The Perfect Bottle: Quality through Chemistry

MGT 256 Final Report

Megan Dang Arvind Kamboh Matthew Barclay Riley Baumgarten

Research Problem

Can we predict the quality of a wine based on its chemical properties?

This question is important for decision-makers in the wine production industry because the quality assessments can impact their product sales. Within the market for vinho verde wines, a highly rated wine may command a high price and high demand while lower ratings may decrease demand and command a lower price. In order to maximize profits, we assume the wine producers should be incentivized to maximize their wine's quality ratings. Because quality ratings can be subjective and difficult to predict, the wine producers may not know how the quality will turn out until they submit their final product for evaluation. If wine producers were able to understand how specific physicochemical variables influence the wine quality, they can set certain targets during the production process and can predict the quality outcome. With this information, wine producers can attempt to maximize their quality ratings and can more easily set prices and estimate demand.

Data Exploration

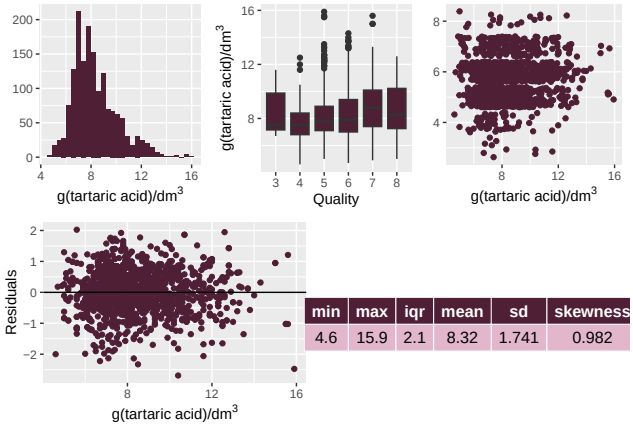
About The Dataset

This dataset comes from the UC Irvine Machine Learning Repository archive (Cortez and Reis 2009). This dataset was uploaded to the archive on October 6th, 2009. The dataset originally included two different datasets that included both red and white wine from Portuguese "Vinho Verde" wine, however to keep our variables to a certain limit and keep the report more focused, we decided to use the Red Wine dataset for our analysis. This dataset includes 12 numerical variables that we chose for our analysis and has 1599 total observations. The data was cleaned prior to our observations so the data has no missing variables.

Predictor Variables

1. fixed acidity (g(tartaric acid)/dm3)

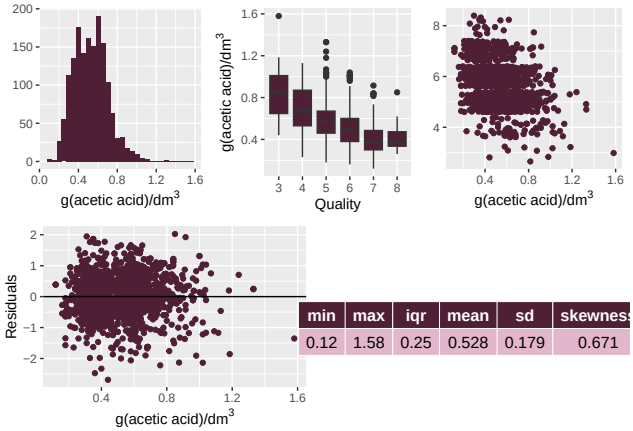
- The most common fixed acids in wines are: tartaric, malic, citric, and succinic, with tartaric acid being measured in this case. In the histogram, we can see a slight positive skew within the relatively normal distribution. The boxplot shows several outliers and as quality increases, there is a small change in fixed acidity but is still somewhat flat. The scatterplot does not show any linear relationships and is mostly clustered. The residual plot shows plots that are mostly randomly distributed, being closely packed in the center and widely scattered around the edges with a few outliers.



min	max	iqr	mean	sd	skewness
4.6	15.9	2.1	8.32	1.741	0.982

2. volatile acidity (g(acetic acid)/dm3)

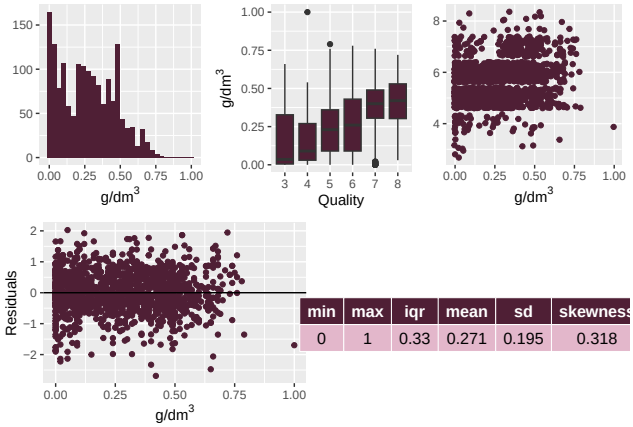
- The primary acid in volatile acidity is acetic acid, which gives wine the smell and taste of vinegar. In the histogram, we can see a mostly normal distribution. The boxplot shows several outliers and as quality increases, there is a negative linear relationship. The scatterplot shows a slight negative linear relationship and is mostly clustered. The residual plot shows plots that are closely packed in the center and widely scattered around the edges with a few outliers.



min	max	iqr	mean	sd	skewnes
0.12	1.58	0.25	0.528	0.179	0.671

3. citric acid (g/dm3)

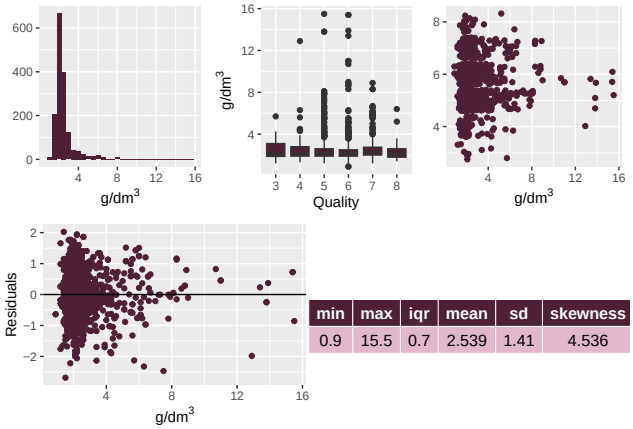
- Citric acid is relatively low in wine, but is added to increase acidity or to counteract tannic haze (iron instability which produces tannis). In the histogram, we can see a somewhat positive skew. The boxplot shows a few outliers and as quality increases, there is a positive linear relationship. The scatterplot shows a cluster in the middle with a spread of plots around the perimeter. The residual plot shows a random distribution.



min	max	iqr	mean	sd	skewness
0	1	0.33	0.271	0.195	0.318

4. residual sugar (g/dm3)

- Residual sugar is the natural grape sugars left over after the alcohol fermentation is completed. In the histogram, we can see a strong positive skew. The boxplot shows many outliers and as quality increases, there is no skew and is flat across quality. The scatterplot shows a cluster in one portion and is slightly random. The residual plot shows a random distribution.

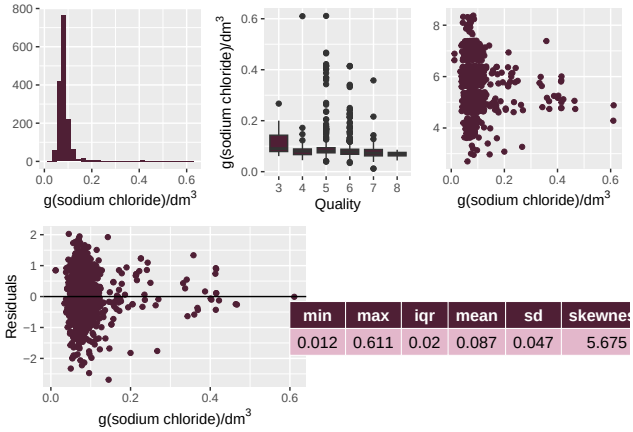


min	max	iqr	mean	sd	skewness
0.9	15.5	0.7	2.539	1.41	4.536

5. chlorides (g(sodium chloride)/dm3)

- Chlorides add to the saltiness of the wine, adjusting the acidity and taste. In the histogram, we can see a strong positive skew. The boxplot shows several outliers and as quality increases, there is a negative

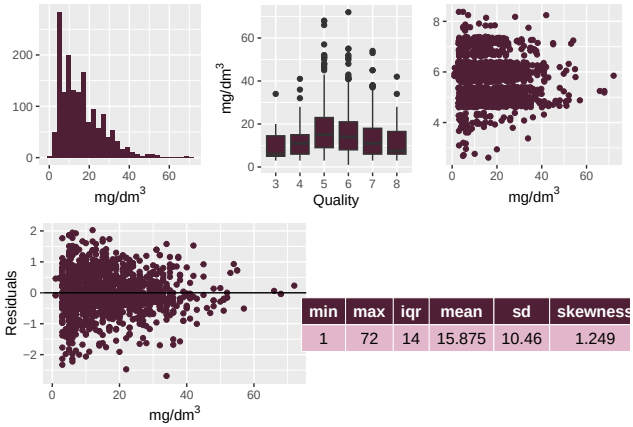
linear relationship. The scatterplot shows a negative linear relationship and is mostly clustered around 0.087 with outliers decreasing from 0.2 to 0.6. The residual plot shows plots that are closely packed in the center and outliers scattered decreasingly toward 0.6 .



min	max	iqr	mean	sd	skewnes
0.012	0.611	0.02	0.087	0.047	5.675

6. free sulfur dioxide (mg/dm3)

- Sulfur dioxide or sulfite, is a natural byproduct of fermentation. Free sulfur dioxide means no other molecules are bonded to the sulfur dioxide. In the histogram, there is a positive skew. The boxplot shows several outliers and as quality increases, there is a flat relationship. The scatterplot shows a slight negative linear relationship and is mostly clustered. The residual plot shows plots that are closely packed in the center and widely scattered around the edges with a few outliers.

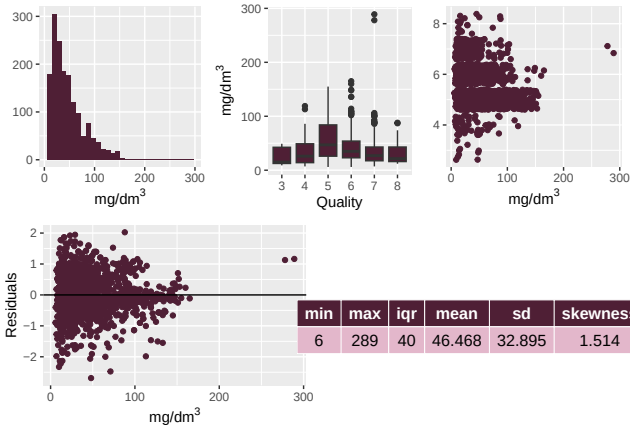


min	max	iqr	mean	sd	skewness
1	72	14	15.875	10.46	1.249

7. total sulfur dioxide (mg/dm3)

- Sulfite is a preservative and enhancer, used to stop fermentation or help protect the wine. Total sulfur dioxide measures all types of sulfur dioxide molecules. In the histogram, there is a positive skew. The boxplot shows several outliers and as quality increases, there is a flat relationship. The scatterplot shows a slight negative linear relationship and is mostly clustered. The residual plot shows plots that

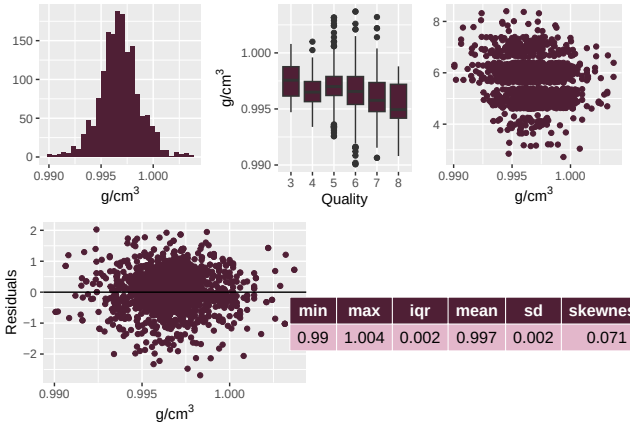
are closely packed in the center and widely scattered around the edges with a few outliers.



min	max	iqr	mean	sd	skewnes
6	289	40	46.468	32.895	1.514

8. density (g/cm3)

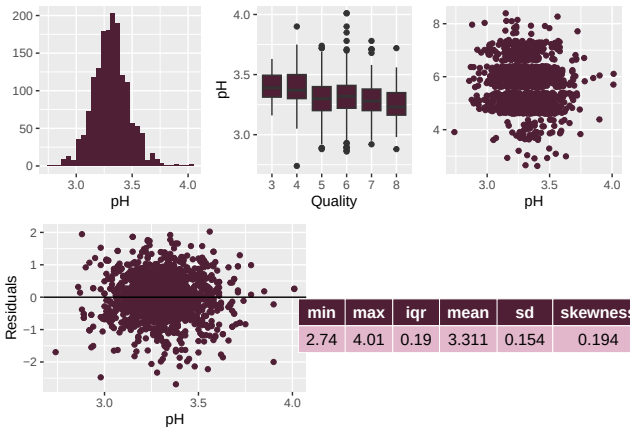
- Density has little effect on wines, but allows for overall control of juice volume in the fermentation process. In the histogram, we can see a normal distribution that has a central tendency. The boxplot shows several positive and negative outliers and as quality increases, there is a slight negative linear relationship. The scatterplot shows a slight negative linear relationship and is mostly clustered. The residual plot shows a random distribution with few outliers.



min	max	iqr	mean	sd	skewnes
0.99	1.004	0.002	0.997	0.002	0.071

9. pH

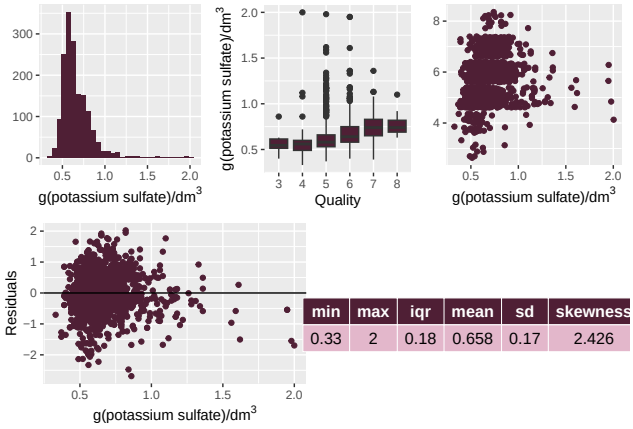
- pH is the acidity level, ranging on a log scale from 0 (very acidic) to 14 (very basic) with 7 being neutral. In the histogram, we can see a normal distribution. The boxplot shows multiple outliers around a rating of 6 quality and a slight negative trend between pH and quality. The scatter plot does not show a strong linear relationship and is mostly clustered. The residual plot shows random residuals that are closely packed in the center and widely scattered around the edges with a few outliers.



min	max	iqr	mean	sd	skewnes
2.74	4.01	0.19	3.311	0.154	0.194

10. sulfates (g(potassium sulfate)/dm3)

- Potassium sulfate is an antioxidant that fights off microbial entities that would spoil the wine. Little is needed, approx. 1/2 teaspoon/gal of wine. In the histogram, we can see a positively skewed distribution. The boxplot shows several outliers throughout the entire distribution, with a majority of them existing around the center, and a positive relationship with quality. The scatter plot is mostly clustered with a slight positive correlation including the few outliers. The residual plot shows plots that are closely packed in the left side of the graph and has a few outliers that exist as the potassium sulfate increases.

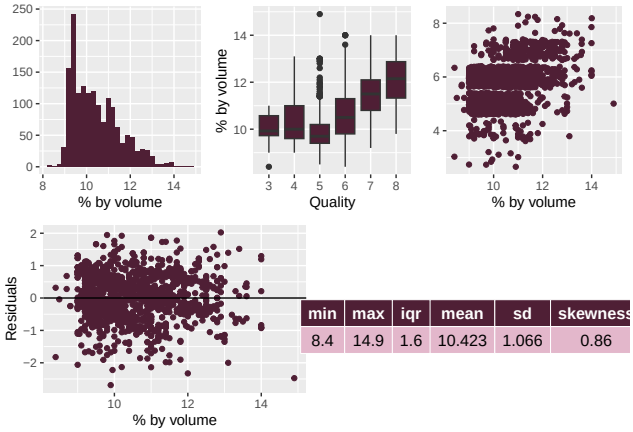


min	max	iqr	mean	sd	skewness
0.33	2	0.18	0.658	0.17	2.426

11. alcohol (% vol.)

- Alcohol, also abbreviated as ABV, is the amount of ethanol in a given volume of liquid, given in a percentage. ABV in wine usually ranges from 5.5% to 25%. Alcohol can affect the "body" of wine which is the weight and richness of the liquid felt when drinking. Higher ABV usually translates to a more "full-bodied" wine. In the histogram, we can see a relatively positive skew in the distribution. The boxplot shows several outliers in the middle of the distribution and a positive relationship with quality at higher quality scores. The scatter plot is mostly clustered with a positive correlation. The

residual plot shows plots that are randomly distributed in the center and widely scattered around the edges with a few outliers.

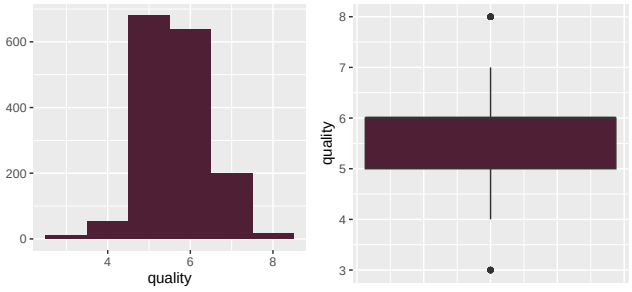


min	max	iqr	mean	sd	skewness
8.4	14.9	1.6	10.423	1.066	0.86

Dependent variable

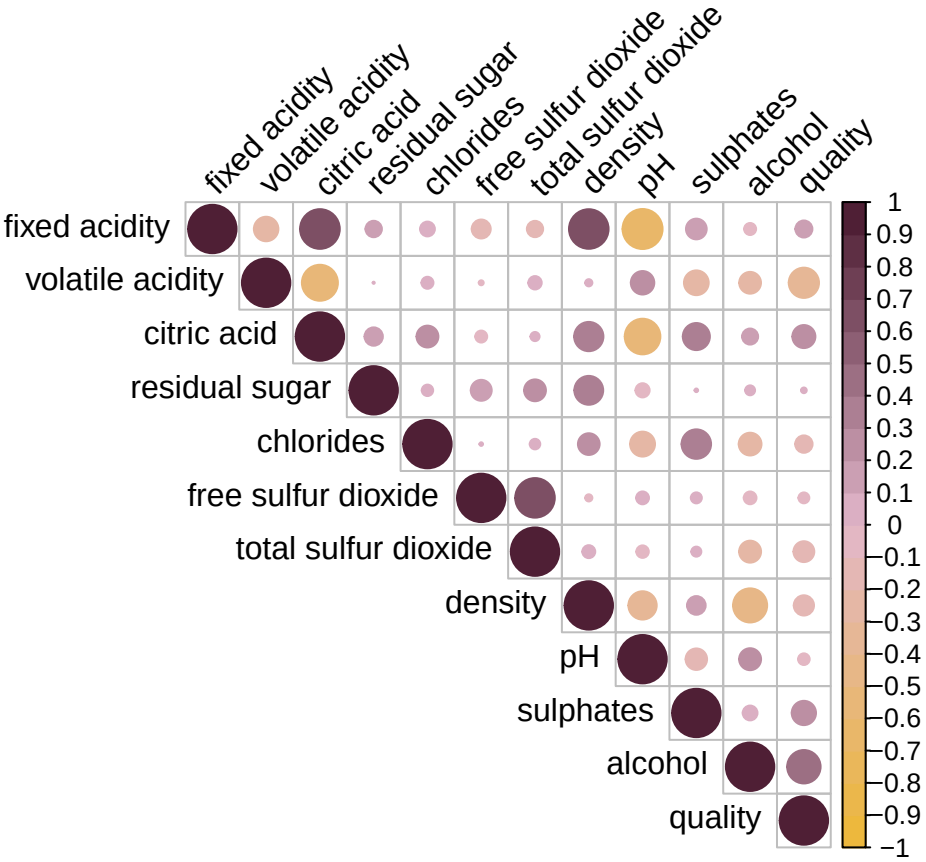
12. quality (score between 0 and 10)

- The quality ratings are determined by the independent Viticulture Commission of the Vinho Verde region (CVRVV) during the certification process and are rated on a scale of 1 to 10, with 10 being the highest quality. The ratings are determined through tasting and comparison to CVRVV standards. In the histogram, quality is normally distributed around a score of 6, showing that most wines of the Portuguese "Vinho Verde" type score just above an average score of 5. The boxplot looks balanced with outlier points on both sides of the distribution.



min	max	iqr	mean	sd	skewness
3	8	1	5.636	0.808	0.218

Correlation



When predicting quality, the correlation between the predictors and quality may foreshadow the results of the linear regression. We see that quality is most correlated with alcohol content and volatile acidity, so we expect that these will be important in the linear regression model. Additionally, there was negative correlation between the predictors volatile acidity, citric acid, fixed acidity, and pH. This is expected, as higher amounts of acid will result in a lower pH. We will keep these correlations in mind when running the linear regression in order to prevent possible multicollinearity.

Modeling Wine Quality

Linear Regression Variable Selection via AIC

Using backwards stepwise AIC variable selection, the linear regression model was found to be most significant using the following equation:

$$Quality = \beta_i + \beta_1 VolatileAcidity + \beta_2 Chlorides + \beta_3 FreeSulfurDioxide + \beta_4 TotalSulfurDioxide + \beta_5 pH + \beta_6 Sulfates + \beta_7 Alcohol$$

Therefore, the stepwise AIC function removed the following predictors as insignificant:

- Fixed Acidity
- Citric Acid
- Residual Sugar
- Density

We will choose to drop these predictors from our final model in order to maximize the significance of the model. The full model AIC was -1375.49 and was minimized to -1380.79 through the stepwise AIC process.

Validation

To check for overfitting in the final model, we used cross validation to test the final model performance on unseen data. The final linear model was validated by randomly partitioning the data into a training-validation and test set using 60% and 40% of the dataset, respectively. The model was trained using 10-fold k-means cross-validation, then the test data was used to predict the wine quality. Comparing the predicted quality values to the observed test quality, the following error was observed:

ME	RMSE	MAE	MPE	MAPE
−0.002	0.634	0.492	−1.345	9.055

Final Linear Model

$Quality = 4.4301 - 1.0127 * VolatileAcidity - 2.0178 * Chlorides + 0.0051 * FreeSulfurDioxide - 0.0035 * TotalSulfurDioxide - 0.4827 * pH + 0.8826 * Sulfates + 0.2893 * Alcohol$

	MLR / Est.	MLR / p
(Intercept)	4.430	<0.001
volatile acidity	-1.013	<0.001
chlorides	-2.018	<0.001
free sulfur dioxide	0.005	0.017
total sulfur dioxide	-0.003	<0.001
pH	-0.483	<0.001
sulphates	0.883	<0.001
alcohol	0.289	<0.001
Num.Obs.	1599	

	MLR / Est.	MLR / p
R2	0.359	
R2 Adj.	0.357	
F	127.555	
RMSE	0.65	

- For all variables, the p-values are less than 0.05 meaning that they are highly significant. The R-squared shows that the model explains 35.9% of the variation.
- Volatile acidity and quality are negatively correlated and for every unit change in volatile acidity, the estimated change in quality is -1.01 assuming all other variables are held constant.
- Chlorides and quality are negatively correlated and for every unit change in chloride, the estimated change in quality is -2.02 assuming all other variables are held constant.
- Free sulfur dioxide and quality are positively correlated and for every unit change in free sulfur dioxide, the estimated change in quality is 0.005 assuming all other variables are held constant.
- Total sulfur dioxide and quality are negatively correlated and for every unit change in total sulfur dioxide, the estimated change in quality is -0.003 assuming all other variables are held constant.
- pH and quality are negatively correlated and for every unit change in pH, the estimated change in quality is -0.48 assuming all other variables are held constant.
- Sulfates and quality are positively correlated and for every unit change in sulfates, the estimated change in quality is 0.88 assuming all other variables are held constant.
- Alcohol and quality are positively correlated and for every unit change in alcohol, the estimated change in quality is 0.29 assuming all other variables are held constant.
- The intercept estimate of 4.43 indicates that a wine with zero sulfites, volatile acidity, pH, etc. would still score a 4.4 out of 10, as unrealistic as that scenario would be.

K-Nearest Neighbor

KNN Model

k	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
38	0.653	0.356	0.52	0.047	0.075	0.028

Similar to the linear regression model, we implemented the K-nearest neighbors (KNN) algorithm on the test dataset using a 10-fold cross-validation methodology to ascertain the most proximate neighbors. Our analysis

revealed that the optimal parameter value for k was determined to be 38, resulting in performance metrics of an RMSE (Root Mean Square Error) of 0.6527, an R^2 (coefficient of determination) of 0.3564, and an MAE (Mean Absolute Error) of 0.5196. These findings contribute valuable insights into the efficacy of the KNN model in our predictive analysis.

KNN Validation

ME	RMSE	MAE	MPE	MAPE
−0.023	0.667	0.535	−2.029	9.91

The KNN model was validated using an unseen test partition of 40% of the dataset. It performed well, with the predicted RMSE and MAE values being close to the training data model indicating there was not significant overfitting.

Conclusions

The Linear Regression Model

In conclusion, our research project focused on analyzing the quality of wine with a particular emphasis on the variables: sulfates, pH, volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, and alcohol.

The linear model we constructed revealed compelling insights into the relationship between these variables and wine quality. The coefficients provided a quantitative understanding of the impact of each variable on the overall quality of the wine. Notably, sulfates, free sulfur dioxide, and alcohol exhibited significant positive relationships with wine quality while volatile acidity, chlorides, and total sulfur dioxide exhibited significant negative relationships with quality.

The significance of our findings is underscored by the high level of statistical significance, with predictor and model p-values well below the conventional threshold. The multiple R-squared value of 0.3595 suggests that our model explains approximately 36% of the variability in wine quality, providing a substantial insight into the influential factors. With quality ratings being highly subjective based on taste preferences, this R^2 is surprisingly high, showing the strength of our model.

The residuals analysis indicated a reasonably good fit of the model, with residuals showing a symmetrical distribution around zero. The residual standard error of 0.6477 further attests to the model's accuracy in predicting wine quality.

In practical terms, winemakers can leverage this model to enhance their production processes by focusing

on optimizing higher sulfate and alcohol levels while keeping volatile acidity, chlorides, and sulfur dioxide levels lower. This research contributes valuable knowledge to the field, providing a quantitative basis for understanding and improving wine quality. Future studies may explore additional variables or refine the model further to enhance its predictive capabilities. Overall, our research contributes to the ongoing pursuit of precision and excellence in winemaking practices.

The KNN Model

The KNN model was successful in minimizing the RMSE to just 0.6528 with an R^2 of 0.3564, indicating this model explains about 36% of the variation in quality ratings. As with the linear regression model, we are proud of the performance of the KNN model in predicting the quality of wine given the nature of taste ratings.

While looking at the specific predictor estimates is not useful for KNN as it is for linear regression, the overall model is useful to apply to real life situations. This KNN model can be used by winemakers to predict the quality of their wines before they are officially rated. By inputting their wines' chemical composition into the model, the 38 closest wines are used to predict the outcome quality.

Comparing The Models

While both models performed strongly, we can compare their strengths especially when facing new, unseen data. The multiple linear regression predicted the test partition at a RMSE of 0.6338 while the KNN model predicted a RMSE of 0.6670. The MLR prediction has a lower RMSE which means there is a lower magnitude of errors in this model. The MAPE (Mean Average Percent Error) of the MLR predictions was 9.05% while the KNN model had a MAPE of 9.91%.

Using these metrics, we find that our multiple linear regression model performed the best of the two models. Because we removed insignificant predictors from our linear model, its higher predictive performance and lower data collection requirement make it a strong candidate for use within the wine industry. Our KNN model should not be forgotten, however, as its data-driven predictions may capture patterns or nonlinear interactions not identified by our linear model.

In conclusion, we recommend that wine industry leaders utilize both of our models when manufacturing and inspecting their wines in order to maximize their quality.

Bibliography

n.d. *Vinho Verde*. Viticulture Commission of the Vinho Verde region (CVRVV). https://portal.vinhoverde.pt/pt/regulamento-do-vinho-verde.

Bache, Stefan Milton, and Hadley Wickham. 2022. *Magrittr: A Forward-Pipe Operator for r*. https://magrittr.tidyverse.org.

Cortez, Cerdeira, Paulo, and J. Reis. 2009. "Wine Quality." UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

Kassambara, Alboukadel. 2023. *Ggpubr: Ggplot2 Based Publication Ready Plots*. https://rpkgs.datanovia.com/ggpubr/.

Komsta, Lukasz, and Frederick Novomestky. 2022. *Moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. https://www.r-project.org.

Kuhn, Max. 2022. *Caret: Classification and Regression Training*. https://github.com/topepo/caret/.

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

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with r*. New York: Springer. http://lmdvr.r-forge.r-project.org.

———. 2023. *Lattice: Trellis Graphics for r*. https://lattice.r-forge.r-project.org/.

Wei, Taiyun, and Viliam Simko. 2021a. *Corrplot: Visualization of a Correlation Matrix*. https://github.com/taiyun/corrplot.

———. 2021b. *R Package 'Corrplot': Visualization of a Correlation Matrix*. https://github.com/taiyun/corrplot.

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

Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*. https://readxl.tidyverse.org.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wilke, Claus O. 2020. *Cowplot: Streamlined Plot Theme and Plot Annotations for Ggplot2*. https://wilkelab.org/cowplot/.