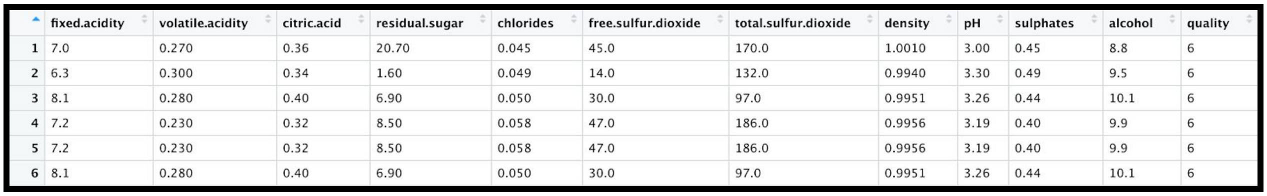
UCI Wine Analysis

Part 1:

I'll be analyzing a wine quality dataset from the University of California, Irvine's Machine Learning Repository. This is actually one of two datasets, which, respectively, are related to red and white variants of the Portuguese "Vinho Verde" wine, which consists of wine from the Minho region of Portugal. Portugal is currently one of the top exporters of wine in the world. The dataset contains the same variables(only for the white wine), where the inputs/independent avariables are physicochemical data, and sensory/output variables are also available (e.g. there is no data about grape type, brand, selling price, etc.). Below is a screenshot of the top 6 rows of the dataset.

```
>head(whitewinequality)
```



	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
1	7.0	0.270	0.36	20.70	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
2	6.3	0.300	0.34	1.60	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
3	8.1	0.280	0.40	6.90	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
4	7.2	0.230	0.32	8.50	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
5	7.2	0.230	0.32	8.50	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
6	8.1	0.280	0.40	6.90	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6

Part 2:

Research Question 1: Is there a significant difference in quality of wines with higher alcohol content (a higher alcohol by volume (ABV) percentage) compared to the population of wines?

The question asks to indicate if there's a difference between quality of wines with higher alcohol content as opposed to wines with lower alcohol content. Running a two sample t-test would be appropriate because I want to test the difference between two population means and see if they're equal or not. I split the dataset into the higher ABV% white wines, "greateralc", and the lesser, "lesseralc". I considered the alcohol content "lesser" if the ABV% was less than the average(~10.5%).

Ho: There is no significant difference in wine quality between lower ABV% wines and wines of a higher ABV%
Ha: Wines with a higher ABV% are greater in quality than wines with a lower ABV%.

```
> avgalc = mean(whitewinequality$alcohol)
> avgalc
[1] 10.51427
> greateralc = whitewinequality[which(whitewinequality$alcohol >= mean(whitewinequality$alcohol)),]
> lesseralac = whitewinequality[which(whitewinequality$alcohol < mean(whitewinequality$alcohol)),]
> t.test(x = greateralc$quality, y = lesseralc$quality, alternative = 'greater')
```

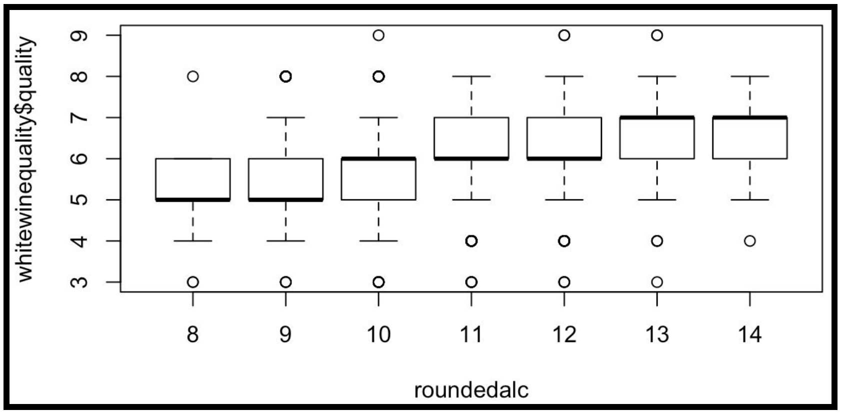
Welch Two Sample t-test

data: lesseralac$quality and greateralc$quality
t = -28.748, df = 4331.5, p-value < **2.2e-16**
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:

```
-Inf -0.6471552
sample estimates:
mean of x mean of y
 5.574771  6.261211
```

Since the p-value, 2.2e-16, is less than alpha(.05), we have significant evidence to reject Ho and indicate that white wines with a higher ABV% are greater in quality than white wines with a lower ABV%.

```
> roundedalc = round(whitewinequality$alcohol)
# I used rounded values for a clearer boxplot
> boxplot(whitewinequality$quality~roundedalc)
```



You notice that for the most part, a trend exists, and a larger alcohol percentage tends to receive a better rating on the quality index.

<u>Research Question 2: Is there a significant difference in quality of wines with a large total sulfur dioxide content versus wines with a smaller total sulfure dioxied content?</u>

The question asks to indicate if there's a difference between quality of wines with higher total sulfure dioxide content as opposed to wines with lower total sulfure dioxide content. Running a two sample t-test would be appropriate because I want to test the difference between two population means and see if they're equal or not. I split the dataset into wines with higher TSD content, "greaterTSD", and the lesser, "lesserTSD". I considered the TSD content "lesser" if it was less than the average(~138.3607).

Ho: There is no significant difference in wine quality between lower total sulfur dioxide wines and wines of higher sulfur dioxide content.
Ha: Wines with a smaller sulfure dioxide content are greater in quality than wines with a larger sulfur dioxide content.

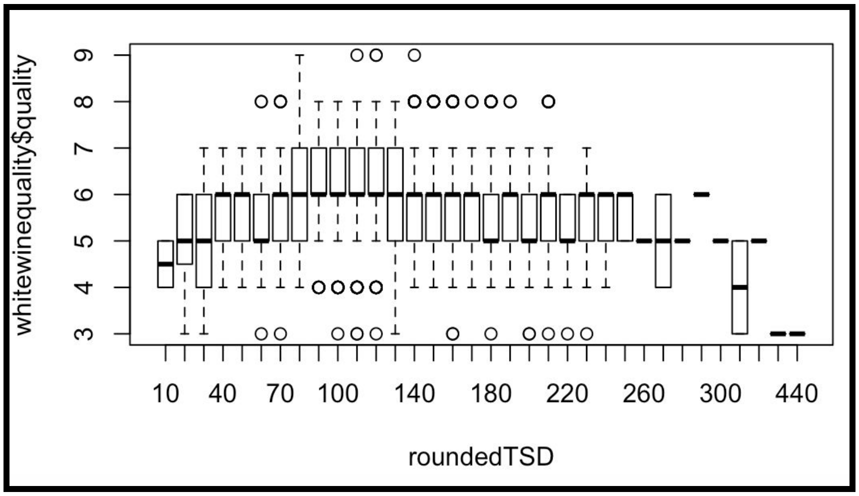
```
> avgTSD = mean(whitewinequality$total.sulfur.dioxide)
> avgTSD
[1] 138.3607
```

```
> greaterTSD = whitewinequality[which(whitewinequality$total.sulfur.dioxide >=
mean(whitewinequality$total.sulfur.dioxide)),]
> lesserTSD = whitewinequality[which(whitewinequality$total.sulfur.dioxide <
mean(whitewinequality$total.sulfur.dioxide)),]
> t.test(x = lesserTSD$quality, y = greaterTSD$quality, alternative = 'greater')
```

```
Welch Two Sample t-test
data:  lesserTSD$quality and greaterTSD$quality
t = 12.558, df = 4889.1, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2628838 0.3601456
sample estimates:
mean of x mean of y
 6.022918  5.711404
```

Since the p-value, 2.2e-16, is less than alpha(.05), we have significant evidence to reject Ho and indicate that wines with a smaller total sulfure dioxide content are greater in quality than wines with a larger total sulfur dioxide content.

```
> roundedTSD = round(whitewinequality$total.sulfur.dioxide, -1)
> boxplot(whitewinequality$quality~roundedTSD)
```



You can notice that wine with a total sulfur dioxide from around 90 to 130 has a higher quality. Everywhere else it's pretty stagnant and the quality isn't very effective. Observing a negative correlation, as the total sulfure dioxide starts getting very large, the quality decreases alot.

```
> cor(whitewinequality$quality,whitewinequality$total.sulfur.dioxide)
[1] -0.1747372
```

This means total sulfur dioxide and quality have a weak inverse relationship.

Research Question 3: Does wine quality differ across pH values ?

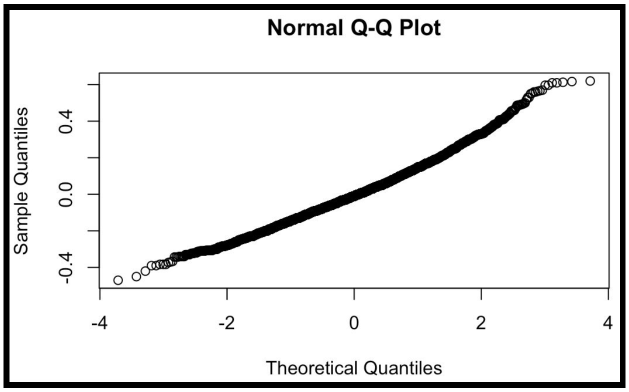
This question wants us to see if statistical differences among the means of two or more groups exist. So using ANOVA is our best bet. For this question we would use

Ho: There is no significant difference in white wine quality between different pH values of the wines.
Ha: There is a significant difference in white wine quality between different pH values of the wines.

```
> summary(aov(formula = pH~quality, data = whitewinequality))
             Df Sum Sq Mean Sq F value   Pr(>F)
quality       1    1.1  1.1038   48.88   3.08e-12 ***
Residuals  4896  110.5  0.0226
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

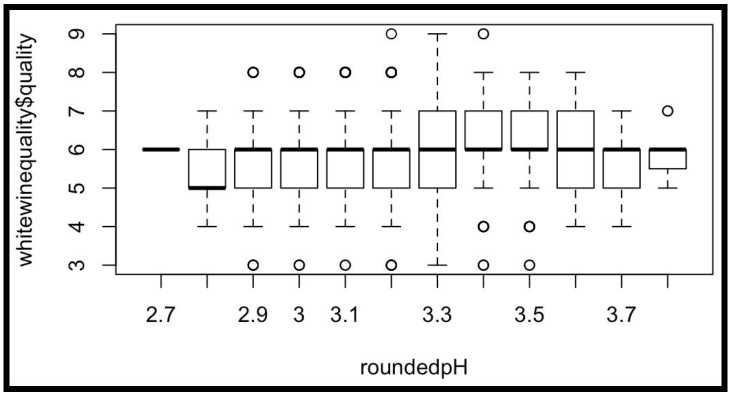
Since the p-value, 3.08e-12, is less than alpha(.05), we have sufficient evidence to reject Ho, and accept that there is a significant difference in white wine quality between different pH values of the wines. As far as assumptions for a one way anova test, to prove normality, you can graph the residuals on a QQplot and observe a linearity, which you do.

```
> wineanova = aov(formula = pH~quality, data = whitewinequality)
> qqnorm(wineanova$residuals)
```

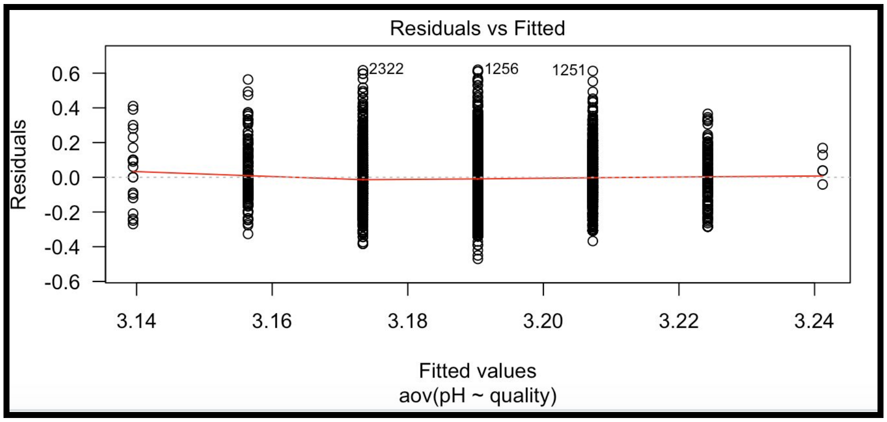


Based on the boxplot below you can see the variance and spread of each group is similar. This way, we can prove homogeneity of variance. Also, plotting the Residuals vs. Fitted Values plot, we can observe that the variances are approximately homogeneous since the residuals are distributed approximately equally above and below zero.

```
> roundedpH = round(whitewinequality$pH,1)
> boxplot(whitewinequality$quality~roundedpH)
```

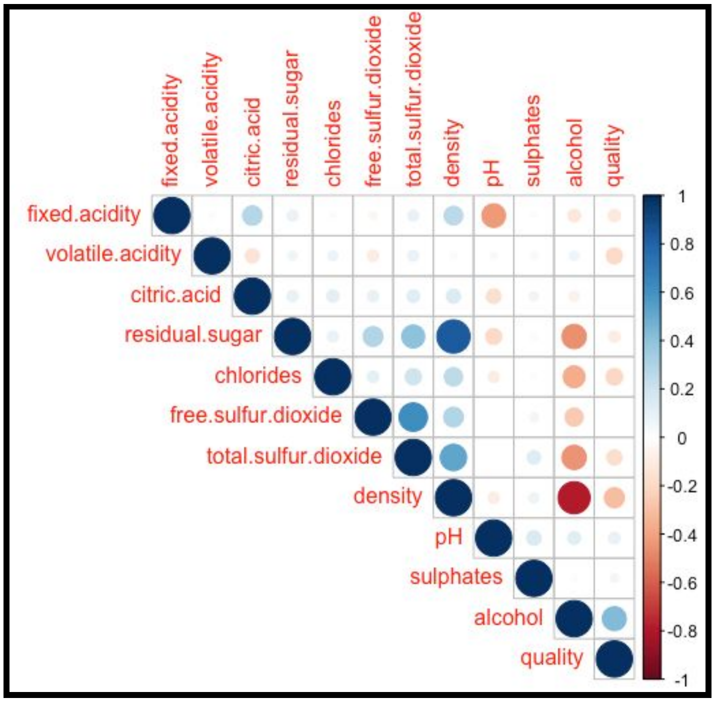


```
> plot(wineanova,1,las=1)
```



Let’s model the correlation relationships between all ingredients

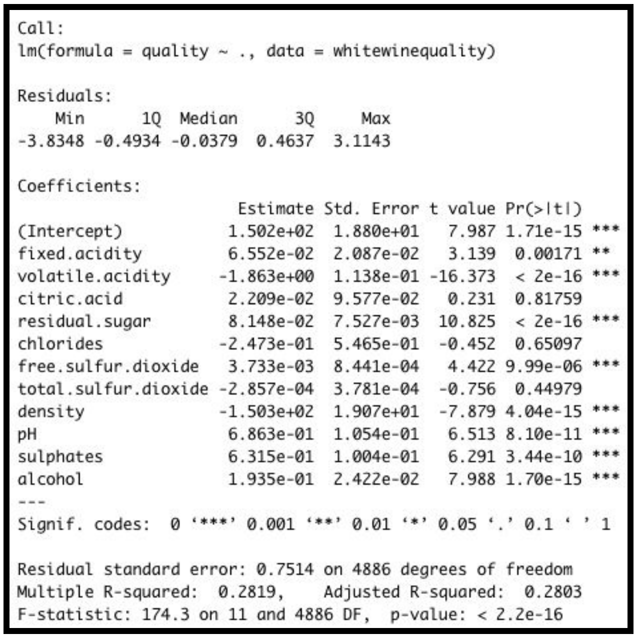
```
> library(corrplot)

> cor.table = cor(whitewinequality)
> corrplot(cor.table)
> corrplot(cor.table, type = 'upper')
```



We’ll now use a multivariable linear regression to:

- Detect which features are statistically significant
- Test interactions between attributes

```
> model_linear <- lm(quality ~., data = whitewinequality)
> summary(model_linear)
```

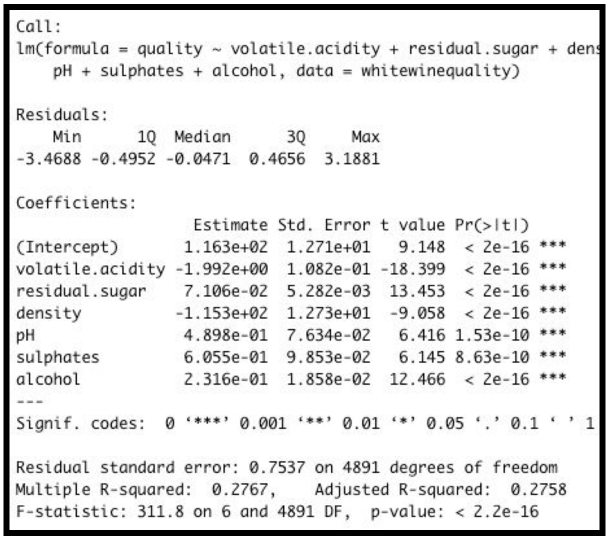


```
Call:
lm(formula = quality ~ ., data = whitewinequality)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8348 -0.4934 -0.0379  0.4637  3.1143 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.502e+02  1.880e+01   7.987 1.71e-15 ***
fixed.acidity         6.552e-02  2.087e-02   3.139  0.00171 ** 
volatile.acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
citric.acid           2.209e-02  9.577e-02   0.231  0.81759    
residual.sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
chlorides            -2.473e-01  5.465e-01  -0.452  0.65097    
free.sulfur.dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
total.sulfur.dioxide -2.857e-04  3.781e-04  -0.756  0.44979    
density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7514 on 4886 degrees of freedom
Multiple R-squared:  0.2819,	Adjusted R-squared:  0.2803 
F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

Out of 11 features affecting quality, 7 are statistically significant. Among these, volatile acidity, residual sugar and density are negatively correlated with wine quality while pH, alcohol and sulphates are positively correlated with wine quality. Free sulfur dioxide was statistically significant but had no real correlation with the wine quality index.

This adjusted model’s R-squared is 28.19%, which means that 28.19% of the wine quality is explained by the model. So now let’s fit the model with our statistically significant variables only.

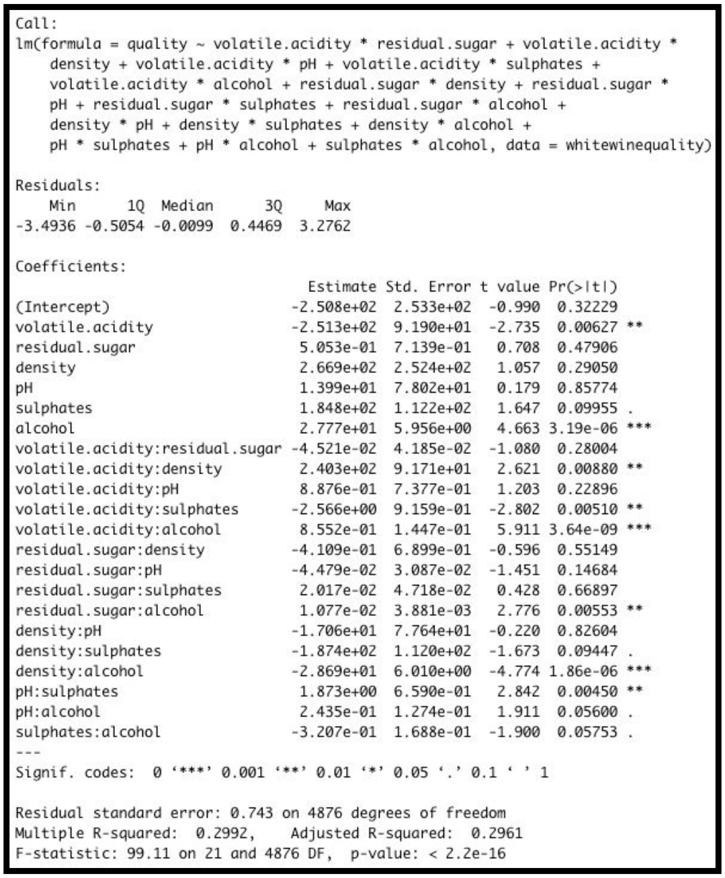
```
> model_linear_sig <- lm(quality ~ volatile.acidity + residual.sugar + density + pH + sulphates + alcohol, data = whitewinequality)
> summary(model_linear_sig)
```



```
Call:
lm(formula = quality ~ volatile.acidity + residual.sugar + dens
    pH + sulphates + alcohol, data = whitewinequality)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4688 -0.4952 -0.0471  0.4656  3.1881 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.163e+02  1.271e+01   9.148  < 2e-16 ***
volatile.acidity -1.992e+00  1.082e-01 -18.399  < 2e-16 ***
residual.sugar    7.106e-02  5.282e-03  13.453  < 2e-16 ***
density          -1.153e+02  1.273e+01  -9.058  < 2e-16 ***
pH                4.898e-01  7.634e-02   6.416 1.53e-10 ***
sulphates         6.055e-01  9.853e-02   6.145 8.63e-10 ***
alcohol           2.316e-01  1.858e-02  12.466  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7537 on 4891 degrees of freedom
Multiple R-squared:  0.2767,	Adjusted R-squared:  0.2758 
F-statistic: 311.8 on 6 and 4891 DF,  p-value: < 2.2e-16
```

This model with only statisticallys significant variables has a similar adjusted R-squared value. Around 27.67%.

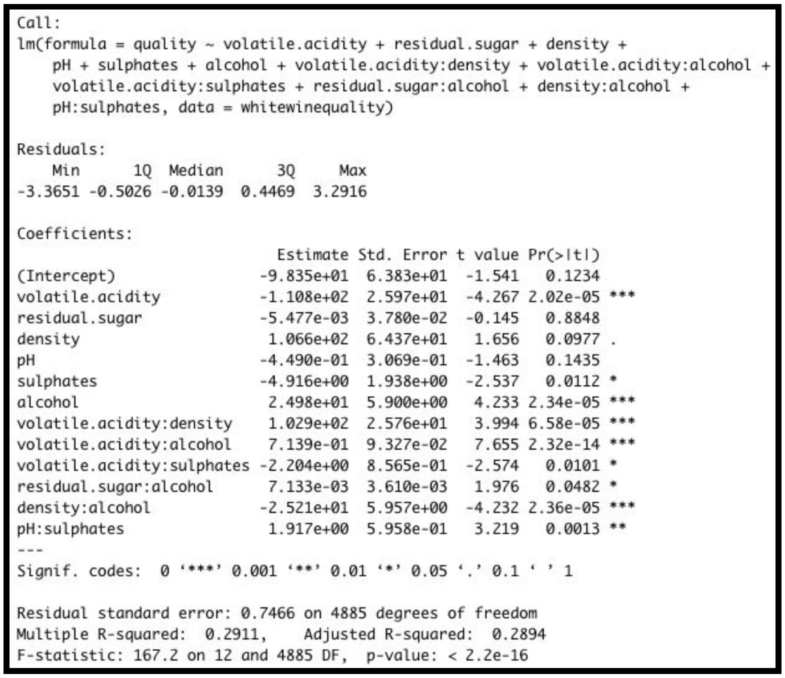
```
> model_linear_inter <- lm(quality ~ volatile.acidity*residual.sugar + volatile.acidity*density + volatile.acidity*pH + volatile.acidity*sulphates + volatile.acidity*alcohol + residual.sugar*density + residual.sugar*pH + residual.sugar*sulphates + residual.sugar*alcohol + density*pH + density*sulphates + density*alcohol + pH*sulphates + pH*alcohol + sulphates*alcohol, data = whitewinequality)
> summary(model_linear_inter)
```



```
Call:
lm(formula = quality ~ volatile.acidity * residual.sugar + volatile.acidity *
    density + volatile.acidity * pH + volatile.acidity * sulphates +
    volatile.acidity * alcohol + residual.sugar * density + residual.sugar *
    pH + residual.sugar * sulphates + residual.sugar * alcohol +
    density * pH + density * sulphates + density * alcohol +
    pH * sulphates + pH * alcohol + sulphates * alcohol, data = whitewinequality)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4936 -0.5054 -0.0099  0.4469  3.2762

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -2.508e+02  2.533e+02  -0.990  0.32229
volatile.acidity               -2.513e+02  9.190e+01  -2.735  0.00627 **
residual.sugar                  5.053e-01  7.139e-01   0.708  0.47906
density                         2.669e+02  2.524e+02   1.057  0.29050
pH                              1.399e+01  7.802e+01   0.179  0.85774
sulphates                       1.848e+02  1.122e+02   1.647  0.09955 .
alcohol                         2.777e+01  5.956e+00   4.663 3.19e-06 ***
volatile.acidity:residual.sugar -4.521e-02  4.185e-02  -1.080  0.28004
volatile.acidity:density        2.403e+02  9.171e+01   2.621  0.00880 **
volatile.acidity:pH             8.876e-01  7.377e-01   1.203  0.22896
volatile.acidity:sulphates     -2.566e+00  9.159e-01  -2.802  0.00510 **
volatile.acidity:alcohol        8.552e-01  1.447e-01   5.911 3.64e-09 ***
residual.sugar:density         -4.109e-01  6.899e-01  -0.596  0.55149
residual.sugar:pH              -4.479e-02  3.087e-02  -1.451  0.14684
residual.sugar:sulphates        2.017e-02  4.718e-02   0.428  0.66897
residual.sugar:alcohol          1.077e-02  3.881e-03   2.776  0.00553 **
density:pH                     -1.706e+01  7.764e+01  -0.220  0.82604
density:sulphates              -1.874e+02  1.120e+02  -1.673  0.09447 .
density:alcohol                -2.869e+01  6.010e+00  -4.774 1.86e-06 ***
pH:sulphates                    1.873e+00  6.590e-01   2.842  0.00450 **
pH:alcohol                      2.435e-01  1.274e-01   1.911  0.05600 .
sulphates:alcohol              -3.207e-01  1.688e-01  -1.900  0.05753 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.743 on 4876 degrees of freedom
Multiple R-squared:  0.2992,    Adjusted R-squared:  0.2961
F-statistic: 99.11 on 21 and 4876 DF,  p-value: < 2.2e-16
```

There seem to be interactions between i) volatile acidity & density, ii)volatile acidity & alcohol, iii)volatile acidity and sulphates, iv)residual sugar and alcohol, v)density and alcohol, and vi) pH and sulfates (at alpha = 0.05). Now let’s fit a new model with all attributes and our statistically signficant interactions.

```
> model_linear_inter2<- lm(quality ~ volatile.acidity + residual.sugar + density + pH + sulphates + alcohol + volatile.acidity:density + volatile.acidity:alcohol + volatile.acidity:sulphates + residual.sugar:alcohol + density:alcohol + pH:sulphates , data = whitewinequality)

> summary(model_linear_inter2)
```



```
Call:
lm(formula = quality ~ volatile.acidity + residual.sugar + density + 
    pH + sulphates + alcohol + volatile.acidity:density + volatile.acidity:alcohol + 
    volatile.acidity:sulphates + residual.sugar:alcohol + density:alcohol + 
    pH:sulphates, data = whitewinequality)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3651 -0.5026 -0.0139  0.4469  3.2916 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -9.835e+01  6.383e+01  -1.541   0.1234    
volatile.acidity           -1.108e+02  2.597e+01  -4.267 2.02e-05 ***
residual.sugar             -5.477e-03  3.780e-02  -0.145   0.8848    
density                     1.066e+02  6.437e+01   1.656   0.0977 .  
pH                         -4.490e-01  3.069e-01  -1.463   0.1435    
sulphates                  -4.916e+00  1.938e+00  -2.537   0.0112 *  
alcohol                     2.498e+01  5.900e+00   4.233 2.34e-05 ***
volatile.acidity:density    1.029e+02  2.576e+01   3.994 6.58e-05 ***
volatile.acidity:alcohol    7.139e-01  9.327e-02   7.655 2.32e-14 ***
volatile.acidity:sulphates -2.204e+00  8.565e-01  -2.574   0.0101 *  
residual.sugar:alcohol      7.133e-03  3.610e-03   1.976   0.0482 *  
density:alcohol            -2.521e+01  5.957e+00  -4.232 2.36e-05 ***
pH:sulphates                1.917e+00  5.958e-01   3.219   0.0013 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7466 on 4885 degrees of freedom
Multiple R-squared:  0.2911,	Adjusted R-squared:  0.2894 
F-statistic: 167.2 on 12 and 4885 DF,  p-value: < 2.2e-16
```

Conclusion for multivariable linear regression model

Lower levels of volatile acidity, density, and residual sugar could generally result in better wine tasting, as well as higher levels of alcohol and pH. Also, the big collinearities mostly involve insignificant variables. Adjusted R-squared is only 29.11%, implying a limited level of fit of the model too.