What affects the quality of wine?

Zhouyi Xue

ISYE-4031-T09 Section T09 (TU/TH, 9:30AM – 10:45PM)

4/28/2022

Summary

We all drink wine. However, what affects the quality of wine? In fact, there are a large number of factors could affect its quality. How do we know which factor is the most significant? Even as a wine producer, which factor should be paying attention to in order to produce high quality wine. With statistical analysis, you can find some answers.

Introduction

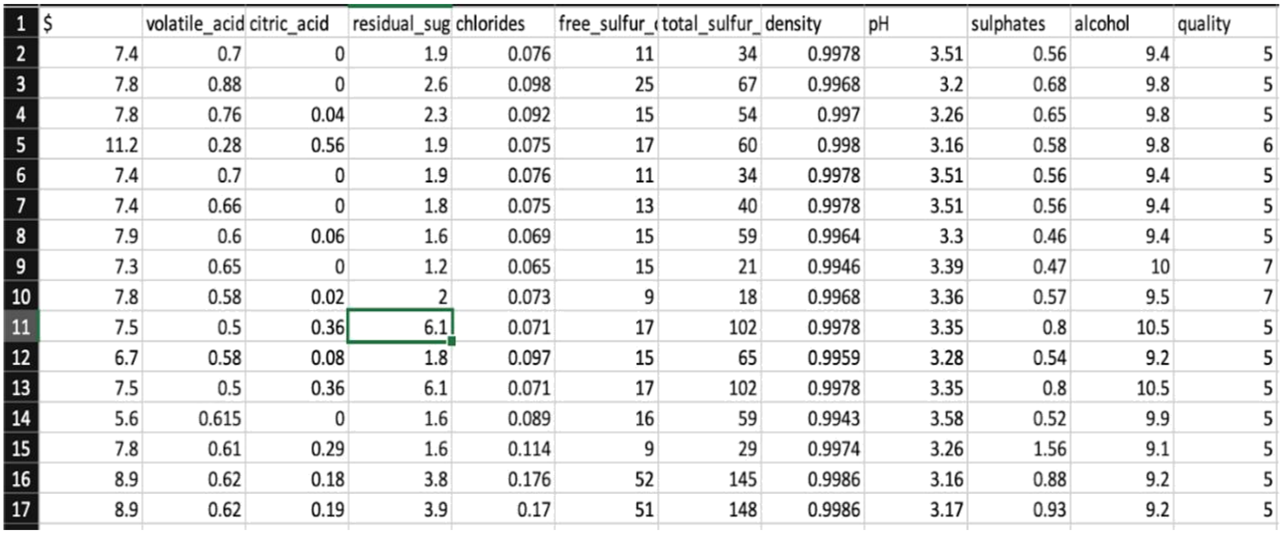
Wine is an essential product in the life of the majority of people. It is one of the most popular go-to products for relaxation after a long day of work. We noticed that people are willing to pay a premium price for a great quality wine. Some wines sell for hundreds of thousands of dollars a bottle, while cheaper options can be as low as two or three dollars a bottle. For this project, our team wants to find out what makes a good wine. We believe that cracking down the secret code for making good quality wine will bring enormous profit in this industry. This project will utilize linear regression analysis to decode the variables that affect the quality of wine.

Goals

For this project, we want to make a linear regression model that predicts the quality of wine. We want to analyze different variables and find out which variables have stronger correlations with our variables as well as the quality of wine.

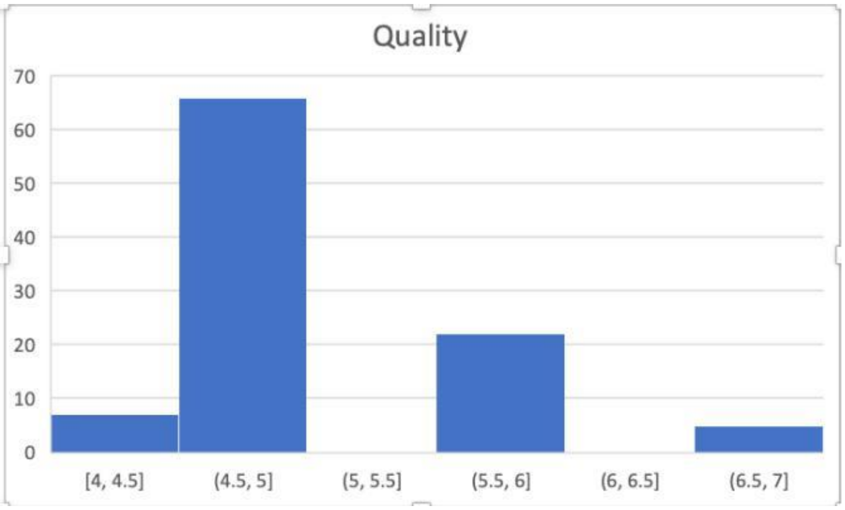
Data Description

In order to make our regression model, we will be using data set collected by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.

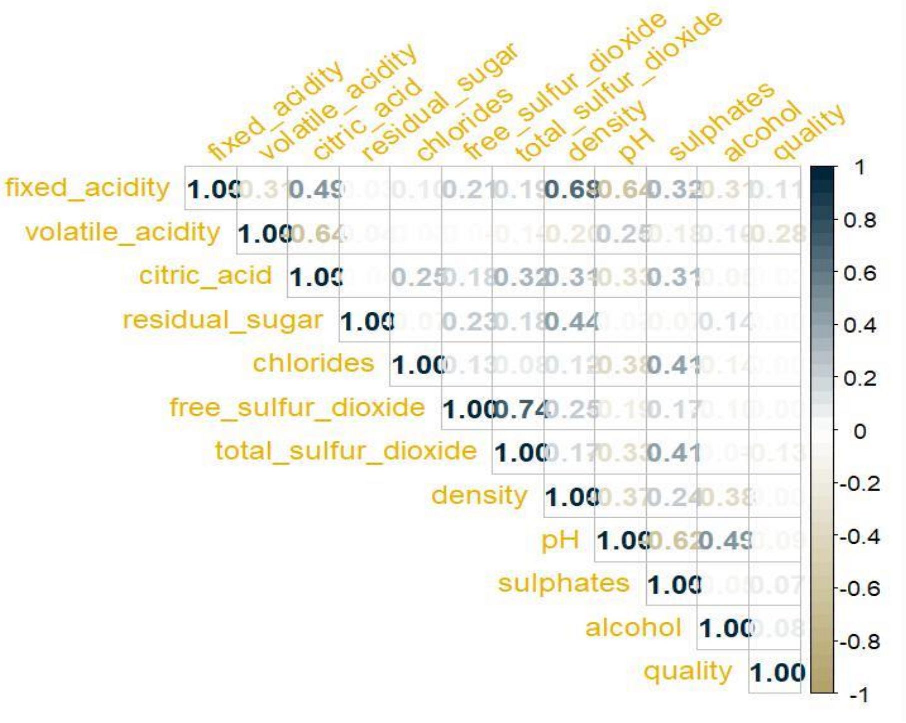


1	$	volatile_acid	citric_acid	residual_sug	chlorides	free_sulfur_(	total_sulfur_	density	pH	sulphates	alcohol	quality
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
3	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
4	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
5	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
6	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
8	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
9	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
10	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
11	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
12	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
13	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
14	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
15	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
16	8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
17	8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5

We have the data of a total of 100 wines. There are a total of 11 input variables and one output variable. The output variable, quality of the wine, is rated between score (0 to 10). All of our variables are quantitative. Below is an example of the data set. We made an initial analysis of the quality ratings, below is the summary of the quality: As we can see, around 60 percent of the wine has quality of 5, and all of our wines have quality ratings between 4 and 7.



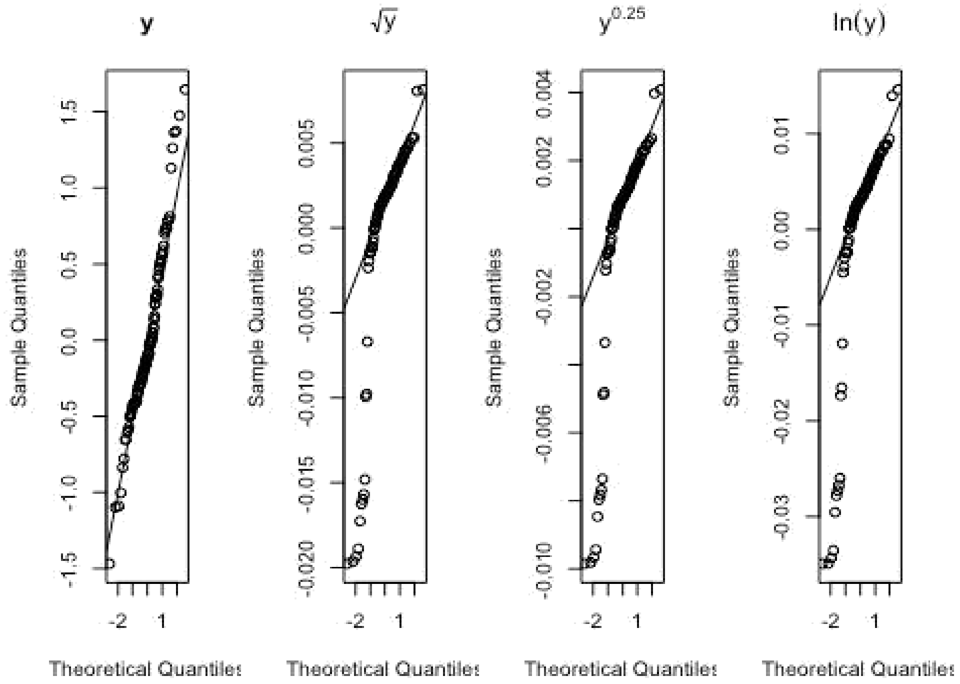
Correlation plot for all variables



We created a correlation ratio graph between each variable as shown above. Steps are referred to Appendix 1. As we can see from the correlation ratio graph, a high correlation ratio representing a strong relationship between two variables. For example, volatile acidity has highest the highest correlation ratio with quality. This means that the quality of wine is affected by its volatile acidity more significantly than any other variables based on this plot. We also found the ratio for density is nearly 0. As wine makers, perhaps density is not the factor that they need to improve on.

Transformation

Generate and compare the normal probability plots of residuals by using 3 transformations and the original quality. Discuss which one seems the best model for the normality assumption. Detailed steps and R code are referred to Appendix 2.



We want to test if our linear regression model is appropriate. From the R output, the original y function seems the best model for the normality assumption since more points fall along with the straight line. Therefore, we will keep using this model for future analysis.

Analysis of outliers and influential points

Here we want to further analyze our model. Make our data clean and find those data not representative. In order to achieve this, we generate the leverage values and outliers.

```
n=length(quality)
k=11
leverage_cutoff=2*(k+1)/n
leverage_cutoff

## [1] 0.24

hat_i=hatvalues(model)
hat_i[hat_i>leverage_cutoff]

 ##          34         39         46         57         82         95         96
## 0.4758528 0.2851497 0.3192137 0.2450082 0.3343995 0.3490665

0.3241504 length(hat_i[hat_i>leverage_cutoff])

## [1] 7
```

From the output, there are eight outliers, 34,39,46,57,82,95,96 are outliers with respect to their x values since their leverage values are larger than that of the cutoff.

b. Find the studentized deleted residuals and outliers.

```
e=resid(model)
SSE=anova(model)[k+1,2]
SDR=e*sqrt((n-k-2)/(SSE*(1-hat_ i)-e^2))
SDR[abs(SDR)>qt(0.995,n-(k+2))]
```

```
##        46        63
## -3.096188 2.874433
```

```
SDR[abs(SDR)>qt(1-0.025,n-(k+2))]
```

```
 ##         8         9        17        38        46        47        63
##  2.474222  2.582513  2.512125  2.199513 -3.096188  2.186182  2.874433
```

From R output, we can see that there is some evidence that 8,9,17,38,46,47,63 are outliers with respect to their y value because its studentized deleted residual is larger than that of 0.025. We can say that there is strong evidence that 46 and 63 are outliers because the studentized deleted residual value are larger than that of 0.005.

c. Find the Cook's distance identify which points are influential.

```
cooksd=cooks.distance(model)
cooksd[cooksd>qf(0.5,k+1,n-k-1)]
```

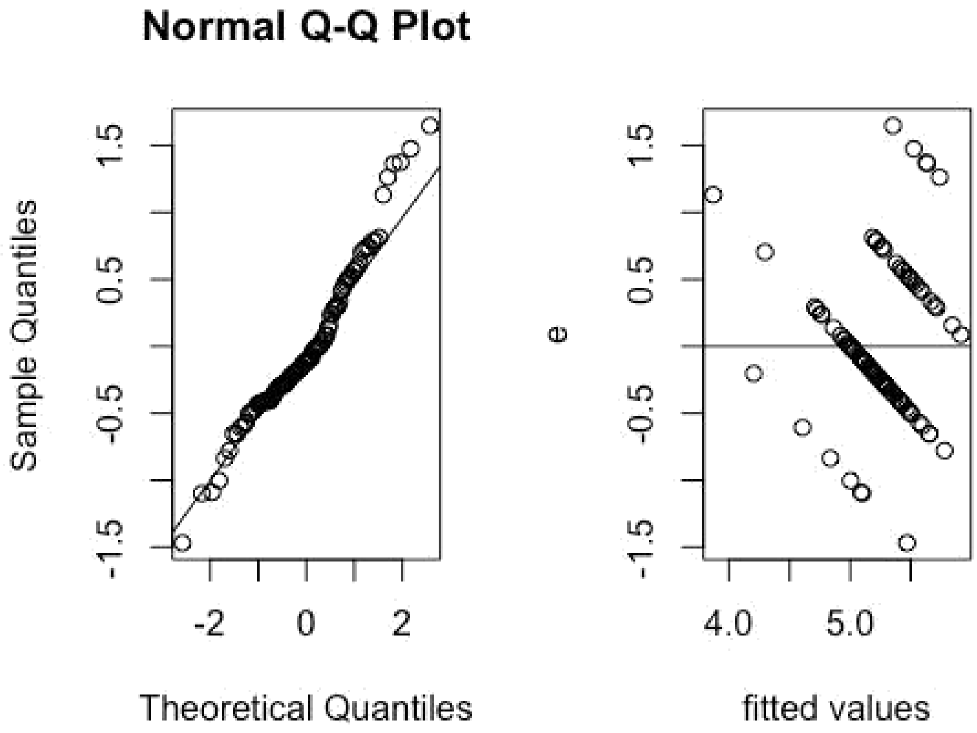
```
## named numeric(0)
```

From R output, we can see that there are no influential points.

Residuals analysis

In this step, we want to see how close each data point is vertically from the graph of the prediction equation of the model. This shows if our data fits the linear model or not.

Fine Normal probability plot of the residuals, and residual plots of e vs fitted values. Steps refer to Appendix 6.



From the plot, we can see most of data is very close to the linear model, which means our linear model is representative supported by these data.

Assumptions Check

Normality assumption:

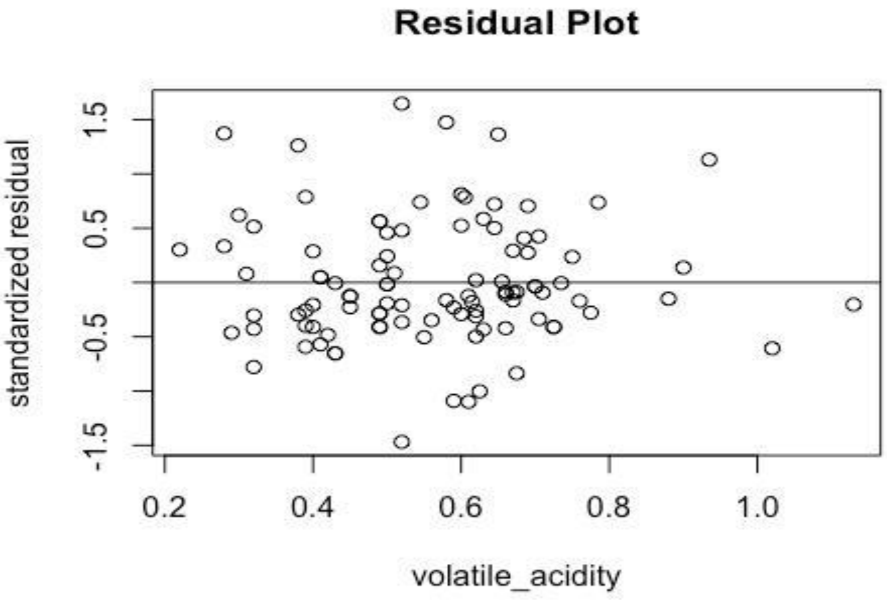
From the normal probability plot, we can see that the model is normal. Normality test:

```
library(nortest)
ad.test(resid(model))
```

```
##
##  Anderson-Darling normality test
##
## data:  resid(model)
## A = 1.5339, p-value = 0.0005607
```

From the Anderson-Darling normality test, we can see that the p value is less than 0.001, meaning there are strong evidence to suggest the model is not normal. This means that the normality assumption is failed.

Constant variance assumption:



The model seems to meet the constant variance assumption (Steps refer to Appendix 7)

Model selection process

Steps method for model selection process (Refer to Appendix 3)

From the both steps of model selection process, we can see the best model is with independent variables volatile_acidity, citric_acid, fixed_acidity, alcohol.

Checks for multicollinearity

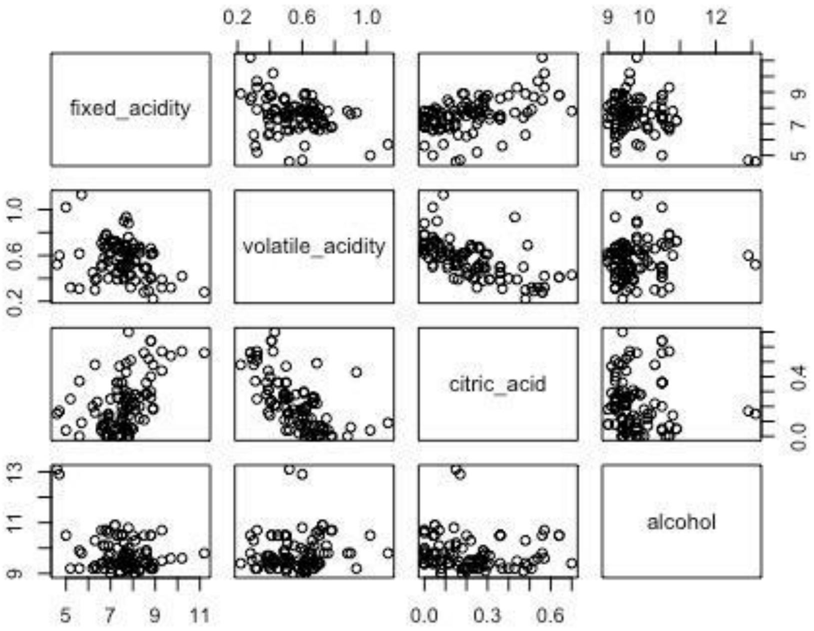
Refer to Appendix 4, From the correlation matrix, we can see that no 2 variables have corelation greater than +-0.9.

```
library(car)

vif(model)

##         fixed_acidity     volatile_acidity          citric_acid
##              6.211184             1.930201             2.827841
##        residual_sugar            chlorides  free_sulfur_dioxide
##              2.414384             1.814310             3.065900
## total_sulfur_dioxide               density                   pH
##              3.930886             5.137787             6.291721
##             sulphates              alcohol
##              2.681954             2.833822
```

```
mean(vif(model))

## [1] 3.558181
```

From the Variance Inflation Factor (VIF) output, we can see that fixed_acidity, density, and pH are greater than 4, which is large, indecating multicollinearity. In addition, the mean of vifs is greater than 1, another indication of multicollinearity.

Correlation plot for best model variables



This scatter plot shows the correlation between the best model variables we found. From the plot, the relationship between each variable is discovered. As we can see, most of these don't have a correlation. However, plot between critic acid and fixed acidity shows that critic acid will likely increase when fixed acidity increases. Another plot of citric acid and volatile acidity shows with citric acid increases, volatile acidity will decrease.

Conclusion:

From our R output and analysis, we can conclude that the quality of wine is major affected by fixed acidity, volatile acidity, citric acid and alcohol amount. Among this, volatile acidity is the most significant factor that affects the quality of wine. Also, the graph that we made showing that the data fits the linear regression model, which means our model is to appropriate analyze the relationship between these factors.

From this statistical analysis, there is one suggestion for winemakers. Statistical evidence supports the statement that volatile acidity plays a very important role in making wine. If you want to produce good wine, you definitely should consider decreasing volatile acidity.

Appendix 1:

Import file

```
input = read.csv(file = "wine.csv", sep=',')
attach(input)
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

```
new_input =cor(input)
TechGold =rgb(179,163,105,max=255)
TechBlue =rgb(0,38,58,max=255)
BuzzGold =rgb(234,179,0,max=255)
tech_col <-colorRampPalette(c(TechGold,"white", TechBlue))(40)
corrplot(new_input,method="number",type="upper",col =tech_col,tl.col =B
uzzGold,tl.srt=40)
```

Appendix 2:

```
input = read.csv(file = "wine.csv", sep=',')
attach(input)
```

```
## The following objects are masked from input (pos = 4):
##
##     alcohol, chlorides, citric_acid, density, fixed_acidity,
##     free_sulfur_dioxide, pH, quality, residual_sugar, sulphate
s,
##     total_sulfur_dioxide, volatile_acidity
```

```
y=quality
y1=sqrt(y)
y2=y^(0.25)
y3=log(y)
model=lm(quality~.,data = input)
model1=lm(y1~., data = input)
model2=lm(y2~., data = input)
model3=lm(y3~., data = input)
par(mfrow=c(1,4))
qqnorm(resid(model),main="y")
qqline(resid(model))
qqnorm(resid(model1),main=bquote(sqrt(y)))
```

```
qqline(resid(model1))
qqnorm(resid(model2),main=bquote(y^0.25))
qqline(resid(model2))
qqnorm(resid(model3),main=bquote(ln(y)))
qqline(resid(model3))
```

Appendix 3：

```
full.lm = lm(quality~., data = input)
min.lm = lm(quality~ 1, data = input)
step_both = step(min.lm, list(upper = full.lm), direction = "both
")
```

```
## Start:  AIC=-7.32
## quality ~ 1
## 
##                        Df Sum of Sq    RSS     AIC
## + volatile_acidity      1   2.48066 10.069 -9.7247
## + residual_sugar        1   1.89622 10.654 -8.5964
## + alcohol               1   1.77404 10.776 -8.3683
## + density               1   1.19546 11.354 -7.3223
## <none>                              12.550 -7.3202
## + citric_acid           1   1.06887 11.481 -7.1005
## + fixed_acidity         1   0.74582 11.804 -6.5456
## + total_sulfur_dioxide  1   0.22061 12.329 -5.6749
## + pH                    1   0.14559 12.404 -5.5536
## + sulphates             1   0.10769 12.442 -5.4926
## + free_sulfur_dioxide   1   0.04984 12.500 -5.3998
## + chlorides             1   0.01184 12.538 -5.3391
## 
## Step:  AIC=-9.72
## quality ~ volatile_acidity
## 
##                        Df Sum of Sq     RSS      AIC
## + residual_sugar        1   2.15334  7.9160 -12.5369
## + density               1   1.55394  8.5154 -11.0771
## <none>                              10.0693  -9.7247
## + alcohol               1   0.91842  9.1509  -9.6375
## + total_sulfur_dioxide  1   0.66458  9.4048  -9.0903
## + citric_acid           1   0.51671  9.5526  -8.7783
## + sulphates             1   0.39468  9.6747  -8.5244
## + chlorides             1   0.26323  9.8061  -8.2545
## + fixed_acidity         1   0.03083 10.0385  -7.7861
## + pH                    1   0.02468 10.0447  -7.7738
## + free_sulfur_dioxide    1   0.00928 10.0601   -7.7432
## - volatile_acidity       1   2.48066 12.5500   -7.3202
##
## Step:  AIC=-12.54
## quality ~ volatile_acidity + residual_sugar
##
##                         Df Sum of Sq     RSS       AIC
## + alcohol                1   2.39753  5.5185  -17.7526
## <none>                                7.9160  -12.5369
## + chlorides              1   0.53404  7.3820  -11.9339
## + density                1   0.37065  7.5454  -11.4960
## + sulphates              1   0.30674  7.6093  -11.3273
## + free_sulfur_dioxide    1   0.21234  7.7037  -11.0807
## + citric_acid            1   0.04546  7.8705  -10.6521
## + fixed_acidity          1   0.04036  7.8756  -10.6391
## + pH                     1   0.01210  7.9039  -10.5675
## + total_sulfur_dioxide   1   0.00009  7.9159  -10.5371
## - residual_sugar         1   2.15334 10.0693   -9.7247
## - volatile_acidity       1   2.73777 10.6538   -8.5964
##
## Step:  AIC=-17.75
## quality ~ volatile_acidity + residual_sugar  + alcohol
##
##                         Df Sum of Sq    RSS       AIC
## <none>                               5.5185  -17.7526
## + free_sulfur_dioxide    1    0.3201 5.1984  -16.9477
## + fixed_acidity          1    0.2291 5.2894  -16.6006
## + pH                     1    0.2155 5.3029  -16.5494
## + citric_acid            1    0.1299 5.3886  -16.2289
## + density                1    0.0107 5.5077  -15.7916
## + sulphates              1    0.0066 5.5119  -15.7766
## + chlorides              1    0.0010 5.5175  -15.7561
## + total_sulfur_dioxide   1    0.0001 5.5184  -15.7528
## - volatile_acidity       1    1.4554 6.9739  -15.0712
## - alcohol                1    2.3975 7.9160  -12.5369
## - residual_sugar         1    3.6324 9.1509   -9.6375
```

Appendix 4:

```
cor(input)
```

```
##                            quality fixed_acidity volatile_acidit
y citric_acid
## quality                1.00000000    0.24377815      -0.4445916
9    0.2918366
## fixed_acidity          0.24377815    1.00000000      -0.4486925
9    0.5833595
## volatile_acidity      -0.44459169   -0.44869259       1.0000000
0   -0.8762944
## citric_acid            0.29183656    0.58335953      -0.8762944
3    1.0000000
## residual_sugar        -0.38870744    0.04047078      -0.0559330
1    0.2192762
## chlorides             -0.03072008    0.14498704      -0.2465920
5    0.3628960
## free_sulfur_dioxide    0.06301540    0.39972619      -0.0807765
3    0.2363574
## total_sulfur_dioxide -0.13258244     0.32306805      -0.2080562
3    0.3939692
## density               -0.30863551    0.58455837      -0.0937810
9    0.3109357
## pH                    -0.10770755   -0.61265501       0.3361931
1   -0.5722523
## sulphates             -0.09263497    0.18864402      -0.1837255
7    0.4863942
## alcohol                0.37597575   -0.02111093      -0.2577592
6    0.3202989
##                      residual_sugar    chlorides free_sulfur_di
oxide
## quality                  -0.38870744 -0.03072008          0.063
01540
## fixed_acidity            0.04047078   0.14498704          0.399
72619
## volatile_acidity         -0.05593301 -0.24659205         -0.080
77653
## citric_acid              0.21927623   0.36289603          0.236
35739
## residual_sugar            1.00000000 -0.12480297          0.242
62167
## chlorides               -0.12480297   1.00000000          0.223
96366
## free_sulfur_dioxide      0.24262167   0.22396366          1.000
00000
## total_sulfur_dioxide     0.55947334   0.17122360          0.885
06505
## density                  0.48951454   0.07671054          0.354
76372
```

```
## pH                       -0.04902412 -0.67003096         -0.428
84854
## sulphates                 0.06051410  0.64499486          0.184
73153
## alcohol                   0.34188292 -0.37769681          0.040
12940
##                         total_sulfur_dioxide     density
 pH    sulphates
## quality                            -0.1325824 -0.30863551 -0.10770
755 -0.09263497
## fixed_acidity                       0.3230680  0.58455837 -0.61265
501   0.18864402
## volatile_acidity                   -0.2080562 -0.09378109  0.33619
311 -0.18372557
## citric_acid                         0.3939692  0.31093573 -0.57225
229   0.48639415
## residual_sugar                      0.5594733  0.48951454 -0.04902
412   0.06051410
## chlorides                           0.1712236  0.07671054 -0.67003
096   0.64499486
## free_sulfur_dioxide                 0.8850650  0.35476372 -0.42884
854   0.18473153
## total_sulfur_dioxide                1.0000000  0.43263151 -0.39377
562   0.20681753
## density                             0.4326315  1.00000000 -0.27475
789   0.28869248
## pH                                 -0.3937756 -0.27475789  1.00000
000 -0.56325618
## sulphates                           0.2068175  0.28869248 -0.56325
618   1.00000000
## alcohol                             0.2418875 -0.17461844  0.20430
212 -0.21064330
##                            alcohol
## quality                 0.37597575
## fixed_acidity          -0.02111093
## volatile_acidity       -0.25775926
## citric_acid             0.32029891
## residual_sugar          0.34188292
## chlorides              -0.37769681
## free_sulfur_dioxide     0.04012940
## total_sulfur_dioxide    0.24188746
## density                -0.17461844
## pH                      0.20430212
## sulphates              -0.21064330
## alcohol                 1.00000000
```

Appendix 5:

```
model = lm(quality~fixed_acidity+volatile_acidity+citric_acid+residual_
sugar+chlorides+free_sulfur_dioxide+total_sulfur_dioxide+density+pH+sul
phates+alcohol, data = input)
pairs(~fixed_acidity+volatile_acidity+citric_acid + alcohol, data = inp
ut)
```

Appendix 6:

```
par(mfrow=c(1,2))
e=resid(model)
qqnorm(e)
qqline(e)
fittedvalues=fitted(model)
plot(fittedvalues,e,xlab="fitted values",ylab="e")
abline(h=0)
```

Appendix 7:

```
e = resid(model)
plot(volatile_acidity,e, xlab="volatile_acidity",ylab = "standardized r
esidual",main = "Residual Plot")
abline(h=0)
```

Reference:

1. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties.
In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

   Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

2. *Example of multiple linear regression in R*. Data to Fish. (2020, April 3). Retrieved April 28, 2022, from https://datatofish.com/multiple-linear-regression-in-r/

3. robk@statmethods.net, R. K.-. (n.d.). Multiple (linear) regression. Quick-R: Multiple Regression. Retrieved April 28, 2022, from https://www.statmethods.net/stats/regression.html

4. *Sign in*. RPubs. (n.d.). Retrieved April 28, 2022, from https://rpubs.com/iabrady/residual-analysis