My wine project

Javier Monedero

Executive summary

In this report, I am going to explore the wineQualityReds data set, which is about quality in red wines. The inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Data from 2009.

Reference: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Input variables (based on physicochemical tests): 1 - fixed acidity (tartaric acid - g / dmˆ3) 2 - volatile acidity (acetic acid - g / dmˆ3) 3 - citric acid (g / dmˆ3) 4 - residual sugar (g / dmˆ3) 5 - chlorides (sodium chloride - g / dmˆ3 6 - free sulfur dioxide (mg / dmˆ3) 7 - total sulfur dioxide (mg / dmˆ3) 8 - density (g / cmˆ3) 9 - pH 10 - sulphates (potassium sulphate - g / dm3) 11 - alcohol (% by volume) Output variable (based on sensory data): 12 - quality (score between 0 and 10)

None missing values.

Loading and cleaning data

Loading data:

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

Seeing the variables' names:

```
names(data)
```

```
##  [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"
```

General information:

```
str(data)
```

```
## '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 ...
```

Data set summary:

```
summary(data)
```

```
##        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
```

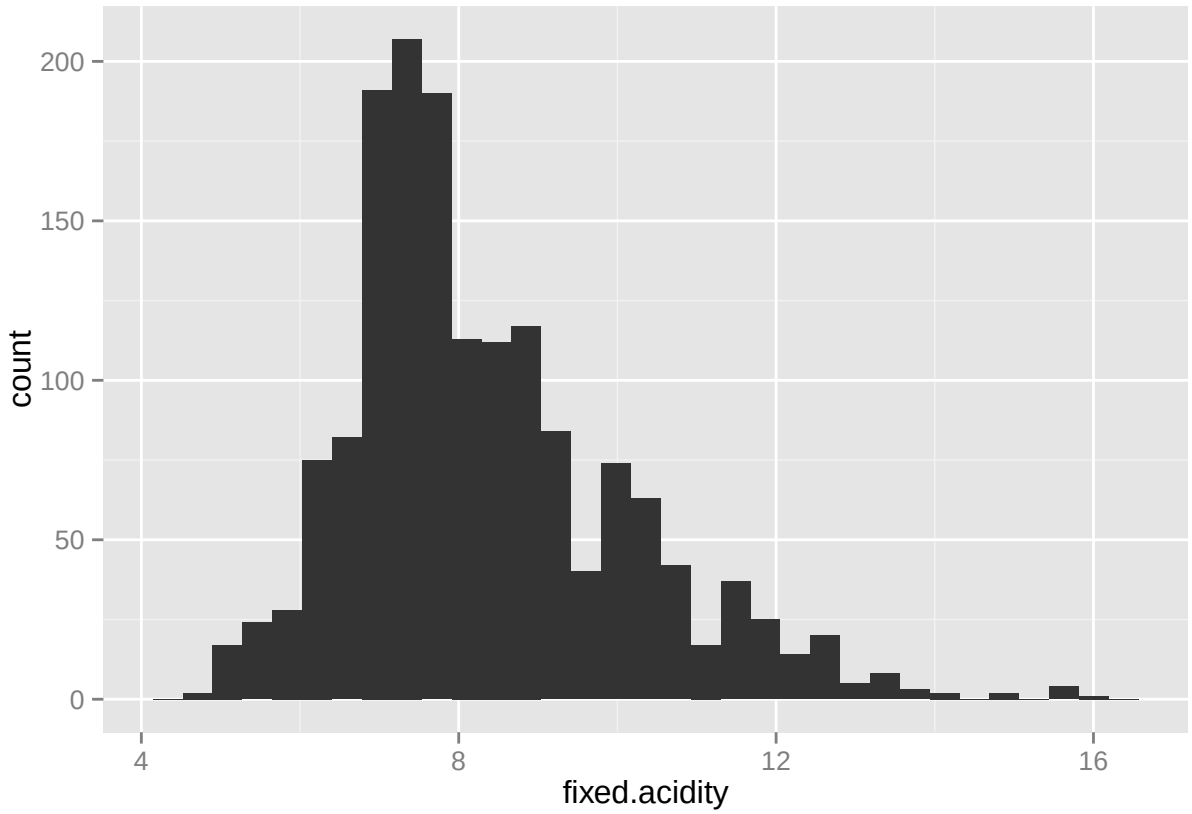
The mean wine quality is 5.636, ranging from 3 to 8. The other characteristics are numeric and thus, I can not comment anything at this level.

```
library(ggplot2)
```

Fixed acidity plot:

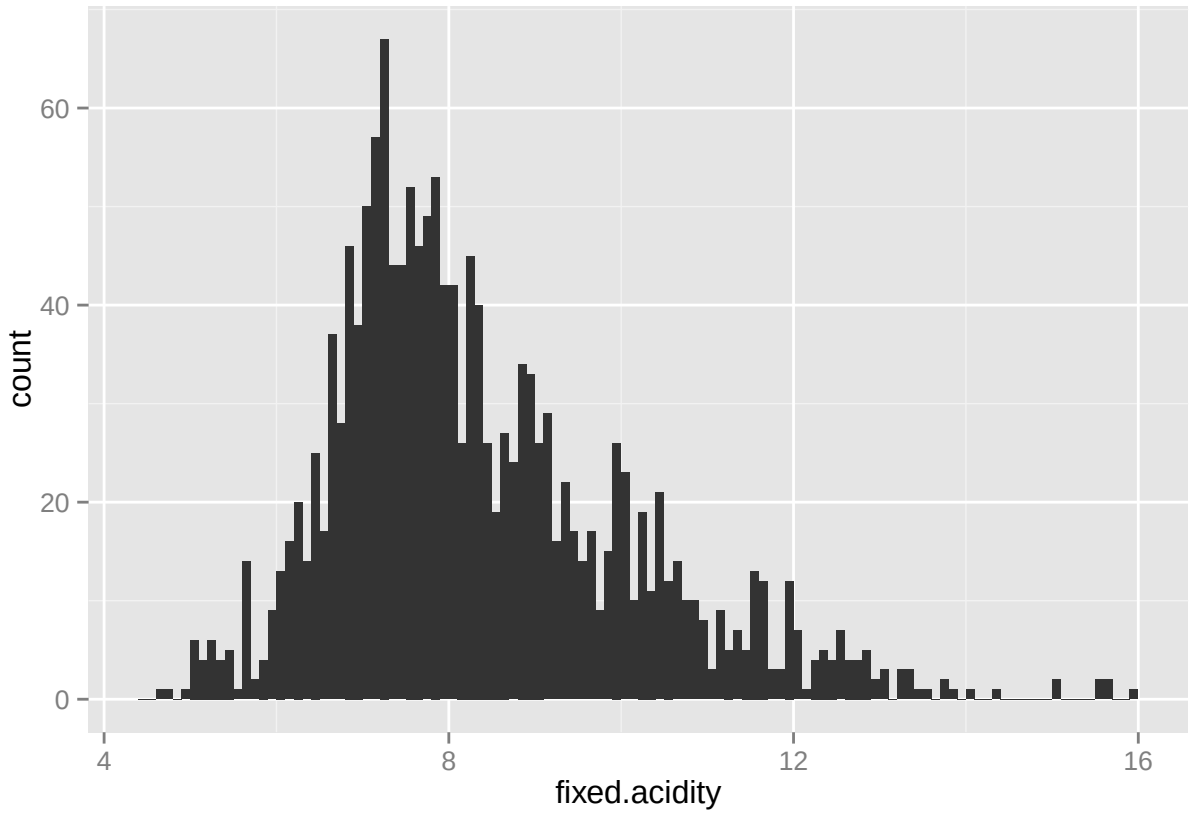
```
qplot(data = data, x = fixed.acidity)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot(data = data, x = fixed.acidity, binwidth = 0.1)
```

```
## Warning: position_stack requires constant width: output may be incorrect
```



Transformed the previous plot to better understand the distribution of fixed acidity. The tranformed fixed acididty distribution appears monomodal with the fixed acidity peaking around 7.

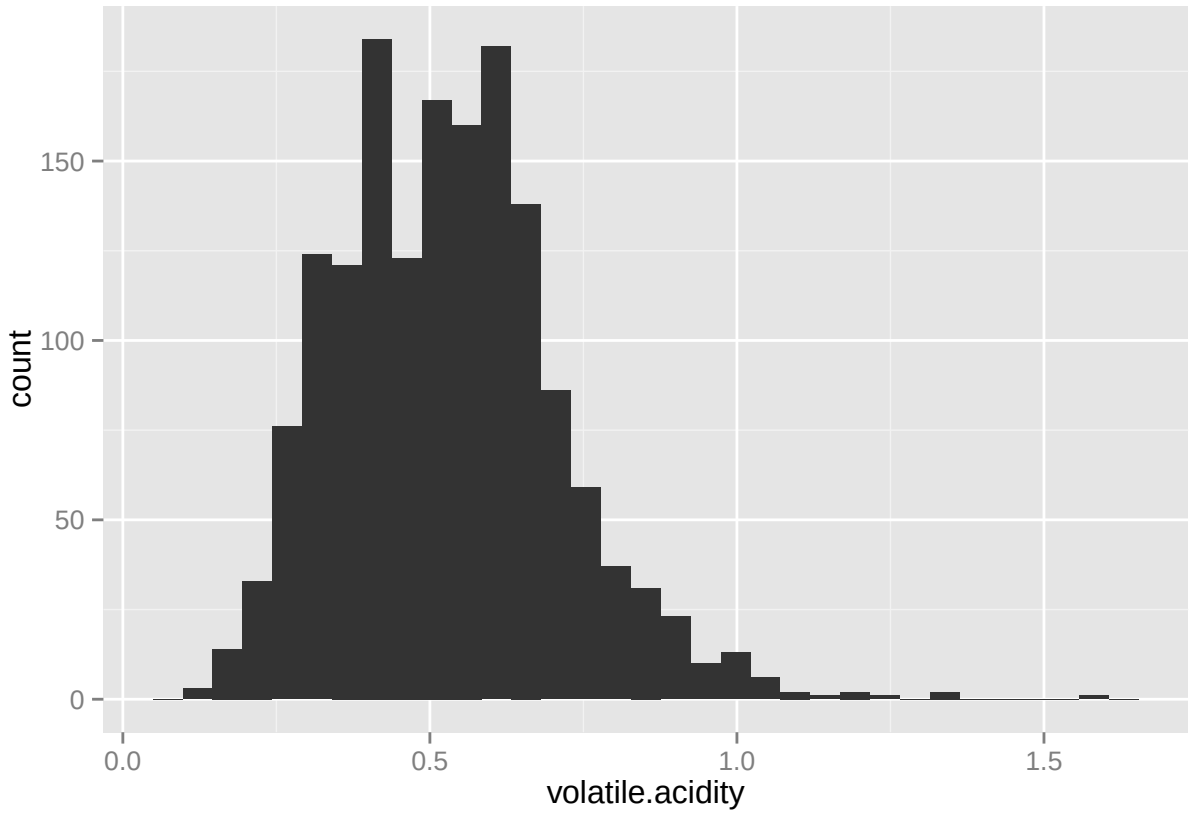
```
summary(data$fixed.acidity)
```

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

Volatile acidity plot:

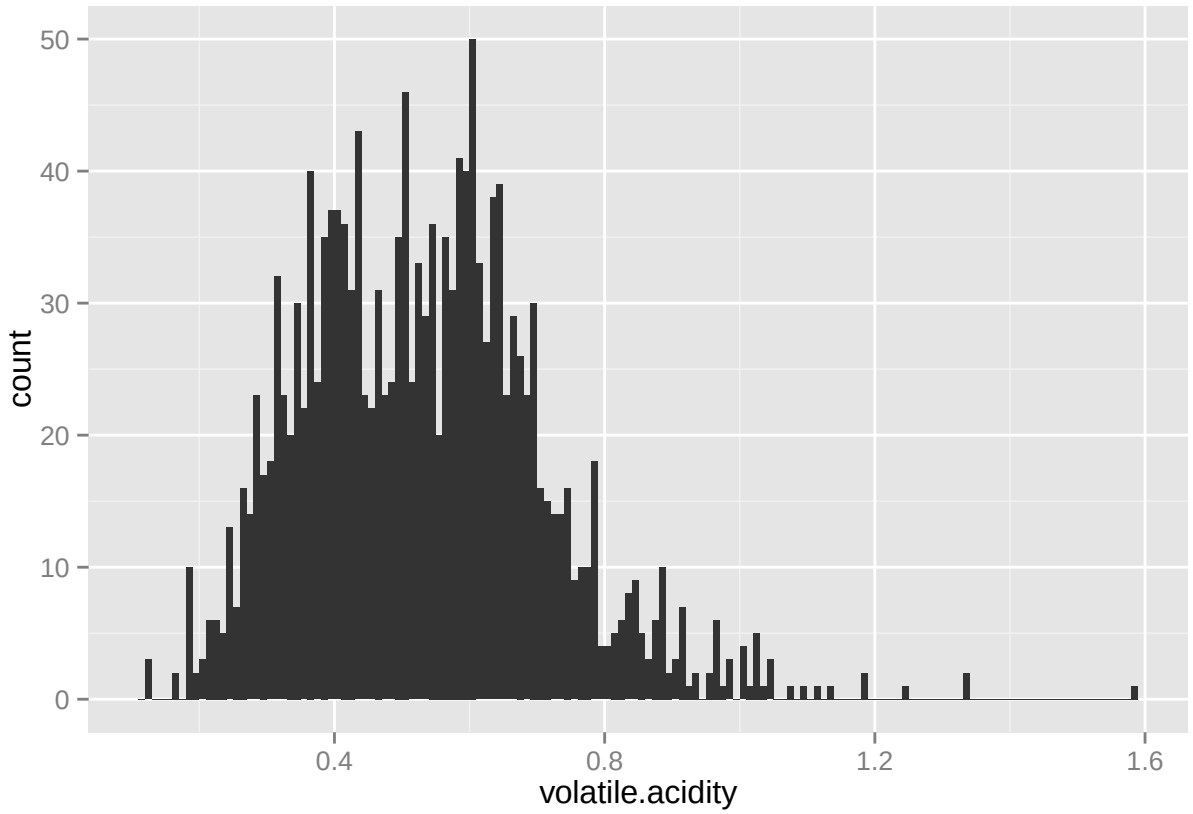
```
qplot(data = data, x = volatile.acidity)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot(data = data, x = volatile.acidity, binwidth = 0.01)
```

```
## Warning: position_stack requires constant width: output may be incorrect
```



This time the transformed plot shows two peaks with the volatile acidity peaking around 0.4, and then reaching its maximum at about 0.6.

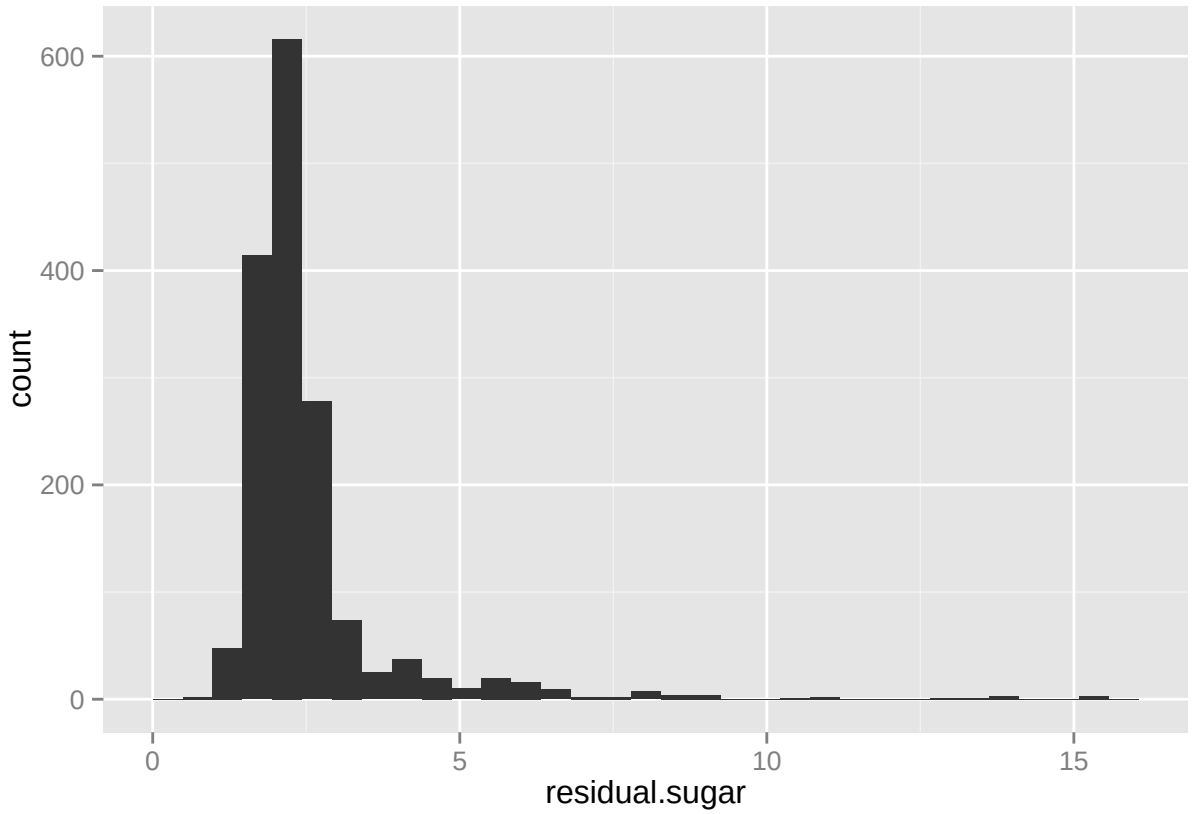
```
summary(data$volatile.acidity)
```

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

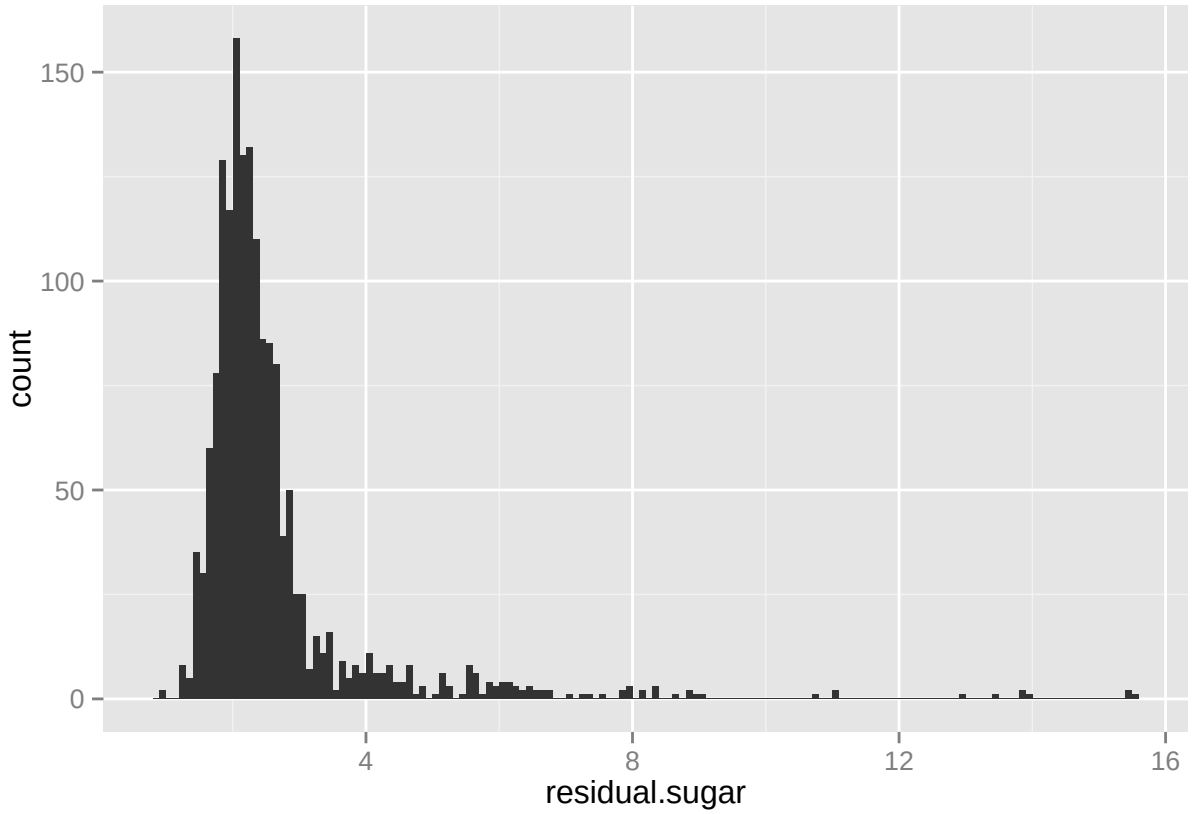
Residual sugar plot:

```
qplot(data = data, x = residual.sugar)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot(data = data, x = residual.sugar, binwidth = 0.1)
```



Residual sugar shows a unique peak around 2.

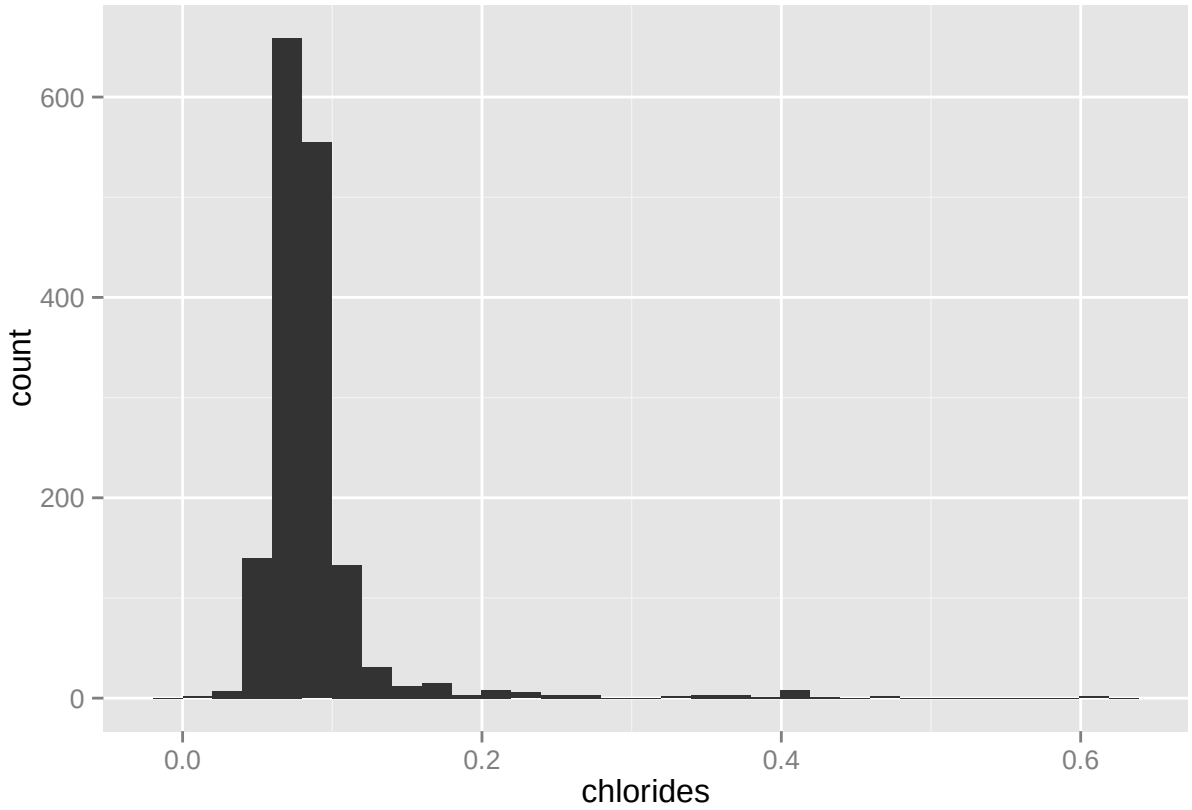
```
summary(data$fresidual.sugar)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

Chlorides plot:

```
qplot(data = data, x = chlorides)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Clearly, there is only one peak at about something a little less than 0.1.

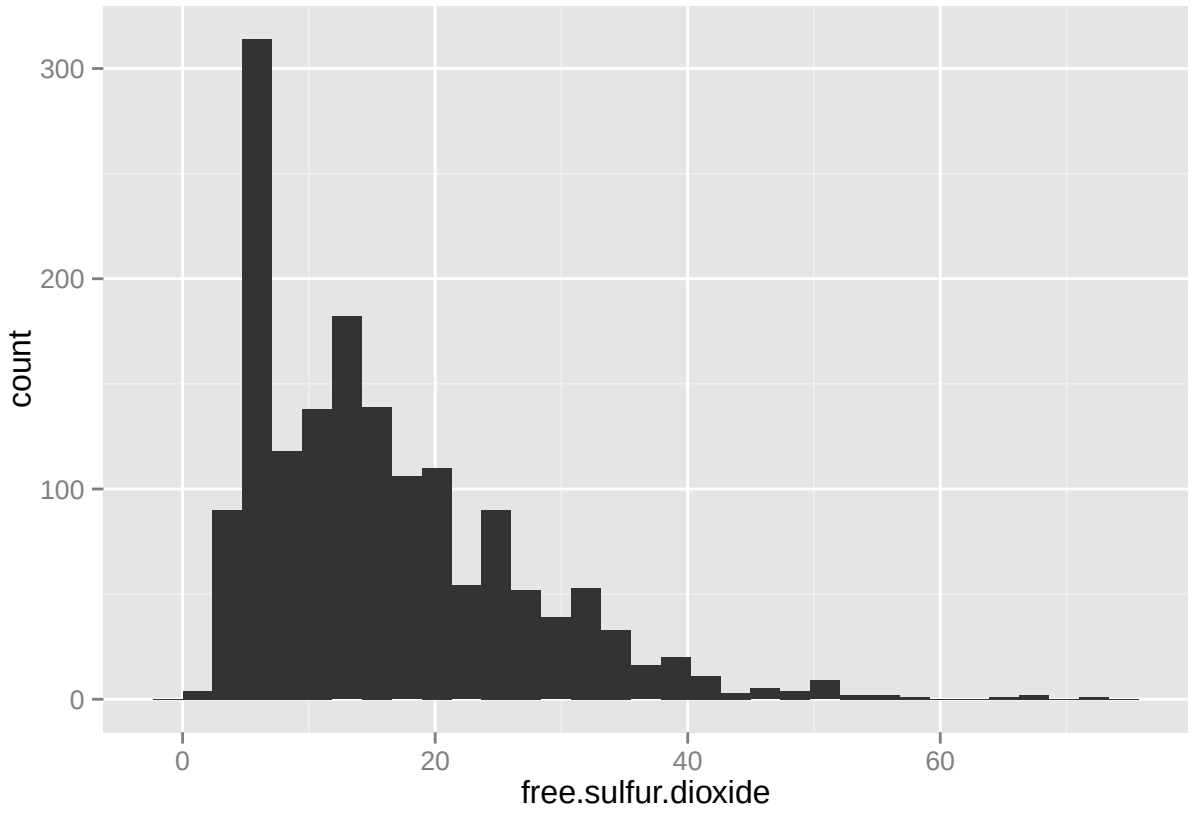
```
summary(data$chlorides)
```

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

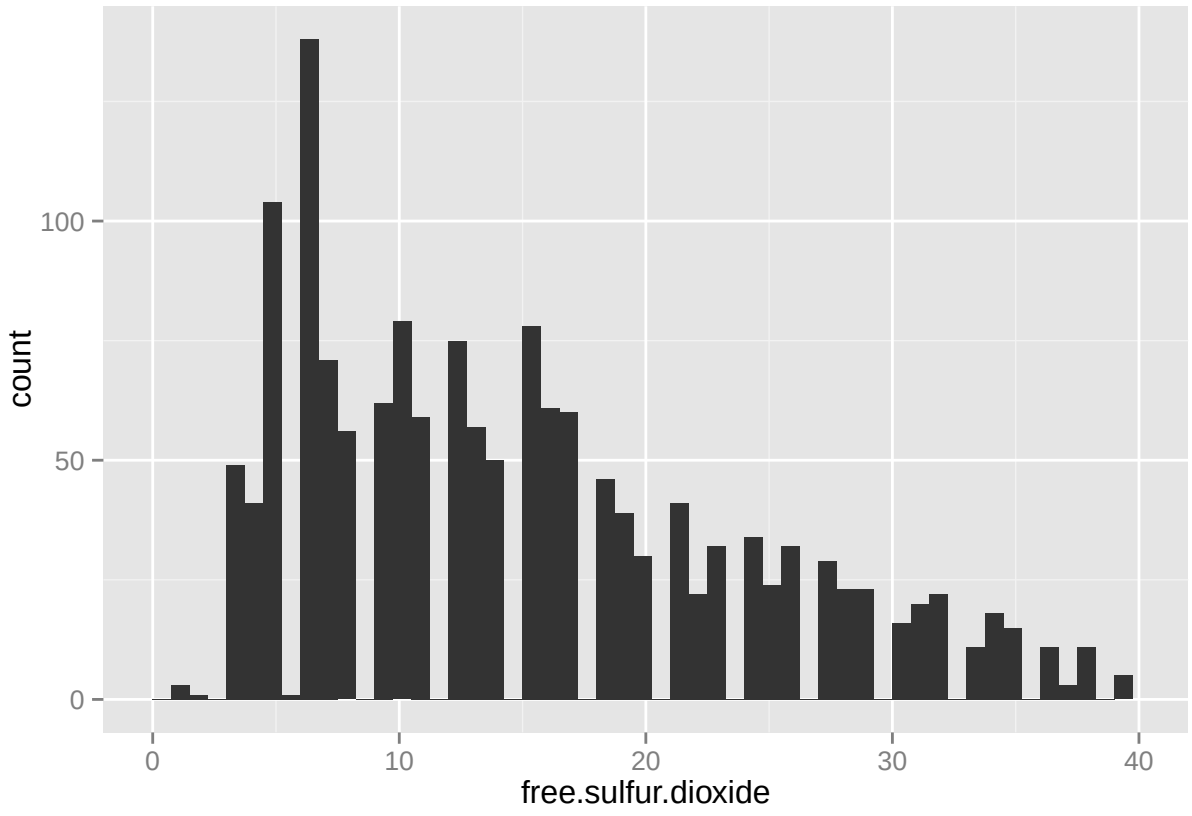
Free sulfur dioxide plot:

```
qplot(data = data, x = free.sulfur.dioxide)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot(data = data, x = free.sulfur.dioxide, binwidth = 0.75) + scale_x_continuous(limits = c(0, 40))
```



Transformed the long tail data to better understand the distribution of free sulfur dioxide. The tranformed free sulfur dioxide distribution appears momodal with the price peaking around 6.

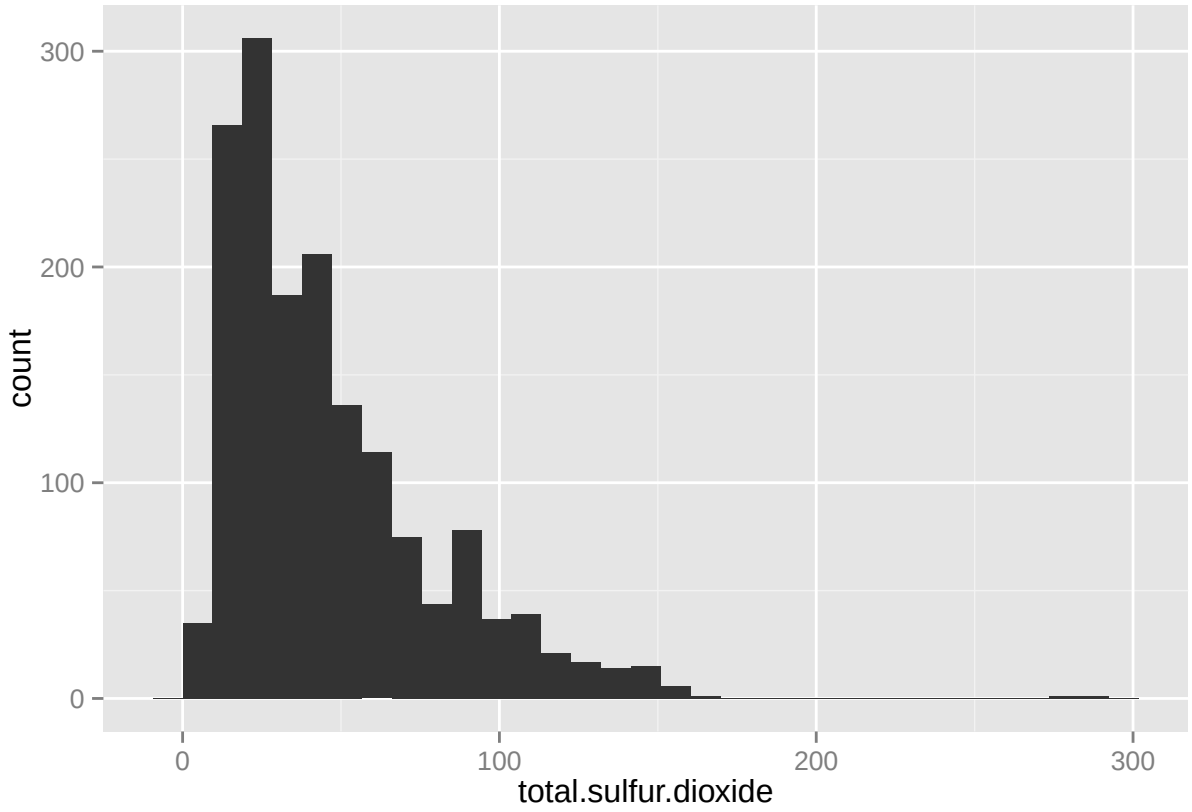
```
summary(data$free.sulfur.dioxide)
```

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

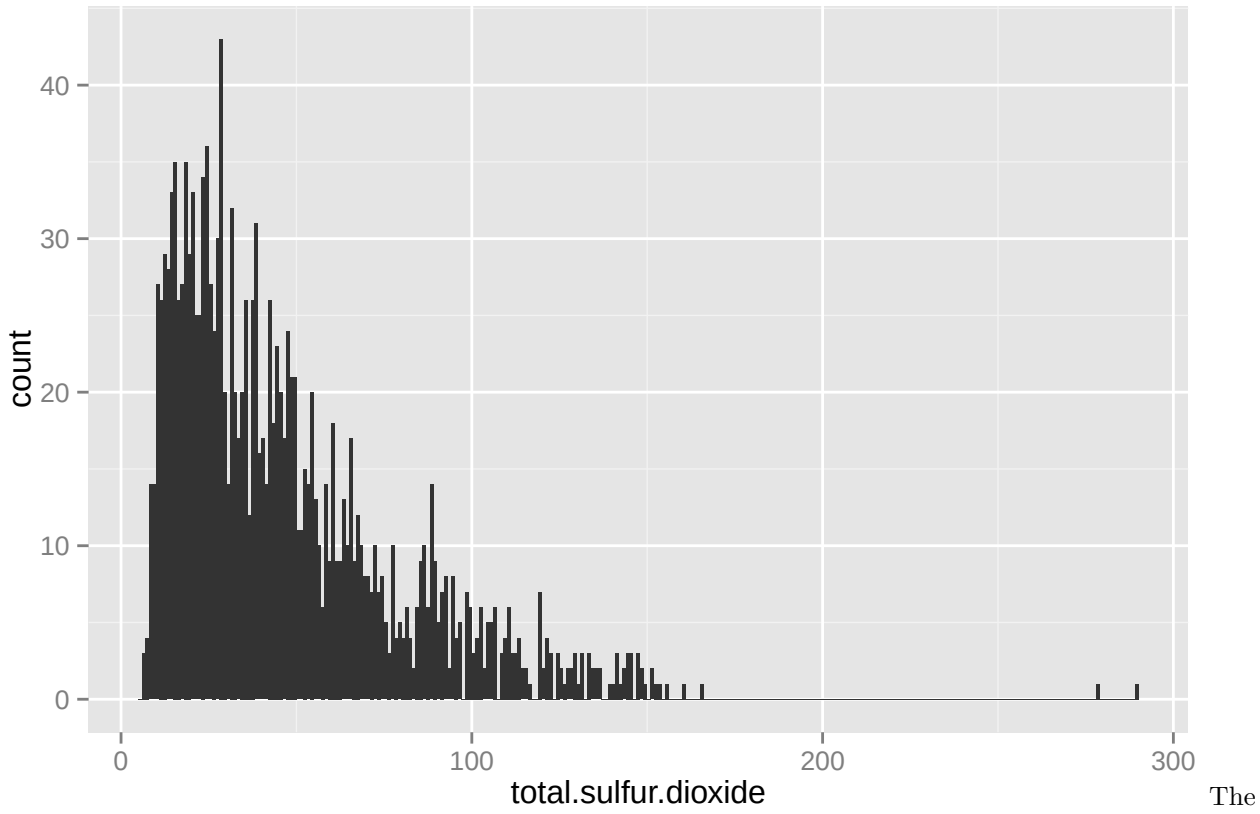
Total sulfur dioxide plot:

```
qplot(data = data, x = total.sulfur.dioxide)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot(data = data, x = total.sulfur.dioxide, binwidth = 1)
```



The total sulfur dioxide distribution follows a distribution with one peak around 30, and there are two point with extreme values: 280 and 290.

```
summary(data$total.sulfur.dioxide)
```

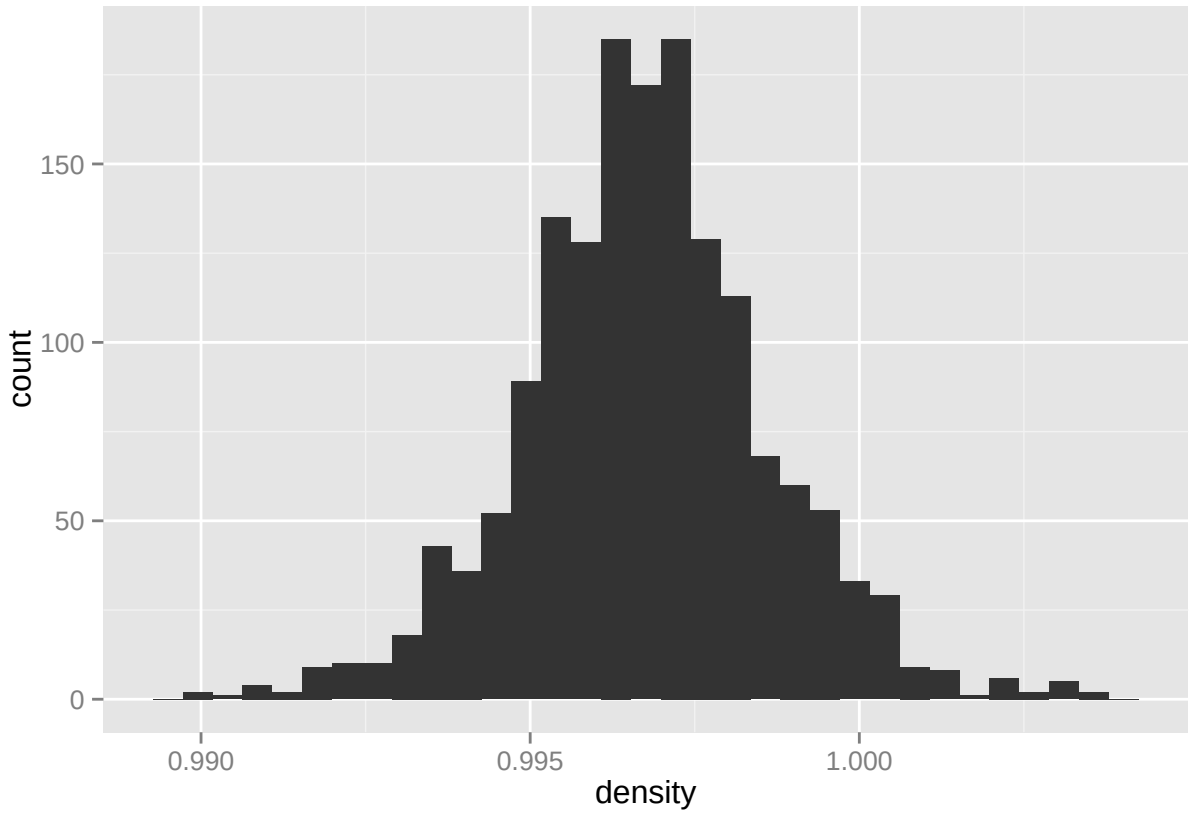
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

Density plot:

```
qplot(data = data, x = density)
```

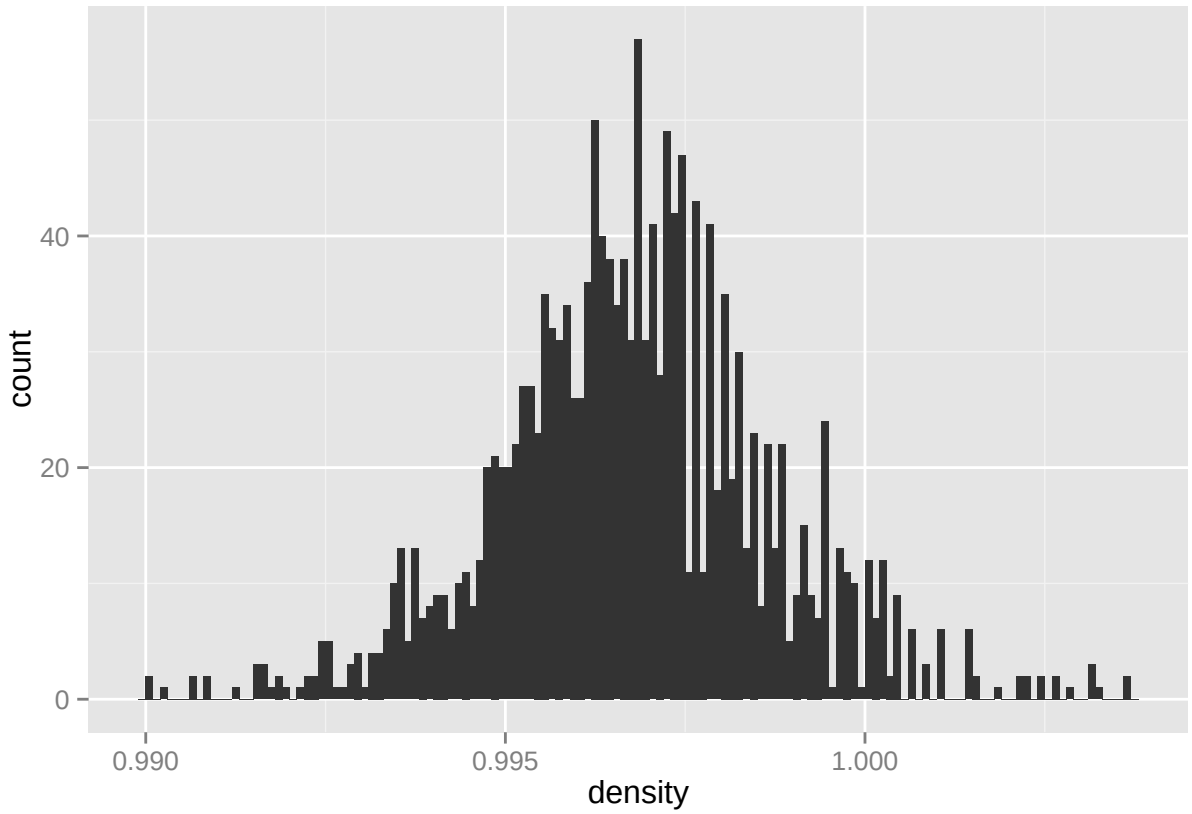
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

```
## Warning: position_stack requires constant width: output may be incorrect
```




```
qplot(data = data, x = density, binwidth = 0.0001)
```

```
## Warning: position_stack requires constant width: output may be incorrect
```



The peak is reached at 0.997.

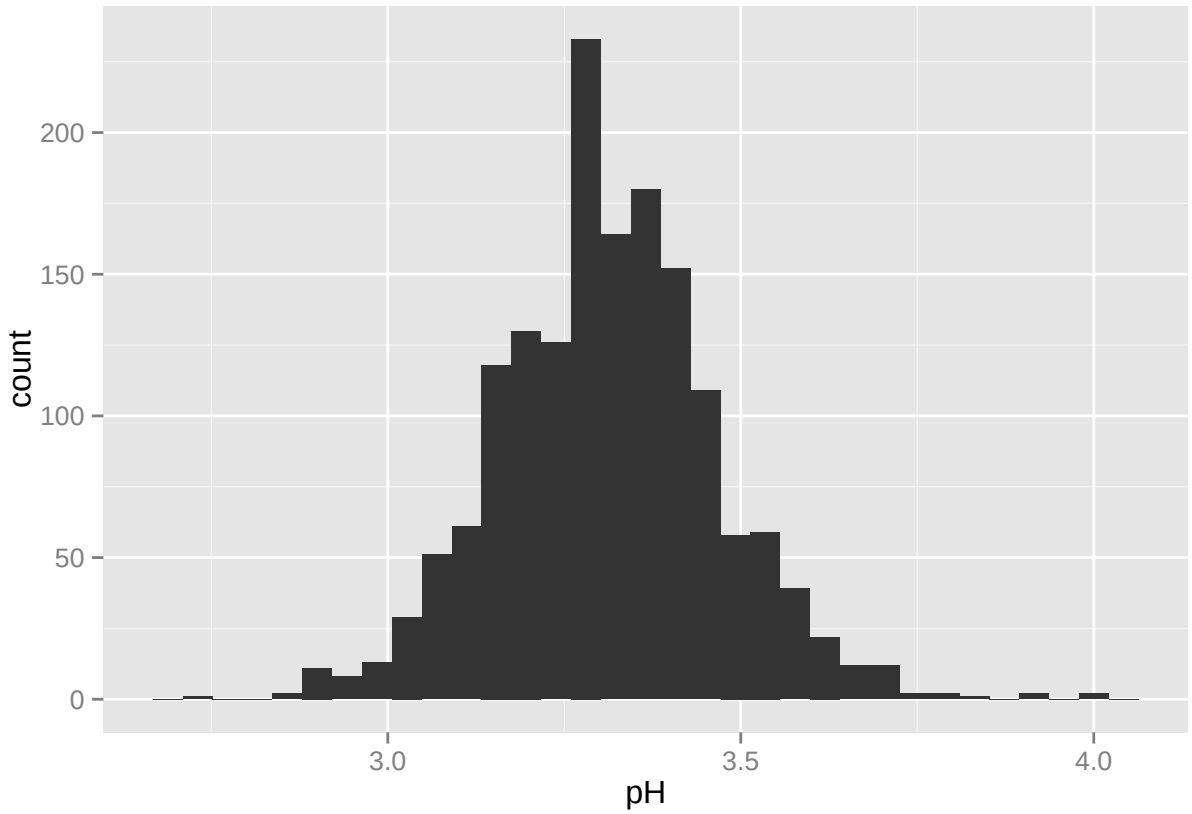
```
summary(data$density)
```

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

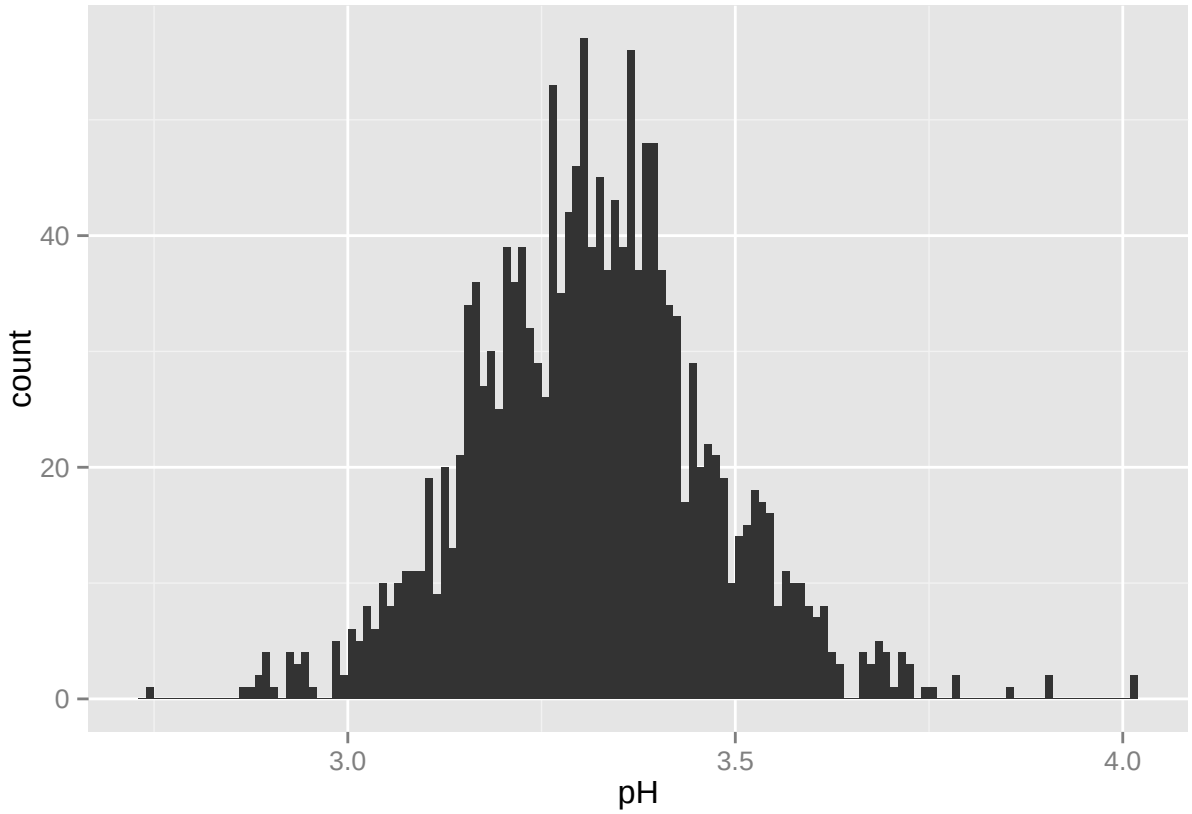
pH plot:

```
qplot(data = data, x = pH)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot(data = data, x = pH, binwidth = 0.01)
```



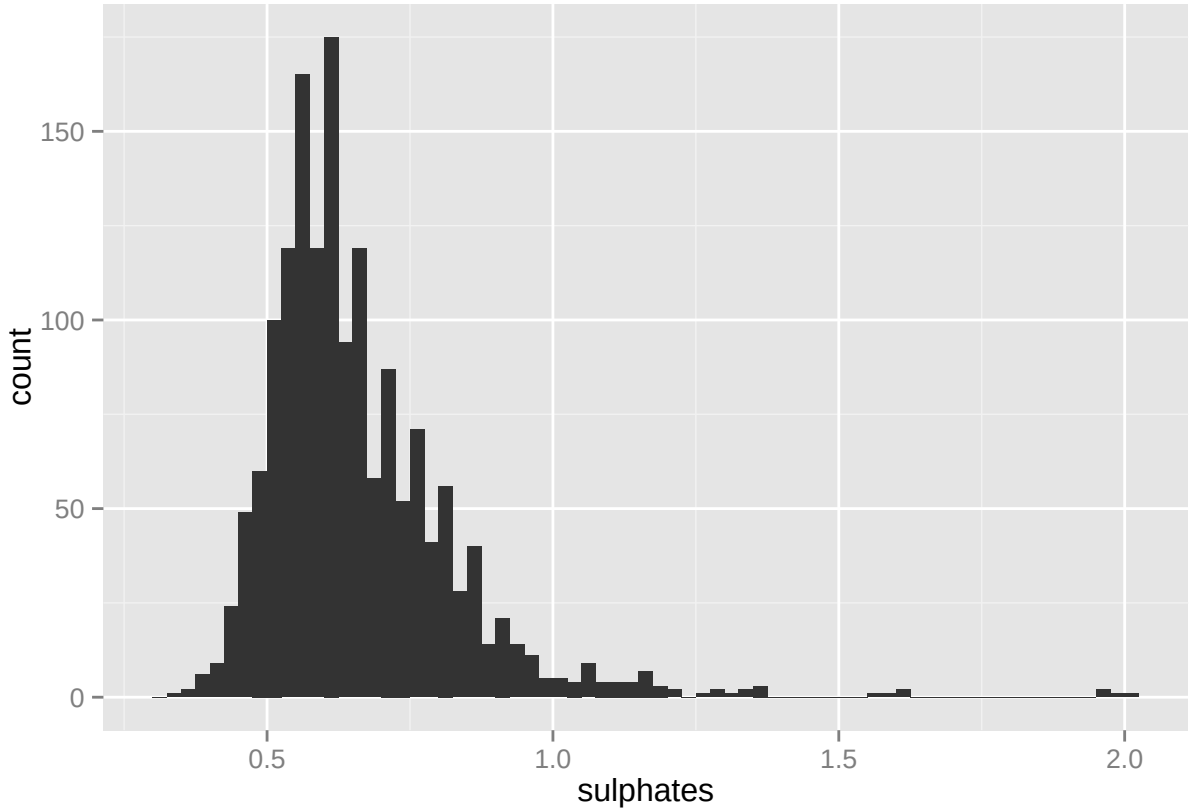
Again, there is only one peak at 3.35.

```
summary(data$pH)
```

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

Sulphates plot:

```
qplot(data = data, x = sulphates, binwidth = 0.025)
```



There is one maximum peak at 0.65.

```
summary(data$sulphates)
```

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

```
summary(ifelse(data$sulphates > 0.5 & data$sulphates < 1, TRUE, FALSE))
```

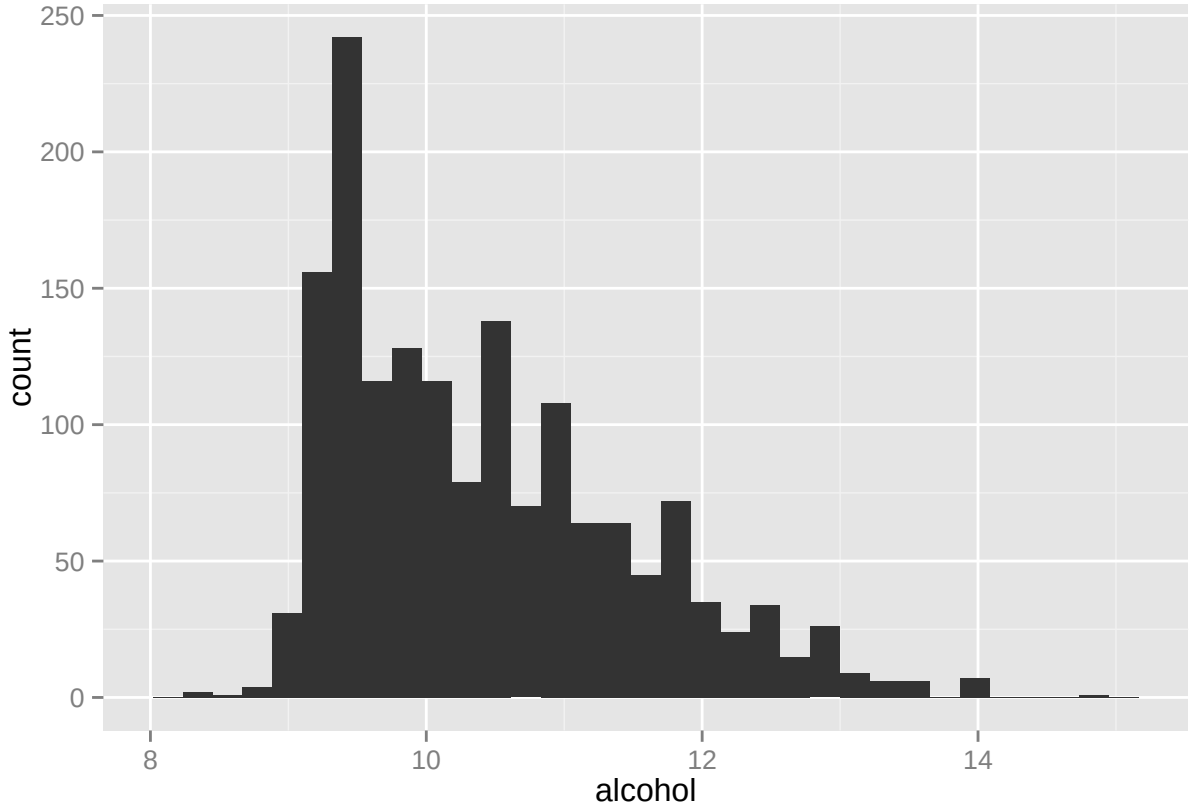
```
##    Mode   FALSE    TRUE    NA's
## logical     237    1362       0
```

The vast majority of the sulphates values are in the range of 0.5 to 1.

Alcohol plot:

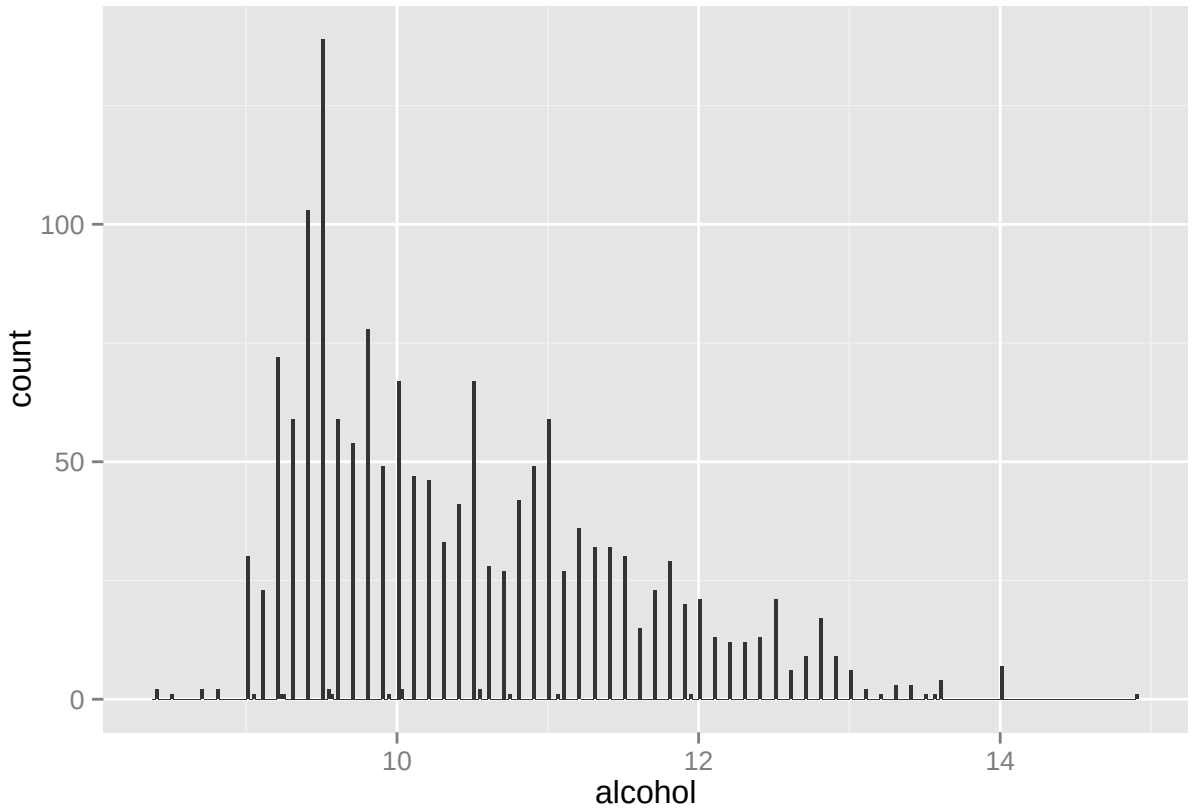
```
qplot(data = data, x = alcohol)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot(data = data, x = alcohol, binwidth = 0.02)
```

```
## Warning: position_stack requires constant width: output may be incorrect
```



It shows one peak at 9.5, with a high varitation for this category represented by the long tail.

```
summary(data$alcohol)
```

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

```
summary(ifelse(data$alcohol > 9 & data$alcohol < 11.5, TRUE, FALSE))
```

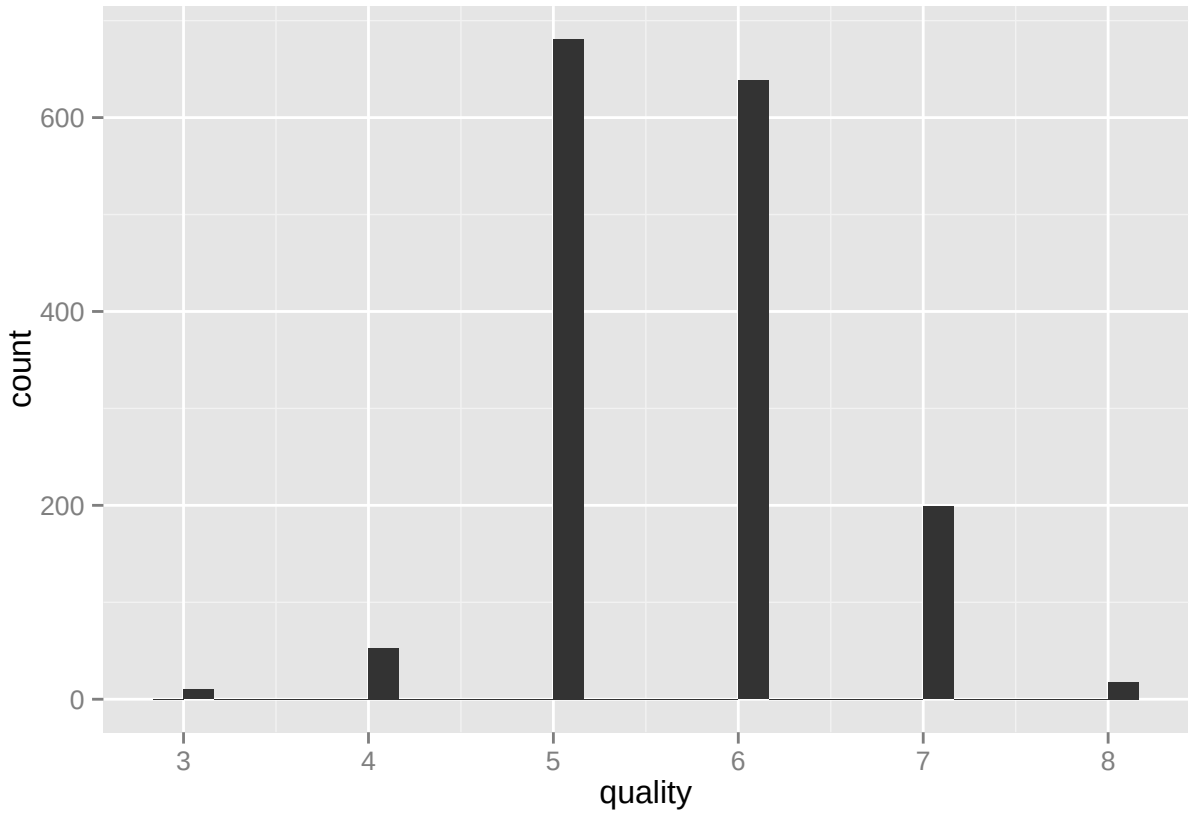
```
##    Mode   FALSE    TRUE    NA's
## logical     317    1282       0
```

More than two thirds of alcohol values are in the range of 9 to 11.5.

Quality plot:

```
qplot(data = data, x = quality)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Since quality values are integers, there is no need to modify this plot.

```
summary(data$quality)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

Due to the fact that the high quality wines tend to be the most pensive, a subset has been produced with the red wines with quality above 6. Here is a summary:

```
best <- subset(data, quality > 6)
summary(best)
```

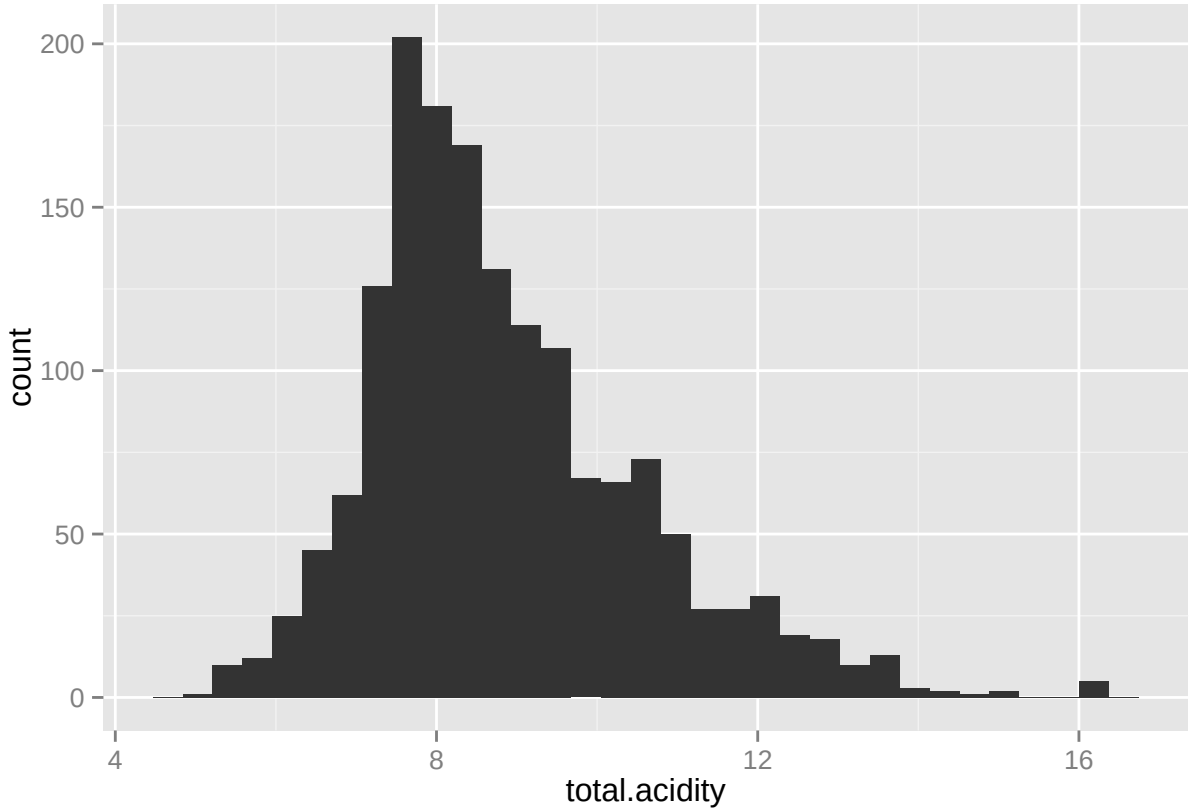
```
##        X          fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   8.0   Min.   : 4.900   Min.   :0.1200   Min.   :0.0000
##  1st Qu.: 482.0   1st Qu.: 7.400   1st Qu.:0.3000   1st Qu.:0.3000
##  Median : 939.0   Median : 8.700   Median :0.3700   Median :0.4000
##  Mean   : 831.7   Mean   : 8.847   Mean   :0.4055   Mean   :0.3765
##  3rd Qu.:1089.0   3rd Qu.:10.100   3rd Qu.:0.4900   3rd Qu.:0.4900
##  Max.   :1585.0   Max.   :15.600   Max.   :0.9150   Max.   :0.7600
##  residual.sugar    chlorides       free.sulfur.dioxide
##  Min.   :1.200   Min.   :0.01200   Min.   : 3.00
##  1st Qu.:2.000   1st Qu.:0.06200   1st Qu.: 6.00
##  Median :2.300   Median :0.07300   Median :11.00
##  Mean   :2.709   Mean   :0.07591   Mean   :13.98
##  3rd Qu.:2.700   3rd Qu.:0.08500   3rd Qu.:18.00
##  Max.   :8.900   Max.   :0.35800   Max.   :54.00
```

```
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  7.00       Min.   :0.9906   Min.   :2.880   Min.   :0.3900
##  1st Qu.: 17.00       1st Qu.:0.9947   1st Qu.:3.200   1st Qu.:0.6500
##  Median : 27.00       Median :0.9957   Median :3.270   Median :0.7400
##  Mean   : 34.89       Mean   :0.9960   Mean   :3.289   Mean   :0.7435
##  3rd Qu.: 43.00       3rd Qu.:0.9973   3rd Qu.:3.380   3rd Qu.:0.8200
##  Max.   :289.00       Max.   :1.0032   Max.   :3.780   Max.   :1.3600
##     alcohol         quality
##  Min.   : 9.20   Min.   :7.000
##  1st Qu.:10.80   1st Qu.:7.000
##  Median :11.60   Median :7.000
##  Mean   :11.52   Mean   :7.083
##  3rd Qu.:12.20   3rd Qu.:7.000
##  Max.   :14.00   Max.   :8.000
```

Total acidity:

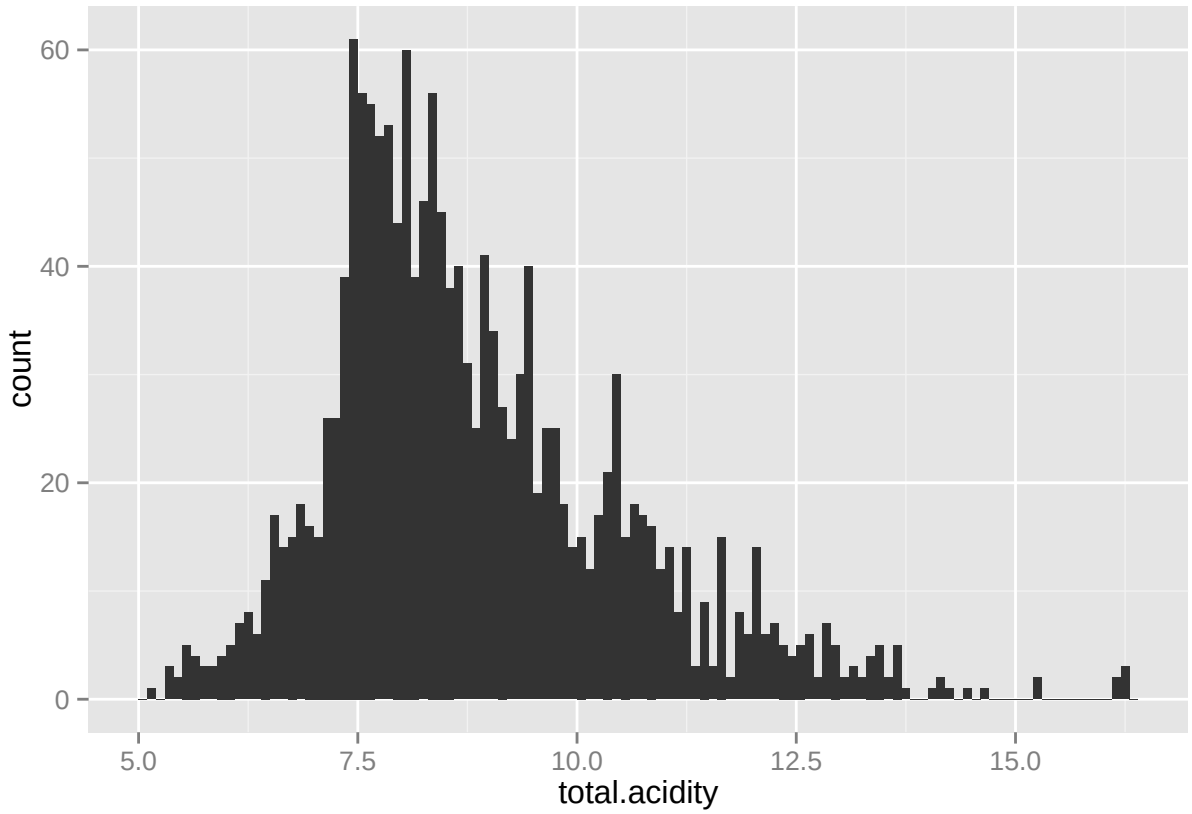
```
data$total.acidity <- data$fixed.acidity + data$volatile.acidity
qplot(data = data, x = total.acidity)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot(data = data, x = total.acidity, binwidth = 0.1)
```

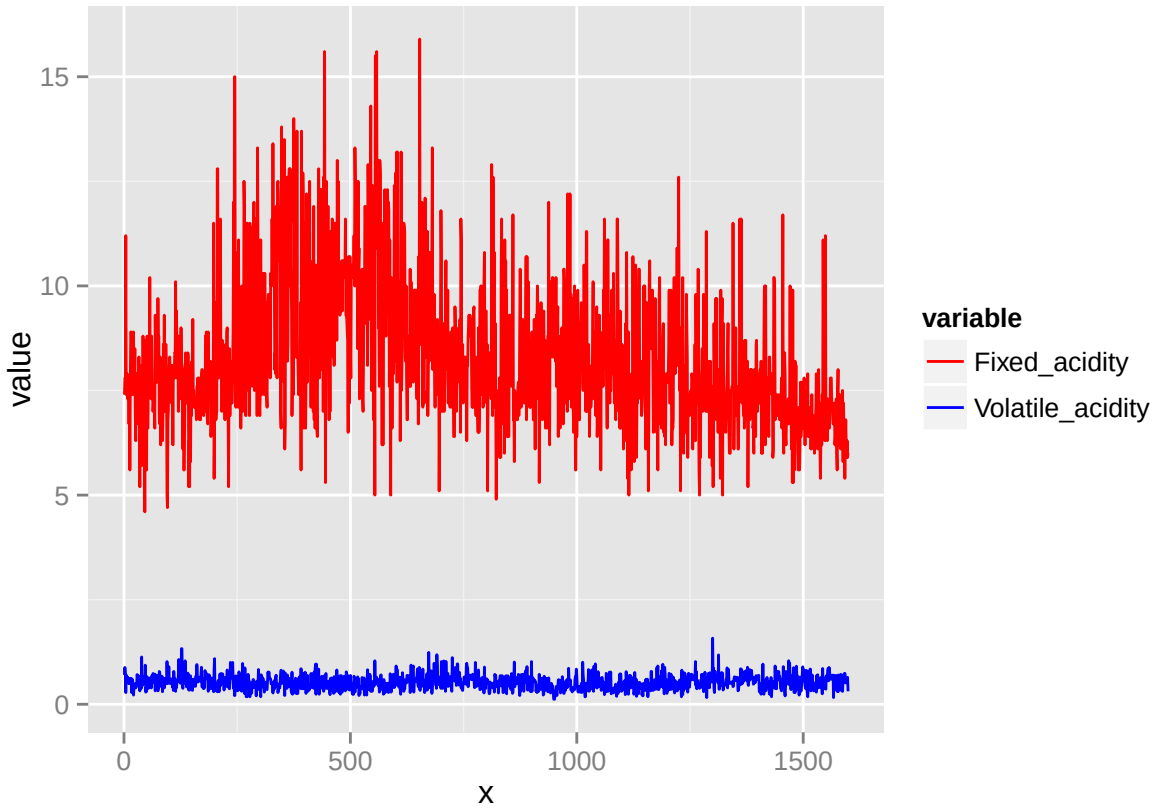
```
## Warning: position_stack requires constant width: output may be incorrect
```




```
summary(data$total.acidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.120   7.680   8.445   8.847   9.740  16.280
```

```
library(reshape2)
mix <- data.frame(x = data$X, Fixed_acidity = data$fixed.acidity, Volatile_acidity = data$volatile.acidi
dat.mix <- melt(mix, id.vars = "x")
ggplot(dat.mix, aes(x, value, colour = variable)) +
    geom_line() +
    scale_colour_manual(values = c("red", "blue"))
```




```
summary(ifelse(data$total.acidity > 7 & data$total.acidity < 10, TRUE, FALSE))
```

```
##    Mode   FALSE    TRUE    NA's
## logical     495    1104       0
```

Most of the total acidity values range between 7 and 10.

Univariate Analysis

What is the structure of your dataset?

There are 1599 red wines and my dataset has 12 variables. All are numeric ones, except for quality, which is integer. There are no NAs.

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

I am going to focus on the three most important features of a good wine: Total acidity (as the sum of fixed acidity plus volatile acidity), Sulphates and Alcohol. Thus, I need to create a new variable called total.acidity as the sum of both acidities.

In addition, I am going to compare results with the subset best, which includes the most quality red wines.

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

All the other variables (citric acid, residual sugar, chlorides, sulfur dioxide and pH) influence the wines quality. Nevertheless, their influence is limited and so, I am not going to consider these facts.

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

Yes. As I mentioned, I created a new variable called total.acidity in order to capture the acidity effect on the wine. Moreover, I decided to create a subset with the most valued red wines as it comes to quality.

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 original data set was already tidy, without NAs, so I avoided this task.

The sulphates qplot seems correct because there was only one maximum peak. This has sense as sulphates are added to control bacteria development while fermentation and it is legaly controlled.

Regarding alcohol, there was a los tail indicating the amount of sugars in the grapes. This also has sense since vineyards are located in a variety of locations with different temperatures and climate conditions, and this influences the sugar content before fermentation, and ultimately the wines alcohol content.

Similarly as before, total acidity showed a unique peak, but again with a long tail. Here, the climate conditions influence the acids content and this explains the plot.

Bivariate Plots Section

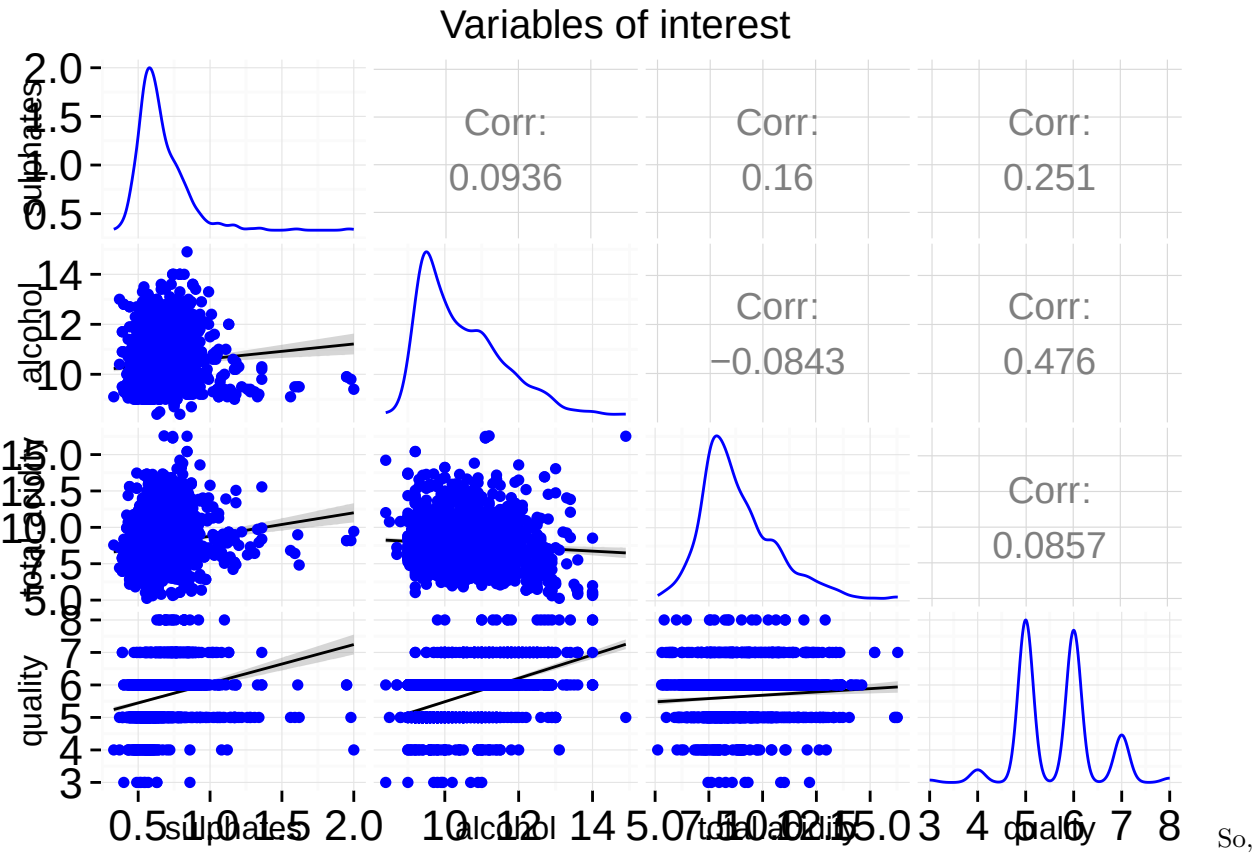
```
cor(data)
```

```
##                              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
## alcohol            0.245122841   -0.06166827     -0.202288027
## quality            0.066452608    0.12405165     -0.390557780
## total.acidity     -0.275247580    0.99482800     -0.156620601
##                   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
## alcohol               0.10990325    0.042075437 -0.221140545
## quality               0.22637251    0.013731637 -0.128906560
## total.acidity         0.62825187    0.117473729  0.102183639
##                     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
## alcohol                  -0.069408354          -0.20565394 -0.49617977
## quality                  -0.050656057          -0.18510029 -0.17491923
## total.acidity            -0.158241719          -0.10760684  0.68488647
##                               pH    sulphates     alcohol     quality
## X                      0.13600533 -0.125306999  0.24512284  0.06645261
## fixed.acidity         -0.68297819  0.183005664 -0.06166827  0.12405165
## volatile.acidity       0.23493729 -0.260986685 -0.20228803 -0.39055778
## citric.acid           -0.54190414  0.312770044  0.10990325  0.22637251
## residual.sugar        -0.08565242  0.005527121  0.04207544  0.01373164
## chlorides             -0.26502613  0.371260481 -0.22114054 -0.12890656
## free.sulfur.dioxide    0.07037750  0.051657572 -0.06940835 -0.05065606
## total.sulfur.dioxide  -0.06649456  0.042946836 -0.20565394 -0.18510029
## density               -0.34169933  0.148506412 -0.49617977 -0.17491923
## pH                     1.00000000 -0.196647602  0.20563251 -0.05773139
## sulphates             -0.19664760  1.000000000  0.09359475  0.25139708
## alcohol                0.20563251  0.093594750  1.00000000  0.47616632
## quality               -0.05773139  0.251397079  0.47616632  1.00000000
## total.acidity         -0.67314051  0.159560329 -0.08426530  0.08570932
##                     total.acidity
## X                     -0.27524758
## fixed.acidity          0.99482800
## volatile.acidity      -0.15662060
## citric.acid            0.62825187
## residual.sugar         0.11747373
## chlorides              0.10218364
## free.sulfur.dioxide   -0.15824172
## total.sulfur.dioxide  -0.10760684
## density                0.68488647
## pH                    -0.67314051
## sulphates              0.15956033
## alcohol               -0.08426530
## quality                0.08570932
```

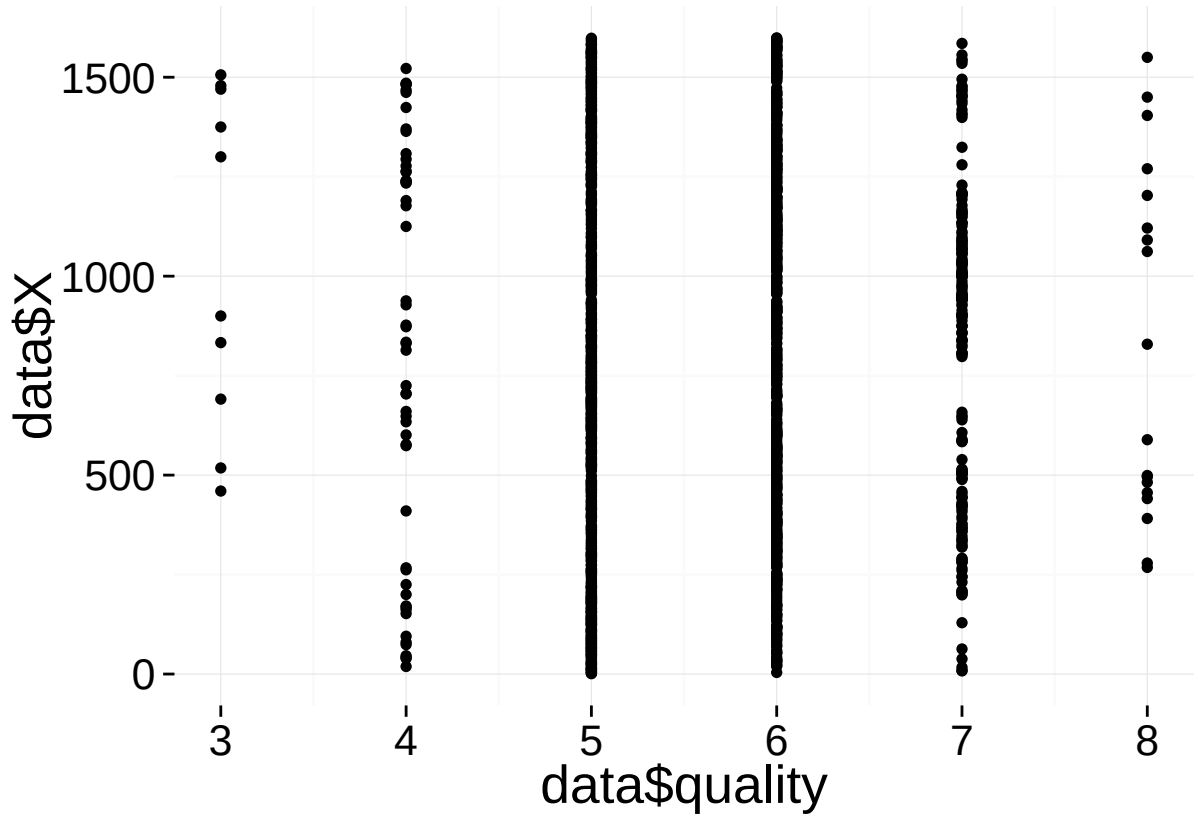
```
## total.acidity           1.00000000
```

Total acidity correlates very good with fixed acidity as shown above. The variable quality seems not to so well with a few variables with the exception of alcohol, meaning it is influenced by a lot of variables. Thus, alcohol, volatile acidity and sulfates are the most influential features in wine quality. Finally, total acidity is the sum of volatile acidity to measure the influence of total acitidy in quality.



So, most of quality is explained by alcohol, sulphates and total.acidity.

```
qplot(data$quality, data$X)
```

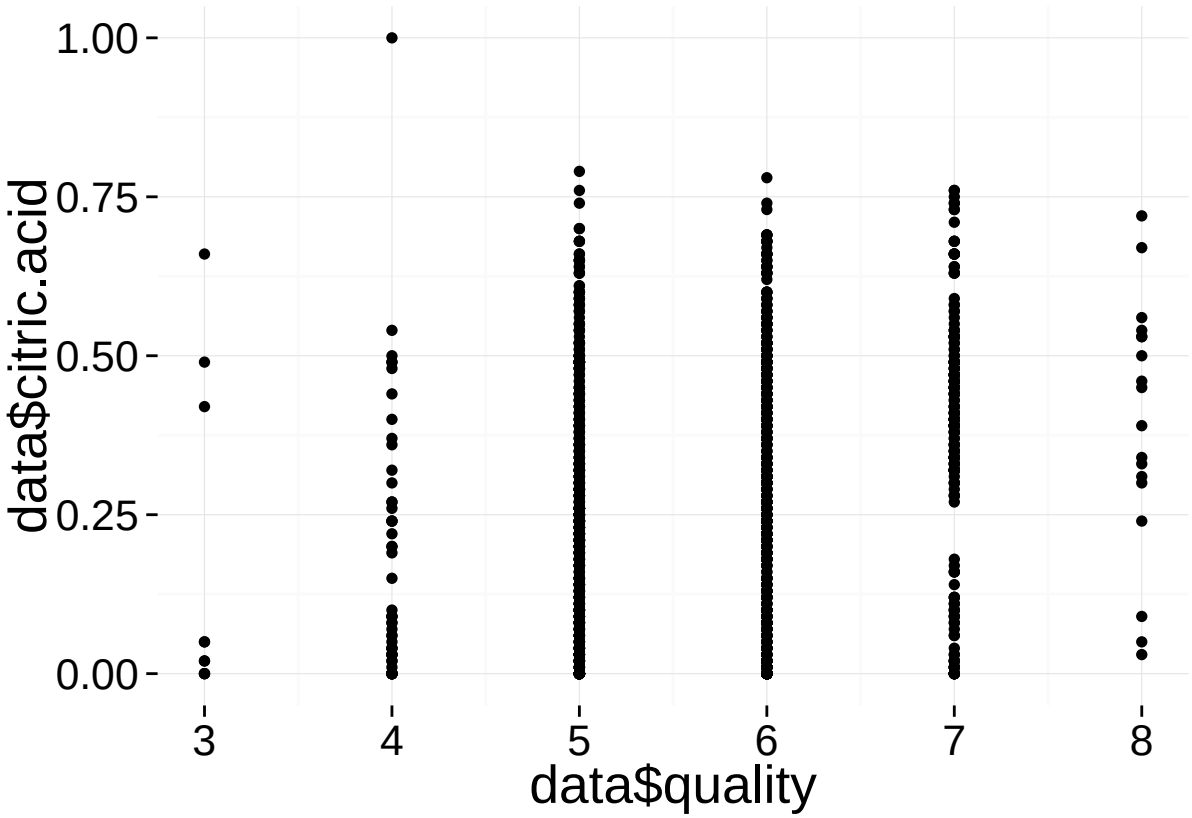


Clearly, the majority of tested wines obtained a quality of 5 or 6.

I want to look closer at scatter plots involving price and some other variables: Total acidity, sulphates and alcohol.

Relation between citric acid and quality:

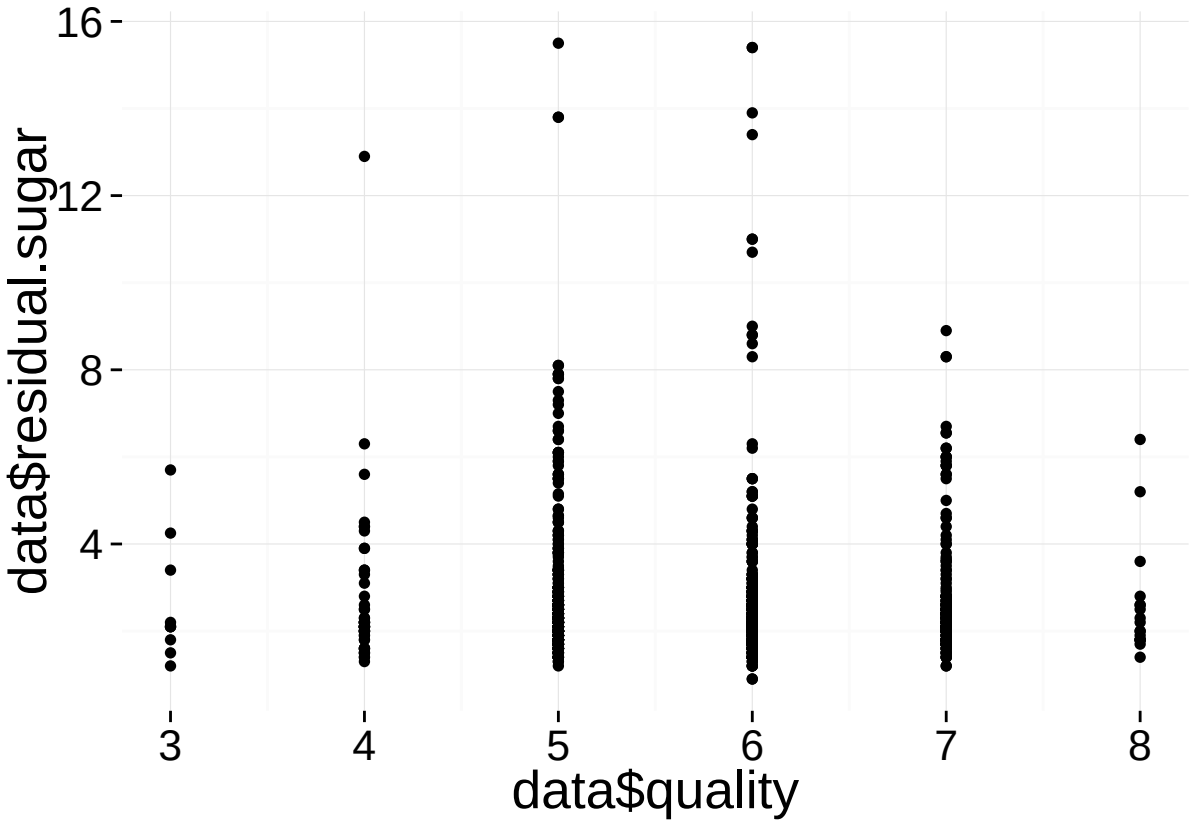
```
qplot(data$quality, data$citric.acid)
```



Wines ranged with 7 and 8 in quality tend to have less citric acid compared with less quality wines.

Relation between residual sugar and quality:

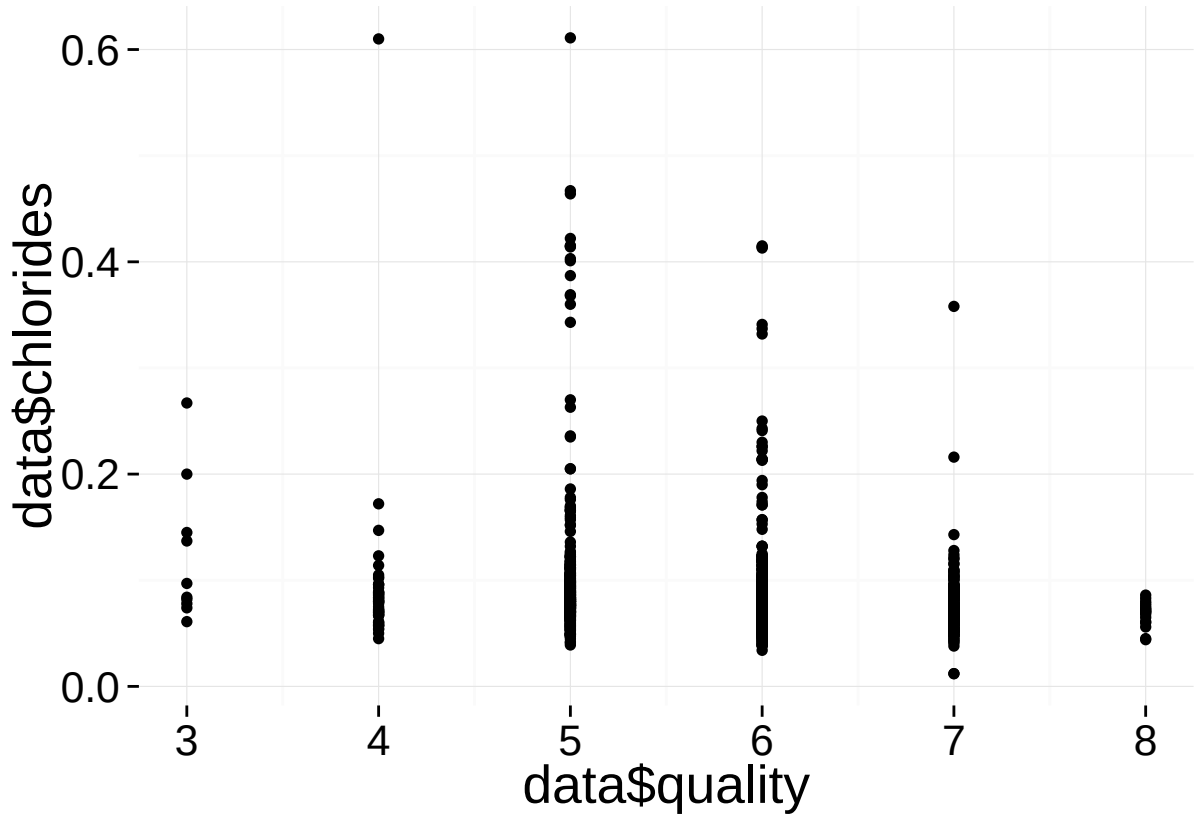
```
qplot(data$quality, data$residual.sugar)
```



Residual sugar levels tend to be low in high quality wines.

Relation between chlorides and quality:

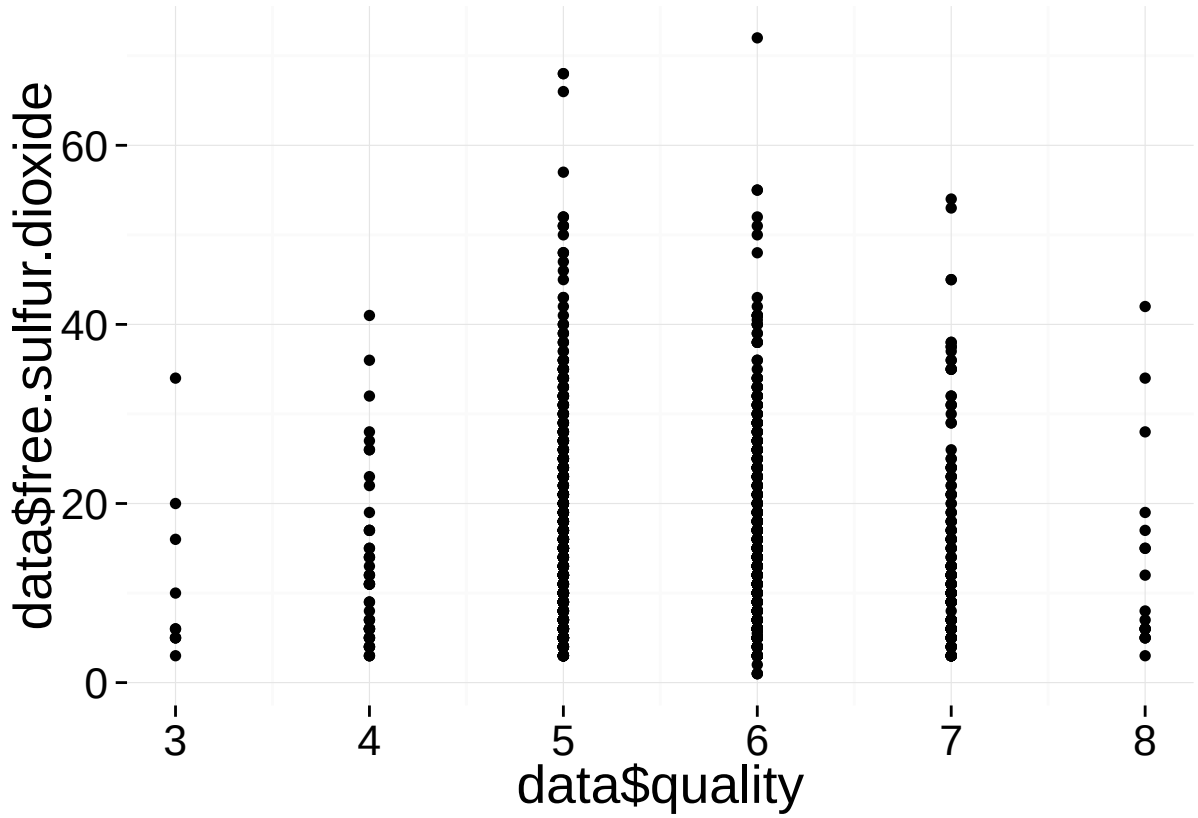
```
qplot(data$quality, data$chlorides)
```



Similarly as with residual sugar, chlorides levels tend to be low in high quality wines.

Relation between free sulfur dioxide and quality:

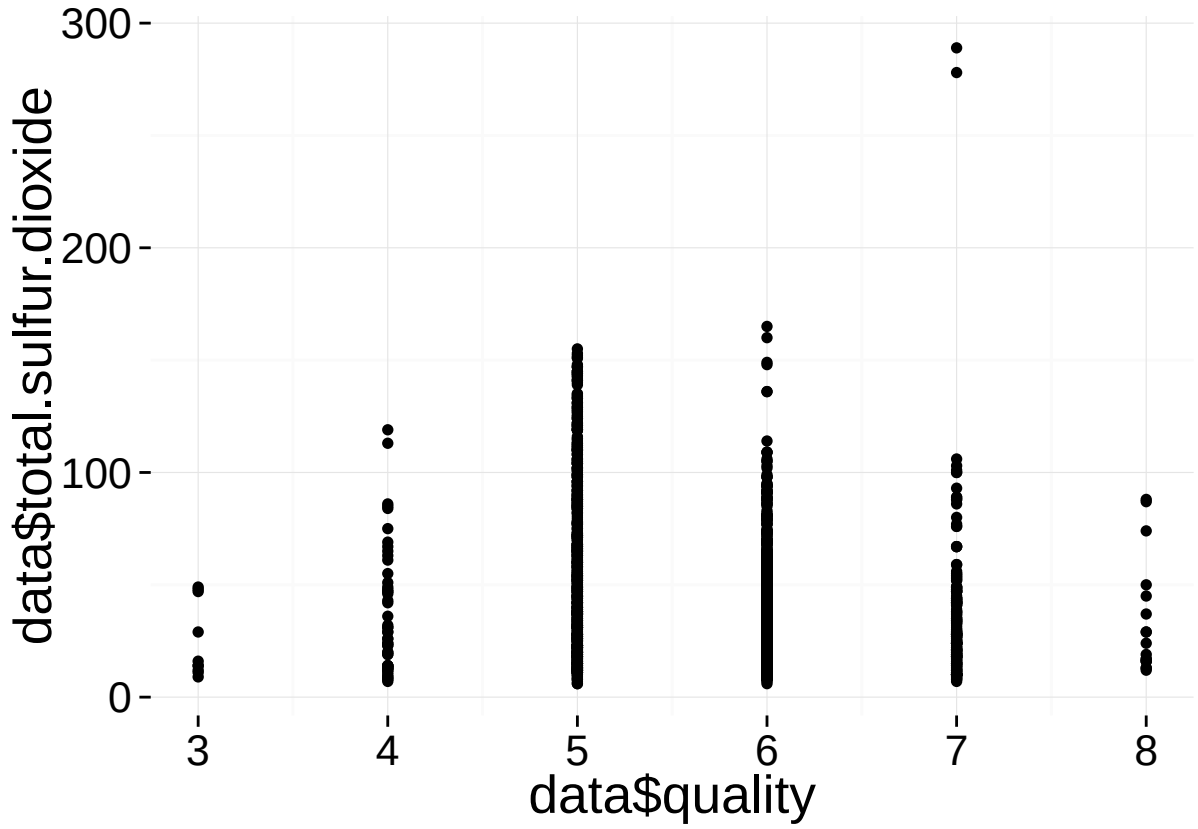
```
qplot(data$quality, data$free.sulfur.dioxide)
```



Medium quality red wines present high content of this kind of sulfur dioxide.

Relation between total sulfur dioxide and quality:

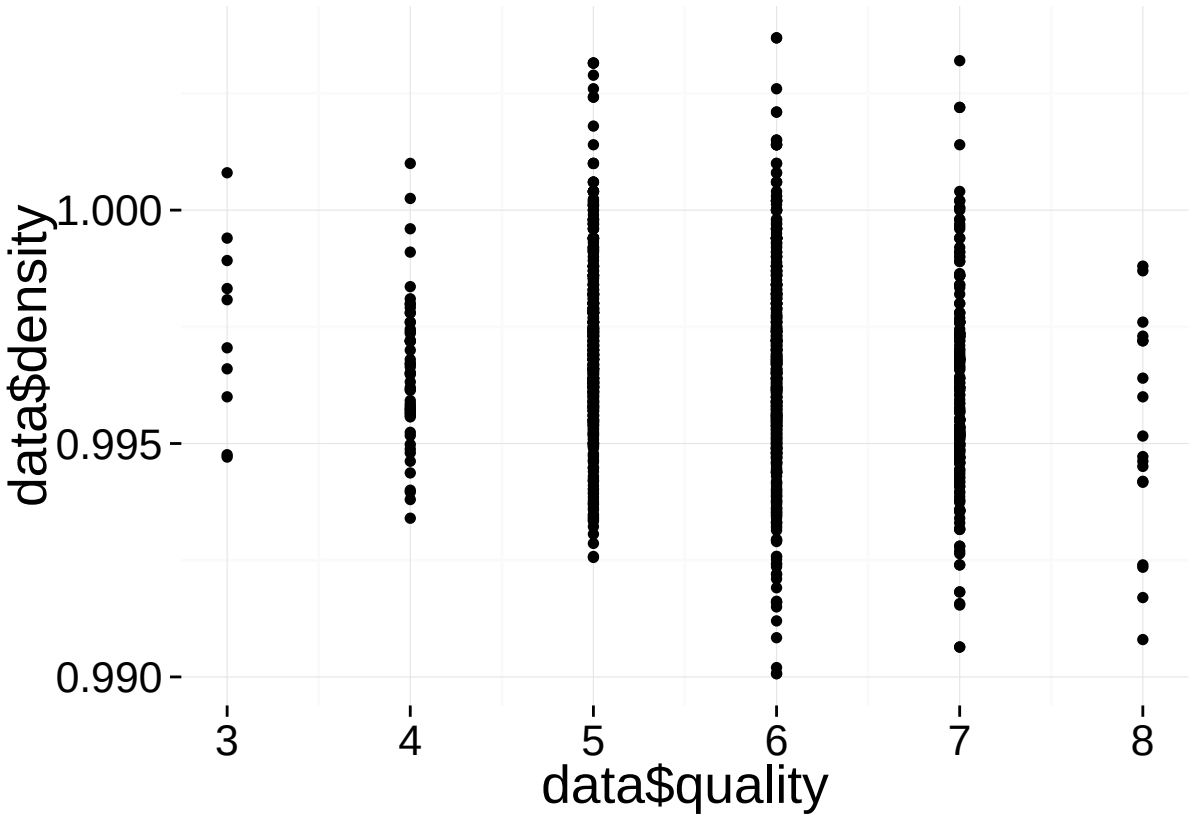
```
qplot(data$quality, data$total.sulfur.dioxide)
```



Obviously, total sulfur dioxide content is higher than the free one. In this case, it presents a normal distribution, reaching its peak at quality 5.

Relation between density and quality:

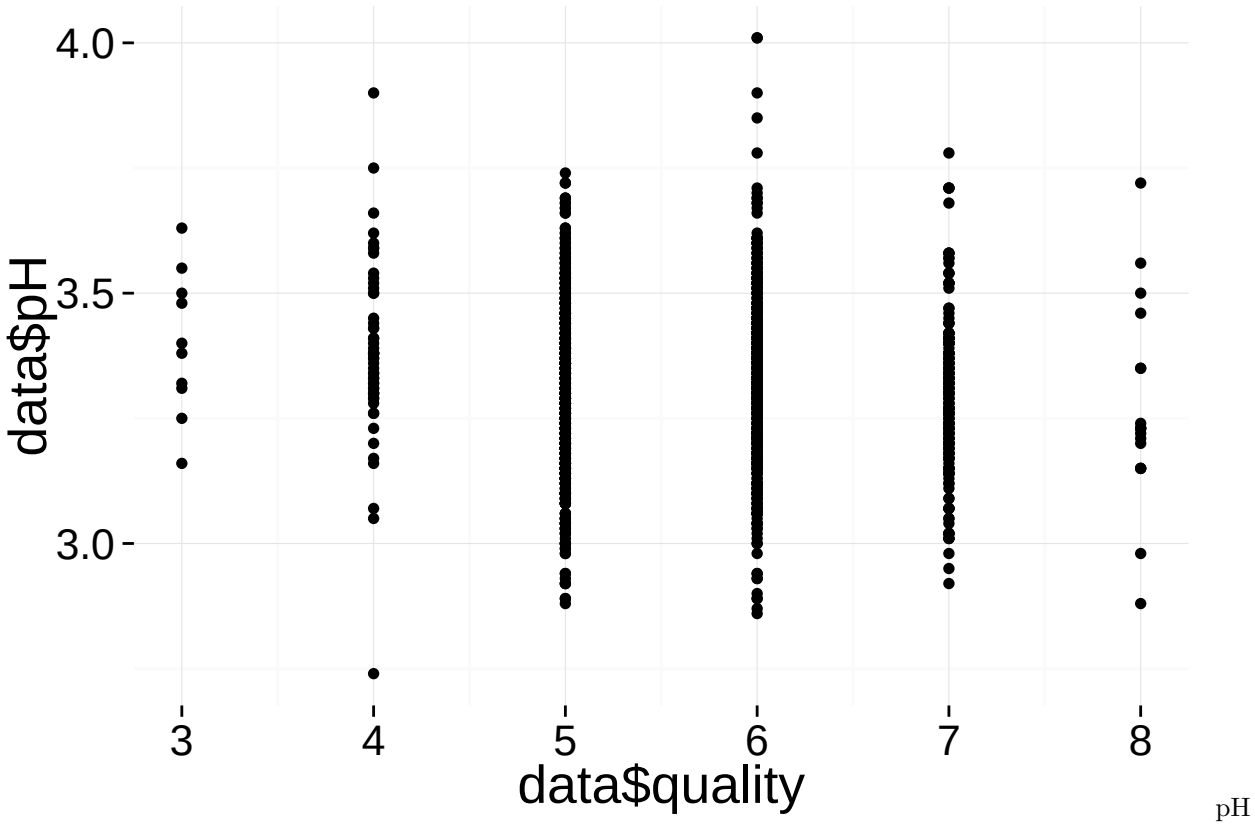
```
qplot(data$quality, data$density)
```



There are no important differences in density regarding quality.

Relation between pH and quality:

```
qplot(data$quality, data$pH)
```

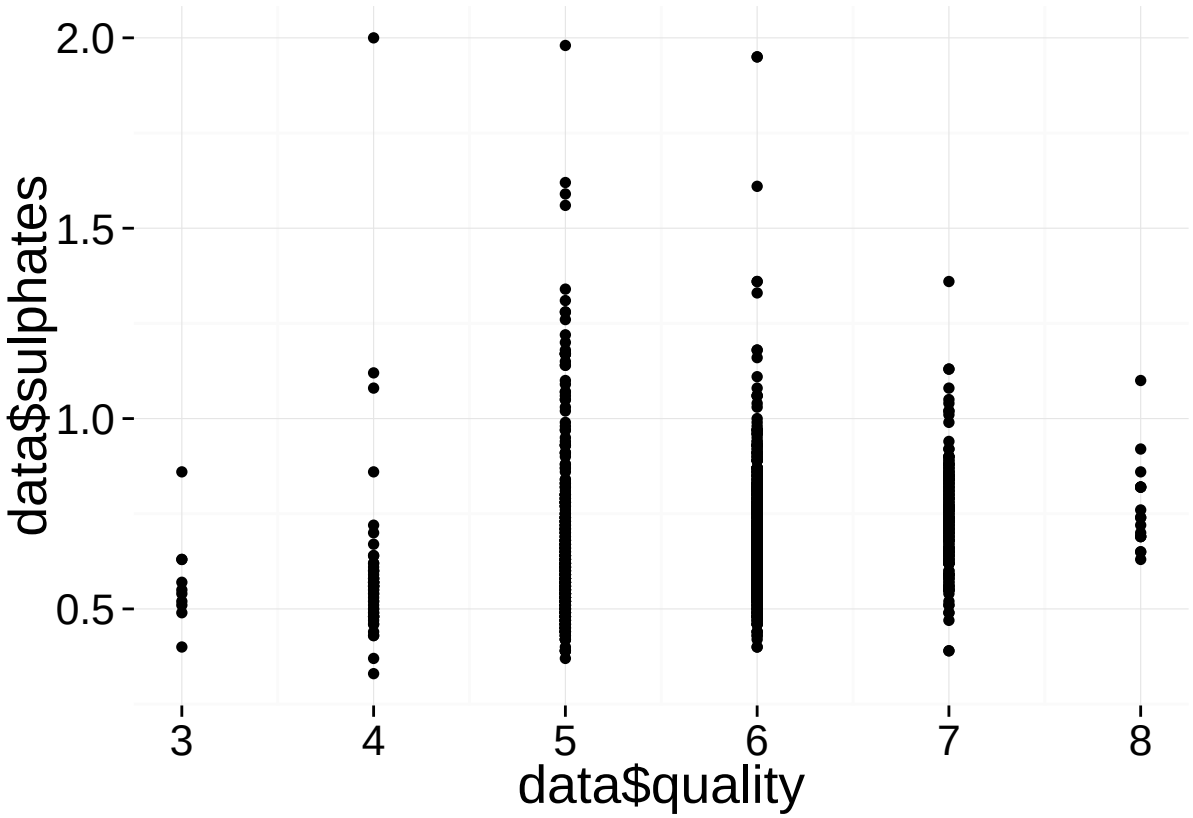


pH

variable does not present important differences between high and low quality red wines.

Relation between sulphates and quality:

```
qplot(data$quality, data$sulphates)
```



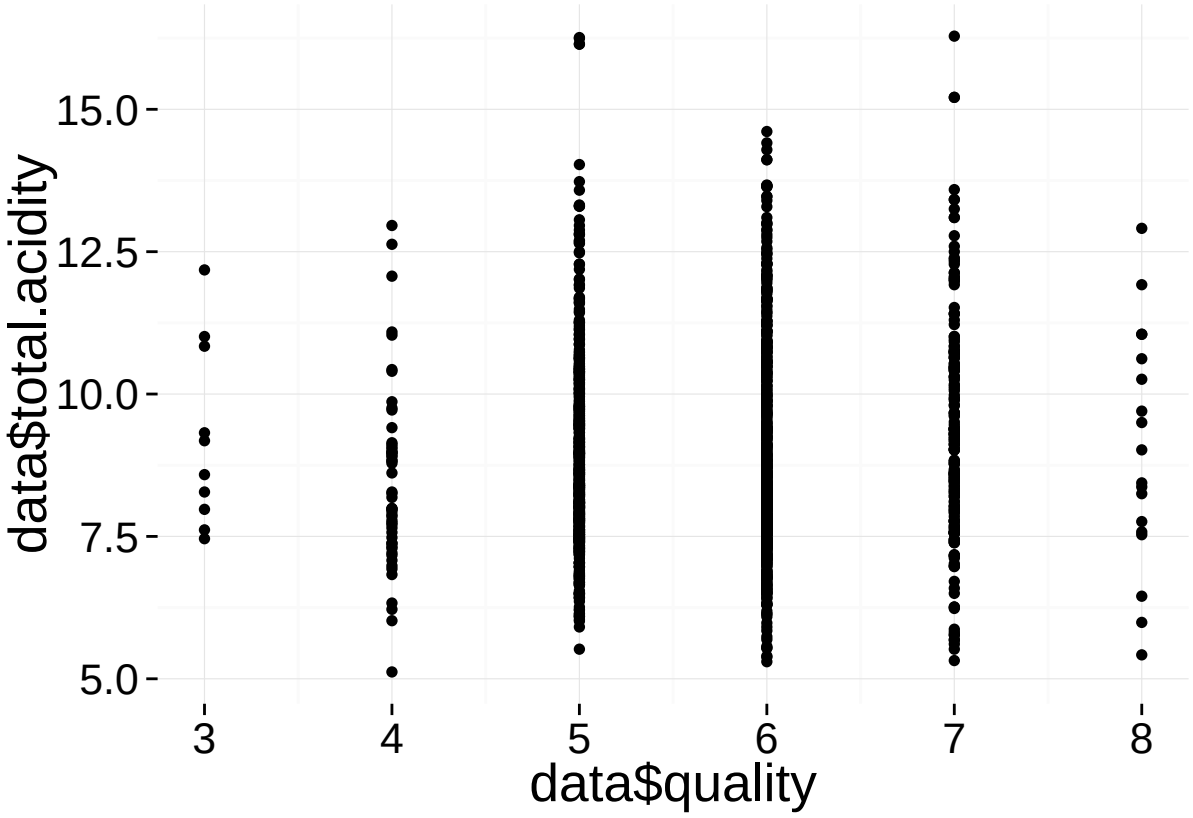

```
by(data$sulphates, data$quality,summary)
```

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

As sulphates content increases, quality also increses.

Relation between total acididty and quality:

```
qplot(data$quality, data$total.acidity)
```




```
by(data$total.acidity, data$quality,summary)
```

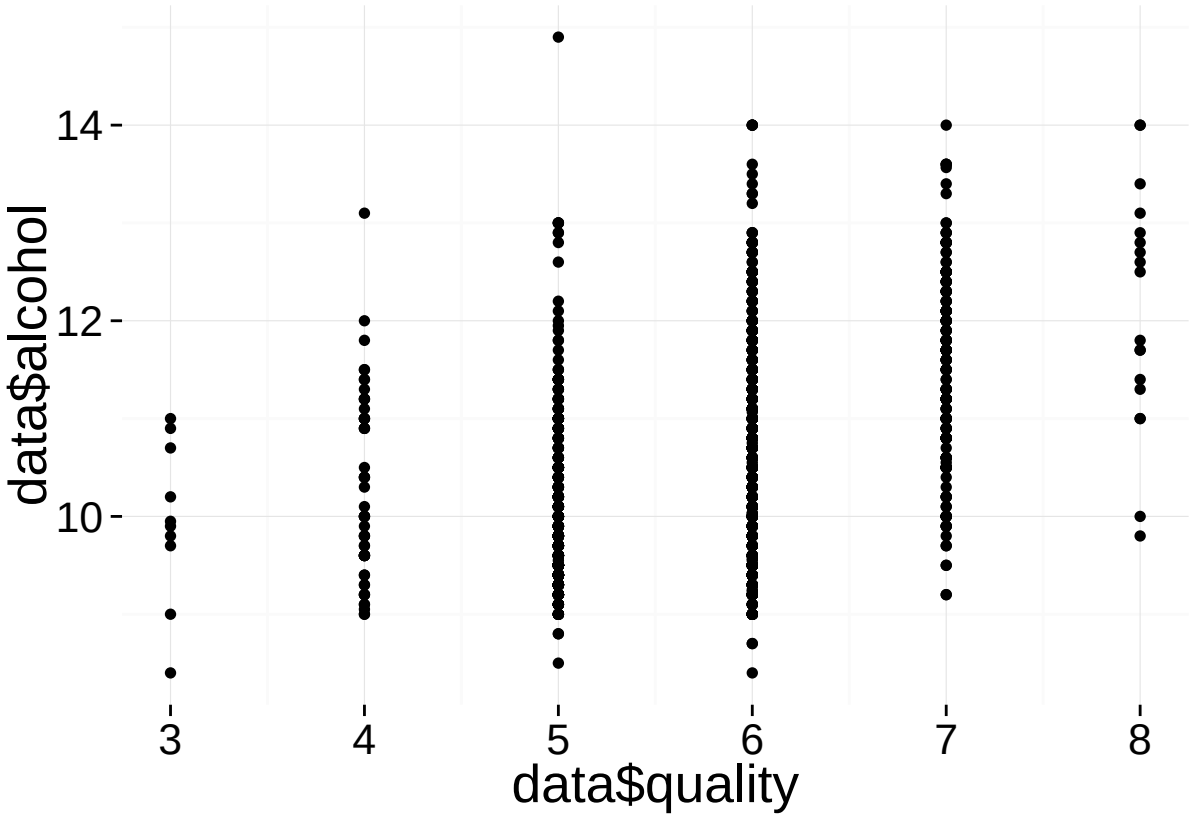
```
## data$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.460   8.051   8.882   9.244  10.460  12.180
## ------------------------------------------------------------
## data$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.120   7.380   8.185   8.473   9.070  12.960
## ------------------------------------------------------------
## data$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.520   7.735   8.390   8.744   9.490  16.260
## ------------------------------------------------------------
## data$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.300   7.605   8.400   8.845   9.881  14.610
## ------------------------------------------------------------
## data$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.320   7.880   9.110   9.276  10.480  16.280
## ------------------------------------------------------------
## data$quality: 8
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.420   7.625   8.730   8.990  10.530  12.910
```

Sulphates content seems to have little variation at least as it comes to the median of the red wines. There is no clear tendency here.

Relation between alcohol and quality:

```
qplot(data$quality, data$alcohol)
```




```
by(data$alcohol, data$quality,summary)
```

```
## data$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.580  11.000
## ------------------------------------------------------------
## data$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## ------------------------------------------------------------
## data$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## ------------------------------------------------------------
## data$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
```

```
## ------------------------------------------------------
## data$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## ------------------------------------------------------
## data$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

There seems to be an improvement in the quality of the wines as alcohol increases, with the better quality in mean equal to 12.09.

Correlation with the best wines:

```
cor(best)
```

```
##                              X fixed.acidity volatile.acidity
## X                  1.000000000   -0.44139589     -0.177963545
## fixed.acidity     -0.441395889    1.00000000     -0.265123947
## volatile.acidity  -0.177963545   -0.26512395      1.000000000
## citric.acid       -0.225954707    0.74527921     -0.494797992
## residual.sugar    -0.154809444    0.19540026      0.089458373
## chlorides         -0.209764325    0.21324228      0.072972680
## free.sulfur.dioxide  0.138761756   -0.15825919      0.017496787
## total.sulfur.dioxide  0.022426070   -0.18242933      0.045418879
## density           -0.505876725    0.78172195      0.008009062
## pH                 0.152452153   -0.77124197      0.342637656
## sulphates         -0.142024749    0.15584018     -0.208231630
## alcohol            0.197971946   -0.39169407      0.074566422
## quality           -0.003861185   -0.04225416      0.037021914
##                   citric.acid residual.sugar   chlorides
## X                 -0.225954707    -0.15480944 -0.20976432
## fixed.acidity      0.745279207     0.19540026  0.21324228
## volatile.acidity  -0.494797992     0.08945837  0.07297268
## citric.acid        1.000000000     0.27744936  0.25312687
## residual.sugar     0.277449363     1.00000000  0.12960269
## chlorides          0.253126870     0.12960269  1.00000000
## free.sulfur.dioxide  -0.070361099     0.01760084 -0.17964142
## total.sulfur.dioxide -0.001172564     0.25239139 -0.22901498
## density            0.516376490     0.34988921  0.34542338
## pH                -0.721071977    -0.18174275 -0.18024991
## sulphates          0.185814219    -0.12511994  0.12901983
## alcohol           -0.106003539     0.07175752 -0.21030843
## quality            0.022655985    -0.02896722 -0.07904467
##                   free.sulfur.dioxide total.sulfur.dioxide      density
## X                         0.138761756          0.022426070 -0.505876725
## fixed.acidity            -0.158259191         -0.182429327  0.781721948
## volatile.acidity          0.017496787          0.045418879  0.008009062
## citric.acid              -0.070361099         -0.001172564  0.516376490
## residual.sugar            0.017600840          0.252391387  0.349889213
## chlorides                -0.179641421         -0.229014982  0.345423384
## free.sulfur.dioxide       1.000000000          0.659703377 -0.104751635
## total.sulfur.dioxide      0.659703377          1.000000000 -0.182686191
## density                  -0.104751635         -0.182686191  1.000000000
```

```
## pH                          0.119722295          0.049723087 -0.449244670
## sulphates                   0.017190565         -0.045562296  0.208764923
## alcohol                     0.008409127          0.136670437 -0.584116886
## quality                    -0.020729255         -0.013372724 -0.112029687
##                                   pH   sulphates      alcohol      quality
## X                        0.15245215 -0.14202475  0.197971946 -0.003861185
## fixed.acidity           -0.77124197  0.15584018 -0.391694065 -0.042254161
## volatile.acidity         0.34263766 -0.20823163  0.074566422  0.037021914
## citric.acid             -0.72107198  0.18581422 -0.106003539  0.022655985
## residual.sugar          -0.18174275 -0.12511994  0.071757520 -0.028967225
## chlorides               -0.18024991  0.12901983 -0.210308431 -0.079044669
## free.sulfur.dioxide      0.11972229  0.01719057  0.008409127 -0.020729255
## total.sulfur.dioxide     0.04972309 -0.04556230  0.136670437 -0.013372724
## density                 -0.44924467  0.20876492 -0.584116886 -0.112029687
## pH                       1.00000000 -0.02700994  0.349997301 -0.042110521
## sulphates               -0.02700994  1.00000000 -0.052292982  0.054698507
## alcohol                  0.34999730 -0.05229298  1.000000000  0.174074808
## quality                 -0.04211052  0.05469851  0.174074808  1.000000000
```

Quality is most influenced by alcohol and density, but these two variables only explain about 30% of total quality in these red wines.

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?

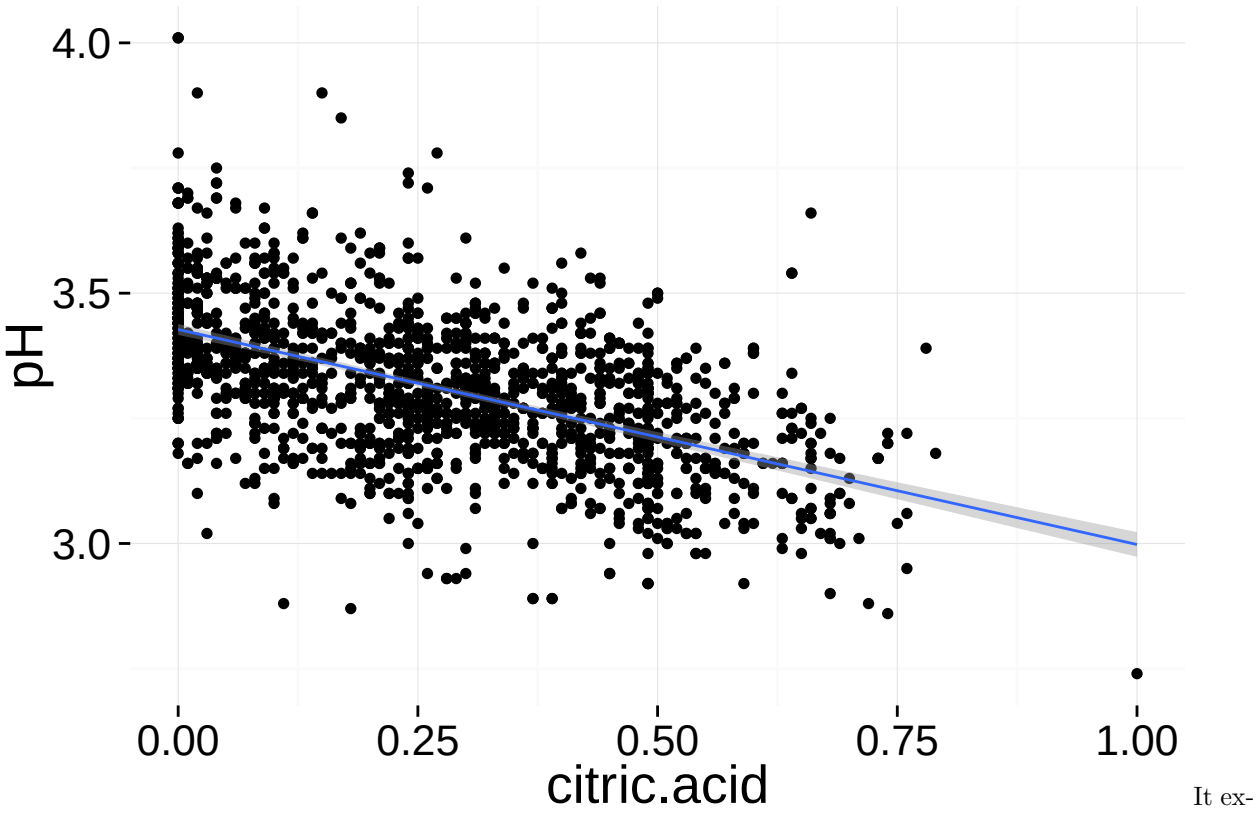
Quality correlates moderately with alcohol and sulphates. Its correlation with total acidity is lower, but higher than the ones with the other variables.

Total acidity and sulphates values are more concentrated in plots in comparison with alcohol content in red wines.

It is clear that the sensory evaluation is influenced by a myriad of variables in the best wines, whereas alcohol, sulphates and total acididty influence quality in the original data set, explaining more than 70% of total quality.

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

```
ggplot(aes( x = citric.acid, y = pH), data = data) + geom_point() + stat_smooth(method = "lm")
```

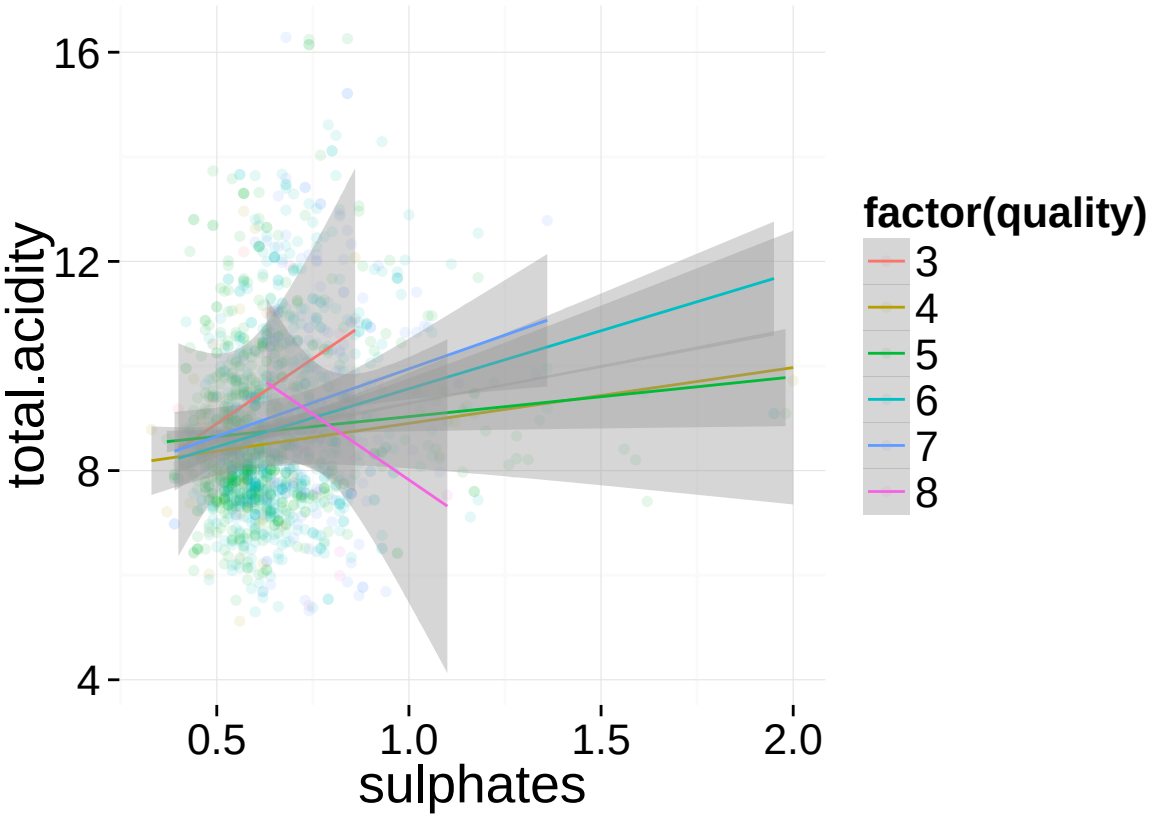


It explains that pH diminishes as citric acid increases.

What was the strongest relationship you found?

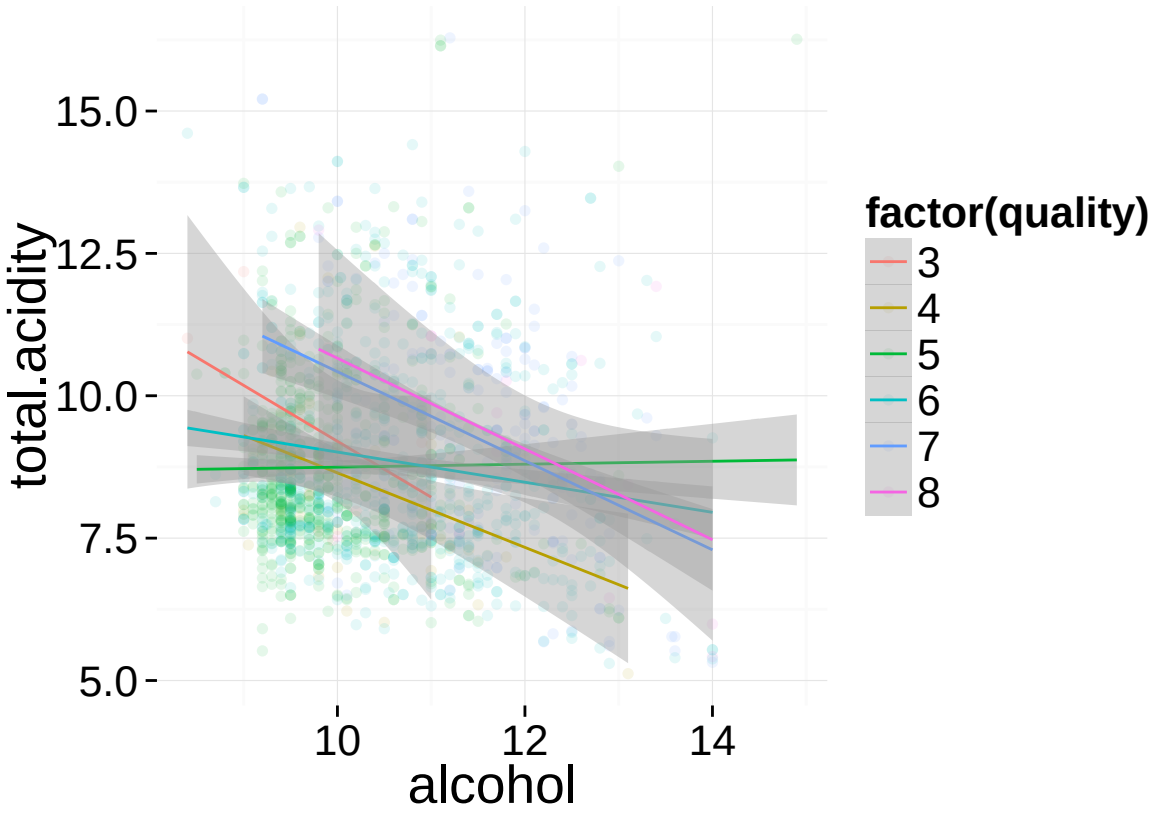
Alcohol correlates with the originald ata set by more than 47%, and 17% with the best red wines. Importantly, sulphures and acidity also contribute to explain the quality variable. Although, this last variable should be taken with caution since it is a subjective one.

Multivariate Plots Section



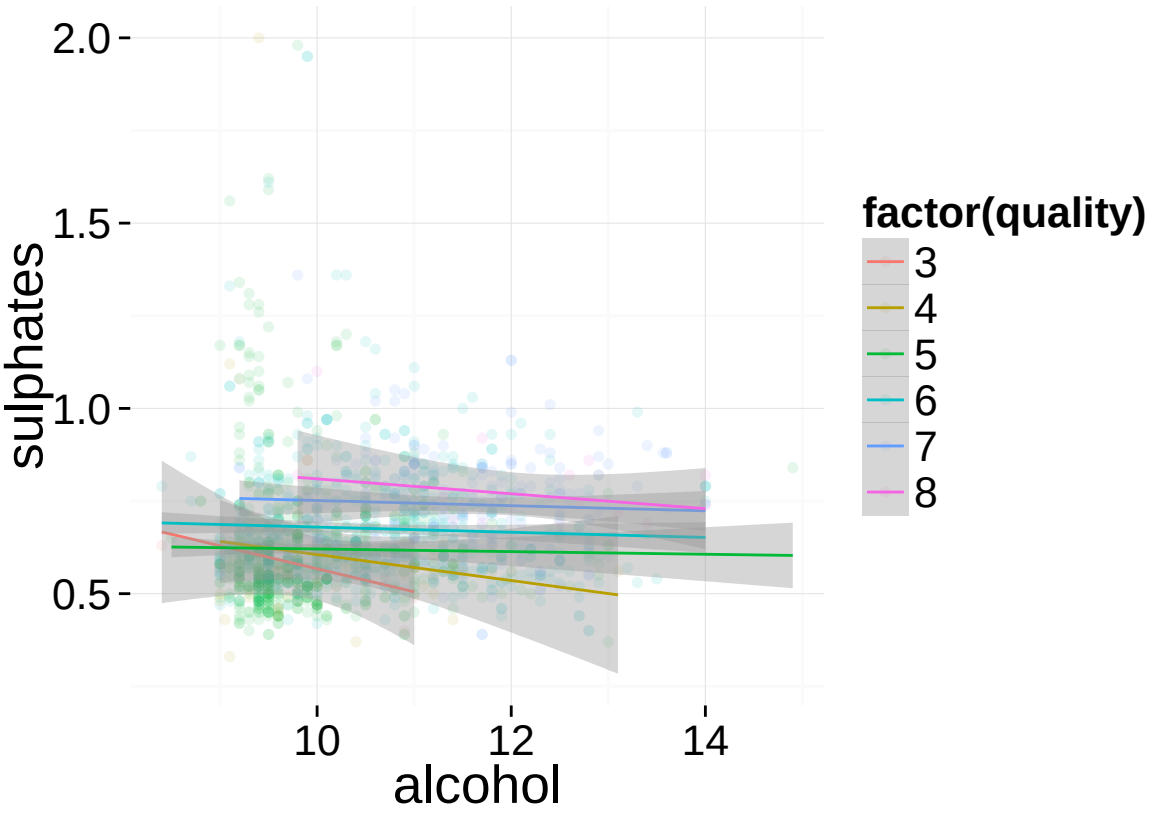
This plot shows the relation between total acidity, sulphates and quality. The positive trends in qualities 3 to 7, indicate that the more sulphates, the more total acidity. This is true except for the quality 8, in which the relation is negative. This means that top quality red wines tend to present small figures in sulphates and total acidity.

```
ggplot(data, aes(alcohol, total.acidity, color = factor(quality)))+
     geom_point(alpha = 1/10) +
     stat_smooth(method = "lm")
```



The relation between alcohol and total.acidity is negative in most cases as it comes to quality. Only in quality equal to 5, it is slightly positive. The best quality wines present an inverse relation between alcohol and total acidity. In other words, the less alcohol, the more total acidity.

```
ggplot(data, aes(alcohol, sulphates, color = factor(quality)))+
    geom_point(alpha = 1/10) +
    stat_smooth(method = "lm")
```

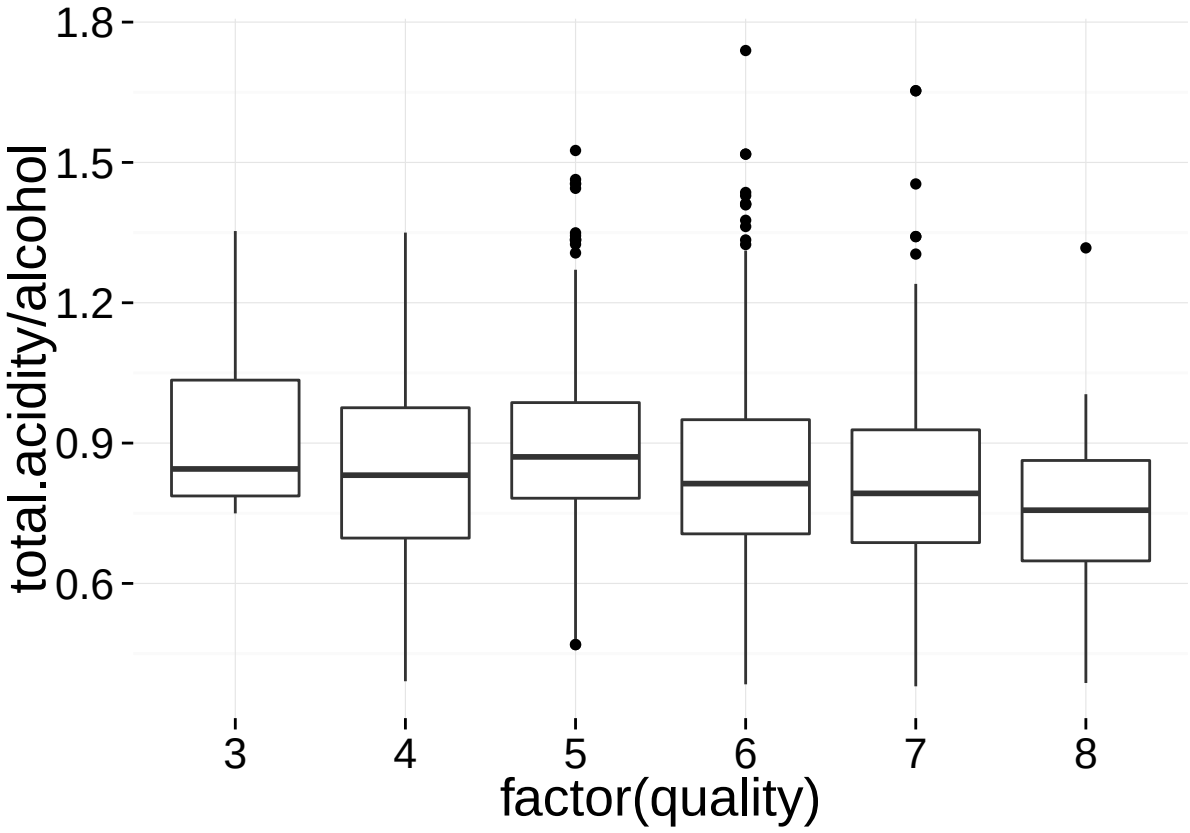


Here, an increment in alcohol content does not seem to produce a statistical increase in sulphates. In fact, in most of the quality values an increment in alcohol produces a reduction in the quality.

I want to observe relations between ttwo variables and their influence on quality:

Total acidity/alcohol vs. quality:

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



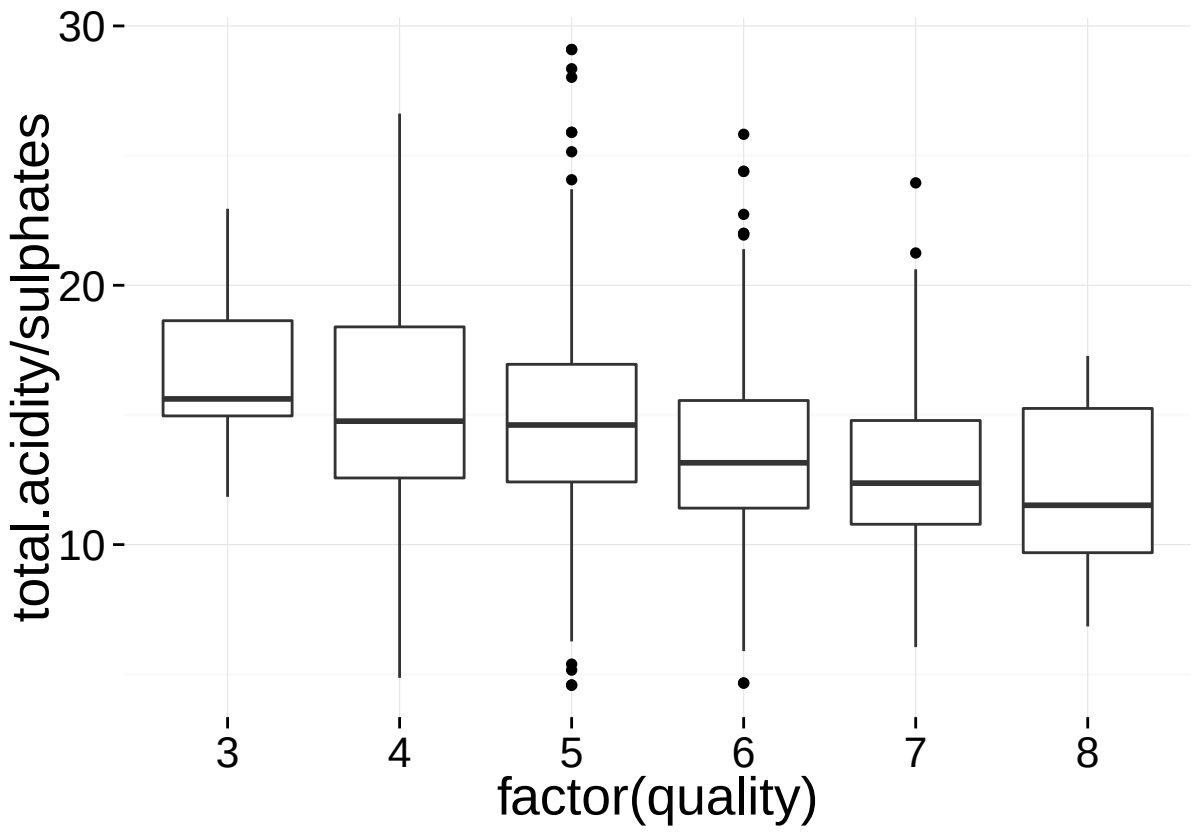

```
data$r1 <- data$total.acidity / data$alcohol
by(data$r1, data$quality, summary)
```

```
## data$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7497  0.7870  0.8447  0.9413  1.0350  1.3530
## ------------------------------------------------------------
## data$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3908  0.6969  0.8314  0.8368  0.9755  1.3500
## ------------------------------------------------------------
## data$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4692  0.7821  0.8705  0.8876  0.9866  1.5260
## ------------------------------------------------------------
## data$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3843  0.7060  0.8134  0.8424  0.9500  1.7390
## ------------------------------------------------------------
## data$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3800  0.6873  0.7925  0.8203  0.9284  1.6530
## ------------------------------------------------------------
## data$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3871  0.6482  0.7565  0.7586  0.8628  1.3170
```

Interestingly, the less median, the better quality.

Total acidity/sulphates vs. quality:

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




```
data$r2 <- data$total.acidity / data$sulphates
by(data$r2, data$quality, summary)
```

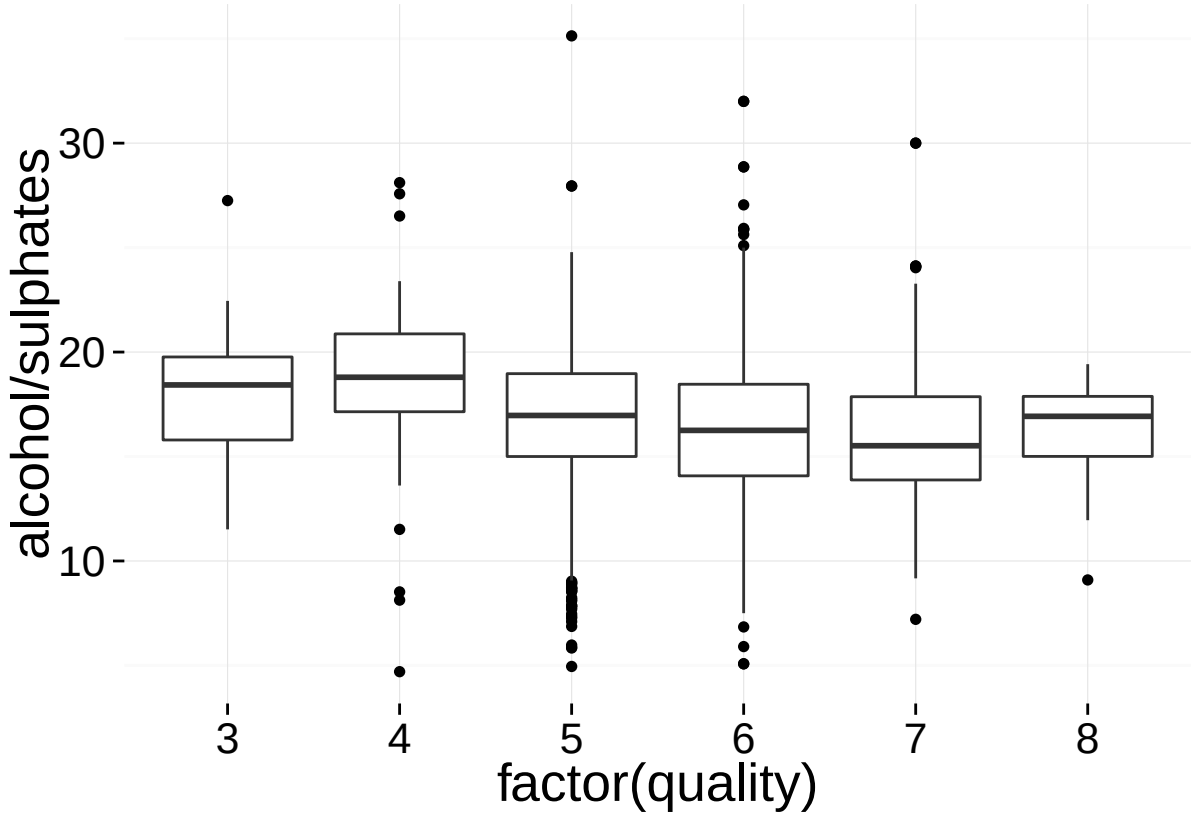
```
## data$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.84   14.96   15.62   16.65   18.63   22.95
## ------------------------------------------------------------
## data$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.86   12.57   14.76   15.24   18.40   26.62
## ------------------------------------------------------------
## data$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.574  12.420  14.610  14.770  16.950  29.090
## ------------------------------------------------------------
## data$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.662  11.410  13.150  13.540  15.560  25.820
## ------------------------------------------------------------
## data$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   6.048  10.780  12.370  12.840  14.790  23.950
## ------------------------------------------------------------
## data$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.845   9.688  11.520  12.040  15.250  17.280
```

Similarly as before, the less median, the better red wine quality.

Alcohol/sulphates vs. quality:

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




```
data$r3 <- data$alcohol / data$sulphates
by(data$r3, data$quality,summary)
```

```
## data$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.51   15.79   18.43   18.24   19.76   27.25
## ------------------------------------------------------------
## data$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.70   17.14   18.79   18.57   20.87   28.11
## ------------------------------------------------------------
## data$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.949  15.000  16.960  16.780  18.970  35.140
## ------------------------------------------------------------
```

```
## data$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.077  14.070  16.250  16.420  18.460  32.000
## ------------------------------------------------------------
## data$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.206  13.880  15.510  16.010  17.860  30.000
## ------------------------------------------------------------
## data$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   9.091  15.000  16.920  16.070  17.880  19.420
```

In this case, there is a slight relation in quality for the ratio alcohol/sulphates

It seems a linear model can be constructed in order to predict the red wine quality based on its alcohol, sulphates and total acidity.

```
linMod <- lm(data$quality ~ data$total.acidity + data$alcohol + data$sulphates)
summary(linMod)
```

```
##
## Call:
## lm(formula = data$quality ~ data$total.acidity + data$alcohol +
##     data$sulphates)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.72564 -0.36283 -0.08247  0.50432  2.25397
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.95530    0.20110   4.750 2.21e-06 ***
## data$total.acidity   0.04466    0.01026   4.351 1.44e-05 ***
## data$alcohol         0.35319    0.01627  21.705  < 2e-16 ***
## data$sulphates       0.91826    0.10326   8.893  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6866 on 1595 degrees of freedom
## Multiple R-squared:  0.2785, Adjusted R-squared:  0.2771
## F-statistic: 205.2 on 3 and 1595 DF,  p-value: < 2.2e-16
```

So, the linear model is: quality = 0.95530 + 0.04466 x Total.acidity + 0.35319 x Alcohol + 0.91826 x Sulphates. The model is not precisse at all with a R squared of 0.2771.

The model can be improved by adding squared and cubed relations:

```
linMod2 <- lm(data$quality ~ data$total.acidity + I(data$total.acidity^2) + data$alcohol + I(data$alcoh
summary(linMod2)
```

```
##
## Call:
## lm(formula = data$quality ~ data$total.acidity + I(data$total.acidity^2) +
```

```
##     data$alcohol + I(data$alcohol^2) + I(data$alcohol^3) + data$sulphates +
##     I(data$sulphates^3))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.82131 -0.36945 -0.06585  0.48603  2.26308
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              37.256713  11.153016   3.341 0.000856 ***
## data$total.acidity        0.207058   0.079437   2.607 0.009231 **
## I(data$total.acidity^2)  -0.008876   0.004055  -2.189 0.028745 *
## data$alcohol            -10.007608   3.015850  -3.318 0.000926 ***
## I(data$alcohol^2)         0.941450   0.271491   3.468 0.000539 ***
## I(data$alcohol^3)        -0.028314   0.008094  -3.498 0.000481 ***
## data$sulphates            2.302177   0.201217  11.441  < 2e-16 ***
## I(data$sulphates^3)      -0.534078   0.067369  -7.928 4.17e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6701 on 1591 degrees of freedom
## Multiple R-squared:  0.3145, Adjusted R-squared:  0.3114
## F-statistic: 104.3 on 7 and 1591 DF,  p-value: < 2.2e-16
```

But stil the R squared is only 0.3114, with the intercept, alcohol and sulphates terms are significant to 0.001.

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?

As it comes tor atios, the two ones in which total acidity was involved, shoed an improvement in wine quality as the ratio reduces. This fact is not shown in the ratio alcohol/sulphates.

Were there any interesting or surprising interactions between features?

Developing the total acidity variable, one gets to the conclusion that fixed volatility contributes more to the final output than the volatile one.

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

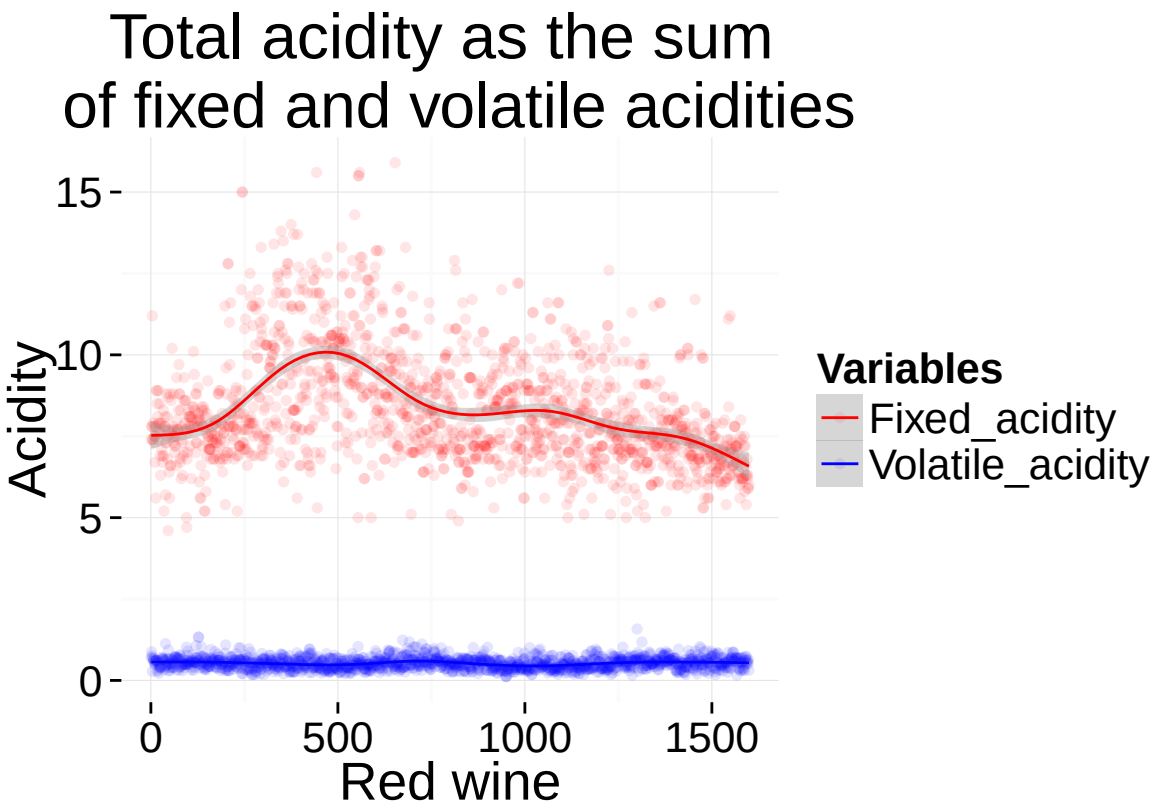
I created the model of total acidity to capture the influence of the sum of the fixed and volatile acidities. It has certain influence on the model, contributing to explain it.

The three plots should show different trends and should be polished with appropriate labels, units, and titles

Final Plots and Summary

Plot One

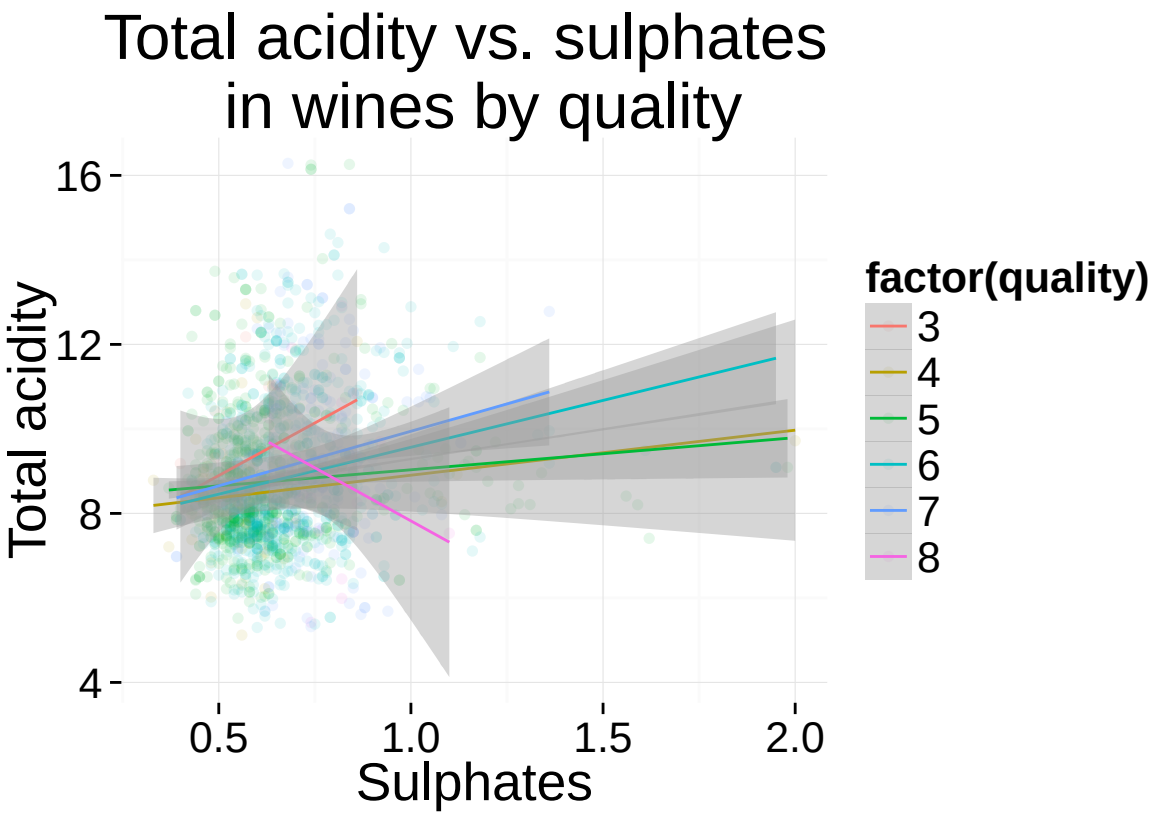
```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam with formula: y ~ s(x, b
```



Description One

This plot shows the influence of each acidity on the total acidity variable. The contribution of fixed acidity is far more important than the one in volatile acidity in every wine.

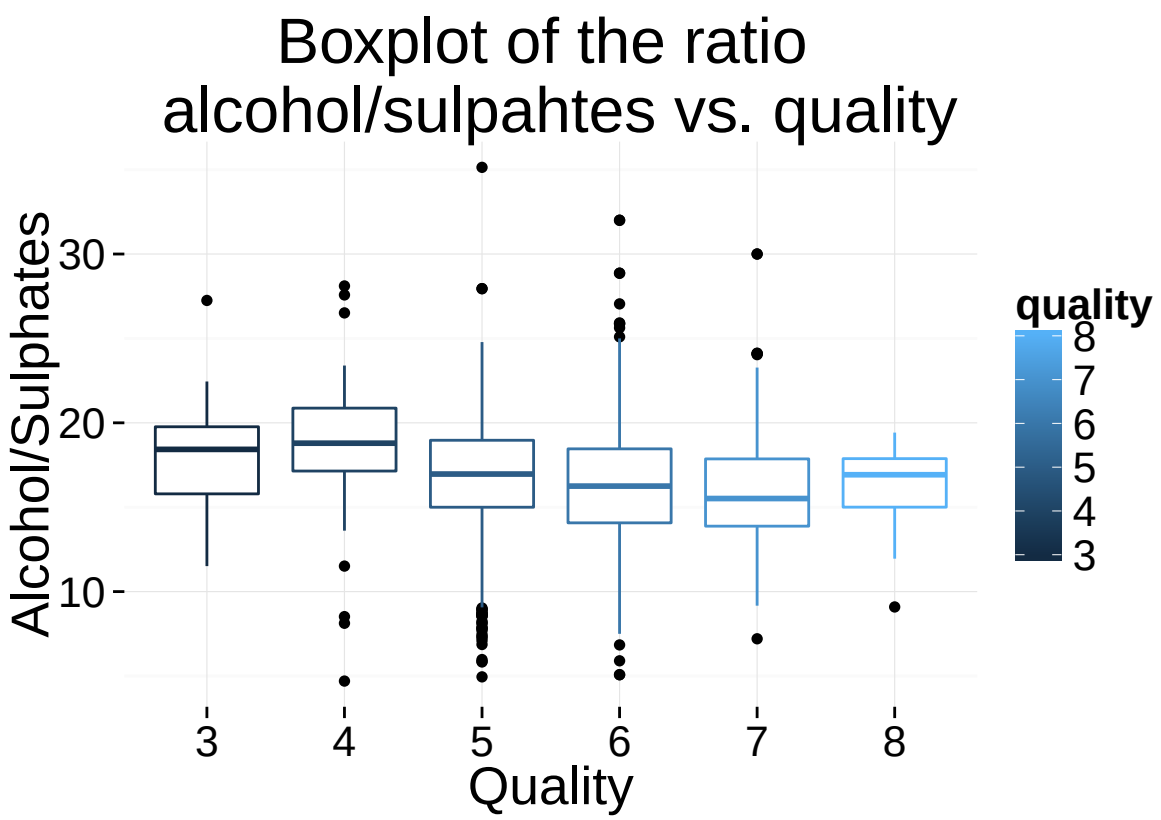
Plot Two



Description Two

Contrary to what can be thought, the positive relation between sulphates and total acidity is negative for wines with quality 8. Conversely, the relation is positive for every other wine. An important fact is thatpoor quality wines (3) show the highest median relation similalry as with the first values in wines with quality 8.

Plot Three



Description Three

The ratio alcohol /sulphates tend to decrease as wine quality increases with the exception of wines with quality 3. Even in the case of the best red wines this is true, since the box is smaller and thus the values are more concentrated.

Reflection

Quality in wines is a confused term. In this data set, it is provided by the opinion of three experts. So, it is a subjective fuature. I have tried to develop a model to explain it in terms of the three most important variables: Total acidity (as the sum of fixed acidity plus volatile acidity), alcochol and sulphates. This were the most influential variables to quality as explained with plots and figures. Nevertheless, their contribution is roughly 30% to the model, even working out polinomial models. This is a clear example on how complex is wine quality and the influence on multiple features. I would like to have a key paramenter in wine quality: tannin content. This, plus alcohol and acidity are the key triad in wines, not to mention additional parameters such as flavor or sabour. Again, the quality in wines is a complex subject. A more recent data would be better to make predictions of wines quality, and comparisons might be made between the other linear models to see if other variables may account for wines quality.