WILL GOSNOLD

Wine Quality Dataset

Exploratory Data Analysis

Introduction

In this EDA, I am going to be looking at a red wine quality dataset from Cortez et al., 2009. It contains physicochemical data about variants of the Portuguese "Vinho Verde" wine. It does not include data about grape varieties or price, but does have reviews of the wines' quality, so this will be an analysis of how the chemical contents of each wine relates to its quality.

Data Summary

The dataset is made up of 1599 instances for various wine varieties with 12 columns describing the following features:

	Feature	Feature Range
1.	fixed acidity	`4.6 - 15.9, continuous`
2.	volatile acidity	`0.12 - 1.58, continuous`
3.	citric acid	`0 - 1, continuous`
4.	residual sugar	`0.9 - 15.5, continuous`
5.	chlorides	`0.012 - 0.611, continuous`
6.	free sulfur dioxide	`1 - 72, continuous`
7.	total sulfur dioxide	`6 - 289, continuous`
8.	density	`0.99007 - 1.00369, continuous`
9.	pH	`2.74 - 4.71, continuous`
10.	sulphates	`0.33 - 2, continuous`
11.	alcohol	`8.4 - 14.9, continuous`
12.	quality	`3, 4, 5, 6, 7, 8 (target)`

While there are only values between 3 and 8 for quality in the dataset, it is a mark out of 10 so theoretically could be higher or lower. In collecting this data, three wine experts performed separate blind taste tests on the wines and scored them individually out of 10. The median score given by these experts is the one given in the dataset.

The physicochemical properties of the wines were measured and recorded in the laboratory and were chosen based on their ease of testing at the end of the production process.

Data Cleaning

To begin, let's have an initial look through the dataset and see what might need to be cleaned up. By checking the data types of the columns, we can see that the data is already represented numerically and thus, we don't need to perform any manipulations in that regard.

All of the features in the dataset are continuous and numerical and the target variable is discrete. We can have a look at a correlation heatmap to see which features are the most highly correlated with our target variable: 'quality'.

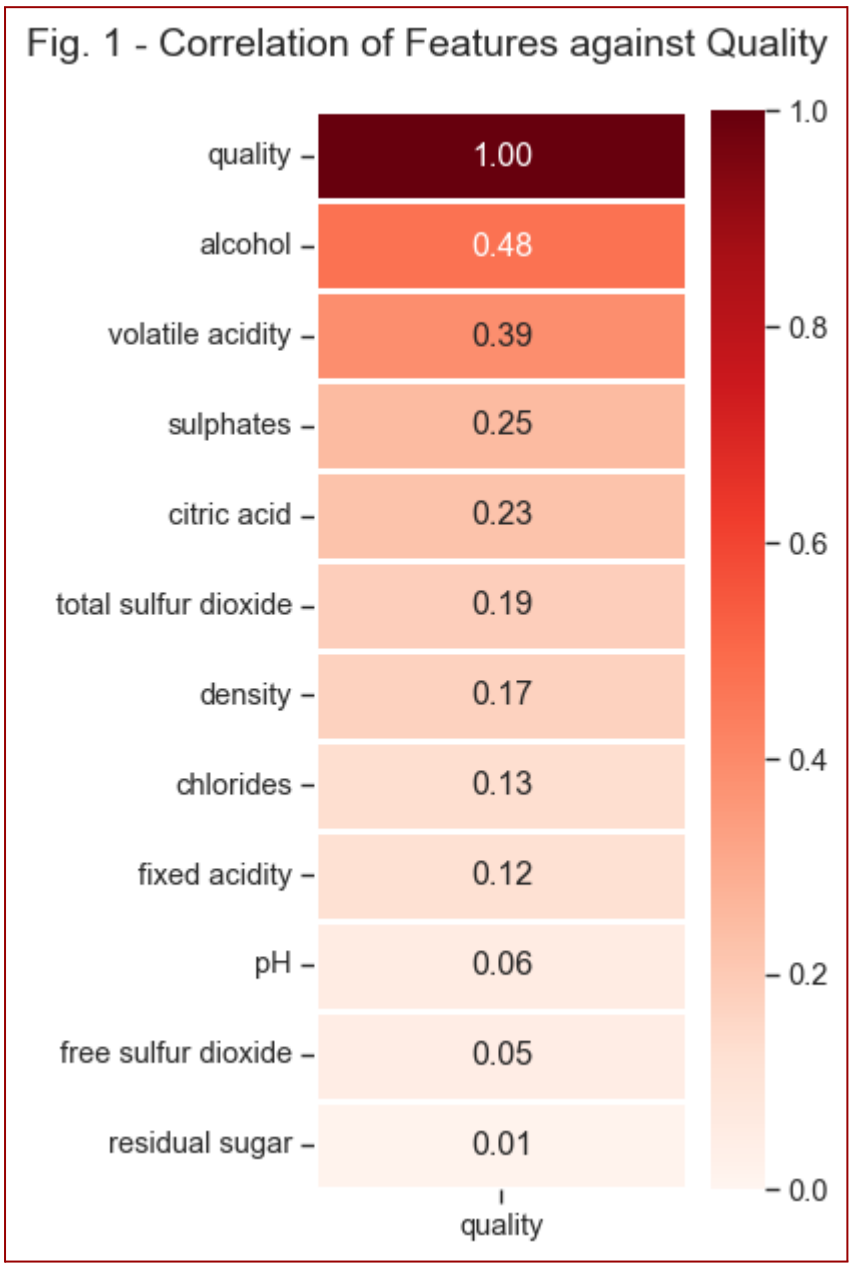


Fig. 1 shows the correlation of each of the features in the dataset compared to 'quality'. Obviously, 'quality' is perfectly correlated with itself and has a score of 1.

Among the other features, those most highly correlated with 'quality' are 'alcohol', volatile acidity', 'sulphates', and 'citric acid'.

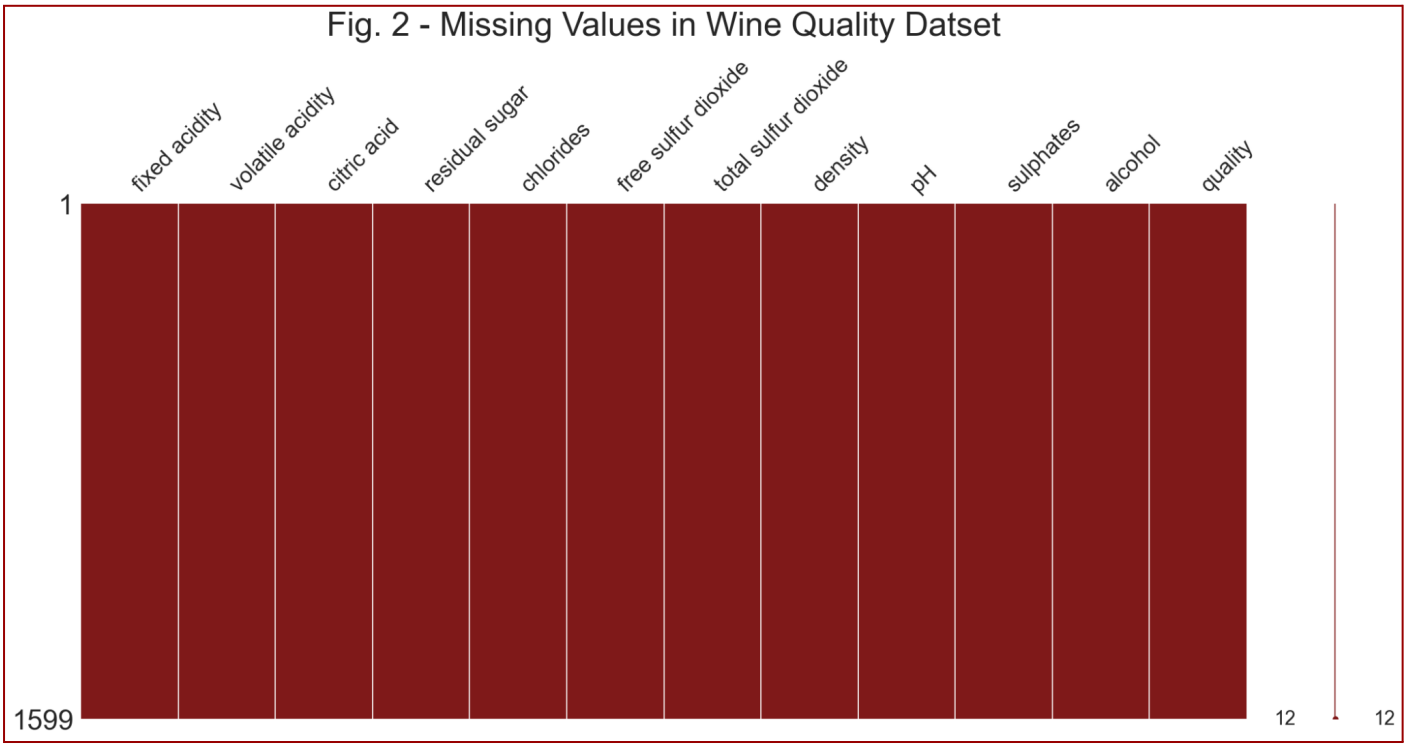
Features such as 'pH', 'free sulfur dioxide', and 'residual sugar' show very low correlation to 'quality'.

We can use this information to help direct our efforts in further analysis, knowing which features have the most linear correlation and which might need to be explored in other ways.

We might want to use this information to drop some of the columns of our dataset if we purely want to explore linear relationships with quality (if planning to perform linear regression for example), but for now let's keep all of the data and see if there are any other non-linear relationships to be found in the data in its current form.

Missing Data

Since the data seems to be in a useable state already and doesn't require any further cleaning, let's have a look now and see if there are any missing values. Fig. 2 (below) shows all NaN values in the dataset.



We would expect to see any missing values represented as horizontal white lines in the bar for that feature. Here, we can see that there are apparently no missing values. This could be because the data has already been cleaned, or because it was collected thoroughly and nothing was missing at any point. We know that the dataset is not using something other than NaN to represent missing values as the column data types are all '*float64*' and therefore must contain either decimal numbers or NaN.

We have been very lucky that this dataset has no missing values and doesn't require any work in the way of data cleaning, so let's move on to some more in depth exploration.

Data Stories and Visualisation

Let's now have a look at the data in more detail and see if we can discover any useful insights that might help indicate 'quality'.

We can begin by exploring the distribution of each of the features to get a sense of the shape of the data, and then also have a look at some scatter plots to see if any clear relationships or patterns exist. After this, we can move on to having a look at a few of the features in a bit more depth to find even deeper compounding relationships between different combinations of the features.

Fixed Acidity

Fixed acidity is a measure of the total amount of acids present in a wine. It is one of the key components that determines a wine's pH and contributes to its overall balance and structure. The main acid found in wine is tartaric acid and it is responsible for the tartness or freshness of the flavour of a wine. It is a naturally occurring acid found in grapes and is present in all wines. The units used to measure this feature are grams per litre (g/L).

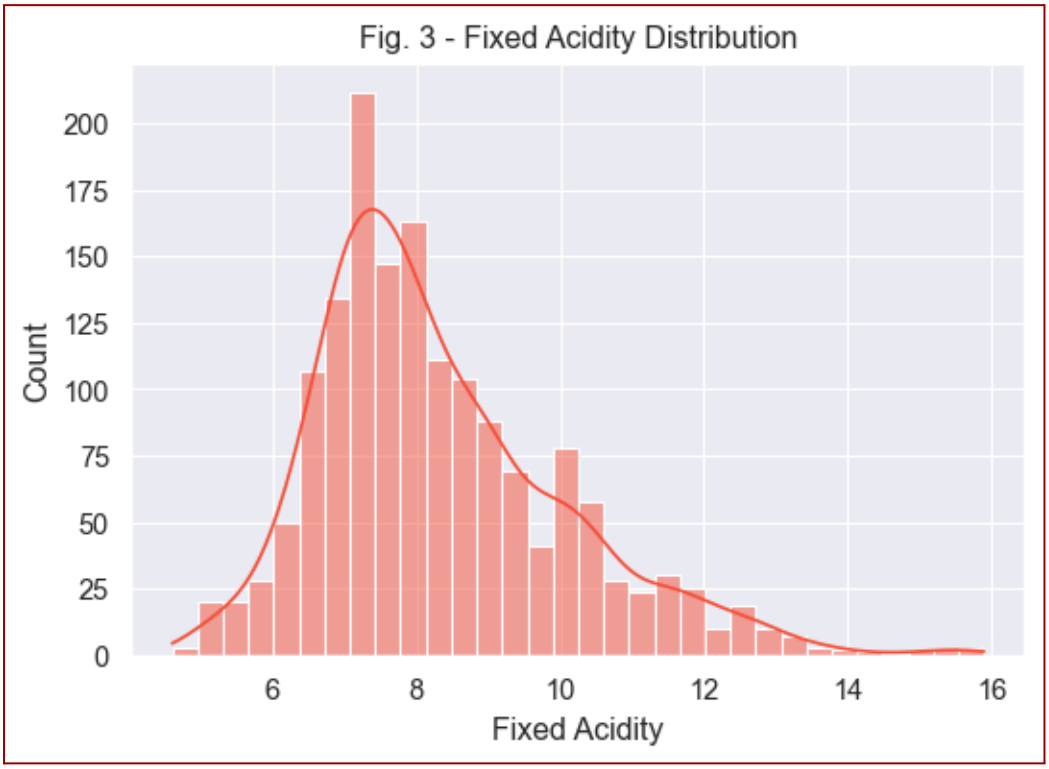
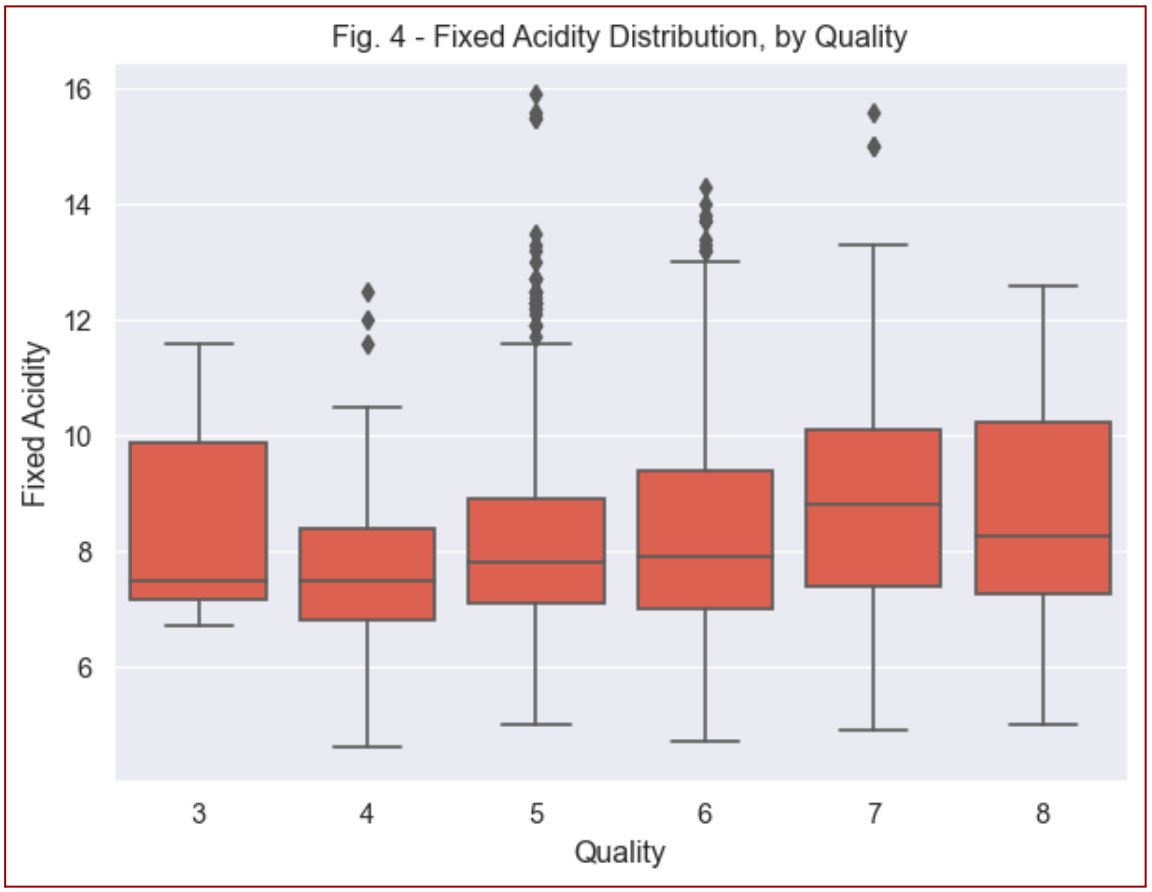


Fig. 3 shows the distribution of the 'fixed acidity' data. We can see that it follows a slightly right-skewed distribution with a peak just above 7g/L.

As such, this data would likely benefit from being normalised before using it in certain types of predictive models.

In Fig. 4, we can see how the fixed acidity data is distributed across the different 'quality' scores.

There doesn't seem to be a clear pattern here, which makes sense given the low correlation coefficient we found earlier of 0.12.



Volatile Acidity

Volatile acidity is a measure of the acetic acid level present in a wine. Acetic acid is responsible for the vinegar-like aroma that can be found in some wines. It is measured here, again, in g/L. The level can be influenced by a number of factors including the variety of grape used, the winemaking process, and the storing/ageing of the wine.

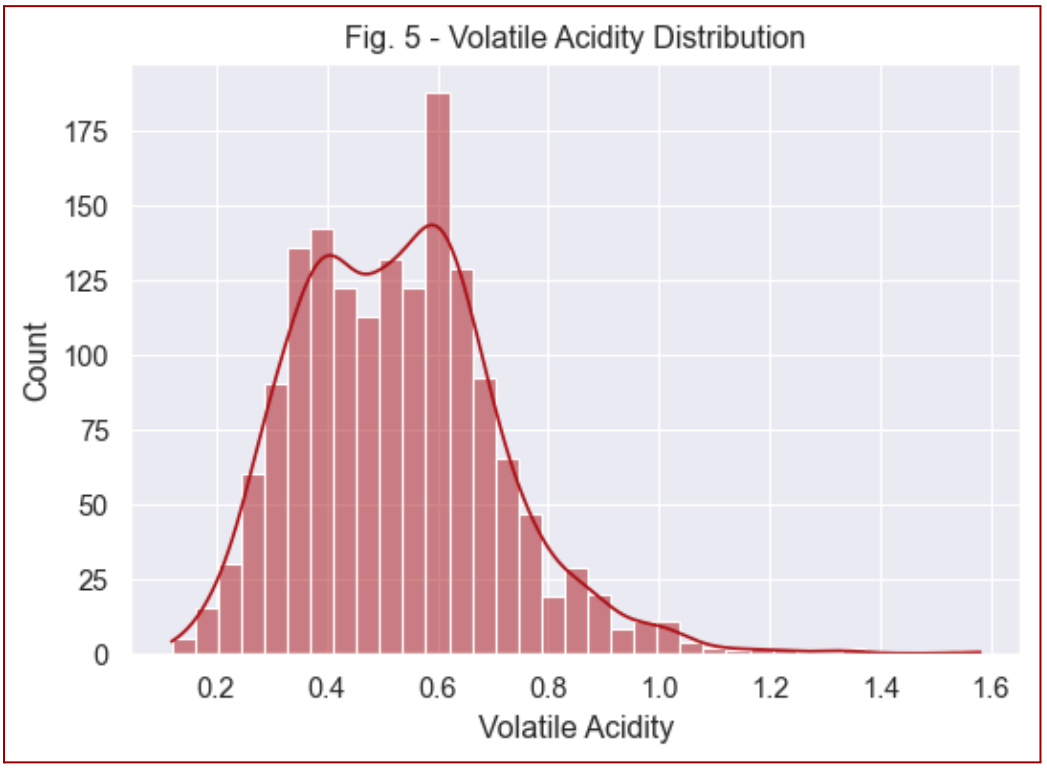
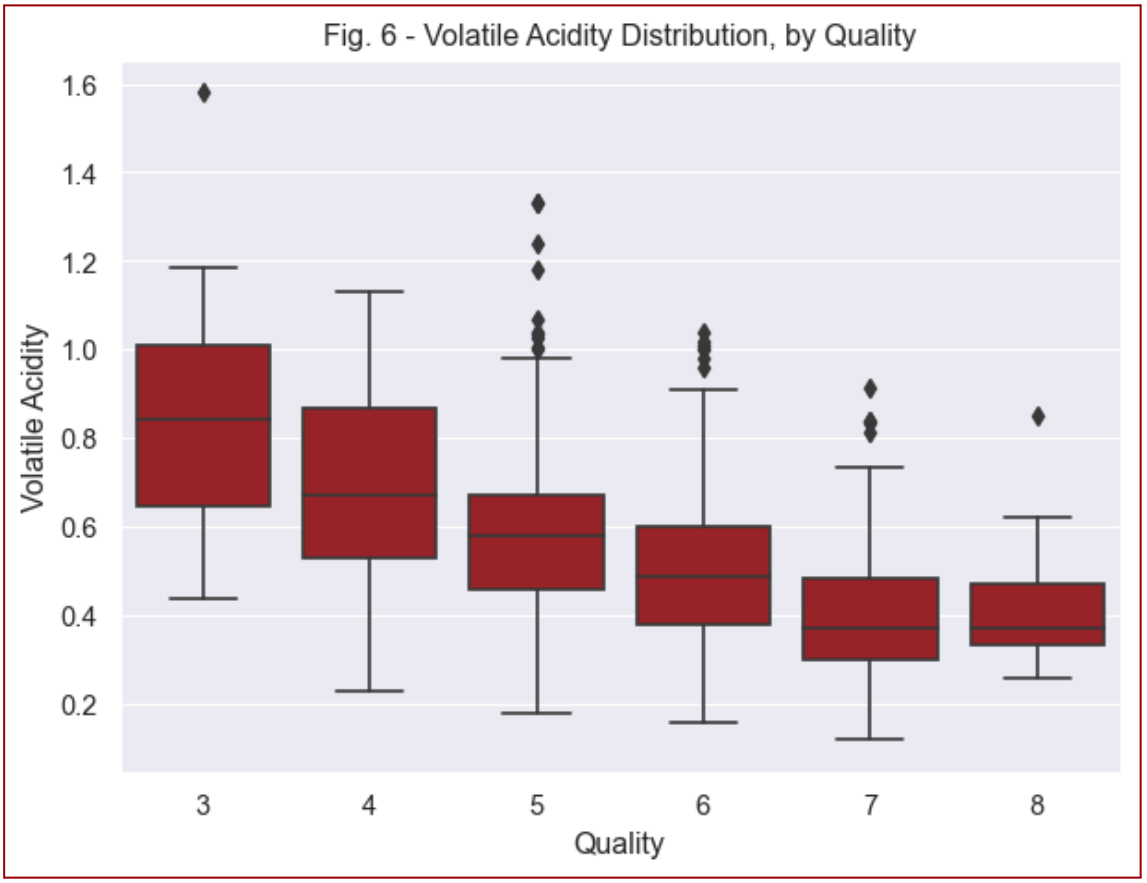


Fig. 5 shows that the distribution of the volatile acidity data follows a slightly right-skewed shape with a peak at around 0.6g/L.

Fig. 6 clearly shows us that lower levels of volatile acidity are generally found in better quality wines.

There seems to be quite a linear relationship here so this would likely be a very useful indicator for many predictive models.



Citric Acid

Citric acid is found in small quantities in wines and can add a 'fresh' flavour if balanced correctly. It is also measured here in g/L. It is found in the grapes used to make the wine and also can be produced during the fermentation process.

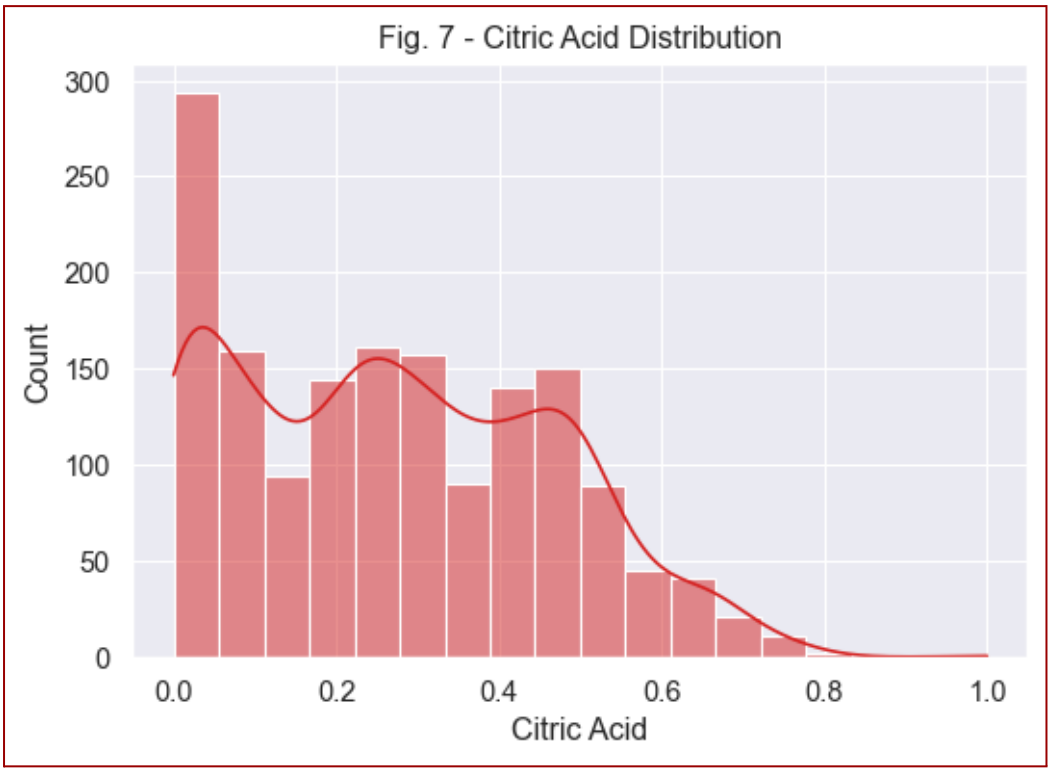
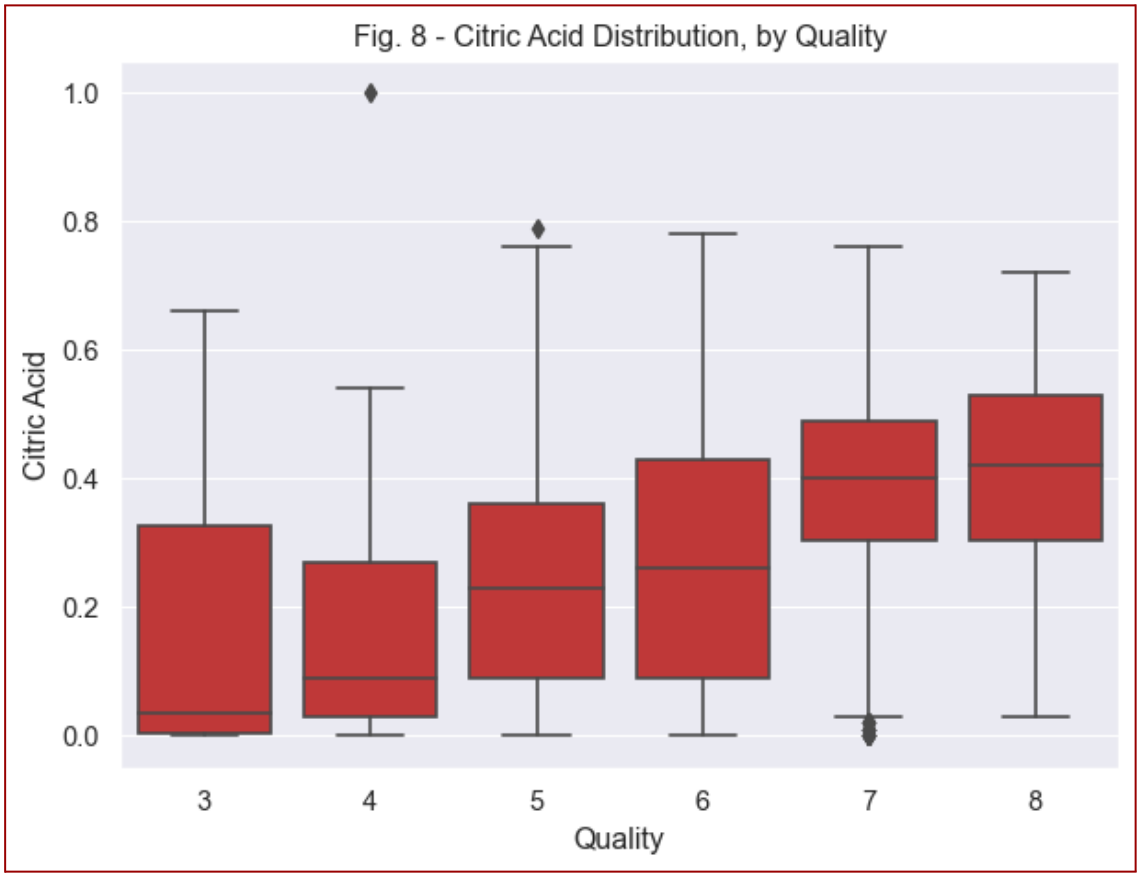


Fig. 7 shows us that the citric acid data doesn't follow a normal distribution at all.

The majority of the wines have extremely low levels of citric acid with the peak just above 0 g/L.

Fig. 8 suggests that there is a positive linear correlation between citric acid levels and wine quality.

This would likely also be a useful predictive indicator to use in various models.



Residual Sugar

Residual sugar is a measure of the levels of sugar remaining in the wine at the end of the fermentation process. It is measured here, again, in g/L and wines with high residual sugar tend to be sweeter while those with lower levels are described as dry.

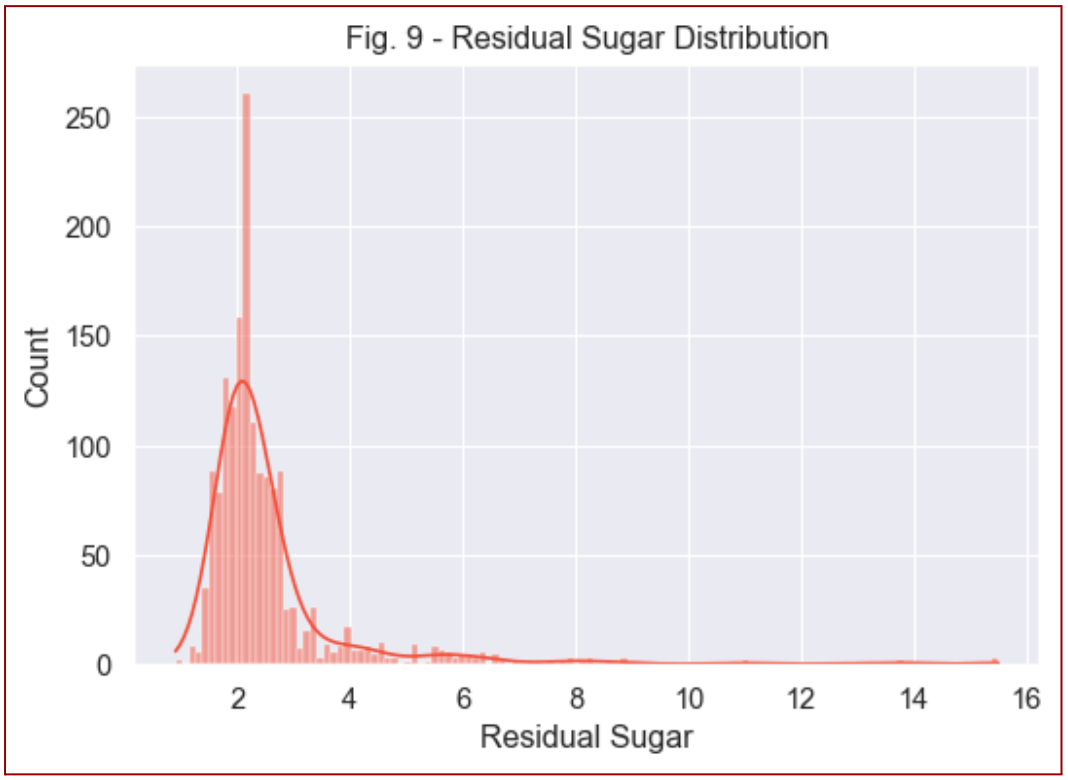
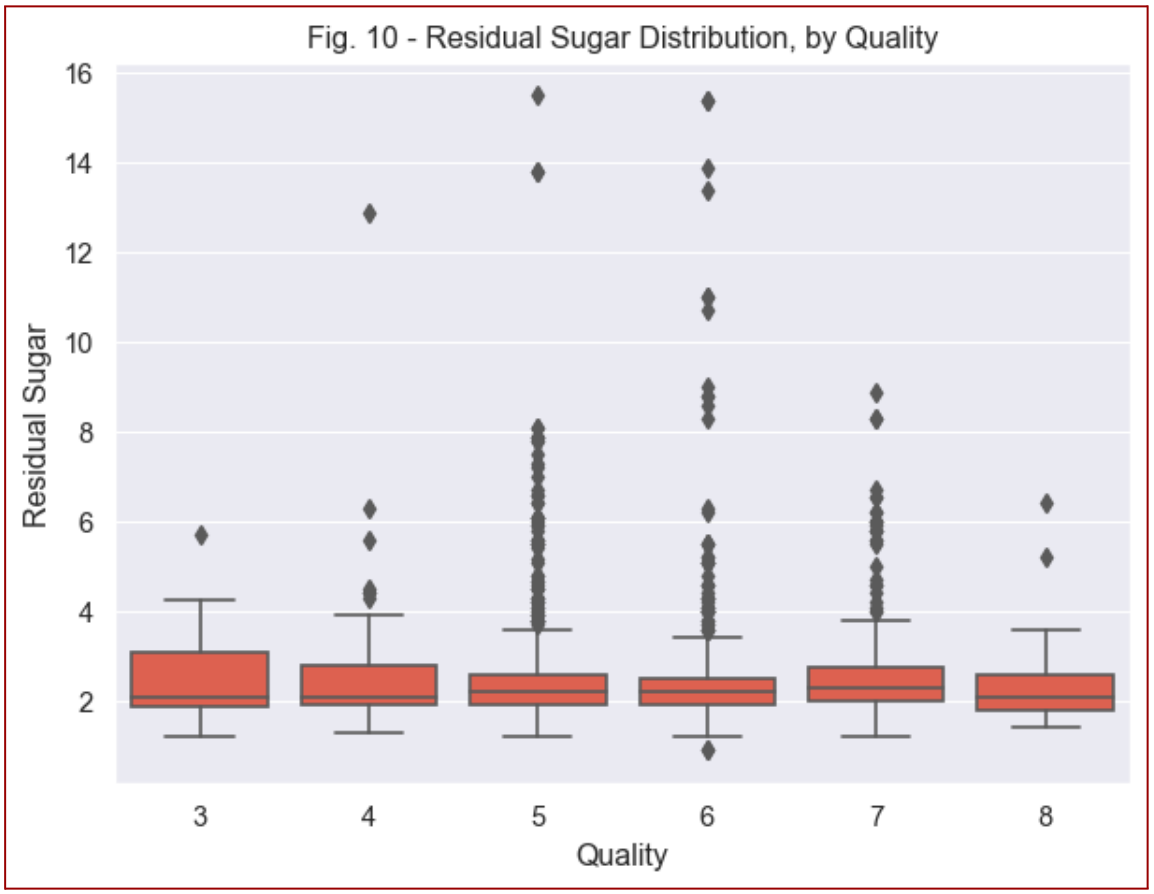


Fig. 9 shows the distribution of the residual sugar data and it suggests that the vast majority of the wines have levels around 2g/L with a few wines having levels much higher (up to over 15g/L).

Fig. 10 confirms that there are many outliers in this data. The vast majority of the wines have similar residual sugar level regardless of their quality.

As such, this probably wouldn't be very useful for linear regression style predictive models.



Chlorides

Chlorides refers to compounds formed when hydrochloric acid reacts with other compounds in the wine. It can be affected by the types of grape used, the soil in which the grapes are grown, and the winemaking process. The levels can affect both the flavour and the mouthfeel of a wine and it is measured here again in g/L.

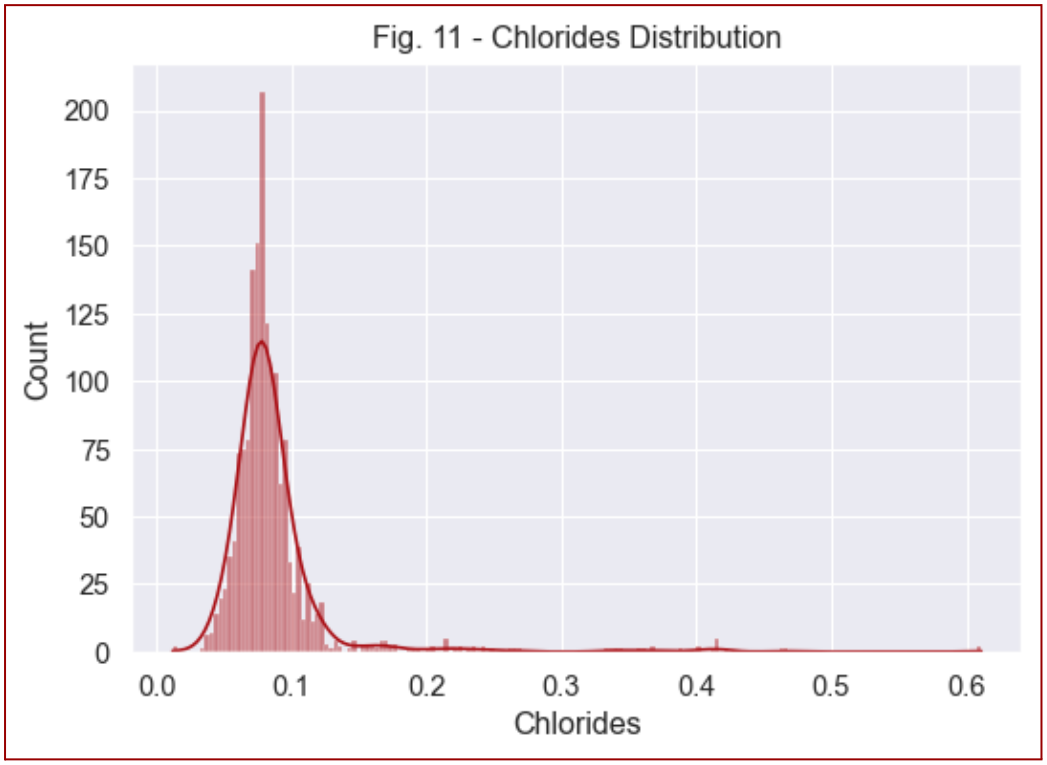
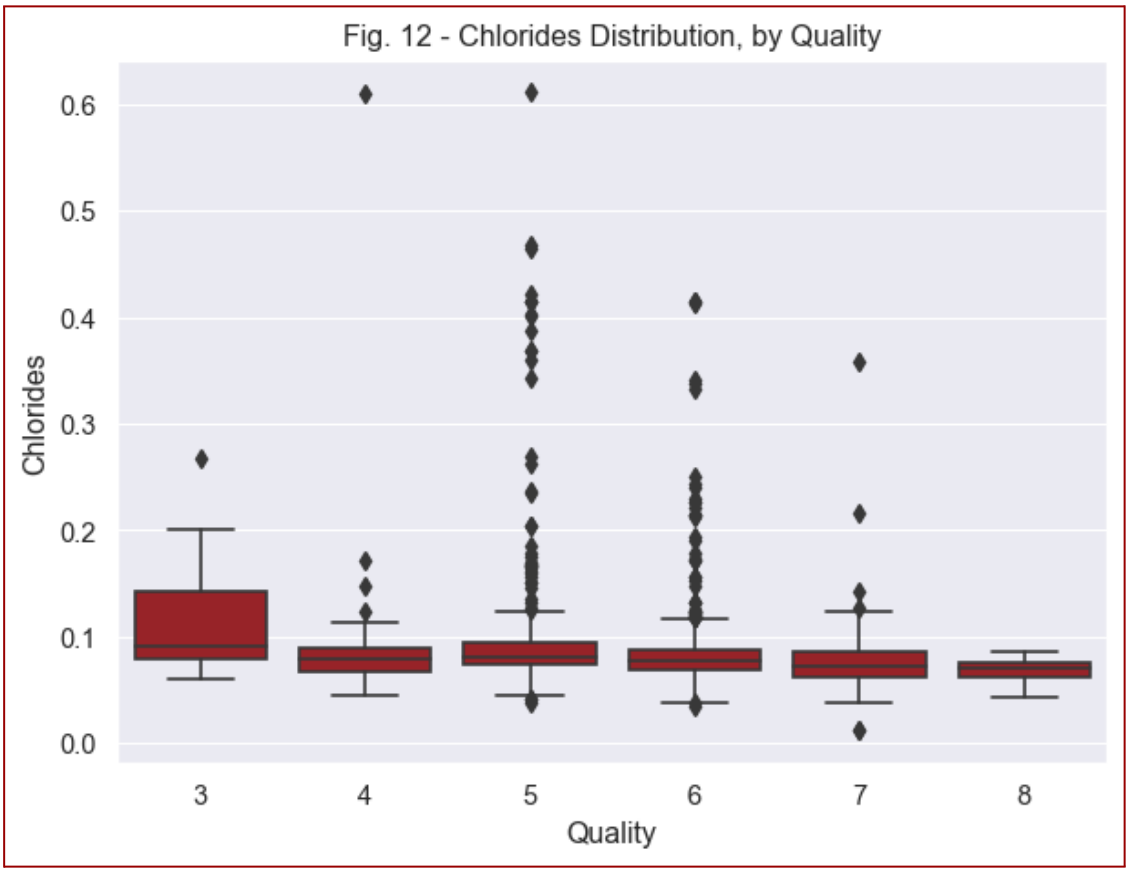


FIg. 11 shows that the chlorides data follows a similar shape to the residual sugar distribution.

The majority of the wines have chloride levels just below 0.1g/L but a few have levels as high as 0.6g/L.

Fig. 12 confirms the concentration of chloride levels just below 0.1g/L across most of the categories.

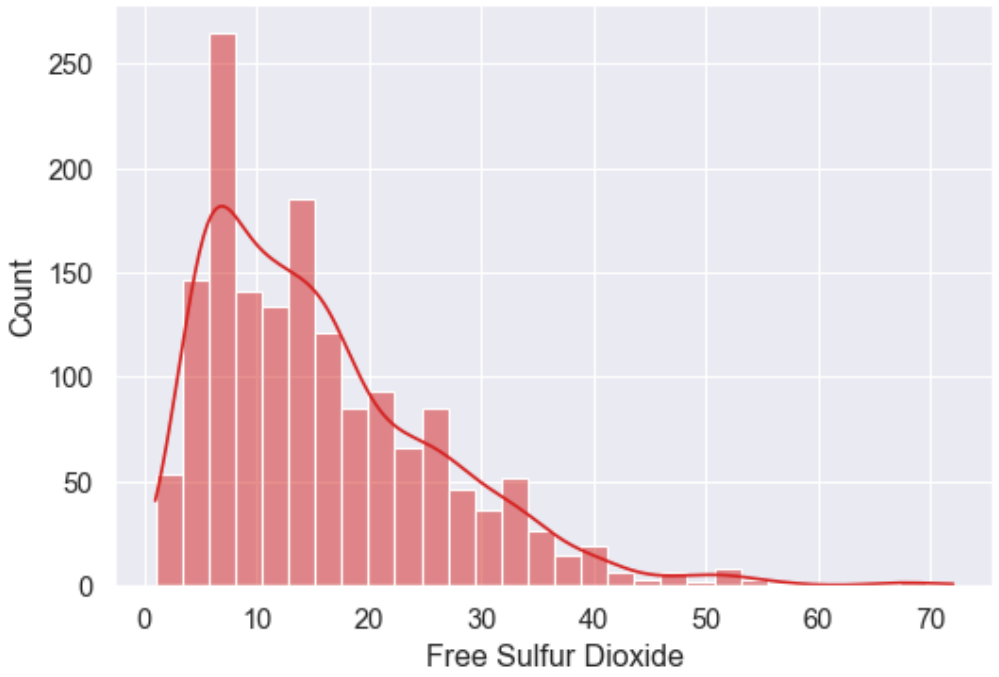
There is a suggestion though, that the lower quality wines have slightly higher chloride levels.



Free Sulfur Dioxide

Free sulfur dioxide is the first of two measurements in this dataset related to sulfur dioxide levels. Sulfur dioxide (SO2) is a chemical compound that is often used in winemaking to help protect wine from spoilage and oxidisation. Levels are measured here in mg/L and can be affected by grape variety, the ripeness of the grapes, and the winemaking process. Sulfur dioxide is described as free when it does not bind to other compounds in the wine (eg. tannins).

Fig. 13 - Free Sulfur Dioxide Distribution



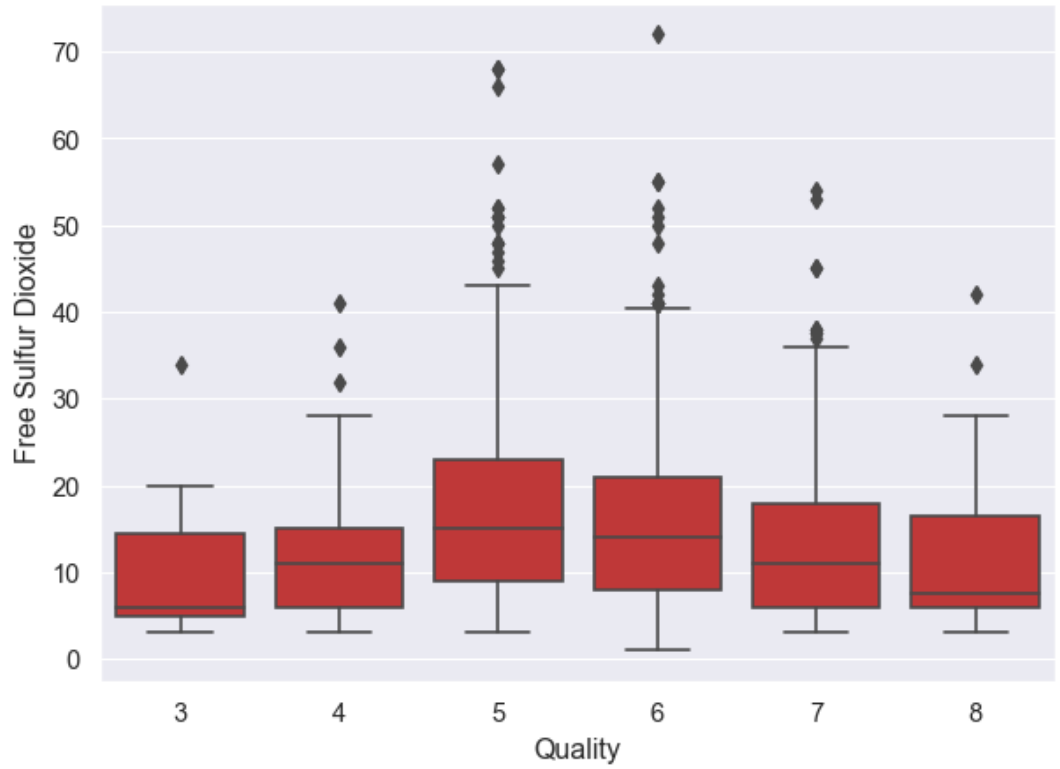
We can see in Fig. 13, that the free sulfur dioxide data has a heavily right-skewed shape with a peak at around 7mg/L.

This data would likely benefit from normalisation before being used in many predictive models.

Fig. 14 shows no clear linear correlation in the data to suggest that higher or lower levels of free sulfur dioxide play a big part in the perceived quality of a wine.

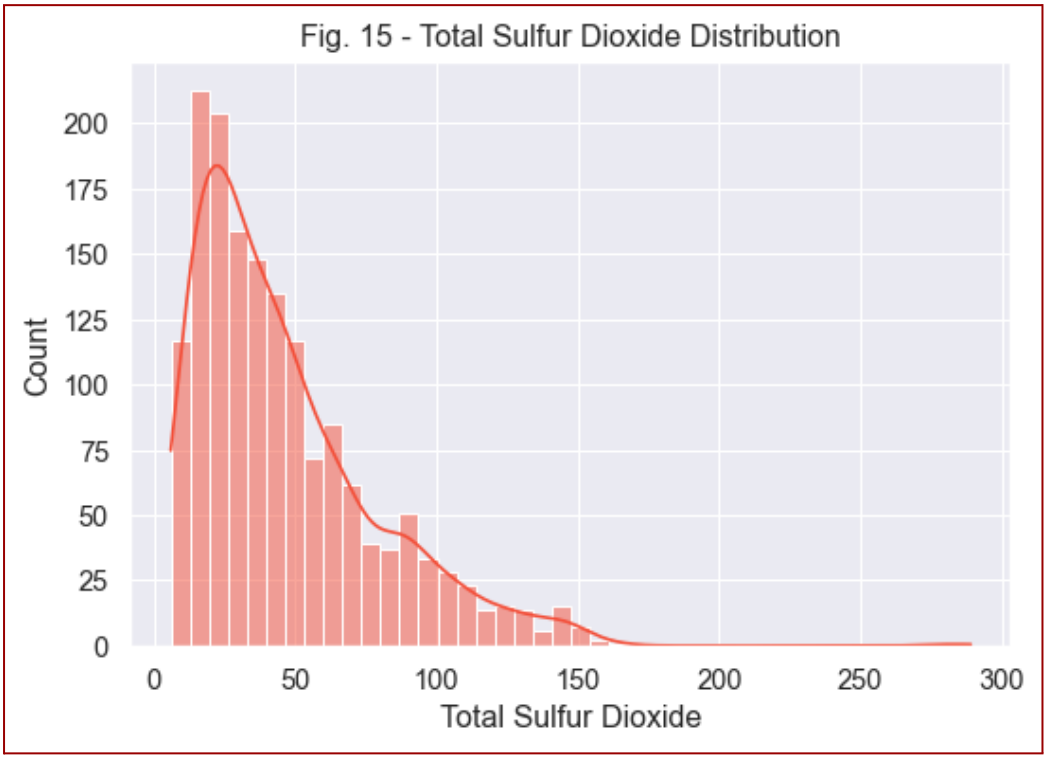
This backs up the information we saw earlier in the correlation heatmap where this feature had a correlation coefficient of only 0.05.

Fig. 14 - Free Sulfur Dioxide Distribution, by Quality



Total Sulfur Dioxide

The total sulfur dioxide data is a measure of the total levels of SO2, both free and bound. This suggests that this data might be highly correlated with the free sulfur dioxide data and this is something we might want to look into further if looking to create linear regression models to prevent multicollinearity which can affect the model performance. This feature is measured here in mg/L.

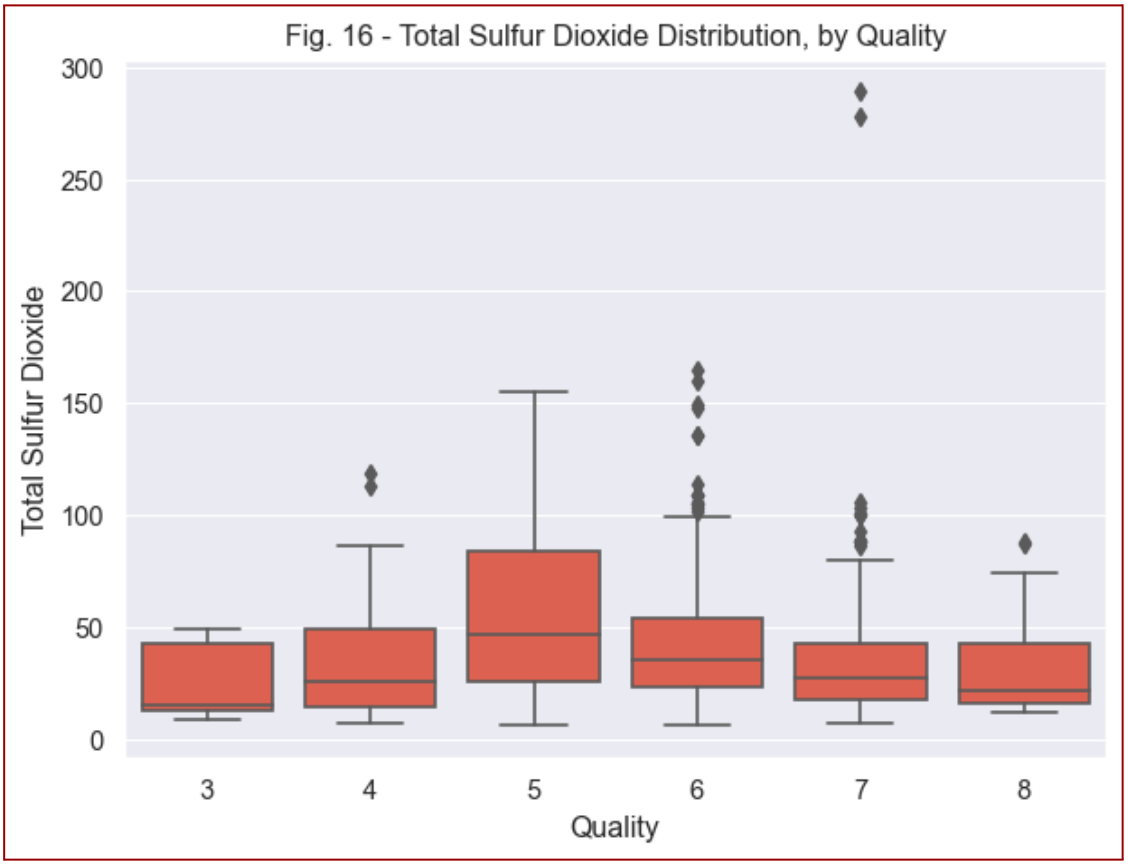


We can see in Fig. 15, that the total sulfur dioxide data has a similarly right-skewed distribution to that of the free sulfur dioxide data.

In fact, it looks to be even more skewed and therefore would again benefit from normalisation.

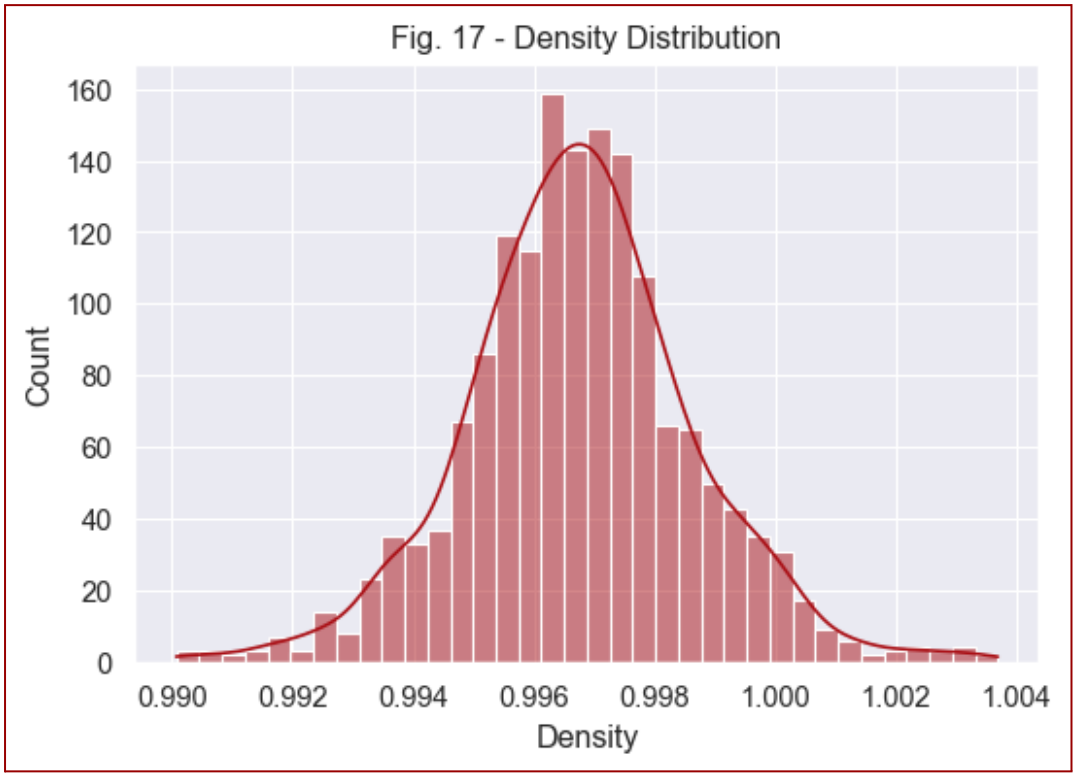
Fig. 16 shows a similar distribution with reference to a wine's quality as we saw with free sulfur dioxide.

It is interesting to note, however, that this feature had a much higher correlation coefficient (0.19 > 0.03), though still would not be considered a strong correlation.



Density

A wine's density is similar to the density of water but can be affected by factors such as alcohol and sugar content. It is measured here as a percentage of the density of water (a density of 1 would indicate the exact same density as water, and 0.8 would indicate it is 80% as dense).

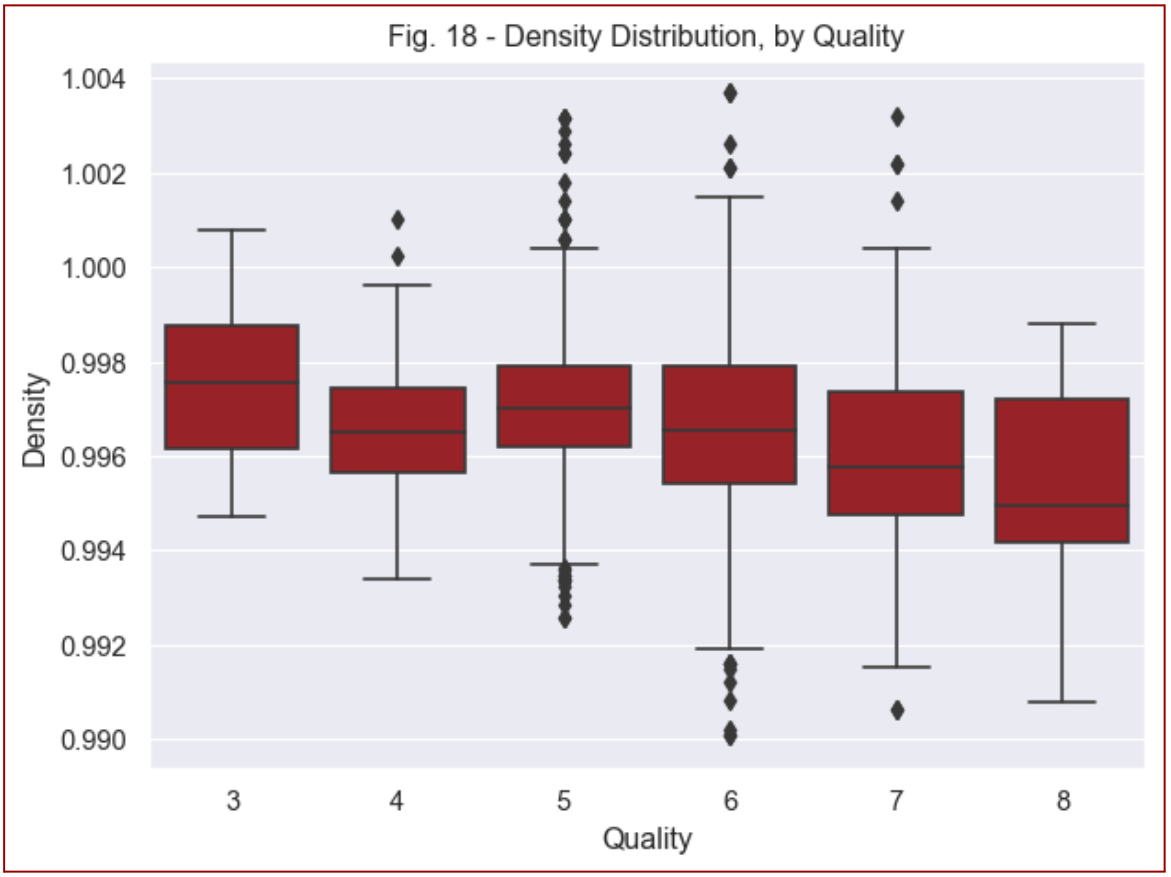


We can see in Fig. 17 that the density data follows a normal distribution with a peak at around 0.997 or 99.7% the density of water.

Most of the wines are less dense than water but there are a few that are more dense. Let's see how this relates to quality.

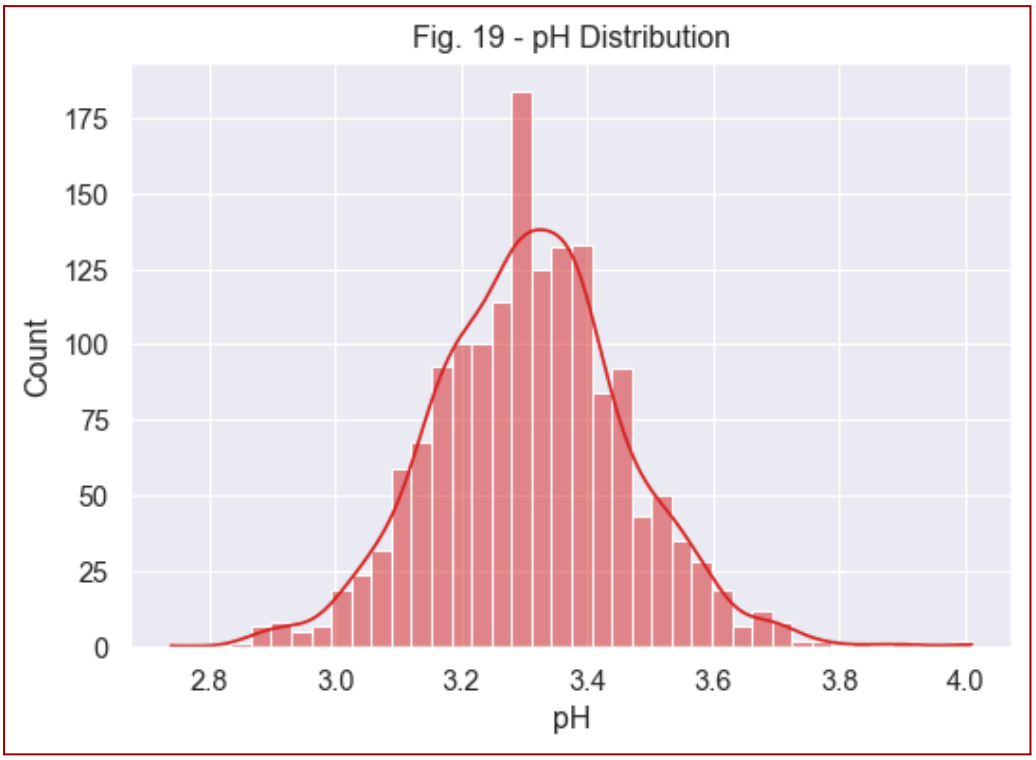
Fig. 18 suggests a slight negative linear correlation, in that as a wine's density decreases, the quality increases.

There would likely be some multicollinearity found between this and 'residual sugar' and 'alcohol' as they directly affect density.



pH

The pH feature is a measure of the acidity of a wine. It is measured on a scale from 0-14 with 0 being the most acidic and 14 being the most basic, a pH of 7 is classed as neutral. The pH of a wine affects both its flavour and its risk of spoilage, so it is carefully monitored during the winemaking process.

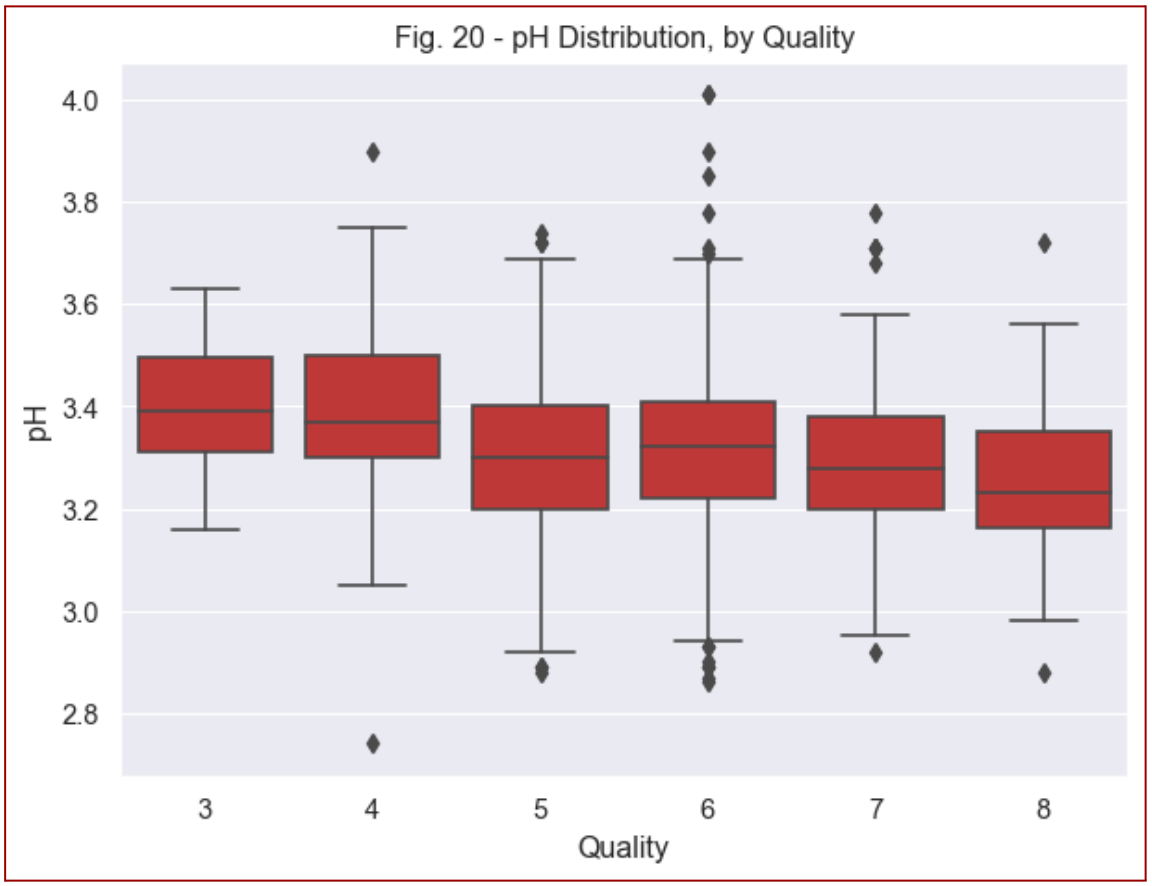


From Fig. 19, we can see that the data follows a normal distribution, peaking at around 3.3.

The majority of the wines have pH levels between 3 and 3.6 with a few outliers beyond that. Let's have a look at how this is distributed against quality.

Fig. 20 shows us that the wines with the lowest 'quality' scores are also the ones with the highest pH.

The more acidic wines (the ones with the lowest pH) are the ones that generally scored the highest.



Sulphates

Sulphates refers to compounds that are produced when SO2 reacts with other compounds in a wine. It again affects both the flavour and the longevity of a wine and can be influenced by grape variety, soil, and the winemaking process. It is measured here in g/L.

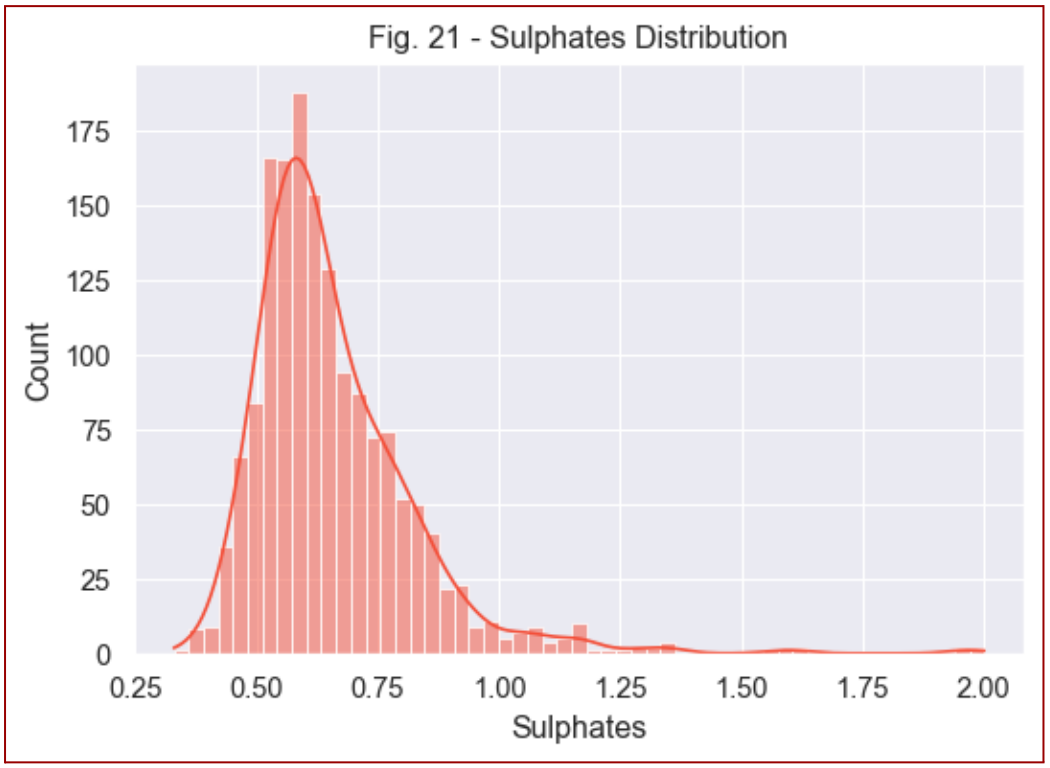
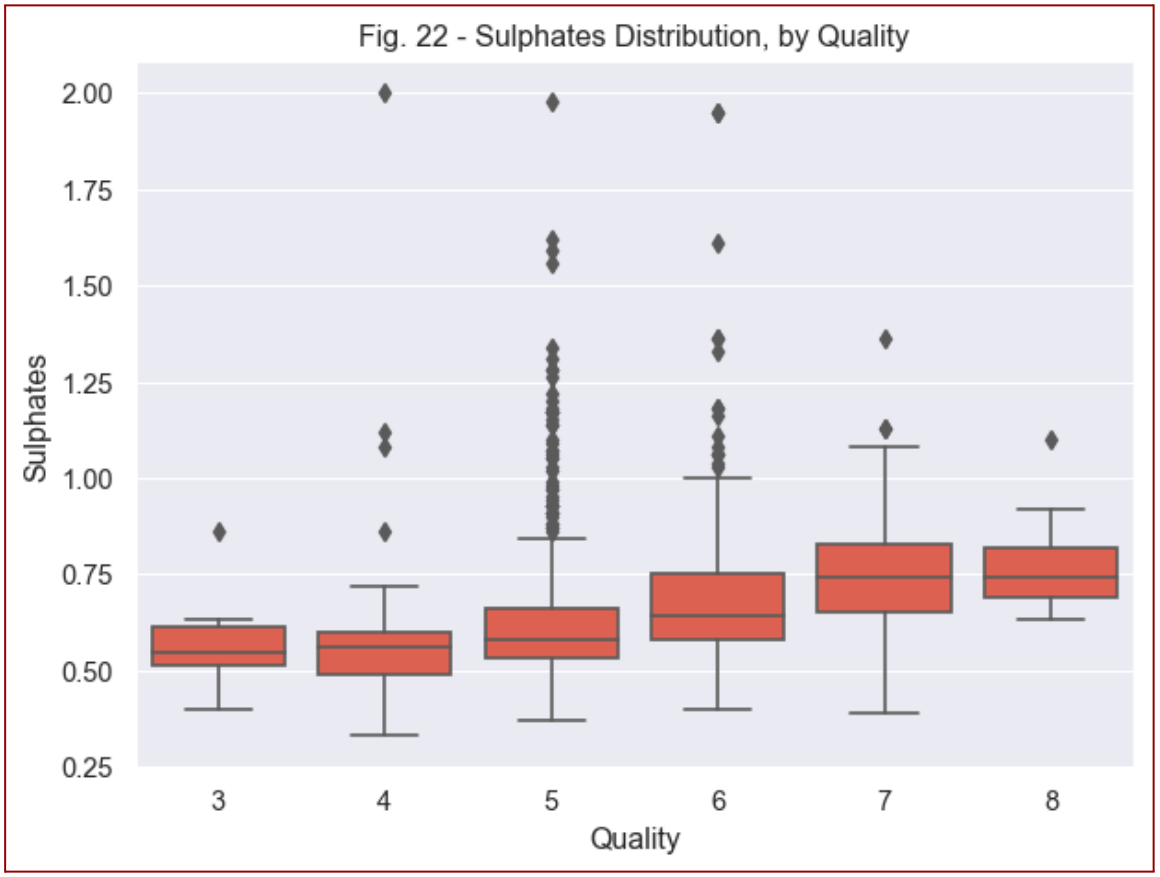


Fig. 21 shows us that the sulphates data has a right-skewed distribution with a peak just above 0.5g/L.

Again, this feature would likely be normalised before using in predictive models.

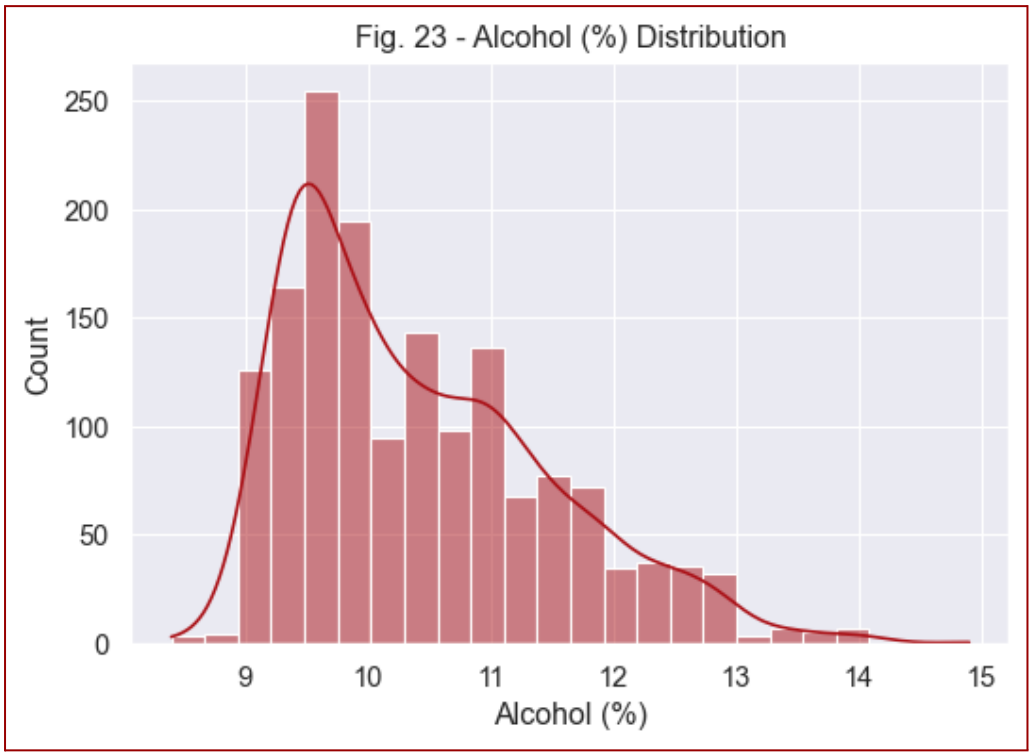
Fig. 22 suggests that higher levels of sulphates are found in the higher quality wines.

This positive linear correlation would likely make it very useful as a predictive indicator of quality.



Alcohol

The alcohol content of a wine refers to the percentage of alcohol by volume (ABV) present in the wine. It is primarily determined by the amount of sugar present in the grapes at harvest and the type of yeast used in the fermentation process.

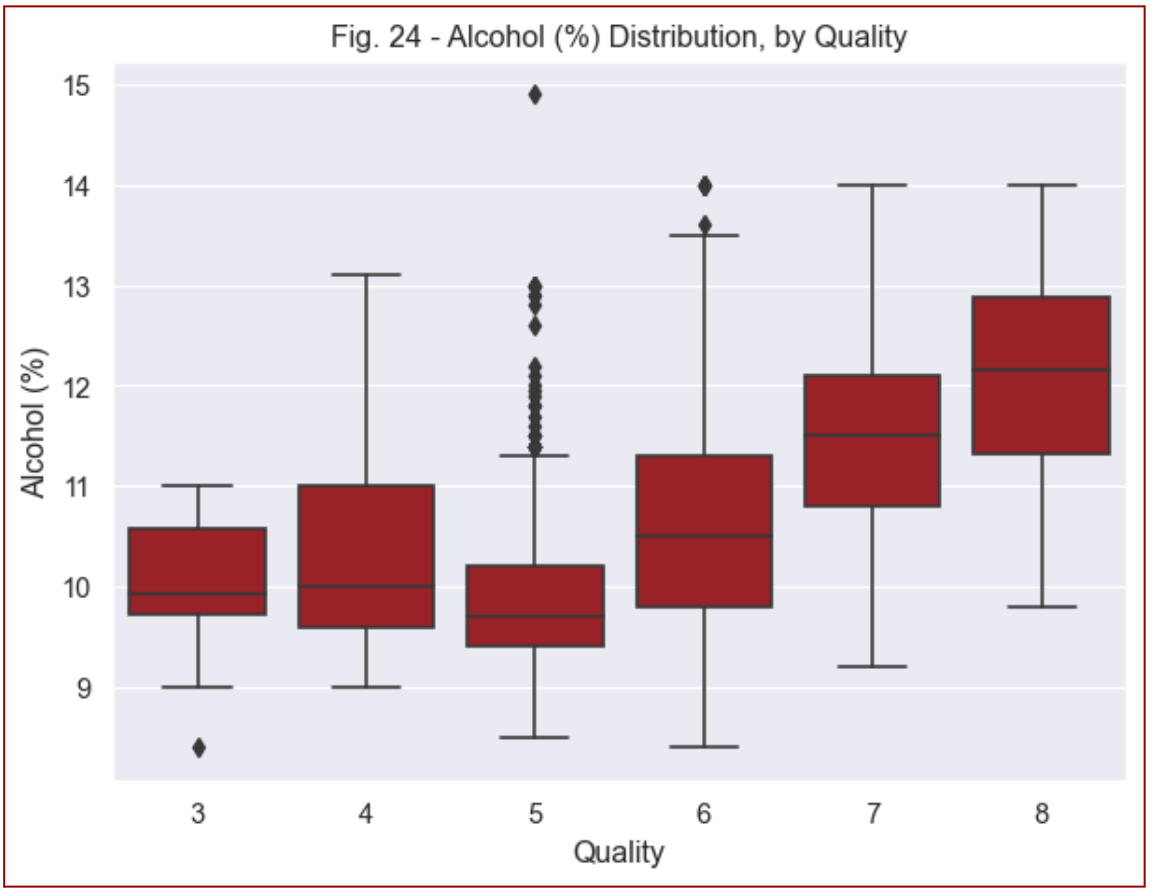


We can see in Fig. 23 that the alcohol content on the wines in the dataset follows a right-skewed distribution.

The data ranges from a little over 8% to just under 15%, let's have a look and see if this might affect the quality.

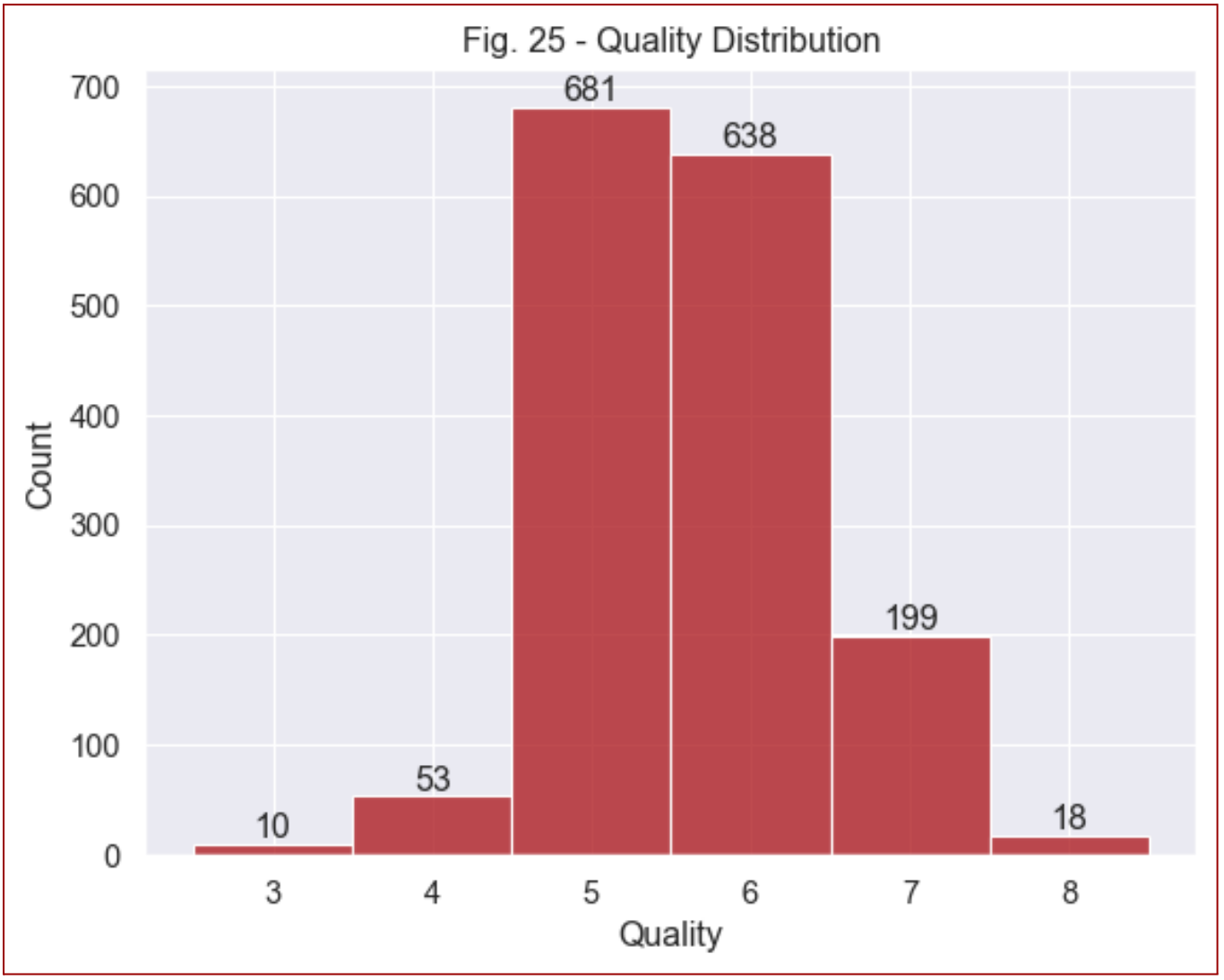
We can see in Fig. 24, that there is a positive correlation between alcohol content and quality.

Generally, wines with a higher alcohol content were perceived as being of a higher quality.



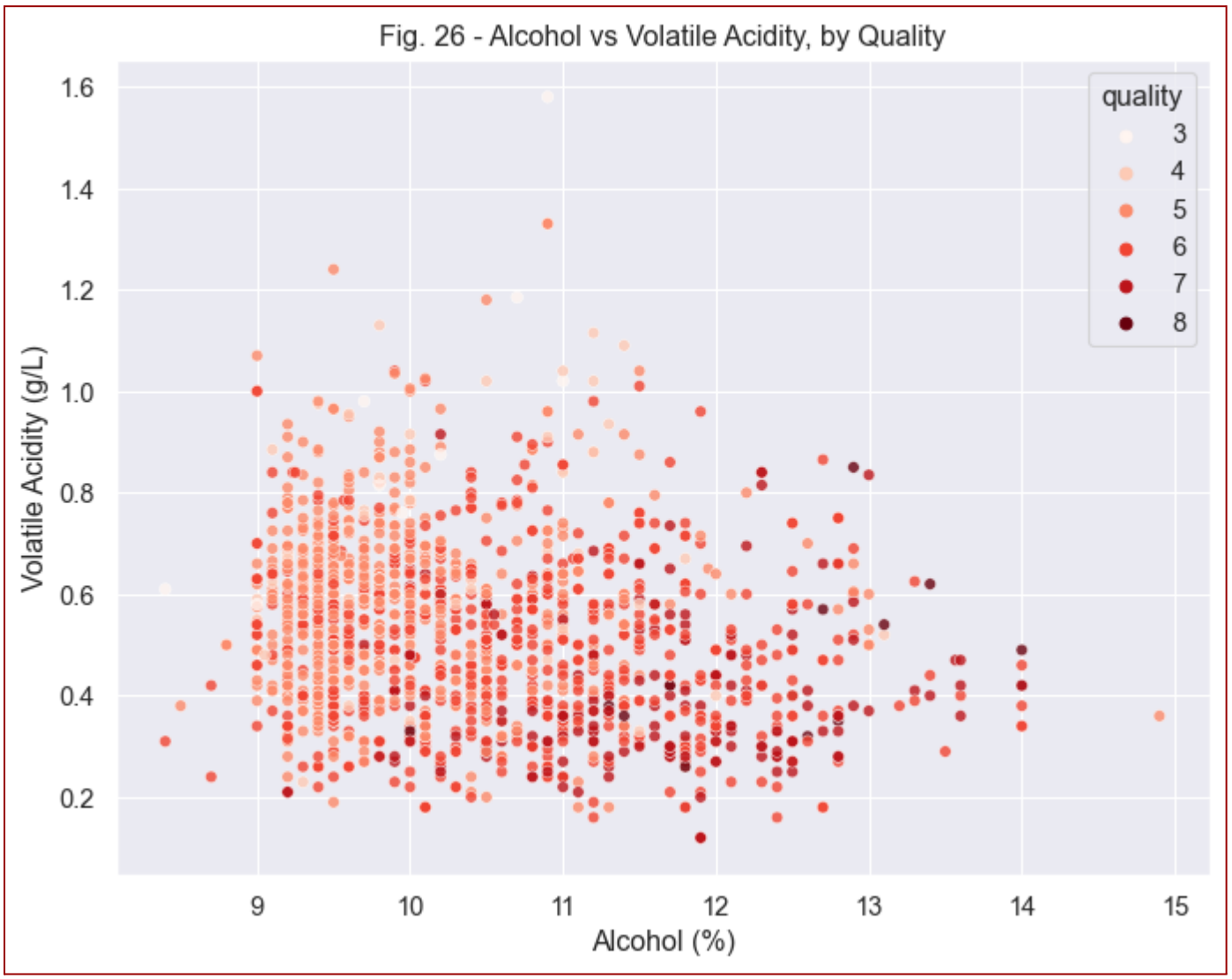
Quality

It is worth noting that some of the scores were awarded more frequently than others. As shown in Fig. 25 below, very few wines were awarded the lowest and the highest scores. Out of 1599 samples, 1319 of them were awarded either 5 or 6. Having such a small sample size for some of the quality categories makes the insights we have gleaned a lot less reliable than they might otherwise be with more data. Still, we have found some patterns that could be used to help predict quality and we can dig a bit deeper to explore some of those relationships in more detail now.



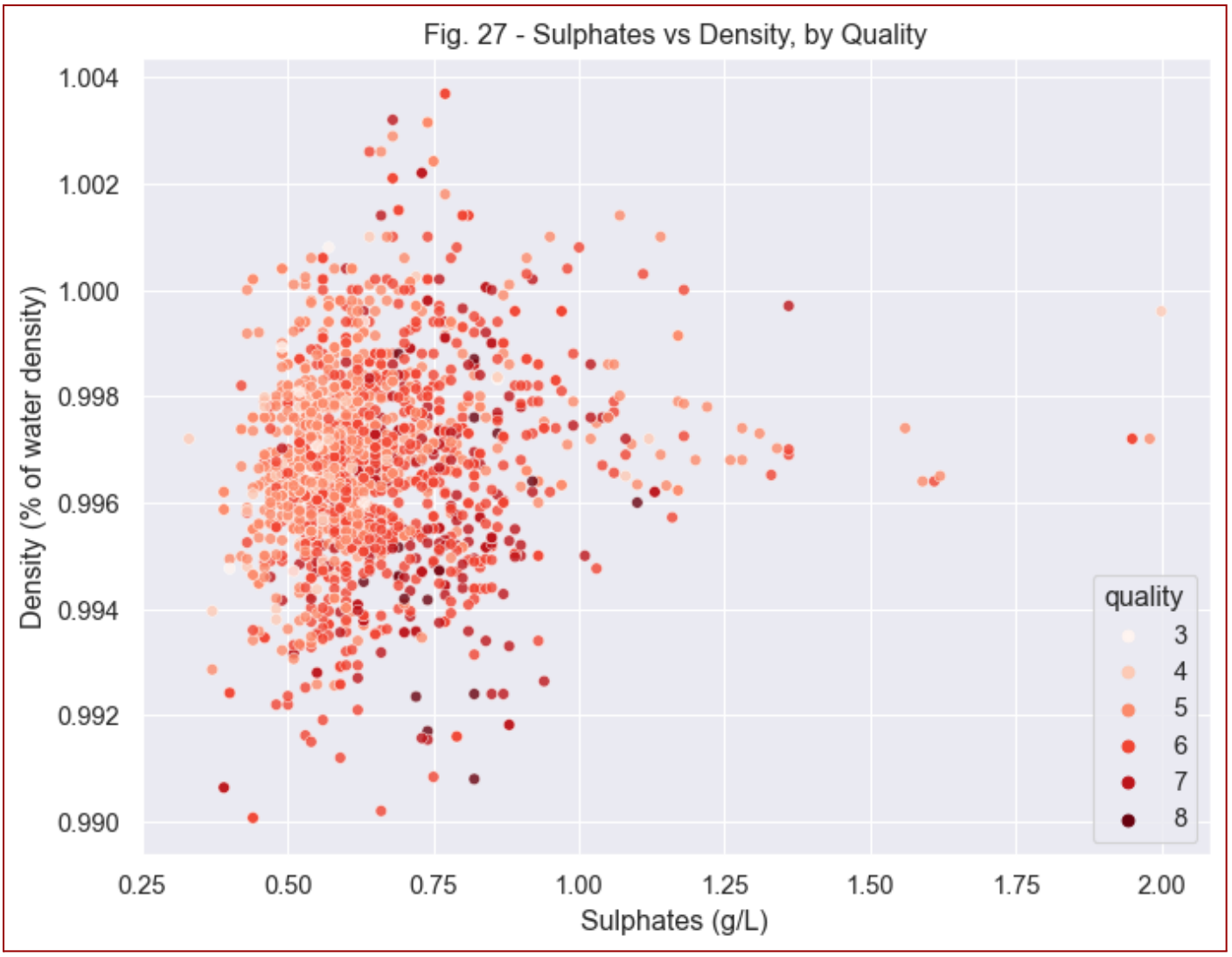
Further Exploration

Let's now have a look at some compound relationships between multiple variables and see if there are any interesting things we can notice. We can start by plotting some of the features that were the most linearly correlated with quality against each other in scatter plots and highlight the quality of each sample.



Above, in Fig. 26, we have plotted the volatile acidity against alcohol content and can see that there is no real linear relationship between these two features. The darker marks represent the higher quality wines and we can see that the marks in the top-left of the cluster seem to be generally paler than those in the bottom right. This suggests that a combination of high alcohol content and a lower volatile acidity indicates a higher quality wine.

Let's now have a look at 2 different features and see if we can glean any similar insights there. Fig. 27 (below) shows the relationship between a wine's sulphate content and density.



Here, we see that again the darker marks are found in the bottom-right of the cluster and that again there is not really a linear relationship between these two features. This suggests that wines with a lower density and higher sulphates are preferred. However, we should note that the wines with the highest sulphate content have lower quality scores, indicating that there is a sweet spot for sulphate levels and that levels either above or below this can have a detrimental effect on the wine's quality.

Let's look at one more pair of features in this way, this time two that we have already looked at: alcohol and sulphates. They both showed some linear correlation with quality so let's see if there are any further insights to be found by looking at them in conjunction with one another.

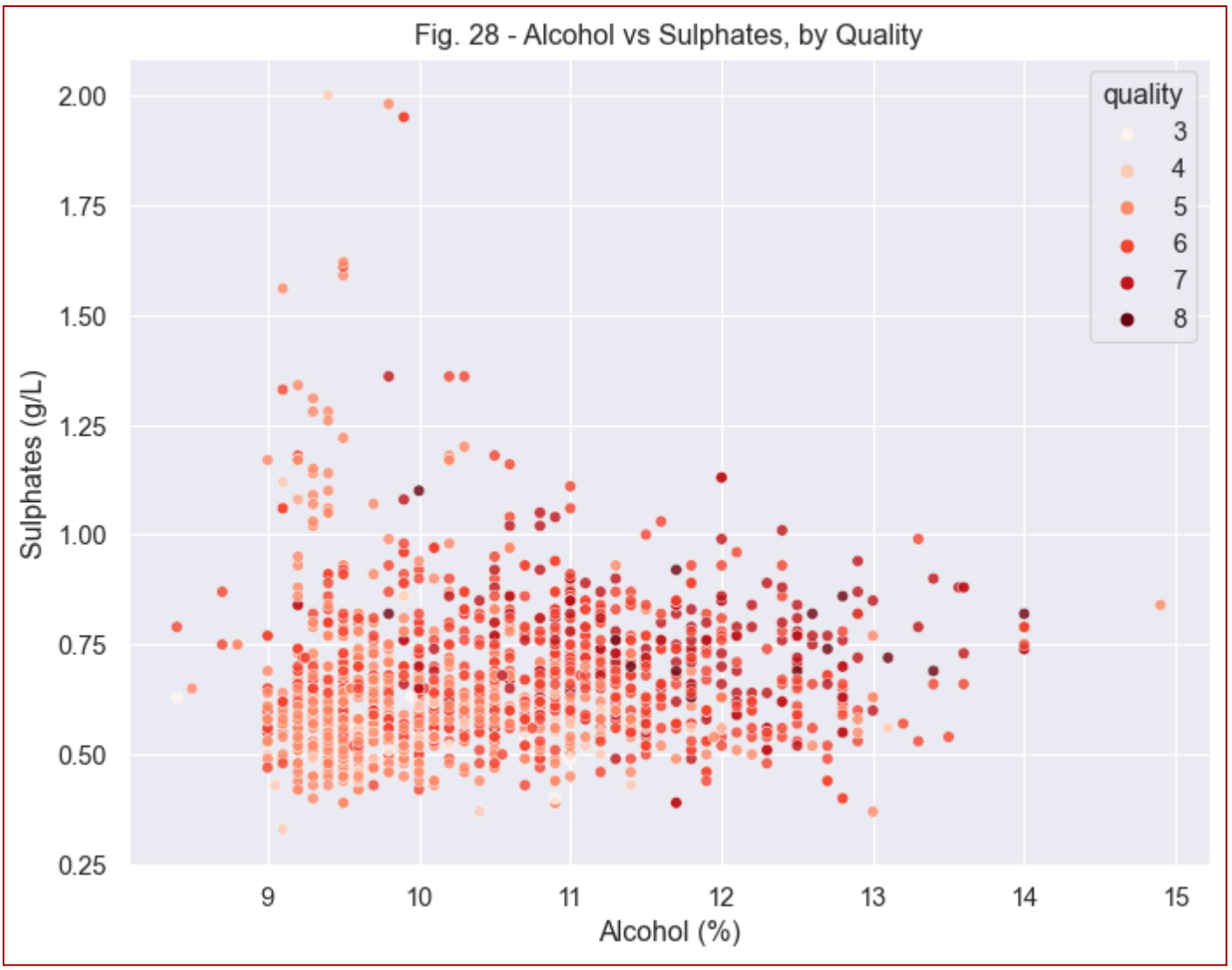
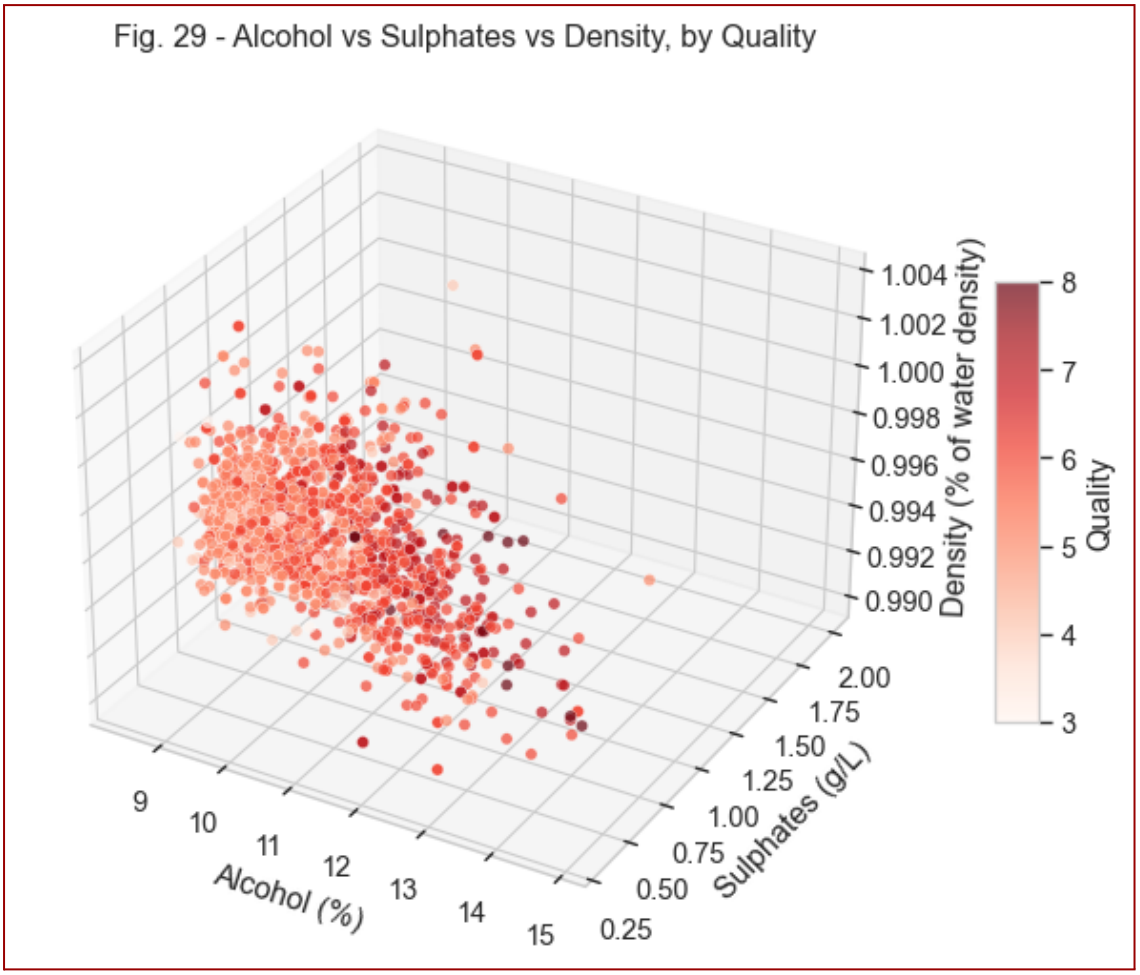


Fig. 28 confirms what we have seen previously, that higher alcohol content wines were scored more favourably in the tests and that the 'sweet spot' for sulphate levels is around 0.6-1.2g/L. Again, there is not much of a linear relationship between the two features themselves, but there is a very slight positive correlation that would indicate that higher alcohol content wines also generally have higher sulphate levels, but this is far from obvious.

We can also view multiple features in conjunction with one another by using a 3-D scatter plot and, to see what this would look like, I have included one below showing 3 of the features that we have explored so far.

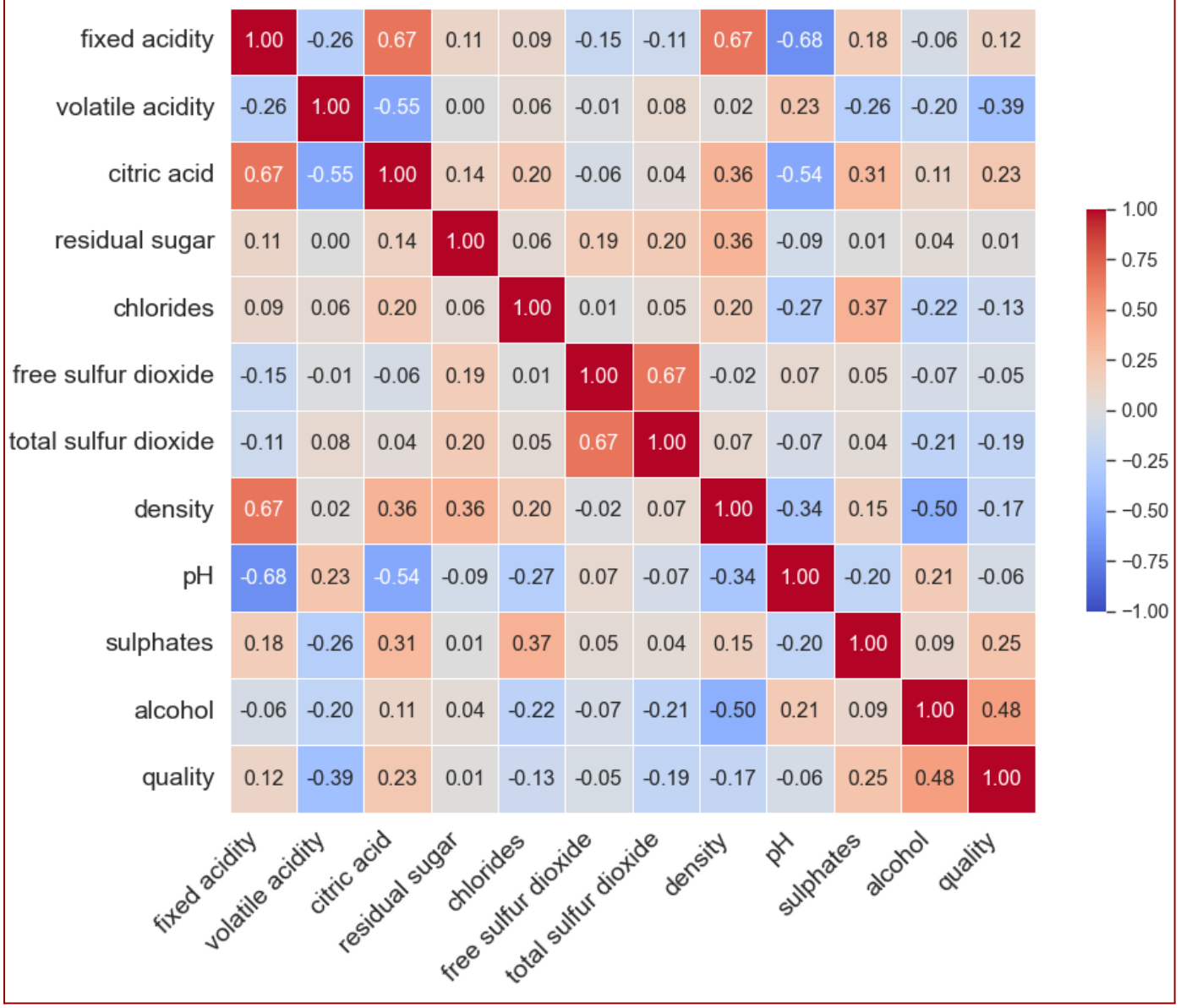


In Fig. 29, we can see confirmation of the patterns that we already found.

The higher quality wines tend to have higher alcohol content, higher sulphates (to a point), and a lower density.

Another way that we can view the correlations between all the different features is in a heatmap like the one that we made earlier to view each feature's correlation with quality, but this time we will include all the correlations between all pairs of features (Fig. 30 below). This would also be helpful in avoiding multicollinearity when selecting which features to use for predictive models by avoiding those that are most correlated with one another.

Fig. 30 - Correlation Heatmap

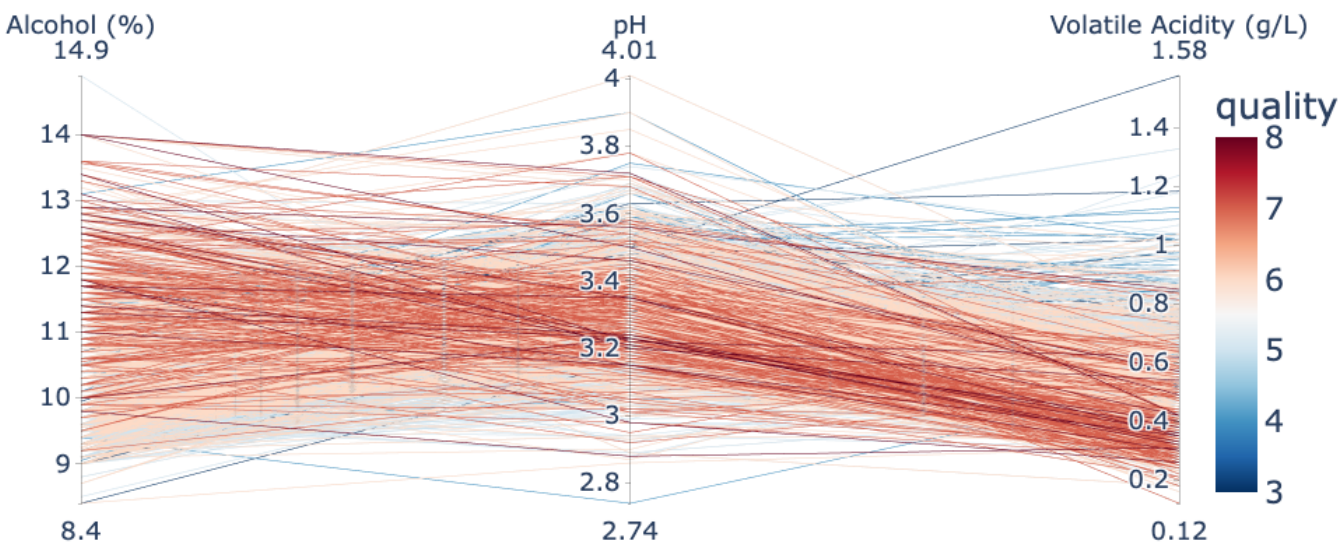


We can see in Fig. 30 that some of the features are highly correlated with each other. For example, pH and fixed acidity have a strong negative correlation meaning that the pH decreases as fixed acidity increases. This makes sense, as the more acidic something is, the lower its pH. There are also noticeable correlations between some other pairs of features, such as: fixed acidity and citric acid, density and fixed acidity, free sulfur dioxide and total sulfur dioxide, and density and alcohol.

We would want to be careful when using combinations of these features together in a multiple regression model to help us optimise interpretability.

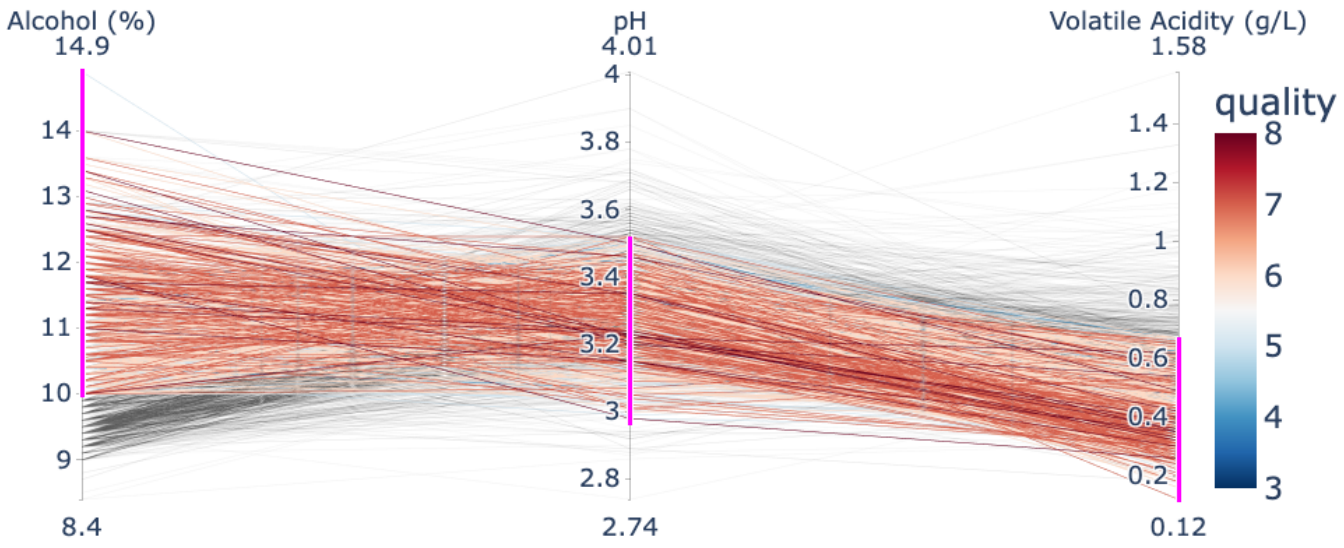
One final way we can visualise the relationships between multiple features in our data is by using a parallel coordinates plot as shown below in Fig. 31-33. Here, we can see the relationships between alcohol, pH, and volatile acidity as they relate to quality.

Fig. 31 - Alcohol vs pH vs Volatile Acidity, by Quality



In these plots, the dark red lines represent the highest quality wines, the paler colours are average wines, and the blue are the lowest quality. We can interact with these plots and create filters to grey out certain ranges of each feature and leave only those that fall within certain ranges. For example, in Fig. 32:

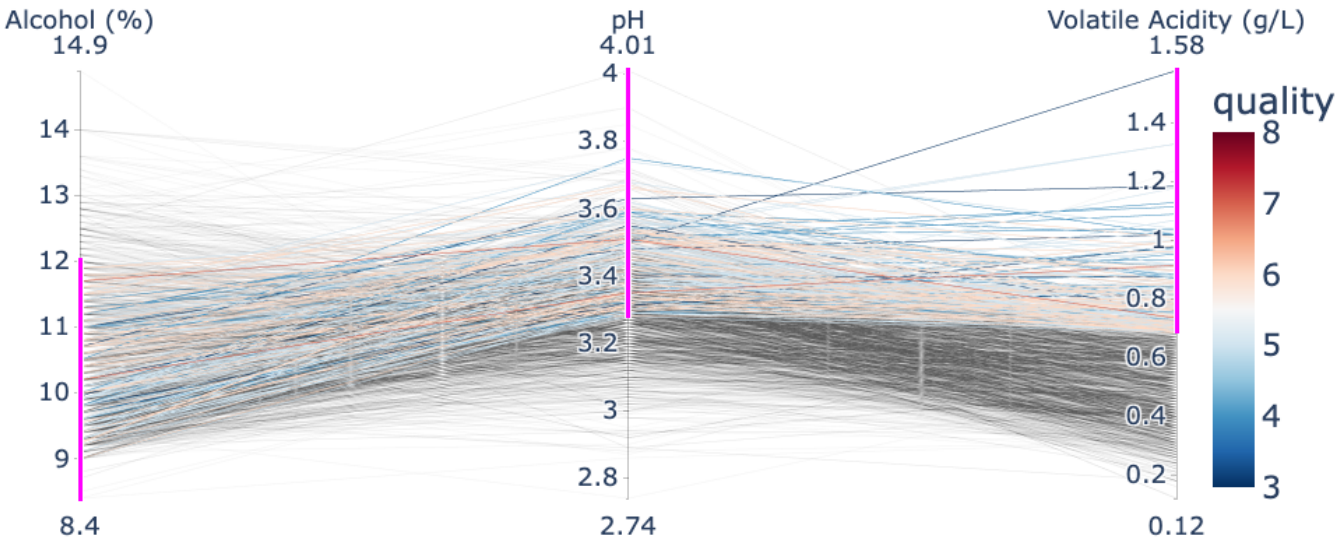
Fig. 32 - Alcohol vs pH vs Volatile Acidity, by Quality



Here, we have filtered for wines that have 10% or higher alcohol content, a pH of between 3 and 3.5, and a volatile acidity of less than 0.7g/L. We can see that, compared to the previous plot, there is a much higher concentration of the dark red lines, with many of the paler lines and blue lines now greyed out.

As confirmation, we can also filter for the opposite and see what that looks like:

Fig. 33 - Alcohol vs pH vs Volatile Acidity, by Quality



This time we have included all wines with an alcohol content of 12% or lower, with a pH of 3.3 or higher, and a volatile acidity above 0.7g/L. We can see a stark contrast between this and Fig. 32 with the coloured lines that have not been filtered out, all representing average or poor wines.

This adds weight to our hypothesis that increased alcohol content, and decreased volatile acidity generally makes a higher quality wine.

Conclusion

After getting a feel for the dataset by checking for distribution patterns and finding correlations and relationships, I would say that the following are some of the main insights that we have found through this exploratory analysis:

- All of the features are numerical as opposed to categorical and very few of them show any strong linear correlation to quality. The highest correlation coefficients in relation to quality are found in the alcohol, volatile acidity, and sulphates features (0.48, 0.39, 0.25 respectively).
- There are many stronger correlations between the features themselves, suggesting the presence of a degree of multicollinearity in the dataset that would have to be addressed before moving on to making a predictive model to try and score wines.
- The alcohol content of a wine has a positive correlation with its quality, generally the “stronger” wines were scored more favourably.
- There is a negative correlation between the volatile acidity (levels of acetic acid) and quality. The lower the volatile acidity measurement, the higher the quality score.
- The level of sulphates in a wine has an effect on its perceived quality, and there seems to be a sweet spot between 0.6-1.2g/L with wines that have levels both higher and lower than this scoring less favourably for quality.
- Combinations of the strongest correlated features reinforce the patterns that we saw previously, but don’t seem to show any major new insights

Next Steps

It would be very interesting to go on to try and build a predictive model to estimate the quality of wines based on the data that we can collect during/after the winemaking process.

We would want to try to select features that provide enough information without having too many that are highly correlated with each other (eg. pH and fixed acidity). Implementing some sort of principal component analysis could also be useful in the feature selection process to try to get the most accurate results.

A lot of the data has a skewed distribution and therefore we would benefit from normalising it before moving on to using it in a predictive model. The data that does follow a normal distribution, such as the pH, would still likely be standardised to optimise model performance.

We could then try a number of different models and compare the results to find something that is effective. We would likely start with some sort of simple linear regression model as a baseline and then perhaps try some other models such as a support vector machine or deep learning methods to get the most accurate estimations.

Final Thoughts

This dataset is clean and thoroughly collected, with no missing data and with accurate numerical measurements that are already formatted correctly for data analysis or machine learning. There are some subtle insights and relationships that can be gleaned through exploration and visualisation, but a clearer picture could be formed through future predictive modelling.

In any case, it seems clear that there are insights in the physicochemical properties of wine that could enhance the winemaking process by giving producers guidelines around certain properties that are common amongst many of the most popular wines. There could also be information here for the discerning buyer who might want a clue as to which wine to choose from the endless supermarket shelves, and that is just as valid an outcome.