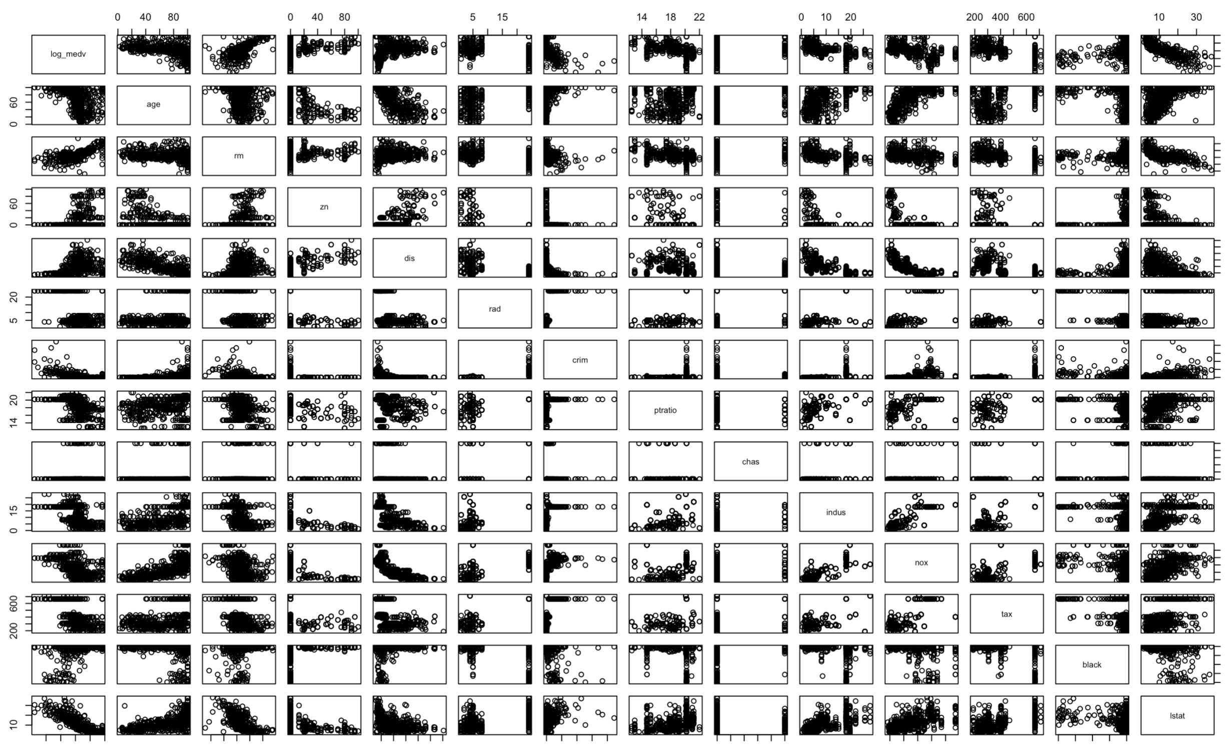
Technical Summary

Data: The Boston data set contains information from 504 geographic areas and 14 attributes for each area. The attributes include:

Attribute	Description
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	nitrogen oxides concentration (parts per 10 million).
rm	average number of rooms per dwelling.
age	proportion of owner-occupied units built prior to 1940.
dis	weighted mean of 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 blacks by town.
lstat	lower status of the population (percent).
medv	median value of owner-occupied homes in $1000s.



The above shown plot is the pairs plot for each attribute with every other one including log(medv).

Hypothesis Testing: A multiple linear regression model is built with all the attributes included and got the t-statistic values and p-values as followed:

```
> bosmodel = lm(log(medv)~., data = Boston) ; summary(bosmodel)

Call:
lm(formula = log(medv) ~ ., data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.73361 -0.09747 -0.01657  0.09629  0.86435 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.1020423  0.2042726  20.081  < 2e-16 ***
crim        -0.0102715  0.0013155  -7.808 3.52e-14 ***
zn           0.0011725  0.0005495   2.134 0.033349 *  
indus        0.0024668  0.0024614   1.002 0.316755    
chas         0.1008876  0.0344859   2.925 0.003598 ** 
nox         -0.7783993  0.1528902  -5.091 5.07e-07 ***
rm           0.0908331  0.0167280   5.430 8.87e-08 ***
age          0.0002106  0.0005287   0.398 0.690567    
dis         -0.0490873  0.0079834  -6.149 1.62e-09 ***
rad          0.0142673  0.0026556   5.373 1.20e-07 ***
tax         -0.0006258  0.0001505  -4.157 3.80e-05 ***
ptratio     -0.0382715  0.0052365  -7.309 1.10e-12 ***
black        0.0004136  0.0001075   3.847 0.000135 ***
lstat       -0.0290355  0.0020299 -14.304  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1899 on 492 degrees of freedom
Multiple R-squared:  0.7896,	Adjusted R-squared:  0.7841 
F-statistic: 142.1 on 13 and 492 DF,  p-value: < 2.2e-16
> vif(bosmodel)
    crim       zn    indus     chas      nox       rm      age 
1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 
     dis      rad      tax  ptratio    black    lstat 
3.955945 7.484496 9.008554 1.799084 1.348521 2.941491 
```

The MLR fit shows an F-statistic value of 142.1 and p-value of 2.2e-16 which very less and indicates to reject the null-hypothesis and accept the alternate hypothesis that there is some relationship between at least one predictor and the response variable.

Though the p-value for the complete model turned out to be less, the individual p-values for predictors such as "indus", "age" showing that they are statistically insignificant.

Adding, the VIF values shown below are high (>5) for "tax" and "rad" showing that these two have high collinearity. Hence, "tax", "rad", "indus", "age" are removed for the next iteration which showed that the "zn's" p-value increased beyond threshold value to be statistically significant. So removed "zn" too.

The **final MLR model** is shown below, which is used to deduce insights:

```
Call:
lm(formula = log(medv) ~ . - age - indus - tax - zn, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.73252 -0.10612 -0.01410  0.09214  0.87773 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.0213020  0.2051665  19.600  < 2e-16 ***
crim        -0.0100087  0.0013294  -7.529 2.44e-13 ***
chas         0.1161515  0.0346669   3.351 0.000868 ***
nox         -0.8712566  0.1395967  -6.241 9.32e-10 ***
rm           0.1014707  0.0162931   6.228 1.01e-09 ***
dis         -0.0442353  0.0065520  -6.751 4.10e-11 ***
rad          0.0056058  0.0016295   3.440 0.000630 ***
ptratio     -0.0426888  0.0049175  -8.681  < 2e-16 ***
black        0.0004319  0.0001088   3.970 8.26e-05 ***
lstat       -0.0288434  0.0019305 -14.941  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1929 on 496 degrees of freedom
Multiple R-squared:  0.7812,	Adjusted R-squared:  0.7773 
F-statistic: 196.8 on 9 and 496 DF,  p-value: < 2.2e-16
```

This final MLR model shown beside has an F-statistic of 196.8 and p-value of 2.2e-16 which indicates to reject the null hypothesis. The individual p-values are also very less to prove that all those are statistically significant. The R^2 value is 0.7812 which is close to 1 and proves that 78% of the variability in the housing price is explained by the predictor variables. RSE is 0.1929 increased from the initial model but not that significant.

The R^2 can further be improved by including interaction terms and non-linear terms but that is not pursued in this project.

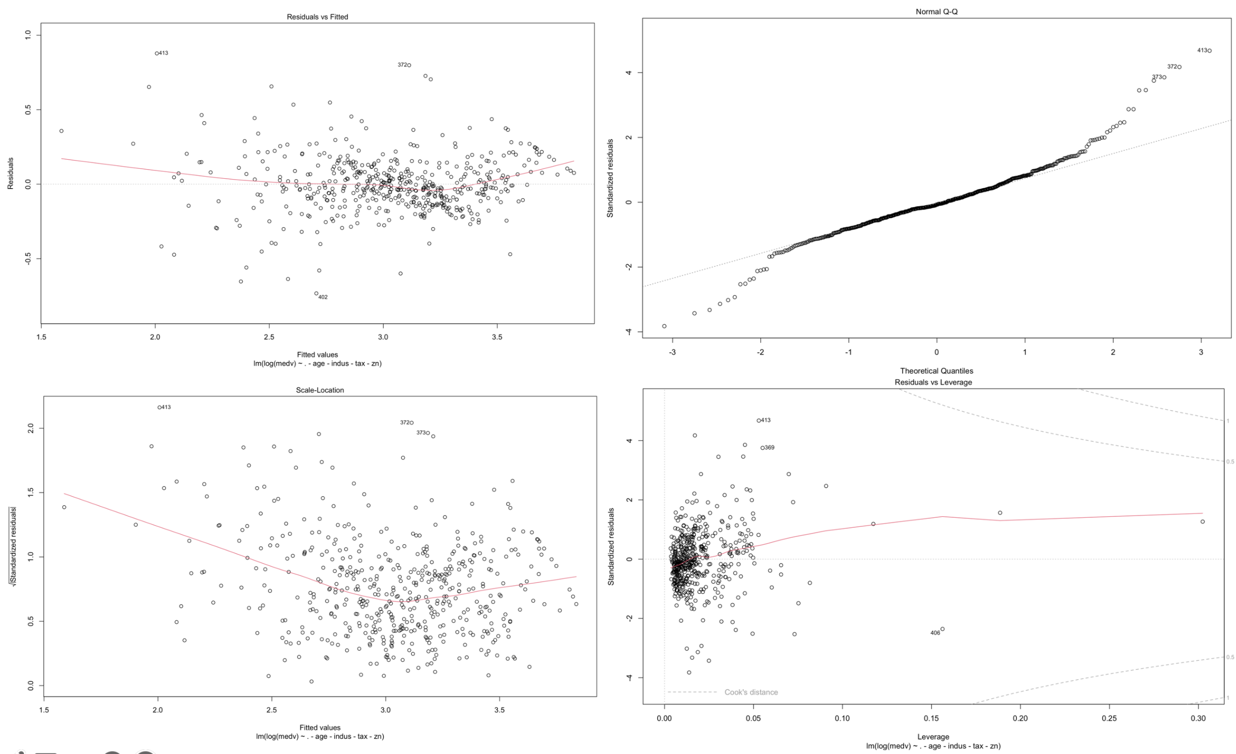
RSE is 0.1929 for the log(medv) which is equivalent to RSE of $1,212, whereas the mean value of the $22,533. The % error is 5.38%.

Confidence Intervals of predictors: 95% confidence intervals are calculated and shown below.

```
> confint(bosmodelfin3)
                    2.5 %       97.5 %
(Intercept)  3.6181994704  4.424404627
crim        -0.0126206205 -0.007396778
chas         0.0480395103  0.184263532
nox         -1.1455302967 -0.596982892
rm           0.0694587687  0.133482632
dis         -0.0571082586 -0.031362256
rad          0.0024042082  0.008807308
ptratio     -0.0523505199 -0.033027166
black        0.0002181651  0.000645729
lstat       -0.0326362734 -0.025050488
```

All the predictors confidence interval limits are narrow and far from zero further proving the significance of each predictor.

Diagnostic Plots: The below shown four plots are the diagnostic plots for the MLR model. Top-Left: Residuals vs Fitted, Top-Right: Normal Q-Q, Bottom-Left: Scale-Location, Bottom-Right: Standardized Residuals vs Leverage Statistic.



Residuals vs Fitted Plot: This plot is between the Residuals and Fitted Values; smooth fit of the residuals shows some non-linearity in the model because the line is not linear. This can be improved by introducing interaction terms or non-linear predictor terms.

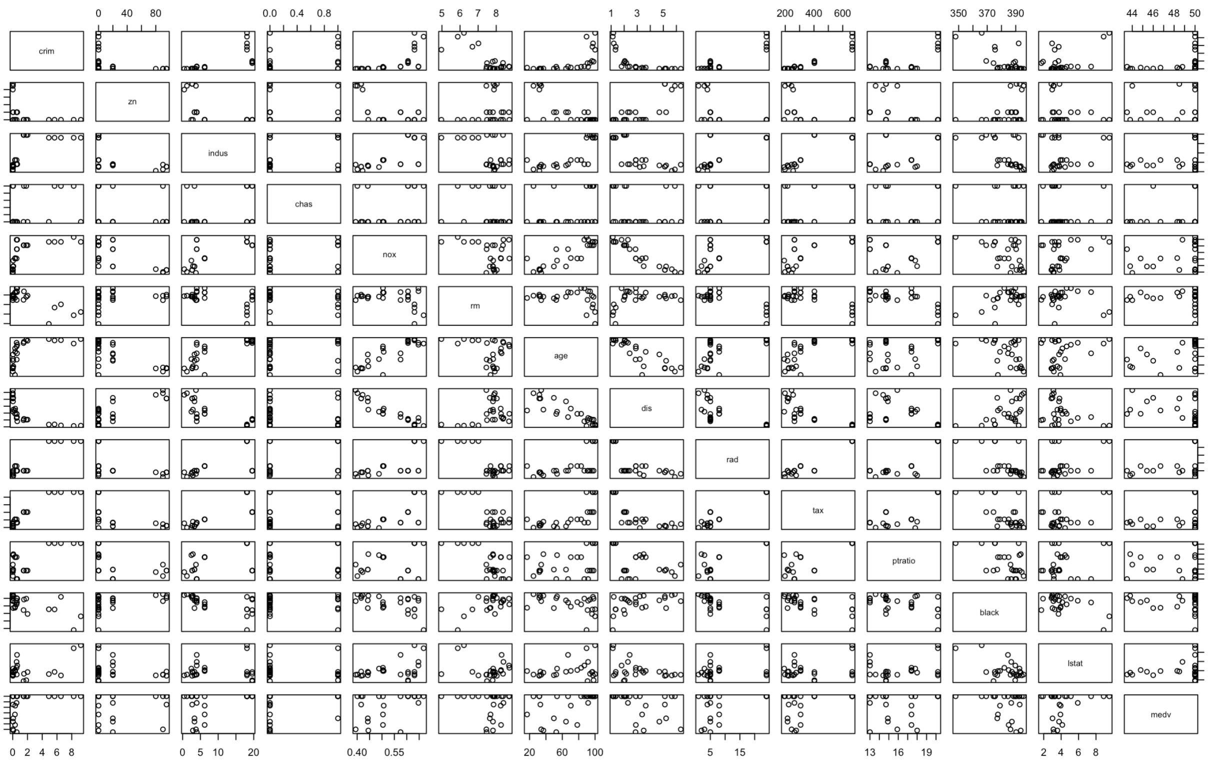
The data shown in the same plot is scattered across and is not showing any pattern like conical. So, this model between the log(medv) and predictors is not showing any homoscedasticity. If a model is built with just the medv as the response, it might have shown that problem.

Normal Q-Q Plot: This plot is between Standardized Residuals and Theoretical Quantiles and is used to show any possible outliers in the data. The safe region is between -2 and +2 and mostly all the data is concentrated between this region only and we can assume the residuals are normally distributed. There are some points outside this region, but they are following the trend.

Scale-Location Plot: This plot is between Sqrt(Standardized Residuals) and Fitted Values. The smooth fit Red Line isn't exactly horizontal across the plot, but it doesn't deviate too wildly at any point. The assumption of equal variance is not likely violated in this case.

Residuals vs Leverage: This is plotted between Standardized Residuals and Leverage Statistic. One point in this plot is going close to the cook's distance, but it doesn't fall outside of the dashed line. This means that there aren't any high leverage points in the dataset.

Top 5% Pair Plots:



The above top 5 percentile plot explains some important insights about the characteristics that will impact the prices. In the top 5 percentile, there is no evidence that the proximity to the Charles River is not that important of a factor as we observed from the complete model. So, the important attributes that increase the value of real estates are the increase in the average number of rooms and decrease in the crime rate and the reduction of nox levels.

Conclusion: The final MLR model is:

log(medv) = 4.0213020 - 0.0100087crim + 0.1161515chas - 0.8712566nox + 0.1014707rm - 0.0442353dis + 0.0056058rad - 0.0426888ptratio + 0.0004319black - 0.0288434*lstat

This is final MLR model fit for the Boston dataset which is used to linearly predict the housing prices with 5.38% RSE. This can be further explored and improved by including interaction terms between the predictor variables or non-linear terms.