Allen Deon Saunders

Wine Quality Analysis

1. INTRODUCTION

You want your wine to age like wine - not vinegar. Understanding what makes a wine top shelf is beneficial to not only vineyards - this information is also valuable to the average person. This report shows the relationships between wine characteristics and quality for red and white variants of the Portuguese "Vinho Verde" wine. For vineyards who produce this wine variety a single report can change their perspective on how their process of winemaking affects wine quality. To have a decent understanding of the characteristics of which properties contribute to quality of wine is to be a head above the rest. The process of making wine differs from company to company but even the best creators are always looking for new insights. The public should have an understanding of what makes our wine delicious.

Using data about the wine from two datasets of red and white variants of the Portuguese "Vinho Verde" wine we will explore how characterics like volatile acid, citric acid, chloride concentration, total sulfates dissolved, pH, and more effect the quality of wine. Potential clients can see how these factors affect the quality of wine. These insights will give a winemaker clarity about how their wine may potentially rank against others of the same wine variety.

1.1 Datasets

Two datasets were created, using red and white wine samples. The inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Missing Attribute Values: None. The details are described in [Cortez et al., 2009].

Variables:

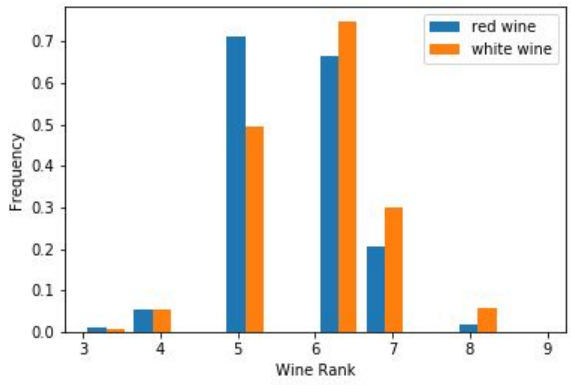
1 - fixed acidity, 2 - volatile acidity, 3 - citric acid, 4 - residual sugar, 5 - chloride, 6 - free sulfur dioxide, 7 - total sulfur dioxide, 8 - density, 9 - pH, 10 - sulphates, 11 - alcohol, 12 - quality (score between 0 and 10)

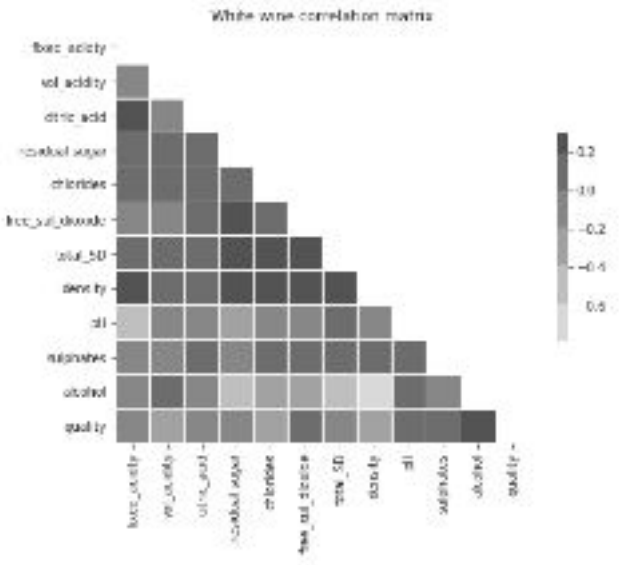
1.2 Data importation and cleaning

The data was imported using PANDAS the python 3rd party open source library used for data analysis. The data was clean and came out of the box with separated values with a semicolon “;” as a separator. No preprocessing was performed. A check to see if there were missing values and there are none.

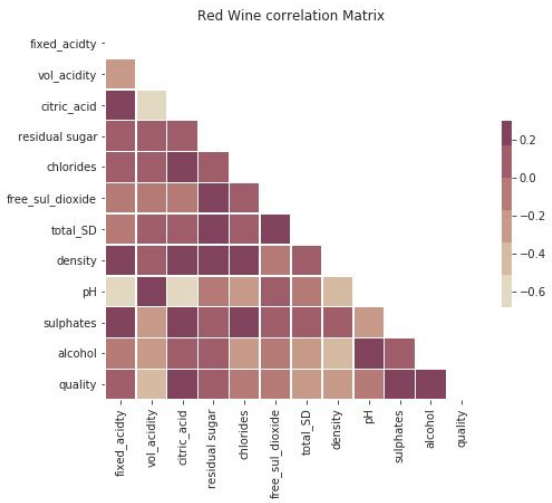
2. DATA EXPLORATION

2.1 Relations





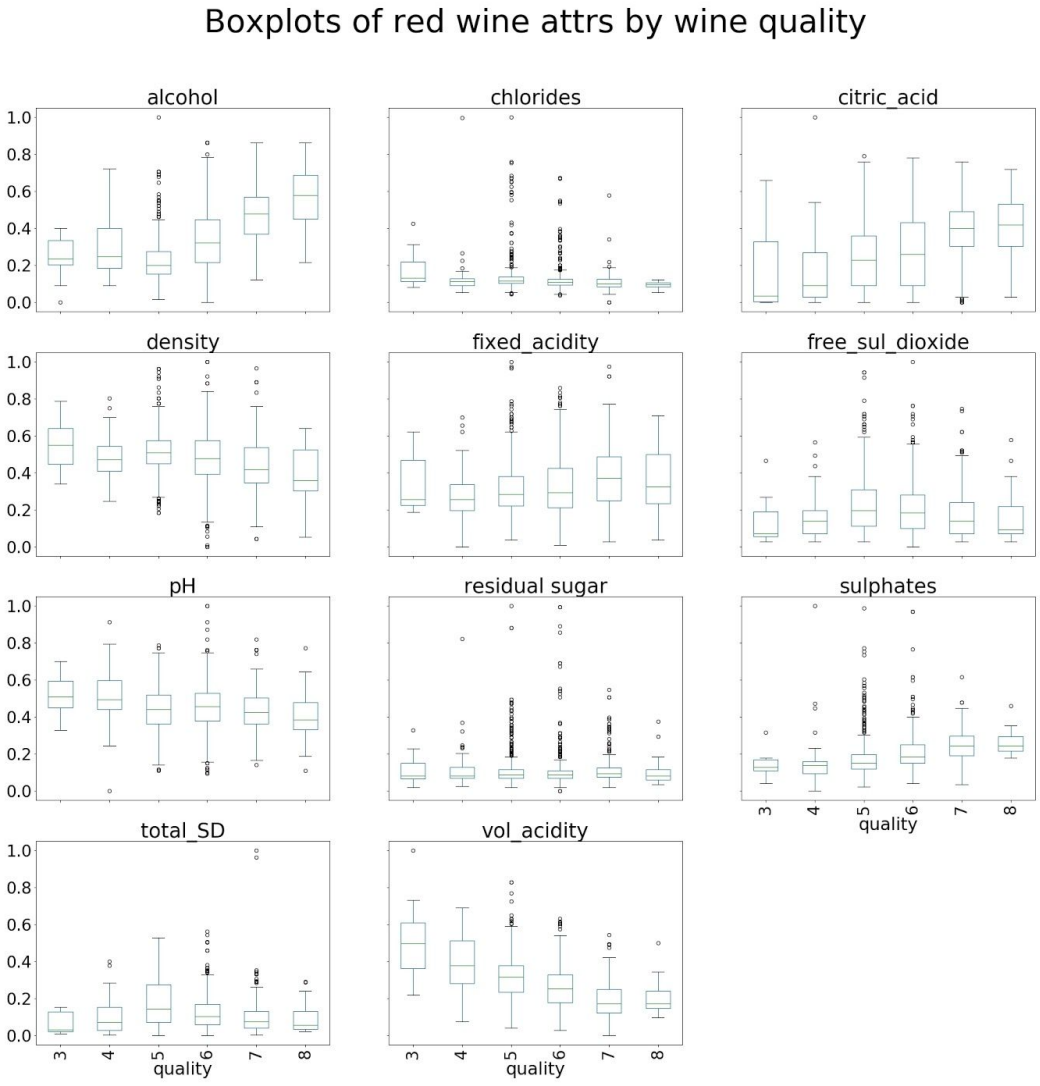
The density was examined with respect to wine quality inspected. We immediately notice the distribution for quality seems normal. Yet, we look for normality in the characterics of the wine with respect to quality. Here it would be best to have a uniform distribution of quality to be able to compare the wine by quality. Because there are a low number of instances of bad and great wine we will pool those together once we start the hypothesis. The groups are low[3,4], mid[5,6], high[>6]. This grouping should eliminate noise from having a small samples of 8,9,3,4 ranking wines. There was also a high variance of among the features with respect to each quality. Normalized boxplots were used to show a side by side comparison of attributes without the scales skewing the message shown by the variance of each feature. Outliers were removed using a interquartile range method before performing any frequentist or Bayesian statistical tactics. To explore the relationships between attributes I used a matrix heatmap for robust visualization.



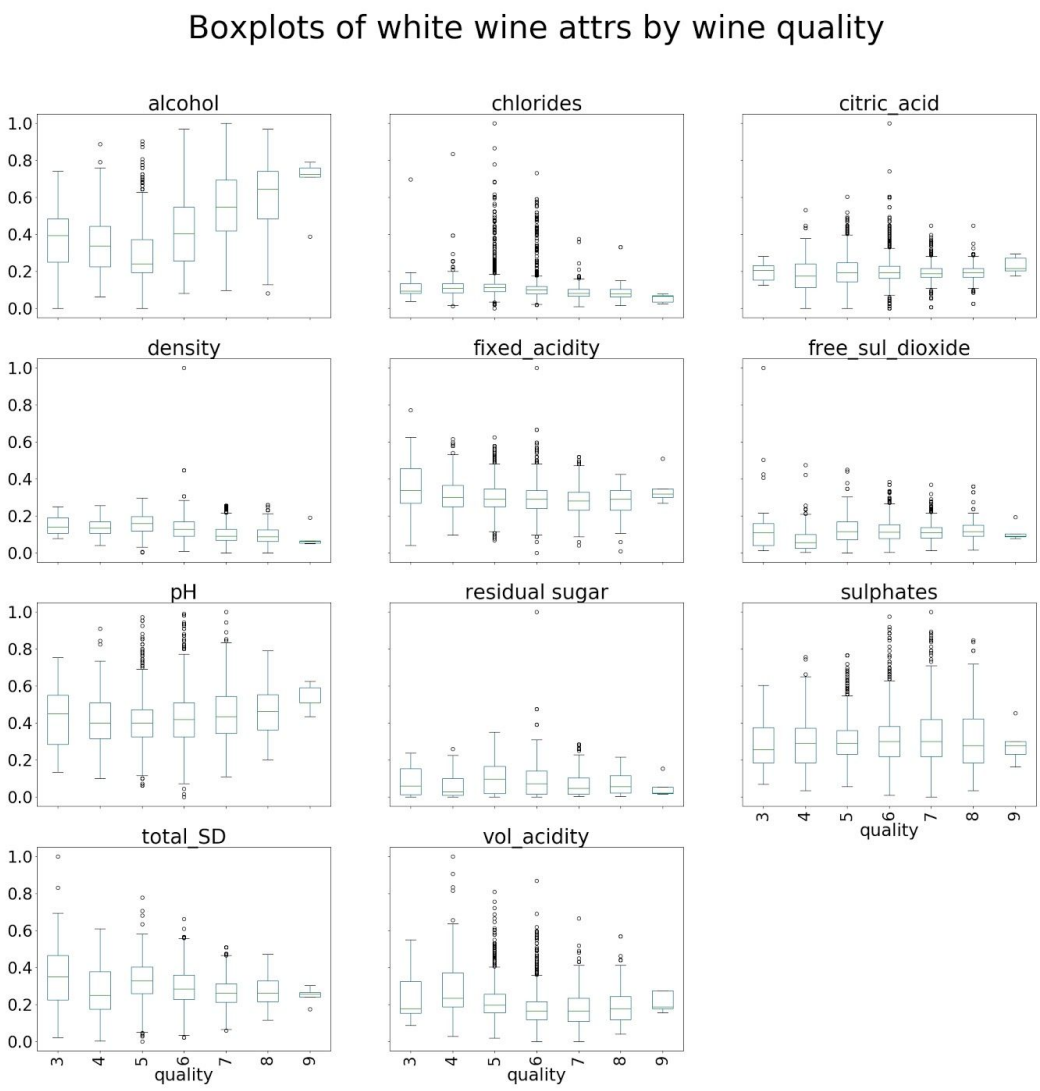
From the matrix plots we can see strong relationships between various combinations of wine attributes for both white and red wine, respectively. The correlations are explored in the regression plots later in the report.

Building a model requires a detailed level of data exploration to show contrast between positive and negative relationship between the target and the attributes. This level of data exploration is required

before any decent model is chosen. For this kind of problem Simple regression is not adequate to predict wine quality based on these features. A simple and dry solution would be to build a fixed effects or random effects model. Anything more simple than these could not do the job. After examining the strongest relationships between attributes I selected pairs or attributes with a threshold using absolute r value of greater than or equal to 0.1



Allen Deon Saunders



The variability was quite large for most of the features so removal of outliers was performed using an interquartile range method before computing any statistical tests. Outliers were removed by selecting any values greater than 3 standard deviations away from the mean. After the outliers were removed a one way ANOVA was computed to show that for each feature, with varying quality, at least one mean was not equal to the rest. From the results of the one way ANOVA we can see that with the exception of residual sugar for red wine all of the attributes of each wine have means that are different. To explore this as a random effects model a comparison among treatment means was executed using a Tukey's test. Also it should be noted that for the white and red wine the higher quality and lower quality had such low numbers that they could not be included in the statistical testing with attributes which had much higher instances. As we can see from the summary all of the groups of wine are fundamental different in quality. And, there is an observed mean difference in the groups. For this reason the dataset was split by quality into groups: low, mid, and high quality. Before attempting to build a model

more complex than basic linear these methods of exploratory data analysis were necessary to uncover which factors affect wine quality.

```
White wine one way ANOVA
Attr: fixed_acidity has a one way ANOVA p-value: 1.1674421647332543e-12
Attr: vol_acidity has a one way ANOVA p-value: 9.092023243656977e-43
Attr: citric_acid has a one way ANOVA p-value: 0.00020681983848938007
Attr: residual sugar has a one way ANOVA p-value: 3.0630453786310768e-21
Attr: chlorides has a one way ANOVA p-value: 1.2450550378908237e-37
Attr: free_sul_dioxide has a one way ANOVA p-value: 9.319482745495827e-13
Attr: total_SD has a one way ANOVA p-value: 2.3997056721219952e-32
Attr: density has a one way ANOVA p-value: 4.083385433988804e-90
Attr: pH has a one way ANOVA p-value: 4.494551992968747e-10
Attr: sulphates has a one way ANOVA p-value: 0.0016456344822937494
Attr: alcohol has a one way ANOVA p-value: 1.1792460124785171e-171


Red wine one way ANOVA
Attr: fixed_acidity has a one way ANOVA p-value: 2.1264300676738575e-06
Attr: vol_acidity has a one way ANOVA p-value: 1.436357968294605e-42
Attr: citric_acid has a one way ANOVA p-value: 1.1309939791709047e-19
Attr: residual sugar has a one way ANOVA p-value: 0.09825353703123613
Attr: chlorides has a one way ANOVA p-value: 0.0002699663035816813
Attr: free_sul_dioxide has a one way ANOVA p-value: 9.639020248327718e-05
Attr: total_SD has a one way ANOVA p-value: 3.91767412687428e-10
Attr: density has a one way ANOVA p-value: 8.805121986213431e-09
Attr: pH has a one way ANOVA p-value: 8.65097637054731e-05
Attr: sulphates has a one way ANOVA p-value: 3.103071132575916e-16
Attr: alcohol has a one way ANOVA p-value: 1.291616119044909e-63
```

Tukey's test for red wines:

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-0.9756	-1.5557	-0.3954	True
H	M	-0.5927	-0.8897	-0.2958	True
L	M	0.3829	-0.1399	0.9056	False

Test above performed against: fixed_acidity

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	0.3187	0.262	0.3753	True
H	M	0.133	0.104	0.162	True
L	M	-0.1856	-0.2367	-0.1346	True

Test above performed against: vol_acidity

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-0.2028	-0.2665	-0.1392	True
H	M	-0.1182	-0.1508	-0.0856	True
L	M	0.0846	0.0272	0.142	True

Test above performed against: citric_acid

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-0.0238	-0.4968	0.4491	False
H	M	-0.2049	-0.447	0.0372	False
L	M	-0.1811	-0.6073	0.2451	False

Test above performed against: residual sugar

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	0.0198	0.0041	0.0355	True
H	M	0.0131	0.005	0.0211	True
L	M	-0.0068	-0.0209	0.0074	False

Test above performed against: chlorides

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-1.9181	-5.4118	1.5756	False
H	M	2.3869	0.5985	4.1752	True
L	M	4.305	1.1567	7.4532	True

Test above performed against: free_sul_dioxide

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-0.445	-11.3469	10.457	False
H	M	14.0575	8.4771	19.638	True
L	M	14.5025	4.6785	24.3265	True

Test above performed against: total_SD

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	0.0007	0.0	0.0013	True
H	M	0.0008	0.0005	0.0012	True
L	M	0.0002	-0.0004	0.0007	False

Test above performed against: density

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	0.0953	0.0438	0.1469	True
H	M	0.0225	-0.0039	0.0489	False
L	M	-0.0728	-0.1193	-0.0264	True

Test above performed against: pH

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-0.1512	-0.2069	-0.0955	True
H	M	-0.0962	-0.1247	-0.0677	True
L	M	0.055	0.0049	0.1052	True

Test above performed against: sulphates

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-1.3022	-1.6291	-0.9752	True
H	M	-1.2653	-1.4327	-1.098	True
L	M	0.0368	-0.2578	0.3315	False

Test above performed against: alcohol

Tukey's test for white wines: H,M,L

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	0.4557	0.2982	0.6132	True
H	M	0.1509	0.0823	0.2196	True
L	M	-0.3048	-0.4539	-0.1558	True

Test above performed against: fixed_acidity

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	0.1106	0.0921	0.1292	True
H	M	0.0117	0.0037	0.0198	True
L	M	-0.0989	-0.1165	-0.0813	True

Test above performed against: vol_acidity

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-0.0184	-0.041	0.0043	False
H	M	0.0118	0.0019	0.0217	True
L	M	0.0302	0.0087	0.0516	True

Test above performed against: citric_acid

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-0.4405	-1.3834	0.5025	False
H	M	1.5362	1.1253	1.9471	True
L	M	1.9767	1.0844	2.869	True

Test above performed against: residual sugar

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	0.0124	0.0084	0.0164	True
H	M	0.0096	0.0078	0.0113	True
L	M	-0.0028	-0.0066	0.001	False

Test above performed against: chlorides

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-7.9166	-11.091	-4.7422	True
H	M	1.4116	0.0282	2.795	True
L	M	9.3282	6.3243	12.3321	True

Test above performed against: free_sul_dioxide

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	4.987	-2.8725	12.8464	False
H	M	17.326	13.9008	20.7511	True
L	M	12.339	4.9017	19.7764	True

Test above performed against: total_SD

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	0.0019	0.0014	0.0025	True
H	M	0.0021	0.0018	0.0023	True
L	M	0.0001	-0.0004	0.0006	False

Test above performed against: density

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-0.0317	-0.06	-0.0035	True
H	M	-0.0344	-0.0467	-0.0221	True
L	M	-0.0027	-0.0294	0.024	False

Test above performed against: pH

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-0.0242	-0.0456	-0.0028	True
H	M	-0.0126	-0.0219	-0.0033	True
L	M	0.0116	-0.0086	0.0318	False

Test above performed against: sulphates

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
H	L	-1.2425	-1.4557	-1.0294	True
H	M	-1.1462	-1.2391	-1.0533	True
L	M	0.0963	-0.1054	0.298	False

Test above performed against: alcohol

Allen Deon Saunders