PROJECT REPORT ON ANALYSIS OF QUALITY OF WINE

Descriptive analytics of dataset of Wine

The data set comprises of measures of content in wine and wine quality.

The project is to predict the best quality of wine. In order to increase the consumption of wine and ultimately resulting in the better sales of wine in market. In this dataset wine quality is defined by the value which lies in the range of 4 to 8 and differ by 1 usually.

The given data set is

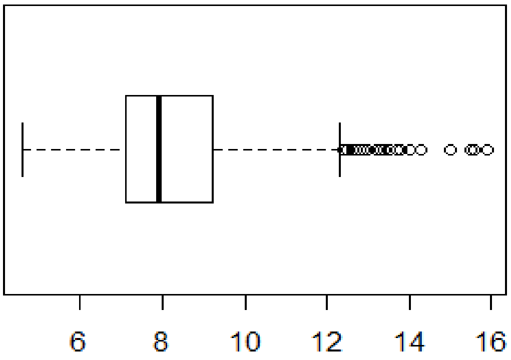
Wine.csv

Variable Identification

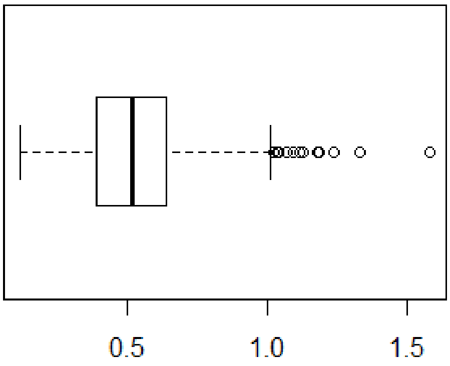
Predictor variables	**Numeric**	**Continuous**
fixed.acidity	fixed.acidity	fixed.acidity
volatile.acidity	volatile.acidity	volatile.acidity
citric.acid	citric.acid	citric.acid
residual.sugar	residual.sugar	residual.sugar
chlorides	chlorides	chlorides
free.sulfur.dioxide	free.sulfur.dioxide	free.sulfur.dioxide
total.sulfur.dioxide	total.sulfur.dioxide	total.sulfur.dioxide
density	density	density
pH	pH	pH
sulphates	sulphates	sulphates
alcohol	alcohol	alcohol
		quality
Response variable	**Integer**	
quality	quality	

#UNIVARIATE ANALYSIS

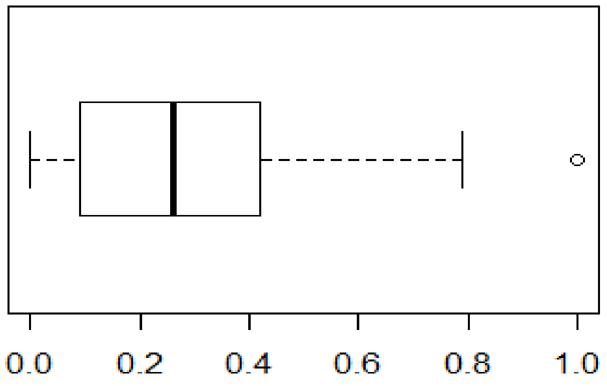
intercept	fixed.acidity
mean	8.3196
median	7.9000
var	3.0314
sd	1.7411
max	15.9000
min	4.6000
range	15.9-4.6
IQR	2.1000
skewness	0.9818
kurtosis	4.1249



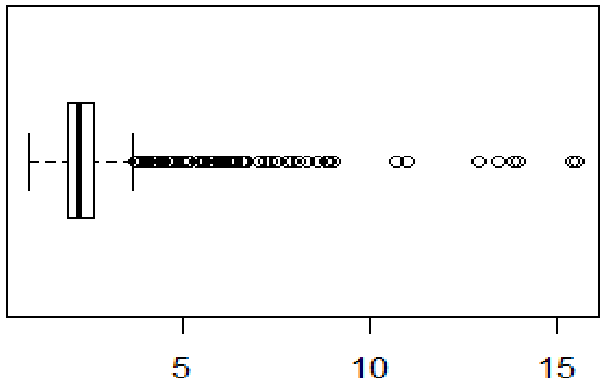
intercept	volatile.acidity
mean	0.5278
median	0.5200
var	0.0321
sd	0.1791
max	1.5800
min	0.1200
range	1.58-0.12
IQR	0.2500
skewness	0.6710
kurtosis	4.2180



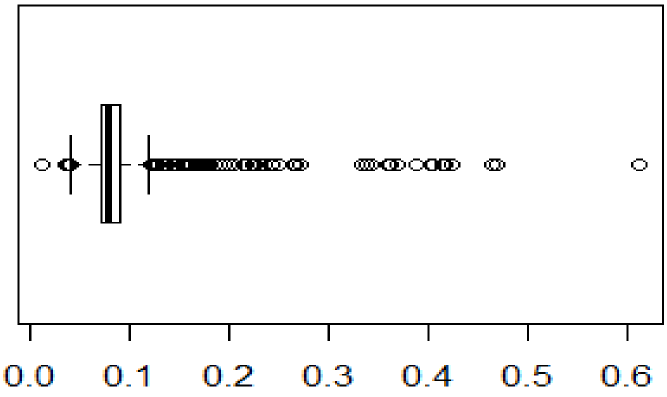
intercept	citric.acid
mean	0.2710
median	0.2600
var	0.0379
sd	0.1948
max	1.0000
min	0.0000
range	1-0
IQR	0.3300
skewness	0.3180
kurtosis	2.2097



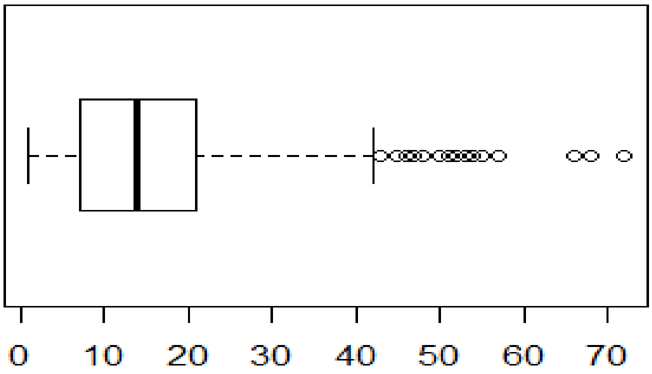
intercept	residual.sugar
mean	2.5388
median	2.2000
var	1.9879
sd	1.4099
max	15.5000
min	0.9000
range	15.5-0.9
IQR	0.7000
skewness	4.5364
kurtosis	31.5244



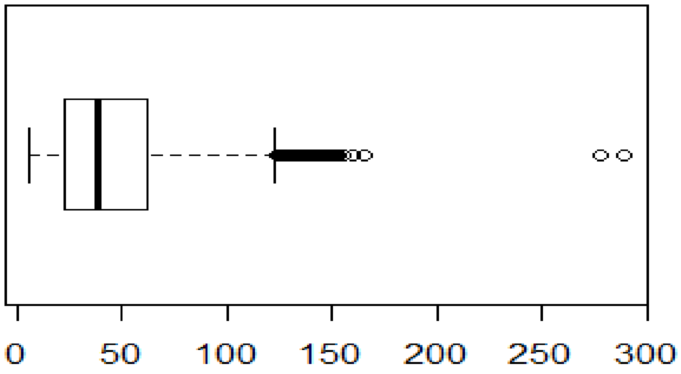
intercept	chlorides
mean	0.0875
median	0.0790
var	0.0022
sd	0.0471
max	0.6110
min	0.0120
range	0.611-0.012
IQR	0.0200
skewness	5.6750
kurtosis	44.5817



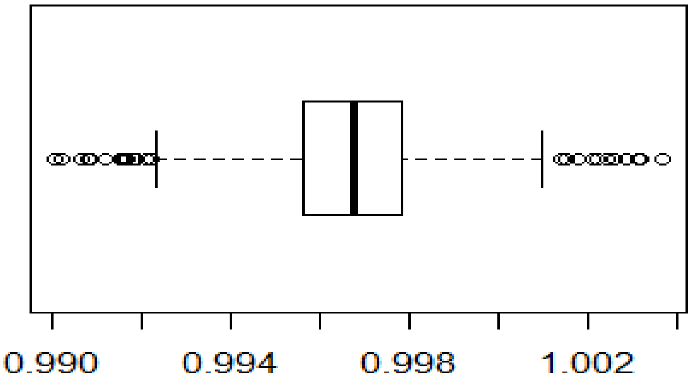
intercept	free.sulfur.dioxide
mean	15.8749
median	14.0000
var	109.4149
sd	10.4602
max	72.0000
min	1.0000
range	72-1
IQR	14.0000
skewness	1.2494
kurtosis	5.0135



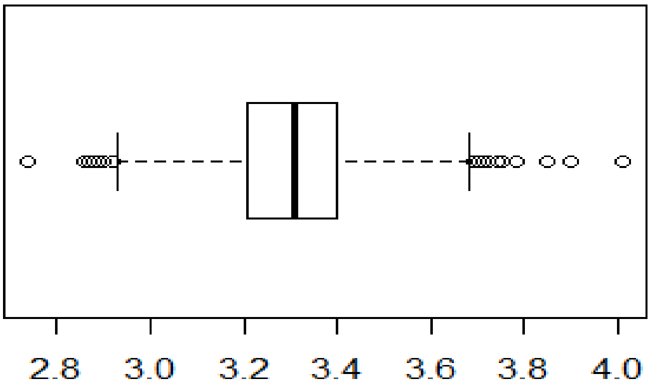
intercept	total.sulfur.dioxide
mean	46.4678
median	38.0000
var	1082.1024
sd	32.8953
max	289.0000
min	6.0000
range	289-6
IQR	40.0000
skewness	1.5141
kurtosis	6.7942



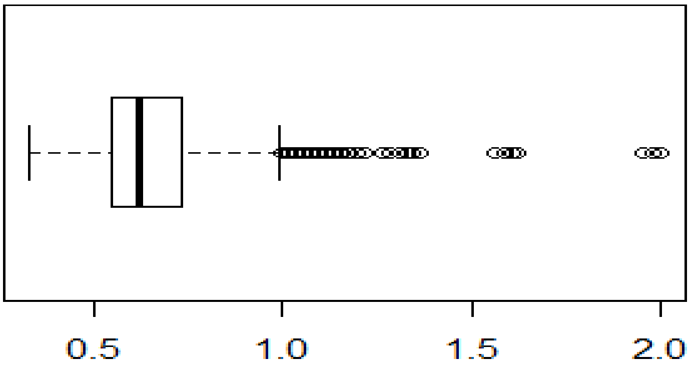
intercept	density
mean	0.9967
median	0.9968
var	0.0000
sd	0.0019
max	1.0037
min	0.9901
range	1.00369-0.99007
IQR	0.0022
skewness	0.0712
kurtosis	3.9274



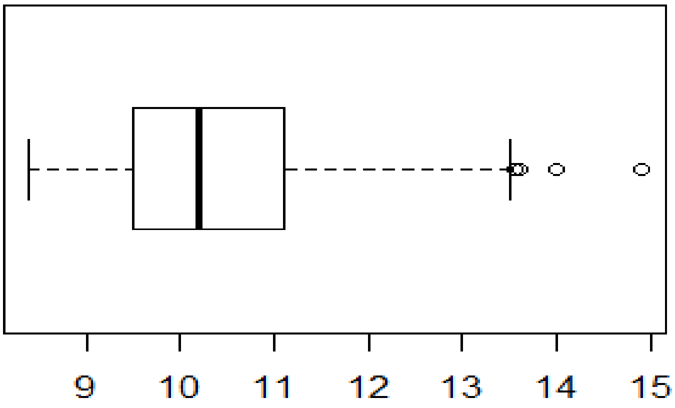
intercept	pH
mean	3.3111
median	3.3100
var	0.0238
sd	0.1544
max	4.0100
min	2.7400
range	4.01-2.74
IQR	0.1900
skewness	0.1935
kurtosis	3.8007



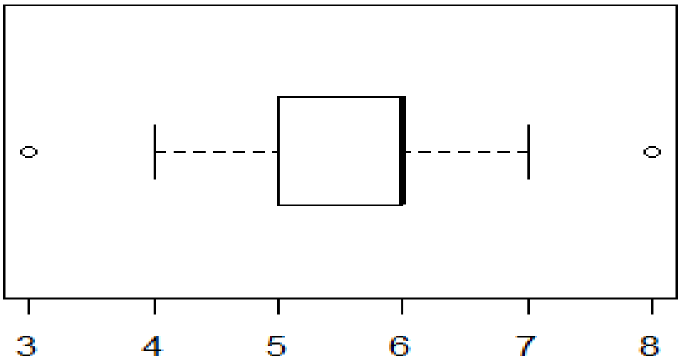
intercept	sulphates
mean	0.6581
median	0.6200
var	0.0287
sd	0.1695
max	2.0000
min	0.3300
range	2-0.33
IQR	0.1800
skewness	2.4264
kurtosis	14.6799



intercept	alcohol
mean	10.4230
median	10.2000
var	1.1356
sd	1.0657
max	14.9000
min	8.4000
range	14.9-8.4
IQR	1.6000
skewness	0.8600
kurtosis	3.1957



intercept	quality
mean	5.6360
median	6.0000
var	0.6522
sd	0.8076
max	8.0000
min	3.0000
range	42437.0000
IQR	1.0000
skewness	0.2176
kurtosis	3.2920



```
#Setting working directory
setwd("C:/Users/User/Documents/Modelling--Wine dataset")

#Reading data from .csv file
mydata = read.csv("winequality-red.csv")

#Checking the head of dataset
head(mydata)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
```

```
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
## 1                  11                   34  0.9978 3.51      0.56
## 2                  25                   67  0.9968 3.20      0.68
## 3                  15                   54  0.9970 3.26      0.65
## 4                  17                   60  0.9980 3.16      0.58
## 5                  11                   34  0.9978 3.51      0.56
## 6                  13                   40  0.9978 3.51      0.56
```

```
##        alcohol       quality
## 1        9.4            5
## 2        9.8            5
## 3        9.8            5
## 4        9.8            6
## 5        9.4            5
## 6        9.4            5
```

```
#Checking the Summary of data to check the disperse in data and also to fi
nd out the NA's value
summary(mydata)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chloridesfree.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
```

```
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
```

```
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

```
#Checking the structure of dataset
str(mydata)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5
...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58
0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069
0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ 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 3.3
6 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0
.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 .
..
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
#Dividing the dataset into train and test data
train_ind = sample(seq_len(nrow(mydata)),round(0.70*nrow(mydata)))


train = mydata[train_ind,]


str(train)

## 'data.frame':    1119 obs. of  12 variables:
##  $ fixed.acidity       : num  12.6 6.7 8.3 6.4 8 10.7 11.5 8.9 11.6 6.8
...
##  $ volatile.acidity    : num  0.31 0.855 0.715 0.57 0.42 0.43 0.3 0.5 0
.32 0.56 ...
##  $ citric.acid         : num  0.72 0.02 0.15 0.14 0.17 0.39 0.6 0.21 0.
55 0.22 ...
##  $ residual.sugar      : num  2.2 1.9 1.8 3.9 2 2.2 2 2.2 2.8 1.8 ...
##  $ chlorides           : num  0.072 0.064 0.089 0.07 0.073 0.106 0.067
0.088 0.081 0.074 ...
##  $ free.sulfur.dioxide : num  6 29 10 27 6 8 12 21 35 15 ...
##  $ total.sulfur.dioxide: num  29 38 52 73 18 32 27 39 67 24 ...
##  $ density             : num  0.999 0.995 0.997 0.997 0.997 ...
##  $ pH                  : num  2.88 3.3 3.23 3.32 3.29 2.89 3.11 3.33 3.
32 3.4 ...
##  $ sulphates           : num  0.82 0.56 0.77 0.48 0.61 0.5 0.97 0.83 0.
92 0.82 ...
##  $ alcohol             : num  9.8 10.8 9.5 9.2 9.2 ...
##  $ quality             : int  8 6 5 5 6 5 6 6 7 6 ...



test = mydata[-train_ind,]


str(test)

## 'data.frame':    480 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 6.7 8.9 8.9 8.1 7.4 8.9 7.6 8.5 .
..
##  $ volatile.acidity    : num  0.7 0.58 0.58 0.62 0.62 0.56 0.59 0.22 0.
39 0.49 ...
##  $ citric.acid         : num  0 0.02 0.08 0.18 0.19 0.28 0.08 0.48 0.31
0.11 ...
##  $ residual.sugar      : num  1.9 2 1.8 3.8 3.9 1.7 4.4 1.8 2.3 2.3 ...
##  $ chlorides           : num  0.076 0.073 0.097 0.176 0.17 0.368 0.086
0.077 0.082 0.084 ...
##  $ free.sulfur.dioxide : num  11 9 15 52 51 16 6 29 23 9 ...
##  $ total.sulfur.dioxide: num  34 18 65 145 148 56 29 60 71 67 ...
##  $ density             : num  0.998 0.997 0.996 0.999 0.999 ...
##  $ pH                  : num  3.51 3.36 3.28 3.16 3.17 3.11 3.38 3.39 3
.52 3.17 ...
##  $ sulphates           : num  0.56 0.57 0.54 0.88 0.93 1.28 0.5 0.53 0.
65 0.53 ...
##  $ alcohol             : num  9.4 9.5 9.2 9.2 9.2 9.3 9 9.4 9.7 9.4 ...
##  $ quality             : int  5 7 5 5 5 5 4 6 5 5 ...
```

```r
#Scaling is done for column no.6 and 7
minmax=function(x){
  newx=(x-min(x))/(max(x)-min(x))
}
train[,6] = minmax(train[,6])
test[,6] = minmax(test[,6])
train[,7] = minmax(train[,7])
test[,7] = minmax(test[,7])


head(train)
```

```
##     fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 645           9.9            0.540        0.45            2.3     0.071
## 1041          7.4            0.965        0.00            2.2     0.088
## 1070          8.0            0.620        0.35            2.8     0.086
## 469          11.4            0.360        0.69            2.1     0.090
## 1599          6.0            0.310        0.47            3.6     0.067
## 368          10.4            0.575        0.61            2.6     0.076


##      free.sulfur.dioxide  total.sulfur.dioxide density   pH   sulphates
## 645           0.21126761           0.12014134  0.99910  3.39      0.62
## 1041          0.21126761           0.09187279  0.99756  3.58      0.67
## 1070          0.38028169           0.16254417  0.99700  3.31      0.62
## 469           0.07042254           0.05300353  1.00000  3.17      0.62
## 1599          0.23943662           0.12720848  0.99549  3.39      0.66
## 368           0.14084507           0.06360424  1.00000  3.16      0.69


##       alcohol   quality
## 645      9.4       5
## 1041    10.2       5
## 1070    10.8       5
## 469      9.2       6
## 1599    11.0       6
## 368      9.0       5
```

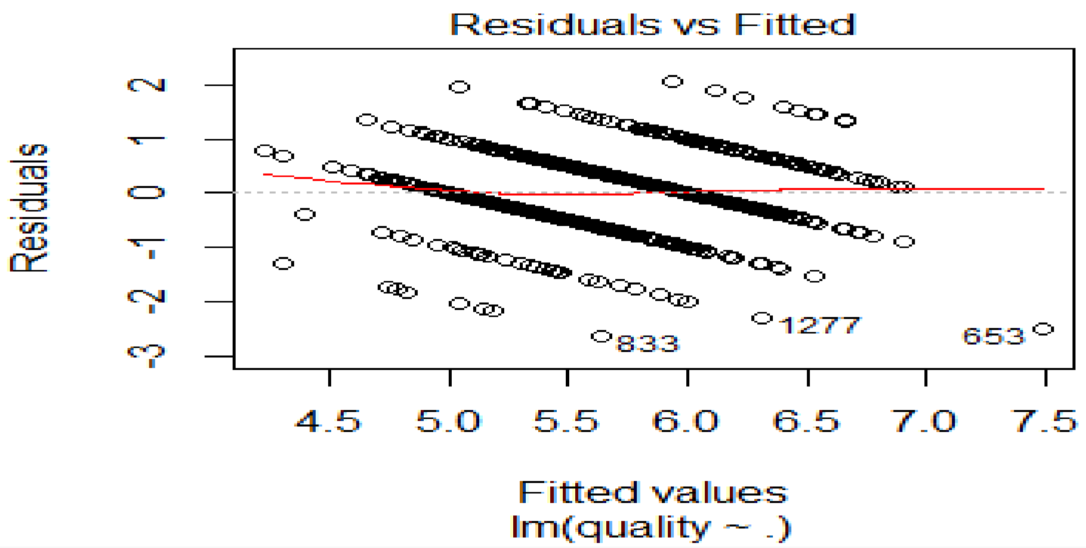
```r
#Developing model considering quality as response variable and rest all as independent variables
fit = lm(quality ~ ., train)
fit
```

```
## 
## Call:
## lm(formula = quality ~ ., data = train)
## Coefficients:
##          (Intercept)         fixed.acidity      volatile.acidity  
##             28.00987               0.03476              -1.10099  
##          citric.acid        residual.sugar             chlorides  
##             -0.07339               0.01665              -1.65735  
##  free.sulfur.dioxide  total.sulfur.dioxide               density  
##              0.18382              -0.73880             -24.47157  
##                   pH             sulphates               alcohol  
##             -0.28215               0.89428               0.27354
```

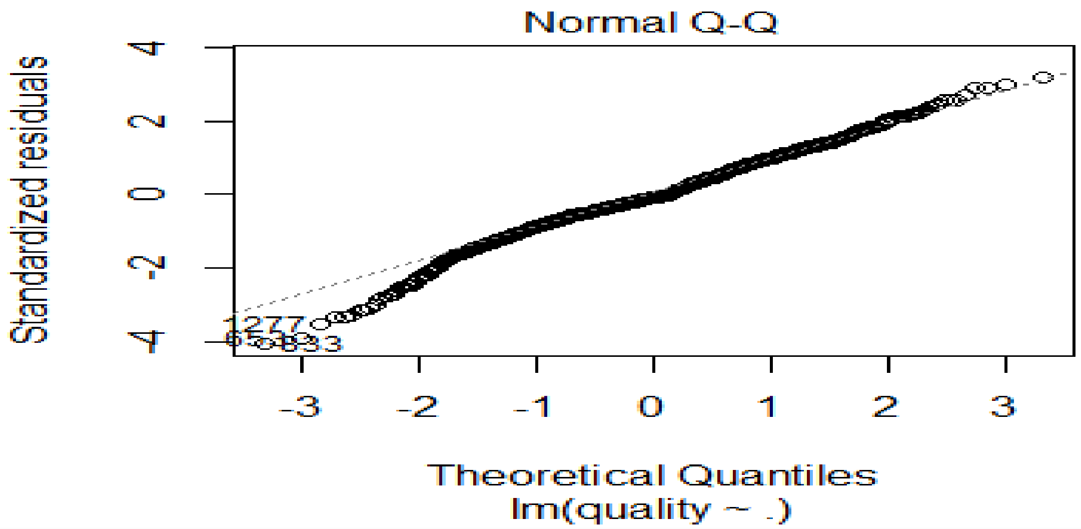
```r
summary(fit)
```

```
## Call:
## lm(formula = quality ~ ., data = train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.70962 -0.37250 -0.05885  0.45868  1.99986 
## 


## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           28.00987   25.85707   1.083  0.27893    
## fixed.acidity          0.03476    0.03142   1.106  0.26883    
## volatile.acidity      -1.10099    0.15290  -7.201 1.11e-12 ***
## citric.acid           -0.07339    0.17439  -0.421  0.67396    
## residual.sugar         0.01665    0.01788   0.931  0.35189    
## chlorides             -1.65735    0.50747  -3.266  0.00112 ** 
## free.sulfur.dioxide    0.18382    0.18858   0.975  0.32989    
## total.sulfur.dioxide  -0.73880    0.25133  -2.940  0.00335 ** 
## density              -24.47157   26.41050  -0.927  0.35434    
## pH                    -0.28215    0.23727  -1.189  0.23464    
## sulphates              0.89428    0.13579   6.586 6.97e-11 ***
## alcohol                0.27354    0.03194   8.563  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.659 on 1107 degrees of freedom
## Multiple R-squared:  0.355,  Adjusted R-squared:  0.3486 
## F-statistic: 55.38 on 11 and 1107 DF,  p-value: < 2.2e-16
```

```
#Plot the model inorder to check all the assumptions of linear regression
is met
plot(fit)

1)Residual vs Fitted

This graph shows that there is a linear relationship between response vari
able and all other dependent variables
```

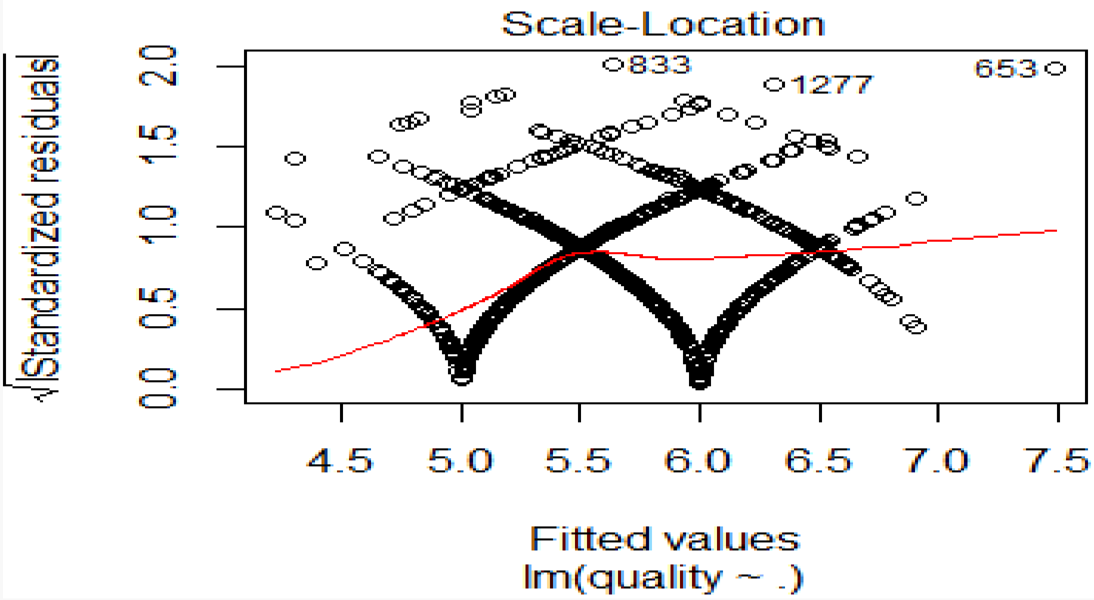


```
2)Normal Q-Q plot

This graph shows that the error are normally distributed near to line
```



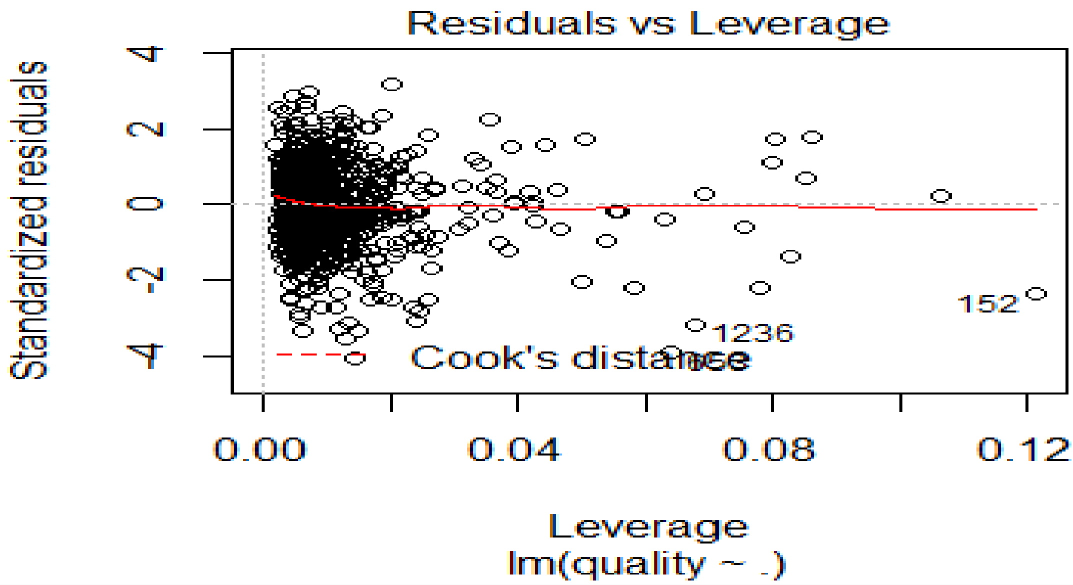
3)Scale-Location Plot

This graph shows that Variance are equally dispersed on the either side of the best fit line



4)Residuals vs Leverage plot

This graph is used to check the outliers and these outliers are actually affecting the predictions and this is done by Cook's Distance. Any value lieing outside the Cook's Distance boundary is considered as outliers.



```
#Attaching library car for using vif function in order to find out the mul
ticollinearity
library(car)
```

```
vif(fit)
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##             7.821364             1.788048             3.039689
##       residual.sugar            chlorides   free.sulfur.dioxide
##             1.756097             1.463939             2.036966
## total.sulfur.dioxide              density                   pH
##             2.202068             6.406207             3.454231
##            sulphates              alcohol
##             1.410045             2.985412
```

```
#Developing model by removing density because it is showing multicollinear
ity with other independent variables
fit1 = lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + resid
ual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + pH +
sulphates + alcohol , train)


fit1
```

```
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dio
xide +
##     pH + sulphates + alcohol, data = train)
##
## Coefficients:
##          (Intercept)         fixed.acidity      volatile.acidity
##             4.061187              0.011739             -1.122825
##          citric.acid        residual.sugar             chlorides
##            -0.076993              0.006732             -1.717835
##  free.sulfur.dioxide  total.sulfur.dioxide                    pH
##             0.199330             -0.757535             -0.409581
##            sulphates               alcohol
##             0.862612              0.295911
```

```
summary(fit1)
```

```
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dio
xide +
##     pH + sulphates + alcohol, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68511 -0.36926 -0.05622  0.46717  2.02133
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           4.061187   0.750101   5.414 7.55e-08 ***
## fixed.acidity         0.011739   0.019232   0.610 0.541728    
## volatile.acidity     -1.122825   0.151066  -7.433 2.12e-13 ***
## citric.acid          -0.076993   0.174336  -0.442 0.658840    
## residual.sugar        0.006732   0.014319   0.470 0.638343    
## chlorides            -1.717835   0.503226  -3.414 0.000664 ***
## free.sulfur.dioxide   0.199330   0.187828   1.061 0.288813    
## total.sulfur.dioxide -0.757535   0.250495  -3.024 0.002551 ** 
## pH                   -0.409581   0.193342  -2.118 0.034361 *  
## sulphates             0.862612   0.131405   6.565 8.00e-11 ***
## alcohol               0.295911   0.020921  14.144  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.659 on 1108 degrees of freedom
## Multiple R-squared:  0.3545, Adjusted R-squared:  0.3486 
## F-statistic: 60.84 on 10 and 1108 DF,  p-value: < 2.2e-16
```

```
vif(fit1)
```

```
##        fixed.acidity     volatile.acidity          citric.acid 
##             2.930595             1.745574             3.038178 
##       residual.sugar            chlorides  free.sulfur.dioxide 
##             1.126572             1.439720             2.020924 
## total.sulfur.dioxide                   pH            sulphates 
##             2.187810             2.293820             1.320691 
##              alcohol 
##             1.280610
```

```
#Thus the answer shows that there is no multicollinearity in all other var
iables because the value is within desired range
```

```
#Checking the step function to remove the irrelevant variables from the mo
del by observing the value of AIC
step(fit1)
```

```
## Start:  AIC=-938.52
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sug
ar + 
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide + 
##     pH + sulphates + alcohol
## 
##                        Df Sum of Sq    RSS     AIC
## - residual.sugar        1     0.001 474.29 -940.52
## - fixed.acidity         1     0.644 474.93 -939.00
## <none>                              474.29 -938.52
## - free.sulfur.dioxide   1     1.528 475.82 -936.92
## - citric.acid           1     1.789 476.08 -936.31
## - pH                    1     2.949 477.24 -933.58
## - chlorides             1     4.647 478.94 -929.61
## - total.sulfur.dioxide  1     4.695 478.98 -929.50
## - sulphates             1    24.348 498.64 -884.50
## - volatile.acidity      1    24.359 498.65 -884.48
## - alcohol               1    72.774 547.06 -780.79
##
## Step:  AIC=-940.52
## quality ~ fixed.acidity + volatile.acidity + citric.acid + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates +
##     alcohol
##
##                        Df Sum of Sq    RSS     AIC
## - fixed.acidity         1     0.643 474.93 -941.00
## <none>                              474.29 -940.52
## - free.sulfur.dioxide   1     1.535 475.82 -938.90
## - citric.acid           1     1.808 476.10 -938.26
## - pH                    1     2.951 477.24 -935.58
## - chlorides             1     4.676 478.97 -931.54
## - total.sulfur.dioxide  1     4.738 479.03 -931.40
## - sulphates             1    24.453 498.74 -886.27
## - volatile.acidity      1    24.493 498.78 -886.17
## - alcohol               1    73.454 547.74 -781.39
##
## Step:  AIC=-941
## quality ~ volatile.acidity + citric.acid + chlorides + free.sulfur.diox
ide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## <none>                              474.93 -941.00
## - citric.acid           1     1.169 476.10 -940.25
## - free.sulfur.dioxide   1     1.713 476.65 -938.97
## - chlorides             1     5.588 480.52 -929.91
## - total.sulfur.dioxide  1     6.127 481.06 -928.66
## - pH                    1     6.480 481.41 -927.84
## - volatile.acidity      1    24.108 499.04 -887.59
## - sulphates             1    25.254 500.19 -885.03
## - alcohol               1    72.816 547.75 -783.38

##
## Call:
## lm(formula = quality ~ volatile.acidity + citric.acid + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates +
##     alcohol, data = train)
##
## Coefficients:
##          (Intercept)      volatile.acidity           citric.acid
##               5.0488               -1.0546               -0.2411
##            chlorides   free.sulfur.dioxide  total.sulfur.dioxide
##              -1.8382                0.3654               -0.9365
##                   pH             sulphates               alcohol
##              -0.6269                1.0224                0.2703
```

```
#Predicting the output of model using the test data
pr = predict(fit1,test)

#Finding the error by taking the difference between actual and predicted
variable
error = predict(fit1,test) - test["quality"]
```

```
#Calculating the Root Mean Square Error value
RMSE = sqrt(mean(error^2))
RMSE

## [1] 0.6388015
```

```
#Removing fixed.acidity and considering density in new model
fit2 = lm(quality ~ volatile.acidity + citric.acid + residual.sugar + free
.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates + alcohol + densit
y + chlorides,train)


fit2
## Call:
## lm(formula = quality ~ volatile.acidity + citric.acid + residual.sugar
+ free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates + alcohol +
density + chlorides, data = train)
##
## Coefficients:
##          (Intercept)      volatile.acidity           citric.acid
##             4.699804             -1.273081             -0.227617
##       residual.sugar   free.sulfur.dioxide  total.sulfur.dioxide
##            -0.006805              0.437494             -1.005749
##                   pH             sulphates               alcohol
##            -0.600516              0.779480              0.297067
##              density             chlorides
##             0.283447             -1.738711
```

```
summary(fit2)

##
## Call:
## lm(formula = quality ~ volatile.acidity + citric.acid + residual.sugar
+
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates +
##     alcohol + density + chlorides, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.69431 -0.37115 -0.03677  0.43496  1.97444
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           4.699804  15.316803   0.307 0.759023    
## volatile.acidity     -1.273081   0.143839  -8.851  < 2e-16 ***
## citric.acid          -0.227617   0.162801  -1.398 0.162353    
## residual.sugar       -0.006805   0.014745  -0.461 0.644538    
## free.sulfur.dioxide   0.437494   0.177864   2.460 0.014057 *  
## total.sulfur.dioxide -1.005749   0.230286  -4.367 1.38e-05 ***
## pH                   -0.600516   0.164023  -3.661 0.000263 ***
## sulphates             0.779480   0.131446   5.930 4.04e-09 ***
## alcohol               0.297067   0.025635  11.589  < 2e-16 ***
## density               0.283447  15.268232   0.019 0.985192    
## chlorides            -1.738711   0.491459  -3.538 0.000420 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6452 on 1108 degrees of freedom
## Multiple R-squared:  0.3785, Adjusted R-squared:  0.3729 
## F-statistic: 67.49 on 10 and 1108 DF,  p-value: < 2.2e-16
```

```
vif(fit2)
```

```
##     volatile.acidity          citric.acid       residual.sugar 
##             1.784047             2.750908             1.412756 
##  free.sulfur.dioxide total.sulfur.dioxide                   pH 
##             1.930184             2.049505             1.601196 
##            sulphates              alcohol              density 
##             1.405257             2.027204             2.274625 
##            chlorides 
##             1.404293
```

```
# In the above model (fit2) the more accurate collinearity value is achieved also the value of R-square is increased and Residual standard error is decreased. So we will consider this model as the best fit model for our dataset



#Predicting the output of model using the test data
pr2 = predict(fit2,test)

#Finding the error
error2 = predict(fit2,test) - test["quality"]


#Calculating the Root Mean Square Error value
RMSE2 = sqrt(mean(error2^2))
RMSE2
```

```
## [1] 0.6641864
```