**CHAPTER I**

**INTRODUCTION**

**1.1 Scope of analysis**

* + - A Chinese automobile company Geely-Auto’s aspires to enter the US market by setting up their managing unit there and producing cars. As a statistical consultant working for an automobile industry, your task is to develop a model to predict the selling price of a car in US Market. Geely Company hopes to use this information to help assess the factors affecting the pricing of cars in the American marketing.
    - Many factors, in addition, to 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**.**

**1.2 Approach of Analysis**

With 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.

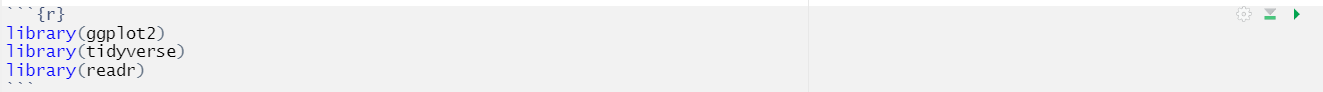
Additionally, the data is used to predict the selling price of the cars through various machine learning approaches. The proposed tool can prove beneficial for the Geely management in making of cars at US country.

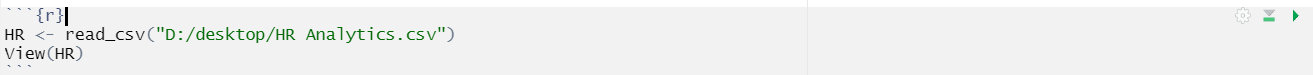
**CHAPTER II**

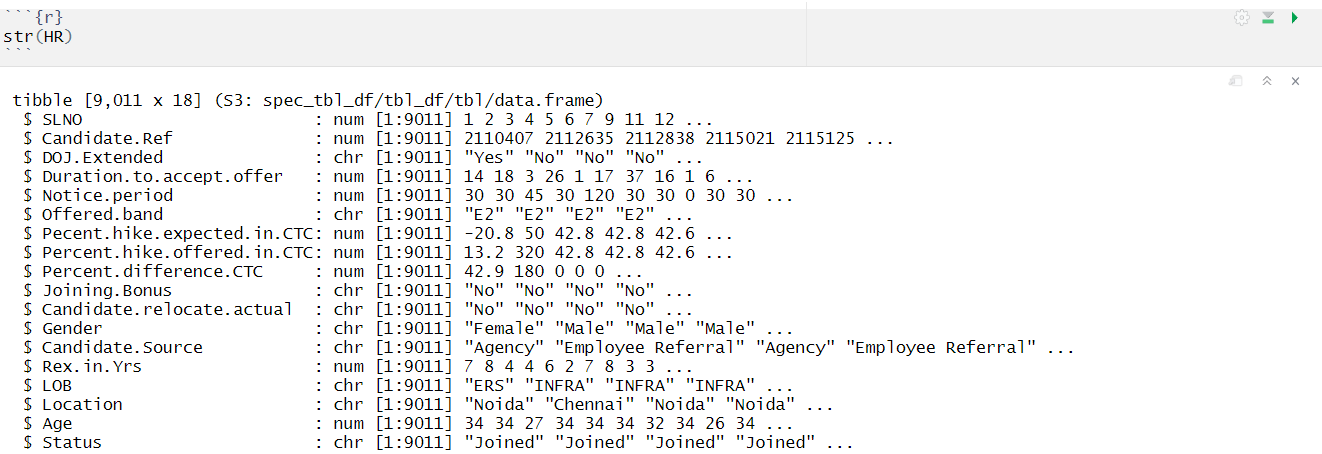
**DATA UNDERSTANDING**

**HR ANALYTICS**

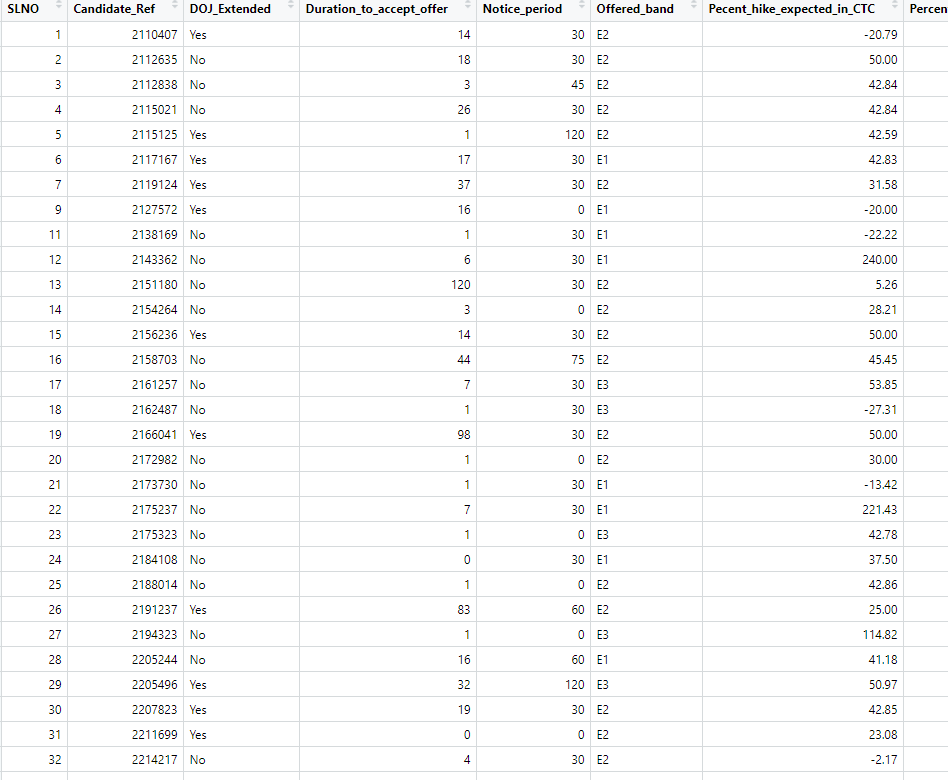
**2.1 Gathering Data**

**Load the relevant Packages**

**Load the dataset**

**Structure of the data**

**2.2 Data description**

 The HR dataset has 9011 rows and 18 columns

The data-set contains the following Variables:

**Candidate reference number**

Unique number to identify the candidate

**DOJ extended**

Binary variable identifying whether candidate asked for date of joining extension (Yes/No)

**Duration to accept the offer**

Number of days taken by the candidate to accept the offer (continuous variable)

**Notice period**

Notice period to be served in the parting company before candidate can join this company (continuous variable)

**Offered band**

Band offered to the candidate based on experience and performance in interview rounds (categorical variable labelled C0/C1/C2/C3/C4/C5/C6)

**Percentage hike (CTC) expected**

Percentage hike expected by the candidate (continuous variable)

**Percentage hike offered (CTC)**

Percentage hike offered by the company (continuous variable)

**Joining bonus**

Binary variable indicating if joining bonus was given or not (Yes/No)

**Gender**

Gender of the candidate (Male/Female)

**Candidate source**

Source from which resume of the candidate was obtained (categorical variables with categories: Employee referral/Agency/Direct)

**REX (in years)**

Relevant years of experience of the candidate for the position offered (continuous variable)

**LOB**

Line of business for which offer was rolled out (categorical variable)

**DOB**

Date of birth of the candidate

**Joining location**

Company location for which offer was rolled out for candidate to join (categorical variable)

**Candidate relocation status**

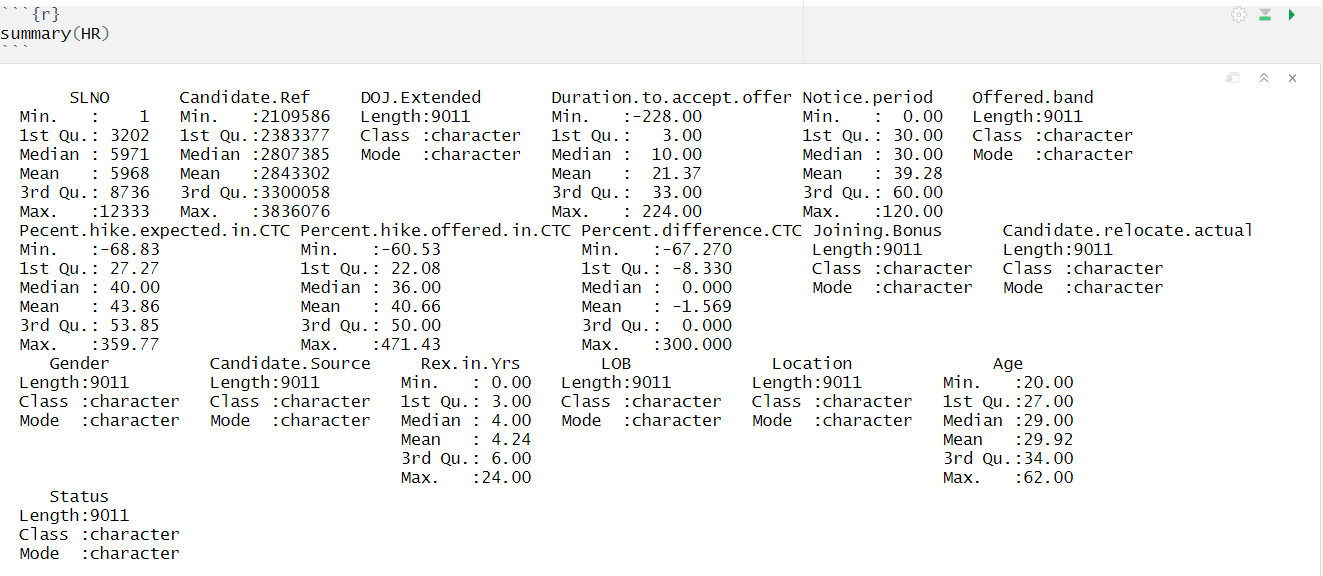
Binary variable indicating whether candidate has to relocate from one city to another city for joining (Yes/No)

**HR status**

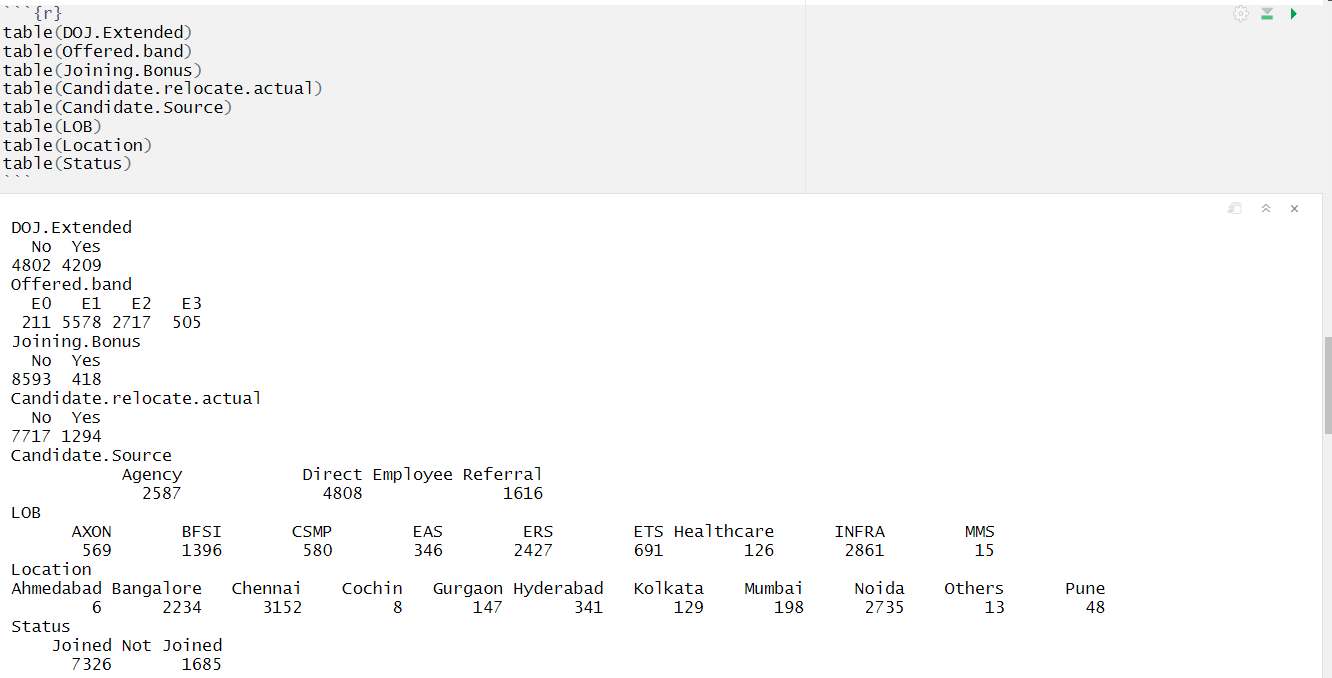
Final joining status of candidate (Joined/Not-Joined)

**2.3 Data Understanding**

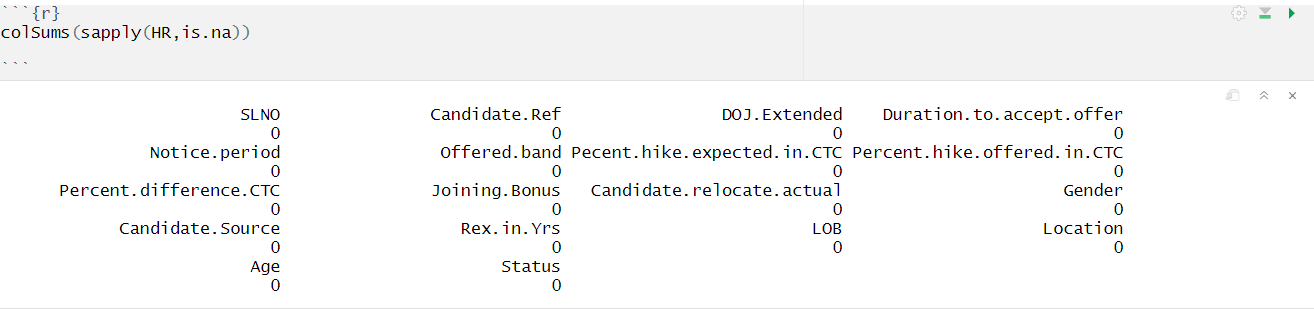
This module explains data understanding. This dataset consist of different columns. Each and every columns we should find the summary () function. This function is used to calculate the average value and determine the maximum, minimum of the column in a data frame.



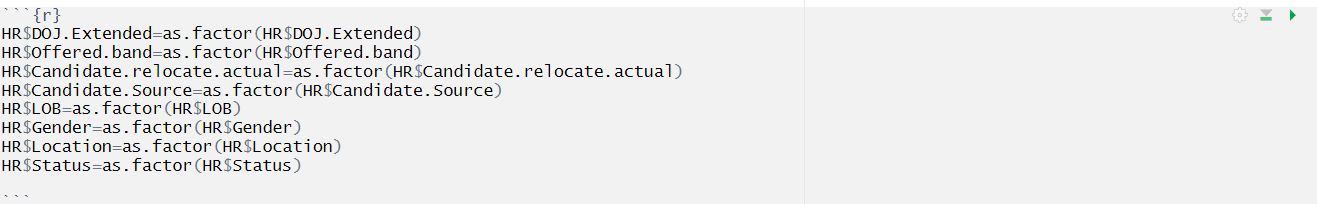
**To show the categories of Categorical variable :**



**NA’s in the dataset**



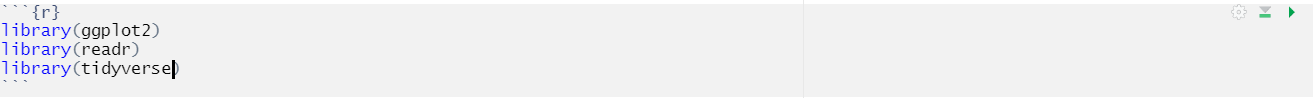
**Converting to factor**



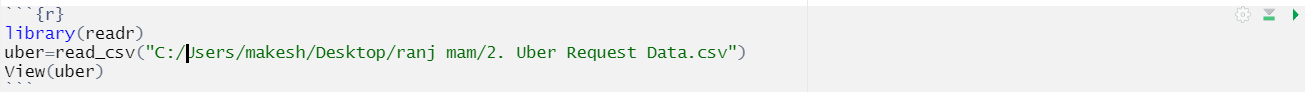
**UBER**

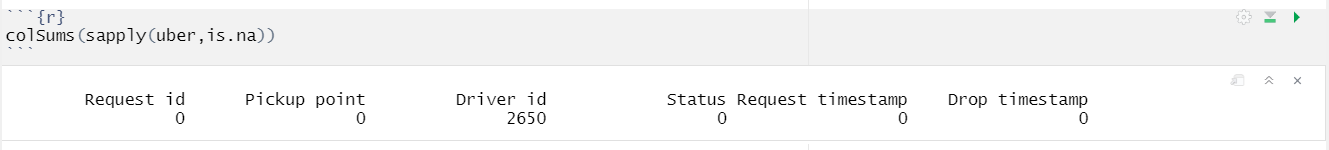
**2.1 Gathering Data**

**Load the relevant Packages**

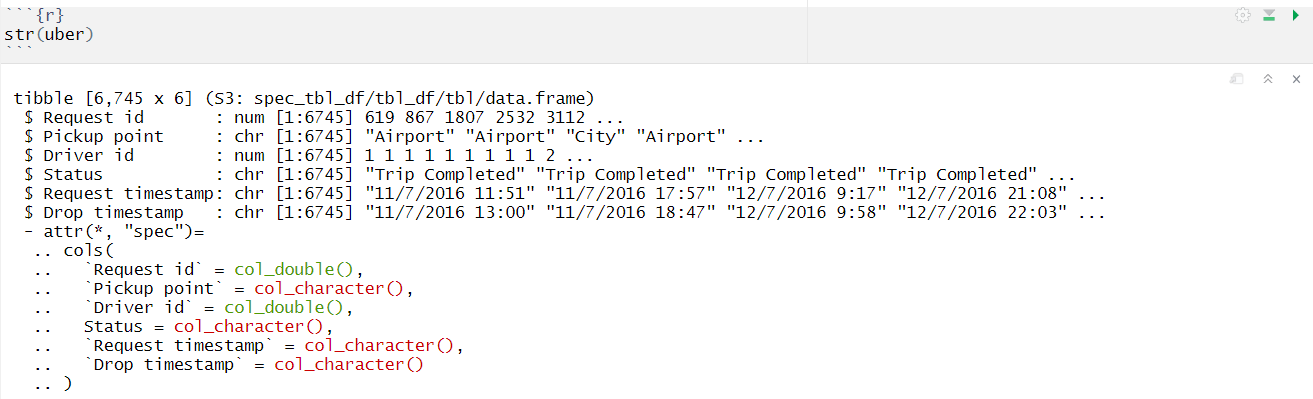


**Load the dataset**



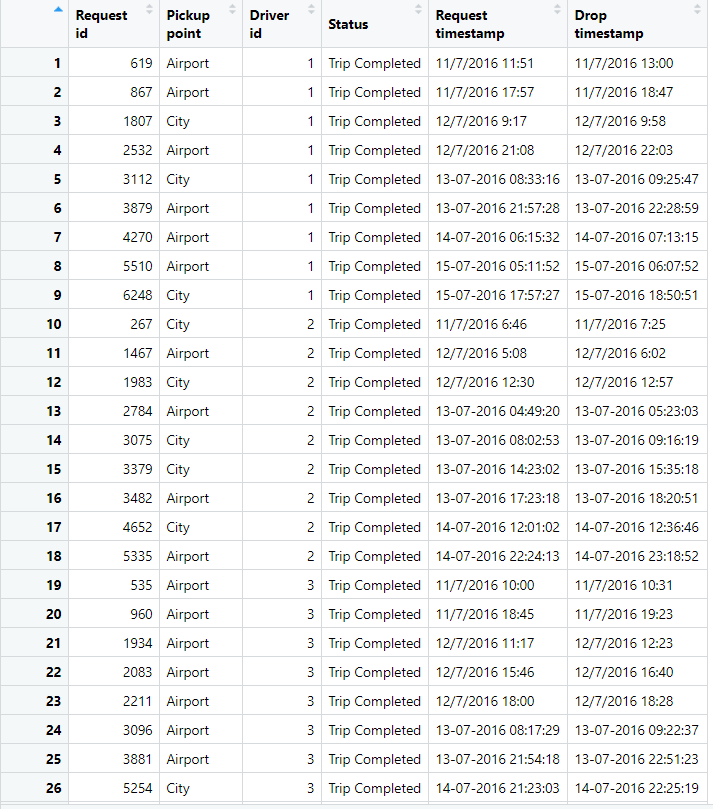
**NA’s in the dataset**

**Structure of the data**



**2.2 Data description**

Uber dataset has 6745 rows and 8columns



The data-set contains the following Variables:

**Request id:**

A unique identifier of the request

**Time of request:**

The date and time at which the customer made the trip request

**Drop-off time:**

The drop-off date and time, in case the trip was completed

**Pick-up point**:

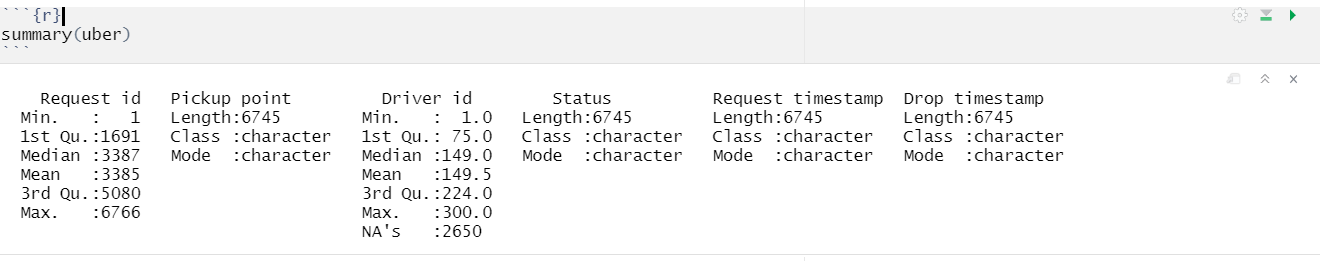
The point from which the request was made

**Driver id:**

The unique identification number of the driver 6. Status of the request: The final status of the trip that can be either completed, cancelled by the driver or no cars available

**2.3 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.

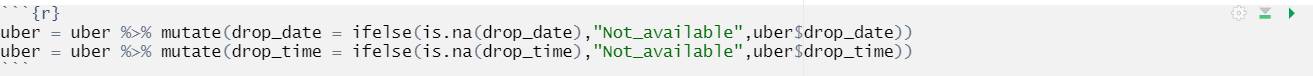


**Data Cleaning**

**Separate the Date and time**



**Change the drive id value to Not\_available**

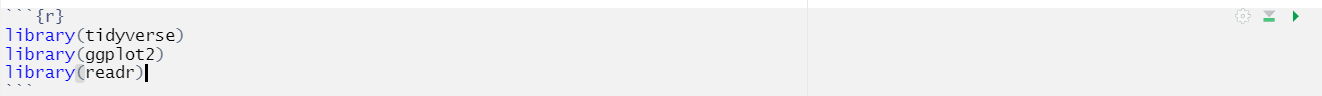




**GRAMENER**

**2.1 Gathering Data**

**Load the relevant Packages**

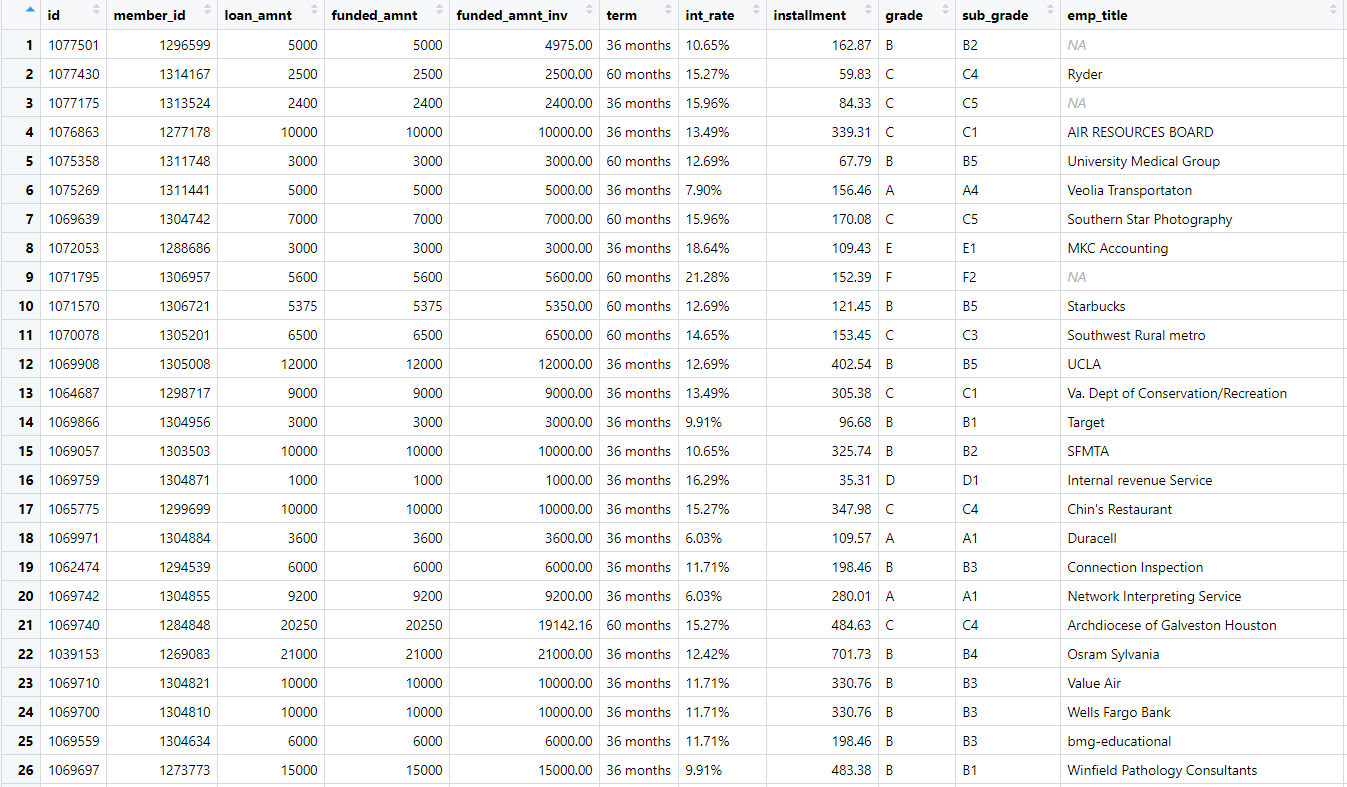


**Load the dataset**



**2.2 Data description**

Gramener dataset has 39717 rows and 111 columns



**Customer’s Demographic Information**

* + - * + Emp\_title
        + Emp\_length
        + Home\_ownership
        + Annual\_inc
        + Verification\_status
        + Addr\_state
        + Zip\_code
        + Title
        + Purpose
        + Desc
        + url

**Loan Characteristics Information**

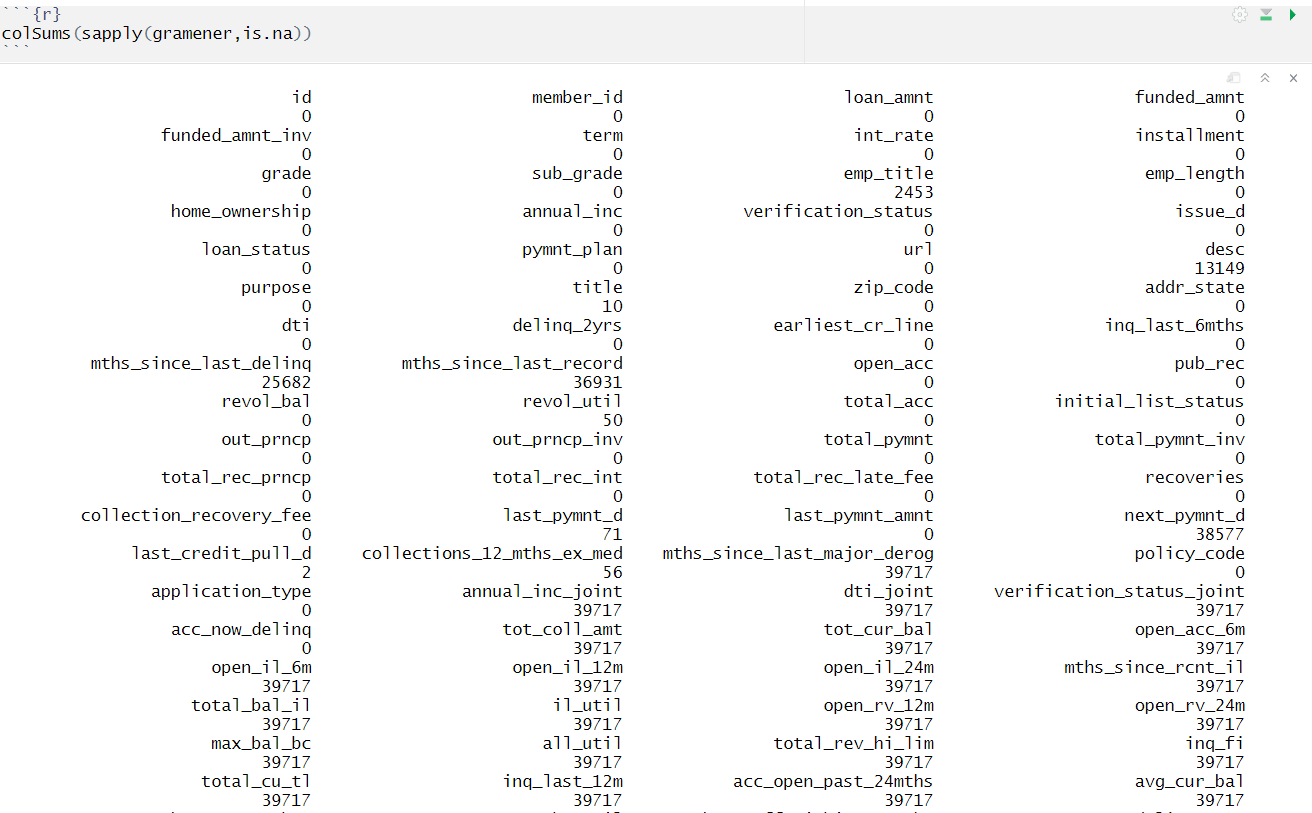
* + - * + Loan amount
        + Funded amount
        + Funded amount invested
        + Interest rate
        + Loan status
        + Loan grade
        + Loan sub\_grade
        + Dti
        + Loan issue date
        + Loan term
        + Installment

**Credit information: Customer Behavior Variable**

* + - * + Delinq\_2yrs
        + Earliest\_cr\_line
        + Inq\_last\_6mths
        + Open\_acc
        + Pub\_rec
        + Revol\_bal
        + Revol\_util
        + Total\_acc
        + Out\_prncp
        + Out\_prncp\_inv
        + Total\_pymnt
        + Total\_pymnt\_inv
        + Total\_rec\_prncp
        + Total\_rec\_int
        + Total\_rec\_late\_fee
        + Recoveries
        + Collection\_recovery\_fee

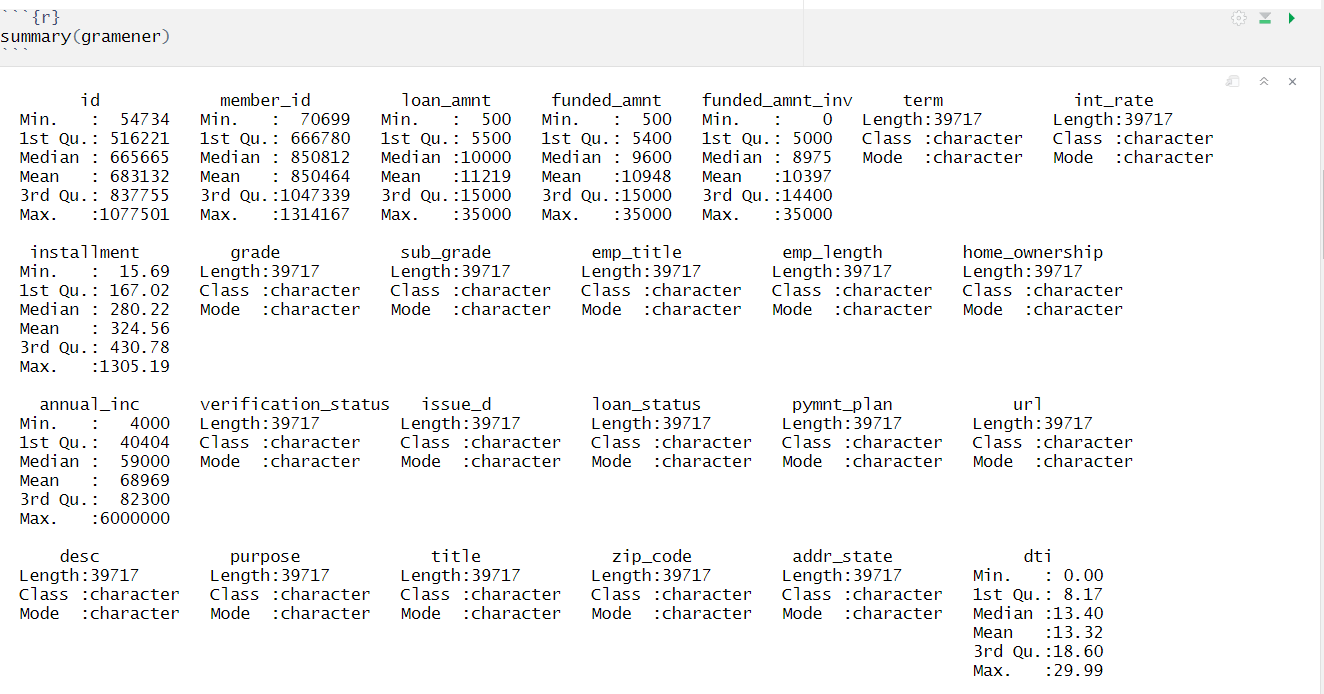
* + - * + Last\_pymnt\_d
        + Last\_pymnt\_amnt
        + Next\_pymnt\_d
        + Last\_credit\_pull\_d
        + Application\_type

**NA’s in the dataset**



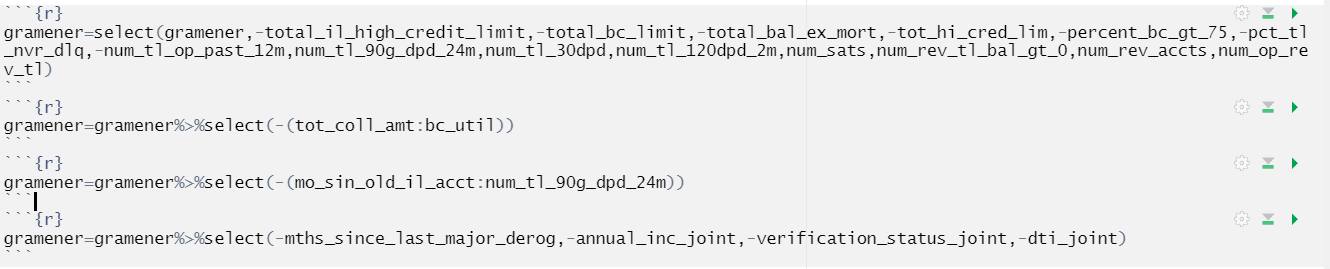
**2.3 Data Understanding**

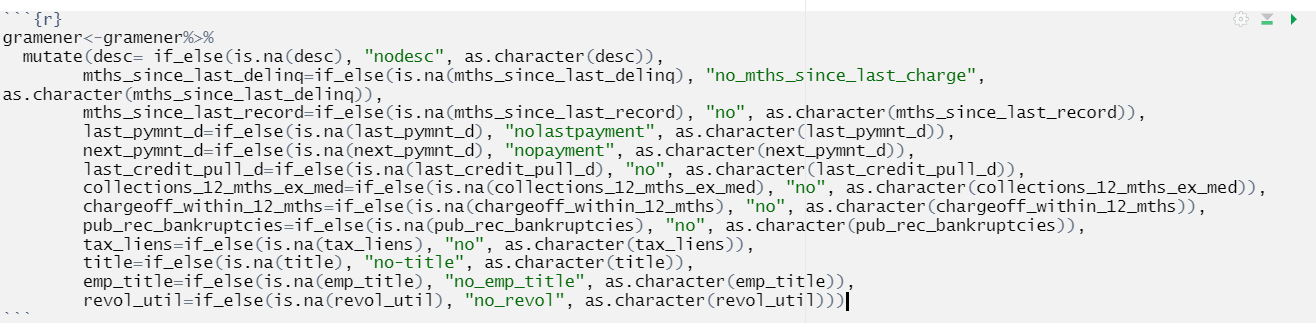
This module explains data understanding. This dataset consist of different columns. Each and every columns we should find the summary () function. This function is used to calculate the average value and determine the maximum, minimum of the column in a data frame.



**Variable names:**



**unselect the unwanted columns**



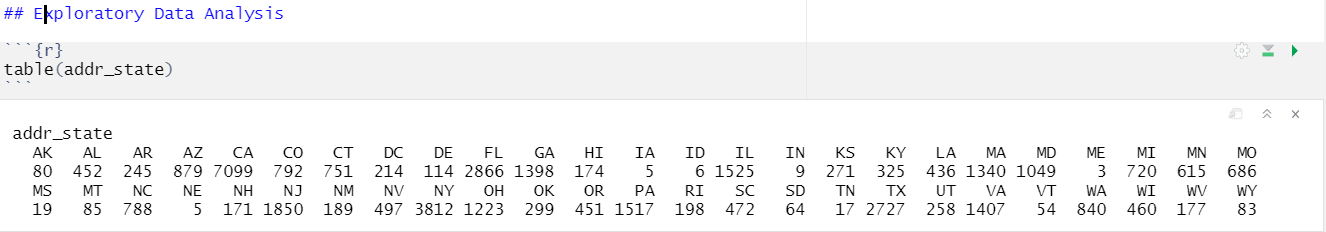
**Check the NA’S**

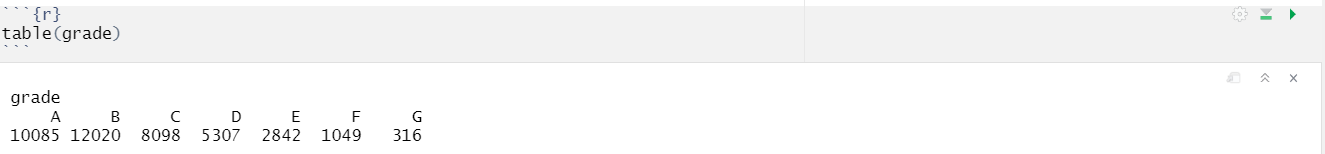


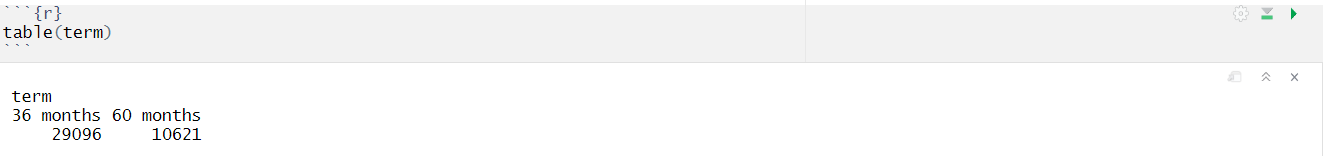
**Structure of the dataset**

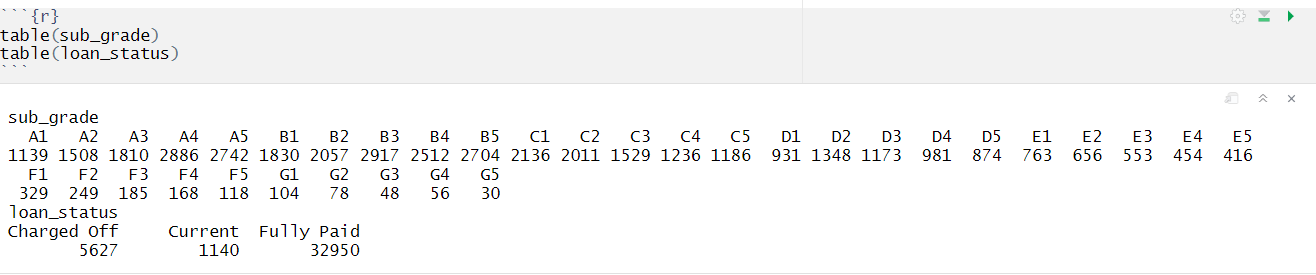


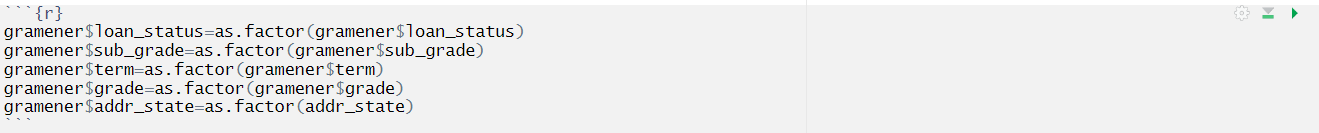
**To check the categories and count of the categories of the variable**







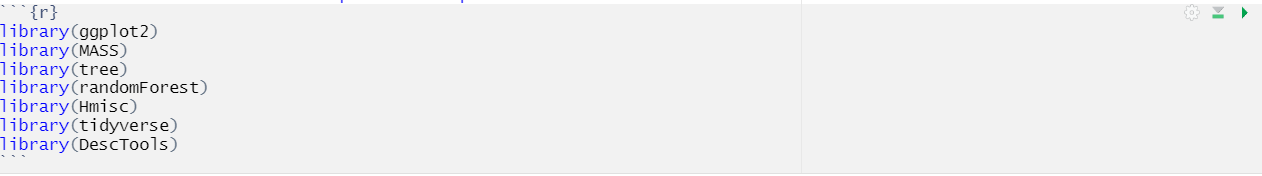




**GEELY**

**2.1 Gathering Data**

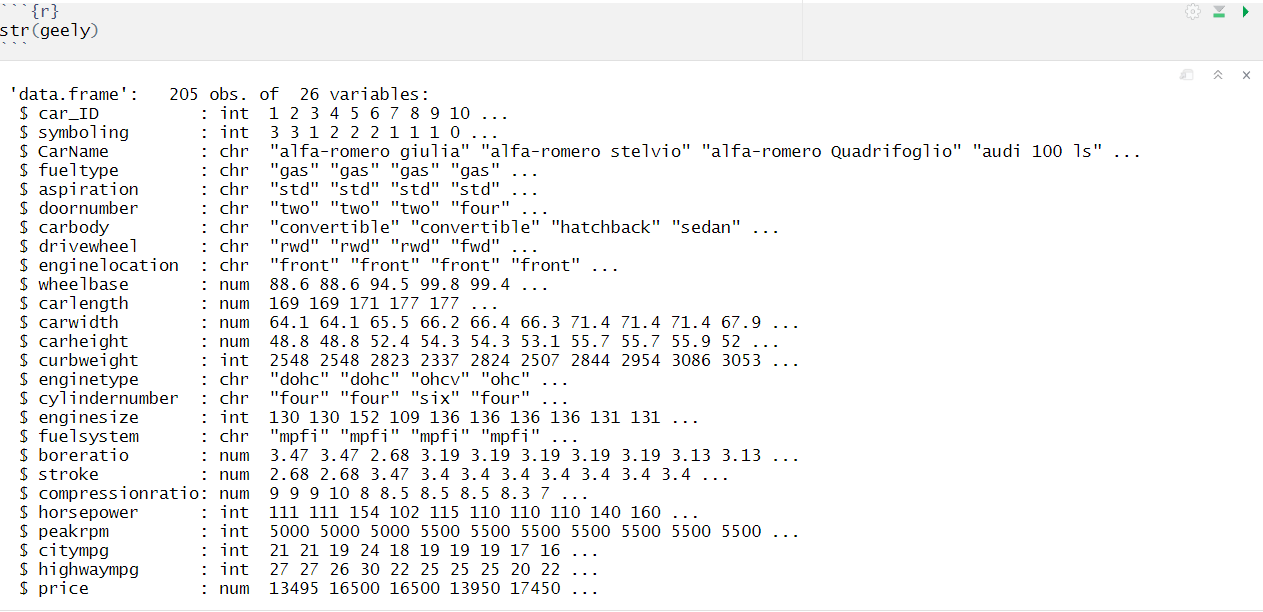
**Load the relevant Packages**



**Load the dataset**

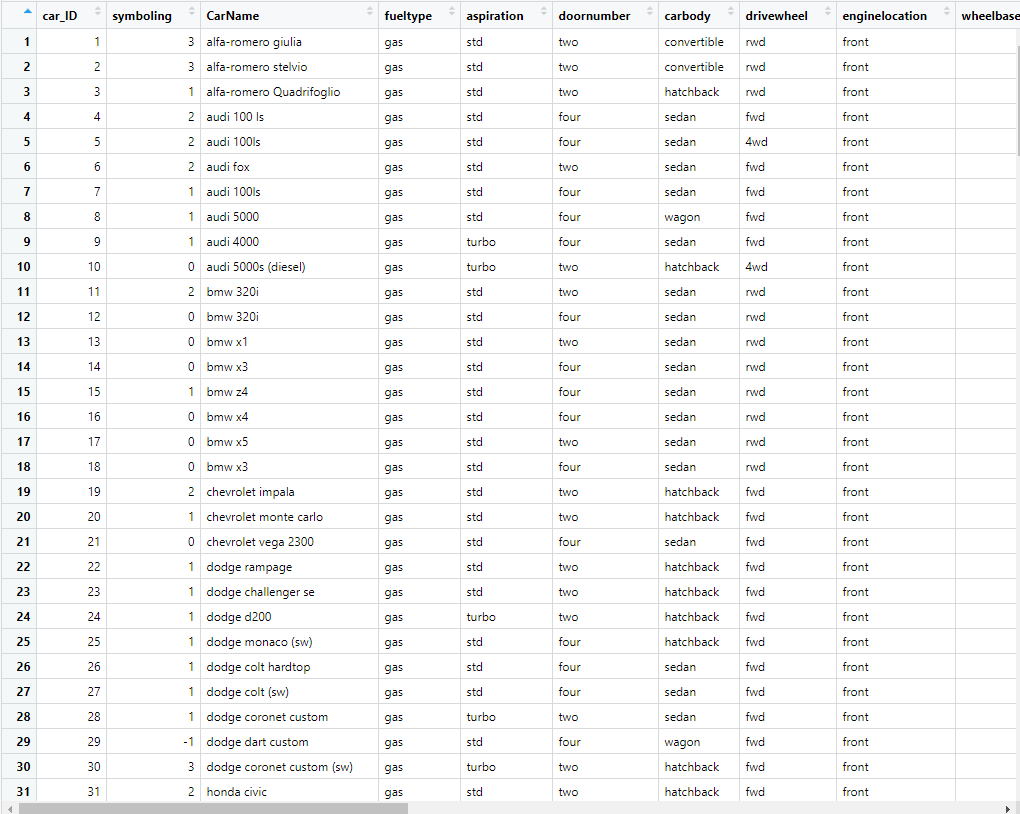


**Structure of the data**



**2.2** **Data description**

The geely dataset has 205 rows 26 column



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)

**Curbweight**

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**

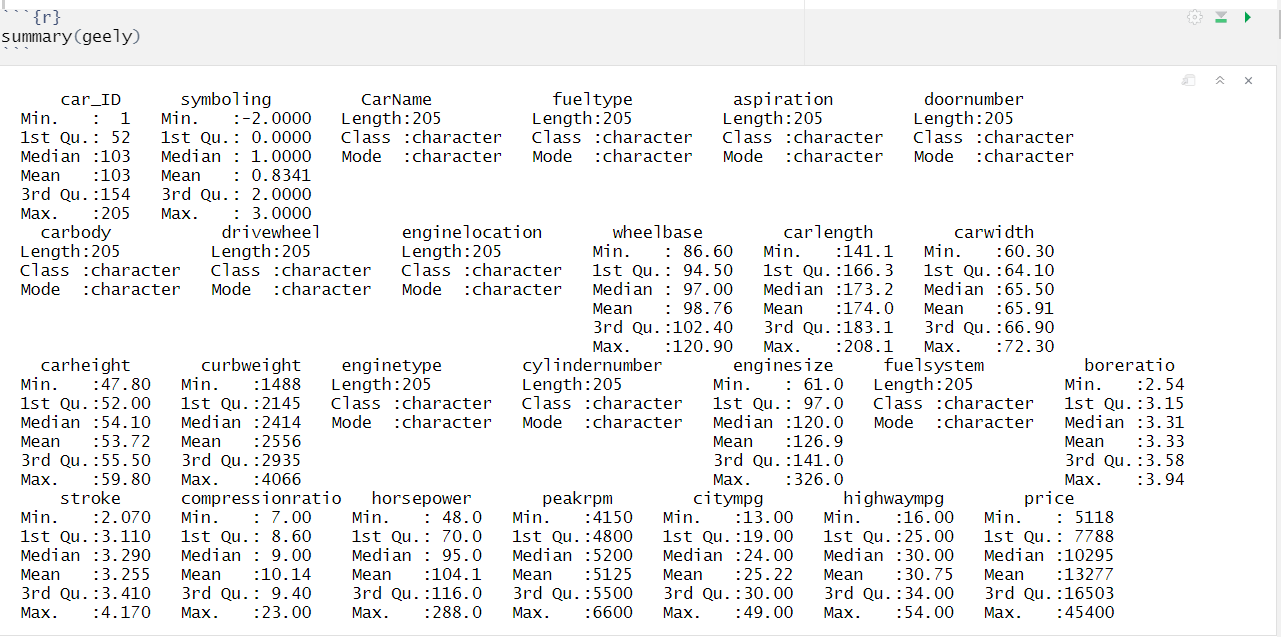
Mileage on highway (Numeric)

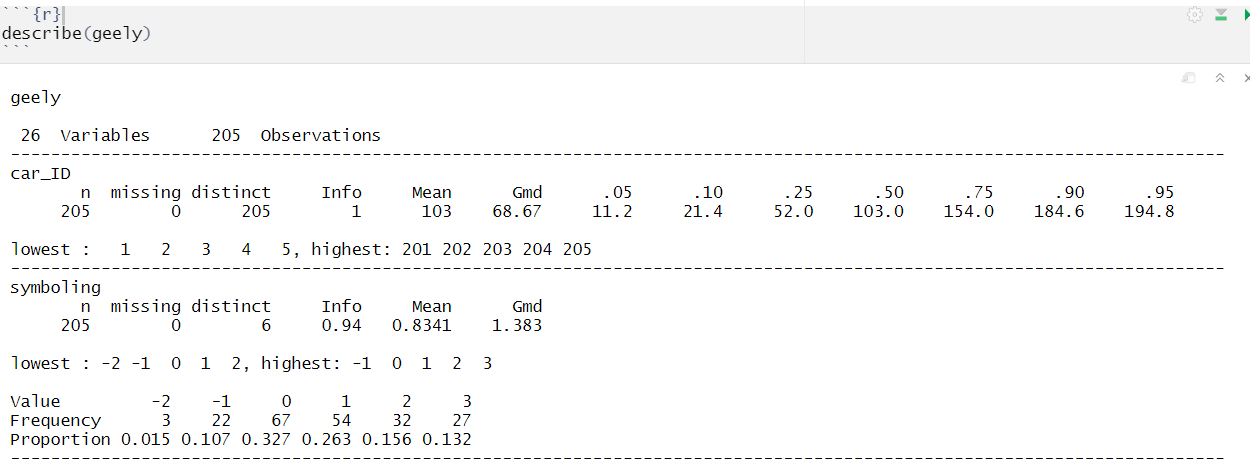
**Price (Dependent variable)**

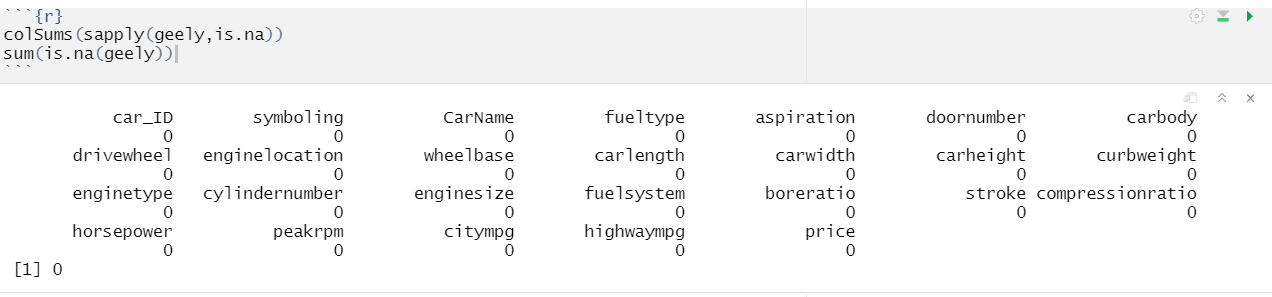
Price of car (Numeric)

**2.3 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.

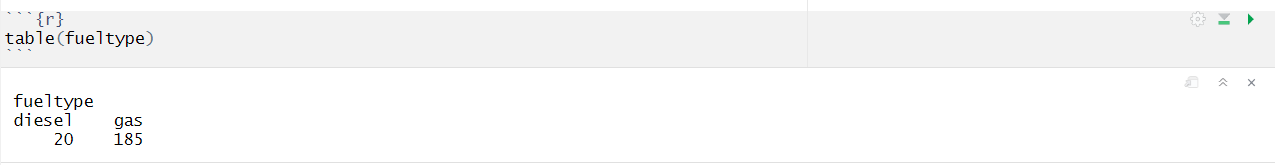


**Describe of dataset**

**NA’s in the dataset**

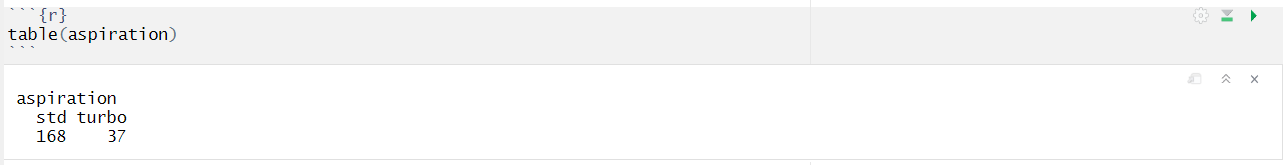
**Fueltype:**

Fueltype is a Categorical variable .There are two fuel type

* + - Diesel
    - Gas

**Aspiration :**

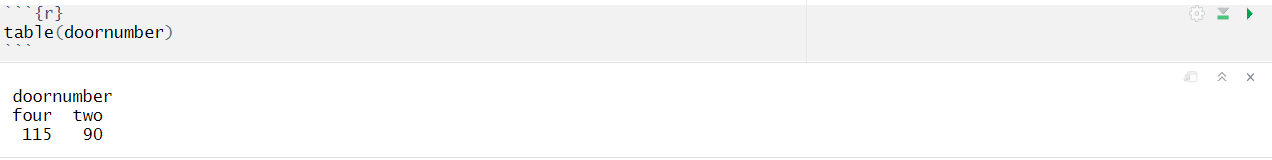
Aspiration is a Categorical variable . It is divided into two

* + - Std
    - Turbo

**Doornumber:**

Doornumberis a Categorical variable . It is divided into two

* + - Four
    - Two



**Carbody :**

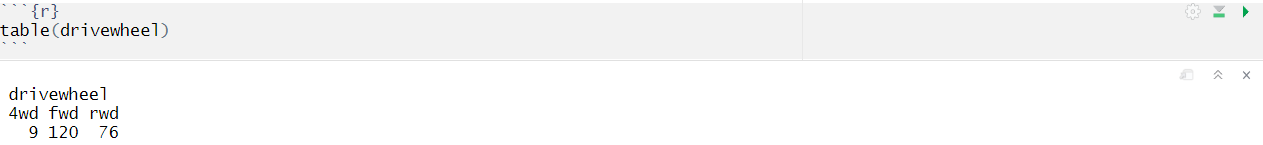
Carbodyis a Categorical variable . It is divided into five

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



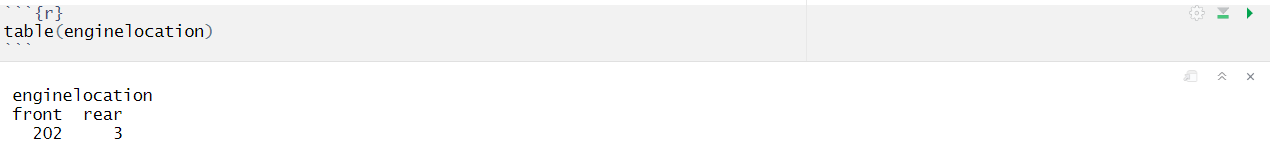
**Drivewheel:**

Drivewheel is a Categorical variable . It is divided into three

* + 4wd
  + Fwd
  + rwd

**Enginelocation :**

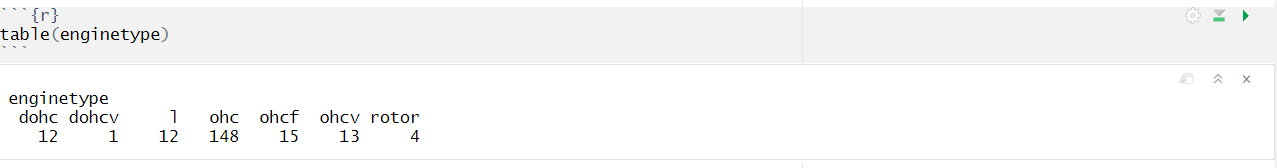
Enginelocation is a Categorical variable . It is divided into two

* + - Front
    - Rear

**Enginetype :**

Enginetype isa Categorical variable . It is divided into seven

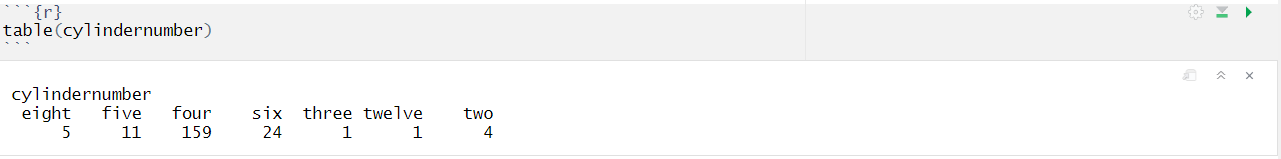
* + - Dohc
    - Dohcv
    - L
    - Ohc
    - Ohcf
    - Ohcv
    - rotor



**Cylindernumber:**

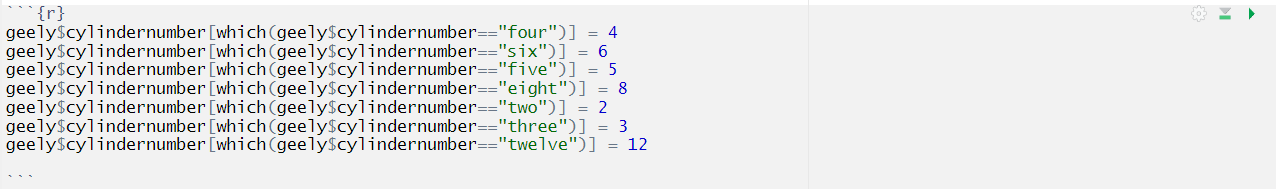
Cylinder numberis a Categorical variable . It is divided into seven

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



**2.4 Data Cleaning**

**Changing the Cylinder number Categorical to numerical**

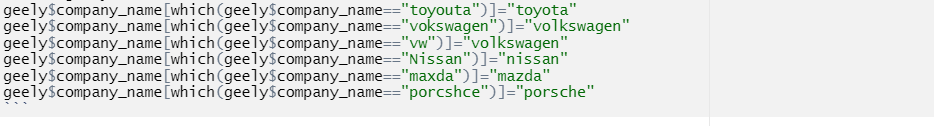


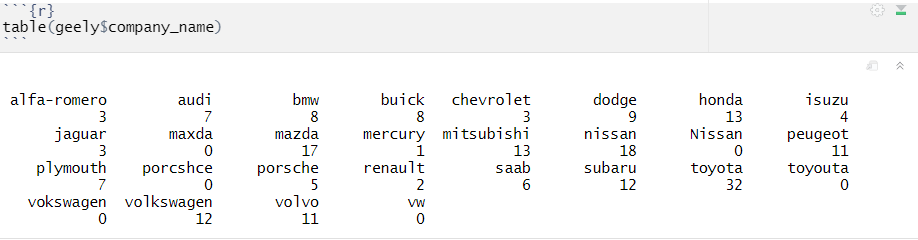
**Separating the Companyname and model**



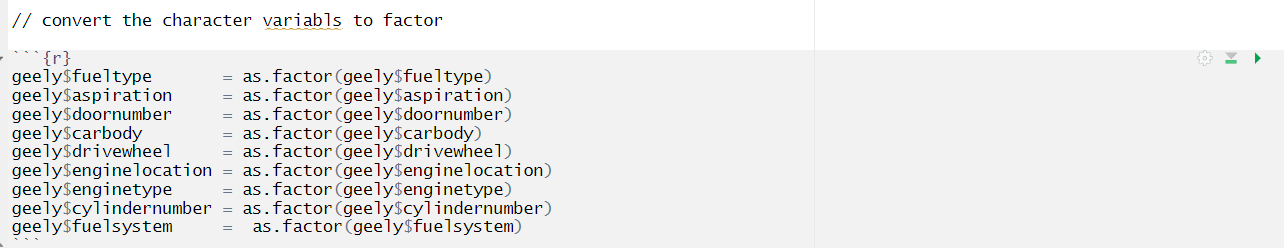
**Changing the companyname as factor**



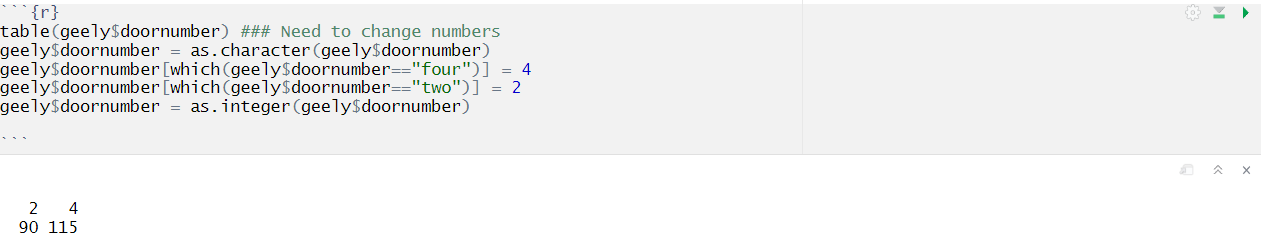
**Changing the company name**



**Convert the character Variable to factor**



**Convert the door number categorical to numeric**



**CHAPTER III**

**EDA USING R**

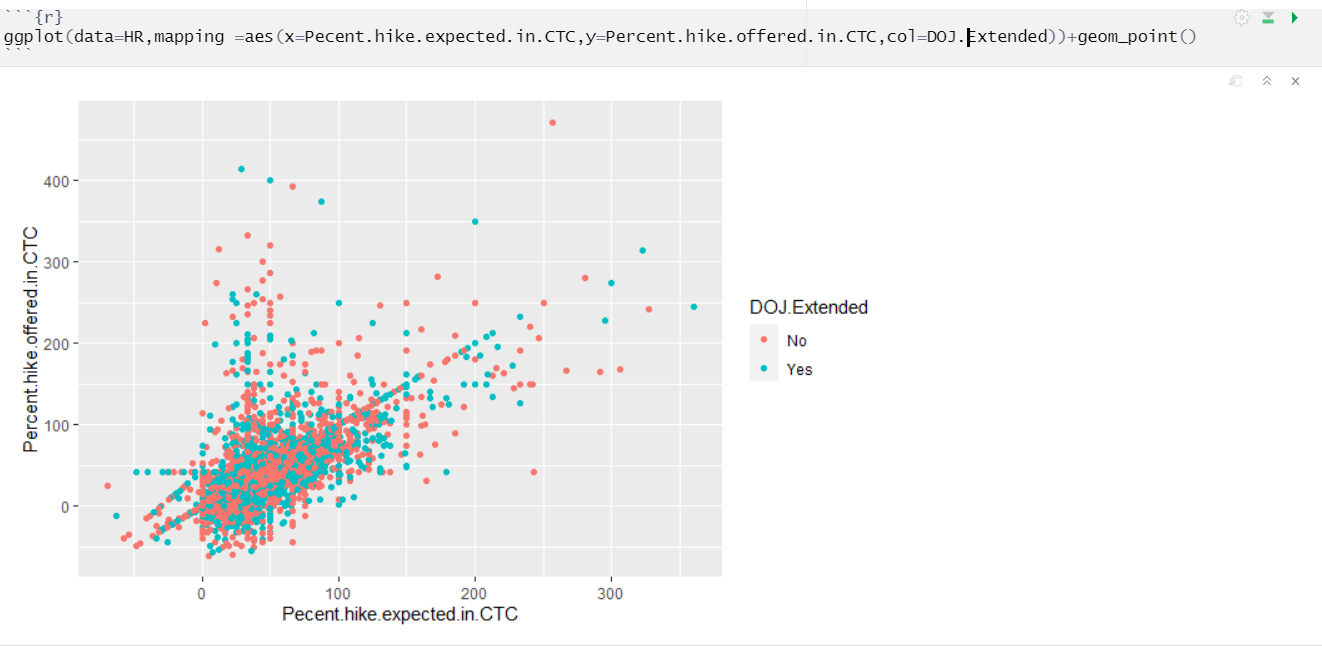
**3.1 Exploratory Data Analysis**

* When you first get your 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). Your EDA should involve creating and reviewing many plots/graphs and considering the patterns and relationships you see.

**HR ANALYTICS**

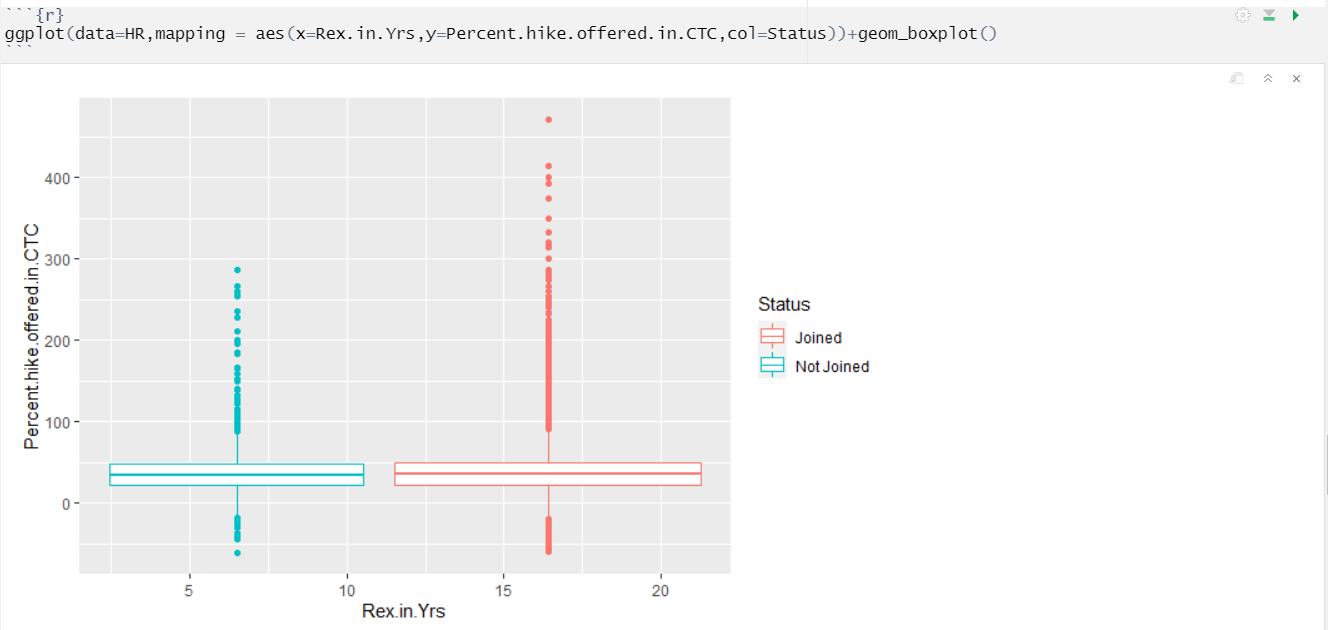
**Plot 1**

This plot clearly explains the companies offered jobs vs company expected jobs according to the DOJ of the job workers.



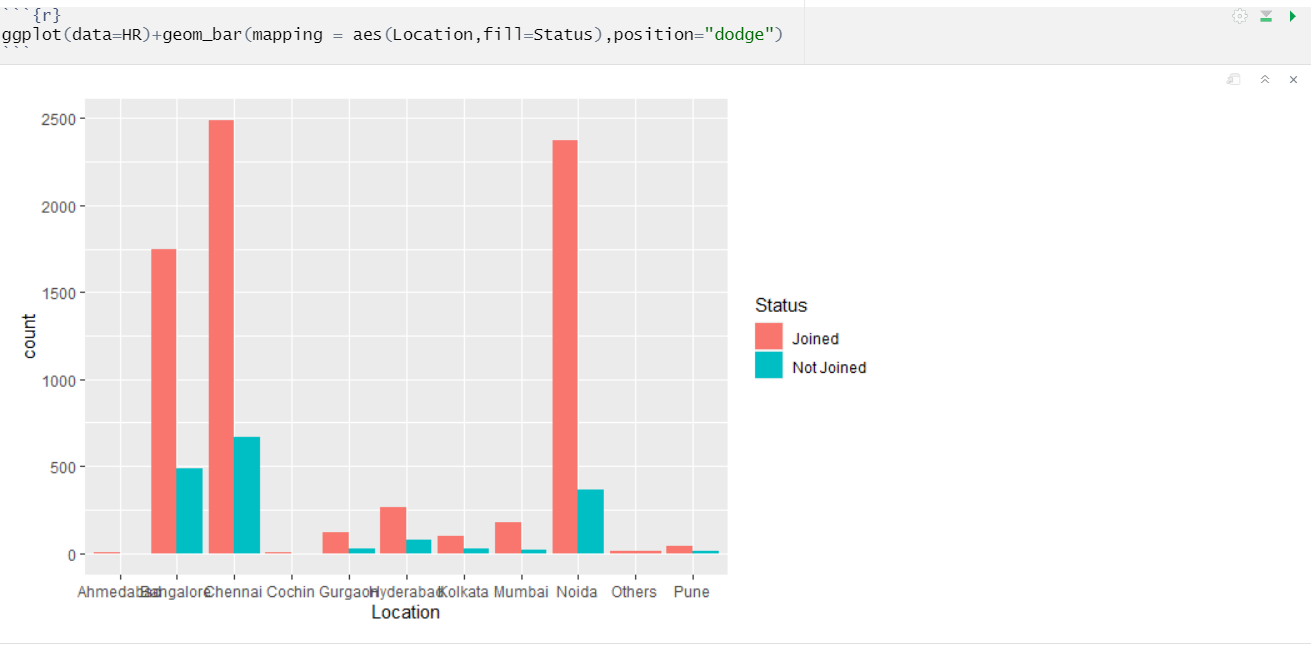
**Plot 2**

Experience of candidates vs percent of the offered jobs .



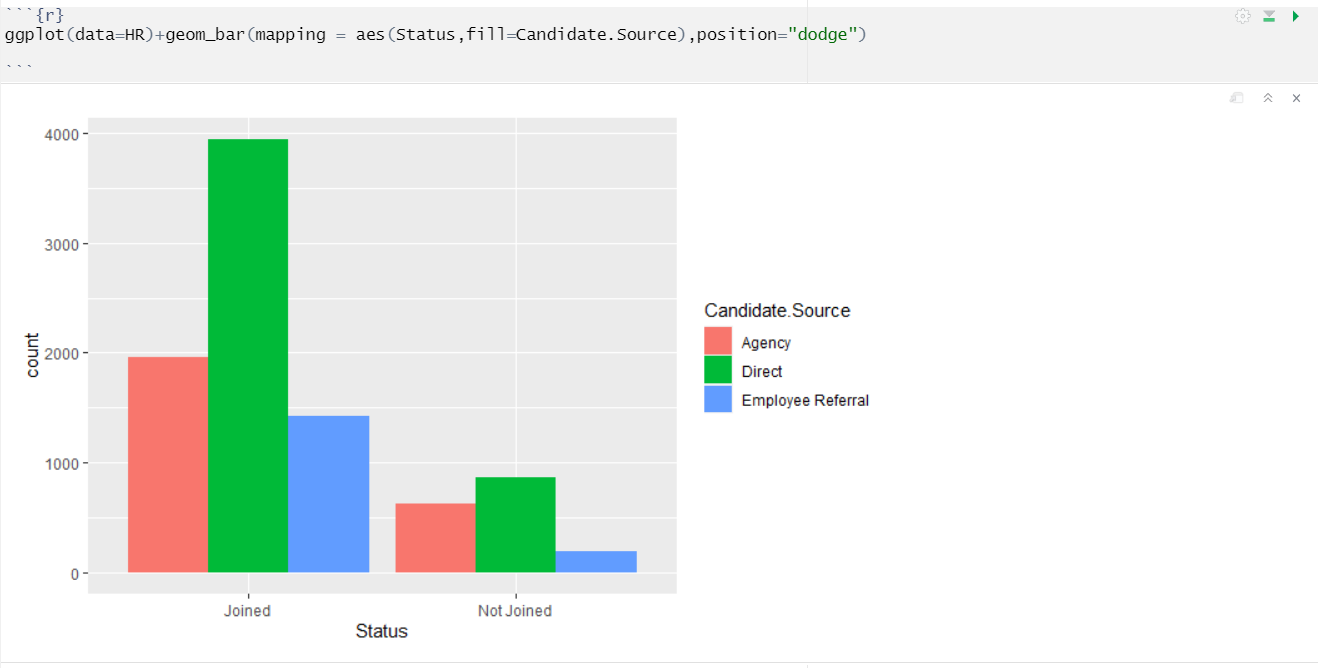
**Plot 3**

The count of joined not joined in each city.



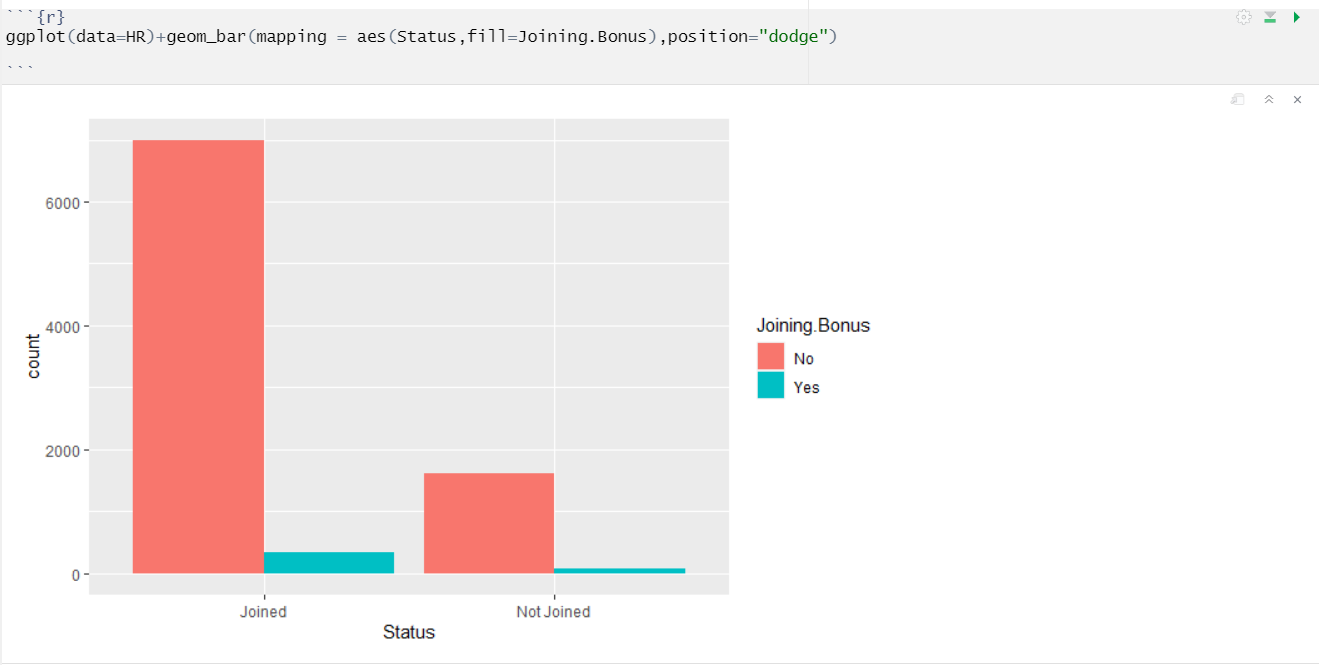
**Plot 4**

The count of joined and joined workers who are all came from agency or direct or employee referral.



**Plot 5**

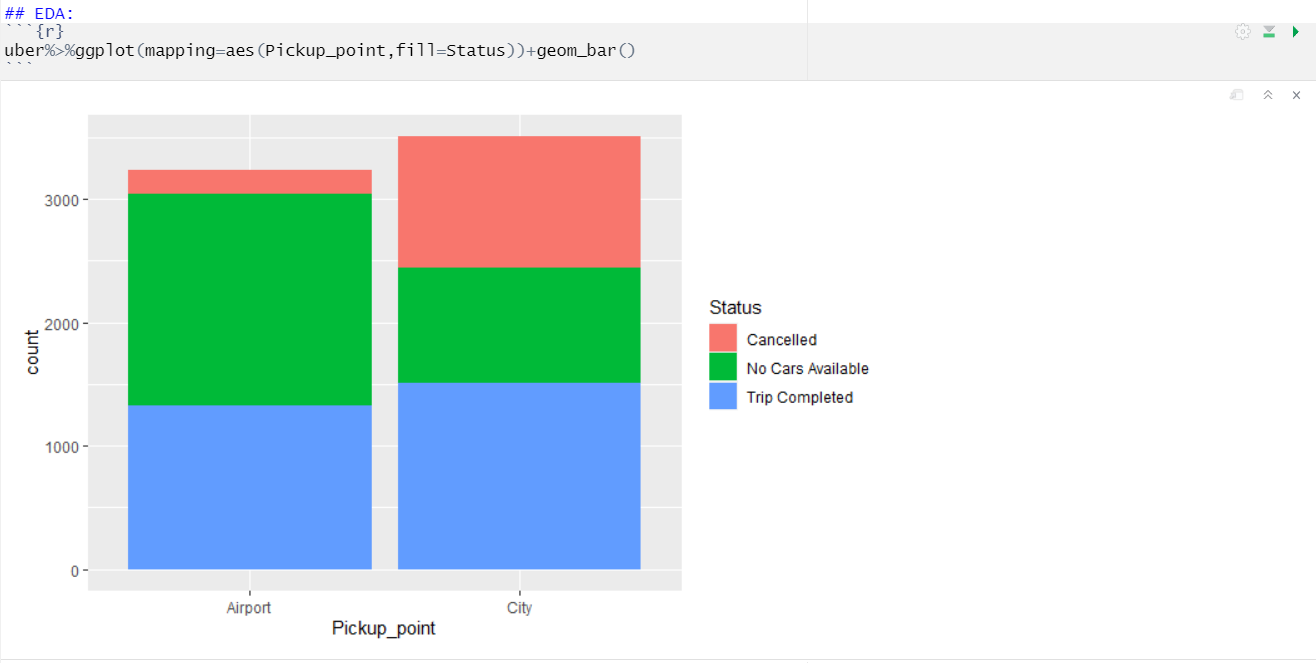
The count of joined and not joined according to the joining bonus.



**UBER**

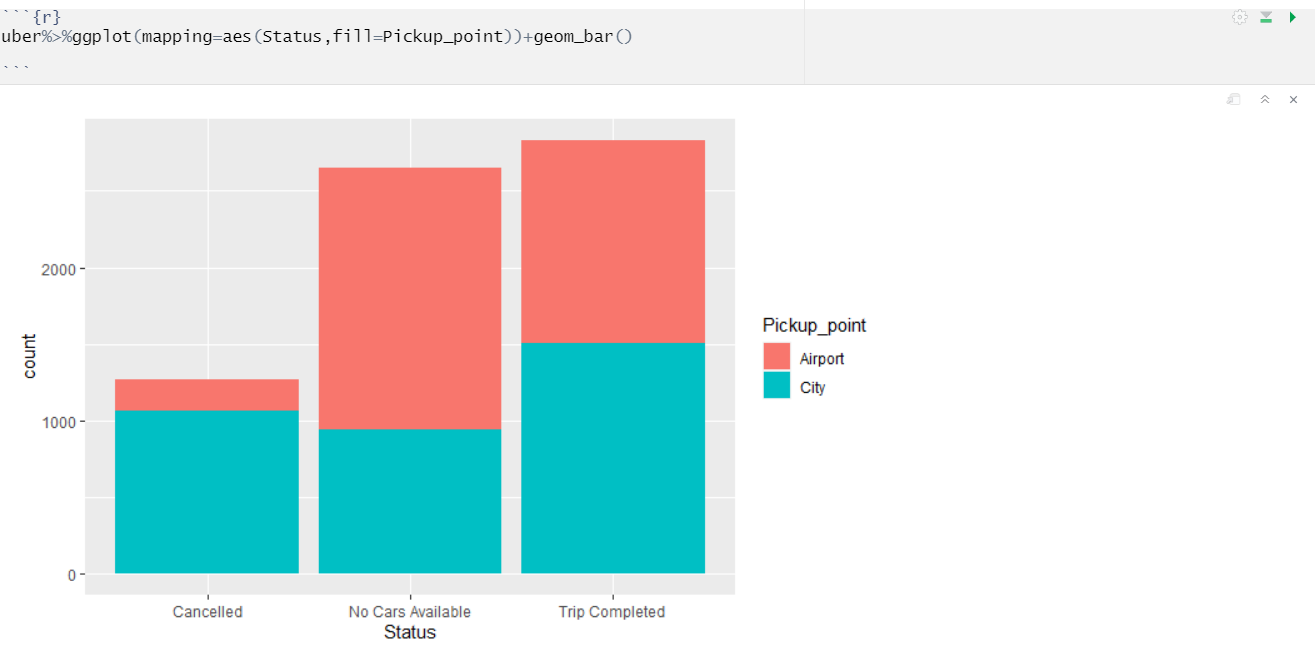
**Plot 1**

This plot clearly explains that the Pickup\_Point count according to the count of Status which divided into three types [Cancelled , Trip Completed , No Cars Available ] .



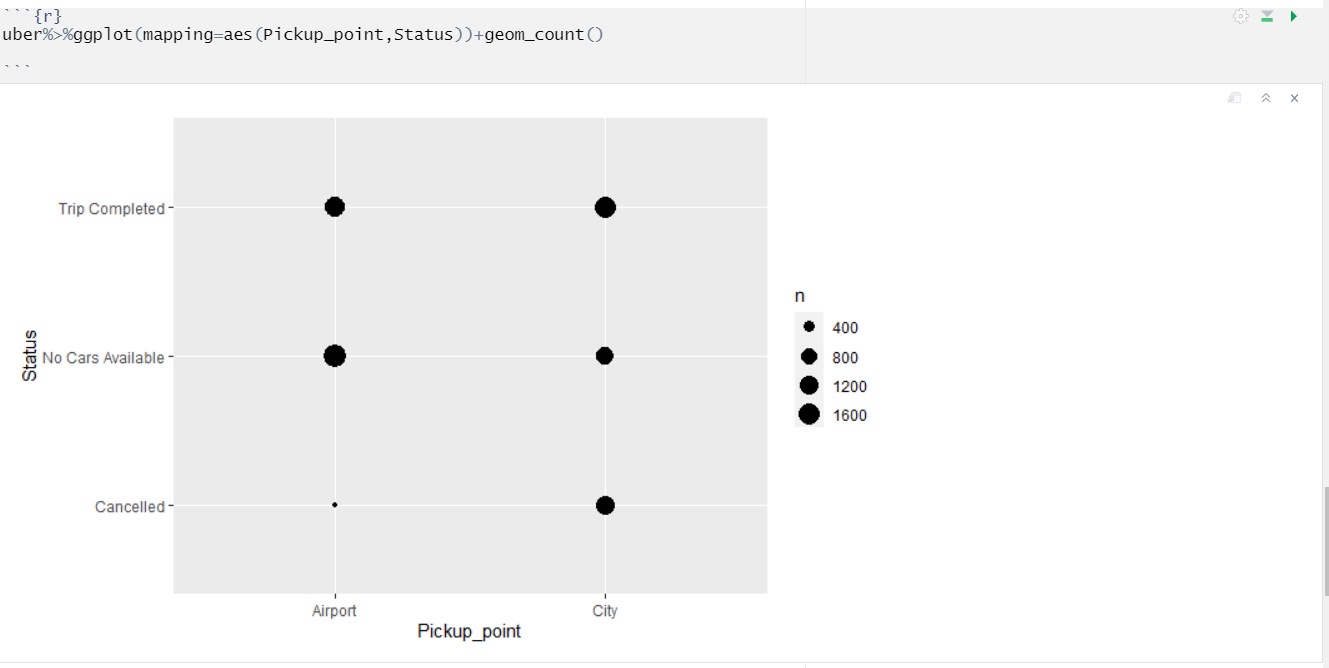
**Plot 2**

This plot clearly explains that the Status count according to the count of Pickup\_count which divided into two types [Airport , City] .



**Plot 3:**

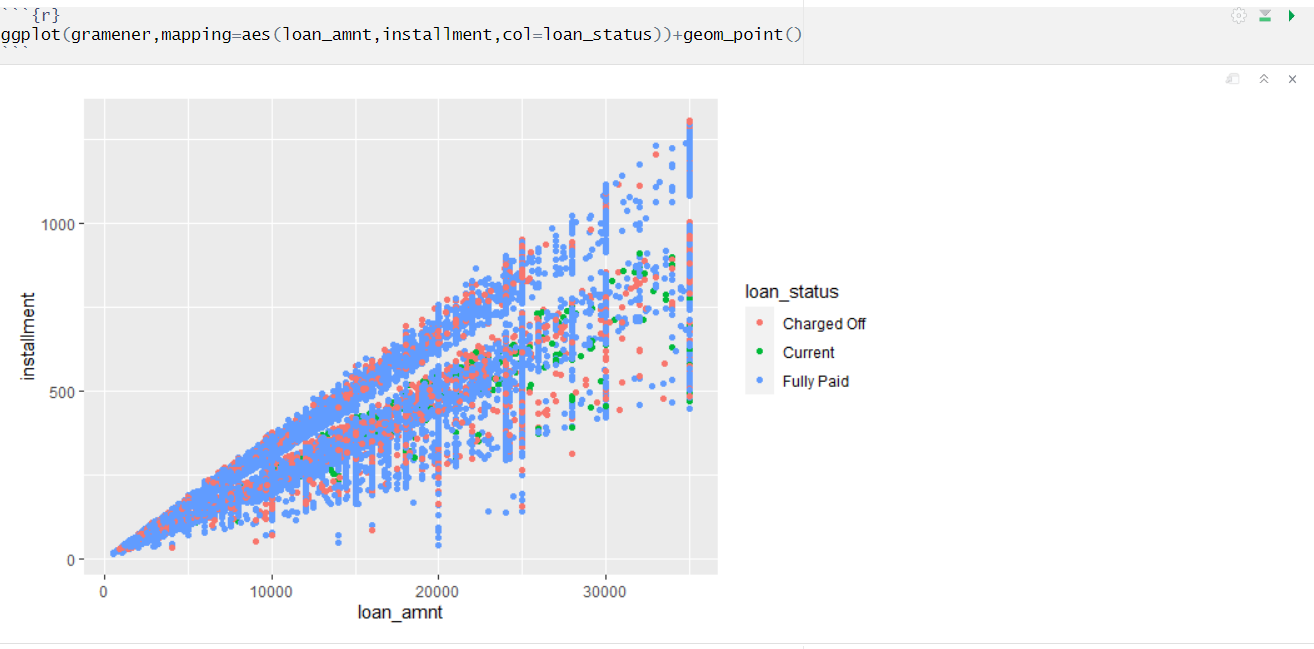
How many Cancelled and number of cars available in airport and city using geom\_count () plot.



**GRAMENER**

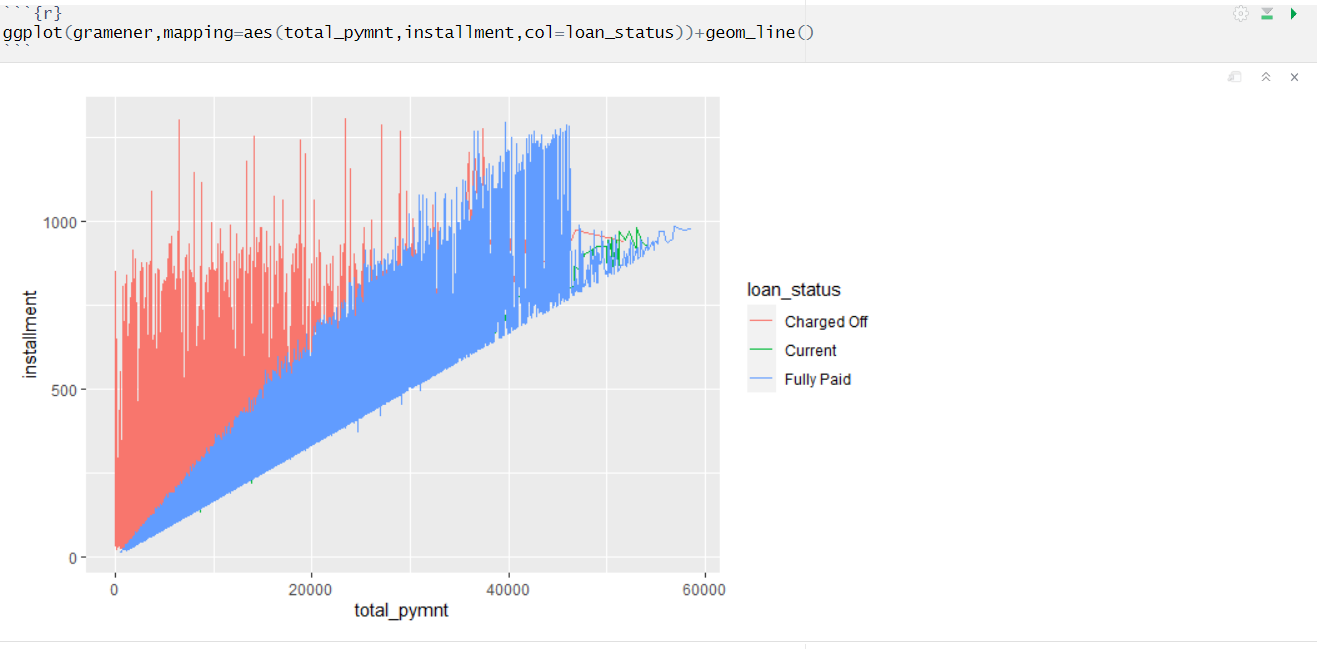
**Plot 1 :**

Loan amount vs installment segregated according to loan\_status



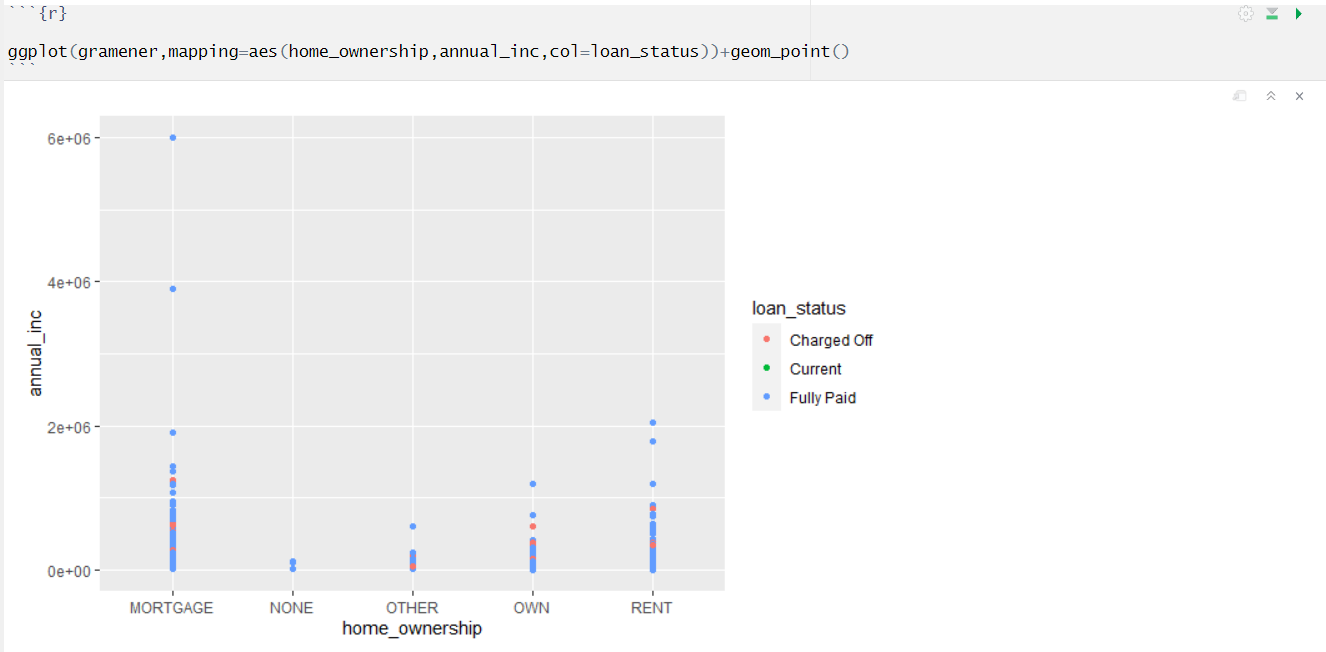
**Plot 2:**

Total payment vs installment due according to loan\_status .



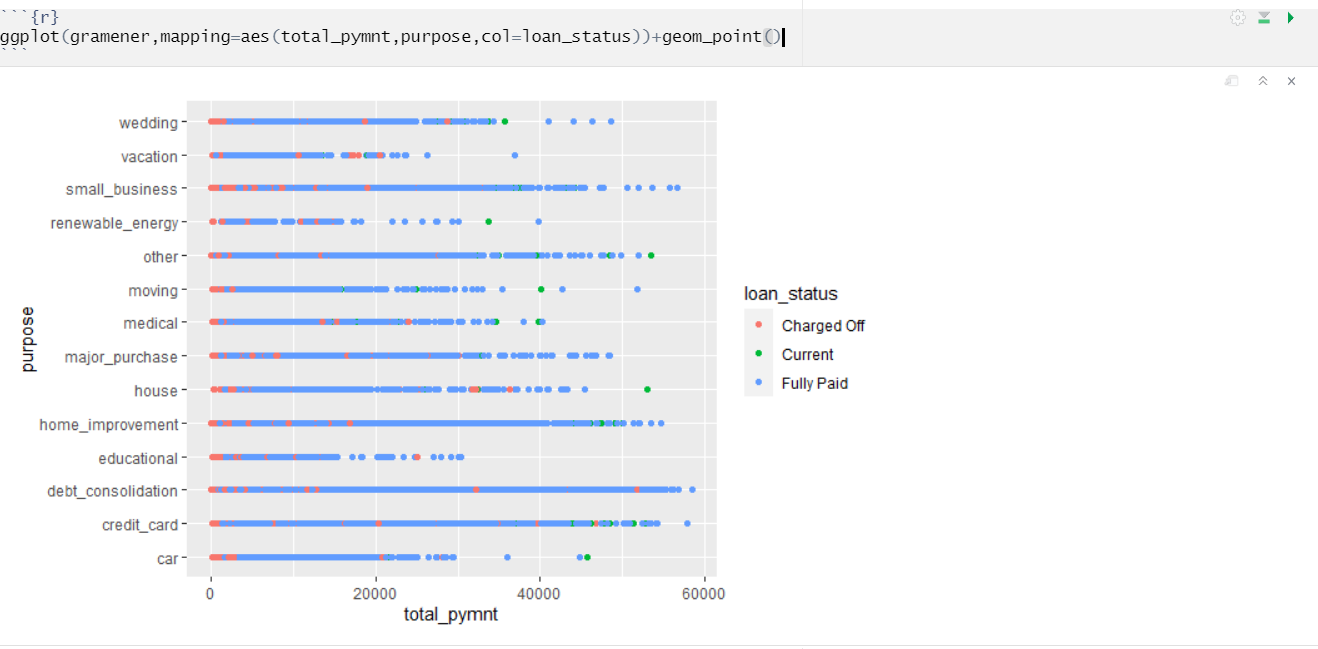
**Plot 3:**

Annual income of home owners .



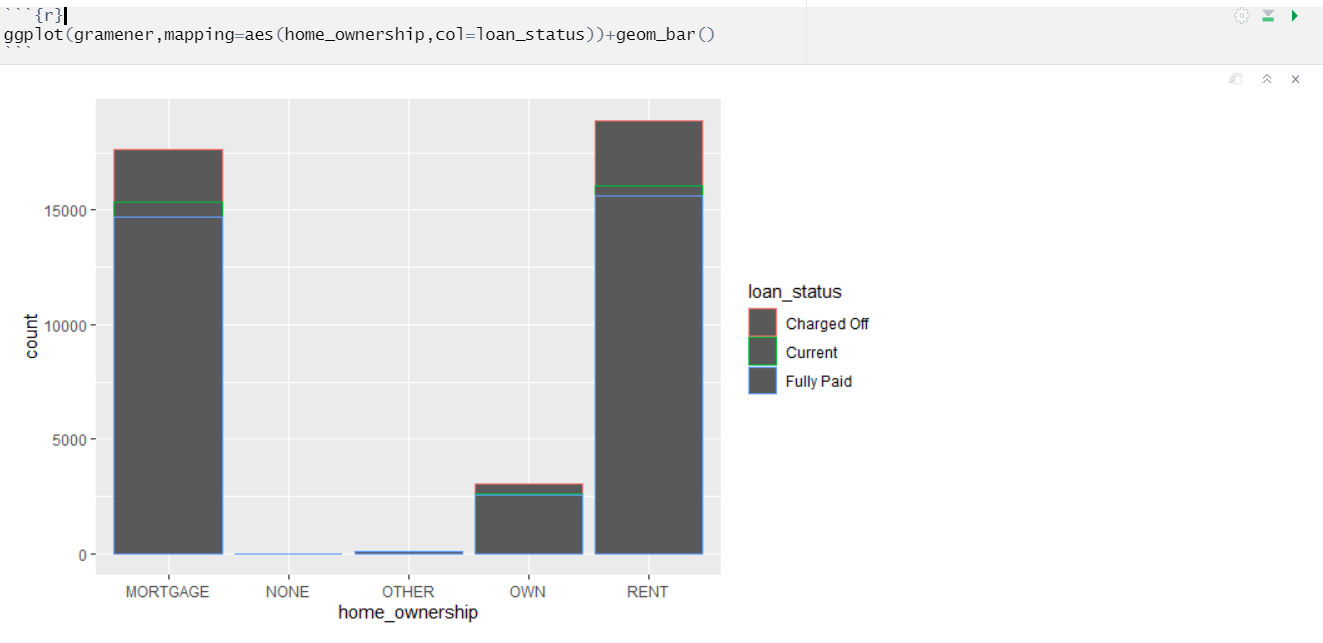
**Plot 4 :**

To check whether the relationship between Total payment and purpose



**Plot 5 :**

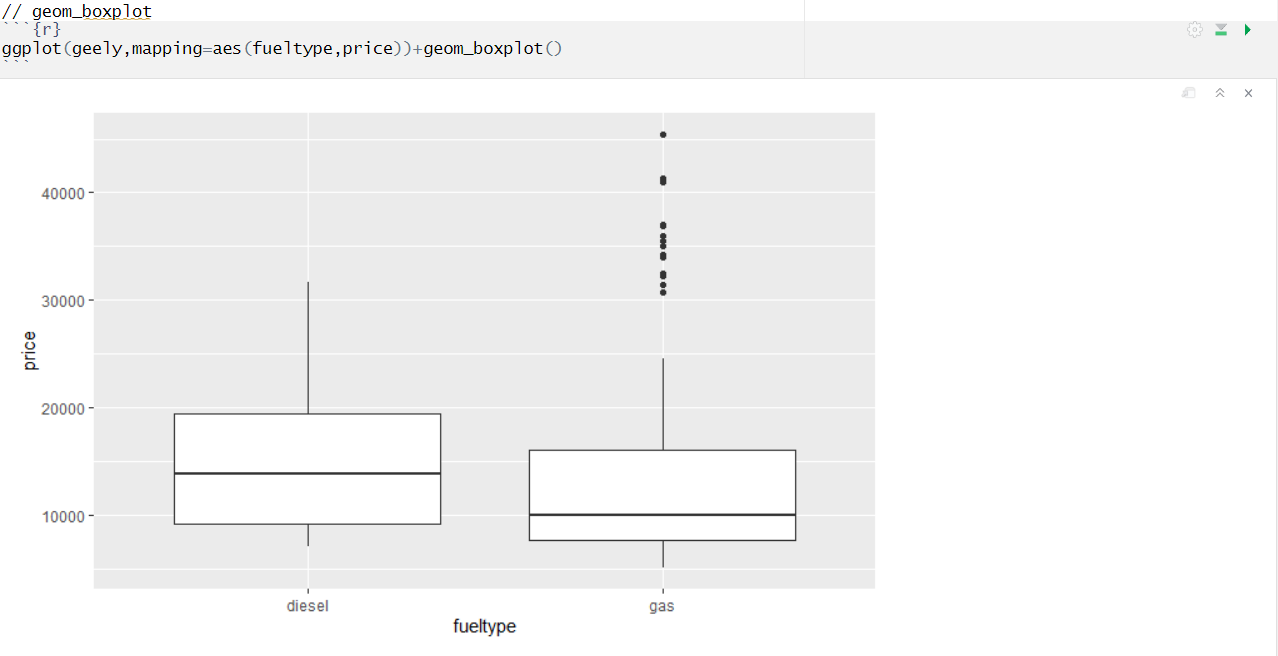
The count of home\_ownership according to Mortgage ,None,other,own and rent .



**GEELY**

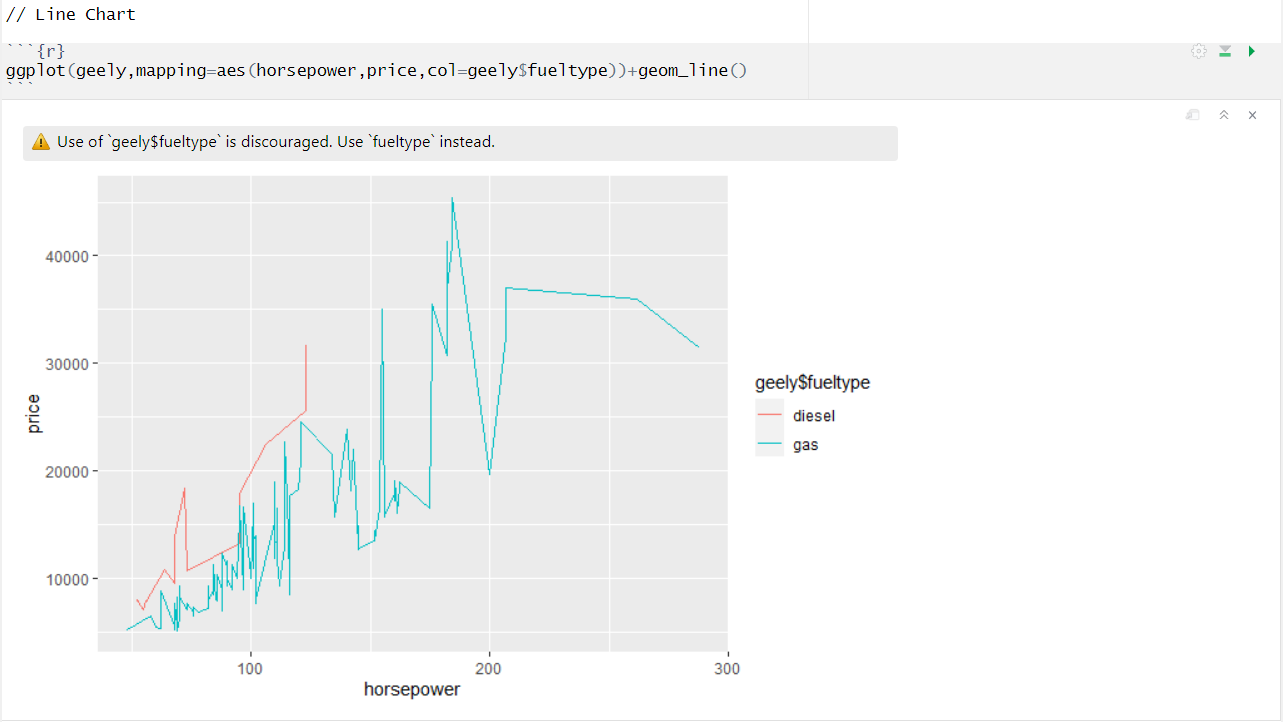
**Plot 1**

To find fuel type & price have a relationship or not



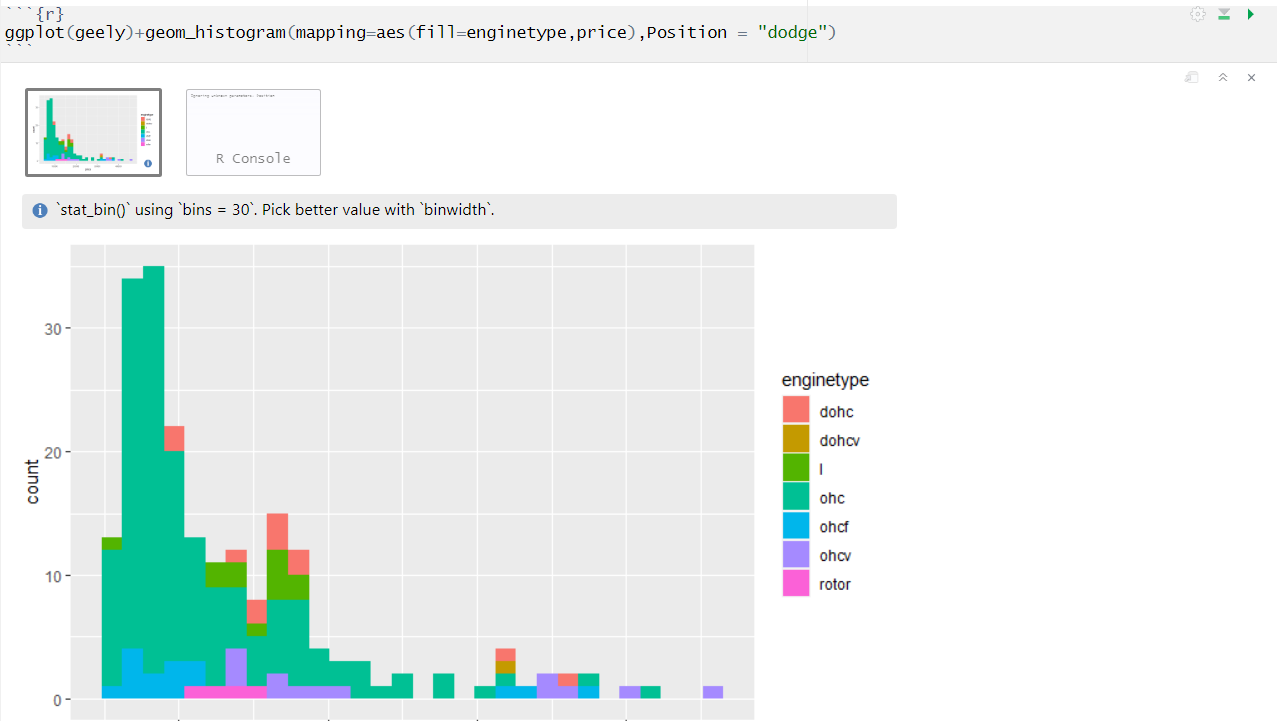
**Plot 2**

Horsepower (vs) price according to fueltype



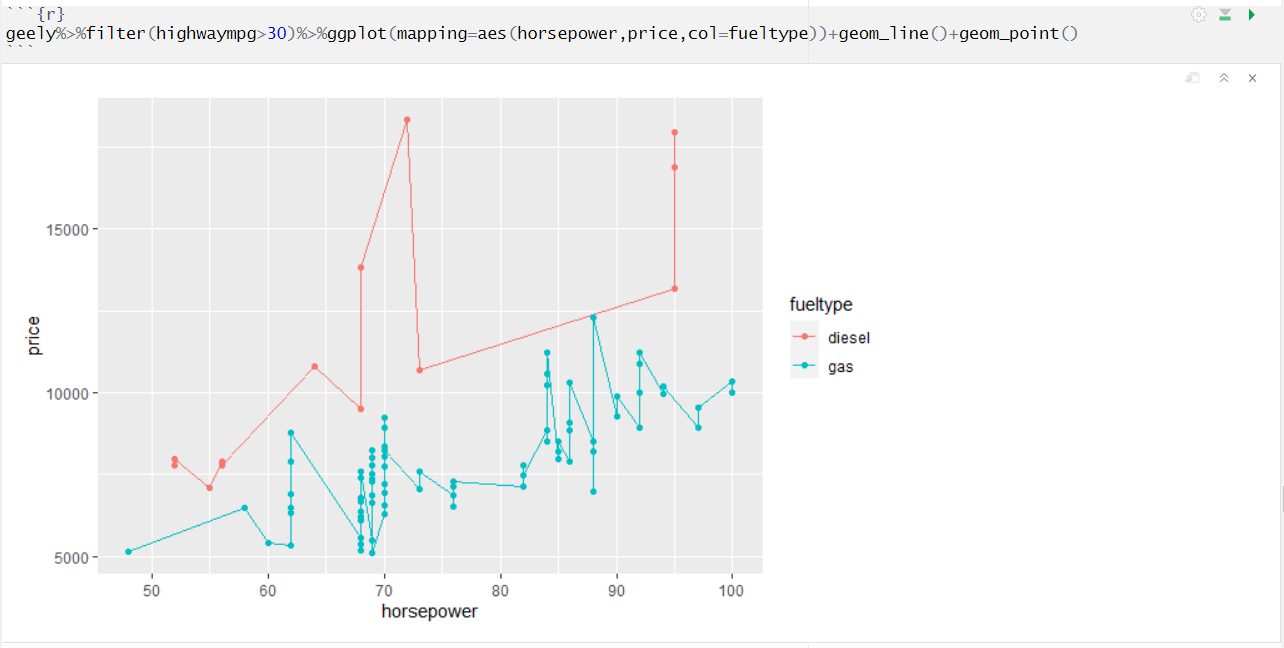
**Plot 3**

Engine type prices



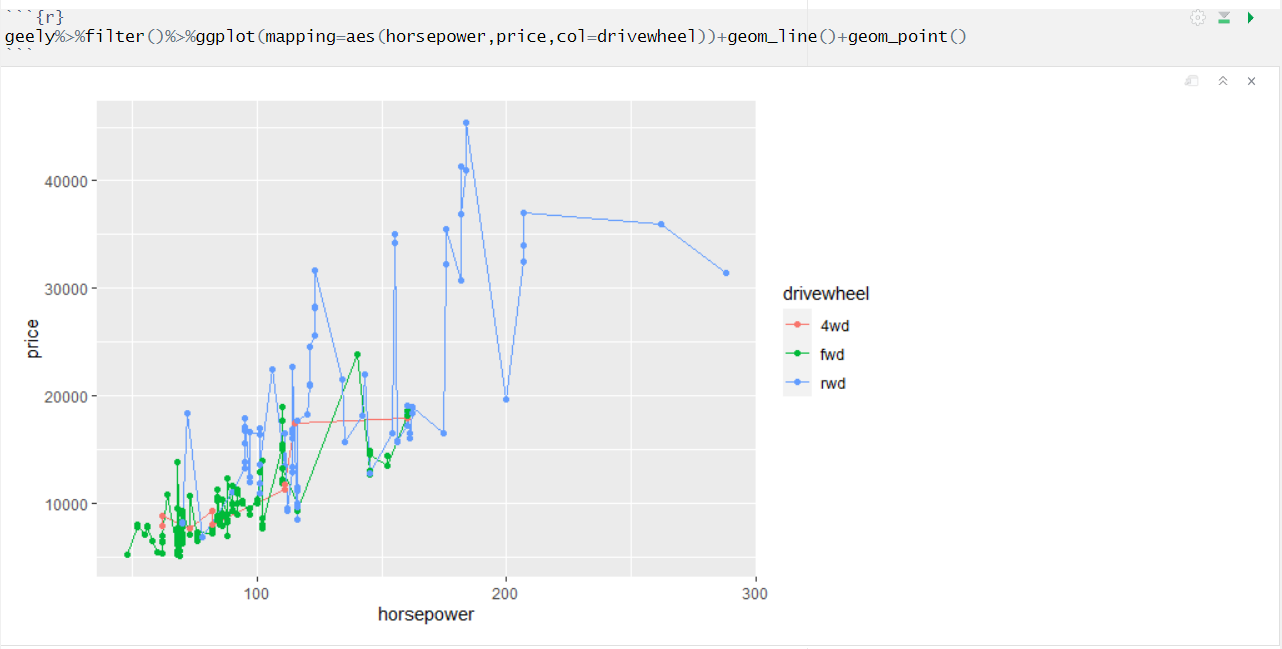
**Plot 4**

Horsepower (vs) price according to that fuel type



**Plot 5**

Horsepower (vs) price according to that wheel type



**CHAPTER IV**

**MODEL BUILDING**

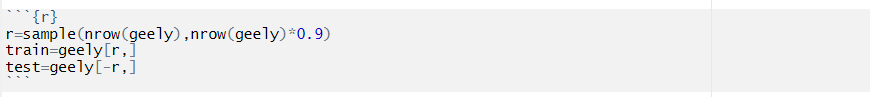
**4.1 Choosing the model**

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

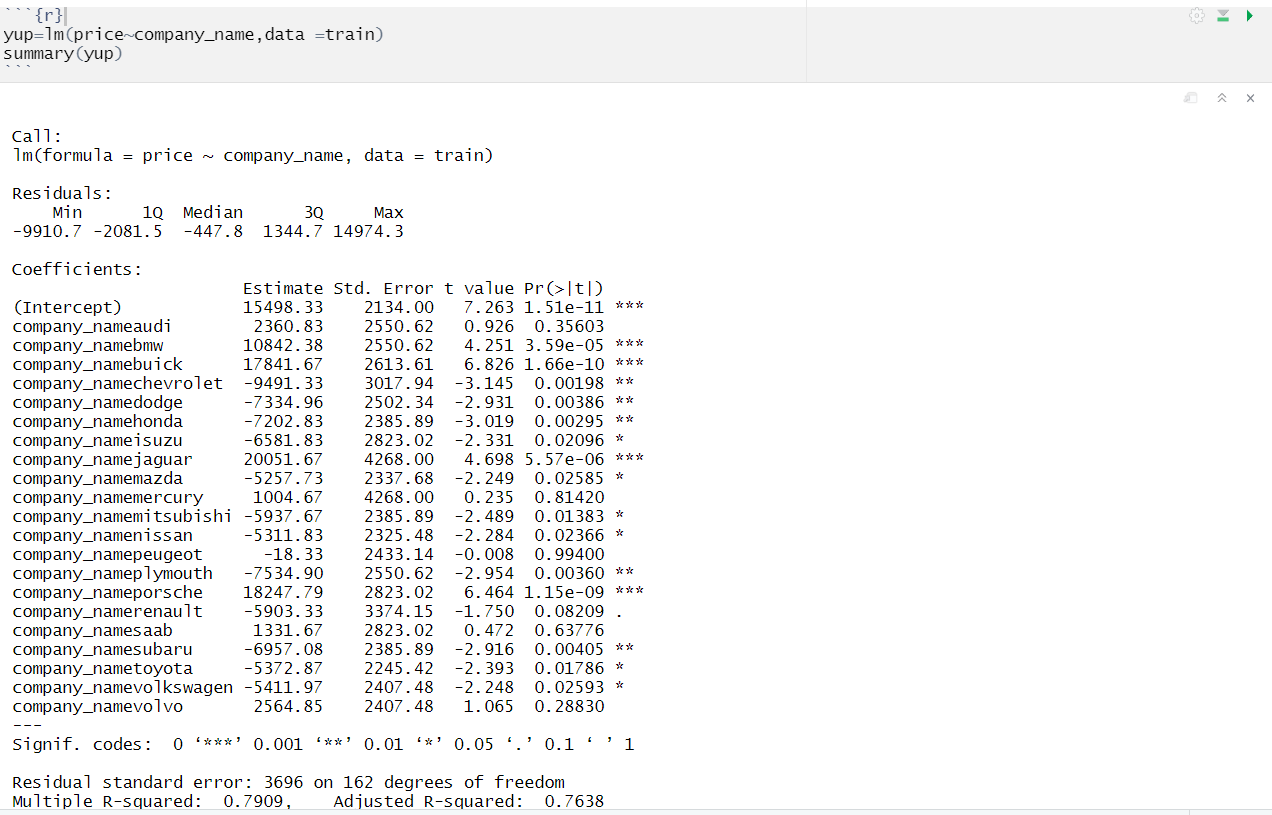
**Divide the dataset into Training data and Testing data**

* Typically, when you 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.

****

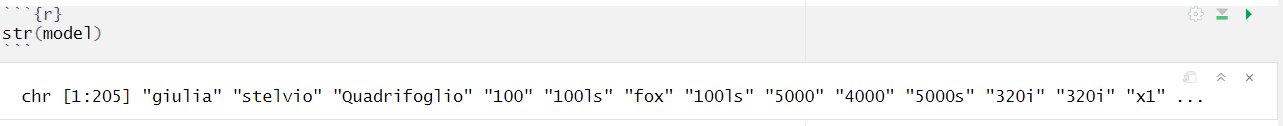
**Linear Regression**

* Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables.
* Does a set of predictor variables do a good job in predicting an outcome (dependent) variable.
* Which variables in particular are significant predictors of the outcome variable
* Three major uses for regression analysis are determining the strength of predictors, forecasting an effect, and trend forecasting.

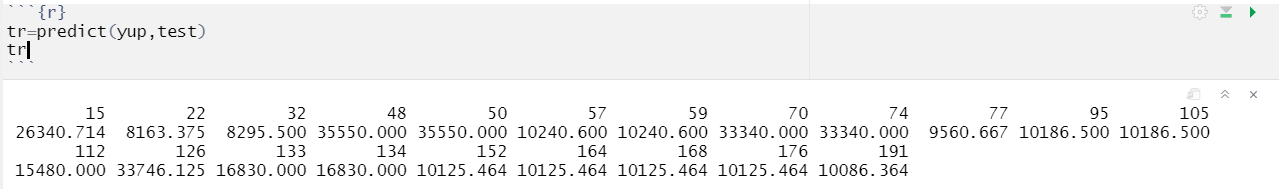


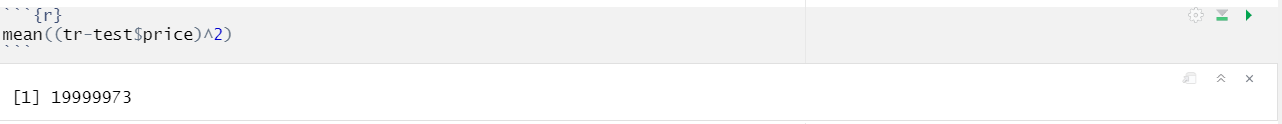
From the model we have learnt BMW,buick,Chevrolet,dodge,iszu,jaguar,mitsubushi,Plymouth,Porsche,Subaru,toyato,Volkswagen are the significant values of the model

* + If bmw value change it will affect in price 10842.38 increase
  + If buick value change it will affect in price 17841.67 increase
  + If chevrolet value change it will affect in price -9491.33 decrease
  + If dodge value it will affect in price 1-7334.96 decrease
  + If honda value it will affect in price -7202.83 decrease
  + If jaguar value it will affect in price 20051.67 increase
  + If porsche value it will affect in price 18247.79 increase
  + If subaru value it will affect in price -6957.08 decrease
  + If toyato value it will affect in price -5372.87 decrease
  + If volkswagen value it will affect in price -5411.97 decrease







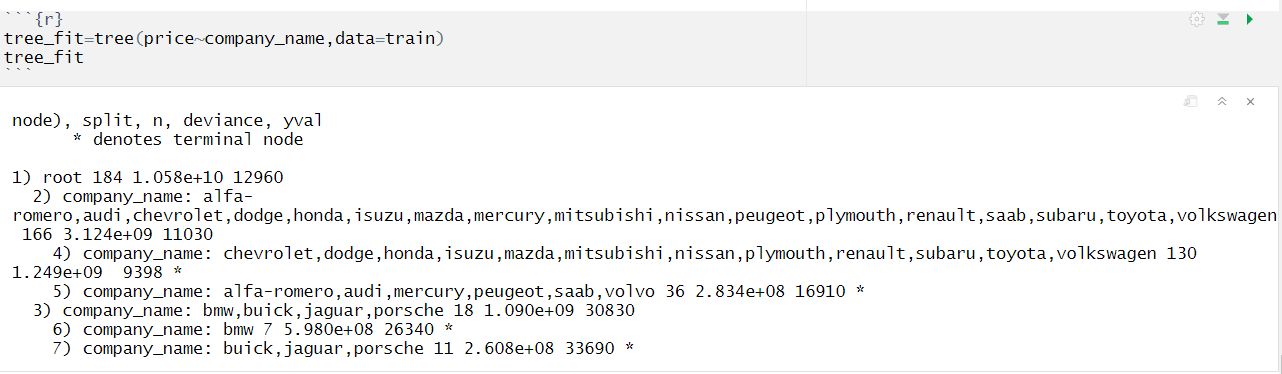


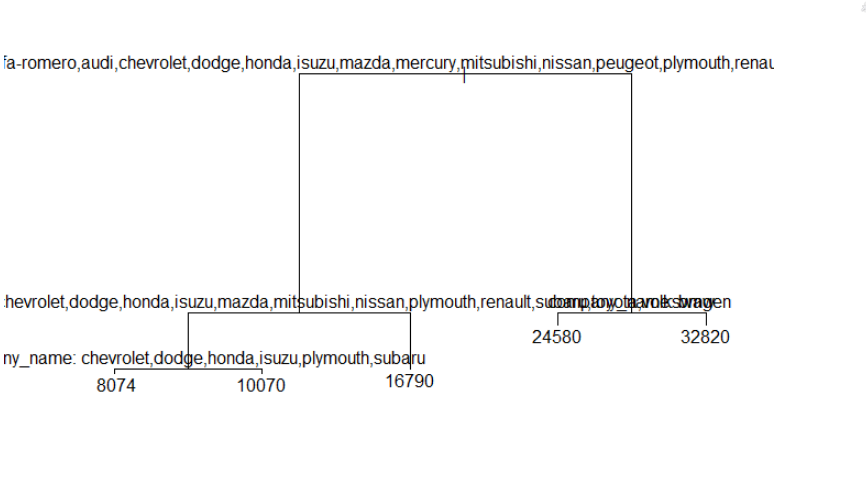
According to Linear regression error rate will be **19999973 .**

**Decision Tree:**

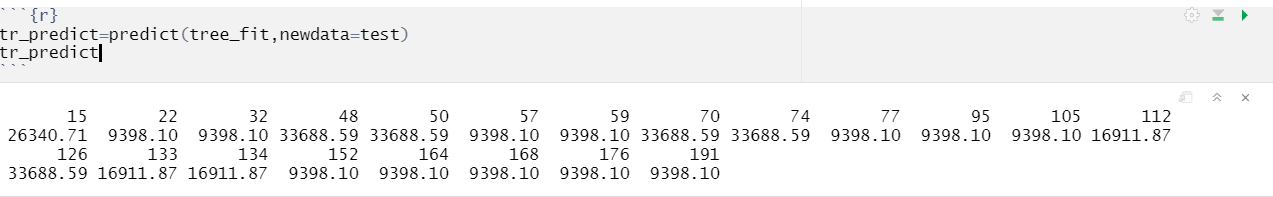
* Decision Tree Algorithm is a supervised Machine Learning Algorithm. It is an approach to predictive analysis that can help you make decisions. Decision tree goes down in a tree-structured format.
* Tree-based methods are simple and useful for interpretation. Decision trees can be applied to both regression 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.

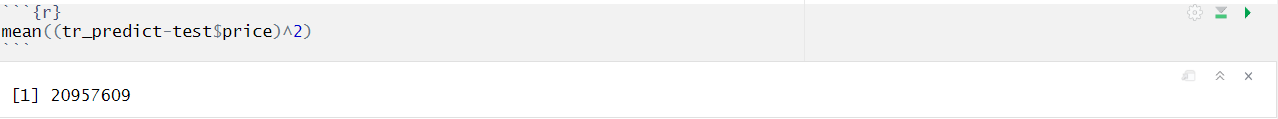
**Tree Building**

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.



* The average price of dodge,Honda,isuzu,mazdawill be 9328
* The average price of nissan,Plymouth,Mitsubishi will be 16630
* The average price of Renault,Subaru,toyato,saab will be 26120
* The average price of Volkswagen,BMW will be 33120
* Totally it consider 4 leaf nodes from the root node.

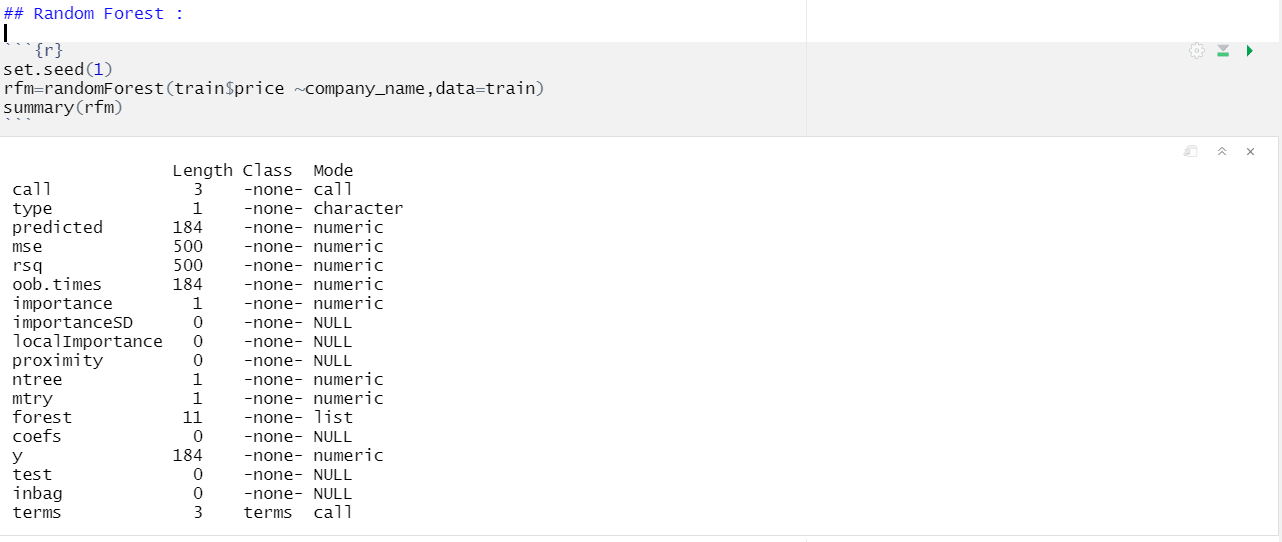


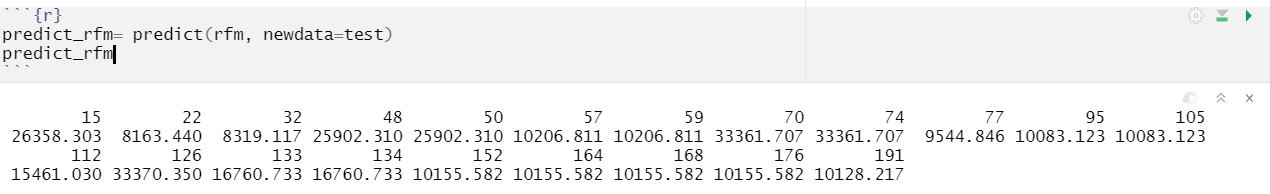


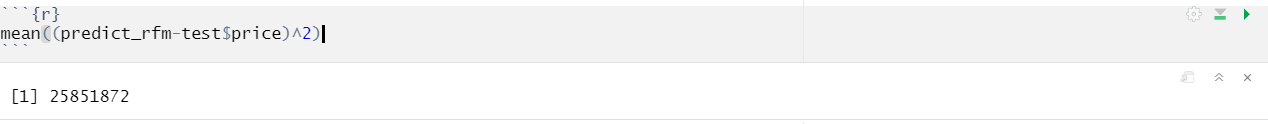
According to decision tree error rate will be **20957609**

**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.







According to random forest error rate will be **25851872**

**CHAPTER V**

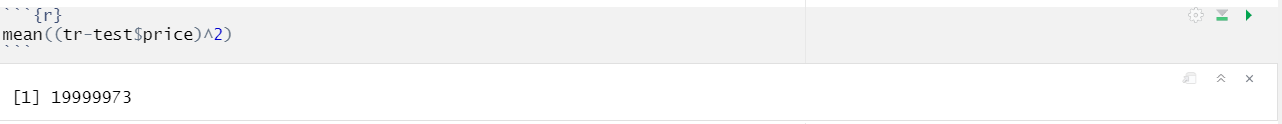
**PERFORMANCE EVALUATION**

* 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.
* For example, classification accuracy for classification (total correct correction divided by the total predictions made multiple by 100 to turn it into a percentage).
* 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.

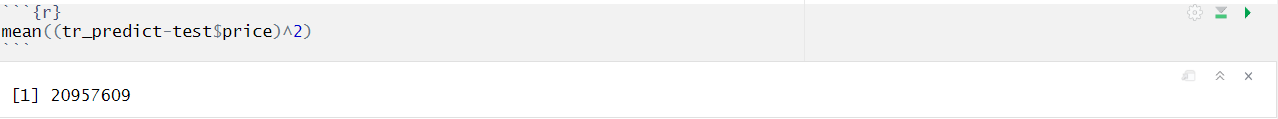
**MSE of Linear Regression**



According to decision tree error rate will be **20957609**

Accuracy rate will be **81%**

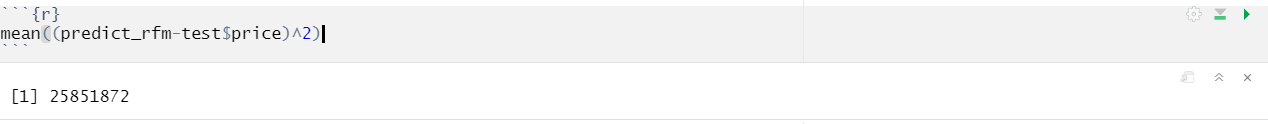
**MSE of Decision Tree**



According to decision tree error rate will be **20957609**

Accuracy rate will be **80%**

**MSE of Random Forest**

According to random forest error rate will be **25851872**

Accuracy rate will be **75%**

**CHAPTER VI**

**CONCLUSION**

* The automotive industry in the United States began in the 1890s and, as a result of the size of the domestic market and the use of mass production, rapidly evolved into the largest in the world. However, the United States was overtaken by Japan as the largest automobile producer in the 1980s, and subsequently by China in 2008.
* American manufacturers produce approximately 8–10 million units annually. Notable exceptions were 5.7 million automobiles manufactured in 2009 (due to [crisis](https://en.wikipedia.org/wiki/Effects_of_the_2008%E2%80%9310_automotive_industry_crisis_on_the_United_States)), while production peaked during the 1970s and early 2000s at levels of 13–15 million units
* A multivariate regression based solution is proposed to calculate selling price of each car available in the U.S.Market.
* The U.S. is currently second among the [largest manufacturer(s) in the world by volume](https://en.wikipedia.org/wiki/List_of_countries_by_motor_vehicle_production).
* According to our prediction random forest tree is the best method to predict cars selling price.

# REFERENCES

[1]V. Lazarov and M. Capota. Churn Prediction. Business Analytics Course. TUM Computer Science, December 2007. <http://home.in.tum.de/~lazarov/files/research/pap> ers/churn- prediction.pdf

1. Carlo Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making,

John Wiley & Sons, Ltd. 2009 ISBN: 978-0-470- 51138-1

1. S. Gotovac. “Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services,” vol. 51, no. 3, pp. 275-283, 2010
2. H. Kim, and C. Yoon, “Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market.” Telecommunications Policy. Vol. 28 No.: PP. 751-765, 2004.

[5]W. Au, C. Chan, and X. Yao, A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction, IEEE transactions on evolutionary computation, 7, 6, 532-545, 2003.

[6] R. Behara, W. Fisher, and J. Lemmink, Modelling and Evaluating Service Quality Measurement Using Neural Networks, International journal of operations and production management, 22, 10, 1162-1185, 2002.