What Makes a Great Wine?

Analyzing 6,000 Red and White Wines to see if the same attributes make them great wines

Objective: Determine What Makes a Great Wine

-We will build a ML classifier that can predict if a wine is high quality or not.

-If the model is accurate, we will then compare the most impactful coefficients to determining the wine's quality.

- Finally, we will assess if the same attributes contribute to high quality red and white wine.

About the Datasets:

-Red Wine: 1599 rows about distinct Portuguese red wines.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

-White Wine: 4898 rows about distinct Portuguese white wines.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Variables in Datasets

Red Wine Info:

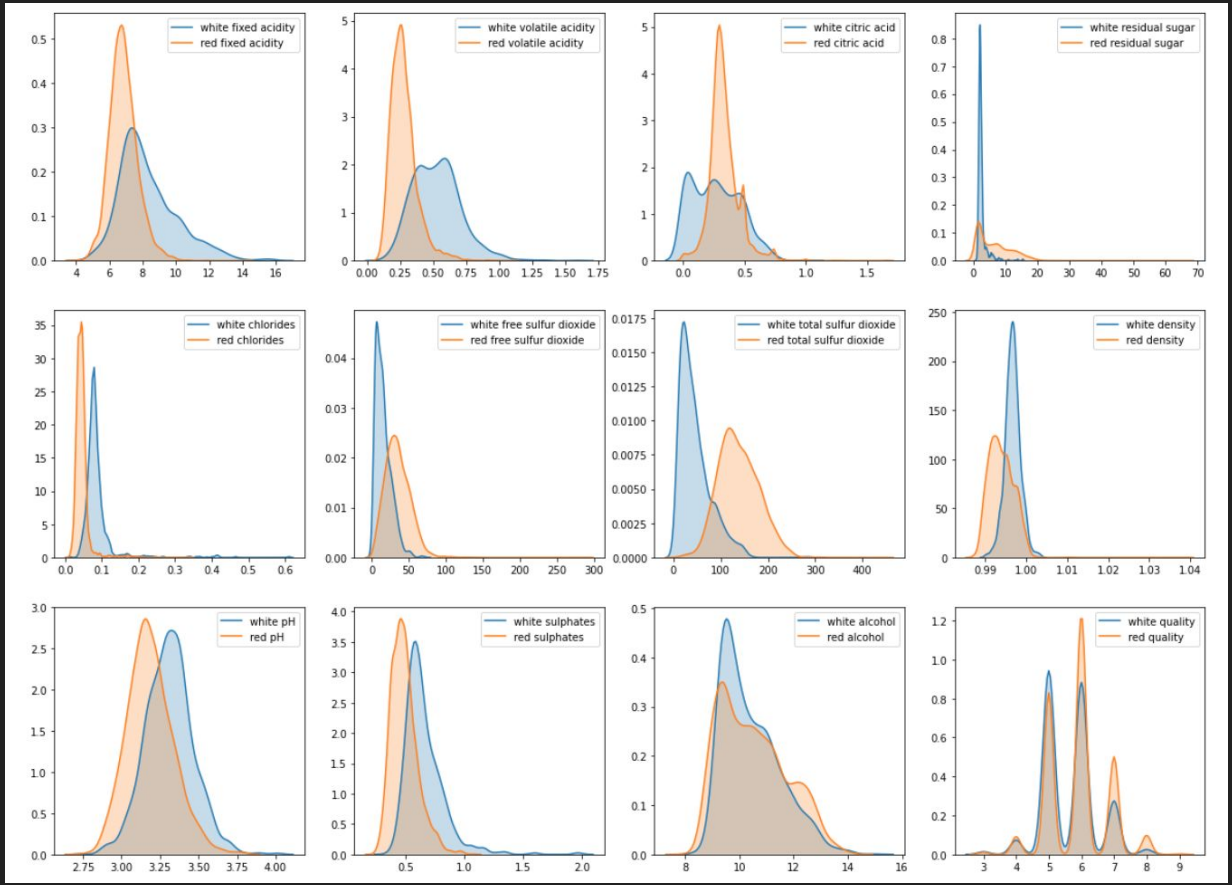
```
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         4898 non-null   float64
 1   volatile acidity      4898 non-null   float64
 2   citric acid           4898 non-null   float64
 3   residual sugar        4898 non-null   float64
 4   chlorides             4898 non-null   float64
 5   free sulfur dioxide   4898 non-null   float64
 6   total sulfur dioxide  4898 non-null   float64
 7   density               4898 non-null   float64
 8   pH                    4898 non-null   float64
 9   sulphates             4898 non-null   float64
 10  alcohol               4898 non-null   float64
 11  quality               4898 non-null   int64
```

White Wine Info:

```
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         4898 non-null   float64
 1   volatile acidity      4898 non-null   float64
 2   citric acid           4898 non-null   float64
 3   residual sugar        4898 non-null   float64
 4   chlorides             4898 non-null   float64
 5   free sulfur dioxide   4898 non-null   float64
 6   total sulfur dioxide  4898 non-null   float64
 7   density               4898 non-null   float64
 8   pH                    4898 non-null   float64
 9   sulphates             4898 non-null   float64
 10  alcohol               4898 non-null   float64
 11  quality               4898 non-null   int64
```

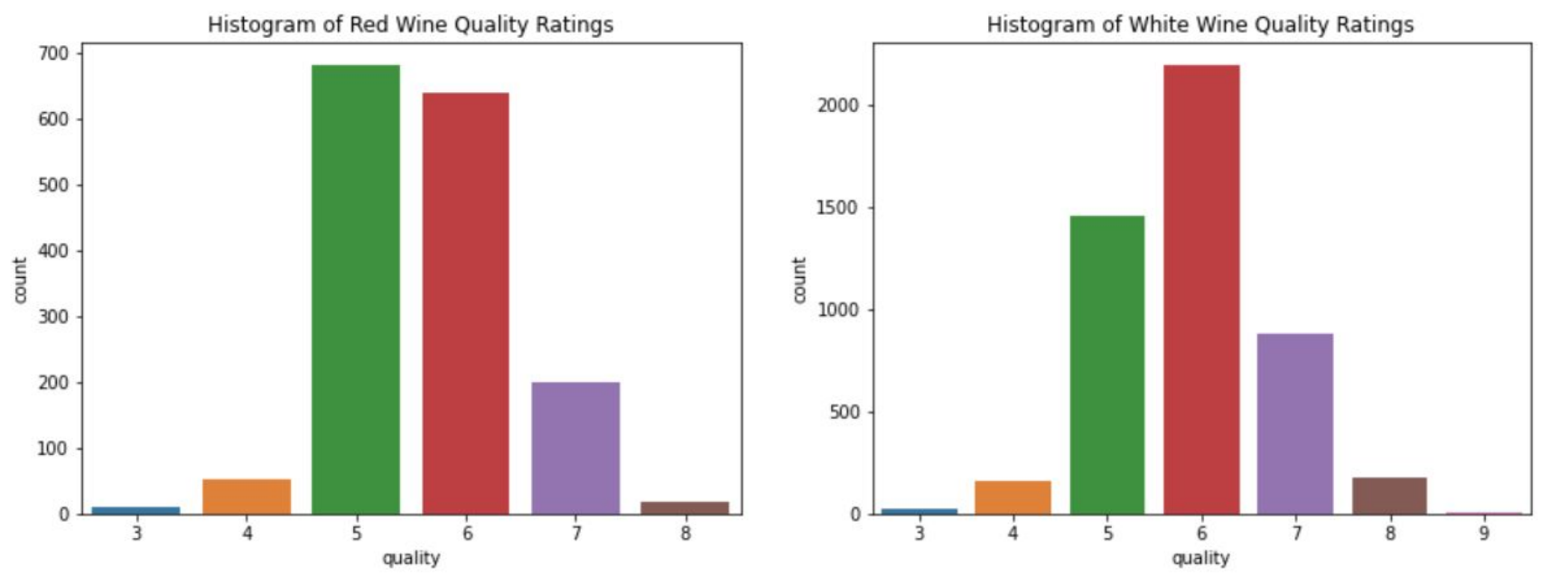
-All columns are continuous numeric variables except for ‘quality’, which is a discrete ordinal numeric variable.

A Quick Comparison of Red vs. White Features



-It is evident that there are significant differences between the red and white distributions.

Red vs. White Quality Scores



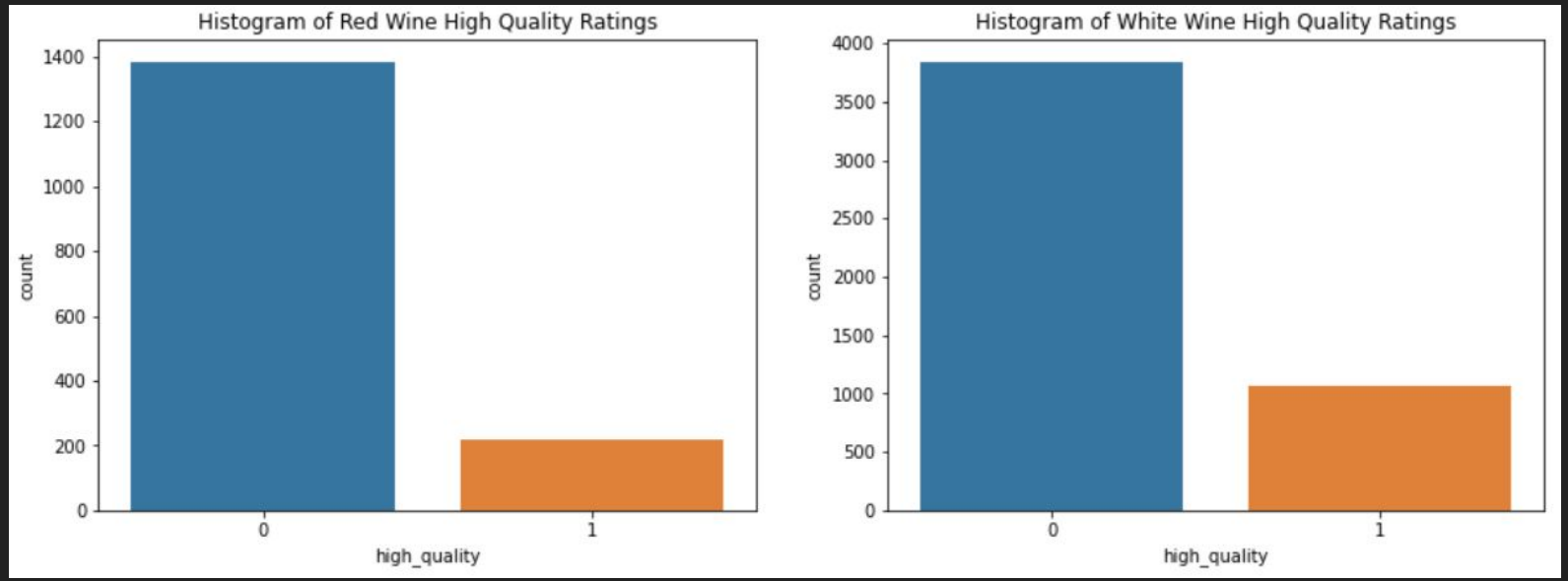
-More scores of 5-6 range for Reds. Whites have a few(5) ratings of 9.

-Overall similar distributions here as well.

Reclassification

For the purpose of this model, we will classify wine's with ratings 7 and above as a high quality wine with a boolean value of 1 and low quality with 0.

Here is the distribution of our new quality feature:



Modeling

- For the purpose of this analysis we will will compare out of the box:
  - Logistic Regression
  - Random Forests
  - XGBoost.

  Given the unbalanced value counts of the high quality wines, we will use F1-score to evaluate these models.

Logistic Regression Classification Report

Red Wine

	precision	recall	f1-score	support
0	0.88	0.98	0.93	342
1	0.62	0.22	0.33	58
accuracy			0.87	400
macro avg	0.75	0.60	0.63	400
weighted avg	0.84	0.87	0.84	400

White Wine

	precision	recall	f1-score	support
0	0.82	0.95	0.88	968
1	0.53	0.21	0.31	257
accuracy			0.80	1225
macro avg	0.68	0.58	0.59	1225
weighted avg	0.76	0.80	0.76	1225

Random Forest Classification Report

Red Wine

	precision	recall	f1-score	support
0	0.92	0.98	0.95	342
1	0.81	0.50	0.62	58
accuracy			0.91	400
macro avg	0.86	0.74	0.78	400
weighted avg	0.90	0.91	0.90	400

White Wine

	precision	recall	f1-score	support
0	0.90	0.97	0.93	968
1	0.83	0.58	0.68	257
accuracy			0.89	1225
macro avg	0.86	0.77	0.81	1225
weighted avg	0.88	0.89	0.88	1225

XGBoost Classification Report

Red Wine

	precision	recall	f1-score	support
0	0.92	0.98	0.95	342
1	0.78	0.50	0.61	58
accuracy			0.91	400
macro avg	0.85	0.74	0.78	400
weighted avg	0.90	0.91	0.90	400

White Wine

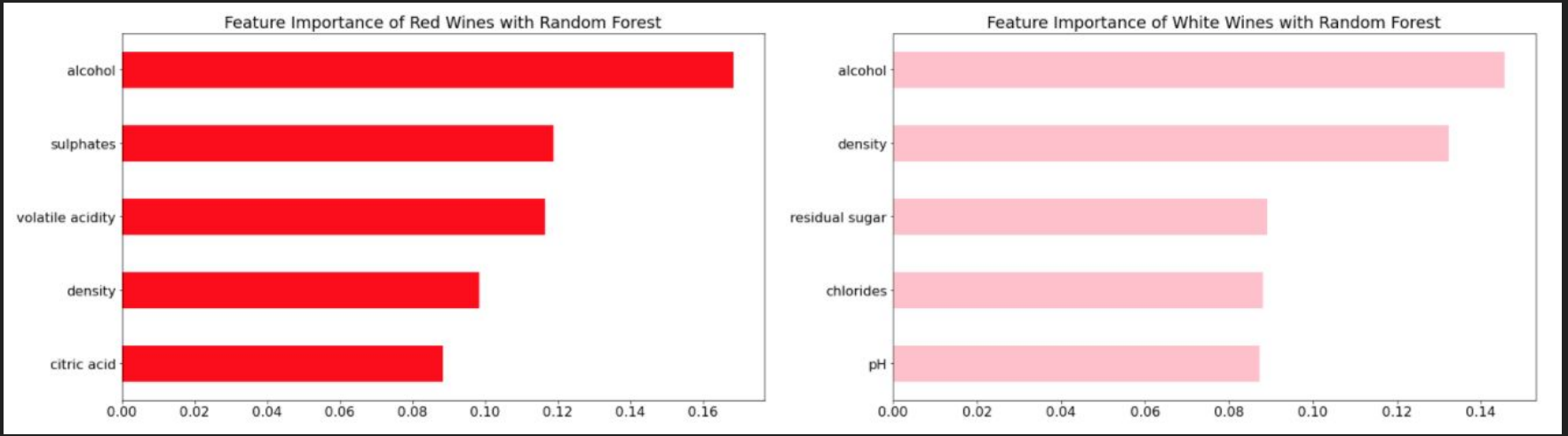
	precision	recall	f1-score	support
0	0.90	0.95	0.93	968
1	0.78	0.61	0.68	257
accuracy			0.88	1225
macro avg	0.84	0.78	0.80	1225
weighted avg	0.88	0.88	0.88	1225

Evaluation

- Random Forest had F1 scores of 0.62(red) and 0.68(white) and XGBoost had F1 scores of 0.61(red) and 0.68(white).

- For further analysis and comparison of results we use the Random Forest Model for the next step of evaluation, comparing the the most important features between red and whites at predictioning high quality wine.

Feature Importance for Predicting Wine Quality



Analysis: For red and white models, alcohol is greatest predictor for quality. Beyond that, for reds, sulphates and volatile acidity are the next most important factors. For whites, density and residual sugar are the next most important features.

Conclusion: From these results we can conclude that beyond higher alcohol content, there are distinct attributes that determine a high quality wine. However, noting the Recall score for Reds was 0.5 but for Whites was 0.58, perhaps having a larger sample size would have increased model performance and highlighted different features as important.

So next time you’re buying a bottle of wine, ask the Sommelier for a red wine high in alcohol, sulphides, and volatile acidity or a white wine high in alcohol, density, and residual sugar!

Cheers!