Modeling Wine Quality

Jason Witry, Jerome Doe, Armand Heydarian, Hang Zhao

March 20, 2019

Contents

Chapter 1: Introduction

While the earliest evidence of wine is debatable topic, its consumption has been and is still enjoyed by people all over the globe (OIV, 2018). Wine Prices in today's market range from a $6 Gazela Vinho Verde bottle to a $7,447 Domaine Leroy Chambertin Grand Cru bottle. One integral part of wine quality is how it tastes, and one may argue this positively correlates to its cost in most cases. However, there is subjectivity associated with taste. We all have different preferences, palates, hormones that strike our taste buds differently, and even some biases when it comes to taste.

This brings up the question of what method is currently available to ensure wine quality? Traditionally, knowledgeable and trained wine professionals known as sommeliers taste and rate wine (Ebeler, 1999). Some of the things they look for are the dryness, sweetness, tears, color, and clarity of a wine. Although sommeliers are great at their job, their services can be cost intensive, time consuming and the sample is often too small. With the potential exception of small sample sizes, these issues can be mitigated by applying data science techniques to determine from a physicochemical perspective why sommeliers favor certain wines over others. The goal of this work is to answer two questions: What properties of wine are the best predictors of wine quality (according to wine certification experts) and what prediction accuracy can we obtain by modeling wine quality with these features in a linear regression?

An alternative data mining approach was used to calculate wine taste preferences using physicochemical properties by the Department of information Systems/R&D Center Algoritmi, University of Minho and Viticulture Commission of the Vinho Verde Region (Cortez et al., 2009). Their analysis covered a fairly large dataset to include data on red and white vinho verde wines. In our effort to identify physical wine properties that relate or predict wine quality, we will be focusing only on the white vinho verde wine data. Our derived variables, which prove to be the strongest predictor of quality, will then be compared to those presented in

the conclusion of the work by the Department of information Systems/R&D Center Algoritmi, University of Minho and Viticulture Commission of the Vinho Verde Region. We will use their predictive accuracy of 88% (Cortez et al., 2009) as a benchmark for our own linear regression.

Chapter 2: Data Description, Variable Description and Exploratory Data Analysis

Data Description

The data used for our investigation contains information on 4,898 white wines from the Minho region of Portugal, dubbed vinho verde. The wine samples were collected from May 2004 to February 2007 by a computerized system (iLab) that automatically handles wine sample testing. Only samples that were tested by the official certification entity, the CVRVV, were used in the data. 11 physicochemical properties of wine were used, listed in the table below. The quality is a number on a scale from 0 (poor) to 10 (excellent) determined by the median rating of at least three sensory assessors.

The table below presents the physicochemical features available in our dataset as well as relevant statistics. While examining the descriptive statistics, the dataset variables are presented with N indicating that there is the same amount of occurrences for each variable, so no data is missing. In the range, there appears to be a heavy skew towards the end of the range, within the 75th to 100th percentile, while it is not as heavily skewed within the 25th percentile. This applies to almost all of our variables.

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
fixed.acidity	4,898	6.855	0.844	4	6.3	7.3	14
volatile.acidity	4,898	0.278	0.101	0.080	0.210	0.320	1.100
citric.acid	4,898	0.334	0.121	0.000	0.270	0.390	1.660
residual.sugar	4,898	6.391	5.072	0.600	1.700	9.900	65.800
chlorides	4,898	0.046	0.022	0.009	0.036	0.050	0.346
free.sulfur.dioxide	4,898	35.308	17.007	2	23	46	289
total.sulfur.dioxide	4,898	138.361	42.498	9	108	167	440
density	4,898	0.994	0.003	0.987	0.992	0.996	1.039
pH	4,898	3.188	0.151	3	3.1	3.3	4
sulphates	4,898	0.490	0.114	0.220	0.410	0.550	1.080
alcohol	4,898	10.514	1.231	8.000	9.500	11.400	14.200
quality	4,898	5.878	0.886	3	5	6	9

Limitations

We see a few limitations to our dataset here; there are only 11 physicochemical properties being studied, while there are many hundreds of chemicals present in wine that may subtly affect the quality. However, these are likely the most significant, as we will see in the Variable Description.

Also, for some of the quality ratings only three certification expert's opinions were used. This is a small sample size that could be a very unreliable prediction of wine quality that may affect the model.

Variable Description

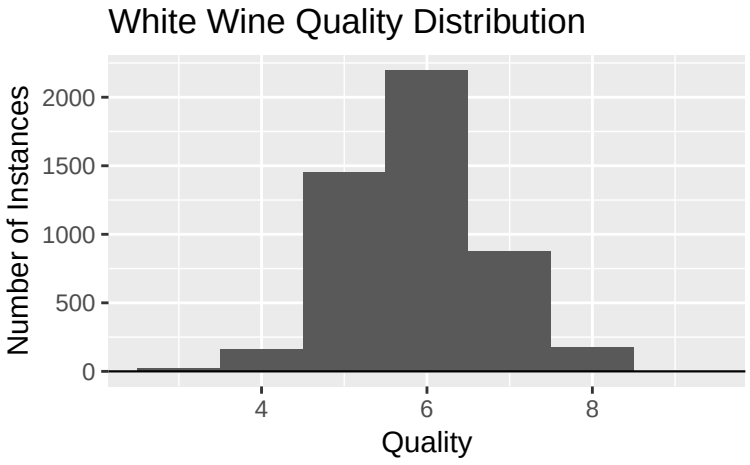
The acids in wine are an important component in both winemaking and the finished product of wine. They have direct influences on the color, balance and taste of the wine as well as the growth of yeast during fermentation and protecting the wine from bacteria. Traditionally total acidity is divided into two groups, the volatile acids and the fixed acids. The citric acid is generally used by winemakers in acidification to boost the wine's total acidity. The strength of acidity is measured according to pH, with most wines having a pH between 2.9 and 3.9. Generally, the lower the pH, the higher the acidity in the wine. (Wikipedia, Acids in Wine)

Among the components influencing how sweet a wine will taste is residual sugar, which typically refers to the sugar remaining after fermentation stops. One rule of thumb in wine tasting is that wines with lower alcohol content will have more residual sugars, because during fermentation yeast converts the sugars to alcohol. The less alcohol is generated by yeast, the more residual sugar there is. This is not a perfect relationship, as a grape with low sugar levels that has all of it's sugar converted to alcohol will produce a wine with low alcohol content and no residual sugar. How sweet a wine will taste is also controlled by the acidity, alcohol levels and chlorides. Sugars and alcohol enhance a wine's sweetness; acids and chlorides counteract it. The density of wine is close to that of water depending on the percent alcohol and sugar content. (Wikipedia, Sweetness in wine)

Sulfur Dioxide (SO_2) is used as an antioxidant and preservative and has become widely used in winemaking. It is present in wine in free and bound states. The total sulfur dioxide is the sum of all the SO_2 in the wine, while free refers to dissolved SO_2 that is not chemically bound to another molecular structure. Excessive amounts of SO_2 can inhibit fermentation and cause undesirable sensory effects. Sulphates are a wine additive which can contribute to sulfur dioxide levels in wine. (Wikipedia, Sulfur dioxide)

Because the physicochemical properties covered in this dataset give the alcohol content of the wine, the Sulphur Dioxide content, the sweetness and acidity, we can say with reasonable certainty that the variables covered here are likely the biggest indicators of the wine quality.

Exploratory Data Analysis

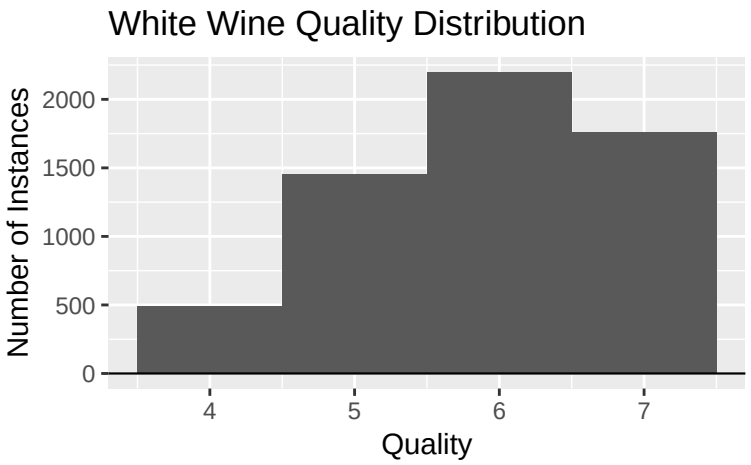


From the above histogram of the quality distribution (our target variable), we can see that it is approximately normal, i.e. with more average ratings than extreme ones. This could lead to issues in trying to fit our model. To mitigate this, we use a simple replication algorithm that takes observations in under-represented classes and simply replicates them a specified number of times. One limitation of this method is that if the under-represented classes are replicates too many times, the model will become severely overfit. Therefore, we adjusted the replicating factor in order to ensure we maintained the normal shape of our data.

Before balancing classes, we note that after testing the linear regression model on the full class set with replication sampling, we discovered that there is simply not enough information to predict the 3, 8 and 9 quality classes with our current methods. Trying to include these qualities decreases the model prediction accuracy significantly, from about 46% to about 30%. Replication oversampling does not provide any new information for fitting, it just ensures that the algorithm sees a similar number of data points for each class to prevent overfitting to over-represented classes. Therefore, we will eliminate the classes 3, 8 and 9 as outliers.

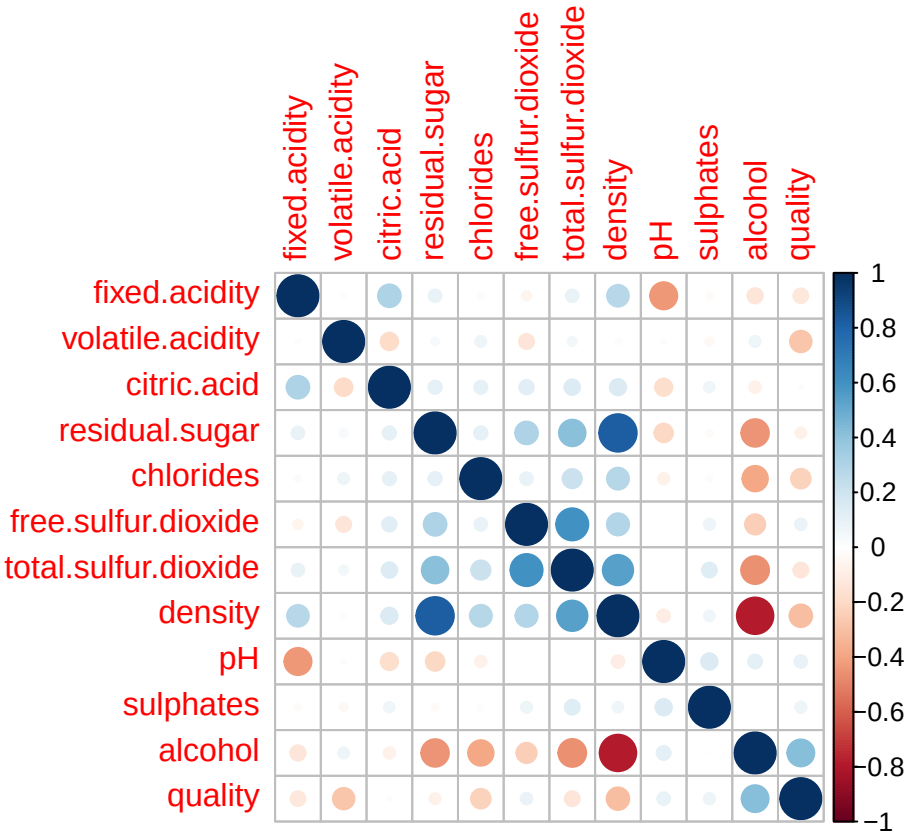
Below is a plot of the quality after the classes have been balanced and outliers removed. We can see that the normal distribution is maintained, but under-represented classes will play a more dominant role in our model.

```
## Warning: Removed 1400 rows containing non-finite values (stat_bin).
```



An initial view into how the variables are related is provided in the correlation plot below. We see that alcohol and density have a strong inverse correlation of -0.80. This makes sense given that alcohol is less dense than wine on average. The residual sugar and density variables are strongly positively correlated at 0.84, and the residual sugar and alcohol content are inversely correlated. These verify our descriptions of the variables above. The total sulfur dioxide and free sulfur dioxide are also fairly strongly correlated, at around 0.61. This makes sense, as both variables measure the same quantity (SO_2) in different states.

We can get a preliminary sense of what variables may be indicators of overall wine quality from this plot as well. Alcohol, for example, has the strongest correlation at 0.42. Density is second strongest at -0.31, followed by volatile acidity (-0.23), chlorides (-0.2), total sulfur dioxide (-0.16), and fixed acidity and residual sugar (-0.10 for both). Constructing a linear model from these variables may provide an indication as to their predictive power. It should be noted that this linear model will not include density, due to the very strong relationship between density and alcohol. It is likely they represent similar information. When it comes down to choosing between these two variables, we prefer alcohol as a predictor because it is a very noticeable quantity in wine that sommeliers look for. Between free sulfur dioxide and total sulfur dioxide, we believe both variables should remain in the dataset for testing. There does not seem to be any physical or sensory reason to prefer one, and the correlation is not nearly as strong as density and alcohol or density and residual sugar.



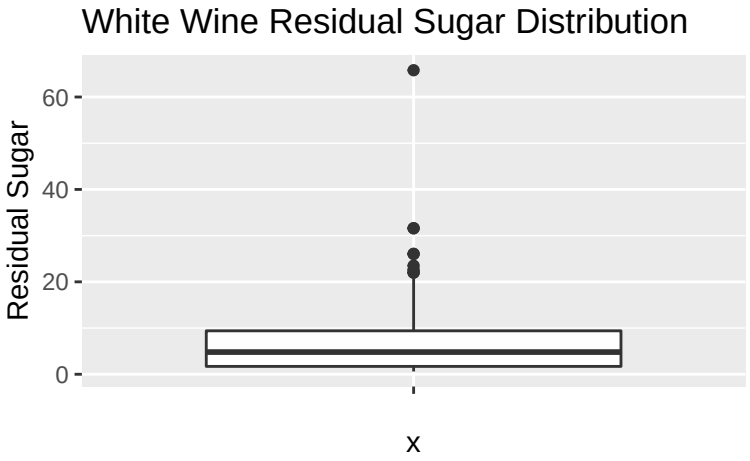
To deal with outliers in our data, we use the OutlierKD function. However, we do not wish to remove all of the outliers from our data and the boxplots below (under Outliers Not Removed) show that some outliers that are $1.5 * IQR$ may not require removal; i.e. they are likely novelty data points. In the residual sugar plot, it appears that a singular data point lies far away from all others. This is an outlier we remove, as opposed to the outliers shown in the box plot below for chlorides. In that box plot, we see all the values above the final quartile are close, and don't deviate far from the body of the plot. These are likely novelty outliers. All of our data fell into one of these two categories, and we removed those outliers that were singular separate values and kept the outliers that were closely clustered and close to the center.

Therefore, we impose a more lenient outlier detection method using the Z-score. Outliers are only removed if they score higher than the Z-score. We use this method for all features with the notable exception of quality, which has already been cleaned.

To further illustrate our outlier removal decisions, below are boxplots for all variables where we removed the outliers outliers were removed, followed by 1 where outliers were not removed.

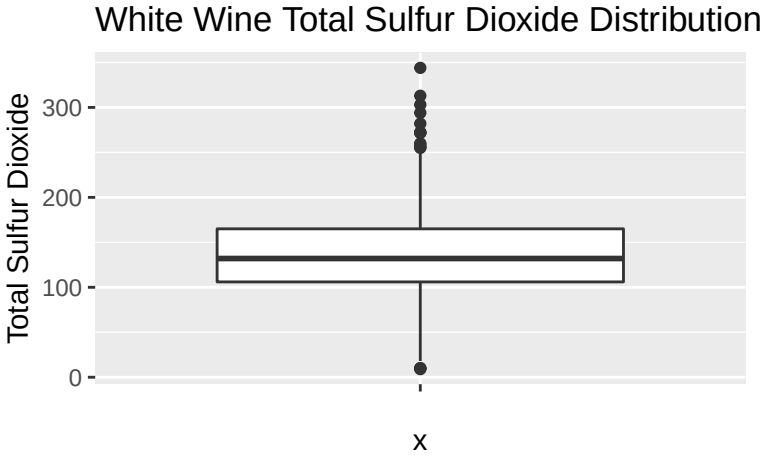
Outliers Removed:

```
## Warning: Removed 1400 rows containing non-finite values (stat_boxplot).
```



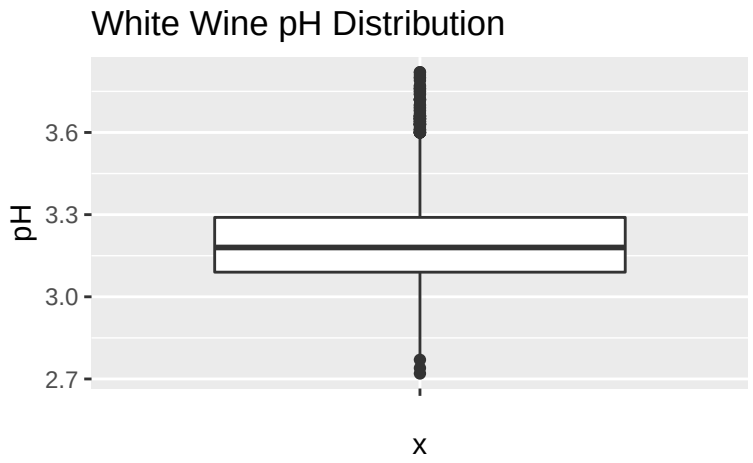

```
## Warning: Removed 1400 rows containing non-finite values (stat_boxplot).
```

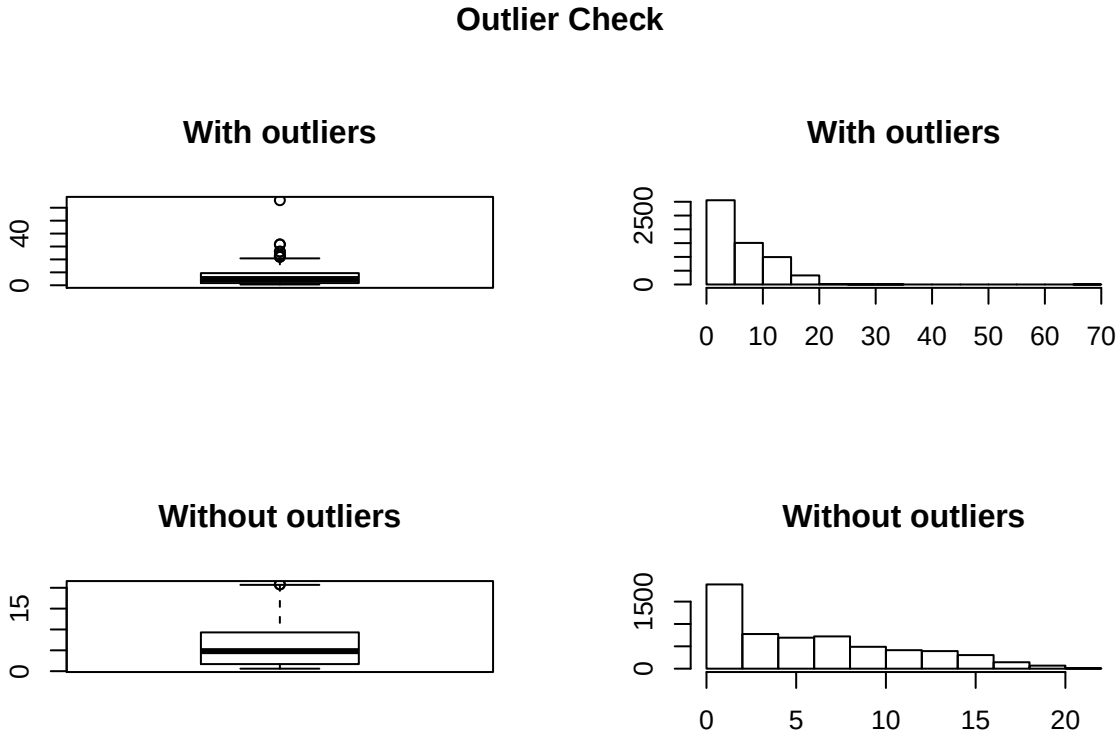

```
## Warning: Removed 1400 rows containing non-finite values (stat_boxplot).
```



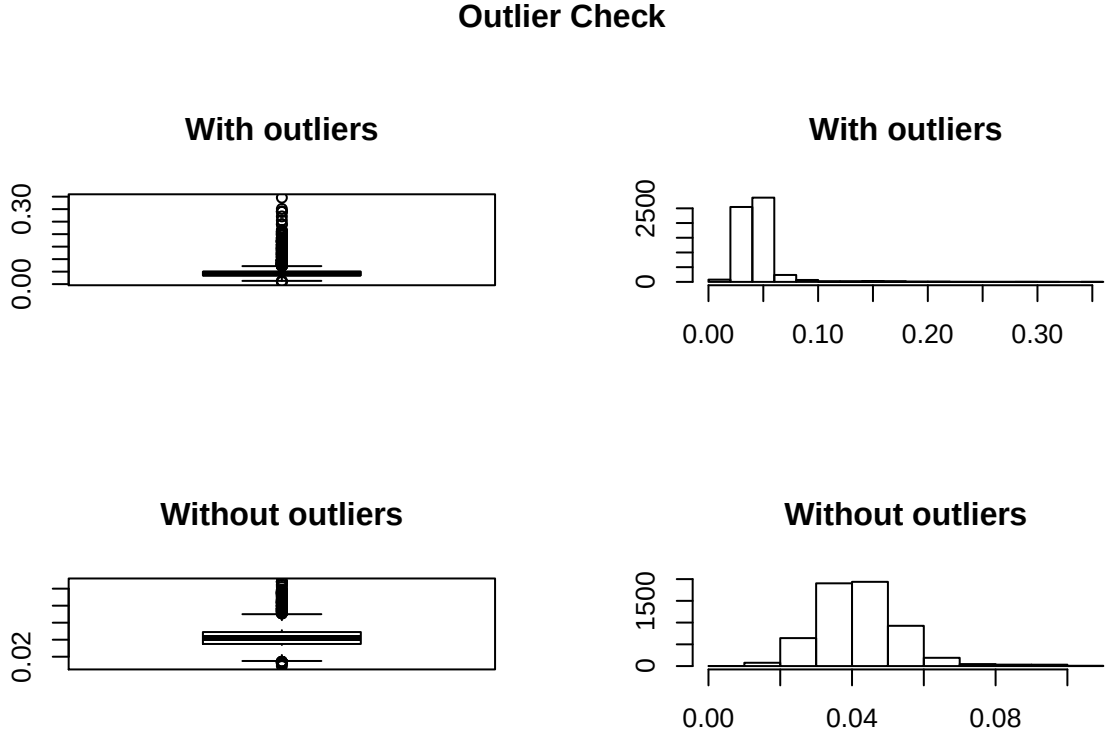
Outliers Not Removed:

```
## Warning: Removed 1400 rows containing non-finite values (stat_boxplot).
```



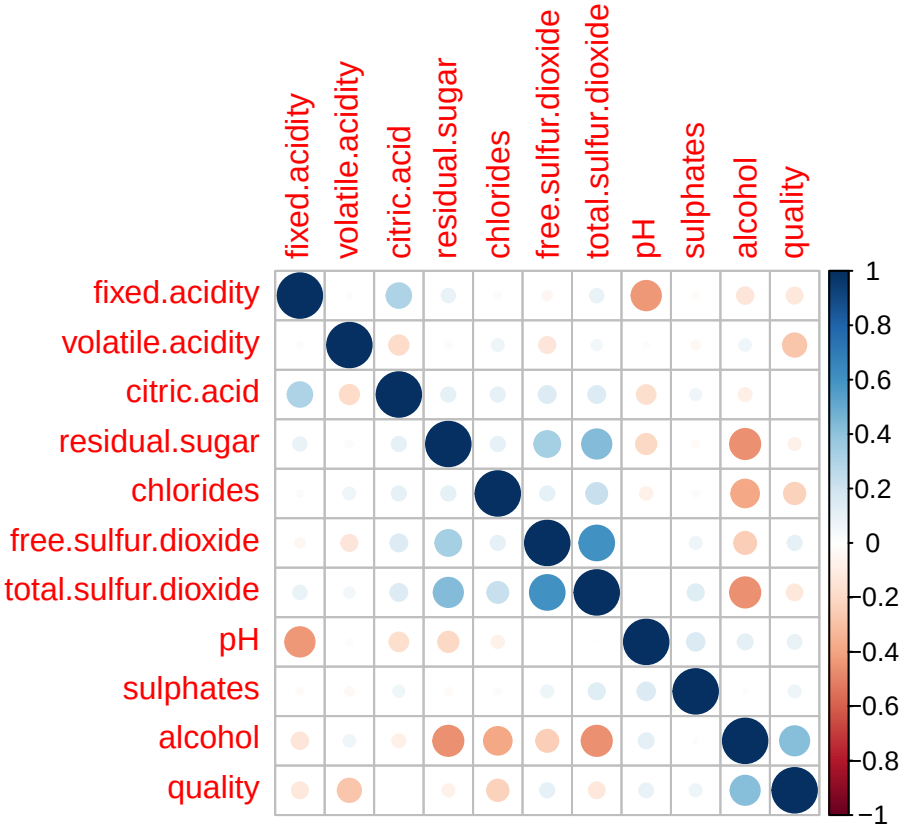


The figure above shows an outlier check with the variable of residual sugar. As the first row indicates there is a heavy skew on the left, between 0 and 15 for the variable when the outliers are included, while the graph on the second row indicates there is a more evenly spread distribution when the outliers are excluded. Many of our variables have are skewed in such a manner, but for this variable it is mainly driven by one large outlier, which should be removed.



Above is an example of an outlier check done on a variable where we did not remove outliers, chlorides. We can see that there is a large skew in this variable, but no one or few data points stick out away from the rest, and so it seems like the outliers are not errors, but novelties.

For the sake of completeness, it is worth noting that our correlation plot is not significantly changed by the removal of outliers. Therefore, we can conclude that the relationships in our model were not significantly affected by them. Note also that at this point, we have dropped the density column.



Chapter 3: Feature Selection and Standardization

ANOVA

Given that our prediction variables are all continuous and the target variable is discrete with more than two classes, we will use the analysis of variance (ANOVA) test to determine which features we will use in our model. The idea is that the ANOVA test will compare sub-distributions of each feature for each quality, and determine if the means are significantly different. The ideal feature would create distinct distributions for each quality rating and a poor feature would create similar distributions for each quality rating. We rate our variables based on the F-statistic to see which are the best predictors. In general, the higher the F-statistic, the better the predictor.

To determine specifically which distributions are significantly different, we use a Tukey Test. The Tukey Test will tell us for which quality ratings the feature variable has significantly different distribution means. For example, if the Tukey Test returns a significant p-value for the distributions between 4-5, then the variable will likely be able to distinguish between the classes. If the p-value is not significant however, the variable may have trouble with the two classes and may pass the error on to the model. The Tukey Test will give us valuable insight as to which quality ratings may confuse the algorithm, as some features may not generate significantly different distributions for all variables.

Our feature selection method differs slightly from (Cortez et al. 2009), who use backward selection governed by sensitivity analysis. Sensitivity analysis is an algorithm that determines the variance of the model output

with respect to each input variable. The importance of the feature will likely be determined by how changing it perturbs the output, that is the features that create the highest variance in the output will be the most important. The procedure consists of holding all other variables at their average values except for the value of interest, and then computing the model variance as follows:

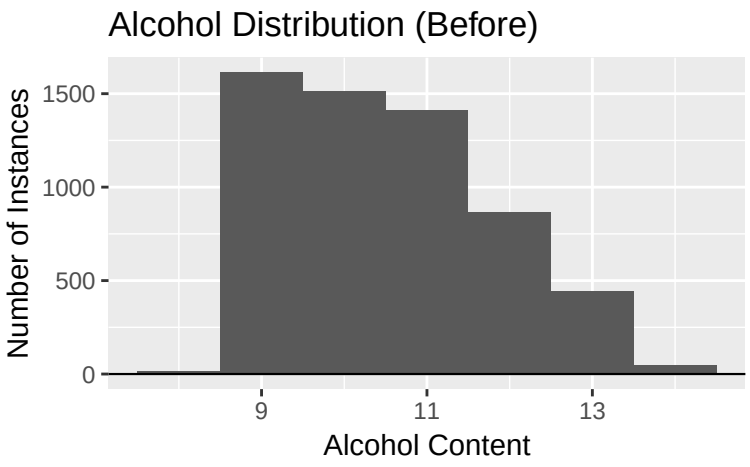
```
##                              fvals
##  [1,] "alcohol"              "562.615508110023"
##  [2,] "volatile.acidity"     "228.04119359909"
##  [3,] "chlorides"            "124.949695716141"
##  [4,] "total.sulfur.dioxide" "119.589974959878"
##  [5,] "free.sulfur.dioxide"  "114.044465631803"
##  [6,] "residual.sugar"       "69.5696187721314"
##  [7,] "fixed.acidity"        "34.2515735214261"
##  [8,] "pH"                   "24.3193886849616"
##  [9,] "citric.acid"          "13.3808154333753"
## [10,] "sulphates"            "11.930568413802"
```

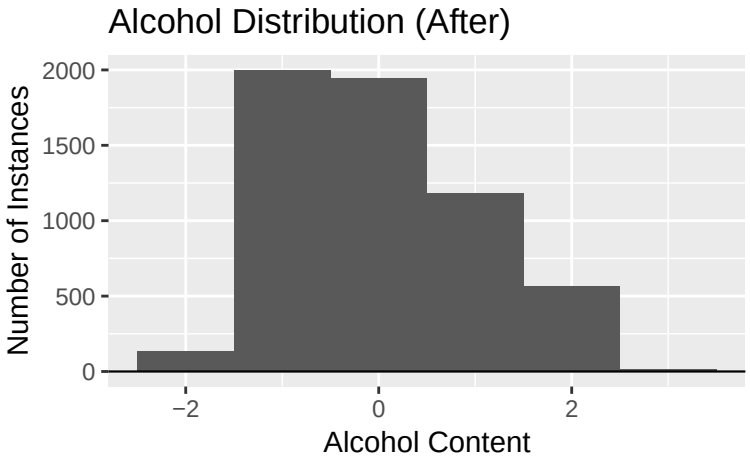
The 5 variables predicted as the most important differ somewhat from those in (Cortez et al., 2009). They found that the 5 most important variables were sulphates, alcohol, residual sugar, citric acid and total sulfur dioxide in that order.

Standardizing Features

In an effort to follow the procedure in (Cortez et al., 2009), we normalize the data by standardizing it to a mean of 0 and a variance of 1 (Hastie et al, 2001), to ensure the model is not sensitive to variables that tend to be large. For example, free sulfur dioxide is larger than any of the other variables used in our fit. Below are two histograms of the alcohol content feature, showing the distribution before and after standardizing as an example.

```
## Warning: Removed 1400 rows containing non-finite values (stat_bin).
```





Chapter 4: Results

Using our 5 best predictor variables according to ANOVA analysis (alcohol, total sulfur dioxide, fixed acidity, volatile acidity, chlorides), we will fit a continuous linear regression to the quality. Given that our quality is a factor variable, and the linear regression will return floating point numbers that may be between the quality ratings, we will use the rating it is closest to as the quality prediction by rounding.

(Cortez et al., 2009) used several models in their approach, including multiple regression and support vector machines (SVM). Given that we also use multiple regression, we will compare our results to that model as well as SVM as it is their best model.

Partitioning Data

Next, we split our data into training and test samples. The training sample will consist of approximately 70% of our dataset, and the test sample will consist of the remaining 30%. We will use the training sample to build our linear regression, and then use the remaining test sample to determine the predictive accuracy.

We tried three fits to the quality, the first of which used all 5 variables predicted by ANOVA analysis. For this model, we saw that the total sulfur dioxide variable was not significant in the fit with a p-value of 0.8151. The coefficient was also very small, and the error was 4 times as large as the coefficient itself. These led us to believe total sulfur dioxide was not a good predictor and needed replacement. The next highest F-statistic ranking was residual sugar, so we used that predictor instead of total sulfur dioxide for our next model. After this regression fit, we saw that chlorides was no longer significant with a p-value of 0.0898. The error in the coefficient was almost half of the coefficient itself, so we replaced chlorides with the next variable on our ANOVA list, fixed acidity.

```
##
## Call:
## lm(formula = quality ~ alcohol + residual.sugar + free.sulfur.dioxide +
##     volatile.acidity + fixed.acidity, data = winequality_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62437 -0.51530  0.05085  0.54833  2.04930
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          5.87253    0.01197 490.800  < 2e-16 ***
## alcohol              0.50166    0.01375  36.482  < 2e-16 ***
## residual.sugar       0.12674    0.01403   9.034  < 2e-16 ***
## free.sulfur.dioxide  0.13186    0.01296  10.172  < 2e-16 ***
## volatile.acidity    -0.27758    0.01209 -22.954  < 2e-16 ***
## fixed.acidity       -0.05919    0.01210  -4.893 1.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7658 on 4092 degrees of freedom
## Multiple R-squared:  0.322,  Adjusted R-squared:  0.3212
## F-statistic: 388.7 on 5 and 4092 DF,  p-value: < 2.2e-16
```

```
## [1] "VIF"
```

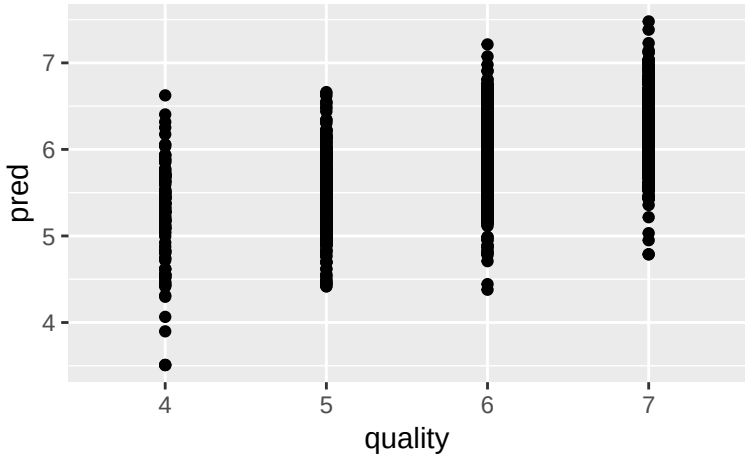
```
##             alcohol      residual.sugar free.sulfur.dioxide
##            1.313096            1.371972            1.180365
##    volatile.acidity       fixed.acidity
##            1.035407            1.032903
```

Our final model consisted of alcohol, volatile acidity, free sulfur dioxide, residual sugar, and fixed acidity. The specifics on our final regression fit are reported above.

Testing Fit

Below is a plot of the prediction versus the quality for the test set. Below that is the percent accuracy, measured by the ratio of predictions that are within 0.5 of the actual quality. For example, if the prediction for a quality 5 wine lies in the range 4.5 <= quality < 5.5, it is a correct prediction.

We can see that the predictions are very spread, and no single class is perfectly predicted. As expected, our model is still not very accurate at predicting a quality of 4. This is likely due to the fact that there is less information on that class than the others, as replication oversampling does not provide any new information.




```
## Percent Correct Predictions:
```

```
## [1] 0.4607062
```

Chapter 5: Conclusion

In conclusion, we have determined the best model for our linear regression with 5 predictors is the above using alcohol content, volatile acidity, fixed acidity, free sulfur dioxide and residual sugar. These variables seem to make sense; sulfur dioxide is a preservative in the wine, which may contribute to freshness. Too much volatile acidity tends to overwhelm the wine with a vinegar flavor, reducing the quality. This variable also has a negative coefficient in our fit, which is expected. Residual sugar contributes to the sweetness of the wine, which is a good indication of quality as people can easily pick up on sweetness. Alcohol content is related to sweetness, but is also a quantity that sticks out on it's own. It is one of the most recognizable characteristics of a wine.

The ANOVA analysis yielded total sulfur dioxide in the top 5 predictors, but it can be seen that the predictor was not significant in the model as the p-value was very large. When we replace total sulfur dioxide with residual sugar, the fit improves marginally, as the R^2 value improves from 0.3064 to 0.3185. So we adopt the model with residual sugar. After trying this regression, we found that chlorides was no longer statistically significant to the fit with a p-value of 0.0898. In addition, the error on the coefficient was around half of the actual value for the coefficient, so we replaced chlorides with fixed acidity, the next best predictor according to our ANOVA analysis. This increased the R^2 value from 0.3185 to 0.3220, and all variables are significant.

In comparison to (Cortez et al., 2009), our accuracy is slightly worse. Strictly considering correct predictions over all predictions, their SVM model returned an accuracy of 62.4% and the multiple regression model an accuracy of 59.1%. Given that for multiple regression, they used 9.2 features on average in their k-fold process. We were able to approach that accuracy with ANOVA feature selection and regression analysis on only 5 features. Our accuracy was slightly lower at 46.07%, but this is still better than random selection and not far off given the reduction in predictors.

Future work would consist of trying different models, such as SVM, to see how well they predict quality based on our selected features. It is likely that the function mapping physicochemical properties to quality is non-linear, which violates our linear regression assumption. However, we were able to get decent results through a simple linear model. It would be useful for a future dataset to have more balanced classes, as this would likely improve our model significantly. Being able to recreate similar results to published work with fewer features would be an enormous benefit to the wine industry, reducing costs for quality control as there would be no need to invest in a system that captures every feature of the wine, or even all 11 covered in our dataset.

References

Organisation Internationale de la Vigne et du Vine (OIV), State of the Vitiviniculture World Market, April 2018. http://www.oiv.int/public/medias/5958/oiv-state-of-the-vitiviniculture-world-market-april-2018.pdf

Ebeler S., Flavor Chemistry — Thirty Years of Progress, Kluwer Academic Publishers, 1999, pp. 409–422, chapter Linking flavour chemistry to sensory analysis of wine.

Cortez, P. et al., Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems 47 (2009) 547-553.

Hastie T.,Tibshirani R.,Friedman J., The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer-Verlag, NY, USA, 2001.

Wikipedia, Acids in Wine, 2019. https://en.wikipedia.org/wiki/Acids_in_wine

Wikipedia, Sweetness in wine, 2019. https://en.wikipedia.org/wiki/Sweetness_of_wine

Wikipedia, Sulfur Dioxide (In winemaking), 2019. https://en.wikipedia.org/wiki/Sulfur_dioxide