**CAR PRICE PREDICTION**

Submitted by

   YUVARAJ V  1P20CS023

In partial fulfilment of the requirement

for the award of the Degree of

**Master of Science in Computer Science**

**Bharathiar University**

**Coimbatore**.

Under the Internal Supervision of

**Dr.S.Suganya. MCA., Ph.D**

**Associate Professor**

**School of Computer Studies-PG**



**Master of Science in Computer Science,**

**Rathnavel Subramaniam College of Arts and Science (Autonomous),**

**Sulur, Coimbatore – 641 402**

**October 2021**

**RATHNAVEL SUBRAMANIAM COLLEGE OF ARTS AND SCIENCE**

**(AUTONOMOUS)**

**AFFLIATED TO BHARATHIAR UNIVERSITY**

**NAAC Re-accredited & ISO 9000:2008 Certified**

**SULUR, COIMBATORE-641402.**



**October 2021**

**Register Number –**

Certified bona fide original record of work done by **-----------------**

**Internal Supervisor                                                                        HOD**

Submitted for the Project Evaluation and Viva voce held on**………………….**

**Internal Examiner                                     External Examiner**

**CERTIFICATE**

This is to certify that the project entitled **“car price prediction”**, submitted by the Department Of Computer Science Rathnavel Subramaniam College of Arts and Science   in partial fulfilment of the requirement for the award of the Degree of Master of Computer Application  is a record of original project work done by **YUVARAJ V (Reg. No: -1P20CS023)** during the period July 2021 – October 2021 of  **her** study in the Department of Computer Science, under my Internal supervision and the project has not formed the award any Degree/Diploma/Associate ship/Fellowship or other similar title to any candidate of any University.

**Internal Supervisor**

       [Dr.S.Suganya]

**DECLARATION**

I am **YUVARAJ V** hereby declare that the project work entitled **CAR PRICE PREDICTION** Submitted to the Department of Computer Science RathnavelSubramaniam College of Arts and Science  in partial fulfilment of requirement for the award of the   Degree of  Master Computer  Application  is a record  of   original   done by me during the period of July 2021-October 2021  under the supervision of  **Dr.S.Suganya,** Associate Professor, School of Computer Studies – PG in RathnavelSubramaniam College of Arts and Science and the project have not formed the basis for the award of any Degree/Diploma/Associate ship/Fellowship or other similar title to any candidate of any University.

**Signature of the Candidate**

**ABSTRACT**

The objective of this analysis is Car Price prediction which aims to identify the relationship between the car price in USA Market and other many factors. In addition, predict the selling price of Cars. Among these, the influence of Company-name, model-name, and fuel-type, engine location, fuel system, highway or city mpg has been highlighted in the diverse field of cars research. Company\_name is the most valuable variable to predict the selling price of a car . The project involves predictions using two  regression techniques Decision tree and Random forest. Car price prediction on a dataset has been done by using the above mention techniques to find out the best among the two.  The motive of this project is to help the US market estimate the car perfectly and to help to predict the exact price. The built model has evaluated with test data and verified that the model predicted the selling price with minimum error rate.

**TABLE OF CONTENTS**

**CHAPTER-I  INTRODUCTION PAGE.NO**

1.1 Predictive Analytics 7

1.2 Overview 7

**CHAPTER-II            GATHERING DATA**

2.1 Data Description 9

2.2 Data Understanding 13

**CHAPTER-III          EXPLORATORY DATA ANALYSIS** 18

**CHAPTER-IV          MODEL BUILDING**

          4.1 Choosing the model 21

          4.2 Training and test dataset 21

          4.3 Model 21

**CHAPTER-V           EVALUATION OF MODEL**

                                  5.1 Model Evaluation 27

**CHAPTER-VI        PREDICTION AND INFERENCE**

6.1 Conclusion 28

**CHAPTER I**

**INTRODUCTION**

**1.1 PREDICTIVE ANALYTICS**

* Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning.
* Companies employ predictive analytics to find patterns in this data to identify risks and opportunities. Predictive analytics is often associated with big data and data science. Companies today are swimming in data that resides across transactional databases, equipment log files, images, video, sensors or other data sources.
* To gain insights from this data, data scientists use deep learning and machine learning algorithms to find patterns and make predictions about future events. These include linear and nonlinear regression, neural networks, support vector machines and decision trees. Learnings obtained through predictive analytics can then be used further within prescriptive analytics to drive actions based on predictive insights.

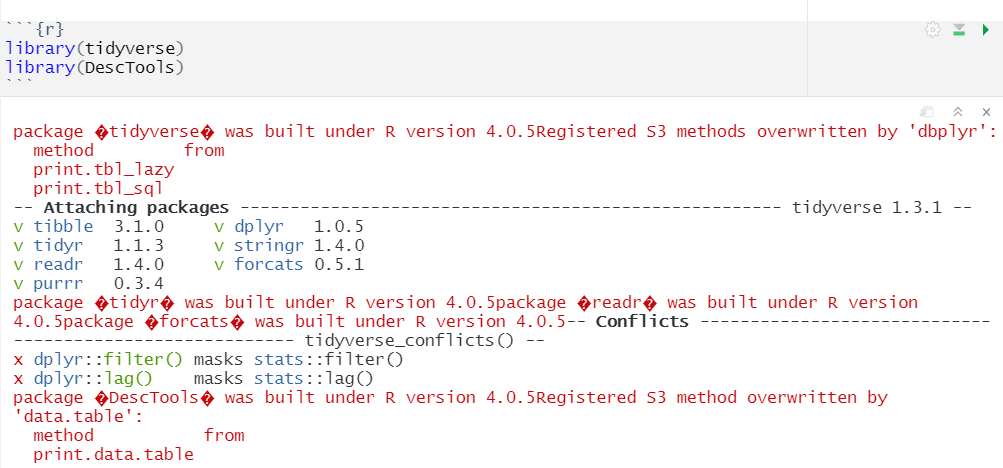
**1.2 OVERIVEW OF THE PROJECT**

* This project is developed the increasing number of cars sold day by day, it has become difficult to manage ( or) extract useful information from the available data of all the cars. The Geely\_data is utilized to keep the data related to cars price prediction. This data is then used for visualizing the specifications of cars which is available in US market.
* Additionally, The data is used to predict the selling price of the cars through various machine learning approaches. The proposed tool can prove beneficial for the Geely management in making of cars at US country. we are using TREE BASED algorithms(decision tree and random forest).in our geely dataset contains 205 observations and 26 variables.

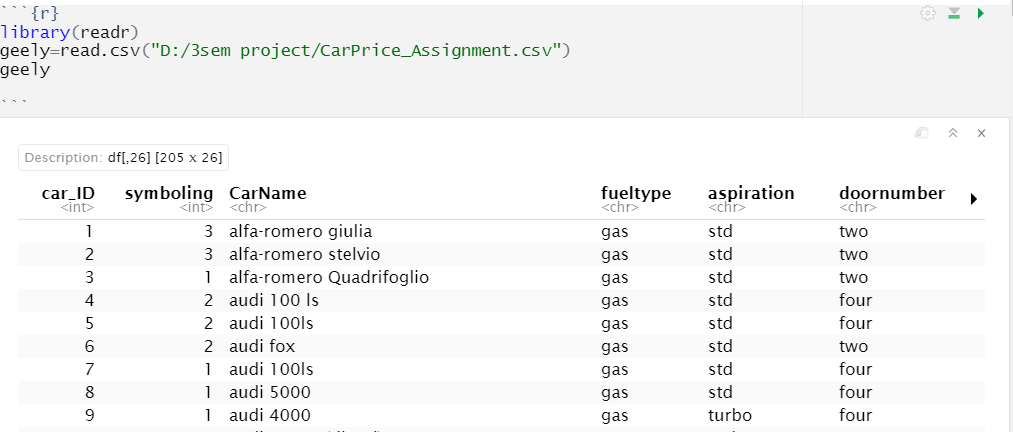
**CHAPTER II**

**GATHERING DATA**

**Load the packages**

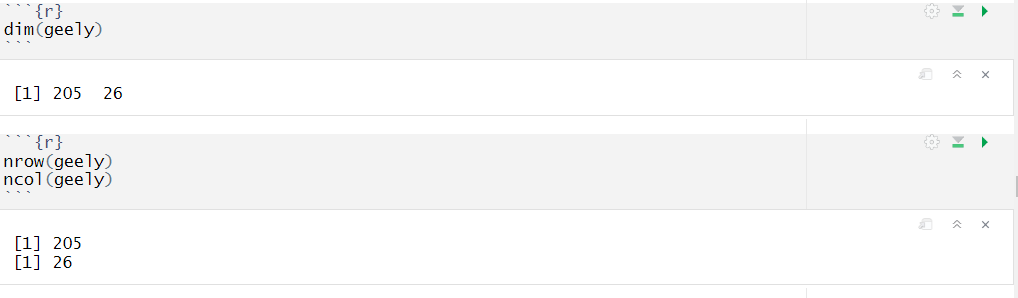


**Load the dataset**

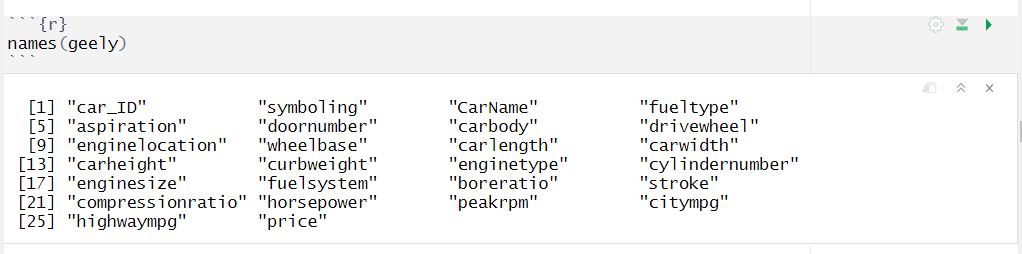


**2.2 .Data description**

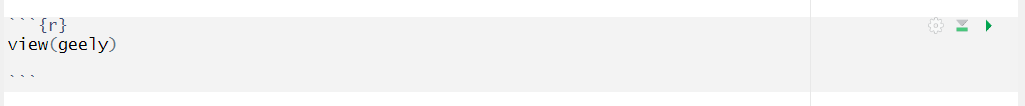
The geely dataset has 205 rows 26 column.

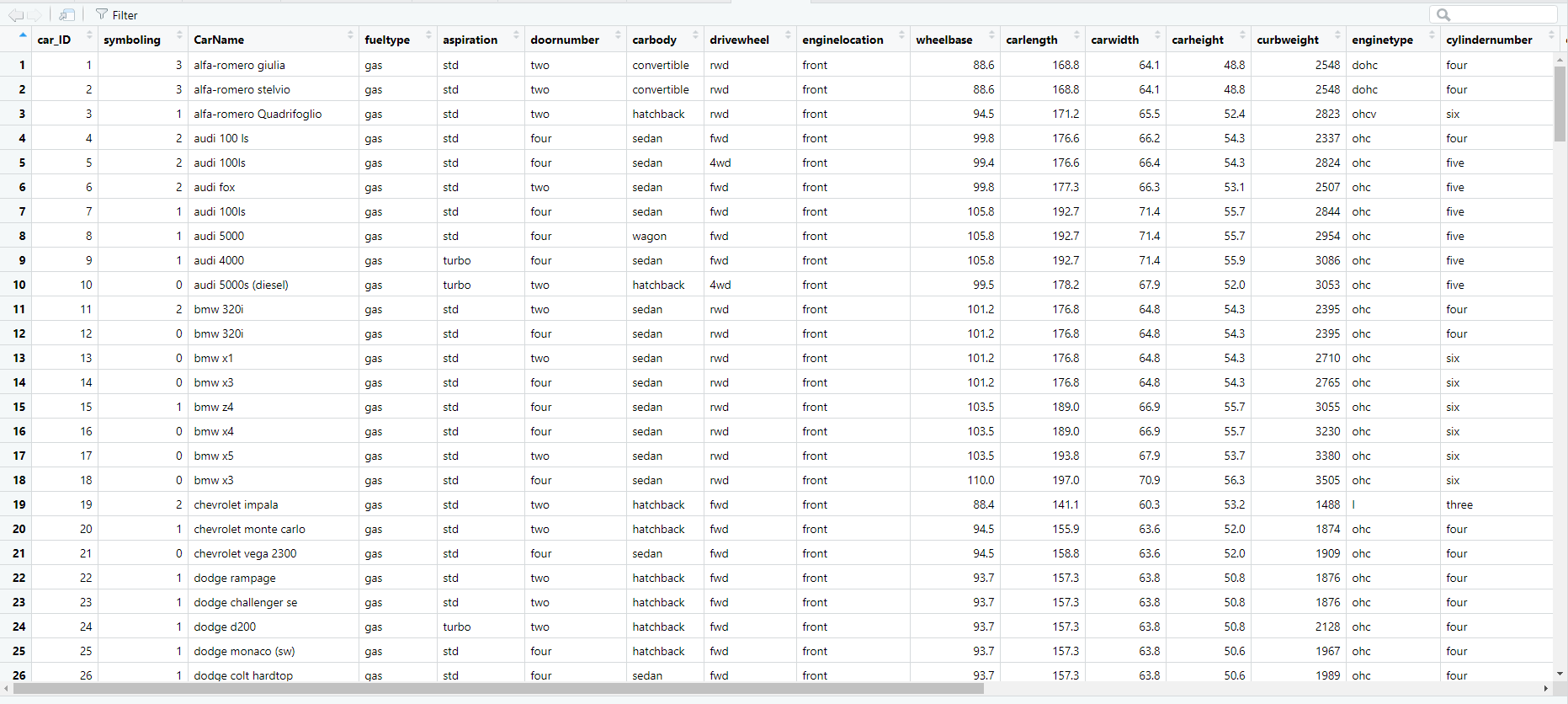


To view the each all variables in this dataset



To view all the contents of a defined object, use the View() function. Behind the scenes, the R calls utils::View() on the input and returns it invisibly. If the input is not a data frame, it is processed using a variant of as.





The data-set contains the following Variables:

**Car\_ID**

Unique id of each observation (Integer)

**Symboling**

Its assigned insurance risk rating, a value of +3 indicates that the auto is risky, -3 that it is probably pretty safe. (Categorical)

**CarCompany**

Name of car company (Categorical)

**Fueltype**

Car fuel type i.e. gas or diesel (Categorical)

**Aspiration**

Aspiration used in a car (Categorical)

**Doornumber**

Number of doors in a car (Categorical)

**Carbody**

Body of car (Categorical)

**Drivewheel**

Type of drive wheel (Categorical)

**Enginelocation**

Location of car engine (Categorical)

**Wheelbase**

Wheelbase of car (Numeric)

**Carlength**

Length of car (Numeric)

**Carwidth**

Width of car (Numeric)

**Carheight**

Height of car (Numeric)

**Carweight**

The weight of a car without occupants or baggage. (Numeric)

**Enginetype**

Type of engine. (Categorical)

**Cylindernumber**

Cylinder placed in the car (Categorical)

**Enginesize**

Size of car (Numeric)

**Fuelsystem**

Fuel system of car (Categorical)

**Boreratio**

Boreratio of car (Numeric)

**Stroke**

Stroke or volume inside the engine (Numeric)

**Compressionratio**

Compression ratio of car (Numeric)

**Horsepower**

Horsepower (Numeric)

**Peakrpm**

Car peak rpm (Numeric)

**Citympg**

Mileage in city (Numeric)

**Highwaympg**

Mileage on highway (Numeric)

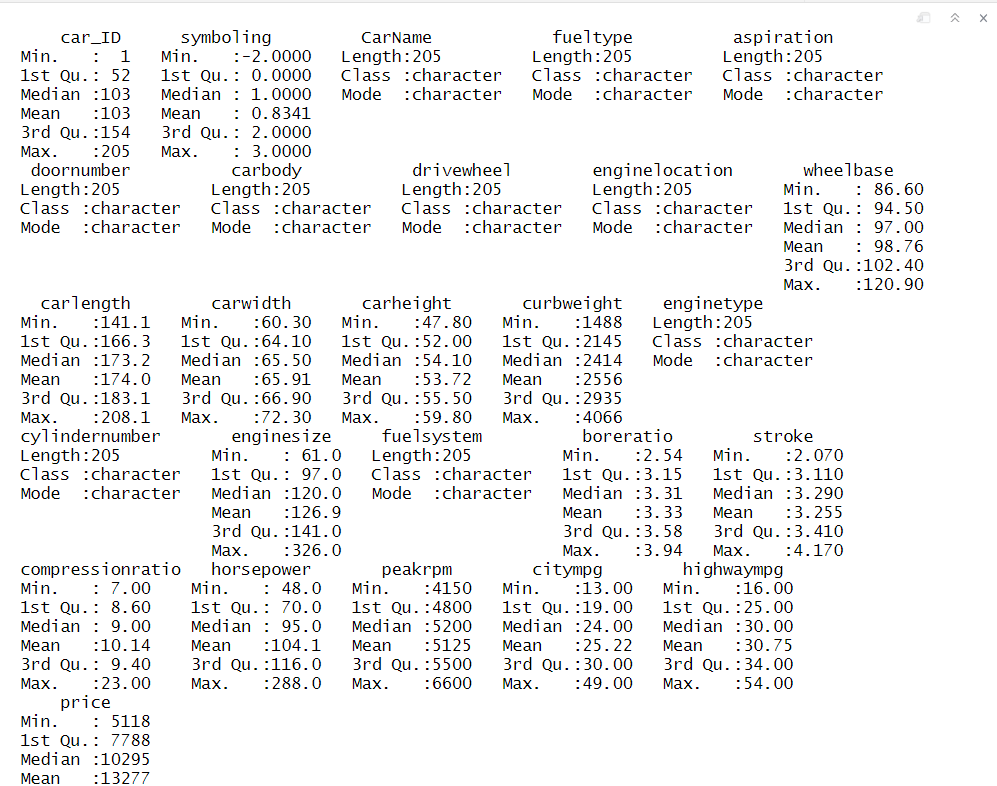
**Price (Dependent variable)**

Price of car (Numeric)

**2.2 Data understanding**

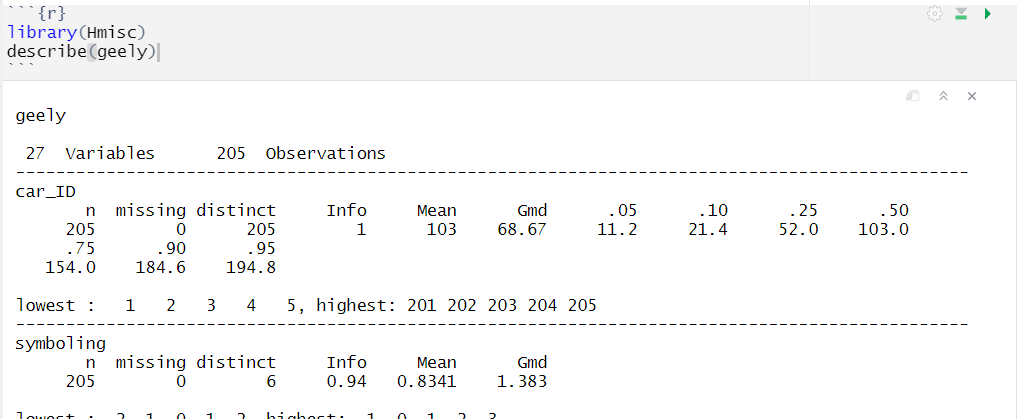
This module explains data understanding. This dataset consist of different columns. Each and every columns we should find the summary () function. This function is used to calculate the average value and determine the maximum, minimum of the column in a data frame.



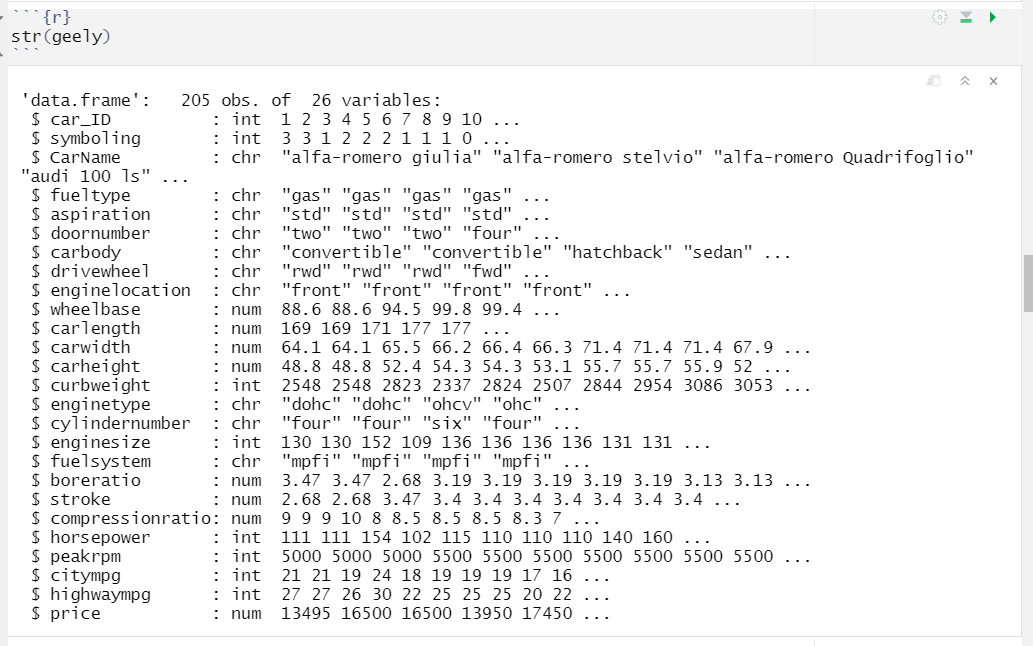


This function determines whether the variable is character, factor, category, binary, discrete numeric, and continuous numeric, and prints a concise statistical summary according to each.

**Describe the data**

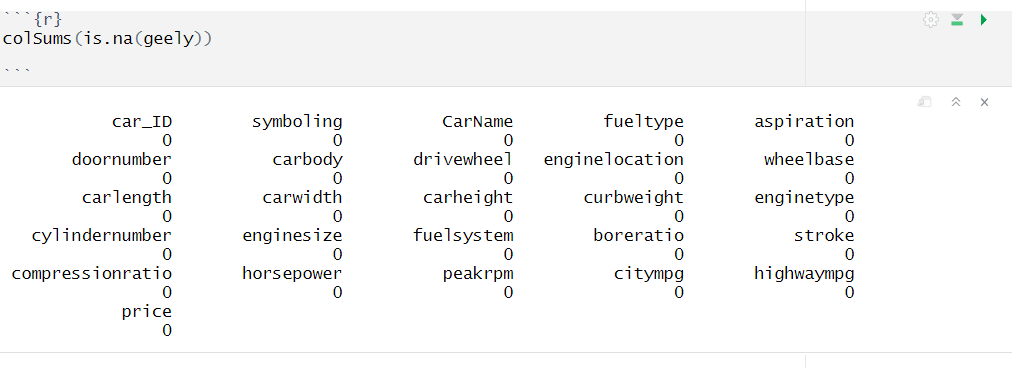


Structure() function is a **simple, yet powerful function that describes a given object with given attributes**.



**To searching the missing values:**

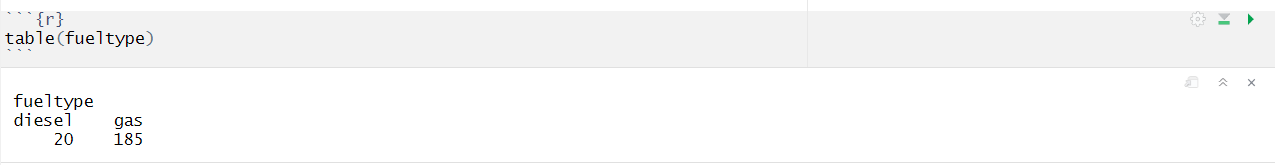
To find missing values you check for NA in R using the is.na() function. This **function returns a value of true and false for each value in a data set**. If the value is NA the is.na() function return the value of true, otherwise, return to a value of false.



No missing values in this datasets.

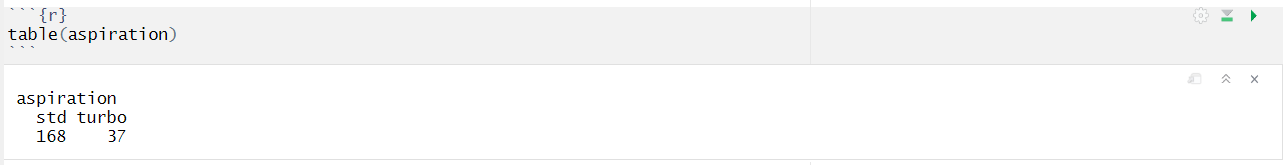
**Fueltype :**

Fueltype is a Categorical variable .There are two fuel type

* + - Diesel
    - Gas

**Aspiration :**

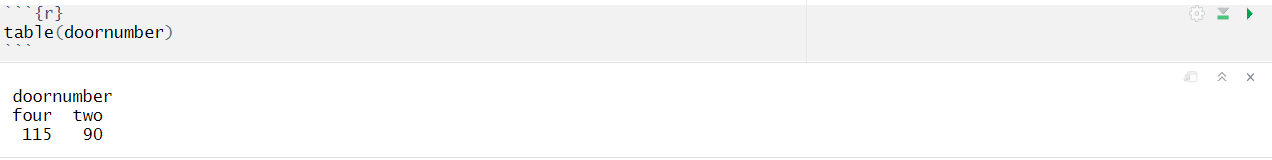
Aspiration is a Categorical variable . It is divided into two

* + - Std
    - Turbo

**Doornumber:**

Doornumberis a Categorical variable . It is divided into two

* + - Four
    - Two



**Carbody :**

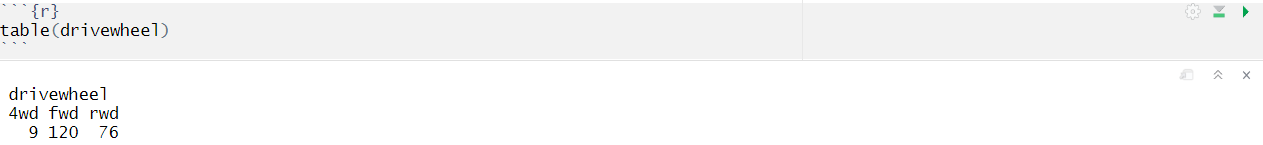
Carbodyis a Categorical variable . It is divided into five

* + - Convertible
    - Hardtop
    - Hatchback
    - Sedan
    - Wagon



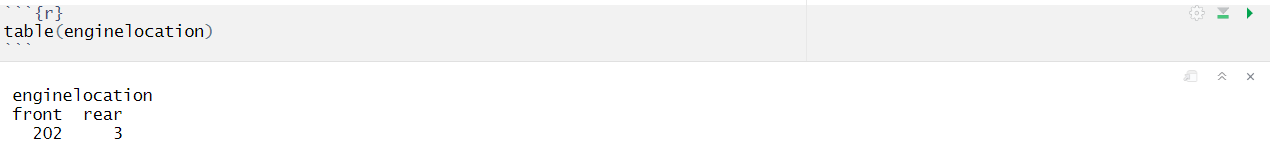
**Drivewheel:**

Drivewheel is a Categorical variable . It is divided into three

* + 4wd
  + Fwd
  + rwd

**Enginelocation :**

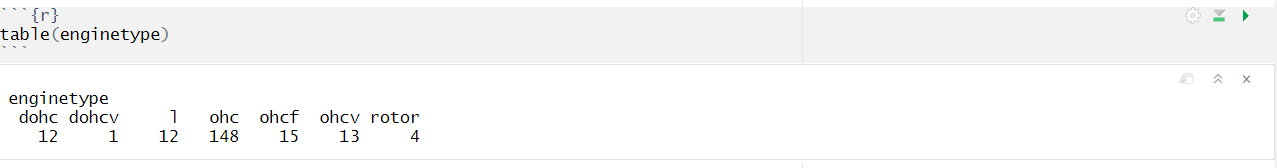
Enginelocation is a Categorical variable . It is divided into two

* + - Front
    - Rear

**Enginetype :**

Enginetype isa Categorical variable . It is divided into seven

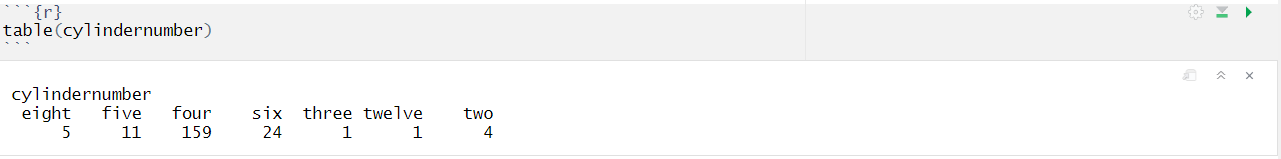
* + - Dohc
    - Dohcv
    - L
    - Ohc
    - Ohcf
    - Ohcv
    - roto



**Cylindernumber:**

Cylinder numberis a Categorical variable . It is divided into seven

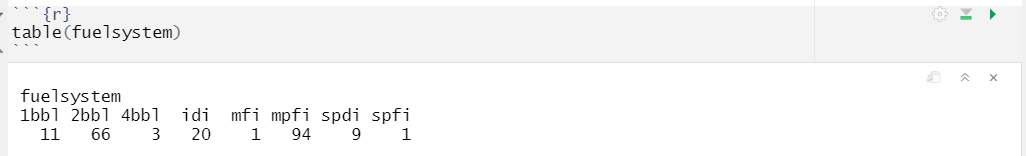
* + - Eight
    - Five
    - Four
    - Six
    - Three
    - Twelve
    - Two



**Fuelsystem**:

Fuel system is a Categorical variable.Its is divided into Eight.

* + - * 1bb1
      * 2bb1
      * 4bb1
      * Idi
      * Mfi
      * Mpfi
      * Spdi
      * spfi



**CHAPTER III**

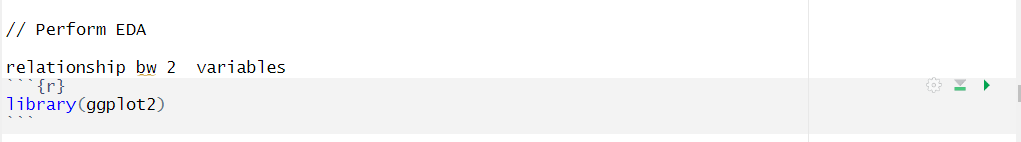
**EXPLORATORY DATA ANALYSIS**

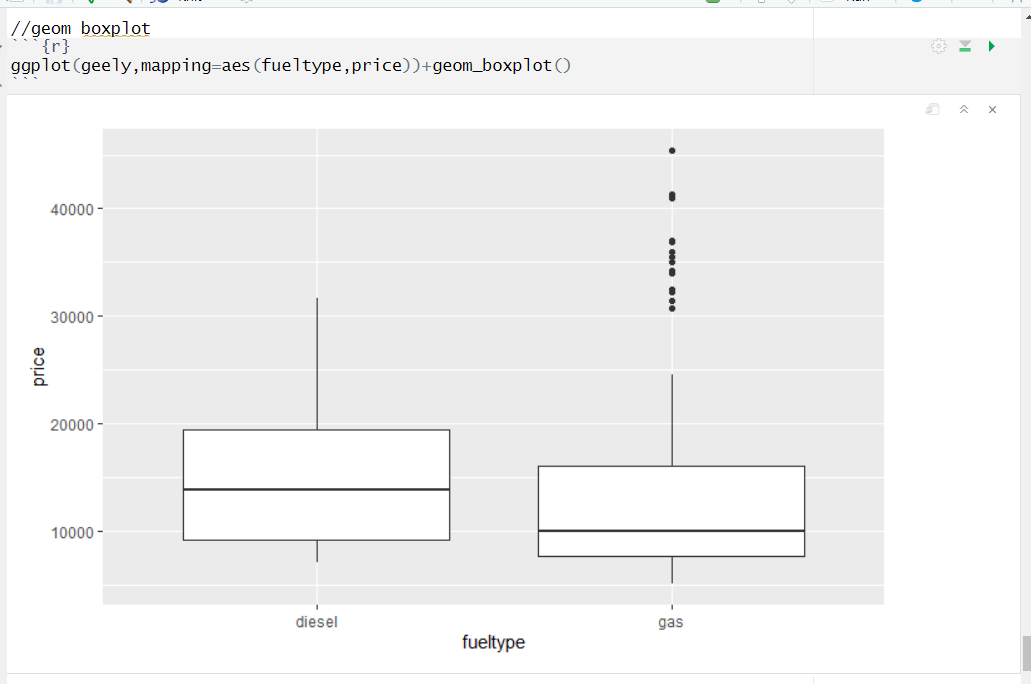
**3.1 Exploratory Data Analysis**

* When we first get our data, it’s very tempting to immediately begin fitting models and assessing how they perform. However, before you begin modeling, it’s absolutely essential to explore the structure of the data and the relationships between the variables in the data set.
* Do a detailed EDA of the geely data set, to learn about the structure of the data and the relationships between the variables in the data set (refer to Data description sheet of geely data). We EDA should involve creating and reviewing many plots/graphs and considering the patterns and relationships you see.

**Plot 1**

Boxplot helps to show the presence of outlier in the particular variable.

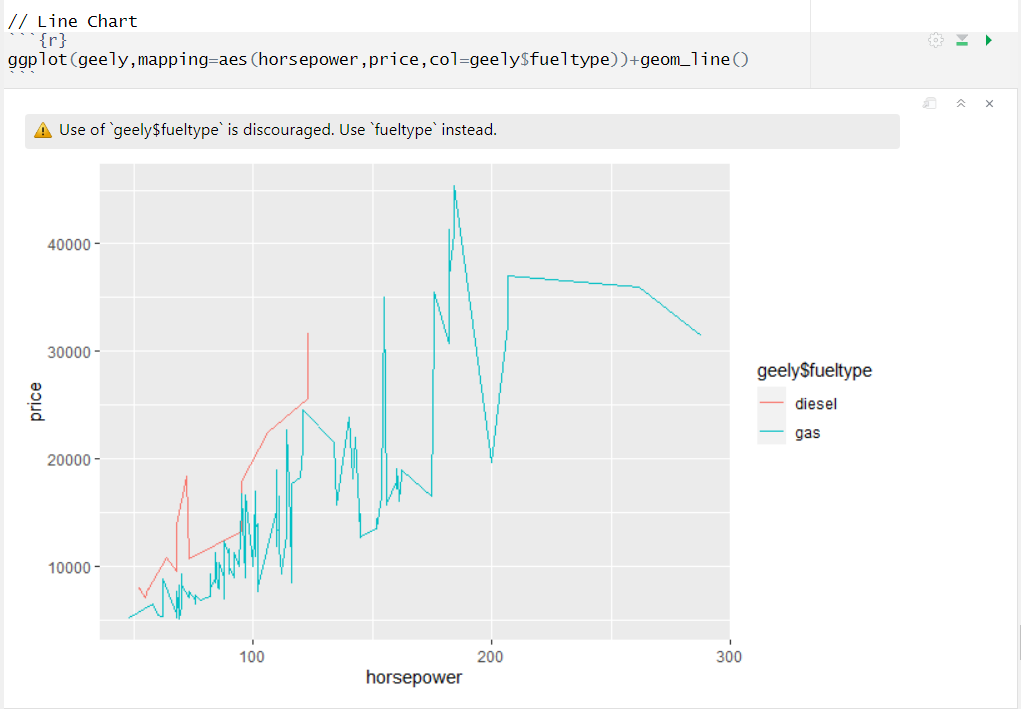




Here, the values above the third quartile, a black line out of the box are considered as Outliers.

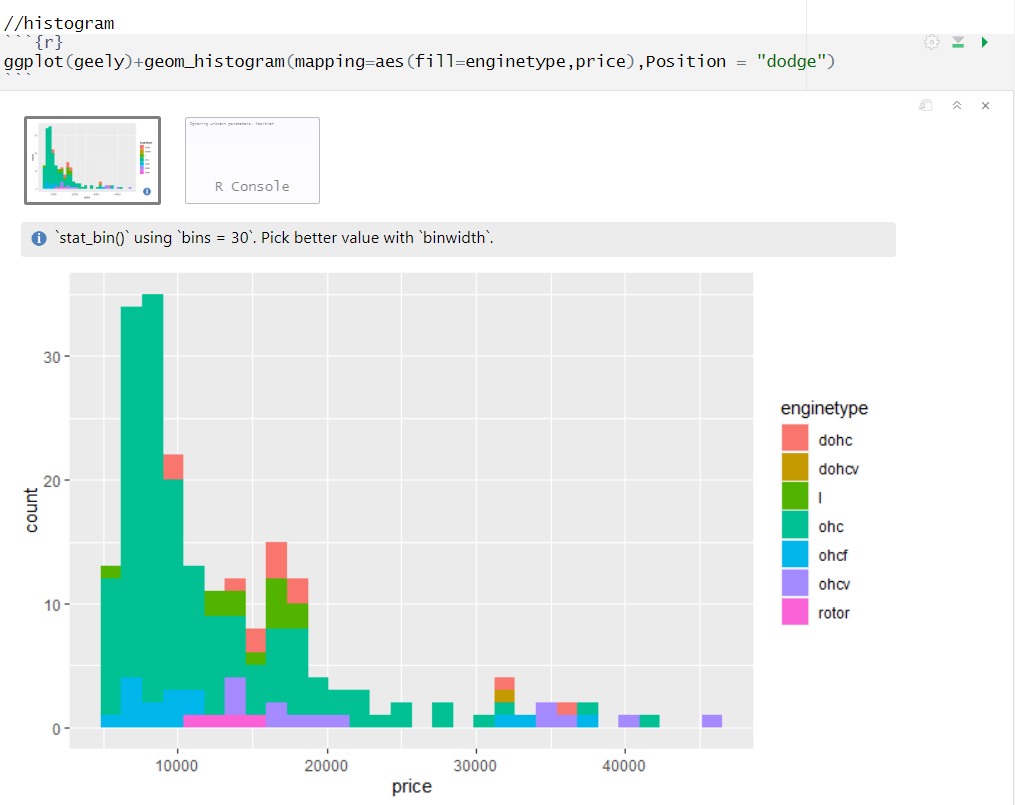
**Plot 2**

Horsepower (vs) price according to fueltype



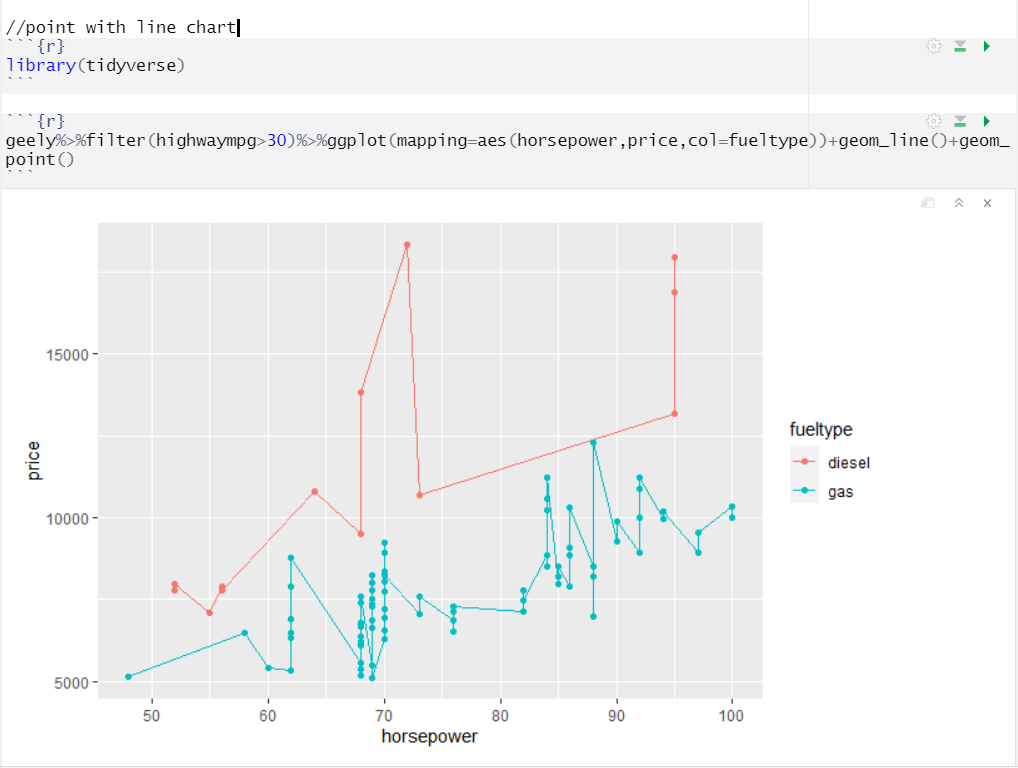
**Plot 3**

Engine type prices



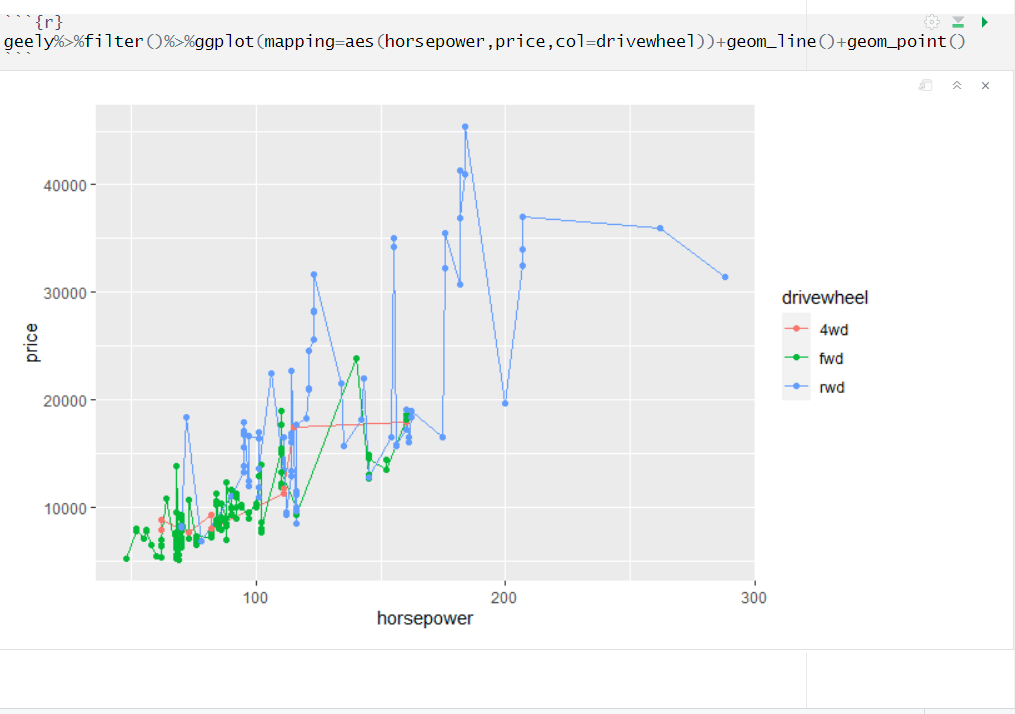
**Plot 4**

Horsepower (vs) price according to that fuel type



**Plot 5**

Horsepower (vs) price according to that wheel type



**CHAPTER IV**

**MODEL BUILDING**

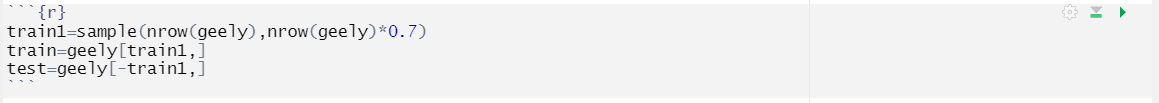
**4.1 Choosing the model**

Choosing a model to use is very essential. You must consider the input and output of your data. For this data:

* The data is labelled, so it’s a supervised learning problem
* The data is looking to predict a number as output, so it’s a regression problem
* So, now we will be looking for Regression model that works on supervised learning problems

**4.2 Training data and Testing data**

* Typically, when we separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing .
* After a model has been processed by using the training set, you test the model by making predictions against the test set.



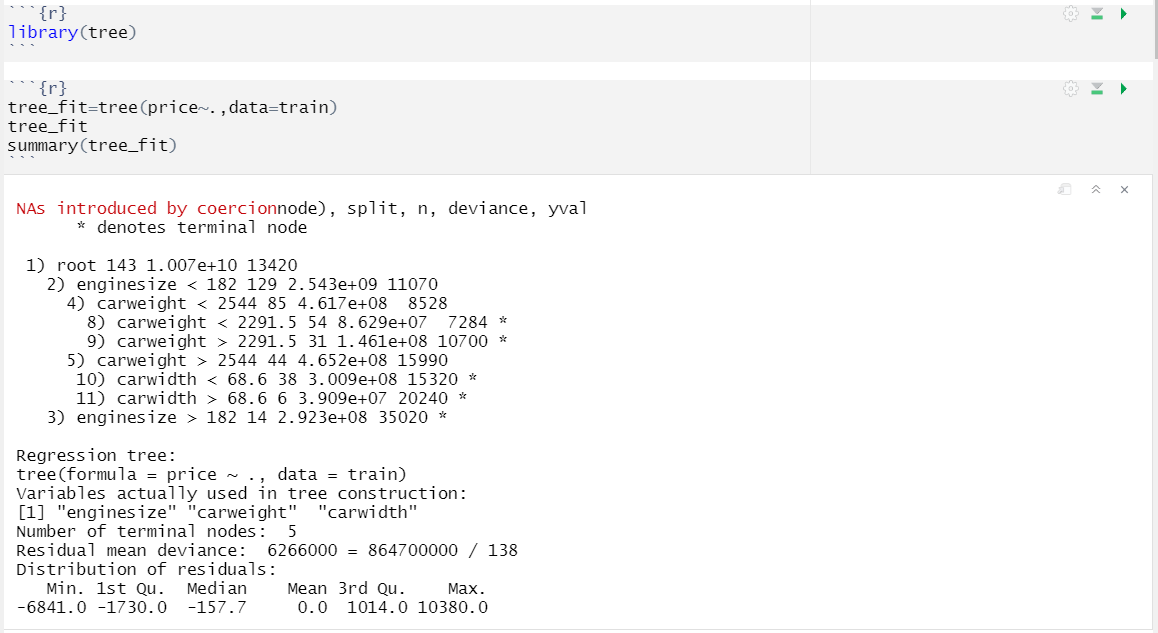
**Decision Tree:**

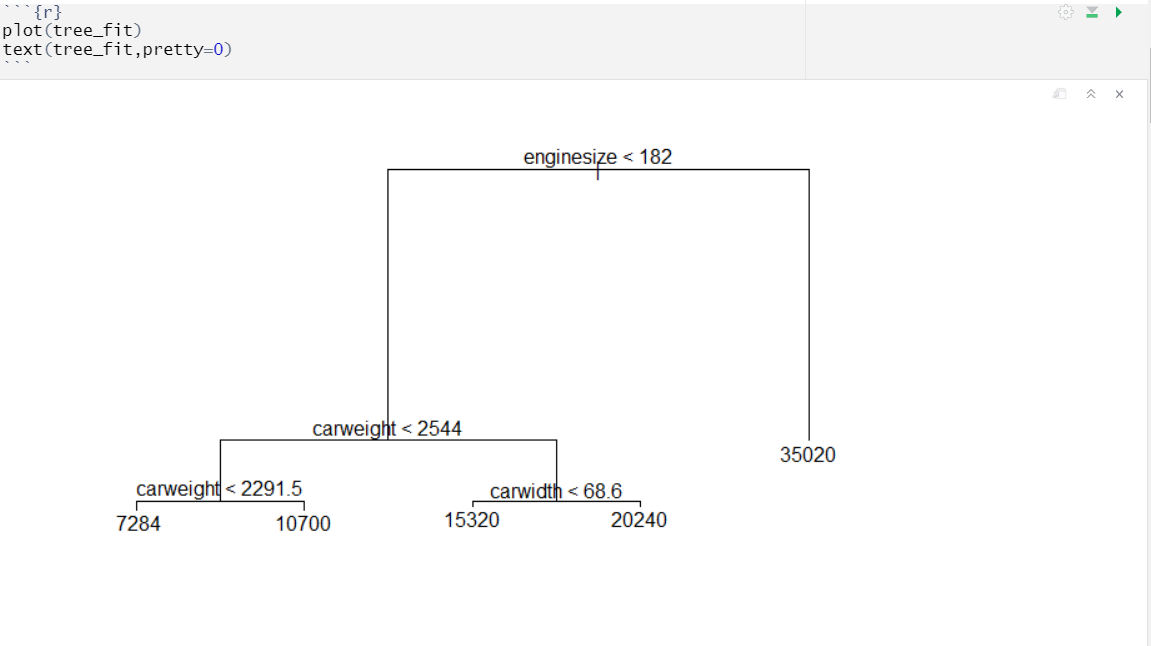
* Decision Tree Algorithm is a supervised Machine Learning Algorithm. It is an approach to predictive analysis that can help you make decisions. Decision tree goes down in a tree-structured format.
* Tree-based methods are simple and useful for interpretation. Decision trees can be applied to both regression and classification problems.
* There are three methods used in decision tree Bagging, random forests, and boosting. These methods grow multiple trees which are then combined to yield a single consensus prediction.

**4.3 Model Building**

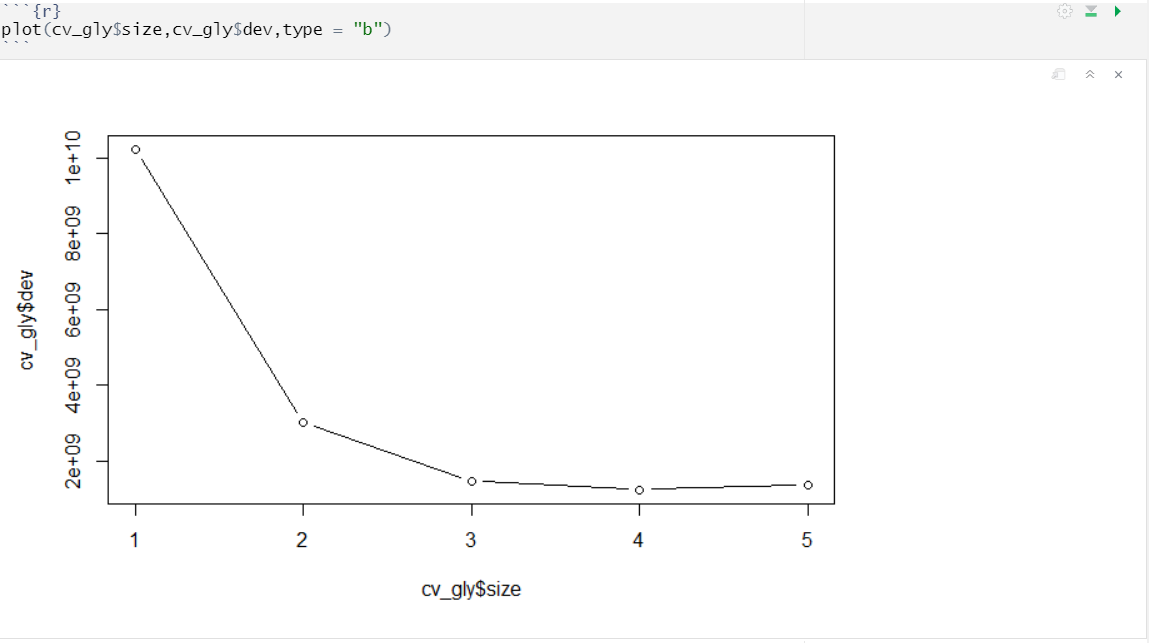
**Tree Building**

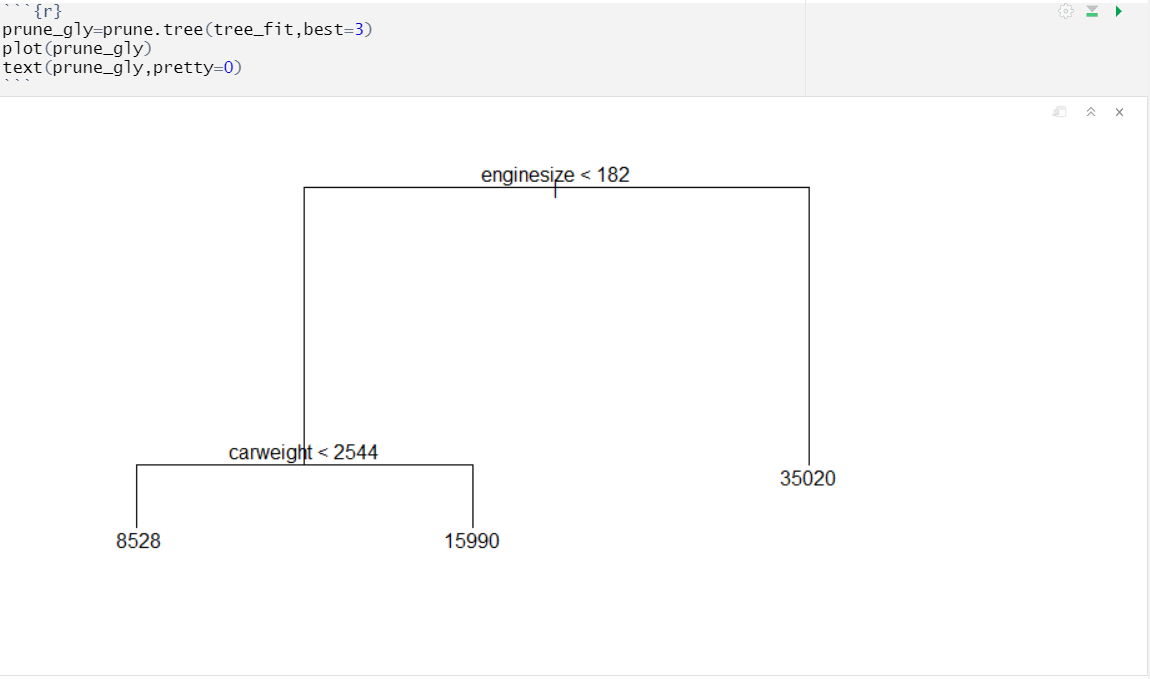
Decision tree is a graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions. It is mostly used in Machine Learning in R

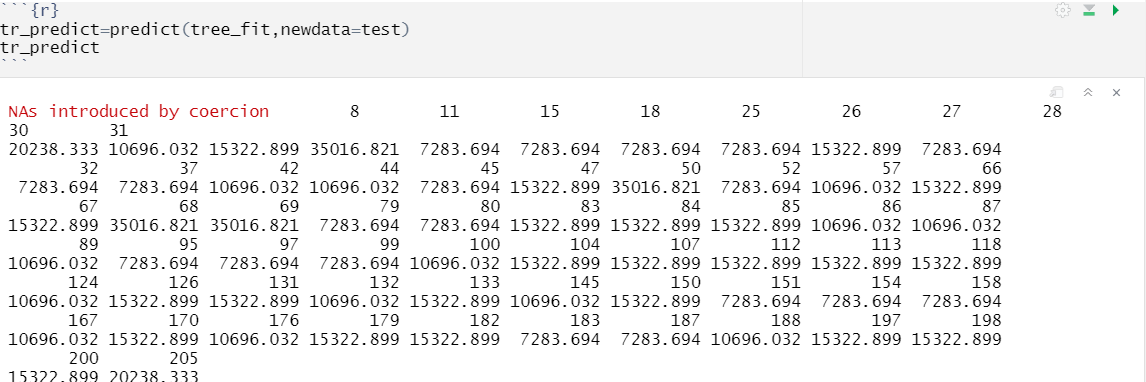






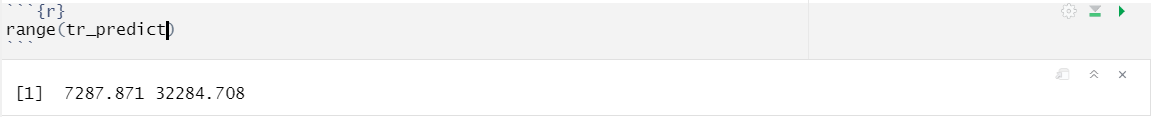




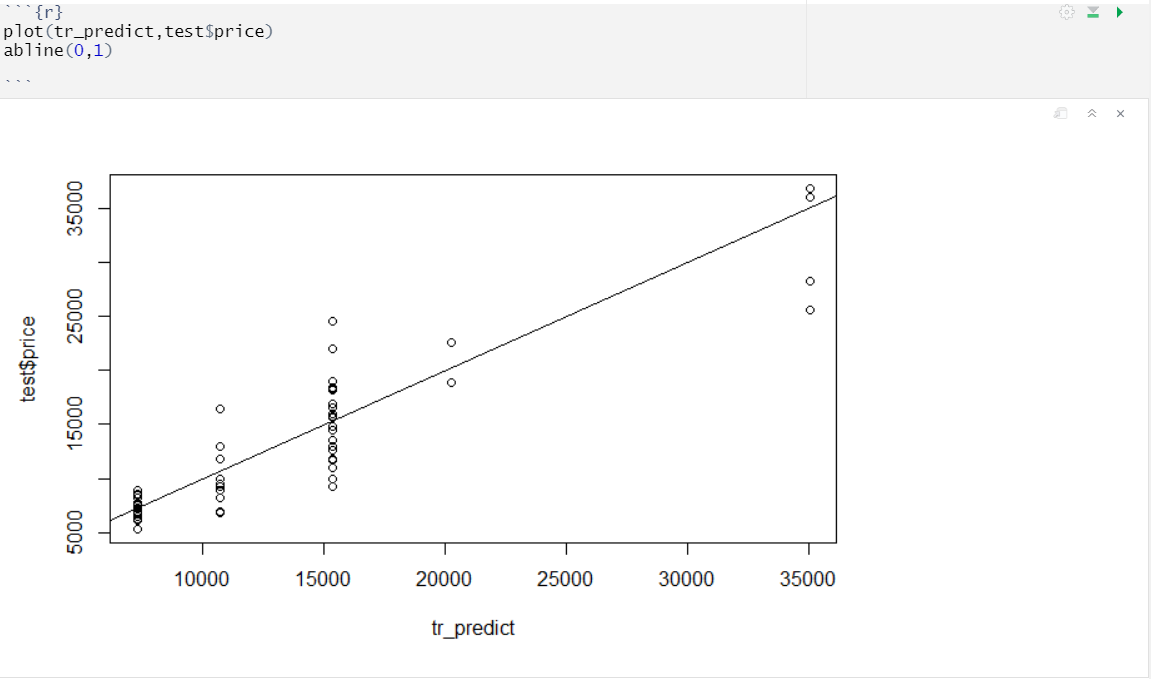


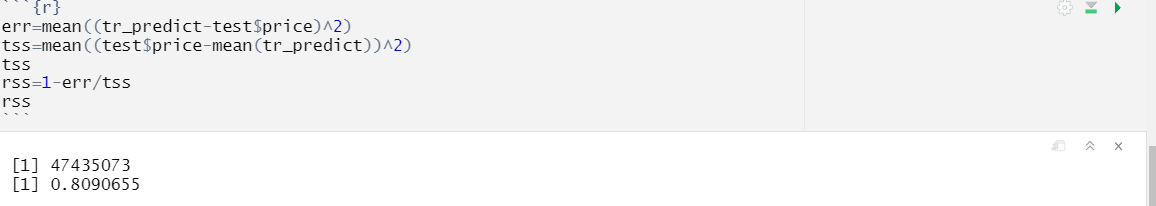
**Range**

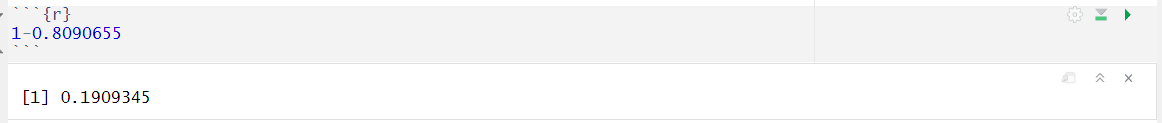
Range function in R returns a vector containing the minimum and maximum of predictions produced for test data.



The model built in the analysis predicts the value of car price varies from 7287.871 to 32284.708.



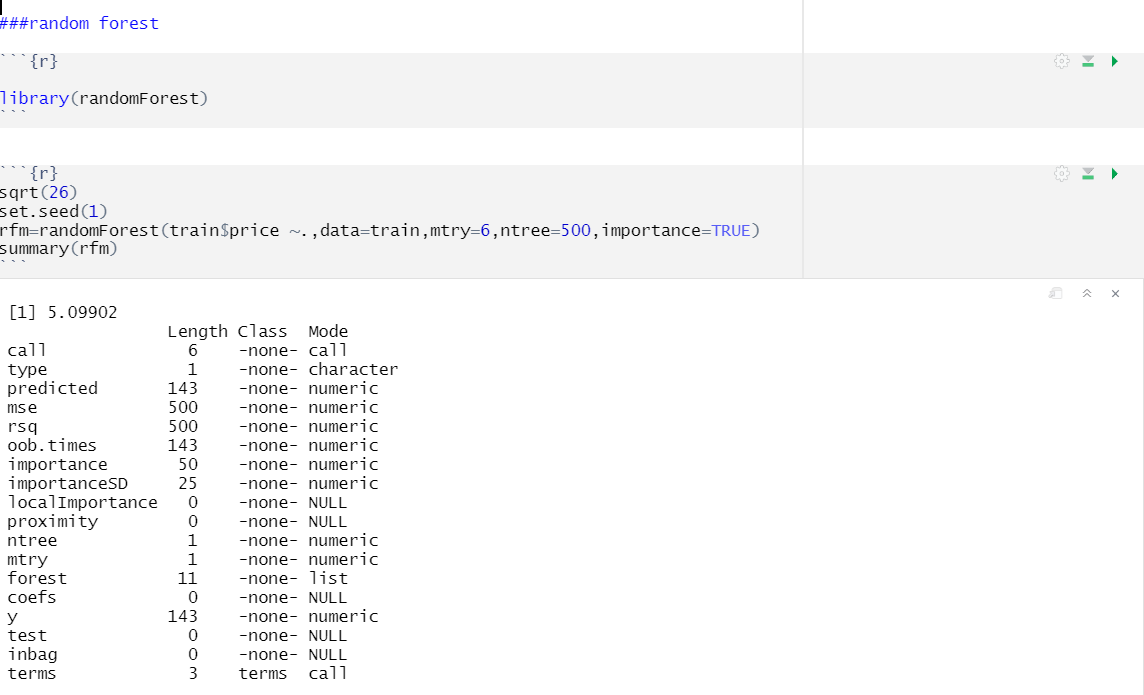


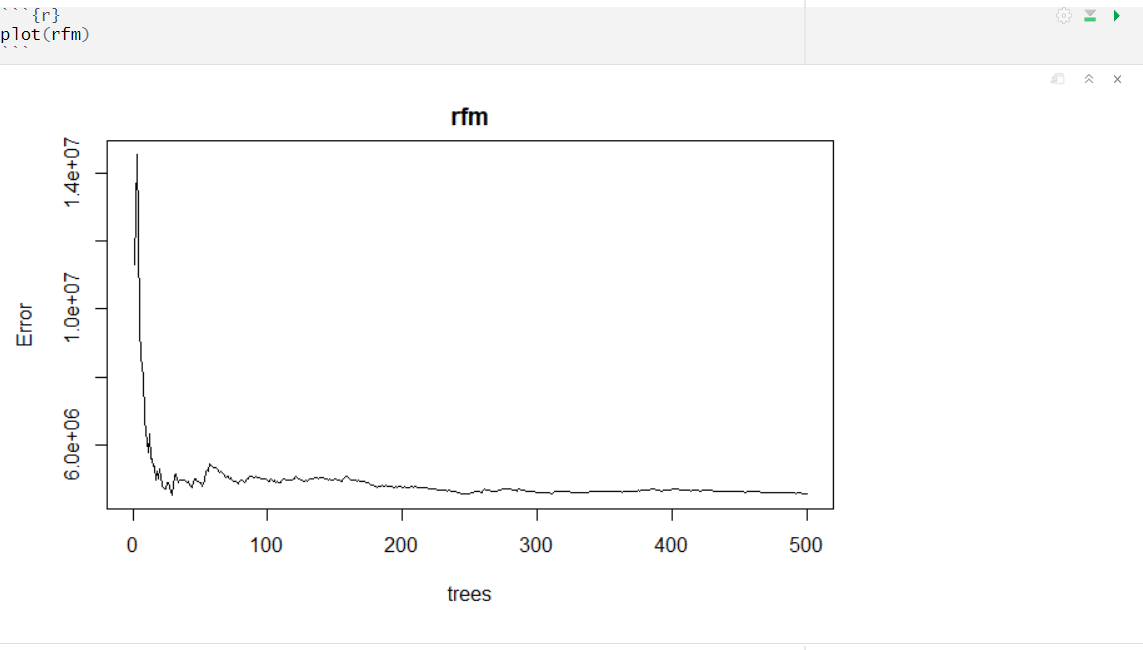


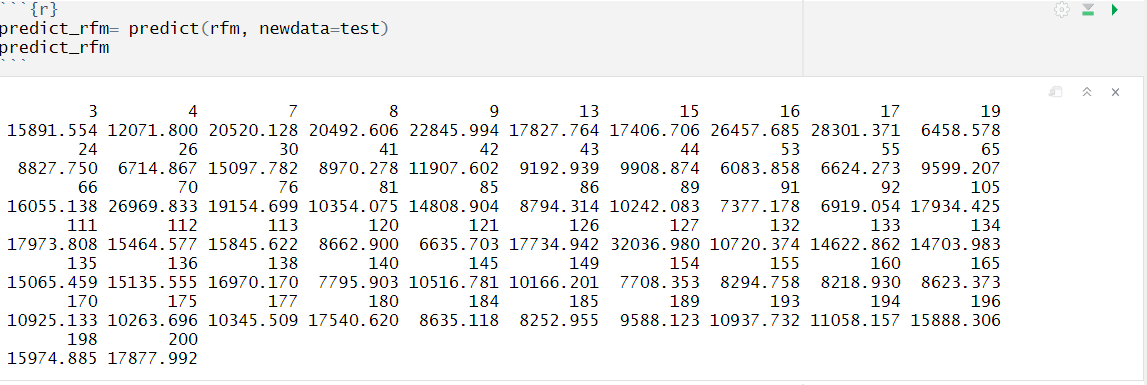
According to decision tree error rate will be 0.1909345

**Random Forest**

* Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.
* Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous.
* Random forests provide an improvement over bagged trees by way of a small tweak that decorrelates the trees. This reduces the variance when we average the trees.
* As in bagging, we build a number of decision trees on bootstrapped training samples.
* But when building these decision trees, each time a split in a tree is considered, a random selection of m predictors is chosen as split candidates from the full set of p predictors.

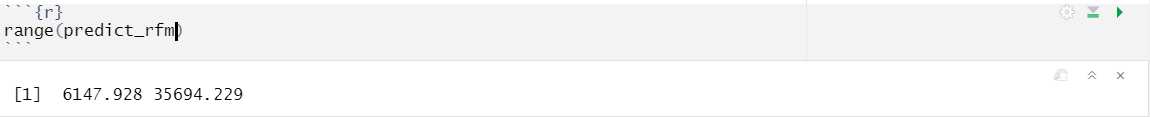




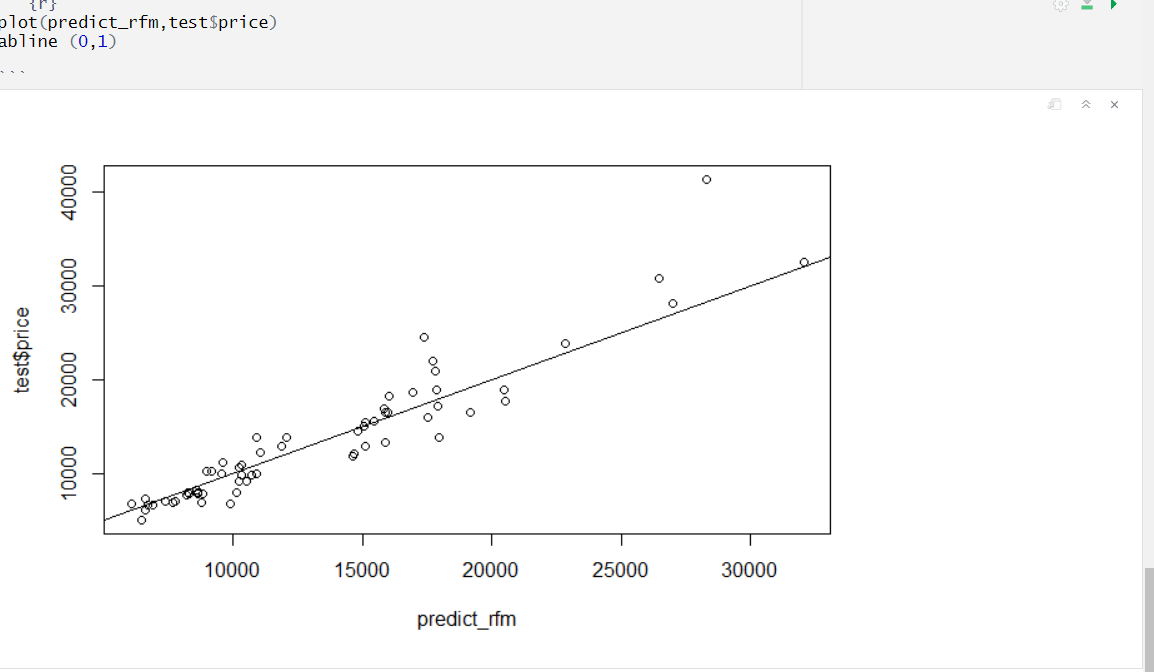


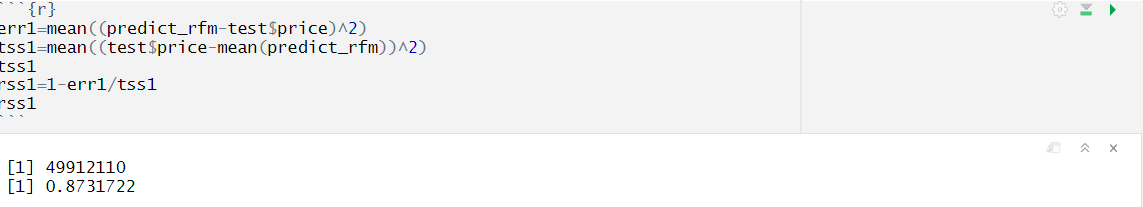
**Range**

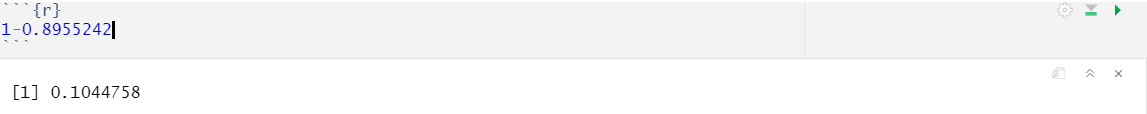
Range function in R returns a vector containing the minimum and maximum of predictions produced for test data.



The model built in the analysis predicts the value of car price varies from 6147.928 to 35694.229.







According to random forest error rate will be 0.104475.

**CHAPTER V**

**PERFORMANCE EVALUATION**

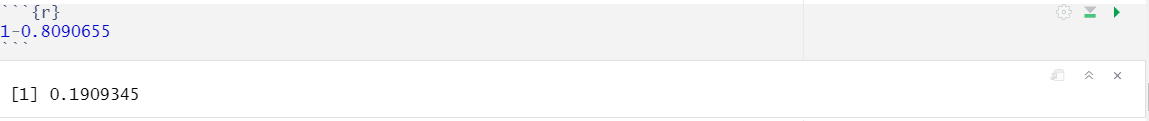
**5.1MODEL EVALUATION**

* Evaluating machine learning algorithm is an essential part of any project. The model may give satisfying results when evaluated using a metric accuracy score but may give poor results when evaluated against other metrics such as logarithmic loss or any other such metric.
* The performance measure is the way to evaluate a solution to the problem. It is the measurement that will make of the predictions made by a trained model on the test dataset. Performance measures are typically specialized to the class of problem that are working with, for example classification, regression, and clustering. Many standard performance measures will give a score that is meaningful to the problem domain.
* Since this project is related to regression model, the commonly used performance measure is mean squared error (MSE). In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

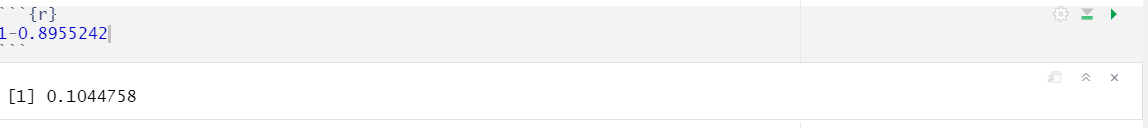
**MSE:**

* Mean squared error is an estimator measures the average of the squares of the errors that is the average squared difference between the estimated value and actual value.

**MSE of Decision Tree**



**MSE of Random Forest**



**CHAPTER VI**

**CONCLUSION**

The built model predicts the car price of the given dataset named geely, with following conclusions.

* The RSS value of decision tree model is 19%
* The RSS value of random forest model is 10%.
* Comparing the two values, RSS value of random forest error rate is small error rate in this dataset.
* As truth, RSS value of lower than this model is best fit of the model.