Red Wine Association: Onima Shah

2025-05-01

Setting up R file

```r
#Downloading necessary packages
#install.packages("tidyverse")
#install.packages("corrplot")
#install.packages("regclass")

#Loading packages
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────────────
tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ─────────────────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```r
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
library(regclass)
```

```
## Loading required package: bestglm
## Loading required package: leaps
## Loading required package: VGAM
## Loading required package: stats4
## Loading required package: splines
## Loading required package: rpart
## Loading required package: randomForest
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
##
## Important regclass change from 1.3:
## All functions that had a . in the name now have an _
## all.correlations -> all_correlations, cor.demo -> cor_demo, etc.
```

Loading file and checking dataset

```r
#Loading csv file
wine <- read.csv("winequality-red.csv")

#Checking dataset
head(wine)
```

```
##   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
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```r
str(wine)
```

```
## 'data.frame':    1599 obs. of  12 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 ...
```

```
summary(wine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar  
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500  
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density      
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901  
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956  
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968  
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967  
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978  
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037  
##        pH          sulphates         alcohol         quality     
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000  
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000  
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000  
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636  
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000  
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

```
names(wine)
```

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"         
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide" 
##  [7] "total.sulfur.dioxide" "density"              "pH"                  
## [10] "sulphates"            "alcohol"              "quality"
```

```
#Checking for missing values
sum(is.na(wine))
```

```
## [1] 0
```

```
colSums(is.na(wine))
```

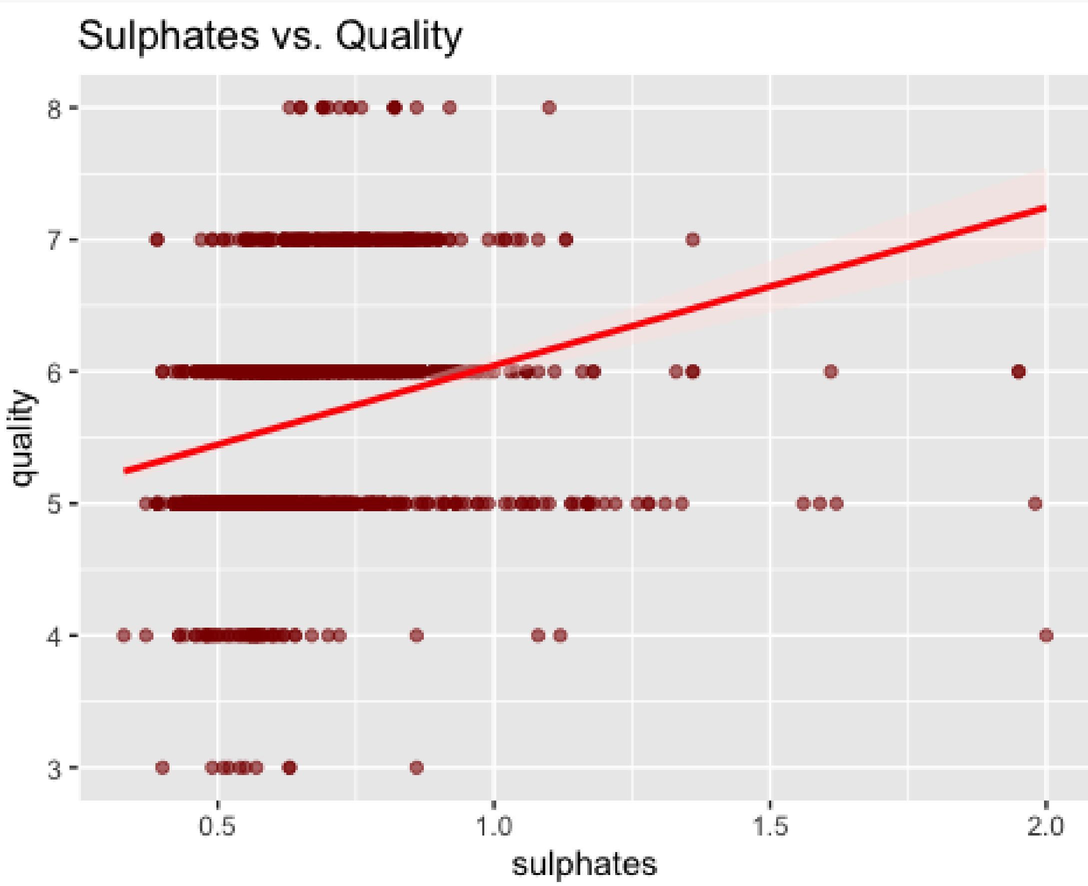
```
##        fixed.acidity     volatile.acidity          citric.acid 
##                    0                    0                    0 
##       residual.sugar            chlorides  free.sulfur.dioxide 
##                    0                    0                    0 
## total.sulfur.dioxide              density                   pH 
##                    0                    0                    0
```

```
##            sulphates              alcohol              quality
##                    0                    0                    0

#Checking for duplicates
sum(duplicated(wine))

## [1] 240

#Should we remove the duplicates??
```

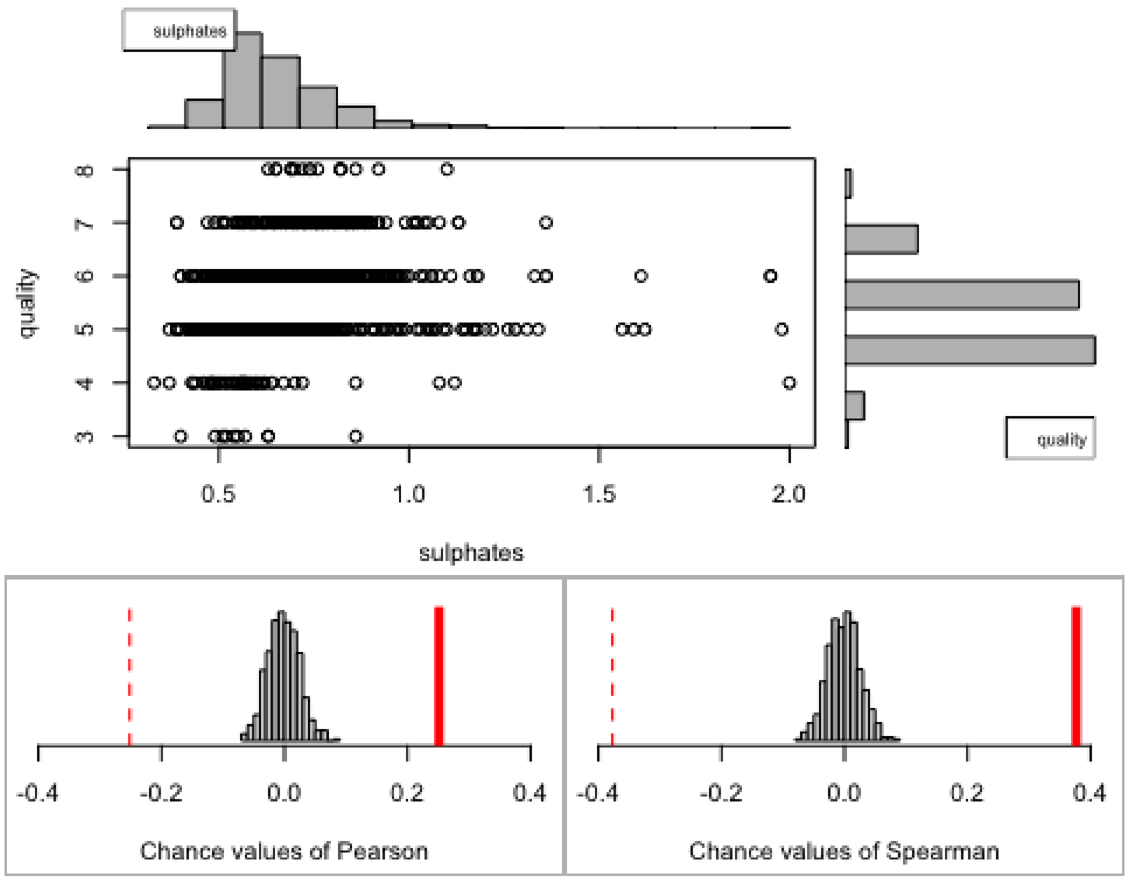
Variables for association analysis: y = quality; x = sulphates, residual.sugar

Sulphates

```
#Graph
ggplot(wine, aes(x = sulphates, y = quality)) +
  geom_point(color = "darkred", alpha = 0.6) +
  geom_smooth(method = "lm", color = "red", fill = "mistyrose", se = TRUE) +
  ggtitle("Sulphates vs. Quality")

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



```r
#Checking for association
associate(quality~sulphates, data = wine)

## Association between sulphates (numerical) and  quality (numerical)
##  using 1599 complete cases
```




```
## Permutation procedure:
##                                  Value Estimated p-value
## Pearson's r                  0.2513971                 0
## Spearman's rank correlation 0.3770602                 0
## With 500 permutations, we are 95% confident that:
##  the p-value of Pearson's correlation (r) is between 0 and 0.007
##  the p-value of Spearman's rank correlation is between 0 and 0.007
## Note:  If 0.05 is in this range, increase the permutations= argument.
##
##
##
## Advice: If stream of points is well described by an ellipse, use Pearson's
r.
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation
## or try logs, e.g. associate( log10(y)~log10(x) )
```

```r
#Checking for correlation against all variables
all_correlations(wine, interest = "sulphates", sorted = "magnitude")
```

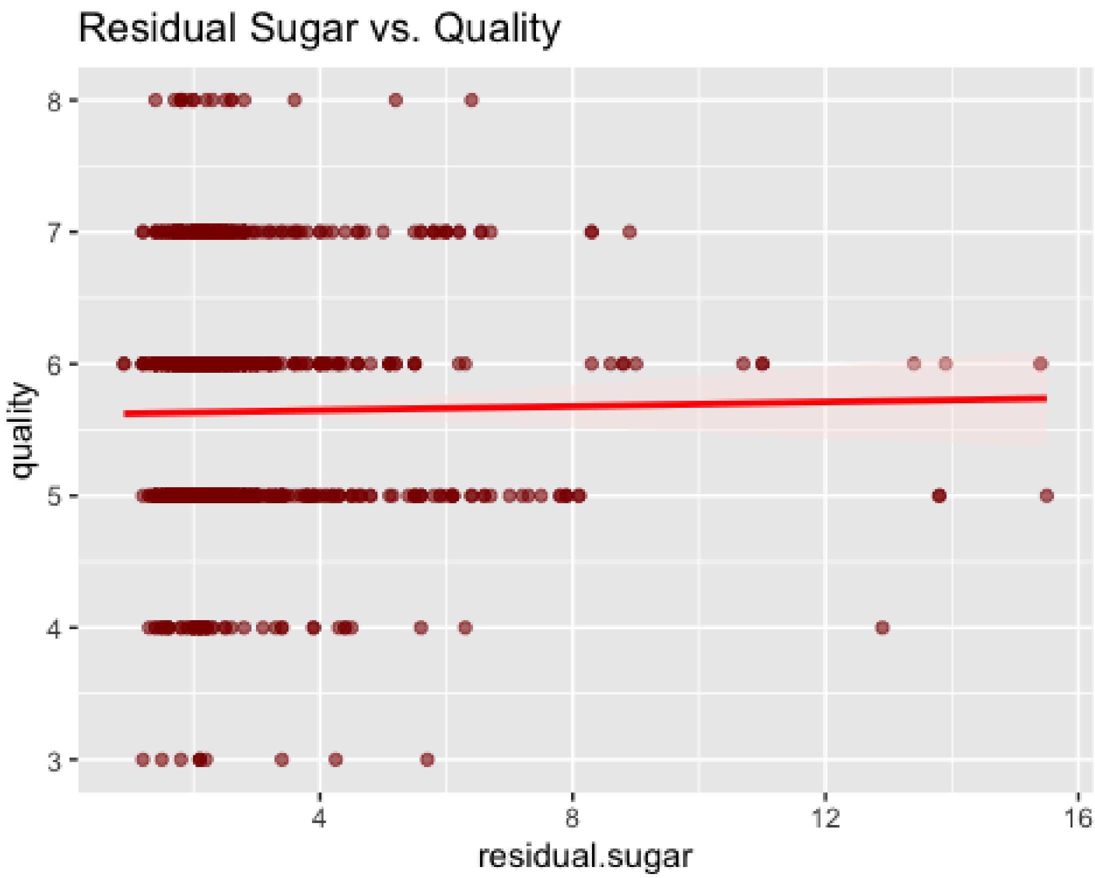
```
##                    var1      var2  correlation         pval
## 1             chlorides sulphates  0.371260481 1.986310e-53
## 2            citric.acid sulphates  0.312770044 1.265262e-37
## 3       volatile.acidity sulphates -0.260986685 2.606926e-26
## 4              sulphates   quality  0.251397079 1.802088e-24
## 5                     pH sulphates -0.196647602 2.106734e-15
## 6          fixed.acidity sulphates  0.183005664 1.648652e-13
## 7                density sulphates  0.148506412 2.418474e-09
## 8              sulphates   alcohol  0.093594750 1.783053e-04
## 9    free.sulfur.dioxide sulphates  0.051657572 3.888321e-02
## 10 total.sulfur.dioxide sulphates  0.042946836 8.601835e-02
## 11        residual.sugar sulphates  0.005527121 8.252134e-01
```

Analysis: According to the Pearson correlation (r = 0.25), there is a moderate positive linear relationship between the quality of wine and sulphates. As sulphates increase, the quality of wine tends to increase slightly, but it is not a very strong correlation. P-values are lower than 0.007 for both quality and sulphates, so the correlations are significant between the two and sulphates are meaningfully associated with the quality of wine, even if the linear relationship is weaker. Sulphates appear to correlate strongly with other variables such as citric acid, suggesting that sulphates is more present in the chemical composition of the wine rather than the quality.

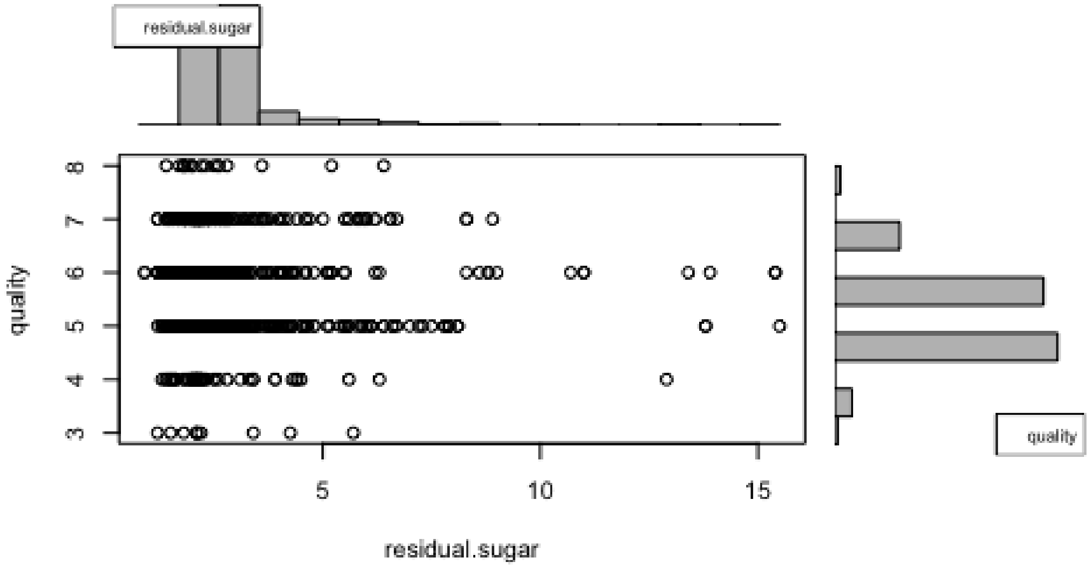
Residual Sugar

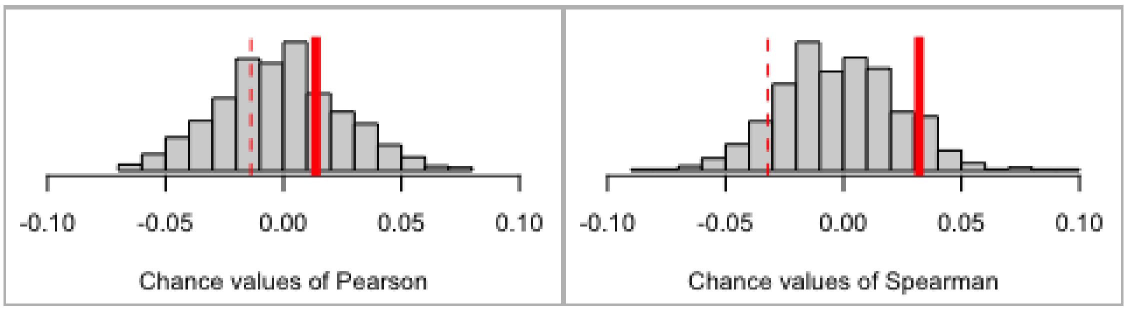
```r
#Graph
ggplot(wine, aes(x = residual.sugar, y = quality)) +
  geom_point(color = "darkred", alpha = 0.6) +
  geom_smooth(method = "lm", color = "red", fill = "mistyrose", se = TRUE) +
  ggtitle("Residual Sugar vs. Quality")
```

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




```
#Checking for association
associate(quality~residual.sugar, data = wine)

## Association between residual.sugar (numerical) and  quality (numerical)
##  using 1599 complete cases
```






```
## Permutation procedure:
##                                   Value Estimated p-value
## Pearson's r                  0.01373164             0.600
## Spearman's rank correlation 0.03204817             0.202
## With 500 permutations, we are 95% confident that:
##  the p-value of Pearson's correlation (r) is between 0.556 and 0.643
##  the p-value of Spearman's rank correlation is between 0.168 and 0.24
## Note:  If 0.05 is in this range, increase the permutations= argument.
##
##
##
## Advice: If stream of points is well described by an ellipse, use Pearson's
r.
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation
## or try logs, e.g. associate( log10(y)~log10(x) )
```

```
#Checking for correlation against all variables
all_correlations(wine, interest = "residual.sugar", sorted = "magnitude")
```

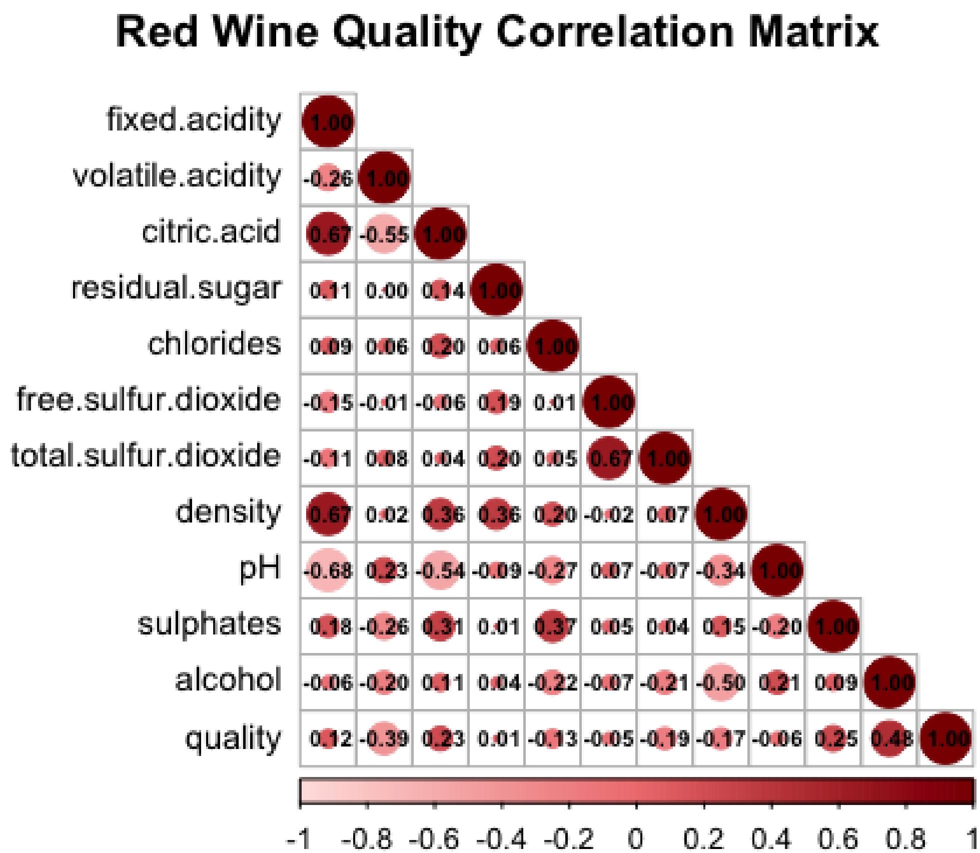
```
##                var1                 var2  correlation         pval
## 1    residual.sugar              density  0.355283371 9.013042e-49
## 2    residual.sugar total.sulfur.dioxide  0.203027882 2.449285e-16
## 3    residual.sugar  free.sulfur.dioxide  0.187048995 4.684735e-14
```

```
## 4       citric.acid       residual.sugar  0.143577162 8.083723e-09
## 5     fixed.acidity       residual.sugar  0.114776724 4.199465e-06
## 6    residual.sugar                   pH -0.085652422 6.065915e-04
## 7    residual.sugar            chlorides  0.055609535 2.617079e-02
## 8    residual.sugar              alcohol  0.042075437 9.258425e-02
## 9    residual.sugar              quality  0.013731637 5.832180e-01
## 10   residual.sugar            sulphates  0.005527121 8.252134e-01
## 11 volatile.acidity       residual.sugar  0.001917882 9.389168e-01
```

Analysis: According to the Pearson correlation (r = 0.0137), there is nearly no linear relationship between the quality of wine and residual sugar. P-values are higher than 0.05, so the correlations are not statistically significant. We fail to reject the null hypothesis, meaning residual sugar is not associated with the quality of wine. In the graph, there is no visible pattern and the data points are concentrated at low sugar levels, with no clear slope. Residual sugar has some correlation with density, as sugar adds weight to the wine. Overall, there is no meaningful correlation between the quality of wine and residual sugar.

Correlations Visualization

```
cor_matrix <- cor(wine)
wine_colors <- colorRampPalette(c("mistyrose", "lightcoral", "darkred"))(200)
corrplot(cor_matrix, method = "circle", type = "lower", col = wine_colors,
tl.col = "black", tl.cex = 0.9, tl.pos = "l", cl.pos = "b", cl.cex = 0.8,
number.cex = 0.6, addCoef.col = "black", title = "Red Wine Quality
Correlation Matrix", mar = c(0, 0, 2, 0))
```



Analysis: This correlation matrix creates a visualization of the linear relationships between the quality of wine and other variables in the chemical properties. The darker red shows a stronger positive correlation, and the lighter colors show weaker correlations. Negative correlations are shown in lighter colors. The bigger circles show stronger correlation, whether it is positive or negative, and smaller circles show weak or no correlation.

Alcohol appears to have the strongest positive correlations with quality at 0.48, meanwhile volatile acidity shows a weak negative correlation with quality at -0.26. Residual sugar and chlorides show little or no correlation with quality. Citric acid and fixed acidity are strongly correlated with each other, and weakly correlated with pH. Residual sugar, chlorides, and sulphates have either little or no direct relationship with quality of wine.

Multi linear regression

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          2.8431068  0.2050732  13.864  < 2e-16 ***
alcohol              0.2953419  0.0160375  18.416  < 2e-16 ***
volatile.acidity    -1.2223102  0.1124774 -10.867  < 2e-16 ***
sulphates            0.7207881  0.1027039   7.018 3.32e-12 ***
citric.acid         -0.0427246  0.1035810  -0.412     0.68    
total.sulfur.dioxide -0.0022182  0.0005126  -4.327 1.60e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6552 on 1593 degrees of freedom
Multiple R-squared:  0.3439,    Adjusted R-squared:  0.3418
```

Analysis: Of the coefficients all are statistically significant except for citric acid using a standard 95% confidence interval ($p < .05$). From the significant coefficients two of them are positive and two are negative with alcohol level and sulfates both providing a positive impact though the model shows that sulphates have a higher impact on the wines quality. From this wine makers could conclude that focusing on a wine that emphasizes sulphates over one that does not would result in a better end product.

On the other hand, volatile acidity has a very negative impact on wine quality. The presence of excess citric acid causes an undesirable taste in the wine and because this coefficient is high it shows that it has a large impact on how wine is made and should be considered for wine makers Total sulfur dioxide value is low indicating that even though it is statistically significant the impact that it has is negligible. This shows that sulfur dioxide may be unnoticeable to people who are drinking the wine and is not something that may need to be considered for wine makers.

The R squared value for this model indicates that it explains an adequate amount of the variance but there may be other factors that contribute to a wines quality outside of what is in the dataset currently such as vineyard or vintage.