Statistics 6021: Project 2

Michael Kolonay (mhk9c) - Jordan Machita (jm8ux) - Stephen Morris (sam3ce) - Siddharth Surapeneni (sss2e)

Vinho da 'Ville: A Charlottesville Wine Story

Background

Vinho da 'Ville is a new winery planning to open next year in Charlottesville. The owners of the winery have identified a niche market segment of locals who enjoy Portuguese wine that is not currently being met by the established wineries. With an unstable global economy, the owners fear that relying on importing the wine from Portugal may leave them vulnerable to trade tensions and other international factors. Thus, they intend to make their own wine that mimics the quality of the Portuguese. While the owners are well-established businesspeople, they know little about the characteristics and qualities that differentiate wine. To fill this knowledge gap, Vinho da 'Ville has hired our expert team of data scientists with two goals in mind. First, they want to understand the various characteristics of Portuguese wine. Second, they want us to build a model that accounts for the various attributes of wine and categorizes the wine as high quality or low quality.

In order to meet the needs of Vinho da 'Ville, we traveled to northern Portugal to see firsthand what sets their wine apart. While our team was there, we obtained two datasets on Portuguese wine, one with the characteristics and quality rating of nearly 1,600 red wines and the other with almost 5,000 white wines. With the necessary data at hand, we have conducted extensive analysis to meet the objectives of Vinho da 'Ville. We have analyzed the various characteristics of the wine and extracted key information about how the characteristics affect quality for both red and white wines separately. We have also conducted analyses aimed at exploring the similarities and differences between Portuguese reds and whites. With this information at hand, we are confident that the owners of Vinho da 'Ville will be able to wine and dine with the most well-known of wine drinkers and leave the impression that they are world-class wine connoisseurs.

However, there is more to running a winery than being an expert on tasting the wine. The owners of Vinho da 'Ville must also be able to make the wine. We have generated two models to aid them in this feat. Both models allow the owners to manipulate the characteristics of the wine to see how they influence quality, but one will be for red wine and the other for white. The two must be approached differently as their various attributes influence quality in different ways as will be described in the Data Background. With these models, Vinho da 'Ville will be able to distinguish their high quality wines from the masses. This is of the utmost importance to the owners as they need to be confident that they are serving high quality wines to their higher end clientele. This is the precision metric, which will be addressed in detail throughout the analysis sections of the report. Meanwhile, they still want to have an extensive collection of lower quality wines that will be of more interest to their everyday, local customers and UVA students. The owners have emphasized that the consequences of misclassifying a low quality wine and serving it to their more important clients are much more severe than mistakenly serving a high quality wine thought to be of lower quality to their regular customers.

These models will allow Vinho da 'Ville to produce a plethora of wines that vary in quality and provide them with the confidence that they will be able to speak to the individualization of each wine created. In the following sections of this report, we give a detailed analysis of the wine characteristics related to quality as well as a description of the models we have generated and how we have tested and refined them. The information provided will equip Vinho da 'Ville with the tools they need to succeed in this niche market.

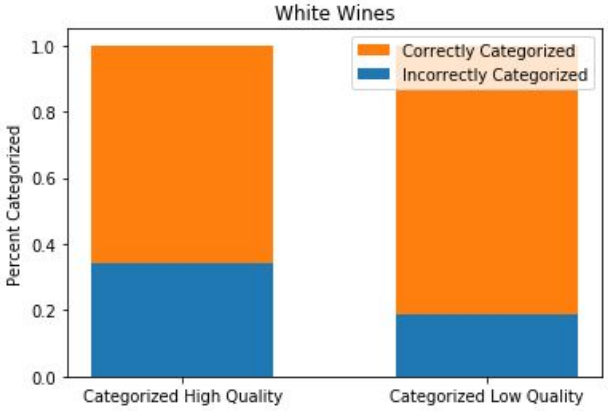
Executive Summary

We conducted extensive analysis and testing of multiple predictive models to arrive at a separate optimum model for each white and red wines. Our models focus on identifying the characteristics of each color that are most significant for assessing a high quality Portuguese wine, defined in this study as having a **quality rating of 7 or above**. With our recommendations, we are confident that you will be able to craft a high quality wine to exceed the expectations of a discerning clientele. Acknowledging that no predictive model can be 100% accurate, we have also tailored thresholds for applying these models which will **minimize the possibility of misclassifying a lower quality wine.** In doing so, we do reduce the available population of high quality wines, but this is an acceptable tradeoff to help ensure that you do not serve an unfitting wine to your most distinguished clients.

White wines: The optimum white wine formula will focus on these four characteristics:

- Alcohol content. The percentage of the wine that is alcohol, sugars converted in the fermentation process. As alcohol content *increases*, so does the probability of producing a high quality wine.
- Residual sugars. The amount of sugars remaining in the wine after fermentation, contributing to the wine's sweetness. As residual sugars *increase*, so does the probability of producing a high quality wine.
- Chlorides. A mineral often paired with sodium that increases the salty flavor of wine. A *reduction* in chlorides will increase the probability of producing a high quality wine.
- Density. Simply the mass by volume, determined by the alcohol, sugar, and glycerol concentration. *Reducing* this increases the probability of producing a high quality wine.

Additionally, we recommend applying this model with a **threshold of 55%** to minimize the chance of selecting a lower quality wine erroneously classified as high quality. This yields an **overall accuracy rate of 80.0%**, but 97.0% of lower quality wines are identified correctly. The downside is that only 20.6% of high quality whites are correctly identified. This may inadvertently introduce more high quality wines into your inventory which will be marketed to a lower cost customer. However, of the wines categorized as high, **65.9% are truly high quality.**



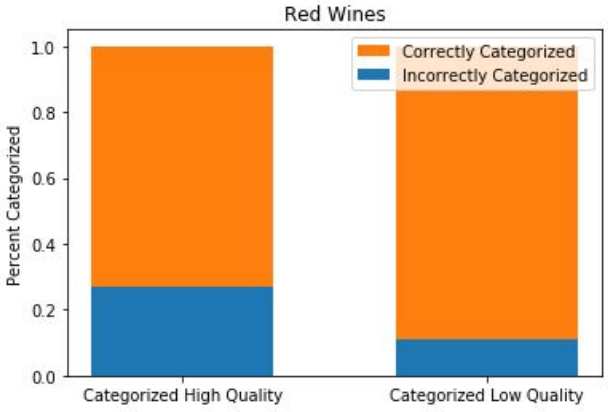
White Wines		
55% Threshold	Categorized as Low	Categorized as High
True Low Quality Wines	1848	**58**
True High Quality Wines	431	**112**

Note: Figures in **green** are categorized correctly; those in **red** are categorized incorrectly.

Red wines: The optimum red wine formula will focus on these characteristics:

- Alcohol content. As above. As alcohol content *increases*, so does the probability of producing a high quality wine.
- Volatile acidity. The acidic elements of wine that are gaseous, sensed as smell rather than taste. With a *reduction* in volatile acidity, the probability of high quality rises.
- Sulphates. A product of sulphur dioxide added during winemaking to preserve the wine from the negative effects of bacteria and oxidation. With an *increase* in sulphates, we increase the probability of producing a high quality wine.

For the reds, we recommend applying our model with a **threshold of 45%** to minimize the chance of wrongly classifying a low quality wine as a high quality wine, while still maintaining a useful inventory size for selection of true high quality wines. With this threshold, we maintain an **overall 88.3% accuracy**, with 98.4% of lower quality wines identified correctly. As before, the tradeoff is that only 26.5% of high quality whites are correctly identified. But most importantly, of the wines categorized as high, **73.2% are truly high quality.**



Red Wines		
45% Threshold	Categorized as Low	Categorized as High
True Low Quality Wines	676	**11**
True High Quality Wines	83	**30**

Note: Figures in **green** are categorized correctly; those in **red** are categorized incorrectly.

Armed with our recommendations and analysis, you can be quite confident in the quality of wines delivered to your best-paying, "high quality" customers!

Data Description

In our analysis, we use two datasets that consist of the same characteristics of wine, one for red wine and the other for white wines. The red wine dataset has nearly 1,600 data points while the white wine set has almost 4,900. Both datasets include eleven physicochemical attributes of the wine as well as its quality rating on a scale of 0 to 10. The physicochemical characteristics are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. In our analysis, we add a binary characteristic of the wine that classifies it as high quality if the wine's quality rating is a 7 or above and low quality if the corresponding quality is a 6 or below. This allows us to delineate between the excellent wines and the more common normal and poor wines. We chose this approach because our client, Vinho da 'Ville, has stressed the necessity of their serving only the most excellent wines to their higher end customers. Thus, we need to be able to predict if a wine would be considered excellent, a quality of 7 or above, or not, a quality of 6 or below.

Data Background

Prior to building our models, we needed to make the critical decision of whether or not the red and white wines should be combined into one dataset for analysis. To make an informed decision, we compared the density distributions of the two dataset for each variable. Our first approach was to compare the histograms of each; however, the size of the datasets did not allow for a meaningful comparison in this way but the density distribution method achieved the same ends while standardizing the scale. These can be seen in Figure 1, where the black lines represent the white wines and the red lines correspond to red wine.

We first look at the distribution of the quality rating, which shows that red wines have the highest frequency of 5s followed by 6s before a steep drop off in 7s and barely any 8s. The white wines peak at a quality rating of 6 and are more densely distributed across the 7s and 8s than their red counterparts. We can see this relationship more clearly when looking at the binary quality category of zero for 6 and below and one for 7 and above. Although both appear to be imbalanced as the majority of the wines are classified as low quality, we can see that red wines have an even lower proportion of high quality bottles than whites.

By analyzing the eleven other variables, we cannot only make comparisons between red and white wines but we can also gain a better understanding of the individual characteristics of each. For the owners of Vinho da 'Ville, being able to speak to these isolated attributes as well as the contrasts will make all the difference in their merit as Portuguese wine connoisseurs. We can see that the alcohol content for reds and whites peaks around 10% but that whites are a bit more evenly distributed. Red wines tend to have a slightly higher pH with a peak around 3.4 while the whites are heavily concentrated around 3.1. Similarly, reds tend to have slightly higher sulphates with the most around 0.6 g/L compared to 0.5 g/L for whites. We next see that the density of red wines is pretty compact between 0.995 and 1.0 g/mL while the density of whites is more spread out but concentrated slightly lower in the 0.990 to 0.995 g/mL range.

Although the numeric values differ, the distribution comparison between total sulfur dioxide and free sulfur dioxide is very similar for reds and whites, indicating that they may be correlated. The sulfur dioxide levels are important as their presence protects the wine from spoilage but may dilute the flavor.

When it comes to total sulfur dioxide, reds are mostly concentrated between 0 and 50 mg/L, whereas whites tend to range from 75 to 200 mg/L. Similarly, the free sulfur dioxide of reds peaks between 0 and 20 mg/L, while whites have a smoother curve from 0 to 80 mg/L with a peak around 35 mg/L. Red wines tend to have higher chlorides, which contribute saltiness to the taste, clustered around 0.09 g/L, while whites converge around 0.05 g/L. The acidity levels in wine contribute to the freshness, tartness, and sourness of the taste. The trends in volatile versus fixed acidity are somewhat similar but differ more in volatile acidity for which whites have a sharp peak around 0.2 g/L and reds have more of a hill spread between 0.2 and 0.8 g/L. Whites also have a sharper peak in fixed acidity around 7 g/L, whereas reds have a smaller peak closer to 8 with a larger right tail heading toward 14 g/L. Red wines tend to have an even spread of citric acid between 0.0 and 0.5 g/L, while whites are more densely correlated around 0.3 g/L but mostly still fall within the same ranges as reds. Lastly, the residual sugar in red wines has a sharp peak around 2 g/L compared to the smaller peak of white wines around 1 g/L, but then whites tend to have a thicker right tail heading out to 15 g/L. The level of residual sugar adds sweetness to the wine that contrasts the tartness of the acidity.

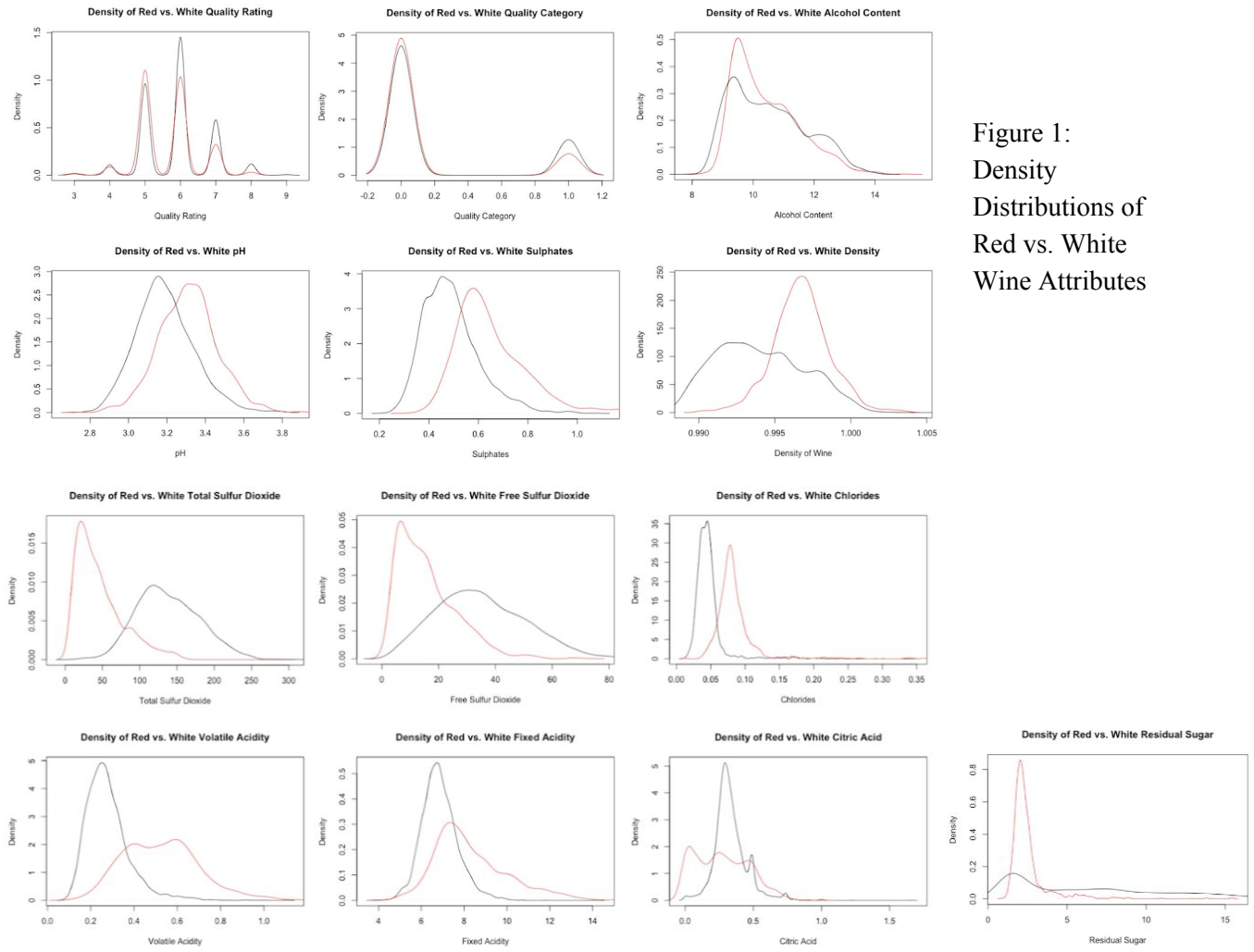


Figure 1: Density Distributions of Red vs. White Wine Attributes

White Wine

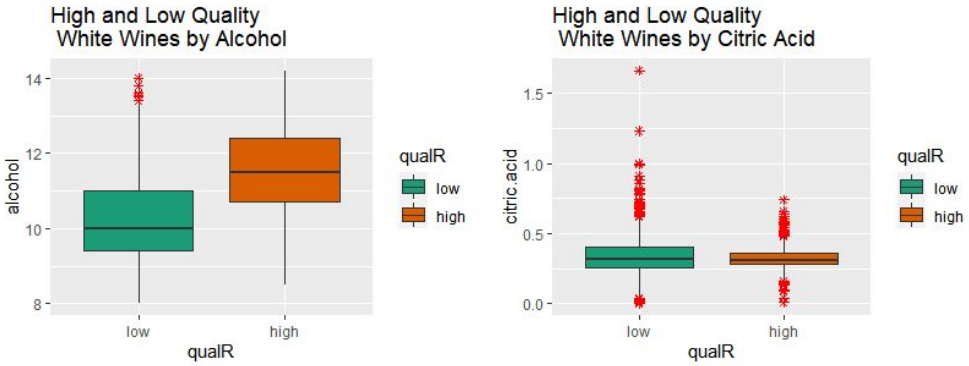
Exploratory Data Analysis

We started our analysis by segmenting the white wines into two categories, those with a quality rating above 6 and those with a quality rating of 6 or lower. This categorization gave us a set of data with 3,838 'low' quality white wines and 1,060 'high' quality white wines.

The goal of the model would be to predict high quality wines (greater than 6) while minimizing the generation of false positives. This adheres to the clients priorities because for them it would be better to sell a wine presented as low when it was in fact high quality than to sell a wine presented as high quality when it was actually a low quality wine. It follows that a consumer would have a more favorable impression of the store if they thought they had gotten a good deal on a wine (the false negatives) rather than if they had felt they were oversold a wine that was billed as a high quality but was in fact low quality (the false positives).

We examined each of the wine characteristics presented in the data to see if there were any trends that we could use to inform our model building. We hypothesized that characteristics that had visually apparent differences in their boxplots would be better predictors when identifying wines as high or low quality.

Figures 2 and 3: Boxplots of Alcohol and Citric Acid by Quality



In Figure 2, we compare wines by alcohol by quality. We can clearly see that the median values are different between low and high quality wines. We suggest that alcohol content would make a good candidate to be further considered for our model. Figure 3 shows the boxplot of wine quality by citric acid concentration. The difference between the two qualities of wine plots was not as pronounced as in alcohol but the different sizes of the interquartile range and the distribution of the outliers make this a candidate for a predictor as well.

Figures 4 and 5: Boxplots of Chlorides and Residual Sugar by Quality

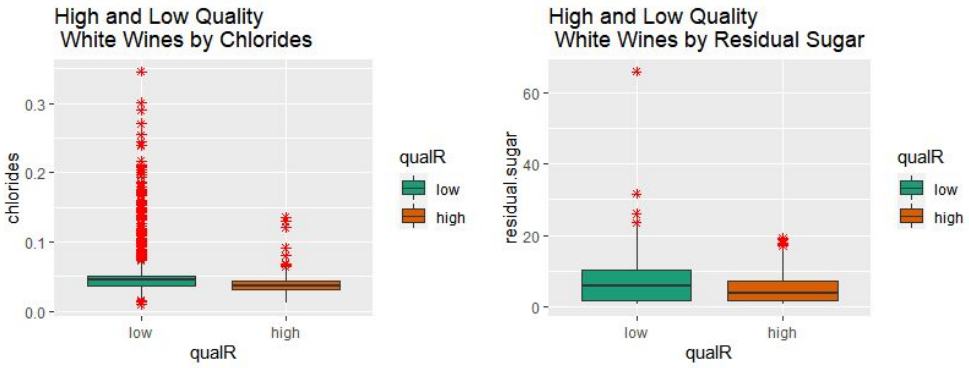
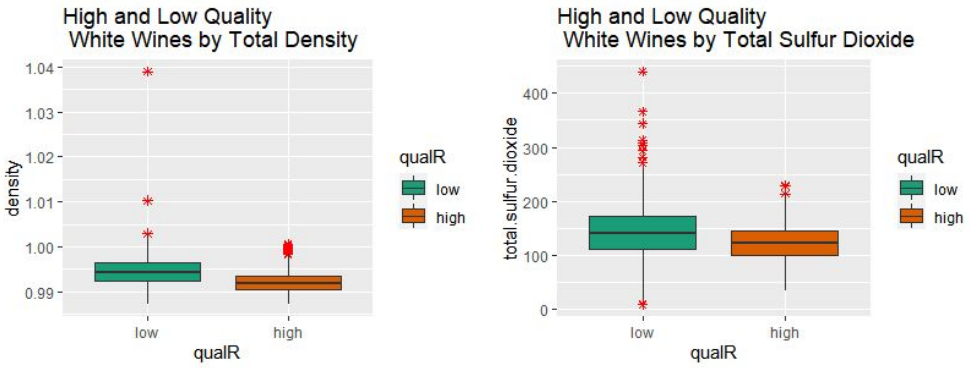


Figure 4 shows the difference between chlorides for high and low quality wines. We can see that the interquartile range is similar, but the median value appears to be different between the two categories. Additionally, because the lower quality has a significant spread of values that fall outside the interquartile range the difference may translate into a good predictor for the model. Figure 5 shows the difference between residual sugar for high and low quality wines. The interquartile range seems to differ and the median appears different between the two categories, thus adding residual sugar to our list of candidate predictors.

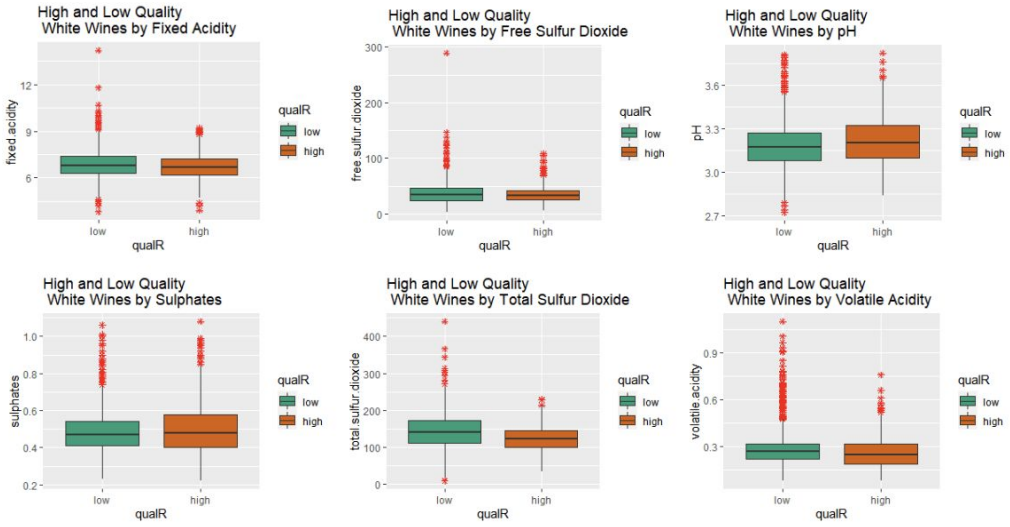
Figures 6 and 7: Boxplots of Density and Total Sulfur Dioxide by Quality



Similarly, Figure 6 shows the differences in density between high and low quality wines. Low quality wines have a distinctly higher median and a larger interquartile range than high quality wines. Figure 7 shows the boxplots for low and high quality wines for total sulfur dioxide. While the medians appear close the interquartile ranges are distinct and lower quality wines have many more outliers than high quality wines. These differences are sufficiently distinct to consider density and total sulfur dioxide as predictors.

The remainder of the characteristics of wine did now show as much variation among the lower and higher quality wines and were not considered good predictors for the model. The boxplots of these characteristics, fixed acidity, free sulfur dioxide, pH, sulphates, total sulfur dioxide, and volatile acidity, can be seen in Figure 8.

Figure 8: Boxplots of Remaining Predictors by Quality



Detailed Analysis

After careful consideration, our team decided to generate a model with two competing methods. It was clear from our EDA that some of the predictors appeared to be more useful than others, but we still wanted to verify that we were not discarding any significant predictors, so we utilized two distinct forms of model generation. The first model was created using a backward elimination method that started with the full set of predictors. The second model used a similar step back methodology but started with the predictors indicated as important in the EDA above: alcohol, citric acid, residual sugar, chlorides, total sulfur dioxide, and density. Both backward elimination methods involved creating a model and dropping one predictor that had the highest standard error and largest p value at a time. The model would then be refit and evaluated again in an iterative process. At each step, an ROC curve, auc value, confusion matrix for a .5 threshold, and corresponding metrics (Sensitivity, Specificity, Accuracy, Ratio of True Positives to Total Predicted Positives) were generated and used to determine if the model was improving as predictors were dropped. We added a column “Precision” which is simply the number of True Positives divided by the sum of True and False Positives. This is highly important to our client as the Precision is how we measure the probability of not mistakenly serving a low quality wine as high quality. Throughout our analysis, we will continue to focus on balancing maximization of the Precision with minimizing reduction of the number of predicted positives. This tradeoff will leave Vinho Da ‘Ville with a significant inventory of wines classified as high quality without the fear that many are truly low quality.

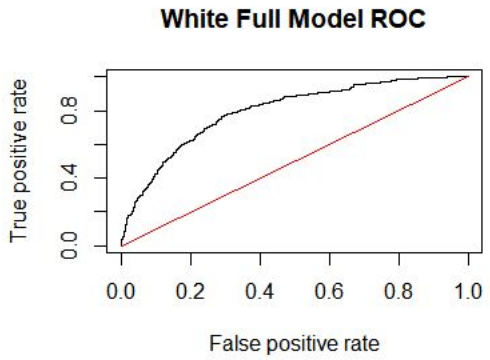
Method 1: Backward Elimination with Full Set of Predictors

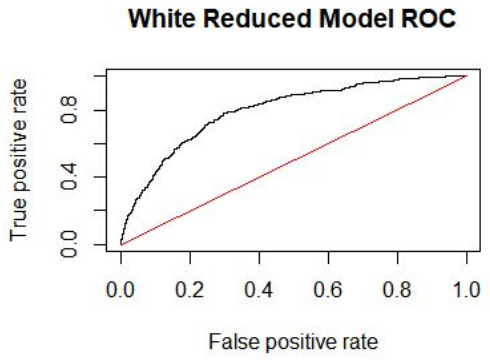
Starting with the full set of predictors in the data (alcohol, fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, and sulphates), models were generated and then refined with a stepwise removal of the predictor with the largest p values and standard deviation for that iteration of the model. We summarize the results of this process in Table 1.

Table 1: Method 1 Iteration Results

Step	AUC	Confusion(.5)	Sensitivity	Specificity	Accuracy	Precision	dropped
0	0.7939588	FALSE TRUE 0 1819 98 1 388 144	0.270677	0.9488785	0.801552	0.5827338	
1	0.7939153	FALSE TRUE 0 1789 117 1 381 162	0.298343	0.9386149	0.796652	0.5806452	alcohol
2	0.7937192	FALSE TRUE 0 1790 116 1 376 167	0.307551	0.9391396	0.799102	.590106	citric.acid
3	**0.7943926**	**FALSE TRUE 0 1789 117 1 376 167**	**0.307551**	**0.9386149**	**0.798693**	**0.5880282**	**Total Sulfur dioxide**

Figures 9 and 10: Method 1 ROC Curves





As can be seen in Table 1 and Figures 9 and 10, the ROC curve did not change much from the start of the model evaluation to the end. At step 3 of the process there were no longer predictors with significant p values. This method resulted in only dropping three predictors, leaving us with eight still in the model, which is more complex than we had anticipated. The measure of precision had also started to decrease so we decided to evaluate the model at different thresholds for the confusion matrix as seen in Table 2.

Table 2: Method 1 Key Metrics

Threshold	Confusion	Sensitivity	Specificity	Accuracy	Precision
.5	FALSE TRUE 0 1789 117 1 376 167	0.3075506	0.9386149	0.7986933	0.5880282
.55	**FALSE TRUE 0 1833 73 1 422 121**	**0.2228361**	**0.9616999**	**0.7978767**	**0.6237113**
.60	FALSE TRUE 0 1863 43 1 451 92	0.1694291	0.9774397	0.798285	0.6814815

Our final summary table for this model can be seen in Table 3 and is with the following predictors: fixed acidity, volatile acidity, residual sugar, chlorides, free sulfur dioxide, density, pH, and sulphates. The

logistic regression equation produced by the summary table and shown below represents the log-odds of a wine being high quality. By exponentiating both sides, we would produce the odds that the wine is high quality. However, the threshold for our confusion matrices represents the required estimated probability of being high quality that a wine needs to surpass for our model to classify it as high quality. The traditional threshold is 50%, meaning that if our model calculates the probability of the wine being high quality to be above 50% then it classifies it as high quality. We adjust this threshold at various times to make it easier or harder for a wine to be classified as high quality and we will justify these decisions as they are made throughout the analysis.

Table 3: Method 1 Logistic Regression Model

Model Summary				
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	7.33E+02	4.83E+01	15.167	< 2e-16 ***
fixed.acidity	6.36E-01	9.06E-02	7.015	2.31e-12 ***
volatile.acidity	-3.22E+00	6.31E-01	-5.101	3.38e-07 ***
residual.sugar	3.44E-01	2.74E-02	12.56	< 2e-16 ***
chlorides	-1.67E+01	5.44E+00	-3.075	0.0021 **
free.sulfur.dioxide	6.58E-03	3.27E-03	2.013	0.0441 *
density	-7.57E+02	4.99E+01	-15.183	< 2e-16 ***
pH	3.70E+00	4.74E-01	7.81	5.70e-15 ***
sulphates	2.17E+00	4.70E-01	4.617	3.88e-06 ***

$$log\left(\frac{\pi}{1-\pi}\right) = 733 + .636\,(fixed.acidity) - 3.22\,(volatile.acidity) + .344\,(residual.sugar) - 16.7\,(chlorides) + .00658\,(free.sulfur.dioxide) - 757\,(density) + 3.7\,(pH) + 2.17\,(sulphates)$$

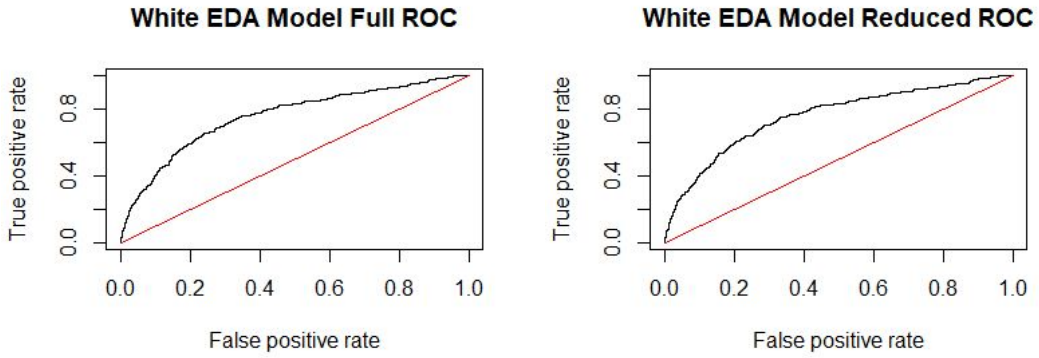
Method 2: Backward Elimination with Predictors from EDA

Starting with the set of predictors chosen from the EDA analysis (alcohol, citric acid, residual sugar, chlorides, total sulfur dioxide, and density) a model was generated followed by stepwise iterations with removal of the predictor with largest p value and standard deviation. The results of this process are displayed in Table 4 and the starting and ending ROC curves are illustrated in Figures 11 and 12.

Table 4: Method 2 Iteration Results

Step	AUC	Confusion(.5)	Sensitivity	Specificity	Accuracy	Precision	dropped
0	0.7572742	FALSE TRUE 0 1804 102 1 391 152	0.279926	0.9464848	0.798693	0.5984252	
1	0.7573061	FALSE TRUE 0 1805 101 1 390 153	0.281768	0.9470094	0.79951	0.6023622	citric acid
2	0.7581912	FALSE TRUE 0 1806 100 1 390 153	0.281768	0.9475341	0.799918	0.6047431	total sulfur dioxide
3	**0.7576863**	**FALSE TRUE 0 1802 104 1 382 161**	**0.296501**	**0.9454355**	**0.801552**	**0.6075472**	**density**

Figures 11 and 12: Method 2 ROC Curves



We noticed that after removing density in step 3, the number of false positives increased from 100 to 104. Since keeping false positives to a minimum is a critical measure in the criteria of generating this model, it was decided to vary thresholds for models at step 2 and step 3 to see which performed better at predicting higher quality wines while keeping the number of false positives low.

Table 5: Step 2 Metrics (alcohol, residual sugar, chlorides, total sulfur dioxide, **density**)

Threshold	Confusion	Sensitivity	Specificity	Accuracy	Precision
.5	FALSE TRUE 0 1806 100 1 390 153	0.281768	0.9475341	0.7999183	0.6047431
.55	**FALSE TRUE** **0 1848 58** **1 431 112**	**0.2062615**	**0.9695698**	**0.8003267**	**0.6588235**

Table 6: Step 3 Metrics (alcohol, residual sugar, chlorides, total sulfur dioxide)

Threshold	Confusion	Sensitivity	Specificity	Accuracy	Precision
.5	FALSE TRUE 0 1802 104 1 382 161	0.2965009	0.9454355	0.8015517	0.6075472
.55	# FALSE TRUE # 0 1847 59 # 1 441 102	0.1878453	0.9690451	0.795835	0.6335404

Utilizing Tables 5 and 6, we can see that the model including density with a threshold of .55 for true predictions produces better results than the model without density. Notably, the model with density included better optimizes our precision metric and gives slightly larger numbers for wines correctly classified as higher quality.

Therefore, the best model to use generated by this method is with the following predictors: alcohol, residual sugar, chlorides, total sulfur dioxide, and density. We summarize the results of this model in Table 7.

Table 7: Method 2 Logistic Regression Model

Coefficients				
	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	114.6922	70.02613	1.638	0.101453
alcohol	0.59823	0.10116	5.914	3.34e-09 ***
residual.sugar	0.09231	0.02717	3.398	0.000680 ***
chlorides	-20.31611	5.23858	-3.878	0.000105 ***
density	-122.94871	69.73427	-1.763	0.077883 .

$$log\left(\frac{\pi}{1-\pi}\right) = 114.69 + 0.59823\,(alcohol) + 0.09231\,(residual.sugar) - 20.31611\,(chlorides) - 122.94871\,(density)$$

Model Selection

Table 8: Comparison of Method 1 and 2 Results

Model	Threshold	Confusion	Sensitivity	Specificity	Accuracy	Precision
Method 1: Full	.55	FALSE TRUE 0 1833 73 1 422 121	0.2228361	0.9616999	0.7978767	0.6237113
Method 2: EDA	**.55**	**FALSE TRUE** **0 1848 58** **1 431 112**	**0.2062615**	**0.9695698**	**0.8003267**	**0.6588235**

We summarize the results of the final models selected by methods 1 and 2 in Table 8. The full model we generated from starting with the entire set of predictors had better sensitivity than the EDA informed model from method 2. However, this sensitivity came at the expense of misclassifying more lower quality wines as higher quality. The full model classified 73 wines as high quality when they were in fact low quality, whereas the EDA model falsely classified 58 wines as high quality. Overall, the EDA model from the second method demonstrated better predictability when considering the criteria of Vinho Da 'Ville. Therefore, we recommend using the EDA model to predict whether a white wine is of low or high quality.

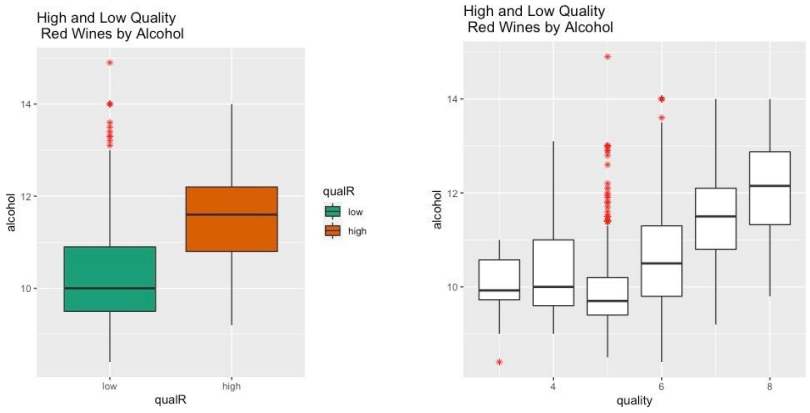
Red Wines

Exploratory Data Analysis

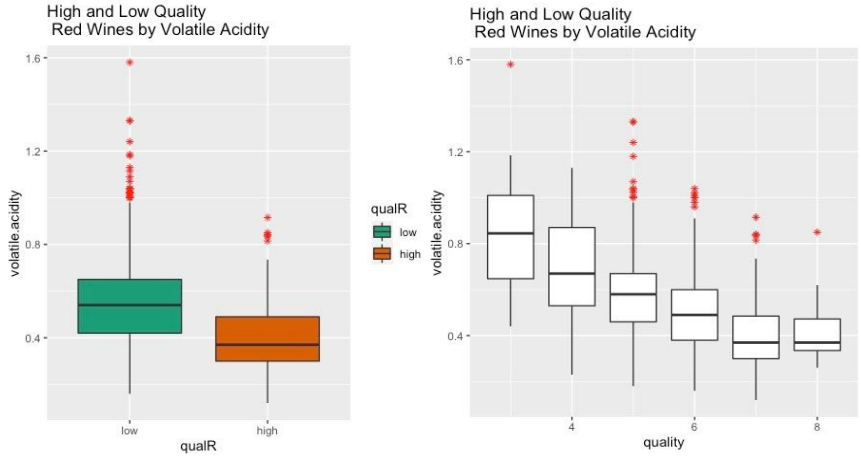
For the Exploratory Data Analysis, we began by plotting boxplots to compare each feature to see how the quality of the red wine varies, similar to our procedure for the white wines. Through these boxplots we realized some potential predictors that could be used to help classify whether a wine is of high quality or not. When comparing the boxplots for low and high quality wines by alcohol in Figures 13 and 14, there was a noticeable difference in the medians of alcohol content for each boxplot indicating that the alcohol content in high quality wines might be slightly higher than the alcohol content in low quality wines.

Volatile acidity also appeared to be a noteworthy predictor due to the evident difference in medians visible in Figures 15 and 16 with low quality red wines tending to have higher volatile acidity.

Figures 13 and 14: Boxplots of Alcohol by Quality

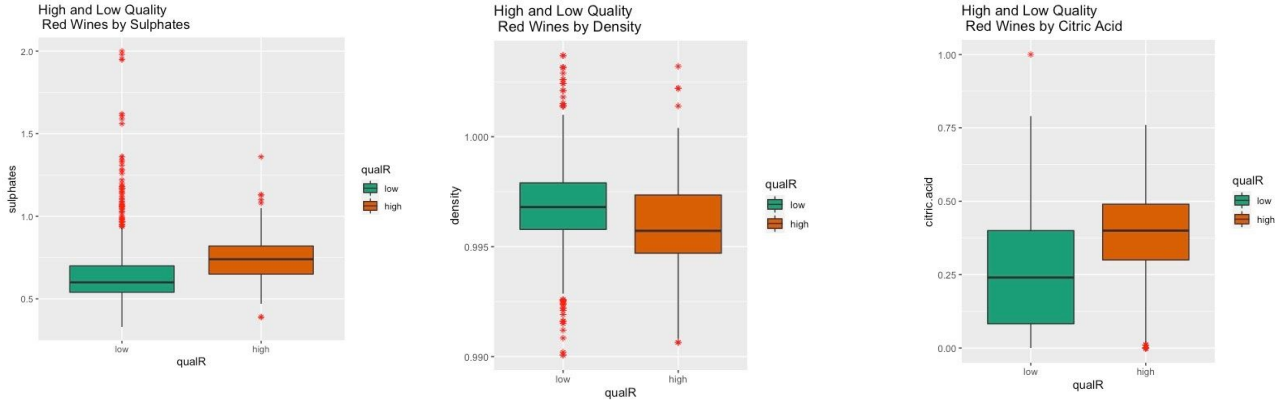


Figures 15 and 16: Boxplots of Volatile Acidity by Quality



We followed the same approach to determine that sulphates, density, and citric acid may be useful predictors. From the boxplot in Figures 17, we can see that the median quantity of sulphates is greater in high quality red wines. On the other hand, Figure 18 reveals that high quality red wines may have lower density. Lastly, the quantity of citric acid appears to increase with quality, which is apparent in Figure 19.

Figures 17,18,and 19: Boxplots of Sulphates, Density, and Citric Acid by Quality



Among the predictors we decided that might not be useful in predicting whether a wine was of high quality or low quality was fixed acidity. While the boxplots in Figures 20 and 21 indicate that the median fixed acidity is higher in high quality wines than in low quality wines, we see in Figure 21 that the level of fixed acidity across the different quality wines is variable. So we considered fixed acidity to not be a potential predictor. The remaining predictors that we did not deem to be significant based on our exploratory data analysis were residual sugar, chlorides, pH, total sulfur dioxide, and free sulfur dioxide. The boxplots for these predictors against quality are shown in Figure 22 and provide a basis for this decision.

Figures 20 and 21: Boxplots of Fixed Acidity by Quality

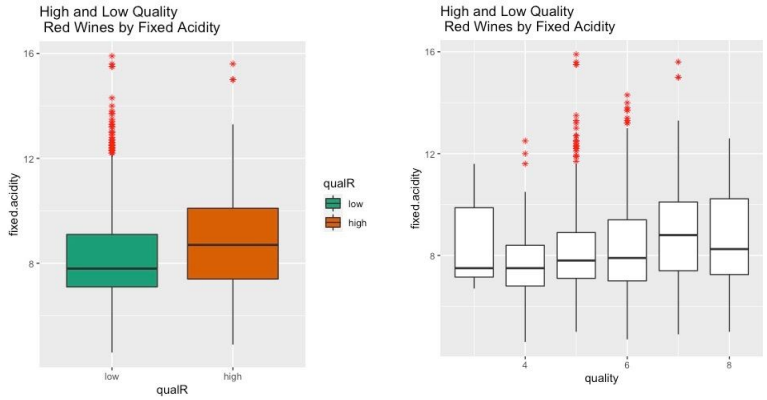
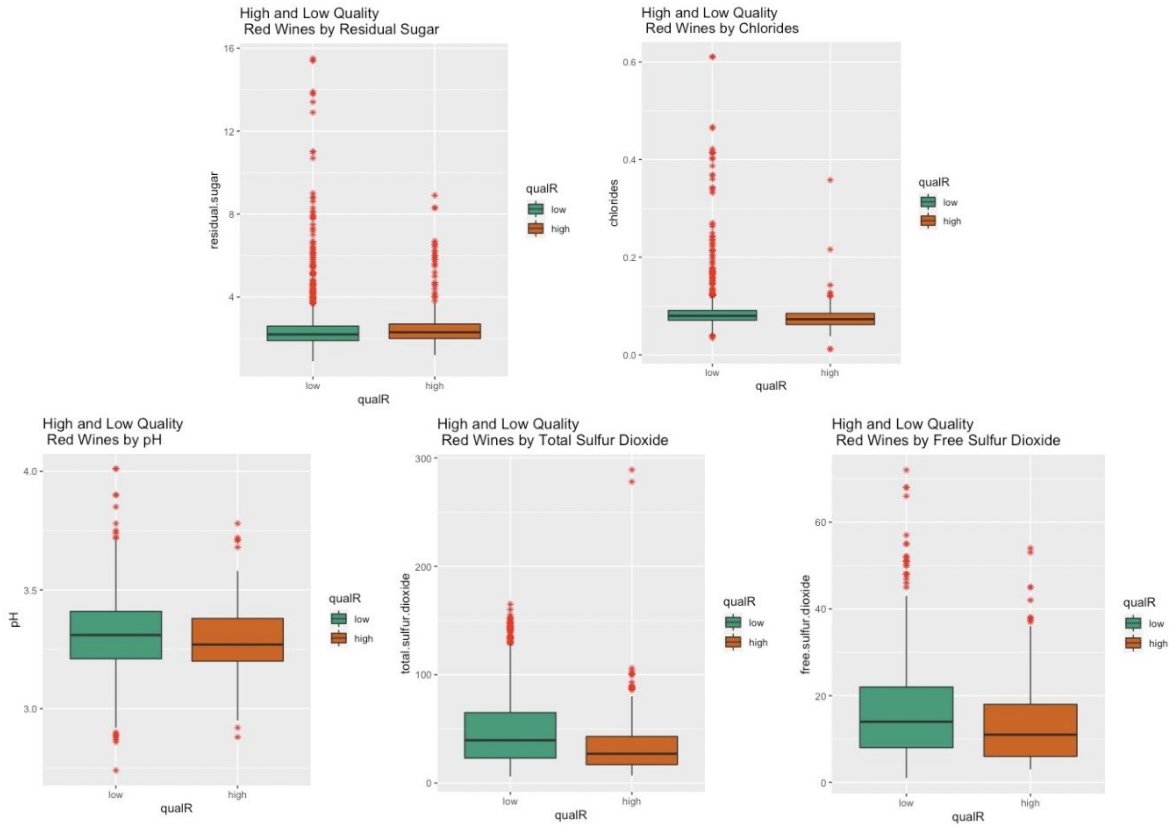


Figure 22: Boxplots of Remaining Predictors by Quality



Based on our EDA, we decided that Alcohol, Volatile Acidity, Sulphates, Density, and Citric Acid may be potential predictors for determining whether the quality of red wine.

Detailed Analysis

For the model building process, our team used two different approaches. One approach was to use the backwards elimination technique where we would fit the full model. Using this full model, we would drop the predictor with the highest insignificant p-value from the results of the Wald tests and we would keep iterating until we reach a model where no predictors can be dropped. The second approach was to fit a model based on the predictors that we deemed useful from our EDA and use the same backwards elimination technique where we would drop predictors based on the results of the Wald's test until we refine it to a model in which all predictors would be deemed useful and are significant.

Method 1: Backward Elimination with Full Set of Predictors

We began our process of backward elimination by fitting a model that included all of the predictors present in our dataset. We then used the same iteration procedure as with white wines whereby we eliminate a single predictor in each step using the Wald test to determine which one to remove. We then refit the model with the remaining predictors and repeat the process until all predictors are deemed significant as indicated by a p-value of less than 0.05. The step-by-step results of this process can be seen in Table 9. The ROC Curves in Figures 23 and 24 show little difference between the first and final models.

This method resulted in dropping pH, free sulfur dioxide, citric acid, and density from the model. We show the summary of this model in Table 10. We then analyzed the key metrics in Table 11 for our final model by this method at various thresholds but we determined that sticking with a threshold of 50% was best for this model based on the needs of our clients. The model had a specificity of 0.973, indicating that 97.3% of low quality wines were correctly classified. The model also resulted in a sensitivity of .3008, meaning that 30.08% of high quality wines were correctly classified. The overall accuracy of the model was 87.88%. Most importantly, the precision of this model was 65.38%. We need to keep in mind that precision is of vital concern for the owners of Vinho Da 'Ville and must be prioritized above other metrics.

Table 9: Method 1 Iteration Results

Step	AUC	Confusion(.5)	Sensitivity	Specificity	Accuracy	Precision	dropped
0	0.8902758	FALSE TRUE 0 672 15 1 76 37	0.3274336	0.9781659	0.88625	0.7115385	
1	0.8902629	FALSE TRUE 0 672 15 1 77 36	0.3185841	0.9781659	0.885	0.7058824	pH
2	0.8900826	FALSE TRUE 0 672 15 1 77 36	0.3185841	0.9753266	0.885	0.7058824	free sulfur dioxide
3	0.8914222	FALSE TRUE 0 673 14 1 77 36	0.3185841	0.9796215	0.88625	0.72	citric acid
4	**0.8895802**	**FALSE TRUE 0 669 18 1 79 34**	**0.300885**	**0.9737991**	**0.87875**	**0.6538462**	**density**

Figures 23 and 24: Method 1 ROC Curves

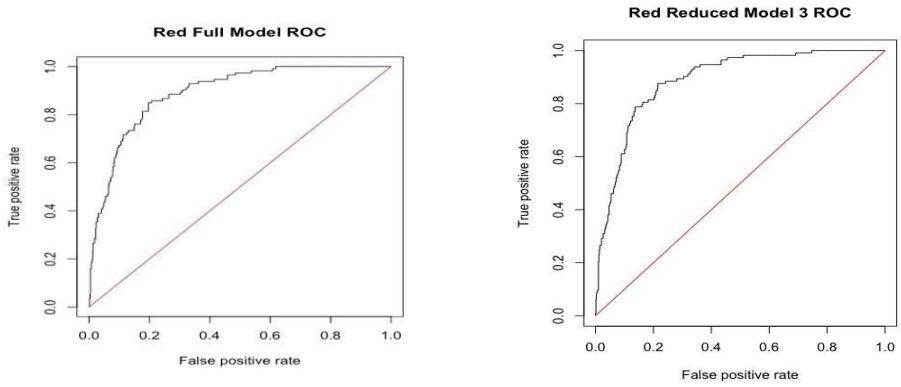


Table 10: Method 1 Logistic Regression Model

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -11.518339   1.770100  -6.507 7.66e-11 ***
alcohol               0.787720   0.113892   6.916 4.63e-12 ***
fixed.acidity         0.167830   0.067056   2.503 0.012321 *
volatile.acidity     -3.007320   0.860154  -3.496 0.000472 ***
residual.sugar        0.200280   0.081706   2.451 0.014237 *
chlorides           -13.176308   5.452411  -2.417 0.015666 *
total.sulfur.dioxide  -0.011942   0.004503  -2.652 0.007998 **
sulphates             3.067516   0.645272   4.754 2.00e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 11: Method 1 Key Metrics

Threshold	Confusion	Sensitivity	Specificity	Accuracy	Precision
.55	FALSE TRUE 0 678 9 1 87 26	0.2300885	0.9868996	0.88	0.7428571
.50	**FALSE TRUE** **0 669 18** **1 79 34**	**0.3008**	**0.973**	**0.87875**	**0.6538**
.45	FALSE TRUE 0 662 25 1 75 38	0.3362832	0.9636099	0.875	0.6031746

Method 2: Backward Elimination with Predictors from EDA

Based on the EDA above, we determined that the predictors Alcohol, Volatile Acidity, Citric Acid, Density, and Sulphates may be useful to predict the quality of the red wine. We followed a similar approach as in method 1 and with white wines, but started with just these five predictors before beginning the iteration process using the Wald test. We summarize these results in Table 12 and concluded that based on this method, only density should be dropped, which resulted in a model with just the predictors alcohol, citric acid, and sulphates. The ROC Curves in Figures 25 and 26 show little variation, which is

confirmed by the slight marginal increase in the AUC seen in Table 12. We analyzed various thresholds for this model in Table 14 and chose 50% because it balanced precision with quantity. The overall accuracy of the model was 87.75%, but as we have previously discussed the most important metric, precision, was 0.7435, indicating that among the wines classified as high quality, 74.35% of them actually had a high quality rating. We show the logistic regression equation summary for this model in Table 13.

Table 12: Method 2 Iteration Results

Step	AUC	Confusion(.5)	Sensitivity	Specificity	Accuracy	Precision	dropped
0	0.8839768	FALSE TRUE 0 678 9 1 91 22	0.1946	0.9869	0.875	0.7096774	
1	**0.8840541**	**FALSE TRUE** **0 679 8** **1 90 23**	**0.2035**	**0.9884**	**0.8775**	**0.7419355**	**density**

Figures 25 and 26: Method 2 ROC Curves

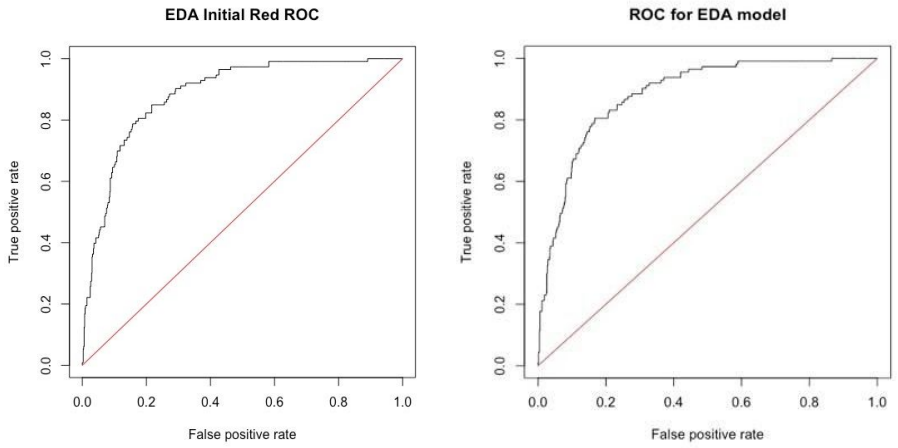


Table 13: Method 2 Logistic Regression Model

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -11.3906     1.4908  -7.641 2.16e-14 ***
alcohol            0.8113     0.1048   7.744 9.65e-15 ***
volatile.acidity  -2.6282     0.9312  -2.822 0.004765 **
citric.acid        1.7234     0.7311   2.357 0.018409 *
sulphates          2.0640     0.5552   3.717 0.000201 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Table 14: Method 2 Key Metrics

Threshold	Confusion	Sensitivity	Specificity	Accuracy	Precision
.55	FALSE TRUE 0 683 4 1 95 18	0.15929	0.99417	0.875	0.8181
.50	**FALSE TRUE** **0 679 8** **1 90 23**	**0.20353**	**0.9884**	**0.8775**	**0.7419355**
.45	FALSE TRUE 0 670 17 1 86 27	0.238938	0.9752547	0.87125	0.613636

Method 3: Adjusted Version of Method 2

After reviewing the results of the first two methods of model generation and weighing in the context, we decided to use a third method to build the model for red wines to then compare to the results of the previous two procedures. We were somewhat surprised that the results of our EDA for red wines did not include residual sugars as we thought this would be a more useful predictor than citric acid due to the sweetness playing a role in quality. Thus, we followed the same procedure as in method 2, but swapped citric acid for residual sugar before beginning the backward elimination process, which started with alcohol, volatile acidity, residual sugar, density, and sulphates. We show the results of this procedure in Table 15 and highlight that density and residual sugar were dropped. We were surprised to see that even by this process where we purposefully included residual sugar, it did not end up being significant in the model given that alcohol, volatile acidity, and sulphates were included. We again show the ROC Curves for the beginning and ending models of this method in Figures 27 and 28, but they appear quite similar. In Table 16, we show the summary of the logistic regression model by this approach.

We follow our standard process of exploring thresholds in Table 17 and we decide to set the threshold at 45%. While lowering the threshold actually reduces our precision to 73%, we must also consider that this increases the number of bottles of wine classified as high quality. This is a necessary tradeoff as we want to avoid significantly reducing the selection of high quality red wines served at Vinho Da 'Ville.

Table 15: Method 3 Iteration Results

Step	AUC	Confusion(.5)	Sensitivity	Specificity	Accuracy	Precision	dropped
0	0.8855741	FALSE TRUE 0 680 7 1 91 22	0.2300885	0.9898108	0.8775	0.7586207	
1	0.88645	FALSE TRUE 0 682 5 1 92 21	0.1858407	0.992722	0.87875	0.8076923	density
2	**0.8874934**	**FALSE TRUE** **0 682 5** **1 94 19**	**0.1681416**	**0.992722**	**0.87625**	**0.7916667**	**residual sugar**

Figures 27 and 28: Method 3 ROC Curves

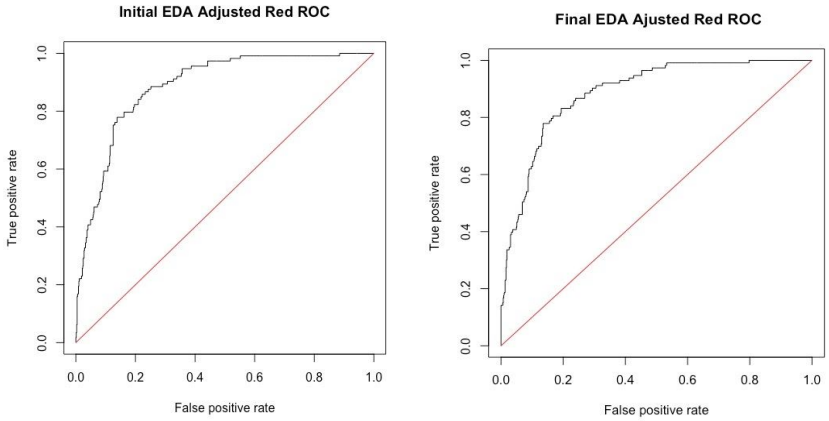


Table 16: Method 3 Logistic Regression Model

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -10.2873     1.3981  -7.358 1.87e-13 ***
alcohol            0.8003     0.1044   7.665 1.79e-14 ***
volatile.acidity  -3.7610     0.8152  -4.614 3.95e-06 ***
sulphates          2.2173     0.5521   4.017 5.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 17: Method 3 Key Metrics

Threshold	Confusion	Sensitivity	Specificity	Accuracy	Precision
0.45	**FALSE TRUE 0 676 11 1 83 30**	**0.2655**	**0.98399**	**0.8825**	**0.7317**
0.50	FALSE TRUE 0 682 5 1 94 19	0.1681416	0.992722	0.87625	0.7916667
0.55	FALSE TRUE 0 685 2 1 97 16	0.1416	0.997	0.87625	0.888

Model Selection

Table 18: Comparison of Methods 1, 2 and 3 Results

Model	Threshold	Confusion	Sensitivity	Specificity	Accuracy	Precision
Method 1: Full	.50	FALSE TRUE 0 669 18 1 79 34	0.3008	0.973	0.87875	0.6538
Method 2: EDA	.50	FALSE TRUE 0 679 8 1 90 23	0.2035	0.9884	0.8775	0.7419355
Method 3: EDA Adjusted	**0.45**	**FALSE TRUE 0 676 11 1 83 30**	**0.2655**	**0.98399**	**0.8825**	**0.7317**

In Table 18, we show a summary of the key metrics of the three best models that resulted from each of the three model generation methods for red wines. Of these three models, we can see that the second has the highest precision at 74%. Nevertheless, we must remember that maximizing precision must be balanced against the number of wines classified as high quality and that it can also make some true positive results false negatives. We can then see that for just a 1% drop in precision, the third model adds ten bottles of wine back into the high quality classification. The benefit of this tradeoff leads us to choose the third model as our final model for predicting whether a red wine is high or low quality.

Conclusion

The final selected models for both red and white wines are effective predictors of whether a given wine tested will be categorized as high quality, defined as a quality rating of 7 or higher. By employing the results of our exploratory data analysis, we were able to arrive at a better predictive model than when starting with a full model for both red and white wines. When testing our models, we optimized the thresholds to maximize precision, while still leaving a sufficient population of wines. This precision was stated as most important to our client, defined as the ratio of true positives to total assessed positives, expressed as *TP/(TP+FP)*. Based on our selected models, we have achieved a precision of 65.9% with white wines and 73.2% with reds as can be seen in Table 19.

With our data analysis, the owners of Vinho Da 'Ville will be well equipped to speak on the similarities and differences of red and white Portuguese wines as well as the characteristics that define each. Similarly, the implementation of our models will give the owners confidence that they are serving high quality wine to their impressive clientele. All the owners must do is read this report to become the newest wine connoisseurs of Charlottesville.

Table 19: Final Results

Type	Predictors	Threshold	Confusion	Sensitivity	Specificity	Accuracy	Precision
White	Alcohol Residual Sugar Chlorides Density	.55	FALSE TRUE 0 1848 58 1 431 112	0.2062615	0.9695698	0.8003267	**0.658823**
Red	Alcohol Volatile Acidity Sulphates	.45	FALSE TRUE 0 676 11 1 83 30	0.2654867	0.9839884	0.8825	**0.731707**