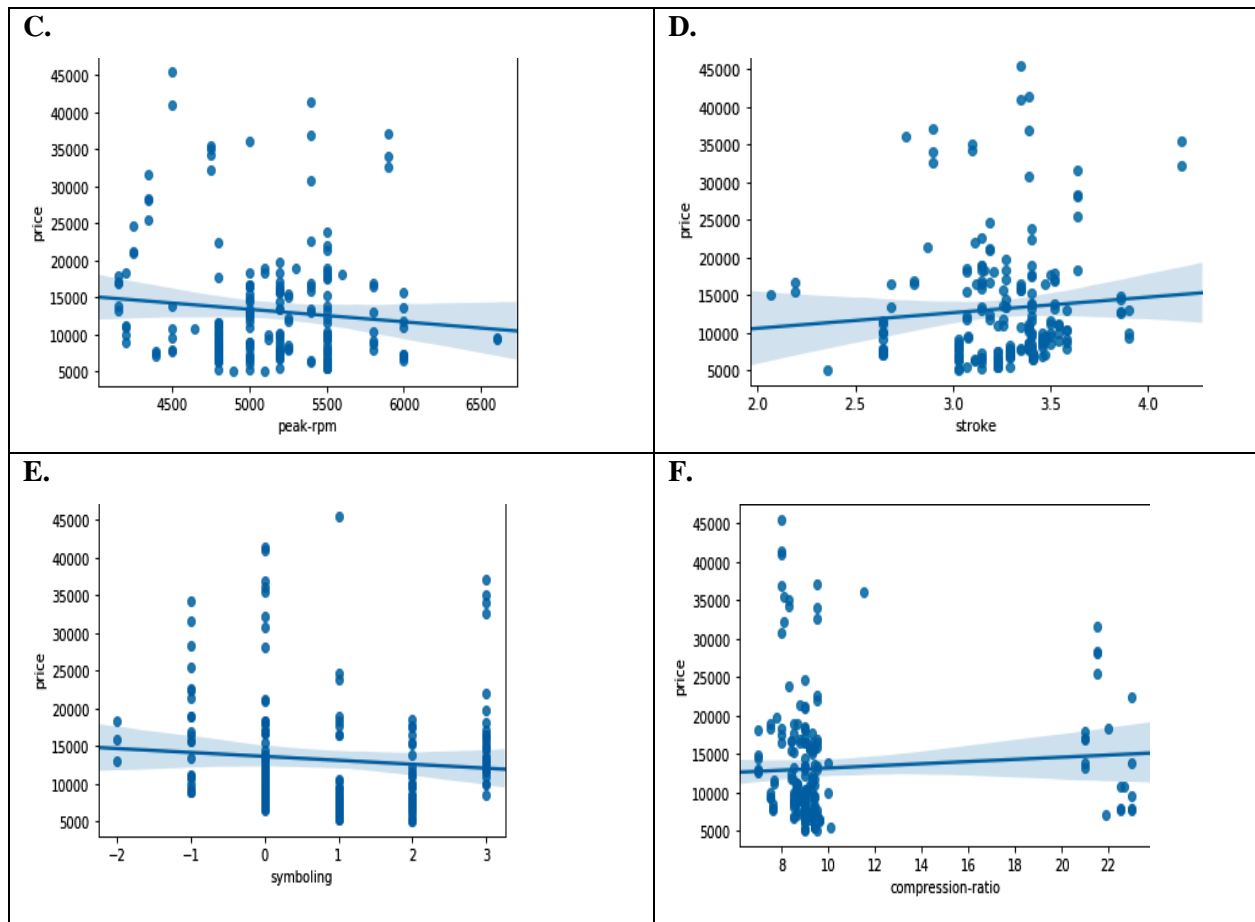
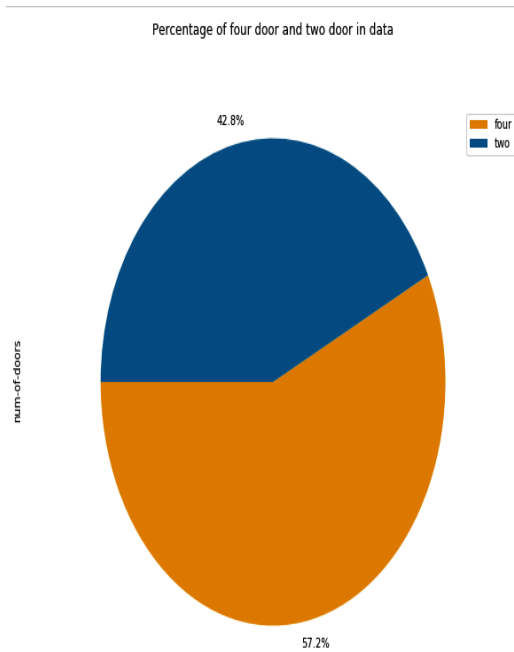


Figure 3.13- Weak linear relationship

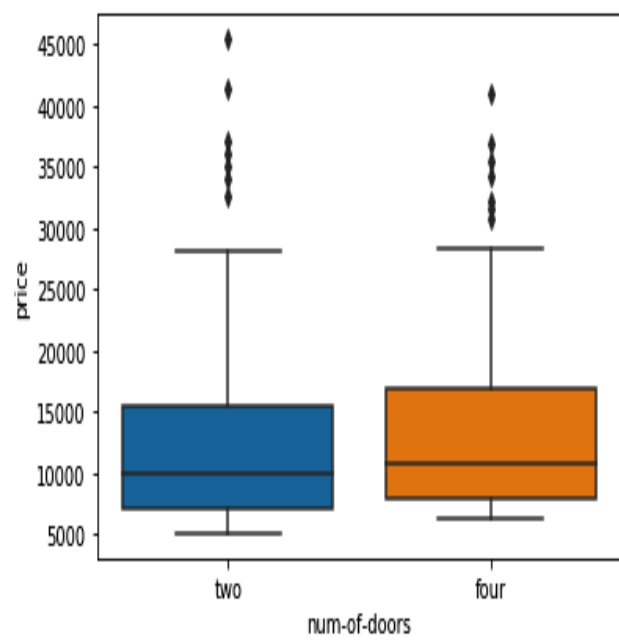
3.4.2. Analysis of Categorical Variables:

These are variables that describe a 'characteristic' of a data unit and are selected from a small group of categories. The categorical variables can have the type "object" or "int64". An excellent way to visualize categorical variables is by using boxplots. Let's look at the relationship between Categorical variables and "price" with the help of box plots and pie-chart to describe the percentage of labels in the variables.

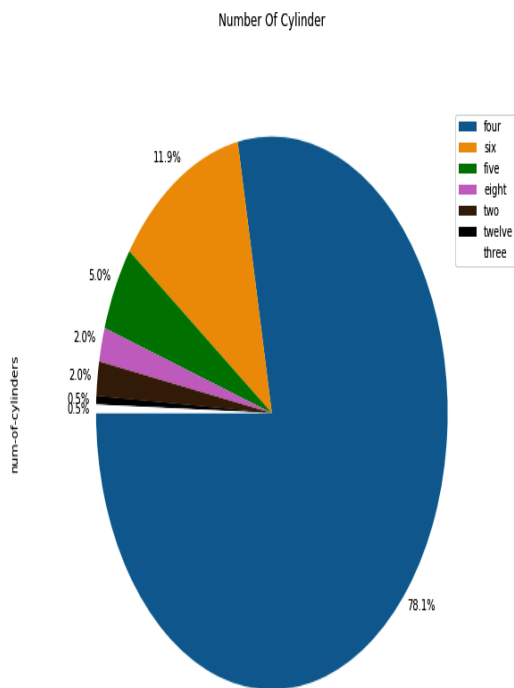
1.A



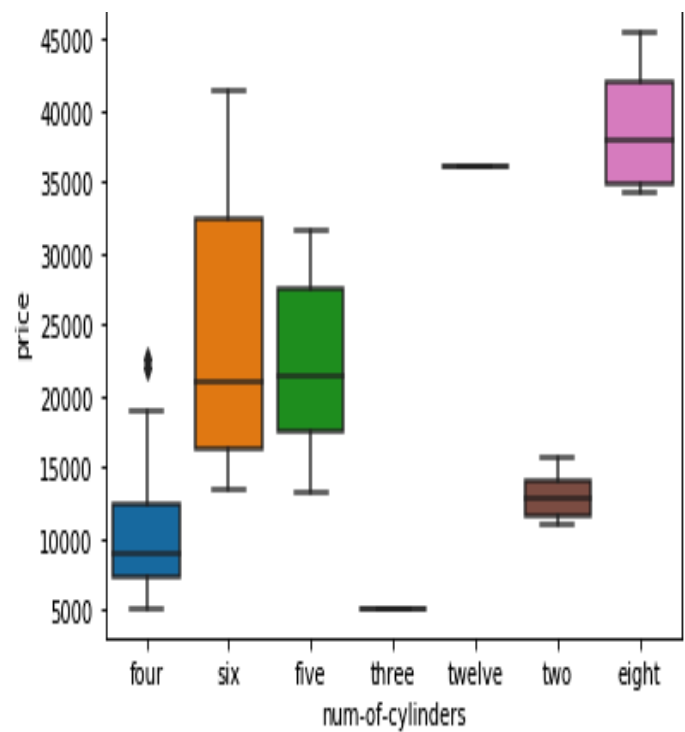
1.B



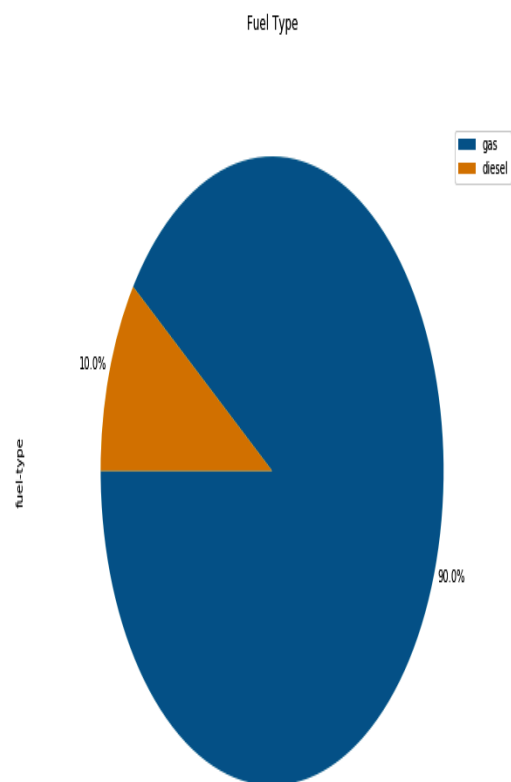
2.A



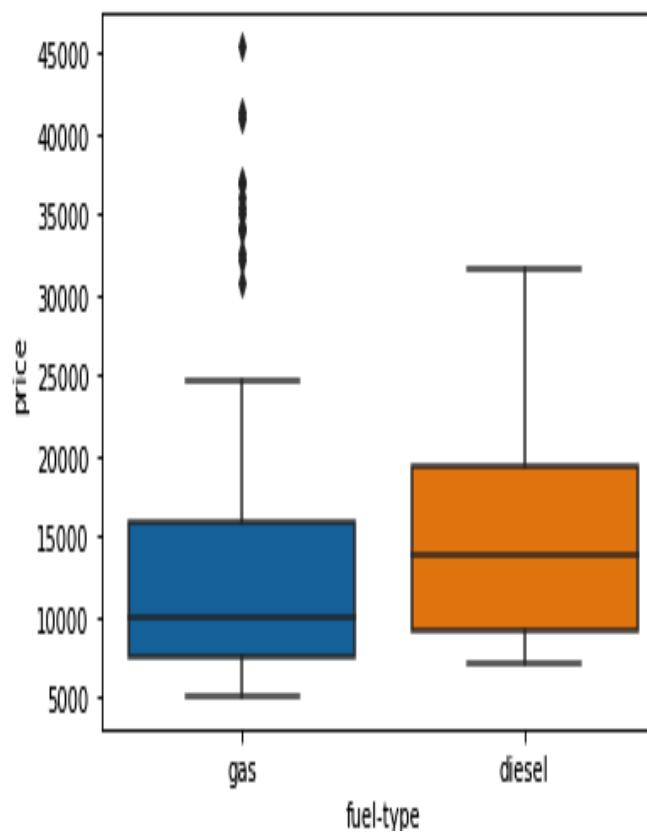
2.B



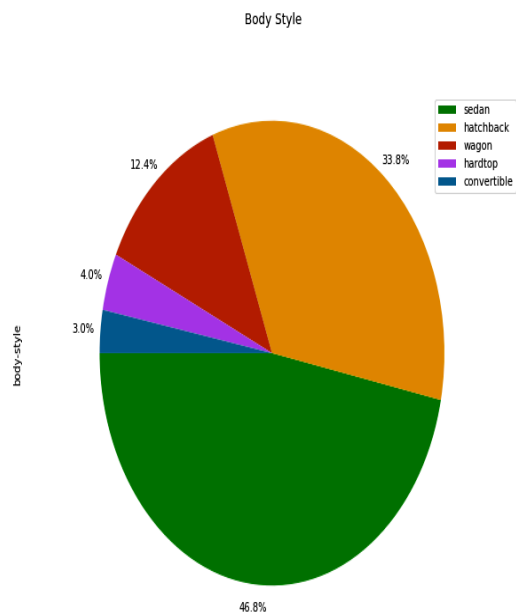
3.A



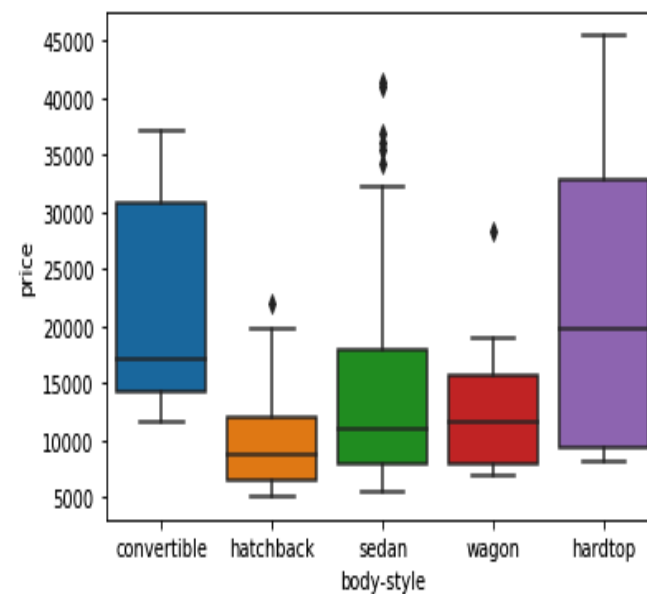
3.B

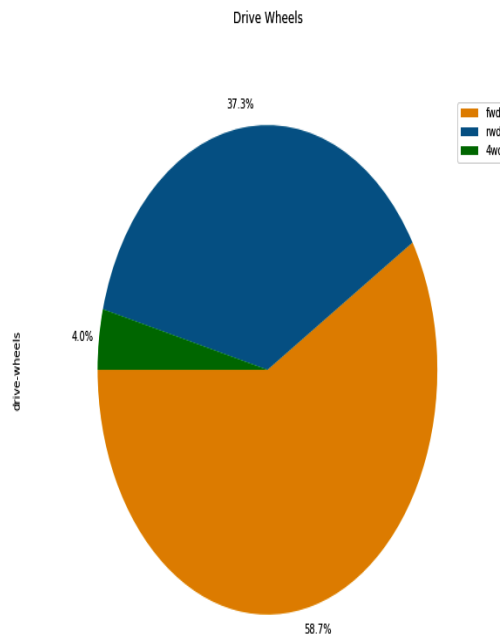
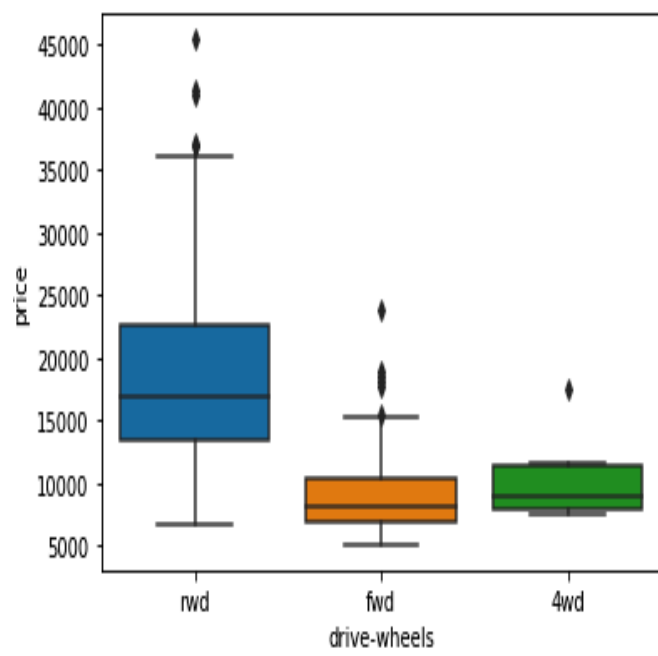
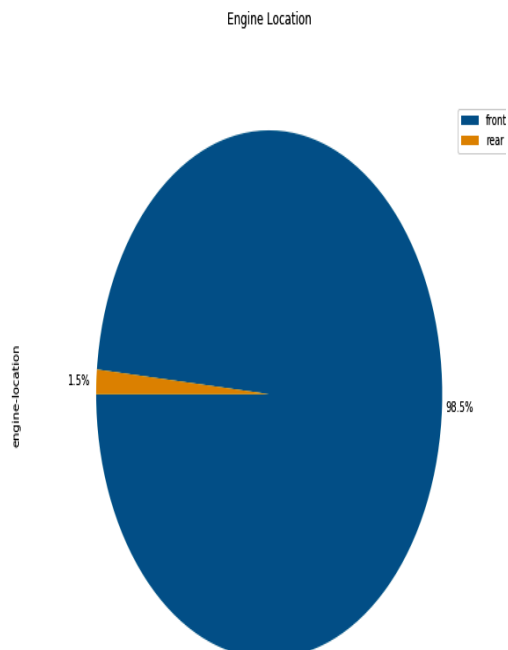
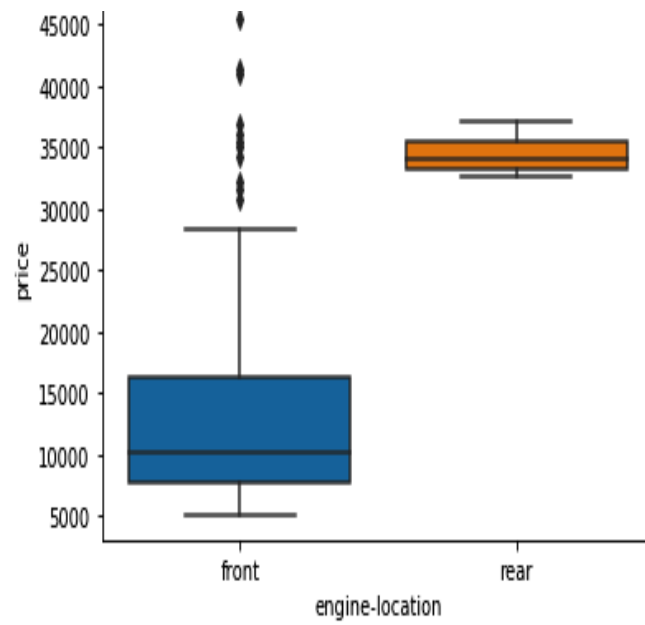


4.A



4.B



5.A**5.B****6.A****6.B**

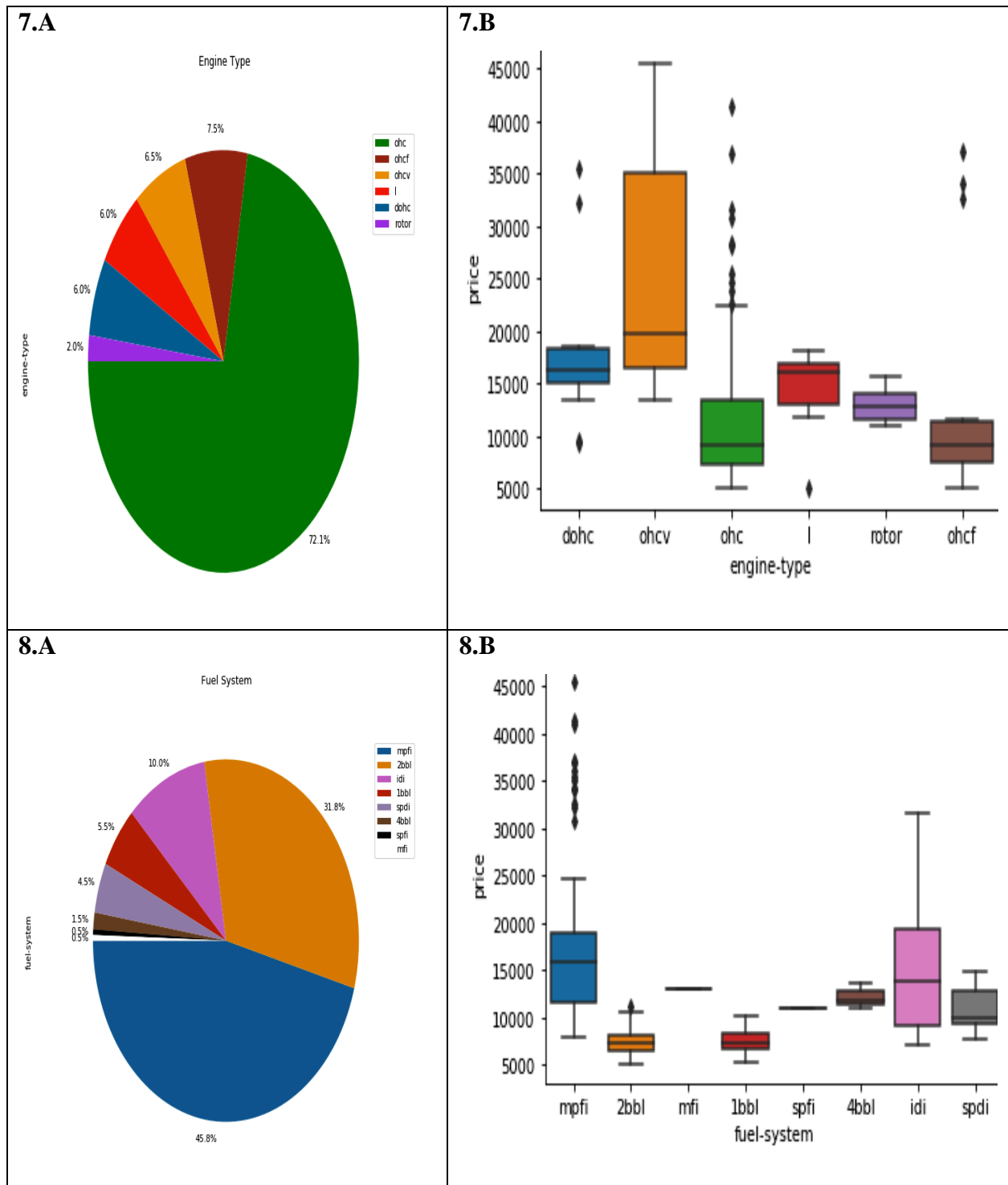


Figure 3.14- Pie-chart and box plots for illustrating the percentage of labels in the variables and the relationship between categorical variables and price:

As we see that the distributions of price with the different all categorical variable, we have found significant result. We will explain through following three cases-

- **Case-1:**

We see that the distributions of price with the most of the categorical variables (**fuel-system, engine-type, body-style, fuel-type, num-of-cylinder, num-of-door**) have significant overlap, and so these variables would not be a good predictor of price.

- **Case-2:**

Here we see that the distribution of price between these two **engine-location** categories, front and rear, are distinct enough but the frequency of the front type is so high with compare to the rear type, only 1.5% car have engine-location in the rear side. So engine-location also would not be a good predictor of price.

- **Case-3:**

Here, we see that the distribution of price between the different **drive-wheels** categories. We found that fwd and 4wd are overlapped, but rwd is distinct from fwd and 4wd. So, drive-wheels could potentially be a predictor of price.

3.4.3. An examination of price trend:

Price is the feature that we are predicting in this project. So, before applying any models taking a look at 'price' variables in the data. By looking at Figure 1, we have the following observation-

- The plot in the histogram is seemed to be right-skewed, meaning that the prices of the most cars in the dataset are low (less than \$20,000).
- There is a significant difference between the mean(\$13,207) and the median(\$10,295) of the price distribution.
- The data points are far spread out from the mean, which indicates a high variance in the prices.(87% of the prices are below \$20,000, whereas the remaining 13% are between \$20,000 and \$45,400.)

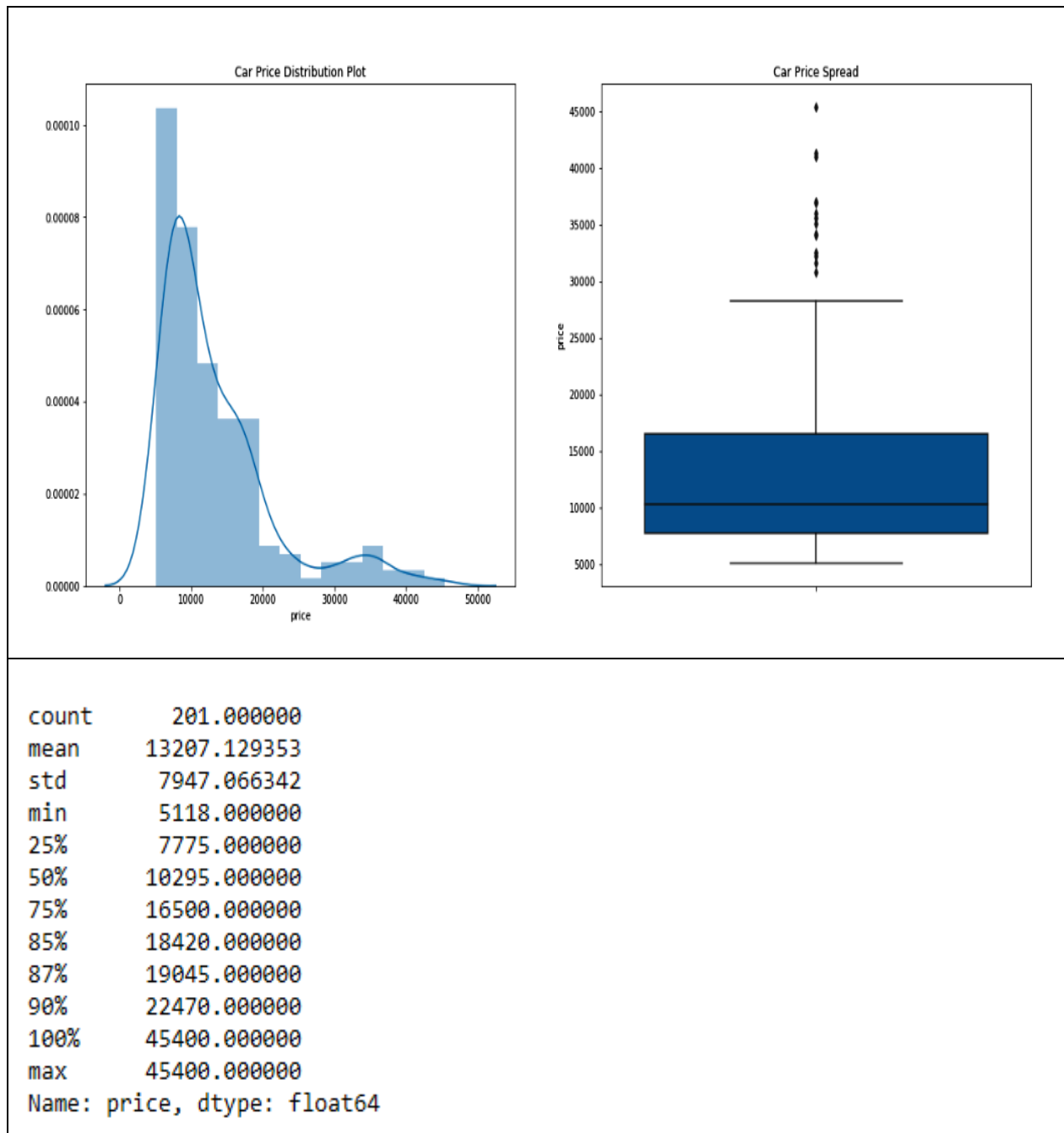


Figure 3.15- Plots for illustrating the price trends through histogram and box plot and also show its percentile, mean and standard deviation

3.4.4. Conclusion: Important variables-

We now have a better idea of what our data looks like and which variables are important to take into account when predicting the car price. We have narrowed it down to the following variables:

- **Continuous numerical variables:**

- Length
- Width
- Curb-weight
- Engine-size
- Horsepower
- City-km/L
- Highway-km/L
- Wheel-base
- Bore

- **Categorical variables:**

- Drive-wheels

3.5. Data Preprocessing:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Before applying the modeling, we should preprocess the data. In our project, according to situation we have to apply following three data preprocessing method-

3.5.1. One Hot Encoding:

After applying analysis, we select 10 variables for development of the model. But one of them is categorical variables. In order to apply machine learning models, we need numeric representation of the features. Therefore, this non-numeric features was transformed into numerical form. For this, we will use the one hot encoding method. By using this method, we splits the three categories of drive-wheels variables into three separate columns named as 'rear-wheel drive' for rwd, 'four-wheel drive' for 4wd and 'front-wheel drive' for fwd and each column has a series of zeros and ones. '1' indicates any particular feature is present, and '0' indicates the feature is not present in the car.

e	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	stroke	compression-ratio	horsepower	peak-rpm	city-km/L	highway-km/L	price	four-wheel drive	front-wheel drive	rear-wheel drive
1-	gas	std	two	convertible	rwd	front	88.6	...	2.68	9.0	111	5000.0	8.94	11.49	13495.0	0	0	1
1-	gas	std	two	convertible	rwd	front	88.6	...	2.68	9.0	111	5000.0	8.94	11.49	16500.0	0	0	1
1-	gas	std	two	hatchback	rwd	front	94.5	...	3.47	9.0	154	5000.0	8.09	11.06	16500.0	0	0	1
di	gas	std	four	sedan	fwd	front	99.8	...	3.40	10.0	102	5500.0	10.21	12.77	13950.0	0	1	0
di	gas	std	four	sedan	4wd	front	99.4	...	3.40	8.0	115	5500.0	7.66	9.36	17450.0	1	0	0

Figure 3.16: The change of categorical variable into numerical variables

3.5.2. Data Standardization:

In our dataset, the fuel consumption columns "city-mpg" and "highway-mpg" are represented by mpg (miles per gallon) unit. We are developing this model for our country that accept the fuel consumption with km/L standard. So we have to transform 'mpg' into 'km/L' standard by using formula-

$$\text{Km/L} = \text{mpg} / 2.35$$

3.5.3. Data Normalization:

We are going to normalize the columns "length", "width" and "height" so their value ranges from 0 to 1.

Replace original value by (original value)/(maximum value).

Now, in the next section, we are going to move into building machine learning models to automate our analysis, feeding the model with variables that meaningfully affect our target variable will improve our model's prediction performance.

3.6. Model development:

In this section, we will develop several models that will predict the price of the car using the variables or features. It is just an estimate but should give us an objective idea of how much the

car should cost. For this purpose, I will use 4 Machine Learning models, Multiple Linear Regression, Polynomial Regression, Decision Tree Regression and Random Forest Regression. In the end, I will compare r^2_score and RMSE for each model and select the best model, which have high r^2_score and low RMSE value. We also go through cross-validation for evaluation of the best model.

3.6.1. Data splitting:

Before going through the modeling process, we randomly split our data into training and testing data using the function **train_test_split**. In this process, 75% of the data was split for the train data and 25% of the data was taken as test data.

Now we applied our four machine learning models and visualize these models through distribution plot and scatter plot. In the scatter plot, have actual price (in \$) on the x-axis and predicted price (in \$) on the y-axis. We draw these plots between actual values and predicted values for both train dataset and test dataset.

3.6.2. Model -1: Multiple Linear Regression:

We have more than one variable in our model to predict car price. So we can use **Multiple Linear Regression**. It is very similar to Simple Linear Regression. The visualization of our model is shown below-

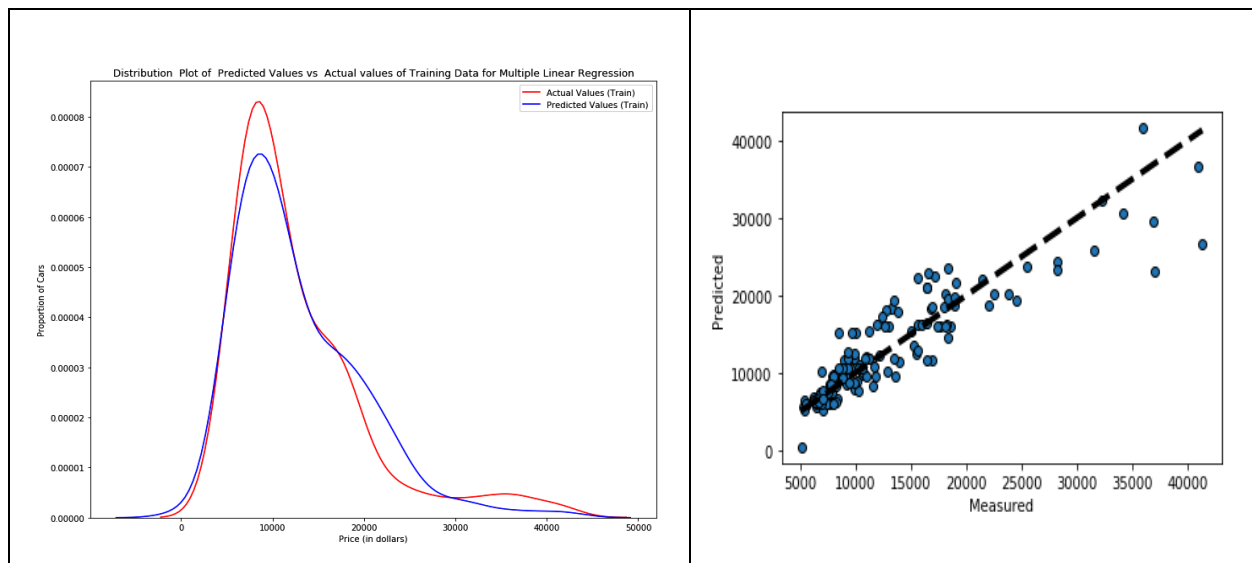


Figure 3.17 - Distribution plot and Scatter plot between actual value and predicted value of train data for Multiple Linear Regression

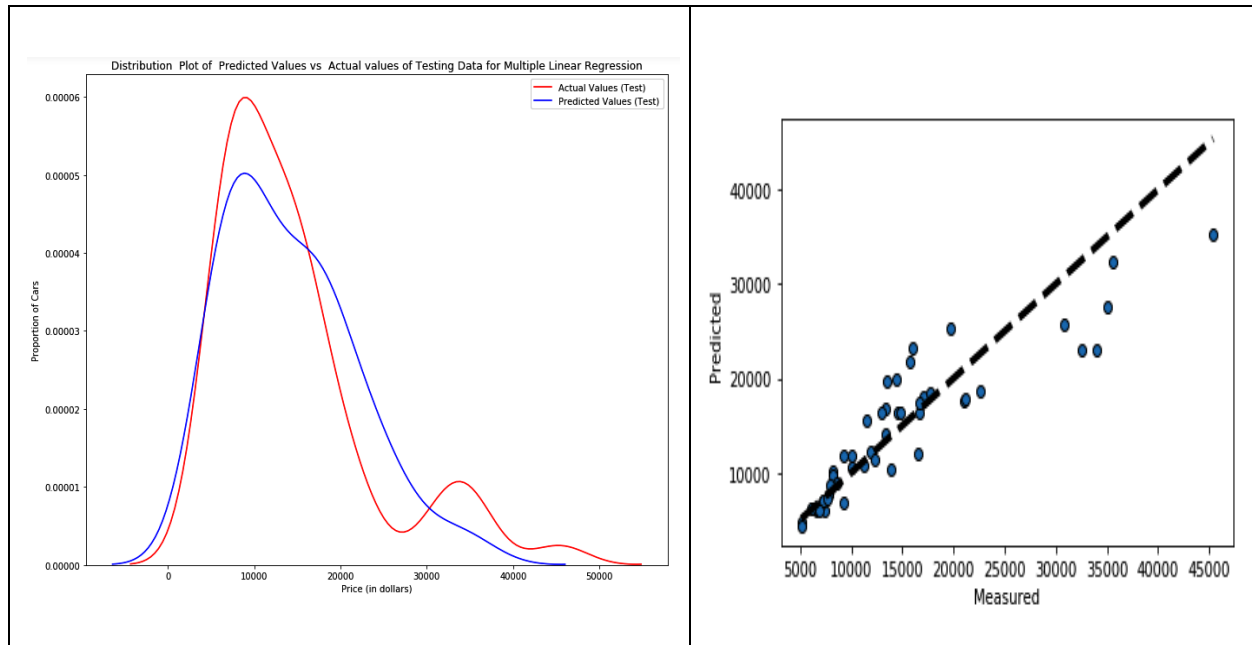


Figure 3.18 - Distribution plot and Scatter plot between actual value and predicted value for test data for Multiple Linear Regression

Inference- After fitting the model for the training dataset shown in figure 3.6, we can understand that the fitted values are reasonably close to the actual values for lower values of the price (\$0 to \$500) and for the higher value of the price it is not fitted well. When the model generates new values from the test data (figure 3.7), we see the distribution of the predicted values is much different from the actual target values. So, there is some work for the improvement of the model.

3.6.3. Model -2: Polynomial Regression:

If the data shows a curvy trend, then linear regression would not produce very accurate results when compared to a non-linear regression. Simply because, as the name implies, linear regression presumes that the data is linear. Polynomial regression fits a curve line to your data. The visualization of our polynomial regression model with the predicted value and actual value is shown below-

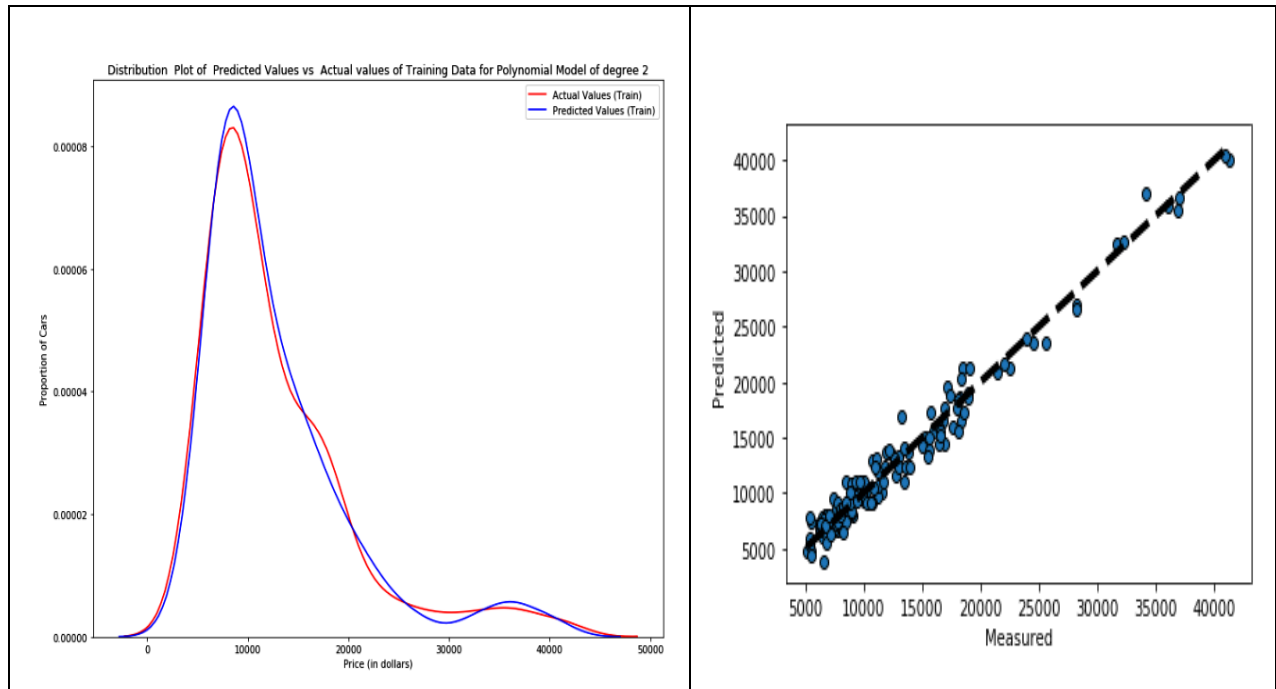


Figure 3.19 - Distribution plot and Scatter plot between actual value and predicted value for train data for Polynomial Regression

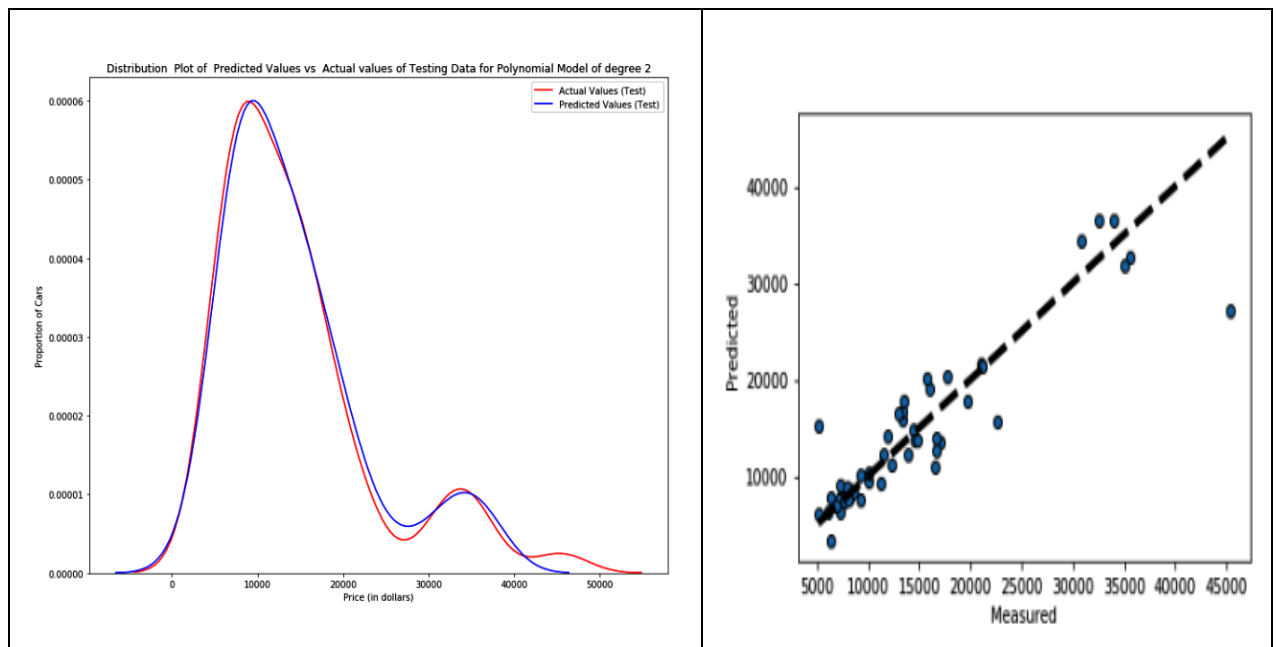


Figure 3.20 - Distribution plot and Scatter plot between actual value and predicted value for test data for Polynomial Regression

Inference- In figure 3.8, for the training dataset, the data was fitted very well for a low price between (\$0 to 800\$), and after that, it shows some deviation. The model seems to be doing well in learning from the training dataset, and it looks better than the multiple linear regression model. When we put the new values from the test data (figure 3.9), we can see the distribution of the predicted values is much closer to the actual target values. We want the best model; therefore, we build two other supervised machine learning model in the next two sections.

3.6.4. Model -3: Decision Tree Regression:

This model builds regression models in the form of a tree structure. It breaks down a dataset into the smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. The visualization of our decision tree regression model for our dataset with the predicted value and actual value is shown below-

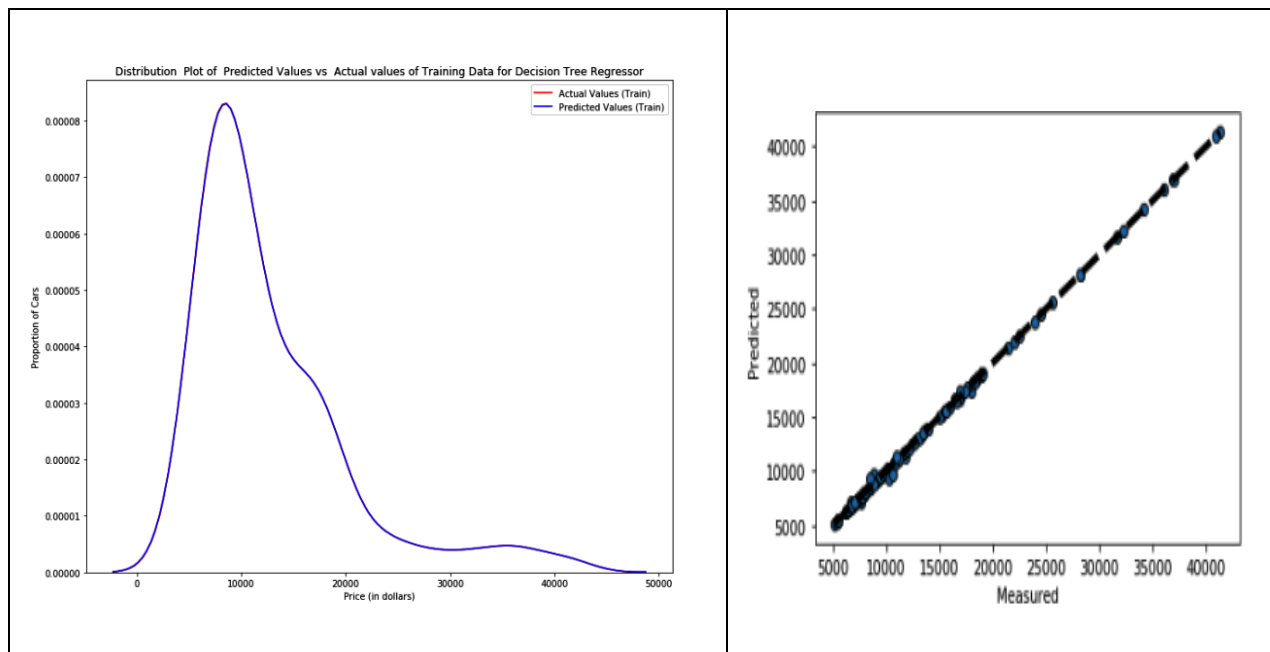


Figure 3.21 - Distribution plot and Scatter plot between actual value and predicted value for train data for Decision Tree Regression

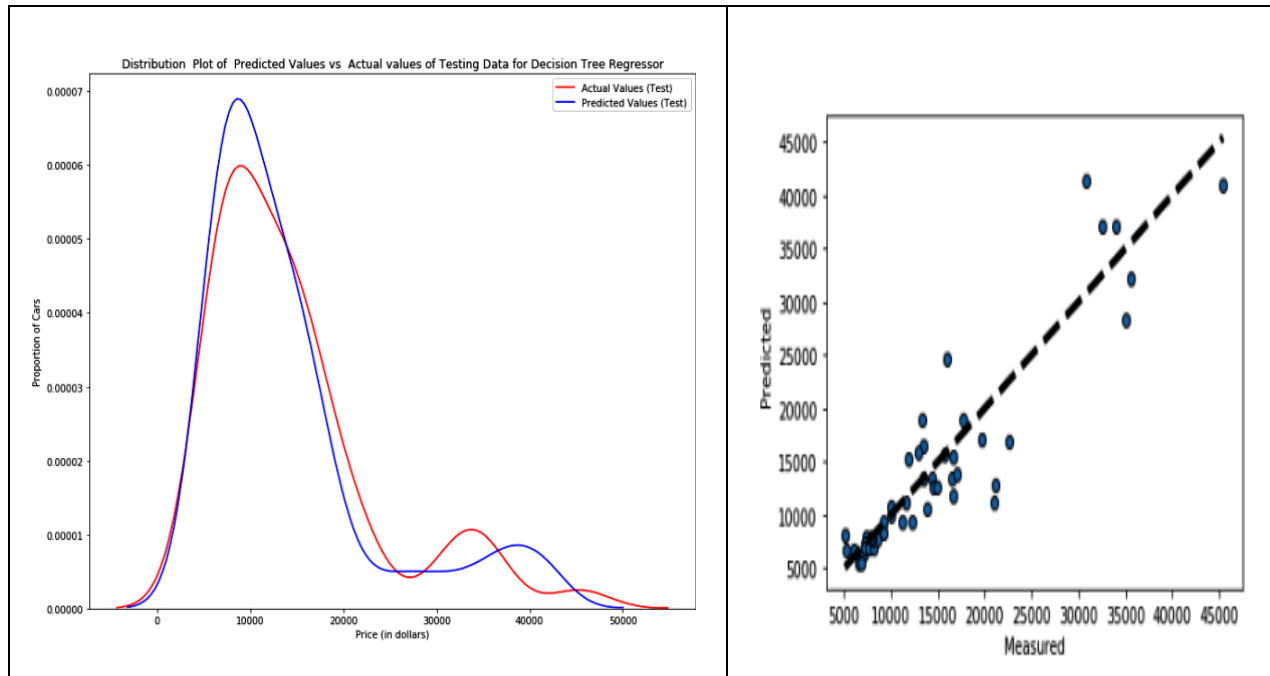


Figure 3.22 - Distribution plot and Scatter plot between actual value and predicted value for test data

Inference- In figure 3.10, for the training dataset, the predicted value and the actual value is fully overlapped to each other in the distribution plot. In the right plot (regression plot), all the values are situated on the straight line in the regression plot and show all the predicted price is equal to their correspondent actual price. So far, the model seems to be doing well with the training dataset. But when the model encounters new data from the testing dataset (figure 3.11), we see the distribution of the predicted values is much different from the actual target values. So, we can say this model shows the overfitting.

3.6.5. Model -4: Random Forest Regression:

Our forth model is random forest regression model. A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models. We randomly perform

row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

The visualization of our random forest regression model for our dataset with the predicted value and actual value is shown below-

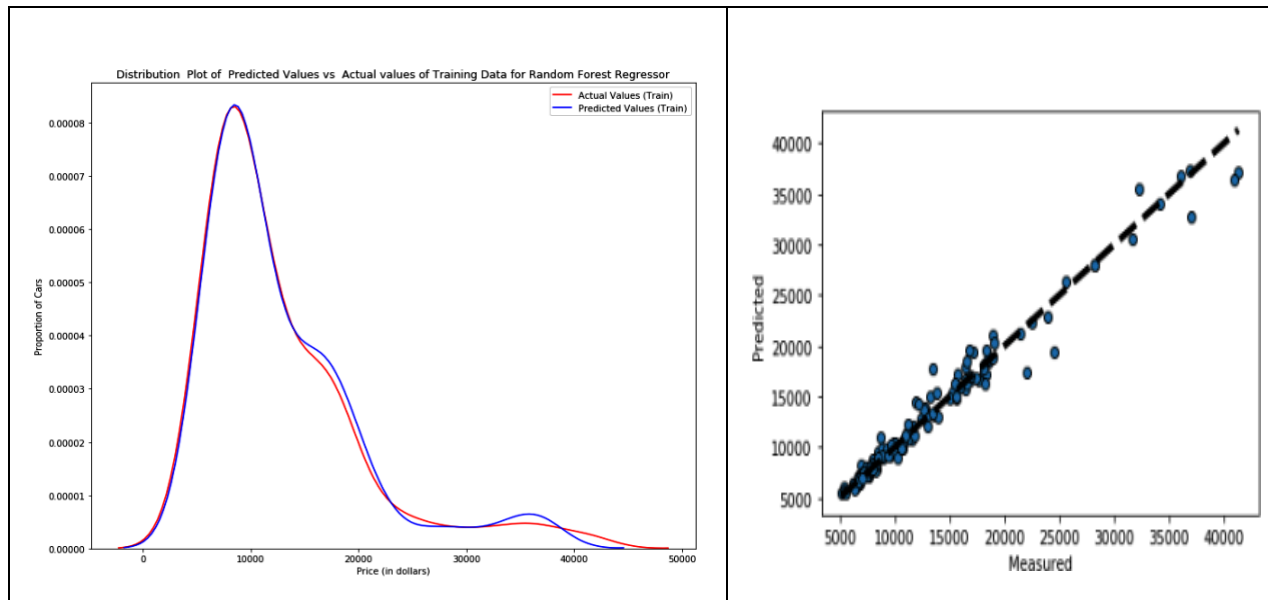


Figure 3.23 - Distribution plot and Scatter plot between actual value and predicted value for train data and Random Forest Regression

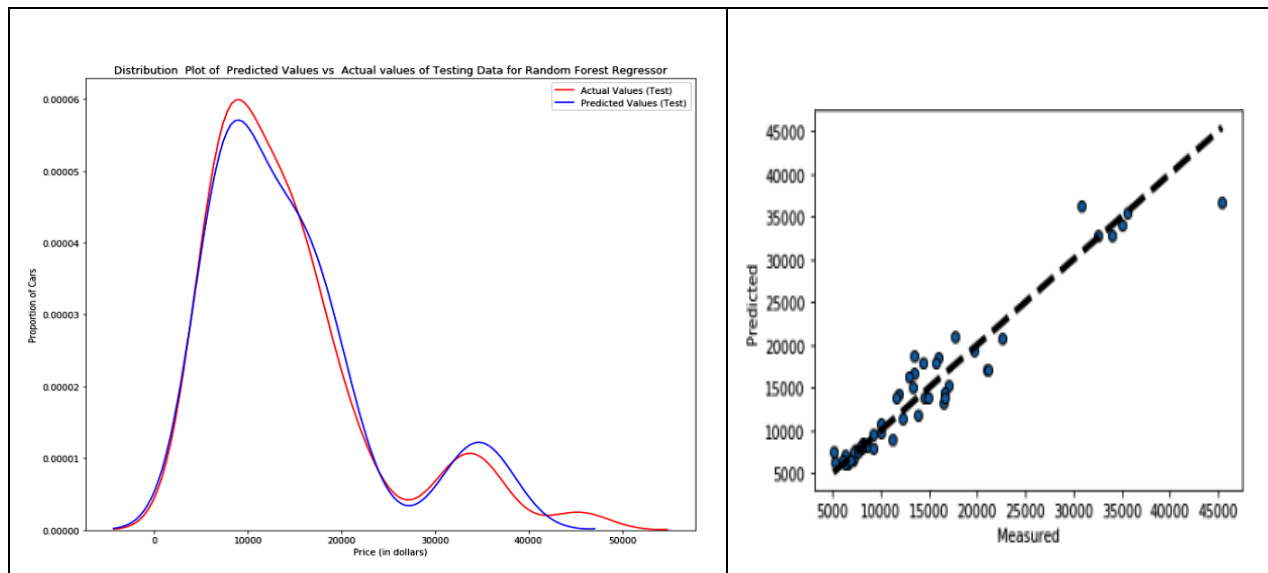


Figure 3.24 - Distribution plot and Scatter plot between actual value and predicted value for test data for Random Forest Regression

Inference- In figure 3.12, for the training dataset, the predicted value and the actual value is almost overlapped to each other in the distribution plot, there are very minor separation between them for the price more than \$1800. In the right plot (regression plot), the price which has value less than \$1800 are very close to the regression line, only very few values are away from the line which have price value more than \$1800 in the regression plot. So far, the model is also seems to be doing well in learning from the training dataset like polynomial regression.

Now we are too confused to select the best model, which gives the best prediction because we cannot all plot looks very. Now we want a quantitative measure to determine how accurate the model is. In the next section (model evaluation), we do all things for the selection of the best model.

3.7. Model evaluation:

In this section we will evaluate all the models by using quantitative measure. Now we are going to use two very important measures that are often used in Statistics to determine the accuracy of a model are:

- R^2 / R-squared
- Root Mean Squared Error (RMSE)

R-squared:

R squared, also known as the coefficient of determination, is a measure to indicate how close the data is to the fitted regression line.

Root Mean Squared Error (RMSE):

The Root Mean Squared Error measures the root of the average of the squares of errors, that is, the difference between actual value (y) and the estimated value (\hat{y}).

3.7.1. General Evaluation:

We have four different models, and we are going to generate R-squared and RMSE values for each model in general way. For determine a good model fit, we will focus on two points-

A. The model with the higher R-squared value is a better fit for the data.

B. The model with the smallest RMSE value is a better fit for the data.

The table shown below, describe the R-squared value and RMSE value for both training data and testing data for all the models. Now it will be easy for us to evaluate the best model, because it has quantitative value.

Model Name	R ² Score	RMSE	
		Train Data	Test Data
Multiple Linear Regression Model	0.8261	3117.2174	3822.5342
Polynomial Regression Model	0.9729	1214.0961	3788.8739
Decision Tree Regressor Model	0.9994	174.3885	3328.8639
Random Forest Regressor Model	0.9804	1011.4426	2684.2192

Table 3.1: R-squares and RMSE value for each model

Inference: As we can see in the table Decision tree regression is the best model for the training dataset because it has a lowest RMSE value and a highest R-squared value while Random forest model good for the test dataset. So, we are not able to get the conclusion that which model is correct. For getting the best model, we will adopt another way of model evaluation. The method that we will use is K-fold cross-validation.

3.7.2. Model Evaluation by using K-fold Cross-Validation:

Cross-validation is a re-sampling procedure used to evaluate machine learning models on a data sample. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. Here we split the data into four equal folds; it may be called 4-fold cross-validation. We use three parts out of four to train the model and use one remaining part for test the model. So, we use each and every fold to test the model and the remaining three parts for training the model. Therefore we have four R-

squared values and four RMSE values for each model. We calculate the mean of both R-squared values and RMSE values for a model. After that, we compare all the scores of the four models and select the best model.

Following tables show the R-squared values and RMSE values for each fold of all models and their mean and standard deviation.

S N.	Model Name	R ² score for 1 st fold	R ² score for 2 nd fold	R ² score for 1 rd fold	R ² score for 4 th fold	Mean	Standard Deviation
1.	Multiple Linear Regression	0.8464	0.7985	0.4373	-0.2554	0.4567	0.4405
2.	Polynomial Regression	0.8464	0.7985	0.4373	-0.2554	0.4567	0.4405
3.	Decision Tree Regression	0.8546	0.8515	0.7886	-0.0636	0.6078	0.3886
4.	Random Forest Regression	0.8774	0.9075	0.5764	0.6680	0.7574	0.1394

Table 3.2: R² score, mean and standard deviation for each model in cross-validation

S N.	Model Name	RMSE Score for 1 st fold	RMSE Score for 2 nd fold	RMSE Score for 1 rd fold	RMSE Score for 4 th fold	Mean	Standard Deviation
1.	Multiple Linear Regression	3715.30	4272.06	5289.93	4813.52	4522.70	589.04
2.	Polynomial Regression	3715.30	4272.06	5289.93	4813.52	4522.70	589.04
3.	Decision Tree Regression	3510.19	3836.52	3775.33	4393.58	3878.90	321.46
4.	Random Forest Regression	3293.66	3415.75	3332.00	2752.24	3198.41	261.35

Table 3.3: RMSE values, mean and standard deviation for each model in cross-validation

Inference: As we saw both tables show above, we can understand that random forest model has the lowest mean of RMSE value and the highest mean of R-squared score for both training and testing dataset among all four models.

3.8. Save Model:

After all the process given above, we save the model which has the lowest RMSE value and the highest mean of R-squared score. We use this saved model for future uses for the prediction of unknown data by using joblib library function of python.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction:

After implementing all the process in chapter 3, we will discuss the result in this chapter. We have performed four machine learning techniques (Multiple Linear Regression, Polynomial Regression, Decision Tree Regression Model and Random Forest) on our automobile dataset. These models can predict the price of the car. We evaluated all the models by finding their R-squared score and RMSE values by applying test dataset. We also use K-fold cross-validation for the evaluation process. After that, we discovered that our random forest regression model fitted very well and has the lowest RMSE value (**3198.41**) and the highest mean of R-squared score (**0.7574**) among all the models.

4.2 Test The Model With New Dataset:

Now we apply new test data to our model and predict the price. We compare the predicted price with the actual price with the help of the distribution plot and scatter plots to check the accuracy of our model.

4.3 Visualization of Results:

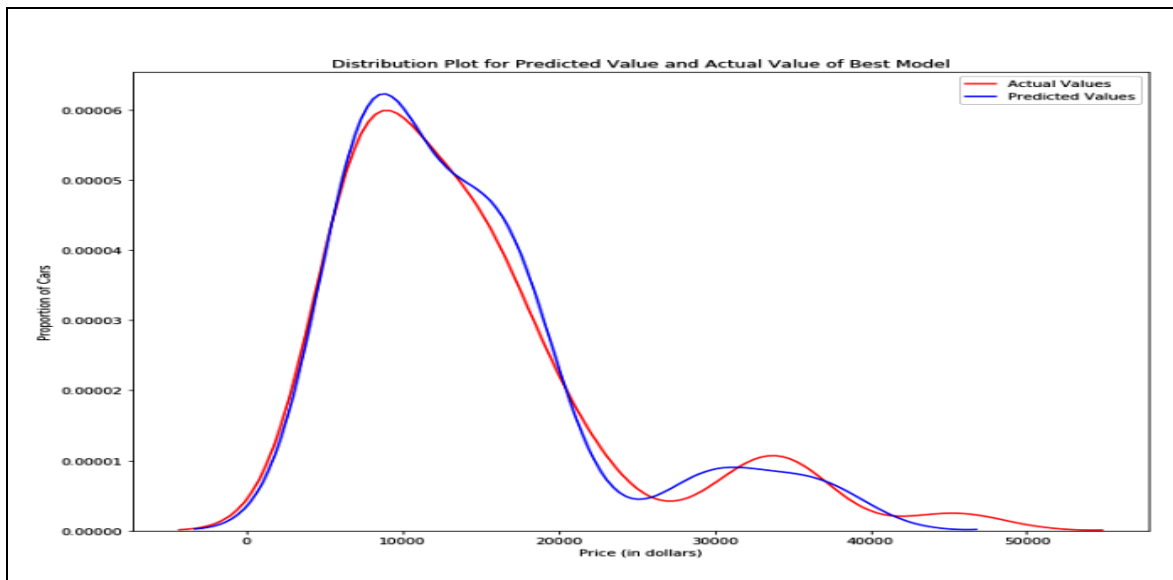


Figure-4.1: Distribution plot for the predicted values and the actual values for model testing

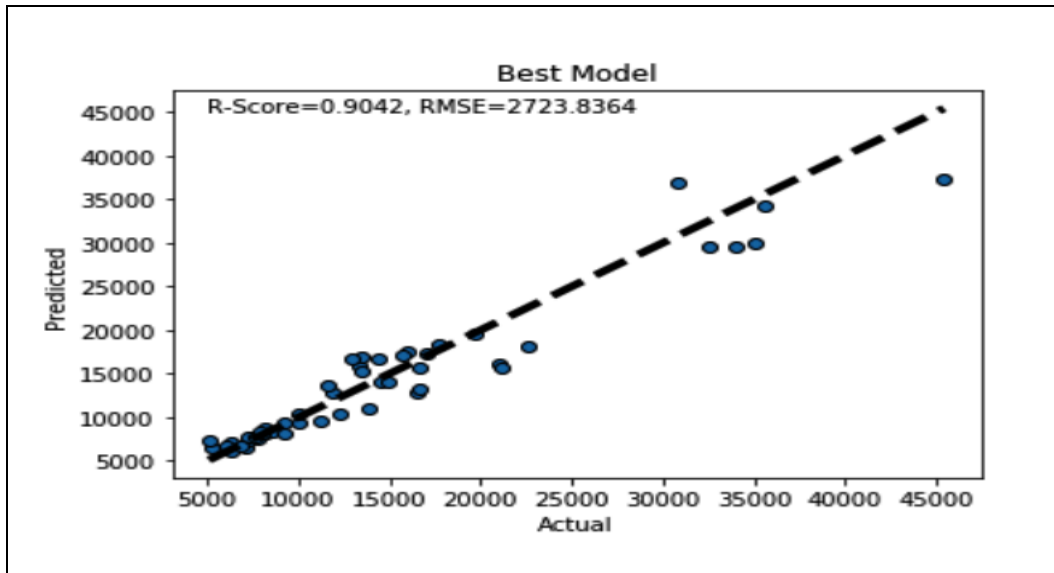


Figure-4.2: Scatter plot of the predicted value with respect to actual value of a particular car in the model testing

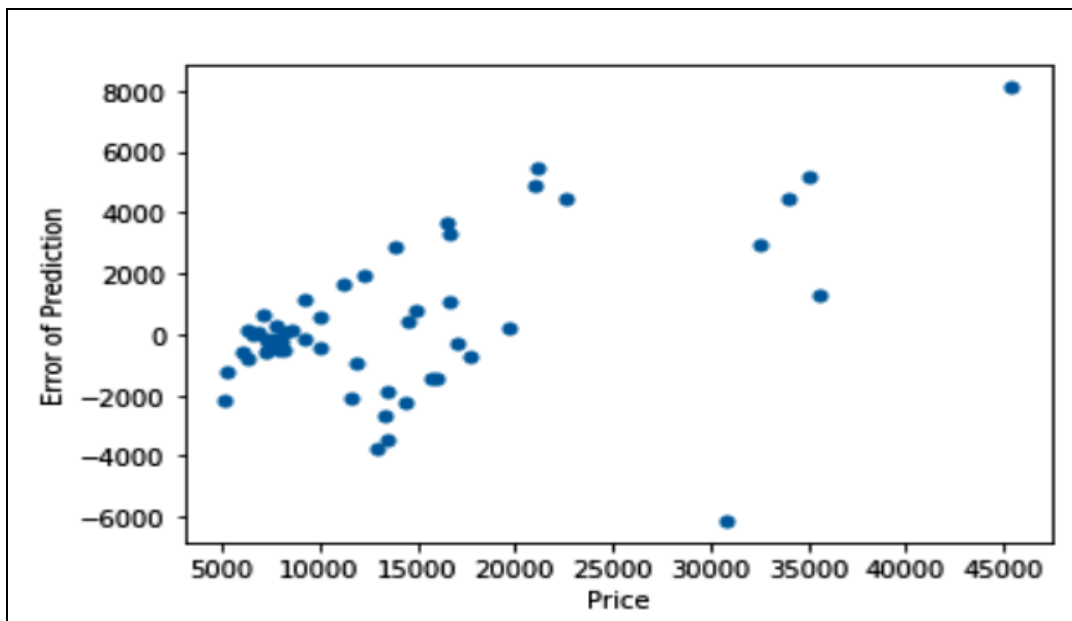


Figure-4.3: Scatter plot for the difference between the predicted value and the actual value of a particular car n the model testing

R-squared Score	0.8969
RMSE	2573.0688

Table-4.1: R² score and RMSE value for test dataset

4.4 Discussion On Result

After applying the test data and analyzing the above figure, we can conclude the following results-

- A.** As we can see in the distribution plot (figure-4.1), the fitted values are very close enough to the actual values
- B.** In the scatter plot shown in the figure-4.2, have the actual price (in \$) on the x-axis and predicted price (in \$) on the y-axis. All dots show the predicted price corresponding to its actual price. After analyzing this plot, we observed that the most of the values (or dots) are either lying on the straight line or too close to it, too few of them are at a very short distance from the line.
- C.** In the scatter plot shown in the figure-4.3, have the actual price (in \$) on the x-axis and the difference between the actual price and predicted price (in \$) on the y-axis. After analyzing this plot, we observed that most of the values (or dots) is very close enough to the zero on the y-axis. Maximum dots show the error between -2000 to 2000 in price. Too few of them have error out of the range -2000 to 2000.
- D.** This model has an R-squared score is 0.8969, and RMSE value is 2573.0688. It is perfect for a good model

After observing the above results, we can say that our model works efficiently and predict the price nearly equal to the actual price of most of the cars.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion:

In this project, I tried to build a model by using data science process. This model can predict the price of cars based on its features or attributes. In the process of building the model, at the first stage, we collect the data, cleaned the data and preprocessed. With the help of the exploratory data analysis and the data visualizations, we explored the data deeply and uncovered the hidden pattern in the data. We find a meaningful relationship between features. After examined the data, we found 12 crucial features out of 26 that play a significant role in deciding the price of the cars. At the last stage, we applied four machine learning models to predict the price of cars in order: multiple linear regression, polynomial regression, random forest regression and decision tree regression. By evaluating all four models, it can be concluded that the random forest regression model is the best model for the prediction for the price of cars. Random Forest, as a regression model, gave the minimum RMSE and maximum R-squared score values.

This model works efficiently for a given data. This model can be used for business purpose. It will give benefit to both buyer and seller, and save money and time of the user.

5.2 Limitations of the Study and Suggestions for Further work:

The main limitation of this project is the low number of data that have been used. We use only 205 data. Gathering more data can yield more robust predictions. As a suggestion for future work, we intend to use more advanced techniques like artificial neural networks, fuzzy logic and genetic algorithms to predict car prices. We can add some other advance features and techniques like CNN and computer vision to predict the price of automobiles other than cars and so on.

REFERENCES:

- [1] Gareth, J., Daniela, W., Trevor, H., & Tibshirani, R. (2013). An Introduction to Statistical
- [2] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In *Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on* (Vol. 2, pp. 682-685). IEEE.
- [3] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, 4(7), 753-764.
- [4] Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. *International Journal of Computer Applications*, 167(9), 27-31.
- [5] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.), Retrieved from: <https://www.cs.waikato.ac.nz/ml/weka/>. [August 04, 2018].
- [6] Ho, T. K. (1995, August). Random decision forests. In *Document analysis and recognition, 1995.proceedings of the third international conference on* (Vol. 1, pp. 278-282). IEEE.
- [7] Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. *Journal of machine learning research*, 2(Dec), 125-137.
- [8] http://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf
- [9] <https://en.wikipedia.org/wiki/Prediction>
- [10] https://www.researchgate.net/publication/332072545_Car_Sales_Prediction_Using_Machine_Learning_Algorithms
- [11] <https://gallery.azure.ai/Experiment/Automobile-price-prediction-256>
- [12] <https://towardsdatascience.com/build-develop-and-deploy-a-machine-learning-model-to-predict-cars-price-using-gradient-boosting-2d4d78fddf09>
- [13] https://en.wikipedia.org/wiki/Data_analysis
- [14] <https://machinelearningmastery.com/k-fold-cross-validation/#:~:text=Cross%2Dvalidation%20is%20a%20resampling,k%2Dfold%20cross%2Dvalidation.>
- [15] https://www.saedsayad.com/decision_tree_reg.htm
- [16] <https://www.geeksforgeeks.org/random-forest-regression-in-python/>
- [17] <https://www.geeksforgeeks.org/random-forest-regression-in-python/>

PLAGIARISM REPORT



Document Information

Analyzed document	17419MCA020.pdf (D76375568)
Submitted	7/15/2020 1:15:00 AM
Submitted by	
Submitter email	msalmanar.777@gmail.com
Similarity	18%
Analysis address	cenlib2014.bhuni@analysis.urkund.com