White Wine Quality

by Michael Ochs

==========

Citation

This dataset is public available for research. The details are described in [Cortez et al., 2009]. Please include this citation if you plan to use this database: 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. Available at: - [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 (http://dx.doi.org/10.1016/j.dss.2009.05.016) - [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf (http://www3.dsi.uminho.pt/pcortez/winequality09.pdf) - [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib (http://www3.dsi.uminho.pt/pcortez/dss09.bib)

Sources: - Created by: Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009

About the Data

In the above reference, two datasets were created, using red and white wine samples. 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 qualitybetween 0 (very bad) and 10 (very excellent). Several data mining methods were applied to model these datasets under a regression approach. The support vector machine model achieved the best results. Several metrics were computed: MAD, confusion matrix for a fixed error tolerance (T), etc. Also, we plot the relative importances of the input variables (as measured by a sensitivity analysis procedure).

This analysis focuses on white wine quality and the characteristics as well as chemical compositions that affect it. There are 4898 data points for white wine with 11+ output attributes

Attribute Information

For more information, read [Cortez et al., 2009].

Input variables (based on physicochemical tests):

1 - fixed acidity (tartaric acid - g / dm^3)

2 - volatile acidity (acetic acid - g / dm^3)

3 - citric acid (g / dm^3)

4 - residual sugar (g / dm^3)

5 - chlorides (sodium chloride - g / dm^3)

6 - free sulfur dioxide (mg / dm^3)

7 - total sulfur dioxide (mg / dm^3)

8 - density (g / cm^3)

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)

Description of attributes:

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3 - citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines

4 - residual sugar: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

5 - chlorides: the amount of salt in the wine

6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7 - total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content

9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

11 - alcohol: the percent alcohol content of the wine

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.0             0.27        0.36           20.7     0.045
## 2 2           6.3             0.30        0.34            1.6     0.049
## 3 3           8.1             0.28        0.40            6.9     0.050
## 4 4           7.2             0.23        0.32            8.5     0.058
## 5 5           7.2             0.23        0.32            8.5     0.058
## 6 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

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

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

Univariate Analysis

Structure Dataset

Dataset is structured by measurements of various chemicals and chemical attributes in white wine.

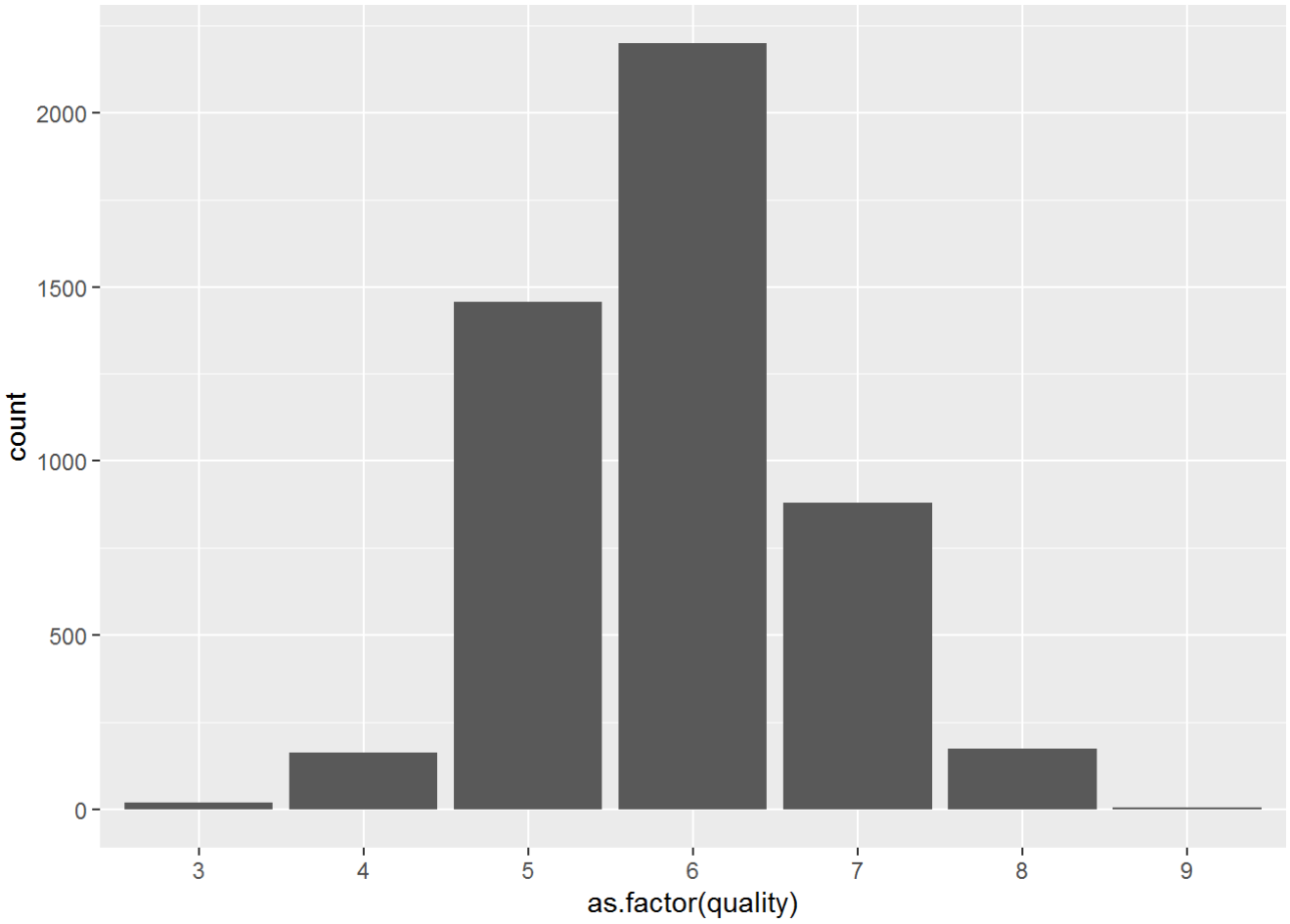
New/Removed Variables

No new variables were created. The ‘X’ data column was deleted as it had no value. It was an iteration of the wine tested.

Main Feature of Interest in Dataset

The main feature of interest is the quality and determining what improves the quality rating of the consumer.

Variable: quality




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000
```

Observation

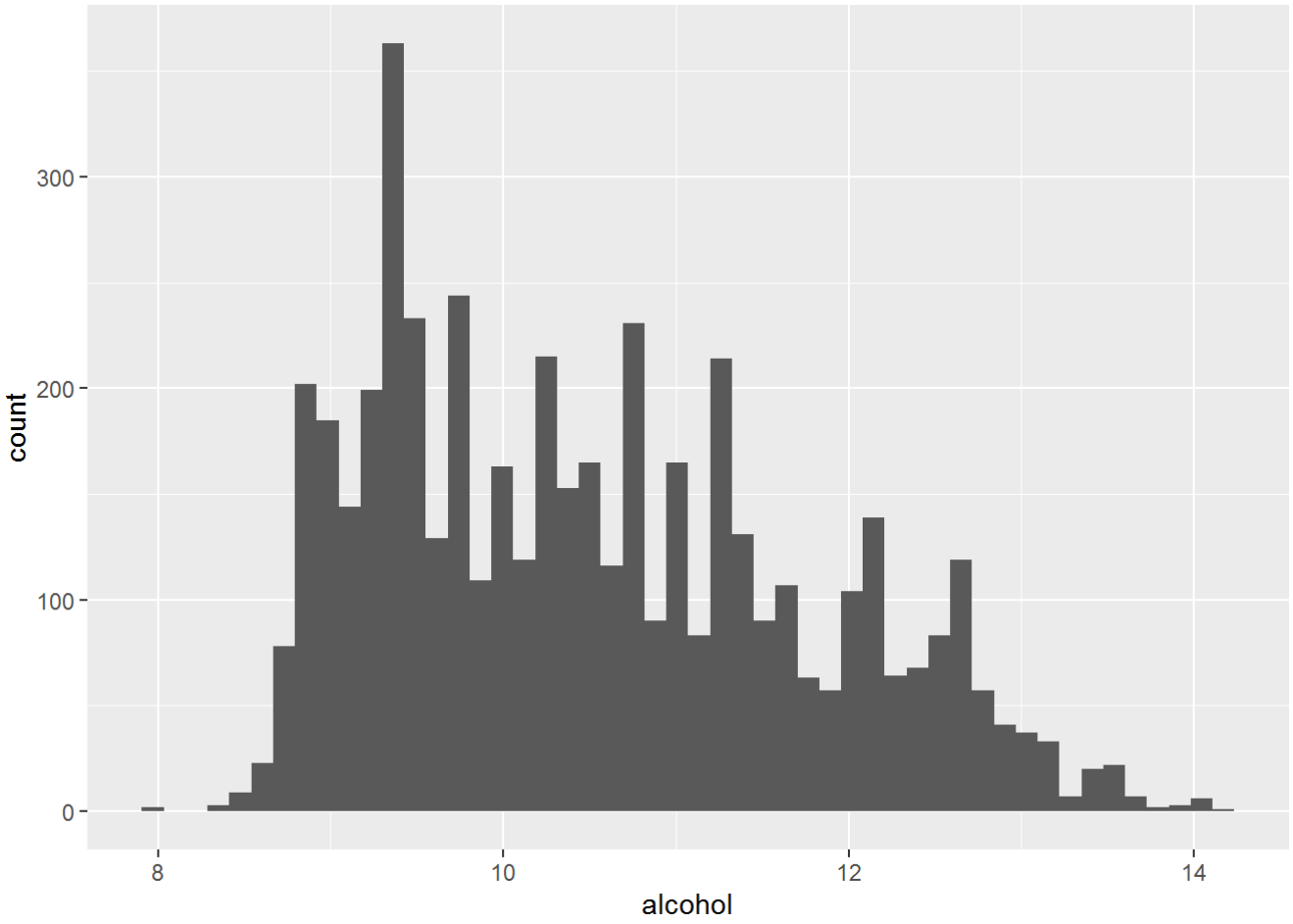
Normal Distribution: [X] Yes, [ ] No

The ‘quality’ histogram showed a fairly tight normal distribution with 80% of the ratings a 5 or 6 out of 10. The x-scale was adjusted to show the quality ratings on a 1 to 10 scale by 1 to further illustrate the narrow range.

Other Features of Interest

The other features of interest are listed below. Histograms and summaries of these variables are included. These variables will be cross examplined for further analysis in follow sections. Differenet binwdiths and scales are used for a cleaner look.

Variable: alcohol




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

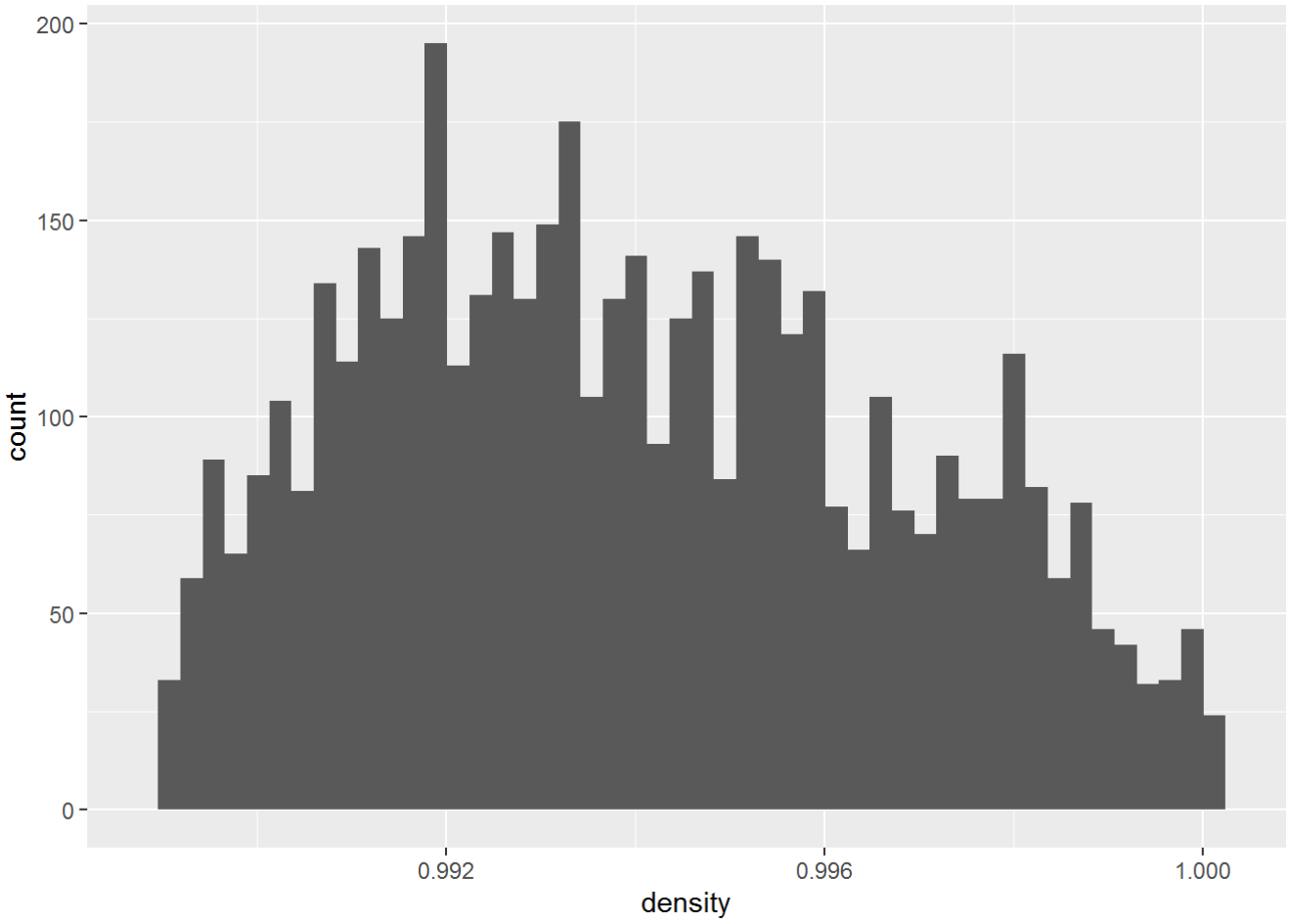
Observation

Normal Distribution: [ ] Yes, [X] No

Distribution: Log Normal, Skewed left

The sqrt scale on the y-axis appeared to be the best fit for the ‘alcohol’ histogram. After trying a few different scales, it made the histogram closer to a normal distribution.

Variable: density




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

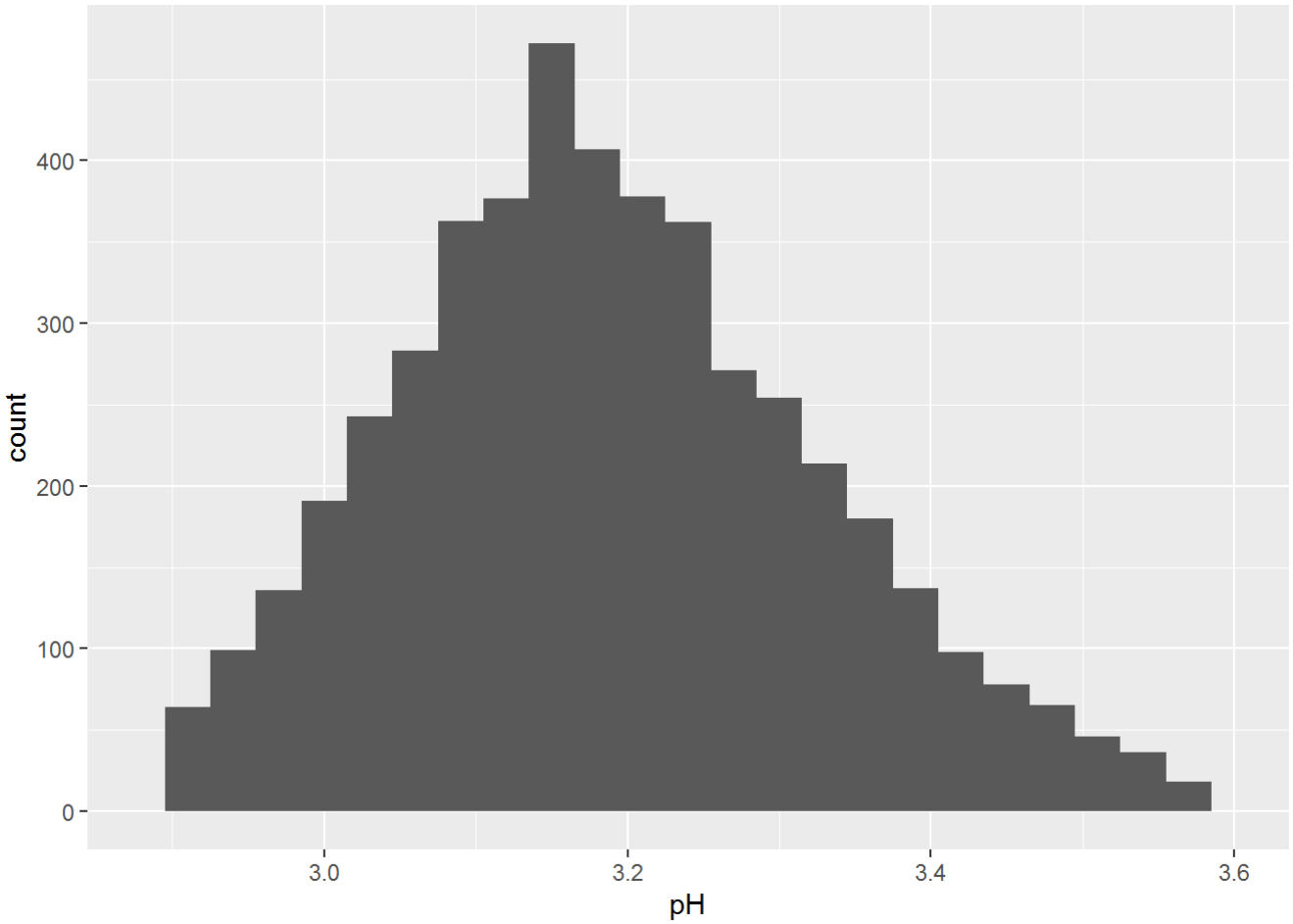
Observation

Normal Distribution: [X] Yes, [ ] No

Distribution: Bimodal

Density appeared to have a very tight range. The 10th percentile to 90th percentile was roughly .992 to .996 respectively. The percentage different between the lowest value and highest value is (1.0390-0.9871)/0.9871 = 5.26%

Variable: pH



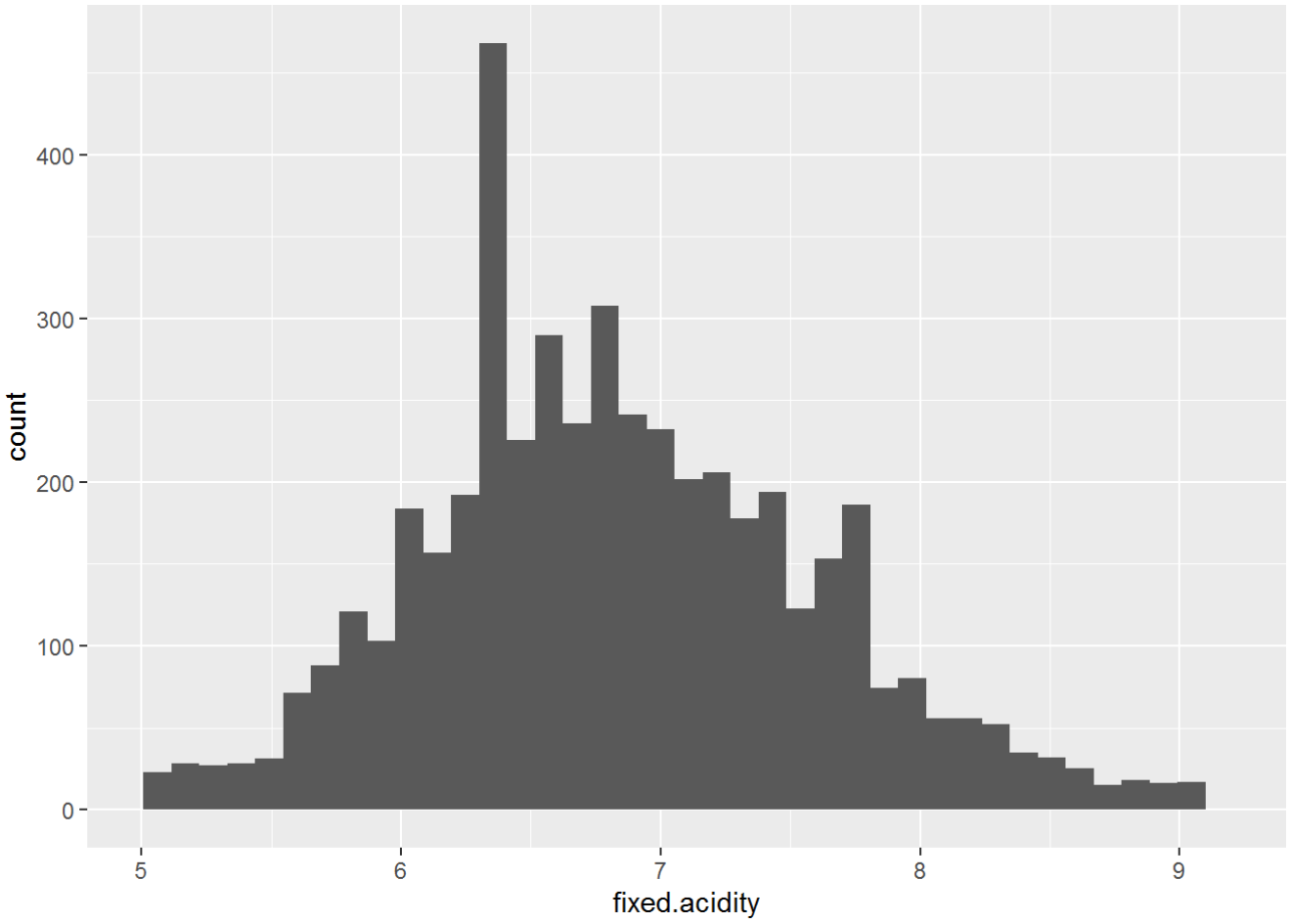

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

Observation

Normal Distribution: [X] Yes, [ ] No

Many iterations were tested for bin count of the pH histogram. A value of ‘50’ yielded the best result as seen above. pH followed a normal distribution pattern. The mean and median were almost identical at 3.180 and 3.188 respectively. Regarding quantiles, 80% of the balues were between 3.09 and 3.28.

Variable: fixed.acidity



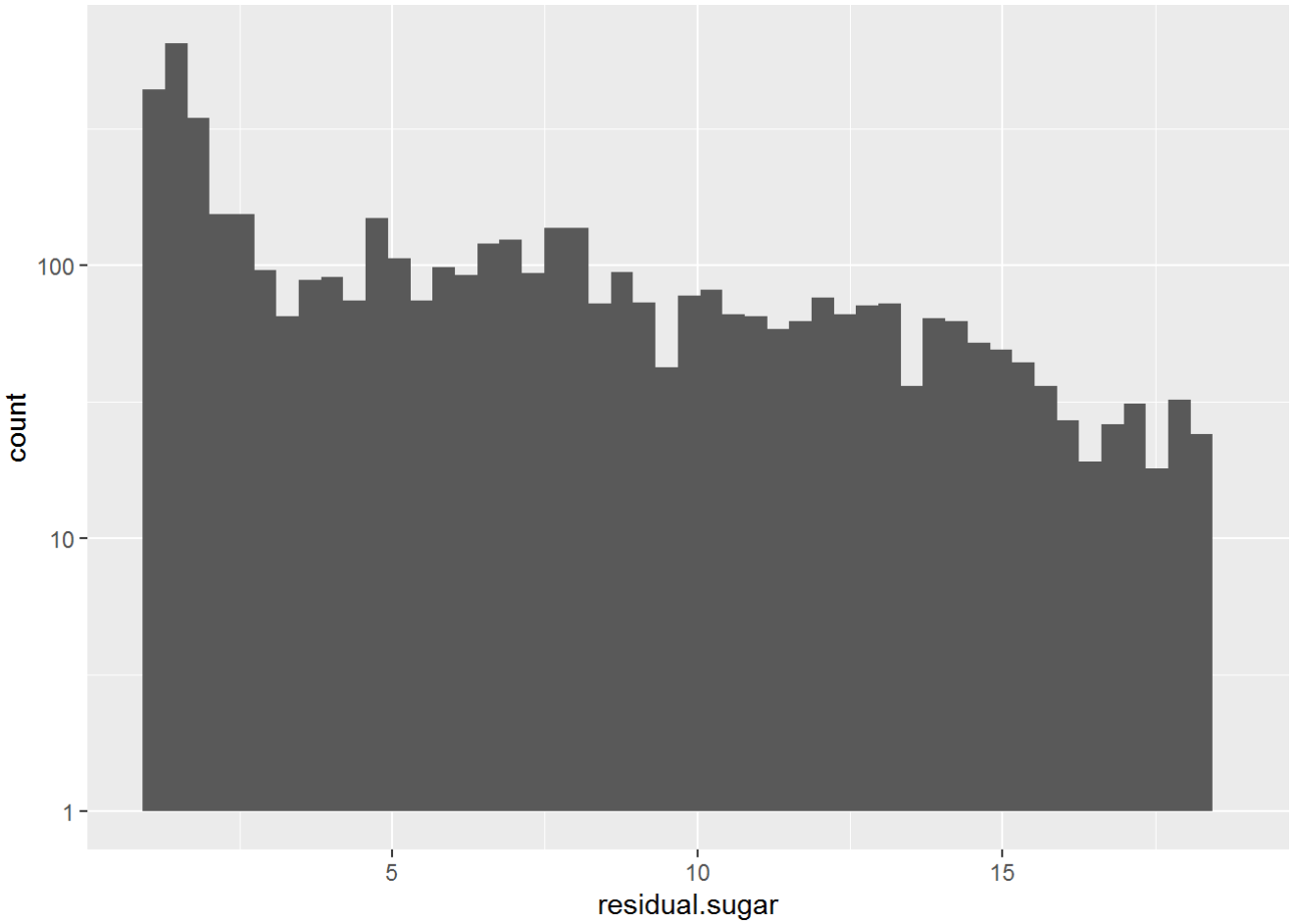

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

Observation

Normal Distribution: [X] Yes, [ ] No

Fixed acidity was a bit wider distribution than pH. At first thought, one might think these had a tighter correlation.

Variable: residual.sugar




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

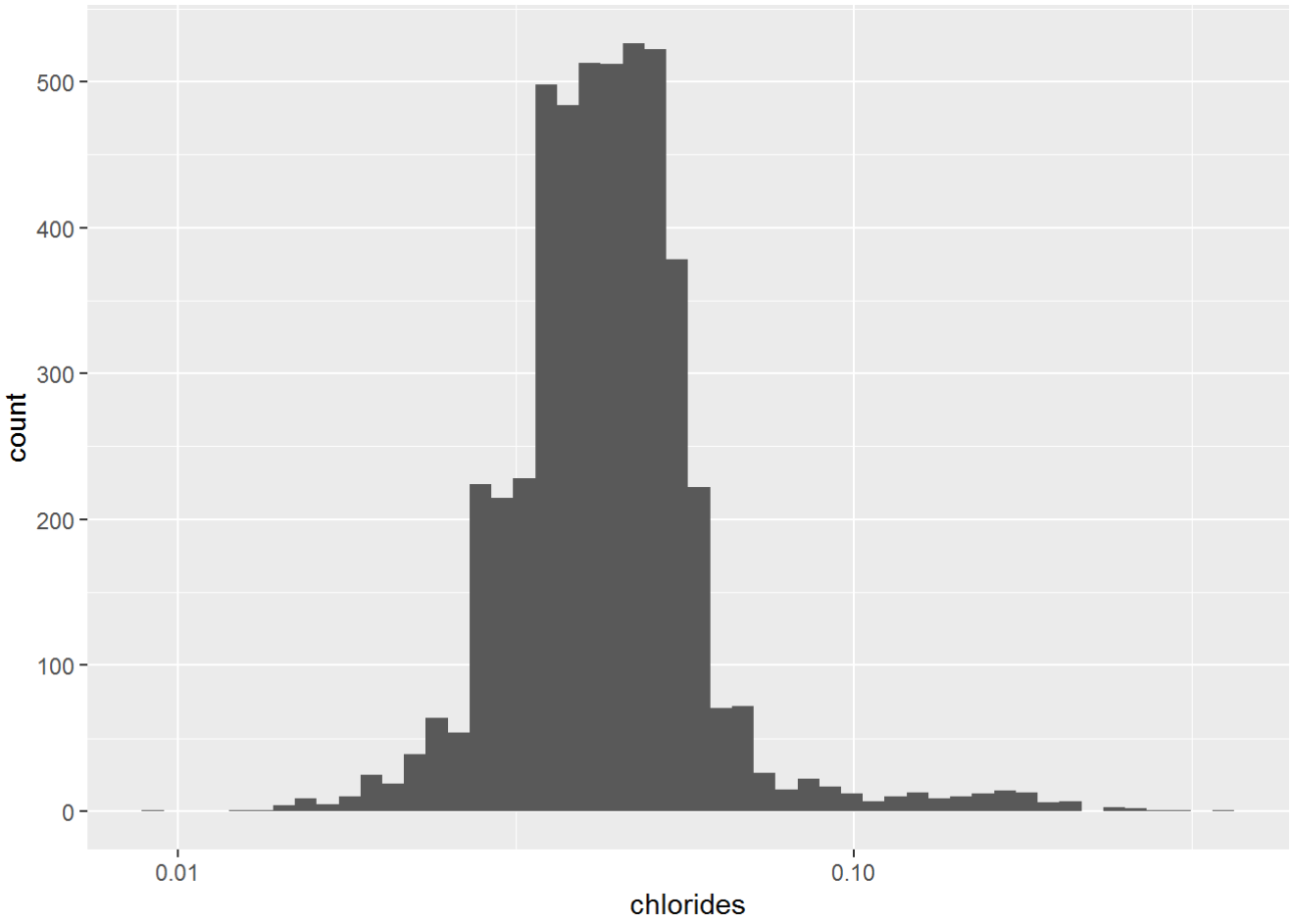
Observation

Normal Distribution: [ ] Yes, [X] No

Distribution: Exponential decline

Residual sugar was the only histogram that had a very different shape than the other variables in this section. The count was high for low levels of residual sugar and gradually reduced–almost linearly–as the sugar levels increased.

Variable: chlorides



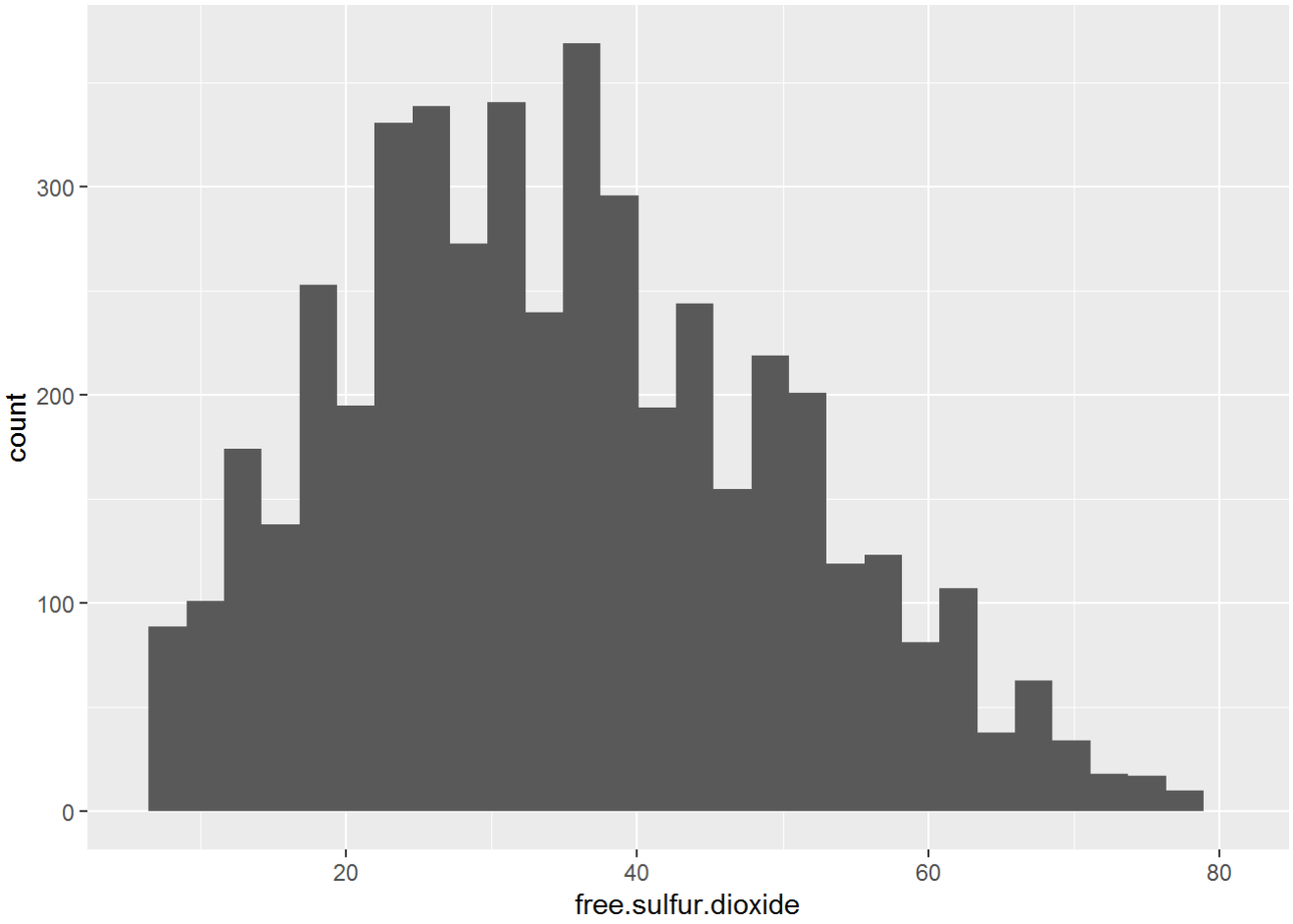

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

Observation

Normal Distribution: [X] Yes, [ ] No

Chloride levels were concentrated around .458. It is interesting to note that the ratio of max to min was 38.4:1.

Variable: free.sulfur.dioxide



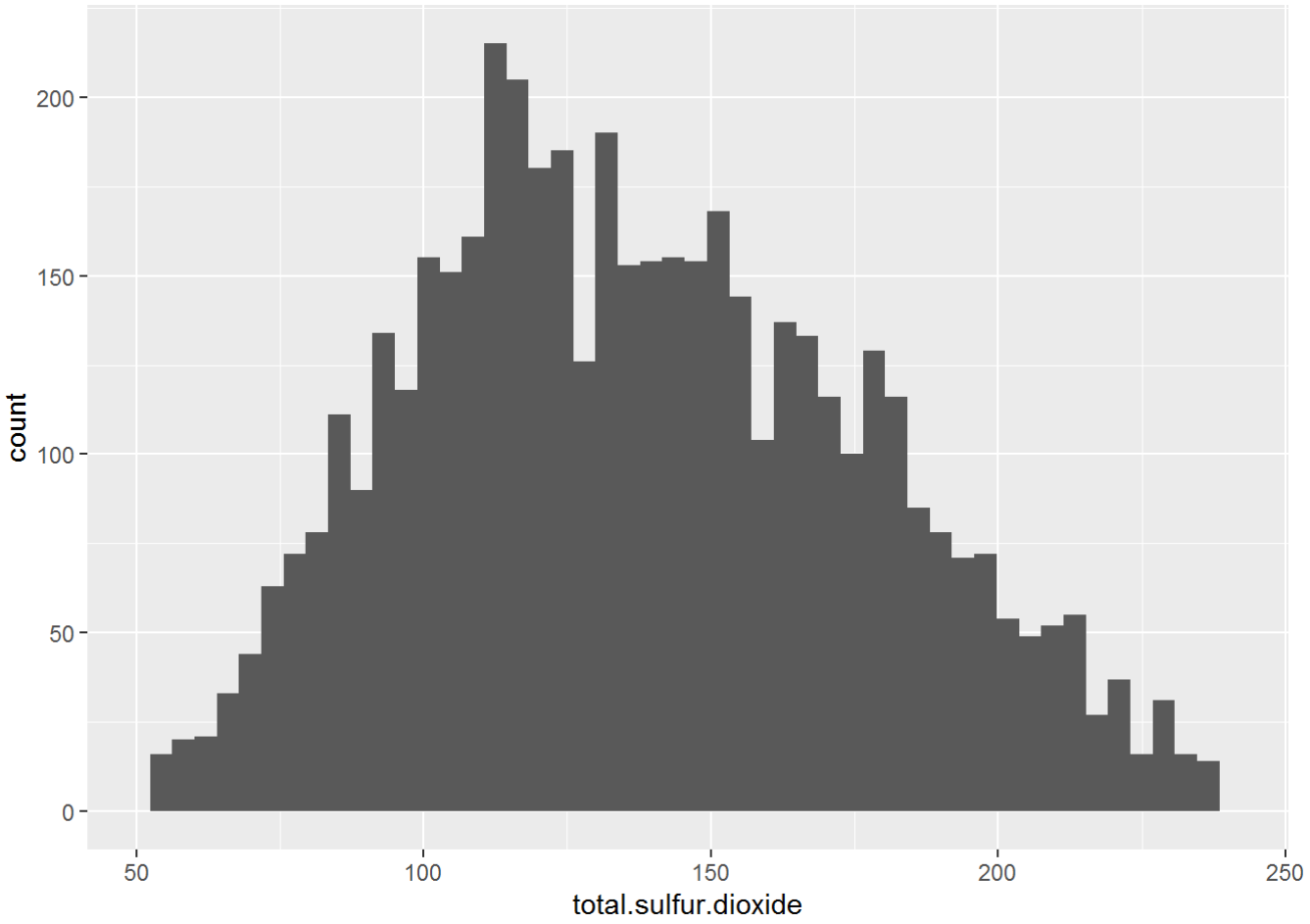

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

Observation

Normal Distribution: [X] Yes, [ ] No

The range of free sulfure dioxide was high.

Variable: total.sulfur.dioxide



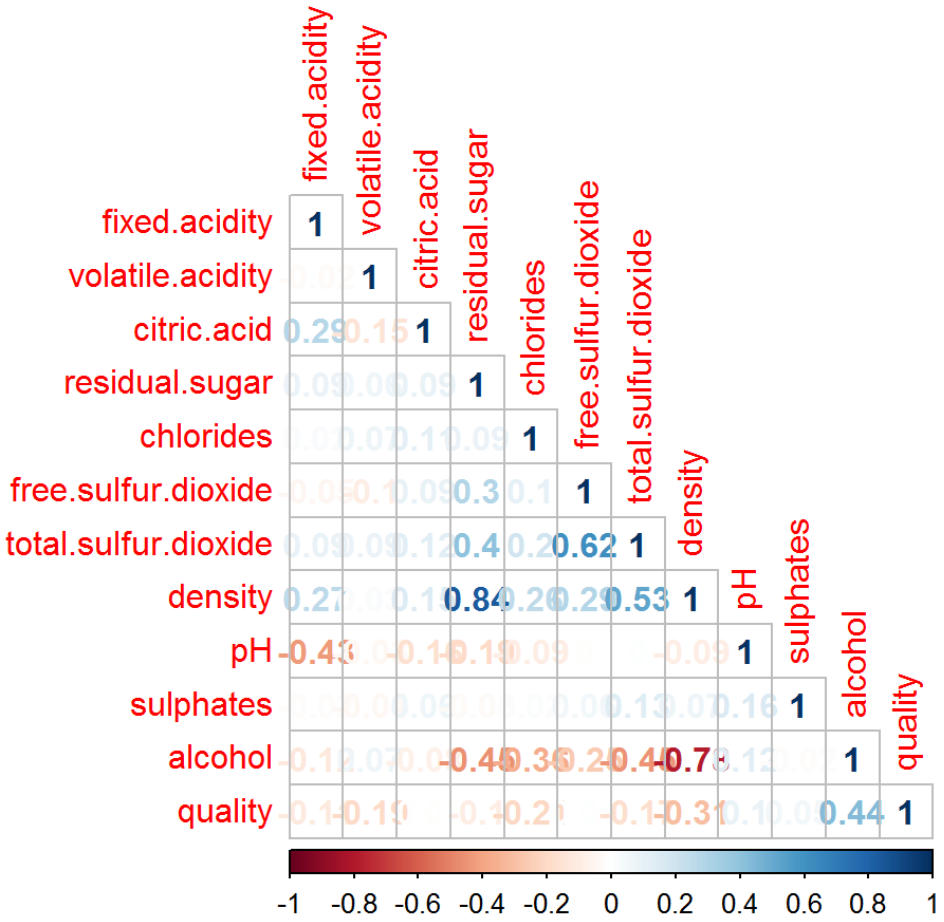

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

Observation

Normal Distribution: [X] Yes, [ ] No

Another high range.

Bivariate Plots Section



Variables with Absolute Correlation Greater than 0.300

free.sulfur.dioxide vs. residual.sugar

total.sulfur.dioxide vs. residual.sugar

total.sulfur.dioxide vs. free.sulfur.dioxide

density vs. residual.sugar

density vs. total.sulfur.dioxide

pH vs. fixed.acidity

alcohol vs. residual.sugar

alcohol vs. chlorides

alcohol vs. total.sulfur.dioxide

alcohol vs. density

quality vs. density

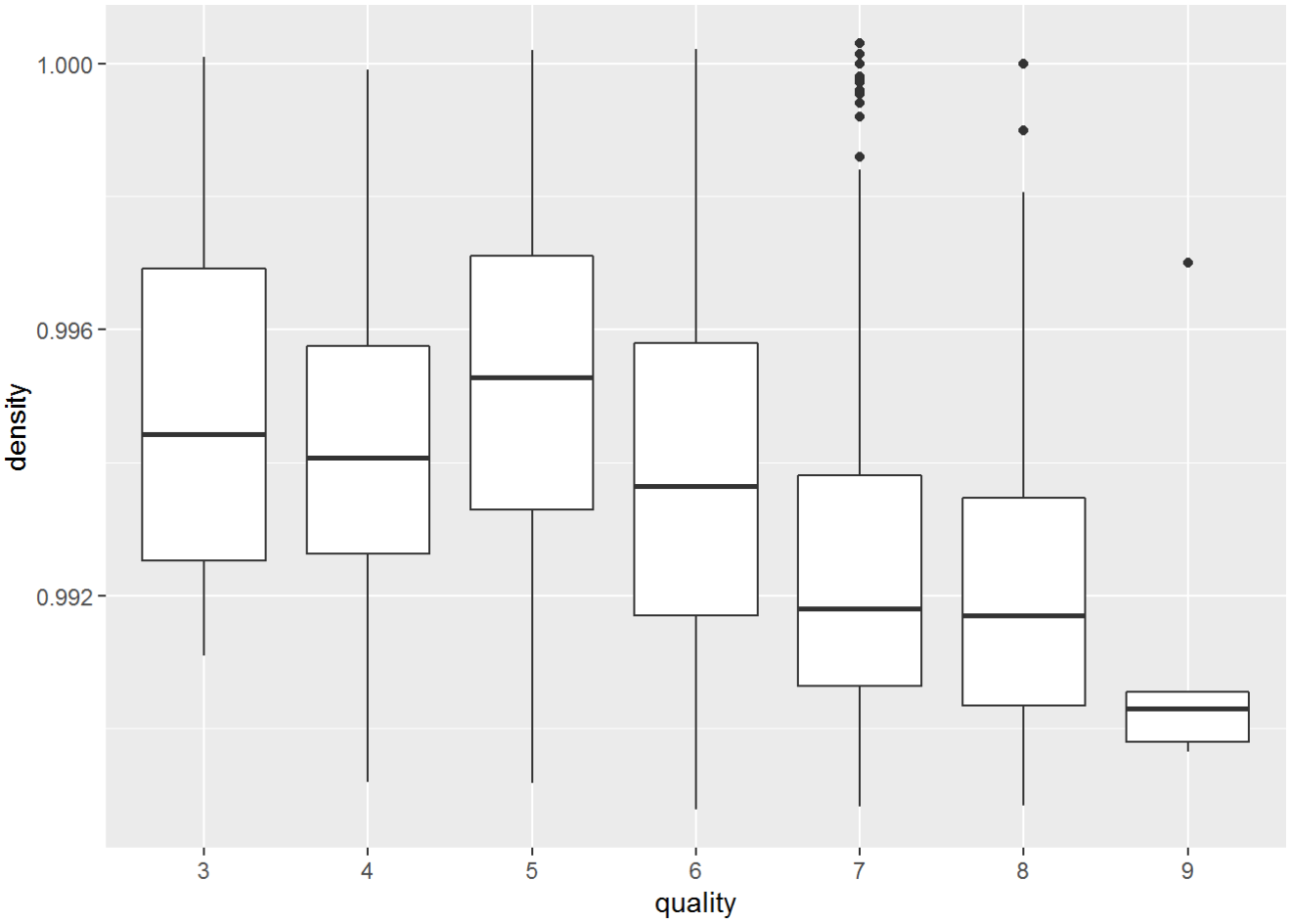
quality vs. alcohol

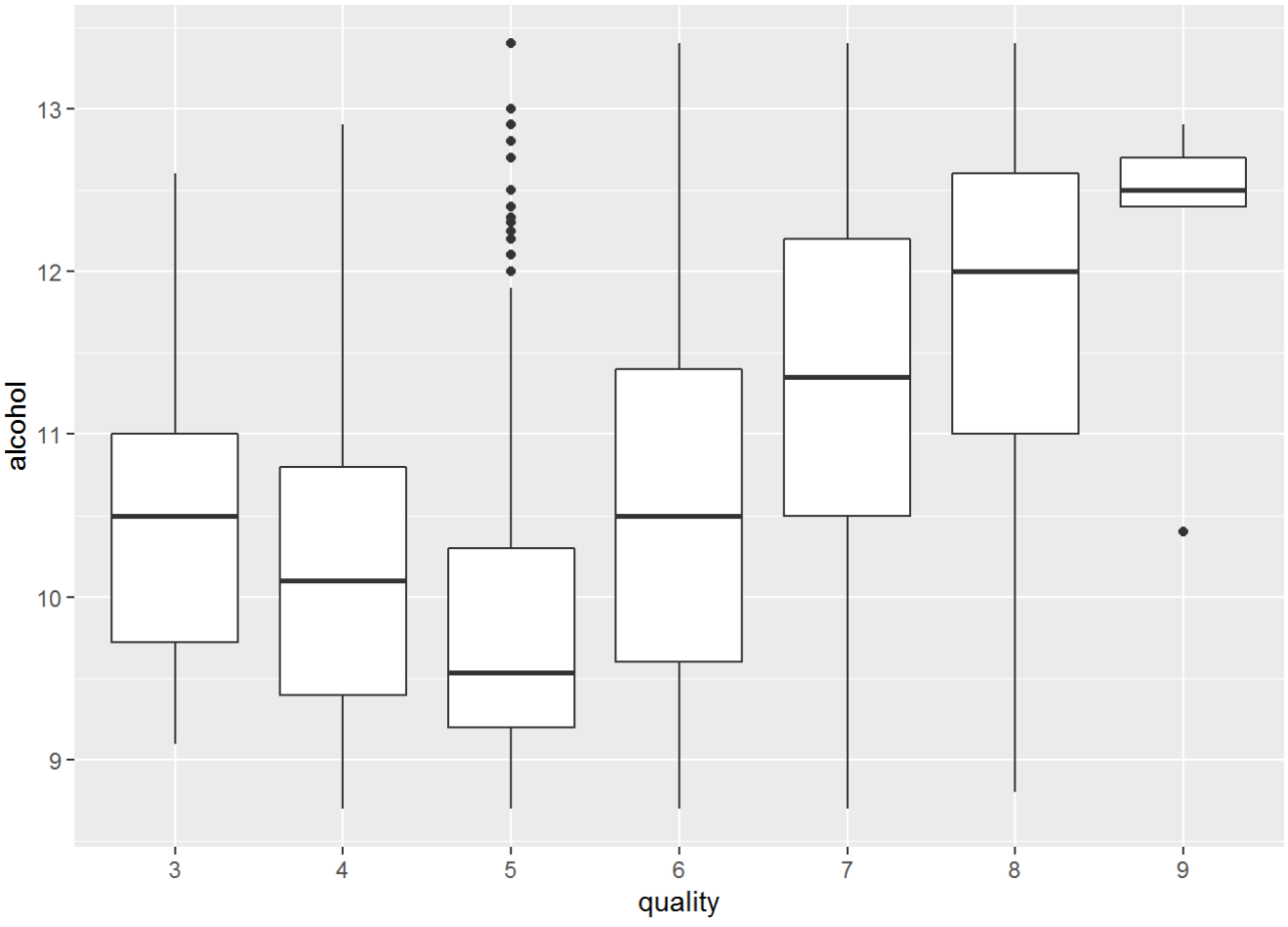
All of these relationships will be plotted for further analysis.

Highest 3 Correlations

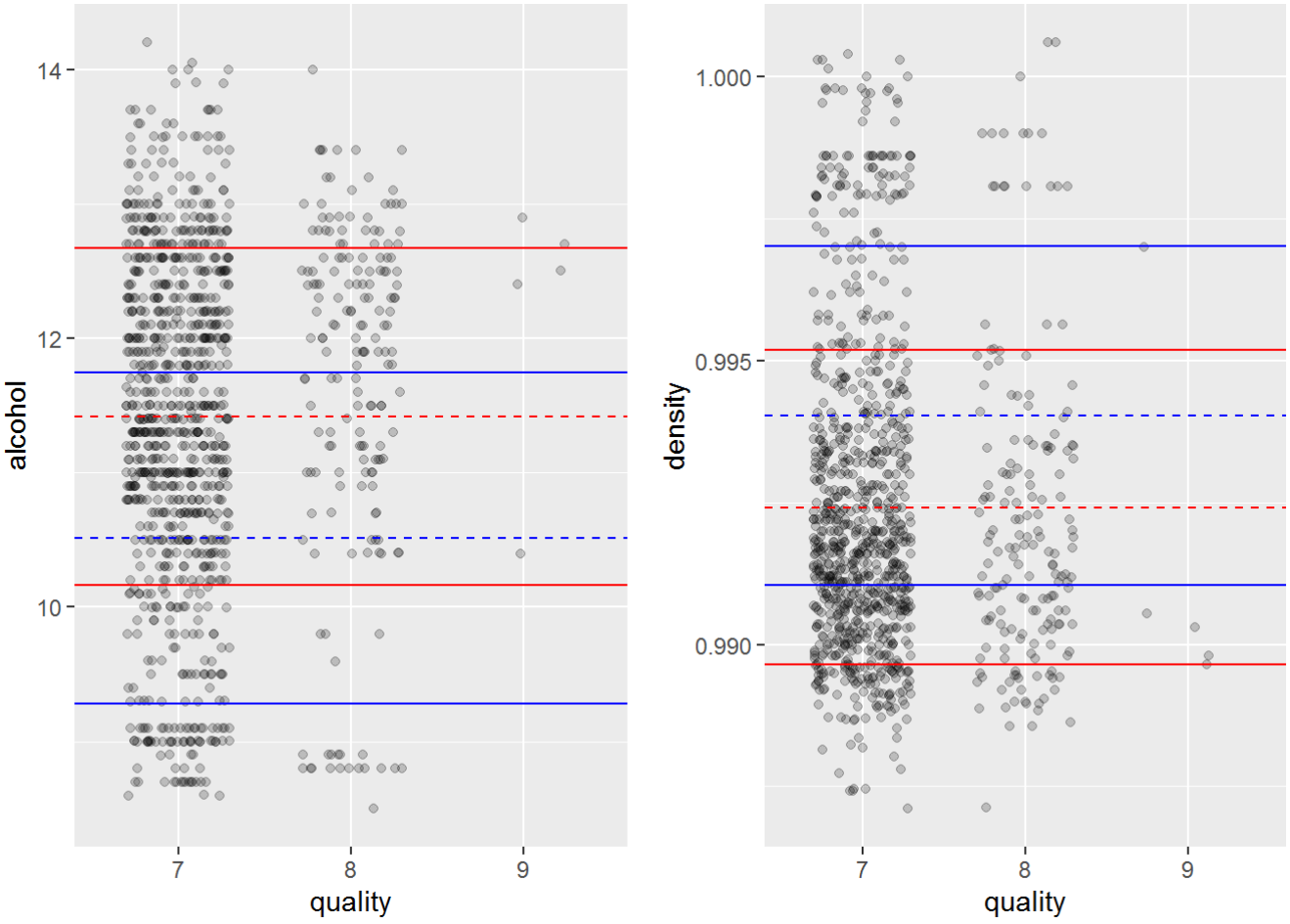
1. density vs. residual.sugar - 0.84
2. alcohol vs. density - -0.78
3. total.sulfur.dioxide vs. free.sulfur.dioxide - 0.62

Bivariate Analysis


1.000
0.996
0.992
density
3
4
5
6
7
8
9
quality


13
12
11
10
9
alcohol
3
4
5
6
7
8
9
quality

These two box plots illustrate two interesting points: 1) Quality increases as density decreases 2) Quality increases as alcohol increases

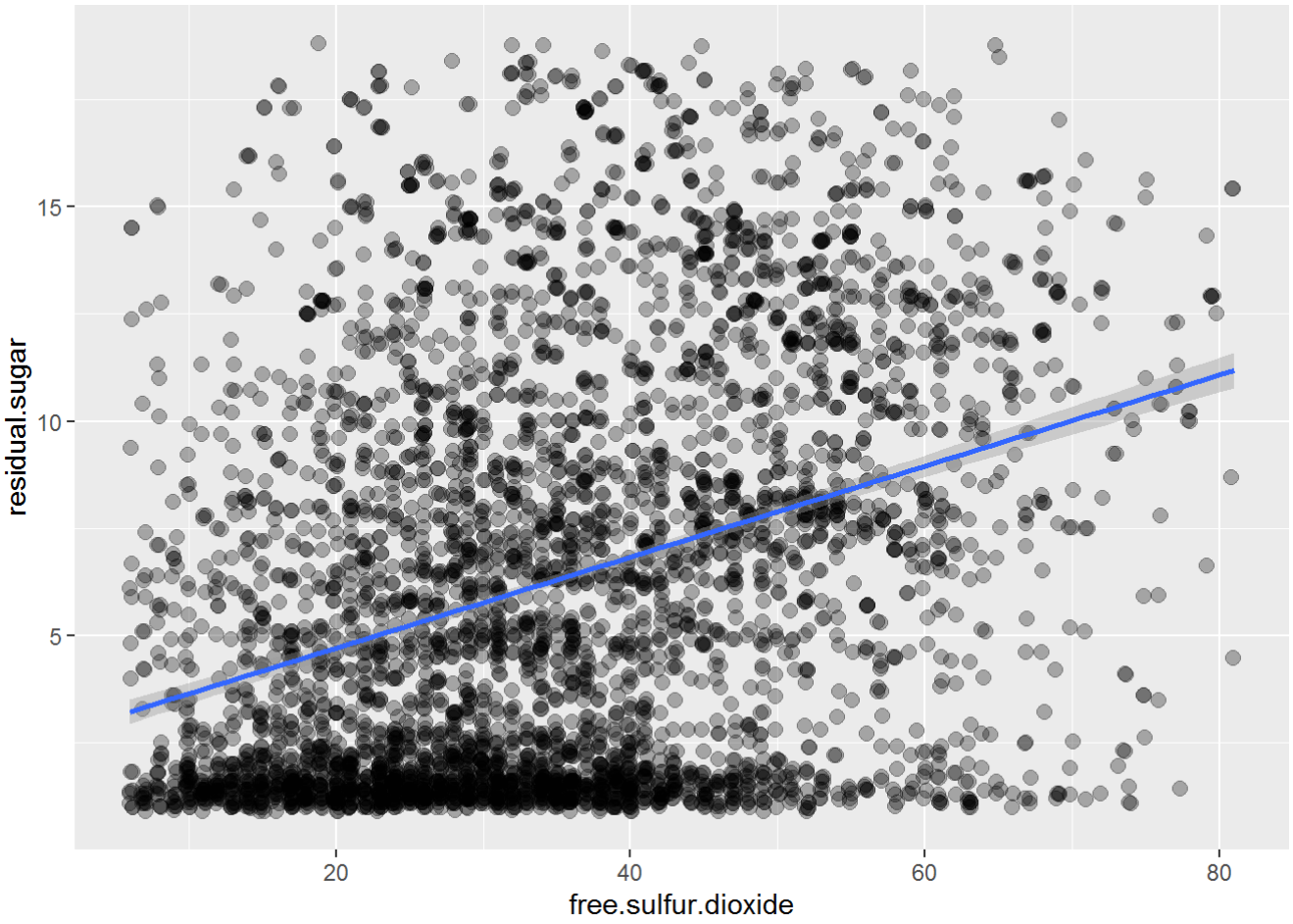


Relationships to Main Features

CORRPLOTwas used to provide some guidance on correlations between 2 variables in the data set. The correlations to quality were exceptionally low. Only 2 variables yielded a |Cor| >= 0.300 with wine quality: Quality vs Density (-0.307) and Quality vs Alcohol (0.436). Many variables seemed to affect density and alcohol.

There were some interesting data points gathered for alcohol and density vs quality. Data points with quality greater than 6 were stored in a separate data. The relationship between alcohol and density vs quality with hlines at the mean, -1 SD, and +1 SD of the entire data set. It was interesting to note that high quality wine had a much higher mean and SD than the entire data set. Density of high quality wine was much lower than the entire data set.

Plot: residual.sugar vs. free.sulfur.dioxide

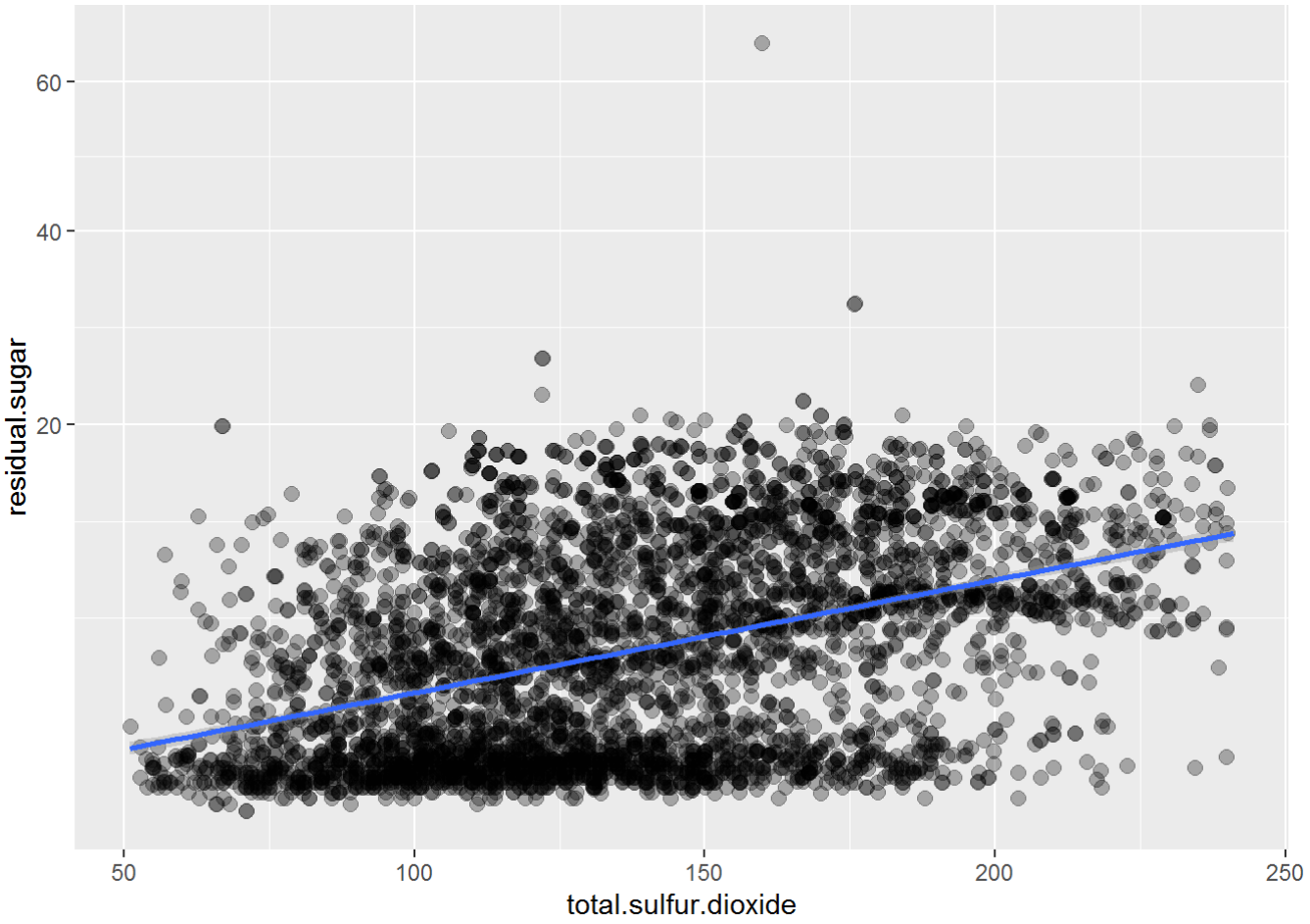



```
## [1] 0.2990984
```

Description

Not a strong correlation. Most data resides at low residual sugar levels.

Plot: residual.sugar vs. total.sulfur.dioxide

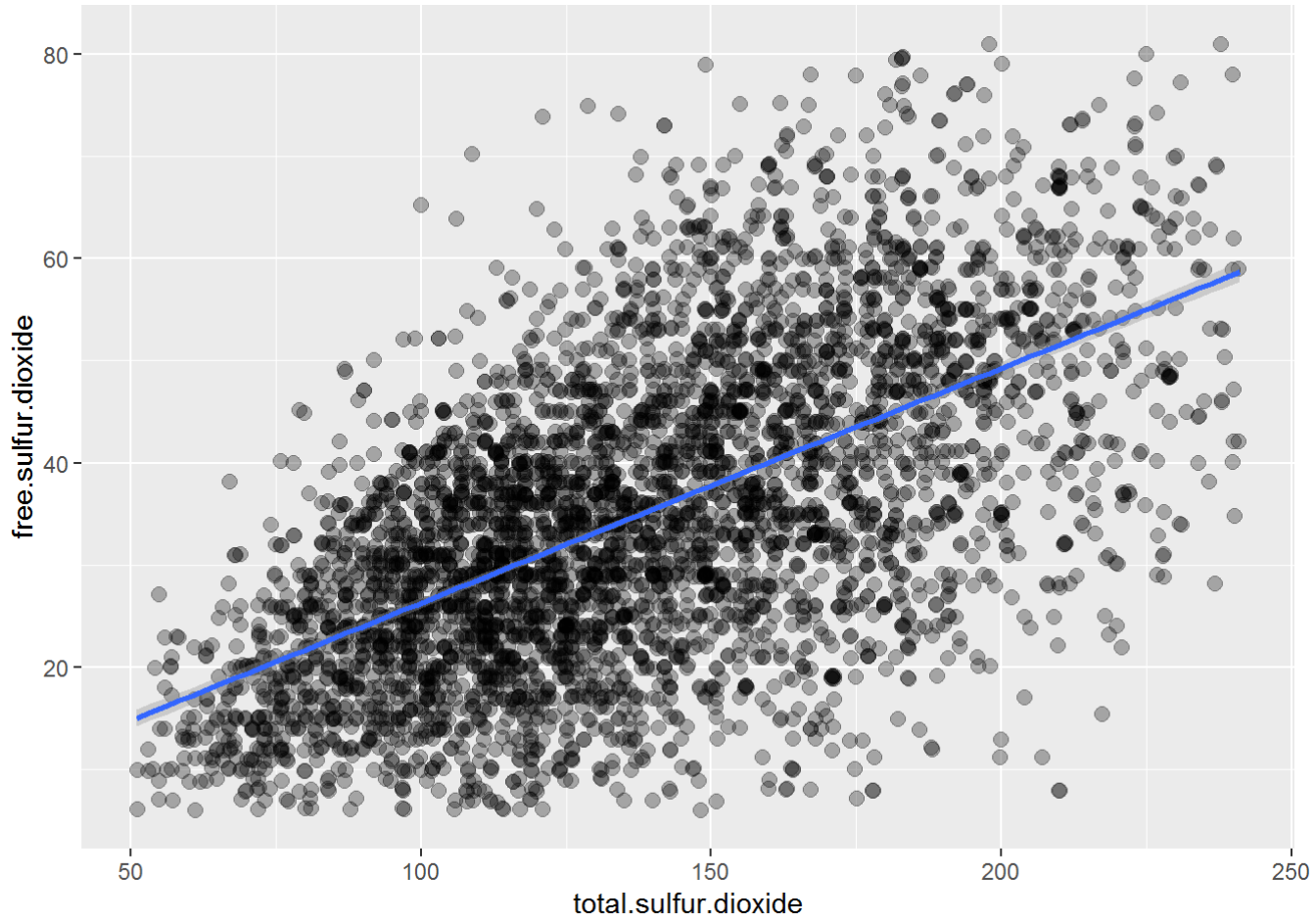



```
## [1] 0.4014393
```

Description

Appears to have some correlation. Again, most residual sugar levels remain low and collected.

Plot: free.sulfur.dioxide vs. total.sulfur.dioxide

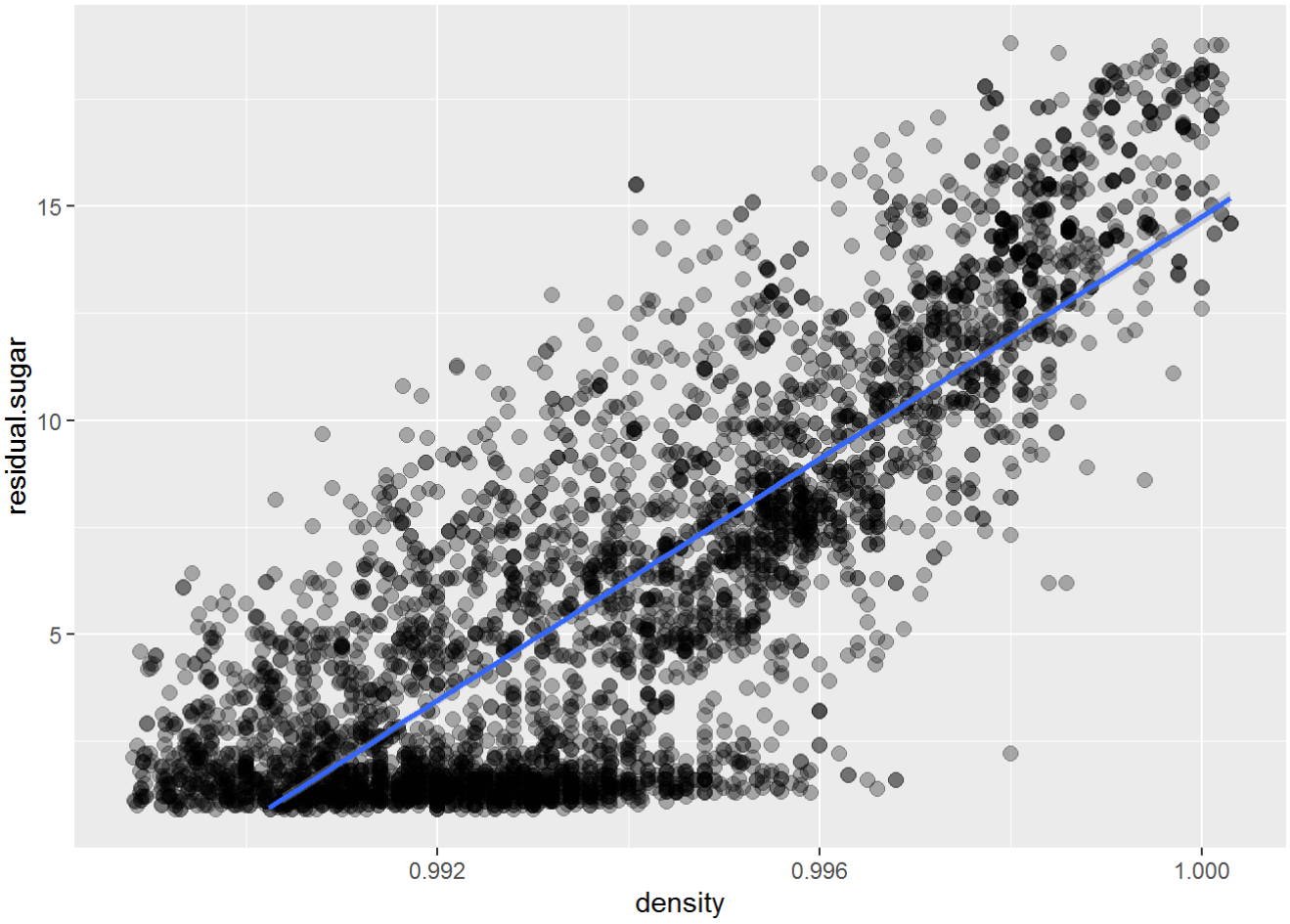



```
## [1] 0.615501
```

Description

Strong correlation. This would be expected since free SO2 levels are summed into total SO2 levels. Thus, as free SO2 level increase, so would the total.

Plot: residual.sugar vs. density

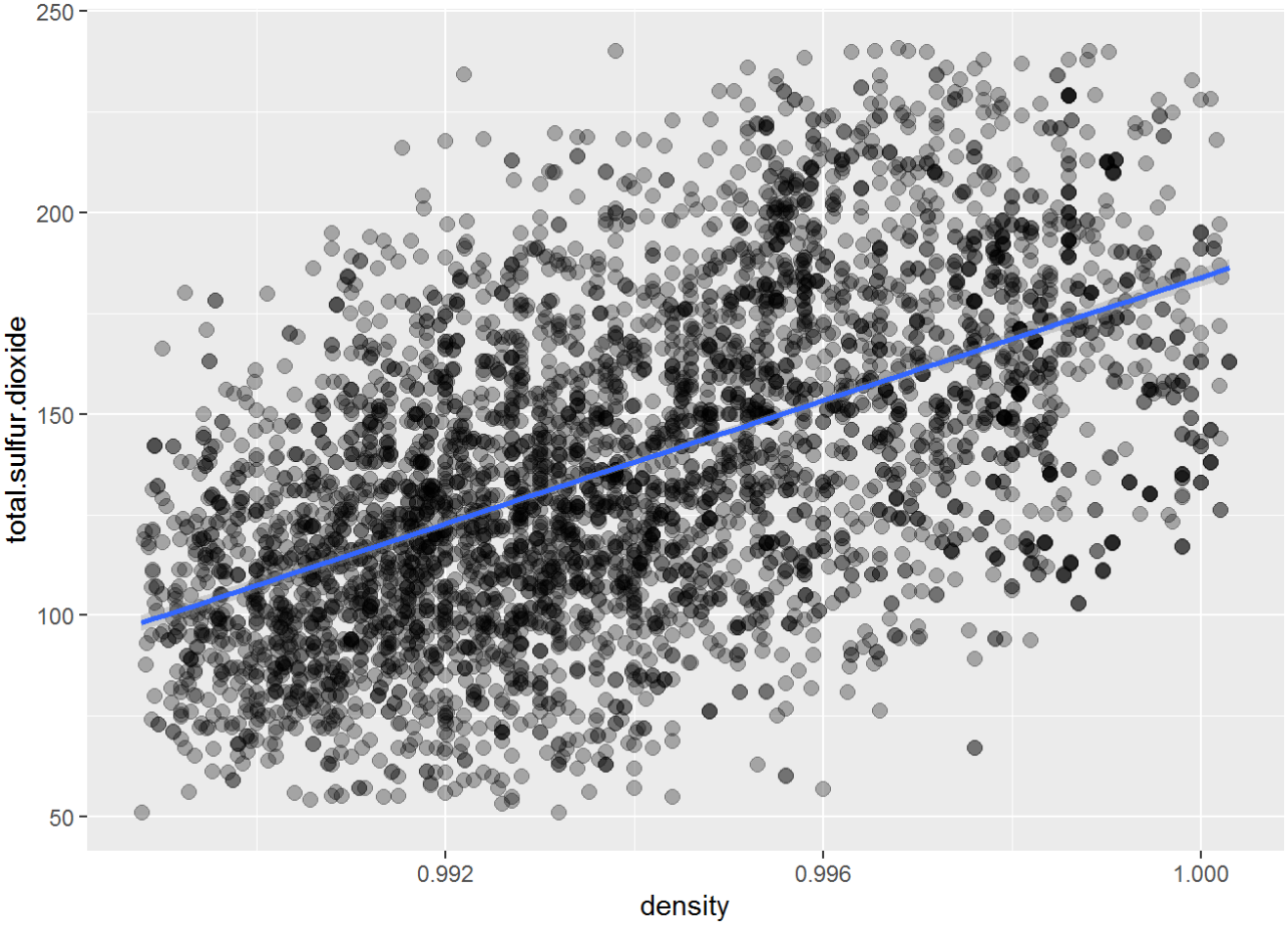



```
## [1] 0.8389665
```

Description

Strong correlation. Intuitively, this would make sense: as residual sugar levels increase, density increases.

Plot: total.sulfur.dioxide vs. density

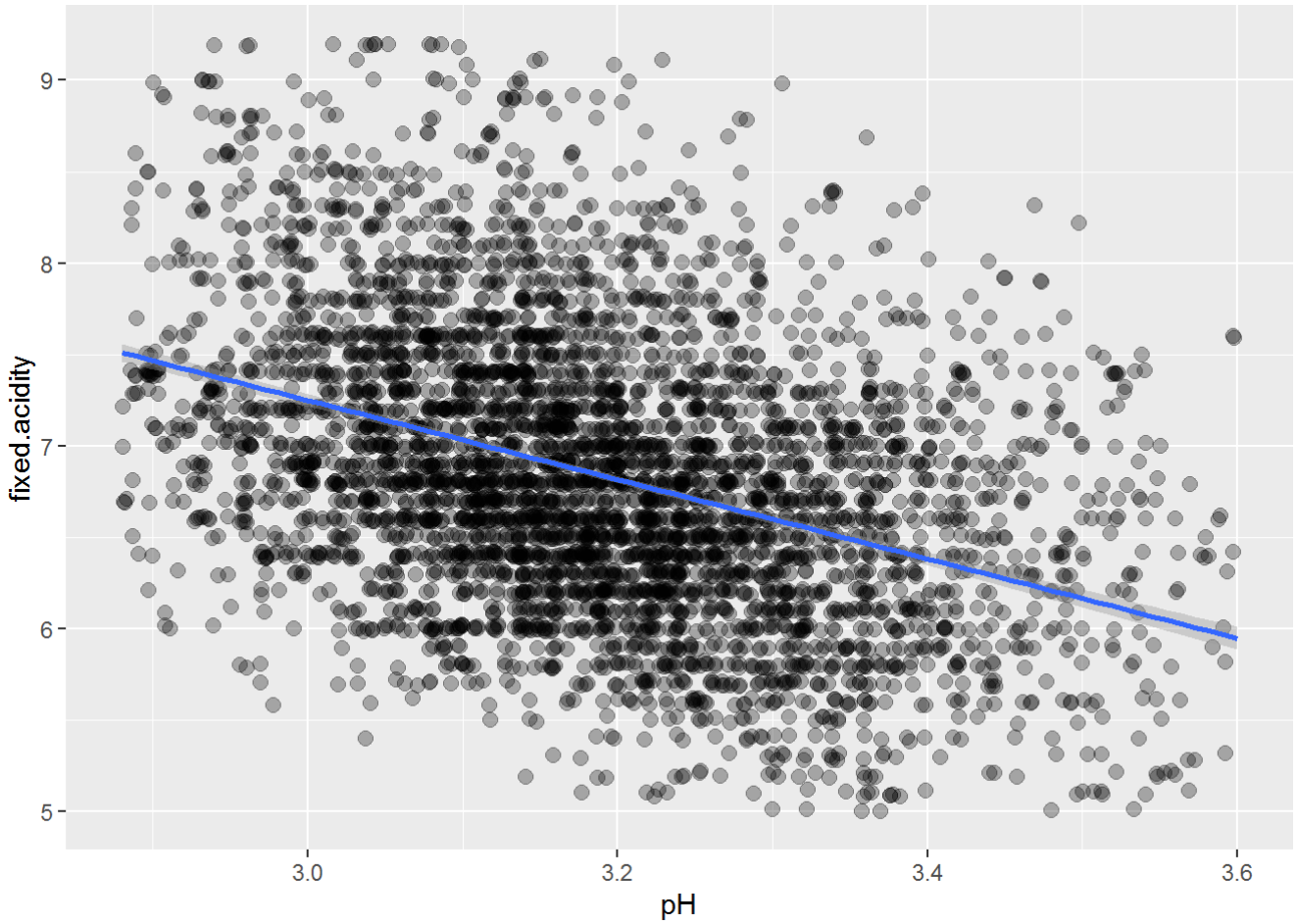



```
## [1] 0.5298813
```

Description

Correlation is fairly strong. This is a bit surprising considering SO2 density 2.63kg/m^3–significantly lower than the density of wine of roughly 995 kg/m^3. It would make sense that they would be inversely proportional.

Plot: fixed.acidity vs. pH

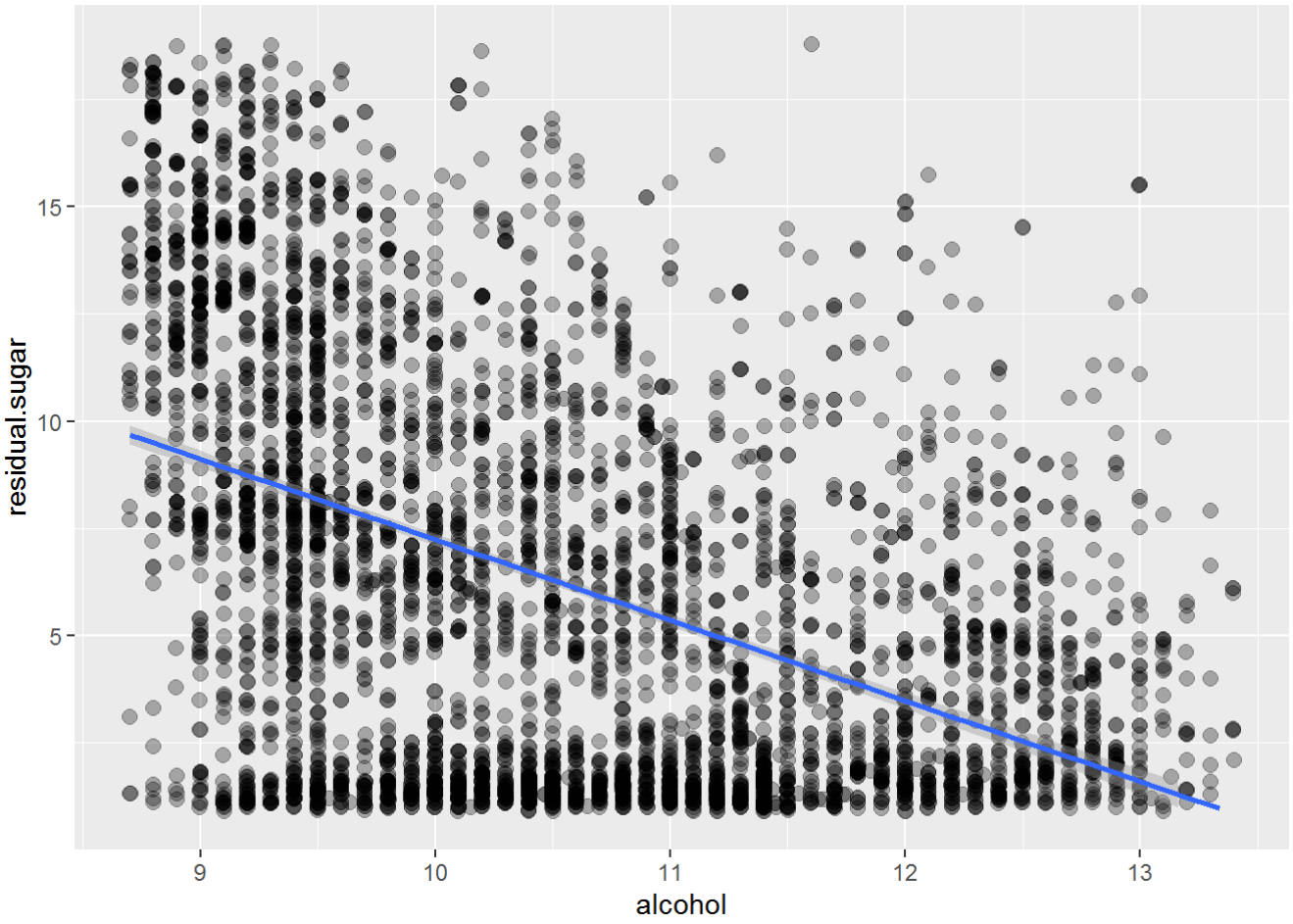



```
## [1] -0.4258583
```

Description

Strong correlation. This makes sense since pH is a measurement of acidity and alkalinity. Thus, pH is lower if the level of acidity is high.

Plot: residual.sugar vs. alcohol

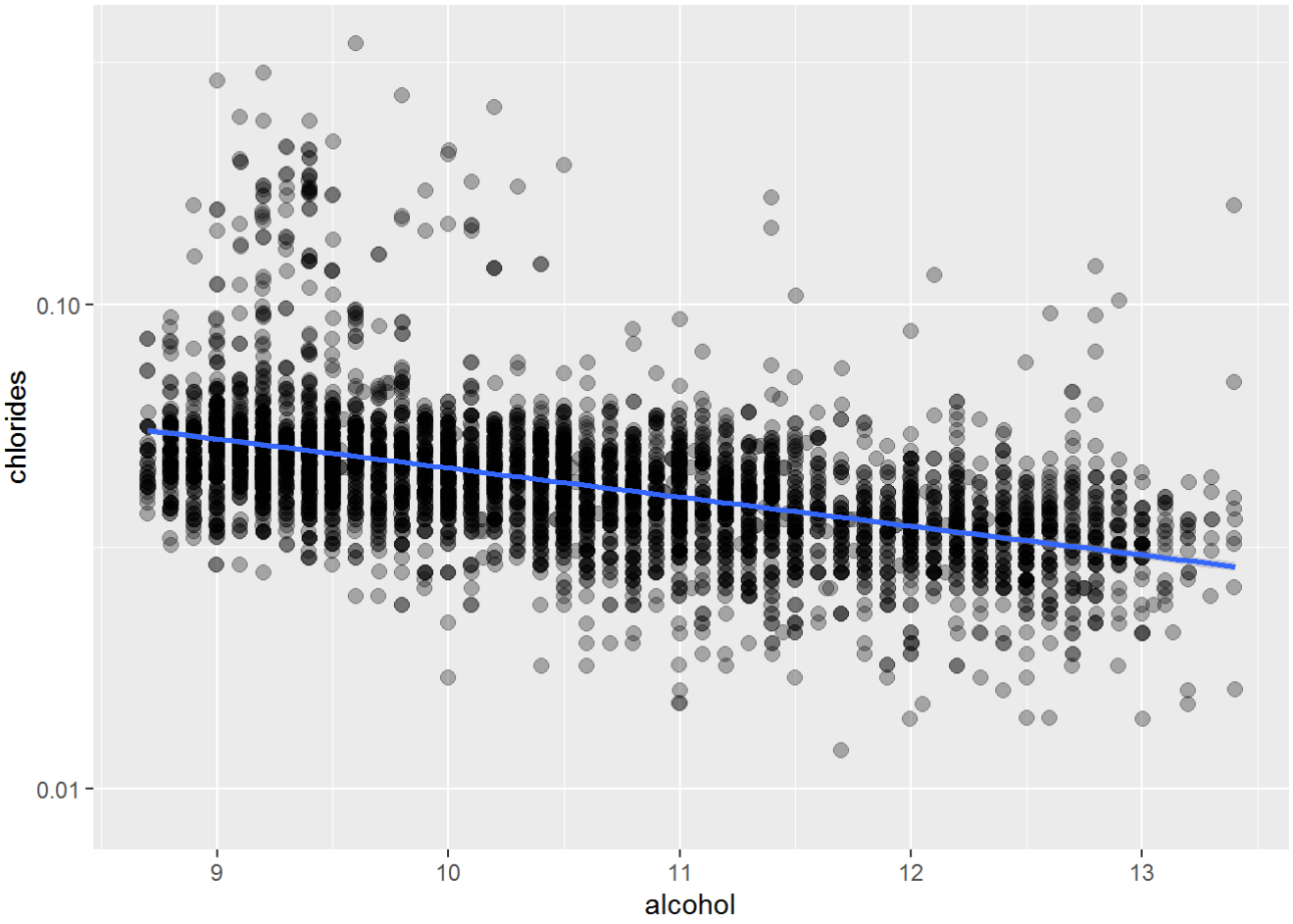



```
## [1] -0.4506312
```

Description

Decent correlation. Alcohol is created from sugar. Alcohol would increase as sugar is consumed by yeast.

Plot: chlorides vs. alcohol

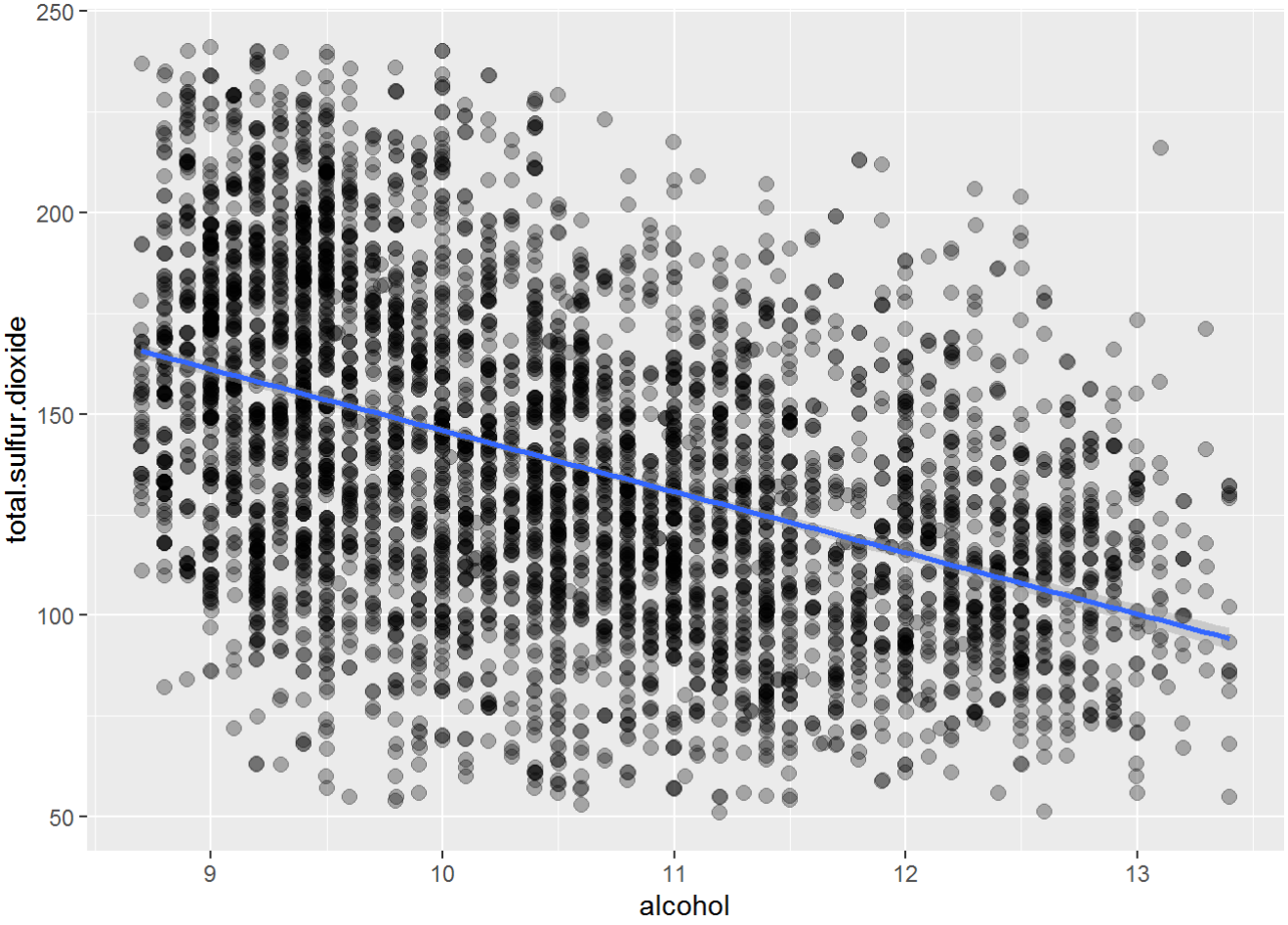



```
## [1] -0.3601887
```

Description

Strong correlation. The level of chlorides appears to gradually reduce as the level of alcohol increases. There aresome outliers.

Plot: total.sulfur.dioxide vs. alcohol

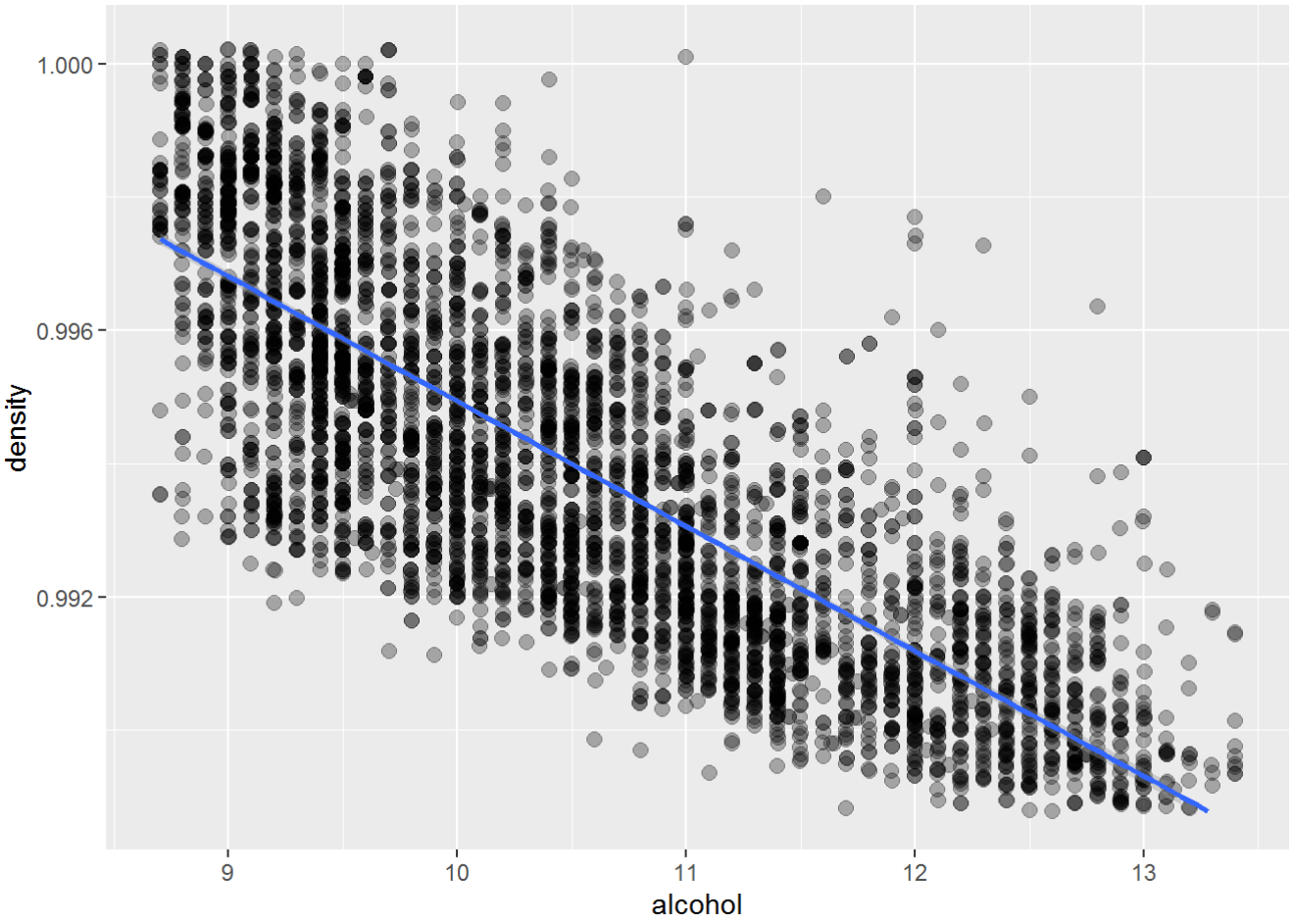



```
## [1] -0.4488921
```

Description

Reasonable correlation. Inversely proportional.

Plot: density vs. alcohol




```
## [1] -0.7801376
```

Description

Strong correlation. Alcohol is a low density chemical. This decreases the density of the wine since wine has a much higher density.

Relationships to Other Features

A CORRPLOT matrix provided quick insight into potential variables to investigate. All pairs of variables with |Correlation| >= 0.300 were plotted as a guide for deeper investigation.

Free SO2 vs. total SO2 had a tighter correlation. This isn't terribly surprising since has free SO2 levels would increase so too would the total SO2 levels.

Fixed acidity had a seemingly predictable relationship with pH. As acidity levels increase pH–a measurement of acidity or alkalinity–would decrease.

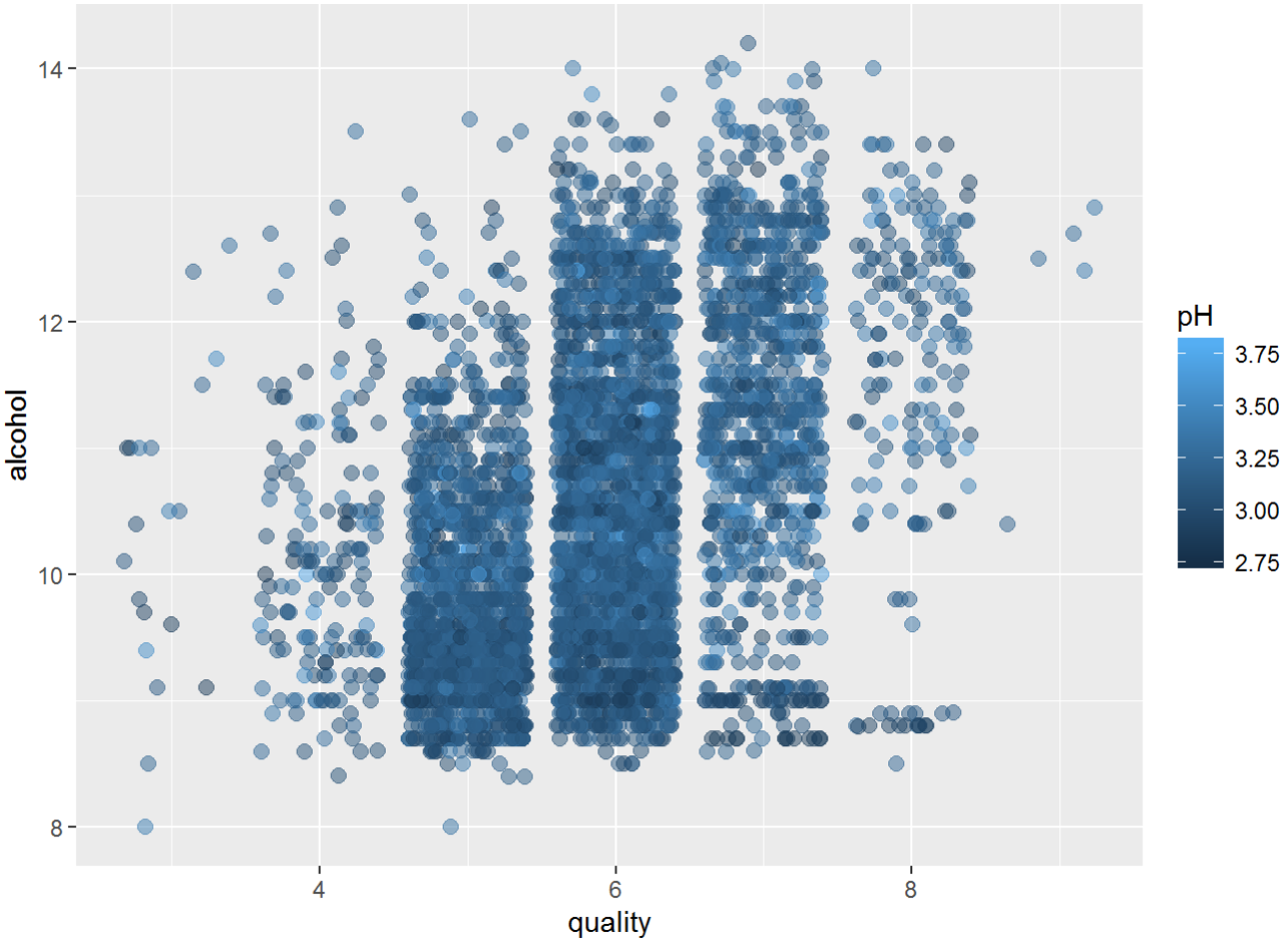
Density vs. Alcohol were very closely related. As alcohol increased density decrease. This makes sense as alcohol is a low density chemical. The 3 variables that correlated the most with alcohol and density (i.e. residual.sugar, chlorides, and total.sulfur.dioxide) will be analyzed further below. These variables as well as alcohol vs. density will be plotted with quality to understand further what makes a great wine.

Strongest Relationship Found

The strongest relationship found was between residual.sugar and density. The correlation between the two variables was 0.839.

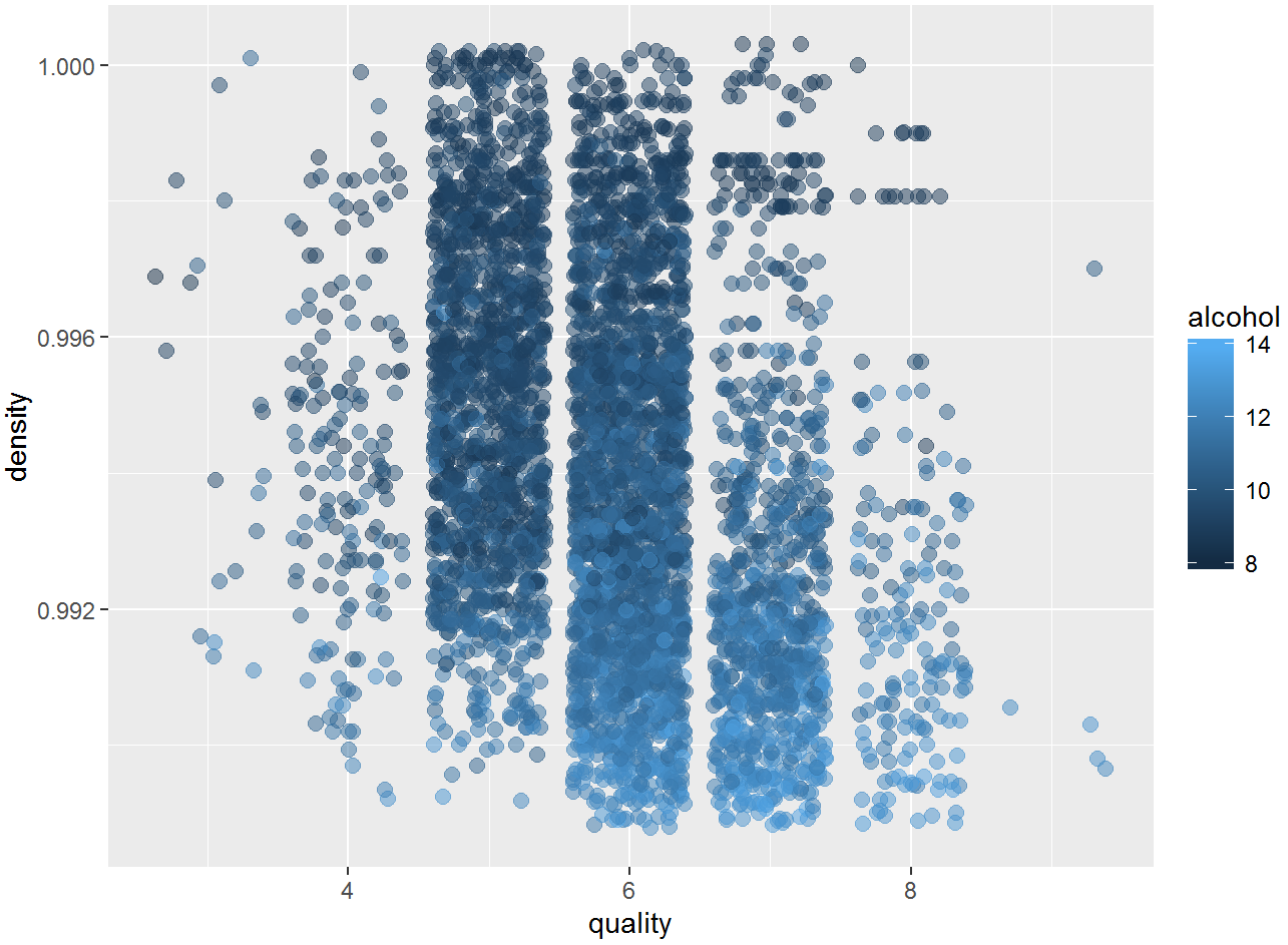
Total SO2 vs. density also had an upward correlation. This would create the notion that density would also be sensitive to free SO2 levels; however, that correlation was a mere 0.29.

Multivariate Plots Section



Description of Alcohol vs. Quality with pH

Plots appear very scattered with not much relation to pH. Some higher acidity wines (low pH) appear to be lower in alcohol level.

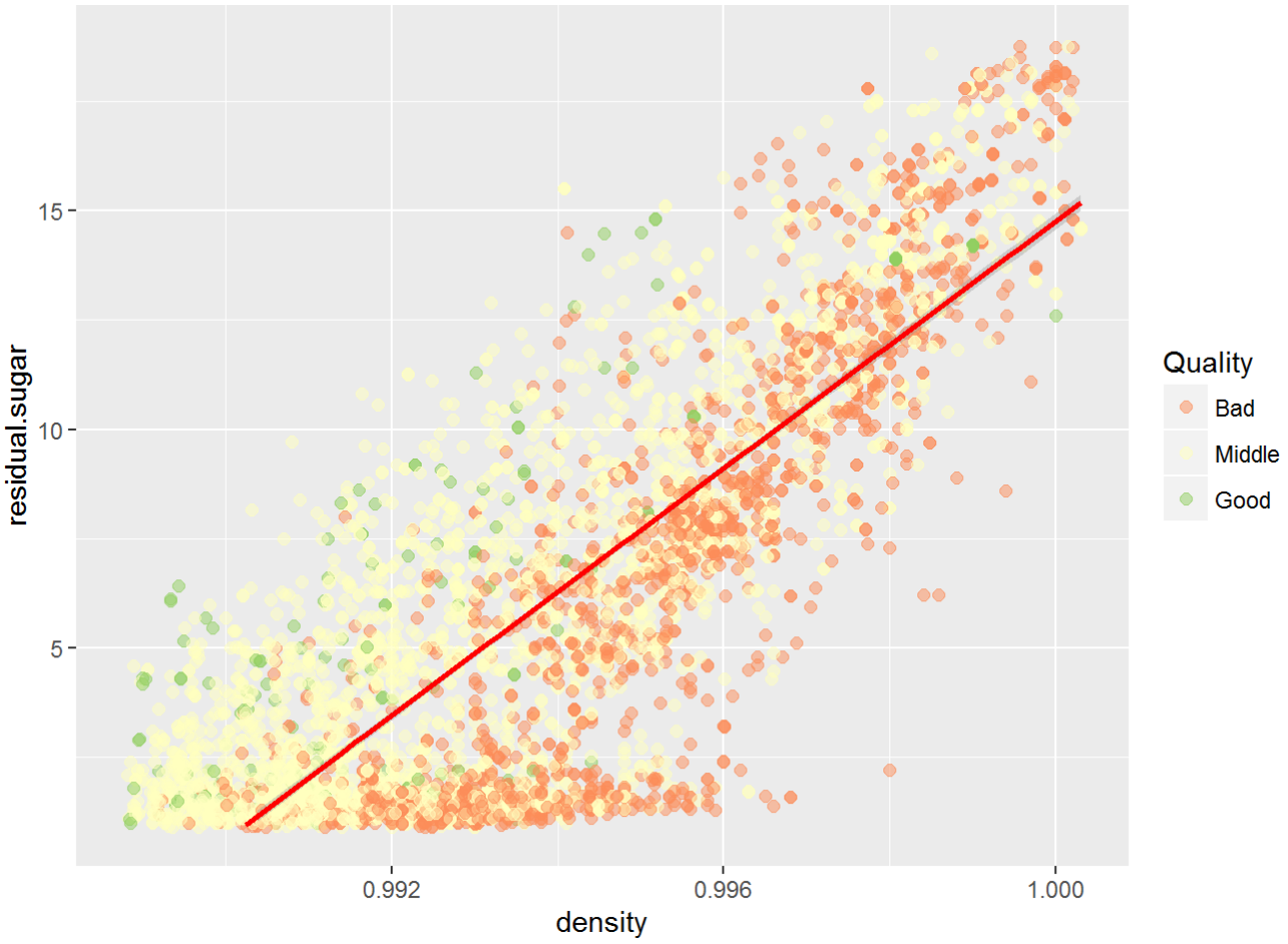


Description of Density vs. Quality with Alcohol

Higher levels of alcohol tend to have lower levels of density. Higher alcohol wines appear to be better quality as well.

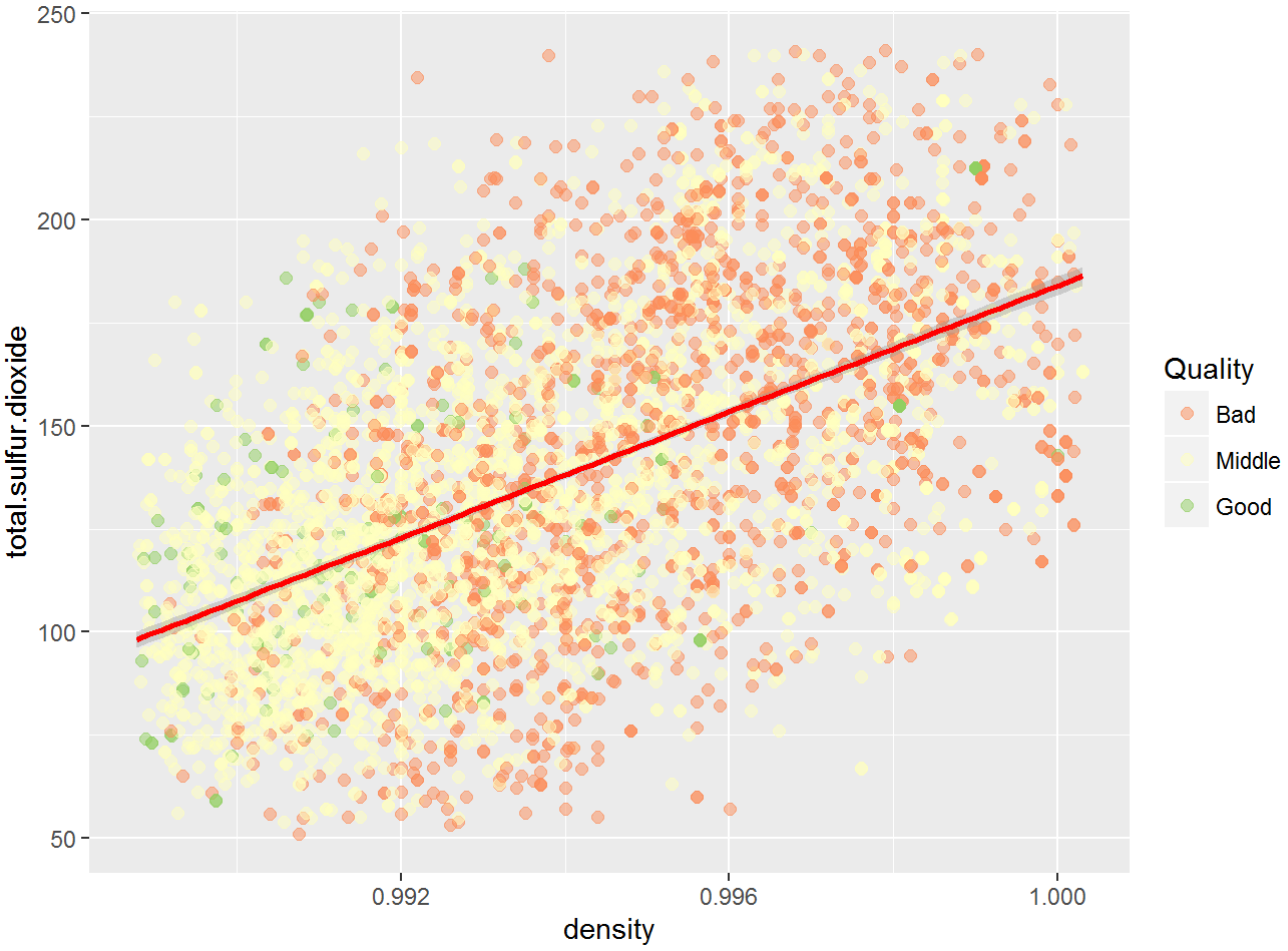
```
## quality.bucket
##    Bad Middle   Good
##   1640   3078    180
```

The table above is a new bucket for quality variables. The bucket will be used to separate wines into more intuitive categories: bad, middle, and good.



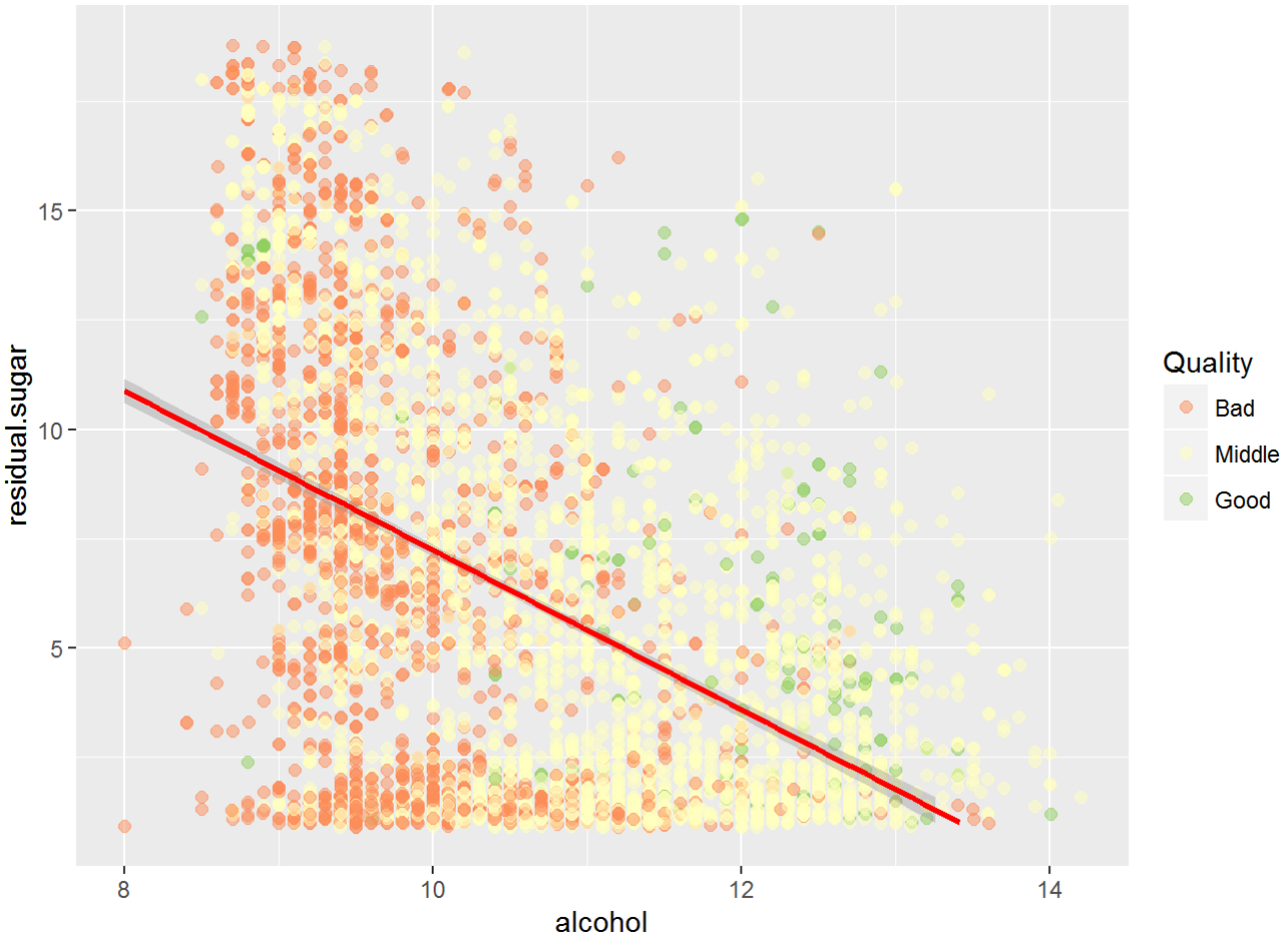
Description of Density vs. Residual Sugar

This is an update to a previous plot with quality colored into the graph. The level of sugar appears to have no affect on bad wines; however, it appears that most good wines have lower levels of sugar.



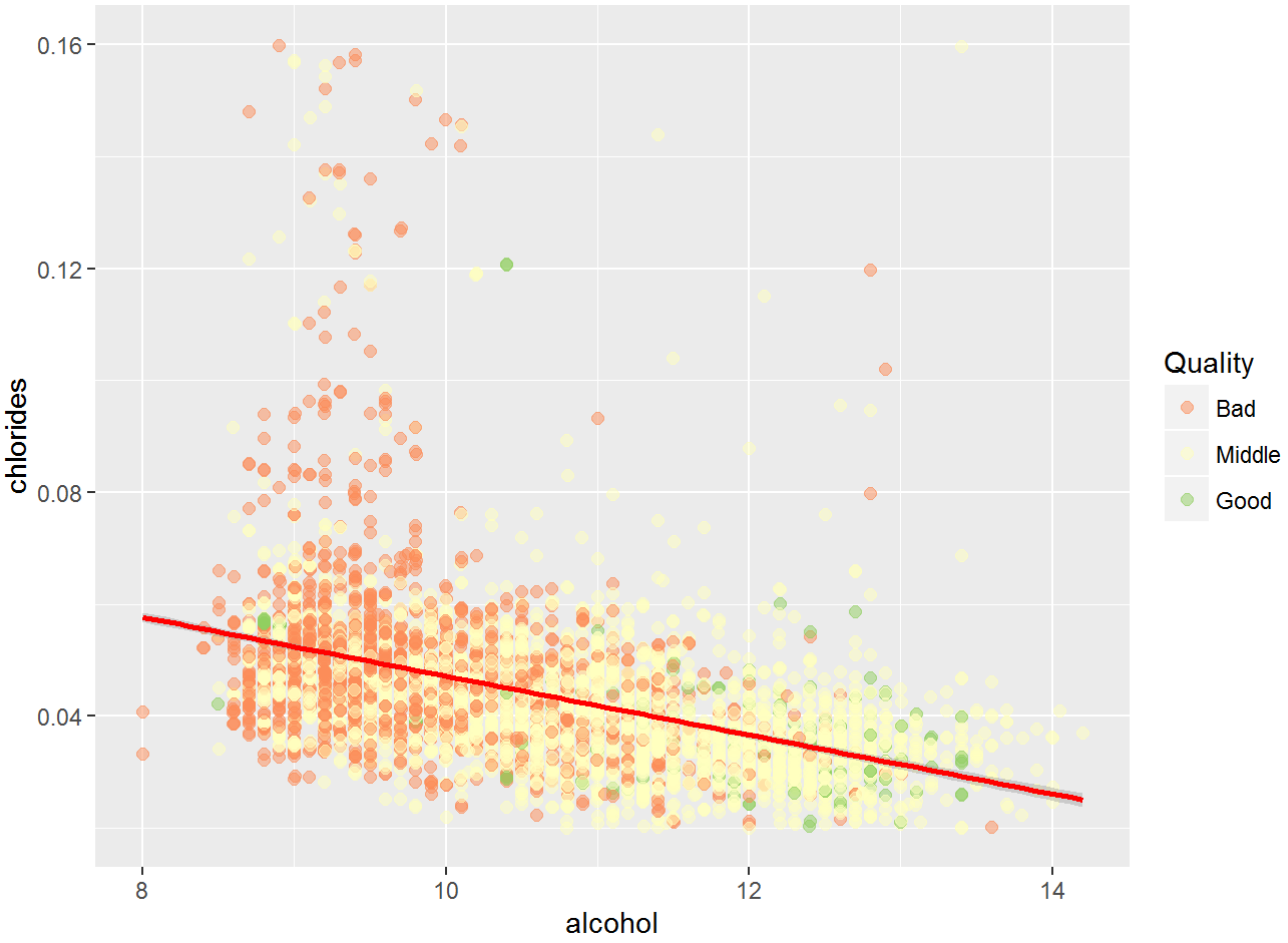
Description of Density vs. Total SO2

A clear trend can be observed that bad wines have higher levels of total SO2 and good wines have lower levels; however, the difference is small.



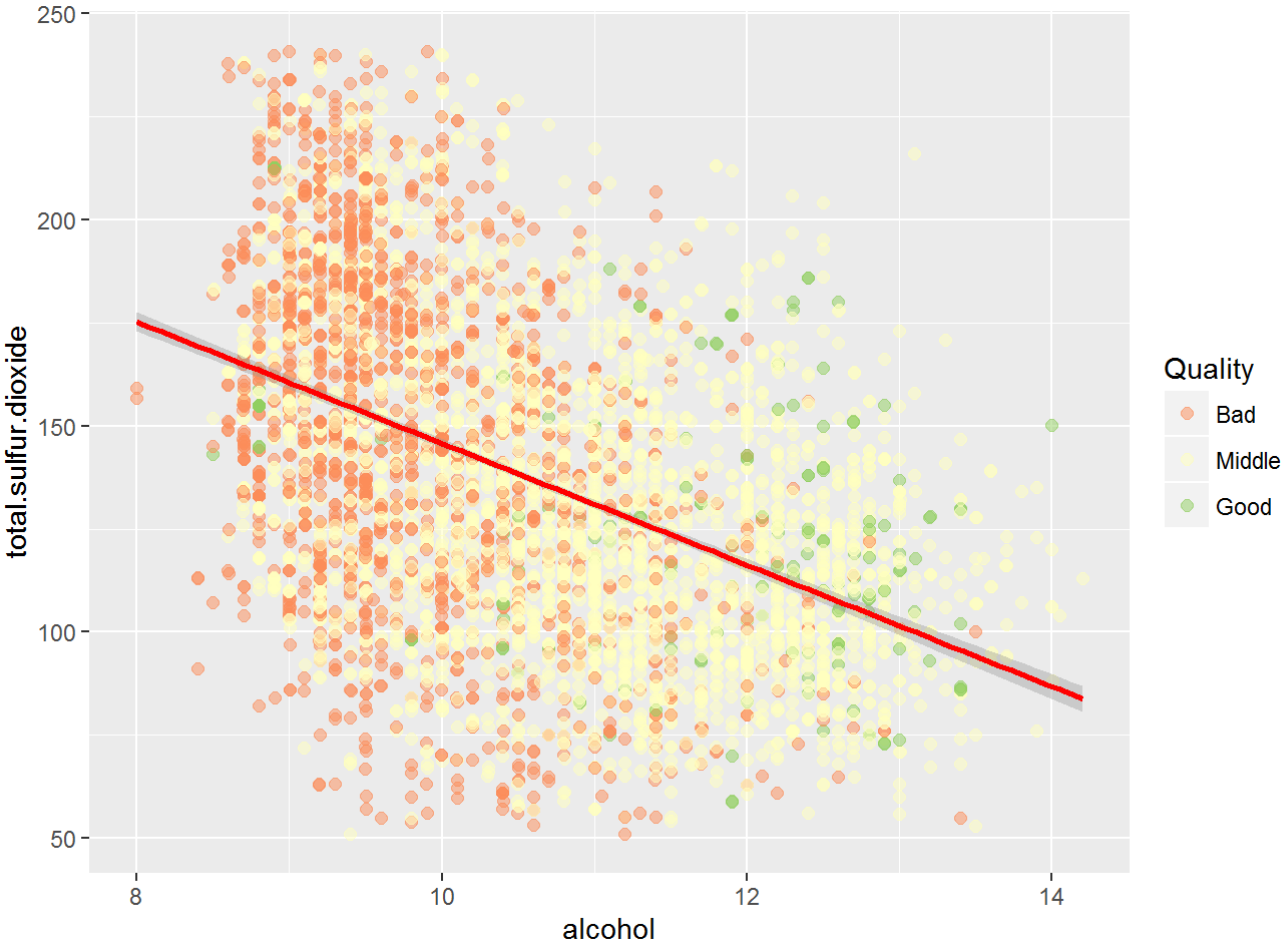
Description of Alcohol vs. Residual Sugar

Bad wines have a large spectrum of residual sugar levels. Good wines appear to have lower levels and with higher levels of alcohol.



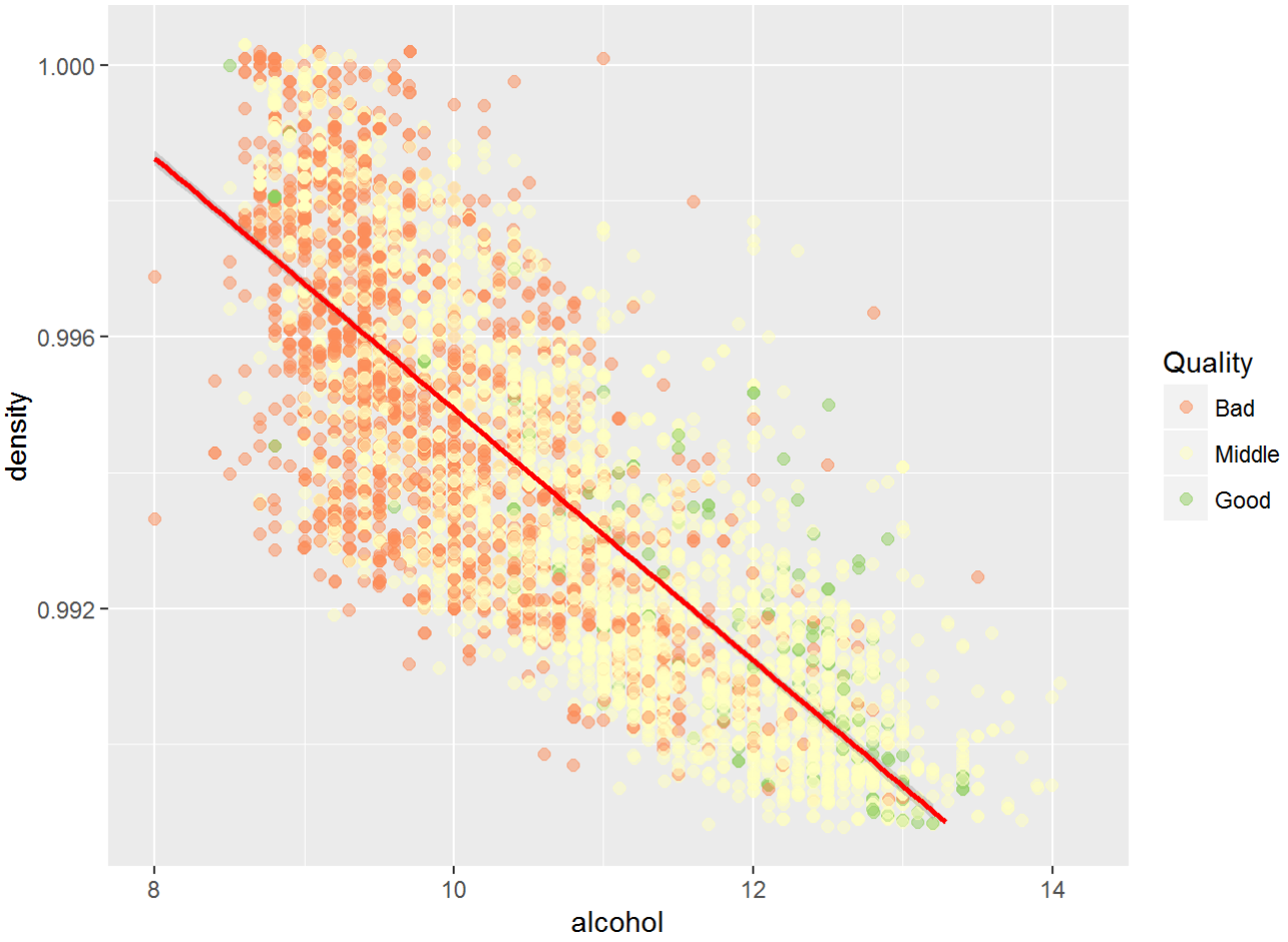
Description of Alcohol vs. Chlorides

Bad wines have a much higher spectrum of chloride levels when compared to good wines.



Description of Alcohol vs. Total SO2

Good wines appear to congregate at lower levels of total SO2. Bad wines appear to have higher levels at total SO2. The difference, however, appears small.



Description of Alcohol vs. Density

Bad wines have a higher density–closer to the density of water at 1,000 kg/m^3. Good wines have more alcohol and lower density.

Multivariate Analysis

Observations

Residual sugar levels increased with density; however, the quality decreased as sugar increased. This would make sense as many lower quality vineyards use sugar as a substitute for rich flavor. Rich flavor is usually derived from a longer fermentation process, premium oak barrels, or closely monitored aging–all things that cost more. Hence why sugar is a cheap subtitute.

Interesting Interactions Between Features

Density and alcohol were inversely proportional. That was evident in plotting them against each other as well as the various components that contirbute to them. The regression lines for residual sugar, chlorides, free sulfur dioxide and total sulfur dioxide all increased as density increased. Conversely, all those same variables decreased as alcohol increased.

Strengths and Limitations of the Model

Geom_smooth was heavily used to outline regression lines of the various relationships. The strength was to highlight the correlation between variables. For example, density vs alcohol has a downward regression line. This would make sense since alcohol has a lower density. Thus, the more alcohol the less dense the wine is overall.

The limitations of the model were especially apparent in total sulfur dioxide vs alcohol. The correlation was -0.449; however, the variance was exceptionally high. Thus, one variable could not confidently be calculated from the other. Calculations of total SO2 are below.

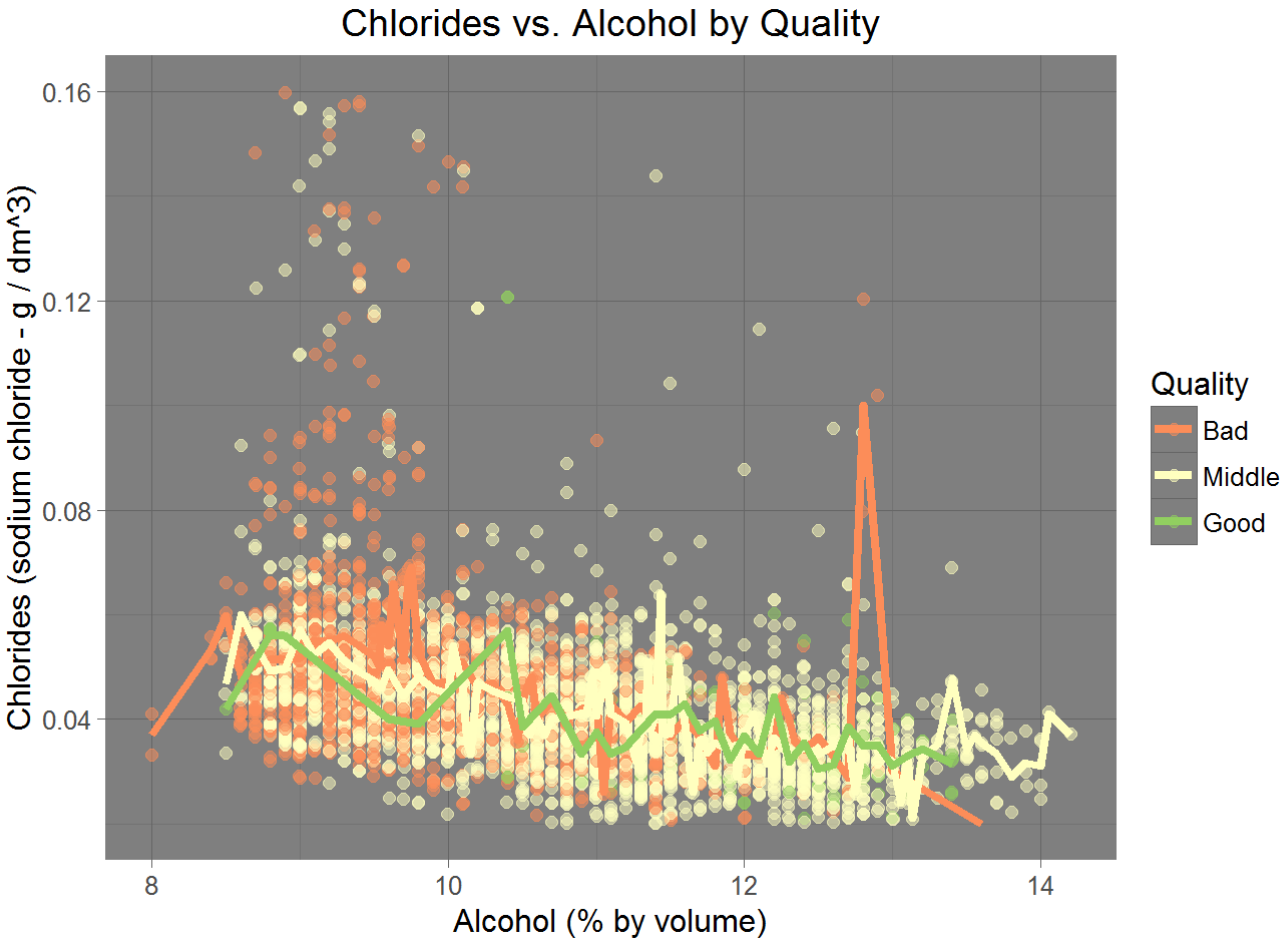
```
## [1] 42.49806
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

```
## [1] 1806.085
```

Final Plots and Summary

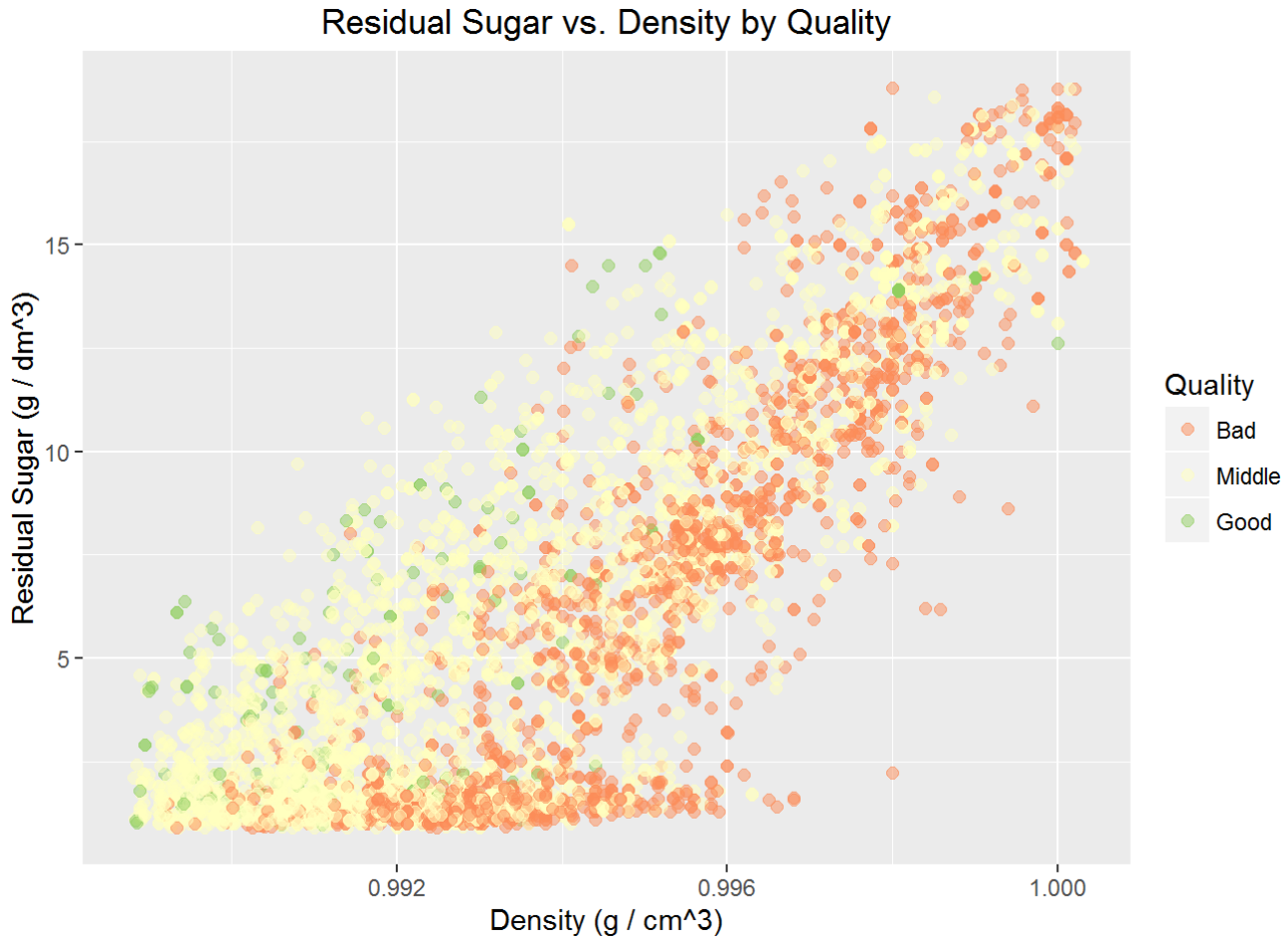
Plot One: Chlorides vs. Alcohol by Quality



Description

We can see in this plot that the chloride variance of bad quality wines is high at low levels of alcohol. Good wines did not appear to have a relatively high chloride variance. High quality wines also appeared to have lower levels of chlorides. Thus, we can reason that high chloride levels are not desirable for high quality wines.

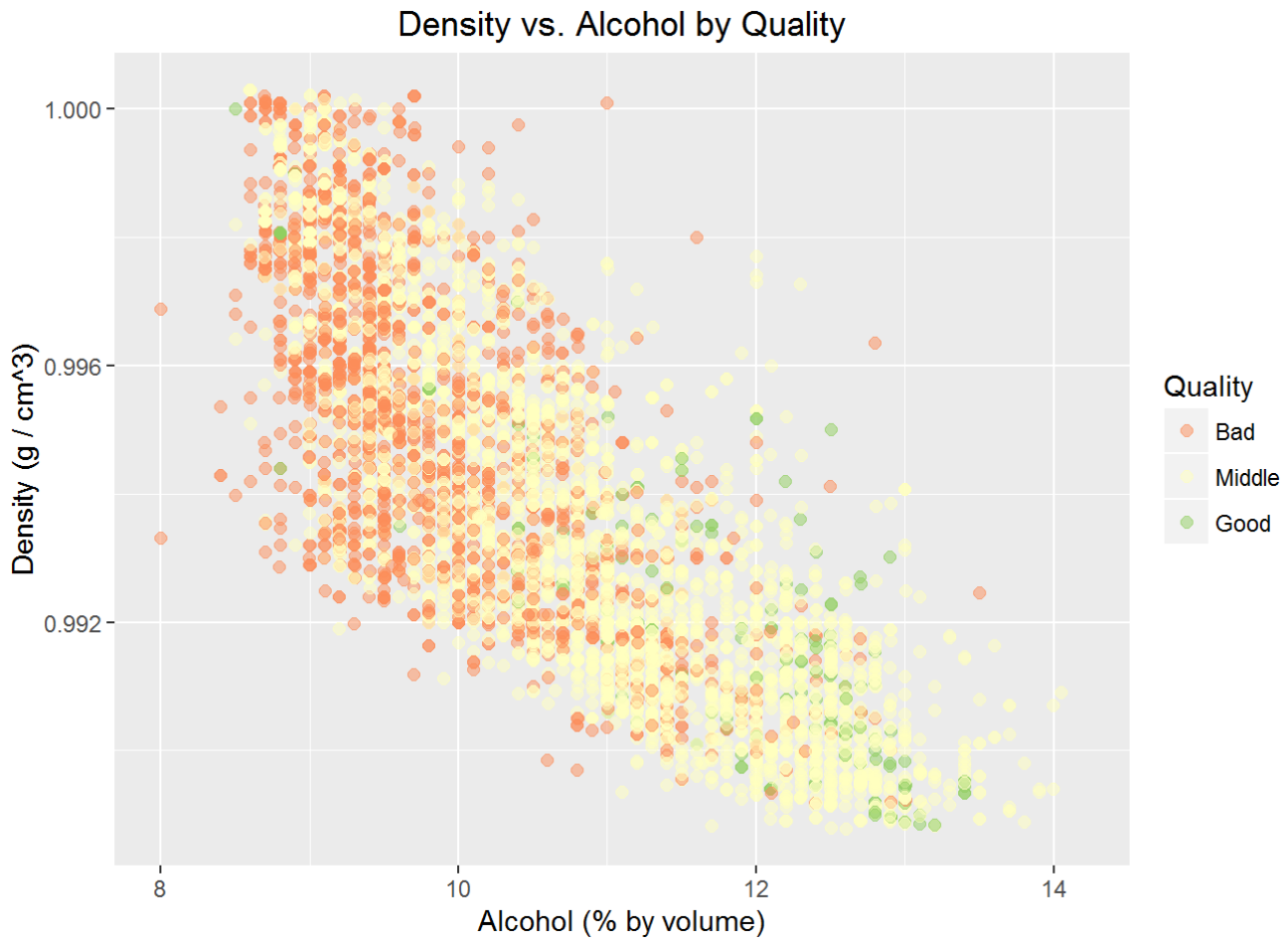
Plot Two: Residual Sugar vs Density by Quality



Description

Residual sugar levels appear to follow a similar trend as chlorides; however, the sensitivity was not as harsh. There are some high quality wines with higher levels of residual sugar. Most high quality wines appear to have lower levels of residual sugar and density. Residual sugar and density are tightly correlated, but did not affect the spectrum of bad wines. That is to say, a bad wine could have low or high levels of residual sugar and density.

Plot Three: Density vs Alcohol by Quality



Description

This graph illustrates a tighter relation with alcohol and density with quality. Also, we can see that density and alcohol have a tigher correlation. It can be observed that as alcohol decreases, density increases couples with a high probability the wine is bad quality. Conversely, good quality wine appears to reside at the high alcohol, lower density end of the spectrum.

Challenges

Correlation in the data was somewhat difficult to determine. Many variables did not appear to have any affect on quality. These include but are not limited to: citric acid, pH, and sulphates. It is important to note that this is not conclusive. Over 28 million liters of wine were produced in 2014[1]. The sample included in this document is roughly 3,700 liters–a significantly small samples size. Thus, if the sample size is increased we may find better correlations.

[1] “World Wine Production by Country”, http://www.wineinstitute.org/files/World_Wine_Production_by_Country_2014_cTradeDataAndAnalysis.pdf (http://www.wineinstitute.org/files/World_Wine_Production_by_Country_2014_cTradeDataAndAnalysis.pdf)

Usage of Analysis

The usage of the analysis in this document can be used boost the quality of wine provided that is the desired outcome. For example, lower levels of residual sugar will have a higher probability of increasing quality.

Reflection

It was interesting to note that the that the quality of a wine was most correlated to density and alcohol levels. Many variables such as fixed acidity and free and total sulfur dioxide levels had some correlation to quality with high levels of variance. It was also noteworthy that volatile acid, citric acid, pH, and sulphates had very little correlation to any other variables in the data set. One would think that the level of acid would correlate to pH since pH is measurement of acidity.

The variable that did seem to affect density and alcohol was residual sugar. The more sugar was present–possibly added–in a wine, the lower the density and higher the alcohol. Both outcomes would lead to a higher quality wine. This correlation makes sense since some lower-cost vineyards substitute a more costly fermentation process with sugar. Also, sugar is used to create alcohol. So, fermentation processes that are longer will have more alcohol and less sugar. More reason to support the theory that better wines are produced over a longer period of time.