Regression analysis of wine quality based on its physicochemical properties*

based on white Vinho Verde wine from Portugal

Ziyao Han

27 April 2022

Abstract

Wine is one of the most widely consumed beverages in the world with significant cultural influence. In this paper, the physicochemical properties of wine will be compared with the perceived quality of wine by wine experts. Even though the preference for wine taste is highly subjective, this study may help identify objective components for wine preference. We found that higher alcohol percentage, higher sulphate, and less acidic wines yield higher quality scores, while sugar percentage does not appear to affect scores significantly.

Keywords: Multiple Linear Regression, Model Selection, Wine Quality, White Wine, Vinho Verde, Physicochemical Properties

1 Introduction

Wine is one of the most widely consumed beverages in the world with significant cultural influence. Wine is made through an arduous process of selecting fruit, processing, fermentation, and ageing with many small intricacies that impact the final product. The sugar and acid levels of the fruit and ageing vessels are just some of the ways that the taste of the final product can be changed (Denig 2020). Wine has historically been consumed as a part of cultural and religious ceremonies throughout the world and is a part of the cultural identities of many countries, especially in Europe (*History of Wine - Wine History and Origins* 2022).

In this report, the dataset that will be used was obtained from UCI's (University of California Irvine) Machine Learning Repository and will be used for statistical modelling and visualization of the data. Some of the physicochemical variables include: Citric acid, residual sugar, chlorides, pH, and alcohol percentage. The wine's quality was graded by various wine experts on a scale of 0 (very bad) to 10 (very excellent) (Cortez et al. 2009). Moreover, this report will use Multiple Linear Regression along with Model selection techniques in order to create an appropriate model for modelling wine quality. Furthermore, this report will interpret the results of the determined model and discuss the possible implications.

The preference for the taste of wine is highly subjective just like with all other foods and drinks. There have been many studying showing that even many experts in blind taste tests cannot accurately differentiate between cheap and expensive wines (Derbyshire 2013). Thus, the possible implications of the study could show the objective physicochemical properties associated with wine quality. Moreover, a goal of this study is to identify dominant physicochemical variables associated with the taste of wine. The results could provide valuable insight into wine production and consumer preference.

The layout of this paper will be presented in several sections. Section 2 will discuss the source of the data and key features along with possible data limitations. Section 3 will present the selected Linear Regression Model along with an explanation of the variables and properties of the model. Next, Section 4 will present

[*] Code and data are available at: https://github.com/HanFrank/Analysis-of-Wine-Quality-based-on-its-physicochemical-properties

the results of the study with appropriate visualizations such as tables and graphs. Finally, Section 5 will be a commentary and analysis of the results along with weaknesses and potential next steps.

2 Data

The data used in this report was obtained through the UCI Machine Learning Repository (Dua and Graff 2017) and provided by (Cortez et al. 2009). The statistical analysis in this report will be done using `R` (R Core Team 2020). The R Packages, `tidyverse` (Wickham et al. 2019) and `dplyr` (Wickham et al. 2021) will be used for data manipulation and cleaning. The graphs and tables for this report will created and formatted with `ggplot2` (Wickham 2016) and `kableExtra` (Zhu 2021). The packages `bookdown` (Xie 2016) and `knitr` (Xie 2014) will be used to format this report.

The original dataset contained 4898 observations of different white Vinho Verde wines with 12 variables. These variables are: Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol level, and a quality score from 0 to 10. The first 11 variables are physicochemical properties that were measured through scientific tests while the quality score was determined by several experts grading each wine from 0 (very bad) to 10 (very excellent).

Table 1 is a summary of the variables in our dataset.

2.1 Model selection:

For our analysis, we will be using contextual reasoning combined with Model Diagnostics and the significance of the parameter estimates. Additionally, the Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) will be used to measure how well variables fit the model. We will not be using an automated variable selection procedure because it ignores the contextual reasoning behind the variable selection and may cause bias (Leeb and Potscher 2005). In general, we would be looking for a smaller model so as to not overfit the model and a smaller model allows for a more interpretable model.

Variables such as volatile acidity, fixed acidity, citric acid, pH all tell a similar story so pH will be used instead of all of these variables. This is also true for free sulphur dioxide, total sulphur dioxide and sulphates. Thus, the sulphates will be used to account for this factor. Density was not used because there is not a good contextual reason to why density would cause a difference in wine quality. Additionally, in the exploratory data analysis we found that the density variable has very small variability between its values, thus would probably not be very explanatory for analysis.

Need another few paragraphs justifying variables + AIC/BIC + model diagnostics

2.2 Data Limitations

Notes need to expand

1. Only white wine
2. Taste is subjective
3. Only 1 type of wine from 1 country
4. Many variables may be correlated (Multicollinearity)
5. Removing variables removes some potential explanatory ability for example citric acid and pH isn't entirely the same thing

3 Model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

where:

- y_i is the response variable, quality score of the wine
- β_0 is the estimated intercept coefficient
- $\beta_1 x_1$ is the estimated intercept and predictor variable for Alcohol percentage
- $\beta_2 x_2$ is the estimated intercept and predictor variable for pH
- $\beta_3 x_3$ is the estimated intercept and predictor variable for Sulphates
- $\beta_4 x_4$ is the estimated intercept and predictor variable for Residual Sugars
- ϵ is the error term

4 Results

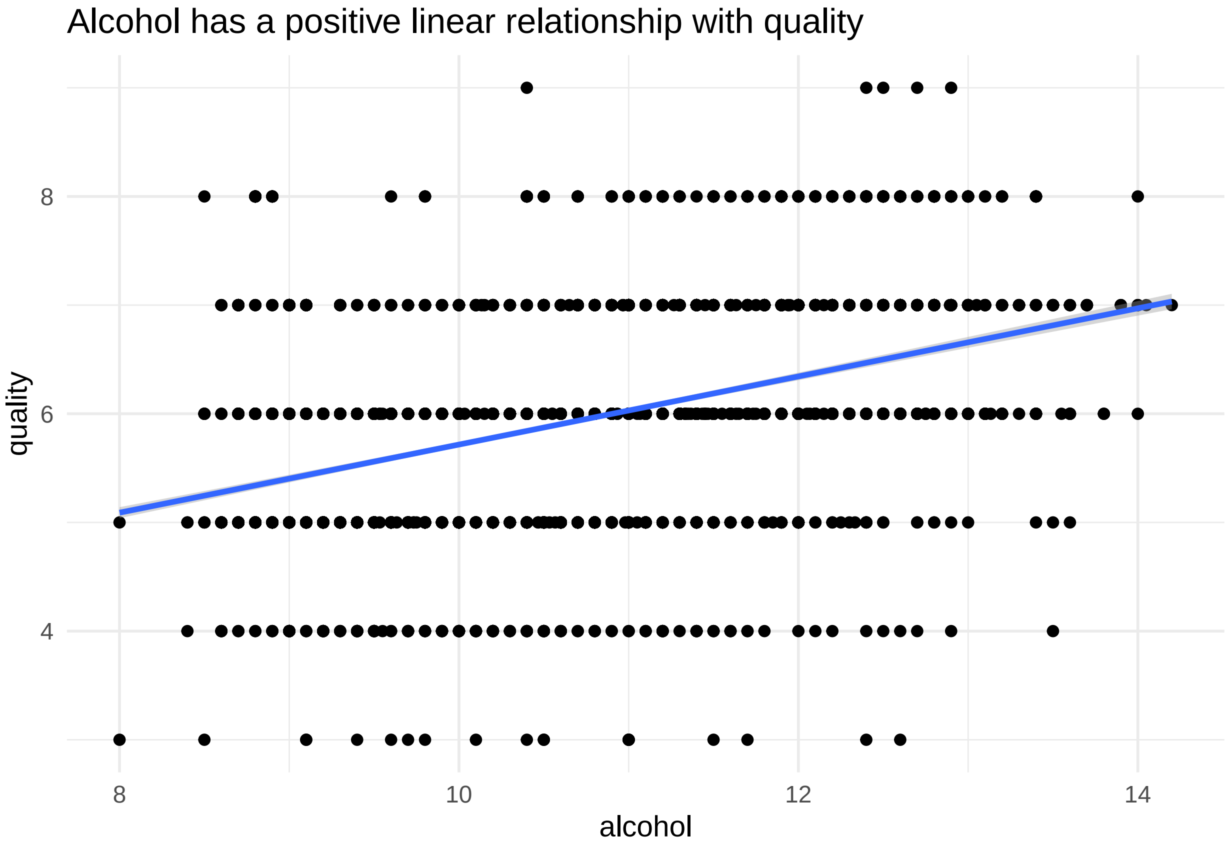


Figure 1: Alcohol vs Quality scores

Figure 1 shows a positive correlation between higher alcohol percentage in the wine and higher scores given by experts.

5 Discussion

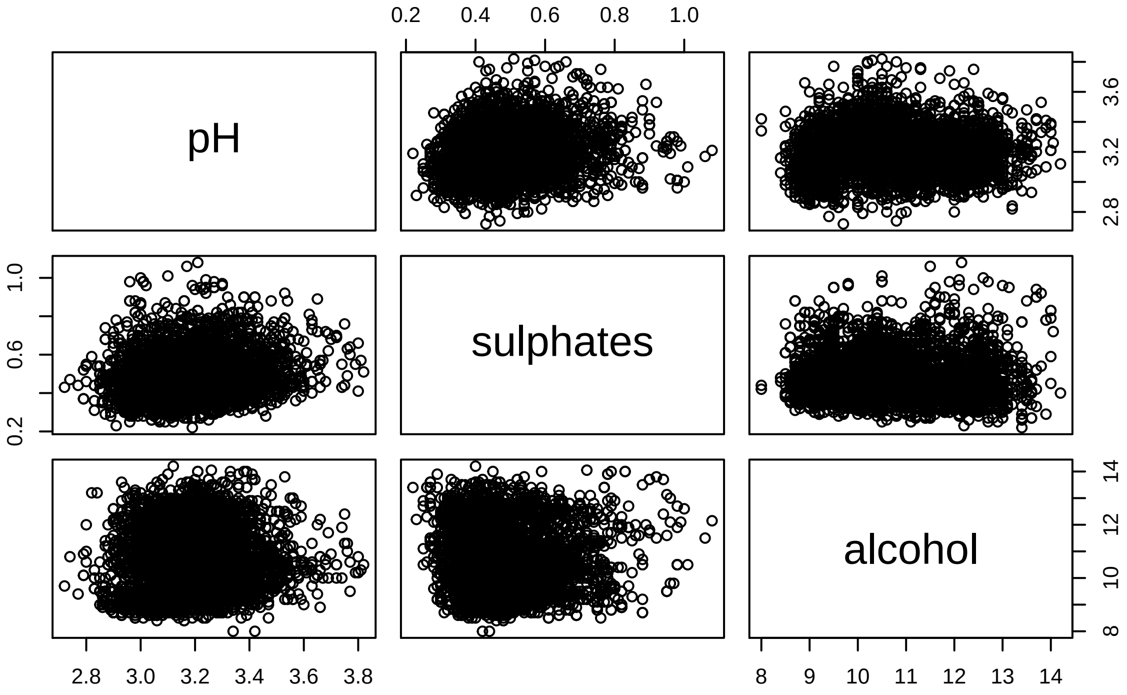
5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Appendix



Residual plots for variables in model

References

Cortez, Paulo, António Cerdeir, Fernando Almeida, Telmo Mato, and José Reis. 2009. "Modeling Wine Preferences by Data Mining from Physicochemical Properties." *In Decision Support Systems, Elsevier* 47 (4): 547–53.

Denig, Vicki. 2020. *How Is Wine Made?*

Derbyshire, David. 2013. *Wine-Tasting: It's Junk Science.* The Guardian.

Dua, Dheeru, and Casey Graff. 2017. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.

History of Wine - Wine History and Origins. 2022. Wine Facts.

Leeb, Hannes, and Benedikt Potscher. 2005. "Model Selection and Inference: Facts and Fiction." *Econometric Theory* 21 (1): 21–59.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.*

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2016. *Bookdown: Authoring Books and Technical Documents with R Markdown.* Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/bookdown.

Zhu, Hao. 2021. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.*