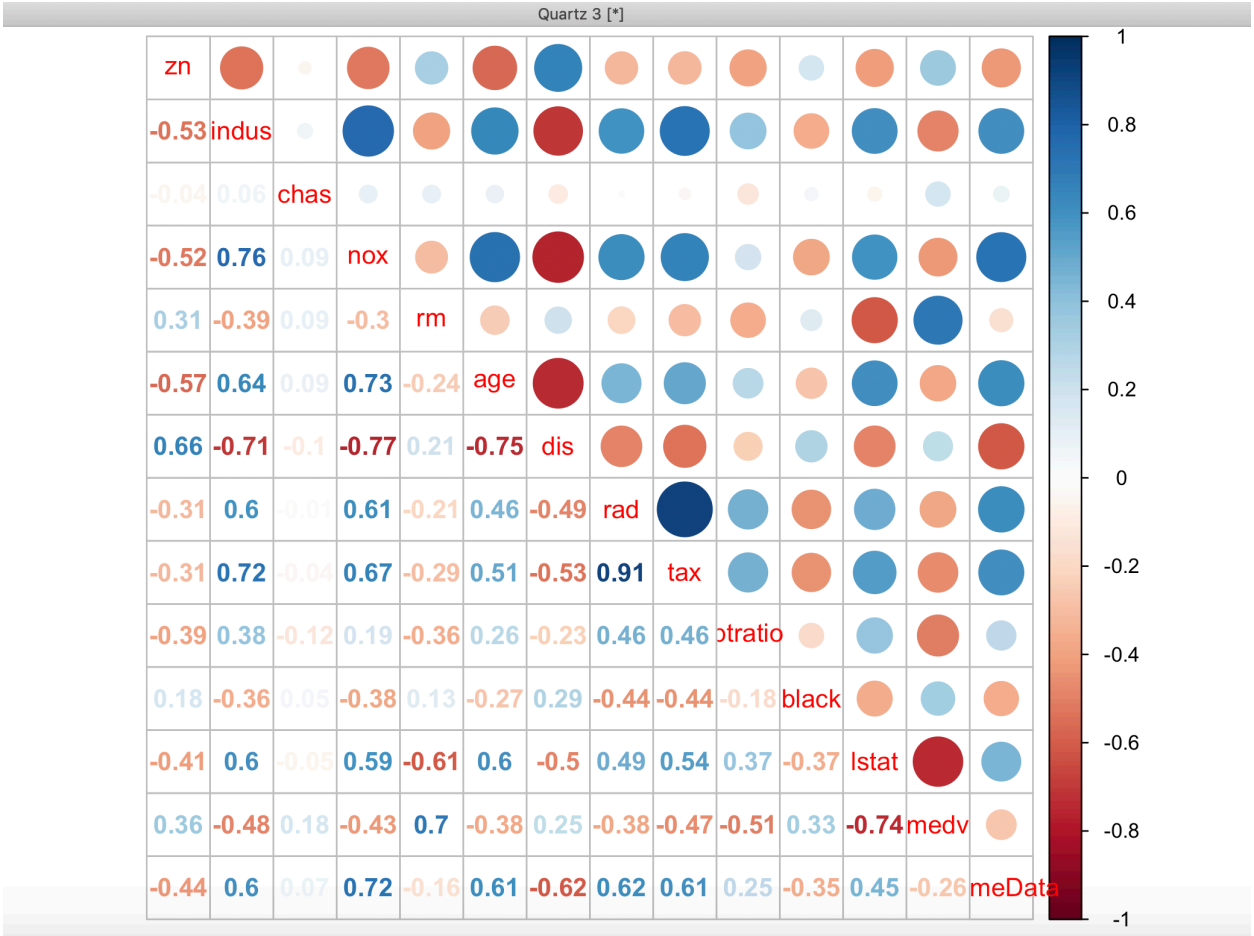
Homework 3 Report

Problem 1:

In this problem, we are given a dataset of Boston. We had to fit the classification models in order to know whether a given suburb has a crime rate above or a below a median.

After loading the data, did a sanity over data by checking the NA values. Then converted the crime feature to categorial value based on median value.

Then splitted data into training and testing set and plotted the correlation matrix to know the relation of other features with crime feature.



As we can see from correlation matrix, some of features indus, nox, age, rad, tax, lstat are strongly correlated to crime feature.

Then fitted the model with indus, nox, age and rad. Here are the results:

```
>
> model1 <- glm(crimeData ~ nox + indus + age + rad, data = BostonData, family = binomial)
>
> probs <- predict(model1, Boston.test, type = "response")
> predModel1 <- rep(0, length(probs))
> predModel1[probs > 0.5] <- 1
> table(predModel1, crimeData.test)
          crimeData.test
predModel1  0  1
         0 68 18
         1  7 59
>
> # linear regression model error
> mean(predModel1 != crimeData.test)
[1] 0.1644737
>
```

The test error for linear regression model is 16%.

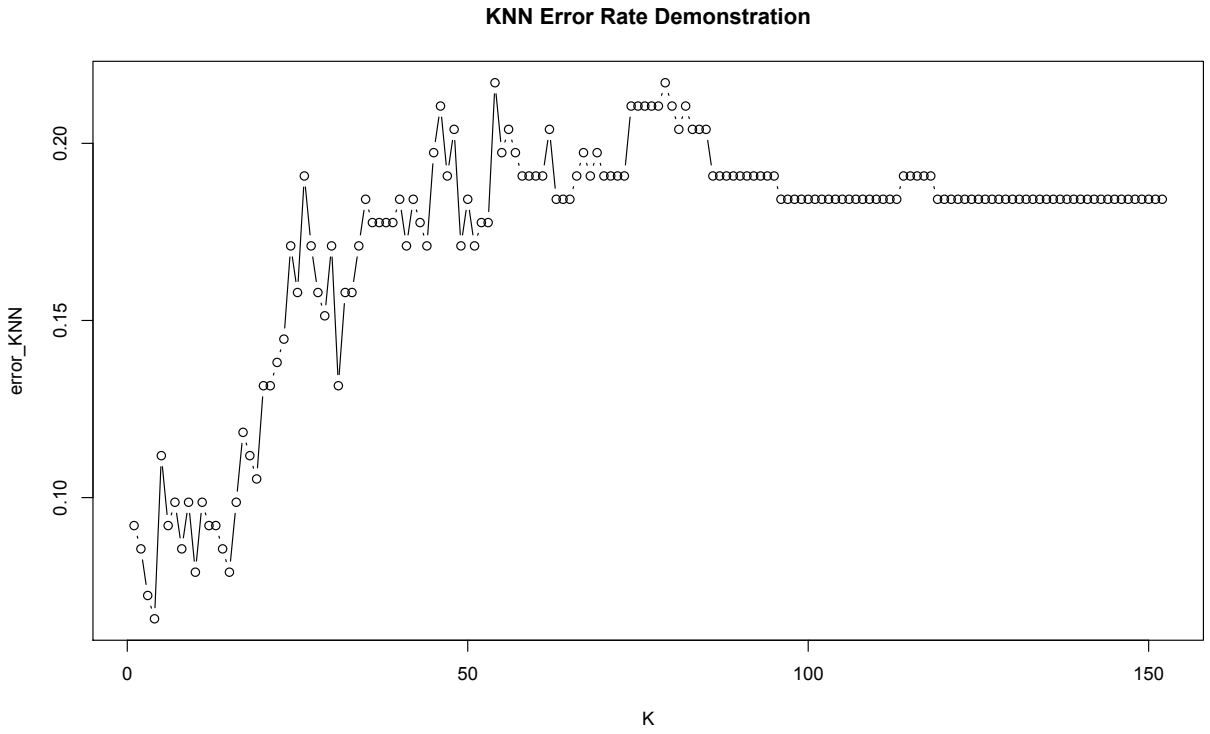
Then the LDA model is fitted.

```
> ldaModel <- lda(crimeData ~ nox + indus + age + rad , data = BostonData)
> predldaModel <- predict(ldaModel, BostonData.test)
> table(predldaModel$class, crimeData.test)
   crimeData.test
     0  1
  0 72 25
  1  3 52
>
> #LDA model error
> mean(predldaModel$class != crimeData.test)
[1] 0.1842105
>
```

The test error for LDA model is 18%.

Then the KNN is performed.

Here is the result for error rate for KNN.



The error rate is minimum for k=4.

And, the minimum error value for k=4 is 0.06578947.

From the above results, we can say that the KNN is performing better than the other features LDA and Linear Regression.

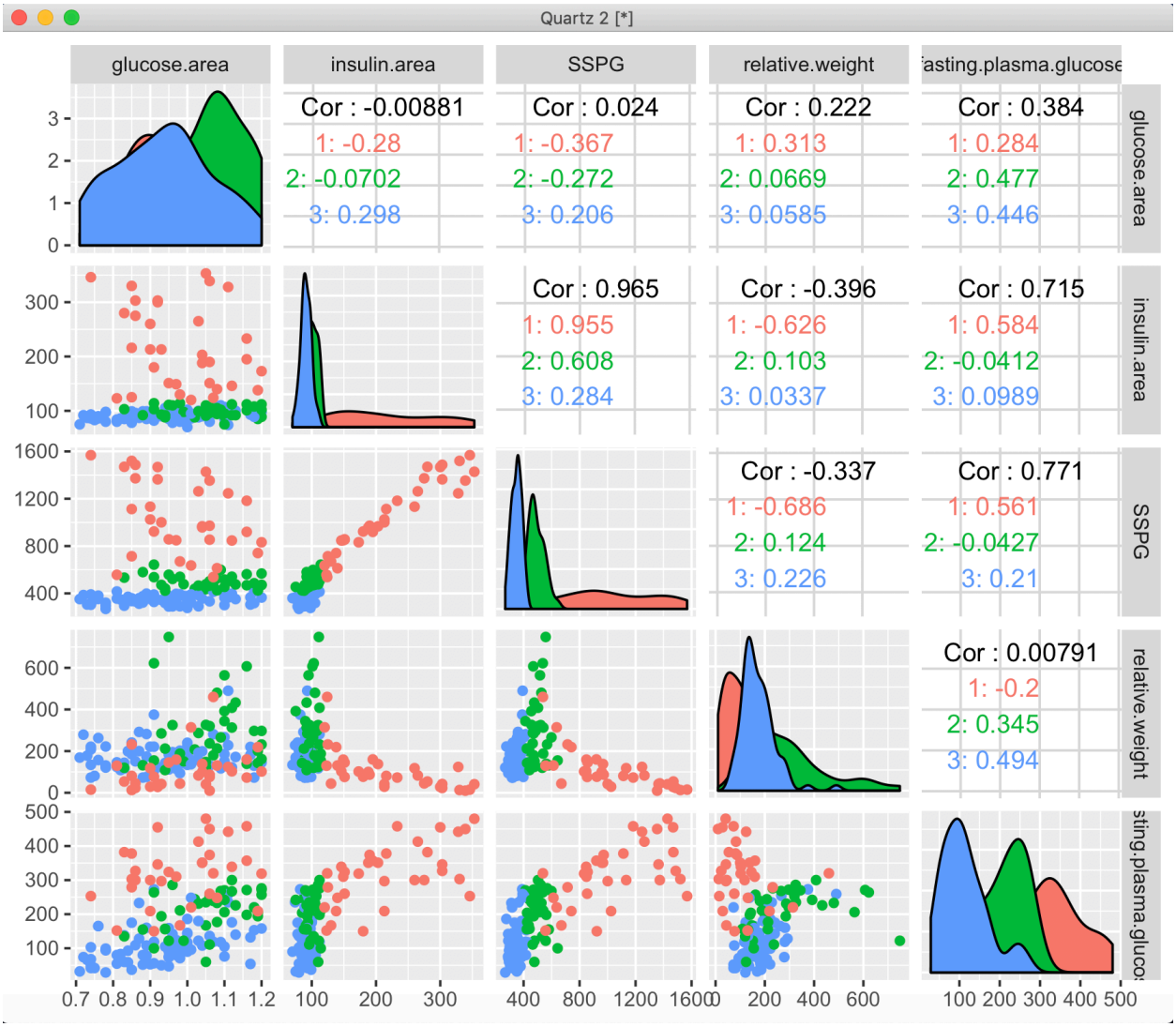
Problem 2:

In this problem, we are given a diabetes dataset. We had to ignore the first three columns of data. After having data, did a sanity check on the data by checking the NA values.
The data has no NA values.

A) Then, plotted a pairwise scatterplot for all five variables with three different colors representing the three different classes.

Here is the result:

All the features are looking normally distributed with right tails.



When looking at the variable SSPG, the distribution of individual classes is different. When we look at the combination of features let say a feature insulin.area and a feature SSPG or glucose.area and insulin.area, it can be inferred that the classes are separable.

When we look at the results of model, we can say that the classes are separable.

B) We performed LDA and QDA to compare the performance of both the model on the dataset.

We get LDA error = 4 and QDA error =5

Error for LDA is less. So it is better for this dataset.

C) We are given the data I.e.
glucose area = 0.98, insulin area =122, SSPG = 544. Relative weight = 186, fasting plasma glucose = 184

To know which class does LDA and QDA assign this.

After predicting, we get

LDA predicted class 3 and QDA predicted class 2.

Problem 4:

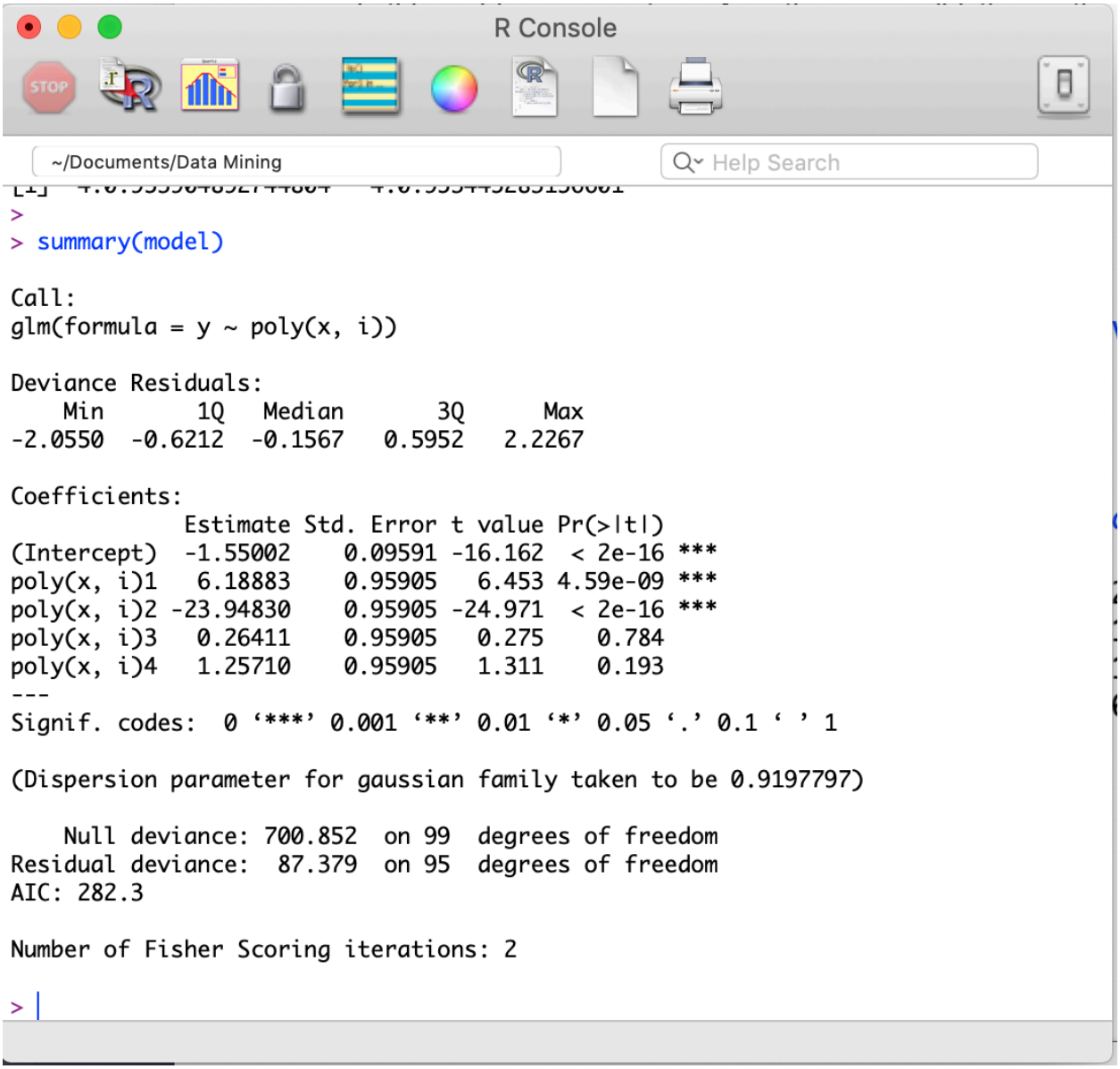
In this problem, we are to perform the cross validation on the simulated dataset.

```
> set.seed(1)
>x=rnorm(100)
 >y=x-2*x^2+rnorm(100)
```

Here is the result from the model:

```
> # creating data frame for computing LOOCV error.
> data = data.frame(x, y);
>
> for(i in 1:4){
+ model = glm(y~poly(x,i))
+ print(paste(i, ":", cv.glm(data, model)$delta, sep = ""))
+ }
[1] "1:7.28816160667281" "1:7.28474411546929"
[1] "2:0.937423637615552" "2:0.937178917181124"
[1] "3:0.95662183010894"  "3:0.956253813731322"
[1] "4:0.953904892744804" "4:0.953445283156601"
>
```

As we can see that the error for 2nd degree polynomial i.e 0.937178917181124 is the minimum. This result was expected because the degree of the polynomial was 2 when we created the data.



Looking at the summary of the model, we see that the coefficients of the model for degree 1 and degree 2 are low which agrees with the cross validation results.

As the coefficients of the model low for the degree 1 and 2 indicates that how well the model fits the data.