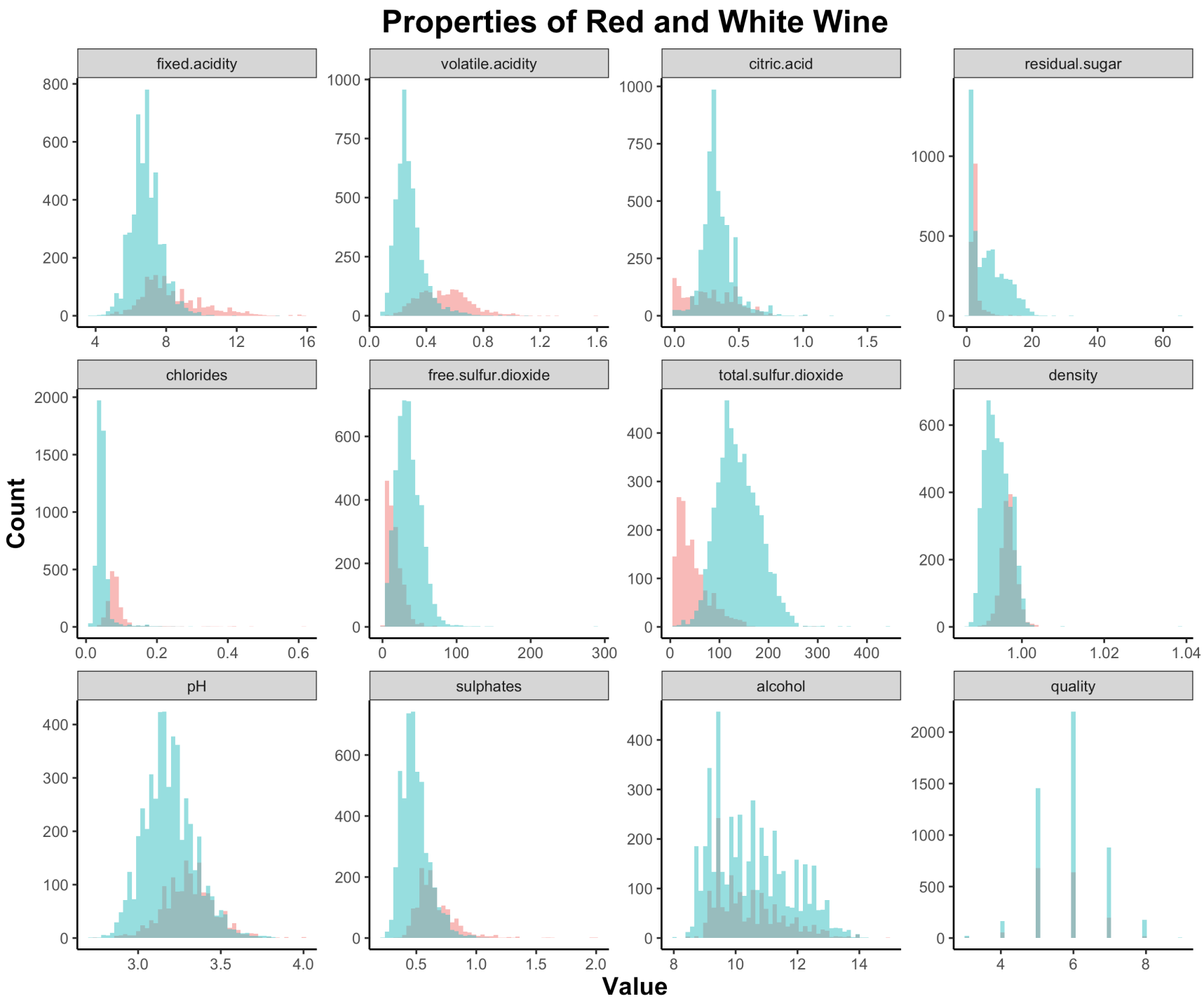
Exploring Wine Quality

Nick Wawee

Questions to Address

- What are the differences in distributions of between red and white wine?
- What are the characteristics of wines that are considered outliers based on their quality scores?
- Are the distributions of each type of wine symmetric before and after power transformation?
- Which variables are statistically different between the untransformed data?
- Which physicochemical properties impact the quality score the most?
  - What do the correlations between quality and physiochemical properties look like?
  - What do the regression coefficients in a multiple linear regression model tell us?

What are the differences between red and white wine?

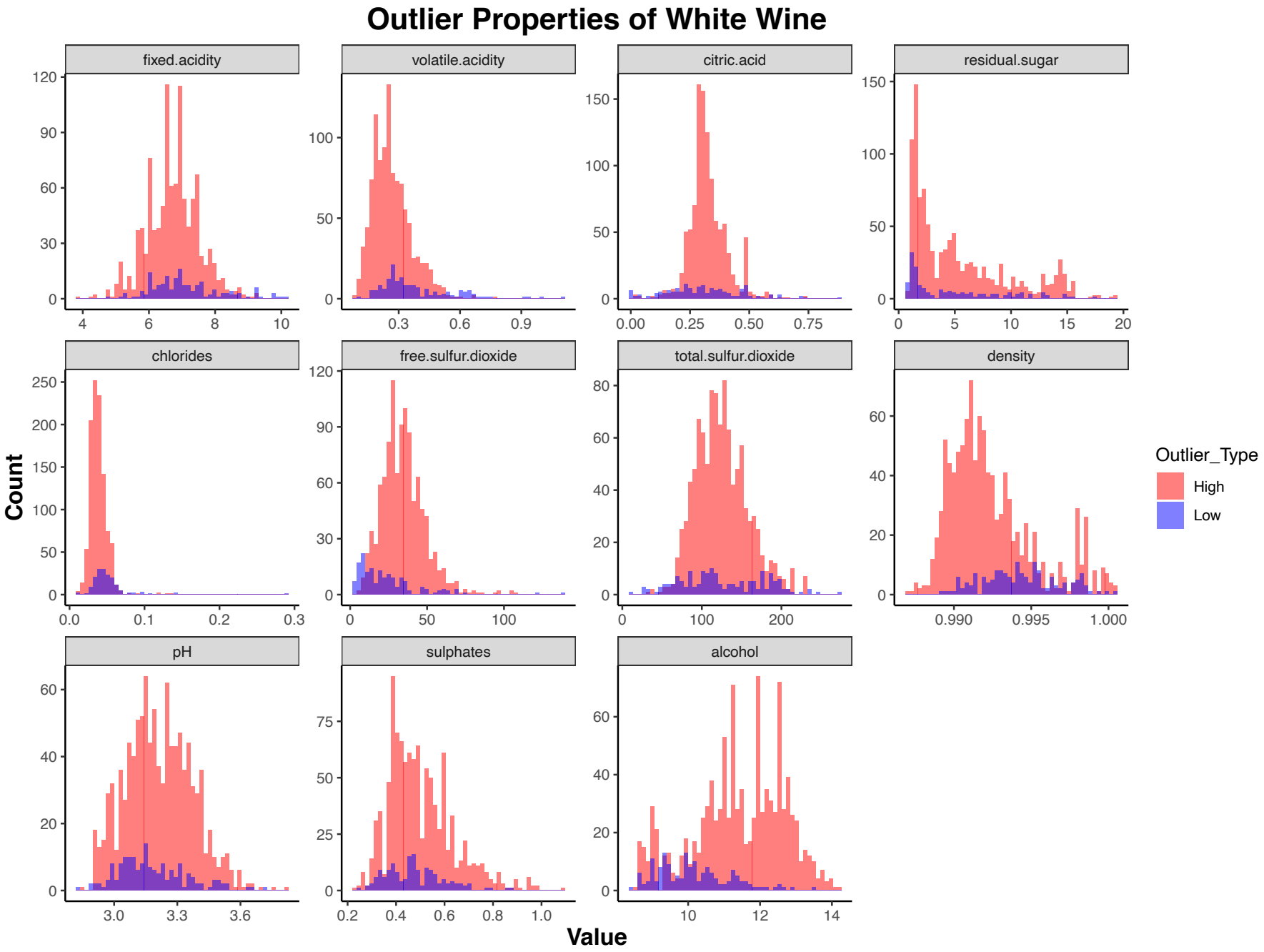


- Red wines have higher fixed and volatile acidity, but show to have lower citric acid and higher pH
- Median sulfur dioxides are higher in white wines compared to reds
- White wines appear to be less dense
- Red wines have subpopulations where sulphate concentration is higher
- White wines have lower chloride concentration
- Some white wines have higher residual sugar content

What are the characteristics of the outlier wines?

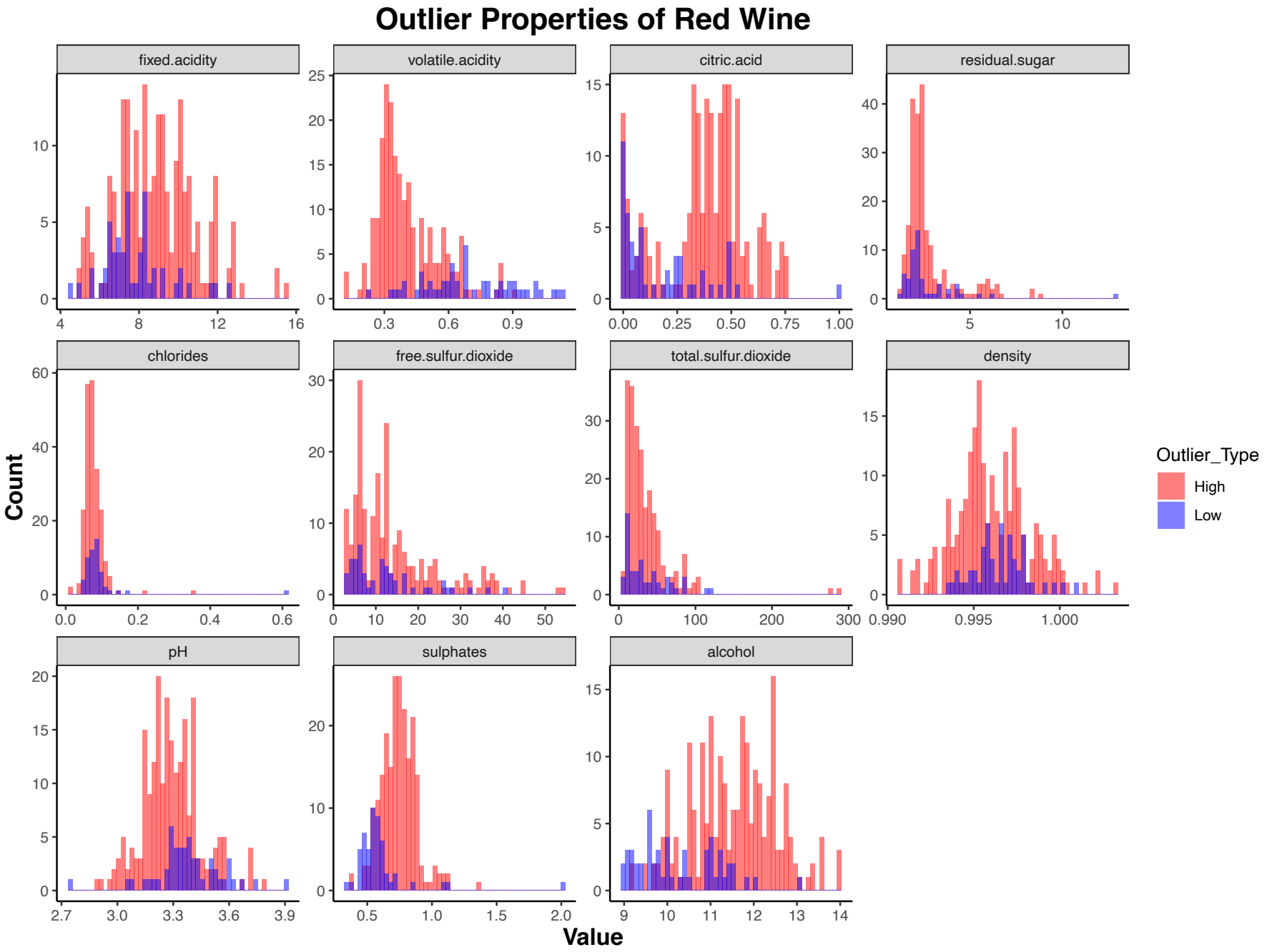
White Outlier Wines

- Outlier wines are defined by having quality scores outsides of the inner fences
- High alcohol content appears to be prevalent in wines with higher quality scores
- White wines that are denser appear to be more prevalent in lower quality wines
- The pH is higher in higher quality wines
- The free sulfur dioxide population is slightly elevated in higher quality wines



Red Outlier Wines

- **Higher quality wines:**
  - Low volatile acidity levels
  - Higher alcohol and sulphate content
  - Higher citric acid content
  - Slightly lower pH
  - Some have elevated total sulfur dioxide content



What does the symmetry look like in the data before and after power transformation?

Original Data

- **Red wine:** fixed acidity, total sulfur dioxide, sulphates, and alcohol are right skewed
- **White wine:** citric acid, residual sugar, total sulfur dioxides, and sulphates are right skewed
- **Both:**
  - Residual sugar, chlorides, free sulfur dioxide, and density appear to have ”squished” distributions because of their outliers
  - Quality levels are left skewed
- Power transformation may cause the distributions to be more symmetric and ”wider” to decrease the effect of outliers

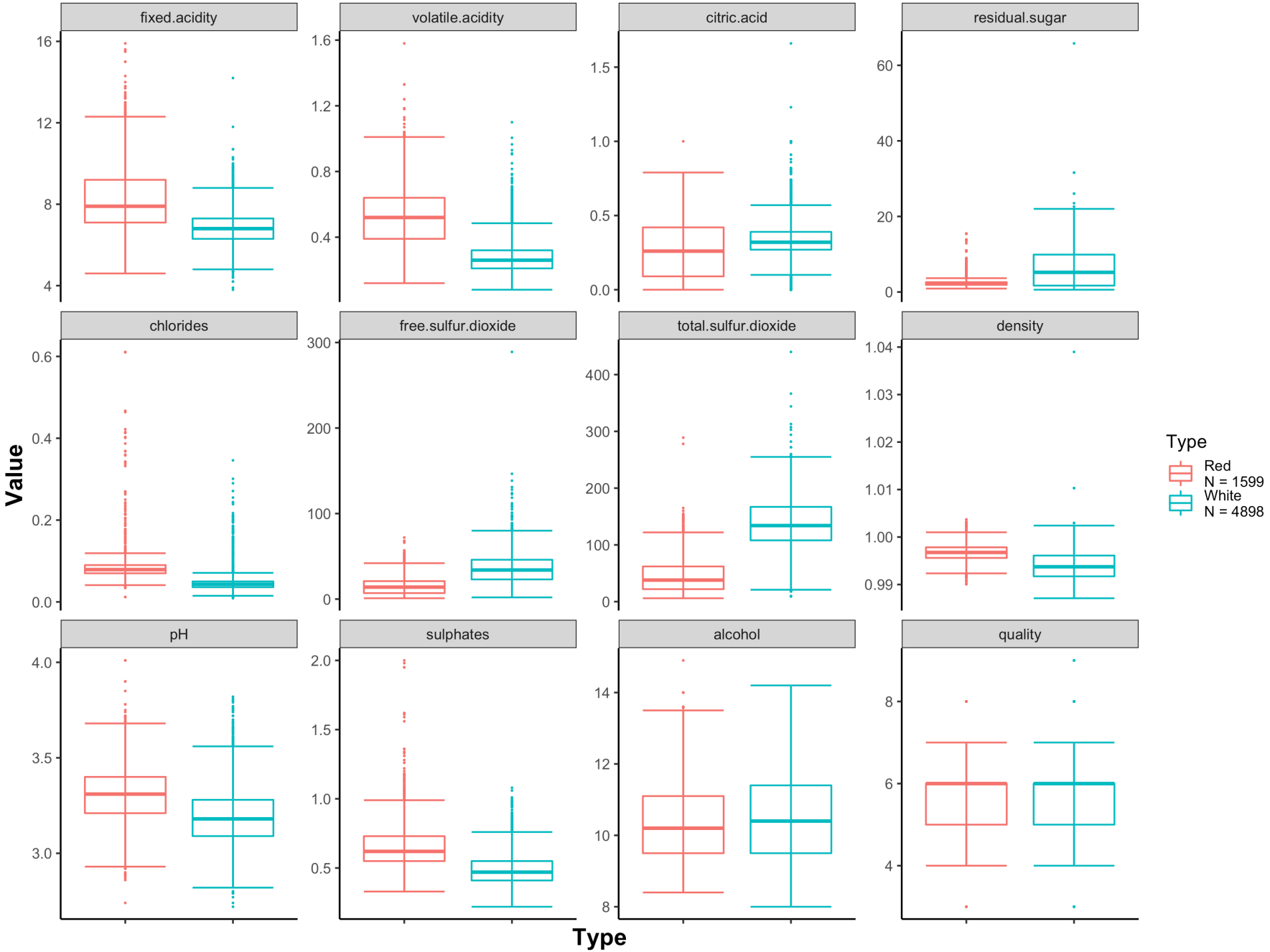


Table 3: Combined Results

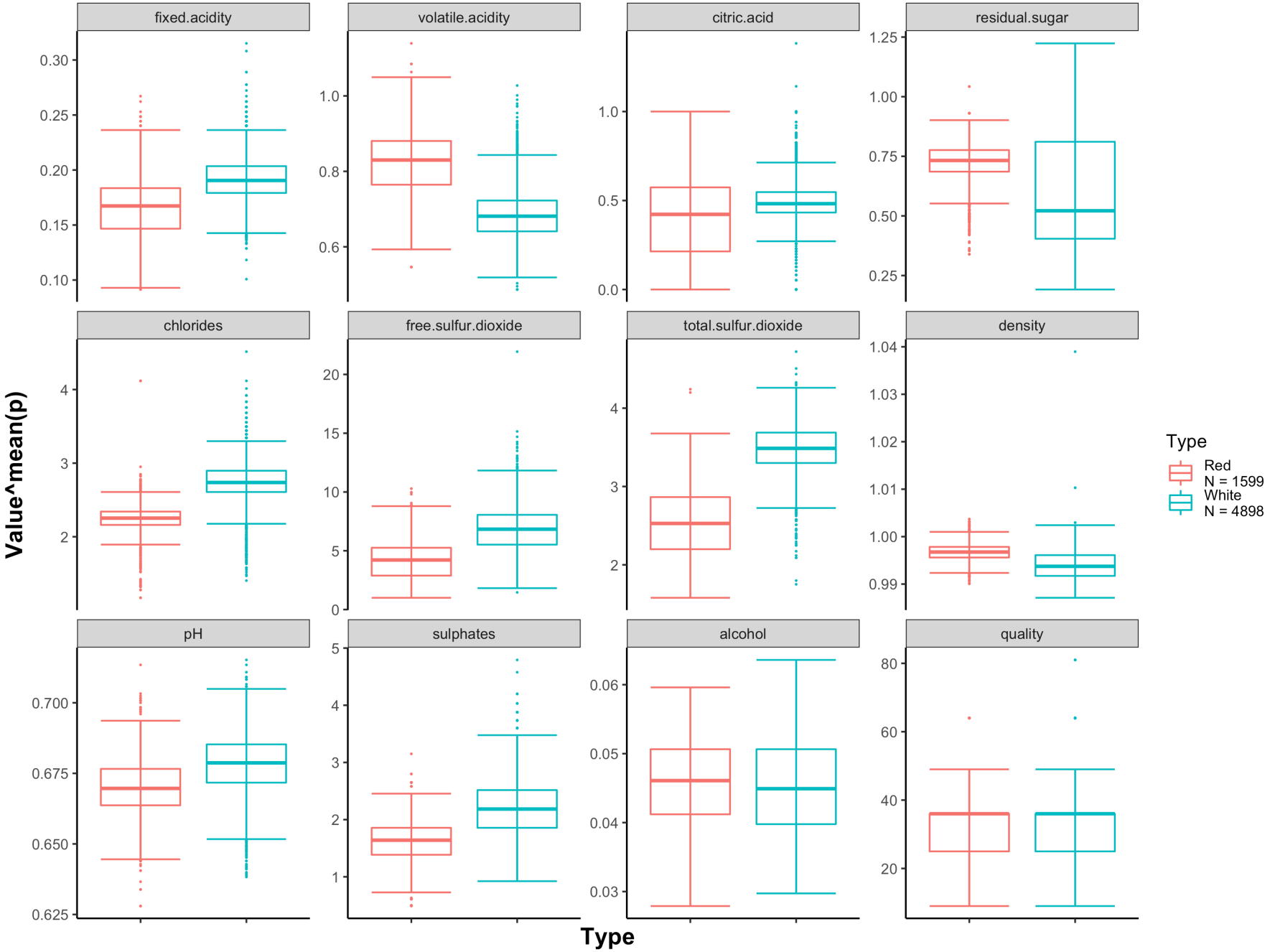
	Mean_Power
fixed.acidity	-0.865
volatile.acidity	0.285
citric.acid	0.640
residual.sugar	-0.395
chlorides	-0.320
free.sulfur.dioxide	0.545
total.sulfur.dioxide	0.255
density	0.000
pH	-0.335
sulphates	-1.035
alcohol	-1.325
quality	2.000

- The datasets were transformed by finding a power at which Hinkley (H) value is closest to zero[1]
  - $H = \frac{mean(x) - median(x)}{IQR(x)}$, x = variable of interest
  - H represents the skewness of the data
    - H < 0, indicates left skew
    - H > 0, indicates right skew
  - Minimum absolute value was taken to ensure there was not any skewness
  - Mean was taken of both powers (red and white) and each variable was transformed by the mean power (see Table 3)

[1]https://github.com/nickwawee/EDA/blob/master/minHinkley

Power Transformed

- **Red wine:** fixed acidity, total sulfur dioxide, sulphates, and alcohol appear to be symmetric
- **White wine:**
  - Total sulfur dioxides, and sulphates are more symmetric
  - Citric acid and residual sugar are still right skewed
- **Both:**
  - Residual sugar, chlorides, free sulfur dioxide, and density appear to be less ”squished”
  - Quality levels are still left skewed



Which variables are statistically different between the untransformed data?

Table 4: t-test Results

	t	p
volatile.acidity	-53.059	0.000
density	-42.709	0.000
sulphates	-37.056	0.000
chlorides	-34.240	0.000
fixed.acidity	-32.423	0.000
pH	-27.775	0.000
alcohol	2.859	0.004
quality	10.149	0.000
citric.acid	12.229	0.000
residual.sugar	47.802	0.000
free.sulfur.dioxide	54.428	0.000
total.sulfur.dioxide	89.872	0.000

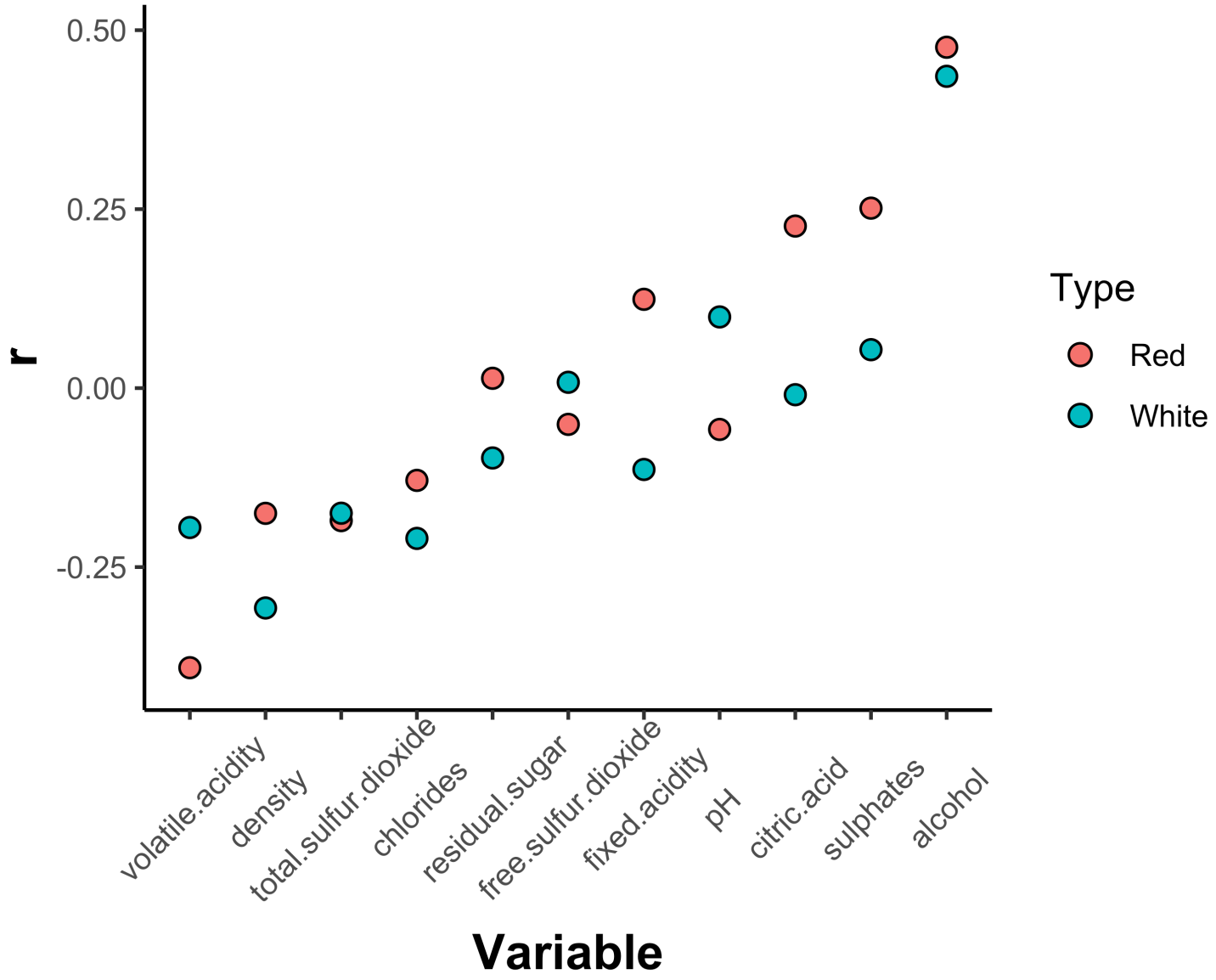
Testing Variables for Differences

- Two-sample independent t test was employed on the red and white wine untransformed variables
  - Untransformed data was used to interpret the differences in the measurement of the properties
  - Red N = 1599, White N = 4898
- All variables turned out to have significantly different means (probably due to large N, small SEM)
- Negative t-values indicate that red wine has a higher mean
- Positive t-values indicate that white wine has a higher mean

Which factors impact the quality score the most?

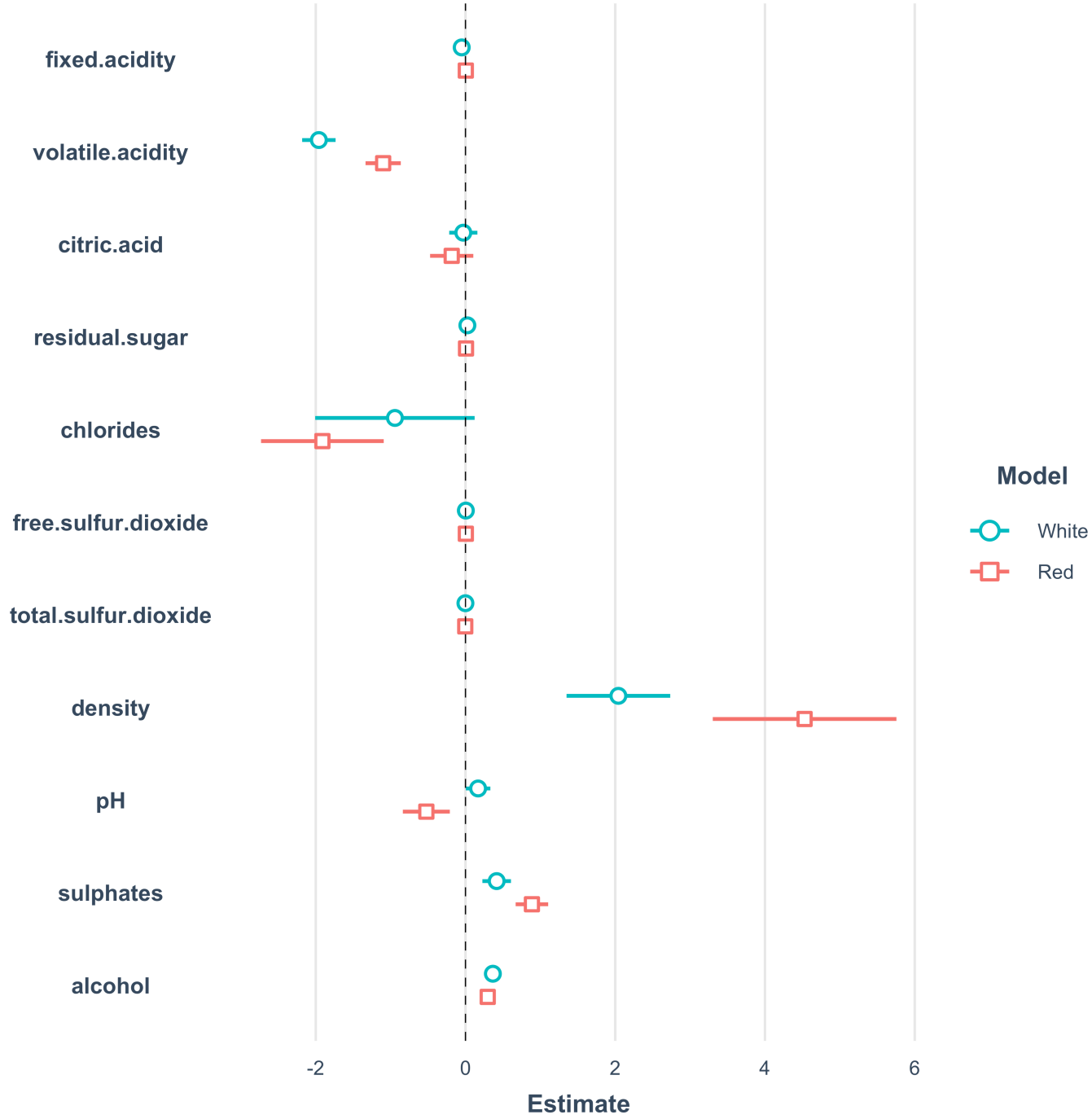
Pearson Correlations Between Wine Quality and Variables

- Alcohol is the most correlated with quality
- Volatile acidity is the most anti correlated between the two wines
- Sulphates and citric acid is slightly correlated with quality for red wines, while they show little correlation in white wines



MLR Regression Coefficients

- **Model Details:**
  - R^2: White = 0.98, Red = 0.99
  - Does not have intercept term
- Density appears to impact quality the most
- Chlorides negatively impact quality for red wines
- Volatile acidity negative impacts quality for white wines
- Sulphates and alcohol content slightly impact quality
- pH positively impacts quality white wines but negatively impacts quality of red wines at a higher magnitude



Conclusions

- Red wines have less sulfur dioxide and more chlorides
- White wines are more acidic (pH is lower) and are less dense
- Some of the distributions are skewed and power transformation makes some distributions not skewed
- Higher quality wines have more alcohol
- According to the MLR models:
  - Density, sulphate content, and alcohol content positively impact quality score for both types of wine
  - Volatile acidity negatively impacts the quality score for both wines
  - pH negatively impacts quality for red wines, but positively impacts white wines
  - Chloride content negatively impacts the quality of red wines

Thank You!

Any Questions?

Appendix: Evaluating Wine Quality

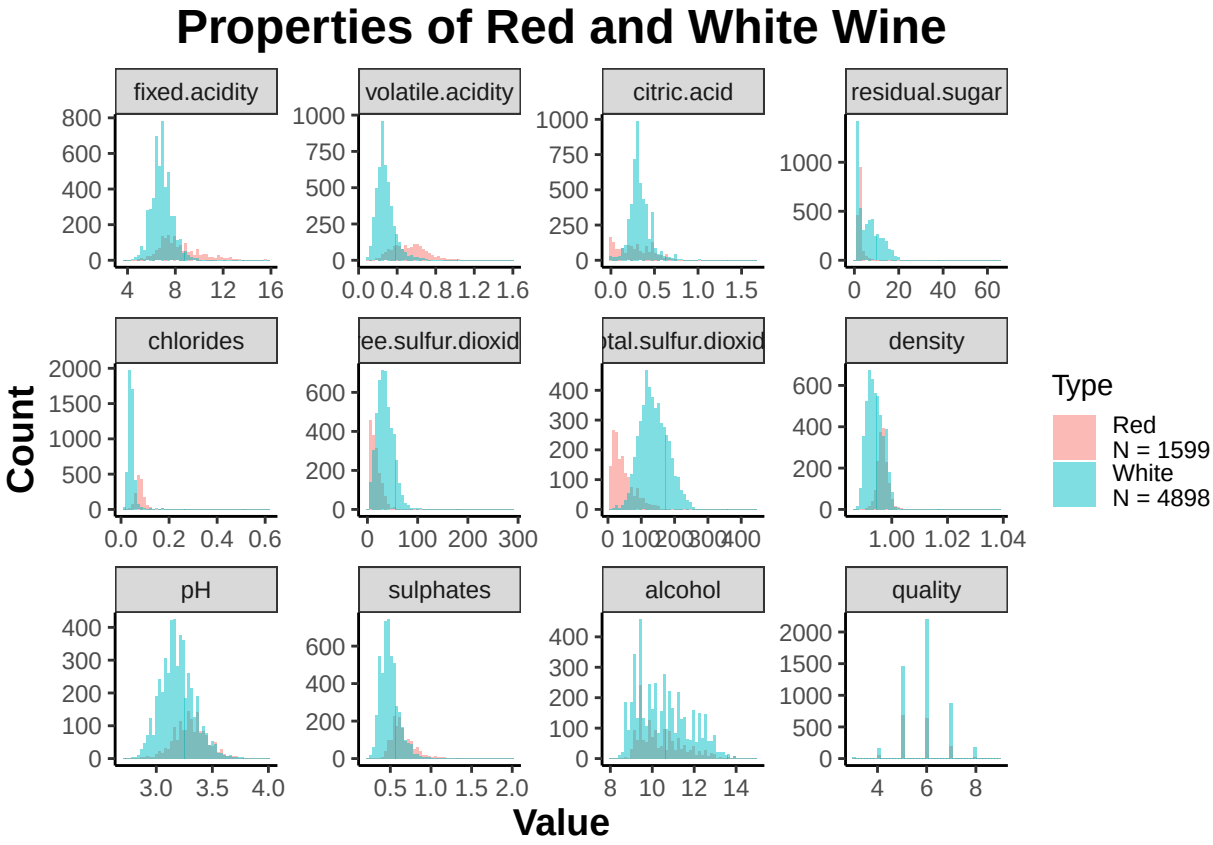
Nick Wawee

12/4/2020

Introduction

This document will explore two datasets that comprises physichemical properties and quality ratings of red and white wines.

What do the distirubtions look like of both red and white wines?



What are the letter values of red and white wine?

```
##
##
## Table: fixed.acidity
##
## |                  |  White|   Red|
## |:-----------------|------:|-----:|
## |Maximum           | 14.200| 15.90|
## |Upper Fourth      |  7.300|  9.20|
## |Median            |  6.800|  7.90|
## |Mean              |  6.855|  8.32|
## |Lower Fourth      |  6.300|  7.10|
## |Minimum           |  3.800|  4.60|
## |Outer Upper Fence |  9.300| 13.40|
## |Inner Upper Fence |  7.800| 10.25|
## |Inner Lower Fence |  5.800|  6.05|
## |Outer Lower Fence |  4.300|  2.90|
##
##
## Table: volatile.acidity
##
## |                  |  White|    Red|
## |:-----------------|------:|------:|
## |Maximum           |  1.100|  1.580|
## |Upper Fourth      |  0.320|  0.640|
## |Median            |  0.260|  0.520|
## |Mean              |  0.278|  0.528|
## |Lower Fourth      |  0.210|  0.390|
## |Minimum           |  0.080|  0.120|
## |Outer Upper Fence |  0.540|  1.140|
## |Inner Upper Fence |  0.375|  0.765|
## |Inner Lower Fence |  0.155|  0.265|
## |Outer Lower Fence | -0.010| -0.110|
##
##
## Table: citric.acid
##
## |                  | White|    Red|
## |:-----------------|-----:|------:|
## |Maximum           | 1.660|  1.000|
## |Upper Fourth      | 0.390|  0.420|
## |Median            | 0.320|  0.260|
## |Mean              | 0.334|  0.271|
## |Lower Fourth      | 0.270|  0.090|
## |Minimum           | 0.000|  0.000|
## |Outer Upper Fence | 0.630|  1.080|
## |Inner Upper Fence | 0.450|  0.585|
## |Inner Lower Fence | 0.210| -0.075|
## |Outer Lower Fence | 0.030| -0.570|
##
##
## Table: residual.sugar
##
```

```
## |                  |   White|    Red|
## |:-----------------|-------:|------:|
## |Maximum           |  65.800| 15.500|
## |Upper Fourth      |   9.900|  2.600|
## |Median            |   5.200|  2.200|
## |Mean              |   6.391|  2.539|
## |Lower Fourth      |   1.700|  1.900|
## |Minimum           |   0.600|  0.900|
## |Outer Upper Fence |  26.300|  4.000|
## |Inner Upper Fence |  14.000|  2.950|
## |Inner Lower Fence |  -2.400|  1.550|
## |Outer Lower Fence | -14.700|  0.500|
##
##
## Table: chlorides
##
## |                  | White|   Red|
## |:-----------------|-----:|-----:|
## |Maximum           | 0.346| 0.611|
## |Upper Fourth      | 0.050| 0.090|
## |Median            | 0.043| 0.079|
## |Mean              | 0.046| 0.087|
## |Lower Fourth      | 0.036| 0.070|
## |Minimum           | 0.009| 0.012|
## |Outer Upper Fence | 0.078| 0.130|
## |Inner Upper Fence | 0.057| 0.100|
## |Inner Lower Fence | 0.029| 0.060|
## |Outer Lower Fence | 0.008| 0.030|
##
##
## Table: free.sulfur.dioxide
##
## |                  |   White|     Red|
## |:-----------------|-------:|-------:|
## |Maximum           | 289.000|  72.000|
## |Upper Fourth      |  46.000|  21.000|
## |Median            |  34.000|  14.000|
## |Mean              |  35.308|  15.875|
## |Lower Fourth      |  23.000|   7.000|
## |Minimum           |   2.000|   1.000|
## |Outer Upper Fence |  92.000|  49.000|
## |Inner Upper Fence |  57.500|  28.000|
## |Inner Lower Fence |  11.500|   0.000|
## |Outer Lower Fence | -23.000| -21.000|
##
##
## Table: total.sulfur.dioxide
##
## |                  |   White|     Red|
## |:-----------------|-------:|-------:|
## |Maximum           | 440.000| 289.000|
## |Upper Fourth      | 167.000|  62.000|
## |Median            | 134.000|  38.000|
## |Mean              | 138.361|  46.468|
```

```
## |Lower Fourth      | 108.000|  22.000|
## |Minimum           |   9.000|   6.000|
## |Outer Upper Fence | 285.000| 142.000|
## |Inner Upper Fence | 196.500|  82.000|
## |Inner Lower Fence |  78.500|   2.000|
## |Outer Lower Fence | -10.000| -58.000|
##
##
## Table: density
##
## |                  | White|   Red|
## |:-----------------|-----:|-----:|
## |Maximum           | 1.039| 1.004|
## |Upper Fourth      | 0.996| 0.998|
## |Median            | 0.994| 0.997|
## |Mean              | 0.994| 0.997|
## |Lower Fourth      | 0.992| 0.996|
## |Minimum           | 0.987| 0.990|
## |Outer Upper Fence | 1.005| 1.002|
## |Inner Upper Fence | 0.998| 0.999|
## |Inner Lower Fence | 0.990| 0.994|
## |Outer Lower Fence | 0.983| 0.991|
##
##
## Table: pH
##
## |                  | White|   Red|
## |:-----------------|-----:|-----:|
## |Maximum           | 3.820| 4.010|
## |Upper Fourth      | 3.280| 3.400|
## |Median            | 3.180| 3.310|
## |Mean              | 3.188| 3.311|
## |Lower Fourth      | 3.090| 3.210|
## |Minimum           | 2.720| 2.740|
## |Outer Upper Fence | 3.660| 3.780|
## |Inner Upper Fence | 3.375| 3.495|
## |Inner Lower Fence | 2.995| 3.115|
## |Outer Lower Fence | 2.710| 2.830|
##
##
## Table: sulphates
##
## |                  | White|   Red|
## |:-----------------|-----:|-----:|
## |Maximum           |  1.08| 2.000|
## |Upper Fourth      |  0.55| 0.730|
## |Median            |  0.47| 0.620|
## |Mean              |  0.49| 0.658|
## |Lower Fourth      |  0.41| 0.550|
## |Minimum           |  0.22| 0.330|
## |Outer Upper Fence |  0.83| 1.090|
## |Inner Upper Fence |  0.62| 0.820|
## |Inner Lower Fence |  0.34| 0.460|
## |Outer Lower Fence |  0.13| 0.190|
```

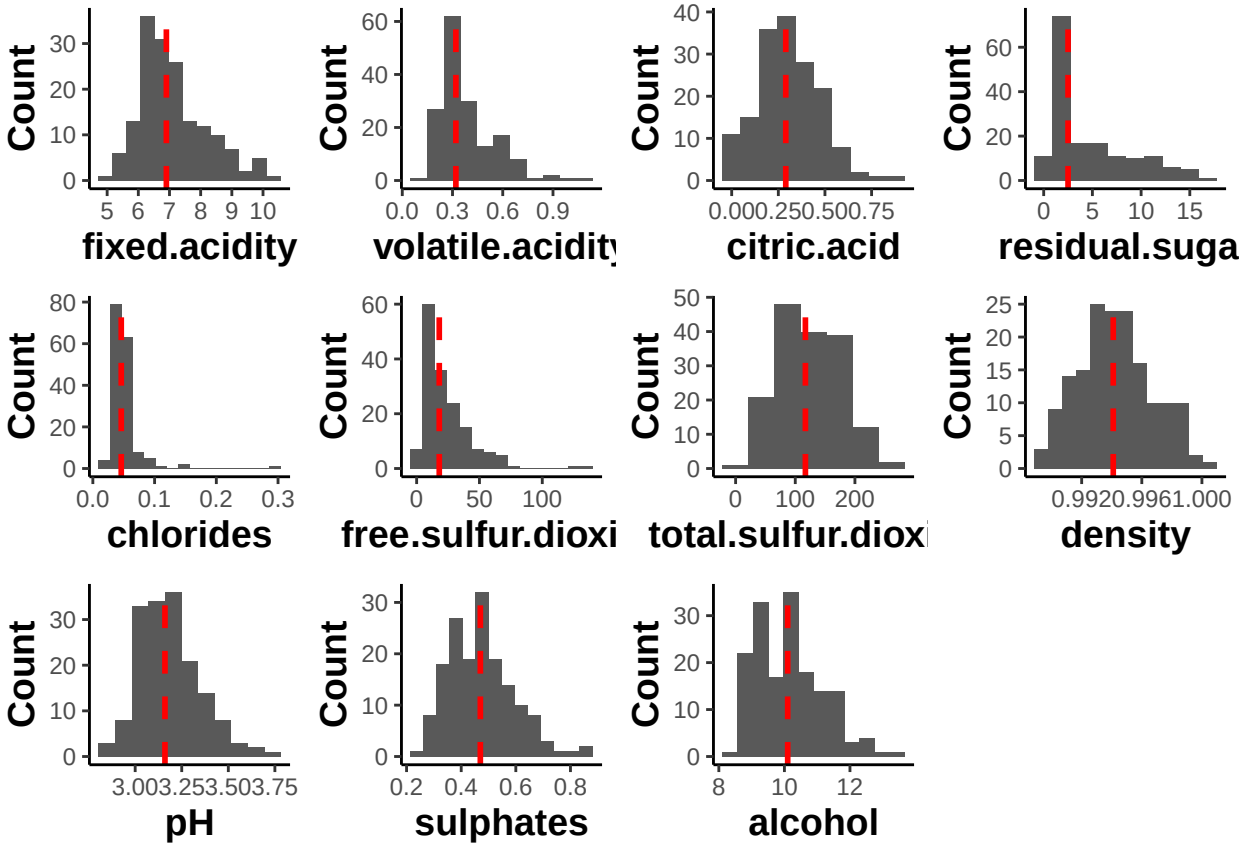
```
##
##
## Table: alcohol
##
## |                  |  White|    Red|
## |:-----------------|------:|------:|
## |Maximum           | 14.200| 14.900|
## |Upper Fourth      | 11.400| 11.100|
## |Median            | 10.400| 10.200|
## |Mean              | 10.514| 10.423|
## |Lower Fourth      |  9.500|  9.500|
## |Minimum           |  8.000|  8.400|
## |Outer Upper Fence | 15.200| 14.300|
## |Inner Upper Fence | 12.350| 11.900|
## |Inner Lower Fence |  8.550|  8.700|
## |Outer Lower Fence |  5.700|  6.300|
##
##
## Table: quality
##
## |                  | White|   Red|
## |:-----------------|-----:|-----:|
## |Maximum           | 9.000| 8.000|
## |Upper Fourth      | 6.000| 6.000|
## |Median            | 6.000| 6.000|
## |Mean              | 5.878| 5.636|
## |Lower Fourth      | 5.000| 5.000|
## |Minimum           | 3.000| 3.000|
## |Outer Upper Fence | 8.000| 8.000|
## |Inner Upper Fence | 6.500| 6.500|
## |Inner Lower Fence | 4.500| 4.500|
## |Outer Lower Fence | 3.000| 3.000|
```

What are the characteristics of quality outliers?

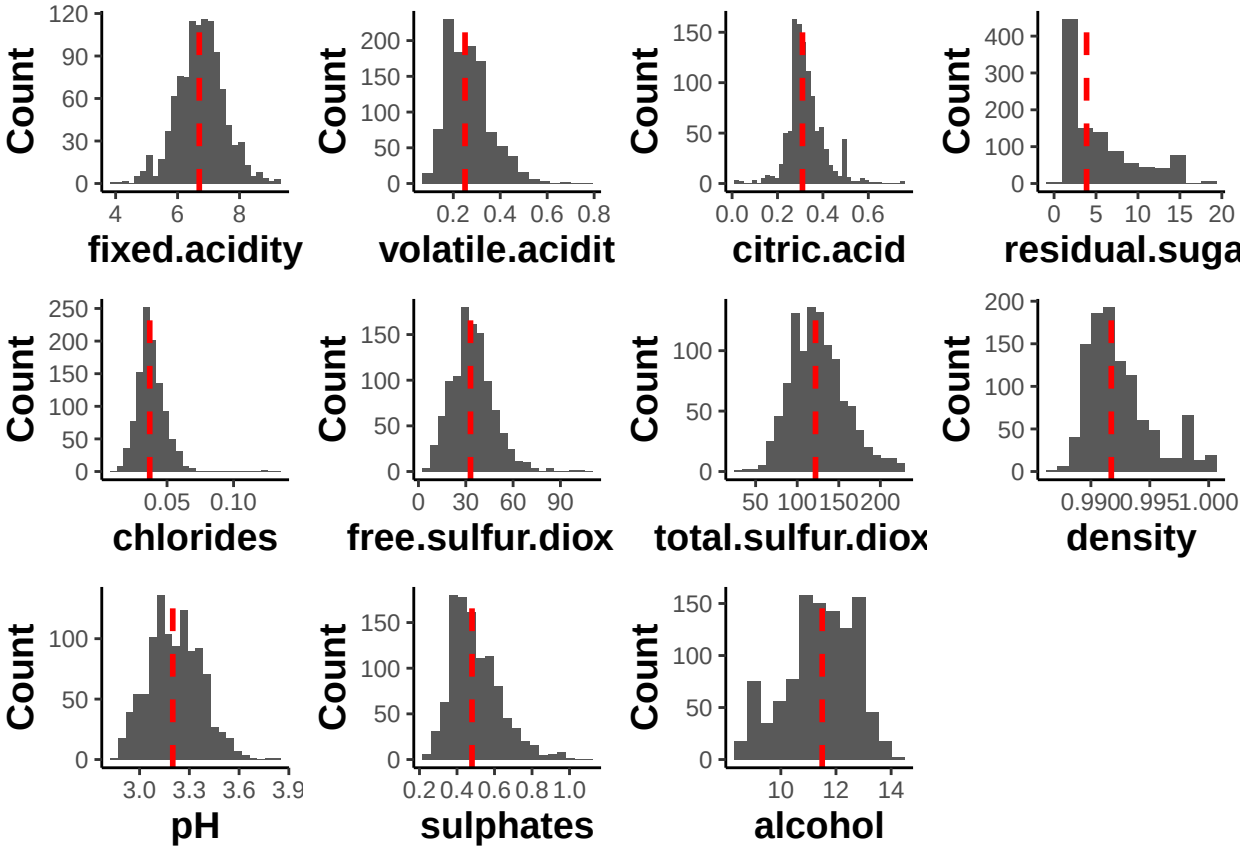
White Wine

There are 1043 mild outliers and 5 extreme outliers. Below will plot the characteristics of the mild outliers.

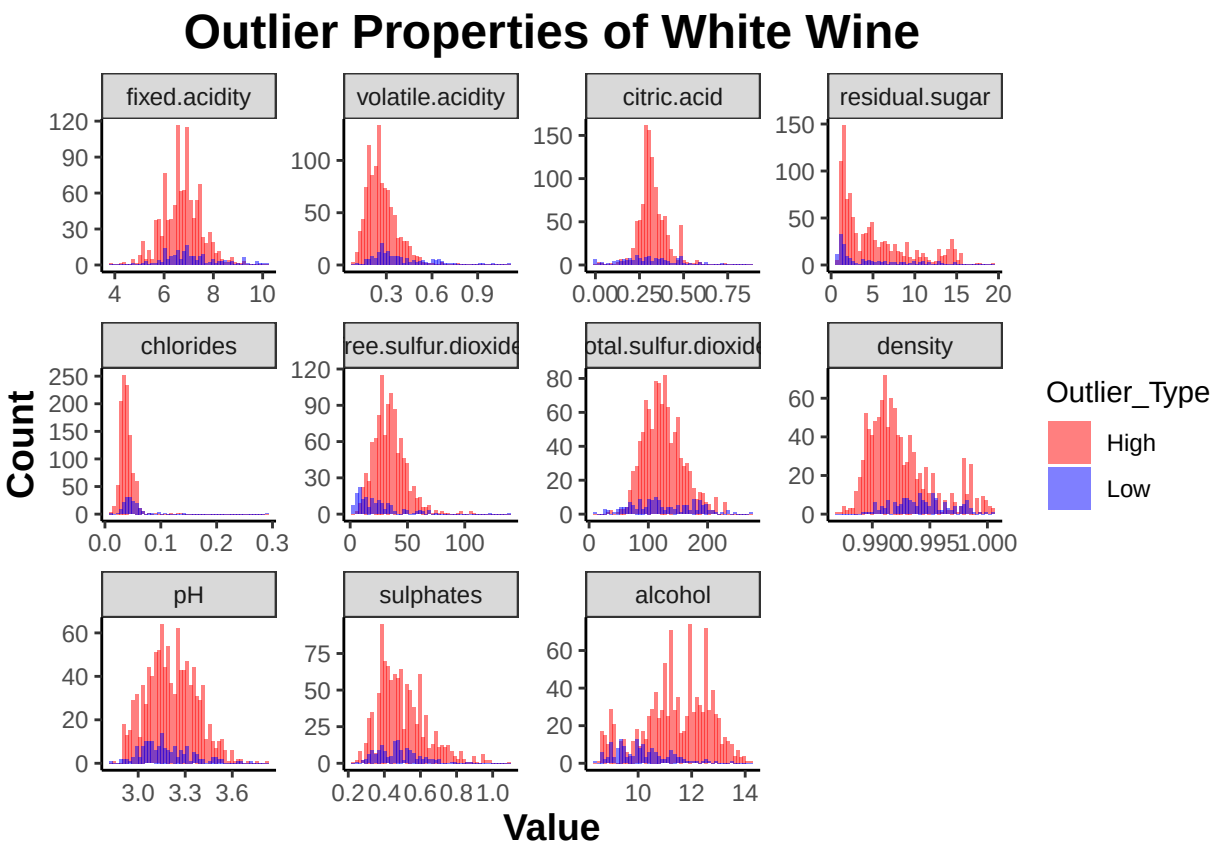
Low Quality


Count
fixed.acidity
volatile.acidity
citric.acid
residual.suga
chlorides
free.sulfur.dioxi
total.sulfur.dioxi
density
pH
sulphates
alcohol

High Quality

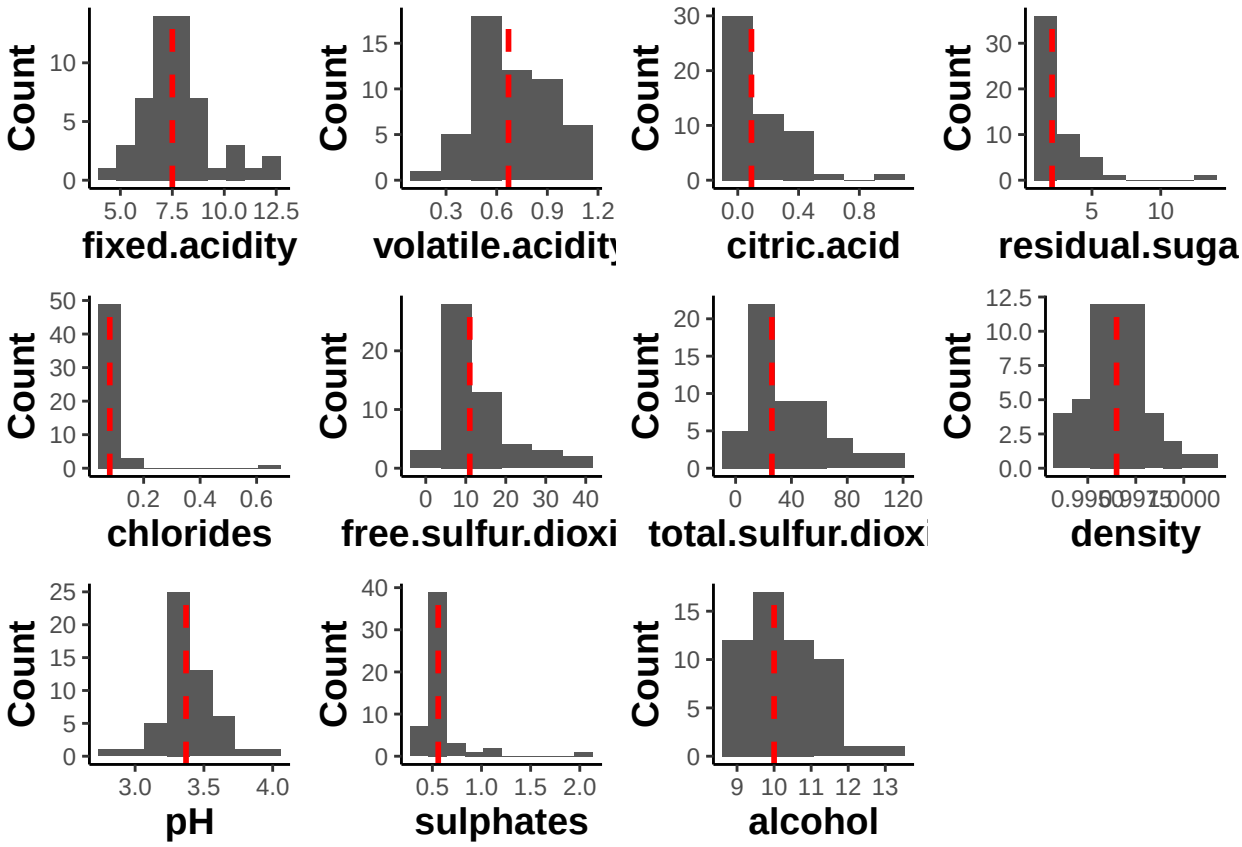


Both

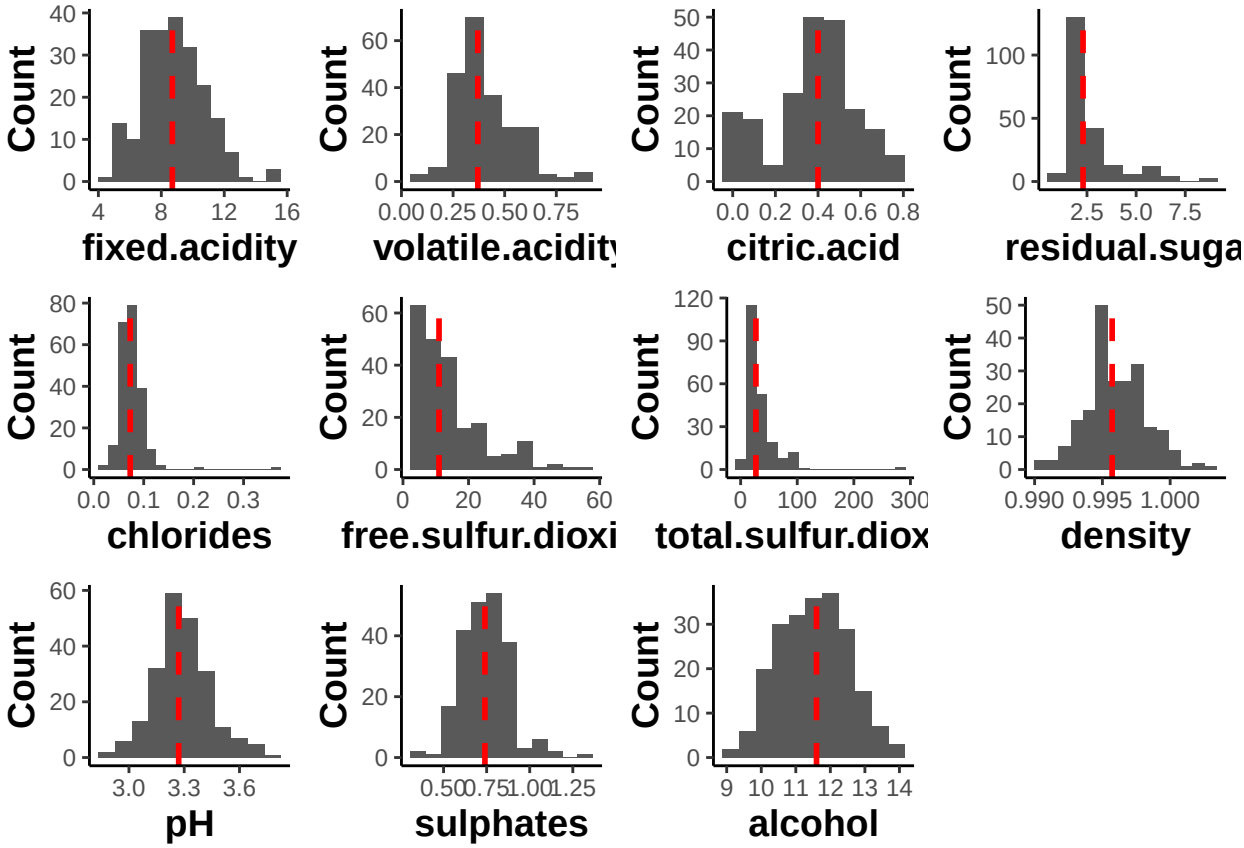

Outlier Properties of White Wine
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
ree.sulfur.dioxide
otal.sulfur.dioxid
density
pH
sulphates
alcohol
Count
Value
Outlier_Type
High
Low

Red Wine

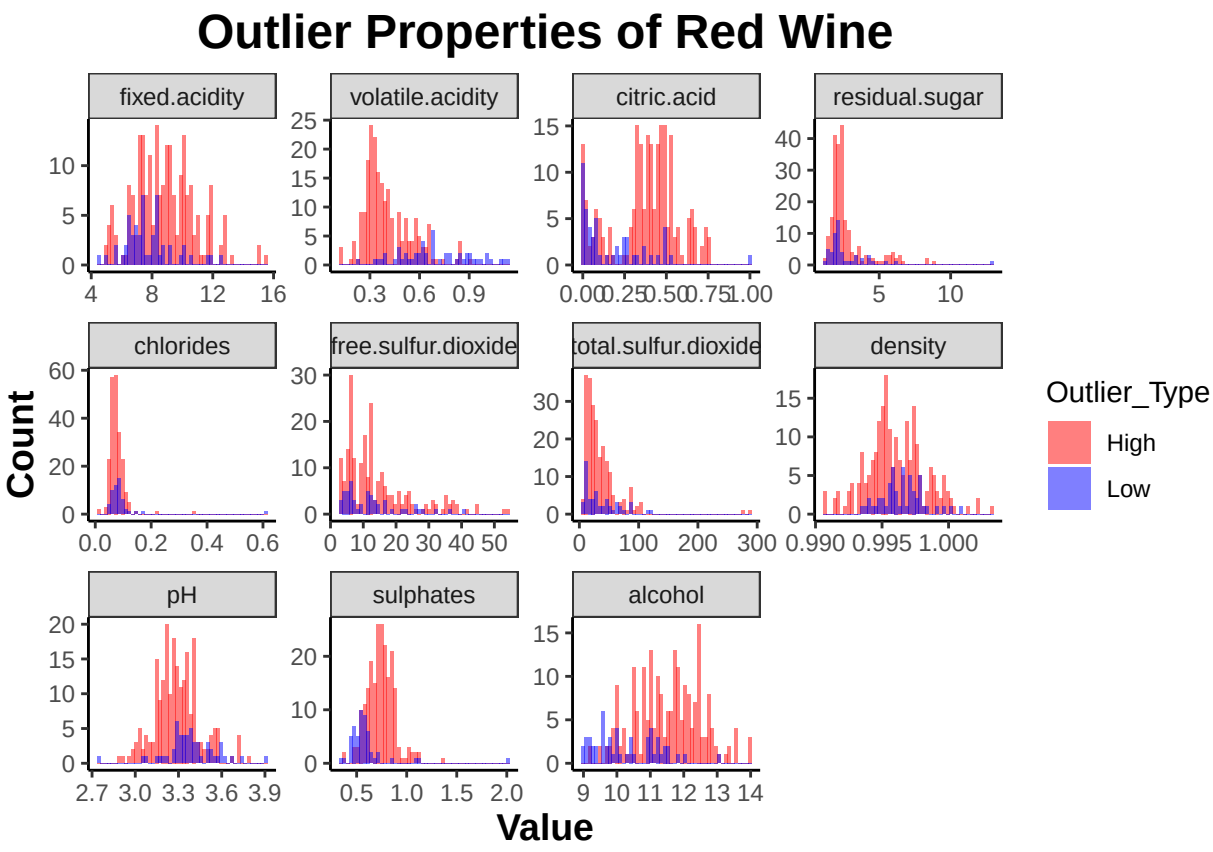
Low Quality



High Quality

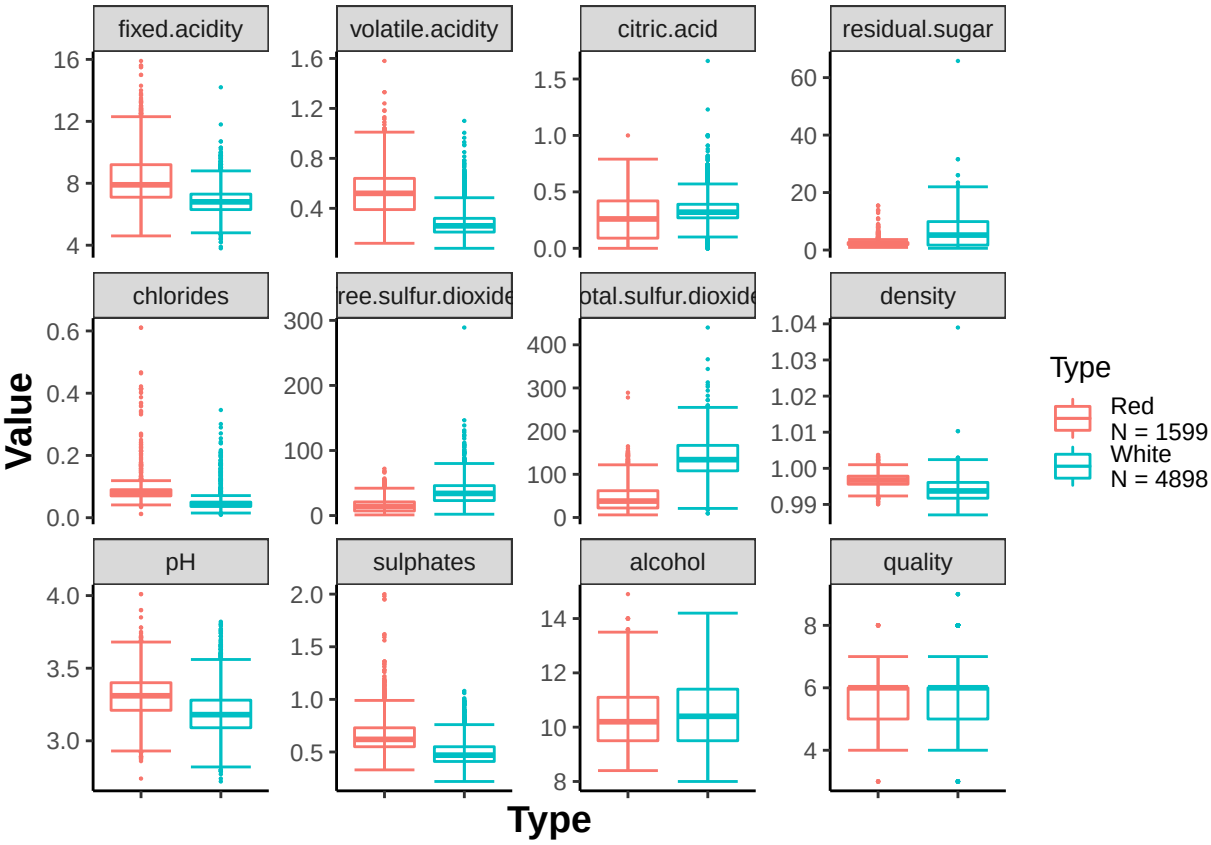


Both


Outlier Properties of Red Wine
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
Count
Value
Outlier_Type
High
Low

What does the symmetry of each variable look like?

Before Transformation



Transforming by Optimizing Hinkley Value

Table 1: Red Wine Results

	min_Hinkley	Power
fixed.acidity	0.0002313	-1.65
volatile.acidity	-0.0000051	0.74
citric.acid	-0.0009887	0.87
residual.sugar	-0.0005884	-1.20
chlorides	0.0004822	-0.68
free.sulfur.dioxide	0.0008651	0.41
total.sulfur.dioxide	0.0002937	0.15
density	-0.0006838	2.00
pH	0.0000015	0.69
sulphates	-0.0000682	-1.43
alcohol	-0.0469389	-2.00
quality	-0.3257718	2.00

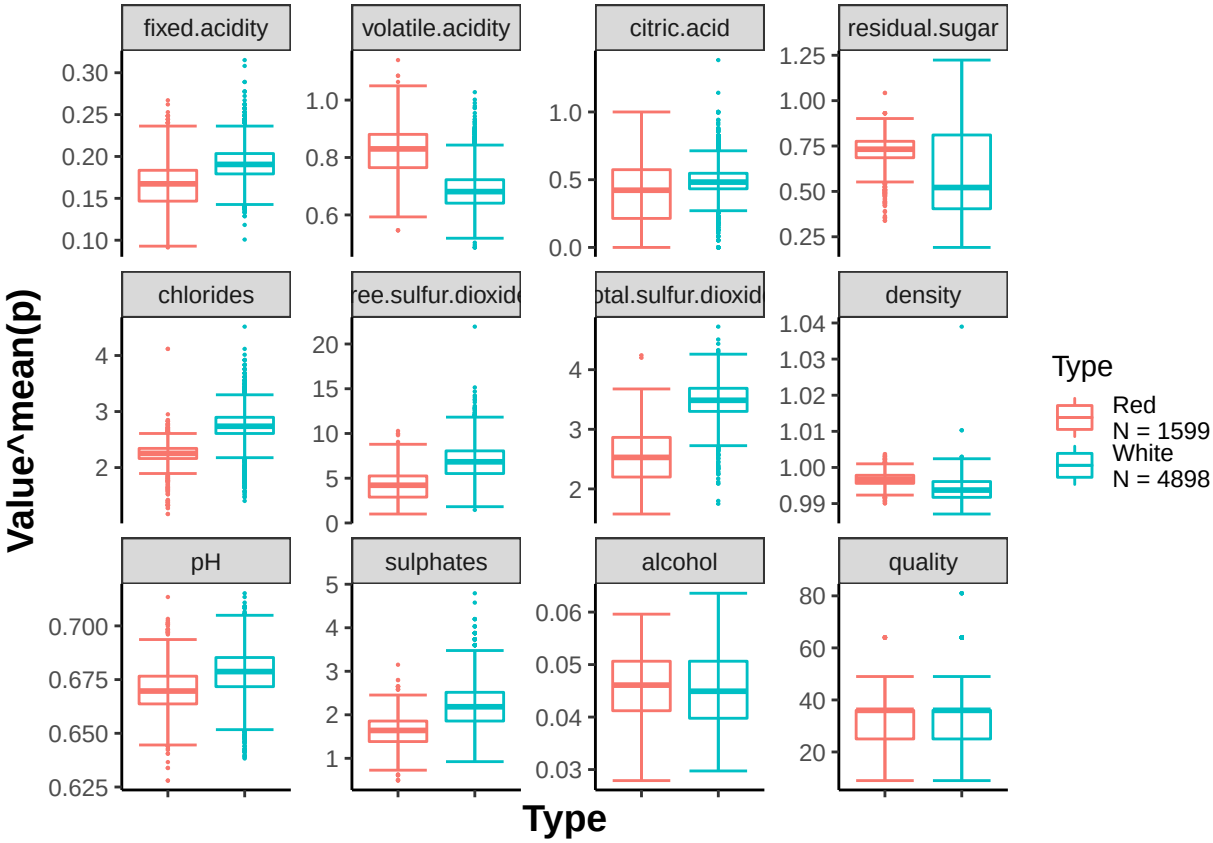
Table 2: White Wine Results

	min_Hinkley	Power
fixed.acidity	-0.0001077	-0.08
volatile.acidity	0.0004603	-0.17
citric.acid	0.0011695	0.41
residual.sugar	0.0001969	0.41
chlorides	-0.0007633	0.04
free.sulfur.dioxide	0.0002793	0.68
total.sulfur.dioxide	0.0000249	0.36
density	-0.0625470	-2.00
pH	0.0000900	-1.36
sulphates	-0.0000883	-0.64
alcohol	0.0001319	-0.65
quality	-0.0605442	2.00

Table 3: Combined Results

	Mean_Power
fixed.acidity	-0.865
volatile.acidity	0.285
citric.acid	0.640
residual.sugar	-0.395
chlorides	-0.320
free.sulfur.dioxide	0.545
total.sulfur.dioxide	0.255
density	0.000
pH	-0.335
sulphates	-1.035
alcohol	-1.325
quality	2.000

Plotting Transformed Variables



Which variables are statistically different between the untransformed data?

Below will employ a two sample t test to test if the populations are equal for each variable. We assume unequal variances despite the power transformation as the IQR spread is not equal in all cases. We will also run a Mood's Median test to test for differences in median rather than mean.

Table 4: t-test Results

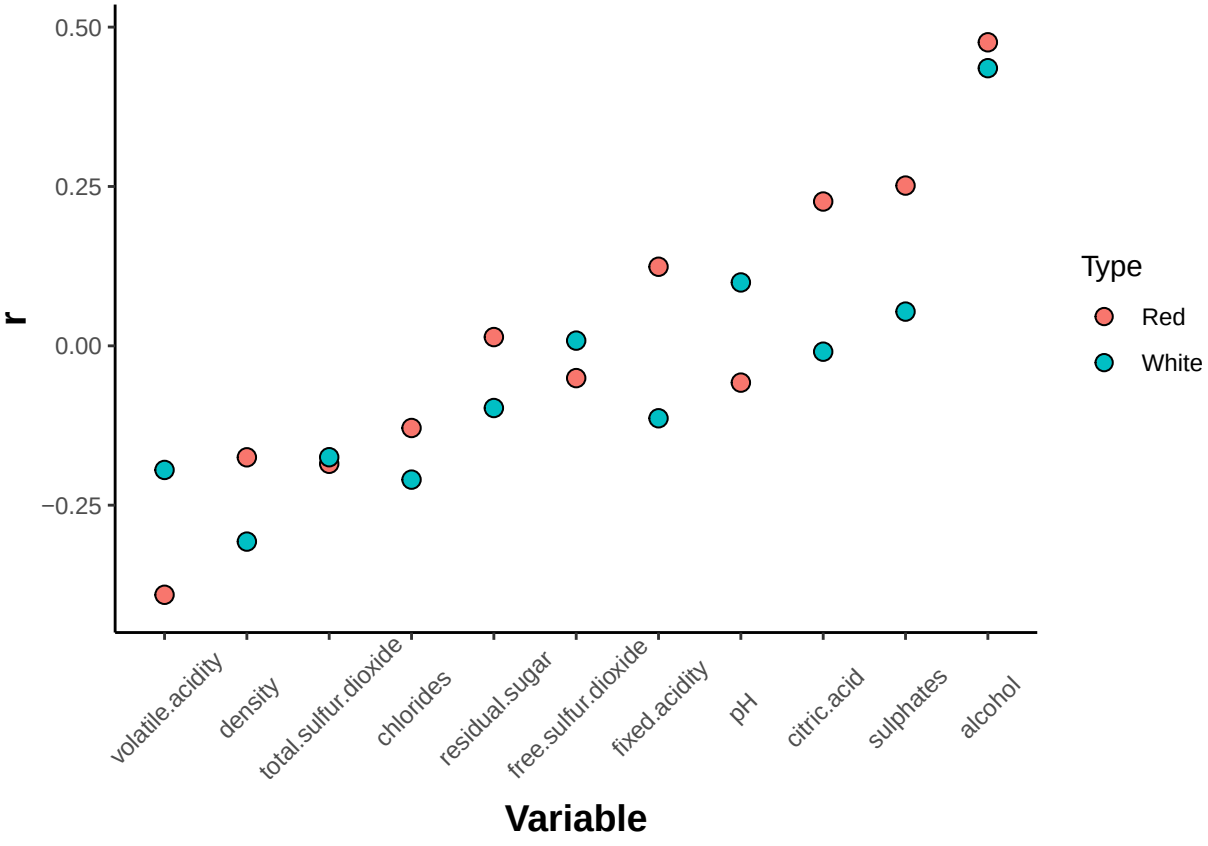
	t	p
volatile.acidity	-53.059	0.000
density	-42.709	0.000
sulphates	-37.056	0.000
chlorides	-34.240	0.000
fixed.acidity	-32.423	0.000
pH	-27.775	0.000
alcohol	2.859	0.004
quality	10.149	0.000
citric.acid	12.229	0.000
residual.sugar	47.802	0.000
free.sulfur.dioxide	54.428	0.000
total.sulfur.dioxide	89.872	0.000

Table 5: Mood's Median Results

	Z	p
chlorides	-42.494	0.000
total.sulfur.dioxide	-41.700	0.000
sulphates	-35.480	0.000
free.sulfur.dioxide	-33.739	0.000
fixed.acidity	-27.227	0.000
pH	-23.440	0.000
quality	-7.051	0.000
citric.acid	-6.897	0.000
alcohol	3.462	0.001
residual.sugar	32.532	0.000
density	33.500	0.000
volatile.acidity	39.328	0.000

Which factors impact the quality score the most?

Examining Correlations with Quality



MLR

Modeling

```
##
## Call:
## lm(formula = quality ~ 0 + fixed.acidity + volatile.acidity +
```

```
##     citric.acid + residual.sugar + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + density + pH + sulphates + alcohol,
##     data = whitedf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9144 -0.4958 -0.0333  0.4675  3.1762
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## fixed.acidity        -0.0505906  0.0150754  -3.356 0.000797 ***
## volatile.acidity     -1.9585102  0.1138903 -17.196  < 2e-16 ***
## citric.acid          -0.0293492  0.0961648  -0.305 0.760229
## residual.sugar        0.0249884  0.0025917   9.642  < 2e-16 ***
## chlorides            -0.9425824  0.5430204  -1.736 0.082660 .
## free.sulfur.dioxide   0.0047908  0.0008390   5.710 1.20e-08 ***
## total.sulfur.dioxide -0.0008776  0.0003731  -2.352 0.018699 *
## density               2.0420461  0.3532997   5.780 7.94e-09 ***
## pH                    0.1683951  0.0835957   2.014 0.044022 *
## sulphates             0.4164536  0.0973279   4.279 1.91e-05 ***
## alcohol               0.3656334  0.0111203  32.880  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7562 on 4887 degrees of freedom
## Multiple R-squared:  0.9839, Adjusted R-squared:  0.9838
## F-statistic: 2.707e+04 on 11 and 4887 DF,  p-value: < 2.2e-16
```

Table 6: White Wine Regression Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
volatile.acidity	-1.959	0.114	-17.196	0.000
chlorides	-0.943	0.543	-1.736	0.083
fixed.acidity	-0.051	0.015	-3.356	0.001
citric.acid	-0.029	0.096	-0.305	0.760
total.sulfur.dioxide	-0.001	0.000	-2.352	0.019
free.sulfur.dioxide	0.005	0.001	5.710	0.000
residual.sugar	0.025	0.003	9.642	0.000
pH	0.168	0.084	2.014	0.044
alcohol	0.366	0.011	32.880	0.000
sulphates	0.416	0.097	4.279	0.000
density	2.042	0.353	5.780	0.000

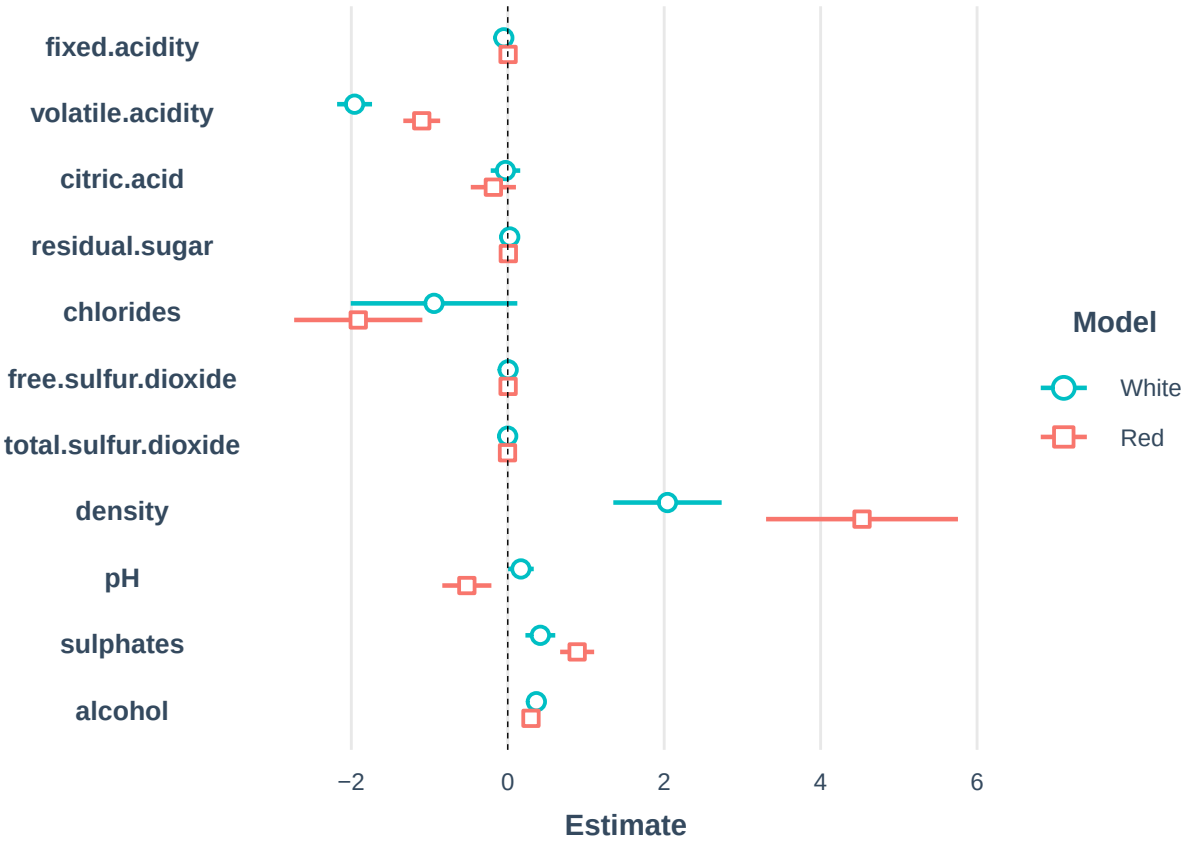
```
##
## Call:
## lm(formula = quality ~ 0 + fixed.acidity + volatile.acidity +
##     citric.acid + residual.sugar + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + density + pH + sulphates + alcohol,
##     data = reddf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.66872 -0.36621 -0.04653  0.45604  2.04187
```

```
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## fixed.acidity          0.0041937  0.0164513   0.255  0.79882
## volatile.acidity      -1.0997431  0.1200969  -9.157  < 2e-16 ***
## citric.acid           -0.1841460  0.1471717  -1.251  0.21103
## residual.sugar         0.0070712  0.0120512   0.587  0.55745
## chlorides             -1.9114188  0.4177542  -4.575 5.12e-06 ***
## free.sulfur.dioxide    0.0045478  0.0021639   2.102  0.03574 *
## total.sulfur.dioxide -0.0033186  0.0007269  -4.565 5.37e-06 ***
## density                4.5291462  0.6253297   7.243 6.82e-13 ***
## pH                    -0.5228983  0.1599968  -3.268  0.00111 **
## sulphates              0.8870761  0.1107998   8.006 2.27e-15 ***
## alcohol                0.2970228  0.0172513  17.217  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1588 degrees of freedom
## Multiple R-squared:  0.9871, Adjusted R-squared:  0.987
## F-statistic: 1.108e+04 on 11 and 1588 DF,  p-value: < 2.2e-16
```

Table 7: Red Wine Regression Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
chlorides	-1.911	0.418	-4.575	0.000
volatile.acidity	-1.100	0.120	-9.157	0.000
pH	-0.523	0.160	-3.268	0.001
citric.acid	-0.184	0.147	-1.251	0.211
total.sulfur.dioxide	-0.003	0.001	-4.565	0.000
fixed.acidity	0.004	0.016	0.255	0.799
free.sulfur.dioxide	0.005	0.002	2.102	0.036
residual.sugar	0.007	0.012	0.587	0.557
alcohol	0.297	0.017	17.217	0.000
sulphates	0.887	0.111	8.006	0.000
density	4.529	0.625	7.243	0.000

Plotting Coefficients


fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
−2
0
2
4
6
Estimate
Model
White
Red