Logistic Regression: Wine Classification (Milestone)

Max Schemitsch

April 10, 2019

Dataset Citation

This dataset is public available for research. The details are described in [Cortez et al., 2009].

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

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

---

Dataset Notes

These two datasets, which I have downloaded from Kaggle, use red and white win samples. The metrics used are objective levels like acidity, pH, chlorides, and sugars. Using these types of levels, a score, or quality, is output on a scale of 0 to 10, with 0 being very poor and 10 being outstanding.

The wine used in this dataset are variants of the Portuguese "Vinho Verde" wine. The Vinho Verde region occupies northwest Portugal, and is one of the largest wine regions on the planet.

Attribute Information

Before diving into the dataset, there are a few bits of information that can be extracted at face value.

Our dataframe has 1599 instances of red wines, and 4898 instances of white wines. Although more than two-thirds of the data is white-wine related, there methods that we will use to isolate both types of wine, and analyze them together.

We have a total of 11 input attributes:

Input variables (based on physicochemical tests):

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

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

3 - citric acid (g / dm^3)

4 - residual sugar (g / dm^3)

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

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

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

8 - density (g / cm^3)

9 - pH

10 - sulphates (potassium sulphate - g / dm3)

11 - alcohol (% by volume)

We also have our singular output attribute:

```
Output variable (based on sensory data):
   12 - quality (score between 0 and 10)
```

The description text file included with this dataframe describes what each attribute means:

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

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

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

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

5 - chlorides: the amount of salt in the wine

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

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

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

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

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

11 - alcohol: the percent alcohol content of the wine

Project Goals & Methods

For this project, there will be two parts of interest.

The first part of this project will look in-depth at the characteristics and correlations of variables.

The primary purpose of this section is to first gain an understanding of the data. Additionally, it will gives us an idea of what variables will be useful when moving to the second part of the project.

The second part of this project will be using multiple logistic regression to create models that determine wine scores.

The skills I have learned throughout my Data Science curriculum will hopefully allow me to find a suitable model.

Load Data & Library

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.4.4
```

```
## corrplot 0.84 loaded
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.4.4
```

```
red = read.csv("https://www.dropbox.com/s/jtfubj8tfqpqsa4/wineReds.csv?dl=1")
white = read.csv("https://www.dropbox.com/s/n1pbwl5fkne2i3k/wineWhites.csv?dl=1")
red["color"]="red"
white["color"]="white"
df = rbind(red, white)
attach(df)
```

Dataframe Characteristics

```
head(df)
```

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
```

```
## 1                  11                    34  0.9978 3.51      0.56     9.4
## 2                  25                    67  0.9968 3.20      0.68     9.8
## 3                  15                    54  0.9970 3.26      0.65     9.8
## 4                  17                    60  0.9980 3.16      0.58     9.8
## 5                  11                    34  0.9978 3.51      0.56     9.4
## 6                  13                    40  0.9978 3.51      0.56     9.4
##   quality color
## 1       5   red
## 2       5   red
## 3       5   red
## 4       6   red
## 5       5   red
## 6       5   red
```

```
names(df)
```

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

Looking at the top of our dataset, tells us our attribute variable names, and a general idea of what their values are like.

(We can also see that the integer attribute that increments the wines is called X.)

```
summary(df)
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.: 813   1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500
##  Median :1650   Median : 7.000   Median :0.2900   Median :0.3100
##  Mean   :2044   Mean   : 7.215   Mean   :0.3397   Mean   :0.3186
##  3rd Qu.:3274   3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900
##  Max.   :4898   Max.   :15.900   Max.   :1.5800   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  1.00
##  1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00
##  Median : 3.000   Median :0.04700   Median : 29.00
##  Mean   : 5.443   Mean   :0.05603   Mean   : 30.53
##  3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00
##  Max.   :65.800   Max.   :0.61100   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300
##  Median :118.0        Median :0.9949   Median :3.210   Median :0.5100
##  Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.5313
##  3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000
##  Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.0000
##     alcohol         quality          color
##  Min.   : 8.00   Min.   :3.000   Length:6497
##  1st Qu.: 9.50   1st Qu.:5.000   Class :character
```

```
##  Median :10.30   Median :6.000   Mode  :character
##  Mean   :10.49   Mean   :5.818
##  3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :14.90   Max.   :9.000
```

The summary of our dataframe gives important values like averages, minimums, and maximums.

Looking at these values, there are a few important notes to make:

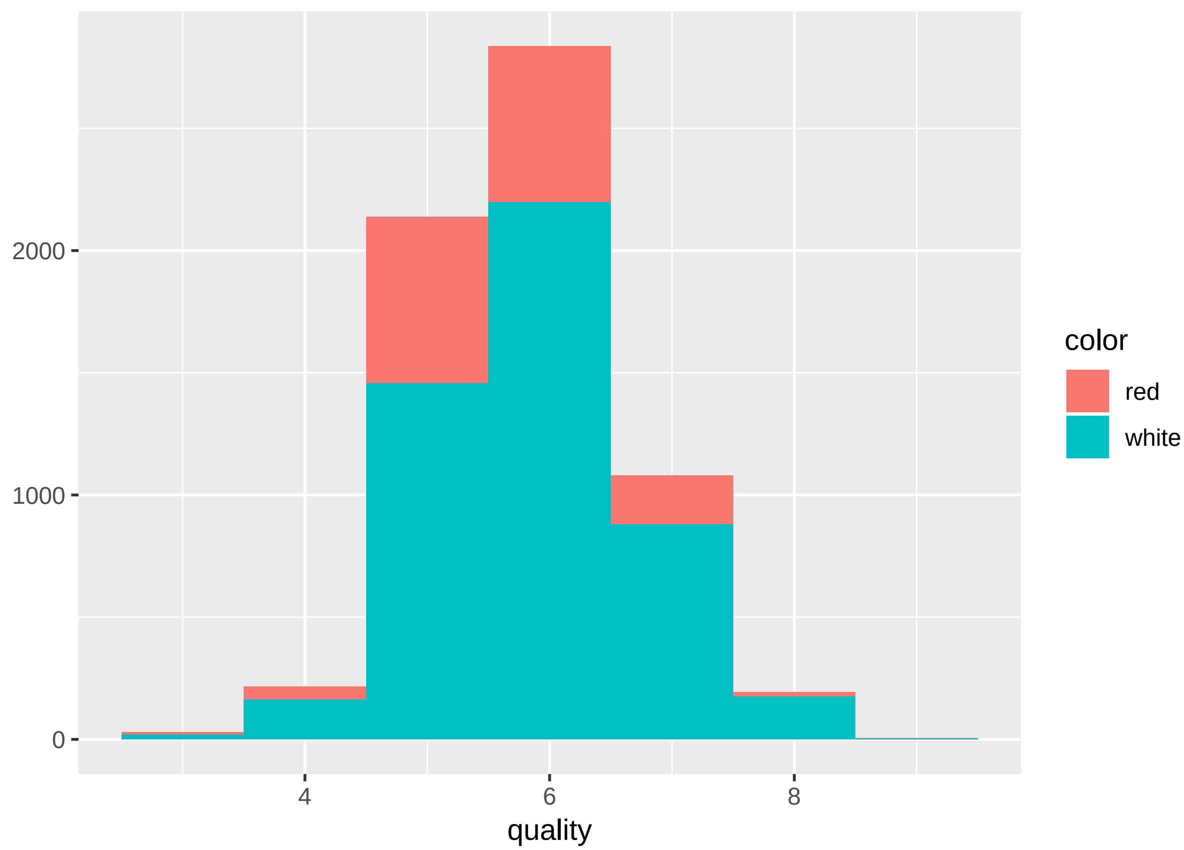
```
1 - Minimum quality is 3 and maximum is 9

2 - The average residual sugar is skewed left: the mean is 5.443, but has a maximum of 65.8.

3 - Similarly, chlorides is skewed. It's range is from 0.009 to 0.611, but has a mean of 0.056.

4 - In the same vein, both free form sulfur dioxide and total sulfur dioxide averages are skewed.
```

First, we can take a look at the distribution of scores and wines:

```r
table(quality)
```

```
## quality
##    3    4    5    6    7    8    9
##   30  216 2138 2836 1079  193    5
```

```r
qplot(quality, data = df, fill = color, binwidth = 1)
```



This shows us that a majority (~6,000) of score values lie between the 5-7 range. There are roughly 200 values for both scores of 4 and scores of 8. Finally, we only have 30 scores of 3 and a miniscule five scores of 9.

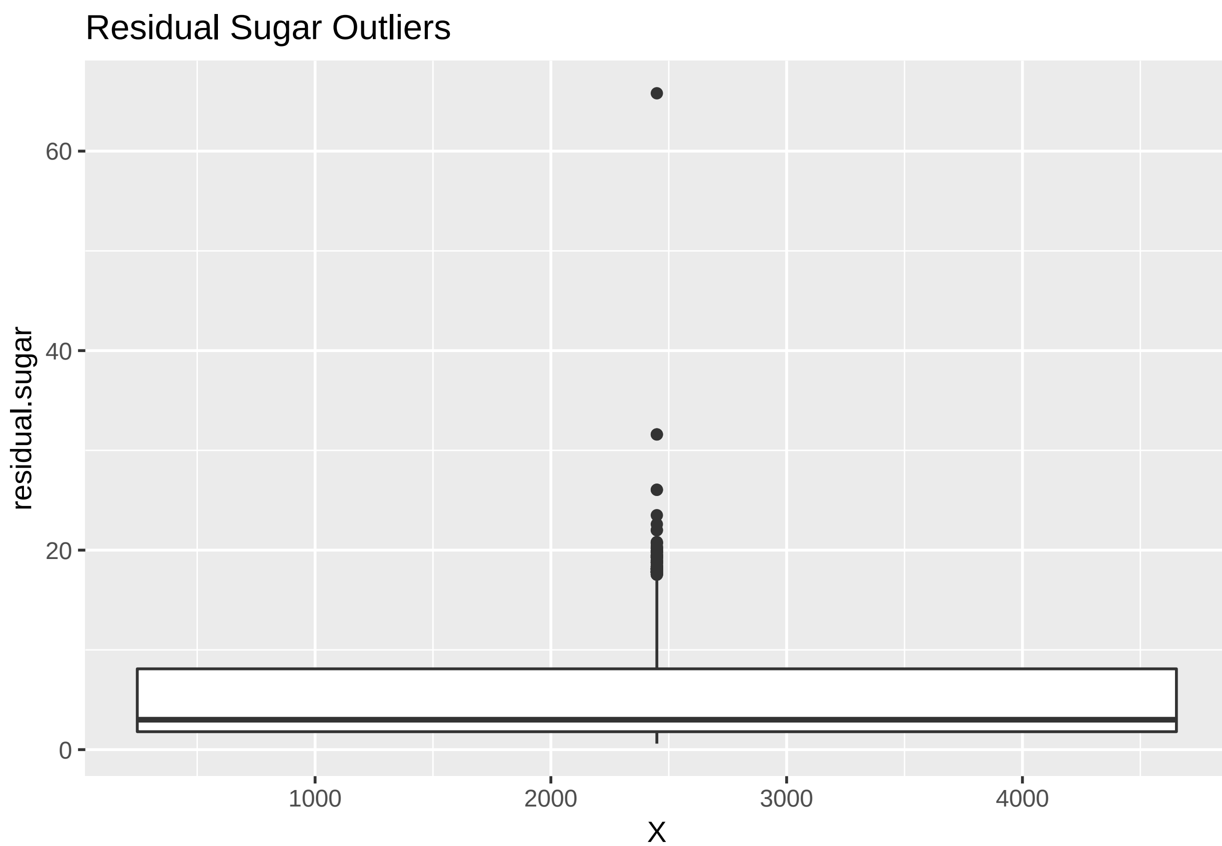
The histogram of wine qualities also shows us the the scores are normally distribued for both red and white wines.

If we want to further validate the data for later regression use, we can check the residuals.

We can also verify the existence of outliers with boxplots:

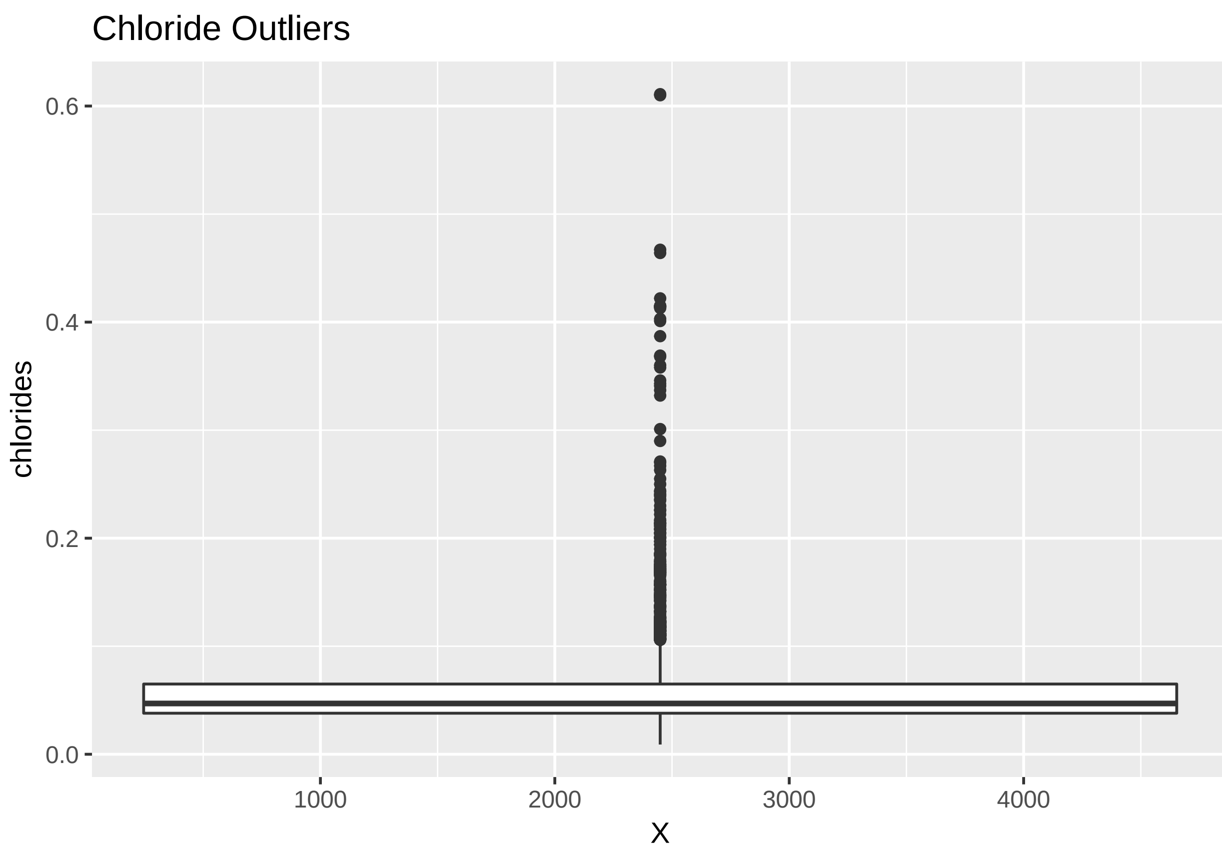
```
ggplot(df, aes(X, residual.sugar))+geom_boxplot() + ggtitle("Residual Sugar Outliers")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



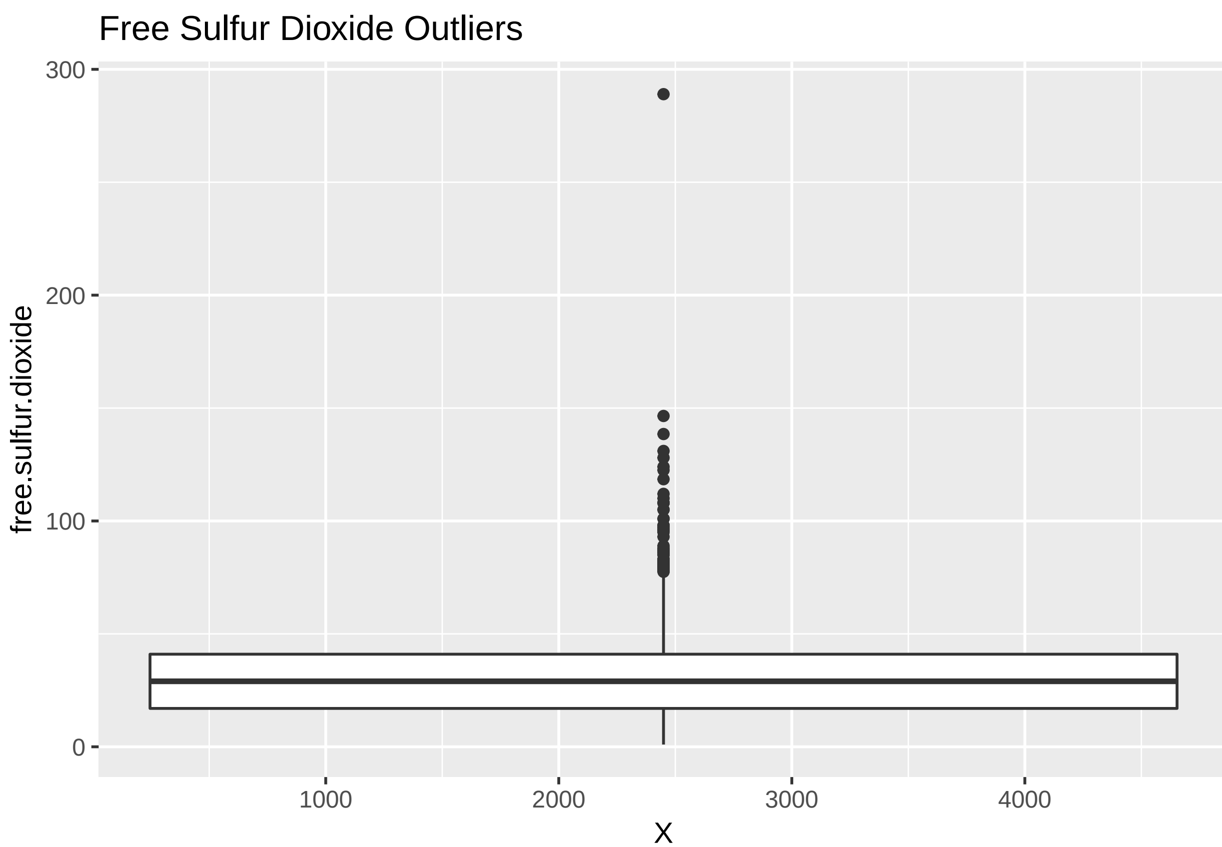

```
ggplot(df, aes(X, chlorides))+geom_boxplot() + ggtitle("Chloride Outliers")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



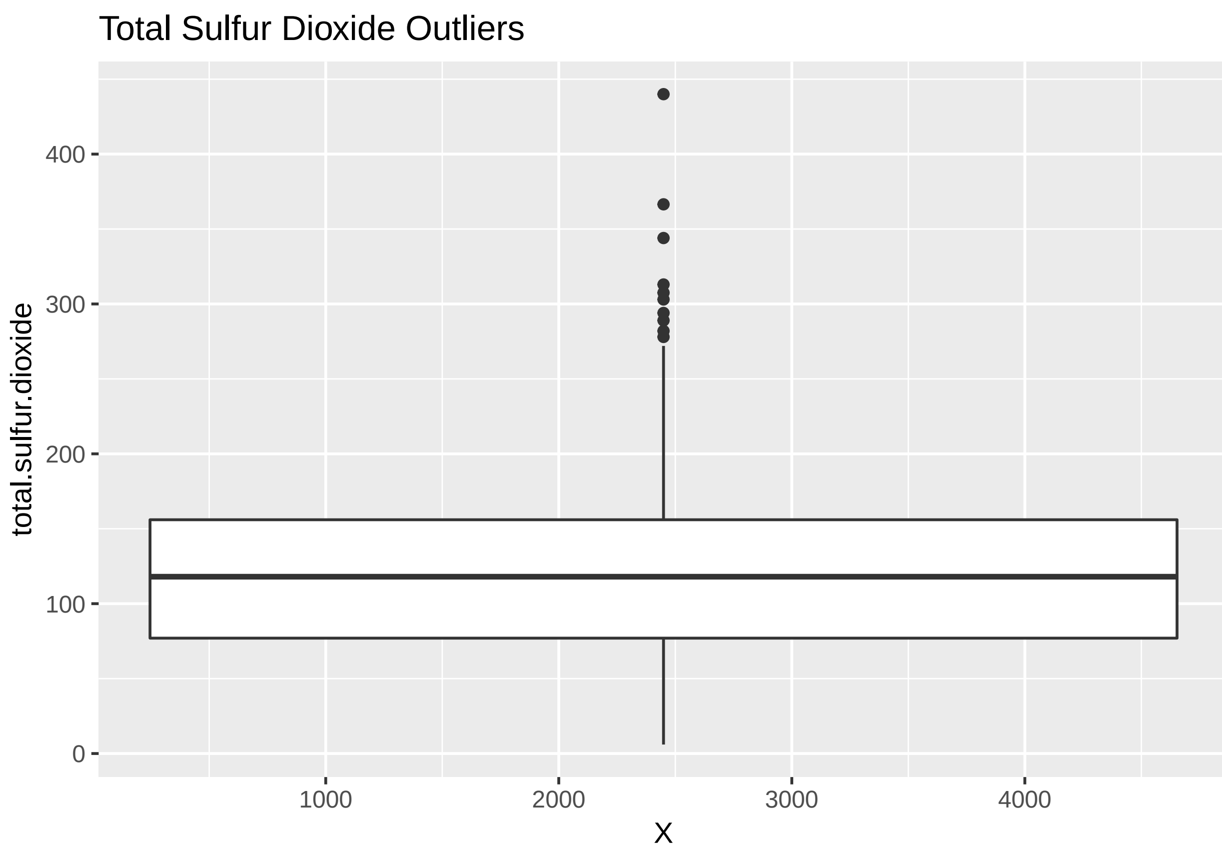

```r
ggplot(df, aes(X, free.sulfur.dioxide))+geom_boxplot() + ggtitle("Free Sulfur Dioxide Outliers")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



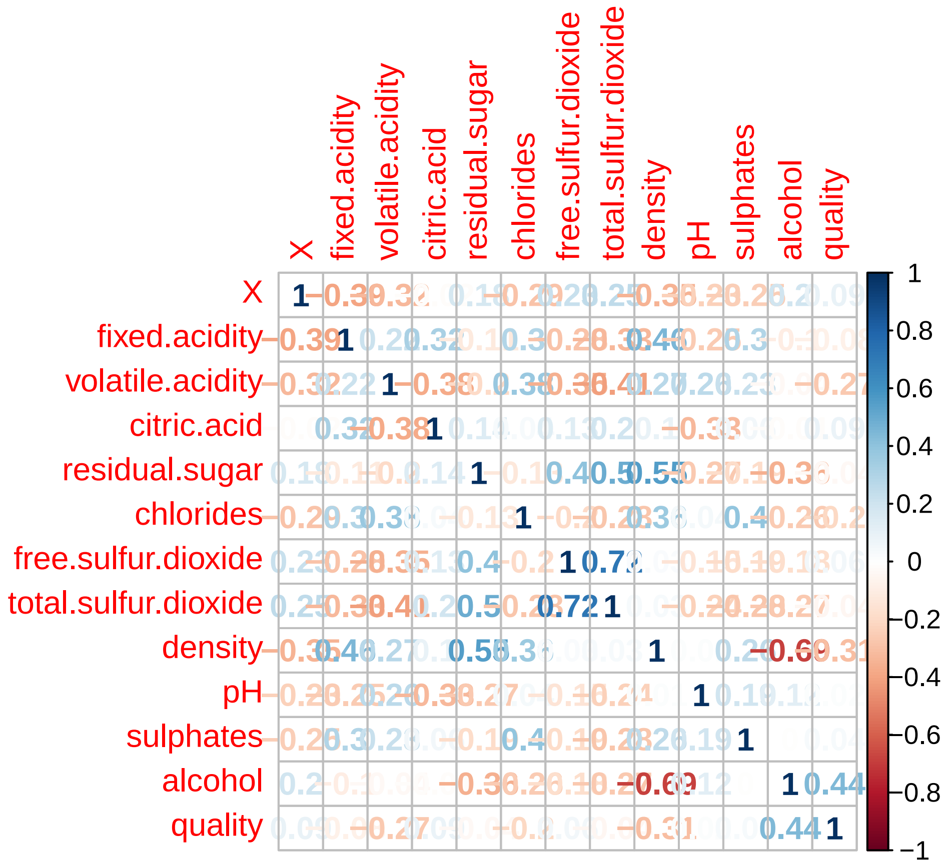

```r
ggplot(df, aes(X, total.sulfur.dioxide))+geom_boxplot() + ggtitle("Total Sulfur Dioxide Outliers")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



Next what we can do, before anything else, is check the correlations of variables. This will help us better understand which characteristics are and aren't related.

```
dfcorr=cor(df[,-14])  # our correlation coefficients
corrplot(dfcorr, method="number")
```



Since we want to predict quality score, we can first look there for variables. A surprising (or unsurprising) seven of the eleven characteristics have a correlation coefficient of between -0.1 and 0.1. This means these variables most likely have very little to do with score output.

Does this mean they won't be included in our modeling process? Not necessarily. Seemingly irrelevant data points can actually improve model accuracy if included.

The other four variables with at least a ± 2 correlation are volatile acidity (-0.27), chlorides (-0.2), density (0.31), and alcohol level (0.44). Before we do any analysis of these variables, we can tell that these will most likely be important to our modelling process later on.

Looking at other variables, we can see some more important correlations (≤ -0.4 or ≥ 0.4):

```
1 - Fixed Acidity & Density (0.46)

2 - Volatile Acidity & Total Sulfur Dioxide (-0.41)

3 - Residual Sugar & Free Sulfur Dioxide (0.4)

4 - Residual Sugar & Total Sulfur Dioxide (0.5)

5 - Residual Sugar & Density (0.55)

6 - Chlorides & Sulfates (0.4)

7 - Free Sulfur Dioxide & Total Sulfur Dioxide (0.72)

8 - Density & Alcohol (-0.69)
```

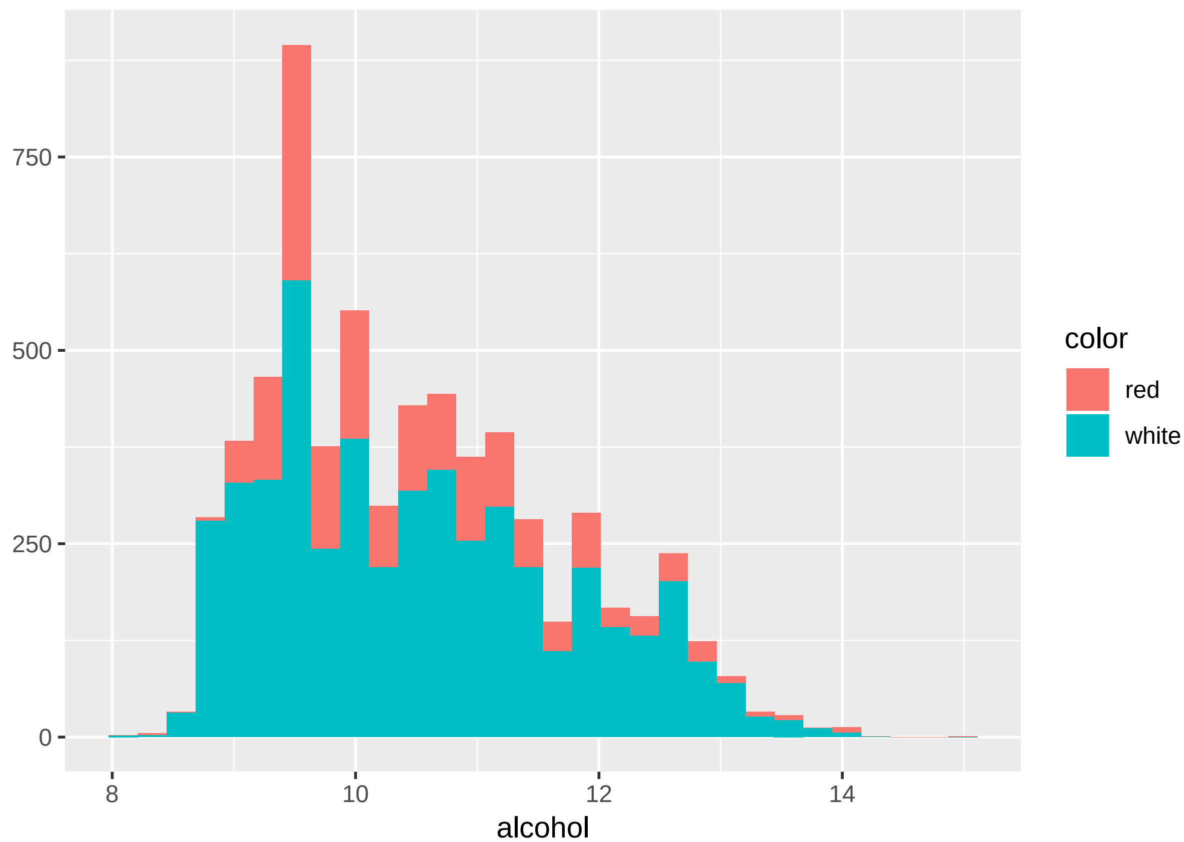
Quality Analysis

```
tapply(alcohol, quality, mean)
```

```
##         3         4         5         6         7         8         9
## 10.215000 10.180093  9.837783 10.587553 11.386006 11.678756 12.180000
```

```
qplot(alcohol, data=df, fill=color)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
tapply(density, quality, mean)
```

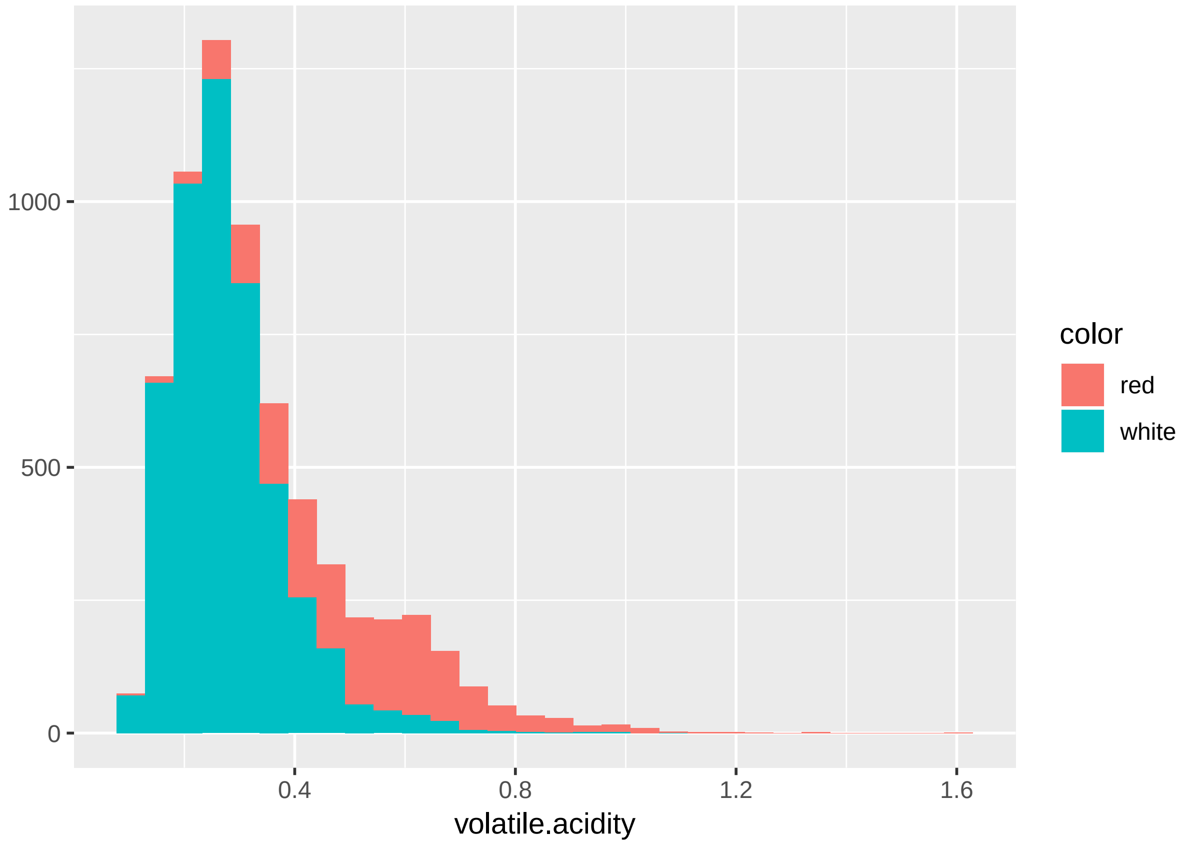
```
##         3         4         5         6         7         8         9
## 0.9957440 0.9948326 0.9958490 0.9945583 0.9931259 0.9925135 0.9914600
```

```
tapply(volatile.acidity, quality, mean)
```

```
##         3         4         5         6         7         8         9
## 0.5170000 0.4579630 0.3896141 0.3138628 0.2887998 0.2910104 0.2980000
```

```r
qplot(volatile.acidity, data=df, fill=color)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
tapply(chlorides, quality, mean)
```

```
##          3          4          5          6          7          8
## 0.07703333 0.06005556 0.06466604 0.05415726 0.04527247 0.04112435
##          9
## 0.02740000
```

Looking at these quality dependent means we can see that quality score generally rises as the level of alcohol increases. This rise isn't true for quality scores of 4 and 5, but qualities 3 & 4 have very similar alcohol levels. Additionally, a third of the wines have a quality score of 5, which could explain the rather low mean alcohol level.

We can see that density stays relatively the same throughout all quality levels. The only trend is that the density level generally decreases as quality goes up.

Volatile acidity levels generally decrease (except at qualities 8 and 9), and chloride levels generally decrease as well, except at quality 5.

We have to remember that red and white wines are characteristically different, so we can check quality on both subsets of our dataframe.

```r
cor(red[,(2:12)], red$quality)
```

```
##                             [,1]
## fixed.acidity         0.12405165
## volatile.acidity     -0.39055778
## citric.acid           0.22637251
## residual.sugar        0.01373164
## chlorides            -0.12890656
## free.sulfur.dioxide  -0.05065606
## total.sulfur.dioxide -0.18510029
## density              -0.17491923
## pH                   -0.05773139
## sulphates             0.25139708
## alcohol               0.47616632
```

```r
cor(white[,(2:12)], white$quality)
```

```
##                              [,1]
## fixed.acidity        -0.113662831
## volatile.acidity     -0.194722969
## citric.acid          -0.009209091
## residual.sugar       -0.097576829
## chlorides            -0.209934411
## free.sulfur.dioxide   0.008158067
## total.sulfur.dioxide -0.174737218
## density              -0.307123313
## pH                    0.099427246
## sulphates             0.053677877
## alcohol               0.435574715
```

We can notice from these corrlations that volatile acidity, citric acid, and sulphates matter more to red wines than white wines.

We can also see that chlorides and density matter more to white wines than red wines.

These correlations are apparent due to the characteristics of each wine and how they're made. We can show that white wines are generally sweeter:

```r
tapply(white$residual.sugar, white$quality, mean)
```

```
##        3        4        5        6        7        8        9
## 6.392500 4.628221 7.334969 6.441606 5.186477 5.671429 4.120000
```

```r
tapply(red$residual.sugar, red$quality, mean)
```

```
##        3        4        5        6        7        8
## 2.635000 2.694340 2.528855 2.477194 2.720603 2.577778
```

This is due to the fermentation process for each wine:

White wine is made only from the juice of the grape; juice is pressed out from the grape, and only that juice is fermented.

In contrast, red wine fermentation utilizes not only the juice of the grape, but also the grapes' skin and pieces of the grape.

While there are other differences in these wines, these differences along with grape type, create the differing acidity and sugar levels between the two.