Red Wine Analysis by Michael Christensen

Univariate Plots Section

```
## [1] "Dataset Variables"
```

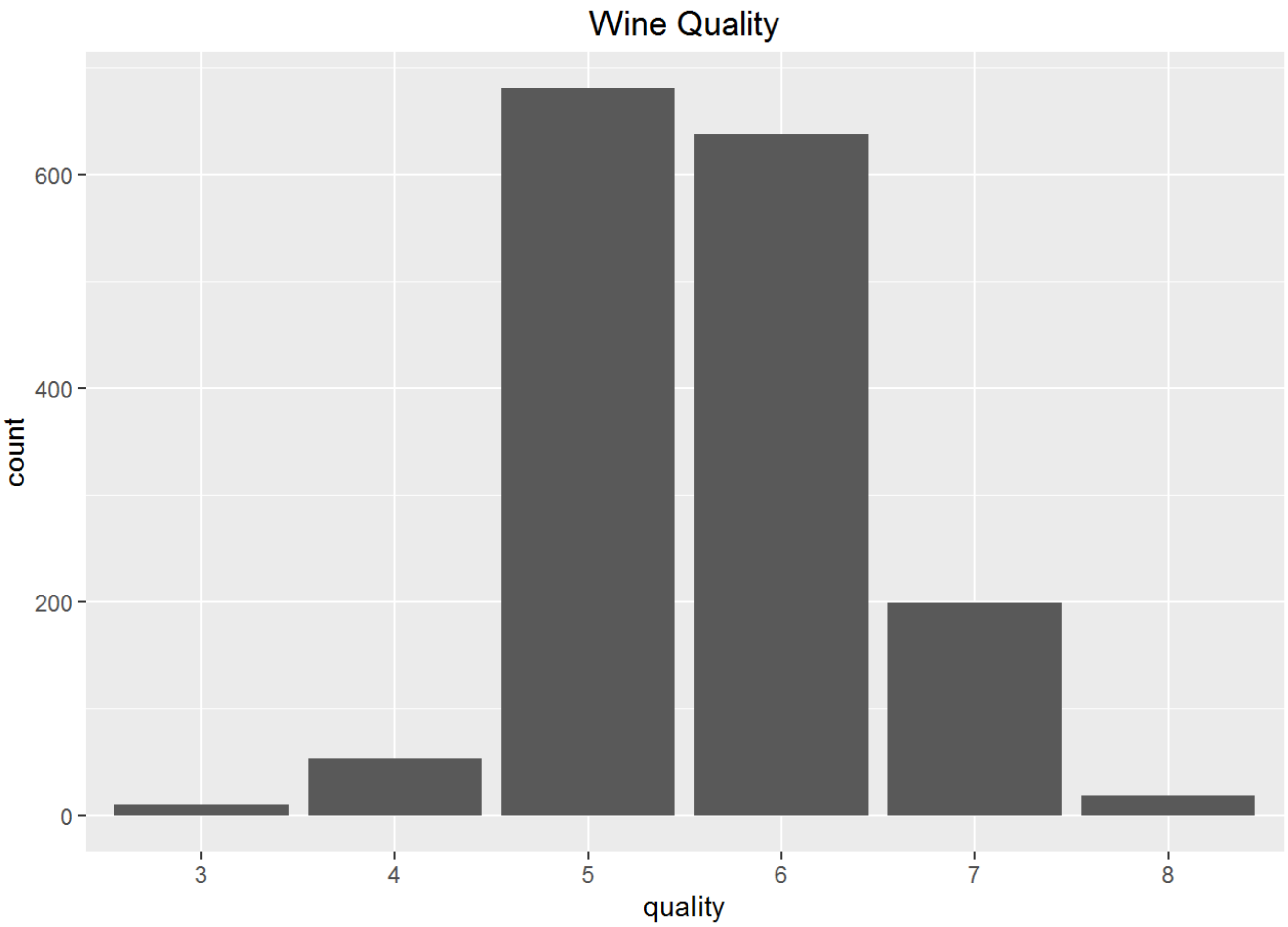
```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"         
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide" 
##  [7] "total.sulfur.dioxide" "density"              "pH"                  
## [10] "sulphates"            "alcohol"              "quality"  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

```
## [1] "Dataset structure"
```

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ 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             : Ord.factor w/ 6 levels "3"<"4"<"5"<"6"<..: 3 3 3 4 3 3 3 5 5 3 ...
##  $ rating              : Ord.factor w/ 3 levels "bad"<"average"<..: 2 2 2 2 2 2 2 3 3 2 ...
```

```
## [1] "Dataset Summary"
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar 
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500  
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00      
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00      
##  Median :0.07900   Median :14.00       Median : 38.00      
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47      
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00      
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00      
##     density             pH          sulphates        alcohol      quality
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40   3: 10  
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   4: 53  
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   5:681  
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42   6:638  
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   7:199  
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90   8: 18  
##      rating    
##  bad    :  63  
##  average:1319  
##  good   : 217  
##                
##                
## 
```

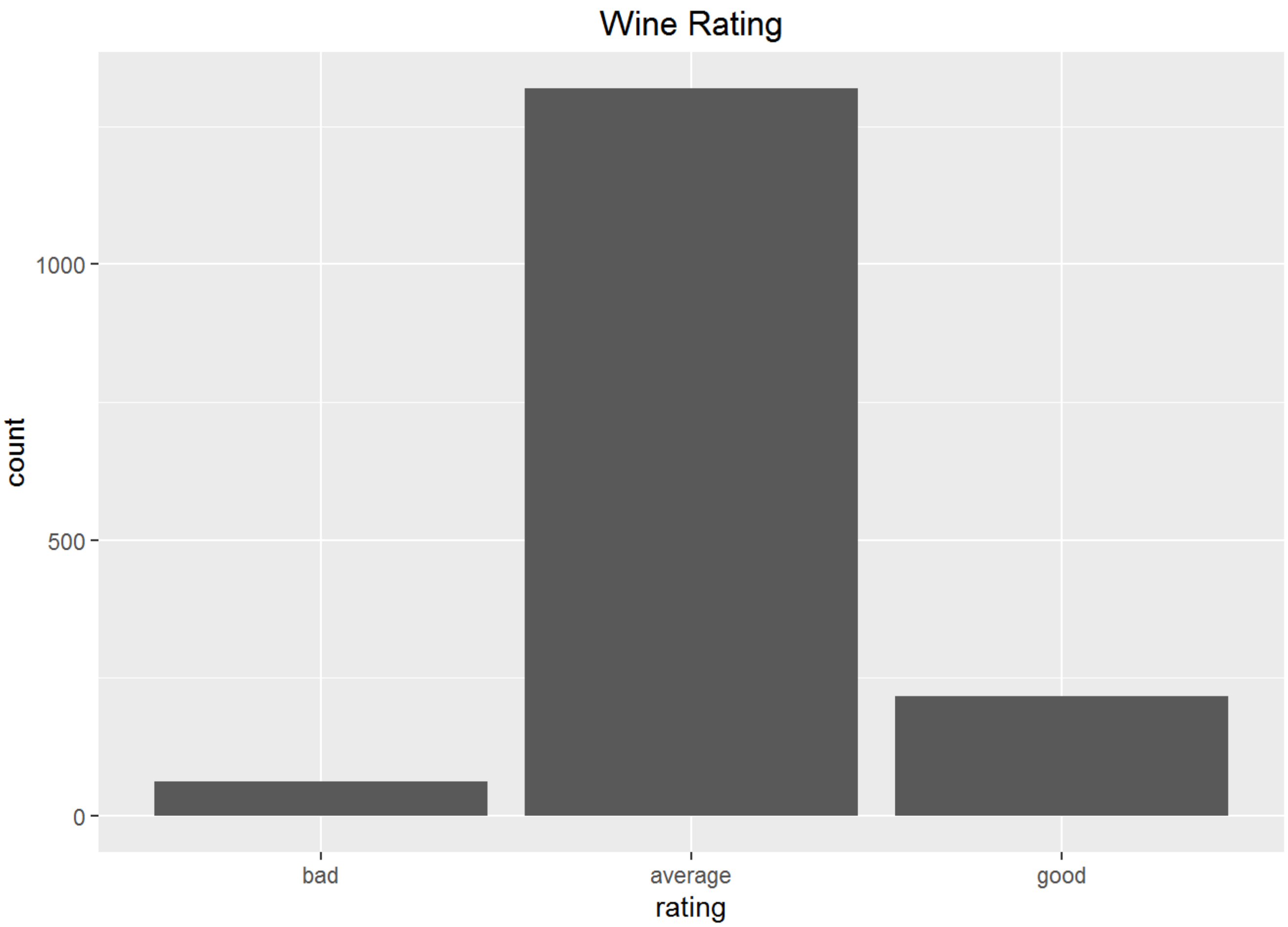
```
## [1] "Explore Each Variable To Better Understand The Data"
```



```
## [1] "Wine Count By Quality Category"
```

```
## 
##   3   4   5   6   7   8 
##  10  53 681 638 199  18
```

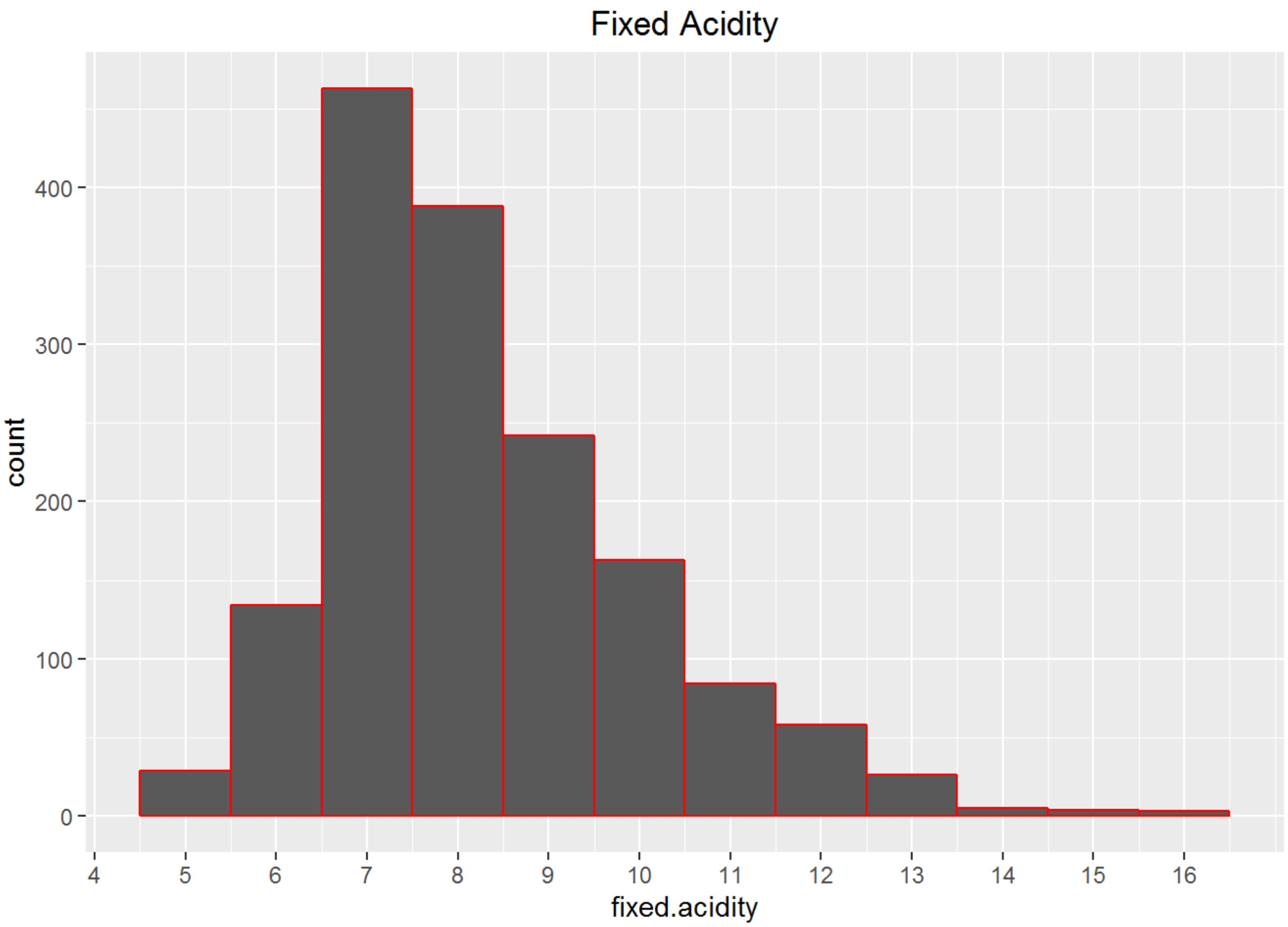
```
## The majority of the wines can be found in the 5 to 7 quality range.
```



```
## [1] "Wine Count By Rating Category"
```

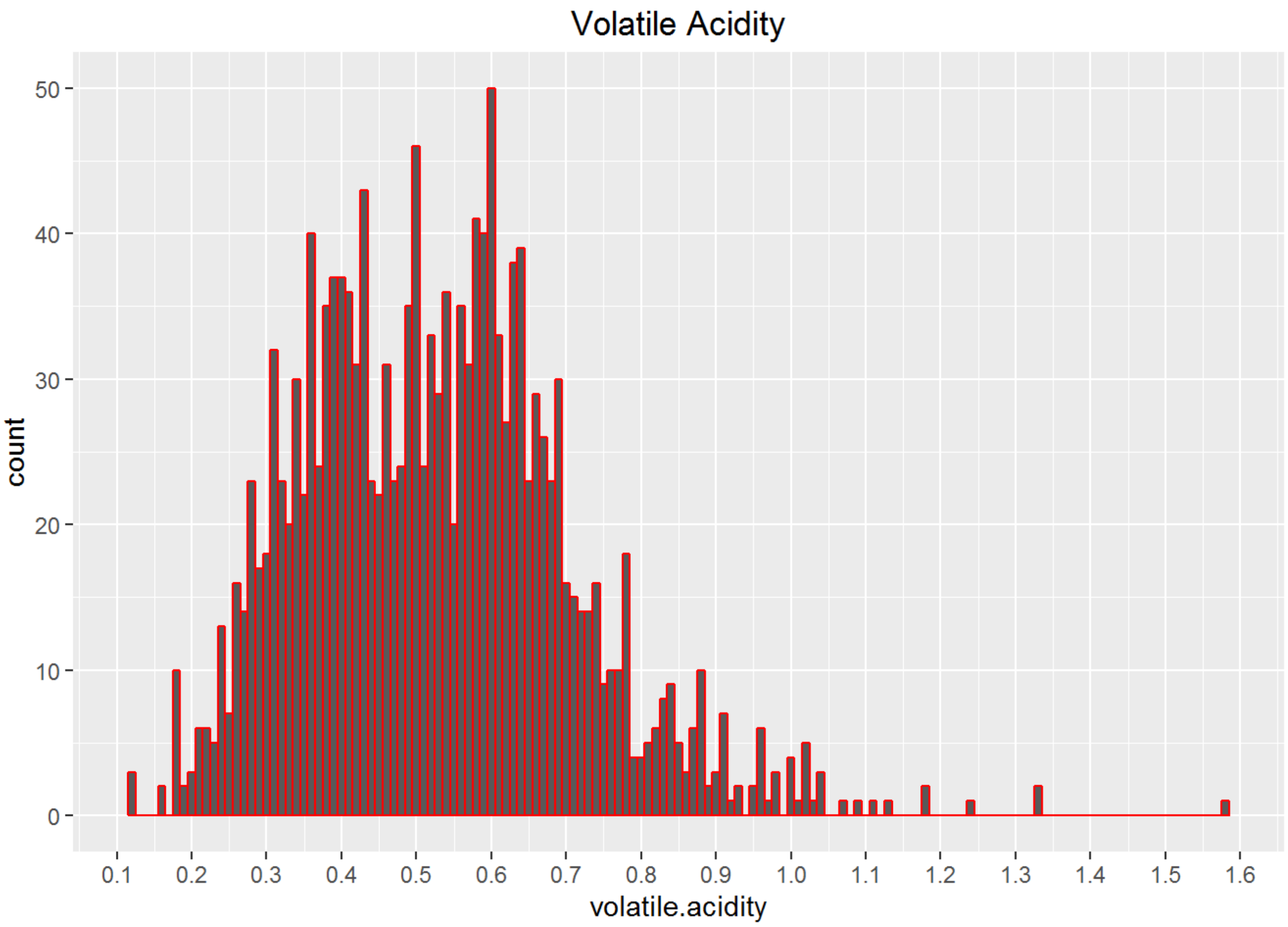
```
##
##     bad average    good
##      63    1319     217
```

```
## Most wines fall in to the average category (better than or equal to 5 and less than 7).
```



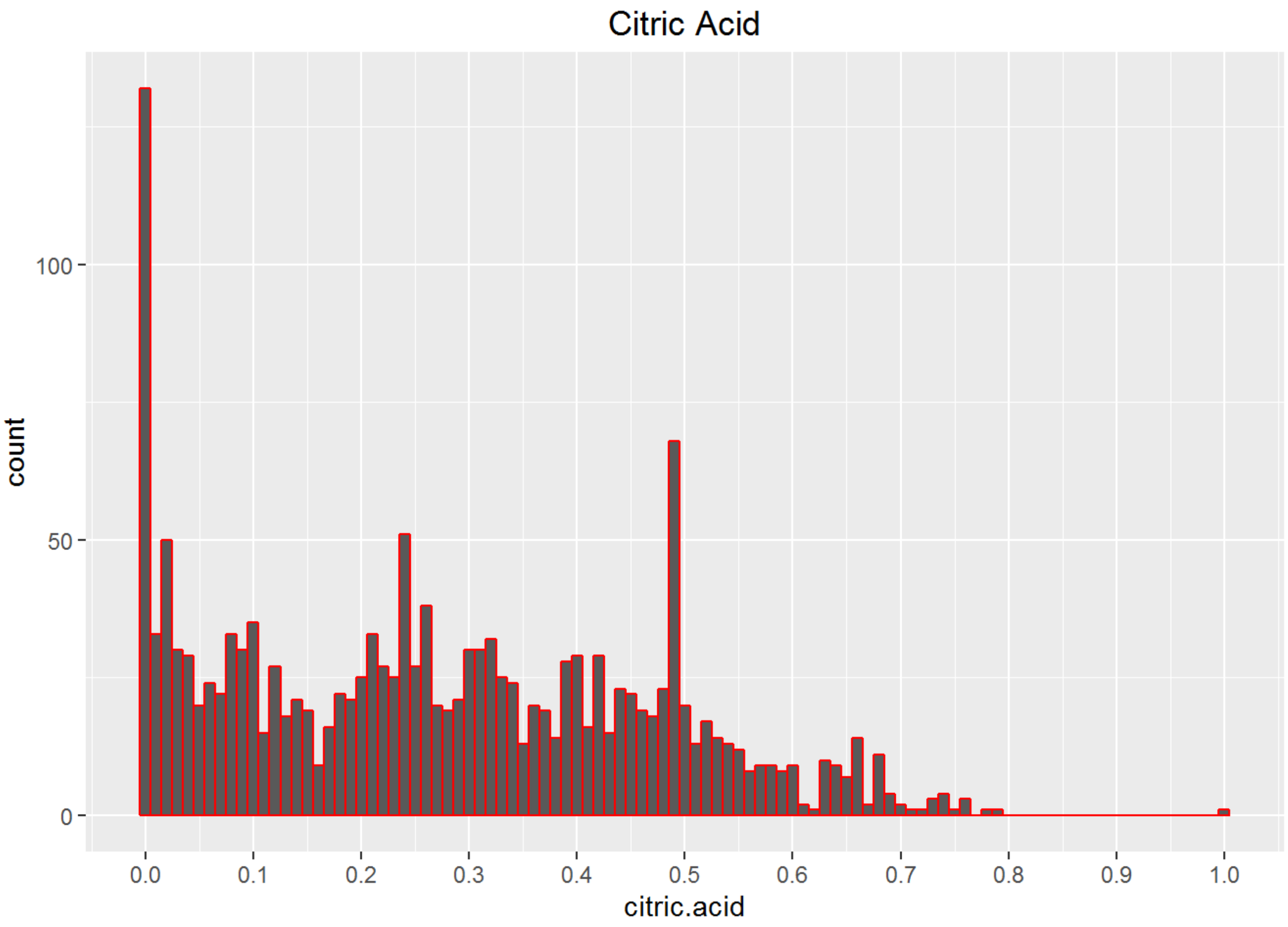
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    4.60    7.10    7.90    8.32    9.20   15.90
```

```
## A large number of wines have a fixed acidity close to the median (7.9 g / dm^3). Outliers are
 responsible for a higher mean than median.
```



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

```
## A bimodal distribution is apparent with peaks at 0.4 and 0.6 g/dm^3. Outliers are also presen
t in this dataset.
```

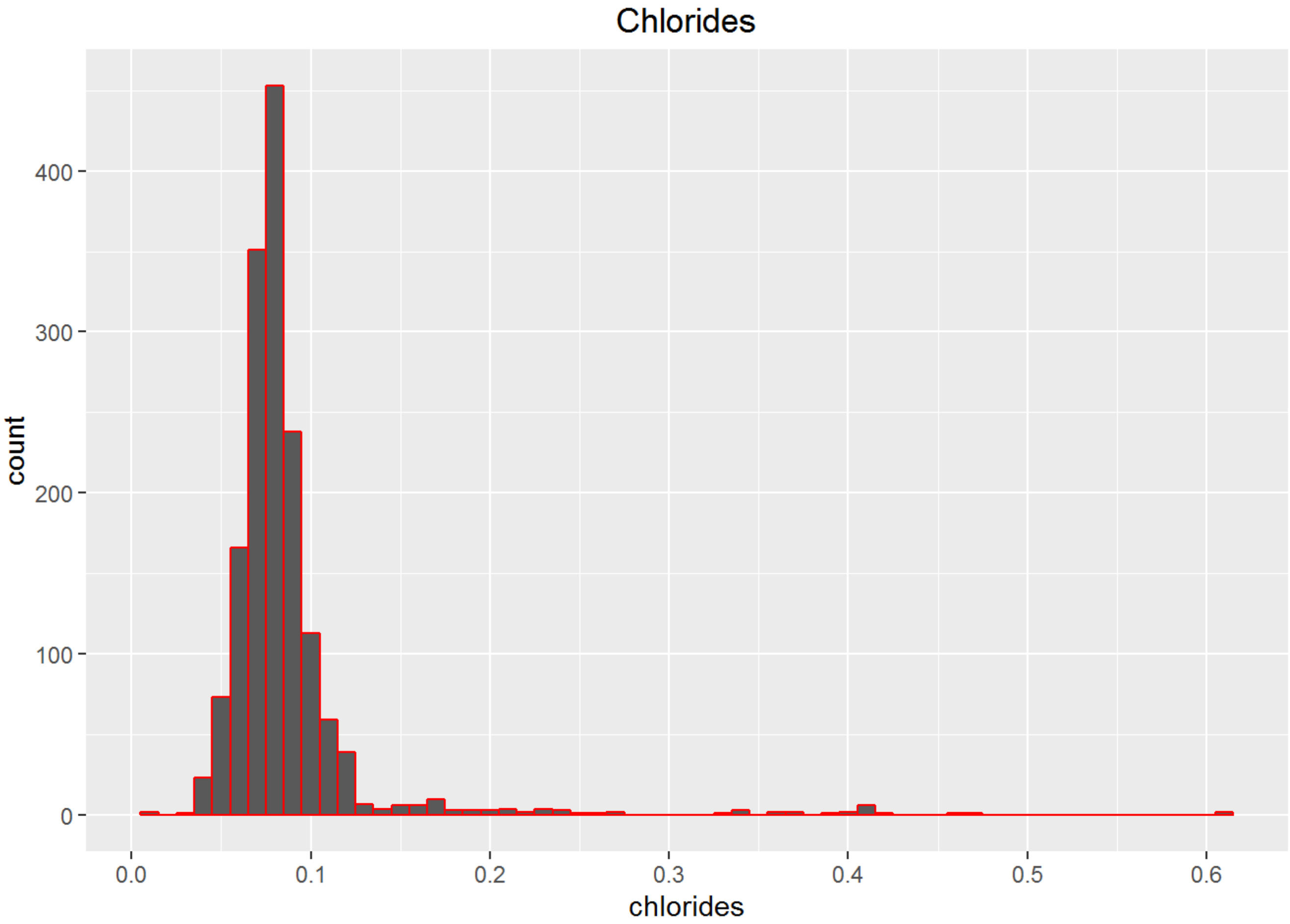


```
## [1] "Number of Samples with citric.acidity=0"
```

```
##       x freq
## 1 FALSE 1467
## 2  TRUE  132
```

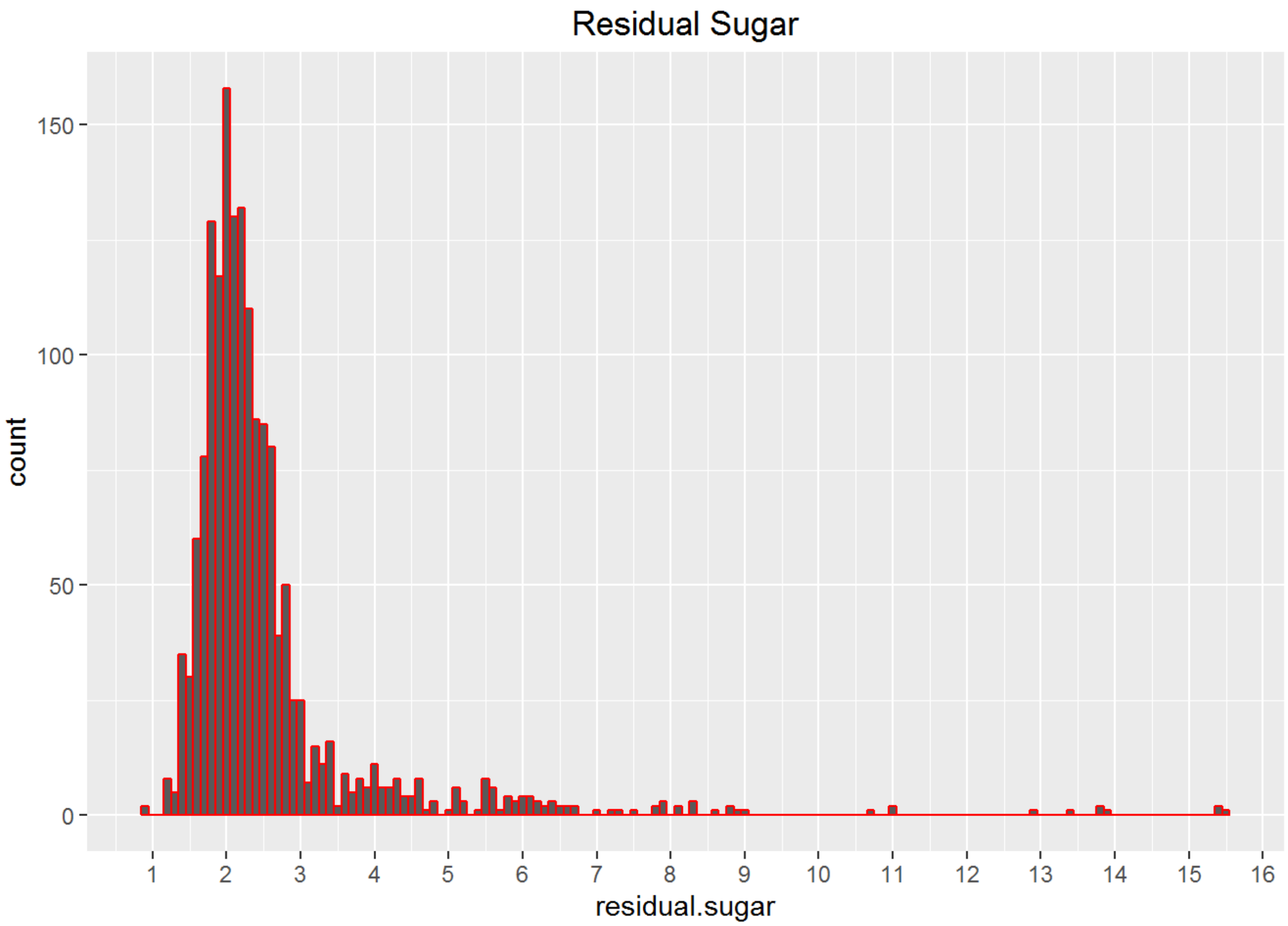
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.090   0.260   0.271   0.420   1.000
```

```
## This is a strange looking distribution. It is noteworthy that 132/1467 of the wines have a va
lue of 0 for citric.acid. There are also peaks at .02, .24, and .49 g/dm^3. One wine sample had
 a citric.acid value equal to 1.
```



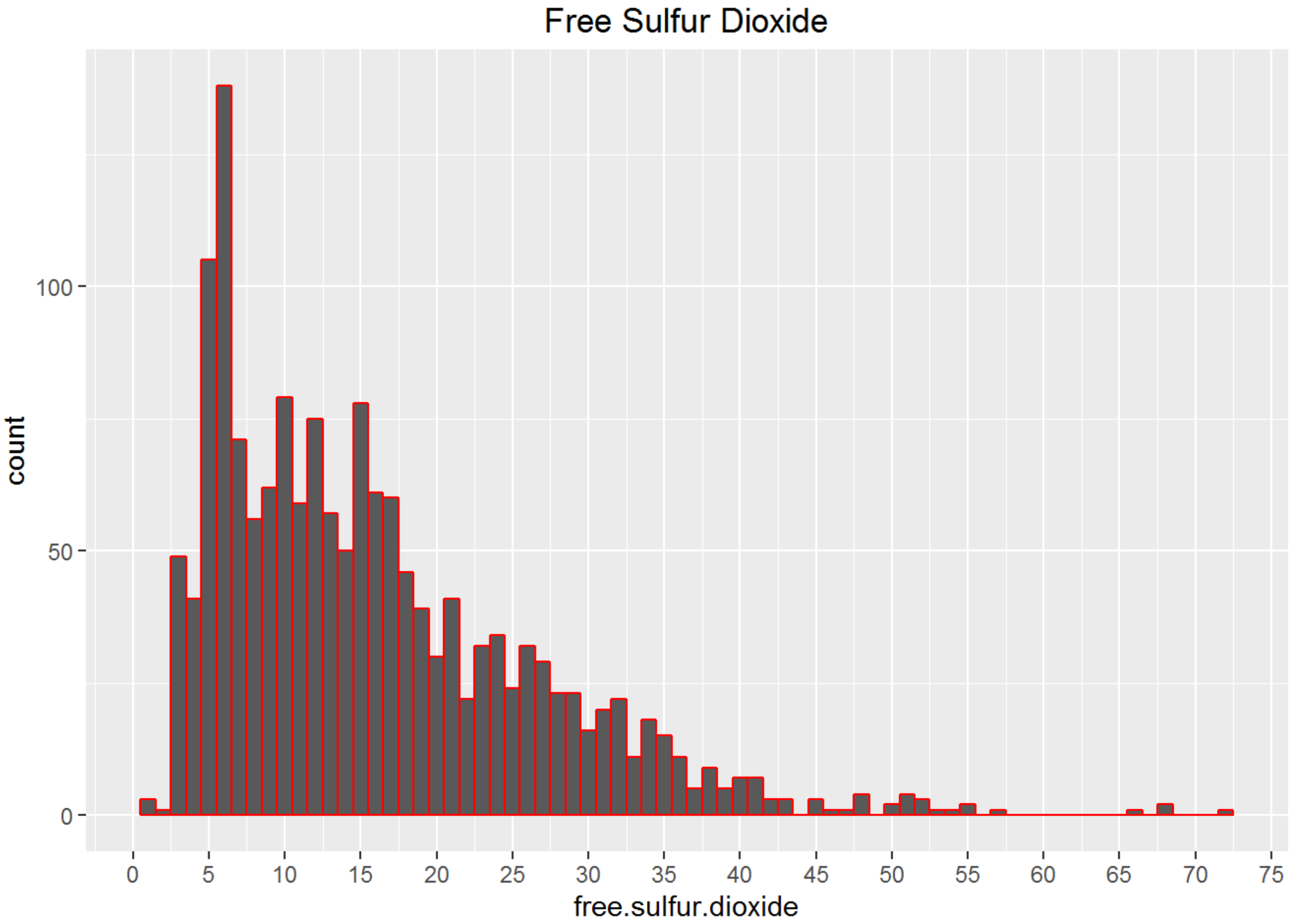
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

```
## There are a high concentration of wines around 0.079 g/dm^3 (the median). Some outliers are p
resent in the higher ranges.
```



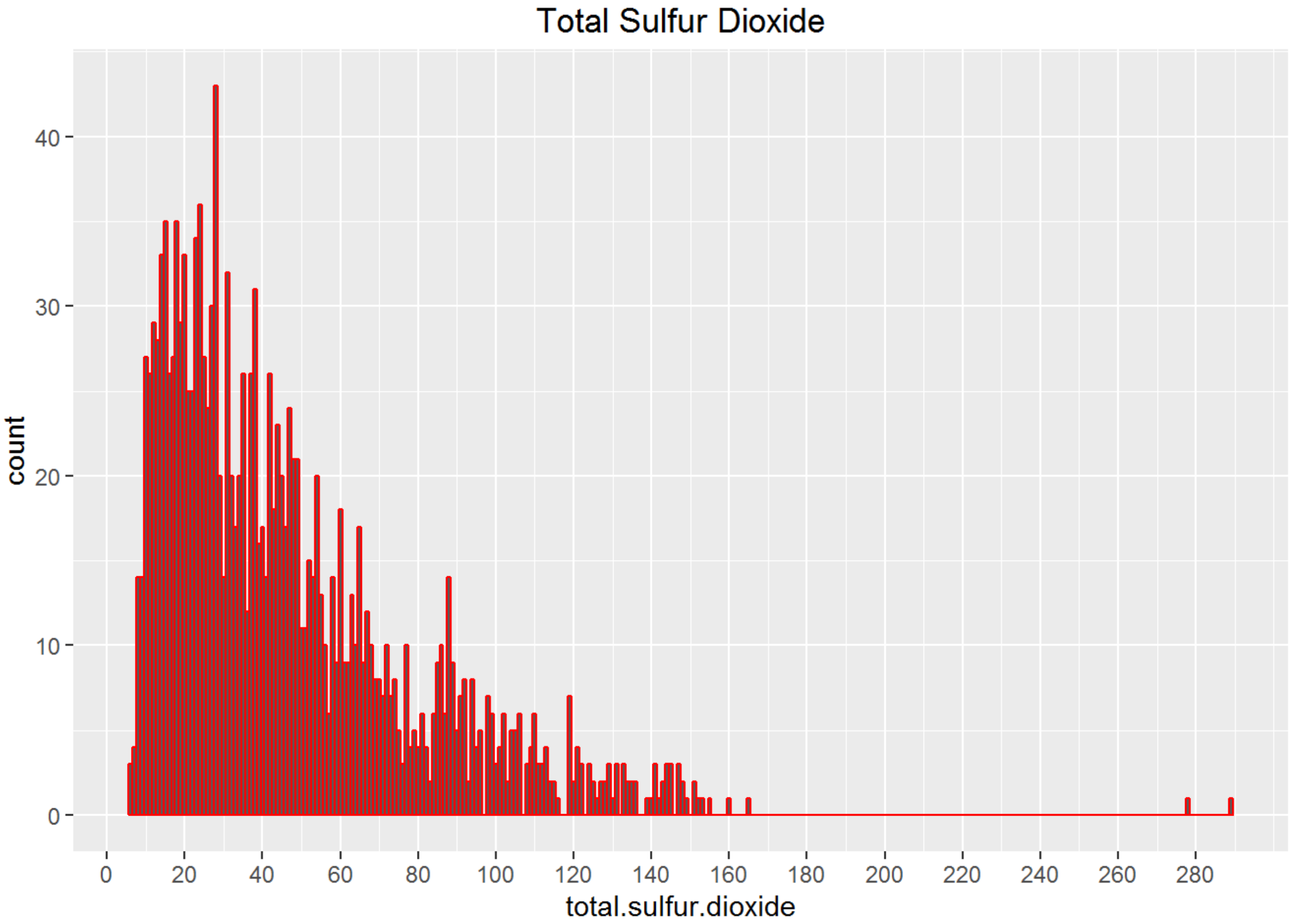
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

```
## There are a high concentration of wines around 2.2 g/dm^3 (the median). Some outliers are pre
sent in the higher ranges.
```



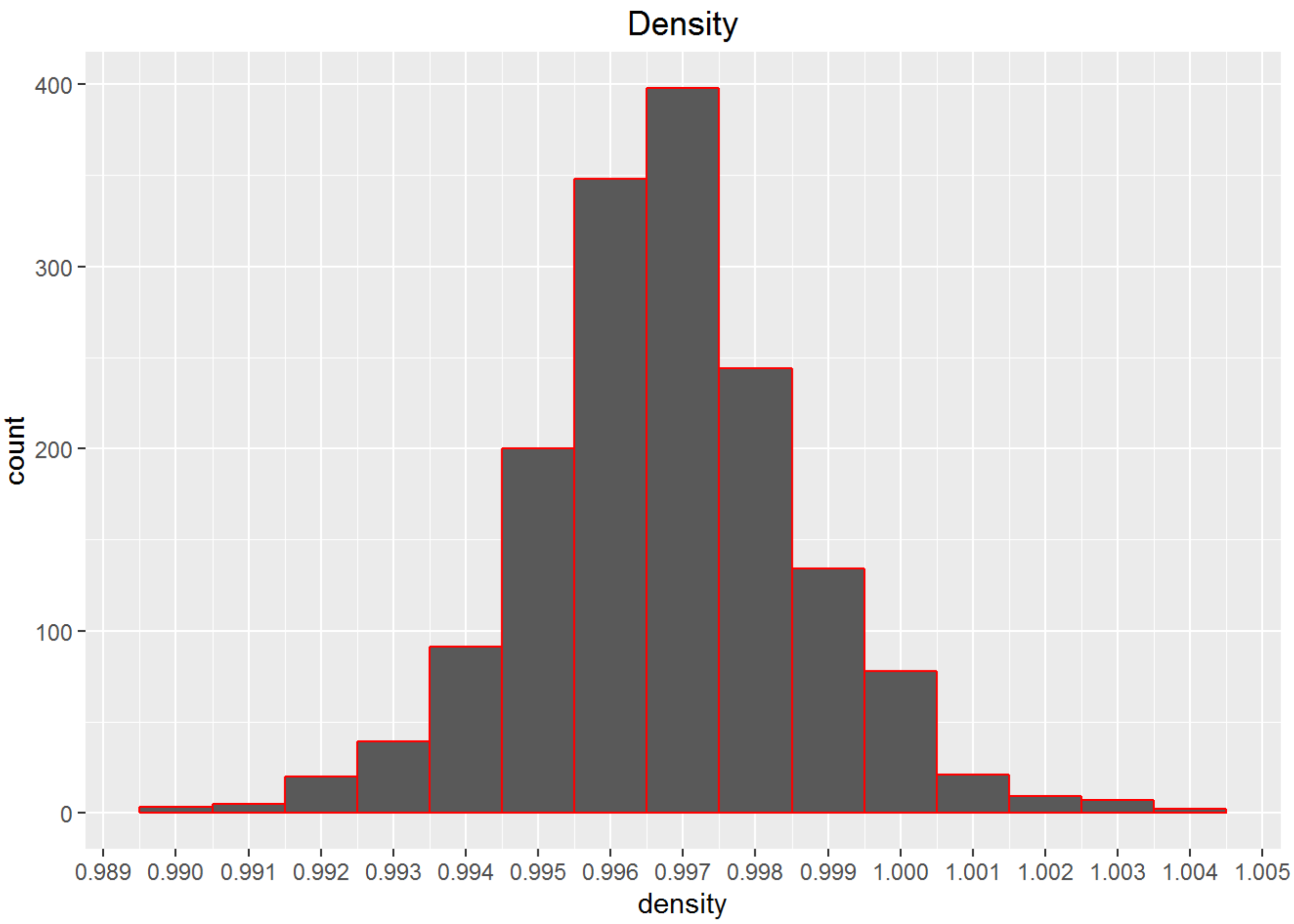
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00    7.00   14.00   15.87   21.00   72.00
```

```
## There is a peak at 6 mg/dm^3. The distribution has a long right tail.
```



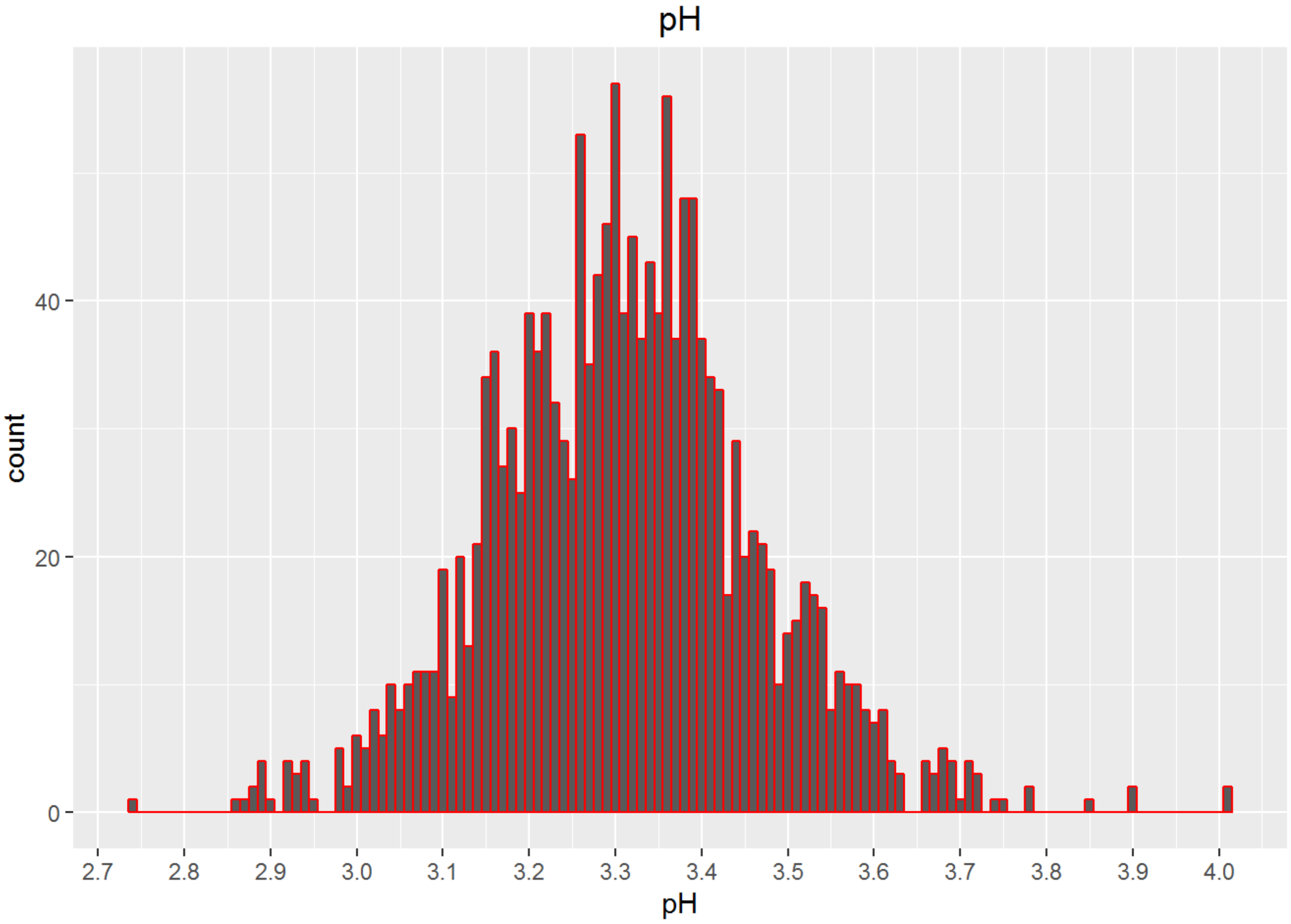
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.00   22.00   38.00   46.47   62.00  289.00
```

```
## This distribution has a long right tail. Some outliers are also apparent.
```



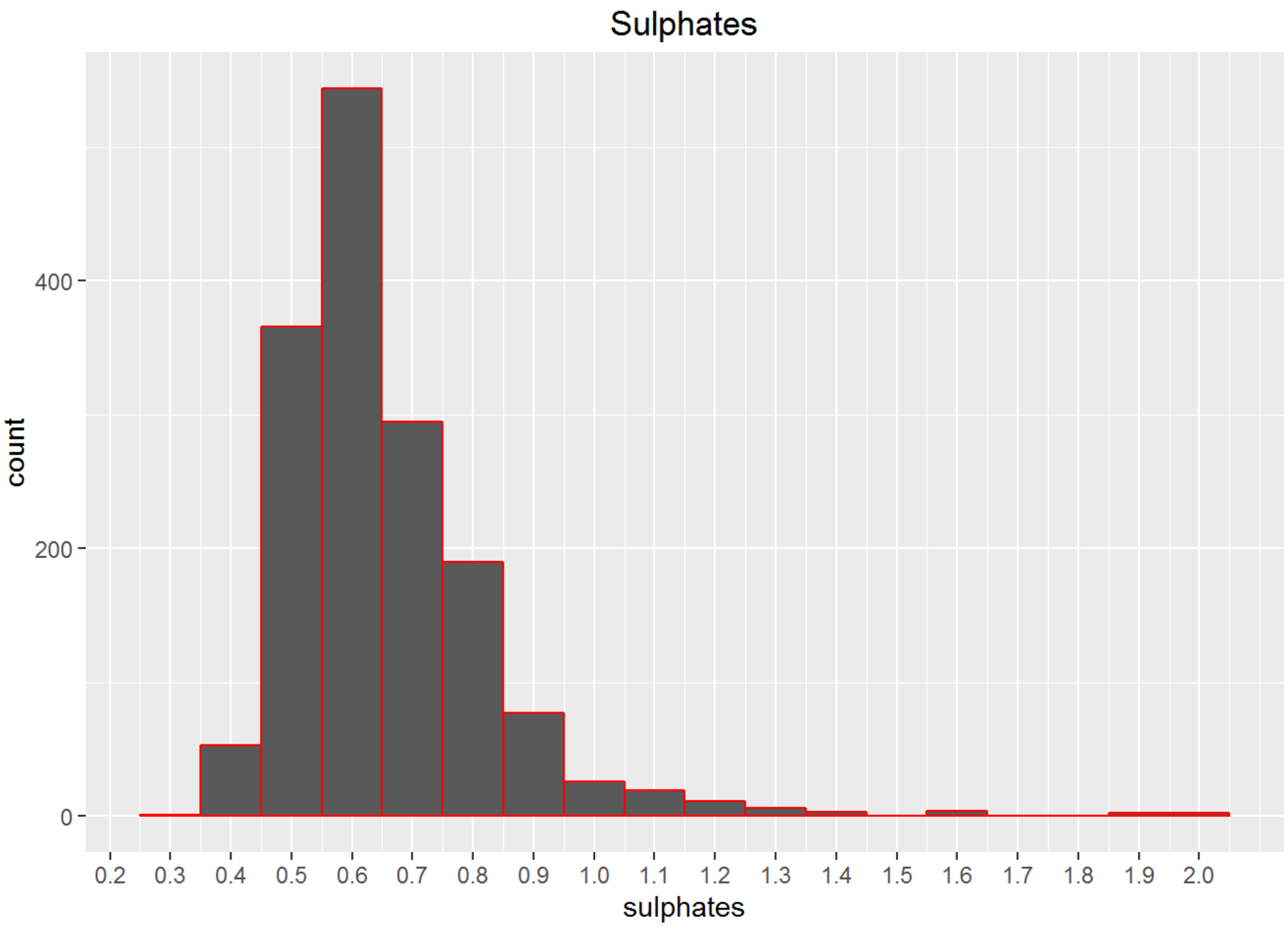
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```

```
## [1] "This distribution looks roughly normal."
```



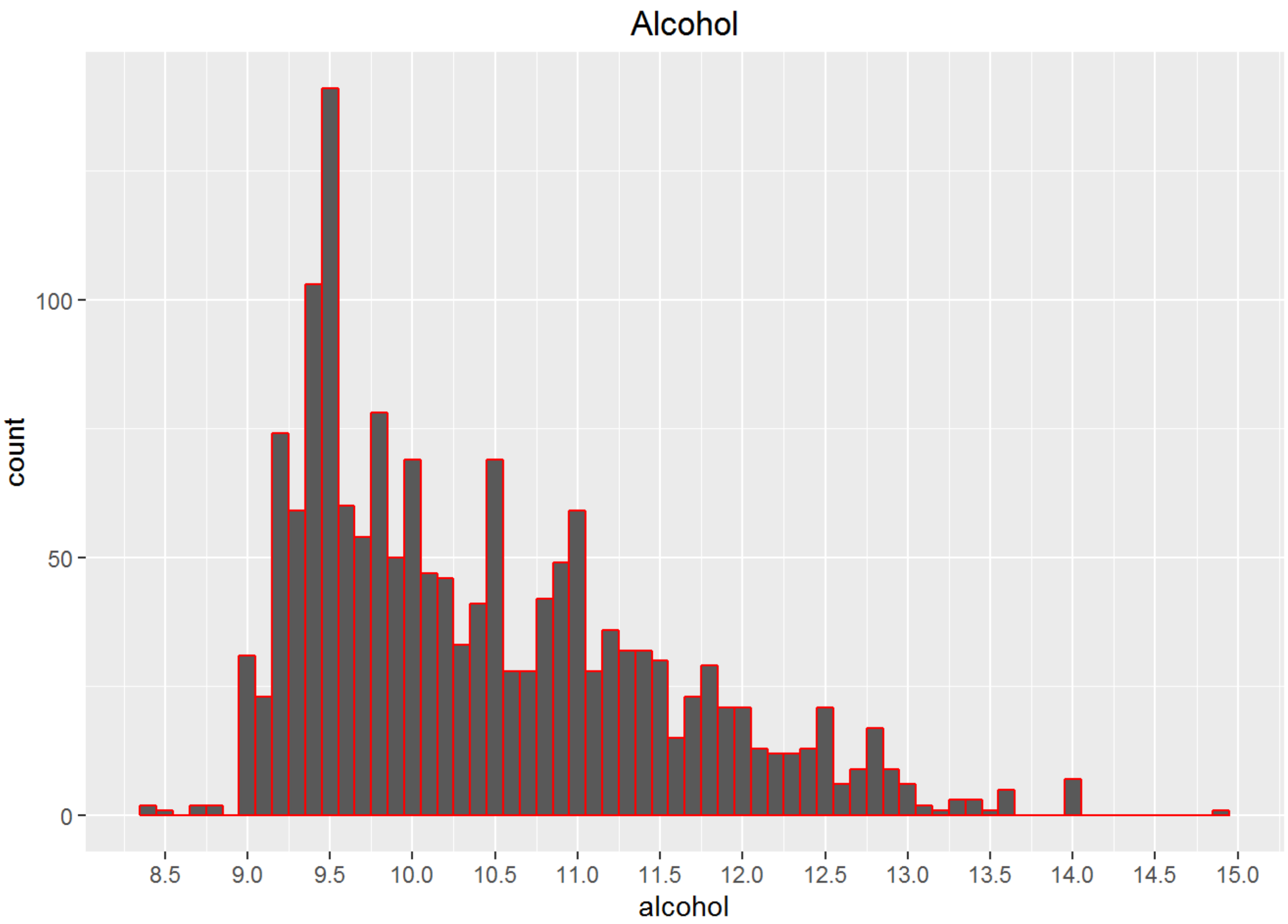
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

```
## This distribution looks roughly normal. However, there are some outliers on both the high and
 low ends.
```



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

```
## A large number of samples fall close to the median (0.62 g/dm3). The shape of the distributio
n is comparable to that for residual.sugar and chlorides.
```



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

```
## The distribution is long-tailed and has a peak around 9.5.
```

Univariate Analysis

What is the structure of your dataset?

This red wine dataset has 12 features (I added a 13th: rating). There are 1599 observations. I transformed the quality feature from an integer type to an ordered factor (categorical variable).The rating is also an ordered factor. The other features are numerical type.

Variable Names– fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol quality, rating

(worst) ——> (best) rating: bad, average, good quality: 3, 4, 5, 6, 7, 8

Other Observations: The median wine quality is 6. There are far more wine samples of average quality than bad or good. In fact, 82.4% of wine samples fall into the average category. Thissampling distribution may make predictive modeling difficult. It is also interesting to note that many variables in this dataset have non-normal distributions with longer right-hand tails.

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

The main feature of interest is quality. Quality and rating are inextricably linked by definition in this project. My goal is to understand what variablesare most closely tied to/influence wine quality.

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

Based on the description of the attributes, I suspect that acidity (fixed and volatile), citric acid, residual sugar, chlorides, free sulphur dioxide, pH, and sulphates will provide a good starting point for investigating what most affects the quality of wine. As a non-wine drinker, however, I understand that exploring the relationships between all the variables will be necessary to give me a baseline for further analysis.

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

Following the lead of some others, I created the categorial variable: rating. This grouping idea should help improve data visualizations later in the project.

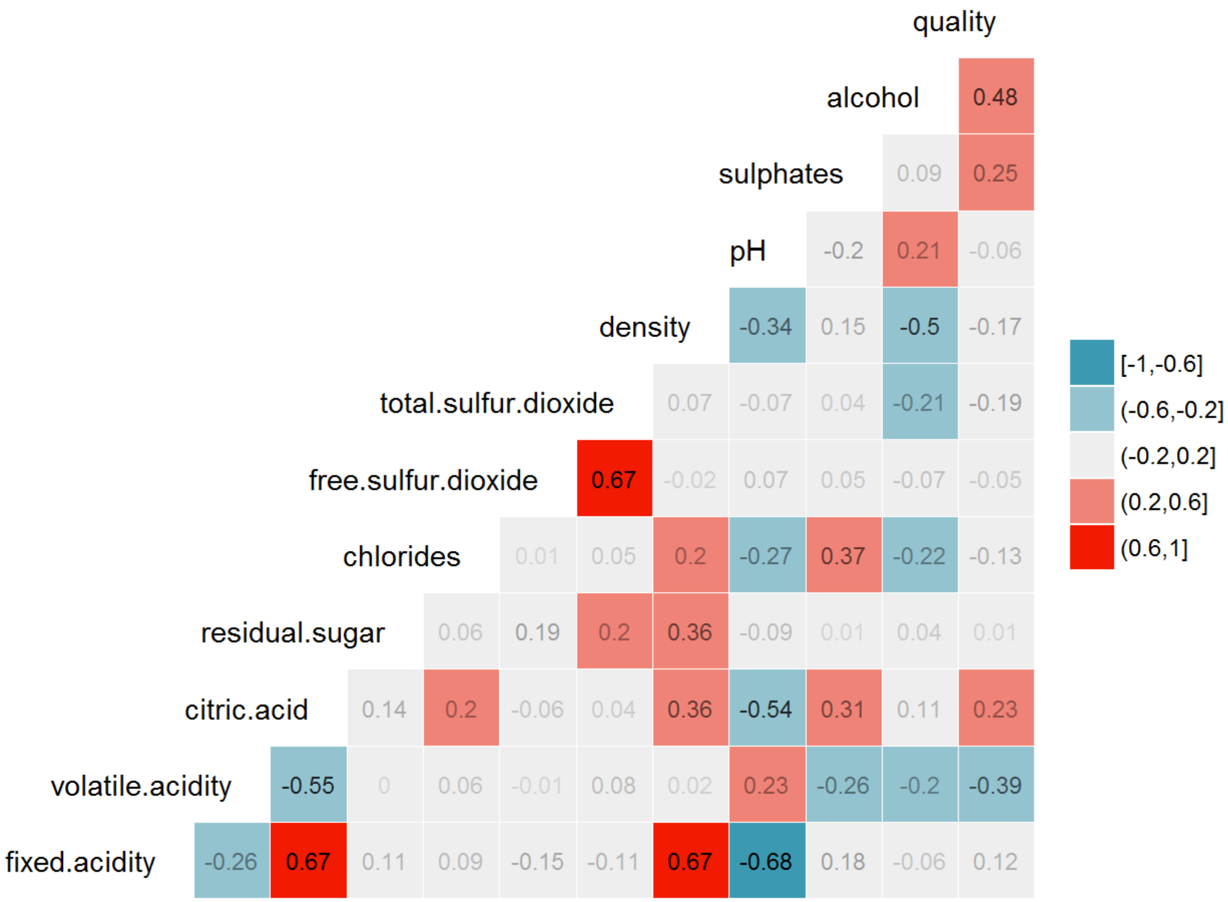
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

The most unusual distributions were that of citric acid and volatile acidity. I did not perform any operation to alter these data. I removed X from the original dataset X since it was only the observation identifier and not useful in this analysis. I adjusted the quality variable to be an ordered factor to help with grouping and visualizations later in the project. At this point, I have not removed outliers or trimmed the data in any way.

Bivariate Plots Section

Correlation Between Variables

```
## [1] "Correlation Between Variables"
```



quality											
alcohol											0.48
sulphates										0.09	0.25
pH									-0.2	0.21	-0.06
density								-0.34	0.15	-0.5	-0.17
total.sulfur.dioxide							0.07	-0.07	0.04	-0.21	-0.19
free.sulfur.dioxide						0.67	-0.02	0.07	0.05	-0.07	-0.05
chlorides					0.01	0.05	0.2	-0.27	0.37	-0.22	-0.13
residual.sugar				0.06	0.19	0.2	0.36	-0.09	0.01	0.04	0.01
citric.acid			0.14	0.2	-0.06	0.04	0.36	-0.54	0.31	0.11	0.23
volatile.acidity		-0.55	0	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.2	-0.39
fixed.acidity	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.12

```
## The correlation diagram helps to quickly visualize the strength of the correlation between va
riables.
```

```
##
## ---------------------------------------------------------------------------
##           &nbsp;            fixed.acidity   volatile.acidity   citric.acid
## -------------------------- --------------- ------------------ -------------
##     **fixed.acidity**             1             -0.2561        **0.6717**
##
##    **volatile.acidity**        -0.2561             1           **-0.5525**
##
##      **citric.acid**         **0.6717**       **-0.5525**           1
##
##     **residual.sugar**         0.1148           0.001918         0.1436
##
##       **chlorides**            0.09371           0.0613          0.2038
##
##  **free.sulfur.dioxide**       -0.1538          -0.0105         -0.06098
##
##  **total.sulfur.dioxide**      -0.1132          0.07647          0.03553
##
##        **density**            **0.668**         0.02203        **0.3649**
##
##           **pH**             **-0.683**          0.2349        **-0.5419**
##
##       **sulphates**             0.183            -0.261        **0.3128**
##
##        **alcohol**            -0.06167          -0.2023          0.1099
##
##        **quality**             0.1241         **-0.3906**        0.2264
## ---------------------------------------------------------------------------
##
## Table: Table continues below
##
##
## -----------------------------------------------------------------------------
##           &nbsp;            residual.sugar   chlorides   free.sulfur.dioxide
## -------------------------- ---------------- ----------- ---------------------
##     **fixed.acidity**            0.1148        0.09371          -0.1538
##
##    **volatile.acidity**        0.001918       0.0613           -0.0105
##
##      **citric.acid**            0.1436        0.2038          -0.06098
##
##     **residual.sugar**             1           0.05561           0.187
##
##       **chlorides**            0.05561           1            0.005562
##
##  **free.sulfur.dioxide**        0.187        0.005562             1
##
##  **total.sulfur.dioxide**       0.203         0.0474         **0.6677**
##
##        **density**            **0.3553**      0.2006          -0.02195
##
##           **pH**               -0.08565       -0.265           0.07038
##
```

```
##       **sulphates**            0.005527     **0.3713**         0.05166
##
##        **alcohol**             0.04208        -0.2211         -0.06941
##
##        **quality**             0.01373        -0.1289         -0.05066
## -----------------------------------------------------------------------------
##
## Table: Table continues below
##
##
## -------------------------------------------------------------------------
##           &nbsp;            total.sulfur.dioxide    density       pH
## -------------------------- ---------------------- ----------- -----------
##     **fixed.acidity**             -0.1132          **0.668**  **-0.683**
##
##    **volatile.acidity**           0.07647           0.02203     0.2349
##
##      **citric.acid**              0.03553         **0.3649**  **-0.5419**
##
##     **residual.sugar**             0.203          **0.3553**   -0.08565
##
##       **chlorides**                0.0474           0.2006      -0.265
##
##  **free.sulfur.dioxide**         **0.6677**        -0.02195     0.07038
##
##  **total.sulfur.dioxide**            1              0.07127    -0.06649
##
##        **density**                0.07127              1      **-0.3417**
##
##           **pH**                  -0.06649        **-0.3417**      1
##
##       **sulphates**               0.04295           0.1485      -0.1966
##
##        **alcohol**                -0.2057         **-0.4962**   0.2056
##
##        **quality**                -0.1851           -0.1749    -0.05773
## -------------------------------------------------------------------------
##
## Table: Table continues below
##
##
## --------------------------------------------------------------
##           &nbsp;            sulphates    alcohol     quality
## -------------------------- ----------- ----------- -----------
##     **fixed.acidity**         0.183     -0.06167     0.1241
##
##    **volatile.acidity**      -0.261      -0.2023   **-0.3906**
##
##      **citric.acid**       **0.3128**    0.1099      0.2264
##
##     **residual.sugar**      0.005527     0.04208     0.01373
##
##       **chlorides**        **0.3713**    -0.2211     -0.1289
##
```

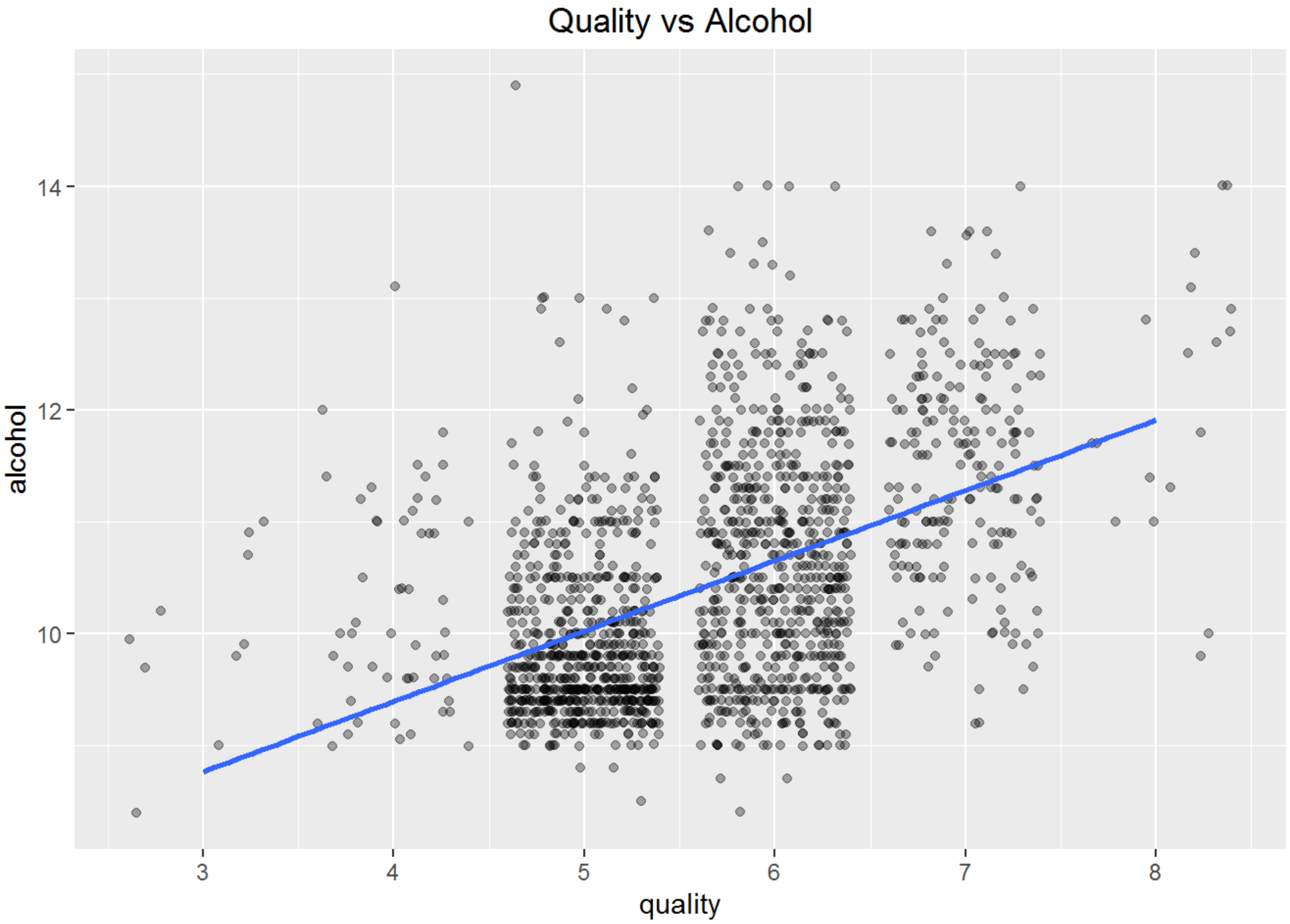
```
##  **free.sulfur.dioxide**     0.05166    -0.06941    -0.05066
##
##  **total.sulfur.dioxide**    0.04295     -0.2057     -0.1851
##
##        **density**           0.1485    **-0.4962**   -0.1749
##
##           **pH**             -0.1966     0.2056     -0.05773
##
##       **sulphates**             1        0.09359     0.2514
##
##        **alcohol**           0.09359        1      **0.4762**
##
##        **quality**           0.2514    **0.4762**       1
## --------------------------------------------------------------
```

```
## Most highly correlated variables with quality--> Alcohol (r=.48), Volatile Acidity (r=-.39),
 Sulphates (r=.25), Citric Acid (r=.23). I was surprised that most variables were only weakly co
rrelated with quality.
```

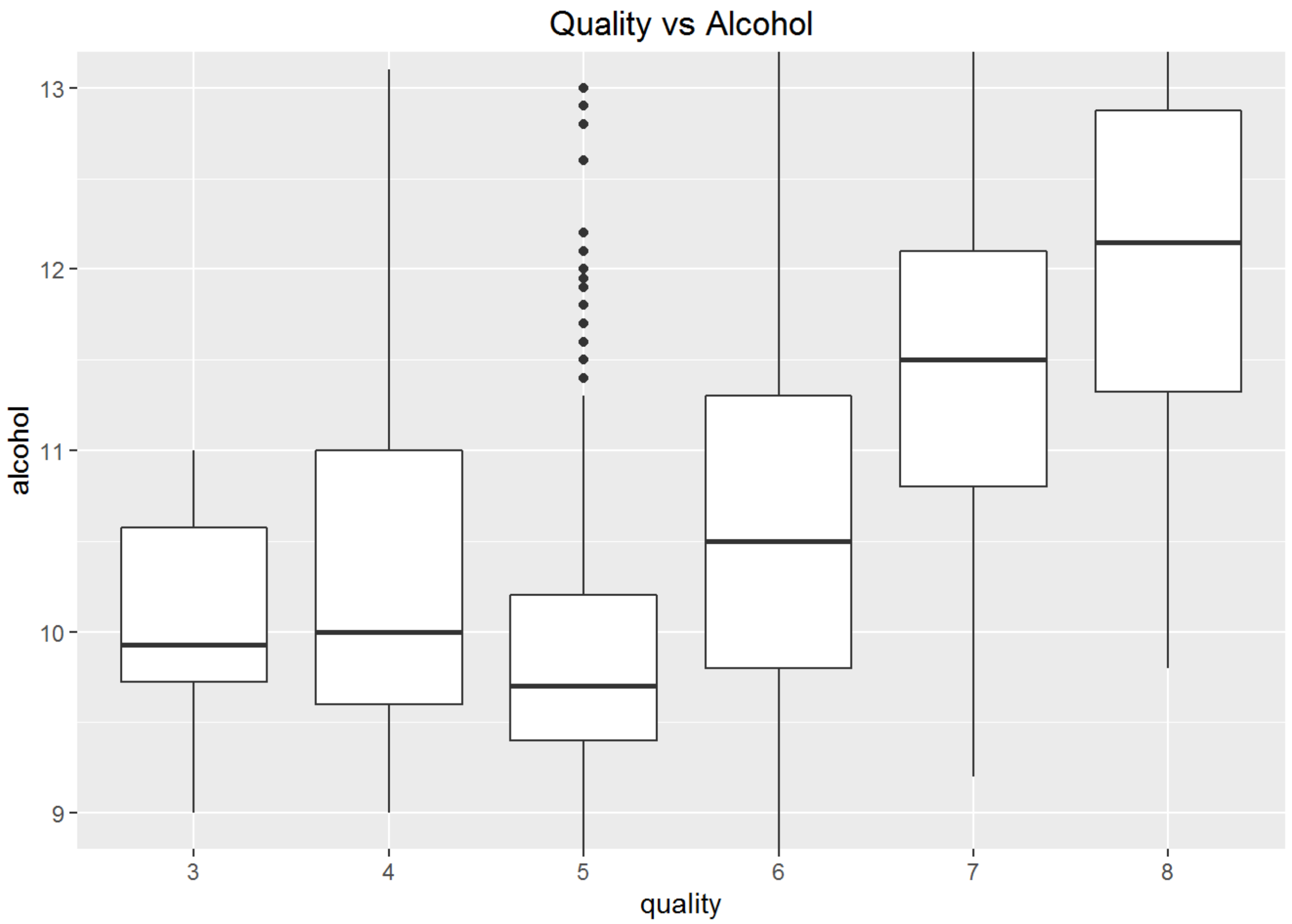
```
## Density, chlorides, and pH have a weak negative correlation with quality. Pearson's r values
 are -.17, -.13, and -.06, respectively
```

```
## Some stronger correlations between other variables were apparent. For example, density and fi
xed acidity are positively correlated (r=.67, df=1597, p<.001). Fixed acidity and pH are negativ
ely correlated (r=-.68, df=1597,p<.001). pH and citric acid are negatively correlated (r=-.54, d
f=1597,p<.001)
```

Quality vs Alcohol



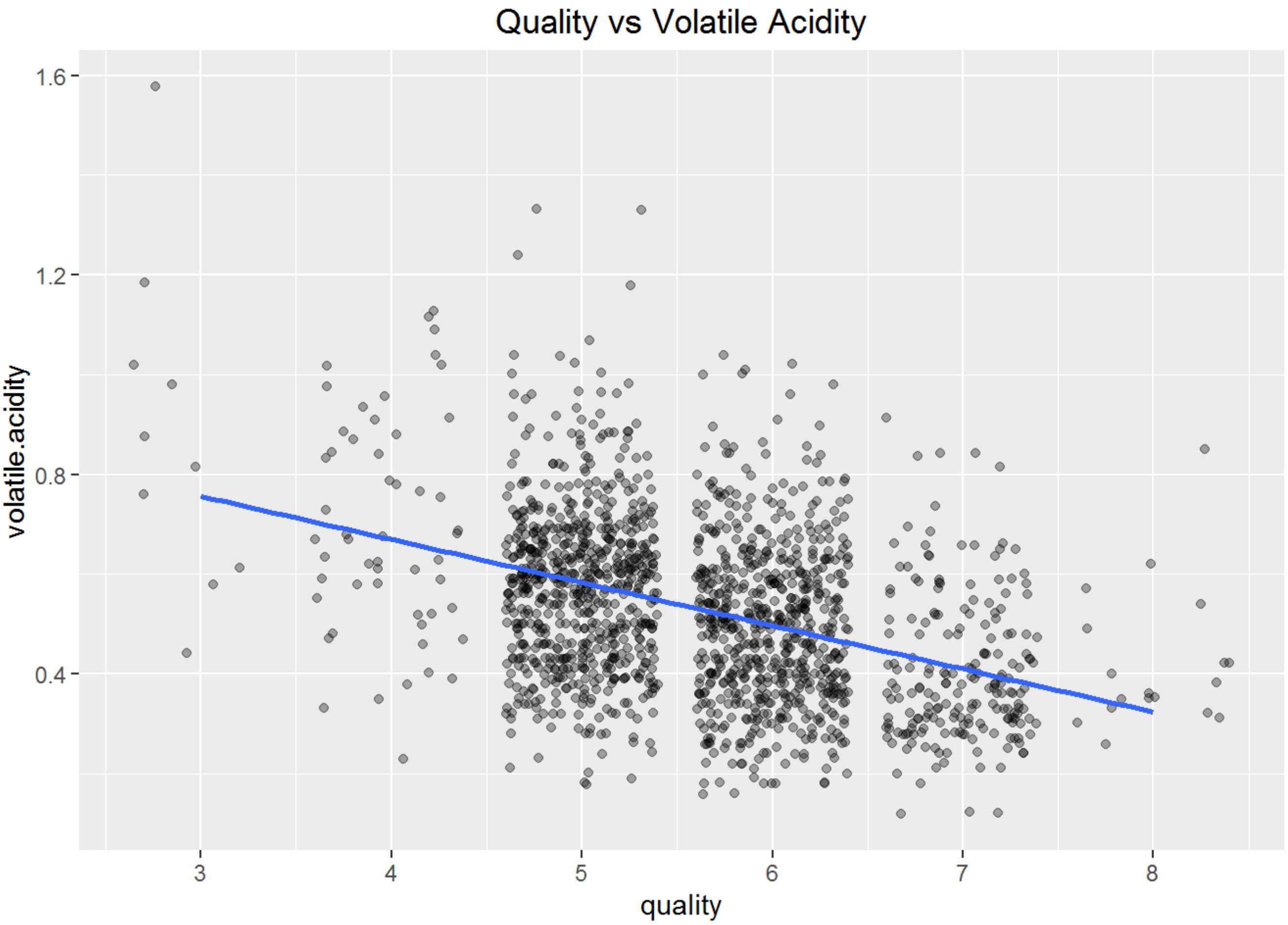
```
## There is a weak positive correlation between alcohol content and quality (r=.48, df=1597, p<.
001, 95% CI: [0.4373540 0.5132081]). This was an interesting finding. I suppose that if a wine t
astes a bit stronger then the perceived quality may be higher. In the scatterplot above, a blue
 linear regression line is overlaid to help visualize the relationship. Points are jittered to h
elp declutter the graphic.
```



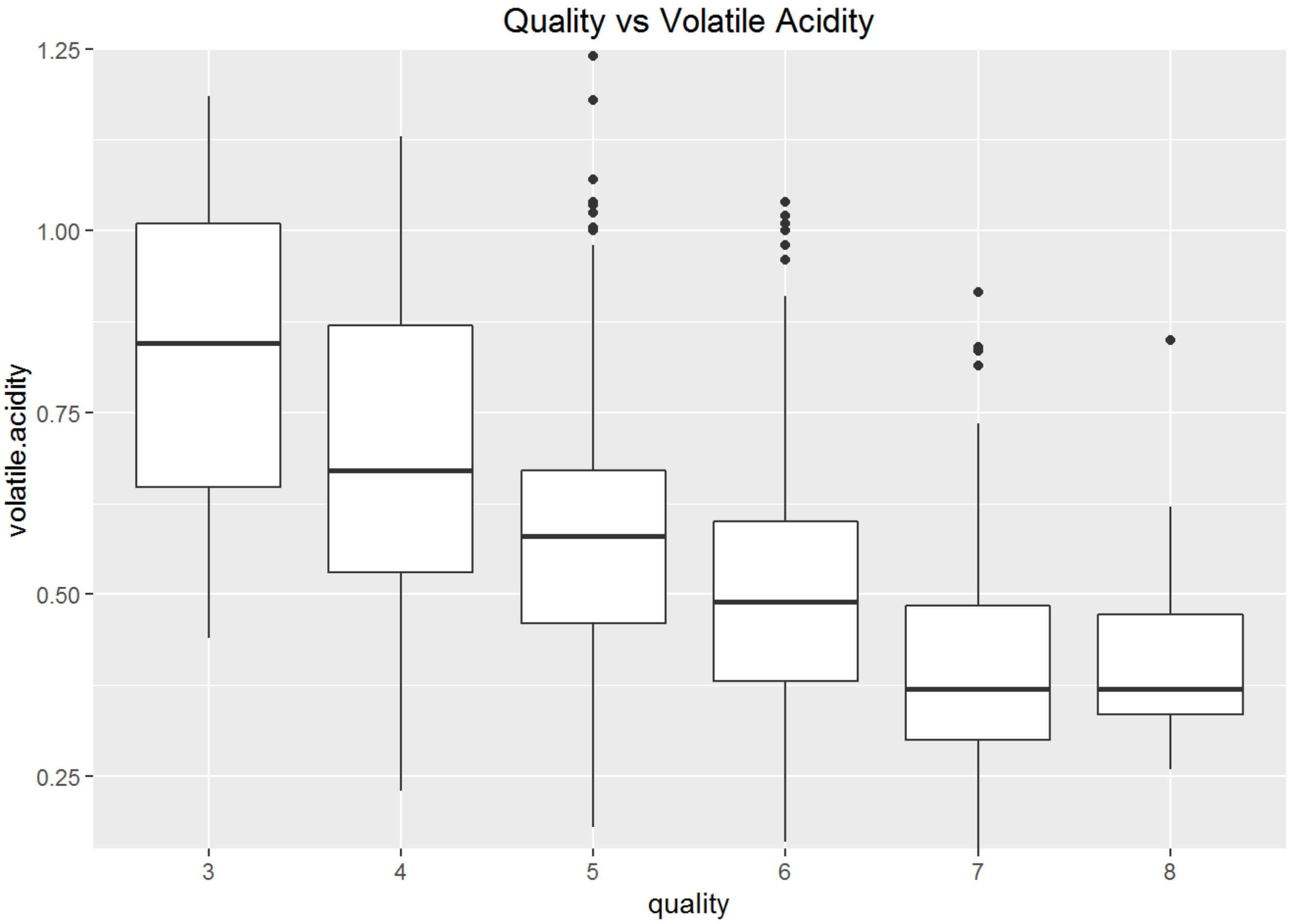
```
## [1] "Summary Statistics"
```

```
## quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   8.400   9.725   9.925   9.955  10.580  11.000 
## -------------------------------------------------------- 
## quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.00    9.60   10.00   10.27   11.00   13.10 
## -------------------------------------------------------- 
## quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     8.5     9.4     9.7     9.9    10.2    14.9 
## -------------------------------------------------------- 
## quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    8.40    9.80   10.50   10.63   11.30   14.00 
## -------------------------------------------------------- 
## quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.20   10.80   11.50   11.47   12.10   14.00 
## -------------------------------------------------------- 
## quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.80   11.32   12.15   12.09   12.88   14.00
## From the boxplots it is apparent that the median value for quality 5 wine is a bit lower than
 the others. It is important to note that this quality category had the highest number of observ
ations (681). Quality 3 and 4 categories only had 10 and 53 observations, respectively. This une
ven sampling is likely introducing some error in my analysis.
```

Quality vs Volatile Acidity

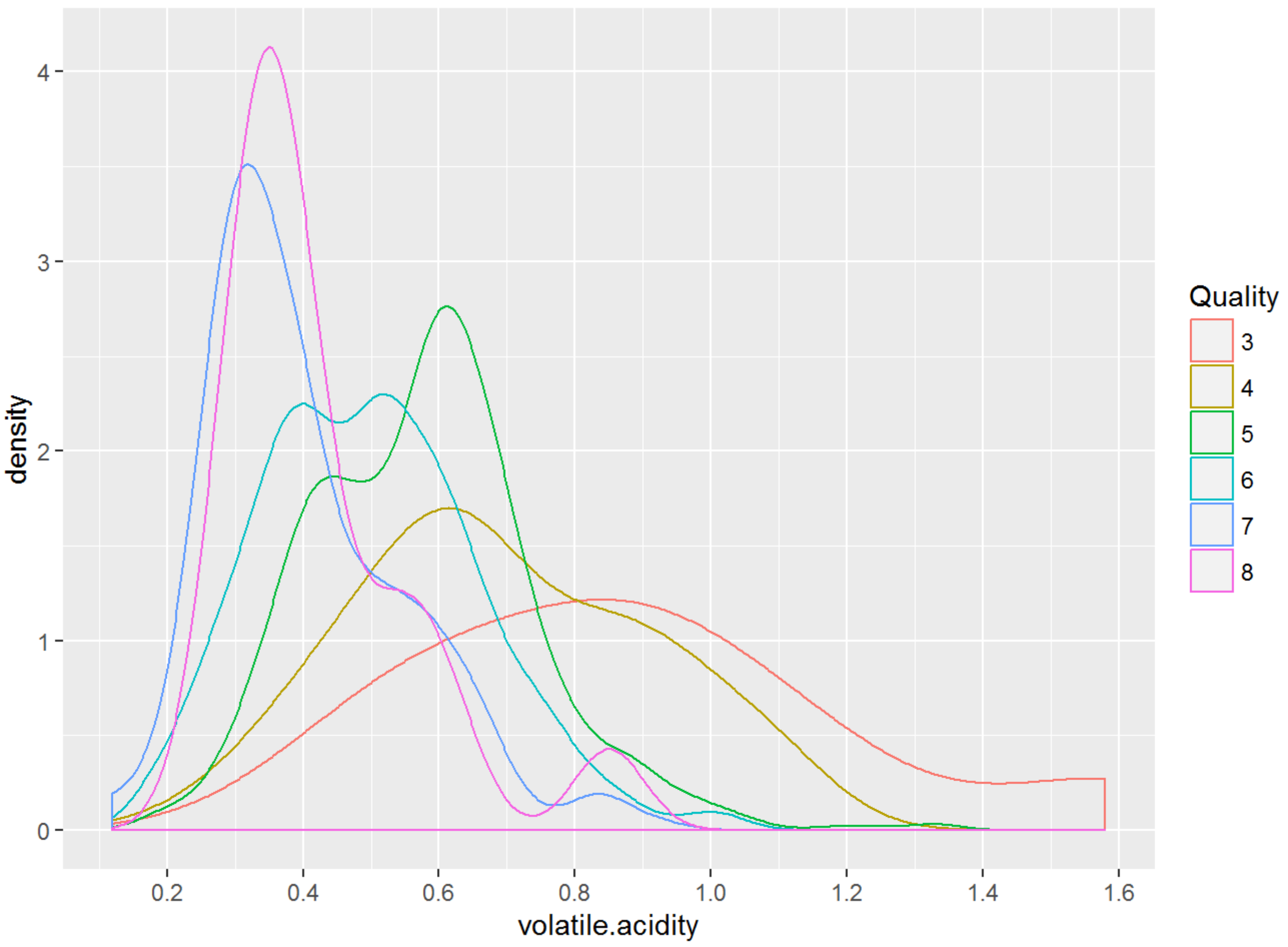


```
## Quality and Volatile Acidity are weakly negatively correlated (r= -.39, df=1597, p<.001, 95%
 CI: [-0.4313210 -0.3482032]). I expected that wines containing more acetic acidic are less plea
sant tasting so this finding is not surprising. In the scatterplot above, a blue linear regressi
on line is overlaid to help visualize the relationship. Points are jittered and their transparen
cy adjusted to help declutter the graphic.
```



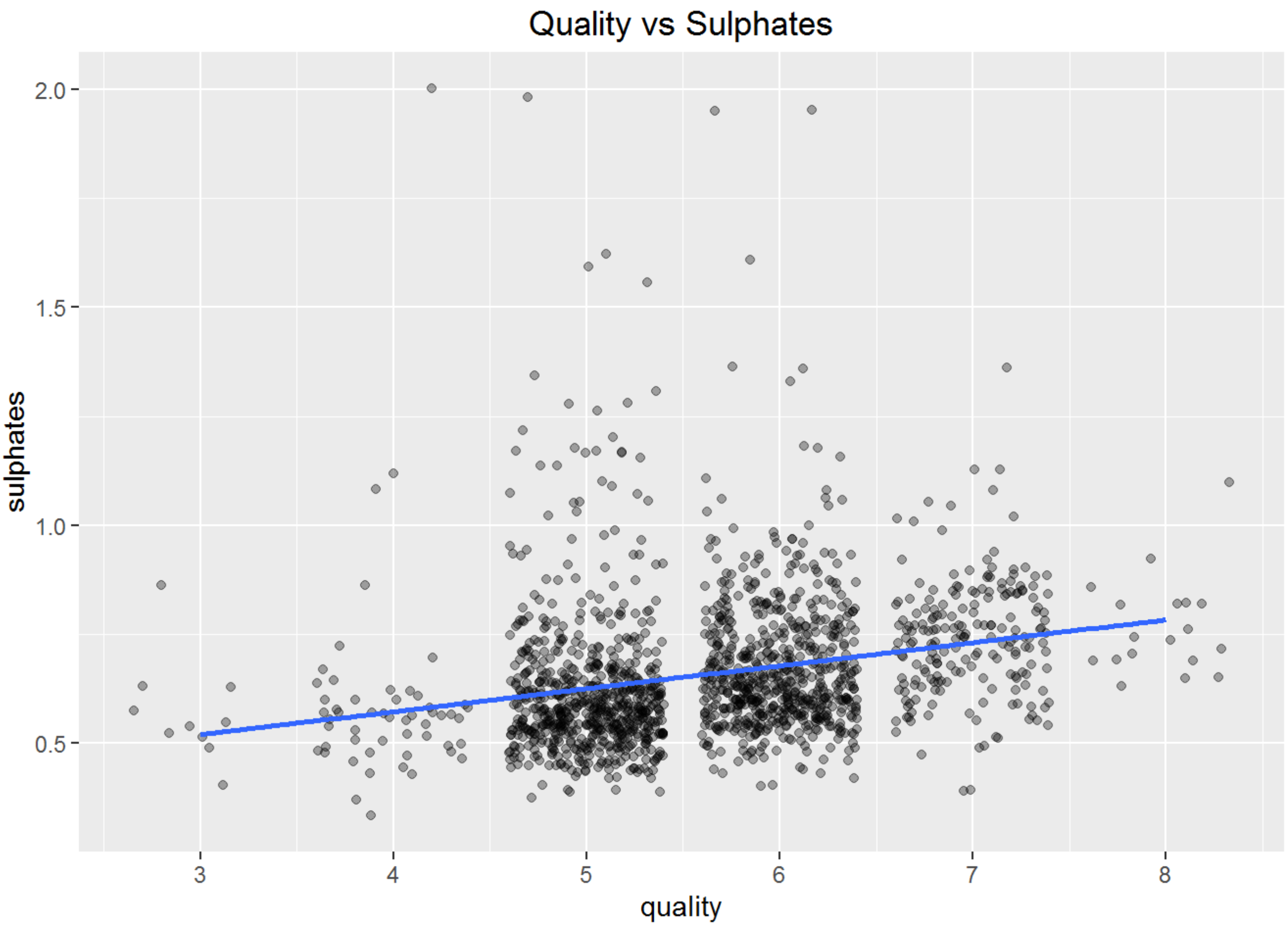
```
## [1] "Summary Statistics"
```

```
## quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4400  0.6475  0.8450  0.8845  1.0100  1.5800 
## -------------------------------------------------------- 
## quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.230   0.530   0.670   0.694   0.870   1.130 
## -------------------------------------------------------- 
## quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.180   0.460   0.580   0.577   0.670   1.330 
## -------------------------------------------------------- 
## quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1600  0.3800  0.4900  0.4975  0.6000  1.0400 
## -------------------------------------------------------- 
## quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1200  0.3000  0.3700  0.4039  0.4850  0.9150 
## -------------------------------------------------------- 
## quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2600  0.3350  0.3700  0.4233  0.4725  0.8500
## The relationship between quality and volatile acidity is illustrated with boxplots. The y-lim
its have been modified to make the graphic easier to read. From the boxplots it is easy to see t
hat in general lower quality wines have higher median volatile acidity.
```

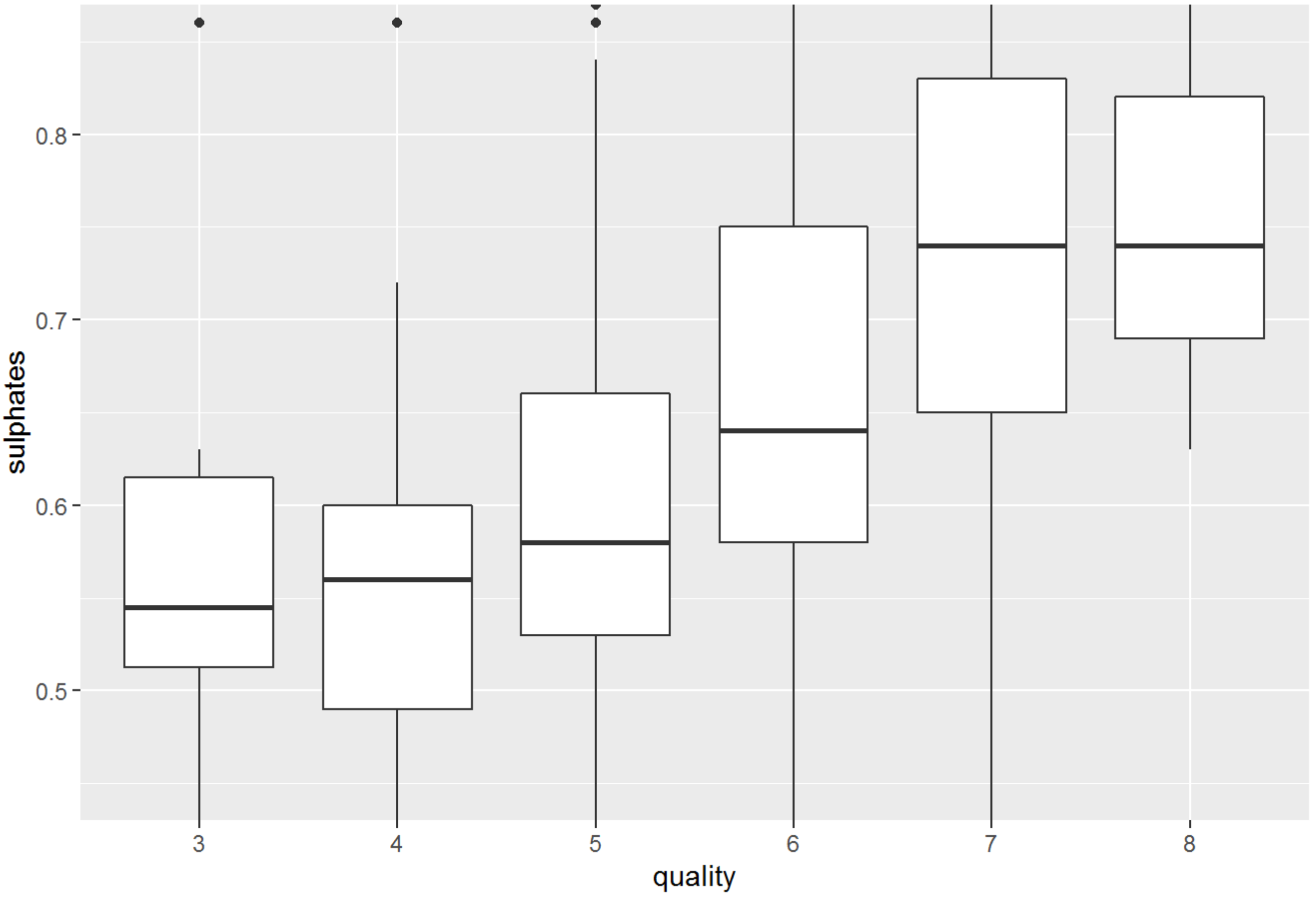


```
## This smoothed density plot has the different quality categories broken out. It is easy to see
 that a large number of wines with a quality of 7 or 8 have a volatile acidity near 0.4 g/dm^3.
 I also see a relative peak for quality 4 and 5 wines around 0.6 g/dm^3. These peaks are represe
nted in the histogram of volatile acidity seen earlier in this project.
```

Quality vs Sulphates



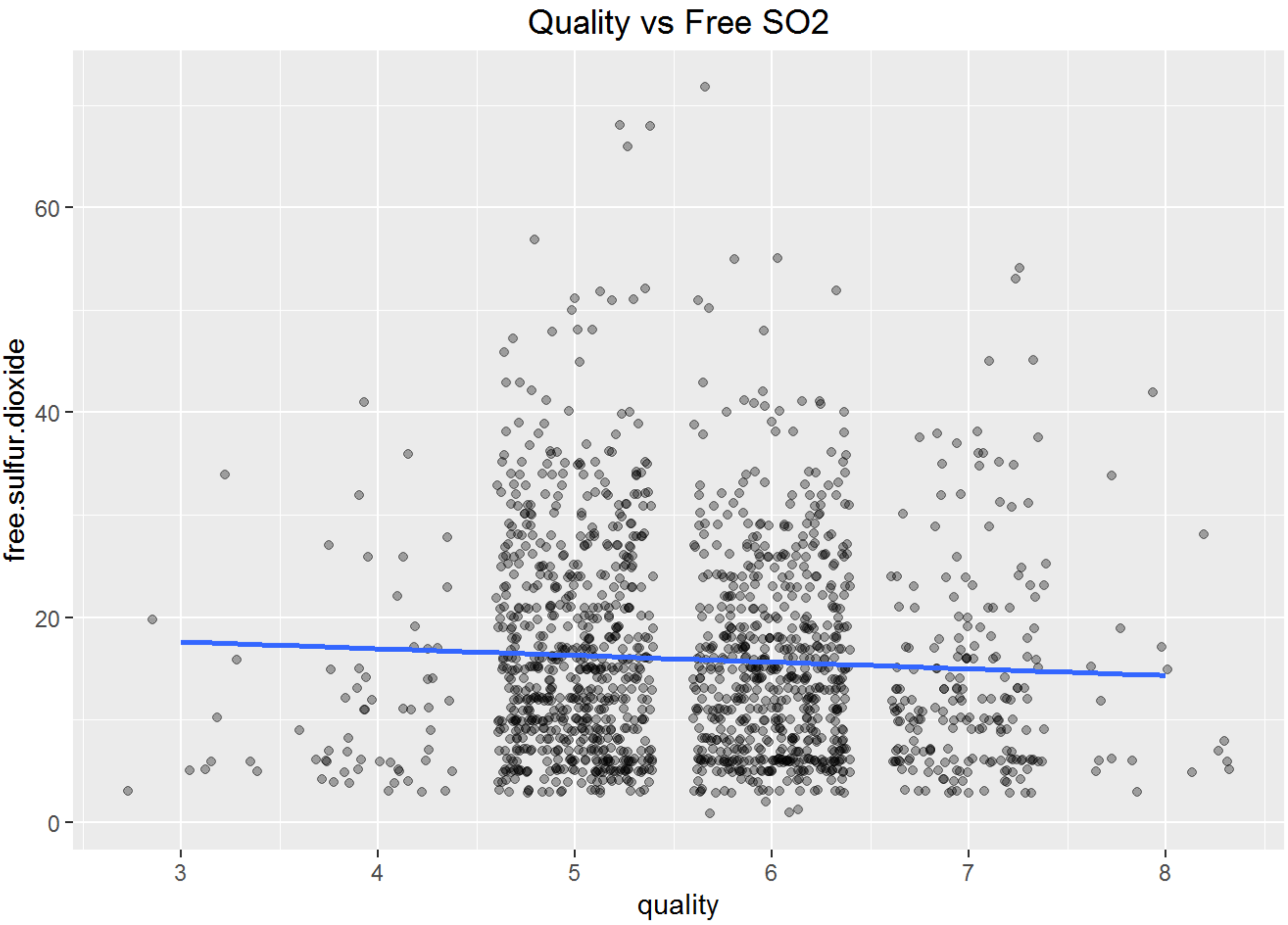
```
## There is a weak positive correlation between quality and sulphates (r=.25, df=1597, p<.001, 9
5% CI: [0.2049011 0.2967610]). This was unexpected especially since sulphates can contribute to
 sulfur dioxide and both free and total sulphur dioxide have a weak negative correlation with qu
ality. According to the variable descriptions, when free SO2 concentrations are over 50 ppm, SO2
 becomes evident in the smell and taste of wine. I will subset the data to look at quality for s
amples with greater than 50 ppm to see if this relationship holds true. In the graphic above, th
e points are jittered and a linear regression line is overlaid in blue.
```



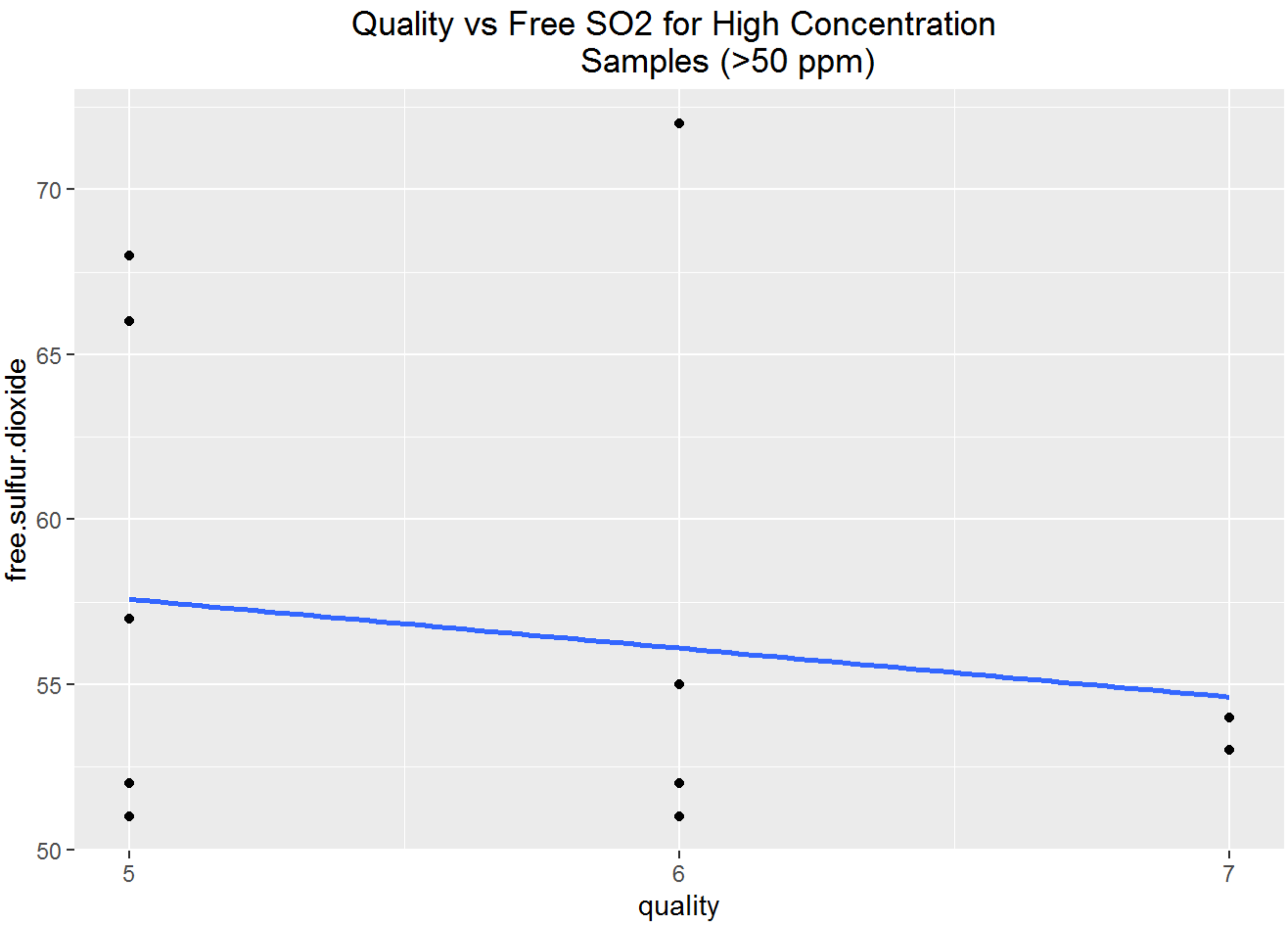
Quality vs Sulphates

```
## [1] "Summary Statistics"
```

```
## quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4000  0.5125  0.5450  0.5700  0.6150  0.8600 
## -------------------------------------------------------- 
## quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3300  0.4900  0.5600  0.5964  0.6000  2.0000 
## -------------------------------------------------------- 
## quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.370   0.530   0.580   0.621   0.660   1.980 
## -------------------------------------------------------- 
## quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4000  0.5800  0.6400  0.6753  0.7500  1.9500 
## -------------------------------------------------------- 
## quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3900  0.6500  0.7400  0.7413  0.8300  1.3600 
## -------------------------------------------------------- 
## quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.6300  0.6900  0.7400  0.7678  0.8200  1.1000
## The boxplots provide a nice summary of the scatterplot. The y-limits have been modified to ma
ke the graphic easier to read. From this graphic it is possible to say that in general high qual
ity wines have higher median sulphate concentrations. There is, however, signficant overlap in t
he interquartile ranges when comparing the wine quality categories.
```

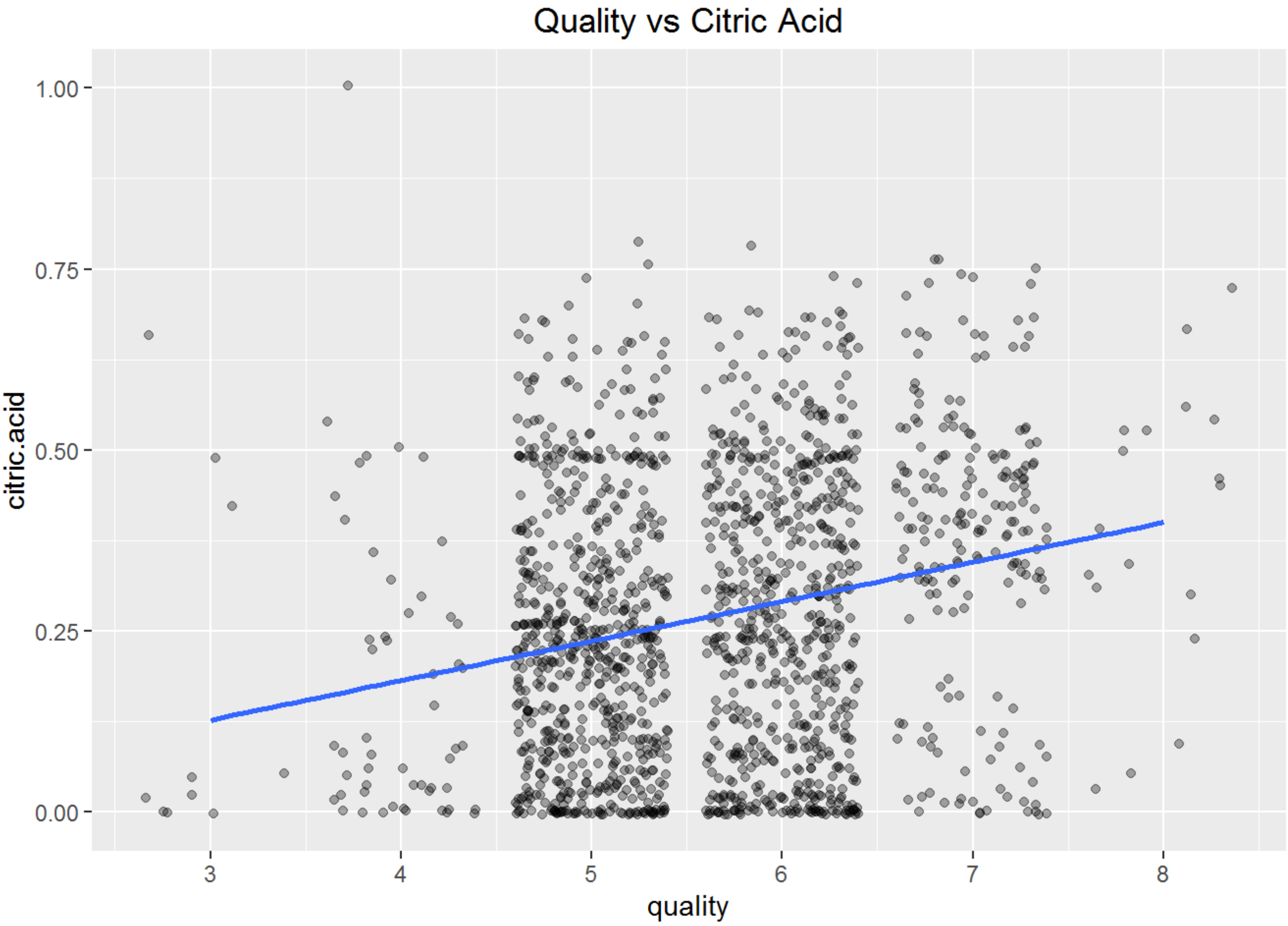


```
## A correlation close to 0 exists between quality and free SO2 (r=-.05, df=1597, p<.05, 95% CI:
 [-0.099430290 -0.001638987]). A linear regression line is overlaid in blue. Points are jittered
 to minimize overplotting.
```

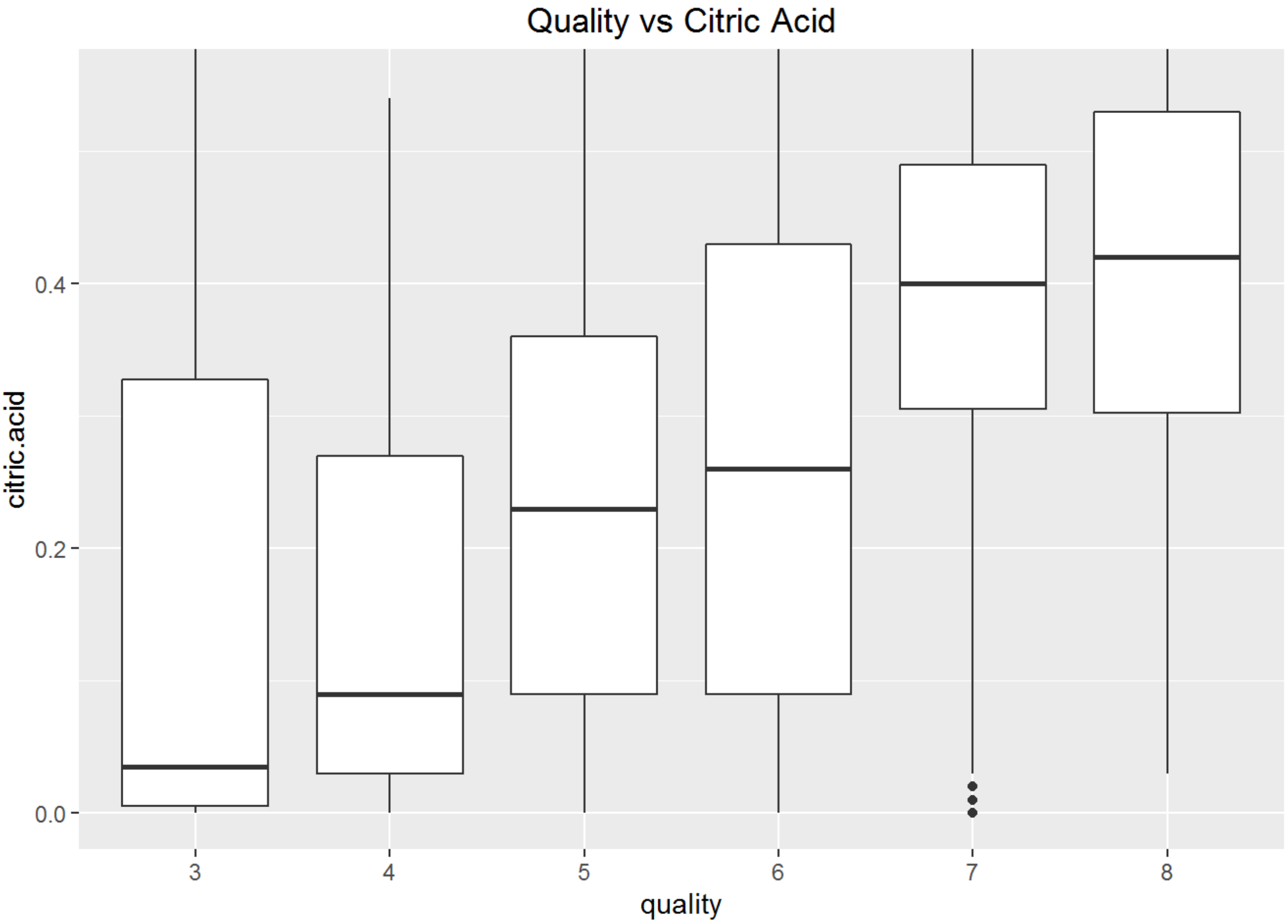


```
## Wine samples with high SO2 (>50 ppm or mg/d^3) were subsetted from the dataset. The correlati
on for quality vs free.sulfur.dioxide for these subsetted samples did not provide conclusive evi
dence that wines with high free SO2 are lower quality (r=-.15, df=14, p=.585). This p-value does
 not meet my .05 signifcance level. I do not jitter the points here since overplotting is not an
 issue. A linear regression line is overlaid in blue.
```

Quality vs Citric Acid



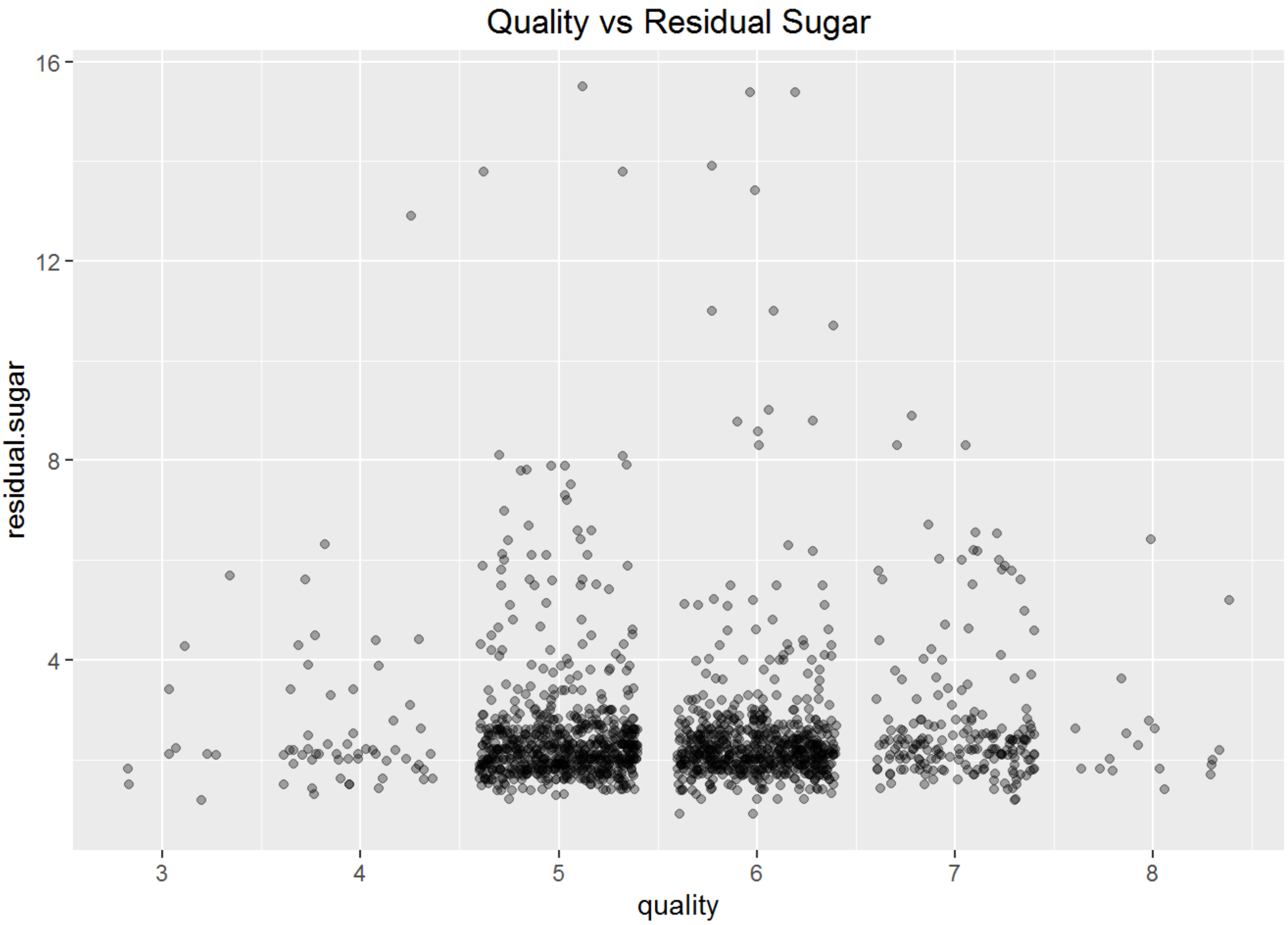
```
## Quality and citric acid are weakly positively correlated (r= .23, df=1597, p<.001, 95% CI:
 [0.1793415 0.2723711]). I presume wines with a higher concentration of citric acid taste freshe
r and may be perceived as being of higher quality. In the scatterplot above, a blue linear regre
ssion line is overlaid to help visualize the relationship. Points are jittered and their alpha p
arameter adjusted to help declutter the graphic.
```



```
## [1] "Summary Statistics"
```

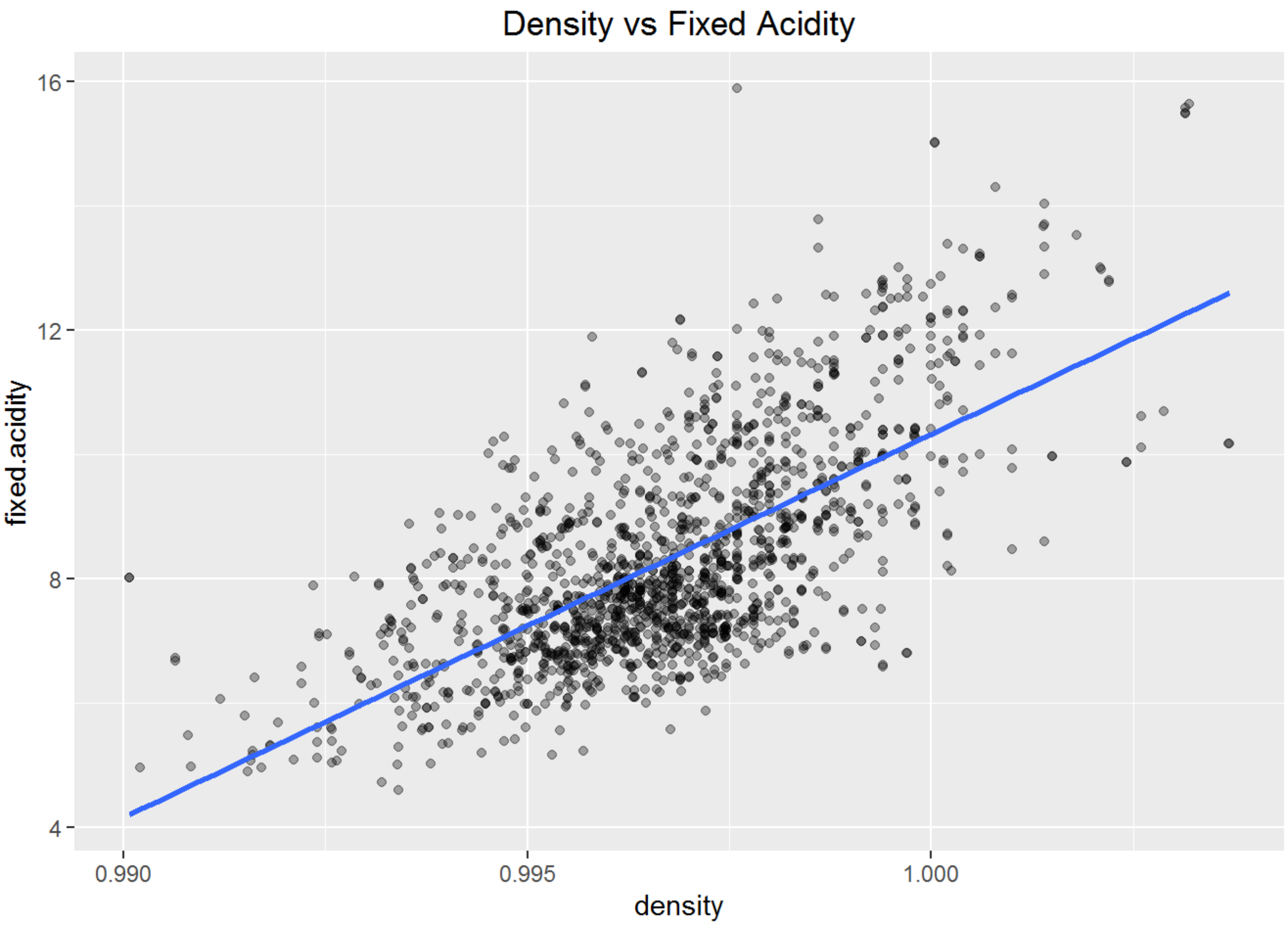
```
## quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4400  0.6475  0.8450  0.8845  1.0100  1.5800 
## -------------------------------------------------------- 
## quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.230   0.530   0.670   0.694   0.870   1.130 
## -------------------------------------------------------- 
## quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.180   0.460   0.580   0.577   0.670   1.330 
## -------------------------------------------------------- 
## quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1600  0.3800  0.4900  0.4975  0.6000  1.0400 
## -------------------------------------------------------- 
## quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1200  0.3000  0.3700  0.4039  0.4850  0.9150 
## -------------------------------------------------------- 
## quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2600  0.3350  0.3700  0.4233  0.4725  0.8500
## The relationship between quality and citric acid is illustrated with boxplots. The y-limits h
ave been modified to make the graphic easier to read. From the boxplots I observe that lower qua
lity wines have lower median citric acid concentrations. Significant overlap between the interqu
artile ranges is evident in the dataset.
```

Quality vs Residual Sugar



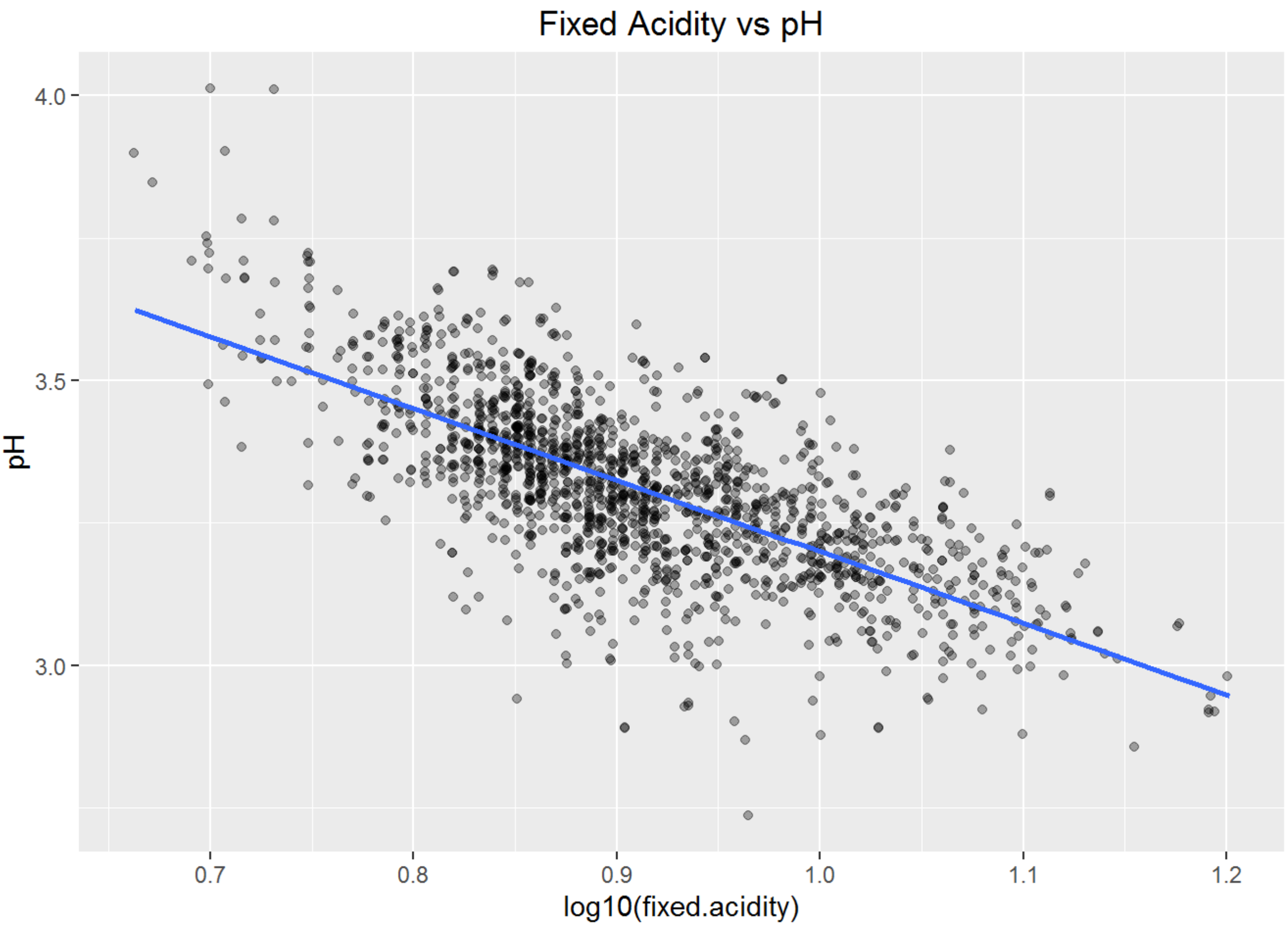
```
## Residual sugar and quality have a correlation close to 0 (r=.01, df= 1597, p<.001, 95% CI: [-
0.03531327  0.06271056]). I was surprised by this. I figured sweeter wines would be seen as lowe
r quality
```

Density and Fixed Acidity



```
## Density is postively correlated with fixed acidity (r=.67, df= 1597, p<.001, 95% CI: [ 0.6399
847 0.6943302]). The three major acids found in wine are tartaric acid, malic acid, and citric a
cid. All of these acids have densities greater than water so the higher the acid concentration i
n the sample the higher the density (with all else held constant).
```

Fixed Acidity and pH



```
## Fixed acidity is negatively correlated with pH (r=-.68, df= 1597, p<.001, 95% CI: [-0.7082857
 -0.6559174]). In the above graphic I transformed fixed acidity using log10 since ph=-log10[H+]
 I expected this relationship since higher acidity means a lower pH with all else held equal.
```

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?

For this project my primary goal is understanding how certain physiochemical variables influence the perceived quality of wine. Out of all the variables, alcohol (r=.48), volatile.acidity (r=-.39), sulphates (r=.25), and citric.acid (r=.23) had the highest correlation with quality. I focus on these below:

I presume that wines with a higher alcohol content had a “fuller” taste and hence were perceived as higher quality.

The negative correlation between volatile acidity and quality was not surprising since higher acetic acid levels (associated with higher volatile acidity) lead to a stronger vinegar taste.

The weak positive correlation between quality and sulphates was interesting. I wasn’t sure what to expect initially. On one hand, sulphates contribute to SO2 levels which act as an antimicrobial and antioxidant. I’m guessing this improves the taste of the wine. On the other hand, high free SO2 levels (>50 ppm) have a negative effect on the smell and taste of wine. This thought process led me to explore the relationship between free SO2 and quality for wine samples with greater than 50 ppm. I found no conclusive relationship, however, that indicacted the expected negative correlation. It is important to note that this was based on a small sample size (n=16).

The weak positive correlation between quality and citric acid was not totally unexpected. It is easy to imagine that a fresher tasting wine may be perceived as being of higher quality.

I was surprised to find that residual sugar concentrations and quality have a correlation close to 0. I figured that sweeter wines would be viewed as lower quality. Maybe there is some other chemistry at play here that I am not aware of.

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

The positive correlation between density and fixed acidity stood out to me. I later learned that the primary acids in wine all have densities greater than water so the higher the acid concentration in the sample the higher the density (with all else held constant).

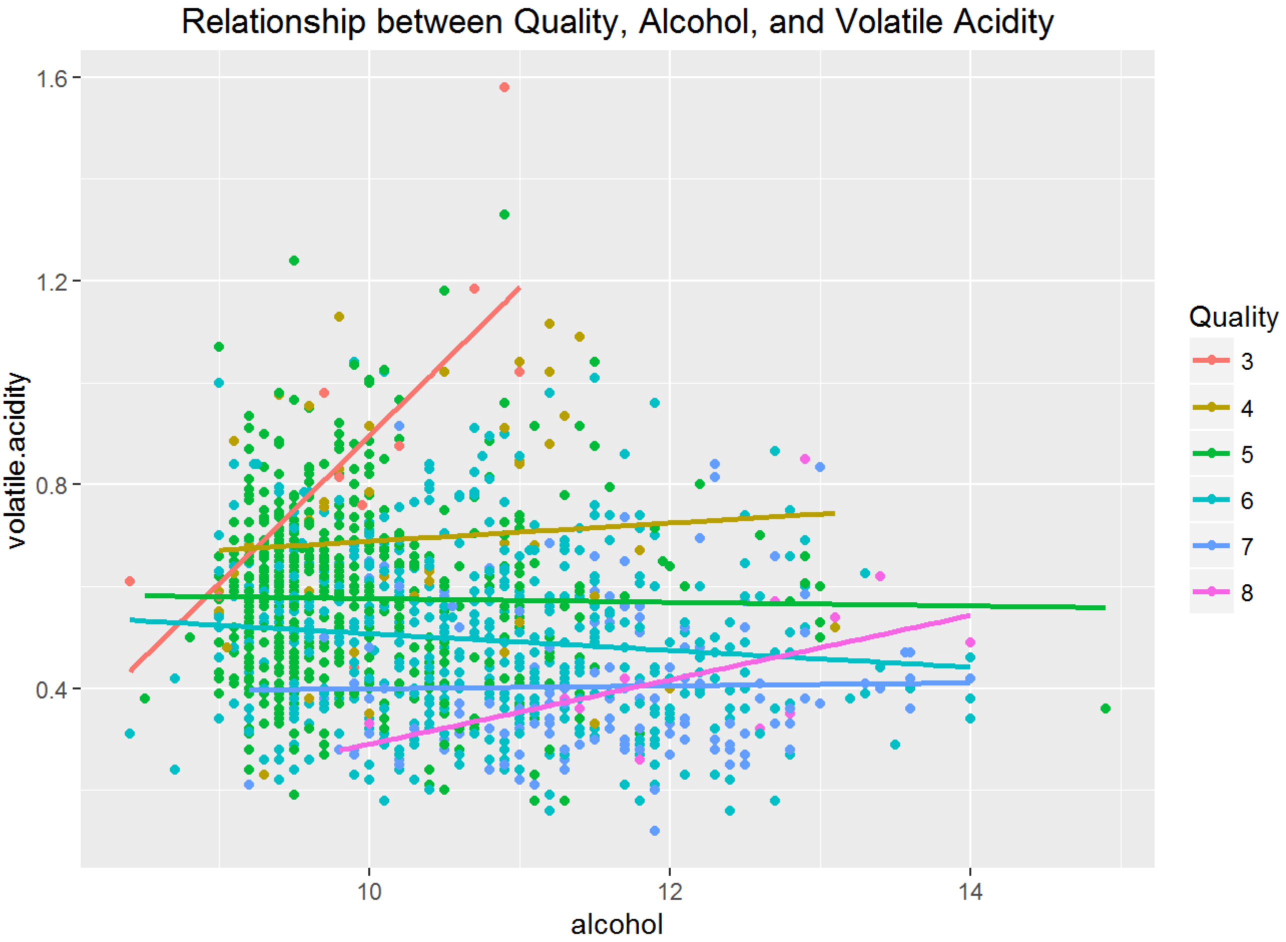
The negative correlation between fixed acidity and pH wasn’t surprising. It is well-understood that as acidity increases pH decreases (with other variables held constant).

What was the strongest relationship you found?

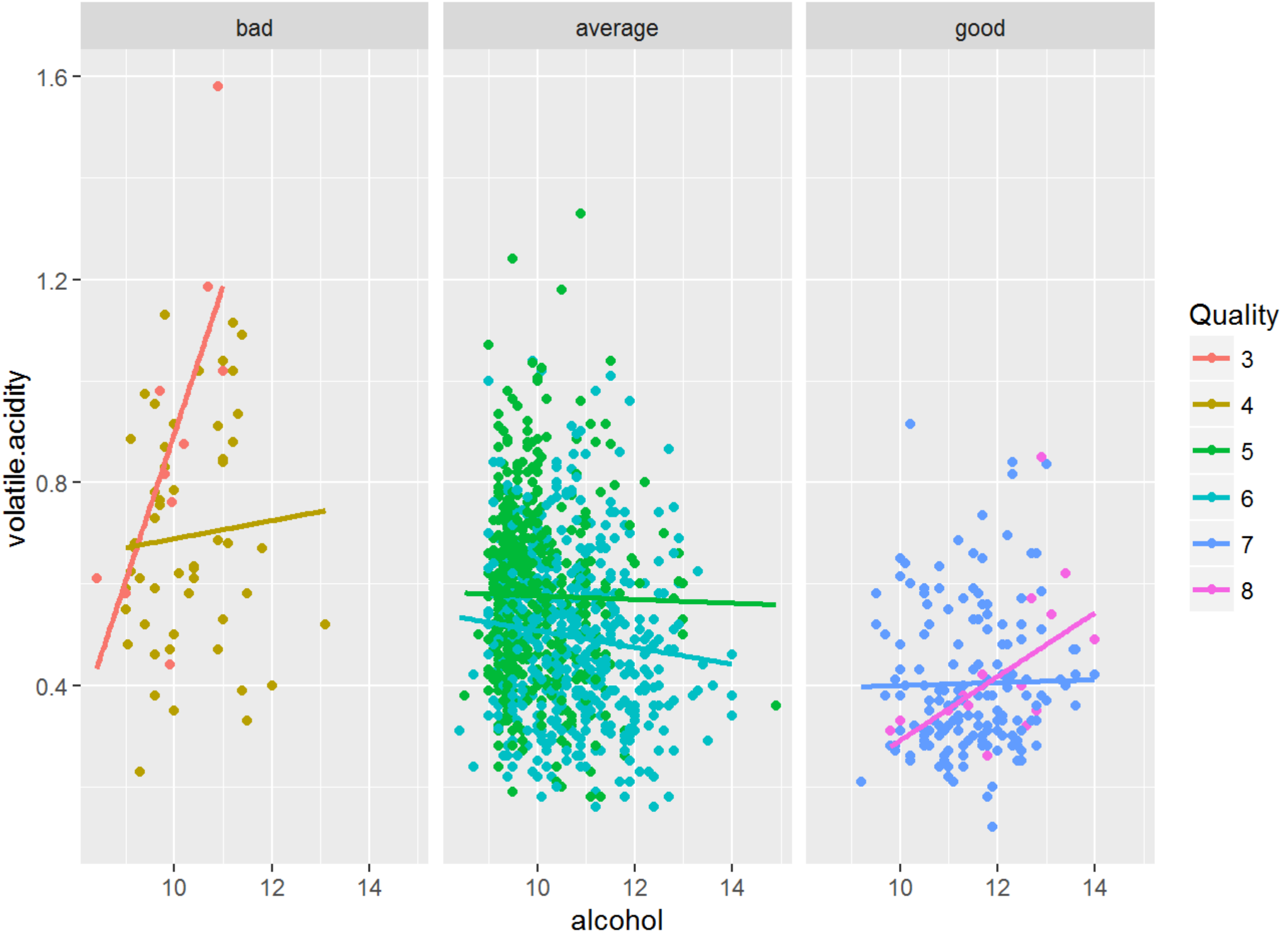
The strongest relationship that I found was between fixed acidity and pH.

Multivariate Plots Section

Explore the Relationship Between Alcohol, Volatile Acidity, and Quality



```
## [1] "Some clustering of similar quality wines is present."
```



```
## [1] "Summary Stats: Alcohol"
```

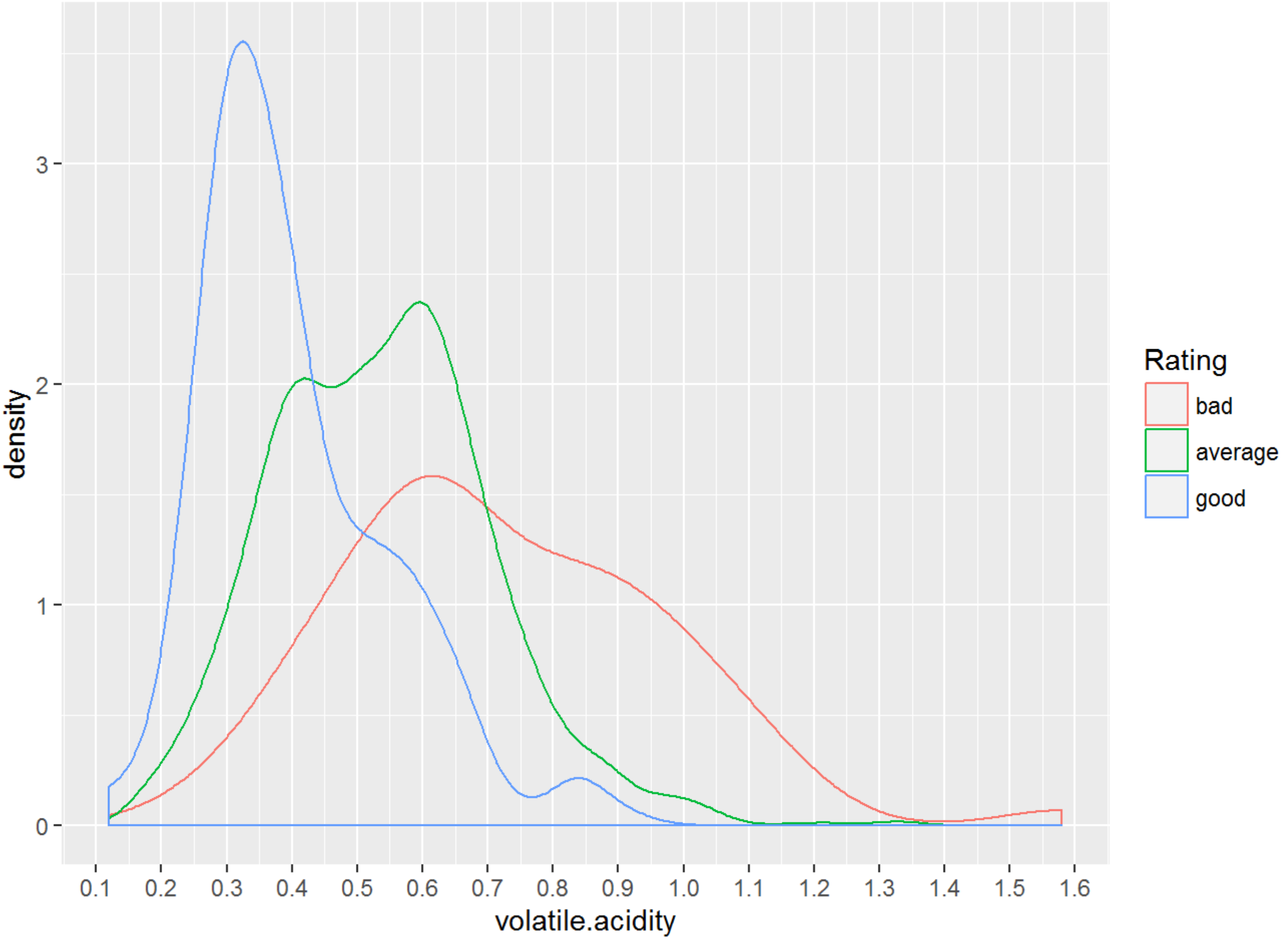
```
## rating: bad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    8.40    9.60   10.00   10.22   11.00   13.10 
## -------------------------------------------------------- 
## rating: average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    8.40    9.50   10.00   10.25   10.90   14.90 
## -------------------------------------------------------- 
## rating: good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.20   10.80   11.60   11.52   12.20   14.00
```

```
## [1] "Summary Stats: Volatile Acidity"
## rating: bad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2300  0.5650  0.6800  0.7242  0.8825  1.5800 
## -------------------------------------------------------- 
## rating: average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1600  0.4100  0.5400  0.5386  0.6400  1.3300 
## -------------------------------------------------------- 
## rating: good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1200  0.3000  0.3700  0.4055  0.4900  0.9150
```

```
## [1] "Volatile Acidity: Variance"
```

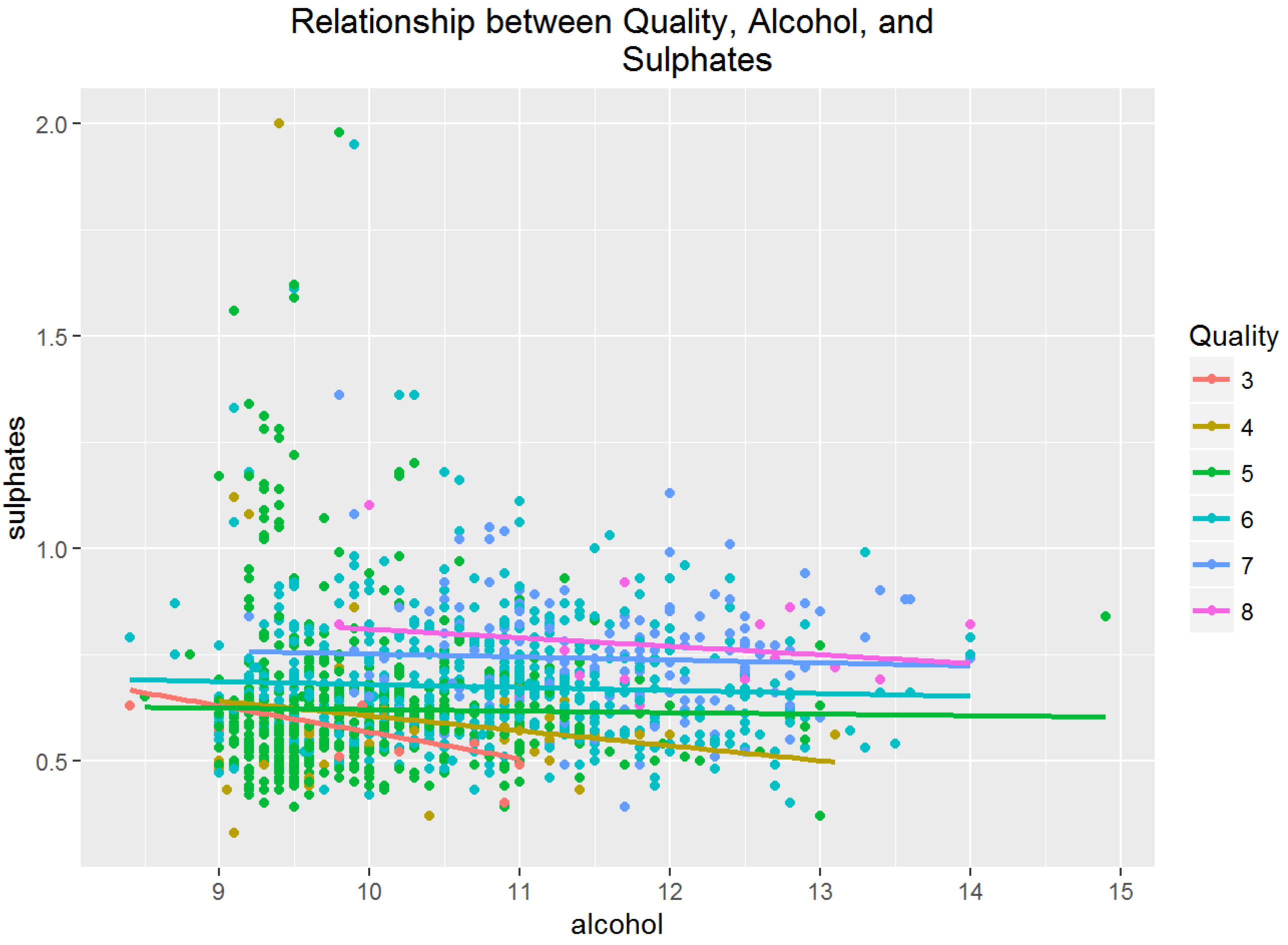
```
## rating: bad
## [1] 0.06148888
## -------------------------------------------------------- 
## rating: average
## [1] 0.02811625
## -------------------------------------------------------- 
## rating: good
## [1] 0.02101419
```

```
## This graphic breaks down the previous graphic into different ratings categories. This makes i
t easier to understand the relationship between quality, alcohol, and volatile acid concentratio
n. With regard to alcohol, it is evident that wines rated as bad (quality<5) have lower median a
lcohol percent by volume (10) than those rate as good (quality>6) which have a median alcohol pe
rcent by volume of (11.6). When looking at the alcohol by volume across all the rating categorie
s, I find an interquartile range of 1.4 for each. With regard to volatile acidity, median acidit
y decreases as rating increases. In general, the highest rated wines have both high alcohol cont
ent and low volatile acidity. It is important to note that wines rated as bad have the distribut
ion with the greatest variance.
```

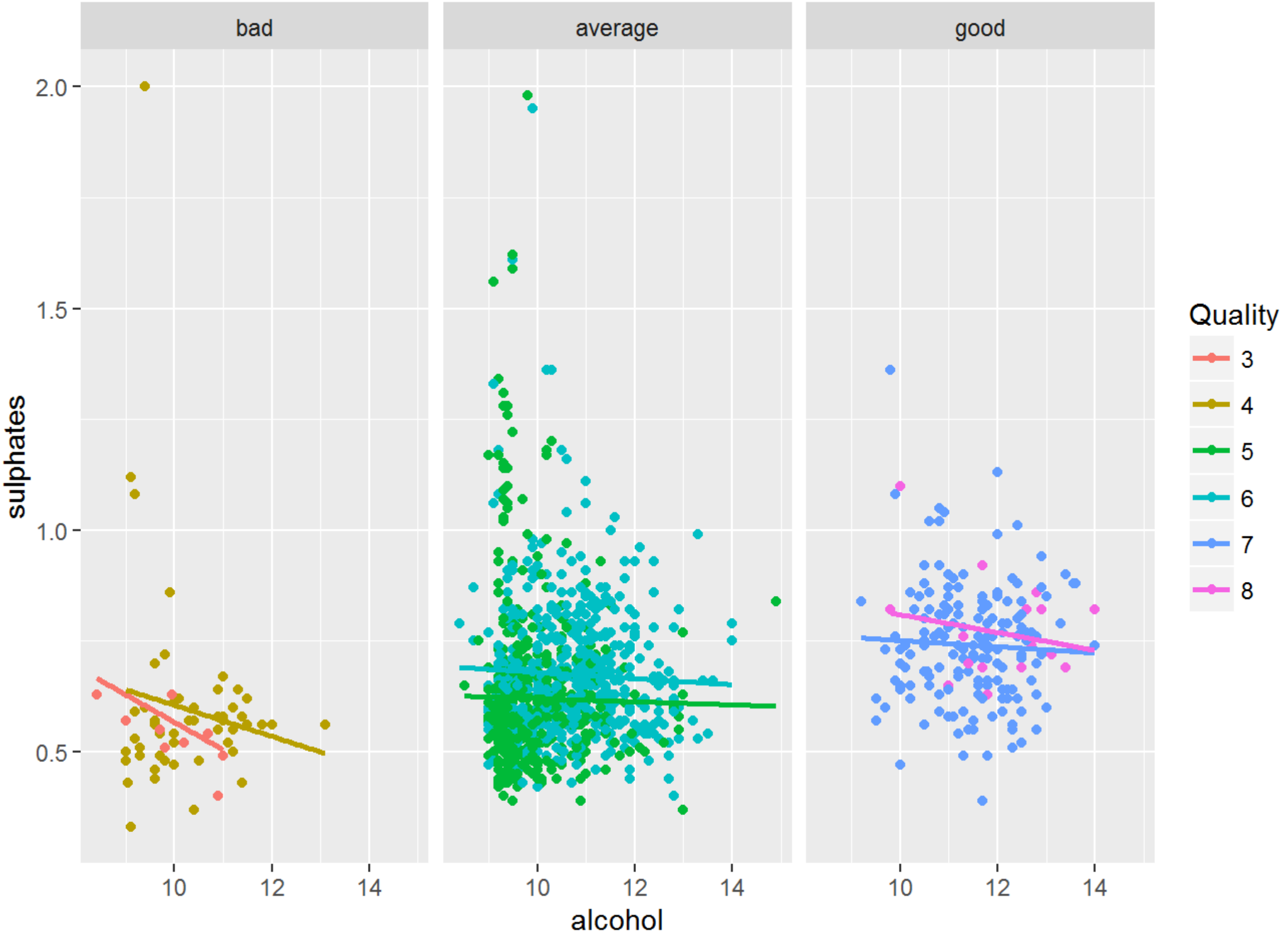


```
## This density plot provides a nice visualization of volatile acidity concentration across the
 3 ratings categories. It is easy to see a quasi-bimodal distribution with most of the highest r
ating wines having a lower concentration of volatile acidity (0.4 g/dm^3) compared to average an
d bad wines which have a peak at about 0.6 g/dm^3.
```

Explore the Relationship Between Alcohol, Sulphates, and Quality



```
## It seems like there is some clustering of samples by wine quality. I need to break this graph
ic down further to explore this.
```



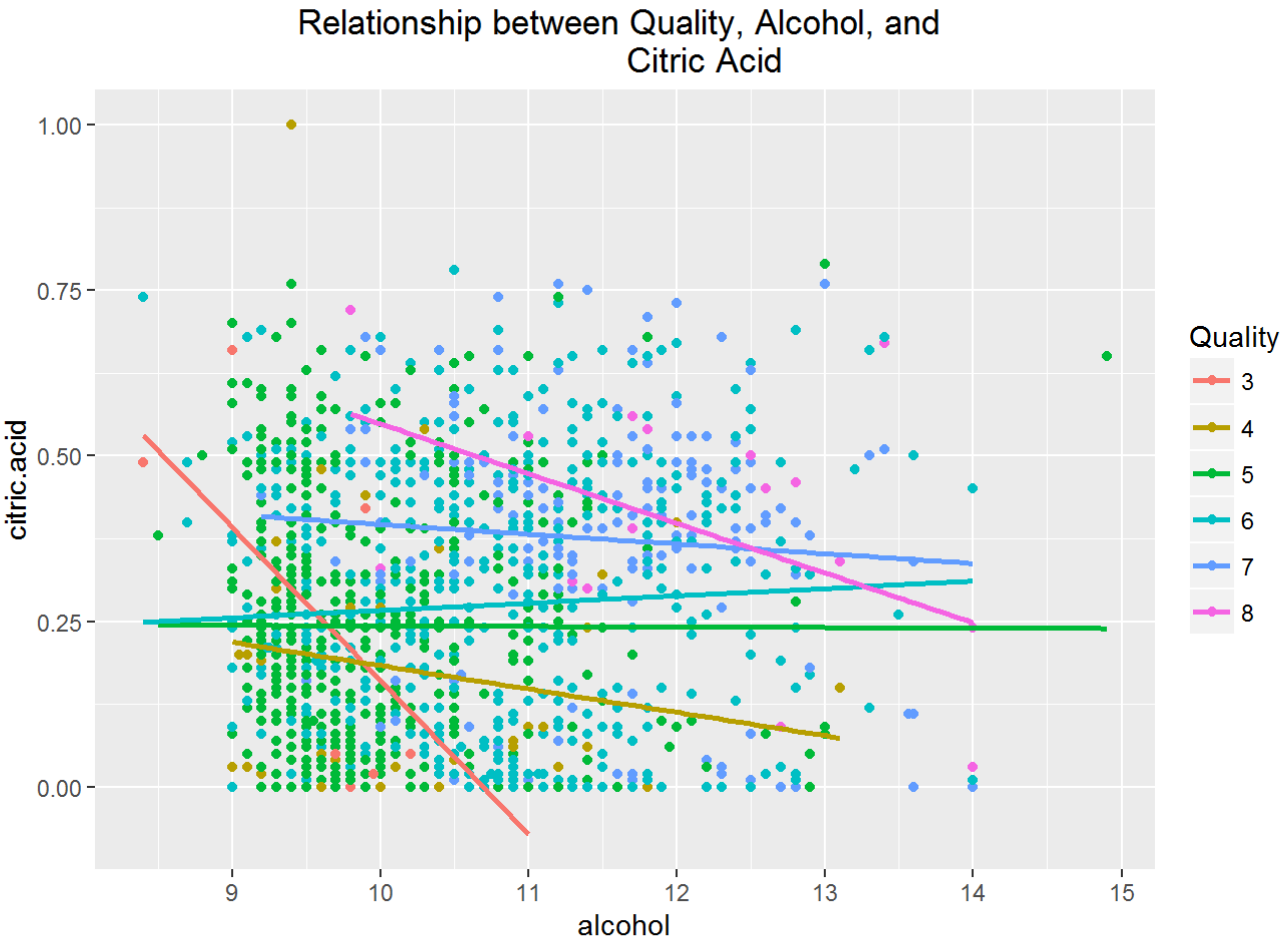
```
## [1] "Summary Stats: Sulphates"
```

```
## rating: bad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3300  0.4950  0.5600  0.5922  0.6000  2.0000 
## -------------------------------------------------------- 
## rating: average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3700  0.5400  0.6100  0.6473  0.7000  1.9800 
## -------------------------------------------------------- 
## rating: good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3900  0.6500  0.7400  0.7435  0.8200  1.3600
```

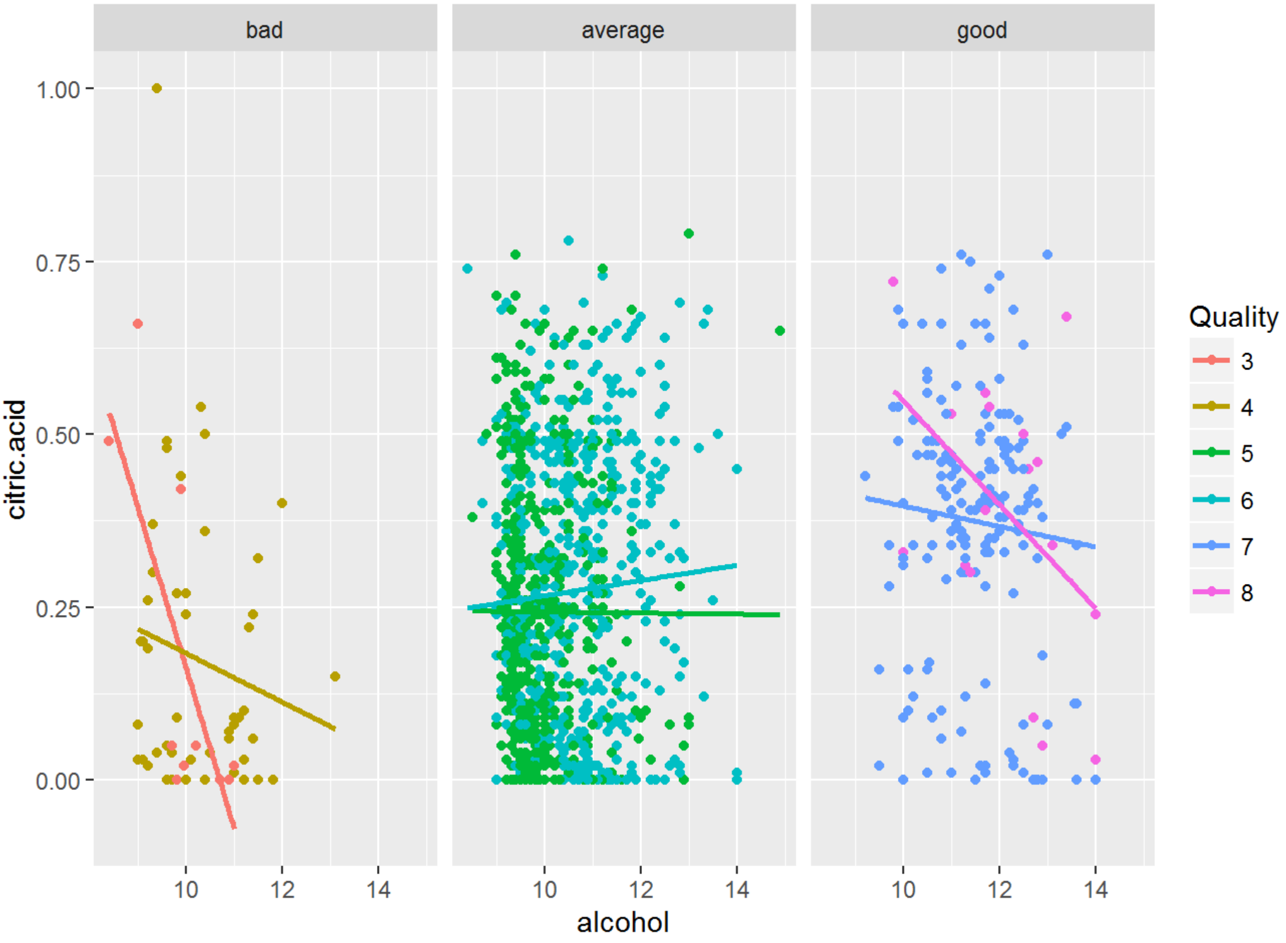
```
## [1] "Sulphates: Variance"
## rating: bad
## [1] 0.05032079
## --------------------------------------------------------
## rating: average
## [1] 0.02800062
## --------------------------------------------------------
## rating: good
## [1] 0.01796624
```

```
## This graphic breaks down the previous graphic by rating to help facilitate exploration of the
 relationship between quality, alcohol, and sulphate concentration. Better rated wines have both
 higher median alcohol content and higher median sulphate concentrations. Wines with a rating of
 bad have the highest sulphate concentration variance.
```

Explore the Relationship Between Alcohol, Citric Acid, and Quality



```
## It seems like for each quality category a wide range of citric acid concentrations exist. I s
aw this earlier with boxplots.
```



```
## [1] "Summary Stats: Citric Acid"
```

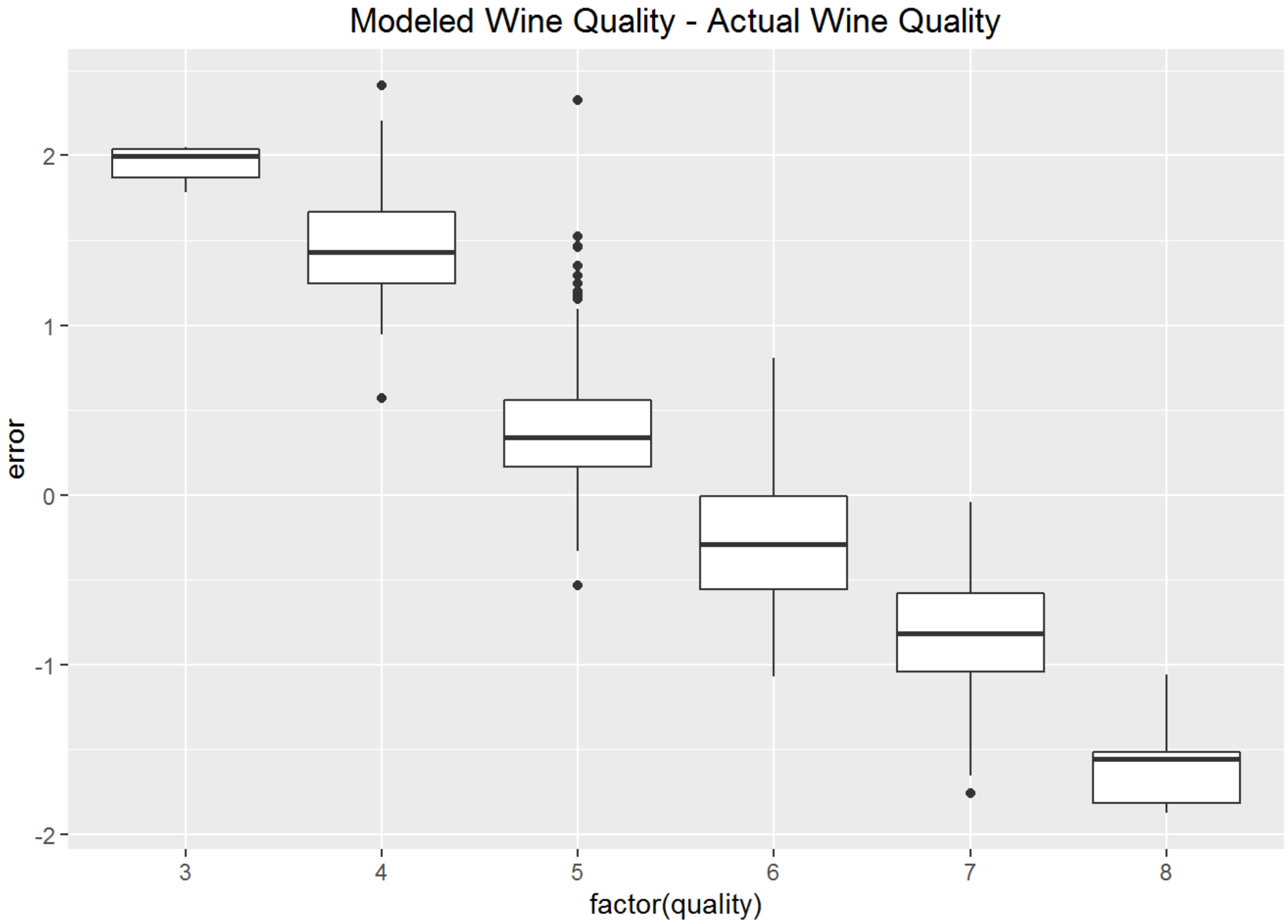
```
## factor(rating): bad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0200  0.0800  0.1737  0.2700  1.0000 
## -------------------------------------------------------- 
## factor(rating): average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0900  0.2400  0.2583  0.4000  0.7900 
## -------------------------------------------------------- 
## factor(rating): good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.3000  0.4000  0.3765  0.4900  0.7600
```

```
## [1] "Citric Acid: Variance"
## factor(rating): bad
## [1] 0.0430171
## --------------------------------------------------------
## factor(rating): average
## [1] 0.03534198
## --------------------------------------------------------
## factor(rating): good
## [1] 0.0378062
```

```
## The high degree of dispersion of data points in all ratings categories is apparent. The varia
nces in citric acid concentrations are comparable between bad, average, and good wines. In gener
al, the best rated wines have high alcohol content and high citric acid.
```

Build a Model

```
## 
## Calls:
## m1: lm(formula = I(quality) ~ I(alcohol), data = wine_numeric)
## m2: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity, data = wine_numeric)
## m3: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + sulphates, 
##     data = wine_numeric)
## m4: lm(formula = I(quality) ~ I(alcohol) + volatile.acidity + sulphates + 
##     citric.acid, data = wine_numeric)
## 
## ================================================================
##                         m1         m2         m3         m4     
## ----------------------------------------------------------------
##   (Intercept)        1.875***   3.095***   2.611***   2.646***  
##                     (0.175)    (0.184)    (0.196)    (0.201)    
##   I(alcohol)         0.361***   0.314***   0.309***   0.309***  
##                     (0.017)    (0.016)    (0.016)    (0.016)    
##   volatile.acidity             -1.384***  -1.221***  -1.265***  
##                                (0.095)    (0.097)    (0.113)    
##   sulphates                                0.679***   0.696***  
##                                           (0.101)    (0.103)    
##   citric.acid                                        -0.079     
##                                                      (0.104)    
## ----------------------------------------------------------------
##   R-squared              0.2        0.3        0.3        0.3   
##   adj. R-squared         0.2        0.3        0.3        0.3   
##   sigma                  0.7        0.7        0.7        0.7   
##   F                    468.3      370.4      268.9      201.8   
##   p                      0.0        0.0        0.0        0.0   
##   Log-likelihood     -1721.1    -1621.8    -1599.4    -1599.1   
##   Deviance             805.9      711.8      692.1      691.9   
##   AIC                 3448.1     3251.6     3208.8     3210.2   
##   BIC                 3464.2     3273.1     3235.7     3242.4   
##   N                   1599       1599       1599       1599     
## ================================================================
```



```
## Residual scatter plot was summarized as boxplots showing actual wine quality subtracted from
 modeled wine quality for a test sample derived from the dataset.
```

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?

For this part of the analysis I explored the relationship between wine quality and the variables that had the highest correlation with it (alcohol, volatile acidity, and sulphates). Higher alcohol and lower volatile acidity wine samples were generally rated as higher quality. The same was true of wines with higher alcohol and higher sulphate concentrations. Wines with higher alcohol and higher citric acid were also percieved as higher quality. Breaking the individual scatter plots down by rating enabled me to visualize these multi-variable relationships with more ease.

Were there any interesting or surprising interactions between features?

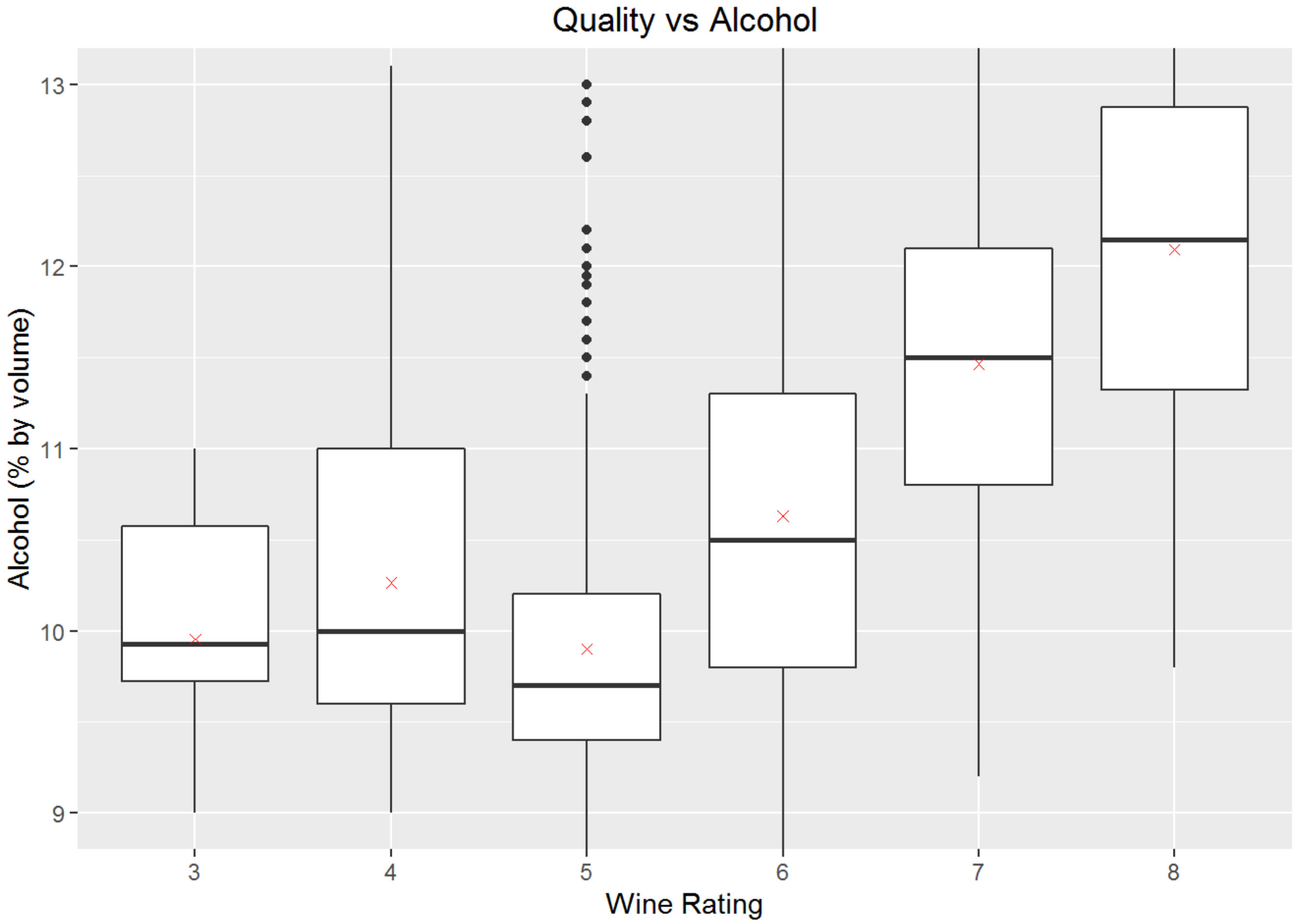
I was suprised that in general higher rated wines had higher sulphate concentrations. Understanding that suphates act as an antimicrobrial could help to explain this finding if the presence of microbrobes makes the wine taste less pleasant. It is clear from the multivariate analysis that the samples with the highest sulphate concentrations did not make the good rating category. These outlier samples fell in the bad and average categories. I presume this was the result of higher free sulphur dioxide concentrations which can negatively affect the smell and taste of wine. My exploration of this expectation in the bivariate analysis section, however, did not provide evidence of this.

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

I created a linear model to predict wine quality utilizing alcohol content, volatile acidity, sulphates, and citric acid concentrations. The r-squared value indicates that this model only expains about 30% of the variance in wine quality. Adding additional physiochemical properties to the model did not improve it. Predicting human behavoir (or in this case the opinion of a wine expert) is difficult and in these types of exercises r-squared values typically fall below 50%. Even though low r-squared values can sometimes be acceptable or even normal, in this case, it is clear from the residual plot I generated that some bias exists in the data. The model seems to overestimate the quality of bad wine (quality = 3 or 4) and underestimate the quality of good wine (quality = 6 or 7). This may suggest that a linear model is not appropriate or that the model is underspecified.

Final Plots and Summary

Plot One (Quality Vs Alcohol)

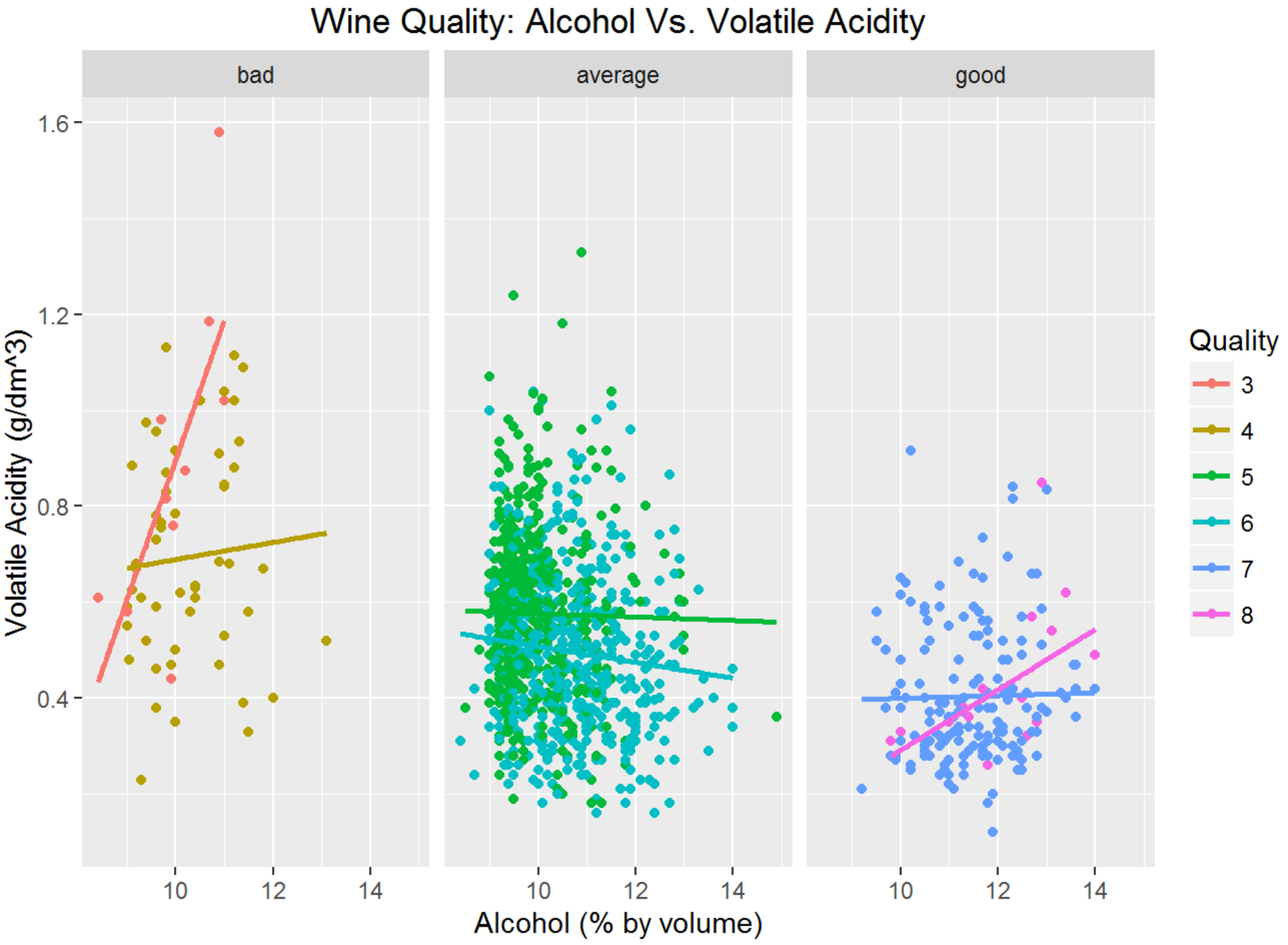


```
## [1] "Summary Statistics"
## quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.580  11.000
## --------------------------------------------------------
## quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## --------------------------------------------------------
## quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## --------------------------------------------------------
## quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
## --------------------------------------------------------
## quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## --------------------------------------------------------
## quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

Description One

These boxplots are broken down by wine quality. The y-limits have been modified to better display the bulk of the data. The mean value for alcohol percent by volume for each individual quality category is overlaid as a red x. Out of all the physicochemical variables in the dataset, alcohol had the strongest correlation with quality (r=.48, df=1597, p<.001, 95% CI: [0.4373540 0.5132081]). It seems that if a wine tastes stronger then the perceived quality is higher.

Plot Two (Quality vs Alcohol and Volatile Acidity)



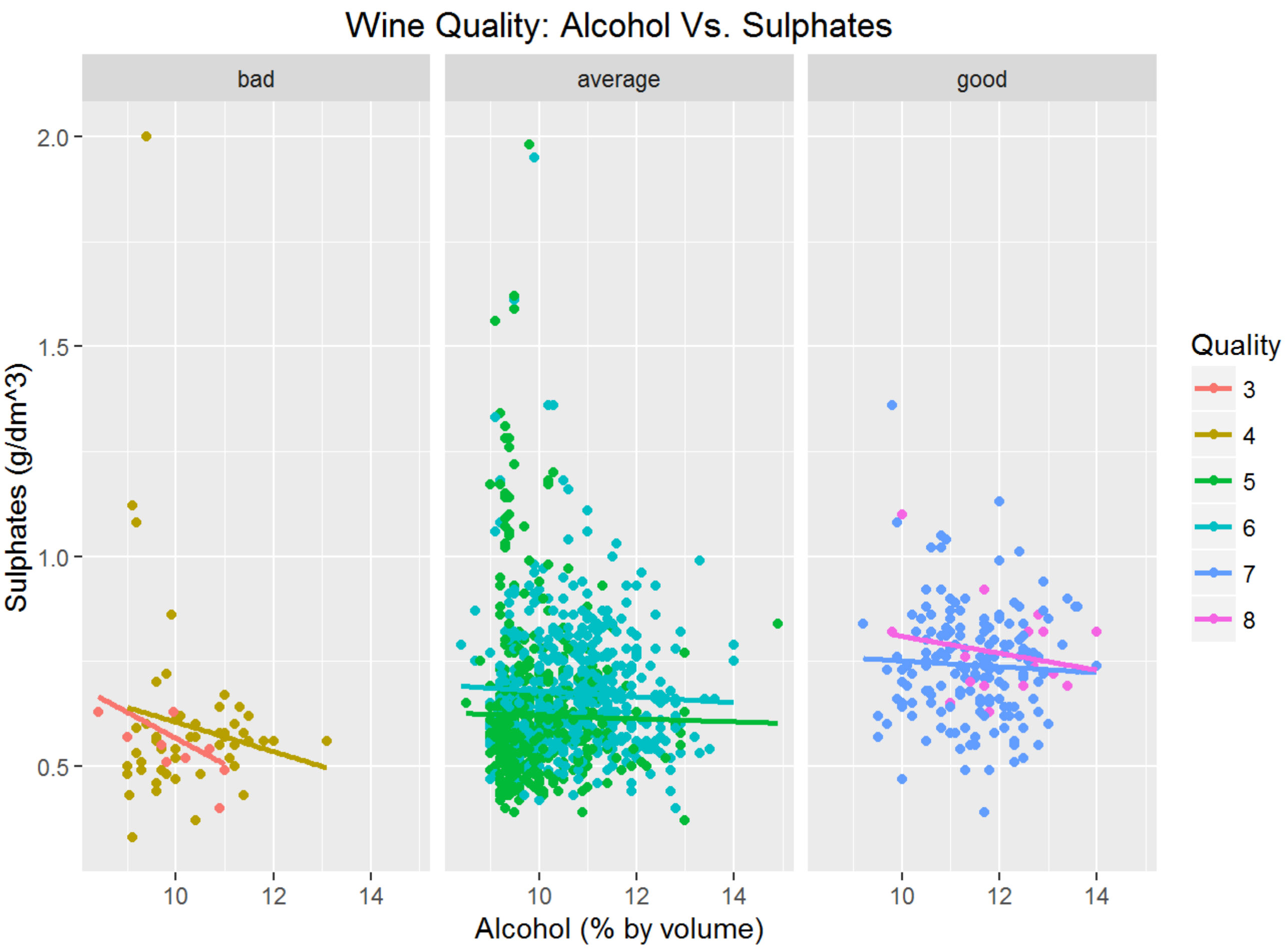
```
## [1] "Summary Stats: Volatile Acidity"
```

```
## quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4400  0.6475  0.8450  0.8845  1.0100  1.5800
## --------------------------------------------------------
## quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.230   0.530   0.670   0.694   0.870   1.130
## --------------------------------------------------------
## quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.180   0.460   0.580   0.577   0.670   1.330
## --------------------------------------------------------
## quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1600  0.3800  0.4900  0.4975  0.6000  1.0400
## --------------------------------------------------------
## quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3000  0.3700  0.4039  0.4850  0.9150
## --------------------------------------------------------
## quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2600  0.3350  0.3700  0.4233  0.4725  0.8500
```

Description Two

This scatter plot is broken down by rating. From the graphic it is easy to see that higher rated wines typically have higher alcohol content and lower volatile acidity. Wines with high volatile acidity have higher concentrations of acetic acid which can give the wine a vinegar taste. Regression lines are overlaid to illustrate the relationship between alcohol content and volatile acidity concentration by quality category.

Plot Three (Quality Vs Alcohol and Sulphates)



```
## [1] "Summary Stats: Sulphates"
## quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4000  0.5125  0.5450  0.5700  0.6150  0.8600 
## -------------------------------------------------------- 
## quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3300  0.4900  0.5600  0.5964  0.6000  2.0000 
## -------------------------------------------------------- 
## quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.370   0.530   0.580   0.621   0.660   1.980 
## -------------------------------------------------------- 
## quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4000  0.5800  0.6400  0.6753  0.7500  1.9500 
## -------------------------------------------------------- 
## quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3900  0.6500  0.7400  0.7413  0.8300  1.3600 
## -------------------------------------------------------- 
## quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.6300  0.6900  0.7400  0.7678  0.8200  1.1000
```

```
## [1] "Variance: Sulphates"
```

```
## rating: bad
## [1] 0.05032079
## -------------------------------------------------------- 
## rating: average
## [1] 0.02800062
## -------------------------------------------------------- 
## rating: good
## [1] 0.01796624
```

Description Three

This scatter plot is broken down by rating. In general, wine samples with higher alcohol content and higher sulphate concentrations were rated better. Individual wine samples with the largest sulphate concentrations can be found in the bad and average categories. The greatest variance in sulphate concentrations exists in the average category. Regression lines are overlaid to illustrate the relationship between alcohol content and sulphate concentration by quality category.

Reflection

The goal of this project was to work with a real-world dataset and get familiar with some exploratory data analysis techniques in R. The redwine dataset used here contains 1599 observations of 12 variables (after removing “X”). The data was tidy in that variables were provided in columns and each row represented an observation. Not much data wrangling was necessary.

To simplify some aspects of my code and to facilitate creating certain figures and summary tables throughout the project, I created a duplicate wine dataframe with variable quantities represented only in a numeric form. I called this dataframe “wine_numeric”. With the original wine dataframe (“wine”) I transformed the quality variable into an ordered factor. I also added an additional “rating” categorical variable to the dataframe to break down quality into bad, average, and good. This helped make complex, cluttered graphics easier to read later in the project. I realize that creating a duplicate dataframe is not best practice since it uses computer resources to store it in memory. I did struggle to some degree switching quickly back and forth between ordered factors and numerical values and having 2 separate dataframes helped to alleviate this issue.

My ultimate goal was to understand which physiochemical properties in the provided dataset had the biggest influence on the perceived quality of wine. I attacked the project head-on, exploring the relationship between all the variables simultaneously with a correlation diagram (corrgram). From this I was able to quickly and easily identify the variables most postitively or negatively correlated with quality. I found that alcohol, volatile acidity, and sulphate concentration had the highest absolute correlation value with quality and I subsequenly explored these variables and relationships thoroughly.

After extensive univariate, bivariate, and multivariate data analysis, I built a linear model to predict wine quality from a given a set of inputs. The final model utilizes the 4 variables with the highest absolute correlation with wine quality (alcohol, volatile acidity, sulphates, and citric acid) as inputs. I tested the model using a subset of the wine dataset and plotted the difference between modeled and actual observations as a scatterplot. From the scatterplot of the residual error it is clear that some bias exists in the data. The model seemed to systematically overestimate the quality of bad wines and underestimate the quality of good wines. Furthermore, the model could only explain about 30% of the total variance in wine quality. These findings suggest that a linear model may not be the best choice for wine quality prediction and that more inputs might be necessary in the model.

In future analysis of wine quality it would be nice to evaluate more input variables. I believe that defining the quality of wine in some quantitative way without a wine expert’s opinion would also make analysis more robust and less prone to personal preference bias.