Wine Quality Prediction

Statistical Analysis

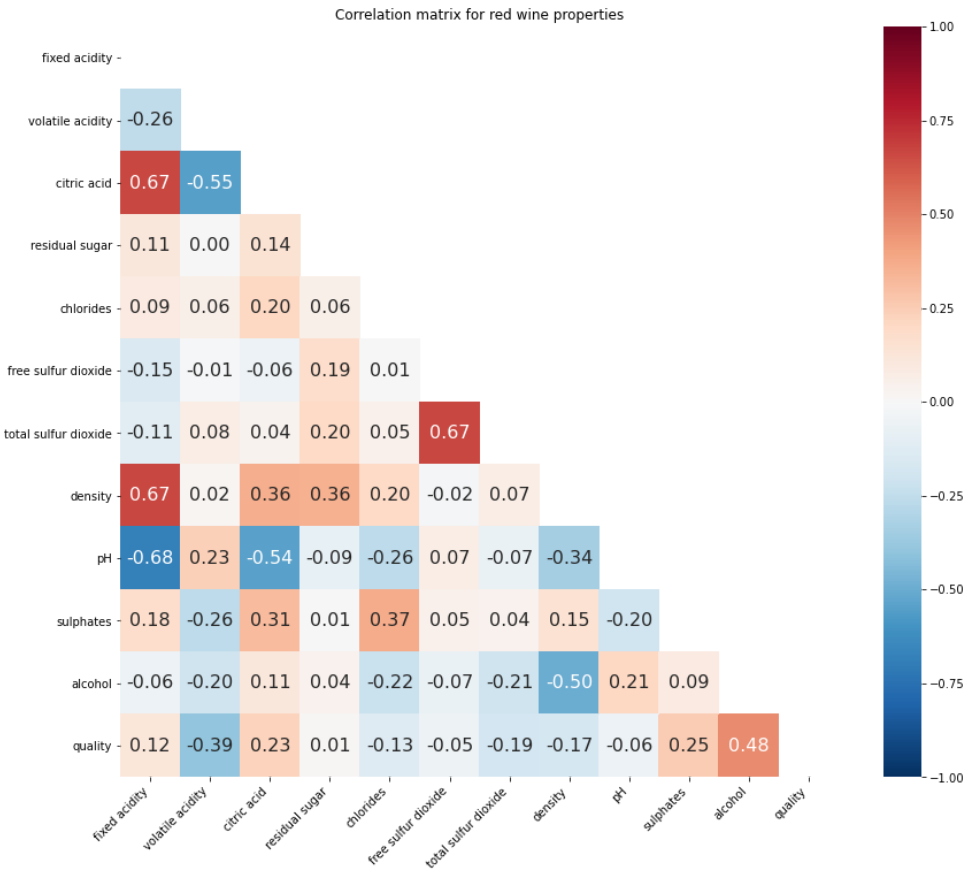
OVERVIEW

Ater visualizing the data trends via EDA, we now start with statistical analysis to check if the relationships are significant. This document overviews all the statistical tests performed on the wine quality dataset.

STATISTICAL ANALYSIS

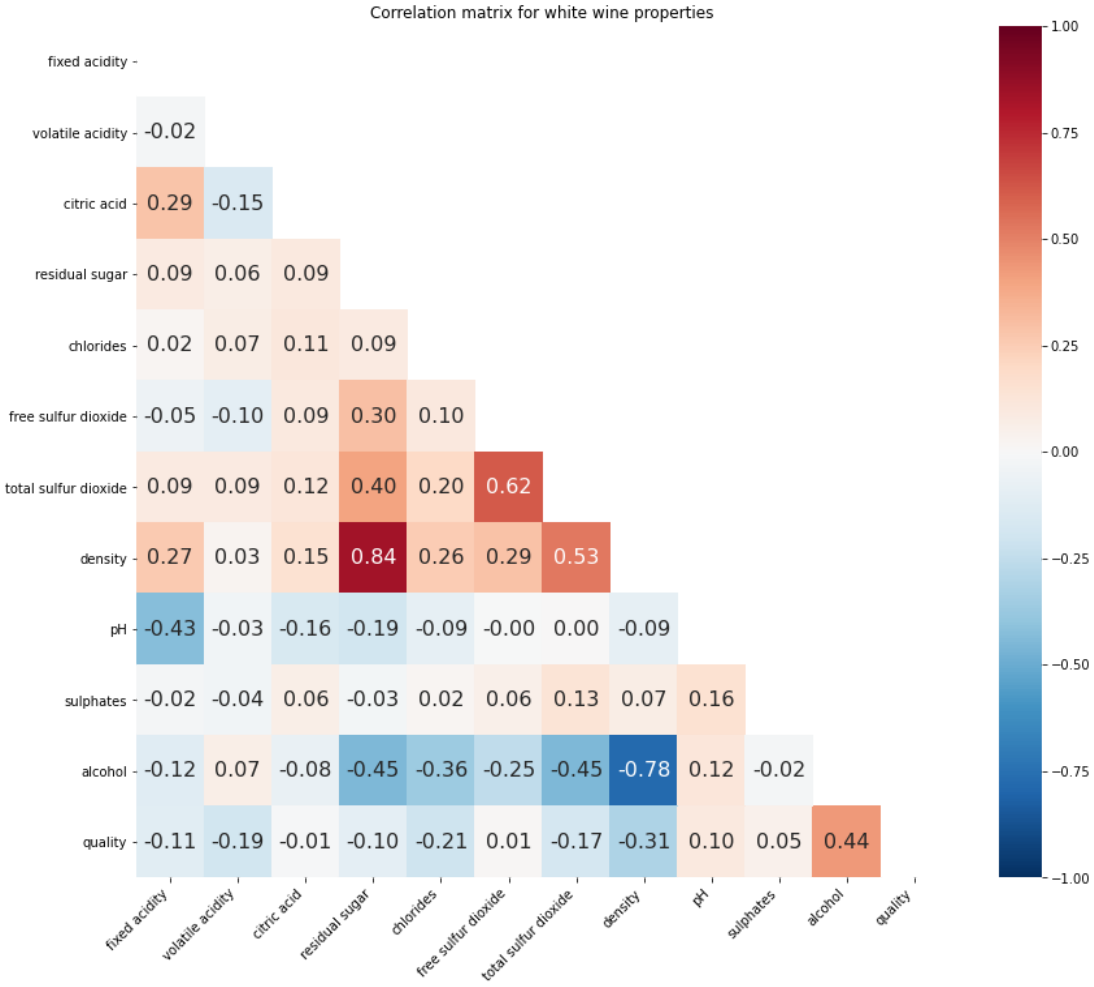
- **Correlation Matrix:** Correlation matrix is a good way to summarize a large amount of data in order to see patterns. We can check if the variables are highly correlated with each other. We can better understand the relationships between our features, which is why I plotted correlation matrices for red and white wines to explore how different features have an effect on the wine quality.

1. Red Wine correlations



The fixed acidity feature of red wines seems to have strong/ moderate correlations with pH, density and citric acid. Free sulfur dioxide has strong correlation with total sulfur dioxide, which might be expected. Alcohol has a correlation of 0.48, which is the maximum correlation any feature has with quality! pH has the least correlation of -0.06 with the quality feature of red wines.

2. White Wine correlations



White wines seem to have strong/moderate correlations between residual sugar, alcohol with density, alcohol and residual sugar, pH and fixed acidity, alcohol and density with quality. We can say there is multicollinearity between density, residual sugar and alcohol. (correlation > 0.7)

- **Tukey Test:** Next, we visualize how the quality rating of wines changes with each feature, seeing if we can find any patterns.We try to find answers for questions like: do highly rated wines have greater alcohol content? Do low rated wines have higher acidity? For this reason, we have divided the quality ratings of red and white wines into 3 categories: Low (1-4), Medium (5-6) and High (7-10) and we make boxplots for all possible statistically

significant features. Significance testing for the box plots below is done via a Tukey test, which compares the means of all treatments to the mean of every other treatment.

1. Density VS Quality

We performed the Tukey test to determine if the relationship between the quality groups is statistically significant.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
high	low	0.0003	0.2788	-0.0002	0.0008	False
high	medium	0.0005	0.001	0.0003	0.0008	True
low	medium	0.0003	0.3221	-0.0002	0.0007	False

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
high	low	0.0019	0.001	0.0014	0.0024	True
high	medium	0.0019	0.001	0.0017	0.0021	True
low	medium	-0.0001	0.9	-0.0005	0.0004	False

The results of the Tukey test for quality groups of red and white wine for density feature are shown in the figures above respectively.

For red wines, the Tukey test indicates that there could be a statistically significant relationship between the medium and high quality groups of red wine, since the Null hypothesis rejection is true. This indicates that the change in density is more marked for red wines rated between 5-7 (medium) and those rated above 7 (high quality).

Tukey test for white wines indicates that there could be a marked change in densities between the high and low quality white wines (which could be expected) and between high and medium quality white wines too. This might mean that overall high quality white wines seem to have more density.

2. Volatile Acidity VS Quality

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
high	low	0.1661	0.001	0.1187	0.2134	True
high	medium	0.1185	0.001	0.0942	0.1427	True
low	medium	-0.0476	0.0243	-0.0902	-0.0049	True

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
high	low	0.0274	0.001	0.0135	0.0412	True
high	medium	0.0074	0.0114	0.0014	0.0135	True
low	medium	-0.02	0.0011	-0.0331	-0.0068	True

The results of the Tukey test for quality groups of red and white wine for volatile acidity feature are shown in the figures above respectively.

For red and white wines, there seems to be a statistically significant relationship between all the quality groups. The volatile acidity for all the quality groups seems to be markedly different from each other! The mean difference is negative for low and medium quality red and white wines. So higher quality red and white wines seem to have more volatile acidity. As the wine quality rating increases, the volatile acidity content might be increasing too!

3. Citric Acid VS Quality

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj  lower   upper  reject
-------------------------------------------------
  high    low  -0.1738 0.001   -0.23 -0.1175   True
  high medium  -0.0869 0.001 -0.1157 -0.0581   True
   low medium   0.0868 0.001  0.0362  0.1375   True
-------------------------------------------------
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj   lower   upper  reject
-------------------------------------------------
  high    low  -0.0346  0.001 -0.0517 -0.0175   True
  high medium  -0.0039 0.4408 -0.0113  0.0036  False
   low medium   0.0308  0.001  0.0146  0.0469   True
-------------------------------------------------
```

The results of the Tukey test for quality groups of red and white wine for citric acid feature are shown in the figures above respectively.

Citric acid content seems to vary significantly between low, medium and high quality groups for red wines. The mean difference is negative for high quality red wines when compared to low and medium quality ones. So maybe high quality red wines have a lower amount of citric acid! The citric acid content for low quality groups of white wines seems to be different when compared to other quality groups.

As with red wines, even for low and high quality white wines, the mean difference is negative, meaning high quality white wines might have lower amounts of citric acid in them! If we extrapolate this logic then it is to be expected that the mean difference is positive for low and medium quality white wines. This means citric acid content is more for lower quality white wines!

4. Alcohol VS Quality

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj  lower   upper  reject
-------------------------------------------------
  high    low  -0.9368 0.001  -1.218 -0.6557   True
  high medium  -0.9368 0.001 -1.0807 -0.7929   True
   low medium      0.0   0.9 -0.2533  0.2534  False
-------------------------------------------------
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj   lower   upper  reject
-------------------------------------------------
  high    low   -0.874  0.001 -1.0668 -0.6812   True
  high medium  -0.8052  0.001 -0.8893 -0.7212   True
   low medium   0.0688 0.6381 -0.1137  0.2512  False
-------------------------------------------------
```

The results of the Tukey test for quality groups of red and white wine for quality feature are shown in the figures above respectively.

The alcohol content for high quality red and white wines seems to vary significantly when compared to other quality groups. There seems to be a negative difference in means indicating that the alcohol content might be lower for higher quality red and white wines!

FEATURE SELECTION AND STATISTICAL ANALYSIS

VIF Scores

I then checked VIF scores for the features. The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to assess accurately the contribution of predictors to a model. The more VIF increases, the less accurate our coefficients will be as indicators of the effect of a variable. In general, a VIF above 10 indicates high correlation and might be a cause for concern. But the VIF scores for the red and white wine dataset were below 5!

Recursive Feature Elimination (RFE) using logistic regression

RFE is a wrapper-type feature selection algorithm. This means that a different machine learning algorithm is given and used in the core of the method, is wrapped by RFE, and used to help select features. RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains.

After scaling the red and white wines dataset within the range of 0 to 1 using sklearn's MinMaxScaler, we implemented sklearn's Recursive Feature Elimination (RFE) to check if any of the wine features are unnecessary or redundant. Features were ranked using RFE's ranking_ and support_ attribute. RFE returned all the 11 features as important. So I built a Logistic Regression model using statsmodels, with all the 11 features.

Logistic Regression using statsmodels

The aim was to build a classifier which can differentiate between good and poor quality wines. For this reason all wines rated 7 and above were classified as "good quality" and the ones below 7 as "poor quality".

Logit Regression Results

Dep. Variable:	good quality	No. Observations:	1599
Model:	Logit	Df Residuals:	1588
Method:	MLE	Df Model:	10
Date:	Mon, 21 Sep 2020	Pseudo R-squ.:	0.3100
Time:	12:42:49	Log-Likelihood:	-438.10
converged:	True	LL-Null:	-634.96
Covariance Type:	nonrobust	LLR p-value:	2.040e-78

	coef	std err	z	P>|z|	[0.025	0.975]
fixed acidity	0.0593	0.081	0.736	0.462	-0.099	0.217
volatile acidity	-3.0907	0.767	-4.031	0.000	-4.593	-1.588
citric acid	0.2658	0.829	0.321	0.749	-1.360	1.891
residual sugar	0.1431	0.062	2.310	0.021	0.022	0.265
chlorides	-10.0446	3.564	-2.819	0.005	-17.029	-3.060
free sulfur dioxide	0.0135	0.012	1.091	0.275	-0.011	0.038
total sulfur dioxide	-0.0176	0.005	-3.480	0.001	-0.028	-0.008
density	-9.9319	3.363	-2.953	0.003	-16.524	-3.340
pH	-0.9416	0.858	-1.098	0.272	-2.623	0.739
sulphates	3.4468	0.527	6.546	0.000	2.415	4.479
alcohol	0.9707	0.090	10.765	0.000	0.794	1.147

Logistic Regression results for unscaled red wine data

After scaling the data, all red wine features except for fixed acidity, citric acid, free sulfur dioxide and density were found to have a statistically significant effect on quality.

Logit Regression Results

Dep. Variable:	good quality	No. Observations:	4898
Model:	Logit	Df Residuals:	4887
Method:	MLE	Df Model:	10
Date:	Mon, 21 Sep 2020	Pseudo R-squ.:	0.1803
Time:	12:39:55	Log-Likelihood:	-2097.1
converged:	True	LL-Null:	-2558.4
Covariance Type:	nonrobust	LLR p-value:	8.521e-192

	coef	std err	z	P>|z|	[0.025	0.975]
fixed acidity	0.0772	0.056	1.380	0.168	-0.032	0.187
volatile acidity	-3.9316	0.482	-8.151	0.000	-4.877	-2.986
citric acid	-0.8858	0.397	-2.228	0.026	-1.665	-0.107
residual sugar	0.0631	0.010	6.309	0.000	0.044	0.083
chlorides	-17.9156	3.884	-4.613	0.000	-25.527	-10.304
free sulfur dioxide	0.0127	0.003	4.186	0.000	0.007	0.019
total sulfur dioxide	-0.0032	0.001	-2.232	0.026	-0.006	-0.000
density	-14.2625	1.332	-10.706	0.000	-16.873	-11.652
pH	1.2800	0.297	4.315	0.000	0.699	1.861
sulphates	1.3010	0.320	4.062	0.000	0.673	1.929
alcohol	0.8604	0.044	19.604	0.000	0.774	0.946

Logistic Regression results for scaled red wine data

For white wines, after scaling the data, except for pH and sulphates, all other features had a significant effect on quality.

Odds Ratio on unscaled data

I then checked the odds ratio for the logistic regression model features. Each estimated coefficient is the expected change in the log odds of wine being of good quality (rated<7) for a unit increase in the corresponding predictor variable holding the other predictor variables constant at certain value. Each exponentiated coefficient is the ratio of two odds, or the change in odds in the multiplicative scale for a unit increase in the corresponding predictor variable holding other variables at a certain value.

sulphates	31.400246
alcohol	2.639836
citric acid	1.304531
residual sugar	1.153877
fixed acidity	1.061138
free sulfur dioxide	1.013586
total sulfur dioxide	0.982516
pH	0.389994
volatile acidity	0.045469
density	0.000049
chlorides	0.000043

sulphates	3.673049e+00
pH	3.596510e+00
alcohol	2.364114e+00
fixed acidity	1.080260e+00
residual sugar	1.065141e+00
free sulfur dioxide	1.012788e+00
total sulfur dioxide	9.968211e-01
citric acid	4.123928e-01
volatile acidity	1.961138e-02
density	6.395571e-07
chlorides	1.657071e-08

Odds ratio for red and white wines (left and right) respectively.

Odds ratio was highest for sulphates in case of red wines and white wines. It was the least for chlorides in case of red wines and white wines.

An odds ratio of 31.40 for sulphates in case of red wines means we will see a 3140% increase in the odds of a red wine being of good quality for a 1 g/dm^3 increase in sulphates! Similarly, there will be a 163% increase in the odds of a red wine being of good quality for a 1 % by volume increase in alcohol content. And a 30% increase in the odds of a red wine being of good quality for a 1 g/dm^3 increase in citric acid. But the p-value was not significant for citric acid indicating that the odds ratio won't be statistically significant.

For white wines, 267% increase in the odds of a white wine being of good quality for a 1 g/dm^3 increase in sulphates. And 259% and 136% increase in odds for good quality white wine for 1 mg/dm^3 increase in pH and 1 % by volume alcohol content! Since the p-value was < 0.01, these results are statistically significant.

CONCLUSION

After extensive statistical tests it is now clear that sulphates and alcohol might be the major features contributing to good quality! We now move on to the machine learning section of the project in order to continue the classification of wines as good or poor quality.