Team 2: Liam Bachelor, Sainofo Fanene, Conor Tradewell

Madhumitha Sivakumaran

DATA 115

29 April 2025

Final Project: Wine Quality Report

**Summary**

The data was gathered by researchers from the University of Minho in Portugal. In the data collection, physicochemical tests were performed for the red and white wine--"Vinho Verde" wine samples. Each sample was sensory rated by wine tasters on quality scale from 1 to 10. The objective lab measurements (acidity, sugar, pH) were paired with subjective quality scores. The goal was to explore how physicochemical attributes relate to perceived wine quality. Cases have happened in predicting the wine quality, understanding which features influence wine taste the most, and benchmark classification and regression models. The dataset was used in several research papers to develop predictive models for wine quality. In the wine quality dataset programming, there are two datasets that are loaded--red wine and white wine. We have combined the two datasets into one data frame by implementing the column, naming it as 'type' and creating a variable named 'wine' which leads to using pd.concat(). We checked if there were any missing values, and it has resulted in no missing values. There were duplicates which ended up dropping. The outliers were checked but not removed. All data types were appropriate as the code was good to execute.

**Features**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Description | Data Type | Typical Range |
| *‘fixed\_acidity`* | Tartaric acid (g/dm³) | float | 4.6-15.9 |
| *`volatile\_acidity`* | Acetic acid (g/dm³) | float | 0.12-1.58 |
| *`citric\_acid`* | Citric acid (g/dm³) | float | 0-1.66 |
| *`residual\_sugar`* | Remaining sugars after fermentation (g/dm³) | float | 0.6-65.8 |
| *`chlorides`* | Salt concentration (g/dm³) | float | 0.009-0.346 |
| *`free\_sulfur\_dioxide`* | Free SO₂ (mg/dm³) | float | 1-289 |
| *`total\_sulfur\_dioxide`* | Total SO₂ (mg/dm³) | float | 6-440 |
| *`density`* | Density of the wine (g/cm³) | float | 0.987-1.039 |
| *`pH`* | Acidity level (0–14 scale) | float | 2.72-4.01 |
| *`sulphates`* | Potassium sulphate content (g/dm³) | float | 0.22-2.0 |
| *`alcohol`* | Alcohol content (% vol) | float | 8.0-14.9 |
| *`quality`* | Quality score (sensory rating from 0 to 10) | int | 3-9 |
| *`type`* | Wine type label (*`'red'`* or *`'white'`*) | object | Red/white |

**Graphical Analysis**

The graphical analysis for this dataset was done in two parts. The first was the creation of a scatterplot matrix for the data. This was able to reveal multiple relationships between the variables in the dataset. One important finding from this was that there were many relationships between variables that were stronger than any relationships with those variables and quality. This includes the positive relationship between residual sugar and density, and the negative relationship between fixed acidity and pH. This demonstrates the difficulty of analyzing these variables independently but does provide the benefit of increasing our awareness of how the relationships between our variables might influence the outcome of our findings. The scatterplot matrix additionally gave indications of which variables influence quality, with the representation of each individual data point showing how much variation there is at each rating and how outliers may be affecting the results. The second part of the graphical analysis was the creation of bar graphs representing the average orientation of each variable at each quality rating. This was able to more easily represent general correlations between each variable and quality ratings. This found a positive quality correlation between citric acid and alcohol, and a negative quality correlation between volatile acidity and chlorides. This was also able to show if a variable made more of a difference at the higher or lower range of ratings, such as average volatile acidity decreasing between the quality ratings of 3 and 6 before remaining steady for the other ratings, or average alcohol concentration only increasing between the quality ratings of 6 and 9.

While the graphical analysis was useful for identifying certain trends and relationships in the data, there are certainly limitations to the methods used. One of these is the lack of quantification, which can obscure the degree of the impact of the relationships indicated. This is mostly made up for in regression analysis. Another limitation is the lack of cross-comparison between the variables impacting quality. While the scatterplot matrix is useful in identifying relationships between these variables, none of the graphical methods used explored how these multi-variable relationships impacted the perceived quality of the wine in the dataset. Methods that we could have used to analyze the impact of multi-variable relationships on quality include combined scatterplots and heatmaps. While these may have revealed new insights about how combinations of variables impact quality, they likely would have had the effect of obscuring the already weak correlations found in our analysis.

**Regression Analysis**

Performing linear regression with quality as the response variable and every other factors as the predictors resulted in a multiple linear regression model where every factor other than citric acid had a significant relationship with quality. However, a large condition number (2.85e+05) suggested a high amount of multicollineary, meaning that our predictor variables may be highly correlated with one another. This caused many factors to be represented as having a significant relationship with quality, when in reality that is not likely to be the case. The very low R-squared value indicated that a linear relationship is not appropriate for this dataset in addition, generally pointing to linear regression being a poor fit for this dataset. In order to try and mitigate the effects of multicollinearity, we performed lasso and ridge regression. These models both had results that differed greatly from the linear regression model, indicating that multicollinearity did have a large effect on our dataset. Within the ridge regression model, volatile acidity, residual sugar, alcohol and density seemed to have the greatest effect on quality, while in the lasso regression model, alcohol and volatile acidity appeared to be the only factors with a significant effect on quality.

Because of the high amount of multicollinearity, regression is limited in its effectiveness, and a linear model appears to be a rather poor representation of the quality, judging by the low R-squared values of each model (0.311 for linear, 0.279 for lasso, 0.315 for ridge). Though different types of regression may help with these problems, in reality it is very possible that there are other factors that are not included within the dataset that are larger influences on the quality of the wine than any of these chemical properties.

**Conclusions**

Based on the results of our analysis, it appears that initial regression results indicate that all factors other than citric acid influence the quality of the wine but in actuality there is a high degree of multicollinearity, meaning that the predictive factors highly influence one another, leading to many factors being represented in regression that should not be. Additionally, a low R-squared value indicates that a linear model is not appropriate for this data. Running lasso and ridge regression to help combat multicollinearity gives a better indication of what factors might be significant predictors of wine quality but still result in a low R-squared. Thus, while we cannot accurately model wine quality based on our factors, the results indicate that the most significant predictors of wine quality are alcohol content and volatile acidity. As the alcohol content increases, so does the perceived quality of the wine, and as the volatile acidity increases, the perceived quality of the wine decreases. Other factors such as residual sugar, sulphates and free sulfur dioxide potentially have a less significant positive impact on quality, while density may have a less significant negative impact on quality. The significance of these factors is not easy to determine, as ridge and lasso regression give differing results, but it is worth noting that they are much less significant than the two primary factors, and their significance may be due to relation to our primary significant factors.

**Future Research**

The results of this research are not particularly useful for several reasons, not the least of which is the problems encountered in regression. Considering that the quality of wine is a somewhat nebulous concept, and can differ from person to person, the way in which it was measured may be less than satisfactory. A more robust system of measurement, averaging a greater number of quality scores given by many different people, might make for a better variable to predict. In addition, there are other factors to consider that may have a greater influence on the quality of the wine other than just the chemical properties, such as age or how it was stored, or factors concerning the grapes used in the wine’s production. Considering more factors and using a more robust system of quality measurement may result in a better model for the quality of wine.