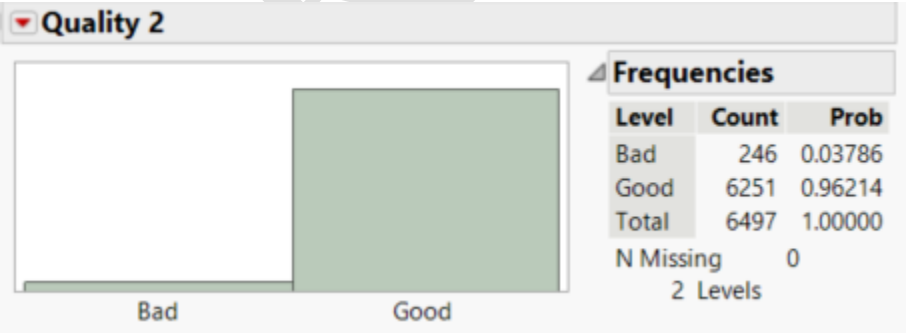
Introduction

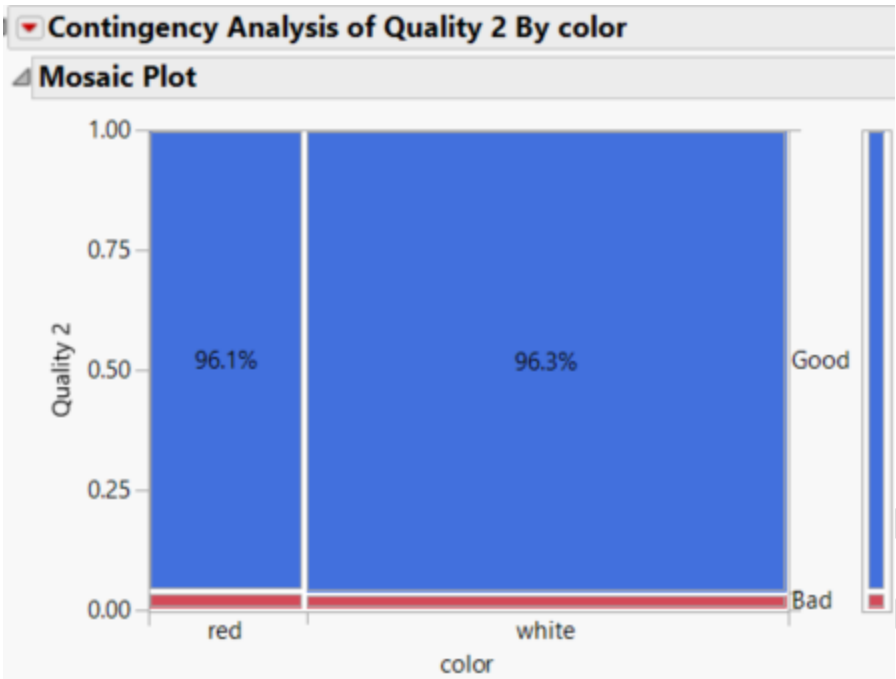
Can we determine if a population will enjoy a wine before they even taste it? What attributes make a wine good or bad? I will run a logistic regression on a dataset to see if a wine's quality can be predicted. This dataset includes 11 characteristics that make a wine unique.

A range of acidic qualities of wine are compiled in this dataset. It contains the level of fixed acidity in the wine, which does not quickly evaporate from the wine as well as volatile acidity. High levels of volatile acidity can lead to the wine tasting similar to vinegar. Citric acid is used to add a fresh quality to wine or to give it a little flavor.

Along with acid, this dataset consists of a range of sulfur dioxides. Free sulfur dioxide prevents bacterial growth in the wine and inhibits the wine from oxidizing. Total sulfur dioxide is the combination of free and bound sulfur dioxide in the wine. The sulfur dioxide should not be noticeable in a wine, but if there is too much free sulfur dioxide it can become perceptible in the smell and taste of the wine. Sulphates are also added to wine to give it antioxidant properties as well as add to the prevention of microorganism growth in the wine.

Other important qualities in wine that sets each one apart is the wine's density, pH level, and alcohol content. The density of wine is normally close to that of water while its pH level stays around three to four, on the more acidic side of the scale. The amount of sugar and salt in wine are also crucial traits. This dataset includes the amount of residual sugar in the wine, or the amount that remains after fermentation is complete. Most wines have at least a little bit of sugar left, but too much can cause the wine to be very sweet. Chlorides are the amount of salt in the wine.





Contingency Table

color	Count / Total % / Col % / Row %	Bad	Good	Total
red		63	1536	1599
		0.97	23.64	24.61
		25.61	24.57	
		3.94	96.06	
white		183	4715	4898
		2.82	72.57	75.39
		74.39	75.43	
		3.74	96.26	
Total		246	6251	6497
		3.79	96.21	

The majority of this dataset are wines that are of good quality. Out of 6,497 observations, 96 percent are of good quality while almost four percent are bad. If we divide the dataset into red wines and white wines, we can also see that there are more white wines in this dataset than red wines. Seventy-five percent of the observations are white wines. Both have a much higher percentage of good wines than bad.

Analysis and Model Comparison

For this predictive analysis, I will use JMP's validation column. Since this dataset does not include time series data, I can employ JMP's ability to randomly split the data into three parts for me. I will designate a 60 percent split for the training data, 20 percent for validation, and the last 20 percent for testing data. I will set the random seed to 123 so this analysis has the ability to be replicated.

After creating a validation column, I will perform five logistic regressions and compare their abilities on test data to choose the best model. First, I will run a nominal logistic that estimates on maximum likelihood as a benchmark to compare against the four penalized logistic regressions. Next, I will complete the penalized regressions of LASSO, adaptive LASSO, elastic net, and adaptive elastic net. All of these models will predict the probability of the wine being good or bad based off the qualities of the wine.

Several of the variables are highly correlated, such as different acidity levels and the pH level of the wine or sugar levels and color of the wine. Due to this, elastic net may prove to be a better model as it will keep all variables with predictive powers whereas LASSO will only keep the one it believes to be informative. LASSO will penalize uninformative variables using $|\beta|$ while elastic net uses both $|\beta|$ and β^2. A big advantage to LASSO is its ability to shrink uninformative variables all the way to zero. Other penalized regressions, like ridge, can only make those variables smaller, not take them all the way down to zero. Elastic net is a combination of LASSO and ridge. The adaptive models for LASSO and elastic net use the weights from the initial maximum likelihood estimates. These adaptive versions can have better asymptotic statistical properties.

To compare how well all five of these models perform on this wine quality dataset, I will look at several factors. While the R^2 value is not informative with categorical data, we still want a higher value for both entropy and generalized R^2. The best model should have a lower Root Mean Squared Error (RMSE), or Brier score. The two more valuable model results to look at are the misclassification rate and most importantly, the area under the curve (AUC).

The misclassification rate measures the overall success of the model. A lower rate means the model is doing better at classifying than models with higher rates. A Receiver Operating Characteristic (ROC) curve plots the model's ability to correctly classify. A model's true positive rate, or ability to predict positives correctly, is on the y-axis and its false positive rate, or amount the model classifies incorrectly, is plotted on the x-axis. The closer the curve is to the y-axis, the

more successful the model because that means it produces less false positives, or incorrect classifications. The AUC, which will be used to determine which model is best, is the measure of this area under the ROC curve.

Measures of Fit for Quality 2

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Fit Nominal Logistic		0.1474	0.1684	0.1376	0.1805	0.0657	0.0372	3898	0.7819
Training	Fit Generalized Lasso		0.1474	0.1684	0.1376	0.1805	0.0657	0.0372	3898	0.7819
Training	Fit Generalized Adaptive Lasso		0.1474	0.1684	0.1376	0.1805	0.0657	0.0372	3898	0.7819
Training	Fit Generalized Elastic Net		0.1474	0.1684	0.1376	0.1805	0.0657	0.0372	3898	0.7819
Training	Fit Generalized Adaptive Elastic Net		0.1474	0.1684	0.1376	0.1805	0.0657	0.0372	3898	0.7819
Validation	Fit Nominal Logistic		0.1546	0.1788	0.1502	0.1891	0.0695	0.0400	1299	0.7731
Validation	Fit Generalized Lasso		0.1546	0.1787	0.1502	0.1891	0.0695	0.0400	1299	0.7731
Validation	Fit Generalized Adaptive Lasso		0.1545	0.1787	0.1502	0.1891	0.0695	0.0400	1299	0.7731
Validation	Fit Generalized Elastic Net		0.1546	0.1787	0.1502	0.1891	0.0695	0.0400	1299	0.7730
Validation	Fit Generalized Adaptive Elastic Net		0.1545	0.1787	0.1502	0.1891	0.0695	0.0400	1299	0.7731
Test	Fit Nominal Logistic		0.1085	0.1228	0.1272	0.1724	0.0621	0.0331	1300	0.8131
Test	Fit Generalized Lasso		0.1087	0.1230	0.1272	0.1724	0.0621	0.0331	1300	0.8131
Test	Fit Generalized Adaptive Lasso		0.1087	0.1230	0.1272	0.1724	0.0621	0.0331	1300	0.8130
Test	Fit Generalized Elastic Net		0.1088	0.1231	0.1272	0.1724	0.0621	0.0331	1300	0.8132
Test	Fit Generalized Adaptive Elastic Net		0.1087	0.1230	0.1272	0.1724	0.0621	0.0331	1300	0.8130

Based on these results from the model comparison, elastic net is the chosen model for predicting wine quality. While most of the models performed almost exactly the same across all measures, elastic net performed slightly better. It has the highest entropy and generalized R^2 values. All models had the same results for RMSE and misclassification rate. Elastic net performed slightly higher with the most important measure, AUC. A random coin toss would create an AUC of 0.50; therefore, any AUC value above 0.50 is seen as a good model because it is better than a random classification. A perfect AUC is a value of one. This elastic net model has a very good AUC of 0.8132.

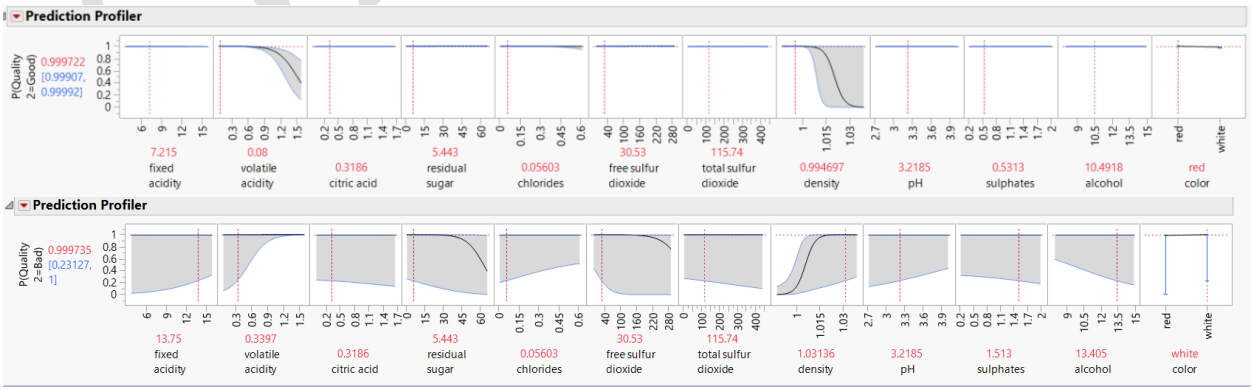
Interpretation

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-320.3024	141.4671	5.1263659	0.0236*	-597.5729	-43.03202
fixed acidity	-0.037997	0.1963462	0.0374496	0.8466	-0.422828	0.3468347
volatile acidity	5.7421429	0.6827833	70.72651	<.0001*	4.4039122	7.0803735
citric acid	-0.383	0.8271741	0.2143894	0.6433	-2.004231	1.2382316
residual sugar	-0.143581	0.0638888	5.0506143	0.0246*	-0.268801	-0.018361
chlorides	3.1922214	3.2989794	0.9363254	0.3332	-3.273659	9.6581022
free sulfur dioxide	-0.027385	0.0201376	1.849279	0.1739	-0.066854	0.0120842
total sulfur dioxide	-0.003279	0.0037008	0.7848975	0.3756	-0.010532	0.0039747
density	319.62015	144.2895	4.9068055	0.0268*	36.817927	602.42237
pH	0.066407	1.0069777	0.004349	0.9474	-1.907233	2.040047
sulphates	-0.426625	0.8729477	0.2388448	0.6250	-2.137571	1.2843212
alcohol	-0.01192	0.1774643	0.0045117	0.9464	-0.359744	0.3359036
color[red-white]	-3.917585	0.614823	40.600983	<.0001*	-5.122616	-2.712554

The elastic net model did not completely remove any variables from the model. However, only four variables are statistically significant. The most significant variables in predicting the quality of wine are volatile acidity and the color of the wine. Density and residual sugar are less statistically significant variables.

The higher the amount of volatile acidity in wine, the more it has a vinegar taste to it; therefore, as the volatile acidity goes up in wine, the more likely the quality will be poor. For color, red wine has better odds of being good than white wine. A wine with a little sugar is predicted to be good. As the residual sugar level drops 0.14 percent, the chance of the wine being of bad quality goes up. As the density of the wine goes up, the greater the chance the wine will be of bad quality. Alcohol is less dense than water, so the level of alcohol in the wine brings the density level down. If the wine has a high density, it will be a thicker liquid and less likely to be of good quality.



A wine has a 99.97 percent chance of being good when it is a red wine with 10 percent alcohol content. The pH level is 3.2 and the density is slightly less than the density of water. It has 5 grams of residual sugar per liter and chlorides, or amount of salt, equaling 0.056. A wine has almost 100 percent chance of being good when the acidity levels are as follows: 7.215 fixed, 0.08 volatile, and 0.3186 citric. As well as the sulfur dioxide levels being 30.53 free, 115.74 total, and 0.5313 sulphates.

All the physiochemical qualities of a wine that has a 99.97 percent chance of being bad are the same, except for five. Citric acid, residual sugar, chlorides, free and total sulfur dioxides, and the pH level can all remain the exact same. Nonetheless, the wine is just as likely to be bad, if the fixed acidity increases to 13.75, the volatile acidity to 0.3397, and the sulphates to 1.513. For the wine to have almost 100 percent chance of being bad, it would be a white wine and the density would be increased to 0.04 greater than the density of water measuring 1.03.

As the volatile acidity, chlorides, density, and pH level of a wine increases, the chance of it being good decreases. Once the density of a wine becomes more dense than water, the chance of it being good declines dramatically. This is especially true if the wine is white. For example, if a white wine has a 1.005 density level, there is a 46.36 percent chance it will be good, but a red wine with the same density has a 97.75 percent chance of being good.

A red wine has a better probability of being good than a white wine. As the residual sugar level goes up, so does the chance for good quality wine. The quality of wine is also positively correlated with fixed acidity, citric acid, free and total sulfur dioxide, sulphates, and alcohol content.

This elastic net model would predict a 98 percent chance of the wine being good and two percent chance of it being bad if the wine is a red wine with nine percent alcohol, 0.99 density level, and a pH of 3.2. The residual sugar level would need to be 4.8 with 0.09 of chlorides and 0.65 of sulphates. This wine has to have 8.9 fixed acidity, 0.9 volatile acidity, and 0.35 citric acid as well as 4 free and 40 total sulfur dioxides.

Variable Importance: Independent Uniform Inputs

Summary Report

Column	Main Effect	Total Effect
density	0.426	0.635
residual sugar	0.113	0.26
volatile acidity	0.085	0.219
free sulfur dioxide	0.081	0.196
color	0.069	0.153
chlorides	0.007	0.014
total sulfur dioxide	0.004	0.007
sulphates	0.001	0.002
citric acid	0.001	0.001
fixed acidity	0.001	0.001
pH	4e-5	8e-5
alcohol	5e-5	8e-5

The most important predictor variable for the quality of wine is the density. The density of a wine explains 63.5 percent of predictions on the quality of the wine. Residual sugar contributes 26 percent to the quality of wine prediction and volatile acidity is 22 percent. The fourth statistically significant variable, color, explains 15 percent of the difference between good or bad quality wine.

In conclusion, the best model to predict the quality of wine using 11 characteristics that make wine unique is an elastic net penalized regression. There are four statistically significant variables that affect the quality of the wine. The color and the volatile acidity level are the most statistically significant variables. If it is a red wine, over white, it has a better chance of being good and as the volatile acidity increases, or gets closer to tasting like vinegar, the chance of the wine being good goes down. The amount of sugar in the wine after fermentation stops is positively correlated with better quality. The most important variable, density, is negatively correlated with good quality. If a wine gets to be more dense than water, the quality dramatically decreases.