Table of Contents

[Goal 2](#_Toc482987199)

[Solution 2](#_Toc482987200)

[Exploratory Data Analysis 2](#_Toc482987201)

[Predict using Multiple regression with all variables 4](#_Toc482987202)

[PRINCIPAL COMPONENT ANALYSIS (PCA) 5](#_Toc482987203)

[Factor Analysis 6](#_Toc482987204)

[Predict using Multiple regression with Factors & RAD 7](#_Toc482987205)

[Conclusion 8](#_Toc482987206)

# Goal

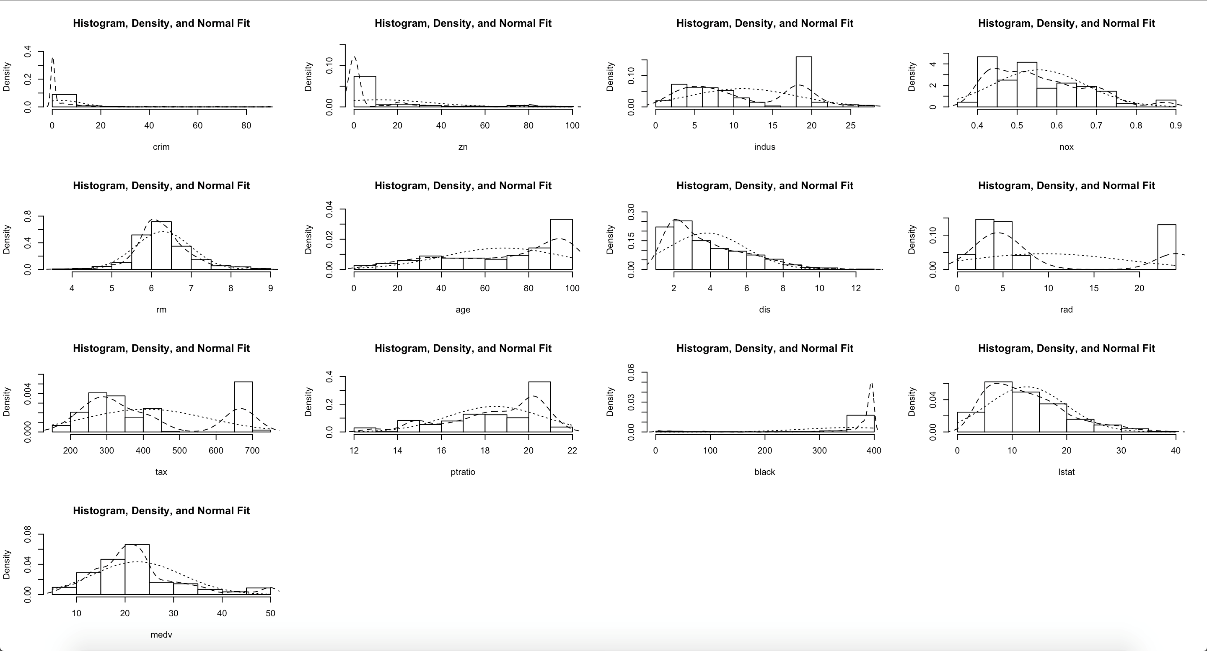
The goal of this assignment is to predict median housing value (medv) using the Boston data set available with the library MASS

# Solution

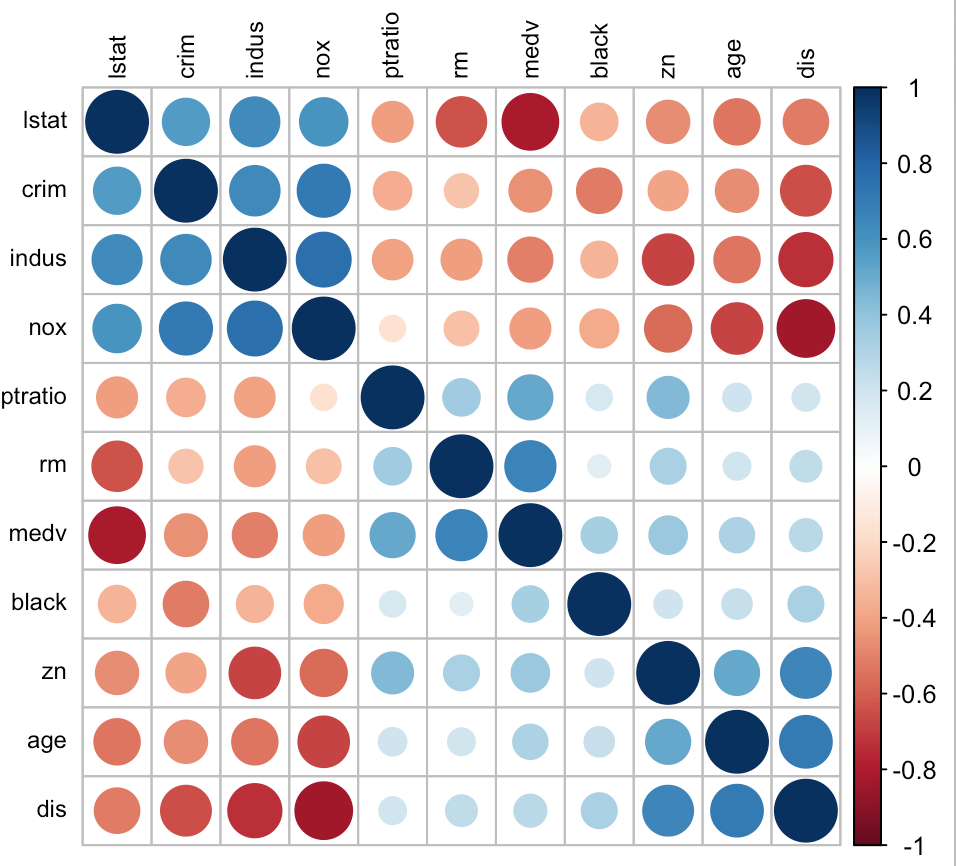
## Exploratory Data Analysis

Important Observations:

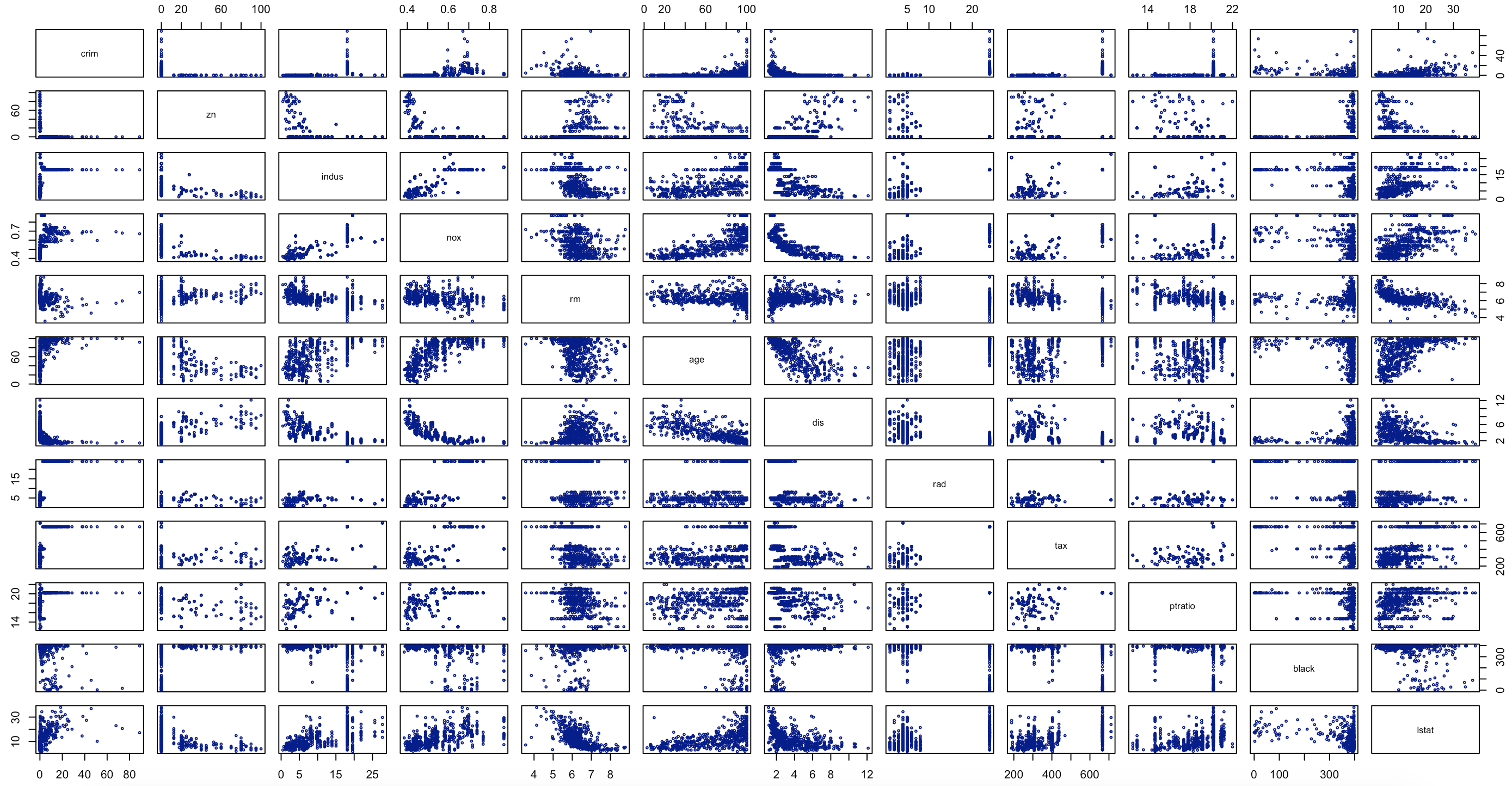
* Variables are not normally distributed. Hence a log transformation would help reduce the skew
  + crim, zn, indus, nox, rm, dis, tax, lstat are positively skewed
  + age, ptratio, black are negatively skewed
  + outliers exist in crim, zn, chas, black



* Variances of the indices are widely different. Hence scaling of variables is necessary for further analysis
* From the correlation plot, it is evident that several variables are strongly correlated. Several pairwise correlations are over 60 – 70%. Hence this data set is a prime candidate for PCA or factor analysis

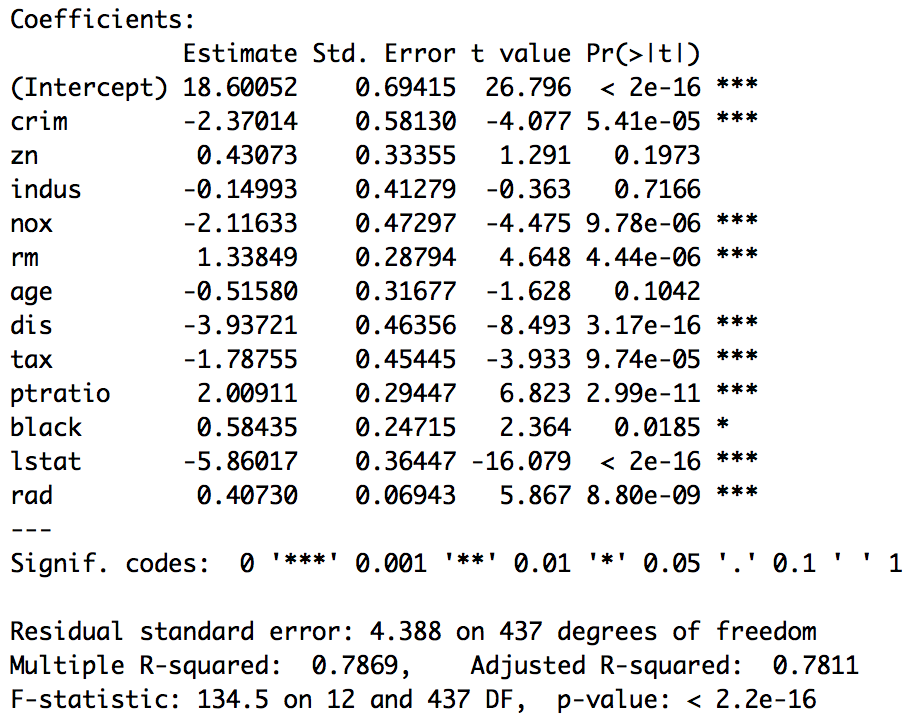


* Scatterplot of the variables also reveals the same story



## Predict using Multiple regression with all variables

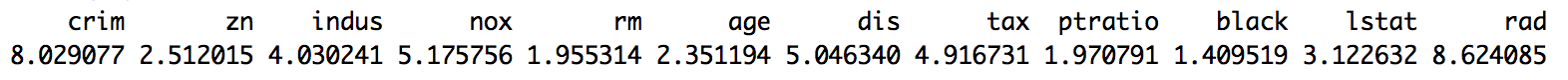
The multiple regression model using all the variables expresses 78% of the variation but some of the variables like zn, indus, age is highly significant should be omitted.



The model is used to predict the mean housing prices for the test data. The root-mean-square-error for the test data is: **3.743696**

However, as the variables are highly correlated we shall either use VIF to eliminate variables one after another or use a dimension reduction technique like PCA or Factor Analysis.

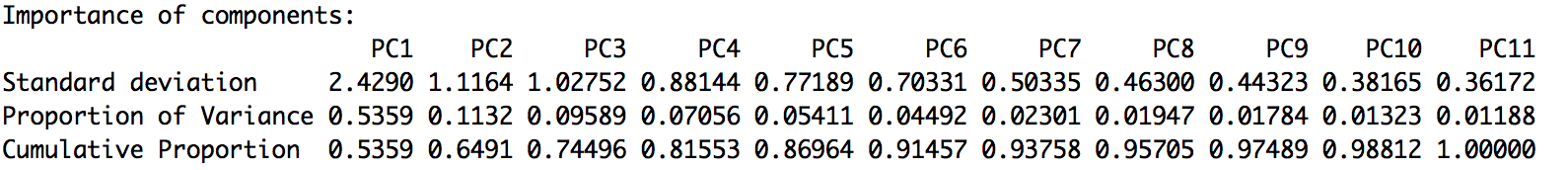
Before we proceed to PCA or Factor Analysis let’s explore the variance inflation factor (VIF). The VIF is > 4 for 6 variables hence strong multi-collinearity.

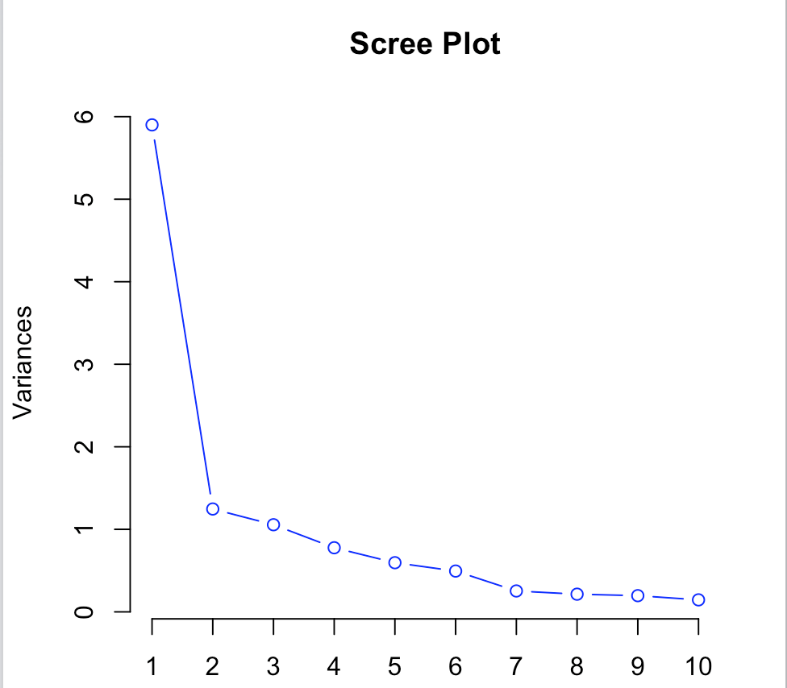


We choose **not to eliminate variables** one after another but choose to **use a dimension reduction technique** like PCA or Factor Analysis to address the affect multi-collinearity.

## PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA (principal component analysis) is done on the training data to determine the number of factors. Summary of the results are shown below.





Important Observations:

* After PC3 from PC4 onwards the eigenvalue falls below 1 but it only explains 74.44%

Of the total variances.

* The first 4 principal components explain almost 82% of the total variances. However, the first 5 and 6 components explain 87% and 91% respectively, even though the eigenvalue is less than 1.

Investigating the correlation between the variables and the first 7 components

Appendix) it is observed that PC1 has high correlation (≥ 0.7) with 17 of the 30 indices, PC2

has high correlation with only one index and PC3 has high correlation with 2 indices. The

other PCs do not have high correlation with any indices at all. This does not render well to

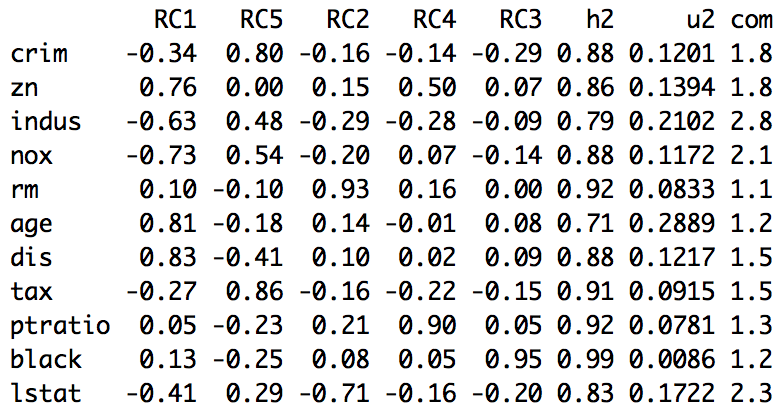
proper interpretation.

As the principal component scores do not explain how the variables can be grouped we use factor analysis as the next stage for our analysis.

We decided to extract 5 factors as they explain 87% of the variation.

## Factor Analysis

We extract 5 factors and observe that with “varimax” rotation the factors are able to better explain and group the variables.



**Factor 1: City Outskirts (RC1)**

- indus: proportion of non-retail business acres per town

- nox: nitrogen oxides concentration (parts per 10 million).

+ zn: proportion of residential land zoned for lots over 25,000 sq.ft.

+ age: proportion of owner-occupied units built prior to 1940

+ dis: weighted mean of distances to five Boston employment centres

**Factor 2: High Alert Zone (RC5)**

crim: per capita crime rate by town

tax: full-value property-tax rate per \$10,000

**Factor 3: Small Dwellings (RC2)**

rm: average number of rooms per dwelling

lstat: lower status of the population (percent)

**Factor 4: High Pupil-Teacher Ratio (RC4)**

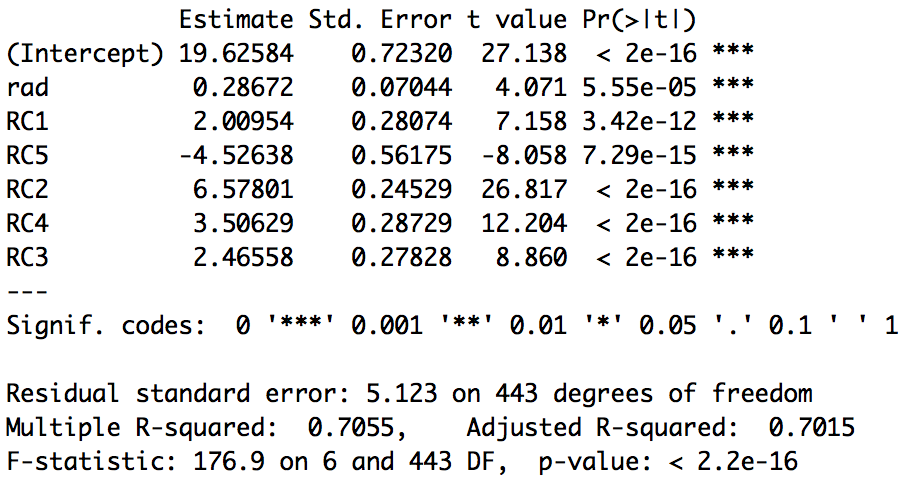
ptratio: pupil-teacher ratio by town

**Factor 5: High Black Population (RC3)**

Black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

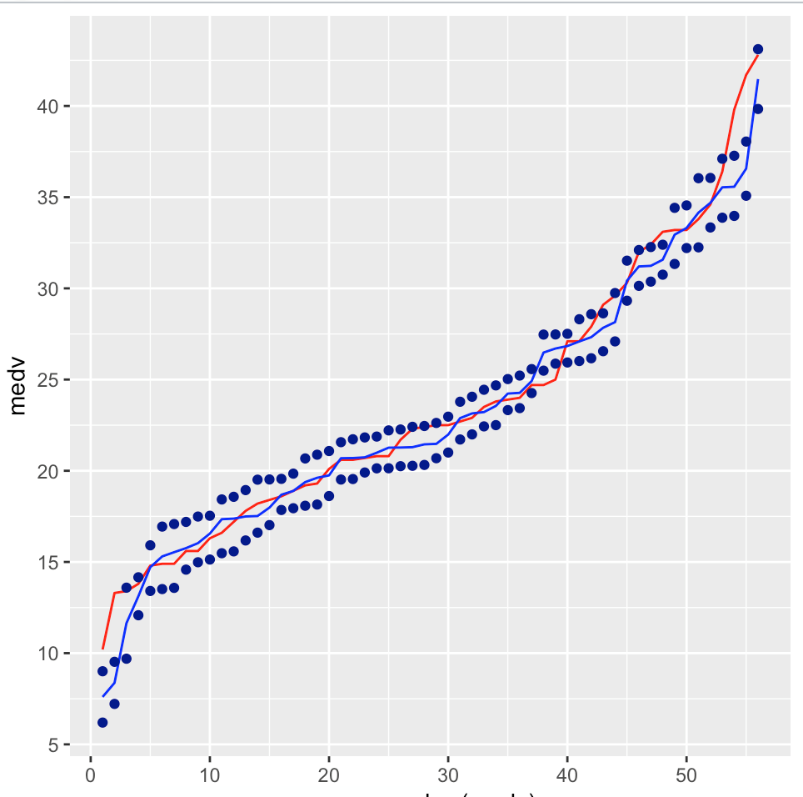
## Predict using Multiple regression with Factors & RAD

The multiple regression model using the factors and rad expresses 70% of the variation. The p-value of each variable is < 0.05 and can be included in the regression model.



The model is used to predict the mean housing prices for the test data. The root-mean-square-error for the test data is: **3.512422**.

The actual values are between the confidence interval of the predicted value. Hence the model is robust.



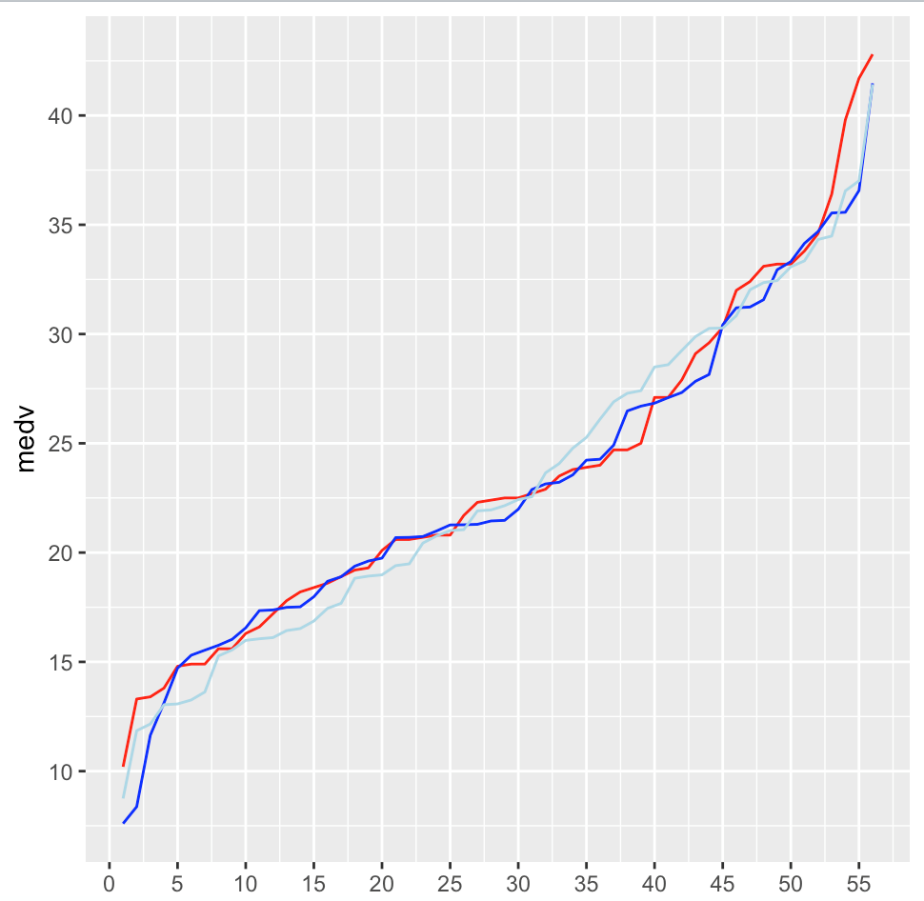
Red – Actual, Blue – Predicted, Dotted Blue – Upper & Lower limit

# Conclusion

The root-mean-square-error for the test data using the multiple regression model with no factors is: **3.743696.** The model explains 78% of the variation but the model is not robust.

The root-mean-square-error for the test data using the multiple regression model with factors is: **3.512422.** The model explains 70% of the variation and it is roubust.

Following is the depiction of the actual value (red), predicted value using the factor (dark blue) and the predicted value without factors (light blue)



The regression model with the below factors best explains the relationship with the mean house prices in Boston and is the robust model. **Hence our recommendation is to use the regression model with factors to predict the mean hour prices.**

* Factor 1: **City Outskirts**
* Factor 2: **High Alert Zone**
* Factor 3: **Small Dwellings**
* Factor 4: **High Pupil-Teacher Ratio**
* Factor 5: **High Black Population**
* Proximity to the highway: **RAD**