Red Wine

Red Team

2023-12-5

```
#install.packages("tidyverse")
#install.packages("caret")
#install.packages("randomForest")
#install.packages("corrplot")
#install.packages("pROC")


library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
#library(caret)
#library(randomForest)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
options(warn=-1)
wines_red <- read.csv("wineQualityReds.csv", header=TRUE)

wines_red$r_quality <- ifelse(wines_red$quality < 6, 0, ifelse(wines_red$quality > 6, 1, NA))
wines_red$good_quality <- wines_red$r_quality == 1

summary(wines_red)
```

```
##        id          fix_acidity     vol_acidity      citric_acid
```

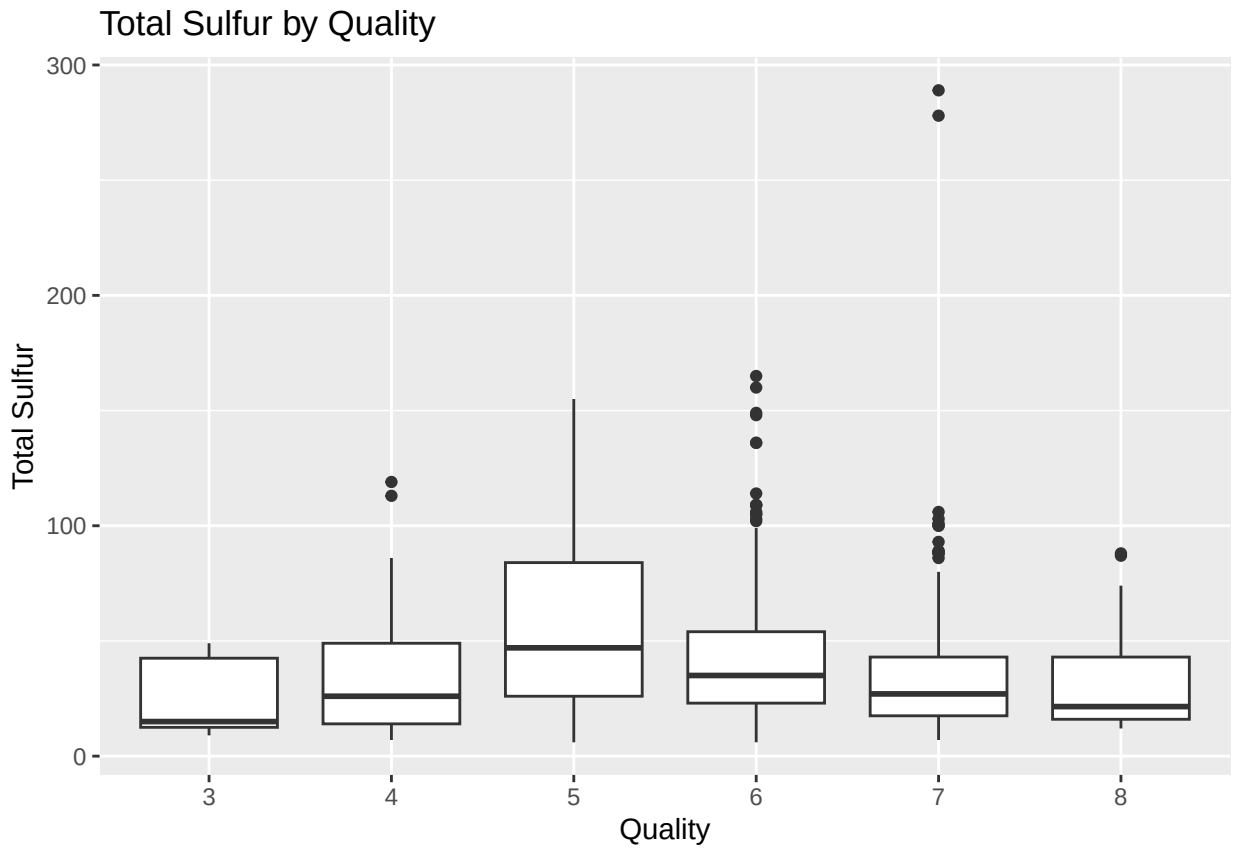
```
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5284   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##
##      sugar          chlorides        free_sulfur     total_sulfur
##  Min.   : 0.900   Min.   :0.01000   Min.   : 1.00   Min.   :  6.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00   1st Qu.: 22.00
##  Median : 2.200   Median :0.08000   Median :14.00   Median : 38.00
##  Mean   : 2.539   Mean   :0.08787   Mean   :15.87   Mean   : 46.47
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00   3rd Qu.: 62.00
##  Max.   :15.500   Max.   :0.61000   Max.   :72.00   Max.   :289.00
##
##     density             pH          sulphates         alcohol
##  Min.   :0.9900   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:1.0000   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :1.0000   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9985   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:1.0000   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0000   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##
##     quality        r_quality      good_quality
##  Min.   :3.000   Min.   :0.0000   Mode :logical
##  1st Qu.:5.000   1st Qu.:0.0000   FALSE:744
##  Median :6.000   Median :0.0000   TRUE :217
##  Mean   :5.636   Mean   :0.2258   NA's :638
##  3rd Qu.:6.000   3rd Qu.:0.0000
##  Max.   :8.000   Max.   :1.0000
##                  NA's   :638
```

```
cor(wines_red)
```

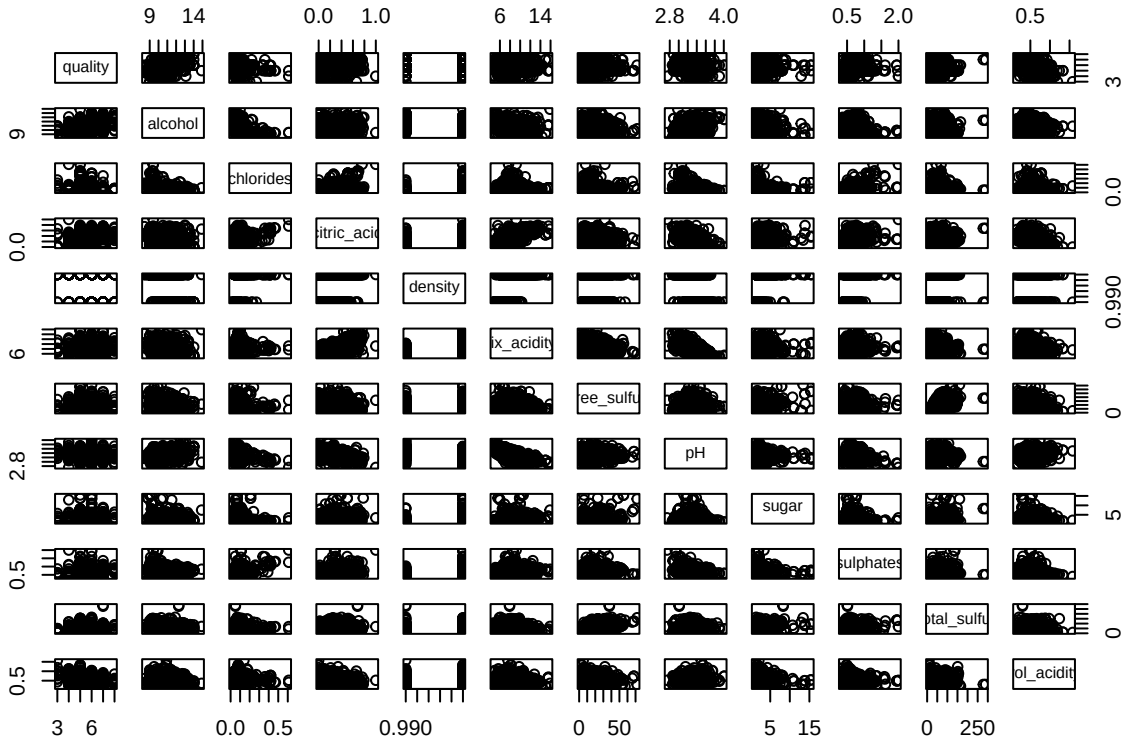
```
##                       id fix_acidity  vol_acidity citric_acid        sugar
## id            1.000000000 -0.26848392 -0.009309731 -0.15355136 -0.031260835
## fix_acidity  -0.268483920  1.00000000 -0.255729479  0.67170343  0.114776724
## vol_acidity  -0.009309731 -0.25572948  1.000000000 -0.55226229  0.002221340
## citric_acid  -0.153551355  0.67170343 -0.552262287  1.00000000  0.143577162
## sugar        -0.031260835  0.11477672  0.002221340  0.14357716  1.000000000
## chlorides    -0.120568197  0.09351597  0.062763615  0.20038577  0.052162455
## free_sulfur   0.090479643 -0.15379419 -0.011000729 -0.06097813  0.187048995
## total_sulfur -0.117849669 -0.11318144  0.076128109  0.03553302  0.203027882
## density      -0.279316525  0.34262182  0.071211144  0.13243017  0.130322333
## pH            0.136005328 -0.68297819  0.234503245 -0.54190414 -0.085652422
## sulphates    -0.125306999  0.18300566 -0.260781896  0.31277004  0.005527121
## alcohol       0.245121139 -0.06167907 -0.202300984  0.10989532  0.042078806
## quality       0.066452608  0.12405165 -0.390509029  0.22637251  0.013731637
## r_quality              NA          NA           NA          NA           NA
## good_quality           NA          NA           NA          NA           NA
##                  chlorides  free_sulfur total_sulfur      density           pH
## id           -0.120568197  0.090479643  -0.11784967 -0.279316525  0.13600533
## fix_acidity   0.093515966 -0.153794193  -0.11318144  0.342621816 -0.68297819
## vol_acidity   0.062763615 -0.011000729   0.07612811  0.071211144  0.23450324
## citric_acid   0.200385774 -0.060978129   0.03553302  0.132430175 -0.54190414
```

```
## sugar         0.052162455  0.187048995   0.20302788  0.130322333 -0.08565242
## chlorides     1.000000000  0.002842304   0.04504124  0.170939059 -0.26325813
## free_sulfur   0.002842304  1.000000000   0.66766645 -0.008452892  0.07037750
## total_sulfur  0.045041242  0.667666450   1.00000000  0.074662064 -0.06649456
## density       0.170939059 -0.008452892   0.07466206  1.000000000 -0.23177121
## pH           -0.263258128  0.070377499  -0.06649456 -0.231771209  1.00000000
## sulphates     0.368424679  0.051657572   0.04294684  0.078331510 -0.19664760
## alcohol      -0.220173780 -0.069400617  -0.20564269 -0.520314592  0.20563509
## quality      -0.131756039 -0.050656057  -0.18510029 -0.210725598 -0.05773139
## r_quality              NA           NA           NA           NA          NA
## good_quality           NA           NA           NA           NA          NA
##                  sulphates     alcohol     quality r_quality good_quality
## id           -0.125306999  0.24512114  0.06645261        NA           NA
## fix_acidity   0.183005664 -0.06167907  0.12405165        NA           NA
## vol_acidity  -0.260781896 -0.20230098 -0.39050903        NA           NA
## citric_acid   0.312770044  0.10989532  0.22637251        NA           NA
## sugar         0.005527121  0.04207881  0.01373164        NA           NA
## chlorides     0.368424679 -0.22017378 -0.13175604        NA           NA
## free_sulfur   0.051657572 -0.06940062 -0.05065606        NA           NA
## total_sulfur  0.042946836 -0.20564269 -0.18510029        NA           NA
## density       0.078331510 -0.52031459 -0.21072560        NA           NA
## pH           -0.196647602  0.20563509 -0.05773139        NA           NA
## sulphates     1.000000000  0.09359460  0.25139708        NA           NA
## alcohol       0.093594599  1.00000000  0.47616445        NA           NA
## quality       0.251397079  0.47616445  1.00000000        NA           NA
## r_quality              NA          NA          NA         1           NA
## good_quality           NA          NA          NA        NA            1
```

```
ggplot(wines_red, aes(x = factor(quality), y = total_sulfur)) +
  geom_boxplot() +
  labs(x = "Quality", y = "Total Sulfur") +
  ggtitle("Total Sulfur by Quality")
```

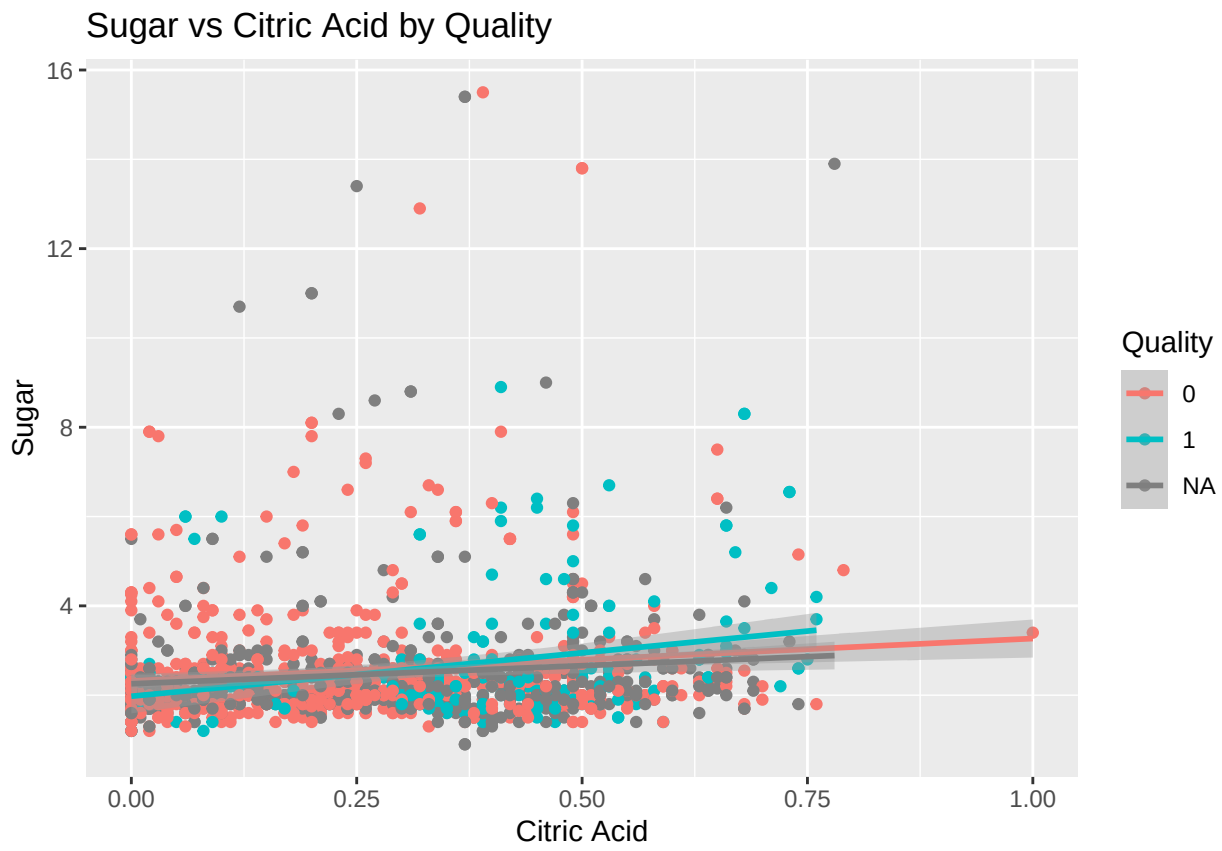



```
pairs(~ quality + alcohol + chlorides + citric_acid + density + fix_acidity +
        free_sulfur + pH + sugar + sulphates + total_sulfur + vol_acidity,
      data = wines_red, diagonal = "histogram")
```




```
ggplot(wines_red, aes(x = citric_acid, y = sugar, color = factor(r_quality))) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Citric Acid", y = "Sugar", color = "Quality") +
  ggtitle("Sugar vs Citric Acid by Quality")
```

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




```
wines_red$alcohol2 <- wines_red$alcohol^2

# Linear Regression
lm_quality_alcohol <- lm(quality ~ alcohol, data = wines_red)
summary(lm_quality_alcohol)
```

```
##
## Call:
## lm(formula = quality ~ alcohol, data = wines_red)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8442 -0.4112 -0.1690  0.5166  2.5888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.87502    0.17471   10.73   <2e-16 ***
## alcohol      0.36084    0.01668   21.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7104 on 1597 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.2262
## F-statistic: 468.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

```
# Multiple Regression
lm_multiple <- lm(quality ~ citric_acid + pH + fix_acidity, data = wines_red)
```

```
summary(lm_multiple)
```

```
##
## Call:
## lm(formula = quality ~ citric_acid + pH + fix_acidity, data = wines_red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.05431 -0.57801  0.09158  0.50784  2.55048
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.696748   0.673294   5.491 4.66e-08 ***
## citric_acid 1.139210   0.137686   8.274 2.70e-16 ***
## pH          0.488645   0.176198   2.773  0.00561 **
## fix_acidity 0.001516   0.017725   0.086  0.93183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7849 on 1595 degrees of freedom
## Multiple R-squared:  0.05722,    Adjusted R-squared:  0.05545
## F-statistic: 32.27 on 3 and 1595 DF,  p-value: < 2.2e-16
```

```
# Polynomial Regression
lm_poly <- lm(quality ~ alcohol + alcohol2, data = wines_red)
summary(lm_poly)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + alcohol2, data = wines_red)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8716 -0.3884 -0.1642  0.5157  2.5852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.53519    1.52764  -0.350  0.72613
## alcohol      0.80894    0.28264   2.862  0.00426 **
## alcohol2    -0.02059    0.01297  -1.588  0.11245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.71 on 1596 degrees of freedom
## Multiple R-squared:  0.228,  Adjusted R-squared:  0.227
## F-statistic: 235.6 on 2 and 1596 DF,  p-value: < 2.2e-16
```

```
# Model Selection
lm_model <- step(lm(quality ~ ., data = wines_red), direction = "both")
```

```
## Start:  AIC=-2213.28
## quality ~ id + fix_acidity + vol_acidity + citric_acid + sugar +
##     chlorides + free_sulfur + total_sulfur + density + pH + sulphates +
##     alcohol + r_quality + good_quality + alcohol2
##
```

```
##
## Step:  AIC=-2213.28
## quality ~ id + fix_acidity + vol_acidity + citric_acid + sugar +
##     chlorides + free_sulfur + total_sulfur + density + pH + sulphates +
##     alcohol + r_quality + alcohol2
##
##                Df Sum of Sq    RSS      AIC
## - sulphates     1      0.00  93.10 -2215.28
## - free_sulfur   1      0.05  93.15 -2214.78
## - citric_acid   1      0.07  93.17 -2214.56
## <none>                       93.10 -2213.28
## - id            1      0.26  93.36 -2212.56
## - total_sulfur  1      0.34  93.44 -2211.76
## - chlorides     1      0.35  93.45 -2211.70
## - sugar         1      0.37  93.47 -2211.45
## - alcohol       1      0.42  93.52 -2210.94
## - alcohol2      1      0.49  93.59 -2210.21
## - fix_acidity   1      0.50  93.60 -2210.12
## - density       1      0.51  93.61 -2210.01
## - pH            1      1.54  94.64 -2199.50
## - vol_acidity   1      2.73  95.82 -2187.55
## - r_quality     1    361.37 454.47  -691.63
##
## Step:  AIC=-2215.28
## quality ~ id + fix_acidity + vol_acidity + citric_acid + sugar +
##     chlorides + free_sulfur + total_sulfur + density + pH + alcohol +
##     r_quality + alcohol2
##
##                Df Sum of Sq    RSS      AIC
## - free_sulfur   1      0.05  93.15 -2216.77
## - citric_acid   1      0.07  93.17 -2216.56
## <none>                       93.10 -2215.28
## - id            1      0.27  93.37 -2214.50
## - total_sulfur  1      0.34  93.44 -2213.74
## - sugar         1      0.38  93.47 -2213.41
## + sulphates     1      0.00  93.10 -2213.28
## - alcohol       1      0.42  93.52 -2212.94
## - chlorides     1      0.43  93.53 -2212.82
## - alcohol2      1      0.49  93.59 -2212.21
## - fix_acidity   1      0.50  93.60 -2212.10
## - density       1      0.53  93.62 -2211.87
## - pH            1      1.54  94.64 -2201.48
## - vol_acidity   1      2.80  95.90 -2188.84
## - r_quality     1    377.56 470.66  -659.99
##
## Step:  AIC=-2216.77
## quality ~ id + fix_acidity + vol_acidity + citric_acid + sugar +
##     chlorides + total_sulfur + density + pH + alcohol + r_quality +
##     alcohol2
##
##                Df Sum of Sq    RSS     AIC
## - citric_acid   1      0.09  93.23 -2217.9
## <none>                       93.15 -2216.8
## - id            1      0.23  93.38 -2216.4
```

```
## + free_sulfur   1      0.05  93.10 -2215.3
## - sugar         1      0.35  93.50 -2215.1
## + sulphates     1      0.00  93.15 -2214.8
## - chlorides     1      0.41  93.56 -2214.6
## - alcohol       1      0.41  93.56 -2214.5
## - fix_acidity   1      0.48  93.62 -2213.9
## - alcohol2      1      0.48  93.63 -2213.8
## - density       1      0.52  93.67 -2213.4
## - total_sulfur  1      1.02  94.17 -2208.3
## - pH            1      1.50  94.65 -2203.4
## - vol_acidity   1      2.96  96.11 -2188.7
## - r_quality     1    382.14 475.28  -652.6
##
## Step:  AIC=-2217.88
## quality ~ id + fix_acidity + vol_acidity + sugar + chlorides +
##     total_sulfur + density + pH + alcohol + r_quality + alcohol2
##
##                Df Sum of Sq    RSS      AIC
## <none>                      93.23 -2217.88
## - id            1      0.22  93.45 -2217.62
## + citric_acid   1      0.09  93.15 -2216.77
## + free_sulfur   1      0.07  93.17 -2216.56
## - sugar         1      0.37  93.61 -2216.02
## + sulphates     1      0.00  93.23 -2215.88
## - alcohol       1      0.43  93.67 -2215.41
## - alcohol2      1      0.50  93.73 -2214.75
## - density       1      0.52  93.76 -2214.50
## - chlorides     1      0.61  93.84 -2213.62
## - total_sulfur  1      0.94  94.17 -2210.28
## - fix_acidity   1      0.97  94.20 -2209.96
## - pH            1      1.46  94.70 -2204.91
## - vol_acidity   1      3.36  96.60 -2185.83
## - r_quality     1    382.12 475.36  -654.46
```

```
summary(lm_model)
```

```
##
## Call:
## lm(formula = quality ~ id + fix_acidity + vol_acidity + sugar +
##     chlorides + total_sulfur + density + pH + alcohol + r_quality +
##     alcohol2, data = wines_red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87092 -0.04362  0.04654  0.13316  0.99253
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.977e-01  3.943e+00  -0.177 0.859596
## id           -3.532e-05  2.364e-05  -1.494 0.135551
## fix_acidity  -2.871e-02  9.147e-03  -3.138 0.001752 **
## vol_acidity  -3.572e-01  6.106e-02  -5.850 6.75e-09 ***
## sugar        -1.518e-02  7.769e-03  -1.953 0.051056 .
## chlorides    -5.270e-01  2.116e-01  -2.490 0.012936 *
## total_sulfur  9.435e-04  3.057e-04   3.086 0.002085 **
```

```
## density       9.213e+00  3.990e+00   2.309 0.021167 *
## pH           -3.829e-01  9.920e-02  -3.860 0.000121 ***
## alcohol      -3.617e-01  1.719e-01  -2.104 0.035683 *
## r_quality     2.126e+00  3.409e-02  62.366  < 2e-16 ***
## alcohol2      1.775e-02  7.877e-03   2.253 0.024492 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3134 on 949 degrees of freedom
##   (638 observations deleted due to missingness)
## Multiple R-squared:  0.8966, Adjusted R-squared:  0.8954
## F-statistic: 747.9 on 11 and 949 DF,  p-value: < 2.2e-16
```

```
# ROC Curve
roc_model <- roc(wines_red$good_quality, wines_red$alcohol)
```

```
## Setting levels: control = FALSE, case = TRUE
```

```
## Setting direction: controls < cases
```

```
plot(roc_model)
```