# Data Mining Report

## 1 Division of labor in groups

**Table1 Division of labor in groups**

| CWID | Name | Contribution | Percent Contribution |
|---|---|---|---|
| A20563419 | Jianing Song | Data processing, create visual charts, and write reports | 50% |
| A20563469 | Laiyuhao Yu | Collect data, create visual charts, and write reports | 50% |

## 2 Dataset description

This data set focuses on the Portuguese "Vinho Verde" wine and revolves around the relationship between the content of various chemicals in the wine and the wine quality, containing 1143 records and 13 fields. It provides rich data support for studying the influencing factors of wine quality and for the construction of predictive models. The properties of the wine provided by the data set are as follows:

**Table2 Wine properties**

| Attribute | Data type | Attribute | Data type |
|---|---|---|---|
| fixed acidity | double | chlorides | double |
| volatile acidity | double | free sulfur dioxide | double |
| citric acid | double | total sulfur dioxide | double |
| residual sugar | double | density | double |
| pH | double | sulphates | double |
| alcohol | double | quality | int |
| Id | int | | |

## 3 Data processing

Upon inspection, we found no duplicate values in the data. There were some missing data, so we used the mean filling method to fill them in. The attribute with a strong correlation with red wine quality, fixed acidity, was selected to draw the boxplot (see part 4 for the boxchart), and some outliers were found. Because the sample size was too small, we chose to correct the outliers manually.

# 4 Data visualization analysis

## 4.1 Drawing method

When mapping the graphs, we used the following libraries:

**Table3 Library used for drawing**

| Related library | function |
|---|---|
| pandas | Data reading and preprocessing |
| numpy | Performed the data preprocessing |
| matplotlib | For the drawing of the parallel coordinate map and the radar map |
| seaborn | For the drawing of the heat map |
| plotly.express | For the drawing of the histograms |
| warnings | For scatter plots and pairwise plots |

We use the flow chart to explain the drawing method and analysis ideas:

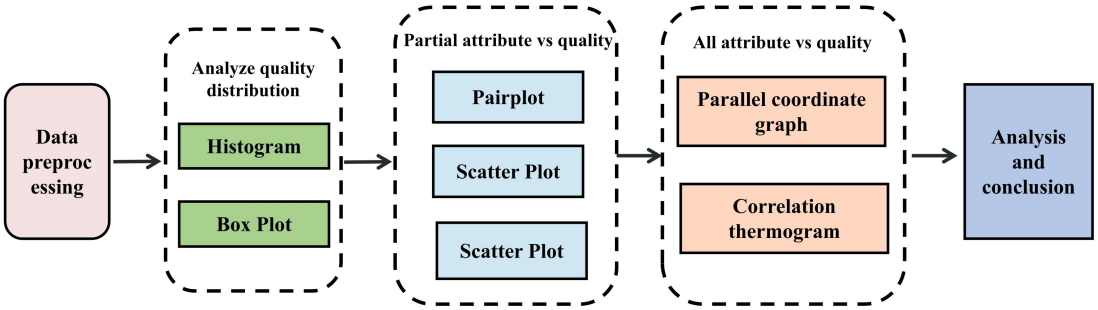


**Figure 1 Drawing method flow chart**

### 4.1 Histogram (Using Plotly)

This is a histogram showing the wine quality distribution in the dataset. Each column in the histogram represents a specific wine quality score (3 to 8 points), and the height of each column represents the number of wines with that score (i.e., frequency of occurrence). Different colors in the figure indicate different quality scores. Such charts help to visually analyze the distribution of wine quality, allowing us to observe which quality score of wine is the most common and which score is scarcer.

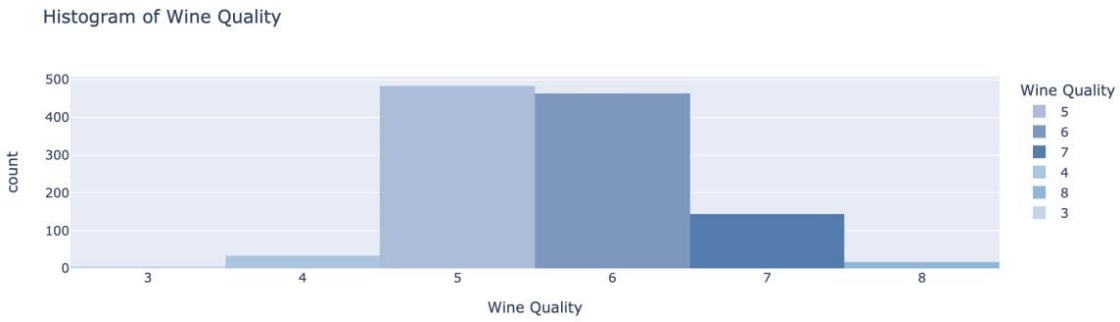


**Figure 2 Histogram**

From the histogram, the wine column bars with quality scores 5 and 6 were significantly higher, showing that they occupied a large proportion in the dataset, indicating a majority of moderate - quality wines. There were fewer wines with quality scores 7 and 4 compared to wines with quality scores of 5 and 6. Wines with quality scores 8 and 3 were the least in number, which had a very small proportion in the data, showing the scarcity of wines at these two extreme scores.

The quality scores of the dataset presented an unbalanced pattern, with medium - quality wines dominating.

### 4.2 Boxplot (Using Seaborn)

This plot is a boxplot showing the distribution of wines with different quality scores on fixed acidity. Each quality score (from 3 to 8 points) corresponds to a boxplot showing the distribution of fixed acidity under this quality score. The median line of the box represents the median (the 50th percentile of the data), and the upper and lower edges of the box represent the upper (75th percentile) and the lower quartile(25th percentile) of the data, respectively. The “whiskers” of the boxplot represent the maximum and minimum of the data, and points beyond the whiskers range indicate outliers.

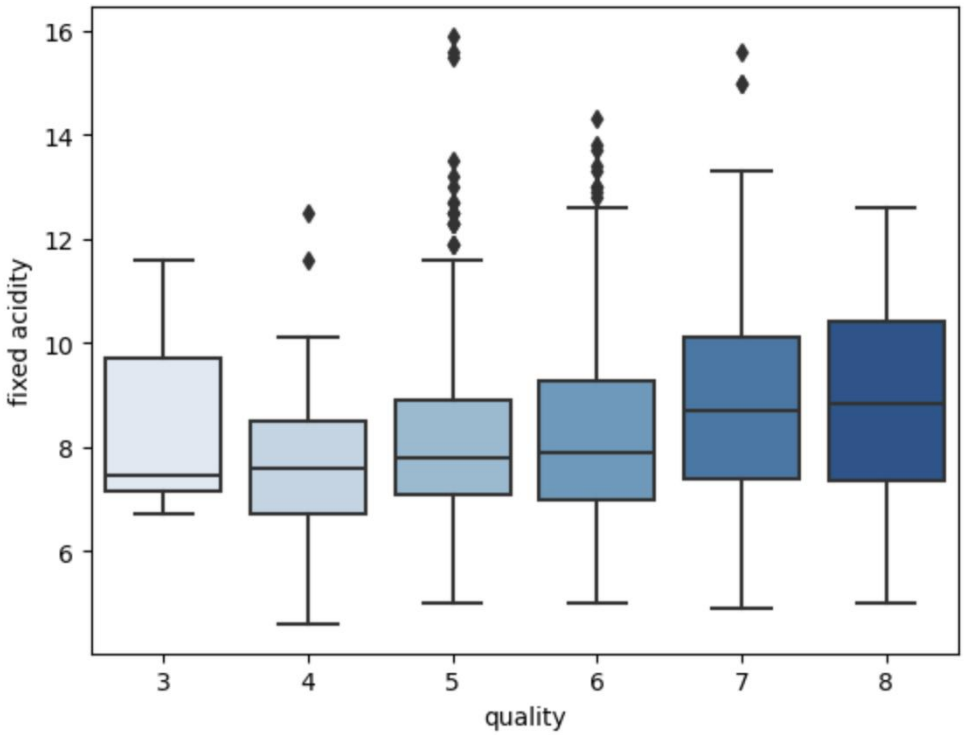


**Figure 3 Boxplot**

Quality scores 3 to 4: The fixed acidity distribution was concentrated and the median was within the lower acidity range (between 7 and 8). This may mean that lower quality wines have lower fixed acidity.

Quality score 5 to 7: With the improvement of quality score, the median of fixed acidity increased slightly, but the wine distribution of these scores is relatively broad and the box is longer, indicating that their fixed acidity has a large range of changes.

Quality score 8: Wine with a quality score of 8 showed a high median (about 8 to 9) at fixed acidity and had a small fluctuation range. The relatively compact distribution indicates that these high-quality wines show less variation in fixed acidity and tend to have higher acidity levels.

Furthermore, a certain number of outliers exist in each quality score group, often located outside the whiskers of the boxplot, indicating extreme values of fixed acidity.

### 4.3 Pairplot (Using Seaborn)

This scatter plot matrix shows the relationship between the three variables, specifically: fixed acidity, alcohol, and quality. Furthermore, each small graph in the scatter plot matrix demonstrates the correlation between these variables. The color of

each dot represents a different quality score.

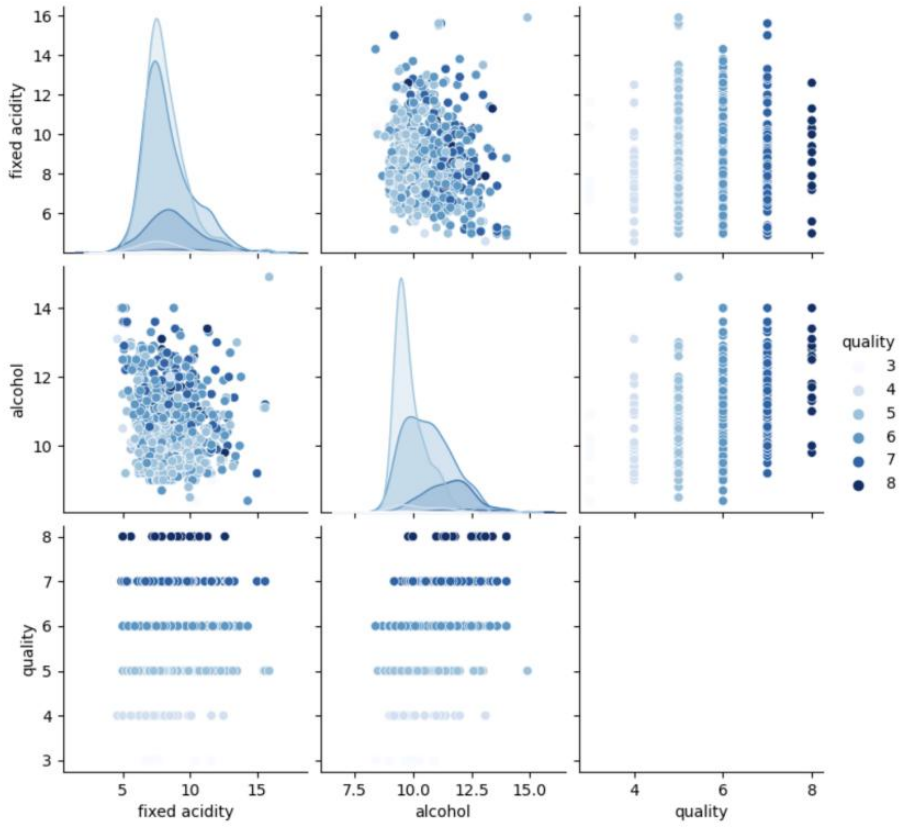


**Figure 4 Pairplot**

Quality vs. fixed acidity: Lower-quality wines (e.g.,3 points, 4 points) have a lower fixed acidity, and most lower-quality wines have a fixed acidity between 5 and 8.

Higher -quality wines (such as 7 and 8 points) have a wide distribution in fixed acidity. Higher - quality wines are not limited to a certain acidity range, showing a certain diversity.

Quality vs. alcohol: In the relationship between quality score and alcohol content, higher-quality wines tend to be associated with higher alcohol content. Lower-quality wines generally have lower alcohol content, showing a possible positive correlation between alcohol content and wine quality.

Distribution of the scatter-plot matrix: From the whole matrix, wines with lower quality scores are mostly concentrated in the range of lower acidity and alcohol content. Wines with higher quality scores are distributed in wider areas, especially in areas with higher alcohol content, indicating that wines with higher alcohol content

have a larger proportion of higher-quality wines.

### 4.4 Scatter Plot (Using Seaborn)

This plot is a scatter plot showing the relationship between the fixed acidity and alcohol of the wine. Each dot represents a wine, and the color represents the quality, from 3 (light) to 8 (dark).

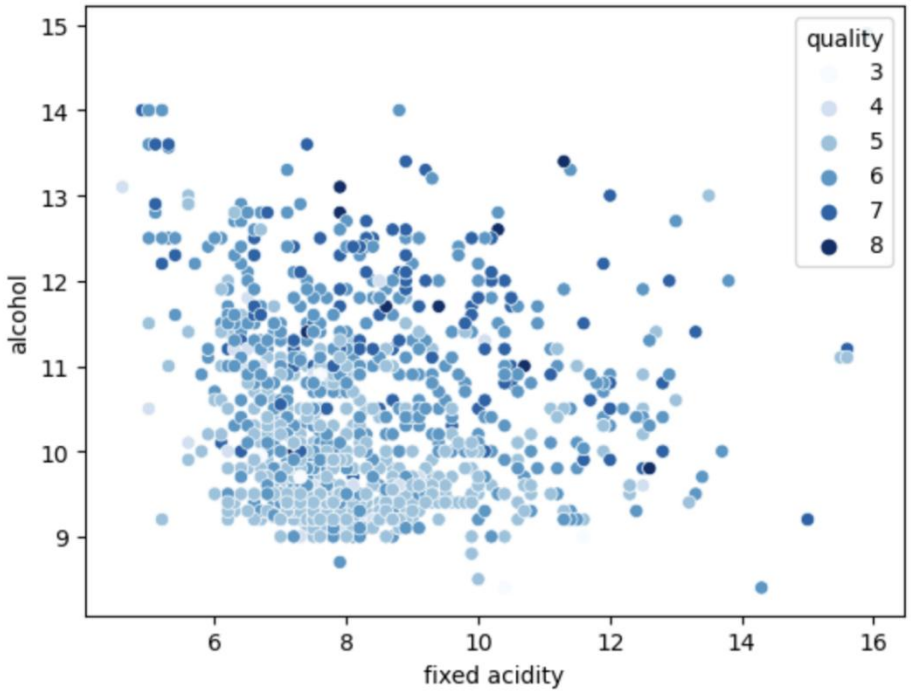


**Figure 5 Scatter Plot**

Quality score vs. acidity and alcohol content: The wines with higher quality scores tend to have darker colors in the figure, mainly in areas with higher alcohol content and moderate acidity. Wine with lower quality scores were mostly concentrated in areas with lower alcohol content and higher fixed acidity.

### 4.5 Parallel coordinate graph

In order to analyze the differences of different attributes of wine with different qualities, we drew the parallel coordinate diagram of various attributes and quality of wine, and the lines of different colors represent the wine samples with different quality scores. Through the changes of the lines on each attribute axis, the relationship between attributes and quality can be analyzed. The parallel coordinate diagram is

shown below:

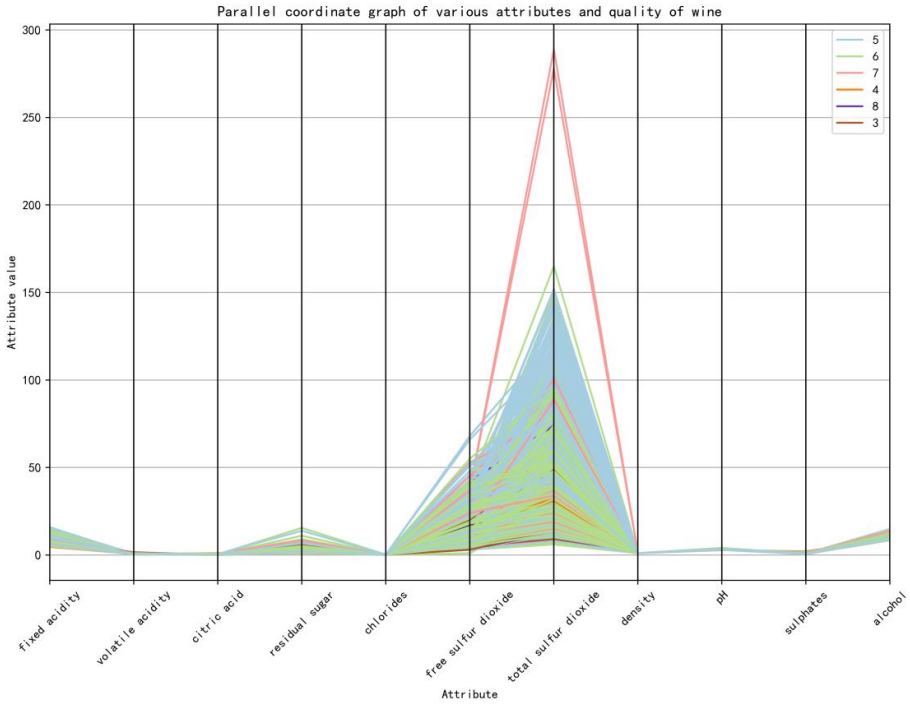


**Figure 6 Parallel coordinate graph**

Based on the free sulfur dioxide and total sulfur dioxide axes, the values of wines with different qualities on the citric acid, volatile acidity and other axes are relatively concentrated.

From the alcohol-axis, wines with higher quality scores have more lines with relatively high alcohol content, indicating that higher alcohol content may be associated with better wine quality. On the volatile acidity-axis, wines with higher quality scores generally have lower values for the corresponding lines, indicating that lower volatile acidity may be more conducive to improving wine quality.

The larger number of lines of light blue (quality score 5) and green (quality score 6) indicates that there are large numbers of wine samples with scores of 5 and 6 in the data set; while fewer lines of purple (8) and brown (3) indicate fewer wine samples of high quality (8) and low quality (3), consistent with the imbalance of the data set category.

### 4.6 Sample radar image

Through the analysis of the parallel coordinate map, we found that some data

with relatively large change ranges but small bases were difficult to distinguish differences in the parallel coordinate map. So we chose fixed acidity, pH, volatile acidity, residual sugar and alcohol, the five attributes, normalized the data and randomly selected a sample radar map, which can intuitively show the contrast between the samples. And the radar map is as follows:

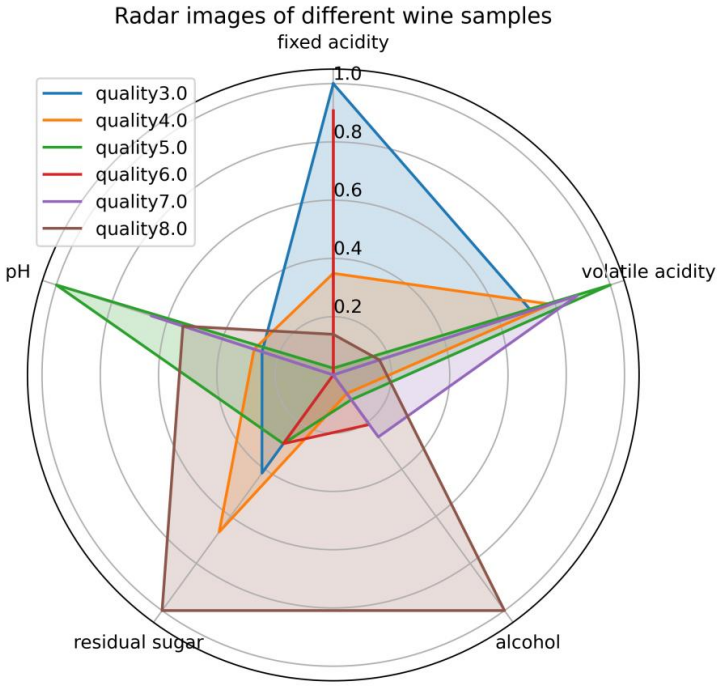


**Figure 7 radar image**

As can be seen from the figure, for the fixed acidity attribute, the samples with quality score 3.0 have higher values, while with the increase of quality score, the fixed acidity value of some samples tends to decrease, showing a negative correlation.

For volatile acidity, overall, the values of the samples in volatile acidity properties generally decrease with higher quality scores, indicating that lower volatile acidity may be negatively correlated with higher wine quality.

Samples with higher quality scores have relatively high values on alcohol attributes, showing that higher alcohol content may have a positive effect on wine quality.

For residual sugar, its numerical distribution is relatively scattered, and there is no obvious correlation with the quality score.

For pH values, samples with higher quality scores have relatively low values on

the pH value attribute, suggesting that lower pH values may be somewhat associated with better wine quality.

### 4.7 Correlation heatmap

To further analyze the influence of different attributes on wine quality, We use Pearson correlation coefficient to calculate the correlation between these attributes and drew the correlation heat map as follows:

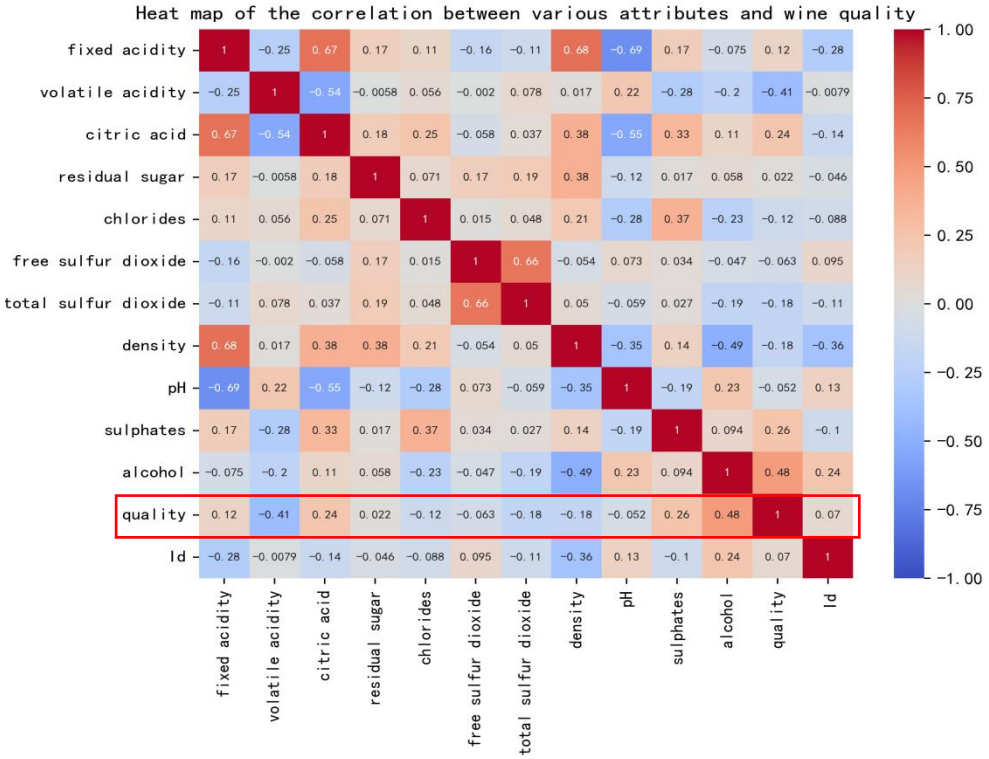


**Figure 8 Correlation heatmap**

According to this thermal map analysis, alcohol, sulphates, citric acid and fixed acidity were positively correlated with wine quality. Among these, alcohol had the highest correlation coefficient of 0.48, indicating that the higher the alcohol content, the better the wine quality may be.

Volatile acidity, total sulfur dioxide, density, chlorides, free sulfur dioxide, and pH were negatively correlated with wine quality. Specifically, the correlation coefficient of volatile acidity was-0.407394, with a large absolute value, indicating that the higher the volatile acidity, the lower the wine quality.

In conclusion, the absolute values of the correlation coefficients of citric acid,

sulphates, alcohol, and volatile acidity are all above 0.2, and we believe that the above four attributes have a great impact on the wine quality.

## 5 Conclusion

By analyzing the data set, we found that several key factors influence the quality of the wine:

Alcohol has a strong positive correlation with wine quality. Higher quality wines generally have higher alcohol content, suggesting that alcohol content may be an important factor in improving wine quality.

Volatile Acidity has a significant negative correlation with wine quality. Lower volatile acidity is often associated with higher wine quality, indicating that excessive acidity may affect the quality of the wine.

Fixed Acidity The impact on wine quality is more complex. The large numerical difference in fixed acidity in different quality wines suggests that this property may have different effects.

Sulfates and Citric Acid are also positively correlated with wine quality, suggesting that these two may play a role in improving wine quality.

We also observed that wines with quality scores of 5 and 6 dominated the dataset, while wines with quality scores of 8 and 3 were scarce. Based on the data, we can infer that the quality of wine can be improved by appropriately increasing the content of Alcohol, Sulfates and Citric Acid and reducing the content of Volatile Acidity.