The Fine Art of Wine: Predicting Quality through Data Science

Mehtab Chhina

Department of Statistics, Simon Fraser University

STAT 310/311: Introduction to Data Science for the Social Science

Wei (Becky) Lin

December 12, 2023

Introduction

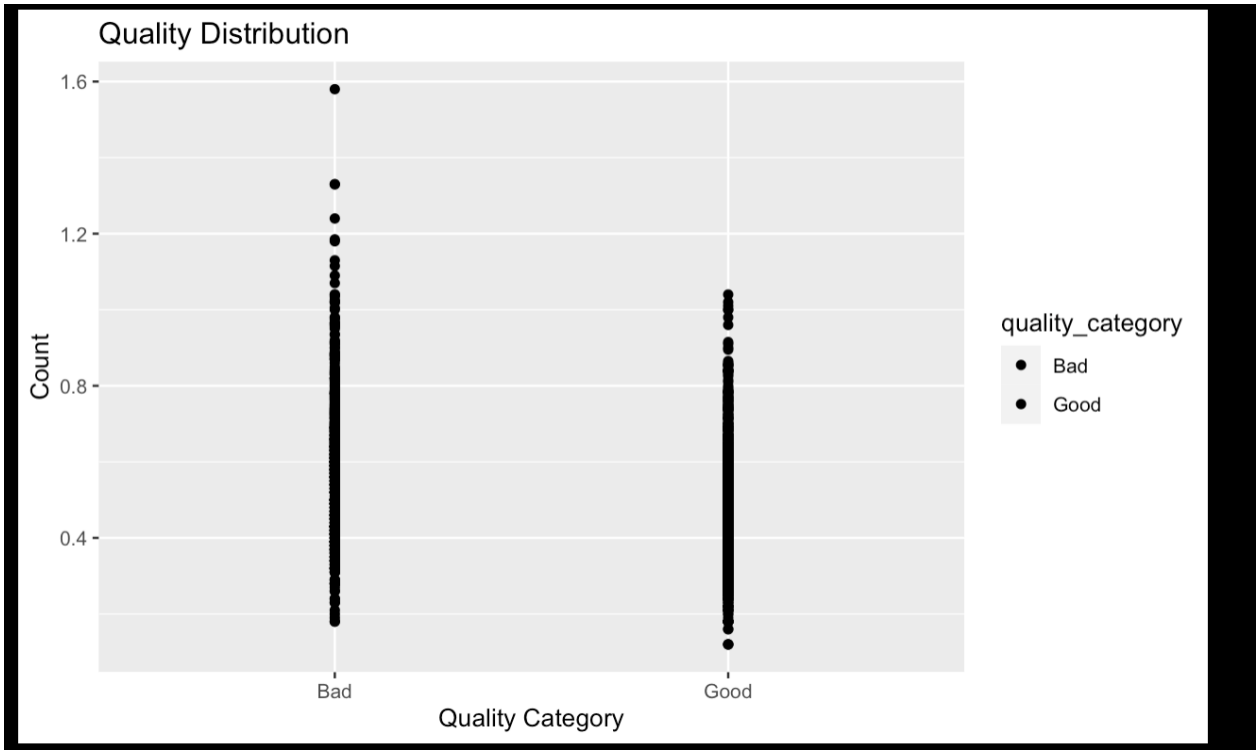
Who doesn't like relaxing with a good glass of wine on a sunny day? Enjoying wine with a great meal is one thing, but have you ever thought about what makes wine "good"? Now, with all the wine information from around the world and data science tools, we can use numbers and stats to figure out if a wine is as good as we think. Wine tasting is a fun activity, but using stats to predict the best wine is something new. Deep dives into wine research have shown us which parts of the wine are important for its quality. In my study, I used tools like Random Forest Regression and Multiple Linear Regression to see if there's a link between how good a wine is (dependent variable) and how much alcohol it has (independent or predictor variable). This research adds to the conversation about whether data science methods are good at picking out fine wines. It goes into detail about how we got our data, the steps we took in our research, and why we chose certain methods to get to the bottom of our main question. We also looked at different models during the first part of our research and then talked about what we found. In the end, we discussed what the study didn't cover and offered some ideas to improve the research.

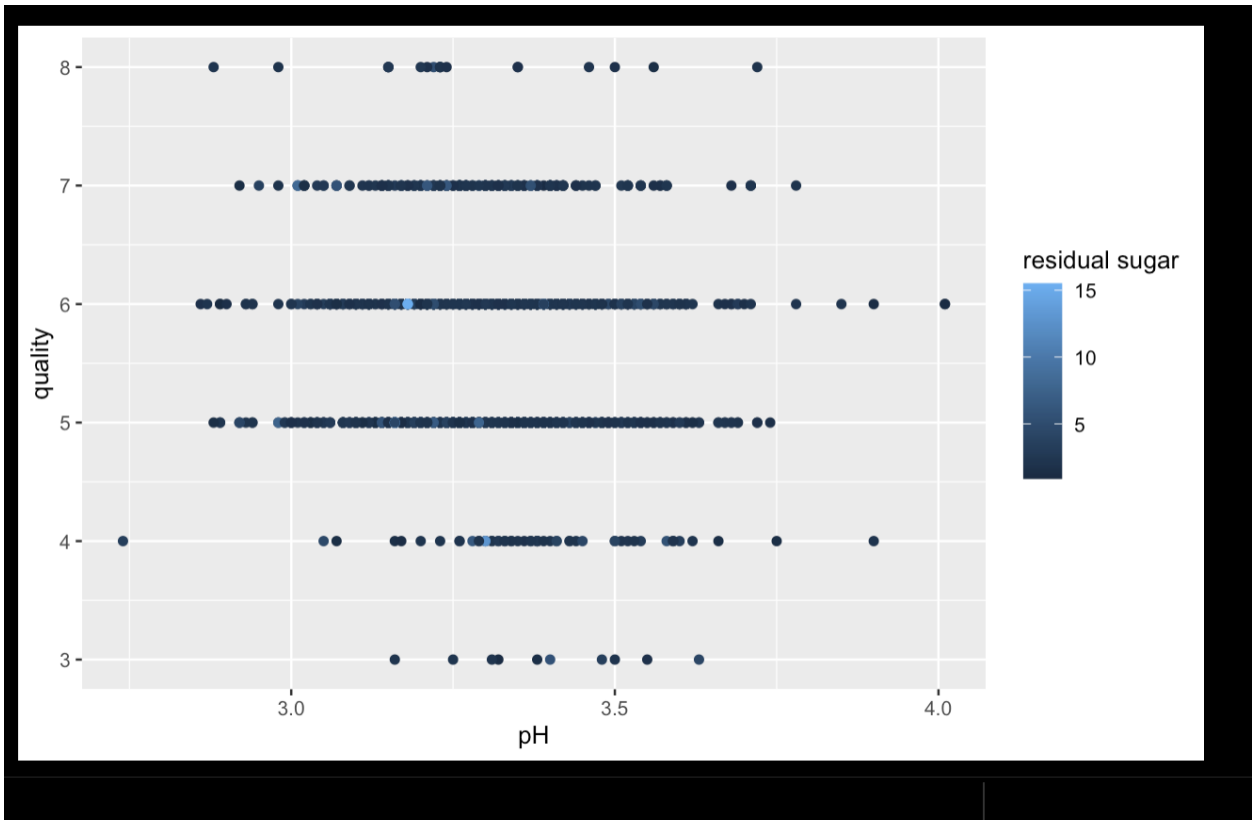
Contribution

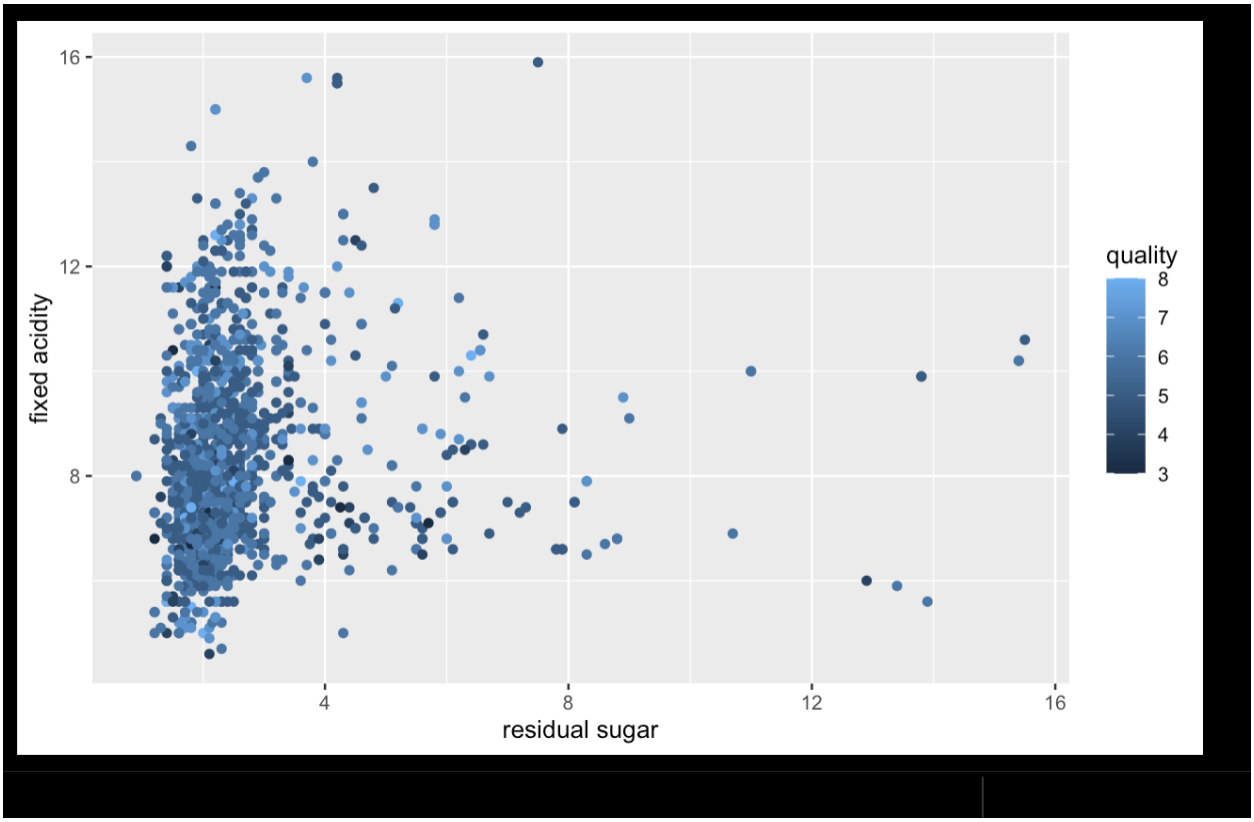
The wine industry is increasingly using data science for quality checks. Previously, wine quality was judged by expert tasters, relying on taste, smell, and appearance. These methods, though traditional, weren't always consistent or widely applicable. The introduction of data science brought a big change. Techniques like Random Forest and Linear Regression started using the chemical makeup of wine, like acidity and alcohol levels, to predict its quality. At first, some people were skeptical. They thought the complex art of wine tasting couldn't be fully understood through data alone, as the human experience and senses seemed irreplaceable. However, it turned out that data science methods were quite good at predicting wine quality, shaking up the old way of relying only on expert opinions. These methods are consistent, unbiased, and can handle large amounts of data, which is useful for both winemakers and buyers. This research adds to the debate by showing that combining data science with traditional tasting creates a better way to understand wine quality. It proves that using both scientific analysis and expert opinions gives a more complete view. This approach enhances how we appreciate wine, blending new methods with old traditions.

Data and Methods

I used a wine dataset from Kaggle for my study, with 12 different things tested in wines, like acidity and sugar. Getting a lot of good data on wine quality is hard, and this set has more not-so-great wines than good ones. When I looked at the data closely, I noticed that the better wines didn't have a lot of sugar, and most of them were acidic, with a pH value of 3.5 or less. Also, the acid levels in the average and the better wines were usually between 4 to 12. To figure out if the amount of alcohol in wine tells us something about how good the wine is, I used two ways of analyzing the data: Multiple Linear Regression and Random Forest Regression. These methods helped me see if there's a clear link between alcohol and wine quality. They're good because they mix classic stats with newer data analysis techniques that can find more complicated patterns.


Quality Distribution
1.6
1.2
0.8
0.4
Count
Bad
Good
Quality Category
quality_category
Bad
Good

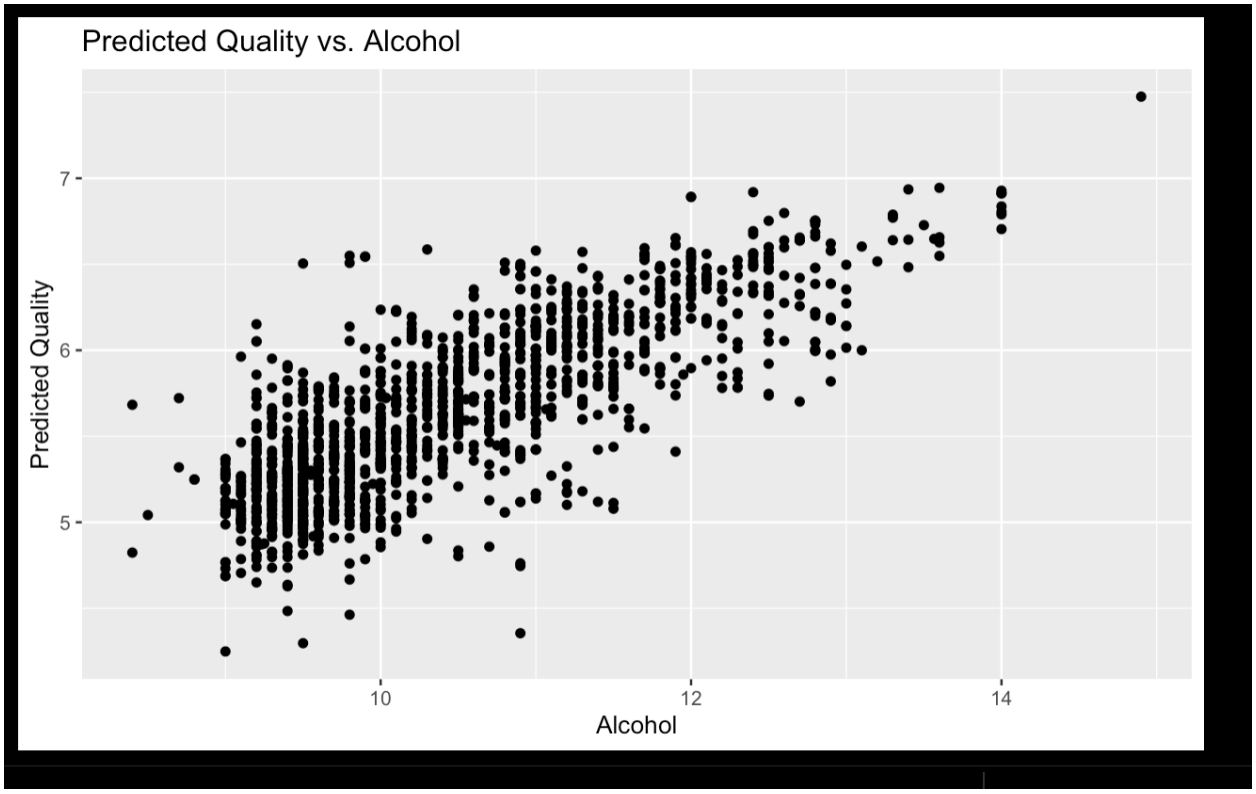

8
7
6
5
4
3
quality
3.0
3.5
4.0
pH
residual sugar
15
10
5

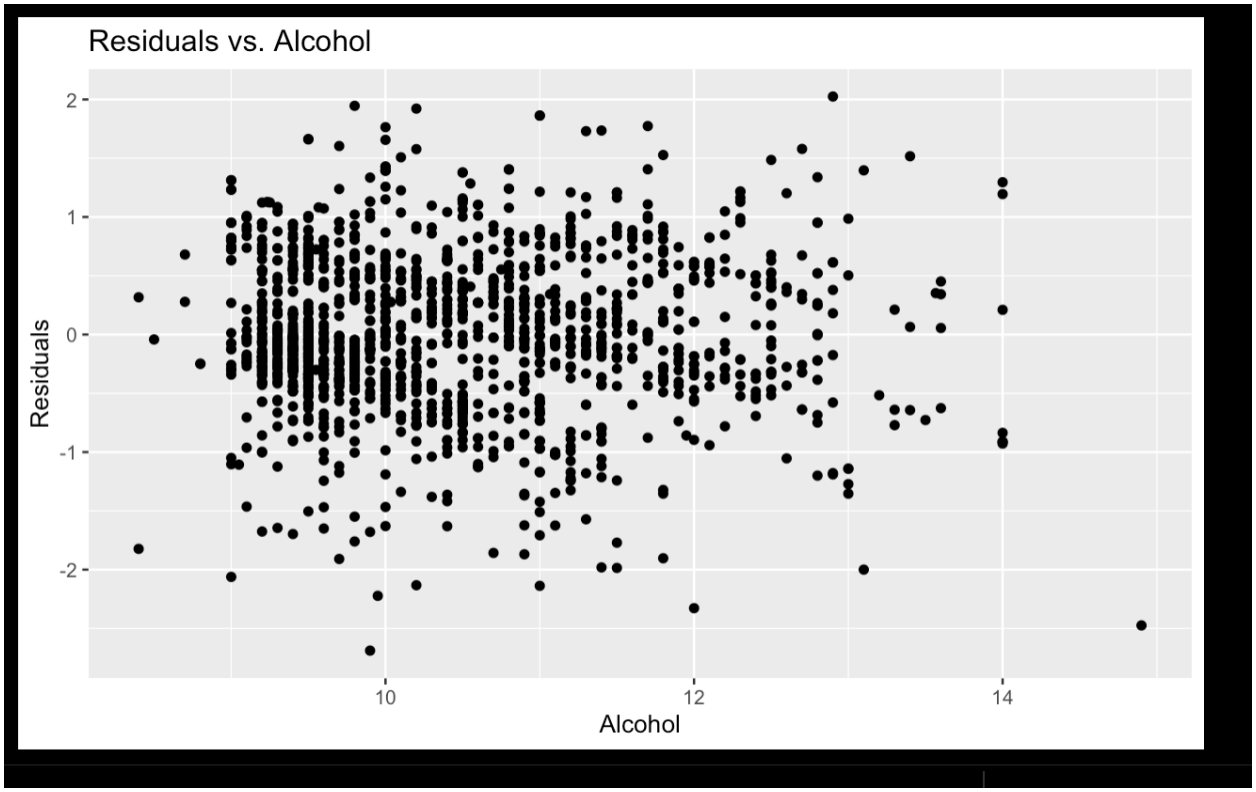


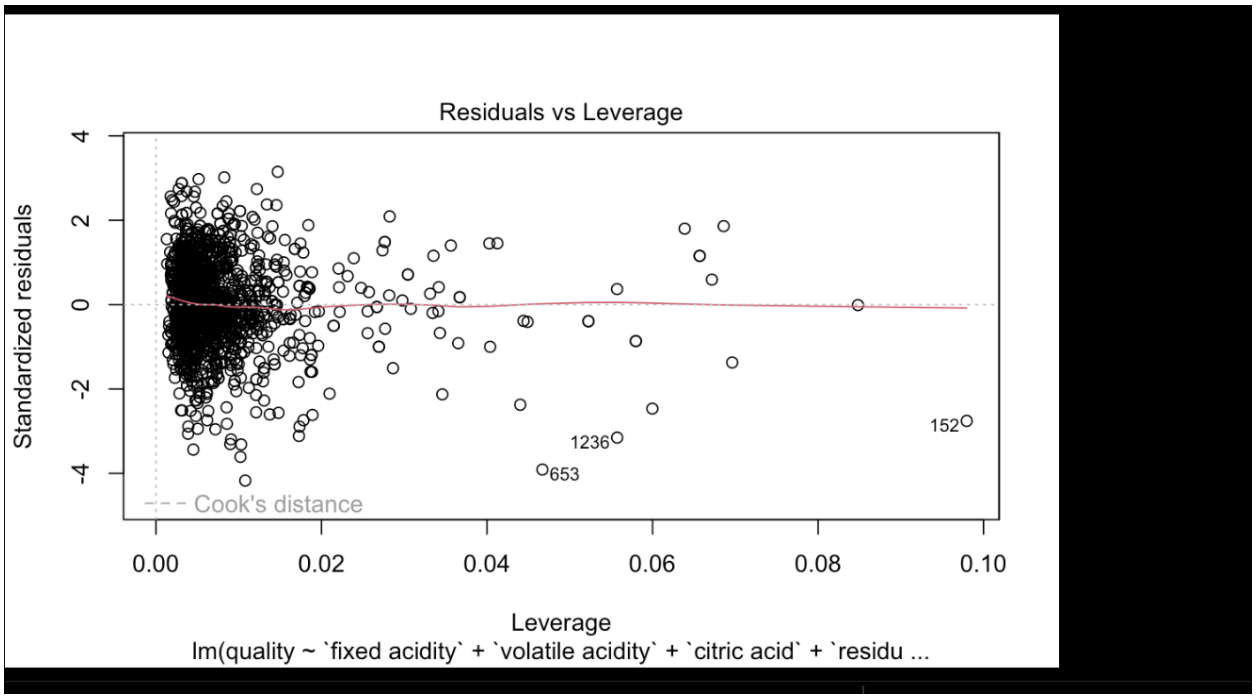

```
fixed acidity   volatile acidity  citric acid    residual sugar
Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
Mean   : 8.32   Mean   :0.5278   Mean   :0.271   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
Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
Median :0.07900   Median :14.00       Median : 38.00
Mean   :0.08747   Mean   :15.87       Mean   : 46.47
3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
Max.   :0.61100   Max.   :72.00       Max.   :289.00
   density             pH          sulphates         alcohol
Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
   quality      quality_category
Min.   :3.000   Length:1599
1st Qu.:5.000   Class :character
Median :6.000   Mode  :character
Mean   :5.636
3rd Qu.:6.000
Max.   :8.000
```

Results

From my tests, I found that red wines with more alcohol often got higher quality ratings from the model. Most of the red wines in the study had alcohol levels between 0 to 11 percent. This is where we got most of our information and based our predictions. The data from Kaggle mentioned there weren't many samples of top-notch red wine—it was mostly just average. When I looked at the data myself, it backed this up, showing that we had a lot more average or lower-quality red wines than the great ones. The quality scores were usually 6 or less. For red wines with an alcohol percentage over 11, our model's predictions might not be as solid because we don't have enough data for these wines or all the factors that could influence their quality. To get our model to be better at predicting high-quality red wine, we'd need more data on those high-quality wines and to look at more factors that can affect wine quality. The dataset focused on "Vinho Verde" from Portugal, so including different kinds of red wine could help make our results stronger and less biased. Right now, the amount of alcohol only explains about 15% of what makes a red wine good according to our model. That means there's a lot more to good wine than just alcohol, and we'd need to consider other things too to figure out what makes some red wines better than others.

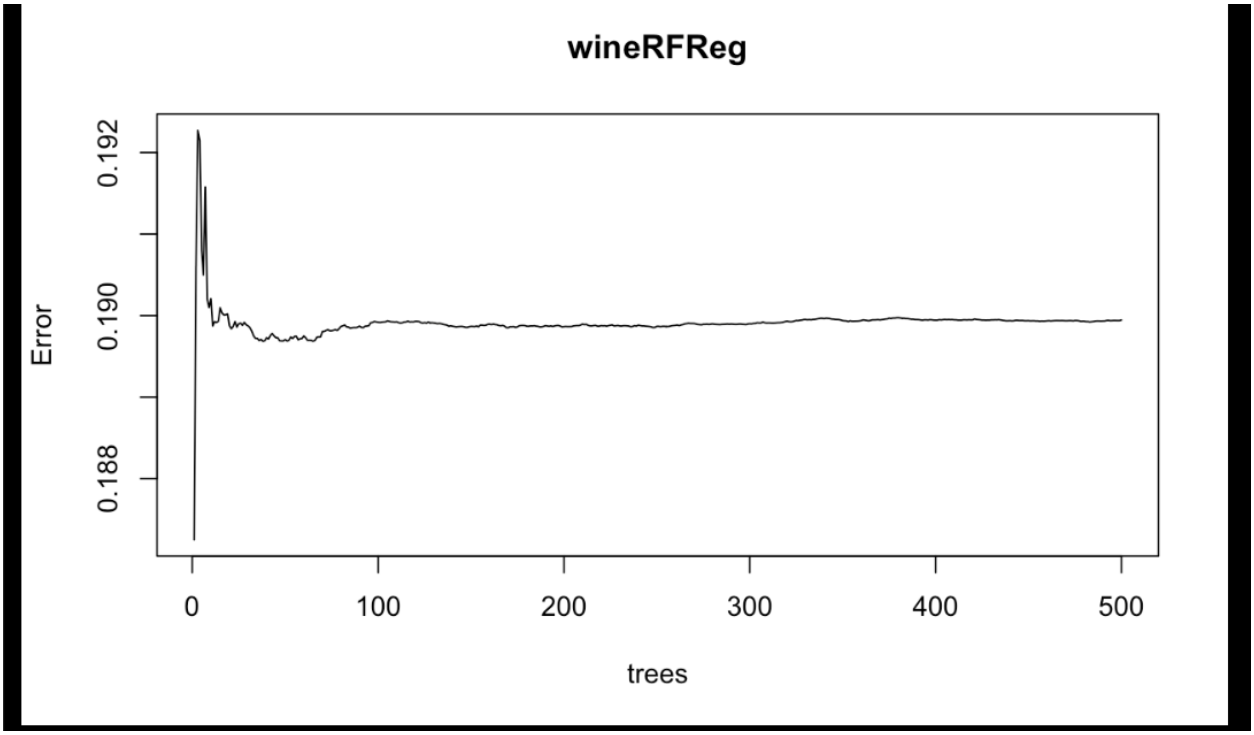



Residuals vs. Alcohol
Residuals
Alcohol
2
1
0
-1
-2
10
12
14


Residuals vs Leverage
Standardized residuals
Leverage
lm(quality ~ `fixed acidity` + `volatile acidity` + `citric acid` + `residu ...
Cook's distance
1236
653
152
4
2
0
-2
-4
0.00
0.02
0.04
0.06
0.08
0.10

```
Call:
 randomForest(formula = quality ~ alcohol, data = train, importance = TRUE,      ntrees = 500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 0.1899448
                    % Var explained: 15.07
```



Limitations

Our study was limited because it only looked at one type of Portuguese red wine, so it doesn't represent all red wines. Also, we found that alcohol level wasn't the only thing that makes a wine good, but our data didn't have many high-quality wines to confirm this. Our models, Multiple Linear Regression and Random Forest Regression might miss other qualities that contribute to a wine's greatness. More data on a variety of wines and improved methods would help us get a complete understanding.

Future Research

In the future, we need to study more kinds of red wines, especially the good ones, to see if what we learned about alcohol and quality is true for all wines. Looking at things like the type of grapes and how

the wine is made could tell us more about what makes a great wine. Using better ways to analyze the data might help us understand more about what makes wine taste so good.

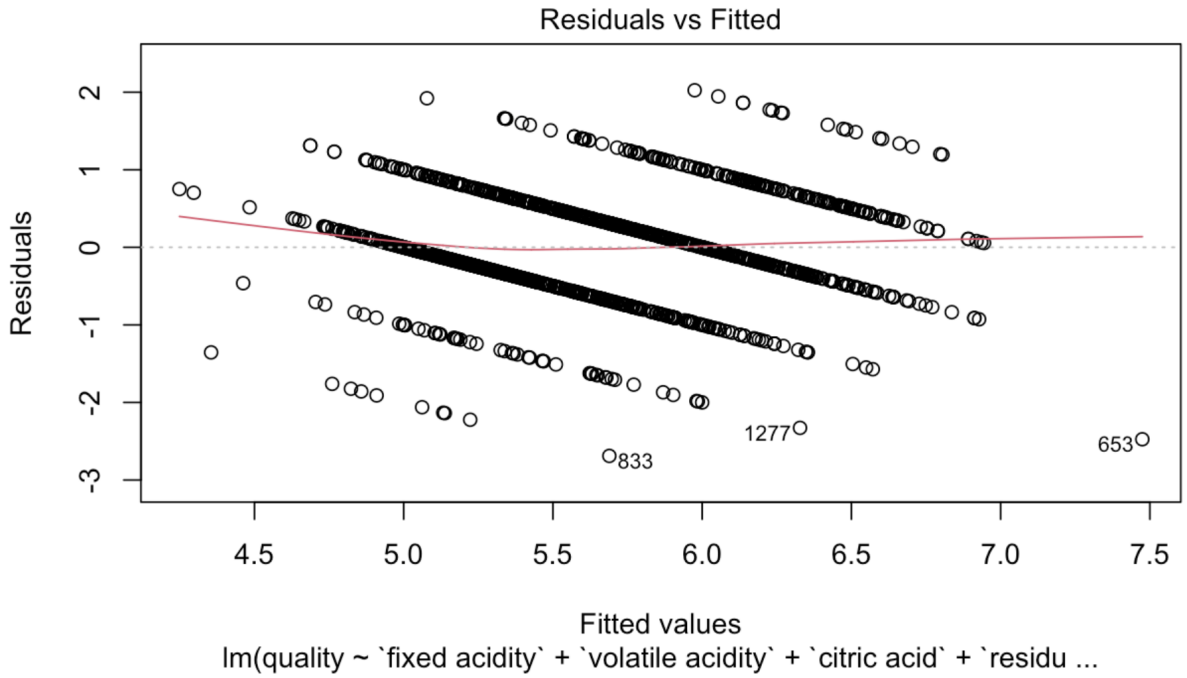
Conclusion

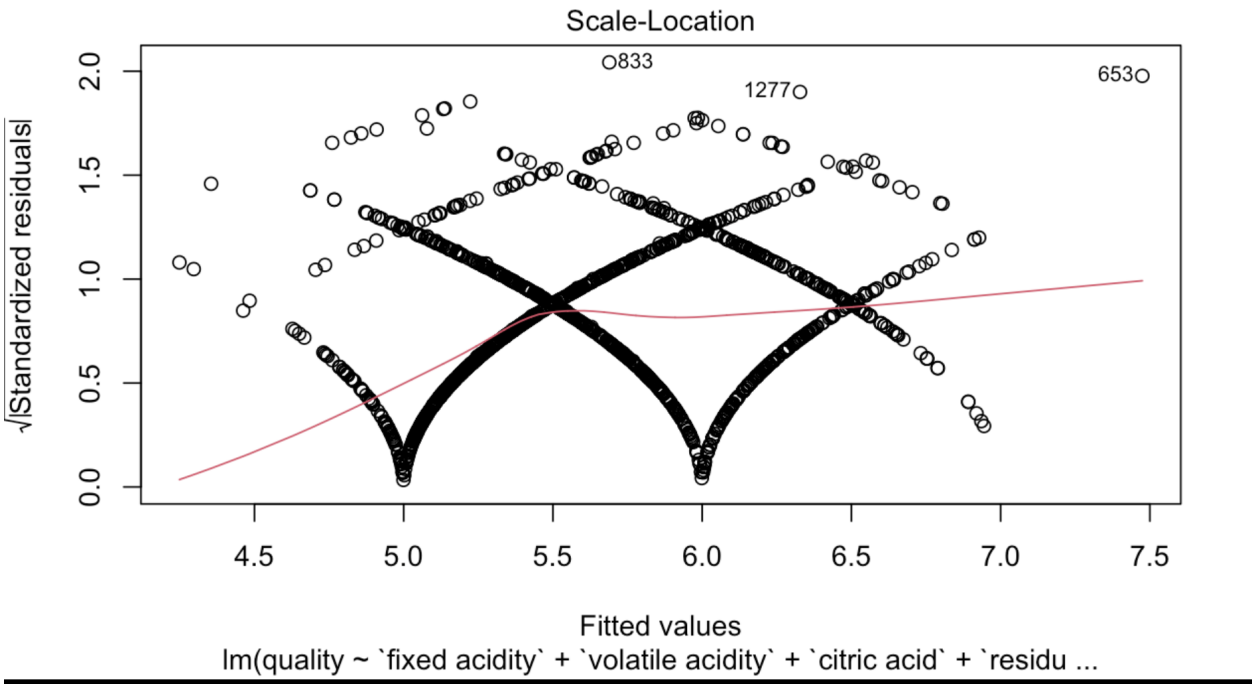
Our study looked at how much alcohol is in red wine to see if it makes wine better. Using data tools, we learned that wines with more alcohol are often seen as better, but we mostly had average wines to study, not the top ones. Alcohol only tells us a little about wine quality, there's more to discover. The data we had was all from one type of Portuguese red wine, so we don't know if this is true for all red wines. We need to look at more types of red wines and other details that make wine good. Combining what data scientists and wine experts know could give us a better way to tell if wine is good. As we get better at gathering and understanding wine data, we'll get better at knowing what makes a wine stand out.

References

Data retrieved from:

https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/data


Residuals vs Fitted
Residuals
2
1
0
-1
-2
-3
4.5
5.0
5.5
6.0
6.5
7.0
7.5
833
1277
653
Fitted values
lm(quality ~ `fixed acidity` + `volatile acidity` + `citric acid` + `residu ...


Scale-Location
833
1277
653
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
Fitted values
lm(quality ~ `fixed acidity` + `volatile acidity` + `citric acid` + `residu ...