Home Work 3

Q1

The Boston dataset has 506 rows. To ensure that our training dataset has a Confidence level of 95% we will ensure at least 385 rows in the dataset and will split our test and train at ~75% of the dataset. We will take the median of the output crime which is 0.2565 and will create an output response variable defining high and low crime rates and fit our classification models to predict this response. The following are the observations from various classification models:

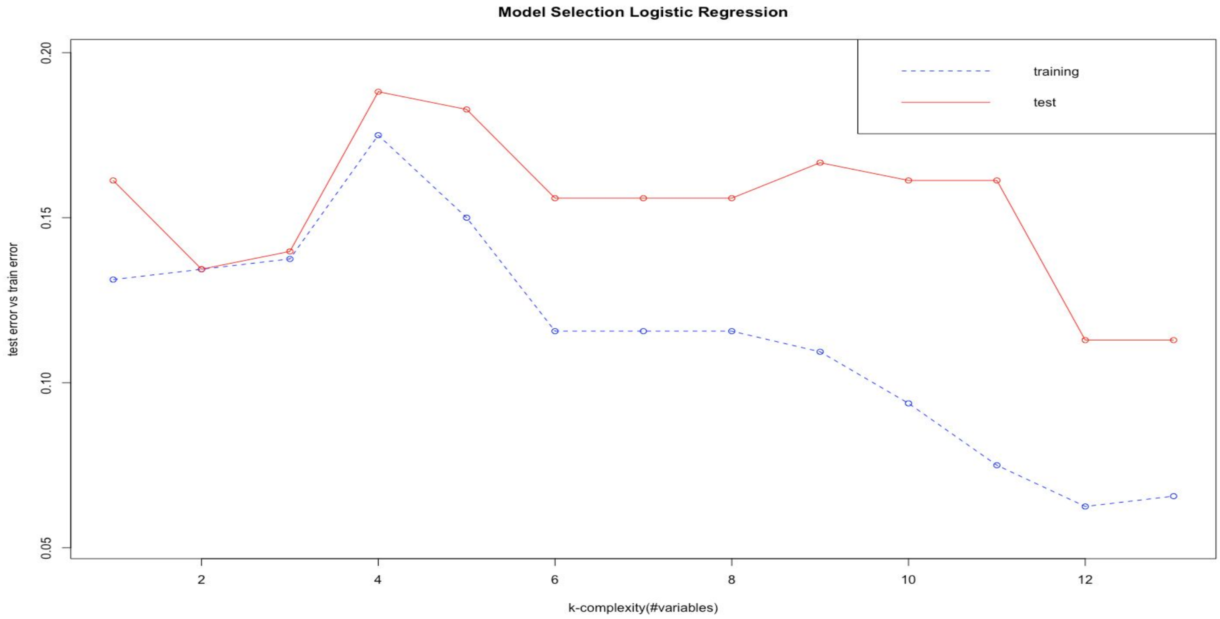
Logistic Regression

There are 8 significant variables zn, nox, age, dis, rad, tax, ptratio, and medv. The following are the metrics from the confusion matrix for train and test:

Dataset	Accuracy	Precision	Recall
Train	0.9196	0.9308	0.9067
Test	0.9166	0.9310	0.9

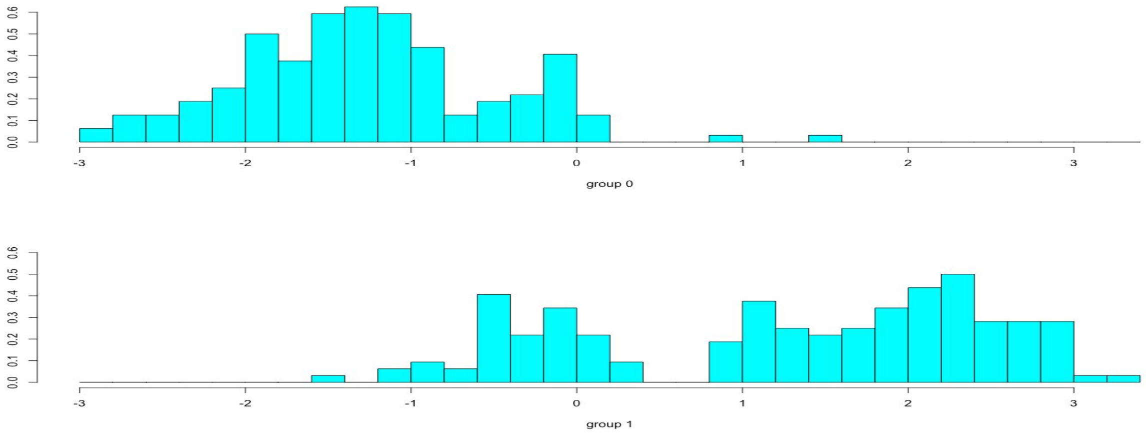
As we can see that the model built is performing pretty well w.r.t. All metrics have a very close difference in train and test accuracies, so it is a pretty stable model for the dataset. We will try to reduce redundant variables so that models are performing well. We run exhaustive subset selection on the dataset and we see that the **minimum value for Cp is a model with 6 variables while minimum BIC is for a model with 4 variables.**

Iterating through various variables in the dataset in the order of significance found via exhaustive subset selection we can see that the model that will best fit Occam's Razor principle will be the one with 3 variables which are nox, rad, and age.



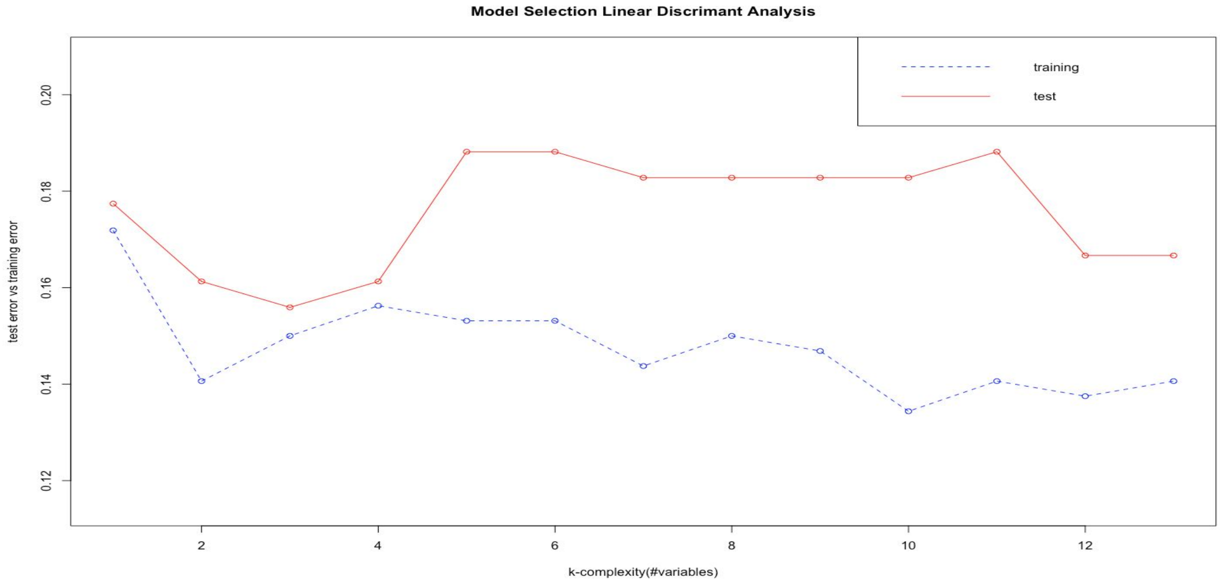
Linear Discriminant Analysis:

The variable with the highest LDA coefficient is nox at 7.91. The following is the graph of the classification of groups into zeroes and ones, along with the values of threshold. As we can see there is some overlap between 0 and 1 for the one linear discriminant.



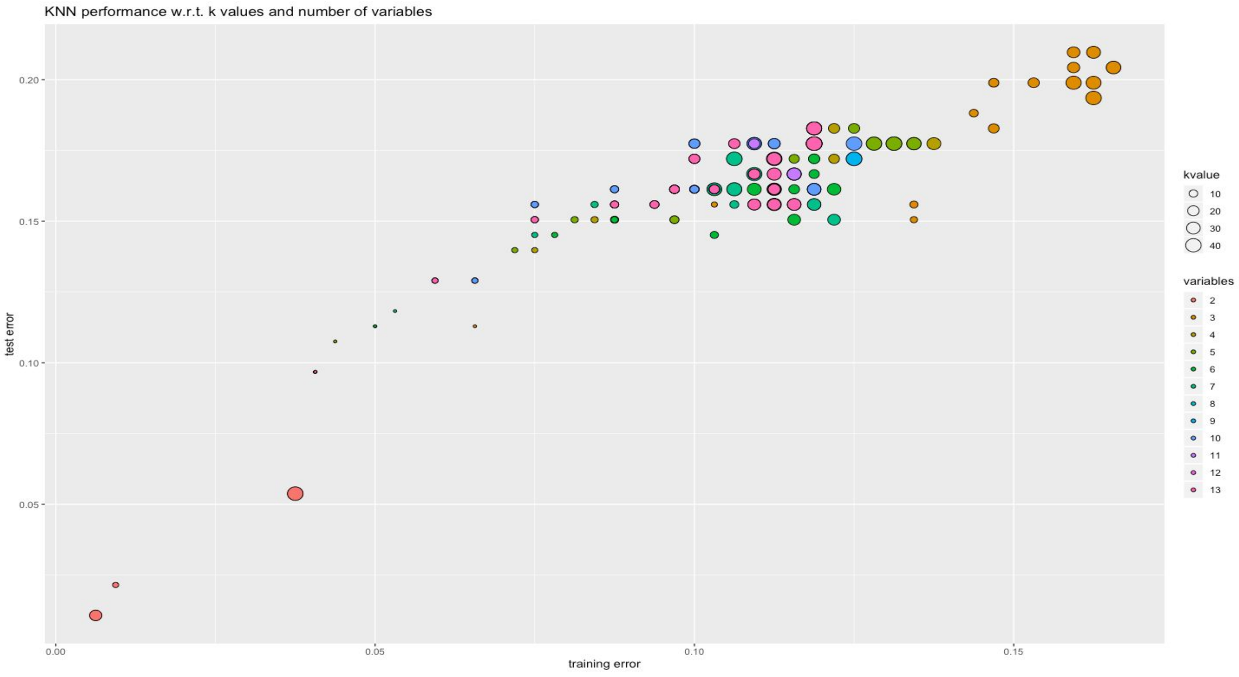
Dataset	Accuracy	Precision	Recall
Train	0.934375	0.9483	0.9187
Test	0.8333	0.9079	0.7419

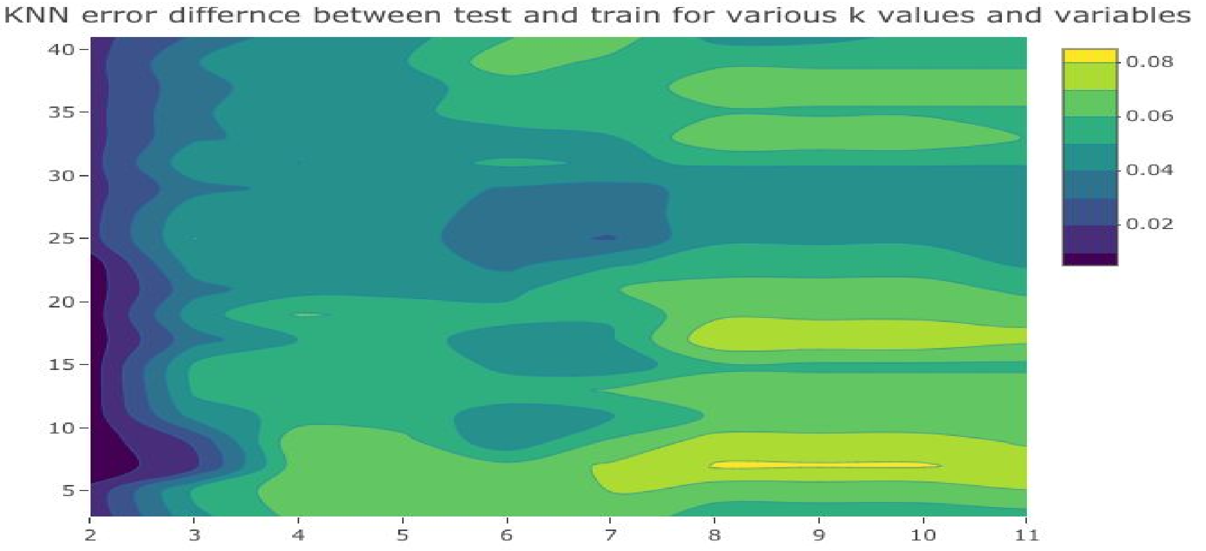
We can see that the test and train accuracies are not very close and thus LDA is not a good model for this dataset. Iterating through various variables in the dataset in the order of significance found via exhaustive subset selection we can see that the model that will best fit Occam's Razor principle will be the one with 3 variables which are nox, rad, and age.



KNN:

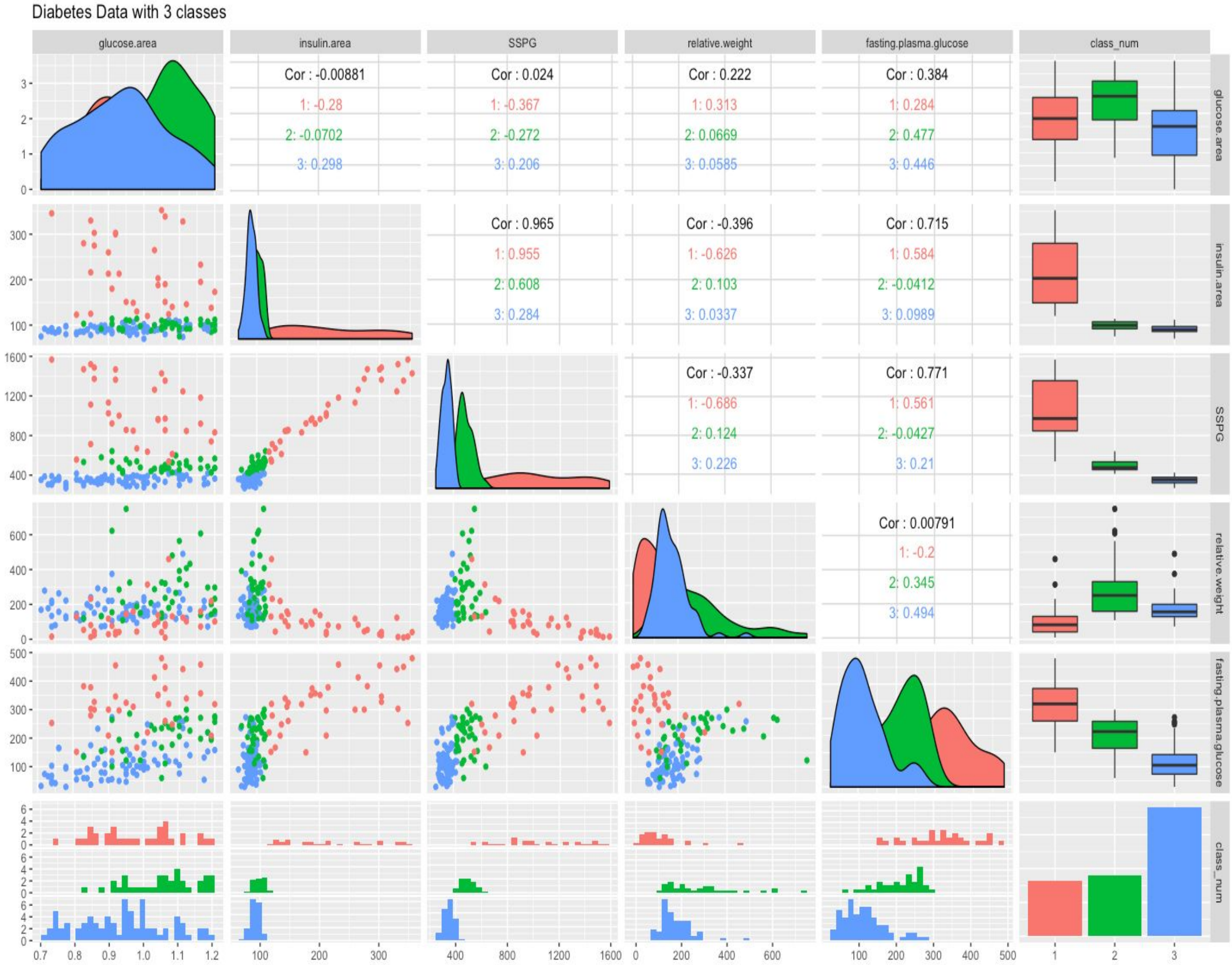
We can see that running KNN for all 13 variables has the same train error and the best test error for k value of 11 as a higher value of k will have lower complexity. Using subset selection we try to iterate through variables in the order of significance in the exhaustive subset selection and various k values and try to arrive at an appropriate set of variables and k value. As the complexity of the model will be defined as a combination of variables and the k value, we will try to minimize the variables and maximize the k-value for low complexity while keeping low test and train error. Interestingly in KNN as we can see at the bottom-left part of the graph that model with 2 variables and k=23 has the best performance for the dataset and the contour plot shows that error difference is minimum for the same set of values





Q2.

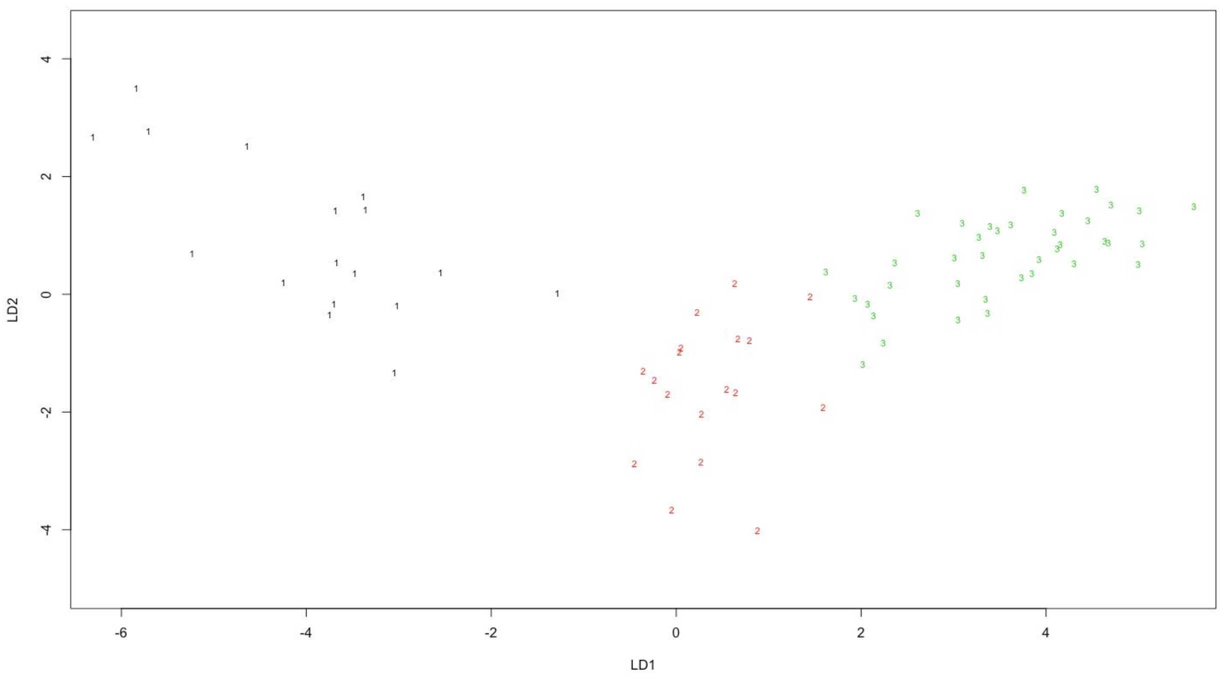
a) The pairwise plots show that distribution of class 1 is anomalous to the other two classes in the dataset for insulin.area and SSPG and hence it has a different covariance matrix compared to the other two classes for these variables thus they're not multivariate normal so the assumption of LDA of common covariance might not suitable and QDA might be a better fit



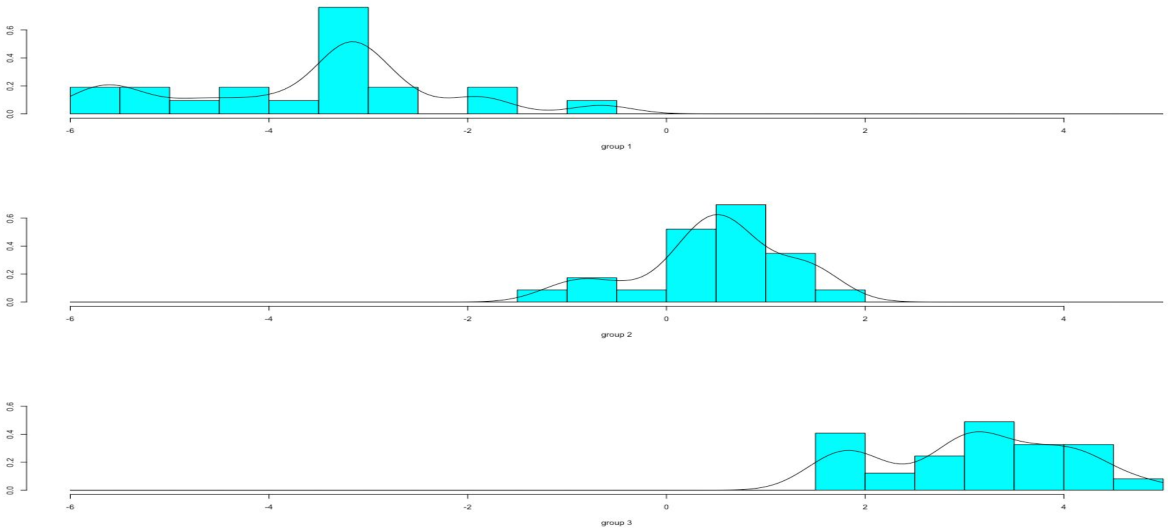
b)

Linear Discriminant Analysis:

We can check for the performance of the 2 linear discriminants that will result as it is a 3 class problem and check for their performance. As we can see from the graph below, LD1 separates the classes quite well but LD2 explains a lot less of the in-between class variance as we can check that the proportion of the trace of eigenvalue matrix for LD1 is 0.899 while for LD2 it is 0.100



We can see from the graph below that the 3 classes have been separated well in the training dataset just using LD1.



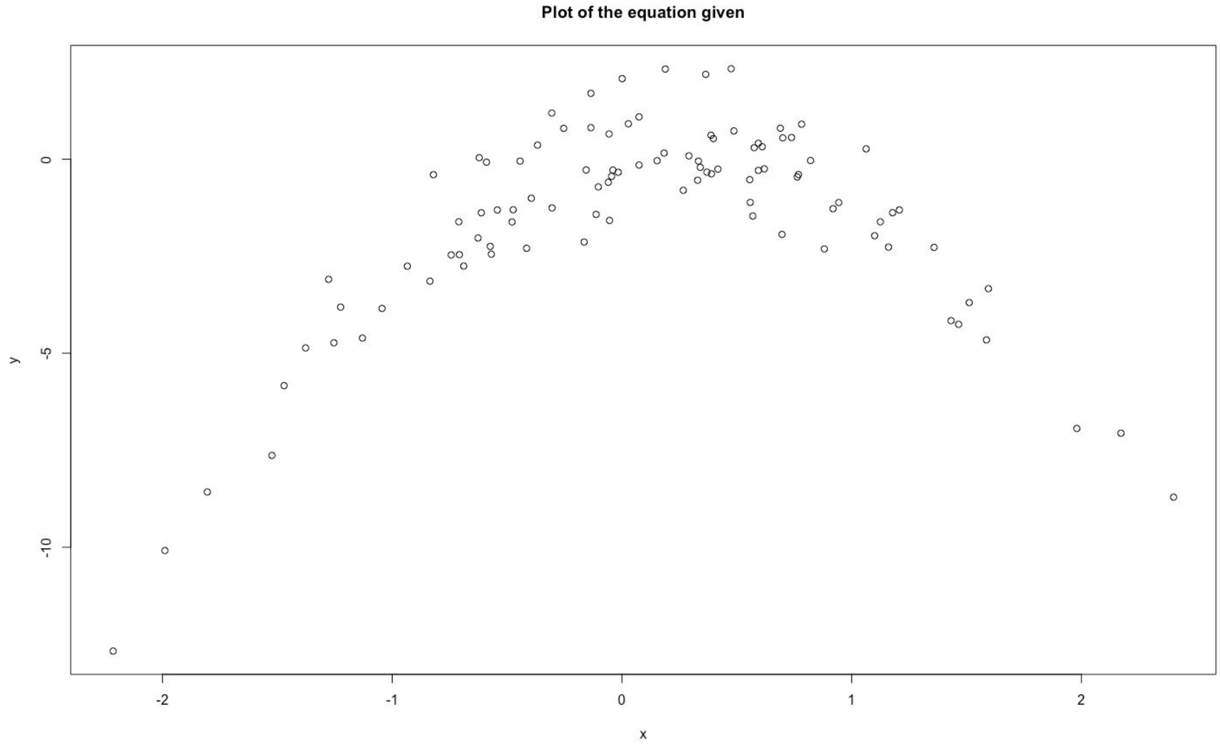
The training accuracy of the model for the dataset is 0.946 and the test accuracy is 0.942

Quadratic Discriminant Analysis:

The training accuracy of the model for the dataset is 0.978 while the test accuracy is 0.942 showing that QDA might be overfitted and has high variance and the LDA might be a better model even if the assumption might be violated as we have few observations in the dataset and hence reducing variance is essential.

c) Entering the values for the data given we can see that QDA predicts class 2 for the give values while LDA predicts 1 for these values which are the mean values of the variables for the dataset.

Q4.



a) The following are the LOOCV errors of the four models:

Model	linear	quadratic	cubic	quartic
Error	7.288	0.937	0.956	0.954

b) As we can see the LOOCV of glm fit for quadratic has the smallest error of 0.937 and the linear one the highest at 7.288. The given equation is a quadratic equation so the best fit for least squares should be one with a quadratic equation.

c) The number of significant coefficients for the model are given below:

Model	linear	quadratic	cubic	quartic
#Significant coefficients	1	2	2	2

These results agree with the conclusions drawn from cross-validation as increasing the degree of the polynomial in the equation has not increased the number of significant variables