**S.P.MANDALI’S**

**RAMNARAIN RUIA AUTONOMUS COLLEGE**

**MATUNGA, MUMBAI-400019**

**PROJECT REPORT**

**ON**

**“ANALYSIS OF USED CARS IN INDIA”**

**UNDER THE GUIDANCE OF**

**“PROF. MR. ABHIJEET GOLE.”**

**PROJECT BY**

**“Mr. PRATIK ANIL BHOSALE.”**

**MSC.CS. [PART II]**

**DEPARTMENT OF**

**COMPUTER SCIENCE AND INFORMATION TECHNOLOGY**

**2019-2020**

**ANALYSIS OF USED CARS IN INDIA & FIND THE PERCENTAGE LOSS FOR SELLER AND PERCENTAGE PROFIT FOR PURCHASER.**

**ACKNOWLEDGEMENT**

I would like to express my thanks to the people who have helped me most throughout my project.

I would like to thank my project guide “**Prof. Abhijeet Gole**” to have permitted me to go ahead with this project and appreciating my work at every stage.

Special thanks of mine goes to my colleagues who helped me out in testing/debugging the project and with whom I exchanged lot of interesting ideas, thoughts and which made it possible to complete my project.

I am highly indebted to Ruia College for their guidance and constant supervision as well as for keeping me track regarding the project & also for their support in completing the project.

At last but not least I wish to avail myself of this opportunity, express a sense of gratitude and love to my friends and my beloved parents for their manual support, strength and help for everything.

**Pratik Anil Bhosale .**

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Topic** | **Page.**  **No** |
| **1** | **TITLE AND ABSTRACT** | **5** |
| **2** | **BACKGROUND AND LITERATURES REVIEW** | **6** |
| **3** | **PROJECT INTRODUCTION AND SCOPE** | **7** |
| **4** | **RESULTS AND INTERPRETATIONS** | **10** |
| **5** | **CONCLUSIONS** | **34** |
| **6** | **FUTURE SCOPE AND DEVELOPMENT** | **36** |
| **7** | **REFERENCES AND APPENDIX** | **37** |
| **8** | **INDEX AND ACRONYMS** | **NA** |

* **TITLE AND ABSTRACT :-**

**ANALYSIS OF USED CARS**

India, the competition between used cars and new cars has reached its peak. All these used cars are generally termed as “**Certified Used Cars**”, as they are provided warranty and certified quality from the car dealers. Different car manufacturers have entered the pre-owned car business with different objectives, and they have different marketing strategies and priorities in entering into the pre-owned car business.

The objective of this study is to find the company wise and model wise total loss for seller (in percent) after selling the car and profit for the customer who brought the car (in percent). The objective is to analyze the current data available to the business and generate insights for future.

The entire analysis was conducted with the help of dataset, weka tool and R studio. Dataset file in CSV format. The analysis is done with the data of 10 years (2010-1019) comprising of 5238 Cars consisting of total 22 brands like BMW,AUDI, MARUTI SUZUKI, HONDA etc. and total 188 models.

In this weanalyze the data company wise and model wise and calculate the total percent of loss for seller and profit percent for purchaser. In this we compare the company and models basis on km driven, mileage, engine, power, fuel types and base on price. And analysis also based on location and year wise.

* **BACKGROUND AND LITERATURES REVIEW :-**

Even as new car sales have slowed down in the recent past, the used cars market has continued to grow over the past year and is larger than the new car market now.

The growth rate of new car sales has slowed owing to a variety of reasons, including cyclical slowdown in auto sale in election years and an overall consumption slowdown in the economy. New car sales grew 2.70% in 2018-19, the slowest growth rate for the industry in four years. In April, passenger vehicle sales saw a sharp decline compared with the same month last year, and domestic sales saw a contraction of 17.07%.

According to the report, 45% of the buyers want a car that is four to five years old. However, 46% of the sellers want to sell their vehicle when it is six to eight years old.

Data analytics can help the used cars system in many ways. Predictive analytics is the process of analyzing the data using automated statistical processes and summarizing results into useful information. The information acquired from the predictive analysis can be very useful to the customers who buy used cars.

* **PROJECT INTRODUCTION AND SCOPE :-**

Used Car's Business is a blooming sector in India, especially with Multiple Car companies having their own Used car showrooms. And its only gonna increase, given the current trend of Indian people moving towards cars!, especially Middle class families, which prefer an Used car instead of a new one.

So there is a strong need to analyze used car data, With the R tool for visualization it is easy to understand and interpret the results of dataset produced by applying various techniques on it. R performs wide variety of functions, such as data manipulation, statistical modeling in graphics.

With the help of predictive modeling technique like Linear Regression, it is easy and significantly efficient to understand the future car prices. Clustering helps to relate observations in the same group to be similar and observations in different groups to be dissimilar.

With Aggregation methods, it is easy to understand the average of company and model wise mileage, engine, km driven, power and cars selling and purchase prices. Using the same we can also understand the year wise average used car sales . Using the method, it is easy to interpret the total cars sales locations in india.

Also using analytical techniques, to get company wise and model wise percentage profit for purchaser and percentage loss for seller. This will also tells the average selling price company wise and model wise.

# SYSTEM DESIGN AND FLOW :-

Dataset: The dataset used csv format.

Dataset origin:- kaggle.com

* **R Studio and R Language :-**

**In this system the Software used is RStudio:-**

As R Studio  is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics

**In this System the Technology used is R :-**

R is an open-source language and environment for statistical computing and data visualization, supporting data manipulation and transformations, as well as sophisticated graphical displays.

* **Packages Used:-**
* **ggplot2:-**

ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts.

* **Plotly:-**

Plotly provides online graphing, analytics, and statistics tools for individuals and collaboration.

* **Dplyr :-**

Dplyr is the next iteration of plyr, focused on tools for working with data frames (hence the d in the name).

* **Shiny :-**

Shiny is a new package from R Studio that makes it incredibly easy to build interactive web applications with R.

* **Shiny dashboard :-**

Create dashboards with Shiny. This package provides a theme on top of Shiny, making it easy to create attractive dashboards.

* **RColorBrewer :-**

Provides color schemes for maps (and other graphics) designed by Cynthia Brewer as described at [http://colorbrewer2.org](http://colorbrewer2.org/).

* **Wordcloud :-**

Plot a cloud of words shared across documents.

* **Tm :-**

A framework for text mining applications within R.

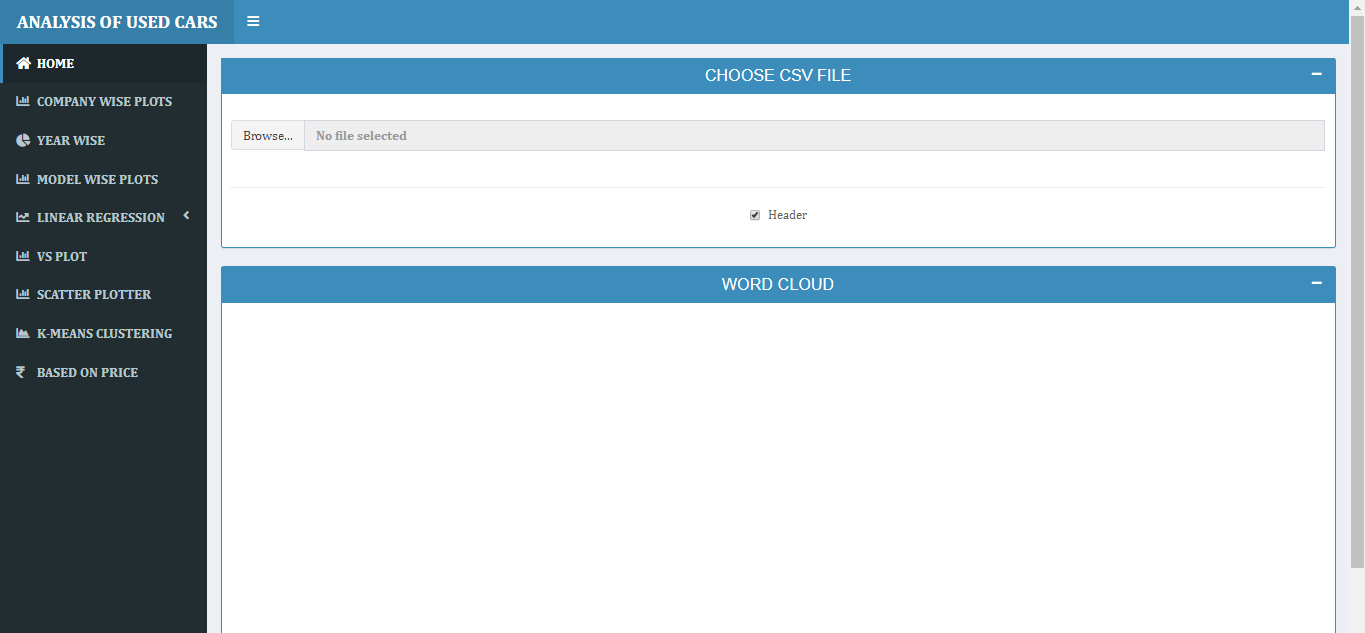
* **Gapminder :-**

The main object in this package is the gapminder data frame or “tibble”. There are other goodies, such as the data in tab delimited form, a larger unfiltered dataset, premade color schemes for the countries and continents, and ISO 3166-1 country codes.

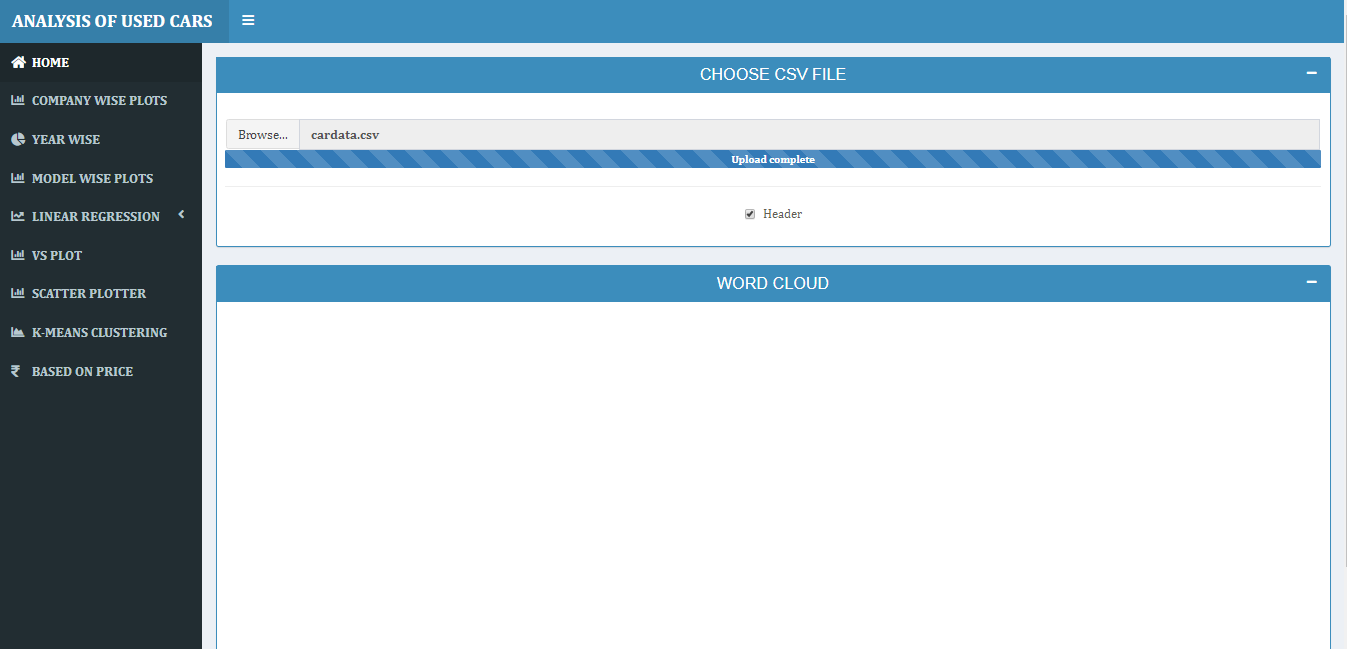
* **Factoextra :-**

[**factoextra**](http://www.sthda.com/english/rpkgs/factoextra) is an R package making easy to extract and  visualize the output of exploratory **multivariate data analyses.**

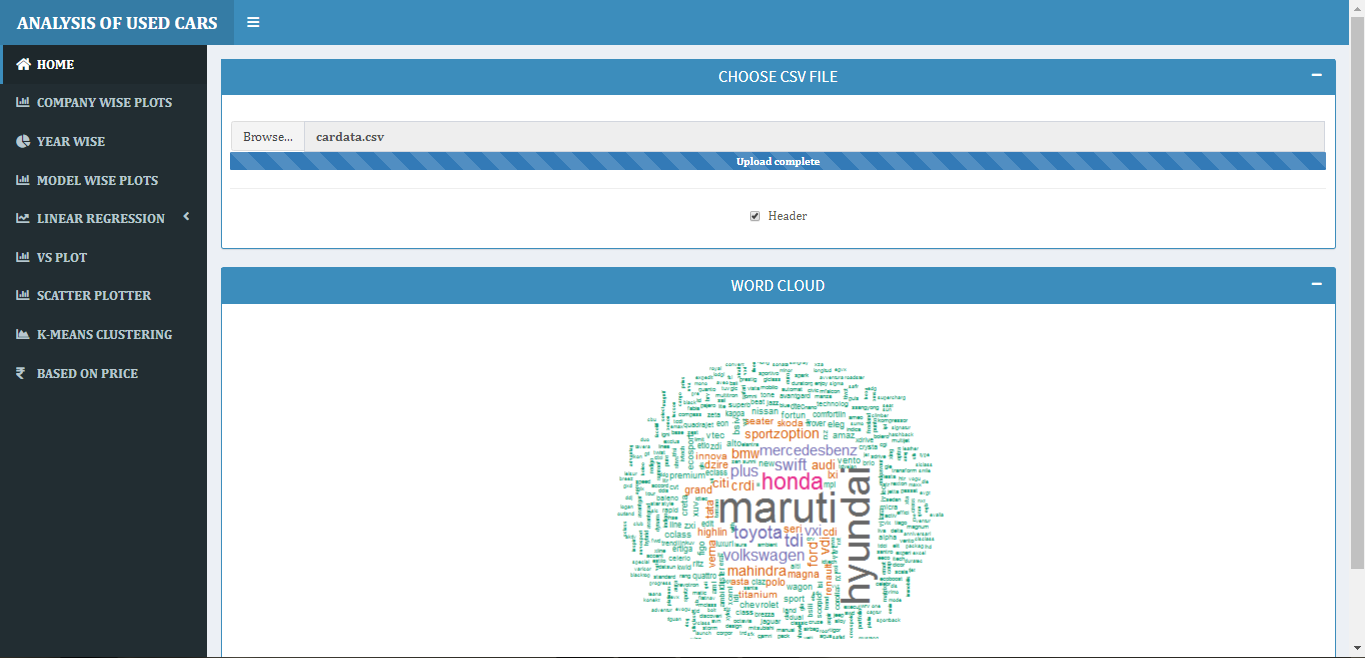
* **RESULTS AND INTERPRETATIONS :-**
  + - **Dashboard.**



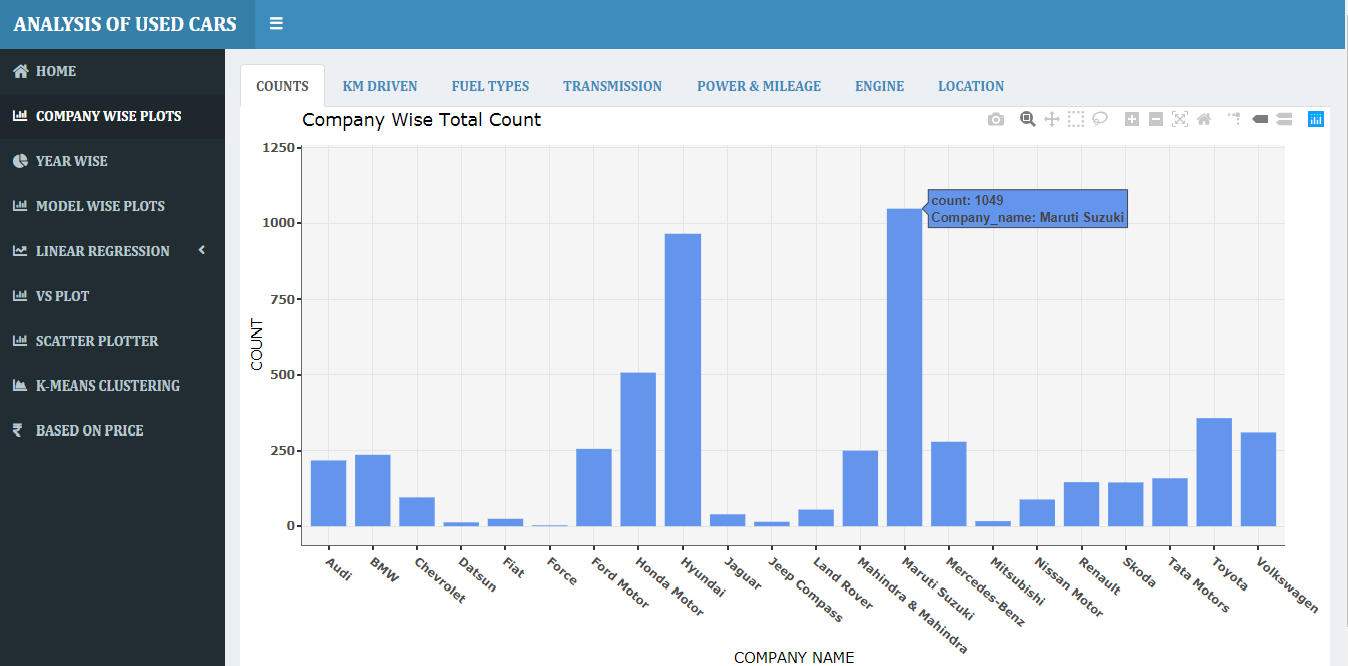
* **Browse File.**



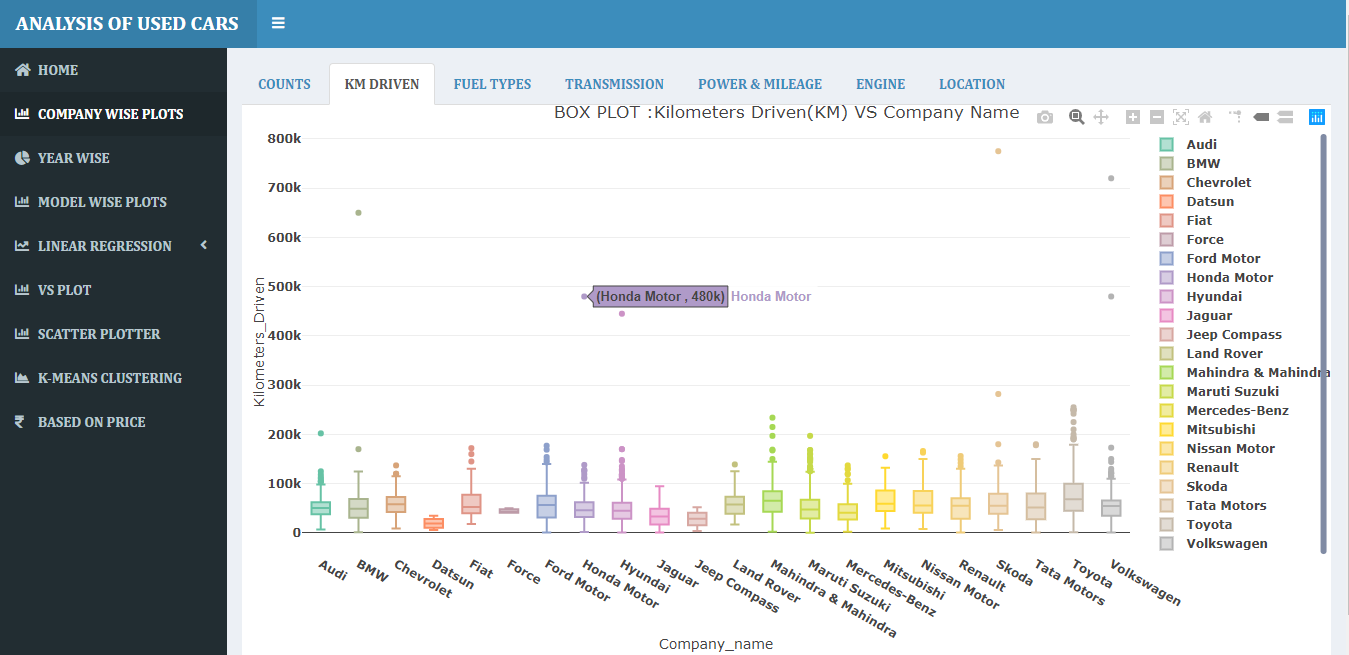
* **After browse file word cloud appears, it interprets the names of the Companies related to the cars.**



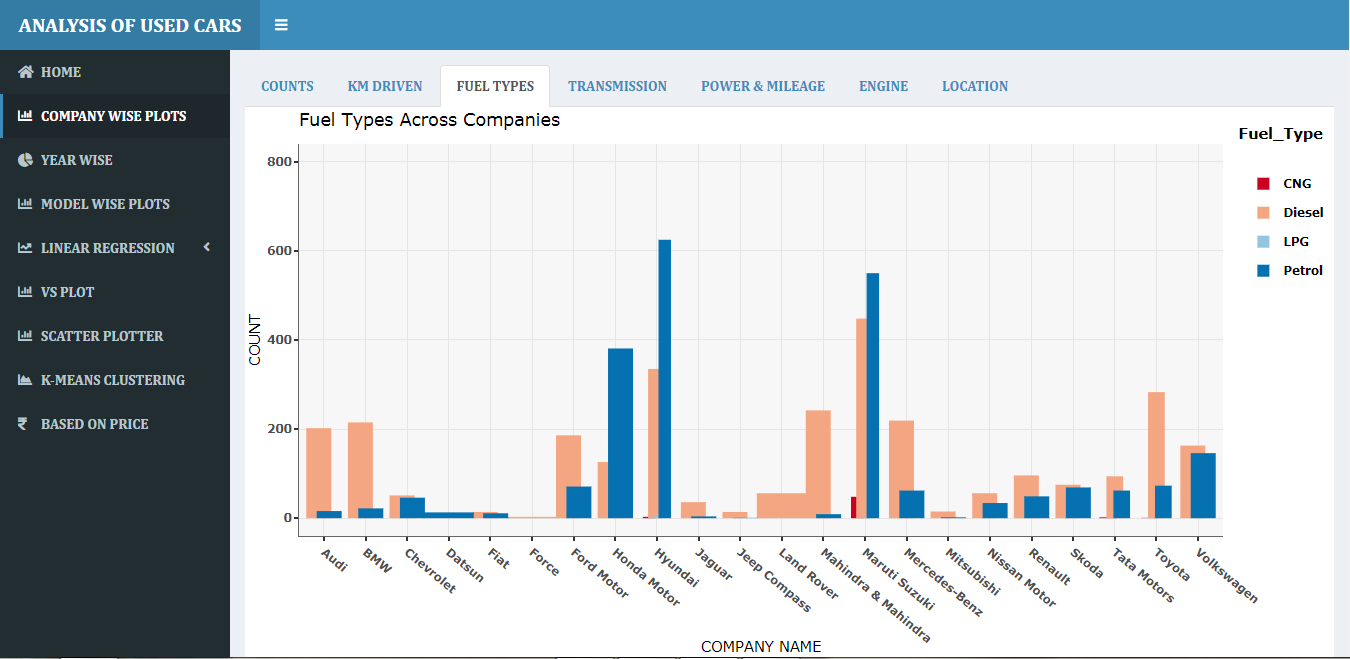
* **Company wise plot :-**
* **Count:-** bar graph with maximum number of cars count of maruti Suzuki and minimum number offorce(3).



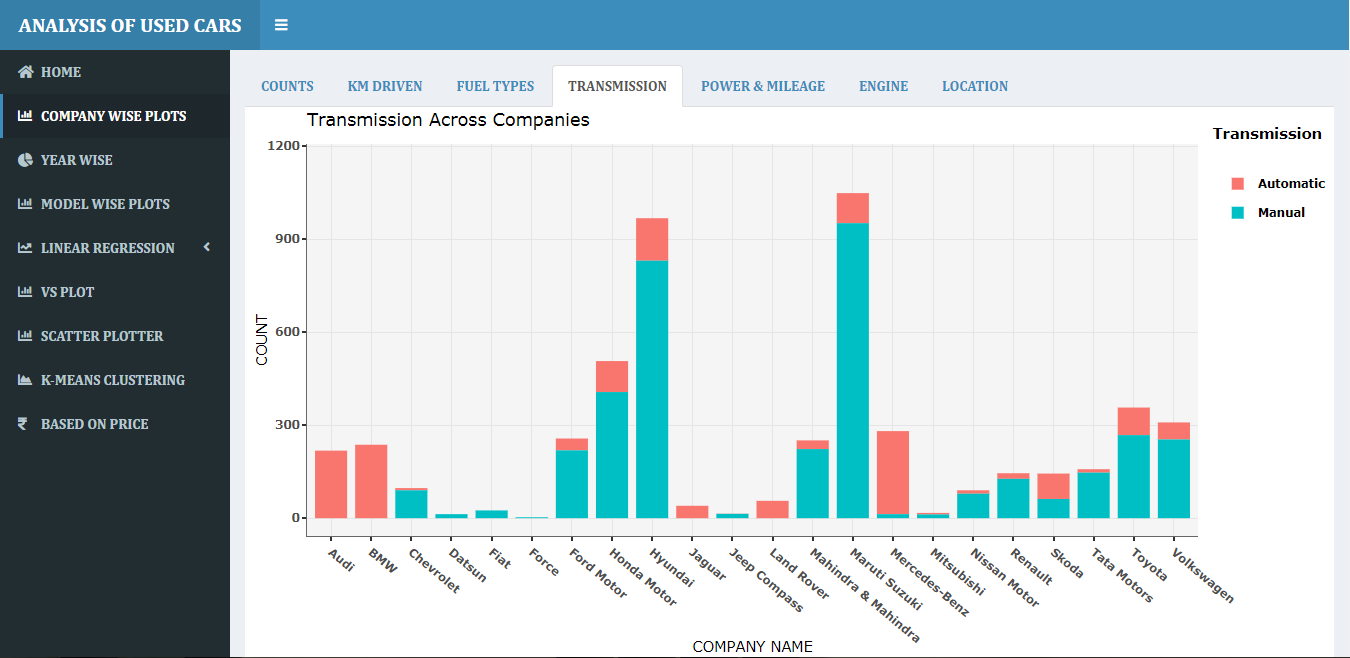
* **Km driven :-** Box plot for km driven.



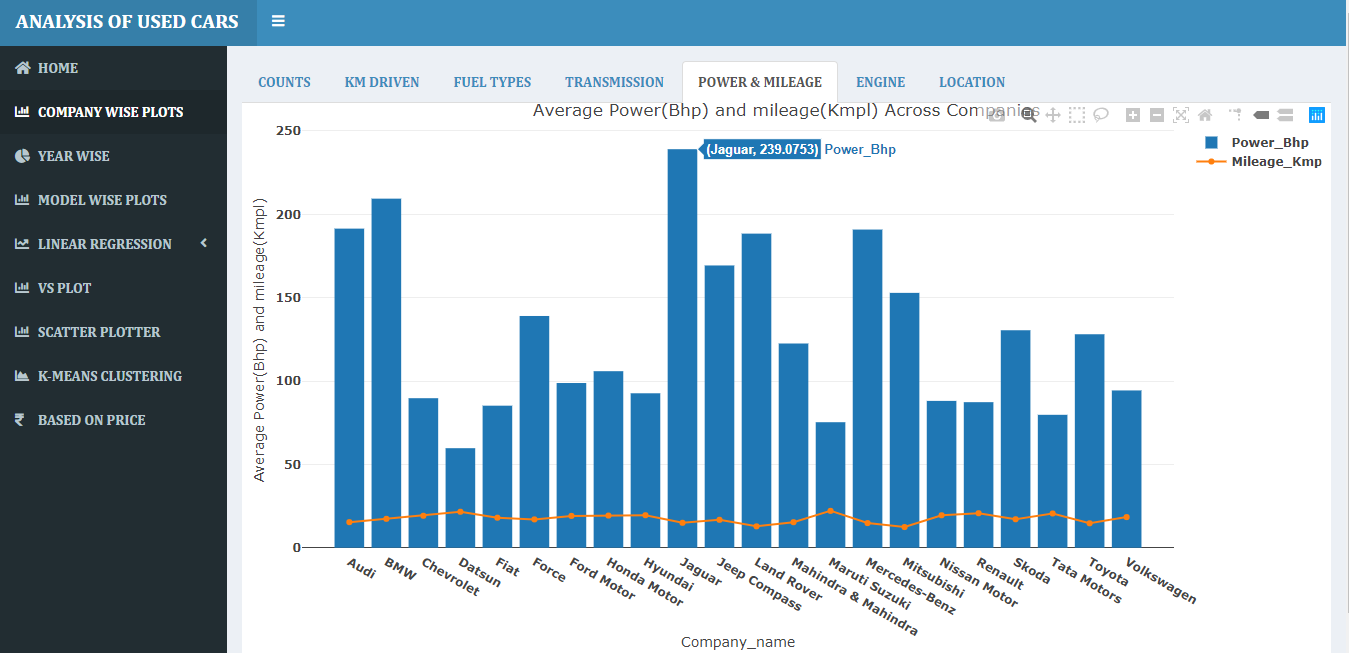
* **Fuel types :-** Histogram of fuel types across companies maximum number of company use petrol and diesel. Company like Maruti, Hyundai , Tata and Toyota use minimum amount of cng and lpg cars.



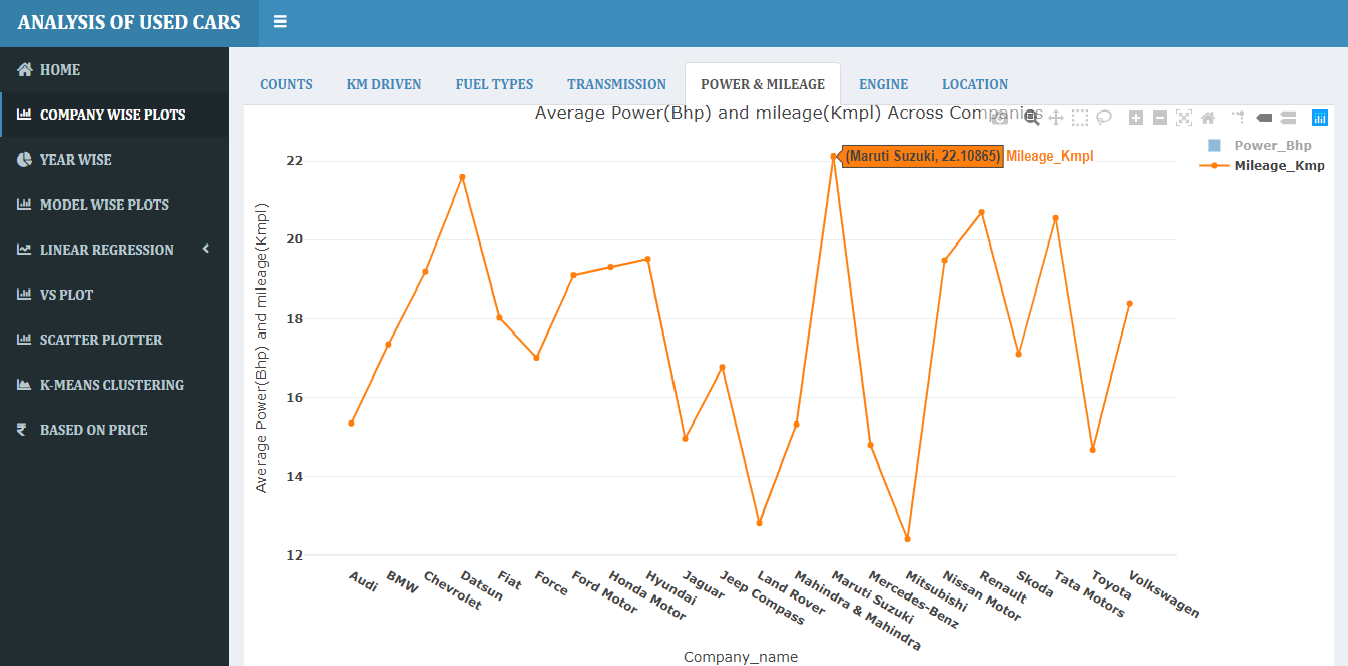
* **Transmission :-** Stacked bar graph ofTransmission across companies(Automatic & Manual). Companies like Bmw, Audi, Jaguar and Land Rover are fully automatic. And other companies use mostly manual transmission.



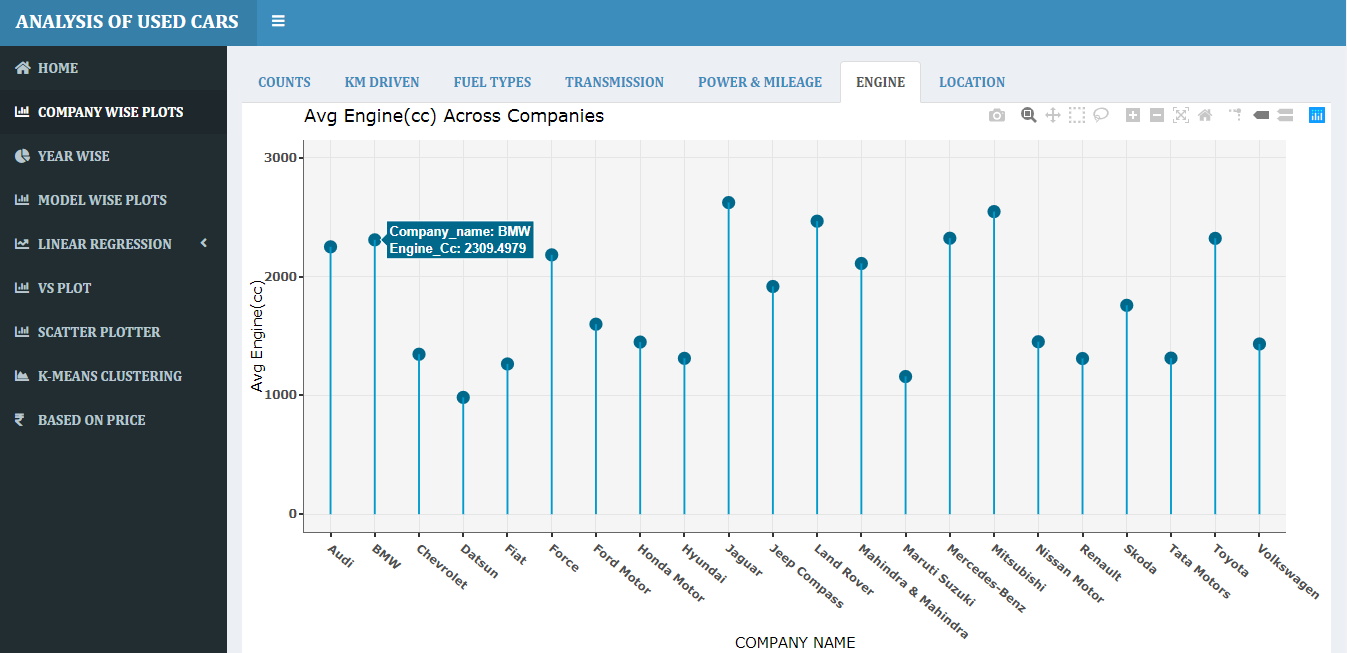
* **Power & Mileage :-** Bar graph for average power and line graph for average mileage across companies. Those companies have more power, that have minimum mileage.

****

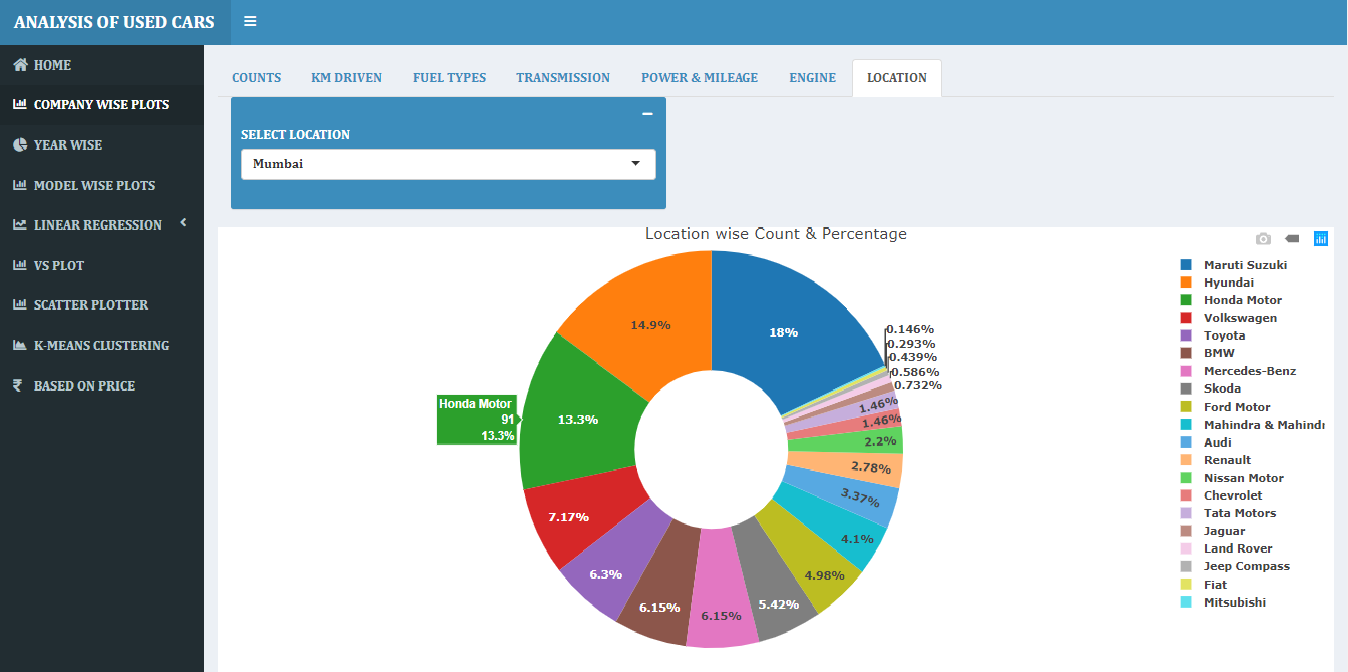
* + - **Mileage :-**Average mileage line graph Company wise. Max mileage gives by Maruti cars whereas minimum mileage gives bye Mitsubishi.

****

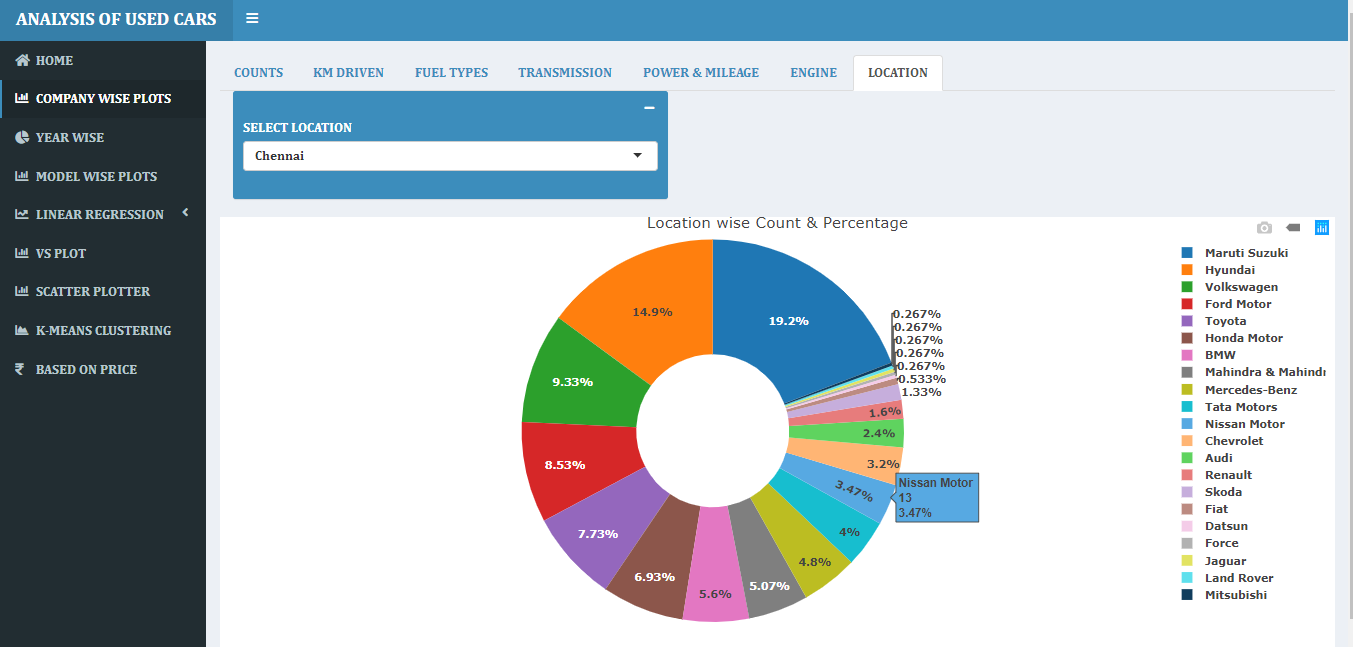
* **Engine :-** Lollipop chart for Average engine across companies. Jaguar have maximum engine cc cars whereas Datsun have less.



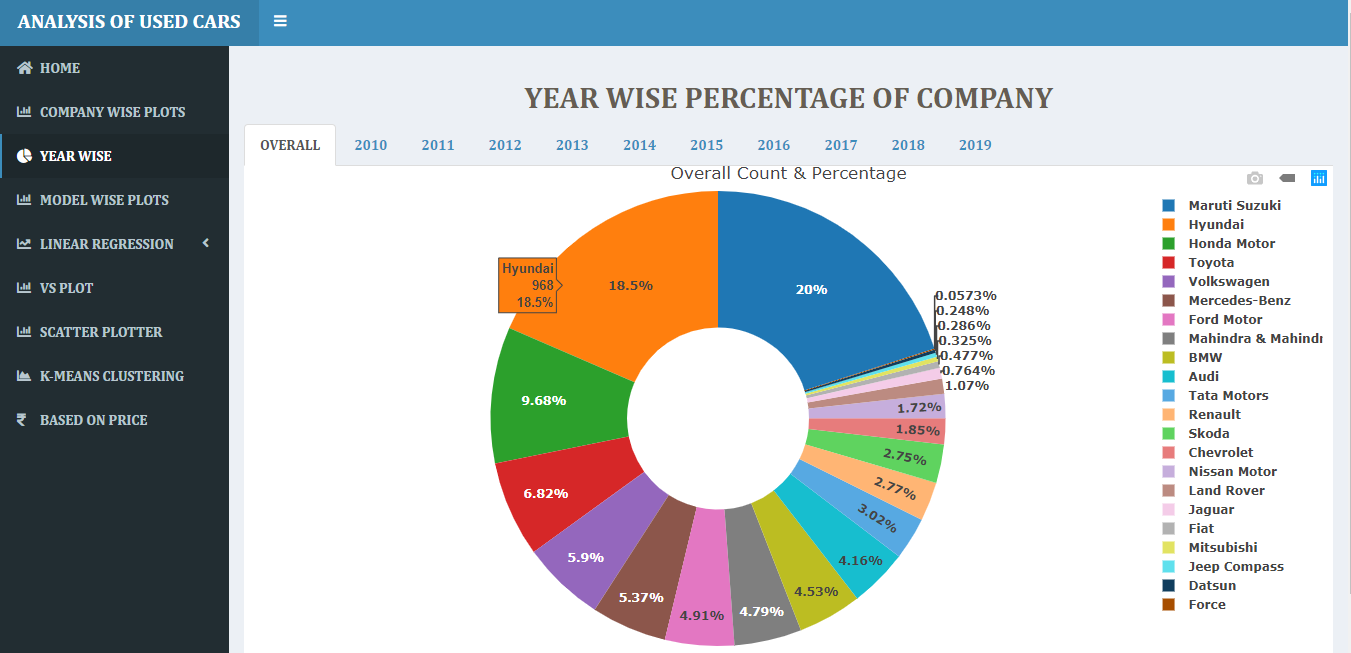
* **Location :-** Select location from drop down list and get the donut chart of location wise count and percentage of companies.
* **Location Mumbai:-**



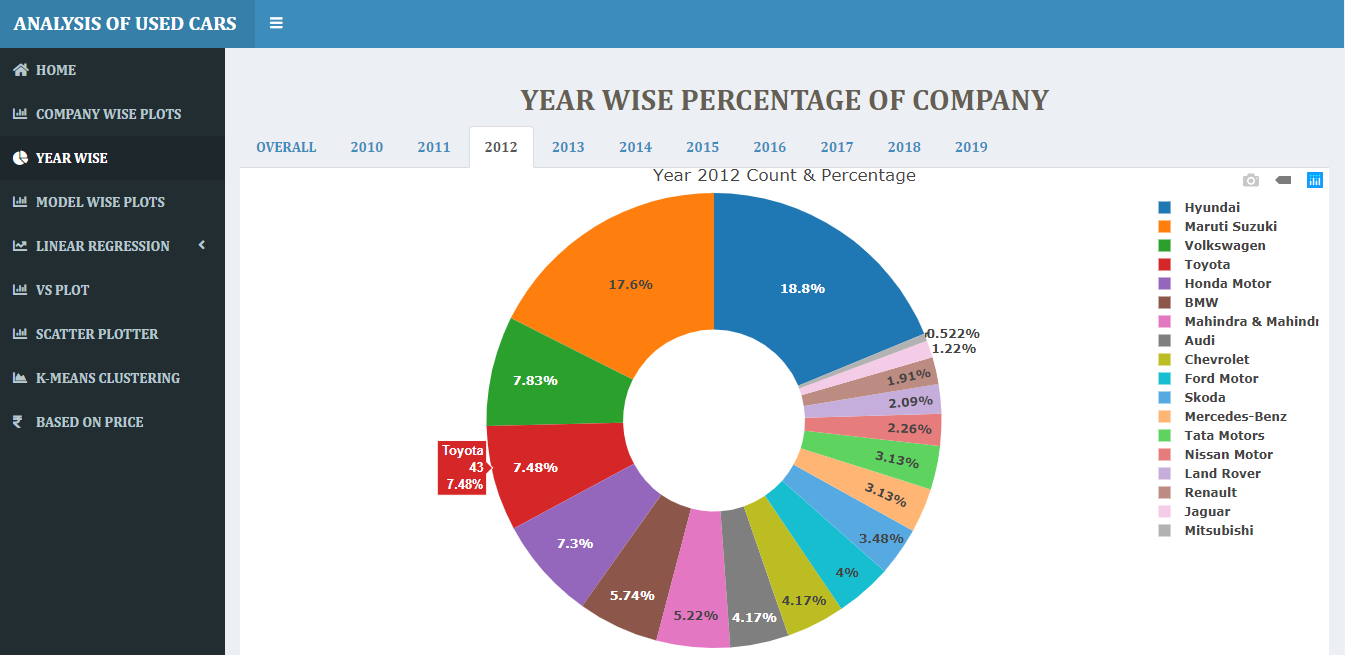
* **Location Chennai:-**



* **Year wise:-** Donut chart of overall and year wise(2010-2019) percentage and count of companies.
* **Overall (From 2010-2019):-**

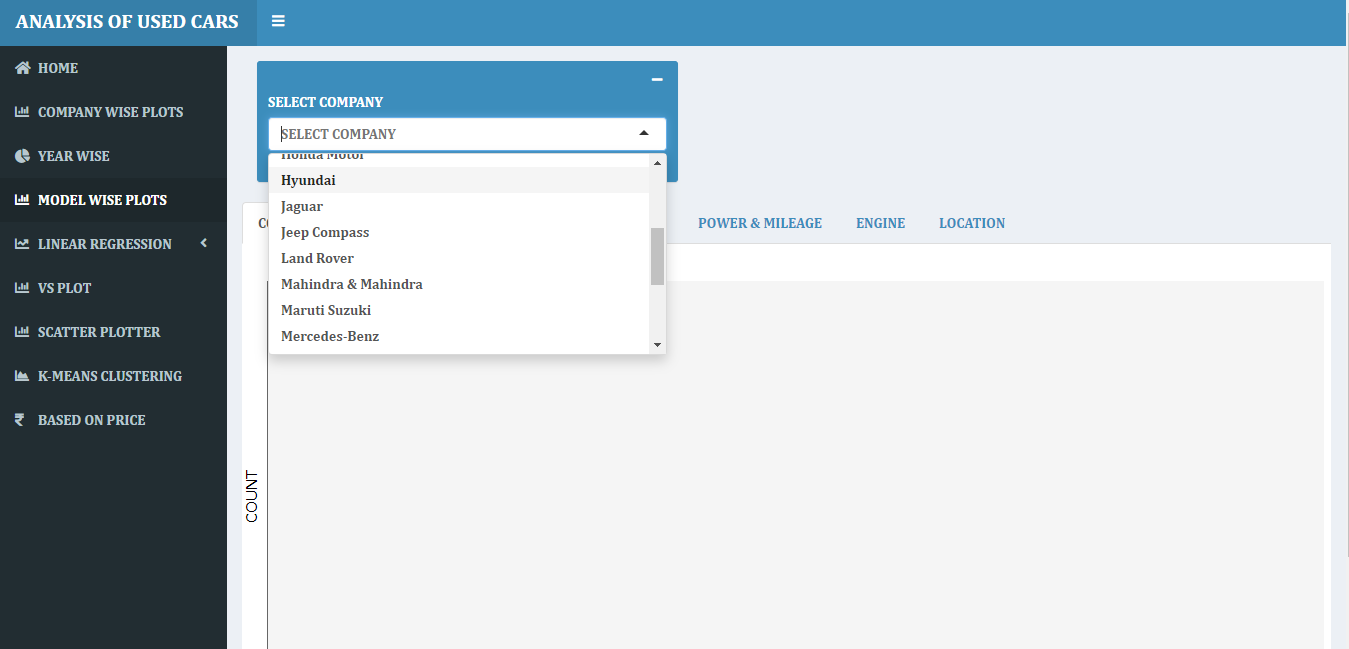


* **Year wise:-**

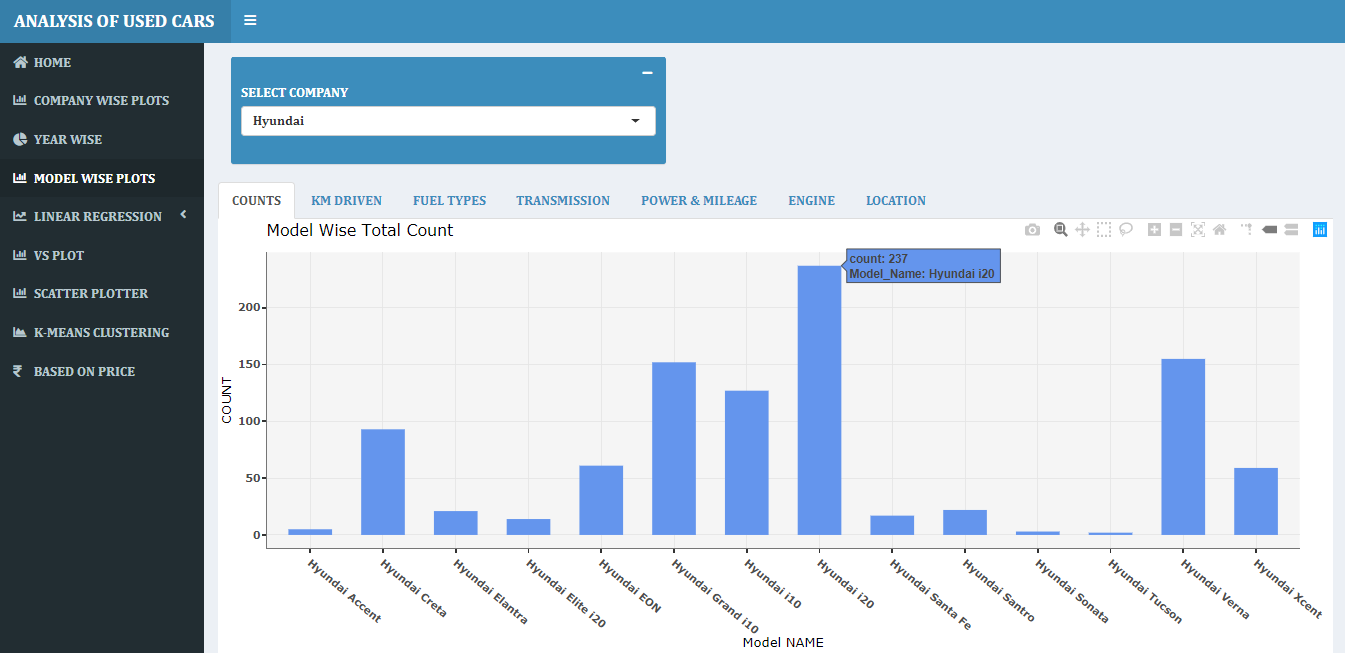
****

* **Model wise plot :-**

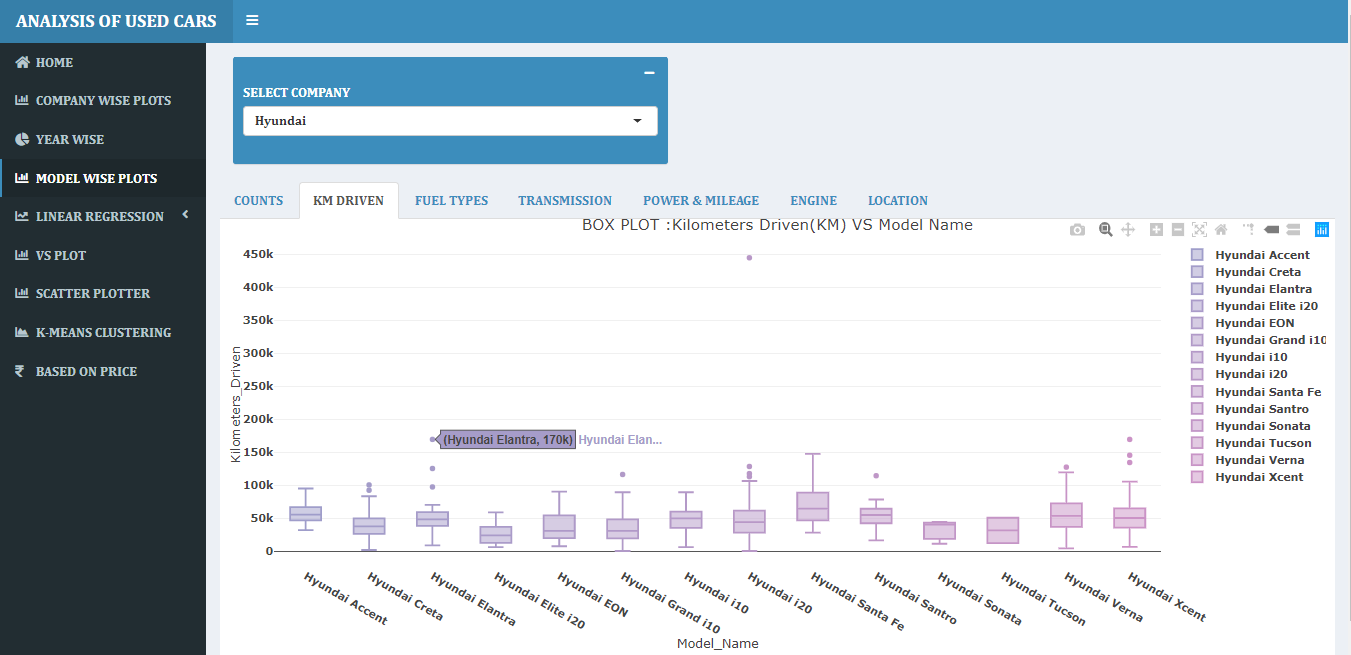
Select company from drop down list, then you get models of selected company.



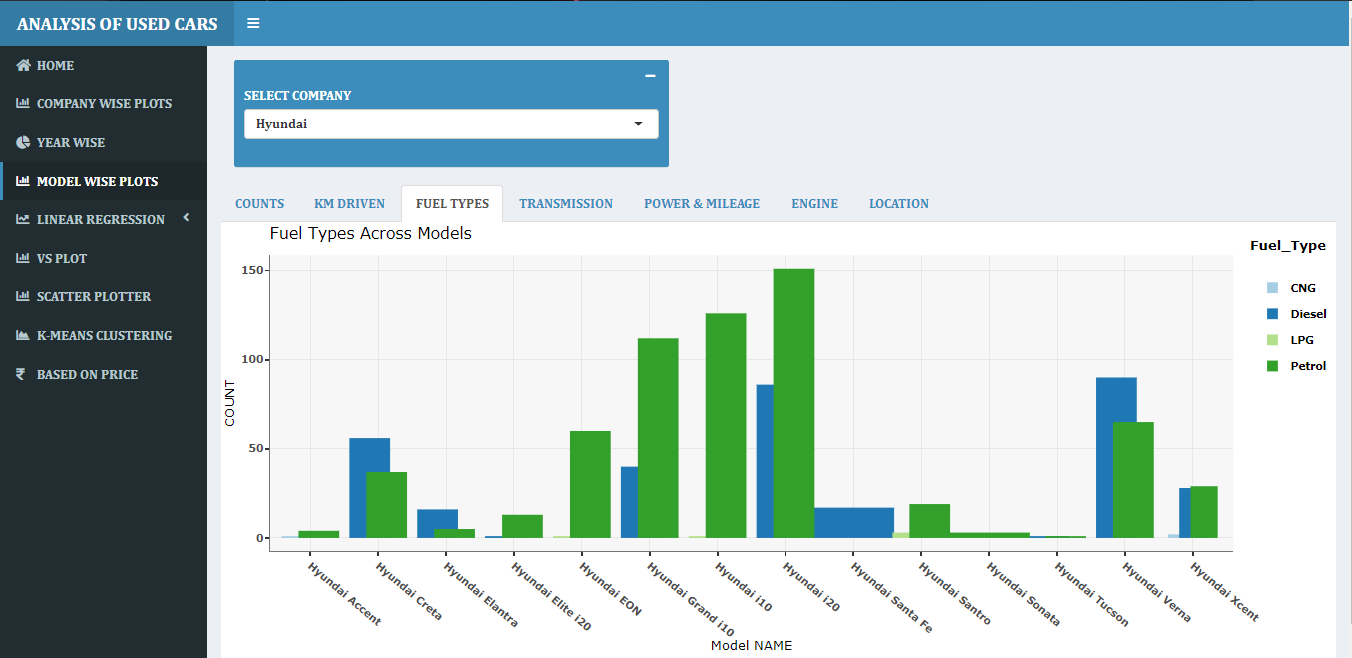
Selected Hyundai company. Then gets total models counts of Hyundai .



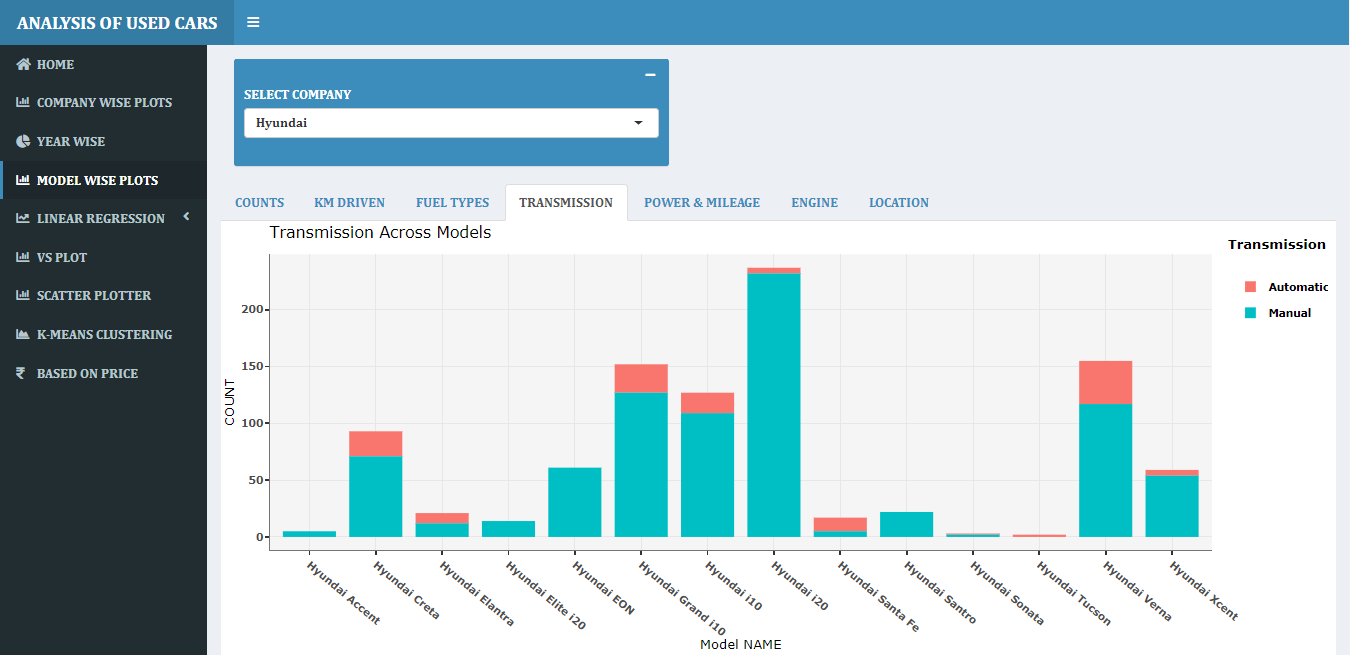
* **Km driven:-** Model wise Box plot of km driven of Hyundai company .



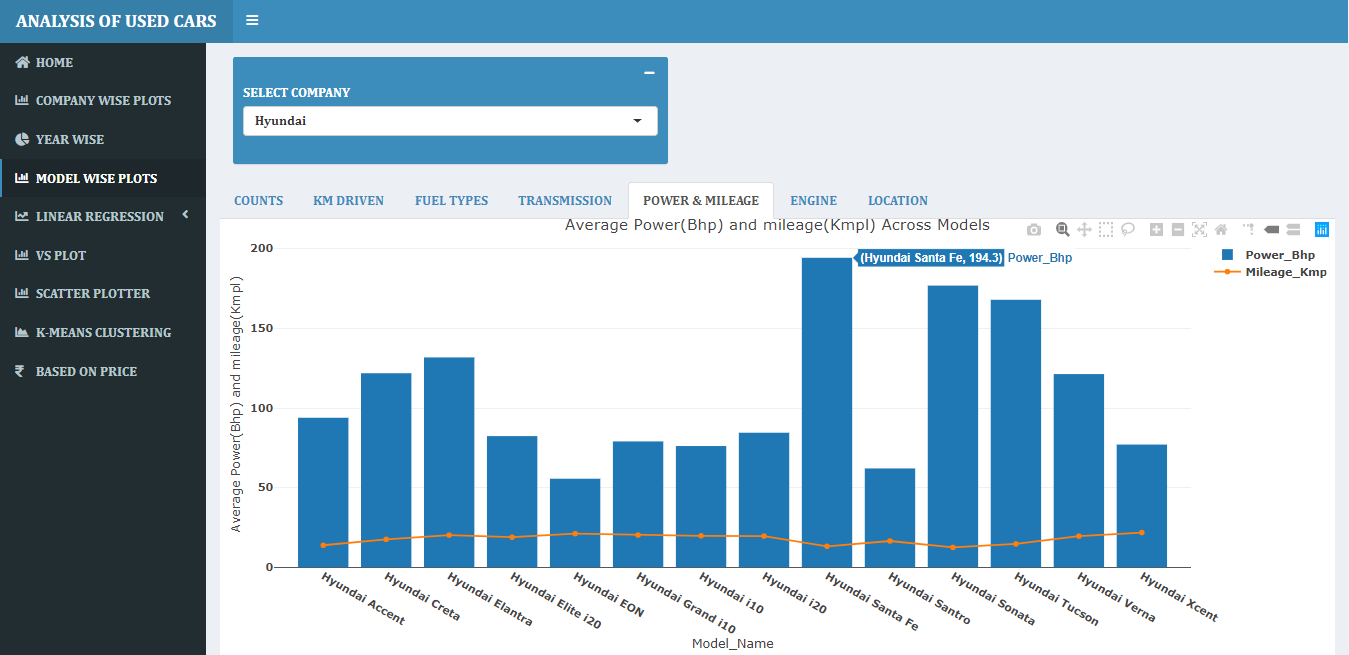
* **Fuel types :-** Histogram of fuel types across models of Hyundai company .(maximum number of Models use petrol and diesel. )



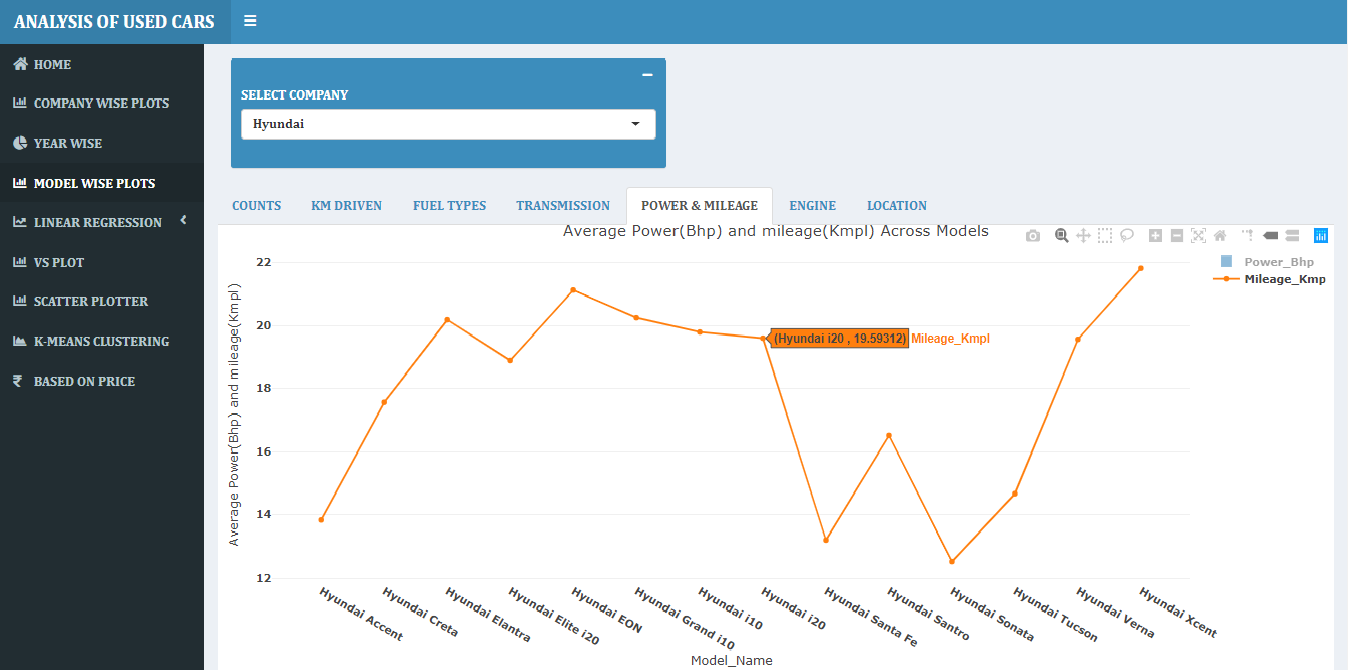
* **Transmission :-** Stacked bar graph ofTransmission across models of Hyundai company(Automatic & Manual).maximum number of Models use Manual Transmission.



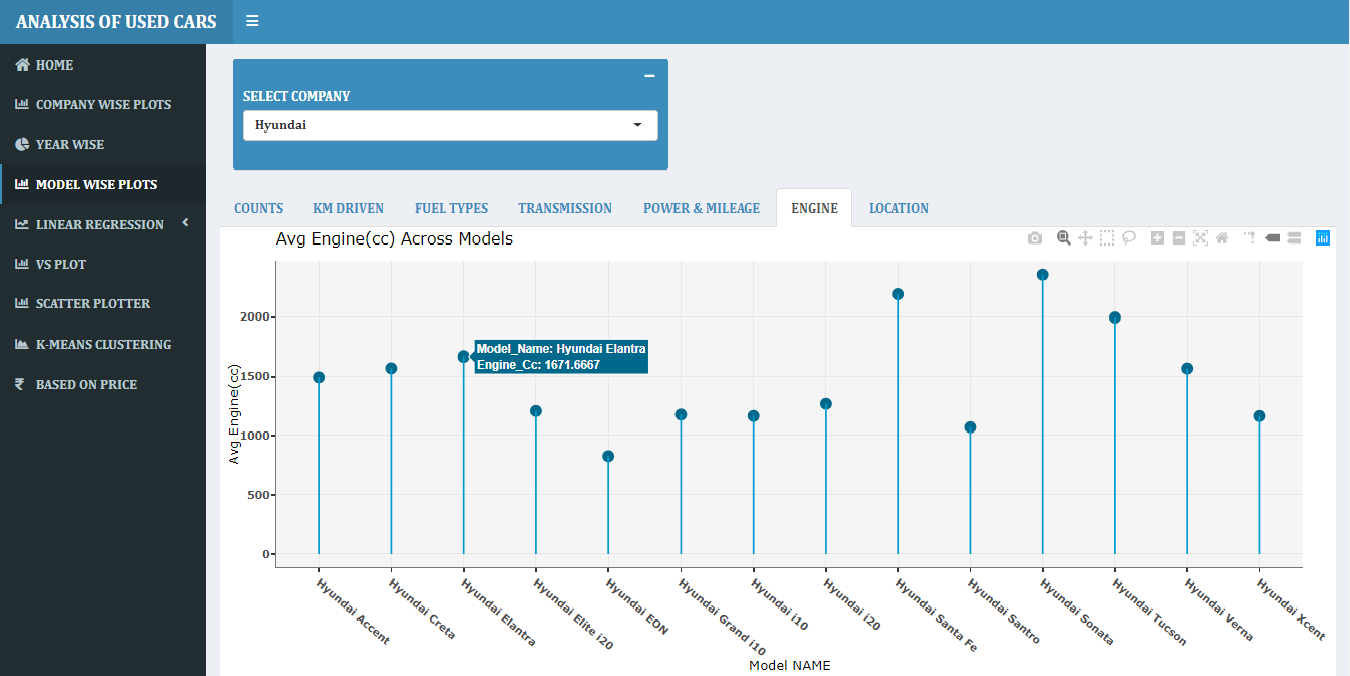
* **Power & Mileage :-** Bar graph for average power and line graph for average mileage across models of Hyundai company.

****

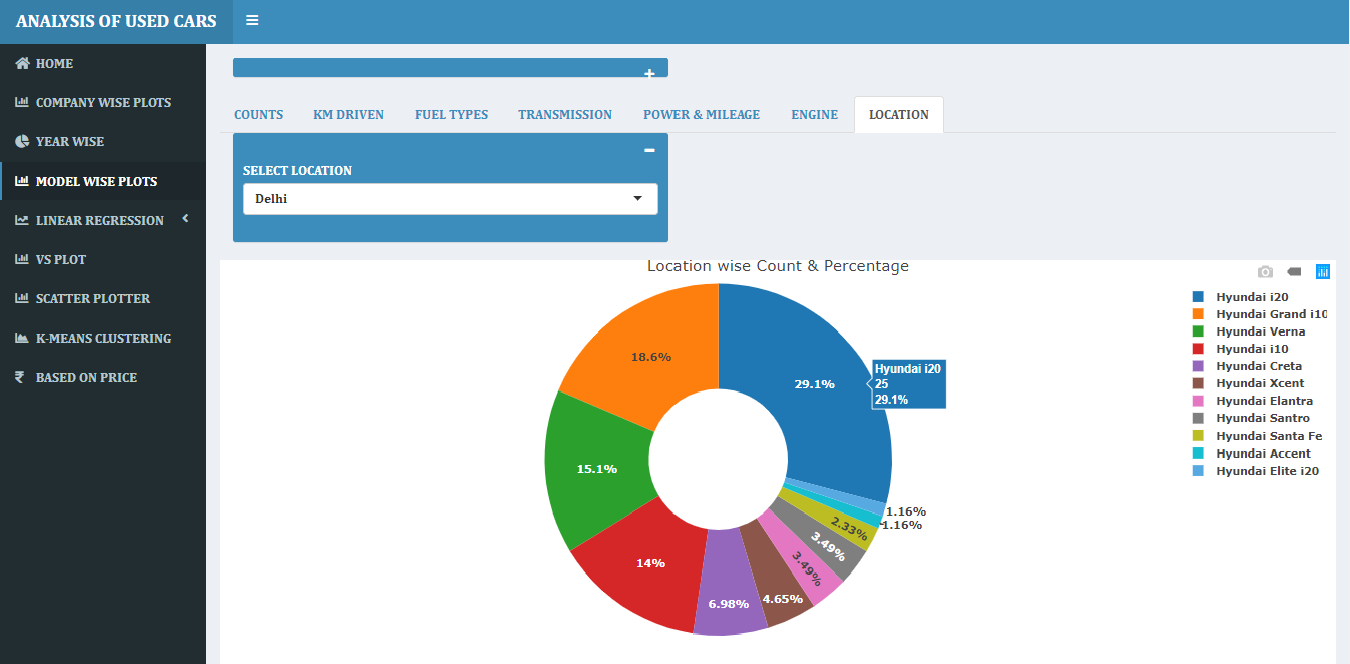
* **Mileage :-** For average mileage line graph. Those models have maximum power that have minimum mileage.

****

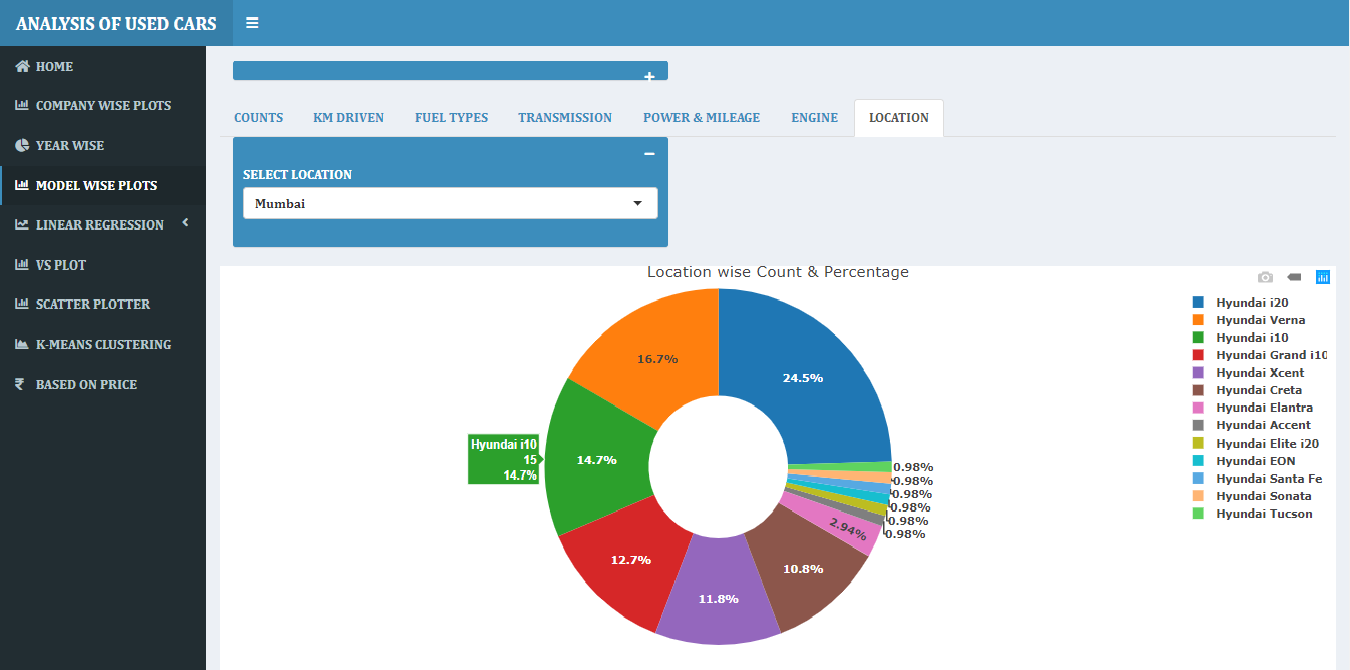
* **Engine :-** Lollipop chart for Average engine across models of Hyundai company.



* **Location :-** Select location from drop down list and get the donut chart of location wise count and percentage models of Hyundai company.
* **Location Delhi:**-

****

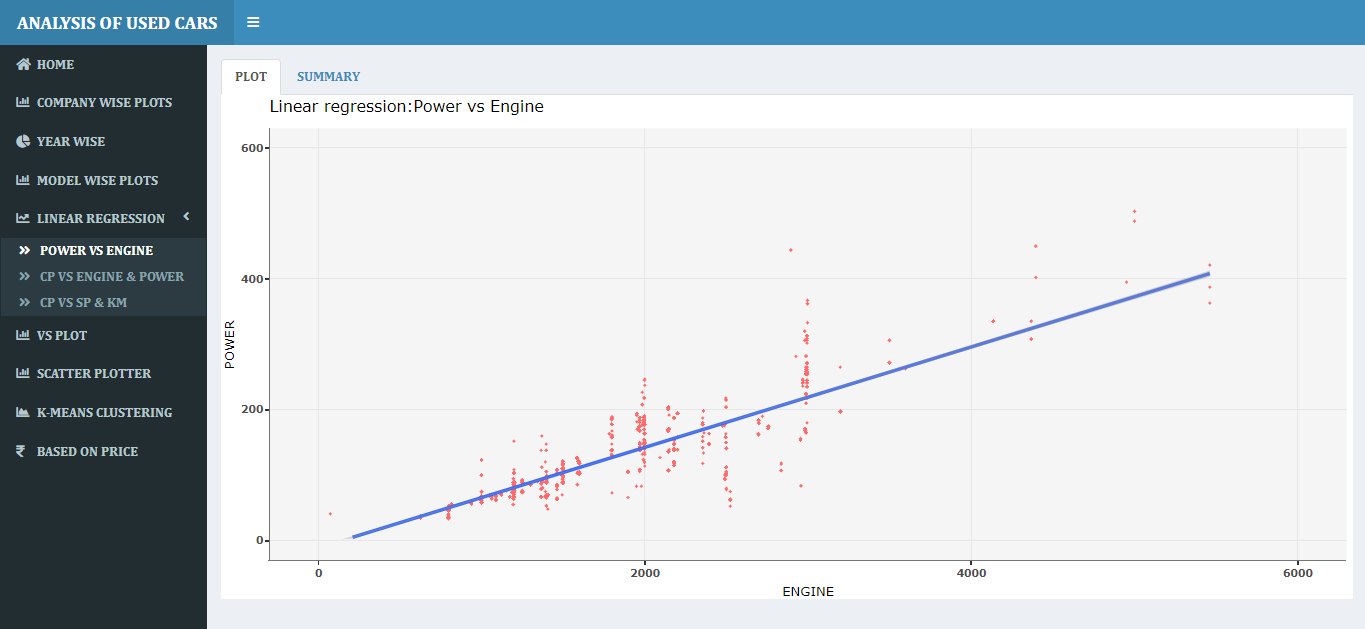
* **Location Mumbai:-**

****

* **Linear Regression:-**
* **Power vs Engine**

Power and engine linearly dependent on each other.

**Plot :-**

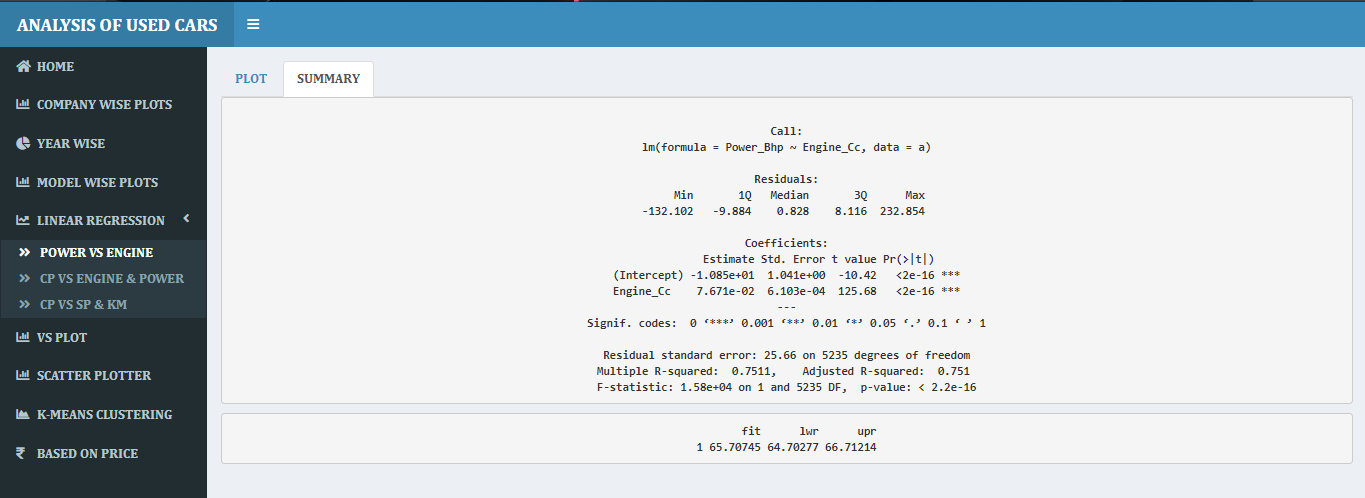


**Summary:-**

The accuracy is 75%, here

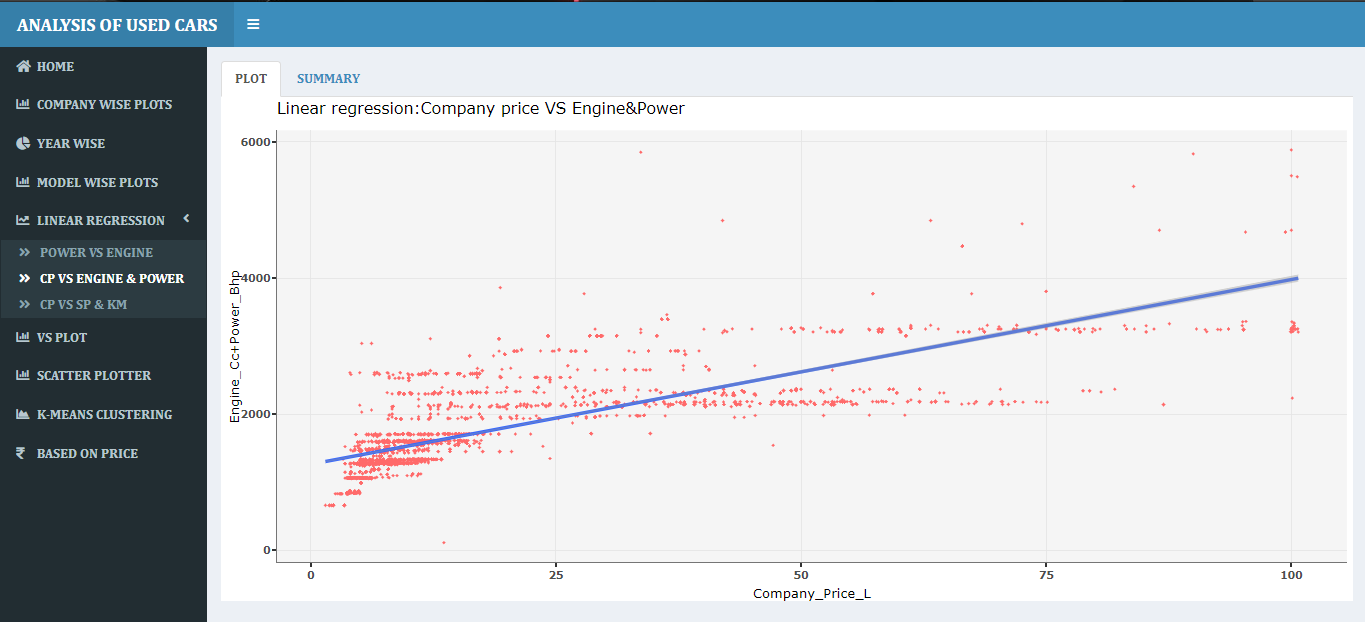
Gives the engine\_cc value and it will predict the value for power.

Engine=998 and predicted value is 65.07 for power(actual=67.04).



* **Company price vs engine & power**
* Company price, engine & power linearly dependent on each other.

**Plot :-**

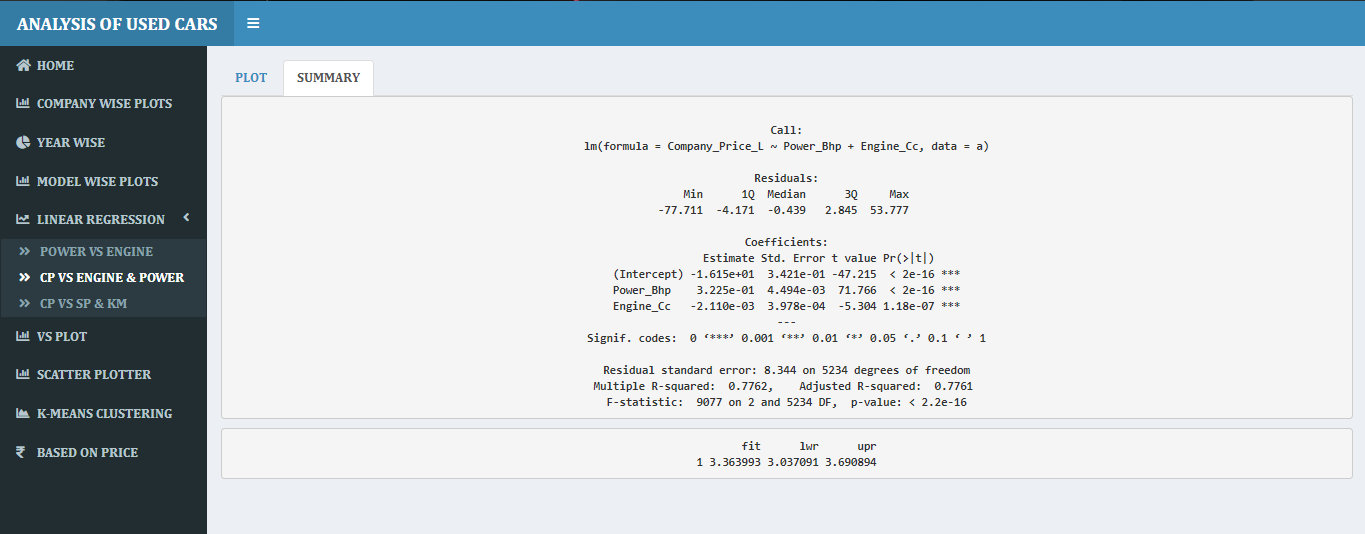


**Summary:-**

The accuracy is 77%, here

Gives the power & engine\_cc value and it will predict the value for Company price.

Power\_Bhp=67.04,Engine\_Cc=998 and predicted value is 3.36L for Company price (actual=4L).



* **Selling price vs Company price & km driven**

Selling price, Company price & km driven linearly dependent on each other.

**Plot :-**

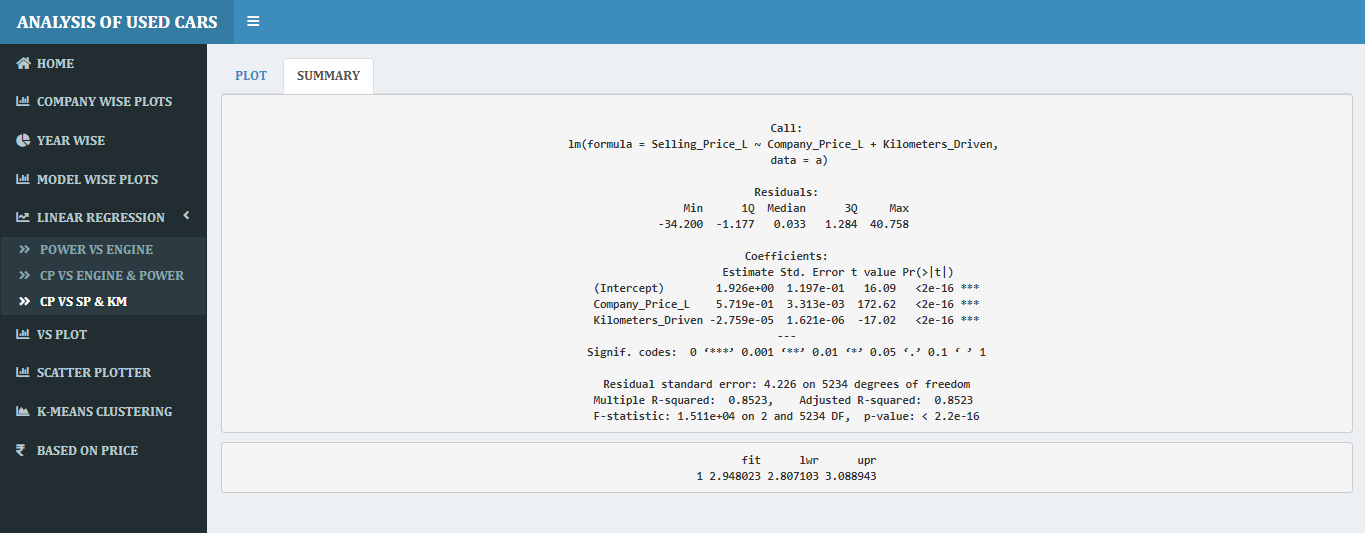


**Summary:-**

The accuracy is 85%, here

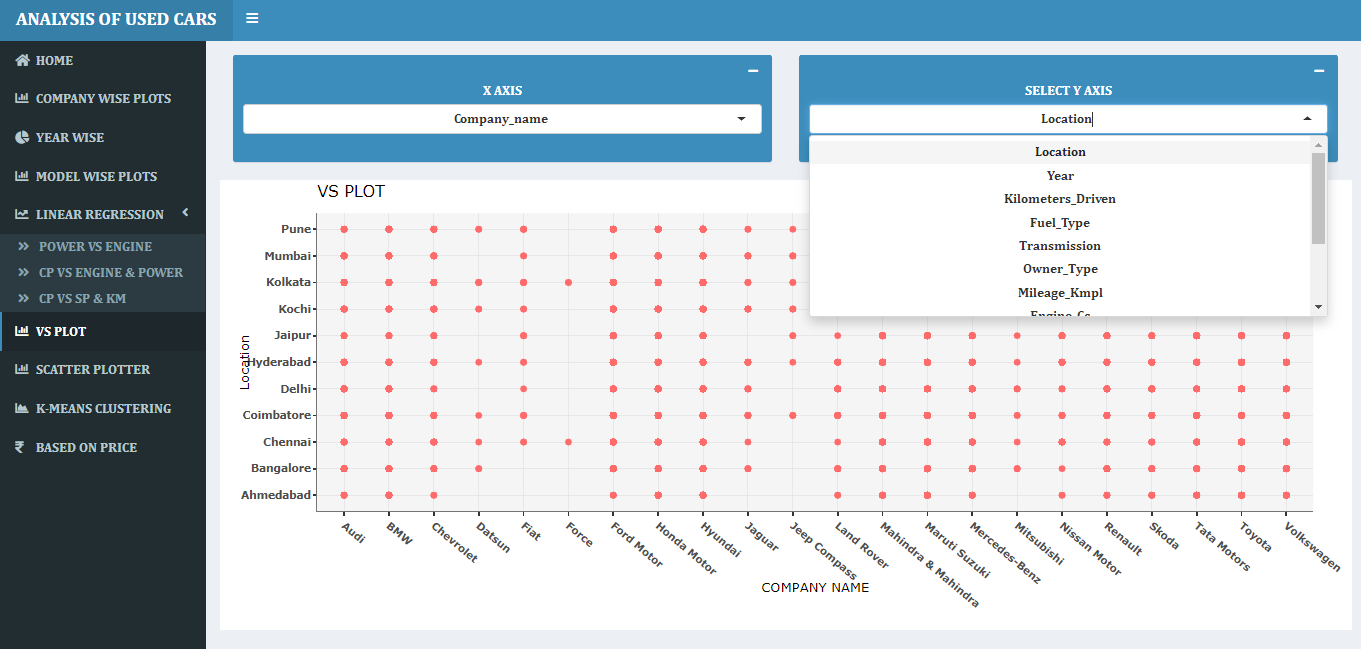
Gives the Company price & km driven value and it will predict the value for Selling price.

Company\_Price\_L=4.2, Kilometers\_Driven=50000 and predicted value is 2.94L for Selling price(actual=2.75L).



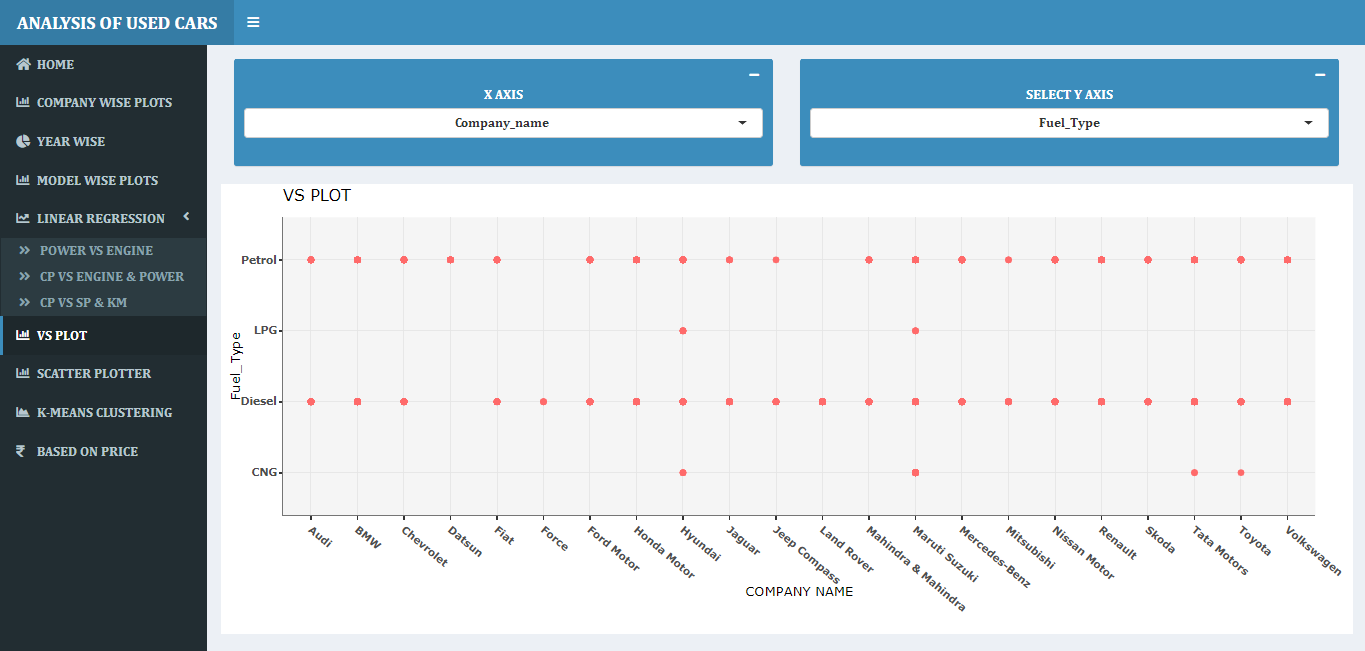
* **VS Plot :-**

x-axis is fixed and choose y-axis from drop down list. (Dot-plot )



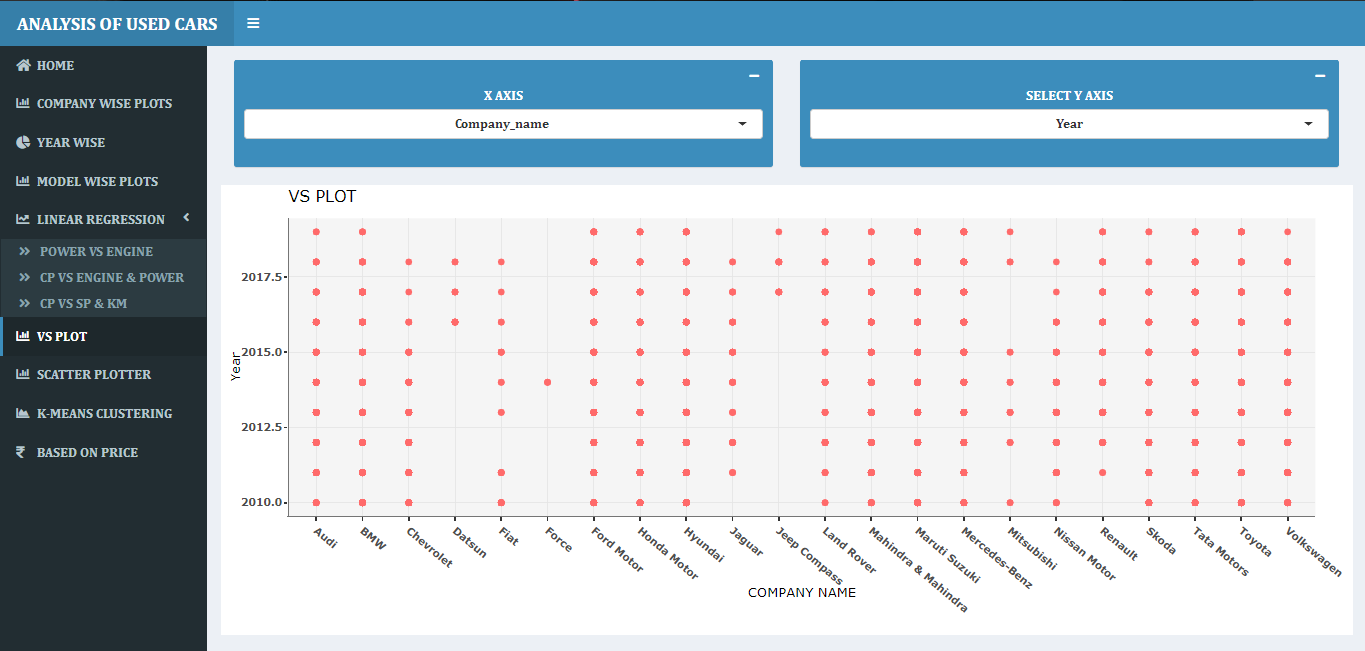
* **Company Name vs Fuel types.**

maximum number of company use petrol and diesel. Company like Maruti, Hyundai , Tata and Toyota use minimum amount of cng and lpg cars



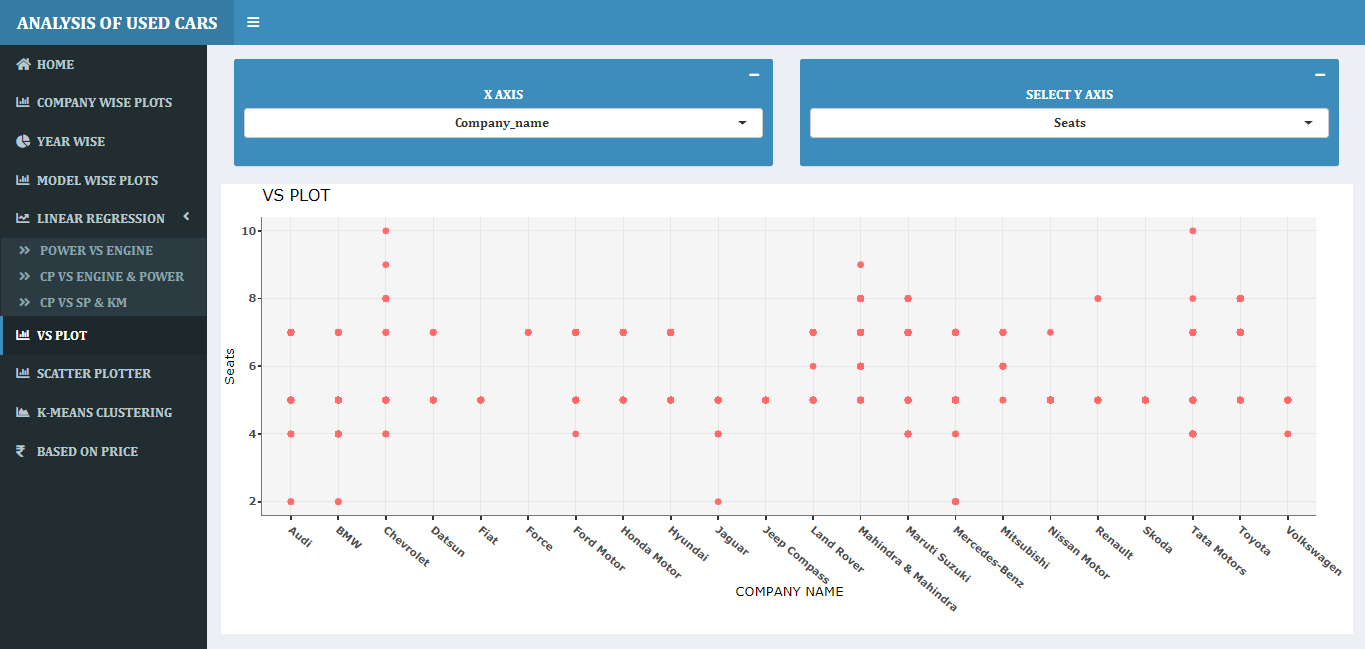
* **Company Name vs Year.**

Force have minimum number of sells.



* **Company Name vs Seats.**

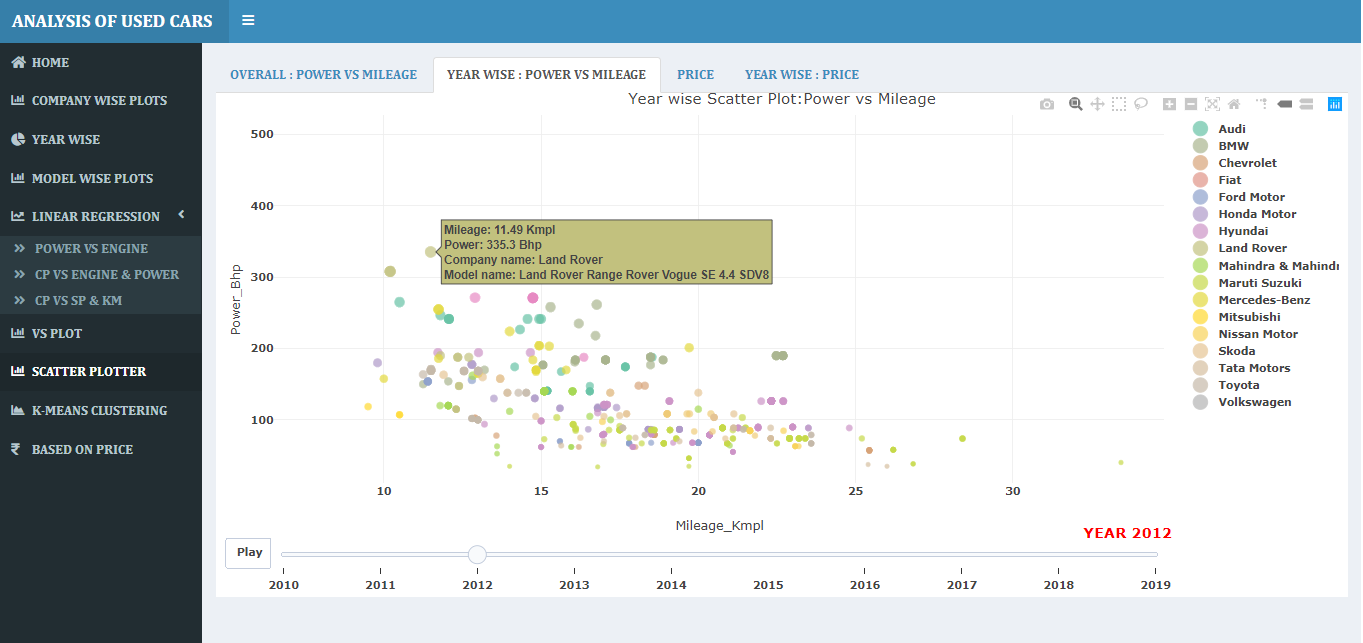
Mostly 4,5 and 7 seater cars Used.



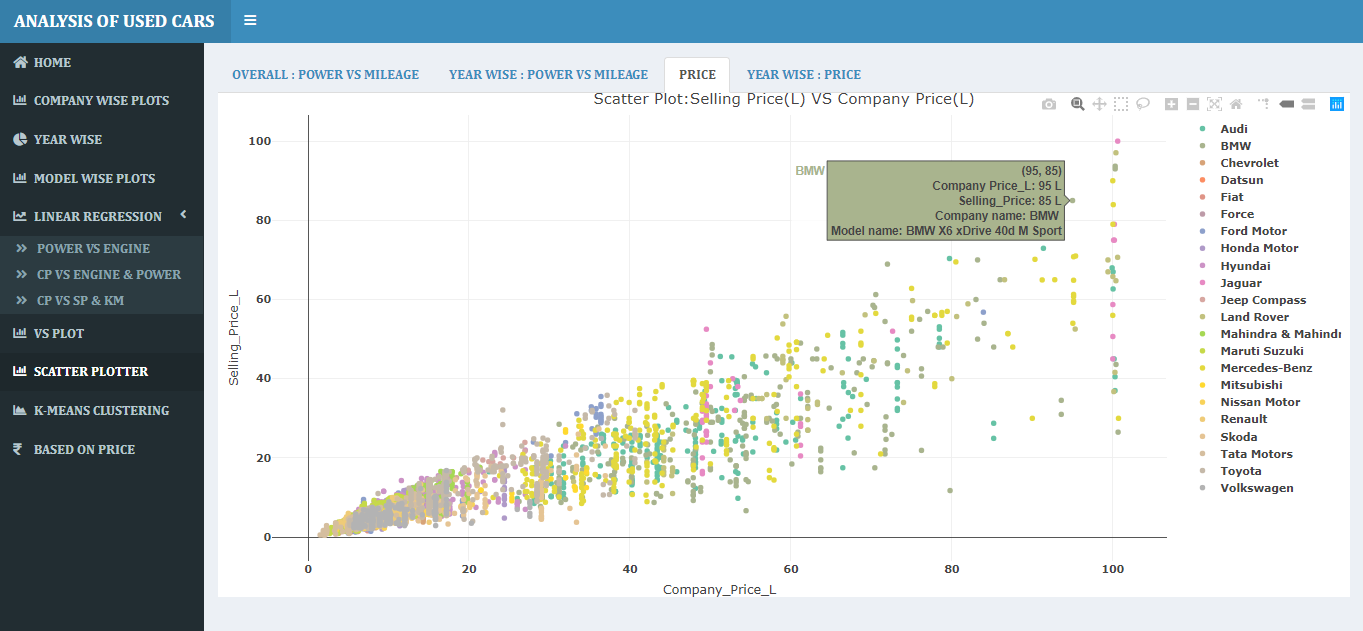
* **Scatter plotter :-**
* **Overall :- Power vs Mileage.(linearly dependent on each other)**



* **Year wise Animated scatter plot for Power vs Mileage.**



* **Scatter plot for Selling price vs company price.**

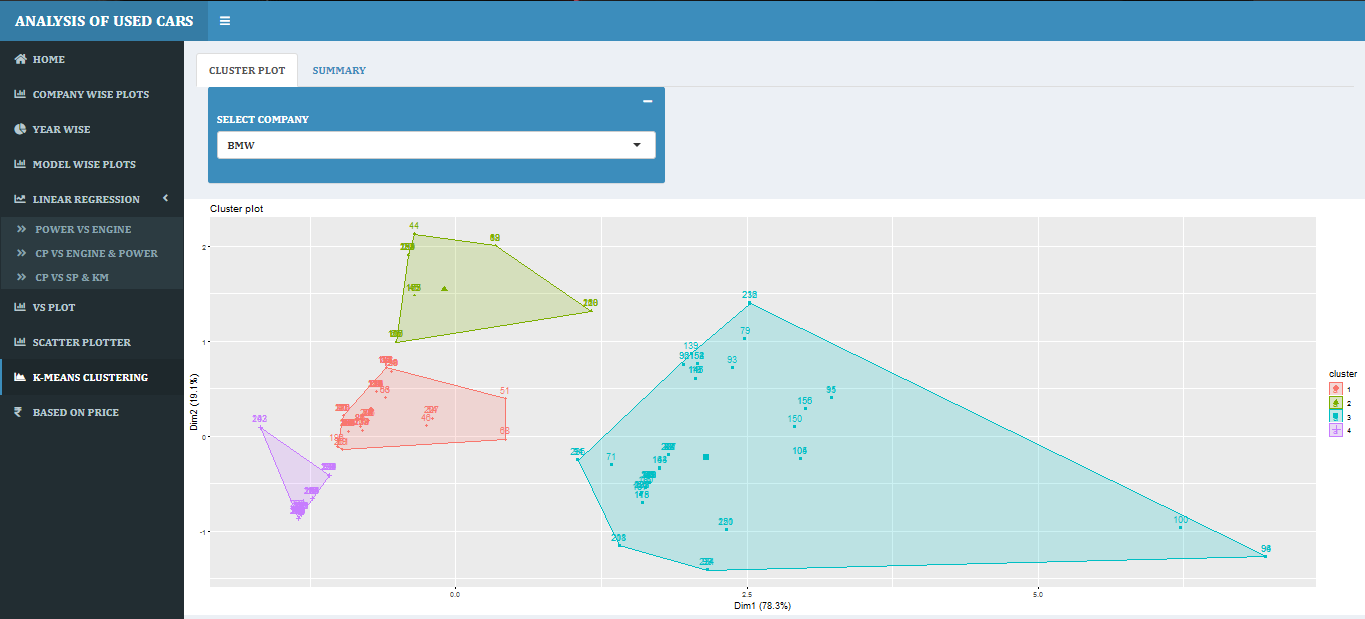


* **Year wise Animated scatter plot for Selling price vs company price. As years increases the cost of vehicles also increases.**



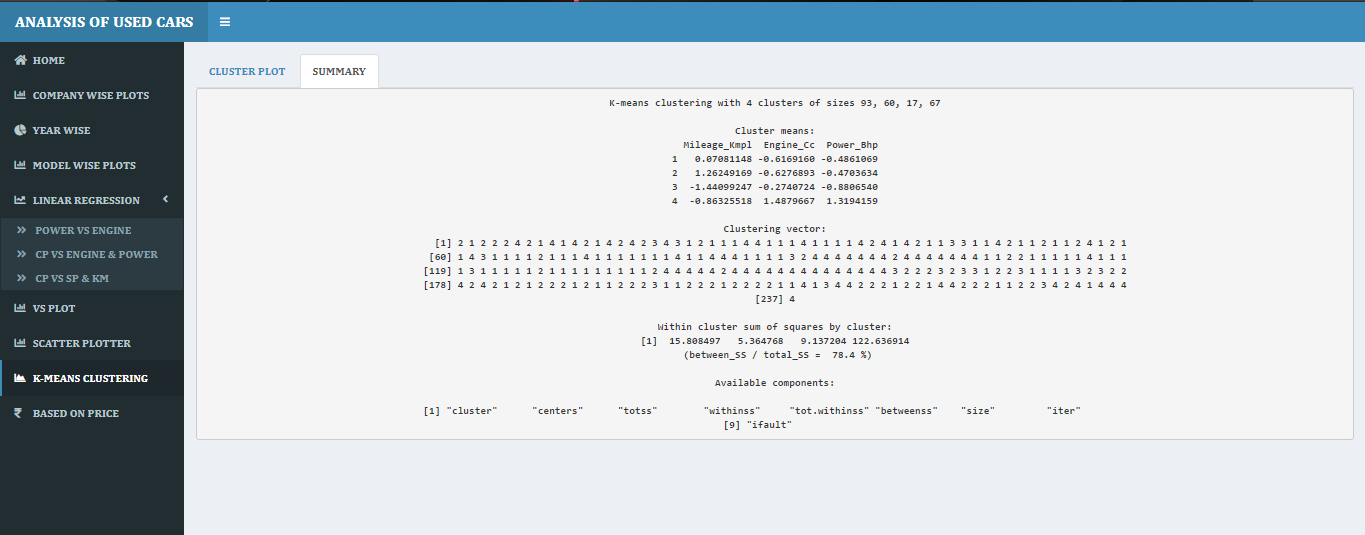
* **K-Means Clustering :-**

Select company name from drop down list ,then get cluster plot cluster size is 4.



* **Summary:-**

Accuracy for BMW company is 79.2% with 4 clusters.



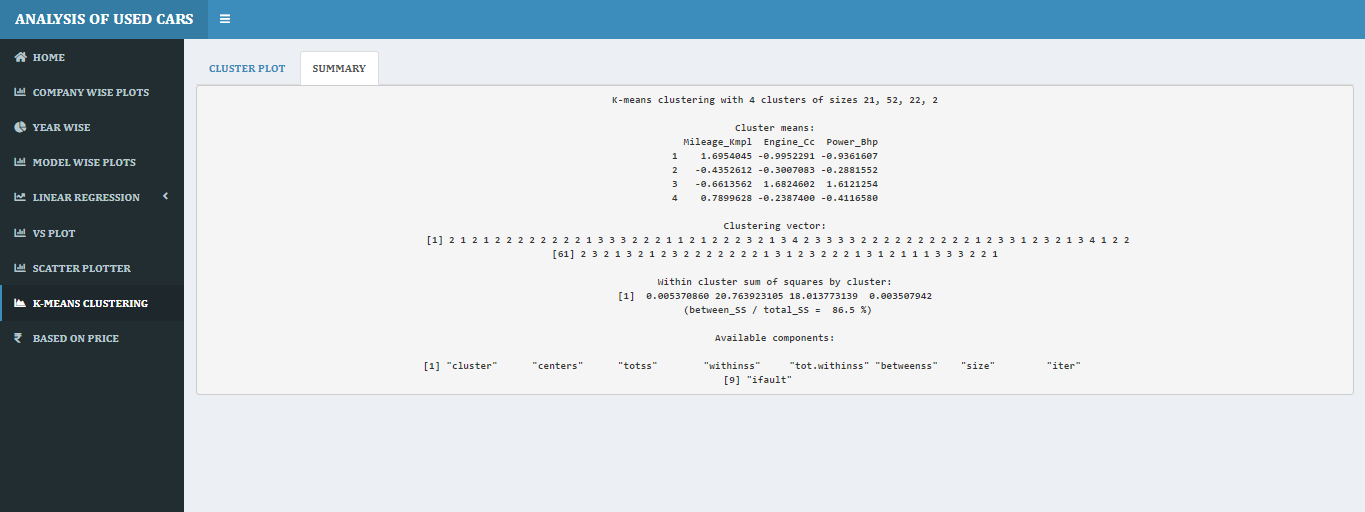
* **Selected Company :- Chevrolet**

K-Means cluster plot for chevrolet with 4 clusters



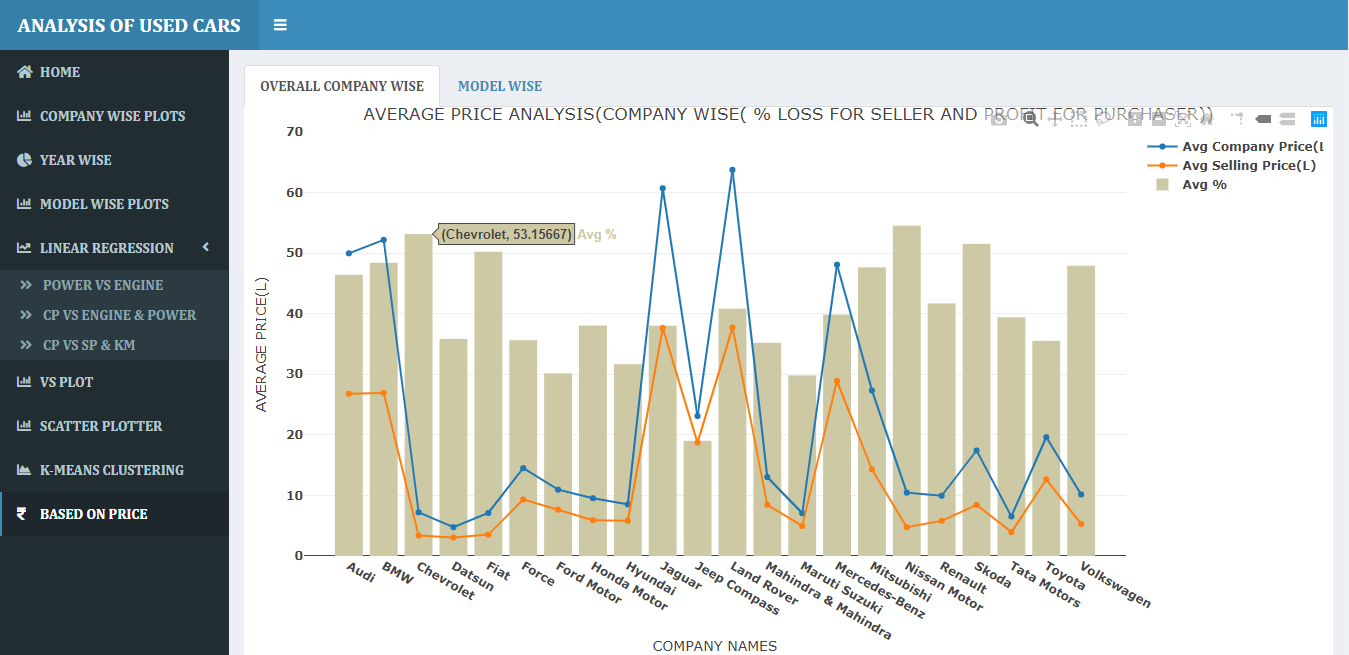
* **Summary:-**

Accuracy for BMW company is 86.5% with 4 clusters.

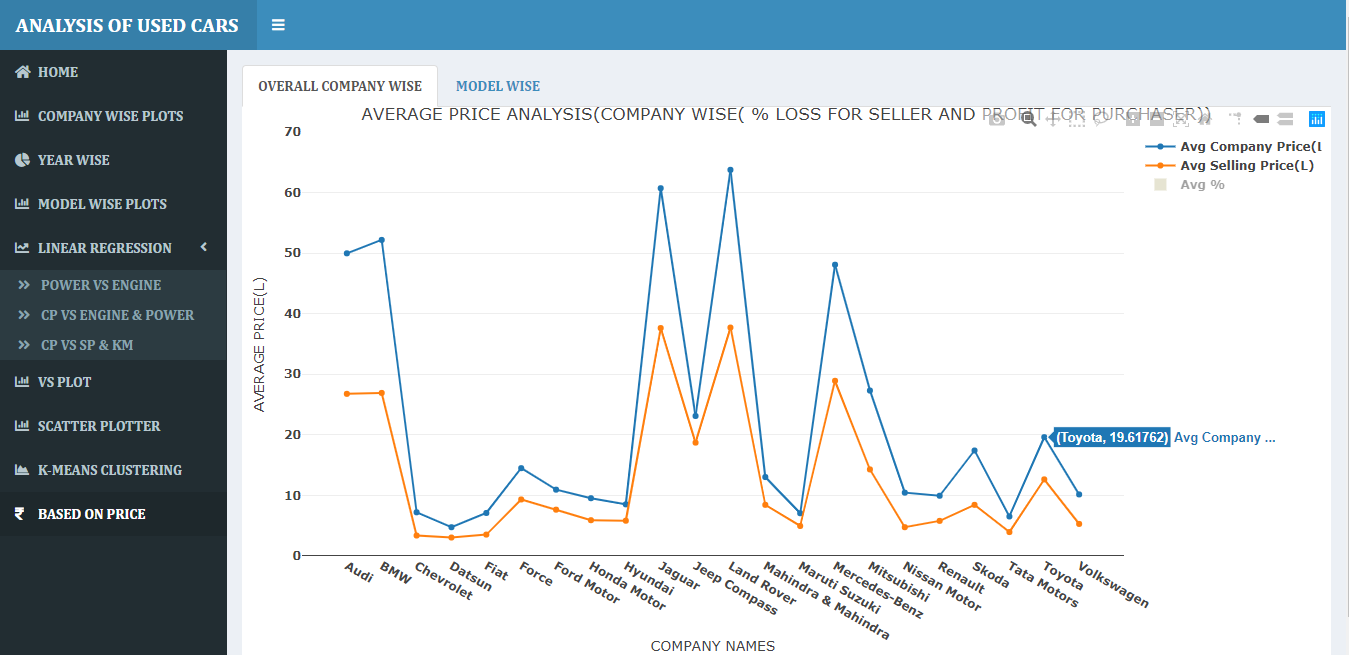


* **Based on price** **:-**
* **Company wise:-**

Bar graph for Avg % loss for seller and profit for purchaser. Line and dot graph for avg. company and avg. selling price.Based on average price, companies like Audi, Bmw, Chevrolet, Fiat, Mitsubishi, Nissan, Skoda & Volkswagen gives average 45% loss for seller and 45 % profit for purchaser.



**Graph for avg. company price and avg. selling price.**

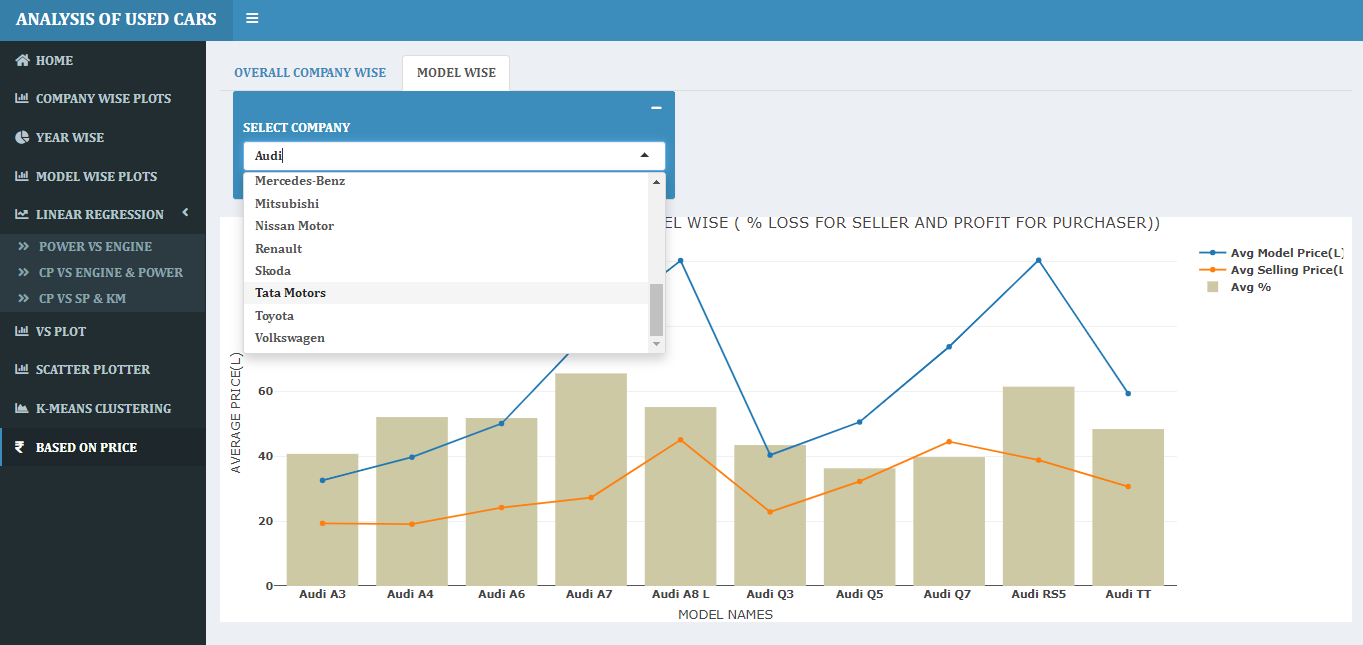


* **Model Wise :-**

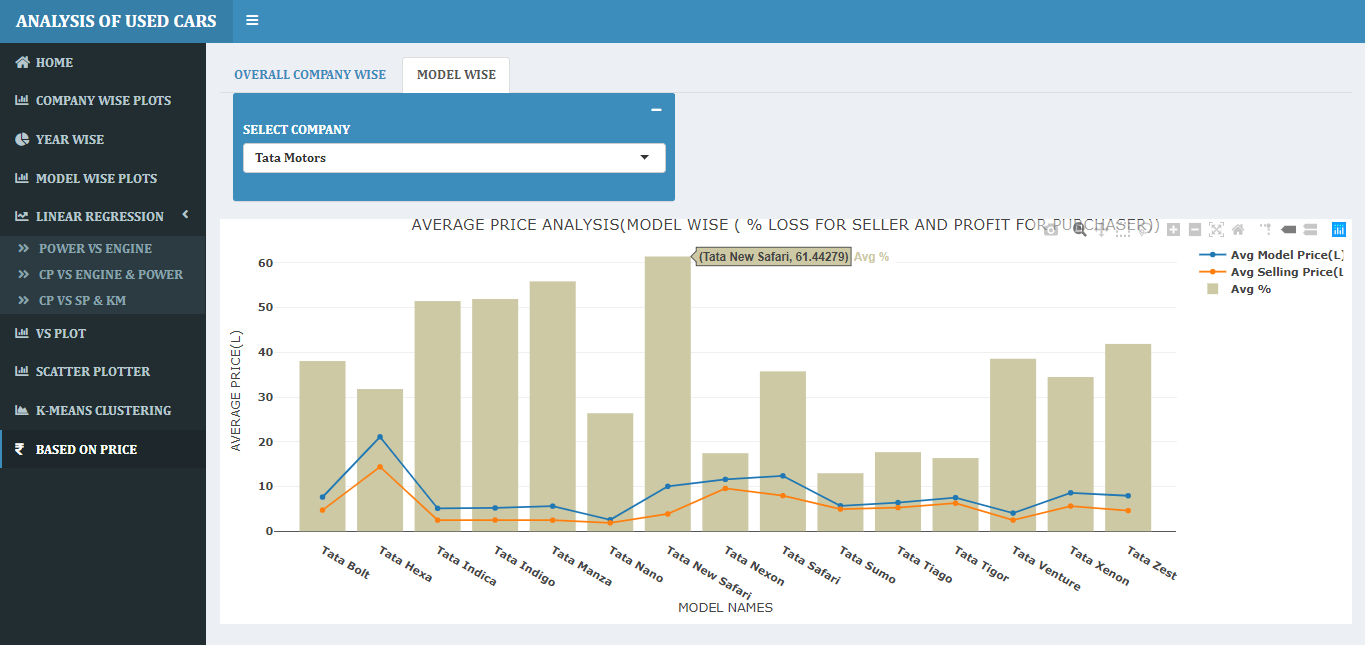
Select company name from drop down list, then get model wise graphs based on price.

Bar graph for Avg % loss for seller and profit for purchaser.

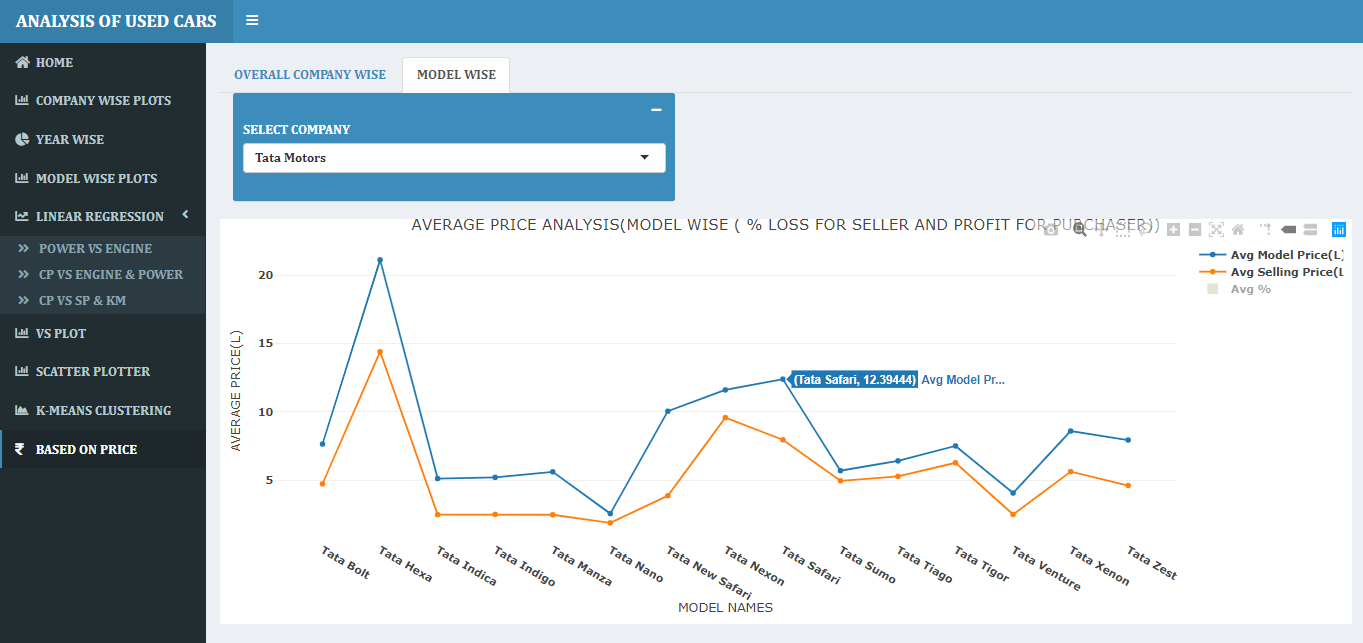
Line and dot graph for avg. company and avg. selling price.

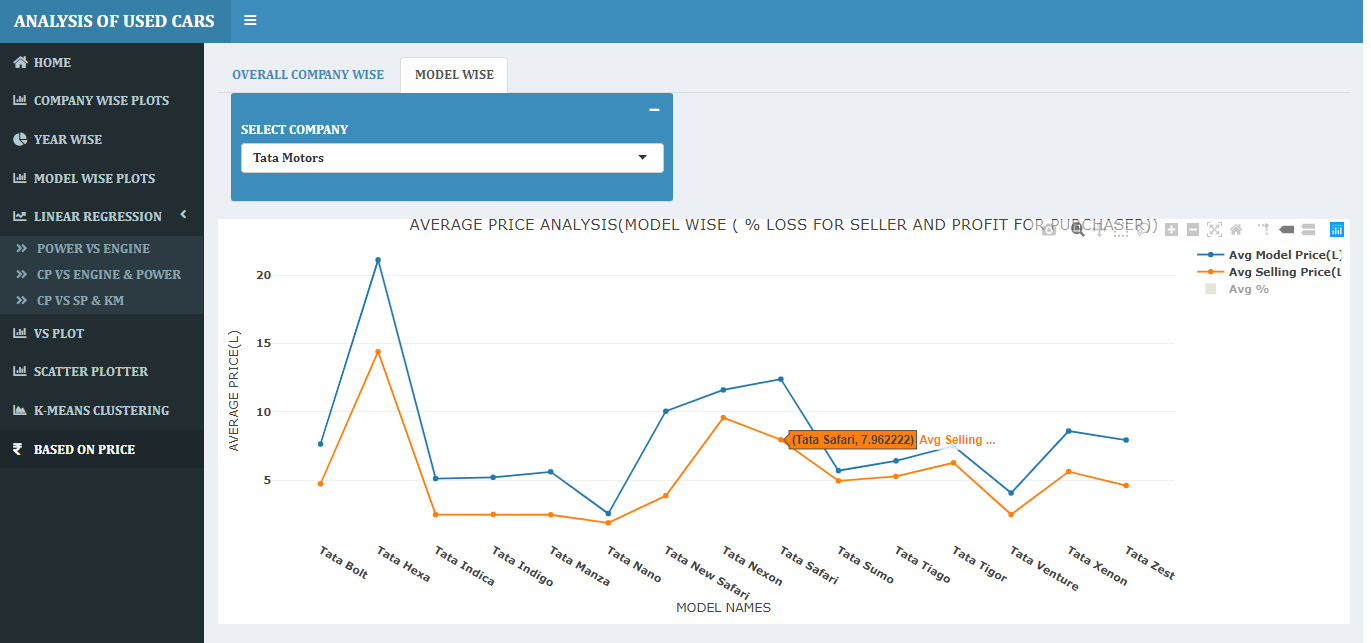
****

**Selected company name Tata motors , then gets model wise graphs**

****

**Line & dot Graph for avg. company price and avg. selling price model wise.**

****

****

* **CONCLUSIONS :-**
* **Linear Regression :-**

1. **Power vs Engine :-**

Power and engine linearly dependent on each other.

The accuracy is 75%, Here gives the engine\_cc value and it will

predict the value for power.

Engine=998 and predicted value is 65.07 for power (actual = 67.04).

1. **Company price vs engine & power :-**

Company price, engine & power linearly dependent on each other.

The accuracy is 77%, here gives the power & engine\_cc value and it will predict the value for Company price.

Power\_Bhp=67.04,Engine\_Cc=998 and predicted value is 3.36L for Company price (actual=4L).

1. **Selling price vs Company price & km driven**

Selling price ,Company price & km driven linearly dependent on each other.

The accuracy is 85%, here gives the Company price & km driven value and it will predict the value for Selling price. Company\_Price\_L=4.2 , Kilometers\_Driven = 50000 and predicted value is 2.94L for Selling price (actual=2.75 L).

* **K-Means clustering :-**

1. I create K-Means clustering for company wise using 3 columns

(Mileage, Engine & Power) with fixed number of cluster size 4 .

1. The accuracy for clustering varies company wise.
2. Accuracy in between (65 to 99), it’s depends on company**.**

* With the growing family incomes and ever-growing middle class society, India is becoming one of the most potential markets meant for used cars.
* Maruti Suzuki has the highest number of sales for used cars and it occupied almost 21% of market follow by Hyundai(19%) and the least number of sales of Datsun(0.2%) and Force(0.05%) respectively.
* Diesel and petrol cars are mostly used, whereas CNG & LPG cars are rarely used.
* Mileage range between 12-23kmpl, avg. minimum mileage gives by land rover(12) and maximum mileage gives by Maruti Suzuki(23).
* Two types of transmission used in cars, Mostly used manual transmission where company like BMW and AUDI are used Automatic.
* Avg. Engine range between 950-2600CC.
* Based on average price, companies like Audi, Bmw, Chevrolet, Fiat, Mitsubishi, Nissan, Skoda & Volkswagen gives average 45% loss for seller and 45 % profit for purchaser.
* Companies like Maruti Suzuki, Honda, Hyundai, Ford, Renault, Datsun, Force gives average 25-35% loss for seller and profit for purchaser.
* The general consensus among the industry is that the Used cars segment may become almost double of the new car market in another five years as is the case in india.
* **FUTURE SCOPE AND DEVELOPMENT :-**

1. This analysis is major useful for customer who want to buy used car for better understanding.
2. The knowledge gain from analysis helps to predict the which company and model gives the more profit to customer who want to buy used car.
3. This analysis is major useful for seller who want to sale car.
4. In future with help of this data we can also analyze the data for electrical cars.
5. Helps to get the location wise year wise sells of used cars in india.

* **REFERENCES AND APPENDIX :-**
* **References:-**
* <https://www.worldwidejournals.com/paripex/recent_issues_pdf/2016/November/November_2016_1478613517__74.pdf>
* <http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/>
* <http://www.rpubs.com/swapnilkura/ClusterAnalysis>
* <https://www.researchgate.net/publication/319306871_Predicting_the_Price_of_Used_Cars_using_Machine_Learning_Technique>
* <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/112123/lycumich.pdf?sequence=1&isAllowed=y>
* **Appendix :-**
* **Linear regression:-**

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1\*x

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

* **K-Means Cluster Analysis :-**

Clustering is a broad set of techniques for finding subgroups of observations within a data set. When we cluster observations, we want observations in the same group to be similar and observations in different groups to be dissimilar. Because there isn’t a response variable, this is an unsupervised method, which implies that it seeks to find relationships between the observations without being trained by a response variable. Clustering allows us to identify which observations are alike, and potentially categorize them therein. K-means clustering is the simplest and the most commonly used clustering method for splitting a dataset into a set of k groups.

K-means clustering is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents the number of groups pre-specified by the analyst. It classifies objects in multiple groups (i.e., clusters), such that objects within the same cluster are as similar as possible (i.e., high intra-class similarity), whereas objects from different clusters are as dissimilar as possible (i.e., low inter-class similarity). In k-means clustering, each cluster is represented by its center (i.e, centroid) which corresponds to the mean of points assigned to the cluster.

**-----------------THANK YOU------------------**