Project

MACT423301- MACT423302- Applied Multivariate Analysis (Spring 2023)

Group members:

Maya Hany Elshweiky 900204233

Omar Ahmed 900201077

Dr. Ali Hadi

22 March 2023

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?
- Can we identify any outliers or anomalies in the dataset that may be influencing the results?
- Can we predict the quality rating of a wine based on its physicochemical properties?
- Can we use cluster analysis to identify any patterns in the dataset that could be useful for wine production or marketing?
- How do the properties of the highest-rated wines differ from those of the lowest-rated wines?

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 split the data into two groups based on the "quality of the wine" variable which is a categorical quantity, any quality value above 6.5 will be in group 1, and any value below 6.5 will be in group 2.

We will use the BACON approach of identifying outliers which will ease the process of analyzing the data and answering our questions. The BACON approach is an "A compromise between robustness and computational efficiency" method discovered by the following authors: Hadi (1992, 1994), JRSS Billor, Hadi, and Velleman (2000), CS&DA, it is a method for analyzing and distinguishing outliers in any given data set. Firstly, we compute the Mahalanobis distance of each observation (it is a fast but not robust method), and we measure the distance euclidean distance between each observation and the mean. Secondly, we compare it to the adjusted chi-square critical value (Billor, Hadi, and Velleman (2000)), if the Mahalanobis distance is smaller than the adjusted chi-square critical value, we put this observation in a new basic subset. The two steps mentioned above are repeated until the last two iterations showing the number of observations in the basic subset are equal; the same is repeated for the non-basic subset until the last two iterations show the same number of variables in the subset. This highlights that the basic subset is free from outliers, and the non-basic subset contains all the outliers in the data. An advantage of this method is that it is computationally efficient and can be used for a large data set with many observations like the "Wine Quality" data set we are using.

In addition to that, we will also be using the Hotelling T2 test in which the test hypotheses are:

- Null hypothesis (H0): the two samples are from populations with the same multivariate mean.

- Alternative hypothesis (H1): the two samples are from populations with different multivariate means.

We compute the value of the T-squared and compare it to a p-value; if the p-value is less than the alpha(significant level), which is 0.05, we reject the null hypothesis.

2. Data Description

Who: 1599 observations of wine quality based on physicochemical tests

When: The data was collected in 2009

Where: The data set 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 (0 or 1) and eleven chemical attributes (quantitative), which are as follows: Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, PH, Sulphates, and Alcohol 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	Chloride, 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 score between 0 (very bad) and 10 (very excellent)
score[1]	Categorical (binary)	-	- Score 1 represents good quality wine in which the quality rating is greater than 6.5 -Score 0 represents bad quality wine in which the quality rating is less than 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. This data is accurate, relevant, and consistent 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 which allow us to analyze different aspects of the data and answer many questions.

[1] Note: score variable was added using excel to split quality rating into two categories good and bad

3. Data Visualization and Analysis

3.1 The Relationship Between Variables

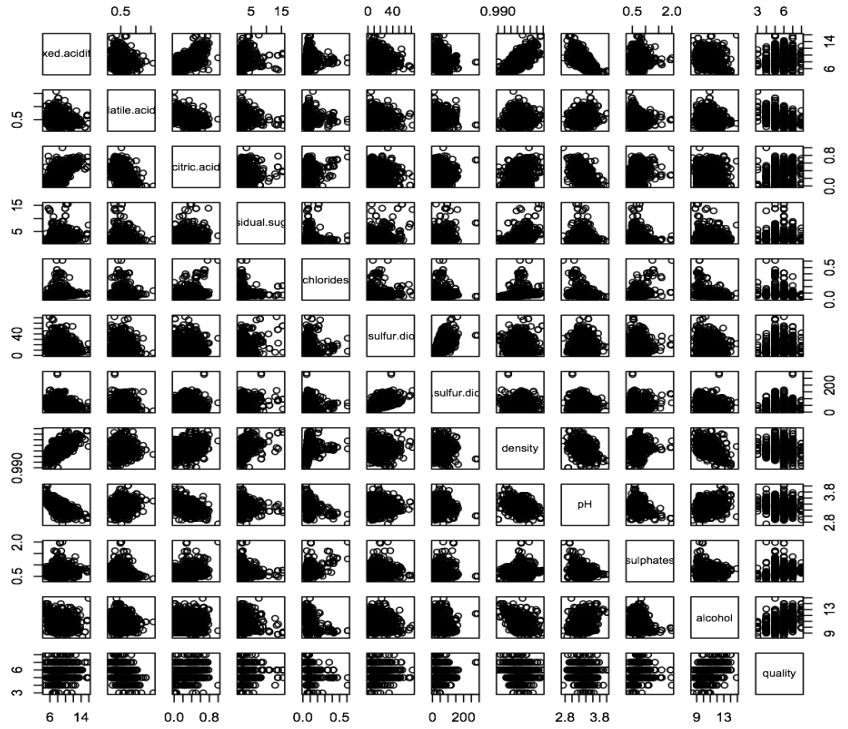


Figure 1: pairs of variables of the whole dataset

As shown in figure 1, we can see that some of the variables have linear relationships such as fixed acidity and density have a positive linear relationship, fixed acidity, and pH have a negative linear relationship, and fixed acidity and citric acid have a positive linear relationship. While for the other graphs, there is not a linear relationship but we cannot conclude that the variables don't have a relationship in some other sort. Also, in other graphs, we can see outliers that affect the relationship. For the quality variable, it is categorical so the graph expectedly does not show any relation.

3.2 Splitting Data into two groups

We split the data into two groups based on the score which is the quality rating of wine where group 1 represents good quality wine in which the quality rating is greater than 6.5 and group 2 represents bad quality wine in which the quality rating is less than 6.5.

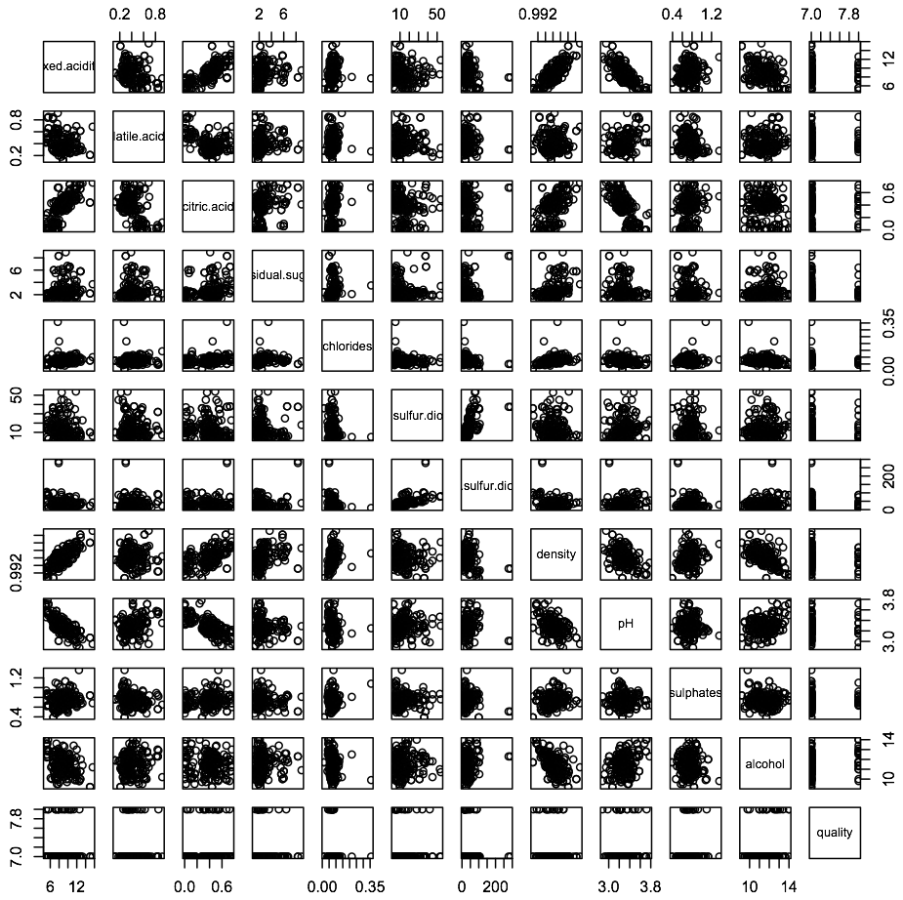


Figure 2: Pairs Graphs of Group 1

As shown in figure 2, we can see that some variables have linear relationships, however there are barely any strong linear relationships seen in this data set, we can highlight that citric acid and fixed acidity have a positive linear relationship, and density and fixed acidity have a strong positive correlation, in addition to that, Ph level and fixed acidity also have a linear relationship

but a negative correlation. In sulfur dioxide for example, we can see that most of the correlations with that variable have an outlier right at the top of the graph, which affects the overall correlation of the graph. That is why it is important to detect and remove outliers, so the overall correlation of the graph is not affected.

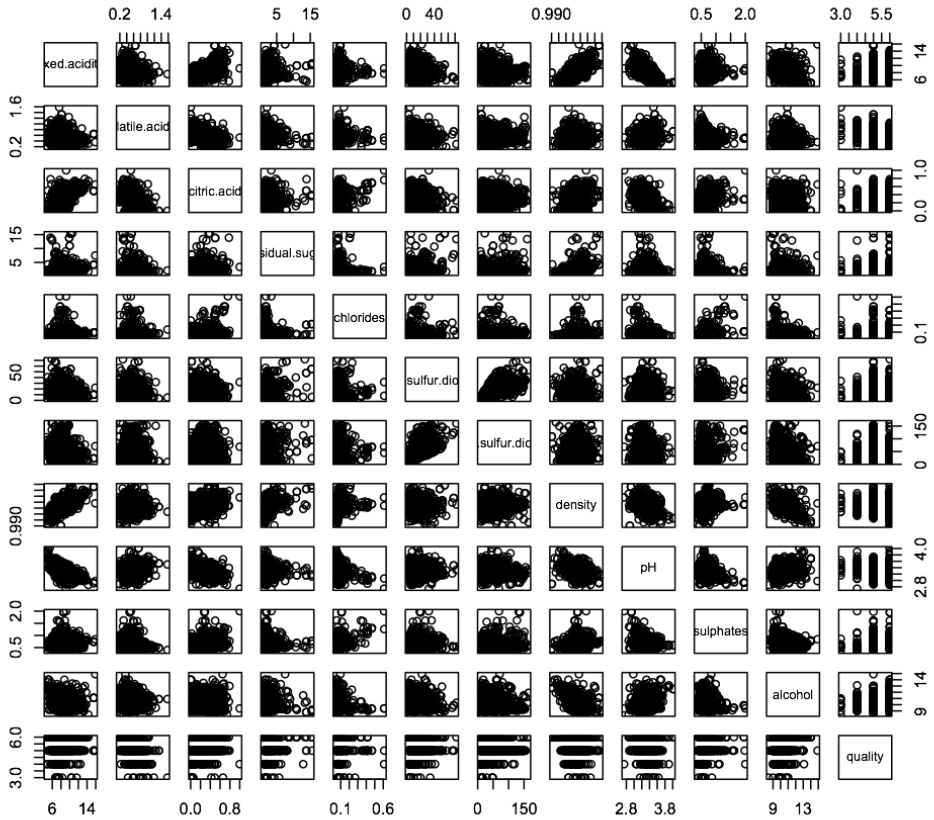


Figure 3: Pairs Graphs of Group 2

As shown in figure 3, we can see that some of the variables have linear relationships such as fixed acidity and density have a positive linear relationship, fixed acidity and pH have a negative linear relationship, fixed acidity and citric acid have a positive linear relationship. While for the other graphs, there is not a linear relationship, we cannot conclude that the variables don't have a relationship in some other sort. Also, in other graphs, we can see outliers that affect the

relationship. For the quality variable, it is categorical so the graph expectedly does not show any relation.

3.3 The BACON Method

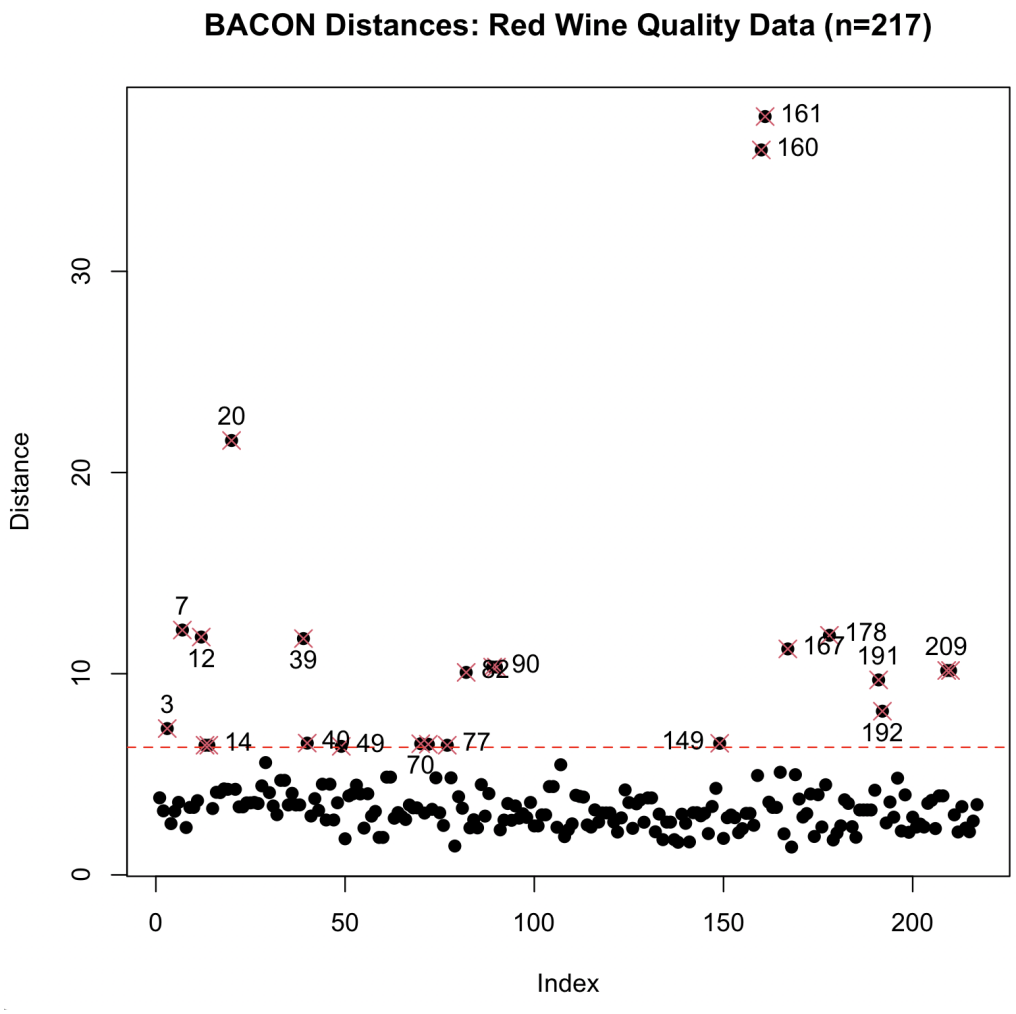


Figure 4: Graph of BACON for Group 1

As seen above, any observations above the red-dotted line are considered as outliers according to the BACON approach, the maximas seen are observations: 20, 160 and 161. Starting with Observation 20, when we looked at the data itself, we found that the observation has very high

levels of chlorides compared to other observations, we also found out that the mean and median of the chlorides variable is 0.08747 and 0.07900, respectively, while observation 20 had a value of 0.341 g / dm^3 of chlorides, which is very far away from the mean and median, so we consider this an outlier. Secondly, we investigated observation 160, which is very far away from the red-dotted line; we found that the fixed acidity level was very low compared to the other observations and the mean and median; the mean and median of the fixed acidity variable are 8.32 and 7.90, respectively, while observation 160 had a fixed acidity value of 6.8, and the minimum value in this variable was 4.9, which shows us that the observation had a fixed acidity level close to the minimum value and away from the mean and median, which caused it to be an outlier. Lastly, spectating observation 161, we found that it had a very deviated value of volatile acidity than the mean and median of the variable, which were 0.5278 and 0.5200 respectively, while observation 161 had a value of 0.950 which is very far away from the mean and median, also, we found that observation 161 had 7.0 g / dm^3 of free sulfur dioxide, while the mean and median of this variable are 15.87 and 14 respectively, which indicates that this observation's value of free sulfur dioxide is an outlier compared to the other observation; however, we do need more information and analysis to detect exactly what is the reason for this observation to be an outlier. Overall, the outlier we detected were very deviated from the adjusted chi-square value, and we tried by inspection to determine the causes that lead to these observations to become outliers, however, we do need more critical analysis to determine why exactly they are outliers, the rest of the points above the red-dotted line are assumed to be outliers, but we are not sure as they are not highly deviated from the adjusted chi-square value, so we also need a more precise tool of analyzing outliers to determine the causes of them being outliers.

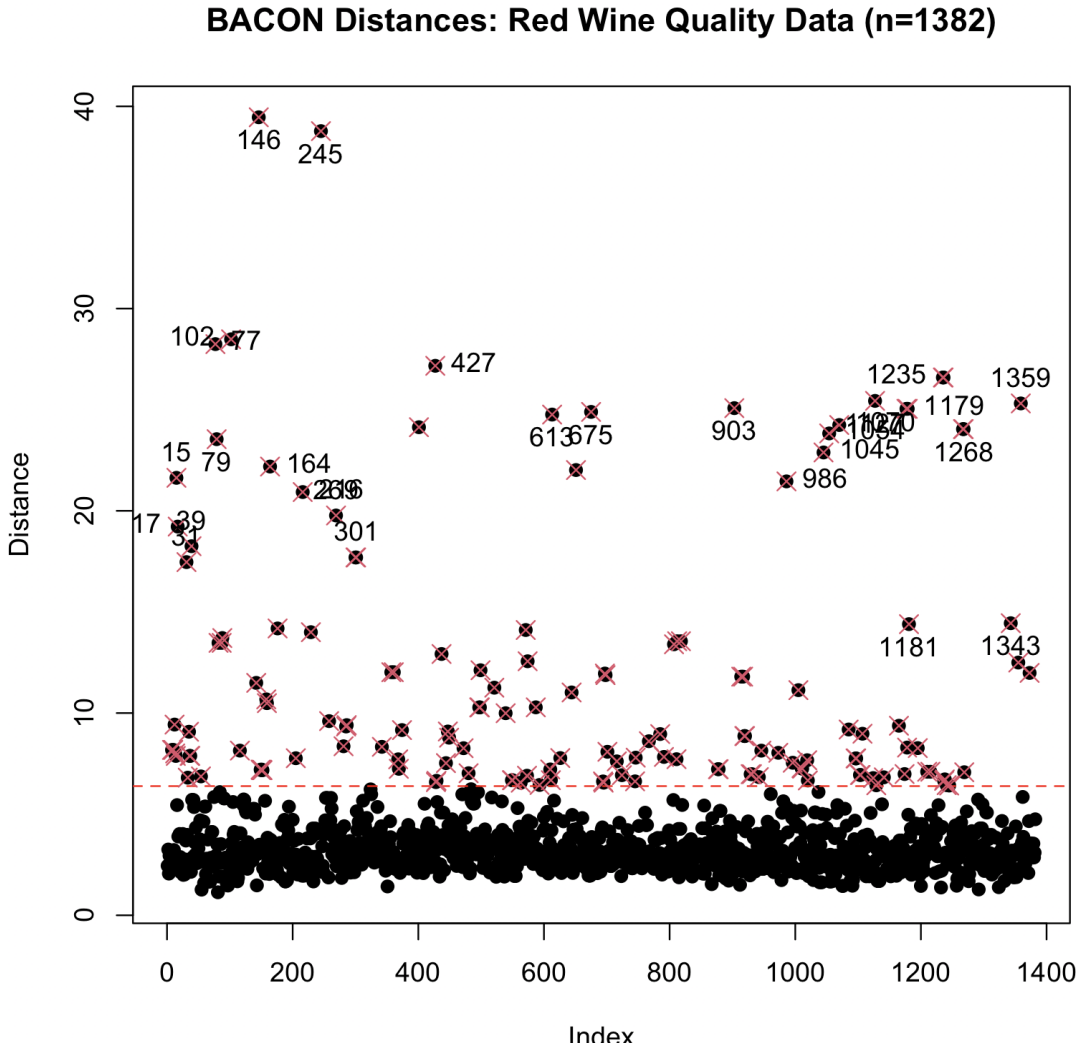


Figure 5: Graph of BACON for Group 2

As shown in figure 5, the BACON distances of group 2 which consists of 1382 observations. The outliers detected by BACON are the points above the red dotted line in which anything below the red dotted line is in the basic subset, and anything above this line is an outlier that is in the non-basic subset. Most of these outliers are close to the original points. To illustrate, the extreme outliers are observations 146, 245, and 77, respectively. To identify the reason for these outliers, we will compute the measure of location using R (summary(df2)) and detect which

variable is affecting the outlier points. Starting with the first extreme outlier, point 146, had a value of 14% at alcohol, and the statistics summary of alcohol is as follows, Minimum: 8.40, 1st Quantile: 9.5, Median:10.00, Mean: 10.25, 3rd Qu.: 10.90, Maximum: 14.90. Therefore observation 146 is affected by outliers for the reason of alcohol mean is 10.25 and the maximum value is 14.9. The second extreme outlier, observation 245, had a value of 15 g / dm^3 of fixed acidity the statistics summary of fixed acidity is as follows, Minimum: 4.600, 1st Quantile: 7.100, Median:7.800, Mean:8.237, 3rd Quantile: 9.100, Maximum:15.900. So as we can see that the mean is 8.237, 1st Quantile is 7.100, and 3rd Quantile is 9.100; so accordingly, we can conclude that observation 245 is identified as an outlier for the reason of in fixed acidity is away from mean and range of values. Lastly, The third extreme outlier, observation 77, comparing it with the values of this observation of each variable with the statistics summary, it can be concluded that all values are within the range; similarly, for the rest of the observations above the red dotted line are close to true values; therefore, these outliers detected by BACON could be outliers for other reasons.

3.4 Hotelling's T2-test

Hotelling T2 the test hypotheses are:

Null hypothesis (H0): the two samples are from populations with the same multivariate mean.

Alternative hypothesis (H1): the two samples are from populations with different multivariate means.

We will compute the value of the T-squared and compare the p-value to alpha; if the p-value is less than the alpha(significance level), which is 0.05, we reject the null hypothesis.

We take two samples from:

- population df1 (group 1) of size 217
- population df2 (group 2) of size 1382

3.4.1 Classical (non-robust) Hotelling T2 test

Hotelling's two sample T2-test

data: df1[, 1:11] and df2[, 1:11]

T.2 = 45.959, df1 = 11, df2 = 1587, p-value < 2.2e-16

As explained above, we can see that the p-value is smaller than alpha, which means that we will reject the null hypothesis H0, and accept the alternative hypothesis; this tells us that both populations have different means hence both populations are not the same.

3.4.2 A robust version Hotelling T2 test

Two-sample Hotelling test

data: df1[, 1:11] and df2[, 1:11]

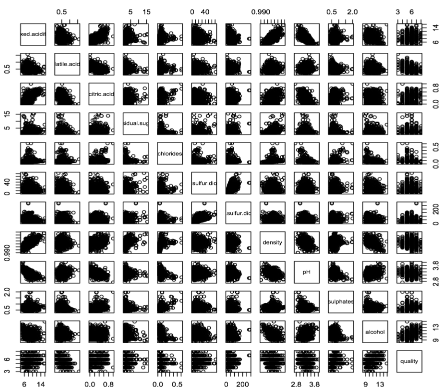
T2 = 508.737, F = 45.959, df1 = 11, df2 = 1587, p-value < 2.2e-16

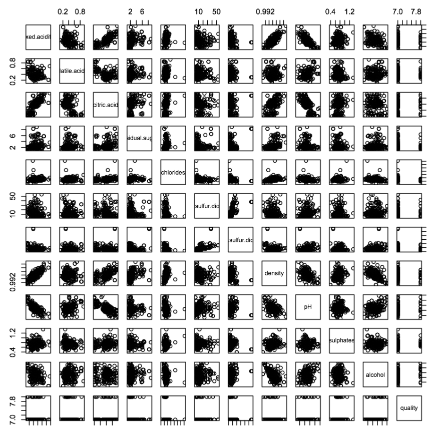
After computing robust version Hotelling T2 of groups 1 and 2, we can see that the value T2 is 508.737 and F value is 45.959 which in this case are not affected by the outliers. The p-value is approximately 0, therefore, we will reject the null hypothesis H0 as the p-value is less than alpha which is 0.05. Accordingly, this means that the two parameters (means) are significantly different.

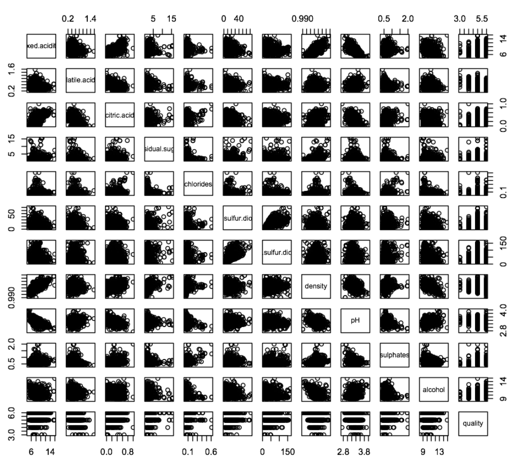
4. Conclusion

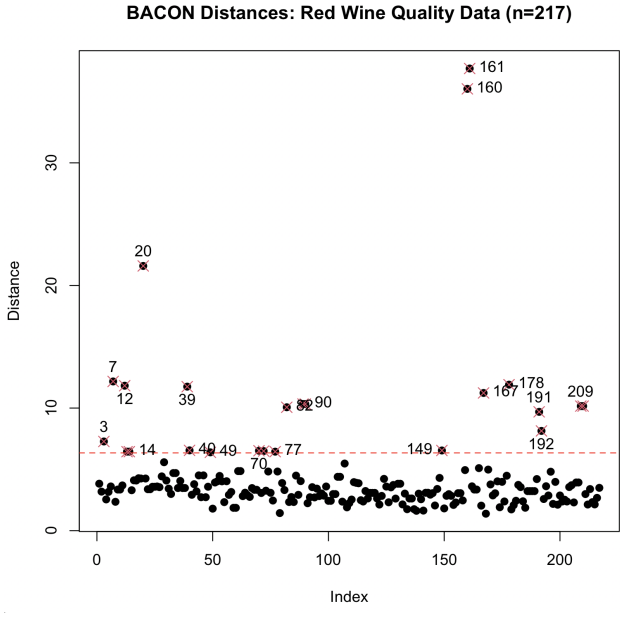
After critical analysis of the data set, using methods such as the BACON approach, Hotelling T2, and finding correlations between the variables, we derived that there were quite some outliers in the data, and we found out the reasons why some of the maximas where outliers, however, we believe that need more analysis for the data in order to precisely detect and determine the reasonings behind these outliers; some of these reasons could be that the sample size is small and the data needs transformation. In addition, we found some correlations between the variables and the quality of the wine and which variables had the most effect on increasing the quality of the wine; we discovered that the closer the variables in each observation are to the means of each observation, the higher the quality of the wine, the further away the observations are from the means of the variables, the lower the quality of the wine. Lastly, after we split the data into two populations depending on the quality of the wine, we discovered that both groups did not have equal means, even though they were from the same data set, which shows that the quality of the wine is greatly determined by the combination of variables and the amounts of components in it. In addition to that, the lack of 'NA' values in the dataset made it much easier to handle data without using many filters.

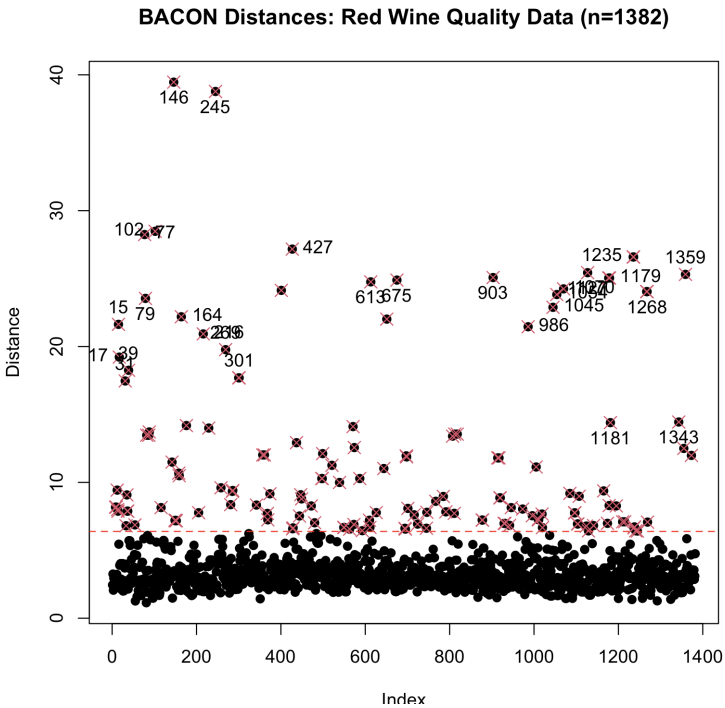
5. **Appendix:**


xed.acidit
latile.acid
citric.acid
sidual.sug
chlorides
sulfur.dio
sulfur.dio
density
pH
sulphates
alcohol
quality


xed.acidit
latile.acid
citric.acid
sidual.sug
chlorides
sulfur.dio
sulfur.dio
density
pH
sulphates
alcohol
quality


xed.acidit
latile.acid
citric.acid
sidual.sug
chlorides
sulfur.dio
sulfur.dio
density
pH
sulphates
alcohol
quality


BACON Distances: Red Wine Quality Data (n=217)
Distance
Index
0
10
20
30
0
50
100
150
200
161
160
20
7
12
39
3
14
40
49
77
70
90
149
167
178
191
192
209


BACON Distances: Red Wine Quality Data (n=1382)
Distance
Index
0
10
20
30
40
0
200
400
600
800
1000
1200
1400
146
245
102
427
15
79
164
17
39
301
613
675
903
986
1045
1179
1235
1268
1359
1181
1343