Wine dataset Multivariate analysis

1 Introduction

1.1 Background

Wine is a product that can vary in price and in quality ; some varieties are akin to cheap consummers goods while others are luxury products. In that context, it can be hard for consummers to identify which are more appropriate for certain occasions, or which are worth spending more or less money on ; similarly, for producers, the task of setting up a price can be made harder. On both sides, certifications are very important and thus, there is a real need to build trustworthy models to evaluate wines quality: it would bring clarity to consumers and recognition to producers. Moreover, by identifying the most important factors, the latter could turn their focus on these aspects and find more efficient ways to improve their wine's rating.

1.2 Variables

Our project is designed to harness the rich insights of a comprehensive public dataset, featuring detailed attributes and quality assessments of wine samples from Minho, a renowned wine-producing region in north-west Portugal. The dataset encompasses an extensive period of data collection, from May 2004 to February 2007, and includes an impressive total of 6,498 records related to both red and white wine variants, dissected across 11 key attributes. This meticulous compilation offers an unparalleled opportunity to delve into the nuances that distinguish wine quality, presenting a foundation for robust analysis and potential advancements in wine science. The dataset provides a granular look at the chemical composition and sensory attributes of wine, each variable casting light on its potential impact on overall quality. These attributes include:

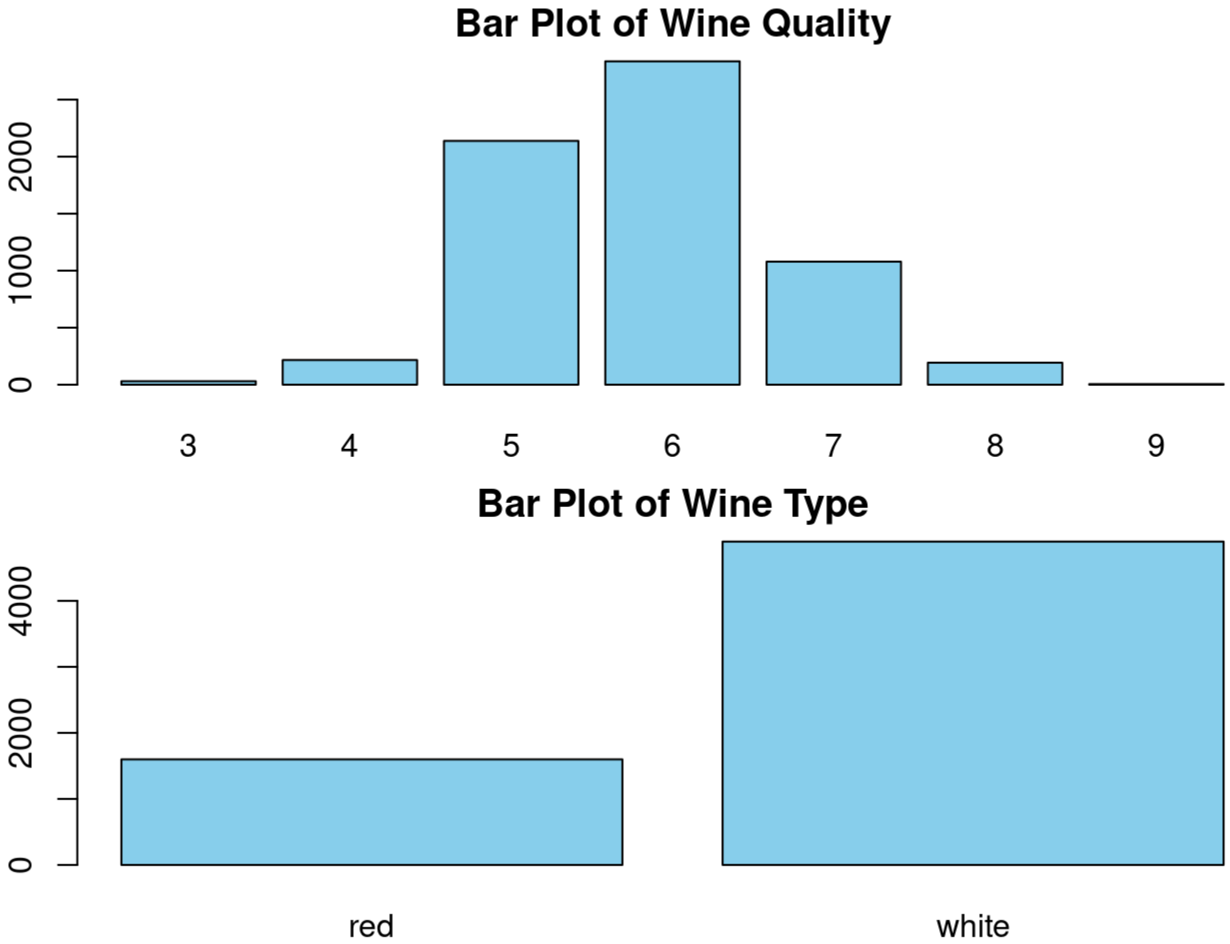
- **Fixed Acidity** (X_1, numeric): Reflects the concentration of nonvolatile acids (such as tartaric acid) that remain fixed during winemaking and contribute to the wine's structure.
- **Volatile Acidity** (X_2, numeric): Measures the volatile acids (like acetic acid), where excessive levels can mar wine with an undesirable vinegar taste.
- **Citric Acid** (X_3, numeric): Although present in small amounts, citric acid can enhance the wine's freshness and flavor profile.
- **Residual Sugar** (X_4, numeric): Indicates the sugar level post-fermentation, affecting sweetness.
- **Chlorides** (X_5, numeric): The measure of salt content in wine, impacting taste.
- **Free Sulfur Dioxide** (X_6, numeric): Represents the portion of sulfur dioxide not bound to other molecules, crucial for preserving wine's freshness and inhibiting microbial growth.
- **Total Sulfur Dioxide** (X_7, numeric): The total concentration of sulfur dioxide, encompassing both free and bound forms, essential for wine longevity.
- **Density** (X_8, numeric): Reflects the wine's density, which correlates with its alcohol and sugar content.
- **pH** (X_9, numeric): A vital indicator of acidity, influencing taste, color, and stability.
- **Sulphates** (X_{10}numeric): Pertains to added sulphates like potassium sulphate, affecting microbial stability and antioxidant properties.
- **Alcohol** (X_{11}, numeric): The alcohol percentage, directly influencing flavor and body.
- **Quality** (X_{12}, discrete): An assessment of wine quality on a scale from 0 to 10, based on sensory

evaluation.

- **Wine Type** (X_{13}, categorical): Distinguishes between red and white wine variants.

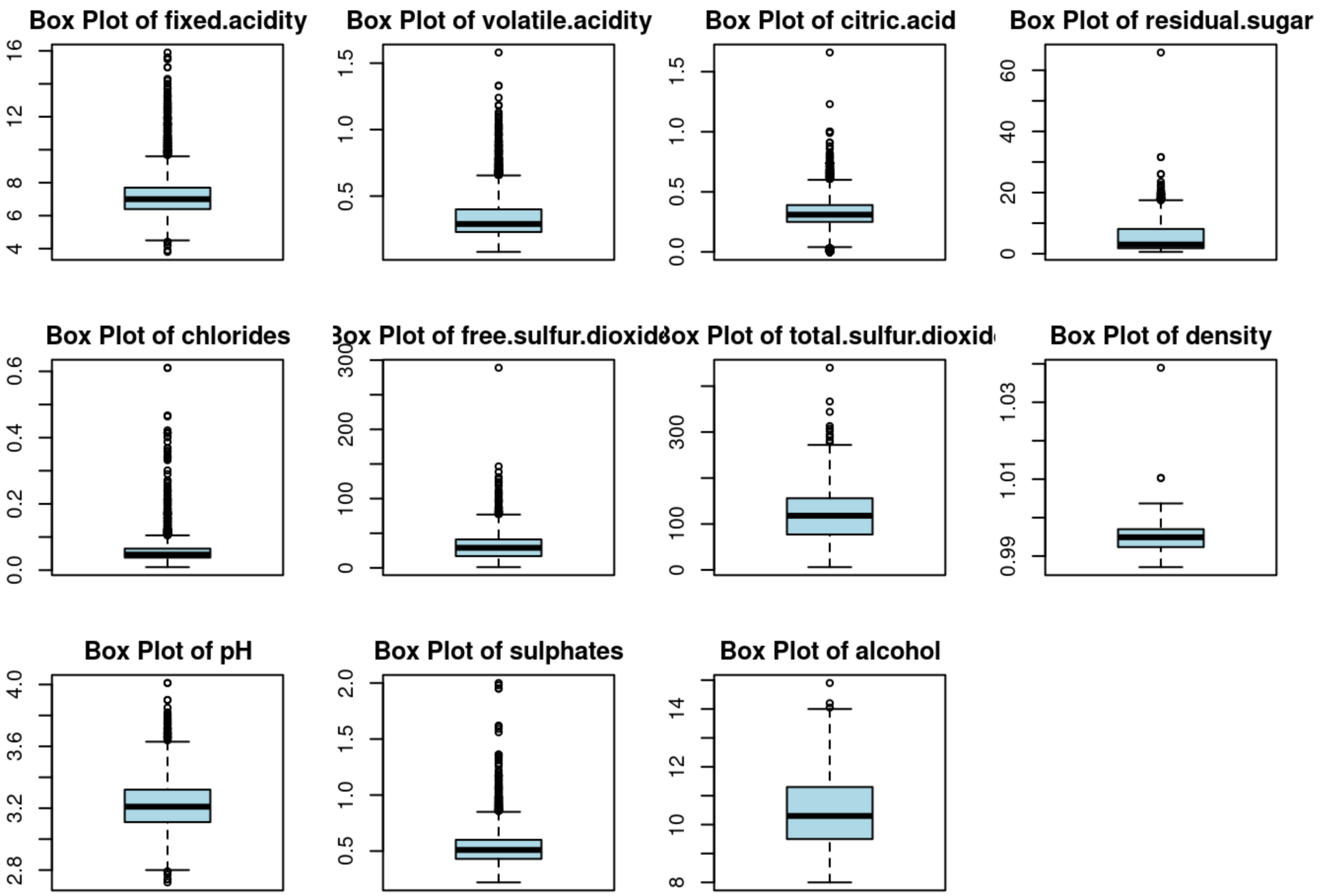
By exploring these attributes, our project aims not only to dissect the complex interplay of factors influencing wine quality but also to contribute valuable insights to the wine industry, enhancing our understanding of what makes a wine stand out.

1.2.1 barplot



White wine is the most common wine type recorded. In terms of wine quality, grades 5, 6 and 7 are the most common. They make up more than half of the data

1.2.2 boxplot



- **Fixed Acidity**: Appears to have a fairly symmetric distribution but with a few high outliers.
- **Volatile Acidity**: Has a median closer to the third quartile, suggesting a slight skew towards lower values.
- **Citric Acid and Residual Sugar**: Both have a few outliers, suggesting some wines have much higher levels of these characteristics than typical.
- **Chlorides**: This variable has a wide spread of outliers, indicating that while most wines have a similar chloride content, a few have much higher amounts.
- **Free Sulfur Dioxide**: Shows a positive skew with a long upper whisker and several outliers, indicating that most wines have lower values with a few exceptions.
- **Total Sulfur Dioxide**: Has many outliers on the higher end, suggesting that most wines have a lower total sulfur dioxide content, but there are several with significantly higher amounts.
- **Density, pH, Sulphates, Alcohol**: These plots also show the variation in wine samples, with a few outliers.

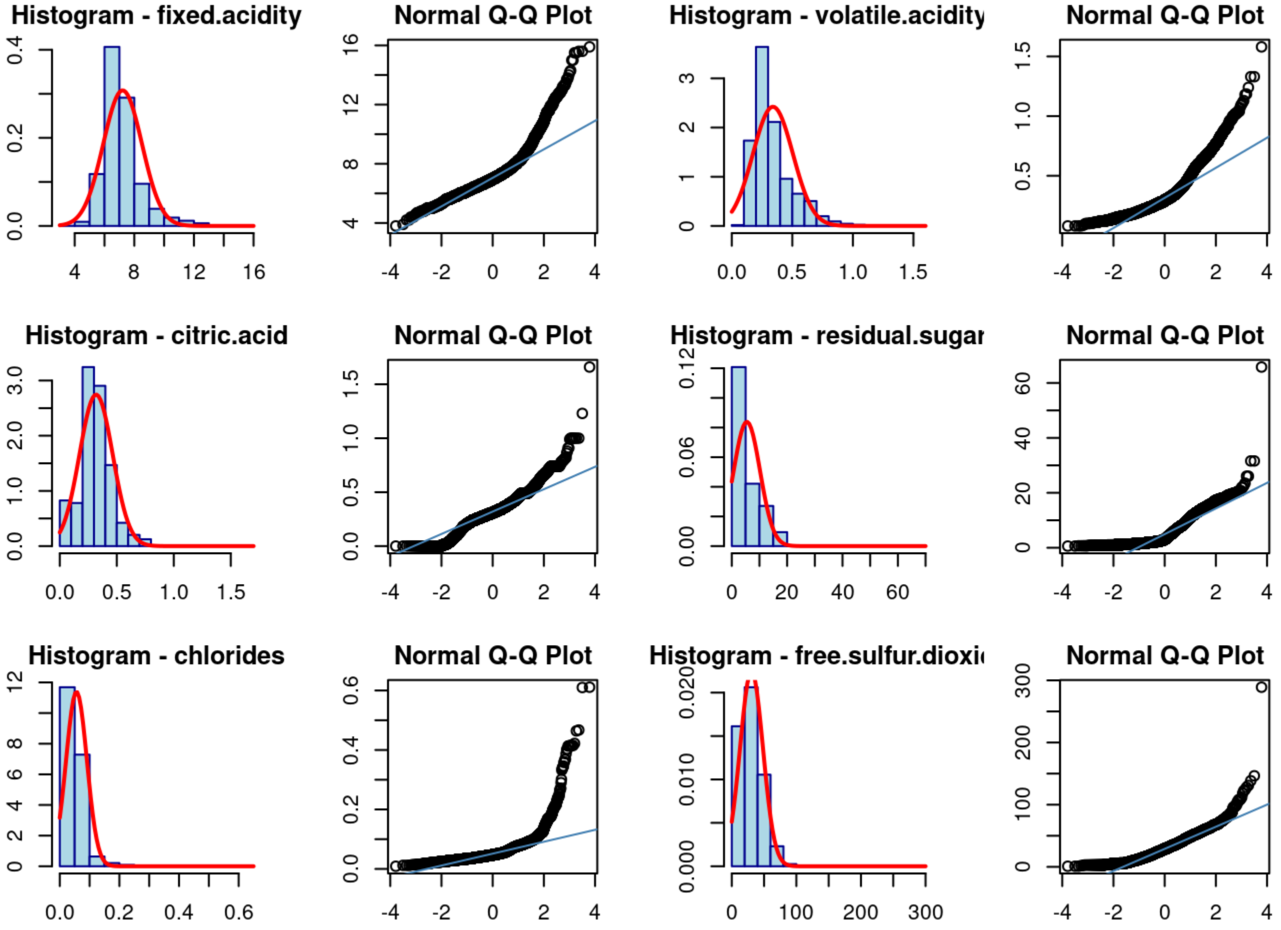
2 Data Normalization

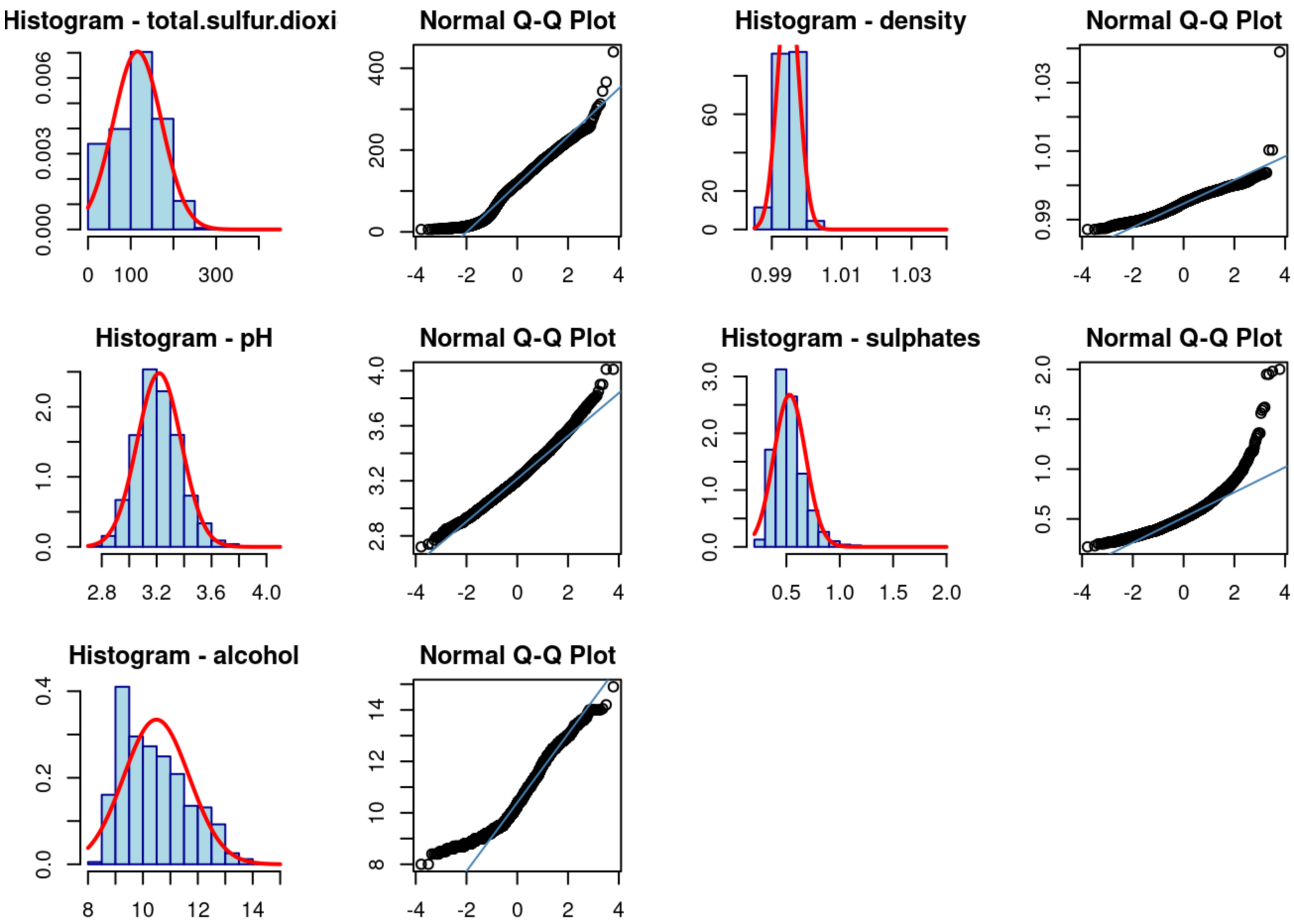
2.1 Univariate normality

For each variable in the dataset, a histogram was generated alongside a superimposed normal distribution curve to visualize the distribution of the data. The mean and standard deviation of each variable were calculated and used to plot the normal curve.

Q-Q plots were also created to assess the quantiles of the wine characteristics against the quantiles of a normal distribution. A line (in steelblue) was added to each plot to provide a reference for assessing normality. The closer the data points are to this line, the more normally distributed they are.

The r_Q values, which represent the correlation between the sample quantiles and the theoretical quantiles of the normal distribution, were calculated using the `cor` function on the Q-Q plot data points.





Upon visual inspection of the histograms and the Q-Q plots, it appears that most variables exhibit some deviation from normality:

1. Fixed Acidity, Volatile Acidity, and Citric Acid show histograms with a single peak, but the distribution is not perfectly symmetrical.
2. Residual Sugar, Free Sulfur Dioxide, and Total Sulfur Dioxide display a right-skewed distribution, indicating the presence of outliers or a long tail to the right.
3. Density and pH are relatively closer to a normal distribution, but slight deviations are evident.
4. Sulphates and Alcohol concentrations show considerable variance from normality, especially at the tails of the distribution.

The r_Q test results are as follows:

- Fixed Acidity: 0.9379
- Volatile Acidity: 0.9358
- Citric Acid: 0.9824
- Residual Sugar: 0.9079
- Chlorides: 0.7861
- Free Sulfur Dioxide: 0.9699
- Total Sulfur Dioxide: 0.9129
- Density: 0.9837
- pH: 0.9957
- Sulphates: 0.9480
- Alcohol: 0.9765

Values closer to 1 suggest a stronger agreement with a normal distribution. All of them are less then cretical value 0.9935 at significance levels = 0.01.

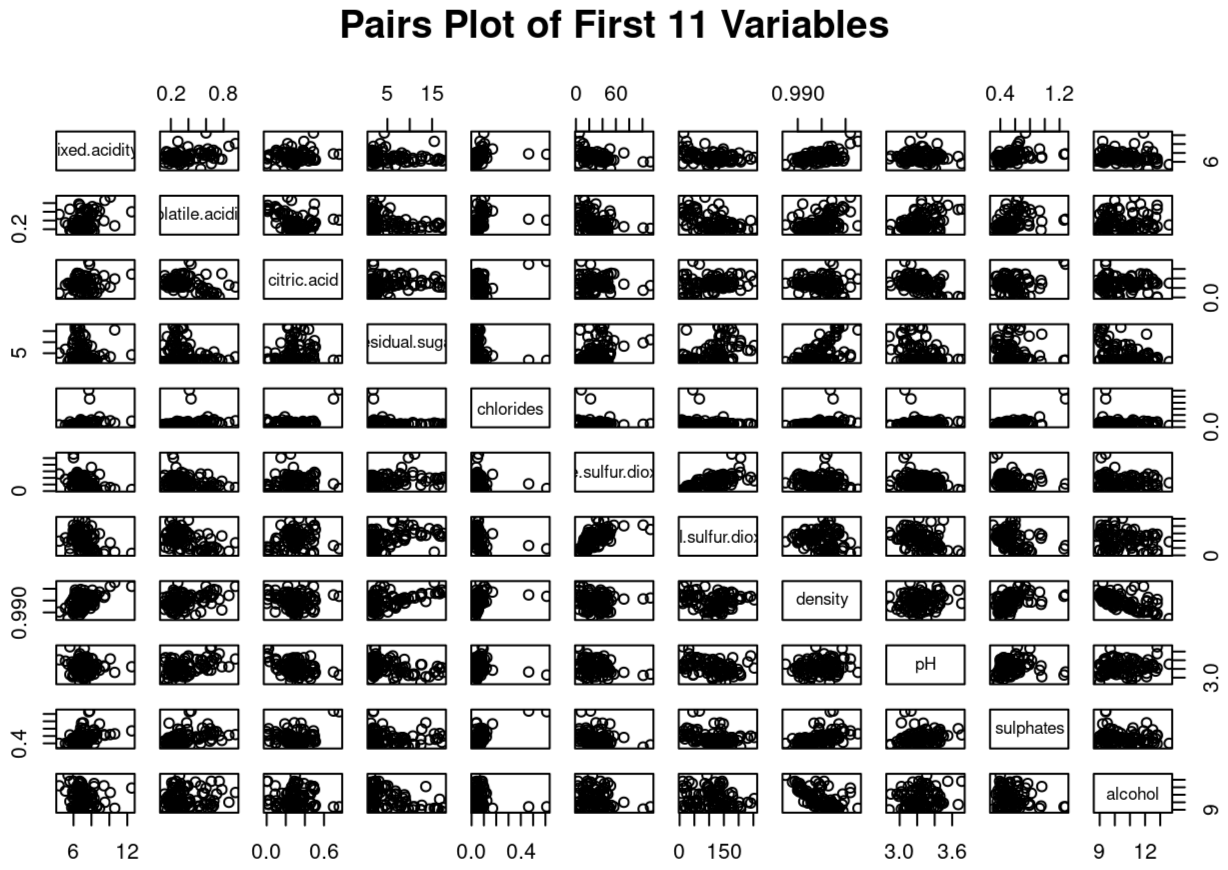
Further investigation may be required to address the non-normality, possibly through transformations or non-parametric methods.

We create a function `transform_check_normality` to apply five transformations (logarithmic, square root, inverse, exponential, and rank) to each variable in the dataset. The Shapiro-Wilk test was employed to check for normality post-transformation. Here is the result for `fixed.acidity` .

The Shapiro-Wilk test results for all the variables under all transformations showed p-values less than the threshold of 0.05, suggesting that none of the transformations resulted in a normal distribution at a 95% confidence level. The only exceptions were for pH under logarithmic and square root transformations, where the p-values indicated a failure to reject the null hypothesis of normality. However, given the stringent criteria for normality (α = 0.01), even these variables would not be considered normally distributed.

2.2 Bivariate Normality

Pairs Plot To investigate the bivariate relationships between the first 11 variables of a wine dataset to assess patterns, correlations, and potential normality, a pairs plot was generated using a random sample of 100 observations from the wine dataset. The plot includes scatterplots for each variable pairing along with histograms for the individual variables.



A

sample of 100 observations was randomly selected from the dataset to ensure variability and computational efficiency. The R programming language was utilized, leveraging its `pairs` function to create a comprehensive grid of scatterplots. Each plot juxtaposed two different variables, while histograms were generated for each individual variable along the diagonal of the grid.

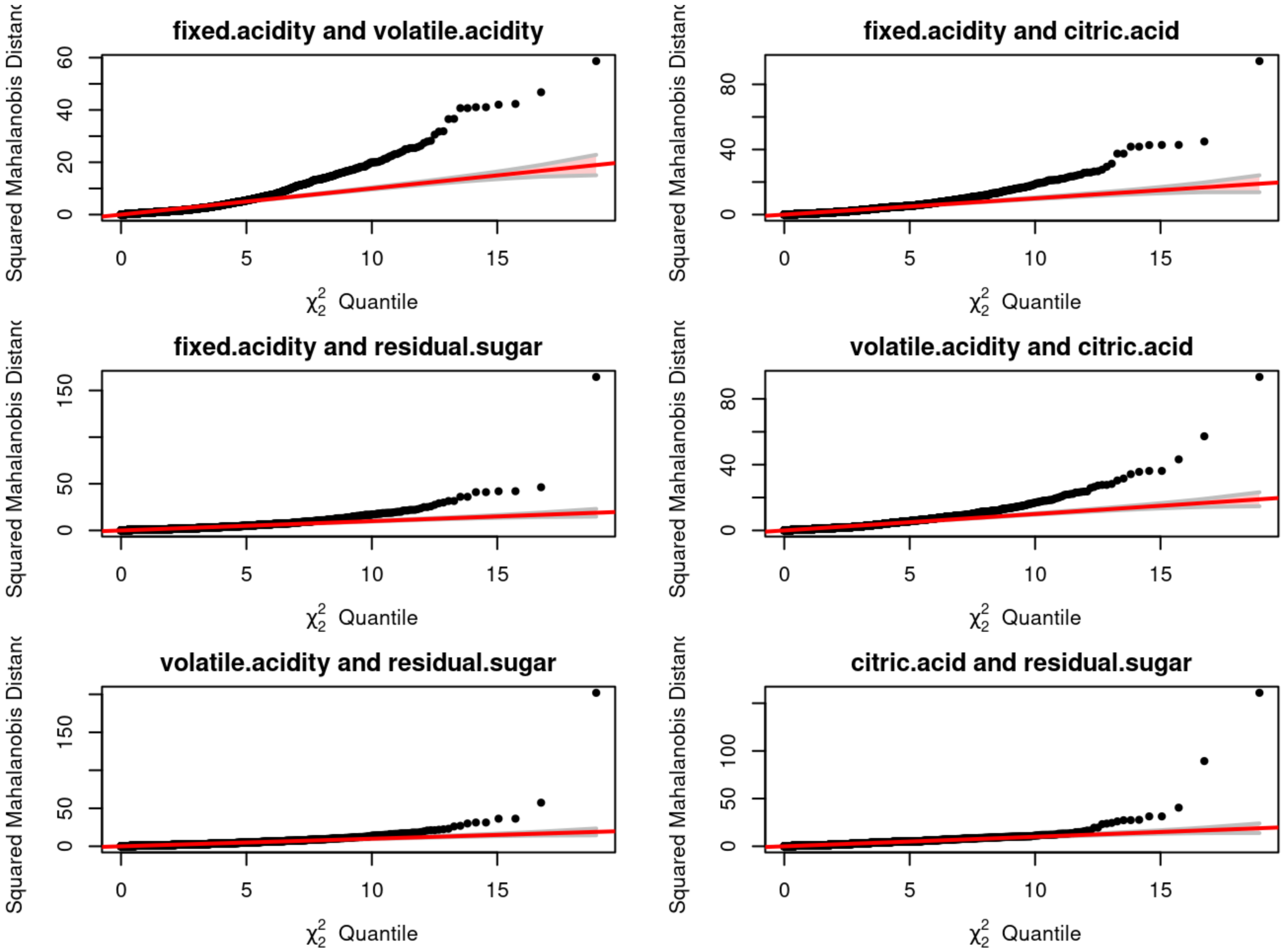
The pairs plot elucidates several notable features within the data:

- **Linear Trends**: Certain pairs, such as ‘fixed.acidity’ and ‘citric.acid’, exhibited a positive linear trend, suggesting a potential direct relationship.
- **Heteroscedasticity**: A few scatterplots indicated heteroscedastic patterns, where the variability of one variable seemed to change along with the level of another variable.
- **Variable Independence**: Some pairs displayed a lack of discernible patterns, indicating possible statistical independence between these variables.
- **Anomalous Observations**: Outliers were present in numerous bivariate scatterplots, potentially impacting correlation and regression analyses.
- **Univariate Distributions**: The histograms revealed varying distributions across the variables, with most deviating from normality, as previously established in the univariate normality assessment.

The pairs plot has provided valuable insights into the complex interplay of wine characteristics. The exploratory analysis has set the groundwork for subsequent, more rigorous statistical testing and modeling.

chi-square Plot

A chi-square plot analysis was conducted to evaluate the association between pairs of variables within a wine dataset. The visual inspection aimed to detect deviations from expected distributions under the assumption of bivariate normality.



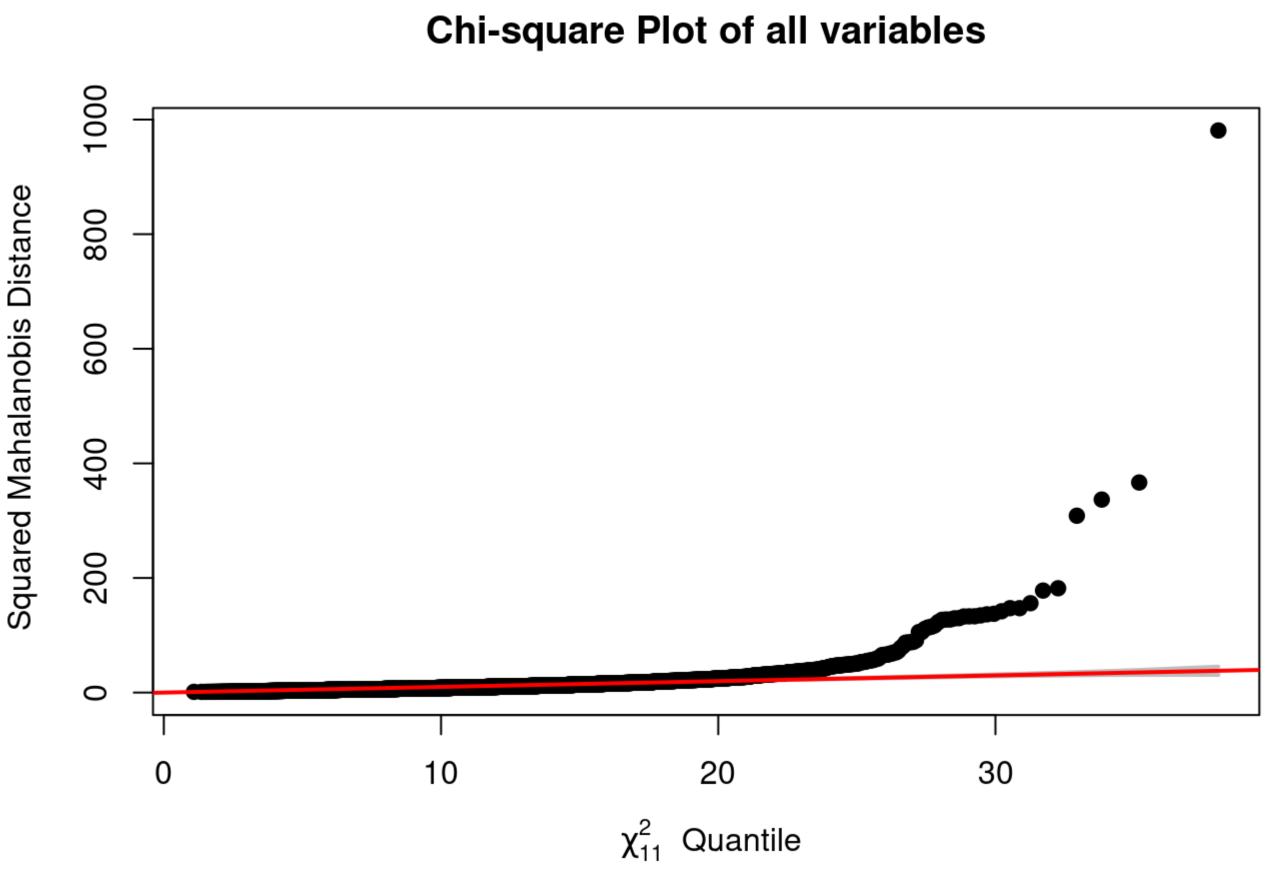
The chi-square plots exhibited the following trends across the variable pairs:

- A general upward trajectory, indicating a positive association between the variables.
- Several plots displayed pronounced deviations from the reference line towards the higher quantiles, signaling potential outliers or non-normal bivariate distributions.

The chi-square plots serve as diagnostic tools for identifying multivariate outliers and assessing the assumption of bivariate normality. The substantial divergence in some plots raises questions about underlying distributions and potential influential factors.

2.3 Check all variables normality

A chi-square plot was constructed for the first 11 variables of the wine dataset using the `cqplot` function.



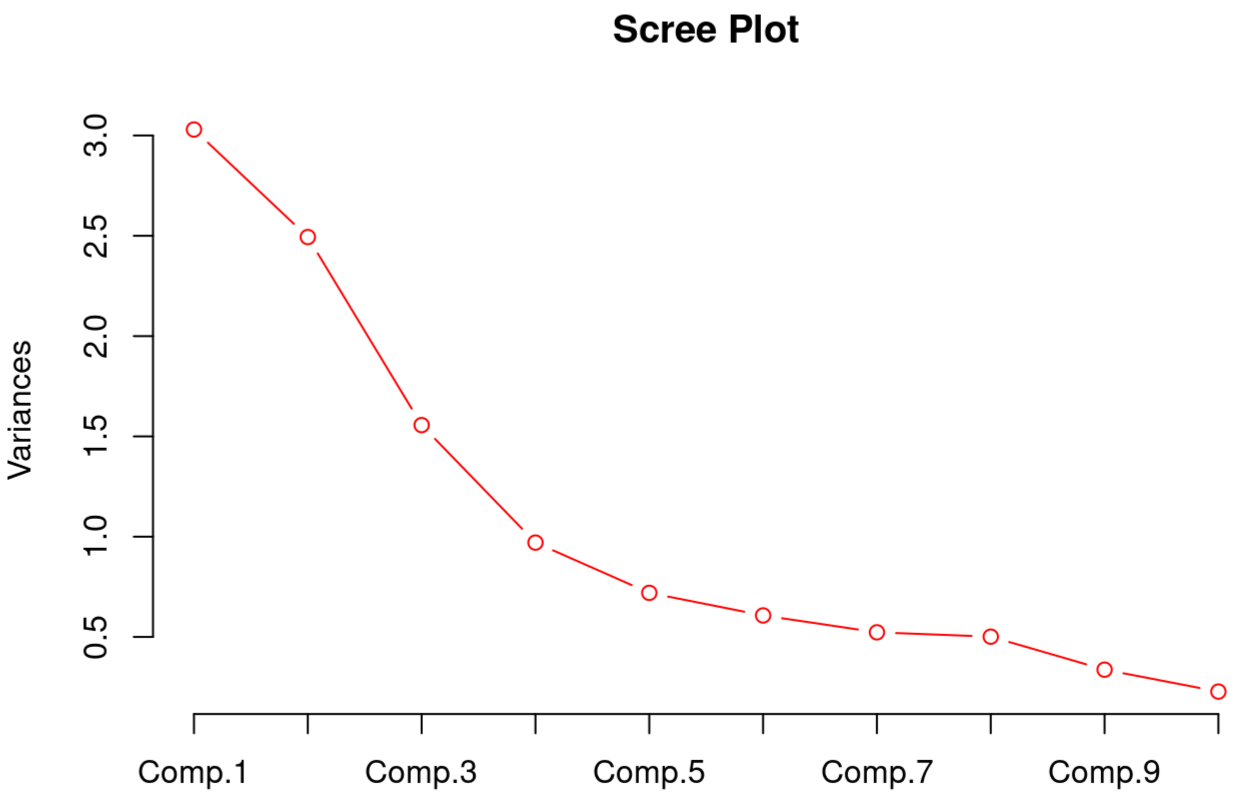
The chi-square plot showed an increase in the squared Mahalanobis distances with increasing quantiles. Notably, several points lay significantly above the reference line, indicating the presence of outliers that deviate from a multivariate normal distribution.

3 Models

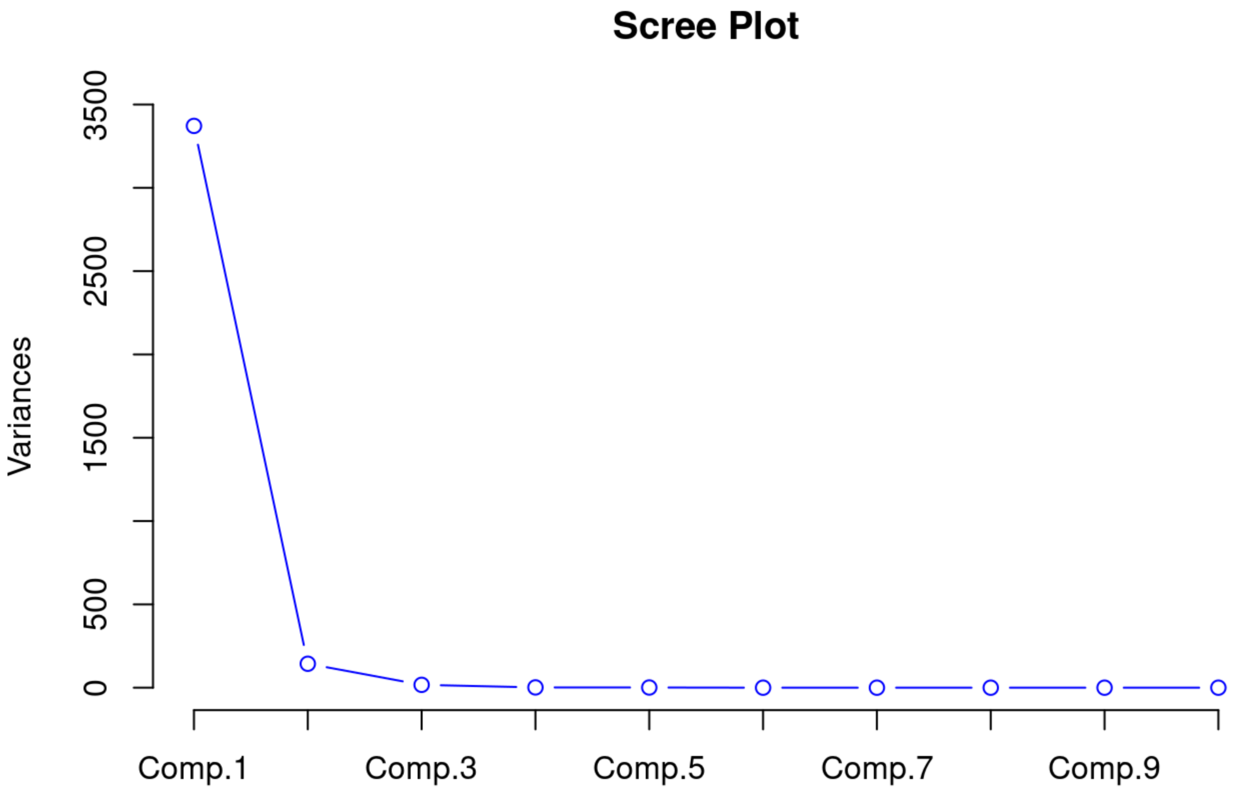
3.1 Data Reduction or Structural Simplification

3.1.1 PCA

In our analysis, we turn to Principal Component Analysis (PCA) as a means to interpret our dataset more effectively by reducing its dimensionality while preserving as much variation as possible. We set a pre-defined cutoff of 90% variation explained to guide our approach selection.



PCA based on the correlation matrix revealed that the first principal component accounts for only 9% of the total sample variation, necessitating the inclusion of the first 7 principal components to explain 90% of the total variance. While this is a substantial proportion, the significant number of principal components required indicates limited reduction in dimensionality. Alternatively, examination of the scree plot suggests a somewhat sharp drop at component 4. However, the first 4 principal components only explain 36% of the total variation, which is suboptimal.



Conversely, PCA based on the covariance matrix demonstrated that the first principal component explains a remarkable 95% of the total sample variance. Consistent with this finding, the scree plot reveals a clear elbow at component 1, indicating that further components contribute minimally to the total variance. Therefore, the first principal component alone can effectively replace the original 11 features.

However, it’s essential to note that results obtained from the covariance matrix may be misleading, as features with larger values disproportionately influence the analysis. Therefore, we opt for the PCA derived from the correlation matrix for its more balanced representation.

In conclusion, PCA based on the correlation matrix offers a more insightful interpretation of our dataset, facilitating a meaningful reduction in dimensionality and providing valuable insights into the underlying relationships among the features.

3.1.2 Factor Analysis

In our investigation of highly correlated features within our dataset, we sought to uncover potential latent factors underlying these correlations. With 11 features, our covariance matrix contains 66 unique entries. Considering a factor model with 2 latent factors, we can reduce the dimensionality of our analysis to 33 parameters, effectively simplifying the interpretation of our data.

Using two-factor models derived from principal component analysis (PCA) and maximum likelihood (ML) methods, we examined the factor loadings to identify the latent features associated with our variables. Applying a cutoff of 0.5 for correlation strength, we gained insights into the relationships between the factors and features.

From the PCA-based factor analysis: - Factor 1 is strongly correlated with volatile acidity (0.663), residual sugar (-0.602), free sulfur dioxide (-0.750), total sulfur dioxide (-0.848), and sulphates (0.512). This latent feature primarily captures variations in these variables. - Factor 2 exhibits strong correlations with fixed acidity (0.531), residual sugar (0.521), density (0.922), and alcohol (-0.734), indicating its association with these variables.

```
## 
## Loadings:
##                      PC1    PC2   
## fixed.acidity         0.416  0.531
## volatile.acidity      0.663  0.186
## citric.acid          -0.265  0.289
## residual.sugar       -0.602  0.521
## chlorides             0.505  0.498
## free.sulfur.dioxide  -0.750  0.114
## total.sulfur.dioxide -0.848  0.138
## density                      0.922
## pH                    0.381 -0.246
## sulphates             0.512  0.303
## alcohol               0.185 -0.734
## 
##                  PC1   PC2
## SS loadings    3.030 2.494
## Proportion Var 0.275 0.227
## Cumulative Var 0.275 0.502
```

Similarly, from the ML-based factor analysis: - Factor 1 shows strong correlations with residual sugar (0.553), density (0.998), and alcohol (-0.688), suggesting that this latent feature primarily represents variations in these variables. - Factor 2 is strongly correlated with volatile acidity (-0.509), residual sugar (0.568), free sulfur dioxide (-0.76), and total sulfur dioxide (-0.888), indicating its association with these variables.

In both methods, the residuals are comparable and produce smaller values, indicating satisfactory model fit. Therefore, we can confidently consider the results obtained from either method in this case, facilitating a deeper understanding of the underlying factors driving the correlations among our features.

3.2 Grouping or Discrimination

3.2.1 Test the equality of covariance matrices (red vs white)

We will now try to find a way to classify our wines between red and white based on their chemical properties and composition and on their quality score.

To chose between a linear or quadratic discriminant analysis approach (lda or qda), we started by checking if the covariance matrices for each groupe (or) were equal. To do so, we performed the following test: H_0: $\sum_{white.wine} = \sum_{red.wine}$ against H_1: $\sum_{white.wine} \neq \sum_{red.wine}$. Define the Box's M test statistic $M =-2ln() $ (where Λ is the likelihood ration test statistics) and $u = [\sum_{l=1}^{g} \frac{1}{n_l-1} - \frac{1}{\sum_{l=1}^{2}(n_l-1)}][\frac{2p^2+3p-1}{6(p+1)(g-1)}]$ where p=11 (number of predictors) and g = 2 (number of classes), we know that $(1-u)M*{p(p+1)(g-1)/2}^2 =* {66}^2 $ is our test statistic. The resulting p-value is 1, thus, we fail to reject H_0: the data does not provide evidence going against the equality of the means (with confidence level higher than 95%). Therefore, a lda model would be more appropriate in this situation.

3.2.2 LDA

wine_type is a binary variable and, for analysis purpose, we code wine_type as follows: red = 1, white = 0. This conversion is made within the dataset before fitting the model. Moreover, to evaluate the performance of our model, we selected a random sample containing 80% of the wine data that will be used as a training set (wine_train) and the other 20% will be used as a validation set (wine_validation). This will allow us to use cross validation. Because we only need quantitative data for these analysis, we will not consider as a predictor.

```
##         0         1 
## 0.9969199 0.9892390
```

```
## [1] 0.9949971
```

After building the model, we want to evaluate its performance in two ways. First, we want to check that it performs well in predicting the data that have been used to build it (wine_train). When generating the cross validation matrix, we can see that more than 99% of white wines and more than 98% of red wines were correctly classified. On average, more than 99% of all predictions were correct. Even though these results are very high, we might fear that such good results could be due to overfitting ; that is the reason why we also need to check its performance on the validation set.

Surprisingly, on new data, the lda model performs even better. More than 99% of the predictions made on both red and white wines are correct and, thus, on average, almost all predictions are also correct.

3.2.3 QDA

```
##         0         1 
## 0.9838254 0.9915450
```

```
## [1] 0.9857583
```

```
##         0         1 
## 0.9840319 0.9798658
```

```
## [1] 0.9830769
```

Even though the linear separation was very satisfying and seemed to be the most appropriate, we might want to check if a quadratic discriminant analysis would also be able to capture the limit between the two categories.

The performance are still very good: between 98 and 99% of the predictions are correct on both white and red wines, when the predictions are made on the training dataset and it goes over 99% with red wine, when the validation set is used. On average, the qda model is correct 98% of the time. This is a highly satisfying result too.

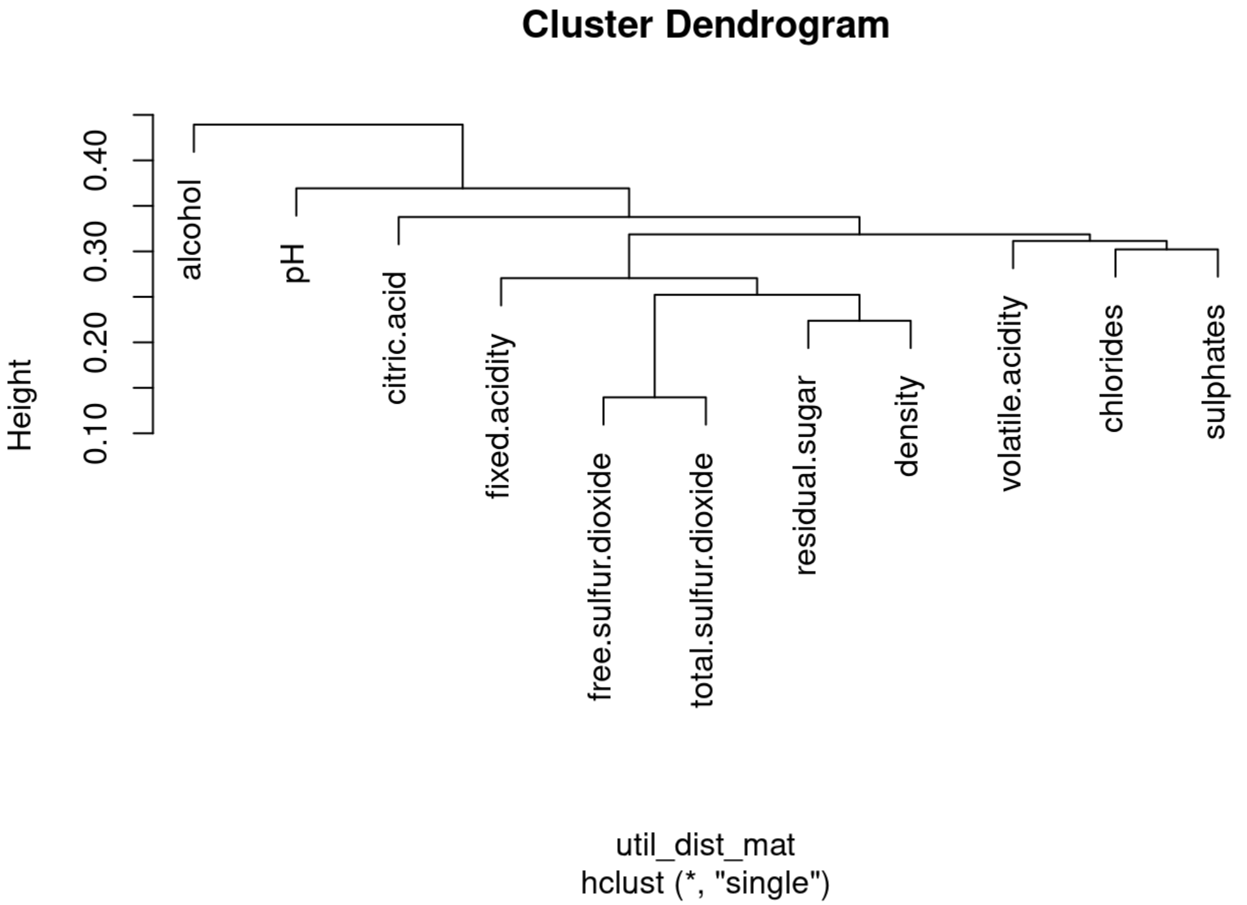
Since we have to make a choice between the two, however, we will select the lda model: not only does it performs better (even though the difference is not substancial), it is also simpler and more appropriate with the data. Indeed, we showed that the red_wine and white_wine covariance matrices, which is also a sign that the lda model is a better choice.

3.2.4 Cluster Analysis

3.2.4.1 Hierarchal Clustering

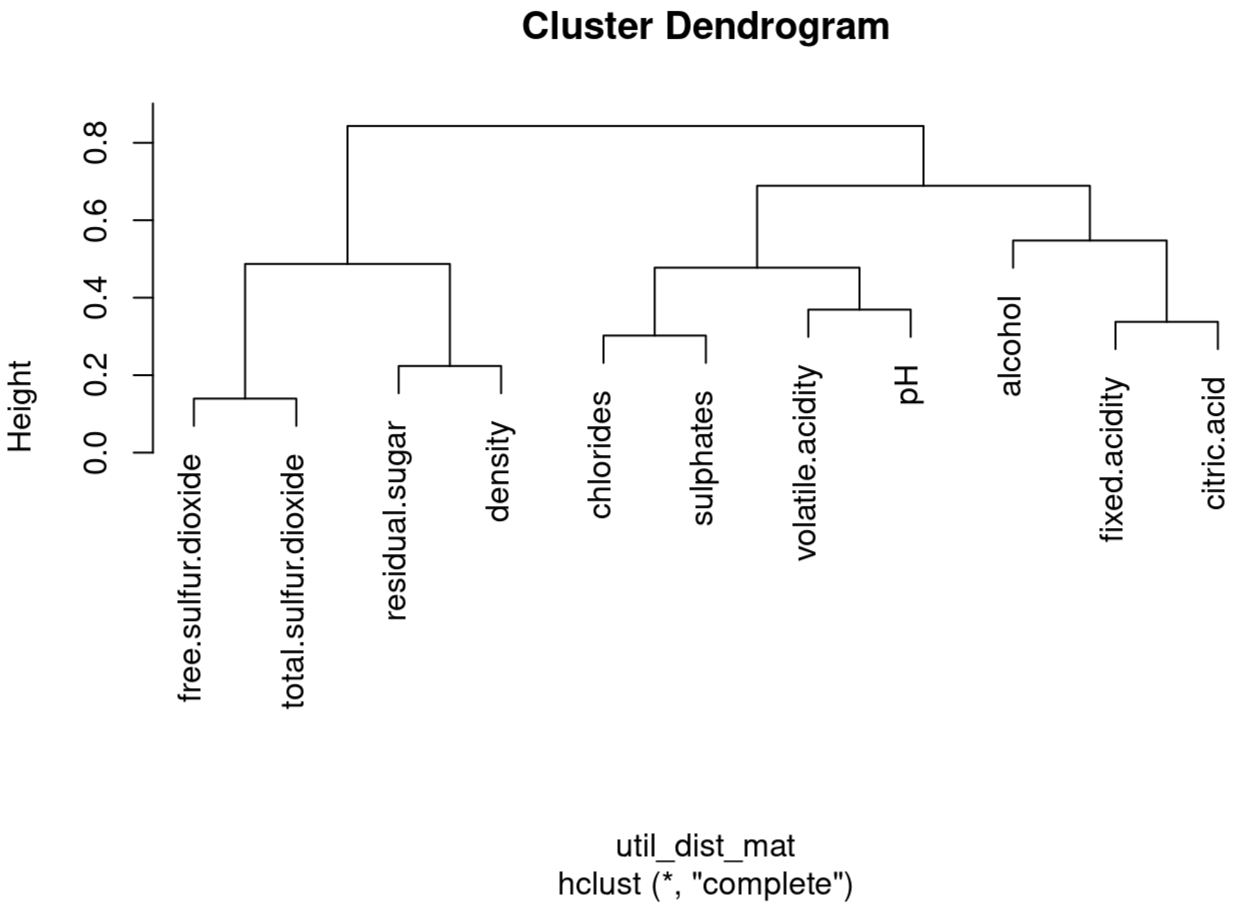
In this paper, hierarchical clustering analysis was performed on a dataset comprising various chemical properties of wine. The analysis utilized four different linkage methods—single, complete, average, and median—to explore the natural groupings of the variables based on their correlations. This section will discuss the results depicted in the deprograms generated by each linkage method.

Single Linkage (Minimum Linkage) In the single linkage method, the distance between two clusters is defined as the minimum distance between any single member of one cluster to any single member of the other cluster. This method tends to produce long, ‘stringy’ clusters. These clusters can sometimes be sensitive to outliers and may not perform well if the natural clusters are not well separated by distance.



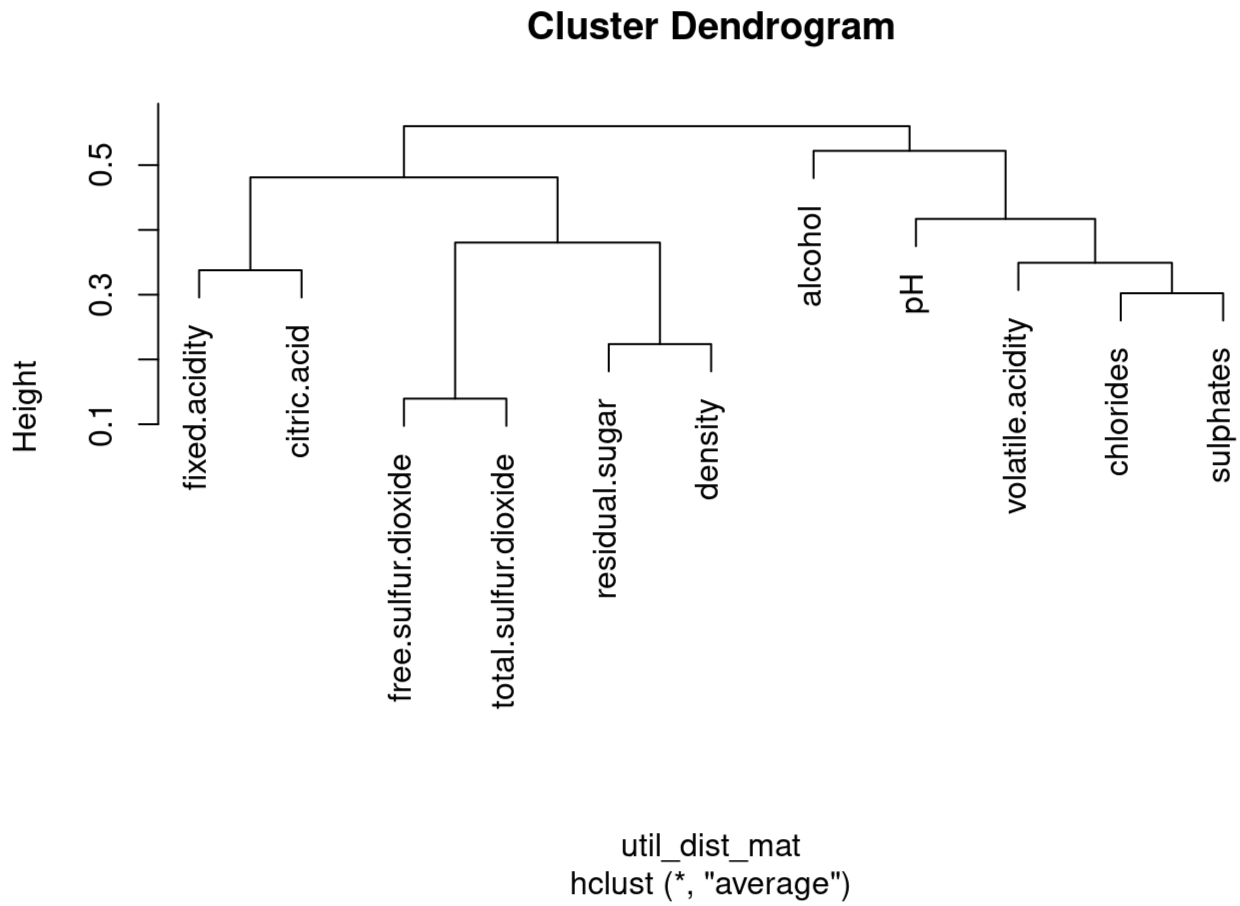
The dendrogram obtained from the single linkage method exhibits a distinct pattern where certain variables, such as ‘alcohol’ and ‘pH’, are the last to join the main cluster. This indicates that these variables have a less strong linear relationship with the other chemical properties in the dataset. Single linkage is known for its tendency to create ‘chains’ where clusters can merge at substantially different levels of similarity. This is observed in the significant height at which the ‘alcohol’ and ‘pH’ join, suggesting these properties behave quite differently from the others or might be influenced by different factors in the winemaking process.

Complete Linkage (Maximum Linkage) The complete linkage method defines the distance between two clusters as the maximum distance between any member of one cluster to any member of the other cluster. This method tends to find compact clusters of approximately equal diameter. It is less susceptible to noise and outliers compared to single linkage.



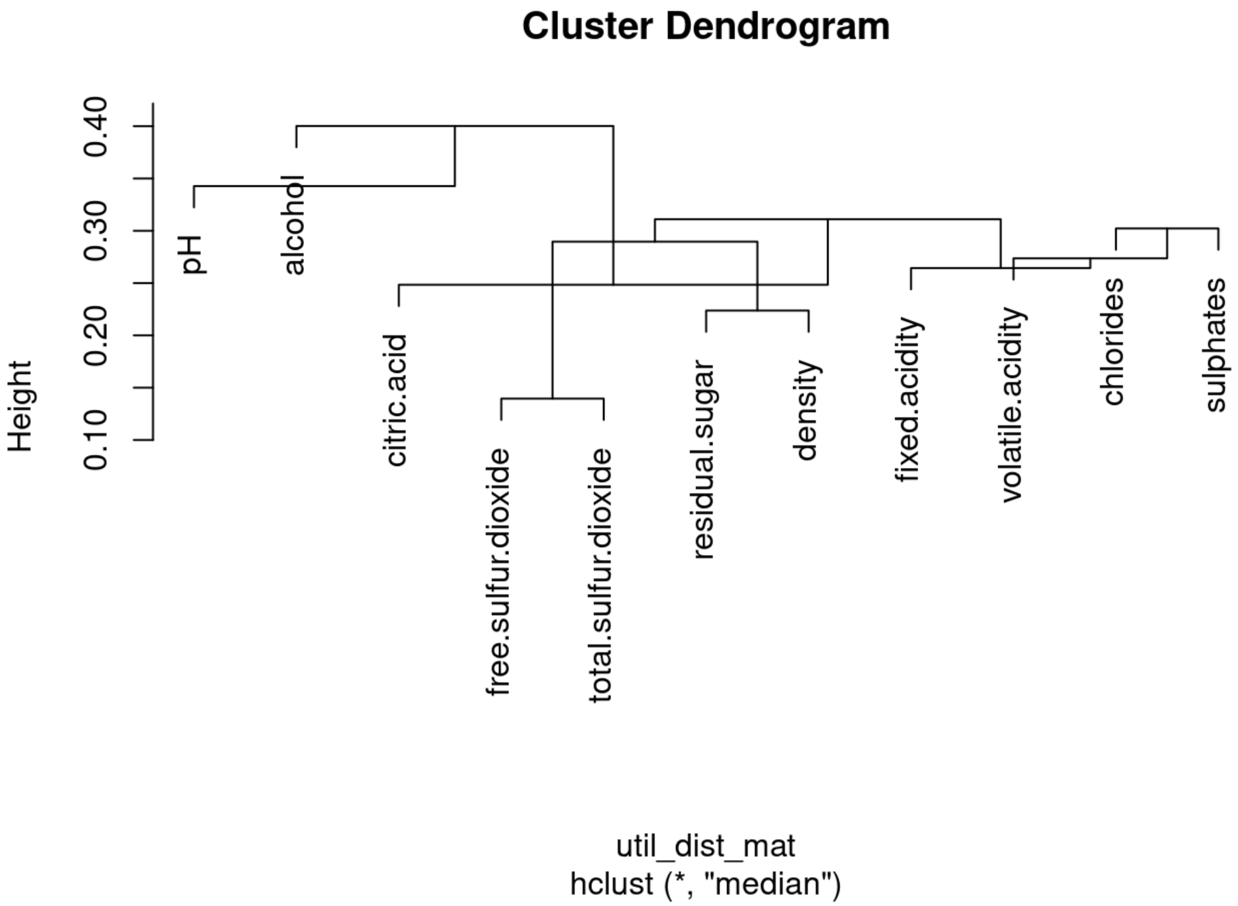
In contrast, the complete linkage dendrogram shows a more balanced hierarchical structure. The variables ‘density’ and ‘chlorides’ form one of the initial clusters, while ‘fixed acidity’, ‘citric acid’, and ‘alcohol’ cluster together at a higher level. This suggests that ‘density’ and ‘chlorides’ share a similar profile regarding their influence on wine characteristics. Complete linkage avoids the chaining effect seen in single linkage, indicating that ‘fixed acidity’, ‘citric acid’, and ‘alcohol’ might share a unique relationship distinct from other variables, potentially in how they contribute to the flavor profile of the wine.

3. Average Linkage (Mean Linkage) The average linkage method defines the distance between two clusters as the average distance between each member of one cluster to every member of the other cluster. It provides a balance between the sensitivity of the single linkage to outliers and the tendency of complete linkage to force clusters to be compact.



The average linkage dendrogram presents a more balanced clustering approach, where the clusters form at intermediate levels of similarity. This method mitigates the influence of outliers and does not force clusters to be overly compact. The cluster comprising ‘fixed acidity’, ‘citric acid’, and ‘free sulfur dioxide’ indicates that these variables may be moderately related, contributing to a common characteristic of the wine, such as its balance between tartness and preservability.

4. Median Linkage Median linkage uses the median distance between elements of the two clusters. It’s a less common method but can sometimes provide a balance that is not as tight as complete linkage and less sensitive to outliers than single linkage.



Finally, the median linkage dendrogram is relatively similar to the average linkage dendrogram but with slight variations. For instance, ‘citric acid’ appears closer to ‘fixed acidity’, suggesting that these variables may share a median level of similarity that is more pronounced than what is captured by average linkage. This can be interpreted as ‘citric acid’ having a more central role in relation to ‘fixed acidity’ in terms of their joint effect on wine properties.

Across all dendrograms, there are consistent patterns worth noting. ‘Alcohol’ and ‘pH’ often appear as outliers, suggesting they have unique roles in wine chemistry not closely related to other measured variables. Conversely, ‘fixed acidity’, ‘citric acid’, and ‘density’ frequently cluster together, which may indicate their combined influence on the acidity and body of the wine.

The results revealed by these hierarchical clustering analyses offer valuable insights into the relationships between the chemical properties of wine. Such information could be instrumental in the field of oenology for improving wine quality control and enhancing production processes. #### Kmeans This part includes a segment on the application of K-means clustering to a wine dataset, which contains various chemical attributes of wines and their types (red and white). K-means clustering was performed with the goal of dividing the dataset into clusters that capture inherent groupings based on the wines’ chemical properties.

For the clustering process, a decision was made to set the number of clusters (k) to 3. The clustering algorithm was initialized 100 times (`nstart = 100`) to ensure convergence to a good solution.

Upon applying the K-means algorithm, the dataset was partitioned into three clusters. When cross-tabulating these clusters with the `wine_type` variable, the following distribution was observed:

```
##
##           1    2    3
##   red    259 1329   11
##   white 2666  289 1943
```

This cross-tabulation reveals distinct patterns:

- **Cluster 1** is predominantly composed of red wine, with a significant but smaller proportion of white wine. This cluster may represent wines with a chemical profile more typical of red wines but still present in a subset of white wines.
- **Cluster 2** consists mainly of white wine, with a few red wines. The properties defining this cluster seem to be closely aligned with those typically found in white wines.
- **Cluster 3** has a majority of white wine but also includes a minimal number of red wines, possibly indicating a unique subset of white wines that share certain characteristics with some red wines.

The clustering results suggest that the inherent chemical characteristics captured in the dataset do align, to some extent, with the traditional wine type classifications of red and white. It is interesting to note that while there is a clear majority of one type of wine in each cluster, there is also representation from the other type. This suggests that the conventional dichotomy between red and white wines may not fully capture the complexity of wine chemistry.

3.3 Canonical Correlation Analysis

Canonical Correlation Analysis is a multivariate statistical method used to examine the relationships between two sets of variables. By identifying pairs of canonical variates—one for each set—that are maximally correlated, CCA helps to understand how one set of variables might predict or relate to another set.

Canonical Variables: - Canonical variables (or variates) are synthetic variables created as linear combinations of the original variables within each set. These are formulated so that their correlations across the sets are maximized. - The canonical variables for each set are derived as follows:

$$u_i = a_{1i}x_1 + a_{2i}x_2 + \ldots + a_{mi}x_m$$

$$v_i = b_{1i}y_1 + b_{2i}y_2 + \ldots + b_{ni}y_n$$

where a_{ji} and b_{ji} are coefficients optimizing the correlation between u_i and v_i.

Selection Criteria: - The number of canonical variates generated is the lesser of the number of variables in the two sets. Each pair of variates is orthogonal (independent) to the others within its own set. - The correlation coefficients (r_i) between the pairs of canonical variates provide a measure of their association.

Variable Sets:

- **Set 1 (Acidity Levels)**: Includes fixed acidity, volatile acidity, citric acid, and pH. These variables were chosen because they represent fundamental aspects of wine’s chemical nature that directly affect its taste, stability, and fermentation process.
- **Set 2 (Sugar and Sulfur Dioxide)**: Includes residual sugar, free sulfur dioxide, and total sulfur dioxide. This group is critical for understanding the wine’s sweetness, preservation, and anti-oxidative properties.
- **Set 3 (Other Chemical Properties)**: Includes chlorides, density, sulphates, and alcohol. These properties influence the wine’s preservation, body, and overall quality.

3.3.1 Acidity Levels & Sugar and Sulfur Dioxide

This analysis aims to explore and quantify the relationships between two critical sets of variables in wine data: **Acidity Levels** (comprising fixed acidity, volatile acidity, citric acid, and pH) and **Sugar and Sulfur Dioxide** (including residual sugar, free sulfur dioxide, and total sulfur dioxide).

```
## [1] 0.55497110 0.20137538 0.03044709
```

```
## $xcoef
##                        [,1]        [,2]       [,3]
## fixed.acidity    -0.7190255 -0.06027978  0.7028952
## volatile.acidity -0.3682300 -0.55872701 -0.9086635
## citric.acid       0.3076262 -0.12840639 -1.0265988
## pH               -0.4254979  0.93111171 -0.2156168
## 
## $ycoef
##                            [,1]        [,2]       [,3]
## residual.sugar       0.08087317 -1.04989404  0.4735366
## free.sulfur.dioxide  0.13898312  0.69940421  1.2592851
## total.sulfur.dioxide 0.85208294  0.03760409 -1.2640234
```

Canonical Correlations: The first three pairs of canonical variates yielded the following correlations:

1. **First pair**: 0.55497110 (moderate correlation)
2. **Second pair**: 0.20137538 (low correlation)
3. **Third pair**: 0.03044709 (negligible correlation)

These results indicate how well combinations of acidity-related variables correlate linearly with combinations of sugar and sulfur dioxide-related variables.

Canonical Coefficients:

- **Acidity Levels Coefficients (X Coefficients)**:
    - **First Canonical Variate**:
        - Fixed Acidity: -0.7190255
        - Volatile Acidity: -0.3682300
        - Citric Acid: 0.3076262
        - pH: -0.4254979
    - **Further Variates** show varying influence of these acidity variables, reflecting different aspects of their interaction with sugar and sulfur dioxide levels.
- **Sugar and Sulfur Dioxide Coefficients (Y Coefficients)**:
    - **First Canonical Variate**:
        - Residual Sugar: 0.08087317
        - Free Sulfur Dioxide: 0.13898312
        - Total Sulfur Dioxide: 0.85208294
    - **Further Variates** exhibit distinct patterns, indicating more complex and less pronounced relationships.

The analysis indicates a significant but moderate initial correlation, driven primarily by total sulfur dioxide and fixed acidity. This suggests that as acidity levels in wine change, they might be closely linked with changes in sulfur dioxide levels, which play a crucial role in wine preservation and stability.

The weaker correlations observed in the second and third variates suggest that while there are additional relationships between these sets of variables, they are less straightforward and possibly influenced by other factors not captured solely by these measurements.

Canonical Correlation Analysis between Acidity Levels and Sugar and Sulfur Dioxide in wine reveals that there are meaningful but complex interactions between these variables. Understanding these relationships can help in optimizing wine production to enhance both taste and longevity, particularly by managing acidity and preservative levels effectively.

3.3.2 Acidity Levels & Other Chemical Properties

The objective of this analysis is to explore and quantify the relationships between two sets of wine-related variables: Acidity Levels (fixed acidity, volatile acidity, citric acid, and pH) and Other Chemical Properties (chlorides, density, sulphates, and alcohol), utilizing Canonical Correlation Analysis (CCA).

```
## [1] 0.68820274 0.25636970 0.16665241 0.01562419
```

```
## $xcoef
##                        [,1]       [,2]       [,3]        [,4]
## fixed.acidity    -0.7482428  0.5611915  0.6568862  0.31191167
## volatile.acidity -0.4369648 -0.9916582 -0.5217603 -0.12422477
## citric.acid      -0.2119665 -0.1849292 -0.4426953 -1.10206743
## pH               -0.3663495  0.8226707 -0.6424545  0.03037317
## 
## $ycoef
##                 [,1]       [,2]       [,3]       [,4]
## chlorides -0.3366398 -0.9933596 -0.4561147 -0.0556623
## density   -1.0419215  0.3113337  0.7055710 -0.7262553
## sulphates -0.1874707  0.6706734 -0.4203028  0.8001993
## alcohol   -0.7362245  0.2939287 -0.3318446 -1.1410195
```

Canonical Correlations: The analysis produced the following canonical correlations between the variate pairs:

1. **First pair**: 0.68820274 (strong correlation)
2. **Second pair**: 0.25636970 (moderate correlation)
3. **Third pair**: 0.16665241 (weak correlation)
4. **Fourth pair**: 0.01562419 (negligible correlation)

These correlations reveal the varying degrees of linear relationships between the constructed variates of the two sets.

Canonical Coefficients:

- **Acidity Levels Coefficients (X Coefficients)**:
    - **First Canonical Variate**: Predominantly influenced by fixed acidity.
    - **Second Canonical Variate**: Volatile acidity is the major contributor.
    - **Third Canonical Variate**: Led by fixed acidity.
    - **Fourth Canonical Variate**: Citric acid is the dominant variable.

- **Other Chemical Properties Coefficients (Y Coefficients)**:
  - **First Canonical Variate**: Density has the most significant negative influence.
  - **Second Canonical Variate**: Chlorides are the main factor.
  - **Third Canonical Variate**: Density plays a pivotal role.
  - **Fourth Canonical Variate**: Alcohol is the major contributor.

The first canonical variate pair shows a strong correlation, indicating a significant relationship between a combination of acidity level variables (especially fixed acidity) and a combination of other chemical properties (dominantly density). This suggests that aspects such as the body and heaviness of wine, which are influenced by density, are closely linked to its acid content.

Subsequent variate pairs show decreasing correlations, with the fourth pair being almost negligible, indicating limited to no linear relationship for that combination of variables.

This Canonical Correlation Analysis has highlighted significant and varied relationships between acidity levels and other chemical properties of wine. The strongest correlation suggests a pivotal interaction between the wine's acidity and its physical characteristics such as density and alcohol content, which can greatly influence winemaking decisions and wine quality assessments.

3.3.3 Sugar and Sulfur Dioxide & Other Chemical Properties

The goal of this analysis is to understand the relationships between two sets of wine-related variables: **Sugar and Sulfur Dioxide** (residual sugar, free sulfur dioxide, total sulfur dioxide) and **Other Chemical Properties** (chlorides, density, sulphates, and alcohol), using Canonical Correlation Analysis.

```
## [1] 0.735553383 0.470342155 0.008172205
```

```
## $xcoef
##                             [,1]        [,2]       [,3]
## residual.sugar        1.1436557 -0.15769078  0.0154966
## free.sulfur.dioxide  -0.1007520 -0.06867401 -1.4420175
## total.sulfur.dioxide -0.3568999  1.11657329  0.9752441
## 
## $ycoef
##                  [,1]       [,2]       [,3]
## chlorides -0.2615726 -0.5746925  0.8495933
## density    1.3357433 -0.5414479 -0.3217285
## sulphates -0.3712604 -0.1999157 -0.2575060
## alcohol    0.4437359 -1.0041868 -0.6642965
```

Canonical Correlations: The canonical correlations for the first three pairs of variates are as follows:

1. **First pair**: 0.735553383 (strong correlation)
2. **Second pair**: 0.470342155 (moderate correlation)
3. **Third pair**: 0.008172205 (negligible correlation)

These values indicate the strength of the linear relationships between the canonical variates from each set.

Canonical Coefficients:

- **Sugar and Sulfur Dioxide Coefficients (X Coefficients)**:

    - **First Canonical Variate**: Dominated by residual sugar.
    - **Second Canonical Variate**: Strong influence from total sulfur dioxide.
    - **Third Canonical Variate**: Largely influenced by total sulfur dioxide and negatively by free sulfur dioxide.
- **Other Chemical Properties Coefficients (Y Coefficients)**:
    - **First Canonical Variate**: Primarily driven by density.
    - **Second Canonical Variate**: Notably influenced by alcohol and chlorides.
    - **Third Canonical Variate**: Dominated by chlorides.

The analysis reveals a strong initial correlation suggesting a significant link between the sugar/sulfur dioxide content and the density of the wine. This could suggest that wines with higher residual sugar potentially exhibit higher densities, which might influence sensory qualities like body and sweetness.

Subsequent variates display moderate to negligible correlations, indicating less pronounced relationships for the other combinations of variables. The second variate suggests an interesting interaction between sulfur dioxide content and alcohol levels, possibly indicating how preservation strategies could be related to alcohol content for balance in winemaking.

This Canonical Correlation Analysis highlights important relationships between the sugar and sulfur dioxide properties of wine and its other chemical characteristics, particularly the strong link between wine sweetness and its physical density. These insights can inform more nuanced approaches to balancing these properties in wine production, enhancing both preservation and sensory qualities.

3.4 Prediction or classification

3.4.1 Logistic regression

As previously stated, wine_type is a categorical variable, which can take the two following values: white or red. In order to predict the type of a certain wine based on its chemical composition and properties, we fit a logistic regression model. It is appropriate since wine_type is a binary variable. Because the result of such a model implicitly depends on a probability score, which will be rounded to the closest between 0 and 1, we coded wine_type as follows: red = 1, white = 0. This conversion was made within the dataset before fitting the model.

We have no previously held knowledge or method that would have permitted us to identify which variables were the most influencial in this situation. Therefore, we first fitted a model that contained all possible predictors. The Student tests performed on each of their associated coefficients indicated that the data showed no evidence that the ones associated to and would not be zero (their p-values were, respectively, 0.1689 and 0.0861, which is higher than 0.05). This implies that these variables do not have a significant impact on the outcome, with 95% confidence ; thus, they could easily be removed from the model. However, these tests are performed individually for each coefficients, hence, we cannot remove both of them at once, since their impact could be changed by the absence of the other. The p-value for the coefficient associated to was higher, indicating a lower significance, therefore, it is the one that we chose to remove first. When the second model was fitted (with all possible predictors but), fixed.acidity was still not significant (for its coefficient, the Student test's p-value was 0.2985>0.05). We then removed it and fitted a new model without nor .

Now, note

$$f(X) = -1645 + 7.10X_2 - 2.83X_3 - 0.871X_4 + 24.3X_5 + 0.0586X_6 - 0.0522X_7 + 1636X_8 + 3.06X_{10} + 1$$

We are trying to predict $Y = X_{13}$, thus, our final model is: $Y = \mathbb{1}_{f(X)>0.5}$.

3.4.2 Multivariate Regression

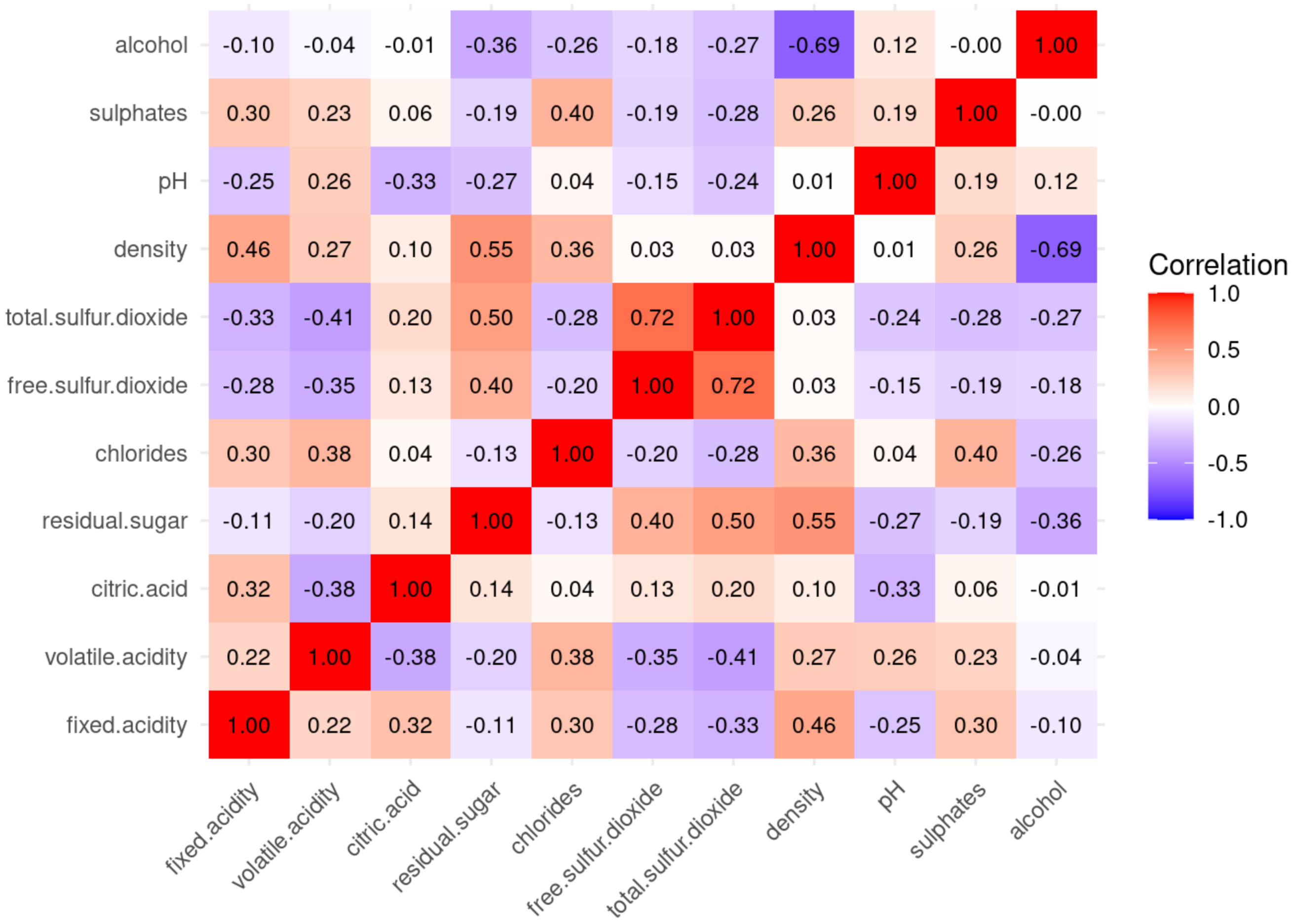
We now want to fit a model to predict a wine's quality score X_{12} based on its chemicals properties and composition. Quality is, in essence a qualitative variable, since it can only take integer values between 0 and 10. However, the range is large enough so we can consider it quantitative. When it comes to prediction, it would even allow for fraction of points, which would be useful in prediction: wine producers would be able to know how far they are from a certain score whereas consumers would know if a wine is in lower or higher range of the said score.

Considering quality as a qualitative variable is also practical for us to chose a model to fit to our data. Indeed, we will try to predict with multivariate regression. Since we want the predictors to be only quantitative as well, we removed upfront, since it is binary. Then, we will use the BOC model selection algorithm in order to remove insignificant predictors and select the ones that would maximize the likelihood function associated to X_{13}. The BIC penalizes large models, to avoid the risk of overparametrization, which is also useful in our case, since we have 13 parameters. This time, we want to predict $Y = X_{12}$ In the end, the chosen model is $Y = 60 + 0.066X_1 - 1.30X_2 + 0.045X_4 + 0.0059X_6 - 0.0025X_7 - 0.59X_8 + 0.48X_9 + 0.74X_{10} + 0.26X_{11}$

After displaying the summary for this model, we confirm that all parameters are indeed significant at level 95% (from the Student tests) and that, overall, it demonstrate strong performance compared to a constant model (small Fisher test p-value). The adjusted R-squared, however, is low (only 0.2906). We tried improving the model by adding or removing parameters, but we were never able to improve this indicator.

3.5 Hypothesis construction and testing

3.5.1 Correlation matrix



Total sulphur dioxide and free sulphur dioxide show a very strong positive correlation of 0.72. Alcohol and density show strong negative correlation of 0.69. Density and residual sugar also show strong correlation.

In our analysis, we explore the impact of two key factors—wine quality and wine type—on our dataset. Wine quality comprises seven distinct populations, while wine type encompasses two populations. Initially, we conduct individual one-way MANOVA analyses for each factor, followed by a comprehensive examination using two-way MANOVA, incorporating both variables simultaneously.

Our investigation employs the entirety of the 11 features available in our dataset as response variables. At each stage of our analysis, we rigorously assess significance levels at 5%, ensuring robustness and reliability in our findings.

3.5.2 MANOVA

One-way MANOVA with wine_type

In our one-way MANOVA analysis, we aimed to discern whether there exists statistically significant evidence indicating variations in all 11 mean responses between the two distinct wine populations: red and white wines. Consequently, our null hypothesis posited that these two wine type populations exhibit no variance.

Our investigation utilized four different tests—Wilks, Hotelling-Lawley, Roy, and Pillai—to evaluate the hypothesis. Remarkably, all four tests yielded an identical p-value of 2.2×10^{-16}. This uniformity in results is unsurprising, given our focus on only two populations in this context.

Therefore, based on our findings, we confidently assert that at least one of the 11 mean responses under consideration differs significantly between the two wine populations examined.

Following our initial findings, we conducted further investigation to identify which of the 11 response variables contributed to the observed differences between the two wine populations. To accomplish this, we performed individual univariate analyses for each response variable, utilizing the Pillai test.

Remarkably, all 11 tests yielded significant results, indicating substantial differences across the various response variables. Notably, while the p-value for alcohol slightly deviated (0.00787), all other tests produced a consistent p-value of 2×10^{-16}.

In light of these outcomes, we confidently conclude that each mean response significantly varies between the two wine populations, underscoring the diverse characteristics and attributes associated with red and white wines.

One-way MANOVA with quality

In our subsequent analysis, we conducted a one-way MANOVA considering all 11 response variables collectively, with wine quality serving as the primary factor. Our null hypothesis posited that the seven wine quality populations exhibit no variation.

Upon evaluation using four different tests—Wilks, Hotelling-Lawley, Roy, and Pillai—we obtained significant results, with a p-value of 2.2×10^{-6} across all tests. This consistent significance underscores the presence of substantial differences among the wine quality populations in terms of the mean responses for at least one of the 11 variables examined.

In light of these findings, we assert that there exists variability in the mean responses across the wine quality populations, highlighting the diverse characteristics and attributes associated with different wine quality levels.

Continuing our analysis, we conducted univariate one-way MANOVA tests to investigate the specific differences detected in the previous step. Out of the 11 response variables examined, all but one yielded significant results. Notably, for the response variable pH, we did not obtain significant evidence to conclude that at least one of the wine quality populations differs from the rest, with a p-value of 0.0593.

This observation suggests that while the majority of response variables exhibit significant variations across the wine quality populations, pH does not demonstrate such differences at a statistically significant level. However, it's essential to interpret this result cautiously, considering its proximity to the conventional significance threshold $(\alpha = 0.05)$. Further exploration or additional analyses may be warranted to fully elucidate the role of pH in distinguishing among wine quality populations.

Two-way MANOVA with quality and wine_type

```
##                             Df  Pillai approx F num Df den Df    Pr(>F)    
## factor(quality)              6 0.45079     47.8     66  38874 < 2.2e-16 ***
## wine_type                    1 0.86261   3695.1     11   6474 < 2.2e-16 ***
## factor(quality):wine_type    5 0.09037     10.8     55  32390 < 2.2e-16 ***
## Residuals                 6484                                             
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In our subsequent analysis, we explored the interaction between two key factors, wine type and wine quality, using a two-way MANOVA approach. Employing the Pillai test, our analysis revealed a significant interaction effect between these factors.

The obtained p-value of 2.2×10^{-16} indicates strong evidence of interactive effects between wine type and wine quality on the mean response of at least one of the 11 variables examined. This suggests that both wine type and wine quality play crucial roles in influencing the characteristics and attributes represented by these variables, and their combined effects contribute to the overall variability observed in the dataset.

In conclusion, our findings underscore the importance of considering both wine type and wine quality concurrently, as they jointly impact the mean responses of the variables under investigation in an interactive manner.

Following the identification of a significant interaction between wine type and wine quality, we proceeded to conduct univariate two-way MANOVA tests to examine whether this interaction effect manifests consistently across all response variables.

Our analyses revealed that for each of the 11 response variables, except for chlorides (p-value 0.106) and alcohol (p-value 0.56262), significant interaction effects were observed. This indicates that the joint influence of wine type and wine quality varies across most response variables, contributing to the overall variability in the dataset.

In the cases of chlorides and alcohol, however, we found no statistically significant evidence to support the presence of interaction effects. Instead, our results suggest that each factor independently affects the mean responses in an additive manner.

In summary, while the interaction between wine type and wine quality significantly impacts most response variables, the effects may differ across individual variables. Nonetheless, the independent effects of each factor remain discernible in certain cases, emphasizing the complexity of their combined influence on the characteristics of the dataset.