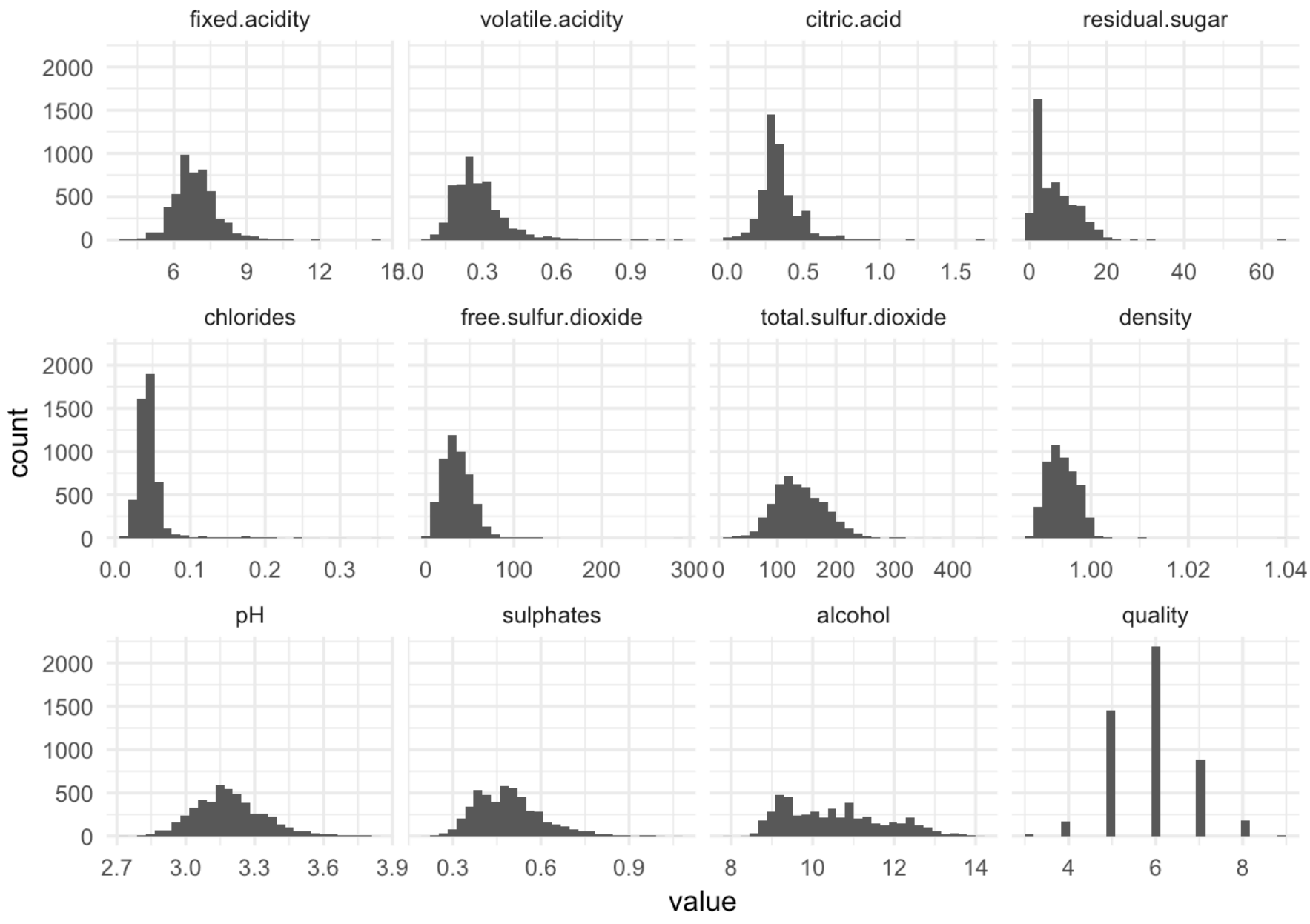
White Wines Exploration

Ben Spinks

> We are exploring the white wine dataset, which contains information on 4898 white wines over 12 variables, including a quality assesment by wine experts.

Univariate Plots Section

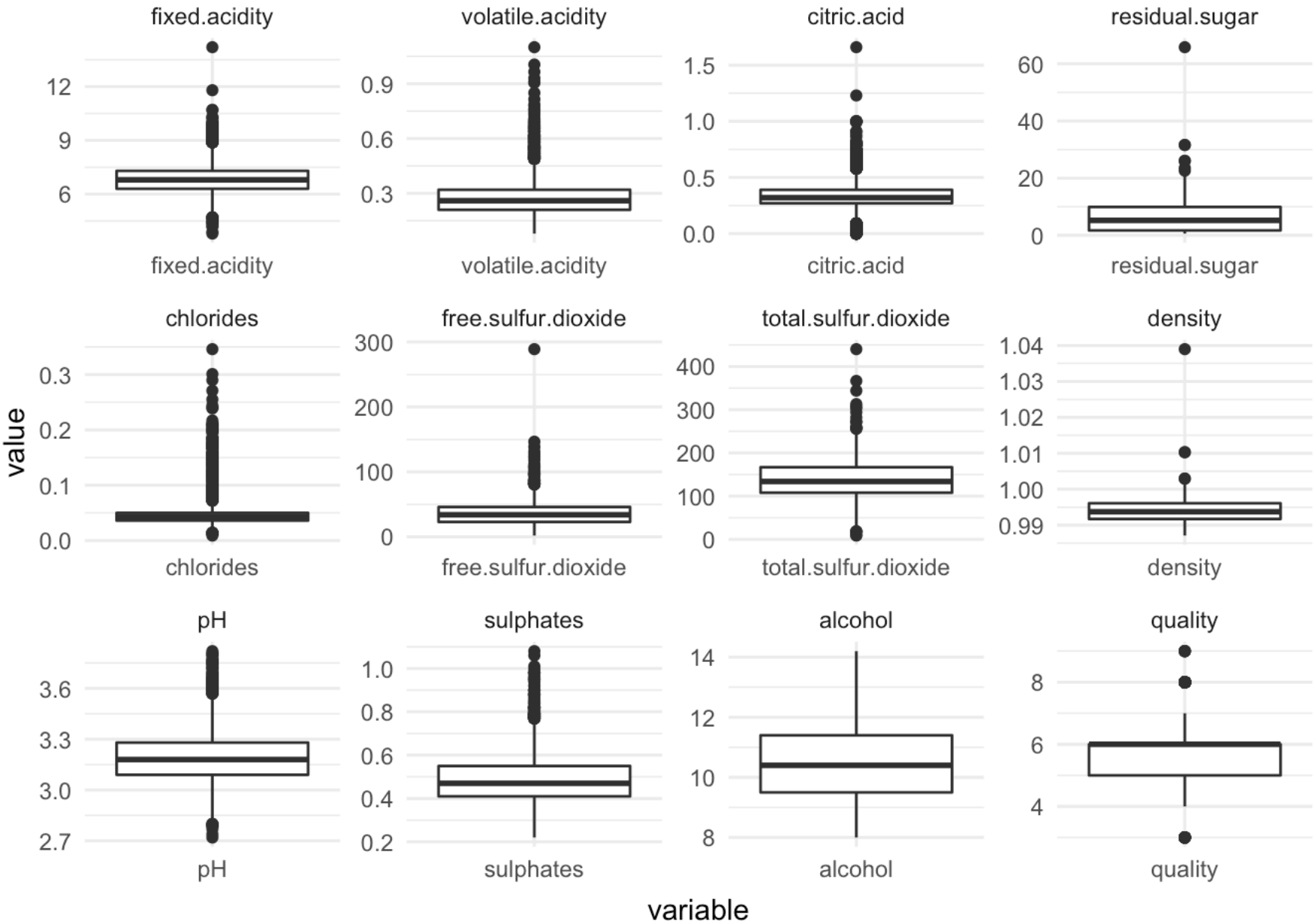
We will first want to take a quick look at the distribution of the data, done so below by plotting histograms of the variables.



There are already some observations to be made,

- Generally most plots show a normal distribution, with perhaps a tendency towards being right skewed
- Some variables are very long tailed, such as chlorides and volatile acidity
- It is clear that the data is not uniform, i.e. there are many more ‘normal’ wines than excellent or poor

To get a better sense of outliers and measures of center we can do the same but with box-plots, as seen below,



Again the long tailed nature of many of the variables can be further seen in variables such as chlorides, volatile acidity and sulphates there are many outliers at the high end of the range while the bulk 75% of the values are tightly packed.

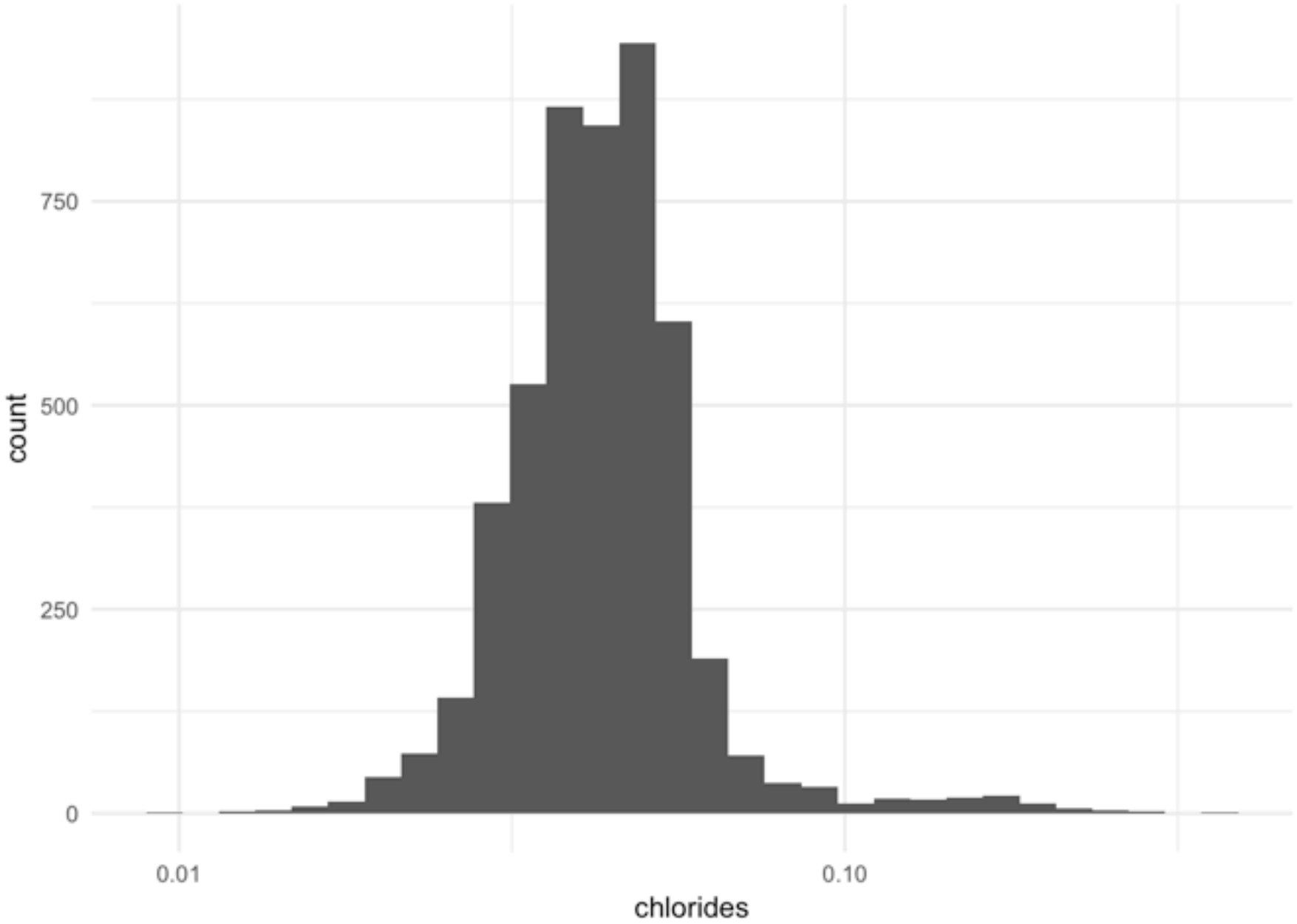
There are also some particularly noticeable outliers, namely in residual sugar, density and free sulfur dioxide, we can check if all these outliers are one wine below,

```
apply(wine[c('residual.sugar', 'density', 'free.sulfur.dioxide')], 2, which.max)
```

```
##      residual.sugar             density free.sulfur.dioxide
##                2782                2782                4746
```

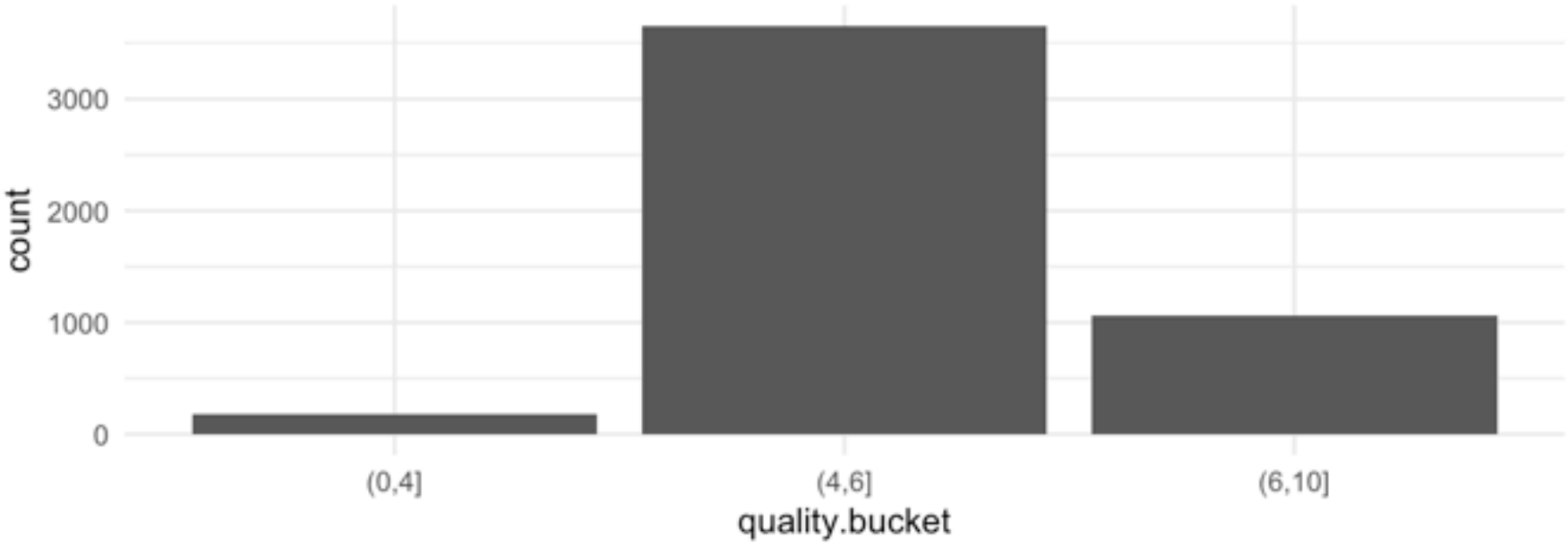
And as we see the residual sugar and density outliers are indeed the same wine sample, and it may be beneficial to exclude this from future graphs.

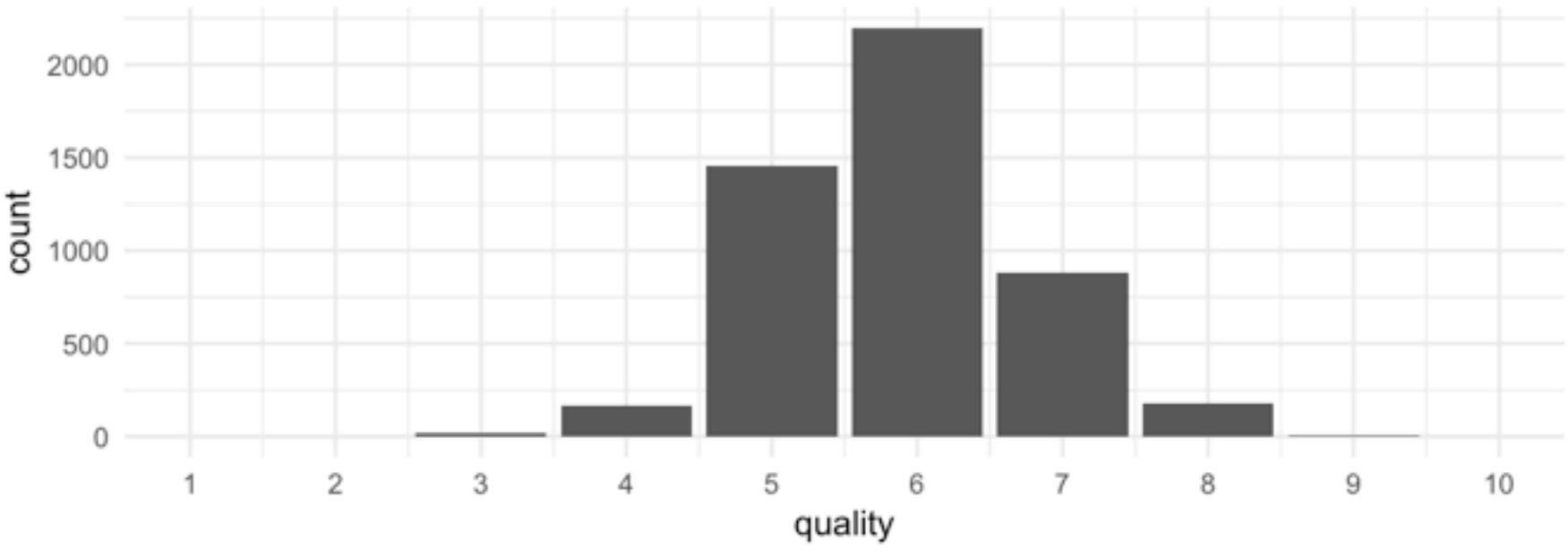
Moving on in regard to the long tailed nature of the chlorides distribution it may be beneficial to apply a log scale to see if the data is log-normally distributed,



And after the transformation the distribution looks much more normal, meaning it may be helpful to apply this transform going forwards.

Another transformation of the data which may be useful going forwards is to create buckets of quality, this will simplify exploring trends relating to quality. Below we create three buckets, 1-4, 5-6, and 7-10. The middle bin being smaller is not significant as the majority of the data will fall into it anyways. We can see how these buckets split the data below.





As expected the middle bucket is much larger than the other two given the normally distributed nature of the quality values.

And as a final note we will look into some general statistics of the variables numerically.

```
##  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      quality.bucket
##  Min.   :3.000   (0,4] : 183
##  1st Qu.:5.000   (4,6] :3655
##  Median :6.000   (6,10]:1060
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

Univariate Analysis

What is the structure of your dataset?

The white wine data-set has data on 4898 wines, with 13 quantitative variables for each.

As seen many of the variables are normally distributed, with a tendency towards being right-skewed.

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

The main feature of interest is quality, as the goal is to explore if certain characteristics of a wine hint towards the evaluation of quality.

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

Just from looking at which distributions are perhaps similar to that of quality it may appear that variables such as alcohol or total sulfur dioxide may be significant.

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

Yes quality bucket was created from the quality variable, splitting it into three groups.

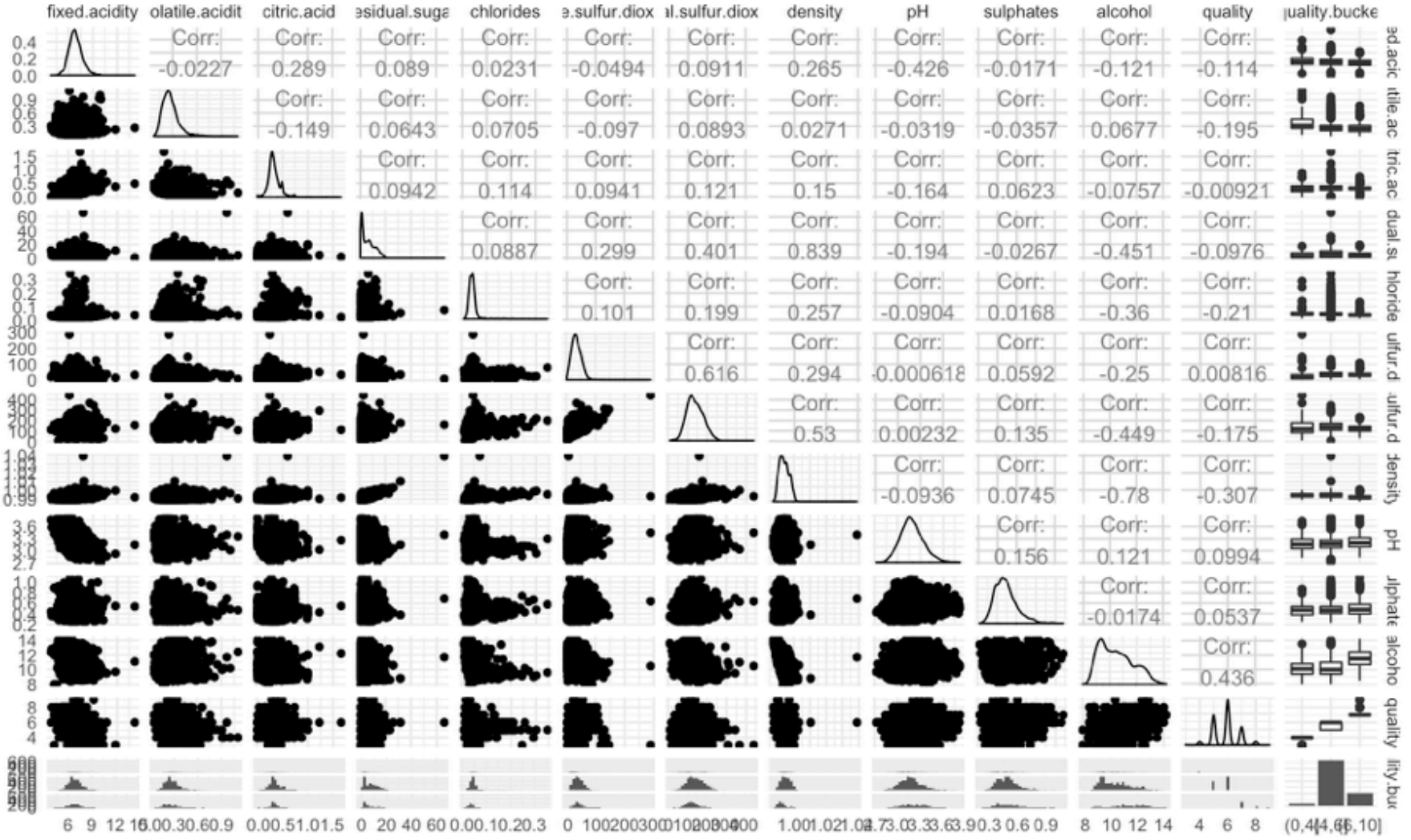
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

As mentioned some distributions were unusual, such as the chloride distribution, when a log transformation was applied however it much more closely resembled a normal distribution, which will be easier to work with going forwards.

There are also a few noticeable outliers in some of the distributions, such as the density/residual sugar outlier, and it's important to be aware of this when plotting these variables or looking for correlations.

Bivariate Plots Section

Using the ggpairs function we can quickly produce a scatter matrix of all variables to look for any distinctly interesting plots or high correlations.



At first glance the most apparently linear plot is that of residual sugar and density, with a correlation of 0.839, apart form that and other density plots it is hard to discern any meaningful relationships.

We can take a more insightful look at the correlation values with a correlation plot as seen below.

Some interesting observations about the correlations:

Quality

- Most strongly correlated with alcohol
- Negative correlations with most other characteristics

Density & Alcohol

- As seen before the correlation between density and residual sugar is one of the strongest relationships in the data-set
- Alcohol is highly negatively correlated with density
- These two variables most often have strong correlations with others in the data-set, which do not often correlate with one another, implying that these two variables are most affected by other characteristics of wine

Other Observations

- Free sulfur dioxide and total sulfur dioxide, as expected, are positively correlated

For ease of reading below we show the top correlations between characteristics and quality,

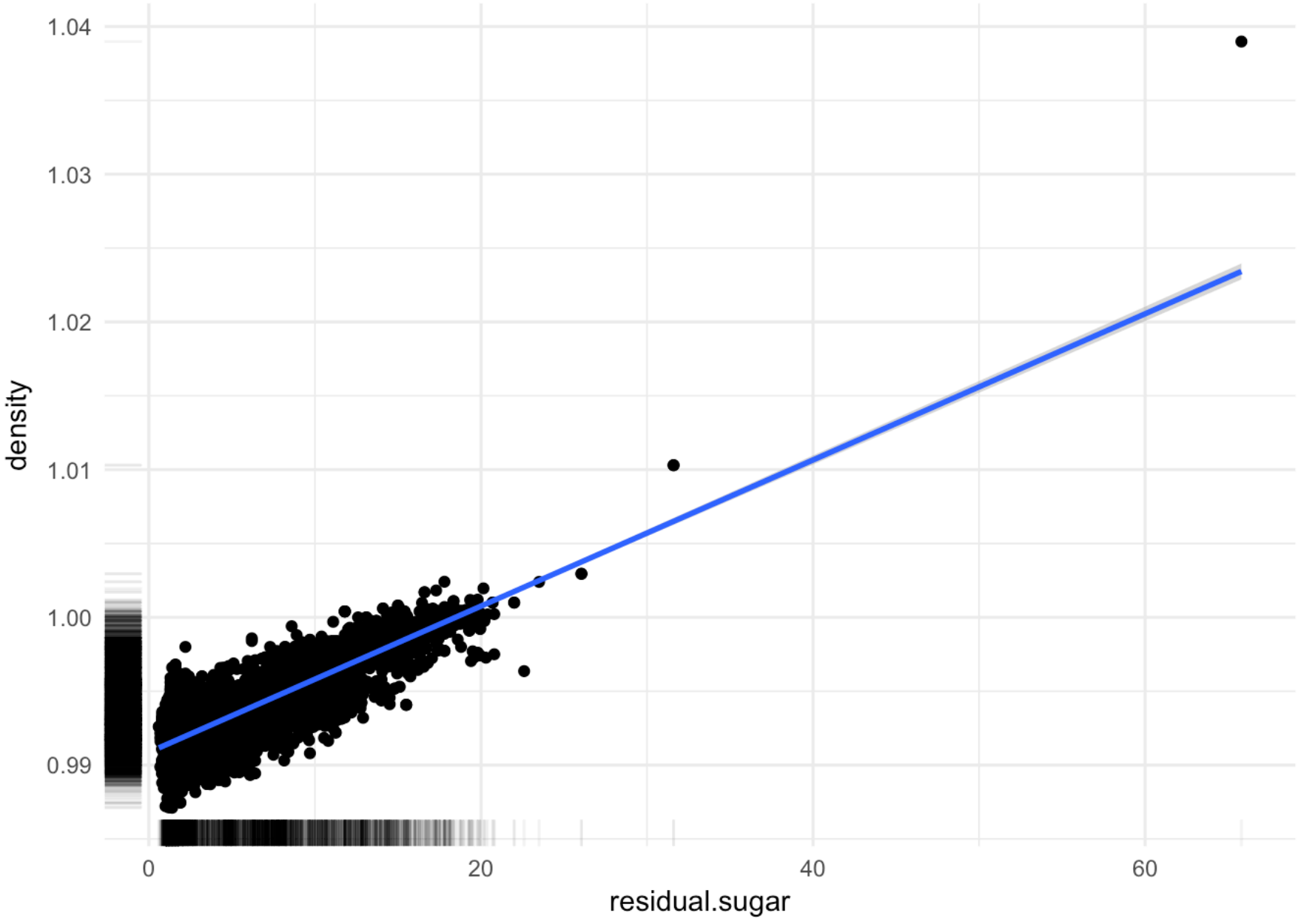
```
##       Var1                Var2 correlation
## 11 quality             alcohol   0.4355747
## 8  quality             density  -0.3071233
## 5  quality           chlorides  -0.2099344
## 2  quality    volatile.acidity  -0.1947230
## 7  quality total.sulfur.dioxide  -0.1747372
## 1  quality       fixed.acidity  -0.1136628
```

And the overall top correlations,

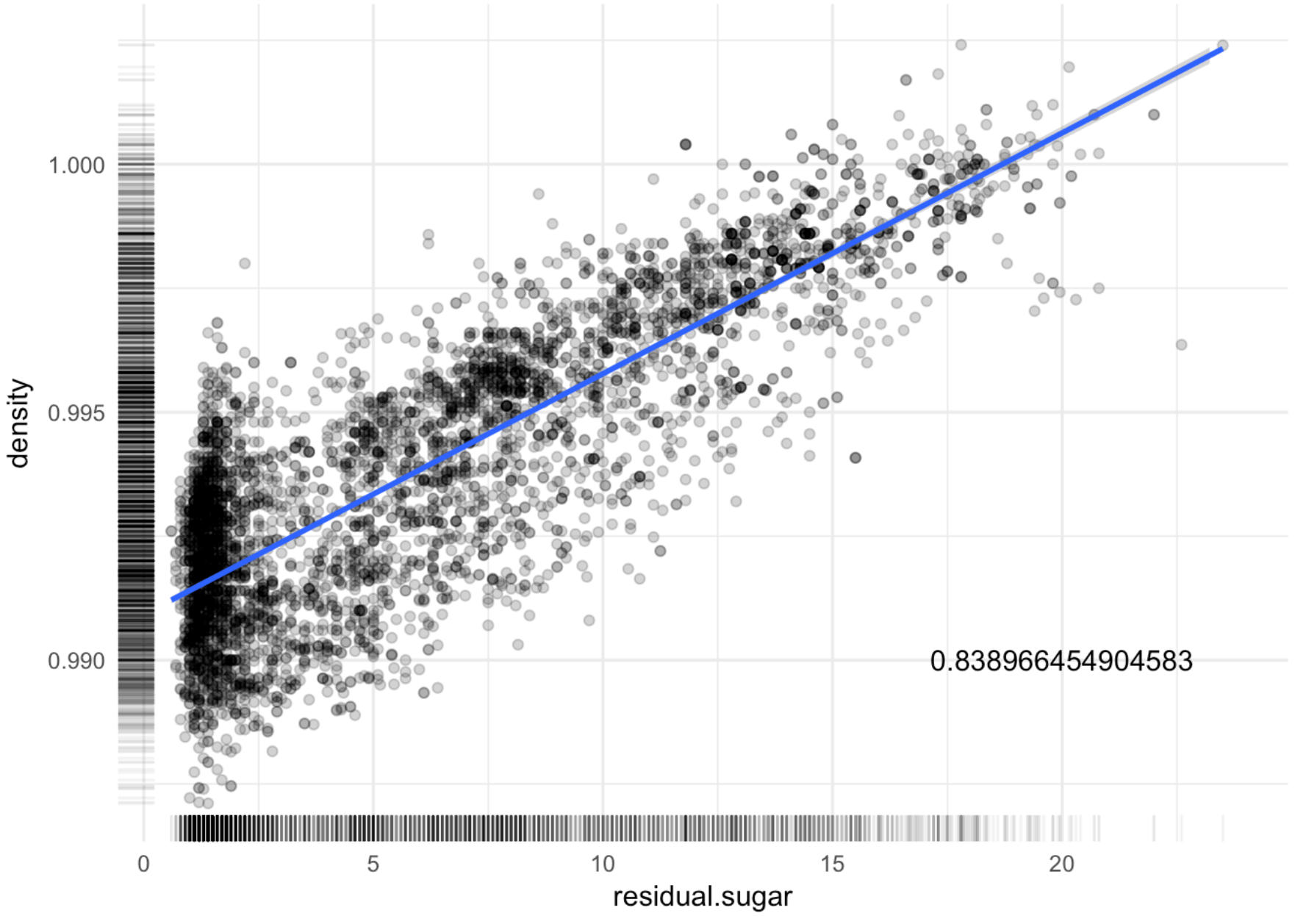
```
##                    Var1                Var2 correlation
## 88         residual.sugar             density   0.8389665
## 128              density             alcohol  -0.7801376
## 78   free.sulfur.dioxide total.sulfur.dioxide   0.6155010
## 91  total.sulfur.dioxide             density   0.5298813
## 124        residual.sugar             alcohol  -0.4506312
## 127 total.sulfur.dioxide             alcohol  -0.4488921
## 143              alcohol             quality   0.4355747
## 97         fixed.acidity                  pH  -0.4258583
## 76         residual.sugar total.sulfur.dioxide   0.4014393
## 125             chlorides             alcohol  -0.3601887
```

Onto plotting, lets first take a closer look at the top correlation from the data-set, density vs. residual sugar.

We'll also add a linear model, representing the correlation, and a geom_rug to get an understanding for how the points are distributed.

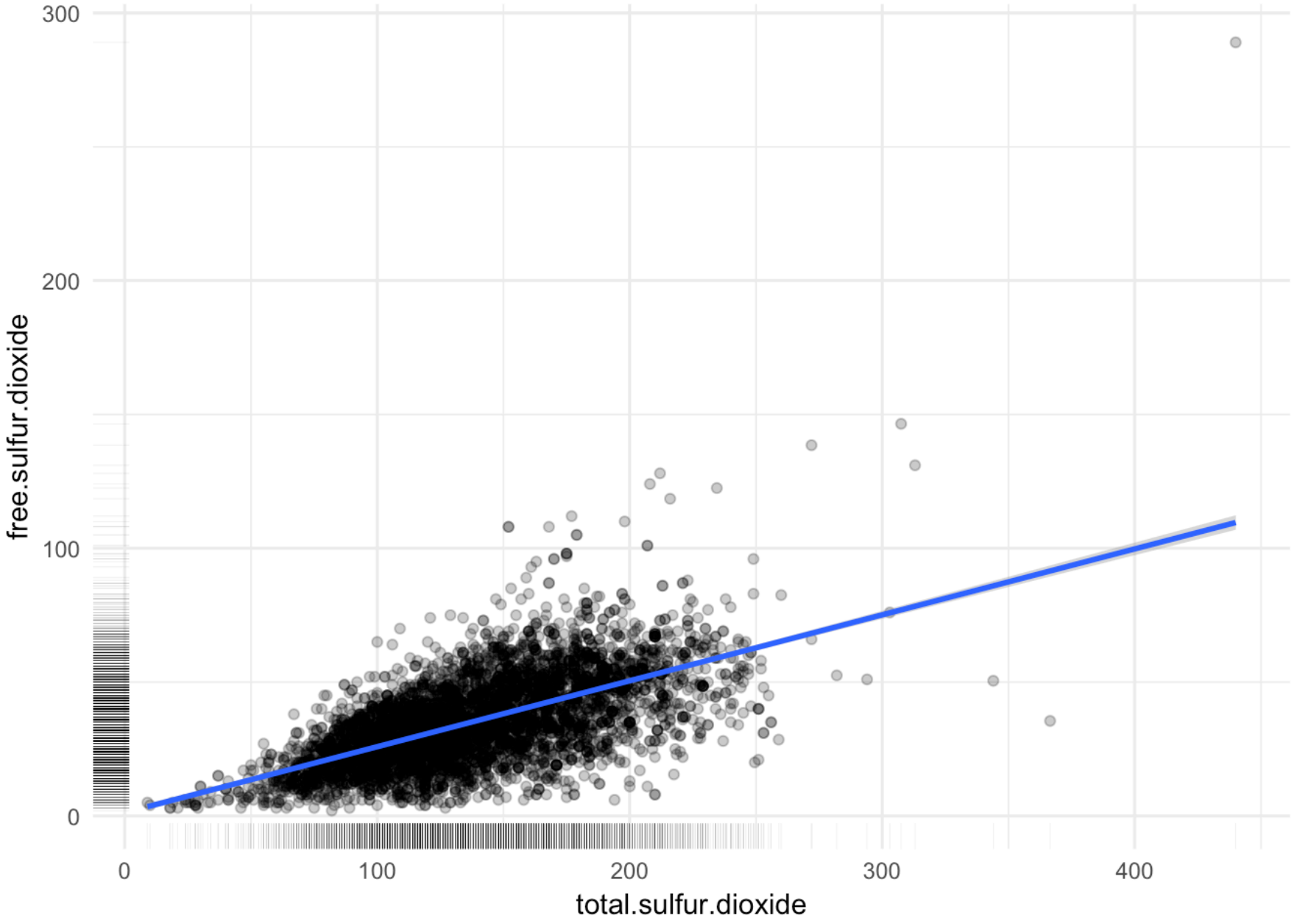


We come across the outlier for both data-sets, far away from the majority of points at approx 1.04 density. We can get a better look at the data by zooming in on the area of interest in the bottom left. Cutting out the top 0.1% of values gives,



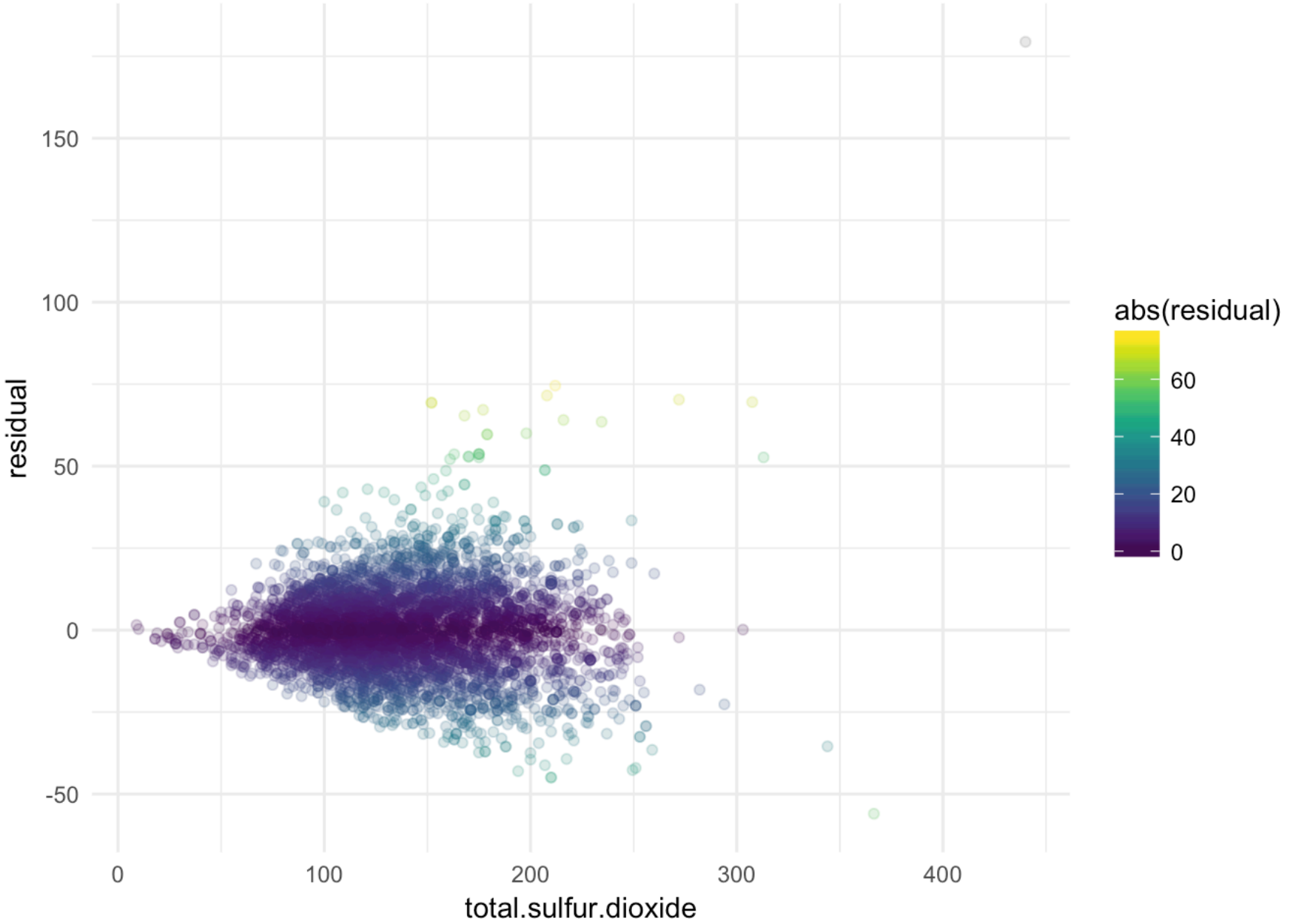
The correlation has been added as an annotation. As seen it is interesting that a majority of points have very low residual sugar, as seen in the earlier uni-variate analysis. From this it is clear that the amount of sugar in a wine has a significant impact in the density of the wine.

Another high correlation relationship is free sulfur dioxide vs. total sulfur dioxide, a relatively obvious relationship.



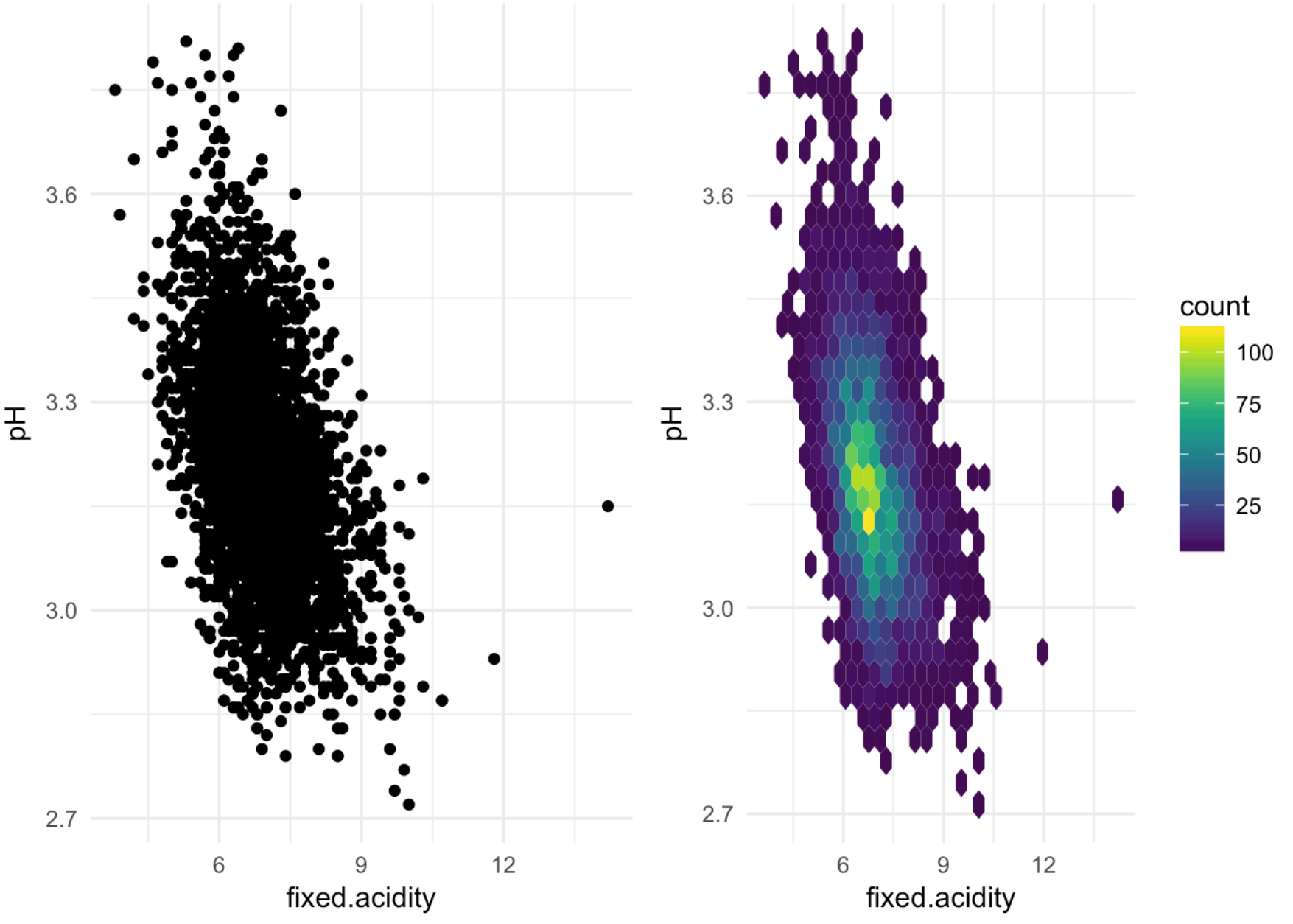
The variation from the linear model in this case shows there are other components in total sulfur dioxide rather than just free sulfur dioxide. It also appears that the prediction becomes less accurate at higher total sulfur dioxide levels.

This can be looked into by plotting residuals of the linear model, as below,



The increasing range of residuals shows the decrease in accuracy of the linear model as total sulfur dioxide increases. This might imply that the other components of total sulfur dioxide also vary increasingly with free sulfur dioxide.

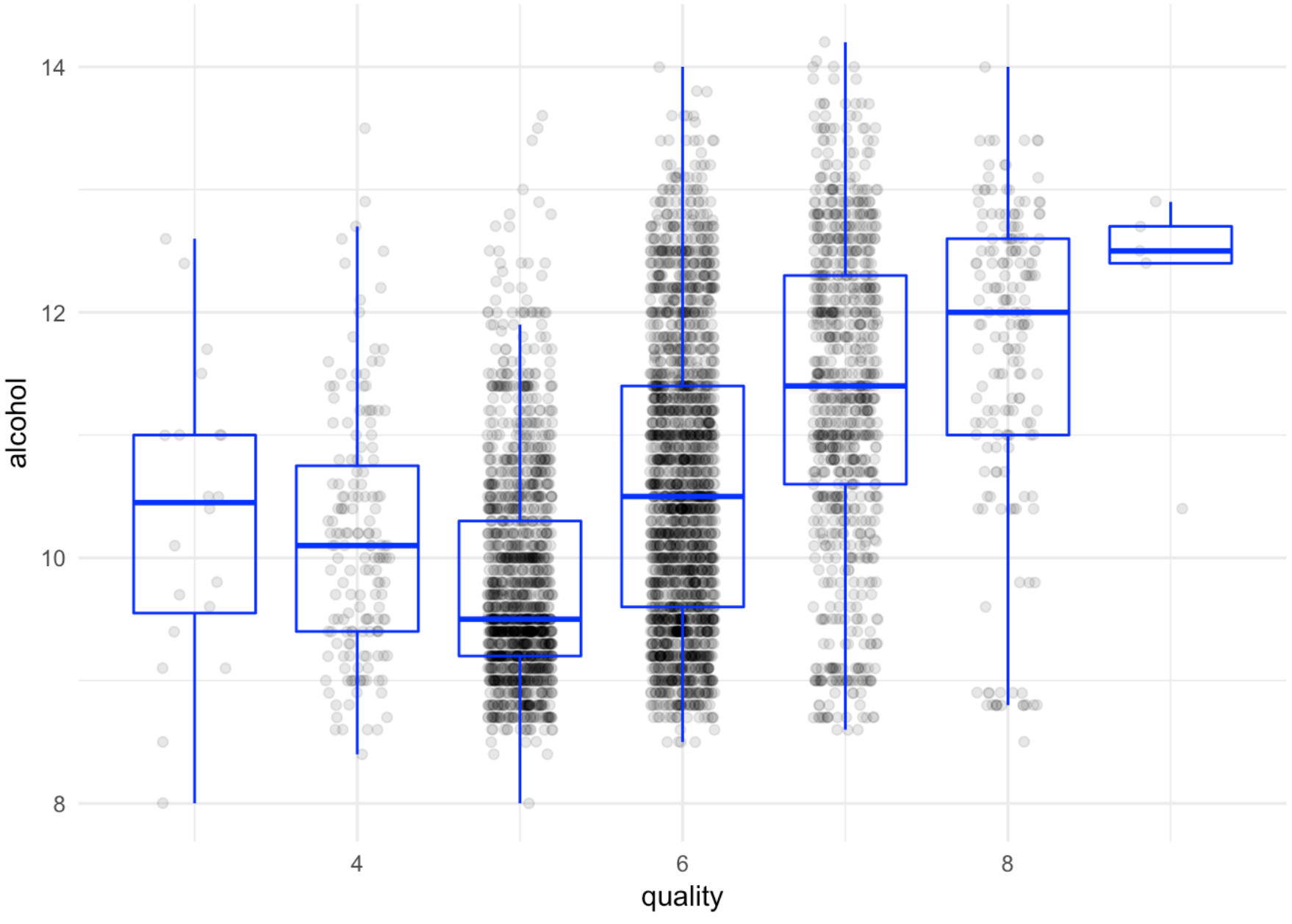
Another apparent relationship would be that of pH and fixed acidity, given that pH is a measure of how acidic a solution is. Below a different method of showing the density of the plot, hex-binning, is used.



There is a fairly strong negative correlation and as seen in the hex-bin plot there is a clear dense area in the center of the plot around 3.16 pH.

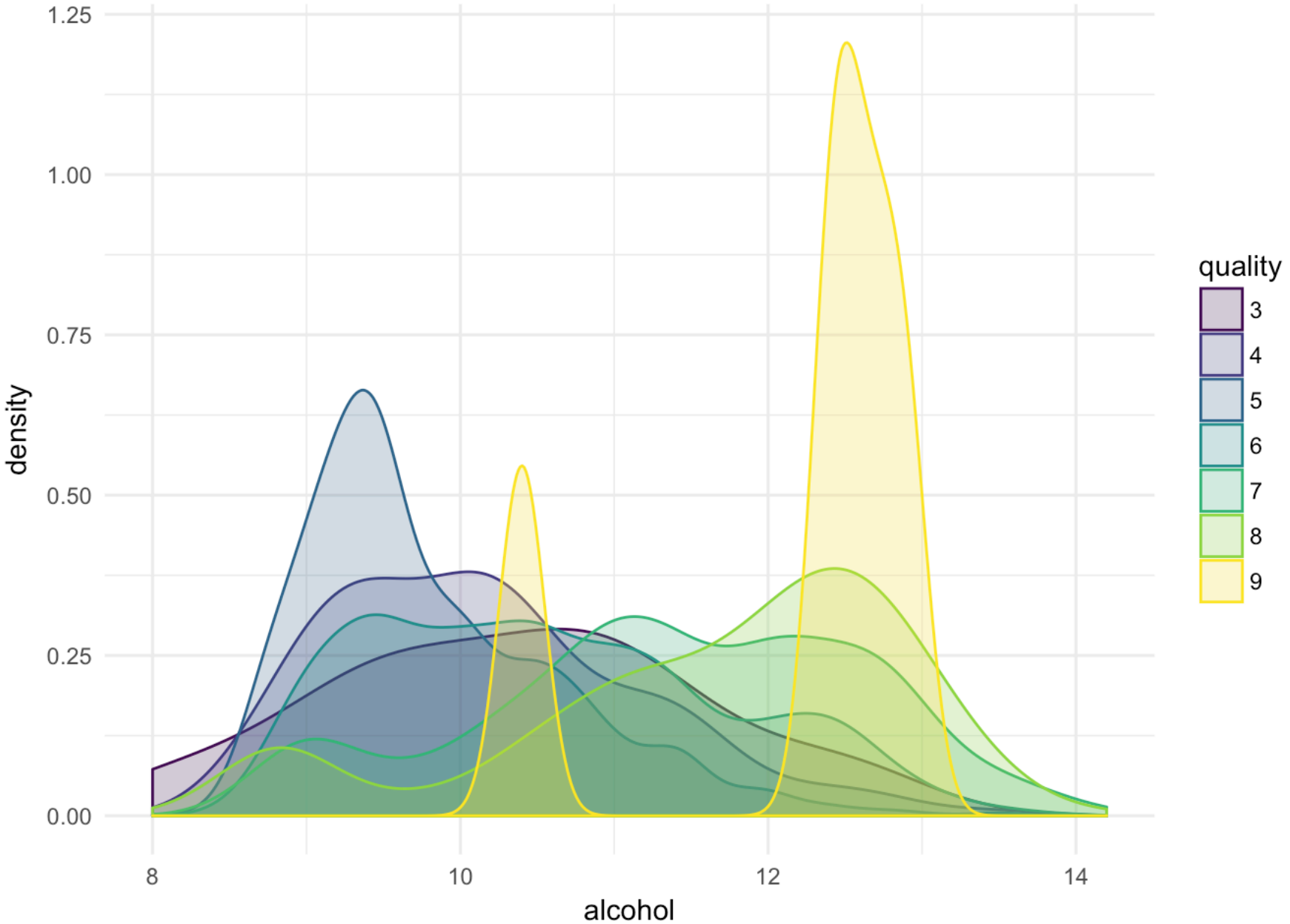
Now looking to the feature variable, quality, we want to examine how some of the more highly correlated features are related.

Firstly a jitter and box-plot of alcohol by quality.



There is a dip in alcohol content as we move from low quality, 3, to mid quality, 5, but after that there is a clear trend of increasing alcohol content with increased quality.

Another way to display this is with a density plot of alcohol contents by quality,

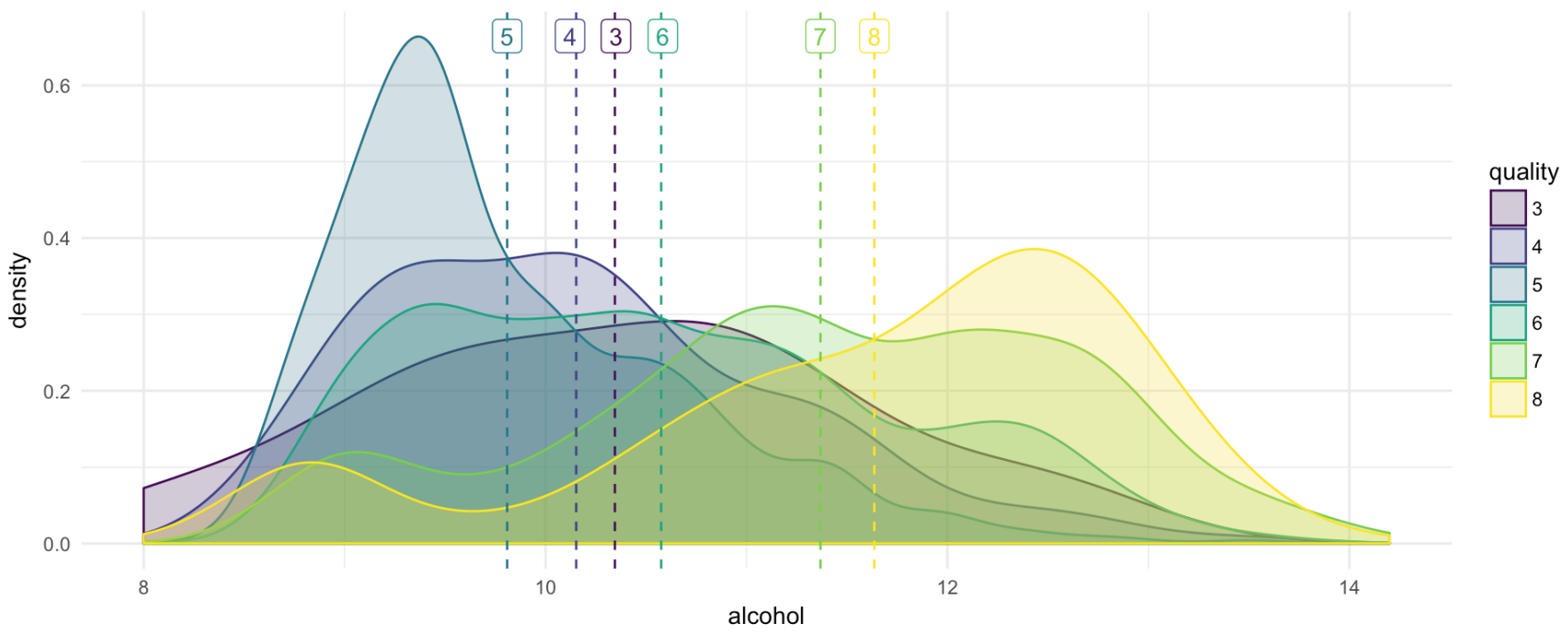


The very high density of quality 9 is somewhat surprising but looking at the number of wines with that quality it becomes apparent the reason,

```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

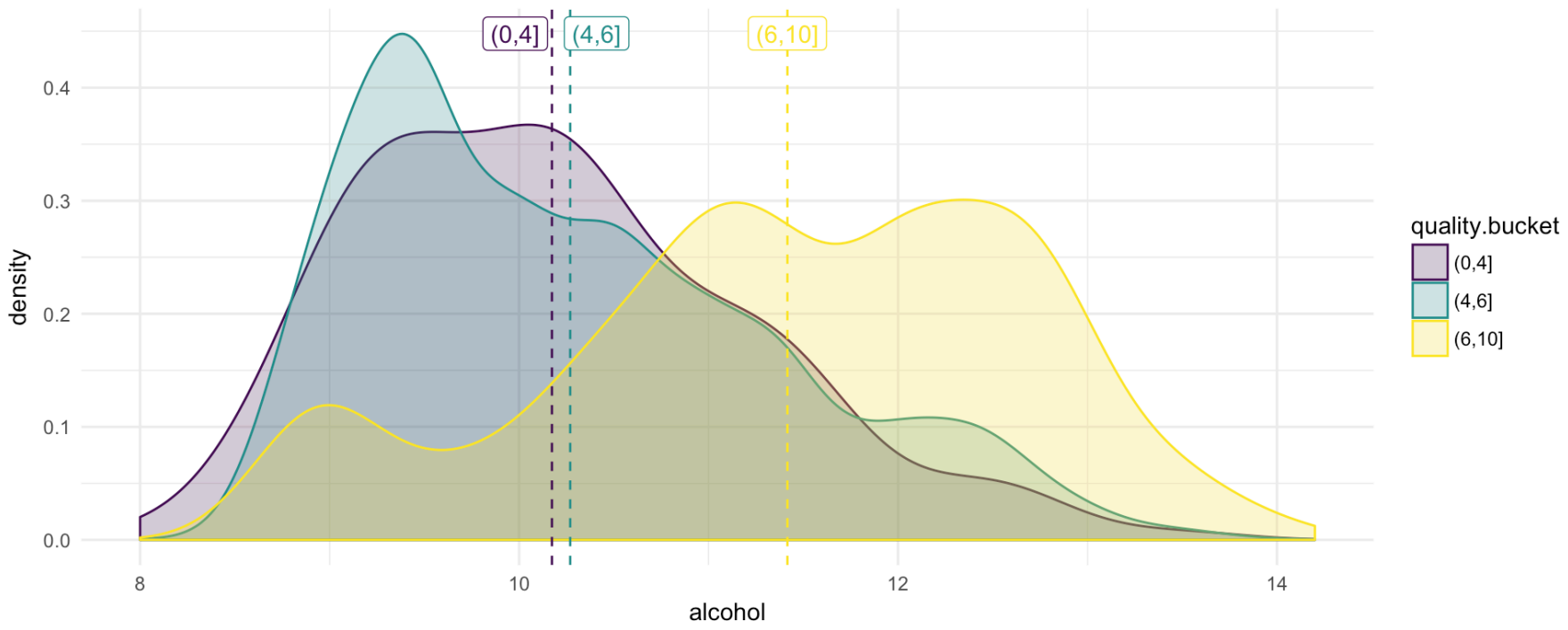
As there are only 5 values it may not be particularly insightful to use quality 9 in the further plots as it is not as representative as other quality ranges.

We can also plot a representation of the mean alcohol value for each quality to better see the trends in distribution. Plotting this, and removing quality 9 we get the following plot,

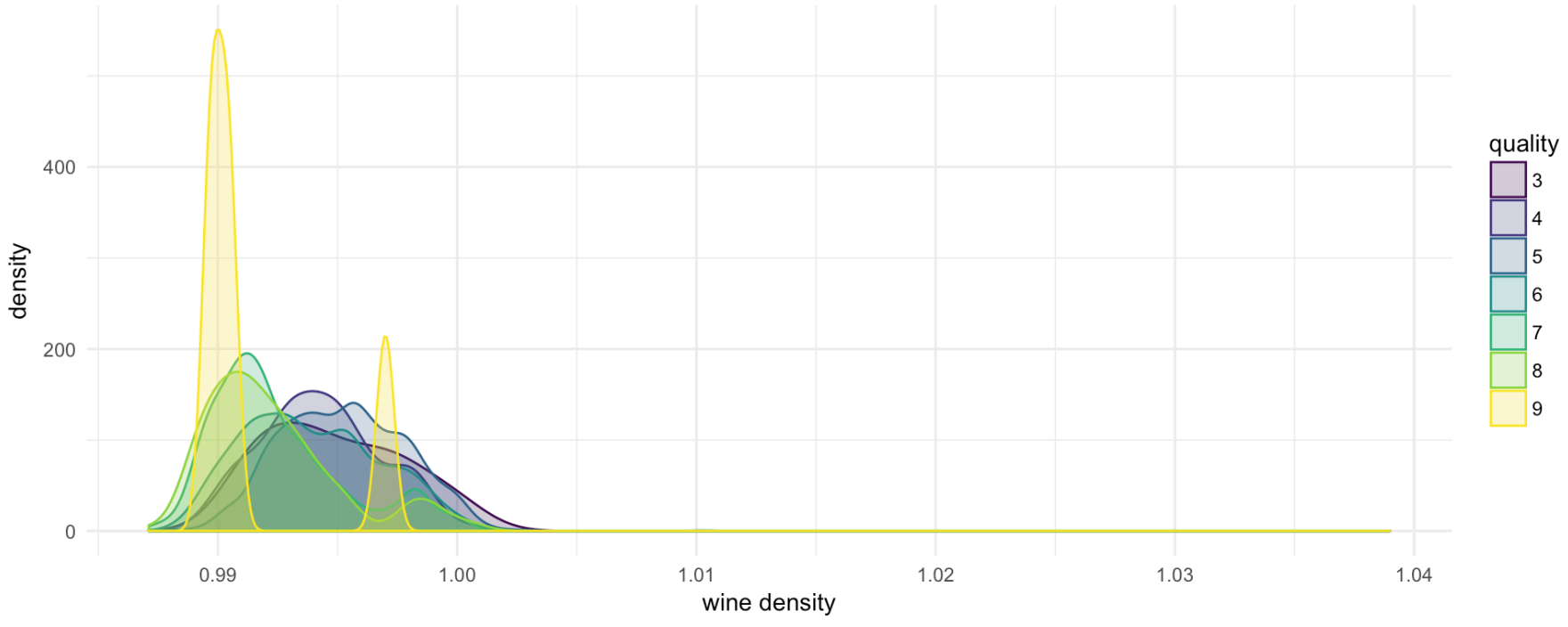


As can be seen there is a observable rightward shift of the distributions as quality increases. Other than this the high density of mid, 5, quality wines at a relatively low alcohol level is interesting.

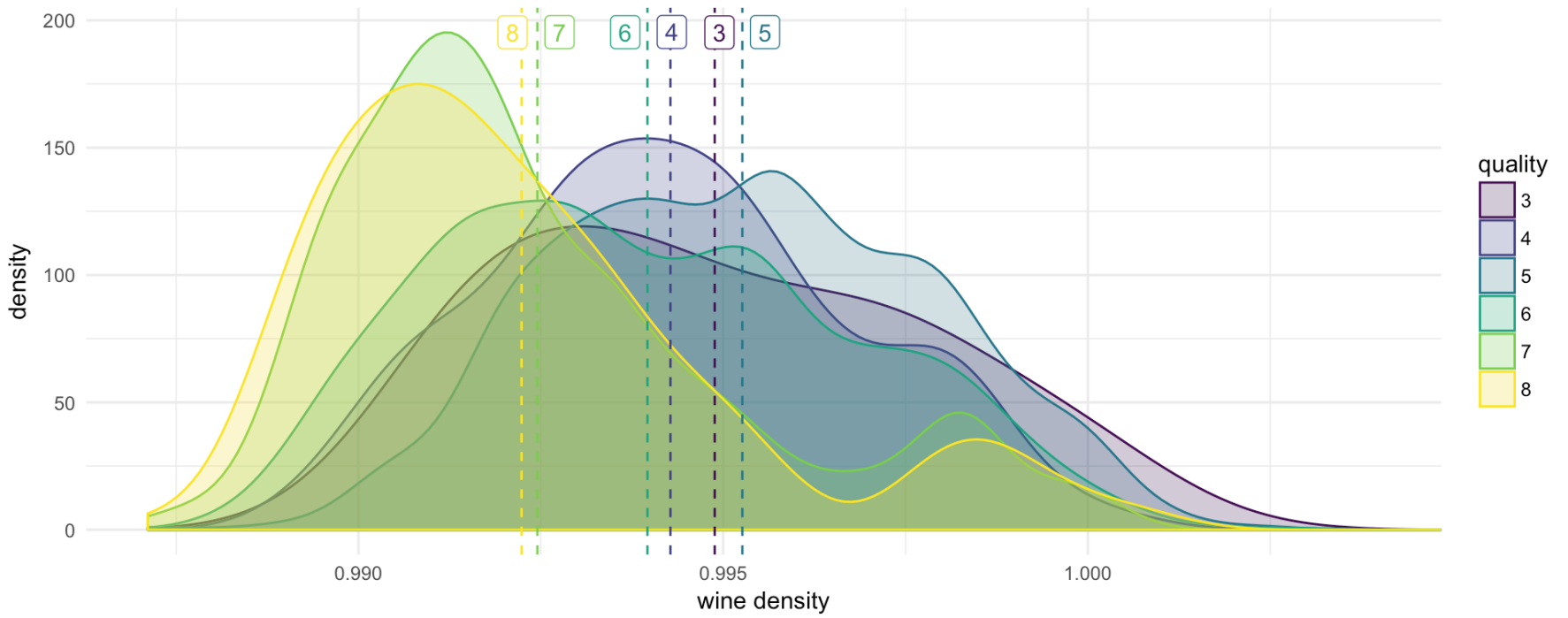
Looking at the same plot but now using the quality buckets created earlier the overall trend is more apparent, and the mean values reflect this.



Moving to the next highest correlation with wine, a similar plot can be made with wine density.



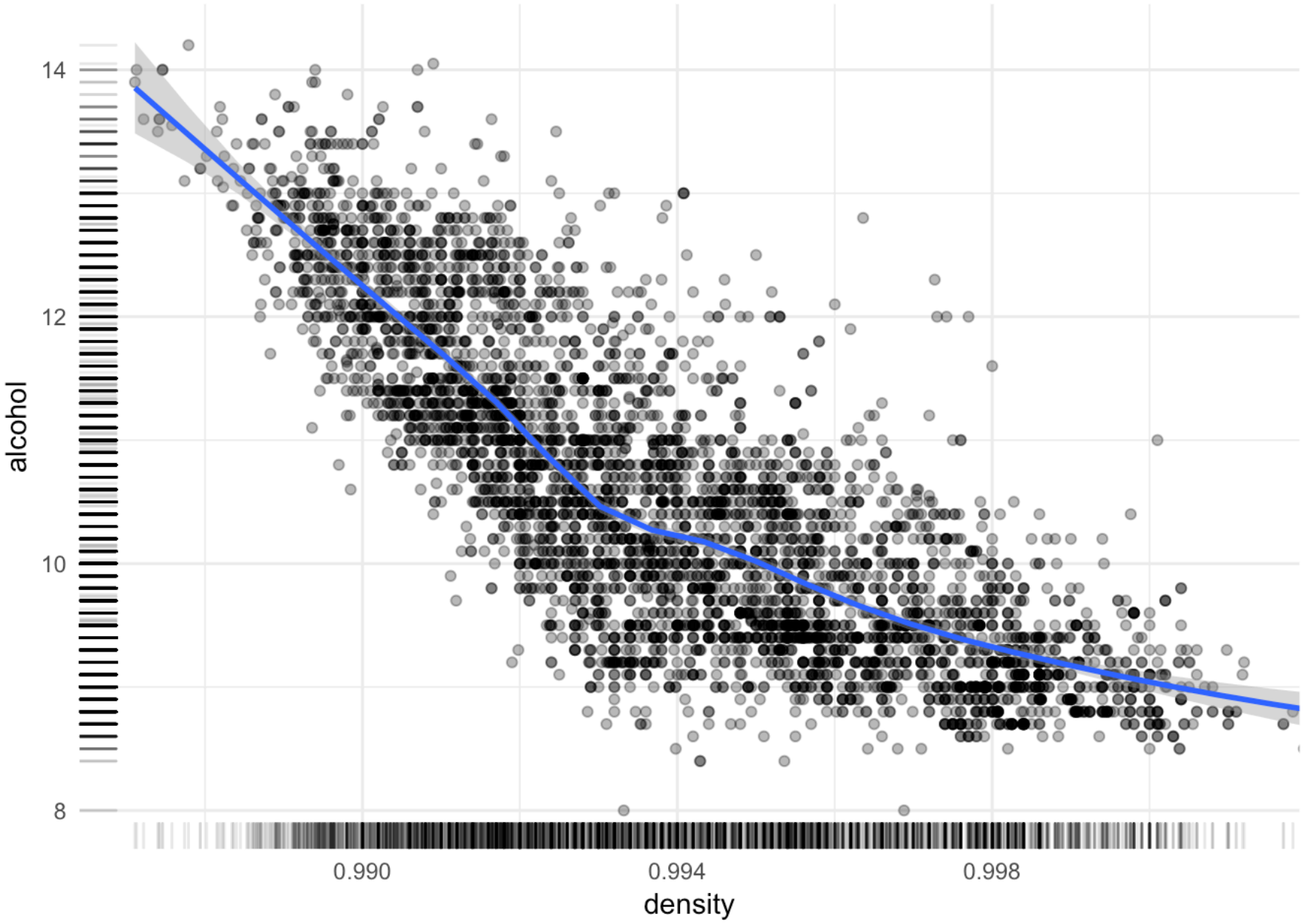
This plot, like the alcohol content is also heavily skewed by the few quality 9 wines, and the large outlier of wine density. Plotting again without quality 9 and trimming the density outlier gives us a better look at the trend.



As seen, as the negative correlation implies there is a leftwards shift of distributions as quality increases. It is again interesting that quality 5 wines go against the trend, with a higher mean density than quality 3 or 4 wines.

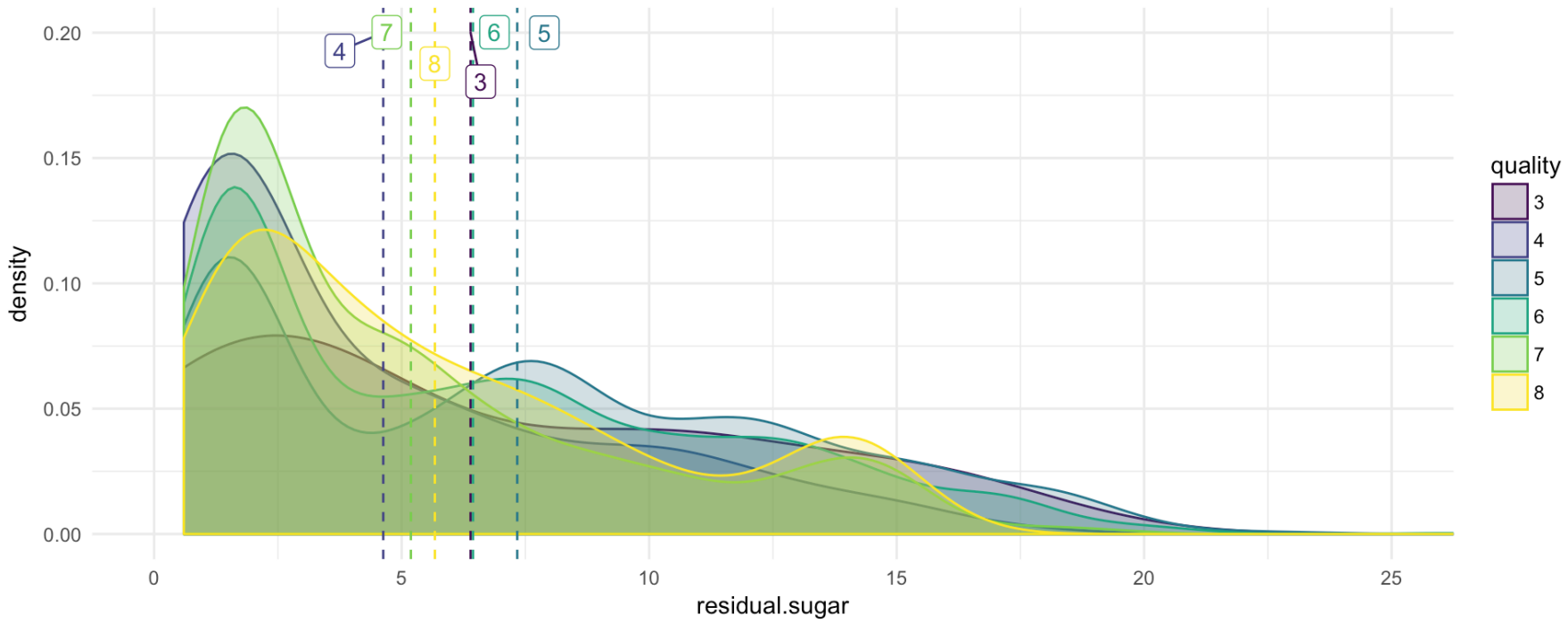
It may be interesting to see how these two variables respond to one another, given that they are the two top correlated variables in respect to quality.

They are plotted below, again omitting the large outlier of density.

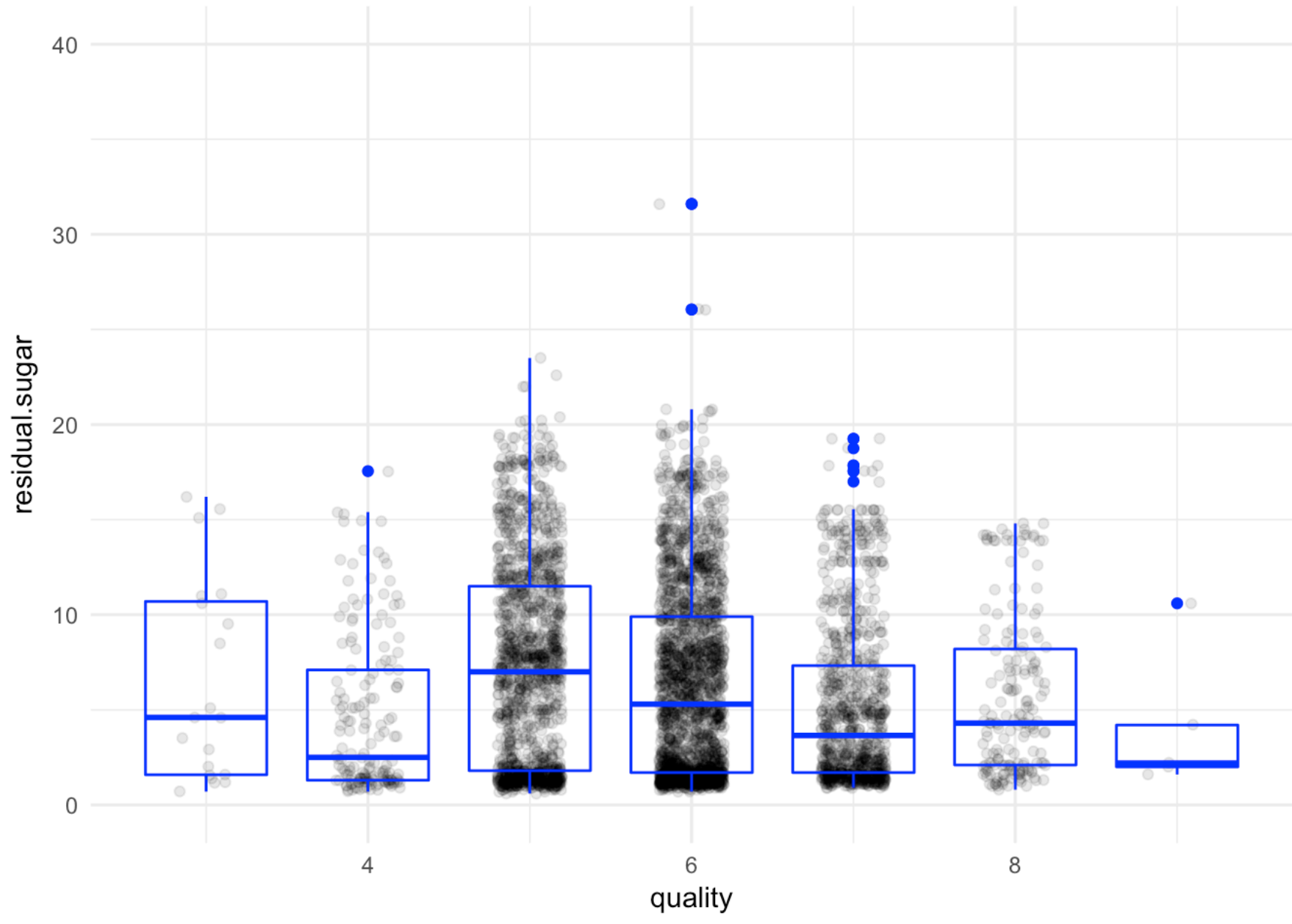


The negative correlation is seen and with a geom smooth the more dense curve through the data can be seen, which looks like it could accurately represent the relationship.

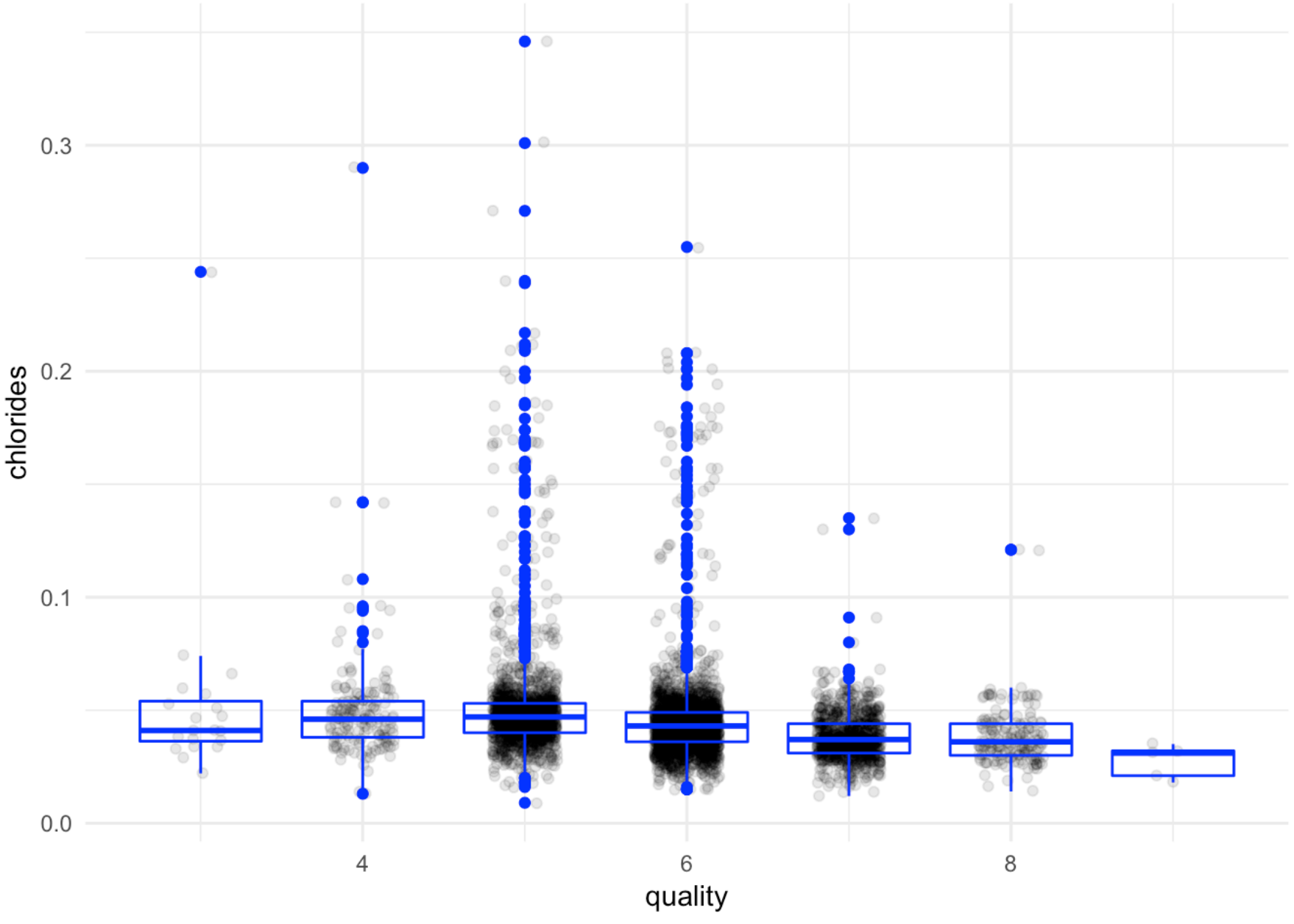
As density is so strongly correlated to residual sugar, and there was a visible trend between density and quality it may be interesting to see if a similar effect can be found between residual sugar and quality.



As seen there is no clear trend, and the random order of the means reflects as such. This reinforces that residual sugar is not a significant determining factor on its own, even given the major impact it has on density. This can be seen in a different style below,

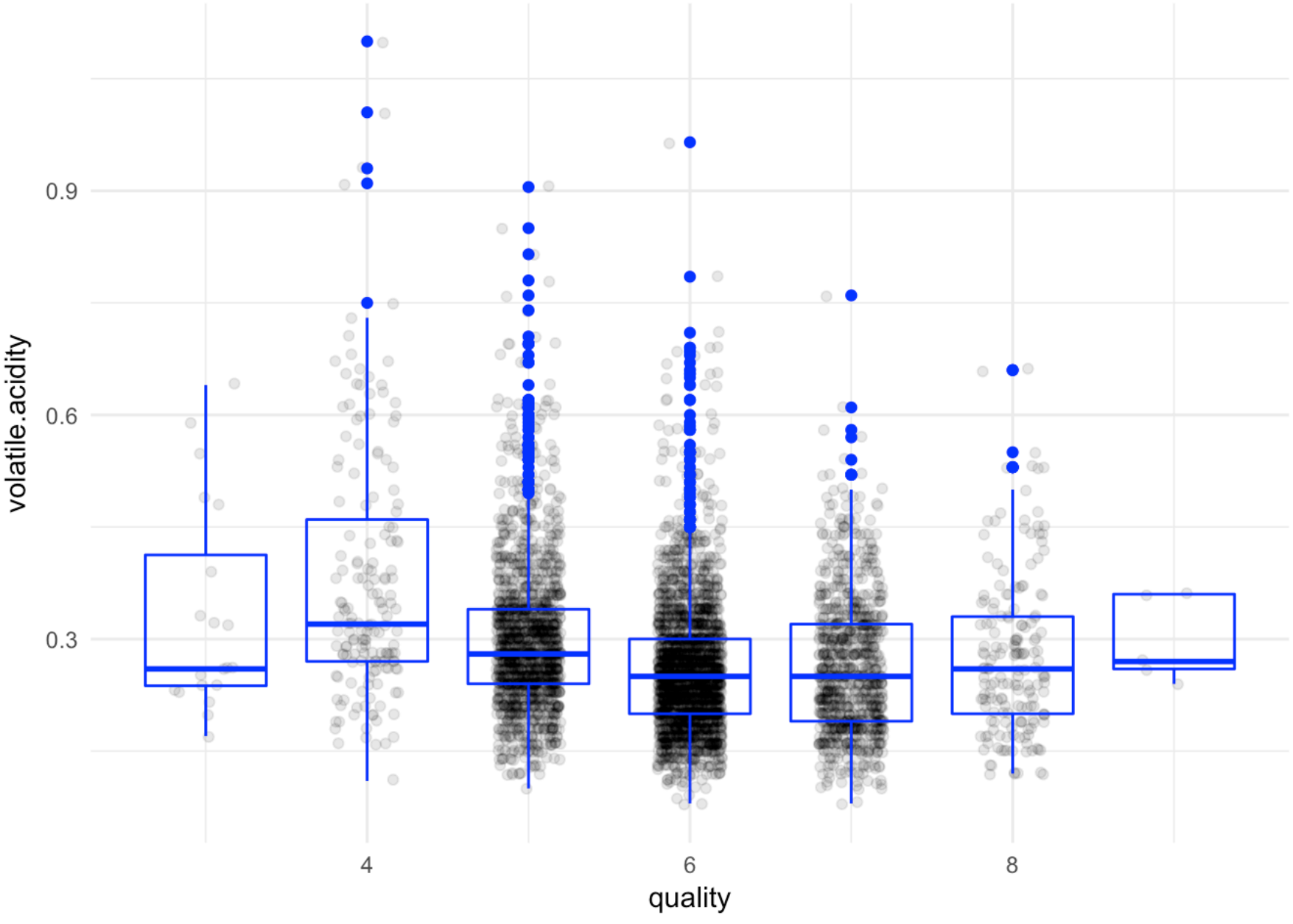


After alcohol and density chlorides are the next most correlated variable to quality, the negative correlation can be observed below,



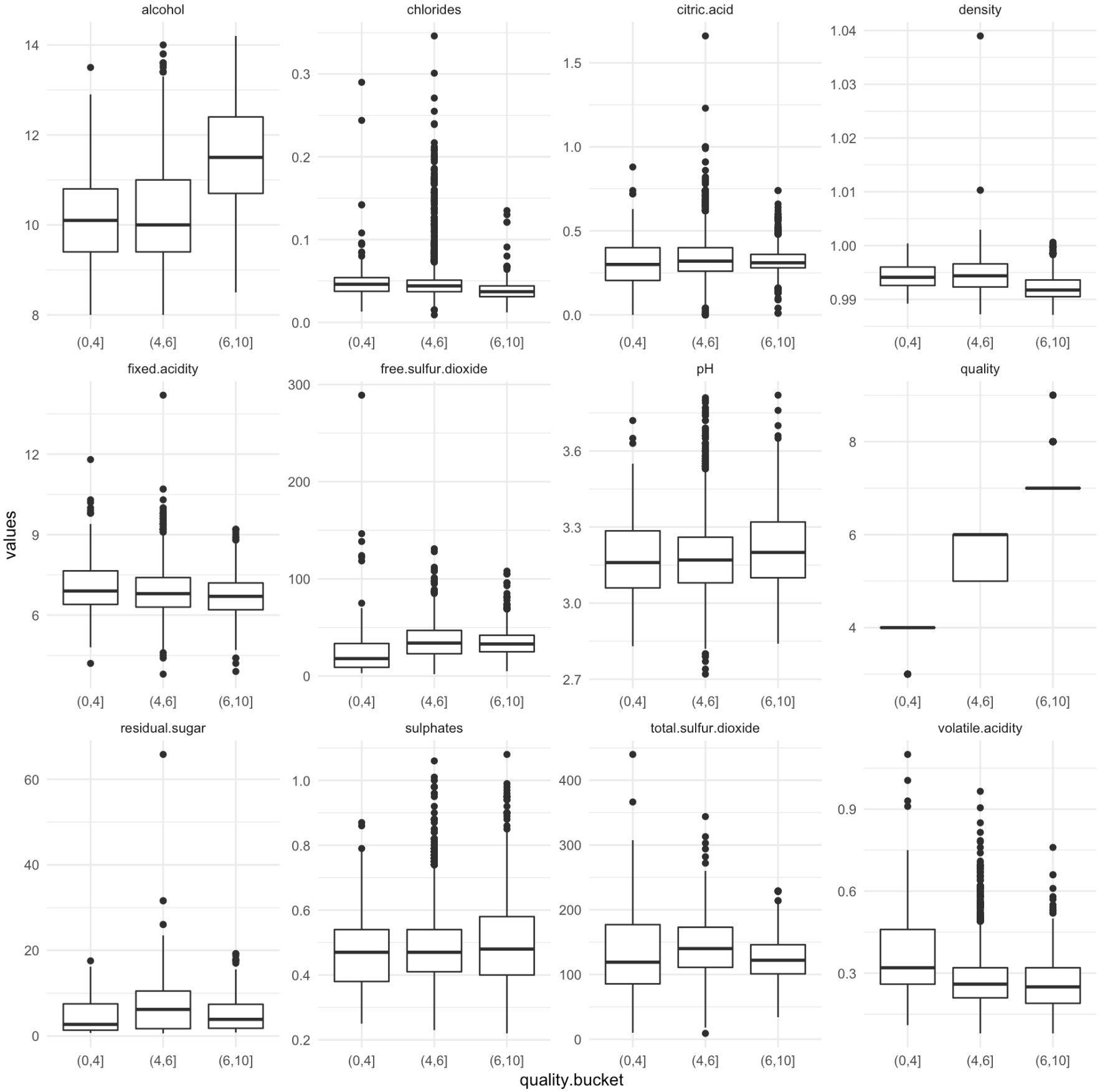
There is a slight observable decrease in chloride content with quality. Also interesting is that the mid range wines have a much wider spread of outliers, which may just be due to there being many more values in that range.

Moving on to volatile acidity,



Here the correlation is very minimal and there is no significant trend in the data.

Now we will perform similar plots but by quality bucket, to see if there are any more clear trends after the data is grouped,



- Firstly the volatile acidity correlation found in the linear model is more clear to see when the data is binned, as the downwards trend is more visible.
- Another promising graph is of fixed acidity, there is a more apparent trend than would be suggested by the correlation of -0.1136628. This implies that it may be useful for determining the quality of a wine.
- Other than that the previously explored correlations are all relatively apparent in these plots

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 data-set?

- Quality increases with increased alcohol content, shown strongly in the binned graph
- Quality increases with wine density

- However other variables that are highly correlated to density such as residual sugar have no observable trend in relation to quality
- With lower chloride values quality tends to increase

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

- There is a strong correlation between density and residual sugar
- A linear model is a good prediction for free and total sulfur dioxide, however its accuracy decreases as total sulfur dioxide increases, as seen in the residuals plot
- As expected there is a negative correlation between pH and fixed acidity
- There appears to be a underlying curve in the relationship between density and alcohol content

What was the strongest relationship you found?

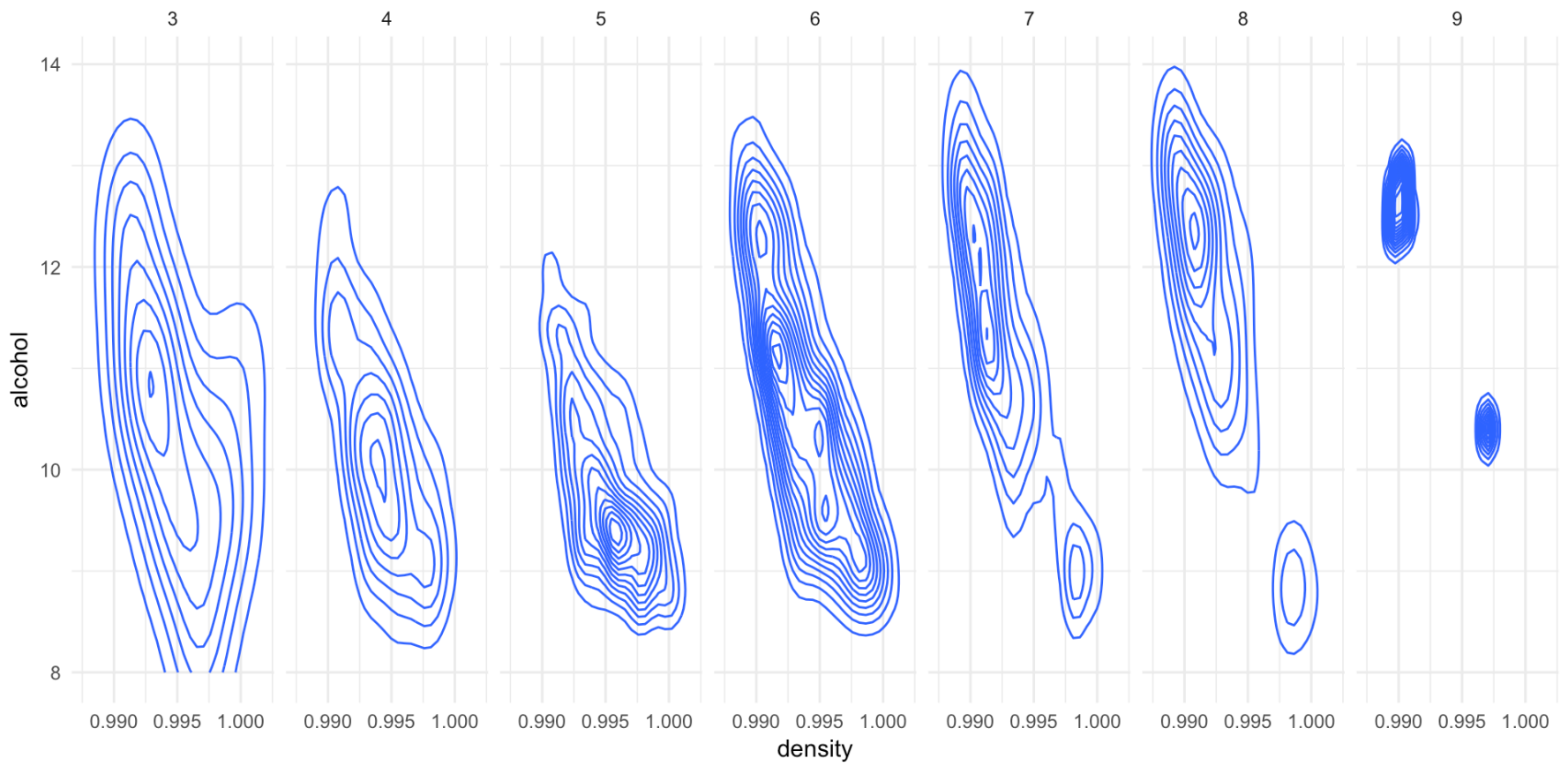
- The strongest relationship was between residual sugar and density

---

Multivariate Plots Section

First we can look at the two most strongly correlated features in relation to quality, density and alcohol.

A density plot will show how the distribution of these values varies with quality.



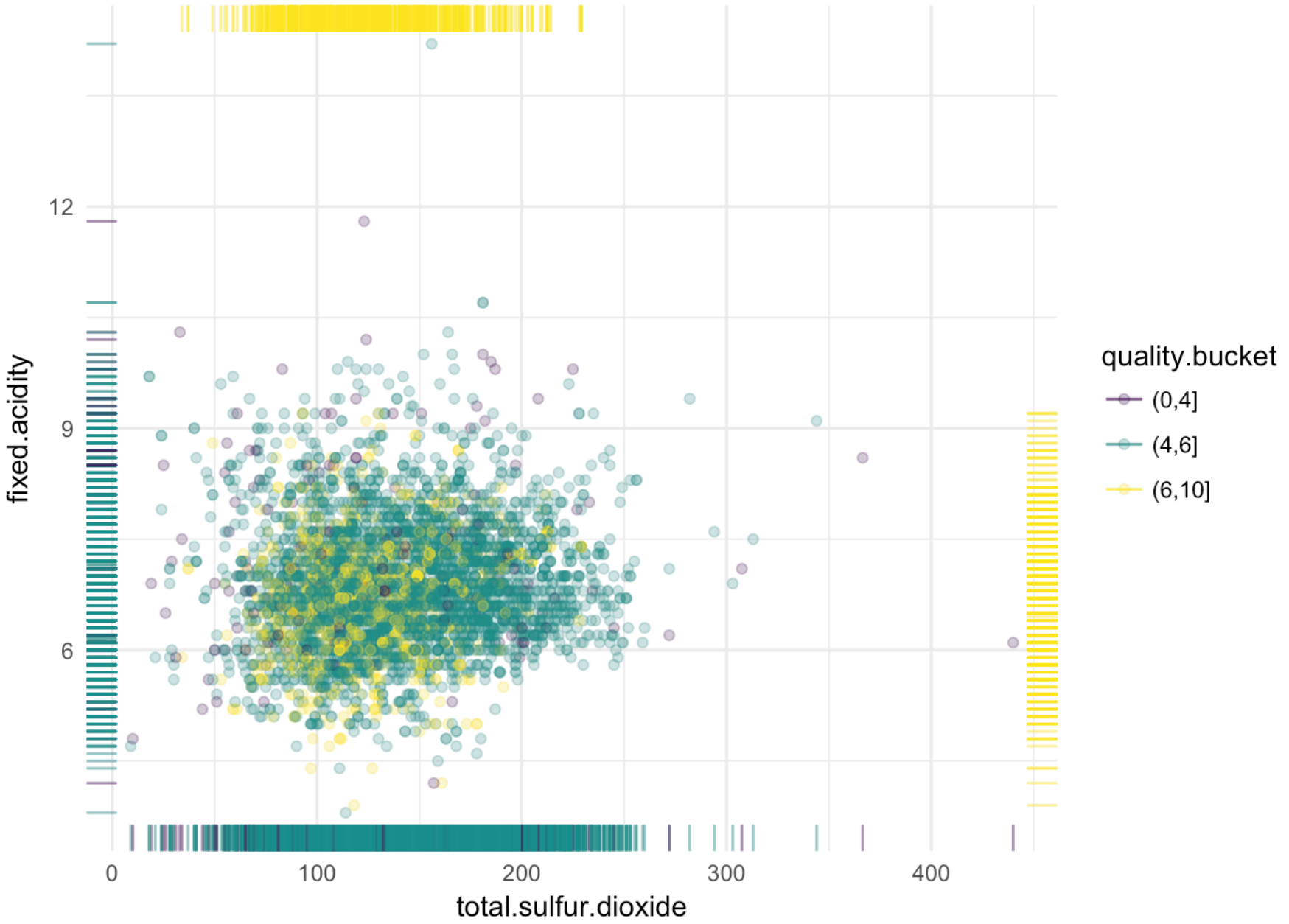
As seen the distribution tends to move upwards and leftwards, as expected from the uni-variate analysis. This trend is more visible in the binned data however,

Here the movement of the density plot, particularly in the high quality bucket is clear to see, as most values are densely packed towards higher alcohol and lower density.

We can now look at the next two more interesting features, chlorides and volatile acidity. Plotting by quality bucket allows us to see the general area that each graph falls. The geom rugs are useful for illustrating the distribution of values, we chose to separate the high quality bucket to better see where they are distributed.

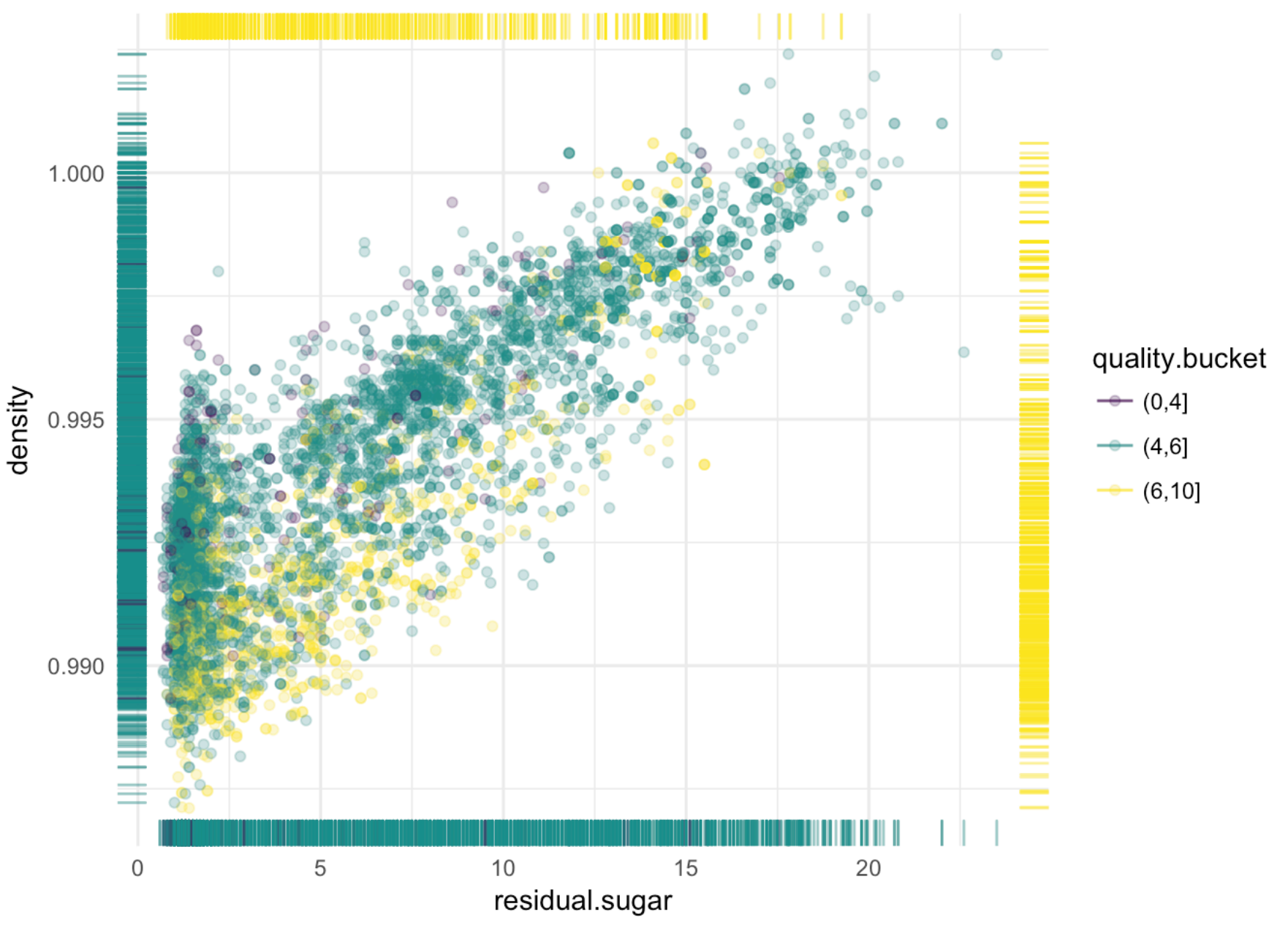
As seen in the points and the rugs the high quality wines are particularly densely group towards low chloride values, whereas mid and low quality wines are much more widely dispersed. A similar tend is observable in volatile acidity, however not to as great an extent.

Looking at two more variables by quality,



Here the trend is not as observable, as in volatile acidity and chlorides, however again the distribution of high quality wines is smaller and more dense than mid and low quality wines, as seen in the rugs.

And looking at the two most highly correlated variables but now by quality,



Here there is an interesting relationship not seen in just the residual sugar plot, in that at a given level of density the high quality wines consistently have higher sugar content than mid and low quality wines. This is seen in how the yellow points underline the trend for most of the graph.

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?

Particularly the relationship between density, alcohol content and quality is strengthened, the trend of movement of the density plot is evidence of such.

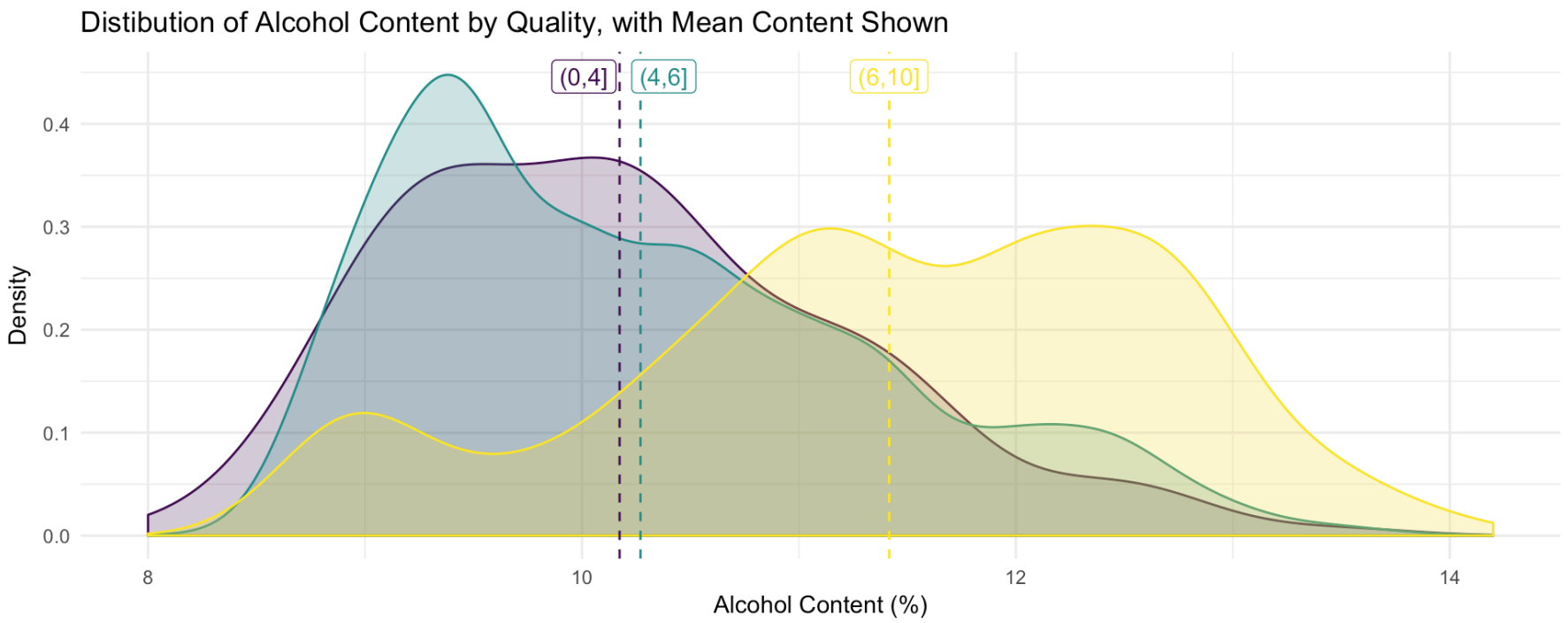
The chlorides and volatile acidity interaction highlighted how higher quality wines are particularly grouped at the lower end of the range for both variables, and they show much less variability than low and mid quality wines.

Were there any interesting or surprising interactions between features?

As mentioned there is an interesting relationship between residual sugar and density not previously observable in just bi-variate analysis. At a given level of density the higher quality wines, predominantly have higher sugar contents.

Final Plots and Summary

Plot One



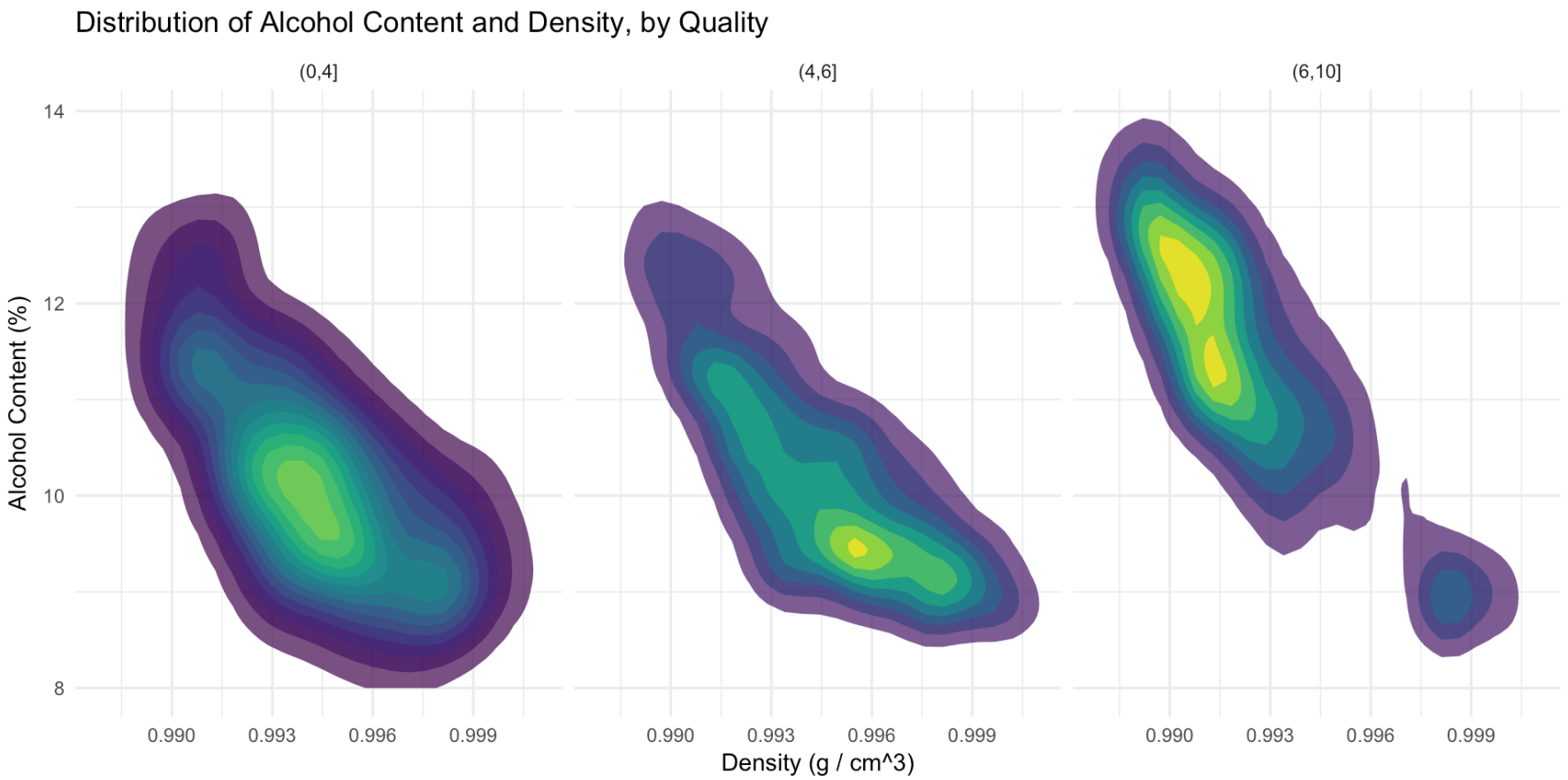
Description One

This plot shows the trend of higher quality wines to have higher alcohol content by showing how the distribution of wines' alcohol contents moves rightwards as quality increases.

The below summary statistics show how higher quality wines have higher average alcohol content. Particularly telling is the first quartile, showing that more than 75% of high quality wines have a higher alochol content than the mean value of low and mid quality wines.

```
## $`(0,4]`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.40   10.10   10.17   10.80   13.50
##
## $`(4,6]`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.40   10.00   10.27   11.00   14.00
##
## $`(6,10]`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   10.70   11.50   11.42   12.40   14.20
```

Plot Two



Description Two

This plot shows how the distribution of density and alcohol content, together trend by alcohol quality. As seen the distribution tends towards higher alcohol content and lower density, particularly for high quality wines. It also illustrates the relationship between alcohol and density, and to some extent the underlying curve through linking these variables, as explored earlier.

The following statistics can give an indication of the centre of these density plots, and in such the same trend towards lower density and higher alcohol can be seen.

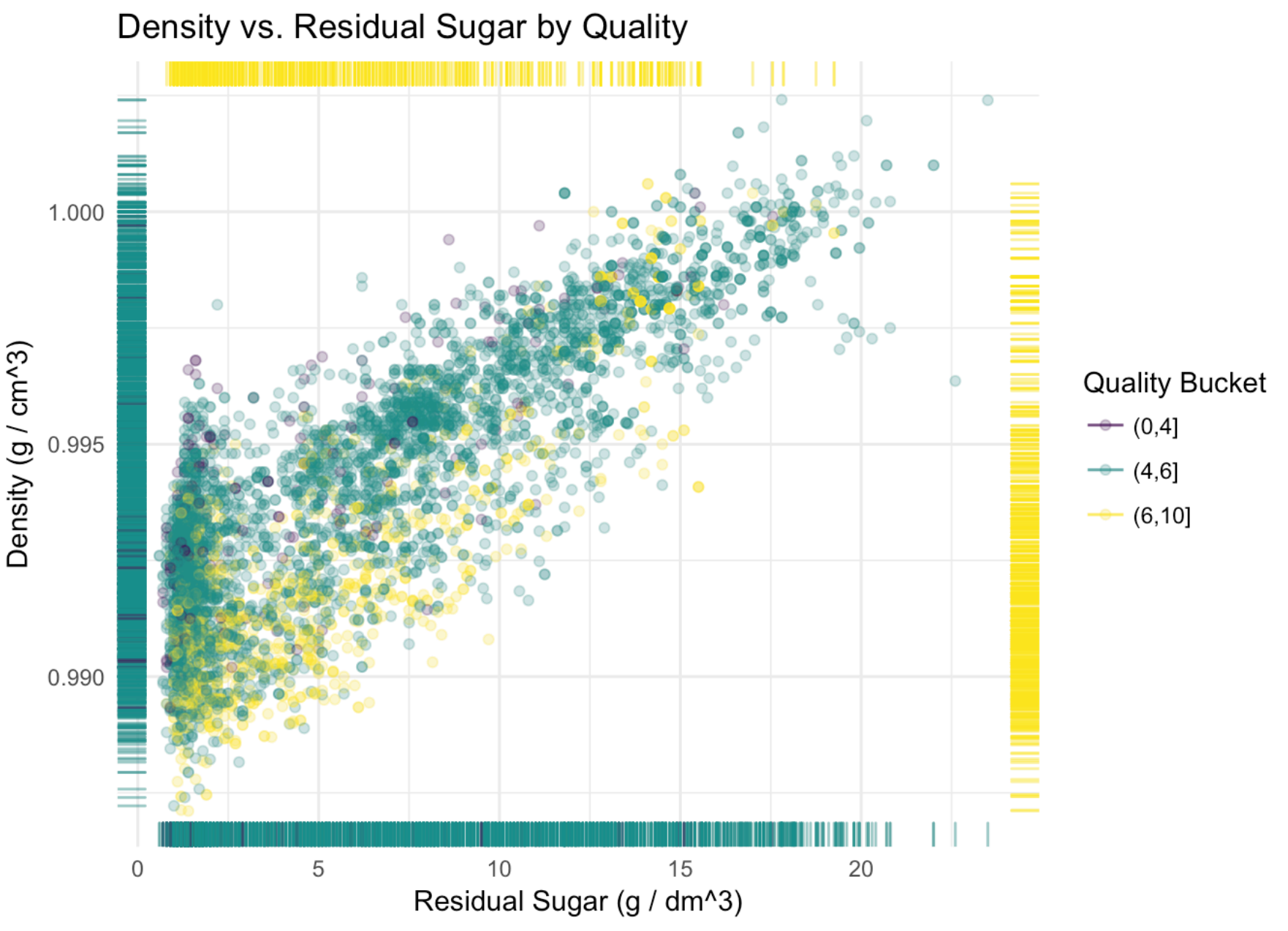
```
## [1] "Density Means"
```

```
##     (0,4]     (4,6]    (6,10]
## 0.9943431 0.9944801 0.9924120
```

```
## [1] "Alcohol Means"
```

```
##    (0,4]    (4,6]   (6,10]
## 10.17350 10.26981 11.41602
```

Plot Three



Description Three

This plot shows an interesting relationship, in showing how for a given density high quality wines tend to have higher residual sugar content. It also illustrates the strong relationship between residual sugar and density.

Reflection

This data-set allowed for exploration over many variables and different possible combinations, and as such relationships which could be found in the data.

The initial uni-variate analysis gave a quick insight into the structure of the data and how the variables were distributed. It also brought attention to possible outliers and non-normal distributions, such as the chloride distribution.

Moving on from that the majority of the bi-variate and multivariate exploration was guided by the correlations calculated by a linear model. I was able to look into the accuracy of one of these linear models, between residual sugar and density, and explored how the accuracy of the model decreases as density increases.

The multi-layered density plots by quality proved to be some of the most insightful and it was pleasing to see how relationships could be seen in this manner.

One aspect which was difficult as first was trying to plot one variable against many, as seen in the box-plots by quality bucket section, however with the help of a web recourse, it became quite an elegant solution, and allowed for quick insight onto the relationships over the binned data.

Moving into the multivariate plots the most surprising and unexpected result (as it wasn't hinted to in the bi-variate analysis) was the relationship between residual sugar and density. I though this was an interesting underlying relationship.

Moving forward I would be interested in comparing this white wine data-set with the red wine data-set, which would allow for a much more detailed categorical exploration. I would also like to create models for the data, and a means to predict the quality of wine given the characteristics, perhaps a linear regression or random forest, and explore which characteristics are necessary for such a model to be accurate.

Sources

- Method to plot one variable against all others: https://drsimonj.svbtle.com/plot-some-variables-against-many-others (https://drsimonj.svbtle.com/plot-some-variables-against-many-others)