Wine Quality

MDA 620 Data Driven Decision Making

Capstone Project

By: Matthew Cozetti

Table of contents:

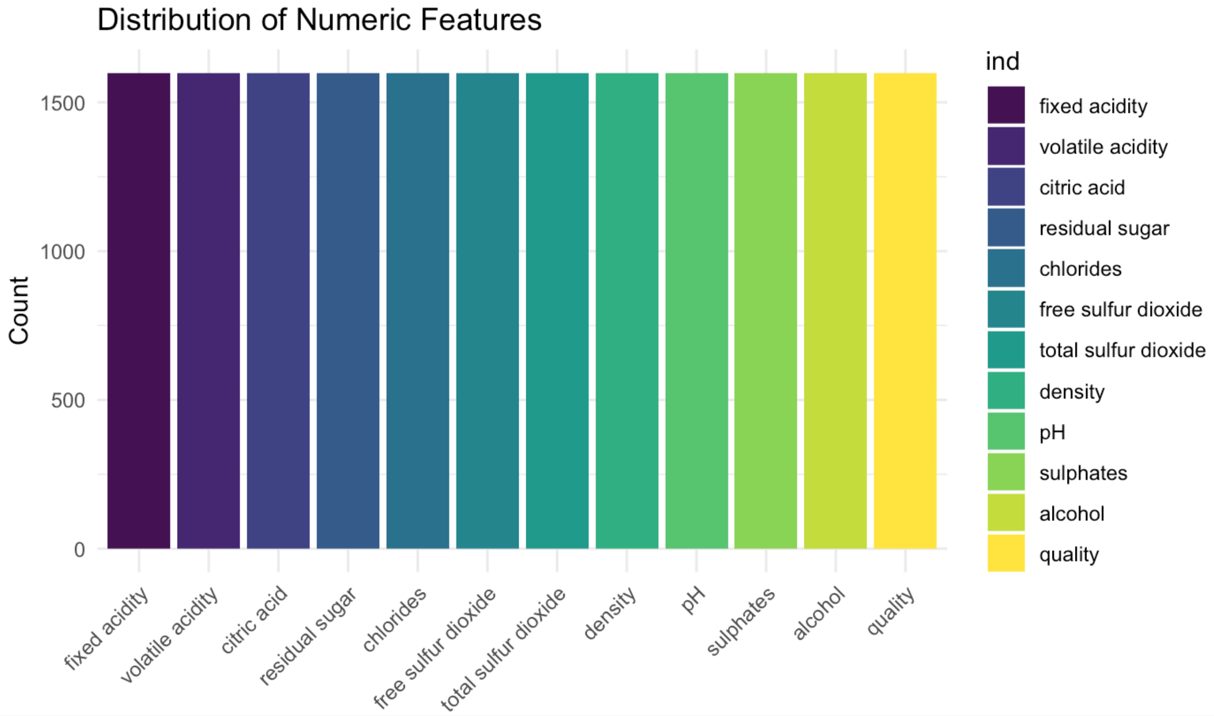
- Background
- Goal of the project
- Descriptive Statistics
- ANOVA method
- Z-score
- Identification of the most important Features
- Conclusion

Background:

The background of my project is to find which variable has the strongest relationship to the quality of wine. In other words, what improves the quality of wine the best. My variables are Fixed acidity, Volatile acidity, Citric acid**,** Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulphates, Alcohol, Quality. I have always been interested in why one wine has a higher quality than the other, so after looking at kaggle I was interested in the dataset and decided to use it as my project.

Problem Scenario/Business Issue & Objective/Goals of the Project:

My goal of the project is to understand how each characteristic (independent variables) affect the quality of the wine (dependent variable). Also to Identify the key features that significantly influence the quality of wine. The overall goal of my project in the business world is to find trends of why a quality of wine is better than another and provide this information to producers and consumers. Knowing this information will assist future producers in how to make a higher quality wine.



Data Exploration/Data Visualization & Data Manipulation:

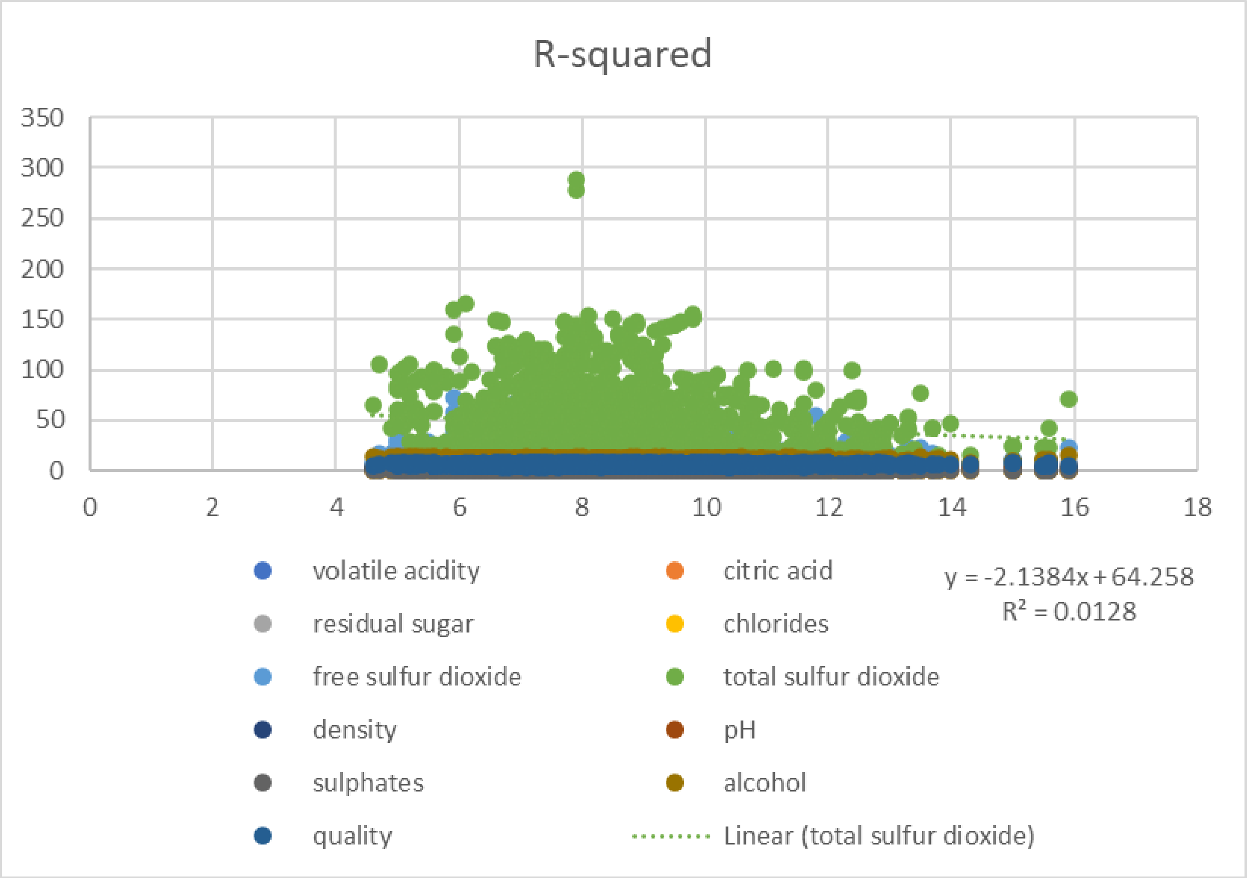
Some descriptive statistics of the data set are as follows,
Means:

Fixed acidity **8.319637**, Volatile acidity **0.527821**, Citric acid **0.270976,** Residual sugar **2.538806**, Chlorides **0.087467,** Free sulfur dioxide **15.87492**, Total sulfur dioxide **46.46779,** Density **0.996747,** pH **3.311113,**
Sulfates **0.658149,** Alcohol **10.42298,** Quality **5.636023**

Modes:
Fixed acidity **7.2**, Volatile acidity **00.6**, Citric acid **0,** Residual sugar **2**, Chlorides **0.08**, Free sulfur dioxide **6**
Total sulfur dioxide **28**, Density **0.9972**, pH**3.3,** Sulphates **0.6**, Alcohol **9.5,** Quality **5**

25th Percentile(Lower quartile):

Fixed acidity **7.1**, Volatile acidity **0.39,** Citric acid **0.09**, Residual sugar **1.9**, Chlorides **0.07,** Free sulfur dioxide **7**, Total sulfur dioxide **22**, Density **0.995,** pH **3.21**, Sulphates **0.55,** Alcohol **9.5,** Quality **5**

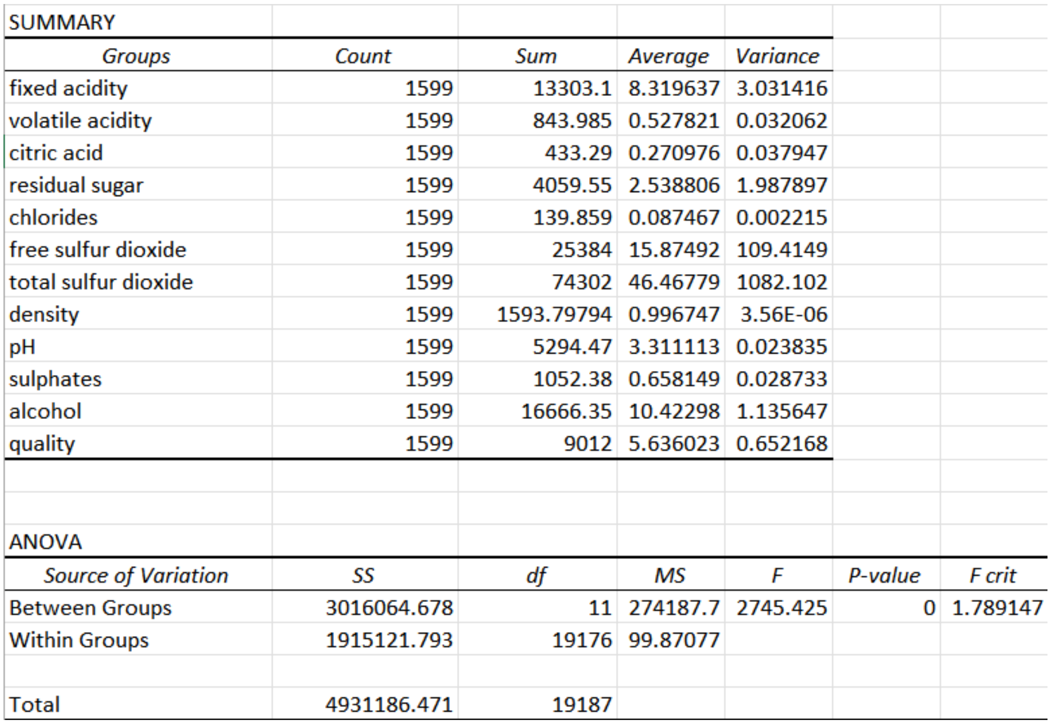


Here is a histogram of the data showing the r-squared value.

Methodology/Model Building & Model Selection:

One method I used was **ANOVA,** which stands for Analysis of Variance, that is a statistical method used to compare the means of three or more groups to determine if there is a significant difference between them. It's particularly important in understanding the relative importance of different features in a context like wine evaluation for several reasons. Another method that I used was the **Random Forest** that can be applied to wine testing in the context of predicting the quality of wine based on various physicochemical properties. This is a classic example of how machine learning, particularly ensemble learning techniques like Random Forest, can be used in the food and beverage industry for quality assessment and classification.

For **ANOVA,** it compares the means of different groups and shows you if there are any statistical differences between the means. As you can see here:



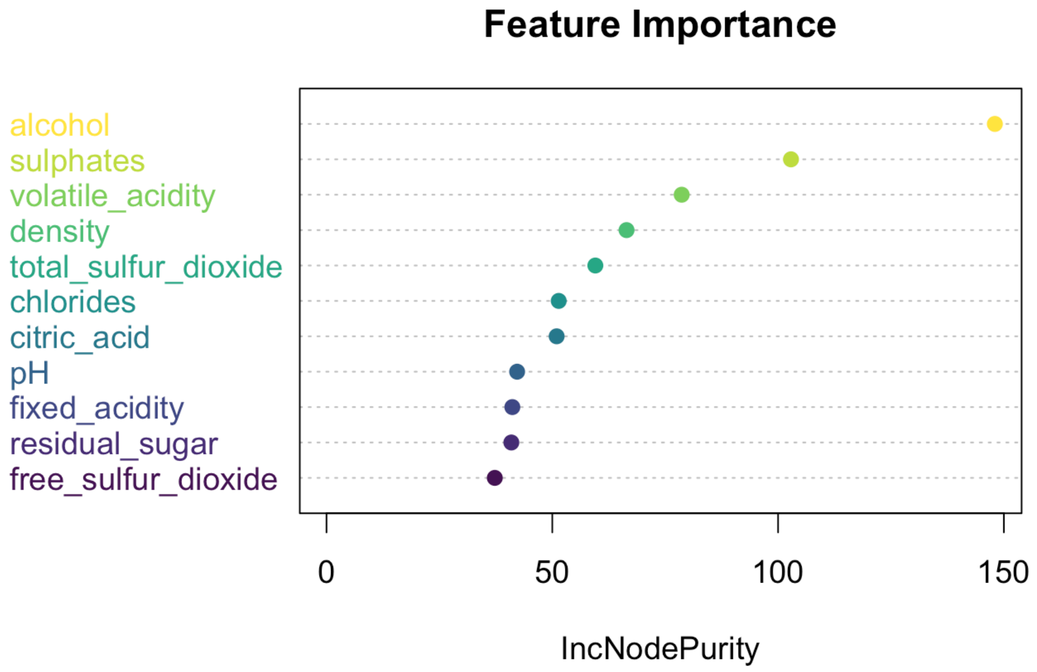
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
fixed acidity	1599	13303.1	8.319637	3.031416
volatile acidity	1599	843.985	0.527821	0.032062
citric acid	1599	433.29	0.270976	0.037947
residual sugar	1599	4059.55	2.538806	1.987897
chlorides	1599	139.859	0.087467	0.002215
free sulfur dioxide	1599	25384	15.87492	109.4149
total sulfur dioxide	1599	74302	46.46779	1082.102
density	1599	1593.79794	0.996747	3.56E-06
pH	1599	5294.47	3.311113	0.023835
sulphates	1599	1052.38	0.658149	0.028733
alcohol	1599	16666.35	10.42298	1.135647
quality	1599	9012	5.636023	0.652168

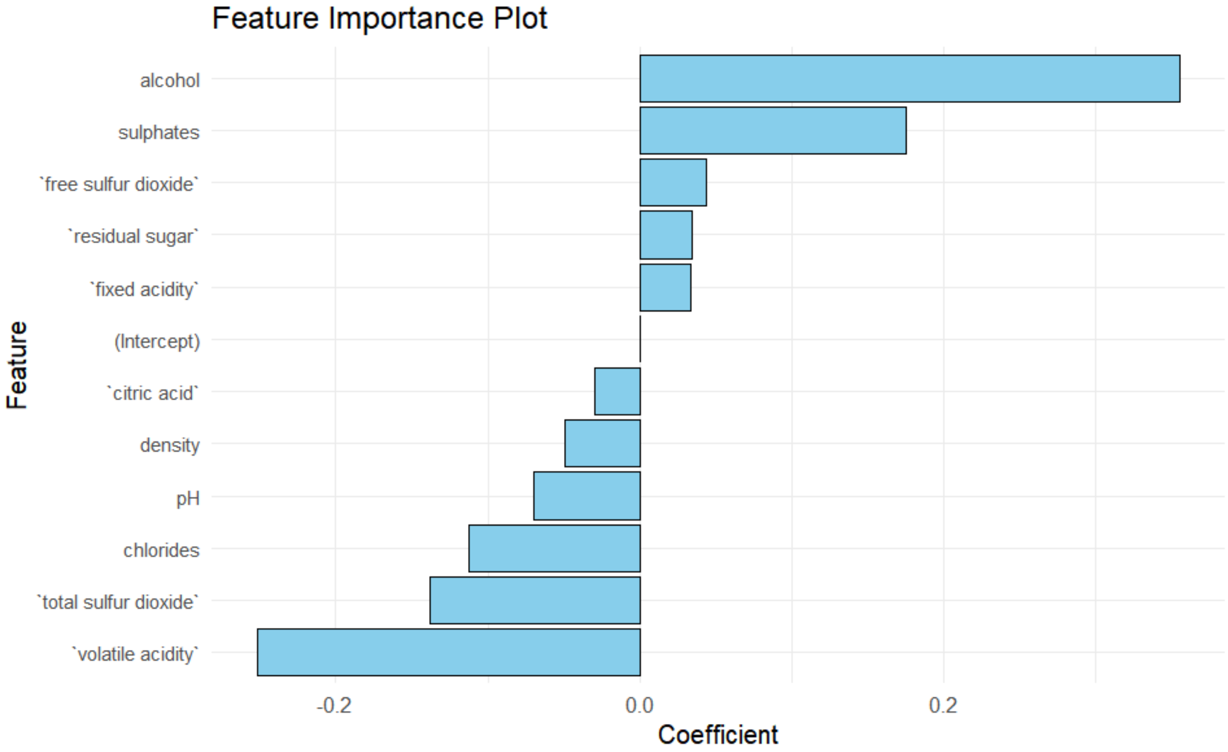
ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3016064.678	11	274187.7	2745.425	0	1.789147
Within Groups	1915121.793	19176	99.87077			
Total	4931186.471	19187				

For **Random Forest**, in this graph it gives a great representation of the importance of the data and which variable affects the price the most:



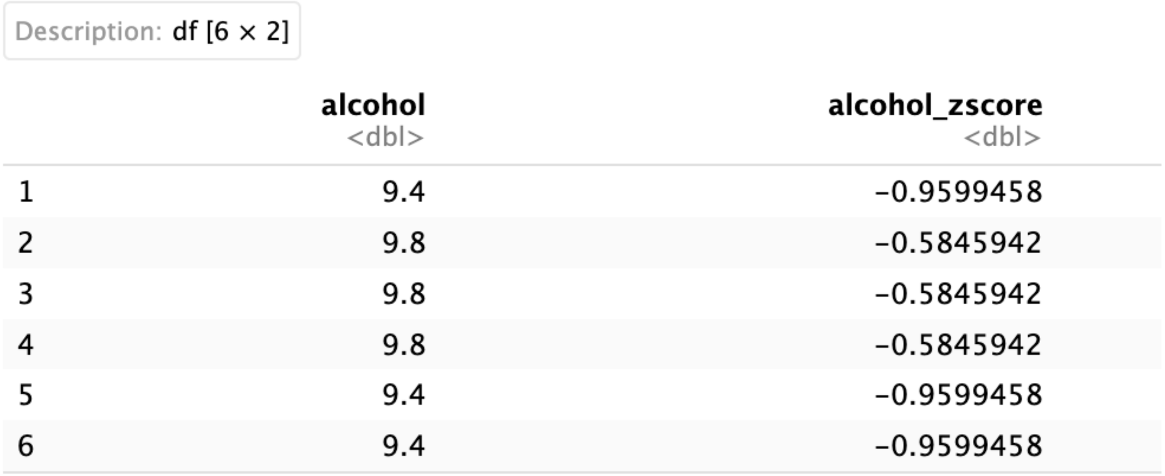
Next I made a graph where it specifically shows the importance between the features of wine and the price.



As you can see alcohol and sulfates have the highest coefficient between themselves and the price.

Lastly, here are the z-scores:

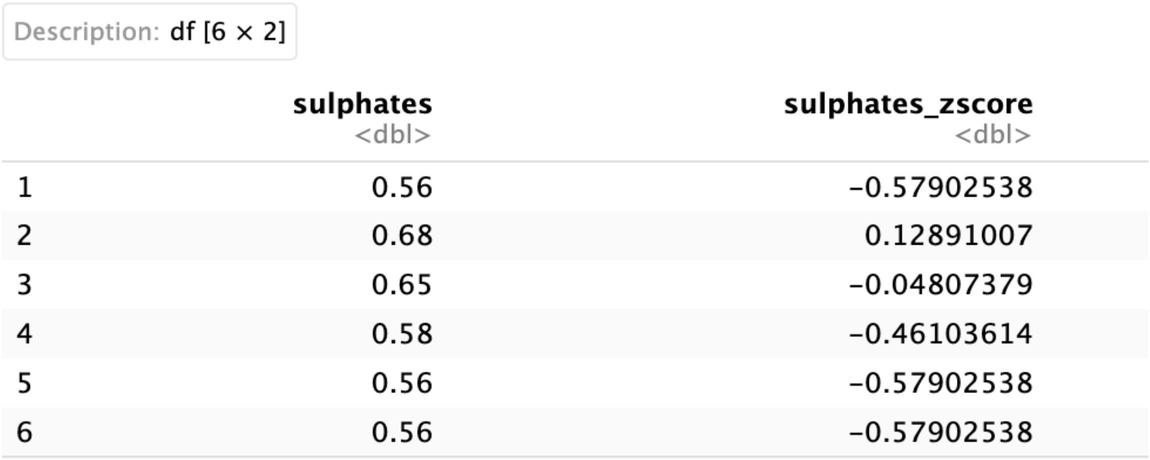
Alcohol



Description: df [6 × 2]

	alcohol <dbl>	alcohol_zscore <dbl>
1	9.4	-0.9599458
2	9.8	-0.5845942
3	9.8	-0.5845942
4	9.8	-0.5845942
5	9.4	-0.9599458
6	9.4	-0.9599458

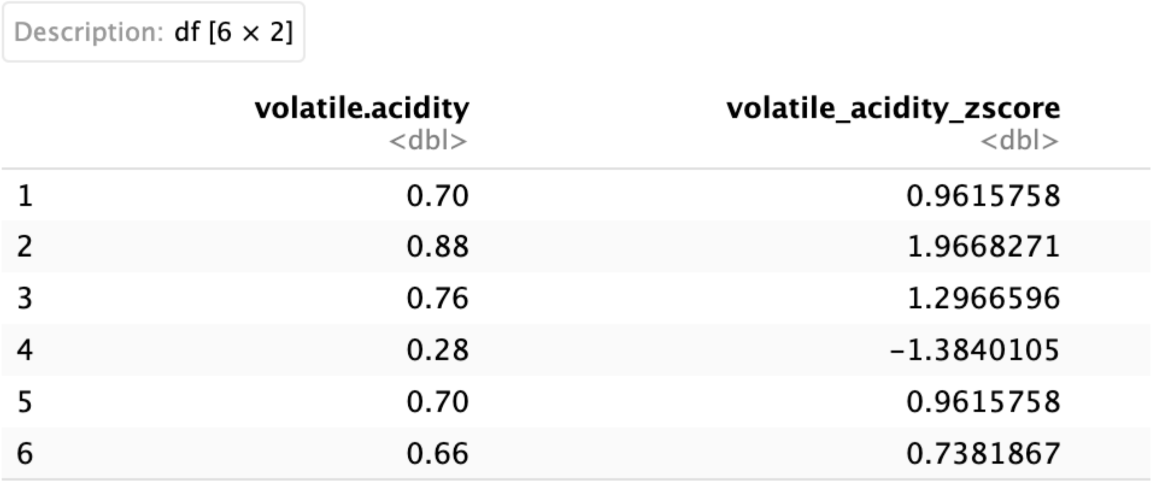
Sulphates



Description: df [6 × 2]

	sulphates <dbl>	sulphates_zscore <dbl>
1	0.56	-0.57902538
2	0.68	0.12891007
3	0.65	-0.04807379
4	0.58	-0.46103614
5	0.56	-0.57902538
6	0.56	-0.57902538

Volatile acidity



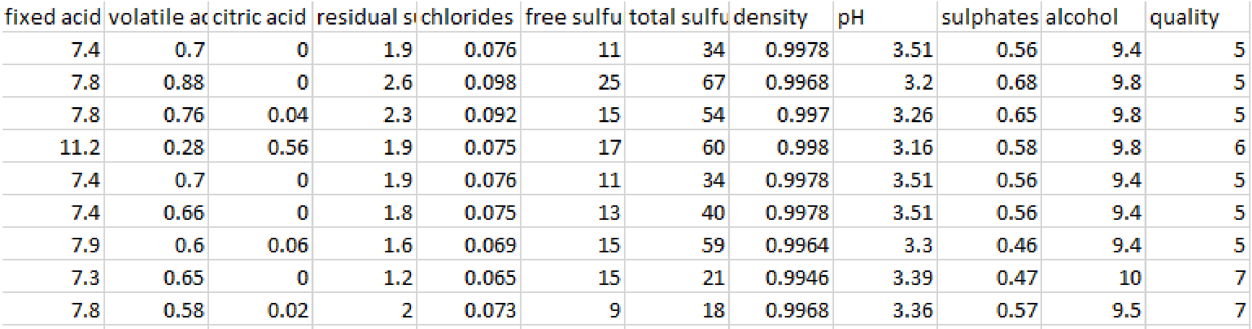
Description: df [6 × 2]

	volatile.acidity <dbl>	volatile_acidity_zscore <dbl>
1	0.70	0.9615758
2	0.88	1.9668271
3	0.76	1.2966596
4	0.28	-1.3840105
5	0.70	0.9615758
6	0.66	0.7381867

Conclusions/Recommendations:

Overall I found that the alcohol content and sulfates had the highest correlation between each other and the price after my study. I would recommend if you are producing the wine and care about the quality of it, you increase the amount of alcohol and sulfates when manufactured. For the consumer, if you are willing to pay a bit extra on wine, then you should look at the alcohol amount and sulfates of it and do your research. A consumer shouldn’t look into other features as much as these two.

Here is the first ten slides of my data set:



fixed acid	volatile ac	citric acid	residual s	chlorides	free sulfu	total sulfu	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7

Bibliography/References/Works Cited:

- Dataset: https://www.kaggle.com/
- Graphs: R studio
- Formulas: Excell
- https://www.investopedia.com/terms/a/anova.asp
- https://towardsdatascience.com/understanding-random-forest-58381e0602d2