Data 621 Assignment 5

Md Jalal Uddin

December 09, 2018

Overview: In this homework assignment, I will explore, analyze and model a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

My goal is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine.

To begin I will install some require packages. I will do the following data analysis procedure for my assignment here. 1. Data Exploration. 2. Data Preparation. 3. Build model. 4.Select model.

At first, I am installing require packages below:

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v ggplot2 3.1.0     v forcats 0.3.0

## -- Conflicts ------------------------------------------------------------------------------------ ti
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

## Loading required package: sp

## Loading required package: raster
```

```
##
## Attaching package: 'raster'

## The following object is masked from 'package:dplyr':
##
##     select

## The following object is masked from 'package:tidyr':
##
##     extract

## Loading required package: lattice

##
## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':
##
##     complete

## The following objects are masked from 'package:base':
##
##     cbind, rbind

##
## Attaching package: 'cowplot'

## The following object is masked from 'package:ggplot2':
##
##     ggsave

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths

##
## Attaching package: 'corrgram'

## The following object is masked from 'package:lattice':
##
##     panel.fill

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

## Loading required package: Rcpp

## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.5, built: 2018-05-07)
## ## Copyright (C) 2005-2018 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##

##
## Attaching package: 'qqplotr'

## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:usdm':
##
##     vif

## The following object is masked from 'package:psych':
##
##     logit

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some

##
## Attaching package: 'MASS'

## The following objects are masked from 'package:raster':
##
##     area, select

## The following object is masked from 'package:dplyr':
##
##     select

## ------------------------------------------------------------
##  Analysis of Geostatistical Data
##  For an Introduction to geoR go to http://www.leg.ufpr.br/geoR
##  geoR version 1.7-5.2.1 (built on 2016-05-02) is now loaded
## ------------------------------------------------------------
```

(1) DATA EXPLORATION:

In data exploration process, I have described the size and the variables of wine training data. To do that I used summary function of R which shows the mean, median,1st quartile, 3rd quartile and also missing values.

```
##     ï..INDEX         TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##
##    CitricAcid      ResidualSugar        Chlorides       FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00
##  Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00
##                    NA's   :616        NA's   :638       NA's   :647
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
##  Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##  NA's   :682                         NA's   :395     NA's   :1210
##     Alcohol       LabelAppeal          AcidIndex          STARS
```

```
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##  NA's   :653                                          NA's   :3359
```

Removing the Index column: Here, I have removed the index column as follows:

It shows that there are 12795 observations and 16 variables. Each wine has 14 potential predictor variables, and 1 response variable. The response variable is "TARGET", which is the number of cases purchased.

Visual Exploration:

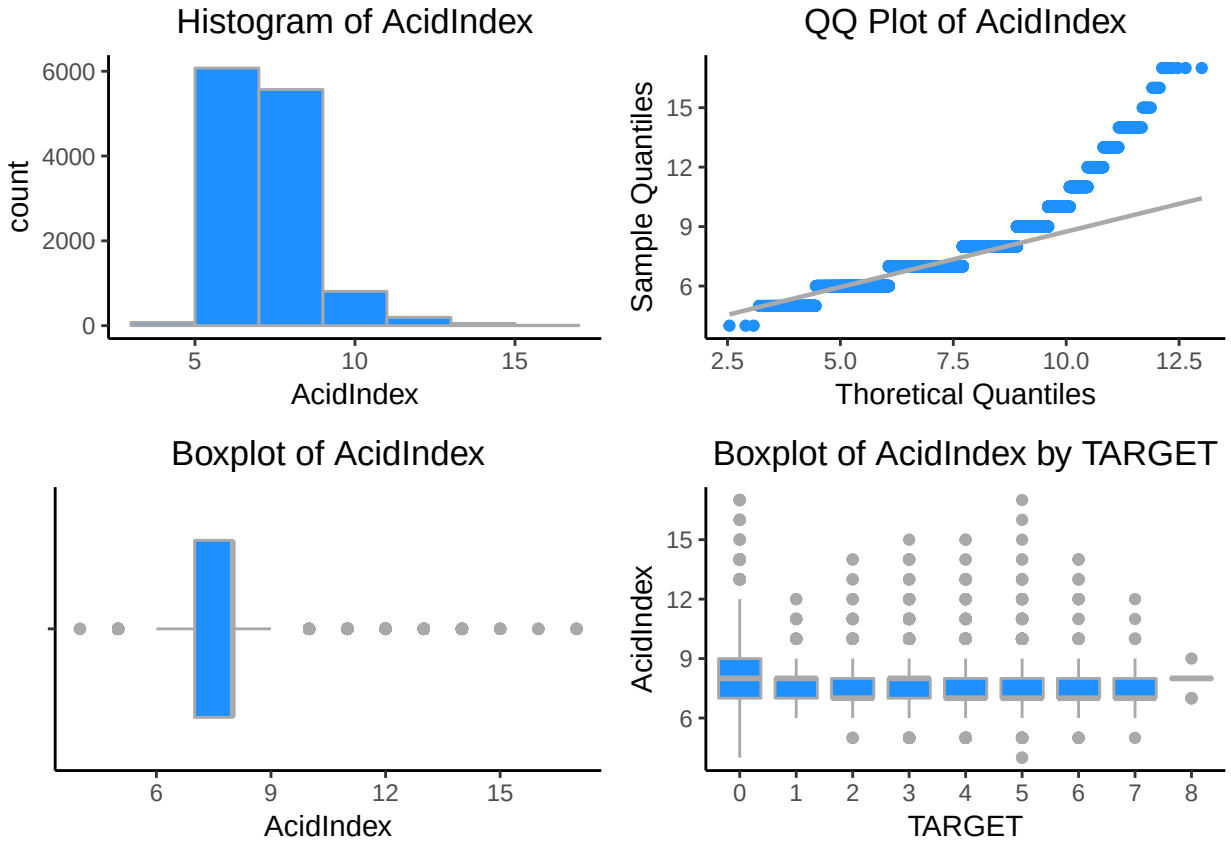
Now I will describe more specifically of the each individual available variables.

Testing Acidity of wine from AcidIndex variables: Proprietary method of testing total acidity of wine by using a weighted average with AcidIndex variable.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.        SD
##  4.000000  7.000000  8.000000  7.772724  8.000000 17.000000  1.323926
##      Skew      Kurt
##  1.648689  8.191373
```

The above result shows that the mean and median of acid index almost same with approximately 1.65 skewness.

Now, I will draw the hitogram, QQ plot, Boxplot of AcidIndex variable. I will also draw the boxplot of the variable by TARGET variable to see how much acidIndex are available in each number of cases.



I will find the information about Alcohol variable. This variable tells us about the Alcohol content.

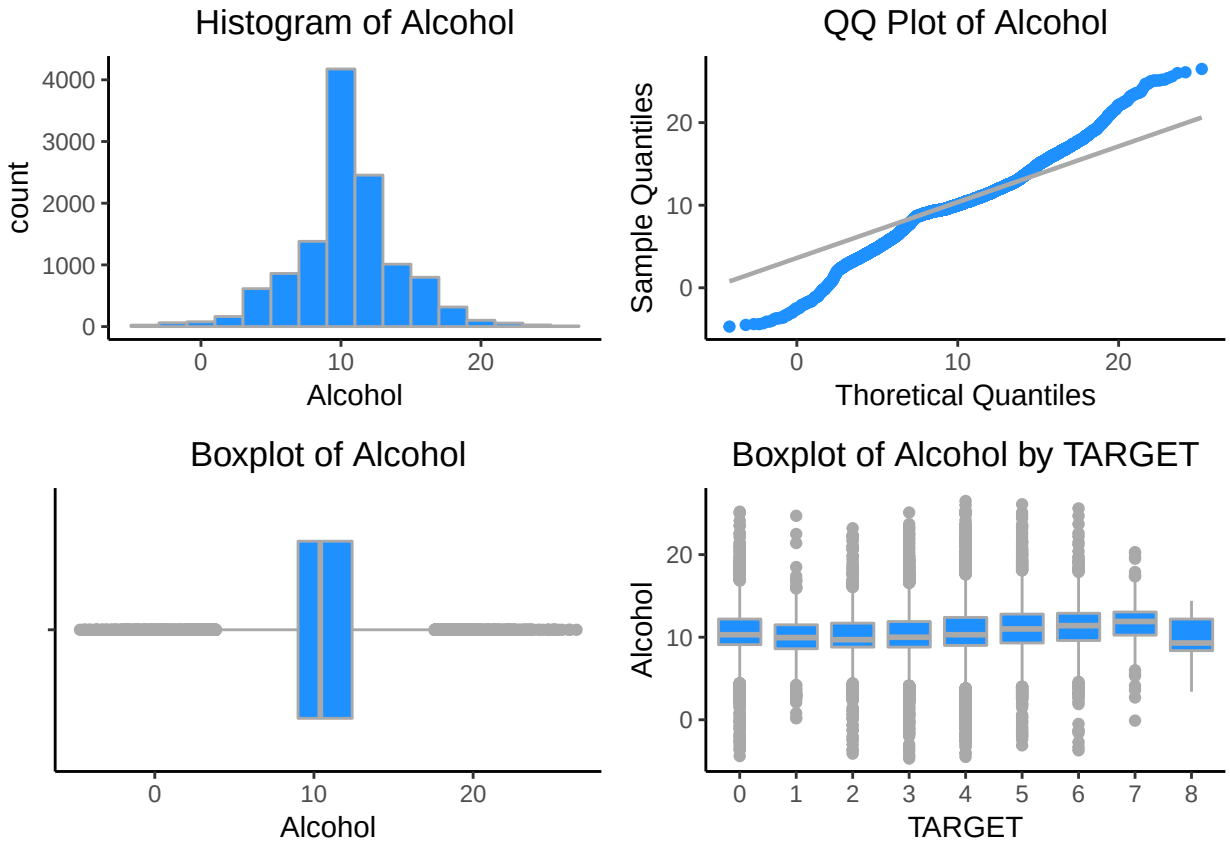
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##  -4.70000   9.00000  10.40000  10.48924  12.40000  26.50000 653.00000
##        SD      Skew      Kurt
##        NA        NA        NA
```

Now, I will draw the hitogram, QQ plot, Boxplot of Alcohol variable. I will also draw the boxplot of the variable by TARGET variable to see how much alcohol are available in each number of cases.

```
## Warning: Removed 653 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 653 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 653 rows containing non-finite values (stat_boxplot).
```



I will describe about Chlorides variable. This variable tells us about the Chloride content of wine.

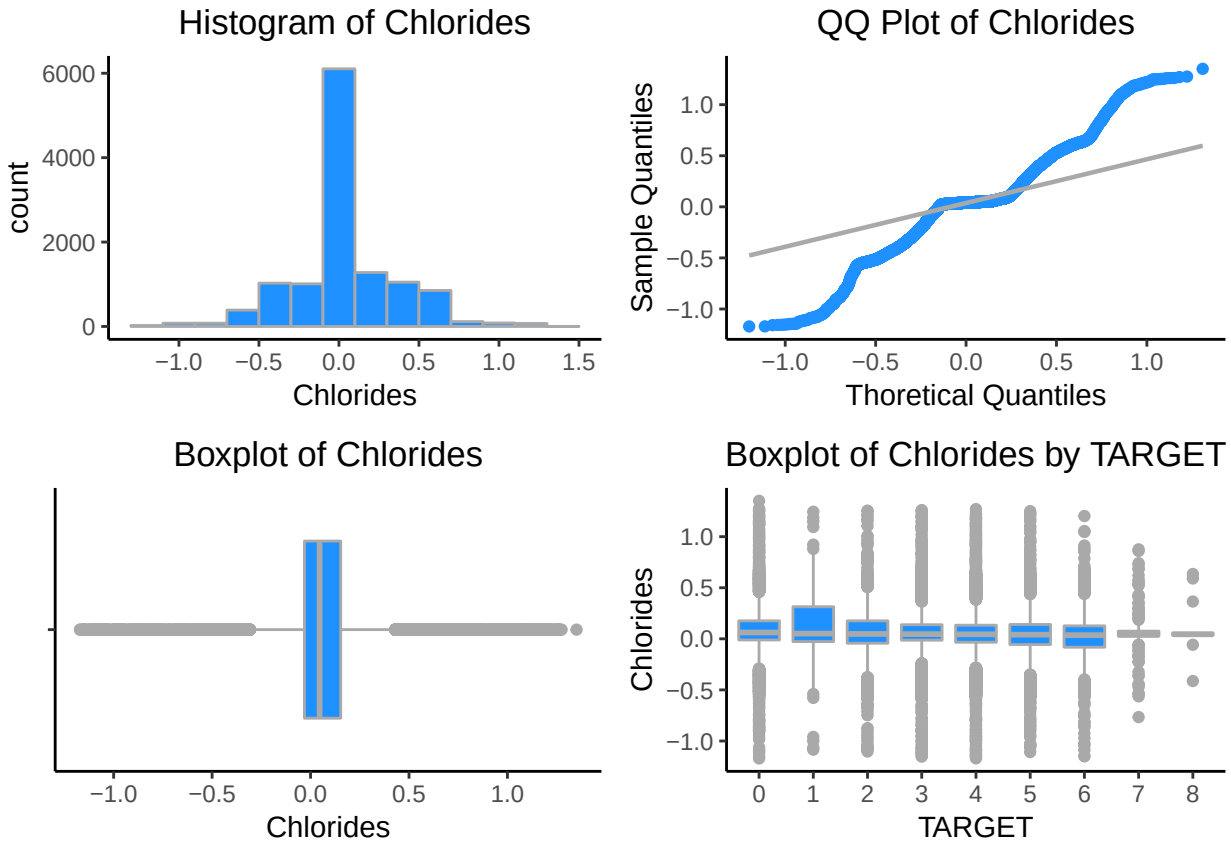
```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
##  -1.17100000  -0.03100000   0.04600000   0.05482249   0.15300000
##         Max.         NA's           SD         Skew         Kurt
##   1.35100000 638.00000000           NA           NA           NA
```

Now, I will draw the hitogram, QQ plot, Boxplot of Chloride variable. I will also draw the boxplot of the variable by TARGET variable to see how muchChloride are available in each number of cases.

```
## Warning: Removed 638 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 638 rows containing non-finite values (stat_boxplot).
```

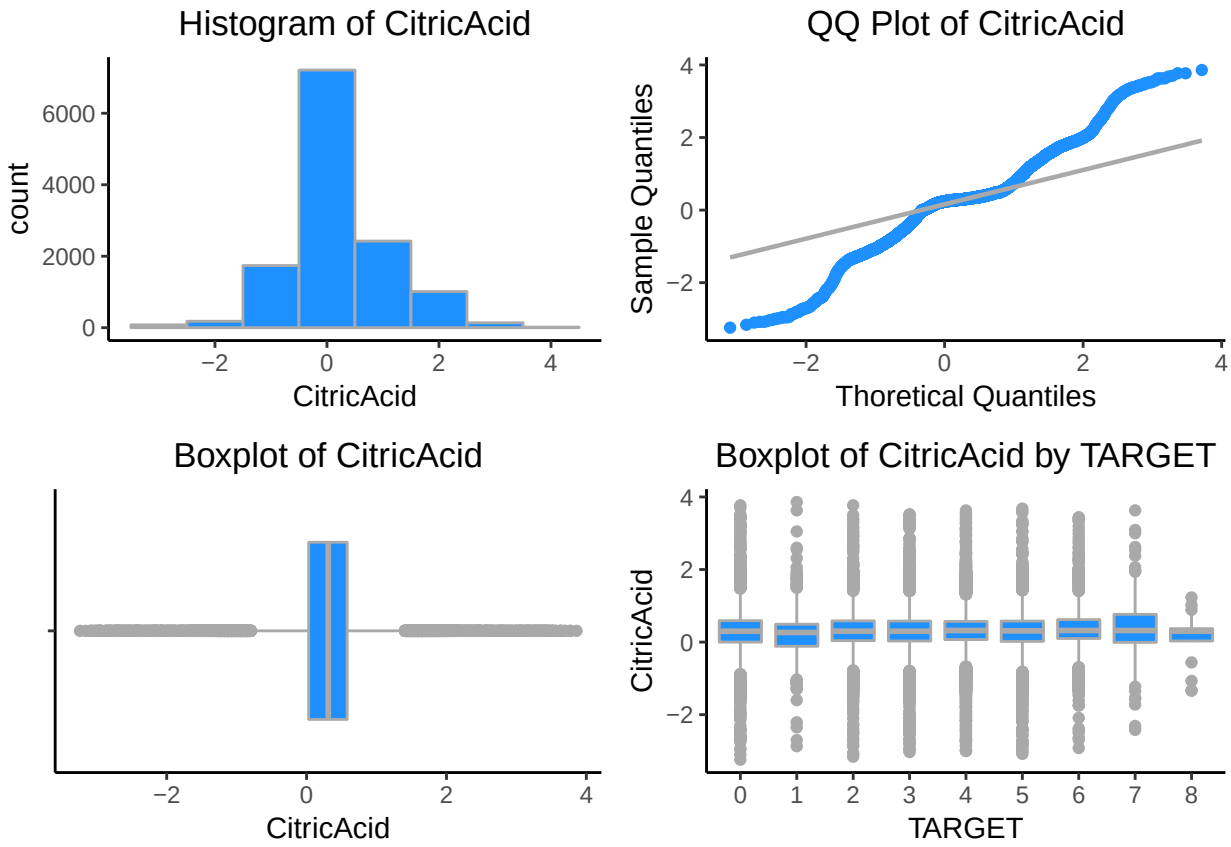
```
## Warning: Removed 638 rows containing non-finite values (stat_boxplot).
```



I will see the information about CitricAcid variable. This variable tells us about the Citric Acid Content of wine.

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## -3.24000000  0.03000000  0.31000000  0.30841266  0.58000000  3.86000000
##           SD         Skew         Kurt
##  0.86207979 -0.05031294  4.83869638
```

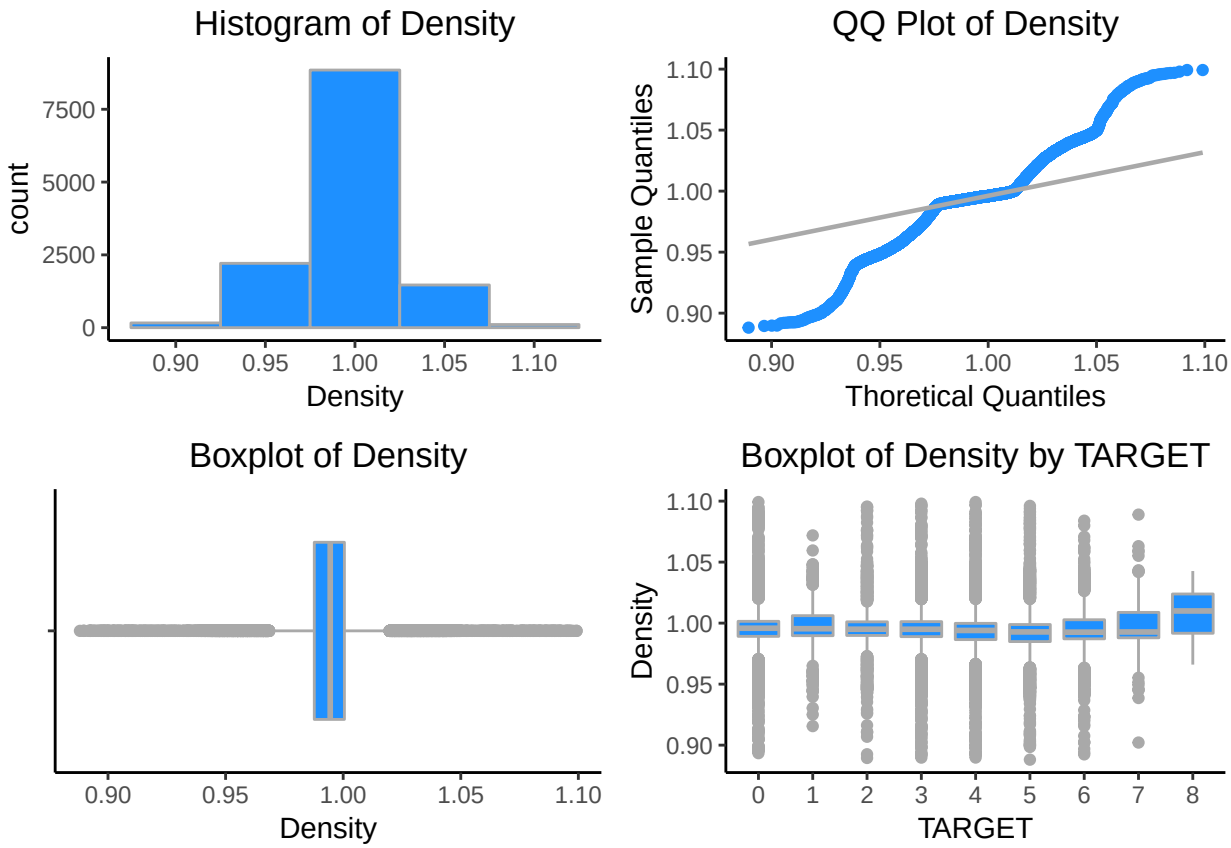
Now, I will draw the hitogram, QQ plot, Boxplot of CitricAcid variable. I will also draw the boxplot of the variable by TARGET variable to see how much CitricAcid are available in each number of cases.



I will find the informaion about Density variable. This variable tells us about the Density of wine.

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##   0.88809000   0.98772000   0.99449000   0.99420272   1.00051500   1.09924000
##           SD         Skew         Kurt
##   0.02653765 -0.01869596  4.90072521
```

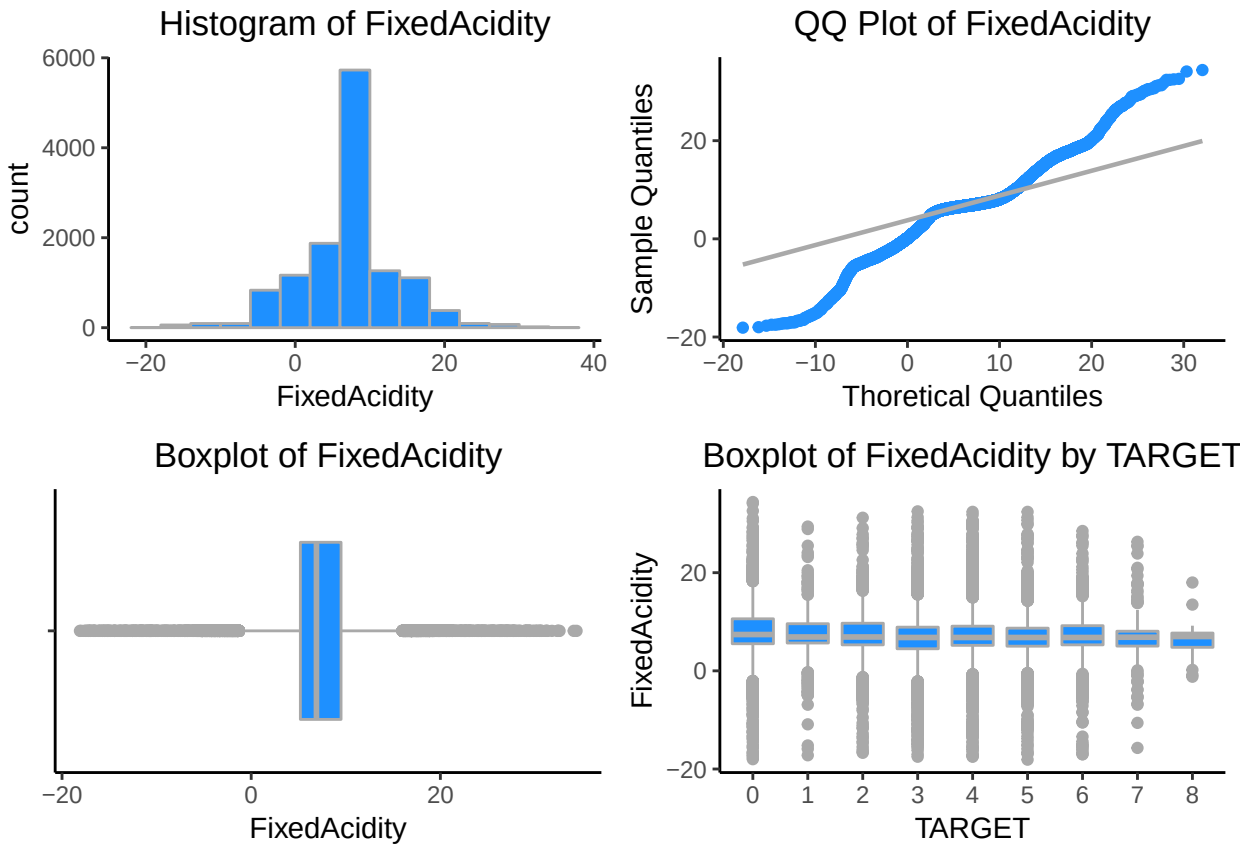
Now, I will draw the hitogram, QQ plot, Boxplot of Density variable. I will also draw the boxplot of the variable by TARGET variable to see how Density are available in each number of cases.



I will find the information about FixedAcidity. This variable tells us about the FixedAcidity of wine.

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## -18.10000000   5.20000000   6.90000000   7.07571708   9.50000000
##         Max.           SD         Skew         Kurt
##  34.40000000   6.31764346  -0.02258861   4.67572951
```

Now, I will draw the hitogram, QQ plot, Boxplot of FixedAcidity variable. I will also draw the boxplot of the variable by TARGET variable to see how much FixedAcidity are available in each number of cases.



I will find the information here abut FreeSulfurDioxide. This variable tells us about the Sulfur Dioxide content of wine.

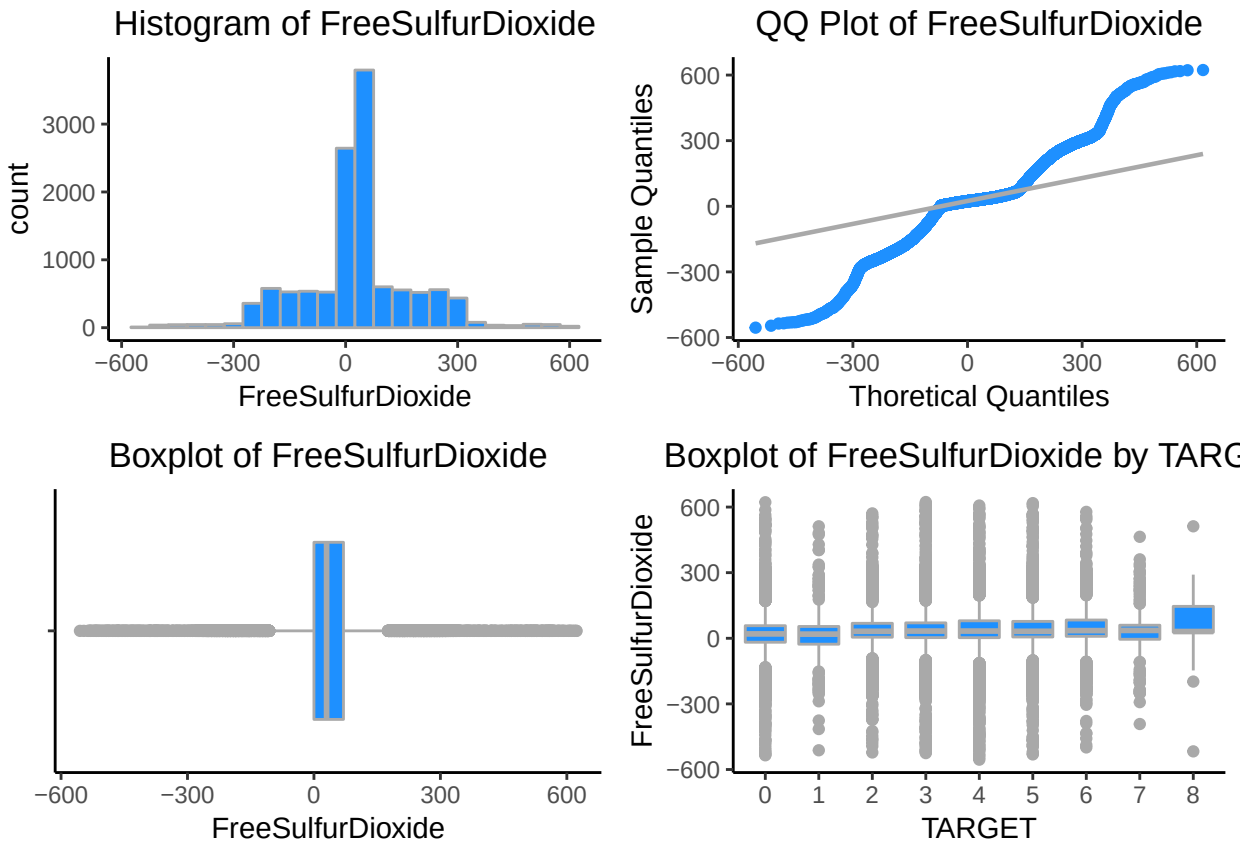
```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -555.00000    0.00000   30.00000   30.84557   70.00000  623.00000
##       NA's         SD       Skew       Kurt
##  647.00000         NA         NA         NA
```

Now, I will draw the hitogram, QQ plot, Boxplot of FreeSulfurDioxide variable. I will also draw the boxplot of the variable by TARGET variable to see how much FreeSulfurDioxide are available in each number of cases.

```
## Warning: Removed 647 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 647 rows containing non-finite values (stat_boxplot).
```

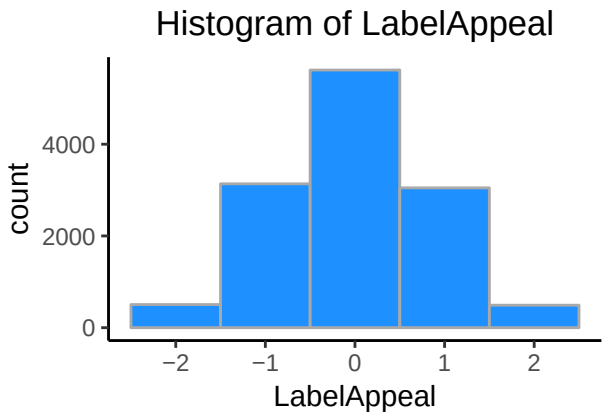
```
## Warning: Removed 647 rows containing non-finite values (stat_boxplot).
```

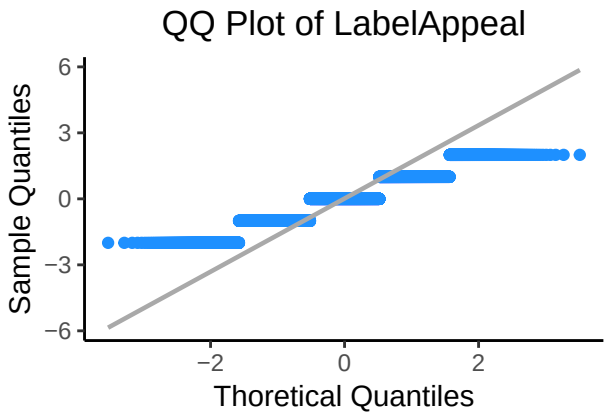


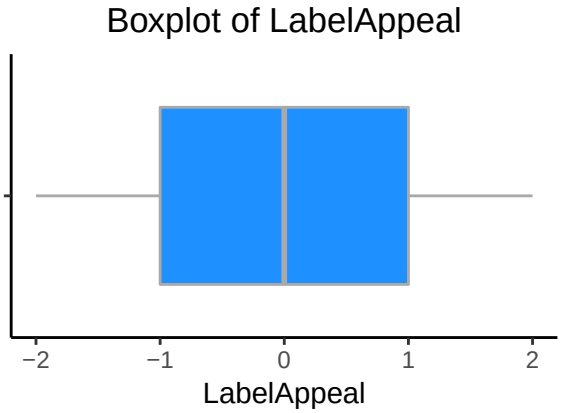
I will find the information about LabelAppeal.This variable shows Marketing Score which indicate the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customers don't like the design. Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales.

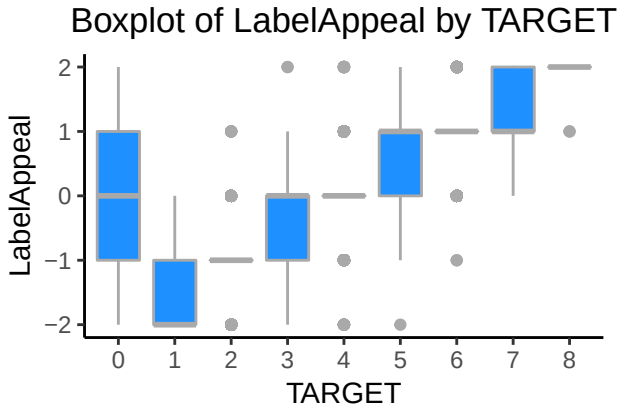
```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## -2.000000000 -1.000000000  0.000000000 -0.009066041  1.000000000
##         Max.           SD         Skew         Kurt
##  2.000000000  0.891089247  0.008430445  2.738136433
```

Now, I will draw the hitogram, QQ plot, Boxplot of LabelAppeal variable. I will also draw the boxplot of the variable by TARGET variable to see how much LabelAppeal were done.









Here, I will find the information about ResidualSugar. This variable tells us about the ResidualSugar of wine.

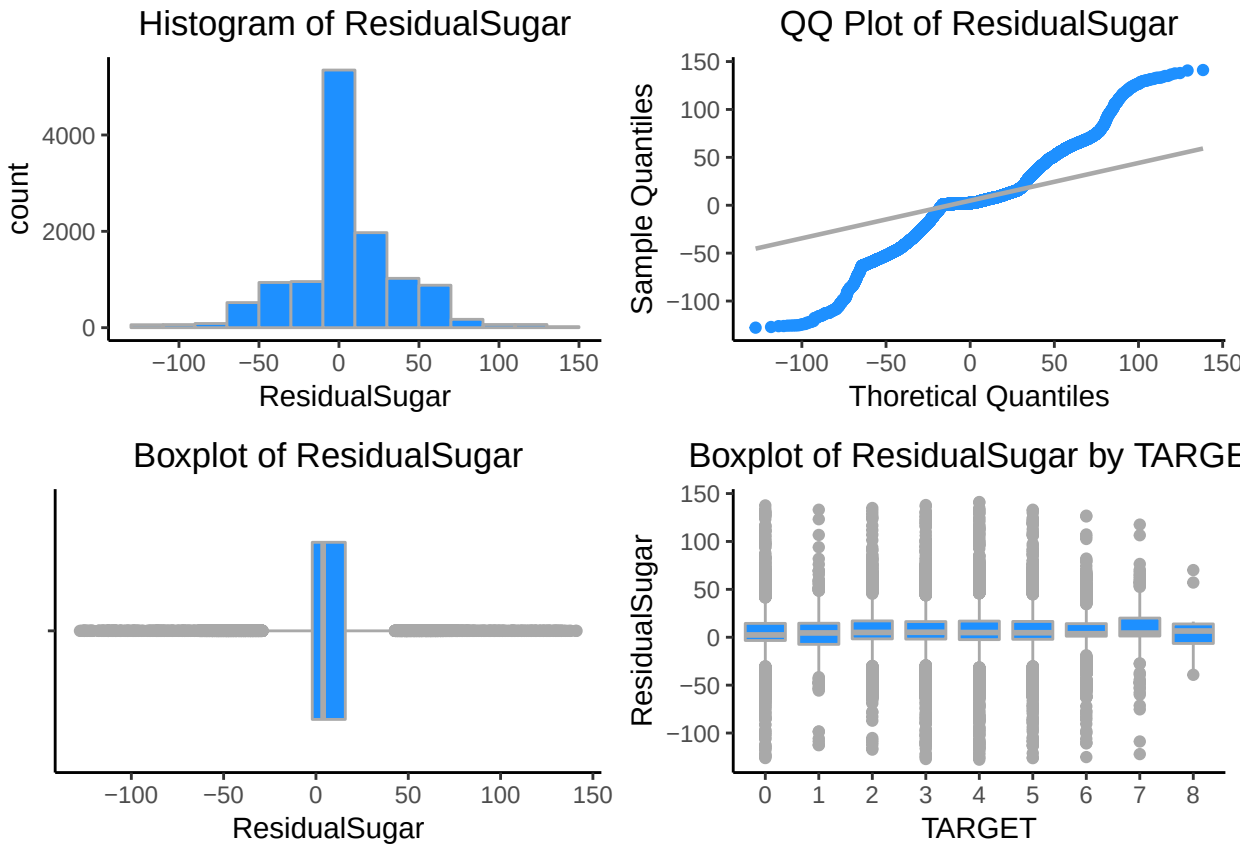
```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## -127.800000   -2.000000    3.900000    5.418733   15.900000  141.150000
##         NA's           SD         Skew         Kurt
##  616.000000           NA           NA           NA
```

Now, I will draw the hitogram, QQ plot, Boxplot of ResidualSugar variable. I will also draw the boxplot of the variable by TARGET variable to see how much ResidualSugar are available in each number of cases.

```
## Warning: Removed 616 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 616 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 616 rows containing non-finite values (stat_boxplot).
```



Here, I will find the information about STARS. This variable tells us about Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor. A high number of stars suggests high sales.

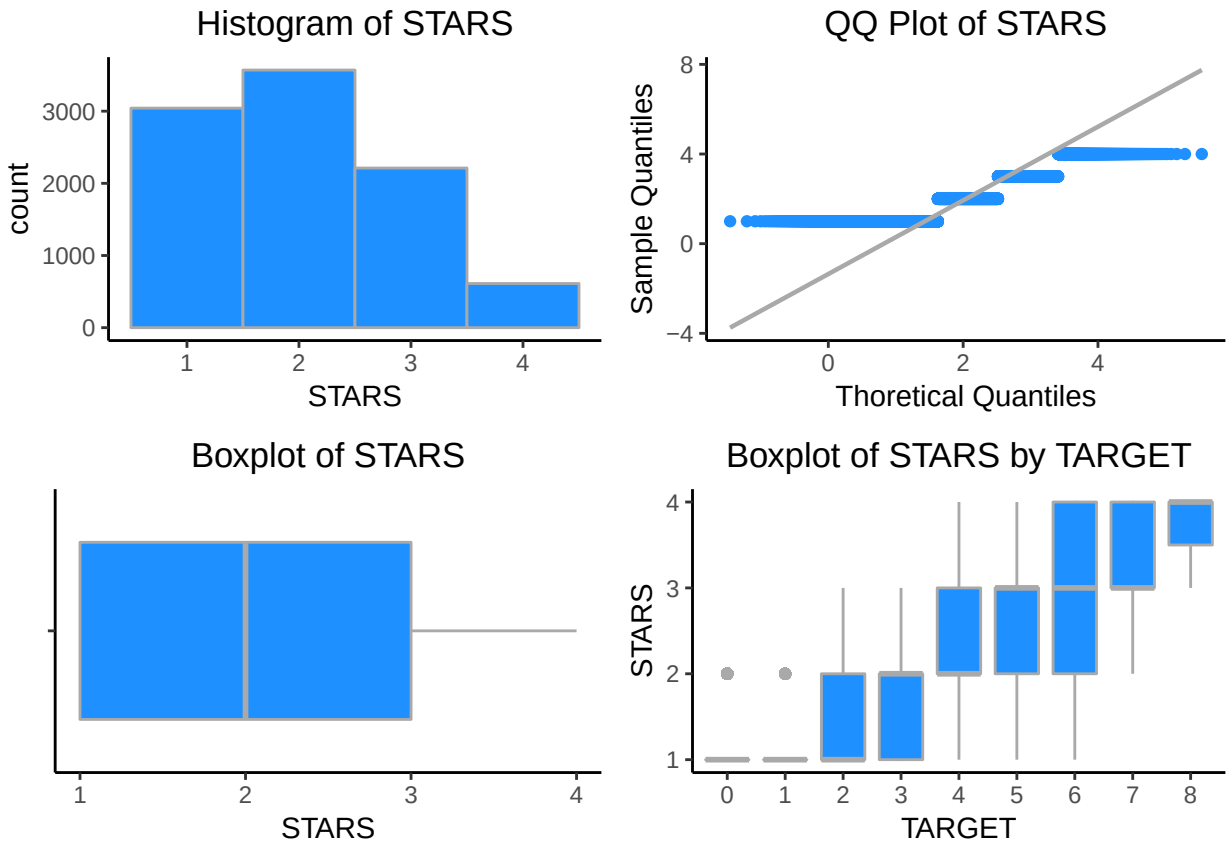
```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##     1.000000     1.000000     2.000000     2.041755     3.000000     4.000000
##         NA's           SD         Skew         Kurt
## 3359.000000           NA           NA           NA
```

Now, I will draw the hitogram, QQ plot, Boxplot of STARS variable. I will also draw the boxplot of the variable by TARGET variable to see how much STARS werer provided.

```
## Warning: Removed 3359 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 3359 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 3359 rows containing non-finite values (stat_boxplot).
```



Here, I will find the information about Sulphates. This variable tells us about the Sulphates content of wine.

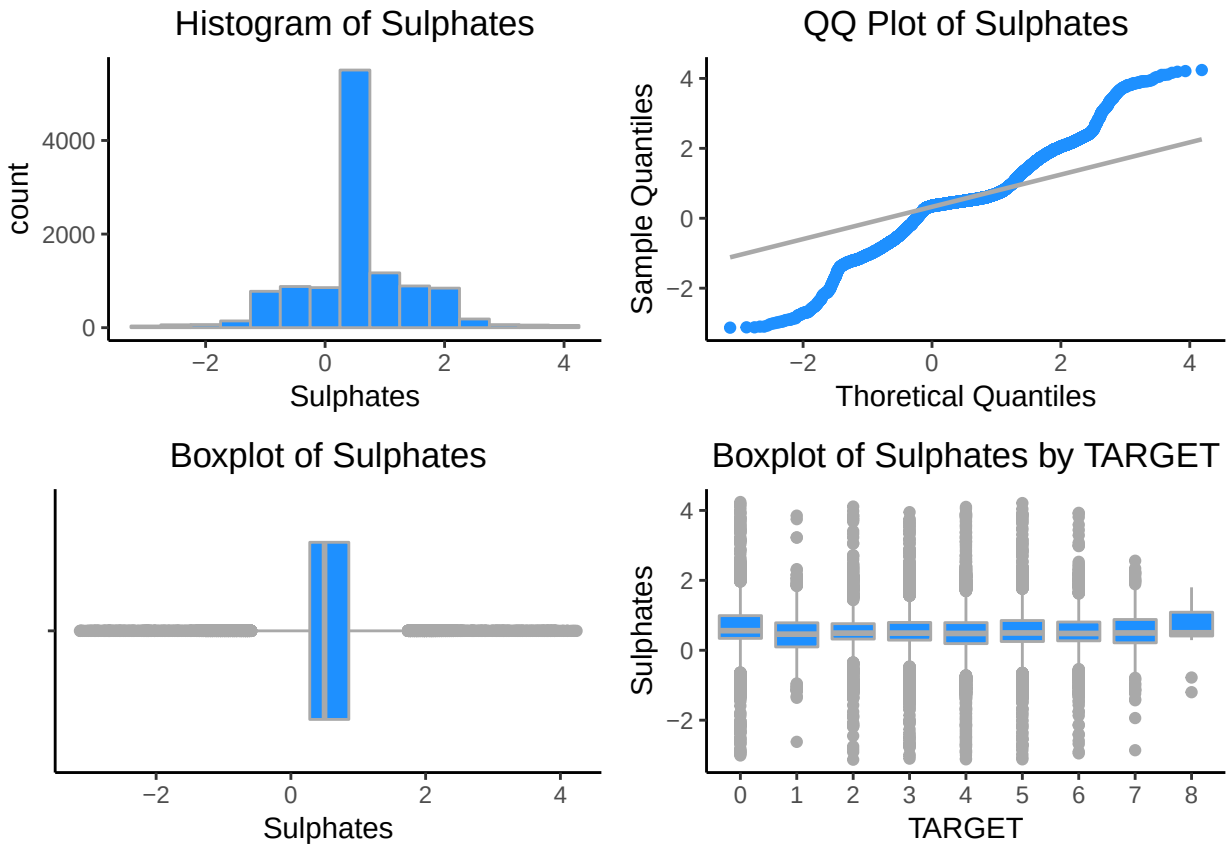
```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
##   -3.1300000    0.2800000    0.5000000    0.5271118    0.8600000
##         Max.         NA's           SD         Skew         Kurt
##    4.2400000 1210.0000000           NA           NA           NA
```

Now, I will draw the hitogram, QQ plot, Boxplot of Sulphates variable. I will also draw the boxplot of the variable by TARGET variable to see how much Sulphates are available in each number of cases.

```
## Warning: Removed 1210 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1210 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 1210 rows containing non-finite values (stat_boxplot).
```



Here, I will find the information about TotalSulfurDioxide. This variable tells us about the Total Sulfur Dioxide of Wine.

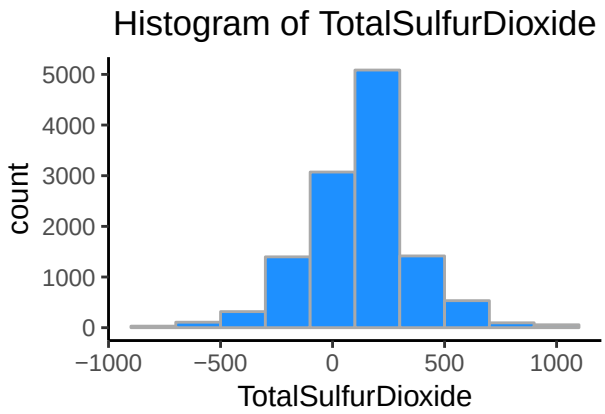
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## -823.0000   27.0000  123.0000  120.7142  208.0000 1057.0000  682.0000
##        SD      Skew      Kurt
##        NA        NA        NA
```

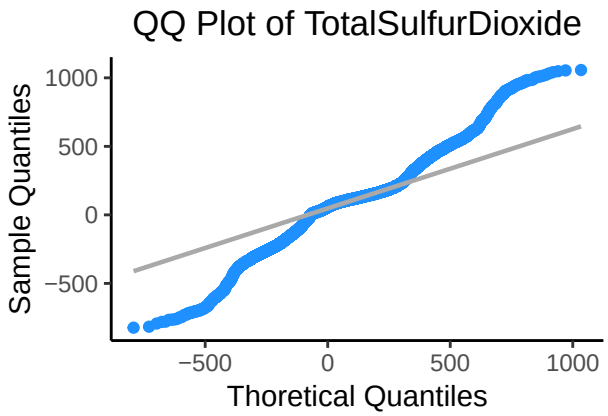
Now, I will draw the hitogram, QQ plot, Boxplot of TotalSulfurDioxide variable. I will also draw the boxplot of the variable by TARGET variable to see how much RTotalSulfurDioxide are available in each number of cases.

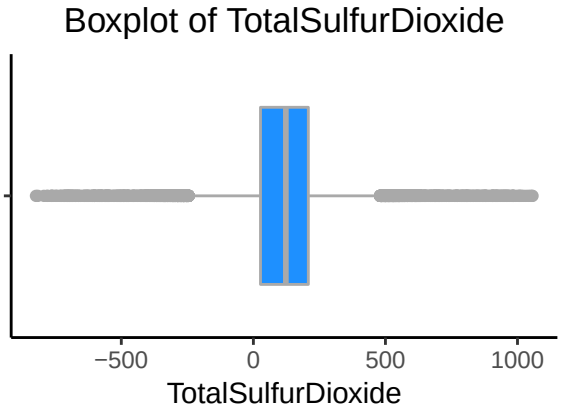
```
## Warning: Removed 682 rows containing non-finite values (stat_bin).
```

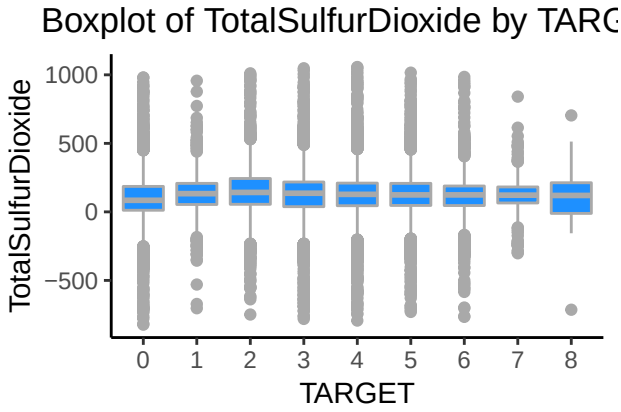
```
## Warning: Removed 682 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 682 rows containing non-finite values (stat_boxplot).
```





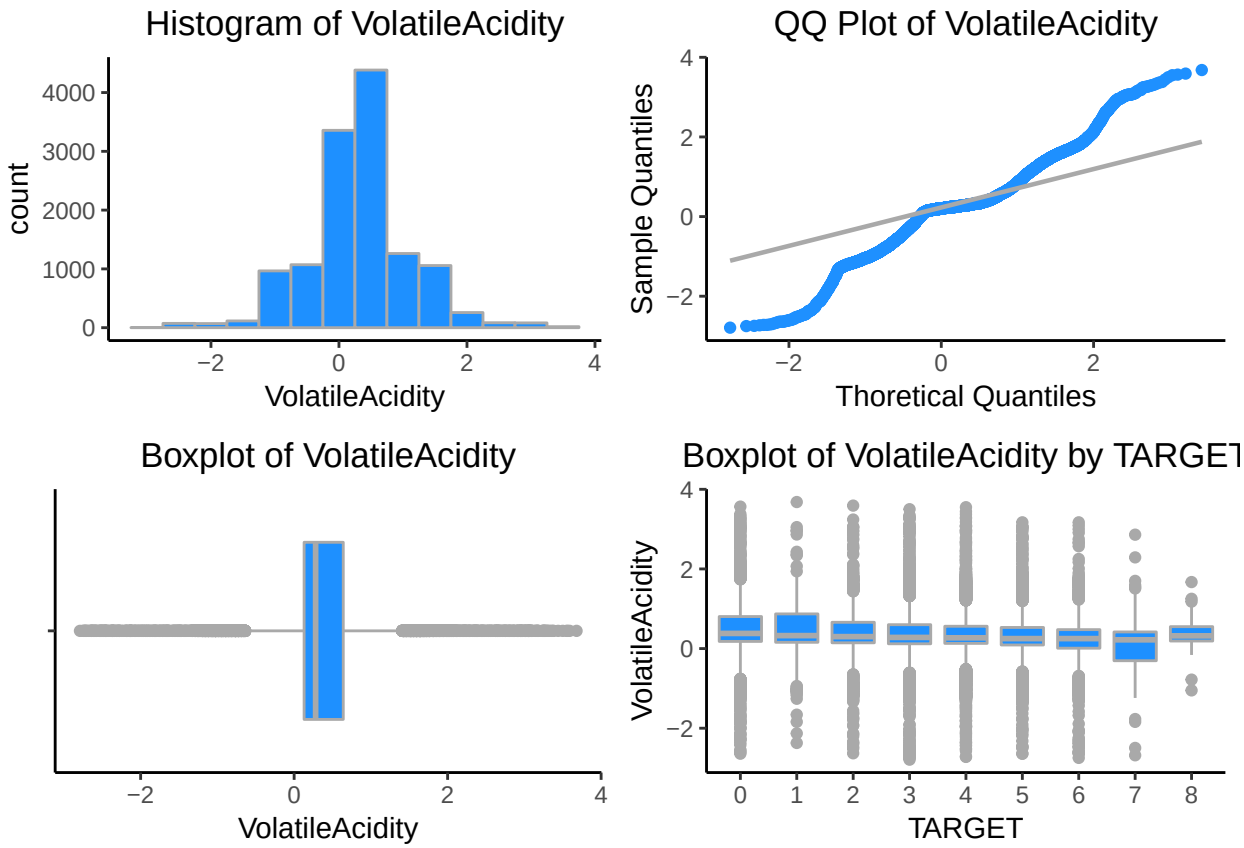




Here, I will find the information about VolatileAcidity. This variable tells us about the VolatileAcidity content of Wine.

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## -2.79000000  0.13000000  0.28000000  0.32410395  0.64000000  3.68000000
##           SD         Skew         Kurt
##  0.78401424  0.02038235  4.83296606
```

Now, I will draw the hitogram, QQ plot, Boxplot of VolatileAcidity variable. I will also draw the boxplot of the variable by TARGET variable to see how much VolatileAcidity are available in each number of cases.



Here, I will find the information about pH. This variable tells us about the pH of Wine.

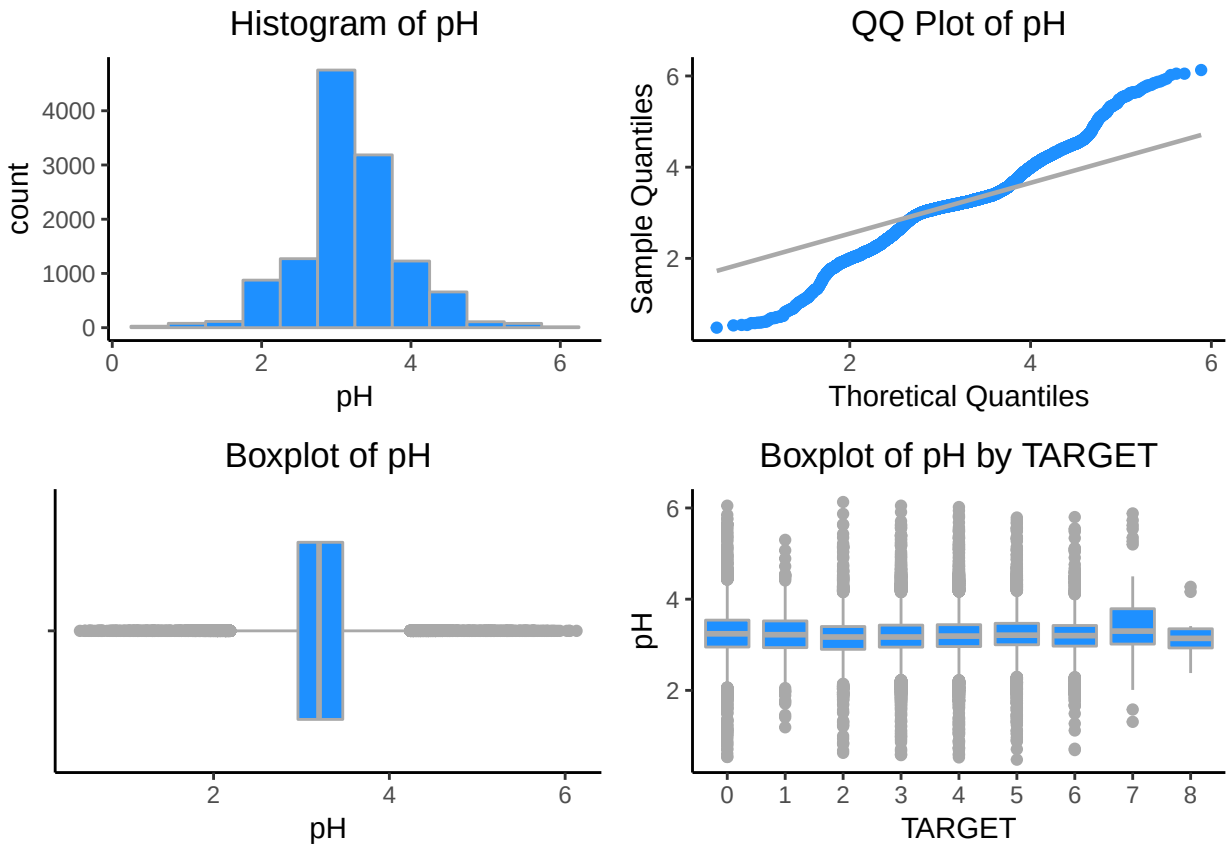
```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
##   0.480000   2.960000   3.200000   3.207628   3.470000   6.130000
##       NA's         SD       Skew       Kurt
## 395.000000         NA         NA         NA
```

Now, I will draw the hitogram, QQ plot, Boxplot of PH variable. I will also draw the boxplot of the variable by TARGET variable to see how much PH are available in each number of cases.

```
## Warning: Removed 395 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 395 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 395 rows containing non-finite values (stat_boxplot).
```



Now, I will find the missing values also called NA's for all the variables.

Finding NA's using NA_Summary:

	Non_NAs	NAs	NA_Percent
TARGET	12795	0	0.0000000
FixedAcidity	12795	0	0.0000000
VolatileAcidity	12795	0	0.0000000
CitricAcid	12795	0	0.0000000
ResidualSugar	12179	616	0.0481438
Chlorides	12157	638	0.0498632
FreeSulfurDioxide	12148	647	0.0505666
TotalSulfurDioxide	12113	682	0.0533021
Density	12795	0	0.0000000
pH	12400	395	0.0308714
Sulphates	11585	1210	0.0945682
Alcohol	12142	653	0.0510356
LabelAppeal	12795	0	0.0000000
AcidIndex	12795	0	0.0000000
STARS	9436	3359	0.2625244

From the result we see that the highest number of missing vaules are in STARS variable which is 3359. It is almost 27%.

Finding the summary of the

```
##  [1] "TARGET"             "FixedAcidity"       "VolatileAcidity"
##  [4] "CitricAcid"         "ResidualSugar"      "Chlorides"
##  [7] "FreeSulfurDioxide"  "TotalSulfurDioxide" "Density"
```

```
## [10] "pH"                "Sulphates"          "Alcohol"
## [13] "LabelAppeal"       "AcidIndex"          "STARS"
```

Finding the correlation between the variables by using cor function.

```
##                         TARGET FixedAcidity VolatileAcidity
## TARGET              1.0000000000 -0.012538100   -0.0759978765
## FixedAcidity       -0.0125380998  1.000000000    0.0190109733
## VolatileAcidity    -0.0759978765  0.019010973    1.0000000000
## CitricAcid          0.0023450490  0.014000376   -0.0234315631
## ResidualSugar       0.0035195999 -0.015429391    0.0015279517
## Chlorides          -0.0304301331 -0.006104447    0.0148489225
## FreeSulfurDioxide   0.0226398054  0.015438463   -0.0114408079
## TotalSulfurDioxide  0.0216020726 -0.023323485   -0.0007434083
## Density            -0.0475989086  0.011574241    0.0130977690
## pH                  0.0002198557 -0.004553886    0.0072030364
## Sulphates          -0.0212203783  0.042229181    0.0015161001
## Alcohol             0.0737771084 -0.013085026    0.0002603082
## LabelAppeal         0.4979464796  0.011375965   -0.0202419713
## AcidIndex          -0.1676430648  0.154167846    0.0250529742
## STARS               0.5546857223 -0.004937345   -0.0402432388
##                       CitricAcid ResidualSugar     Chlorides
## TARGET              0.0023450490   0.003519600 -0.0304301331
## FixedAcidity        0.0140003760  -0.015429391 -0.0061044471
## VolatileAcidity    -0.0234315631   0.001527952  0.0148489225
## CitricAcid          1.0000000000  -0.009843146 -0.0335608661
## ResidualSugar      -0.0098431456   1.000000000  0.0041215692
## Chlorides          -0.0335608661   0.004121569  1.0000000000
## FreeSulfurDioxide   0.0121132485   0.021959113 -0.0204924876
## TotalSulfurDioxide -0.0099174506   0.017030939  0.0004188605
## Density            -0.0169919691  -0.007120841  0.0206724860
## pH                 -0.0007581304   0.017563769 -0.0179702278
## Sulphates          -0.0144237270  -0.002705775  0.0026187777
## Alcohol             0.0169864284  -0.018943324 -0.0228849573
## LabelAppeal         0.0153315666  -0.004579308 -0.0063870237
## AcidIndex           0.0545838104  -0.020301890 -0.0017134096
## STARS               0.0071401699   0.019665541 -0.0063242568
##                    FreeSulfurDioxide TotalSulfurDioxide      Density
## TARGET                    0.022639805       0.0216020726 -0.047598909
## FixedAcidity              0.015438463      -0.0233234848  0.011574241
## VolatileAcidity          -0.011440808      -0.0007434083  0.013097769
## CitricAcid                0.012113248      -0.0099174506 -0.016991969
## ResidualSugar             0.021959113       0.0170309394 -0.007120841
## Chlorides                -0.020492488       0.0004188605  0.020672486
## FreeSulfurDioxide         1.000000000       0.0134616726 -0.008663509
## TotalSulfurDioxide        0.013461673       1.0000000000  0.023167955
## Density                  -0.008663509       0.0231679548  1.000000000
## pH                       -0.002008516      -0.0034227601 -0.002019229
## Sulphates                 0.026829029       0.0025040509 -0.010609294
## Alcohol                  -0.023867458      -0.0168515467 -0.006128355
## LabelAppeal               0.014960087      -0.0027237419 -0.018094403
## AcidIndex                -0.014733717      -0.0221292631  0.047778830
## STARS                    -0.015390398       0.0220949002 -0.028492455
##                                pH    Sulphates       Alcohol   LabelAppeal
```

```
## TARGET              0.0002198557 -0.021220378  0.0737771084  0.4979464796
## FixedAcidity       -0.0045538857  0.042229181 -0.0130850260  0.0113759650
## VolatileAcidity     0.0072030364  0.001516100  0.0002603082 -0.0202419713
## CitricAcid         -0.0007581304 -0.014423727  0.0169864284  0.0153315666
## ResidualSugar       0.0175637691 -0.002705775 -0.0189433242 -0.0045793083
## Chlorides          -0.0179702278  0.002618778 -0.0228849573 -0.0063870237
## FreeSulfurDioxide  -0.0020085157  0.026829029 -0.0238674577  0.0149600871
## TotalSulfurDioxide -0.0034227601  0.002504051 -0.0168515467 -0.0027237419
## Density            -0.0020192285 -0.010609294 -0.0061283546 -0.0180944026
## pH                  1.0000000000  0.010449255 -0.0122034469  0.0002181758
## Sulphates           0.0104492547  1.000000000  0.0108443299  0.0037686996
## Alcohol            -0.0122034469  0.010844330  1.0000000000 -0.0006449123
## LabelAppeal         0.0002181758  0.003768700 -0.0006449123  1.0000000000
## AcidIndex          -0.0537128921  0.031071782 -0.0558919056  0.0103009840
## STARS              -0.0044002985 -0.023135130  0.0648544864  0.3188970216
##                      AcidIndex        STARS
## TARGET             -0.16764306  0.554685722
## FixedAcidity        0.15416785 -0.004937345
## VolatileAcidity     0.02505297 -0.040243239
## CitricAcid          0.05458381  0.007140170
## ResidualSugar      -0.02030189  0.019665541
## Chlorides          -0.00171341 -0.006324257
## FreeSulfurDioxide  -0.01473372 -0.015390398
## TotalSulfurDioxide -0.02212926  0.022094900
## Density             0.04777883 -0.028492455
## pH                 -0.05371289 -0.004400299
## Sulphates           0.03107178 -0.023135130
## Alcohol            -0.05589191  0.064854486
## LabelAppeal         0.01030098  0.318897022
## AcidIndex           1.00000000 -0.095482582
## STARS              -0.09548258  1.000000000
```

Now we will see the TARGET Variable. This variable tells us about the number cases purchased.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    StdD    Skew    Kurt
##    0.00    2.00    3.00    3.03    4.00    8.00    1.93   -0.33    2.12
```

(2) DATA PREPARATION: We will prepare our data by fixing missing values, transform the data etc. First, i will split the data into training and test.

We will now use the mice package to impute missing values.

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
```

```
##
##  iter imp variable
```

```
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
```

There is very low correlation between AcidIndex and TARGET. Therefore, I will do log transformation on AcidIndex to build our model as follows.

(3) BUILD MODELS:

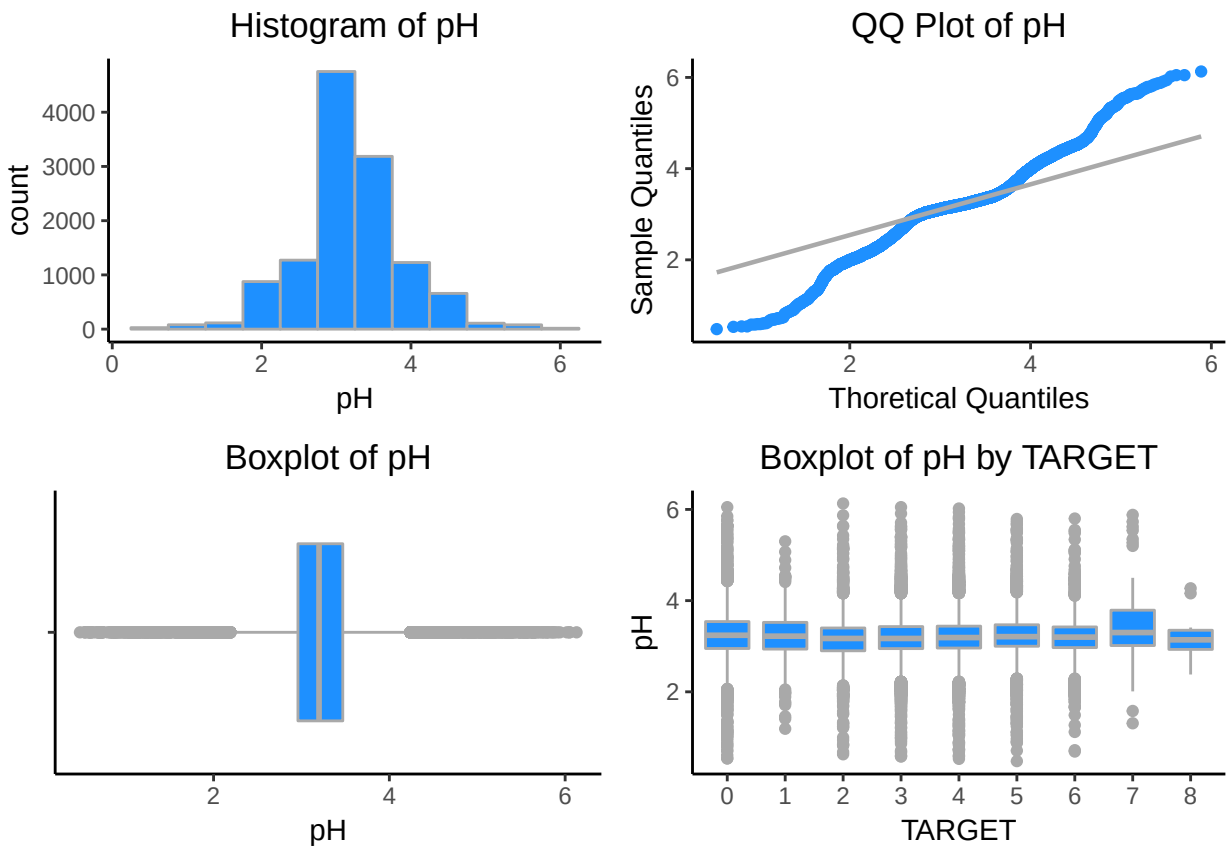
Now, I will build different models by using the wine training data. At first, I will build Poisson model without imputations.

1.Poisson model without imputations: Here is the model and the summary of the model:

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = wine_train1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2128  -0.2757   0.0647   0.3766   1.6981
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.608e+00  2.796e-01   5.750 8.90e-09 ***
## FixedAcidity         6.705e-04  1.177e-03   0.570  0.56901
## VolatileAcidity     -2.750e-02  9.283e-03  -2.963  0.00305 **
## CitricAcid          -3.835e-03  8.519e-03  -0.450  0.65259
## ResidualSugar        1.828e-05  2.152e-04   0.085  0.93232
## Chlorides           -3.764e-02  2.314e-02  -1.627  0.10377
## FreeSulfurDioxide    5.671e-05  4.892e-05   1.159  0.24630
## TotalSulfurDioxide   2.230e-05  3.177e-05   0.702  0.48274
## Density             -4.025e-01  2.749e-01  -1.464  0.14326
## pH                   2.307e-04  1.085e-02   0.021  0.98303
## Sulphates           -5.984e-03  7.973e-03  -0.751  0.45293
## Alcohol              3.262e-03  2.004e-03   1.628  0.10360
## LabelAppeal          1.730e-01  8.858e-03  19.530  < 2e-16 ***
## AcidIndex           -4.967e-02  6.666e-03  -7.451 9.28e-14 ***
## STARS                1.929e-01  8.328e-03  23.160  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4720.5  on 5143  degrees of freedom
## Residual deviance: 3242.8  on 5129  degrees of freedom
##   (5093 observations deleted due to missingness)
## AIC: 18545
##
## Number of Fisher Scoring iterations: 5
```

Now, I will draw the histogram, qq plot, and the box plot if PH and also the boxplot of PH by TARGET variable.

```
## Warning: Removed 395 rows containing non-finite values (stat_bin).

## Warning: Removed 395 rows containing non-finite values (stat_boxplot).

## Warning: Removed 395 rows containing non-finite values (stat_boxplot).
```

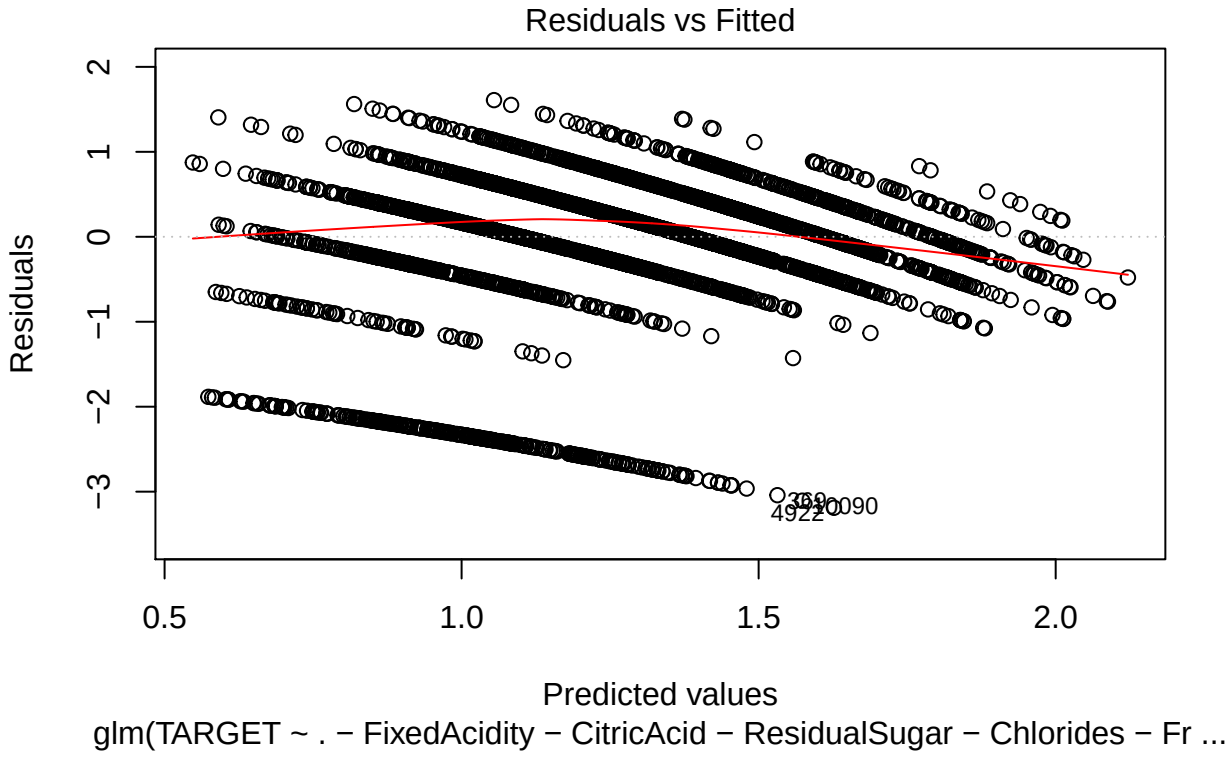


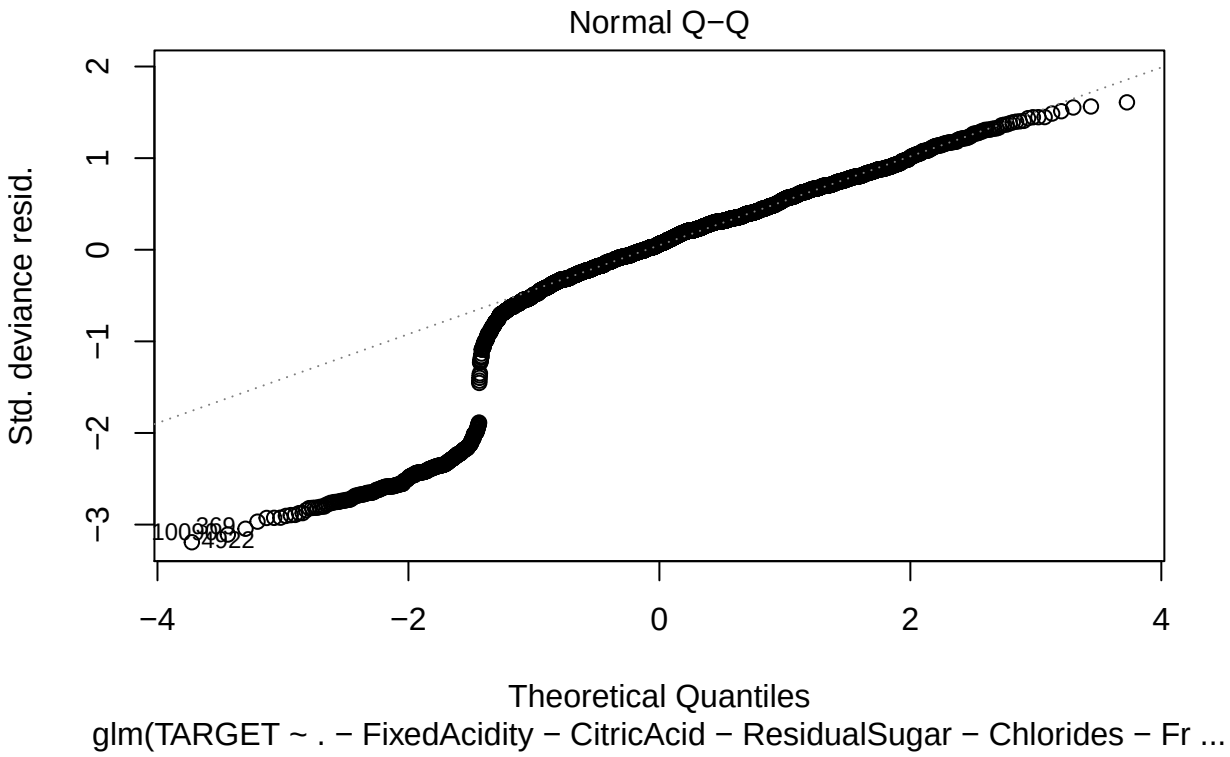
2. Poisson model without imputations and only with significant variables: Next, I have build the Poisson model without imputations and only with significant variables. The model and the summary of the model is as follows:

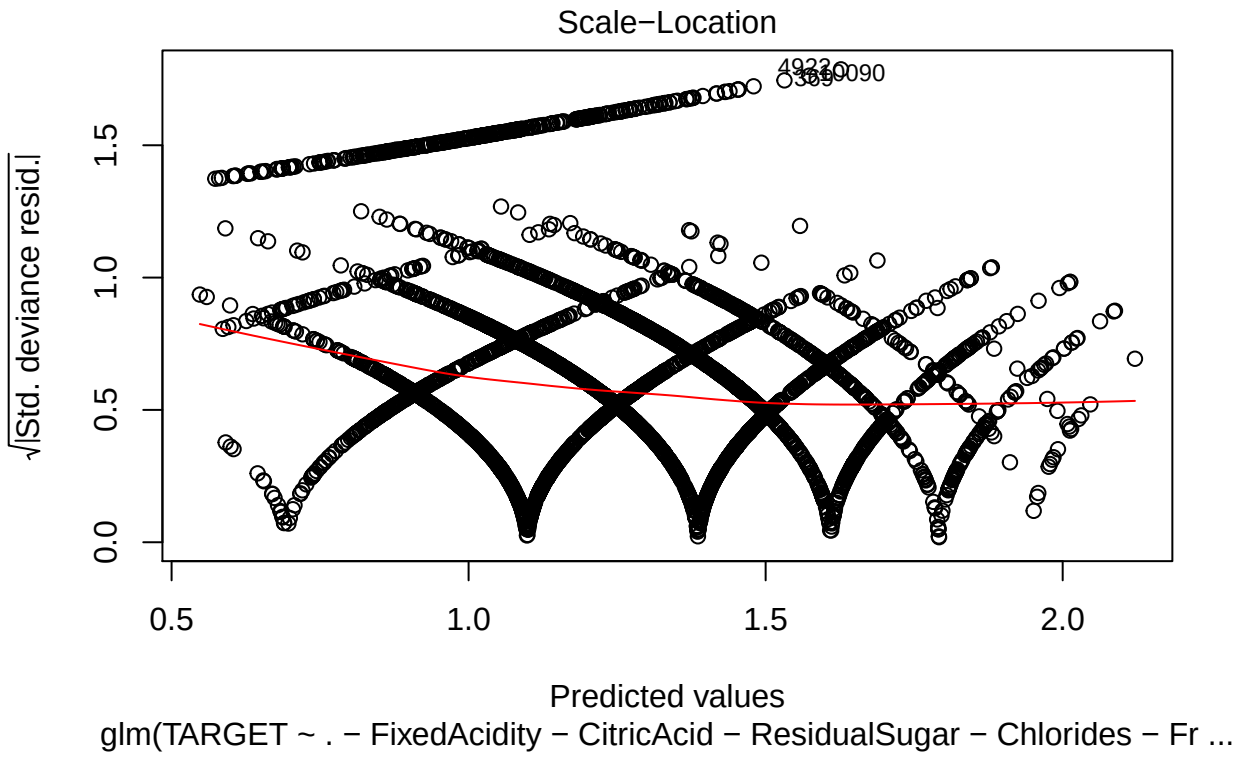
```
##
## Call:
## glm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##     Chlorides - FreeSulfurDioxide - TotalSulfurDioxide - Density -
##     pH - Sulphates - Alcohol, family = poisson, data = wine_train1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1898  -0.2777   0.0622   0.3764   1.6086
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.251442   0.054724  22.868  < 2e-16 ***
## VolatileAcidity -0.027581   0.009278  -2.973  0.00295 **
## LabelAppeal       0.173177   0.008853  19.562  < 2e-16 ***
```

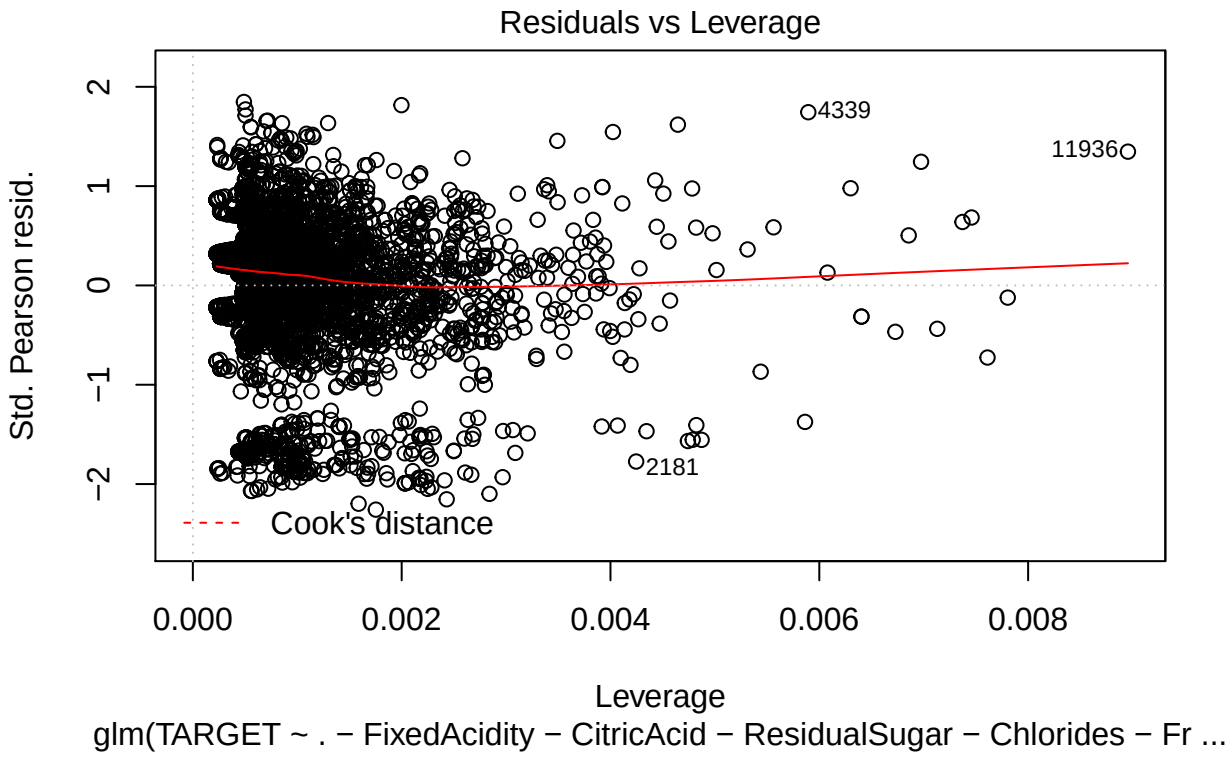
```
## AcidIndex       -0.050616   0.006553  -7.724 1.13e-14 ***
## STARS            0.194208   0.008292  23.421  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4720.5  on 5143  degrees of freedom
## Residual deviance: 3253.1  on 5139  degrees of freedom
##   (5093 observations deleted due to missingness)
## AIC: 18535
##
## Number of Fisher Scoring iterations: 5
```

I have plotted my Poisson model without imputations and only with significant variables with Residuals vs Fitted, Normal Q-Q, Scale-Location, and finally Residuals vs Leverage as follows:




Normal Q−Q
Std. deviance resid.
2
1
0
−1
−2
−3
−4
−2
0
2
4
Theoretical Quantiles
glm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Chlorides − Fr ...


Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
Predicted values
glm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Chlorides − Fr ...



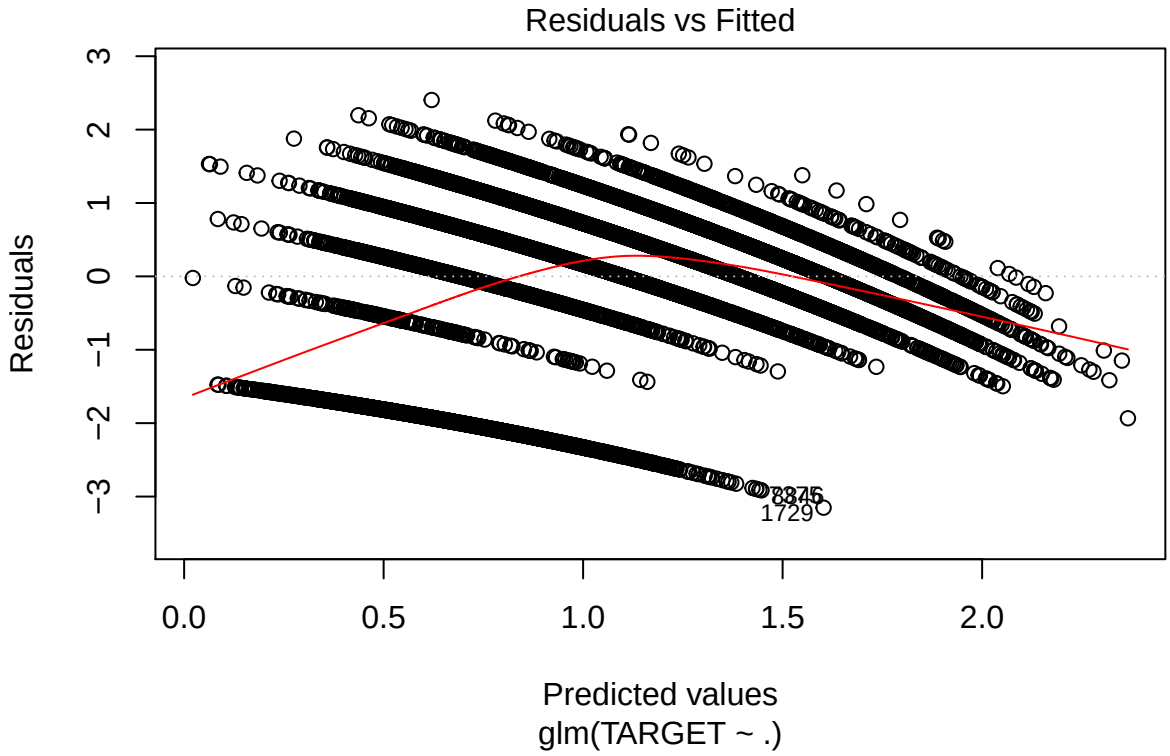
3. Poisson model with Imputation: Here, I have drawn the poisson model with imputation as follows.

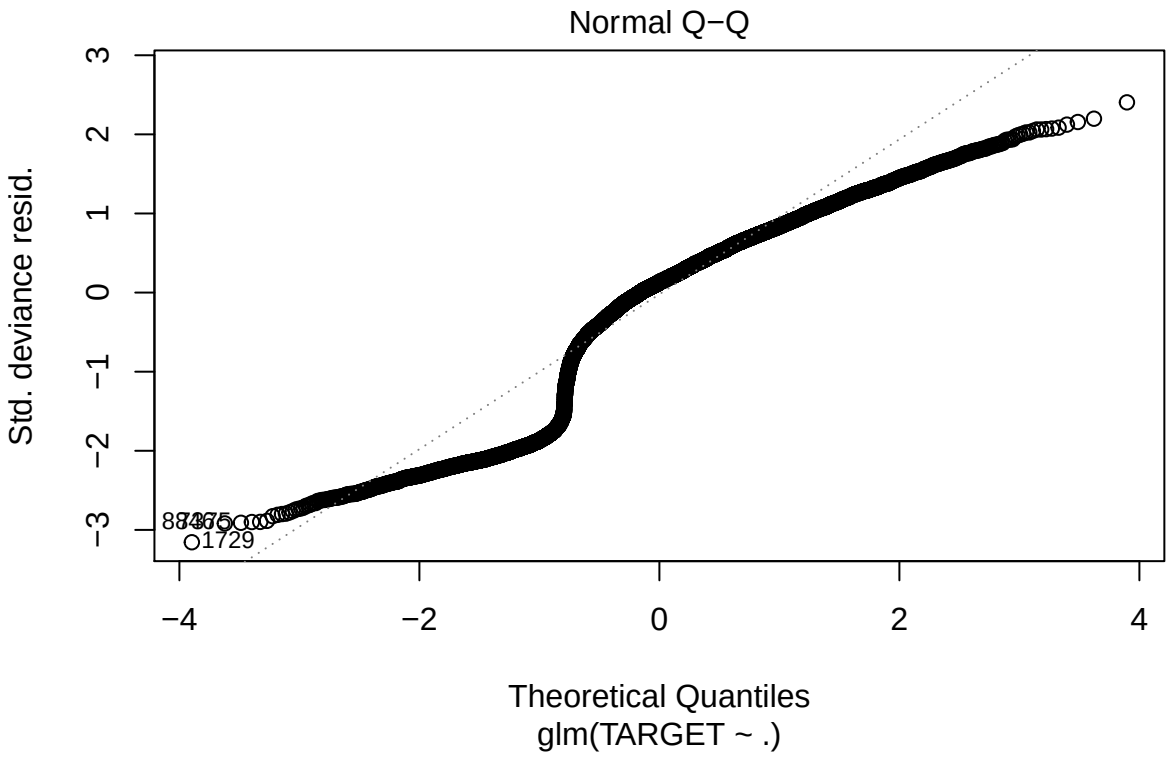
The summary of the model have provided bellow.

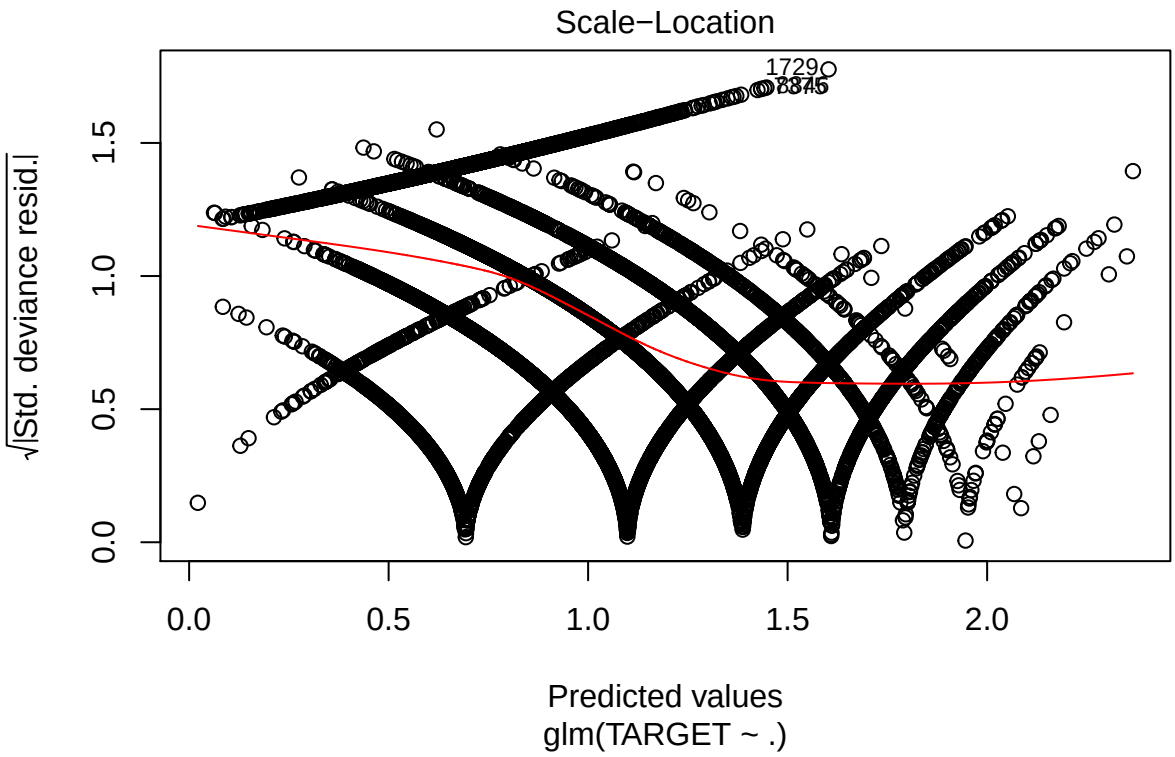
```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = wine_train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1516  -0.6809   0.1304   0.6390   2.4033
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.382e+00  2.277e-01  10.463  < 2e-16 ***
## FixedAcidity        -1.332e-04  9.197e-04  -0.145  0.88487
## VolatileAcidity     -4.351e-02  7.275e-03  -5.982 2.21e-09 ***
## CitricAcid           8.883e-03  6.576e-03   1.351  0.17679
## ResidualSugar        1.508e-04  1.675e-04   0.900  0.36797
## Chlorides           -6.506e-02  1.791e-02  -3.633  0.00028 ***
## FreeSulfurDioxide    1.143e-04  3.804e-05   3.005  0.00266 **
## TotalSulfurDioxide   8.709e-05  2.446e-05   3.560  0.00037 ***
## Density             -4.047e-01  2.141e-01  -1.890  0.05876 .
## pH                  -1.788e-02  8.407e-03  -2.126  0.03347 *
## Sulphates           -1.327e-02  6.163e-03  -2.153  0.03129 *
```

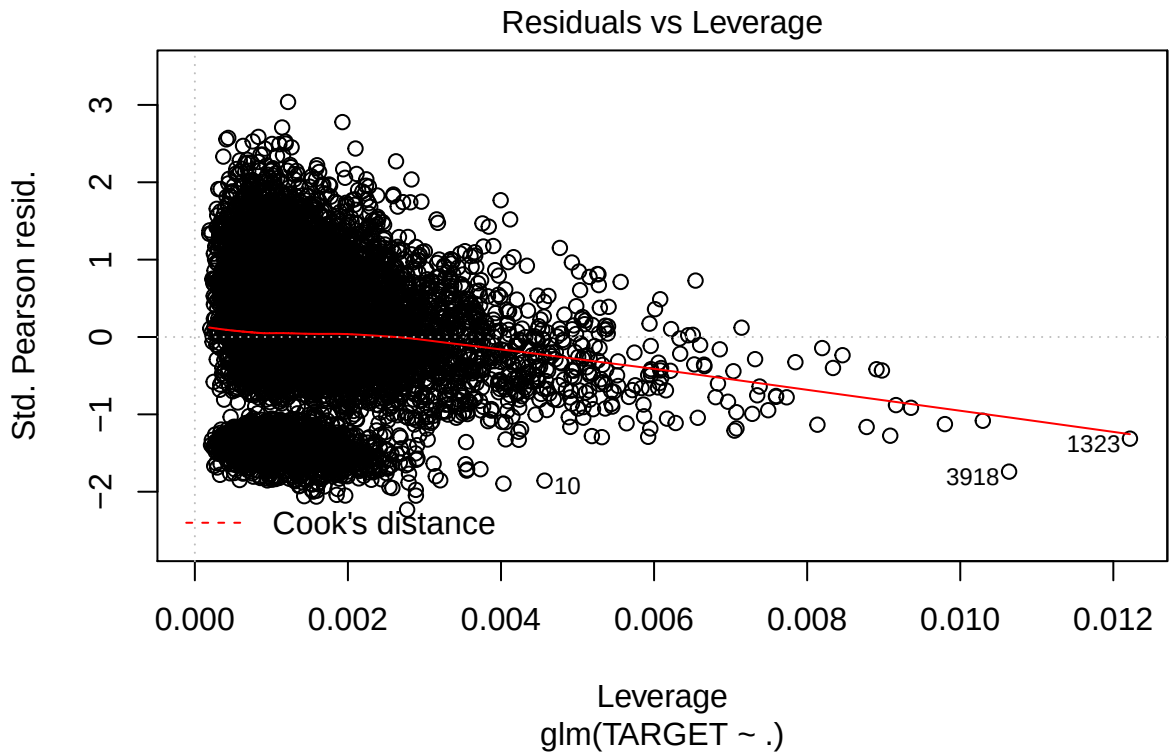
```
## Alcohol             2.690e-03  1.546e-03   1.740  0.08187 .
## LabelAppeal         1.432e-01  6.783e-03  21.107  < 2e-16 ***
## AcidIndex          -7.622e-01  4.005e-02 -19.029  < 2e-16 ***
## STARS               3.401e-01  6.252e-03  54.395  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 18291  on 10236  degrees of freedom
## Residual deviance: 12830  on 10222  degrees of freedom
## AIC: 38418
##
## Number of Fisher Scoring iterations: 5
```

I have plotted my Poisson model with Imputation by Residuals vs Fitted, Normal Q-Q, Scale-Location, and finally Residuals vs Leverage as follows:




Normal Q−Q
Std. deviance resid.
3
2
1
0
−1
−2
−3
1729
−4
−2
0
2
4
Theoretical Quantiles
glm(TARGET ~ .)


Scale−Location
1729
8345
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
Predicted values
glm(TARGET ~ .)



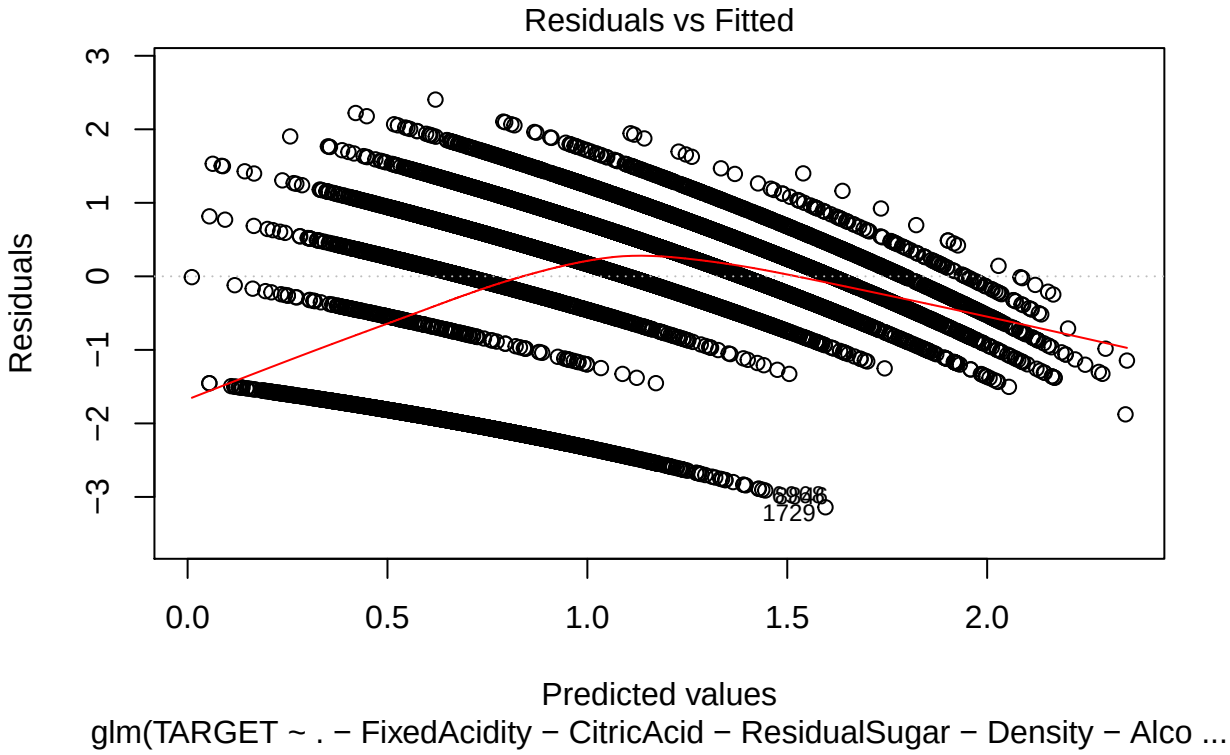
4.Poisson model with imputations and only significant variables: Here, I am building poisson model with imputations and also with only significant variables.

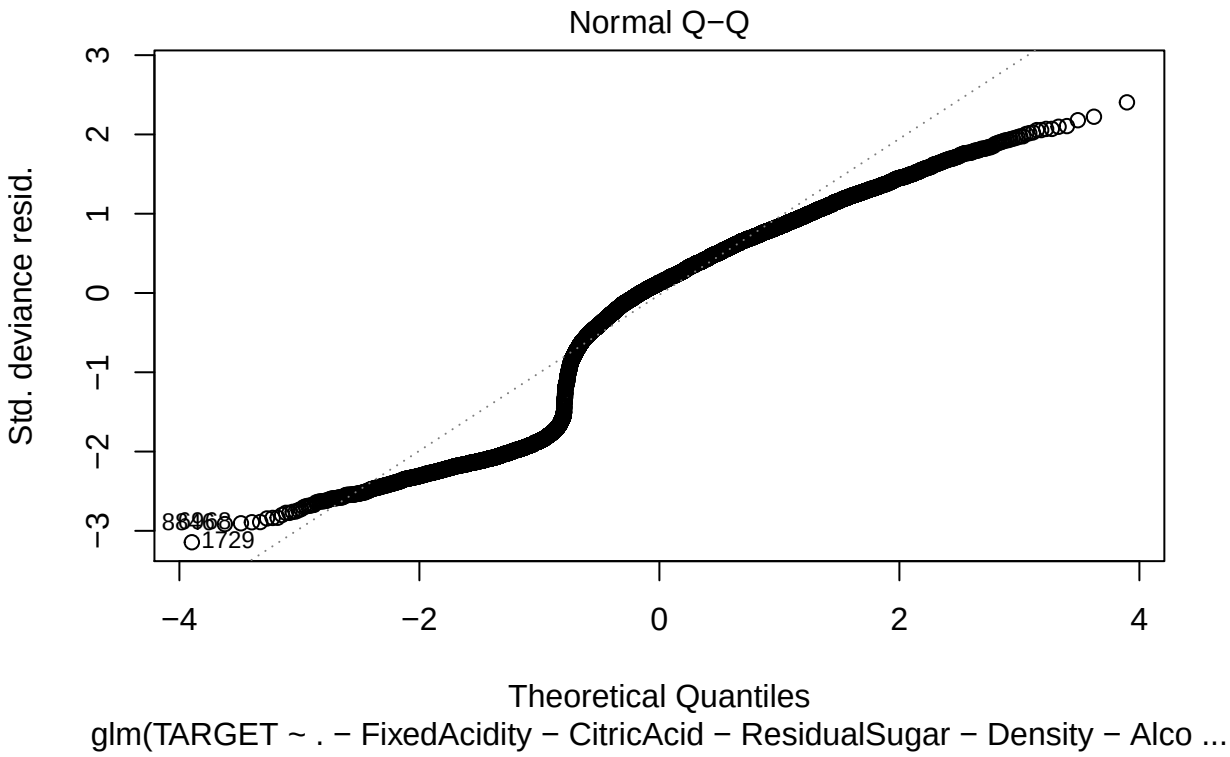
My model and the summary of the model is as below:

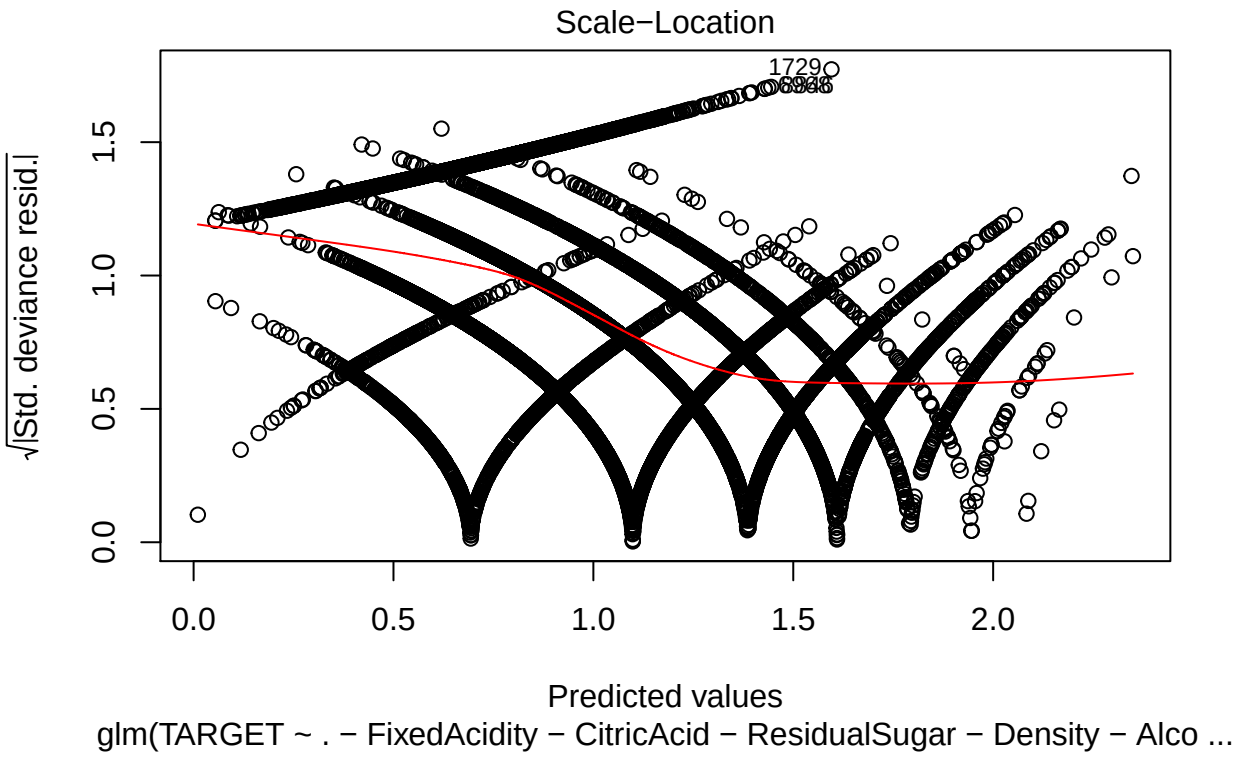
```
##
## Call:
## glm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##     Density - Alcohol, family = poisson, data = wine_train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1405  -0.6852   0.1288   0.6412   2.4039
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.019e+00  8.848e-02  22.820  < 2e-16 ***
## VolatileAcidity     -4.388e-02  7.273e-03  -6.033 1.61e-09 ***
## Chlorides           -6.711e-02  1.790e-02  -3.750 0.000177 ***
## FreeSulfurDioxide    1.119e-04  3.802e-05   2.943 0.003256 **
## TotalSulfurDioxide   8.560e-05  2.442e-05   3.505 0.000457 ***
## pH                  -1.818e-02  8.404e-03  -2.164 0.030488 *
## Sulphates           -1.327e-02  6.157e-03  -2.155 0.031143 *
## LabelAppeal          1.433e-01  6.783e-03  21.120  < 2e-16 ***
## AcidIndex           -7.665e-01  3.941e-02 -19.448  < 2e-16 ***
## STARS                3.410e-01  6.237e-03  54.673  < 2e-16 ***
## ---
```

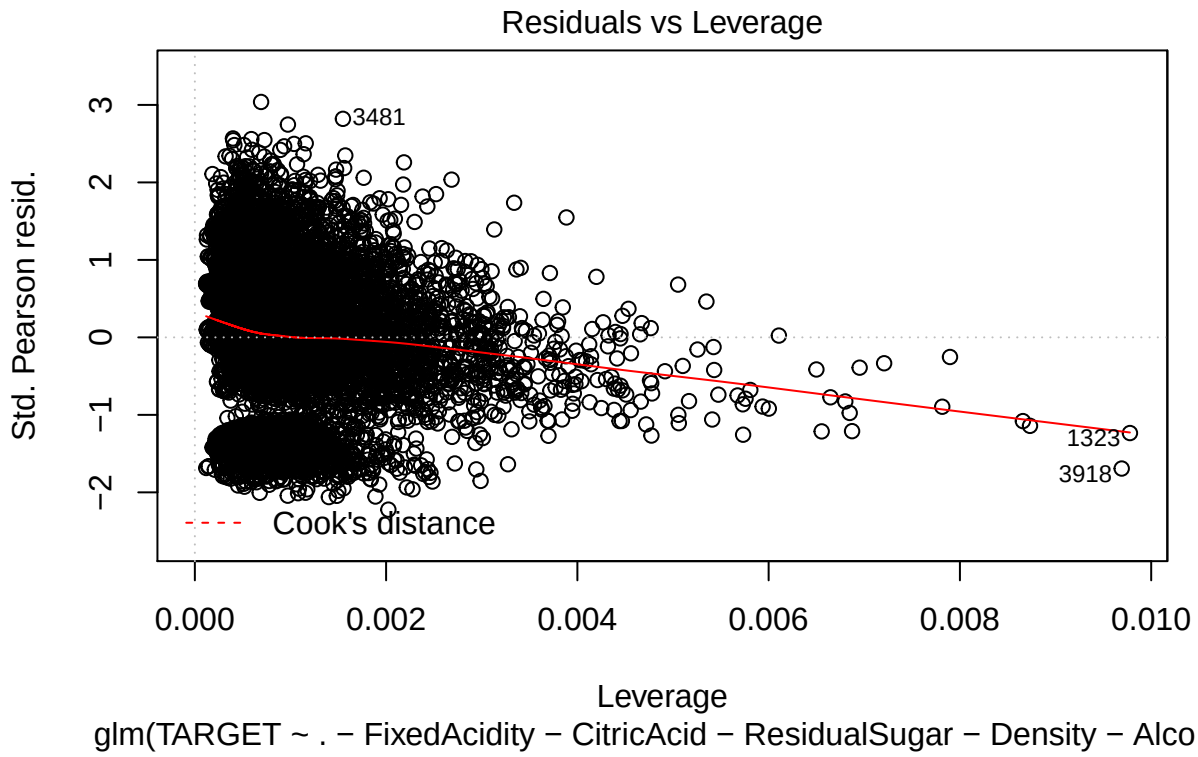
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 18291  on 10236  degrees of freedom
## Residual deviance: 12839  on 10227  degrees of freedom
## AIC: 38417
##
## Number of Fisher Scoring iterations: 5
```

I have plotted my Poisson model with imputations and only significant variables by Residuals vs Fitted, Normal Q-Q, Scale-Location, and finally Residuals vs Leverage as follows:




Normal Q−Q
Std. deviance resid.
3
2
1
0
−1
−2
−3
1729
−4
−2
0
2
4
Theoretical Quantiles
glm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − Alco ...


Scale−Location
1729
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
Predicted values
glm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − Alco ...



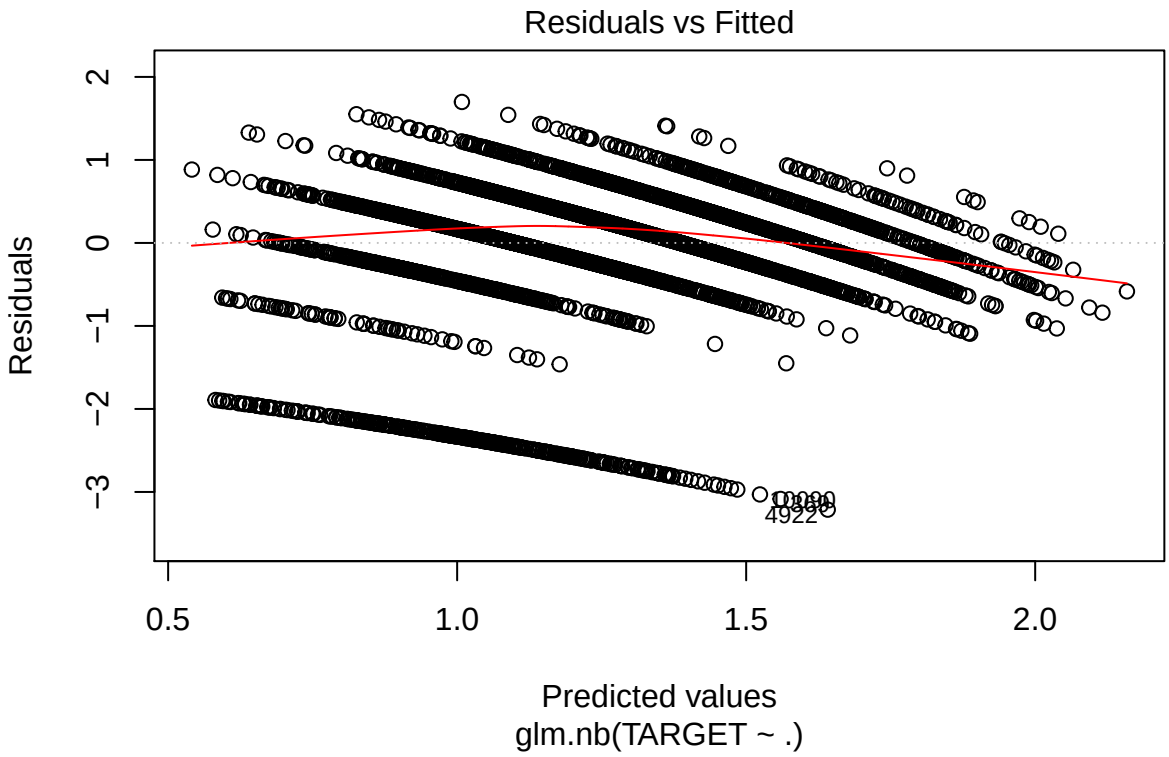
5. Negative Binomial model without imputations: Now, I will build negative binomial model without imputation form the training data.

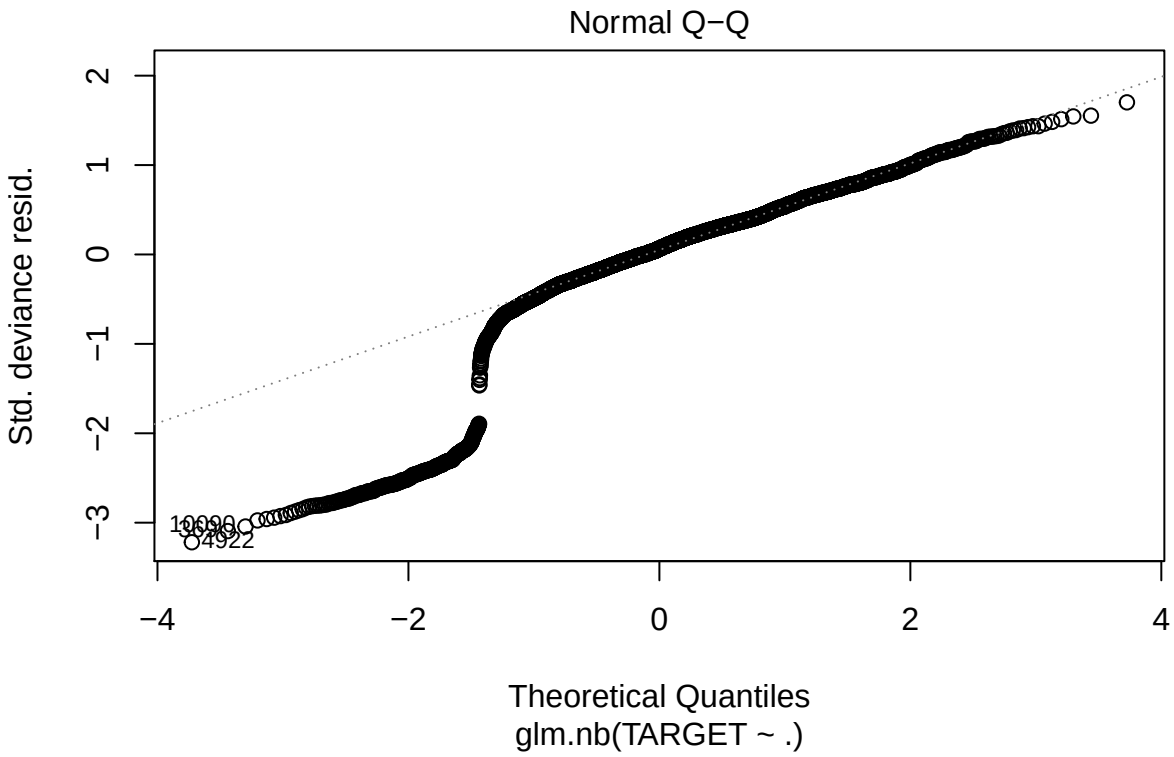
The model and the summary of the model is as below:

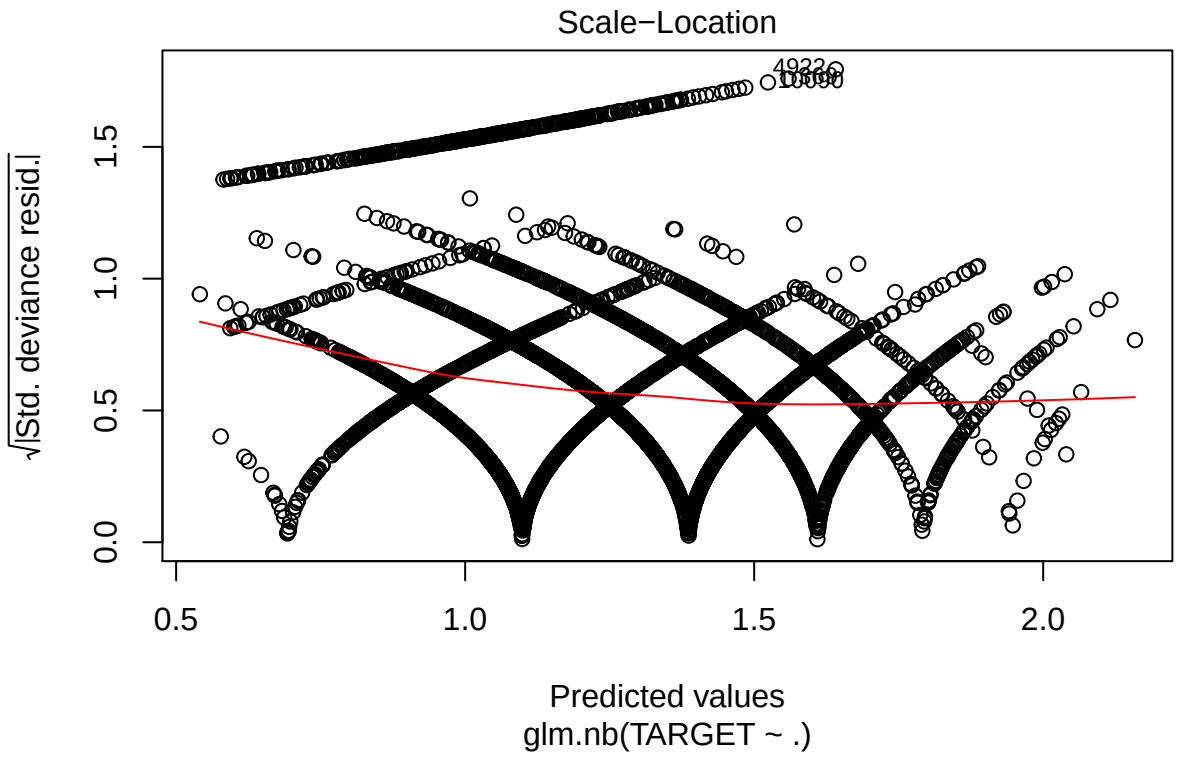
```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace > : iteration
## limit reached

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace > : iteration
## limit reached


##
## Call:
## glm.nb(formula = TARGET ~ ., data = wine_train1, init.theta = 138898.9107,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2127  -0.2757   0.0647   0.3766   1.6981
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.608e+00  2.796e-01   5.750 8.91e-09 ***
## FixedAcidity        6.705e-04  1.177e-03   0.570  0.56900
## VolatileAcidity    -2.750e-02  9.283e-03  -2.963  0.00305 **
```

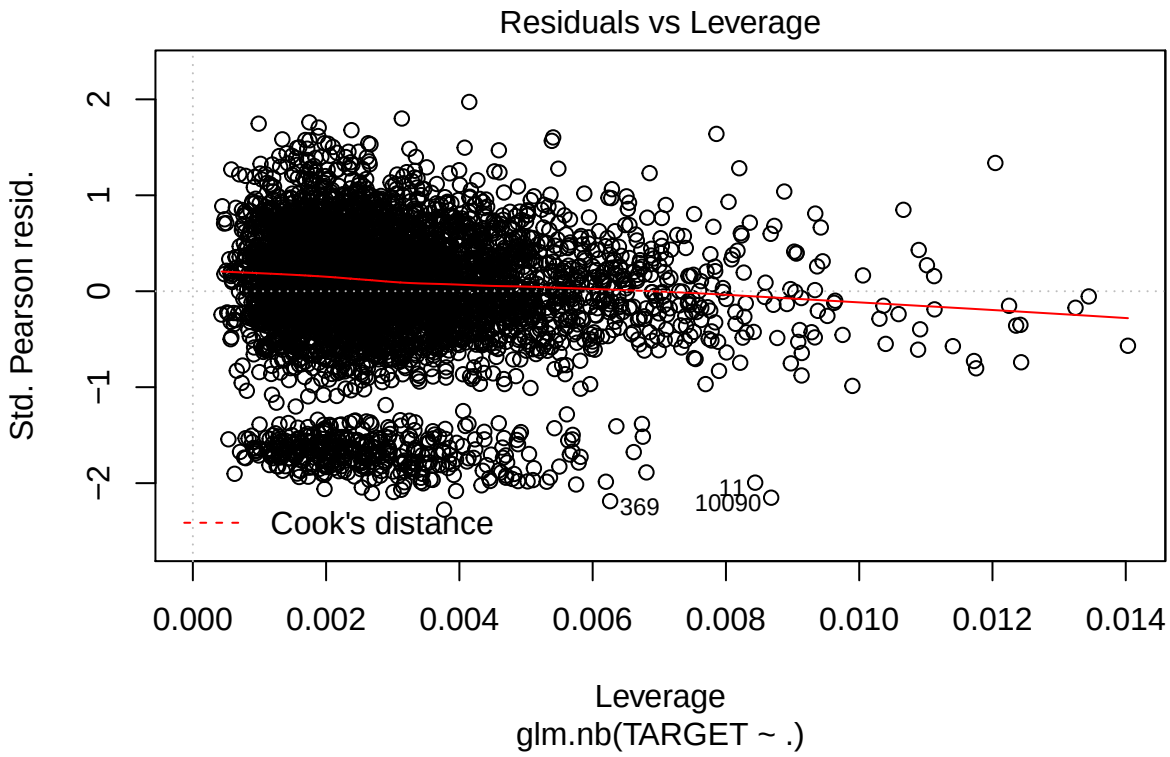
```
## CitricAcid         -3.835e-03  8.519e-03  -0.450  0.65259
## ResidualSugar       1.828e-05  2.152e-04   0.085  0.93231
## Chlorides          -3.764e-02  2.314e-02  -1.627  0.10378
## FreeSulfurDioxide   5.671e-05  4.892e-05   1.159  0.24630
## TotalSulfurDioxide  2.230e-05  3.177e-05   0.702  0.48275
## Density            -4.025e-01  2.750e-01  -1.464  0.14326
## pH                  2.307e-04  1.085e-02   0.021  0.98303
## Sulphates          -5.984e-03  7.973e-03  -0.751  0.45293
## Alcohol             3.262e-03  2.004e-03   1.628  0.10360
## LabelAppeal         1.730e-01  8.858e-03  19.529  < 2e-16 ***
## AcidIndex          -4.967e-02  6.666e-03  -7.451 9.28e-14 ***
## STARS               1.929e-01  8.328e-03  23.160  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(138898.9) family taken to be 1)
##
##     Null deviance: 4720.4  on 5143  degrees of freedom
## Residual deviance: 3242.7  on 5129  degrees of freedom
##   (5093 observations deleted due to missingness)
## AIC: 18547
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  138899
##           Std. Err.:  259921
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -18515.07
```

Here, I have plotted the Negative Binomial model without imputation by Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage as below:


Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
0.5
1.0
1.5
2.0
4922
Predicted values
glm.nb(TARGET ~ .)


Normal Q−Q
Std. deviance resid.
2
1
0
−1
−2
−3
−4
−2
0
2
4
4922
Theoretical Quantiles
glm.nb(TARGET ~ .)


Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
Predicted values
glm.nb(TARGET ~ .)



6. Negative Binomial without imputations and only significant variables: Now, I am building my negative binomial model without imputations and only with significant variables as follows:

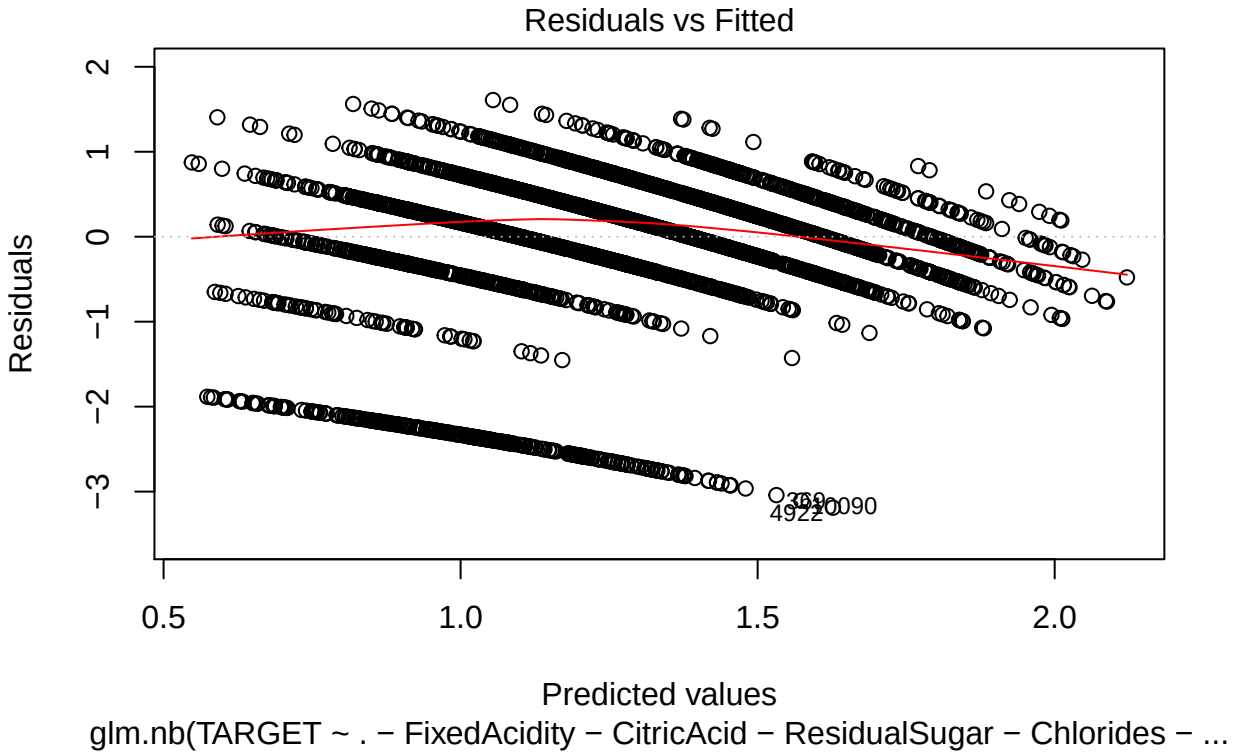
```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace > : iteration
## limit reached

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace > : iteration
## limit reached
```

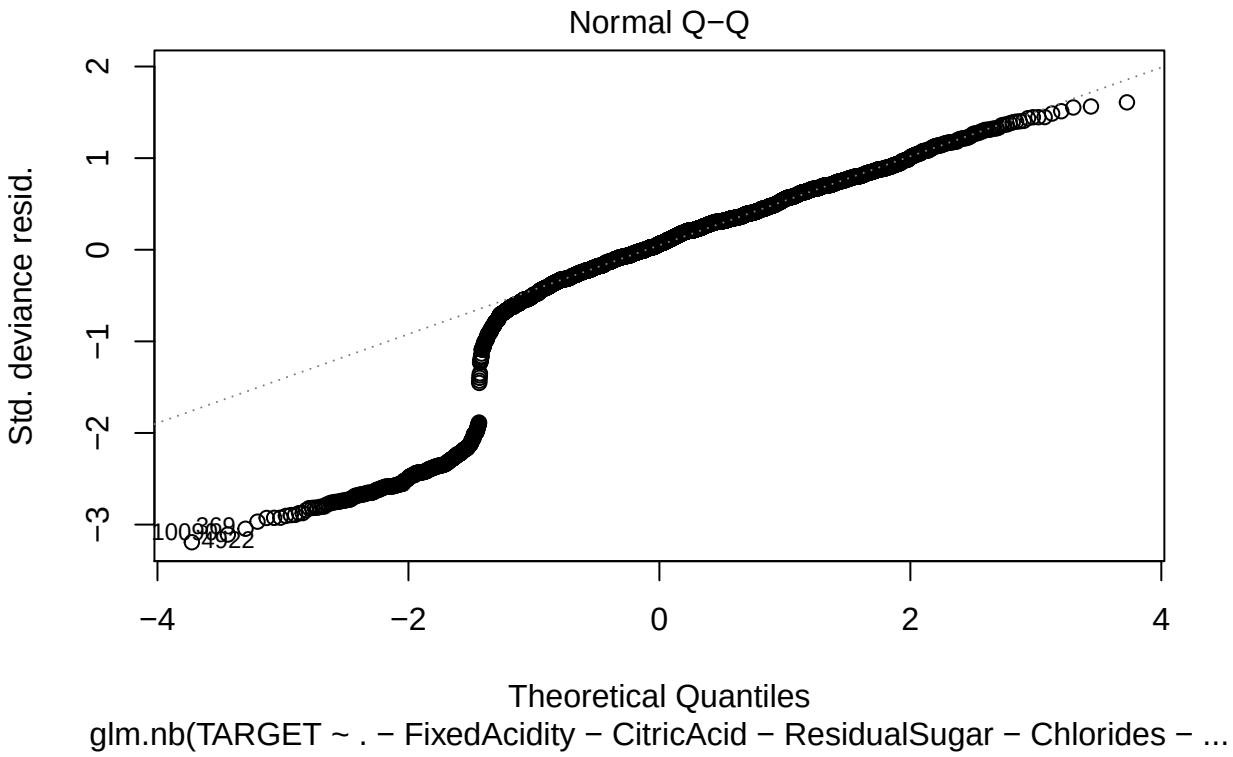
The summary of my negative binomial model without imputations is:

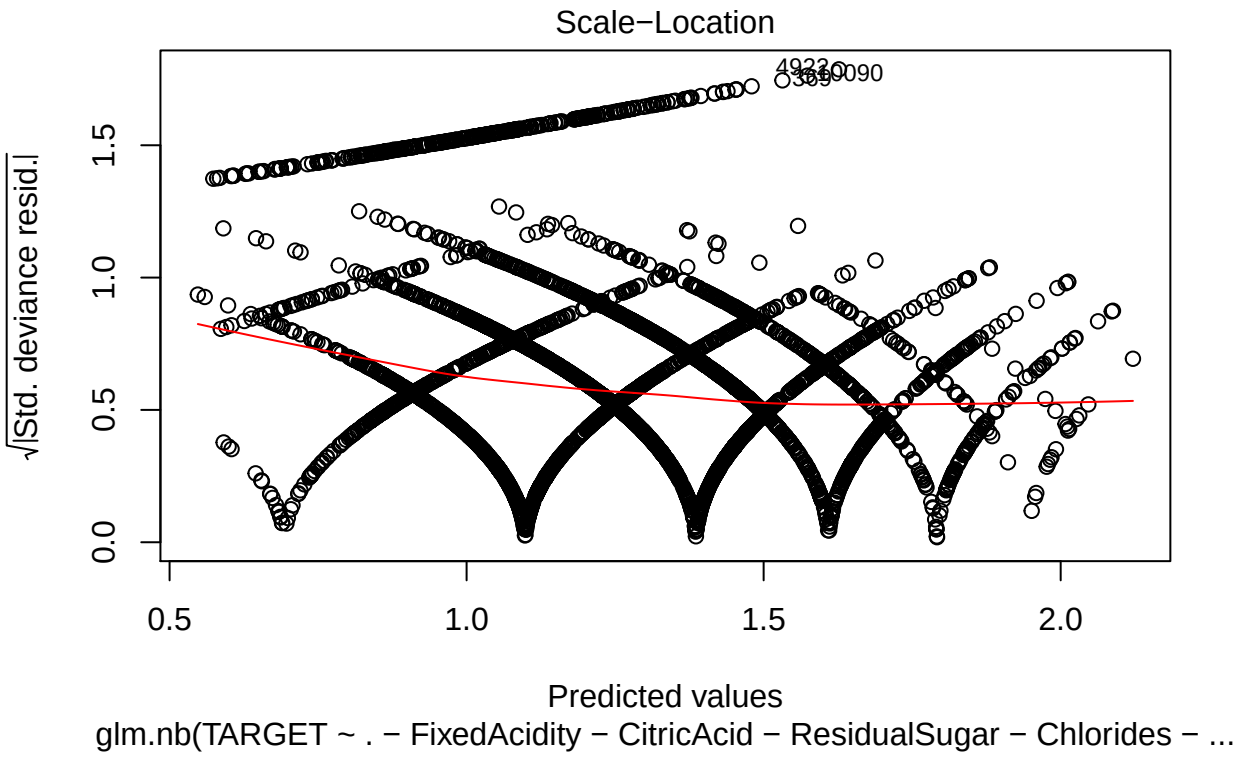
```
##
## Call:
## glm.nb(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##     Chlorides - FreeSulfurDioxide - TotalSulfurDioxide - Density -
##     pH - Sulphates - Alcohol, data = wine_train1, init.theta = 138402.5261,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1898  -0.2777   0.0622   0.3764   1.6086
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.251443   0.054725  22.868  < 2e-16 ***
```

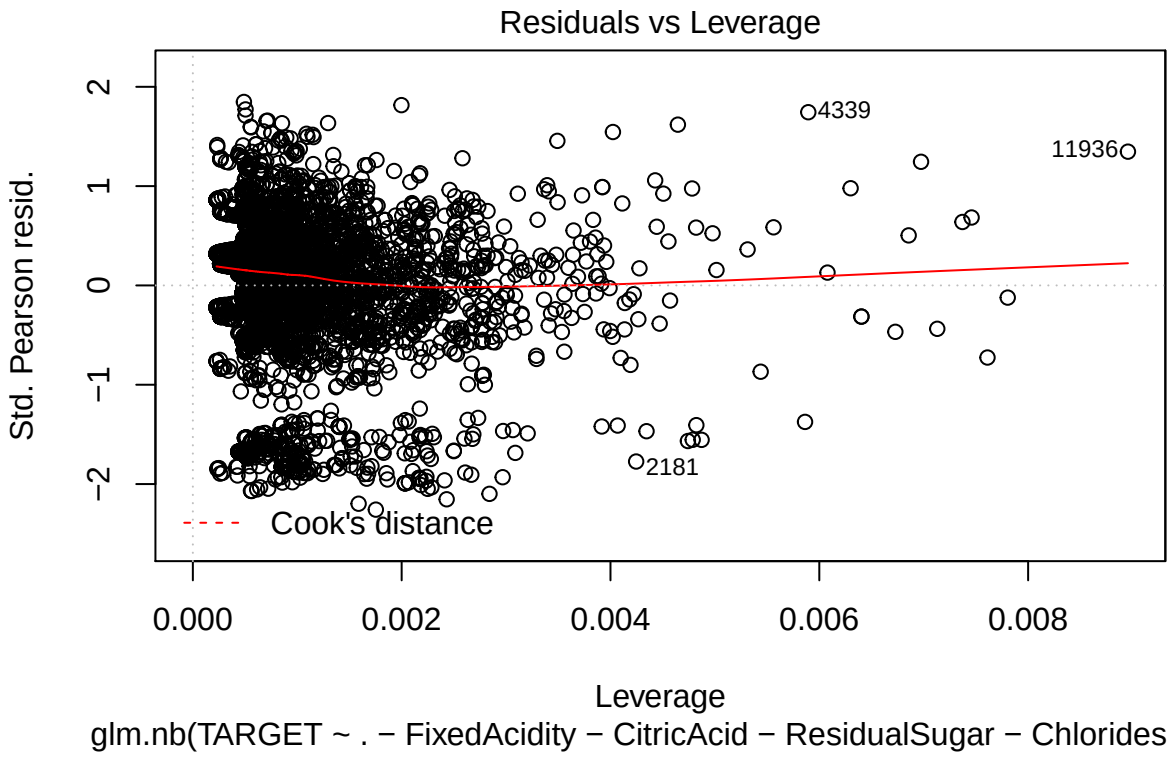
```
## VolatileAcidity -0.027581   0.009279  -2.973  0.00295 **
## LabelAppeal      0.173177   0.008853  19.562  < 2e-16 ***
## AcidIndex       -0.050616   0.006553  -7.724 1.13e-14 ***
## STARS            0.194209   0.008292  23.421  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(138402.5) family taken to be 1)
##
##     Null deviance: 4720.4  on 5143  degrees of freedom
## Residual deviance: 3253.0  on 5139  degrees of freedom
##   (5093 observations deleted due to missingness)
## AIC: 18537
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  138403
##           Std. Err.:  258834
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -18525.37
```

Here, I have plotted the Negative Binomial model without imputation and only with significant variables by Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage as below:




Normal Q−Q
Std. deviance resid.
2
1
0
−1
−2
−3
−4
−2
0
2
4
Theoretical Quantiles
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Chlorides − ...


Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
Predicted values
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Chlorides − ...



7. Negative Binomial with imputations: Now, I will build a negative binomial model with imputation here.

```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace > : iteration
## limit reached
```

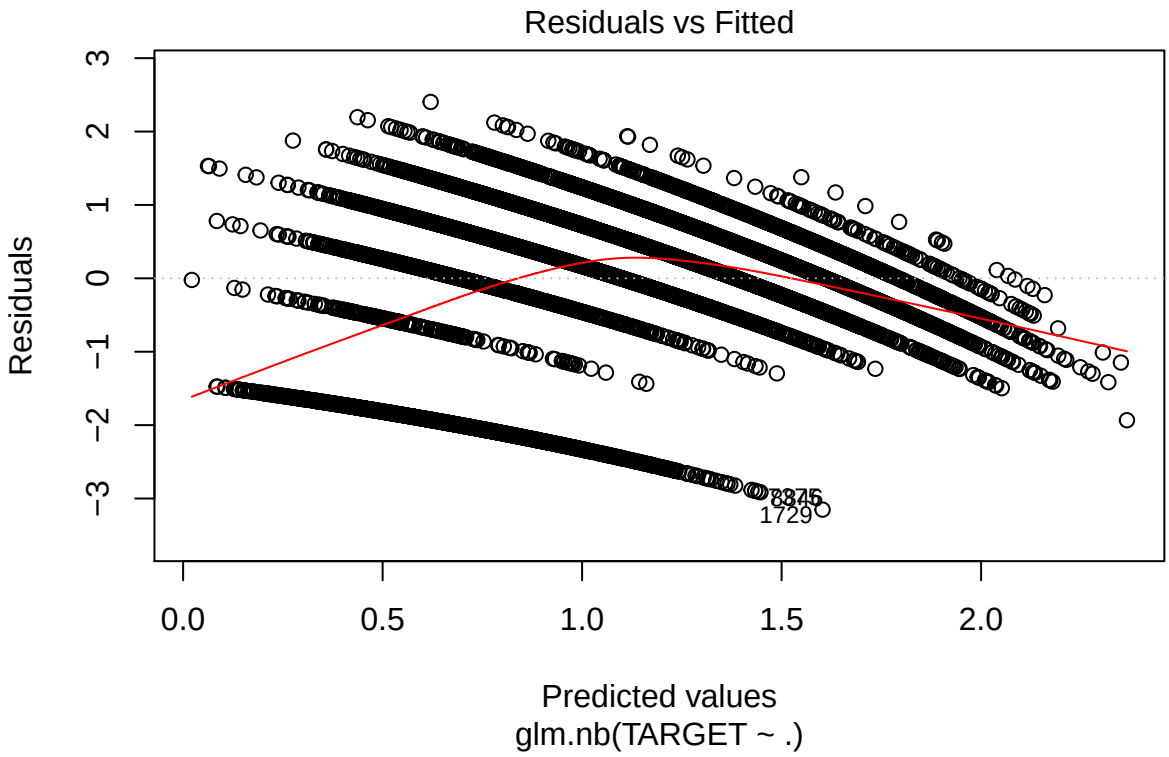
```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace > : iteration
## limit reached
```

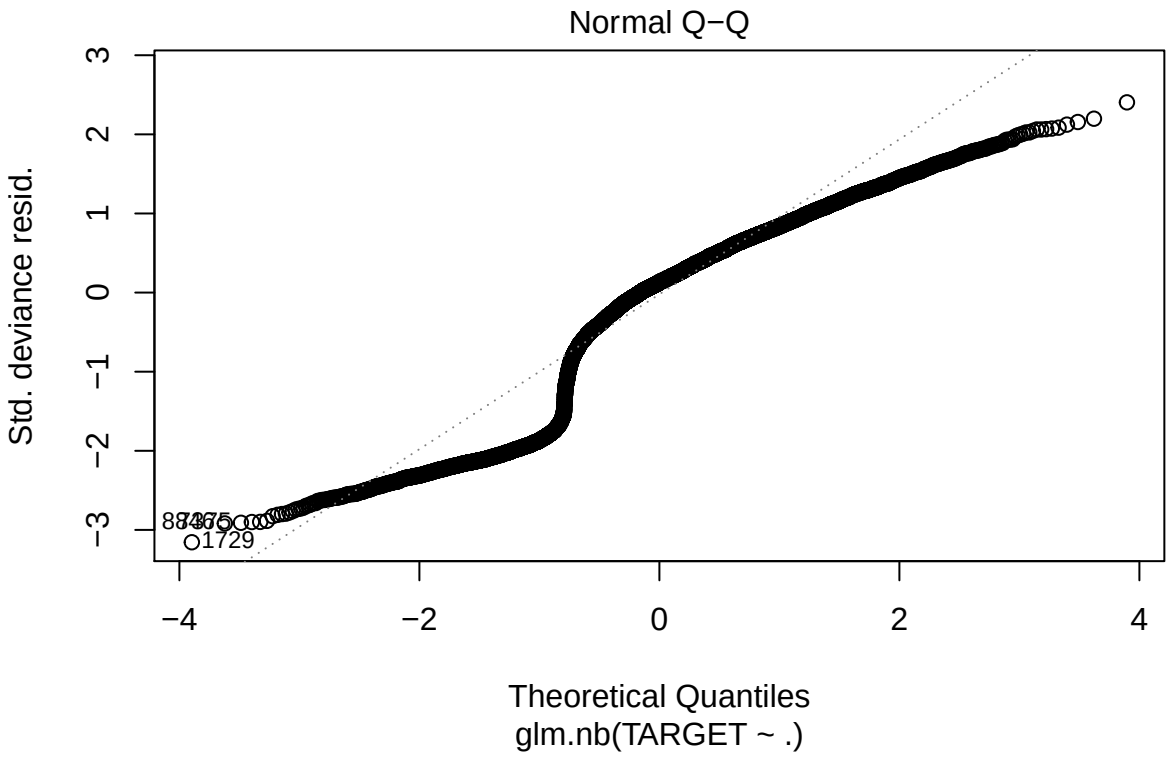
Summary of the negative binomial model with imputation.

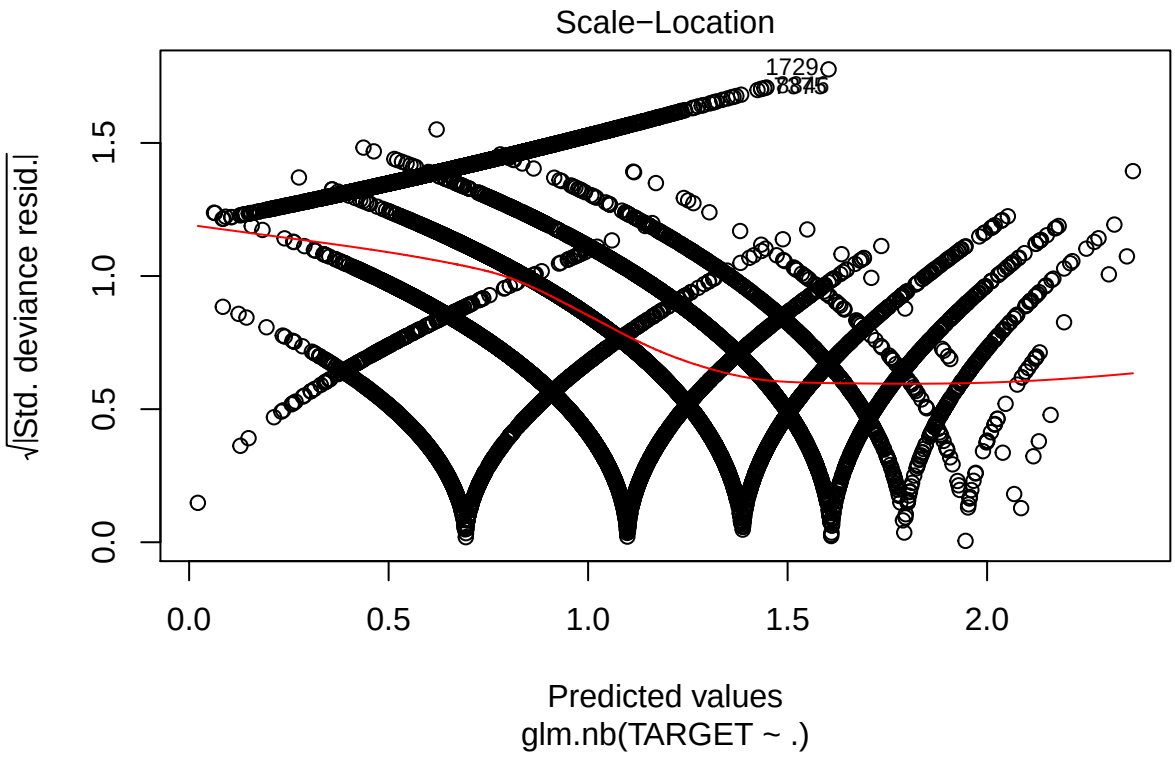
```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = wine_train2, init.theta = 48897.24324,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1515  -0.6808   0.1304   0.6390   2.4032
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.382e+00  2.277e-01  10.463  < 2e-16 ***
## FixedAcidity       -1.332e-04  9.197e-04  -0.145 0.884879
## VolatileAcidity    -4.351e-02  7.275e-03  -5.981 2.21e-09 ***
```

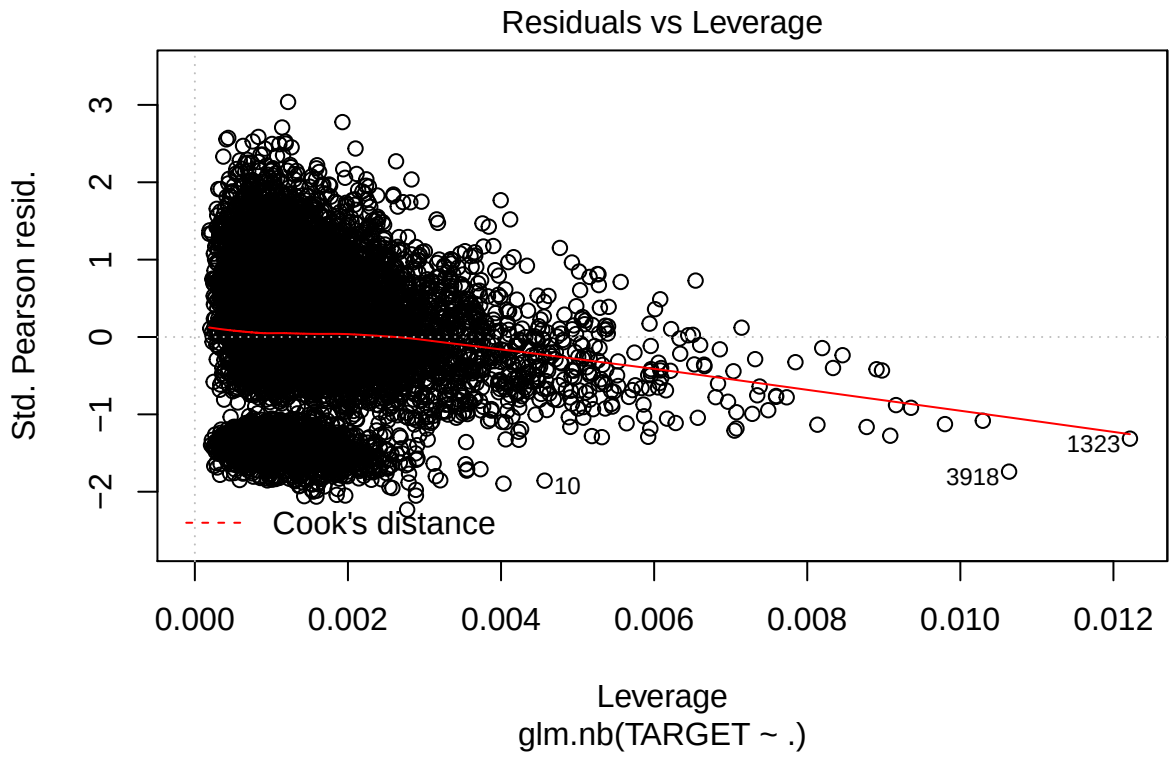
```
## CitricAcid          8.883e-03  6.577e-03   1.351 0.176804
## ResidualSugar       1.508e-04  1.675e-04   0.900 0.367960
## Chlorides          -6.506e-02  1.791e-02  -3.633 0.000280 ***
## FreeSulfurDioxide   1.143e-04  3.804e-05   3.005 0.002657 **
## TotalSulfurDioxide  8.709e-05  2.446e-05   3.560 0.000371 ***
## Density            -4.047e-01  2.141e-01  -1.890 0.058762 .
## pH                 -1.788e-02  8.407e-03  -2.126 0.033466 *
## Sulphates          -1.327e-02  6.164e-03  -2.153 0.031286 *
## Alcohol             2.690e-03  1.546e-03   1.740 0.081887 .
## LabelAppeal         1.432e-01  6.783e-03  21.106  < 2e-16 ***
## AcidIndex          -7.622e-01  4.005e-02 -19.029  < 2e-16 ***
## STARS               3.401e-01  6.252e-03  54.393  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48897.24) family taken to be 1)
##
##     Null deviance: 18290  on 10236  degrees of freedom
## Residual deviance: 12830  on 10222  degrees of freedom
## AIC: 38420
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48897
##           Std. Err.:  63448
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -38388.3
```

Here, I have plotted the Negative Binomial model with imputation by Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage as below:


Residuals vs Fitted
Residuals
3
2
1
0
−1
−2
−3
0.0
0.5
1.0
1.5
2.0
8345
1729
Predicted values
glm.nb(TARGET ~ .)


Normal Q−Q
Std. deviance resid.
3
2
1
0
−1
−2
−3
1729
−4
−2
0
2
4
Theoretical Quantiles
glm.nb(TARGET ~ .)


Scale−Location
1729
8345
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
Predicted values
glm.nb(TARGET ~ .)



8. Negative Binomialmodel with imputations and only with significant variables: Now, I will again build a negative binomial model with imputations and only significant variables here.

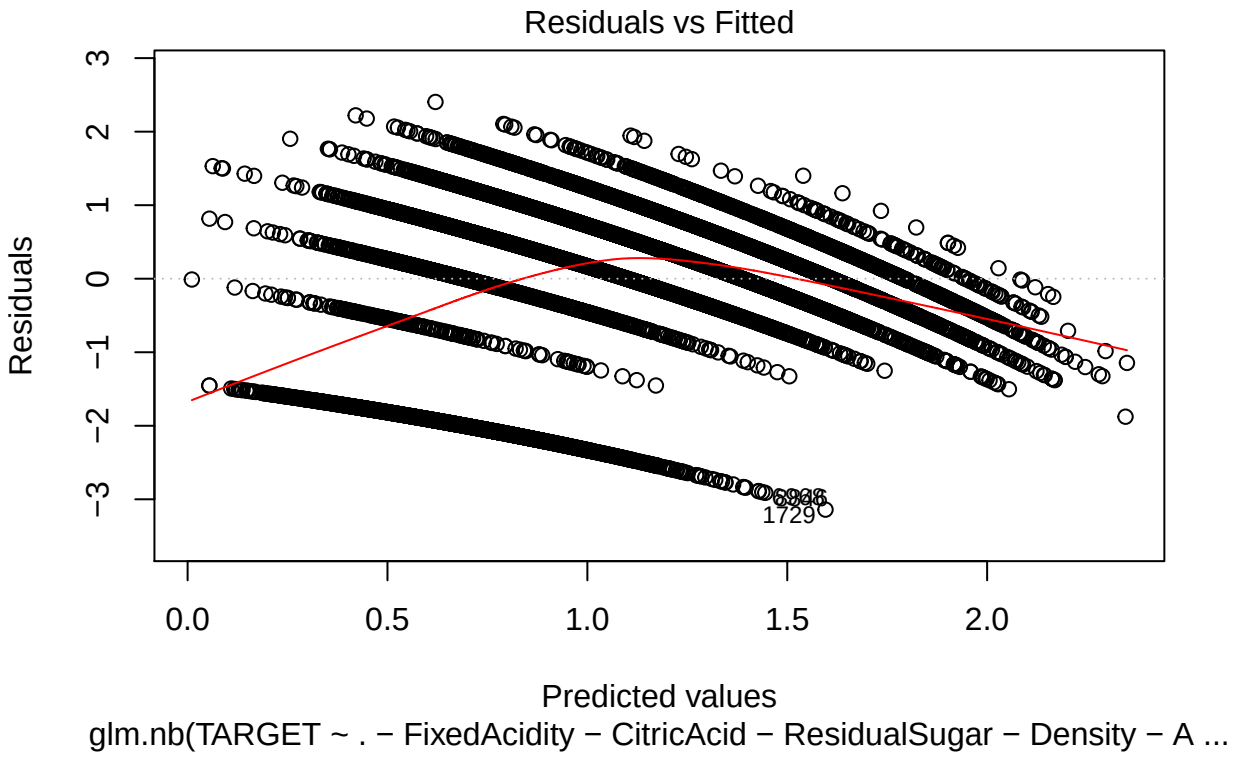
```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace > : iteration
## limit reached

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace > : iteration
## limit reached
```

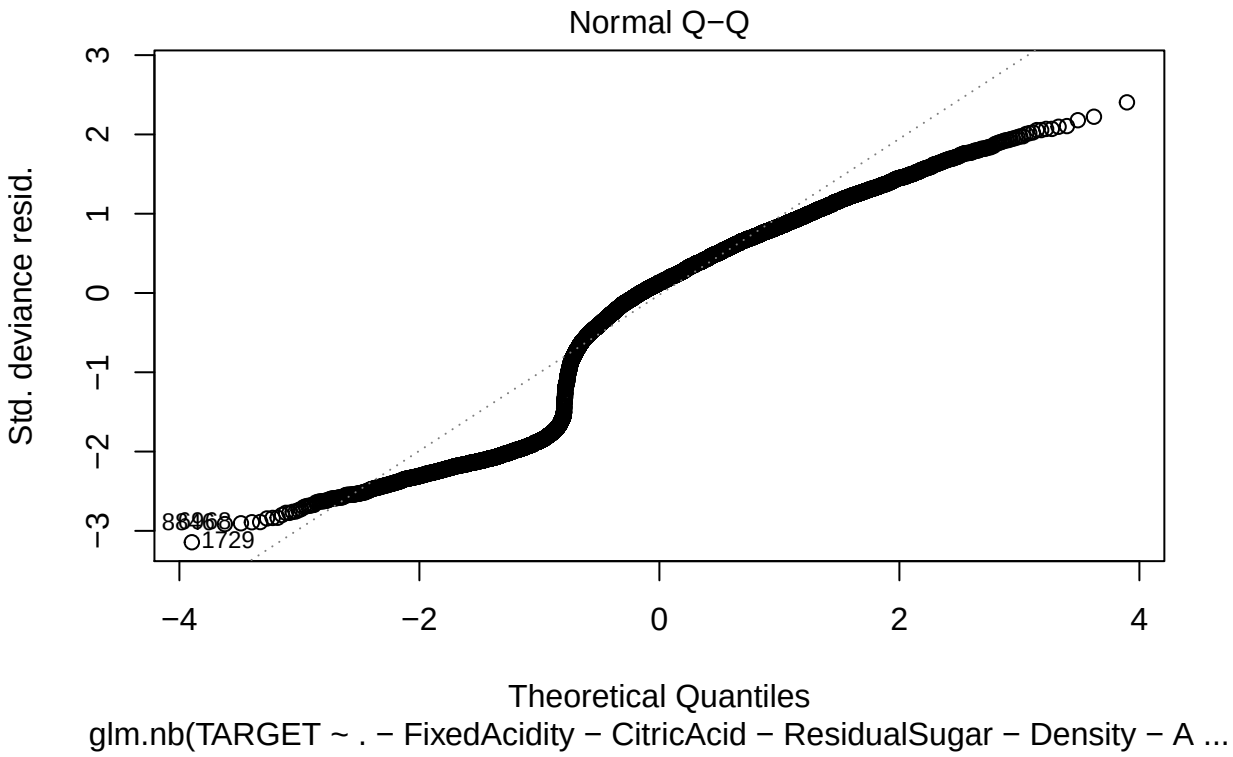
The summary of the negative binomial model with imputations and only with significat variables are as follows:

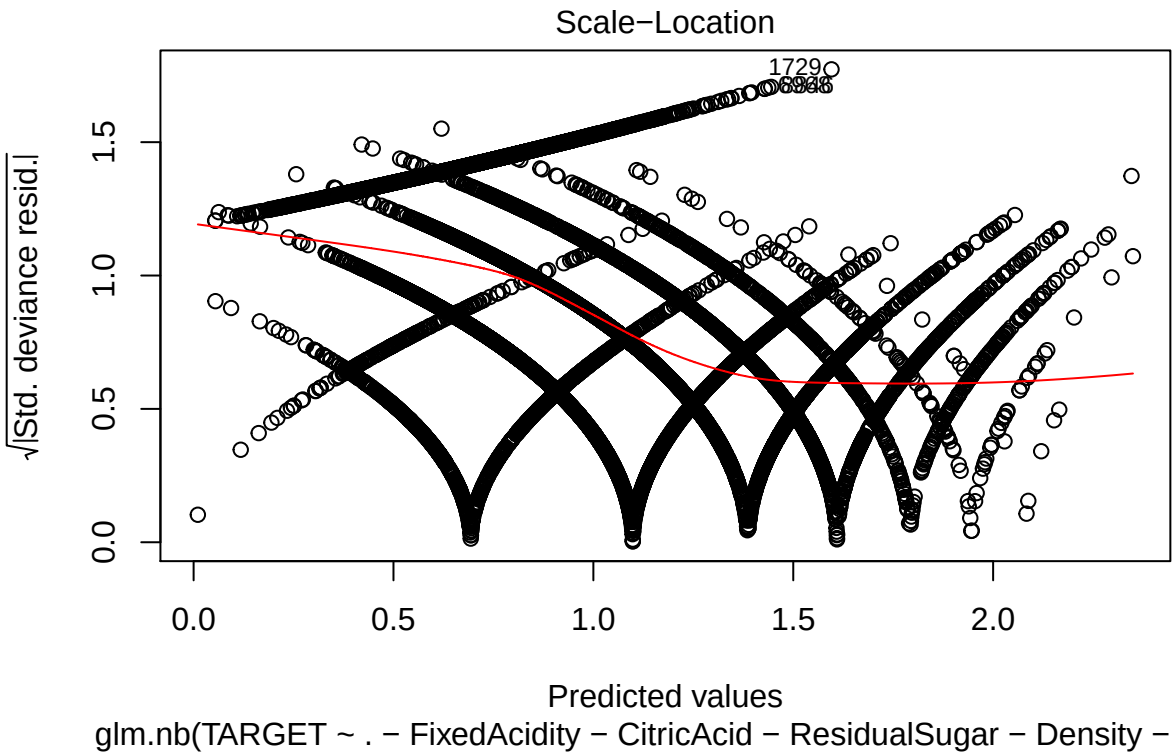
```
##
## Call:
## glm.nb(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##     Density - Alcohol, data = wine_train2, init.theta = 48805.90033,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1405  -0.6852   0.1288   0.6412   2.4038
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.019e+00  8.849e-02  22.820  < 2e-16 ***
```

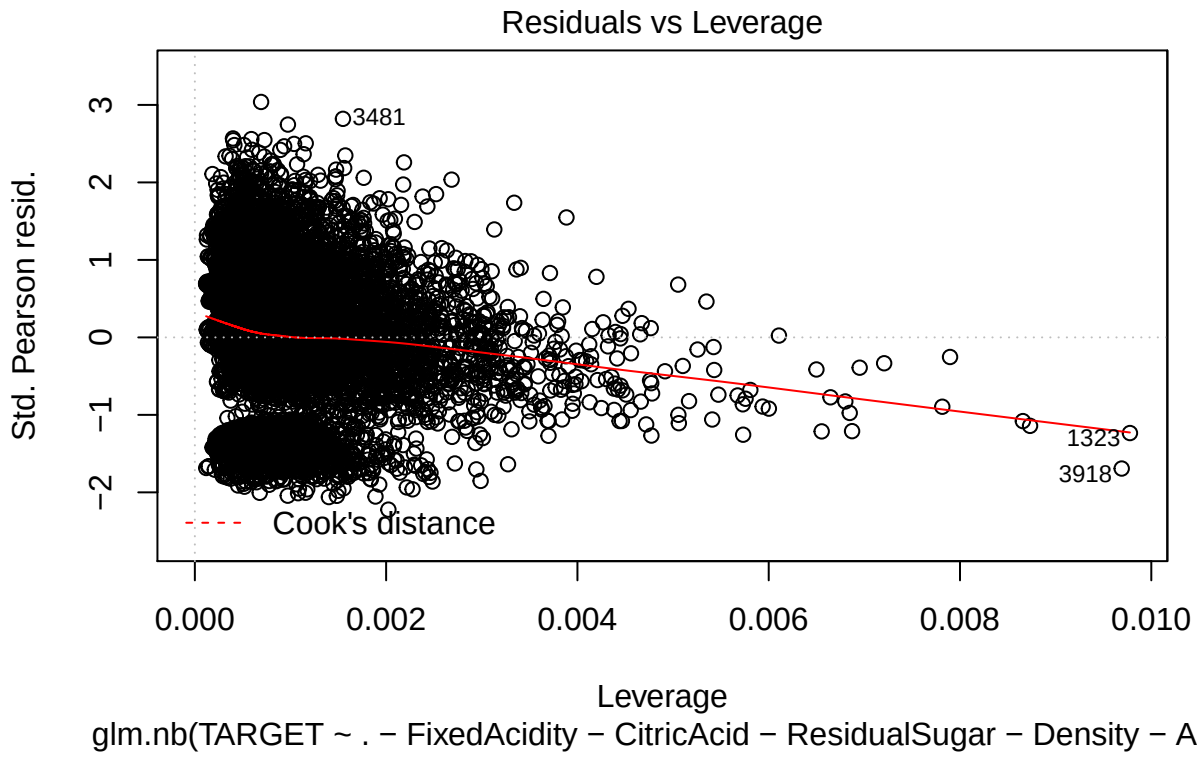
```
## VolatileAcidity    -4.388e-02  7.273e-03  -6.033 1.61e-09 ***
## Chlorides          -6.711e-02  1.790e-02  -3.750 0.000177 ***
## FreeSulfurDioxide   1.119e-04  3.802e-05   2.942 0.003257 **
## TotalSulfurDioxide  8.561e-05  2.443e-05   3.505 0.000457 ***
## pH                 -1.818e-02  8.404e-03  -2.164 0.030489 *
## Sulphates          -1.327e-02  6.157e-03  -2.155 0.031144 *
## LabelAppeal         1.433e-01  6.783e-03  21.119  < 2e-16 ***
## AcidIndex          -7.665e-01  3.941e-02 -19.447  < 2e-16 ***
## STARS               3.410e-01  6.237e-03  54.671  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48805.9) family taken to be 1)
##
##     Null deviance: 18290  on 10236  degrees of freedom
## Residual deviance: 12839  on 10227  degrees of freedom
## AIC: 38420
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48806
##           Std. Err.:  63368
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -38397.65
```

Here, I have plotted the Negative Binomial model with imputation and only with significant variables by Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage as below:


Residuals vs Fitted
Residuals
Predicted values
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − A ...
1729
3
2
1
0
−1
−2
−3
0.0
0.5
1.0
1.5
2.0


Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − A ...
1729


Scale−Location
1729
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
Predicted values
glm.nb(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar − Density − A ...

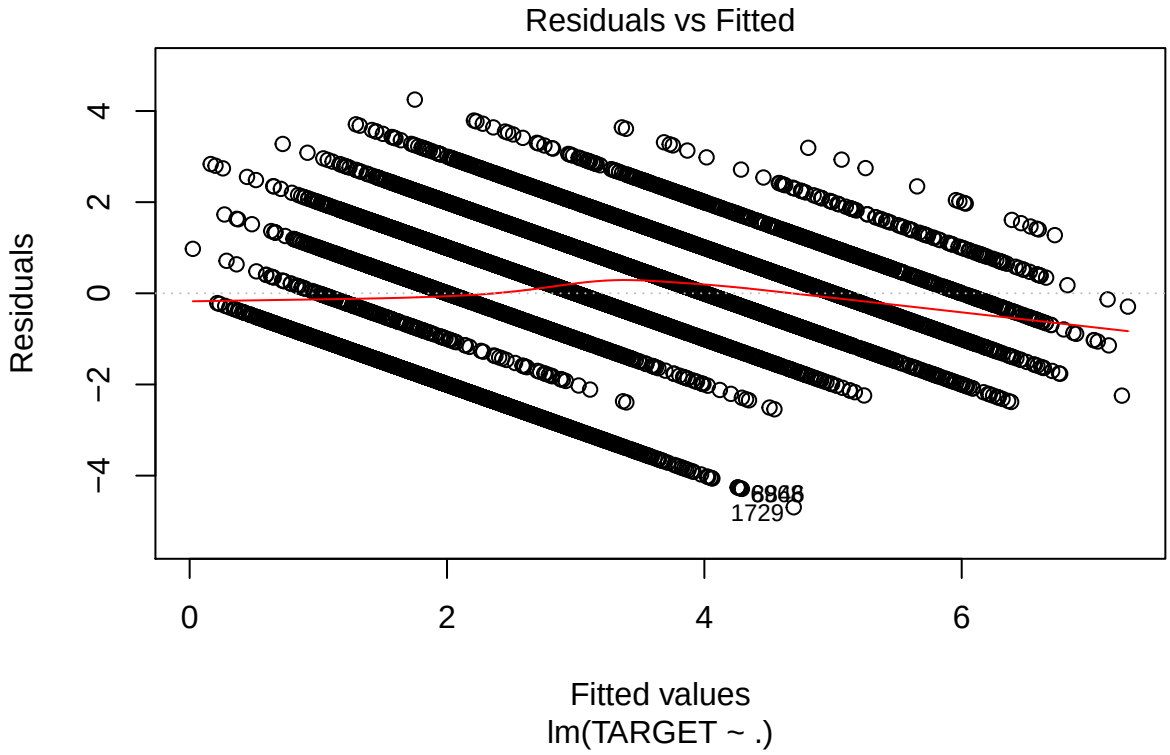


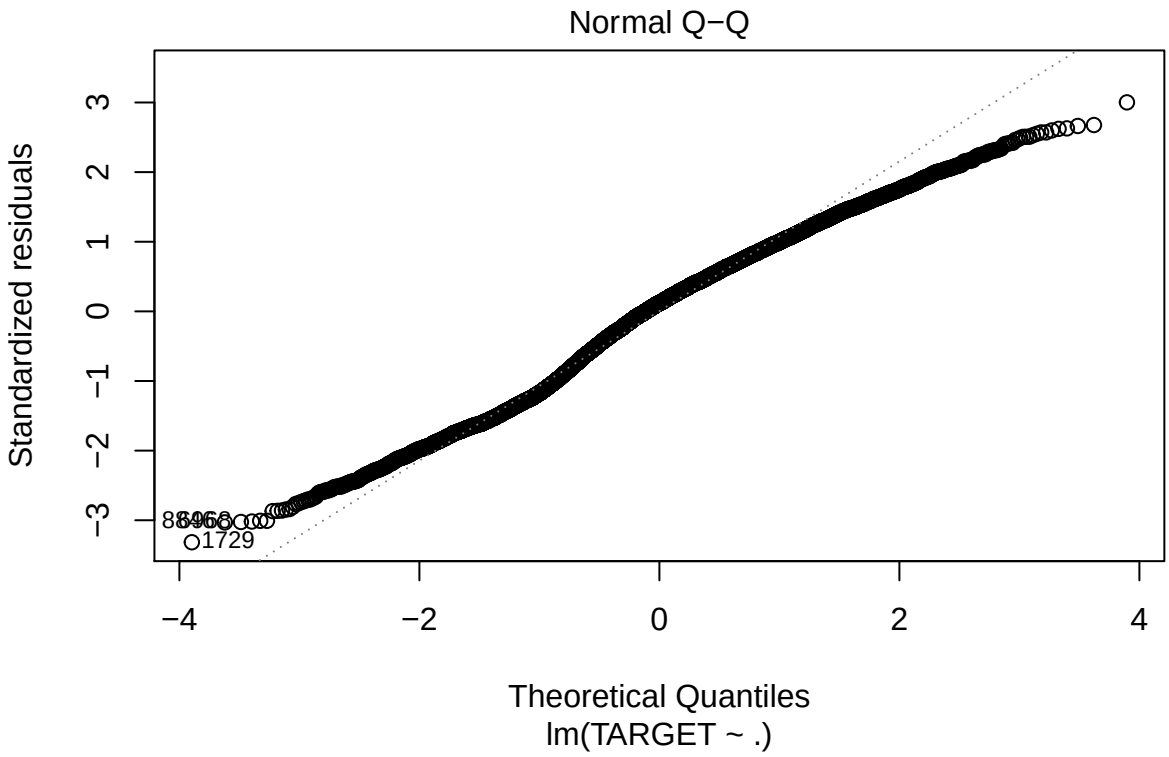
9. Linear Model with imputations: Now, I am building a linear model with imputations and also the summary of the model as follows.

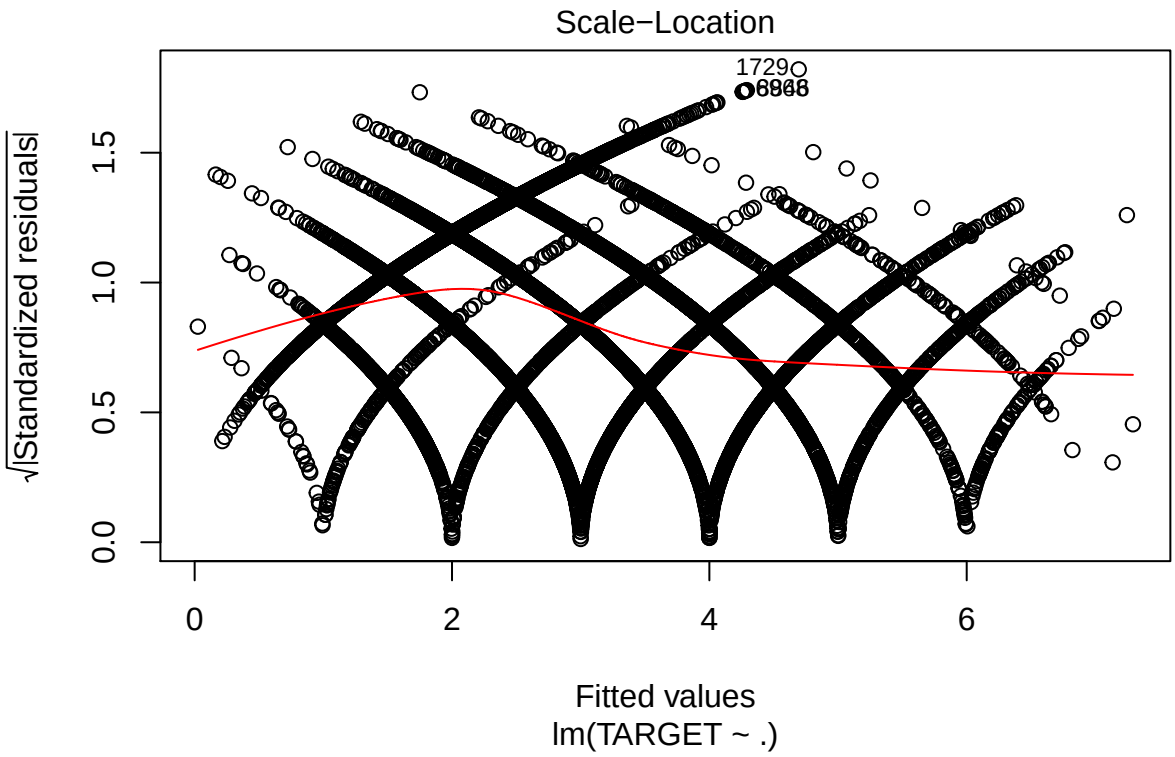
```
##
## Call:
## lm(formula = TARGET ~ ., data = wine_train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6944 -1.0191  0.1692  1.0335  4.2502
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.149e+00  5.564e-01  11.052  < 2e-16 ***
## FixedAcidity        -1.428e-04  2.255e-03  -0.063  0.94952
## VolatileAcidity     -1.265e-01  1.792e-02  -7.056 1.82e-12 ***
## CitricAcid           2.771e-02  1.630e-02   1.699  0.08927 .
## ResidualSugar        4.479e-04  4.138e-04   1.083  0.27904
## Chlorides           -1.956e-01  4.398e-02  -4.448 8.77e-06 ***
## FreeSulfurDioxide    2.930e-04  9.398e-05   3.117  0.00183 **
## TotalSulfurDioxide   2.365e-04  6.006e-05   3.938 8.28e-05 ***
## Density             -1.099e+00  5.263e-01  -2.088  0.03678 *
## pH                  -4.064e-02  2.071e-02  -1.962  0.04978 *
## Sulphates           -3.621e-02  1.519e-02  -2.384  0.01713 *
## Alcohol              1.131e-02  3.782e-03   2.991  0.00279 **
```

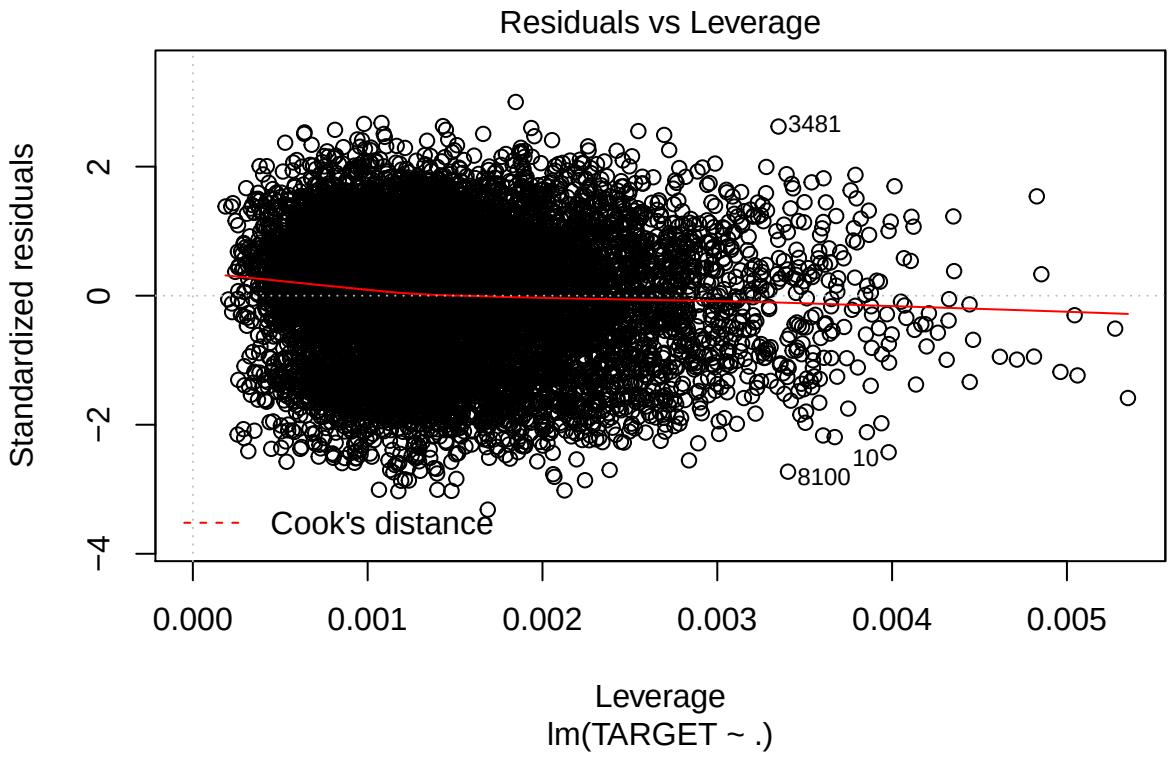
```
## LabelAppeal          4.379e-01  1.644e-02  26.633  < 2e-16 ***
## AcidIndex           -2.041e+00  9.250e-02 -22.067  < 2e-16 ***
## STARS                1.162e+00  1.665e-02  69.754  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.417 on 10222 degrees of freedom
## Multiple R-squared:  0.4598, Adjusted R-squared:  0.4591
## F-statistic: 621.5 on 14 and 10222 DF,  p-value: < 2.2e-16
```

Here, I have plotted the linear model with imputations by Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage as below:




Normal Q−Q
Standardized residuals
−3
−2
−1
0
1
2
3
1729
−4
−2
0
2
4
Theoretical Quantiles
lm(TARGET ~ .)


Scale−Location
1729
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
2
4
6
Fitted values
lm(TARGET ~ .)



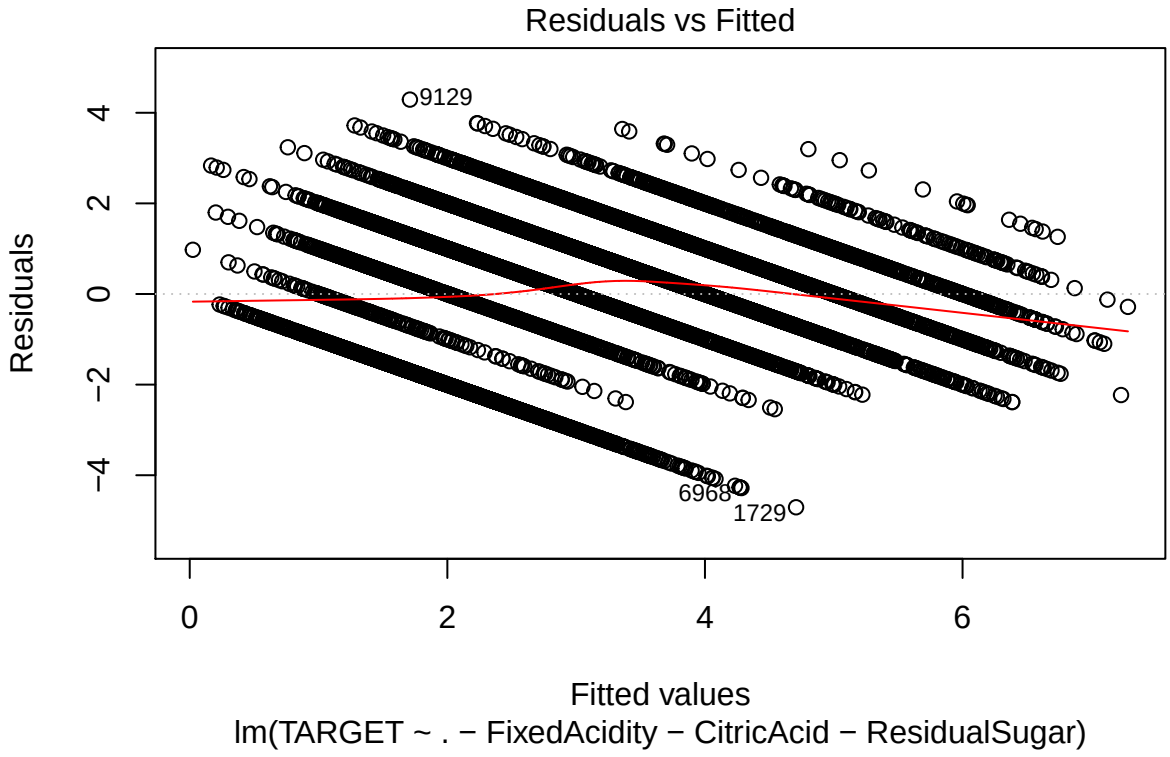
10. Linear Model with imputations and only significant variables: Now, I am building a Linear Model with imputations and only significant variables.

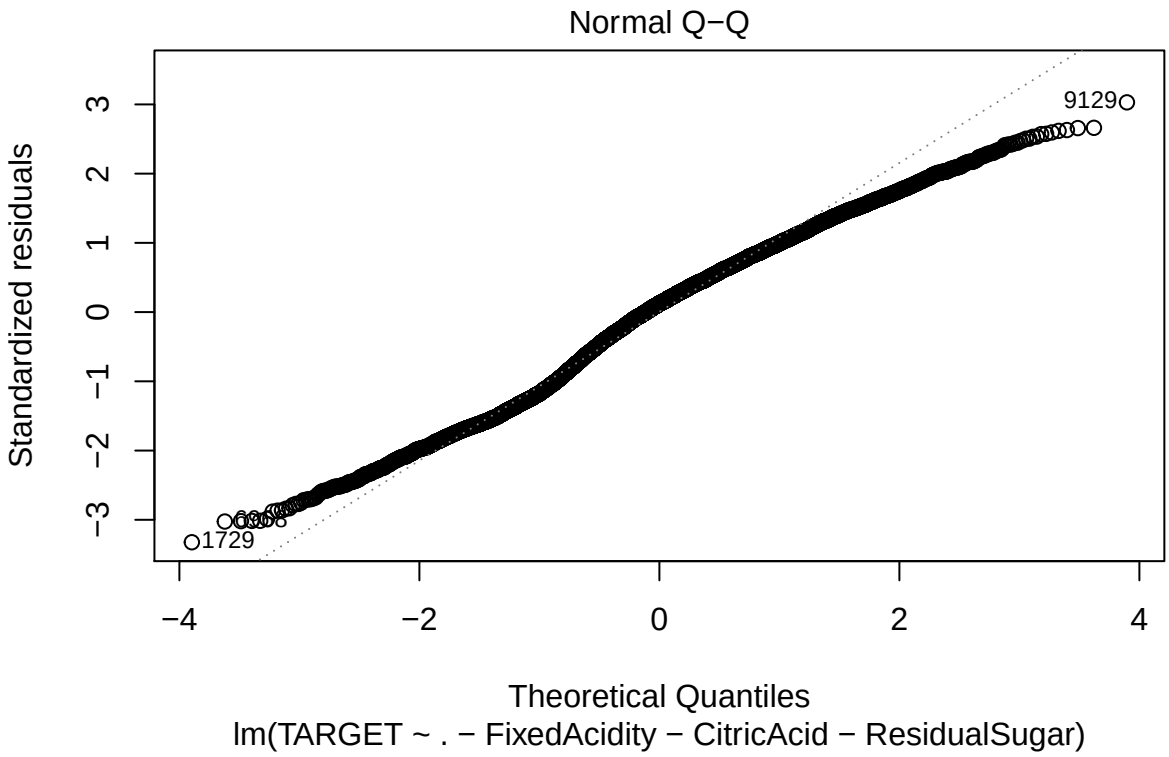
The sumary of the Linear Model with imputations and only significant variables are as follows:

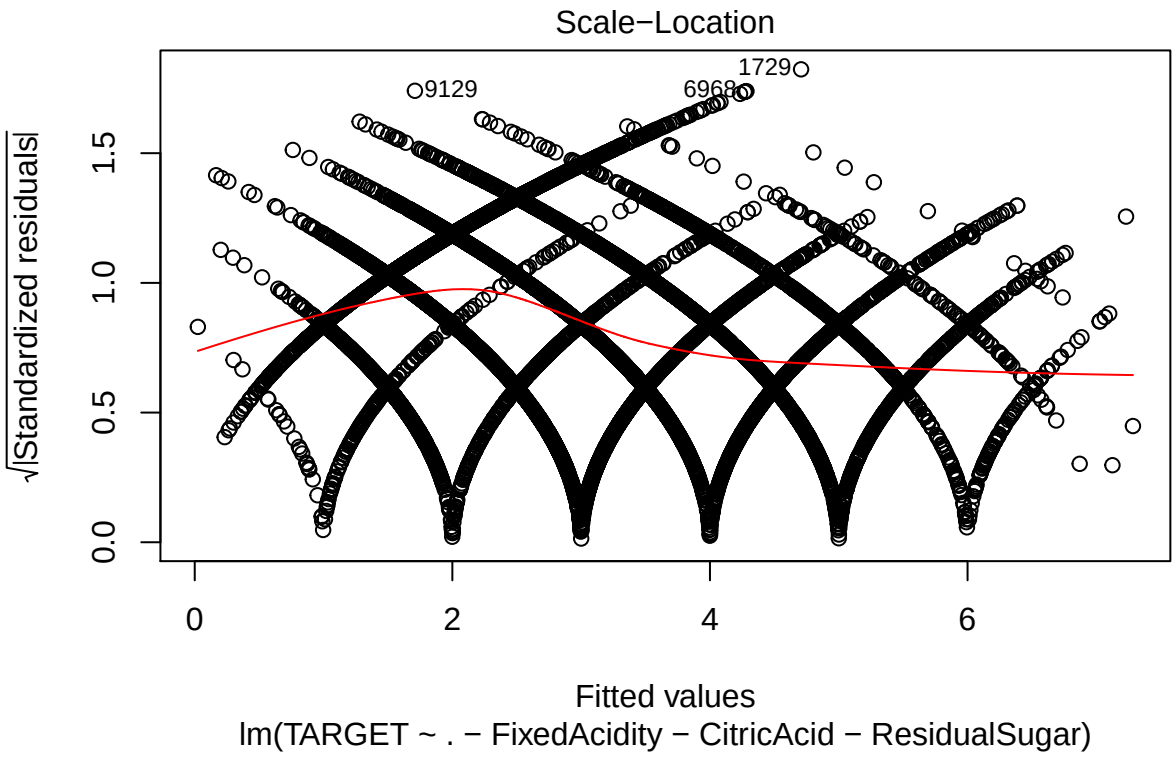
```
##
## Call:
## lm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar,
##     data = wine_train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7075 -1.0195  0.1718  1.0343  4.2907
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.139e+00  5.563e-01  11.036  < 2e-16 ***
## VolatileAcidity    -1.273e-01  1.792e-02  -7.104 1.30e-12 ***
## Chlorides          -1.970e-01  4.397e-02  -4.479 7.57e-06 ***
## FreeSulfurDioxide   2.939e-04  9.397e-05   3.128  0.00177 **
## TotalSulfurDioxide  2.389e-04  6.003e-05   3.980 6.94e-05 ***
## Density            -1.101e+00  5.263e-01  -2.093  0.03638 *
## pH                 -4.059e-02  2.071e-02  -1.960  0.05008 .
## Sulphates          -3.699e-02  1.517e-02  -2.437  0.01481 *
## Alcohol             1.136e-02  3.781e-03   3.005  0.00266 **
```

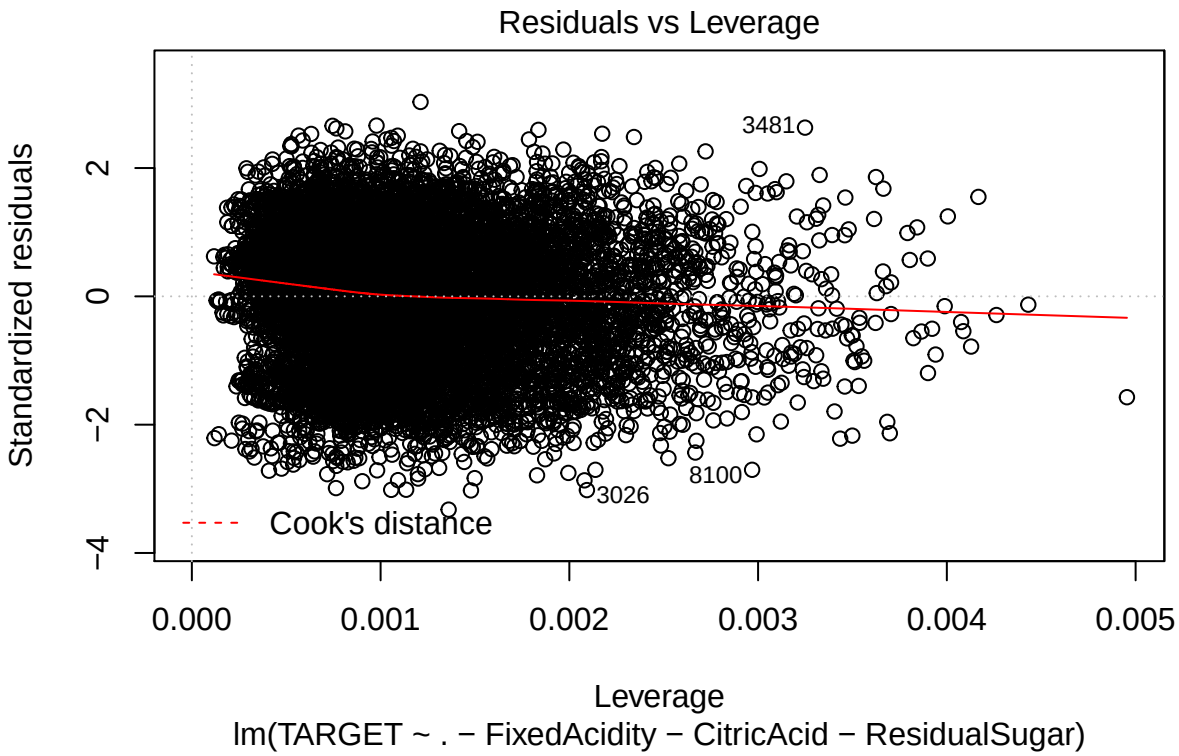
```
## LabelAppeal          4.379e-01  1.644e-02  26.632  < 2e-16 ***
## AcidIndex           -2.031e+00  9.085e-02 -22.351  < 2e-16 ***
## STARS                1.162e+00  1.665e-02  69.794  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.417 on 10225 degrees of freedom
## Multiple R-squared:  0.4596, Adjusted R-squared:  0.459
## F-statistic: 790.6 on 11 and 10225 DF,  p-value: < 2.2e-16
```

Here, I have plotted the Linear Model with imputations and only significant variables by Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage as below:




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar)
9129
1729


Scale−Location
9129
6968
1729
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
2
4
6
Fitted values
lm(TARGET ~ . − FixedAcidity − CitricAcid − ResidualSugar)



Output of the Model using different test: Now, I will see the output of the Models using test data.

I will use the squared loss to validate the model.

Poisson model with imputations:

```
## [1] 6.852209
```

Poisson model with imputations and only significant variables:

```
## [1] 6.854547
```

Negative Binomial with imputations:

```
## [1] 6.852205
```

Negative Binomial with imputations and only significant variables:

```
## [1] 6.854543
```

Linear Model with imputations:

```
## [1] 2.029061
```

Linear Model with imputations and only significant variables:

```
## [1] 2.030002
```

MODEL SELECTION: From the above different test, I find following: 1.Poisson model with imputations has modelValidation value 6.852 2.Poisson model with imputations and only with significant variable has the modelvalidation value 6.854 3.Negative Binomial with imputations has the modelvalidation value 6.852 4.Negative Binomial with imputations and only with significant variable has the modelvalidation value 6.854 5.Linear model with imputations has the modelvalidation value 2.02 6. Linear model with imputations and only with significant variable has the modelvalidation value 2.03

From the above models, I would like to go with Linear model with imputations and only with significant variable model based on the test result. Since, Linear Model with imputations and only significant variables uses less variables and is parsimonious. Also the R2 looks fine. The squared root is also fine.

Prediction: We will use the same method to impute and use log transformation for AcidIndex.

```
##        IN          TARGET         FixedAcidity     VolatileAcidity     CitricAcid
##  Min.   :    3   Mode:logical   Min.   :-18.200   Min.   :-2.8300   Min.   :-3.1200
##  1st Qu.: 4018   NA's:3335      1st Qu.:  5.200   1st Qu.: 0.0800   1st Qu.: 0.0000
##  Median : 7906                  Median :  6.900   Median : 0.2800   Median : 0.3100
##  Mean   : 8048                  Mean   :  6.864   Mean   : 0.3103   Mean   : 0.3124
##  3rd Qu.:12061                  3rd Qu.:  9.000   3rd Qu.: 0.6300   3rd Qu.: 0.6050
##  Max.   :16130                  Max.   : 33.500   Max.   : 3.6100   Max.   : 3.7600
##
##  ResidualSugar        Chlorides        FreeSulfurDioxide TotalSulfurDioxide    Density
##  Min.   :-128.300   Min.   :-1.15000   Min.   :-563.00   Min.   :-769.00    Min.   :0.8898
##  1st Qu.:  -2.600   1st Qu.: 0.01600   1st Qu.:   3.00   1st Qu.:  27.25    1st Qu.:0.9883
##  Median :   3.600   Median : 0.04700   Median :  30.00   Median : 124.00    Median :0.9946
##  Mean   :   5.319   Mean   : 0.06143   Mean   :  34.95   Mean   : 123.41    Mean   :0.9947
##  3rd Qu.:  17.200   3rd Qu.: 0.17100   3rd Qu.:  79.25   3rd Qu.: 210.00    3rd Qu.:1.0005
##  Max.   : 145.400   Max.   : 1.26300   Max.   : 617.00   Max.   :1004.00    Max.   :1.0998
##  NA's   :168        NA's   :138        NA's   :152       NA's   :157
##        pH          Sulphates          Alcohol       LabelAppeal         AcidIndex
##  Min.   :0.600   Min.   :-3.0700   Min.   :-4.20   Min.   :-2.00000   Min.   : 5.000
##  1st Qu.:2.980   1st Qu.: 0.3300   1st Qu.: 9.00   1st Qu.:-1.00000   1st Qu.: 7.000
##  Median :3.210   Median : 0.5000   Median :10.40   Median : 0.00000   Median : 8.000
##  Mean   :3.237   Mean   : 0.5346   Mean   :10.58   Mean   : 0.01349   Mean   : 7.748
##  3rd Qu.:3.490   3rd Qu.: 0.8200   3rd Qu.:12.50   3rd Qu.: 1.00000   3rd Qu.: 8.000
##  Max.   :6.210   Max.   : 4.1800   Max.   :25.60   Max.   : 2.00000   Max.   :17.000
##  NA's   :104     NA's   :310       NA's   :185
##      STARS
##  Min.   :1.00
##  1st Qu.:1.00
##  Median :2.00
##  Mean   :2.04
##  3rd Qu.:3.00
##  Max.   :4.00
##  NA's   :841
```

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
```

```
## Warning: Number of logged events: 1
```

Conclusion: By following the procedure of Data exploration, data preparation, building model and selecting model, I have come to a conlcusion that Linear model with imputations and only with significant variable model is the best model based on our wine traing data set. Since, this model has less variables and is parsimonious and also Modelvalidation value also lower. The R value and the square root value of the model looks great.