Aim: To build a Multiple Linear Regression model to predict the quality of wine.

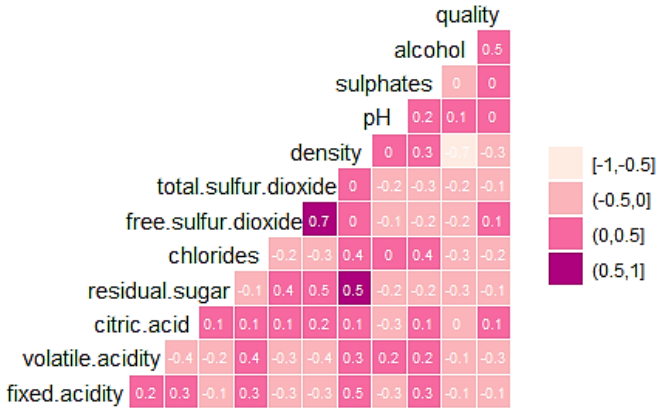
Exploring the data

There are two datasets, where each dataset has 12 features such as quality of wine, alcohol level, sugar and so on. All the features have continuous values except the *quality* of wine which is discrete variable bounded in the range 1-10 (more the better). The red wine dataset has 1599 samples and white wine dataset has 4899 samples. We have merged the two datasets into new dataset "wine" and added a factor variable with two levels (red, white) to denote the color of wine. There were no missing values in the dataset, but there are duplicate values (1177) in the dataset.

Visual inspection of scatter plots and inspection of minimum and maximum data points revealed no outliers. We continued our analysis using entire dataset by **removing duplicate values;** checked using *wine[duplicated(wine),].*

```
## 'data.frame'          : 5320 obs. of  13 variables:
##  $ fixed.acidity      : num  7.4 7.8 7.8 11.2 7.4 7.9 7.3 7.8 7.5 6.7 ...
##  $ volatile.acidity   : num  0.7 0.88 0.76 0.28 0.66 0.6 0.65 0.58 0.5 0.58 ...
##  $ citric.acid        : num  0 0 0.04 0.56 0 0.06 0 0.02 0.36 0.08 ...
##  $ residual.sugar     : num  1.9 2.6 2.3 1.9 1.8 1.6 1.2 2 6.1 1.8 ...
##  $ chlorides          : num  0.076 0.098 0.092 0.075 0.075 0.069 0.065 0.073 0.071 0.097 ...
##  $ free.sulfur.dioxide: num  11 25 15 17 13 15 15 9 17 15 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 40 59 21 18 102 65 ...
##  $ density            : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                 : num  3.51 3.2 3.26 3.16 3.51 3.3 3.39 3.36 3.35 3.28 ...
##  $ sulphates          : num  0.56 0.68 0.65 0.58 0.56 0.46 0.47 0.57 0.8 0.54 ...
##  $ alcohol            : num  9.4 9.8 9.8 9.8 9.4 9.4 10 9.5 10.5 9.2 ...
##  $ quality            : int  5 5 5 6 5 5 7 7 5 5 ...
##  $ color              : Factor w/ 2 levels "red","white": 1 1 1 1 1 1 1 1 1 1 ...
```

Correlation between variables



We have used correlation matrix to select the predictor variables that have influence on the quality of wine. From the above, features like **residual sugar and density**, and **density and alcohol** are correlated with each other. Also, none of the variables display high correlation ranging from 0.8~1.

Multiple Linear Regression

Assumptions to be tested *Prior* to constructing the model

1) All predictor variables must be quantitative or categorical, and outcome must be quantitative, continuous, and unbounded. All predictor variables are quantitative while **color** is a categorical variable (with 2 levels). **Quality** is an ordinal categorical variable in the range 1-10. However, **for the purpose of analysis we are assuming** that the Quality of wine is a continuous interval variable.

2) Predictor variables are not highly correlated with any other variable in the dataset.
As seen by the correlation matrix above, we proceed by considering our predictor variables as not being highly correlated with any other predictors in the dataset.

3) Predictors are uncorrelated with external variables.
We have not been able to test whether there are external variables. We thus assume that these are the only variables under consideration.

4) Checking non-zero variance between quality and color groups (Using Levene Test)

```
## Levene's Test for Homogeneity of Variance (center = median)
##                      Df       F value        Pr(>F)
## group                 1        3.9541      0.04681*
##                    5318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result F= 3.9541, p=0.04681 is non-significant for the wine quality at 0.01 level of significance (the value in the Pr (>F) column is more than .01). This indicates that the variances are similar between groups and the homogeneity of variance assumption is applicable.

Building the Regression model

n1: Model with all predictor variables
n2: Model after removing citric acid which has high p-value and cannot be trusted in the model
n3: Model after removing density; VIF = 22 (>10) and tolerance = 0.04 (<0.1), which indicates a serious problem [1]
n4: removing fixed acidity which has high p-value and cannot be trusted in the model

Note: R-squared will be reduced as we are reducing the number of predictor variables, but it was important to remove variables with high multicollinearity and p values > 0.05 level of significance

Comparing the 4 models

```
##   Calls:
##   n1: lm(formula = quality ~ ., data = wine)
##   n2: lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sulfur.dioxide +
##   total.sulfur.dioxide + density + pH + sulphates + alcohol + color, data = wine)
##   n3: lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sulfur.dioxide +
##   total.sulfur.dioxide + pH + sulphates + alcohol + color, data = wine)
##   n4: lm(formula = quality ~ volatile.acidity + residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##   pH + sulphates + alcohol + color, data = wine)
##
##   ================================================================================
##                                      n1              n2              n3              n4
##   --------------------------------------------------------------------------------------------------
##   (Intercept)                    99.938  ***     98.701  ***      1.734  ***      1.709  ***
##                                 (15.155)        (15.131)        (0.339)         (0.255)
##   fixed.acidity                   0.077  ***      0.082  ***     -0.001
##                                  (0.017)         (0.017)         (0.010)
##   volatile.acidity               -1.344  ***     -1.388  ***     -1.442  ***     -1.441  ***
##                                  (0.088)         (0.082)         (0.082)         (0.081)
##   citric.acid                     0.124
##                                  (0.088)
##   residual.sugar                  0.054  ***      0.054  ***      0.016  ***      0.016  ***
##                                  (0.006)         (0.006)         (0.003)         (0.003)
##   chlorides                      -0.917  **       -0.827  *       -1.075  **       -1.073  **
##                                  (0.350)         (0.344)         (0.343)         (0.343)
##   free.sulfur.dioxide             0.006  ***      0.006  ***      0.006  ***      0.006  ***
##                                  (0.001)         (0.001)         (0.001)         (0.001)
##   total.sulfur.dioxide           -0.002  ***     -0.002  ***     -0.002  ***     -0.002  ***
##                                  (0.000)         (0.000)         (0.000)         (0.000)
##   density                       -99.617  ***    -98.369  ***
##                                 (15.370)        (15.346)
##   pH                              0.617  ***      0.605  ***      0.224  **       0.229  ***
##                                  (0.099)         (0.098)         (0.079)         (0.068)
##   sulphates                       0.746  ***      0.752  ***      0.620  ***      0.619  ***
##                                  (0.084)         (0.084)         (0.081)         (0.081)
##   alcohol                         0.230  ***      0.234  ***      0.340  ***      0.340  ***
##                                  (0.019)         (0.019)         (0.010)         (0.010)
##   color: white/red               -0.322  ***     -0.315  ***     -0.111  *        -0.109  *
##                                  (0.061)         (0.061)         (0.052)         (0.047)
##   --------------------------------------------------------------------------------------------------
##   R-squared                       0.311           0.311           0.306           0.306
##   N                               5320            5320            5320            5320
##   ================================================================================
##   Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

Summary for the chosen regression model: n4

```
##  Call:
##  lm(formula = quality ~ volatile.acidity + residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + pH +
##  sulphates + alcohol + color, data = wine)
##
##  Residuals:
##       Min        1Q    Median        3Q       Max
##  -4.0417   -0.4498   -0.0259    0.4621    3.1442
##
##  Coefficients:                Estimate  Std. Error  t value  Pr(>|t|)
##  (Intercept)                 1.7092885   0.2553039    6.695  2.38e-11 ***
##  volatile.acidity           -1.4412048   0.0813728  -17.711   < 2e-16 ***
##  residual.sugar              0.0163727   0.0026758    6.119  1.01e-09 ***
##  chlorides                  -1.0729989   0.3426619   -3.131   0.00175 **
##  free.sulfur.dioxide         0.0064696   0.0008341    7.756  1.04e-14 ***
##  total.sulfur.dioxide       -0.0018309   0.0003490   -5.246  1.62e-07 ***
##  pH                          0.2286043   0.0681470    3.355   0.00080 ***
##  sulphates                   0.6191738   0.0810273    7.642  2.53e-14 ***
##  alcohol                     0.3401657   0.0099541   34.173   < 2e-16 ***
##  colorwhite                 -0.1085062   0.0466369   -2.327   0.02002 *
##  ---
##  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Residual standard error: 0.7338 on 5310 degrees of freedom
##  Multiple R-squared:  0.3056, Adjusted R-squared:  0.3044
##  F-statistic: 259.6 on 9 and 5310 DF, p-value: < 2.2e-16
```

Predicting the Quality of Wine using our Regression Model (n4)

Predicted value for quality of wine with prediction intervals against volatile.acidity=1, residual.sugar=3, chlorides=0.1, free.sulfur.dioxide=7, total.sulfur.dioxide=16, pH=3.43, sulphates=0.46, and alcohol=10,color='red'

```
##         fit      lwr      upr
##  1 4.696485 3.255251 6.137719
```

In the given dataset, real value = 5, predicted ≈ 4.696. Thus, a close representation within 95% of prediction interval.

Checking the assumptions for multiple regressions (*Post* model building)

1) *No perfect multicollinearity (predictor variables should not correlate highly)*

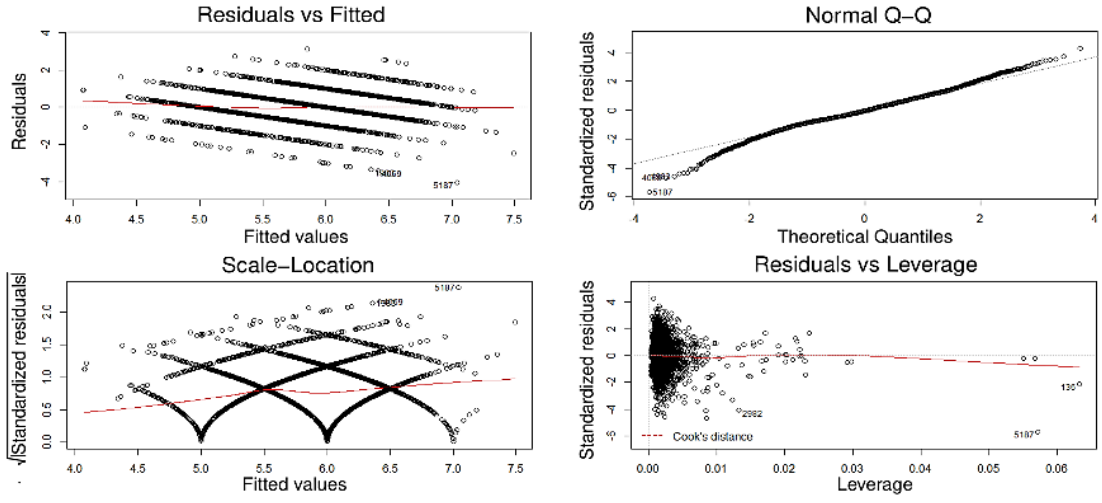
a) *VIF value is <10 and Tolerance is greater than 0.1 and 0.2*

	Variables	Tolerance	VIF	Variables	Tolerance	VIF
1	volatile.acidity	0.5400178	1.851791	residual.sugar	0.6980801	1.432500
3	chlorides	0.6343762	1.576352	free.sulfur.dioxide	0.4588976	2.179136
5	total.sulfur.dioxide	0.2577924	3.879091	pH	0.8473803	1.180108
7	sulphates	0.6875628	1.454413	alcohol	0.7263417	1.376762
9	colorwhite	0.2446393	4.087651			

b) *Average is closer to 1, in our case it is 2.11 that indicates no high correlation i.e. variables do not corelate highly.*

```
##  [1]  2.113089
```

2) *Residuals are Linear, Normal and Homoscedastic (constant variance)*



From the plots we observe the following:
a) The residuals are linear: Residuals vs Fitted is used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship.
b) The residuals are Normal: Normal Q-Q is used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line which is almost normal in our case.
c) Homogeneity of variance not observed: Scale-Location (or Spread-Location) is used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity. This is not the case in our dataset, where we have a heteroscedasticity problem.
d) Residuals vs Leverage is used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. Further analysis done below using Cook's distance.

3) ***Residuals are independent (via Durbin-Watson test)***

```
##  lag  Autocorrelation  D-W Statistic  p-value
##    1       0.1136399       1.772705        0
##  Alternative hypothesis: rho != 0
```

Null Hypothesis: Linear Regression residuals of wine are uncorrelated.
Alternate Hypothesis: Linear Regression residuals of wine are autocorrelated.
The Durbin-Watson test for independent errors was significant at 5% level of significance. Despite d=1.77 which doesn't imply autocorrelation, a significantly small p-value (p=0) casts doubt on the validity of the null hypothesis and indicates autocorrelation among residuals. This implies that the model has not accounted for all signals and thus, it consists of signal plus noise.

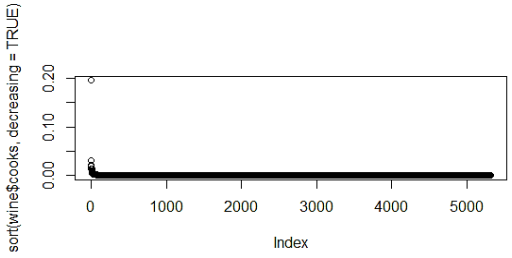
4) ***Checking for outliers and influential points***

a) *Taking standardized residuals to check for outliers*

```
##  [1]  301  16
```

301 observations lie above or below 1.96 standard deviations. As this represents 5.6% of the observations, expected if the residuals are normal i.e., 5% of data is expected to be outside of two standard deviations[1], we do not consider any of these observations as outliers and continued with all 301 observations included in the model.

b) **Cook's distance** *to find out the influential cases*




```
##  [1]  0.1951897
```

Maximum cook's distance is 0.195 (<< threshold value of 1). So, we conclude that there are no influential cases.

Conclusion and Future Work

1. Multiple Linear Regression model was built to predict the quality of wine using the significant variables: volatile acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, alcohol and color.
2. Backward elimination method was used to create this regression model (all predictors entered simultaneously without any order) and then removing predictors that were not significant or failed multicollinearity assumption. All the incorporated predictor variables have an influence on the wine quality at 5% level of significance.
3. Multiple R-squared value is 0.3056 and Adjusted R- squared value is 0.3044. Hence, our model explains 30.44% variance in the Wine Quality and generalizes the population as well. The remaining 69.56% remains unexplained.
4. One standard deviation change in the value of alcohol brings 0.458 (using lm.beta(n4)) standard deviation of change in wine quality. Therefore, alcohol has highest impact on quality of wine.
5. Further work: Finding the quality and type of grape that was used to make wine and storing conditions as well.

References

[1] *Andy Field, Jeremy Miles, Zor Field,* Discovering Statistics using R, Chapter 7 - Regression, *SAGE Publication Ltd.*
[2] *Paulo Cortez*, *Ant´onio Cerdeira*, Modeling wine preferences by data mining from physicochemical properties