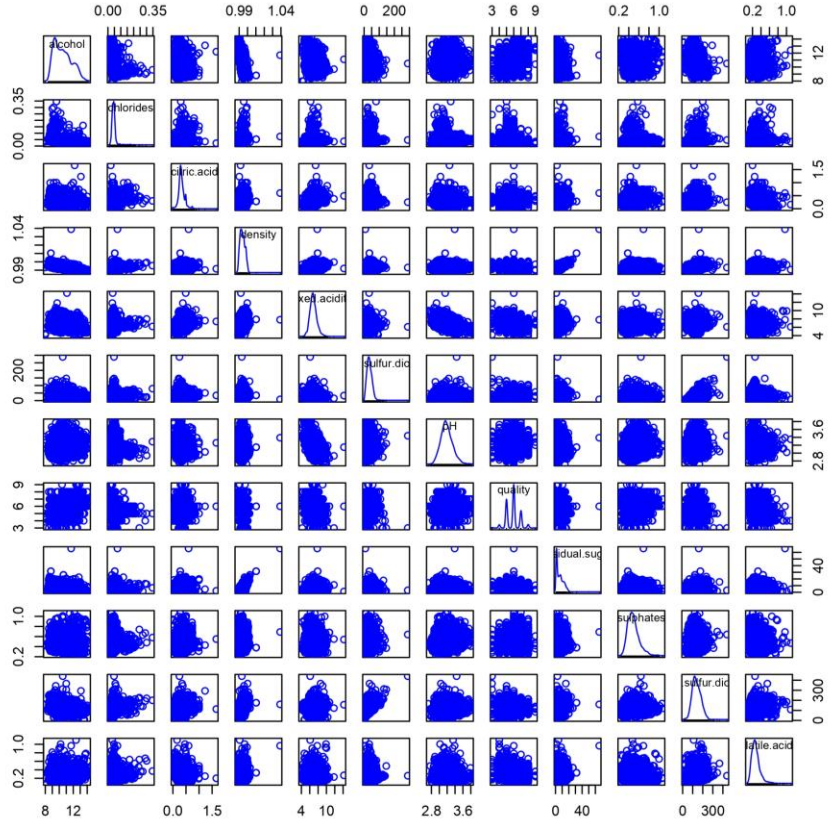
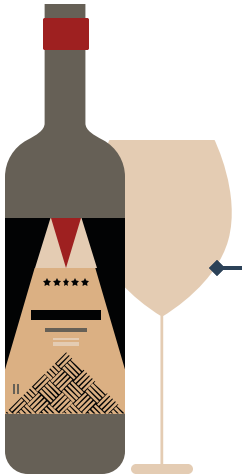
Wine Quality Analysis

Group 6 : ChihHao Yuan, Ching Yu Hsu

1a)- Overall view of relationship of Quality with all variables



1b)- Strong Relationships between pairs of variables

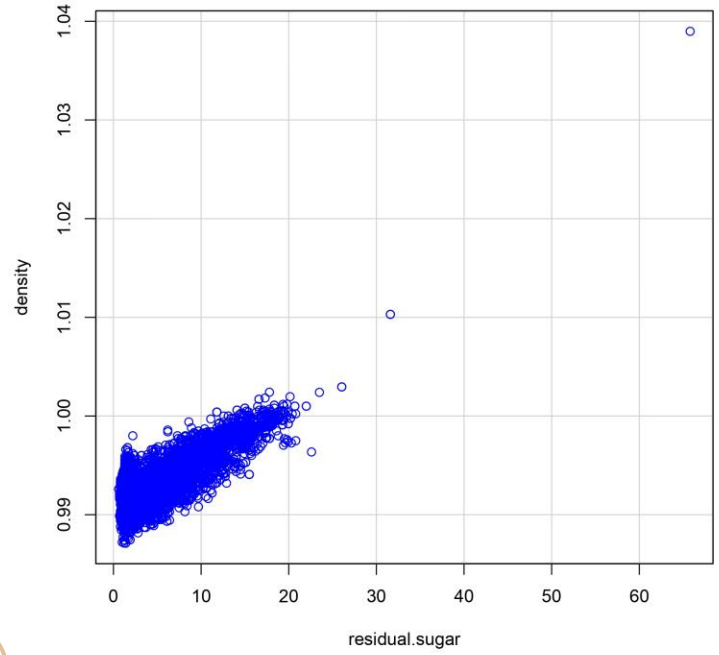


	fixed_acidi	atile_acic	itric_aci	idual_su	chlorides	sulfur_di	sulfur_di	density	pH	sulphates	alcohol	quality
fixed_acidity	1											
volatile_acidity	-0.0203	1										
citric_acid	0.2874	-0.1485	1									
residual_sugar	0.088	0.0641	0.0946	1								
chlorides	0.0236	0.0727	0.113	0.0897	1							
free_sulfur_dioxide	-0.0492	-0.097	0.093	0.2992	0.1014	1						
total_sulfur_dioxide	0.0913	0.09	0.122	0.4013	0.1994	0.6156	1					
density	0.2655	0.0273	0.1504	0.8388	0.2581	0.2943	0.5296	1				
pH	-0.4243	-0.0346	-0.1615	-0.1931	-0.0892	0.0007	0.0036	-0.0924	1			
sulphates	-0.0165	-0.035	0.0644	-0.0261	0.0174	0.0597	0.134	0.0745	0.1556	1		
alcohol	-0.1219	0.0678	-0.0766	-0.4514	-0.3601	-0.2503	-0.4487	-0.7806	0.1213	-0.0169	1	
quality	-0.1122	-0.1967	-0.0101	-0.0949	-0.2108	0.0077	-0.1746	-0.3055	0.0973	0.0542	0.4354	1

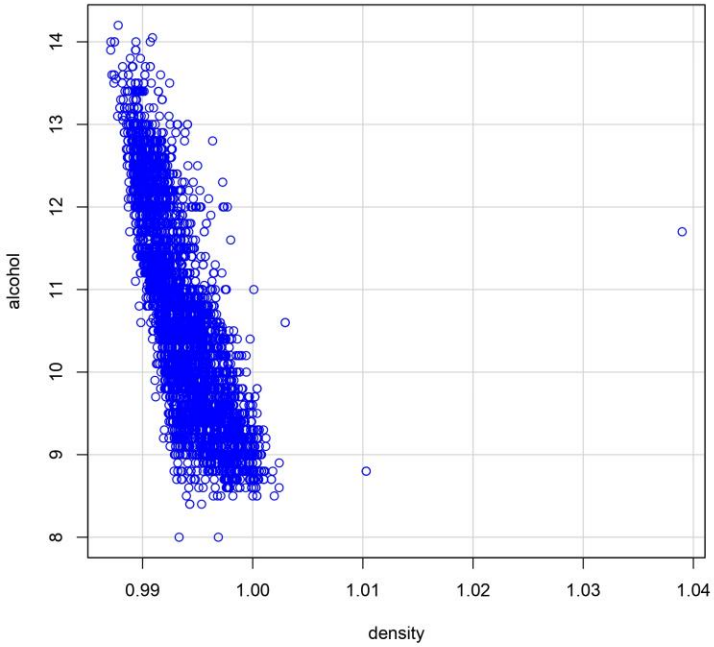
Strong Relationship :

- Residual Sugar & Density (0.8388)
- Density & Alcohol (-0.7806)

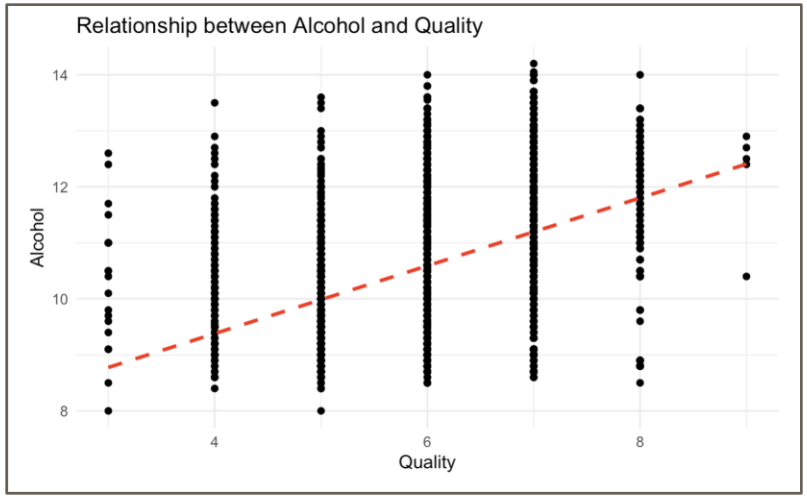
Residual Sugar & Density

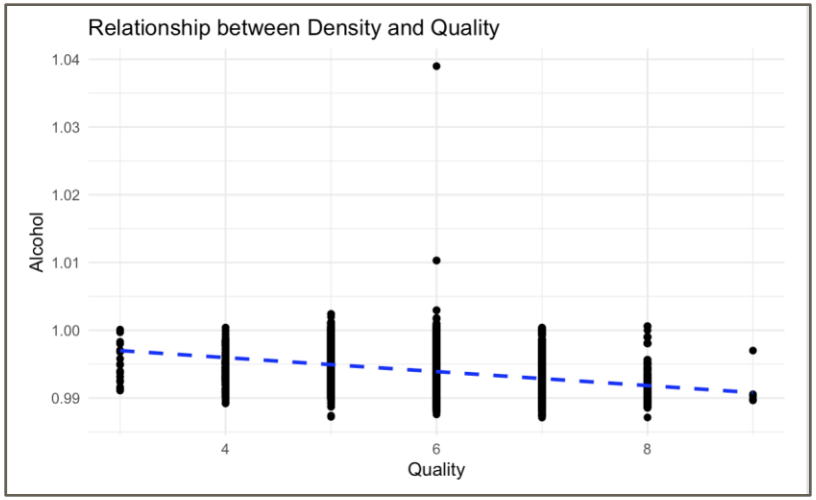


Density & Alcohol



1c)- Strong Relationships between variables and Quality





Alcohol and Quality (0.4354)

Density and Quality (-0.3055)

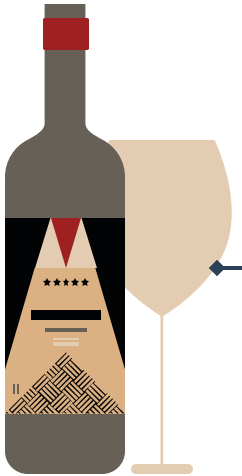
2a)- Correlation Analysis on all variables

	alcohol	chlorides	citric.acid	density	fixed.acidity	free.sulfur.dioxide
alcohol	1.00000000	-0.36013794	-0.07658043	-0.78056110	-0.12185088	-0.2503290067
chlorides	-0.36013794	1.00000000	0.11303445	0.25812554	0.02358781	0.1013509952
citric.acid	-0.07658043	0.11303445	1.00000000	0.15036777	0.28743931	0.0930179405
density	-0.78056110	0.25812554	0.15036777	1.00000000	0.26552503	0.2943267220
fixed.acidity	-0.12185088	0.02358781	0.28743931	0.26552503	1.00000000	-0.0492264843
free.sulfur.dioxide	-0.25032901	0.10135100	0.09301794	0.29432672	-0.04922648	1.0000000000
pH	0.12128283	-0.08921982	-0.16152679	-0.09239691	-0.42434340	0.0006720835
quality	0.43538304	-0.21075374	-0.01007915	-0.30548119	-0.11217222	0.0077470492
residual.sugar	-0.45139835	0.08974945	0.09460236	0.83884109	0.08797948	0.2991767998
sulphates	-0.01686942	0.01741200	0.06441847	0.07453693	-0.01649249	0.0596785709
total.sulfur.dioxide	-0.44873047	0.19936904	0.12202260	0.52956375	0.09134678	0.6156008531
volatile.acidity	0.06777176	0.07268643	-0.14847268	0.02732609	-0.02030097	-0.0969818015

	pH	quality	residual.sugar	sulphates	total.sulfur.dioxide
alcohol	0.1212828342	0.435383037	-0.45139835	-0.01686942	-0.448730472
chlorides	-0.0892198193	-0.210753740	0.08974945	0.01741200	0.199369044
citric.acid	-0.1615267923	-0.010079145	0.09460236	0.06441847	0.122022601
density	-0.0923969142	-0.305481195	0.83884109	0.07453693	0.529563747
fixed.acidity	-0.4243434038	-0.112172217	0.08797948	-0.01649249	0.091346778
free.sulfur.dioxide	0.0006720835	0.007747049	0.29917680	0.05967857	0.615600853
pH	1.0000000000	0.097291537	-0.19305111	0.15555281	0.003552351
quality	0.0972915370	1.000000000	-0.09492275	0.05424105	-0.174596855
residual.sugar	-0.1930511125	-0.094922745	1.00000000	-0.02608008	0.401274621
sulphates	0.1555528065	0.054241050	-0.02608008	1.00000000	0.133955384
total.sulfur.dioxide	0.0035523514	-0.174596855	0.40127462	0.13395538	1.000000000
volatile.acidity	-0.0346121859	-0.196657025	0.06411095	-0.03496529	0.089975165

	volatile.acidity
alcohol	0.06777176
chlorides	0.07268643
citric.acid	-0.14847268
density	0.02732609
fixed.acidity	-0.02030097
free.sulfur.dioxide	-0.09698180
pH	-0.03461219
quality	-0.19665702
residual.sugar	0.06411095
sulphates	-0.03496529
total.sulfur.dioxide	0.08997516
volatile.acidity	1.00000000

2b)- Strongest Correlation with Quality



	fixed_acidi	atile_acic	itric_aci	idual_su	chlorides	sulfur_di	sulfur_di	density	pH	sulphates	alcohol	quality
fixed_acidity	1											
volatile_acidity	-0.0203	1										
citric_acid	0.2874	-0.1485	1									
residual_sugar	0.088	0.0641	0.0946	1								
chlorides	0.0236	0.0727	0.113	0.0897	1							
free_sulfur_dioxide	-0.0492	-0.097	0.093	0.2992	0.1014	1						
total_sulfur_dioxide	0.0913	0.09	0.122	0.4013	0.1994	0.6156	1					
density	0.2655	0.0273	0.1504	0.8388	0.2581	0.2943	0.5296	1				
pH	-0.4243	-0.0346	-0.1615	-0.1931	-0.0892	0.0007	0.0036	-0.0924	1			
sulphates	-0.0165	-0.035	0.0644	-0.0261	0.0174	0.0597	0.134	0.0745	0.1556	1		
alcohol	-0.1219	0.0678	-0.0766	-0.4514	-0.3601	-0.2503	-0.4487	-0.7806	0.1213	-0.0169	1	
quality	-0.1122	-0.1967	-0.0101	-0.0949	-0.2108	0.0077	-0.1746	-0.3055	0.0973	0.0542	0.4354	1

Strongest Correlation with Quality :

- Alcohol (0.4354)

3a)- Linear Regression Analysis on Quality with all variables

```
lm_wine <-lm(quality ~.,
             data = wine_clean)
```

```
                        Estimate   Std. Error t value Pr(>|t|)
(Intercept)          150.4023750   18.8214891   7.991 1.66e-15 ***
alcohol                0.1945220    0.0242590   8.019 1.33e-15 ***
chlorides             -0.2463374    0.5475908  -0.450  0.65283
citric.acid            0.0085846    0.0959947   0.089  0.92875
density             -150.4893251   19.0925627  -7.882 3.95e-15 ***
fixed.acidity          0.0680313    0.0209167   3.252  0.00115 **
free.sulfur.dioxide    0.0036553    0.0008456   4.323 1.57e-05 ***
pH                     0.6769471    0.1056245   6.409 1.60e-10 ***
residual.sugar         0.0822458    0.0075373  10.912  < 2e-16 ***
sulphates              0.6376236    0.1004976   6.345 2.43e-10 ***
total.sulfur.dioxide  -0.0002729    0.0003789  -0.720  0.47137
volatile.acidity      -1.8897521    0.1143956 -16.519  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7508 on 4858 degrees of freedom
  (28 observations deleted due to missingness)
Multiple R-squared:  0.2831, Adjusted R-squared:  0.2815
F-statistic: 174.4 on 11 and 4858 DF,  p-value: < 2.2e-16
```

Significant Variables (8):

- **Alcohol, Density, Fixed Acidity, Free Sulfur Dioxide, pH, Residual Sugar, Sulphates, Volatile Acidity**

3b)- Variance Inflation Analysis (VIF) on model in 3a

`fixed acidity`	`volatile acidity`	`citric acid`
2.688538	1.142385	1.163777
`residual sugar`	chlorides	`free sulfur dioxide`
12.618623	1.236103	1.787738
`total sulfur dioxide`	density	pH
2.238406	28.208165	2.194498
sulphates	alcohol	
1.137704	7.706420	

Residual Sugar and Density have a VIF greater than 10

→ A VIF greater than 10 suggests high multicollinearity.

3c)- Linear Regression only with significant variables (8)

```
Coefficients:
                      Estimate  Std. Error t value Pr(>|t|)
(Intercept)        154.4057700  18.1215581   8.521  < 2e-16 ***
alcohol              0.1939417   0.0241207   8.040 1.11e-15 ***
density           -154.5895534  18.3663600  -8.417  < 2e-16 ***
fixed_acidity        0.0702599   0.0204804   3.431 0.000607 ***
free_sulfur_dioxide  0.0032810   0.0006778   4.841 1.33e-06 ***
pH                   0.6860435   0.1036235   6.621 3.97e-11 ***
residual_sugar       0.0836402   0.0072986  11.460  < 2e-16 ***
sulphates            0.6343620   0.1000864   6.338 2.54e-10 ***
volatile_acidity    -1.9116676   0.1100439 -17.372  < 2e-16 ***
---
```

Direction:

- Positive impact : alcohol, fixed acidity, free sulfur dioxide, pH, residual sugar, sulphates
- Negative impact : density(-154.6), volatile acidity(-1.9)

Neural network Data Preparation

Training/Test Data

```r
set.seed(42)
wine_index <- sample(nrow(wine_norm),
                     0.7 * nrow(wine_norm),
                     replace = FALSE)

wine_train <- wine_norm[wine_index,]
wine_test <- wine_norm[-wine_index,]
```

Function definition

```r
#functions definition
normalize <- function(x){return((x-min(x))/(max(x)-min(x)))}
denormalize <- function(y,x){return(y*(max(x)-min(x))+min(x))}
```

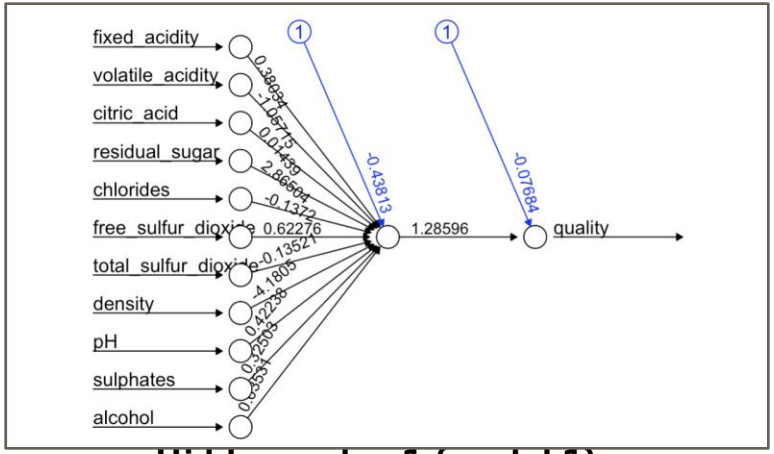
```r
head(wine_norm)
```

Description: df [6 × 12]

	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	chlorides <dbl>
1	0.3076923	0.1862745	0.2168675	0.30828221	0.1068249
2	0.2403846	0.2156863	0.2048193	0.01533742	0.1186944
3	0.4134615	0.1960784	0.2409639	0.09662577	0.1216617
4	0.3269231	0.1470588	0.1927711	0.12116564	0.1454006
5	0.3269231	0.1470588	0.1927711	0.12116564	0.1454006
6	0.4134615	0.1960784	0.2409639	0.09662577	0.1216617

4a)- Neural Network with quality using all variables

```r
set.seed(42)
wine_net <- neuralnet(quality ~.,
                      data = wine_train,
                      hidden = 1,
                      lifesign = "minimal",
                      linear.output = TRUE,
                      threshold = 0.05)
```



Hidden node =1 (model 1)

Accuracy = 0.5081178

```r
wine_test_wo_quality <- wine_test[, -ncol(wine_test)]
```

```r
wine_net.results <- neuralnet::compute(wine_net, wine_test_wo_quality)
```

```r
pred_quality_norm <- denormalize(wine_net.results$net.result, wine_norm$quality)
```

```r
actual_quality_denorm <- denormalize(wine_test$quality, wine_norm$quality)
```

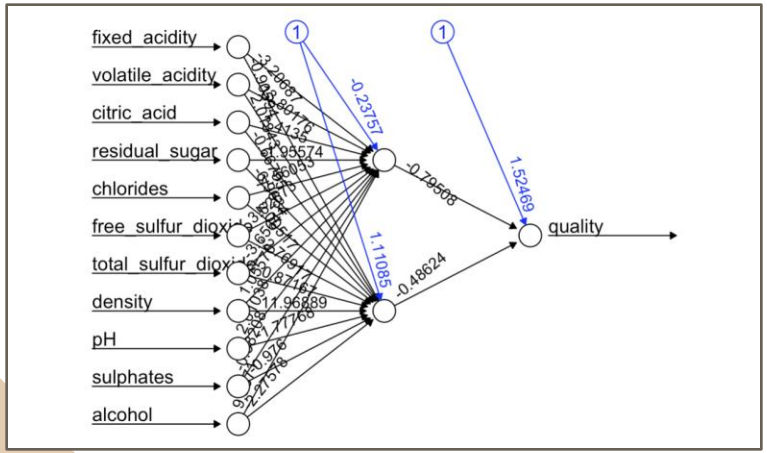
```r
node1_results <- data.frame(actual = actual_quality_denorm,
                            prediction = pred_quality_norm)
```

```r
cor(node1_results$actual, node1_results$prediction)
```

```
[1] 0.5081178
```

4a)- Neural Network with quality using all variables

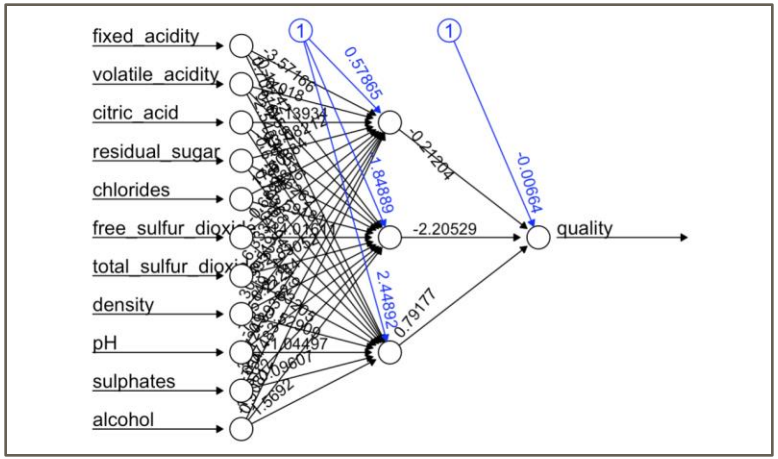
```r
set.seed(42)
wine_net2 <- neuralnet(quality ~.,
                       data = wine_train,
                       hidden = 2,
                       lifesign = "minimal",
                       linear.output = TRUE,
                       threshold = 0.05)
```



Hidden node =2 (model 2)

Accuracy = 0.538973

```r
set.seed(42)
wine_net3 <- neuralnet(quality ~.,
                       data = wine_train,
                       hidden = 3,
                       lifesign = "minimal",
                       linear.output = TRUE,
                       threshold = 0.05)
```

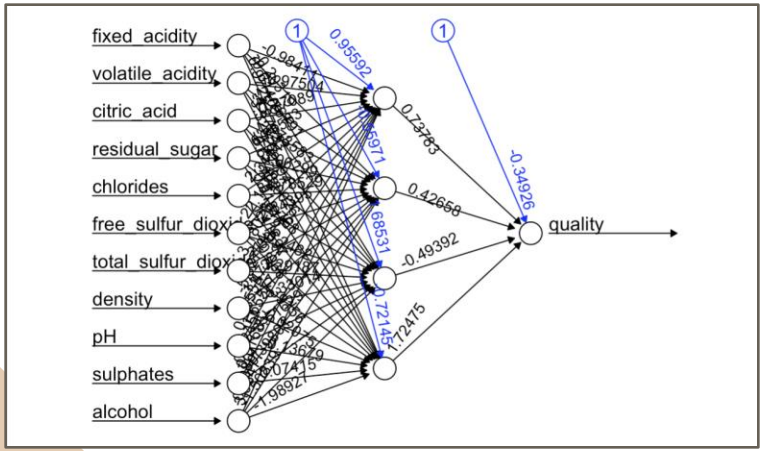


Hidden node =3 (model 3)

Accuracy = 0.5482323

4a)- Neural Network with quality using all variables

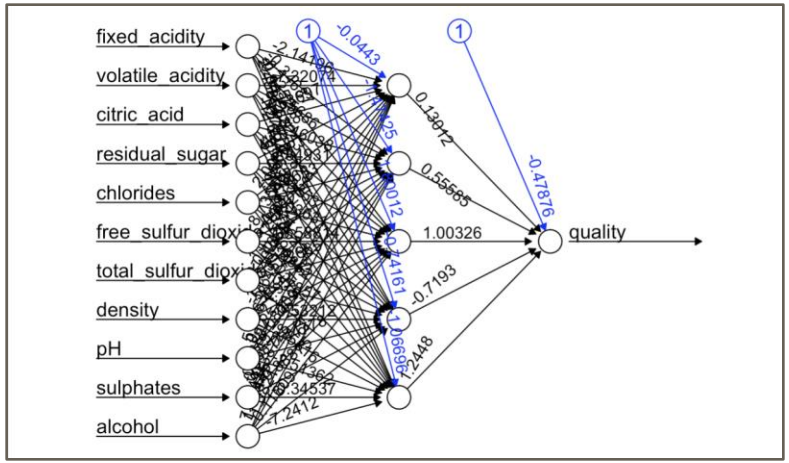
```r
set.seed(42)
wine_net4 <- neuralnet(quality ~.,
                       data = wine_train,
                       hidden = 4,
                       lifesign = "minimal",
                       linear.output = TRUE,
                       threshold = 0.05)
```



Hidden node =4 (model 4)

Accuracy = 0.5636129

```r
set.seed(42)
wine_net5 <- neuralnet(quality ~.,
                       data = wine_train,
                       hidden = 5,
                       lifesign = "minimal",
                       linear.output = TRUE,
                       threshold = 0.05)
```

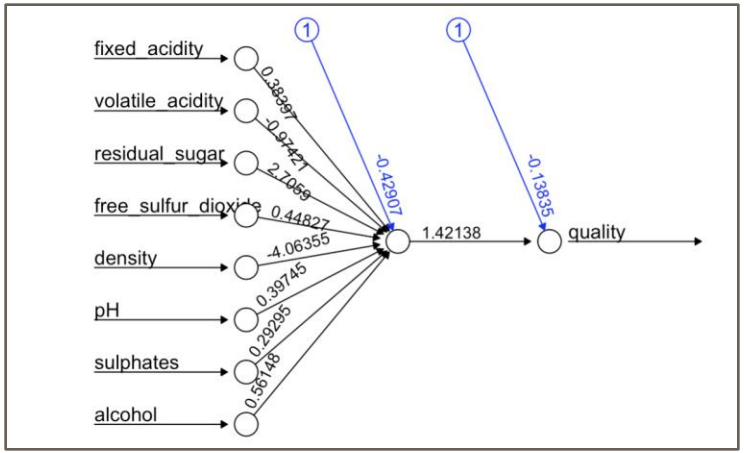


Hidden node =5 (model 5)

Accuracy = 0.5727752

4b)- Neural Network with quality using significant variables

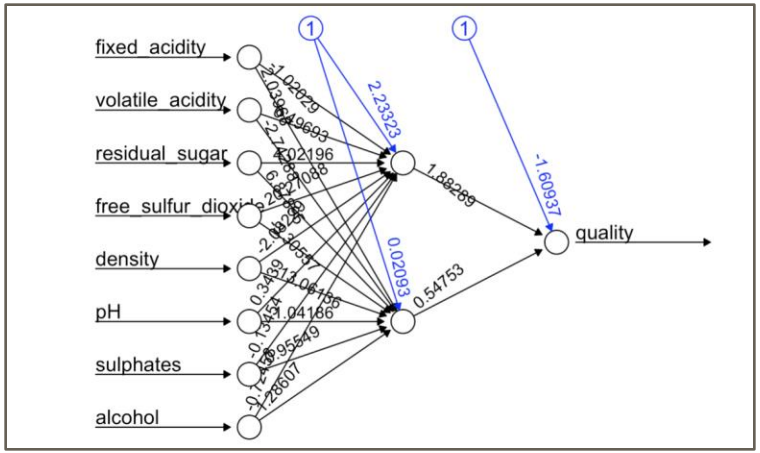
- Hidden node =1 (model 6)



Accuracy = 0.5099308

```
set.seed(42)
wine_net_w_sig <- neuralnet(quality ~ fixed.acidity + volatile.acidity + residual.sugar +
free.sulfur.dioxide + density + pH + sulphates + alcohol,
                            data = wine_train,
                            hidden = 1,
                            lifesign = "minimal",
                            linear.output = TRUE,
                            threshold = 0.05)
```

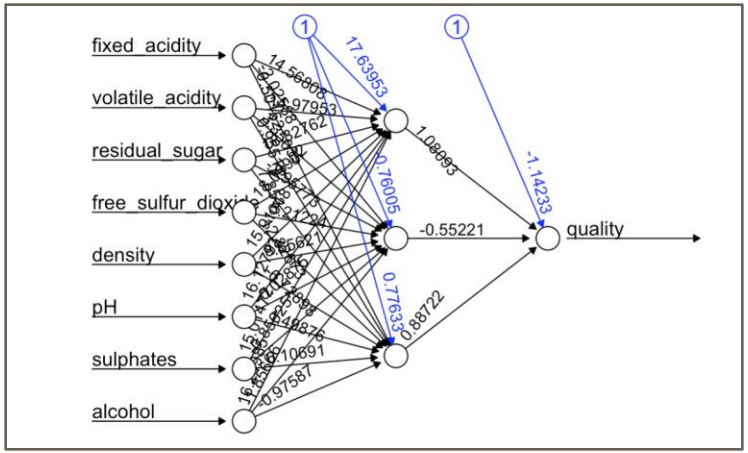
- Hidden node =2 (model 7)



Accuracy = 0.5404927

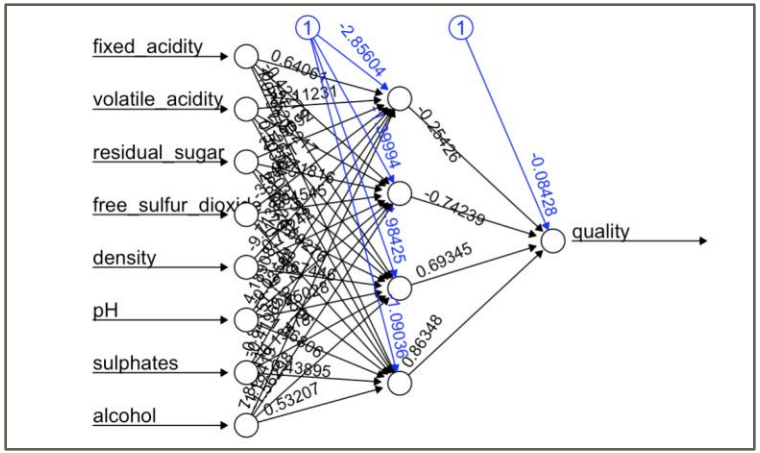
4a)- Neural Network with quality using significant variables

- Hidden node =3 (model 8)



Accuracy = 0.5431191

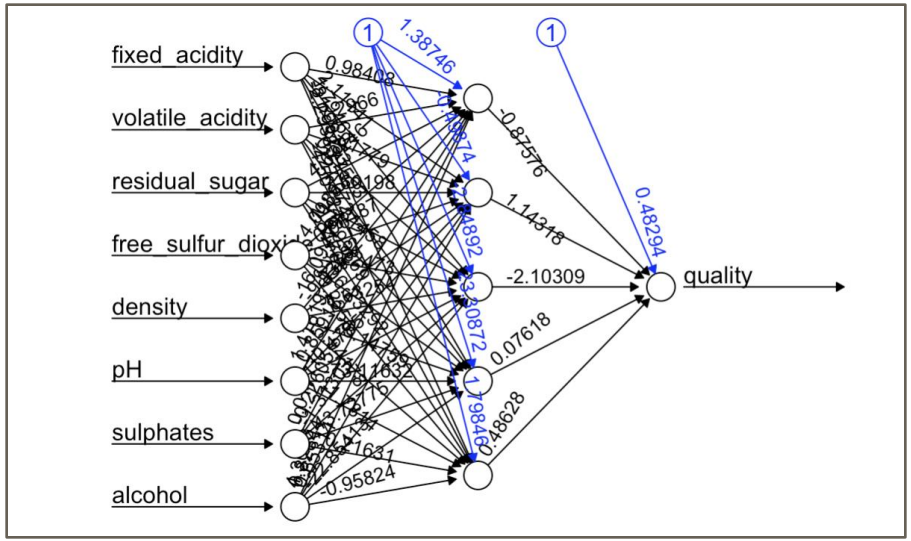
- Hidden node =4 (model 9)



Accuracy = 0.5631

4a)- Neural Network with quality using significant variables

- Hidden node =5 (model 10)



Accuracy = 0.5579772

Summary the accuracy of all models

	Variable	Hidden Node	Accuracy
Model 1	All	1	0.5081178
Model 2	All	2	0.538973
Model 3	All	3	0.5482323
Model 4	All	4	0.5636129
Model 5	All	5	0.5727752
Model 6	Significant Variable	1	0.5099308
Model 7	Significant Variable	2	0.5404927
Model 8	Significant Variable	3	0.5431191
Model 9	Significant Variable	4	0.5631
Model 10	Significant Variable	5	0.5579772

5)- List of Lessons Learned

a. Did the correlation analysis give insight into the results later found in the linear regression?

→ No. But it give insight into the results of VIF.

b. Which linear regression model helped in identifying the best neural network?

→ The models using only significant variables do not show a substantial difference in accuracy compared to models using all variables.

5)- List of Lessons Learned

c. Would the VIF analysis lead you to question your results?

→ Yes,

The variable "density" has a VIF of 28.21, which suggests that it can be highly linearly predicted by other variables, indicating strong multicollinearity.
Similarly, "residual.sugar" has a VIF of 12.62.
Additionally, correlation analysis may reveal that "density" and "residual.sugar" are strongly correlated with each other, which might explain their high VIF values.

Therefore, we may consider removing one of these variables, as the remaining variables may already capture the information they contribute to the neural network model.

Since neural networks are nonlinear models, multicollinearity may not directly harm model performance. However, its actual impact is uncertain—thus, it is important to conduct experiments to test whether removing these variables improves the model's accuracy.

THANK!

Wine Quality Analysis

GROUP 6:
ChihHao Yuan
Ching Yu Hsu