Wine Analysis

Report

AUTHOR
team_e

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

variable	mean	sd	units	descriptions
fixed.acidity	8.320	1.741	g(tartaric acid)/dm³	wine's natural acids
volatile.acidity	0.528	0.179	g(acetic acid)/dm³	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/dm³	Boosts the acidity of wine during fermentation
residual.sugar	2.539	1.410	g/dm³	natural grape sugars left in a wine after the alcoholic fermentation finishes.
chlorides	0.087	0.047	g(sodium chloride)/dm³	adds to the saltiness of a wine
free.sulfur.dioxide	15.875	10.460	mg/dm³	helps protect the wine from oxidation and spoilage
total.sulfur.dioxide	46.468	32.895	mg/dm³	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/cm³	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)/dm³	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

Chemical descriptions and units[6]

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):
    - This model fits a linear regression model using all available chemical predictors, including volatile acidity, fixed acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, and sulfates.Note that this does not include density, alcohol%, or pH.

To do a comparative statistical analysis, adjusted R^2 and R^2 were used. These predictor statistics were chosen as they quantify the proportion of variance in the response variable that is explained by the predictor variables in the regression model which is exactly what this study aims to uncover. The tidy() and glance() functions were used to provide statistical summaries and model evaluation metrics, facilitating the interpretation of the regression results.

Results:

Our top model based on adjusted R^2 was our second model, the mix model. This model's adjusted R^2 value was 0.399, which was the highest of all the models we tested. This makes sense, as it parallels what was said in the research paper of the data set. What we have found helps answer our own research question: Which factors and properties can accurately predict the quality of wine?

The interactive nature of the model helps predict the quality of the wine as it takes into account how each variable interacts with the other. They do not exist in isolation. This interaction can be modeled by the equation.

Equation:

$$\begin{aligned}\hat{y} = &-696.7092 + 882.2127 \times \text{sulphates} + 211.4495 \times \text{pH} + 6.9559 \times \text{total sulfur} \\ &+ 68.9782 \times \text{alcohol} + 1445.558 \times \text{volatile acidity} \\ &+ 6533.450 \times \text{chlorides} - 268.4295 \times (\text{sulphates} \times \text{pH}) \\ &- 86.2912 \times (\text{sulphates} \times \text{alcohol}) - 20.7689 \times (\text{pH} \times \text{alcohol}) - \dots\end{aligned}$$

Due to the complex nature of the interactive model, the entirety of the equation is too long to be put. Instead, the beginning of the equation is put to show the interactions between the beginning predictors. The remainder of the equation can be simply obtained using information taken from Figure 3.

Model Outcomes

The model outcomes show our top three models, including the one with the highest adjusted R^2.

Figure 2: acidity-alc% Fit: Linear regression model predicting wine quality based on volatile acidity and alcohol content

term	estimate	std.error	statistic	p.value
(Intercept)	2.5848	0.5373	4.8108	0.0000
volatile.acidity	-0.3671	1.0089	-0.3638	0.7160
alcohol	0.3621	0.0503	7.1984	0.0000
volatile.acidity:alcohol	-0.0968	0.0957	-1.0121	0.3116

R Squared
0.317

Adjusted R Squared
0.316

Volatile Acidity and Alcohol Content Model

Figure 3: Mix Fit: This model does an additive linear regression predicting quality from sulphates, pH, total sulfur dioxide, and alcohol.

term	estimate	std.error	statistic	p.value
(Intercept)	-696.709	245.439	-2.839	0.005
sulphates	882.213	334.321	2.639	0.008
pH	211.449	74.617	2.834	0.005
total.sulfur.dioxide	6.956	3.805	1.828	0.068
alcohol	68.978	23.803	2.898	0.004
volatile.acidity	1445.558	463.911	3.116	0.002
chlorides	6533.450	3095.463	2.111	0.035
sulphates:pH	-268.429	102.069	-2.630	0.009
sulphates:total.sulfur.dioxide	-9.735	5.111	-1.905	0.057
pH:total.sulfur.dioxide	-2.092	1.167	-1.793	0.073
sulphates:alcohol	-86.291	33.028	-2.613	0.009
pH:alcohol	-20.769	7.217	-2.878	0.004
total.sulfur.dioxide:alcohol	-0.689	0.368	-1.874	0.061
sulphates:volatile.acidity	-1763.054	678.805	-2.597	0.009
pH:volatile.acidity	-432.801	139.839	-3.095	0.002
total.sulfur.dioxide:volatile.acidity	-18.085	7.079	-2.555	0.011
alcohol:volatile.acidity	-141.547	45.420	-3.116	0.002
sulphates:chlorides	-8005.032	4051.036	-1.976	0.048
pH:chlorides	-1956.867	939.812	-2.082	0.037
total.sulfur.dioxide:chlorides	-61.765	56.934	-1.085	0.278
alcohol:chlorides	-660.822	304.855	-2.168	0.030
volatile.acidity:chlorides	-14695.624	5573.193	-2.637	0.008
sulphates:pH:total.sulfur.dioxide	2.960	1.570	1.886	0.060
sulphates:pH:alcohol	26.310	10.050	2.618	0.009
sulphates:total.sulfur.dioxide:alcohol	0.970	0.507	1.912	0.056
pH:total.sulfur.dioxide:alcohol	0.207	0.113	1.840	0.066
sulphates:pH:volatile.acidity	532.597	206.276	2.582	0.010
sulphates:total.sulfur.dioxide:volatile.acidity	22.929	10.016	2.289	0.022
pH:total.sulfur.dioxide:volatile.acidity	5.361	2.143	2.501	0.012
sulphates:alcohol:volatile.acidity	171.466	67.305	2.548	0.011
pH:alcohol:volatile.acidity	42.277	13.644	3.098	0.002
total.sulfur.dioxide:alcohol:volatile.acidity	1.806	0.687	2.629	0.009
sulphates:pH:chlorides	2415.863	1238.107	1.951	0.051
sulphates:total.sulfur.dioxide:chlorides	88.882	73.331	1.212	0.226

term	estimate	std.error	statistic	p.value
pH:total.sulfur.dioxide:chlorides	18.304	17.518	1.045	0.296
sulphates:alcohol:chlorides	819.797	406.220	2.018	0.044
pH:alcohol:chlorides	197.790	92.339	2.142	0.032
total.sulfur.dioxide:alcohol:chlorides	6.194	5.637	1.099	0.272
sulphates:volatile.acidity:chlorides	17692.226	7892.411	2.242	0.025
pH:volatile.acidity:chlorides	4368.875	1674.438	2.609	0.009
total.sulfur.dioxide:volatile.acidity:chlorides	175.945	96.608	1.821	0.069
alcohol:volatile.acidity:chlorides	1495.867	550.176	2.719	0.007
sulphates:pH:total.sulfur.dioxide:alcohol	-0.295	0.155	-1.901	0.058
sulphates:pH:total.sulfur.dioxide:volatile.acidity	-6.911	3.045	-2.270	0.023
sulphates:pH:alcohol:volatile.acidity	-51.746	20.376	-2.540	0.011
sulphates:total.sulfur.dioxide:alcohol:volatile.acidity	-2.280	0.987	-2.311	0.021
pH:total.sulfur.dioxide:alcohol:volatile.acidity	-0.534	0.207	-2.576	0.010
sulphates:pH:total.sulfur.dioxide:chlorides	-26.678	22.756	-1.172	0.241
sulphates:pH:alcohol:chlorides	-247.355	123.803	-1.998	0.046
sulphates:total.sulfur.dioxide:alcohol:chlorides	-9.159	7.381	-1.241	0.215
pH:total.sulfur.dioxide:alcohol:chlorides	-1.833	1.734	-1.057	0.291
sulphates:pH:volatile.acidity:chlorides	-5315.298	2400.513	-2.214	0.027
sulphates:total.sulfur.dioxide:volatile.acidity:chlorides	-224.301	128.090	-1.751	0.080
pH:total.sulfur.dioxide:volatile.acidity:chlorides	-51.552	29.388	-1.754	0.080
sulphates:alcohol:volatile.acidity:chlorides	-1815.935	788.373	-2.303	0.021
pH:alcohol:volatile.acidity:chlorides	-444.120	164.757	-2.696	0.007
total.sulfur.dioxide:alcohol:volatile.acidity:chlorides	-17.997	9.508	-1.893	0.059
sulphates:pH:total.sulfur.dioxide:alcohol:volatile.acidity	0.686	0.299	2.298	0.022
sulphates:pH:total.sulfur.dioxide:alcohol:chlorides	2.747	2.288	1.201	0.230
sulphates:pH:total.sulfur.dioxide:volatile.acidity:chlorides	66.917	39.431	1.697	0.090
sulphates:pH:alcohol:volatile.acidity:chlorides	544.987	238.945	2.281	0.023
sulphates:total.sulfur.dioxide:alcohol:volatile.acidity:chlorides	23.190	12.699	1.826	0.068
pH:total.sulfur.dioxide:alcohol:volatile.acidity:chlorides	5.262	2.887	1.823	0.069
sulphates:pH:total.sulfur.dioxide:alcohol:volatile.acidity:chlorides	-6.907	3.902	-1.770	0.077

R Squared
0.423

Adjusted R Squared
0.399

Mix Model,

Figure 4: All-Predictors Fit: This model does an additive linear regression using every chemical variable:

R Squared
0.479

Adjusted R Squared
0.38

Volatile Acidity and Alcohol Content Model

Discussion

In conclusion, looking at the variables of red wine and how those can predict its quality led us to explore different combinations of predictors and their outcomes. Using linear regression models and their adjusted R^2 values, we were able to come up with our best model that predicts the quality of red wine.

In addition, our analysis does have some limitations. Although we tried to use as many different combinations of predictive variables that we could, there were only so many that we could use. In addition, we couldn't fit the interactive model's output for the last model due to its length. This points to a limitation in our work: the factors that led to which predictors we would use had to do with the logistics of being able to include it in our write-up and presentation, and what value it brought to out analysis.

One way our analysis could be improved is by comparing additive and interactive models. In this write-up, we only included interactive models as the variables in the data set interacted with one another to determine wine quality. However, comparing additive to interactive linear regression models could provide key insights on how the variables interact with each other that is missing in this analysis.

Using this analysis, future work can build off of the red wine data set. We also had access to a white wine data set, and looking at the interactions between those two would have opened up a whole new world of possibilities in terms of comparing predictive values to determine wine quality. Unfortunately, the same limitation came up that we did not have the space or time to complete them, but future work may include comparing white wine to red wine, and determing which may be easier to produce at a high quality.

References

1. Economic impact study of the American wine industry. WineAmerica. (2022, October 27). https://wineamerica.org/economic-impact-study

2. Dunham, J. (2022, September 21). *2022 economic impact study of the American wine industry methodology*. Wine-America. https://wineamerica.org/economic-impact-study/2022-american-wine-industry-methodology

3. Published Jan Conway, & 3, A. (2024, April 3). *Global: Wine market size 2018-2028*. Statista. https://www.statista.com/statistics/922403/global-wine-market-size/

4. Cortez, Paulo, Cerdeira, A., Almeida, F., Matos,T., and Reis,J.. (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

5. Goode, J. (2021). *The science of wine: From vine to glass*. University of California Press.

6. Cortez, P. et al. “Modeling wine preferences by data mining from physicochemical properties.” Decis. Support Syst. 47 (2009): 547-553.