# An Analysis of Wine by Travis Seal

**Intro:** What you will see below is an analysis of wine. This analysis is a free associative exploration of a great beverage.

**Our Data :** The dataset that we are going to explore are data about the natural characteristics of wine. There are 13 total variables displayed below. The ones of particular interest are those relating to pH, and acidity levels: fixed.acidity volatile.acidity citric.acid, and pH. There are 1599 observations.

# Univariate Plots Section

**Summary of entire dataset:**

## X fixed.acidity volatile.acidity citric.acid   
## Min. : 1.0 Min. : 4.60 Min. :0.1200 Min. :0.000   
## 1st Qu.: 400.5 1st Qu.: 7.10 1st Qu.:0.3900 1st Qu.:0.090   
## Median : 800.0 Median : 7.90 Median :0.5200 Median :0.260   
## Mean : 800.0 Mean : 8.32 Mean :0.5278 Mean :0.271   
## 3rd Qu.:1199.5 3rd Qu.: 9.20 3rd Qu.:0.6400 3rd Qu.:0.420   
## Max. :1599.0 Max. :15.90 Max. :1.5800 Max. :1.000   
## residual.sugar chlorides free.sulfur.dioxide  
## Min. : 0.900 Min. :0.01200 Min. : 1.00   
## 1st Qu.: 1.900 1st Qu.:0.07000 1st Qu.: 7.00   
## Median : 2.200 Median :0.07900 Median :14.00   
## Mean : 2.539 Mean :0.08747 Mean :15.87   
## 3rd Qu.: 2.600 3rd Qu.:0.09000 3rd Qu.:21.00   
## Max. :15.500 Max. :0.61100 Max. :72.00   
## total.sulfur.dioxide density pH sulphates   
## Min. : 6.00 Min. :0.9901 Min. :2.740 Min. :0.3300   
## 1st Qu.: 22.00 1st Qu.:0.9956 1st Qu.:3.210 1st Qu.:0.5500   
## Median : 38.00 Median :0.9968 Median :3.310 Median :0.6200   
## Mean : 46.47 Mean :0.9967 Mean :3.311 Mean :0.6581   
## 3rd Qu.: 62.00 3rd Qu.:0.9978 3rd Qu.:3.400 3rd Qu.:0.7300   
## Max. :289.00 Max. :1.0037 Max. :4.010 Max. :2.0000   
## alcohol quality   
## Min. : 8.40 Min. :3.000   
## 1st Qu.: 9.50 1st Qu.:5.000   
## Median :10.20 Median :6.000   
## Mean :10.42 Mean :5.636   
## 3rd Qu.:11.10 3rd Qu.:6.000   
## Max. :14.90 Max. :8.000

**Summary of Volatile Acidity:** Lets trim down the data by focusing on a few variables. Below we will explore the alcoholic content of wine:

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1200 0.3900 0.5200 0.5278 0.6400 1.5800

**Summary of fixed acidity**

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.60 7.10 7.90 8.32 9.20 15.90

**Summary of pH**

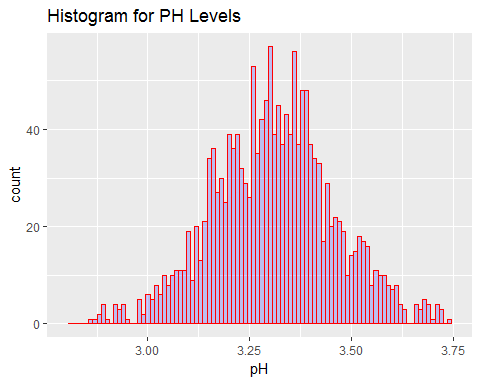
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.740 3.210 3.310 3.311 3.400 4.010

# Univariate Analysis

**Variable Distribution**

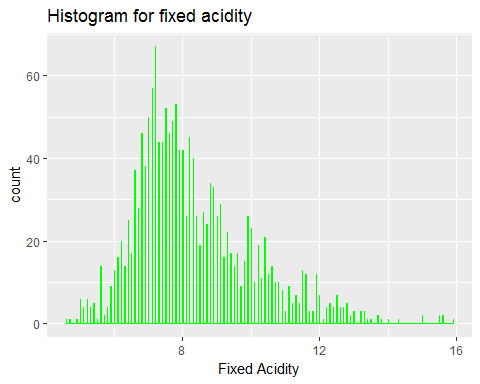
We are going to take a look at the distribution of the 'pH' variable. Basically I am wondering about the frequency of 'pH' in wine for the given dataset.

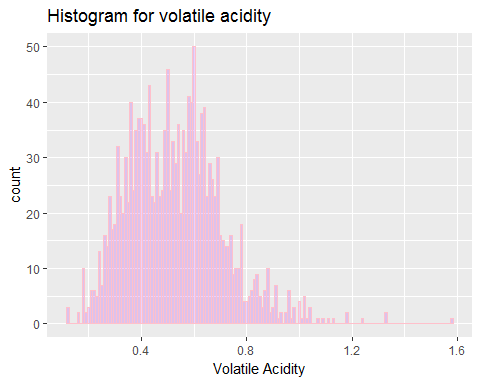
**Observations:** The concentration of acidity is between 3 and just below 3.5. I trimmed the tail to increase our focus on the majority of the data.

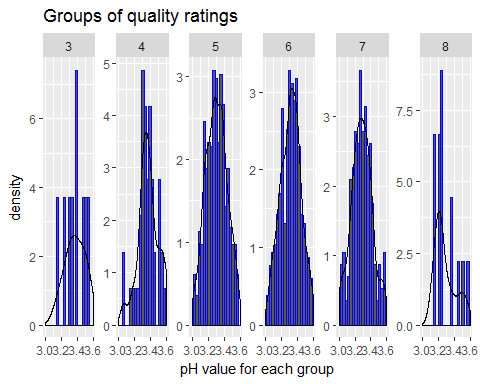


Here I trimmed the x axis so that we can focus on the majority of the data.

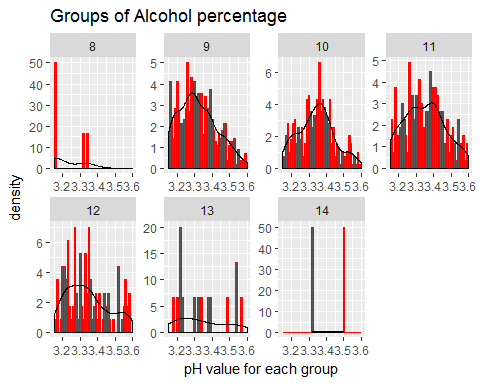
**Observations:** Its pretty clear that, even though there it is not a linear relationship, there is a general trend of increase density as pH level increases. This climaxes at ~3.35, then starts to descend at a rate that is **really** close the the rate of ascension.







*observation* This graph shows that 5 of the 6 graphs are normally distrobuted. Graph 8, or those with the highest quality rating lean to the left, indicating a tendancy to like wine with a greater level of acidity



**Observations** This graph shows the pH for groupings of concentration of alcohol content. It is interesting to note that in group with 8% content have most concentrated observations at a pH value of ~3, while the group with 9% have a normal (symmetric) distribution. Because alcohol can act as a base or an acid, it will impossible to determine using the existing data if the alcohol content is the reason for the shifts in distribution. Because graphs 9-11 are similar in shape, and 8,14 are *very* different, I think the changes may be due to the sample data.

### What is the structure of our dataset?

**Structure of my data** In our adventure so far, we have only seen one variable at a time. Thus the structure of the data is a 1 dimensional histogram that shows the frequency of occurrence and spread of the variable. Our chosen variables have a regular distribution *(normal distribution)*

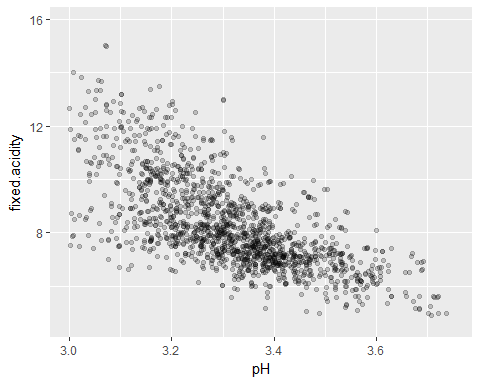
**The main features** of interest in our dataset are the type and category of acids as they relate to wine,quality, and its pH level.

### Summary:

**Imperfections** Our data is not perfect, and we can shape our data to make it appear nice to work with by trimming outliers, or changing the bin size. When I grouped the data by alcohol percentage, only half of my graphs turned out 'normal' looking.

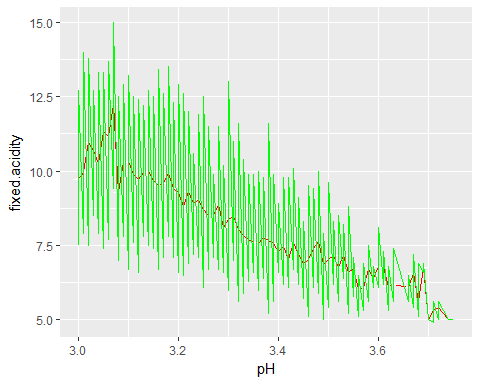
# Bivariate Plots Section

**Exploring Relationships**: Acidity is just acidity, right? Turns out there's more to it. Below is a scatter plot showing the basic relationship between 'fixed acidity' and its pH value.



**Summary** This scatterplot shows us an emerging trend among *all* the data points. It will not show you the *overall relationship* between the two variables.

df.fix\_acid\_by\_ph <- group\_by(df, pH)  
 df.fc\_by\_ph <- summarise(df.fix\_acid\_by\_ph,  
 fixCountMean = mean(fixed.acidity),  
 fixCountMedian = median(fixed.acidity),  
 n = n())  
  
df.fix\_acid\_by\_ph <- arrange(df.fix\_acid\_by\_ph,pH)  
  
  
ggplot(aes(x = pH, y = fixed.acidity),data = df) +  
 xlim(3,3.75)+  
 geom\_line(stat = 'summary', fun.y = mean,color ='red')+  
 geom\_hex(stat = 'summary', fun.x = quantile, probs = .1, color = 'white')+  
 geom\_line(stat = 'summary', fun.y = quantile, probs = .9, color = 'green')



**Observations:** I broke the pH over quality. Graphs 4-7 are similar in distribution, and 3 and 8 are dissimilar.

# Bivariate Analysis

**Tip**: As before, summarize what you found in your bivariate explorations here. Use the questions below to guide your discussion.

### 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?

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

### What was the strongest relationship you found?

# Multivariate Plots Section

**Tip**: Now it's time to put everything together. Based on what you found in the bivariate plots section, create a few multivariate plots to investigate more complex interactions between variables. Make sure that the plots that you create here are justified by the plots you explored in the previous section. If you plan on creating any mathematical models, this is the section where you will do that.

# 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?

### Were there any interesting or surprising interactions between features?

### OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

# Final Plots and Summary

**Tip**: You've done a lot of exploration and have built up an understanding of the structure of and relationships between the variables in your dataset. Here, you will select three plots from all of your previous exploration to present here as a summary of some of your most interesting findings. Make sure that you have refined your selected plots for good titling, axis labels (with units), and good aesthetic choices (e.g. color, transparency). After each plot, make sure you justify why you chose each plot by describing what it shows.

### Plot One

### Description One

### Plot Two

### Description Two

### Plot Three

### Description Three

# Reflection

**Tip**: Here's the final step! Reflect on the exploration you performed and the insights you found. What were some of the struggles that you went through? What went well? What was surprising? Make sure you include an insight into future work that could be done with the dataset.

**Tip**: Don't forget to remove this, and the other **Tip** sections before saving your final work and knitting the final report!