Exploration of Red Wine Quality by Jeremy Yenko

```
rw <- read.csv('wineQualityReds.csv')
```

This analysis will focus on the many chemical properties that have the potential to influence the quality of red wines.

Preliminary exploration

Variables in our dataset

```
str(rw)
```

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.
5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.06
5 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.3
5 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57
0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

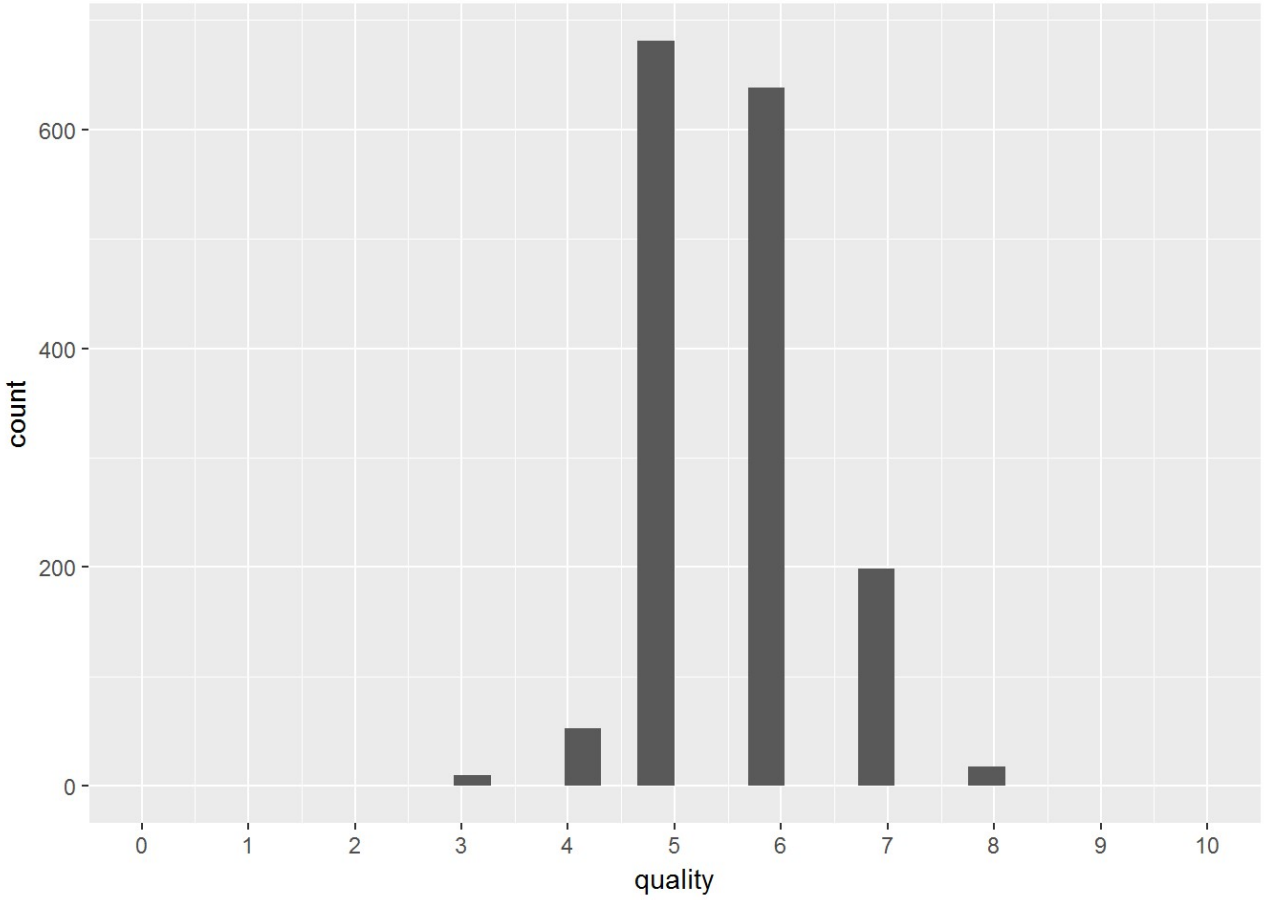
Above, we can see a number of chemical factors present that can influence the quality of red wine.

At the bottom of this list lies our dependent variable: quality.

Quality is measured on a scale of 0-10 with ratings being provided by a minimum of three wine experts for each wine.

Overview of wine quality

```
qplot(data = rw, x = quality) +
  scale_x_continuous(limits = c(0, 10), breaks = seq(0, 10, 1))
```



As seen above, the quality ratings of the wines in our study appear to have a normal distribution.

Alcohol content and wine quality

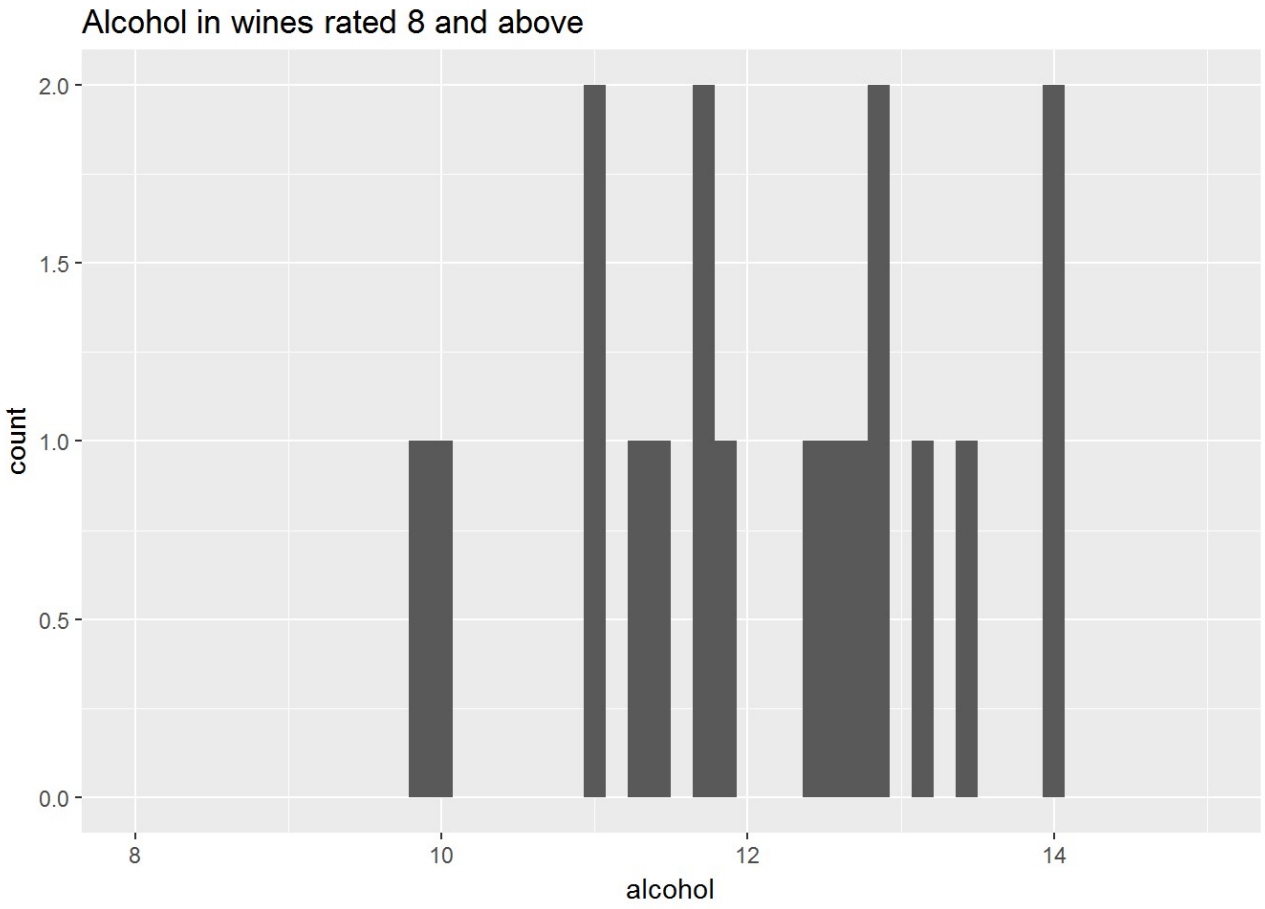
> One of the first aspects of red wine we will want to examine is alcohol content. Below is a summary of the alcohol content of the wines in our dataset:

```
summary(rw$alcohol)
```

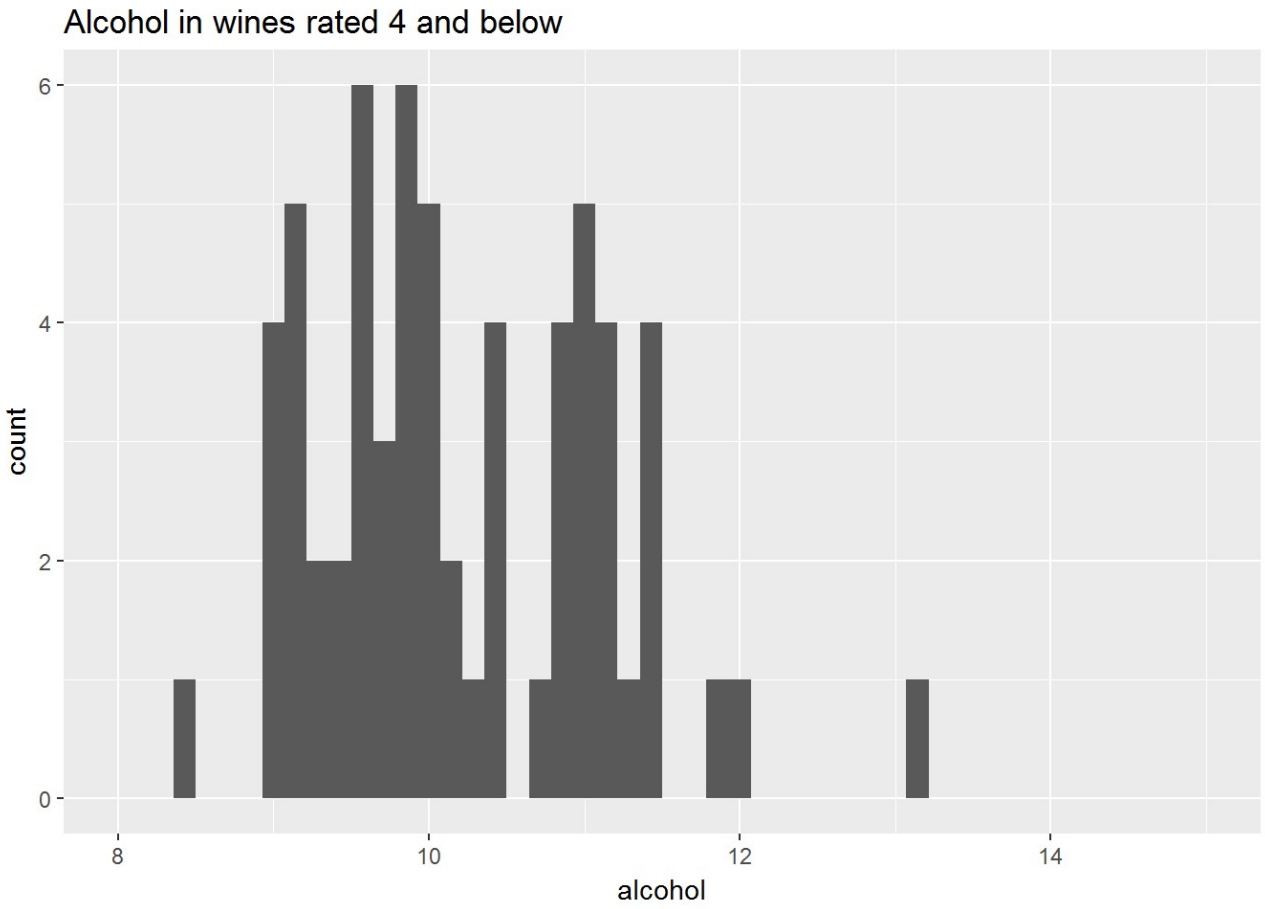
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

As seen above, our wines have a mean alcohol content of 10.42. Since the mean is close to the median, this suggests we do not have many outliers and that the mean is generally a reliable measure for the alcohol content variable.

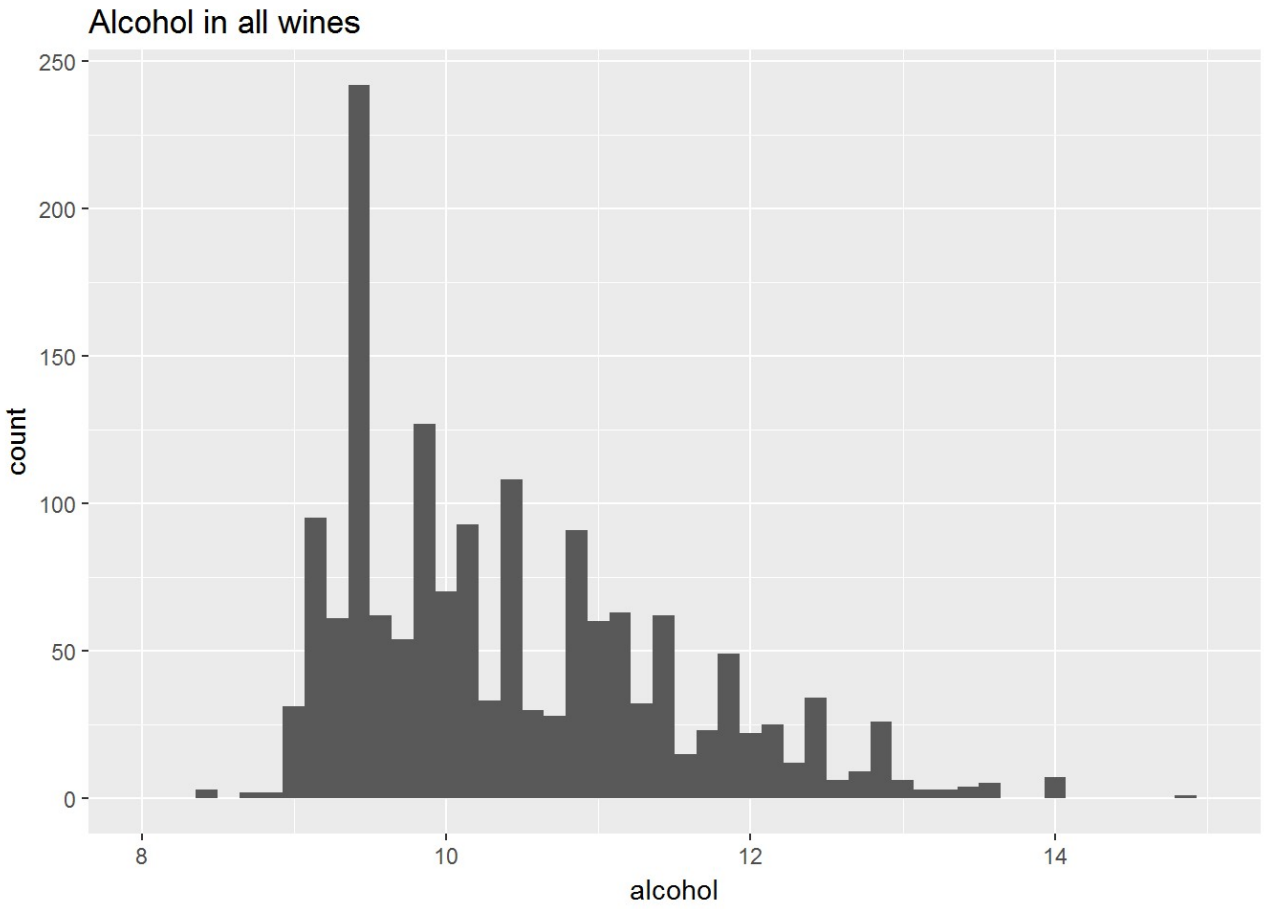
```
rw_hq <- subset(rw, quality >= 8)
qplot(data = rw_hq, x=alcohol, bins=50, main="Alcohol in wines rated 8 and abov
e") +
  xlim(8,15)
```




```
rw_lq <- subset(rw, quality <= 4)
qplot(data=rw_lq, x=alcohol, bins=50, main="Alcohol in wines rated 4 and belo
w") +
  xlim(8,15)
```




```
qplot(data = rw, x=alcohol, bins=50, main="Alcohol in all wines") +
  xlim(8,15)
```



These two plots show that the higher quality wines tend to have a greater alcohol content than the lower quality wines. For all wines, the distribution appears positively skewed.

Now we will examine the correlation between alcohol and wine quality:

```
cor.test(rw$alcohol, rw$quality)
```

```
##
##  Pearson's product-moment correlation
##
## data:  rw$alcohol and rw$quality
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

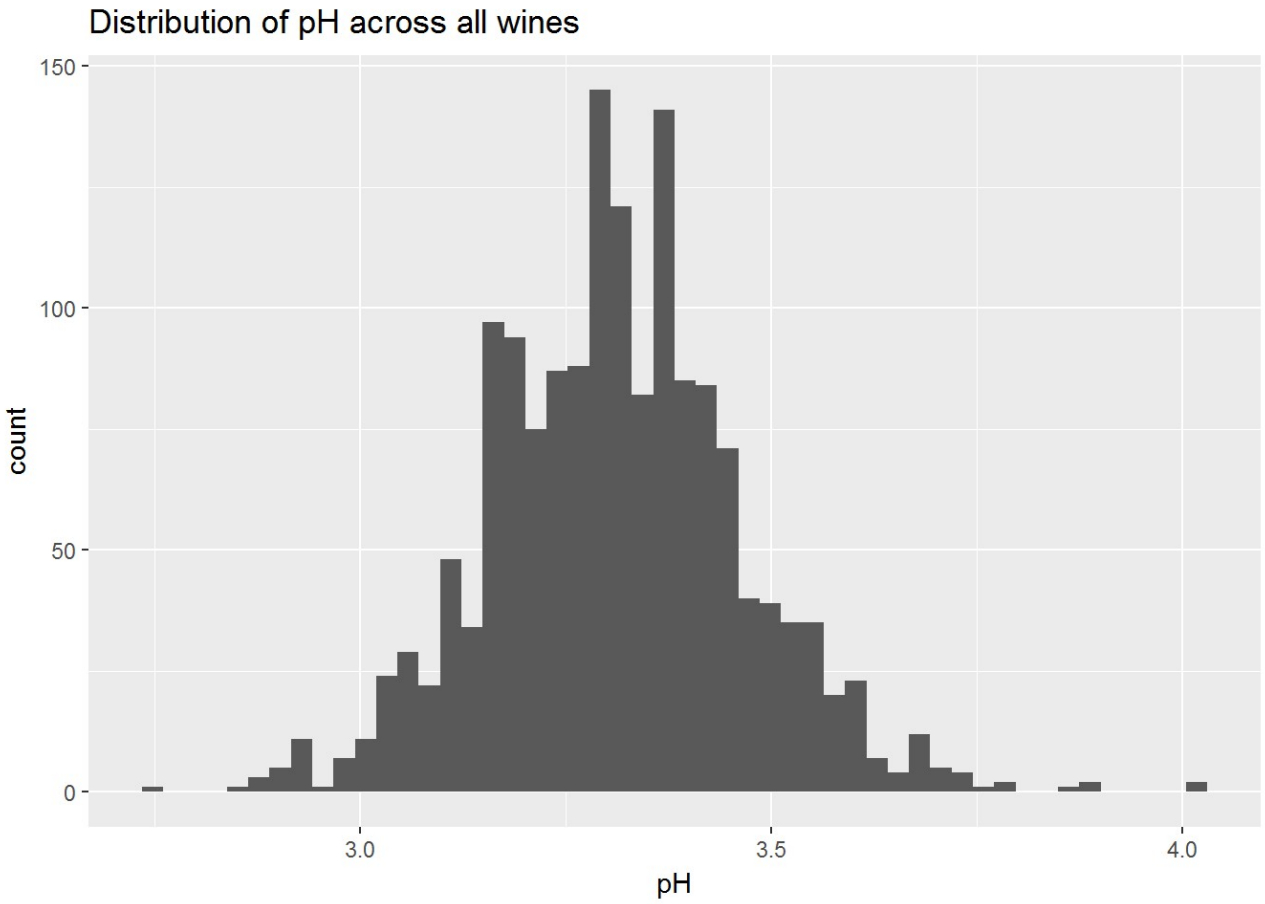
We can see that there is a positive correlation between alcohol content and wine quality. Although this correlation is not notably strong, the low p-value suggests it is a reliable measure.

pH and wine quality

> Another factor that can influence wine quality is pH. We will investigate whether or not pH influences wine quality.

First, we will look at the distribution of pH across all of the wines in our dataset.

```
qplot(data = rw, x=pH, bins=50, main="Distribution of pH across all wines")
```



The pH values of our wines appear to follow a normal distribution.

```
summary(rw$pH)
```

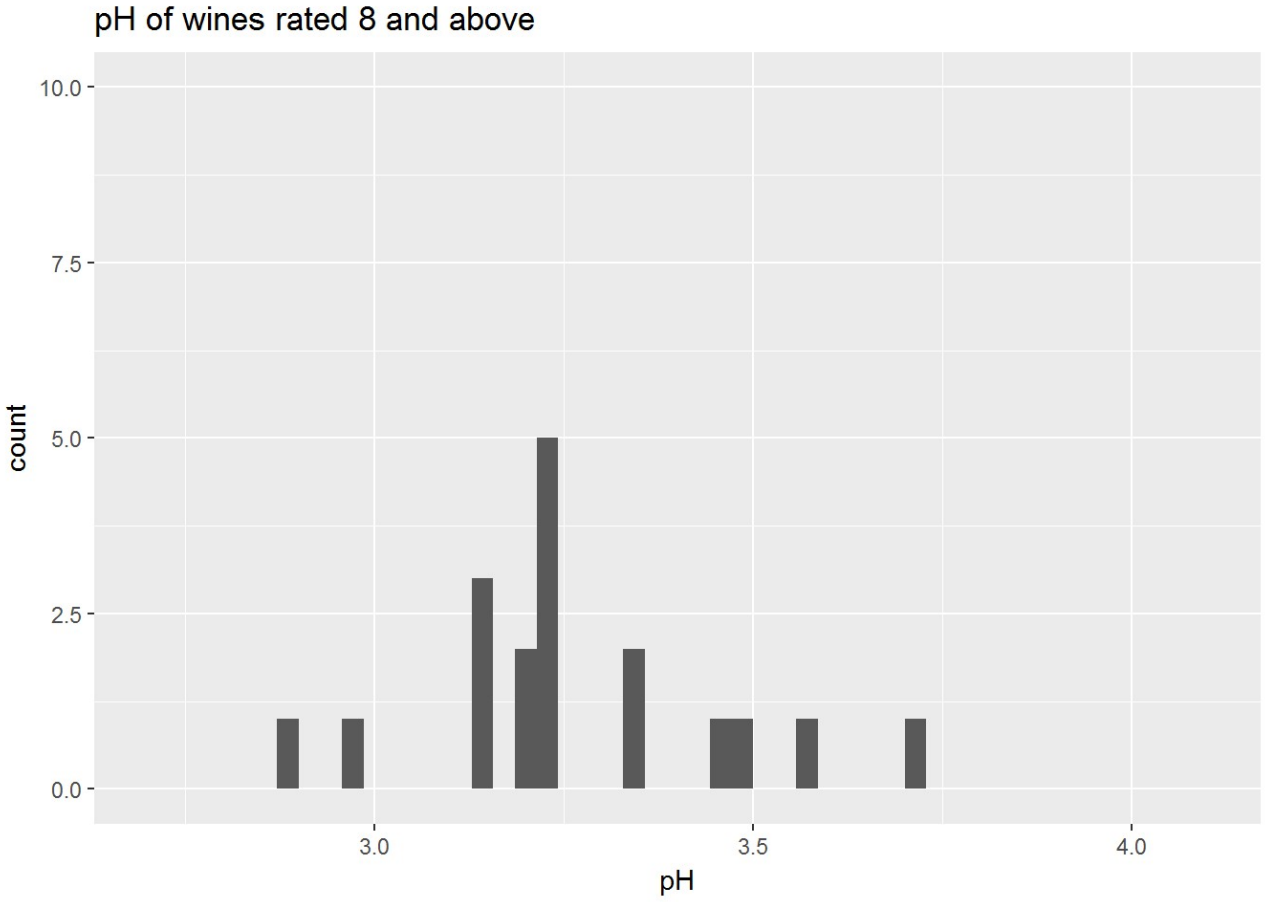
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

All of our wines have a pH well below 7. The lowest pH is 2.740 and the highest pH is 4.010. The mean and median are close, indicating the mean is generally a reliable measure for the pH of our wines.

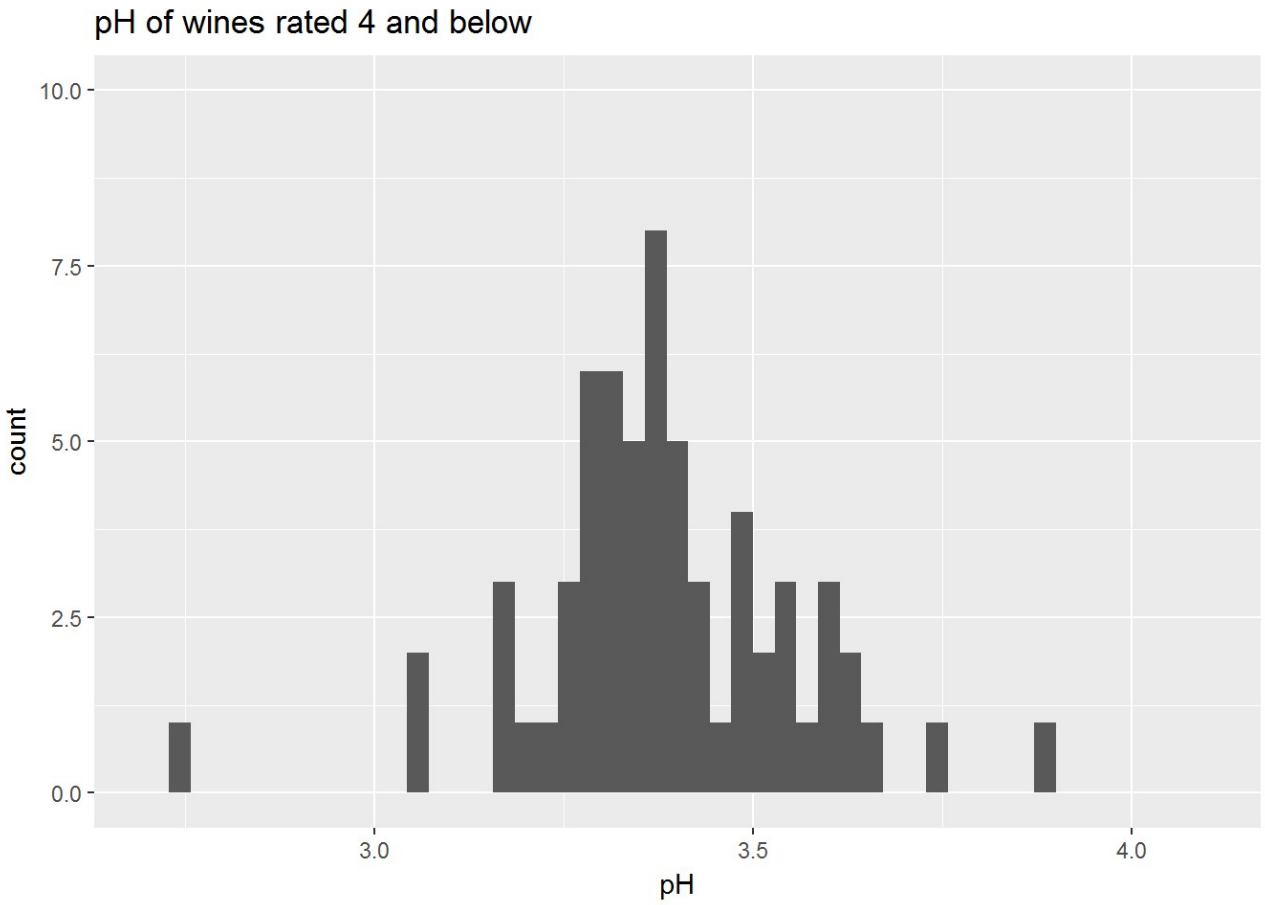
We will now investigate the relationship between pH and quality.

```
rw_pH_hq <- subset(rw, quality>=8)
rw_pH_lq <- subset(rw, quality<=4)

qplot(data=rw_pH_hq, x=pH, bins=50, main="pH of wines rated 8 and above") +
  xlim(2.7,4.1) +
  ylim(0,10)
```




```
qplot(data=rw_pH_lq, x=pH, bins=50, main="pH of wines rated 4 and below") +
  xlim(2.7,4.1) +
  ylim(0,10)
```



From these plots, it appears that higher quality wines tend to have a lower pH, while lower quality wines appear to have a higher pH.

Now we will examine the correlation between pH and wine quality.

```
cor.test(rw$pH, rw$quality)
```

```
##
##  Pearson's product-moment correlation
##
## data:  rw$pH and rw$quality
## t = -2.3109, df = 1597, p-value = 0.02096
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.106451268 -0.008734972
## sample estimates:
##         cor
## -0.05773139
```

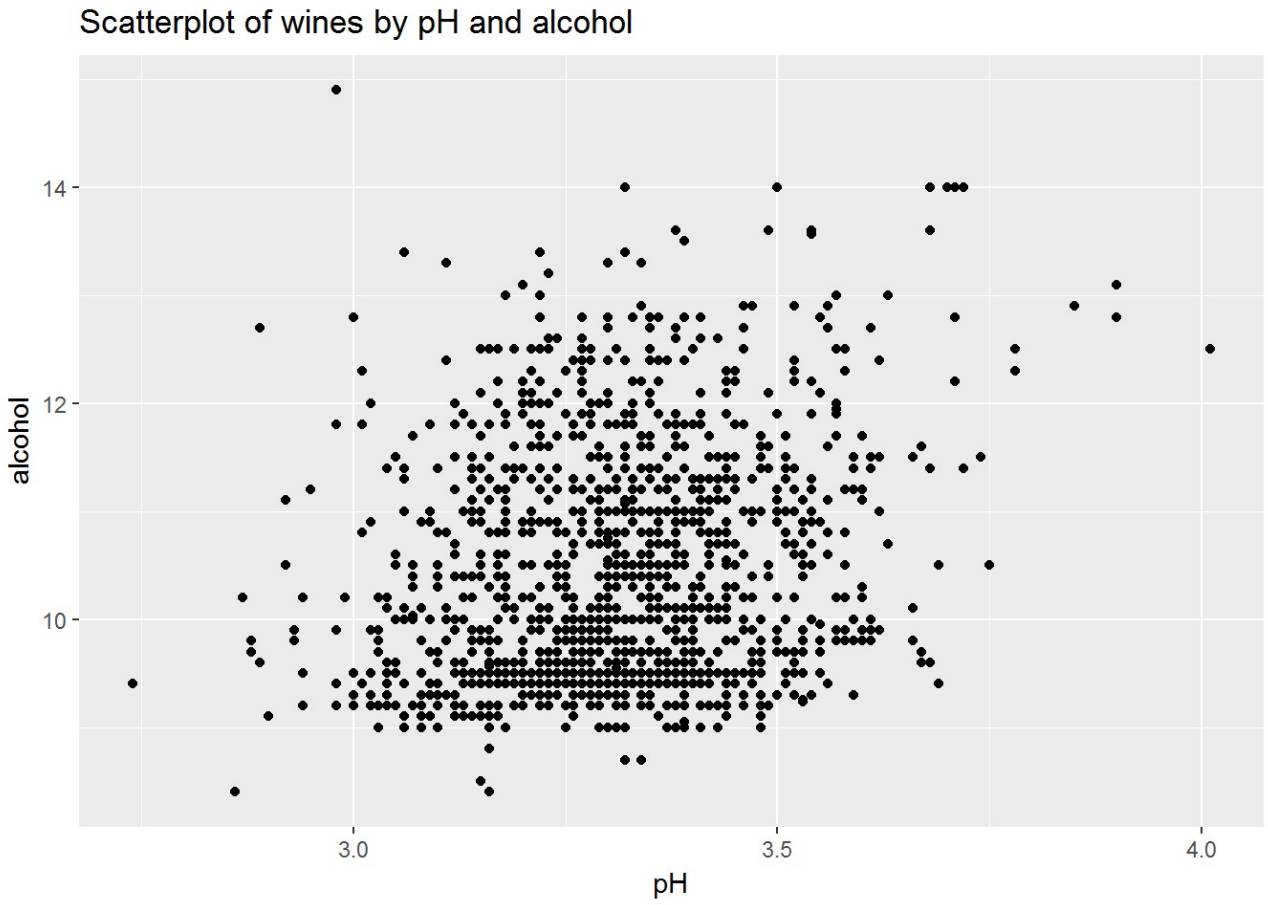
It appears that there is a weak negative correlation between pH and quality.

> From the above analysis of alcohol content and pH with regard to quality, it is clear that alcohol content more heavily influences wine quality than does pH. However, it will be interesting to investigate alcohol content and pH together.

pH and alcohol content

First, let us examine if there is a relationship between alcohol content and pH.

```
qplot(data=rw, x=pH, y=alcohol, main="Scatterplot of wines by pH and alcohol")
+
  geom_point()
```



From this scatterplot, it does not appear that there is a strong correlation between alcohol content and pH.

```
cor.test(rw$alcohol, rw$pH)
```

```
##
##  Pearson's product-moment correlation
##
## data:  rw$alcohol and rw$pH
## t = 8.397, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1582061 0.2521123
## sample estimates:
##       cor
## 0.2056325
```

It appears that there is a weak positive correlation between alcohol content and pH with a low p-value. This would suggest that a higher alcohol content may correlate with a higher pH. However, it would be well-advised to take this measure with a grain of salt.