Data Analysis Project

STAT 420, Fall 2024, C. Easton, A. Roh, T. Toter, A. Treptow

December xx, 2024

Contents

Introduction

In this data analysis study we applied the principles we learned during the semester to iterate on a linear regression model.

For our case study, we chose a "Wine Quality" dataset from the UC Irvine Machine Learning Repository. The data relates to red and white variants of the Portuguese vinho verde wine samples. We drew the data from the following site:

https://archive.ics.uci.edu/dataset/186/wine+quality

Each row in the dataset of wine samples contains a record of 11 numerically measured physicochemical attributes, such as acidity, residual sugar, chlorides, and pH. We combined two datasets (one for white wine and one for red wine), resulting in an additional categorical attribute for wine type.

The 12 attributes listed above served as our source predictor variables. A final attribute from the dataset measures quality, and serves as our response variable. Each quality measurement is a subjectively-assigned integer, ranging from 1 to 10. Our goal was to build a model that could use the objectively measured predictors as inputs to estimate how a human would rate each wine.

Methods

Setup

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
```

Load and Examine the Data

```
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
red_wine = read.csv("winequality-red.csv", sep = ";")
white_wine = read.csv("winequality-white.csv", sep = ";")
# Add categorical variables for wine type
red_wine$type = "Red"
white_wine$type = "White"
wine_data = bind_rows(red_wine, white_wine) # Combine the two datasets
wine_data$type = as.factor(wine_data$type)
str(wine_data)
```

```
## ’data.frame':    6497 obs. of  13 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
##  $ type                : Factor w/ 2 levels "Red","White": 1 1 1 1 1 1 1 1 1 1 ...
```

Remove Outliers

A few data points departed noticeably from the vast majority of the rest of the data. In the real world, we would carefully exam these points and consider why they might be abberations. However, that type of analysis is outside the scope of this modeling exercise, so we deviated from normal practices and simply removed the outliers to allow us to pursue a model that handles the remainder of the data.

TODO: (Add any additional commentary from Alexander.)

```
res_sugar_6580 = which(wine_data$residual.sugar == 65.80)
free_sulf_dio_289 = which(wine_data$free.sulfur.dioxide == 289.0)
dens_10103 = which(wine_data$density == 1.0103)
remove_idx = c(res_sugar_6580, free_sulf_dio_289, dens_10103)
wine_data = wine_data[-remove_idx, ]
nrow(wine_data)
```

```
## [1] 6493
```

Fit a Full Additive Model

We began by creating an additive model using all predictors (without transformations). This model served as a baseline to judge improvements for upcoming iterations.

```
full_add_model = lm(quality ~ ., data = wine_data)
summary(full_add_model)
```

```
##
## Call:
## lm(formula = quality ~ ., data = wine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6211 -0.4695 -0.0416  0.4568  3.0248
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.255e+02  1.569e+01   7.995 1.52e-15 ***
## fixed.acidity         1.030e-01  1.667e-02   6.180 6.80e-10 ***
## volatile.acidity     -1.487e+00  8.118e-02 -18.324  < 2e-16 ***
## citric.acid          -6.694e-02  7.955e-02  -0.841   0.4001
## residual.sugar        6.784e-02  6.273e-03  10.815  < 2e-16 ***
## chlorides            -7.348e-01  3.337e-01  -2.202   0.0277 *
## free.sulfur.dioxide   5.702e-03  7.782e-04   7.327 2.63e-13 ***
## total.sulfur.dioxide -1.346e-03  3.244e-04  -4.149 3.38e-05 ***
## density              -1.249e+02  1.590e+01  -7.853 4.73e-15 ***
## pH                    5.701e-01  9.309e-02   6.125 9.63e-10 ***
## sulphates             7.477e-01  7.643e-02   9.782  < 2e-16 ***
## alcohol               1.985e-01  1.983e-02  10.013  < 2e-16 ***
## typeWhite            -4.152e-01  5.907e-02  -7.029 2.29e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7312 on 6480 degrees of freedom
## Multiple R-squared:  0.2995, Adjusted R-squared:  0.2983
## F-statistic: 230.9 on 12 and 6480 DF,  p-value: < 2.2e-16
```

Log Transformed Predictors

The predictor histograms showed that some of the predictors appear to be normally distributed, but other predictors are skewed. We experimented with a full model using log transformed variables for the ones that appeared especially skewed. We also tried a model with log transformed variables and interactive terms.

```
model_add = lm(quality ~ ., data = wine_data)
model_log = lm(quality ~ fixed.acidity + log(volatile.acidity) + citric.acid +
                 residual.sugar + chlorides + log(free.sulfur.dioxide) +
                 total.sulfur.dioxide + density + (pH) + (sulphates) +
                 (alcohol) + type, data = wine_data)
model_log_int = lm(quality ~ (fixed.acidity + log(volatile.acidity) +
                                citric.acid + residual.sugar + chlorides +
                                log(free.sulfur.dioxide) +
                                total.sulfur.dioxide + density + (pH) +
                                (sulphates) + (alcohol) + type)^2,
                   data = wine_data)
```

Fit a Full Interaction Model

Next, we created an interaction model using all predictors (without transformations).

```
full_int_model = lm(quality ~ .^2, data = wine_data)
summary(full_int_model)
```

```
##
## Call:
## lm(formula = quality ~ .^2, data = wine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2817 -0.4634 -0.0226  0.4378  2.9827
##
## Coefficients:
##                                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               -5.113e+02  2.908e+02  -1.758 0.078739
## fixed.acidity                              4.804e+00  7.175e+00   0.670 0.503145
## volatile.acidity                          -2.842e+01  1.142e+02  -0.249 0.803505
## citric.acid                               -8.157e+01  1.201e+02  -0.679 0.497164
## residual.sugar                            5.248e+00  1.256e+00   4.178 2.97e-05
## chlorides                                -1.105e+03  5.069e+02  -2.180 0.029289
## free.sulfur.dioxide                      -4.293e+00  1.376e+00  -3.119 0.001824
## total.sulfur.dioxide                      5.283e-01  4.911e-01   1.076 0.282030
## density                                   5.246e+02  2.908e+02   1.804 0.071325
## pH                                        1.180e+02  7.404e+01   1.594 0.110878
## sulphates                                 7.002e+01  1.168e+02   0.599 0.548874
## alcohol                                   1.996e+01  6.157e+00   3.242 0.001193
## typeWhite                                 1.635e+02  6.045e+01   2.705 0.006848
## fixed.acidity:volatile.acidity           -5.610e-02  1.357e-01  -0.413 0.679347
## fixed.acidity:citric.acid                -8.851e-02  1.254e-01  -0.706 0.480142
## fixed.acidity:residual.sugar              8.185e-03  3.994e-03   2.049 0.040463
## fixed.acidity:chlorides                  -1.901e+00  5.482e-01  -3.468 0.000528
## fixed.acidity:free.sulfur.dioxide        -9.136e-04  1.423e-03  -0.642 0.520926
## fixed.acidity:total.sulfur.dioxide       -1.118e-04  5.631e-04  -0.198 0.842691
## fixed.acidity:density                    -5.383e+00  7.111e+00  -0.757 0.449053
## fixed.acidity:pH                          2.455e-01  6.643e-02   3.696 0.000221
## fixed.acidity:sulphates                   2.255e-01  1.199e-01   1.881 0.060014
## fixed.acidity:alcohol                    -1.515e-02  1.339e-02  -1.131 0.257945
## fixed.acidity:typeWhite                   7.880e-02  7.812e-02   1.009 0.313197
## volatile.acidity:citric.acid              1.059e+00  5.649e-01   1.875 0.060865
## volatile.acidity:residual.sugar          -6.485e-02  4.828e-02  -1.343 0.179264
## volatile.acidity:chlorides                2.535e+00  2.546e+00   0.996 0.319368
## volatile.acidity:free.sulfur.dioxide      9.882e-03  7.380e-03   1.339 0.180631
## volatile.acidity:total.sulfur.dioxide     5.365e-03  2.725e-03   1.969 0.049008
## volatile.acidity:density                  1.999e+01  1.160e+02   0.172 0.863171
## volatile.acidity:pH                       8.379e-01  8.029e-01   1.044 0.296708
## volatile.acidity:sulphates               -1.146e-01  6.550e-01  -0.175 0.861185
## volatile.acidity:alcohol                  4.455e-01  1.444e-01   3.085 0.002044
## volatile.acidity:typeWhite               -1.177e+00  4.008e-01  -2.937 0.003321
## citric.acid:residual.sugar               -5.847e-02  4.648e-02  -1.258 0.208482
## citric.acid:chlorides                     3.360e+00  2.271e+00   1.480 0.139036
## citric.acid:free.sulfur.dioxide           7.994e-03  6.348e-03   1.259 0.207979
## citric.acid:total.sulfur.dioxide         -1.314e-03  2.440e-03  -0.538 0.590327
## citric.acid:density                       7.932e+01  1.215e+02   0.653 0.513742
## citric.acid:pH                           -7.930e-02  7.443e-01  -0.107 0.915158
## citric.acid:sulphates                    -9.245e-01  7.022e-01  -1.317 0.187993
## citric.acid:alcohol                       3.035e-01  1.541e-01   1.970 0.048911
## citric.acid:typeWhite                     6.663e-01  4.565e-01   1.460 0.144449
## residual.sugar:chlorides                 -6.104e-01  2.348e-01  -2.599 0.009361
## residual.sugar:free.sulfur.dioxide       -1.914e-03  5.435e-04  -3.522 0.000432
## residual.sugar:total.sulfur.dioxide       4.070e-04  2.034e-04   2.001 0.045475
## residual.sugar:density                   -5.061e+00  1.244e+00  -4.067 4.83e-05
## residual.sugar:pH                        -3.359e-02  2.970e-02  -1.131 0.258191
## residual.sugar:sulphates                  1.434e-03  4.808e-02   0.030 0.976208
## residual.sugar:alcohol                   -1.367e-03  4.324e-03  -0.316 0.751947
## residual.sugar:typeWhite                  1.061e-02  2.865e-02   0.370 0.711240
## chlorides:free.sulfur.dioxide             7.326e-03  2.821e-02   0.260 0.795102
## chlorides:total.sulfur.dioxide           -6.946e-03  1.525e-02  -0.455 0.648810
## chlorides:density                         1.163e+03  5.121e+02   2.271 0.023189
## chlorides:pH                             -1.164e+01  3.920e+00  -2.969 0.002997
## chlorides:sulphates                      -7.342e+00  1.961e+00  -3.744 0.000183
## chlorides:alcohol                         3.016e-01  7.056e-01   0.427 0.669037
## chlorides:typeWhite                      -4.112e-01  2.511e+00  -0.164 0.869931
## free.sulfur.dioxide:total.sulfur.dioxide -1.565e-04  1.413e-05 -11.078  < 2e-16
## free.sulfur.dioxide:density               4.262e+00  1.393e+00   3.059 0.002228
## free.sulfur.dioxide:pH                   -3.964e-03  7.628e-03  -0.520 0.603267
## free.sulfur.dioxide:sulphates             1.652e-02  6.351e-03   2.601 0.009321
## free.sulfur.dioxide:alcohol               6.427e-03  1.841e-03   3.492 0.000483
## free.sulfur.dioxide:typeWhite             3.487e-02  5.088e-03   6.854 7.83e-12
## total.sulfur.dioxide:density             -5.131e-01  4.988e-01  -1.029 0.303718
## total.sulfur.dioxide:pH                  -1.317e-03  3.272e-03  -0.402 0.687343
## total.sulfur.dioxide:sulphates           -1.204e-02  2.649e-03  -4.545 5.59e-06
## total.sulfur.dioxide:alcohol             -7.620e-04  6.589e-04  -1.157 0.247509
## total.sulfur.dioxide:typeWhite            2.150e-03  1.438e-03   1.495 0.135035
## density:pH                               -1.210e+02  7.382e+01  -1.639 0.101360
## density:sulphates                        -7.503e+01  1.181e+02  -0.635 0.525147
## density:alcohol                          -1.971e+01  6.294e+00  -3.131 0.001748
## density:typeWhite                        -1.703e+02  6.130e+01  -2.779 0.005469
## pH:sulphates                              1.993e+00  6.533e-01   3.051 0.002291
## pH:alcohol                               -7.377e-02  1.152e-01  -0.640 0.521966
## pH:typeWhite                              1.997e+00  5.163e-01   3.867 0.000111
## sulphates:alcohol                        -1.107e-01  1.417e-01  -0.781 0.434795
## sulphates:typeWhite                       1.417e-01  4.724e-01   0.300 0.764249
## alcohol:typeWhite                        -2.238e-01  8.722e-02  -2.566 0.010301
##
## (Intercept)                              .
## fixed.acidity
## volatile.acidity
## citric.acid
## residual.sugar                           ***
## chlorides                                *
## free.sulfur.dioxide                      **
## total.sulfur.dioxide
## density                                  .
## pH
## sulphates
## alcohol                                  **
## typeWhite                                **
## fixed.acidity:volatile.acidity
## fixed.acidity:citric.acid
## fixed.acidity:residual.sugar             *
## fixed.acidity:chlorides                  ***
## fixed.acidity:free.sulfur.dioxide
## fixed.acidity:total.sulfur.dioxide
## fixed.acidity:density
## fixed.acidity:pH                         ***
## fixed.acidity:sulphates                  .
## fixed.acidity:alcohol
## fixed.acidity:typeWhite
## volatile.acidity:citric.acid             .
## volatile.acidity:residual.sugar
## volatile.acidity:chlorides
## volatile.acidity:free.sulfur.dioxide
## volatile.acidity:total.sulfur.dioxide    *
## volatile.acidity:density
## volatile.acidity:pH
## volatile.acidity:sulphates
## volatile.acidity:alcohol                **
## volatile.acidity:typeWhite              **
## citric.acid:residual.sugar
## citric.acid:chlorides
## citric.acid:free.sulfur.dioxide
## citric.acid:total.sulfur.dioxide
## citric.acid:density
## citric.acid:pH
## citric.acid:sulphates
## citric.acid:alcohol                     *
## citric.acid:typeWhite
## residual.sugar:chlorides                **
## residual.sugar:free.sulfur.dioxide      ***
## residual.sugar:total.sulfur.dioxide     *
## residual.sugar:density                  ***
## residual.sugar:pH
## residual.sugar:sulphates
## residual.sugar:alcohol
## residual.sugar:typeWhite
## chlorides:free.sulfur.dioxide
## chlorides:total.sulfur.dioxide
## chlorides:density                       *
## chlorides:pH                            **
## chlorides:sulphates                     ***
## chlorides:alcohol
## chlorides:typeWhite
## free.sulfur.dioxide:total.sulfur.dioxide ***
## free.sulfur.dioxide:density             **
## free.sulfur.dioxide:pH
## free.sulfur.dioxide:sulphates           **
## free.sulfur.dioxide:alcohol             ***
## free.sulfur.dioxide:typeWhite           ***
## total.sulfur.dioxide:density
## total.sulfur.dioxide:pH
## total.sulfur.dioxide:sulphates          ***
## total.sulfur.dioxide:alcohol
## total.sulfur.dioxide:typeWhite
## density:pH
## density:sulphates
## density:alcohol                         **
## density:typeWhite                       **
## pH:sulphates                            **
## pH:alcohol
## pH:typeWhite                            ***
## sulphates:alcohol
## sulphates:typeWhite
## alcohol:typeWhite                       *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6984 on 6414 degrees of freedom
## Multiple R-squared:  0.3673, Adjusted R-squared:  0.3597
## F-statistic: 47.75 on 78 and 6414 DF,  p-value: < 2.2e-16
```

Compare the Adjusted R-Squared for Various Models

```
summary(model_add)$adj.r.squared
```

```
## [1] 0.2982511
```

```
summary(model_log)$adj.r.squared
```

```
## [1] 0.3091152
```

```
summary(full_int_model)$adj.r.squared
```

```
## [1] 0.359651
```

```
summary(model_log_int)$adj.r.squared
```

```
## [1] 0.3682492
```

Variance Inflation Factors

We calculated the Variance Inflation Factors (VIF) to help us identify issues of multicollinearity between predictors.

```
if (!require(car)) install.packages("car")
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(car)
vif_values = vif(full_add_model)
vif_values
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##             5.674492             2.165464             1.622715
##       residual.sugar            chlorides  free.sulfur.dioxide
##            10.453938             1.660500             2.242120
## total.sulfur.dioxide              density                   pH
##             4.062701            26.474789             2.720556
##            sulphates              alcohol                 type
##             1.570941             6.790581             7.865940
```

Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC) Modelling

```r
model_bac_aic = step(model_log_int, trace = 0)
model_bac_bic = step(model_log_int, k = log(nrow(wine_data)), trace = 0)
model_both_aic = step(model_log_int, direction = "both", trace = 0)
summary(model_bac_aic)$adj.r.squared
```

```
## [1] 0.3694712
```

```r
summary(model_bac_bic)$adj.r.squared
```

```
## [1] 0.3659214
```

```r
summary(model_both_aic)$adj.r.squared
```

```
## [1] 0.3694712
```

Analysis of Variance (ANOVA)

```r
anova(model_bac_aic, model_log_int)
```

```
## Analysis of Variance Table
##
## Model 1: quality ~ fixed.acidity + log(volatile.acidity) + citric.acid +
##     residual.sugar + chlorides + log(free.sulfur.dioxide) + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + type + fixed.acidity:residual.sugar +
##     fixed.acidity:chlorides + fixed.acidity:density + fixed.acidity:pH +
##     fixed.acidity:sulphates + fixed.acidity:alcohol + log(volatile.acidity):citric.acid +
##     log(volatile.acidity):residual.sugar + log(volatile.acidity):log(free.sulfur.dioxide) +
##     log(volatile.acidity):pH + log(volatile.acidity):sulphates +
##     log(volatile.acidity):alcohol + log(volatile.acidity):type +
##     citric.acid:residual.sugar + citric.acid:sulphates + citric.acid:alcohol +
##     citric.acid:type + residual.sugar:chlorides + residual.sugar:log(free.sulfur.dioxide) +
##     residual.sugar:total.sulfur.dioxide + residual.sugar:density +
##     chlorides:density + chlorides:pH + chlorides:sulphates +
##     log(free.sulfur.dioxide):total.sulfur.dioxide + log(free.sulfur.dioxide):density +
##     log(free.sulfur.dioxide):pH + log(free.sulfur.dioxide):alcohol +
##     log(free.sulfur.dioxide):type + total.sulfur.dioxide:sulphates +
##     total.sulfur.dioxide:type + density:pH + density:alcohol +
##     density:type + pH:sulphates + pH:alcohol + pH:type + alcohol:type
## Model 2: quality ~ (fixed.acidity + log(volatile.acidity) + citric.acid +
##     residual.sugar + chlorides + log(free.sulfur.dioxide) + total.sulfur.dioxide +
##     density + (pH) + (sulphates) + (alcohol) + type)^2
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1   6442 3094.3
## 2   6414 3086.9 28    7.4786 0.555  0.972
```

```r
anova(model_bac_bic, model_log_int)
```

```
## Analysis of Variance Table
##
## Model 1: quality ~ fixed.acidity + log(volatile.acidity) + citric.acid +
##     residual.sugar + chlorides + log(free.sulfur.dioxide) + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + type + fixed.acidity:chlorides +
##     fixed.acidity:pH + log(volatile.acidity):citric.acid + log(volatile.acidity):pH +
##     log(volatile.acidity):alcohol + citric.acid:alcohol + citric.acid:type +
##     residual.sugar:chlorides + residual.sugar:log(free.sulfur.dioxide) +
##     residual.sugar:total.sulfur.dioxide + residual.sugar:density +
##     chlorides:density + chlorides:pH + chlorides:sulphates +
##     log(free.sulfur.dioxide):total.sulfur.dioxide + log(free.sulfur.dioxide):density +
##     log(free.sulfur.dioxide):alcohol + log(free.sulfur.dioxide):type +
##     total.sulfur.dioxide:sulphates + density:pH + density:alcohol +
##     density:type + pH:sulphates + pH:type + alcohol:type
## Model 2: quality ~ (fixed.acidity + log(volatile.acidity) + citric.acid +
##     residual.sugar + chlorides + log(free.sulfur.dioxide) + total.sulfur.dioxide +
##     density + (pH) + (sulphates) + (alcohol) + type)^2
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   6455 3118.1
## 2   6414 3086.9 41    31.179 1.5801 0.01066 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

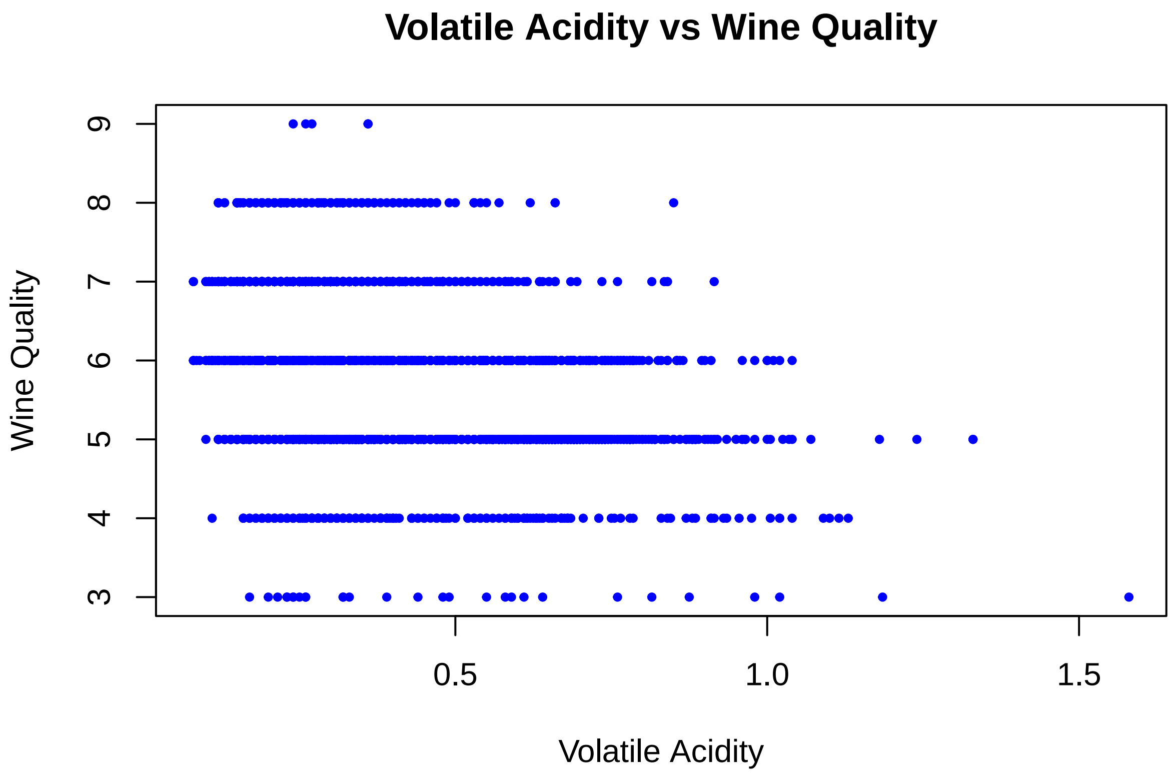
```r
anova(model_both_aic, model_log_int)
```

```
## Analysis of Variance Table
##
## Model 1: quality ~ fixed.acidity + log(volatile.acidity) + citric.acid +
##     residual.sugar + chlorides + log(free.sulfur.dioxide) + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + type + fixed.acidity:residual.sugar +
##     fixed.acidity:chlorides + fixed.acidity:density + fixed.acidity:pH +
##     fixed.acidity:sulphates + fixed.acidity:alcohol + log(volatile.acidity):citric.acid +
##     log(volatile.acidity):residual.sugar + log(volatile.acidity):log(free.sulfur.dioxide) +
##     log(volatile.acidity):pH + log(volatile.acidity):sulphates +
##     log(volatile.acidity):alcohol + log(volatile.acidity):type +
##     citric.acid:residual.sugar + citric.acid:sulphates + citric.acid:alcohol +
##     citric.acid:type + residual.sugar:chlorides + residual.sugar:log(free.sulfur.dioxide) +
##     residual.sugar:total.sulfur.dioxide + residual.sugar:density +
##     chlorides:density + chlorides:pH + chlorides:sulphates +
##     log(free.sulfur.dioxide):total.sulfur.dioxide + log(free.sulfur.dioxide):density +
##     log(free.sulfur.dioxide):pH + log(free.sulfur.dioxide):alcohol +
##     log(free.sulfur.dioxide):type + total.sulfur.dioxide:sulphates +
##     total.sulfur.dioxide:type + density:pH + density:alcohol +
##     density:type + pH:sulphates + pH:alcohol + pH:type + alcohol:type
## Model 2: quality ~ (fixed.acidity + log(volatile.acidity) + citric.acid +
##     residual.sugar + chlorides + log(free.sulfur.dioxide) + total.sulfur.dioxide +
##     density + (pH) + (sulphates) + (alcohol) + type)^2
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1   6442 3094.3
## 2   6414 3086.9 28    7.4786 0.555  0.972
```

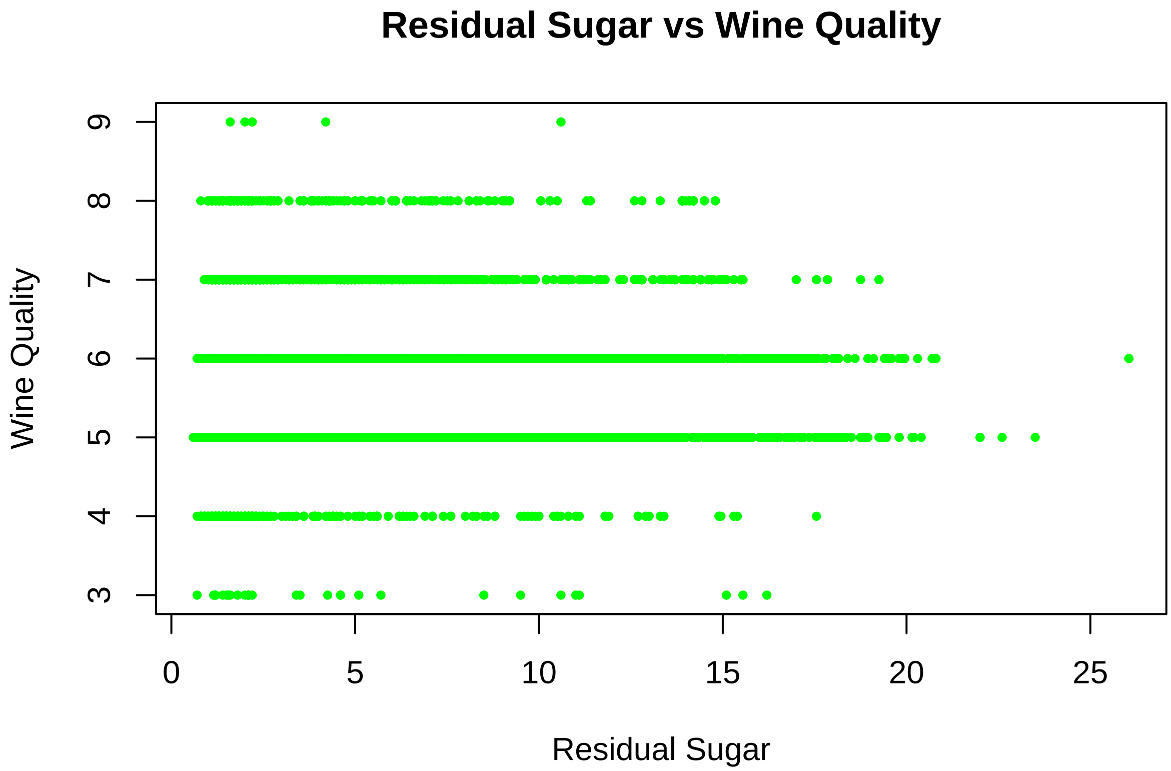
Results

Predictor Scatterplots

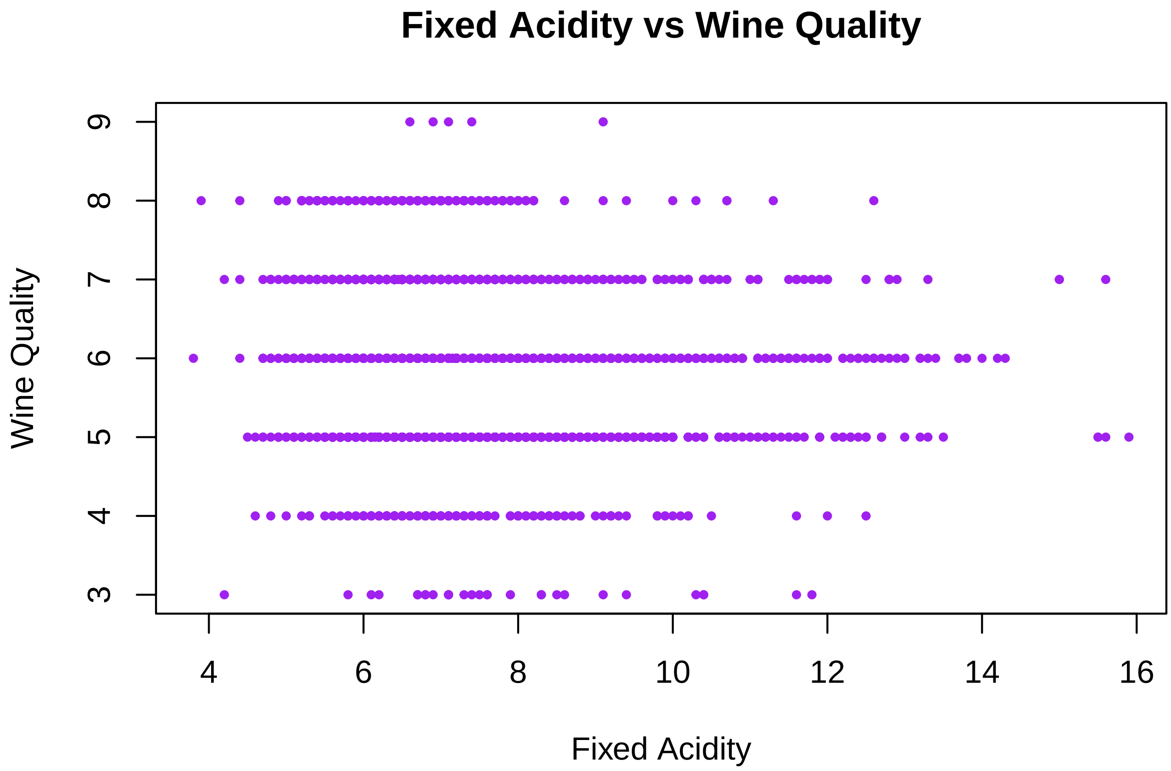
```r
plot(wine_data$volatile.acidity, wine_data$quality,
     main = "Volatile Acidity vs Wine Quality",
     xlab = "Volatile Acidity",
     ylab = "Wine Quality",
     col = "blue", pch = 19, cex = 0.5)
```



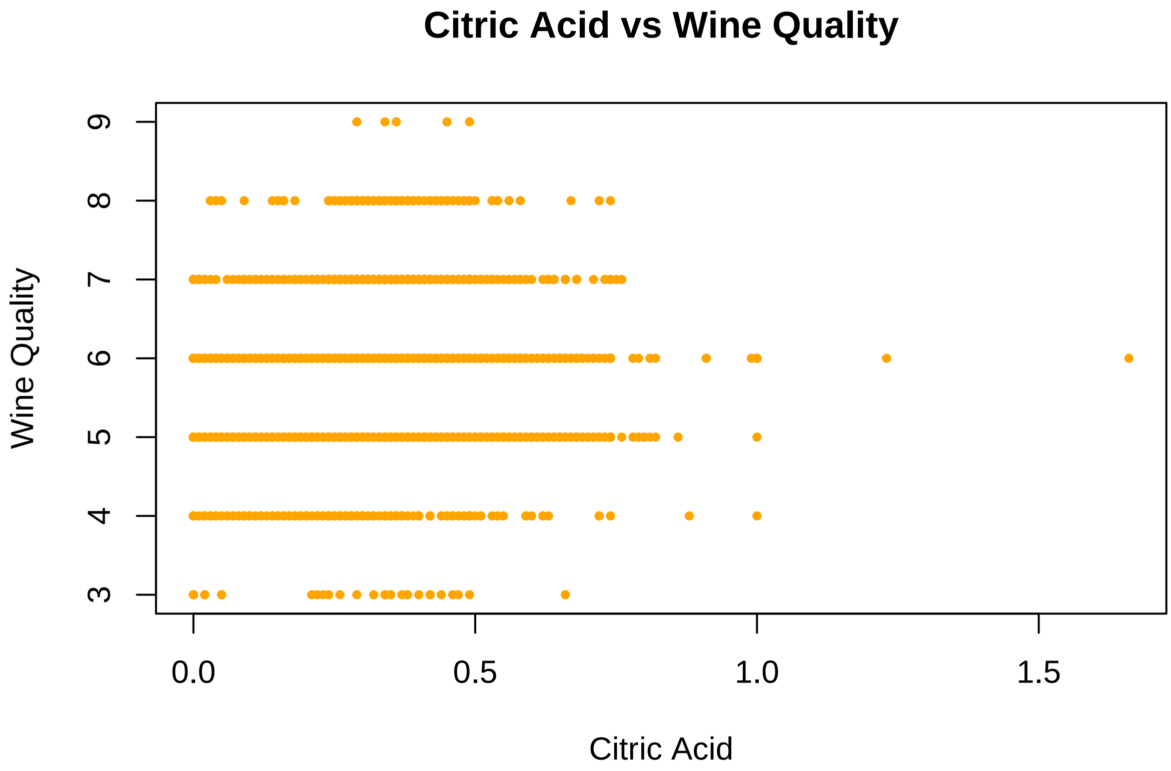
```r
plot(wine_data$residual.sugar, wine_data$quality,
     main = "Residual Sugar vs Wine Quality",
     xlab = "Residual Sugar",
     ylab = "Wine Quality",
     col = "green", pch = 19, cex = 0.5)
```



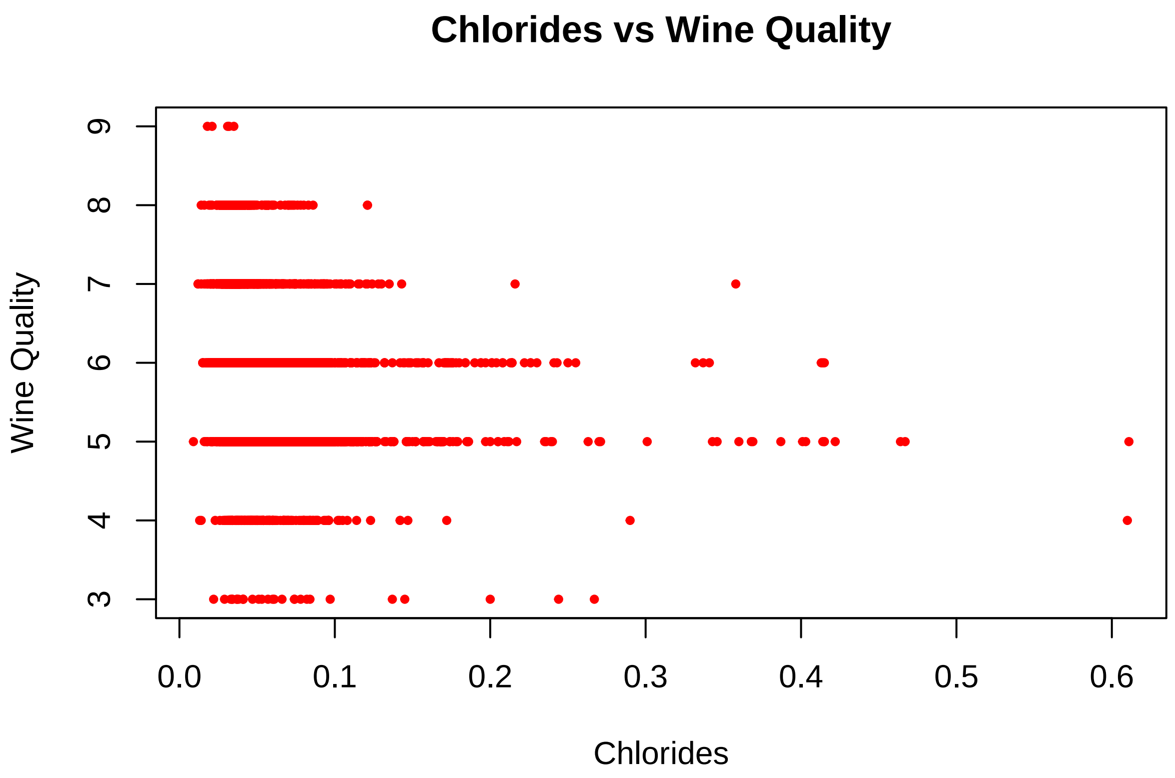
```
plot(wine_data$fixed.acidity, wine_data$quality,
     main = "Fixed Acidity vs Wine Quality",
     xlab = "Fixed Acidity",
     ylab = "Wine Quality",
     col = "purple", pch = 19, cex = 0.5)
```



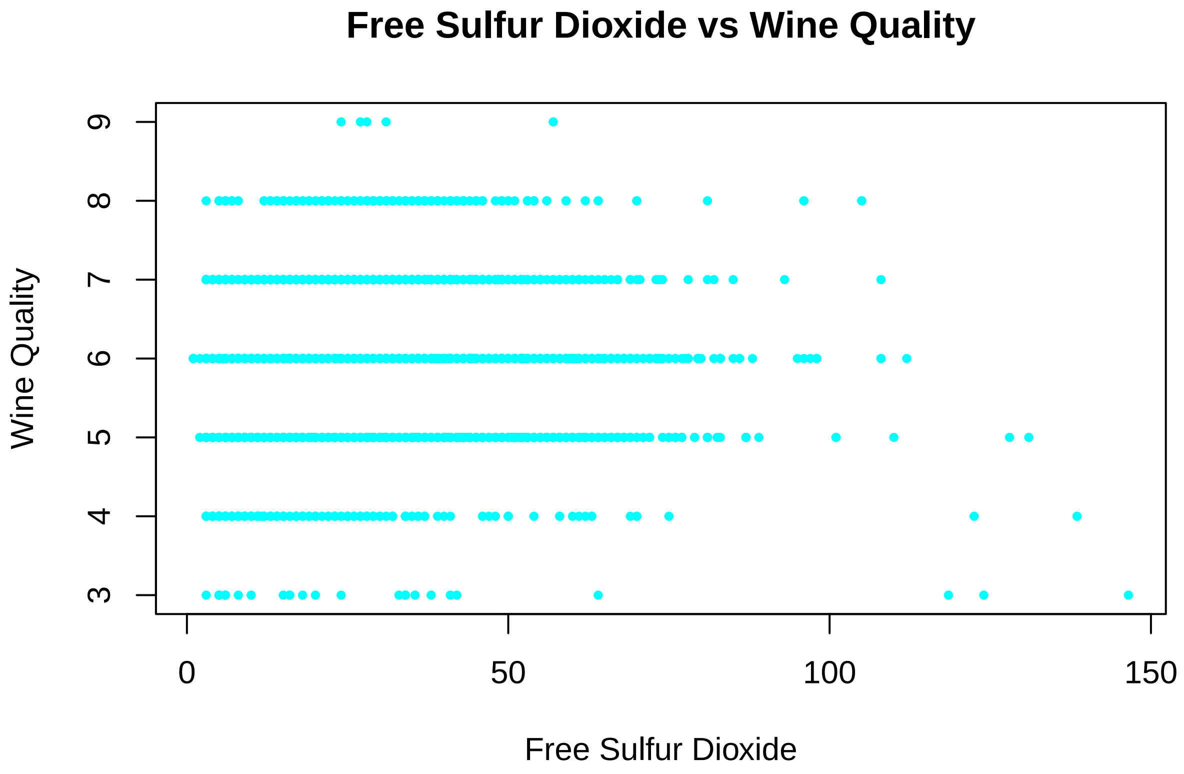
```r
plot(wine_data$citric.acid, wine_data$quality,
     main = "Citric Acid vs Wine Quality",
     xlab = "Citric Acid",
     ylab = "Wine Quality",
     col = "orange", pch = 19, cex = 0.5)
```



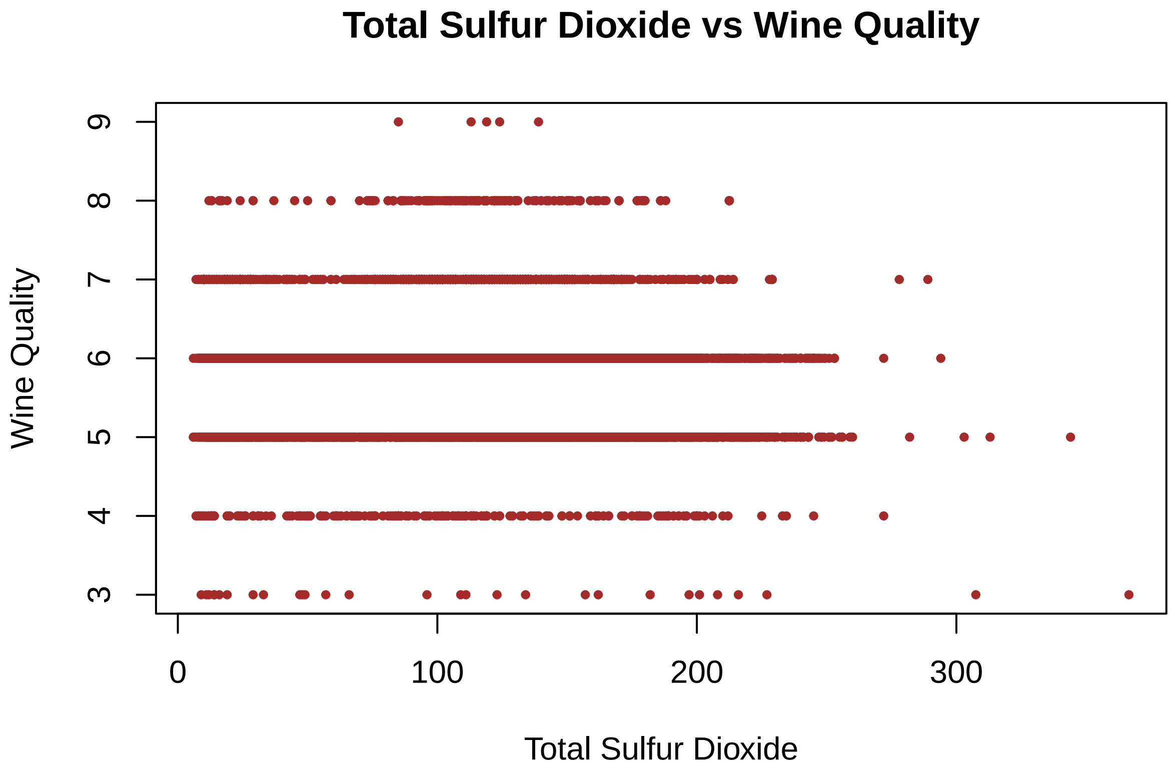
```r
plot(wine_data$chlorides, wine_data$quality,
     main = "Chlorides vs Wine Quality",
     xlab = "Chlorides",
     ylab = "Wine Quality",
     col = "red", pch = 19, cex = 0.5)
```



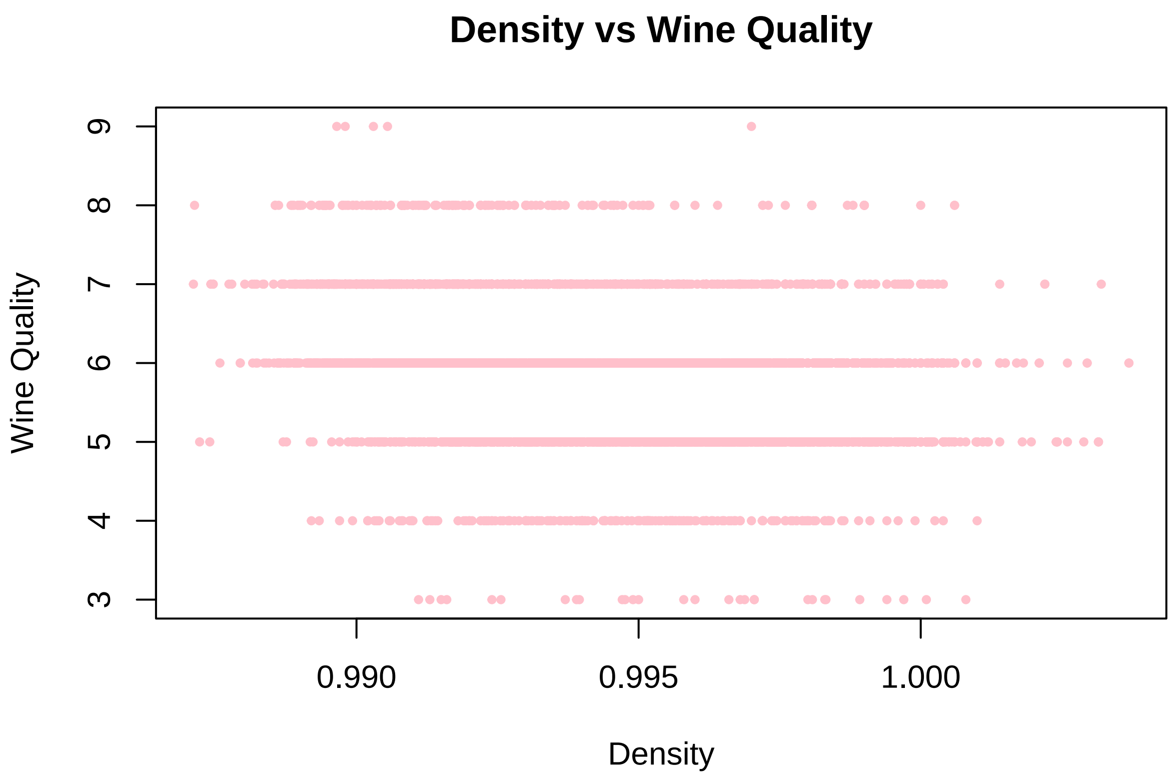
```r
plot(wine_data$free.sulfur.dioxide, wine_data$quality,
     main = "Free Sulfur Dioxide vs Wine Quality",
     xlab = "Free Sulfur Dioxide",
     ylab = "Wine Quality",
     col = "cyan", pch = 19, cex = 0.5)
```



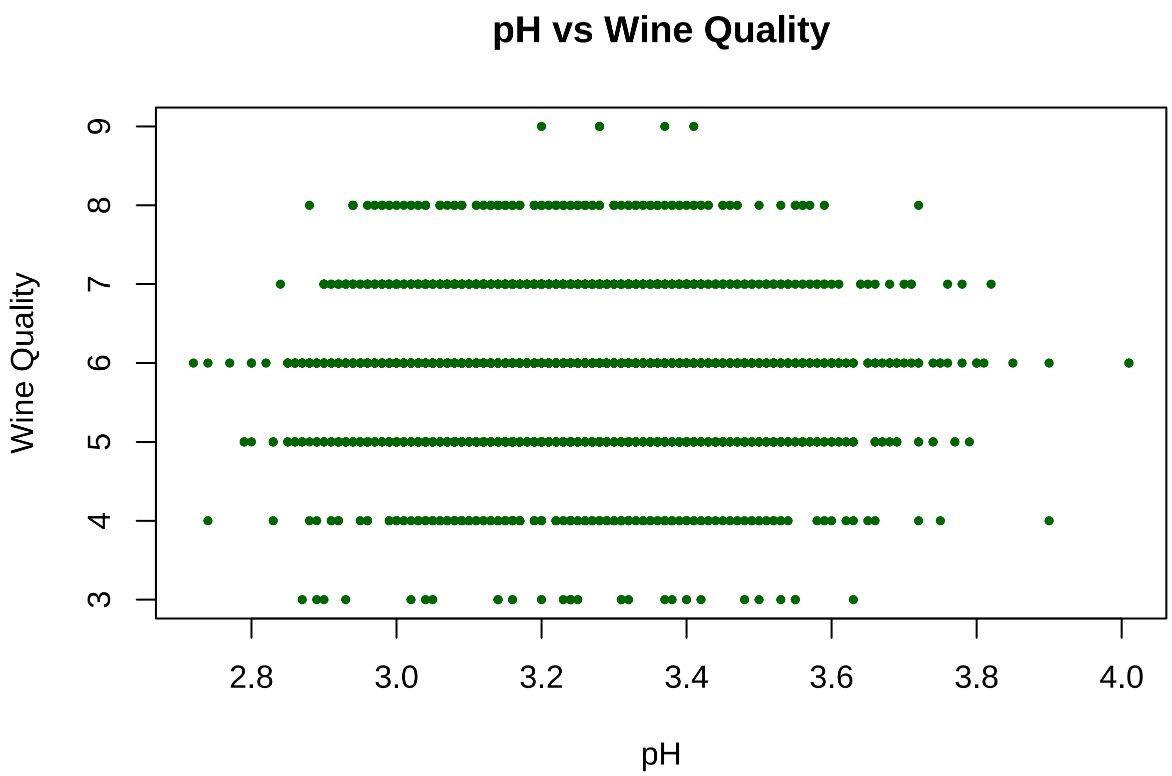
```
plot(wine_data$total.sulfur.dioxide, wine_data$quality,
     main = "Total Sulfur Dioxide vs Wine Quality",
     xlab = "Total Sulfur Dioxide",
     ylab = "Wine Quality",
     col = "brown", pch = 19, cex = 0.5)
```



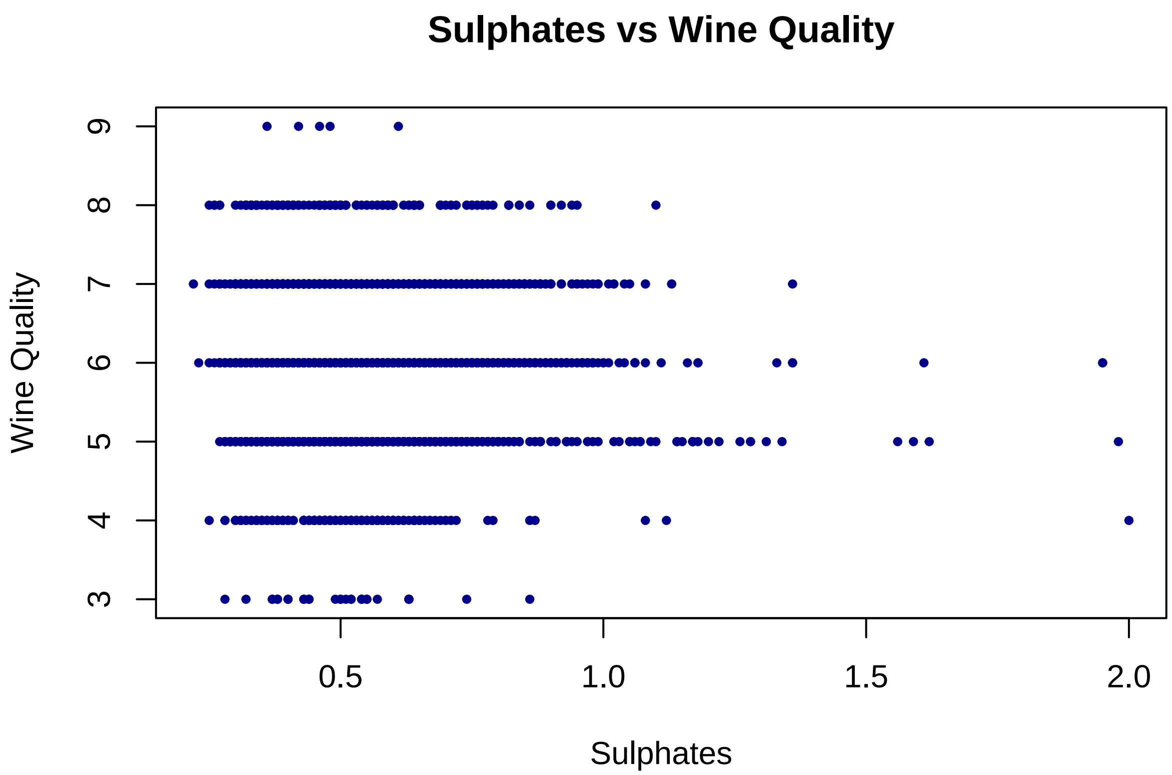
```r
plot(wine_data$density, wine_data$quality,
     main = "Density vs Wine Quality",
     xlab = "Density",
     ylab = "Wine Quality",
     col = "pink", pch = 19, cex = 0.5)
```



```r
plot(wine_data$pH, wine_data$quality,
     main = "pH vs Wine Quality",
     xlab = "pH",
     ylab = "Wine Quality",
     col = "darkgreen", pch = 19, cex = 0.5)
```

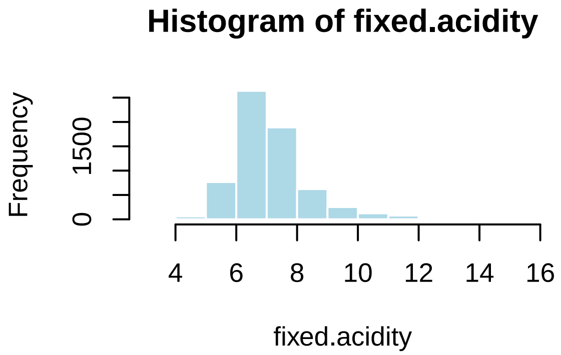


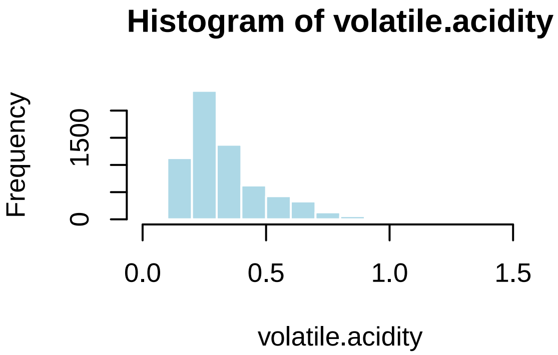
```
plot(wine_data$sulphates, wine_data$quality,
     main = "Sulphates vs Wine Quality",
     xlab = "Sulphates",
     ylab = "Wine Quality",
     col = "darkblue", pch = 19, cex = 0.5)
```

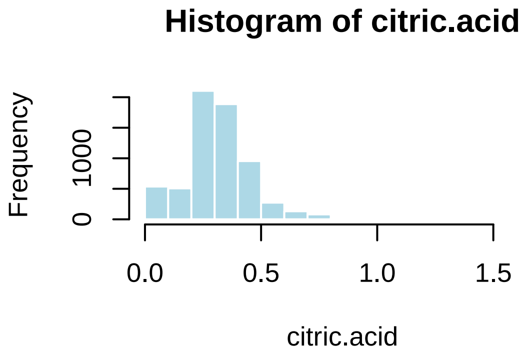


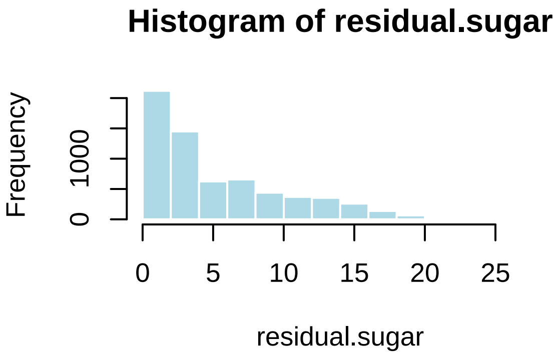
Predictor Histograms

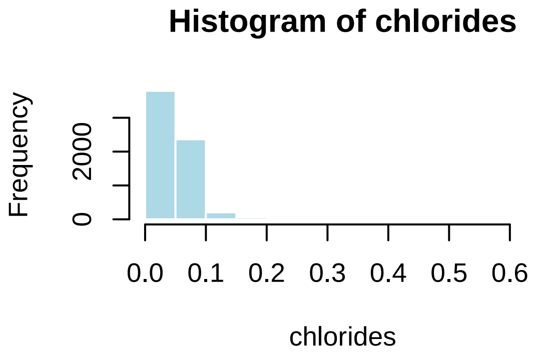
```
numeric_columns = wine_data[sapply(wine_data, is.numeric)]
par(mfrow = c(2, 2))
sapply(names(numeric_columns), function(column) {
  hist(numeric_columns[[column]], main = paste("Histogram of", column),
       xlab = column, col = "lightblue", border = "white")
})
```

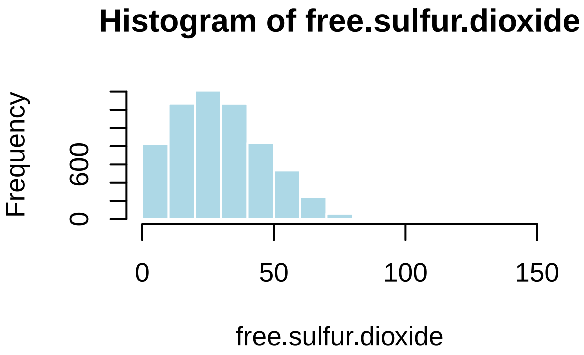

Histogram of fixed.acidity
Frequency
0
1500
4
6
8
10
12
14
16
fixed.acidity

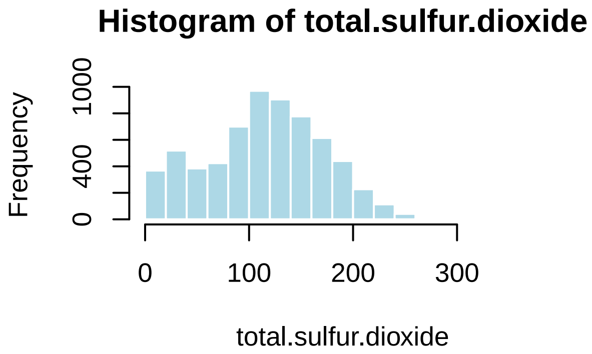

Histogram of volatile.acidity
Frequency
0
1500
0.0
0.5
1.0
1.5
volatile.acidity

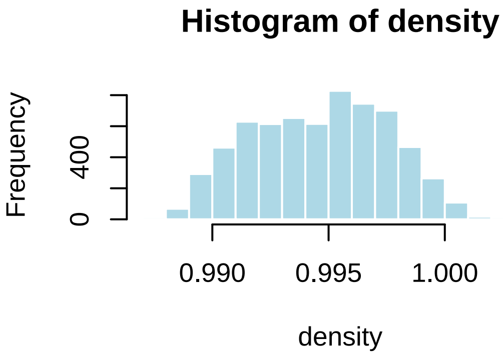

Histogram of citric.acid
Frequency
0
1000
0.0
0.5
1.0
1.5
citric.acid

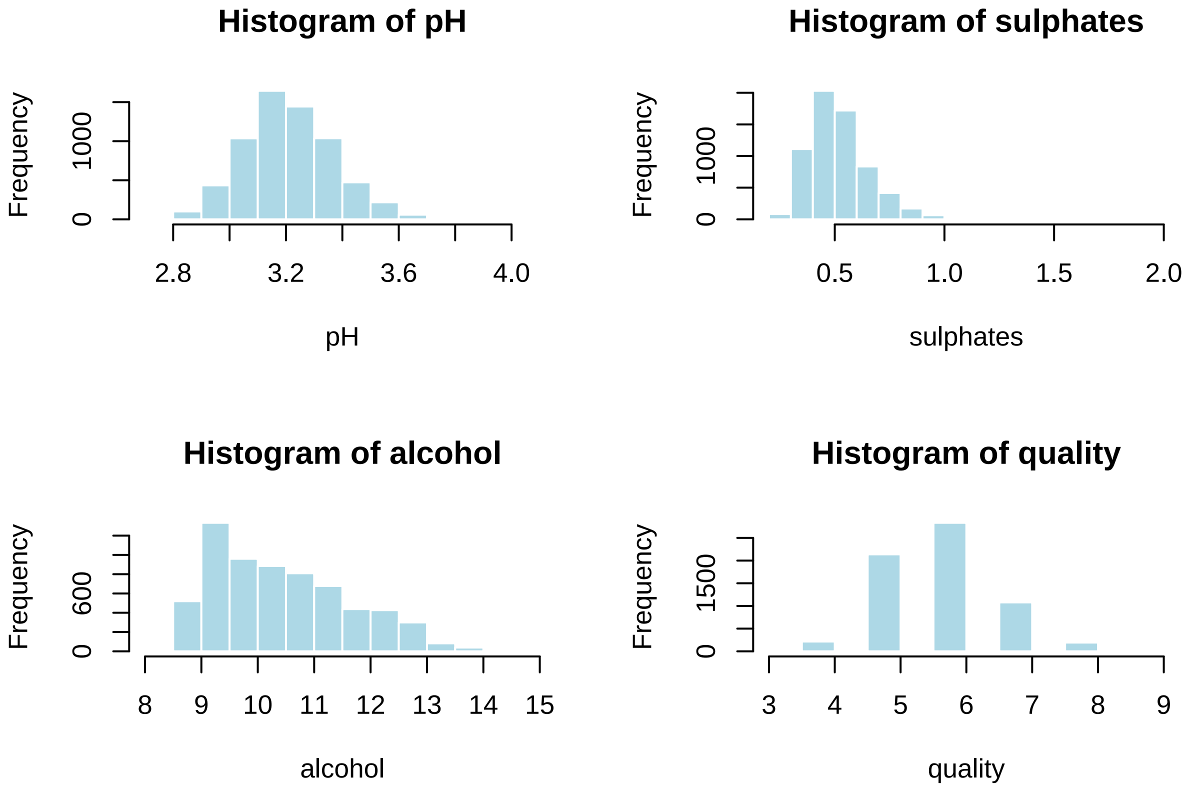

Histogram of residual.sugar
Frequency
0
1000
0
5
10
15
20
25
residual.sugar


Histogram of chlorides
Frequency
0
2000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides


Histogram of free.sulfur.dioxide
Frequency
0
600
0
50
100
150
free.sulfur.dioxide


Histogram of total.sulfur.dioxide
Frequency
0
400
1000
0
100
200
300
total.sulfur.dioxide


Histogram of density
Frequency
0
400
0.990
0.995
1.000
density



```
##          fixed.acidity                volatile.acidity
## breaks   integer,14                   numeric,17
## counts   integer,13                   integer,16
## density  numeric,13                   numeric,16
## mids     numeric,13                   numeric,16
## xname    "numeric_columns[[column]]" "numeric_columns[[column]]"
## equidist TRUE                         TRUE
##          citric.acid                  residual.sugar
## breaks   numeric,18                   numeric,15
## counts   integer,17                   integer,14
## density  numeric,17                   numeric,14
## mids     numeric,17                   numeric,14
## xname    "numeric_columns[[column]]" "numeric_columns[[column]]"
## equidist TRUE                         TRUE
##          chlorides                    free.sulfur.dioxide
## breaks   numeric,14                   numeric,16
## counts   integer,13                   integer,15
## density  numeric,13                   numeric,15
## mids     numeric,13                   numeric,15
## xname    "numeric_columns[[column]]" "numeric_columns[[column]]"
## equidist TRUE                         TRUE
##          total.sulfur.dioxide         density
## breaks   numeric,20                   numeric,18
## counts   integer,19                   integer,17
## density  numeric,19                   numeric,17
## mids     numeric,19                   numeric,17
## xname    "numeric_columns[[column]]" "numeric_columns[[column]]"
## equidist TRUE                        TRUE
##          pH                          sulphates
## breaks   numeric,15                  numeric,19
## counts   integer,14                  integer,18
## density  numeric,14                  numeric,18
## mids     numeric,14                  numeric,18
## xname    "numeric_columns[[column]]" "numeric_columns[[column]]"
## equidist TRUE                        TRUE
##          alcohol                     quality
## breaks   numeric,15                  numeric,13
## counts   integer,14                  integer,12
## density  numeric,14                  numeric,12
## mids     numeric,14                  numeric,12
## xname    "numeric_columns[[column]]" "numeric_columns[[column]]"
## equidist TRUE                        TRUE
```

Predictor Correlation Heatmap

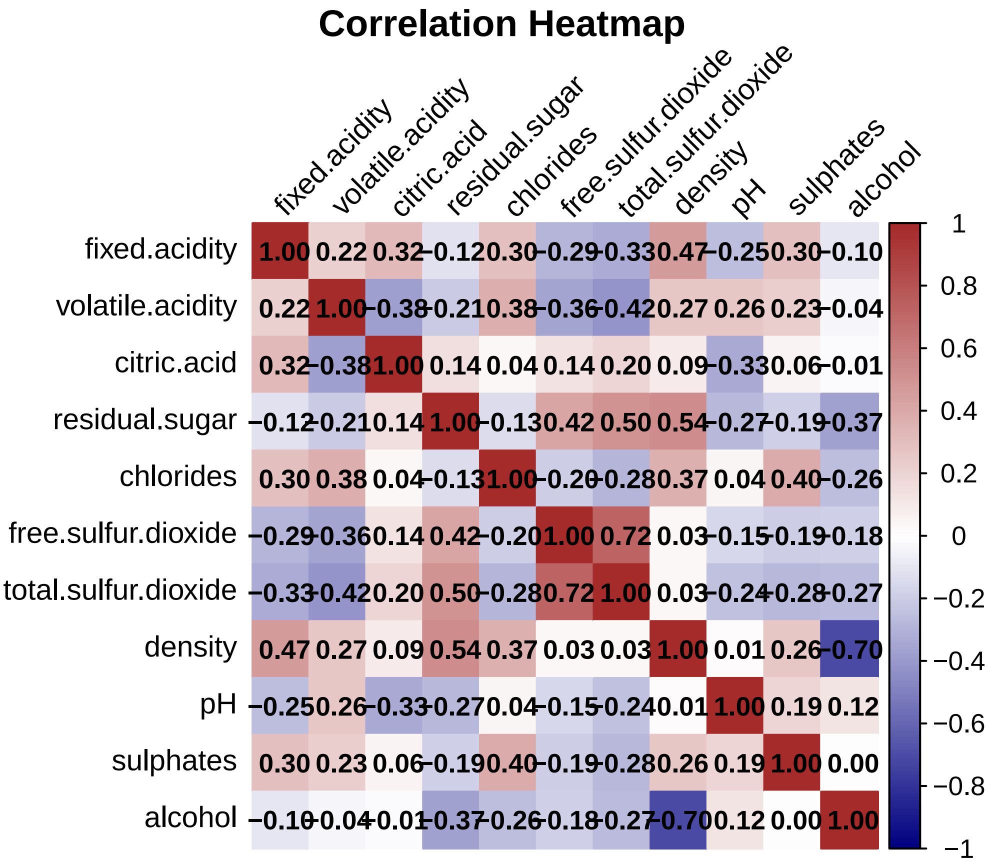
```r
if (!require(corrplot)) install.packages("corrplot")
```

```
## Loading required package: corrplot
```

```
## corrplot 0.95 loaded
```

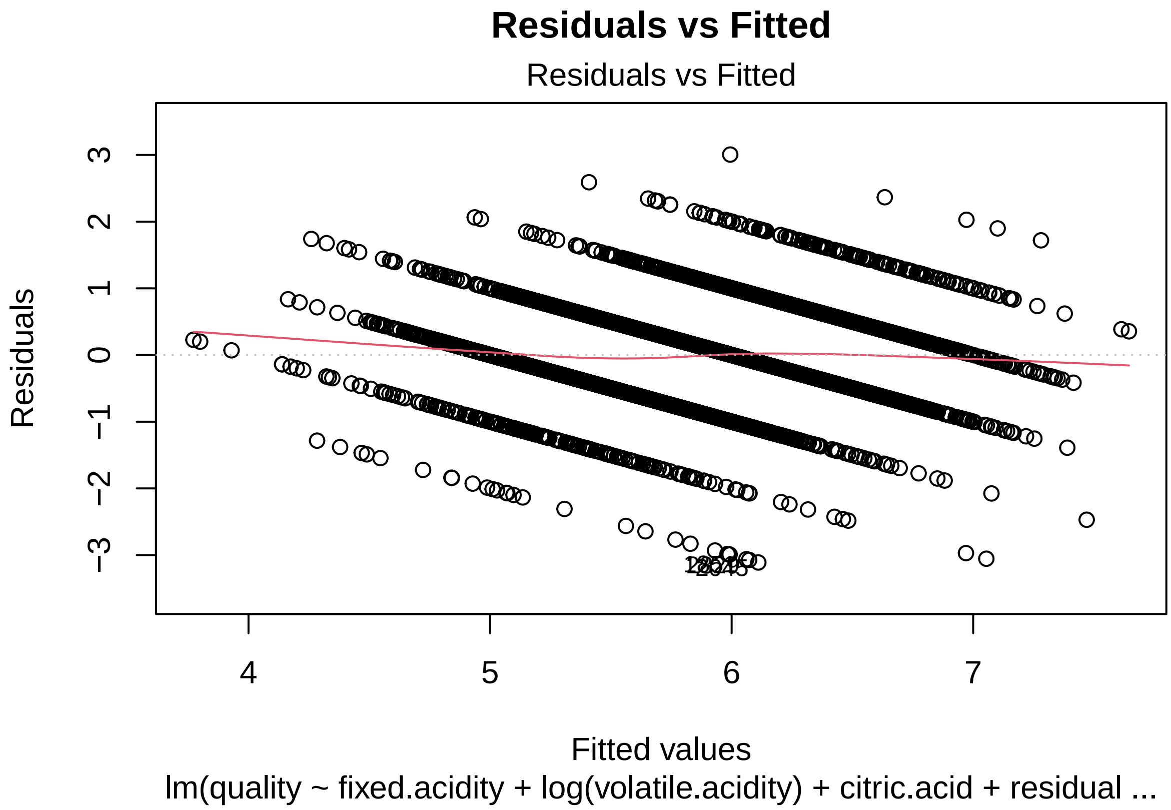
```r
library(corrplot)

numeric_data = wine_data %>%
  select(-type, -quality) # Considers numeric predictors only
cor_matrix = cor(numeric_data, use = "complete.obs")
par(mar = c(0, 0, 5, 0))
corrplot(cor_matrix,
         method = "color",
         addCoef.col = "black",
         tl.cex = 0.9,
         tl.col = "black",
         tl.srt = 45,
         number.cex = 0.8,
         main = "Correlation Heatmap",
         cl.cex = 0.8,
         cl.ratio = 0.2,
         mar = c(0, 0, 1, 0),
         col = colorRampPalette(c("navy", "white", "brown"))(200)
)
```

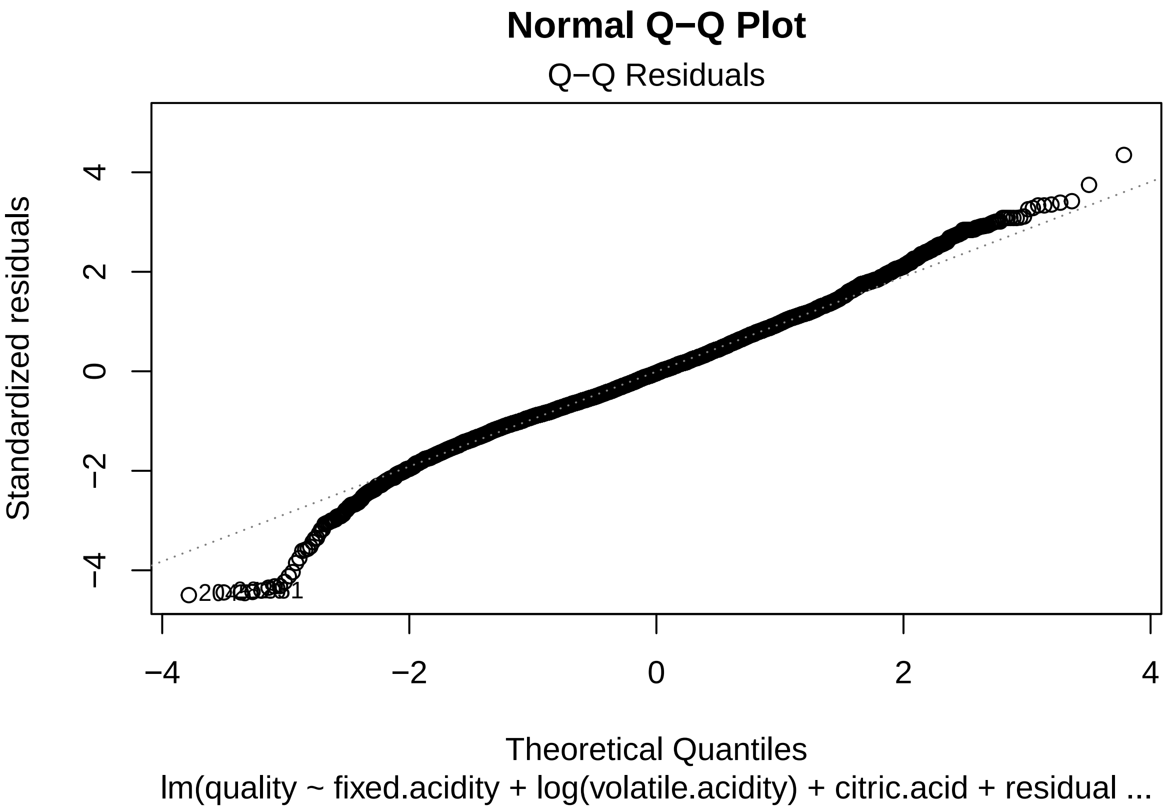


Linear Model Assumption Diagnostic Plots

```r
plot(model_both_aic, which = 1, main = "Residuals vs Fitted")
```



```
plot(model_both_aic, which = 2, main = "Normal Q-Q Plot")
```



Discussion

After considering several different models, our best model was…

TODO: (Describe the top model.)

This model achieved…

TODO: (Describe the performance of this model, using relevant values pulled from the model summaries.)

The steps and techniques we followed to arrive at this model included building a preliminary full additive model, transforming skewed predictors, experimenting with polynomial terms and interactions, and eliminating non-significant predictors. With each new candidate model, we analyzed how it compared to its predecessor models to assess whether we were headed in a productive direction.

Importantly, we did a final check on our final model to ensure that it obeyed the linear model assumptions.