Project 2

MACT423301- MACT423302- Applied Multivariate Analysis (Spring 2023)

Group members:

Maya Hany Elshweiky 900204233

Omar Ahmed 900201077

Dr. Ali Hadi

Table of contents:

Red Wine Quality Data Analysis

1. Statement of the problem:

1.1 What question(s) can be answered by the analysis of the data?

The following questions can be answered from analyzing the data:

- Which factors have the strongest correlation with wine quality ratings?
- Does Fisher Linear Discriminant Analysis (FLDA) separate our data correctly with regards to its quality?
- How good/accuarate is our analysis and method of discrimination
- Can we predict the quality rating of a wine based on its physicochemical properties?

1.2 Background Information

We will analyze the "Wine Quality" data set to assess the questions above, the Data set contains the following variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, PH level, sulfates, alcohol level and quality of wine; the units of measurements for the listed variables is mentioned below; all of the variables are quantitative variables except for the "quality of wine" variable. We will add another categorical variable named "score" and this will be based on the quality of wine. We will split the data into **three groups;** very bad, bad and good. We will investigate the questions mentioned above by applying Fisher Linear Discriminant Analysis (FLDA). Firstly, after splitting the data, we will check if our data is normally distributed or not, if not, we will scale them and will try to standardize our data. Noting that FLDA assumes that the covariances are equal. Then we will apply our discriminant analysis on our data. Starting with FLDA, our function will give us an output for the discriminant analysis and the misclassification error, which is the internal validation, and it allows us to check the validity of our method (how accurate it is), the misclassification error is (the total misclassified observations) divided by the number of observations. Then we will check the validity using external validation (Leave one out); the leave one out method uses n-1 of the observations to derive a rule and then applies this rule to predict our left out observation's class, we repeat this method n times. We will then use FLDA2, which is a special case of FLDA and is used when we have two groups and two variables only, we will use extract groups 1 and 3 and apply FLDA2, our output will include the discriminant analysis, misclassification error and the plot of our data.

2. Data Description

Who: 1599 observations of wine quality based on physicochemical tests

When: The data was collected in 2009

Where: The data was obtained from kaggle.com and the Citation for the data set is found below:

Source : https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

Citation: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

What: There are 13 variables in the data, for each we have the wine quality (scored between 0 and 10), the score of wine quality (1 or 2 or 3) and eleven chemical attributes (quantitative), and their description is shown below in the table.

Variable	**Type**	**Unit of measurement**	**Description**
fixed acidity	Quantitative (Numeric)	tartaric acid - g / dm^3	Fixed acids, numeric from 3.8 to 15.9
volatile acidity	Quantitative (Numeric)	acetic acid - g / dm^3	Volatile acids, numeric from 0.1 to 1.6
citric acid	Quantitative (Numeric)	g / dm^3	Citric acids, numeric from 0.0 to 1.7
residual sugar	Quantitative (Numeric)	g / dm^3	residual sugar, numeric from 0.6 to 65.8
chlorides	Quantitative (Numeric)	sodium chloride - g / dm^3	numeric from 0.01 to 0.61
free sulfur dioxide	Quantitative (Numeric)	mg / dm^3	Free sulfur dioxide, numeric: from 1 to 289
total sulfur dioxide	Quantitative (Numeric)	mg / dm^3	Total sulfur dioxide, numeric: from 6 to 440

density	Quantitative (Numeric)	g / dm^3	Density, numeric: from 0.987 to 1.039
pH	Quantitative (Numeric)	0-14 scale (0 being most acidic, 14 most alkaline)	pH, numeric: from 2.7 to 4.0
sulphates	Quantitative (Numeric)	(potassium sulphate - g / dm3	Sulfates, numeric: from 0.2 to 2.0
alcohol	Quantitative (Numeric)	% by volume	the percent alcohol content of the wine, numeric: from 8.0 to 14.9
quality	Categorical	0-10 scale	Wine quality rating between 0 (very bad) and 10 (very excellent)
score[1]	Categorical (binary)	-	**- Score 1** represents **very bad quality wine** in which the quality rating is lower than 4 **-Score 2** represents **bad quality wine** in which the quality

[1] Note: score variable was added using excel to split quality rating into three categories

			rating is between 4 and 6.5 **-Score 3** represents **good wine quality** in which the quality rating exceeds 6.5

2.1 Why did we choose this Dataset?

We choose the Red Wine Quality Dataset as it satisfies the requirements of a sufficient dataset to be examined. This dataset consists of 1599 observations and 13 different variables. We added a variable named “score”. This data is accurate and relevant as it contains all information necessary for data analysis and does not contain any “NA’s”. Therefore, the data is suitable for multivariate analysis as it contains different variables.

3. Data Visualization and Analysis

3.1 Splitting Data into three groups

We split the data into three groups based on the quality rating of wine where group 1 represents very bad quality wine, group 1 is classified as any quality that is below the value of 4, group 2 represents bad quality wine which is based on values of quality from 4 to 6.5, lastly group 3 represents good quality wine and it includes any value of quality that is above 6.5. In group 1, there were 10 observations, group 2 contained 1372 observations, and group 3 consisted of 217 observations.

3.2 The Relationship Between Variables

As shown in figure 1 below, we can see that many of the variables have linear relationships with each other, for example, fixed acidity and density have positive linear relationship, while fixed acidity and PH have a negative linear relationship. There are also several variables that do not have linear relationships with each other, such as fixed acidity and sulfur dioxide, residual sugar and sulfur dioxide, residual sugar and fixed acidity, and so much more. In addition to that, we can see the separation of the three groups, there are many overlaps between the groups, and in some plots we cannot see the group colored in black, but there is a small proportion of the plots that the data are separated well in, such as volatile acidity and chlorides. Lastly, we can see that there are some outliers in the data set, as we can see points on the graph that are on the edges of the plot, and are far away from the rest of the data.

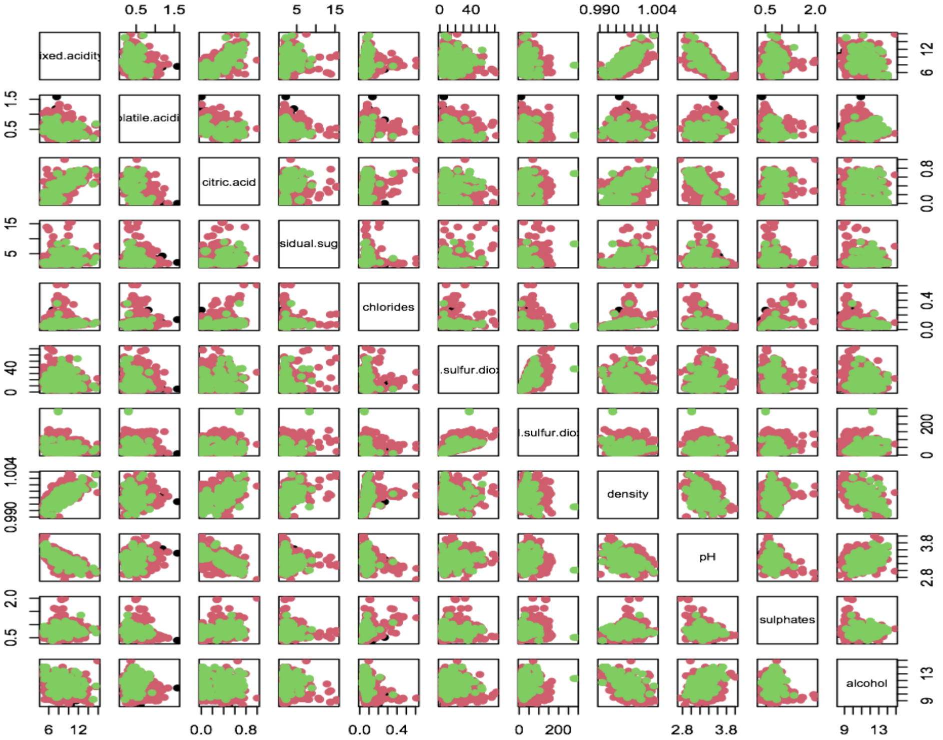


Figure 1: pairs of variables of the three groups before scaling the data

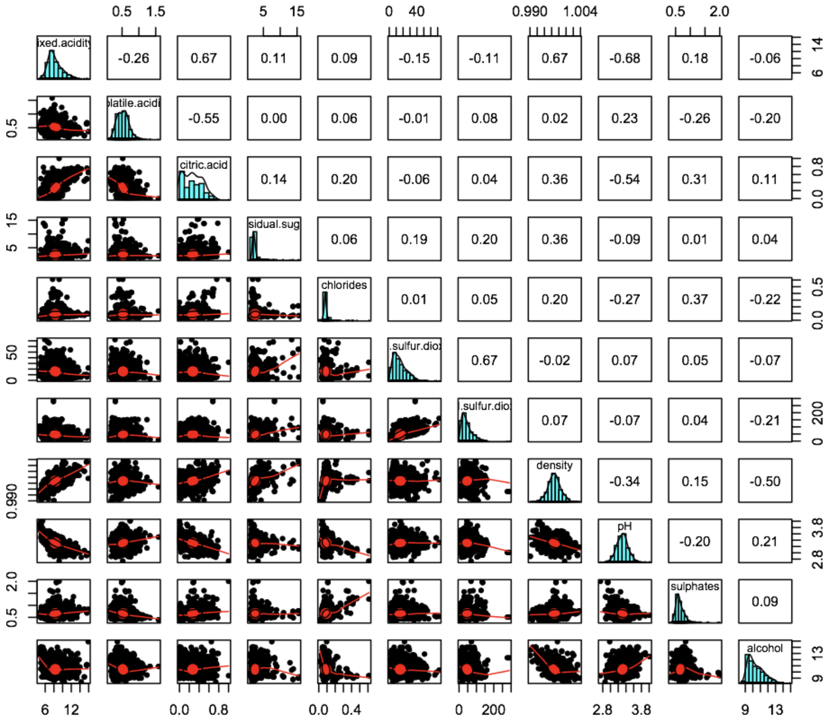


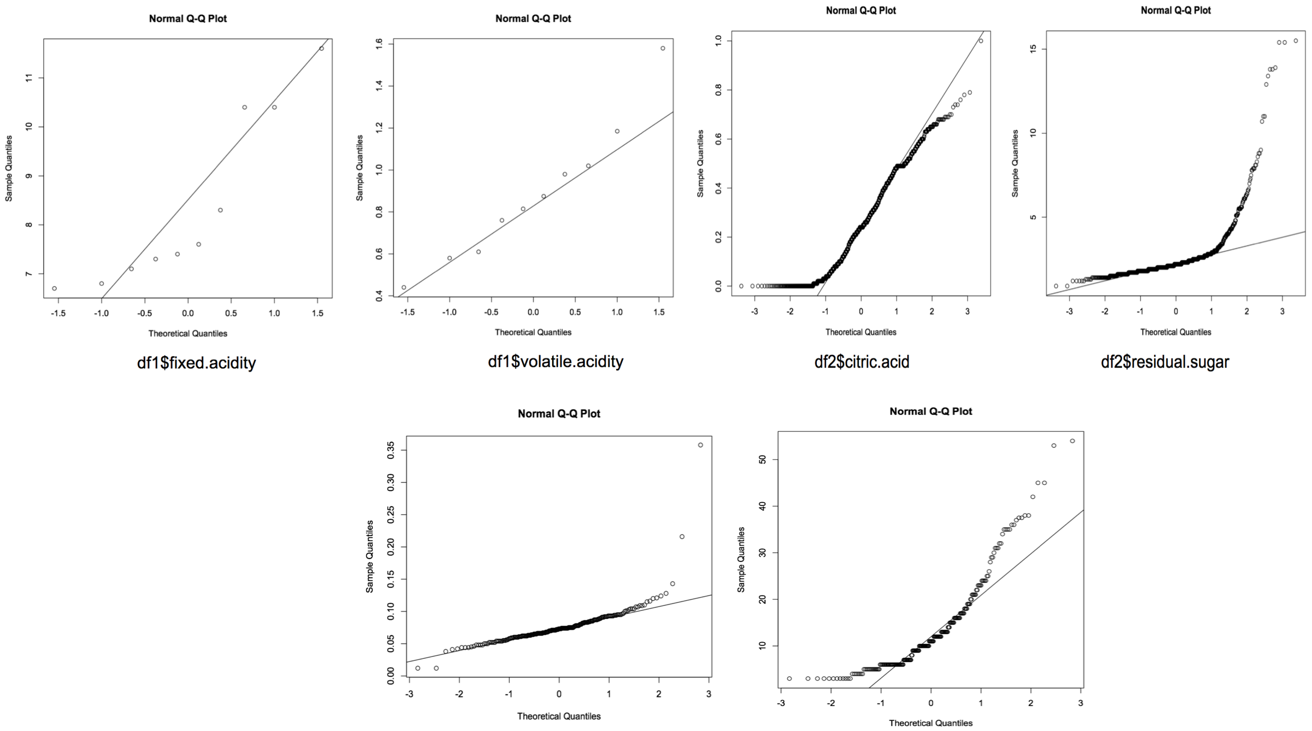
Figure 2: scatter plot of matrices, correlation and histogram

This is a very important graph for our analysis; firstly, it shows the correlation between the variables and shows their relationship, secondly, it plots the histograms of the variables on the diagonals, which will help us analyze which distributions the variables follow, for example, we can see that the density variable and the PH variable follows a normal distribution; Sulfur dioxide follows a gamma distribution. The plot also shows the correlation between the variables, as shown above, the correlation numbers are not that high, which indicates that not many variables have strong correlations, the strongest positive correlation value we obtained was 0.67, which is the correlation between fixed acidity and density, the weakest positive correlation was 0.01, which is the correlation between chlorides and sulfur dioxide. And finally the strongest negative correlation was between fixed acidity and Ph level with a value of -0.68. These values indicate that the relationship or dependency between the variables are not that strong and that they are to some extent independent.

3.3 Testing Normality of the Data

As shown in the previous analysis not all variables have the same distribution; therefore we will be using the Q-Q plot to test normality of the data of the three different groups.

Figure 3: Q-Q plots of different 6 variables from the three different groups are shown below



df1$fixed.acidity

df1$volatile.acidity

df2$citric.acid

df2$residual.sugar

df3$chlorides

df3$free.sulfur.dioxide

As shown in the qq plots above not all variables follow normal distribution as the data is not normally distributed, the points deviate from the qq line such as but not limited to citric.acid, residual.sugar, and free.sulfur.dioxide. Therefore, we need to scale the dataset so that points on the Normal QQ plot shows that the data is normally distributed, the points will fall on qq line, and that provides an indication of normality of the dataset.

As illustrated below in figure 4, after scaling the points and the graph itself do not change and remain the same, however, the scale is the only factor that changes, which allows all our variables to have the same mean which is 0 and standard deviation which is 0.99 approximately 1. Accordingly, our data is ready for implementing Fisher Linear Discriminant Analysis (FLDA) as the method assumes that the variance of the three groups are equal.

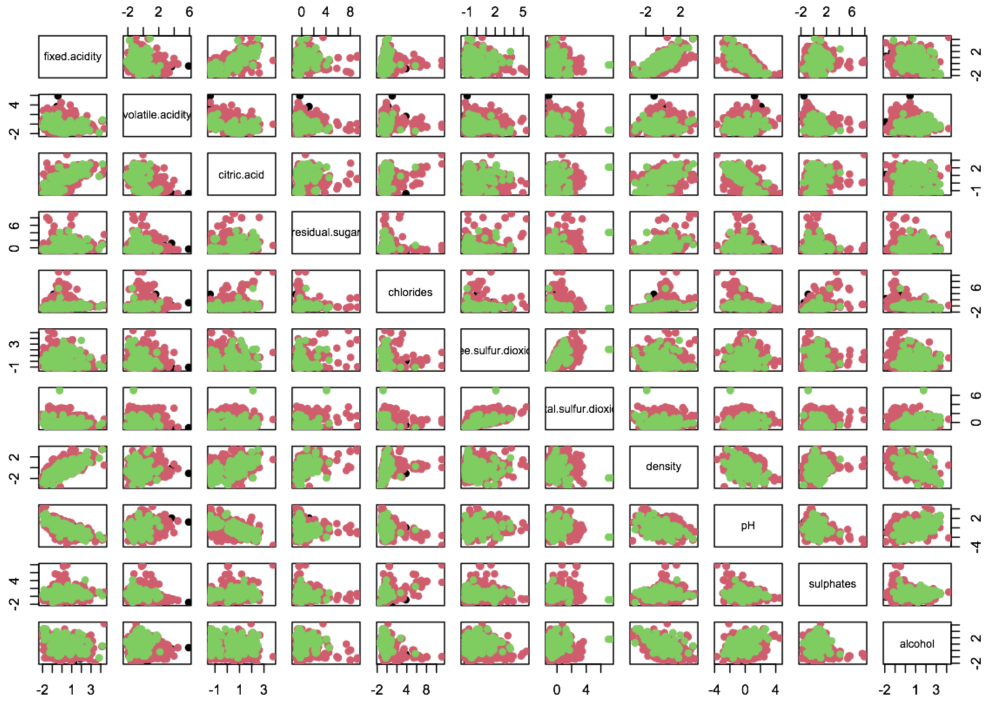


Figure 4: pairs of variables of the three groups after scaling the data

4. Discriminant Analysis

4.1 Using two functions in R (lda and predict)

There are two functions in R (lda and predict) that are useful for performing discriminant analysis. The input is the data matrix X that contains all observations (for all groups) in our case its z = scale(x[,1:11], center = TRUE, scale = TRUE) and a class variable class which score variable. The output of lda is as follows prior probabilities of groups: 1 (0.006253909), 2 (0.858036273), and 3 (0.135709819), Proportion of trace: LD1 (0.9035) and LD2 (0.0965) , the group means of all variables, and Coefficients of linear discriminants: LD1 and LD2. While the output of predict is b$posterior: an n x k matrix containing the posterior probabilities for each observation, b$class: a vector of length n containing the classes to which each observation has been assigned, and b$x: the scores of test data on up to dimension of discriminant variables.

4.2 Fisher Linear Discriminant Analysis (FLDA)

Fisher Linear Discriminant:

Correct Class Classification	Incorrect Class Classification			Total
	1	2	3	
1	2	8	0	10
2	6	1303	63	1372
3	0	130	87	217
Total	8	1441	150	1599

Error Rate = 12.94559 %

Figure 4: FLDA output

We used Fisher Linear Discriminant Analysis (FLDA) for the entire data set, and allowed the FLDA to classify the observations into 3 groups. As shown in the above figure, we can see that FLDA gives the confusion matrix, where we can see the correctly classified observations and the misclassified observations. The internal validation table gives us the correctly classified observations in the diagonals and the misclassified observations in the off-diagonals. In group 1 we got 2 correctly classified observations, and 6 observations from group 2 that were misclassified into group 1; in group 2, 1303 observations were classified correctly, 8 observations from group 1 were misclassified into group 2, and 130 observations from group 3 were misclassified into group 2; Lastly, in group 3, 87 observations were classified correctly, and 63 observations from group 2 were misclassified to group 3, overall, we got an error rate of 12.95% which tells us the overall misclassification error rate which is reasonable rate, it is calculated by((the total number of misclassified observations)/(Total number of observations))*100.

4.3 External Validation for FLDA (Leave one out)

External validation:

Correct Class Classification	Incorrect Class Classification			Total
	1	2	3	
1	2	8	0	10

2	6	1302	64	1372
3	0	135	82	217
Total	8	1445	146	1599

Missclassification Error = FC/(n+ m) = (8+64+6+135)/1599*100 = 13.3208 %

Figure 5: External Validation for FLDA output

As shown in the above figure, we can see that External Validation for FLDA (Leave one out) gives the confusion matrix, where we can see the correctly classified observations and the misclassified observations. The external validation table gives us the correctly classified observations in the diagonals and the misclassified observations in the off-diagonals. In group 1 we got 2 correctly classified observations, and 6 observations from group 2 that were misclassified into group 1; in group 2, 1302 observations were classified correctly, 8 observations from group 1 were misclassified into group 2, and 135 observations from group 3 were misclassified into group 2; Lastly, in group 3, 82 observations were classified correctly, and 64 observations from group 2 were misclassified to group 3, overall, we got an error rate of 13.32% which tells us the overall misclassification error rate which is reasonable rate, but higher than internal validation error rate, it is calculated by((the total number of misclassified observations)/(Total number of observations))*100.

4.4 FLDA 2

Fisher Linear Discriminant Analysis 2; p = 2 and k = 2

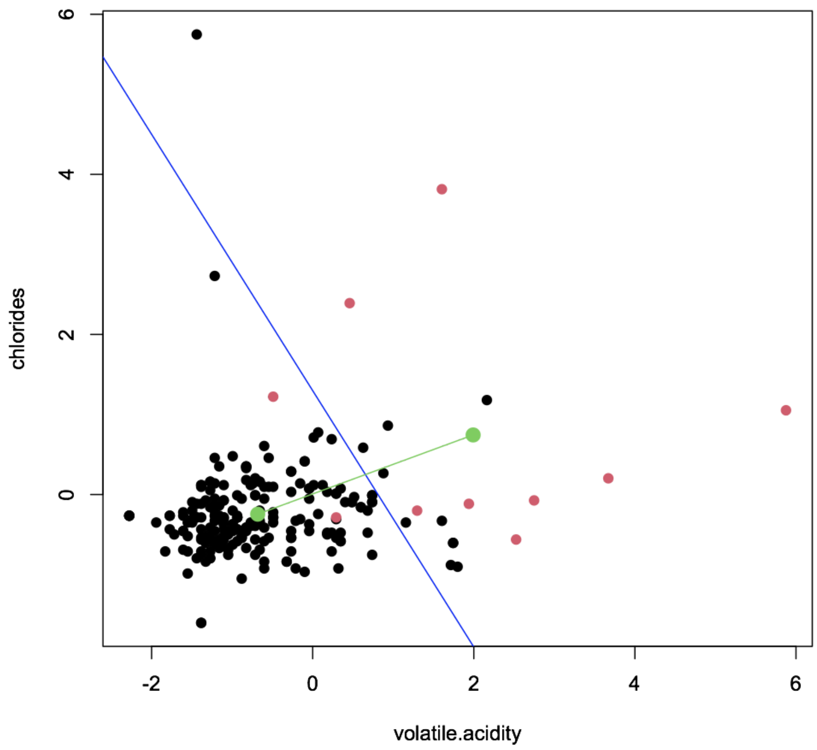


Figure 6: FLDA 2 plot for p = 2 and k = 2

In FLDA2, our input included 2 groups and 2 variables, we chose the 2 variables based on the pairs plot for the groups, we chose the variables that had the best separation of groups which were chlorides and volatile.acidity, then we used it for Fisher Linear Discriminant Analysis 2 The graph above shows the separation of the 2 groups by the blue line.

Correct Class Classification	Incorrect Class Classification		Total
	1	2	
1	206	11	217
2	8	2	10
Total	214	13	227

Error Rate = 5.726872 %

Figure 7: Internal validation for FLDA2

As shown in the above table, we can see the number of correctly classified and misclassified observations in the two groups; similar to FLDA1's confusion matrix, the diagonals of the matrix are the correctly classified observations, and the off-diagonals are the misclassified observations. Starting with group 1, 206 observations were correctly classified and 8 observations from group 2 were misclassified in group 1; moving to group 2, only 2 observations were classified correctly, and 11 observations from group 1 were misclassified into group 2. The error rate here is 5.73%, which is almost half of the error rate in FLDA1, which tells us that more groups were correctly classified using FLDA2 than FLDA1.

4.5 Multinomial

Intenral Validation	**External validation**
results 1 2 3 1 1 9 0 2 2 1327 43 3 0 142 75	**rslt** 1 2 3 1 1 9 0 2 2 1322 48 3 0 148 69
Error Rate = 12.25766%	**Error Rate** = 12.94559 %

Figure 8: Internal and External validation for Multinomial

As illustrated in the table above, we used the multinomial which is a Discriminant Analysis Method and special case of the Generalized Linear Models. Comparing multiomial with FLDA, FLDA reported 12.94559 % error rate in internal validation and 13.3208 % in external validation; therefore, in our data multinomial performs better than FLDA as it reported 12.25766% error rate in internal validation and 12.94559 %in external validation.

4.6 Projection

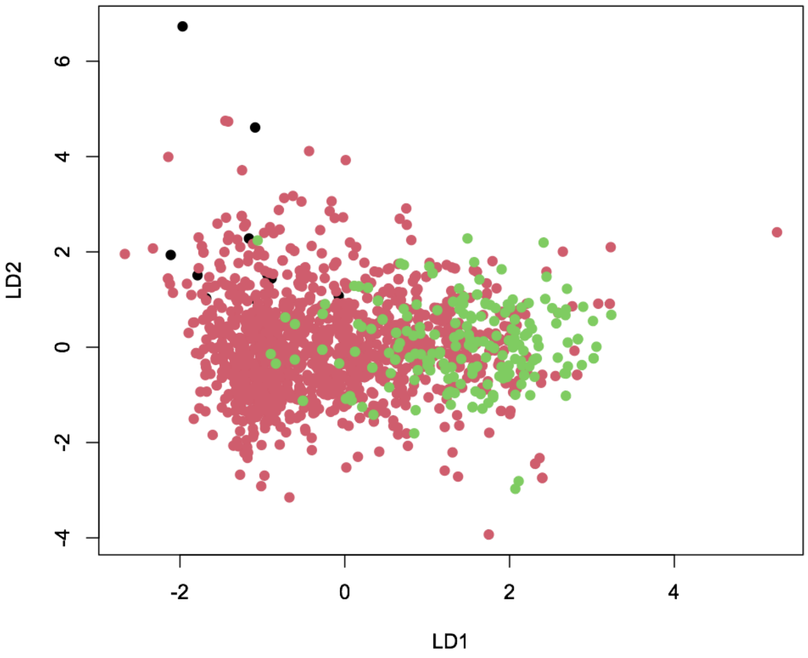


Figure 9: projection using b$x

As shown in figure 9, the projection that we are looking for is the one that maximizes the seperation between groups. However, in our case since the groups' size are not balanced and there might be other effecting factors, so it seems that the separation of groups is not the best.

5. Clustering

5.1 Background information

Another discriminant analysis method is the clustering method; this method works on finding clusters (groups) within the data set. Observations with similar characteristics are combined together into a cluster, we always try to minimize the within cluster dispersion and maximize in between cluster dispersion to reach efficiency. In our data, we will apply the clustering algorithm to determine if there are groups within our data that contain observations with similar characteristics, we will use 2 methods, the hierarchical method which focuses on combining observations **closest** to each other, and the k-means method which does the same thing but we already know the number of clusters before starting, so it makes it more computationally efficient than hierarchical for big data sets.

5.2 hierarchical

The hierarchical method of clustering ensures that we minimize within class dispersion and maximize between class dispersion. The method puts each observation in a cluster of its own, then computes the distances between each cluster, clusters with the smallest distances between them are combined into one cluster, and the process is repeated until there are no more combinations of clusters. We did this method four times using different methods to measure the distances between observations and distances between clusters.

First Trial: Distances between observation: Euclidean ; Distances between clusters: Ward

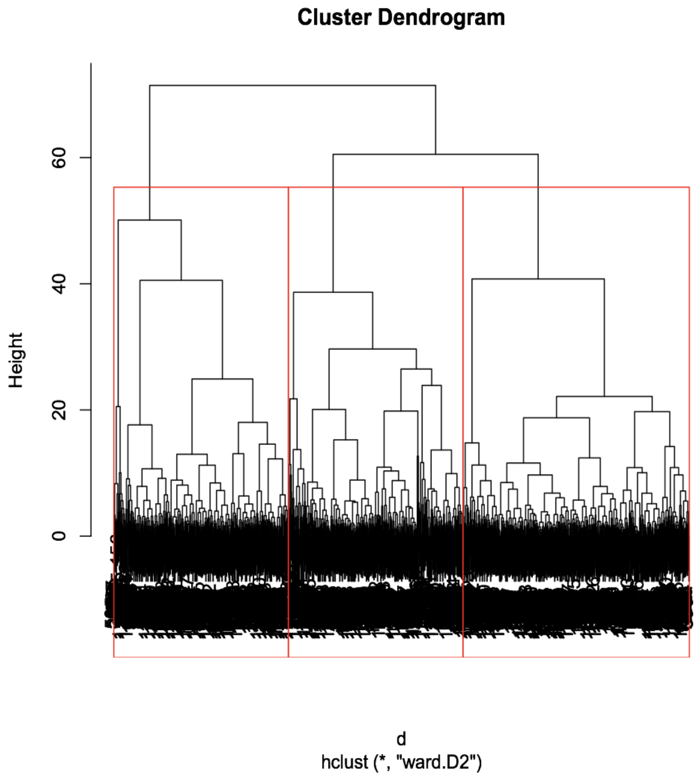


Figure 10: Cluster Dendrogram 1

As seen in the figure above, we can see all the different observations that have been grouped by clusters, if we cut the dendrogram at height = 50, we will find that we have 3 clusters, which is the number of groups we obtained in our previous discriminant analysis, which tells us that that this method of clustering is efficient for discriminant analysis. It also tells us that the Ward method for measuring distances between clusters is a good method, as the clusters are well separated and the between class variation is good and visible.

Second Trial:Distances between observation: Euclidean ; Distances between clusters: Complete

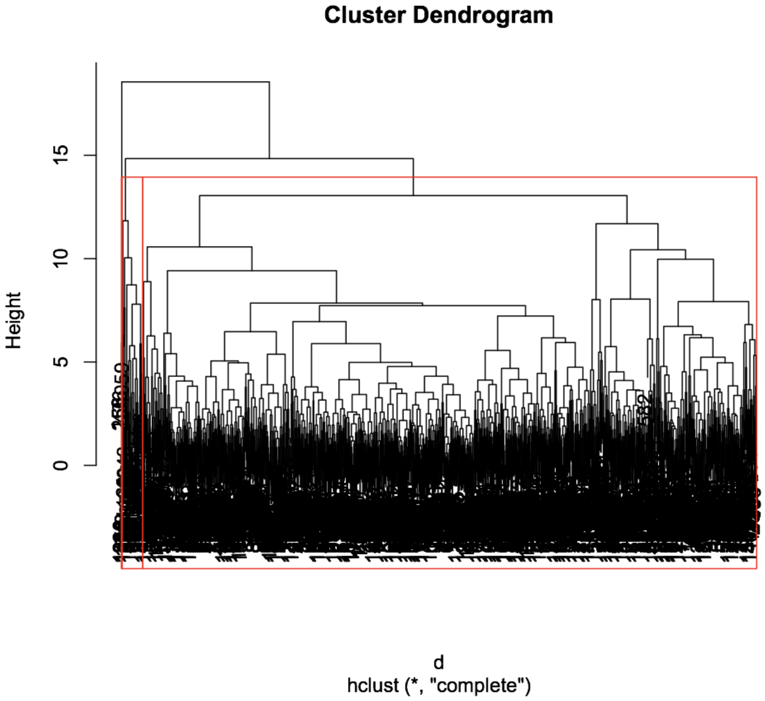


Figure 11: Cluster Dendrogram 2

In Figure 7, we used Euclidean distance to measure the distance between observations and complete method for measuring distance between clusters; if we cut at height = 13, we will get 3 clusters, however, the distances between the clusters is not maximized, as the first 2 groups are very close to each other, and were not combined into one cluster, which tells us that the complete method for measuring distance between clusters is not efficient, or that the combination of Euclidean distance and Complete method are not a good fit for this data, and does not give us a efficient clustering algorithm.

Third Trial:Distances between observation: Manhattan ; Distances between clusters: Ward

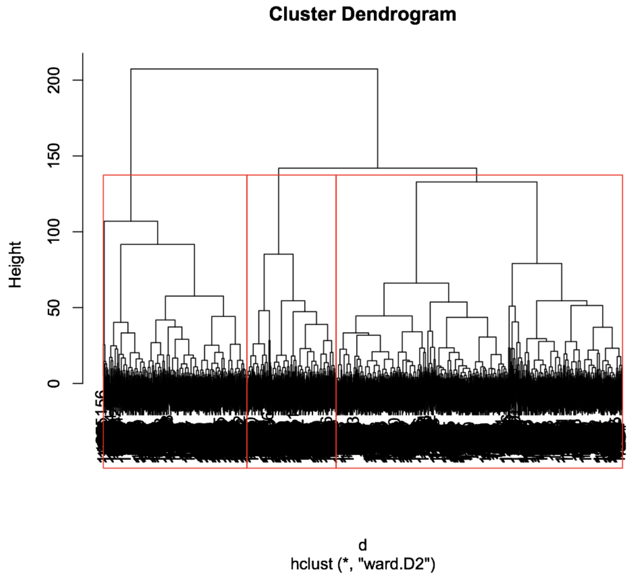


Figure 12: Cluster Dendrogram 3

This figure represents efficient clustering, as the clusters are well separated, if we cut the dendrogram at height = 125, we will obtain 3 clusters, which is the same as the first trial. If we cut at height = 110, we get four clusters that are also very good seperated. So this is trial resulted in a successful clustering algorithm.

Fourth Trial:Distances between observation: Manhattan ; Distances between clusters: Complete

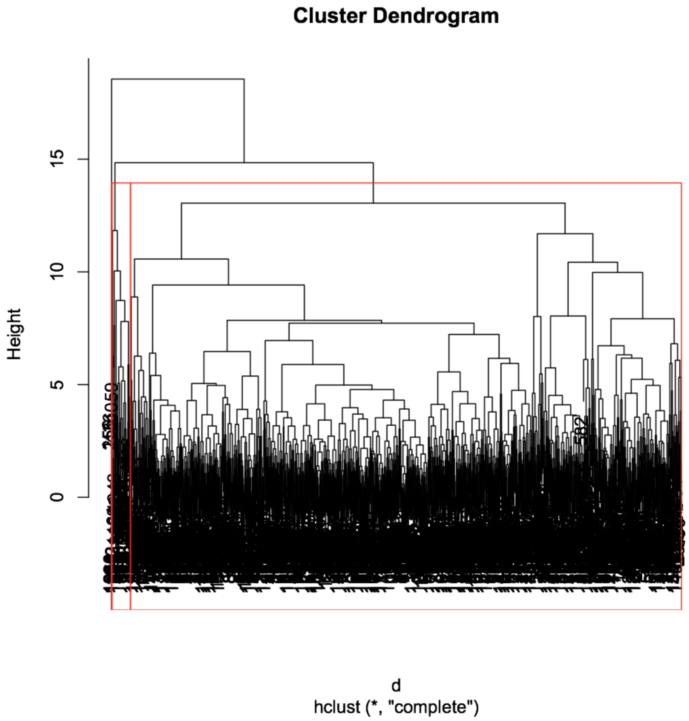


Figure 13: Cluster Dendrogram 4

The figure above represents the last trial; similar to trial 2, the groups are not very well separated, if we cut the dendrogram at height = 13, we get three clusters however, the first 2 groups are very close to each other, which highlights that the between class variation is not effective and another method of clustering should be used.

5.3 Goodness of Fit Test

The goodness of fit test measures the efficiency or the accuracy of our clustering algorithms, it uses the within class variation and the between class variation to give us a value that allows us to determine if our clustering algorithm is good or not. Our goal is to minimize the within class variation and maximize the in between class variation. R^2=tr(BSS)/tr(TSS) where TSS+WSS+BSS , so we want to maximize this value, R^2 lies from 0 to 1; from our data, we got a value of 1, which indicates that our clustering algorithm was efficient as it maximized the distances between clusters and minimized the distance between observations.

5.4 K-means

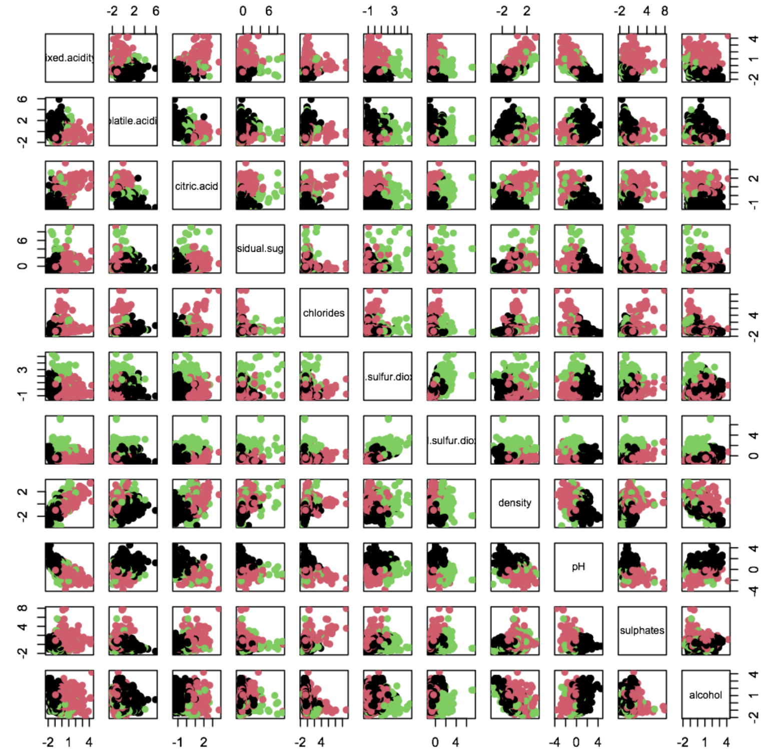


Figure 14: pairs after K-means clustering

The K-means method of clustering is similar to the hierarchical clustering method but, we know the number of clusters in advance. The method plots k random points, then measures the distance between the observations and the centroids of the k random points, the points closer to the centroids are combined into a cluster. The process is then repeated until all observations are in a cluster and the new cluster centroids are computed. As shown in the figure, the three groups are separated in each variable, some clusters overlap, like citric acid and alcohol together, sulfates and alcohol. And some clusters are separated such as citric acid and sulfur dioxide, fixed acidity and sulfur dioxide. Overall, the k-means clustering method was not that efficient in our data, as the between cluster variation is minimal as seen in the pairs graph. It worked on some variables, and failed on others. So the hierarchical method of clustering is more efficient for our data.

5.5 Determining the Number of Clusters

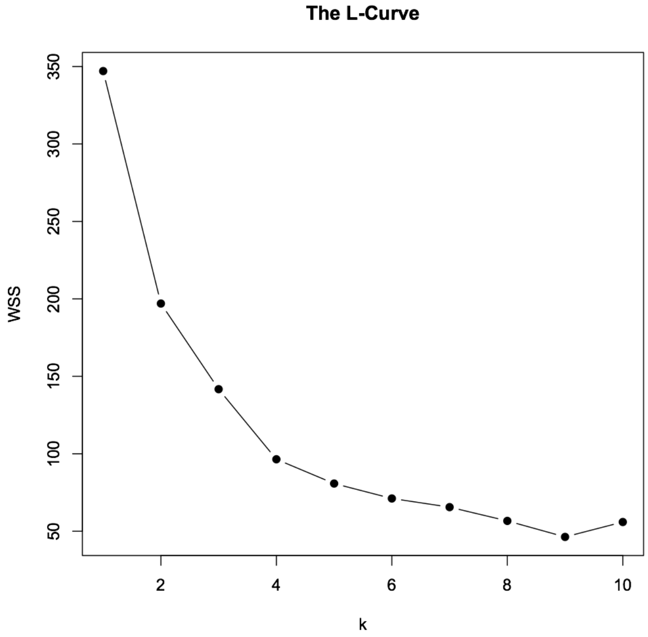


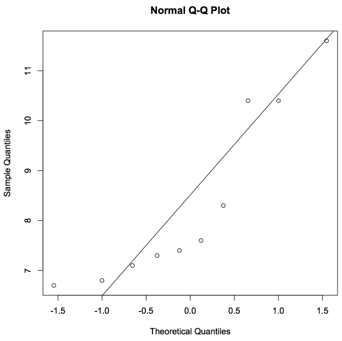
Figure 15: The L- curve

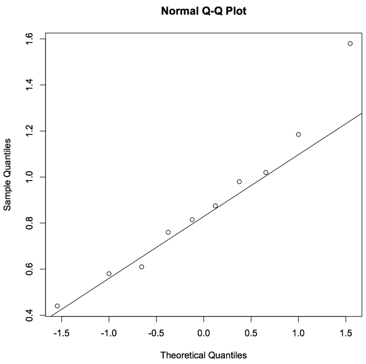
The L-curve suggests that the number of clusters in our data is 4, as at k = 4 the line’s slope starts to flatten which indicates that the within class variation has come to a minimum, however, if we look closely, we will find that the optimal/most efficient number of clusters that minimize the within class variation is k = 9, which is a lot of clusters, but the difference between WSS in k = 4 and k = 9 is not significant, so we can assume that the number of clusters in our data is four.

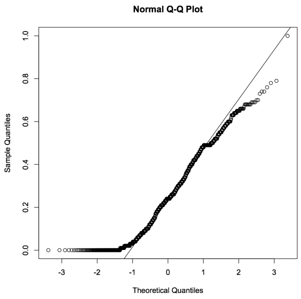
6. Conclusion

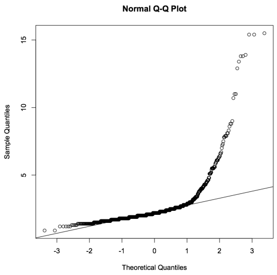
In conclusion, our analysis has shown that for discriminant analysis both FLDA and multinomial showed good results, however, the multinomial was better as it had lower error rate for both internal and external validation. There were some limitations found in our analysis. First when we splitted the data we choose the variable quality to split the based on wine quality rating and according to the data description Score 1 represents very bad quality wine in which the quality rating is lower than 4, Score 2 represents bad quality wine in which the quality rating is between 4 and 6.5, Score 3 represents good wine quality in which the quality rating exceeds 6. Accordingly, by splitting the data the three groups' size was not balanced in which in group 1, there were 10 observations, group 2 contained 1372 observations, and group 3 consisted of 217 observations. Therefore, this have impacted the results of our analysis especially in discriminant analysis. In addition to that, we found out that external validation was better than internal validation, as it excludes one observation out, and uses the rest of the observations to derive a rule that we test on the observation we left out, so it gave more accurate and unbiased results. Finally, in regards to clustering, we found out that the hierarchical method of clustering was better than the k-means clustering method as it gave us a more accurate number of clusters and is more efficient than the k-means method as it maximized the between class variation and minimized the in between class variation. When we conducted the goodness of fit test, we got an R^2 value of 1, which highlighted the efficiency of our clustering method. Overall, we believe that clustering was the best discriminant analysis method for our dataset, as it gave us the correct number of groups (three).

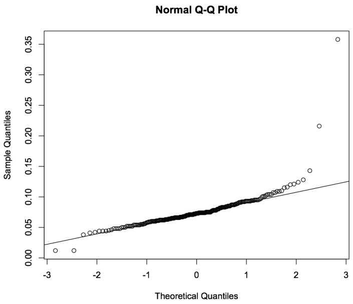
7. Appendix

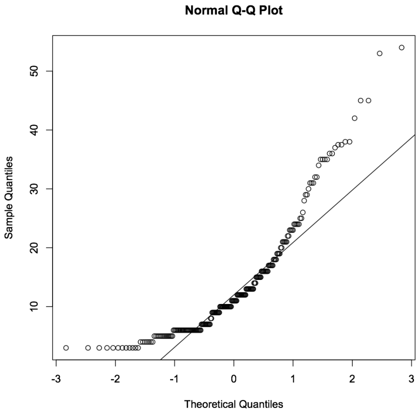

Normal Q-Q Plot
Sample Quantiles
7
8
9
10
11
Theoretical Quantiles
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

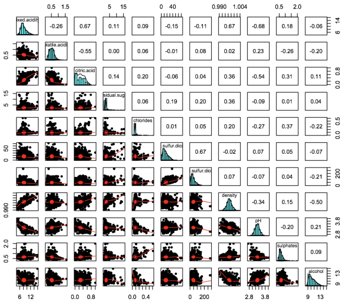

Normal Q-Q Plot
Sample Quantiles
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Theoretical Quantiles
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

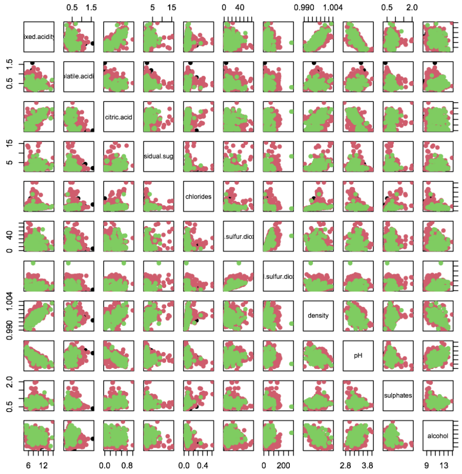

Normal Q-Q Plot
Sample Quantiles
0.0
0.2
0.4
0.6
0.8
1.0
Theoretical Quantiles
-3
-2
-1
0
1
2
3

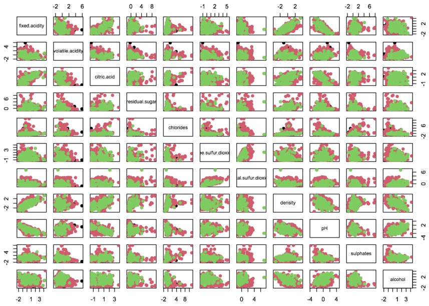

Normal Q-Q Plot
Sample Quantiles
5
10
15
Theoretical Quantiles
-3
-2
-1
0
1
2
3

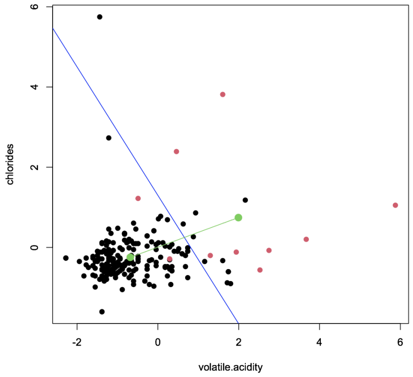

Normal Q-Q Plot
Sample Quantiles
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Theoretical Quantiles
-3
-2
-1
0
1
2
3

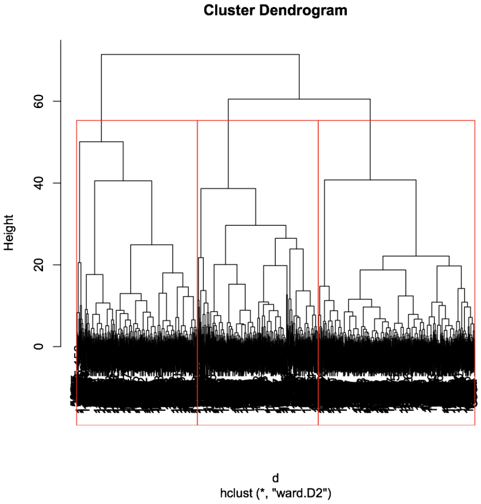

Normal Q-Q Plot
Sample Quantiles
10
20
30
40
50
Theoretical Quantiles
-3
-2
-1
0
1
2
3

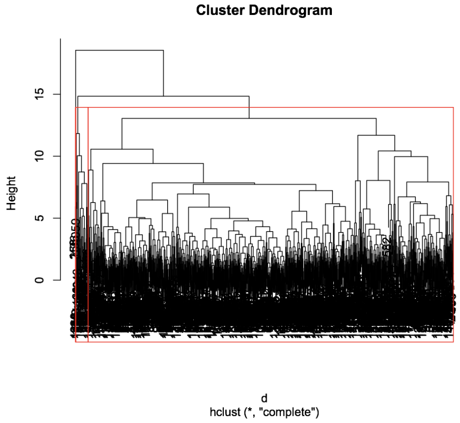

ixed.acidity
-0.26 0.67 0.11 0.09 -0.15 -0.11 0.67 -0.68 0.18 -0.06
olatile.acidi
-0.55 0.00 0.06 -0.01 0.08 0.02 0.23 -0.26 -0.20
citric.acid
0.14 0.20 -0.06 0.04 0.36 -0.54 0.31 0.11
sidual.sug
0.06 0.19 0.20 0.36 -0.09 0.01 0.04
chlorides
0.01 0.05 0.20 -0.27 0.37 -0.22
.sulfur.diox
0.67 -0.02 0.07 0.05 -0.07
.sulfur.diox
0.07 -0.07 0.04 -0.21
density
-0.34 0.15 -0.50
pH
-0.20 0.21
sulphates
0.09
alcohol
0.5 1.5
5 15
0 40
0.990 1.004
0.5 2.0
6 12
0.0 0.8
0.0 0.4
0 200
2.8 3.8
9 13

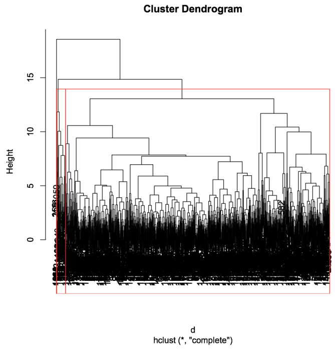

ixed.acidity
olatile.acidi
citric.acid
sidual.sug
chlorides
.sulfur.diox
.sulfur.diox
density
pH
sulphates
alcohol
0.5 1.5
5 15
0 40
0.990 1.004
0.5 2.0
6 12
0.0 0.8
0.0 0.4
0 200
2.8 3.8
9 13

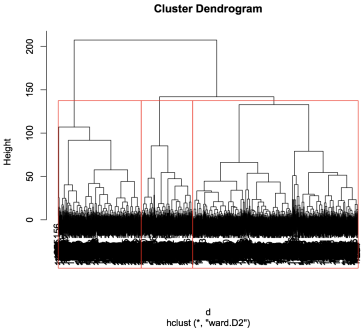

fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
e.sulfur.dioxi
al.sulfur.dioxi
density
pH
sulphates
alcohol

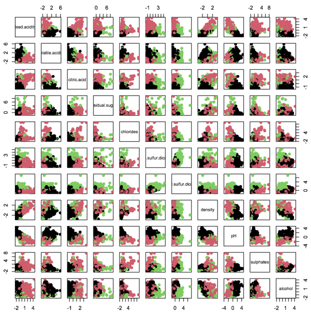

chlorides
0
2
4
6
volatile.acidity
-2
0
2
4
6

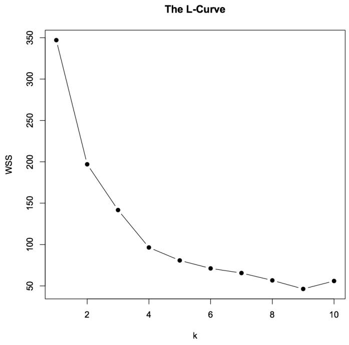

Cluster Dendrogram
Height
0
20
40
60
d
hclust (*, "ward.D2")


Cluster Dendrogram
Height
0
5
10
15
d
hclust (*, "complete")


Cluster Dendrogram
Height
0
5
10
15
d
hclust (*, "complete")


Cluster Dendrogram
Height
0
50
100
150
200
d
hclust (*, "ward.D2")


xed.acidity
latile.acidi
citric.acid
sidual.sug
chlorides
.sulfur.dio
l.sulfur.dio
density
pH
sulphates
alcohol
-2 2 6
0 6
-1 3
-2 2
-2 4 8
-2 1 4
-1 2
-2 4
0 4
-4 0 4
-2 1 4


The L-Curve
WSS
50
100
150
200
250
300
350
k
2
4
6
8
10