Wine Quality Prediction

Data Story

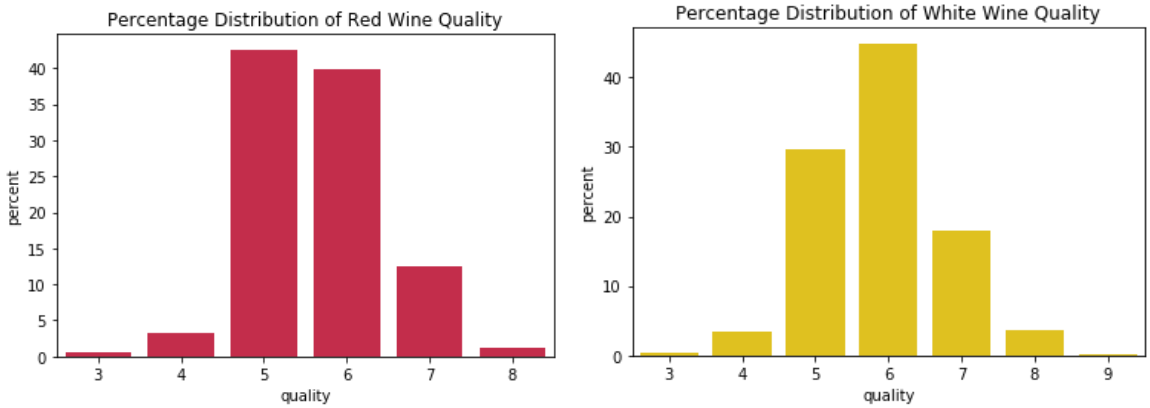
OVERVIEW

After data wrangling, the dataset is now ready for Exploratory Data Analysis! This document reviews the interesting trends in the data, how they were investigated and the resulting visualizations and conclusions.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) which will help us visualize the trends and behavior of the wine quality dataset and ultimately give us necessary pointers as we dive in getting statistical significance for the features and finally build a model for quality prediction of wines.

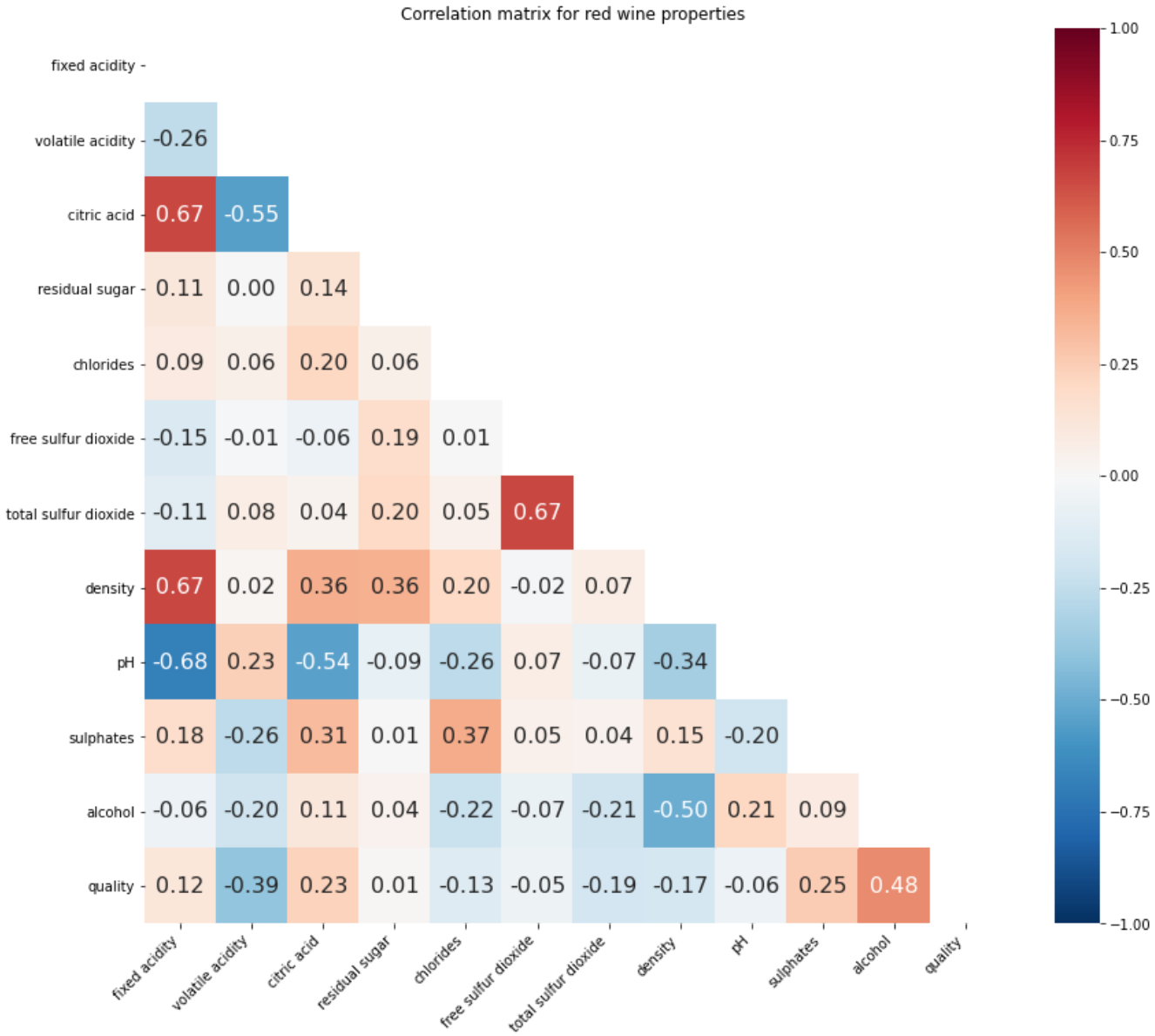
- **Bar Charts:** Let us first visualize the quality distribution of these wines.



As we can see, a majority of red and white wines (about 85%) have a quality rating of 5 and 6. There are no 9s for red wine, very few wines (less than 1%) with a quality rating of 3, and only 2% (for red wines) with a rating of 8. Since we have very few wines of exceptional quality or poor quality, it might add some bias or make it tricky to build a model which accurately selects wines of high (7 or above rating) or low quality (rating below 3).

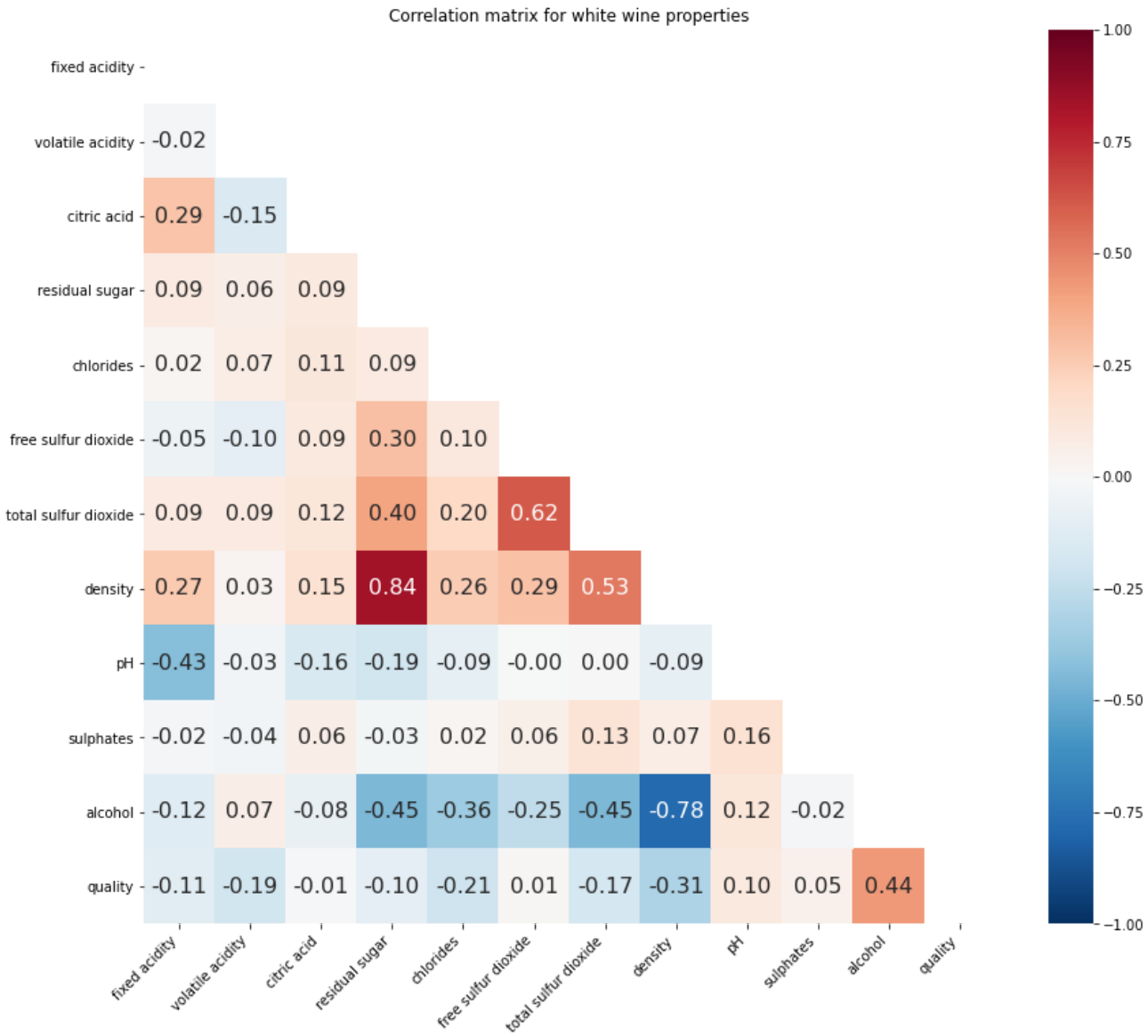
- **Correlation Matrix:** Correlation matrix is a good way to summarize a large amount of data in order to see patterns. We can check if the variables are highly correlated with each other. We can better understand the relationships between our features, which is why I plotted correlation matrices for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

1. Red Wine correlations



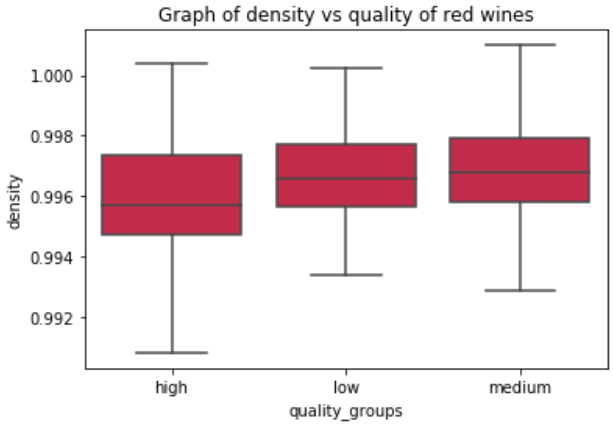
The fixed acidity feature of red wines seems to have strong/ moderate correlations with pH, density and citric acid. Free sulfur dioxide has strong correlation with total sulfur dioxide, which might be expected. Alcohol has a correlation of 0.48, which is the maximum correlation any feature has with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

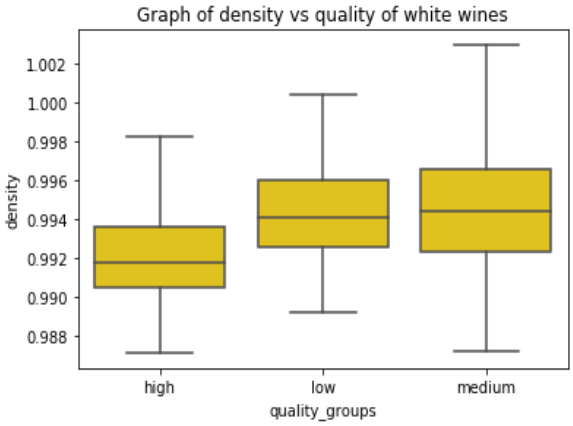
2. White Wine correlations



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)

- **Box Plots:** 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.
    1. **Density VS Quality**

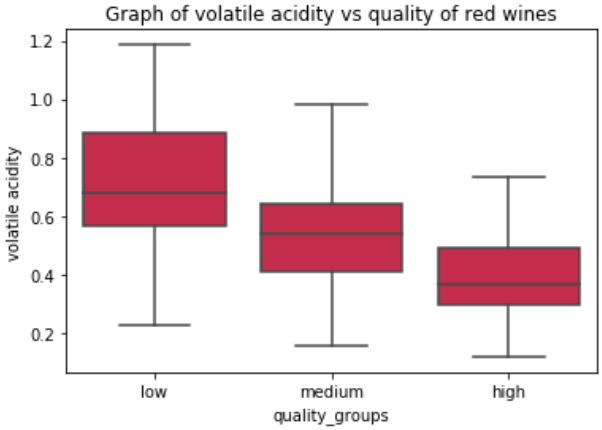


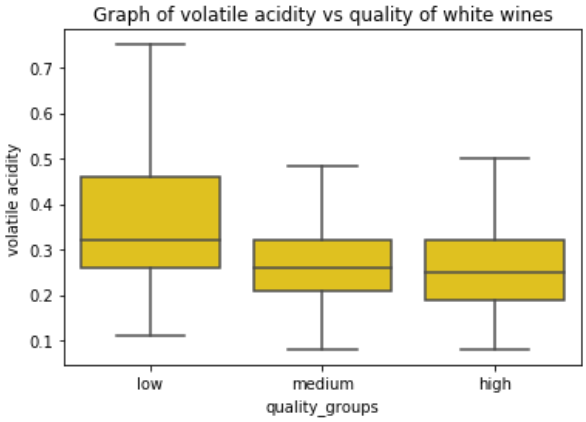


Background: Sweeter wines generally have higher densities. High quality red wines seem to have slightly lesser densities compared to low and medium quality ones. Lesser density could be an indicator of more acidity/lesser pH.

Analysis: The density of white wines does not vary much with respect to quality groups, with high quality white wines having only slightly lower density, as is the case with red wines. There are 3 outliers in medium quality white wines who have significantly higher densities.

2. Volatile Acidity VS Quality



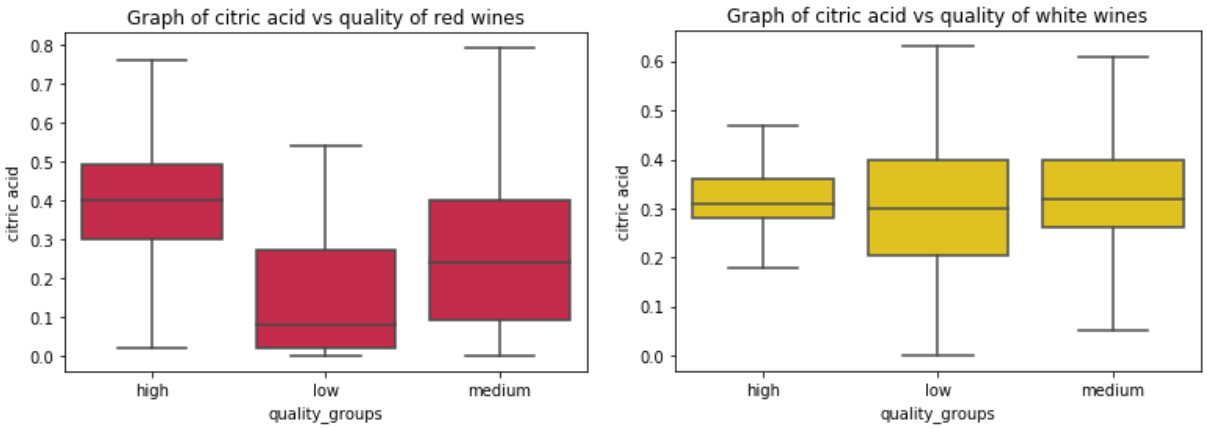


Background: Volatile acidity could be an indicator of spoilage, or errors in the manufacturing processes — caused by things like damaged grapes or wine exposed to air. This causes acetic acid bacteria to enter and thrive, and give rise to unpleasant tastes and smells. It is reasonable to see that volatile acidity content is lesser in high quality red wines. It is most in low quality red wines. There is one outlier in low and high quality groups.

Analysis: The change in volatile acidity with quality groups is not as distinct as it is with red wines. For white wines, the volatile acidity is only slightly higher for low quality compared to medium quality wines. Even with high quality white wines the volatile acidity is not significantly lower but only slightly lower with fewer outliers. It is said that wine experts can often tell the volatile acidity

just by smelling it. However, it seems that in the case of white wines, it might be a bit difficult to distinguish because of these smaller changes. All groups were found to have statistically significantly different means ($p < .005$).

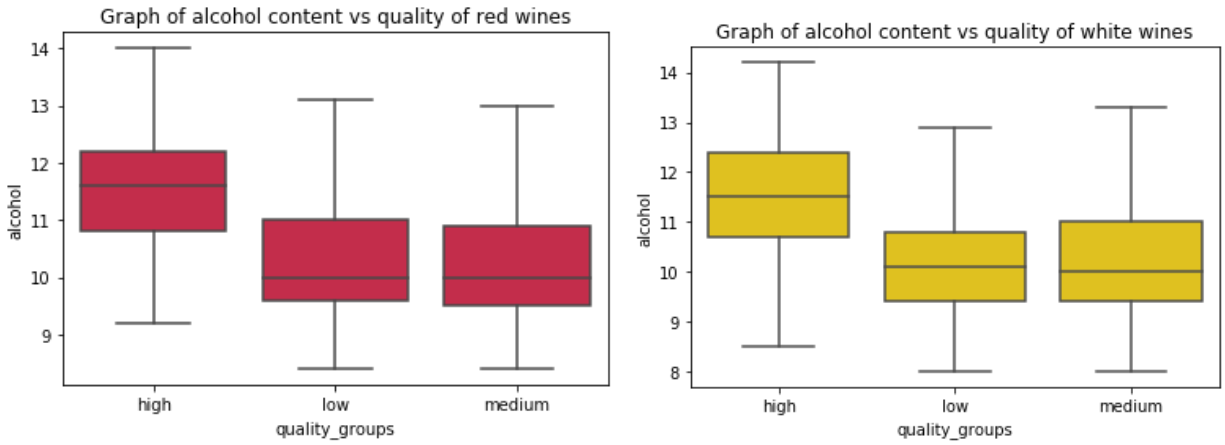
3. Citric Acid VS Quality



Background: Citric acid is generally found in very small quantities in wine grapes. It acts as a preservative and is added to wines to increase acidity, complement a specific flavor or prevent ferric hazes. Citric acid content seems to be slightly higher in high quality red wines. It can be added to finished wines to increase acidity and give a “fresh” flavor.

Analysis: Citric acid content does not vary a lot according to quality groups for white wines. For red wine, however, citric acid seems to have a strong positive relationship with wine quality. There are however, a few outliers in each quality group, which could be investigated to understand better what causes them to have such deviation from the group.

4. Alcohol VS Quality



Background: Alcohol produces a desirable psychedelic effect that many find enjoyable in their wine. Too much can have undesirable side effects.

Analysis: High quality red wines (8 and above) have a higher alcohol content. Low and medium quality wines might have roughly the same range of alcohol content. There seem to be a few outliers in medium quality red wines, where despite high alcohol content, their quality rating is below 8. Alcohol content is slightly more in high quality white wines.

CONCLUSION

After this extensive EDA, it seems like density, citric acid, volatile acidity and alcohol might be important in helping us predict the quality of wines! We now move on to the statistical analysis section of the project in order to check if the trends we saw in EDA are significant/ statistically meaningful.