Predicting Wine Quality From Physicochemical Properties Using Linear Regression

Mohith Baskaran

University of British Columbia

April 16, 2025

1.1 Introduction

The dataset we chose to conduct our analysis is the Wine Quality dataset, which is publicly available from the UCI Machine Learning Repository. It contains the physicochemical and quality related attributes of red wine variants of the Portuguese "Vinho Verde" wine. The dataset

consists of 12 variables for each wine sample, including several continuous chemical properties and a quality score assigned by wine tasters. We will be exploring the following variables:

Response Variable:

- Quality: An integer score between 0 and 10 assigned to each wine sample, representing perceived quality (output target).

Explanatory Variables:

- fixed_acidity: Tartaric acid concentration (g/dm^3)
- volatile_acidity: Acetic acid concentration (g/dm^3)
- citric_acid: Citric acid concentration (g/dm^3)
- residual_sugar: Remaining sugar concentration (g/dm^3) after wine fermentation
- chlorides: Sodium chloride (salt) content (g/dm^3)
- free_sulfur_dioxide: Free form of SO_2 concentration (mg/dm^3)
- total_sulfur_dioxide: Total amount of SO_2 (free + bound) concentration (mg/dm^3)
- density: Wine density (g/cm^3)
- pH: Acidity level (pH level)
- sulphates: Potassium sulphate concentration (g/dm^3)
- alcohol: Alcohol content (Percentage %)

1.2 Motivation & Research Question

Wine is one of the most widely consumed beverages around the world, with its quality often influencing consumer preference, pricing, and production practices. As students interested in data analysis, we were curious about how measurable chemical properties of wine translate into perceived quality scores. Among the various factors, alcohol content and volatile acidity are frequently referenced in discussions around taste and preservation. This dataset gives us an opportunity to explore the relationship between these two variables and wine quality through a statistical lens. Based on this curiosity, we aim to investigate the following question:

How does alcohol and volatile acidity concentration affect wine quality?

2.1 Exploratory Data Analysis:

2.1.1 Covariates:

For the exploratory phase of our analysis, we first generated summary statistics and distribution plots for the dataset to identify any potential anomalies or irregularities (see Figures 8,9). All variables appeared well-behaved, exhibiting ranges and patterns consistent with expectations. Also no extreme outliers or irregular groupings were detected. Next, we created boxplots of wine quality across equal-width bins for each continuous predictor to assess whether linear modeling may be appropriate. This approach allowed us to visually inspect how the response variable, quality, varies with respect to each covariate. From these boxplots, we identified that fixed acidity showed little to no variation in the median quality across their binned values (see figure 1). This lack of trend suggested limited explanatory power, and thus we chose to exclude this variable from our final model. In contrast, other predictors such as alcohol, volatile, acidity, and sulphates displayed visible shifts in quality medians, indicating potential relevance for modeling.

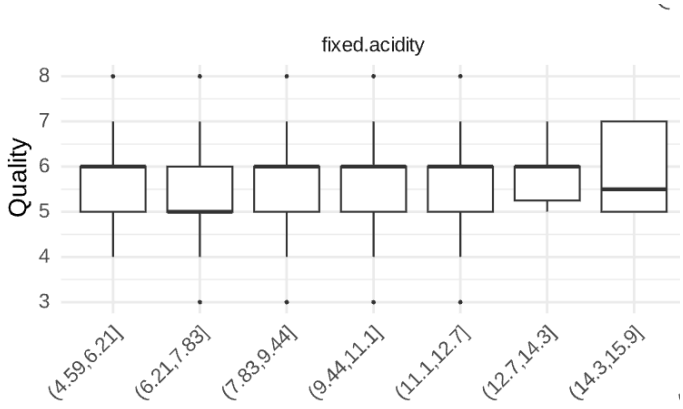

Figure 1. Boxplots of fixed acidity against Quality

2.1.2 Exploratory Data Analysis: Multicollinearity

To assess potential multicollinearity between predictors, we generated a correlation heatmap of the selected variables. While most pairwise correlations were low, a moderate correlation of 0.67 was observed between **free sulfur dioxide** and **total sulfur dioxide**, this correlation was an indicator that multicollinearity may be a potential problem that needed to be investigated (see Figure 2).

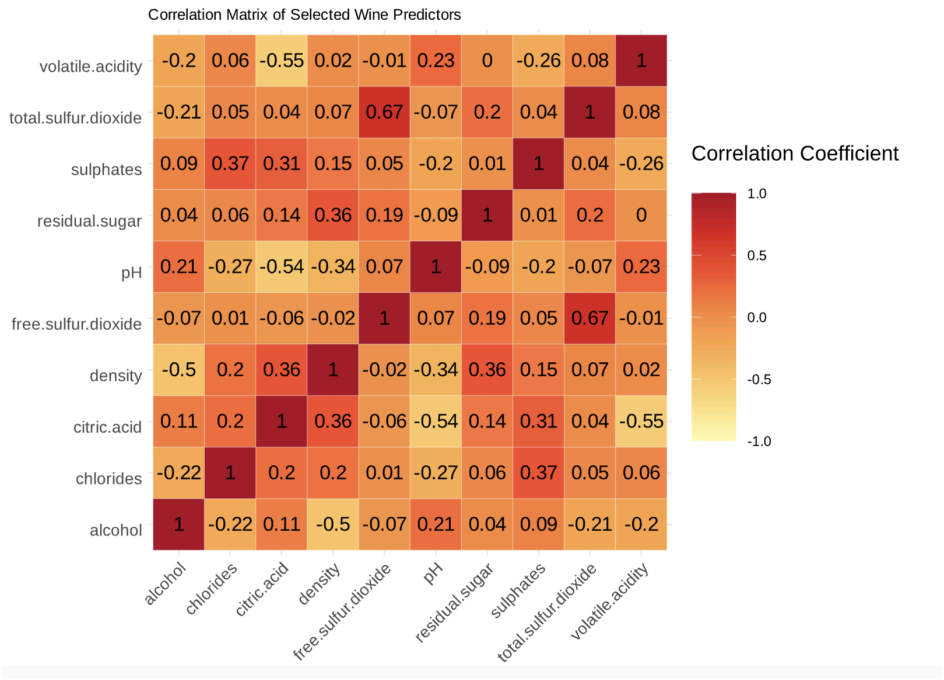

Figure 2: Correlation Heatmap

To find out whether or not there would be a problem we find the Variance Inflation Factor (VIF) for all covariates. The VIF value we found for **free sulfur dioxide** and **total sulfur dioxide** was 1.939 and 2.069 respectively. A large VIF is one generally larger than 10 as rule of thumb, our VIF values are well below that so we have no cause for concern. Likewise this is true for all covariates as we observe no VIF values greater than 3 (see Figure 3).

volatile.acidity: 1.785 **citric.acid:** 2.781 **residual.sugar:** 1.386 **chlorides:** 1.401 **free.sulfur.dioxide:** 1.939
total.sulfur.dioxide: 2.069 **density:** 2.43 **pH:** 1.611 **sulphates:** 1.396 **alcohol:** 2.136

Figure 3: VIF Values For Covariates

2.2 Model Analysis:

2.2.1 Linear Regression

To investigate how the predictors affect wine quality, we applied a linear regression model. While the response variable quality is recorded as integers ranging from 0 to 10, we chose to treat it as continuous for this analysis. This approach was taken as the outcome represents ordered values along a scale, and it allows us to apply linear methods that are both simple and interpretable. Each increase in wine quality score can reasonably be viewed as a consistent step

in perceived quality, which supports the decision to use a continuous modeling framework. This enables us to better understand how changes in factors like alcohol content or acidity are associated with overall wine quality.

2.2.2 Model Selection

To determine the most appropriate model for predicting wine quality, we applied both forward and backward stepwise selection to our full set of covariates, excluding **fixed acidity**, which was removed during exploratory analysis due to minimal variation on the response variable. Interestingly, both selection methods resulted in the same final model, indicating a stable set of important predictors. The selected model retained volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol (see Figure 4). Throughout the selection process, variables such as citric.acid and residual.sugar were removed. Given the agreement between selection methods and the interpretability of the resulting model, we chose to proceed with this set of predictors for our final analysis.

```
lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol, data = wine_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68918 -0.36757 -0.04653  0.46081  2.02954

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
volatile.acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
free.sulfur.dioxide   0.0050774  0.0021255   2.389    0.017 *
total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6477 on 1591 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3567
F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

Figure 4: Summary of Final Model

2.2.3 Model Diagnostics

We observe a banded pattern in the residuals vs. fitted plot, which is expected given that the response variable, wine quality, is discrete and integer-valued (see Figure 5). Although linear regression is typically used for continuous outcomes, the residuals remain randomly scattered

around zero, which is acceptable in this context and does not indicate major violations of model assumptions. In the QQ-plot for the final model, most points fall along the reference line, with the exception of the left tail, which shows slight deviation (see Figure 6). This suggests mild left-tailed non-normality in the residuals. While this violates the assumption of normally distributed errors, attempts to address the non-normality using log and square root transformations only worsened the heavy tail, so we proceed with the linear model while acknowledging this as a potential limitation.

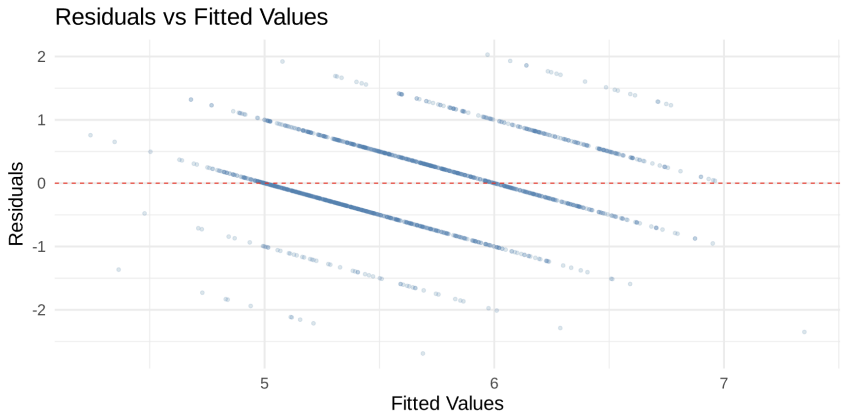

Figure 5: Residual vs Fitted Values

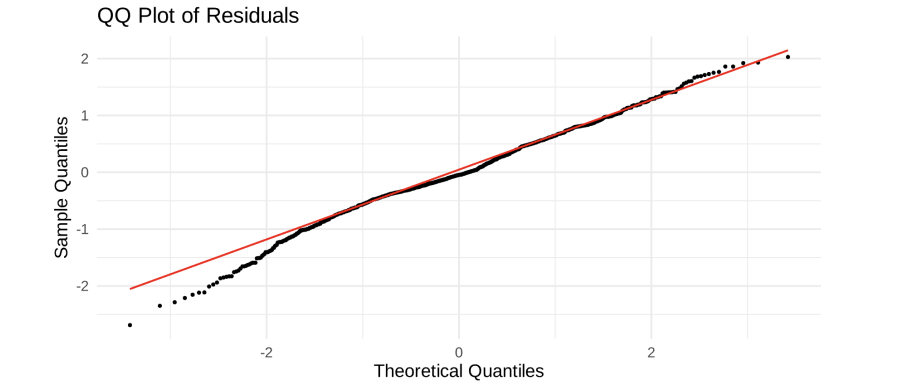

Figure 6: QQ-Plot of Residuals

2.2.4 Correlation Diagnosis:

We once again check the Variance Inflation Factors (VIF) for all of our predictors in the final model. Every VIF is below 2, indicating that no predictor is strongly explained by a linear combination of the others, so multicollinearity is negligible (see Figure 7).

volatile.acidity: 1.242 **chlorides:** 1.333 **free.sulfur.dioxide:** 1.883 **total.sulfur.dioxide:** 1.944 **pH:** 1.255 **sulphates:** 1.322 **alcohol:** 1.22

Figure 7 : VIF Values For Covariates In The Final Model

2.2.5 Reduced Model Interpretation:

Our fitted model suggests that all of the chosen covariates show very strong statistical significance, with the exception being free sulfur dioxide levels being slightly less significant compared to the other covariates in the model (see Figure 4).

Holding all other variables fixed, we find that:

- An increase of 1 g/dm³ in **volatile acidity** levels is associated with a decrease of -1.013 in expected wine quality.
- An increase of 1 g/dm³ in **sodium chloride** concentration denotes a decrease of -2.018 in expected wine quality.
- An increase of 1 mg/dm³ in **free sulfur dioxide** increases the expected wine quality by 0.005.
- An increase of 1 mg/dm³ in **total sulfur dioxide** levels denotes a decrease in expected wine quality by -0.003, which surprisingly contradicts the relationship of wine quality with free sulfur dioxide levels, despite the aforementioned pair-wise correlation value.
- An increase of 1 unit in **pH** levels seems to represent a decrease in expected wine quality by -0.483.
- An increase of 1 g/dm³ in the concentration of **potassium sulphate** increases the mean wine quality by 0.883.
- An increase of 1% in **alcohol** concentration denotes an increase of 0.289 in mean wine quality.

The intercept of the fitted model is 4.43, which means that the expected wine quality value is 4.43 when the values of all other covariates are 0 (see Figure 4).

3.0 Conclusion:

Our finding that there exists a positive relationship between alcohol percentage and mean wine quality is fascinating, as the wide perception seems to suggest the opposite, that the quality of wine should not be associated with alcohol levels at all, but would change the perception and flavour of the wine (Jordão et al., 2015). Some would even suggest that higher alcohol content deteriorates wine quality, as it causes wine to feel warmer and less balanced compared to their

lower alcohol counterparts (King, Dunn, & Heymann, 2013). A potential cause of this observed relationship is that wine with higher alcohol content usually exhibits stronger and more intense flavors, which may align with what most consumers would prefer, or that the selection of wine samples from our dataset consists of higher quality wine with greater alcohol percentage than those with less alcohol percentage.

Similar to alcohol content, our findings on the negative relationship between volatile acidity and wine quality should also be of note, as the general perception seems to agree with this. Excessive volatile acidity is commonly considered a fault and is associated with vinegar-like aromas and spoilage (Vilela-Moura et al., 2011), which likely explains the observed negative correlation with wine quality in our data.

3.1 Limitations:

- As noted in the beginning, our study is limited to wine samples that originate from the red wine variants of the Portuguese "Vinho Verde" wine, and therefore may not be a good representation of the general relationship between wine quality and the variables in question for most types of wine.
- It should be mentioned that taste is the least understood of the human senses (Smith & Margolskee, 2001), making wine classification is a difficult task. Furthermore taste is subjective leading to biases in the quality of the wines reported, as there is no standardized scoring. This can also be seen in the max and min of the scores reported as there is no wine rated lower than 3 or higher than 8.
- We assumed that our wine samples are independent of each other, but this assumption may not hold as our sample consists of variants of the same type of wine. As such, there may be similarities which may cause the independence assumption to be violated.
- The response variable (quality score of 0-10 integer scale) possibly violates the linear regression assumption of continuity. Patterns of the residual plot (banding in residuals vs. fitted) and slightly heavy tails in the QQ-plot suggest the model may not be suitable for the discrete nature of the outcome (see Figures 5,6).

- Our model has relatively low R^2 (0.3595) and adjusted R^2(0.3567) implying that much variance in wine quality remains unexplained by our selected properties. Other sensory attributes (e.g. aroma, tannins) may contribute to the model's explanatory power.
- Log and square root transformations failed to correct the heavy tails in the residuals, worsening them instead potentially highlighting a violation of the normality assumption.

3.2 Future Research:

Future studies may focus on a broader selection of wine samples, and may focus on the level of balance of the variables that contribute to higher wine quality (e.g. the well-balancedness of the contents of wine), instead of emphasizing on the distinct wine constituents. Additionally, future research may also explore variables outside the chemical properties of wine, such as the age of a wine sample, therefore providing us a clearer understanding of the characteristics of a high quality wine. Future studies may also expand the dataset to include diverse wine types (e.g., white, sparkling) and regions (e.g., France, Argentina), which would test the robustness of observed relationships and disentangle cultural or production-specific biases (Parr et al., 2020) from universal trends. Investigations may be helpful to resolve counterintuitive findings and to reinforce the explanatory power and credibility of the model. Integrate other relative data such as soil composition, harvest may enrich models and bridge the gap between our physicochemical measures and holistic quality assessments.

Submission: Mohith Baskaran Submitted The R Code & Dataset On Canvas

Appendix

```
fixed.acidity   volatile.acidity  citric.acid    residual.sugar
Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
  chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
      pH          sulphates         alcohol         quality
Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
Median :3.310   Median :0.6200   Median :10.20   Median :6.000
Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Figure 8: Summary Statistics For Covariates

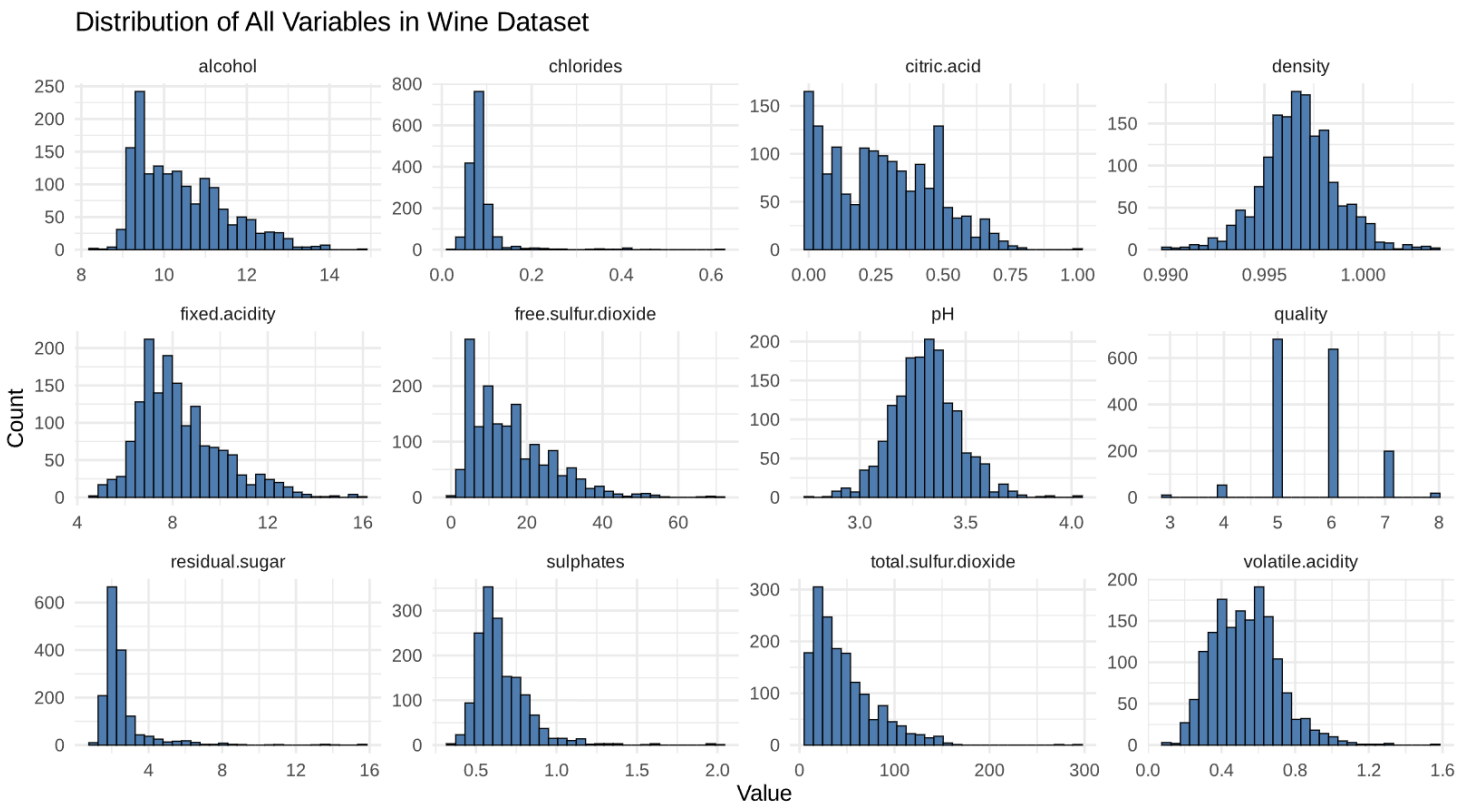

Figure 9: Distribution of Covariates

References

Jordão, A., Vilela, A., & Cosme, F. (2015). From Sugar of Grape to Alcohol of Wine: Sensorial Impact of Alcohol in Wine. *Beverages*, *1*(4), 292–310. https://doi.org/10.3390/beverages1040292

King, E. S., Dunn, R. L., & Heymann, H. (2013). The influence of alcohol on the sensory perception of red wines. *Food Quality and Preference*, *28*(1), 235–243. https://doi.org/10.1016/j.foodqual.2012.08.013

Parr, W. V., & Rodrigues, H. (2020). Cross-Cultural Studies in Wine Appreciation. *Handbook of Eating and Drinking*, 1467–1490. https://doi.org/10.1007/978-3-030-14504-0_168

Smith, D. V., & Margolskee, R. F. (2001). Making Sense of Taste. *Scientific American*, *284*(3), 32–39. https://doi.org/10.1038/scientificamerican0301-32

Vilela-Moura, A., Schuller, D., Mendes-Faia, A., Silva, R. D., Chaves, S. R., Sousa, M. J., & Côrte-Real, M. (2010). The impact of acetate metabolism on yeast fermentative performance and wine quality: reduction of volatile acidity of grape musts and wines. *Applied Microbiology and Biotechnology*, *89*(2), 271–280. https://doi.org/10.1007/s00253-010-2898-3