Red Wine Characterstics by Perceived Quality by David Broadwater

Overview: This tidy data set contains 1,599 red wines with 11 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

Guiding Question: Which chemical properties influence the quality of red wines?

```
# Load all of the packages that you end up using
# in your analysis in this code chunk.

# Notice that the parameter "echo" was set to FALSE for this code chunk.
# This prevents the code from displaying in the knitted HTML output.
# You should set echo=FALSE for all code chunks in your file.
install.packages(c("plyr", "ggplot2", "scales", "gtable","RColorBrewer"),
                 repos = "http://cran.us.r-project.org")
library(ggplot2)
library(ggthemes)
theme_set(theme_minimal(18))
library(dplyr)
library(gridExtra)
library(tidyr)
library(RColorBrewer)
```

```
# Load the Data
wine <- read.csv('wineQualityReds.csv', header = T)
# Remove X from the dataset, since it's just an observation identifier
wine <- subset(wine, select = - X)

# Create function to find the most frequently occuring value for each
# variable. See `resources_used.txt` for code source.
Mode <- function(x) {
  ux <- unique(x)
  tab <- tabulate(match(x, ux)); ux[tab == max(tab)]
}
```

Univariate Plots

```
summary(wine)
str(wine)
```

Most of the wines were rated 5 or 6 on a 10 point scale, with 75% rated a 6 or below. The ratings in the dataset ranged from 3-8 out of a possible 10 point scale. I wonder if these ratings are typical for red wines, or just this set in particular. I'm surprised that none were rated higher than 8, and that so few were rated as an 8. After reading a bit more about how the ratings were calculated (by taking the median of the ratings from three different wine experts), I can see how there could be

fewer ratings at the extremes of the spectrum.

The density values appear to have a small amount of variance, while it looks like there is much more variance in the residual sugar, free sulfur dioxide, and total sulfur dioxide values. I'm interested to see how these all relate to quality and each other. First, I'll take a look at the distributions of each of the variables to get a feel for each.

Explore Variable Distributions

```
ggplot(aes(x = fixed.acidity),
       data = wine) +
  geom_histogram(binwidth = 0.1)

summary(wine$fixed.acidity)
Mode(wine$fixed.acidity)
```

Fixed acidty was slightly skewed toward higher acidities, but otherwise looked somewhat normal with a long tail extending out to a max value of 15.9 g/dm^3. The median value was 7.9 g/dm^3, and 7.2 g/dm^3 was the most frequently occuring value.

```
ggplot(aes(x = volatile.acidity),
       data = wine) +
  geom_histogram(binwidth = 0.01)

ggplot(aes(x = volatile.acidity),
       data = wine) +
  geom_histogram(binwidth = 0.01) +
  scale_x_continuous(
    limits = c(quantile(wine$volatile.acidity, 0.01),
               quantile(wine$volatile.acidity, 0.99)),
    breaks = seq(0.25, 1, 0.1))

theme_set(theme_minimal(18))

summary(wine$volatile.acidity)
Mode(wine$volatile.acidity)
```

Volatile acidity has a bimodal distribution with peaks around 0.40 and 0.60 g/dm^3. Since there were a couple outliers, I adjusted the x-axis scale a bit to get a better look at the overall distribution. The median volatile acidity value was 0.52 g/dm^3 with a maximum value (and outlier) of 1.58 g/dm^3.

```
wine$total.acidity <- with(wine, fixed.acidity + volatile.acidity)

ggplot(aes(x = total.acidity),
       data = wine) +
  geom_histogram(binwidth = 0.1)

summary(wine$total.acidity)
```

I think it could be useful to see how the total acidity (defined here as the sum of fixed and volatile acidities) relates to the other features, so I've added it to the dataset as `total.acidity`. Here is the distribution of total acidities, which looks similar to the fixed acidity distribution, as expected (since the volatile acidity values are much smaller than the fixed acidity values). The median total acidity was 8.445 g/dm^3, an increase of 0.545 g/dm^3 over the median fixed acidity.

```
wine$volatile.acidity.ratio <- with(wine, volatile.acidity / total.acidity)

ggplot(aes(x = volatile.acidity.ratio),
       data = wine) +
  geom_histogram(binwidth = 0.001)

summary(wine$volatile.acidity.ratio)
```

Now that we know the total acidity, we can calculate the ratio of the volatile acidity to the total acidity for each wine. I added this to the dataset as `volatile.acidity.ratio` and plotted the distribution here. This looks similar to the bimodal distribution seen in `volatile.acidity'. Volatile acidities make up a small portion of the total acidity (which makes sense, if it can cause unpleasant vinegar-like flavors). The median volatile acidity ratio was 0.062 (6.2%), and 75% of the wines had a ratio below 0.079.

```
ggplot(aes(x = citric.acid),
       data = wine) +
  geom_histogram(binwidth = 0.01) +
  scale_x_continuous(breaks = seq(0, 1, 0.25))

summary(wine$citric.acid)
```

Citric acid also had a unique distribution; the most common value was 0.00 (132 wines), with 0.49 as the next most common value (68 wines). I wonder why 0.49 g/dm^3 is such a common value compared to the rest of the values, along with 0.02 (50 wines), and 0.24 (51 wines). It seems odd that so many wines had the exact same non-zero values, especially compared to the surrounding values. These citric acid concentation values seem to fall in about the same range as the volatile acidity values, although they are more skewed in general toward smaller concentrations. The median citric acid concentration was 0.26 g/dm^3, and 75% of the wines had a concentration below 0.42 g/dm^3.

```
wine$citric.acid.ratio <- with(wine, citric.acid / total.acidity)

ggplot(aes(x = citric.acid.ratio),
       data = wine) +
  geom_histogram(binwidth = 0.001)

summary(wine$citric.acid.ratio)
```

Similar to what we did with volatile acidity, let's look at the ratio of citric acid to the total acidity. I added it to the dataset as `citric.acid.ratio` and plotted its distribution here. Unsurprisingly, overall it looks very similar to the citric acid distribution (without the distinct peaks, which were lost in calculating the ratio). The citric acid ratios are smaller than the volatile acidity ratios, with 75% of the wines having a citric acid ratio of 0.043 (which is very close to the 1st Quartile value of 0.042 for the volatile acidity ratios). The median citric acid ratio was 0.031.

```
ggplot(aes(x = residual.sugar),
       data = wine) +
  geom_histogram(binwidth = 0.1)

ggplot(aes(x = residual.sugar),
       data = wine) +
  geom_histogram(binwidth = 0.1) +
  xlim(1, 5)

summary(wine$residual.sugar)
```

Residual sugar had a normal distribution centered around 2.2 g/dm^3 (which also was the median value), with a long tail that extended out to 15.5 g/dm^3. I adjusted the x-axis limits a bit to get a better look at the main distribution.

```
ggplot(aes(x = chlorides),
       data = wine) +
  geom_histogram(binwidth = 0.01)

ggplot(aes(x = chlorides),
       data = wine) +
  geom_histogram(binwidth = 0.001) +
  xlim(0, quantile(wine$chlorides, 0.98))

summary(wine$chlorides)
```

Similarly, chlorides had a normal distribution centered around 0.07 g/dm^3 with a long tail of values that extended out to 0.61 g/dm^3. I also adjusted the x-limits a bit here to get a better feel for the main distribution of values, which appeared mostly normal. The median chloride value was 0.079 g/dm^3.

```
ggplot(aes(x = free.sulfur.dioxide),
       data = wine) +
  geom_histogram(binwidth = 1)

ggplot(aes(x = free.sulfur.dioxide),
       data = wine) +
  geom_histogram(binwidth = 1) +
  xlim(0, quantile(wine$free.sulfur.dioxide, 0.99))
```

```
summary(wine$free.sulfur.dioxide)
Mode(wine$free.sulfur.dioxide)
count(subset(wine,
             free.sulfur.dioxide == Mode(free.sulfur.dioxide)))
```

The free sulfur dioxide values were skewed toward lower concentrations, with 6.0 g/dm^3 as the most common value (138 wines) but a maximum value of 72.0 g/dm^3. The median free sulfur dioxide concentration was 14.0 g/dm^3, and 75% of the wines had a value less than 21 g/dm^3.

```
ggplot(aes(x = total.sulfur.dioxide),
       data = wine) +
  geom_histogram(binwidth = 1)

ggplot(aes(x = total.sulfur.dioxide),
       data = wine) +
  geom_histogram(binwidth = 1) +
  xlim(0, quantile(wine$total.sulfur.dioxide, 0.99))

summary(wine$total.sulfur.dioxide)
Mode(wine$total.sulfur.dioxide)
count(subset(wine, total.sulfur.dioxide >280))
```

Total sulfur dioxide also followed the same pattern, and was heavily skewed toward smaller values. A concentration of 28 g/dm^3 was most common (43 wines), although the median value was 38 g/dm^3. There were some big outliers, with 2 values greater than 280 g/dm^3, while 75% of the wines had values below 62 g/dm^3.

```
wine$free.sulfur.dioxide.ratio <- with( wine,
  free.sulfur.dioxide / total.sulfur.dioxide)

ggplot(aes(x = free.sulfur.dioxide.ratio),
       data = wine) +
  geom_histogram(binwidth = 0.005)

summary(wine$free.sulfur.dioxide.ratio)
Mode(wine$free.sulfur.dioxide.ratio)
count(subset(wine,
             free.sulfur.dioxide.ratio == Mode(free.sulfur.dioxide.ratio)))
```

I wonder what the ratio of free sulfur dioxide to total sulfur dioxide looks like. I've added it to the dataset as `free.sulfur.dioxide.ratio` and plotted its distribution here. Interestingly, the free sulfur dioxide ratio was skewed toward ratios below 0.500 in general (75% of the wines had values below 0.485), while the most common ratio was actually 0.500 (68 wines). I'm not sure why there were so many wines with a ratio of 0.500, especially compared to the surrounding values. The median ratio was 0.375.

```
ggplot(aes(x = density),
       data = wine) +
  geom_histogram(binwidth = 0.0001)

summary(wine$density)
Mode(wine$density)
count(subset(wine, density == Mode(density)))
```

Density had a normal-looking distribution (slightly skewed toward higher densities), with a median value of 0.9968 g/cm^3. The most frequently occurring value was 0.9972 g/cm^3 (36 wines). I'll be interested to see this broken out by quality rating.

```
ggplot(aes(x = pH),
       data = wine) +
  geom_histogram(binwidth = 0.01)

summary(wine$pH)
1-count(subset(wine, pH > 4 | pH <3))/1599
```

The pH values appear to be pretty normally distributed. The background info for this dataset stated that most wines have a pH between 3 and 4, which holds here, with 98% of the wines falling within that range, and 50% falling between 3.21 and 3.40. The median pH was 3.31.

```
ggplot(aes(x = sulphates),
       data = wine) +
  geom_histogram(binwidth = 0.01)

ggplot(aes(x = sulphates),
       data = wine) +
  geom_histogram(binwidth = 0.01) +
  xlim(min(wine$sulphates), quantile(wine$sulphates, 0.99))

summary(wine$sulphates)
```

The sulphate values also had a long tail, probably related to the long tail we saw in the total sulfur dioxide distribution (since our background info states that sulphates contribute to the overall sulfur dioxide levels). After adjusting the x-axis limits, the main distribution of values appears mostly normal, but skewed toward larger values. The median sulphate concentration was 0.620 g/dm^3, with 75% of the values falling below 0.73 g/dm^3. The maximum value (and outlier) was 2.00 g/dm^3. I suspect that value is related to the max-value outlier seen in the total sulfur dioxide values.

```
ggplot(aes(x = alcohol),
       data = wine) +
  geom_histogram(binwidth = 0.1)

summary(wine$alcohol)
Mode(wine$alcohol)
count(subset(wine, alcohol == Mode(alcohol)))
```

The alcohol values were skewed toward larger percentages, with 50% of the wines between 10.2 and 14.9 % by volume. The most frequently occurring alcohol percentage was 9.5 (139 wines), which was also the 1st Quartile value.

```
ggplot(aes(x = quality),
       data = wine) +
  geom_histogram(binwidth = 1)

summary(wine$quality)
summary(factor(wine$quality))
1-count(subset(wine, quality > 6 | quality <5))/1599
```

The quality values were somewhat normally distributed, but slightly skewed toward higher ratings. The median rating was 6 (638 wines), but the most frequently occurring rating was 5 (681 wines). 82% of the wines had one of those two ratings. Only 18 wines were rated 8, and the ratings only varies between 3 and 8 (on a 0-10 scale), again likely due to the way the ratings were calculated (the median rating from the three wine experts).

Univariate Analysis

What is the structure of your dataset?

There are 1599 different observations (of wines) of 13 different wine characteristics. All of the variables except X (the observation identifier) and quality (the rating from 0-10 given by the wine tasters) were floating numerical values and were measured values.

What is/are the main feature(s) of interest in your dataset?

The main feature of interest is the `quality` rating. I'm trying to determine how the other features potentially influence it.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

I think that the acidity features (volatile acidity, fixed acidity,and citric acid), residual sugar, sulfur dioxide features (free and total sulfur dioxide), pH, and density will have the biggest impacts on overall quality. I'm not much of a wine drinker though, so my unfamiliarity with chlorides and sulphates (in relation to wine) are probably adding bias to my assumptions. Even though there seems to be a small amount of variance in the densities, I think that could impact what is perceived as "mouthfeel", so I think density could be significant. Sugar, acidity, and pH seem like they would impact how well "balanced" a wine is perceived

to be, since something that is very tart or very sweet might be off-putting, if nothing else because they would taste different than what a wine drinker is used to or expecting. Volatile acidity is mentioned as creating a vinegar-like taste in high-concentrations, so I suspect that will also affect taste and quality. Lastly, since the dataset description mentions that large concentrations of free sulfur dioxide become apparent in the nose and taste of a wine, I think it is very likely to have an impact on perceived quality, although I'm not sure if it would be in a positive or negative way.

Did you create any new variables from existing variables in the dataset?

I created a few different variables for this dataset. First, I created a variable called `total.acidity` by summing the volatile and fixed acidities. In order to further explore the relationship between free and total sulfur dioxide, I created a variable called `free.sulfur.dioxide.ratio` by finding the ratio of free sulfur dioxide to total sulfur dioxide. Similarly, I created `volatile.acidity.ratio` and `citric.acid.ratio` by calculating the ratio of volatile acidity and citric acid (respectively) to the total acidity mentioned above.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

Since the data was already in a very tidy format, no additional formatting was necessary for the univariate analysis portion. I did remove the `x` variable from the dataset, since that simply represented the row number and didn't provide any value to the analysis.

Volatile acidity had a bimodal distribution with peaks around 0.40 and 0.60 g/dm^3. I'm curious to see how that looks separated by quality ratings.

Citric acid also had a unique distribution; the most common value was 0.00 (132 wines), with 0.49 as the next most common value (68 wines). There were a few other values that occurred much more frequently than others, which was odd since they weren't close together at all.

The total sulfur dioxide and sulphate distributions both had some large outliers, which I suspect are related to one another.

Otherwise, most of the distributions of the variables were fairly normal, or at least not too unusual.

Bivariate Plots Section

Correlation Between Variables

```
round(cor(wine, use = 'everything'), digits = 3)
```

First, let's look at the correlation values for each pair of variables to get a feel for potential relationships in the data. There doesn't seem to be much strong correlation in the dataset, particularly relating to quality, which is a bit surprising to me. The strongest correlation value involving quality was with alcohol (0.476), followed by volatile acidity (-0.391), volatile acidity ratio (-0.347), sulphates (0.251), and citric acid (0.226). There could also be other chemical properties not measured here that could affect wine quality and taste, or the relationships could be more complex than what could be captured by Pearson's Correlation Coefficient. None of the new variables I created had strong correlations with quality, and all had weaker correlations with quality than at least one of their "parent" variables.

Since we're most concerned with how the chemical properties of wine affect perceived quality, I'll focus on relationships involving quality first. For each boxplot I made quality a factored variable.

Quality vs Alcohol

```
ggplot(aes(x = quality, y = alcohol), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter')
```

Alcohol had the strongest correlation with quality. This isn't too surprising to me since I'd imagine a higher alcohol content

would be related to a higher concentration of flavor. Lower concentrations of alcohol would likely have more of a "watery" mouthfeel and might not be perceived has being of a high quality.

This also matches up with my own experiences with beer (my brother is a professional brewmaster), where I've experienced more complex flavors in beers with higher alcohol contents. I've also experienced this with bourbon, where alcohol content impacts my own perceived quality in general, with most of my favorite bourbons having higher proofs in general than bourbons I didn't care for.

I suspect that part of this is because a beer/wine/spirit with a higher alcohol content would need to be fairly "balanced" in its flavors to mask or hide the alcohol a bit more. I've experienced some beers and bourbons that had a higher than normal alcohol content, but tasted very harsh and were unpleasant because the alcohol "burn" was overwhelming. I suspect the brewer/distiller/winemaker would also be able to identify that taste before it ever got to market (if they were truly being objective) and adjust their product accordingly, perhaps to water it down a bit more.

```
with(wine, by(alcohol, factor(quality), summary))
with(wine, by(alcohol, factor(quality), IQR))

ggplot(aes(x = factor(quality), y = alcohol), data = wine) +
  geom_boxplot()

ggplot(aes(x = factor(quality), y = alcohol), data = wine) +
  geom_boxplot() +
  ylim(min(wine$alcohol), quantile(wine$alcohol, 0.99))
```

I made quality a factored variable to create a boxplot. The limits were adjusted in the second boxplot to remove some of the outliers. Here the relationship between quality and alcohol is much easier to see. It's important to remember that there aren't as many wines rated highly, but there definitely appears to be a positive relationship between alcohol and quality. The lowest rated wines (with a quality of 3) all had alcohol values less than or equal to 11%, while roughly 75% of the highly rated (quality of 7 or 8) wines had alcohol values greater than 11% abv. Wines with a quality of 8 had the largest inter-quartile range (1.55 % by volume), while wines rated as 5 had quite a few outliers at high alcohol values, but also had the lowest inter-quartile range (0.8 % by volume). With the exception of wines rated as a 5, there is a clear positive relationship between alcohol and quality.

```
ggplot(aes(x = alcohol), data = wine) +
  geom_histogram(aes(fill = factor(quality)), binwidth = 0.1) +
  scale_color_brewer(type = 'div', name="Quality") +
  scale_fill_discrete(name="Quality")
```

Here's a stacked histogram of alcohol value colored by quality rating. Note that since the counts are stacked on top of one another (i.e., the maximum count value for each bin is the sum of ALL of the counts for each quality rating in that bin), it's the relative count of each rating within each bin that is important. Put another way, it shows how the quality ratings are distributed for each bin of the histogram created in the Univariate section. It's very apparent how many more wines are rated as 4 or 5 (82.4%) than all of the other ratings. The wine with the highest alcohol content (14.9 % by volume) had a rating of 5 (perhaps because it was "unbalanced"), while otherwise most of the wines with higher alcohol contents had ratings in the 6-8 range. On the other end of the spectrum, one of the two wines with the lowest alcohol contents (8.4 % by volume) was rated as a 3, while the other was rated as a 6. There definitely seems to be some other factors influencing quality, as I would suspect.

```
ggplot(aes(x = alcohol, color = factor(quality)), data = wine) +
  geom_density() +
  scale_color_discrete(name="Quality")
```

Finally, let's look at some density plots of the alcohol percentages broken out by quality. In general, the same trends we saw in the boxplots hold here. It's still a bit strange that the wines rated as 5 are the mostly skewed toward smaller alcohol percentages, but that explains the low median value from the boxplots.

Quality vs Volatile Acidity

```
with(wine, by(volatile.acidity, factor(quality), summary))

ggplot(aes(x = quality, y = volatile.acidity), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter')
```

Since volatile acidity had the second strongest correlation with quality, let's take a look at a scatter plot of volatile acidity vs. quality. I added jitter and transparency to prevent overplotting. It definitely looks like there is a negative correlation between the two, as indicated in the correlation for the two variables (-0.39). I think the boxplot will provide a better feel for the relationship since quality is really more of a factored variable.

```
ggplot(aes(x = factor(quality), y = volatile.acidity), data = wine) +
  geom_boxplot()

ggplot(aes(x = factor(quality), y = volatile.acidity), data = wine) +
  geom_boxplot() +
  ylim(min(wine$volatile.acidity), quantile(wine$volatile.acidity, 0.99))
```

Here the inverse relationship is much more apparent. There is a definite trend in lower volatile acidity levels as wine quality increases; the median volatile acidity level drops with each successive increase in quality rating, with the exception of 7 and 8, where the median stays the same. Since we know from the background info that high levels of volatile acidity can cause the wine to taste like vinegar, this inverse relationship between volatile acidity and quality makes sense.

```
ggplot(aes(x = volatile.acidity, color = factor(quality)), data = wine) +
  geom_density() +
  scale_color_discrete(name="Quality")
```

These density plots of volatile acidity by quality rating explain the bimodal distribution seen in the histogram of volatile acidity values. The lower rated wines (3-6) explain the peak in values around we saw around 0.60 g/dm^3, while the higher rated wines explain the peak in values around 0.40 g/dm^3. The wines rated in the middle of the scale (5-6), also had slight bimodal distributions themselves.

Quality vs Sulphates

```
with(wine, by(sulphates, factor(quality), summary))

ggplot(aes(x = quality, y = sulphates), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter')
```

The next strongest correlation value for quality was with sulphates concentrations, so lets take a look at those plotted together. Here I again added jitter and some transparency to prevent overplotting. There does appear to be a trend toward higher sulphate levels in higher rated wines. There is a large amount of variance in the sulphates values for wines rated as 5 or 6. Let's look at the boxplots for more insight.

```
ggplot(aes(x = factor(quality), y = sulphates), data = wine) +
  geom_boxplot()

ggplot(aes(x = factor(quality), y = sulphates), data = wine) +
  geom_boxplot() +
  ylim(min(wine$sulphates), quantile(wine$sulphates, 0.95))
```

Here the positive relationship between sulphates and quality is more apparent, but it is also clear there are a large number of outliers for the wines rated as 5 or 6. This likely drove down the correlation value. The median sulphate concentration increases with each quality rating (again, except for 7 and 8, where it remains the same). We know from the background info that sulphates act as an antimicrobial, so perhaps the microbes they are killing have an adverse affect on the perceived quality.

```
ggplot(aes(x = sulphates, color = factor(quality)), data = wine) +
  geom_density() +
  scale_color_discrete(name="Quality")
```

It's a bit harder to see the relationship between quality and sulphates in this density plot, but the peak densities mostly increase with rating, matching the trend we saw in the boxplots. The plots for ratings 3-5 and 7-8 almost completely overlap each other, respectively (not counting the outliers).

```
with(wine, by(citric.acid, factor(quality), summary))
```

```
ggplot(aes(x = quality, y = citric.acid), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter')
```

Finally, let's look at quality and citric acid plotted against each other. Wow, there is a lot a variance in these values, but I can see a slight positive trend, which coincides with the positive correlation value between the two (0.226).

```
ggplot(aes(x = factor(quality), y = citric.acid), data = wine) +
  geom_boxplot()

ggplot(aes(x = factor(quality), y = citric.acid), data = wine) +
  geom_boxplot() +
  ylim(min(wine$citric.acid), quantile(wine$citric.acid, 0.99))
```

The boxplot makes this positive relationship much clearer, with the median citric acid concentrations increasing steadily with each successive quality rating, from a median value of 0.0350 g/dm^3 for wines rated 3, up to a median value of 0.420 g/dm^3 for wines rated 8. The median values almost seem to be grouped together in rating pairs (3-4, 5-6, 7-8). I'm surprised by the overall variance across the values.

```
ggplot(aes(x = citric.acid, color = factor(quality)), data = wine) +
  geom_density() +
  scale_color_discrete(name="Quality")
```

There are some really unusual distributions in the plots. They are consistent with the observation earlier that certain values (0, 0.24, and 0.49) were more likely than others. This is especially apparent for wines rated 5 and 6, where there are distinct local maxima in the density plots around those values. Despite the unusual peaks, it is clear that higher rated wines tended to have larger citric concentrations than lower rated wines.

Other Relationships: Acidty and pH

Next let's take a look at some of the other relationships in the dataset. First, lets look at fixed acidity vs pH, which (unsurprisingly) had one of the the strongest overall correlations in the dataset (-0.683).

```
ggplot(aes(x = fixed.acidity, y = pH), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter')

ggplot(aes(x = fixed.acidity, y = pH), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter') +
  coord_trans(x = 'log10') +
  stat_smooth(method = 'lm', formula = y ~ log10(x))
```

Going back to my high school chemistry days, I remember that acidity and pH are related. The background info for the dataset confirms this, stating that pH "describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic)." According to Wolfram, the relationship is described by pH = -log10[H+], "where [H+] represents the concentration of hydrogen ions in units of moles per liter of volume". Now that we know the nature of the relationship, lets try a log10 transformation on the fixed acidity concentration. After the transformation, the relationship is much clearer. I also added a smoother (using a line of the form `y ~ log10(x)`), so it's much easier to see that the relationship between fixed acidity and pH aligns with our knowledge of chemistry.

Acidity Relationships

```
ggplot(aes(x = fixed.acidity, y = volatile.acidity), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter')
```

I'm curious to see what the relationships between the different types of acidities look like. Are wines with higher fixed acidity more likely to have higher volatile acidity as well? How does citric acid relate to either?

There doesn't appear to be a very strong relationship between fixed acidity and volatile acidity, which is in line with its relatively weak correlation (-0.256). Perhaps this is because the volatile acidity concentrations are so much smaller than the fixed acidity values (likely because high levels of volatile acidity can make the wine taste like vinegar). I added some jitter and transparency again here to prevent overplotting.

```
ggplot(aes(x = fixed.acidity, y = citric.acid), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter') +
  stat_smooth(method = 'lm')
```

Citric acid and fixed acidity had one of the stronger correlations in the dataset (0.672), which is apparent here. There is definitely a positive relationship between the two, as expected (since citric acid presumably contributes to the fixed acidity concentration).

```
ggplot(aes(x = fixed.acidity, y = density), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter') +
  stat_smooth(method = 'lm')
```

Density and fixed acidity also had a fairly strong correlation (0.685), so let's see what they look like plotted against each other. It is clear there is a positive relationship between the two, perhaps because the acids present in the wine have a density greater than water. According to Wolfram|Alpha, acetic acid (or volatile acidity in this dataset) has a density of 1.049 g/cm^3, while citric acid has a density of 1.665 g/cm^3, so at least two of the known acids present in red wine have densities greater than water (1.000 g/cm^3).

Sulfur Dioxide

```
ggplot(aes(x = total.sulfur.dioxide, y = free.sulfur.dioxide), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter') +
  stat_smooth(method = "lm")
```

Finally, let's look at the relationship between free and total sulfur dioxide. They also had a relatively strong correlation (0.668), as expected (since free sulfur dioxide is a part of the total sulfur dioxide concentration).

```
ggplot(aes(x = total.sulfur.dioxide, y = sulphates), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter') +
  stat_smooth(method = "lm")
```

Surprisingly, there doesn't appear to be much of the same relationship (or any relationship at all) when it comes to total sulfur dioxide and sulphates (which also contribute to the total sulfur dioxide concentration). The correlation score between them was a very weak 0.043.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

The main feature of interest in the dataset, quality, had relatively strong relationships (based on correlation scores) with four of the features: alcohol, density, volatile acidity, and sulphates. The strongest correlation value involving quality was with alcohol (0.476), followed by volatile acidity (-0.391), volatile acidity ratio (-0.347), sulphates (0.251), and citric acid (0.226).

Alcohol and quality had the strongest correlation score, and there was a clear positive relationship between the two in the boxplots. Other than a slight dip for wines rated as a 5, the median alcohol values steadily increased with each rating. I suspect that higher alcohol wines have more concentrated flavor in general.

Volatile acidity had an inverse relationship with quality, and variance decreased with each increase in rating. Since high levels of volatile acidity can lead to vinegar-like flavors, this decrease in median values and variance as ratings increase isn't surprising.

Like alcohol, sulphates had a positive relationship with quality. The variance in sulphates concentrations increased with rating in general, as did the median values. The wines rated as 5 or 6 had a lot of outliers. Sulphates act as an anitmicrobial and antioxidant, so a positive relationship between quality and sulphates makes sense (assuming that microbials can lead to undesired flavors).

Lastly, citric acid also had a positive relationship with quality. Variance decreased in general as ratings increased, but the variance was surprisingly large across all of the ratings. However, there was a significant increase in median values as ratings increased (by more than a factor of 10 between ratings of 3 and 8). There also weren't very many outliers.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

I noticed a few different interesting relationships involving fixed acidity and some of the other variables. Fixed acidity and pH had one of the strongest relationships in the dataset (a correlation of -0.673), which was unsurprising because pH is a way to measure how acidic or basic something is. Some quick research revealed that pH is given is calculated via the expression pH = -log10[H+], where [H+] is the concentration of hydrogen ions. Transforming the fixed acidity values via a log10 transform yielded a negative linear relationship, as expected.

Fixed acidity and volatile acidity weren't strongly correlated (-0.256), but fixed acidity and citric acid were (0.671), likely because citric acid contributes to the overall fixed acidity concentrations. Density and fixed acidity were also strongly correlated (0.668), probably due to the fact that some of the acids contributing to the fixed acidity levels (including citric acid) have a larger density than water.

Finally, total sulfur dioxide and sulphates weren't as strongly correltated (0.043) as I expected them to be, considering that sulphates contribute to the total sulfur dioxide concentration.

What was the strongest relationship you found?

Ignoring the (uninteresting) strong relationships involving the new variables and the original variables used to create them, the strongest relationship was between total acidity and density (0.685), followed by pH and fixed acidity (-0.683), pH and total acidity (-0.673), fixed acidity and citric acid (0.671), density and fixed acidity (0.668), and free sulfur dioxide and total sulfur dioxide (0.667). Unsurprisingly, the strongest overall relationship was between fixed acidity and total acidity, with a correlation of 0.995 between them.

Multivariate Plots Section

Quality and Acidity

Total Acidity vs Quality

```
with(wine, by(total.acidity, factor(quality), summary))
with(wine, by(total.acidity, factor(quality), IQR))

ggplot(aes(x = factor(quality), y = total.acidity), data = wine) +
  geom_boxplot()

ggplot(aes(x = factor(quality), y = total.acidity), data = wine) +
  geom_boxplot() +
  ylim(min(wine$total.acidity), quantile(wine$total.acidity, 0.98))
```

Now let's examine some of the acidity relationships a bit deeper. Total acidity and quality were very weakly correlated overall (0.086), and that's evident in the boxplots. With the exception of the lowest rated wines, the inter-quartile ranges increased with rating. I suspect the weak correlation is due in part to the opposite effects citric acid and acetic acid have on quality, keeping the overall acidity values roughly level.

```
with(wine, by(volatile.acidity.ratio, factor(quality), summary))

ggplot(aes(x = factor(quality), y = volatile.acidity.ratio), data = wine) +
  geom_boxplot()
```

The volatile acidity ratios mostly follow the trends seen in the volatile acidity values earlier, which makes sense since there weren't any strong trends between total acidity and quality. There was a slight increase in median ratios (0.039 to 0.045) between wines rated 7 and 8, as seen in the total acidity values. Most wines had volatile acidity ratios less than 0.100 in general, while the highest rated wines (6-8) had a majority of their ratios below 0.075.

```
with(wine, by(citric.acid.ratio, factor(quality), summary))

ggplot(aes(x = factor(quality), y = citric.acid.ratio), data = wine) +
  geom_boxplot()

ggplot(aes(x = factor(quality), y = citric.acid.ratio), data = wine) +
  geom_boxplot() +
  ylim(min(wine$citric.acid.ratio), quantile(wine$citric.acid.ratio, 0.99))
```

The citric acid ratios are much smaller than the volatile acidity ratios, with most of the ratios falling below 0.05. Otherwise, this looks very similar to the citric acid boxplot seen earlier, except with more outliers.

Total Acidity vs pH

```
ggplot(aes(x = total.acidity, y = pH), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter') +
  coord_trans(x = 'log10') +
  stat_smooth(method = 'lm', formula = y ~ log10(x))

theme_set(theme_minimal(15))

p1 <- ggplot(aes(x = fixed.acidity, y = pH), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter') +
  coord_trans(x = 'log10') +
  stat_smooth(method = 'lm', formula = y ~ log10(x))

p2 <- ggplot(aes(x = total.acidity, y = pH), data = wine) +
  geom_point(alpha = 1/2, position = 'jitter') +
  coord_trans(x = 'log10') +
  stat_smooth(method = 'lm', formula = y ~ log10(x))

grid.arrange(p1, p2, nrow = 1)

theme_set(theme_minimal(18))

summary(lm(wine$pH ~ log10(wine$fixed.acidity)))
summary(lm(wine$pH ~ log10(wine$total.acidity)))
```

Now we can plot the total acidity against pH as we did before with fixed acidity. I wonder how they look side-by-side. Unsurprisingly, they are very similar, since the same physical relationship between acidity and pH holds. Additionally, fixed and total acidity only differ by the volatile acidity concentrations, which are very small compared to both. The R-squared values for each of the linear regression models were very close as well; `fixed.acidity` had a slightly better fit, with an R-squared value of 0.4989, while `total.acidity' had an R-squared value of 0.4787. However, the fact they are so much less than 1.0 suggests there are other relationships not captured by the models.

```
ggplot(aes(x = volatile.acidity, y = alcohol, color = factor(quality)),
       data = wine) +
  geom_point(position = 'jitter', size = 2) +
  scale_color_brewer(type = 'div',
    guide = guide_legend(title = 'Quality'))
```

Now let's look at the two variables with the strongest correlations with quality plotted against each other and colored by quality. I used a color scheme which accentuates the lowest and highest rated wines to help clarify the relationships present. While there are some exceptions, it is easy to see two main regions: the lowest quality wines tended to have lower alcohol percentages and higher volatile acidity concentrations, while the higher quality wines had higher alcohol percentages and lower volatile acidity concentrations, in general.

```
ggplot(aes(x = volatile.acidity, y = citric.acid, color = factor(quality)),
       data = wine) +
  geom_point(position = 'jitter', size = 2) +
  scale_color_brewer(type = 'div',
    guide = guide_legend(title = 'Quality'))
```

Finally, we can create a similar plot to examine volatile acidity and citric acid colored by quality. Here there isn't as quite as clear of a delineation between the low and high rated wines, but it does look somewhat similar to the previous plot because

citric acid and alcohol both have a positive correlation with quality. The highest rated wines tended to have higher citric acid concentrations and low volatile acidity concentrations, and the lower rated wines tended to have lower citric acid concentrations and higher volatile acidity concentrations.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

Most of the relationships from this part of the analysis were consistent with what was seen in the earlier sections. There were significant differences in the distributions of volatile acidity concentrations by quality rating, as seen in the density plot. The "average" (5-6) rated wines also had bimodal distributions, which helped explain the overall bimodal distribution we saw earlier.

The citric acid ratio, volatile acidity ratio, and total acidity variables I created didn't add much value to determining which chemical characterics influence wine quality. This was mainly because total acidity didn't have a strong correlation with quality. Total acidity did have the expected relationship with pH, however.

Looking at alcohol plotted against volatile acidity and colored by quality rating helped to visualize the strongest relationships involving quality, even though alcohol and volatile acidity were weakly correlated (-0.202) themselves. The highest quality wines tended to have high alcohol percentages and low volatile acidity concentrations.

Were there any interesting or surprising interactions between features?

Citric acid had some really interesting density plots when colored by quality rating. There were certain values (mainly 0.00, 0.24, and 0.49) which appeared much more frequently than other values, and it's not clear why. There were distinct peaks in the density plots at each of those values, especially for the wines rated 5 and 6. Perhaps it could be related to some sort of measurement or rounding error (since they roughly occur at multiples of 0.25).

OPTIONAL: Did you create any models with your dataset?

I did not, mainly because none of the relationships seemed strong enough to justify creating a model other than for the sake of creating one.

Final Plots and Summary

Plot One

```
# Change theme presets to defaults to accomodate plot text better
theme_set(theme_minimal())

ggplot(aes(x = citric.acid, color = factor(quality)), data = wine) +
  geom_density() +
  scale_color_discrete(name="Quality") +
  ggtitle("Density of Citric Acid Concentration by Quality Rating in Wine") +
  xlab("Citric Acid (g/dm^3)") +
  ylab("Density")
```

Description One

Citric acid concentration and quality rating had a positive relationship, with higher rated red wines tending to have higher citric acid concentrations. This is likely due to the fact that citric acid is known to add "freshness" and flavor to wines, which are both desirable. Many wines (8.26%) had no measurable citric acid at all; however, all of the highest rated wines had a citric acid concentration of at least 0.03 g/dm^3. There were also a few values which appeared much more frequently than others (at 0.02, 0.24, and 0.49 g/dm^3), especially among the "average" rated wines (quality ratings of 4-5), and it's not clear why.

This could be due to a measurement or rounding error.

Plot Two

```
ggplot(aes(x = factor(quality), y = alcohol), data = wine) +
  geom_boxplot() +
  guides(fill=FALSE) +
  ggtitle("Wine ABV by Quality") +
  ylab("Alcohol (% by volume)") +
  xlab("Wine Quality Rating")
```

Description Two

Alcohol had the strongest correlation with red wine quality (0.476) among all of the chemical properties measured. The lowest rated wines (with a quality of 3) all had alcohol values less than or equal to 11%, while roughly 75% of the highly rated (quality of 7 or 8) wines had alcohol values greater than 11% abv. With the exception of wines rated as a 5, there was a clear positive relationship between alcohol and quality. This makes sense since I'd expect a higher alcohol content would be related to a higher concentration of flavor. Lower concentrations of alcohol would likely have more of a "watery" mouthfeel in comparison and might not be perceived has being of a high quality.

Plot Three

```
ggplot(aes(x = volatile.acidity, y = alcohol, color = factor(quality)),
       data = wine) +
  geom_point(position = 'jitter', size = 2) +
  scale_color_brewer(type = 'div',
    guide = guide_legend(title = 'Quality')) +
  ggtitle("Wine ABV and Volatile Acidity by Quality") +
  ylab("Alcohol (% by volume)") +
  xlab("Volatile Acidity (g/dm^3)")
```

Description Three

Alcohol by volume and volatile acidity were the two chemical properties most closely related to quality in red wine. Alcohol had a positive relationship with quality, perhaps due to a higher concentration of flavor in wines with higher alcohol percentages. Volatile acidity had a negative relationship with quality rating, due to the fact that higher concentrations can lead to undesirable vinegar-like flavors. As evidenced by the two distinct regions in the plot, the lowest quality wines tended to have lower alcohol percentages and higher volatile acidity concentrations, while the higher quality wines had higher alcohol percentages and lower volatile acidity concentrations, in general.

Reflection

The red wine data set includes chemical property information and blind taste-test quality ratings for 1,599 red wines from Portugal. My goal was to determine which chemical properties had the strongest effect on perceived red wine quality. I started by examining each of the 14 variables in the dataset to look for any interesting distributions and to get a feel for the ranges of values. I also created a few new variables by taking ratios or sums of a few select variables, which I later found didn't add much value. Then, I calculated the correlation coefficients for each combination of variables in order to determine the strengths of the relationships between the variables, particularly those involving quality.

Alcohol, volatile acidity, sulphates, and citric acid had the strongest correlations with quality. I was surprised that pH, density, and residual sugar didn't have a big impact on quality. I also noticed a familiar relationship between pH and fixed and total acidity based on the way that pH is actually measured. Finally, I used density plots and scatter plots colored by quality rating to better understand the multivariate relationships between the chemical properties and quality. Overall, none of the relationships with quality were particularly strong, and didn't suggest that a simple model would be useful in this case.

The new variables I created didn't really add value, and I was perplexed about a few of the citric acid values which appeared more often than others. The biggest difficulty was handling the complexity of the dataset. It was hard to keep track of all of the different relationships at play, and to determine where to focus next. Because there were so many different potential directions to go with the analysis, it was hard for me to balance my desire to be thorough and consistent with trying to focus

only on the important informaton. It was a good taste of the difficulties of dealing with complex datasets. The boxplots and density plots helped the most to visualize the relationships within the data. I definitely gained a new appreciation for them.

There are a number of different ways to expand and improve this analysis. The dataset could be expanded to include more wines of this category, especially among the higher and lower rated wines (since they had relatively small populations). I would have also been interested to see how price factored in as well, even if the testers didn't know it ahead of time. I also think the dataset could be expanded to include other varieties of red wine from other regions to see how those results compare or contrast. Lastly, a machine learning algorithm woulc be interesting to run on the dataset, partiluarly to help predict quality rating.