Red Wine Analysis By : Shruti Rawat

```
## [1] "E:/DATA ANALYTICS/PROJECT2"
```

```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"
```

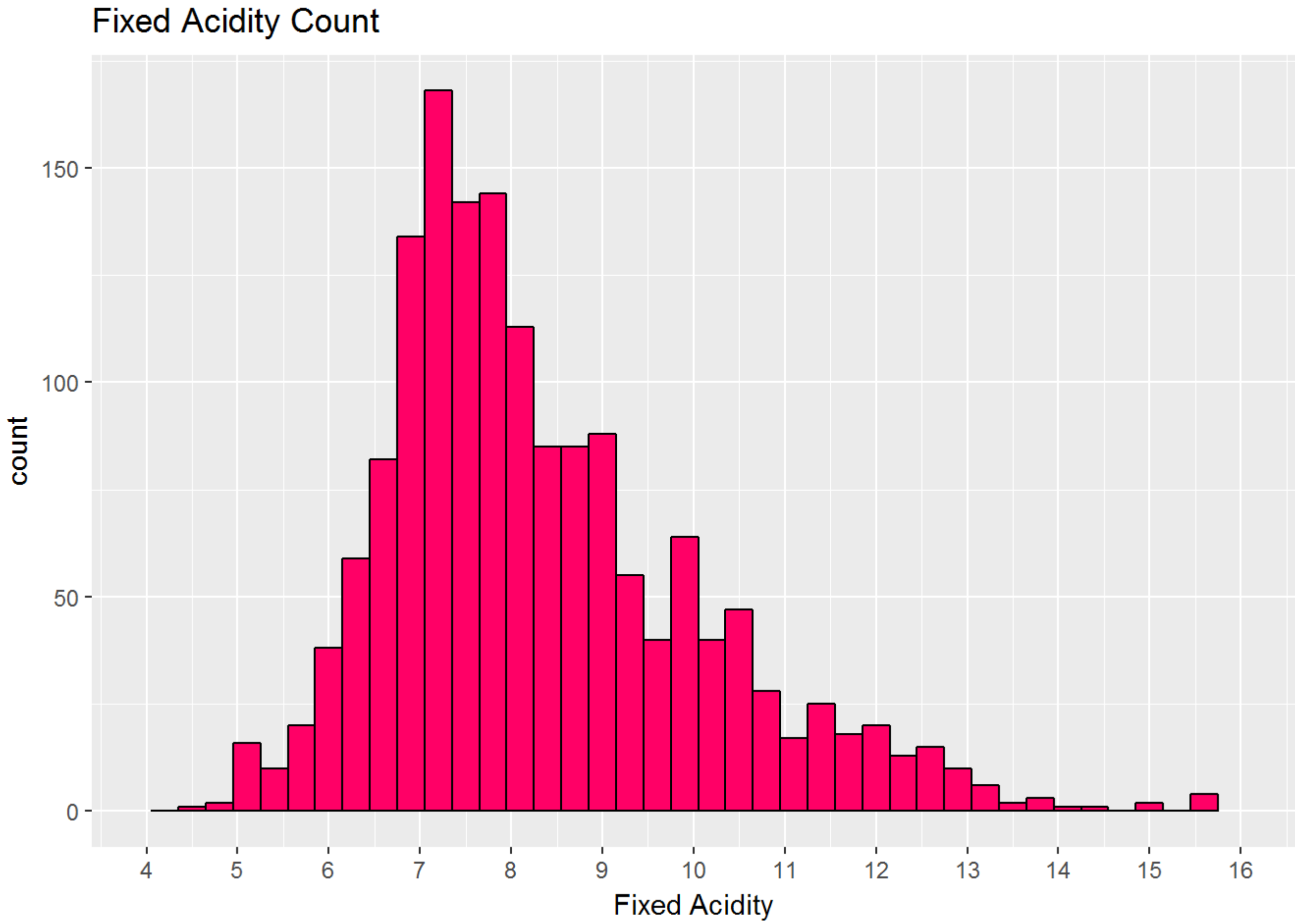
UNIVARIATE PLOTS SECTION

```
## '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 0.065 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 3.35 ...
##  $ 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 ...
```

```
##        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
```

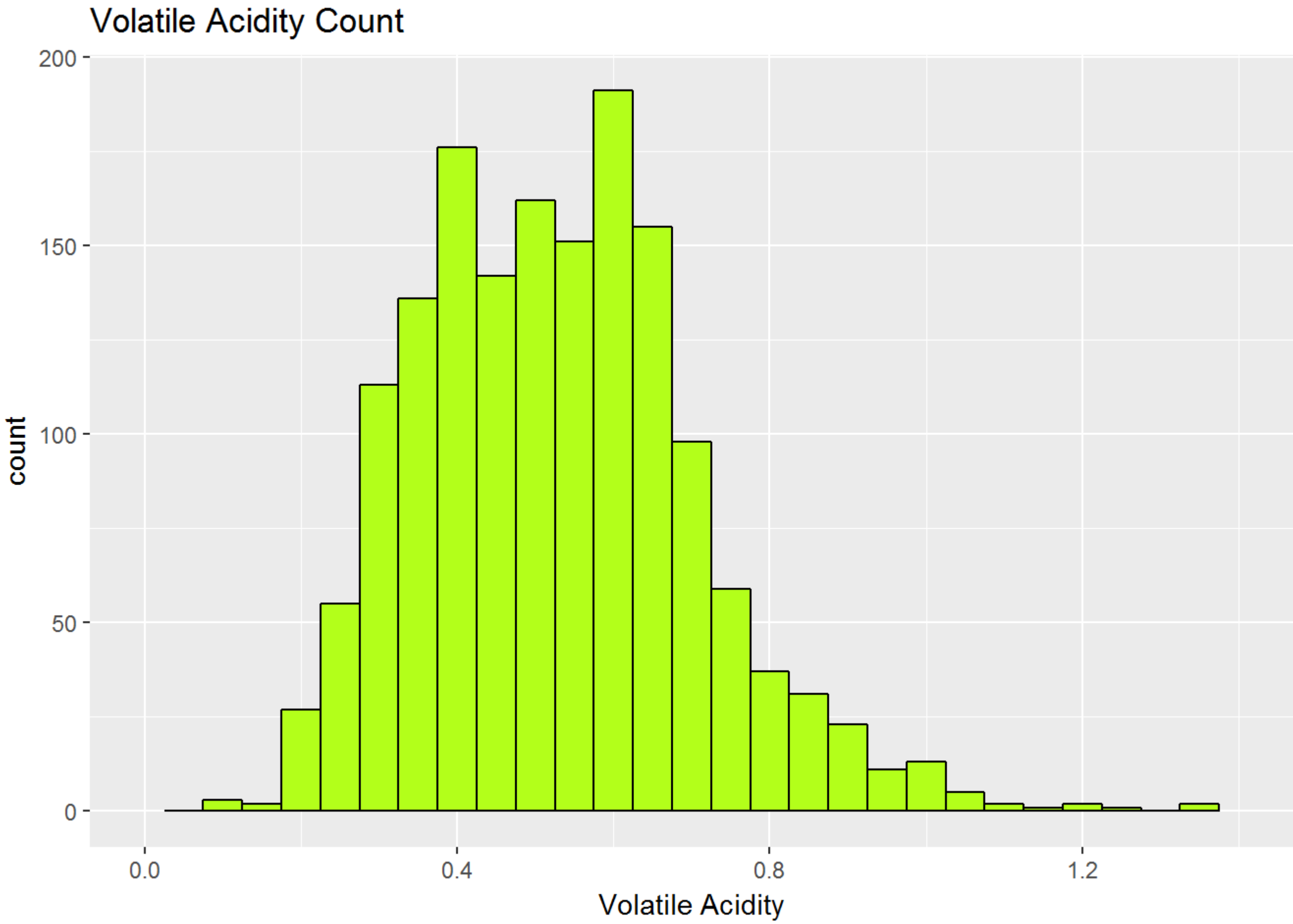
Our data set contains 1599 observations and 13 variables.

```
ggplot(aes(x=fixed.acidity),data=pr)+
  geom_histogram(binwidth=0.3,color=I('black'),fill=I('#ff0066'))+
  scale_x_continuous(limits=c(4,16),breaks=seq(4,16,1))+
  xlab('Fixed Acidity')+
  ggtitle('Fixed Acidity Count')
```



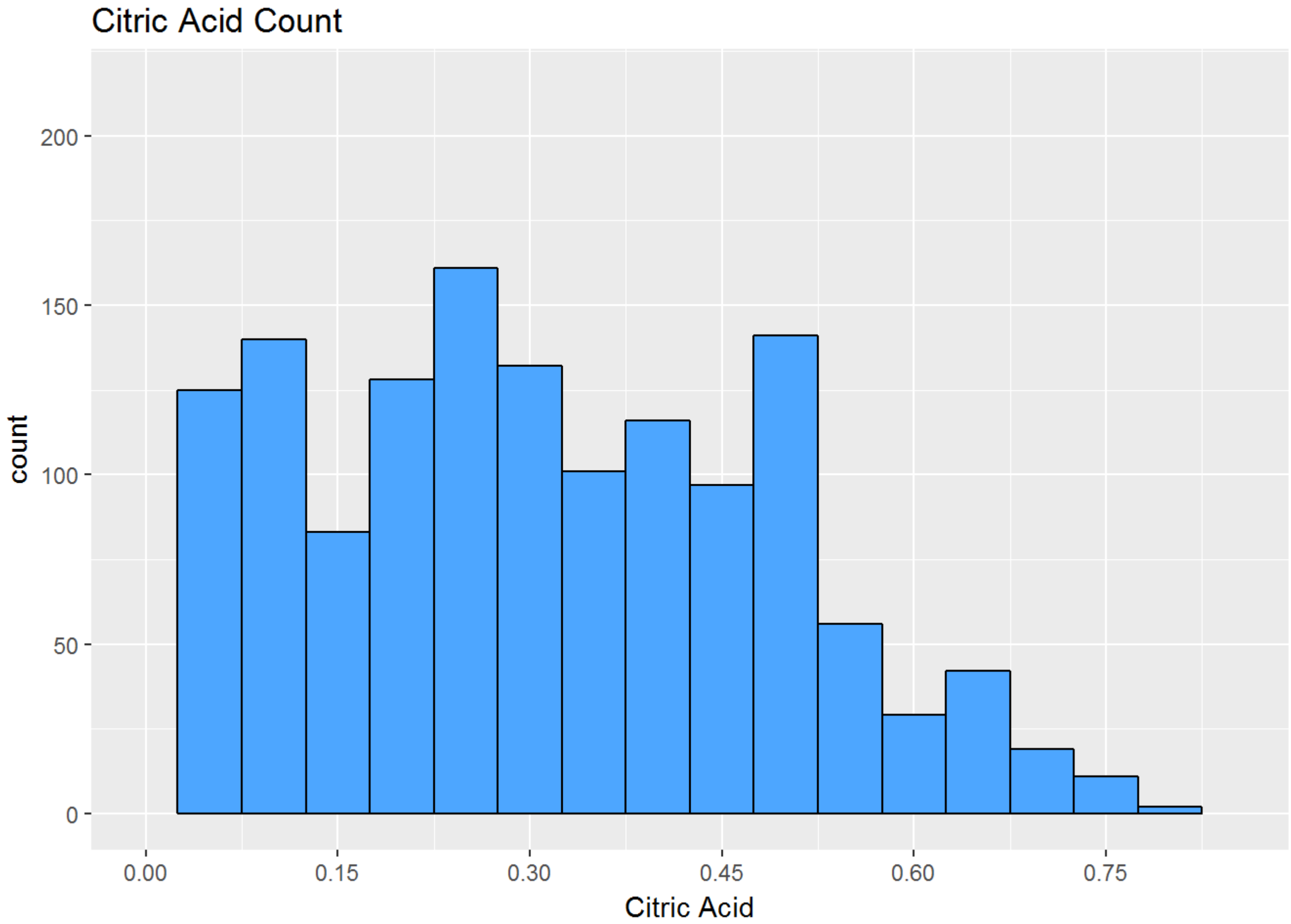
Fixed Acidity is a long tailed histogram plot which has a Median of 7.90 and Mean of 8.32.

```
ggplot(aes(x=volatile.acidity),data=pr)+
  geom_histogram(binwidth=0.05,color=I('black'),fill=I('#b3ff1a'))+
  scale_x_continuous(limits=c(0,1.4),breaks=seq(0,1.4,0.4))+
  xlab('Volatile Acidity')+
  ggtitle('Volatile Acidity Count')
```



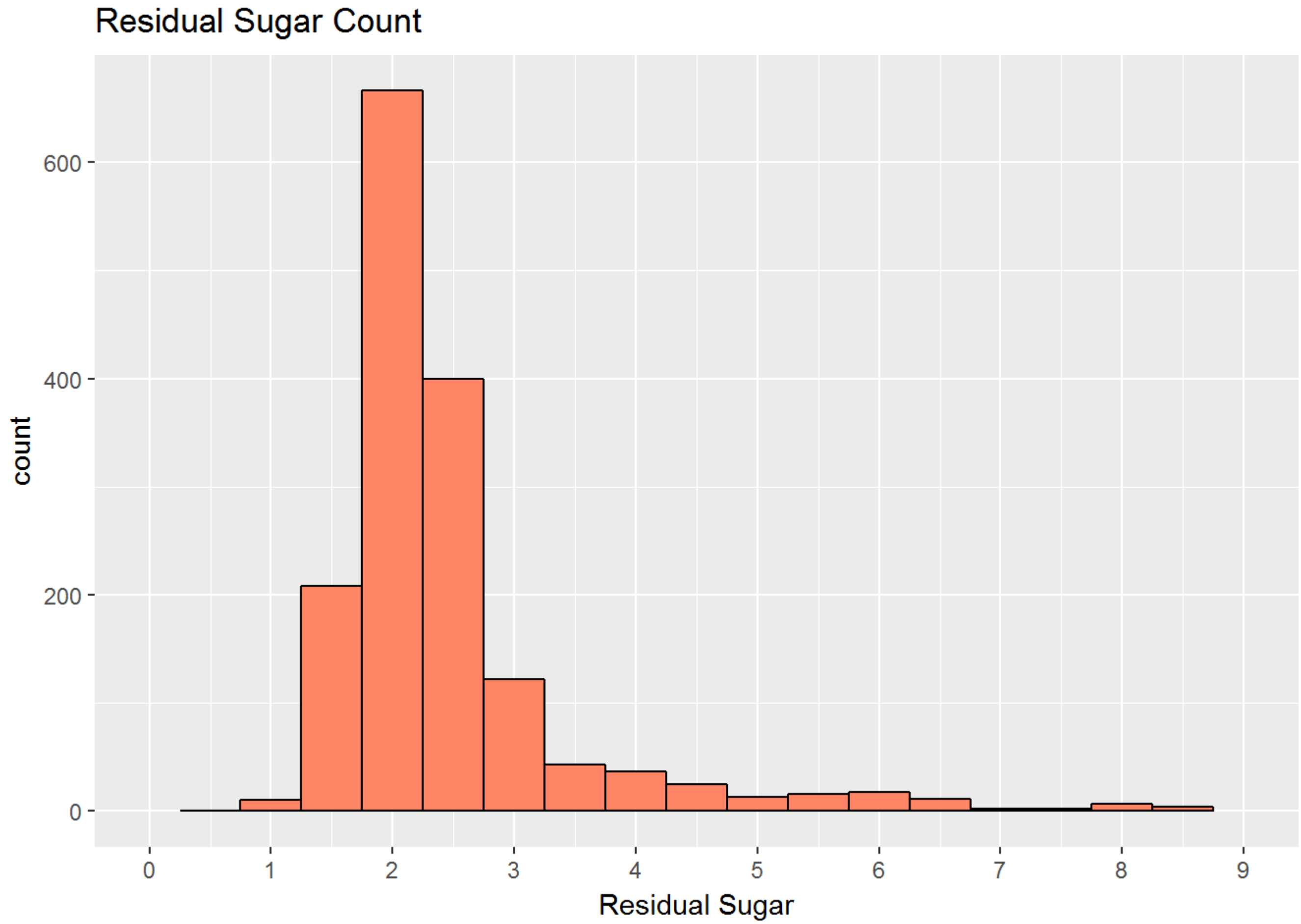
Volatile Acidity is a normal histogram plot as well as long tailed that has a Median of 0.5200 and Mean of 0.5278.

```
ggplot(aes(x=citric.acid),data=pr)+
  geom_histogram(binwidth=0.05,color=I('black'),fill=I('#4da6ff'))+
  scale_x_continuous(limits=c(0,0.85),breaks=seq(0,0.85,0.15))+
  xlab('Citric Acid')+
  ggtitle('Citric Acid Count')
```



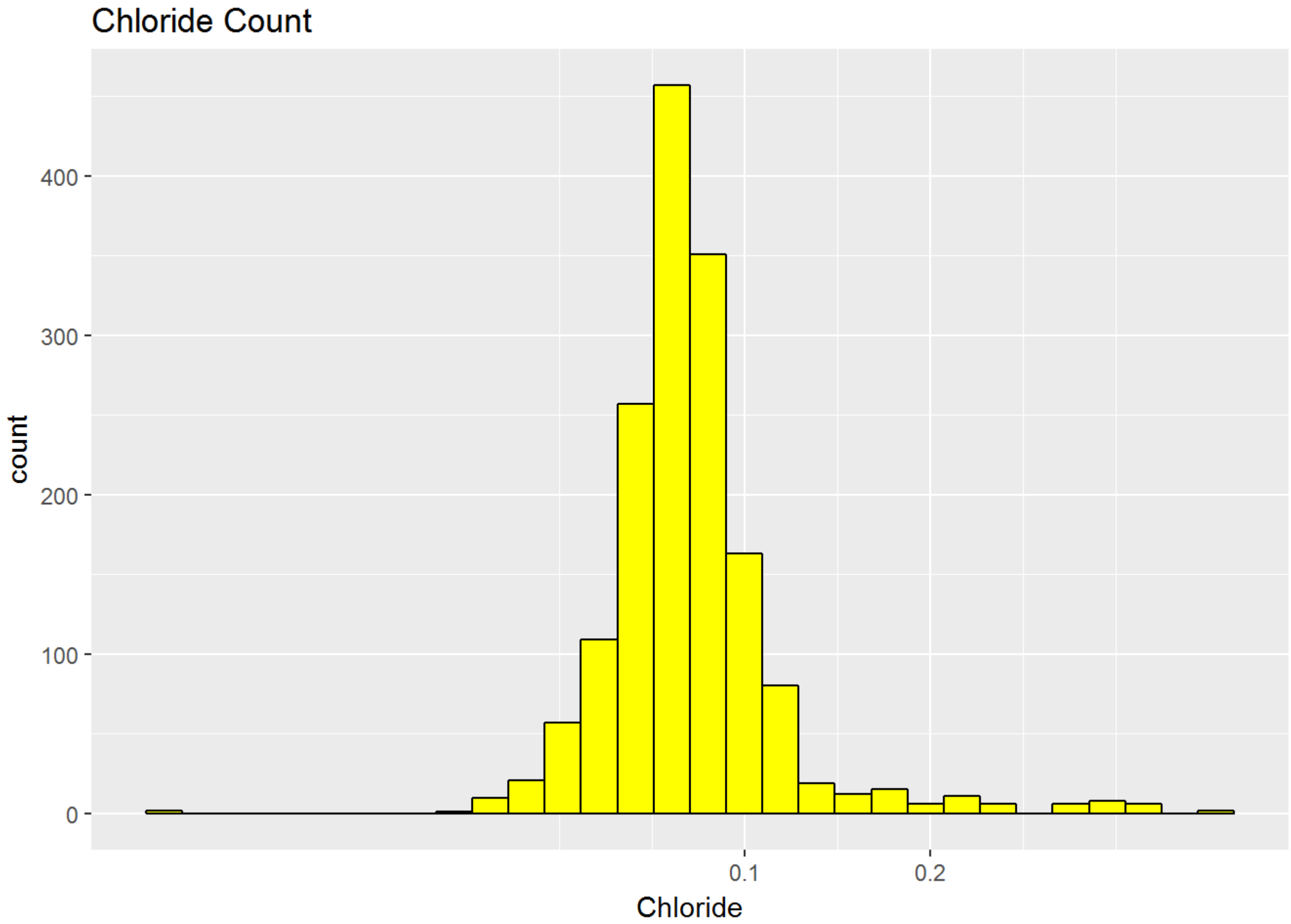
Citric Acid has a Median of 0.260 and Mean of 0.271.

```
ggplot(aes(x=residual.sugar),data=pr)+
  geom_histogram(binwidth=0.5,color=I('black'),fill=I('#ff8566'))+
  scale_x_continuous(limits=c(0,9),breaks=seq(0,9,1))+
  xlab('Residual Sugar')+
  ggtitle('Residual Sugar Count')
```



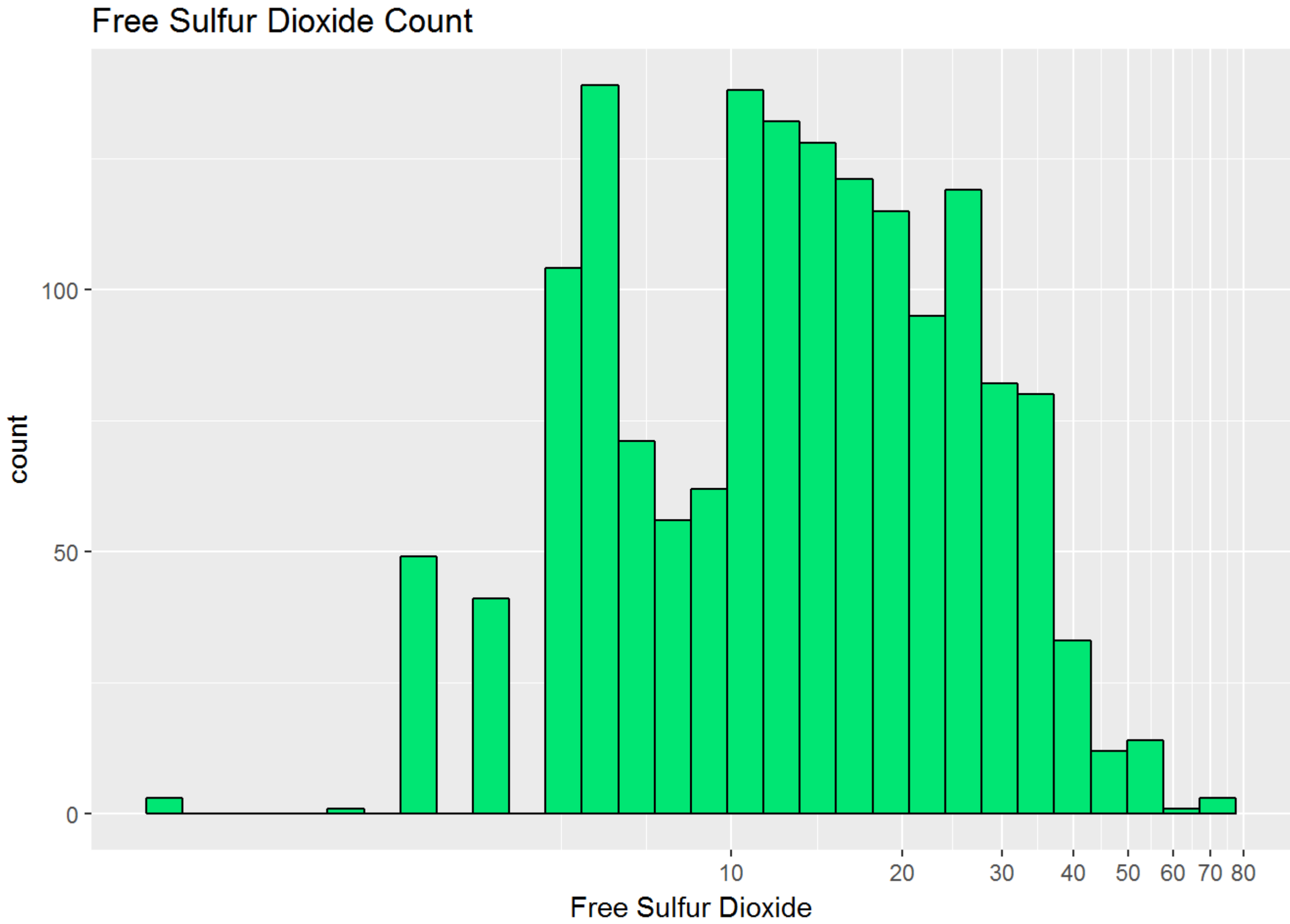
Residual Sugar has a long tailed histogram plot with a Median of 2.200 and Mean of 2.539.

```
ggplot(aes(x=chlorides),data=pr)+
  geom_histogram(color=I('black'),fill=I('#ffff00'))+
  scale_x_log10(breaks=seq(0,0.2,0.1))+
  xlab('Chloride')+
  ggtitle('Chloride Count')
```

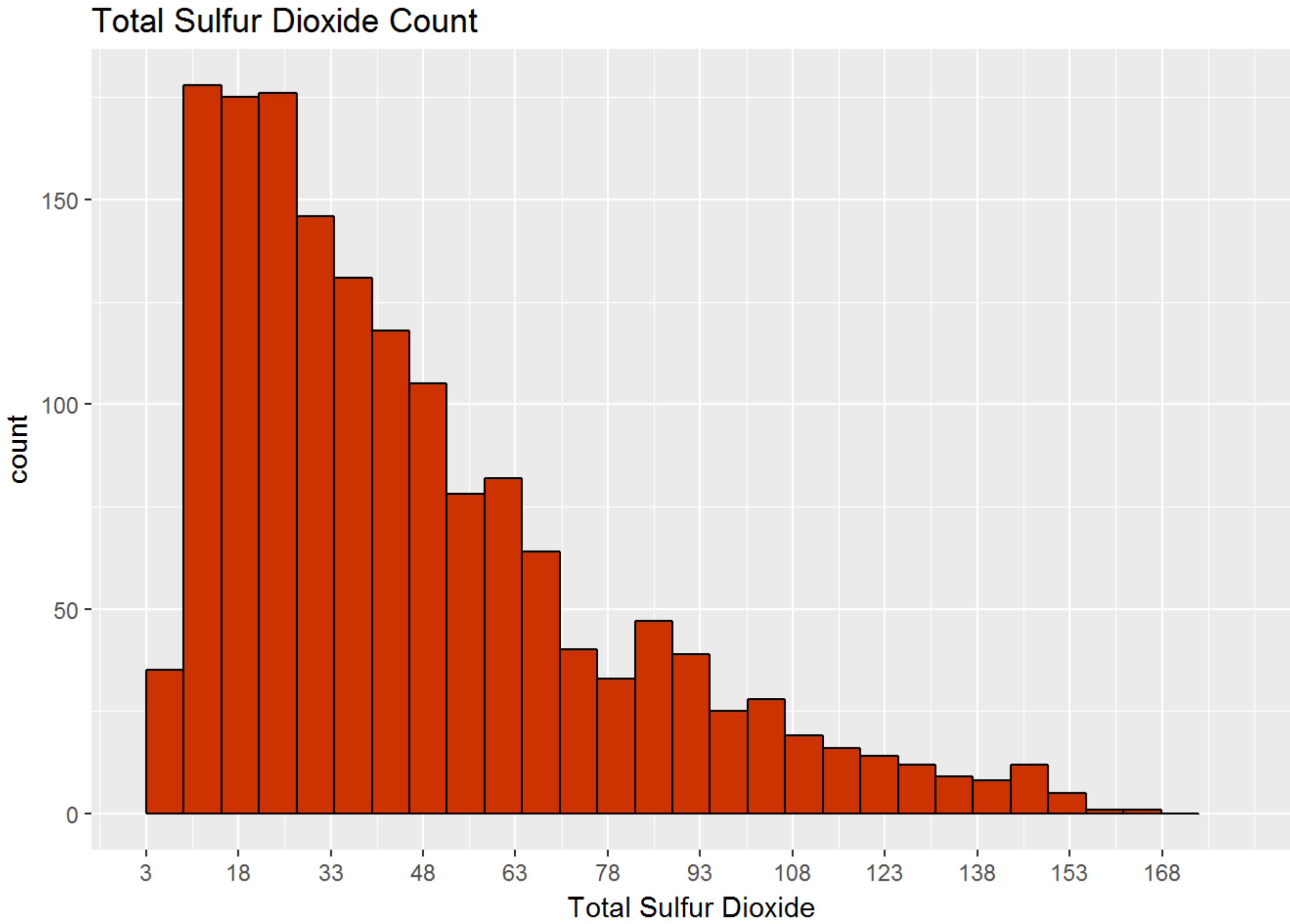


Chlorides have a long tailed histogram plot with a Median of 0.07900 and Mean of 0.08747.

```
ggplot(aes(x=free.sulfur.dioxide),data=pr)+
  geom_histogram(color=I('black'),fill=I('#00e673'))+
  #scale_x_continuous(limits=c(0,80),breaks=seq(0,80,10))+
  scale_x_log10(breaks=seq(0,80,10))+
  xlab('Free Sulfur Dioxide')+
  ggtitle('Free Sulfur Dioxide Count')
```

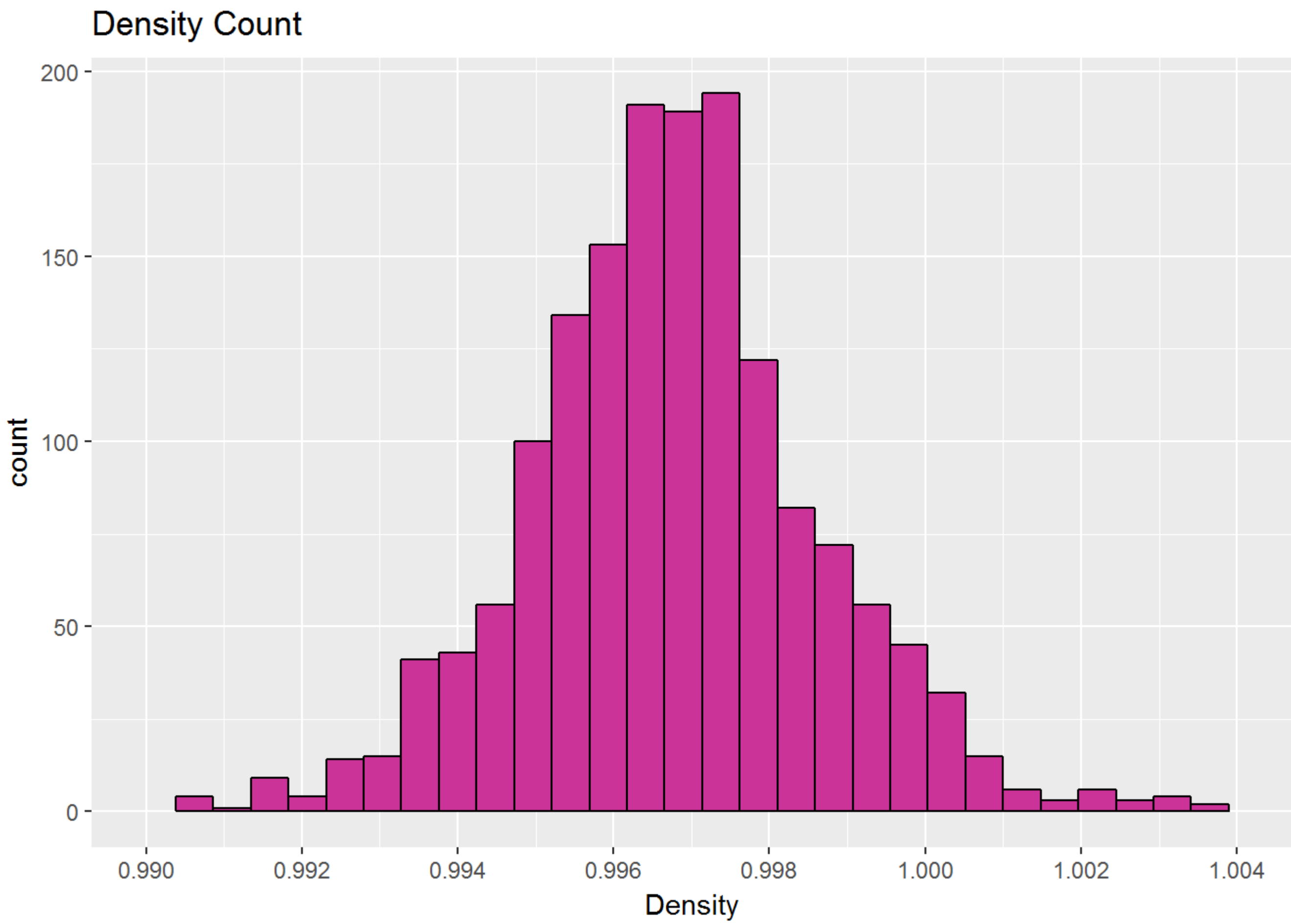


Free Sulphur Dioxide has a Median of 14.00 and Mean of 15.87.



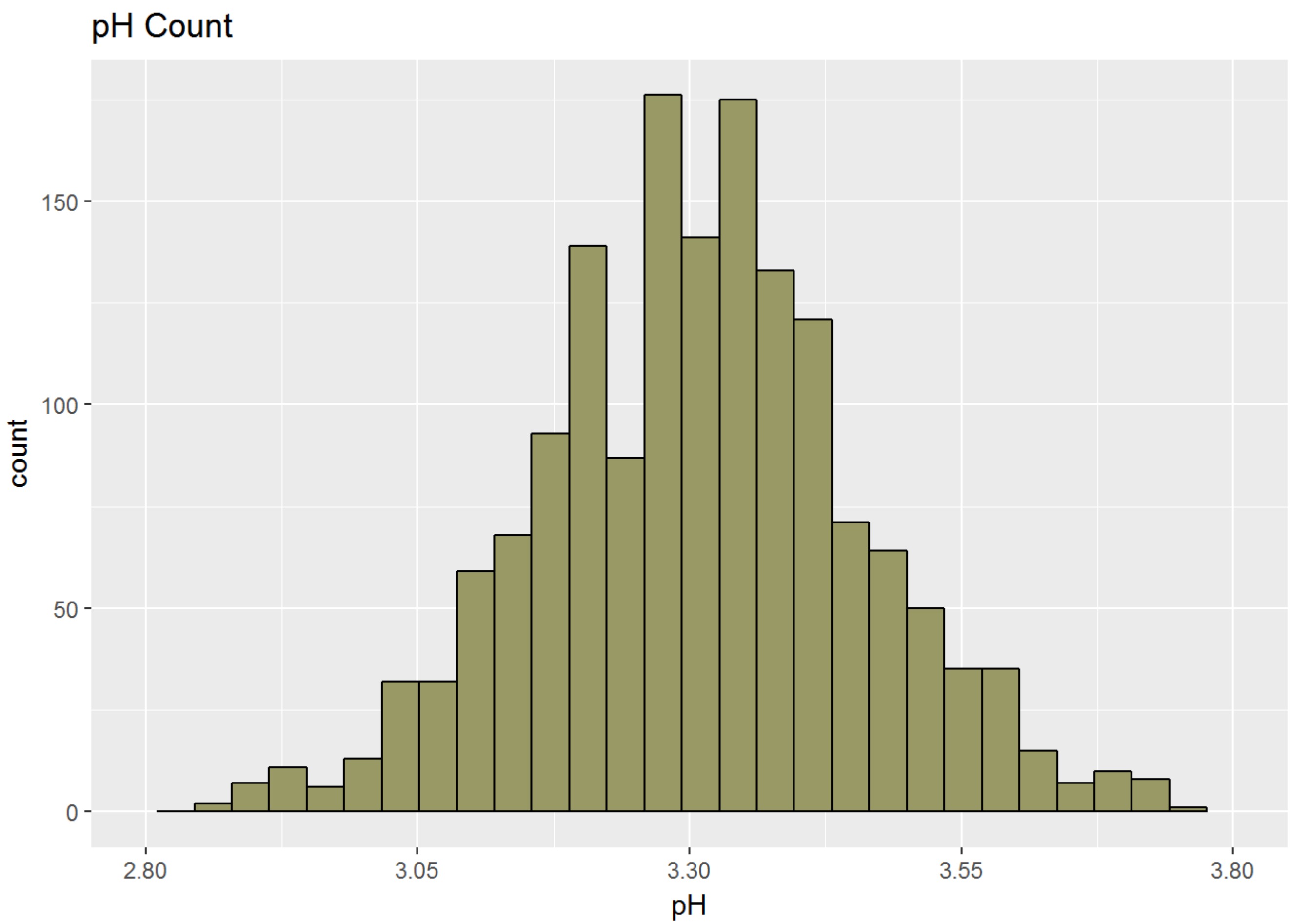
Total Sulfur Dioxide has a Median of 38.00 and Mean of 46.47.

```
ggplot(aes(x=density),data=pr)+
  geom_histogram(color=I('black'),fill=I('#cc3399'))+
  scale_x_continuous(limits=c(0.990,1.004),breaks=seq(0.990,1.004,0.002))+
  xlab('Density')+
  ggtitle('Density Count')
```



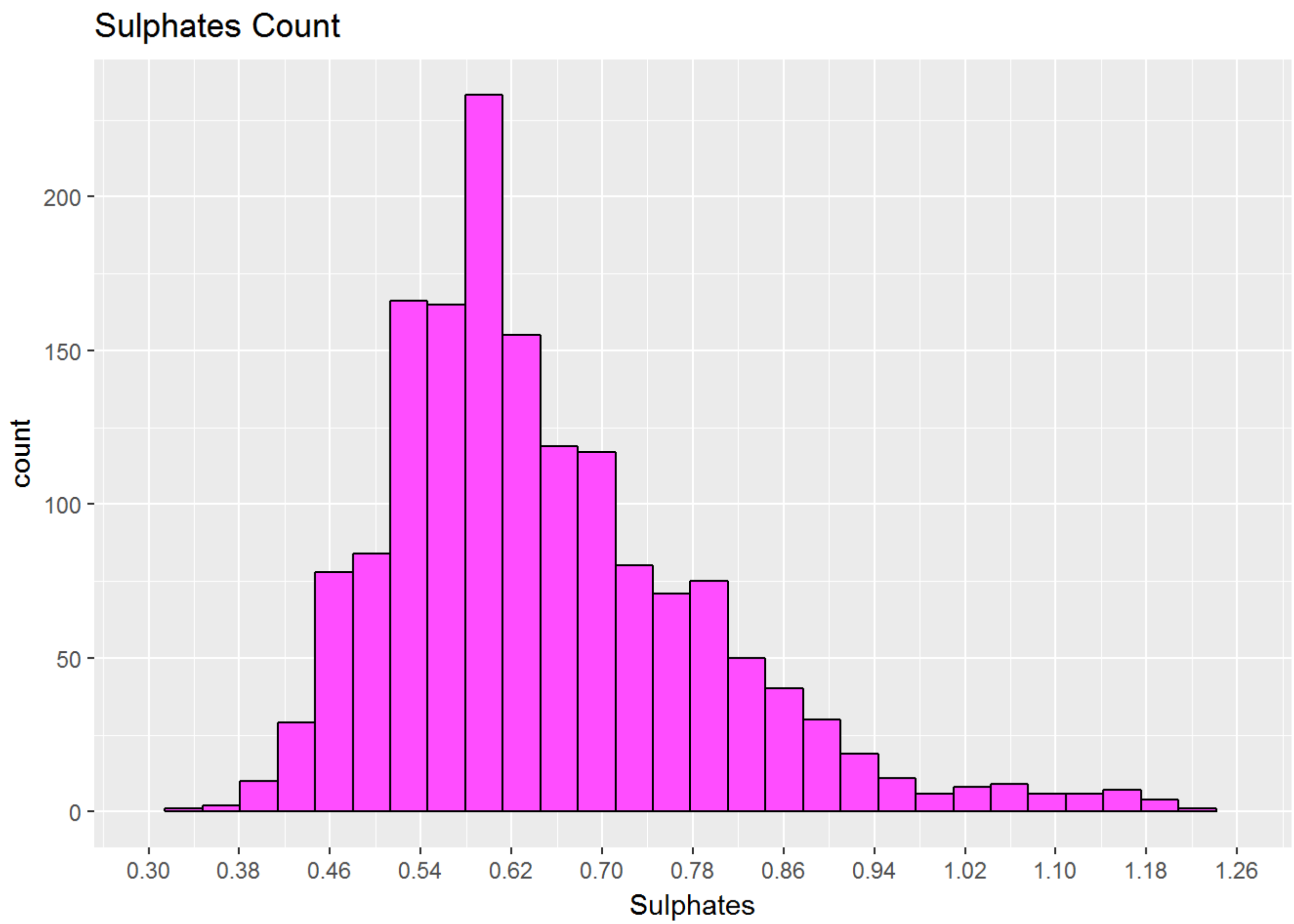
Density is a normal histogram plot that has a Median of 0.9968 and Mean of 0.9967.

```
ggplot(aes(x=pH),data=pr)+
  geom_histogram(color=I('black'),fill=I('#999966'))+
  scale_x_continuous(limits=c(2.80,3.80),breaks=seq(2.80,3.80,0.25))+
  xlab('pH')+
  ggtitle('pH Count')
```



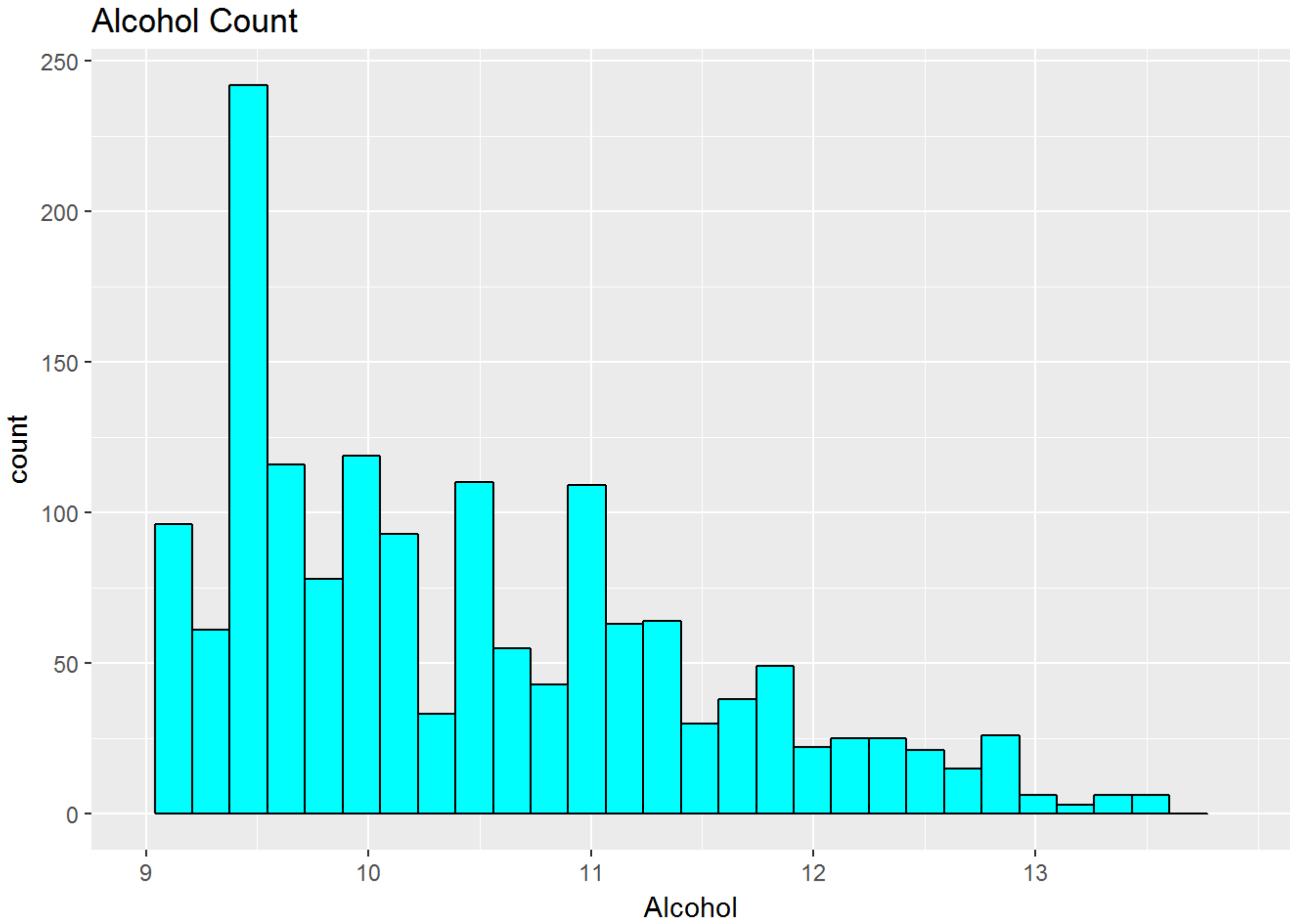
pH is a normal histogram plot that has a Median of 3.310 and Mean of 3.311.

```
ggplot(aes(x=sulphates),data=pr)+
  geom_histogram(color=I('black'),fill=I('#ff4dff'))+
  scale_x_continuous(limits=c(0.3,1.26),breaks=seq(0.3,1.26,0.08))+
  xlab('Sulphates')+
  ggtitle('Sulphates Count')
```



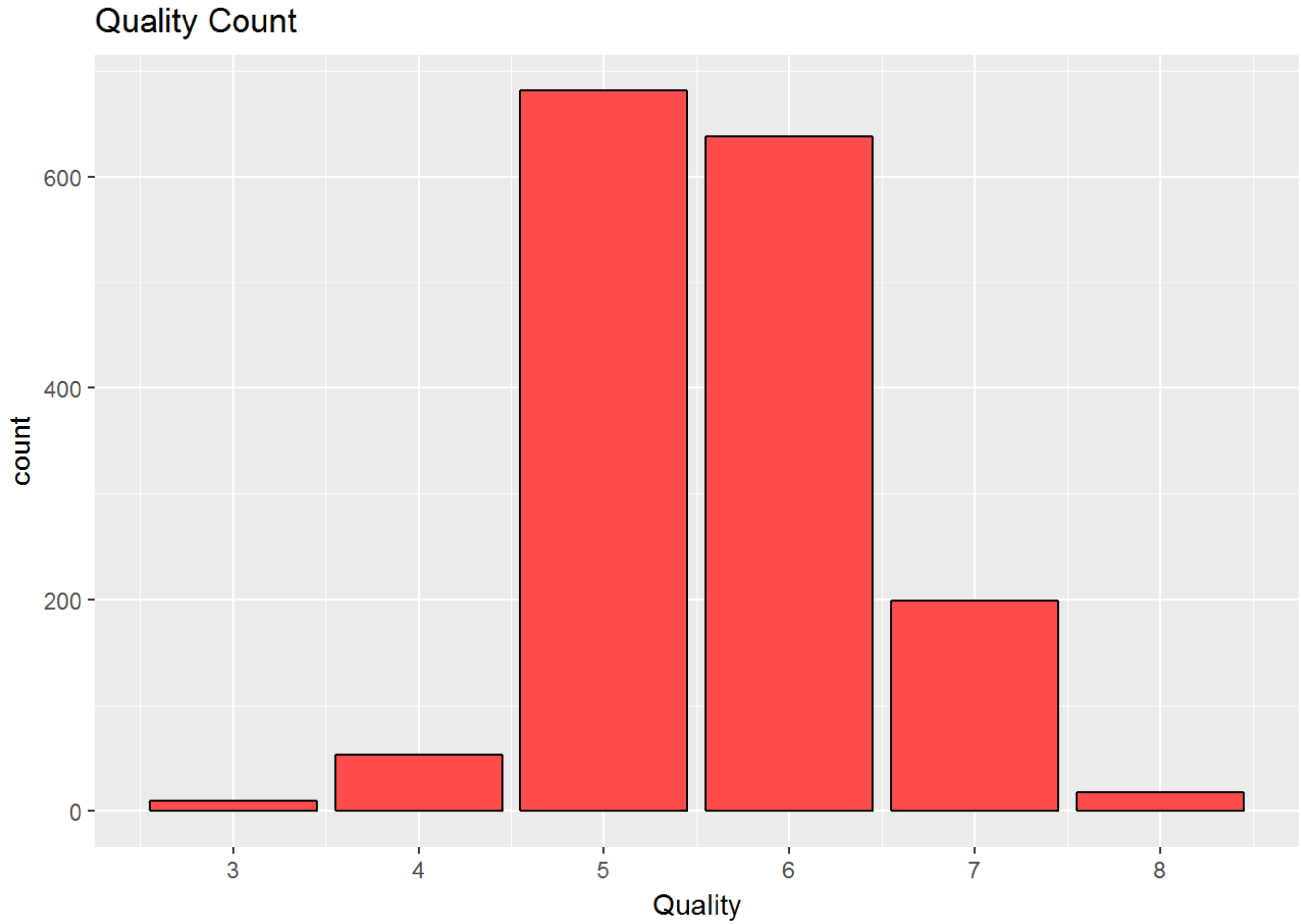
Sulphates have a Median of 0.6200 and Mean of 0.6581.

```
ggplot(aes(x=alcohol),data=pr)+
  geom_histogram(color=I('black'),fill=I('#00ffff'))+
  scale_x_continuous(limits=c(9,13.9),breaks=seq(9,13.9,1))+
  xlab('Alcohol')+
  ggtitle('Alcohol Count')
```



Alcohol has a Median of 10.20 and Mean of 10.42.

```
ggplot(aes(x=quality),data=pr)+
  geom_bar(color=I('black'),fill=I('#ff4d4d'))+
  scale_x_continuous(breaks=seq(2,9,1))+
  xlab('Quality')+
  ggtitle('Quality Count')
```



Quality has a Median of 6.000 and Mean of 5.636.

Univariate Analysis

What is the structure of your dataset?

The red whine dataset contains 13 variables and 1599 observations. All the variables of num data type except X and quality variable.

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

Quality of the wine is the main focus here. The quality of the wine is dependent on the content of the alcohol, citric acid and sulphates in general. These variables make the quality of the wine from better to best. Also, if the volatile acidity variable content is high then we risk the quality of or wine.

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

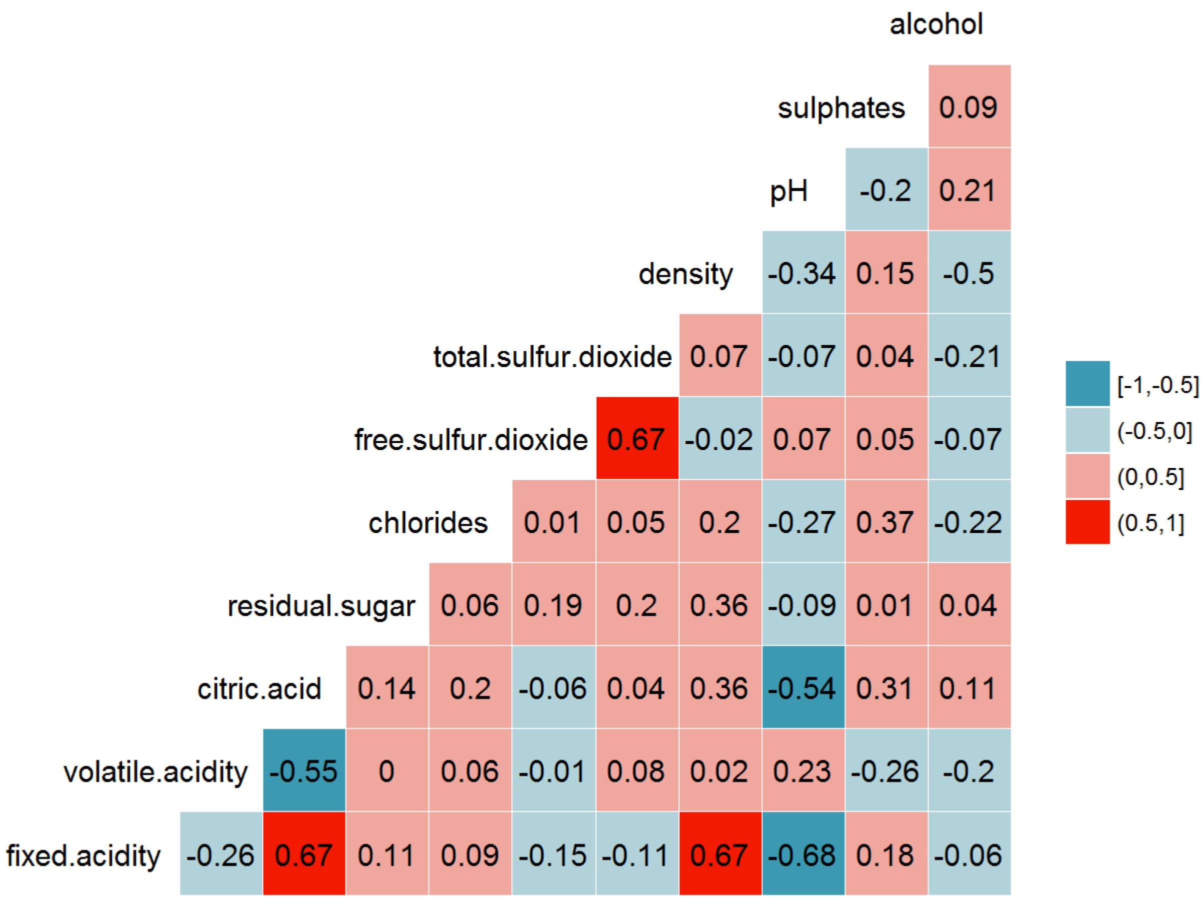
No I did not create any new variable from the existing variable.

Did you perform any operations on the data to tidy, adjust, or change the formof the data? If so, why did you do this?

Yes, I log transformed the chlorides because they were were skewed and to get to bulk of data log transformation was performed.

BIVARIATE PLOTS SECTION

```
pr1 <- select(pr,2,2:12)
ggcorr(pr1,label = TRUE,label_round=2,label_size=4,nbreaks=4,size=4,hjust=0.85,layout.exp=2)
```

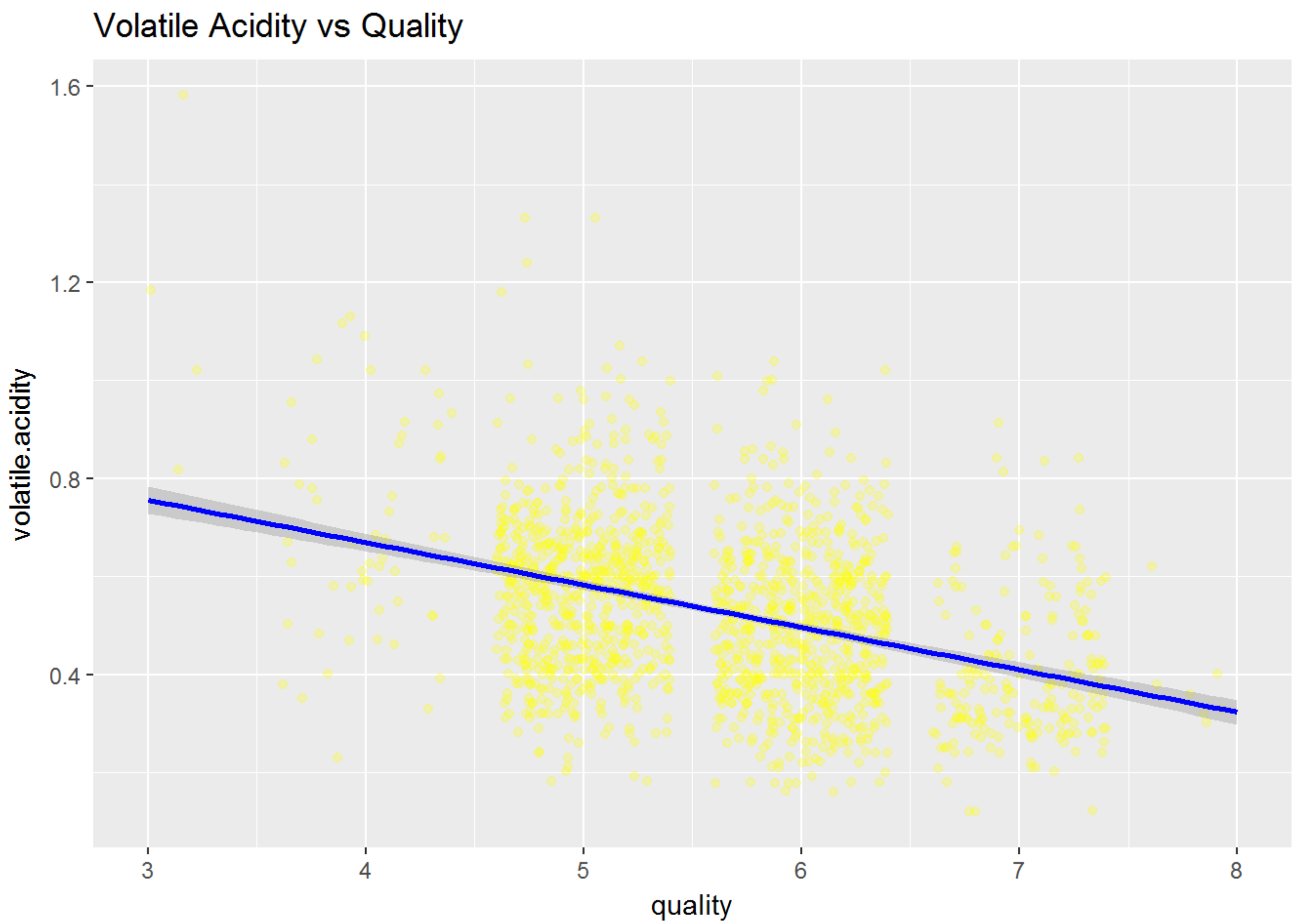


From the above results, we see that alcohol,citric acid and suphates are positively correlated variable, where as alcohol has the greatest value among them. Volatile acidity stands out as a negatively correlated variable and has the greatest value among every negatively correlated variable. In this we take values equal or greater than 2 in both negative and positive correlated variables.

```
ggplot(aes(x=quality,y=alcohol),data=pr)+
  geom_jitter(color=('red'),alpha=1/2)+
  scale_x_continuous(breaks=seq(3,8,1),limit=c(3,8))+
  geom_smooth(method='lm',color='blue')+
  ggtitle('Alcohol vs Quality')
```

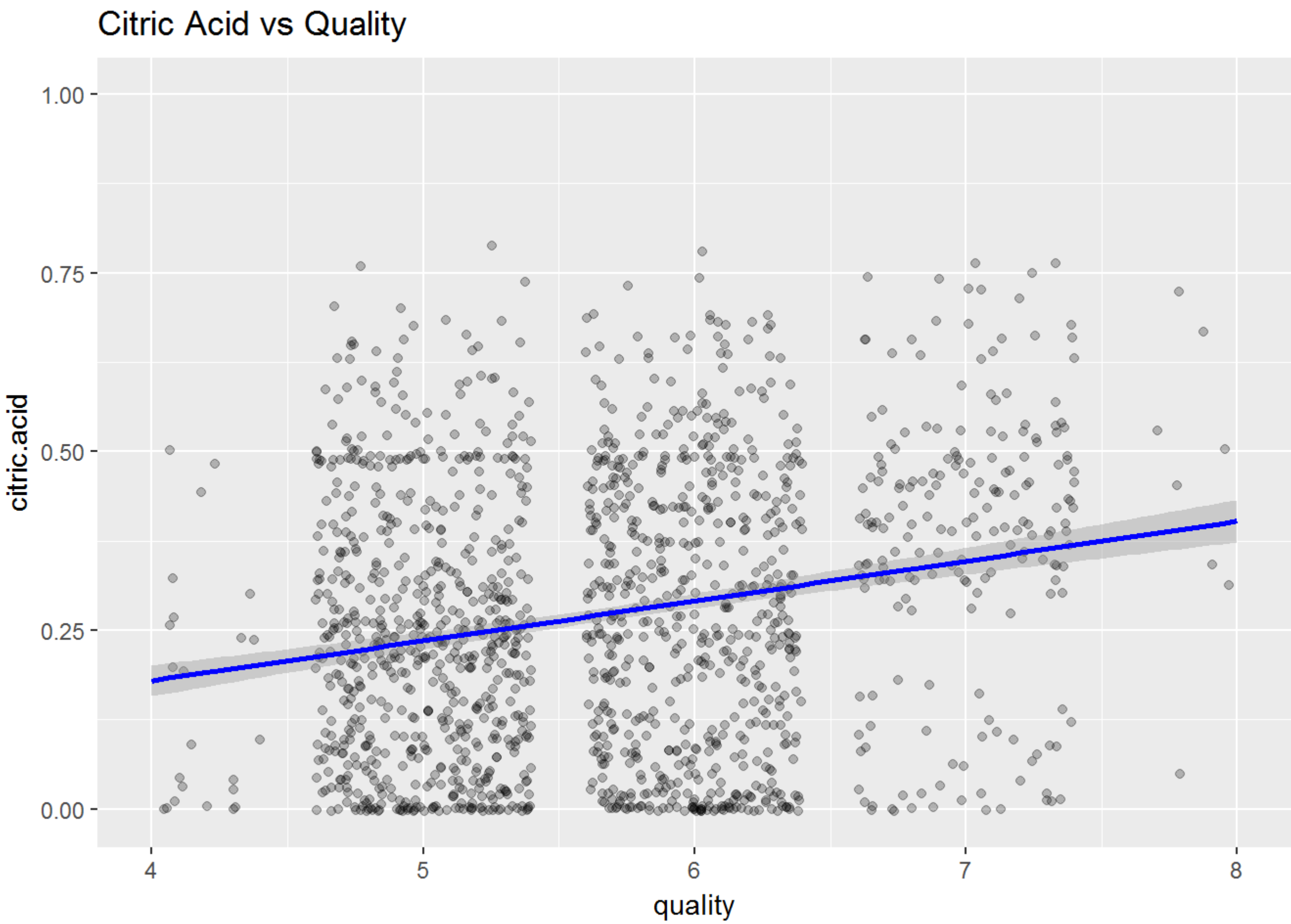
In this plot of Alcohol vs Quality we see that alcohol shows a positive linear relationship with the quality of wine. Thus, proving that high content alcohol gives us a high quality wine.

```
ggplot(aes(x=quality,y=volatile.acidity),data=pr)+
  geom_jitter(color=('yellow'),alpha=1/4)+
  scale_x_continuous(breaks=seq(3,8,1),limit=c(3,8))+
  geom_smooth(method='lm',color='blue')+
  ggtitle('Volatile Acidity vs Quality')
```



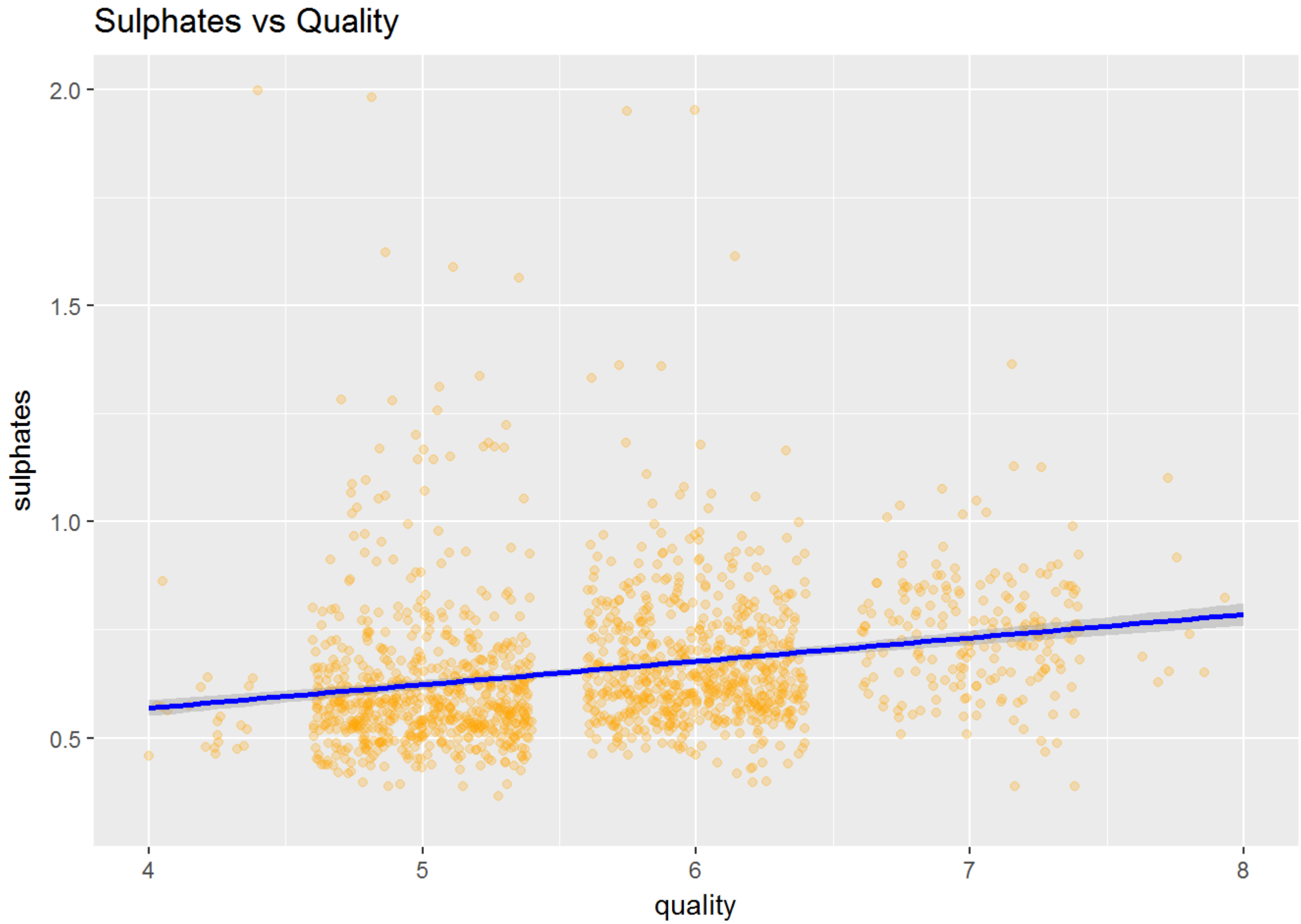
In this plot of Volatile Acidity vs Quality we see that Volatile Acidity shows a negative linear relationship with the quality of wine. Thus, proving that low content volatile acid gives us a high quality wine.

```
ggplot(aes(x=quality,y=citric.acid),data=pr)+
  geom_jitter(color=('black'),alpha=1/4)+
  scale_x_continuous(breaks=seq(4,8,1),limit=c(4,8))+
  geom_smooth(method='lm',color='blue')+
  ggtitle('Citric Acid vs Quality')
```



In this plot of Citric Acid vs Quality we see that citric acid shows a positive linear relationship with the quality of wine. Thus, proving that high content citric acid gives us a high quality wine.

```
ggplot(aes(x=quality,y=sulphates),data=pr)+
  geom_jitter(color=('orange'),alpha=1/4)+
  scale_x_continuous(breaks=seq(4,8,1),limit=c(4,8))+
  geom_smooth(method='lm',color='blue')+
  ggtitle('Sulphates vs Quality')
```



In this plot of Sulphates vs Quality we see that sulphates shows a positive linear relationship with the quality of wine. Thus, proving that high content sulphates gives us a high quality wine.

```
set.seed(1599)
View(pr)
pr_subset <- pr[,c(1:12)]
cor(pr_subset[, c(1:11)])
```

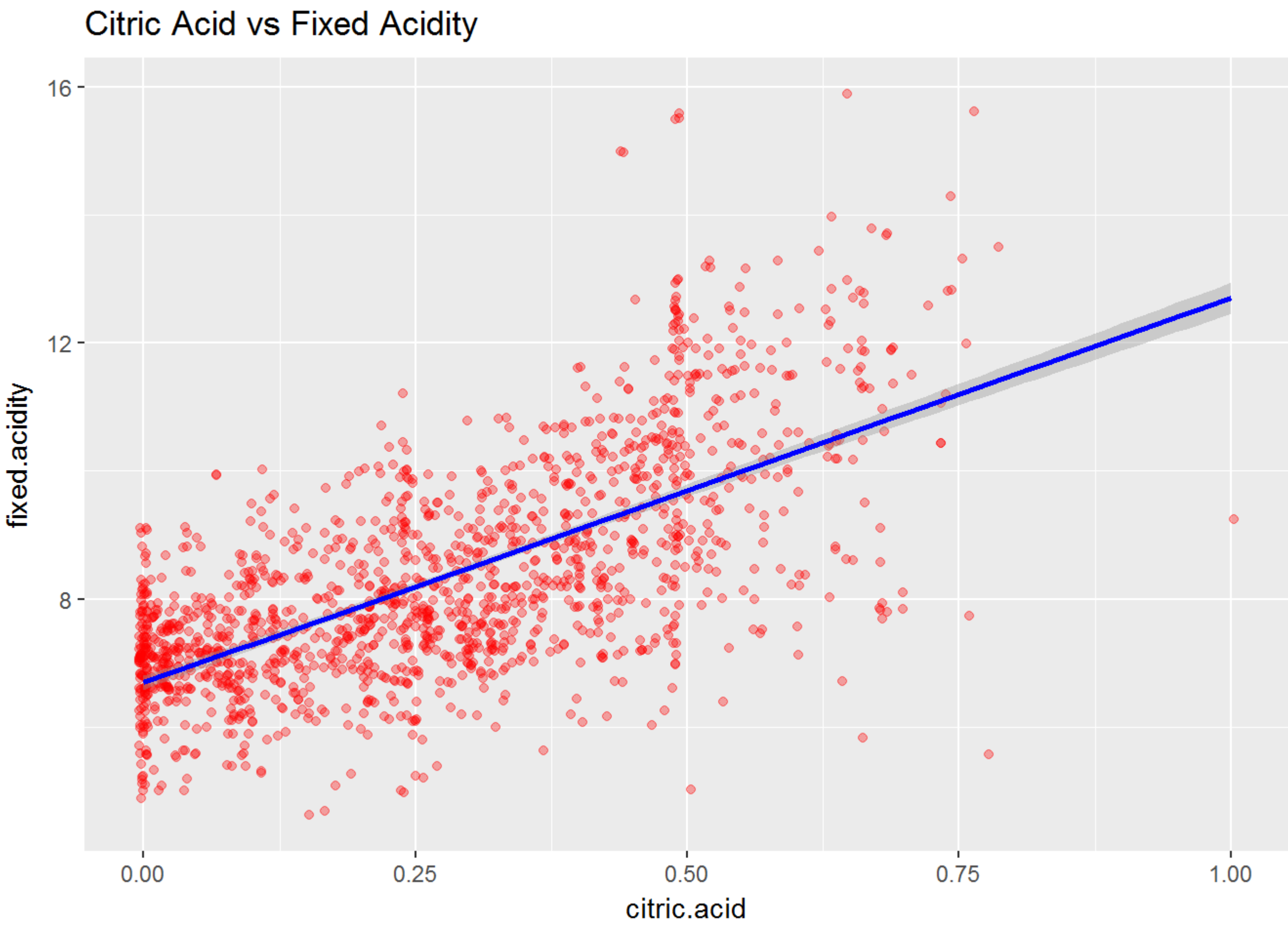
```
##                               X fixed.acidity volatile.acidity
## X                    1.000000000   -0.26848392     -0.008815099
## fixed.acidity       -0.268483920    1.00000000     -0.256130895
## volatile.acidity    -0.008815099   -0.25613089      1.000000000
## citric.acid         -0.153551355    0.67170343     -0.552495685
## residual.sugar      -0.031260835    0.11477672      0.001917882
## chlorides           -0.119868519    0.09370519      0.061297772
## free.sulfur.dioxide  0.090479643   -0.15379419     -0.010503827
## total.sulfur.dioxide -0.117849669   -0.11318144      0.076470005
## density             -0.368372087    0.66804729      0.022026232
## pH                   0.136005328   -0.68297819      0.234937294
## sulphates           -0.125306999    0.18300566     -0.260986685
##                     citric.acid residual.sugar    chlorides
## X                   -0.15355136   -0.031260835 -0.119868519
## fixed.acidity        0.67170343    0.114776724  0.093705186
## volatile.acidity    -0.55249568    0.001917882  0.061297772
## citric.acid          1.00000000    0.143577162  0.203822914
## residual.sugar       0.14357716    1.000000000  0.055609535
## chlorides            0.20382291    0.055609535  1.000000000
## free.sulfur.dioxide -0.06097813    0.187048995  0.005562147
## total.sulfur.dioxide  0.03553302    0.203027882  0.047400468
## density              0.36494718    0.355283371  0.200632327
## pH                  -0.54190414   -0.085652422 -0.265026131
## sulphates            0.31277004    0.005527121  0.371260481
##                     free.sulfur.dioxide total.sulfur.dioxide     density
## X                           0.090479643          -0.11784967 -0.36837209
## fixed.acidity              -0.153794193          -0.11318144  0.66804729
## volatile.acidity           -0.010503827           0.07647000  0.02202623
## citric.acid                -0.060978129           0.03553302  0.36494718
## residual.sugar              0.187048995           0.20302788  0.35528337
## chlorides                   0.005562147           0.04740047  0.20063233
## free.sulfur.dioxide         1.000000000           0.66766645 -0.02194583
## total.sulfur.dioxide        0.667666450           1.00000000  0.07126948
## density                    -0.021945831           0.07126948  1.00000000
## pH                          0.070377499          -0.06649456 -0.34169933
## sulphates                   0.051657572           0.04294684  0.14850641
##                              pH    sulphates
## X                    0.13600533 -0.125306999
```

```
## fixed.acidity        -0.68297819  0.183005664
## volatile.acidity      0.23493729 -0.260986685
## citric.acid          -0.54190414  0.312770044
## residual.sugar       -0.08565242  0.005527121
## chlorides            -0.26502613  0.371260481
## free.sulfur.dioxide   0.07037750  0.051657572
## total.sulfur.dioxide -0.06649456  0.042946836
## density              -0.34169933  0.148506412
## pH                    1.00000000 -0.196647602
## sulphates            -0.19664760  1.000000000
```

Through this we get the correlation between all the variables with each other.

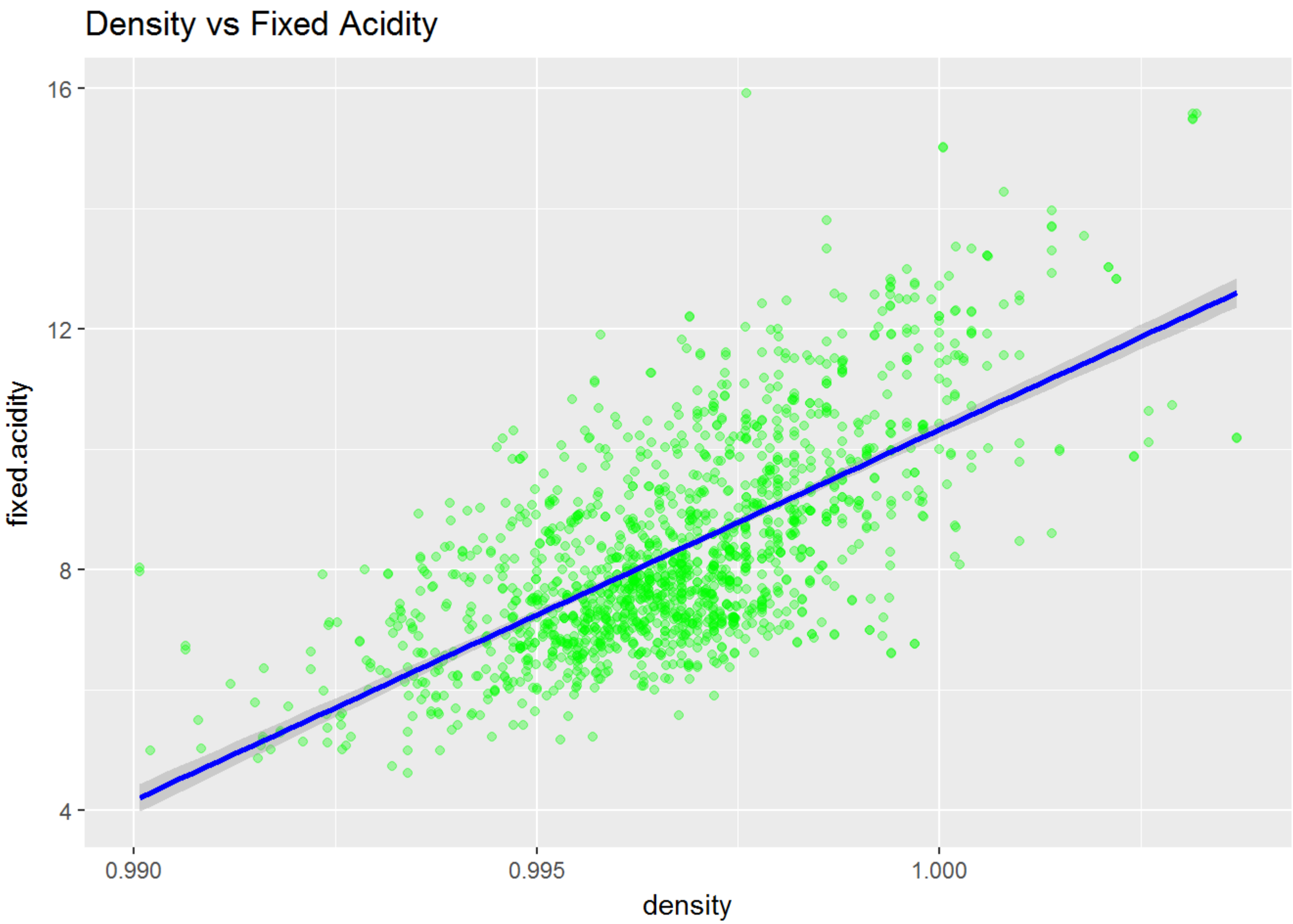
Now we take variables those have correlation value greater than 5,with either negatively or positively correlated variables. Thus we have 1)citric acid and fixed acidity with value 0.6717 2)density and fixed acidity with value 0.6680 3)pH and fixed acidity with value -0.6829 4)citric acid and volatile acidity with value -0.5524 5)total sulfur dioxide and free sulfur dioxide with value 0.6676

```
ggplot(aes(x=citric.acid,y=fixed.acidity),data=pr)+
  geom_jitter(color=('red'),alpha=1/3)+
  geom_smooth(method='lm',color='blue')+
  ggtitle('Citric Acid vs Fixed Acidity')
```



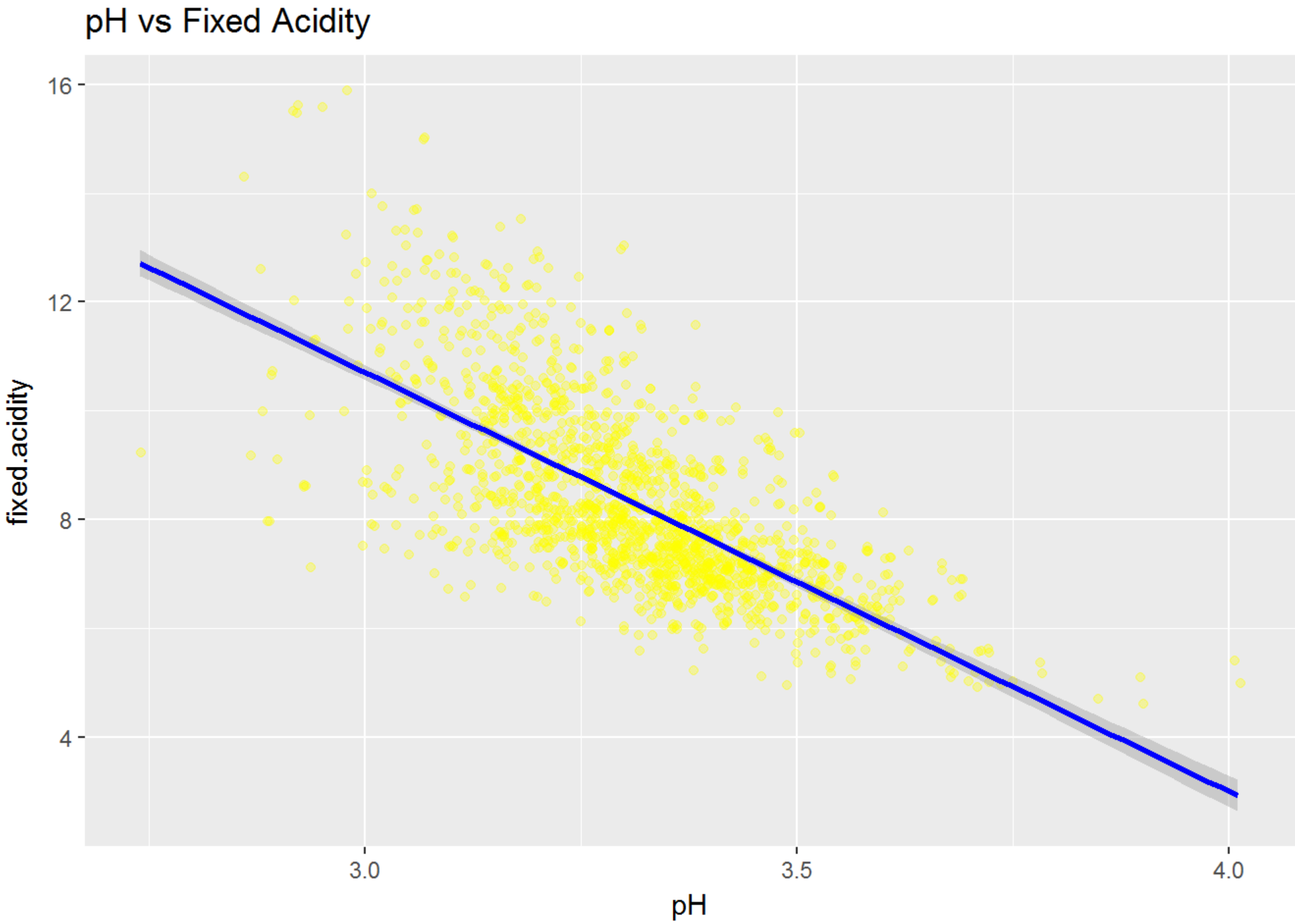
This plot of Citric Acid vs Fixed Acidity indicates that the correlation between the two variables is very strong i.e. 0.6717 and has a positive linear relationship.

```
ggplot(aes(x=density,y=fixed.acidity),data=pr)+
  geom_jitter(color=('green'),alpha=1/3)+
  geom_smooth(method='lm',color='blue')+
  ggtitle('Density vs Fixed Acidity')
```



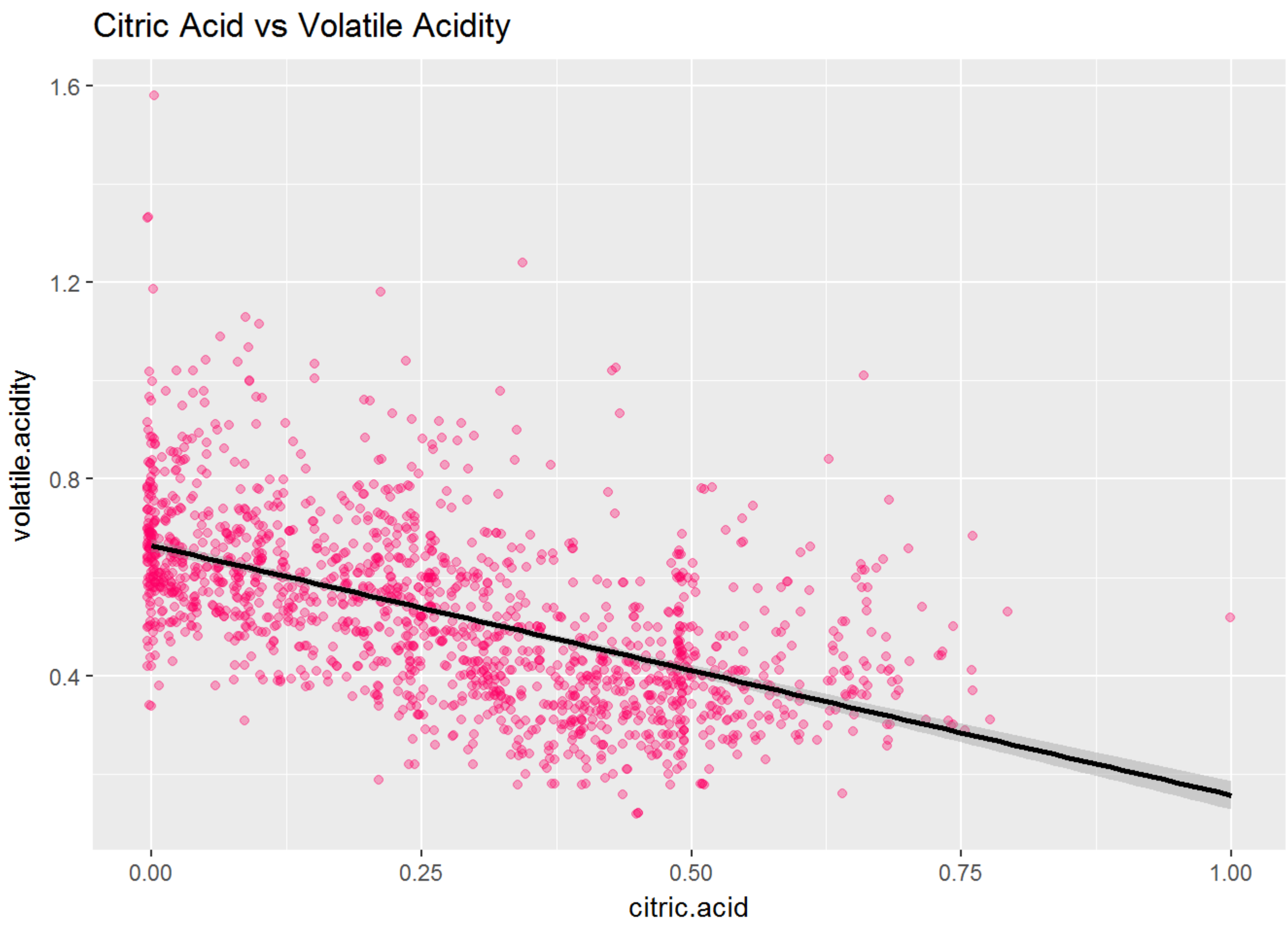
This plot of Density vs Fixed Acidity indicates that the correlation between the two variables is very strong i.e. 0.6680 and has a positive linear relationship.

```
ggplot(aes(x=pH,y=fixed.acidity),data=pr)+
  geom_jitter(color=('yellow'),alpha=1/3)+
  geom_smooth(method='lm',color='blue')+
  ggtitle('pH vs Fixed Acidity')
```



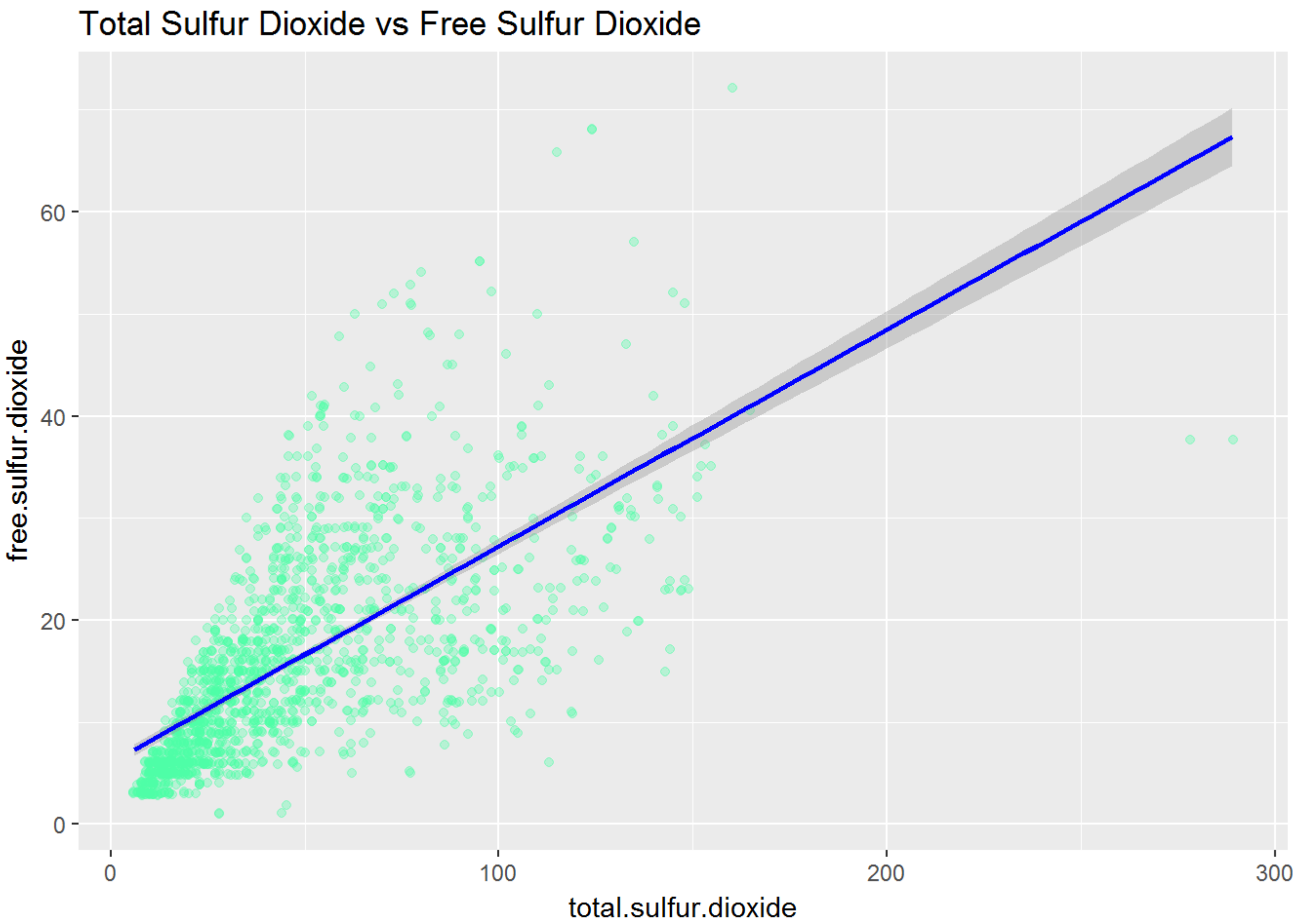
This plot of pH vs Fixed Acidity indicates that the correlation between the two variables is very strong i.e. -0.6829 and has a negative linear relationship.

```
ggplot(aes(x=citric.acid,y=volatile.acidity),data=pr)+
  geom_jitter(color=('#ff0066'),alpha=1/3)+
  geom_smooth(method='lm',color='black')+
  ggtitle('Citric Acid vs Volatile Acidity')
```



This plot of Citric Acid vs Volatile Acidity indicates that the correlation between the two variables is very strong i.e. -0.5524 and has a negative linear relationship.

```
ggplot(aes(x=total.sulfur.dioxide,y=free.sulfur.dioxide),data=pr)+
  geom_jitter(color=('#4dffa6'),alpha=1/3)+
  geom_smooth(method='lm',color='blue')+
  ggtitle('Total Sulfur Dioxide vs Free Sulfur Dioxide')
```



This plot of Total Sulfur Dioxide vs Free Sulfur Dioxide indicates that the correlation between the two variables is very strong i.e. 0.6676 and has a positive linear relationship.

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 histograms of variables with respect to quality did not reveal much information except that alcohol strongly affects wine quality.Through correlation test we created scatterplots that showed strong positive and negativecorrelation of variable with other. Alcohol is the positive correlation variable where as volatile acidity is the negative correlation variable. Therefore, we gather that when the quantity of variables like alcohol, citric acid and sulphates are high then the wine quality becomes high as well. And when the quantity of volatile acidity is low then also our quality of wine becomes high.

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

While using correlation to test the correlation between variables with each other we get that 1)citric acid and fixed acidity with value 0.6717 2)density and fixed acidity with value 0.6680 3)pH and fixed acidity with value -0.6829 4)citric acid and volatile acidity with value -0.5524 5)total sulfur dioxide and free sulfur dioxide with value 0.6676

What was the strongest relationship you found?

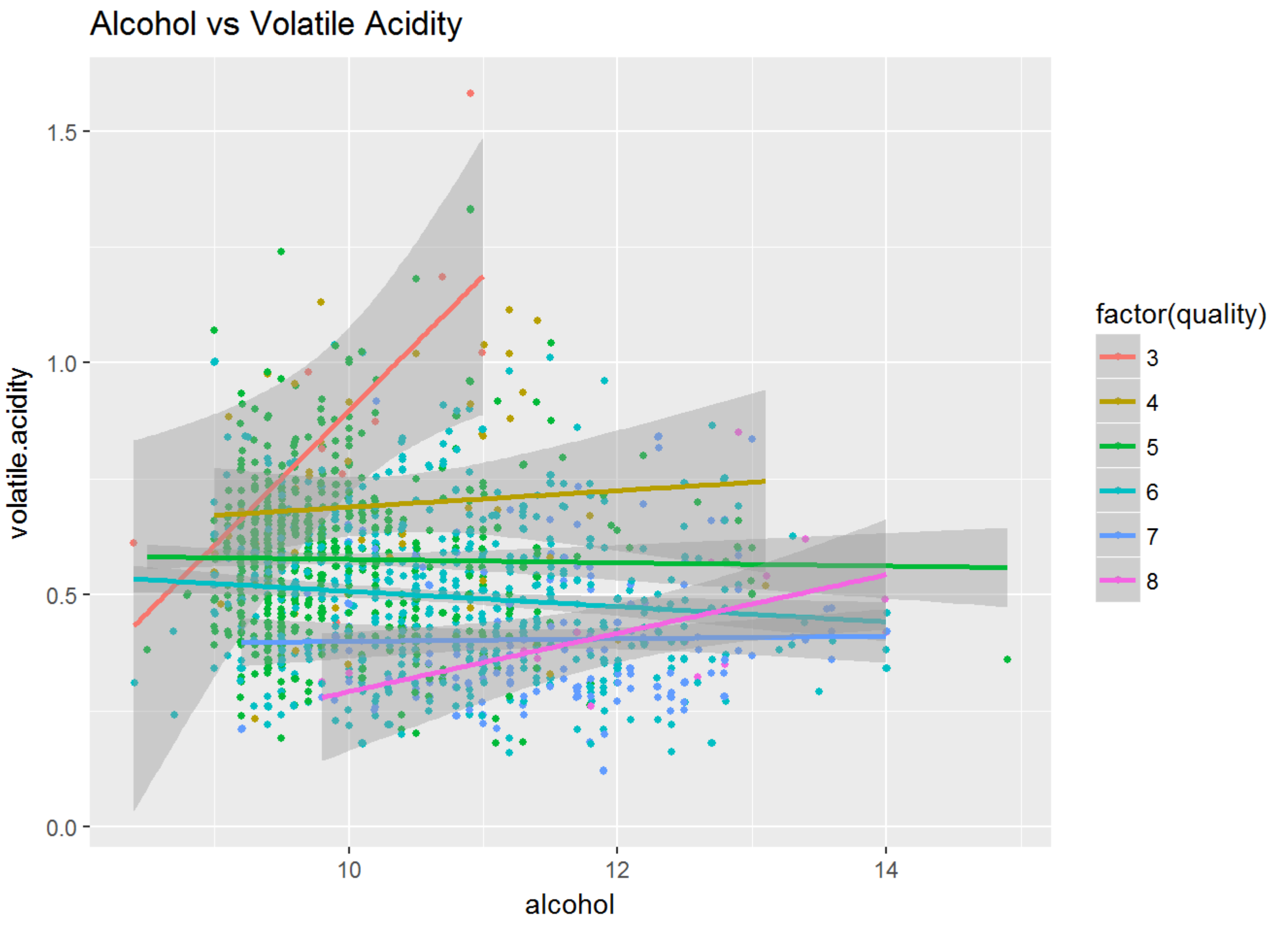
Throug correlation we found that citric acid and fixed acidity show us a strong positive correlation of 0.6717 whereas pH and fixed acidity show a strong negative correlation -0.6829

MULTIVARIATE PLOT SECTION

```
names(pr$quality)
```

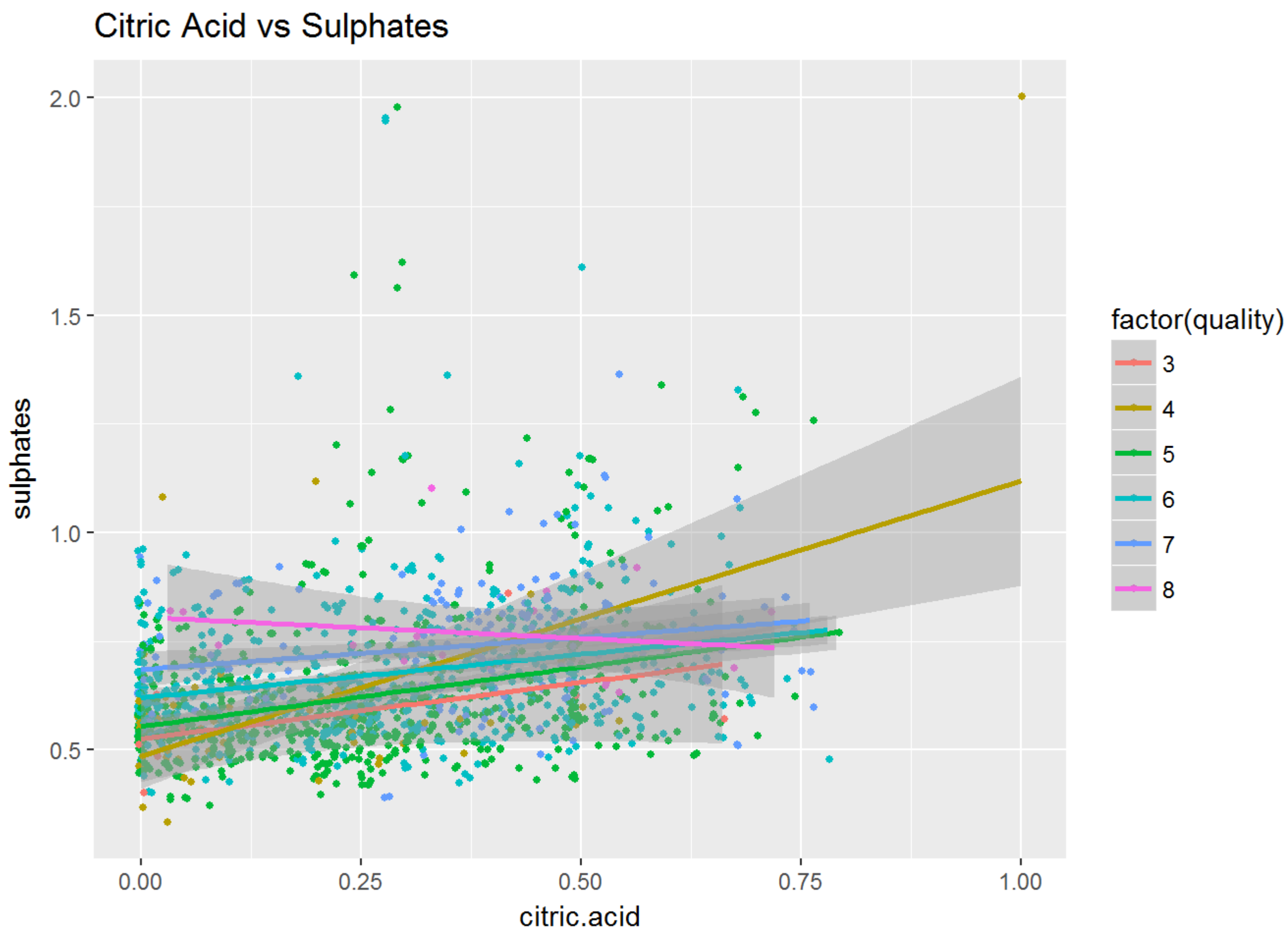
```
## NULL
```

```
ggplot(data=pr,aes(x=alcohol,y=volatile.acidity,color=factor(quality)))+
  geom_jitter(size = 1) +
  geom_smooth(method='lm')+
  ggtitle('Alcohol vs Volatile Acidity')
```



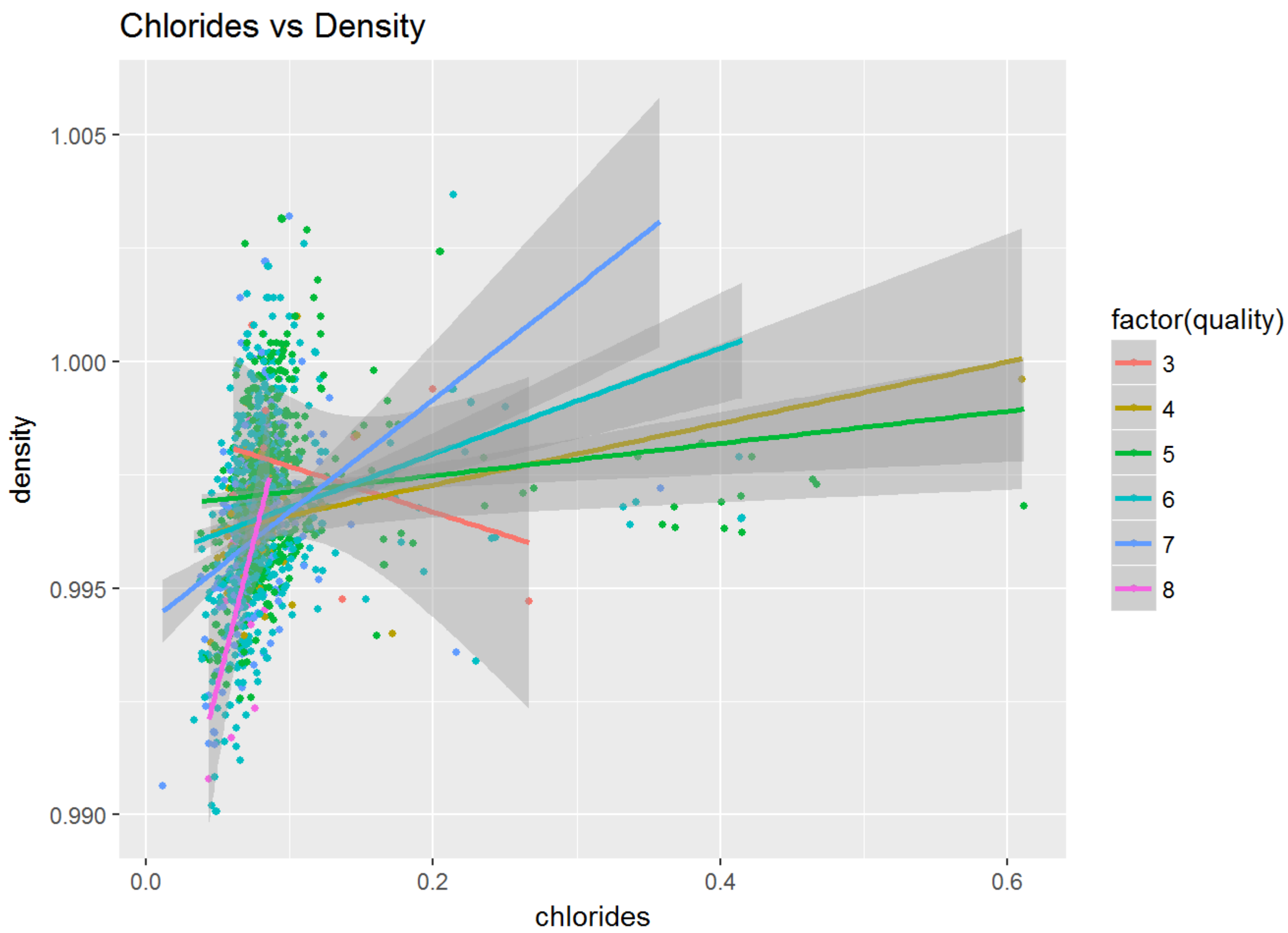
As Alcohol and Volatile Acidity have the highest correlation with quality. Thus, proving our point that, wine is of better quality when there is more alcohol and less volatile acidity.

```
ggplot(data=pr,aes(x=citric.acid,y=sulphates,color=factor(quality)))+
  geom_jitter(size = 1) +
  geom_smooth(method='lm')+
  ggtitle('Citric Acid vs Sulphates')
```



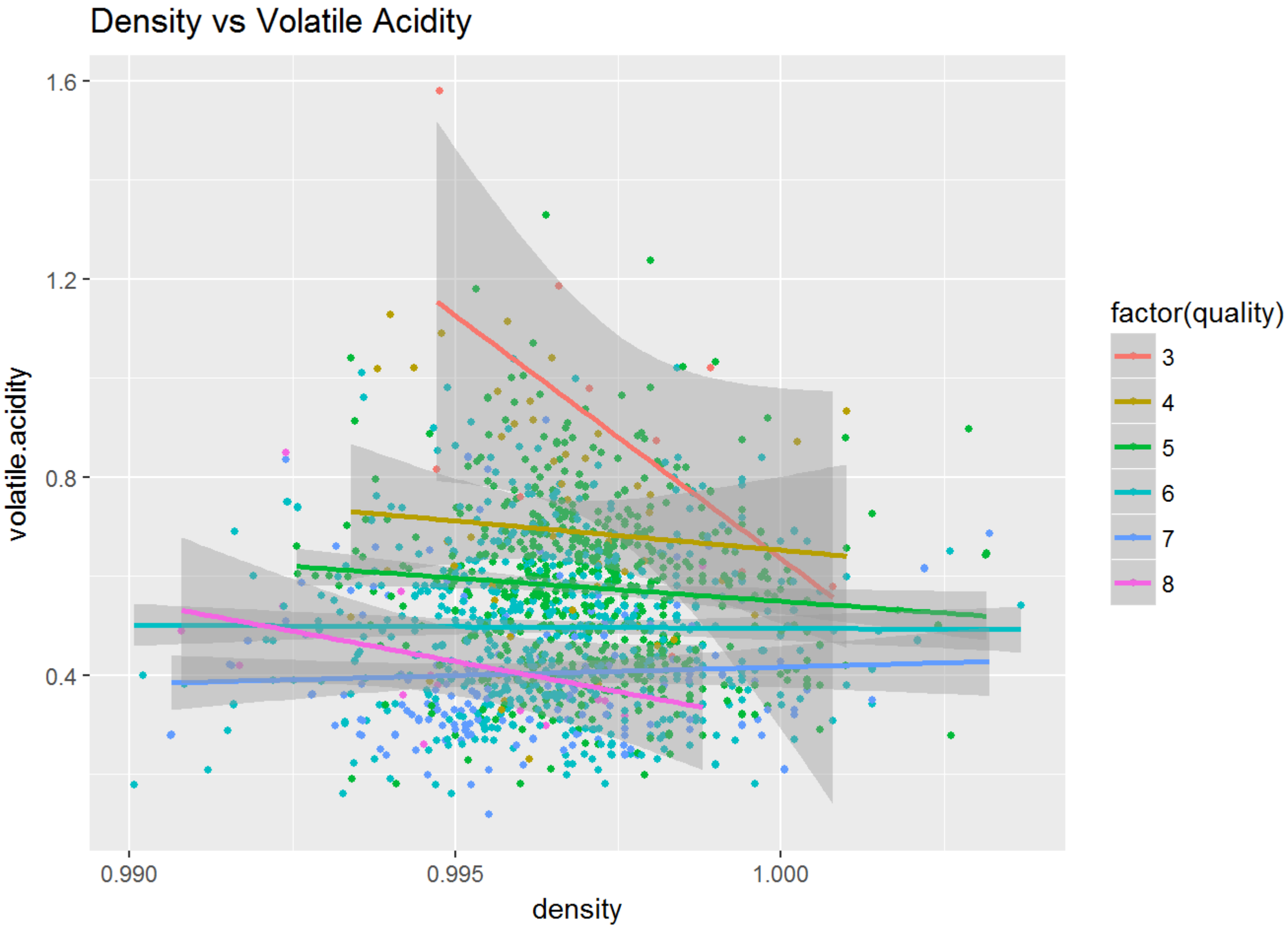
Sulphates and citric acid are the next two most correlated variables with quality. This plot is not quite a clear as the previous one. It appears that wine improves in quality with more sulphates.

```
ggplot(data=pr,aes(x=chlorides,y=density,color=factor(quality)))+
  geom_jitter(size = 1) +
  geom_smooth(method='lm')+
  ggtitle('Chlorides vs Density')
```



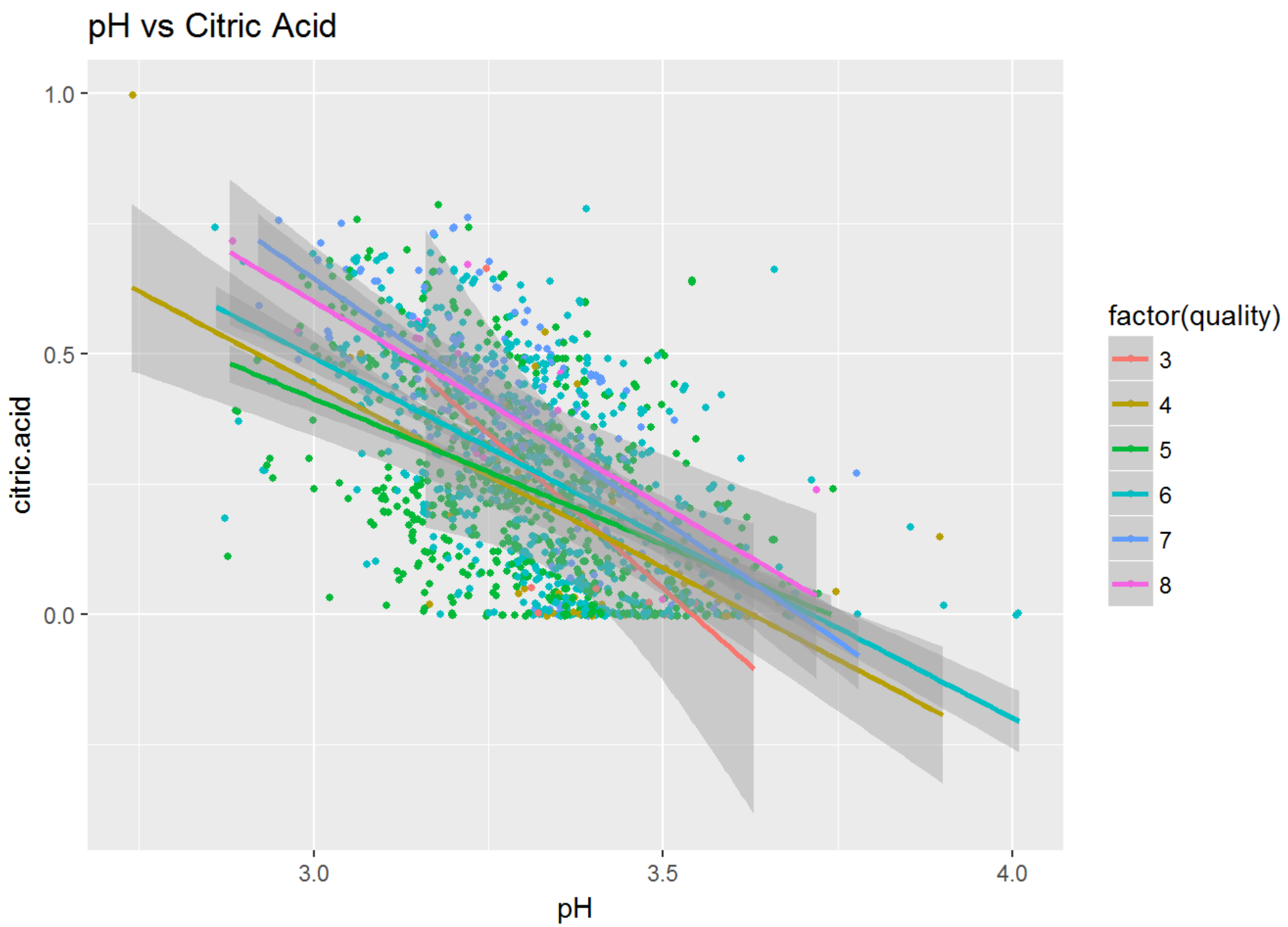
Chlorides and density have the weakest correlations with quality. But it appears that wine is best when it is less dense and has fewer chlorides.

```
ggplot(aes(y=volatile.acidity,x=density,color=factor(quality)),data=pr)+
  geom_jitter(size = 1) +
  geom_smooth(method='lm')+
  ggtitle('Density vs Volatile Acidity')
```



Density and Volatile Acidity are least correlated with each other. From the plot we gather that quality is highest when volatile acidity and density are lower.

```
ggplot(aes(y=citric.acid,x=pH,color=factor(quality)),data=pr)+
  geom_jitter(size = 1) +
  geom_smooth(method='lm')+
  ggtitle('pH vs Citric Acid')
```



As, pH has the weakest correlation with quality, and citric acid has the highest correlation with pH thus we draw this plot.

Multivariate 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?

In this analysis, the variables that were highly or least correlated were taken and faceted by quality variable as we as with each other. The good quality wines tend to have high content of citric acid, sulphates and alcohol and low content of volatile acid and density. The plot of pH vs Fixed Acidity shows us that the pH plays little role in the quality of wine.

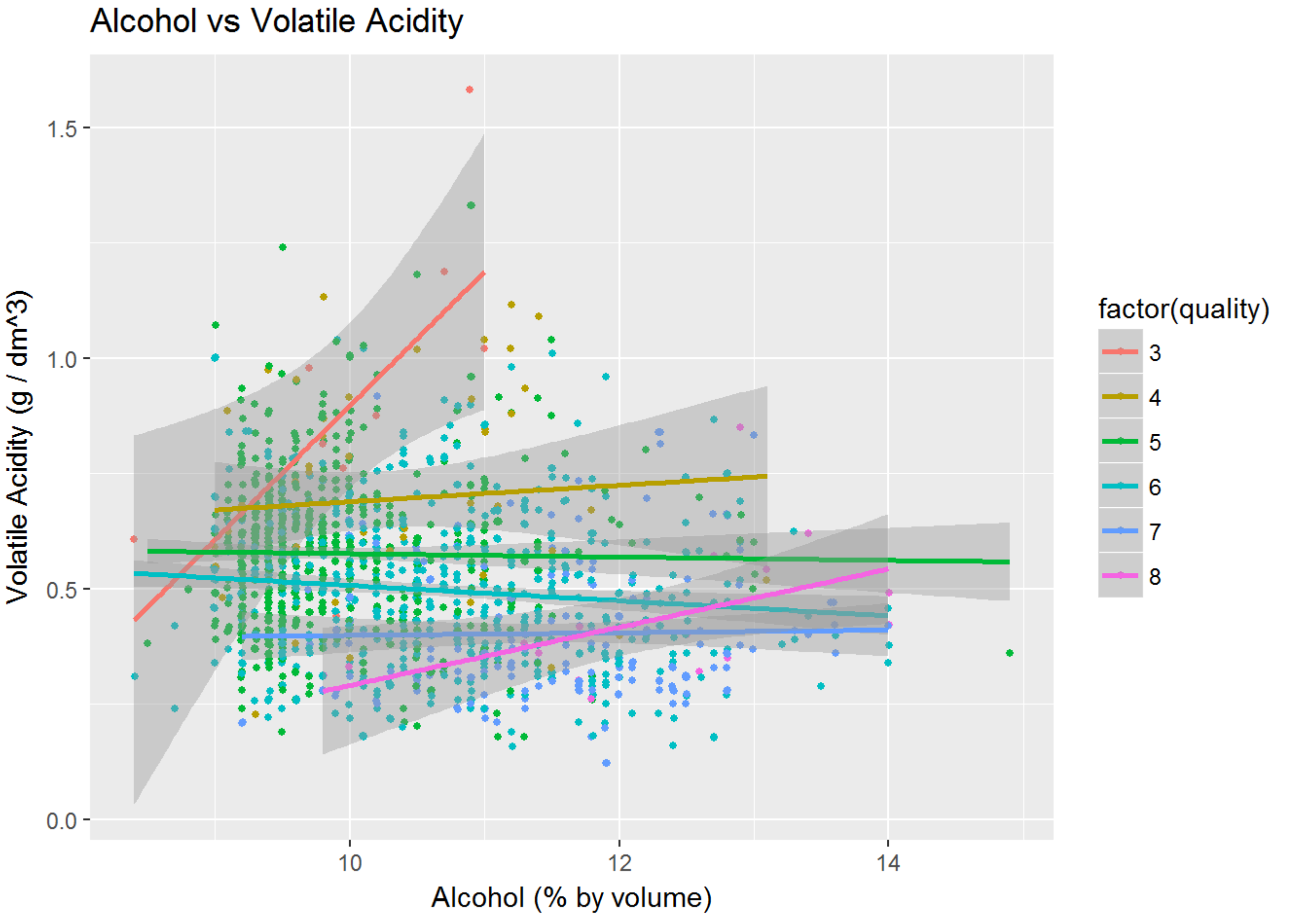
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

The surpising fact was that pH does not play a significant role in quality of wine, whereas every good quality wine has a pH level of 3-4 for a good wine.

FINAL PLOTS AND SUMMARY

Plot1

```
ggplot(data=pr,aes(x=alcohol,y=volatile.acidity,color=factor(quality)))+
  geom_jitter(size = 1) +
  geom_smooth(method='lm')+
  xlab('Alcohol (% by volume)')+
ylab('Volatile Acidity (g / dm^3)')+
  ggtitle('Alcohol vs Volatile Acidity')
```

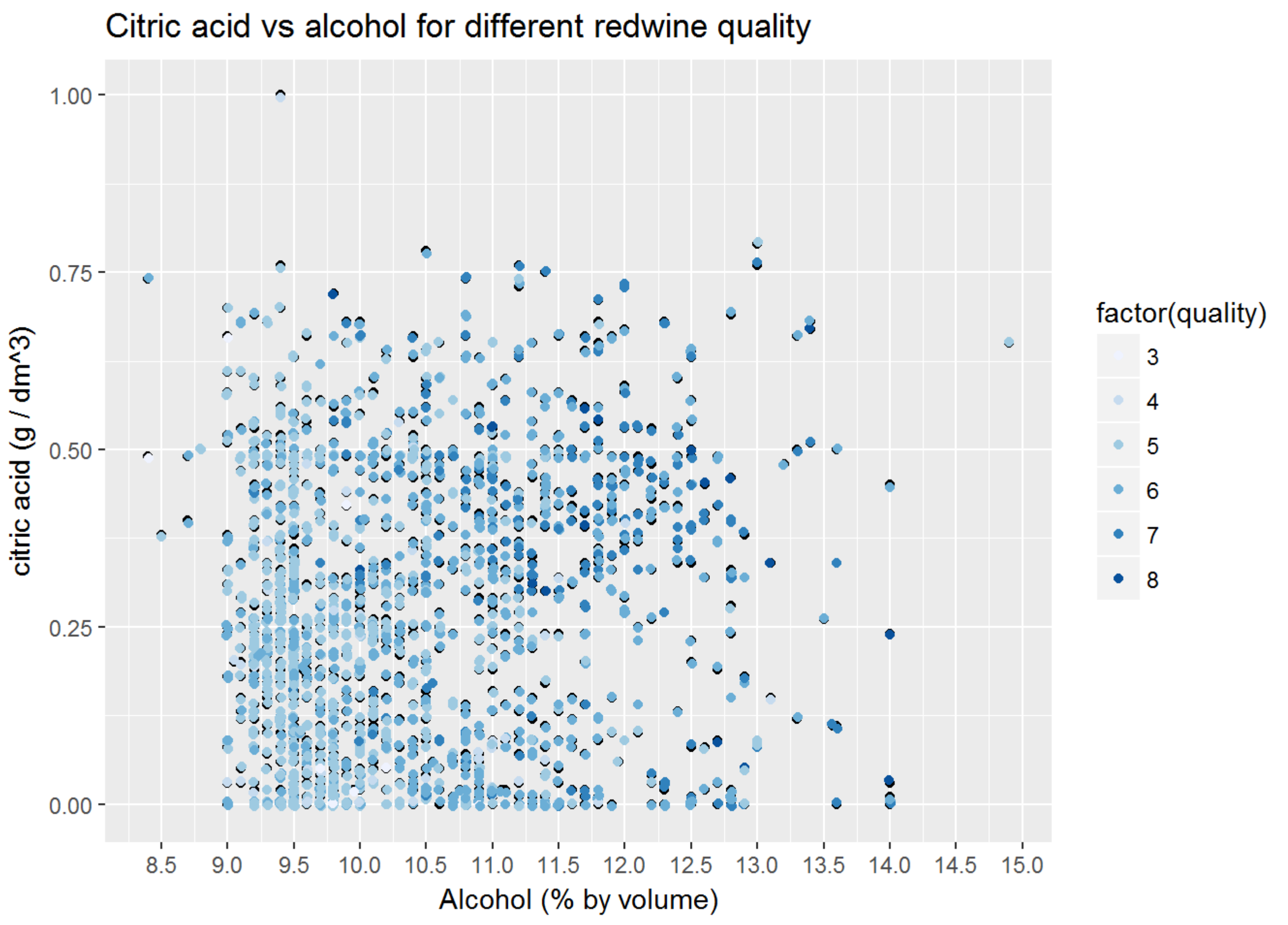


Description

In this plot we take alcohol and volatile acidity because these two variables effect the wine quality highly and Alcohol and Volatile Acidity have the highest correlation with quality.

Plot 2

```
ggplot(data=pr, aes(x = alcohol, y = citric.acid)) +
  geom_point() +
  geom_jitter(position = position_jitter(), aes(color = factor(quality))) +
  scale_color_brewer(type='seq') +
  xlab("Alcohol (% by volume)")+ylab("citric acid (g / dm^3)")+
  scale_x_continuous(breaks = seq(0,15,0.5))+
  ggtitle("Citric acid vs alcohol for different redwine quality")
```

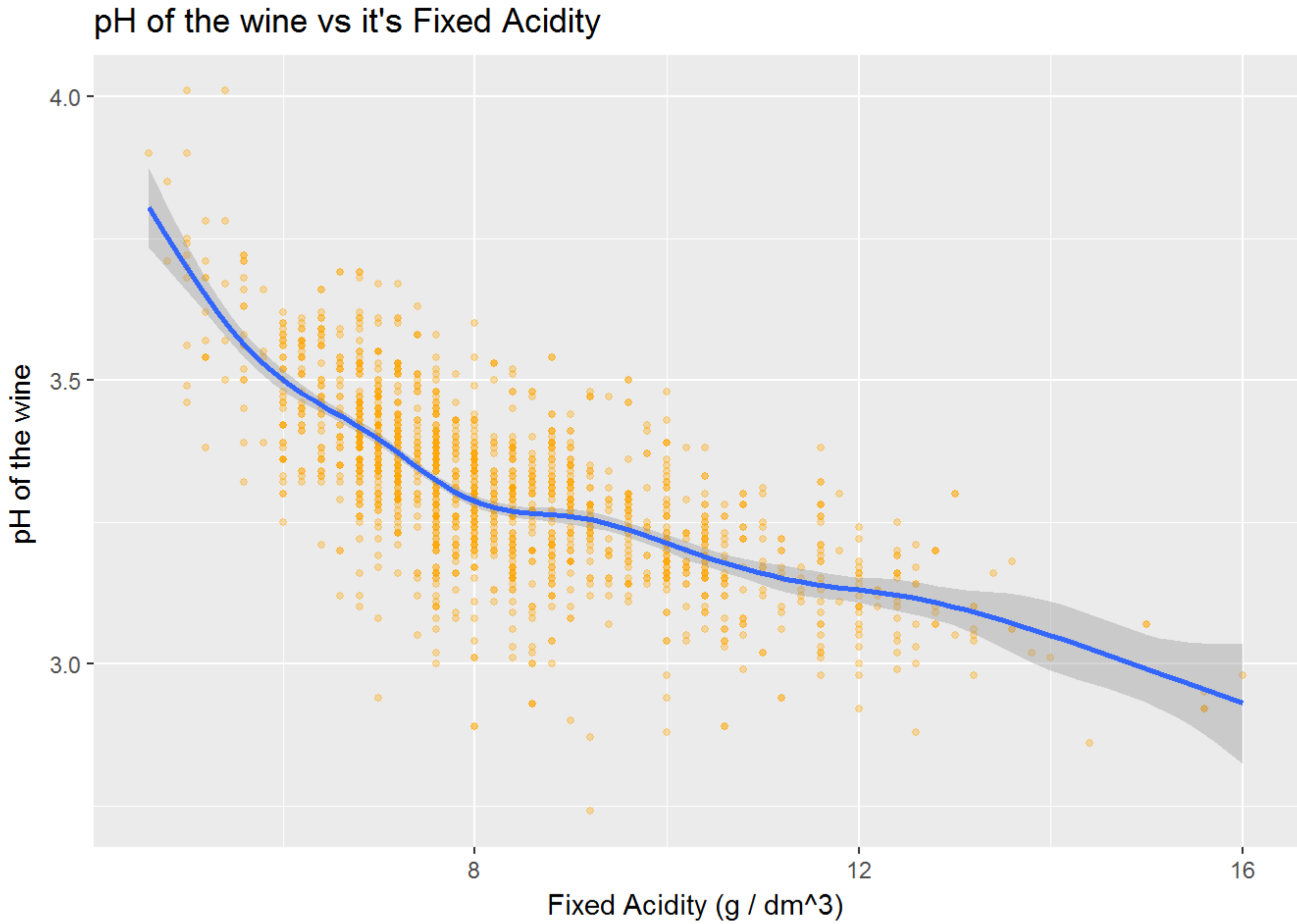


Description

This plot discusses the relation of alcohol and citric acid concentration on quality. Values above 11% by volume of alcohol and 0.25 of citric acid result in rating A wines which are quality 7 and above.

Plot 3

```
ggplot(data = pr, aes(round(fixed.acidity / .2) * .2, pH)) +
  geom_point(color = I('orange'), alpha = 1 / 3, size = 1) +
  geom_smooth() +
  xlab ("Fixed Acidity (g / dm^3)") +
  ylab ("pH of the wine ") +
  ggtitle("pH of the wine vs it's Fixed Acidity")
```



Description

We all know that ph and acid content have a relation among each other. A liquid is consumable only at a particular level of acidity. This graph help us show that innate relation between pH value of the wine and it’s Fixed Acidity. With enough correlation, we can say that the wine data do have some states that have same relation among which ph and acidity as it would have been normally accepted.

Reflection

This is a red wine data set with 1599 observations and 13 varaibles. The dataset is tidy and without attribute values. Quality is the output variable and remaing are the input variable which decide the yielding of a good quality wine. The struggles were that most of red wines in dataset have moderate rating i.e 5,6,7 and the low and high quality wines have very less data points hence bulk of data lied in moderate range hence there was a lot of overplotting in moderate section and very few information(data points) were available for low and high quality wines. Some of the data was skewed hence it needed to be transformed to get to the bulk of data, categorical variables needed to be created to carry out the analysis neatly. The most surprising relations I found were Volatile acidity, alcohol,citric acid and sulphates have a lot of effect in deciding the quality of wine. the density of wines lies in same range og 0.99 to 1. Later in analysis we found out that that low levels of volatile acidity and high percentage of alcohol yields better red wine. Future work can be done on dataset if more opinion of different people were taken to determine the quality of white wine as different people will rate wines how they felt about the taste and smell and then we may be able to make better predictions and analyse it more approprietly.