Mithun Muralidhar

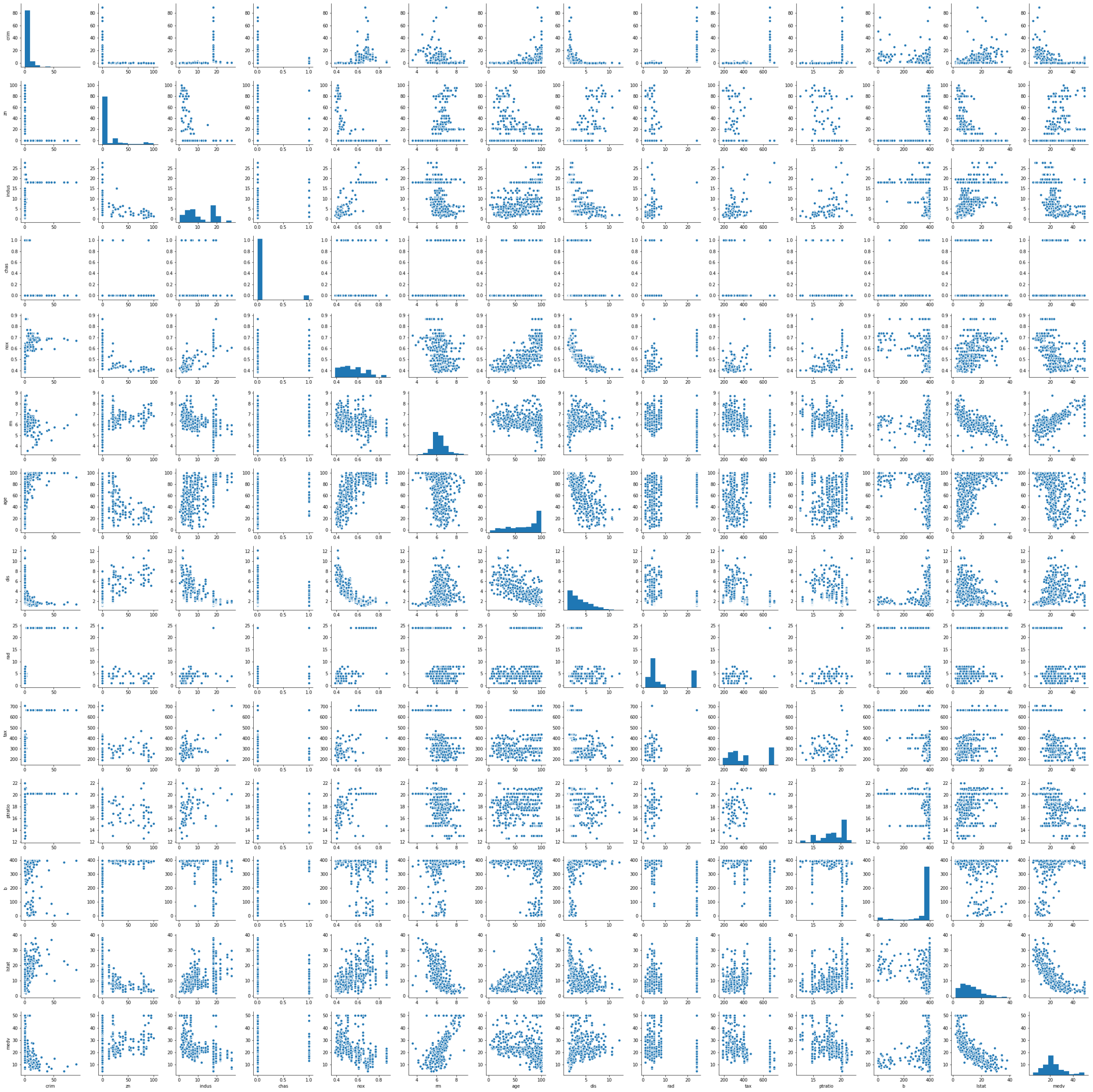
1211309824

Problem 2: Prediction of the Housing price

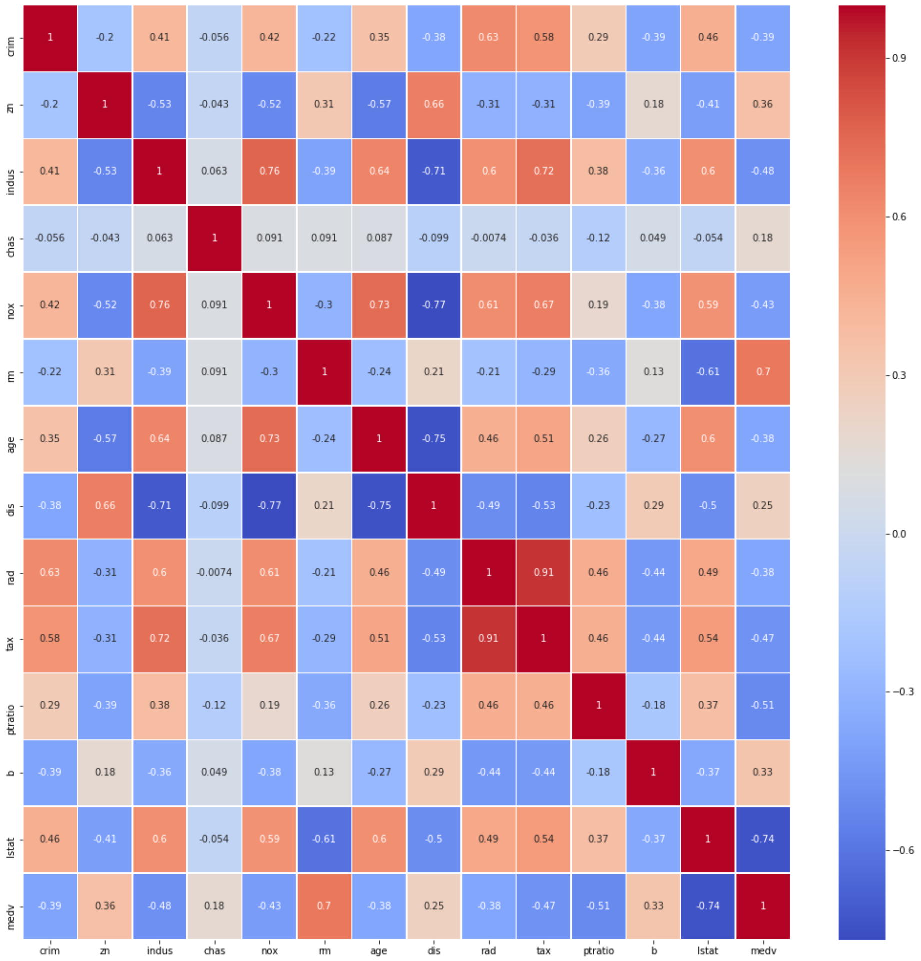
Question 1

Please plot the correlation between all the input variables and output variables pairs. Please identify the first three pairs with strongest correlation (either positive or negative)

Here is the correlation plot between all the input variables and output variables.



I generated the heatmap to visualize the correlation computation between all variables. Here is the heatmap below



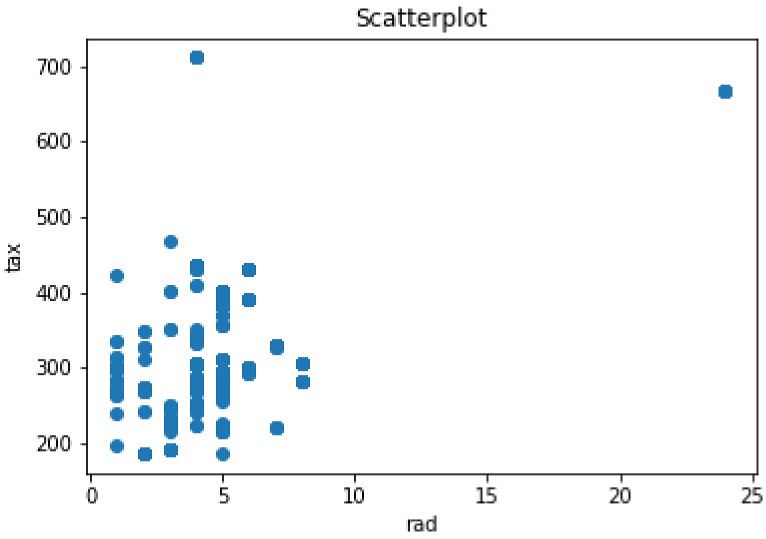
From the heatmap, the top three correlating variables are:

1) rad (index of accessibility to radial highways) and tax (full-value property-tax rate) This pair has a strong positive correlation of 0.91.This makes sense, because tax rate for a particular propertly with access to radial highways will generally be high as this property gives convenience for transportation.

2) nox (nitric oxides concentration) and dis (weighted distances to five Boston employment centres) This pair has a moderately strong negative correlation of -0.77. This conveys an information that Boston employment centres environment typically carry less nitric oxide concentration. Yes, these centres need a good hygiene conditions for work.

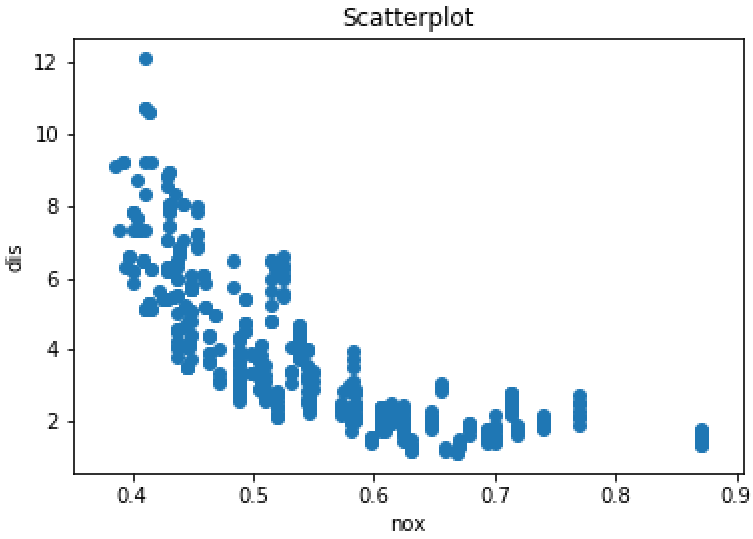
3) nox (nitric oxides concentration) and indus(proportion of non-retail business acres per town) This pair has a moderately strong correlation of 0.76. Non-retail business such as factories with high acre of land will tend to emit nitric oxide during their production work.

Scatterplot of rad vs tax



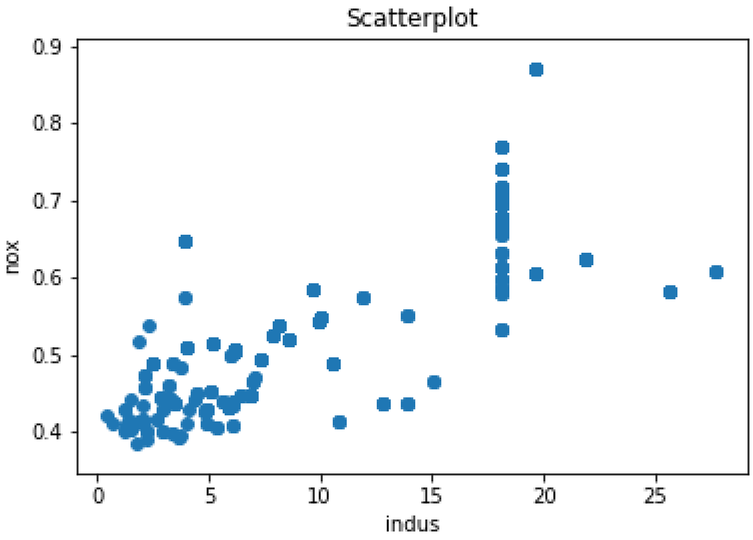
I notice something unusual in this plot considering a strong correlation value of 0.96. This plot is not so linear in nature, the points are scattered and form a cluster. There are outliers seen as well.

Scaterplot of nox vs dis



This plot matches well with a negative correlation of -0.77. As dis value decreases, nox value decreases.

Scatterplot of indus vs nox

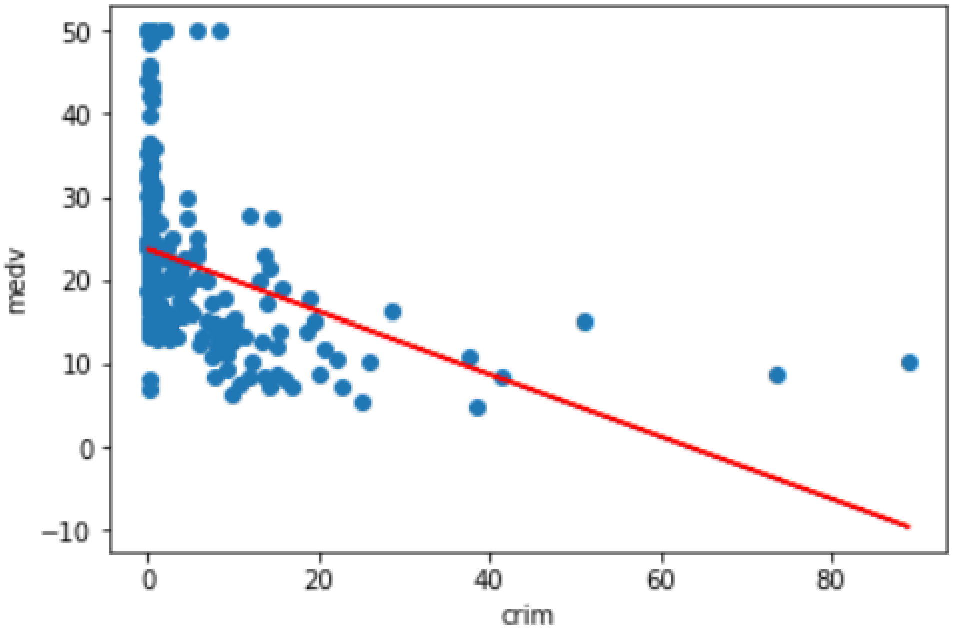


This plot shows a positive moderate correlation between these variables.The observed correlation is 0.76, as indus value increases, nox value increases. But at indus value 18 approximately there is outlier of nox values seen.

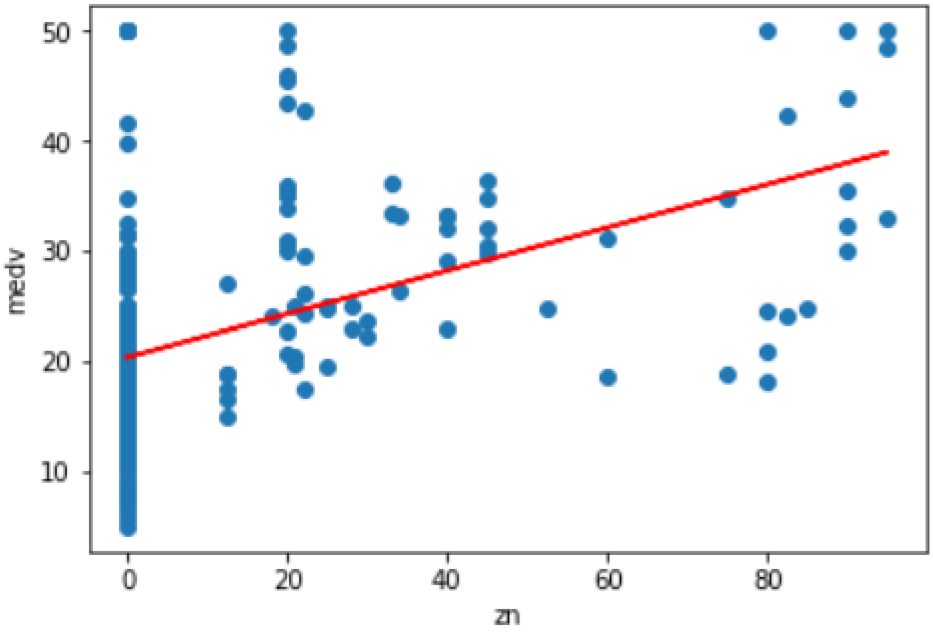
Question 2

Please conduct the simple linear regression of the response 'MEDV' with other individual variables. Please split the data into training and testing and evaluate the testing accuracy for each model. Please generate the plot for each model and report the Residual Sum of Square for each model.

Here is the scatterplot of all individual variables with response variable and their relative sum of squares and accuracy is shown below their image.



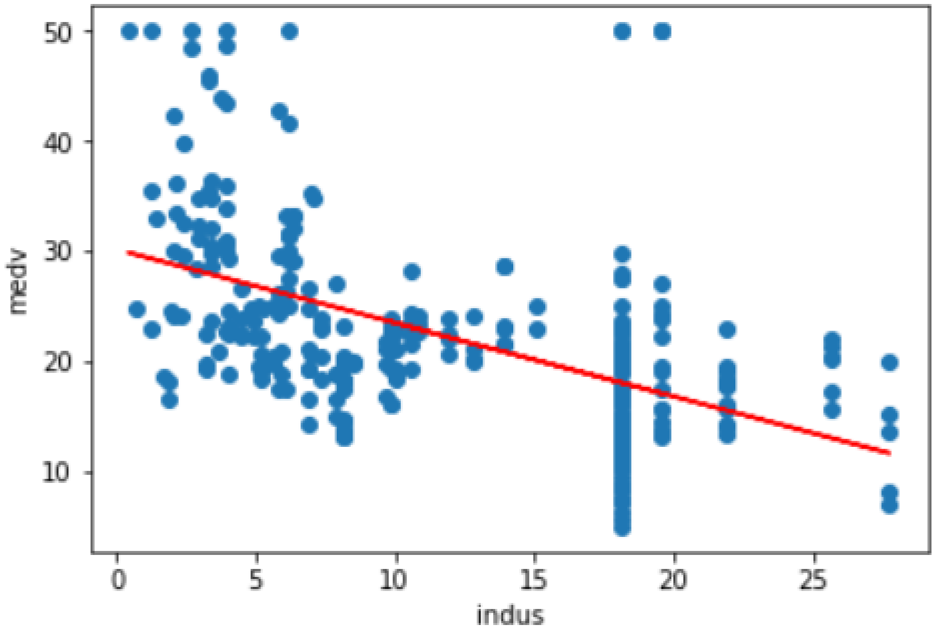
```
crim

Accuracy: 94.20222636135533

RSS: 13739.035962734059
```



zn

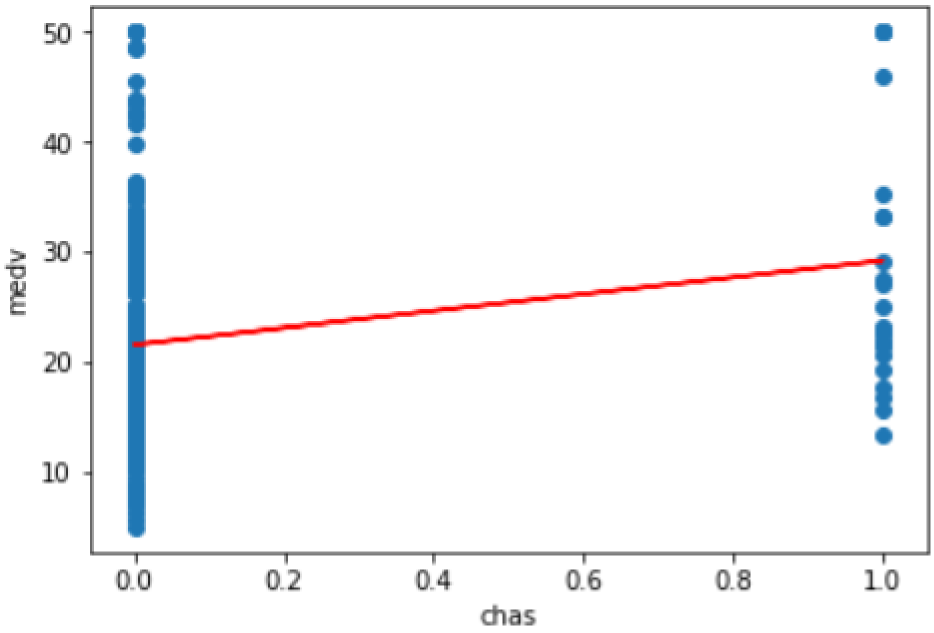
Accuracy: 93.59516931018823

RSS: 17214.844500546158

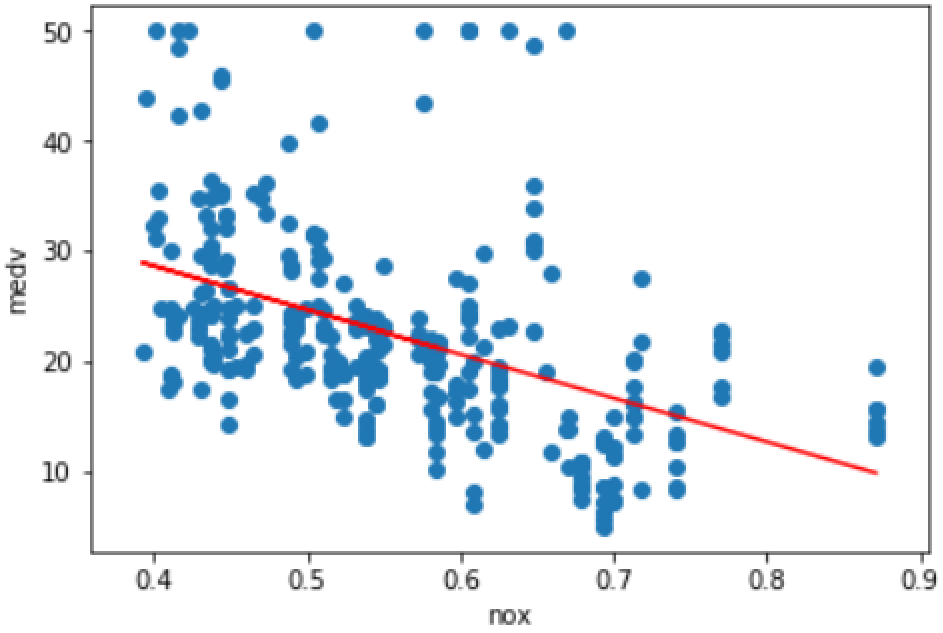


indus

```
Accuracy: 94.35492892686007

RSS: 12921.466465870732
```



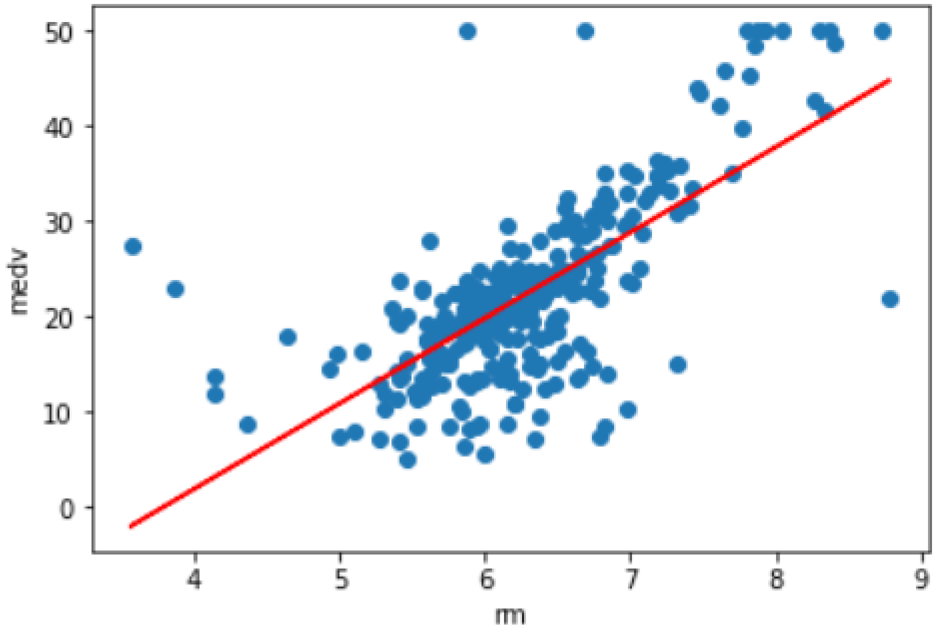
```
chas

Accuracy: 93.63773014286508

RSS: 16610.531499391593
```



```
nox
```

Accuracy: 93.97558125627751

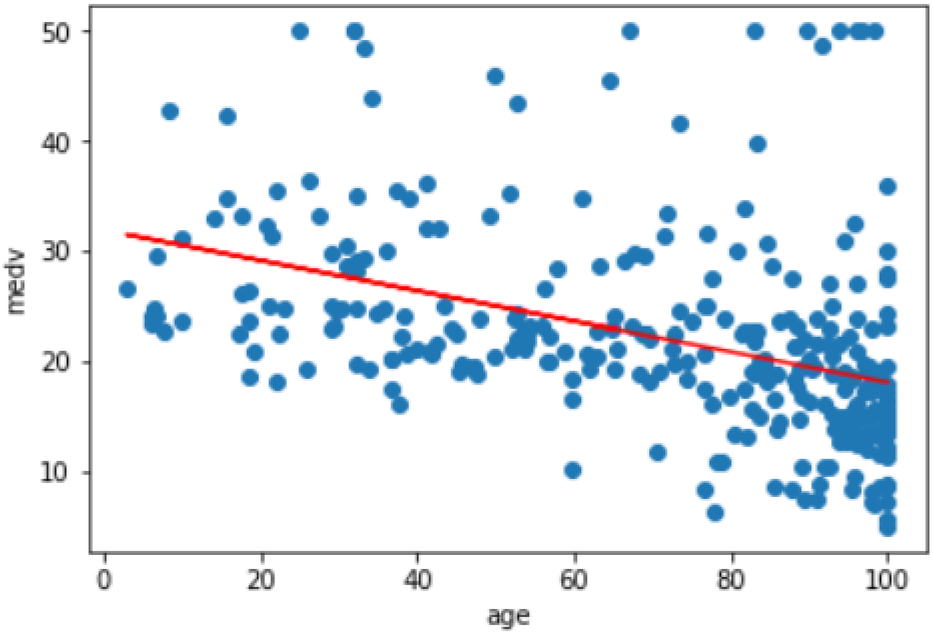
RSS: 14444.50496829556



rm

Accuracy: 95.63045479715126

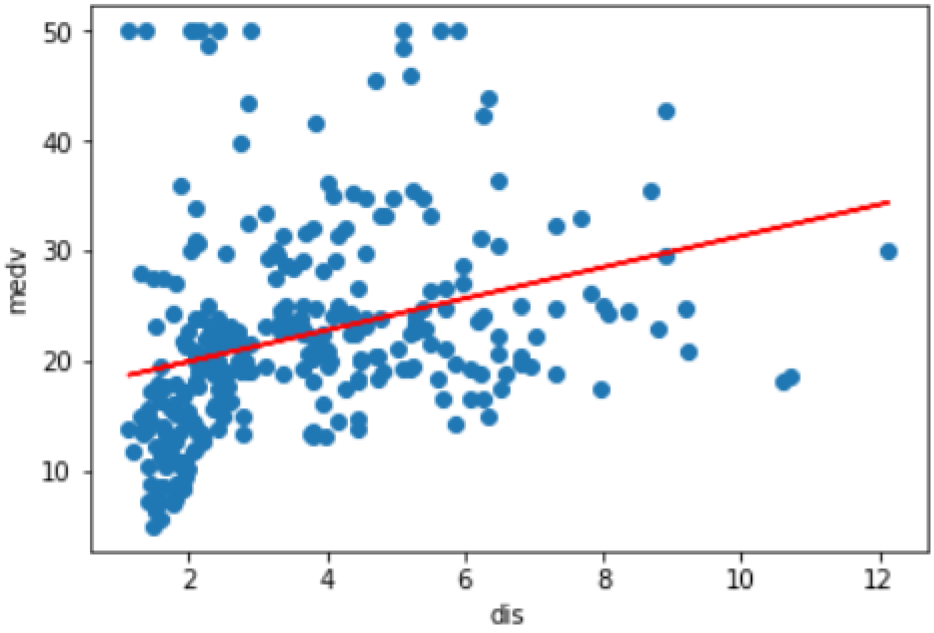
RSS: 8890.66208703505



age

Accuracy: 94.13456860198215

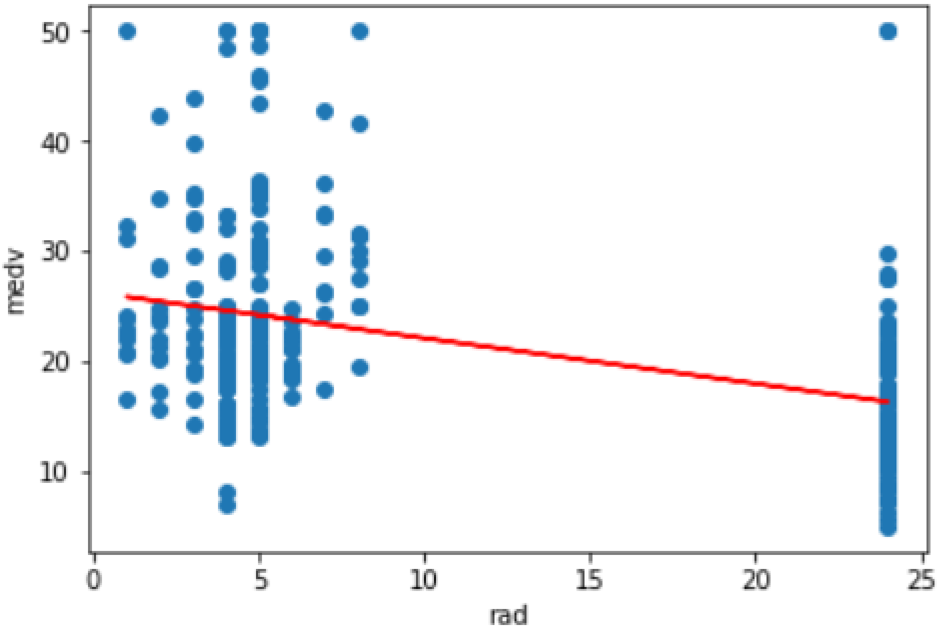
RSS: 14897.445022016602



dis

Accuracy: 93.6891183760721

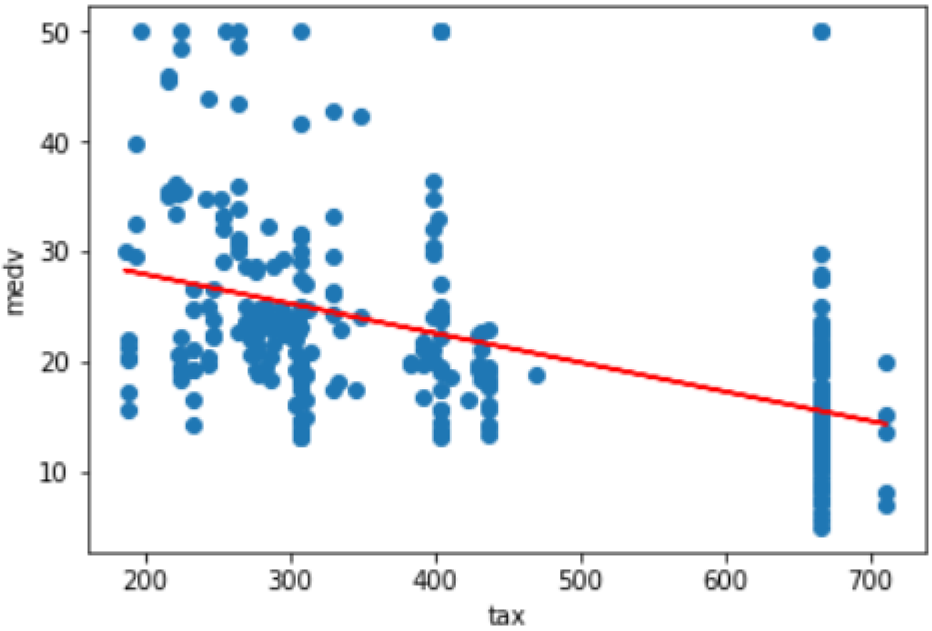
RSS: 16369.87225056796



rad

Accuracy: 94.20574982259042

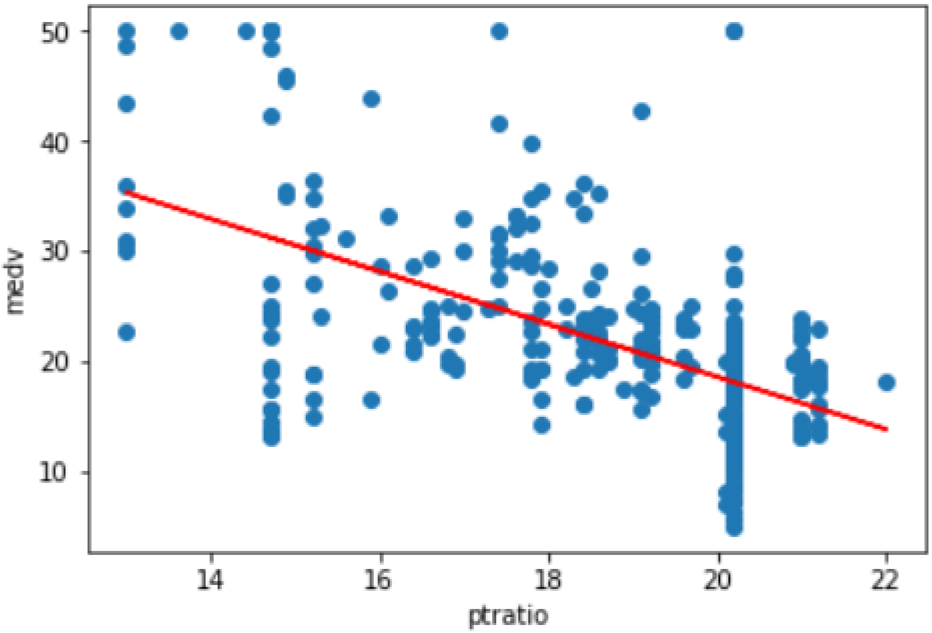
RSS: 14151.055427171103



tax

Accuracy: 94.35585387772795

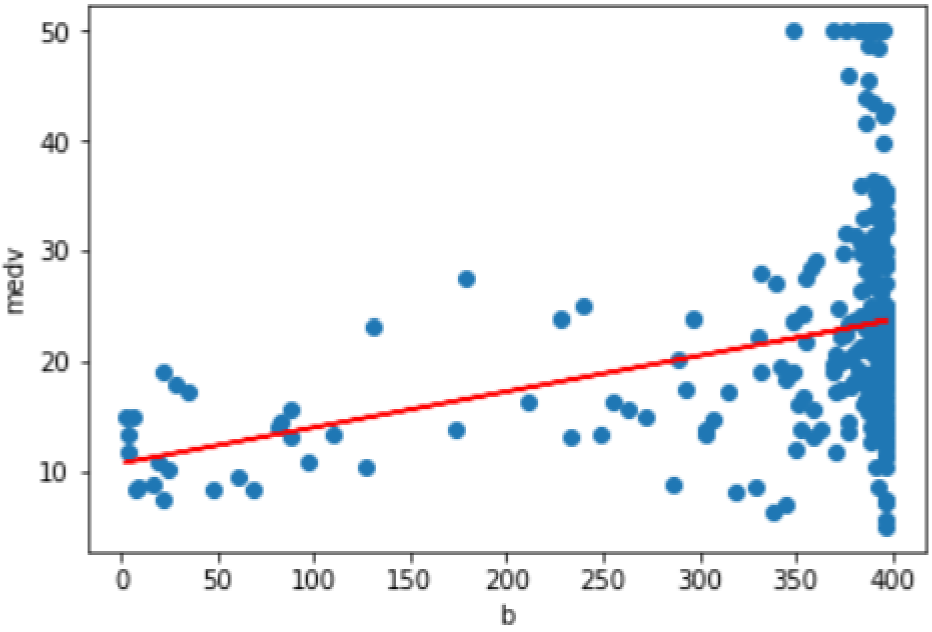
RSS: 13340.680212282861



ptratio

Accuracy: 94.07208701407002

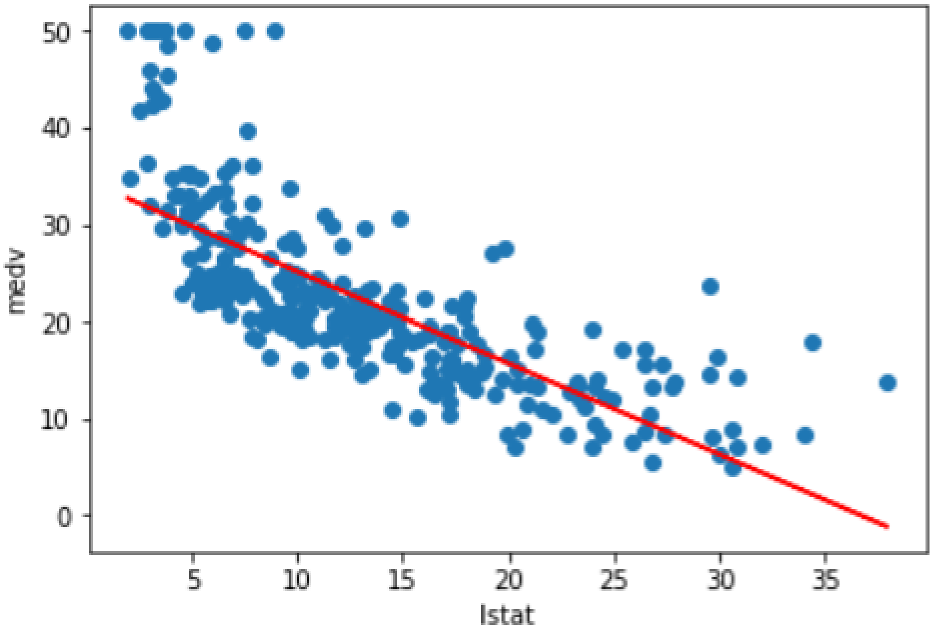
RSS: 13895.713319504957



b

Accuracy: 94.07901528635317

RSS: 14606.134194556247



lstat

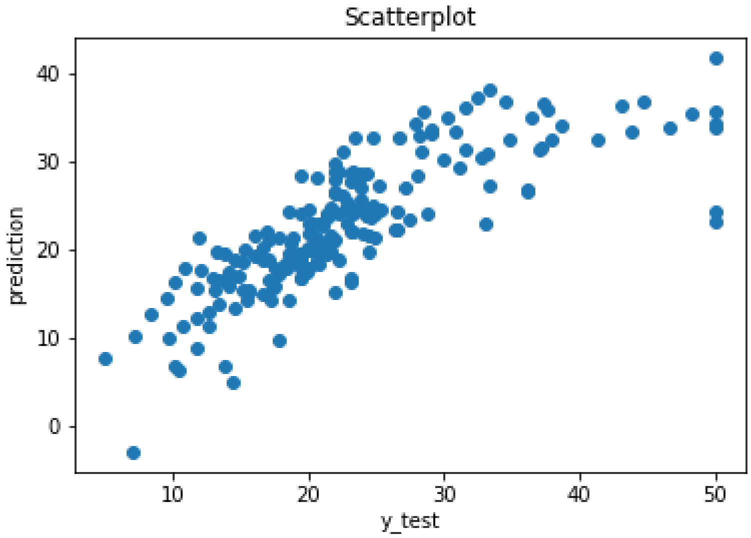
Accuracy: 95.55324320206526

RSS: 7566.53359295479

Question 3

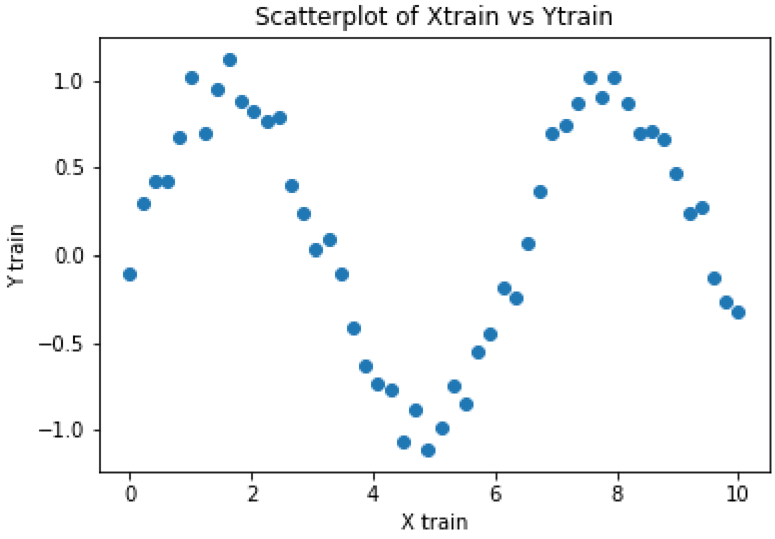
Please use all the other input variables for multiple linear regression to predict the response 'MEDV' with all the input variables and evaluate the testing accuracy

Here is the scatterplot of xtest vs ytest for multiple regression



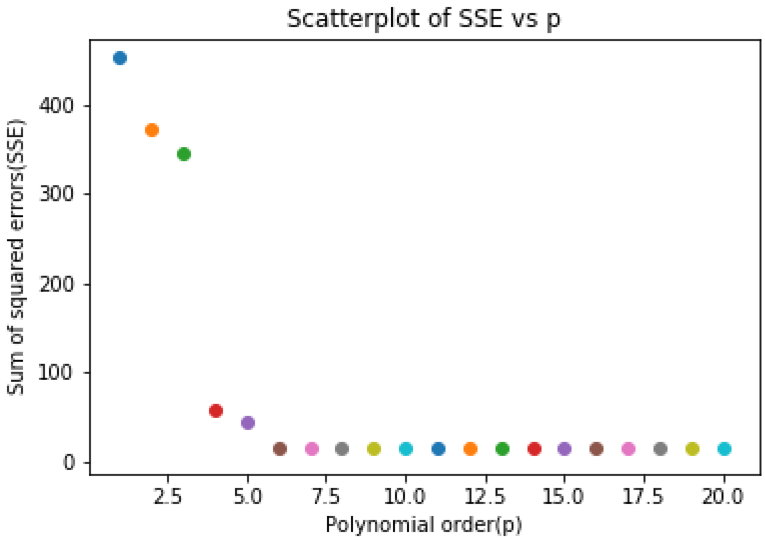
Problem 3: Polynomial Regression Model

Part 1: Visualize the data



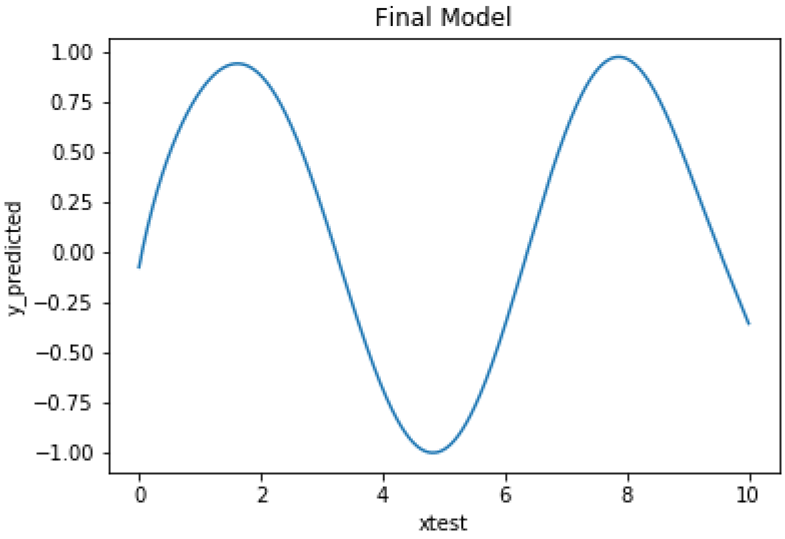
Part 3: Training and Testing Split : Visualize the Testing accuracy

Here is the scatterplot of sum of squared errors for each of the polynomial model



Part 5: Visualization of your final model

I chose polynomial order of 8 as my final model considering Bias-Variance trade off and also for the 8th order as calculated before, the sum of squared error was the least. So having least bias and moderate variance for the 8th order polynomial model will give me a better estimation as most of the values will be focused close to true value(bulls's eye) with moderate deviation between points. This 8th order model gives the best prediction among all the orders.


Final Model
y_predicted
xtest