Spencer Yeh 04/28/2020

Introduction:

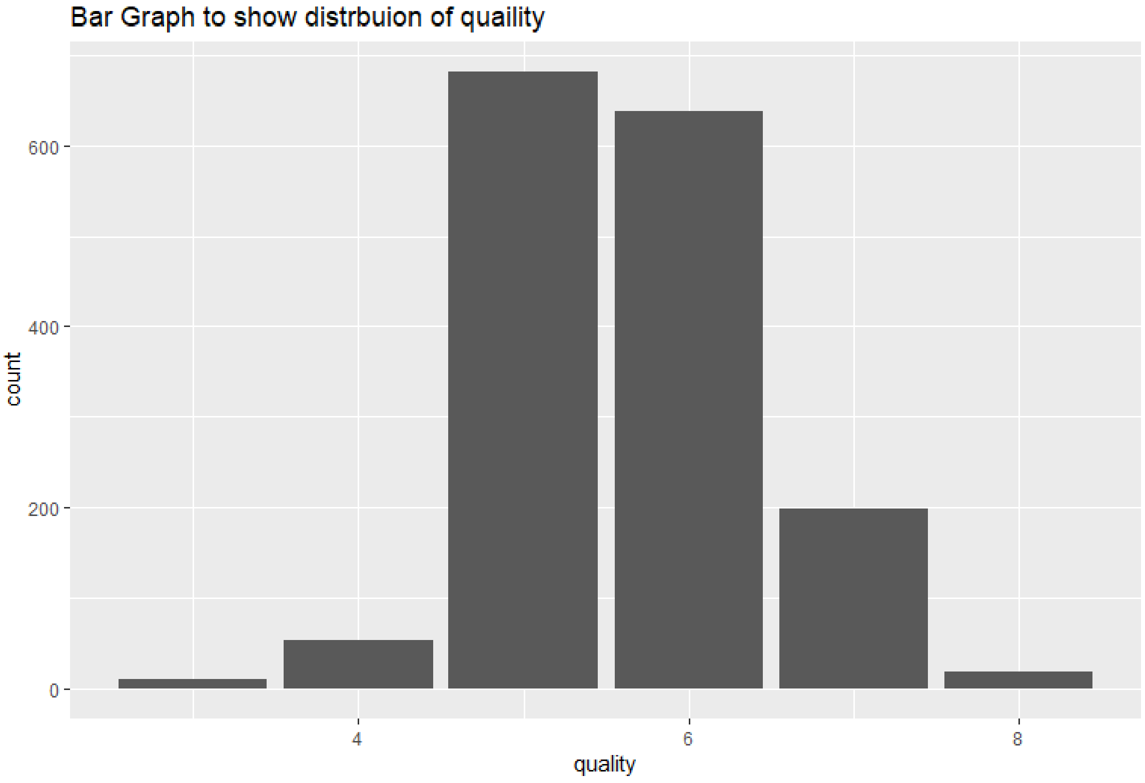
As a child I never understood why my parents would pay so much for a bottle of wine. While I still do not have an answer to my parent's spending habits; I do understand that wine is a popular drink with a long tradition that is integral to many cultures around the world. According to the wineinstitute.org, a website that promotes the California wine industry, the U.S consumed about 805 million gallons of wine in 2018. As a result, there are many different types of wine each with their own ingredients, varying age, and brewing methods. These differences in how the wine is made ultimately affect the quality of wine. What I hope to gain from this project is to further understand how the contents in wine ultimately affect its quality.

Data Description:

The project is utilizing a data set on the red wine Vino Verde. This data set was found on Kaggle.com, a website focused on providing free data sets for people to use, but it is sourced from the University of Minho from Portugal. The data offers 11 predictor variables such as alcohol, sulfates, acidity, and sulfur content that supposedly affect the quality of wine. Using these input variables, we hope to predict a single output variable, the quality of red wine. The input variables were collected using physicochemical (physical and chemical) tests while the output variable was based on a blind sensory test rated by sommeliers. In the data set the red wine is ordered based on quality and are not balanced, thus there are more normal grade wine than high grade wine or low-grade wine in the data set.

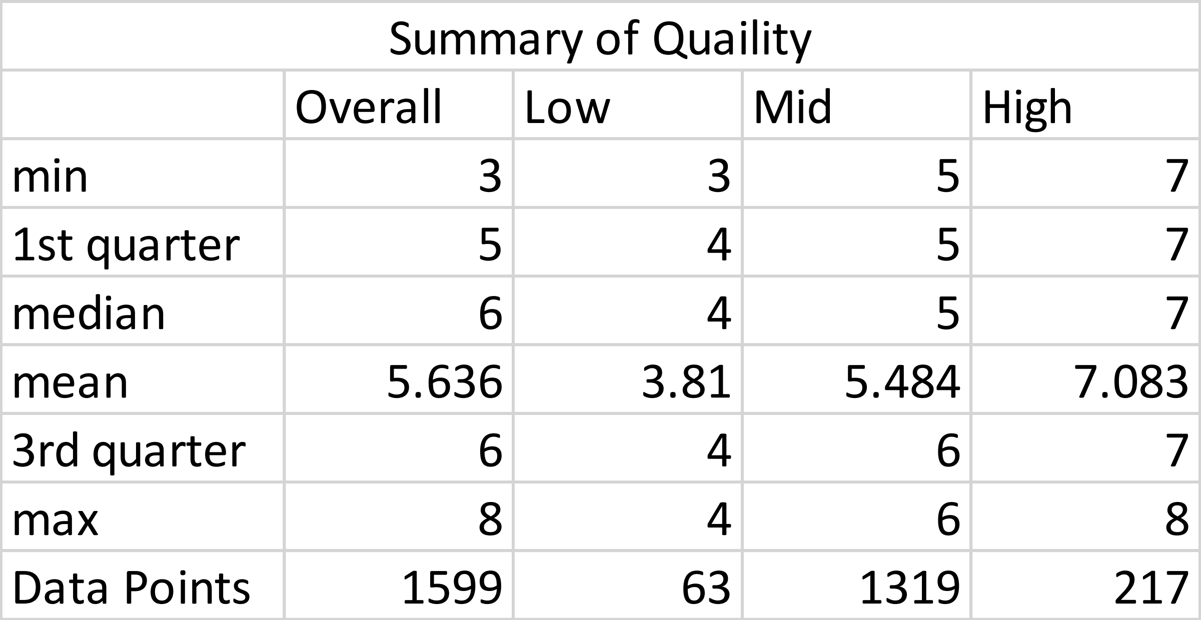
Data Manipulation:

While there was no missing data, which made it significantly easier to work with, the wine quality was ordered and not balance. The wine was given one of the 6 quality ratings 3 as the lowest quality wine and 8 as the highest quality wine. Furthermore, there were far more data for wine with a quality between 5 & 6 than, wine between 3 & 4 quality and 7 & 8. Therefor it was natural to group wine qualities 3-4 and classify them as low-grade wine, wine quality 5-6 as a mid-grade wine, and 7-8 quality wine as high-grade wine. Furthermore, there were less than 30 points of data for wine classified with a quality of 3 or 8, thus combining it with wine quality 4 and 7 respectively, allow these groups to better fit the shape of a population distribution inorder to compare it to the other grades of wine. Below is a graph to show how the wine quality is ordered and not balanced



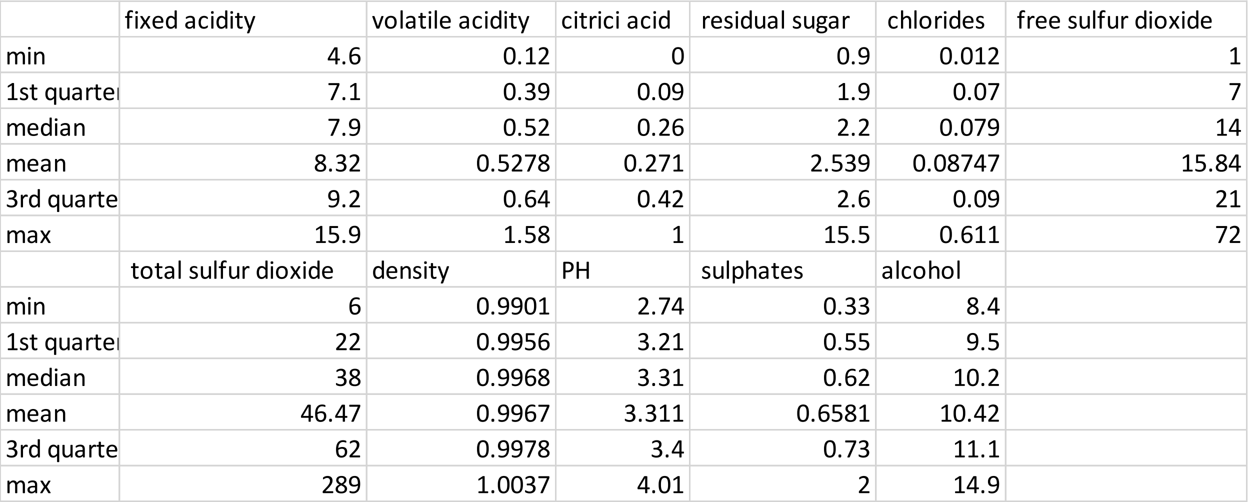
Data Summary:

Below Is a general summary based on the grade of wine. As seen by the summary there are far more mid-grade wines than the high-grade or low-grade wine.



Summary of Quaility				
	Overall	Low	Mid	High
min	3	3	5	7
1st quarter	5	4	5	7
median	6	4	5	7
mean	5.636	3.81	5.484	7.083
3rd quarter	6	4	6	7
max	8	4	6	8
Data Points	1599	63	1319	217

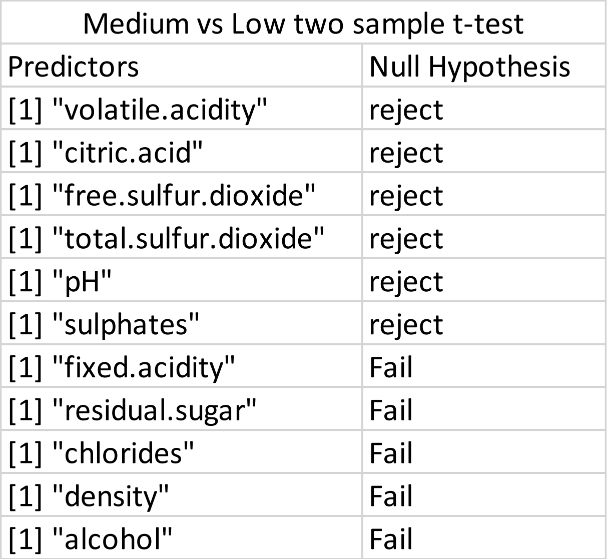
Below is a simple summary of our predictors that would affect the quality of wine.



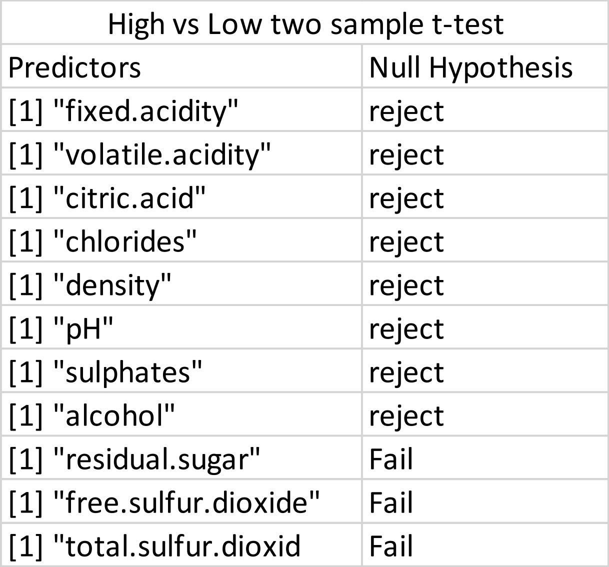
	fixed acidity	volatile acidity	citrici acid	residual sugar	chlorides	free sulfur dioxide
min	4.6	0.12	0	0.9	0.012	1
1st quarter	7.1	0.39	0.09	1.9	0.07	7
median	7.9	0.52	0.26	2.2	0.079	14
mean	8.32	0.5278	0.271	2.539	0.08747	15.84
3rd quarte	9.2	0.64	0.42	2.6	0.09	21
max	15.9	1.58	1	15.5	0.611	72
	total sulfur dioxide	density	PH	sulphates	alcohol	
min	6	0.9901	2.74	0.33	8.4	
1st quarter	22	0.9956	3.21	0.55	9.5	
median	38	0.9968	3.31	0.62	10.2	
mean	46.47	0.9967	3.311	0.6581	10.42	
3rd quarte	62	0.9978	3.4	0.73	11.1	
max	289	1.0037	4.01	2	14.9	

Analysis:

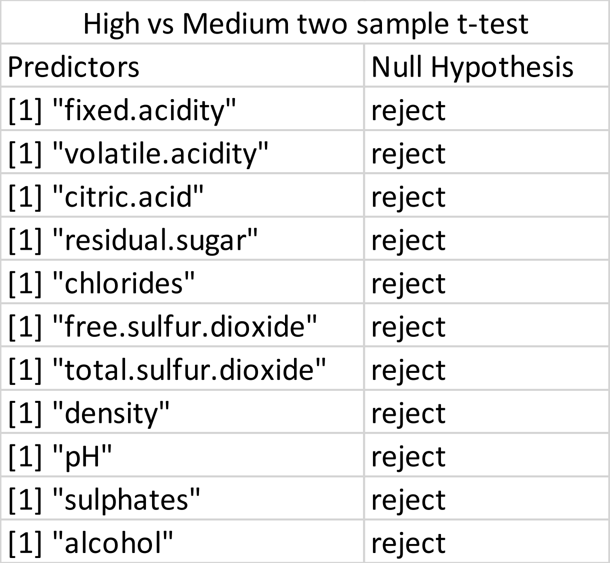
The hypothesis of the study was to understand what predictors affected wine quality. The first step in testing the hypothesis was to see if there was a significant difference in the concentration of predictors found between different grades of wine. Therefore, we used a two-sample t-test which allowed us to compare two different populations by comparing their sample means. We are using a 95% confidence interval, where we are 95% confident in the answer we get. We have a null hypothesis that there is no difference between sample means, and if we reject the null hypothesis that means there is a difference between predictors.



Medium vs Low two sample t-test	
Predictors	Null Hypothesis
[1] "volatile.acidity"	reject
[1] "citric.acid"	reject
[1] "free.sulfur.dioxide"	reject
[1] "total.sulfur.dioxide"	reject
[1] "pH"	reject
[1] "sulphates"	reject
[1] "fixed.acidity"	Fail
[1] "residual.sugar"	Fail
[1] "chlorides"	Fail
[1] "density"	Fail
[1] "alcohol"	Fail



High vs Low two sample t-test	
Predictors	Null Hypothesis
[1] "fixed.acidity"	reject
[1] "volatile.acidity"	reject
[1] "citric.acid"	reject
[1] "chlorides"	reject
[1] "density"	reject
[1] "pH"	reject
[1] "sulphates"	reject
[1] "alcohol"	reject
[1] "residual.sugar"	Fail
[1] "free.sulfur.dioxide"	Fail
[1] "total.sulfur.dioxid	Fail



High vs Medium two sample t-test	
Predictors	Null Hypothesis
[1] "fixed.acidity"	reject
[1] "volatile.acidity"	reject
[1] "citric.acid"	reject
[1] "residual.sugar"	reject
[1] "chlorides"	reject
[1] "free.sulfur.dioxide"	reject
[1] "total.sulfur.dioxide"	reject
[1] "density"	reject
[1] "pH"	reject
[1] "sulphates"	reject
[1] "alcohol"	reject

Looking at the results we are 95% confident that there is a significant statistical difference between all the predictors of quality between the high-grade and medium- grade wine. In the two-sample t-test between high-grade and low-grade wine, nine out of the eleven predictors we are 95% confident that they are statistically different. While in two sample t-test between medium-grade and low-grade wine only six out of the eleven predictors we are 95% confident that they are statistically different. Using these two-sample t-test we are able to determine that there is a significant difference between the predictors.

After computing the two-sample t-tests our next step is building a model to figure out what are the best predictors in predicting the quality of wine. To figure out the best predictors for predicting wine we will use a multiple linear regression. The multiple linear regressions allow us to predict the relationship between two or more variables. To find our multiple linear regression we will use the backwards selection method. That means we will put all our predictors in a model and slowly eliminate the worst predictors one by one until we have a solid model. Computing our model, we get our formula

$$\text{Quality} = 4.430099 + 0.289303*\text{alcohol} + 0.882665*\text{sulphates} - 0.482661*\text{pH}$$

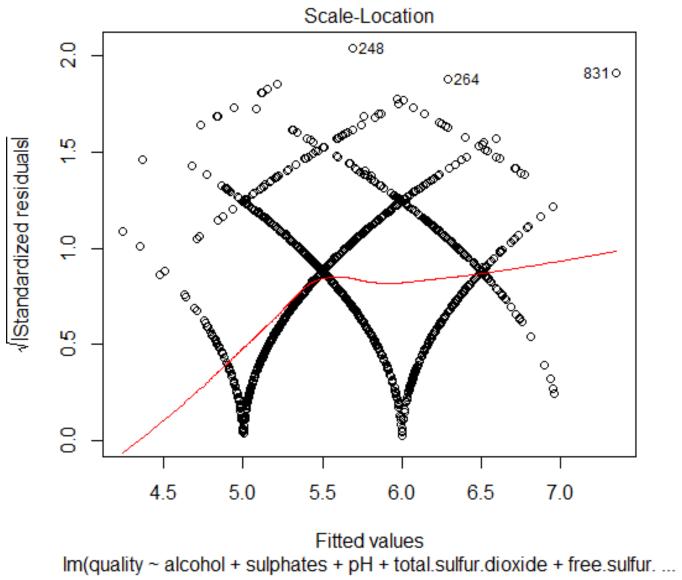
$$- 0.003482 * (\text{Total sulfur Dioxide}) + 0.005077(\text{Free sulfur Dioxide}) - 1.012753(\text{volatile acidity})$$

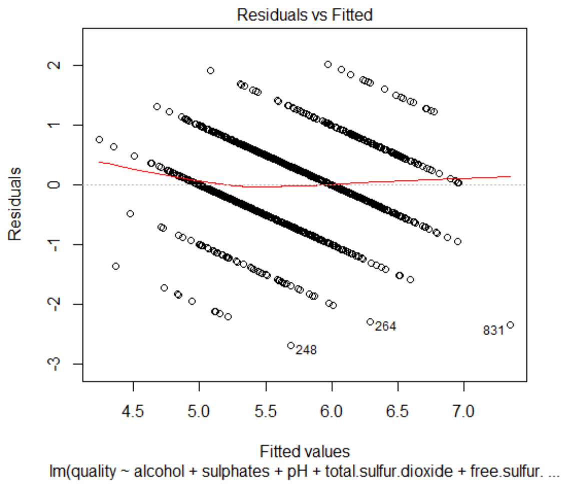
$$+ \text{epsilon}$$

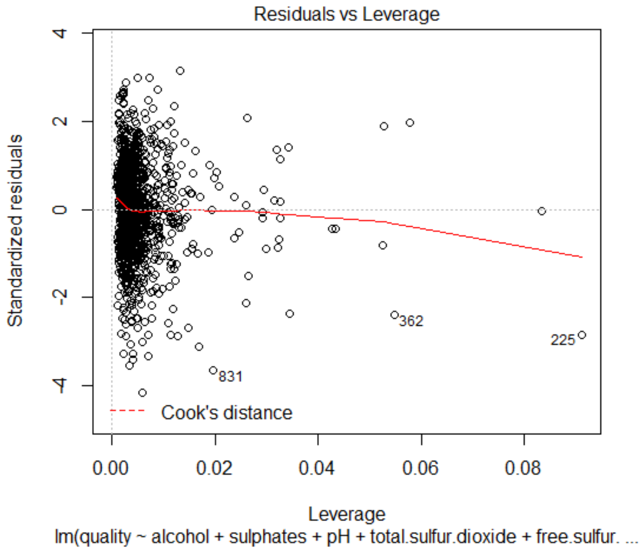
Looking at our model we determined statistically that alcohol, sulphates, pH, total Sulfur Dioxide, free Sulfur Dioxide, and volatile acidity affects the quality of wine.

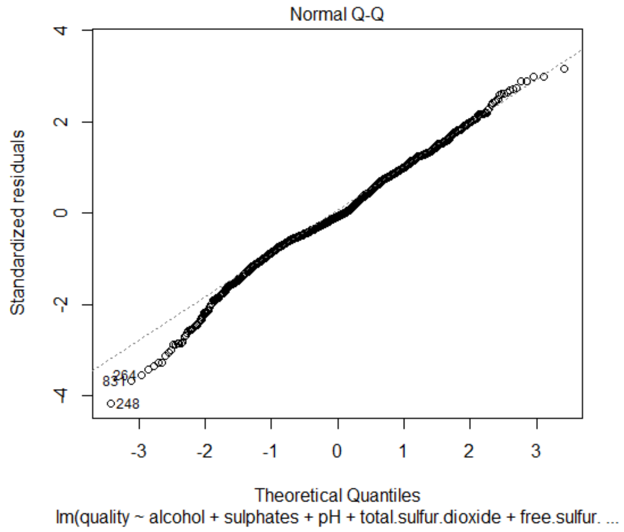
<u>Conclusion:</u>

Looking over our analysis, we were able to determine that there was a significant difference in the quality of a wine and its contents using a two-sample t-test. There was a strong difference between high-grade wine and medium grade-wine, showing us that there is a strong statistical difference between each grade of wine. After that we used a multiple linear regression model to determine what predictors influence the quality of wine. We determined that alcohol, sulphates, pH, total Sulfur Dioxide, free Sulfur Dioxide, and volatile acidity affects the quality of wine. Something that could use improvement on this analysis is to improve the multiple linear regression. Looking at the Scale-location graph and the Residuals vs fitted graph, which can be seen below, shows a weird patter, as seen below, meaning that there is better method that can be used to better identify what affects the quality of wine. I also would be interested to see how much a quality of wine affects its price. However, I was unable to do that because it was not included in the data set.









Appendix:

Data set found on Kaggle: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

U.s Wine consumption data https://wineinstitute.org/our-industry/statistics/us-wine-consumption

Code is attached in blackboard