Lab 3

Krista Bogan & Jessica Booth

2022-10-13

Part 1

Part 2

Business Understanding

Data Preparation

- As a part of data preparation, we downloaded the red wine and white wine csv files and opened them in excel to view them. We noticed that the csv files were separated by semicolons rather than split into columns.

```
##   fixed.acidity.volatile.acidity.citric.acid.residual.sugar.chlorides.free.sulfur.dioxide.total.sulfu
## 1                                                                                                7
## 2                                                                                                7
## 3                                                                                             7.8;
## 4                                                                                            11.2;
## 5                                                                                                7
## 6                                                                                                7.
```

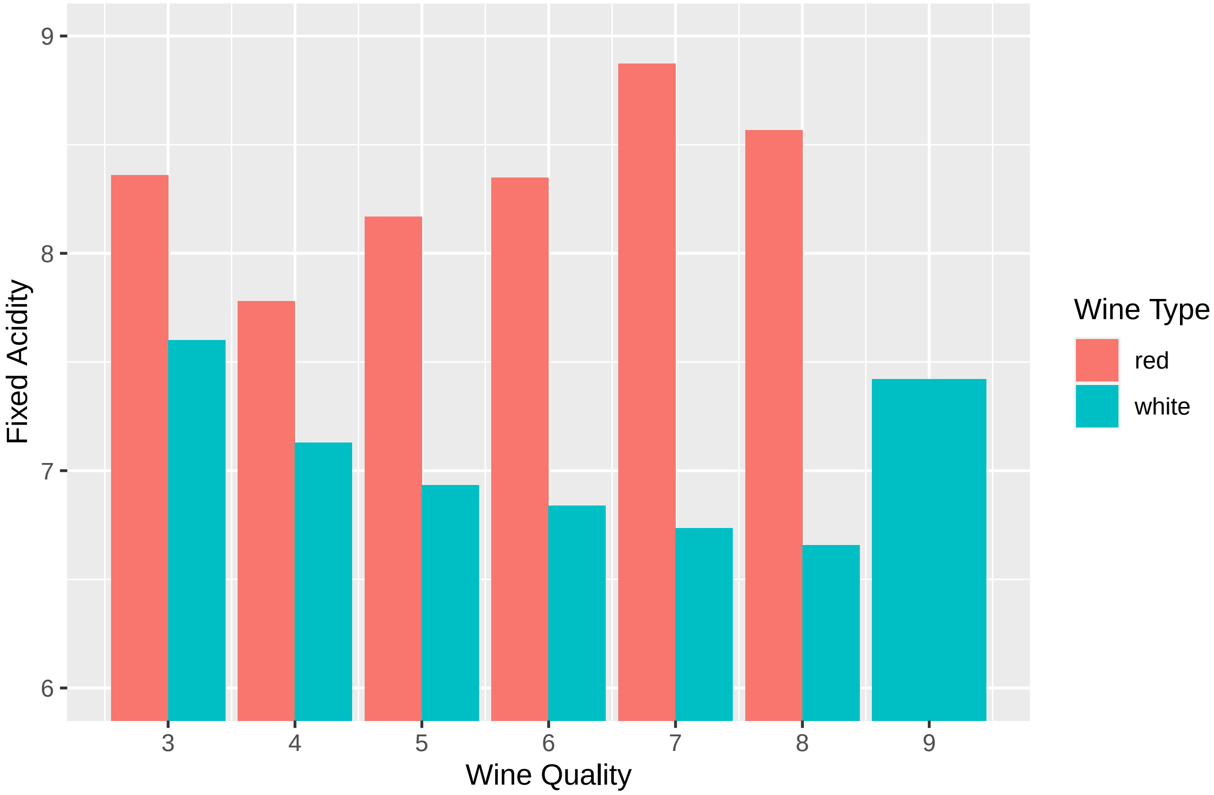
- Using Microsoft Excel, we separated the file into columns using the *Text to Column* function, separating the values by semicolon. Upon doing this for both the red and white wine datasets, we combined them, and added a column for "wine type". It is also important to note that we replaced any spaces in our column names with "_" so it is easy to use in our code.

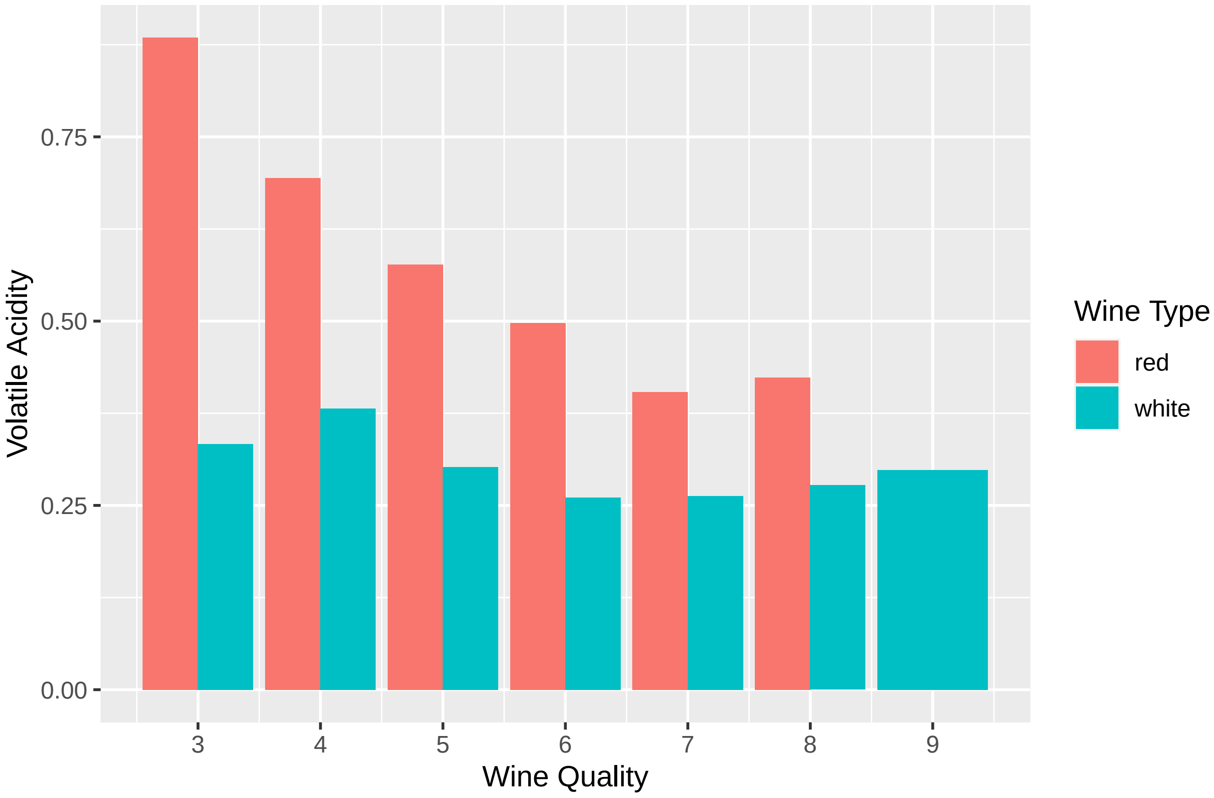
```
##   fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 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 wine_type
## 1       5       red
## 2       5       red
## 3       5       red
## 4       6       red
## 5       5       red
## 6       5       red
```

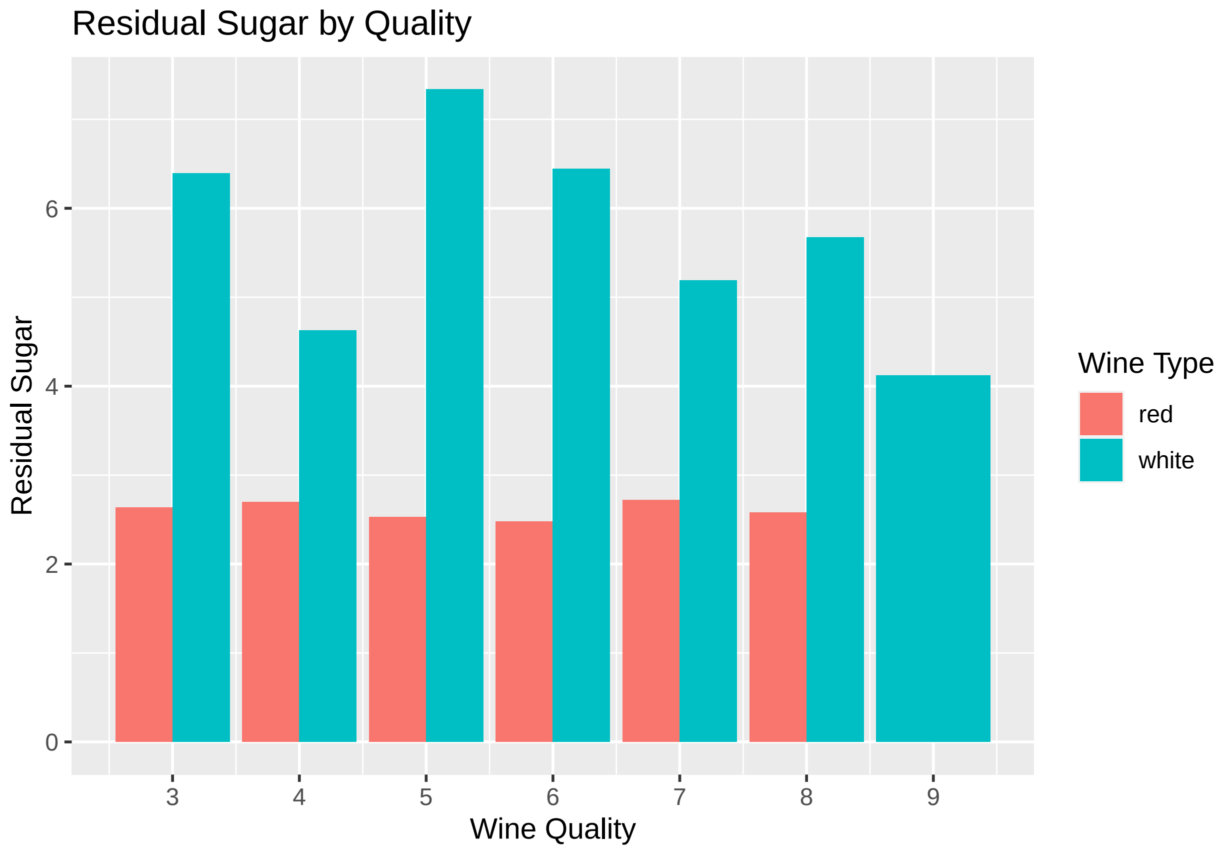
- Before creating any visualizations of the data, we wanted to run summary statistics to see which variables effectively impact the quality of wine. We see from these summary statistics that citric acid is not an effective predictor for wine quality.

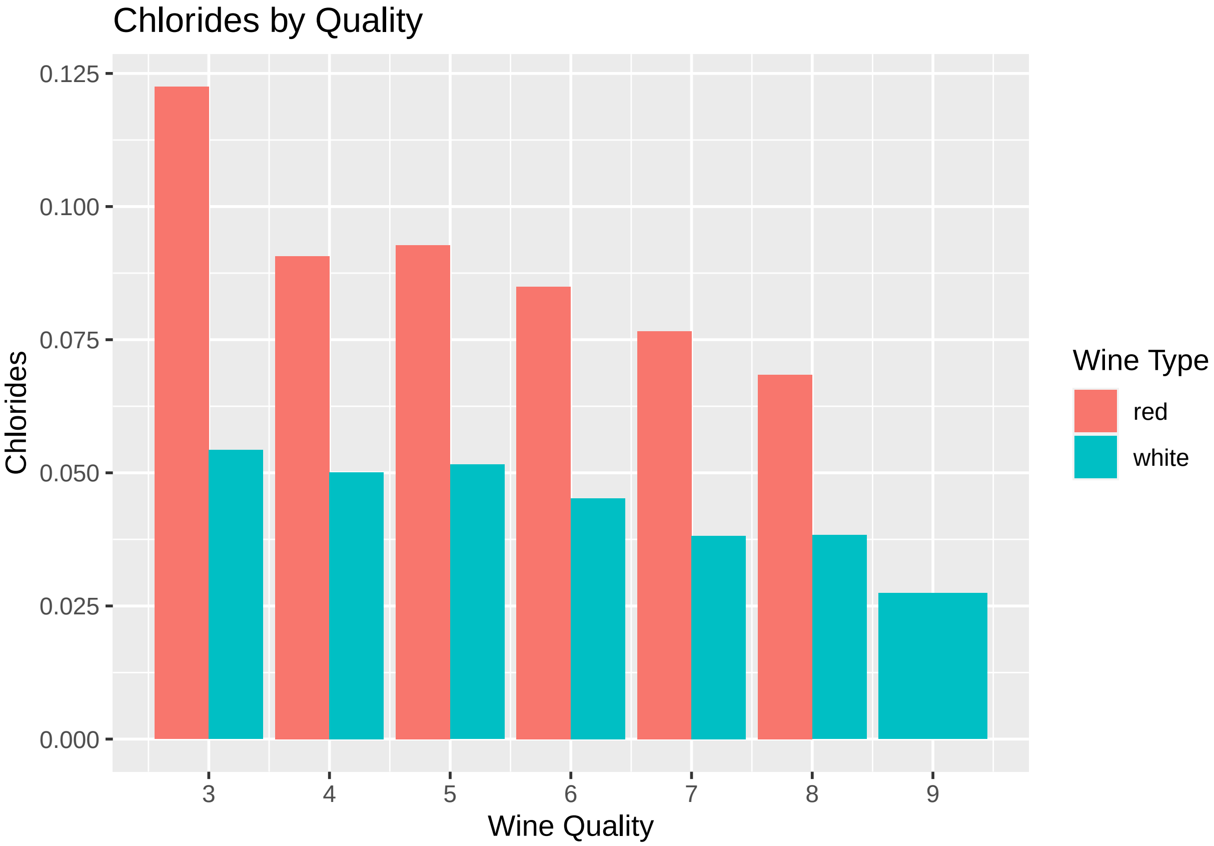
```
##
## Call:
## lm(formula = quality ~ fixed_acidity + volatile_acidity + citric_acid +
##     residual_sugar + chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
##     density + pH + sulphates + alcohol + wine_type, data = winequalitysplit)
##
## Residuals:
##     Min      1Q  Median      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 ***
## wine_typewhite       -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
```

- Now that we know which variables have a significant correlation value for our dependent variable, wine quality, we can start plotting.

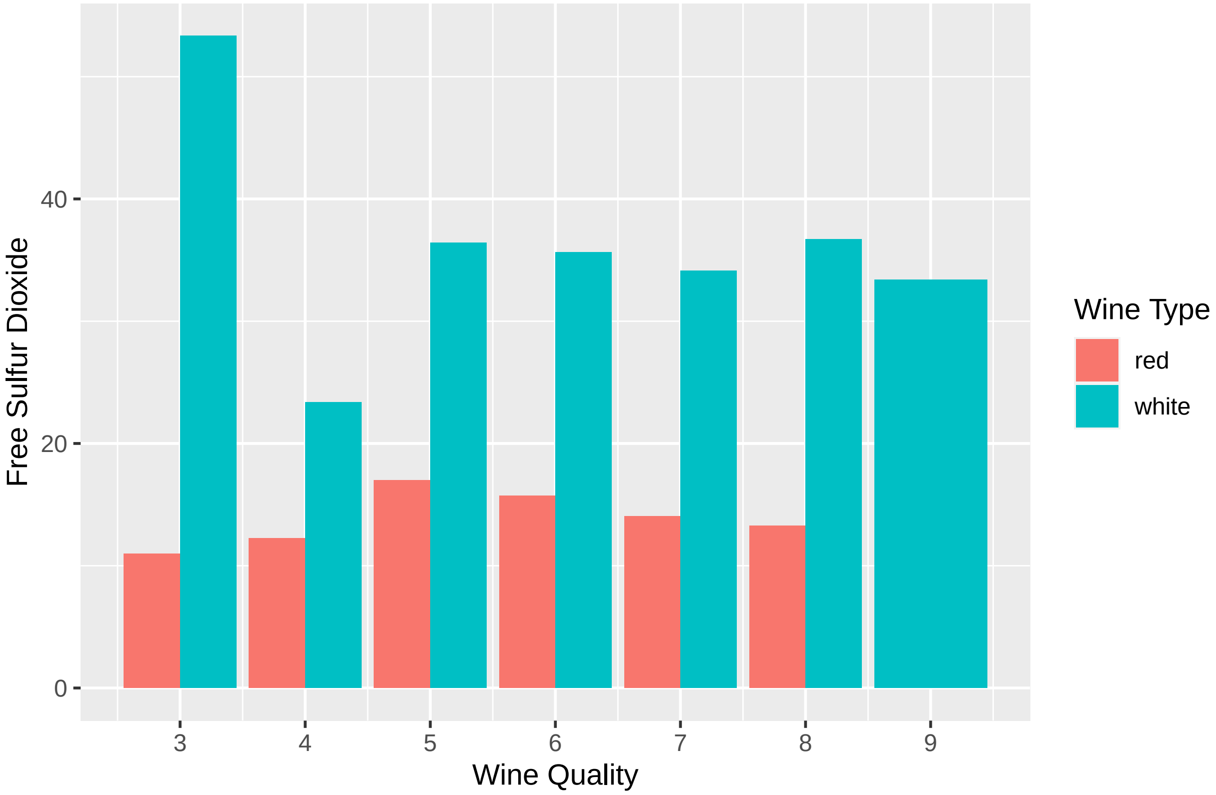

Fixed Acidity by Quality


Volatile Acidity by Quality

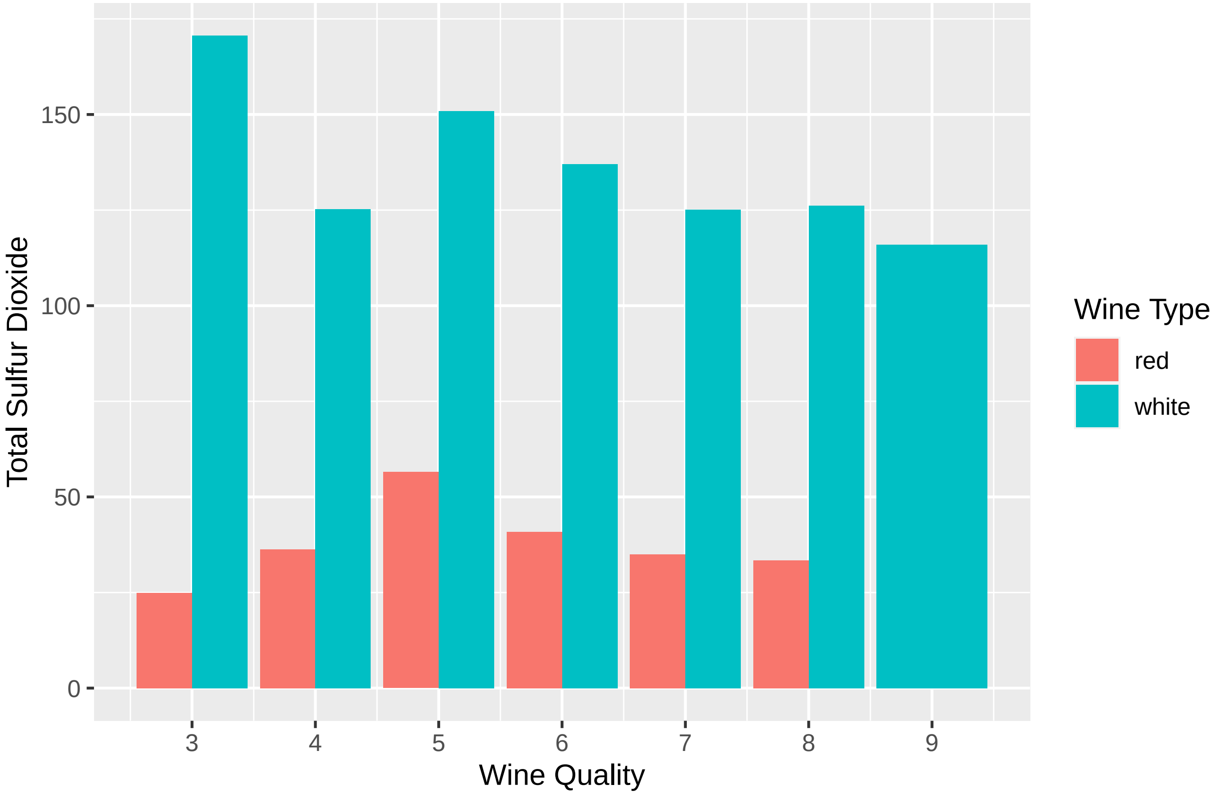

Residual Sugar by Quality
Residual Sugar
0
2
4
6
3
4
5
6
7
8
9
Wine Quality
Wine Type
red
white

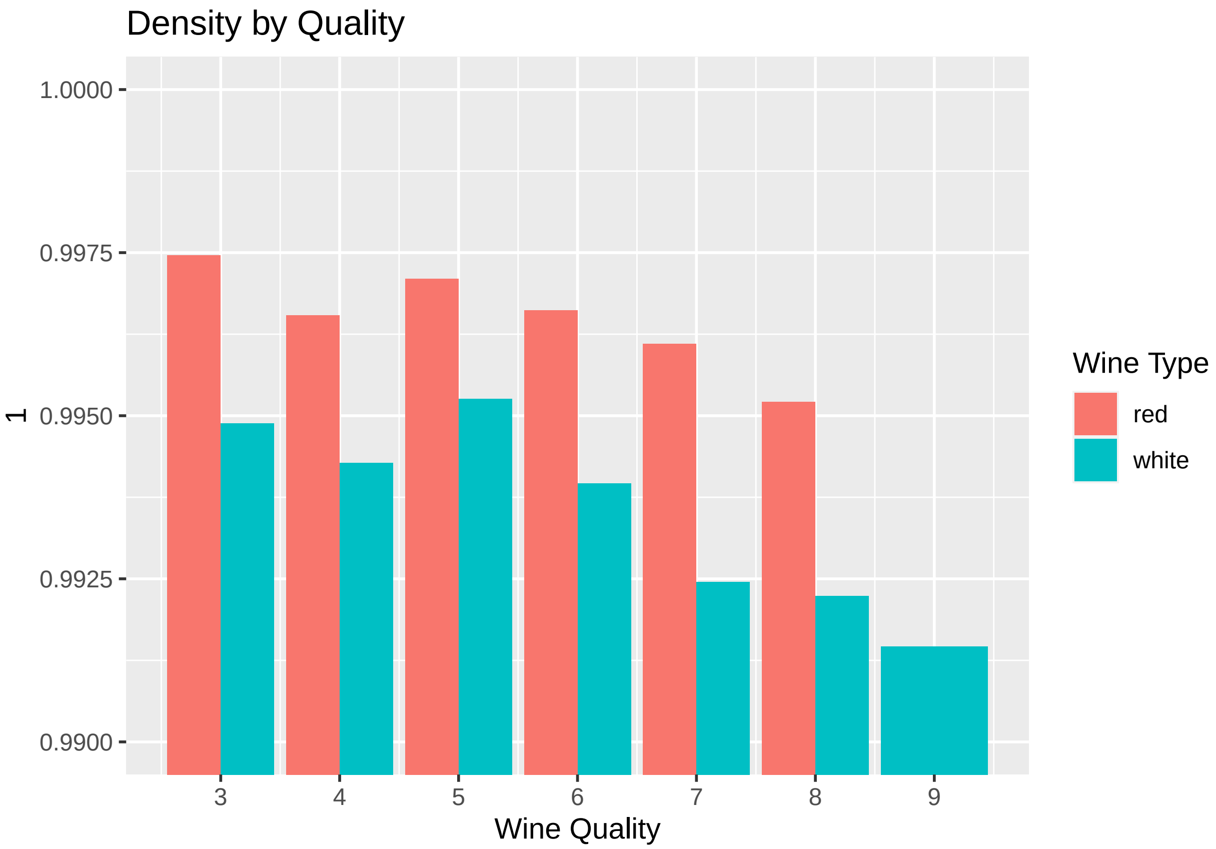

Chlorides by Quality
Chlorides
0.000
0.025
0.050
0.075
0.100
0.125
3
4
5
6
7
8
9
Wine Quality
Wine Type
red
white

Free Sulfur Dioxide by Quality

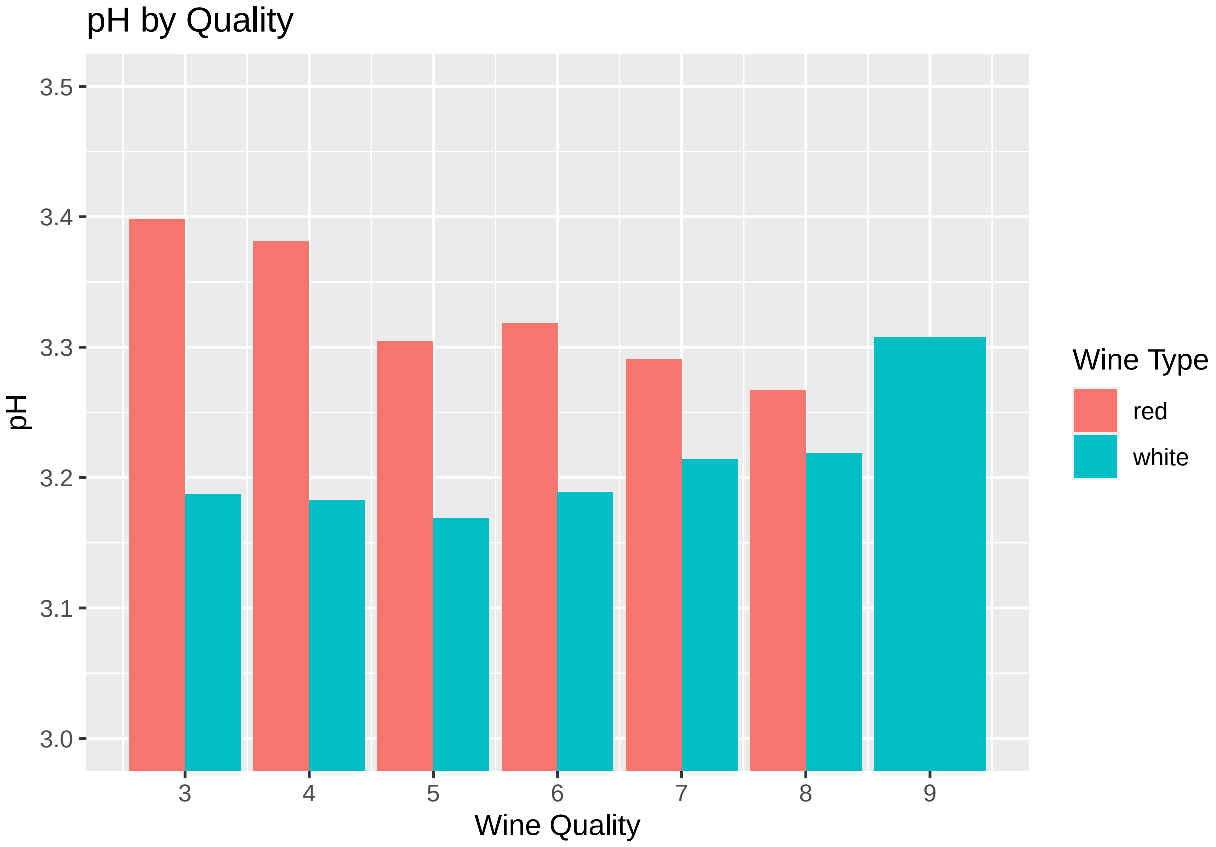


Total Sulfur Dioxide by Quality

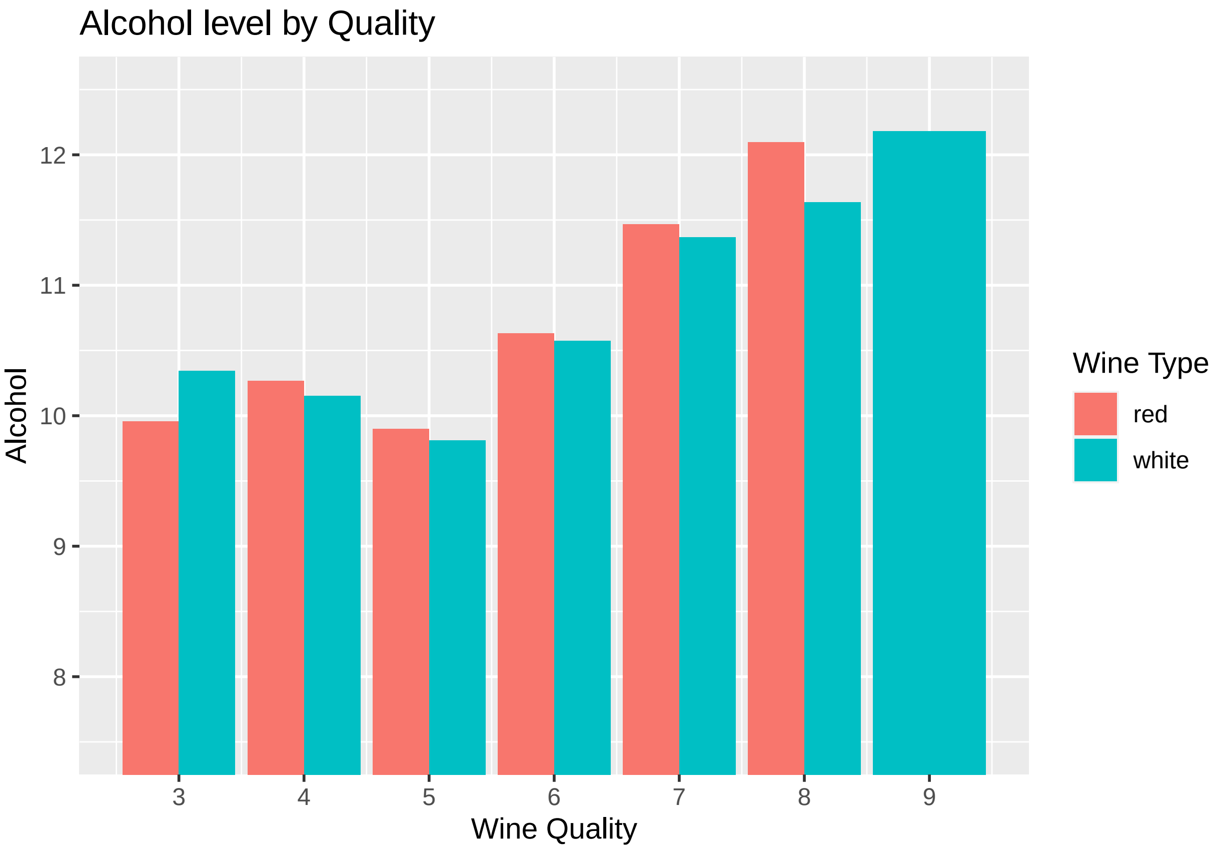




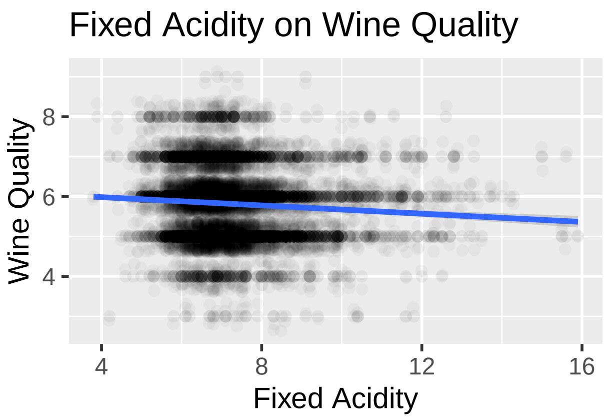
Density by Quality

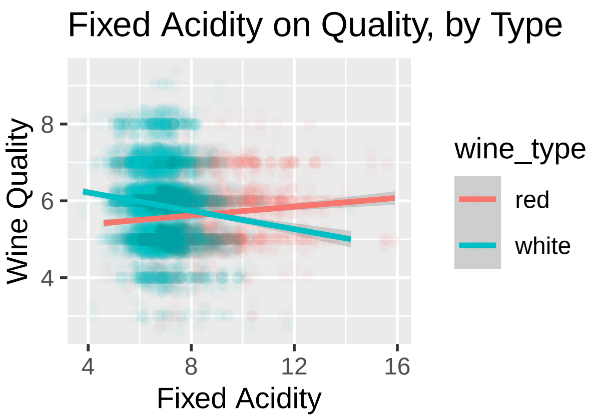


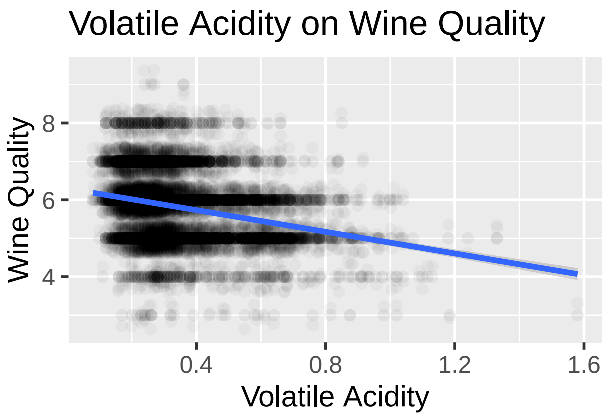
pH by Quality

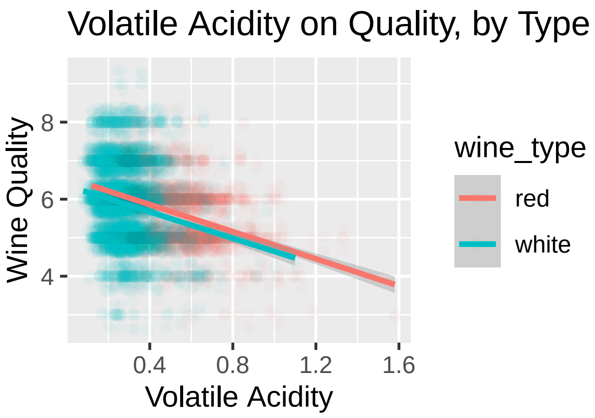


```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

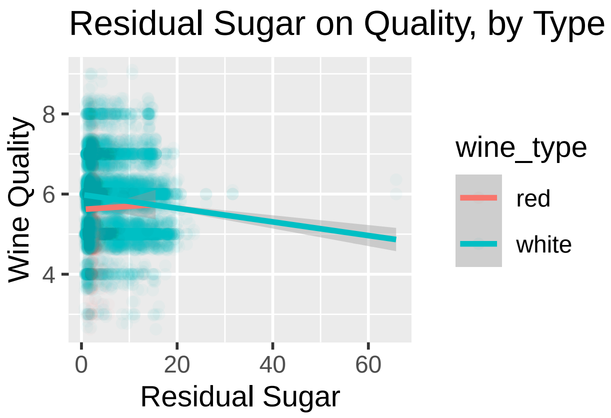

Fixed Acidity on Wine Quality

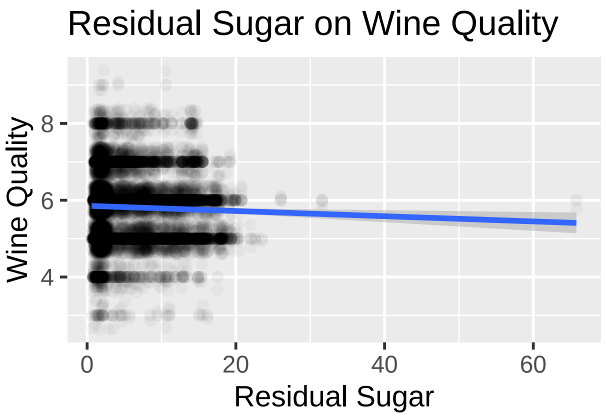

Fixed Acidity on Quality, by Type


Volatile Acidity on Wine Quality


Volatile Acidity on Quality, by Type

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```


Residual Sugar on Quality, by Type


Residual Sugar on Wine Quality

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.