Quality of Wine

By Angel Franco, Thisuri Molligoda, Karl Nika, Jordan Wesseln

Abstract

In our collegiate search of the best possible wine to consume on "Wine Wednesday," we wanted to determine what made a bottle "good". For the sake of our research, we decided that we would hold our standards slightly higher than average. If wine quality was 7 or above on a scale of 0 to 10, it's considered "good wine". From there we can analyze the data in order to determine what predictors contribute the most to making one bottle of wine better than another. Since different members of the group have different preferences, we also wanted to see if the same factors that make white wine "good", make red wine "good".

Introduction

Our dataset is the Wine Quality dataset from UCI Machine Learning Repository with a sample size of around 6500 observations. We have 11 predictors with the first being fixed acidity which are tartaric acid, malic acid, citric acid, and succinic acid. Volatile acidity refers to the steam distillable acids present in wine, primarily acetic acid but also lactic, formic, butyric, and propionic acids. Citric acid is the acid added to wine to add acidity, complement a flavor, or prevent ferric hazes, and makes the wine seem fresher. Residual sugar is from natural grape sugars leftover in a wine after the alcoholic fermentation finishes. Chlorides are the amount of salt in the wine. Free sulfur dioxide is the free form of sulfur dioxide which exists in equilibrium between molecular sulfur dioxide and bisulfite ion, it prevents microbial growth and the oxidation of the wine. Total sulfur dioxide is the amount of free and bound forms of sulfur dioxide which depending on the concentration can influence the smell and taste of the wine. The density of the wine is close to water depending on the percentage of alcohol and the amount of sugar. pH describes how acidic or basic the wine is. Sulfates are wine additives that contribute to the sulfur dioxide and act as an antimicrobial agent and an antioxidant. Alcohol is the percentage of alcohol in the wine. Color is the color of the wine which is a binary variable with the wine being red or white, all the other variables are continuous. The response variable is quality, which is the perceived quality of the wine, and it is a binary variable comparing good and bad wine. Our primary focus of this project was to check the significance of each predictor mentioned above against the quality of the two types of wine in order to come up with the best model for the dataset.

Methodology

Using statistical software (SAS), we calculated the mean and the standard deviation for each predictor variable mentioned above, against the response variable for both types of wine which was the Quality of Wine. The Wine Type which is shown in the table below (figure 1), was not a variable used in the calculations but simply used to compare the findings between Red and White Wine. We then used T-Tests to estimate p-values for the overall quality of wine for each predictor variable. We also created a correlation matrix to see how the predictors were related and checked for multicollinearity. Since our response variable is binary, we decided to run a logistic regression model on our data so we could see predictor p-values and use the results

to create a model for our data. Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome that is measured with a dichotomous variable. We used the Stepwise variable selection method to help us remove any predictors from our model that were insignificant to predict the result of the response variable(quality). We checked interactions between predictors to see how the effects of the predictors would change compared to other predictor levels. Lastly, we created a residual plot and ran goodness of fit tests to see the accuracy of our model and to whether we should be using a linear model or move on to something more complex.

Exploratory Analysis

In our exploratory analysis, we first compared the predictors in our model between those of high quality (those with a quality rating of 7 or greater), to those of low quality (rating of below 7), to get an initial idea of the relationship between the response variable which is the quality of wine. We ran individual T-tests for each predictor against the response variable and calculated the confidence intervals which is shown in the following graph (Figure 1). A few observations that jumped out at us were that alcohol and density had very distinct intervals between high and low quality and when each was plotted against quality it was bimodal. The variability of the data is notably high for most predictors, and we could see that all variables are significant for the change in the quality of wine except Free Sulfur Dioxide which had a p-value of 0.1859.

Predictors	Wine Type	High Quality=1	Low Quality=0	Overall Quality		P-value
				High=1	Low=0	
Fixed Acidity	White	6.7251 ± 0.7688	6.8906 ± 0.8601	7.0857 ± 1.3428	7.2470 ± 1.2830	0.0001
	Red	8.8470 ± 2.000	8.2368 ± 1.6827			
Volatile Acidity	White	0.2653 ± 0.0941	0.2818 ±0.1023	0.2892 ± 0.1170	0.3520 ± 0.1721	<0.0001
	Red	0.4055 ± 0.1450	0.5470 ± 0.1763			
Citric Acid	White	0.3261 ± 0.0803	0.3364 ± 0.1300	0.3346 ± 0.1100	0.3147 ± 0.1525	<0.0001
	Red	0.3765 ± 0.1944	0.2544 ± 0.1897			
Residual Sugar	White	5.2615 ± 4.2908	6.7035 ± 5.2250	4.8277 ± 4.0638	5.5938 ± 4.9013	<0.0001
	Red	2.7088 ± 1.3630	2.5121 ± 1.4158			
Chlorides	White	0.0382 ± 0.0111	0.0479 ± 0.0235	0.0446 ± 0.0210	0.0588 ± 0.0371	<0.0001
	Red	0.0759 ± 0.0285	0.0893 ± 0.0491			
Free Sulfur Dioxide	White	34.5505 ± 13.7971	35.5173 ± 17.7878	31.0552 ± 15.3442	30.3957 ± 18.2887	0.1859
	Red	13.9816 ± 10.2346	16.1722 ± 10.4677			
Total Sulfur Dioxide	White	125.2 ± 32.7248	142.0 ± 44.1454	109.9 ± 47.1262	117.2 ± 58.5064	<0.0001
	Red	34.8894 ± 32.5722	48.2858 ± 32.5856			
Density	White	0.9924 ± 0.00277	0.9945 ± 0.00289	0.9930 ± 0.00301	0.9951 ± 0.00285	<0.0001
	Red	0.9960 ± 0.00220	0.9969 ± 0.00181			
pH	White	3.2151 ± 0.1572	3.1808 ± 0.1484	3.2277 ± 0.1591	3.2163 ± 0.1611	0.0223
	Red	3.2888 ± 0.1545	3.3146 ± 0.1541			
Sulphates	White	0.5001 ± 0.1330	0.4870 ± 0.1082	0.5415 ± 0.1615	0.5288 ± 0.1454	0.0102
	Red	0.7435 ± 0.1340	0.6448 ± 0.1706			
Alcohol	White	11.4160 ± 1.2552	10.2652 ± 1.1006	11.4334 ± 1.2156	10.2615 ± 1.6074	<0.0001
	Red	11.5180 ± 0.9982	10.2510 ± 0.9697			

Figure 1

Correlation matrix (Combined dataset)

The correlation matrix below shows the relationship between each predictor for the total wine data set. Although this may be slightly skewed because Red and White wines seem to have different main predictors, we can still see some high correlation trends in the table. These include the relationship between Residual Sugar and Density, Total Sulfur Dioxide and Free Sulfur Dioxide and between Density and Alcohol which are highlighted below in yellow. The alcohol and density having a negative relationship make sense, as alcohol is less dense than water, so we expect the wine to become less dense as there is more alcohol. The total and free sulfur dioxide would be expected to have a positive correlation. As more sulfur, in general, is going to increase the amounts seen for both of those predictors. We expect to see the overall density increase as more sugar is added. This is expected from the sugar dissolving into the wine, making it denser than wine without added residual sugars.

Pearson Correlation Coefficients, N = 6497

Prob > |r| under H0: Rho=0

	FxdAcid	VolAcid	CitAcid	ResSug	Chlo	FreeSD	TotalSD	Dens	PH	Sulph	Alcoh
FxdAcid	1	0.21901 <.0001	0.32444 <.0001	-0.11198 <.0001	0.29819 <.0001	-0.28274 <.0001	-0.32905 <.0001	0.45891 <.0001	-0.2527 <.0001	0.29957 <.0001	-0.09545 <.0001
VolAcid	0.21901 <.0001	1	-0.37798 <.0001	-0.19601 <.0001	0.37712 <.0001	-0.35256 <.0001	-0.41448 <.0001	0.2713 <.0001	0.26145 <.0001	0.22598 <.0001	-0.03764 0.0024
CitAcid	0.32444 <.0001	-0.37798 <.0001	1	0.14245 <.0001	0.039 0.0017	0.13313 <.0001	0.19524 <.0001	0.09615 <.0001	-0.32981 <.0001	0.0562 <.0001	-0.01049 0.3977
ResSug	-0.11198 <.0001	-0.19601 <.0001	0.14245 <.0001	1	-0.12894 <.0001	0.40287 <.0001	0.49548 <.0001	0.55252 <.0001	-0.26732 <.0001	-0.18593 <.0001	-0.35941 <.0001
Chlo	0.29819 <.0001	0.37712 <.0001	0.039 0.0017	-0.12894 <.0001	1	-0.19504 <.0001	-0.27963 <.0001	0.36261 <.0001	0.04471 0.0003	0.39559 <.0001	-0.25692 <.0001
FreeSD	-0.28274 <.0001	-0.35256 <.0001	0.13313 <.0001	0.40287 <.0001	-0.19504 <.0001	1	0.72093 <.0001	0.02572 0.0382	-0.14585 <.0001	-0.18846 <.0001	-0.17984 <.0001
TotalSD	-0.32905 <.0001	-0.41448 <.0001	0.19524 <.0001	0.49548 <.0001	-0.27963 <.0001	0.72093 <.0001	1	0.03239 0.009	-0.23841 <.0001	-0.27573 <.0001	-0.26574 <.0001
Dens	0.45891 <.0001	0.2713 <.0001	0.09615 <.0001	0.55252 <.0001	0.36261 <.0001	0.02572 0.0382	0.03239 0.009	1	0.01169 0.3463	0.25948 <.0001	-0.68675 <.0001
PH	-0.2527 <.0001	0.26145 <.0001	-0.32981 <.0001	-0.26732 <.0001	0.04471 0.0003	-0.14585 <.0001	-0.23841 <.0001	0.01169 0.3463	1	0.19212 <.0001	0.12125 <.0001
Sulph	0.29957 <.0001	0.22598 <.0001	0.0562 <.0001	-0.18593 <.0001	0.39559 <.0001	-0.18846 <.0001	-0.27573 <.0001	0.25948 <.0001	0.19212 <.0001	1	-0.00303 0.8071
Alcoh	-0.09545 <.0001	-0.03764 0.0024	-0.01049 0.3977	-0.35941 <.0001	-0.25692 <.0001	-0.17984 <.0001	-0.26574 <.0001	-0.68675 <.0001	0.12125 <.0001	-0.00303 0.8071	1

Figure 2

The results of the correlation matrix are as follows.

- The moderately positive correlation of 0.553 between Residual Sugar and Density.
- The strong positive correlation of 0.721 between Total Sulfur Dioxide and Free Sulfur Dioxide which means one of the predictors can be removed from our final model.
- A fairly strong negative correlation of -0.687 between Density and Alcohol.

Regression Analysis (White, Red and Combined dataset)

In our regression analysis, we ran a logistic regression model for the quality of the wine against all predictors to get an idea of the effect of each predictor. The following tables show the results of the regression analysis with the 'Estimate' column giving us the effect of the predictor towards the response variable. For example, the estimate for volatile acid for White wine is -3.785 which tells us that there is a negative relationship between the quality and volatile acid (highlighted in blue, figure 3.1). What this means is that when the amount of volatile acidity increases in the wine, the quality of wine decreases considerably. We can see that the Wald test statistic for this particular variable is quite large at 60.04 and a p-value of approximately zero which tells us that this particular variable is highly significant in predicting the quality of the wine. Another predictor worth noting is the level of Alcohol in White wine. The estimate shows that there is a positive relationship with the quality of the wine, however, it is not a significant effect since the Wald test statistic is small resulting in a large p-value (highlighted in purple, figure 3.1).

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	636.2	94.1147	45.6915	<.0001
FxdAcid	1	0.5521	0.0905	37.1913	<.0001
VolAcid	1	-3.7845	0.4884	60.035	<.0001
CitAcid	1	-0.7376	0.401	3.3839	0.0658
ResSug	1	0.2952	0.0356	68.616	<.0001
Chlo	1	-12.6246	3.8155	10.9477	0.0009
FreeSD	1	0.00864	0.00313	7.628	0.0057
TotalSD	1	-0.00027	0.00151	0.0322	0.8576
Dens	1	-659.1	95.3939	47.7365	<.0001
PH	1	3.3429	0.4268	61.3364	<.0001
Sulph	1	2.1677	0.3475	38.9152	<.0001
Alcoh	1	0.1423	0.1139	1.5621	0.2114

Figure 3.1 White wine

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	242.8	108.1	5.0475	0.0247
FxdAcid	1	0.275	0.1253	4.8168	0.0282
VolAcid	1	-2.581	0.7843	10.8298	0.001
CitAcid	1	0.5678	0.8385	0.4585	0.4983
ResSug	1	0.2395	0.0737	10.5474	0.0012
Chlo	1	-8.8163	3.3649	6.8648	0.0088
FreeSD	1	0.0108	0.0122	0.7822	0.3765
TotalSD	1	-0.0165	0.00489	11.4093	0.0007
Dens	1	-257.8	110.4	5.4528	0.0195
PH	1	0.2242	0.9984	0.0504	0.8223
Sulph	1	3.7499	0.5416	47.9397	<.0001
Alcoh	1	0.7533	0.1316	32.7644	<.0001

Figure 3.2 Red wine

Therefore, analyzing the rest of the predictors we can see that the variables Citric acid, Total Sulfur Dioxide (highlighted in yellow, figure 3.1) and Alcohol are insignificant variables and that they should be removed from our final model. Using the same analysis for the Red Wine, we see that the variables Citric acid, Free Sulfur Dioxide, and PH levels are insignificant to predict the quality of the wine (figure 3.2).

Lastly, for the combined dataset of White and Red wine, we can see that there is only one variable that is insignificant (highlighted in yellow, figure 3.3), therefore, we can conclude from the regression analysis that all other variables are significant to predict the change in the quality of wine for the combined dataset. The results of the regression analysis are given below (figure 3.3).

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	282.9	52.4543	29.0909	<.0001
FxdAcid	1	0.4382	0.0642	46.5916	<.0001
VolAcid	1	-3.3615	0.3741	80.7498	<.0001
CitAcid	1	-0.3201	0.345	0.8606	0.3536
ResSug	1	0.1761	0.0219	64.5189	<.0001
Chlo	1	-6.0932	2.3331	6.8204	0.009
FreeSD	1	0.0127	0.00289	19.3767	<.0001
TotalSD	1	-0.00587	0.00116	25.6714	<.0001
Dens	1	-303.8	53.5043	32.2313	<.0001
PH	1	2.3871	0.352	45.9775	<.0001
Sulph	1	2.4853	0.2848	76.1674	<.0001
Alcoh	1	0.5791	0.0695	69.3714	<.0001

Figure 3.3 Total Wine

Stepwise Variable Selection (Red, White and Combined dataset)

Summary of Stepwise Selection							
Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	Alcoh		1	1	265.2827		<.0001
2	Sulph		1	2	68.5137		<.0001
3	VolAcid		1	3	52.219		<.0001
4	TotalSD		1	4	15.3862		<.0001
5	Chlo		1	5	6.795		0.0091
6	FxdAcid		1	6	7.9181		0.0049
7	ResSug		1	7	4.8622		0.0275
8	Dens		1	8	6.9294		0.0085

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	226.8	91.6292	6.1239	0.0133
FxdAcid	1	0.2812	0.0803	12.2635	0.0005
VolAcid	1	-2.9128	0.6467	20.2872	<.0001
ResSug	1	0.2328	0.0701	11.0364	0.0009
Chlo	1	-8.4408	3.2589	6.7084	0.0096
TotalSD	1	-0.0136	0.00345	15.5702	<.0001
Dens	1	-240.9	92.022	6.8558	0.0088
Sulph	1	3.6987	0.5287	48.9506	<.0001
Alcoh	1	0.7823	0.112	48.7661	<.0001

Figure 4.1 Red wine

Stepwise variable selection was done in order to effectively conclude which of the predictors has the largest effect on the quality of the wine, for red, white, and combined. We see that Citric Acid, Free Sulfur Dioxide, and PH are removed from the contributing factors of red wine (Figure 4.1). From the final model table given above, we can see that Sulphates and Alcohol (highlighted in pink) were the most significant towards the change in Quality of wine after all predictors have been accounted for.

When the Stepwise variable selection was conducted for the White wine, we can see that Alcohol, Total Sulfur Dioxide, and Citric Acid are removed from the list of main predictors of White wine quality (figure 4.2). We should also note that Alcohol was first entered into the model as it was the most significant variable that was associated with the quality of wine, but was removed when Fixed Acidity was entered into the model. We can conclude from the table below (figure 4.2) that the most significant variables were Residual Sugar, Density and PH Levels (highlighted in orange) after all other predictors have been accounted for.

Summary of Stepwise Selection

Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	Alcoh		1	1	726.5025		<.0001
2	VolAcid		1	2	72.3505		<.0001
3	ResSug		1	3	27.7528		<.0001
4	PH		1	4	24.4204		<.0001
5	Chlo		1	5	21.3042		<.0001
6	Sulph		1	6	15.2229		<.0001
7	Dens		1	7	21.8136		<.0001
8	FxdAcid		1	8	36.9885		<.0001
9		Alcoh	1	7		0.8612	0.3534
10	FreeSD		1	8	10.9362		0.0009

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	740.5	34.4846	461.0643	<.0001
FxdAcid	1	0.6006	0.0635	89.4718	<.0001
VolAcid	1	-3.5493	0.4601	59.5036	<.0001
ResSug	1	0.3298	0.0193	293.5574	<.0001
Chlo	1	-12.4747	3.773	10.9314	0.0009
FreeSD	1	0.00805	0.00247	10.6193	0.0011
Dens	1	-764.5	35.5811	461.5997	<.0001
PH	1	3.6571	0.3301	122.7326	<.0001
Sulph	1	2.2661	0.3284	47.6181	<.0001

Figure 4.2 White wine

When also ran the Stepwise selection on our combined wine dataset and see-saw that all predictors besides Citric Acid were added into the model. Once added, none of the predictors were then removed from the model. From the Wald scores, we can see that Volatile Acidity, Sulfates, Alcohol Levels, and Residual Sugar (highlighted in blue) were the most significant in predicting the change in Quality of wine in this test.

Summary of Stepwise Selection

Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	Alcoh		1	1	990.6582		<.0001
2	VolAcid		1	2	162.1664		<.0001
3	Sulph		1	3	50.5235		<.0001
4	ResSug		1	4	38.2431		<.0001
5	Chlo		1	5	12.4121		0.0004
6	TotalSD		1	6	9.3381		0.0022
7	FreeSD		1	7	20.6843		<.0001
8	PH		1	8	6.8079		0.0091
9	FxdAcid		1	9	14.221		0.0002
10	Dens		1	10	31.7853		<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	284.4	52.3598	29.5072	<.0001
FxdAcid	1	0.424	0.0623	46.3645	<.0001
VolAcid	1	-3.2393	0.3497	85.8029	<.0001
ResSug	1	0.1758	0.0219	64.4401	<.0001
Chlo	1	-6.2357	2.3194	7.2282	0.0072
FreeSD	1	0.0129	0.00289	19.861	<.0001
TotalSD	1	-0.00602	0.00115	27.5859	<.0001
Dens	1	-305.2	53.4103	32.6581	<.0001
PH	1	2.3967	0.3516	46.4616	<.0001
Sulph	1	2.4743	0.2845	75.6333	<.0001
Alcoh	1	0.5708	0.0689	68.7124	<.0001

Figure 4.3 Total Wine

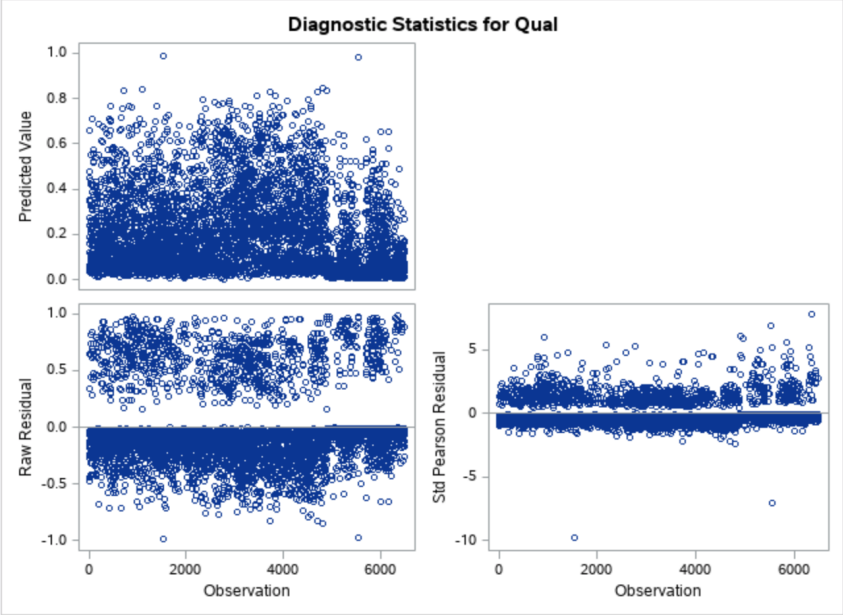
Significance of Interactions

In this section, we tested the significance of interactions towards the quality of wine for each of Red and White Wine datasets. We saw in the previous section certain predictors showing more significance towards the change in Quality of wine compared to other predictors that were entered into the model. We used these significant predictors and created interactions amongst them to see if it resulted in a more accurate model. The interactions for the datasets are as follows,

- *Red Wine*
  - *Sulph*Alcoh*
- *White Wine*
  - *ResSug*Dens*
  - *ResSug*PH*
  - *Dens*PH*

We repeated the Stepwise variable selection method with the interactions included for the White Wine and we were able to find that only the interaction of Residual Sugar and PH with a p-value of less than 0.0001 was significant towards the change in the quality of wine. All other predictors remained significant even after adding the interaction variables. For the Red Wine, we used the two most significant variables and checked their interaction. Once the interaction term was added, Alcohol became insignificant with a p-value of 0.9848 and Sulphates decently significant with a p-value of 0.0363. From this result, we can conclude that this particular interaction term did not help make the model for Red Wines more accurate. The results of this test is provided in the reference pages (page 12).

Residual plots



Hosmer and Lemeshow GOF Tests

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
47.5852	8	<.0001

Red and White Wine GOF

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
9.7635	8	0.2820

Red Wine GOF

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
29.6534	8	0.0002

White Wine GOF

With our residual plots, we see that most of our predictions were between two standard deviations away from the actual result. There are some outliers though. When we compare that with the goodness of fit tests, we can see that our model is predicted well for Red wine, but not for White wine. Since we are using a combined White and Red wine dataset, it would make sense that our model would not be very accurate for both types because they have many factors that are quite different depending on the type of wine.

The Hosmer and Lemeshow Goodness-of-Fit tests show that only the Red Wine model has a good fit since it has a low chi-squared value and a p-value higher than 0.05. A possible explanation for why the Red and White model was not a good fit is that we combined the data for Red and White wine when each was its own data set. As for White Wine, it is possible that the same predictors that make Red Wine good don't make White Wine good, and a different set of predictors might help make a better model. Overall it's nice to see that at least one of the models has a good fit.

Conclusions

The predictors that had the highest effect on quality for ***red wines*** are ***alcohol*** and ***sulfates***. For ***white wines***, the most significant predictors are ***residual sugar***, ***density***, and ***pH***. For the ***mixed wine*** dataset, the most significant predictors are ***volatile acidity***, ***sulfates***, and ***alcohol***. We have a ***strong negative relationship between alcohol and density*** and a ***strong positive relationship also between density and residual sugar.*** When we tested the interactions for our most significant predictors, the main effects became insignificant. Therefore, we conclude that it is quite difficult to find the best combination of predictors that is significantly associated with the higher Quality of Wine. Since an overwhelming majority of our wines had quality levels between 5-7, having more uniform quality levels would be needed to further assess the question of which of our predictors contribute the most to its quality.

References and Appendix

Sources

Agresti, Alan. *An Introduction to Categorical Data Analysis*. 3rd ed., Wiley, 2019.

https://www.medcalc.org/manual/logistic_regression.php

SAS Code

```
Data WineQualityWhite;
Infile '/folders/myfolders/Crime/Wine Quality.csv'delimiter=','firstobs=2;
Input FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD Dens PH Sulph Alcoh Quality;
If Quality>7 or Quality=7 then Qual=1;
else Qual=0;
run;
        Proc logistic;
        Model Qual(event="1")=FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD Dens
        PH Sulph Alcoh;
        run;
        Proc logistic;
        Model Qual(event="1")=FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD
        Dens PH Sulph Alcoh /selection=stepwise;
        run;
        Proc logistic;
        Model Qual(event="1")=FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD
        Dens PH Sulph Alcoh Dens*ResSug Dens*PH ResSug*PH / selection=stepwise;
        run;
                Proc ttest data=WineQualitywhite;
                Class Qual;
                Var FxdAcid VolAcid CitAcid;
                run;
                Proc ttest data=WineQualitywhite;
                Class Qual;
                Var ResSug Chlo FreeSD TotalSD;
                run;
                Proc ttest data=WineQualitywhite;
                Class Qual;
                Var Dens PH Sulph Alcoh;
                run;
Data WineQualityRed;
Infile '/folders/myfolders/Crime/winequality-red.csv'delimiter=','firstobs=2;
Input FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD Dens PH Sulph Alcoh Quality;
If Quality>7 or Quality=7 then Qual=1;
Else Qual=0;
```

```
run;
        Proc logistic;
        Model Qual(event="1")=FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD
        Dens PH Sulph Alcoh;
        Run;
        Proc logistic;
        Model Qual(event="1")=FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD
        Dens PH Sulph Alcoh /selection=stepwise;
        run;
        Proc logistic;
        Model Qual(event="1")=FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD
        Dens PH Sulph Alcoh Sulph*Alcoh / selection=stepwise;
        run;
                Proc ttest data=WineQualityred;
                Class Qual;
                Var FxdAcid VolAcid CitAcid;
                run;
                Proc ttest data=WineQualityRed;
                Class Qual;
                Var ResSug Chlo FreeSD TotalSD;
                run;
                Proc ttest data=WineQualityred;
                Class Qual;
                Var Dens PH Sulph Alcoh;
                run;
Data WineQualityWhiteRed;
Infile '/folders/myfolders/Crime/Wine Quality White+Red.csv'delimiter=','firstobs=2;
Input FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD Dens PH Sulph Alcoh Quality;
If Quality>7 or Quality=7 then Qual=1;
Else Qual=0;
Run;
        Proc    corrdata=WineQualityWhiteRed       nomissPlots=matrix(NVAR=10)
        Plots(maxpoints=none);
        Var FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD Dens PH Sulph Alcoh;
        run;
        Proc logistic;
        Model Qual(event="1")=FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD Dens
        PH Sulph Alcoh;
        run;
        Proc logistic;
        Model Qual(event="1")=FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD
        Dens PH Sulph Alcoh /selection=stepwise;
        run;
        odsnoproctitle;
        odsgraphics/imagemap=on;
```

```
proc genmod data=WORK.WINEQUALITYWHITERED descending
plots=(predicted
resraw(index)stdreschi(index));
modelQual=FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD Dens
PH Sulph
Alcoh /dist=binomial link=logit;
run;
        procttestdata=WineQualityWhiteRed;
        classQual;
        varFxdAcid VolAcid CitAcid;
        run;
        procttestdata=WineQualityWhiteRed;
        classQual;
        varResSug Chlo FreeSD TotalSD;
        run;
        procttestdata=WineQualityWhiteRed;
        classQual;
        varDens PH Sulph Alcoh;
        run;
```

The following code is to get the Hosmer-Lemeshow GOF test:

```
proc logistic;
model Qual(event="1")=FxdAcid VolAcid CitAcid ResSug Chlo FreeSD TotalSD Dens PH
Sulph Alcoh/ lackfit;
run;
```

Significance of Interactions (From page 7):

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	758.3	35.0672	467.6384	<.0001
FxdAcid	1	0.6436	0.0642	100.6283	<.0001
VolAcid	1	-3.5495	0.4625	58.9052	<.0001
ResSug	1	1.7181	0.1858	85.5025	<.0001
Chlo	1	-12.1083	3.8182	10.0567	0.0015
FreeSD	1	0.0085	0.00249	11.6004	0.0007
Dens	1	-790.4	36.3066	473.9688	<.0001
PH	1	6.0116	0.4598	170.9142	<.0001
Sulph	1	2.289	0.3338	47.0291	<.0001
ResSug*PH	1	-0.4344	0.0578	56.4306	<.0001

White Wine

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.5674	4.523	0.6221	0.4303
VolAcid	1	-3.3442	0.6385	27.4342	<.0001
ResSug	1	0.1357	0.0608	4.9892	0.0255
Chlo	1	-9.2696	3.6572	6.4243	0.0113
TotalSD	1	-0.0143	0.00352	16.3913	<.0001
PH	1	-1.6992	0.6359	7.1403	0.0075
Sulph	1	-11.3794	5.4342	4.3849	0.0363
Alcoh	1	0.00692	0.363	0.0004	0.9848
Sulph*Alcoh	1	1.3999	0.5109	7.508	0.0061

Red Wine

Individual Report

Some of the difficulties early on were trying to figure out what to do with the data, whether we wanted to only do red or white, combine them, or do both separately. We ended up doing the last two since we wanted to know what was different between the wines and what would happen to the wine overall. One other problem was making the dataset readable since it was in one column for each row, so we had to make each point its own cell.

The way we interacted was pretty normal for a group project, we set a day to work on the proposal together and made a group chat to talk about what we need to do for the project. We mostly worked individually except for some parts of the presentation where we needed to see each other face to face. Using Google docs and slides helped since it has a way to chat with the people that are editing at the same time. It was a good learning experience since it was my first project in a statistics class and it was good to know how everyone has different methodologies for doing a project like this. I feel that everyone in the project did what they had to do and everyone put in an equal amount of work towards the project. For the presentation I talked the least and I feel like I could’ve expanded on my slides but got stage fright and justed talked really fast.