100634271\_1.R

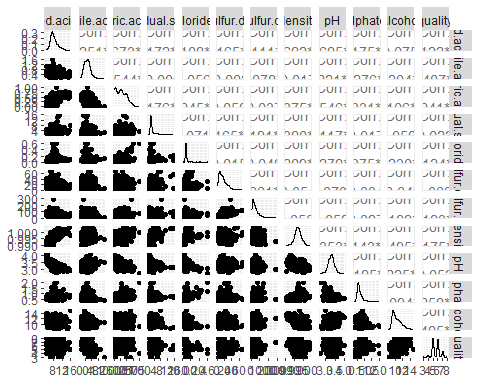
USER

2022-11-27

# ANALYSIS OF WINE QUALITY DATASET  
  
# load the necessary libraries  
library(GGally)  
library(car)  
library(tidyverse)  
  
# Force R not to use scientific notations  
options(scipen = 99)  
  
# Load the wine quality dataset  
# It is the first data statistical analyses will be performed on  
  
wine\_raw <- read.csv(  
 "C:\\Users\\USER\\Desktop\\100634271\\Wine Quality.csv")  
  
str(wine\_raw)

## 'data.frame': 1143 obs. of 13 variables:  
## $ fixed.acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 6.7 ...  
## $ volatile.acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.58 ...  
## $ citric.acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.08 ...  
## $ residual.sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 1.8 ...  
## $ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.097 ...  
## $ free.sulfur.dioxide : num 11 25 15 17 11 13 15 15 9 15 ...  
## $ total.sulfur.dioxide: num 34 67 54 60 34 40 59 21 18 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 3.51 3.3 3.39 3.36 3.28 ...  
## $ sulphates : num 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.54 ...  
## $ alcohol : num 9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 9.2 ...  
## $ quality : int 5 5 5 6 5 5 5 7 7 5 ...  
## $ Id : int 0 1 2 3 4 5 6 7 8 10 ...

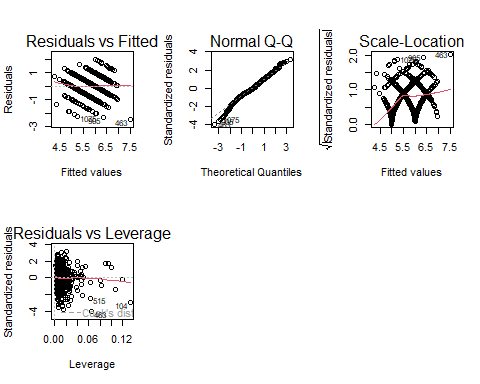
# Dataset Description  
# The wine quality dataset consists of 13 variables which are numeric in nature  
# One of the variables is however an id, which provides the   
# identification of each row in the dataset.  
# Quality is the dependent, while the rest of   
# the variables are independent variables.  
  
# Select all other columns except the Id column  
  
wine <- wine\_raw%>%select(c(fixed.acidity,volatile.acidity,citric.acid,  
 residual.sugar,chlorides,free.sulfur.dioxide,  
 total.sulfur.dioxide,density,  
 pH,sulphates,alcohol,quality))  
  
  
# Pairplots of the variables  
# This will enable us see the distribution of the variables  
# as well as the correlations among them  
ggpairs(wine)



# Build the linear model  
wine\_model<- lm(quality ~ ., data = wine)  
summary(wine\_model)

##   
## Call:  
## lm(formula = quality ~ ., data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.49977 -0.36903 -0.04658 0.43956 2.00117   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.5494703 24.7729578 0.870 0.384551   
## fixed.acidity 0.0229746 0.0302535 0.759 0.447770   
## volatile.acidity -1.1291020 0.1407270 -8.023 0.00000000000000256 \*\*\*  
## citric.acid -0.1318627 0.1730049 -0.762 0.446105   
## residual.sugar 0.0135117 0.0184568 0.732 0.464278   
## chlorides -1.7081606 0.4973883 -3.434 0.000616 \*\*\*  
## free.sulfur.dioxide 0.0023694 0.0025529 0.928 0.353547   
## total.sulfur.dioxide -0.0027849 0.0008386 -3.321 0.000926 \*\*\*  
## density -17.4492955 25.2858434 -0.690 0.490284   
## pH -0.4082028 0.2228714 -1.832 0.067280 .   
## sulphates 0.8752001 0.1335177 6.555 0.00000000008442293 \*\*\*  
## alcohol 0.2801430 0.0312562 8.963 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6405 on 1131 degrees of freedom  
## Multiple R-squared: 0.3742, Adjusted R-squared: 0.3682   
## F-statistic: 61.49 on 11 and 1131 DF, p-value: < 0.00000000000000022

# Examine the diagnostics plots to see that the assumptions are not violated.   
par(mfrow = c(2, 3))  
plot(wine\_model)  
par(mfrow = c(1, 1))



# MULTICOLINEARITY TEST  
# The vif function from the car package is used.  
  
vif(wine\_model)

## fixed.acidity volatile.acidity citric.acid   
## 7.780540 1.778704 3.222840   
## residual.sugar chlorides free.sulfur.dioxide   
## 1.743237 1.538470 1.906045   
## total.sulfur.dioxide density pH   
## 2.103748 6.595115 3.393307   
## sulphates alcohol   
## 1.440741 3.184642

# the vif for each of the independent variable is less than 10.  
# Therefore, there is no serious collinearity among them.