Data Analytics
Assignment -3

Name : R Siva Girish SRN : PES1201700159

Dataset :White Wine Quality Prediction

Dataset link :

https://www.kaggle.com/sgus1318/winedata#winequality_white.csv

❖ About the Dataset

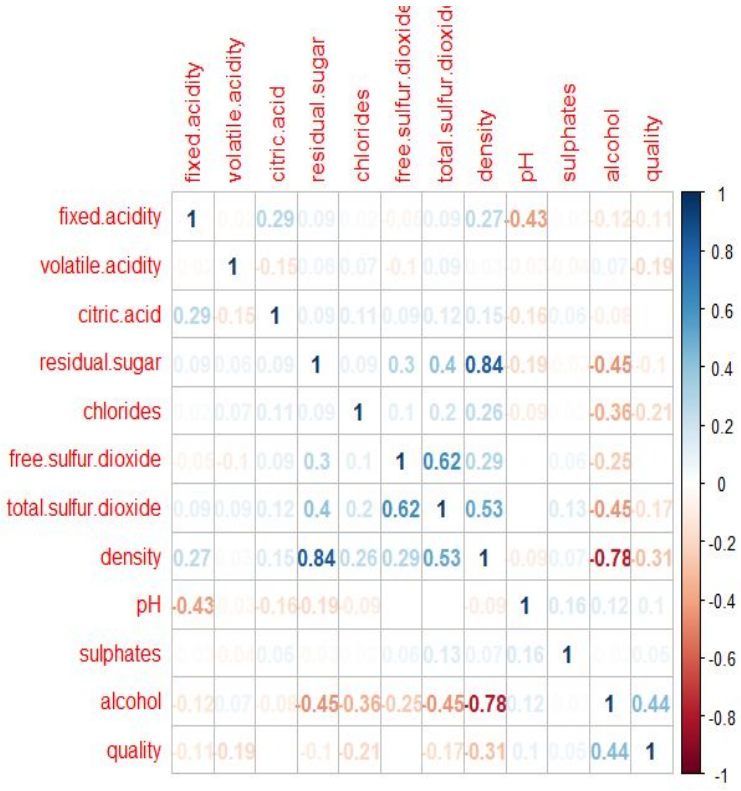
➢ Dataset contains 4898 rows and 12 columns

```
> head(wine)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol quality
1           7.0             0.27        0.36           20.7     0.045                  45                  170  1.0010 3.00      0.45     8.8       6
2           6.3             0.30        0.34            1.6     0.049                  14                  132  0.9940 3.30      0.49     9.5       6
3           8.1             0.28        0.40            6.9     0.050                  30                   97  0.9951 3.26      0.44    10.1       6
4           7.2             0.23        0.32            8.5     0.058                  47                  186  0.9956 3.19      0.40     9.9       6
5           7.2             0.23        0.32            8.5     0.058                  47                  186  0.9956 3.19      0.40     9.9       6
6           8.1             0.28        0.40            6.9     0.050                  30                   97  0.9951 3.26      0.44    10.1       6
>
```

❖ Multivariate Analysis

- ➢ Multivariate Analysis deals with the statistical analysis of data collected on more than one dependent variable.
- ➢ These variables may be correlated with each other, and their statistical dependence is often taken into account when analyzing such data.
- ➢ While performing multivariate analysis on variance (MANOVA) we get important parameters known as pillai score.
- ➢ Based on the Significance of the pillai score we can either choose to keep the independent variable or not but correlation also has to be checked.
- ➢ Usually whenever we eliminate a variable from our model based on it's pillai score we tend to observe an increase in the value of adjusted R-Square.
- ➢ If that's not the case then the multivariate regression model is not very useful.

❖ Correlation Plot



- ❏ Based on the correlation plot we can pick our dependent variables.
- ❏ Since the goal is to predict the quality of wine we fix that as one of the dependent variables and alcohol content has a higher correlation with quality compared to the others.
- ❏ Therefore our dependent variables are alcohol and quality.

Code For creating model :

```
mymodel<-lm(cbind(wine$alcohol,wine$quality) ~ wine$fixed.acidity +
wine$volatile.acidity + wine$citric.acid + wine$residual.sugar +
wine$chlorides + wine$free.sulfur.dioxide + wine$total.sulfur.dioxide +
wine$density + wine$pH + wine$sulphates , data=wine)
summary(mymodel)
coef(mymodel)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4021 -0.2534 -0.0353  0.2189 16.1786

Coefficients:
                              Estimate Std. Error  t value Pr(>|t|)
(Intercept)                  6.909e+02  5.066e+00  136.367  < 2e-16 ***
wine$fixed.acidity           5.210e-01  9.819e-03   53.061  < 2e-16 ***
wine$volatile.acidity        8.504e-01  6.609e-02   12.866  < 2e-16 ***
wine$citric.acid             3.721e-01  5.631e-02    6.608 4.30e-11 ***
wine$residual.sugar          2.427e-01  2.777e-03   87.409  < 2e-16 ***
wine$chlorides              -2.023e-01  3.228e-01   -0.627  0.53085
wine$free.sulfur.dioxide    -3.461e-03  4.961e-04   -6.976 3.44e-12 ***
wine$total.sulfur.dioxide    6.472e-04  2.231e-04    2.901  0.00374 **
wine$density                -6.983e+02  5.208e+00 -134.088  < 2e-16 ***
wine$pH                      2.461e+00  5.132e-02   47.950  < 2e-16 ***
wine$sulphates               1.022e+00  5.745e-02   17.791  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4437 on 4887 degrees of freedom
Multiple R-squared:  0.8702,    Adjusted R-squared:   0.87
F-statistic:  3278 on 10 and 4887 DF,  p-value: < 2.2e-16


Response wine$quality :

Call:
lm(formula = `wine$quality` ~ wine$fixed.acidity + wine$volatile.acidity +
    wine$citric.acid + wine$residual.sugar + wine$chlorides +
    wine$free.sulfur.dioxide + wine$total.sulfur.dioxide + wine$density +
    wine$pH + wine$sulphates, data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7257 -0.4926 -0.0425  0.4668  5.9997

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.839e+02  8.633e+00  32.880  < 2e-16 ***
wine$fixed.acidity           1.663e-01  1.673e-02   9.940  < 2e-16 ***
wine$volatile.acidity       -1.699e+00  1.126e-01 -15.082  < 2e-16 ***
wine$citric.acid             9.408e-02  9.596e-02   0.980 0.326893
wine$residual.sugar          1.284e-01  4.732e-03  27.145  < 2e-16 ***
wine$chlorides              -2.864e-01  5.500e-01  -0.521 0.602574
wine$free.sulfur.dioxide     3.063e-03  8.454e-04   3.624 0.000294 ***
wine$total.sulfur.dioxide   -1.605e-04  3.802e-04  -0.422 0.672838
wine$density                -2.854e+02  8.875e+00 -32.158  < 2e-16 ***
wine$pH                      1.162e+00  8.746e-02  13.292  < 2e-16 ***
wine$sulphates               8.292e-01  9.791e-02   8.470  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7562 on 4887 degrees of freedom
Multiple R-squared:  0.2725,    Adjusted R-squared:  0.271
F-statistic:   183 on 10 and 4887 DF,  p-value: < 2.2e-16
```

- Based on the summary we notice that the R square value of alcohol content is around 87% which is very good but the R squared value for quality is around 27% which is not acceptable.
- So we perform a manova test to infer whether we can drop a few variables from our model and thereby increase its efficiency.

Code For Multivariate Analysis on Variance:

```
Mymodelfit<-manova(mymodel)
summary(Mymodelfit)
```

```
> m1<-manova(mymodel)
> summary(m1)
                          Df  Pillai approx F num Df den Df    Pr(>F)
wine$fixed.acidity         1 0.10857    297.5      2   4886 < 2.2e-16 ***
wine$volatile.acidity      1 0.08827    236.5      2   4886 < 2.2e-16 ***
wine$citric.acid           1 0.00848     20.9      2   4886 9.283e-10 ***
wine$residual.sugar        1 0.60536   3747.4      2   4886 < 2.2e-16 ***
wine$chlorides             1 0.46048   2085.1      2   4886 < 2.2e-16 ***
wine$free.sulfur.dioxide   1 0.06284    163.8      2   4886 < 2.2e-16 ***
wine$total.sulfur.dioxide  1 0.30648   1079.6      2   4886 < 2.2e-16 ***
wine$density               1 0.75912   7699.0      2   4886 < 2.2e-16 ***
wine$pH                    1 0.34450   1284.0      2   4886 < 2.2e-16 ***
wine$sulphates             1 0.06837    179.3      2   4886 < 2.2e-16 ***
Residuals               4887
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

- Based on the manova test we obtain the pillai scores for all the values.
- We notice that free.Sulphur.dioxide has the least significant pillai score as well as its correlation with both alcohol as well as quality is less.Hence we can choose to ignore it in our model and get higher accuracy.

Code For creating model :

```
#Remove free.Sulphur.dioxide
mymodel<-lm(cbind(wine$alcohol,wine$quality)~wine$fixed.acidity+wine$volatile.acidity+wine$citric.acid+wine$residual.sugar+wine$chlorides+wine$total.sulfur.dioxide+wine$density+wine$pH+wine$sulphates,data=wine)
summary(mymodel)
coef(mymodel)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4205 -0.2559 -0.0325  0.2177 16.1715

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)               6.855e+02  5.032e+00  136.221  < 2e-16 ***
wine$fixed.acidity        5.263e-01  9.838e-03   53.493  < 2e-16 ***
wine$volatile.acidity     9.480e-01  6.491e-02   14.606  < 2e-16 ***
wine$citric.acid          3.613e-01  5.656e-02    6.388 1.84e-10 ***
wine$residual.sugar       2.396e-01  2.753e-03   87.013  < 2e-16 ***
wine$chlorides           -2.706e-01  3.242e-01   -0.835    0.404
wine$total.sulfur.dioxide -2.749e-04  1.806e-04   -1.522    0.128
wine$density             -6.930e+02  5.177e+00 -133.869  < 2e-16 ***
wine$pH                   2.463e+00  5.157e-02   47.753  < 2e-16 ***
wine$sulphates            1.028e+00  5.773e-02   17.808  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4459 on 4888 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8687
F-statistic:  3601 on 9 and 4888 DF,  p-value: < 2.2e-16


Response wine$quality :

Call:
lm(formula = `wine$quality` ~ wine$fixed.acidity + wine$volatile.acidity +
    wine$citric.acid + wine$residual.sugar + wine$chlorides +
    wine$total.sulfur.dioxide + wine$density + wine$pH + wine$sulphates,
    data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5303 -0.4933 -0.0501  0.4725  6.0060

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.886e+02  8.545e+00  33.773   <2e-16 ***
wine$fixed.acidity        1.617e-01  1.670e-02   9.679   <2e-16 ***
wine$volatile.acidity    -1.785e+00  1.102e-01 -16.197   <2e-16 ***
wine$citric.acid          1.036e-01  9.604e-02   1.079   0.2806
wine$residual.sugar       1.312e-01  4.675e-03  28.071   <2e-16 ***
wine$chlorides           -2.260e-01  5.504e-01  -0.410   0.6815
wine$total.sulfur.dioxide 6.557e-04  3.066e-04   2.138   0.0325 *
wine$density             -2.901e+02  8.789e+00 -33.006   <2e-16 ***
wine$pH                   1.161e+00  8.757e-02  13.257   <2e-16 ***
wine$sulphates            8.241e-01  9.802e-02   8.407   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7571 on 4888 degrees of freedom
Multiple R-squared:  0.2705,    Adjusted R-squared:  0.2692
F-statistic: 201.4 on 9 and 4888 DF,  p-value: < 2.2e-16
```

- ❖ Conclusion:
  - ❑ Based on the above stats and the stats observed previously we notice that eliminating the variable from our model based on significance of pillai score is not giving us proper results.Hence a multivariate regression model in this case is unable to solve our problem.We must use a different model.