Final Report

Analysis White Wine Quality Dataset

Group 2
Palaka Venkata Gangadhar Naveen, Dixit Ashwin Arun, Liu Rayna, Li Zhuoya

Introduction and Background

Wine is playing an increasingly important role in food culture. Both red and white wines are favored by the public. The primary reason is wine enhances the flavors of the food, especially when you have a wonderful marriage of food and high-quality wine. For example, the crispness in white wines brings out the light, delicate flavors of fish, pork, and chicken *(Ron Saikowski, 2016)*. Since, selecting different quality wine to pair with food will make or break the flavor combination, distinguishing and understanding the quality of wine is useful. Also, distinguishing and understanding the quality of wine is important to winemaking, wine collectors, and sellers, who are keener to get high-quality wines.

Wine quality, as Maynard Amerine, a pioneering researcher in the cultivation, fermentation, and sensory evaluation of wine once said, is easier to detect than define *(Jackson, 2017)*. Since defining the wine quality is a matter of perception and it's subjective, detect wine quality in terms of its chemistry will be more well-founded.

Objectives and Goals

In this project, we will do research on different characteristics in the white wine such as alcohol, chlorides, citric acid, density, etc. to gain a better understanding on how each contributes to the quality. Our goal is to classify the quality of wine by detecting the content of chemical characteristics in wine for providing customers and sellers with white wine quality prediction and reference before tasting it.

Questions

Before we can classify the quality of the white wine, we need to have a general understanding of the distribution of the quality of the white wine, and need to know which chemical characteristics are important, and which are not so important. We decided to start our data exploration by looking at the distribution of the white wine quality.

Question 1: What is the distribution of white wine quality?

Question 2: What chemical characteristics have a significantly influence on the quality of the white wine?

Question 3: What is the range of contents of each important chemical characteristics at different quality?

Question 4: What is the good combination of five important chemical characteristics producing a good quality white wine?

Datasets

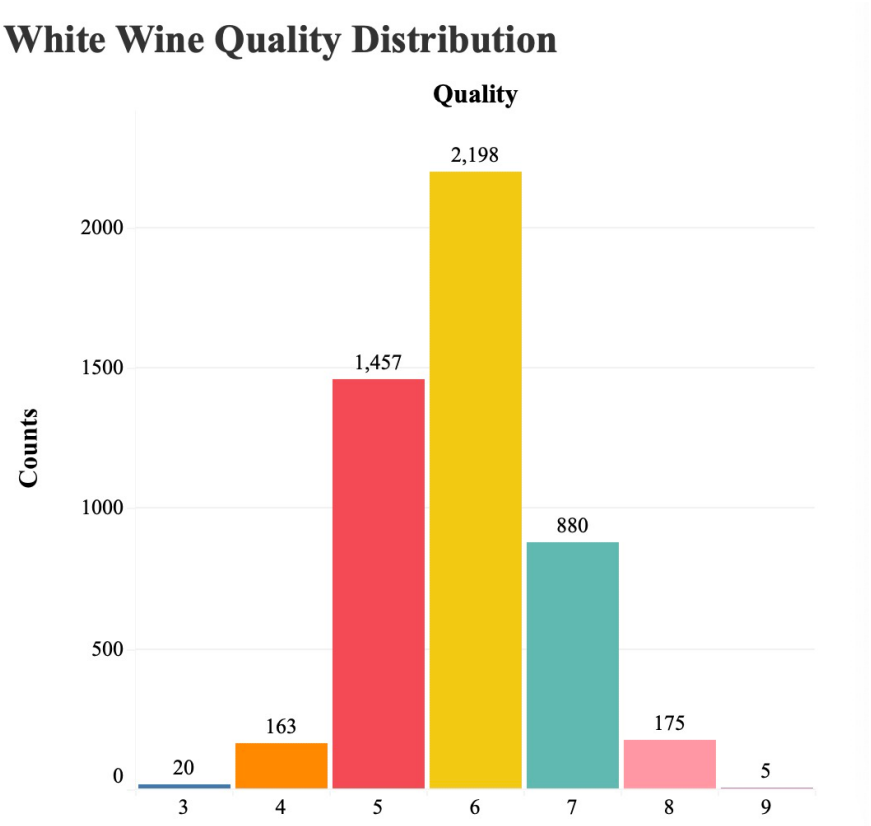
This dataset relates the white vinho verde wine. Vinho verde exclusively produced in the demarcated region of Vinho Verde in northwestern Portugal, it is only produced from the indigenous grape varieties of the region, preserving its typicity of aromas and flavors as unique in the world of wine (Cvrvv, 2021). This dataset contains 4,898 observations, 11 input variables based on physicochemical tests: alcohol, chlorides, citric acid, free sulfur dioxide, pH, residual sugar, total sulfur dioxide, volatile acidity, fixed acidity, density, sulphate, and 1 output variable based on sensory data: quality. This is a clean dataset, there are no null / missing value, and can be used directly for exploratory visual analysis.

The dataset is found from UCI Machine Learning Repository with link Wine Quality-white.

Visualization Components

Question 1: What is the distribution of white wine quality?

Figure 1:



As seen in the histogram, the classes of quality are ordered and not balanced among 4,898 samples of white wine. Most of the white wines' quality was between 5 and 7, there were few exceptionally bad wines (below quality 4) and excellent wines (above quality 8), with no low scores of 1 and 2 and no perfect score of 10. This reflects that most white wines' quality levels were in the mid-range. Next, we are going to explore what chemical characteristics affect the quality of the white wine.

Question 2: What chemical characteristics have a significantly influence on the quality of the white wine?

Figure 2:

Correlation Matrix

Showing correlation coefficient between two chemical characteristics of the wine

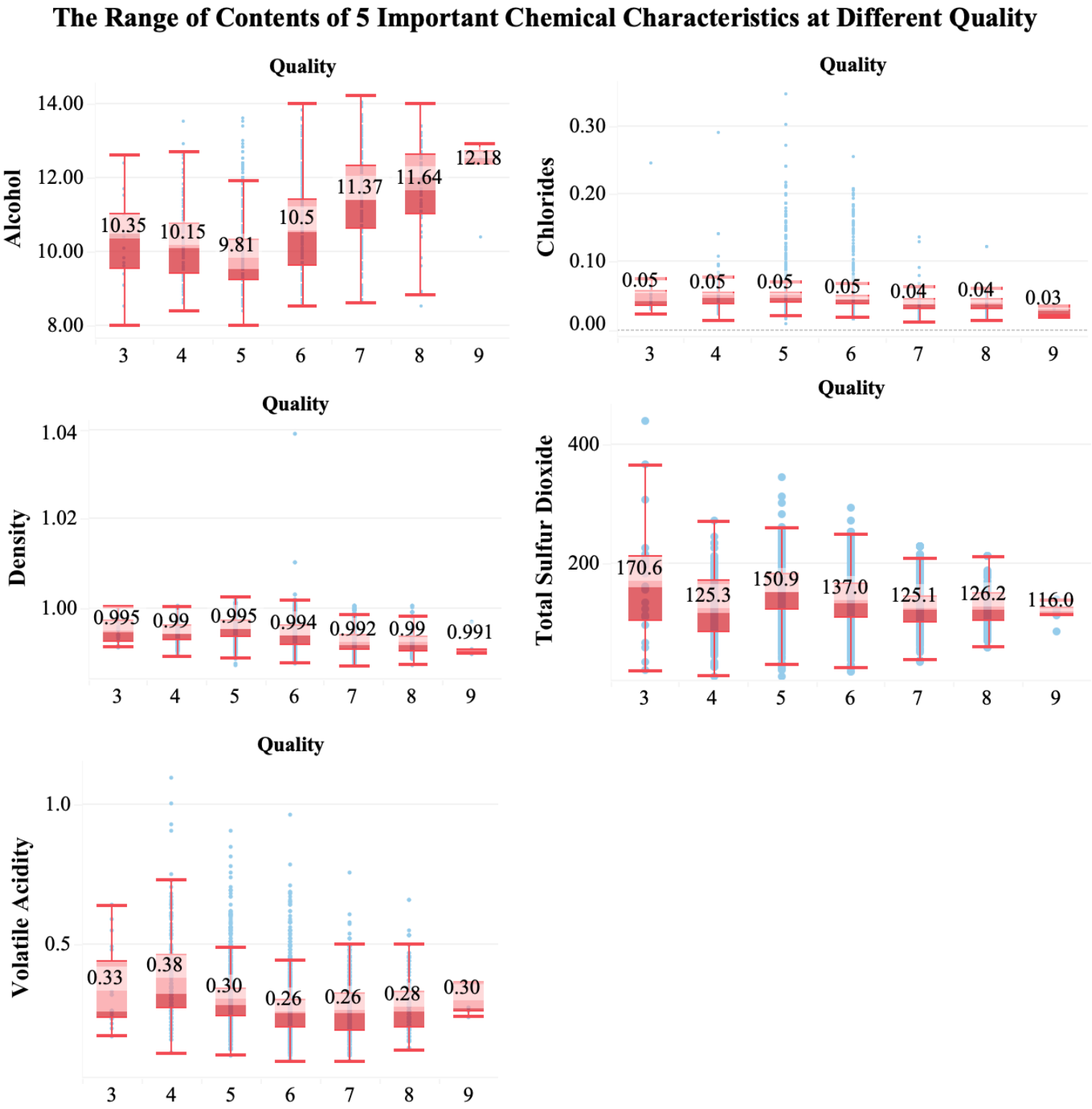
Chemical Characteries

Chemical Character..	alcohol	chlorides	citric acid	density	fixed acidity	free sulfur dioxide	pH	quality	residual sugar	sulphates	total sulfur dioxide	volatile acidity
alcohol		-0.36	-0.08	-0.78	-0.12	-0.25	0.12	0.44	-0.45	-0.02	-0.45	0.07
chlorides	-0.36		0.11	0.26	0.02	0.10	-0.09	-0.21	0.09	0.02	0.20	0.07
citric acid	-0.08	0.11		0.15	0.29	0.09	-0.16	-0.01	0.09	0.06	0.12	-0.15
density	-0.78	0.26	0.15		0.27	0.29	-0.09	-0.31	0.84	0.07	0.53	0.03
fixed acidity	-0.12	0.02	0.29	0.27		-0.05	-0.43	-0.11	0.09	-0.02	0.09	-0.02
free sulfur dioxide	-0.25	0.10	0.09	0.29	-0.05		0.00	0.01	0.30	0.06	0.62	-0.10
pH	0.12	-0.09	-0.16	-0.09	-0.43	0.00		0.10	-0.19	0.16	0.00	-0.03
quality	0.44	-0.21	-0.01	-0.31	-0.11	0.01	0.10		-0.10	0.05	-0.17	-0.19
residual sugar	-0.45	0.09	0.09	0.84	0.09	0.30	-0.19	-0.10		-0.03	0.40	0.06
sulphates	-0.02	0.02	0.06	0.07	-0.02	0.06	0.16	0.05	-0.03		0.13	-0.04
total sulfur dioxide	-0.45	0.20	0.12	0.53	0.09	0.62	0.00	-0.17	0.40	0.13		0.09
volatile acidity	0.07	0.07	-0.15	0.03	-0.02	-0.10	-0.03	-0.19	0.06	-0.04	0.09	

In the correlation matrix of all variables, it shows us the correlation coefficients between white wine quality and each characteristic. The strength of relationship between two variables represented by color, which is stronger with darker color and weaker with lighter color. By looking at the result, there are five chemical characteristics showing stronger relationship with quality, that are alcohol, chlorides, density, total sulfur dioxide, and volatile acidity. Therefore, we could see that these five chemical characteristics have a significantly influence on the quality of the white wine. Next, we take a close to look at the range of contents of these five important chemical characteristics.

Question 3: What is the range of contents of each important chemical characteristics at different quality?

Figure 3:

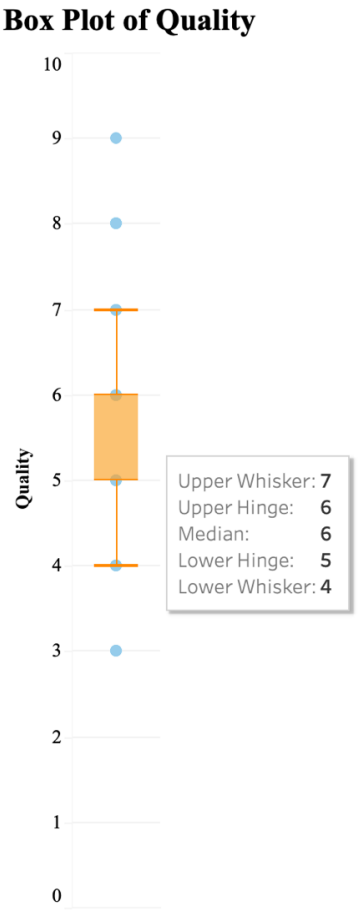


The boxplots were constructed to compare the variability of five important characteristics across different quality ratings. Although the number of lower quality wines (rating 3 or 4) is very less as shown in Figure 1 bar chart, it showed the higher variability among five boxplots, while higher quality wines (rating 8 or 9) showed the lower variability with few outliers in some plots. For example, in above figure, it showed that the higher quality wines (rating 8 or 9) had the lower variability of total sulfur dioxide whereas the lower quality wines (rating 3 or 4) had the higher variability of total sulfur dioxide. Therefore, it suggests that high-quality wines tended to have a consistent range of total sulfur dioxide, as well as other four chemical characteristics.

Next, we want to define the quality of the white wine in three groups (Bad, Fair, and Good), explore the differences in average contents of five important chemical characteristics in three groups, then find the good combination of five important chemical characteristics producing a good quality white wine.

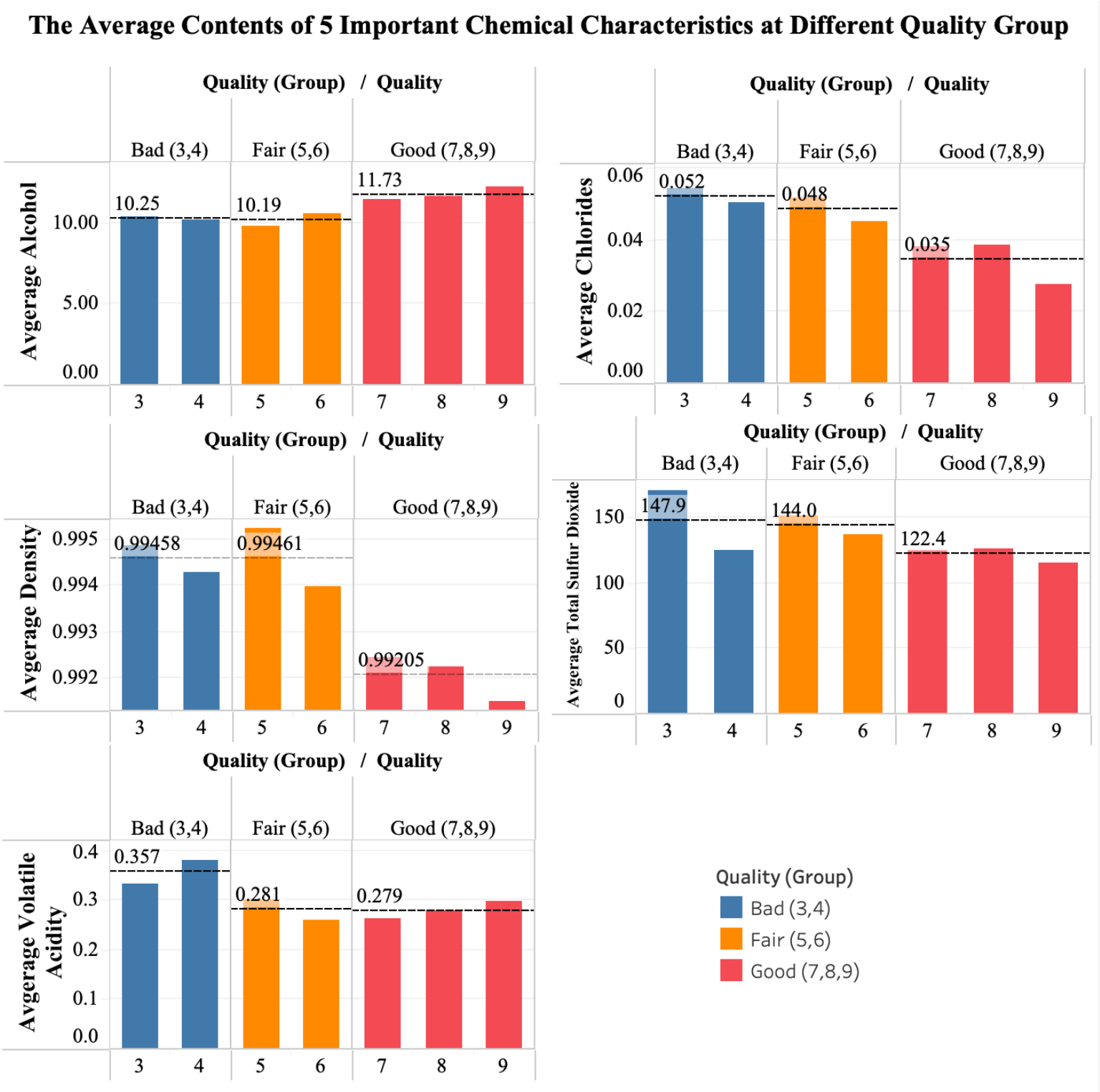
Question 4: Define the quality of the white wine in three groups (Bad, Fair, and Good). What are the average contents of each important chemical characteristics at each group?

Figure 4:



Firstly, we defined the quality of the white wine in three groups: bad, fair, and good based on quality distribution. From the boxplot shown above, we define the quality rating lower than lower hinge as “Bad” (3,4), the quality rating between lower hinge and upper hinge as “Fair” (5,6), the quality rating higher than upper hinge as “Good” (7,8,9).
Next, we compared the differences in the average contents of important chemical characteristics between the group.

Figure 5:


The Average Contents of 5 Important Chemical Characteristics at Different Quality Group
Quality (Group) / Quality
Bad (3,4)
Fair (5,6)
Good (7,8,9)
Avgerage Alcohol
10.25
10.19
11.73
10.00
5.00
0.00
3
4
5
6
7
8
9
Average Chlorides
0.052
0.048
0.035
0.06
0.04
0.02
0.00
Avgerage Density
0.99458
0.99461
0.99205
0.995
0.994
0.993
0.992
Avgerage Total Sulfur Dioxide
147.9
144.0
122.4
150
100
50
0
Avgerage Volatile Acidity
0.357
0.281
0.279
0.4
0.3
0.2
0.1
0.0
Quality (Group)
Bad (3,4)
Fair (5,6)
Good (7,8,9)

In the above, it showed us average contents that are specific by five important chemical characteristics across the different groups, which made the change between group more clearly by using bar graphs and line charts together. For the good wine (red color) had high average alcohol (about 11.73 units), low average chlorides (about 0.035 units), low average density (about 0.99205 units), low average total sulfur dioxide (about 122.4 units), and low average volatile acidity (about 0.279 units) relatively.

It is known that high volatile acidity may bring an irritating smell of wine, the proper amount of alcohol content helps to balance the sweetness and acidity of the wine and high total sulfur dioxide could affect the taste of wine. Therefore, the quality of wine was increased as the volatile acidity was lowered with more fruity-smelling than those with a sharper smell, alcohol content is greater, total sulfur dioxide is lower with a little bit mineral-smell. In other words, we can see that good-quality wine has its own preference relatively for these five critical chemical characteristics like three contents mentioned above.

Conclusions and Suggestions

To summarize some key findings, we created bar chart, correlation matrix, and box plot in tableau. In summary, most wines were rated at the mid-range quality (5-7), there were few poor or exceptional quality wines. The alcohol, density, chlorides, total sulfur dioxide, and volatile acidity are known as five important chemical characteristics to have a significantly influence on the quality of white wine. Then we looked further at the range of these five important chemical characteristics, we found that they were less variability in higher-quality wines. We then tried to find the average contents of these five important chemical characteristics in good-quality white wine. This numerical analysis can provide collectors, buyers, and sellers of white wine with a prediction and reference of good white wine quality prior to tasting.

The TWBX Tableau File with dashboard makes it easy for white wine collectors, buyers, and sellers to explore and analysis white wine information by updating latest data.

References

UCI. Wine Quality Data Set.
[Dataset in UCI Machine Learning Repository]. http://archive.ics.uci.edu/ml/datasets/Wine+Quality
[Dataset for white wine]. http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

Paulo Cortez, University of Minho, Guimarães, Portugal,
http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal

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.

Jackson, R. S. (2017). Nature and Origins of Wine Quality. Wine Quality - an overview | ScienceDirect Topics. Retrieved October 31, 2021, from https://www.sciencedirect.com/topics/food-science/wine-quality.

Ron Saikowski / Houston Wine Walk. (2016, September 27). Does food taste better with wine? Chron. Retrieved October 31, 2021, from https://www.chron.com/neighborhood/article/Does-food-taste-better-with-wine-9303954.php.

Cvrvv. (n.d.). *About vinho verde*. Vinho Verde. Retrieved November 8, 2021, from https://www.vinhoverde.pt/en/about-vinho-verde