EDA

Timur Guler

8/10/2021

Report:

EXEC SUMMARY [TBD]

Our goals: Our team took on the business challenge of the boutique wine store purchasing agent who wishes to identify high-quality red and white wines based on their physical characteristics. As a selective buyer, the agent wishes to only select the upper echelon of wines, and to minimize the number of low quality wines accidentally purchased. In addition to selecting high quality wines, the purchasing agent also wishes to learn which factors determine wine quality, and if and how these differ between white and red wines.

Our team used two data sets to answer these questions. One contained physical and chemical characteristics of red wine (1599 observations), and the other contained physical and chemical characteristics of white wine (4898 observations). This wine type imbalance, while not large enough to prevent meaningful analysis, was critical to keep in mind throughout our process.

Each data set contained 13 variables - an index column, a discrete 1-10 measure of quality, and the following physical/chemical characteristics of the wine:

- fixed acidity
- volatile acidity
- citric acidity
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

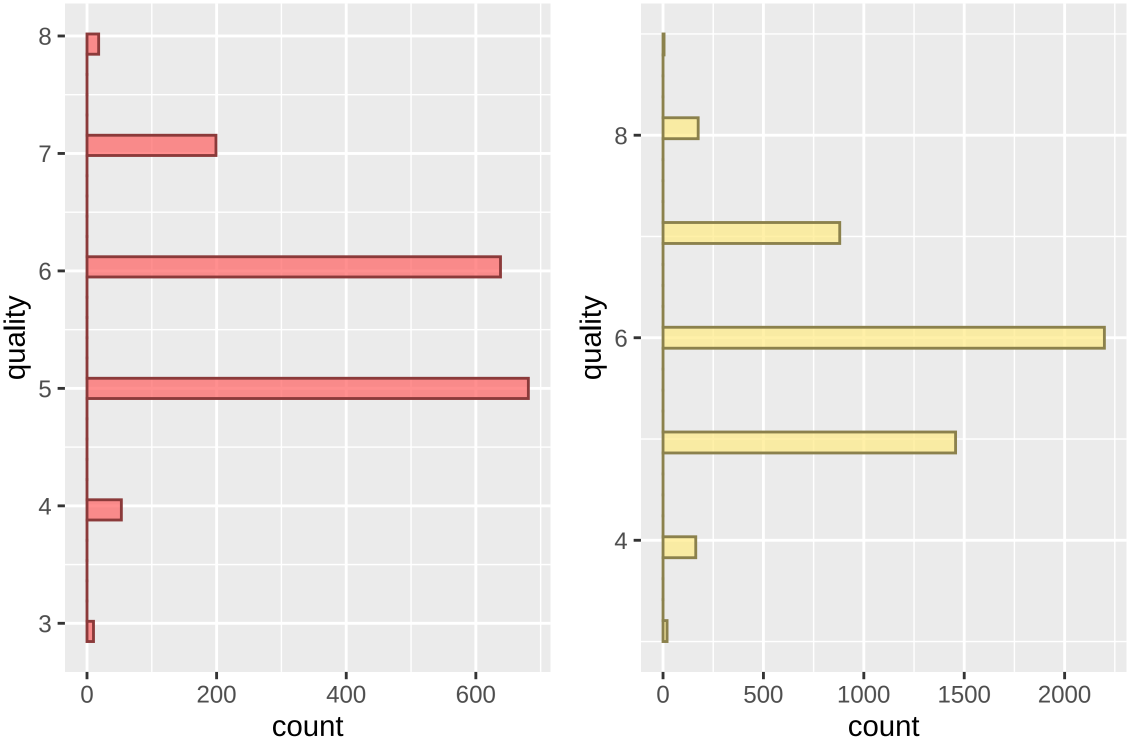
We removed the index column, as well as free sulfur dioxide, due to its high correlation with total sulfur dioxide in both red and white wine. Next, we performed a test-train split for both red and white wines with a fixed seed of 1, using 80% of each group for training data and the remaining 20% as testing data. Further discussions of "the data" during the EDA, model building, and model assumption validation stages will refer to the training data.

One fundamental issue in our analysis was the determination of a cutoff point for "high quality" wine. Based on the distribution for both red and white wines, we picked a cutoff quality rating of 7. This encompassed 13.5% of red wines and 21.4% of white wines. The same cutoff was used for both wine types for several reasons:

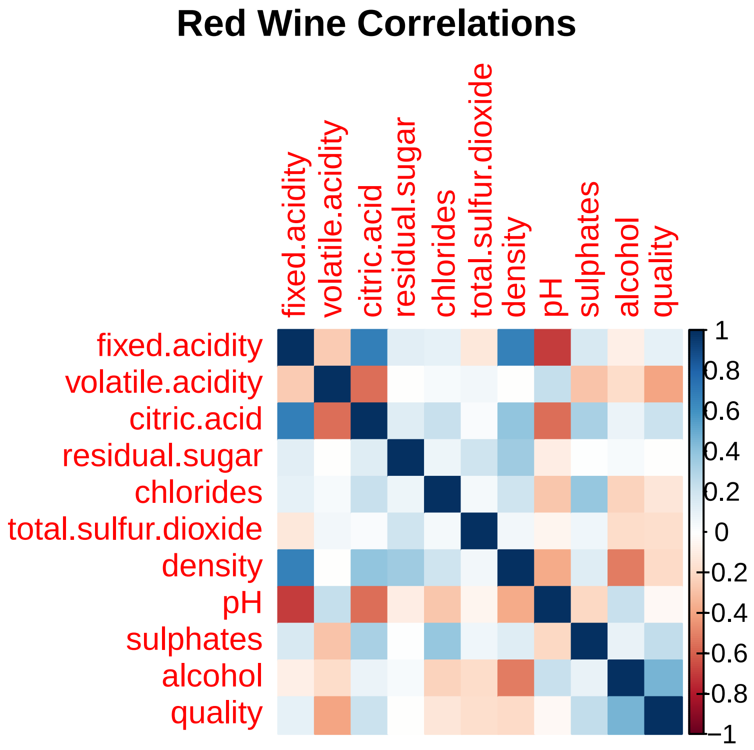
1. These proportions represent a level of distinction without making rarity a statistical concern
2. No determination had been made at this stage as to whether separate models would be used for white and red wine

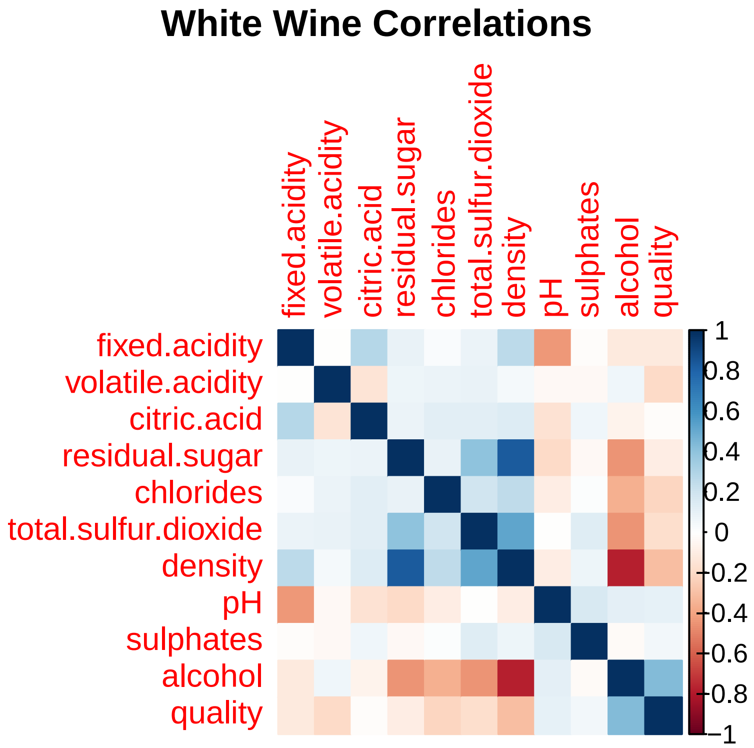
3. While there was a statistically significant difference in mean quality levels (p=2.2e-16), the actual difference was not practically meaningful (5.877909 for whites vs. 5.636023 for reds).

We also added the variable “color” to identify the wines, and created a combined training data set in addition to the separate red and white training sets.



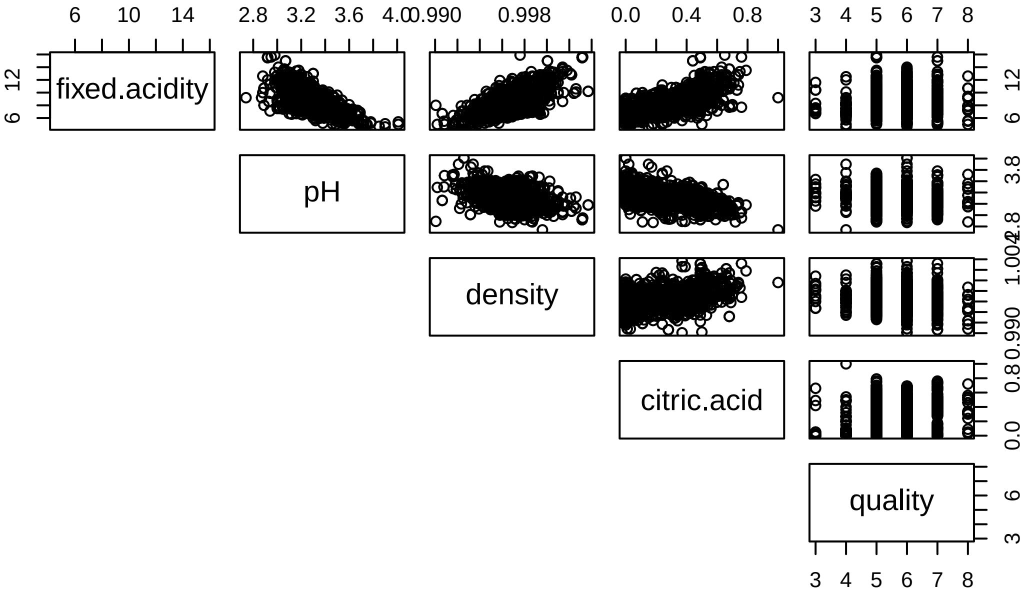
Our team had several goals during the EDA process. Our first goal was to understand the relationships between variables for both red and white wines. We accomplished this by examining correlations and scatterplots.


Red Wine Correlations

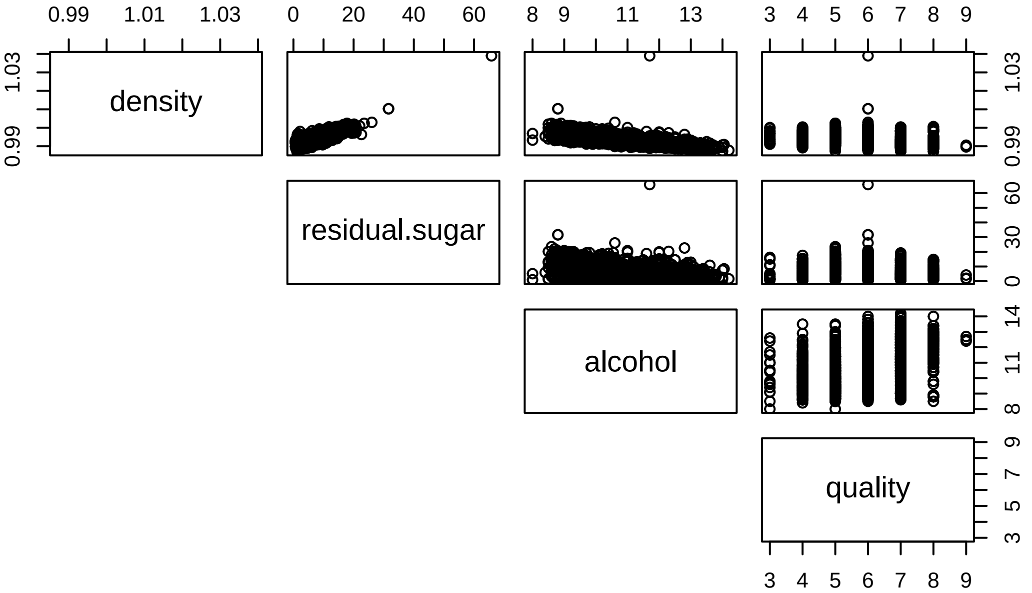

White Wine Correlations

This shows us that correlations between variables differ somewhat between white and red wines, and that each type of wine has some strong relationships between variables, which we will need to consider when looking for multicollinearity (e.g. fixed acidity with pH, density, and citric acid for red wines, density with residual sugar and alcohol in white wines). Looking at these relationships in more detail with scatterplots, we see that they are fairly tight, and somewhat linear.

Red Wine Scatterplots



White Wine Scatterplots

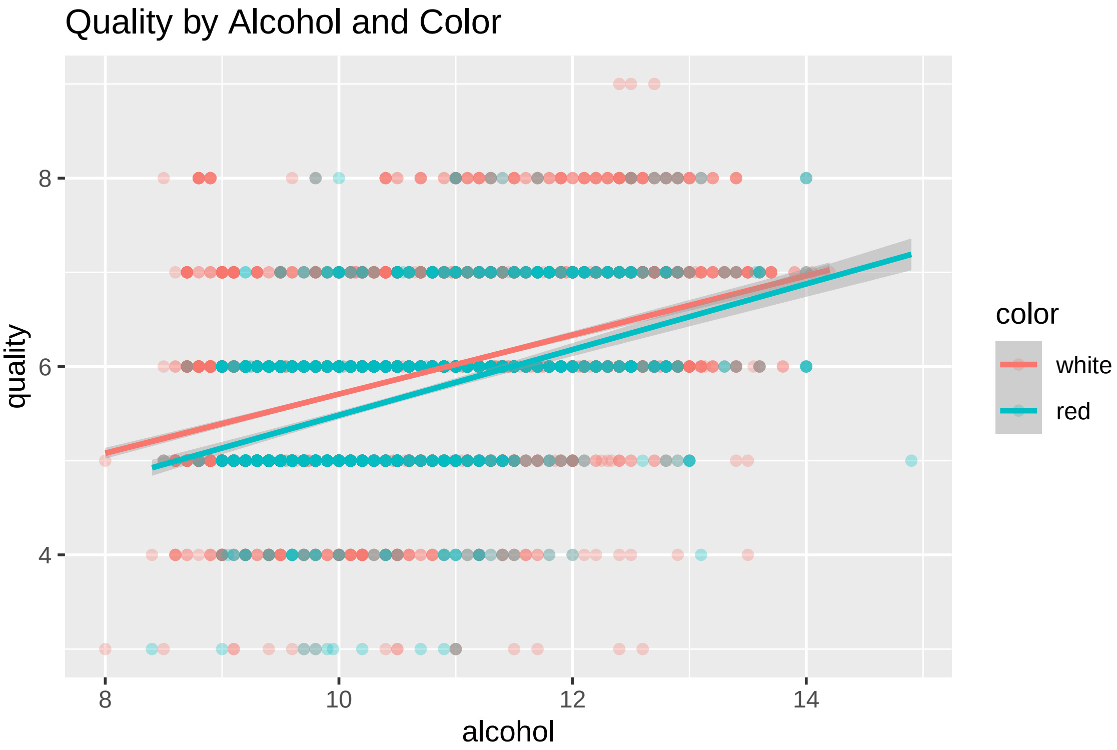


In addition to a general understanding of the variables' distributions and their relationships with each other, we were specifically interested in:

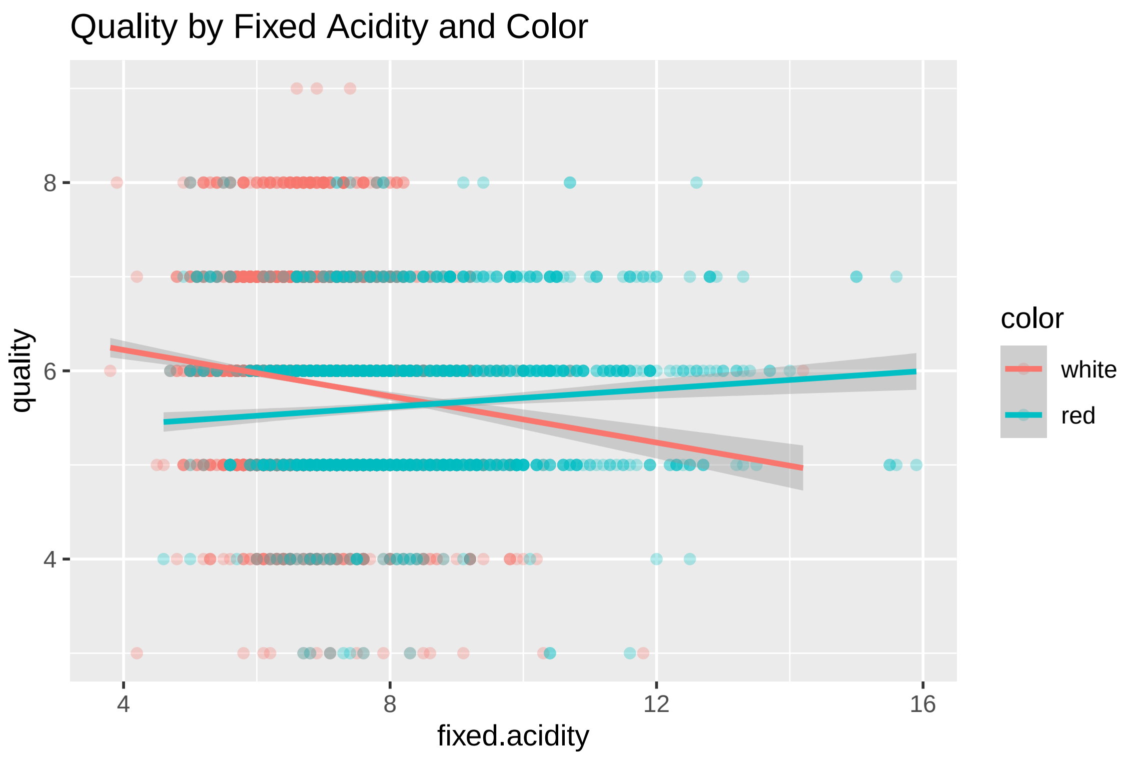
- the relationships between the physical/chemical characteristics of wine and quality
- whether it would be more effective to build a combined model or separate models for red and white

First, we looked at scatterplots of quality vs. physical characteristic by color using best-fit regression lines, as well as boxplots of quality category vs physical characteristic by color. This led to three key takeaways:

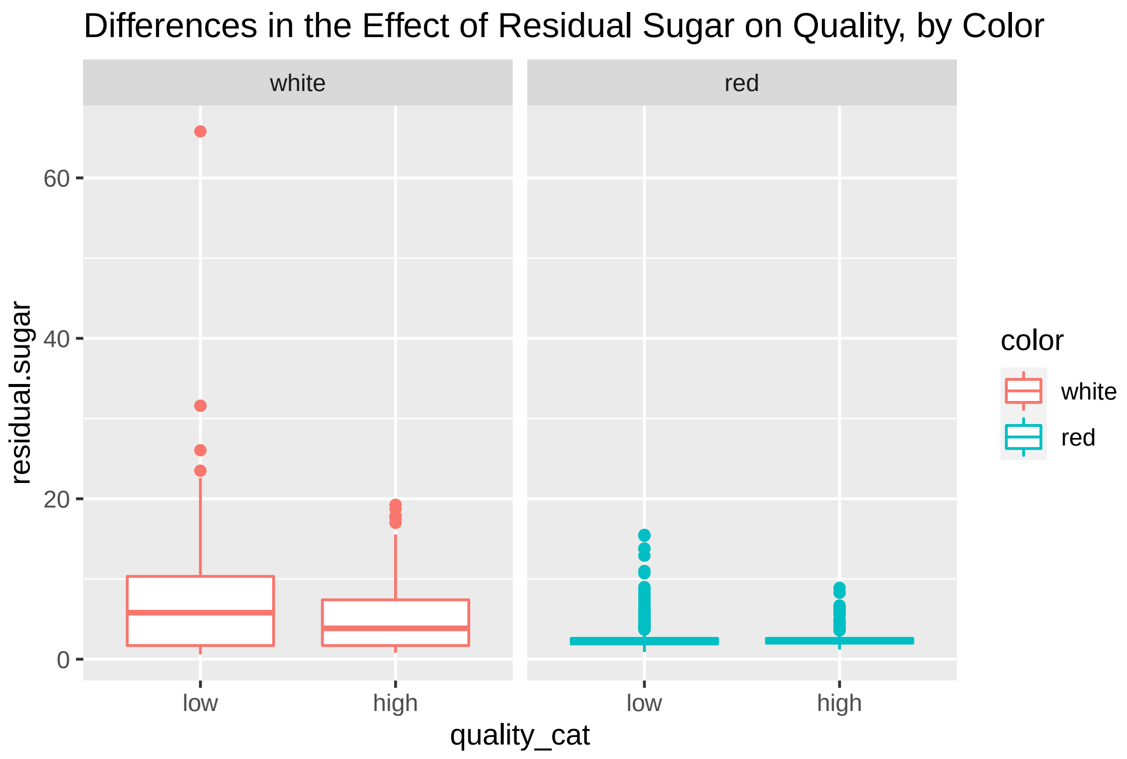
1. While several characteristics seem to have relationships with quality, a visual analysis show low “tightness of fit” for each potential predictor, meaning that it will likely take the combination of several weak predictors to yield a good prediction model



2. Some physical characteristics (notably fixed acidity, citric acidity, chlorides, pH, sulphates) have differing effects on quality in red vs white wines



3. The range for some physical characteristics (notably residual sugar) is markedly different between red and white wines, leading to potential extrapolation issues if a combined model is used.



Based on takeaways (2) and (3), as well as the difference in sizes between the red and white data sets, our team determined that it would be best to use two separate models for predicting high vs. low quality - one for red wines, and one for white wines.

red wine model:

- various measures point to 7 predictor model
- fit model and review assumptions (influential points, linearity assumption, multicollinearity)
- p value of 7 predictor vs intercept only - useful
- 7 more useful than full model

——————-test data—————— - look at roc, auc

- confusion matrix - show multiple thresholds, end up achieving desired business goals

white wine model:

repeat process

————discussion——————

relate back to research question