Portuguese Vinho Verde Red and White Wines

Nikki Aaron, Colleen Callahan,
Michael Pajewski, Pantea Ferdosian

About the Data Set

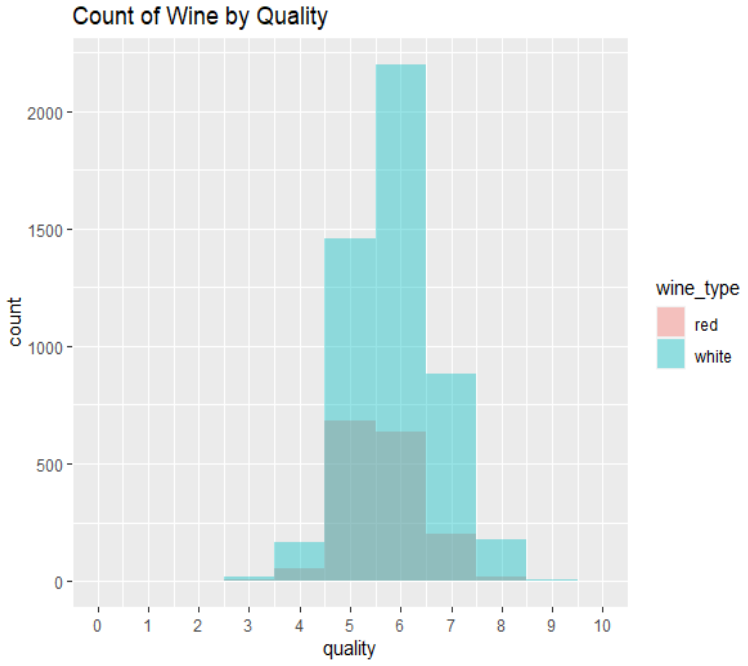
- Released by Paulo Cortez of the University of Minho, Guimarães, Portugal
- Quality and composition of Portuguese Vinho Verde wines
- Vinho Verde is slightly fizzy, high acidity fruit flavored
- Most often of the white variety
- Reds sometimes available in limited supply

Data Set Variables

Variables	**Description**
Fixed Acidity	Most acids contained in wine are fixed or nonvolatile
Volatile acidity	Acids that readily evaporate, high levels lead to an unpleasant vinegar taste
Citric acid	Citric acid can add ‘freshness’ and flavor to wines
Residual sugar	Sugar remaining after fermentation between 1 gram/liter and 45 grams/liter
Chlorides	Amount of salt in the wine
Free sulfur dioxide	Prevents microbial growth and the oxidation of wine
Total sulfur dioxide	In low concentrations is undetectable, but over 50 ppm becomes evident in the nose and taste of the wine
Density	Depends on the percent content of water, alcohol, and sugar
PH	ph is on a scale from 0 (acidic) to 14 (basic) most wines are between 3-4 on the pH scale. Related to acidity, but does not vary nearly as much.
Sulphates	antimicrobial and antioxidant
Alcohol (% volume)	Percent alcohol content of the wine
Quality	Qualitative score between 0-10

About the Data

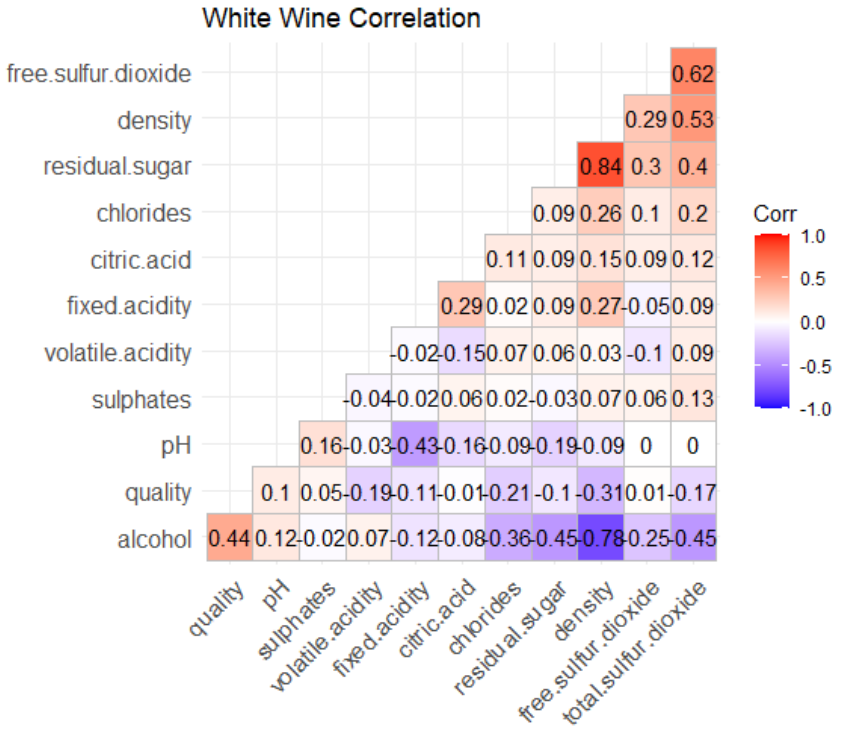
- ~5000 white, 1600 red
- Measures of acidity, alcohol, sugar, antioxidants, and salt
- Possible outliers in residual sugar, total sulfur dioxide, sulphates, and chlorides
- Quality scores 5-7 on average, none higher than 9 or below 3
- Goal of our analysis



Correlation for White Wine

The following variables are highly correlated.

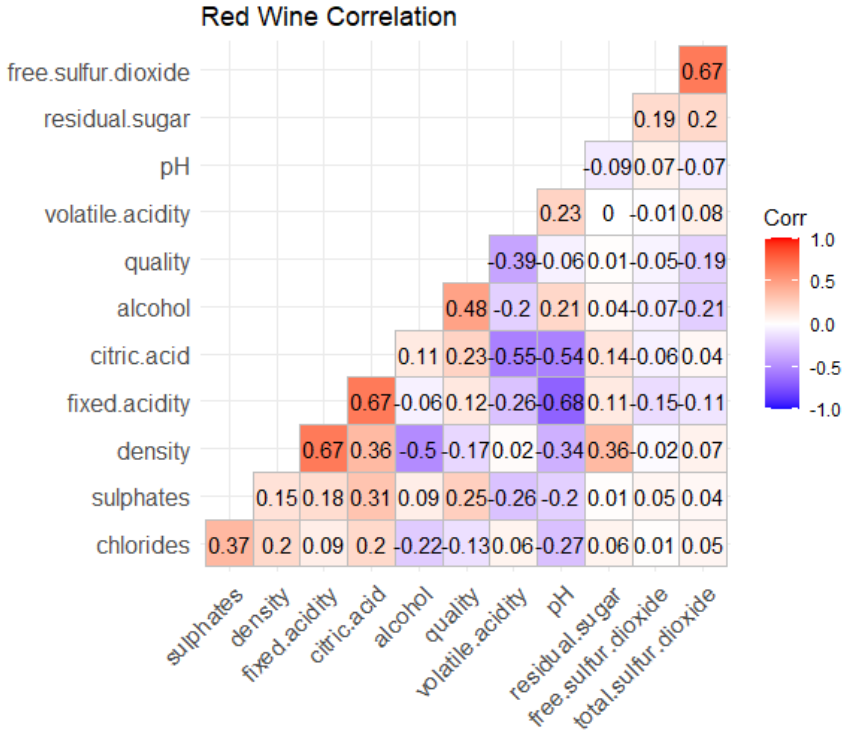
1. Quality and alcohol
2. Residual sugar and density
3. Alcohol and density
4. Free sulfur dioxide and total sulfur dioxide



Correlation for Red Wine

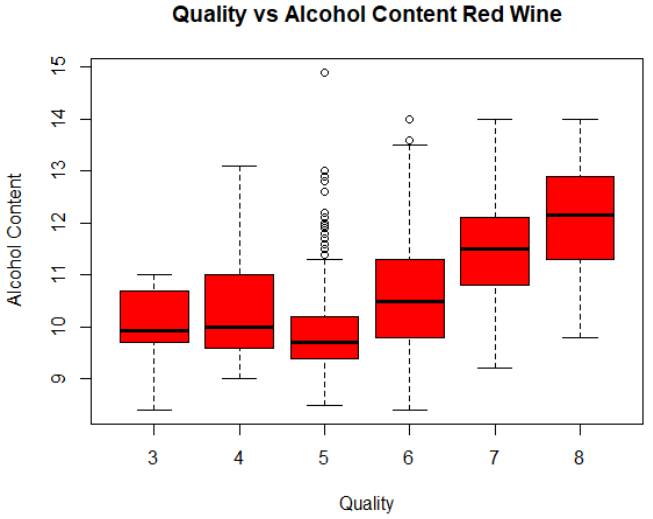
The following variables are highly correlated.

1. Quality and alcohol
2. Density and fixed acidity
3. Fixed acidity and citric acid
4. Volatile acidity and citric acid
5. Fixed acidity and ph
6. Free sulfur dioxide and total sulfur dioxide.



Observations

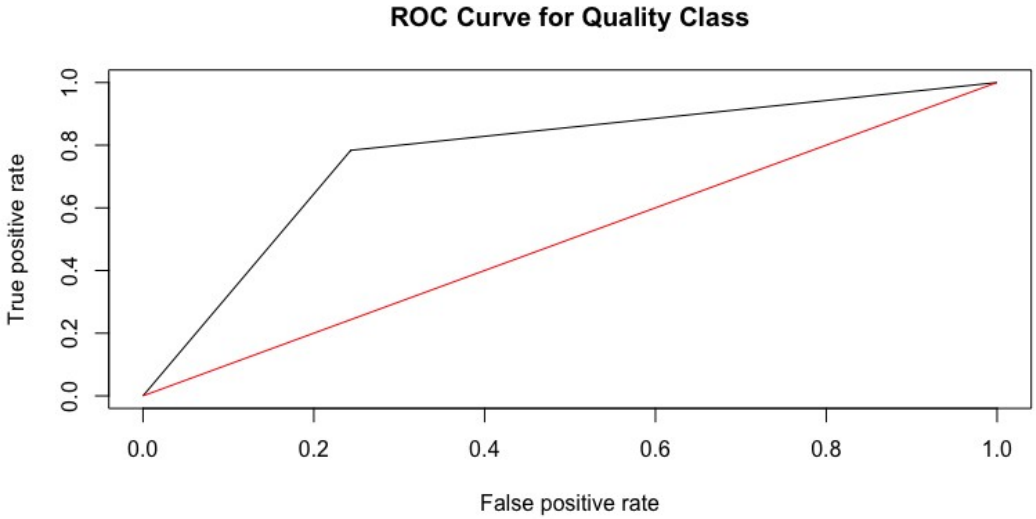
- Multicollinearity in measures of acidity, composition by volume, and sulfur dioxide
- Alcohol content is most notable predictor of quality, more is better
- High volatile acids more often low quality



fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides
4.942690	2.025418	1.599768	7.567620	1.611537
free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates
2.187147	3.018782	16.149845	2.528557	1.556976
alcohol	quality			
4.912481	1.403671			

Wine Quality Logistic Regression (Red)

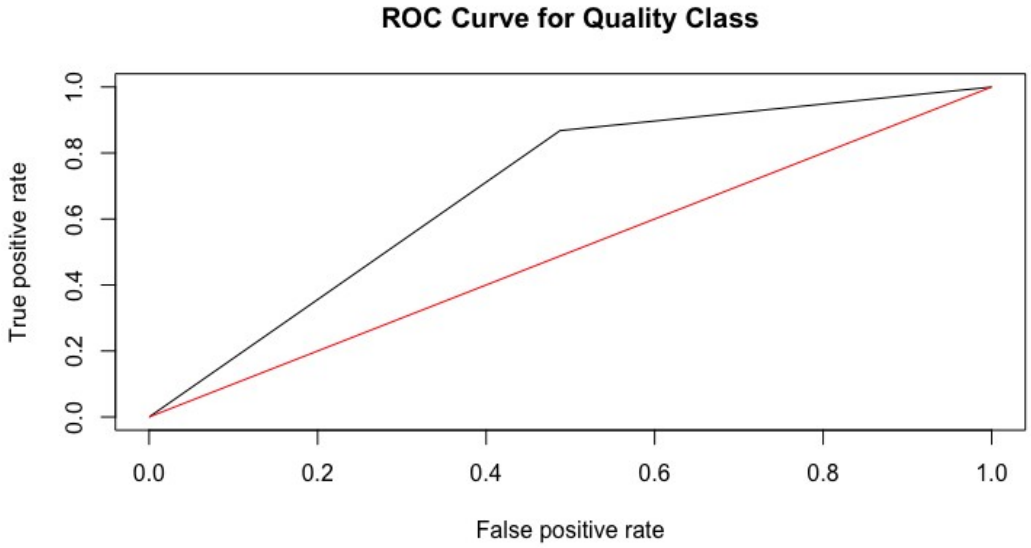
- Quality score
  - “Bad”: 1-5
  - “Good”: 6-10
- Quality class regressed on
  - Alcohol
  - Free Sulfur Dioxide
  - Sulphates
  - Volatile Acidity
  - Total Sulfur Dioxide
  - Chlorides
- ΔG^2: 864
- P-value: 0



$$\log(\pi/1-\pi) = -7.38 + 0.82*\text{alcohol} - 3.13*\text{volatile acidity} - 4.28*\text{chlorides} - 0.02*\text{total sulfur dioxide} + 0.02*\text{free sulfur dioxide} + 2.42*\text{sulphate}$$

Wine Quality Logistic Regression (White)

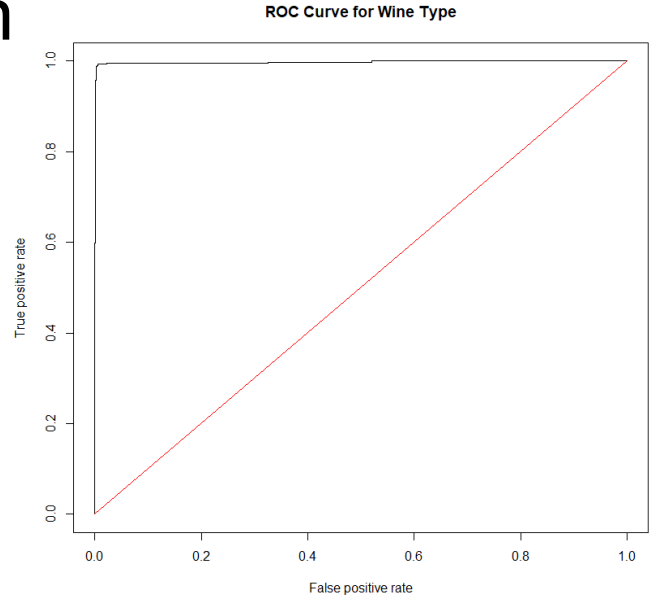
- Quality score
  - “Bad”: 1-5
  - “Good”: 6-10
- Quality class regressed on
  - Alcohol
  - Free Sulfur Dioxide
  - Sulphates
  - Volatile Acidity
  - Density
  - Residual Sugar
  - pH
- ΔG^2: 922
- P-value: 0



$$\log(\pi/1-\pi) = 201.6 - 6.59*\text{volatile acidity} + 0.14*\text{residual sugar} + 0.02*\text{free sulfur dioxide} - 213.6*\text{density} + 0.97*\text{pH} + 0.34*\text{sulphates} + 0.80*\text{alcohol}$$

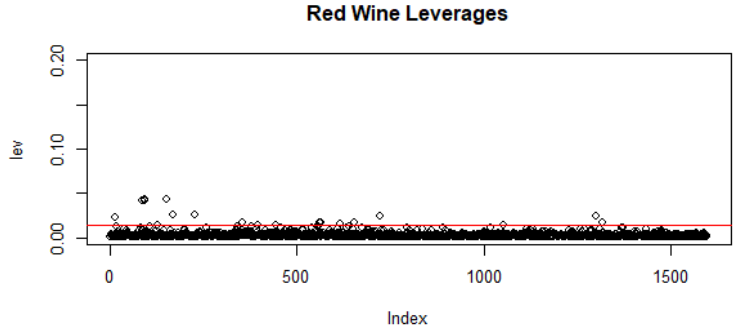
Wine Type Logistic Regression

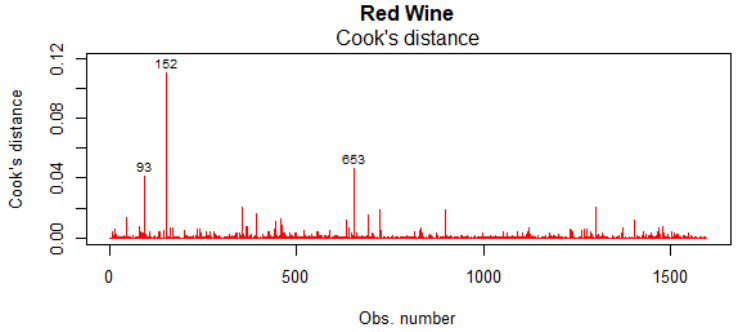
- Type class regressed on
  - Density
  - Residual sugar
  - Total sulfur dioxide
  - Volatile acidity
  - Chlorides
  - Sulphates
  - Alcohol
  - Free sulfur dioxide
- ΔG^2: 4742
- P-value: 0

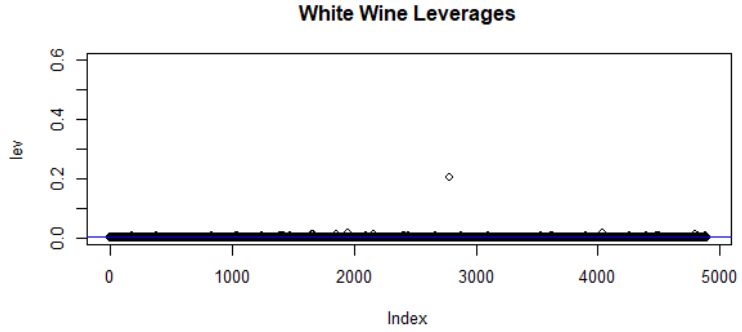


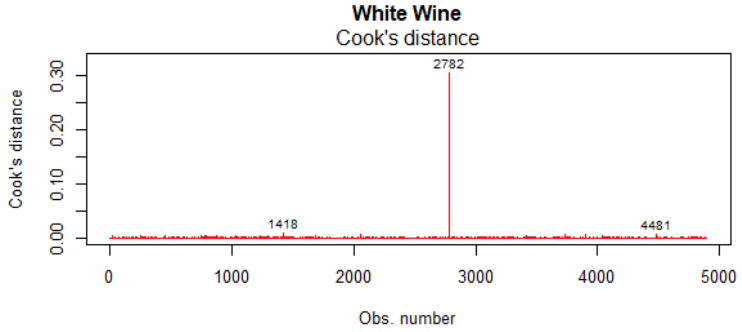
$$\log(\pi/1-\pi) = -1456 + 1447*\text{density} - 0.89*\text{residual sugar} - 0.06 * \text{total sulfur dioxide} + 7.46 * \text{volatile acidity} + 22.02*\text{chlorides} + 3.55*\text{sulphates} + 1.49*\text{alcohol} + 0.071*\text{free sulfur dioxide}$$

Leverage and Cooks - Quality

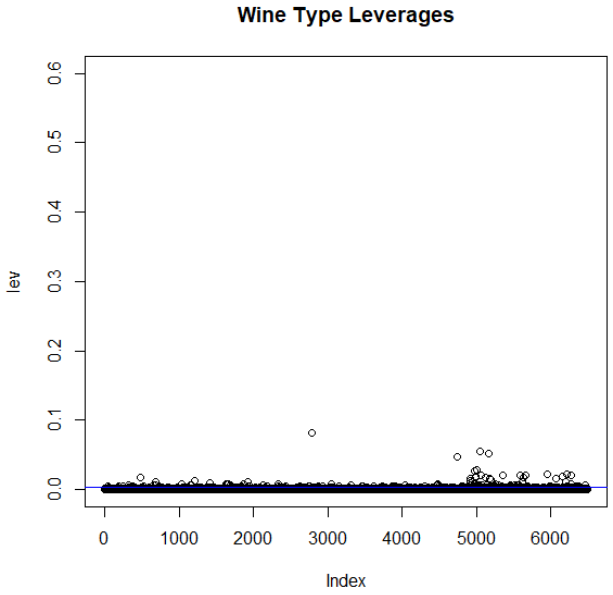


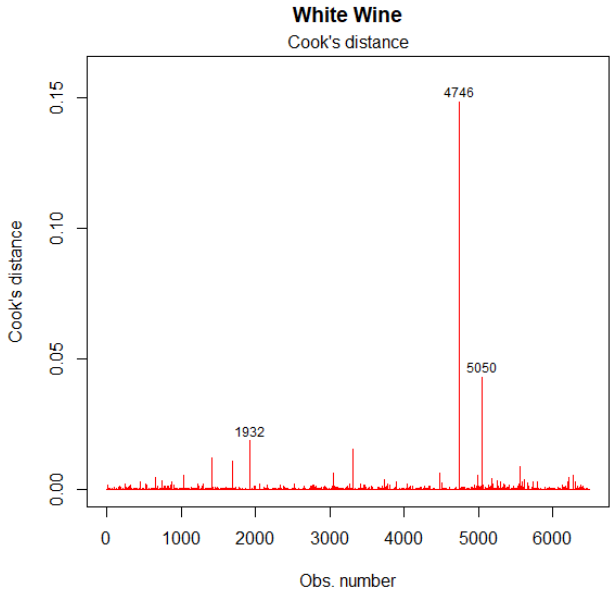






Leverage and Cooks - Wine Type





Confusion Matrices

A model from this data set performs very well at predicting wine type with 99% accuracy and few errors.

However, accuracy for predicting quality is much worse -- only 76-77%. The predictors in this data set are not excellent predictors of quality, but predict correctly more often than not.

Wine Type

Accuracy: 99%	Precision: 99%	False Positive Rate: 0.3%
False Negative Rate: 3.2%	Recall: 99%	

Red Quality

Accuracy: 77%	Precision: 79%	False Positive Rate: 30%
False Negative Rate: 28%	Recall: 78%	

White Quality

Accuracy: 75%	Precision: 89%	False Positive Rate: 47%
False Negative Rate: 30%	Recall: 77%	

Log Odds Predictions

	Red Wine Quality	**White Wine Quality**	**Wine Type**
Delta G-squared p-value	0	0	0
Sample 1 Prediction (Actual: Good Quality Red)	95% Probability Good	--	99.6% Probability Red
Sample 2 Prediction (Actual: Good Quality White)	--	95% Probability Good	1.5% Probability Red

All models are useful for prediction according to the 0 p-values computed from delta G-squared.

We can see from two sample points that odds prediction is correctly identifying actual values with high probabilities.