EXECUTIVE SUMMARY

The linear regression model is built on the Boston Housing dataset and is used to predict the logarithmic value of the median house price in various suburbs of Boston. The attributes used for this prediction are the following:

1. `crim`: Per capita crime rate in that suburb. This denotes the number of crimes happening to one person in one year on an average. This varied vastly with 0.006 being the least value and 88.9 being the highest. As one can expect increase in crime rate decreases the median value of house price.
2. `chas`: This indicates if a property is adjacent to the Charles River or not. If a house is adjacent to the river, its value is 1 and 0 otherwise. We expect this to be a positively correlated with the house price and the same be seen from the model.
3. `rm`: This denotes the average number of rooms per dwelling. The least is 3.5 and its maximum value is 8.8. As the average number of rooms increase, the median house price also increases as expected.
4. `nox`: Nitrogen Oxides concentration in the units of parts per 10 million is given by this attribute. This has a minimum of 0.38 parts per 10 million and a highest amount of 0.87 parts per 10 million
5. `age`: This gives proportion of owner-occupied units built prior to 1940. More on this is discussed below
6. `dis`: This is the weighted mean of distances to five Boston employment centers. Houses closer to the cites have high price and that is also seen in the model
7. `ptratio`: This is the pupil-teacher ratio by town where least value is 12.6 and highest is 22. Lower the number better the quality of education expected
8. `lstat`: Denotes percentage of population with lower socio-economic status. Least is 1.73 and highest is 37.97. Usually, areas with lower `lstat` have high prices. And that is seen from the correlation plots also

These attributes together explain 80% of the variance in data and the predicted result is off by $1200 on average when the mean house price is around $22,500 for the whole lot. Among the different attributes mentioned above, those with the highest significance are per capita crime rate, percentage of lower status of the population, average number of rooms and pupil-teacher ratio. Surprisingly, age of the house does not seriously affect the price of the house and is not as significant as other factors. Older houses may not be appealing to some but for aesthetic reason, it might interest a section of people who look for the historical charm. And for this mixed reason, it does not show a strong relation with price.

A good baseline scenario would be a house with `ptratio` less than 20, `dis` less than 3.2 miles in average and `nox` less than 0.5 parts per 10 million, the price range for this scenario is $26,500 - $40,000

We can also notice the attributes that dictate the houses with highest price. It is observed that houses with the top 5% (around $48,000 on average) price are mostly with the following attribute characteristics (Appendix A):

a. Very low crime rate.
b. More number of average rooms in the dwelling
c. Less distance from employment centers
d. Closer to radial highways
e. Lesser property tax value
f. Lower value of people with lower status (suburbs with high standards of living)

As for the tradeoffs (Appendix B), we can get houses in great locations for a price less than the average (which is $22,500) if we are willing to compromise in some factors. These can look like:

a. If you are willing to stay a bit far from the major employment centers, you can get house for less even though there is very low crime rate, and low Nox values. Or you can get a house for low price close to cities if you can settle for less rooms.
b. If you can compromise on the pupil-teacher ratio, you can get houses close to the employment centers and close to radial highways at low prices.

TECHNICAL SUMMARY

The Boston Dataset that we used has 506 observations and 14 different attributes among which is our response variable, median value of house price in a suburb (`medv`), that needs to be predicted. The rest 13 are independent predictors. There are no null or missing values in the dataset. I have used a linear regression method to estimate the logarithmic values of median house prices.

Exploratory Data Analysis:

After preprocessing the data, we performed Exploratory data analysis to examine the distribution of all the variables and look for possible outliers and leverage points. Upon initial examination, a few points that are considered are possibility of outliers in various attributes like `crim, black, zn, rm`. These can either be misrepresented data or the actual behavior of real-world data.

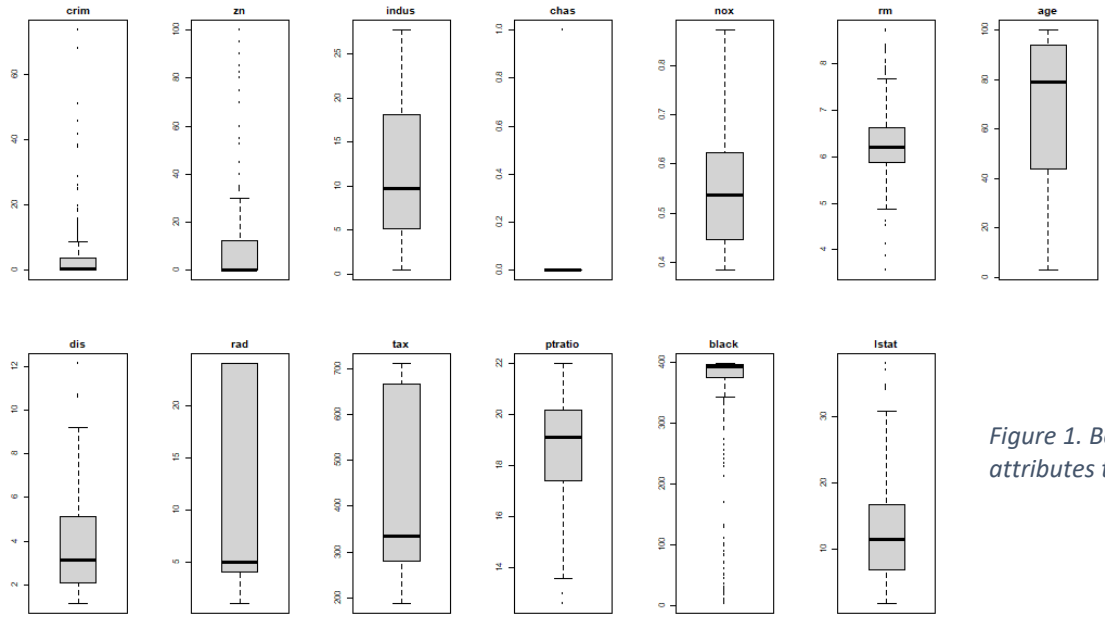


Figure 1. Box plots of the attributes to visualize the outliers

I also checked for the correlations between the predictors and the response variable to check if there are any obvious relations between them. One important observation from this is that `rad` and `tax` are highly correlated and VIF (variance inflation factor) values later also indicate the existence of high collinearity.

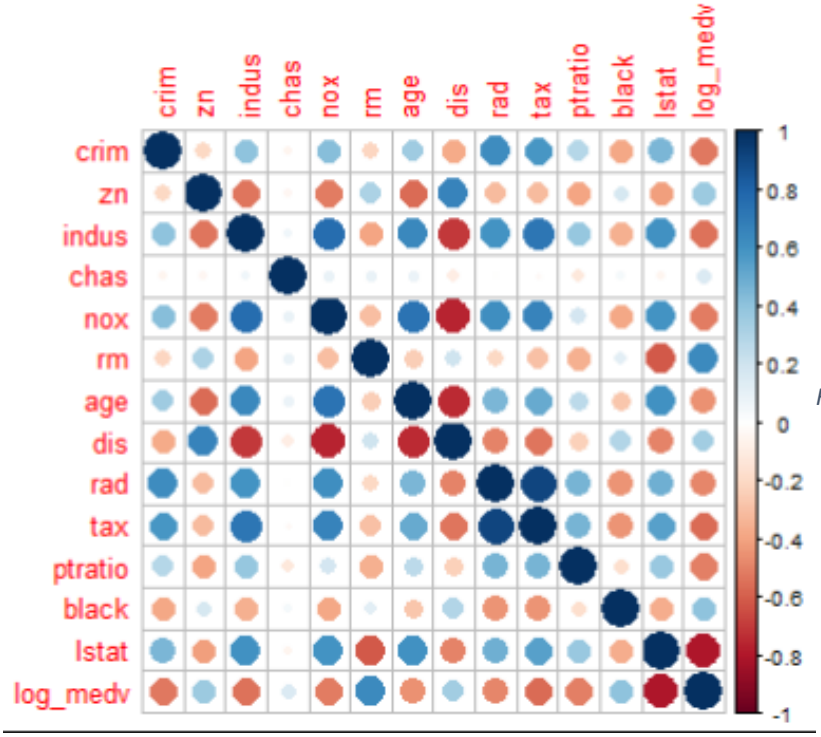


Figure 2. Correlation plot

Collinearity among predictors:

After making the preliminary observations, I proceeded to split the dataset into training and testing datasets. I have used the 80:20 split to leave aside 20% of the data hidden away from the model and use only the 80% of the data in training it. This gives an accurate picture of performance of the model on unseen real-world data. This leaves the training data with 405 observations and testing data 101 observations. Then we fit a linear regression model for all the attributes to check the significance of each predictor on the response variable. From the summary of the model, we have found out that `indus`, `age` and `zn` have very less significance since their `Pr(>|t|)` value is very high indicating the lack of confidence on them.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.1095470  0.2358067  17.428  < 2e-16 ***
crim        -0.0116291  0.0016362  -7.107 5.65e-12 ***
zn           0.0011302  0.0006127   1.845  0.06585 .
indus        0.0019209  0.0027095   0.709  0.47879
chas         0.1098571  0.0379356   2.896  0.00399 **
nox         -0.7036637  0.1723929  -4.082 5.42e-05 ***
rm           0.0839338  0.0190521   4.405 1.36e-05 ***
age         -0.0001242  0.0005888  -0.211  0.83311
dis         -0.0511733  0.0091306  -5.605 3.95e-08 ***
rad          0.0164543  0.0028852   5.703 2.32e-08 ***
tax         -0.0007018  0.0001624  -4.322 1.96e-05 ***
ptratio     -0.0351690  0.0059990  -5.862 9.71e-09 ***
black        0.0004190  0.0001199   3.494  0.00053 ***
lstat       -0.0295088  0.0022370 -13.191  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1921 on 391 degrees of freedom
Multiple R-squared:  0.791,     Adjusted R-squared:  0.784
F-statistic: 113.8 on 13 and 391 DF,  p-value: < 2.2e-16
```

Figure 3. Model summary of initial fit

The table at eh side shows the coefficients and `Pr(>|t|)` values along with the `R-squared` values. Then upon inspecting the VIF (Variance inflation factor) values, we can see that `rad` and `tax` have values greater than 5 indicating strong collinearity. To tackle this collinearity, I have introduced an interaction term of `rad:tax`. This combines the effect of both the predictors into one cohesive predictor without any information loss. Doing this also increases the `R-squared` value which further validates our conclusion.

Nonlinearity of Residuals:

After tacking the collinearity in the model and fitting the model once again, we can observe from the residual vs fitted plot that the residuals do not follow the linearity assumption.

To deal with this nonlinearity in our linear regression, we have introduced a higher order polynomial into the model. To do that we once again turn to our scatter plots which in the interest of space have been excluded from this report. Upon inspecting the plots, I identified `lstat` and `rm` are good candidates to introduce non linearity because of their relationship with `log_medv`. But the introduction of rm^2 only reduced the `R-squared` value where as introducing $lstat^2$ increased the R-squared value from 0.78 to 0.80. Now after adding this $lstat^2$ term our residuals show linearity to great extent.

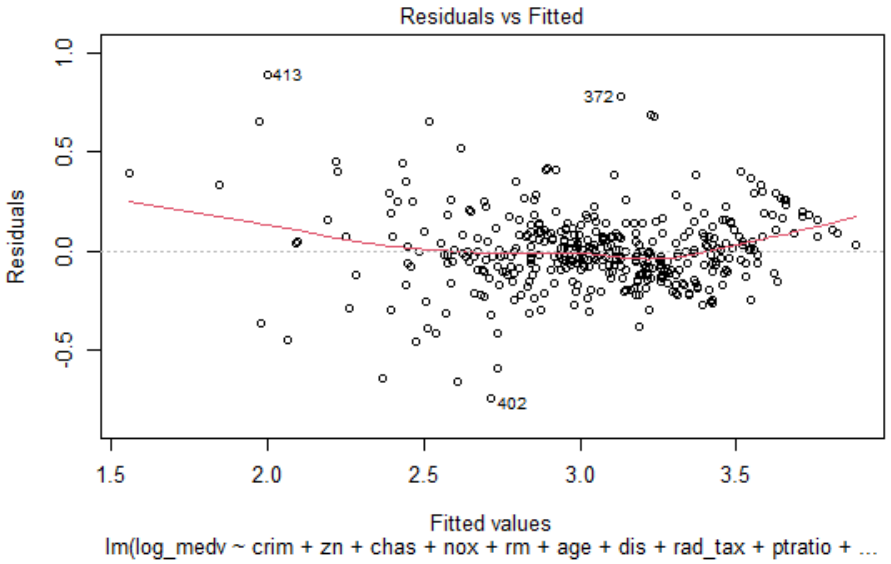


Figure 4. Non-linearity in the residuals

Outliers and leverage points:

Now it is time to deal with the outliers in the hope of increasing the fit. From the preliminary analysis using boxplots, we can see a lot of values falling outside the 95-percentile line. We can also see a huge variation in median and mean values of `crim, zn, black` which indicate a possibility of a few extreme values. I have

decided to delete these outliers which are outside 99-percentile to start. Upon cleaning the data with this constraint, the effect it has on R-squared and `rmse` (root mean square error) on test data is different than expected. Removing these "outliers" only increased the `rmse` value. This brought me to a conclusion that these extreme values are not outliers but the natural order of real-world data. So we are not cleaning this data.

Final Model parameters:

The final model parameters that I have chosen are `crim, zn, chas, nox, rm, age, dis, rad:tax, ptratio, black, lstat, lstat`^{2}. The following is the summary of my final model.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.198e+00  2.288e-01  18.350  < 2e-16 ***
crim        -1.331e-02  1.601e-03  -8.312 1.56e-15 ***
zn          -4.911e-05  5.882e-04  -0.083 0.933500
chas         1.243e-01  3.613e-02   3.440 0.000645 ***
nox         -7.307e-01  1.579e-01  -4.627 5.05e-06 ***
rm           7.085e-02  1.828e-02   3.875 0.000125 ***
age          7.008e-04  5.805e-04   1.207 0.228090
dis         -4.021e-02  8.625e-03  -4.662 4.30e-06 ***
ptratio     -3.170e-02  5.800e-03  -5.465 8.25e-08 ***
black        3.669e-04  1.157e-04   3.172 0.001633 **
lstat       -7.042e-02  5.894e-03 -11.947  < 2e-16 ***
I(lstat^2)   1.153e-03  1.543e-04   7.469 5.30e-13 ***
rad:tax      8.915e-06  2.445e-06   3.646 0.000303 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1848 on 392 degrees of freedom
Multiple R-squared:  0.8061,    Adjusted R-squared:  0.8001
F-statistic: 135.8 on 12 and 392 DF,  p-value: < 2.2e-16

root mean square (rmse):  0.1997425
```

Figure 5. Model summary of the final model along with its rmse

Froom the summary, we can clearly see that `zn` and `age` does not show significant relationship to the response variable that is `log_medv`. The rest all predictors display satisfactory levels of significance. The predictors `crim, zn, nox, dis, ptratio, lstat` all show negative correlation with the price of the houses. Which is expected because increase in crime rate, pollution, distance from major cities always decrease the cost of the property. Similarly, the predictors `chas, rm` show positive correlation with the response variable. Which means that increasing the number of rooms or having the house adjacent to Charles River can increase the price, which is exactly what one might expect.

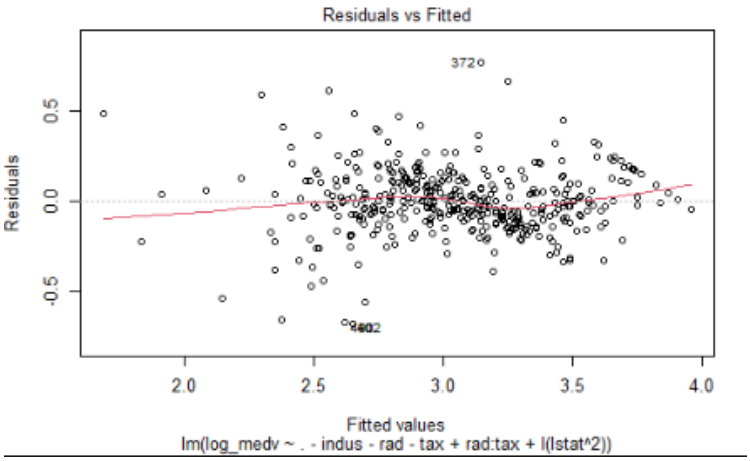


Figure 6. Residuals vs Fitted plot of final model

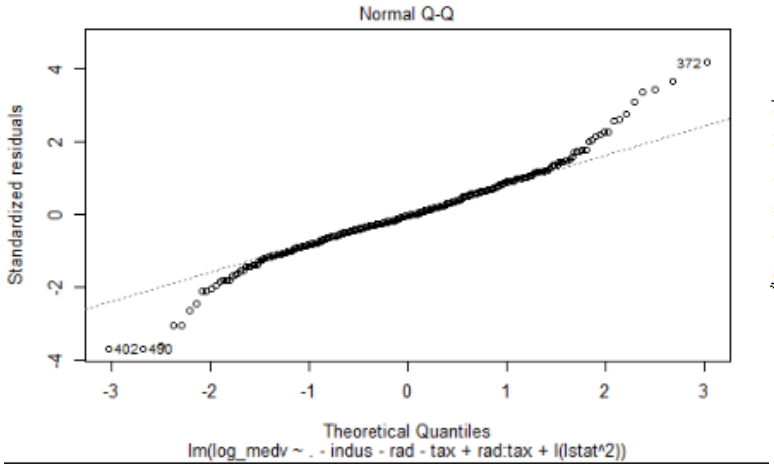


Figure 7. Normal Q-Q plot of the final model

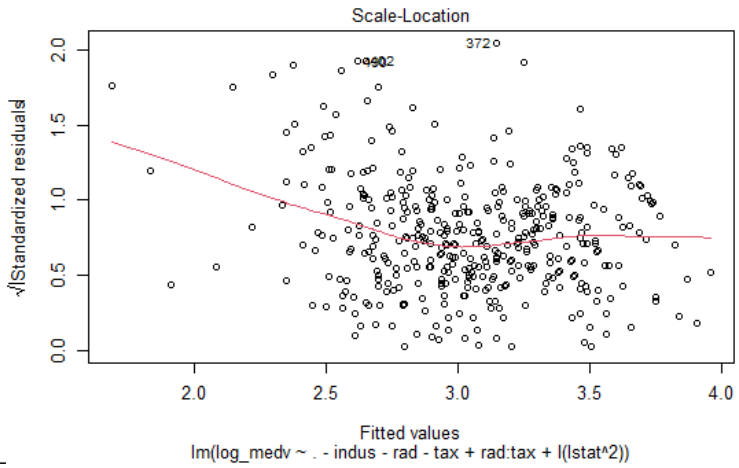


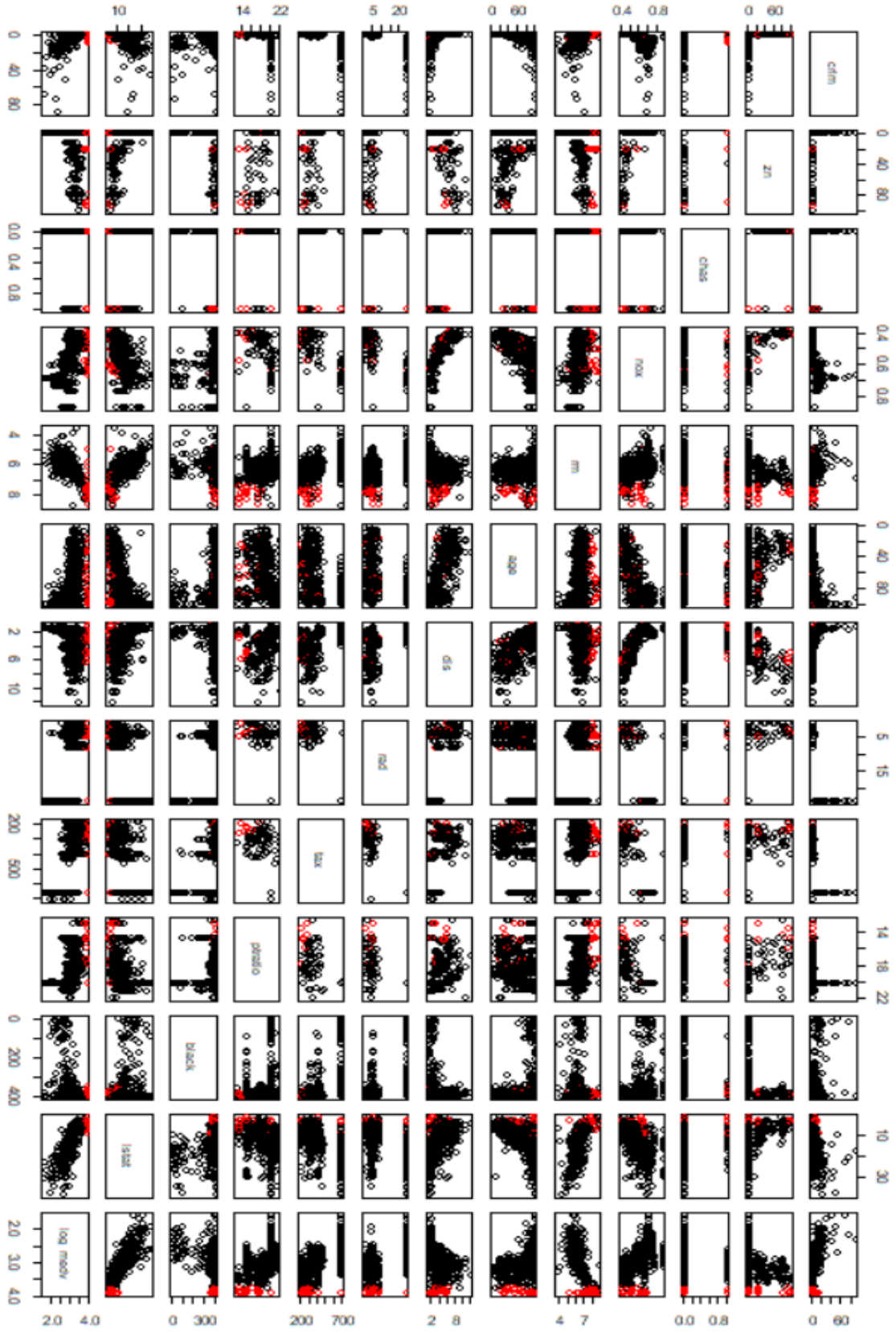
Figure 6. Scale-Location plot of the final model

The above diagnostic plots characterize the final model. The residuals vs fitted plot shows linearity of the residuals which is an indication that the model can explain the information contained in the predictors sufficiently. The Normal Q-Q plot show the gaussian nature of the residuals except for the extreme values. The scale-location plot shows that the homoscedasticity condition of the residuals is met sufficiently as we can see constant variance among the residuals.

We can also see from the summary that the Adjusted R-squared value is 0.8001 which means this particular model can predict 80.01% percent of the variability in the data, which is not that bad. The training standard error is reported to be 0.1848 where as we get a testing standard error of 0.199 i.e., the standard deviation of the residuals is 0.199 which is less than $1200. Considering the average house price is $22,500 a standard deviation of $1200 is not that bad.

APPENDIX

A) This plot shows the correlation between all the predictors and response variable with the red points being the top 5 percent of the houses.



B) This plot shows the correlation between all the predictors and response variable with the red points being the houses below mean price and blue being the houses above the mean