Wine Analysis

Yasmine Abdel-Rahman

Red Wine Quality Analysis

Yasmine Abdel-Rahman, Frankie Stillo, Theo Prosise

Introduction and Data

Wine is a commodity integral to the United States. It contributes over 276 billion dollars to the United States gross domestic product (GDP) accounting for nearly 1.3% of the GDP in 2022[1]. The wine industry employs over 1 million US citizens generating nearly 40.1 billion dollars in wages annually[2]. It is a staple of not just the US economy but internationally as well responsible for 400 billion dollars in revenue each year which includes the millions of tourists that travel seeking quality wine[3]. Hence, illuminating what makes a wine quality and reproducible has been subject to many studies. Previous studies have investigated the influence of grape variety, time, terroir, winemaking techniques, and aging and fermentation conditions on the quality and rating of wine, however, less investigated is the wine's chemical makeup at the time of consumption. This study aims to investigate the chemical compositions, qualities and other physicochemical measures of red wine that contribute to a high-rating of a wine connoisseur. The data[4] used in this analysis was obtained solely from Vinho Verde wines which are wines created from Northwest Portugal. The chemical data was obtained prior to the certification of wine. Certification is a process that ensures the wine is safe for human consumption and determines the quality of the wine for sale and consumption. The dataset contains 12 numerical variables and 1,599 observations each representing a bottle of wine. The variables contained within the dataset and their descriptions are visible in Figure 1. The dataset is ordered and not balanced meaning that there are more average rated wines than high and low quality wines. This project aims to produce a model that can accurately predict wine quality based on chemical properties. According to Dr. Jamie Goode, a wine researcher and journalist,

""[In winemaking] acidity is most crucial. It provides structure, balance, and freshness, ensuring that a wine doesn't taste flabby or overly sweet. Acidity is like the backbone of a wine; without it, a wine lacks definition and vitality.[5]"

An aspect of this dataset worth investigating further is the correlation between pH levels and wine ratings. Developing analytical models to discern the relationship between pH and wine scores, alongside examining the chemical composition's impact on pH, holds promise for enhancing comprehension of wine quality through a chemical perspective.

Our research question is: Which factors and properties can accurately predict the quality of wine?

Figure 1

Table 1: Chemical descriptions and units[6]

variable	mean	sd	units	descriptions
fixed.acidity	8.320	1.741	g(tartaric acid)/dm^3	wine's natural acids
volatile.acidity	0.528	0.179	g(acetic acid)/dm^3	measure of the wine's gaseous acids that contributes to the smell and taste of vinegar in wine
citric.acid	0.271	0.195	g/dm^3	Boosts the acidity of wine during fermentation
residual.sugar	2.539	1.410	g/dm^3	natural grape sugars left in a wine after the alcoholic fermentation finishes.
chlorides	0.087	0.047	g(sodium chloride)/dm^3	adds to the saltiness of a wine
free.sulfur.dioxide	15.875	10.460	mg/dm^3	helps protect the wine from oxidation and spoilage
total.sulfur.dioxide	46.468	32.895	mg/dm^3	portion of SO_2 that is free in the wine plus the portion that is bound to other chemicals in the wine
density	0.997	0.002	g/cm^3	helps determine the alcohol content level of the final wine
pH	3.311	0.154	NA	can affect aroma, flavor, carbon dioxide absorption, tartrate precipitation, color, age-ability, fermentation rate, stability, and malolactic fermentation
sulphates	0.658	0.170	g(potassium sulphate)/dm^3	food preservative used to maintain the flavor and freshness of wine
alcohol	10.423	1.066	vol.%	Alcohol Content
quality	5.636	0.808	NA	Score given by experts

Methodology

RStudio Analysis

RStudio was used for the analytical portion of this study and collaboration was facilitated through the use of GitHub.

The following R packages were utilized for data manipulation, statistical analysis, and presentation:

- **tidymodels**: A collection of packages for modeling using the tidyverse principles.
- **tidyverse**: A collection of packages for data manipulation and visualization.
- **dplyr**: A package for data manipulation and transformation.
- **knitr**: A package for dynamic report generation in R Markdown

Regression Model Specification

Linear regression models were constructed to predict wine quality based on various physicochemical factors. The response variable, quality, was treated as continuous, aligning with the ordinal nature of the quality scores assigned by experts. Predictor variables considered included fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content.

An interactive method was chosen as each predictor has confounding effects on items such as alcohol %, density, and pH. An interactive model would account for these differences by examining the relationship between predictors. In the case of our data, our variables affect the outcome of the red wine.

A trial and error method was used to determine the chosen models. For the top predictors model, we chose the variables based off a research paper[6] published with the data set. Via experimentation with the variables, models with high-predictor statistics were selected for further analysis.

Model Fitting

Three of our top models were fitted to the data for analysis:

1. Volatile Acidity and Alcohol Content Model (vol-aciditiy-alc%-model):
    - This model fits a linear regression model predicting wine quality based on volatile acidity and alcohol content.
2. Top Predictors Model (mix-model):
    - This model does an interactive linear regression predicting quality from sulphates, pH, total sulfur dioxide, and alcohol.
3. All Predictors Model (all-model):