HOMEWORK 3
STATISTICAL DATA MINING

PRIYA MURTHY
50248887

QUESTION 1

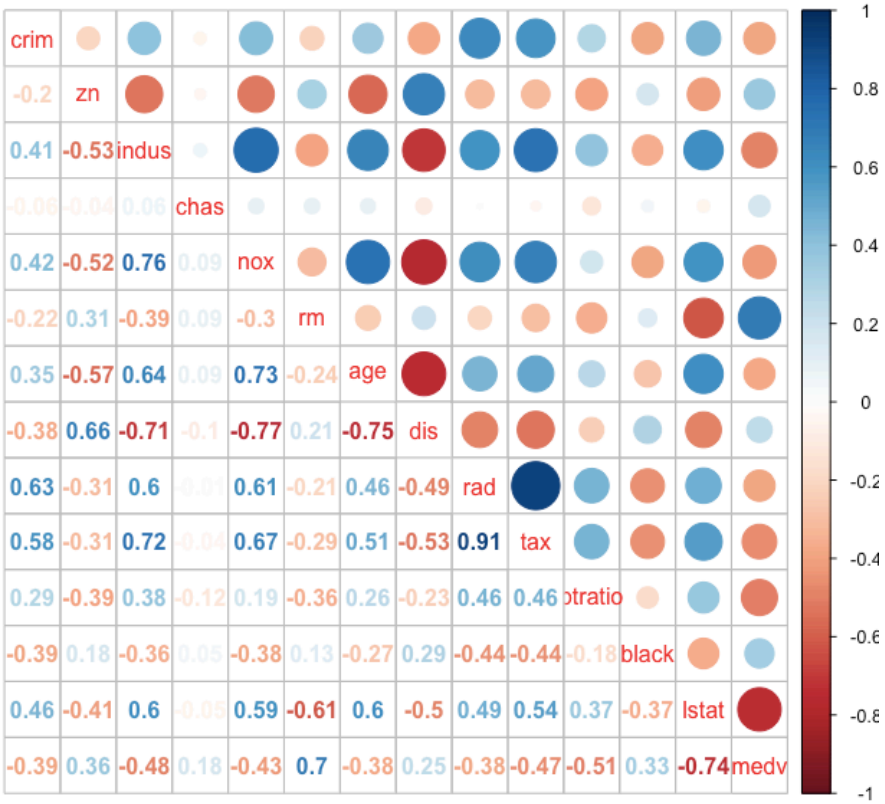
ANSWER 1

Steps:

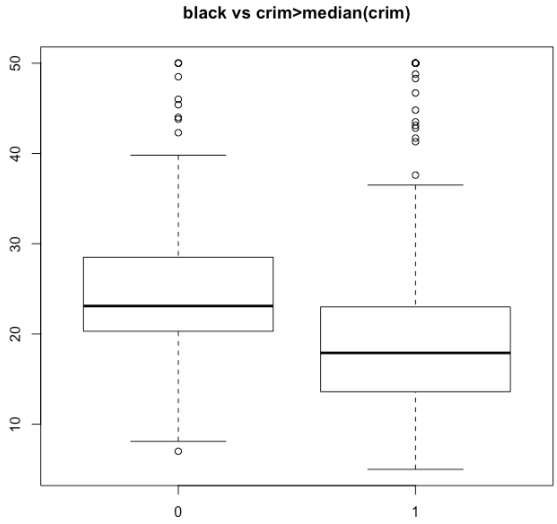
Step 1: Create dataset, found data set where crime rate above or below the median

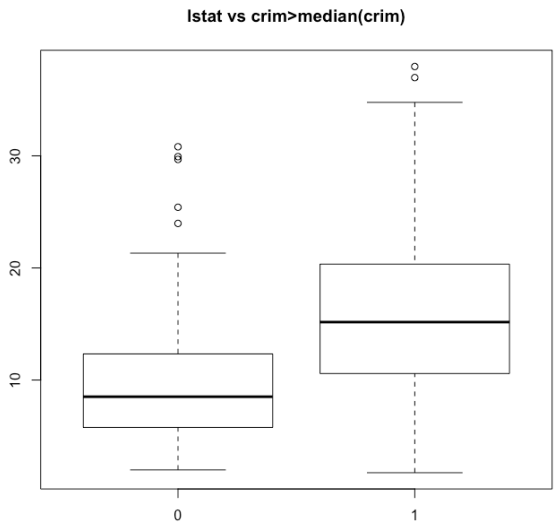
Step 2: Analyzing the dataset:

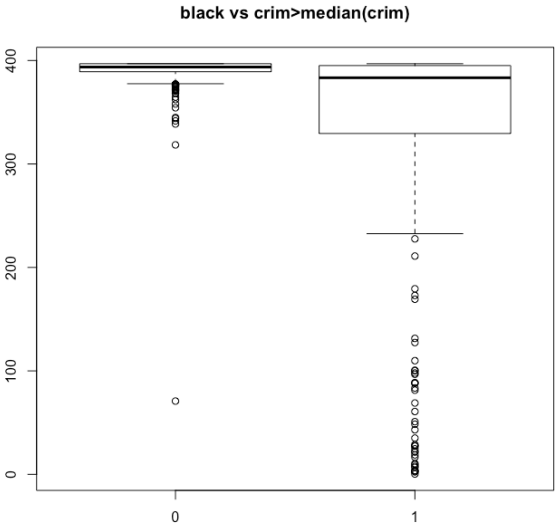
Correlation between different values:

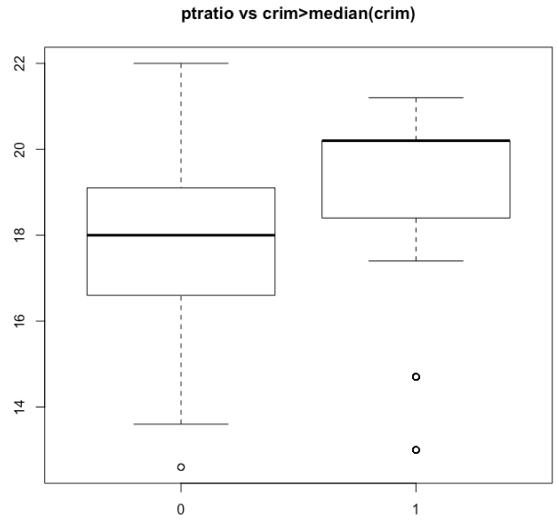


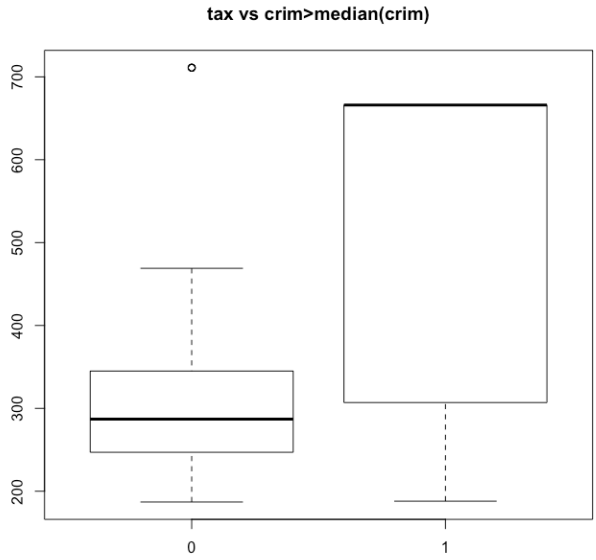
Step 3: Make boxplots to find relation between the different predictors

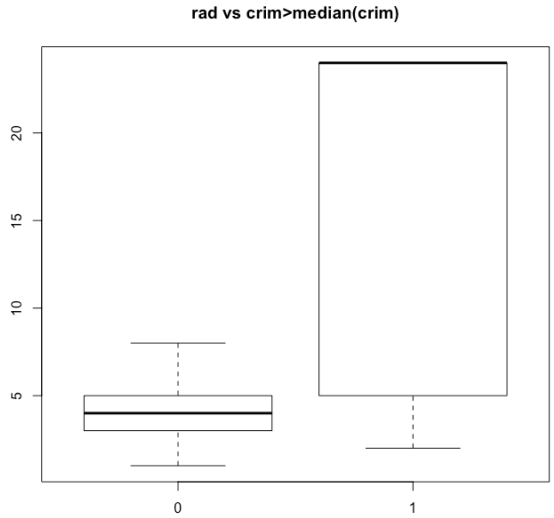


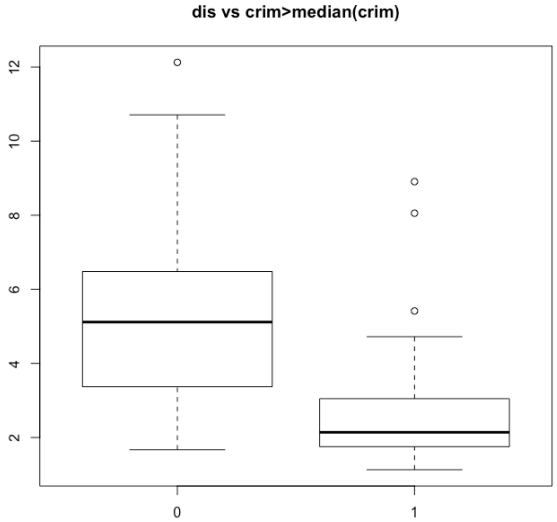


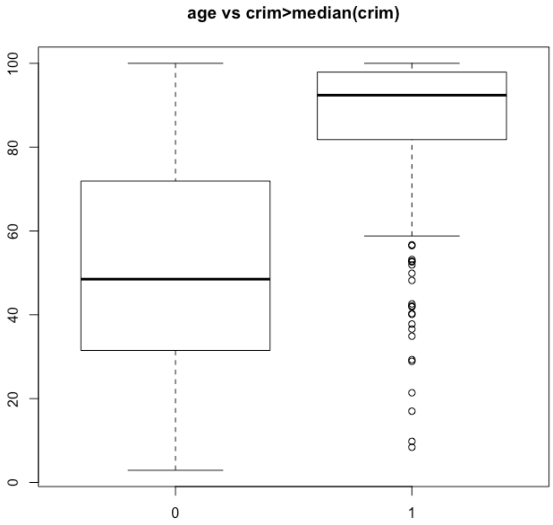


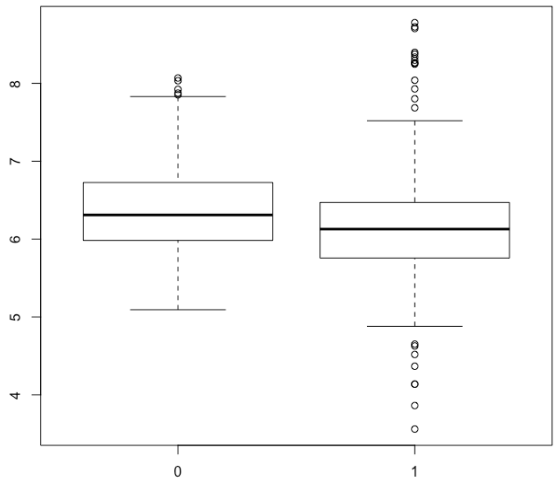


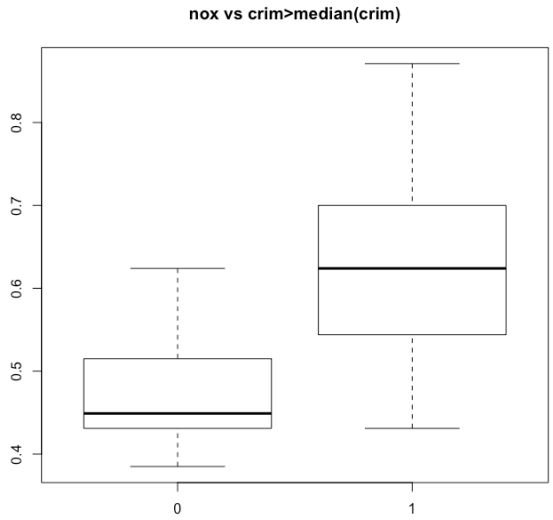

tax vs crim>median(crim)
700
600
500
400
300
200
0
1

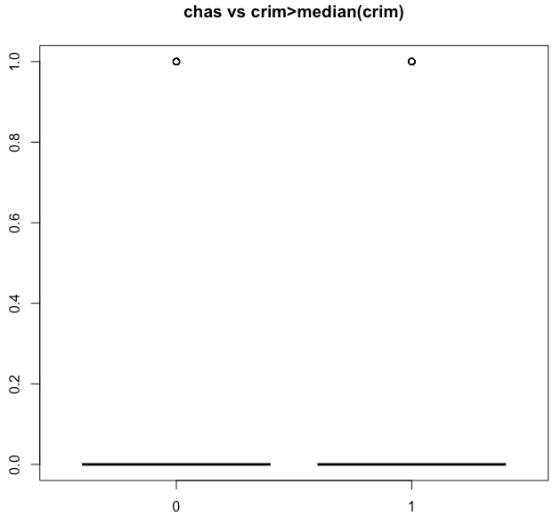

rad vs crim>median(crim)
20
15
10
5
0
1

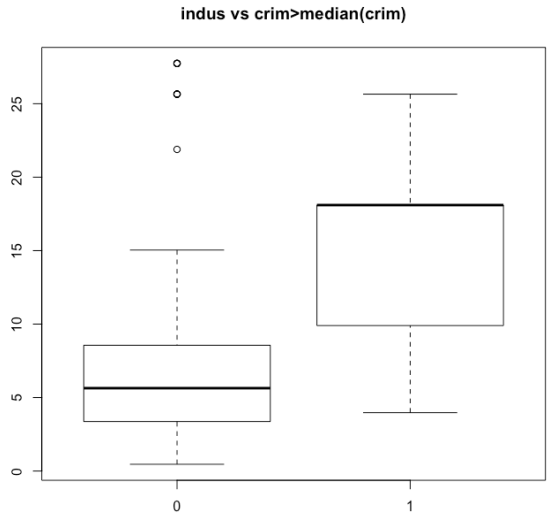

dis vs crim>median(crim)
12
10
8
6
4
2
0
1

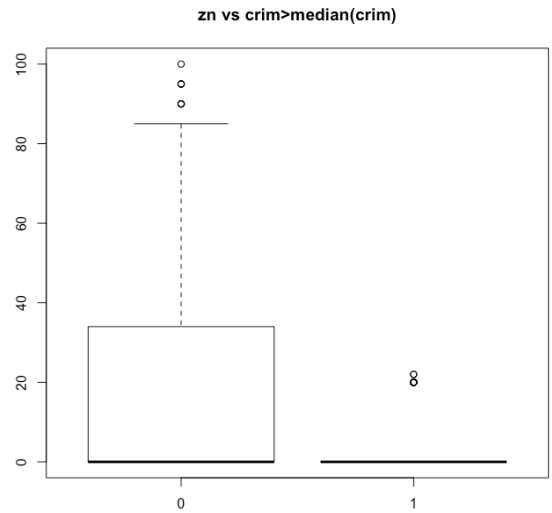

age vs crim>median(crim)
100
80
60
40
20
0
0
1


rm vs crim>median(crim)
8
7
6
5
4
0
1


nox vs crim>median(crim)
0.8
0.7
0.6
0.5
0.4
0
1







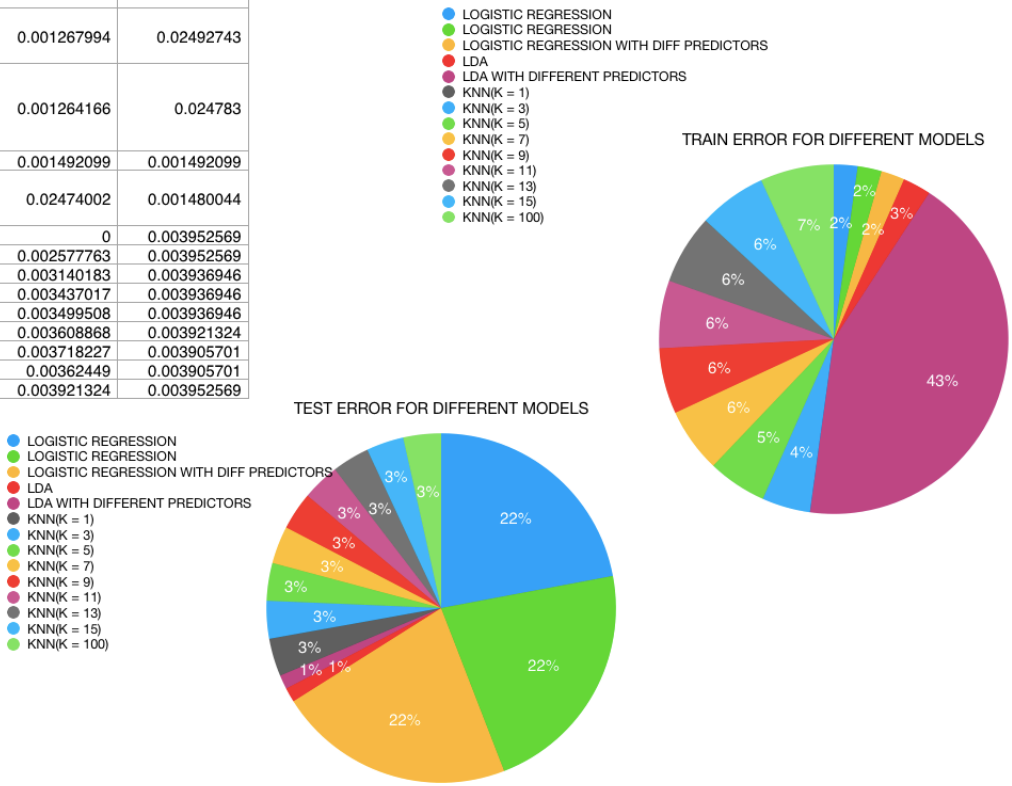
Looking at the above boxplots, we found the predictors which have more relation with the predictor crim.

We can conclude that lstat, ptratio, tax, rad are highly correlated with crim and zn and chas are not. Similarly, we can conclude for other predictors as well.

Step 4: Applied logistic regression, LDA and KNN for different combination of predictors.

COMPARING ERROR RATE

MODEL	TRAIN ERROR	TEST ERROR
LOGISTIC REGRESSION	0.001267994	0.02492743
LOGISTIC REGRESSION WITH DIFF PREDICTORS	0.001264166	0.024783
LDA	0.001492099	0.001492099
LDA WITH DIFFERENT PREDICTORS	0.02474002	0.001480044
KNN(K = 1)	0	0.003952569
KNN(K = 3)	0.002577763	0.003952569
KNN(K = 5)	0.003140183	0.003936946
KNN(K = 7)	0.003437017	0.003936946
KNN(K = 9)	0.003499508	0.003936946
KNN(K = 11)	0.003608868	0.003921324
KNN(K = 13)	0.003718227	0.003905701
KNN(K = 15)	0.00362449	0.003905701
KNN(K = 100)	0.003921324	0.003952569



ACCURACY OF DIFFERENT MODELS

MODEL	TEST ACCURACY	TRAIN ACCURACY
LOGISTIC REGRESSION	97.50726	99.8732
LOGISTIC REGRESSION WITH DIFF PREDICTORS	97.5217	99.87358
LDA	97.52509	99.85079
LDA WITH DIFFERENT PREDICTORS	99.852	97.526
KNN(K = 1)	99.60474	100
KNN(K = 3)	99.60474	99.74222
KNN(K = 5)	99.60631	99.68598
KNN(K = 7)	99.60631	99.6563
KNN(K = 9)	99.60631	99.65005
KNN(K = 11)	99.60787	99.63911
KNN(K = 13)	99.60943	99.62818
KNN(K = 15)	99.60943	99.63755
KNN(K = 100)	99.60474	99.60787

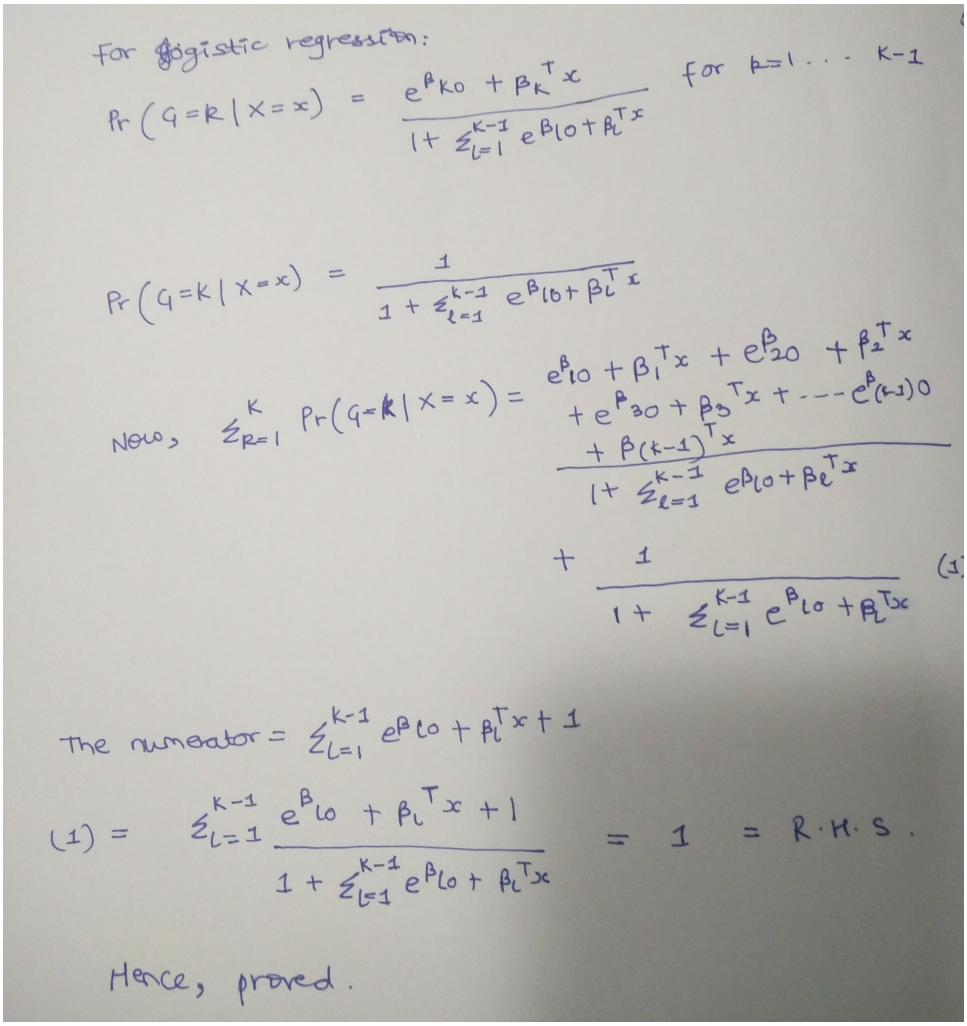
In the above figure, for the model with different predictors, we added predictors lstat,ptratio and tax and removed chas and zn.

Analysis:
Looking at the above data:

- Logistic Regression: The test accuracy of the model increases when predictors are taken according to the relation between predictors and crim rate.
- LDA: In this case also the test accuracy increases significantly when we build the model according to the different subsets of predictors.
- KNN: For knn, Train accuracy is 100 when K = 1, Test accuracy is maximum for K = 13 and K =15.

QUESTION 3
ANSWER 3(A)

We are given that sum of posterior probabilities of classes is equal to one.
Proof:



ANSWER 3(B)

Given: $p(X)$ = [exp ($\beta 0 + \beta 1X$)] / [1 + exp ($\beta 0 + \beta 1X$)]

1-p(X) = 1- [exp ($\beta 0 + \beta 1X$)] / [1 + exp ($\beta 0 + \beta 1X$)]
= [1 + exp ($\beta 0 + \beta 1X$) - exp ($\beta 0 + \beta 1X$)] / [1 + exp ($\beta 0 + \beta 1X$)]
= 1/ [1 + exp ($\beta 0 + \beta 1X$)]

Therefore, (X) / [1-p(X)] = [exp ($\beta 0 + \beta 1X$) * 1 + exp ($\beta 0 + \beta 1X$)]] /[1 + exp ($\beta 0 + \beta 1X$)]

$p(X)$ / [1-p(X)] = exp ($\beta 0 + \beta 1X$)

QUESTION 4

ANSWER 4(A)

The LOOCV errors that result from fitting the following four models using least squares are given below:

For model 1:
$Y = \beta0 + \beta1X + \varepsilon$
Error = 7.288162

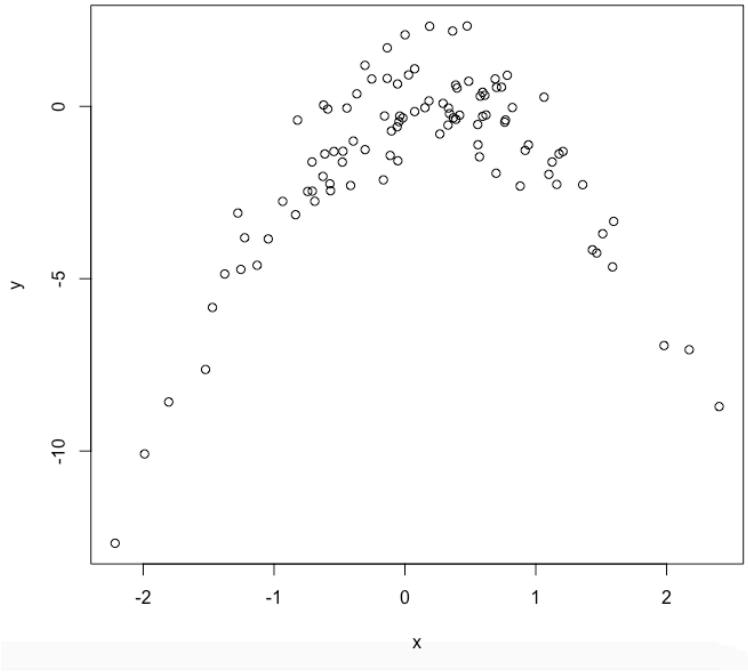
For model 2:
$Y = \beta0 + \beta1X + \beta2X2 + \varepsilon$
Error = 0.9374236

For model 3:
$Y = \beta0 + \beta1X + \beta2X2 + \beta3X3 + \varepsilon$
Error = 0.9566218

For model 4:
$Y = \beta0 + \beta1X + \beta2X2 + \beta3X3 + \beta4X4 + \varepsilon$
Error = 0.9539049

ANSWER 4(B)

1. From the above observation, Model 2 has the smallest LOOCV error.
2. Let's make a scatterplot to observe the relationship between X and Y.



We can see that the plot obtained is a quadratic plot, i.e the relation between x and y is quadratic.

3. The above plot also justifies as to why the quadratic polynomial has the lowest error rate, as it matches the true form of Y.

ANSWER 4(C)

Looking at the summaries of different models, we can conclude that the p values show statistical significance of linear and quadratic terms, which is same as the Cross-Validation results.
The other terms (cubic and 4^{th} power) are not statistically significant.

QUESTION 5

ANSWER 5(A)

We know,
p = 1 feature and X is uniformly distributed on [0,1]
We wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observations.
Case 1: If x belongs to [0.05,0.95], then observations used will be in the interval [x-0.05, x+0.05] which represents a fraction of 10%.
Case 2: If x<0.05, observations used will be in the interval [0,x+0.05], this represents a fraction of (100x +5)%.
Case 3: If x >0.95, observations used will be in the interval [0.95+x,1], then the fraction of observations we will use is (105 – 100x) %.
Calculating the average:
$\int_{0.05}^{0.95} 10\,dx + \int_{0}^{0.05}(100x+5)dx + \int_{0.95}^{1}(105-100x)dx$ = 9+ 0.375+0.375 = 9.75

Thus, the fraction of available observations we will use to make the predictions is: 9.75%.

If we ignore the cases x<0.05 and x>0.95,
Then, the fraction of available observations we will use to make the predictions is 10%.

ANSWER 5(B)

We are given that,
p = 2 features, X1 and X2 and both are uniformly distributed on [0,1].
Looking at the example, If X1 and X2 are independent, then the fraction of available observations we will use to make the predictions is: 9.75%×9.75%=0.950625%

ANSWER 5(C)

We are given that, p = 100 features, and observations are uniformly distributed on [0,1] similar to the above two observations.
Thus, the fraction of available observations that we will use to make the predictions is: $0.10^{100} * 100 = 10^{-98\%}$ which is approximately equal to 0%.

ANSWER 5(D)

In the above three cases, we saw that as p increases linearly, the observations that are geometrically nearby decrease exponentially.
Thus, the fraction of available observations that we used to make predictions can be given by $(9.75\%)^p$.
Thus, when p → infinity.
lim $(9.75\%)^p$ = 0.
p->infinity