Vino Analysis by Lorenz

Intro

Vino is everywhere! Especially white wine is the go the beverage during the summer as it is served best ice cold. And let's face it: it is extremely hard to determine the quality of a wine, and most people are not able to do it (myself included)!

> **What makes a quality vino?**

About the Data

The dataset contains data from 4898 variants of the Portuguese "Vinho Verde" (white wine). Overall, there are 12 variables in the dataset. The data is taken from chemical analysis (11 variables) and 1 subjective evaluation by wine experts for the input variable 'quality'. 'quality' represents the median of at least 3 evaluations made by wine experts. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

Overview of variables and description of their meaning.

Variable	Descripton
1. fixed acidity (g / dm^3 = liter(l))	tartaric acid
2. volatile acidity (g / l)	high level can lead to vinegar taste
3. citric acid (g / l)	small quantities, can add freshness, flavor to wines
4. residual sugar (g / l)	sugar remaining after fermentation stops. It's rare to find wines with less than 1 g / l and wines with more than 45 g / l are considered sweet
5. chlorides (sodium chloride - g / l)	amount of salt
6. free sulfur dioxide (mg / l)	prevents microbial growth and the oxidation of wine
7. total sulfur dioxide (mg / l)	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 (g / cm^3)	density of wine is close to that of water depending on the percent alcohol and sugar content
9. pH	scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
10. sulphates (g / l)	a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
11. alcohol (percentage by volume)	percentage of alcohol

Variable	Descripton
12. quality	score between 0 and 10

Variables can be grouped in 3 categories:

Category 1:

Acidity (1, 2, 3, 9)

Category 2:

Sugar & Salts (4, 5, 6, 7, 10)

Category 3:

Misc - density (8), alcohol percentage (11) and quality (12)

Predict Vino Quality

The variable of interest (dependent variable) is ‘quality’, since this is the ultimate measure of wine. In the end, I would like to know which constitutes to look for when buying the next vino. Therefore, I will investigate what variables might influence the quality. Finally, I will build a model that can predict wine quality based on its substances.

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

All independent variables are numerics. ‘quality’ is an integer. First, I removed X, since it is just a counter variable.

Then, I factored ‘quality’ and named the new variable factor.quality since this is a categorical variable.

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ quality.factor      : Factor w/ 7 levels "3","4","5","6",..: 4 4 4 4 4 4 4 4 4 4 ...
```

To get a first impression of the data, I run the summary statistic.

```
##  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.factor
##  Min.   :3.000   3:  20
##  1st Qu.:5.000   4: 163
##  Median :6.000   5:1457
##  Mean   :5.878   6:2198
##  3rd Qu.:6.000   7: 880
##  Max.   :9.000   8: 175
##                  9:   5
```

Most of the variables, except residual sugar, have means and medians pretty close to each other. Additionally, their max values are far from the third quartile (except pH). The distributions of these variables might be normal with outliers on the right tail.

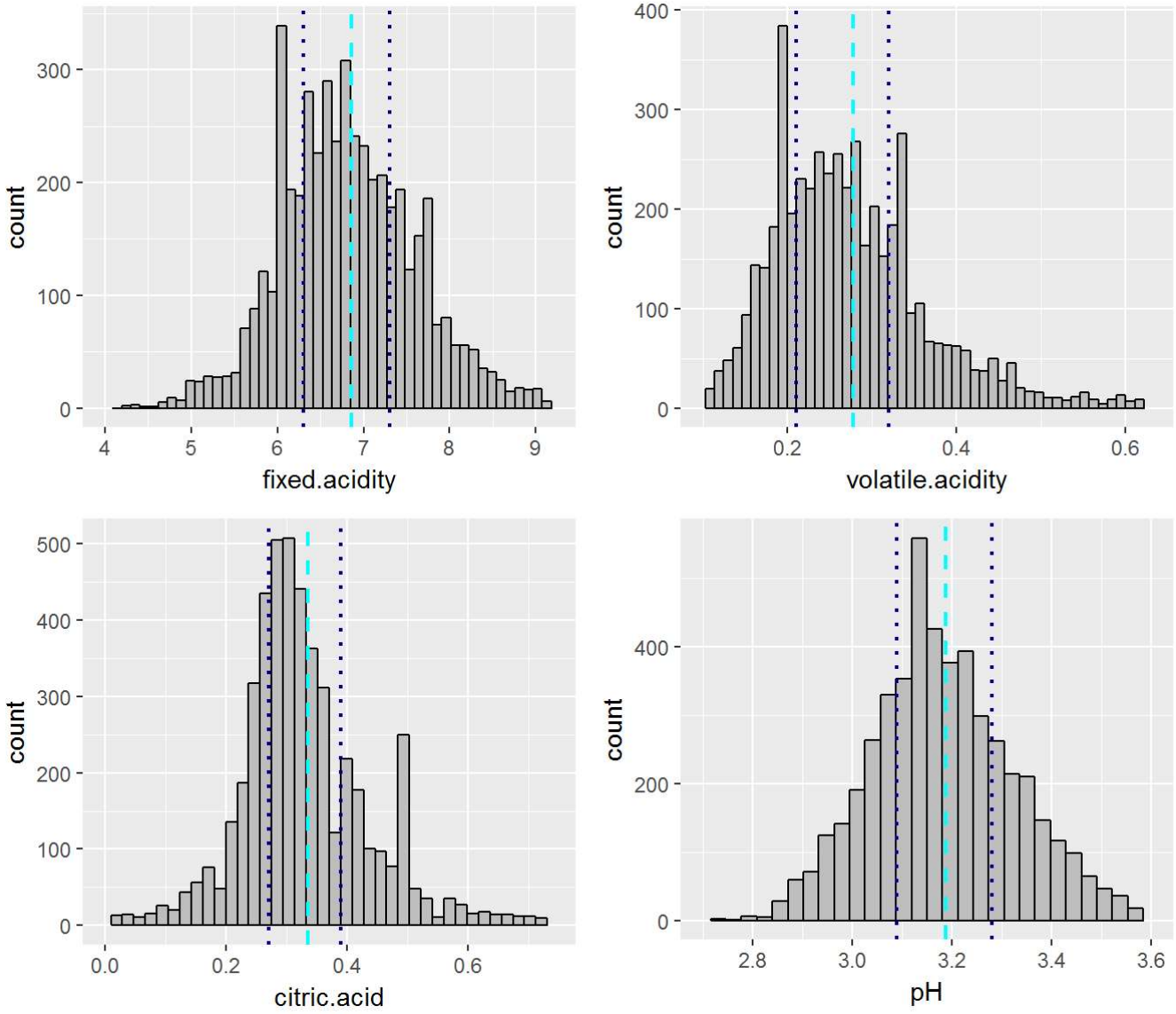
The average alcohol percantage is 10.40% with a mininum of 8% and a maximum of 14.20%. The average sugar amount is 6.391 g/ l with a minimum of 0.60 and maximum of 65.80 g/ l. Density varies between 0.9871 and 1.0390 g / cm^3.

Surprisingly, none of the wines received a perfect quality score of 10, since the highest level for quality observed in the data is 9. At the same time, none of the wines received a 'very poor' quality score of 0 either, since the lowest level observed is 3. It is also interesting to observe that there are very few samples for quality level 3 (only 20) or quality level 9 (only 5). This paucity of data at very low and very high quality scores might make it difficult to draw any statistically significant conclusions about the extremes of the quality scale.

Univariate Plots & Analysis

To get a better idea of the distribution of the variables, I created the following histograms for all 3 groups of variables.

Histograms: Acidity

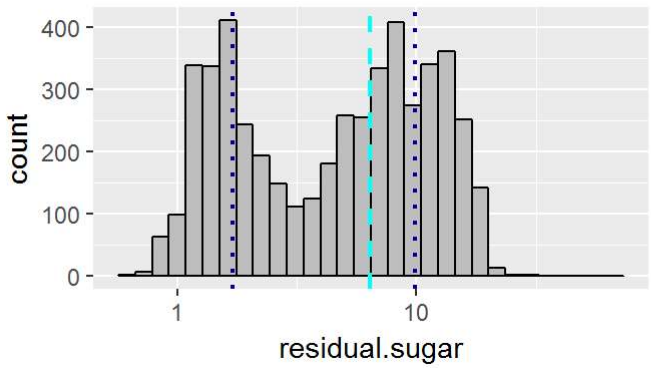


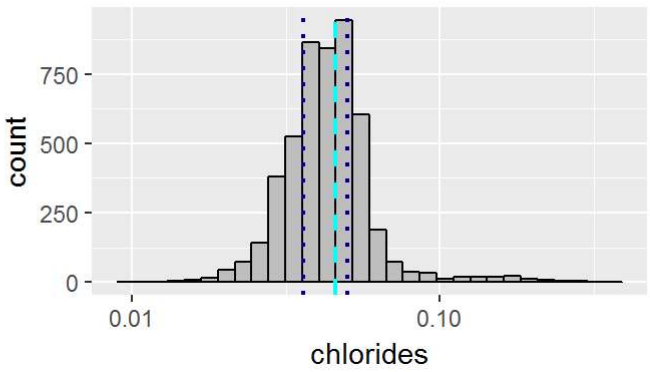
All variables for acidity look reasonably normally distributed. For all variables, there is some positive skewing.

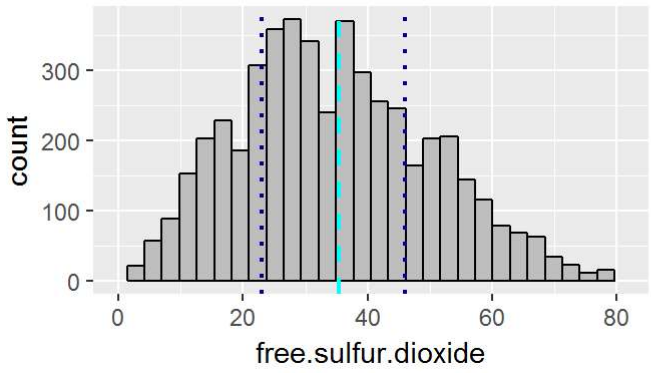
To get a better feeling for the distribution of the bulk of the data, I cut off the upper 1% quantile of the data and choose a bette fitting starting x-value. Additionally, I adjusted the bin sizes and included the mean in the each graph.

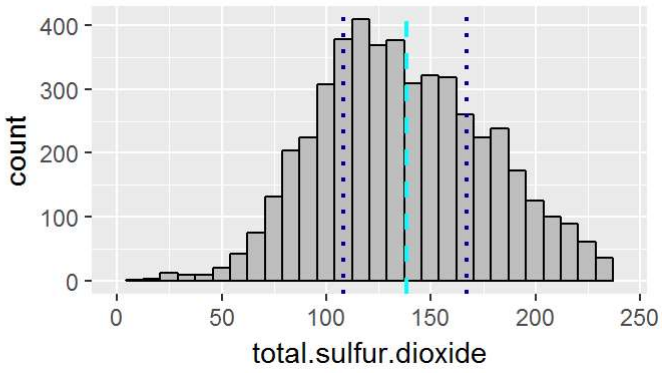
All variables appear to be approximately normally distributed, except volatile.acidity which remains slightly right-skewed. There is one interesting 'spike' in the citric acid profile, near a value of 0.5. There could be something about the wine production process that generates an unusual citric acid profile.

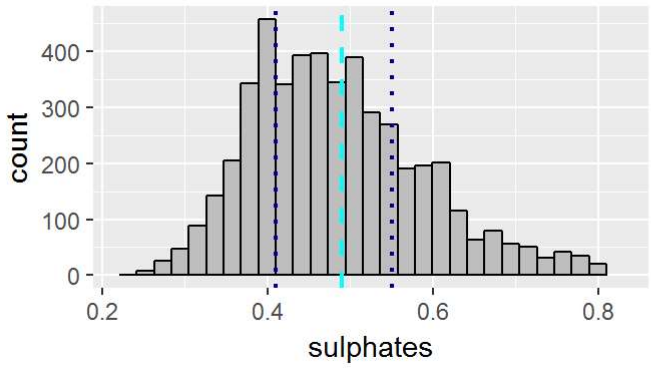
Histograms: Sugar & Salts









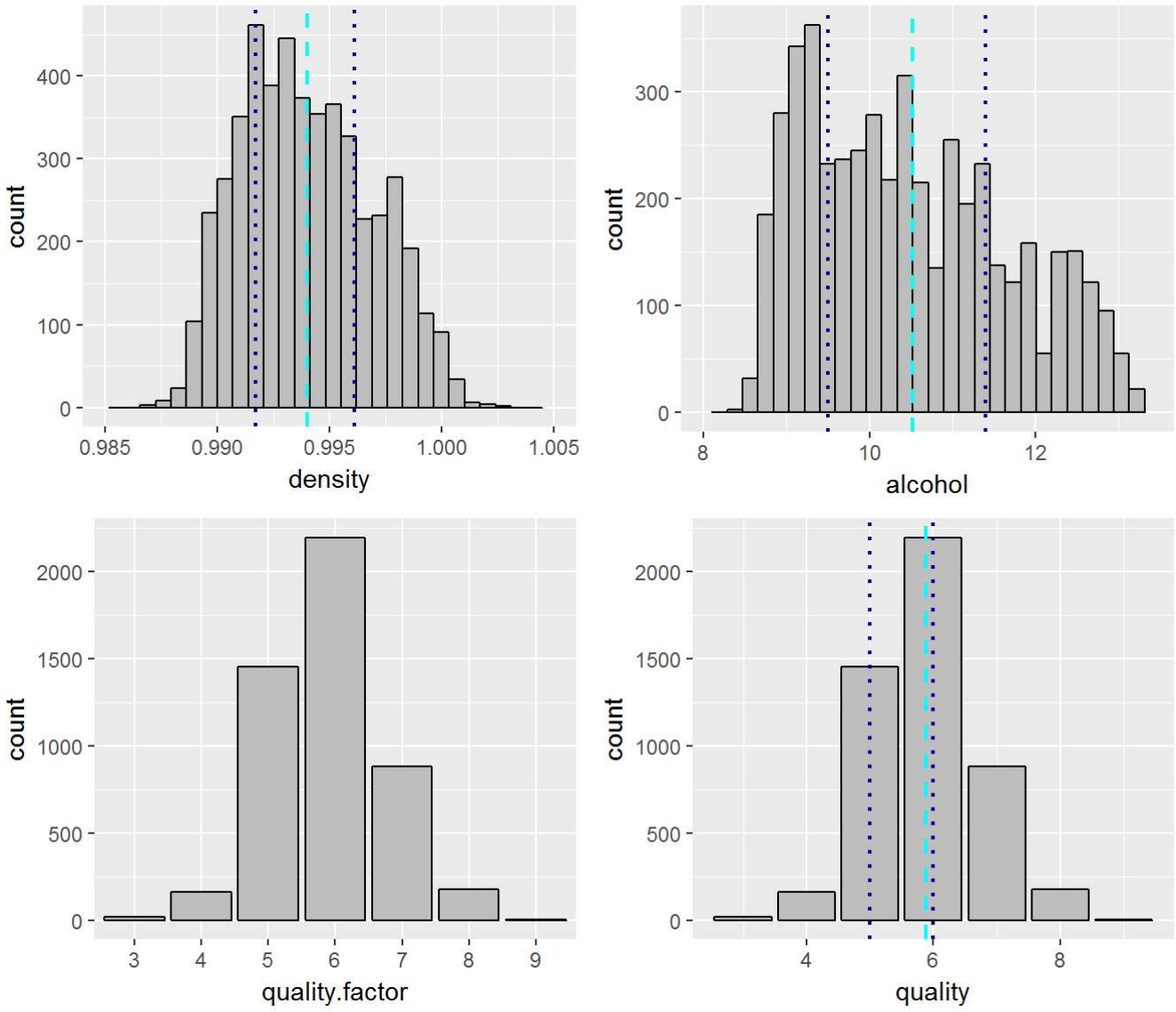


All histograms for sugar & salts look again reasonably normally distributed. For all variables, there is some positive skewing.

Again, I cut off the upper 1% quantile of the data, adjusted the bin size and choose a log scale for residual sugar and chlorides.

It appears that all variables are normally distributed, with the exception of residual sugar. Residual sugar's distribution is positively skewed and looks even bimodally distributed.

More Histograms: Density, Alcohol, Quality



After, cutting off the upper 1% quantile of the data and adjusting the bin size, density and alcohol appear to be approximately normally distributed. Both variables are positively skewed, with alcohol being more positively skewed.

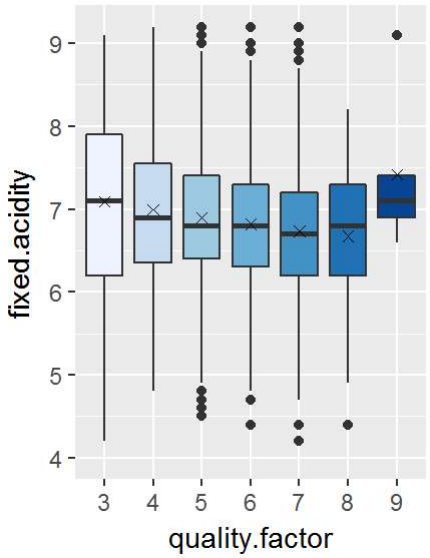
Regarding quality categories, most of the wine in the dataset has a quality ranking of 6 followed by 5 and 7.

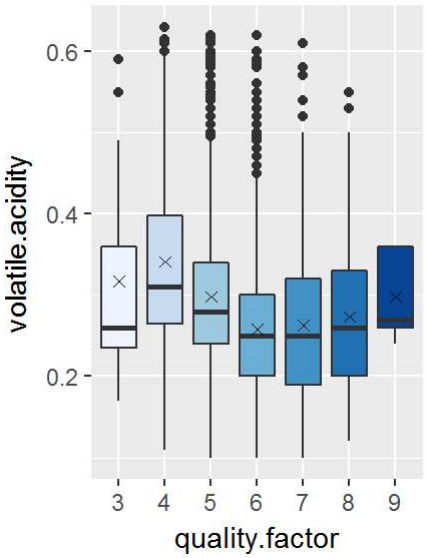
Bivariate Plots Section & Analysis

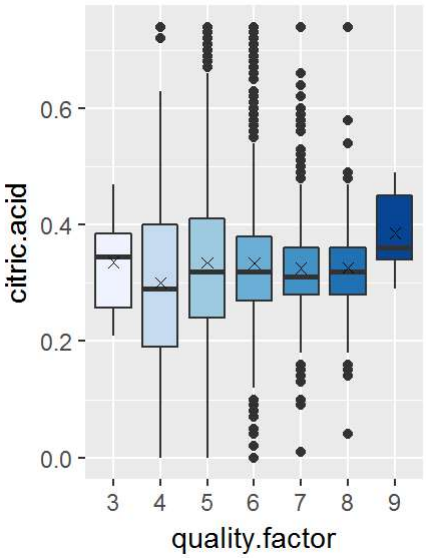
Data Relationships

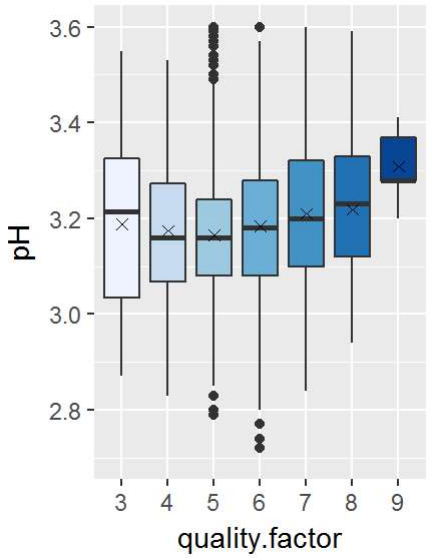
I will kickoff my bivariate analysis with boxplots. Therefore, I created boxplot visualizations for the 11 variables against the dependent variable quality.factor to get a first idea on possible relationships.

I started to look at boxplots for acidity followed by sugar & salts and density plus alcohol percentage.

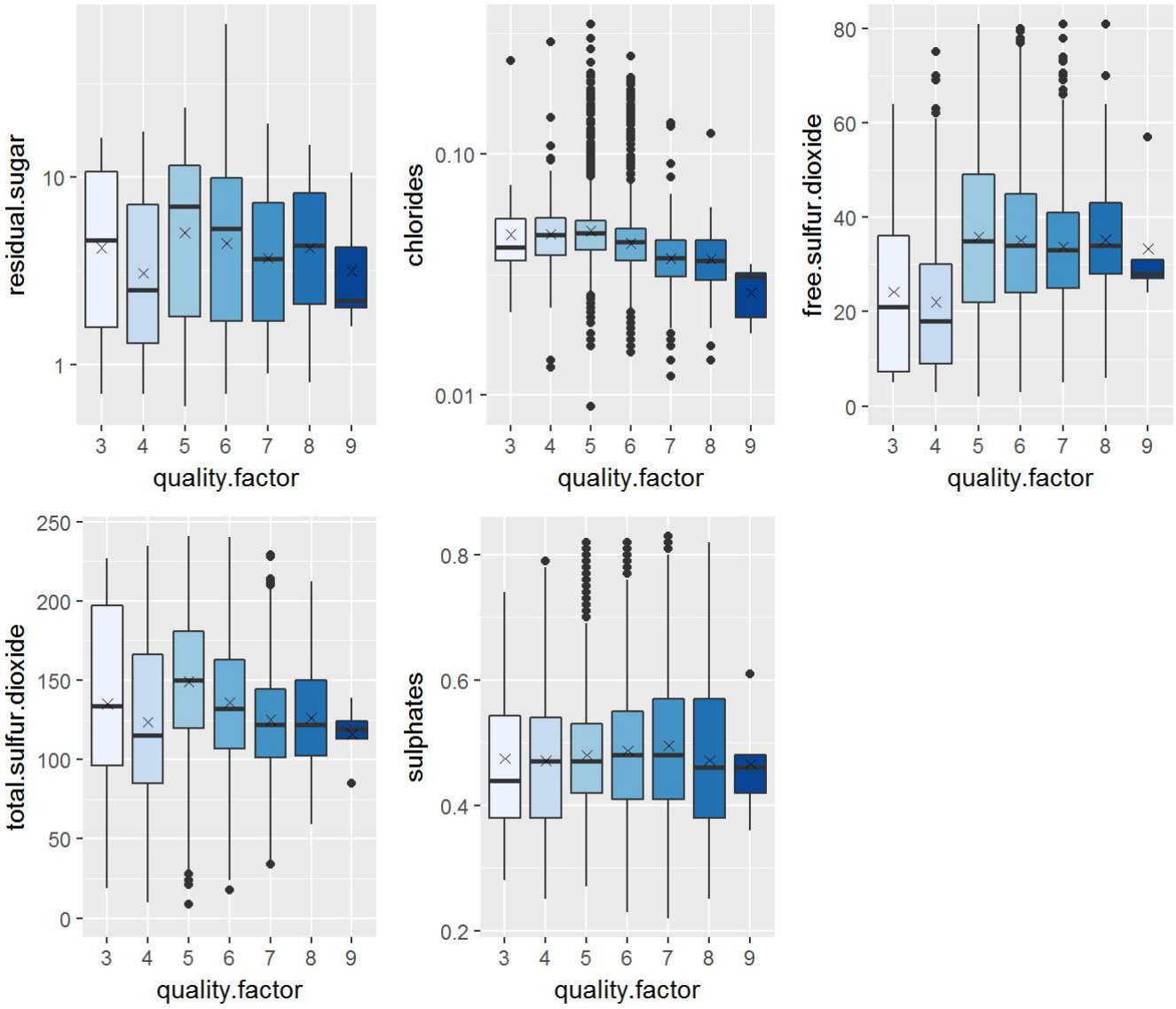




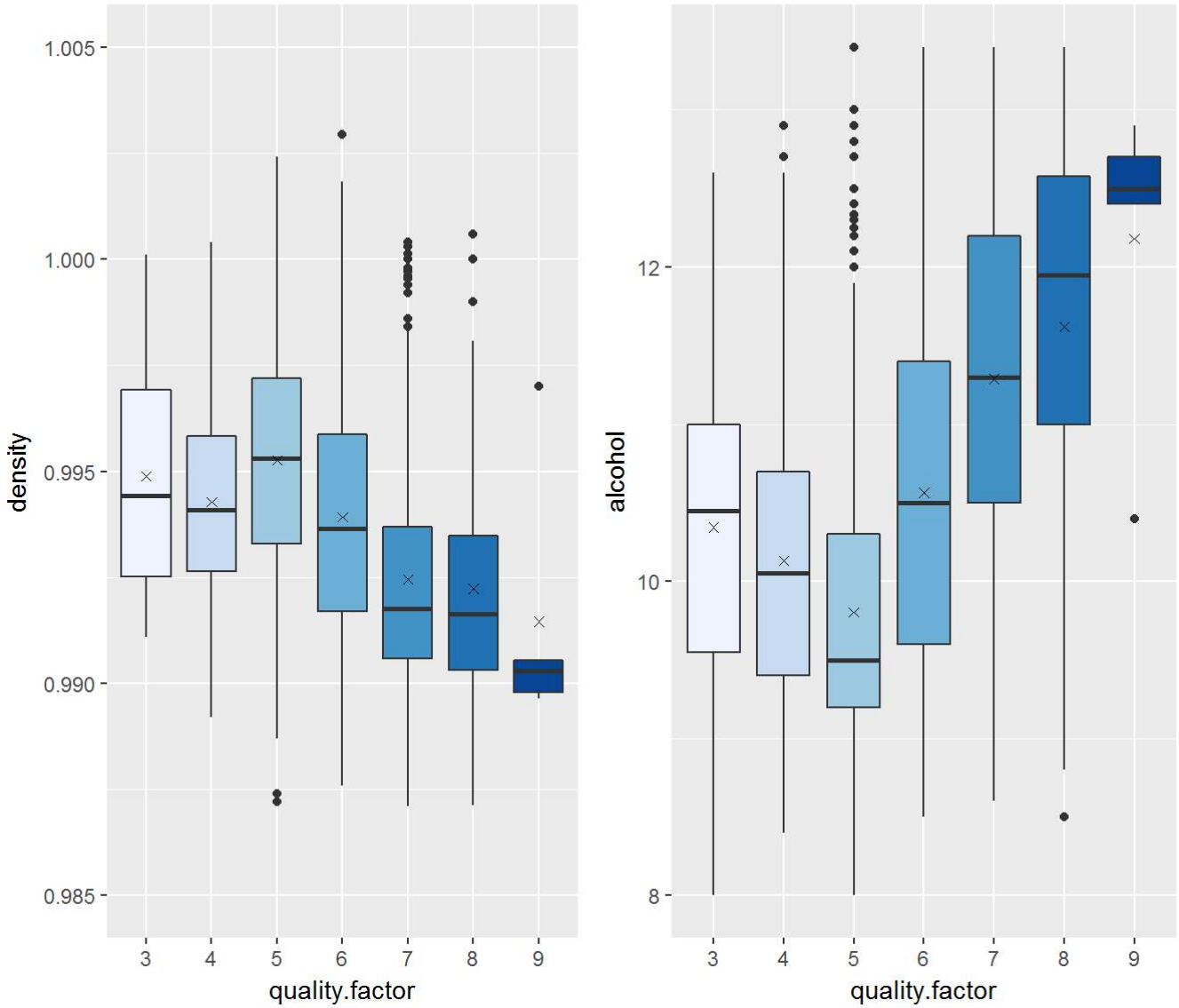




Looking at acidity variables against quality, there are no clear/ sticking out correlations. It looks like wines within the quality category 9 have a higher citric acid amount. However, the amount of observations within quality category 9 is really small (only 5 wines).



For sugar & salts the boxplots illustrate similar findings. Again, no clear trends when measuring sugar & salts against quality categories.

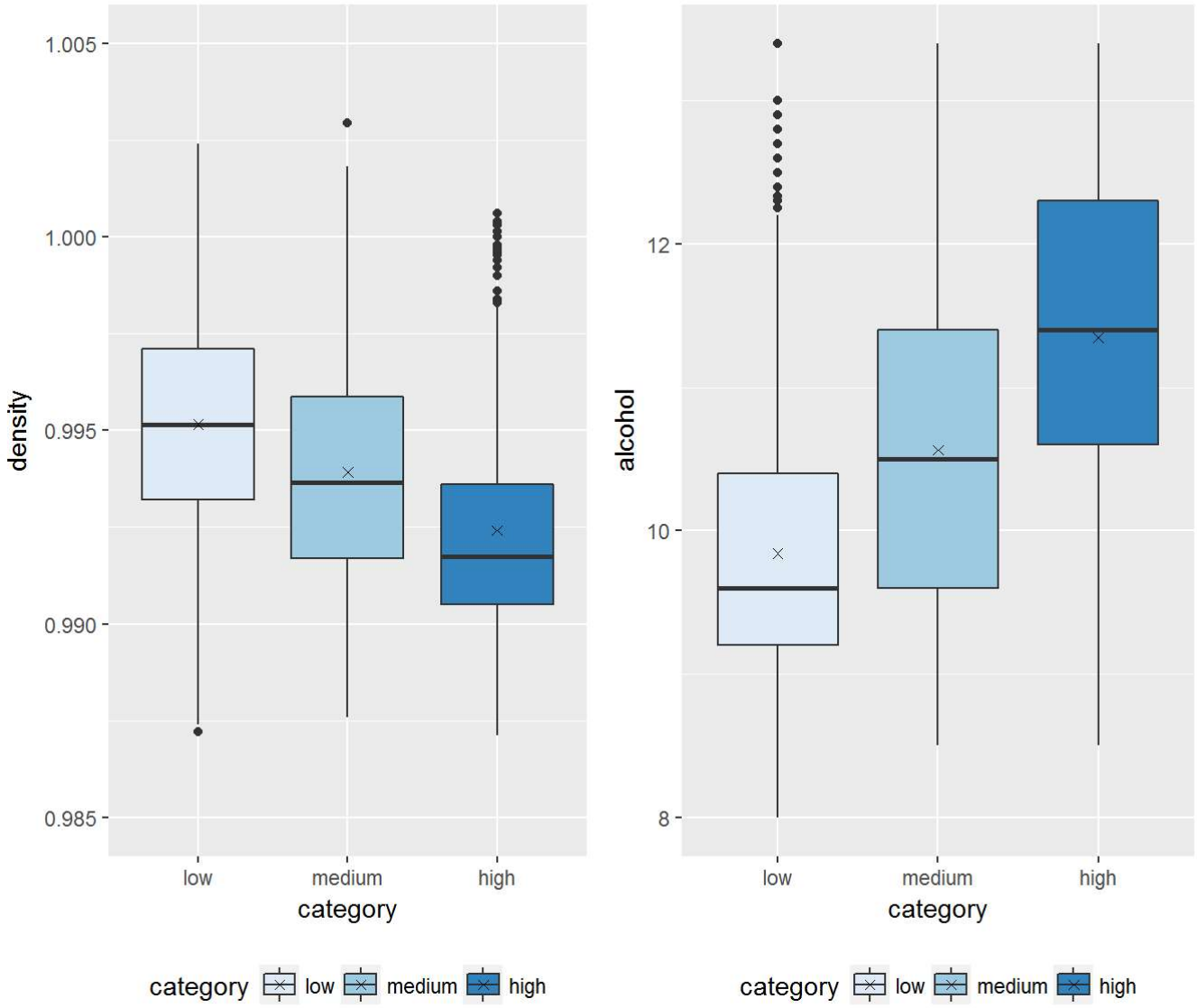


Now turning to density and alcohol percentage, the data writes a different story. It seems that with the increase of quality density is decreasing. On the other hand, for higher quality categories, the alcohol percentage seems to be rising as well. Does more alcohol make more quality vino?

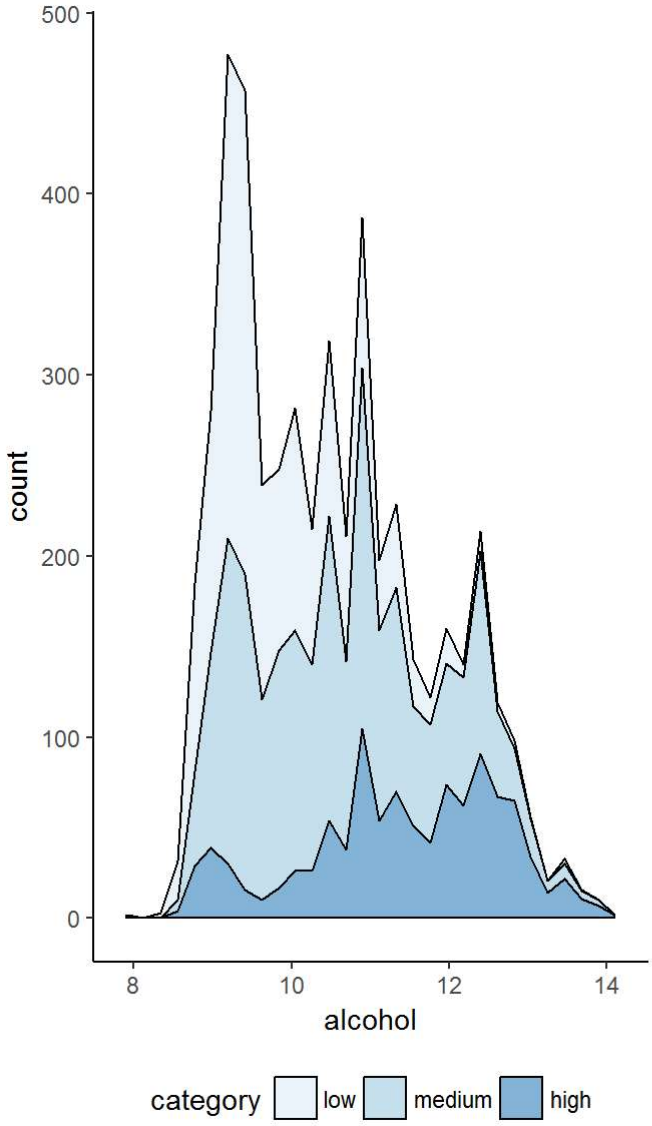
Trying to find out more about this trend, I will create a new categorical variable (category) to group the existing 9 quality categories in the good old “high”, “medium” and “low” levels. “Low” consists of qualtiy scores (3, 4, 5), “medium” of (6) and “high” of (7, 8, 9). I chose this grouping to get a more evenly distributed and sufficiently big number of wines in each category to analyse this trend in more detail.

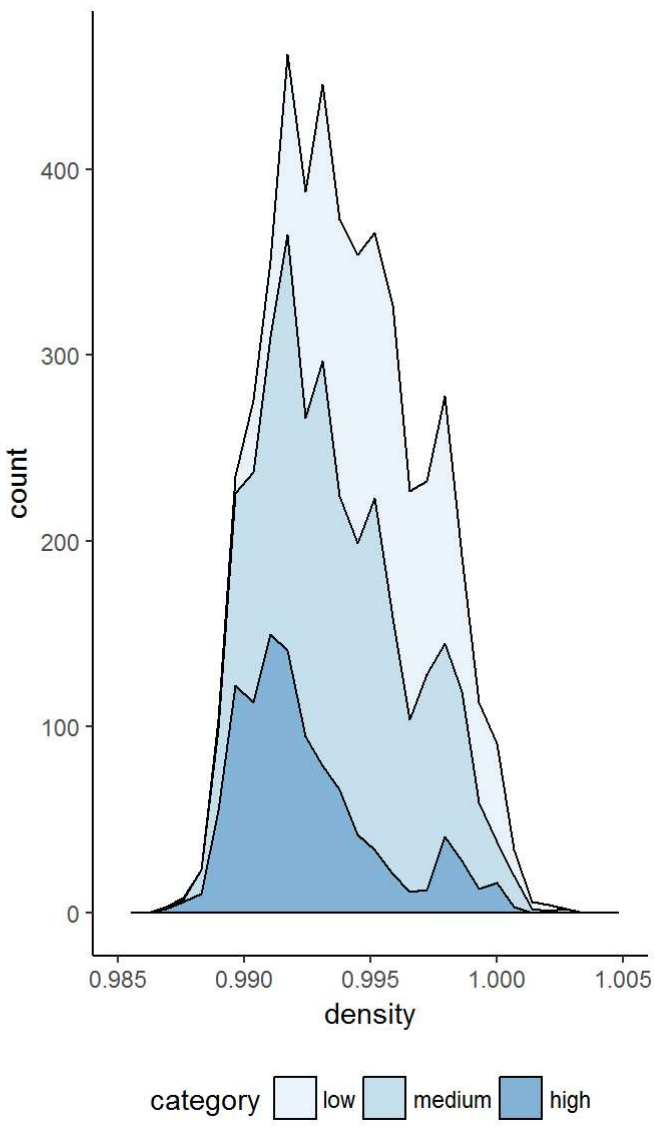
```
##    low medium   high
##   1640   2198   1060
```

Now I will apply the same visualization but using the new variable category.



The new boxplots visualize even better, the negative correlation between density and quality and the positive correlation between alcohol and quality.



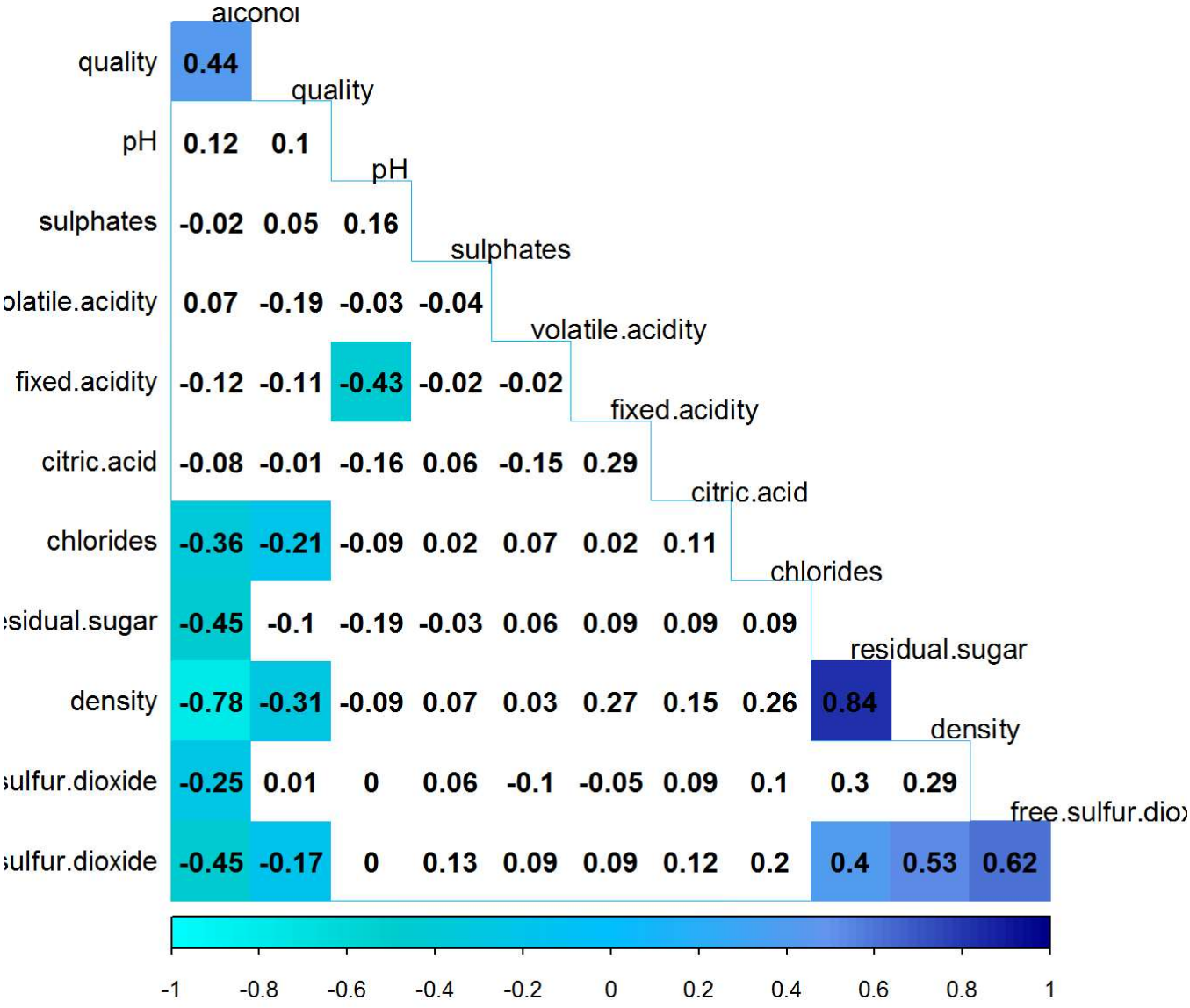


The areaplots (frequency polygons) also support the above observed trend. However, it looks like we have a lot of wines with an alcohol percentage of approx. 9%, 11% and 12.5% across all quality categories since we see peaks at all those alcohol levels. This could be due to a specific type of wine (the same production process). We can observe similar peaks when looking at density.

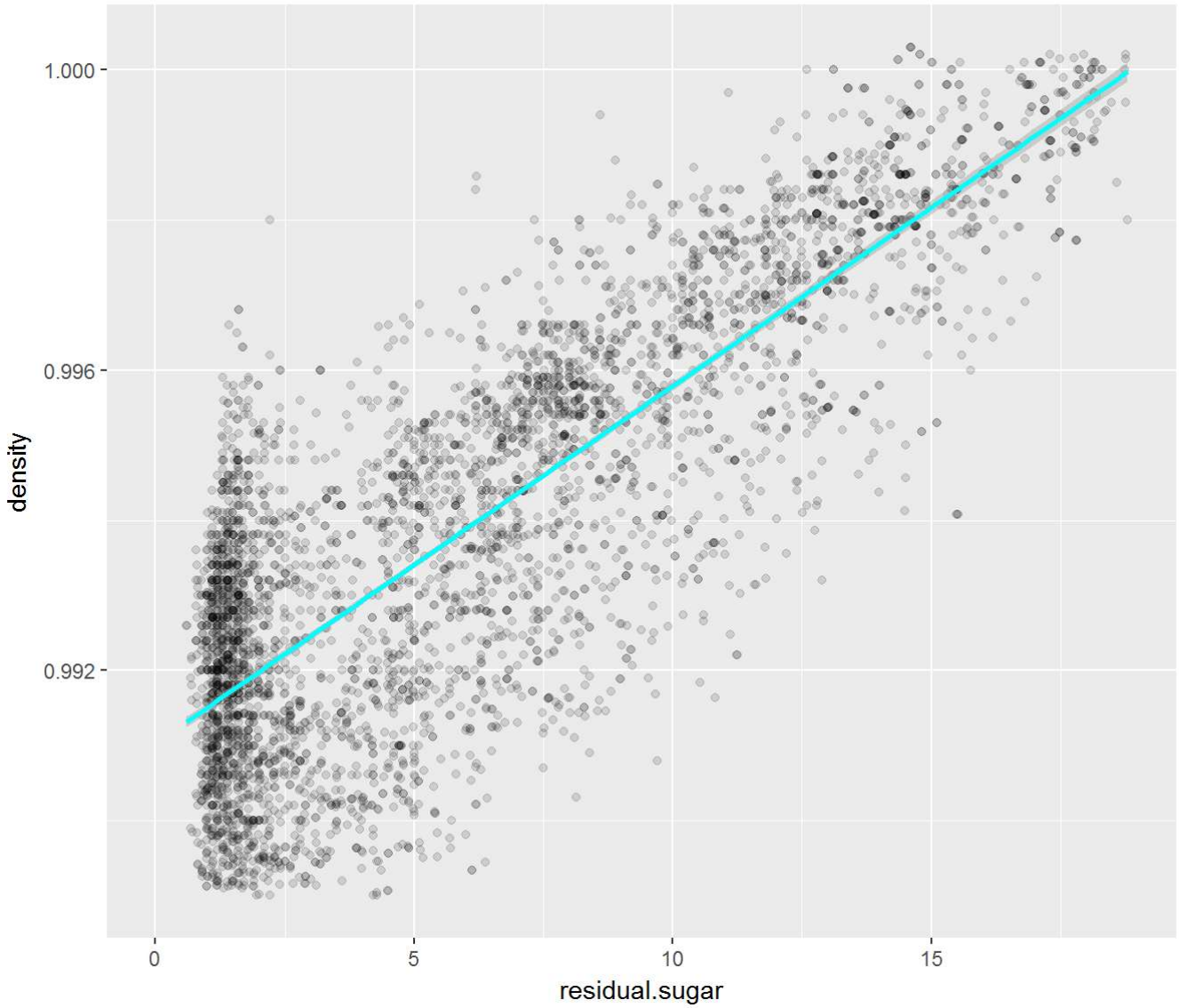
Digging deeper into relationships of all variables, it is always useful to look at correlation coefficients.

In the below figure, correlations with p-value > 0.05 are considered as insignificant. In this case the correlation coefficient values not colored (are neglected due to low statistical significance/ interdependencies with other variables).

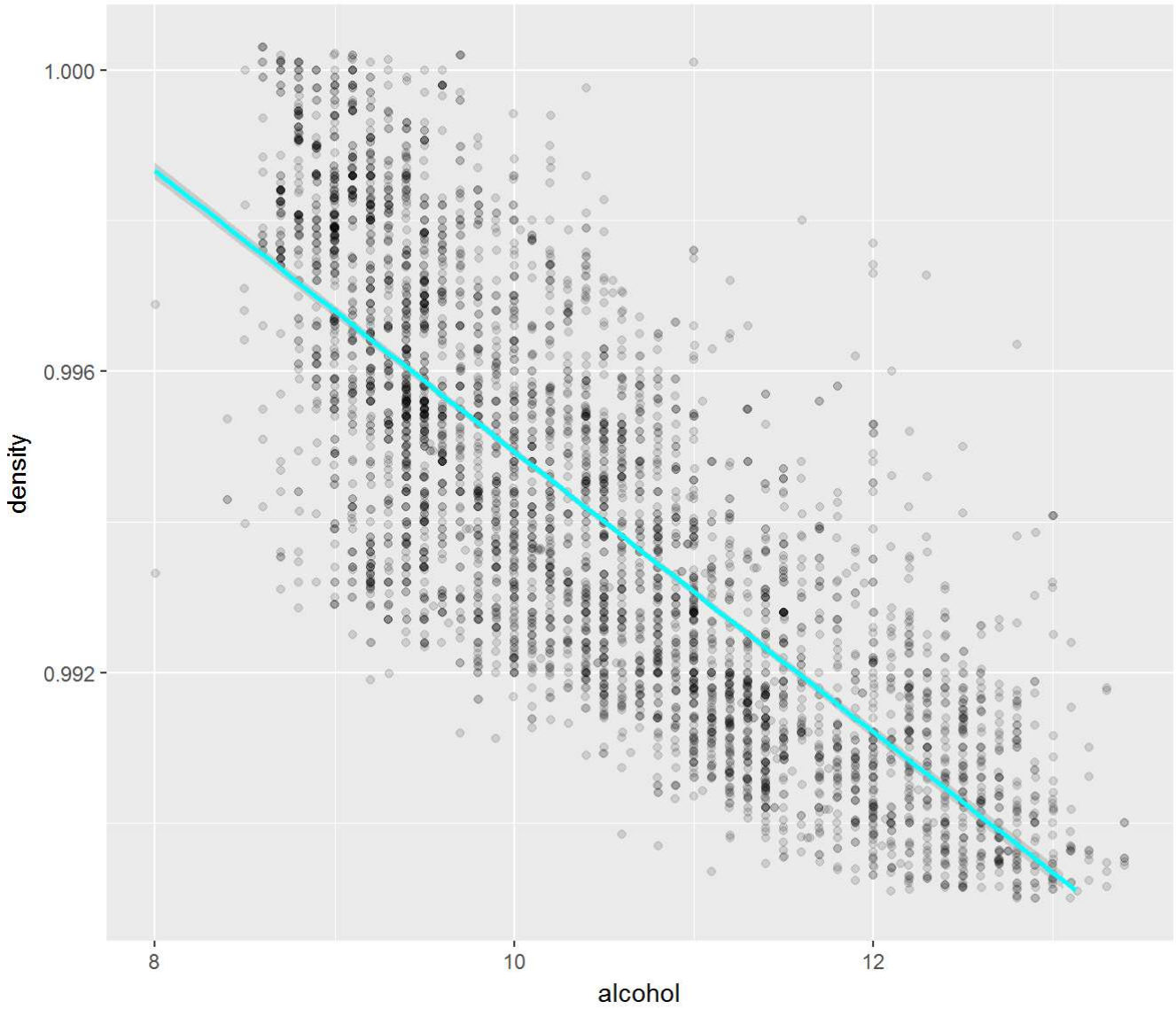
```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity           0.00000000        0.7923645  0.08791755
## volatile.acidity        0.79236447        0.0000000  0.22962896
## citric.acid             0.08791755        0.2296290  0.00000000
## residual.sugar          0.39379324        0.9899498  0.62939340
## chlorides               0.67177019        0.8181973  0.58068121
## free.sulfur.dioxide     0.90142285        0.4597770  0.77621901
##                       residual.sugar chlorides
## fixed.acidity             0.39379324 0.6717702
## volatile.acidity          0.98994984 0.8181973
## citric.acid               0.62939340 0.5806812
## residual.sugar            0.00000000 0.2805576
## chlorides                 0.28055762 0.0000000
## free.sulfur.dioxide       0.06716009 0.4354052
```



This chart displays the highest correlation coefficient (0.84) between density and sugar. This means that these variables have a very strong positive linear relationship. Additionally, we see a strong negative correlation between density and alcohol (-0.78). There is also a moderatly positive relationship between alcohol and quality (highest coefficient with 0.44), which seems to add up with our observations from earlier. I would like to visualize these relationships in the following.



In this chart, we clearly see the positive linear relationship between density and sugar. But why is density increasing with the amount of sugar?

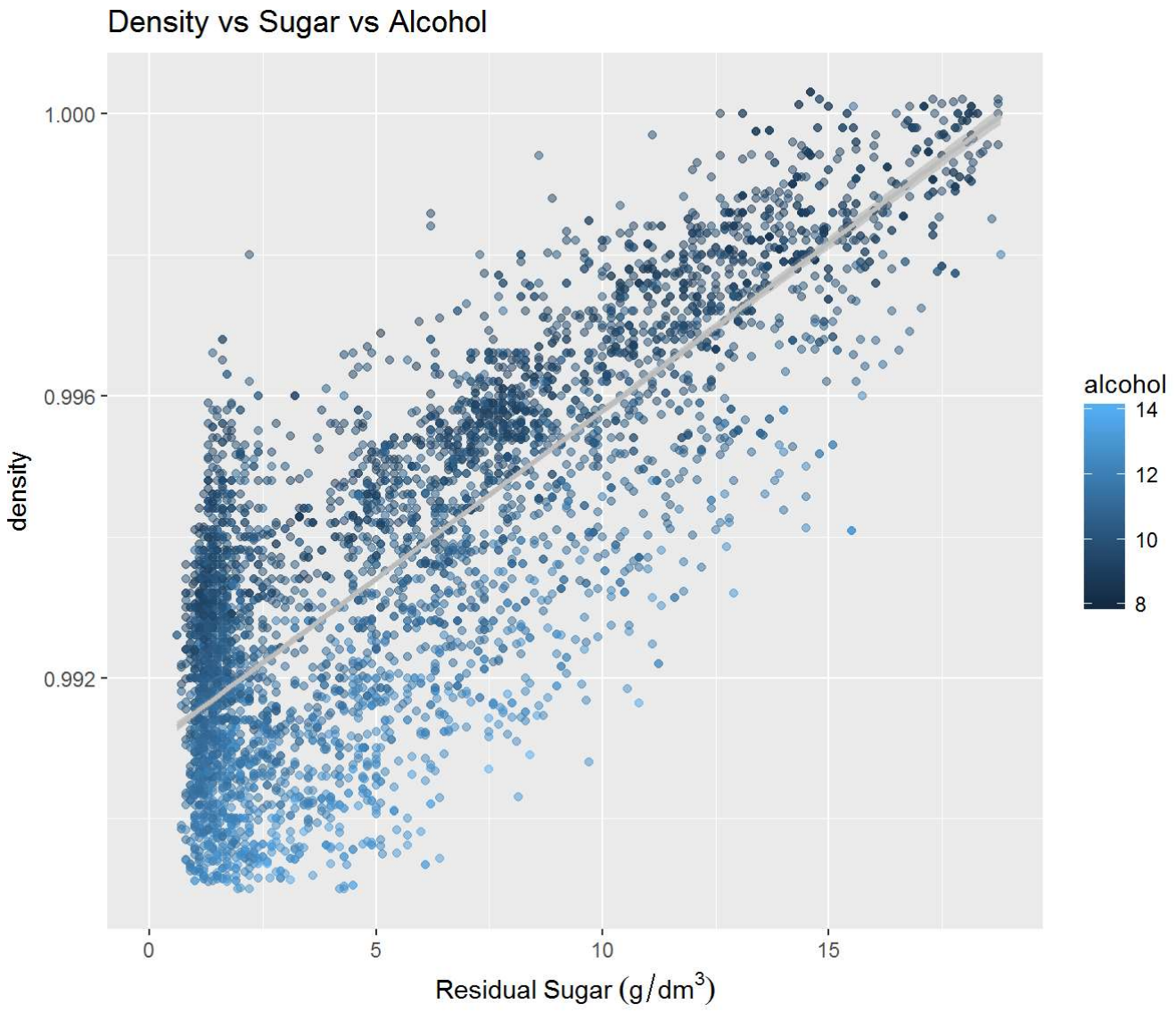


In this graph, we see the clearly negative linear relationship between density and alcohol. Why is density decreasing with higher alcohol levels?

Multivariate Plots Section & Analysis

Time for Multivariate Stuff#

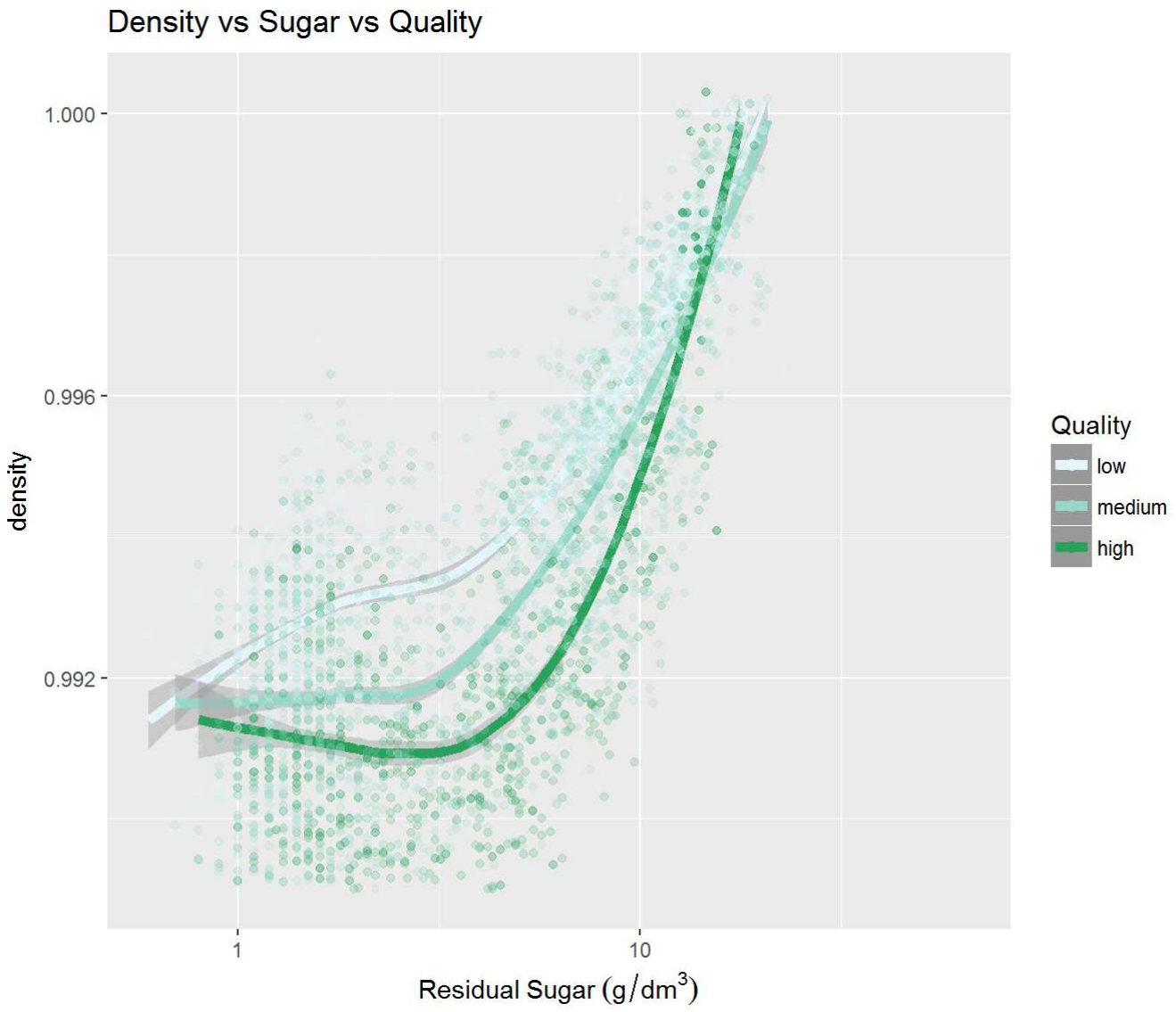
Looking at density, sugar and alcohol at the same time might answer the questions from above.



This visualization lets us observe 2 things: First, sweater wine has more density. Second, wine with the same sweetness has larger volume of alcohol with lower density. In other words, given a value of residual sugar, density increases as alcohol decreases.

The reason for this is in some extent due to the fermentation process of winemaking, in which sugar is consumed to generate alcohol. Since alcohol is less dense than water and sugar is more dense than water, this process makes the density of the wine decrease.

But how does density and sugar affect quality?



This chart shows how quality relates with density and residual sugar. The quality levels have been grouped to improve visibility (category). We can observe, that for a given residual sugar concentration, quality increases as density increases. The same occurs if you fix density and increase residual sugar.

Predictive Model

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = wine)
## m2: lm(formula = quality ~ alcohol + density, data = wine)
## m3: lm(formula = quality ~ alcohol + density + chlorides, data = wine)
## m4: lm(formula = quality ~ alcohol + density + chlorides + volatile.acidity,
##     data = wine)
## m5: lm(formula = quality ~ alcohol + density + chlorides + volatile.acidity +
##     total.sulfur.dioxide, data = wine)
## m6: lm(formula = quality ~ alcohol + density + chlorides + volatile.acidity +
##     total.sulfur.dioxide + fixed.acidity, data = wine)
## m7: lm(formula = quality ~ alcohol + density + chlorides + volatile.acidity +
##     total.sulfur.dioxide + fixed.acidity + pH, data = wine)
## m8: lm(formula = quality ~ alcohol + density + chlorides + volatile.acidity +
##     total.sulfur.dioxide + fixed.acidity + pH + residual.sugar,
##     data = wine)
## m9: lm(formula = quality ~ alcohol + density + chlorides + volatile.acidity +
##     total.sulfur.dioxide + fixed.acidity + pH + residual.sugar +
##     free.sulfur.dioxide, data = wine)
## m10: lm(formula = quality ~ alcohol + density + chlorides + volatile.acidity +
##     total.sulfur.dioxide + fixed.acidity + pH + residual.sugar +
##     free.sulfur.dioxide + sulphates, data = wine)
##
## ============================================================================================
================================================
##                             m1         m2          m3          m4          m5          m6
     m7           m8           m9          m10
## --------------------------------------------------------------------------------------------
------------------------------------------------
##   (Intercept)           2.582***  -22.492***  -21.150***  -35.573***  -30.759***  -43.308***
-43.543***   130.584***   117.752***   149.901***
##                        (0.098)     (6.165)     (6.162)     (6.010)     (6.295)     (6.493)
  (6.510)     (17.934)     (18.125)     (18.760)
##   alcohol               0.313***    0.360***    0.343***    0.389***    0.391***    0.407***
   0.408***     0.222***     0.232***     0.194***
##                        (0.009)     (0.015)     (0.015)     (0.015)     (0.015)     (0.015)
  (0.015)      (0.023)      (0.023)      (0.024)
##   density                          24.728***   23.671***   38.217***   33.251***   46.423***
  46.805***  -130.265***  -117.369***  -149.987***
##                                    (6.079)     (6.074)     (5.926)     (6.234)     (6.458)
  (6.501)     (18.195)     (18.385)     (19.029)
##   chlorides                                    -2.382***   -1.300*     -1.370*     -1.383*
  -1.399**     -0.267       -0.348       -0.234
##                                               (0.558)     (0.542)     (0.543)     (0.540)
  (0.541)      (0.546)      (0.545)      (0.543)
##   volatile.acidity                                        -2.043***   -2.070***   -2.108***
  -2.112***    -2.021***    -1.920***    -1.868***
##                                                          (0.111)     (0.111)     (0.111)
  (0.111)      (0.110)      (0.112)      (0.112)
##   total.sulfur.dioxide                                                0.001*      0.001*
   0.001*       0.001**     -0.000       -0.000
##                                                                     (0.000)     (0.000)
  (0.000)      (0.000)      (0.000)      (0.000)
```

```
##   fixed.acidity                                                                -0.099***
  -0.103***     0.044*       0.044*       0.066**
##                                                                               (0.014)
  (0.015)      (0.021)      (0.020)      (0.021)
##   pH
  -0.042        0.665***     0.642***     0.684***
##
  (0.081)      (0.105)      (0.105)      (0.105)
##   residual.sugar
               0.075***     0.069***     0.081***
##
              (0.007)      (0.007)      (0.008)
##   free.sulfur.dioxide
                           0.004***     0.004***
##
                          (0.001)      (0.001)
##   sulphates
                                        0.632***
##
                                       (0.100)
## ------------------------------------------------------------------------------------------
-------------------------------------------------
##   R-squared                0.190       0.192       0.195       0.248       0.249       0.257
      0.257       0.273        0.276        0.282
##   adj. R-squared           0.190       0.192       0.195       0.247       0.248       0.256
      0.256       0.272        0.275        0.280
##   sigma                    0.797       0.796       0.795       0.768       0.768       0.764
      0.764       0.756        0.754        0.751
##   F                     1146.395     583.290     396.315     402.956     324.034     281.812
    241.554     229.523      207.062      191.810
##   p                        0.000       0.000       0.000       0.000       0.000       0.000
      0.000       0.000        0.000        0.000
##   Log-likelihood       -5839.391   -5831.127   -5822.011   -5657.292   -5654.027   -5627.454
  -5627.322   -5573.700    -5563.613    -5543.767
##   Deviance              3112.257    3101.773    3090.247    2889.234    2885.385    2854.246
   2854.093    2792.280     2780.802     2758.359
##   AIC                  11684.782   11670.255   11654.021   11326.584   11322.054   11270.908
  11272.645   11167.399    11149.225    11111.534
##   BIC                  11704.272   11696.241   11686.504   11365.563   11367.530   11322.880
  11331.114   11232.365    11220.688    11189.493
##   N                     4898        4898        4898        4898        4898        4898
   4898        4898         4898         4898
## ==========================================================================================
=================================================
```

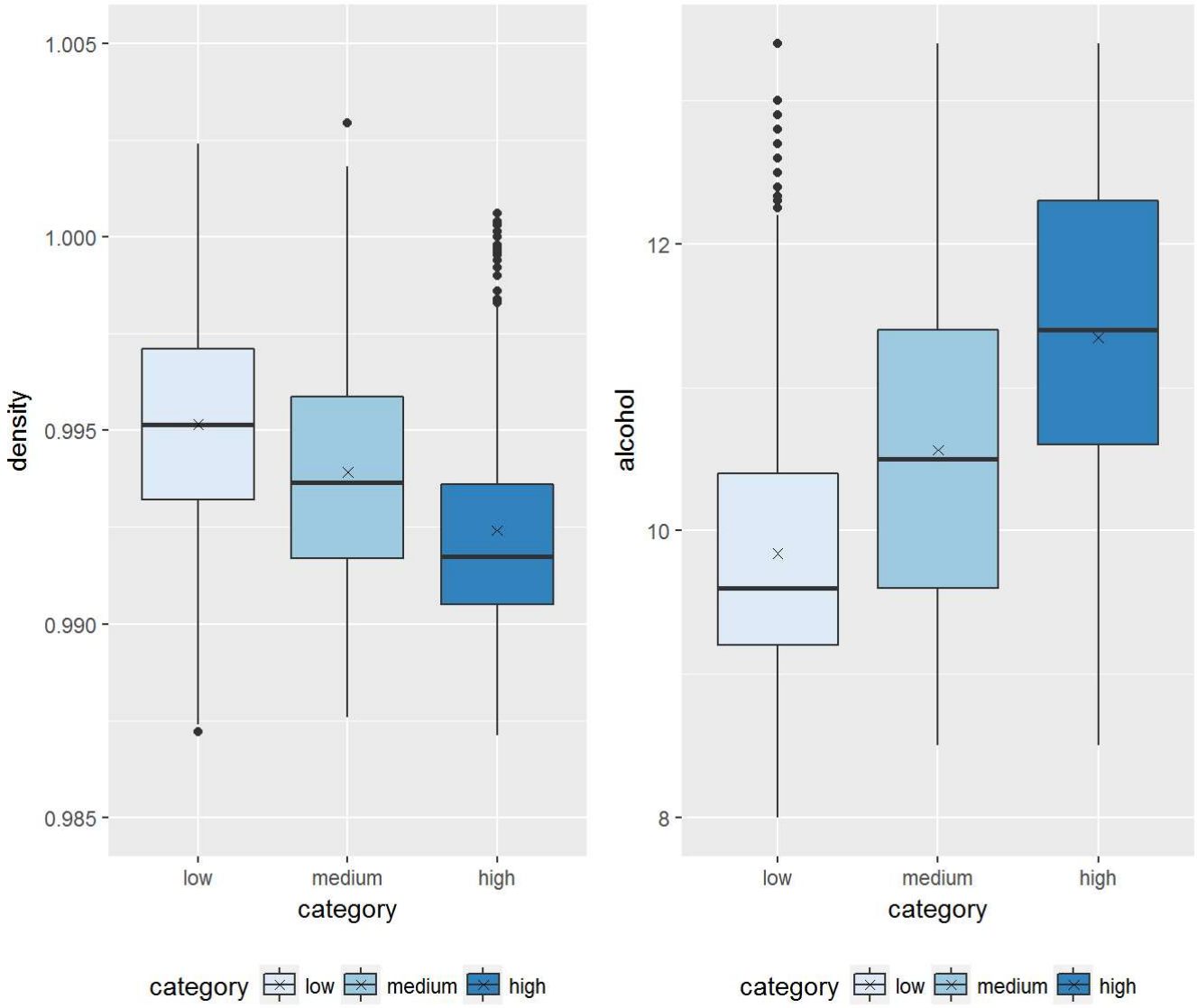
The linear model above can be used for predicting the quality of wine given all substances/ ingredients. However, the R-squared value for the model is 0.282, which is a very low value and means that the model can only explain 28% of the variance of the data.

Further limitations: Since the dataset only contained wines in the 3 to 9 quality range, these models would be unreliable at identifying wines outside of this range. Second, the models are only valid for Portuguese “Vinho Verde” wines. A new model would likely be needed for each wine type.

Final Plots and Summary

Favorites

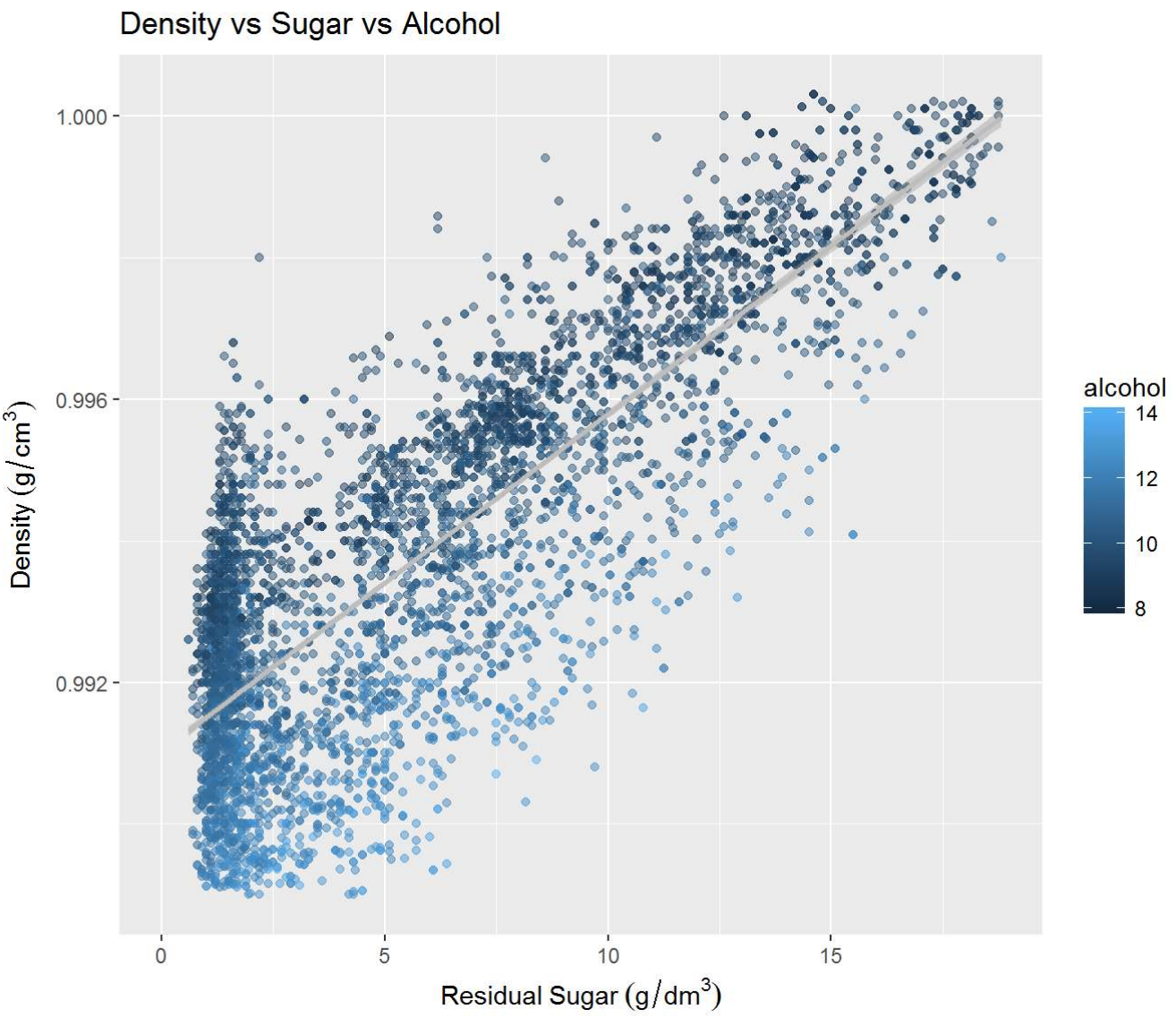
Plot One



Description One

In this visualization, I did boxplots of density and alcohol percentage against the self-build white-wine-quality-levels “high”“,”medium“”, “low”. These categories group the subjective wine quality ratings of wine experts into categories with enough data in each category to find trends/ characteristics for each group.

These boxplots visualize, the negative relationship between density and quality and the positive correlation between alcohol and quality. We see that wines of the higher quality level seem to have more alcohol and less density. #### Plot Two



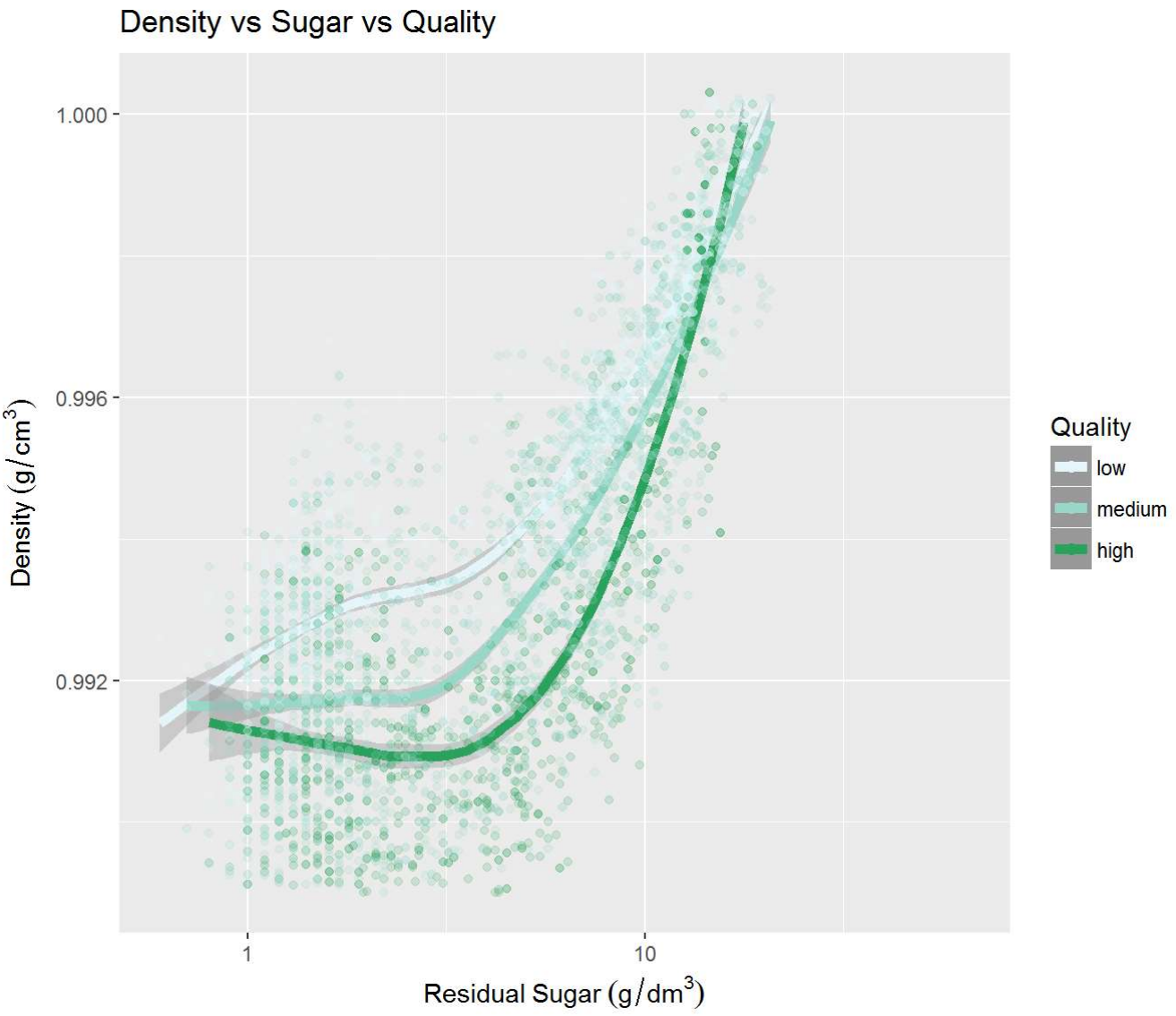
Description Two

In this graph I plotted wine density against residual sugar, while including alcohol percenatage as color for each data point. This visualization lets us observe 2 things: First, sweater wine has more density. Second, wine with the same sweetness has larger volume of alcohol with lower density. In other words, given a value of residual sugar, density increases as alcohol decreases.

The reason for this is in some extent due to the fermentation process of winemaking, in which sugar is consumed to generate alcohol. Since alcohol is less dense than water and sugar is more dense than water, this process makes the density of the wine decrease.

But how does density and sugar affect quality?

Plot Three



Description Three

To answer the question raised above I plotted density against sugar, while including the self-build quality levels to improve visibility (category).

This chart shows how quality relates with density and residual sugar. We observe, that for a given residual sugar concentration, quality increases as density increases. The same occurs if you fix density and increase residual sugar.

Reflection

I structured my vino analysis as follows: at the very beginning I listed all variables with the respective description to better understand what I am dealing with. This is very important because it prepares to ask the right questions. I decdided to group the variables into 3 categories: Acidity, Sugar & Salts and Misc. I applied this structure throughout my analysis to make it easier to follow.

Kicking off the univariate analysis, I run some summary statistics to get a first glance at the data and its distribution. Interestingly, none of the wines received the best or the worst quality score. All wines are ranked between 3 and 9. Additionally, in the dataset we have only a view wines with these extreme quality scores: for quality level 3 only 20 (0.4%) and for quality level 9 only 5 (0.1%). This is an important observation, since I knew at

this point that maybe I would have to adjust the categories to get better insights into characteristics of different wine quality levels. I continued with histograms to get a better feel for the data distribution. After cutting off the upper 1% quantile, adjusting the starting x-value and choosing a log-scale for some variables, all variables look reasonably normally distributed. For a view variables, there is some positive skewing. After visualizing the quality categories the uneven distribution of wines per quality category becomes more tangible.

Moving on to the bivariate stuff, I created boxplots that measure the 11 variables against the dependent variable quality trying to find factors influencing wine quality. Looking at acidity and sugar & salt variables against quality, there seems to be no clear/ sticking out relationships. However, it seems that that with the increase of quality, density is decreasing. And for higher quality categories, the alcohol percentage seems to be rising as well. To dig deeper into that finding, I created new quality categories (high, medium, low) to get more data for each category. Visualizing density and alcohol percentage against the new quality categories even underlines the earlier findings and makes them even more visible. Trying to find out more about the relationships between the data, I created a correlation coefficient matrix. I found a very strong positve linear relationship between density and sugar which I visualized with a graph. Further, we see a strong negative correlation between density and alcohol which I also displayed in a graph. There is also a moderatly positive relationship between alcohol and quality, which seems to add up with our observations from earlier.

In my multivariate section, I looked at the relationship between density, sugar and alcohol since I found out that there is a moderate to strong correlation going on between these variables. My first visualization shows that, wine with more sugar has more density. Moreover, holding the suguar amount constant, wines seem to have a larger volume of alcohol with lower density. In other words, given a value of residual sugar, density increases as alcohol decreases. This could be due to the fermentation process of winemaking, in which sugar is consumed to generate alcohol. Since alcohol is less dense than water and sugar is more dense than water, this process makes the density of the wine decrease. But how does density and sugar affect quality? To answer that question I created a chart that shows how quality relates with density and residual sugar. The quality levels have been grouped to improve visibility (category). We observe, that for a given residual sugar concentration, quality increases as density increases. The same occurs if you fix density and increase residual sugar.

Finally, I run a linear regression to build a model for predicting wine quality. The model can be used for predicting the quality of wine given all substances/ ingredients. However, the R-squared value for the full model (m10) is 0.282, which is a relatively low value and means that the model can explain 28% of the variance of the data. The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line. This is ok because the process of evaluating wines is very subjective, and experts can be biased by their histories and preferences, making the relation between quality and the other variables harder to predict.

However, every variable seems to have a statistically significant influence, since adding variables to the model (m1 to m10) increases the R-value from 0.190 to 0.282. This underlines that it might be interesting to look at the relationships of other variables in the future. For example, looking at total sulfur dioxide vs alcohol (correlation coefficient: -0.45) or density (0.53) or sugar (0.4) might be interesting. Additionally, analyzing chlorides vs quality (-0.36) might also be worth looking into.

Having a bigger dataset with more wines from the extremes of the quality spectrum (i.e. many more quality 0-3 and quality 9-10 wines) might reveal more of the characteristics of quality vino. Further, it might be interesting to try the prediction model for different wine types. The model would probably have to be adjusted. One could create a new model for the other wine type and then compare the models.

The analysis process is a very valuable experience, since I practiced plotting various types of charts, handling overplotting and choosing the best chart type to convey the intended message. I experiemented a lot with the charts I used and it was fun to do so.

References

https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf)

http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram (http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram)

https://stackoverflow.com/questions/43362420/length-of-dimnames-2-not-equal-to-array-extent-when-using-corrplot-function (https://stackoverflow.com/questions/43362420/length-of-dimnames-2-not-equal-to-array-extent-when-using-corrplot-function)

http://www.sthda.com/english/articles/32-r-graphics-essentials/133-plot-one-variable-frequency-graph-density-distribution-and-more/ (http://www.sthda.com/english/articles/32-r-graphics-essentials/133-plot-one-variable-frequency-graph-density-distribution-and-more/)

http://www.sthda.com/english/wiki/ggplot2-area-plot-quick-start-guide-r-software-and-data-visualization (http://www.sthda.com/english/wiki/ggplot2-area-plot-quick-start-guide-r-software-and-data-visualization)
https://ggplot2.tidyverse.org/reference/geom_smooth.html (https://ggplot2.tidyverse.org/reference/geom_smooth.html)

https://stats.idre.ucla.edu/r/faq/how-can-i-explore-different-smooths-in-ggplot2/ (https://stats.idre.ucla.edu/r/faq/how-can-i-explore-different-smooths-in-ggplot2/)

https://stackoverflow.com/questions/11014804/plotting-multiple-smooth-lines-from-a-dataframe (https://stackoverflow.com/questions/11014804/plotting-multiple-smooth-lines-from-a-dataframe)

http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/ (http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/)

http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization (http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization)

http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf (http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf)