Wine Quality Prediction

Mel Sparks

5/6/2022

Predicting Wine Quality

Winemakers want to take their vintages to the next level by using science and data analytics to get that extra edge. While many of them do not require extensive redesigns to their time-honored traditions, most would still benefit from knowing what the data has to offer so that they can continue to make informed desitions. (Item 1)

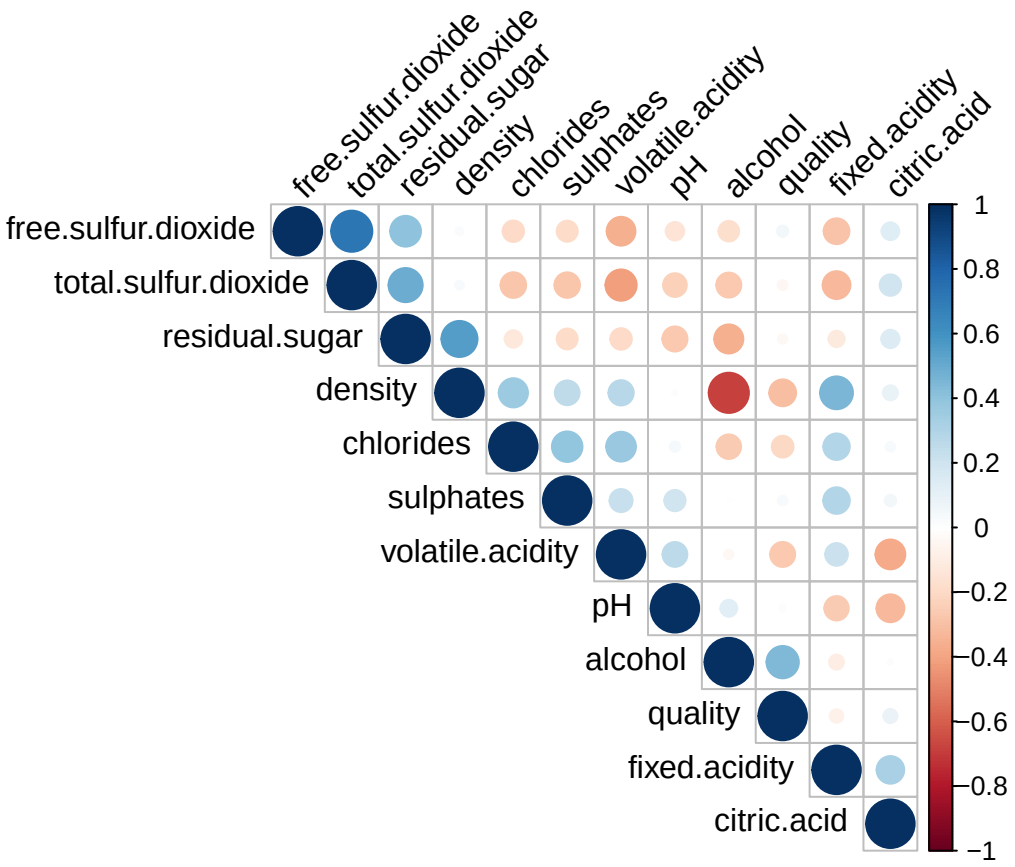
These datasets were retrieved from the University of California, Irving Machine Learning Repository. (Item 2) It contains information on wines from the year 2009 for Portugese "Vinho Verde" variety. There is information on 1599 red wines and 4898 white wines available from that year (for a total of 6497). The data on these vintages include 13 variables primarily regarding the chemical composition of the wines (density, pH, residual sugar, etc), and also includes an overall quality rating and a designator that I added to note if the wine was red or white. (Item 4)

Data Preparation

The datasets I am using have no missing values. This is likely due to the fact that each variable is a measurement that was recorded at the time it was taken. Additionally, as previously stated, I added a variable so that I could combine the two datasets and still keep track of which wines were red and white. However, when I imported the datasets, only the "quality" variable was received as "numeric". All of the variables except for the "type" variable I added should be of the numeric class. I didn't have any particular issue with the column names, so I didn't reformat any of them. I also didn't want to limit any of my data, so I deliberately made the choice to not do that. (Item 5)

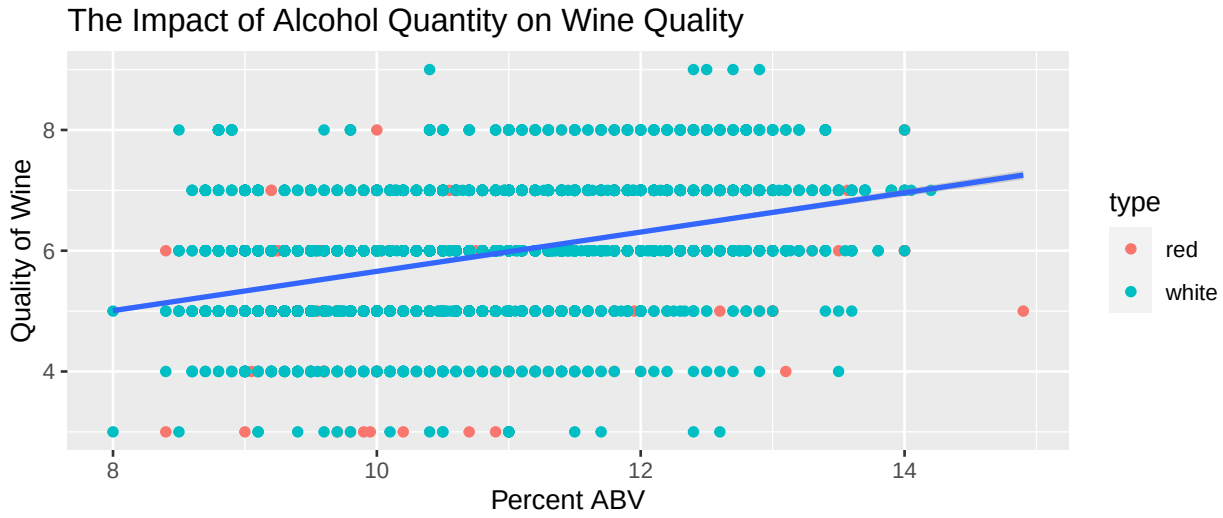
Modeling

One of the first things I wanted to look at were the potential correlations in the data. I'm primarily interested in correlations directly related to quality, but a general overview will help me spot relationships between the variables that could result in multicollinearity and endogeneity. (Item 6, 8)

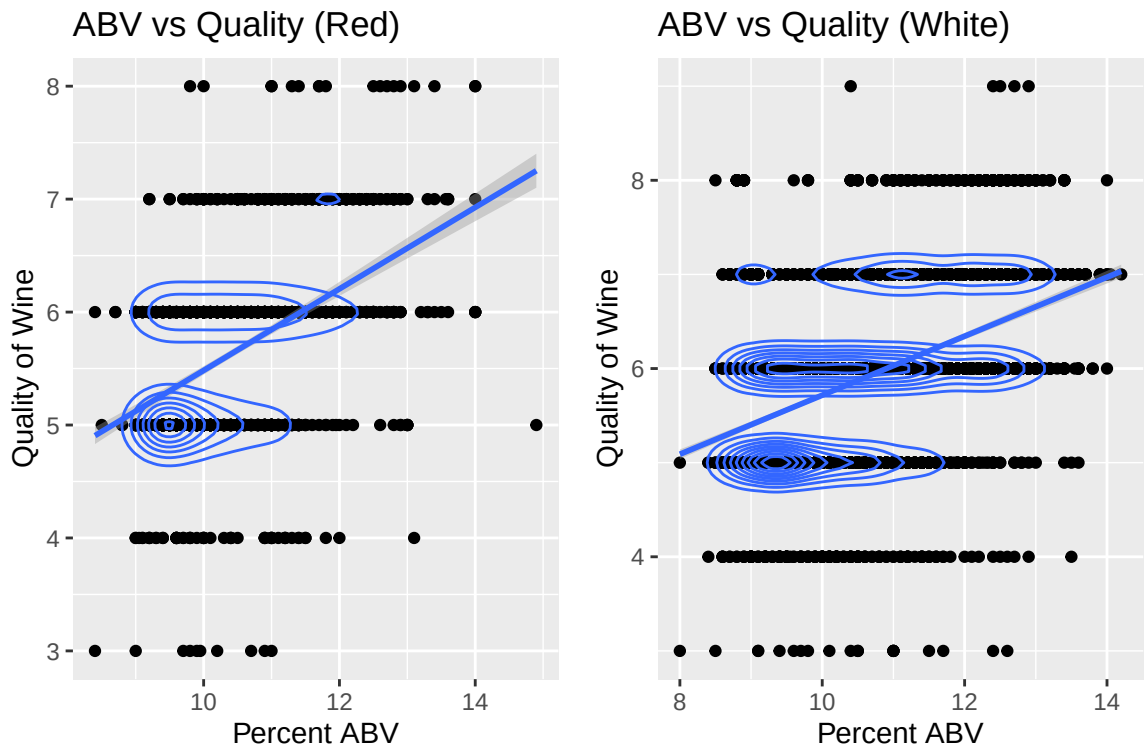


What we can see in this correlogram is that most of the provided qualities have little to no positive impact on the overall quality of the wine. At most, we can see that there are indications that an increase in density, quantity of chlorides, and volatile acidity have a slightly negative correlation to quality. Additionally, we can see that alcohol quantity (ABV) has a slight positive correlation with quality. (Item 7, 8)

However, when we plot quality and alcohol against each other, we can't discern a clear pattern in graph; I added the regression line in case points were falling on top of each other and getting lost in the visualization. (Item 7, 8)



I also noticed that the white wine dominates this scatter plot and I want to see what it looks like when the wine types are put on their own graphs.



What we can ascertain from these two graphs is that there are a considerable number of wines in the lower ABV range that only achieve a quality rating between 5-6. Wines that came in at a 7 or 8 still have a variety of ABVs available, but do seem to lean towards the higher alcohol content when you take the high density of lower quality wines into account. (Item 6, 8)

Conclusions

We can see from the very little work that has been done so far that there's a reason why wine-making is generally considered an art. Firstly, none of the recorded variables seem to have a significant impact that tells us, "If you increase *this*, you'll also increase quality." That being said, we did look at the one variable that may have some positive correlation, which was alcohol content (ABV). The higher levels of quality ratings were accompanied by generally higher alcohol contents. That being said, our data shows plenty of data points where the wine was exceptionally strong, but rated quite low.

There is still room for a little more work, however, when looking at how lower values of density, chlorides, and volatile acidity could also increase overall quality rating. It may still be possible to work with these four variables in concert to help see improvements to quality ratings.