

The background of the slide is a close-up, slightly blurred image of numerous wine corks. The corks are light brown and feature various markings, including vine motifs, the word 'CHATEAU', and dates like 'MP 26/15'.

# Unlocking the Wine Code

A Wine Quality Study

By Chris Atwood

# Introduction

What makes wine good? If you were going to make a wine, what components do you need to ensure that your wine tastes good? Can we examine the composition of wines that we enjoy and unlock the code for the perfect wine? Let's take a look.



# Problem Statements

- What are the most important components that determine the quality of a wine?
- How does a quality red wine differ from a quality white wine?



# The Data

The wine quality dataset comes from the [UCI Machine Learning Repository](#). It consists of 2 datasets for the Portuguese Vinho Verde wine. First is the red wine data set that has 1599 wines with quality rankings from 3 (lowest) to 8 (highest). The second set is the white wine dataset of 4898 wines with quality ranking of 3 (lowest) to 9 (highest). Each of these datasets include the quantities of 12 variables :

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol
- Quality

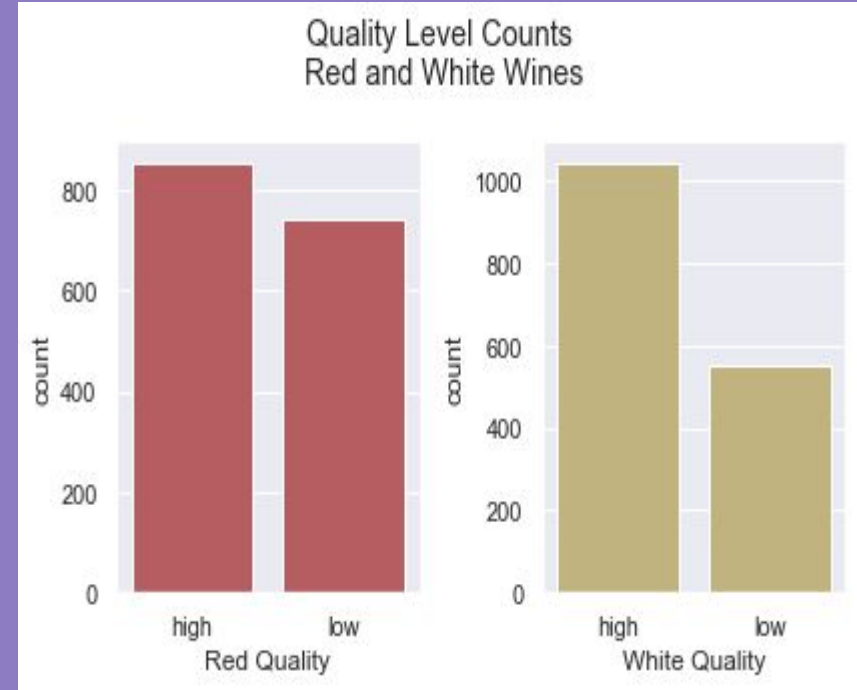


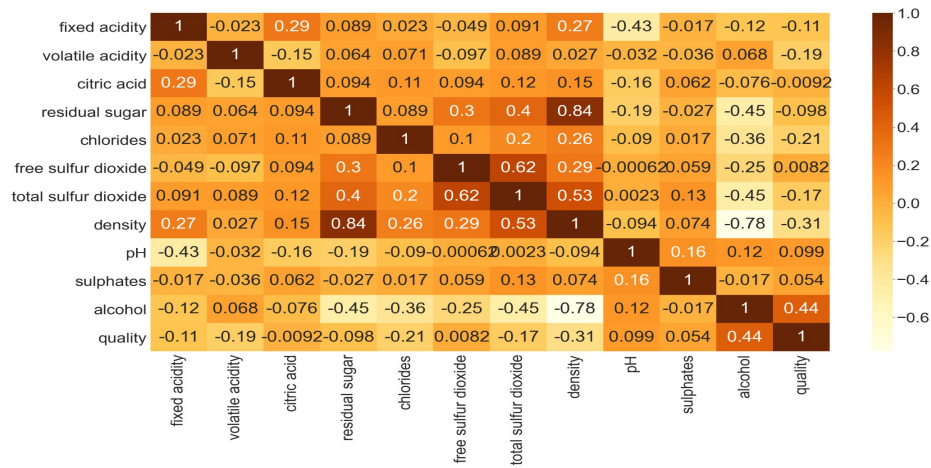
# Exploratory Data Analysis

# Wine Quality Groupings

Both the red wines and the white wines were grouped by their quality level. Wines with a quality ranking of 6 or higher were placed in the High group and wines with a quality ranking of 5 or lower were placed into the low group. The breakdown was:

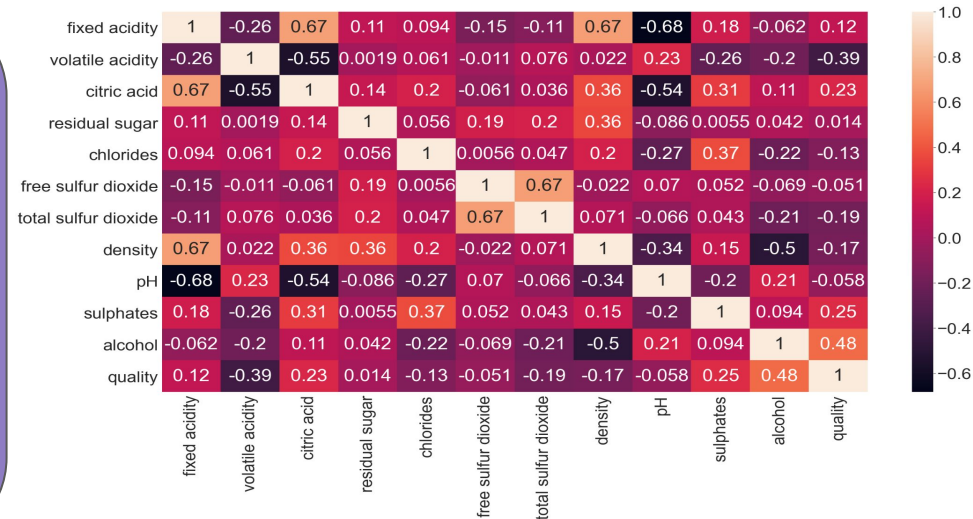
- 3,258 High Quality White Wines
- 1,640 Low Quality White Wines
- 855 High Quality Red Wines
- 744 Low Quality Red Wines



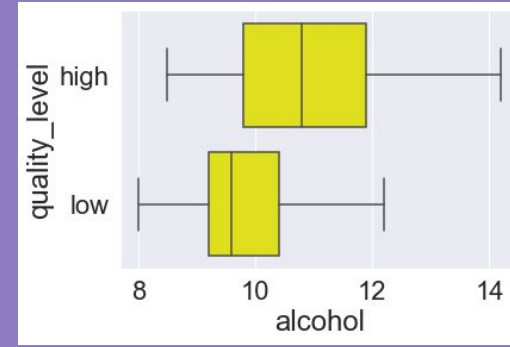
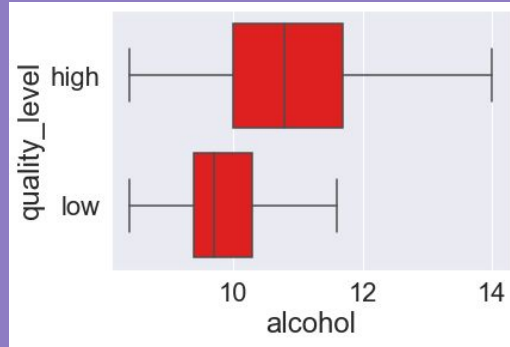


# Feature Correlations

Quality level has the strongest correlation with alcohol content in both kinds of wines. Red wines also have a strong positive correlation between quality and citric acid and sulphates, and a strong negative correlation between quality and volatile acidity. White wines have a strong strong negative correlation between quality and density, chlorides, and volatile acidity.




# Alcohol



High levels of alcohol content clearly produces a higher quality of wine. Alcohol is much more viscous than water and helps to give wine it's "Mouth Feel" . It helps to balance against other attributes such as tannins that are found in red wines and higher levels of sugars found in white wines.

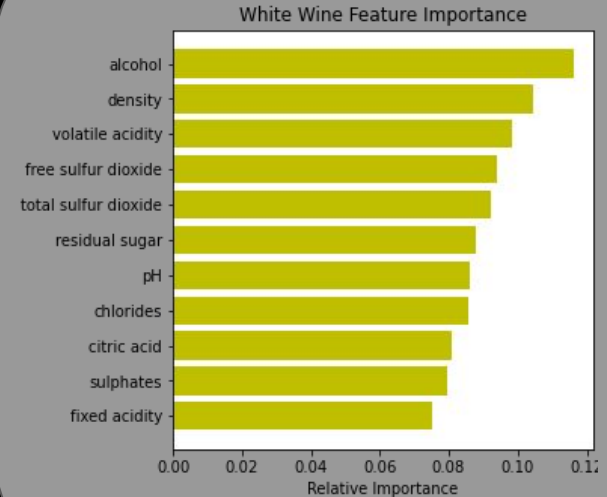
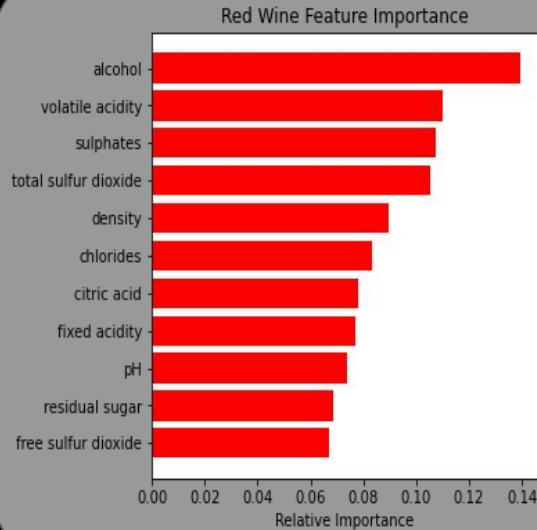




# Machine Learning

# Feature Importance

Using the feature importance method of the Random Forest Classifier, we see that as expected alcohol was the most important feature in predicting a wines quality level. This is followed by Volatile Acidity, Sulphates, and Total Sulfur Dioxide in Red Wine and Density and Volatile Acidity in White Wines.





# Model Selection

## Red Wine

Three different models were tested to best predict the quality of red wines. GridsearchCV was used to find the best parameters

Model Selection Red Wines		
Model	ROC-AUC Score	Optimal Parameters
<b>Random Forest</b>	0.8239	N_estimators: 500 Max depth: 6 Min_samples_leaf: 1 Min_samples_split: 10 Bootstrap: True
<b>K Neighbors Classifier</b>	0.7799	N_neighbors: 324
<b>Logistic Regression</b>	0.8086	C: 1 Max_iter: 100

Model Selection White Wines		
Model	ROC-AUC Score	Optimal Parameters
Random Forest	0.8512	N_estimators: 500 Max depth: 10 Min_samples_leaf: 4 Min_samples_split: 2 Bootstrap: True
K Neighbors Classifier	0.7961	N_neighbors: 24
Logistic Regression	0.7999	C: 100 Max_iter: 100

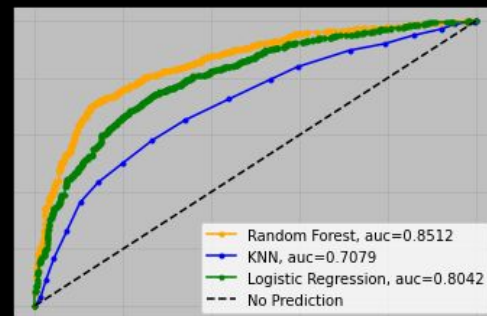
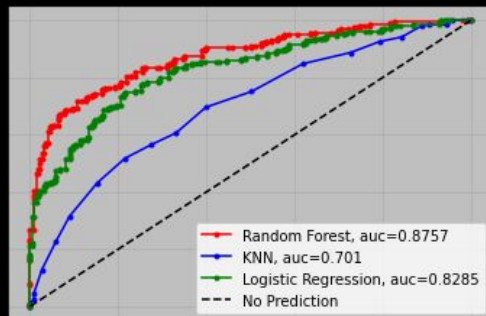
# Model Selection

## White Wine



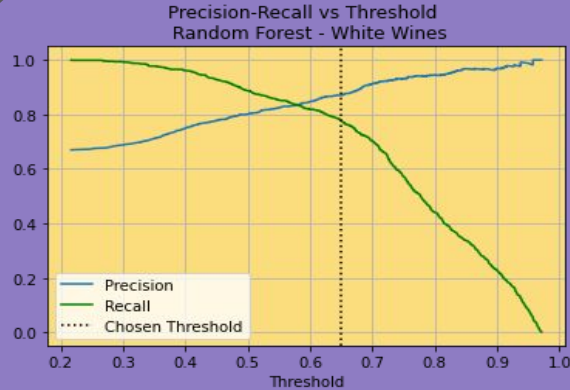


# ROC/AUC Curves



# Thresholding

White Wine



## Classification Report

Random Forest

	Precision	Recall	Threshold
True	0.87	0.79	0.64

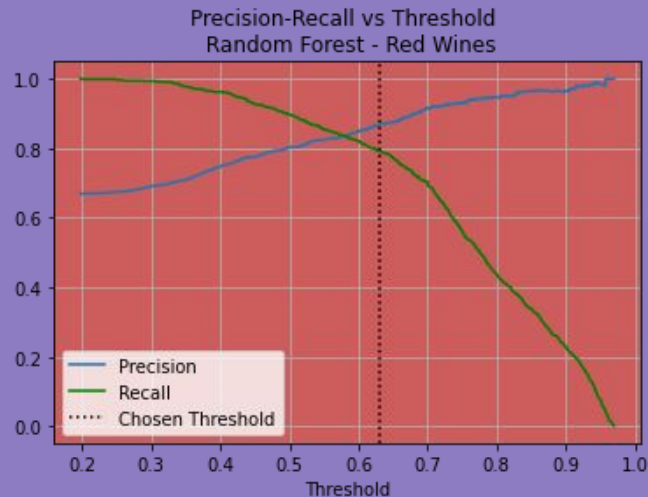
Precision and recall curves for the Random Forest model on white wines show that a threshold value of 0.64 will return the best balance of precision and recall

Classification report			
Random Forest			
	Precision	Recall	Threshold
True	0.87	0.80	0.63

# Thresholding

## Red Wine

Precision and recall curves for the Random Forest model on red wines show that a threshold value of 0.63 will return the best balance of precision and recall



# Conclusion

- High alcohol content is the single biggest determining feature for wine quality.
- In red wines, volatile acidity, sulphates and total sulfur dioxide are other major components that determine the quality of the wine.
- In white wines, density and volatile acidity are the next biggest determining components of its quality.