Does alcohol percentage have any real bearing on the quality of wine?

Dataset Description:

There are two multivariate datasets, representing red and white wine variants of the Portuguese region "Vinho Verde." The categories in both files are identical. The categories are Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates, Alcohol (%), and Quality. There are 1599 instances for red wine and 4898 for white wine. There are no missing values in either dataset. Additionally, looking at the number of instances above, there are roughly 3x as many white wine entries as red wine and at no point do we see a red wine above 8 for quality.

Fixed Acidity: According to Nierman (2004), "The predominant fixed acids found in wines are tartaric, malic, citric, and succinic. Their respective levels found in wine can vary greatly but in general one would expect to see 1,000 to 4,000 mg/L tartaric acid, 0 to 8,000 mg/L malic acid, 0 to 500 mg/L citric acid, and 500 to 2,000 mg/L succinic acid."

Volatile Acidity: The definition according to Neeley (2004), "Volatile acidity refers to the steam distillable acids present in wine, primarily acetic acid but also lactic, formic, butyric, and propionic acids."

Citric Acid: "A tricarboxylic acid C6H8O7 occurring in cellular metabolism, obtained especially from lemon and lime juices or by fermentation of sugars, and used chiefly as a flavoring (Merriam-Webster.com dictionary, n.d.)."

Residual Sugar: "Residual sugar (or RS) refers to the sugars left unfermented in a finished wine. It is measured by grams of sugar per liter (g/l) (Wu, 2020, para 1)." This will affect the wines overall sweetness.

Chlorides: "A compound of chlorine with another element or group *especially*: a salt or ester of hydrochloric acid (Merriam-Webster.com dictionary, n.d.)." With regards to wine, this is talking about the amount of salt in the wine.

Free Sulfur Dioxide: "The FSO2 and the pH of your wine determine how much SO2 is available in the active, molecular form to help protect the wine from oxidation and spoilage (Moroney, 2018, para 1)."

Total Sulfur Dioxide: "Total Sulfur Dioxide (TSO2) is the portion of SO2 that is free in the wine plus the portion that is bound to other chemicals in the wine such as aldehydes, pigments, or sugars (Moroney, 2018, para 3)."

Density: "Density is defined as the mass, or weight, per volume of a material. In the case of liquids, density is often measured in units of g/mL, which is the weight in grams (g) of each milliliter (mL) of liquid. The density of wine is primarily determined by the concentration of alcohol, sugar, glycerol, and other dissolved solids (Density, n.d.)."

pH: "PH is the measure of the degree of relative acidity versus the relative alkalinity of any liquid, on a scale of 0 to 14, with 7 being neutral. Winemakers use pH as a way to measure ripeness in relation to acidity (Vinifera, 2009, para 3)."

Sulphates: "A salt or ester of sulfuric acid (Merriam-Webster.com dictionary, n.d.)." This is an additive that also acts as an antioxidant/antimicrobial. This will add to the overall SO_2 level.

Alcohol (%): The percentage of alcohol content in the wine. Also known as ABV (Alcohol by Volume), defined as, "a measure of the concentration of alcohol in an alcoholic beverage (Merriam-Webster.com dictionary, n.d.)."

Quality: A score between 0 (low quality) and 10 (high quality).

Why this dataset:

Every year my wife and I draw an envelope from our anniversary jar that contains an adventure we'll randomly accomplish over the course of the year. Our pull in August 2020 was with regards to wine, specifically, "Become a Wine Sommelier/Or Just Learn to Like Wine." I selected this dataset because it relates to something I am currently doing in my life and hope that it can teach me something in the process of exploring the data.

Data Processing:

The first thing that had to be addressed was utilizing "Text to Column" in Excel to separate the data by the ";" delimiter in each of the .csv files. After that I started scanning the data looking for things to fix. Both datasets had duplicate rows that I removed. I found it very unlikely that there would be two unique wines with the exact same properties across the board. Red wine went from 1599 instances to 1354. White wine went from 4898 instances to 3956. I capitalized the first letter of each word in the category titles to make it look a little better.

When snapping the column widths to fit the data, I noticed that the alcohol column was wider than I expected initially. When investigating I found that there were some longer values that didn't conform to societal norm. When you see a bottle of wine in the store, you would typically see the percentage out to the tenths place. In this original data there were a few out four places past that. You won't see a bottle in real life that says "10.03333%", so I formatted all of the values in that column to one place after the decimal.

When looking at the smallest and largest values for each respective category of the two wines, I observed that there were a few outliers. I removed two outliers from each data set. The outliers in the white wine data set were the rows that had a Residual Sugar value of 65.8, more than twice the next value, and the row that had Free Sulfides at 289, with the next highest value being at 146.5. The outliers in the red wine data set were the rows that had Total Sulfides at 289 and 278 respectively, more than 100 higher than the next value.

My last step is broken into two parts. First, I added a column to the beginning of each dataset labeled "Variety." For each respective type of wine, I typed either White or Red. I then combined the two files into one. There wasn't a need to keep them separate, and I left the opportunity to still separate the data if necessary.

The Question:

Does alcohol percentage have any real bearing on the quality of wine? This question is meant to evaluate whether or not the alcohol percentage affects the overall quality of the wine using results generated from the 1354 red wines and the 3956 white wines. These datasets both have the points needed to address this question since quality and alcohol percentage are two of the compiled categories.

Exploratory Analysis Preamble:

There are a few things to define before I delve into the analysis to maximize understanding.

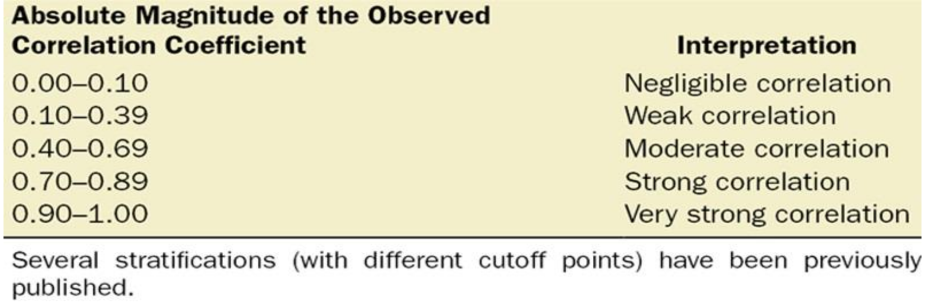
Correlation: "A relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone (Merriam-Webster.com dictionary, n.d.)."

Wine Alcohol Content: Written by Monico (2020), "Although a standard serving of wine is 5 ounces and generally contains between 11-13% alcohol by volume, not all wines are created equal. The same goes for the amount poured, whether it's at a restaurant or at home with friends. White wine generally is on average 10% ABV; however, it can range from as little as 5% to as much as 14%.

Moscato white wines have less alcohol, at 5-7%, while pinot grigio wines may contain 12-13% alcohol, and chardonnay may have 13-14.5%. Red wine has more alcohol, ranging from 12% to 15%. Pinot noir and Boudreaux contain 13-14% ABV, Malbec wines contain 13.5-15%, and some Californian zinfandels and Australian shiraz wines can have ABVs as high as 16-18% (paras 8-9)."

After combing through the data and using this definition, I've defined 11-13% as average for this dataset.

For interpreting the correlation coefficients you'll see, I utilized the following chart:



Absolute Magnitude of the Observed Correlation Coefficient	Interpretation
0.00–0.10	Negligible correlation
0.10–0.39	Weak correlation
0.40–0.69	Moderate correlation
0.70–0.89	Strong correlation
0.90–1.00	Very strong correlation

Several stratifications (with different cutoff points) have been previously published.

The formula for the Pearson Correlation Coefficient (what I utilized with Python) is:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

$\mathrm{r} = \text{correlation coefficient}$

$\mathrm{x_i} = \text{values of the x} - \text{variable in a sample}$

$\bar{\mathrm{x}} = \text{mean of the values of the x} - \text{variable}$

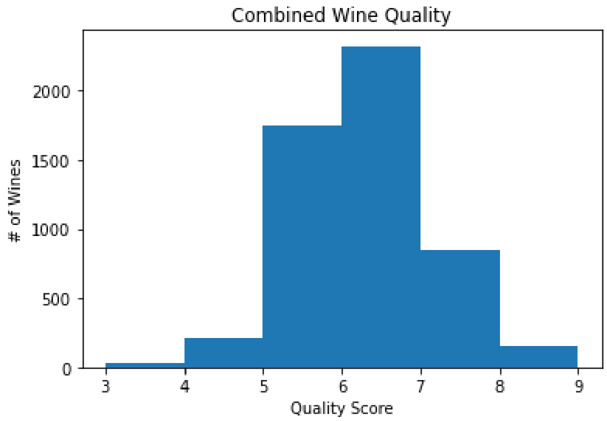
$\mathrm{y_i} = \text{values of the y} - \text{variable in a sample}$

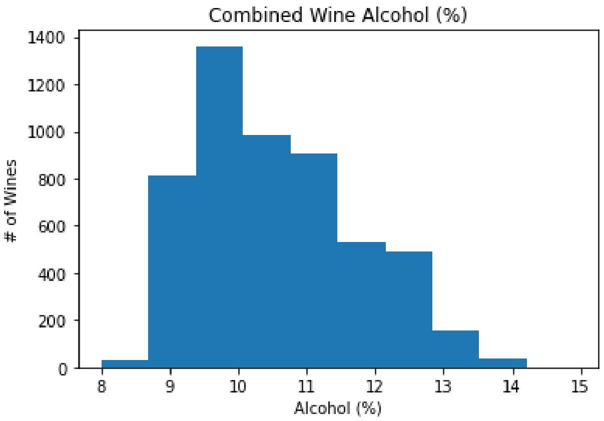
$\bar{\mathrm{y}} = \text{mean of the values of the y} - \text{variable}$

Going further, I define Low-Quality as having a score equal to or less than 5 and I define High-Quality as being greater than 5. Now that we've defined what we are using, we can move on to the actual analysis.

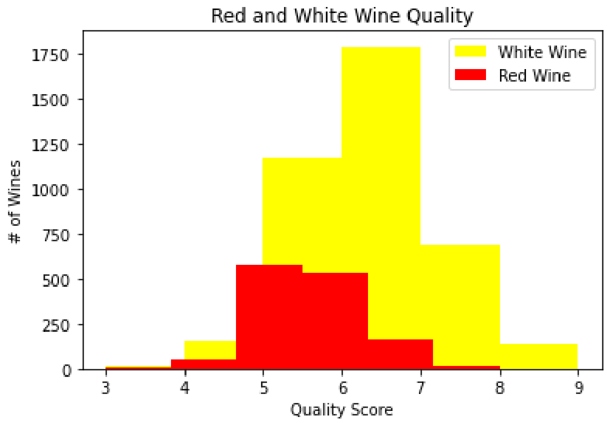
Exploratory Analysis:

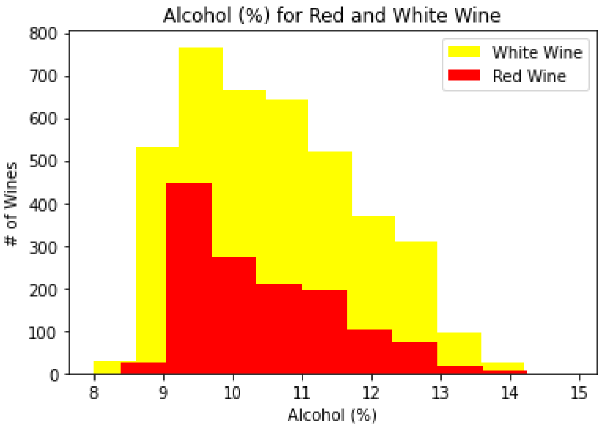
It is important to know what we are looking at and what our data is comprised of. Below are a few histograms. The first two we'll see represent the combined dataset. As you can see in the histogram on the left, the bulk of the wine falls between 5 and 8 for a quality score. There are no values below 3 or above 9. On the right you'll see the alcohol percentage that appears skewed right, but it should be noted that there aren't any wines below 8%.



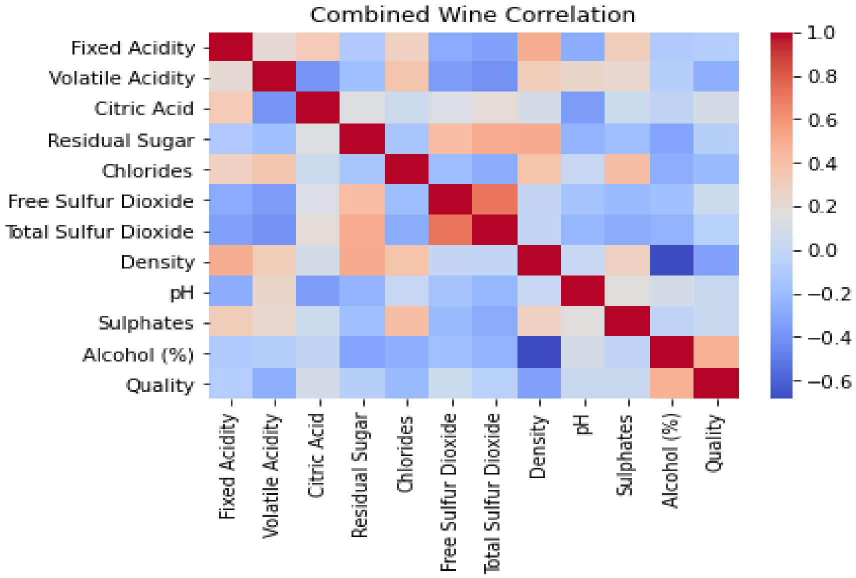


These next two histograms are a little different. The one on the left represents Red and White Wine Quality. Here you can see the disparity with regards to how much more white wine data we have over red. You'll also notice that the bulk of the red wine quality scores are in the 5-6 range while the white wine scores are in the 6-7 range. One the right is a representation of alcohol percentage for red and white wine. Again, you'll see that there are clearly more white wine data points. Barring that though, you will notice that the shape of the histogram is very similar for white and red wine. The bulk of the wine is between 9 and 11.5%. It completely stops at 8% and trails off starting around 11-12%.

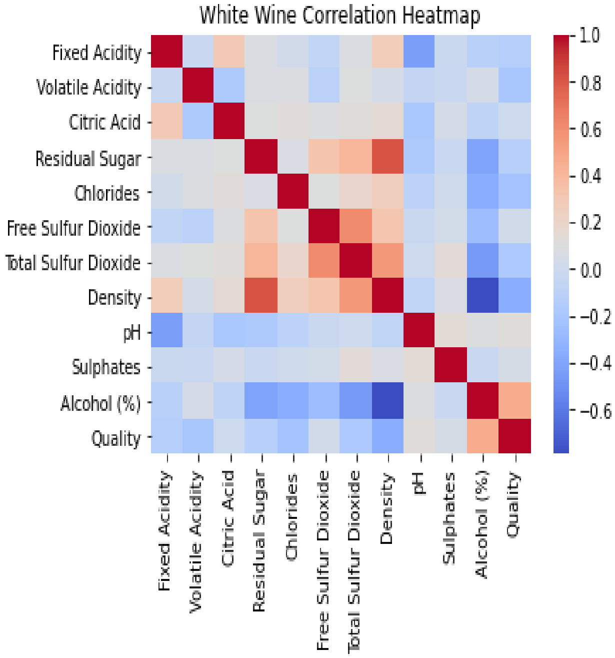


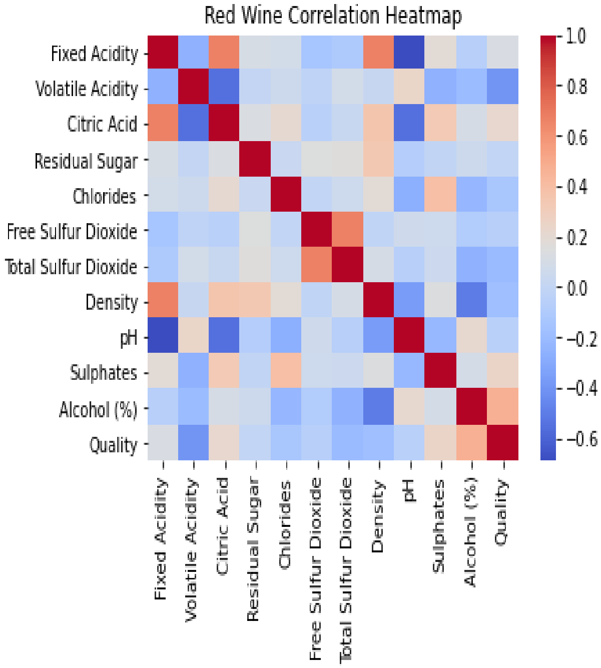


Directly below you can see that alcohol percentage and quality have a positive correlation. Specifically, that correlation is 0.469144. We can break this down further though. If you recall the chart from earlier, this would be considered a moderate positive correlation.

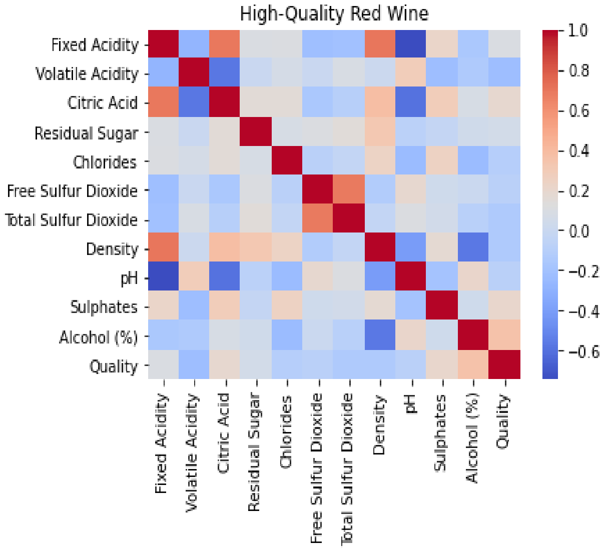


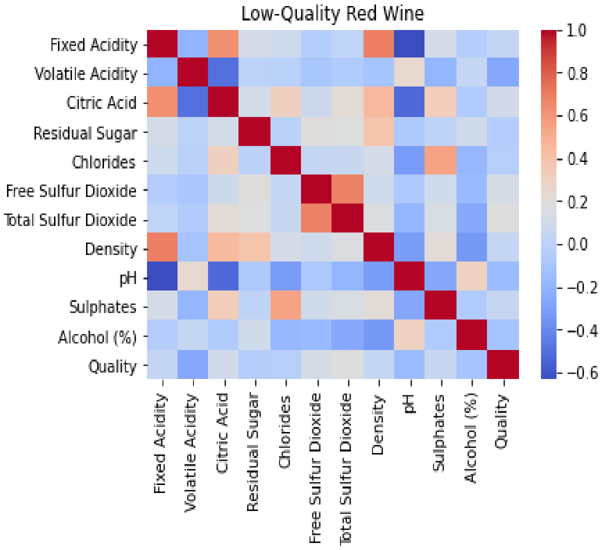
When looking at a Heatmap for just white wine, the correlation between alcohol percentage and quality is sitting at 0.463379, not far from the combined dataset. Bottom right, you'll notice this Heatmap is specific to red wine. Looking at the correlation between alcohol percentage and quality for just red wine brings a slightly higher value, sitting at 0.476720. Both are considered moderate positive correlations.



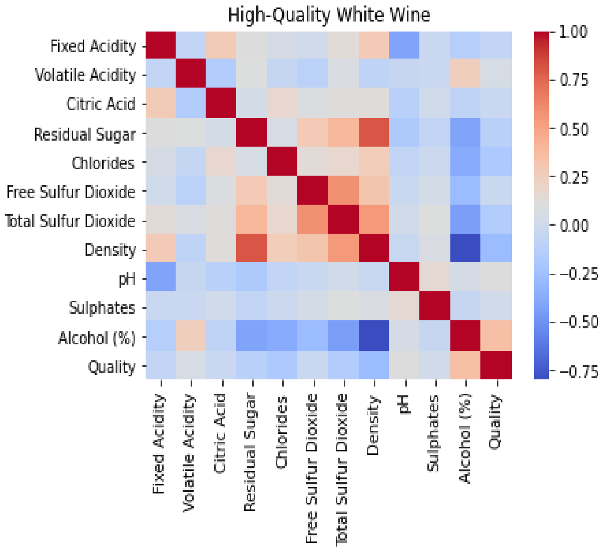


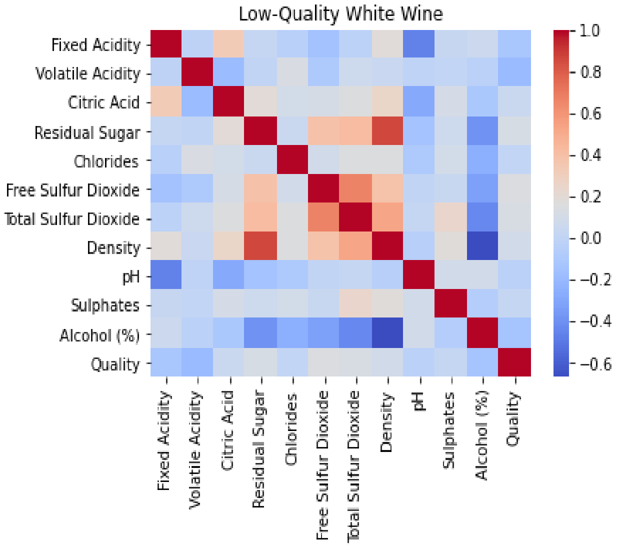
We can break it down even further than that though by breaking up Red and White wine into High- and Low-Quality data frames. When looking at High-Quality Red Wine, the correlation between alcohol percentage and quality is 0.569489 (moderate positive correlation). For Low-Quality Red Wine, it is -0.318789 (weak negative correlation).





For High-Quality White Wine, the correlation between alcohol percentage and quality is 0.699420 (moderate-strong positive correlation). The Low-Quality White Wine correlation is -0.266476 (weak negative correlation).

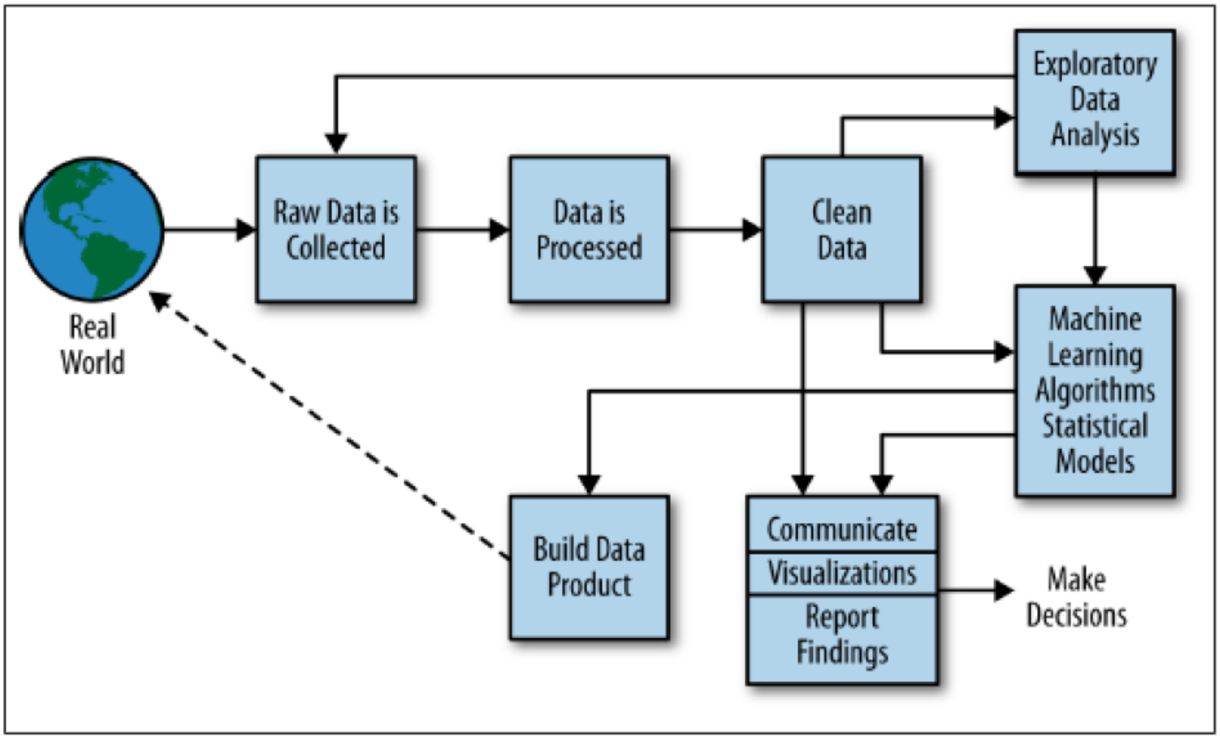




Some preliminary conclusions I've drawn are that there is a moderate positive correlation between alcohol percentage and overall quality. When looking at just High-Quality Red Wine, we see that there is still a moderate positive correlation, but it is higher than the combined data set. The same can be said for High-Quality White Wine, except that it is even higher than the Red Wine and is borderline a strong correlation. Both Low-Quality Red and White Wines have weak negative correlations. This leads me to believe that for a wine to be of High-Quality, it should have a higher alcohol content. Too high though and it drops again. There is a goldilocks zone that needs to be respected.

Methodology/Rationale:

I followed aspects of ‘The Process’ introduced in Introduction to Data Analytics based off of the Data Science Process from the Doing Data Science book written by Cathy O’Neil and Rachel Schutt.



Gathering the raw data was easy. It was provided, all I had to do was choose a dataset that seemed interesting. I then went through the steps of processing and cleaning the data. Once I had a relatively clean set, I started to explore around with what I had. It was then that I noticed a few outliers and went back to my data to clean it up a little more. Once I was confident in my data being clean and relatively balanced, I started creating visualizations to answer my overall question.

Since correlation is what I’m trying to show, I knew that heatmaps would be a useful visual representation. I also made the decision to break the data up into a few sub data frames. Those are High-Quality Red Wine, Low-Quality Red Wine, High-Quality White Wine, and Low-Quality White Wine. I also constructed heatmaps utilizing the combined data without breaking it down into smaller chunks. Histograms were also useful for seeing how this data was laid out.

Pros of going through and using ‘The Process’ were that it gave some structure to this whole process when I was feeling lost. I just went back and broke it into steps and repeated as necessary. The focus I got from a clear structure while walking through not always clear data was helpful. Some cons with this approach were that not everything necessarily applied to me, so I was jumping around from time to time and more of using it as a guideline. Instead of cleaning up the. csv’s by hand in excel, I could have utilized tools in python to make it go faster. However still being relatively new to this, I felt more comfortable doing it in excel by hand. It also gave me more practice in utilizing tools I had never touched before this class.

Final Conclusions:

There are a few things this data showed me. But the most important thing gleaned is the answer to the question, “Does alcohol percentage have any real bearing on the quality of wine?” The answer could be straight-forward, but it isn’t. Obviously, yes, alcohol percentage does play a roll. Too little is bad, too much is bad. There is a science involved in making good wine. But when we look at this data set, of wines that are meeting the average I talked about earlier in the Exploratory Analysis Preamble, alcohol percentage doesn’t seem to be a strong way of measuring quality in wine. So, with this data that I have, I’m going to say no, I don’t believe that alcohol percentage (as long as it is average at a minimum) has any real bearing on the quality of wine. If the data was more even, and truly representative of all wine, that answer could change. There isn’t really a way to know without event more data points to sift through.

Additional Analyses:

There are a few other analyses that would have been interesting to carry out. I would have liked to have seen how price and quality correlated. And if alcohol content had any bearing on the price as well. For that, the only additional data I would have needed is price in a standard unit.

Another analysis would have involved more varieties, specifically adding Ports and Rosé. This would have required two more datasets covering both of these, just like what was utilized with Red and White wine respectively.

Since we know this wine is from the Portuguese region “Vinho Verde”, comparing this to other regions could show variations in all of these categories that may be region-specific.

Specific to port wines, I’d like to explore density and alcohol percentage. I know that the more alcohol the wine has, the less dense it is, but every port I’ve tasted felt thick, like I was consuming a syrup.

For any additional data, I would want it to be as even as possible. This data was interesting to wrangle because one data set was essentially 3x the size of the other one.

References:

1. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
   Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
2. Nierman, D. (2004). *Fixed Acidity*. UC Davis. Retrieved December 6, 2020, from https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity
3. Neeley, E. (2004). *Volatile Acidity*. UC Davis. Retrieved December 6, 2020, from https://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity
4. Merriam-Webster. (n.d.). Citric Acid. In *Merriam-Webster.com dictionary*. Retrieved December 3, 2020, from https://www.merriam-webster.com/dictionary/citric%20acid
5. Merriam-Webster. (n.d.). Sulfate. In *Merriam-Webster.com dictionary*. Retrieved December 3, 2020, from https://www.merriam-webster.com/dictionary/sulfate
6. Wu, S. (2020, July). *What is Residual Sugar in Wine?* Decanter. Retrieved December 7, 2020 from https://www.decanter.com/learn/residual-sugar-46007/
7. Merriam-Webster. (n.d.). Chloride. In *Merriam-Webster.com dictionary*. Retrieved December 3, 2020, from https://www.merriam-webster.com/dictionary/chloride
8. Moroney, M. (2018, Feb 27). *Total Sulfur Dioxide – Why it Matters, Too!* Iowa State University. https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too
9. https://www.etslabs.com/analyses/DEN
10. Vinifera (2009, Apr 15). *What do "pH" and "TA" numbers mean to a wine?* Wine Spectator. Retrieved December 8, 2020, from https://www.winespectator.com/articles/what-do-ph-and-ta-numbers-mean-to-a-wine-5035
11. Merriam-Webster. (n.d.). Correlation. In *Merriam-Webster.com dictionary*. Retrieved December 3, 2020, from https://www.merriam-webster.com/dictionary/correlation
12. Merriam-Webster. (n.d.). Alcohol By Volume. In *Merriam-Webster.com dictionary*. Retrieved December 3, 2020, from https://www.merriam-webster.com/dictionary/alcohol%20by%20volume
13. *Density*. (n.d.). ETS Labs. Retrieved December 7, 2020, from https://www.etslabs.com/analyses/DEN
14. Schober, P., Boer, C., Schwarte, L. (2018, May). *Correlation Coefficients: Appropriate Use and Interpretation.* Retrieved December 7, 2020, from https://journals.lww.com/anesthesia-analgesia/fulltext/2018/05000/correlation_coefficients__appropriate_use_and.50.aspx
15. Schutt, R., & O'Neil, C. (2013, Oct). *Doing Data Science.* O'Reilly Media, Inc.
16. Monico, N. (2020, Jul 24). *Alcohol by Volume: Beer, Wine, & Liquor.* American Addiction Centers. Retrieved December 13, 2020, from https://www.alcohol.org/statistics-information/abv/