Joel Munoz

STAT 43000

Final Project

May 3, 2025

**Factors Impacting the Quality of Wine**

**Introduction**

Wine is one of the most popular forms of alcohol purchased on the market. As such, there is a high demand for bottles containing the highest quality of wine. Since the process of creating wine is a lengthy one with many factors, I thought it would be interesting to see which factors can influence the quality of wine.

The goal of this project is to answer the question “What are the most significant variables that affect the quality of wine?” The dataset selected comes from the UCI website. It consists of 6,497 instances of wines, 4,989 of which are white and 1,599 red. Since the data was originally separated between white and red, I combined the two datasets and added a variable “color” to identify its color.

**> #import**

**> white = read.csv("C:/Users/Joel/Downloads/wine+quality/winequality-white.csv", header = T, sep = ";")**

**> red = read.csv("C:/Users/Joel/Downloads/wine+quality/winequality-red.csv", header = T, sep = ";")**

**> #add column "color"**

**> white$color = "white"**

**> red$color = "red"**

**> dim(white)**

**[1] 4898 13**

**> dim(red)**

**[1] 1599 13**

**> #combine**

**> data = rbind(white, red)**

**> dim(data)**

**[1] 6497 13**

**Variables**

There are a total of 13 variables. Their names and definitions will be found below.

**Predictor Variables**

* **fixed acidity** - Predominantly made up of acids such as tartaric, malic, citric, and succinic. They are what give wines the sourness or tartness.
* **volatile acidity** - refers to the steam distillable acids in wines, made up mostly of acetic acid as well as lactic, fromic, butyric, and propionic acids. They give wine their vinegary taste, and too much can lead to a lower quality wine. Also prevents spoilage
* **citric acid**- added for a fresh, citrusy flavor
* **residual sugar**- the sugars that were left in the wine from the grape, gives wine a sweeter taste
* **chlorides**- the level of chloride in a wine depends on the location and climate of where the original grape vine grew. Gives wine a saltier taste.
* **free sulfur dioxide**- to protect against oxidation and microbial spoilage.
* **density-** can be indicative of the sugar/acid to alcohol ratio
* **pH**- acidity, lower score is more acidic and higher score is more basic
* **sulphates**- similar to free sulfur dioxide in use and chemical composition
* **alcohol -** the % of alcohol content

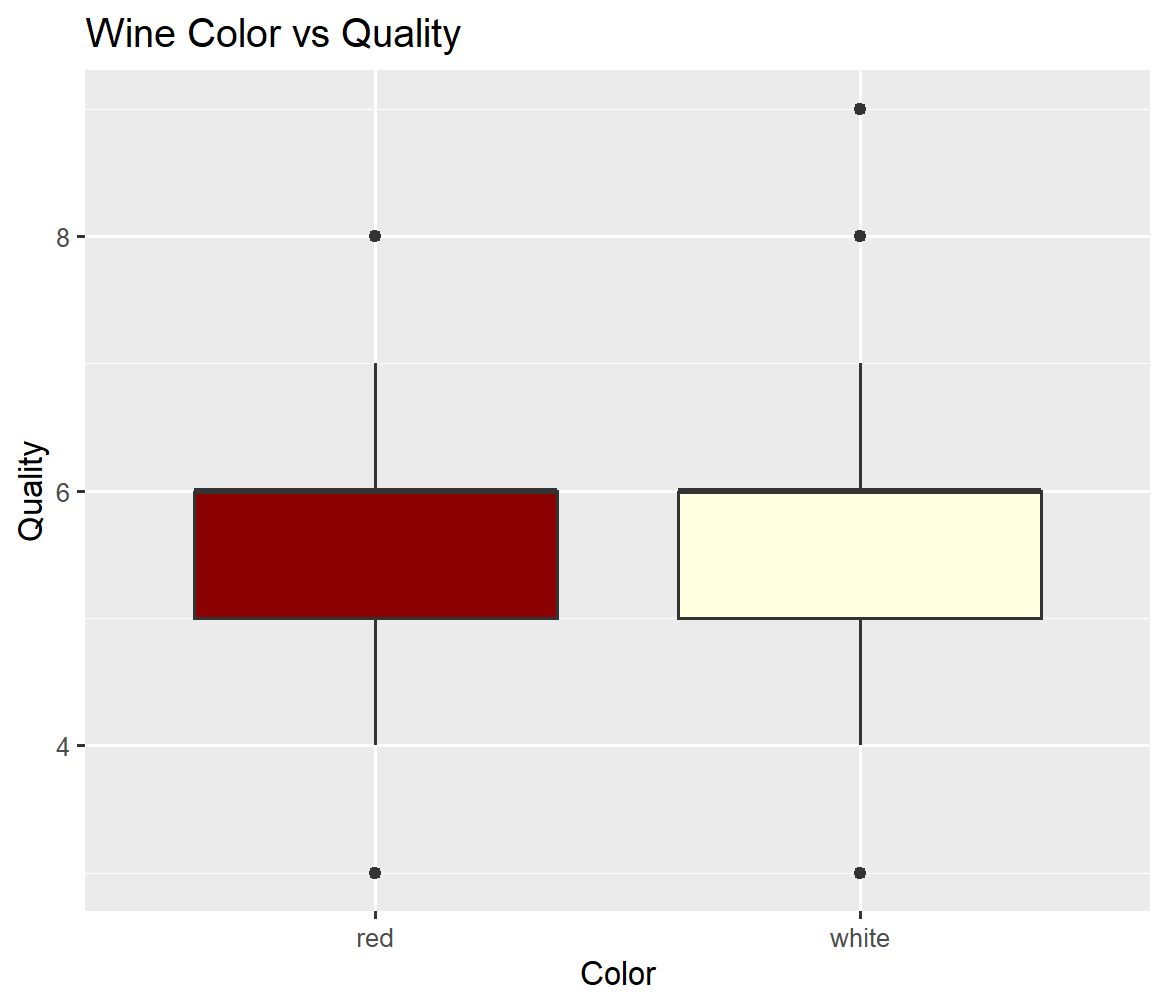
**Response Variable**

* **quality -** score between 1 - 10, 1 being the lowest quality and 10 being the highest

**Graphical Analysis and Hypothesis Testing**

Next is to visualize some of the data to get a better understanding of it, highlight graphs that seem more significant, and confirm or deny any assumptions made from the graphs.

The following is a boxplot showing quality vs wine color.



Visually, it seems that there is almost no difference in average quality between the red and white wines. However, as the following T test shows, there is a slight difference in mean quality between the two types.

**Welch Two Sample t-test**

**data: quality by color**

**t = -10.149, df = 2950.8, p-value < 2.2e-16**

**alternative hypothesis: true difference in means between group red and group white is not equal to 0**

**95 percent confidence interval:**

**-0.2886173 -0.1951564**

**sample estimates:**

**mean in group red mean in group white**

**5.636023 5.877909**

As we can see, we have a p-value of < 0.01, which means we can conclude there is a difference in mean quality between the colors. In fact, there appears to be a significant difference in the means of almost each variable between colors. To save space, I will just show the mean difference of each variable.

Fixed acidity:

mean in group red mean in group white

8.319637 6.854788

Volatile acidity:

mean in group red mean in group white

0.5278205 0.2782411

Citric acid:

mean in group red mean in group white

0.2709756 0.3341915

Residual sugar:

mean in group red mean in group white

2.538806 6.391415

Chlorides:

mean in group red mean in group white

0.08746654 0.04577236

Free sulfur dioxide:

mean in group red mean in group white

15.87492 35.30808

Total sulfur dioxide:

mean in group red mean in group white

46.46779 138.36066

Density:

mean in group red mean in group white

0.9967467 0.9940274

pH:

sample estimates:

mean in group red mean in group white

3.311113 3.188267

Sulphates:

mean in group red mean in group white

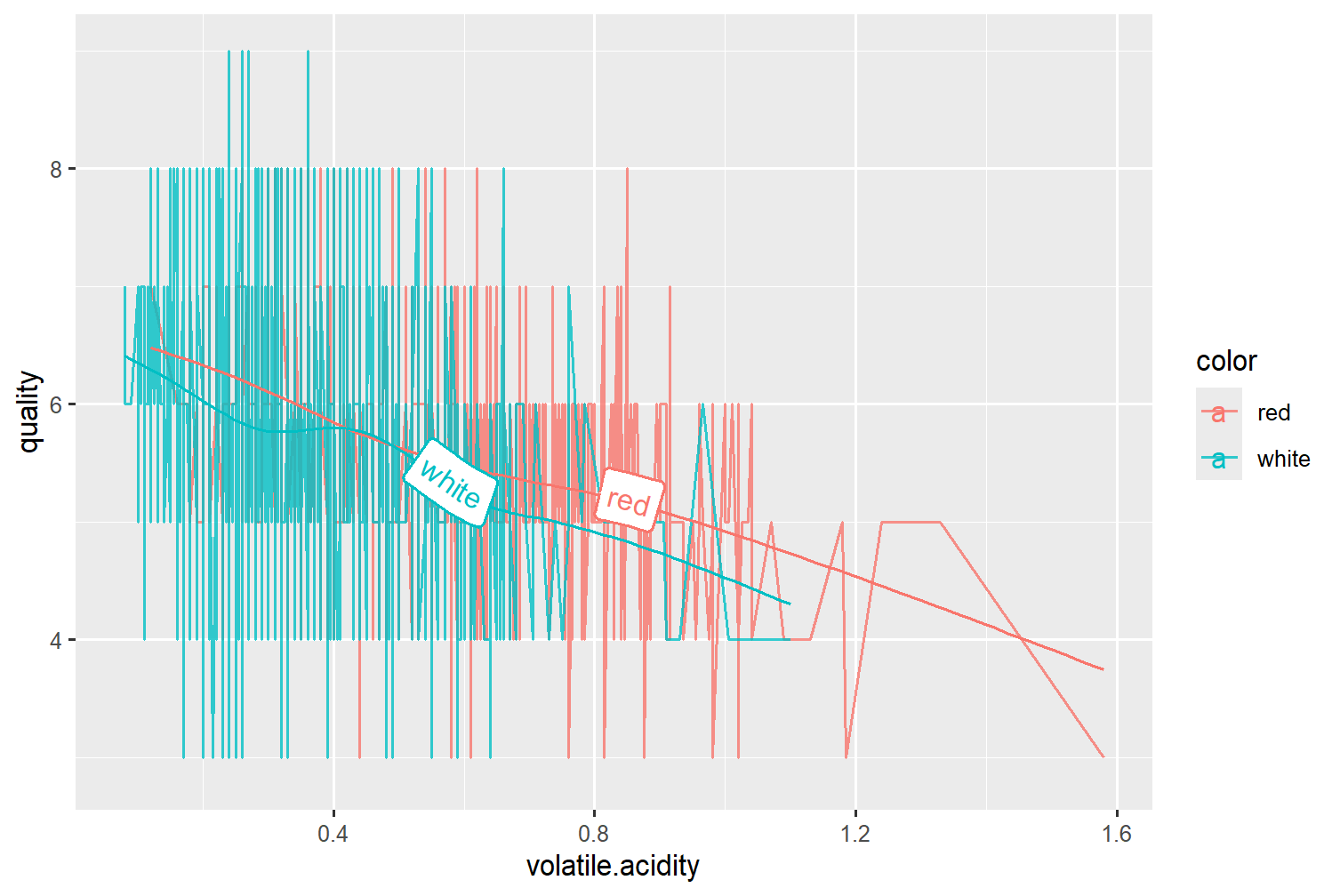
0.6581488 0.4898469

Alcohol:

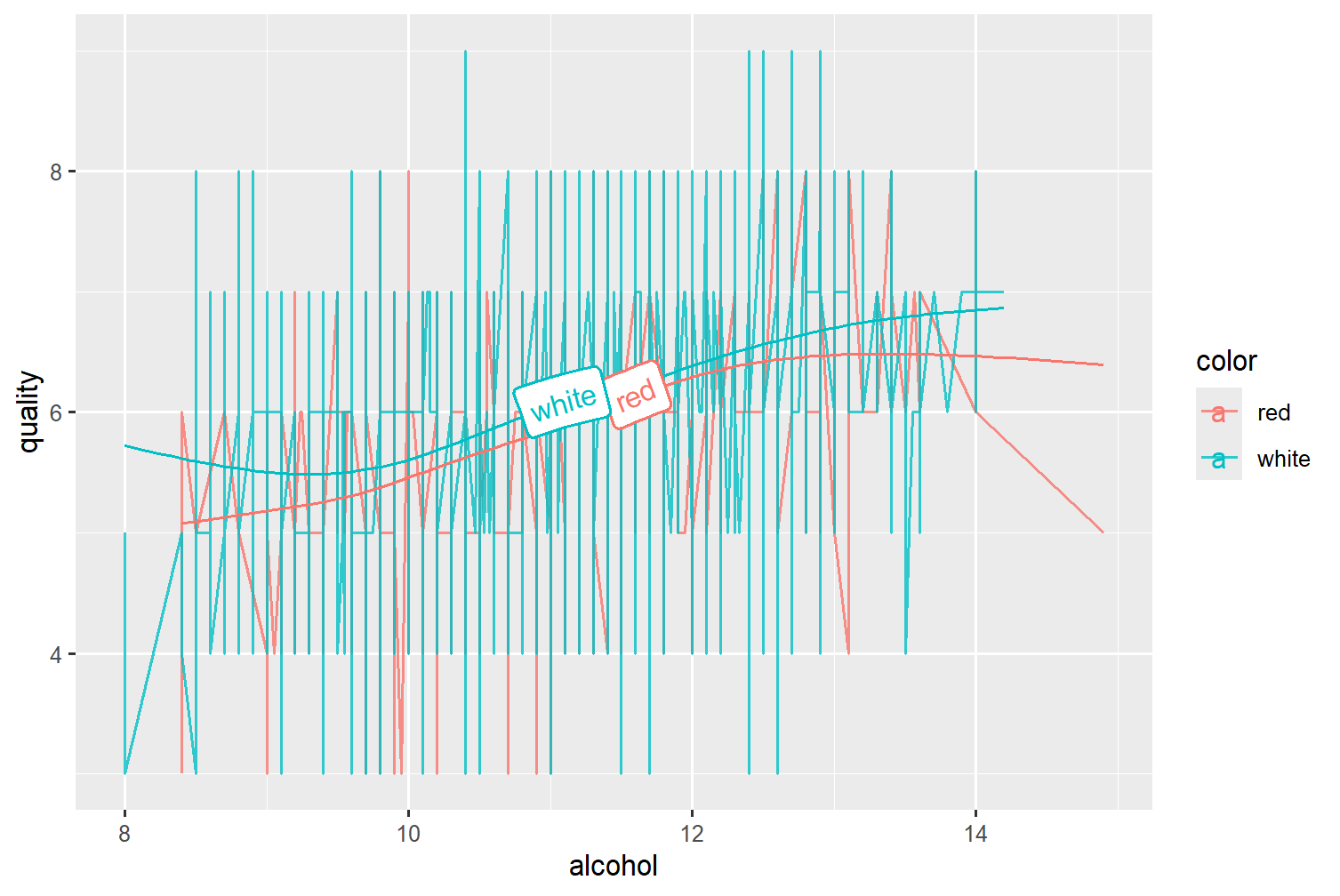
mean in group red mean in group white

10.42298 10.51427

All T tests had a p-value of less than 2.2e-16 except for alcohol, which had a p-value of 0.004278. This is still significant enough to conclude that there is a significant difference in mean, but is noticeably less significant than all other variables.

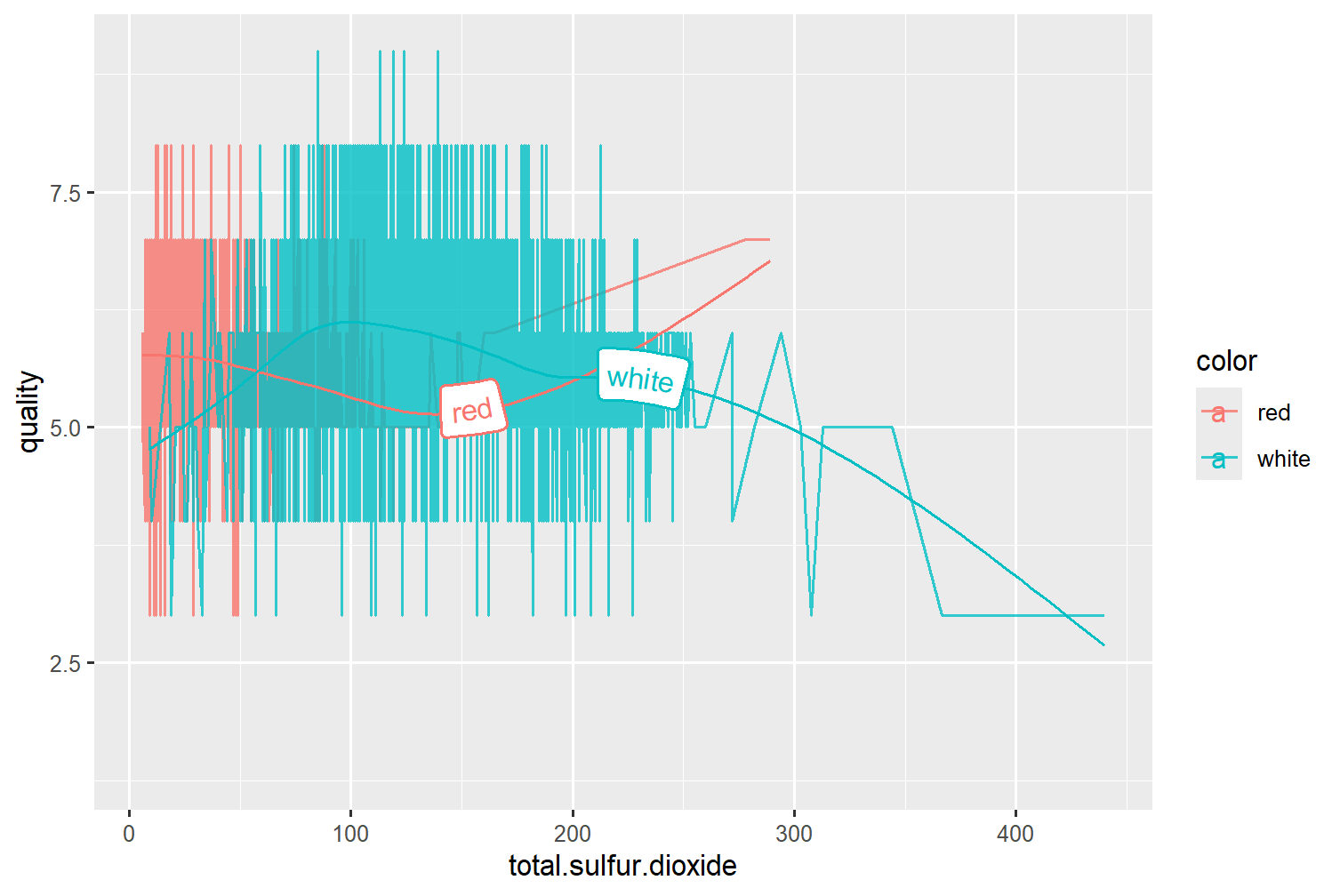
Here we will start showing some line graphs along with a mean line. The following shows quality vs volatile acidity.

As we can see, it appears that quality takes a large spike when volatile acidity increases. Let's look at a graph that seems to show the opposite effect.



This shows quality vs alcohol content. There seems to be a shared relation among white and red wines that a higher alcohol content relates to a higher quality of wine, though the increase seems to taper off when the alcohol reaches around 13%.

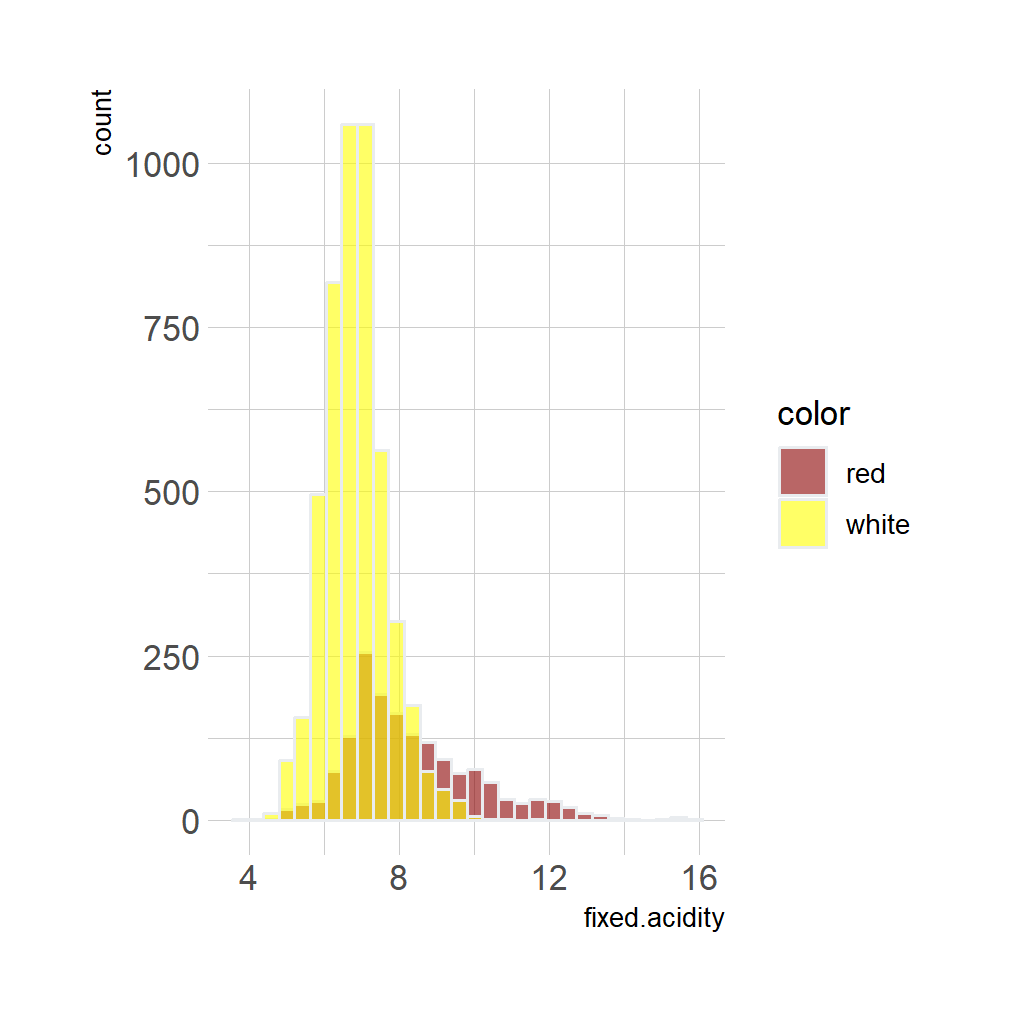
Another quite interesting visual is the line plot of quality vs total sulfur dioxide.



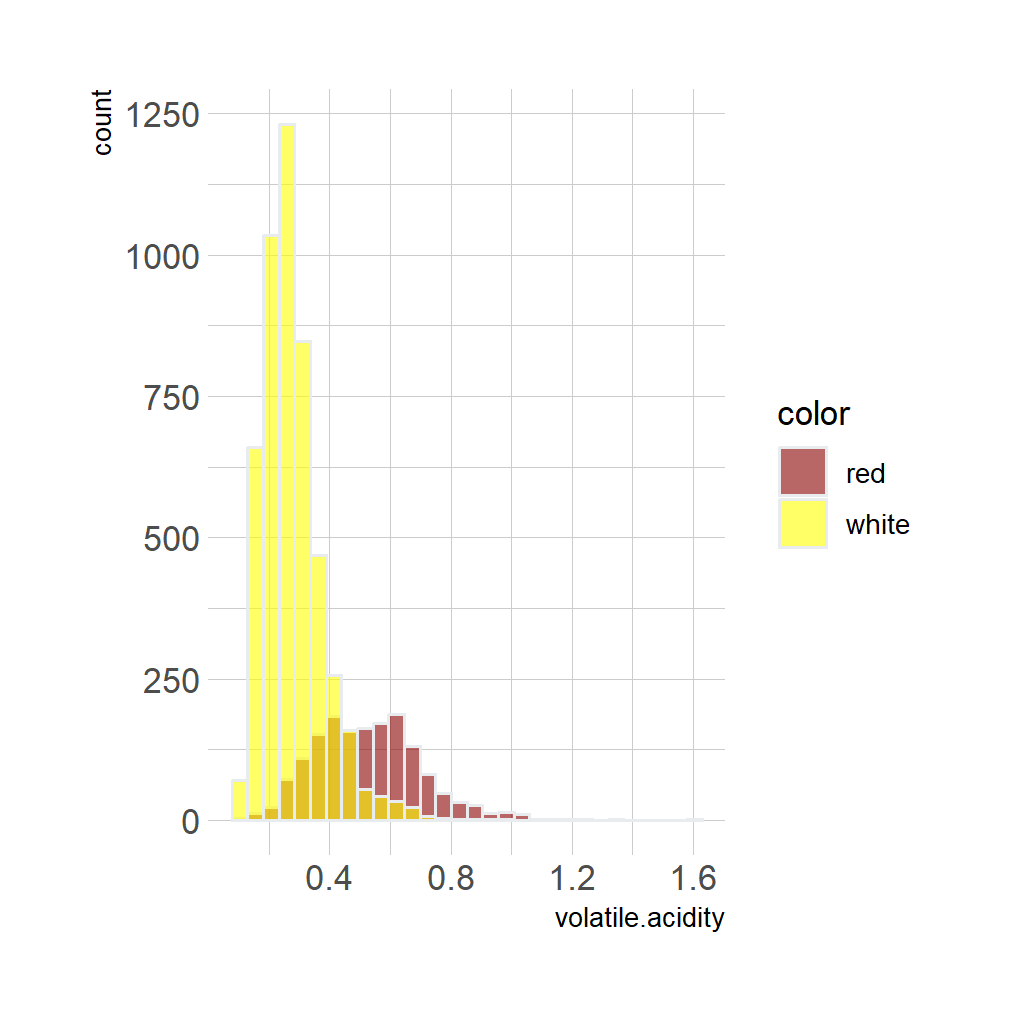
Though white wines seem to have a higher max total sulfur dioxide, We can see a clear inverse effect of total sulfur dioxide based on color. The red wines have a concave up relationship between quality and total sulfur dioxide, while the white wine has a mostly concave down relationship.This may be indicative that the same variable can have different or opposite effects depending on the color of a wine.

Next I will show some histograms detailing the differences in variable distribution among the two wine colors. This is mostly to demonstrate that we may need a different model for each color.

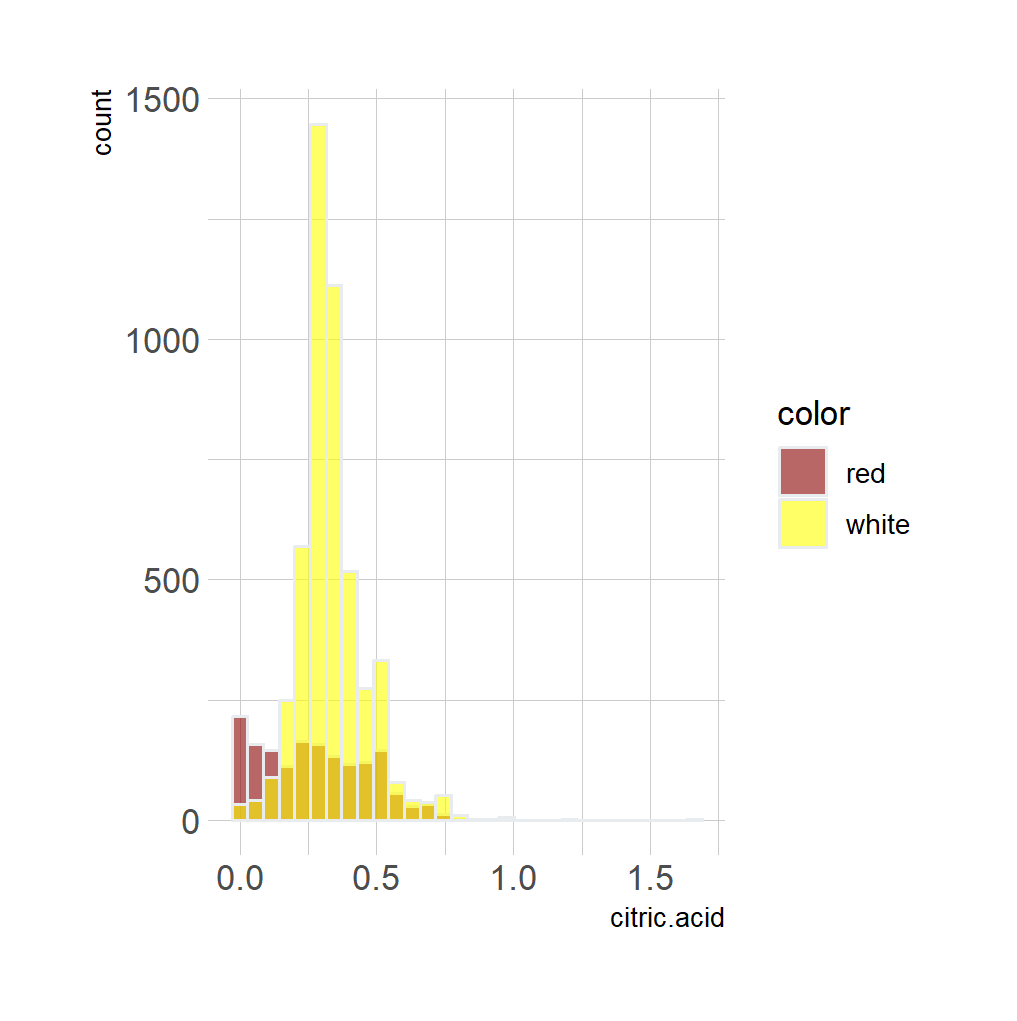
**Wine Fixed Acidity grouped by Color**



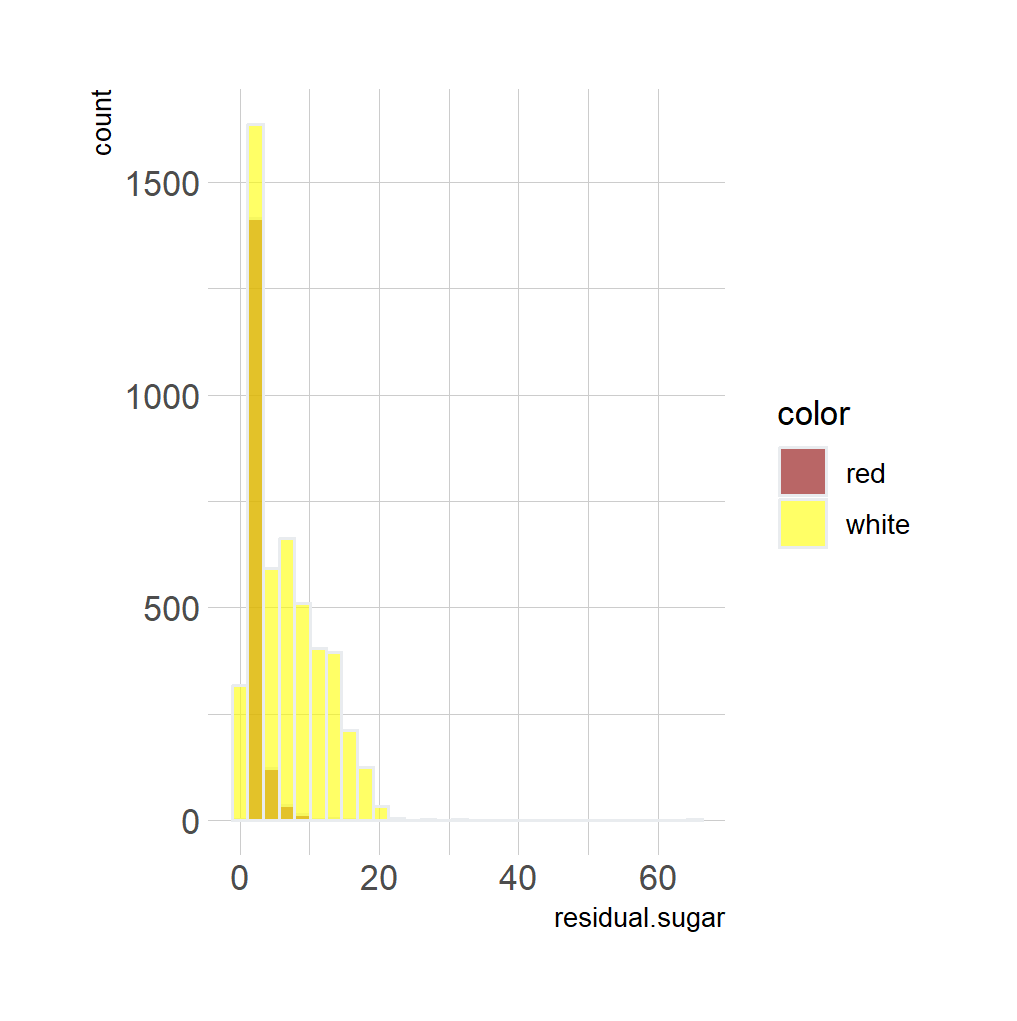
**Wine Volatile Acidity grouped by Color**

****

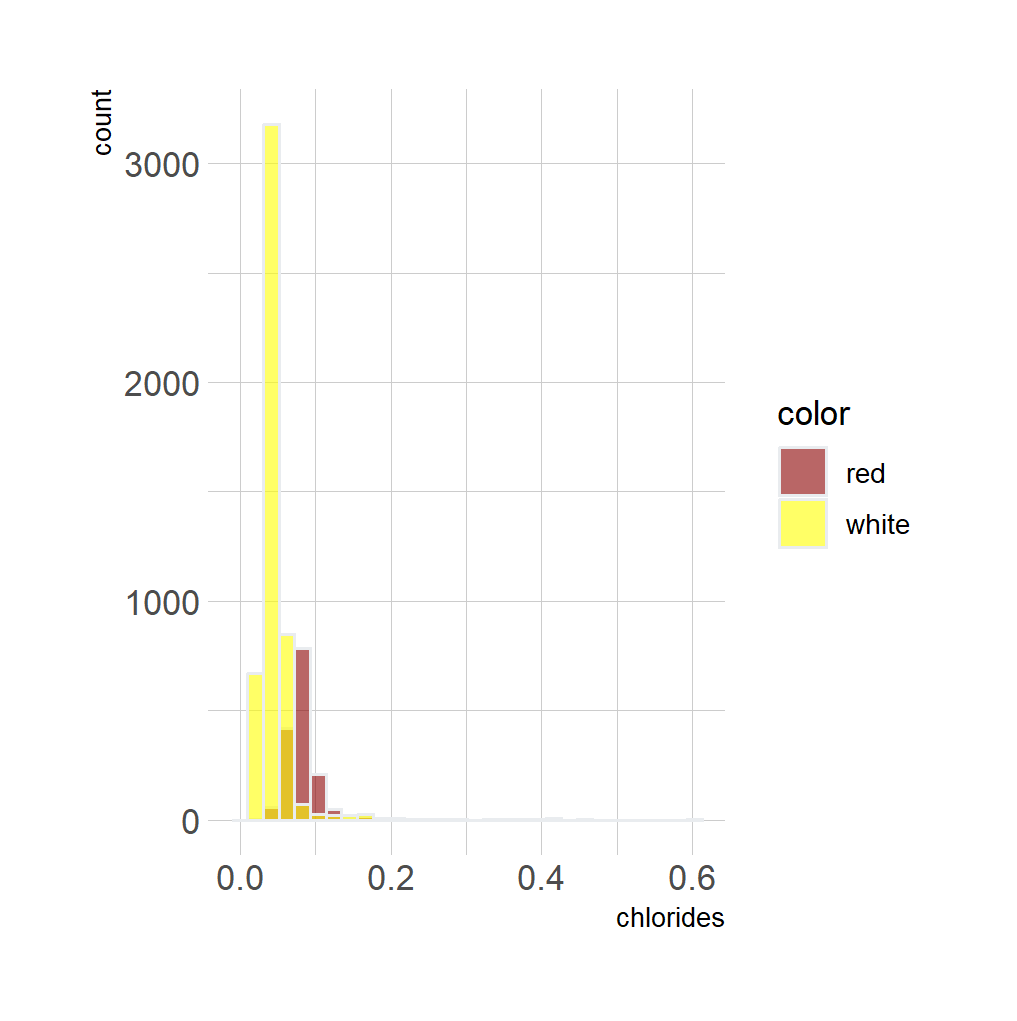
**Wine Citric Acid grouped by Color**

****

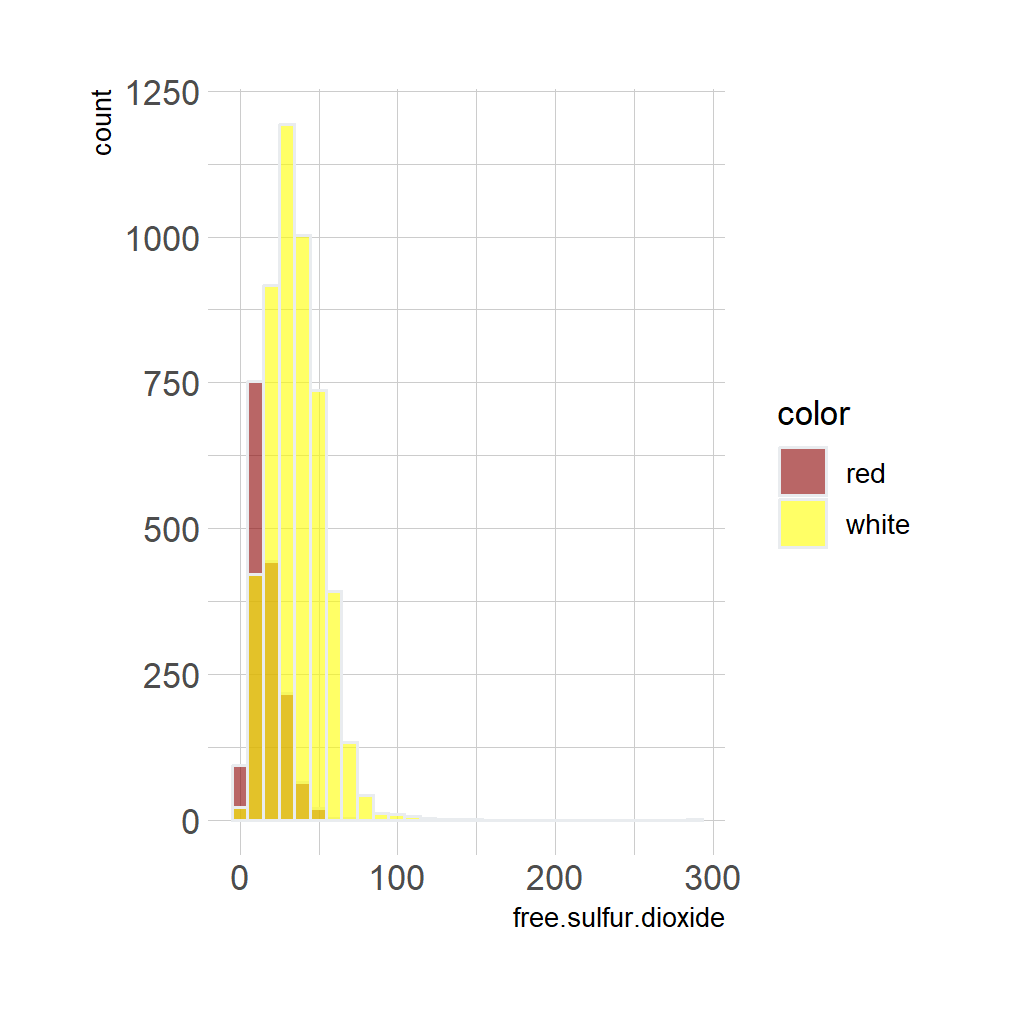
**Wine Residual Sugar grouped by Color**

****

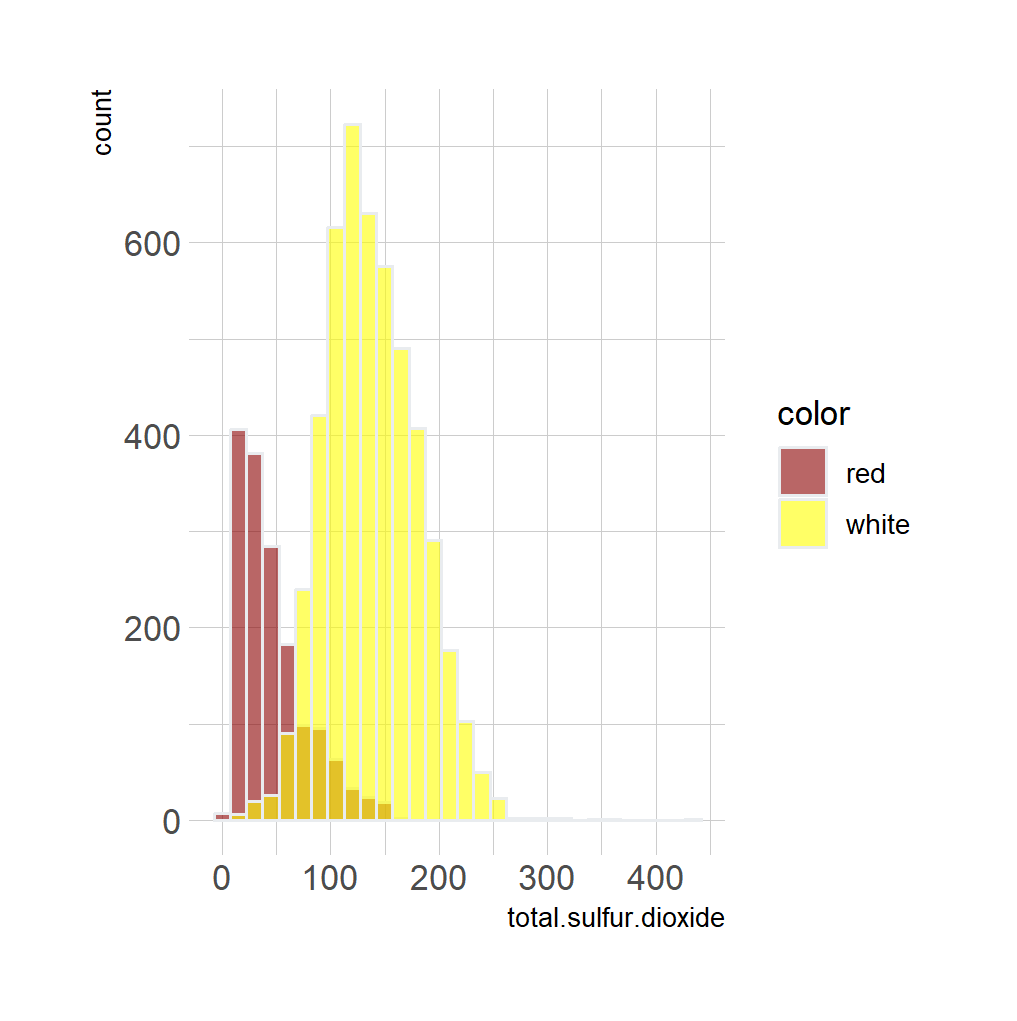
**Wine Chlorides grouped by Color**

****

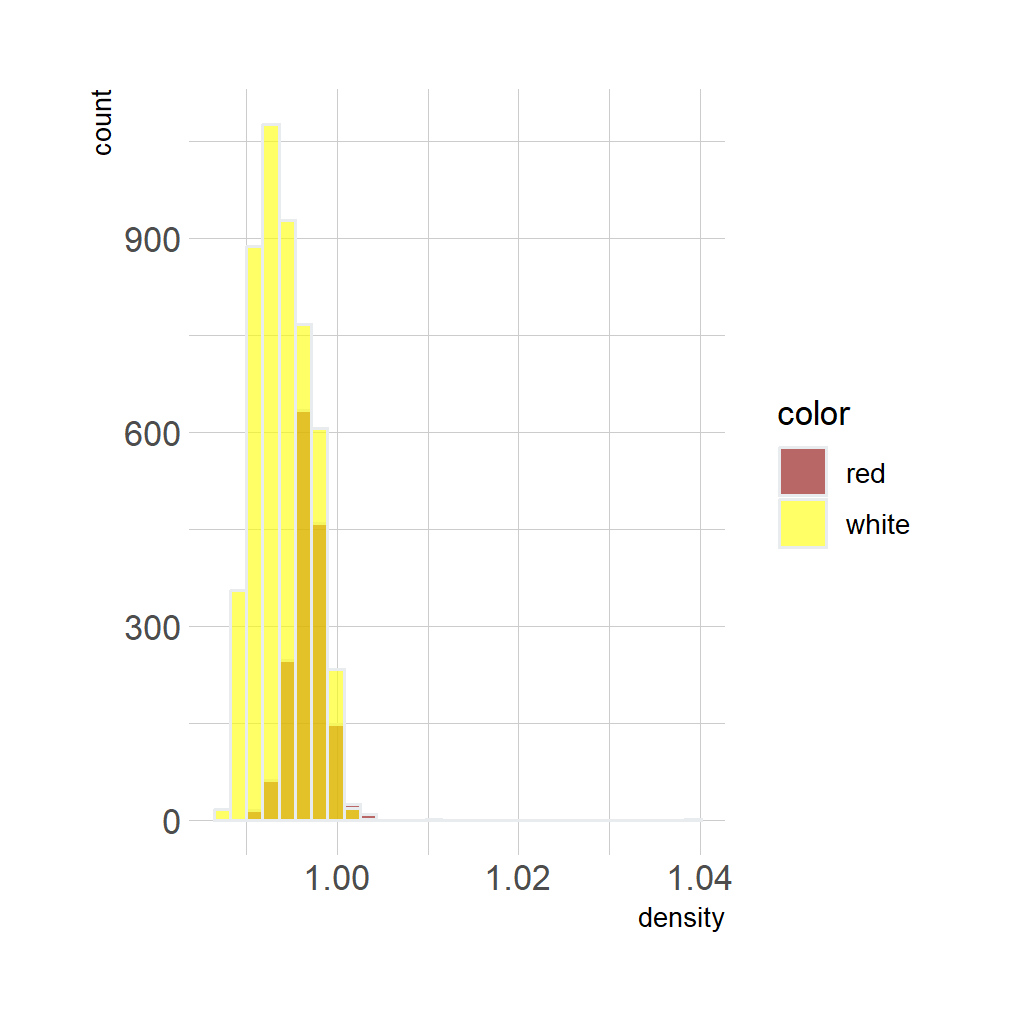
**Wine Free Sulfur Dioxides grouped by Color**

****

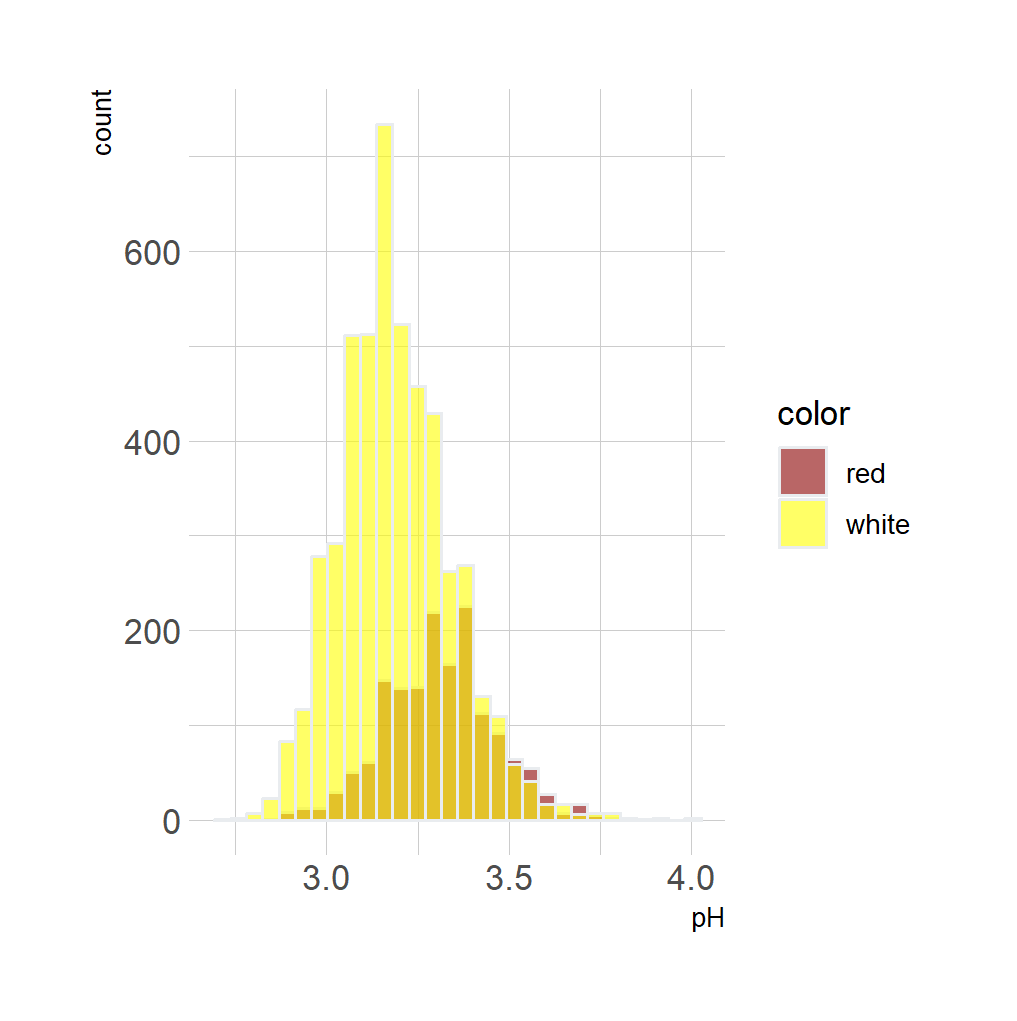
**Wine Total Sulfur Dioxides grouped by Color**

****

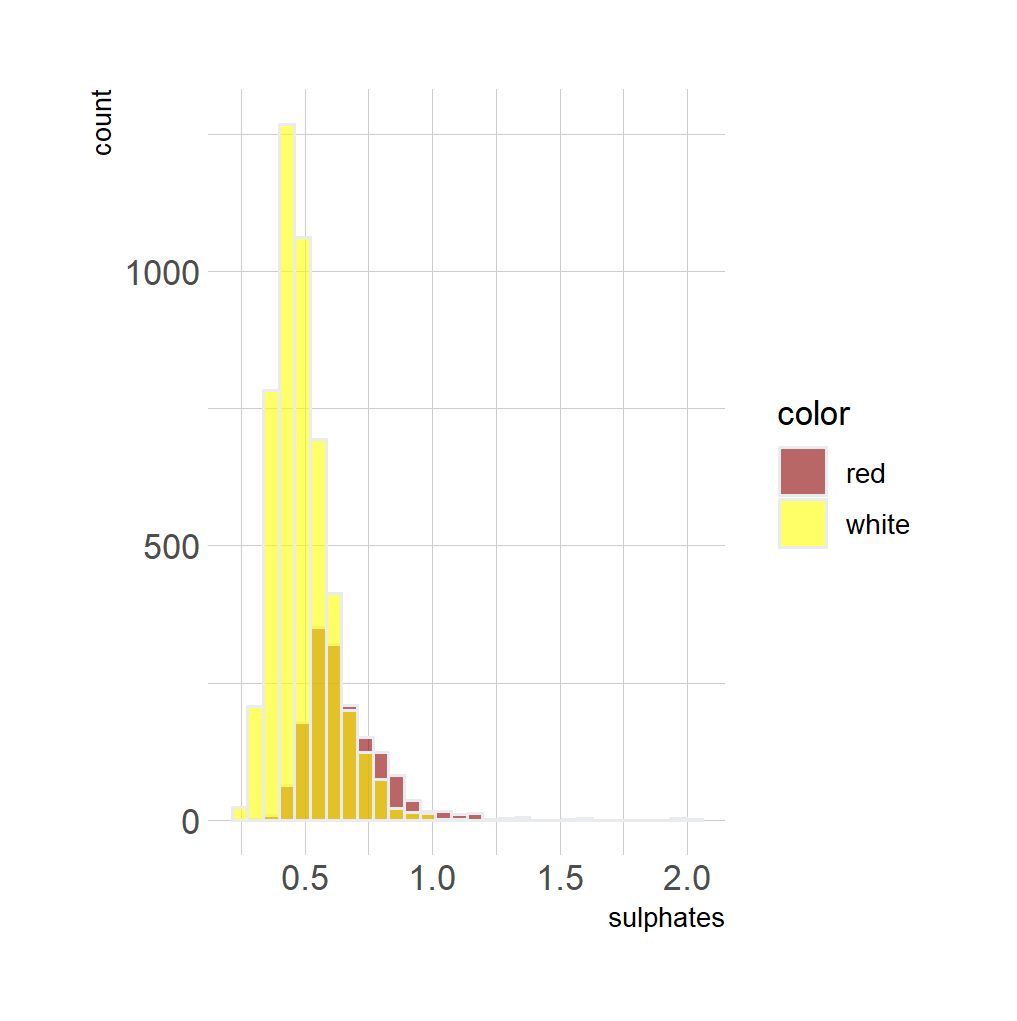
**Wine Density grouped by Color**

****

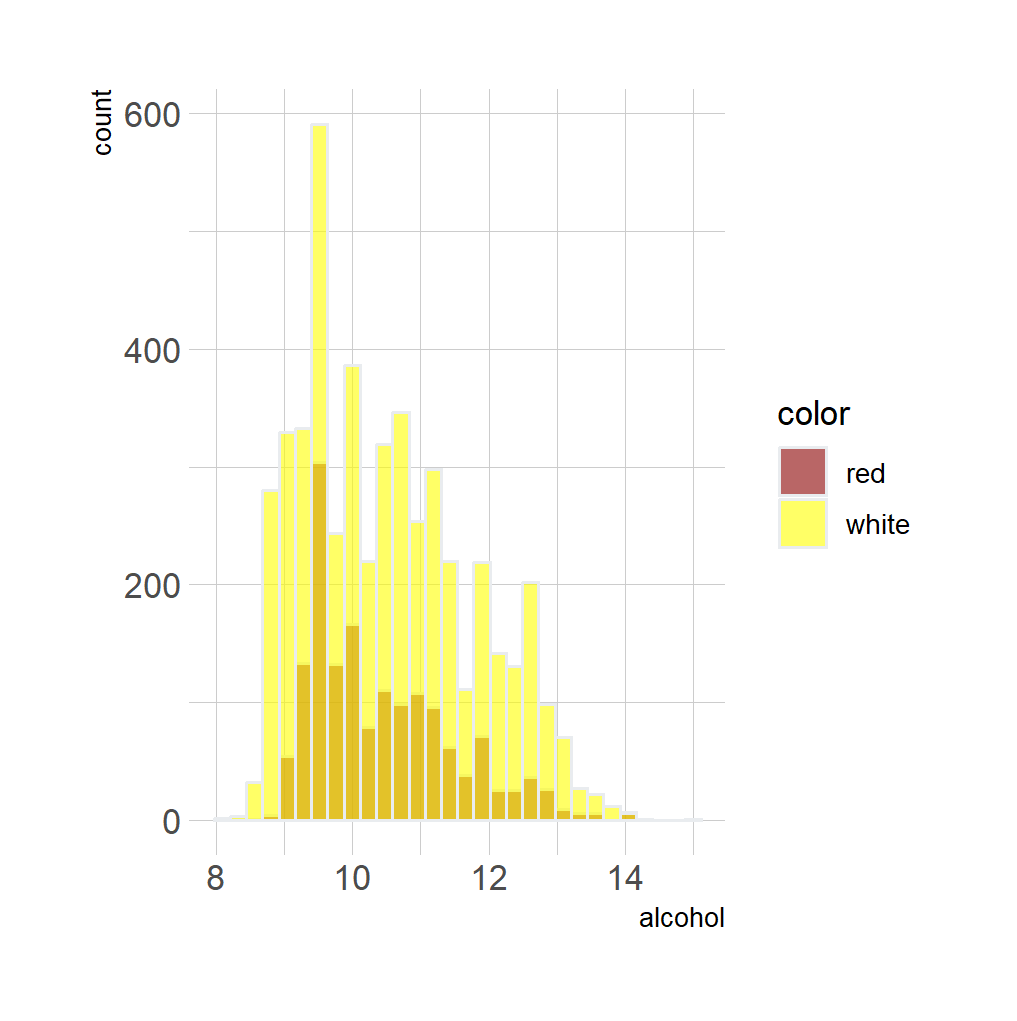
**Wine PH grouped by Color**

****

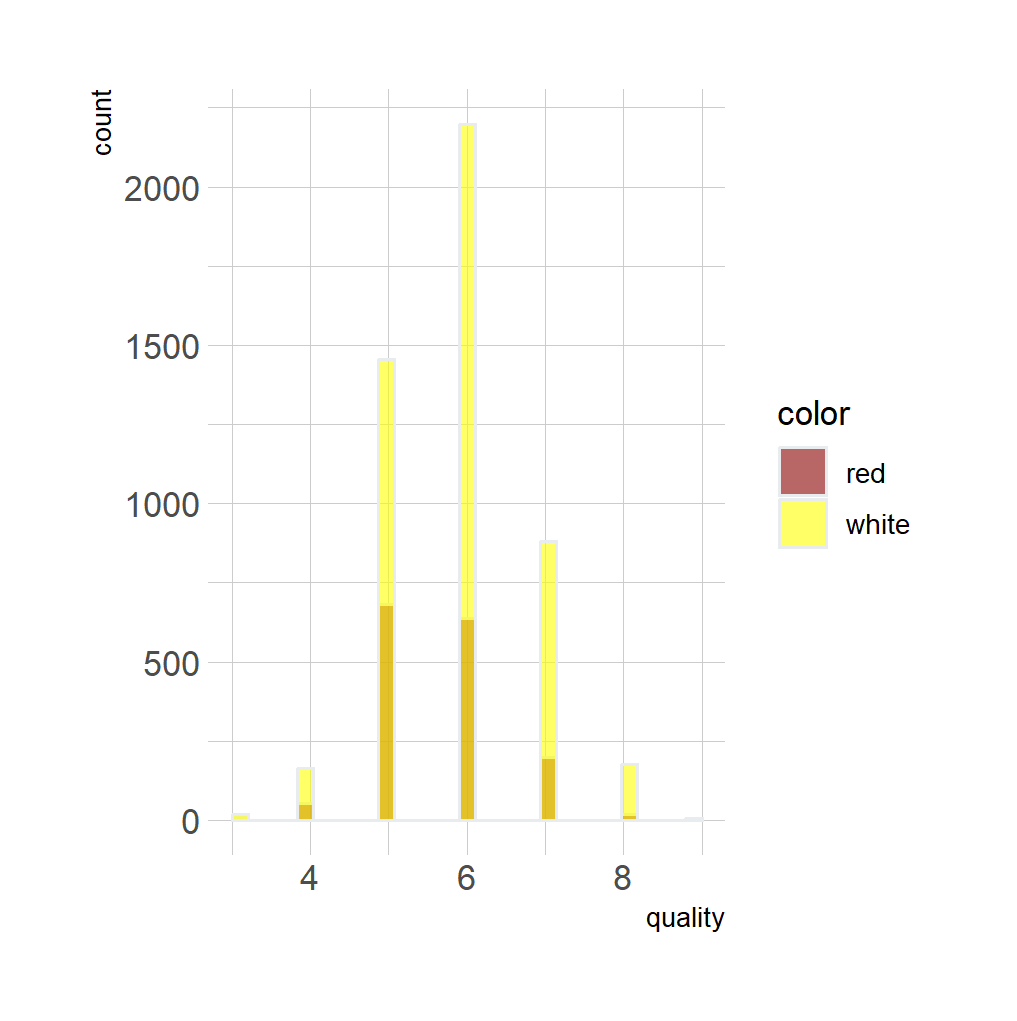
**Wine Sulphates grouped by Color**

****

**Wine Alcohol grouped by Color**

****

**Wine Quality grouped by Color**



**Modeling**

When I first started using this dataset, I was using the data frame that had both the white and red data mashed into one. Upon modeling however, I learned that this may have actually hampered the model’s ability to accurately represent the actual data. I found this to be true for both linear and logistic regression models.

**MULTIPLE LINEAR MODEL:**

**> summary(lmodel)**

**Call:**

**lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +**

**residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +**

**density + pH + sulphates + alcohol + color)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-3.7796 -0.4671 -0.0444 0.4561 3.0211**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 1.048e+02 1.414e+01 7.411 1.42e-13 \*\*\***

**fixed.acidity 8.507e-02 1.576e-02 5.396 7.05e-08 \*\*\***

**volatile.acidity -1.492e+00 8.135e-02 -18.345 < 2e-16 \*\*\***

**citric.acid -6.262e-02 7.972e-02 -0.786 0.4322**

**residual.sugar 6.244e-02 5.934e-03 10.522 < 2e-16 \*\*\***

**chlorides -7.573e-01 3.344e-01 -2.264 0.0236 \***

**free.sulfur.dioxide 4.937e-03 7.662e-04 6.443 1.25e-10 \*\*\***

**total.sulfur.dioxide -1.403e-03 3.237e-04 -4.333 1.49e-05 \*\*\***

**density -1.039e+02 1.434e+01 -7.248 4.71e-13 \*\*\***

**pH 4.988e-01 9.058e-02 5.506 3.81e-08 \*\*\***

**sulphates 7.217e-01 7.624e-02 9.466 < 2e-16 \*\*\***

**alcohol 2.227e-01 1.807e-02 12.320 < 2e-16 \*\*\***

**colorwhite -3.613e-01 5.675e-02 -6.367 2.06e-10 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.7331 on 6484 degrees of freedom**

**Multiple R-squared: 0.2965, Adjusted R-squared: 0.2952**

**F-statistic: 227.8 on 12 and 6484 DF, p-value: < 2.2e-16**

**MULTIPLE LOGISTIC MODEL:**

**> summary(gmodel)**

**Call:**

**glm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +**

**residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +**

**density + pH + sulphates + alcohol + color)**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 1.048e+02 1.414e+01 7.411 1.42e-13 \*\*\***

**fixed.acidity 8.507e-02 1.576e-02 5.396 7.05e-08 \*\*\***

**volatile.acidity -1.492e+00 8.135e-02 -18.345 < 2e-16 \*\*\***

**citric.acid -6.262e-02 7.972e-02 -0.786 0.4322**

**residual.sugar 6.244e-02 5.934e-03 10.522 < 2e-16 \*\*\***

**chlorides -7.573e-01 3.344e-01 -2.264 0.0236 \***

**free.sulfur.dioxide 4.937e-03 7.662e-04 6.443 1.25e-10 \*\*\***

**total.sulfur.dioxide -1.403e-03 3.237e-04 -4.333 1.49e-05 \*\*\***

**density -1.039e+02 1.434e+01 -7.248 4.71e-13 \*\*\***

**pH 4.988e-01 9.058e-02 5.506 3.81e-08 \*\*\***

**sulphates 7.217e-01 7.624e-02 9.466 < 2e-16 \*\*\***

**alcohol 2.227e-01 1.807e-02 12.320 < 2e-16 \*\*\***

**colorwhite -3.613e-01 5.675e-02 -6.367 2.06e-10 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**(Dispersion parameter for gaussian family taken to be 0.5374377)**

**Null deviance: 4953.7 on 6496 degrees of freedom**

**Residual deviance: 3484.7 on 6484 degrees of freedom**

**AIC: 14418**

**Number of Fisher Scoring iterations: 2**

**> nagelkerke(gmodel)**

**$Models**

**Model: "glm, quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol + color"**

**Null: "glm, quality ~ 1"**

**$Pseudo.R.squared.for.model.vs.null**

**Pseudo.R.squared**

**McFadden 0.137040**

**Cox and Snell (ML) 0.296535**

**Nagelkerke (Cragg and Uhler) 0.321200**

**$Likelihood.ratio.test**

**Df.diff LogLik.diff Chisq p.value**

**-12 -1142.6 2285.2 0**

**$Number.of.observations**

**Model: 6497**

**Null: 6497**

**$Messages**

**[1] "Note: For models fit with REML, these statistics are based on refitting with ML"**

**$Warnings**

**[1] "None"**

Both the linear and logistic models have relatively poor R-squared values, indicating that the models do not sufficiently represent the data. I initially thought that this simply meant that the data was unfit for a model, but I then had the idea to split the data frame into two, one containing only red wines and the other white- just like how the data came from UCI. This seemed to improve the R-squared value, but not by that much.

**MULTIPLE LINEAR MODEL OF WHITE WINE**

**> summary(whiteLmodel)**

**Call:**

**lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +**

**residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +**

**density + pH + sulphates + alcohol, data = white)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-3.8348 -0.4934 -0.0379 0.4637 3.1143**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 1.502e+02 1.880e+01 7.987 1.71e-15 \*\*\***

**fixed.acidity 6.552e-02 2.087e-02 3.139 0.00171 \*\***

**volatile.acidity -1.863e+00 1.138e-01 -16.373 < 2e-16 \*\*\***

**citric.acid 2.209e-02 9.577e-02 0.231 0.81759**

**residual.sugar 8.148e-02 7.527e-03 10.825 < 2e-16 \*\*\***

**chlorides -2.473e-01 5.465e-01 -0.452 0.65097**

**free.sulfur.dioxide 3.733e-03 8.441e-04 4.422 9.99e-06 \*\*\***

**total.sulfur.dioxide -2.857e-04 3.781e-04 -0.756 0.44979**

**density -1.503e+02 1.907e+01 -7.879 4.04e-15 \*\*\***

**pH 6.863e-01 1.054e-01 6.513 8.10e-11 \*\*\***

**sulphates 6.315e-01 1.004e-01 6.291 3.44e-10 \*\*\***

**alcohol 1.935e-01 2.422e-02 7.988 1.70e-15 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.7514 on 4886 degrees of freedom**

**Multiple R-squared: 0.2819, Adjusted R-squared: 0.2803**

**F-statistic: 174.3 on 11 and 4886 DF, p-value: < 2.2e-16**

**MULTIPLE LINEAR MODEL OF RED WINE**

**> summary(redLmodel)**

**Call:**

**lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +**

**residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +**

**density + pH + sulphates + alcohol, data = red)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-2.68911 -0.36652 -0.04699 0.45202 2.02498**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 2.197e+01 2.119e+01 1.036 0.3002**

**fixed.acidity 2.499e-02 2.595e-02 0.963 0.3357**

**volatile.acidity -1.084e+00 1.211e-01 -8.948 < 2e-16 \*\*\***

**citric.acid -1.826e-01 1.472e-01 -1.240 0.2150**

**residual.sugar 1.633e-02 1.500e-02 1.089 0.2765**

**chlorides -1.874e+00 4.193e-01 -4.470 8.37e-06 \*\*\***

**free.sulfur.dioxide 4.361e-03 2.171e-03 2.009 0.0447 \***

**total.sulfur.dioxide -3.265e-03 7.287e-04 -4.480 8.00e-06 \*\*\***

**density -1.788e+01 2.163e+01 -0.827 0.4086**

**pH -4.137e-01 1.916e-01 -2.159 0.0310 \***

**sulphates 9.163e-01 1.143e-01 8.014 2.13e-15 \*\*\***

**alcohol 2.762e-01 2.648e-02 10.429 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.648 on 1587 degrees of freedom**

**Multiple R-squared: 0.3606, Adjusted R-squared: 0.3561**

**F-statistic: 81.35 on 11 and 1587 DF, p-value: < 2.2e-16**

**MULTIPLE LOGISTIC MODEL FOR RED WINE**

**> summary(redGmodel)**

**Call:**

**glm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +**

**residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +**

**density + pH + sulphates + alcohol, data = red)**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 2.197e+01 2.119e+01 1.036 0.3002**

**fixed.acidity 2.499e-02 2.595e-02 0.963 0.3357**

**volatile.acidity -1.084e+00 1.211e-01 -8.948 < 2e-16 \*\*\***

**citric.acid -1.826e-01 1.472e-01 -1.240 0.2150**

**residual.sugar 1.633e-02 1.500e-02 1.089 0.2765**

**chlorides -1.874e+00 4.193e-01 -4.470 8.37e-06 \*\*\***

**free.sulfur.dioxide 4.361e-03 2.171e-03 2.009 0.0447 \***

**total.sulfur.dioxide -3.265e-03 7.287e-04 -4.480 8.00e-06 \*\*\***

**density -1.788e+01 2.163e+01 -0.827 0.4086**

**pH -4.137e-01 1.916e-01 -2.159 0.0310 \***

**sulphates 9.163e-01 1.143e-01 8.014 2.13e-15 \*\*\***

**alcohol 2.762e-01 2.648e-02 10.429 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**(Dispersion parameter for gaussian family taken to be 0.4199185)**

**Null deviance: 1042.17 on 1598 degrees of freedom**

**Residual deviance: 666.41 on 1587 degrees of freedom**

**AIC: 3164.3**

**Number of Fisher Scoring iterations: 2**

**> nagelkerke(redGmodel)**

**$Models**

**Model: "glm, quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol, red"**

**Null: "glm, quality ~ 1, red"**

**$Pseudo.R.squared.for.model.vs.null**

**Pseudo.R.squared**

**McFadden 0.185555**

**Cox and Snell (ML) 0.360552**

**Nagelkerke (Cragg and Uhler) 0.396138**

**$Likelihood.ratio.test**

**Df.diff LogLik.diff Chisq p.value**

**-11 -357.5 714.99 3.2945e-146**

**$Number.of.observations**

**Model: 1599**

**Null: 1599**

**$Messages**

**[1] "Note: For models fit with REML, these statistics are based on refitting with ML"**

**$Warnings**

**[1] "None"**

**> #logistic model of only red wine**

**> redGmodel = glm(quality~fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol, data = red)**

**> summary(redGmodel)**

**Call:**

**glm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +**

**residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +**

**density + pH + sulphates + alcohol, data = red)**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 2.197e+01 2.119e+01 1.036 0.3002**

**fixed.acidity 2.499e-02 2.595e-02 0.963 0.3357**

**volatile.acidity -1.084e+00 1.211e-01 -8.948 < 2e-16 \*\*\***

**citric.acid -1.826e-01 1.472e-01 -1.240 0.2150**

**residual.sugar 1.633e-02 1.500e-02 1.089 0.2765**

**chlorides -1.874e+00 4.193e-01 -4.470 8.37e-06 \*\*\***

**free.sulfur.dioxide 4.361e-03 2.171e-03 2.009 0.0447 \***

**total.sulfur.dioxide -3.265e-03 7.287e-04 -4.480 8.00e-06 \*\*\***

**density -1.788e+01 2.163e+01 -0.827 0.4086**

**pH -4.137e-01 1.916e-01 -2.159 0.0310 \***

**sulphates 9.163e-01 1.143e-01 8.014 2.13e-15 \*\*\***

**alcohol 2.762e-01 2.648e-02 10.429 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**(Dispersion parameter for gaussian family taken to be 0.4199185)**

**Null deviance: 1042.17 on 1598 degrees of freedom**

**Residual deviance: 666.41 on 1587 degrees of freedom**

**AIC: 3164.3**

**Number of Fisher Scoring iterations: 2**

**> nagelkerke(redGmodel)**

**$Models**

**Model: "glm, quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol, red"**

**Null: "glm, quality ~ 1, red"**

**$Pseudo.R.squared.for.model.vs.null**

**Pseudo.R.squared**

**McFadden 0.185555**

**Cox and Snell (ML) 0.360552**

**Nagelkerke (Cragg and Uhler) 0.396138**

**$Likelihood.ratio.test**

**Df.diff LogLik.diff Chisq p.value**

**-11 -357.5 714.99 3.2945e-146**

**$Number.of.observations**

**Model: 1599**

**Null: 1599**

**$Messages**

**[1] "Note: For models fit with REML, these statistics are based on refitting with ML"**

**$Warnings**

**[1] "None"**

**MULTIPLE LOGISTIC MODEL FOR WHITE WINE**

**> summary(whiteGmodel)**

**Call:**

**glm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +**

**residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +**

**density + pH + sulphates + alcohol, data = white)**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 1.502e+02 1.880e+01 7.987 1.71e-15 \*\*\***

**fixed.acidity 6.552e-02 2.087e-02 3.139 0.00171 \*\***

**volatile.acidity -1.863e+00 1.138e-01 -16.373 < 2e-16 \*\*\***

**citric.acid 2.209e-02 9.577e-02 0.231 0.81759**

**residual.sugar 8.148e-02 7.527e-03 10.825 < 2e-16 \*\*\***

**chlorides -2.473e-01 5.465e-01 -0.452 0.65097**

**free.sulfur.dioxide 3.733e-03 8.441e-04 4.422 9.99e-06 \*\*\***

**total.sulfur.dioxide -2.857e-04 3.781e-04 -0.756 0.44979**

**density -1.503e+02 1.907e+01 -7.879 4.04e-15 \*\*\***

**pH 6.863e-01 1.054e-01 6.513 8.10e-11 \*\*\***

**sulphates 6.315e-01 1.004e-01 6.291 3.44e-10 \*\*\***

**alcohol 1.935e-01 2.422e-02 7.988 1.70e-15 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**(Dispersion parameter for gaussian family taken to be 0.5645372)**

**Null deviance: 3841.0 on 4897 degrees of freedom**

**Residual deviance: 2758.3 on 4886 degrees of freedom**

**AIC: 11113**

**Number of Fisher Scoring iterations: 2**

**> nagelkerke(whiteGmodel)**

**$Models**

**Model: "glm, quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol, white"**

**Null: "glm, quality ~ 1, white"**

**$Pseudo.R.squared.for.model.vs.null**

**Pseudo.R.squared**

**McFadden 0.127604**

**Cox and Snell (ML) 0.281870**

**Nagelkerke (Cragg and Uhler) 0.304614**

**$Likelihood.ratio.test**

**Df.diff LogLik.diff Chisq p.value**

**-11 -810.88 1621.8 0**

**$Number.of.observations**

**Model: 4898**

**Null: 4898**

**$Messages**

**[1] "Note: For models fit with REML, these statistics are based on refitting with ML"**

**$Warnings**

**[1] "None"**

Let’s condense the information we’ve gotten so far.

**R-Squared for Models**

|  | Red | White | Combined Data (initial model) |
| --- | --- | --- | --- |
| Multiple Linear  (adj r2) | 0.3561 | 0.2803 | 2.952 |
| Multiple Logistic | 0.186 - 0.396 | 0.128 - 0.305 | 0.137 - 0.321 |

An interesting note to make is that while I did attempt to use stepwise regression to improve the model, Removing any variables no matter their significance, decreased the R-squared value. This told me that every variable had some impact on the quality of the wine.

Looking at the table now, we can clearly see the model for red wine improved overall compared to the combined model. On the other hand, The white wine model suffers a very slight reduction in r-squared compared to the combined model. To me, this tells me that there is something regarding the red wine data set that allows more accurate models to be formed from it. Perhaps this has to do with the size of the data sets - after all, the red wine only comprises about 25% of the dataset, while white comprises the other 75%. Overall, it would appear that the multiple logistic model for the red wine has the highest r-squared, making it the most fit for its dataset.

**Most Significant Variables**

When trying to calculate the most significant variables, I found that white and red wines give different weights to their variables. In other words, A variable that increases the quality of white wine may actually decrease the quality of red wine. The following will display the most significant variables impacting an average red wine and an average white wine.

**> print(avgRedWine[order(abs(avgRedWine$redcoefs), decreasing = T),])**

**varNames redWeights**

**8 density -17.82183062**

**11 alcohol 2.87882794**

**9 pH -1.36980753**

**10 sulphates 0.60306178**

**2 volatile.acidity -0.57215744**

**1 fixed.acidity 0.20790774**

**5 chlorides -0.16391230**

**7 total.sulfur.dioxide -0.15171734**

**6 free.sulfur.dioxide 0.06923053**

**3 citric.acid -0.04948015**

**4 residual.sugar 0.04145869**

**> print(avgWhiteWine[order(abs(avgWhiteWine$whitecoefs), decreasing = T),])**

**varNames whiteWeights**

**8 density -1.498110e+02**

**9 pH 2.272417e+00**

**11 alcohol 2.016847e+00**

**2 volatile.acidity -9.833296e-01**

**1 fixed.acidity 5.451026e-01**

**10 sulphates 4.156210e-01**

**4 residual.sugar 2.068619e-01**

**6 free.sulfur.dioxide 5.926108e-02**

**5 chlorides -2.163048e-02**

**7 total.sulfur.dioxide -1.327585e-02**

**3 citric.acid 5.985851e-03**

As we can see, the most impactful variables for red wine is as follows in decreasing order (+ means that, generally, quality will increase as variable increases, - means quality will generally decrease) :

1. Density -
2. Alcohol +
3. PH -
4. Sulphates +
5. Volatile acidity -
6. Fixed acidity +
7. Chlorides -
8. Total sulfur dioxide-
9. Free sulfur dioxide +
10. Citric acid -
11. Residual sugar +

And for white wine;

1. Density -
2. PH +
3. Alcohol +
4. Volatile acidity -
5. Fixed acidity +
6. Sulphates +
7. Residual sugar +
8. Free sulfur dioxide +
9. Chlorides -
10. Total sulfur dioxide -
11. Citric acid +

As you can see, though the weights of each variable are quite similar, there are subtle differences between white and red wines and how they are impacted by the variables. Generally, higher density will cause the wine's quality to degrade, while higher alcohol content will boost the quality, and these two variables have the most impact seemingly. Red wines seem to benefit from having a lower PH, while white wines benefit from a higher PH. Citric acid, though not as significant, shares this trait of having an opposite effect depending on wine, benefiting white wine while degrading red.

**Conclusion**

In retrospect, I wish I had chosen an easier dataset to work with. This dataset seemed to present a lot of trouble for me, since we did not necessarily cover ordinary distributions much in class. Since my response variable was an ordinal variable, I fear I may have not done the appropriate procedures for analyzing this dataset. If I could offer myself some constructive criticism, it would be to not rely purely on class notes and assignments to learn how to do my own statistical analysis. I feel if I spent more time looking online on how to deal with subjects I was not taught in class I could have done a better job on this project. However, it is finals season after all and I only have so much time to dedicate to each class.

**References:**

**Dataset:** <https://archive.ics.uci.edu/dataset/186/wine+quality>

References I used for information about the predictor variables:

<https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity>

<https://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity>

<https://whicherridge.com.au/blog/what-is-residual-sugar-in-wine/#:~:text=Residual%20sugar%20or%20'RS'%20is,to%20the%20addition%20of%20sugar>!

<https://www.oiv.int/standards/compendium-of-international-methods-of-wine-and-must-analysis/annex-d/annex-d-advices/level-of-sodium-and-chlorides-ions-in-wines>

<https://extension.okstate.edu/fact-sheets/understanding-free-sulfur-dioxide-fso2-in-wine.html>

<https://pmc.ncbi.nlm.nih.gov/articles/PMC10489813/#:~:text=The%20density%20of%20the%20wine,wines%20according%20to%20quality%20parameters>.