Introduction to Data Science (DSGA-1001) Capstone Project

Group Name: Sommeliers

Michael Healy (mbh425)
Seonhye Yang (sy3420)
Bhavna Thakar (bt2227**)**

Introduction

The purpose of this project was to use data to answer questions in a meaningful and insightful way. We decided to use data that captured attributes of wine and the data used in this analysis comes from the paper *Modeling wine preferences by data mining from physicochemical properties* (Cortez et.al). Because we are analyzing both red and white wines, two datasets are being used both separately. Each dataset consists of several rows corresponding to individual wines, and twelve columns corresponding to attributes of each wine. The wines captured in the rows are the red and white variants of the Portuguese wine *vine verde*. The structure and purpose of each column is as follows:

1. Fixed Acidity: The amount of acids in wine; most acids found in wine are nonvolatile and thus, do not evaporate easily
2. Volatile Acidity: The amount of acetic acid in wine, at high levels can result in a vinegar taste
3. Citric Acid: Adds a fresh flavor to the wine and is found in small quantities
4. Residual Sugar: The amount of sugar remaining after fermentation—generally the range for this metric is 1 gram/liter to 45 grams/liter. Above 45 grams/liter a wine is considered sweet.
5. Chlorides: The amount of salt found in the wine
6. Free Sulfur Dioxide: the free form of sulfur dioxide prevents microbial growth and wine oxidation; it exists in equilibrium between molecular sulfur dioxide (dissolved gas) and bisulfite ion
7. Total Sulfur Dioxide: The amount of free and bound forms of sulfur dioxide; generally it is found at low concentrations and remains undetectable but at concentrations over 50 ppm, there is a sulfurous smell and taste to the wine
8. Density: The density of wine relative to that of water; the two densities are similar depending of the percent alcohol and sugar content of the wine
9. pH: How acidic or basic a wine is on a scale from 0 (acidic) to 14 (basic); most wines fall in the range of 3-4
10. Sulphates: When added to wine, sulphates contribute to sulfur dioxide levels and have antimicrobial and antioxidant properties
11. Alcohol: The percent alcohol of the wine
12. Quality: A qualitative score given to each wine that falls between 0 and 10 based on subjective sensory data

Often data will have missing or redundant values, and, in those cases, it needs to be cleaned. While this data did not need to be cleaned of null values or repeated values, when answering questions that are not red- or white-specific, the two datasets had to be combined manually.

Question 1

The question we want to answer first is do white wine variants of this wine differ significantly from red wine variants. This question is important in dictating the rest of the experiment as it reveals whether the wine datasets can/should be considered separately or combined. All of the variables except pH are not normally distributed (**Appendix A**). As a note, because quality is a rating, we do not consider a normal distribution for it.

To answer this question thoroughly, we used hypothesis testing to first determine whether there is a difference between white and red wine quality distributions, and if their attributes were significantly different. To do this, we used an alpha value of 0.001 and used a Kolmogorov-Smirnov test and Mann-Whitney U tests, respectively.

For distribution comparison, the KS test had a p-value of 1.0 which indicates that we cannot reject the null hypothesis—that the shapes are the same. This result was surprising as there is a visible difference between the cumulative distributions, but perhaps the KS test caught that the overall shape is nearly the same, ignoring scaling, and reported back such a p-value. The histograms of the white and red wine quality data and quality distributions are shown **Figure 1** and **2**.

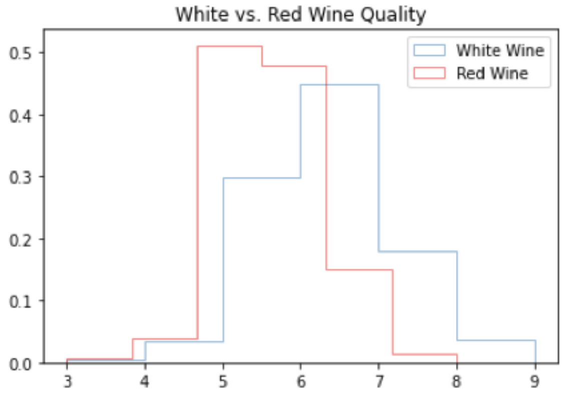


Figure 1

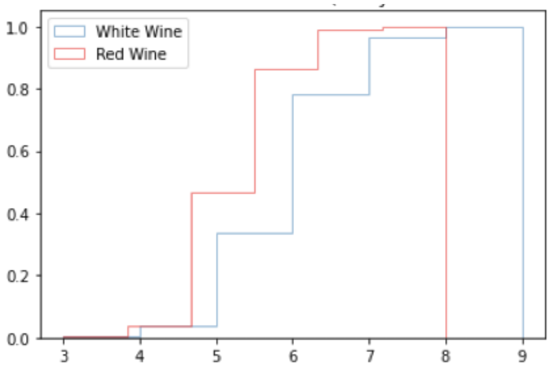


Figure 2

Table 1

Attribute	Significant Difference or Fail to Reject Null	p-value reported from Mann-Whitney U
Fixed Acidity	Difference	1.44e-255
Volatile Acidity	Difference	0.0
Citric Acid	Difference	1.312e-38
Residual Sugar	Difference	5.633e-95
Chlorides	Difference	0.0
Free Sulfur Dioxide	Difference	0.0
Total Sulfur Dioxide	Difference	0.0
Density	Difference	1.503e-237
pH	Difference	5.471e-162
Sulphates	Difference	0.0
Alcohol	Fail	0.182
Quality	Difference	3.634e-23

The results for the attribute differences are shown in **Table 1.** Despite the KS test not showing a significant difference between the two wine types, we felt the need to verify further if the red and white wine datasets were significantly different or not. To do this, we ran Mann-Whitney U tests to compare each attribute between the wine types: each attribute had a p-value output that allowed us to determine whether each attribute has a significant difference on wine quality, or whether we fail to reject the null hypothesis that they are similar (**Table 1)**. P-values below our alpha value of 0.001 indicate that we can reject the null hypothesis and the corresponding attributes have a significant difference on wine quality. All attributes except alcohol were reported as significantly different. Thus, we claim that the white and red wine variants are significantly different from each other, requiring them to be kept separate for future testing.

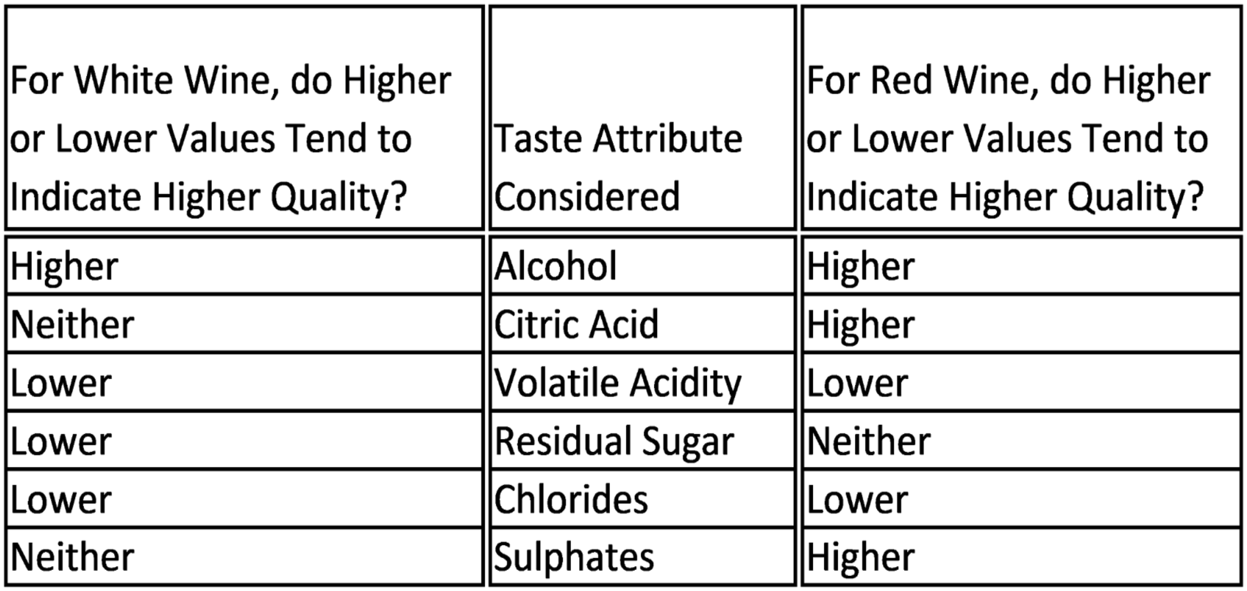
Question 2

The second question we aim to answer using the datasets is whether taste-affecting attributes have a significant difference in quality for higher vs. lower values. The taste attributes we are considering are alcohol percentage, citric acid, volatile acidity, residual sugar, chlorides, and sulphates. Again, we decided to use hypothesis testing to answer this question.

When answering our first question, we concluded that there was a significant difference between white and red wine qualities, but not between the distributions of these qualities. Despite there being no significant difference between distributions, we did the tests separately for each wine type, rather than combining the two datasets, due to the results from Question 1.

To determine whether wines with a higher alcohol percentage had higher quality ratings than those with lower alcohol percentages, we first did a median-split based on alcohol content. We used the median because it is a more robust metric of central tendency. After splitting the data into high alcohol and low alcohol groups, we did a Mann-Whitney U test on their respective quality scores to determine whether there was a significant difference. A Mann-Whitney U test was done due to the nonparametric nature of the quality data. The alpha value was set to 0.001 for all tests. This was repeated for each attribute in the data (alcohol percentage, citric acid, volatile acidity, residual sugar, chlorides, and sulphates), and done separately for both white and red wines. If the p-value was lower than the alpha value of 0.001, we were able to reject the null hypothesis and conclude there the attribute significantly affected taste. If the p-value was greater than the alpha value, we failed to reject the null hypothesis, and could not determine if the attribute impacted taste significantly. Many of the attributes indicate quality differently in red and white wine, which implies that separating the two sets of data, rather than combining them, was a good way to accurately answer our question. The results are reported in **Table 2.**

Table 2



For White Wine, do Higher or Lower Values Tend to Indicate Higher Quality?	Taste Attribute Considered	For Red Wine, do Higher or Lower Values Tend to Indicate Higher Quality?
Higher	Alcohol	Higher
Neither	Citric Acid	Higher
Lower	Volatile Acidity	Lower
Lower	Residual Sugar	Neither
Lower	Chlorides	Lower
Neither	Sulphates	Higher

Question 3

For the third question, we would like to see if we can predict wine type given the attributes and quality score. To do this, we decided to use logistic regression analysis. It made more sense to do a logistic regression as opposed to a machine learning model because we are already given a rather small number of attributes, so doing a PCA to reduce dimensions would result in new dimensions that would be less insightful and unnecessary given the comparatively small size of our data.

To do a logistic regression analysis, the first step was to split the data into training and testing data using an 80% to 20% split. This was done separately for white and red wine to make sure that we have an equal proportion of representation, so one type of wine is not overshadowing the other. Then, we combined the training data and testing data, respectively, after making a new column to track which type of wine is which. A logistic model was built to classify white and red wine—white was classified as 0 and red as 1. The model was then tested using the training data. An accuracy score was also computed by determining the number of correct predictions out of the number of total predictions. Our model's accuracy score came out to be 0.979 on the test data.

Table 3

Wine Attribute	Indicative of this wine type:
fixed acidity	red
volatile acidity	red
citric acid	white
residual sugar	white
chlorides	red
free sulfur dioxide	red
total sulfur dioxide	white
density	red
pH	red
sulphates	red
alcohol	white
quality	red

The results of our logistic regression model are shown in **Table 3**. For each wine attribute, our model's coefficients show whether it is indicative of one type or the other: if a coefficient is negative, then it pushes the result to 0 (white), and to 1 (red) for positive coefficients. Intuitively, these results make sense. For example, for attributes like citric acid and residual sugar, it makes sense that they are more indicative of white wine, as generally white wines are fresher and sweeter in flavor. For attributes like density and total sulfur dioxide, it makes sense that they are more indicative of red wines. Generally red wines are heavier than whites and contain tannins which serve as a stabilizing agent. Because of this, less sulfur dioxide is needed in red wine production to protect it from microbial contamination and oxidation.

Conclusion

To summarize, the three questions we set out to answer about our data were as follows:

- Are wine types (red and white) significantly different?
- Do taste-affecting attributes have a significant difference in quality for higher vs. lower values?
- Can we predict the wine type given attributes and quality score?

For the first question, we used hypothesis testing to determine whether there were significant differences in attributes between red and white wine types, and between red and white wine quality distributions. For quality score distribution, we were unable to determine a significant difference between wine types. However, we determined a significant difference between most attributes for red and white data, indicating that for all attributes other than alcohol percentage, there was a significant difference between wine types (**Table 1**).

For the second question, we used hypothesis testing to establish whether taste-affecting attributes determine significant differences in quality between higher vs. lower values. The attributes tested were alcohol percentage, citric acid, volatile acidity, residual sugar, chlorides, and sulphates. It was determined that these attributes affected red and white wines differently (**Table 2**).

For the third question, we used a regression analysis. We created a logistic regression model that determined which wine type each attribute is indicative of with an accuracy score of 0.979. The model used binary classification to determine, for each attribute, whether the results were closer to 0 (white) or 1 (red). We then reported which attributes were assigned to which wine type by our model (**Table 3**).

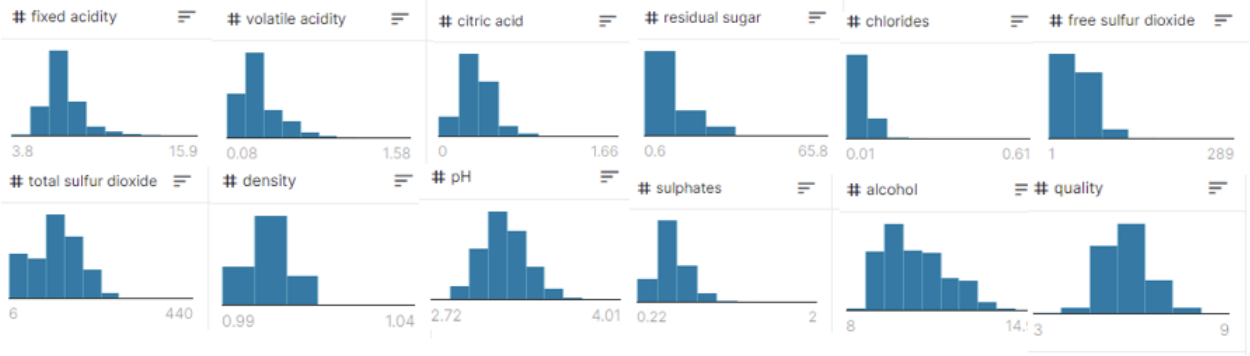
Overall, we were able to draw some conclusions about our data using our results from each of the three questions. It is evident that red and white wine have significantly different properties and attributes. Our data does have some limitations, however. For one, the wines listed in our datasets are of the Portuguese *vino verde* variety, so we are unable to make conclusions that capture red wine and white wine on a wider scale. Instead, we can only make inferences about red and white varieties of *vino verde.* Our results may be different for other varieties of red and white wines. Also, our analysis was limited by the number of samples we had (rows) and the number of attributes that were tested (columns). A richer dataset would result in more accurate results. In addition, we made assumptions about the shape of the distributions, deciding that most attributes were not normally distributed and for this reason, we used nonparametric significance tests for all cases in our hypothesis testing sections (Questions 1 and 2).

Citations

Cortez, Paulo, et al. “Modeling Wine Preferences by Data Mining from Physicochemical Properties.” *Decision Support Systems*, North-Holland, 9 June 2009, https://www.sciencedirect.com/science/article/pii/S0167923609001377.

Appendix A

The histograms for each attribute are shown below, for showing the distributions of their data for red and white wine combined


# fixed acidity
# volatile acidity
# citric acid
# residual sugar
# chlorides
# free sulfur dioxide
# total sulfur dioxide
# density
# pH
# sulphates
# alcohol
# quality