Monster Lab Data Presentation: Wine Preferences

Julianne Ammirati

February 16, 2017

The code used to generate this presentation is available here: https://github.com/JulianneA/EDA_w_modeling.git.

Portugese White Wine Dataset Summary

This dataset contains preference information on 4898 different Portugese white varietals covering the following variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality. Overall, preference scores (coded as the *quality* variable) range from a low score of 3 to a high score of 9 with a mean score of 5.88 and a median score of 6 (SD=0.89). The dataset does not contain any missing values.

Summary Statistics for all variables are below:

	N	Min	Max	Mean	Std Dev	Median
Quality	4898	3	9	5.88	0.89	6
Fixed Acidity	4898	3.8	14.2	6.85	0.84	6.8
Volatile Acidity	4898	0.08	1.1	0.28	0.1	0.26
Citric Acid	4898	0	1.66	0.33	0.12	0.32
Residual Sugar	4898	0.6	65.8	6.39	5.07	5.2
Chlorides	4898	0.009	0.346	0.05	0.02	0.043
Free Sulfur Dioxide	4898	2	289	35.31	17.01	34
Total Sulfur Dioxide	4898	9	440	138.36	42.5	134
Density	4898	0.987	1.039	0.994	0.003	0.994
pH	4898	2.72	3.82	3.19	0.15	3.18
Sulphates	4898	0.22	1.08	0.49	0.11	0.47
Alcohol	4898	8	14.2	10.51	1.23	10.4

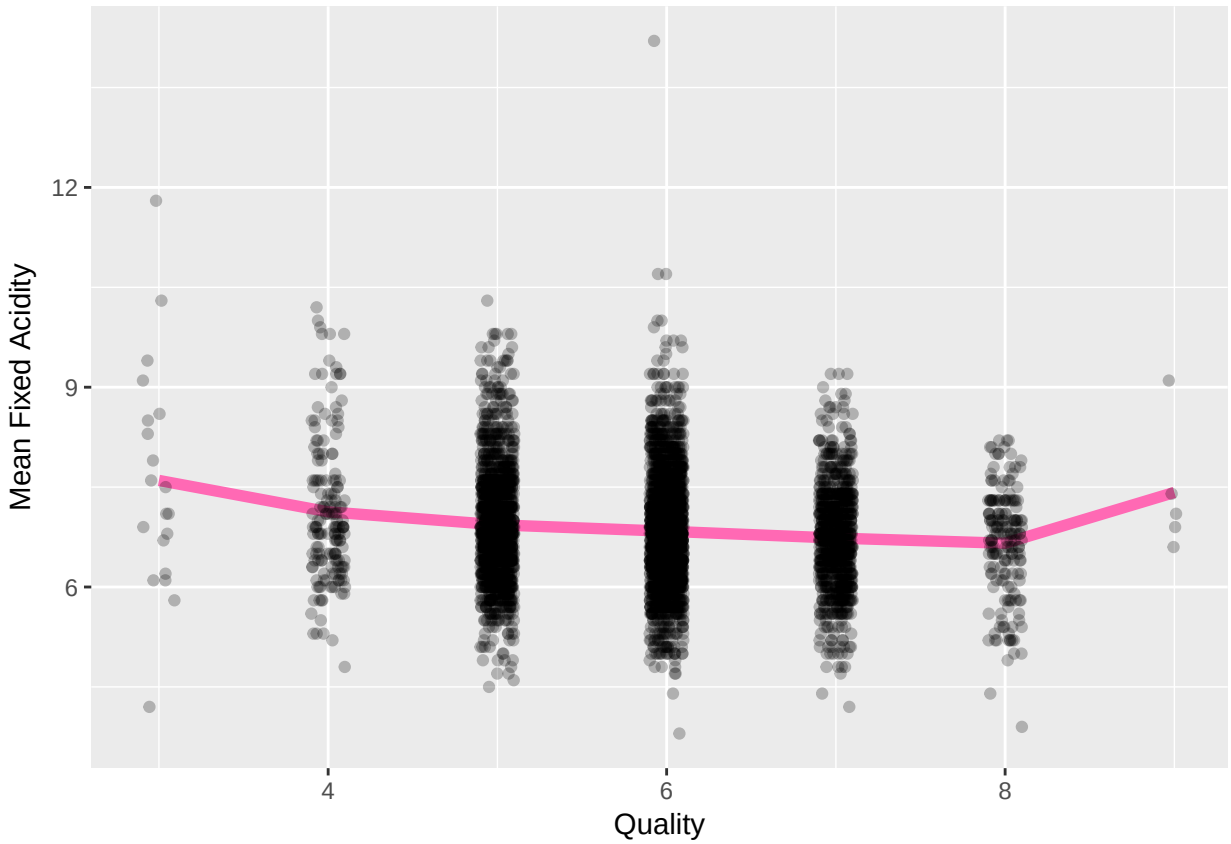
Evaluating the Mean Values for Wine Characteristics Based on Quality Rating

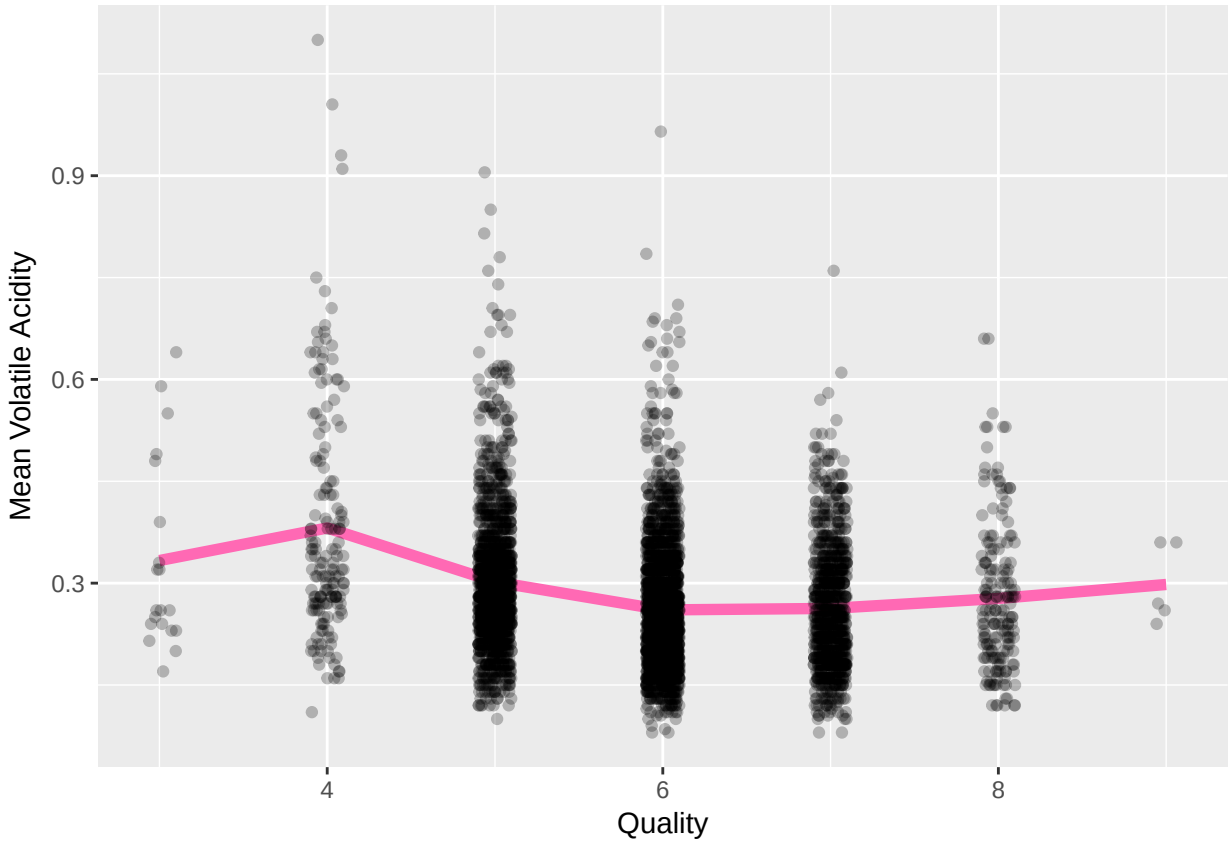
Our variable of interest is the wine's quality, and how this variable is impacted by the other characteristics, such as alcohol ocntent or pH. Below is a table of the mean value of each variable when the wines are grouped by quality rating. Each value is rounded to the nearest hundredth, except density measures.

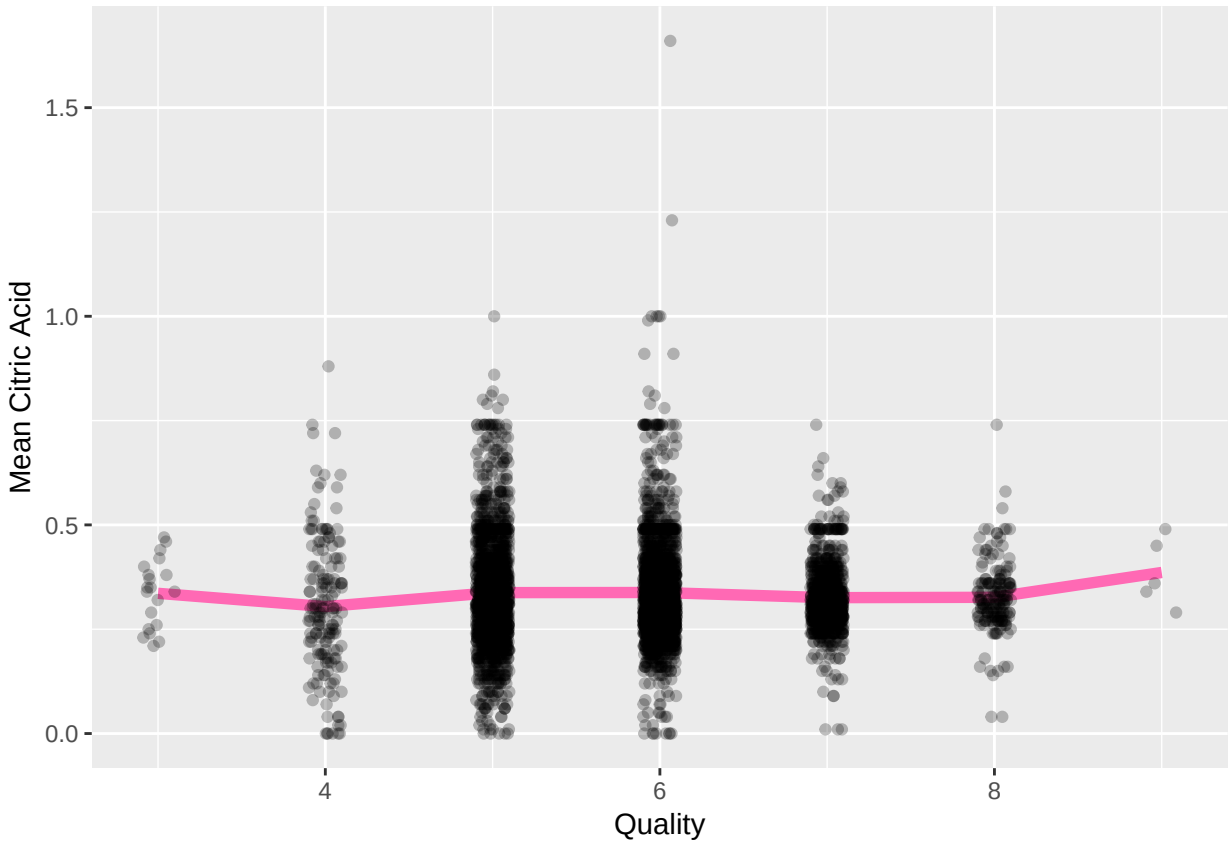
Quality	3	4	5	6	7	8
Fixed Acidity	7.6	7.13	6.93	6.84	6.73	6.66
Volatile Acidity	0.33	0.38	0.3	0.26	0.26	0.28
Citric Acid	0.34	0.3	0.34	0.34	0.33	0.33
Residual Sugar	6.39	4.63	7.33	6.44	5.19	5.67
Chlorides	0.05	0.05	0.05	0.05	0.04	0.04
Free Sulfur Dioxide	53.33	23.36	36.43	35.65	34.13	36.72
Total Sulfur Dioxide	170.6	125.28	150.9	137.05	125.11	126.17
Density	10.35	10.15	9.81	10.58	11.37	11.64
pH	3.19	3.18	3.17	3.19	3.21	3.22
Sulphates	0.47	0.48	0.48	0.49	0.5	0.49
Alcohol	0.9949	0.9943	0.9953	0.994	0.9925	0.9922

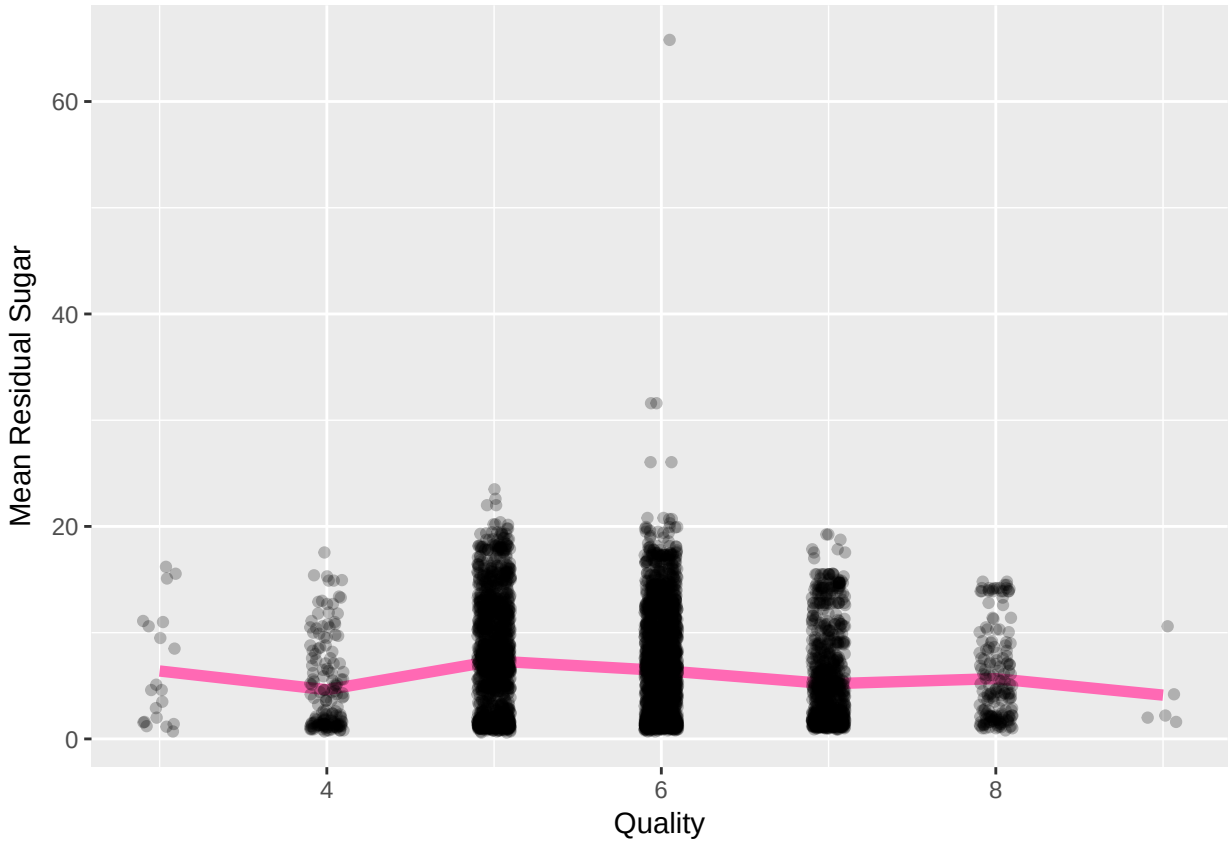
Means of All Independent Variables by Quality Value

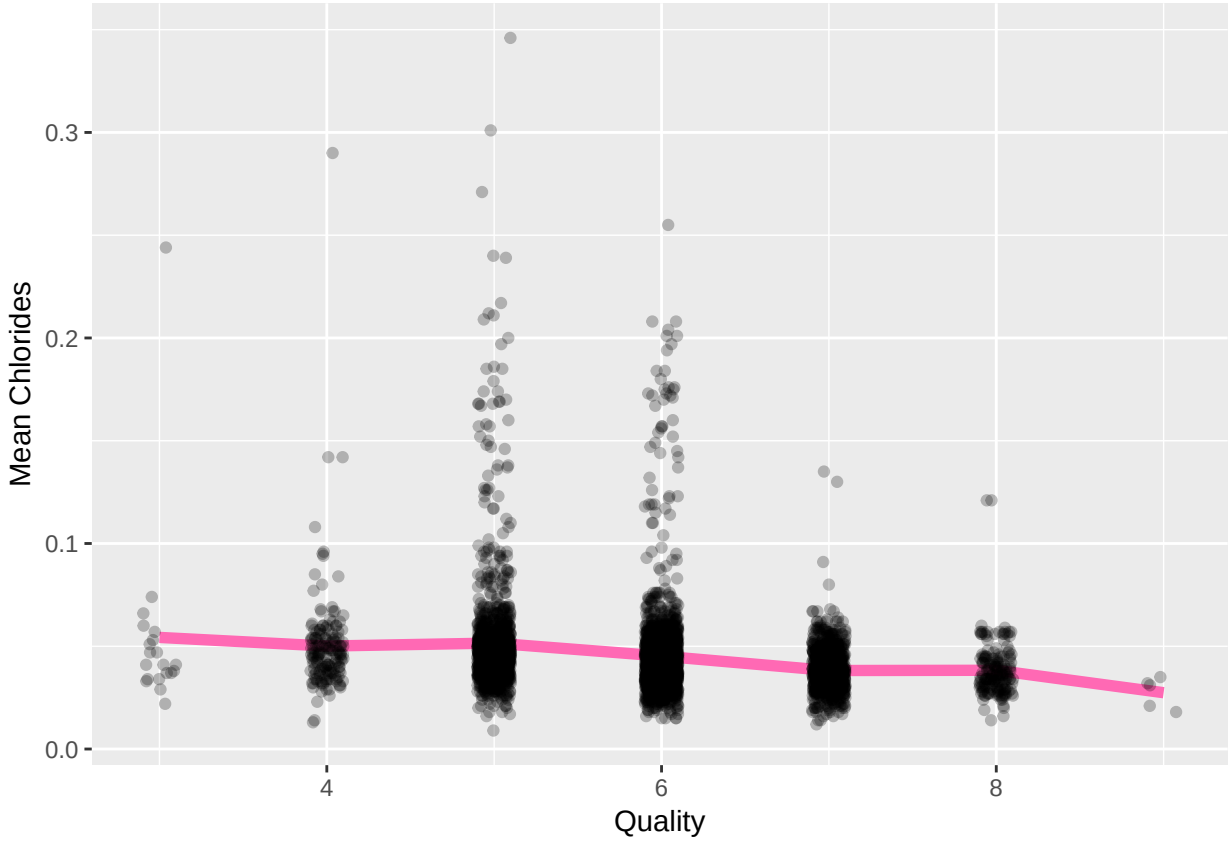
These graphs are histograms of all data points for the given characteristic. The hot-pink overlay tracks the mean value at each quality level for that characteristic, which is purely exploratory and not meant to display significant patterns in the data.

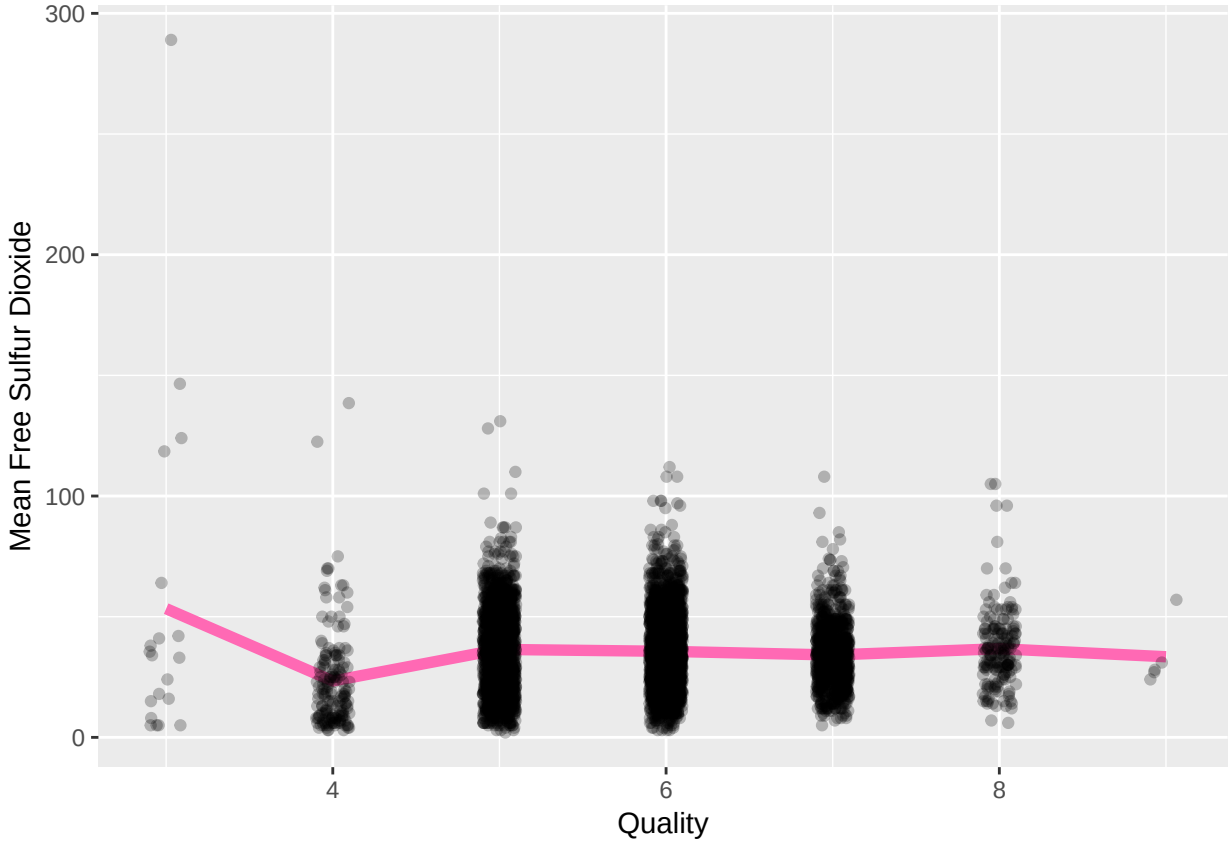

Mean Fixed Acidity
12
9
6
4
6
8
Quality

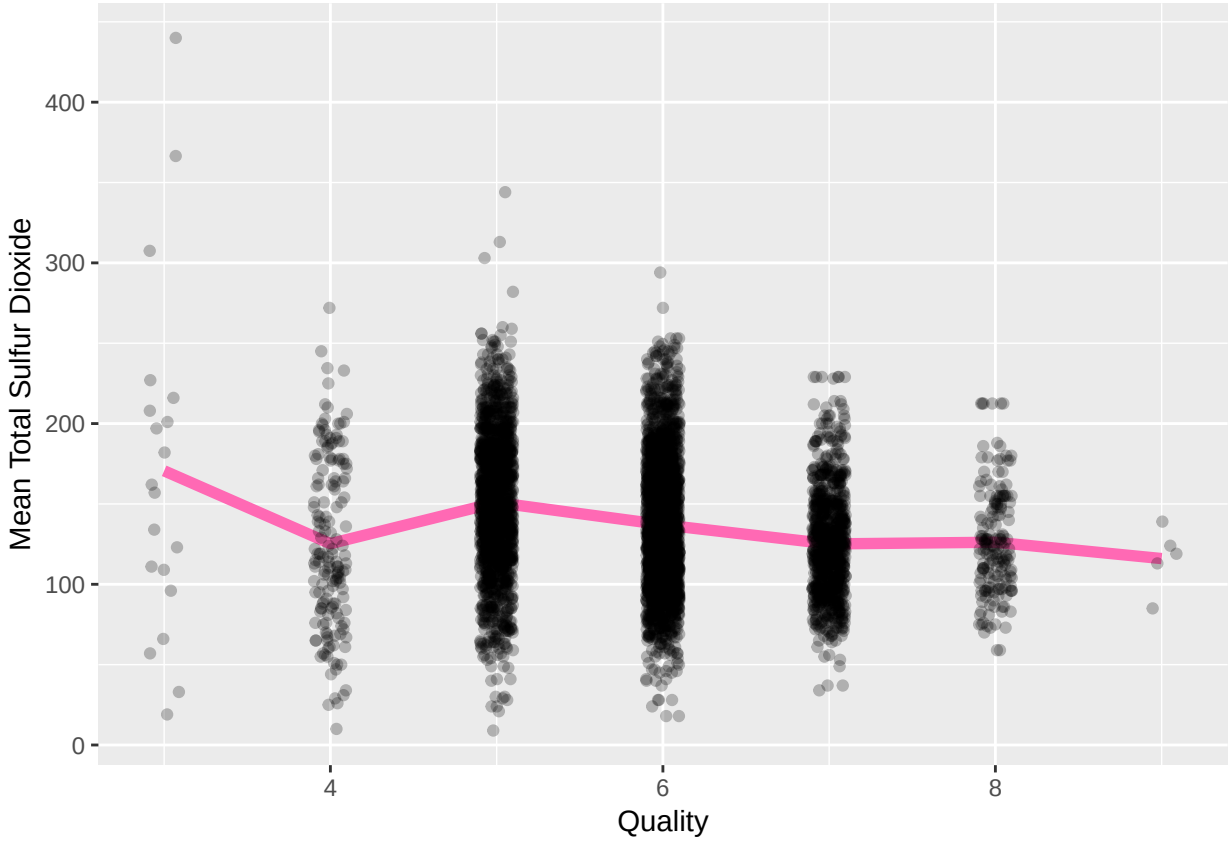

Mean Volatile Acidity
0.9
0.6
0.3
4
6
8
Quality

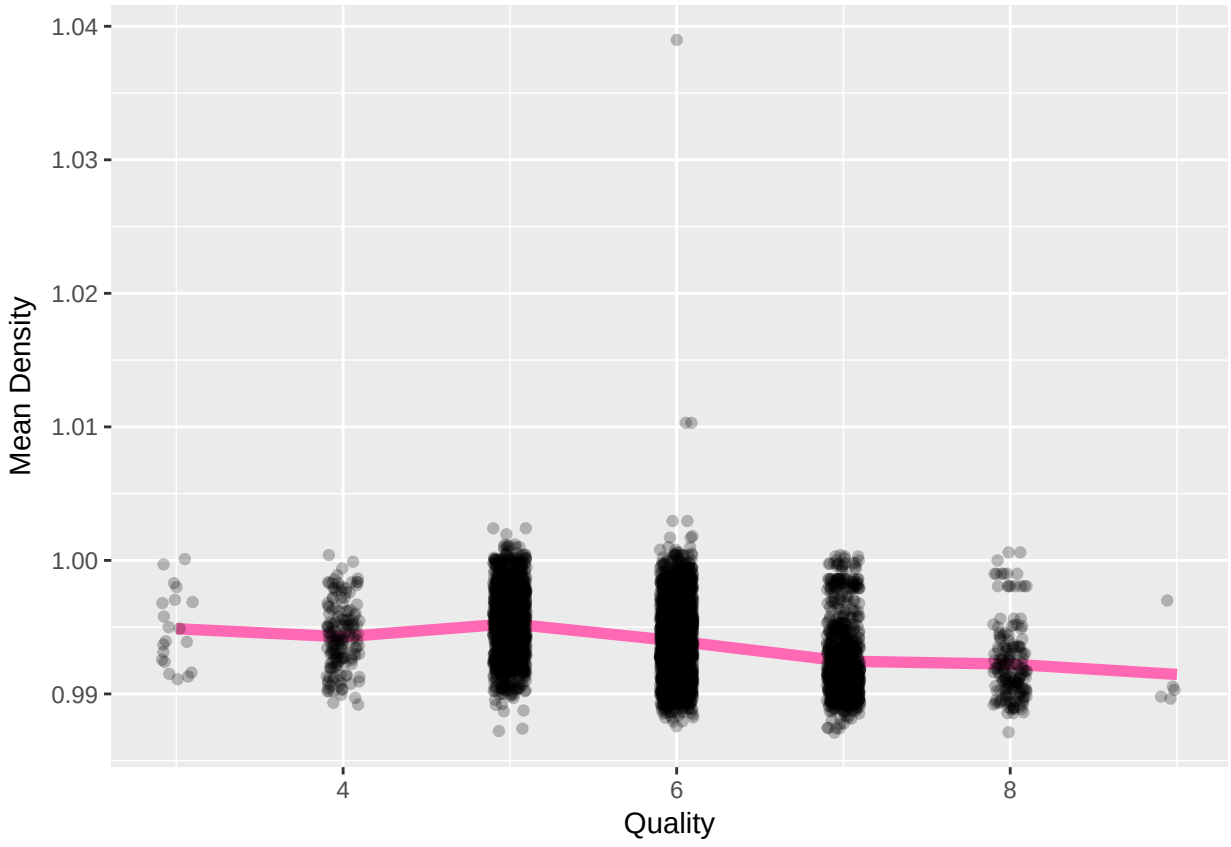

Mean Citric Acid
1.5
1.0
0.5
0.0
4
6
8
Quality

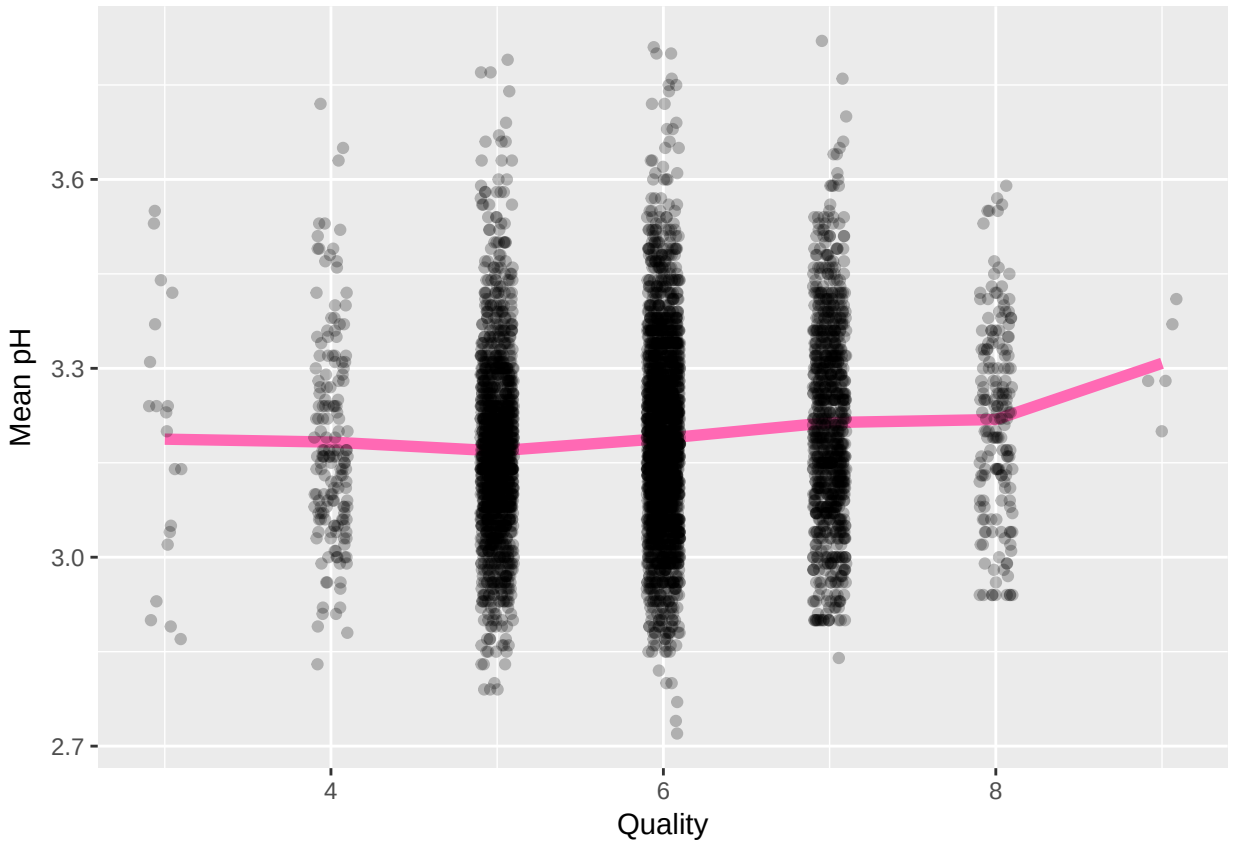

Mean Residual Sugar
60
40
20
0
4
6
8
Quality

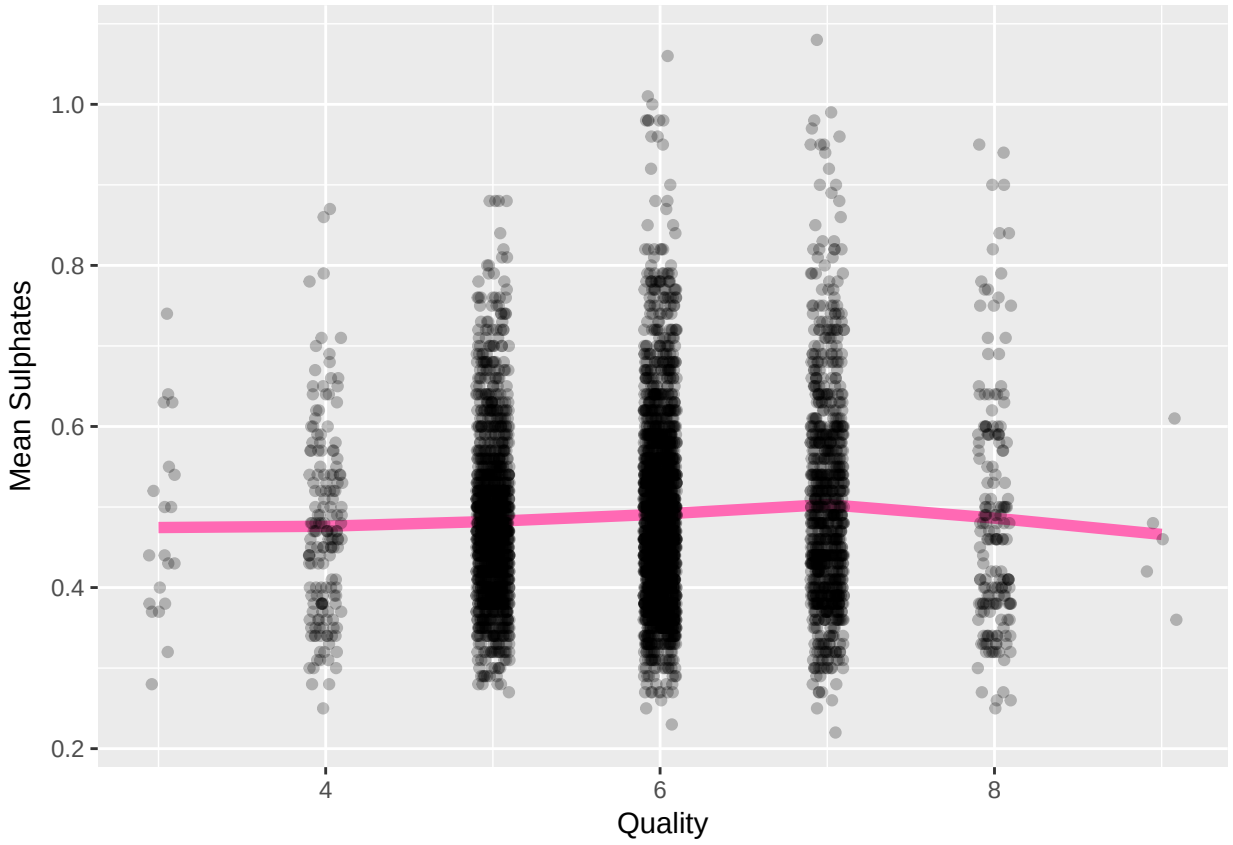

Mean Chlorides
0.3
0.2
0.1
0.0
4
6
8
Quality

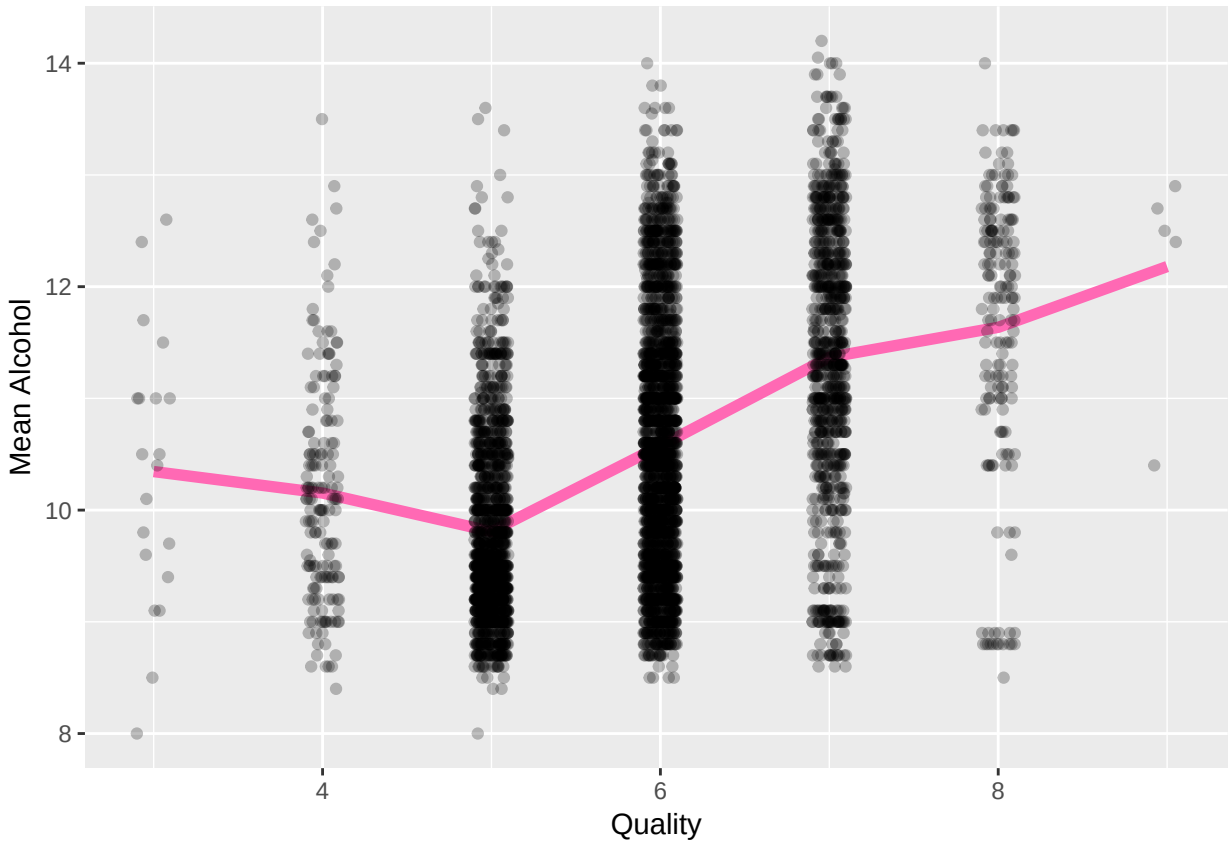

Mean Free Sulfur Dioxide
300
200
100
0
4
6
8
Quality


Mean Total Sulfur Dioxide
400
300
200
100
0
4
6
8
Quality


Mean Density
1.04
1.03
1.02
1.01
1.00
0.99
4
6
8
Quality


Mean pH
3.6
3.3
3.0
2.7
4
6
8
Quality


Mean Sulphates
1.0
0.8
0.6
0.4
0.2
4
6
8
Quality



Univariate Modeling

Based on the histographs and line graphs above, alcohol content appears to be the strongest candidate for a significant linear relationship with the quality measure.

Exploratory simple linear modeling was conducted to complete a preliminary examination of the independent variable by quality relationship. While many of the linear models were found to be significant, this is likely the result of the sample size rather than true linear relationship between variables.

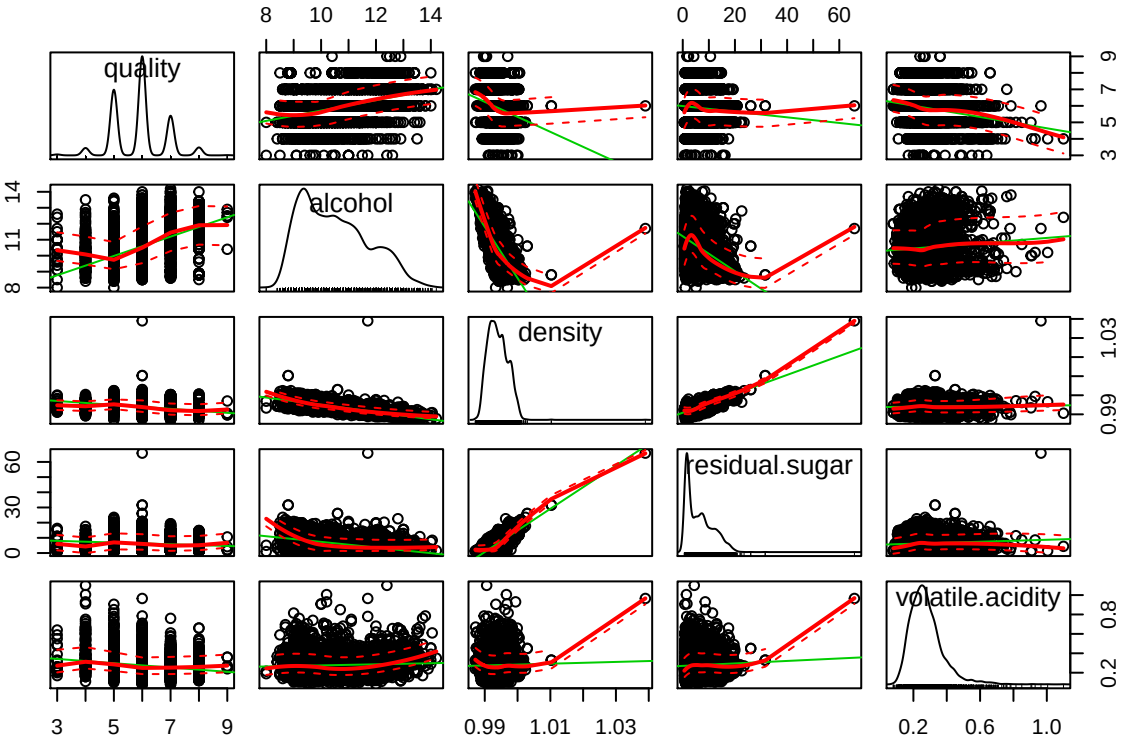
From this univariate analysis, six variables were found to have a significant linear relationship ($p<0.001$) with wine quality. Volatile acidity, residual sugar, chlorides, total sulfur dioxide, density, and alcohol content.

- For each increase of 1 unit in **Fixed Acidity**, the quality score of the wine decreases by 0.12
- For each increase of 1 unit in **Volatile Acidity**, the quality score of the wine decreases by 1.75
- For each increase of 10 units in **Residual Sugar**, the quality score of the wine decreases by 0.16
- For each increase of 1/10 unit in **Chlorides**, the quality score of the wine decreases by 8.31

- For each increase of 100 unit in **Total sulfur dioxide**, the quality score of the wine decreases by 0.40
- For each increase of 1/100 unit in **Density**, the quality score of the wine decreases by 0.90
- For each increase of 1 unit in **pH**, the quality score of the wine increases by 0.59
- For each increase of 1 unit in **Sulphates**, the quality score of the wine increases by 0.40
- For each increase of 1 unit in **Alcohol content**, the quality score of the wine increases by 0.31

Multivariate Modeling

Scatterplot Matrix of variables in Wine dataset:



It is clear from the initial plots that outliers are an issue in this dataset.

It initially seems as though alcohol content has some significant relationship to the quality of the wine. MLR models were generated to evauate which characteristicis improved the variation in quality accounted for by the model (optimized for adjusted R-squared value). Models were iterated on to continue to improve R-squared, with additional variables being gradually added to the strongest initial model: quality ~ alcohol.

- At most, these models were only able to account for approximately 26% of the variation in quality among the various wines.

Optimized model for this dataset

The final model included alcohol, density, volatile acidity, and residual sugar. This model had an adjusted R-square of 26.78% with the training dataset.

- For each increase of 1 unit in **alcohol content**, the quality score of the wine increases by 0.29, as long as density, volatile acidity and residual sugar were kept constant ($p<0.001$)
- For each increase of 1/100 unit in **density**, the quality score of the wine decreases by 0.70, as long as alcohol content, volatile acidity and residual sugar were kept constant ($p<0.001$)
- For each increase of 1 unit in **volatile acidity**, the quality score of the wine decreases by 2.08, as long as alcohol content, density and residual sugar were kept constant ($p<0.001$)
- For each increase of 10 units in **residual sugar**, the quality score of the wine increases by 0.53, as long as alcohol content, density and volatile acidity were kept constant ($p<0.001$)

Conclusion

Based on the final multivariare regression model using these variables, it is clear that we must reject the null hypothesis. The most comprehensive model that was pulled from this data included four variables in the regression equation: residual sugar, volatile acidity, alcohol content, and density. These four variables interact to predict quality via the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{1}$$

$$quality = \beta_0 + [.29 * (alcohol)] + [-70.33 * (density)] + [-2.08 * (volatileacidity)] + [0.05 * (residual_sugar)] \tag{2}$$

These findings suggest that as alcohol content increases, the quality of the wine is likely to improve as well. Density and volatile acid have inverse relationships with qualtiy, so when these variables increase, the quality will most likely decrease. Finally, residual sugar, like alcohol, has a positive linear relationship with wine quality, so sugar amount and quality are more likely to change value in the same direction.

It is important to note that models using this data can predict 26% of the variation, at best. In order to improve predictive value of models, it is necessary to clean the data and determine whether certain variables require transformations before performing regression modeling.