# SIT718 – Real Word Analytics

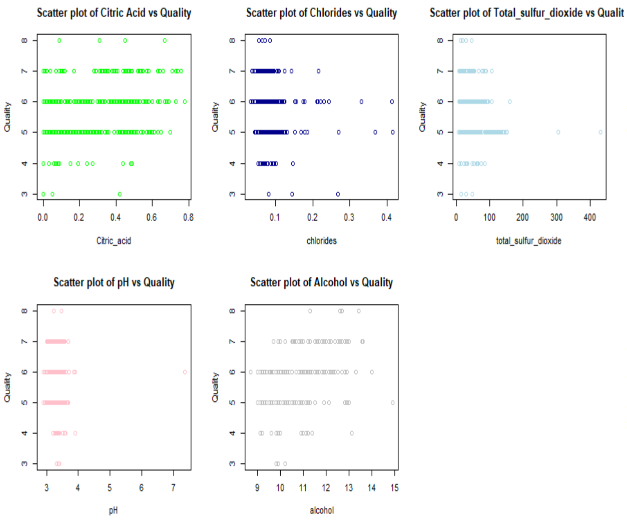
# Assignment 2 – Using aggregation functions for data analysis

By: CHAOYI (BARRY) CHEN

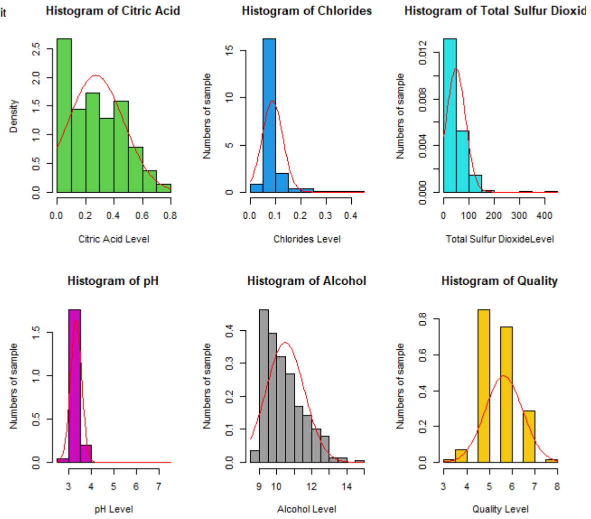
Student ID: 220358865

## Q1. Understand the data

### Scatter plots of each variable



### Histograms of each variable



Correlation between each variable and variable of interest (quality)

a) **Citric Acid vs Quality**- the **correlation is 0.2668687**, indicated that the correlation between Citric Acid and Quality is **Low Positive Correlation.**
b) **Chlorides and Quality**- the **correlation is -0.1033161**, indicated that the correlation between Chlorides and Quality is **Low Negative Correlation.**
c) **Total sulfur dioxide vs Quality**- the **correlation is -0.2156311**, indicated that the correlation between Total sulfur dioxide and Quality is **Low Negative Correlation**.
d) **pH vs Quality** - the **correlation is -0.01423063**, indicated that the correlation between pH and Quality is **Low Negative Correlation**.
e) **Alcohol and Quality** - the **correlation is 0.4782609**, indicated that the correlation between Alcohol and Quality is **Low Positive Correlation**.

The lowest correlation was -0.01423063 which was between pH level and the quality of the wine.

It is important to check whether variable is normally distributed or symmetrical histogram, if the histogram is asymmetry, then it means that variable is having few extremely large or small data that moving the Mean from the Median of the variable. We would like the Mean to be close or equals to Median, so we would look at the skewness of each histogram. If the skewness is 0, indicate it is symmetry, however if the value further away from 0 (in both positive and negative direction), the histogram would be asymmetry. The approximately symmetric skewness is within -0.5 to 0.5 interval[1].

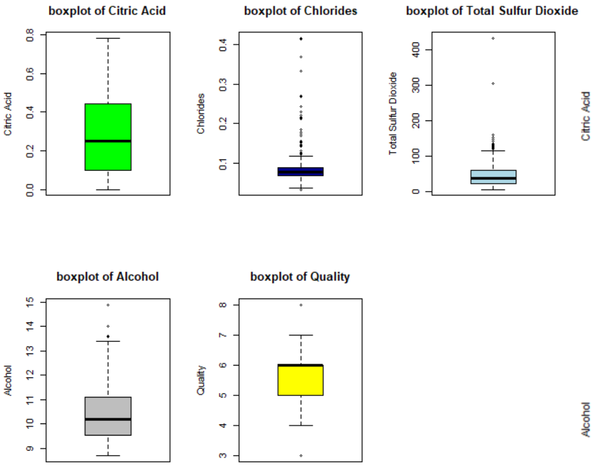
a) **Citric Acid -** The **skewness of Histogram of Citric Acid is 0.34**, indicated the histogram is approximately symmetric, which the mean is very close to the median value.
b) **Chlorides** - The **skewness of Histogram of Chlorides is 4.5**, indicated the histogram is far away from symmetry and skewed to the right, which affected by few extremely large values.
c) **Total Sulfur Dioxide** - The **skewness of Histogram of Total Sulfur Dioxide is 3.65**, like the Chlorides variable, the histogram is far away from symmetry and skewed to the right, which affected by few extremely large values.
d) **pH** - The **skewness of Histogram of pH is 10.17**, indicated the histogram is highly asymmetrical, it skewed to the right by many extremely large values, the mean would be largely different from it median.
e) **Alcohol** - The **skewness of Histogram of Alcohol is 0.87**, indicated the histogram is a little bit far away from symmetry, but it still slightly skewed to the right which the mean is slightly bigger than median.
f) **Quality** - The **skewness of Histogram of Quality is 0.17**, indicated the histogram is very close to symmetry and the variable is normally distributed.

[1] SaiGayatri Vadali, 30 Dec 2017, “Day 8: Data transformation – Skewness, normalization and much more”, viewed Aug 2021, https://medium.com/@TheDataGyan/day-8-data-transformation-skewness-normalization-and-much-more-4c144d370e55

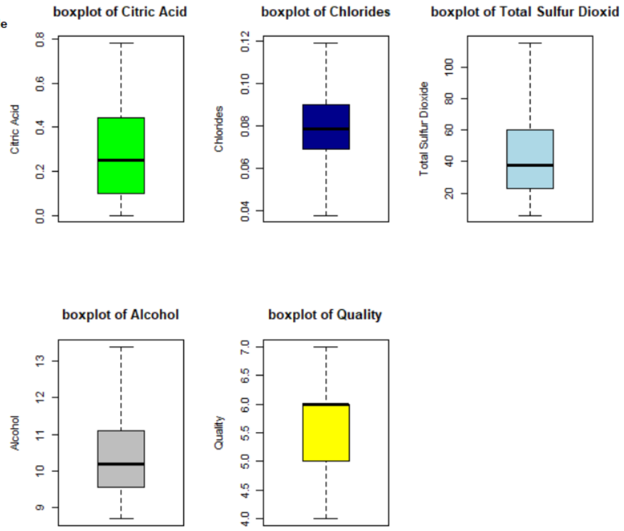
## Q2. Transform the data

I have selected 4 variables (citric acid, chlorides, total sulfur dioxide and alcohol) since their histogram skewness are closer to 0 and have higher correlation with variable of interest (quality) compared to pH variable. Now we could extract them from the dataset and handle outliers of each variable prior to data transformation,

**Boxplots of variables to identify outliers**

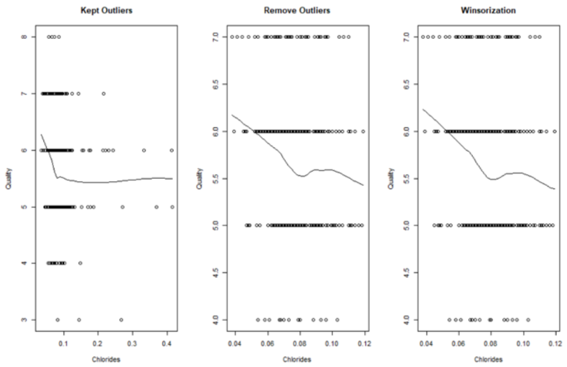


**Boxplots of variables after winsorized outliers**



According to the boxplots on the left above, we could see no outliers in citric acid variable boxplot, but alcohol, total sulfur dioxide and quality variable all showing a few outliers, and chlorides have the most outliers base on the boxplots. We also found that total 63 observations from the dataset have been identified as outliers using interquartile range (IQR) criterion[2], which around 14% of all data. Either we remove or keeping these outliers, its most likely these outliers will impact the accuracy of our model which need to use these data for prediction, and since I also have no expertise regarding to wine quality on deciding which outlier should be kept, therefore I proposed Winsorization which is converting them into either 1st quartile or 3rd quartile of each variable[3].

It's also essential to plot graphs of each variable to visualize what effect if we either keep these outliers, remove them or winsorize them[4]. For example, the Chlorides variable scatter plots with different outlier handling methods on the right, clearly shows these outliers would impact its linear relationship with Quality, which the quality did not trending down as more chlorides contained in the wine if keeping all outliers, and by Winsoriztion, we could have this variable maintain the similar relationship with quality as it would have its outliers removed, without losing any data due to outliers, therefore I chosen Winsorization as it would be a better option on handling outliers in this case. The box plot on the top right showing all outliers have been winsorized.



Note: After outliers handling, the Pearson correlation between quality and citric acid, chlorides are still positive, and the correlation with total sulfur dioxide, alcohol are still showing negative.

[2] Shivam Chaudhary, 28 Sep 2019, "Why "1.5" in IQR method of outlier detection?", viewed Aug 2021, https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097

[3] M Nayoumi, 24 Feb 2021, "Scaling outliers using winsorizing",viewed Aug 2021, https://m-nayoumi.medium.com/scaling-outliers-using-winsorizing-ecd6901187cb

[4] Kayla Ferguson, 06 March 2018, "When should you delete outliers from a dataset?", viewed Aug 2021, https://humansofdata.atlan.com/2018/03/when-delete-outliers-dataset/

**Data transformation of each variable**

a) **Citric Acid** – since this variable has skewness of 0.34, we can use could just use Z-score standardisation to convert it to unit interval, the use polynomial method to reduce the skewness to around -0.0002284564. Negation is not needed as Citric Acid has positive linear relationship with Quality as identified in Q1.

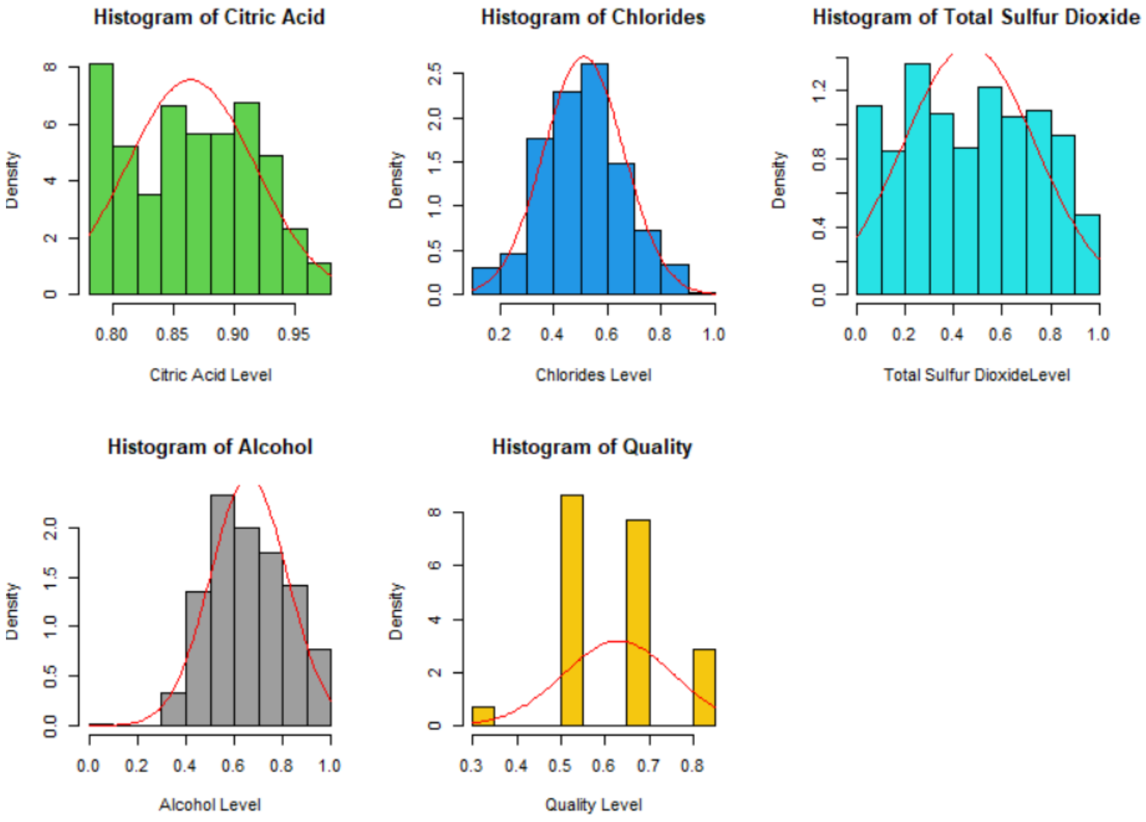
b) **Chlorides** – since this variable has negative correlation with quality and it does not in unit interval, therefore we use negation function first to maintain dataset consistency. The variable has skewness of 0.0367 after negation, therefore we could just use Z-score standardisation to convert it to unit interval. After that we would perform polynomial transformation to further reduce the skewness of the data to 0.000596.

c) **Total Sulfur Dioxide** – since this component has negative correlation with quality, therefore we will use negation function first to convert the data, so lesser this component the better the quality. After we used Min-Max normalization method to convert to unit interval (skewness of this data is around -0.82 therefore Z-scores method not appliable). Then we could use polynomial method to reduce the skewness to around 0.001498176.

d) **Alcohol** - since this component has positive relationship with quality variable, negation is not needed. We checked the skewness of this data is around 0.72 therefore Z-scores method not appliable, so we could use Min-Max normalization method to convert it to unit interval. Then we applied polynomial method to reduce the skewness of the data to around -0.0009473035.

e) **Quality** – since this variable has around 0.2437 skewness which the data is close to normal distribution, we applied Z-scores standardization method and scale it to unit interval. Then we applied the polynomial method to reduce the skewness to around -0.00357144.

Histogram of each variable after data transformation



All histograms showing data are normally distributed after transformation and within unit interval [0,1].

**Q3 Build models and investigate the importance of each variable.**

(i-iii) Results from fitting functions to the data learn the parameters, all figures have been rounded up to 4 decimal places for easier comparison.

Table 1 – Summary of error measures and correlation coefficients

| | WAM | WPM with p=0.1 | WPM with p=6 | OWA | Choquet Integral |
|---|---|---|---|---|---|
| RMSE | 0.1192 | 0.1266 | 0.1197 | 0.1239 | 0.1153 |
| Av. Abs error | 0.0920 | 0.0964 | 0.0896 | 0.0952 | 0.0877 |
| Pearson correlation | 0.4578 | 0.4253 | 0.4157 | 0.4134 | 0.4774 |
| Spearman correlation | 0.4949 | 0.4921 | 0.4529 | 0.4398 | 0.5180 |

Note: The best value high lined in yellow in different model
RMSE = Root mean squared error, the lesser the better
Av.abs error = Average absolute error, the lesser the better
For Pearson and Spearman correlation measurements, the higher correlation the better

Table 2 – Summary of the weight/parameters and other useful information

| | WAM | WPM with p=0.1 | WPM with p=6 | OWA | Choquet Integral |
|---|---|---|---|---|---|
| Importance of v1 | 0.2278 | 0.3392 | 0.0307 | 0.2382 | 0.1685 |
| Importance of v2 | 0.2477 | 0.2187 | 0.6638 | 0.1610 | 0.3076 |
| Importance of v3 | 0.1140 | 0.0379 | 0.1188 | 0.2966 | 0.1244 |
| Importance of v4 | 0.4104 | 0.4042 | 0.1866 | 0.3041 | 0.3995 |
| Orness | | | | 0.5556 | 0.4568 |

Note: the highest importance high lined in yellow in different model
v1 = Citric acid variable
v2 = Chlorides variable
v3 = Total sulfur dioxide variable
v4 = Alcohol variable
importance = Weight measurement of each variable from WAM, WPM and OWA, but for Choquet integral we will use the Shapley value

(iv) Data comparison and interpretation

a. From table 1, we can see the best model would be the Choquet integral model comparing to other 4 models. Although our best model has great accuracy by having the least difference between the predicted value and true value (measure by RMSE and Av abs error), the correlation between each variable and variable of interest could be better since it only has 0.4774 and 0.5180 in correlation measure (the perfect positive relationship would have 1).

b. The importance of each variable
From table 2, we could see variable v4 (Alcohol) ranked the highest importance in all models. If we just focus on our best model, variable v2 (Chlorides) ranked the 2nd best with similar importance as v1, following by v1 then v3.

c. Interaction between variables
To check the interaction between variables, we could extract the binary number fm.weights figures from the Choquet integral model report, this fuzzy measure of each variable set would tell us how effective each variable contributed to make high quality wine and how effective if they working together as a set.

| No. | Binary Numbers | Variable-set | Fuzzy Measure |
|---|---|---|---|
| 1 | 0001 | V1 | 0.2758 |
| 2 | 0010 | V2 | 0.2758 |
| 3 | 0011 | V1,2 | 0.2758 |
| 4 | 0100 | V3 | 0 |
| 5 | 0101 | V1,3 | 0.2758 |
| 6 | 0110 | V2,3 | 0.4127 |
| 7 | 0111 | V1,2,3 | 0.4127 |
| 8 | 1000 | V4 | 0 |
| 9 | 1001 | V1,4 | 0.7072 |
| 10 | 1010 | V2,4 | 0.8192 |
| 11 | 1011 | V1,2,4 | 0.8192 |
| 12 | 1100 | V3,4 | 0.4953 |
| 13 | 1101 | V1,3,4 | 0.7072 |
| 14 | 1110 | V2,3,4 | 1 |
| 15 | 1111 | V1,2,3,4 | 1 |

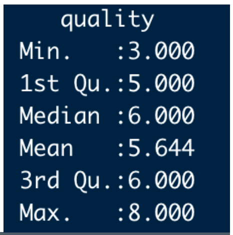
Note: we ignore the null set (v0), and fuzzy measure figure rounded up to 4 decimal places.

1) **Set v1,2** - Variable set is **Redundant,** since set value 0.2758 is less than sum of the variable set 0.5516
2) **Set v1,3** - Variable set is **Additive or no interaction**, since set value 0.2758 is equal to sum of the variable set 0.2758
3) **Set v2,3** - Variable set is **Complementary**, since set value 0.4127 is more than sum of the variable set 0.2758
4) **Set v1,2,3** - Variable set is **Redundant**, since set value 0.4127 is less than sum of the variable set 0.5516
5) **Set v1,4** - Variable set is **Complementary**, since set value 0.7072 is more than sum of the variable set 0.2758
6) **Set v2,4** - Variable set is **Complementary**, since set value 0.8192 is more than sum of the variable set 0.2758
7) **Set v1,2,4** - Variable set is **Complementary**, since set value 0.8192 is more than sum of the variable set 0.5516
8) **Set v3,4** - Variable set is **Complementary**, since set value 0.4953 is more than sum of the variable set 0
9) **Set v1,3,4** - Variable set is **Complementary**, since set value 0.7072 is more than sum of the variable set 0.2758
10) **Set v2,3,4** - Variable set is **Complementary**, since set value 1 is more than sum of the variable set 0.2758
11) **Set v1,2,3,4** - Variable set is **Complementary**, since set value 1 is more than sum of the variable set 0.5516

d. Our best model the Choquet Integral model has orness of 0.4568 which is less than 0.5, so it is likely favour lower inputs.

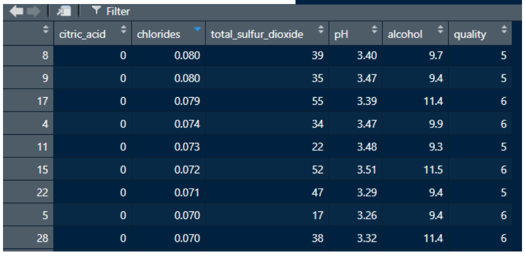
**4. Use your model for prediction**

(i) The predicted quality of wine from given inputs by our model is 5.
(ii) The predicted value equals to the 1st quartile of wine quality in our original dataset, which is also 5, therefore the result is within range of quality variable of the dataset.



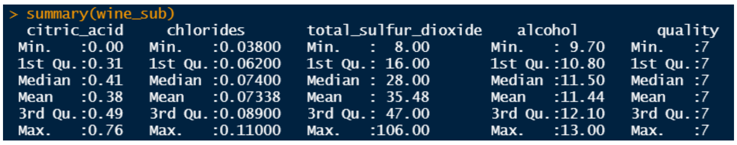
```
   quality
Min.   :3.000
1st Qu.:5.000
Median :6.000
Mean   :5.644
3rd Qu.:6.000
Max.   :8.000
```

On the other hand, since the given inputs are x1=0, x2=0.075, x3=41 x4=3.53 and x5=9.3, from the screenshot of the original dataset above, we can see the quality of wine with similar variable values is normally score 5, such as row 11 and 22. Therefore the result is reasonable.



| | citric_acid | chlorides | total_sulfur_dioxide | pH | alcohol | quality |
|---|---|---|---|---|---|---|
| 8 | 0 | 0.080 | 39 | 3.40 | 9.7 | 5 |
| 9 | 0 | 0.080 | 35 | 3.47 | 9.4 | 5 |
| 17 | 0 | 0.079 | 55 | 3.39 | 11.4 | 6 |
| 4 | 0 | 0.074 | 34 | 3.47 | 9.9 | 6 |
| 11 | 0 | 0.073 | 22 | 3.48 | 9.3 | 5 |
| 15 | 0 | 0.072 | 52 | 3.51 | 11.5 | 6 |
| 22 | 0 | 0.071 | 47 | 3.29 | 9.4 | 5 |
| 5 | 0 | 0.070 | 17 | 3.26 | 9.4 | 6 |
| 28 | 0 | 0.070 | 38 | 3.32 | 11.4 | 6 |

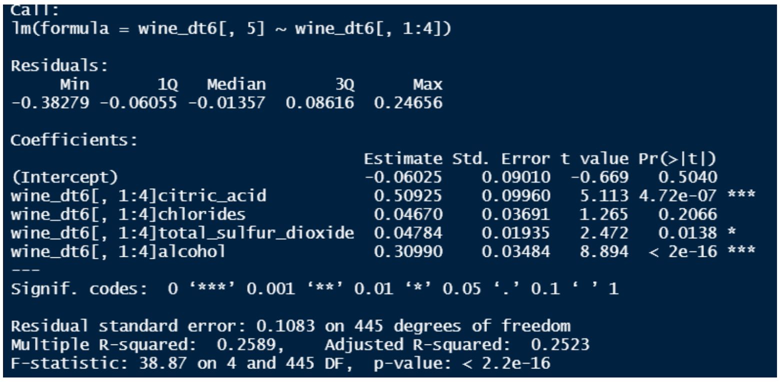
(iii) From Q3 (iv)(b), we know variables v2 (chlorides) and v4 (alcohol) are most important on predicting the quality of wine. From Q3(c) fuzzy measure, we also learnt that variable set v2,4 is the only set that with only 2 variables but having high effectiveness (0.8192 fuzzy measure). Therefore, in the condition that the wine contain low chlorides (negative correlation with quality) and high alcohol, higher quality wine would be occurred. The subset below contained data that has highest quality (dataset with outliers handled), we can see these wines all contains very low chlorides (Mean 0.07) and very high alcohol (Mean 11.44). Both total sulfur dioxide and citric acid have little to none effect on the quality.



```
> summary(wine_sub)
  citric_acid      chlorides        total_sulfur_dioxide    alcohol          quality
 Min.   :0.00   Min.   :0.03800   Min.   :  8.00         Min.   : 9.70   Min.   :7
 1st Qu.:0.31   1st Qu.:0.06200   1st Qu.: 16.00         1st Qu.:10.80   1st Qu.:7
 Median :0.41   Median :0.07400   Median : 28.00         Median :11.50   Median :7
 Mean   :0.38   Mean   :0.07338   Mean   : 35.48         Mean   :11.44   Mean   :7
 3rd Qu.:0.49   3rd Qu.:0.08900   3rd Qu.: 47.00         3rd Qu.:12.10   3rd Qu.:7
 Max.   :0.76   Max.   :0.11000   Max.   :106.00         Max.   :13.00   Max.   :7
```

## Q5. Comparing with a linear regression model

(i) The summary below shows the performance of our linear regression model

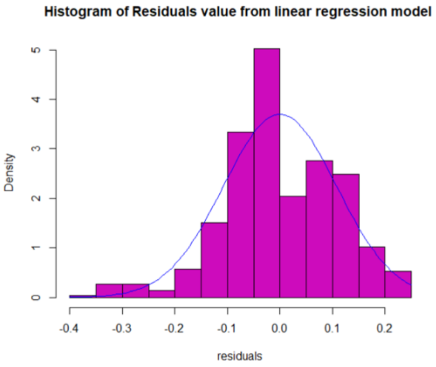


```
Call:
lm(formula = wine_dt6[, 5] ~ wine_dt6[, 1:4])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.38279 -0.06055 -0.01357  0.08616  0.24656 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         -0.06025    0.09010  -0.669   0.5040    
wine_dt6[, 1:4]citric_acid           0.50925    0.09960   5.113 4.72e-07 ***
wine_dt6[, 1:4]chlorides             0.04670    0.03691   1.265   0.2066    
wine_dt6[, 1:4]total_sulfur_dioxide  0.04784    0.01935   2.472   0.0138 *  
wine_dt6[, 1:4]alcohol               0.30990    0.03484   8.894  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1083 on 445 degrees of freedom
Multiple R-squared:  0.2589,	Adjusted R-squared:  0.2523 
F-statistic: 38.87 on 4 and 445 DF,  p-value: < 2.2e-16
```

Below I will list and explain some important figures from the summary

1) **Residual values**, these are the distance between the predicted value and true value. If the residuals are normally distributed around 0 then our model fit the data pretty good[5]. The Histogram on the right show the residuals value seems to be symmetrical around 0, it also has skewness of -0.3710673 which is between -0.5 and 0.5, indicate our model fit the data well.

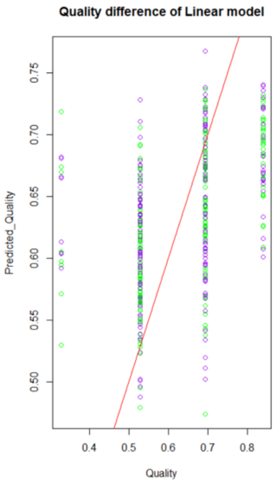
2) **Pr(>|t|) or p-values**, this is the probability of value larger or equal to t values if there are no relationship between these variables and the quality of wine, its p-value is less than 0.05 then we will reject this assumption, otherwise it would be true. From our summary, it seems v1 and 4 both have strong relationship with quality since they have 95% probability of adding meaningful addition to the model, following by v3. However, v2 seems to have no relationship with quality as its p-value larger than 0.05.

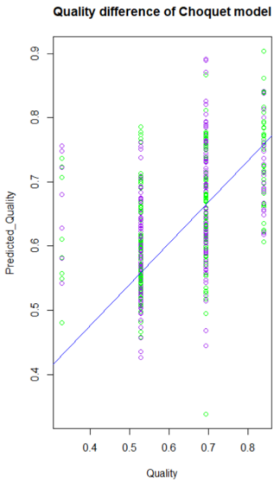


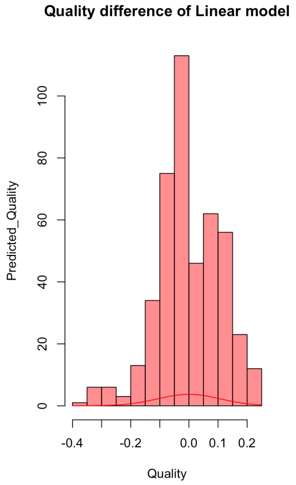
3) **Multiple/Adjusted R-squared error,** this value measure how close our model to the data, it has range between 0 and 1, the higher the value the better fit. Our model only has 0.2589 of this value, indicating our model and data may not having a strong fit.
4) **F-statistic,** this value show whether this was strong relationship between our predictor variables (considering all 4 variables at once in this case) and target variable, the higher the value the stronger the relationship. Our model only has 38.87 of this value, indicate the relationship is weak.
5) **P-value,** this is the p-value of the whole model. From the summary, our model p-value is 2.2e^(-16), this indicated our model fit the data well and also having a strong positive linear correlation between variables.

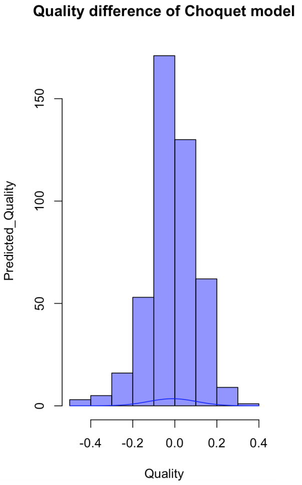
[5] Vik Paruchuri, 16 May 2018, “Using Linear Regression for Predictive Modelling in R”, viewed Aug 2021, https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/

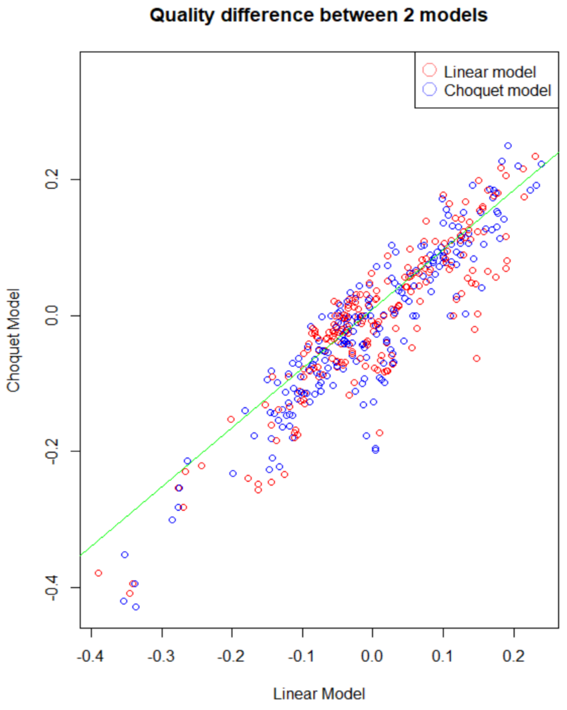
### (ii) Visualization of our models

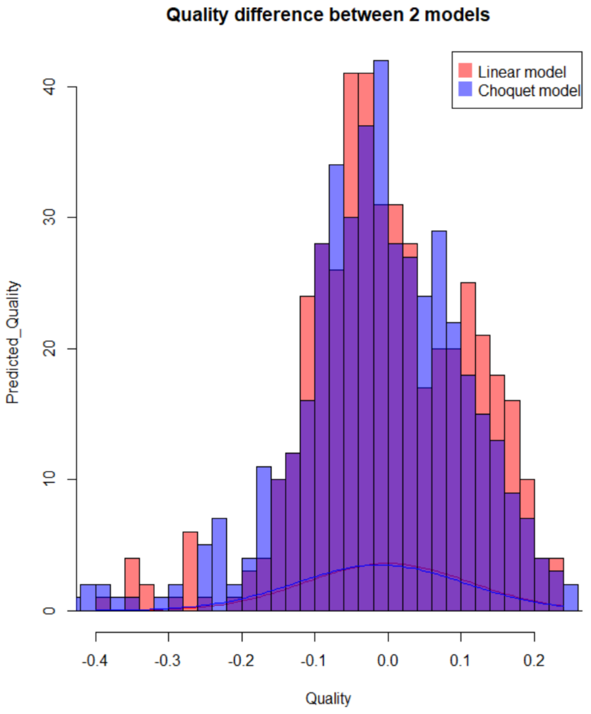












### (iii) Comments on finding comparing Linear model with Choquet model

Based on the individual scatter plots of each model above, it seems that linear model has slightly better fit than Choquet model, the Pearson correlation measure shows the Linear model has higher Pearson correlation measure (0.5088649) compares to the Choquet mode (0.4774369). Our Linear model histogram also has -0.3855657 skewness which is more symmetrical than our Choquet mode which has -0.4820037 skewness.

If we calculate the difference between the actual quality value and predicted quality values and plot them together with scatter plot, we can see a strong positive linear relationship between both models, the Pearson correlation measure score (0.9012929) also show that 2 model produced very similar result. With histogram, linear model produced most the results close to 0 which means most of the values predicted are closer to the true value. The Choquet model has the quality differences widely spread out, so some of the value difference has become quite large as shown on the histogram. Therefore, the linear regression model could be a better model than the Choquet integral model according to the measures above.

However, this was based on assumption of our relatively small dataset (with only 450 observations) and the reliability of the data processing method, we will need further testing with more data and/or examine the way we process the data, to see whether the result would change. The advantage of the Choquet model is that it accounts for the interaction between variables using fuzzy measure, but the limitation of the model is its prediction was heavily relied on the weight elements generated from the fuzzy measure which could create bias. For the linear regression model, the advantage is its simplicity, no need to work out the weight elements such as in Choquet model, the limitation is also clear, we might not fully understand the data purely from their linear relationship, and real-world data are likely not being linear.