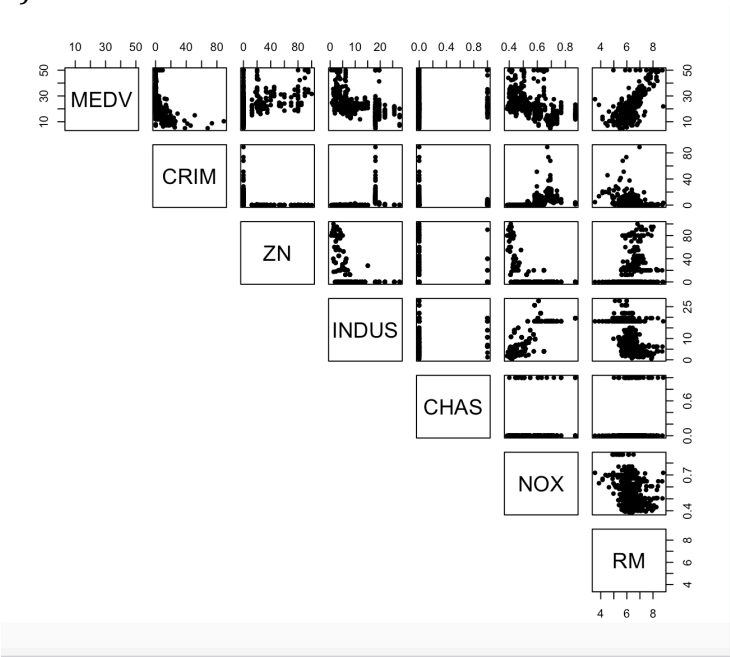
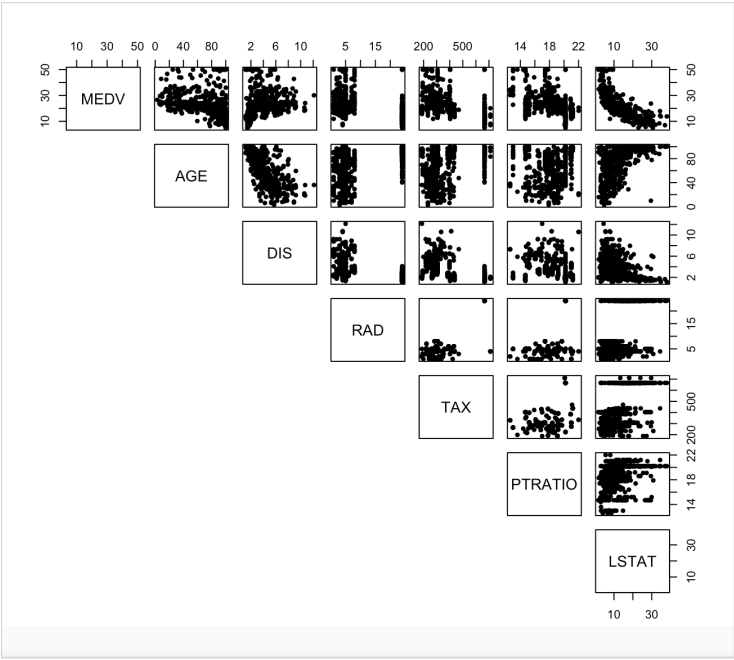
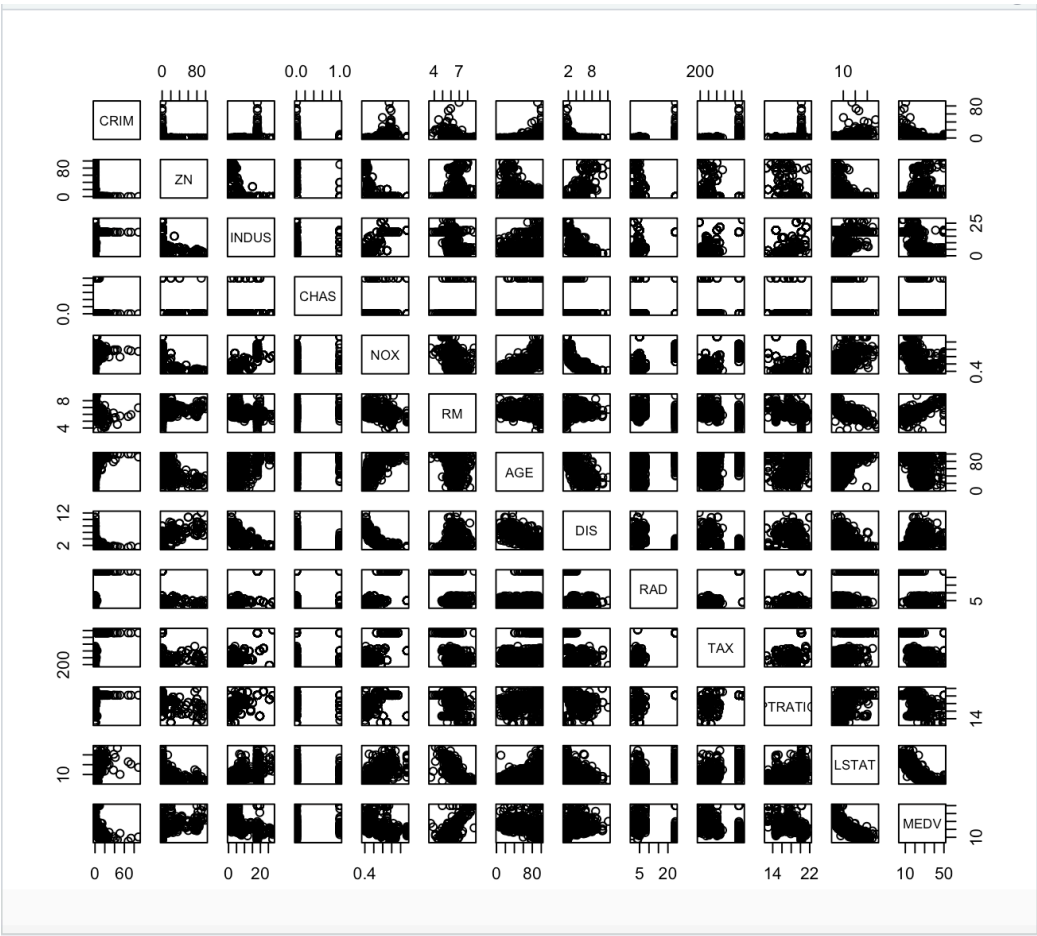
Karishma Parekh
V# 00971328

In this project we took MEDV as a dependent variable and other 12 are independent variables. It has 0.4493 R-squared and p-value is 2.2×10^{-16}.

1)







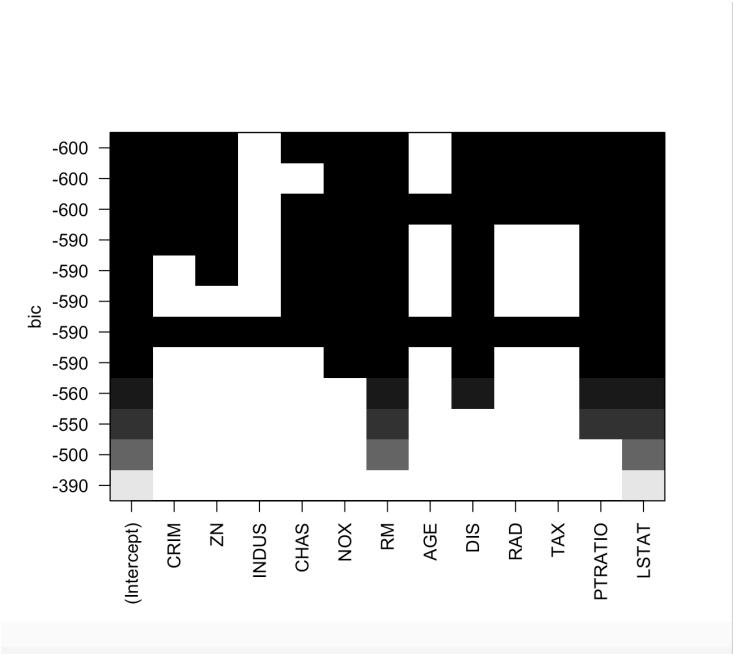
We may conclude that RM, PTRATIO, and LSTAT are the most important in estimating MEDV. From the plot, we can also see multicollinearity amongst dependent variables. Over-fitting can occur when two variables are highly connected; hence this type of variable should be removed.
From the plot we can conclude that MEDV and LSTAT are negatively correlated, because LSTAT is measures the percentage of lower status. Then, looking at the graph matrix, we can observe that INDUS, NOX, and TAX appear to be positively correlated in the bottom triangle, on the left of the graphs. That would imply that these three measurements are all measuring the same thing.
The highest positive correlations are between RAD and TAX, INDUS and NOX. And negative between DIS and AGE, DIS and NOX.

2) DIS and RAD seems to have the most serious variance inflation. RAD has 9.195493 and DIS has 7.029796. So, we removed that.

3) From Best subset regression and stepwise selection (forward, backward, both), we see that all variables except INDUS and AGE are significant.

```
>Subset selection object
>Call: regsubsets.formula(MEDV ~ ., data = BostonHousing, nbest = 1,
   nvmax = 13)
12 Variables  (and intercept)
      Forced in Forced out
CRIM       FALSE     FALSE
ZN         FALSE     FALSE
INDUS      FALSE     FALSE
CHAS       FALSE     FALSE
NOX        FALSE     FALSE
RM         FALSE     FALSE
AGE        FALSE     FALSE
DIS        FALSE     FALSE
RAD        FALSE     FALSE
TAX        FALSE     FALSE
PTRATIO    FALSE     FALSE
LSTAT      FALSE     FALSE
1 subsets of each size up to 12
Selection Algorithm: exhaustive
         CRIM ZN  INDUS CHAS NOX RM  AGE DIS RAD TAX
1  ( 1 ) " "  " " " " "  " "  " " " " " " " " " " " "
2  ( 1 ) " "  " " " " "  " "  " " "*" " " " " " " " "
3  ( 1 ) " "  " " " " "  " "  " " "*" " " " " " " " "
4  ( 1 ) " "  " " " " "  " "  " " "*" " " "*" " " " "
5  ( 1 ) " "  " " " " "  " "  "*" "*" " " "*" " " " "
6  ( 1 ) " "  " " " " "  "*"  "*" "*" " " "*" " " " "
7  ( 1 ) " "  "*" " " "  "*"  "*" "*" " " "*" " " " "
8  ( 1 ) "*"  "*" " " "  "*"  "*" "*" " " "*" " " " "
9  ( 1 ) "*"  "*" " " "  " "  "*" "*" " " "*" "*" "*"
10  ( 1 ) "*"  "*" " " "  "*"  "*" "*" " " "*" "*" "*"
11  ( 1 ) "*"  "*" " " "  "*"  "*" "*" "*" "*" "*" "*"
12  ( 1 ) "*"  "*" "*"  "*"  "*" "*" "*" "*" "*" "*"
         PTRATIO LSTAT
1  ( 1 ) " "    "*"
2  ( 1 ) " "    "*"
3  ( 1 ) "*"    "*"
4  ( 1 ) "*"    "*"
5  ( 1 ) "*"    "*"
6  ( 1 ) "*"    "*"
7  ( 1 ) "*"    "*"
8  ( 1 ) "*"    "*"
9  ( 1 ) "*"    "*"
10  ( 1 ) "*"    "*"
11  ( 1 ) "*"    "*"
12  ( 1 ) "*"    "*"
```

Karishma Parekh
V# 00971328



4) The new model exclude RAD and DIS is much better. The fitted model is MEDV = 0.07711 CRIM + 0.02334 ZN + (- 1.06130) INDUS + 11.55175 CHAS + 1.74100 NOX + (- 0.02349) RM + (- 1.12131) AGE + 0.61061 TAX + 0.28627 PTRATIO + (- 0.17636) LSTAT.

The null hypothesis H_0: The coefficients associated with the variables are zero.

The alternate hypothesis H_1: The coefficients are not.

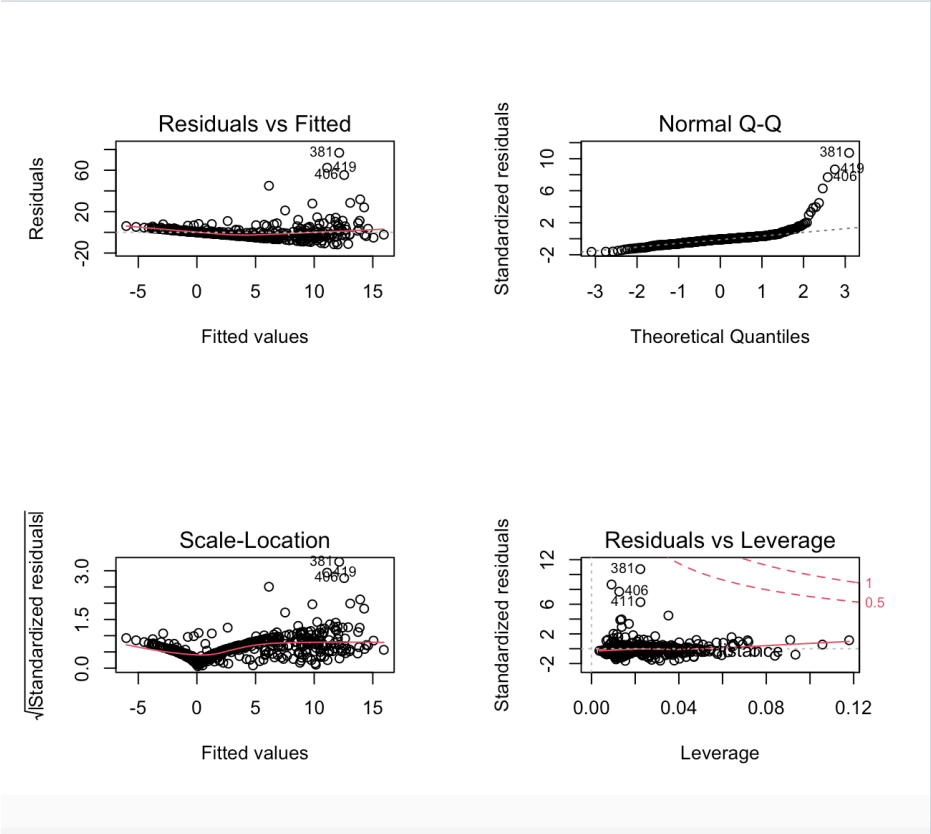
P-value has 3 stars that mean coefficient is very high significance.

R- squared is 0.3049.

F-statistic value is 21.71.

Karishma Parekh
V# 00971328

5)



- The variance is not completely constant; therefore the assumption of constant variance is not fully satisfied.
- The q-q plot shows that the data is not completely normal and skewed to the right.
- There is no autocorrelation in the model.
- There are no outliers seen