Exercise of ISLR

Yamuna Dhungana

1. Question 4.7.3 pg 168 This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a classspecific mean vector and a class specific covariance matrix. We consider the simple case where $p = 1$; i.e. there is only one feature. Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma_k^2)$. Recall that the density function for the one-dimensional normal distribution is given in (4.11). Prove that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic. Hint: For this problem, you should follow the arguments laid out in Section 4.4.2, but without making the assumption that $\sigma_1^2 = ... = \sigma_K^2$.

We may see that finding k for which $p_k(x)$ is largest is equivalent to finding k for which

If we look at 4.4.2 and start working from there. So, the function of 4.4.2 is

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} exp\big(-\frac{1}{2\sigma^2}(x-\mu_k)^2\big)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} exp\big(-\frac{1}{2\sigma^2}(x-\mu_l)^2\big)}$$

Here, the denominator is equal to all $1..k_{th}$ classes, so we can ignore it,

$$f'_x = \pi_k \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2)$$

And taking natural logarithm it becomes:

$$f''_x = \ln \pi_k + \ln(\frac{1}{\sqrt{2\pi}\sigma}) + \ln exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2)$$

or,

$$f''_x = \ln(\frac{\pi_k}{\sqrt{2\pi}\sigma_k}) - \frac{1}{2\sigma_k^2}(x-\mu_k)^2$$

$$f''_x = \ln(\frac{\pi_k}{\sqrt{2\pi}\sigma_k}) - \frac{(x^2 - 2x\mu_k + \mu_k^2)}{2\sigma_k^2}$$

Here, the first term is a distinct constant for each class, and the second one is the quadratic function of x.

2. Question 4.7.5 pg 169 We now examine the differences between LDA and QDA.

a. If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

We can expect QDA to perform better in the training data because QDA is more flexible Since LDA does not overfit therefore, it performs better with the test data.

b. If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

If the decision boundary is non-linear, we would expect QDA to perform better for both training and the test set because of its flexibility.

c. In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

As the sample size n increases, we expect the test prediction accuracy of QDA to improve relative to LDA because of its higher flexibility.

d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer

I think the statement is False. If the Bayes decision boundary is linear, we would expect higher test error in QDA than LDA. After all, QDA overfits and overfitted data produces higher test errors.

3. Continue from Homework #3 Question 4.7.10(e-i) pg 171

e. Repeat (d) using LDA.

```
## [1] "Statistics for the LDA"
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    9  5
##   Up     34 56
## [1] "Model Accuracy (Percentage):"
## [1] 62.5
## [1] "True Positive Rate, TPR (percentage):"
## [1] 91.8
## [1] "False Postive Rate, FPR (percentage):"
## [1] 79.07
```

In this case too I have used a threshold of 0.5. The accuracy of the model is 62.5% that means 37.5% is the test error for the LDA model. By looking at the confusion table, 9 of the down data was correctly predicted, and 56 of Up data was predicted correctly. We also can say that when the market goes up the prediction is correct is for 91.8 % (56/(56+5)).The market when the market goes down we were correct for 20.9% of the time (9/(9+34)). The false-positive rate for the model is 79.07.

f. Repeat (d) using QDA

```
## [1] "Statistics for the QDA"
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    0  0
##   Up     43 61
## [1] "Model Accuracy (Percentage):"
## [1] 58.65
## [1] "True Positive Rate, TPR (percentage):"
## [1] 100
## [1] "False Postive Rate, FPR (percentage):"
## [1] 100
```

For the QDA model, I used thresholds 0.5 which is the same as the logistic and LDA models.However, unlike them, QDA predicts all the data for up leaving zero predictions for the down which is not good. The result of zero in the down row leaves the true positive and False positive rates at 100%. The accuracy of the model is still above 50% that is 58.65 %.

g. Repeat (d) using KNN with K = 1

```
## [1] "Confusion Matrix:"
##       trues
## model  Down Up
##   Down   21 29
##   Up     22 32
## [1] "Model Accuracy (Percentage):"
```

```
## [1] 50.96
## [1] "True Positive Rate, TPR (percentage):"
## [1] 52.46
## [1] "False Postive Rate, FPR (percentage):"
## [1] 51.16
```

When the K= 1 in the KNN model we have 50.96% of the model accuracy which means half of the data was predicted incorrectly. The true positive rate is 52.46 and the False positive rate is 51.16%. The Test errors of the model are comparatively low from the other models.Hence we can say that KNN did not do a good job in the model when K = 1.

h. Which of these methods appears to provide the best results on this data?

By looking at the test error rates, we see that logistic regression and LDA have the minimum error rates, followed by QDA and KNN. Hence we can say Logistic and LDA performed better.

i. Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    4  6
##   Up     39 55
## [1] "Model Accuracy (Percentage):"
## [1] 56.73
## [1] "True Positive Rate, TPR (percentage):"
## [1] 90.16
## [1] "False Postive Rate, FPR (percentage):"
## [1] 90.7
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    9  5
##   Up     34 56
## [1] "Model Accuracy (Percentage):"
## [1] 62.5
## [1] "True Positive Rate, TPR (percentage):"
## [1] 91.8
## [1] "False Postive Rate, FPR (percentage):"
## [1] 79.07
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    7  8
##   Up     36 53
## [1] "Model Accuracy (Percentage):"
## [1] 57.69
## [1] "True Positive Rate, TPR (percentage):"
## [1] 86.89
## [1] "False Postive Rate, FPR (percentage):"
## [1] 83.72
```

```
## [1] "Confusion Matrix:"
```

```
##
## preds  Down Up
##   Down    7  8
##   Up     36 53
## [1] "Model Accuracy (Percentage):"
## [1] 57.69
## [1] "True Positive Rate, TPR (percentage):"
## [1] 86.89
## [1] "False Postive Rate, FPR (percentage):"
## [1] 83.72
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    6  5
##   Up     37 56
## [1] "Model Accuracy (Percentage):"
## [1] 59.62
## [1] "True Positive Rate, TPR (percentage):"
## [1] 91.8
## [1] "False Postive Rate, FPR (percentage):"
## [1] 86.05
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down   23 33
##   Up     20 28
## [1] "Model Accuracy (Percentage):"
## [1] 49.04
## [1] "True Positive Rate, TPR (percentage):"
## [1] 45.9
## [1] "False Postive Rate, FPR (percentage):"
## [1] 46.51
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down   20 25
##   Up     23 36
## [1] "Model Accuracy (Percentage):"
## [1] 53.85
## [1] "True Positive Rate, TPR (percentage):"
## [1] 59.02
## [1] "False Postive Rate, FPR (percentage):"
## [1] 53.49
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    7  4
##   Up     36 57
## [1] "Model Accuracy (Percentage):"
## [1] 61.54
## [1] "True Positive Rate, TPR (percentage):"
## [1] 93.44
```

```
## [1] "False Postive Rate, FPR (percentage):"
## [1] 83.72
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    0  0
##   Up     43 61
## [1] "Model Accuracy (Percentage):"
## [1] 58.65
## [1] "True Positive Rate, TPR (percentage):"
## [1] 100
## [1] "False Postive Rate, FPR (percentage):"
## [1] 100
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    0  0
##   Up     43 61
## [1] "Model Accuracy (Percentage):"
## [1] 58.65
## [1] "True Positive Rate, TPR (percentage):"
## [1] 100
## [1] "False Postive Rate, FPR (percentage):"
## [1] 100
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    7 10
##   Up     36 51
## [1] "Model Accuracy (Percentage):"
## [1] 55.77
## [1] "True Positive Rate, TPR (percentage):"
## [1] 83.61
## [1] "False Postive Rate, FPR (percentage):"
## [1] 83.72
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down   23 36
##   Up     20 25
## [1] "Model Accuracy (Percentage):"
## [1] 46.15
## [1] "True Positive Rate, TPR (percentage):"
## [1] 40.98
## [1] "False Postive Rate, FPR (percentage):"
## [1] 46.51
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    3 11
##   Up     40 50
```

```
## [1] "Model Accuracy (Percentage):"
## [1] 50.96
## [1] "True Positive Rate, TPR (percentage):"
## [1] 81.97
## [1] "False Postive Rate, FPR (percentage):"
## [1] 93.02
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down   31 42
##   Up     12 19
## [1] "Model Accuracy (Percentage):"
## [1] 48.08
## [1] "True Positive Rate, TPR (percentage):"
## [1] 31.15
## [1] "False Postive Rate, FPR (percentage):"
## [1] 27.91
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down   32 44
##   Up     11 17
## [1] "Model Accuracy (Percentage):"
## [1] 47.12
## [1] "True Positive Rate, TPR (percentage):"
## [1] 27.87
## [1] "False Postive Rate, FPR (percentage):"
## [1] 25.58
```

```
## [1] "Confusion Matrix:"
##
## preds  Down Up
##   Down    7  3
##   Up     36 58
## [1] "Model Accuracy (Percentage):"
## [1] 62.5
## [1] "True Positive Rate, TPR (percentage):"
## [1] 95.08
## [1] "False Postive Rate, FPR (percentage):"
## [1] 83.72
```

```
## [1] "########################################"
## [1] "K = 1"
## [1] "Confusion Matrix:"
##       trues
## model  Down Up
##   Down   21 29
##   Up     22 32
## [1] "Model Accuracy (Percentage):"
## [1] 50.96
## [1] "True Positive Rate, TPR (percentage):"
## [1] 52.46
## [1] "False Postive Rate, FPR (percentage):"
## [1] 51.16
```

```
## [1] "########################################"
## [1] "########################################"
## [1] "K = 3"
## [1] "Confusion Matrix:"
##       trues
## model  Down Up
##   Down   15 20
##   Up     28 41
## [1] "Model Accuracy (Percentage):"
## [1] 53.85
## [1] "True Positive Rate, TPR (percentage):"
## [1] 67.21
## [1] "False Postive Rate, FPR (percentage):"
## [1] 65.12
## [1] "########################################"
## [1] "########################################"
## [1] "K = 5"
## [1] "Confusion Matrix:"
##       trues
## model  Down Up
##   Down   15 21
##   Up     28 40
## [1] "Model Accuracy (Percentage):"
## [1] 52.88
## [1] "True Positive Rate, TPR (percentage):"
## [1] 65.57
## [1] "False Postive Rate, FPR (percentage):"
## [1] 65.12
## [1] "########################################"
## [1] "########################################"
## [1] "K = 10"
## [1] "Confusion Matrix:"
##       trues
## model  Down Up
##   Down   18 17
##   Up     25 44
## [1] "Model Accuracy (Percentage):"
## [1] 59.62
## [1] "True Positive Rate, TPR (percentage):"
## [1] 72.13
## [1] "False Postive Rate, FPR (percentage):"
## [1] 58.14
## [1] "########################################"
## [1] "########################################"
## [1] "K = 20"
## [1] "Confusion Matrix:"
##       trues
## model  Down Up
##   Down   20 20
##   Up     23 41
## [1] "Model Accuracy (Percentage):"
## [1] 58.65
## [1] "True Positive Rate, TPR (percentage):"
## [1] 67.21
```

```
## [1] "False Postive Rate, FPR (percentage):"
## [1] 53.49
## [1] "########################################"
## [1] "########################################"
## [1] "K = 50"
## [1] "Confusion Matrix:"
##       trues
## model  Down Up
##   Down   20 24
##   Up     23 37
## [1] "Model Accuracy (Percentage):"
## [1] 54.81
## [1] "True Positive Rate, TPR (percentage):"
## [1] 60.66
## [1] "False Postive Rate, FPR (percentage):"
## [1] 53.49
## [1] "########################################"
## [1] "########################################"
## [1] "K = 75"
## [1] "Confusion Matrix:"
##       trues
## model  Down Up
##   Down   12 13
##   Up     31 48
## [1] "Model Accuracy (Percentage):"
## [1] 57.69
## [1] "True Positive Rate, TPR (percentage):"
## [1] 78.69
## [1] "False Postive Rate, FPR (percentage):"
## [1] 72.09
## [1] "########################################"
## [1] "########################################"
## [1] "K = 100"
## [1] "Confusion Matrix:"
##       trues
## model  Down Up
##   Down   10 12
##   Up     33 49
## [1] "Model Accuracy (Percentage):"
## [1] 56.73
## [1] "True Positive Rate, TPR (percentage):"
## [1] 80.33
## [1] "False Postive Rate, FPR (percentage):"
## [1] 76.74
## [1] "########################################"
```

Here, I have made 8 different combinations of the models and later performed LDA and QDA. The confusion matrices and the test errors are printed above. For LDA the accuracy of the second model looks good with the model accuracy of 62.5%. Also, we can see that the model has a higher True positive rate and comparatively lesser false positive rate. Likewise, for QDA, the First two models predict all zeros for the down, which is bad. The accuracy of model 8 looks good with the high accuracy and high true positive rates. The value of K for the KNN model is chosen randomly and for this specific model, It looks like the higher value of K gives the high accurate model with fewer test errors.

4. Continue from Homework #3 Question 4.7.11(d,e,g) pg 172

d. Perform LDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

```
## [1] "Confusion Matrix:"
##      trues
## preds  0  1
##     0 49  2
##     1 12 54
## [1] "Model Accuracy (Percentage):"
## [1] 88.03
## [1] "True Positive Rate, TPR (percentage):"
## [1] 96.43
## [1] "False Postive Rate, FPR (percentage):"
## [1] 19.67
```

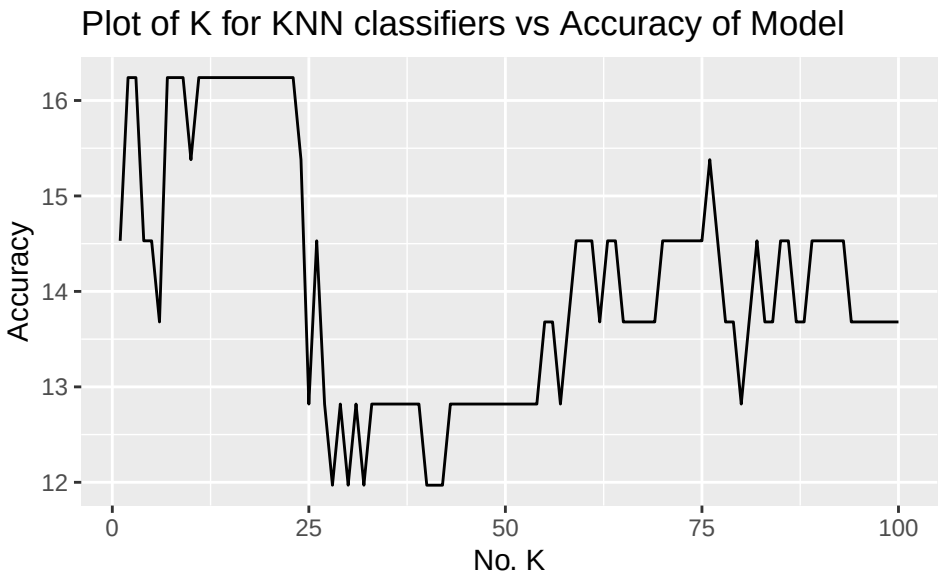
I found out that the variables cylinders, weight, displacement, horsepower were mostly associated with the mpg01, therefore, performing LDA with the same variable. The confusion matrix shows impressive results. The accuracy of the model is 92.31% which is very good for the model. Likewise, we have a true positive rate of 91.8% which is also a good result.Additionally, the preferable false positive rate is less than 10% and we have 7.14%. Hence we can say that LDA performed well for this model.

e. Perform QDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

```
## [1] "Confusion Matrix:"
##      trues
## preds  0  1
##     0 52  2
##     1  9 54
## [1] "Model Accuracy (Percentage):"
## [1] 90.6
## [1] "True Positive Rate, TPR (percentage):"
## [1] 96.43
## [1] "False Postive Rate, FPR (percentage):"
## [1] 14.75
```

Here I used the same function to perform QDA, The results of the QDA somewhat similar to the LDA model. The accuracy of the QDA model is 90.6 which is generally good but is still less than LDA. The true positive rate for the model is 86.89 % however, we have less false positive rate than the LDA. And, a less false positive rate is always preferred.

g. Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?




```
##   k   acc
## 2 2 16.24
## 3 3 16.24
## 7 7 16.24
```

For the value of K, I chose to run 1 to 100 values of K and perform the test errors. By looking at the test error I will choose the optimal value of K that gives the highest accuracy. To illustrate, the value of K and accuracy is drawn. Additionally, to make it clear I have printed the 3 accuracies of the value.

5. Read the paper "Statistical Classification Methods in Consumer Credit Scoring: A Review" posted on D2L. Write a one page (no more, no less) summary.

   ```
   Summary of "Statistical Classification Methods in Consumer Credit Scoring: A Review"
   ```

The Paper "Statistical Classification Methods in Consumer Credit Scoring: A Review" was published by D.J Hand and WE Henley in October 1996. The paper is about the statistical method that is used to classify the applicants for credit into good and bad risk classes.

The paper is mainly focused on the methods for classifying an applicant according to their payment behavior which is default and not defaults. Also, including other problems associated with the credit industry. During the application process, applicants are suggested to provide the information, based on that information, the probability that the applicant will pay or will not pay is determined. Based on this decision the application is either accepted or rejected. Whereas,to analyze the mutualism and whether to grant credit or not is analyzed by the bank through classification method. Different statistical tool that has been used to determine the process are discriminant analysis, linear regression, logistic regression, and decision tree.

Based on the data collected from the customer's payment behavior, the bank decides to classify them into two classes- good and bad risk. However, the common classes that have been using in the credit industry are of three classes- good, bad, and intermediate and use only two of the extreme classes to determine the scorecard. Using this data, new applicants are classified as good or the bad risk (This is different from classifying new applicants as 'bad', 'good' or 'not yet known'- and then seeking further information on the last group). For example, good risks might be those borrowers who have never been in arrears, bad risks might be those who have been three or more consecutive repayments in arrears at some point during the period in question and indeterminate might be those who have been in arrears either for one or for two consecutive repayments. The bank is only interested in whether providing the loan might or might not be profitable. Never in arrears defines a good risk and would be profitable and three months in arrears can be considered as a bad risk and would not be profitable. Likewise, depending upon the economic changes at the

time of seeking a loan, indeterminate may or may not turns profitable. For this indeterminate, the standard method sets a rule which will identify whether the deal will be profitable or unprofitable.

The author mentioned that credit scoring has a large database and consists of 100000 or more applicants and as many as 100 variables. Using these data decision of selecting the applicant is randomly selected whose risk as low as around 5%. Different model's performance was examined. A different model such as discriminant analysis, regression, logistic regression, recursive partitioning, neural networks, smoothing nonparametric methods, and more were analyzed. Each of the models performed relatively similarly beside the neural network which was a bit vast.

6. Explore this website that contains open data sets that are used in machine learning. Select a data set that has classification as a Default Task and describe, in your own words, the task, including a description of the data set. Look for data sets that are amenable to the analyses we have learned thus far. Pay attention to the characteristics of the data with selecting an analysis method. I do not expect you to do the analysis for this homework, but feel free to if you want!

`Analysis of quality of wine`

The dataset used in this project is a red wine quality dataset. This dataset consists of 12 variables and 1599 observations. The dataset consists of a collection of variables that may have affected the quality of the wine. I am aiming to find the variable(s) which contribute the most to the quality of the wine. We are also trying to predict a wine's quality. I have choosen this data because it is similar to the data we analysed.

Exploring basic data statistics

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
```

```
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

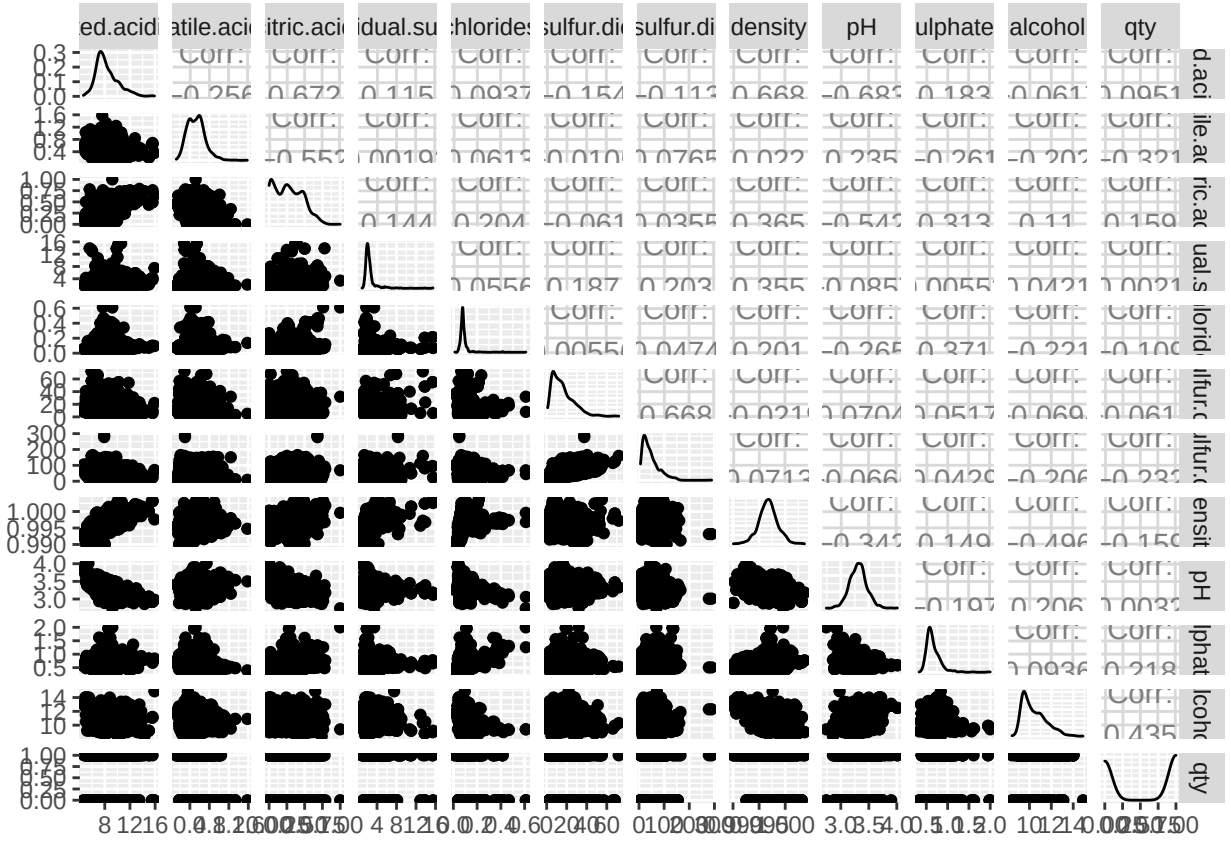
The quality of the wine which is rated from 1 to 10 initially. Here, I have changed the quality of wine that is less than or equal to 5 as low and mention in the data as Zero(0) and quality of wine greater than 5 as high and mentioned in data as one (1).

Now, I want to find which variable is mostly correlated with the wine data.

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity           1.00000000     -0.256130895  0.67170343    0.114776724
## volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882
## citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162
## residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000
## chlorides               0.09370519      0.061297772  0.20382291    0.055609535
## free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995
## total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882
## density                 0.66804729      0.022026232  0.36494718    0.355283371
## pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422
## sulphates               0.18300566     -0.260986685  0.31277004    0.005527121
## alcohol                -0.06166827     -0.202288027  0.10990325    0.042075437
## qty                     0.09509349     -0.321440854  0.15912941   -0.002160450
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.093705186        -0.153794193          -0.11318144
## volatile.acidity      0.061297772        -0.010503827           0.07647000
## citric.acid           0.203822914        -0.060978129           0.03553302
## residual.sugar        0.055609535         0.187048995           0.20302788
## chlorides             1.000000000         0.005562147           0.04740047
## free.sulfur.dioxide   0.005562147         1.000000000           0.66766645
## total.sulfur.dioxide  0.047400468         0.667666450           1.00000000
## density               0.200632327        -0.021945831           0.07126948
## pH                   -0.265026131         0.070377499          -0.06649456
## sulphates             0.371260481         0.051657572           0.04294684
## alcohol              -0.221140545        -0.069408354          -0.20565394
## qty                  -0.109493996        -0.061756744          -0.23196298
##                          density           pH    sulphates     alcohol
## fixed.acidity         0.66804729 -0.682978195  0.183005664 -0.06166827
## volatile.acidity      0.02202623  0.234937294 -0.260986685 -0.20228803
## citric.acid           0.36494718 -0.541904145  0.312770044  0.10990325
## residual.sugar        0.35528337 -0.085652422  0.005527121  0.04207544
## chlorides             0.20063233 -0.265026131  0.371260481 -0.22114054
## free.sulfur.dioxide  -0.02194583  0.070377499  0.051657572 -0.06940835
## total.sulfur.dioxide  0.07126948 -0.066494559  0.042946836 -0.20565394
## density               1.00000000 -0.341699335  0.148506412 -0.49617977
## pH                   -0.34169933  1.000000000 -0.196647602  0.20563251
## sulphates             0.14850641 -0.196647602  1.000000000  0.09359475
## alcohol              -0.49617977  0.205632509  0.093594750  1.00000000
## qty                  -0.15910997 -0.003263984  0.218071663  0.43475120
##                              qty
## fixed.acidity         0.095093490
## volatile.acidity     -0.321440854
## citric.acid           0.159129408
## residual.sugar       -0.002160450
## chlorides            -0.109493996
```

```
## free.sulfur.dioxide  -0.061756744
## total.sulfur.dioxide -0.231962976
## density              -0.159109969
## pH                   -0.003263984
## sulphates             0.218071663
## alcohol               0.434751205
## qty                   1.000000000
```

```
##                  qty              alcohol            sulphates
##                 TRUE                 TRUE                FALSE
##          citric.acid        fixed.acidity       residual.sugar
##                FALSE                FALSE                FALSE
##                   pH  free.sulfur.dioxide            chlorides
##                FALSE                FALSE                FALSE
##              density total.sulfur.dioxide     volatile.acidity
##                FALSE                FALSE                 TRUE
```



I have decided to find those variables whose correlation coefficient is greater than 0.3. From the correlation looks like volatile acidity and alcohol seem mostly correlated with the quality of the wine. Alcohol is positively correlated with the positive correlation whereas, volatile acidity has the negative correlation coefficient.

Splitting data

I have split the data into training and testing in the ratio of 60% to 40% with the library function caTools. I have decided to run the 3 models to check the impact of variables on the quality of the wine. First model is Logistic regression.

With Logistic Regression

```
##
## Call:
## glm(formula = qty ~ volatile.acidity + alcohol, family = binomial,
##     data = tr.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4328  -0.9220   0.3437   0.9038   2.2863
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -8.52994    0.95064  -8.973  < 2e-16 ***
## volatile.acidity -2.94369    0.45148  -6.520 7.03e-11 ***
## alcohol           0.99290    0.08844  11.227  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1286.0  on 932  degrees of freedom
## Residual deviance: 1029.2  on 930  degrees of freedom
## AIC: 1035.2
##
## Number of Fisher Scoring iterations: 4
```

```
##
## preds   0   1
##     0 247 103
##     1  72 244
## [1] "True Positive Rate, TPR (percentage):"
## [1] 70.32
## [1] "False Postive Rate, FPR (percentage):"
## [1] 22.57
```

Since volatile acidity and alcohol are mostly associated with the quality of the data. I have fitted the logistic model with the same. From the logistic model, it appears that both the volatile acidity and alcohol are statistically significant. The estimated coefficient of volatile acidity is -3.02073 that means, when the other predictors in the model are constant, we would expect a mean decrease in log-odds by the unit increase in quality of the wine. Also, The estimated coefficient of alcohol is 1.10115 that means, when the other predictors in the model are constant, we would expect a mean increase in log-odds by the unit increase in quality of the wine. In the confusion matrix of the logistic regression, the test accuracy of the model is 72.11%, and the true the positive rate of the model is 70.9 and the False positive rate of the model is 26.52 which is good.

With LDA

```
## [1] "Statistics for the LDA"
```

```
## [1] "Confusion Matrix:"
##
## preds   0   1
##     0 251 105
##     1  68 242
## [1] "Model Accuracy (Percentage):"
```

```
## [1] 74.02
## [1] "True Positive Rate, TPR (percentage):"
## [1] 69.74
## [1] "False Postive Rate, FPR (percentage):"
## [1] 21.32
```

The LDA model shows that the logistic regression and LDA have similar results. with LDA model accuracy, true positive and false positive rates are almost the same.

With QDA

```
## [1] "Statistics for the QDA"
```

```
## [1] "Confusion Matrix:"
##
## preds   0   1
##     0 262 133
##     1  57 214
## [1] "Model Accuracy (Percentage):"
## [1] 71.47
## [1] "True Positive Rate, TPR (percentage):"
## [1] 61.67
## [1] "False Postive Rate, FPR (percentage):"
## [1] 17.87
```

With the QDA model, The model accuracy is 71.06 which is a little less than the other models. The true positive rate is 65.54 which is also less than the other models however, the false positive rate is 22.68 which is 5% and 2% less than the other models and is considered better.