1.)

At first Added a new column to Boston dataset that indicates crime rate above Median.

(1- Indicates crime rate above median; 0 – indicates crime rate below median)

Then created a training set using 75% of Boston data and the rest will be test data.

<u>Performed Logistic Regression on the whole data (without sub setting):</u>

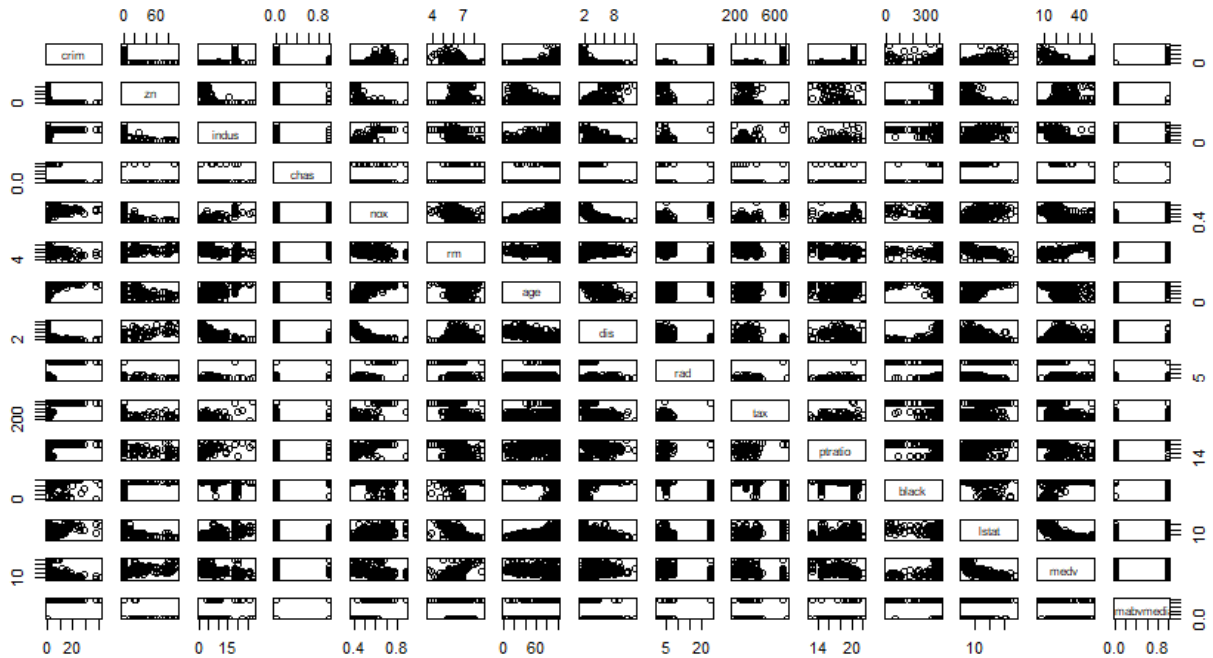
The obtained test error = `0.07874016`

The table obtained for the Test Predictions vs crimrateabvmedian (new added column for dataset)

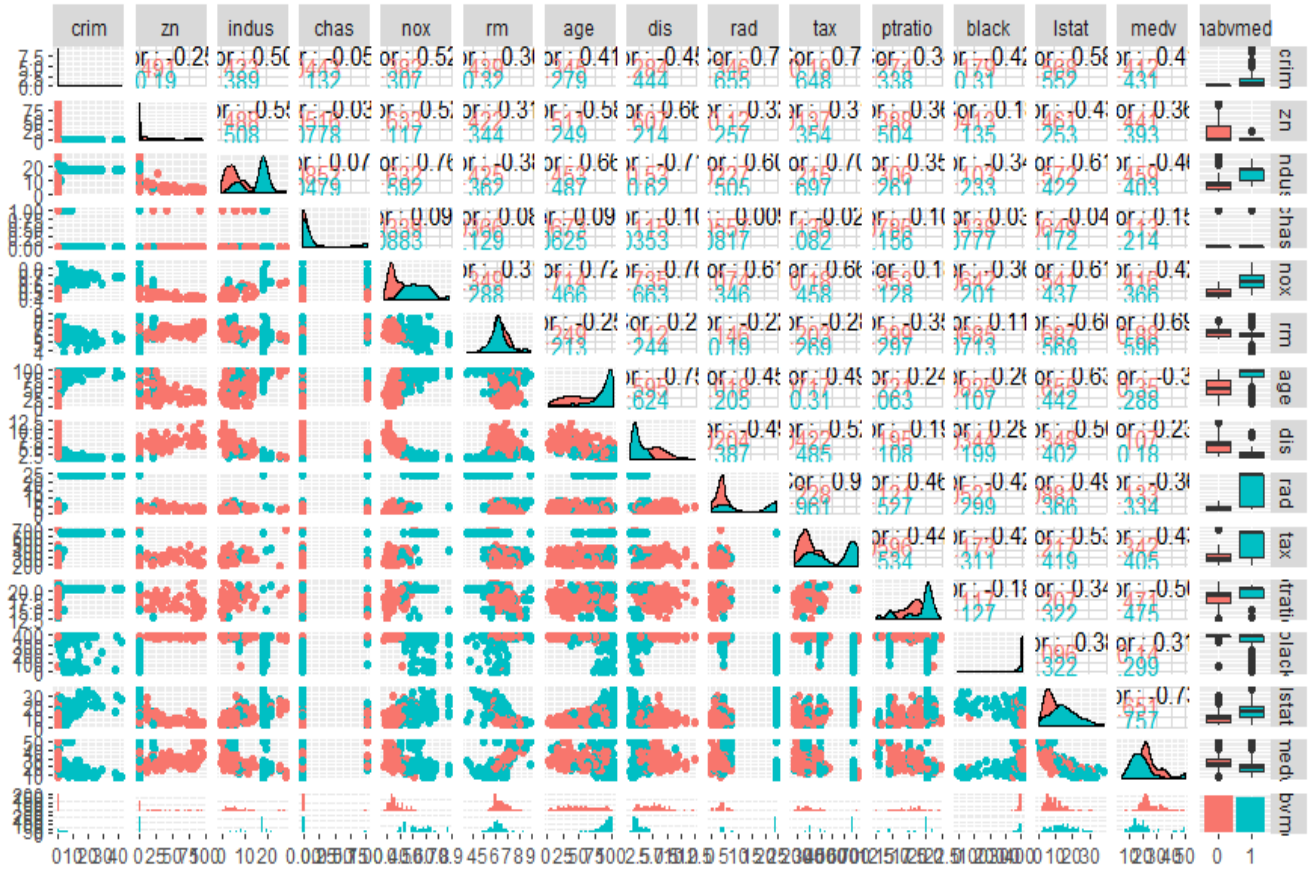
	0	1
0	55	7
1	3	62

Now subset the predictors observing different plots.

<u>The pairwise plot for the Boston training set is:</u>



Pairwise plot for the training set (using crime above median as the color):



Exploring black in the above graph the dominant predictors are:

Ptratio, rad, dis, nox, zn, lstat, rm

Also calculated the correlation coefficient for the data. The correlation of indus, nox, rad, medv, age, tax, ptratio looks good.

So sub setting “indus, nox, rad, medv, age, tax, ptratio” predictors.

Logistic Regression (on subset predictors):

The obtained test error = `0.08661417`

The table obtained for the Test Predictions vs crimrateabvmedian (new added column for dataset)

	0	1
0	57	10
1	1	59

LDA (without sub setting):

The obtained test error = 0.1732283

LDA (on subset predictors):

The obtained test error = 0.1417323

KNN (on subset predictors):

K=1; The obtained test error = 0.1102375
K=2; The obtained test error = 0.1181102
K=3; The obtained test error = 0.1102362
K=4; The obtained test error = 0.1259843
K=5; The obtained test error = 0.1181102

Observing the above errors, K=3 has the least error (Minimum error).

Finally,

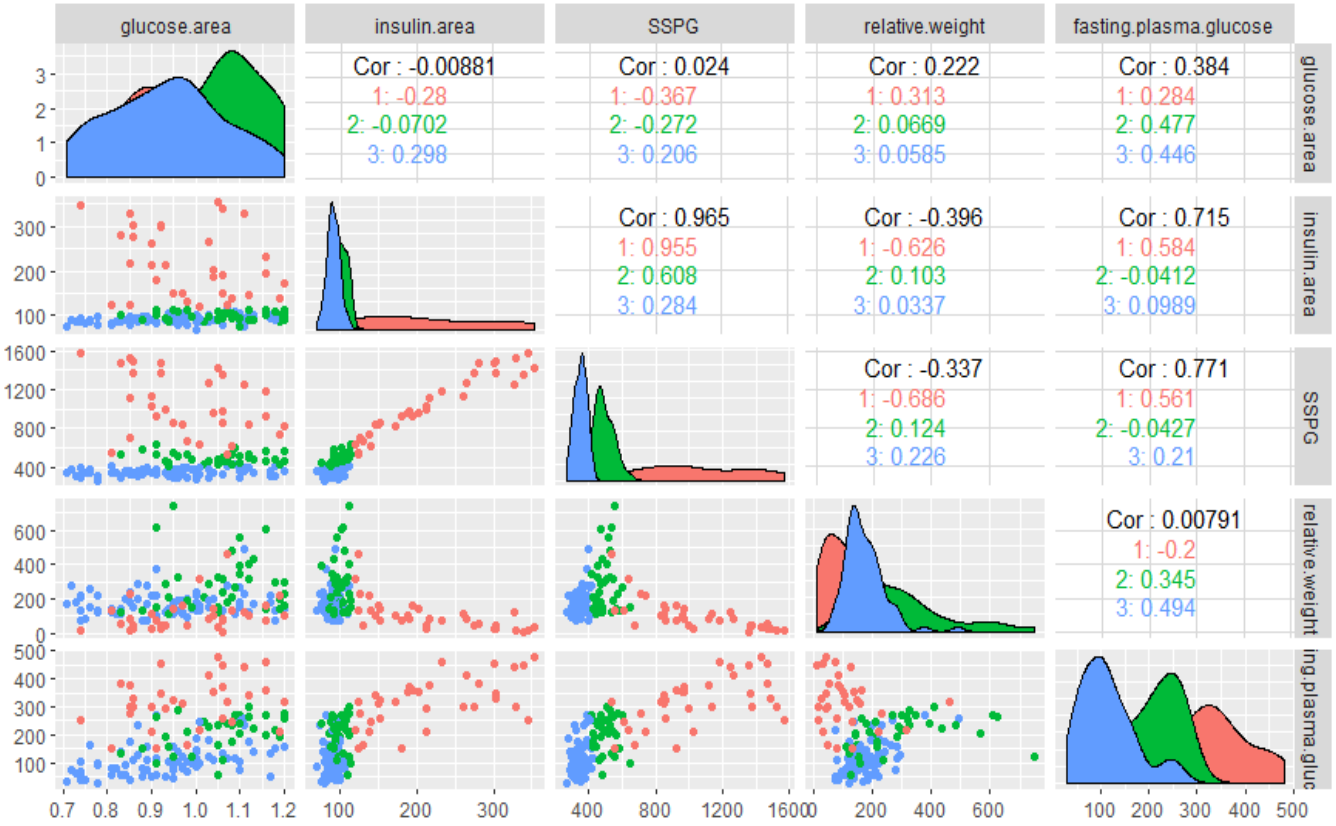
For logistic regression we have a test error rate of 8.66%

For LDA we have a test error rate of 14%

For KNN we have a test error rate of 11.02%(for k=3)

2.) Disregarded the first three columns. Named the other columns as "glucose.area", "insulin.area ", "SSPG", "relative.weight", "fasting.plasma.glucose", "classnumber"

a.) pairwise scatterplots with colors representing the three different classes :



From above plot we can observe that the classes are separable. We can distinguish between different classes easily in the above graph. From this evidence we can say that the classes have different Covariance Matrices.

b.)

LDA:

The test error rate = 0.09090909

The squared error = 4

QDA:

The test error rate = 0.1136364

The squared error = 5

From above values we can observe the performance of LDA is better than QDA.

C.)

LDA assign this individual to class ‘3’

QDA assign this individual to class ‘2’

Show that,

3(a). Sum of posterior probabilities of classes is equal to 1.
Show that this holds for $k = K$.

Let us Prove this by Induction.

For $k=1$, Probability of that class being choosen $= \frac{n}{n} = 1$.

For $k=2$, there are two classes.
The Probability of any of the class being choosen is:

$$\text{Sum} = \frac{k}{n} + \frac{n-k}{n}$$

where k - no. of observations from class 1.
$n-k$ - no. of observations from class 2.

$$\text{Probability} = \frac{k + n - k}{n} = 1.$$

For $k = t-1$, there are $t-1$ classes.

$$\therefore \text{Sum of Posterior Probabilities} = \frac{1}{n} \sum_{i=1}^{t-1} k_i + (n - k_i)$$

$$= 1.$$

Hence By Induction, Sum of Posterior Probabilities of classes is equal to one.

3.
(b.)

$$P(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

$$P(X)\left[1 + \exp(\beta_0 + \beta_1 X)\right] = \exp(\beta_0 + \beta_1 X).$$

$$P(X) + P(X)\exp(\beta_0 + \beta_1 X) = \exp(\beta_0 + \beta_1 X).$$

$$-P(X)\exp(\beta_0 + \beta_1 X) + \exp(\beta_0 + \beta_1 X) = P(X)$$

$$\therefore \exp(\beta_0 + \beta_1 X)\left[1 - P(X)\right] = P(X).$$

$$\therefore \boxed{\frac{P[X]}{1 - P[X]} = \exp(\beta_0 + \beta_1 X).}$$

Hence, using little bit of Algebra we have proved both the Models are equivalent.

4.)

a.)

LOOCV error for:

first degree polynomial = `9.217431`

second degree polynomial = `1.094918`

third degree polynomial = `1.101478`

fourth degree polynomial = `1.115254`

b.)

The quadratic polynomial (second degree) had the lowest LOOCV test error rate and yes I expected that because the true data is of quadratic form.

c.)

observing the summary of both the fourth degree and second degree polynomial.

If we look at the quadratic fit, we can see that both the quadratic term (second degree) and the linear term (first degree) are significant. This is what we expect. Yes, these results agree with the conclusions drawn from the cross-validation.