Analysis of Wine Quality

Christine Lee

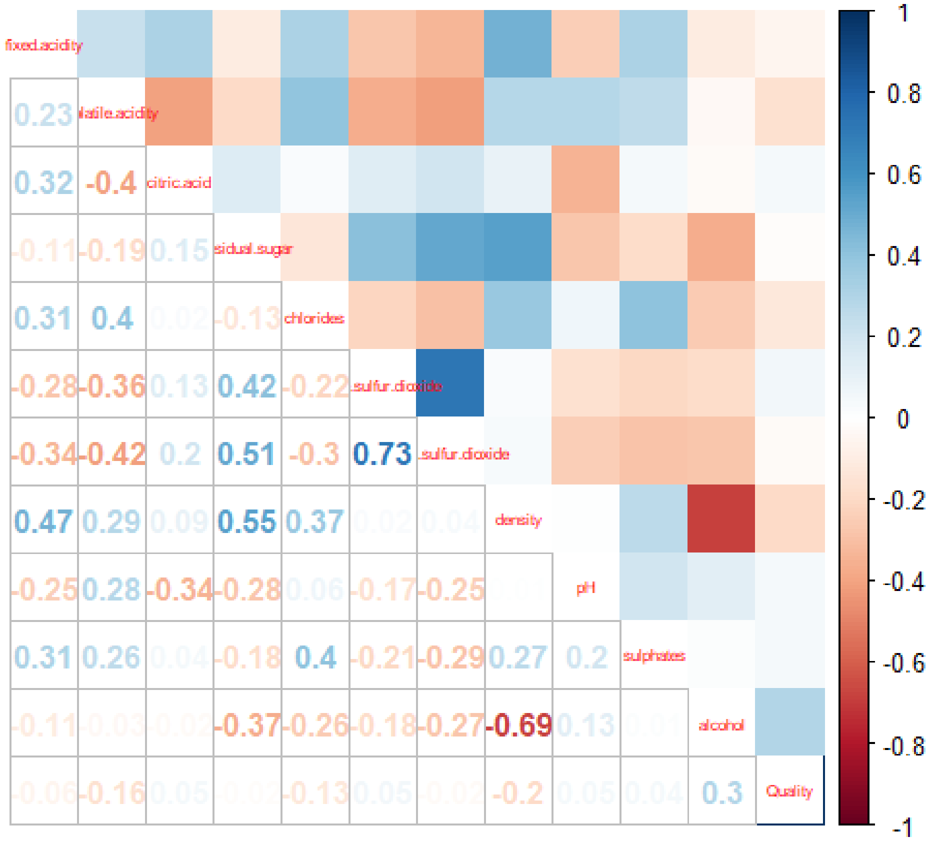
University of California, Los Angeles

Analysis of Wine Quality

Initial Approach

To start off, the Tidyverse package was used to read in the dataset. This dataset contains fourteen variables, twelve of which will be used as predictors for wine quality. There is one categorical predictor and eleven numerical predictors. The categorical predictor is the wine color and the numerical predictors are various levels of different factors of wine such as acidity, pH, alcohol, and more.

Graph 1: Correlation Matrix Plot



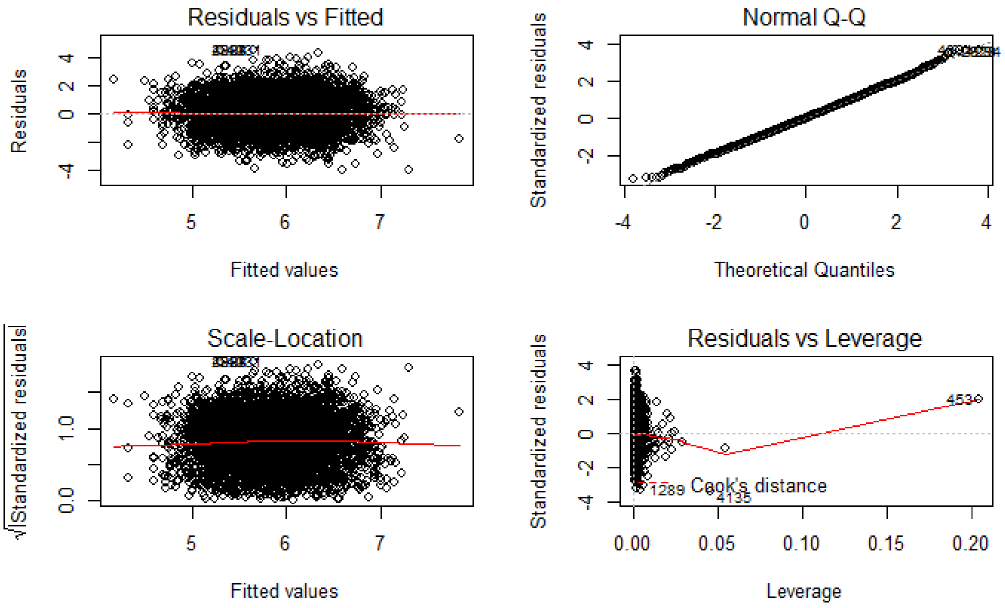
The first step in analysis was to look at correlation plots between the predictors. From the correlation plots we can see that *free.sulfur.dioxide* and *total.sulfur.dioxide* are strongly correlated. *Density* and *sulfates* are also strongly correlated. It is also observed that *Quality* and *alcohol* are strongly associated.

Output 1: Full Model Summary

```
Call:
lm(formula = Quality ~ Wine.Color + fixed.acidity + volatile.acidity + 
    citric.acid + residual.sugar + chlorides + free.sulfur.dioxide + 
    total.sulfur.dioxide + density + pH + sulphates + alcohol, 
    data = wine)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0502 -0.8456 -0.0244  0.8431  4.5320 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.169e+02  2.313e+01   5.055 4.41e-07 ***
Wine.ColorW          -2.075e-01  9.278e-02  -2.237 0.025319 *  
fixed.acidity         1.238e-01  2.584e-02   4.791 1.69e-06 ***
volatile.acidity     -1.298e+00  1.326e-01  -9.788  < 2e-16 ***
citric.acid          -1.039e-01  1.321e-01  -0.786 0.431605    
residual.sugar        7.259e-02  9.665e-03   7.510 6.63e-14 ***
chlorides            -1.582e-01  5.729e-01  -0.276 0.782431    
free.sulfur.dioxide   7.052e-03  1.263e-03   5.585 2.43e-08 ***
total.sulfur.dioxide -1.944e-03  5.313e-04  -3.660 0.000254 ***
density              -1.183e+02  2.345e+01  -5.044 4.68e-07 ***
pH                    1.009e+00  1.483e-01   6.801 1.12e-11 ***
sulphates             9.015e-01  1.239e-01   7.278 3.77e-13 ***
alcohol               2.131e-01  2.967e-02   7.180 7.67e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.235 on 6987 degrees of freedom
Multiple R-squared:  0.134,	Adjusted R-squared:  0.1325 
F-statistic: 90.09 on 12 and 6987 DF,  p-value: < 2.2e-16
```

The initial model observed was the full model with all the predictors. From the linear model summary, the predictors that have the most significance can be noted, with the lowest p-values. The adjusted R-squared value of 0.1325 indicates low goodness of fit.

Graph 2-5: Residual Plots



Residual plots were examined, but no action was taken to remove outliers. This is because removing points from the dataset would not help predicting the new dataset.

Analysis

Output 2: VIF of Full Model

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
         Wine.Color       fixed.acidity    volatile.acidity         citric.acid
           7.376676            5.168157            2.276289            1.666528
     residual.sugar           chlorides free.sulfur.dioxide total.sulfur.dioxide
           9.621136            1.717019            2.288343            4.174535
            density                  pH           sulphates             alcohol
          22.747253            2.564374            1.574993            5.722709
```

From VIF analysis, it can be observed that residual.sugar, density, wine.color, and alcohol have values greater than 5. To fix the model, interaction terms will be used.

Output 3: Inverse Response of Full Model

```
Power Transformations to Multinormality
                     Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
alcohol                -0.4785       -0.50      -0.6396      -0.3174
fixed.acidity          -0.4070       -0.41      -0.4795      -0.3346
volatile.acidity       -0.0577       -0.06      -0.1014      -0.0141
residual.sugar          0.2551        0.26       0.2297       0.2806
free.sulfur.dioxide     0.3432        0.33       0.3173       0.3691
total.sulfur.dioxide    0.6577        0.66       0.6298       0.6856
pH                      0.4457        0.50       0.1177       0.7738
sulphates              -0.3561       -0.33      -0.4219      -0.2903
density               -49.1893      -49.19     -52.2019     -46.1767
```

From the lambdas, the predictors can be transformed accordingly. With the transformations and new interaction terms, a new model can be formed.

New Model

Output 4: New Model Summary (parts omitted)

```
Call:
lm(formula = Quality ~ Wine.Color + Wine.Color:I(residual.sugar^(0.5)) + 
    Wine.Color:alcohol + Wine.Color:I((fixed.acidity)^(-0.5)) + 
    I(log(volatile.acidity)) + I((fixed.acidity)^(-0.5)) + I((free.sulfur.dioxide)^(0.5)) + 
    total.sulfur.dioxide + I((pH)^(0.5)) + I((pH)^(0.5)):total.sulfur.dioxide + 
    I((fixed.acidity)^(-0.5)):citric.acid + I((free.sulfur.dioxide)^(0.5)):total.sulfur.dioxide + 
    I(density^(-50)):I(sulphates^(-0.5)) + I(density^(-50)):I((pH)^(0.5)))

Residual standard error: 1.223 on 6983 degrees of freedom
Multiple R-squared:  0.1509, Adjusted R-squared:  0.149 
F-statistic: 77.57 on 16 and 6983 DF,  p-value: < 2.2e-16
```

The new model has an increased R-squared value of 0.149. AIC and BIC tests were performed to try to minimize predictors to prevent overfitting. The AIC and BIC tests resulted in a model with a slightly higher R-squared value of 0.1491, so the new model was used.

Output 5: Final Model Summary (parts omitted)

```
Call:
lm(formula = Quality ~ Wine.Color + I(log(volatile.acidity)) + 
    I((fixed.acidity)^(-0.5)) + I((free.sulfur.dioxide)^(0.5)) + 
    total.sulfur.dioxide + I((pH)^(0.5)) + Wine.Color:I(residual.sugar^(0.5)) + 
    Wine.Color:alcohol + Wine.Color:I((fixed.acidity)^(-0.5)) + 
    total.sulfur.dioxide:I((pH)^(0.5)) + I((free.sulfur.dioxide)^(0.5)):total.sulfur.dioxide + 
    I(density^(-50)):I(sulphates^(-0.5)) + I((pH)^(0.5)):I(density^(-50)))

Residual standard error: 1.223 on 6984 degrees of freedom
Multiple R-squared:  0.1509, Adjusted R-squared:  0.1491 
F-statistic: 82.75 on 15 and 6984 DF,  p-value: < 2.2e-16
```

Output 6: VIF of Final Model

```
                                                           GVIF Df GVIF^(1/(2*Df))
Wine.Color                                           296.520571  1       17.219773
I(log(volatile.acidity))                               1.927186  1        1.388231
I((fixed.acidity)^(-0.5))                              8.623238  1        2.936535
I((free.sulfur.dioxide)^(0.5))                         6.494050  1        2.548343
total.sulfur.dioxide                                2554.180985  1       50.538906
I((pH)^(0.5))                                          8.550421  1        2.924110
Wine.Color:I(residual.sugar^(0.5))                   144.660326  2        3.468066
Wine.Color:alcohol                                   781.971712  2        5.288077
Wine.Color:I((fixed.acidity)^(-0.5))                 214.010156  1       14.629086
total.sulfur.dioxide:I((pH)^(0.5))                  2473.425545  1       49.733545
I((free.sulfur.dioxide)^(0.5)):total.sulfur.dioxide   22.082258  1        4.699176
I(density^(-50)):I(sulphates^(-0.5))                   4.267518  1        2.065797
I((pH)^(0.5)):I(density^(-50))                        31.384033  1        5.602145
```