CS614 Final

Jacob Gilbreath

Introduction

Wine is one of the most popular and oldest alcoholic beverages in the world, and even more popular in the US compared to the rest of the world. The sheer quantity of wines available and the vast price range may make it difficult to decide on what is best, but what is it that makes these wines so different and why is it that two bottles of wine that may look one in the same are on two different levels of quality. For this project, I will be analyzing the quality of wines based on its physiochemical properties.

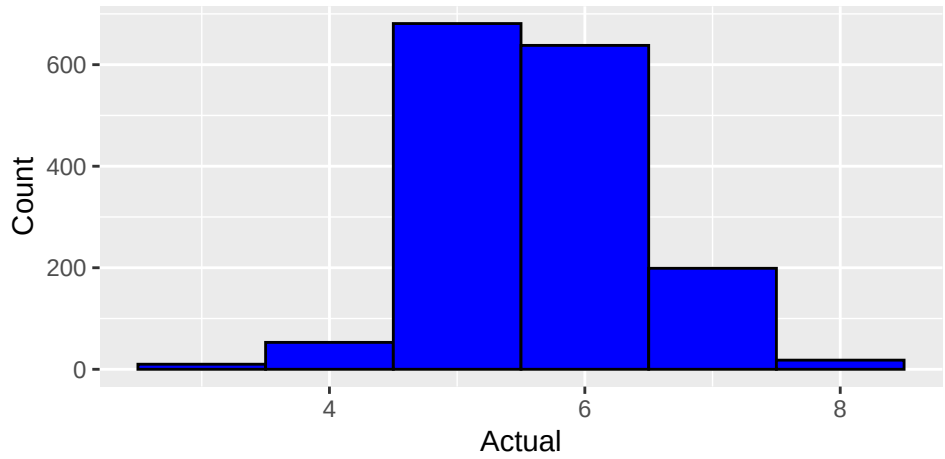
For this analysis, I have used the Wine Quality data set that takes physiochemical properties of 1,599 different wines from the Northern region of portugal. The data set includes 11 different physiochemical attributes, including PH, sulfur dioxide, volatile acidity, fixed acidity, citric acids, residual sugar, chlorides, free and total sulfur dioxides, density, sulphates and alcohol. Many of these attributes do not affect the taste of the wine, but are still important metrics in wine since they contribute to the anti-microbial properties and the prevention of oxidation of the wine which causes a brown hue and poor odor. Many of these values have little variation in them, such as density, where others have high variation, such as alcohol (measured in ABV). These attributes are used as our input variables for my analyses and a twelfth output variable, quality, is what I would like to predict. The quality variable is measured independently of the other variables, which are measured using scientific tests after the wine is made, and is an average of three scores on the quality of the wine. The three scores are based on the taste and smell of the wine which are averaged and rounded to the nearest integer to get the quality score that is used in the provided data set.

The first question I will answer is which of these physiochemical properties contributes the most to the overall quality of wine and to what degree. Since many consumers have their own wine preferences and know what they like in a wine, these results will generally not interest them. However, a wine maker, or vintner, may reference these results as a guide to making a hgiher quality wine with specific properties.

Next, I will take a look at the wines that are "outliers" with respect to their overall quality and test which properties determine whether a wine is poor or great. This result can be useful for consumers as it can be used to tell whether or not you are over-paying for a "poor" wine or if you are getting a great deal on a "great" wine. For vintners or distributors this result can be equally important when deciding how to make and what most to contribute to the wine in order to make it a high quality wine.

Methods

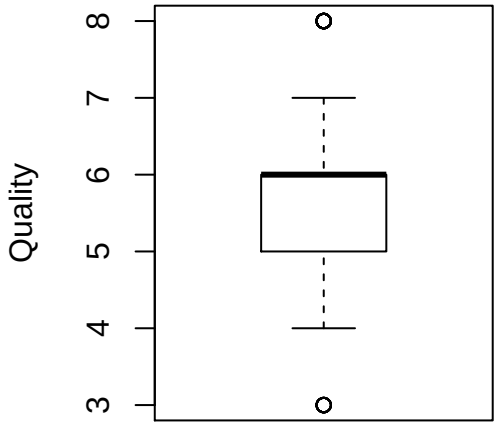
In order to get a first look at the data and see the trend of the quality variable I first designed a histogram:



The histogram gave me a general idea of where all of the quality scores were landing and where I should expect my predictions to end up as well. We can pretty clearly see that the vast majority of wines fell in the 5 or 6 category with a slight left-skewdness, meaning there were more higher quality wines that lower quality wines.

After noticing the trend, I ran a linear regression model that predicted quality for each of the 11 predictor variables and took note of each R^2 value as well as the associated p-values. After observing how the individual variables predicted quality, I took a look at the combinations of several predictor variables as additional parameters in the linear regression model. This method was a lot of guess and checking and I slowly made progress in maximizing the R^2 while keeping the p-values below the desired 0.05. This method proved long and arduous, but still yielded the results I was looking for.

In order to determine the differences between "great" and "poor" wines, I first had to determine what I would consider a great wine versus a poor wine. To do this I did some outlier identification on the quality variable to see if any values would be considered outliers. By just using the base R boxplot function, it was able to identify the wines rated as an 8 or a 3 would be outliers.



Once the outliers were determined, I could define the wines that scored an 8 as great and those that scored a 3 as poor. The next step was to extract the rows of my dataset that had wine scores of a 3 or 8. Once I had this subset of my data, I created more linear regression models with each of the predictor attributes individually. Once I identified those with the greatest correlation with the quality, I found the best combination of variables that yielded the largest R^2 value with minimal p-values.

As you mentioned, I took a look at a logistic regression model to predict probability that each wine would be "great" or "poor," but since the linear regression had worked almost perfectly in classifying the wines, I kept it as it was.

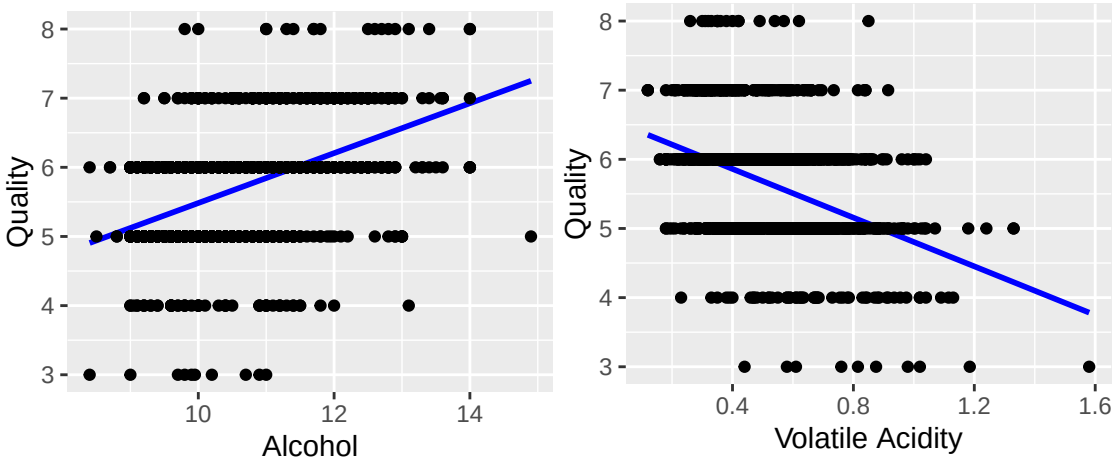
Results

Research Question 1

The individual linear regression models showed that the alcohol percentage (ABV) of the wines had the greatest correlation with quality. The next highest correlation came from volatile acidity. Although both of these showed very low correlation, they still held the highest, by a consiberable margin, compared to the rest. Where the other variables had R^2 values that were less than 0.01, meaning less than 1% of the data could be explained with the model, alcohol's value fell around 0.2263 and volatile acidity fell around 0.152. Again, although these are not very significant in terms of predicting the quality, they still do quite better than the other variables.

The model showed a positive correlation between the quality and the alcohol level of the wine. Whether this is an accurate predictor of the actual quality of the wine I will leave up to you, but I believe it can be interpreted one of three ways and perhaps a combination of the three. The first is that the wine raters that gave it a quality score were slightly inebriated after trying the wine and confused that feeling with an actually delicious wine. Or the second would be that the alcohol level actually makes the wine taste or smell better compared to the other ingredients in the wines, and thus earned it a higher score. The last interpretation would be that the testers knew the alcohol level of the wine before rating and thus held a subconcious bias to give it a higher score. The answer to these interpretations depends largely on how the wine testers were given the wine to be tested as well as the testers themselves.

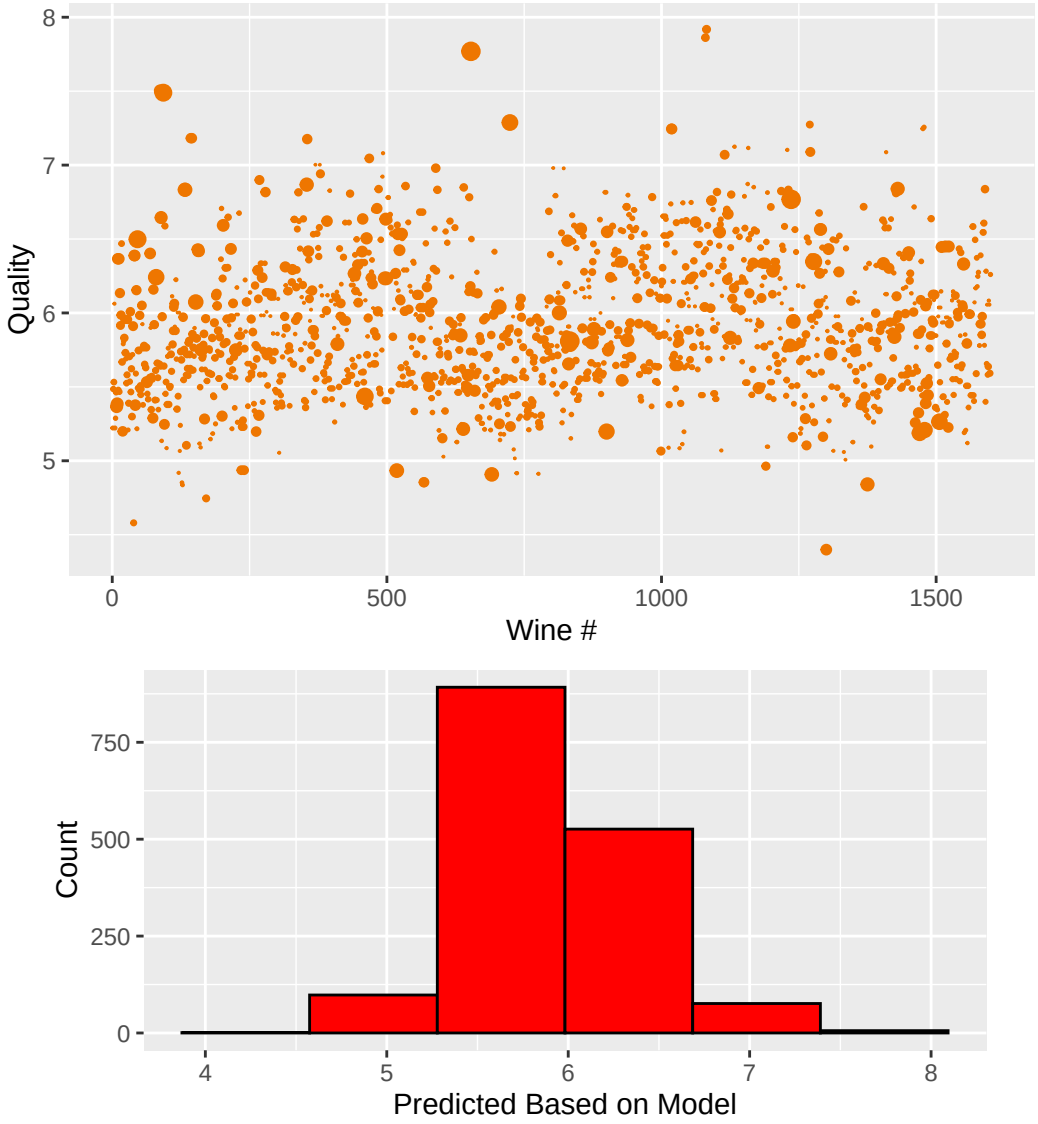
With regards to volatile acidity, the model predicted a negative correlation with quality. Since volatile acidity, gaseuous acidity in the form of acetic acid, contributes to the smell and taste of vinegar, this would make sense. As most people dislike vinegar with regards to smell and taste, it would be only natural that the quality of the wine would decrease as this feature became more and more present.



ALthough the slight correlation is present in the figures, it is also easy to tell that these are not strong predictors of the overall quality of the wine. Therefore, I decided to build the model with multiple variables to get the most accurate model to predict wine quality. Despite my best efforts, the largest R^2 value I could come up with was 0.3567, which could only account for about 36% of the data, but still better than any of the individual scores. This model included volatile acidity, chlorides, total sulfur dioxide, free sulfur dioxide, PH, sulphates, and alcohol. Alcohol, sulphates, and free sulfur dioxide all contributed positively to the quality of the wine, whereas the rest contributed negatively. This model yielded a residual standard error of 0.6477 which is relatively large when considering this is almost an entire quality "grade" difference.

The predicted quality values for each wine are given in the top figure below with a size correlating to the error from the actual. Larger points indicate that a predicted quality was much farther than their actual quality than a smaller point. The figure on the bottom shows a histogram of the predicted values. When this figure is compared to the histogram of the actual qualities as above, we can see the predicted values lose quite a bit of the range they had, which wasn't much to begin with. The predicted values In fact, the

minimum of the predicted values is 4.24, an entire 1.24 away from the actual minimum. When looking at this value in the top figure, we can actually see that this point is quite large and is therefore quite far from the actual quality of that specific wine.



Research Question 2

When removing the median data and only being left with the "great" and "poor" wines as defined above, the individual regressions showed a strong correlation between volatile acidity, sulphates, and alcohol in terms of their R^2 values, but had questionable associated p-values, save alcohol. However, after combining all the predictor variables into one linear regression model, I saw that all but two variables had p-values less than 0.05, PH with a p-value of 0.025 and alcohol with 0.0075. When I ran the linear regression with just the two predictor variables, it yielded an R^2 of 0.7506 and extremely significant p-values (*See Below.*). This seemed to be the best fitted and most accurate difference between "great" and "poor" wines.

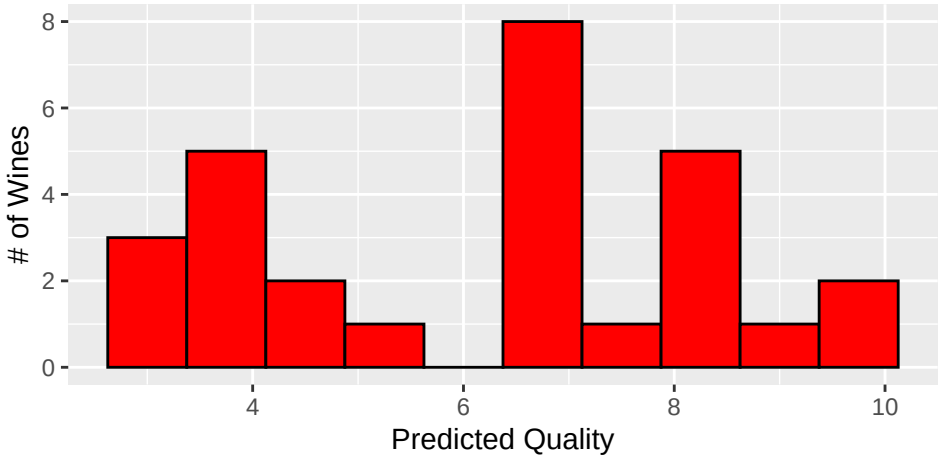
```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  14.260273  4.1478696  3.437975 2.061972e-03
## both$pH      -7.091001  1.2743833 -5.564261 8.722713e-06
## both$alcohol  1.363866  0.1617266  8.433156 8.866068e-09
```

ALthough this model was worth exploring, I decided to shift the values a bit to make interpretation easier. This included subtracting the minimum of PH and alcohol from their respective variables. The model below is the same as above, except for where we can see that the intercept has shifted from 14.26 all the way down to 5.29.

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  5.294661  0.6682470  7.923209 2.800227e-08
## alc          1.363866  0.1617266  8.433156 8.866068e-09
## ph          -7.091001  1.2743833 -5.564261 8.722713e-06
```

The model here estimates that if PH and ABV are at their minimum, 2.88 and 8.4% respectively, then the wine will receive a quality rating of 5.29, or 5 after rounding. With each percentage increase in ABV, it is predicted that the score will increase by 1.36 which suggests positive correlation yet again between alcohol level and wine quality. The PH level, however, shows quite a large negative correlation, suggesting that for each increase of 1 to the PH level from the minimum will cause the quality to decrease by 7 points. The range of PH values for this subset of the data is only 0.84, and therefore the maximum this will decrease the overall quality is 5.88 points.

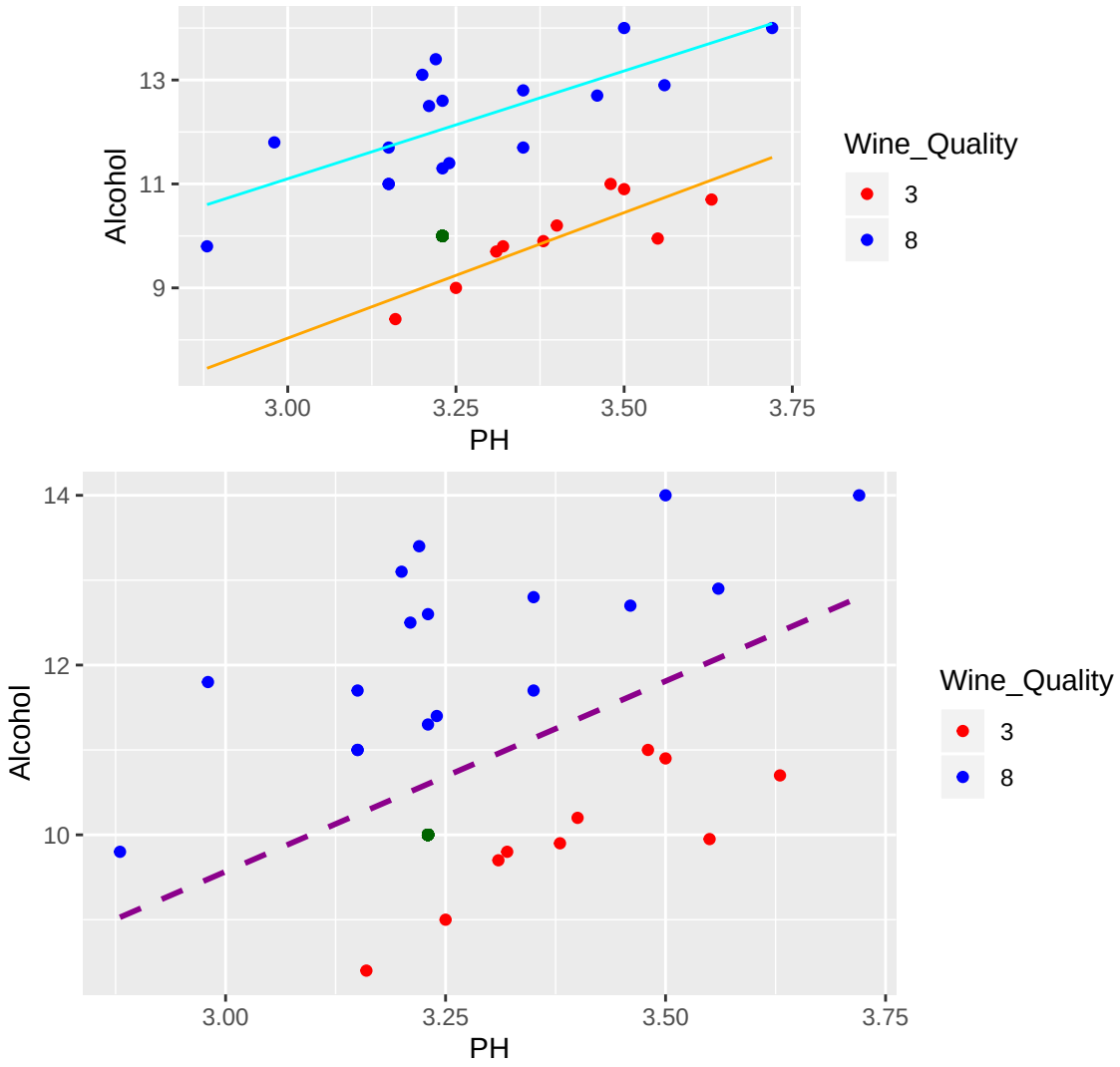
When I found the predicted values, I of course didn't find very many exact 8 or 3 quality wines, but I did notice a bimodality in the distribution of the scores as seen in the figure below. Notice there are no values that fall at or around 6. With this in mind, I took all values greater than 6 and classified those as our "great" wines and all values than 6 as "poor" wines.



To check the accuracy of my predictions based on the model, I counted the number of "poor" and "great" wines and found that there was 1 great wine that got mis-classified as a poor wine. To determine which value was mis-classified, I just needed to find which wine had an absolute error between predicted and actual quality that was greater than 2.

The graph below is a plot of Alcohol vs. PH with a color indicating whether it was rated an 8 or a 3. The green dot is the mis-classified point and the lines are the linear regression lines for each of the ratings of wine, respectively. The graph seems to have a consiberable gap between the great and the poor wines, so I wanted to try to get a somewhat accurate measurement of what this line could be.

To estimate what this dividing line could be, I simply took the average of the intercept and the slope of the individual regression lines and formed a new line that fell between the 2 custers.This gave an estimated intercept of -3.898 and a slope of 4.488.



With the addition of this line, there is a clear difference between the "great" and the "poor" wines, save the one outlier. However, we can see that the one mis-classified point does seem to be rather close to the predicted "quality line." From this, we can assume that in order for a wine to be great, it must maintain an alcohol level that sufficiently washes away the acidity. As mentioned earlier, the wines with higher PH were losing quality as it increased from the minimum to a maximum of five points. However, in the figure we can see the point that is losing 5 quality points is actually still considered a great wine due to the amount of alcohol.

Conclusion

Through the research and exploration of both research questions, I determined that wine quality is mainly a product of how much alcohol is in the wine with some slight influence coming from acidity. However, as we saw, alcohol trumps acidity in terms of greatness. This result could be skewed based on who was testing the wine or if it was even the same person. I think that how this quality data was collected is a major part of the interpretation of these results. Were these blind taste tests done by random people, or was it the same three testers over a period in time. In either case, I believe some sort of bias occurs when determining wine quality.

I think that in the future, wines from more than one vineyard should be used in order to get a more general solution for wines. I think the use of just a single vineyards wine very much skews the results as certain values will inevitably be nearer to each other since it may be the same wine, but in different batches. Future

tests could remove some variables that have little to no variety, like density which falls between 1.08 and 1.09, since it will prove extremely difficult to discern anything significant from this.

Overall, as a consumer, I would recommend the purchase of the wine with the highest alcohol by volume and the lowest price since that is what determines both the quality and greatness of a wine.