# Modeling Portuguese Red Wine Quality via Physicochemical Properties

Name: Nicholas Gunner
NetID: nrg42

## Introduction

For this project, I analyzed a dataset [1] of tasting ratings of a Portuguese red wine called ‘Vinho Verde’. The dataset contained 1,599 observations and 11 physicochemical properties of these wines along with a quality score ranging from 0 to 10.

My goal in conducting this analysis was to investigate the ways that linear regression modeling could predict perceived wine quality using these properties. Additionally, I wanted to interpret the results in a way that might be useful for winemakers.

## Methodology

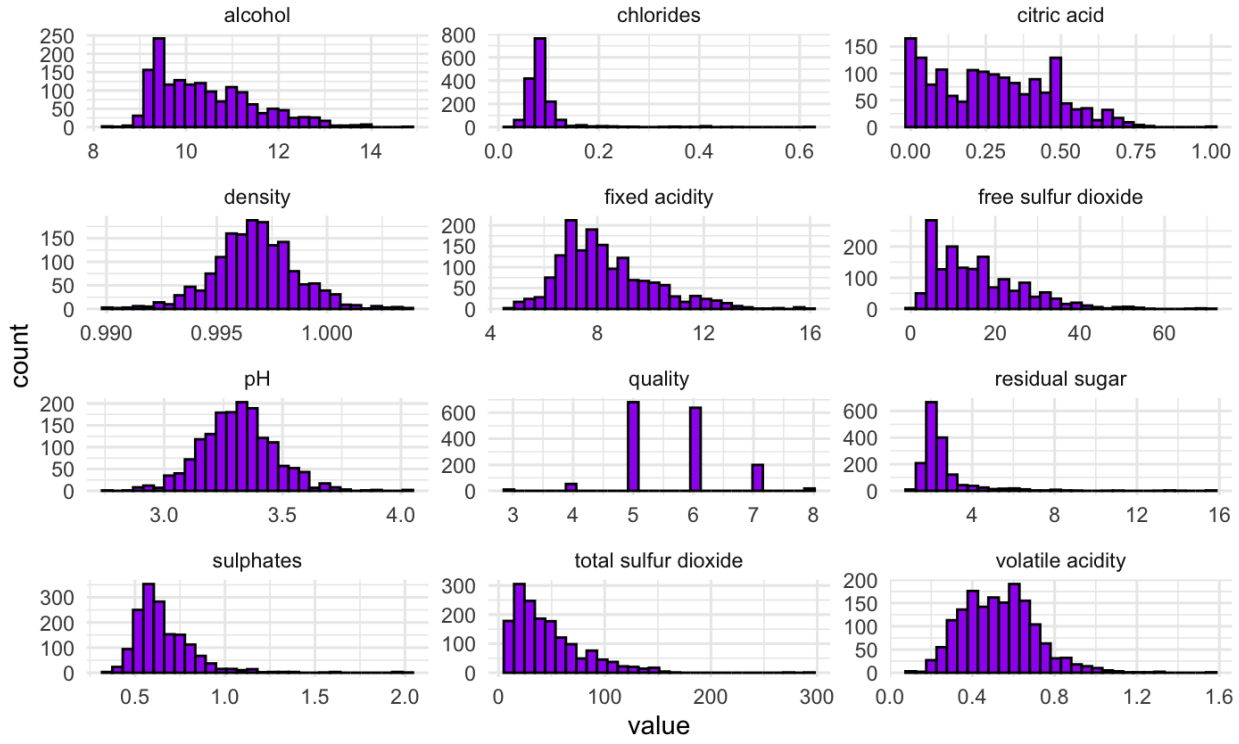
The variables in this dataset were as follows:

| **Variable** | **Description** |
|---|---|
| Fixed acidity | Acids that are fixed and remain through the fermentation process. |
| Volatile acidity | Acids that are often associated as ‘vinegar-like’ and evaporate easily. |
| Citric acid | Acid that is associated with ‘freshness’ in wine. |
| Residual sugar | Sugars that remain in wine after the fermentation process and contribute sweetness. |
| Chlorides | Salts that can be found in wines. |
| Free sulfur dioxide | Available sulfur dioxide in the wine that acts as a preservative and limits microbial growth. |
| Total sulfur dioxide | The total sulfur dioxide content in a wine including those that are bound to other compounds. |
| Density | The mass per volume of wine. Can be associated with ‘mouthfeel’. |
| pH | The overall pH measure of acidity in the wine. |
| Sulphates | An additive that acts as a preservative as well as being associated |

| Variable | Description |
|---|---|
| | with wine structure and ‘mouthfeel’. |
| Alcohol | The percentage of the wine that is alcohol. |
| Quality | Subjective scores between 1 and 10 submitted by tasters. |

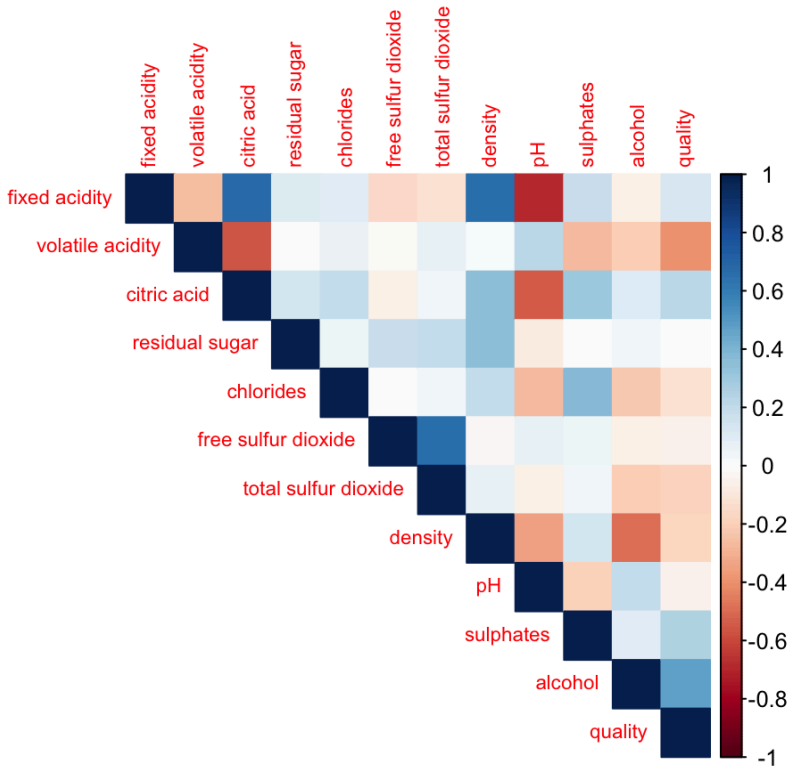
## Data Preparation and Exploratory Analysis

I began my analysis by exploring the structure and distribution of the dataset as a whole. Using Rstudio, I summarized each variable and visualized their distributions with histograms and boxplots. Several variables – including chlorides, residual sugar, sulphates, and total sulfur dioxide – were right-skewed, suggesting that log transformation might improve model performance.




*In the histograms created above, most variables display a normal, bell-shaped curve with a few notable exceptions. Chlorides, residual sugar, sulphates, and total sulfur dioxide show right tails. It is notable that residual sugar does not display a positive or negative correlation with quality.*

A correlation matrix was generated to help identify relationships among predictors and the outcome variable (Quality). For instance, alcohol had a strong positive correlation with quality, while volatile acidity showed a strong negative relationship. I also identified potential multicollinearity between pH, fixed acidity, and citric acid.




*In the correlation plot, we can quickly identify some key variables that have an impact on quality of wines. Specifically, volatile acidity, sulphates, total sulfur dioxide, and alcohol stick out. We also notice that pH has a strong relationship with fixed acidity and citric acid, suggesting that these may be dependent variables.*

## Assumption Checking and Cleaning

I fit an initial model using all predictors and then checked for linear regression assumptions:

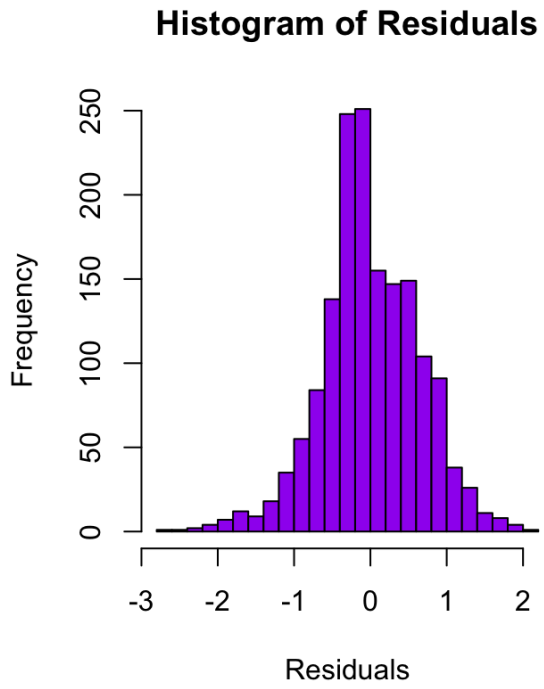
**Linearity**: Scatterplots and trend lines indicated linear relationships for several predictors.

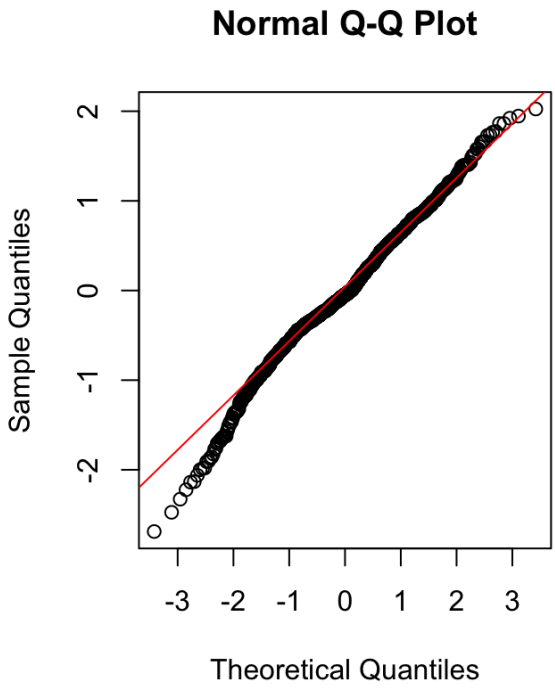
**Normality of Residuals**: A histogram and Q-Q plot suggested approximately normal residuals, with slight deviation in the tails.

**Homoscedasticity**: Residuals were fairly evenly spread across fitted values.

**Independence**: The ACF plot showed signs of autocorrelation, likely due to interactions between pH, fixed acidity, and citric acid which all interact with each other. Furthermore, alcohol and sugar directly impact density in wine. These variables were determined to violate independence assumptions. To address this issue, only pH and alcohol were retained from these mentioned variables in the final model.

**Multicollinearity**: VIF values flagged high collinearity in fixed acidity and density. Based on this observation as well as domain knowledge, I excluded fixed acidity, citric acid, density, and free sulfur dioxide from my final model.




*Residuals are normally distributed and the normal Q-Q Plot is also acceptably normal due to the points primarily following the red line. We do see some deviation on the lower ends of the Q-Q plot but this doesn't seem too dramatic.*

## Variable Selection Techniques

I used two variable selection techniques:

**Stepwise regression** which selected a six-variable model:

```
                     Df Sum of Sq    RSS     AIC
<none>                             669.93 -1377.1
+ `residual sugar`    1     0.284 669.65 -1375.7
- pH                  1     5.919 675.85 -1365.0
- `total sulfur dioxide` 1  9.233 679.16 -1357.2
- chlorides           1    10.647 680.58 -1353.8
- sulphates           1    27.445 697.38 -1314.9
- `volatile acidity`  1    44.972 714.90 -1275.2
- alcohol             1   125.812 795.74 -1103.9
```

**Lasso regression** which confirmed similar variables as important by shrinking others toward zero:

```
                                s1
(Intercept)             4.209006381
`volatile acidity`     -1.038517292
`residual sugar`        0.007366408
chlorides              -1.917411522
`total sulfur dioxide` -0.002361866
pH                     -0.404571809
sulphates               0.870127666
alcohol                 0.287905538
```

## Final Model and Validation

My final model included: volatile acidity, chlorides, total sulfur dioxide, pH, sulphates, and alcohol. I validated this model using 10-fold cross-validation which yielded consistent RMSE and R-squared values compared to the training model.

*Quality = 4.296 - 1.038(Volatile Acidity) - 2.002(Chlorides) - 0.0024(Total Sulfur Dioxide) - 0.435(pH) + 0.889(Sulphates) + 0.291(Alcohol)*

| | **R-Squared** | **RMSE** |
|---|---|---|
| **Final Model** | 0.3572 | 0.647278 |
| **Cross-validation** | 0.3529 | 0.650207 |

# Results

The final model produced an R-squared of 0.3572 and adjusted R-squared of 0.3548, which explained approximately 35.7% of the variation in wine quality scores. The model's RMSE was around 0.65, suggesting that predictions are typically within just over half of a point on the 0-10 quality scale. Cross-validation confirmed these results, supporting the model's generalizability.

All predictors were statistically significant ($p < 0.001$). Alcohol had the strongest positive effect with a predicted increase of 0.29 per increased percent. Volatile acidity had the strongest negative effect with a predicted decrease in score of 1.04 per additional unit. Sulphates were also positively associated with quality, whereas chlorides, total sulfur dioxide, and pH had negative effects.

# Discussion

Subjective ratings of wine have high variability. However, despite this constraint, our model supports known understanding of winemaking. Higher alcohol and sulphates enhance wine favorability and lead to higher perceived quality. Volatile acidity, which can be associated with

spoilage, reduces quality. High chlorides, sulfur dioxide preservative, and pH also correspond to lower quality ratings.

The subjectivity of quality scores, a lack of information about the tasting process, and a limited sample size were limitations in this analysis. Additionally, while transformations were considered, they did not significantly improve performance and were excluded for the sake of simplicity.

## Conclusion

This project demonstrates that linear regression, as well as proper variable selection, can provide valuable insights into the makeup of quality red Portuguese wines. Our final model focussed on simplicity and key variables that might be useful for wine producers. Future work might explore additional chemical properties of wines, expand to wines from wider production regions, include a wider variety of winemaking styles (white, rose, etc.), and document taster data such as age, nationality, etc.

By studying the multivariate impacts of chemistry on perceived wine quality, winemakers can uncover actionable insights that help them make informed decisions and improve their craft.

## References

1.) UCI Machine Learning Repository: Red Wine Quality Dataset. https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009