Wine Quality Analysis

Contents

Initial Data Cleaning

Now that I've imported the data, I'm going to review the basic shape and structure to evaluate if any transformations need to be performed.

shape of white wines dataframe

```
dim(winesWhite)
```

```
## [1] 4898   13
```

shape of red wines dataframe

```
dim(winesRed)
```

```
## [1] 1599   13
```

structure of white wines dataframe

```
str(winesWhite)
```

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

structure of red wines dataframe

```
str(winesRed)
```

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

It looks like the white and red wine datasets have the same shape and structure. I can combine them into a single dataset to make analysis easier. However, I'm going to make a few transformations before I combine the two datasets. First, I'll remove the unnecessary index column (X). Then I'll create a new column called 'type' so I know which dataset the record came from after combining the two dataframes.

```
# drop the 'X' column - it's not necessary for this analysis
winesRed <- winesRed %>% select(-X)
winesWhite <- winesWhite %>% select(-X)

# add a column to indicate from which dataset the record originated
winesRed$type <- 'red'
winesWhite$type <- 'white'

# combine red and white wine datasets
winesAll <- rbind(winesRed, winesWhite)
```

The columns need to be adjusted to better conform to R best practices. I'm going to loop through the column names, and replace any instance of '.' with '_'.

```
# loop through column names and replace '.' with '_'
for(i in 1:ncol(winesAll)) {
    varName <- colnames(winesAll[i])
    newVarName <- str_replace_all(varName, '\\.', '_')
    colnames(winesAll)[i] <- newVarName
    rm(varName, newVarName)
}

# check new column names
colnames(winesAll)
```

```
##  [1] "fixed_acidity"        "volatile_acidity"     "citric_acid"
##  [4] "residual_sugar"       "chlorides"            "free_sulfur_dioxide"
##  [7] "total_sulfur_dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
## [13] "type"
```

Next, I need to convert the type and quality variables to categorical variables (factors). It's important to perform this step on any variables that have set levels; it makes plotting these variables much cleaner.

```
# type as factor
winesAll$type <- factor(winesAll$type)

# quality as factor
l <- c('1', '2', '3', '4', '5', '6', '7', '8', '9', '10')
winesAll$quality_factor <- factor(winesAll$quality, levels = l)
```

I'll take one more look at th structure to make sure everything looks clean and tidy.

```
str(winesAll)
```

```
## 'data.frame':    6497 obs. of  14 variables:
##  $ fixed_acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile_acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric_acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual_sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free_sulfur_dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
```

```
##  $ total_sulfur_dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
##  $ type                : Factor w/ 2 levels "red","white": 1 1 1 1 1 1 1 1 1 1 ...
##  $ quality_factor      : Factor w/ 10 levels "1","2","3","4",..: 5 5 5 6 5 5 5 7 7 5 ...
```

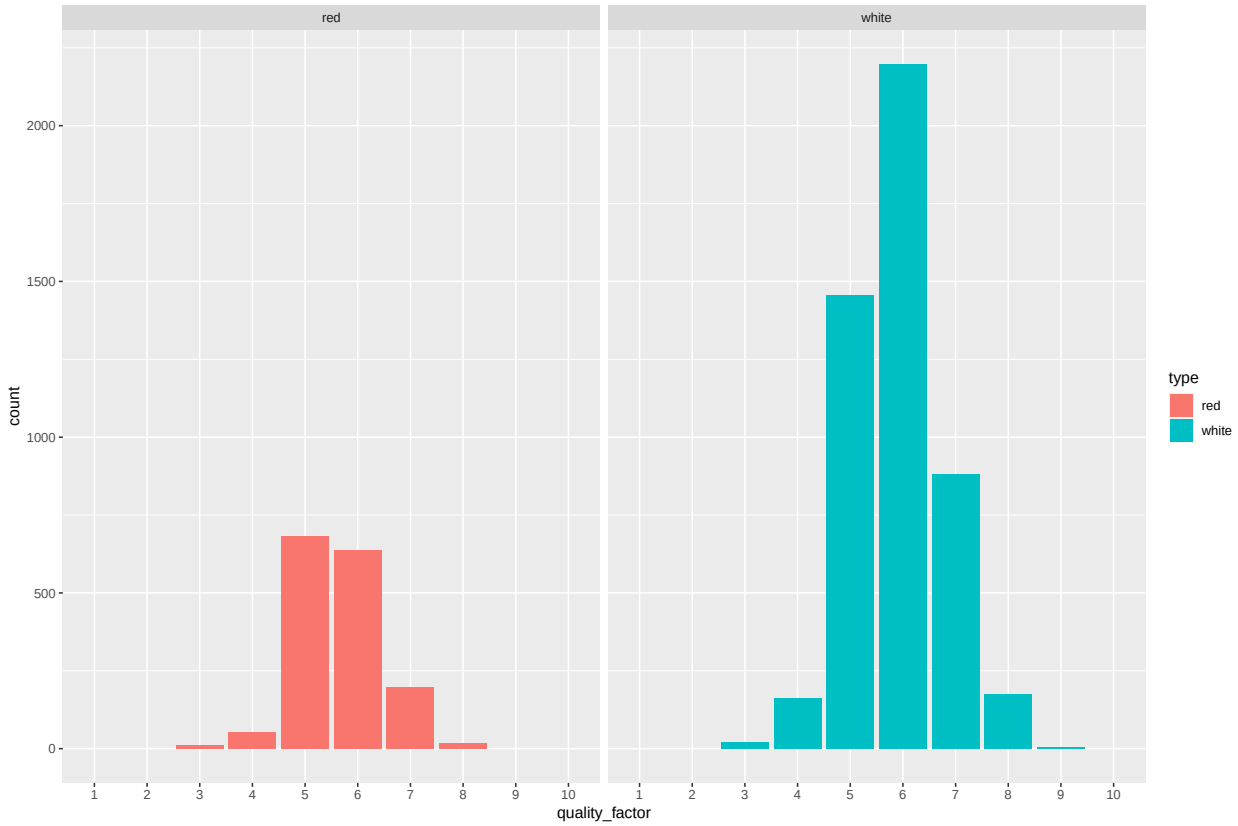
Univariate Plots

In this first section, I'll explore individual variables.

Histogram/Density Plots + Summary

Quality

```
ggplot(winesAll) +
    aes(quality_factor) +
    geom_histogram(stat = 'count', aes(fill = type)) +
    facet_wrap(~type) +
    scale_x_discrete(limits = c(1))
```



It looks like, for both red wine and white wine, the quality scores are grouped around 5 and 6. Let's check the mean and median for each type.

```
winesAll %>%
    group_by(type) %>%
    summarize(
        count = n(),
        mean = mean(quality, na.rm = TRUE),
        median = median(quality, na.rm = TRUE)
        )
```

```
## # A tibble: 2 x 4
##   type  count  mean median
##   <fct> <int> <dbl>  <dbl>
## 1 red    1599  5.64      6
## 2 white  4898  5.88      6
```

There is another interesting observation from the quality histogram: even though quality is reported to be on a scale of 1-10, most of the quality scores are clustered between 4 and 8. Let's take a closer look at that distribution.

```
winesAll %>%
    group_by(quality_factor, .drop = FALSE) %>%
    summarise(count = n()) %>%
    mutate(percent = round(count / sum(count), 4)*100)
```

```
## # A tibble: 10 x 3
##    quality_factor count percent
##    <fct>          <int>   <dbl>
##  1 1                  0    0
##  2 2                  0    0
##  3 3                 30    0.46
##  4 4                216    3.32
##  5 5               2138   32.9
##  6 6               2836   43.6
##  7 7               1079   16.6
##  8 8                193    2.97
##  9 9                  5    0.08
## 10 10                 0    0
```

The distribution for wine quality observed here highlights that quality is a subjective score assigned by humans. If quality were normally distributed, we would see *some* quality scores on the extremes; but in this dataset, there are no observations below 3 or above 9.

```
winesAll %>%
    mutate(quality_group = case_when(
        quality_factor %in% c(1,2,10)  ~ 'none',
        quality_factor %in% c(3,4)     ~ '3-4',
        quality_factor %in% c(5,6,7)   ~ '5-7',
        quality_factor %in% c(8,9)     ~ '8-9',
    )) %>%
    filter(quality_group != 'none') %>%
```

```
    group_by(quality_group) %>%
    summarise(count = n()) %>%
    mutate(percent = round(count / sum(count), 4)*100)
```

```
## # A tibble: 3 x 3
##   quality_group count percent
##   <chr>         <int>   <dbl>
## 1 3-4             246    3.79
## 2 5-7            6053   93.2
## 3 8-9             198    3.05
```

As I suspected - the vast majority of quality scores are between 5 and 7. 93.17% of the observations had an assigned quality score of 5, 6, or 7. Only 3.05% of the observations had a quality score of 8 or 9, and there were only 3.79% with a score of 3 or 4.

In the next section, I'll create a summary and two combined histogram/density plots for each variable in the wines dataset (one for continuous scale and one for log10). To avoid using repetitive code, I'll write a function to handle the summary and plots.

```
PLOT_HISTOGRAM_DENSITY <- function(strVar, limStart=99, limEnd=99)
{
    # create vector based on string variable input
    winesVector <- unlist(winesAll[c(strVar)], use.names = FALSE)

    # adjust default limit start
    if (limStart == 99) {
        limStart <- min(winesVector)
    }

    # adjust default limit end
    if (limEnd == 99) {
        limEnd <- max(winesVector)
    }

    # summary
    summaryOut <- summary(winesVector)

    # calculate number of class intervals (Sturge's Rule)
    # https://www.statisticshowto.com/choose-bin-sizes-statistics/
    k <- 1 + 3.322 * log(6497)

    # calculate bin width - scale_x_continuous
    r <- max(winesVector)- min(winesVector)
    bw <- r / k

    # histogram - scale_x_continuous
    contHist <- ggplot(winesAll) +
        aes(winesVector) +
        geom_histogram(binwidth = bw, aes(fill = type)) +
        facet_wrap(~type) +
        labs(title = 'scale_x_continuous', x = strVar)

    # density plot - scale_x_continuous
```

```
    contDens <- ggplot(winesAll) +
        aes(winesVector, fill = type) +
        geom_density(alpha = 0.5) +
        scale_x_continuous(limits = c(limStart, limEnd)) +
        labs(title = 'scale_x_continuous', x = strVar)

    # calculate bin width - scale_x_log10
    mx <- max(log10(winesVector))
    mn <- min(log10(winesVector))
    bw_log <- (mx - mn) / k

    # set bw_log to static number if result is infinite number
    if ( is.infinite(bw_log) ) {
        bw_log <- 0.069
    }

    # histogram - scale_x_log10
    logHist <- ggplot(winesAll) +
        aes(winesVector) +
        geom_histogram(binwidth = bw_log, aes(fill = type)) +
        facet_wrap(~type) +
        scale_x_log10() +
        labs(title = 'scale_x_log10', x = strVar)

    # density plot - scale_x_log10
    logDens <- ggplot(winesAll) +
        aes(winesVector, fill = type) +
        geom_density(alpha = 0.5) +
        scale_x_log10() +
        labs(title = 'scale_x_log10', x = strVar, y = 'density')

    # print summary and plots
    grid.arrange(contHist, contDens)
    grid.arrange(logHist, logDens)
    return(summaryOut)
}
```

I'm including additional plots for scale_x_log10 because many of these variables are not normally distributed. Plotting them on a log10 scale will often, but not necessarily, result in a normally distributed plot.

Also, I'm guessing that the two sulfur dioxide (SO2) variables could be more informative if compared. Before I generate these plots, I'm going to create another new variable that represents the ratio of free SO2 to total SO2 - percent_free_SO2.

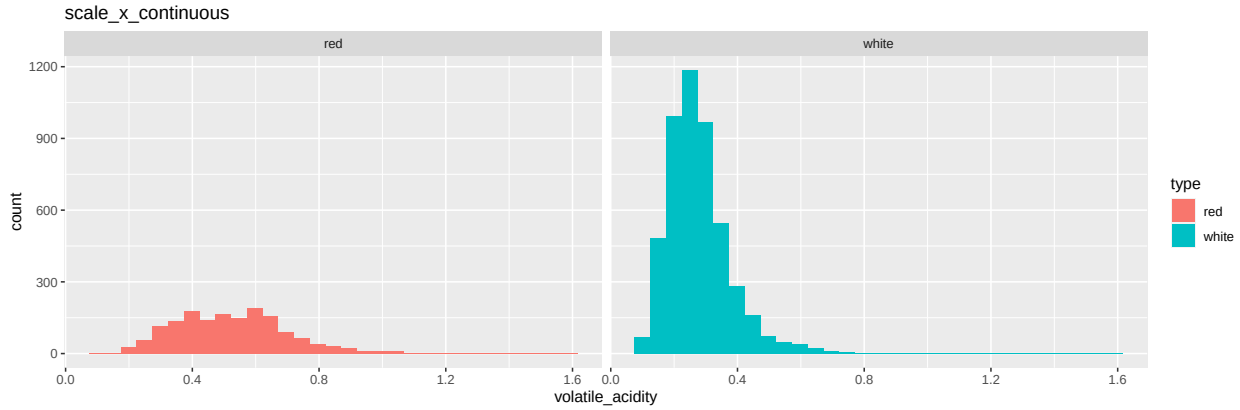
```
newVar <- winesAll$free_sulfur_dioxide / winesAll$total_sulfur_dioxide
winesAll$percent_free_SO2 <- newVar
rm(newVar)
```

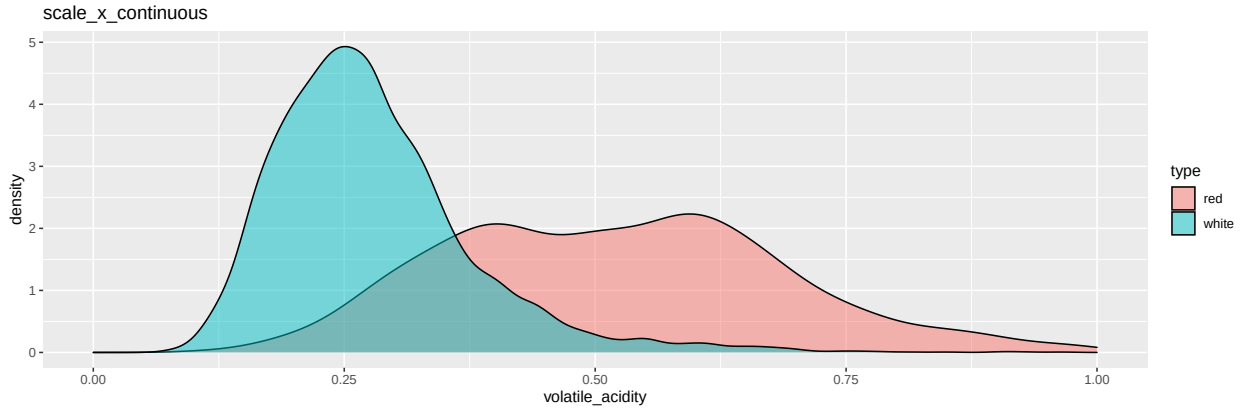
Fixed Acidity

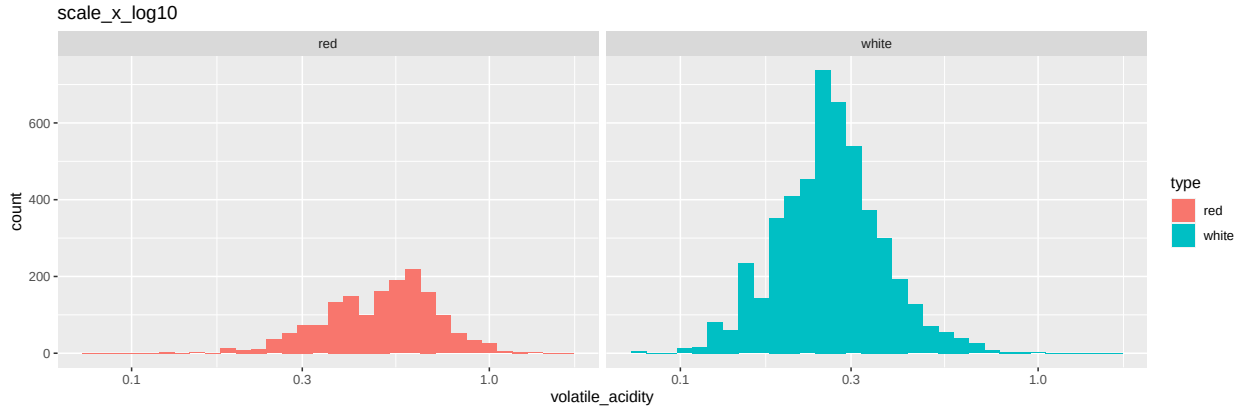
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.400   7.000   7.215   7.700  15.900
```

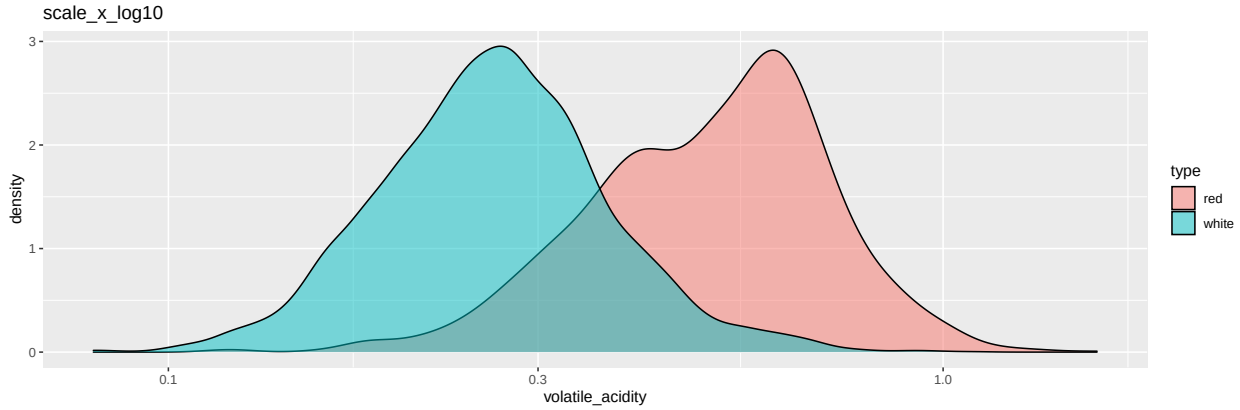
The plots for fixed acidity have a positive skew, but otherwise look fairly normal.

Volatile Acidity









```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2300  0.2900  0.3397  0.4000  1.5800
```

The volatile acidity distribution for white wine has a positive skew, and red wine has an interesting bimodal distribution that may warrant additional investigation.

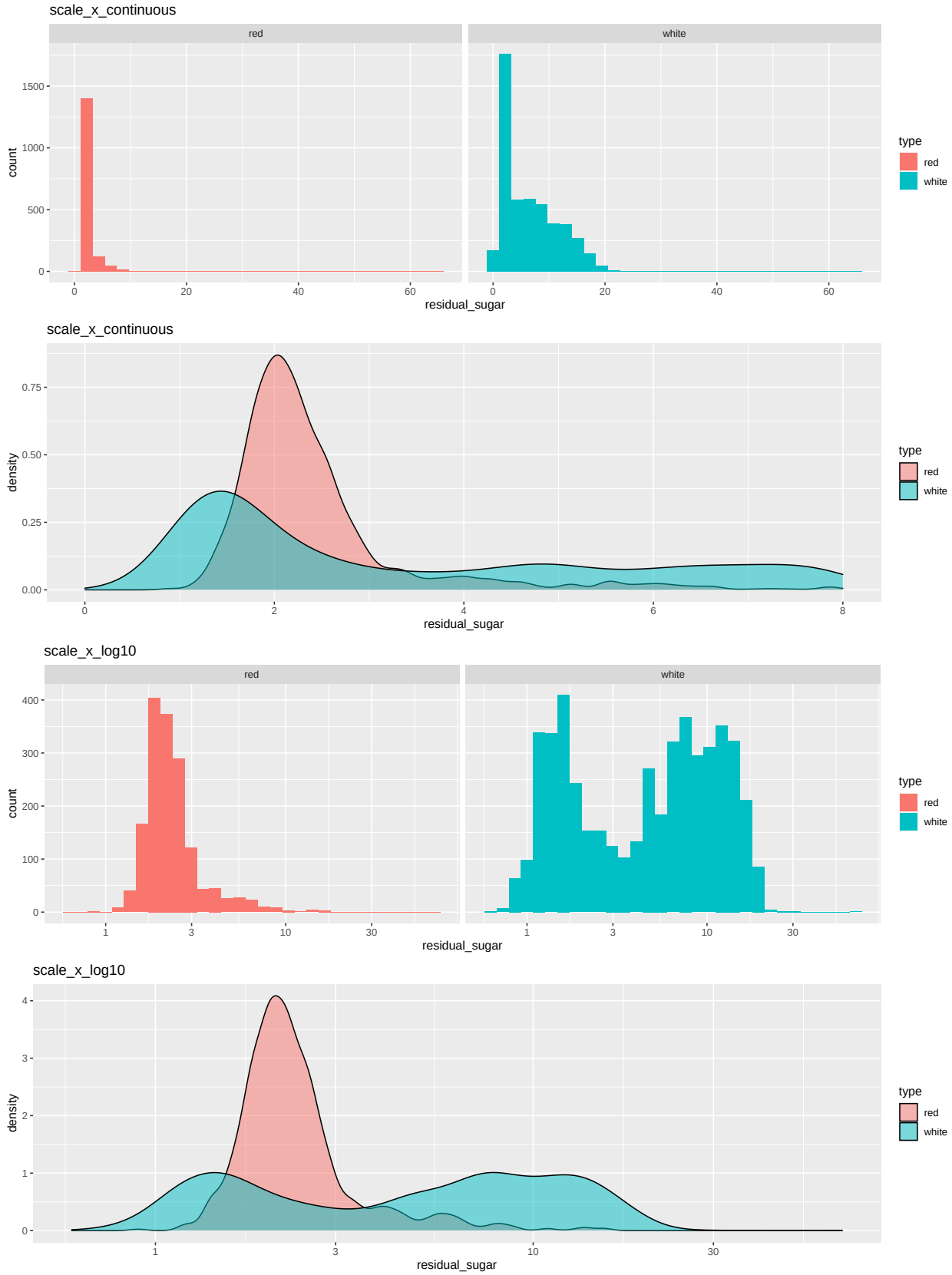
Citric Acid


scale_x_continuous
red
white
count
citric_acid
type
red
white
scale_x_continuous
density
citric_acid
type
red
white
scale_x_log10
red
white
count
citric_acid
type
red
white
scale_x_log10
density
citric_acid
type
red
white

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2500  0.3100  0.3186  0.3900  1.6600
```

Compared to white wine, the citric acid distribution for red wine is flat. The peaks in the distribution for white wine at 0.25, 0.5, and 0.75 are interesting. Plotting on a log10 scale created a more normal distribution, but the peaks observed in the white wine plot are still present.

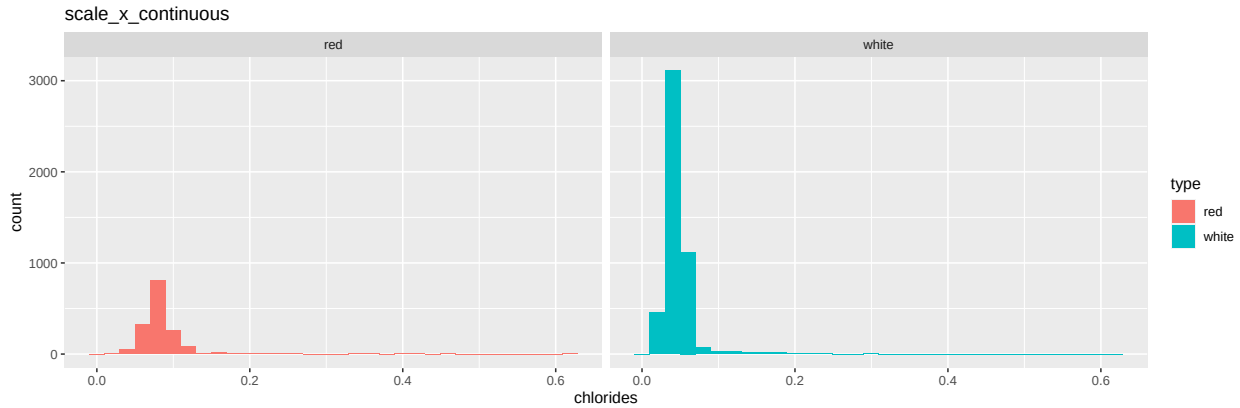
Residual Sugar

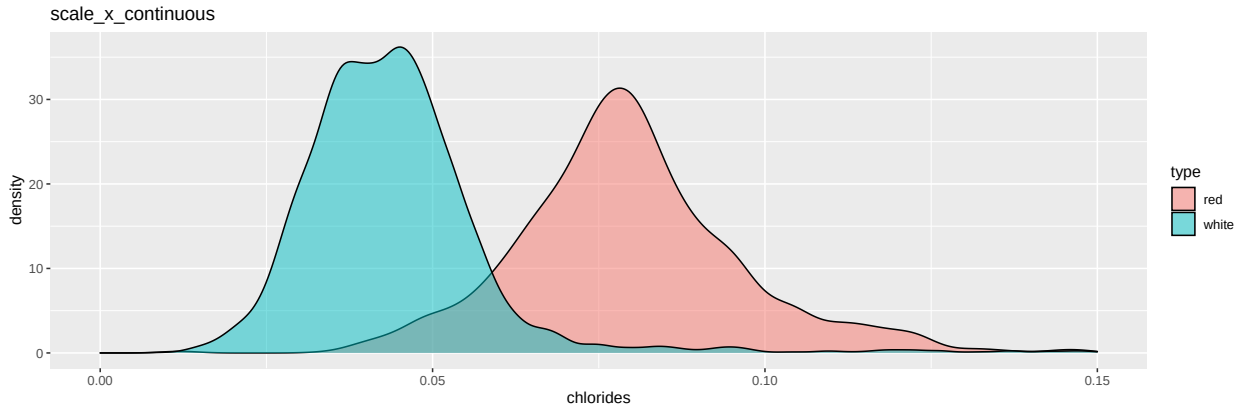


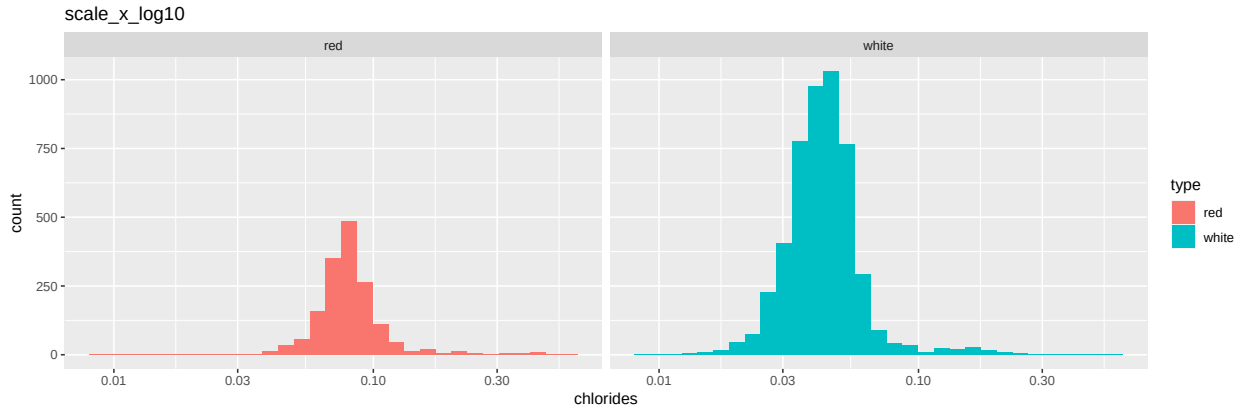
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.800   3.000   5.443   8.100  65.800
```

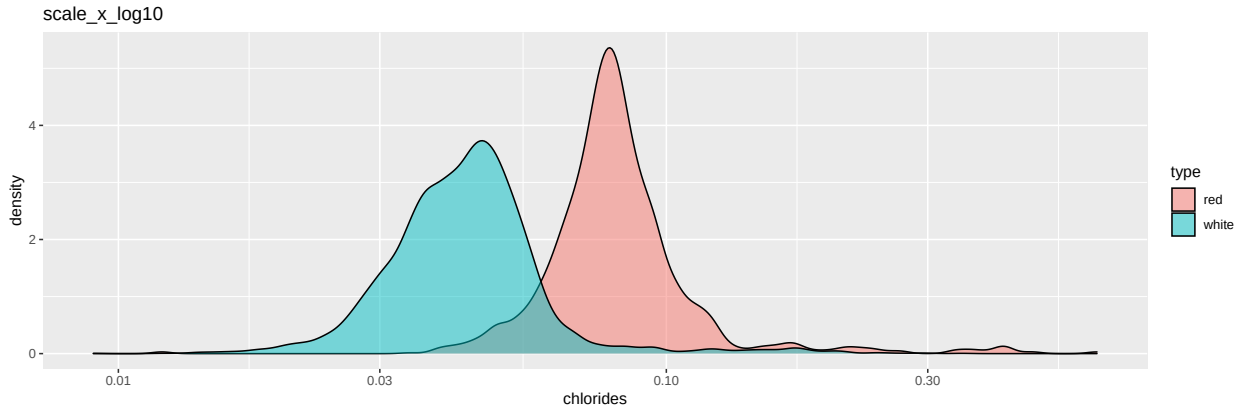
The distribution of residual sugar in white wine is very spread out and both types of wine have a positive skew. After plotting on a log10 scale, a bimodal distribution is apparent in the white wine sample.

Chlorides


scale_x_continuous
red
white
count
chlorides
type
red
white


scale_x_continuous
density
chlorides
type
red
white

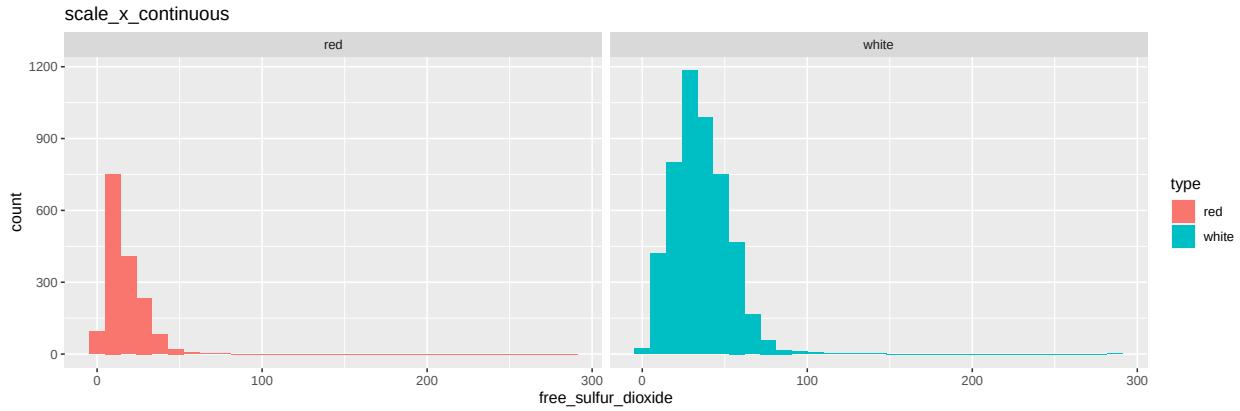

scale_x_log10
red
white
count
chlorides
type
red
white

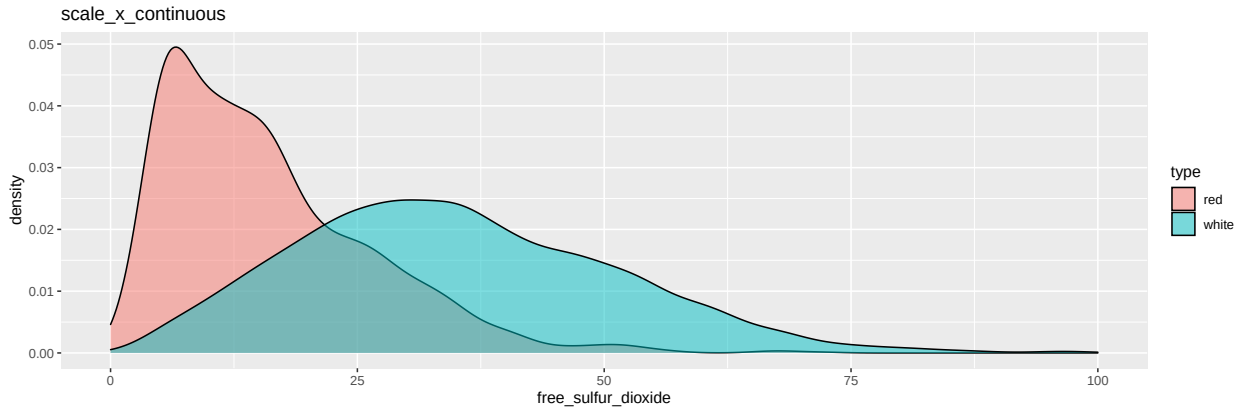

scale_x_log10
density
chlorides
type
red
white

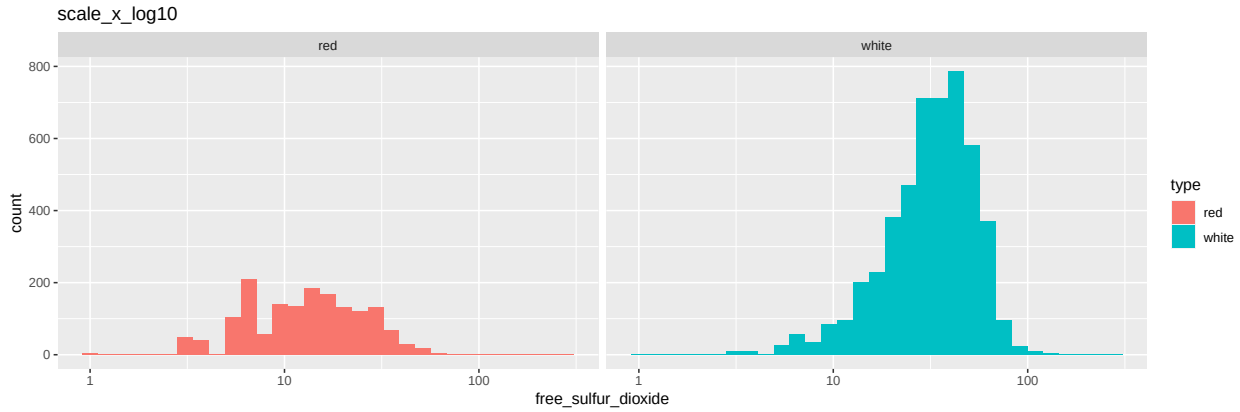
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03800 0.04700 0.05603 0.06500 0.61100
```

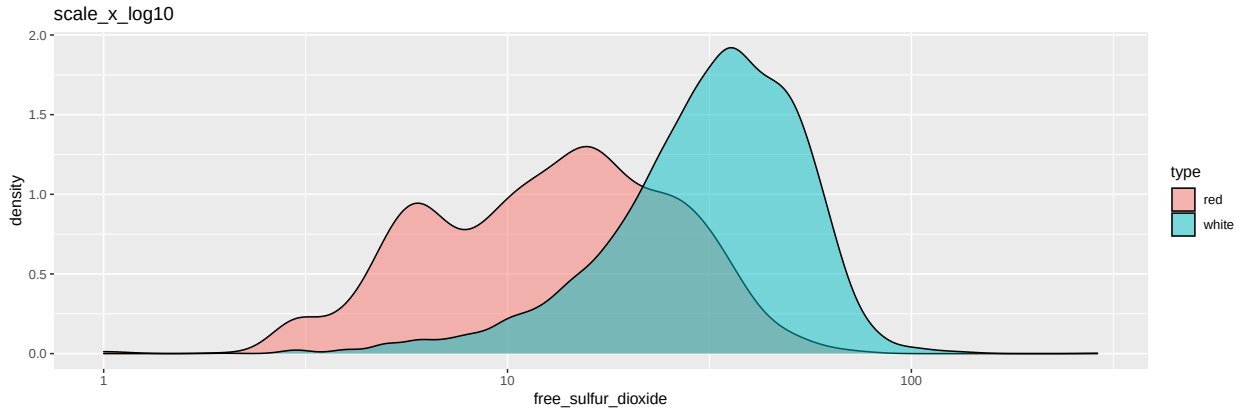
The density distribution for both white and red wine are quite normal.

Free Sulfur Dioxide





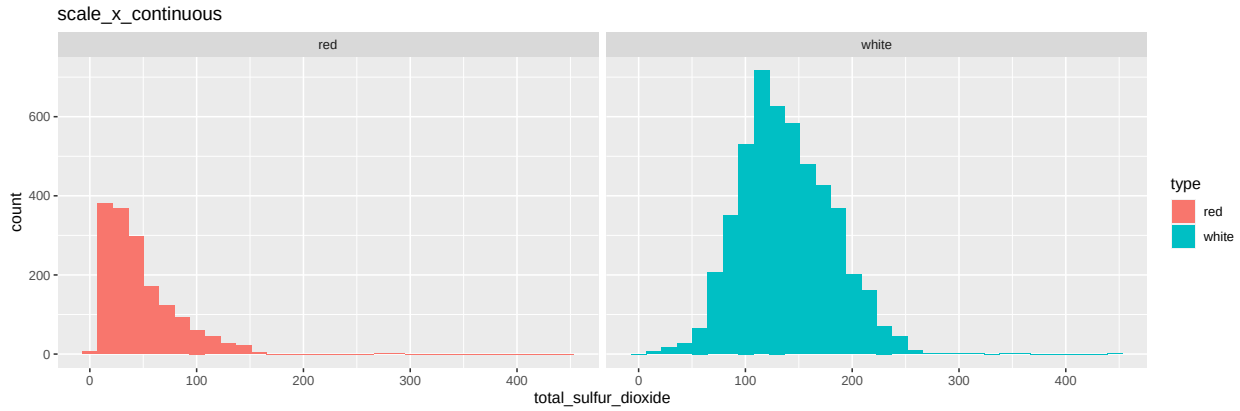


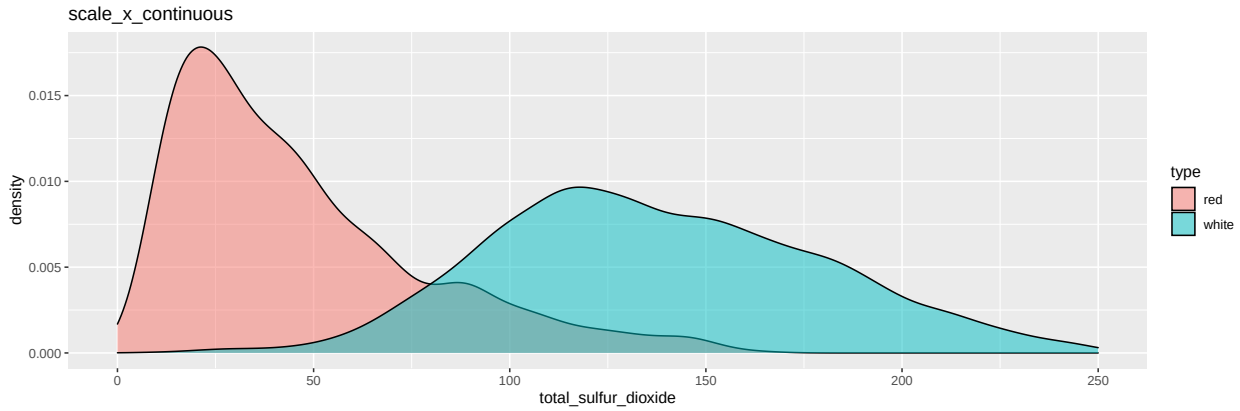


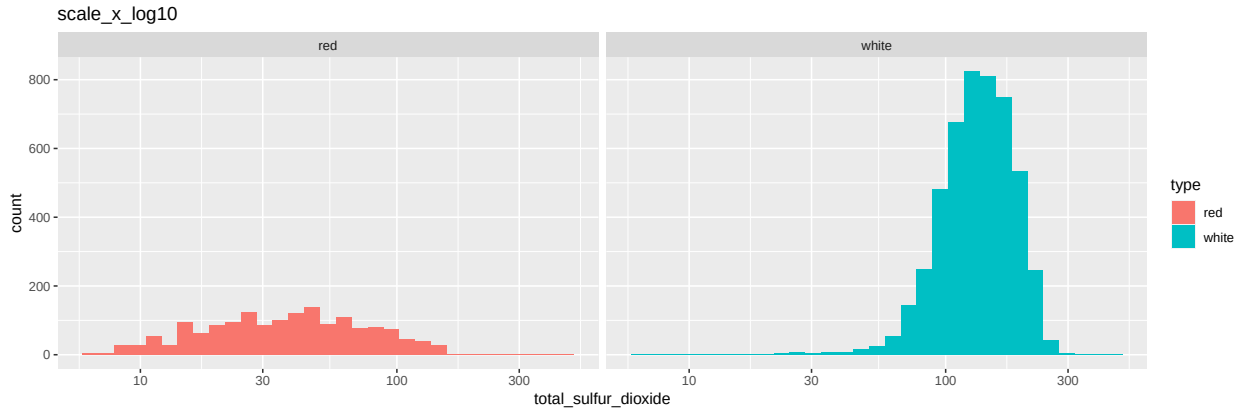
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   17.00   29.00   30.53   41.00  289.00
```

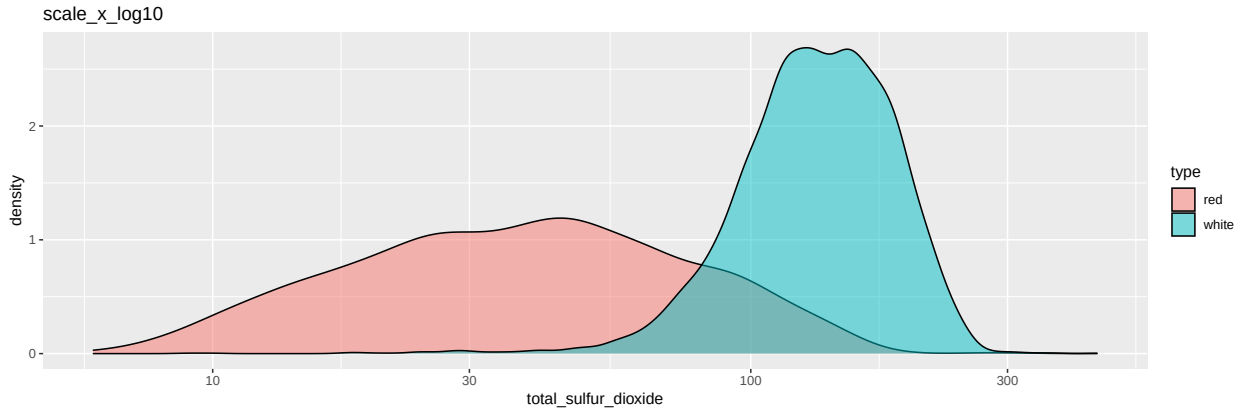
The free sulfur dioxide distribution for both wine types has a positive skew, but this trend is more pronounced in the red wine sample. Plotting on a log10 scale brings the plots closer to a normal distribution, but the distribution for white wine has a negative skew, while the red wine sample has several peaks and valleys.

Total Sulfur Dioxide





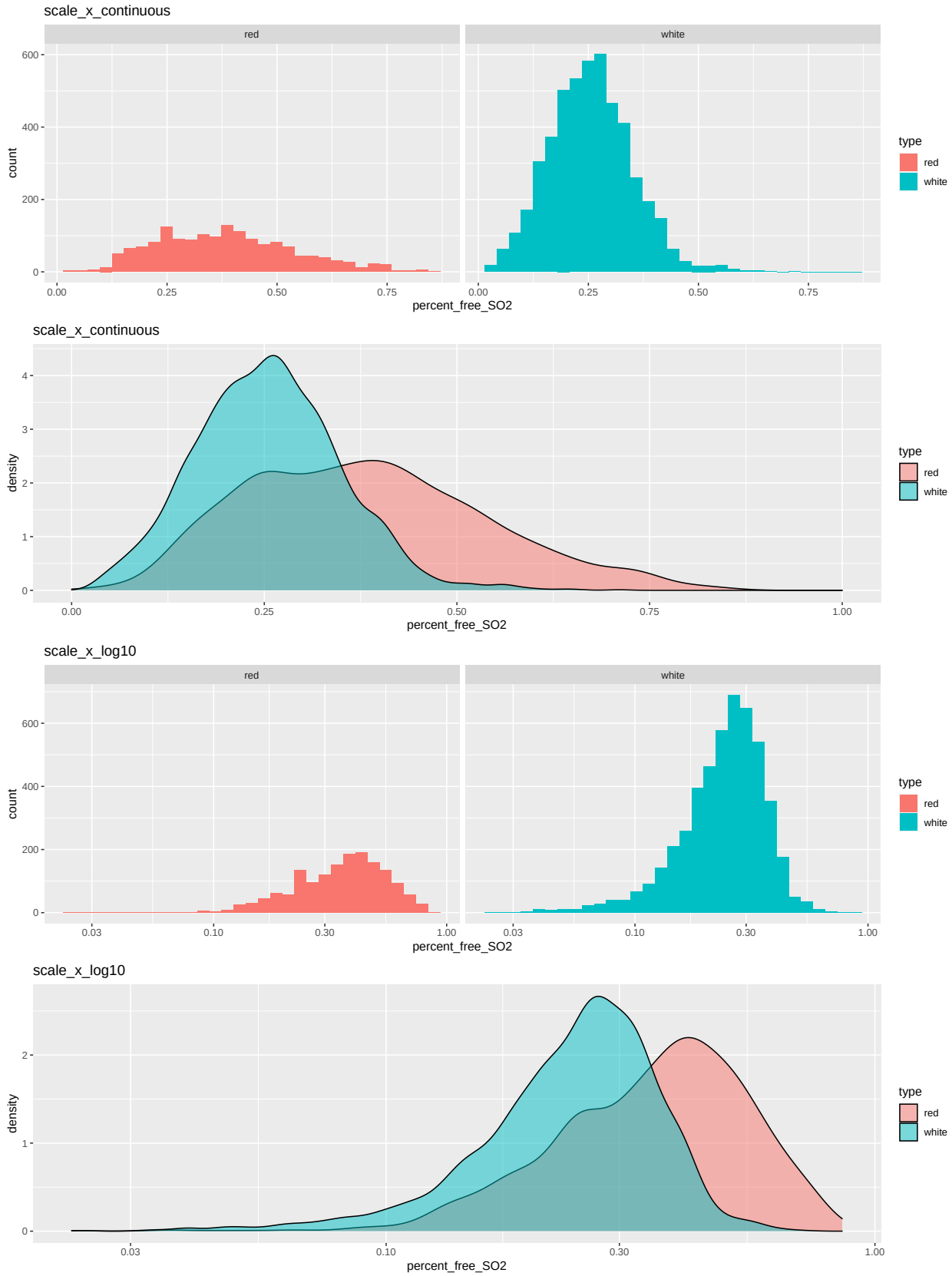




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     6.0    77.0   118.0   115.7   156.0   440.0
```

The total sulfur dioxide distribution for both types of wine has a positive skew, but plotting on a log10 scale largely fixes this.

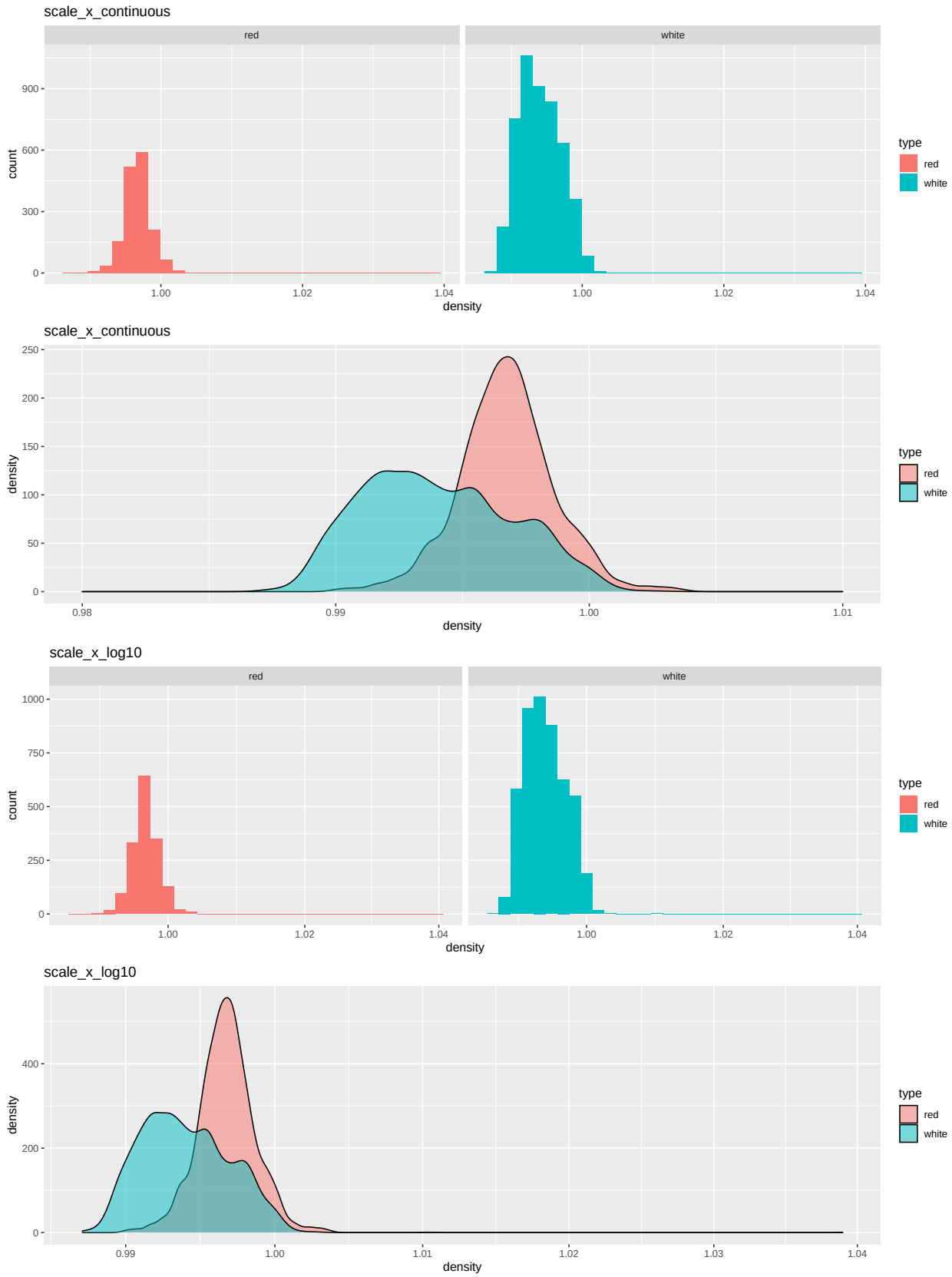
Percent Free Sulfur Dioxide



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02273 0.20207 0.26977 0.28677 0.34884 0.85714
```

As suspected, combining free sulfur dioxide and total sulfur dioxide into a new comparative variable produced a more normal distribution. This variable is not improved by the log10 plots.

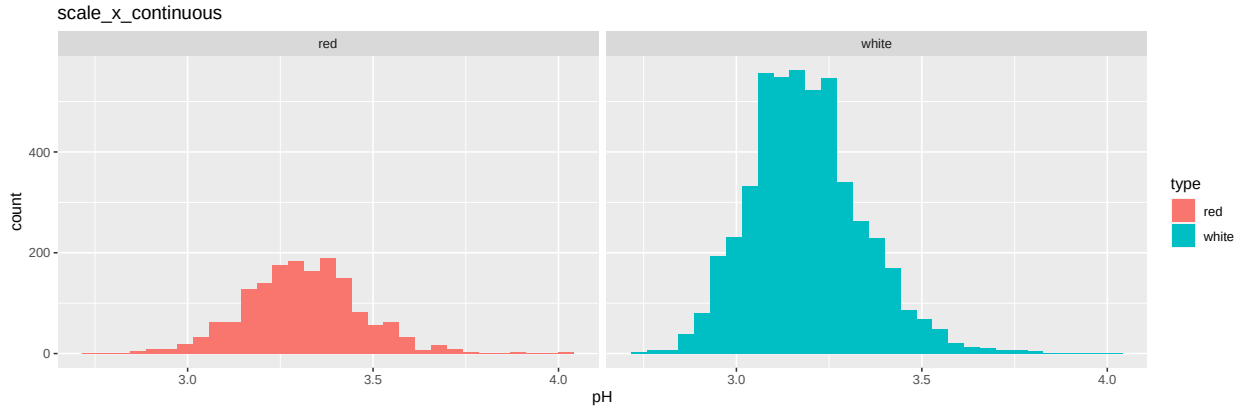
Density

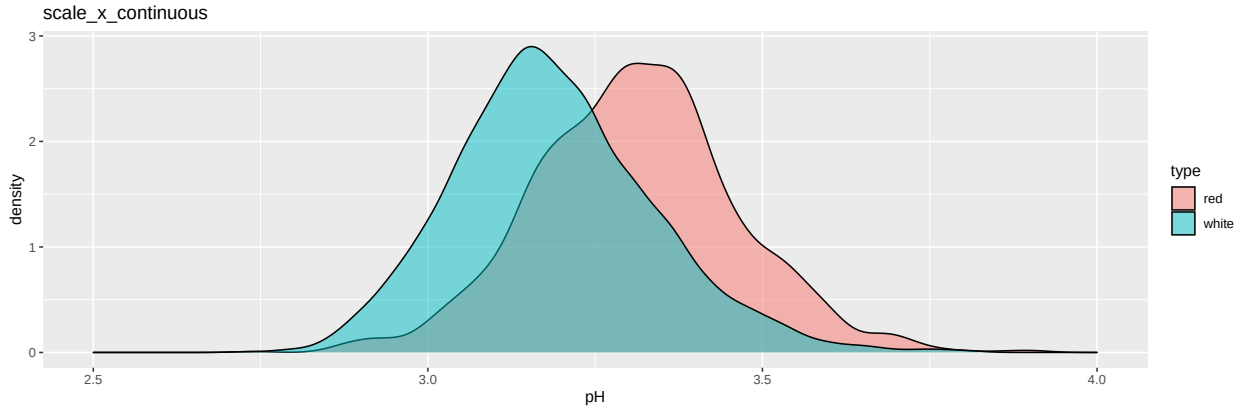

scale_x_continuous
red
white
count
density
type
red
white
scale_x_continuous
density
density
type
red
white
scale_x_log10
red
white
count
density
type
red
white
scale_x_log10
density
density
type
red
white

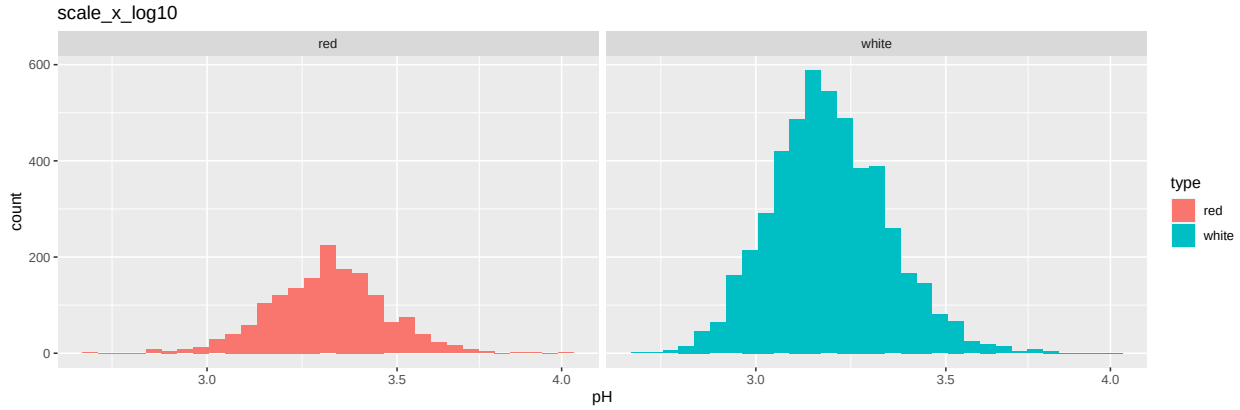
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9923  0.9949  0.9947  0.9970  1.0390
```

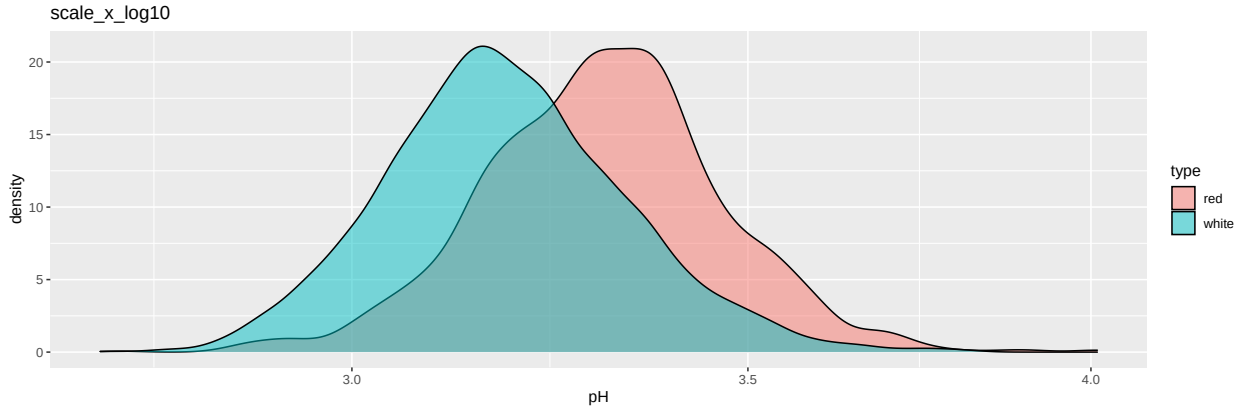
The distribution of the density variable is close to normal, but the white wine sample has a positive skew.

pH


scale_x_continuous
red
white
count
400
200
0
3.0
3.5
4.0
pH
type
red
white


scale_x_continuous
density
3
2
1
0
2.5
3.0
3.5
4.0
pH
type
red
white

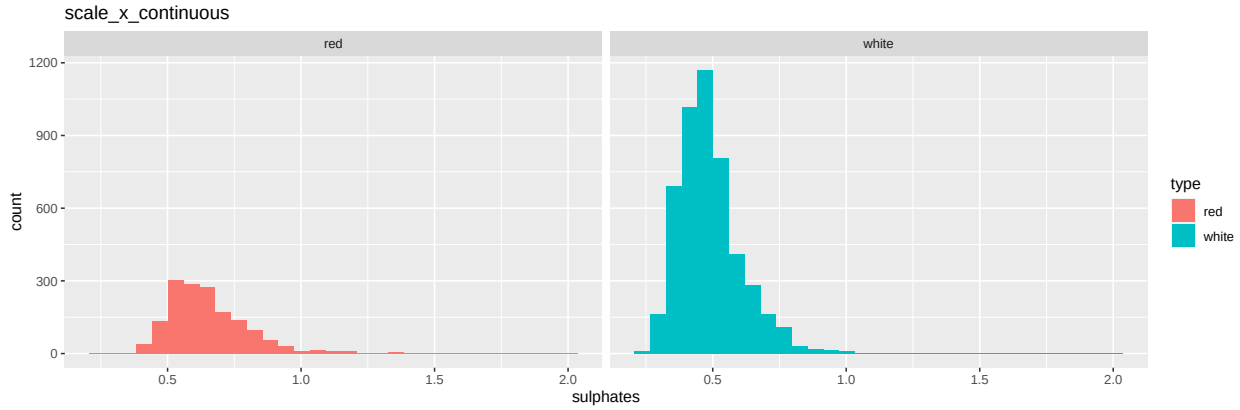

scale_x_log10
red
white
count
600
400
200
0
3.0
3.5
4.0
pH
type
red
white

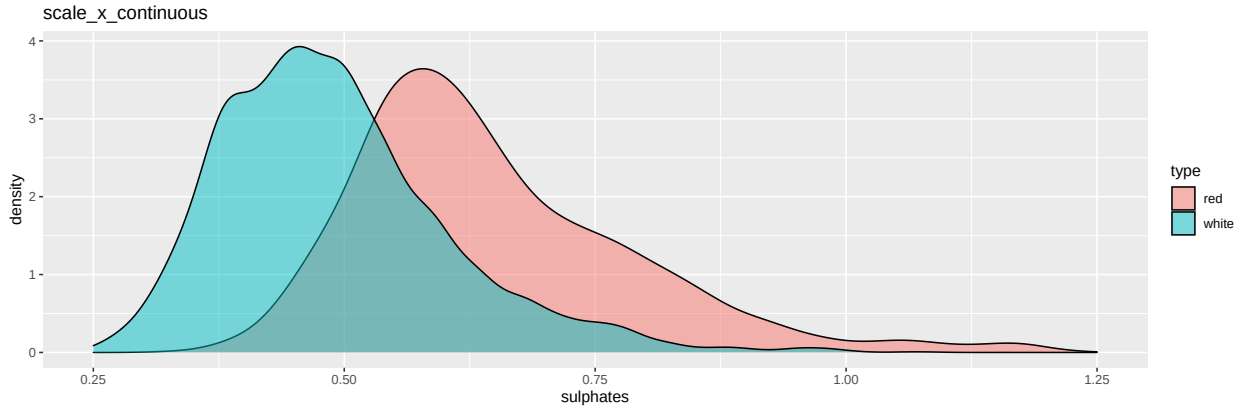

scale_x_log10
density
20
15
10
5
0
3.0
3.5
4.0
pH
type
red
white

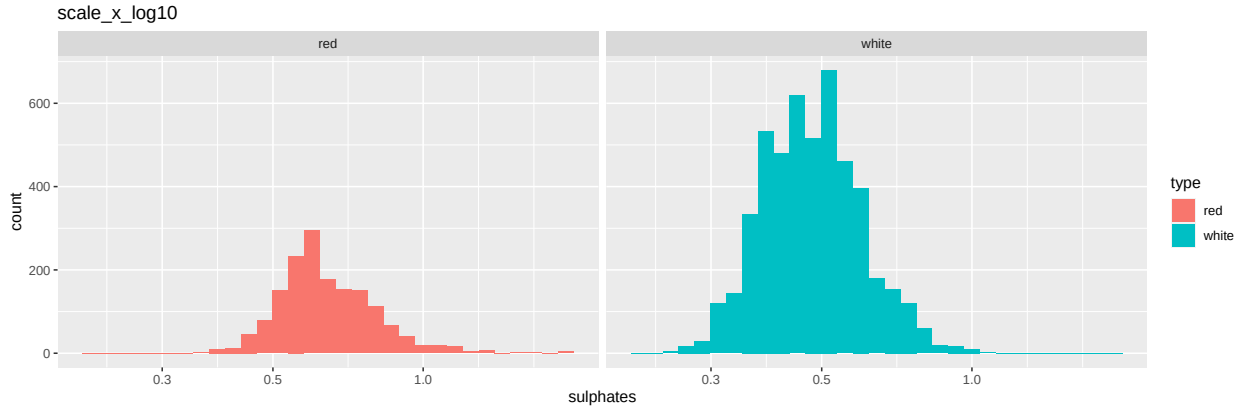
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.110   3.210   3.219   3.320   4.010
```

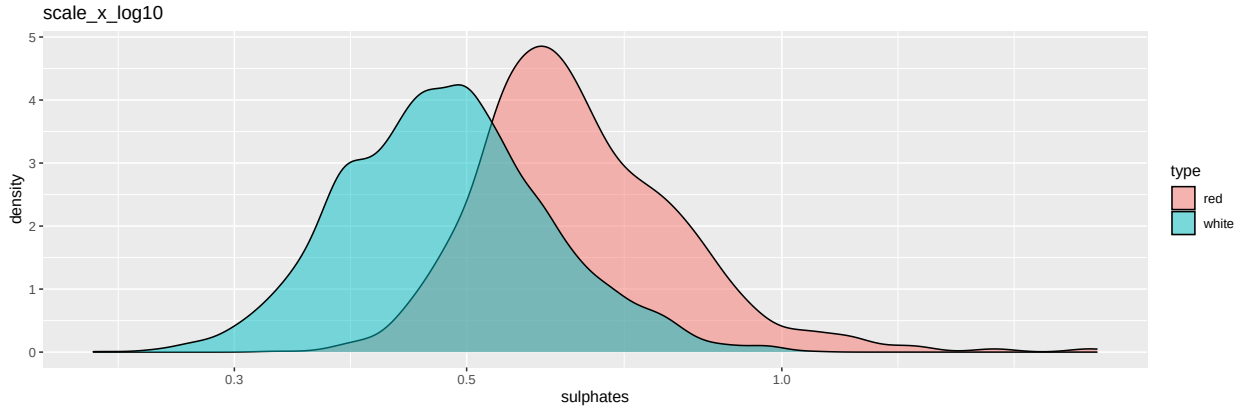
These plots are close to perfectly normal, but I would expect a measurement like pH to have a normal distribution since it's a logarithmic function (https://en.wikipedia.org/wiki/PH).

Sulphates


scale_x_continuous
red
white
count
sulphates
type
red
white


scale_x_continuous
density
sulphates
type
red
white

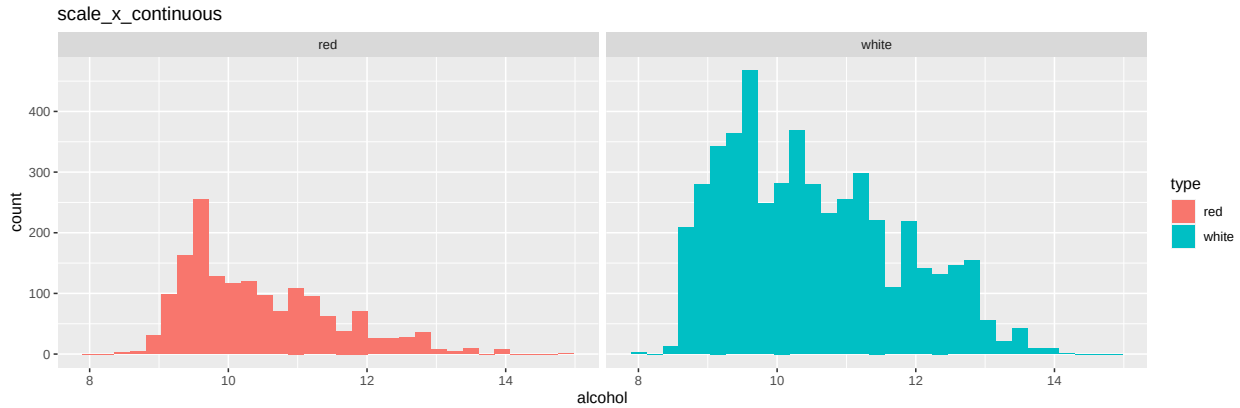

scale_x_log10
red
white
count
sulphates
type
red
white

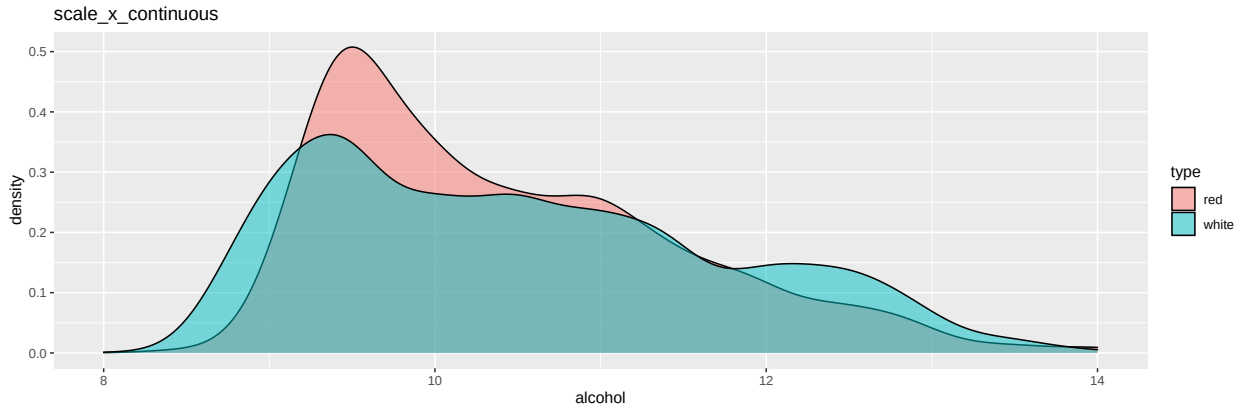

scale_x_log10
density
sulphates
type
red
white

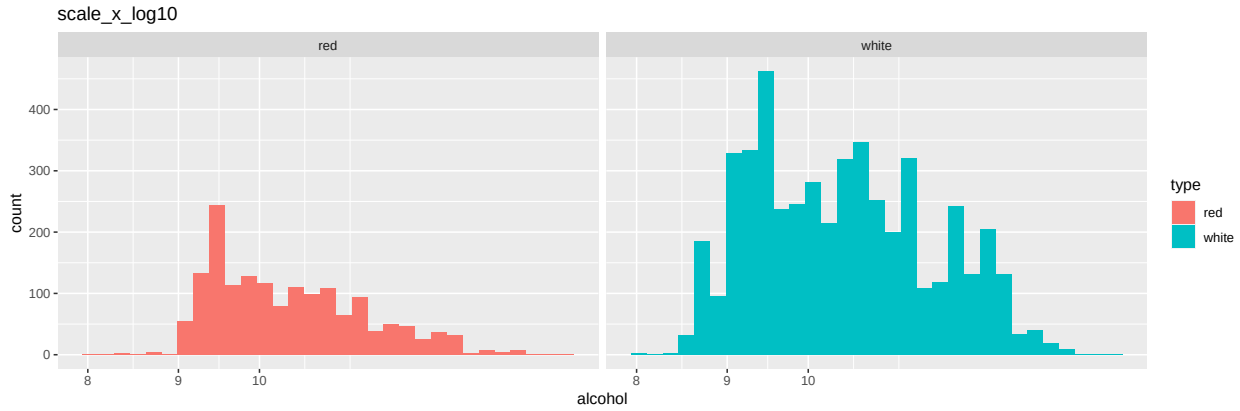
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4300  0.5100  0.5313  0.6000  2.0000
```

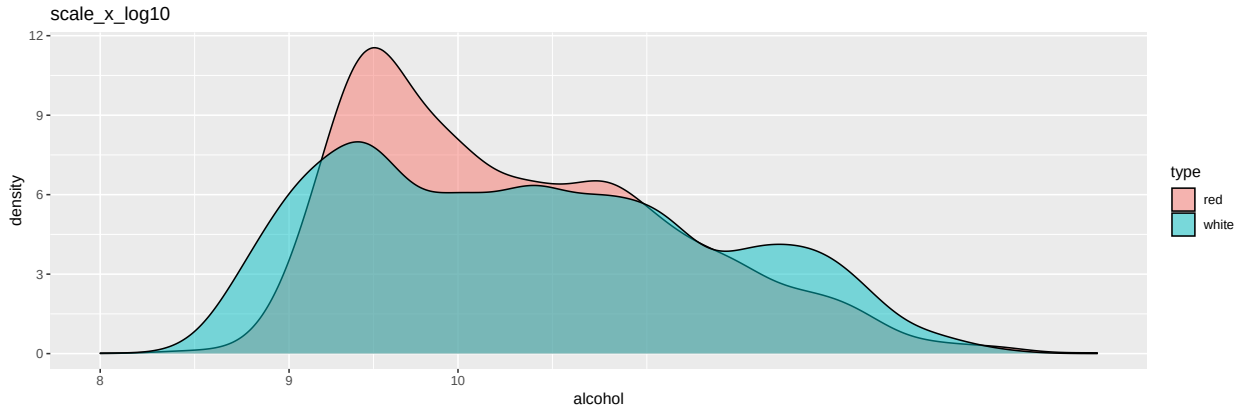
The distribution of sulphates for both types of wine have a positive skew. Plotting them on a log10 scale brings them much closer to a normal distribution.

Alcohol


scale_x_continuous
red
white
count
alcohol
type
red
white


scale_x_continuous
density
alcohol
type
red
white


scale_x_log10
red
white
count
alcohol
type
red
white


scale_x_log10
density
alcohol
type
red
white

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.30   10.49   11.30   14.90
```

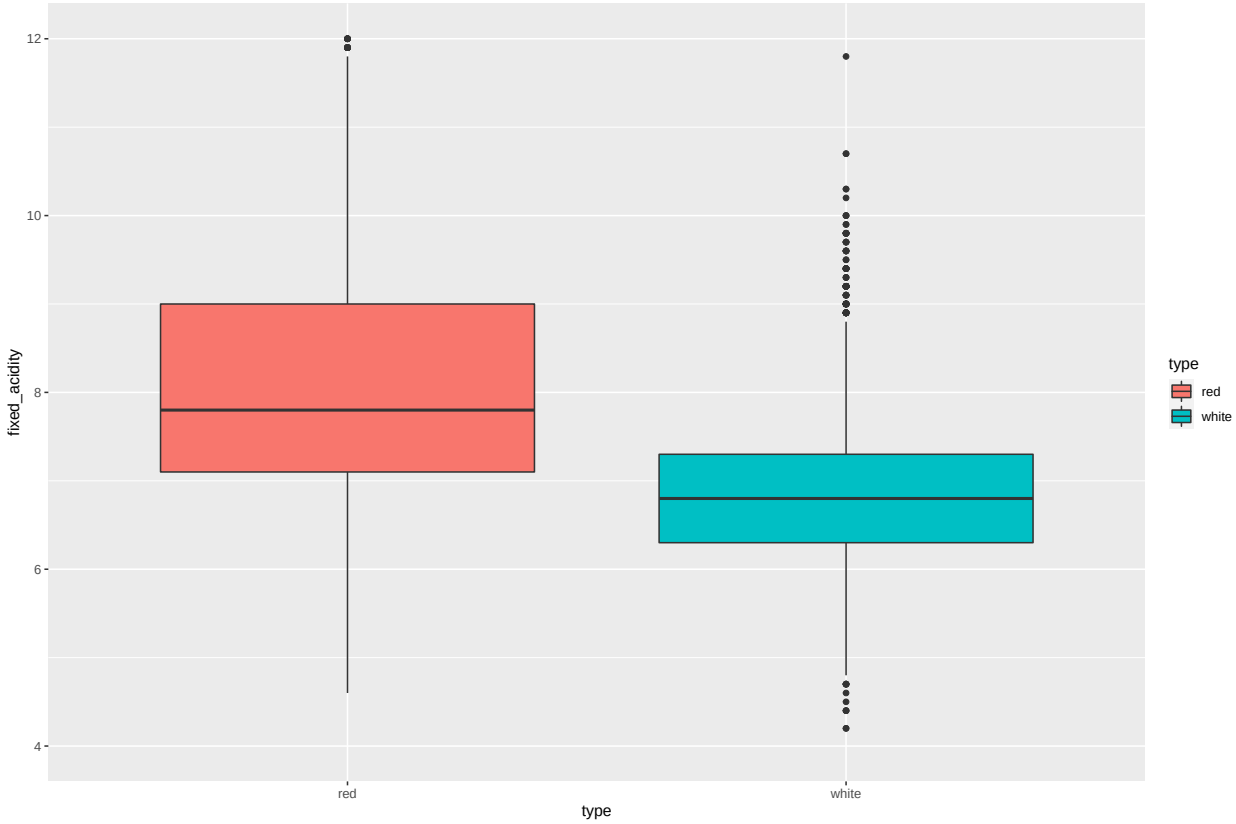
The distribution of alcohol content in both red and white wine have a very pronounced positive skew, and plotting them on a log10 scale doesn't do much to normalize them. This variable definitely needs additional analysis.

Boxplots

Next, I'll create a boxplot for each variable so I can quickly visualize mean values, the distribution of each variable within the dataset, and any outliers in the data. Just like I did for the histogram/density plots, I'll leverage a function to reduce redundant code.

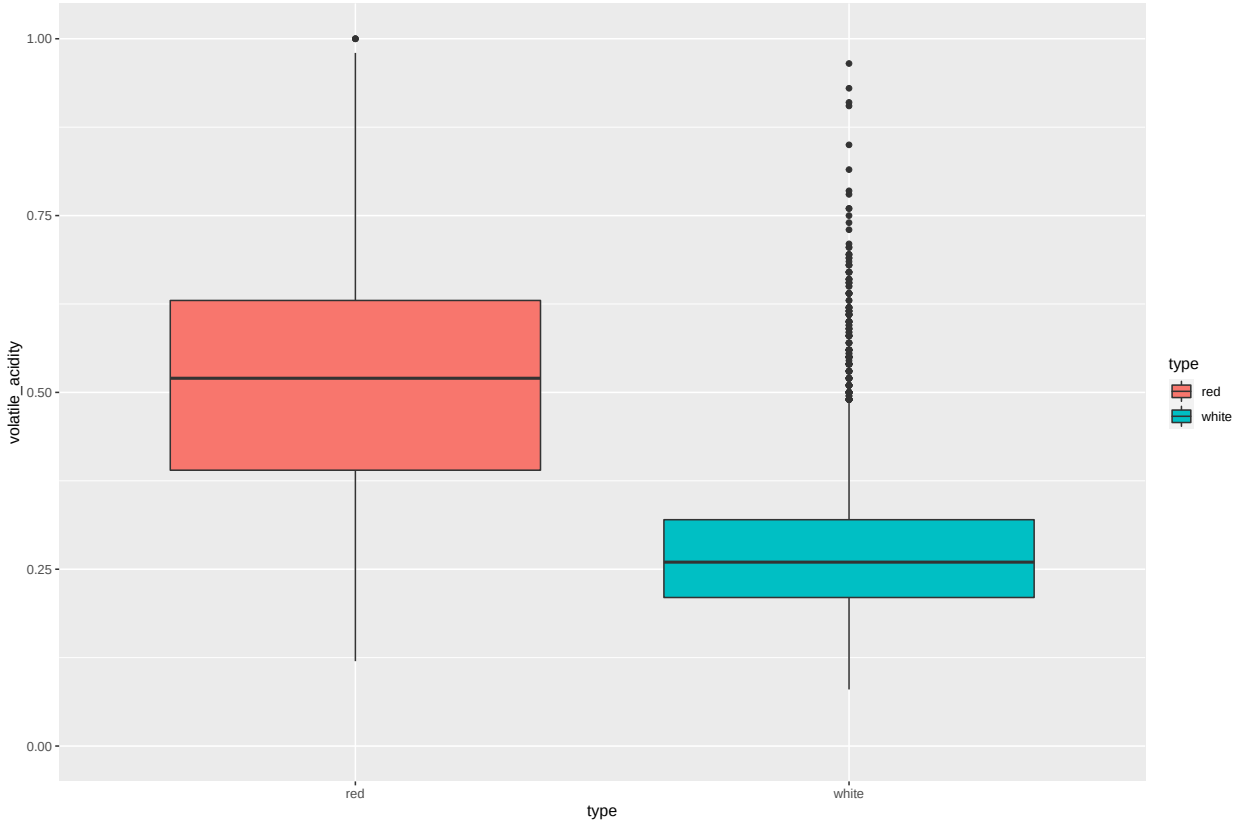
```
PLOT_BOXPLOT <- function(strVar, limStart=99, limEnd=99)
{
    # create vector based on string variable input
    winesVector <- unlist(winesAll[c(strVar)], use.names = FALSE)

    # adjust default limit start
    if (limStart == 99) {
        limStart <- min(winesVector)
    }

    # adjust default limit end
    if (limEnd == 99) {
        limEnd <- max(winesVector)
    }

    # create boxplot
    ggplot(winesAll) +
        aes(x = type, y = winesVector, fill = type) +
        geom_boxplot() +
        scale_y_continuous(limits = c(limStart, limEnd)) +
        labs(x = 'type', y = strVar)
}
```

Fixed Acidity



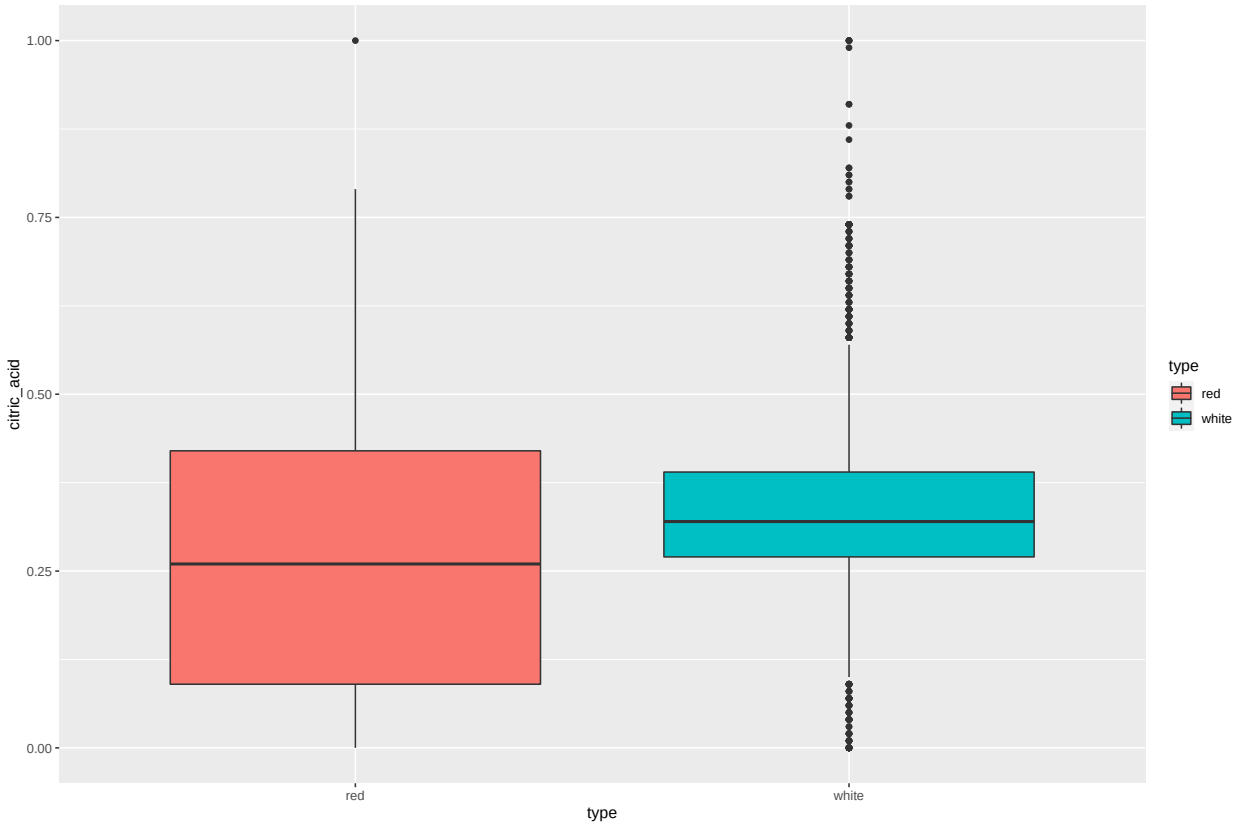
The mean fixed acidity is higher in the red wines sample than the white wines sample, and there are more outliers in the white wines sample.

Volatile Acidity



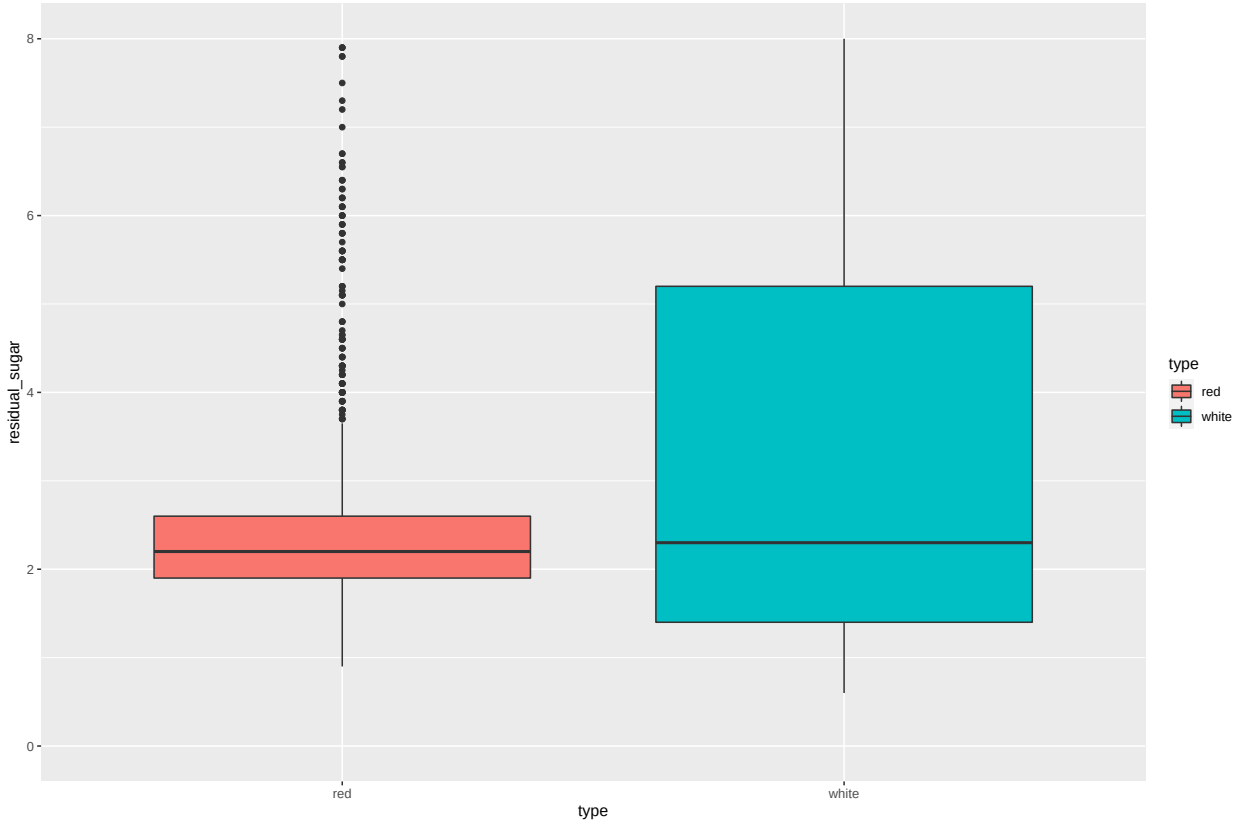
The mean volatile acidity is higher in the red wines sample than the white wines sample, and there are a lot of outliers above the mean in the white wines sample.

Citric Acid



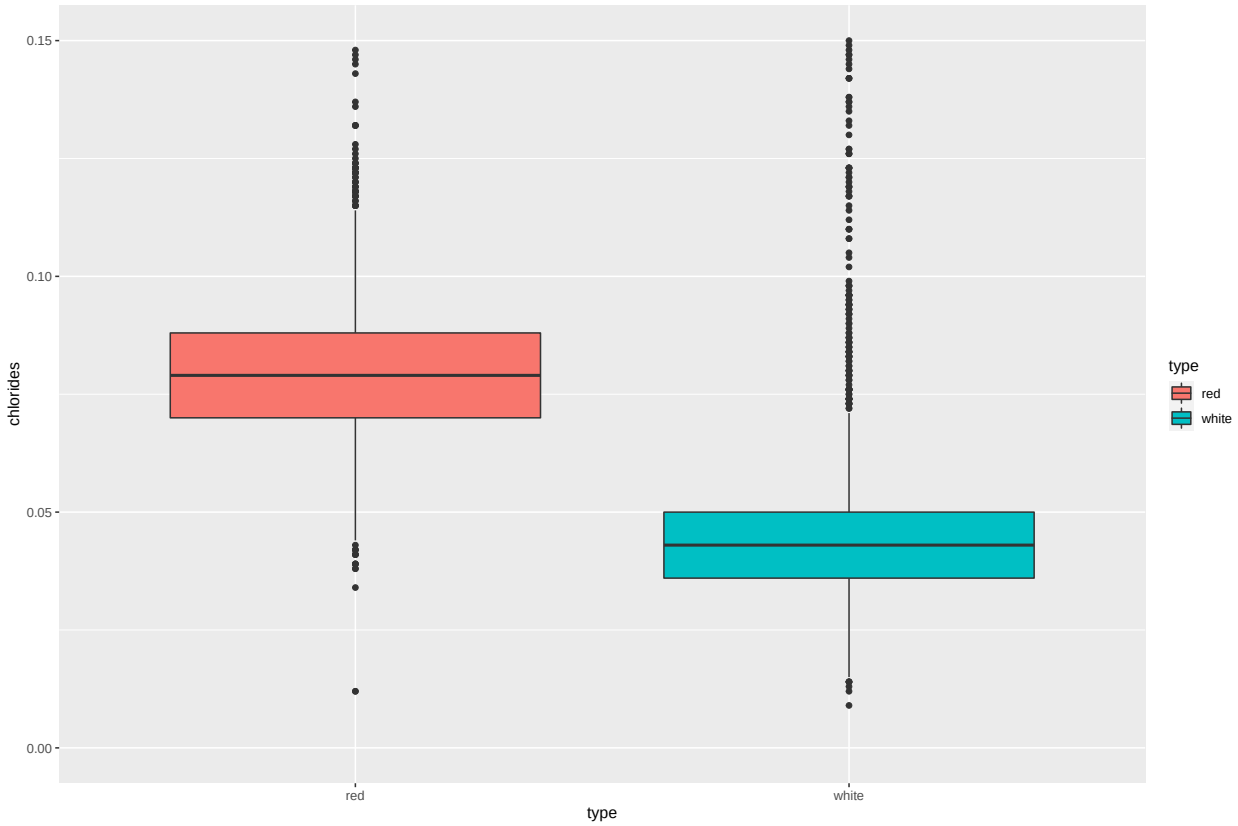
The means are pretty close for citric acid when comparing the two samples, but is slightly higher in the white wines sample. One possible emerging trend is the presence of outliers in the white wines sample that aren't evident in the red wines sample.

Residual Sugar



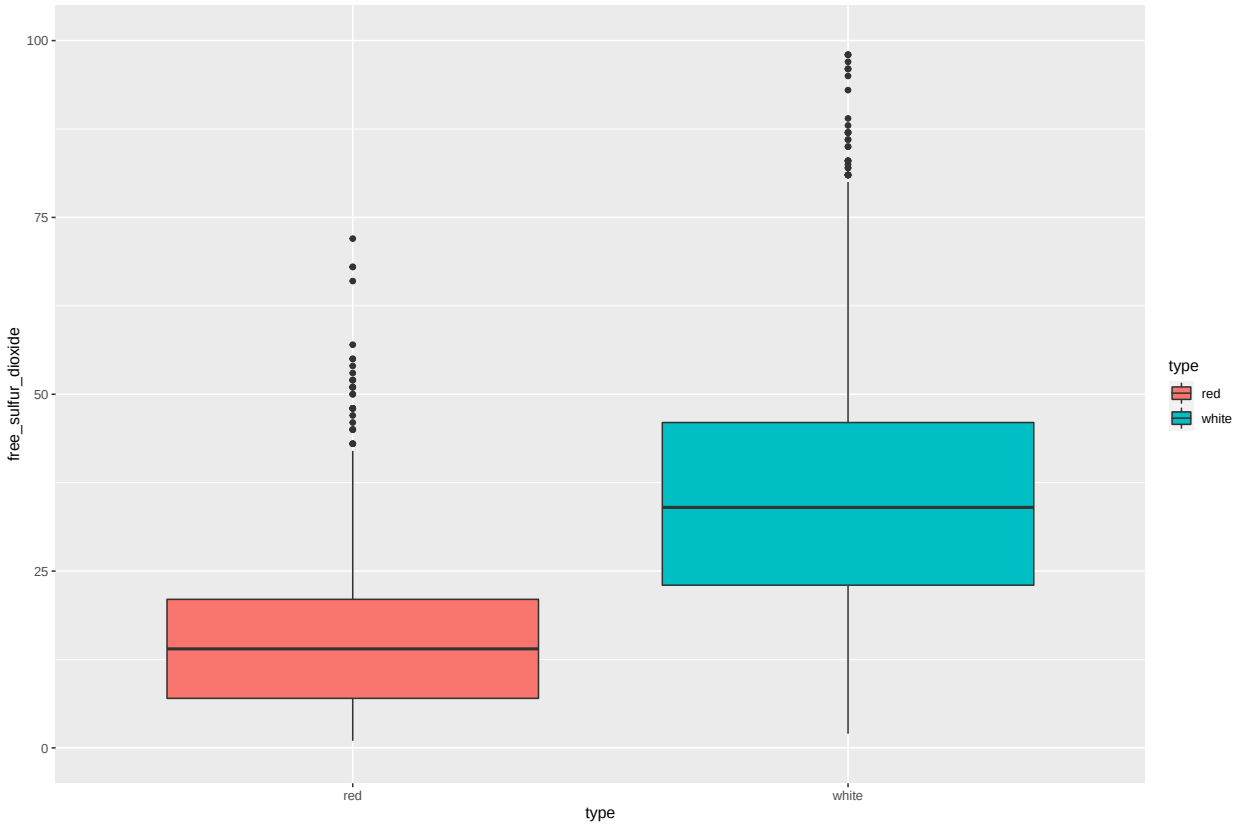
The mean residual sugar for both samples is about the same when comparing both types of wine. However, the white wines dataset has no major outlier events, but the red wines sample has a number of outliers above the mean. I was quickly dissuaded of my idea that the white wines sample contained outliers while the red wines dataset did not. This is a good example of why you plot all the variables.

Chlorides



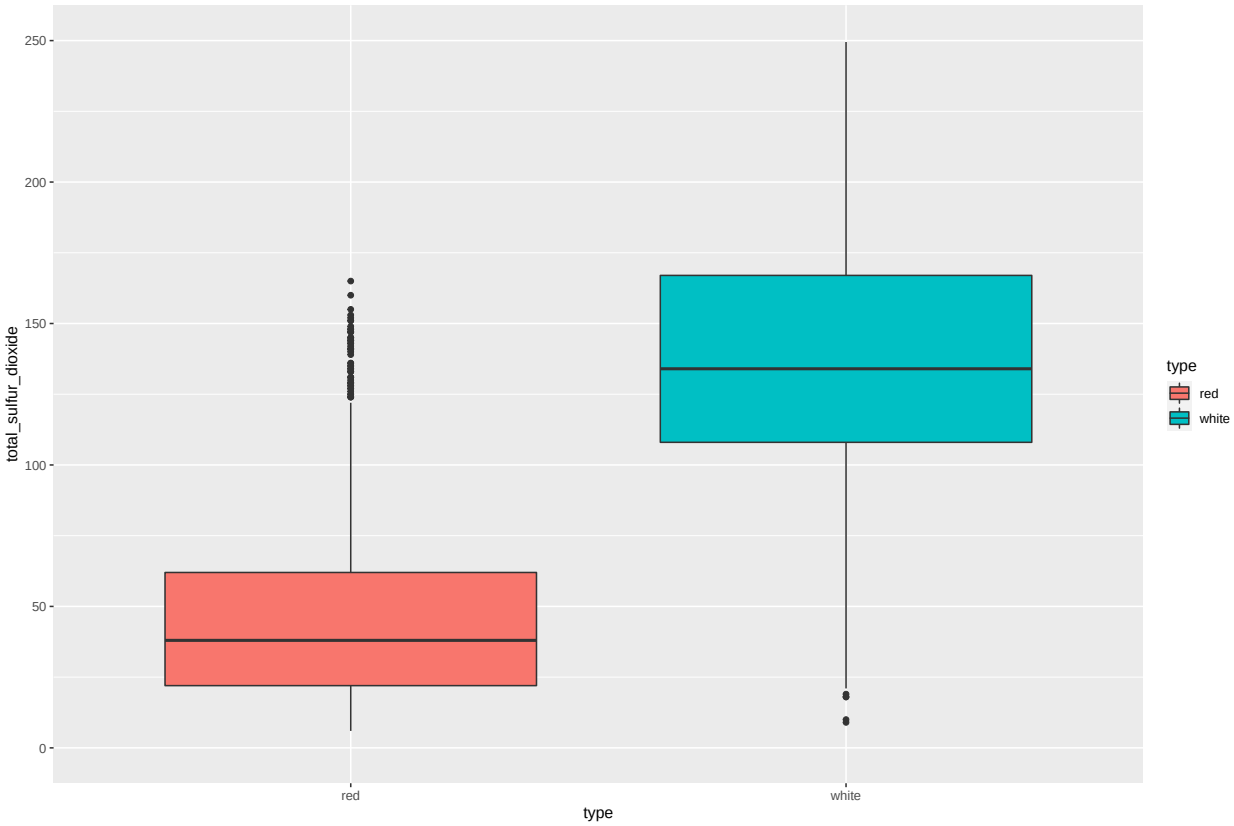
There are outliers above and below the mean for both sample, and the mean chlorides measurement is lower in the white wines dataset compared to the red wines sample.

Free Sulfur Dioxide



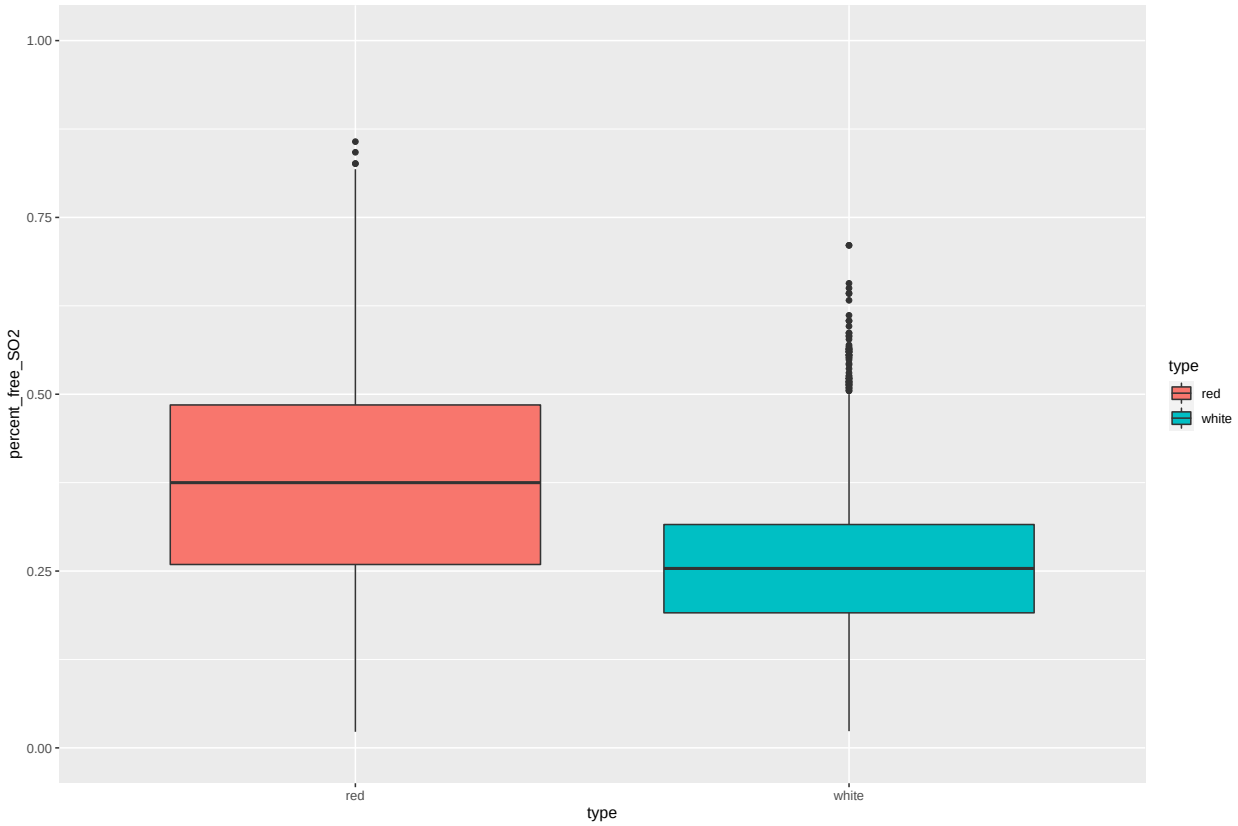
The mean free sulfur dioxide measured in the red wines sample is lower than in the white wines sample, and there are outliers above the means for both wine types.

Total Sulfur Dioxide



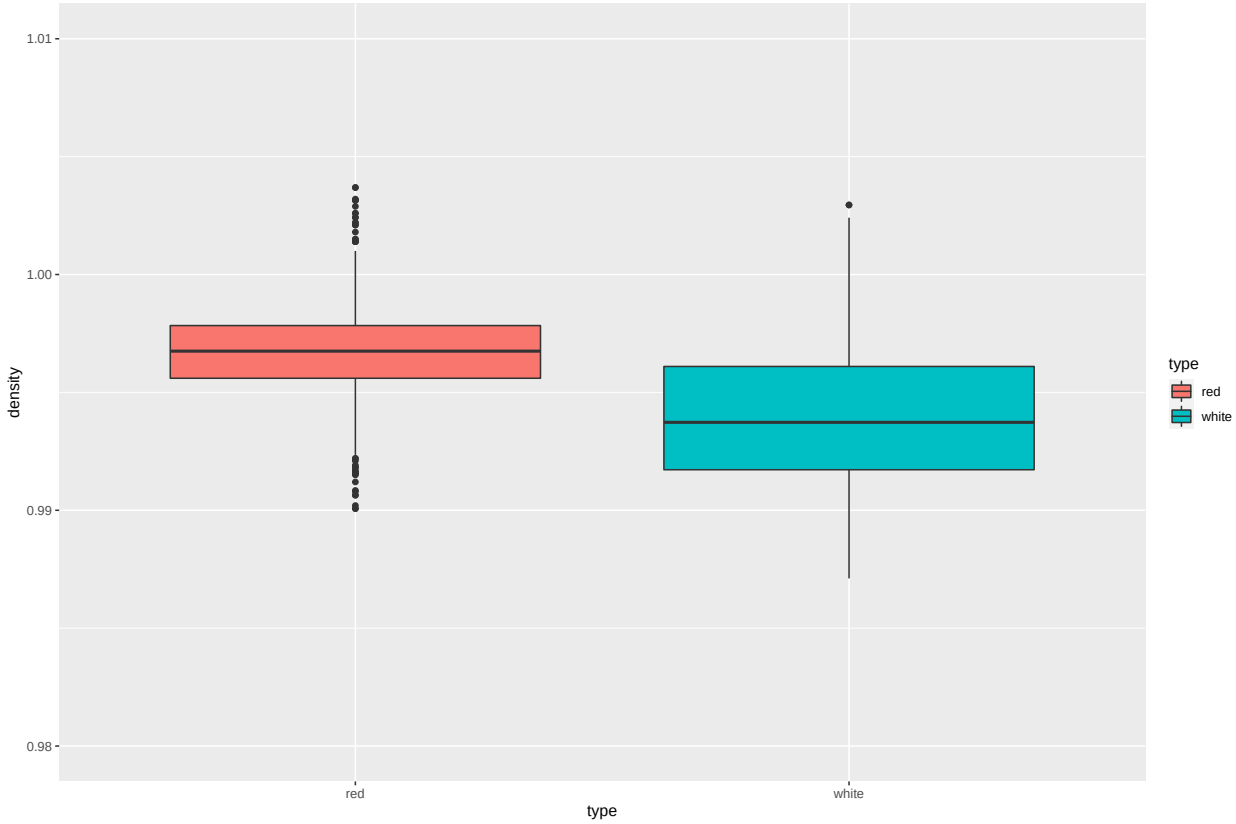
The mean total sulfur dioxide measured in the white wines sample is higher than in the red wines sample. The red wines sample has a number of outliers above the mean, and the white wines sample has a few outliers below the mean.

Percent Free Sulfur Dioxide



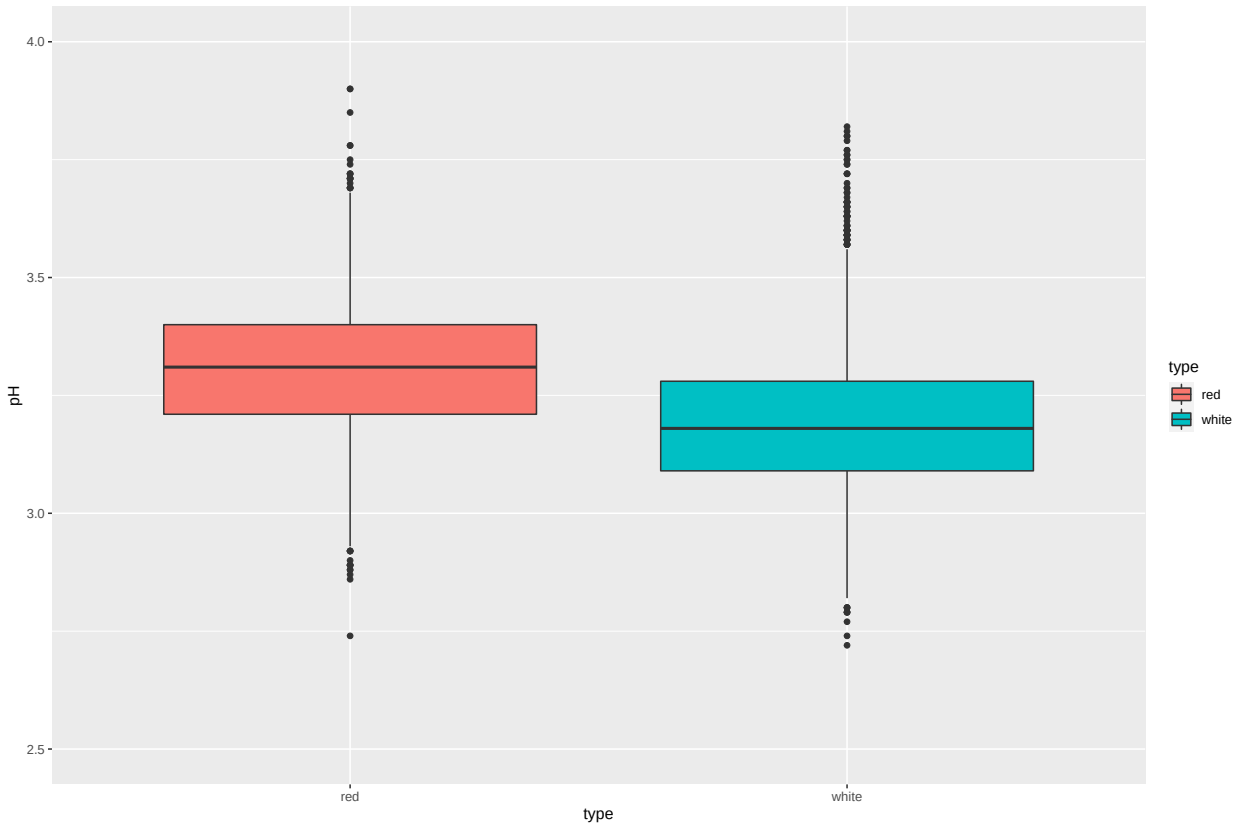
There are a number of outliers above the mean for both wine types, and the mean for the red wines sample is higher than the mean for the white wines sample.

Density



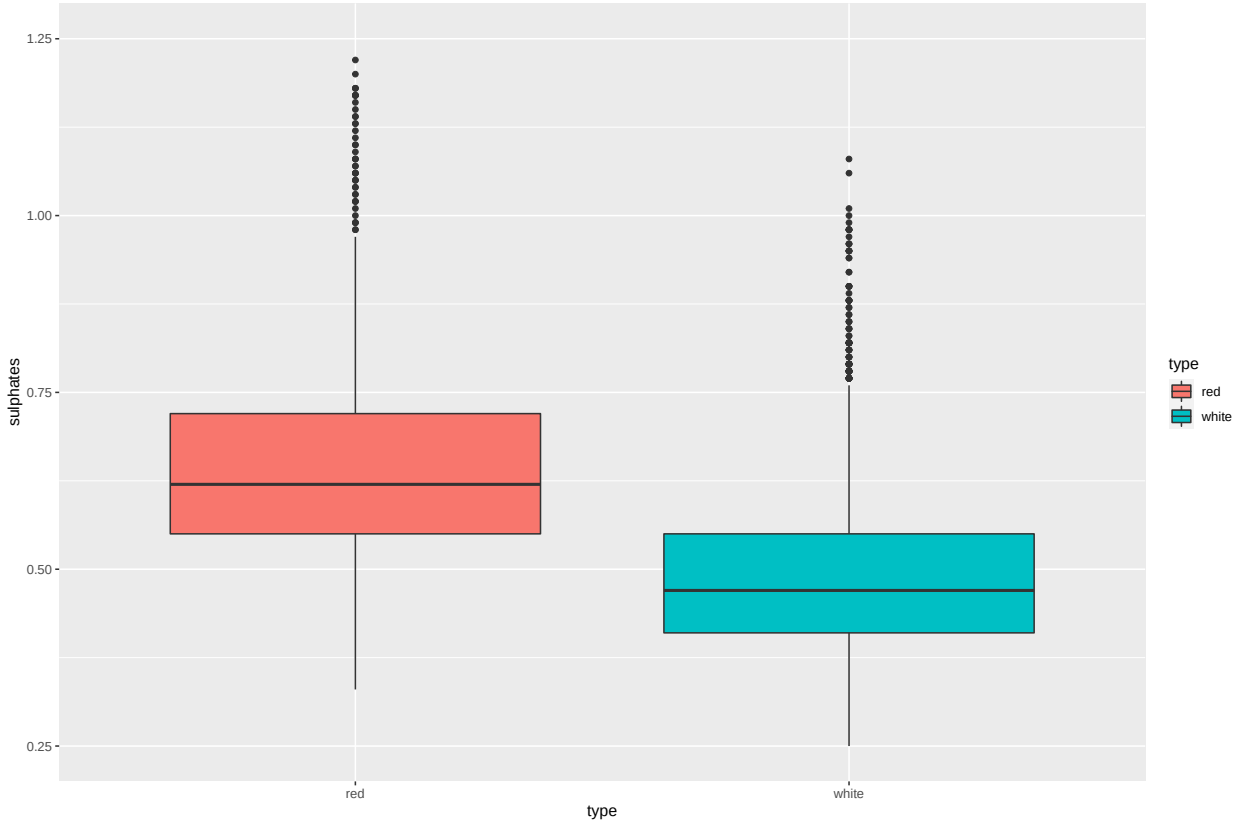
The mean density for the red wines sample is higher than that of the white wines sample. There are outliers above and below the mean in the red wines sample, and one lone outlier above the mean in the white wines sample.

pH



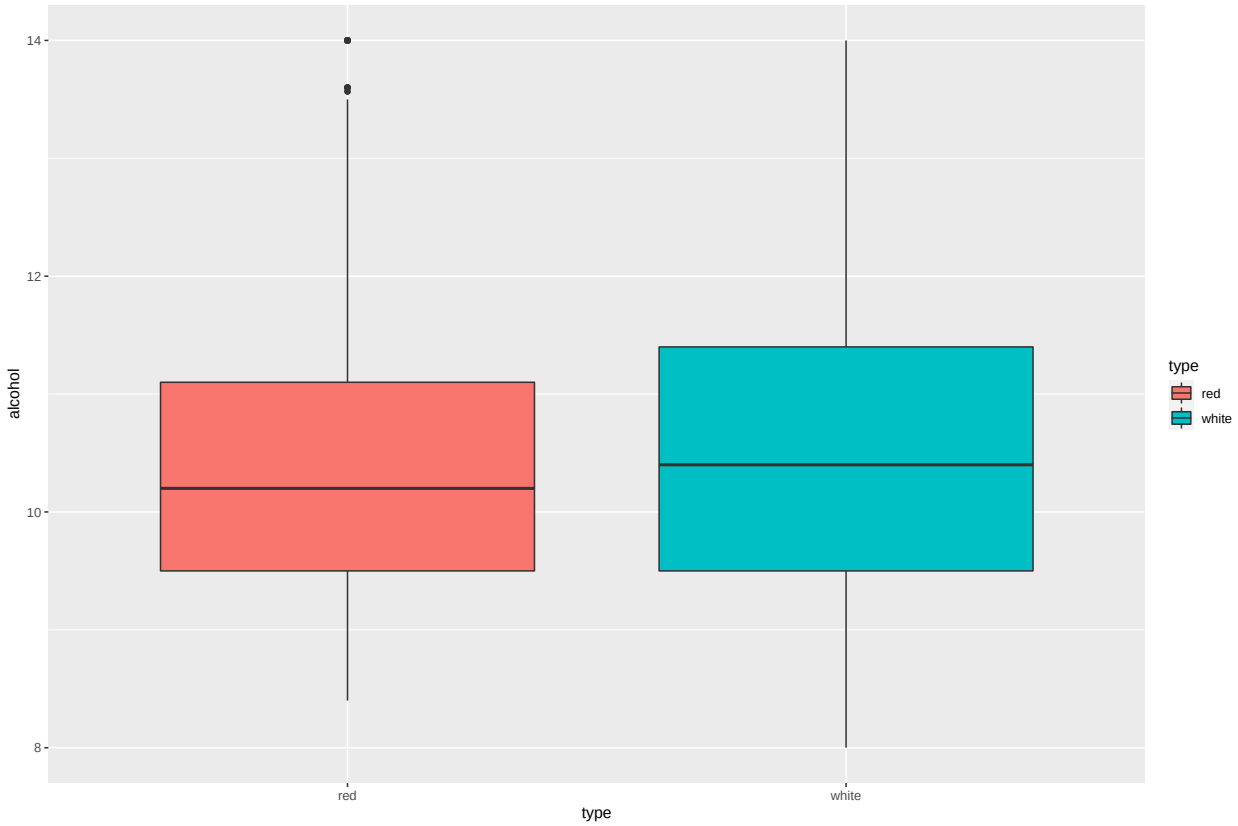
There are a number of outliers both above and below the mean in the red wines sample; the same trend is observed in the the white wines sample. The mean pH for white wines is lower than the mean pH for red wines.

Sulphates



The mean sulphates measurement for red wine is higher than the mean for white wine, and there are a number of outliers above the mean for both samples.

Alcohol



There are a couple outliers above the mean in the red wines sample, and no major outliers for the white wines sample. I'm surprised to observe that the mean alcohol content for white wine is slightly higher than that of red wine. My baseless assumption before this analysis was that red wines had a higher alcohol content.

Univariate Analysis

What is the structure of your datasets?

There are 1,599 observations in the red wines dataset and 4,898 observations in the white wines dataset. Each dataset originally had 12 variables (fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH, sulphates, alcohol, and quality).

Variable Descriptions

1. Fixed Acidity (tartaric acid - g / dmˆ3)
    - most acids involved with wine or fixed or nonvolatile
2. Volatile Acidity (acetic acid - g / dmˆ3)
    - the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. Citric Acid (g / dmˆ3)

- found in small quantities, citric acid can add 'freshness' and flavor to wines

4. Residual Sugar (g / dmˆ3)
    - 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 (sodium chloride - g / dmˆ3)
    - the amount of salt in the wine
6. Free Sulfur Dioxide (mg / dmˆ3)
    - 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 (mg / dmˆ3)
    - 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 / cmˆ3)
    - 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 (potassium sulphate - g / dm3)
    - a wine additive which can contribute to sulfur dioxide gas (S02) levels
    - acts as an antimicrobial and antioxidant
11. Alcohol (% by volume)
    - the percent alcohol content of the wine
12. Quality (score between 1 and 10)
    - based on sensory data, and is highly subjective

Initial Observations

- Red wines have a higher volatile acidity, higher pH, and more chlorides than white wines.
- White wines have more free and total sulfur dioxide than red wines, but red wines have a slightly higher percentage of free sulfur dioxide when compared to white wines.
- Red wines are more dense and have more sulphates (on average) than white wines.
- There is no immediately obvious disparity between the two types of wine when comparing alcohol content.

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

The primary features of interest in this dataset is quality. I'd like to find out what variables are most predictive of the perceived quality of wine.

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

I suspect alcohol and acidity are both big factors in quality, but I will explore the other variables to find out.

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

I created a new variable (type) as a factor when combining the two datasets so white and red wines could be distinguished during analysis. I also created quality_factor, which is a factorized version of the quality variable. This variable will make it possible to group, fill, and facet_wrap by the quality score on my plots in the remainder of this document. Finally, I created a new variable that represents the ratio of free SO2 to total S02 - percent_free_SO2.

Of the variables 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?

Many of the variables have unusual distributions, which is why I included a log10 version of the histogram/density plot for each variable. Plotting them on a log10 scale will often, but not necessarily, result in a more normal distribution. I could have done this just for those that had abnormal distributions, but including the log10 plot in my function makes it easy to quickly review both continuous and log10 versions for each variable.

Also, I adjusted the quality factor variable to include scores that are not in the dataset (1, 2, 10). Even though no records received those scores, I think it's important to display them because the quality measure is on a 1-10 scale.

I also had to include a section in my PLOT_HISTOGRAM_DENSITY function to account for infinite binwidth values in the log10 plots which, obviously, do not plot correctly. If the result of my binwidth calculation was 'Inf', I set bw_log to a default binwidth defined within my function.

Abnormal Distributions

- Volatile Acidity
  - The distribution for white wine has a positive skew, and a bimodal distribution is observed for white wine.
  - Plotting on a log10 scale brought the plots closer to a normal distribution, but red wine is still bimodal.
- Citric Acid
  - The distribution for red wine is fairly flat, and there are peaks in the distribution of white wine at 0.25, 0.5, and 0.75.
  - Plotting on a log10 scale created a more normal distribution, but the peaks observed in white wines are still present.
- Residual Sugar
  - After ploting on a log10 scale, a bimodal distribution is observed in the white wine sample.
- Free Sulfur Dioxide
  - The distribution for both wine types have a positive skew, but this trend is more pronounced in the red wine sample.
  - Plotting on a log10 scale brings the plots closer to normal distribution, but the distribution for white wine now has a negative skew, while the red wine sample has several peaks and valleys.

- Percent Free Sulfur Dioxide
  - Combining free sulfur dioxide and total sulfur dioxide into a new comparative metric resulted in a more normal distribution than either that of free sulfur dioxide or total sulfur dioxide.
  - This variable was not improved by plotting on a log10 scale.
- Sulphates
  - The distribution for both types of wine have a positive skew.
  - Plotting on a log10 bring them much closer to a normal distribution.
- Alcohol
  - The distribution for both white wine and red wine have a prominent positive skew.
  - Plotting on a log10 scale doesn't do much to normalize the plots.
  - I will dig deeper on alcohol and it's relationship to quality more in the rest of this article.

Bivariate Plots

In this section, I'll examine the relationship between variables with correlation and scatter plots.

Correlation Plots

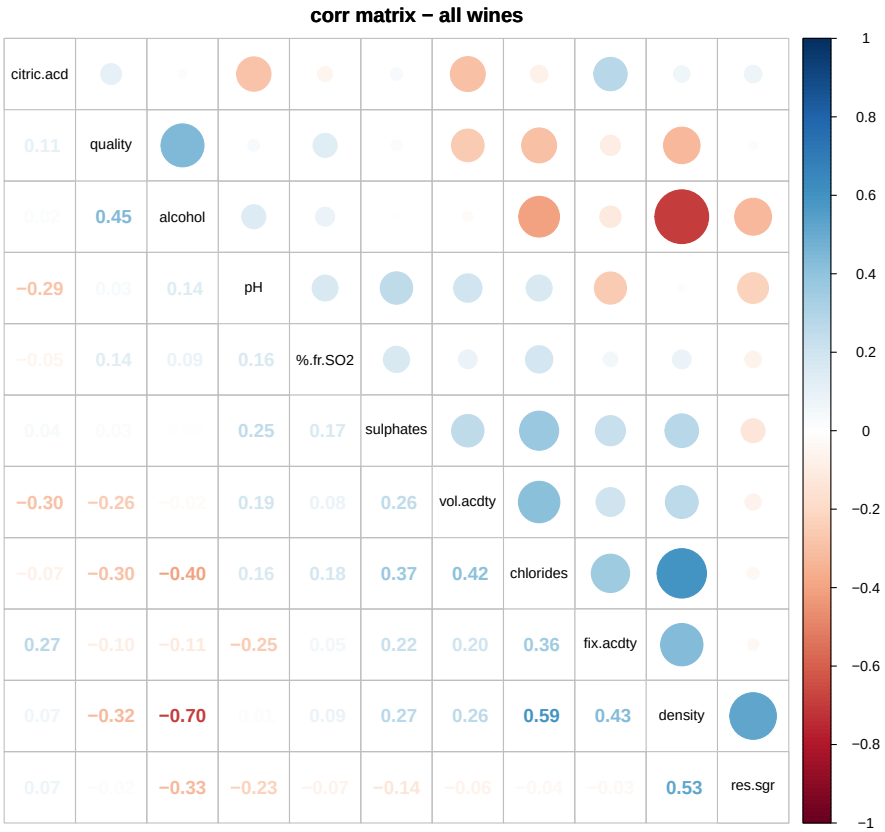
I'll continue to leverage user-defined-functions to create my plots.

```
PLOT_CORR_MATRIX <- function(subType='all')
{
    # create subset of just variables that will be tested
    winesSubset <- winesAll %>%
        select(-c(quality_factor, free_sulfur_dioxide, total_sulfur_dioxide))

    # set title to default
    plotTitle <- 'corr matrix - all wines'

    # white wine only
    if (subType == 'white') {
        winesSubset <- winesSubset %>% filter(type == 'white')
        plotTitle <- 'corr matrix - white wines'
    }

    # red wine only
    if (subType == 'red') {
        winesSubset <- winesSubset %>% filter(type == 'red')
        plotTitle <- 'corr matrix - red wines'
    }

    # remove type
    winesSubset <- winesSubset %>% select(-type)

    # adjust column names so they fit on a chart
    colnames(winesSubset) <- c('fix.acdty', 'vol.acdty', 'citric.acd',
                               'res.sgr', 'chlorides', 'density', 'pH',
```

```
                                'sulphates', 'alcohol', 'quality', '%.fr.SO2')

    # create correlation matrix
    corrMatrix <- cor(winesSubset, method = 'spearman')

    # plot correlation matrix
    corrplot.mixed(
        corrMatrix,
        order = 'AOE',
        tl.col = 'black',
        tl.cex = 0.8,
        number.cex = 1.1,
        title = plotTitle,
        mar=c(0,0,1,0))
}
```

All Wines

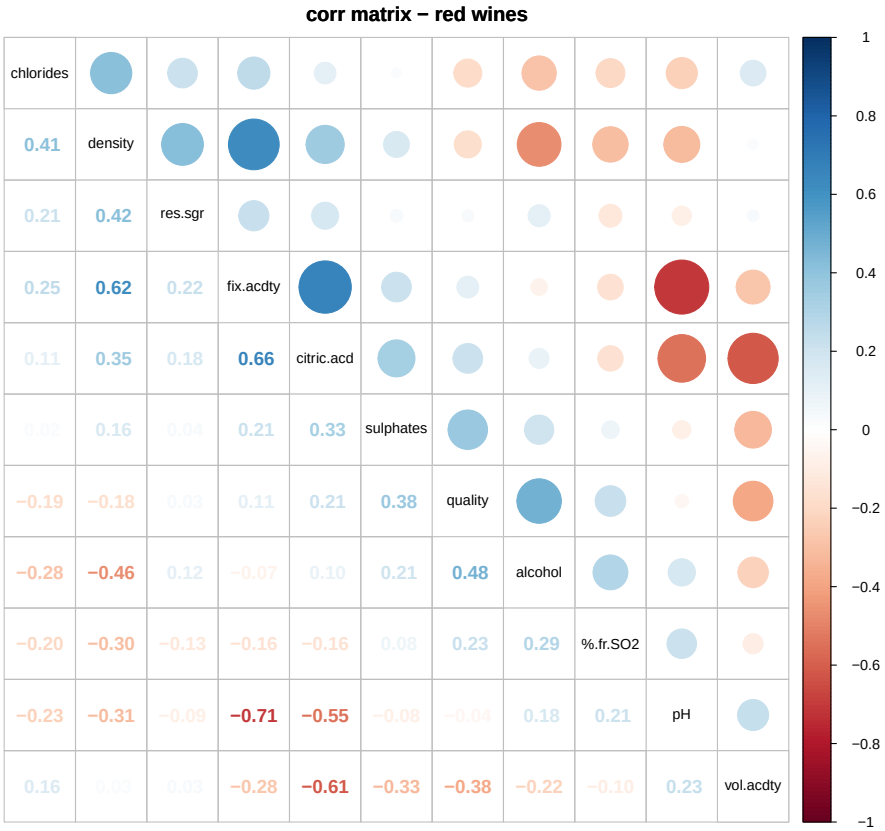
First, I'll plot the correlation matrix for the combined dataset of red wines and white wines.



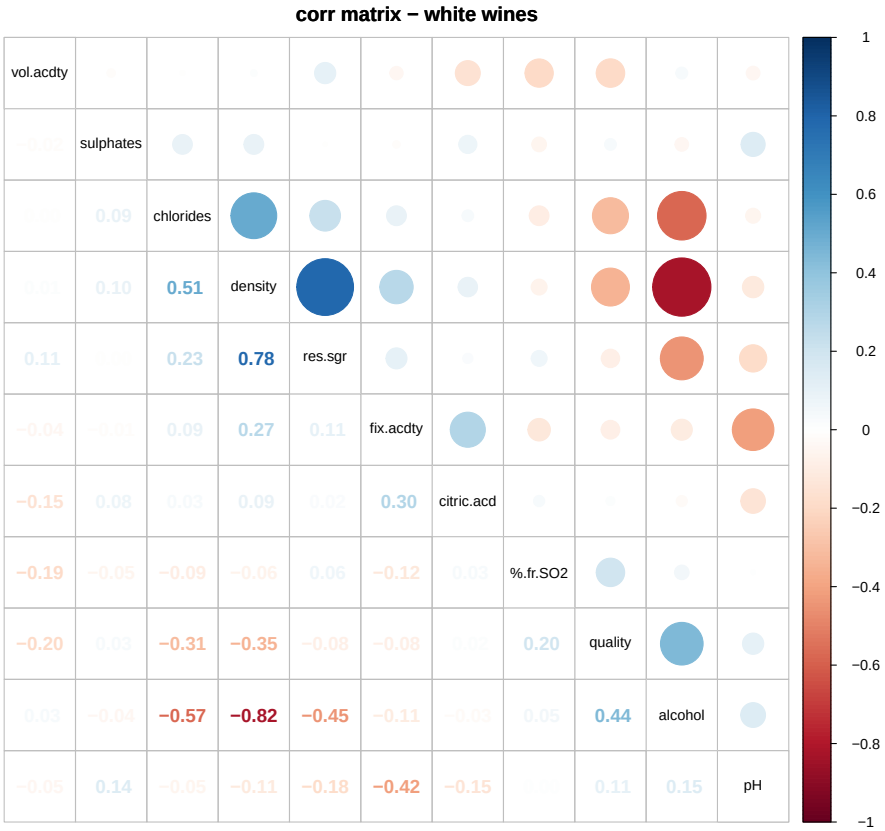
For the combined dataset, it looks like there is a moderate positive correlation between quality and alcohol. Also, there appears to be a moderate negative correlation between quality and three variables: volatile acidity, chlorides, and density.

Next, I'm going to plot the correlation matrices for two samples of data (white wines and red wines). I want to examine both wine types separately to see if the patterns observed in the combined dataset are consistent between red wines and white wines.

Red Wines


corr matrix – red wines
chlorides
density
res.sgr
fix.acdty
citric.acd
sulphates
quality
alcohol
%.fr.SO2
pH
vol.acdty
0.41
0.21 0.42
0.25 0.62 0.22
0.11 0.35 0.18 0.66
0.02 0.16 0.04 0.21 0.33
−0.19 −0.18 0.03 0.11 0.21 0.38
−0.28 −0.46 0.12 −0.07 0.10 0.21 0.48
−0.20 −0.30 −0.13 −0.16 −0.16 0.08 0.23 0.29
−0.23 −0.31 −0.09 −0.71 −0.55 −0.08 −0.04 0.18 0.21
0.16 0.03 0.03 −0.28 −0.61 −0.33 −0.38 −0.22 −0.10 0.23
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

White Wines



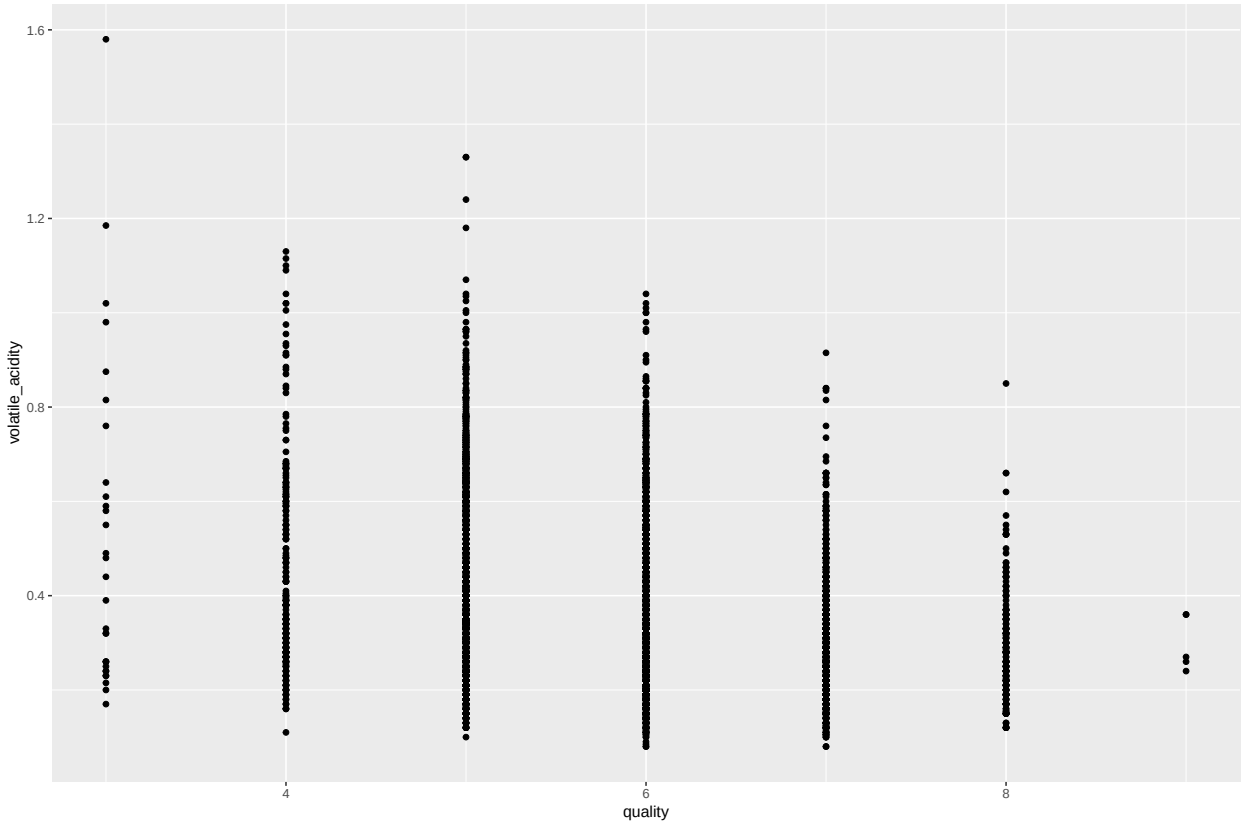
For both the white wines sample and red wines sample, there is a moderate positive correlation between quality and alcohol. It looks like that relationship is consistent across all three matrices. Also, a moderate positive correlation exists between quality and sulphates in the red wines sample that is not observed in the white wines sample.

Scatter Plots

After reviewing the correlation plots, there are a few relationships I want to take a closer look at with scatter plots.

Quality vs. Volatile Acidity

```
ggplot(winesAll) +
    aes(x = quality, y = volatile_acidity) +
    geom_point()
```



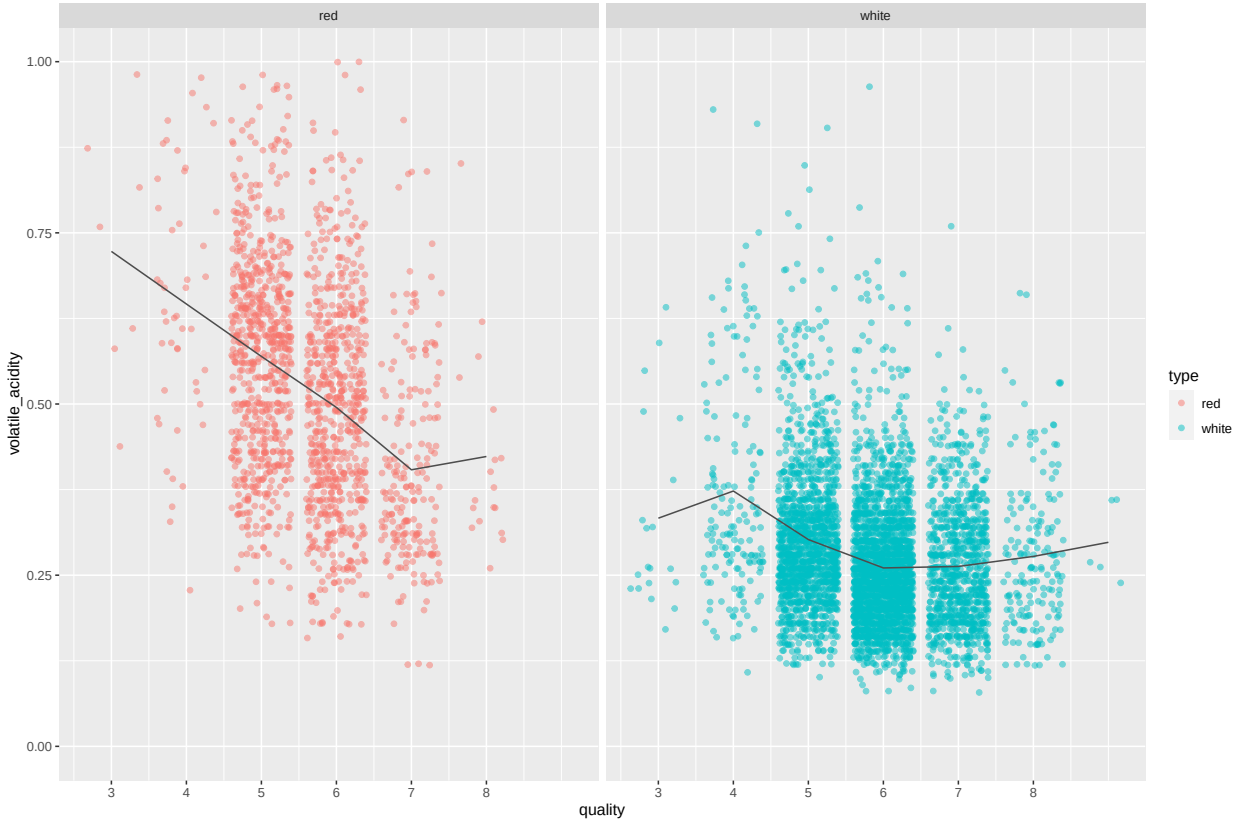
The first iteration of my scatter plot is not very informative. Since the quality score is a categorical variable - meaning all of the results fall into defined levels - all the points are condensed and grouped in vertical lines. This plot looks more like a bar chart than a scatter plot. I should be able to clean this plot up by adding jitter to the points, layering another plot on top to visualize the mean, and plotting the different types of wine separately.

I'll create another user-defined-function to generate these scatter plots.

```
PLOT_SCATTERPLOT <- function(strVar, limStart=99, limEnd=99)
{
    # create vector based on string variable input
    winesVector <- unlist(winesAll[c(strVar)], use.names = FALSE)

    # adjust default limit start
    if (limStart == 99) {
        limStart <- min(winesVector)
    }

    # adjust default limit end
    if (limEnd == 99) {
        limEnd <- max(winesVector)
    }

    # create scatter plot
    ggplot(winesAll) +
        aes(x = quality, y = winesVector, color = type) +
        geom_jitter(alpha = 0.5) +
        geom_line(stat = 'summary', fun = mean, color = 'gray30') +
```

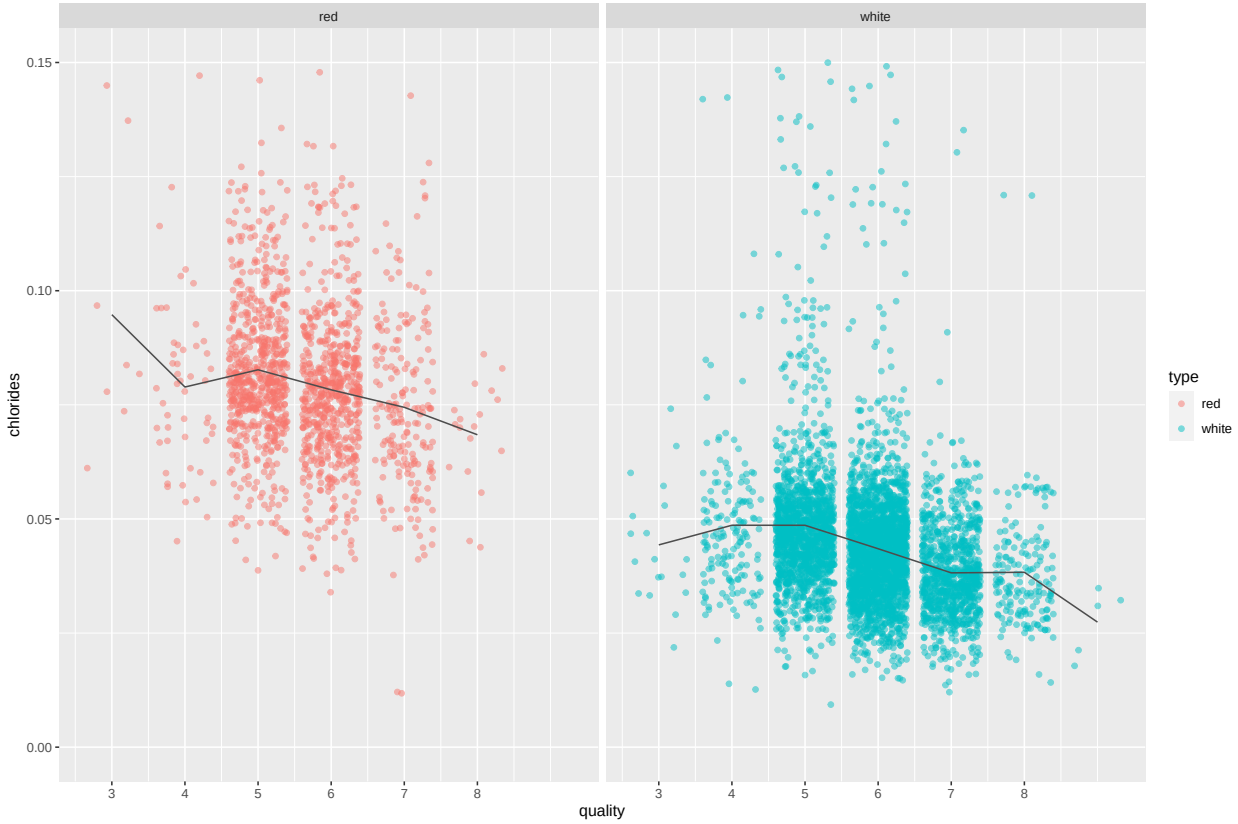
```
        facet_wrap(~type) +
        scale_x_continuous(breaks = c(3,4,5,6,7,8)) +
        scale_y_continuous(limits = c(limStart, limEnd)) +
        labs(y = strVar)
}
```

Quality vs. Volatile Acidity II



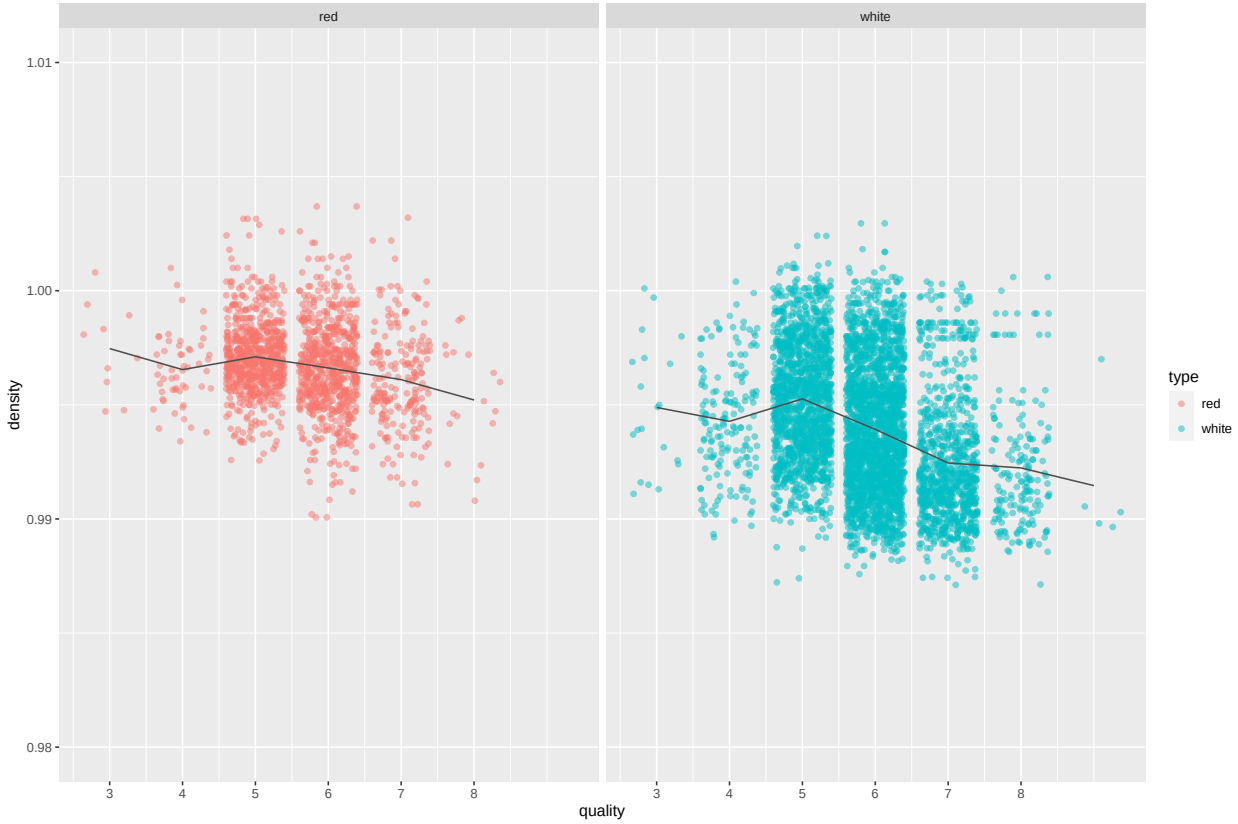
A negative correlation exists between quality and volatile acidity. This variable is more impactful to the perceived quality of red wine than it is for white wine.

Quality vs. Chlorides



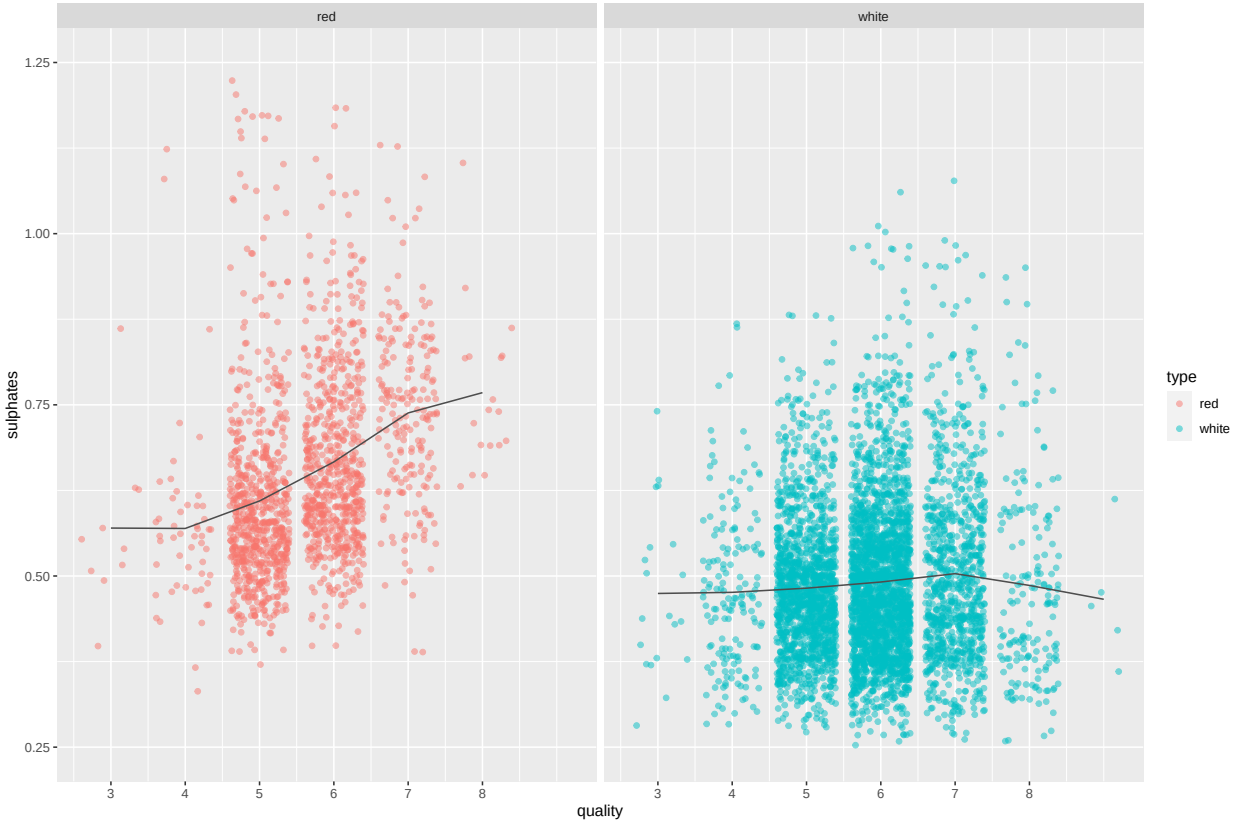
A negative correlation exists between quality and chlorides for both types of wine.

Quality vs. Density



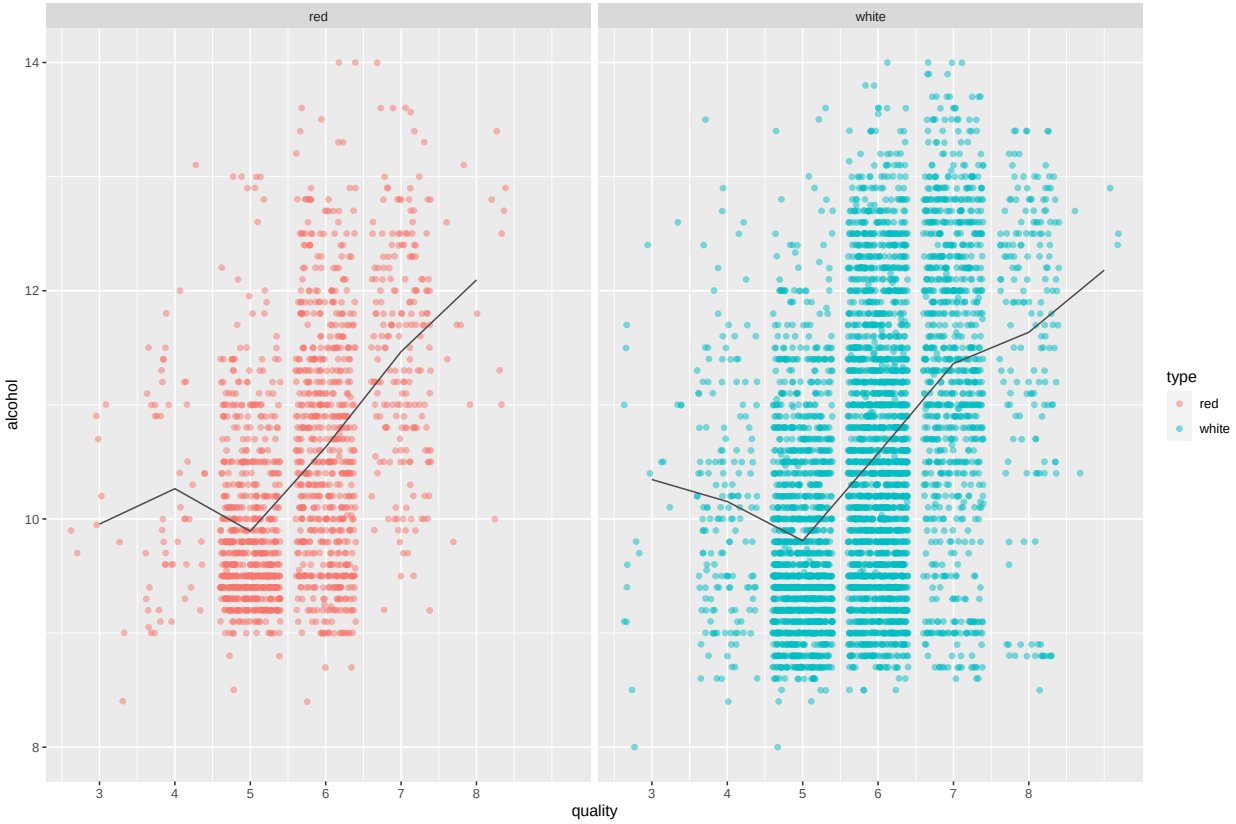
A negative correlation exists between quality and density. This variable is more impactful to the perceived quality of white wine than it is for red wine.

Quality vs. Sulphates



A positive correlation exists between quality and sulphates for red wine. However, no such correlation is observed for white wine.

Quality vs. Alcohol



A positive correlation exists between quality and alcohol content for both white and red wine. Interestingly, this is only true for wines rated at a quality of 5 or higher.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

My primary variables of interest were quality and alcohol. The relationship between alcohol content and perceived quality is interesting, and there is a positive correlation between quality and alcohol for all three datasets:

- All Wines: 0.45
- White Wines: 0.44
- Red Wines: 0.48

However, this is only true for wines rated at a quality of 5 or higher. For wines rated at a quality lower than 5, there is a weak negative correlation.

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

I explored a few other variables' relationship to quality.

Volatile acidity and quality have a clear negative correlation for all three datasets, but this variable is more impactful to the perceived quality of red wine than it is to that of white wine:

- All Wines: -0.26
- White Wines: -0.20
- Red Wines: -0.38

A negative correlation exists between quality and chlorides for all three datasets:

- All Wines: -0.30
- White Wines: -0.31
- Red Wines: -0.19

Density and quality are negatively correlated, but this variable is more impactful to the perceived quality of white wine than it is to that of red wine:

- All Wines: -0.32
- White Wines: -0.35
- Red Wines: -0.18

A moderate positive correlation exists between quality and sulphates for red wine. However, no such correlation is observed for white wine.

- White Wines: 0.03
- Red Wines: 0.38

What was the strongest relationship you found?

The strongest relationship observed was a negative relationship between density and alcohol for both the combined dataset (-0.70) and the white wines subset (-0.82). For the red wine subset, the strongest relationship observed was a negative relationship between fixed acidity and pH (-0.71). However, neither of these relationships are particularly surprising or interesting.

More interestingly are the relationships between quality and the other variables. For the combined datasets, and for both subsets (white and red), the strongest relationship perceived quality has with any other variable is a positive correlation with alcohol content.

- All Wines: 0.45
- White Wines: 0.44
- Red Wines: 0.48

Multivariate Plots

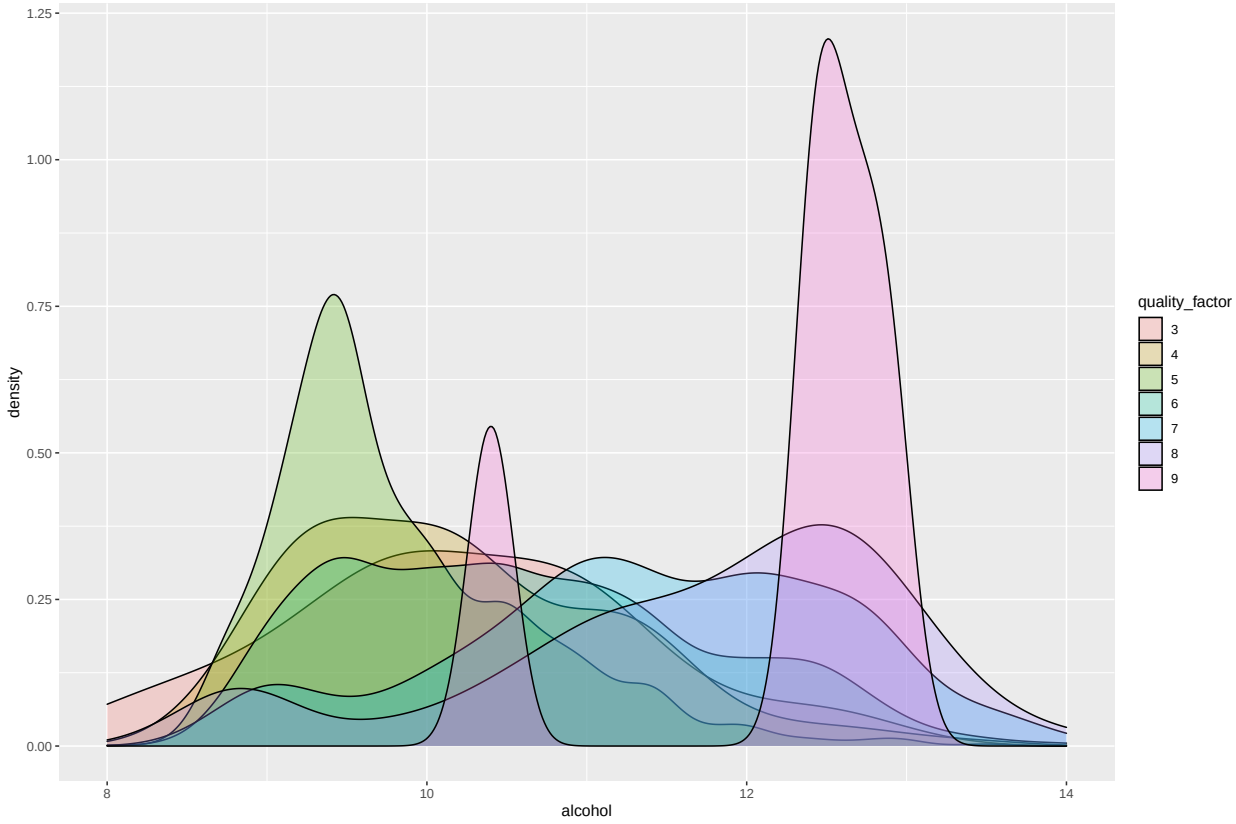
Multi-Variable Density Plots

In this section, I'm going to narrow my focus and closely analyze quality. Specifically, how the different quality ratings compare to a few other key variables: alcohol, sulphates, density, volatile acidity, and chlorides.

Again, I'll create a function to generate the plots for this section.

```
PLOT_DENSITY_BY_FILL <- function(strVar, fillVar, limStart=99, limEnd=99)
{
    # create vector based on string variable input
    winesVector <- unlist(winesAll[c(strVar)], use.names = FALSE)
    groupVector <- unlist(winesAll[c(fillVar)], use.names = FALSE)

    # adjust default limit start
    if (limStart == 99) {
        limStart <- min(winesVector)
    }

    # adjust default limit end
    if (limEnd == 99) {
        limEnd <- max(winesVector)
    }

    # create density plot
    ggplot(winesAll) +
        aes(winesVector, fill = groupVector) +
        geom_density(alpha = 0.25) +
        scale_x_continuous(limits = c(limStart, limEnd)) +
        labs(x = strVar, fill = fillVar)
}
```

Density by Quality - Alcohol

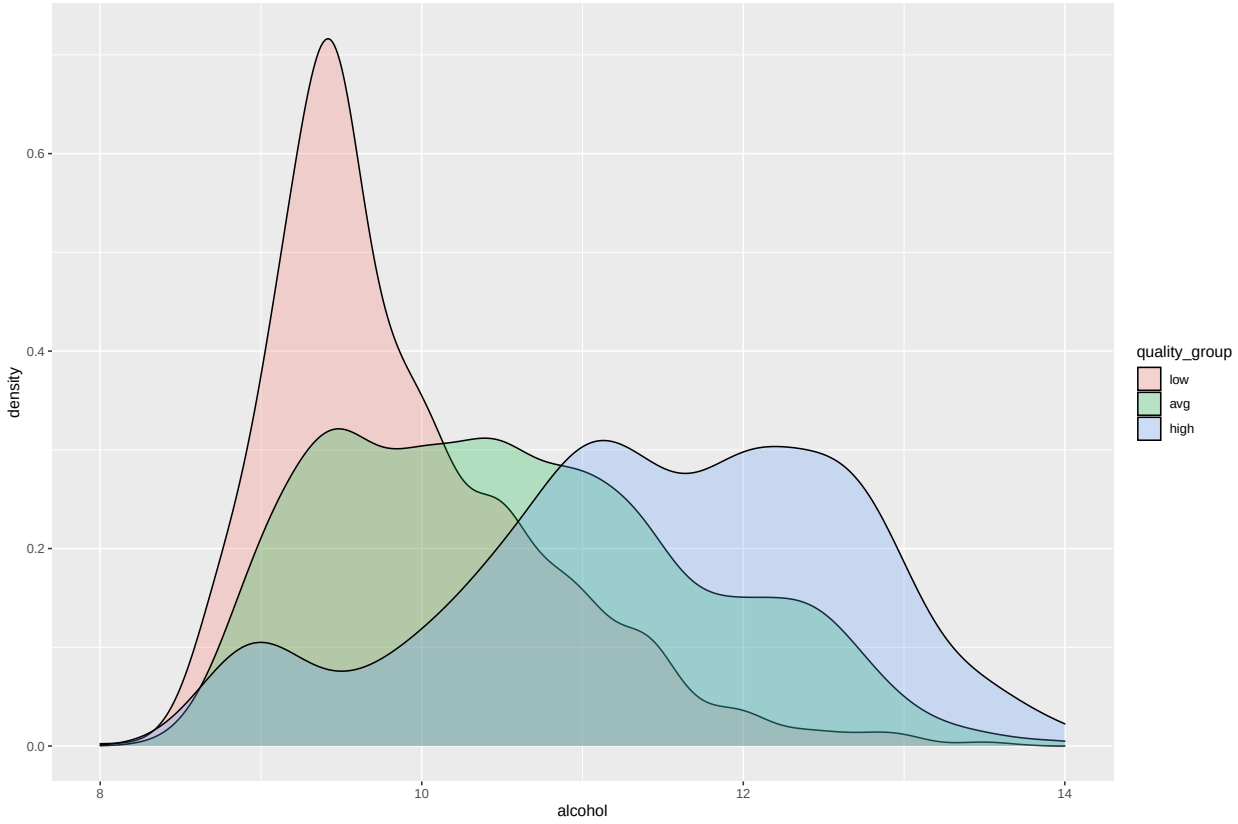


Because there are so many different factors on this plot, it's a bit hard to see what's going on here. To clean this up, I'm going to group qualities scores together into a new factor. I'll bucket quality scores into three groups with roughly the same number of observations: low(3, 4, 5); avg(6); and high(7, 8, 9).

```
# create quality group
winesAll <- winesAll %>%
    mutate(quality_group = case_when(
        quality < 6  ~ 'low',
        quality > 6  ~ 'high',
        TRUE         ~ 'avg'
    ))

# convert to factor for plots
winesAll$quality_group <- factor(winesAll$quality_group,
                                 levels = c('low', 'avg', 'high'))
```

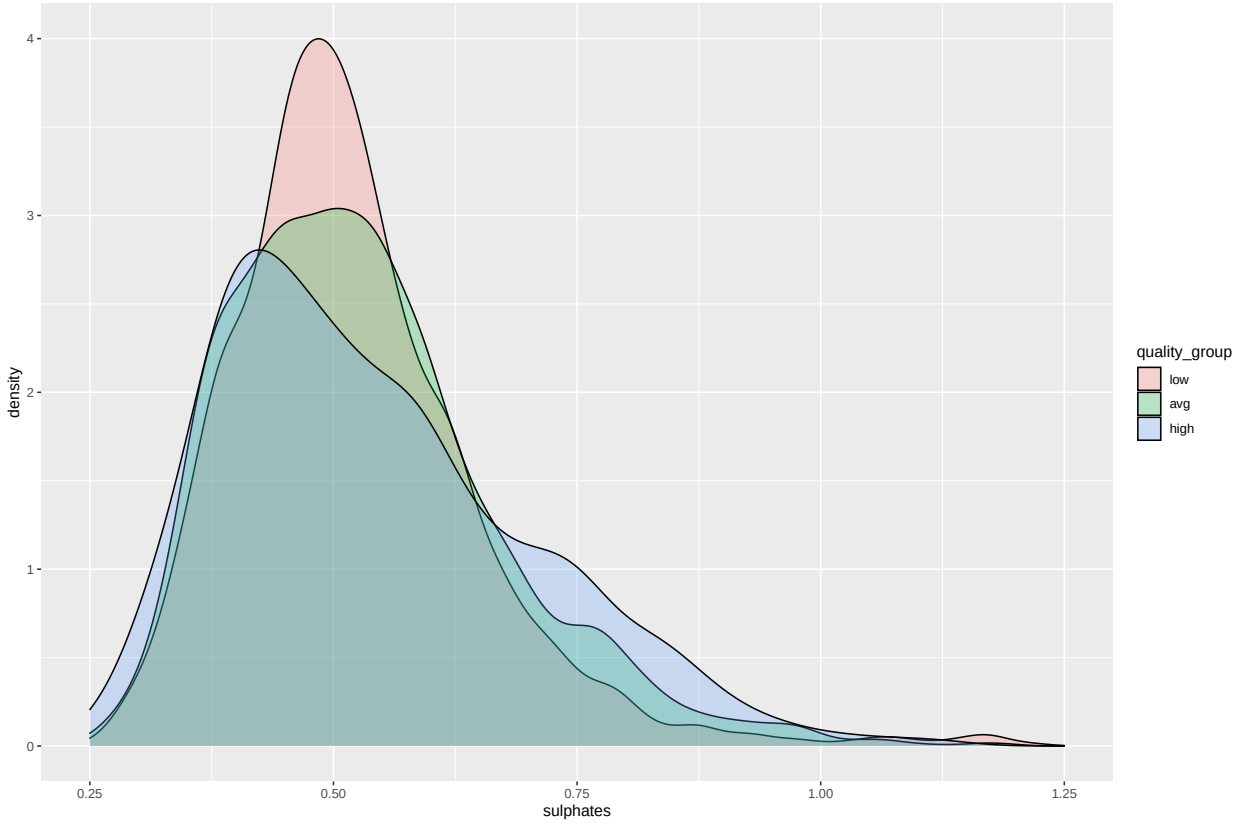
Now I can try creating these density plots again, but with fewer factors on which to group.

Density by Quality - Alcohol II



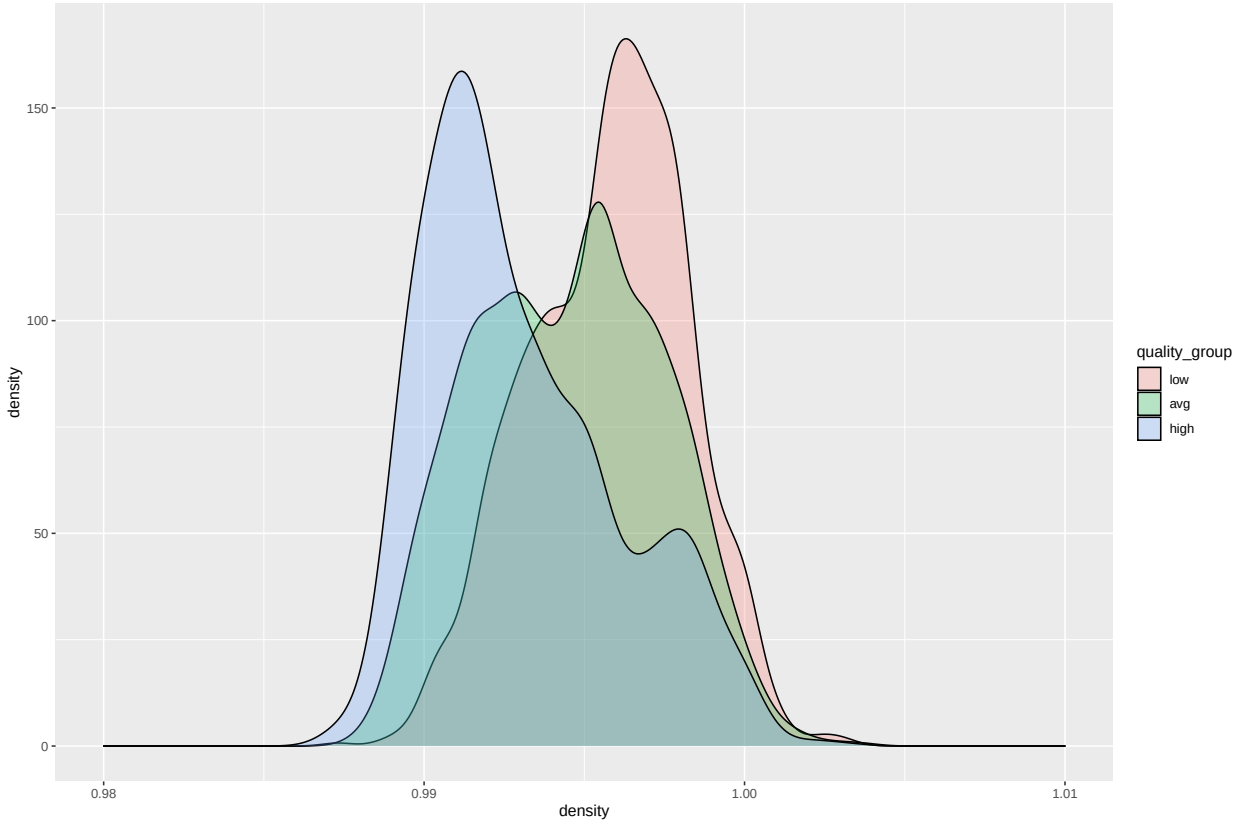
You can see in this plot that lower alcohol content correlates with a lower quality score. The avg quality scores are have a positive skew, but are mostly in the middle of the chart, and the high quality scores are clustered to the right of the plot where the alcohol content is higher.

Density by Quality - Sulphates



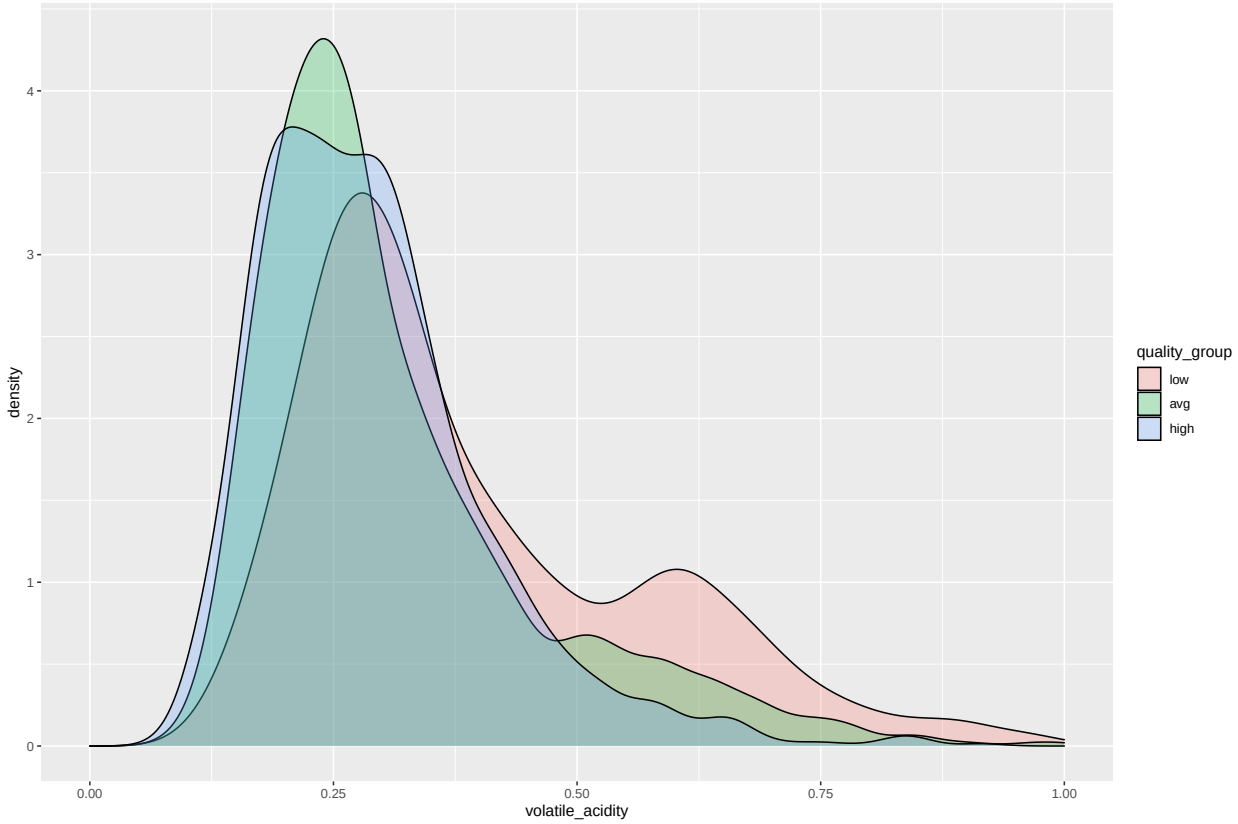
As we move from low to high, the peaks of the density plot get lower, and the plots become more skewed (positive).

Density by Quality - Density



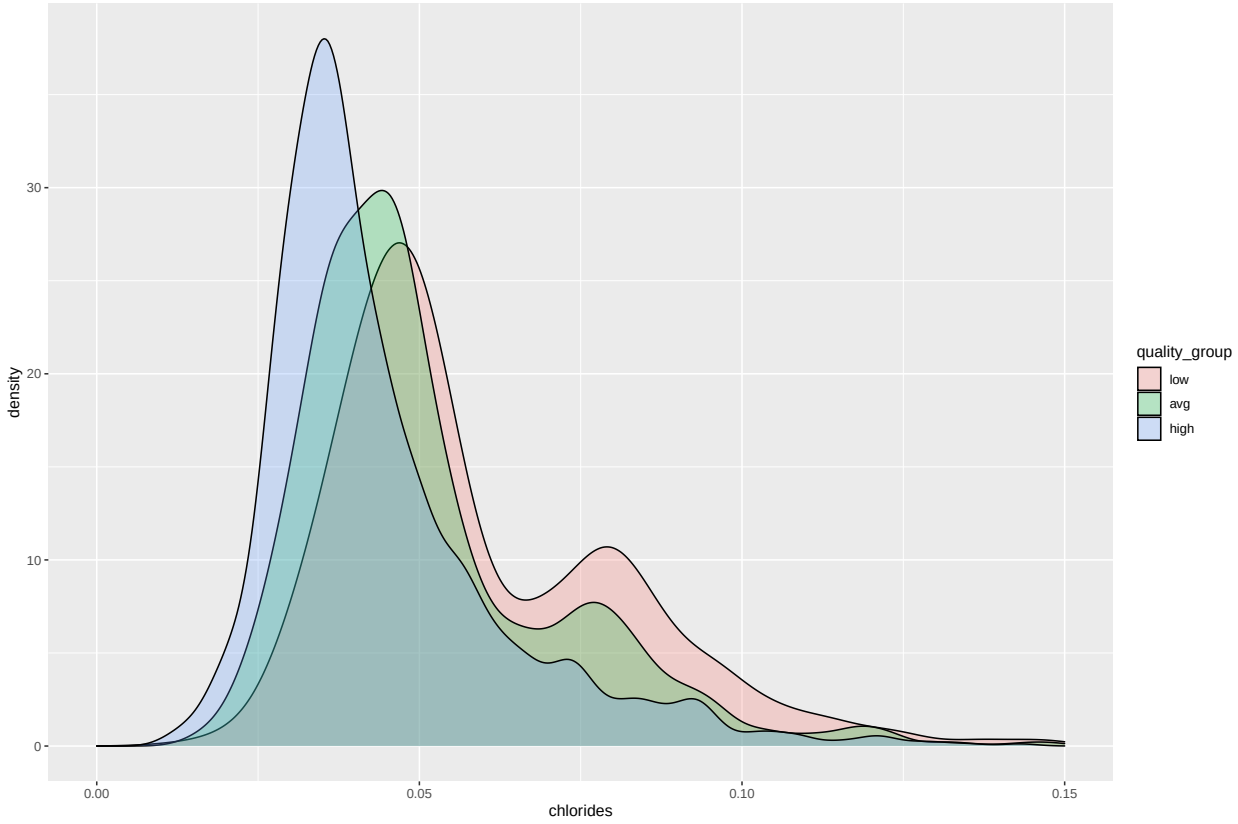
This is an interesting visualization displaying the relationship between density and quality. There are are two peaks flanking the plot for average quality wine. The peak on the left (lower density) is for the higher quality wine, and the peak on the right (higher density) is for lower quality wine.

Density by Quality - Volatile Acidity



Wine perceived to be of lower quality has a more pronounced positive skew, and there are two humps in the plot for that group that aren't present in the high quality wine plot.

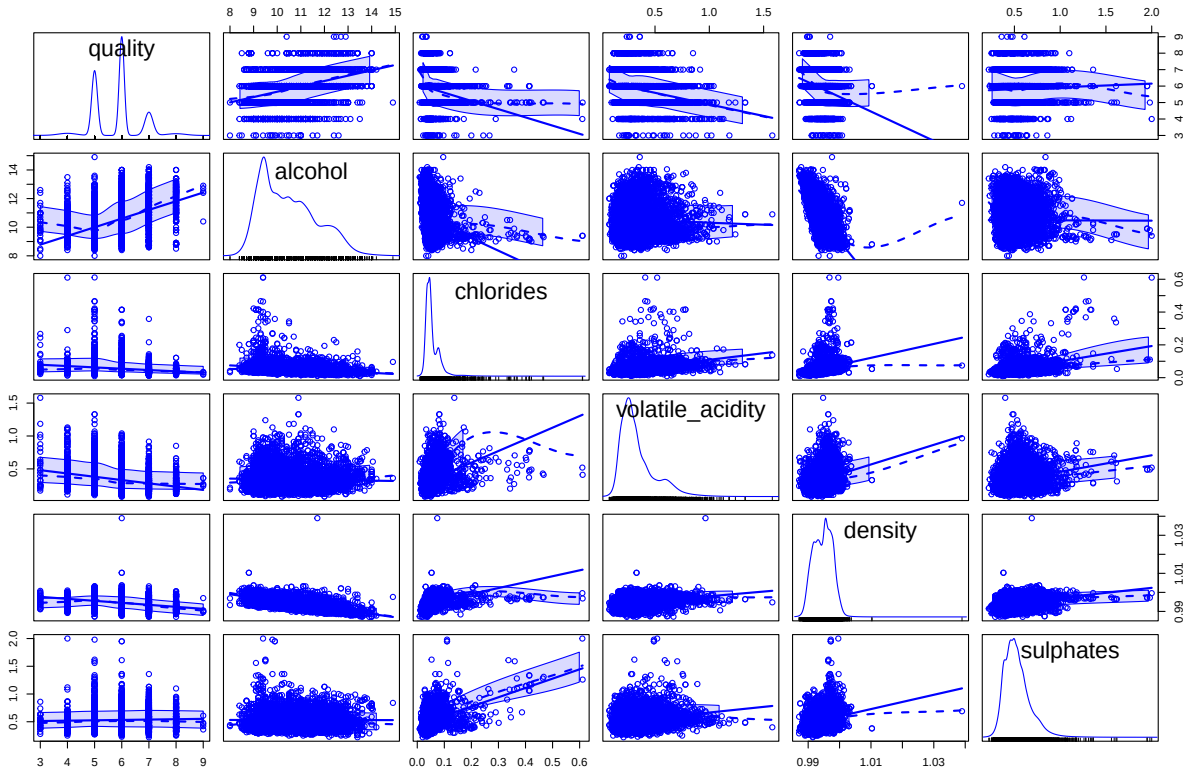
Density by Quality - Chlorides



This plot has a similar trend to the plot for volatile acidity. The plot for the lower quality wines group is more positively skewed, and it is bimodal. However, that trend isn't observable for the higher quality wines group.

Scatter Plot Matrix

Before I move on, I'm going to take a look at a plot I don't like using - the scatter plot matrix. It's extremely popular, so I'd like to demonstrate why I avoid using it.



While this kind of plot does look at multiple variables in different ways, all in one visualization, it's difficult to analyze when you have more than a few variables. In my example, I paired the dataset down to 6 variables, and still the plot is a jumbled mess. It's much more productive to step through the variables and relationships in a more thoughtful way; deciding what visualization would best fit your the goals of your analysis.

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?

In this section, I narrowed my focus and used density plots to analyze the relationships between a few key variables and quality. To do this, I created a new variable called quality_group that is comprised of three factors: low, avg, and high. I'll bucketed quality scores into three groups with roughly the same number of observations: low(3, 4, 5); avg(6); and high(7, 8, 9).

I decided to use this methodology after observing that there was an increase in perceived quality, but only for those wines rated at a quality of 5 or higher.

Were there any interesting or surprising interactions between features?

The most interesting interaction between variables is definitely the correlation between alcohol and quality. The most surprising relationship I observed was between quality and density. In that visualization there

are are two peaks flanking the plot for average quality wine. The peak on the left (lower density) is for the higher quality wine, and the peak on the right (higher density) is for lower quality wine.

Final Plots and Summary

Plot One - Layered Scatter & Line

```
# create clean type variable for visualization
winesAll <- winesAll %>%
    mutate(cType = case_when(
        type == 'white'  ~ 'White Wine',
        type == 'red'    ~ 'Red Wine'
    ))

# factorize clean type variable
winesAll$cType <- factor(winesAll$cType, levels = c('White Wine', 'Red Wine'))

# create layered scatter and line plot
s1 <- ggplot(winesAll) +
    aes(
        x = quality_factor,
        y = alcohol,
        color = cType, group = 1
        ) +

    geom_jitter(
        position = position_jitter(height = 0.1, width = 0.5),
        alpha = 0.3,
        size = 0.5
        ) +

    geom_line(
        stat = 'summary',
        fun = mean,
        color = I('#005429'),
        size = 0.75,
        alpha = 0.9) +

    facet_wrap(~cType) +
    theme_gray() +

    theme(
        legend.position = 'none',
        plot.title = element_text(size = 15L, face = 'bold', hjust = 0.5)
        ) +

    labs(
        x = 'Quality (score between 1 and 10)',
        y = 'Alcohol (% by volume)',
```

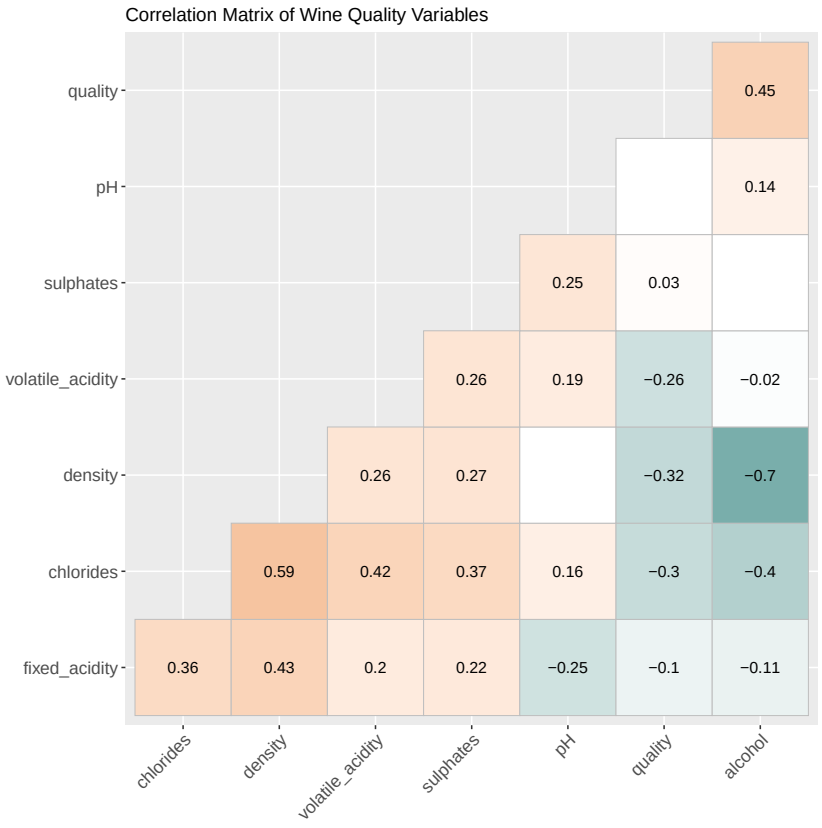
```
        title = 'Alcohol Content by Quality Score & Wine Type',
        caption = 'n = 6479'
        )

# change color scale
s1 <- s1 + scale_color_aaas()

# output visualization
s1
```

This plot uses a scatter (jitter) plot to chart alcohol content by quality score. Because the x-axis variable (quality) is categorical, I decided to use jitter to spread out the data points. If I don't use jitter, the points would plot as single lines centered on each quality score. This visualization also layers a line plot on top of the scatter plot to represent the mean alcohol by volume for each quality score. Adding this line helps visualize the overall trends.

Plot Two - Correlation Matrix

```
# create subset of just variables that will be tested
winesSubset <- winesAll %>%
    select( c(quality, alcohol, volatile_acidity, chlorides,
              density, sulphates, fixed_acidity, pH) )

# create correlation matrix
corrMat <- cor(winesSubset, method = 'spearman')
```

```r
# create matrix of p-values
pValMat <- cor_pmat(winesSubset)


# define title
plotTitle <- 'Correlation Matrix of Wine Quality Variables'

# plot correlation matrix
ggcorrplot(corrMat,
           hc.order = TRUE,
           outline.col = 'gray',
           type = 'lower',
           lab = TRUE,
           insig = 'blank',
           p.mat = pValMat,
           tl.srt = 45,
           show.legend = FALSE,
           title = plotTitle,
           ggtheme = ggplot2::theme_gray,
           colors = c('#3A8C89', 'white', '#E79C60')
           )
```

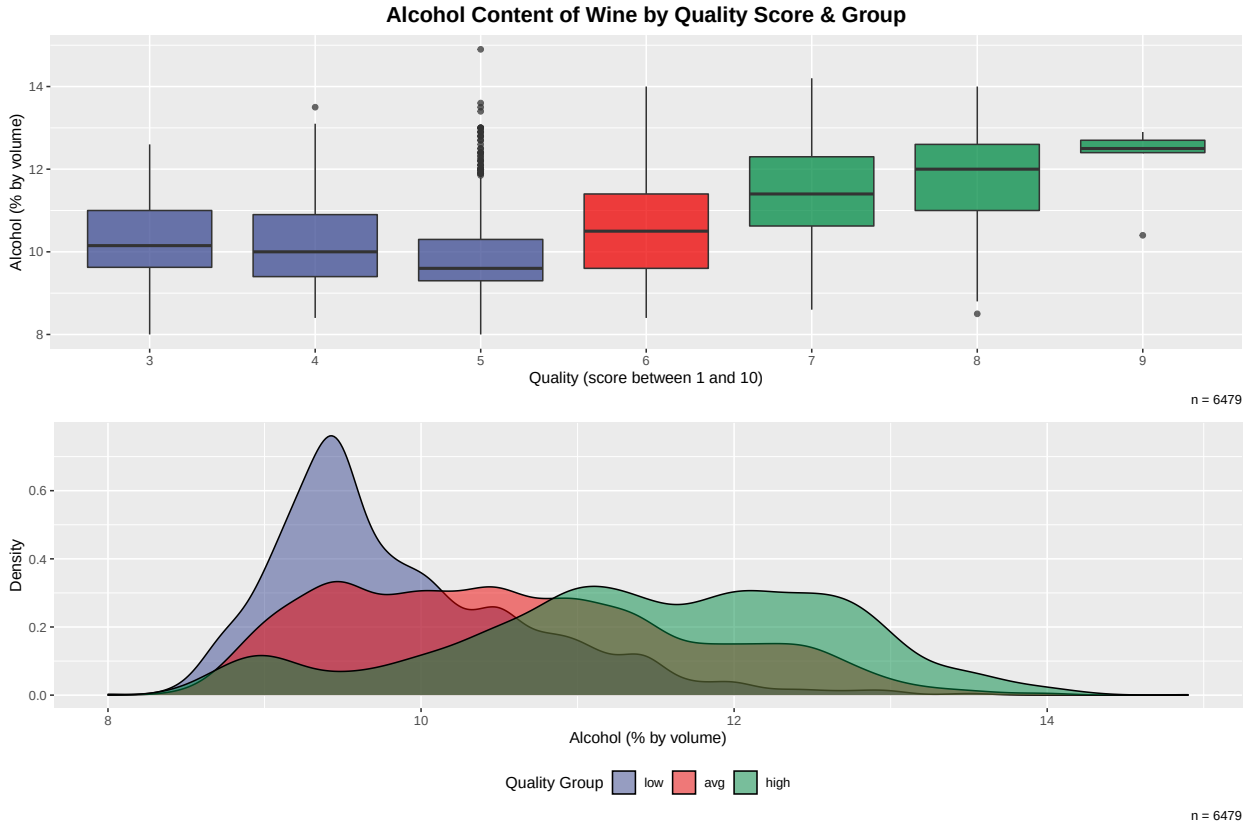


This plot is a cleaned-up correlation matrix that visualizes the relationships between key variables in this dataset. I set the cells to show up blank when the p-value is larger than the significant level (0.05). These cells are insignificant, and can be removed from the visualization to reduce clutter.

Plot Three - Box & Density

```
# create first plot - box
g1 <- ggplot(winesAll) +
    aes(
        x = quality_factor,
        y = alcohol,
        fill = quality_group
        ) +

    geom_boxplot(
        shape = 'circle',
        alpha = 0.75
        ) +

    theme_gray() +

    theme(
        legend.position = 'none',
        plot.title = element_text(size = 15L, face = 'bold', hjust = 0.5)
        ) +

    labs(
        x = 'Quality (score between 1 and 10)',
        y = 'Alcohol (% by volume)',
        title = 'Alcohol Content of Wine by Quality Score & Group',
        caption = 'n = 6479'
        )

# create second plot - density
g2 <- ggplot(winesAll) +
    aes(
        x = alcohol,
        fill = quality_group
        ) +

    geom_density(
        adjust = 0.8,
        alpha = 0.5
        ) +

    theme_gray() +
    theme(legend.position = 'bottom') +

    labs(
        x = 'Alcohol (% by volume)',
        y = 'Density',
        fill = 'Quality Group',
        caption = 'n = 6479'
        )

# change color scale
g1 <- g1 + scale_fill_aaas()
```

```
g2 <- g2 + scale_fill_aaas()

# combine plots into one visualization
grid.arrange(g1, g2)
```



This plot visualizes the relationship between alcohol content and quality score. In the top chart, the individual quality scores are displayed as a boxplot. In the bottom chart, the quality scores are grouped into low, avg, and high; then those groups are displayed as a density plot. Since I aligned the colors between each chart, it's easy to see which quality scores belong in which group.

Reflection

After I combined the red wine and white wine datasets, there were about 6,500 samples with a dozen variables to explore.

I began by creating a summary and a combined histogram/density plot for each variable in the wines dataset. After observing some abnormal distributions in the first few plots, it quickly became apparent that many of the variables were not normally distributed. After discovering this, I went back and included additional plots for scale_x_log10 in my function before continuing my analysis. Adding the log10 scale helped move some variables to a more normal distribution, but not all of them. This dataset is relatively small, and I think that makes it difficult to reach normal distribution in some cases.

Next, I created box plots so I could quickly visualize mean values, the distribution of each variable within the dataset, and any skewed data or outliers.

In the bivariate plots section I employed correlation matrix plots and scatterplots to explore the relationship between variables. I was primarily focused on the relationship each variable had with quality. There were a few interesting correlations, but the strongest indicator of quality was alcohol content.

After that, I narrowed my focus and used density plots to analyze the relationships between a few key variables and quality groups I created. Each of the groups (low, medium, high) had roughly the same number of observations in them, and using this methodology made it easier to visualize the positive correlation between alcohol and perceived quality.

One serious limitation to this project is the quality score used as a key metric. Even though quality was meant to be on a scale of 1-10, 93% of the observations had an assigned quality score of 5, 6, or 7; and there were no observations with a quality score below 3 or above 9. Also, I discovered that using a categorical variable with only a few potential outcomes is good for grouping results, but terrible for using some more advanced plotting techniques, like scatterplots. I was able to overcome this somewhat by layering plots and leveraging geom_jitter, but it made the multivariate plots section of this research challenging.

If I wanted to explore this dataset and these questions more in the future, I would seek out additional data. This project would have benefited from more observations - preferably many more. It also would benefit from additional numeric variables that are related to quality. I suspect price, or perceived price (if real price is hidden from the subjects during wine grading), plays a role in the quality metric - maybe price and alcohol content are correlated.

Or perhaps the one-true indicator of quality in wine is alcohol content, and there's nothing else to discover on this subject.
