**House Price Prediction**

***Data Science Course Project***

AY 2020-2021

Semester 1

Contents

|  |  |  |
| --- | --- | --- |
| Sr. No. | Title | Page No. |
| 1 | Introduction | 1 |
| 2 | Univariate Analysis | 2 |
| 3 | Bivariate Analysis | 8 |
| 4 | Model Building & Prediction - 1 | 19 |
| 5 | Factor Analysis | 25 |
| 6 | Model Building & Prediction - 2 | 32 |

* Introduction

House Price Prediction is one the most important factors in Real Estate. This project focuses on predicting prices of houses. This project uses Regression model to predict house price. In this project a very famous dataset “Boston” has been used. Below is a screenshot showing a small part of the dataset.

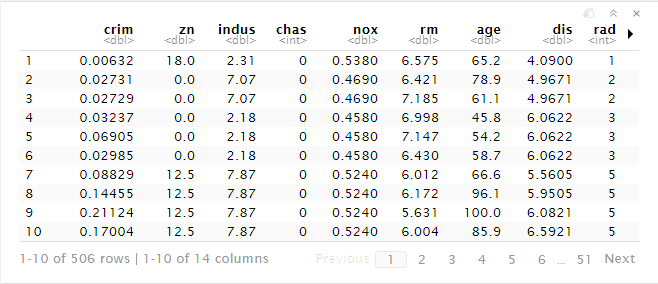


Figure 1: Boston Dataset - Part 1

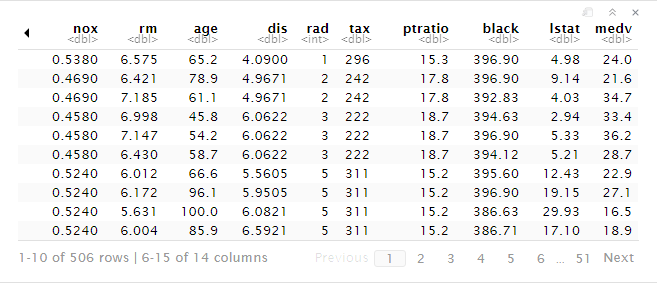


Figure 2: Boston Dataset - Part 2

* Description of Dataset
  + CRIM: Per capita crime rate by town
  + ZN: Proportion of residential land zoned for lots over 25,000 sq. ft.
  + INDUS: Proportion of non-retail business acres per town
  + CHAS: Charles river dummy variable (= 1 if tract bounds river; 0 = otherwise)
  + NOX: Nitric Oxide concentration (parts per 10 million)
  + RM: Average number of rooms per dwelling
  + AGE: Proportion of owner-occupied units built prior to 1940
  + DIS: Weighted distances to five Boston employment centres
  + RAD: Index of accessibility to radial highways
  + TAX: Full-value property tax rate per $10,000
  + PTRATIO: Pupil-teacher ratio by town
  + BLACK: 1000(Bk - 0.63)2, where Bk is the proportion of Afro-American descent by town
  + LSTAT: Percentage of lower status of the population
  + MEDV: Median value of owner-occupied homes in $1000s

Here, medv is the feature that we are going to predict.

Below is the screenshot of the summary of the dataset.

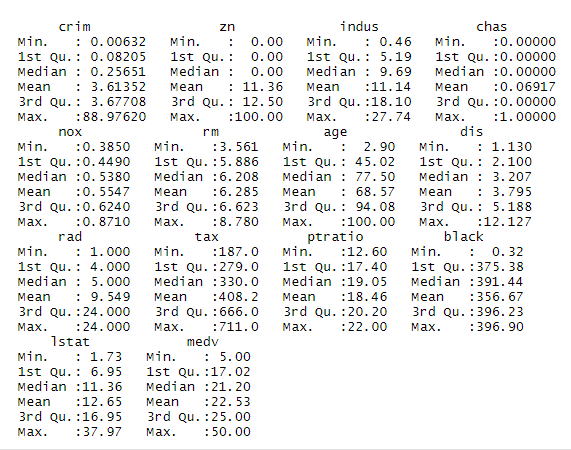


Figure 3: Summary of Boston Dataset

* Univariate Analysis

Ggplot2 library has been used to draw the histograms for each feature of the dataset. Histogram gives us the frequency of all the values. Univariate analysis has been done to see how the independent variables are distributed.

1. CRIM:

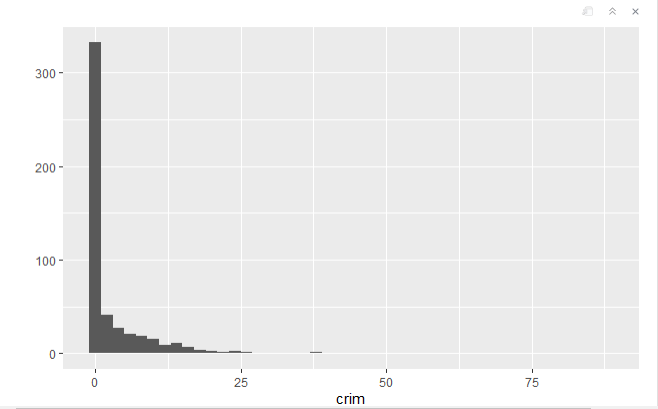


Figure 4: Histogram for CRIM

*Here, we can observe that the frequency of values of crime rate near to rather zero (0) are the most i.e. above 300. The frequency of crime rate per capita being more than 1 is very less. The frequency decreases as the crime rate increases.*

1. ZN:

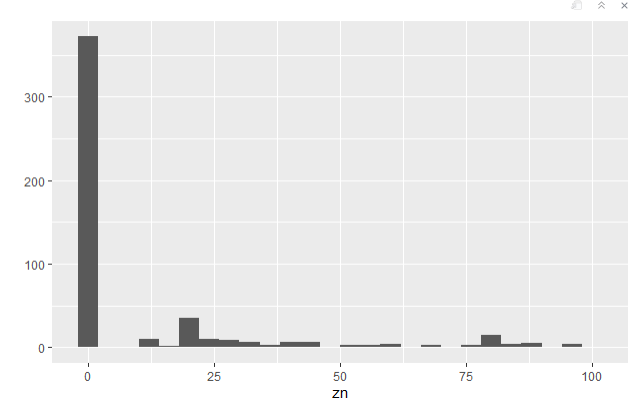


Figure 5: Histogram for ZN

*Here, we can observer that the frequency of zn having value 1 is the most i.e. above 350. The frequency of zn being more than 1 is very less.*

1. INDUS:

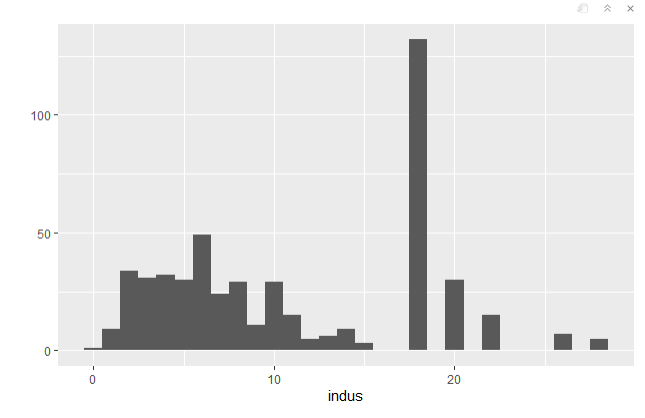


Figure 6: Histogram for INDUS

*Here, we can observe that the frequency of indus having value 17-18-19 is the most i.e. above 125.*

1. CHAS:

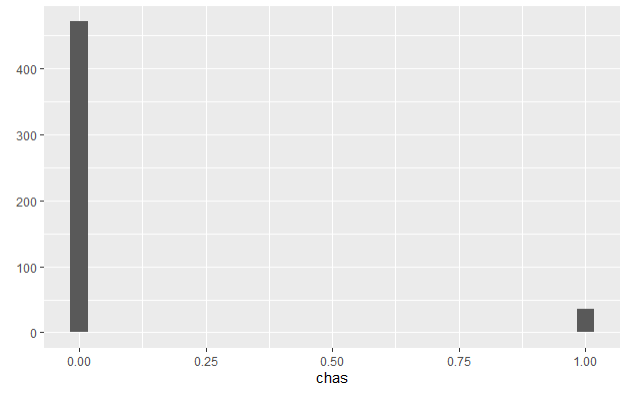


Figure 7: Histogram for CHAS

*Here, we can observe that the frequency of chas having value 0 is the most i.e. above 450. Comparatively, chas having value 1 is less i.e. around 45.*

1. NOX:

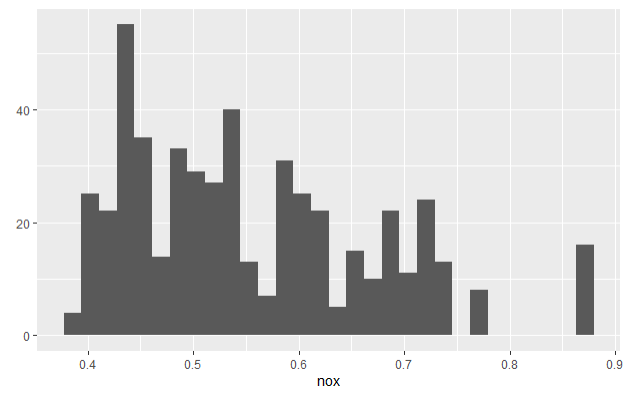


Figure 8: Histogram for NOX

*Here, we can observe that the distribution for NOX is right skewed.*

1. RM:

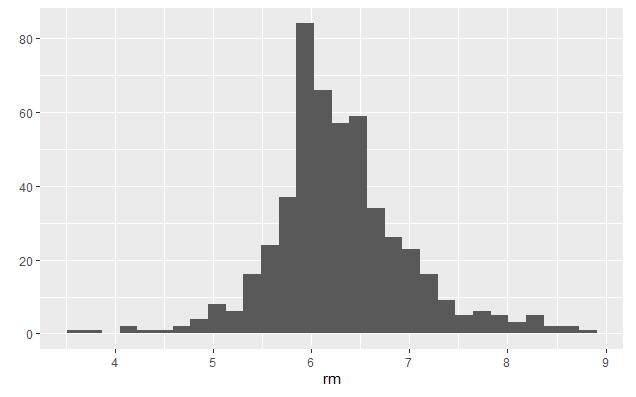


Figure 9: Histogram for RM

*Here, we can observe that the distribution for RM is normal.*

1. AGE:

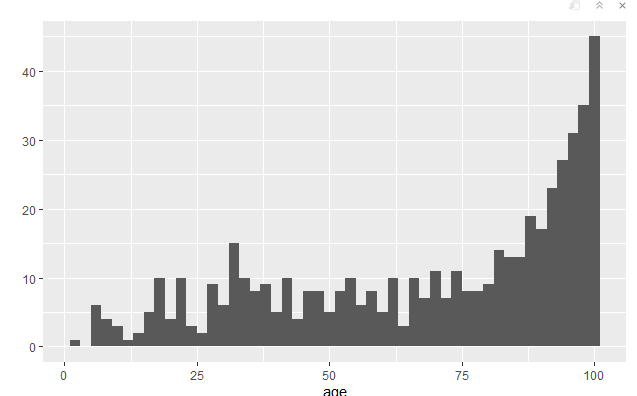


Figure 10: Histogram for Age

*Here, we can observe that the distribution for AGE is left skewed. The extreme values to the right have high frequency.*

1. DIS:

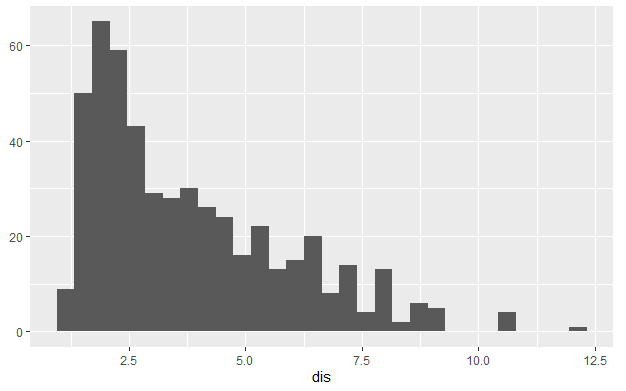


Figure 11: Histogram for DIS

*Here, we can observe that the distribution for DIS is right skewed. The extreme values to the left have high frequency.*

1. RAD:

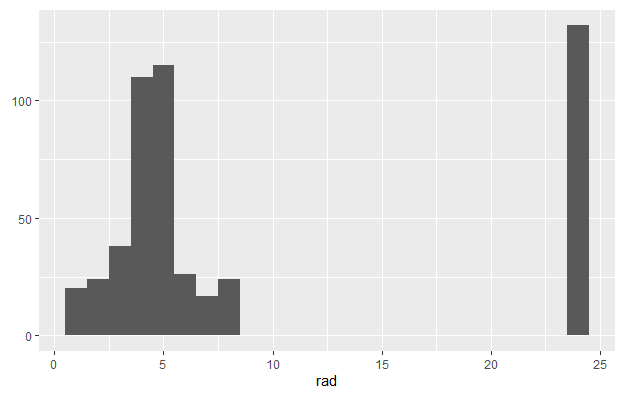


Figure 12: Histogram for RAD

*Here, RAD is a factor. Frequency of RAD having values near to 24 is more. Also RAD having value near to 5 has high frequency.*

1. TAX:

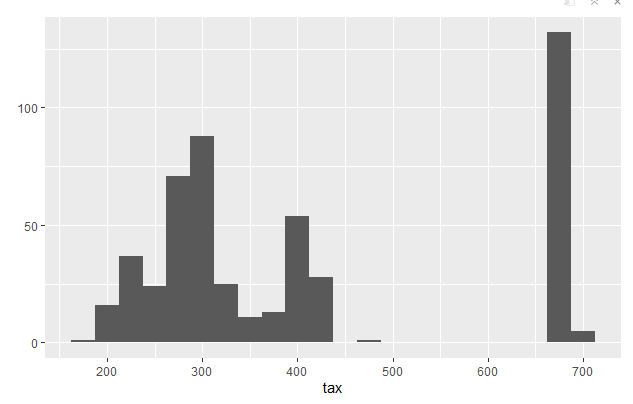


Figure 13: Histogram for TAX

*Here, we can observe that TAX having values between 660-680 has a high frequency.*

1. PTRATIO:

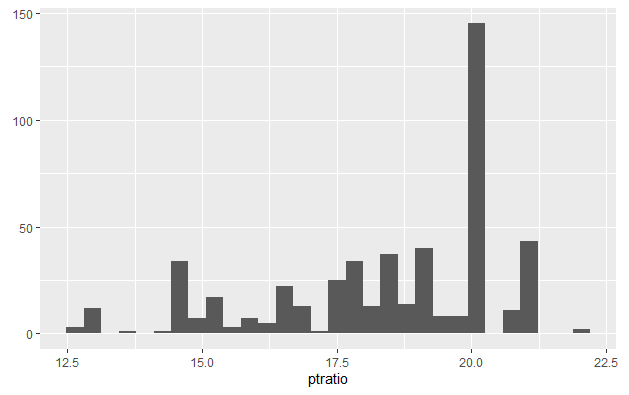


Figure 14: Histogram for PTRATIO

*Here, we observe that ptratio having value 20 has the highest frequency.*

1. BLACK:

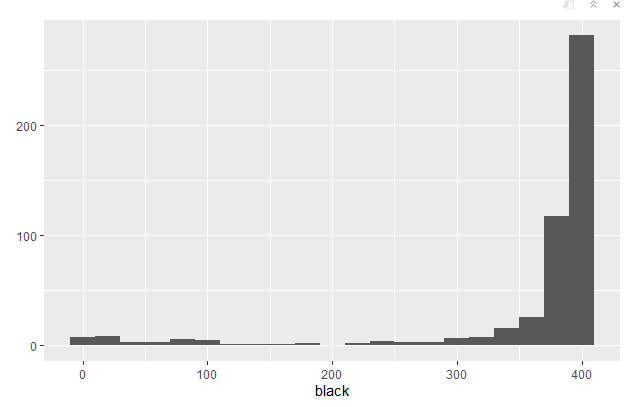


Figure 15: Histogram for BLACK

*Here, we observe that black having value 400 has the highest frequency.*

1. LSTAT:

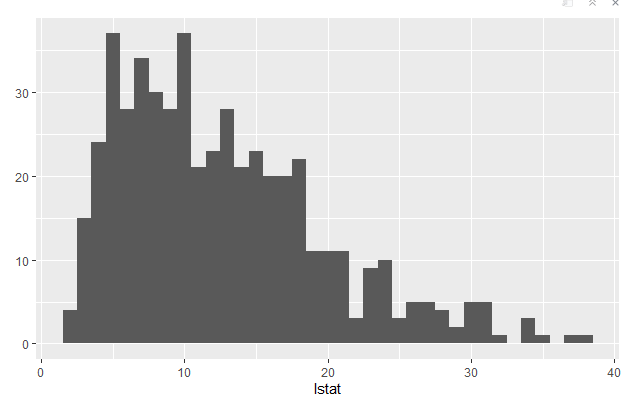


Figure 16: Histogram for LSTAT

*Here, that the distribution for LSTAT is right skewed. Also the frequency of LSTAT having values 5 and 11 is high.*

1. MEDV

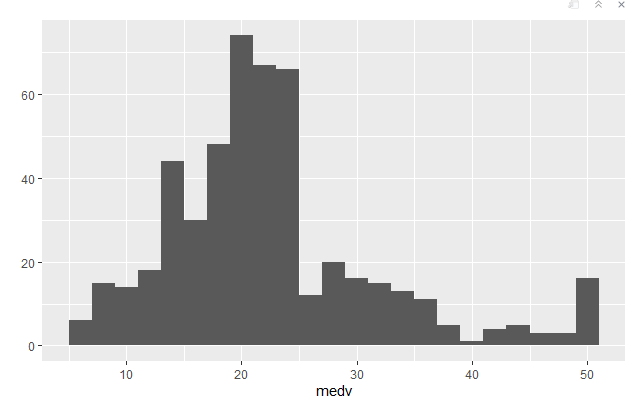


Figure 17: Histogram for MEDV

*Here, it is observed that the distribution for MEDV is nearly normal.*

* Bivariate Analysis

In this project bivariate analysis has been done to find the correlation between all the independent variables and MEDV. With this, it becomes easy to analyse which features affect the Median Price most and which features can be dropped or features who don’t have much significance. Model building is the process of developing a probabilistic model that best describes the relationship between the dependent and independent variables.

1. CRIM & MEDV

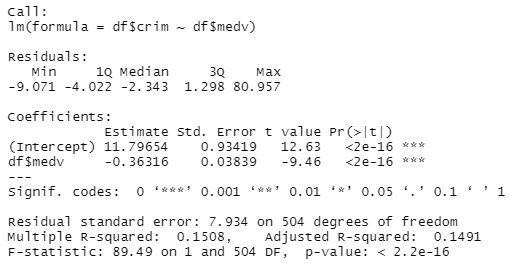


Figure 18: Summary for CRIM & MEDV Linear Model

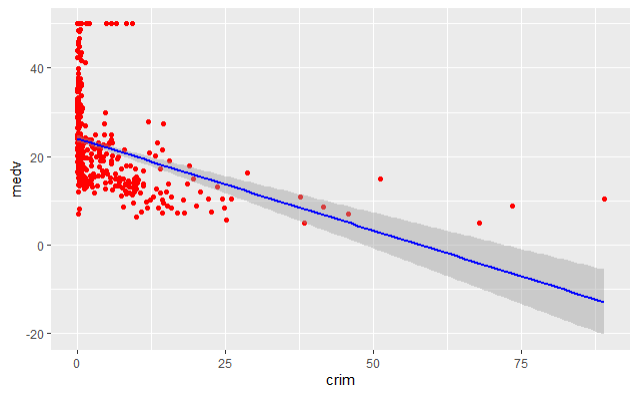


Figure 19: Linear Model between CRIM & MEDV

*Here, it is observed that there is negative correlation between CRIM & MEDV. As the crime rate increases the Median price decreases which is quite obvious.*

1. ZN & MEDV

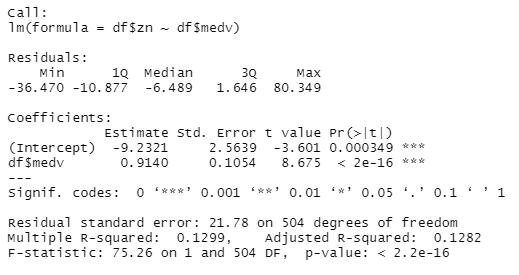


Figure 20: Summary for ZN & MEDV Linear Model

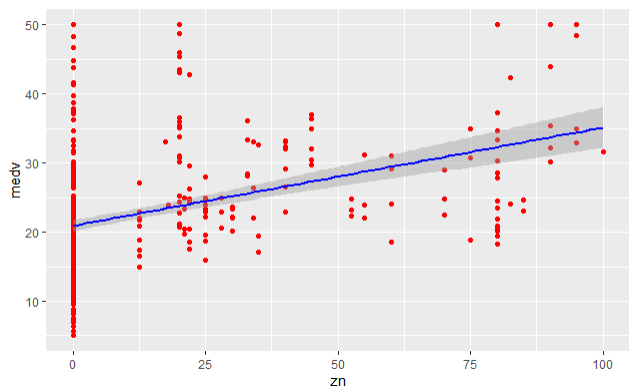


Figure 21: Linear Model between ZN & MEDV

*Here, it is observed that there is positive correlation between ZN & MEDV. As the proportion of land zoned increases the Median price increases.*

1. INDUS & MEDV

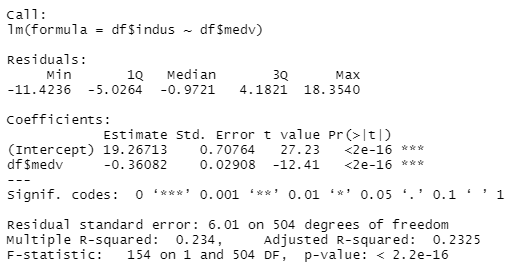


Figure 22: Summary for INDUS & MEDV Linear Model

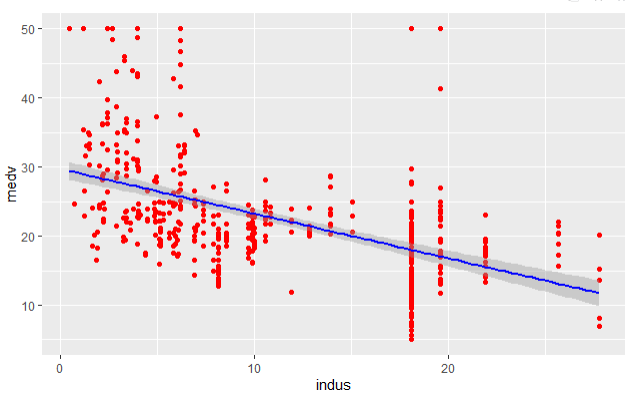


Figure 23: Linear Model between INDUS & MEDV

*Here, it is observed that there is negative correlation between INDUS & MEDV. As the proportion of non-retail business increases the price decreases.*

1. CHAS & MEDV

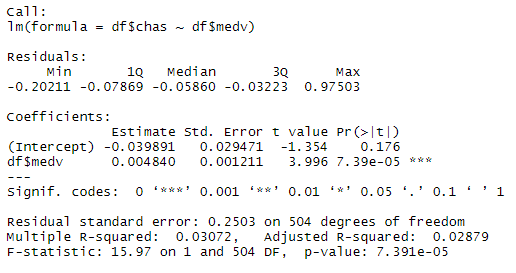


Figure 24: Summary for CHAS & MEDV Linear Model

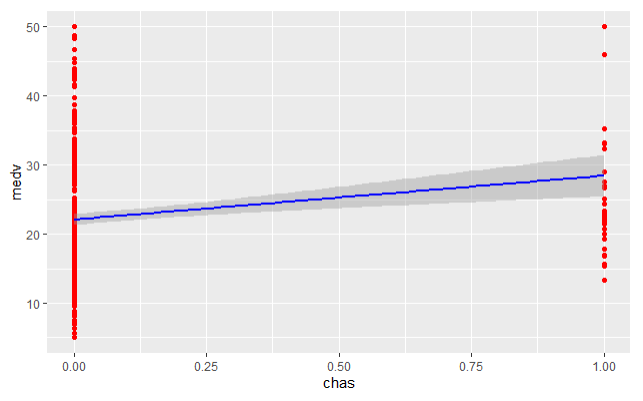


Figure 25: Linear Model between CHAS & MEDV

*Here, as CHAS is just a factor and its value is either 0 or 1, it is difficult to say if MEDV and CHAS do have a correlation. But by looking at the summary, it is observed that CHAS has a significant role in determining the prices.*

1. NOX & MEDV

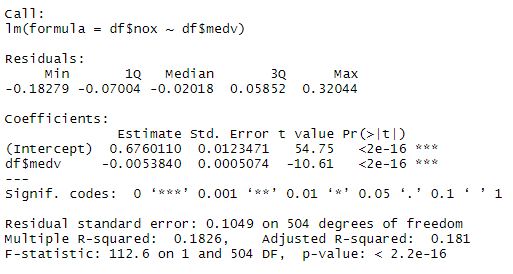


Figure 26: Summary for NOX & MEDV Linear Model

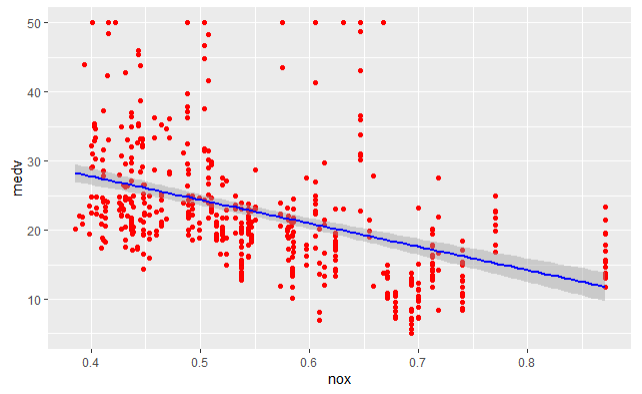


Figure 27: Linear Model between NOX & MEDV

*Here, it can be observed that there is negative correlation between NOX and MEDV. As the nitric oxide concentration increases the price decreases which is again quite obvious.*

1. RM & MEDV

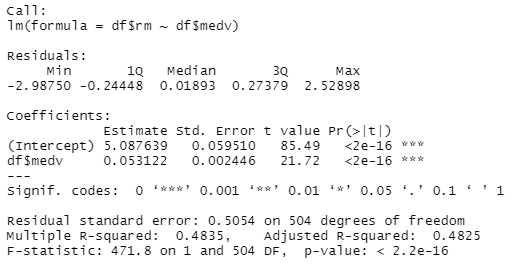


Figure 28: Summary for RM & MEDV Linear Model

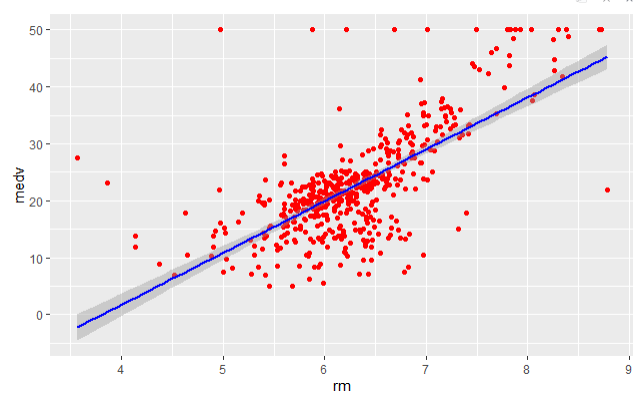


Figure 29: Linear Model between RM & MEDV

*Here, it is observed that there is positive correlation between RM & MEDV. This correlation is the highest that we have seen so far.*

1. AGE & MEDV

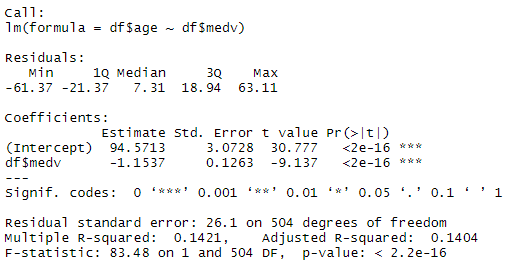


Figure 30: Summary for AGE & MEDV Linear Model

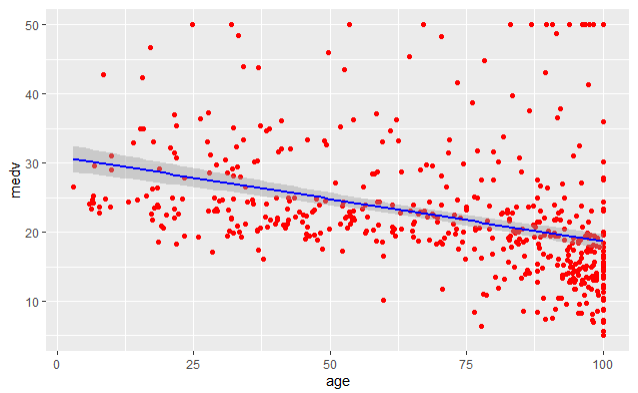


Figure 31: Linear Model between AGE & MEDV

*Here, it can be observed that there is negative correlation between AGE & MEDV. There is slight declination, as the AGE increases, price decreases.*

1. DIS & MEDV

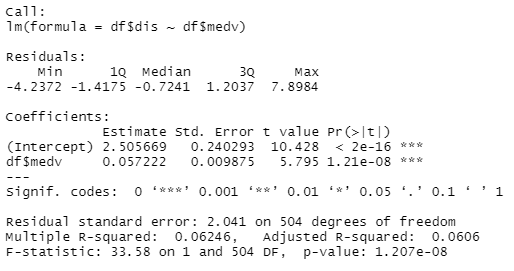


Figure 32: Summary for DIS & MEDV Linear Model

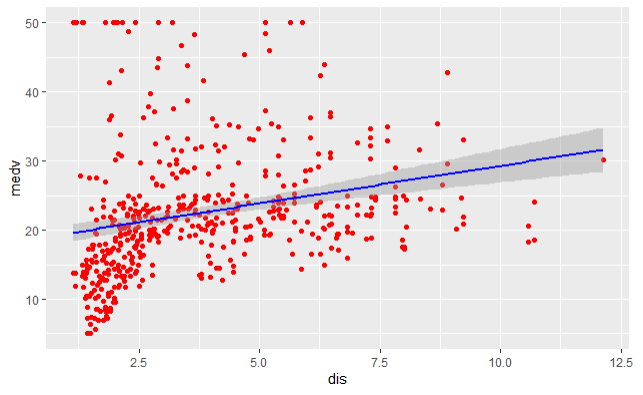


Figure 33: Linear Model between DIS and MEDV

*Here, it is observed that there is a positive correlation between DIS and MEDV. As the distance to main 5 employment centres increases the median price also increases, which is slightly surprising.*

1. RAD & MEDV

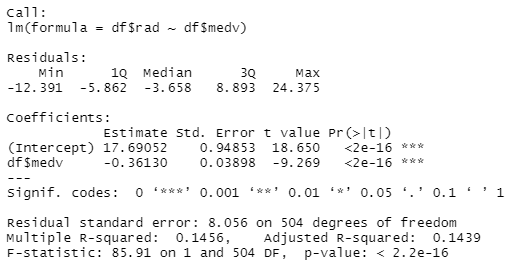


Figure 34: Summary for RAD & MEDV Linear Model

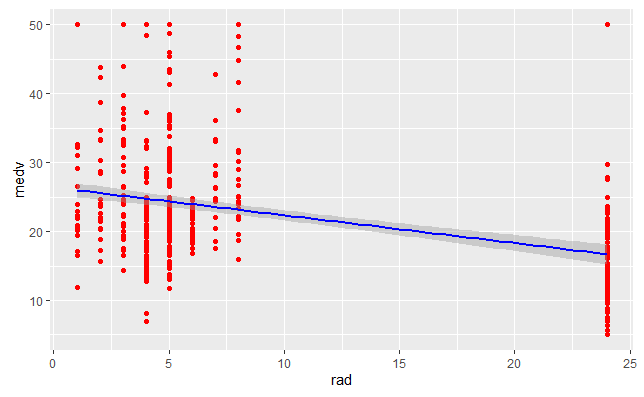


Figure 35: Linear Model between RAD & MEDV

*Here, RAD is also a factor. Looking at the Linear Model we can derive that there is definitely a correlation but to what extent is not clear. The summary shows that the correlation is quite significant.*

1. TAX & MEDV

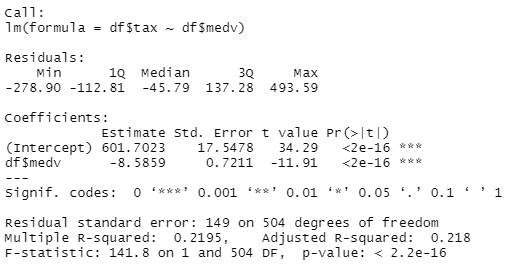


Figure 36: Summary for TAX & MEDV Linear Model

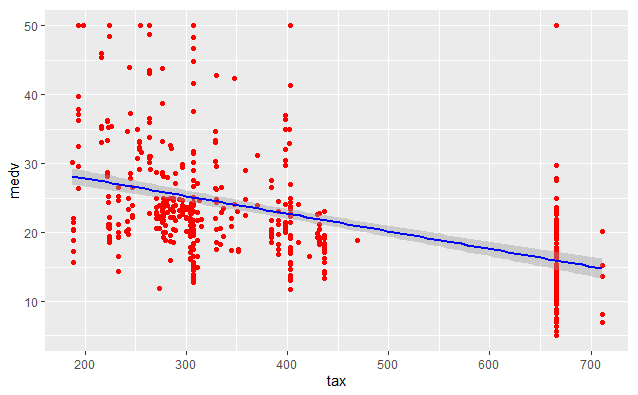


Figure 37: Linear Model between TAX & MEDV

*Here, it is observed that there is a negative correlation between TAX and MEDV. As the property tax increases the median price decreases, which is again quite obvious.*

1. PTRATIO & MEDV

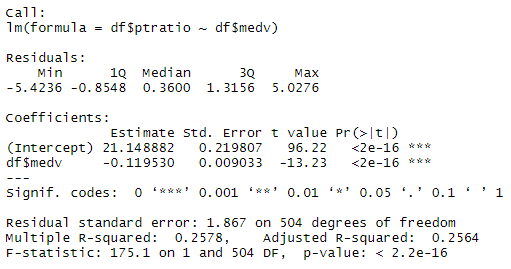


Figure 38: Summary for PTRATIO & MEDV Linear Model

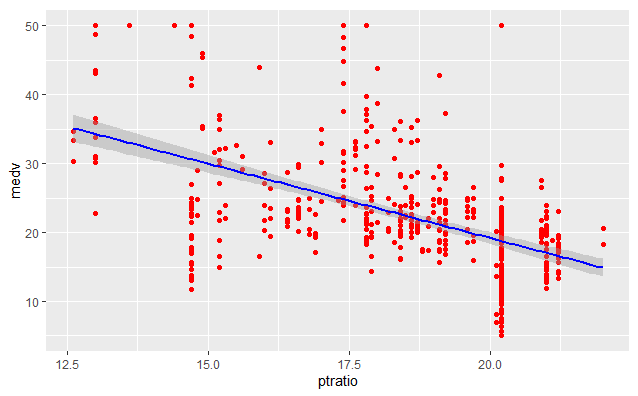


Figure 39: Linear Model between PTRATIO & MEDV

*Here, it can be observed that there is negative correlation between PTRATIO & MEDV. As the Pupil Teacher Ratio increases the median price decreases which seem slightly unconventional.*

1. BLACK & MEDV

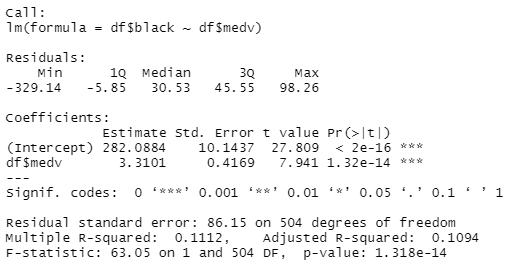


Figure 40: Summary for BLACK & MEDV Linear Model

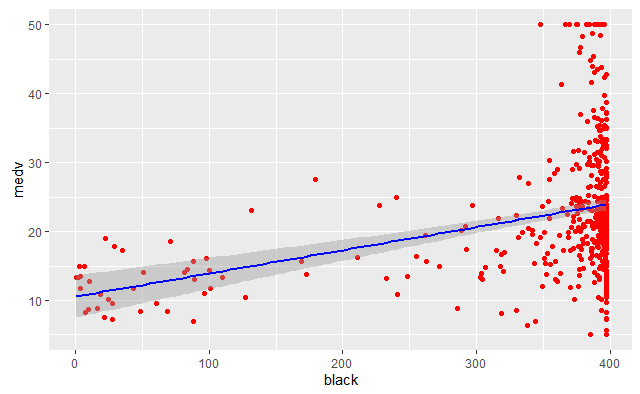


Figure 41: Linear Model between BLACK & MEDV

*Here, it is observed that there is positive correlation between BLACK & MEDV. As the number of Afro-Americans increases the price also increases, which is quite surprising as we are talking about USA, where there is racial discrimination.*

1. LSTAT & MEDV

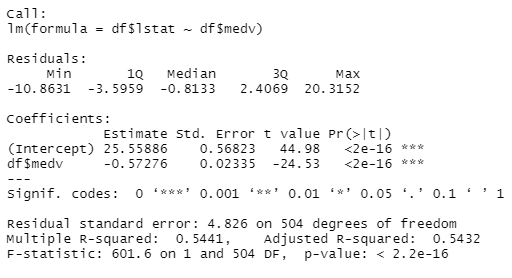


Figure 42: Summary for LSTAT & MEDV Linear Model

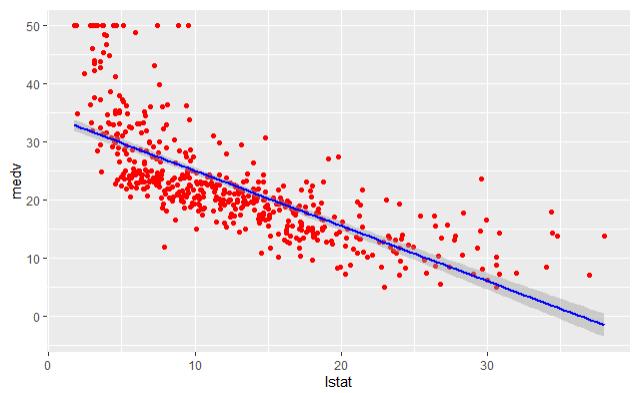


Figure 43: Linear Model between LSTAT & MEDV

*Here, it can be observed that there is negative correlation between LSTAT & MEDV. As the lower status of people increases the median price obviously decreases.*

* Correlation:

Just to confirm the correlation that we have drawn from the linear models, a correlation has been drawn.

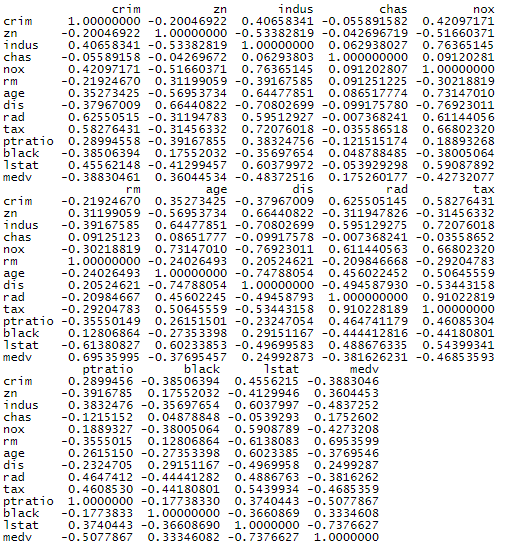


Figure 44: Correlation between All the Features

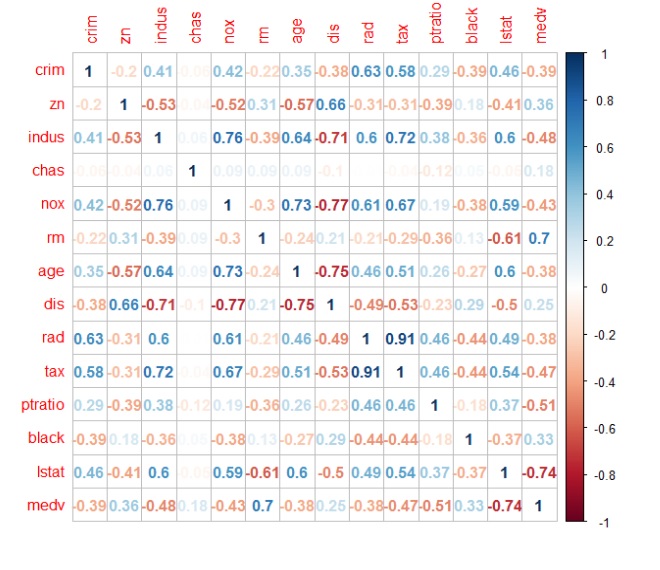


Figure 45: Correlation Plot between All the Features

From the Linear Models as well as from the Correlation plots that have been drawn, it is quite evident that all the features do contribute in predicting the Median price.

From the Correlation plot, it is also evident that there is also a lot of collinearity in the independent variables as well. Despite this collinearity we will get a high R-squared, but it will cause problems in the Regression model. A simple way to remove this collinearity is Factor Analysis.

Just to show the difference, models have been created before and after doing Factor Analysis.

* Model Building & Prediction – Part 1

Before building the models, dataset has been divided into train dataset and test dataset in ratio 0.75.

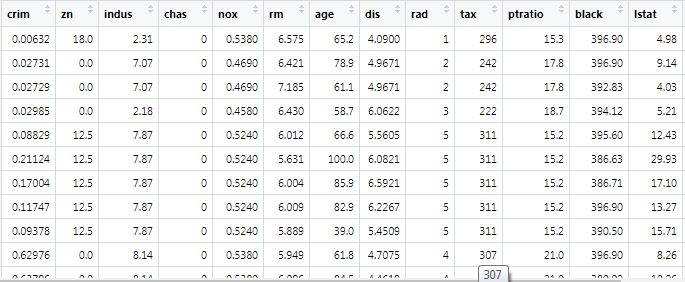


Figure 46: Screenshot of train dataset

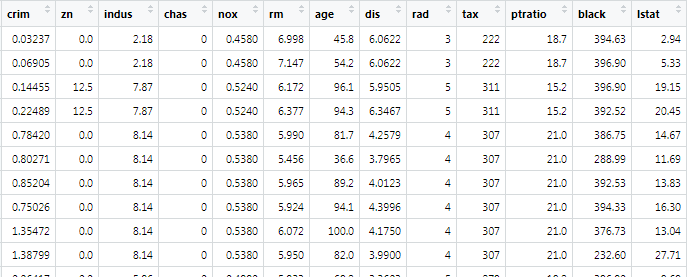


Figure 47: Screenshot of test dataset

Here, a Multiple Linear Regression model has been built. Multiple Regression is a statistical method that uses several explanatory variables to predict the outcome of a response variable.

**Below is the first MLR model: -**

In this model we have taken “medv” of the boston\_train dataset as the dependent variable and all other features of the dataset as the independent variables.

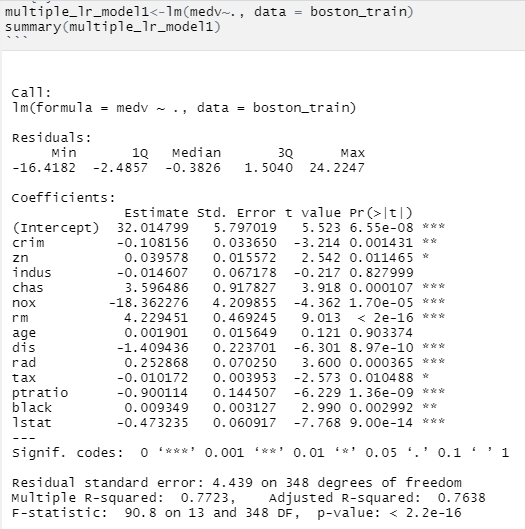


Figure 48: MLR Model 1

*If we look at the summary of multiple\_lr\_model1 first the minimum, median, maximum, 1Q and 3Q of residuals are given. Then the coefficients of estimate, std. error, t-value, p-value are given. In this model Indus and age features/variables have high p-value so their significance level is low in this model. The adjusted R-squared value is 0.7638.*

**Below is the second MLR Model: -**

In this model we have taken “medv” of the boston\_train dataset as the dependent variable and all other features excluding “chas” and adding factors of chas using factor () function as the independent variables.

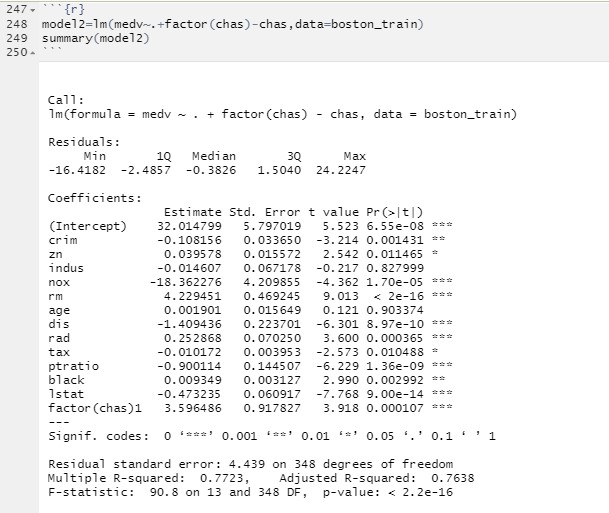


Figure 49: MLR Model 2

*If we look at the summary of the model the model the adjusted R-squared is equal to 0.7638.*

**Below is the third MLR Model: -**

In third model we taken “medv” of the boston\_train dataset as the dependent variable and all other features along with combination of “chas” an “dis” i.e. chas:dis as the independent variables.

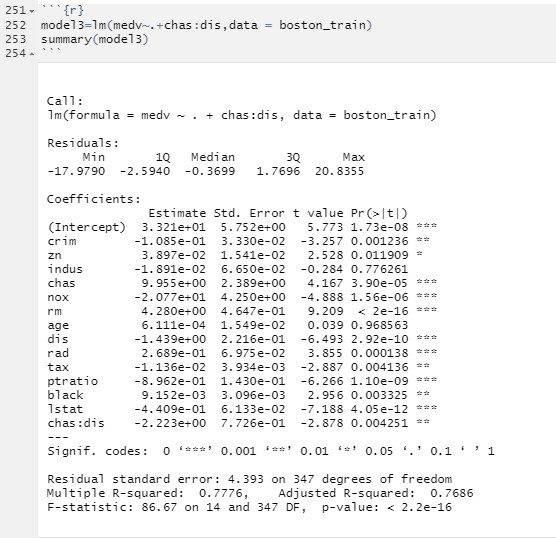


Figure 50: MLR Model 3

*If we look at the summary of the model the model the adjusted R-squared is equal to 0.7686.*

**Below is the fourth MLR Model: -**

In fourth model we have taken “medv” of the boston\_train dataset as the dependent variable and all other features excluding “indus” and combination of “chas” an “dis” i.e. chas:dis as the independent variables.

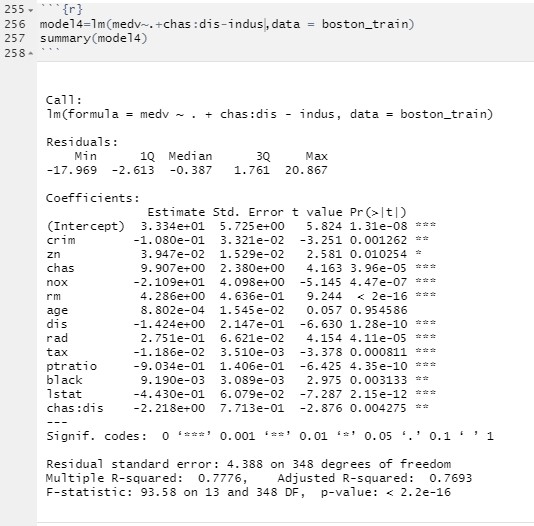


Figure 51: MLR Model 4

*If we look at the summary of the model the model the adjusted R-squared is equal to 0.7693, we have selected this model for prediction.*

**Below are the Prediction results: -**

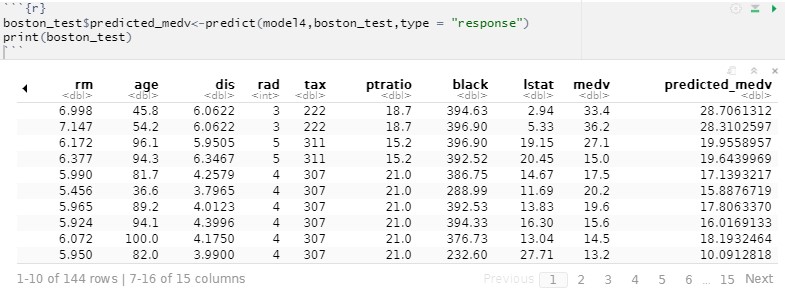


Figure 52: Prediction results – 1

*Now, the accuracy for this fourth model is 80.90%*

Here, we have also used the decision tree to predict the results.

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 and Data Mining applications using R.

**Below is the Decision Tree Model: -**

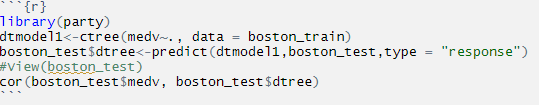
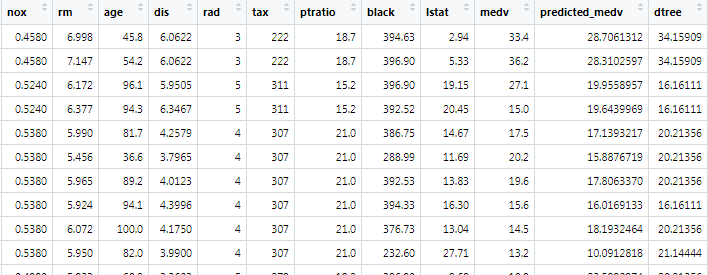
****

Figure 53: Decision Tree Model 1

**Below are the Decision Tree Prediction Results: -**



*Here, the accuracy for this model is 88.71%.*

* **Factor Analysis**

Variable Inflation Factor has been used to confirm the collinearity.

High Variable Inflation Factor (VIF) is a sign of multicollinearity. There is no formal VIF value for determining the presence of multicollinearity; however, in weaker models, VIF value greater than 3.0 may be a cause of concern.

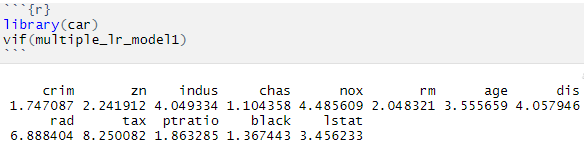


Figure 54: VIF 1

*From the VIF values, we can infer that variables INDUS, NOX, AGE, DIS, RAD and TAX are a cause of concern.*

**Remedial Measures: -**

Perform an analysis design like principal component analysis (PCA)/ Factor Analysis on the correlated variables.

**Factor Analysis: -**

Now let’s check the factorability of the variables in the dataset.  
First, let’s create a new dataset by taking a subset of all the independent variables in the data and perform the Kaiser-Meyer-Olkin (KMO) Test.

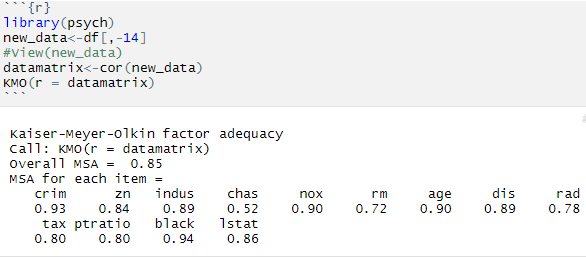
****

Figure 55: KMO Test

*Since MSA > 0.5, we can run Factor Analysis on this data.*

Bartlett’s test of sphericity should be significant

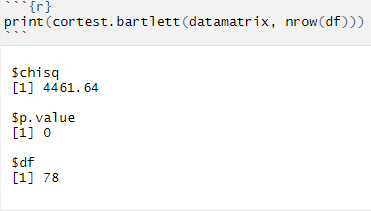


Figure 56: Bartlett's Test

*The Kaiser-Meyer Olkin (KMO) and Bartlett’s Test measure of sampling adequacy were used to examine the appropriateness of Factor Analysis. The approximate of Chi-square is 4461.64 with 78 degrees of freedom, which is significant at 0.05 Level of significance. The KMO statistic of 0.85 is also large (greater than 0.50). Hence Factor Analysis is considered as an appropriate technique for further analysis of the data.*

**Scree plot using ggplot: -**

One way to determine the number of factors or components in a data matrix or a correlation matrix is to examine the “scree” plot of the successive eigenvalues.

The scree plot graphs the Eigenvalue against each factor.

Selection of factors from the scree plot can be based on:  
1. Kaiser-Guttman normalization rule says that we should choose all factors with an eigenvalue greater than 1.  
2. Bend elbow rule.

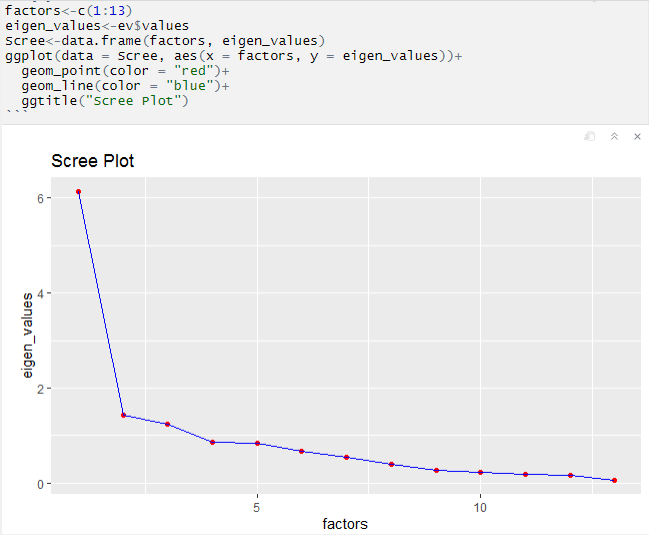


Figure 57: Scree Plot

*So as per the elbow or Kaiser-Guttman normalization rule, we are good to go ahead with 3 factors.*

Let’s use 3 factors to perform the factor analysis.

**Factor analysis using fa method: -**

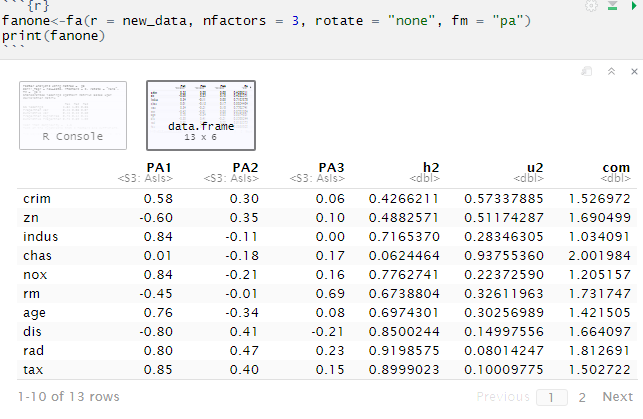
****

Figure 58: Fa Method 1

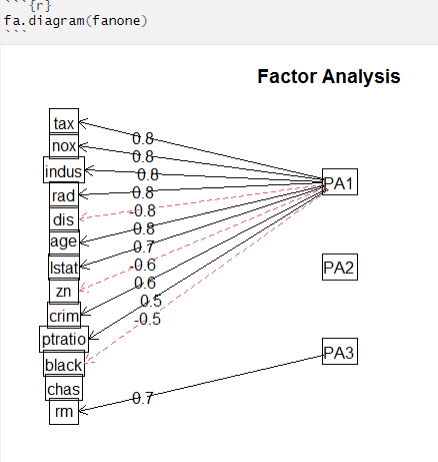


Figure 59: Fa Method 1 Diagram of Factors

*Here, we have used the oblique rotation method, which has resulted in the factors being correlated.*

Let’s use orthogonal rotation i.e. Virmax method which will result in the factors being uncorrelated.

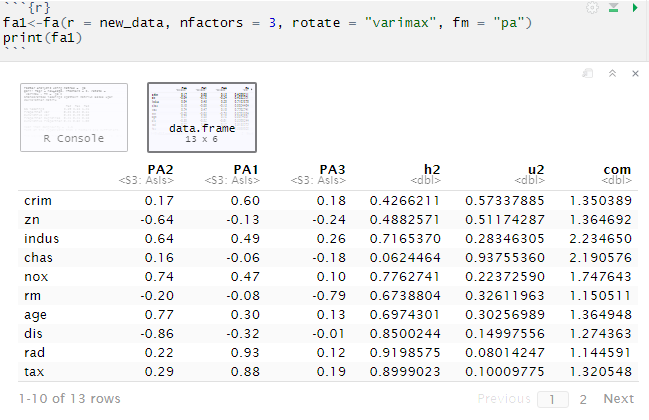


Figure 60: Fa Method 2

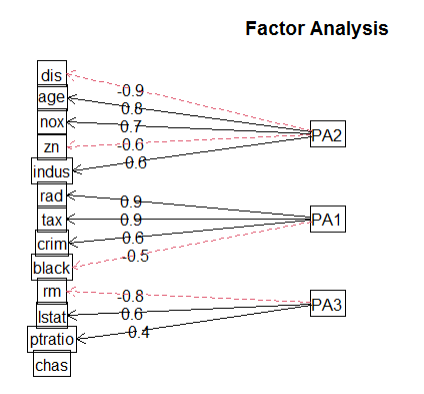


Figure 61: Fa Method 2 Diagram of Factors

**Regression analysis using the factors scores as the independent variable:**Let’s combine the dependent variable and the factor scores into a dataset and label them.

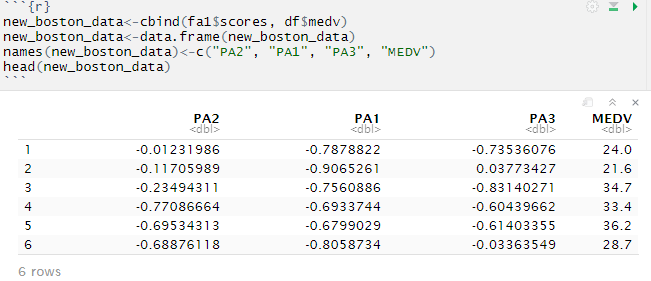


Figure 62: New, dataset of factors

Let’s split the dataset into training and testing dataset (75:25)



Figure 63: Splitting

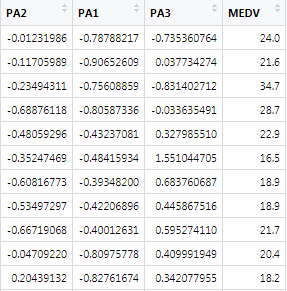


Figure 64: New Train dataset

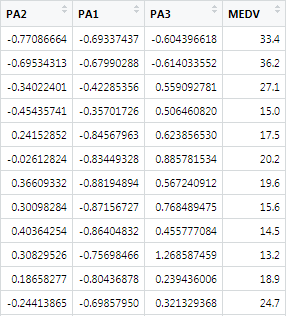


Figure 65: New Test Dataset

* **Model Building & Prediction 2: -**

In this model we have taken “medv” of the new\_train dataset as the dependent variable and all other features of the dataset as the independent variables.

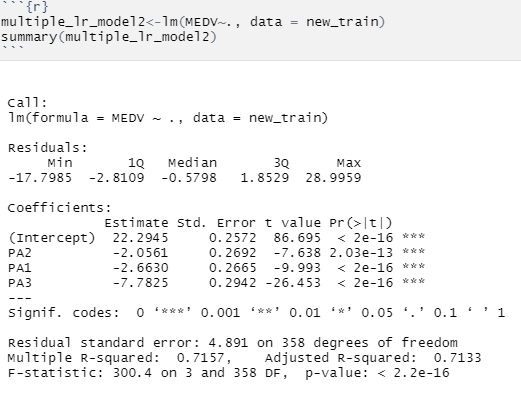


Figure 66: MLR Model 1 after Factor Analysis

Firstly, we will take the vif of this this model to check whether the collinearity has been eliminated or not.

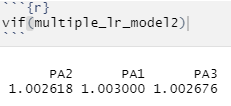


Figure 67: VIF for Model after Factor Analysis

*Here, we can see that the VIF has reduced drastically after the factor analysis.*

**Below are the Prediction results: -**

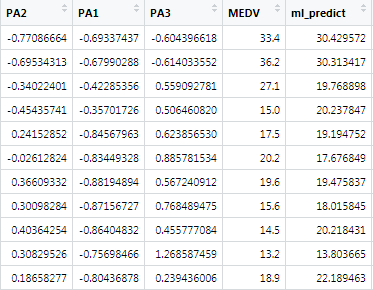
**

Figure 68: Results for MLR Model after Factor Analysis

*Now the accuracy for the model here is 77.56%.*

Here we have also used the decision tree model

**Below is the decision tree model: -**

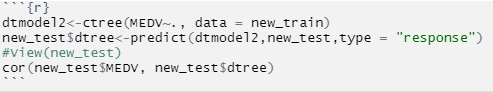
****

Figure 69: Decision Tree Model 2

**Below are the Decision Tree Prediction Results: -**

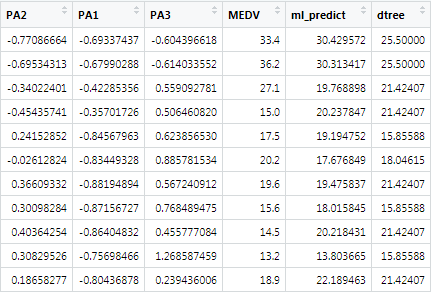


Figure 70: Results for Decision Model 2

The Accuracy for this model is 79.24%.

* **Conclusion:**
  + We have successfully predicted the results using Multiple Regression model and also the Decision Tree Model.
  + We have also done the Factor Analysis and reduced the collinearity between the independent variables.