What Makes a Good Wine?

Analyzing and Predicting Wine Quality

Group 3: Tasheena Narraidoo, Michael Shi, Reynaldo Pena

Introduction

Wine has become increasing popular with a bigger segment of the population. Quality assessment is important to maintain or even improve the quality of wine production based on its best predictors.

Research Questions:

1. Can we find a more simple way to assess wine quality and type?
2. Which variables are the most important in determining wine quality?
3. What is the best way to predict wine quality?

Preliminary Analysis

Description of Original Variables:

fixed.acidity: Fixed acidity makes wines taste sour. Wines with low acidity levels are "flat"	**volatile.acidity:** Volatile acidity can "spoil" the wine, though low levels may add to its complexity
citric.acid: Adds acidity to the wine and may also add some "freshness" to the wine	**residual.sugar**: Determines sweetness of wine
chlorides: Amount of salt in the wine	**free.sulfur.dioxide**: Prevents microbial growth and the oxidation of wine
total.sulfur.dioxide: Amount of free and bound forms of SO2	**density**: Alcohol has low density and sugar has high density, so sweet wines have high density and dry wines have low density
pH: Acidity of wine	**sulphates**: additive that contributes to S02 levels
alcohol: alcohol content of wine	**quality** rating of the wine [0(worst) – 10(best)]

Q1. Principal Component Analysis (PCA)

We performed a PCA analysis in an attempt to reduce the dimensionality of our data sets.

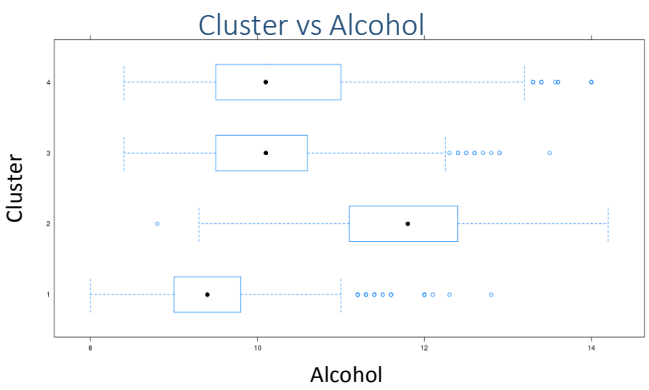
	PC1	PC2	PC3	PC4
fixed.acidity	-0.448	0.426	-0.5989	0.1488
volatile.acidity	-0.689	0.171	0.3583	0.0828
citric.acid	0.255	0.235	-0.7535	-0.0574
residual.sugar	0.556	0.558	0.0967	-0.1163
chlorides	-0.547	0.439	-0.0599	-0.1709
free.sulfur.dioxide	0.737	0.181	0.1268	-0.3135
total.sulfur.dioxide	0.827	0.234	0.1298	-0.1367
density	-0.161	0.903	0.0661	-0.1557
pH	-0.363	-0.249	0.5212	-0.4874
sulphates	-0.523	0.195	-0.2161	-0.6079
alcohol	-0.103	-0.802	-0.2728	-0.0827
quality	0.153	-0.483	-0.3790	-0.4884

PC1: Acidity and salts inversely related with SO2
PC2: Density and sugar strongly inversely related to alcohol
PC3: Fixed acidity, citric acid negatively related to pH
PC4: measure of sulfur dioxide, sulphates, pH, and quality

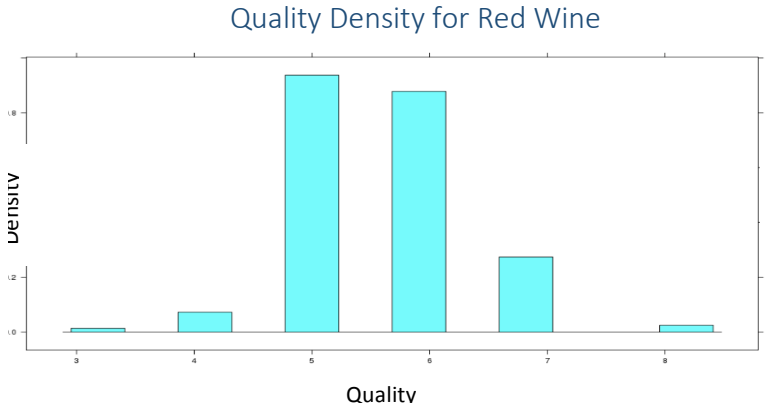
In our PCA, we successfully reduced the dimensionality of our data set into 4 principal components. Looking at these components, we can see some natural grouping in both wine type and quality.

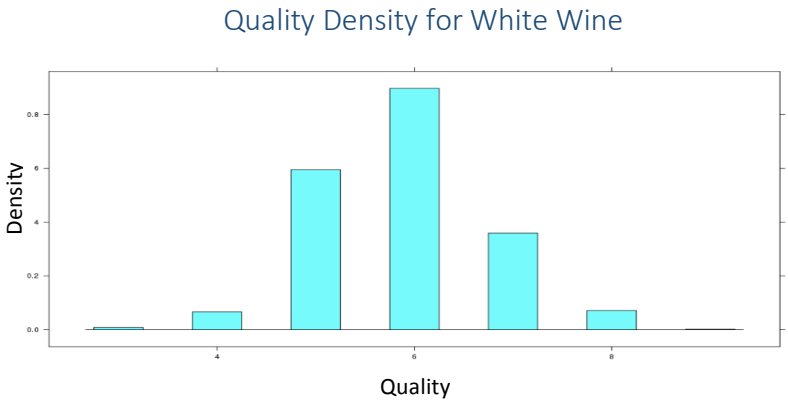
Q2. Cluster Analysis

Our cluster analysis solution showed that alcohol is the most important factor in determining wine quality. The pictures below shows cluster 2 as the cluster with highest alcohol content and also the one with highest quality.



Wine color also seems to affect quality, as shown by the graphs below. Red wine has a large number of observations at quality 5 and 6, while white wine has a large number of observations at quality 6 and 7.





Q3. Random Forest & Support Vector Machine

Random Forest is an ensemble method where many classification trees are generated using bootstrapping. Support Vector Machine is a classification method using machine learning.

Here is a summary of our prediction models based on the above methods:

Accuracy Level		
	Random Forest	**Support Vector Machine**
Red Wine	72.7%	96.7%
White Wine	71.1%	99.7%

Conclusion

It would seem that white wines with higher alcohol content have higher quality. Our model could be used to improve the training of oenology students and set wine prices.

[1] UCI Machine Learning Repository: Wine Quality Data Set <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>