
Critical Thinking Group 4: DATA621 Homework 5

Table of Contents

TEAM Members:	2
Overview	2
Deliverables	3
Data Exploration	3
Data Preparation	Error! Bookmark not defined.
Build Models	Error! Bookmark not defined.
Select Models	Error! Bookmark not defined.
Make Predictions	59
Appendix	59

TEAM Members:

Rajwant Mishra

Priya Shaji

Debabrata Kabiraj

Isabel Ramesar

Sin Ying Wong

Fan Xu

Overview

In this homework assignment, you 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.

Your objective 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. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You can only use the variables given to you (or variables that you derive from the variables provided).

Below is a short description of the variables of interest in the data set:

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET	Number of Cases Purchased	None
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average	
Alcohol	Alcohol Content	
Chlorides	Chloride content of wine	
CitricAcid	Citric Acid Content	
Density	Density of Wine	
FixedAcidity	Fixed Acidity of Wine	
FreeSulfurDioxide	Sulfur Dioxide content of wine	
LabelAppeal	Marketing Score indicating 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.
ResidualSugar	Residual Sugar of wine	
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor	A high number of stars suggests high sales
Sulphates	Sulfate content of wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	

Deliverables

A write-up of your solutions submitted in PDF format.

Assigned predictions (number of cases of wine sold) for the evaluation data set.

Data Load

We have two datasets.

- One is the `wine training dataset`, which includes 14 candidate predictors, 1 response variable and 12795 observations.
- Other one is the `wine evaluation dataset`, which also includes 14 candidate predictors, 1 response variable but 16129 observations.

We are going to study their missing values, data types and data statistics.

Training Dataset

	TARGET	INDEX	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide
1	3	1	3.2	1.16	-0.98	54.2	-0.567	
2	3	2	4.5	0.16	-0.81	26.1	-0.425	
3	5	4	7.1	2.64	-0.88	14.8	0.037	
4	3	5	5.7	0.385	0.04	18.8	-0.425	
5	4	6	8	0.33	-1.26	9.4		
6	0	7	11.3	0.32	0.59	2.2	0.556	
7	0	8	7.7	0.29	-0.4	21.5	0.06	
8	4	11	6.5	-1.22	0.34	1.4	0.04	
9	3	12	14.8	0.27	1.05	11.25	-0.007	
10	6	13	5.5	-0.22	0.39	1.8	-0.277	

Showing 1 to 10 of 12,795 entries

```
## Observations: 12,795
## Variables: 16
## $ TARGET          <dbl> 3, 3, 5, 3, 4, 0, 0, 4, 3, 6, 0, 4, 3, 7, 4, 0, ...
## $ INDEX           <dbl> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, ...
## $ FixedAcidity    <dbl> 3.2, 4.5, 7.1, 5.7, 8.0, 11.3, 7.7, 6.5, 14.8, 5...
## $ VolatileAcidity <dbl> 1.160, 0.160, 2.640, 0.385, 0.330, 0.320, 0.290, ...
## $ CitricAcid      <dbl> -0.98, -0.81, -0.88, 0.04, -1.26, 0.59, -0.40, 0...
## $ ResidualSugar   <dbl> 54.20, 26.10, 14.80, 18.80, 9.40, 2.20, 21.50, 1...
## $ Chlorides        <dbl> -0.567, -0.425, 0.037, -0.425, NA, 0.556, 0.060, ...
## $ FreeSulfurDioxide <dbl> NA, 15, 214, 22, -167, -37, 287, 523, -213, 62, ...
## $ TotalSulfurDioxide <dbl> 268, -327, 142, 115, 108, 15, 156, 551, NA, 180, ...
## $ Density          <dbl> 0.99280, 1.02792, 0.99518, 0.99640, 0.99457, 0.9...
## $ pH               <dbl> 3.33, 3.38, 3.12, 2.24, 3.12, 3.20, 3.49, 3.20, ...
## $ Sulphates        <dbl> -0.59, 0.70, 0.48, 1.83, 1.77, 1.29, 1.21, NA, 0...
## $ Alcohol          <dbl> 9.9, NA, 22.0, 6.2, 13.7, 15.4, 10.3, 11.6, 15.0...
## $ LabelAppeal      <dbl> 0, -1, -1, -1, 0, 0, 1, 0, 0, 1, 0, 1, 2, 0, ...
```

```
## $ AcidIndex      <dbl> 8, 7, 8, 6, 9, 11, 8, 7, 6, 8, 5, 10, 7, 8, 9, 8...
## $ STARS        <dbl> 2, 3, 3, 1, 2, NA, NA, 3, NA, 4, 1, 2, 2, 3, NA,...
```

Evaluation Dataset

	IN	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide
1	3		5.4	-0.86	0.27	-10.7	0.092	
2	9		12.4	0.385	-0.76	-19.7	1.169	
3	10		7.2	1.75	0.17	-33	0.065	
4	18		6.2	0.1	1.8	1	-0.179	
5	21		11.4	0.21	0.28	1.2	0.038	
6	30		17.6	0.04	-1.15	1.4	0.535	
7	31		15.5	0.53	-0.53	4.6	1.263	
8	37		15.9	1.19	1.14	31.9	-0.299	
9	39		11.6	0.32	0.55	-50.9	0.076	
10	47		3.8	0.22	0.31	-7.7	0.039	

Showing 1 to 10 of 3,335 entries

```
## Observations: 3,335
## Variables: 16
## $ IN           <dbl> 3, 9, 10, 18, 21, 30, 31, 37, 39, 47, 60, 62, 63...
## $ TARGET        <lgl> NA, ...
## $ FixedAcidity <dbl> 5.4, 12.4, 7.2, 6.2, 11.4, 17.6, 15.5, 15.9, 11....
## $ VolatileAcidity <dbl> -0.860, 0.385, 1.750, 0.100, 0.210, 0.040, 0.530...
## $ CitricAcid    <dbl> 0.27, -0.76, 0.17, 1.80, 0.28, -1.15, -0.53, 1.1...
## $ ResidualSugar <dbl> -10.70, -19.70, -33.00, 1.00, 1.20, 1.40, 4.60, ...
## $ Chlorides     <dbl> 0.092, 1.169, 0.065, -0.179, 0.038, 0.535, 1.263...
## $ FreeSulfurDioxide <dbl> 23, -37, 9, 104, 70, -250, 10, 115, 35, 40, NA, ...
## $ TotalSulfurDioxide <dbl> 398, 68, 76, 89, 53, 140, 17, 381, 83, 129, 583, ...
## $ Density       <dbl> 0.98527, 0.99048, 1.04641, 0.98877, 1.02899, 0.9...
```

```

## $ pH <dbl> 5.02, 3.37, 4.61, 3.20, 2.54, 3.06, 3.07, 2.99, ...
## $ Sulphates <dbl> 0.64, 1.09, 0.68, 2.11, -0.07, -0.02, 0.75, 0.31...
## $ Alcohol <dbl> 12.30, 16.00, 8.55, 12.30, 4.80, 11.40, 8.50, 11...
## $ LabelAppeal <dbl> -1, 0, 0, -1, 0, 1, 0, 0, 0, 1, 0, 0, -1, ...
## $ AcidIndex <dbl> 6, 6, 8, 8, 10, 8, 12, 7, 12, 7, 8, 10, 9, 8, 9, ...
## $ STARS <dbl> NA, 2, 1, 1, NA, 4, 3, NA, NA, NA, 1, NA, 2, NA, ...

```

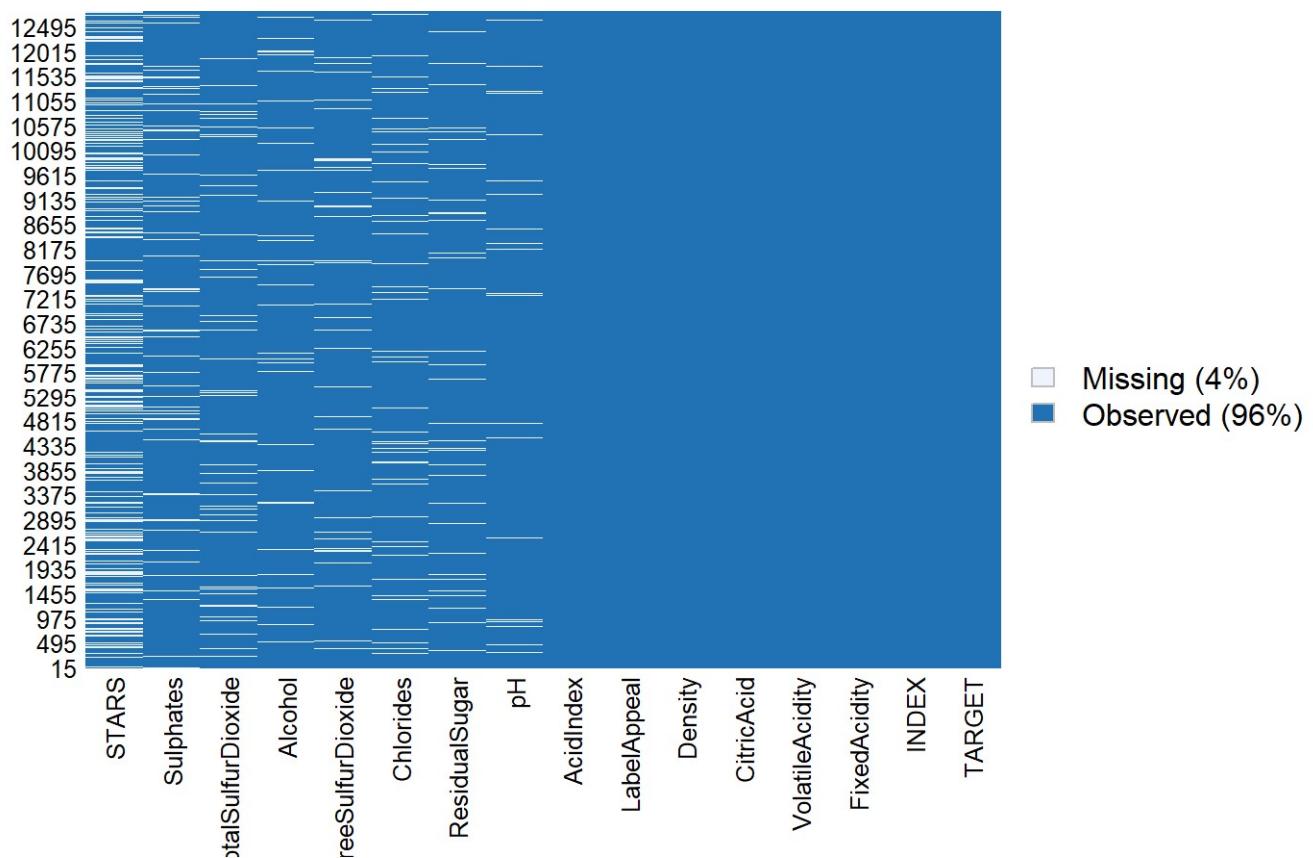
Missing Values & Data Type Check

In the [wine training dataset](#), there are 14 candidate predictors and 1 response variable with 12,795 observations. In the [wine evaluation dataset](#), there are 14 candidate predictors with 3,335 observations. Both datasets have no missing values (eg: NA, NULL or '').

Among the 12 candidate predictors, 3 are categorical ([LabelAppeal](#), [AcidIndex](#), [STARS](#)), the other 11 are continuous numerical. The response variable [TARGET](#) is categorical.

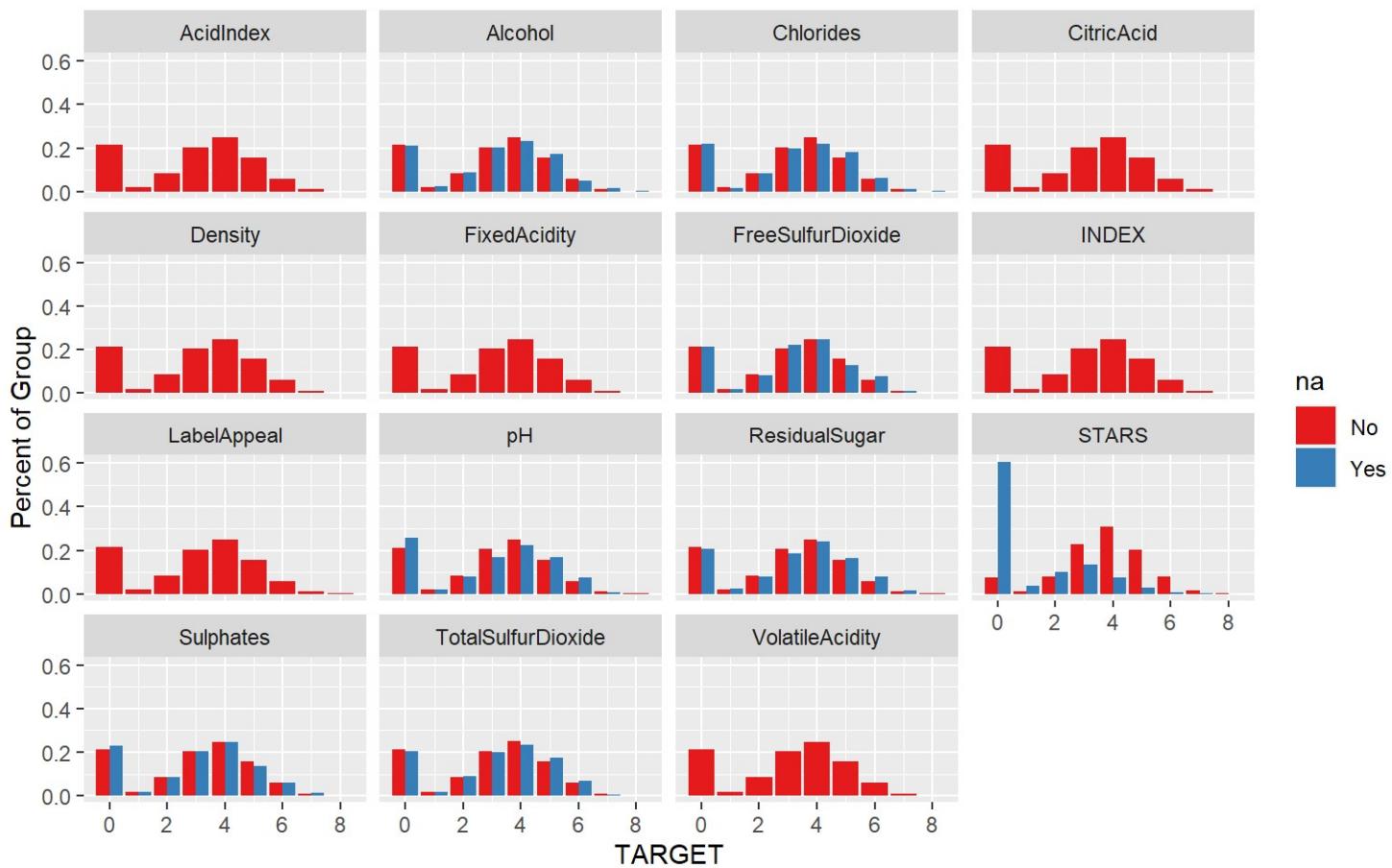


Missing vs Observed Values in Training Data



	Non_NAs	NAs	NA_Percent
TARGET	12795	0	0.0000000
INDEX	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

	Non_NAs	NAs	NA_Percent
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



Below is the summary of the datasets and some inference of it.

1. It seems there are Null values in the predictor variables but none in response variables.
2. Each variables are in different scale.

Data Statistics Summary

A binary logistic regression model is built using the `training set`, therefore the `training set` is used for the following data exploration.

The data types in the raw dataset are all 'doubles', however the counter `INDEX` and the response variable `target` are categorical.

The statistics of all variables are listed below:

```
## TARGET      FixedAcidity    VolatileAcidity   CitricAcid
## Min. :0.000  Min. :-18.100  Min. :-2.7900  Min. :-3.2400
## 1st Qu.:2.000 1st Qu.: 5.200  1st Qu.: 0.1300  1st Qu.: 0.0300
## Median :3.000 Median : 6.900  Median : 0.2800  Median : 0.3100
## Mean   :3.029 Mean   : 7.076  Mean   : 0.3241  Mean   : 0.3084
## 3rd Qu.:4.000 3rd Qu.: 9.500  3rd Qu.: 0.6400  3rd Qu.: 0.5800
## Max.  :8.000  Max.  :34.400  Max.  : 3.6800  Max.  : 3.8600
##
## ResidualSugar   Chlorides   FreeSulfurDioxide TotalSulfurDioxide
## Min. :-127.800  Min. :-1.1710  Min. :-555.00  Min. :-823.0
## 1st Qu.: -2.000  1st Qu.: -0.0310  1st Qu.:  0.00  1st Qu.: 27.0
## Median : 3.900  Median : 0.0460  Median : 30.00  Median : 123.0
## Mean   : 5.419  Mean   : 0.0548  Mean   : 30.85  Mean   : 120.7
## 3rd Qu.: 15.900 3rd Qu.: 0.1530  3rd Qu.: 70.00  3rd Qu.: 208.0
## Max.  : 141.150 Max.  : 1.3510  Max.  : 623.00  Max.  : 1057.0
## NA's   :616       NA's  :638       NA's  :647       NA's  :682
## Density        pH          Sulphates   Alcohol
## Min. :0.8881  Min. :0.480   Min. :-3.1300  Min. :-4.70
## 1st Qu.:0.9877 1st Qu.:2.960   1st Qu.: 0.2800  1st Qu.: 9.00
## Median :0.9945 Median :3.200   Median : 0.5000  Median :10.40
## Mean   :0.9942 Mean   :3.208   Mean   : 0.5271  Mean   :10.49
## 3rd Qu.:1.0005 3rd Qu.:3.470   3rd Qu.: 0.8600  3rd Qu.:12.40
## Max.  :1.0992 Max.  :6.130   Max.  : 4.2400  Max.  :26.50
## NA's   :395       NA's  :1210     NA's  :653
## LabelAppeal    AcidIndex   STARS
## Min. :-2.000000 Min. : 4.000  Min. :1.000
## 1st Qu.:-1.000000 1st Qu.: 7.000  1st Qu.:1.000
## Median : 0.000000 Median : 8.000  Median :2.000
## Mean   : -0.009066 Mean   : 7.773  Mean   :2.042
## 3rd Qu.: 1.000000 3rd Qu.: 8.000  3rd Qu.:3.000
## Max.  : 2.000000 Max.  :17.000  Max.  :4.000
## NA's   :3359
```

The statistics of TARGET Variable.

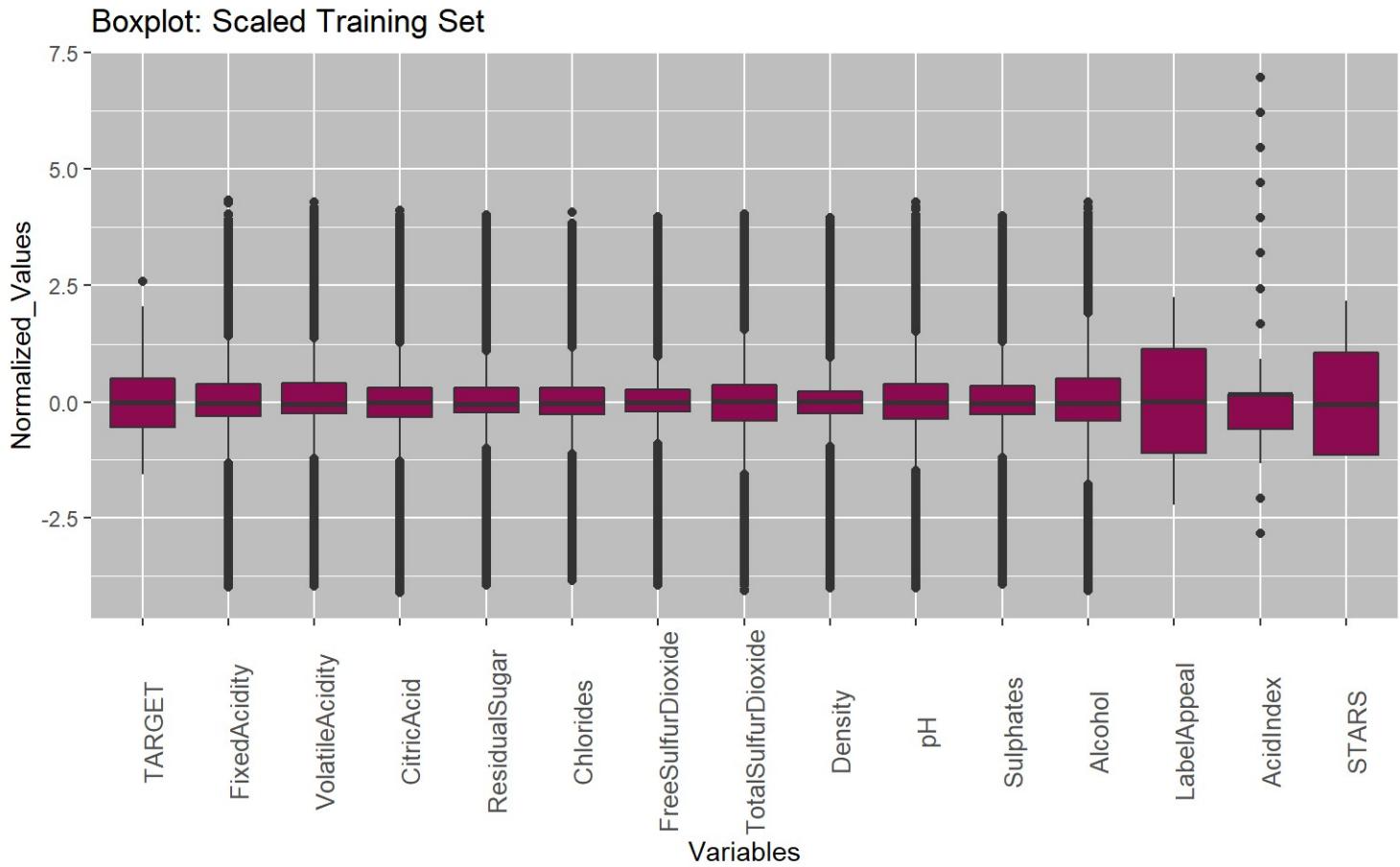
- **TARGET:** Number of Cases Purchased as Actual

	##	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	-0.88

Data Exploration

The first step we did was to import the data from GitHub, remove the index and look at the structure of the data.

Outliers

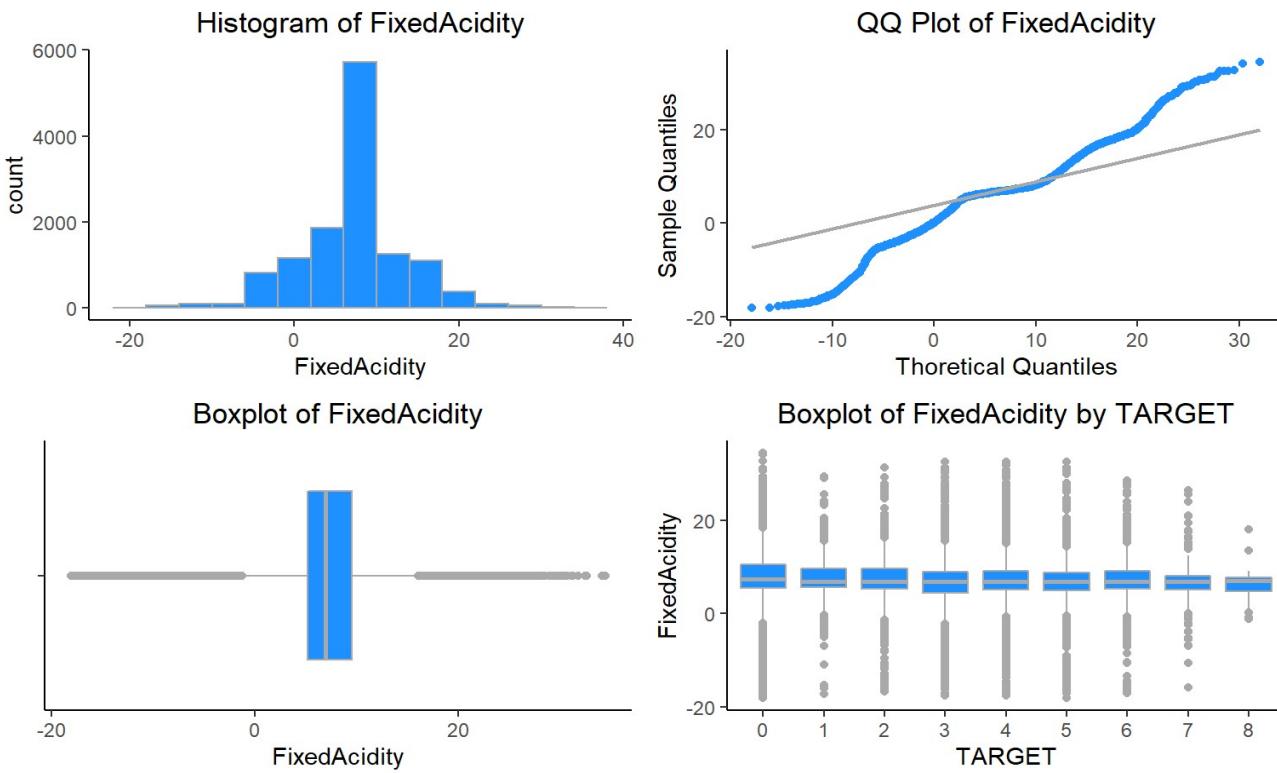


The box plot below shows that outliers exist in variables `FixedAcidity`, `VolatileAcidity`, `CitricAcid`, `ResidualSugar`, `Chlorides`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `Density`, `pH`, `Sulphates`, `Alcohol`, `LabelAppeal` and `AcidIndex`. We use scaled training set to draw the box plot to show the corresponding outliers by ratio.

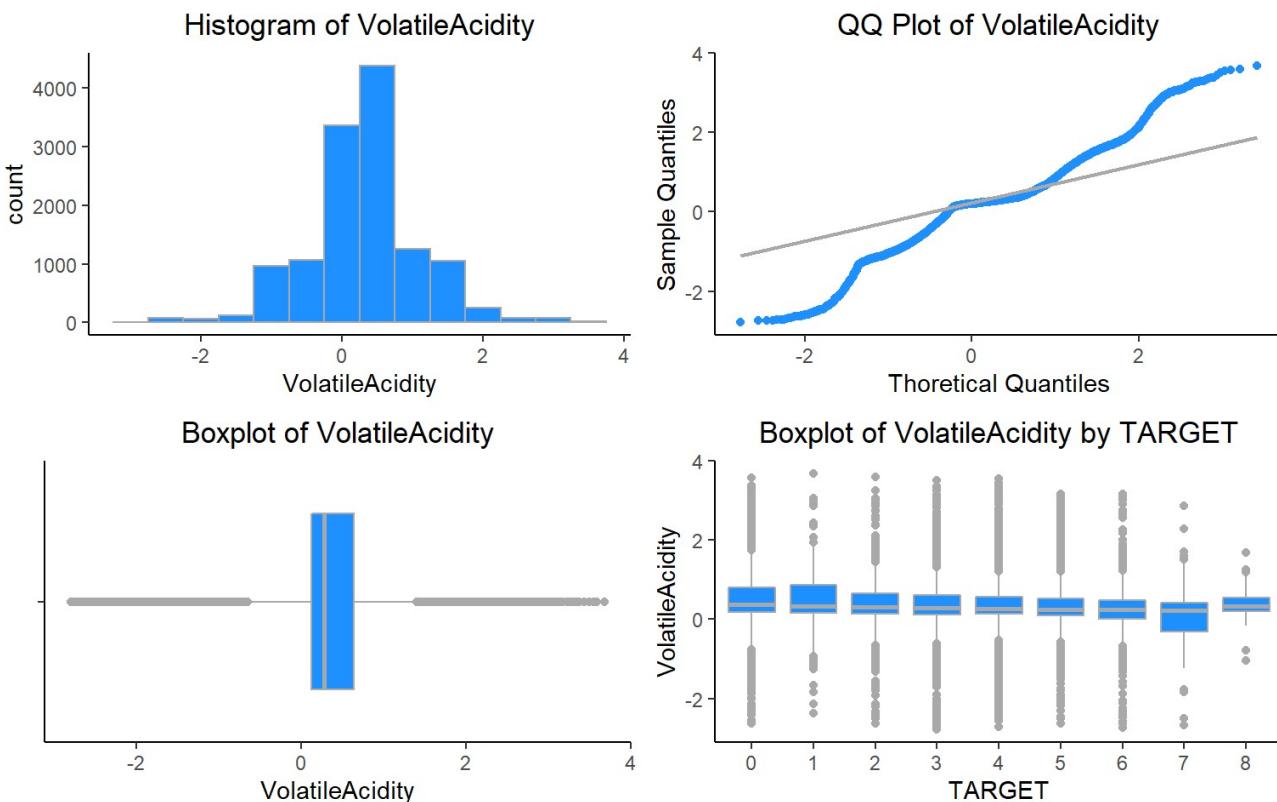
Histogram of attributes

Lets analyze the attributes of the dataset individually in terms of Histogram,, QQ Plot and BoxPlot in comparison to TARGET

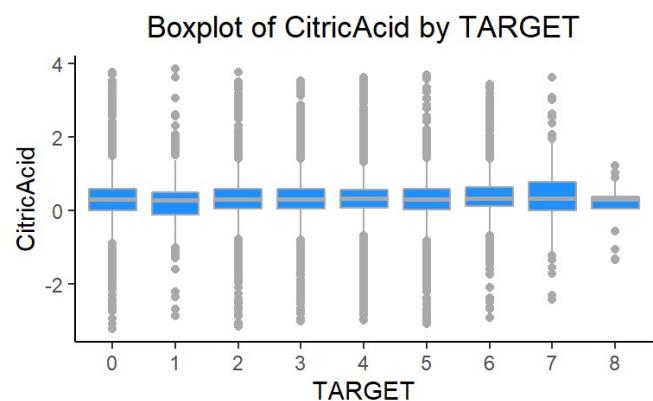
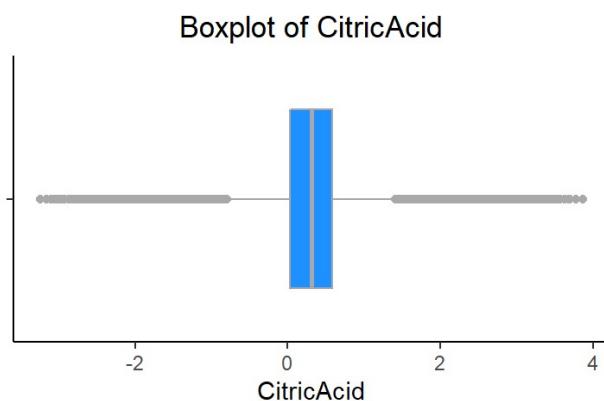
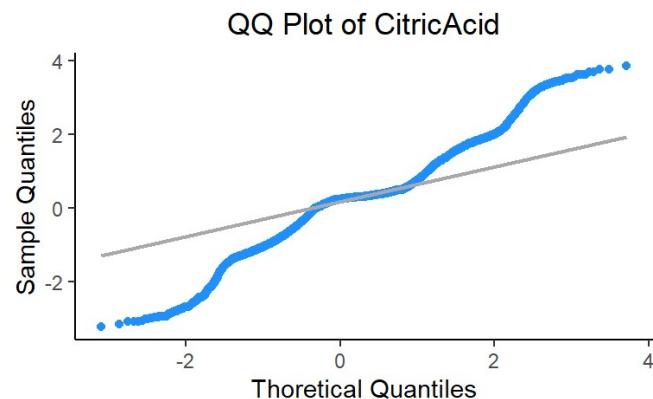
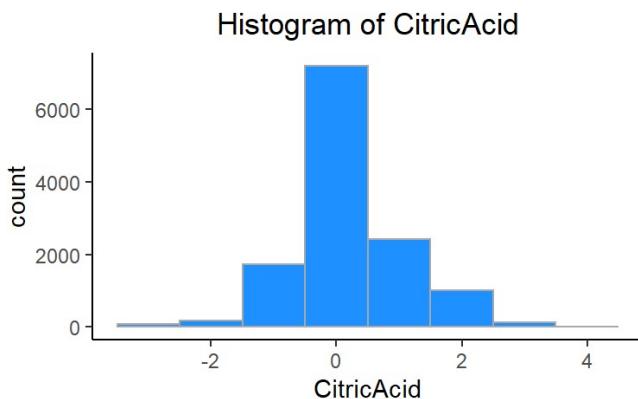
- **FixedAcidity:** This variable tells us about the FixedAcidity of wine.



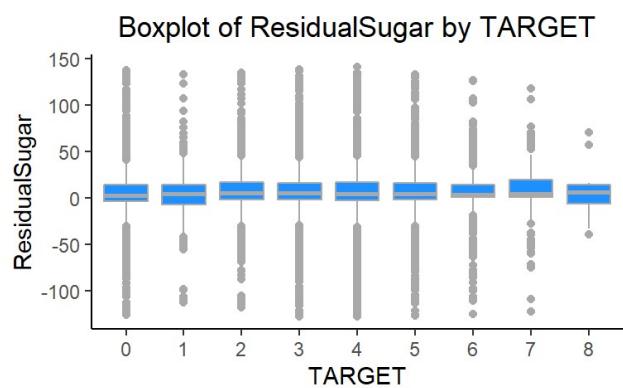
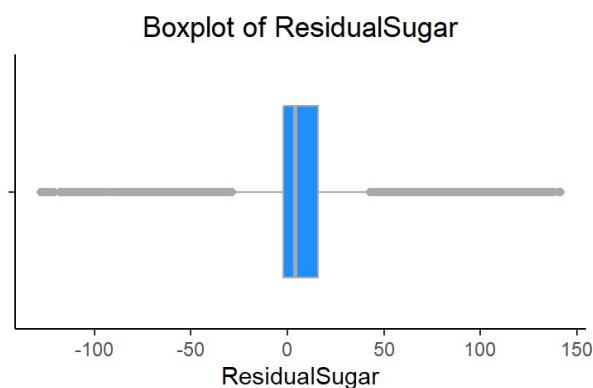
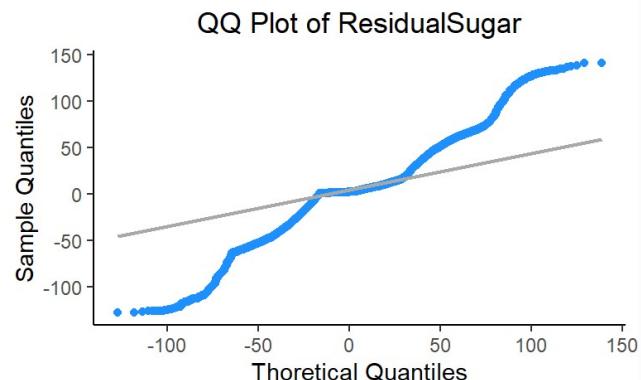
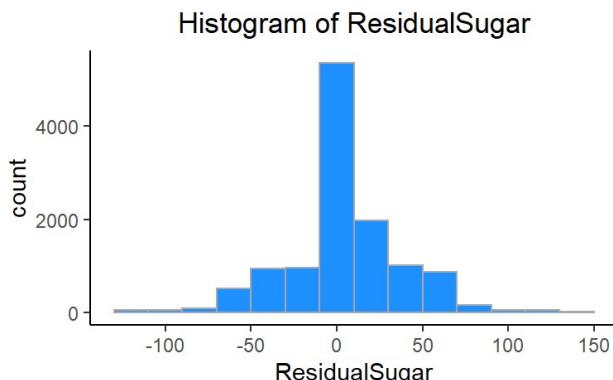
- **VolatileAcidity:** This variable tells us about the VolatileAcidity content of Wine.



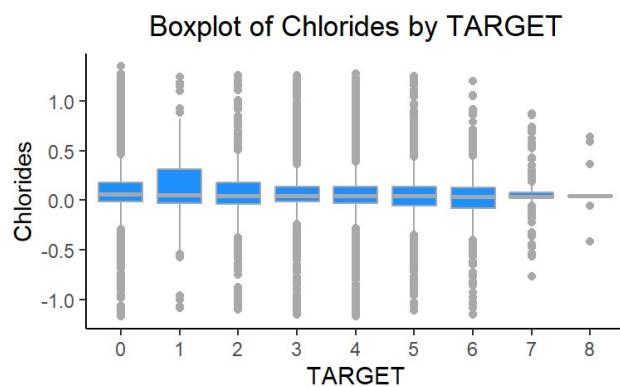
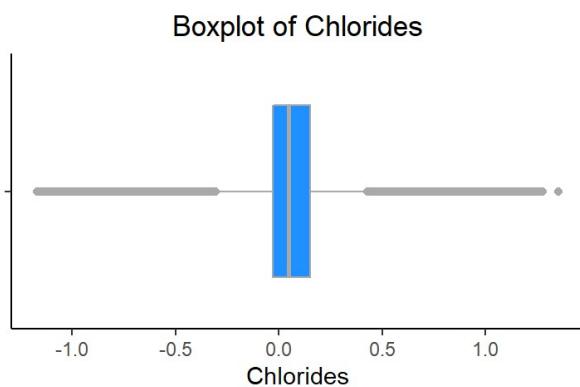
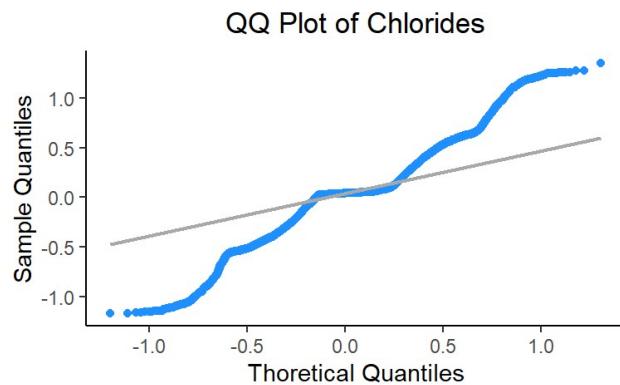
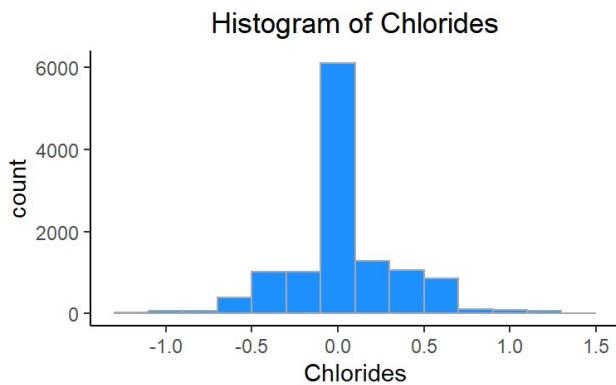
- **CitricAcid:** This variable tells us about the Citric Acid Content of wine.



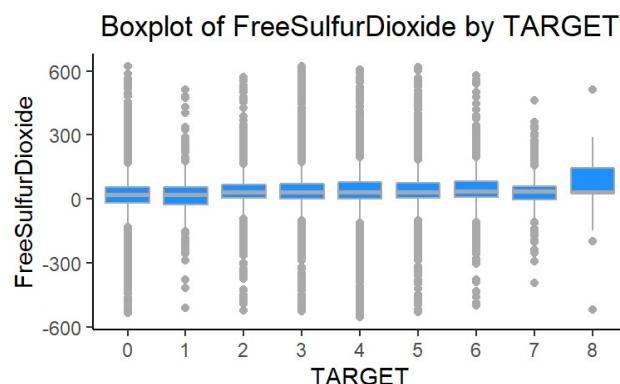
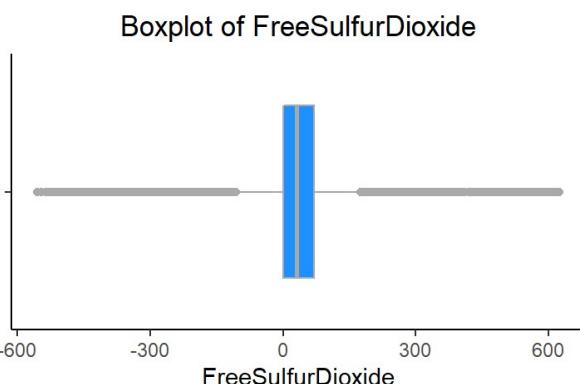
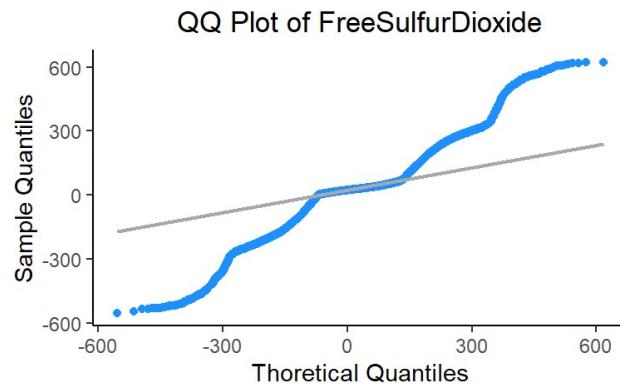
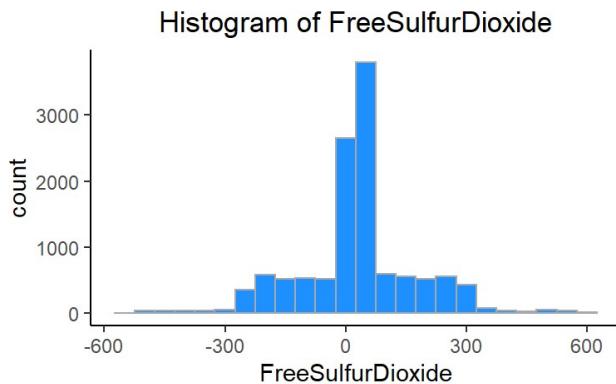
- **ResidualSugar:** This variable tells us about the ResidualSugar of wine.



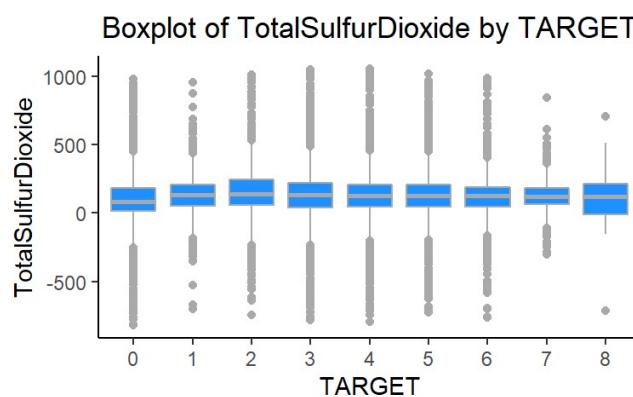
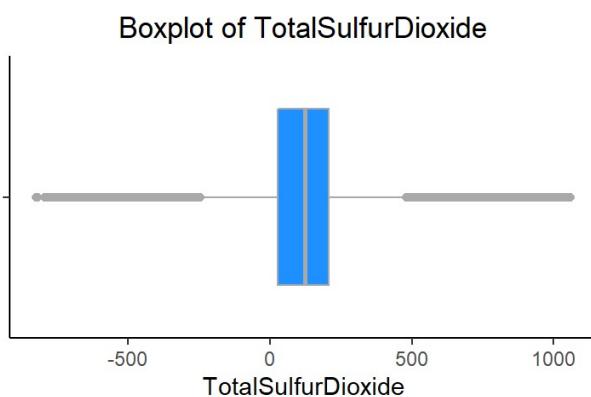
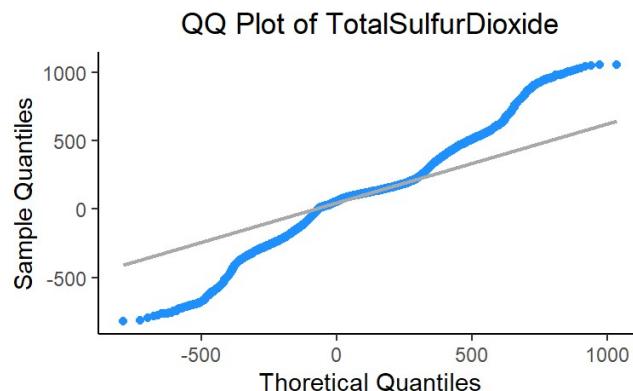
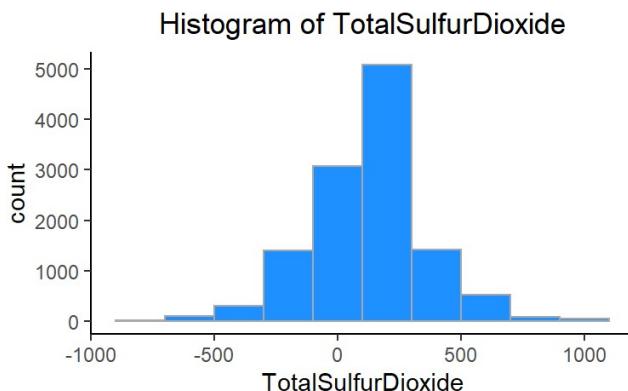
- **Chlorides:** This variable tells us about the Chloride content of wine.



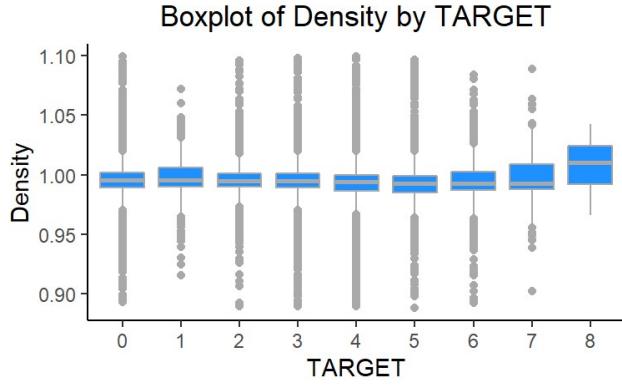
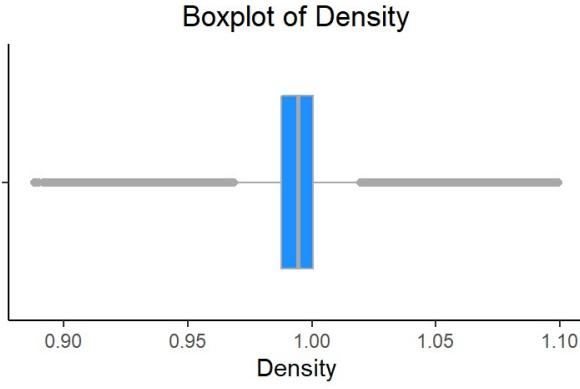
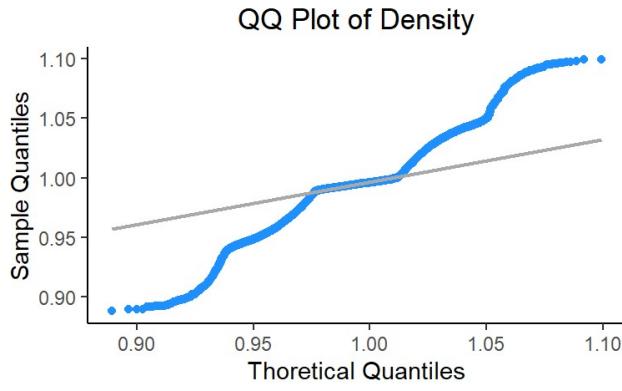
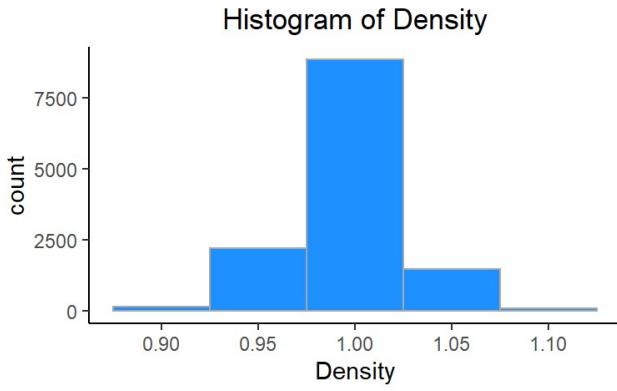
- **FreeSulfurDioxide :** This variable tells us about the Sulfur Dioxide content of wine.



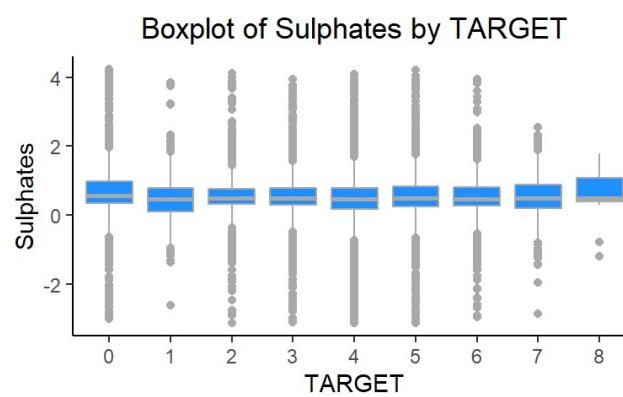
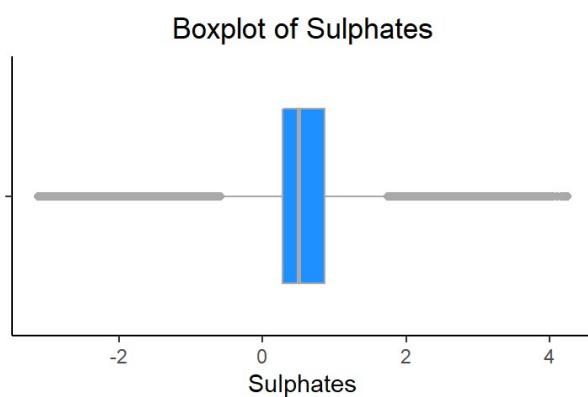
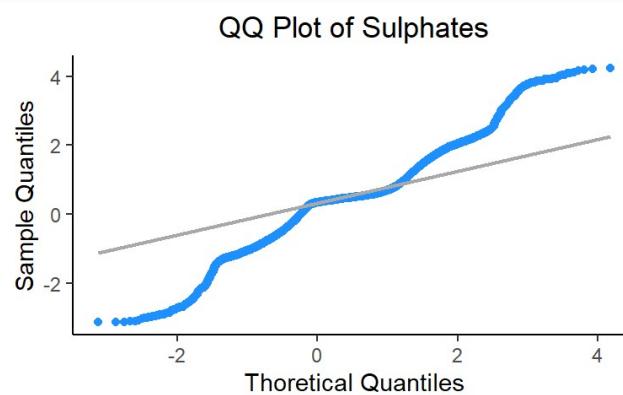
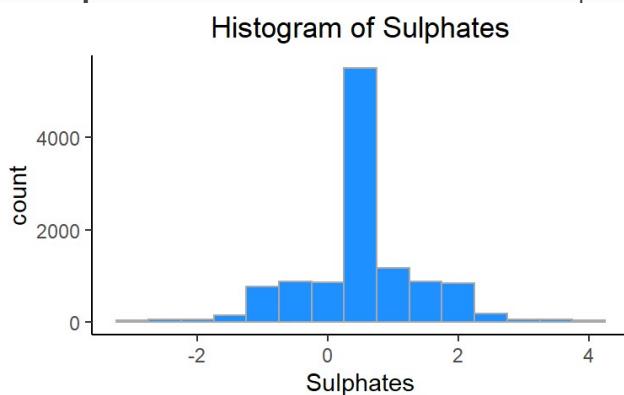
- **TotalSulfurDioxide** : This variable tells us about the Total Sulfur Dioxide of Wine.



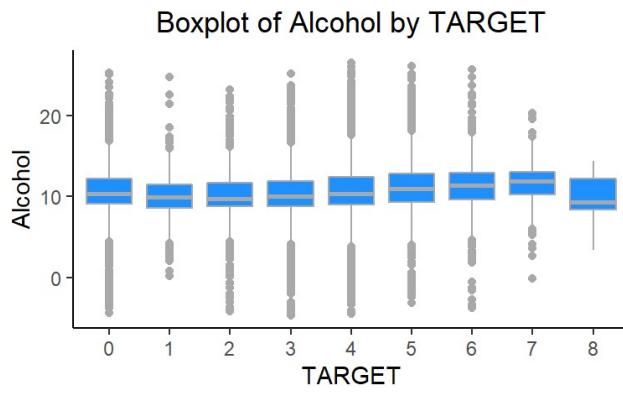
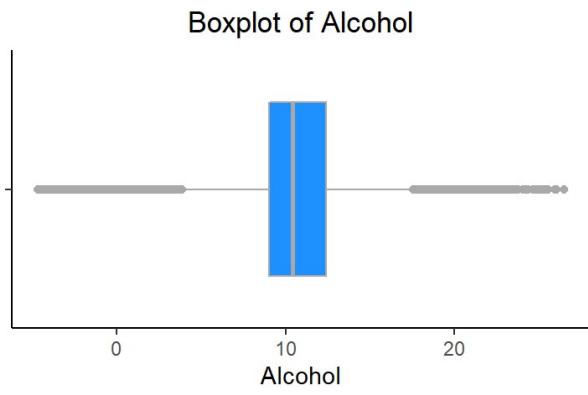
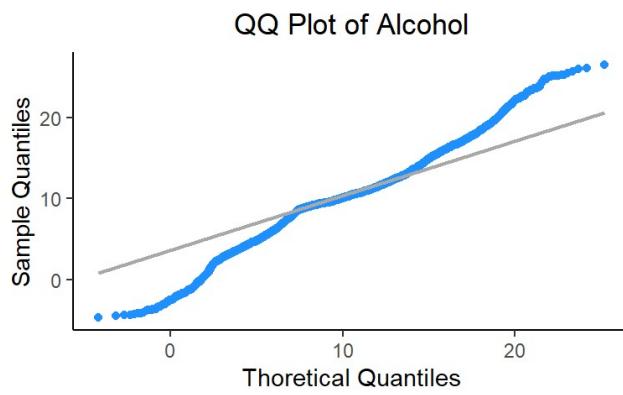
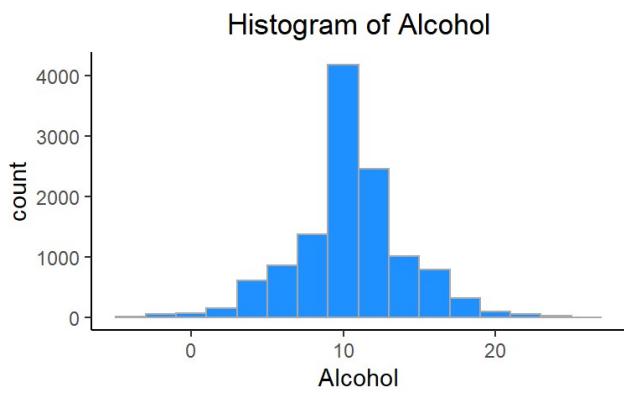
- **Density**: This variable tells us about the Density of wine.



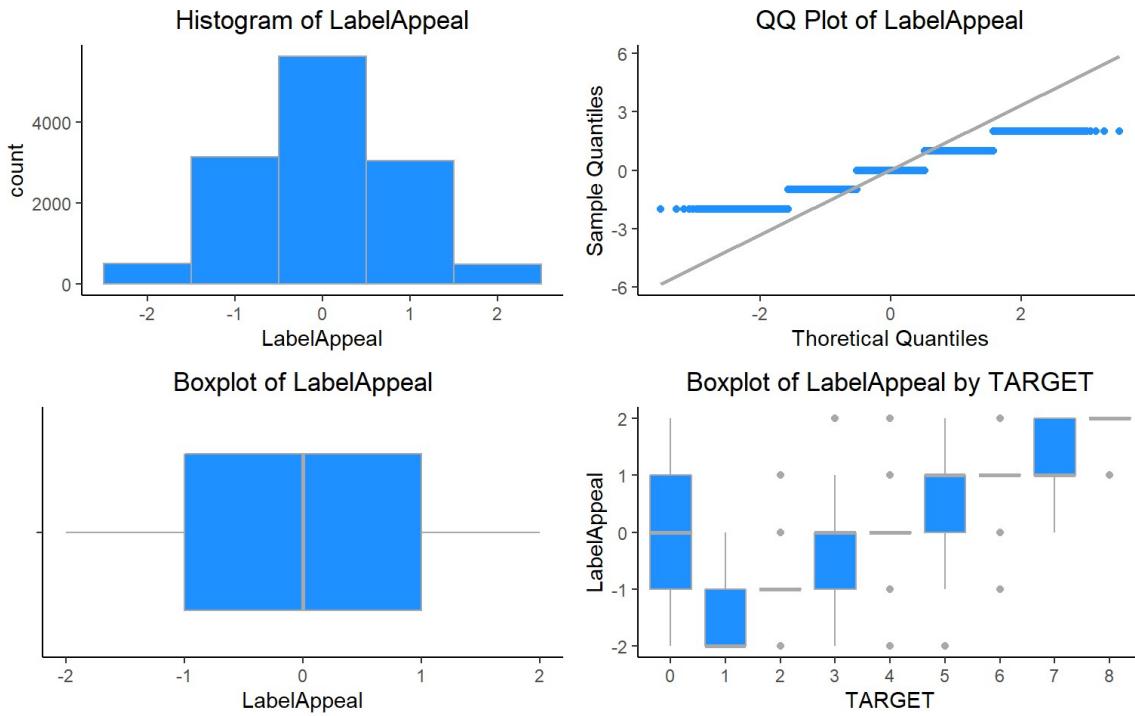
- **Sulphates:** This variable tells us about the Sulphates content of wine.



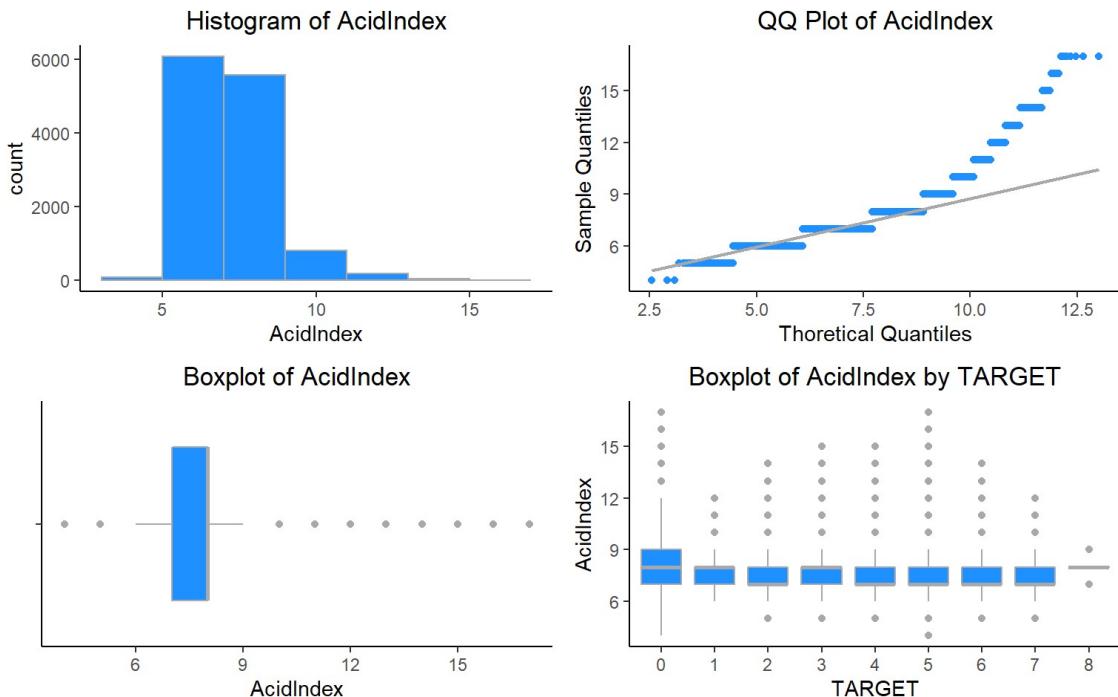
- **Alcohol:** This variable tells us about the Alcohol content.



- **LabelAppeal:** Marketing Score indicating 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.

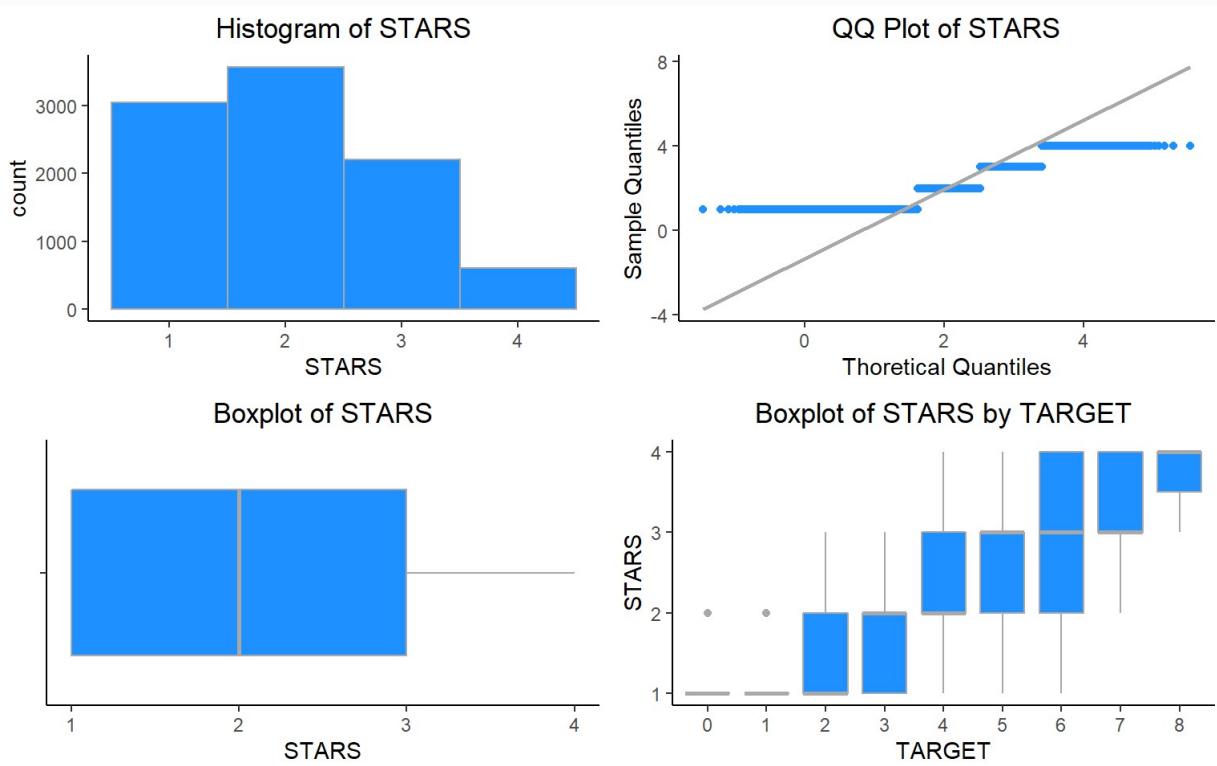


- **AcidIndex:** Proprietary method of testing total acidity of wine by using a weighted average.

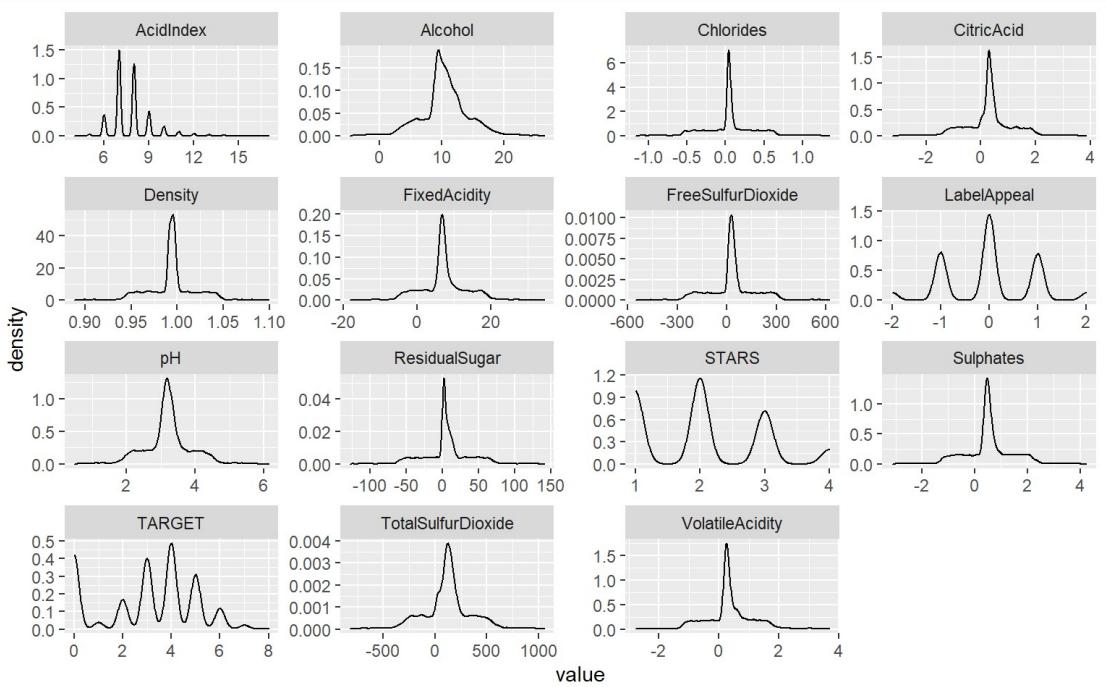


AcidIndex seems to be skewed slightly. It also has a skew of 1.68. This is not enough to worry about.

- **STARS:** Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor. A high number of stars suggests high sales.



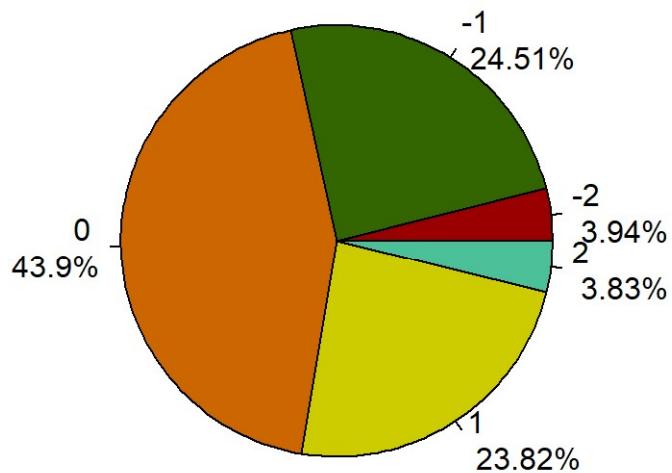
Density Plot



The scaled histogram and density plots show that variables `AcidIndex` is right skewed; `AcidIndex`, `STARS`, `LabelAppeal` and `TARGET` have multimodal distribution; while most others seem to be normally distributed due to the bell curve they display.

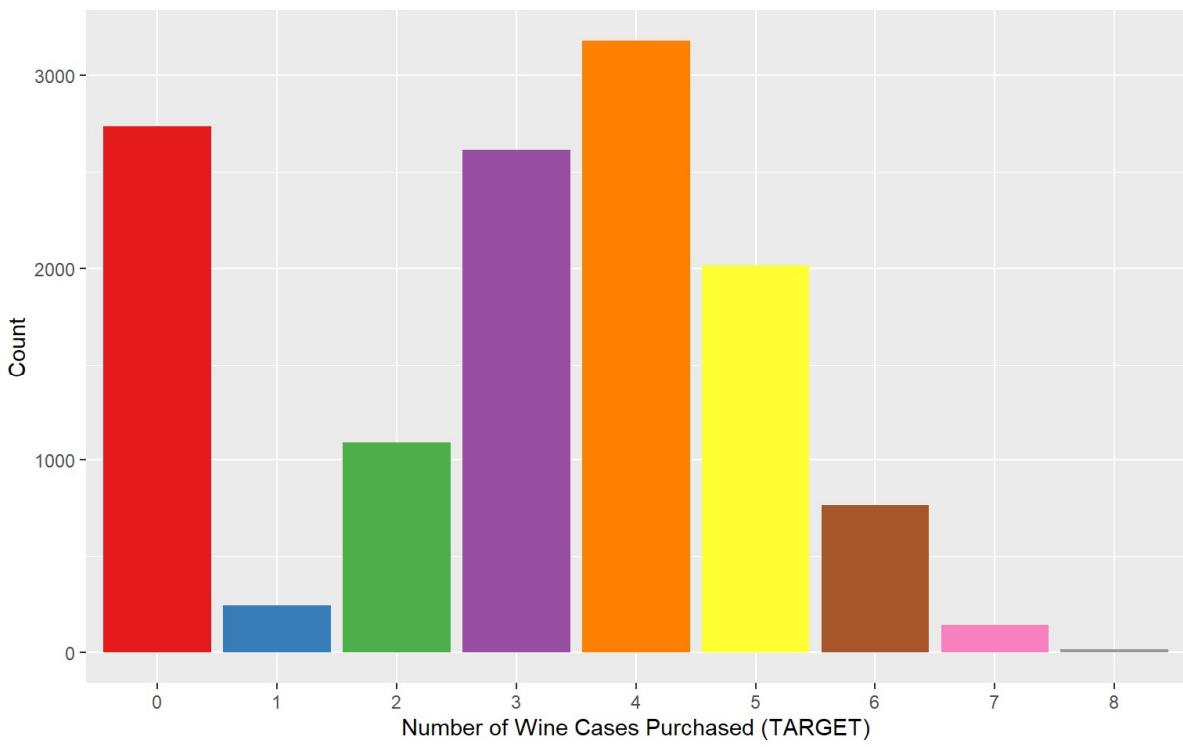
Label Scores

Marketing Scores Proportioned



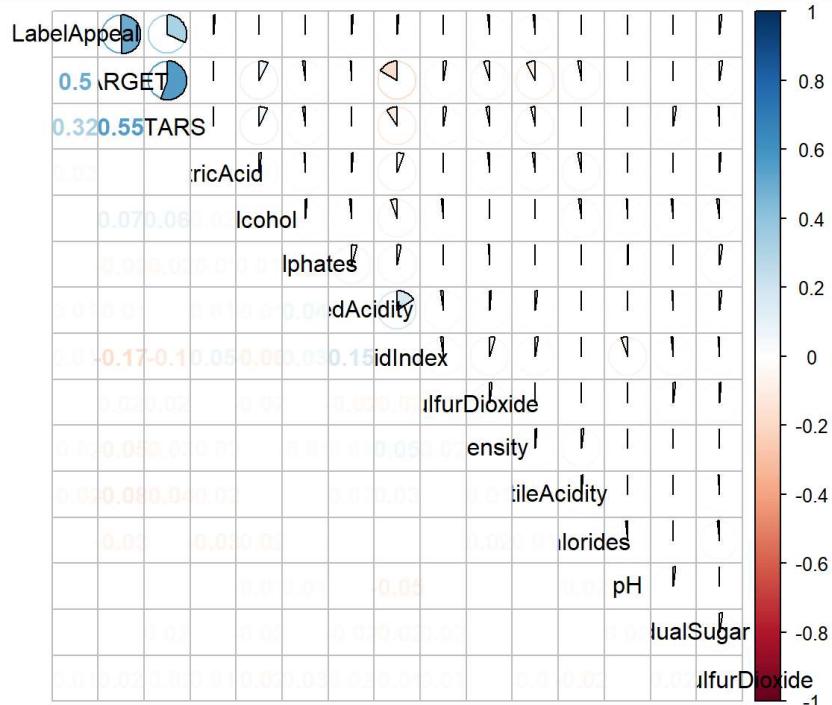
Univariate Analysis

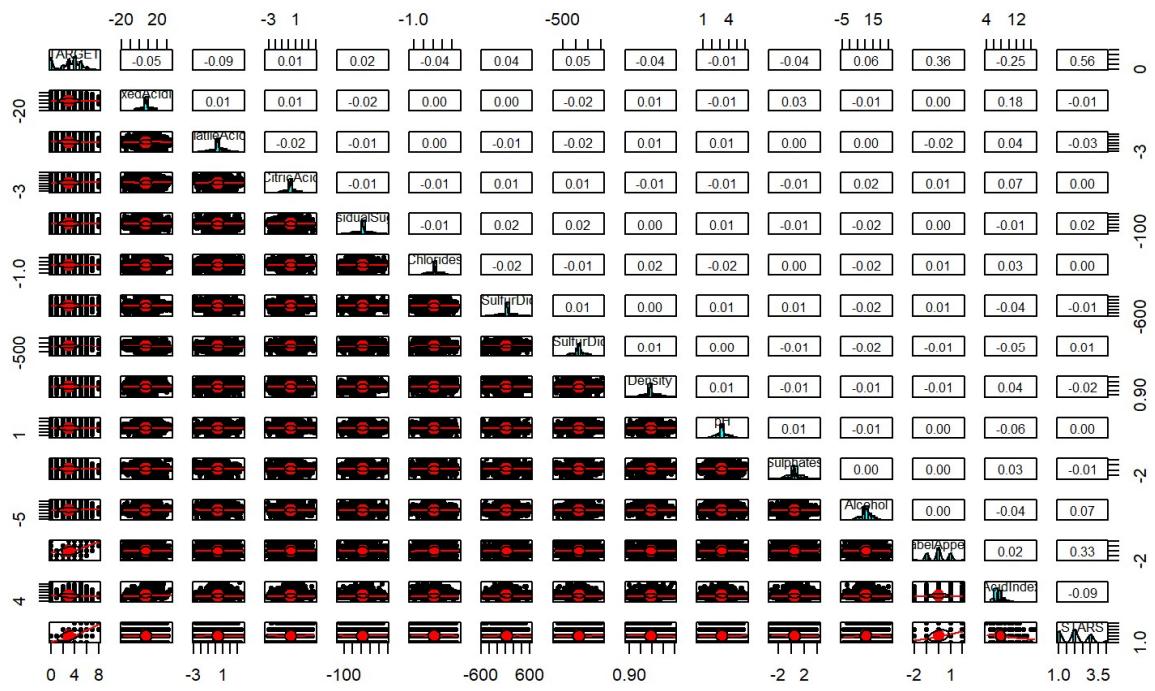
Response Variable



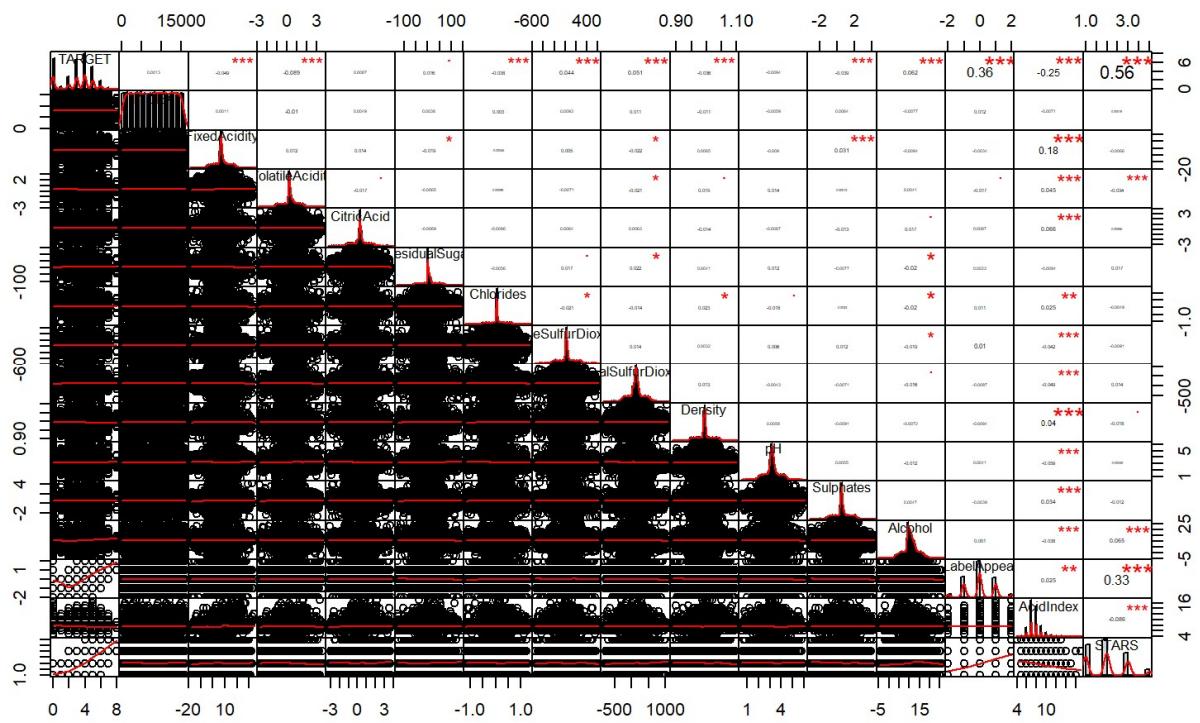
Correlation Plot

We implement a correlation matrix to better understand the correlation between variables in the dataset.

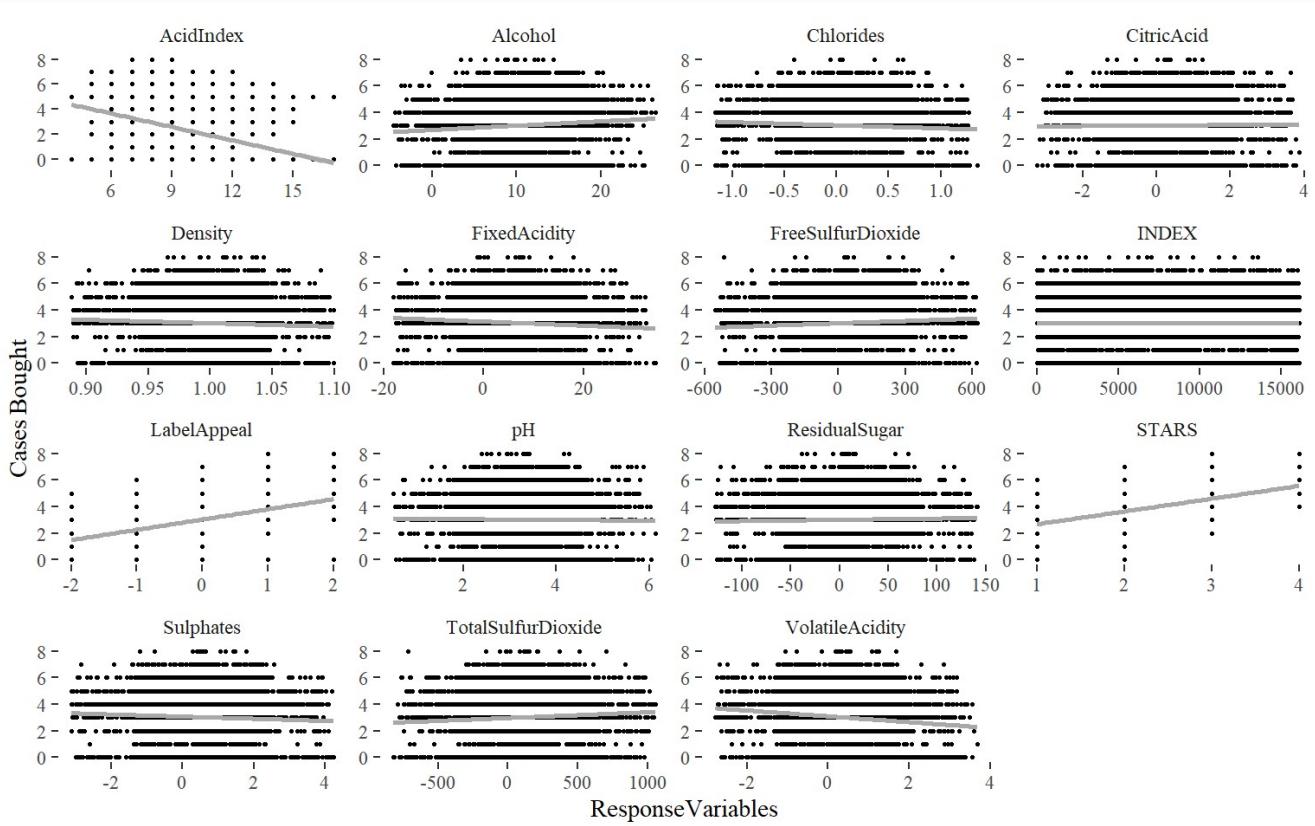




The correlation matrix below shows that the response variable **TARGET** has strong positive relationship ($>=0.6$) with variables **FixedAcidity**, **CitricAcid**, **ResidualSugar**, **Density**, **Alcohol**.



Scatter plots against **TARGET**:



There don't seem to be any crazy patterns here. It mostly looks linear which is a good sign for us. **STARS** and **LabelAppeal** look like they have the greatest correlation.

Variable <chr>	Correlation_vs_Response <dbl>	FixedAcidity <dbl>	VolatileAcidity <dbl>	CitricAcid <dbl>
TARGET	1.000000000	-0.049010939	-0.08879321	0.008684633
FixedAcidity	-0.049010939	1.000000000	0.01237500	0.014240478
VolatileAcidity	-0.088793212	0.012375004	1.000000000	-0.016953371
CitricAcid	0.008684633	0.014240478	-0.01695337	1.000000000
ResidualSugar	NA	NA	NA	NA
Chlorides	NA	NA	NA	NA
FreeSulfurDioxide	NA	NA	NA	NA
TotalSulfurDioxide	NA	NA	NA	NA
Density	-0.035517502	0.006476528	0.01473492	-0.013952174
pH	NA	NA	NA	NA

1-10 of 15 rows | 1-5 of 16 columns

Previous 1 2 Next

Consolidated Data Dictionary

As a summary of the data exploration process, a data dictionary is created below:

Variable	Missing_Value	Mean	Median	Max	Min	SD	Correlation_vs_Response
TARGET	No	NA	NA	NA	NA	NA	1.00
INDEX	No	NA	NA	NA	NA	NA	0.00
FixedAcidity	No	7.08	6.90	34.40	-18.10	6.32	-0.05
VolatileAcidity	No	0.32	0.28	3.68	-2.79	0.78	-0.09
CitricAcid	No	0.31	0.31	3.86	-3.24	0.86	0.01
ResidualSugar	No	NA	NA	NA	NA	NaN	NA
Chlorides	No	NA	NA	NA	NA	NaN	NA
FreeSulfurDioxide	No	NA	NA	NA	NA	NaN	NA
TotalSulfurDioxide	No	NA	NA	NA	NA	NaN	NA
Density	No	0.99	0.99	1.10	0.89	0.03	-0.04
pH	No	NA	NA	NA	NA	NaN	NA
Sulphates	No	NA	NA	NA	NA	NaN	NA
Alcohol	No	NA	NA	NA	NA	NaN	NA
LabelAppeal	No	-0.01	0.00	2.00	-2.00	0.89	0.36
AcidIndex	No	7.77	8.00	17.00	4.00	1.32	-0.25
STARS	No	NA	NA	NA	NA	NaN	NA

DATA PREPARATION

Lets first split the data into training and test.

MICE (Multivariate Imputation by Chained Equations) package helps in inspecting, imputing, diagonise, analyze, pool the result, and generate simulated incomplete data

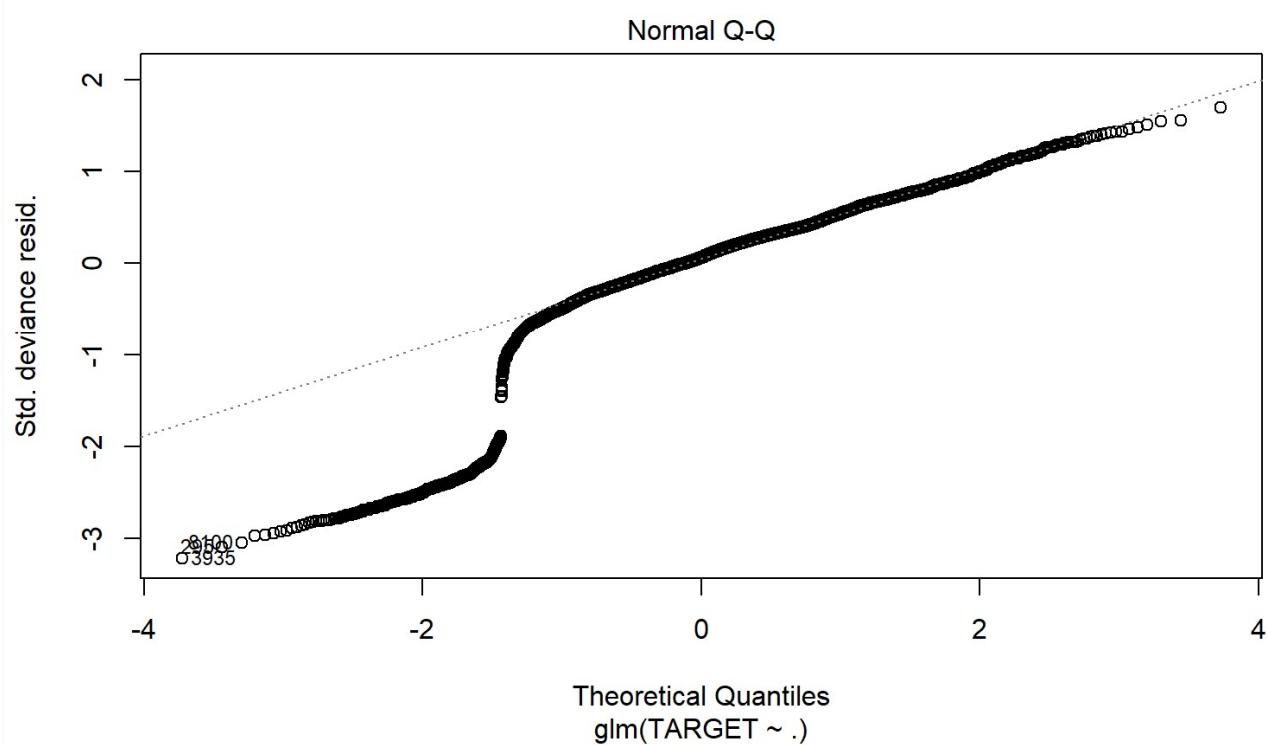
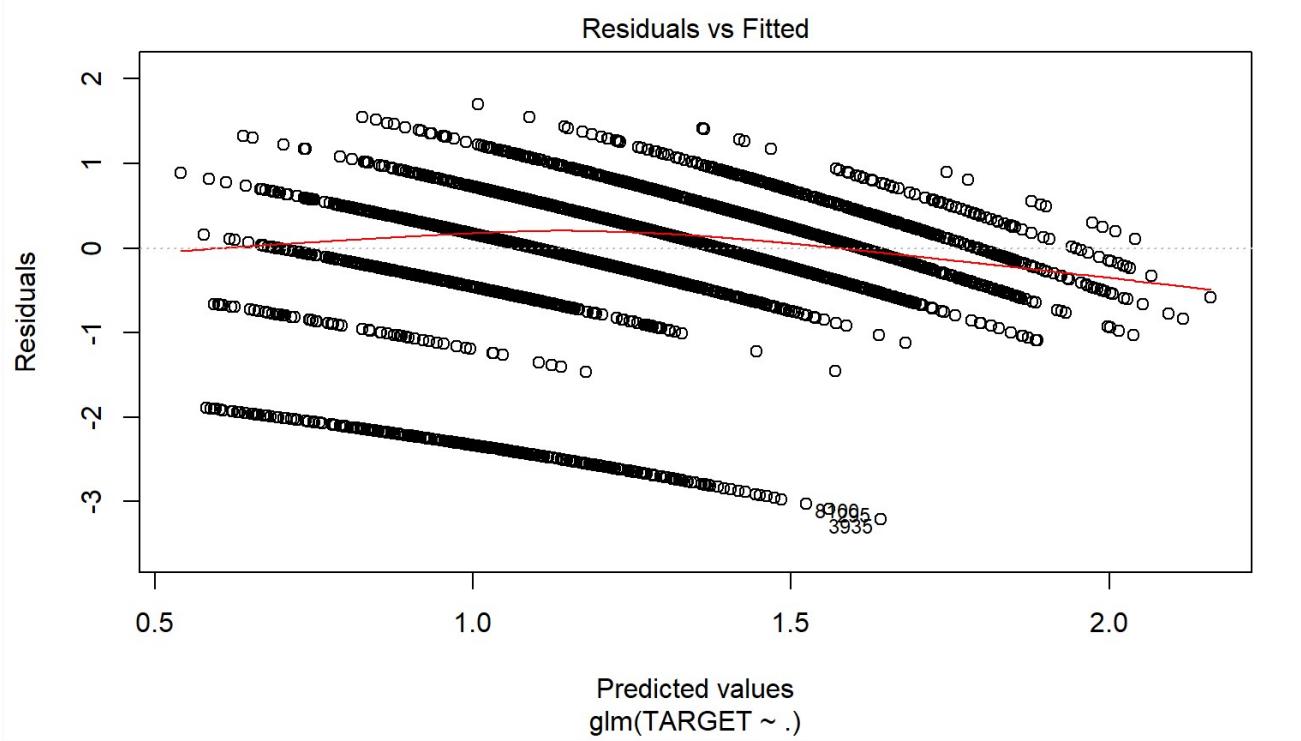
Given the low correlation between AcidIndex and TARGET it might not make a huge difference, however, we will log transform it to test.

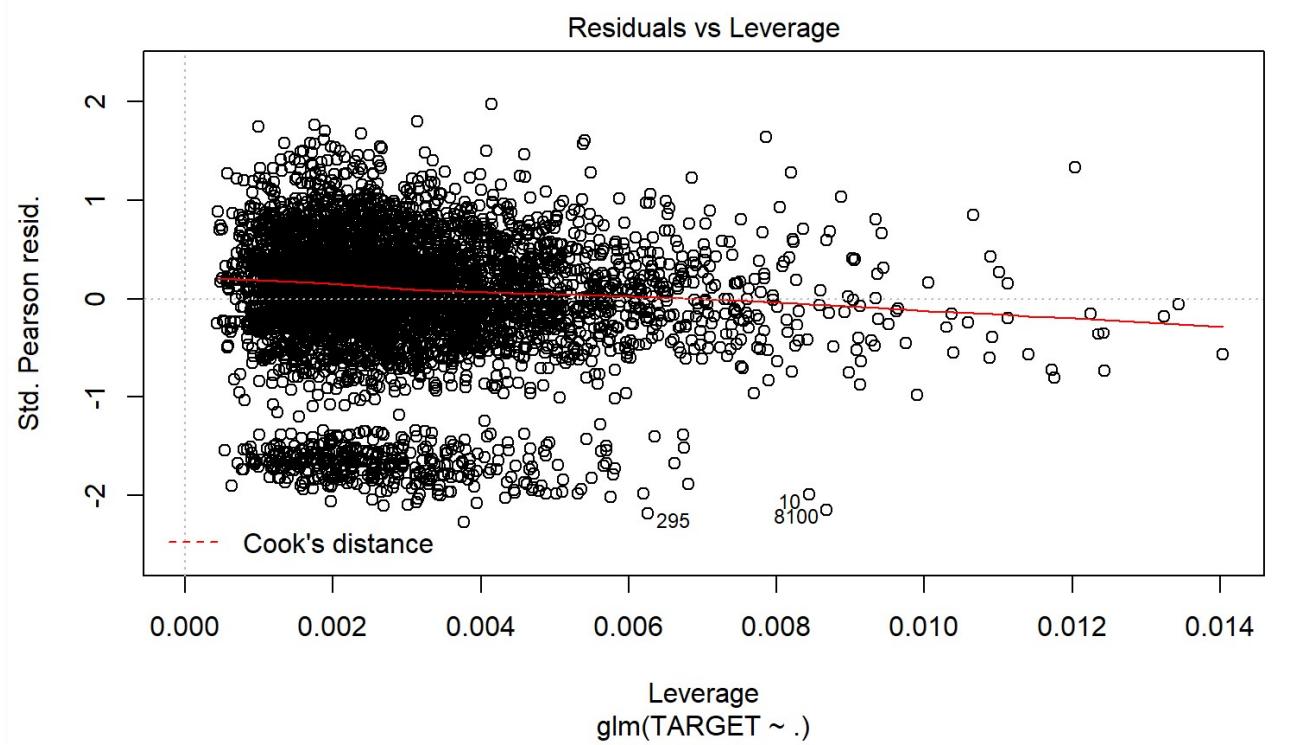
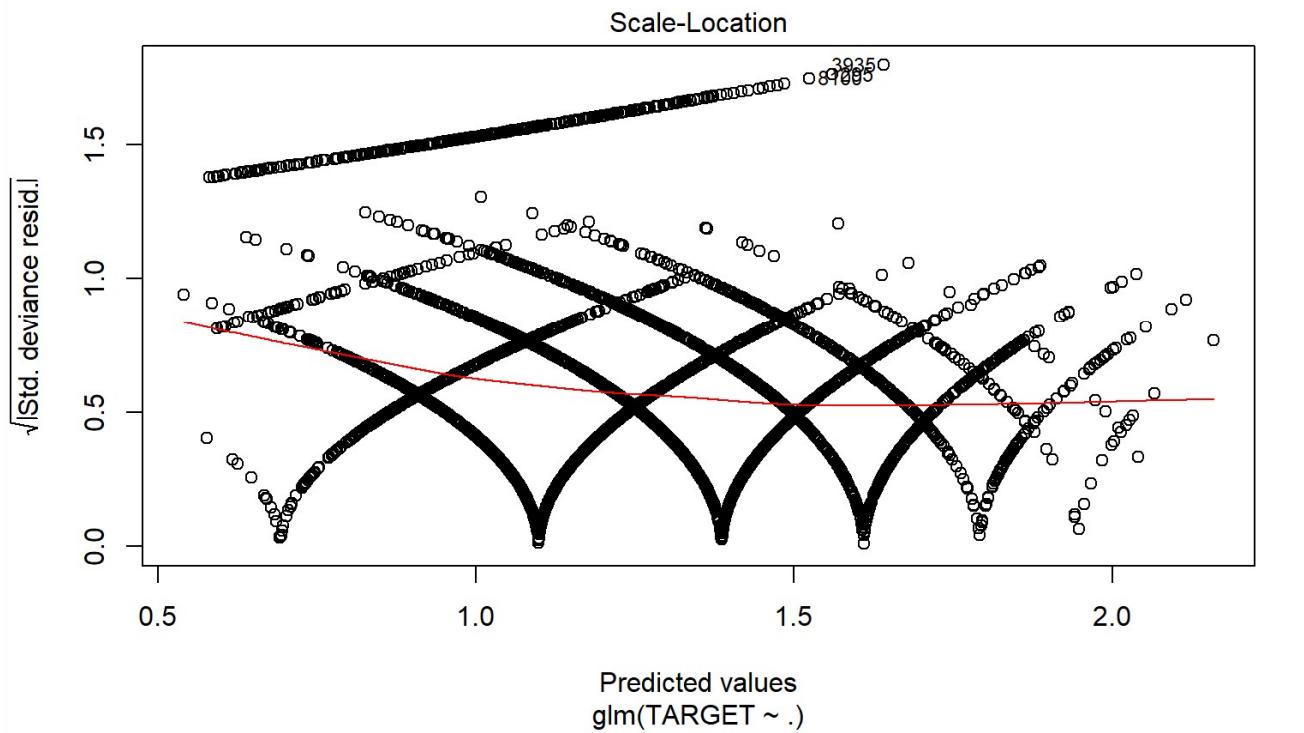
BUILD MODELS

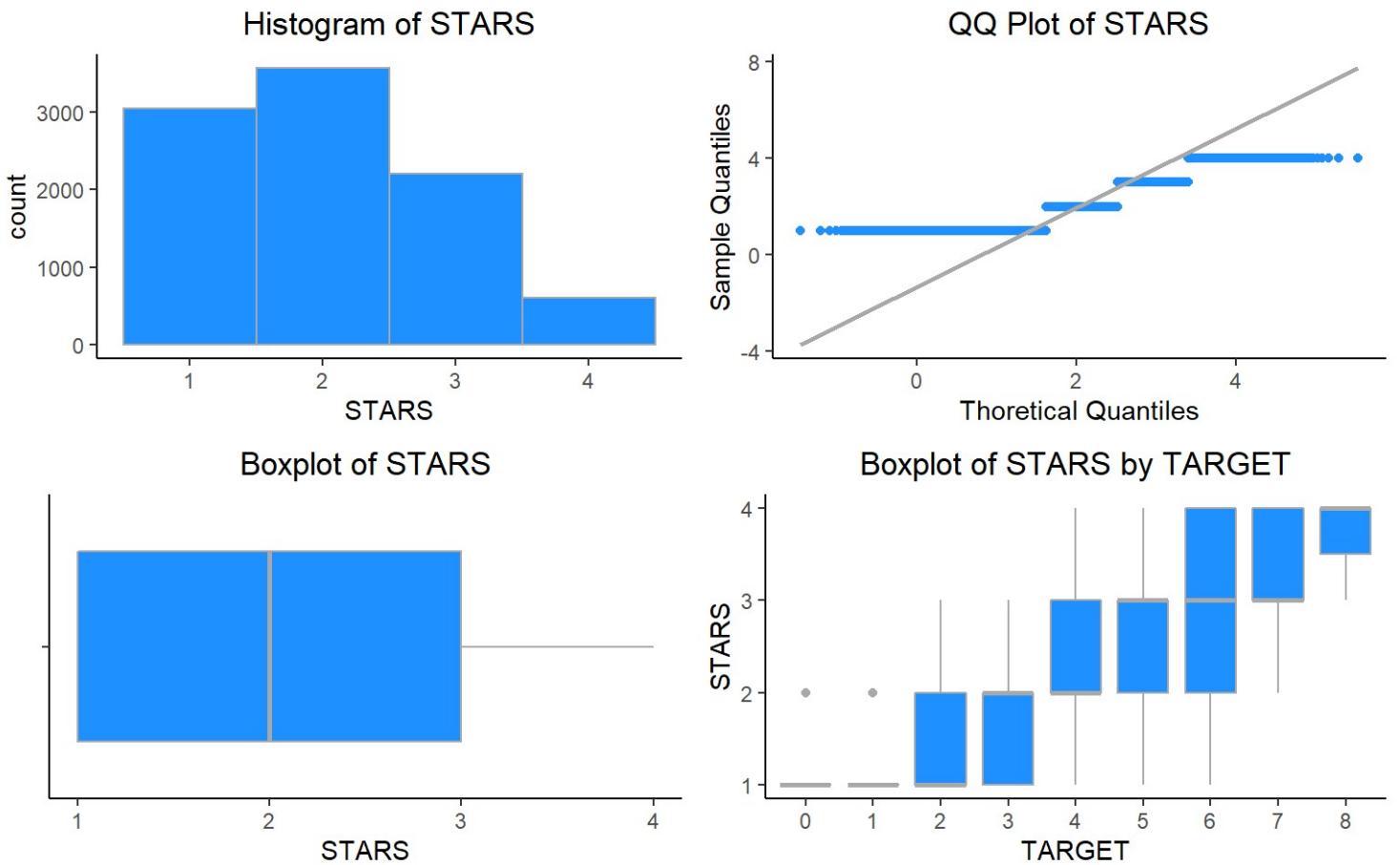
Model I: Poisson Model

Model 1: Poisson Model without imputations

```
##  
## 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
```







Model 2: Poisson Model without imputations and only significant variables

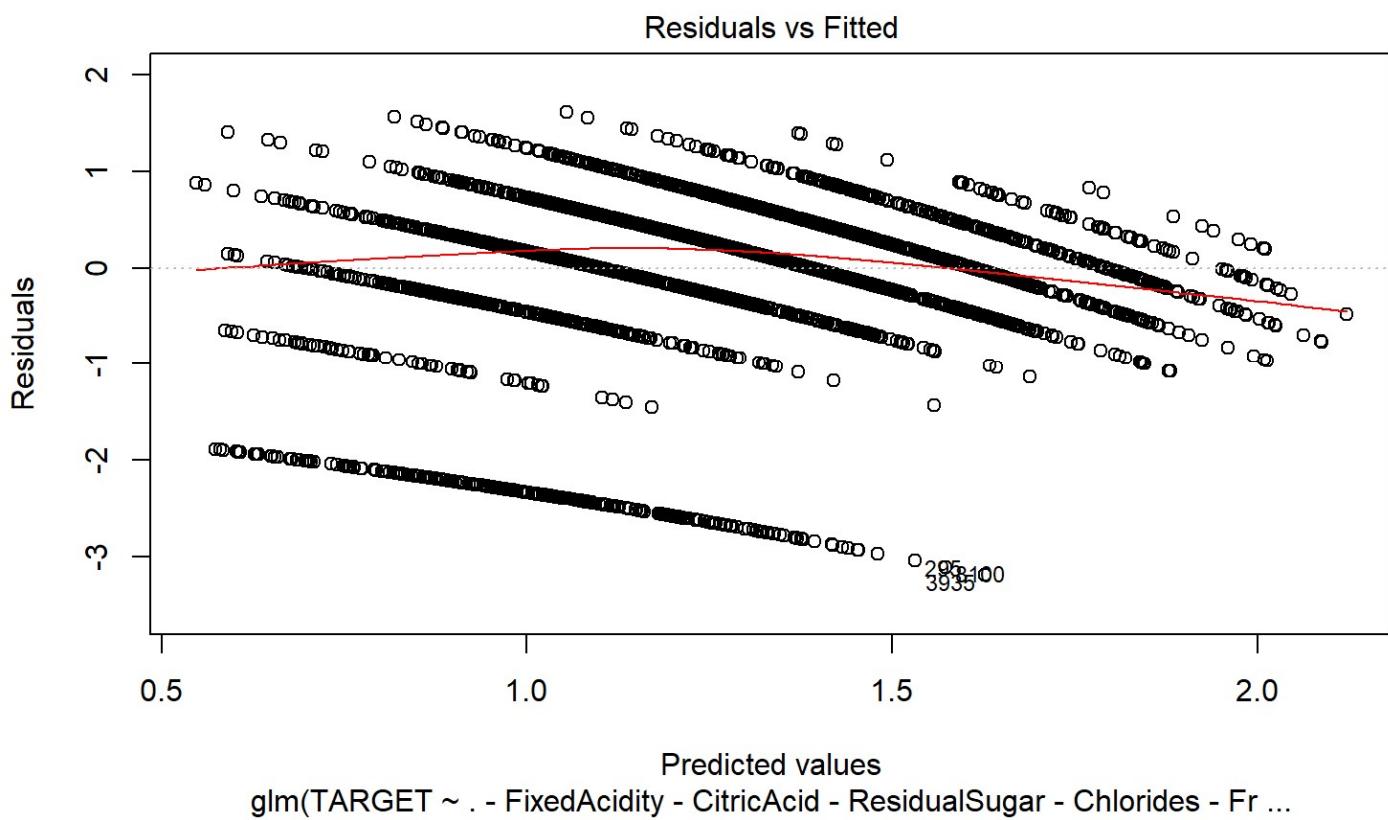
```

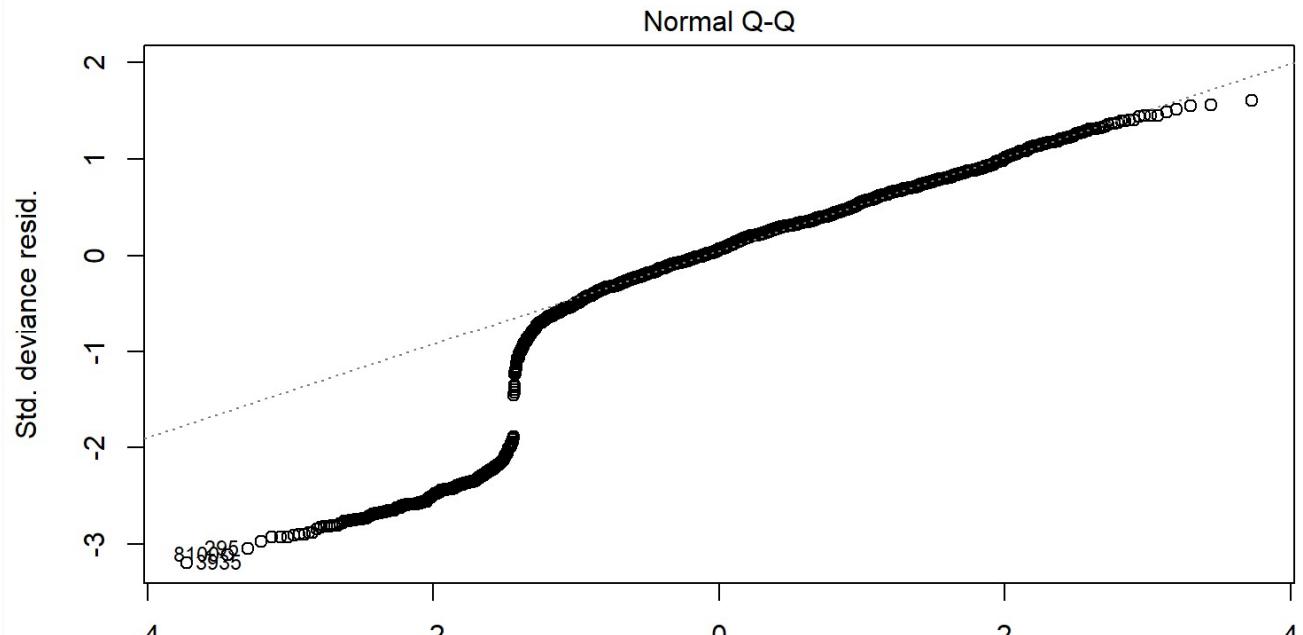
## 
## Call:
## glm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##   Chlorides - FreeSulfurDioxide - TotalSulfurDioxide - Density -
##   pH - Sulphates - Alcohol, family = poisson, data = wine_train1)
## 
## Deviance Residuals:
##    Min      1Q      Median      3Q      Max 
## -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
## 
```

```

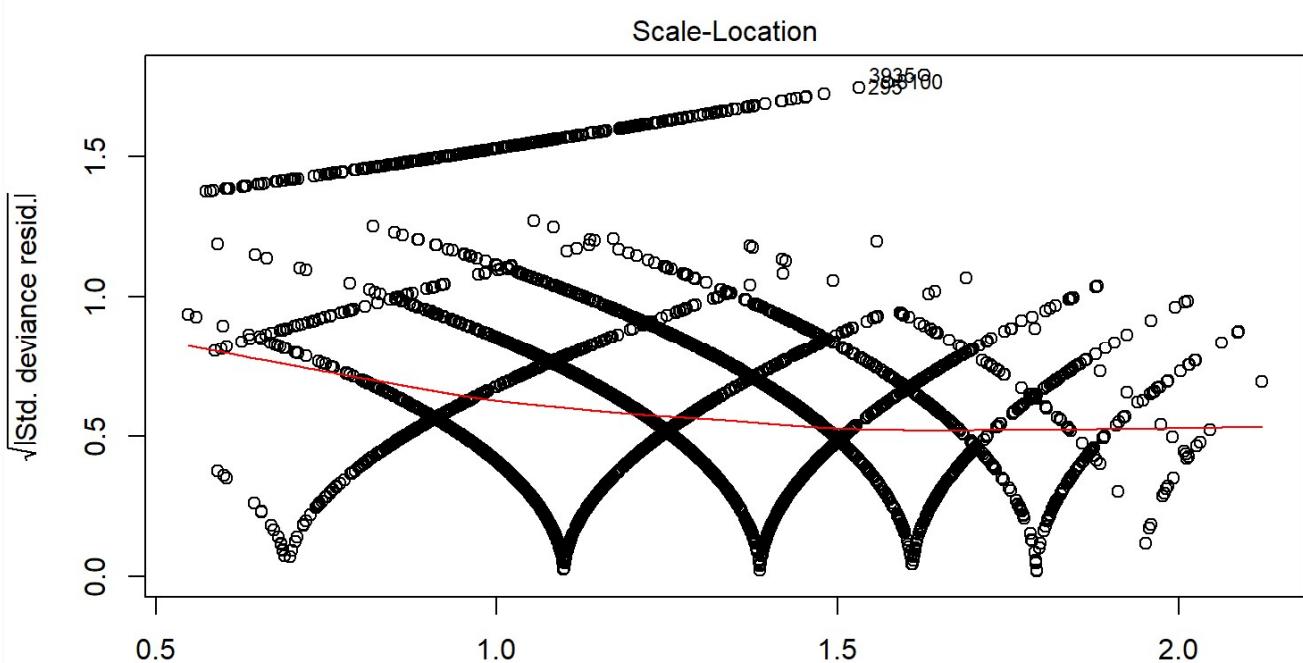
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 4720.5 on 5143 degrees of freedom
## Residual deviance: 3253.1 on 5139 degrees of freedom
## (5093 observations deleted due to missingness)
## AIC: 18535
##
## Number of Fisher Scoring iterations: 5

```

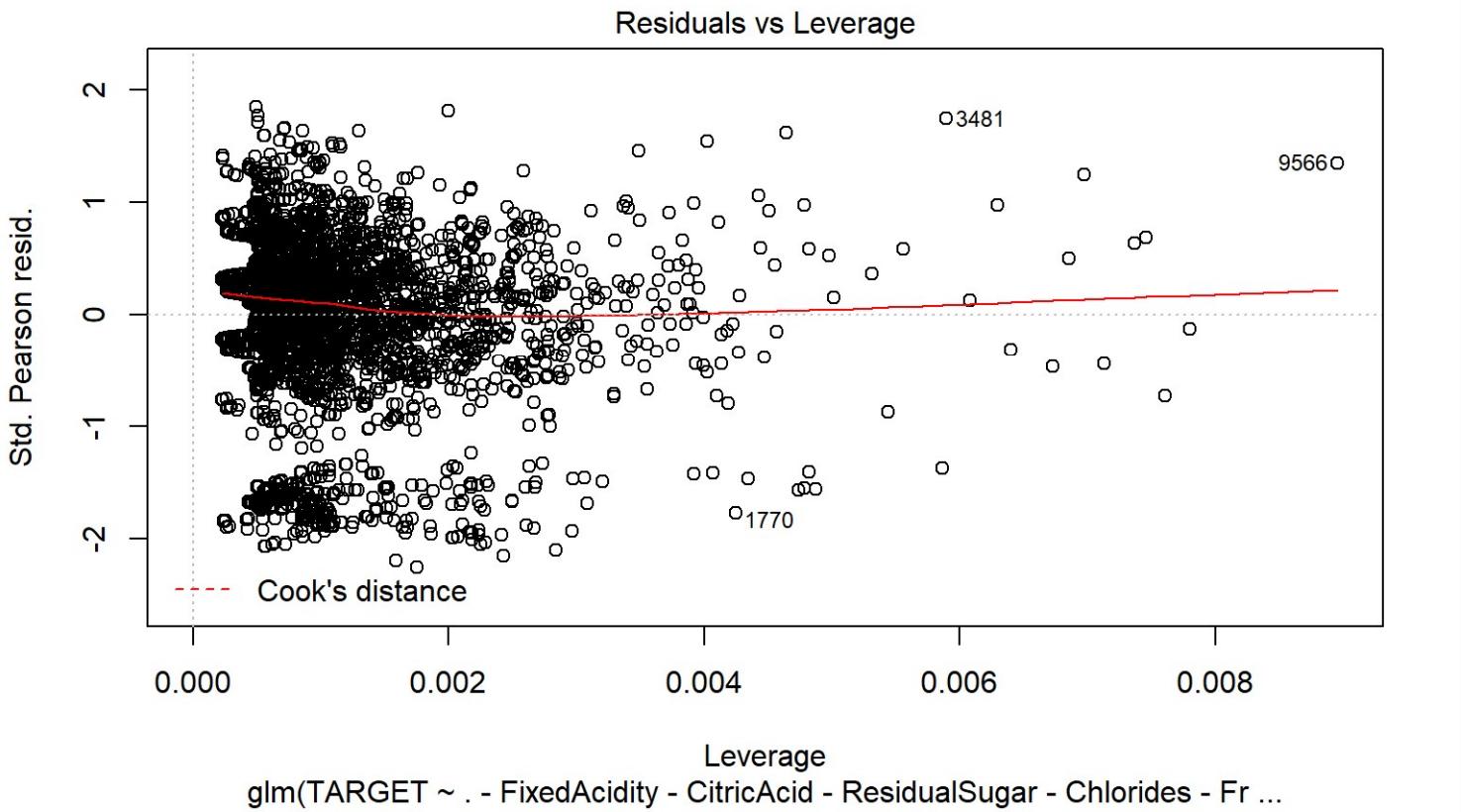




Theoretical Quantiles
glm(TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar - Chlorides - Fr ...)



Predicted values
glm(TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar - Chlorides - Fr ...)



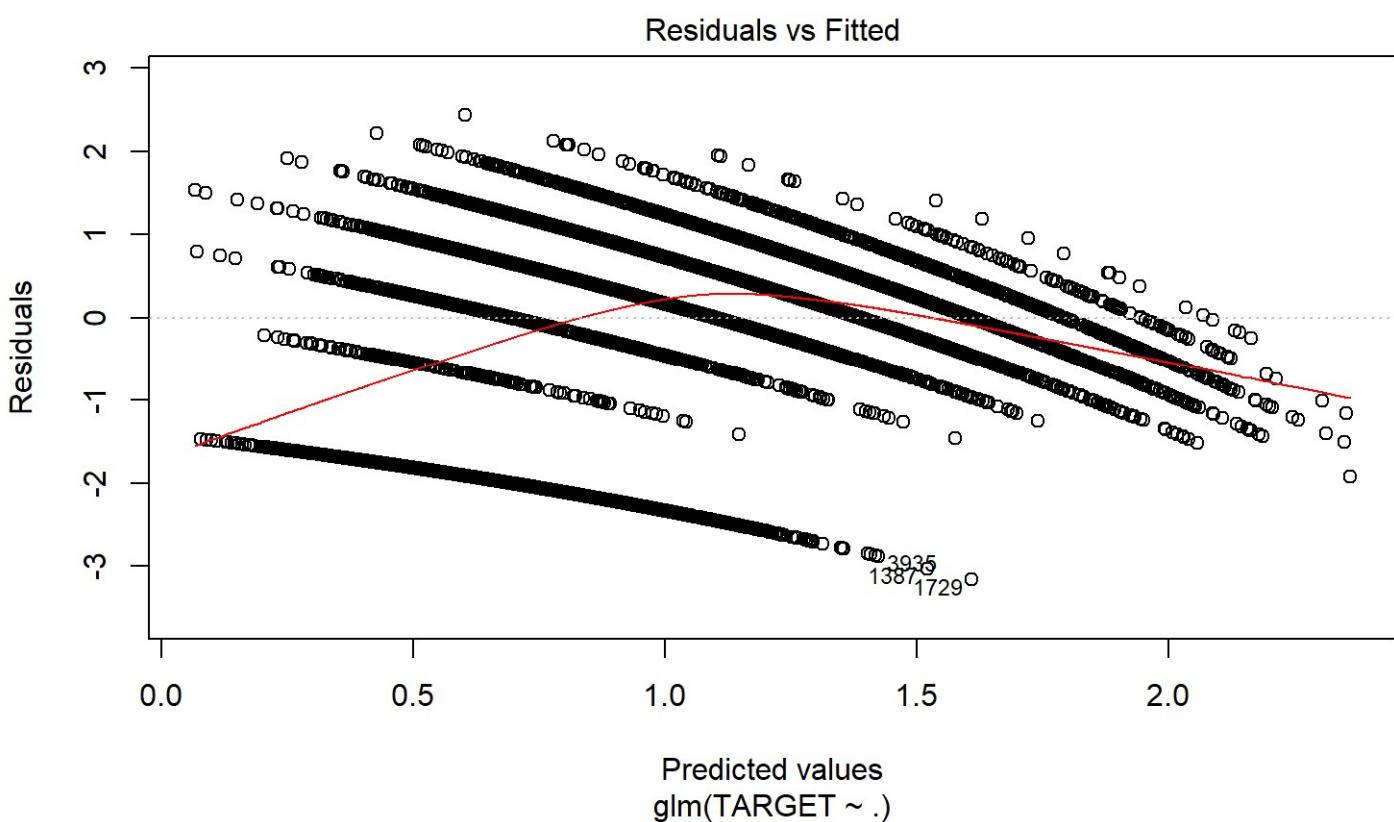
Model 3: Poisson Model with imputations

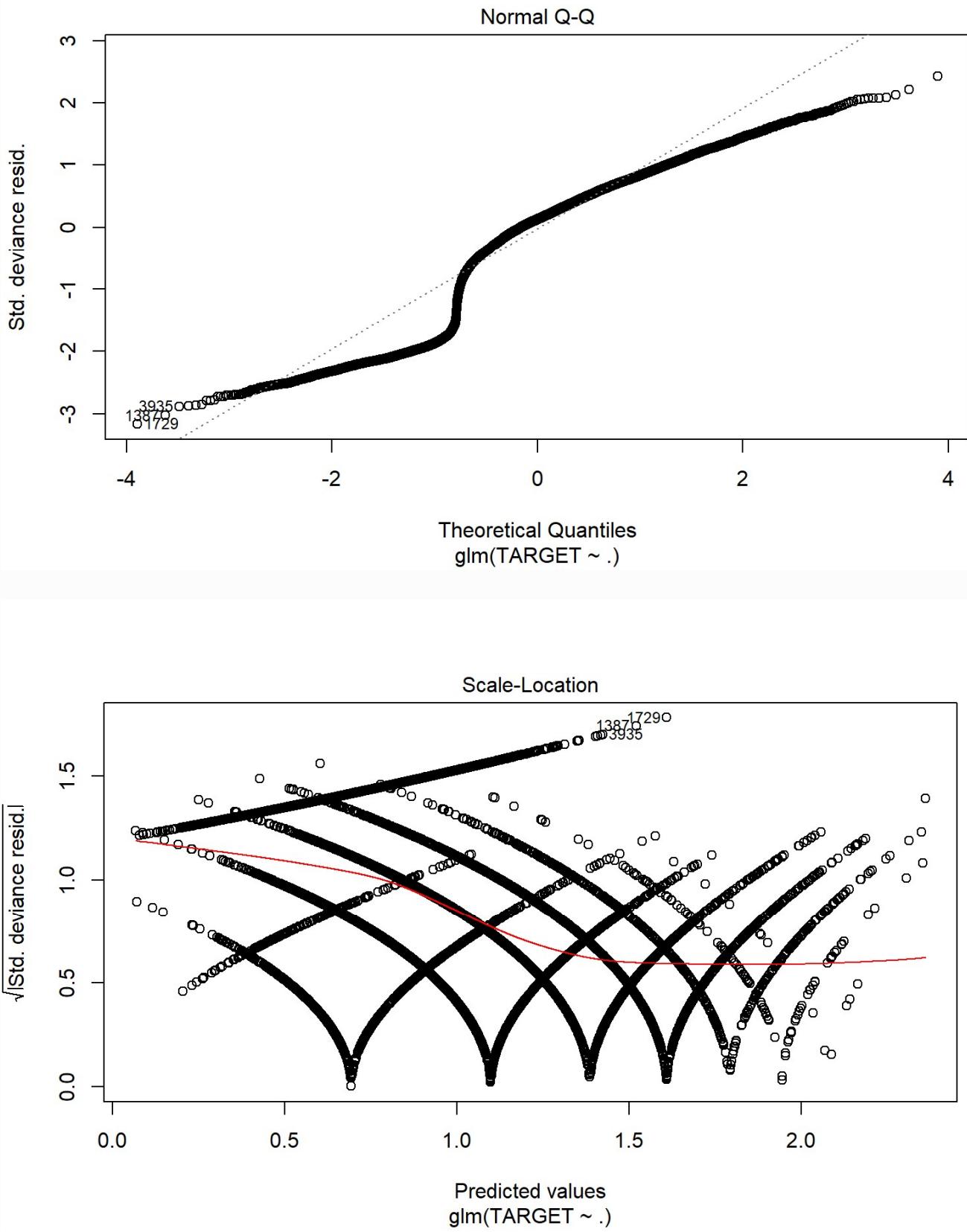
```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = wine_train2)
##
## Deviance Residuals:
##   Min     1Q    Median     3Q    Max 
## -3.1630 -0.6739  0.1305  0.6337  2.4320
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 2.337e+00 2.281e-01 10.242 < 2e-16 ***
## FixedAcidity 2.250e-04 9.190e-04  0.245 0.806610  
## VolatileAcidity -4.313e-02 7.286e-03 -5.919 3.23e-09 ***
## CitricAcid    8.534e-03 6.573e-03  1.298 0.194168  
## ResidualSugar 1.271e-04 1.675e-04  0.759 0.448033  
## Chlorides     -6.572e-02 1.790e-02 -3.673 0.000240 ***
## FreeSulfurDioxide 1.336e-04 3.804e-05  3.512 0.000444 ***
## TotalSulfurDioxide 9.235e-05 2.460e-05  3.754 0.000174 ***
## Density      -3.404e-01 2.144e-01 -1.588 0.112379  
## pH          -1.962e-02 8.417e-03 -2.331 0.019744 *
```

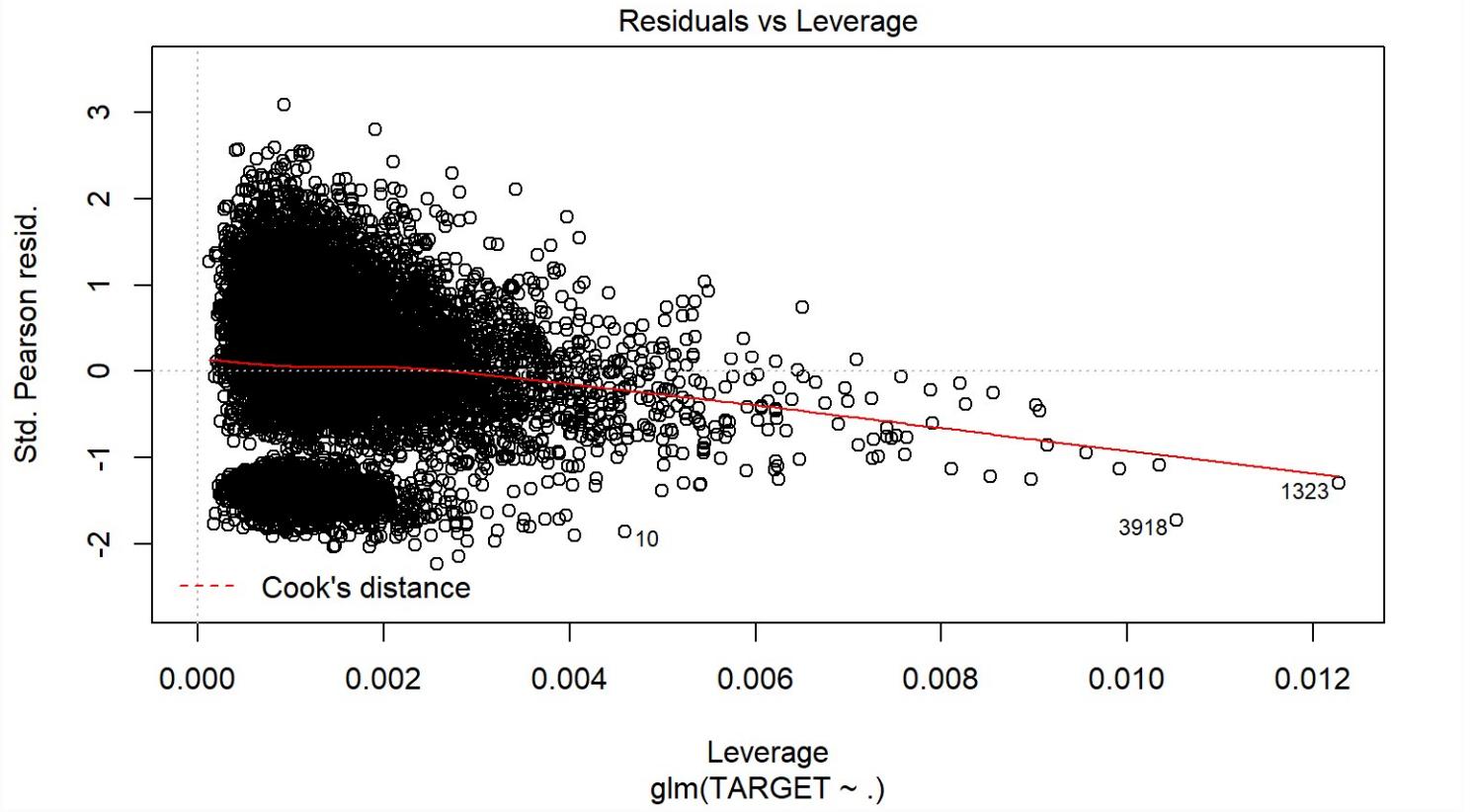
```

## Sulphates -1.569e-02 6.157e-03 -2.549 0.010805 *
## Alcohol 2.951e-03 1.554e-03 1.898 0.057632 .
## LabelAppeal 1.409e-01 6.798e-03 20.724 < 2e-16 ***
## AcidIndex -7.709e-01 3.998e-02 -19.280 < 2e-16 ***
## STARS 3.407e-01 6.270e-03 54.337 < 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: 12829 on 10222 degrees of freedom
## AIC: 38417
##
## Number of Fisher Scoring iterations: 5

```







Model 4: Poisson Model with imputations and only significant variables

```

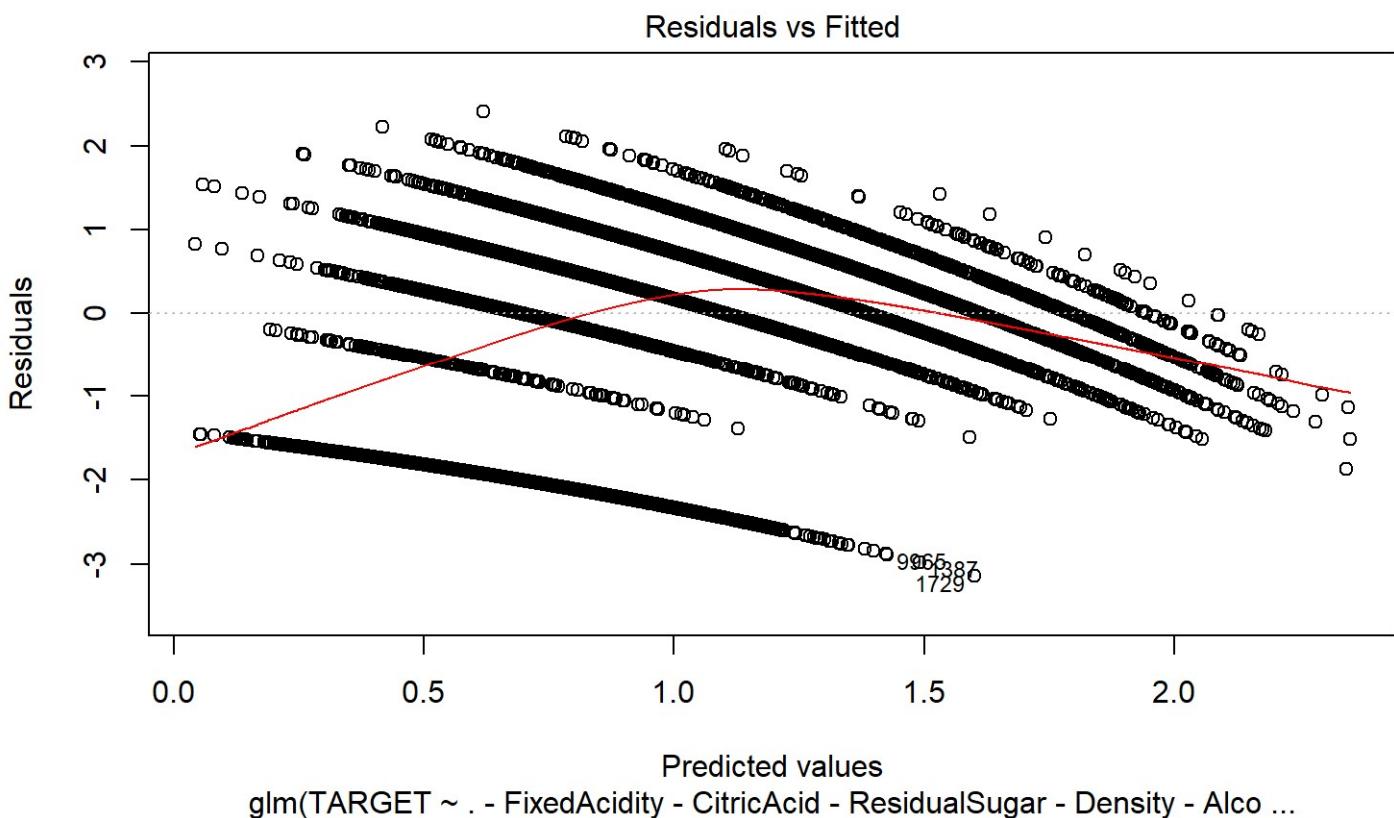
## 
## Call:
## glm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##   Density - Alcohol, family = poisson, data = wine_train2)
## 
## Deviance Residuals:
##   Min     1Q     Median      3Q     Max  
## -3.1469 -0.6828  0.1295  0.6379  2.4054
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 2.038e+00 8.840e-02 23.052 < 2e-16 ***
## VolatileAcidity -4.348e-02 7.284e-03 -5.969 2.39e-09 ***
## Chlorides    -6.725e-02 1.789e-02 -3.760 0.000170 ***
## FreeSulfurDioxide 1.316e-04 3.801e-05 3.461 0.000537 *** 
## TotalSulfurDioxide 9.150e-05 2.458e-05 3.723 0.000197 *** 
## pH          -1.991e-02 8.415e-03 -2.366 0.018003 *  
## Sulphates    -1.563e-02 6.153e-03 -2.540 0.011086 *  
## LabelAppeal   1.409e-01 6.798e-03 20.727 < 2e-16 ***
## AcidIndex    -7.729e-01 3.936e-02 -19.636 < 2e-16 ***

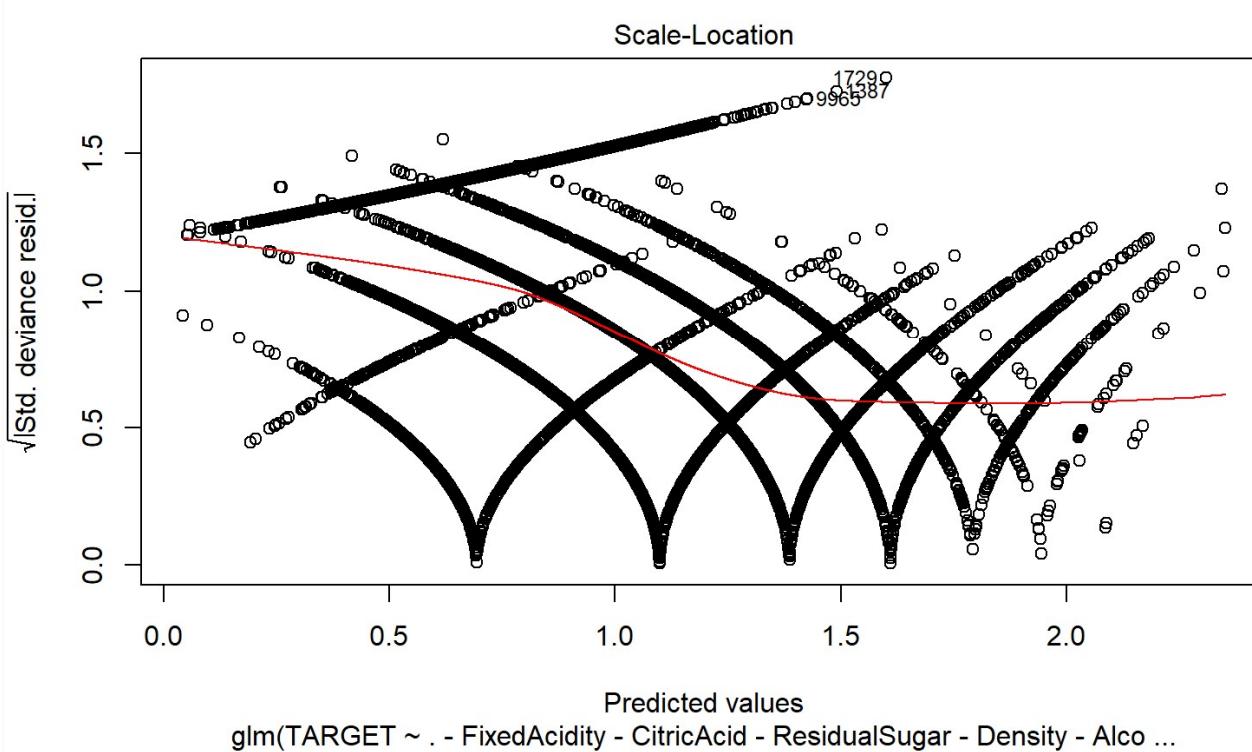
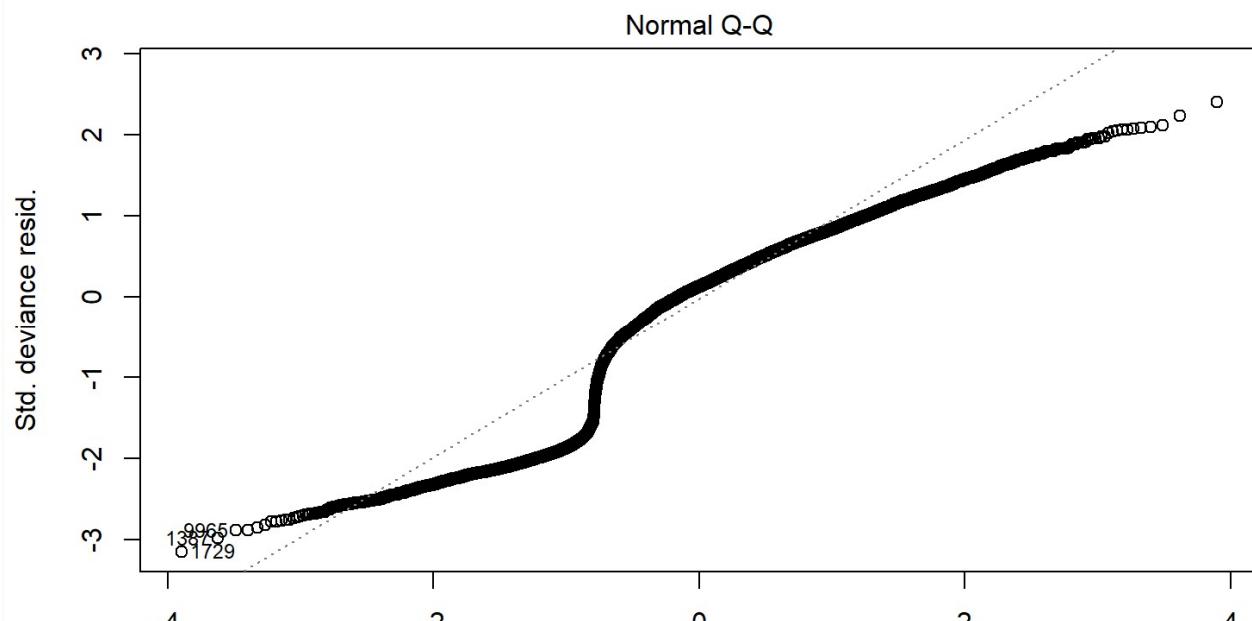
```

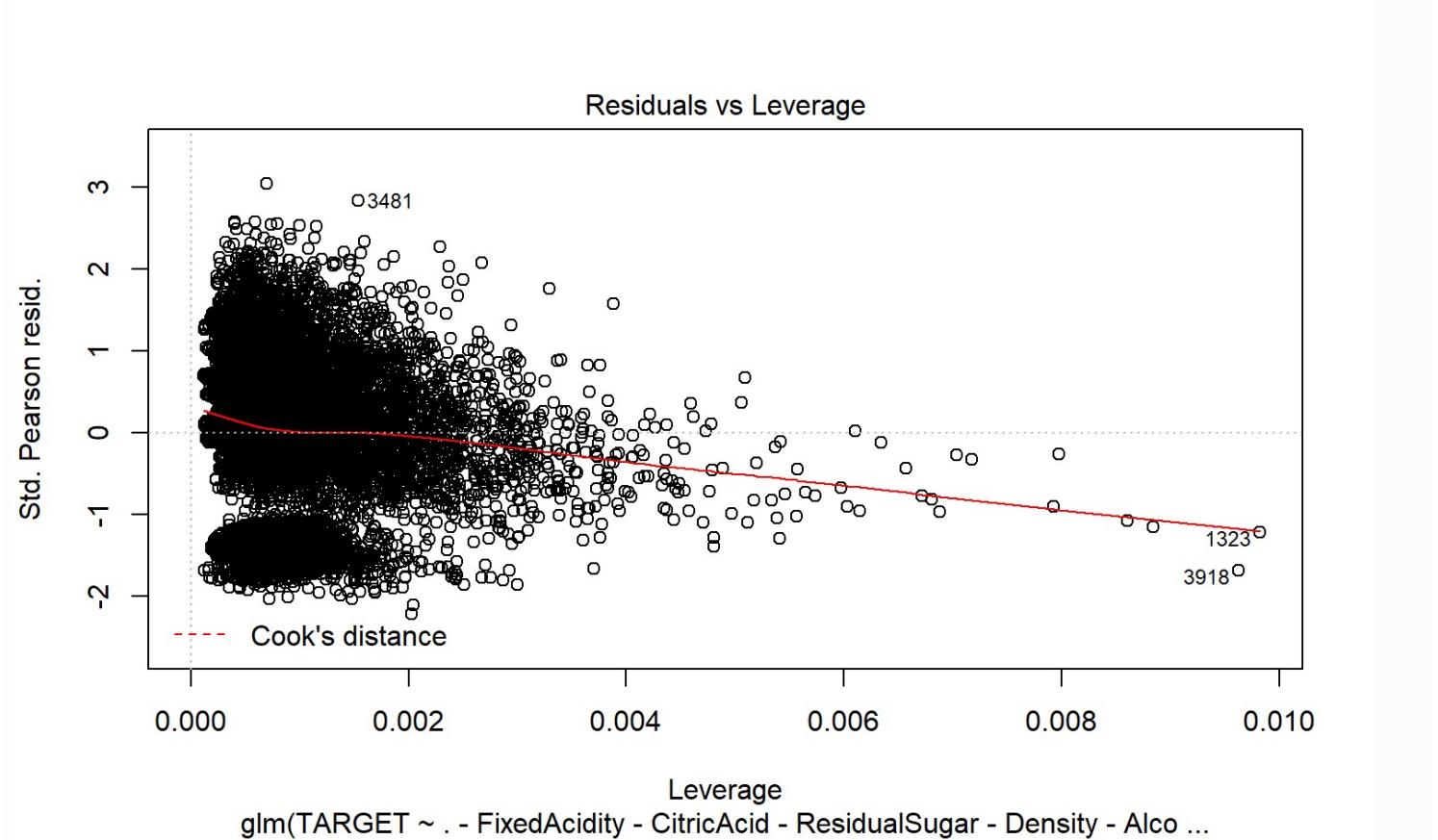
```

## STARS      3.417e-01 6.255e-03 54.634 < 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: 12837 on 10227 degrees of freedom
## AIC: 38415
##
## Number of Fisher Scoring iterations: 5

```







Model II: Negative Binomial

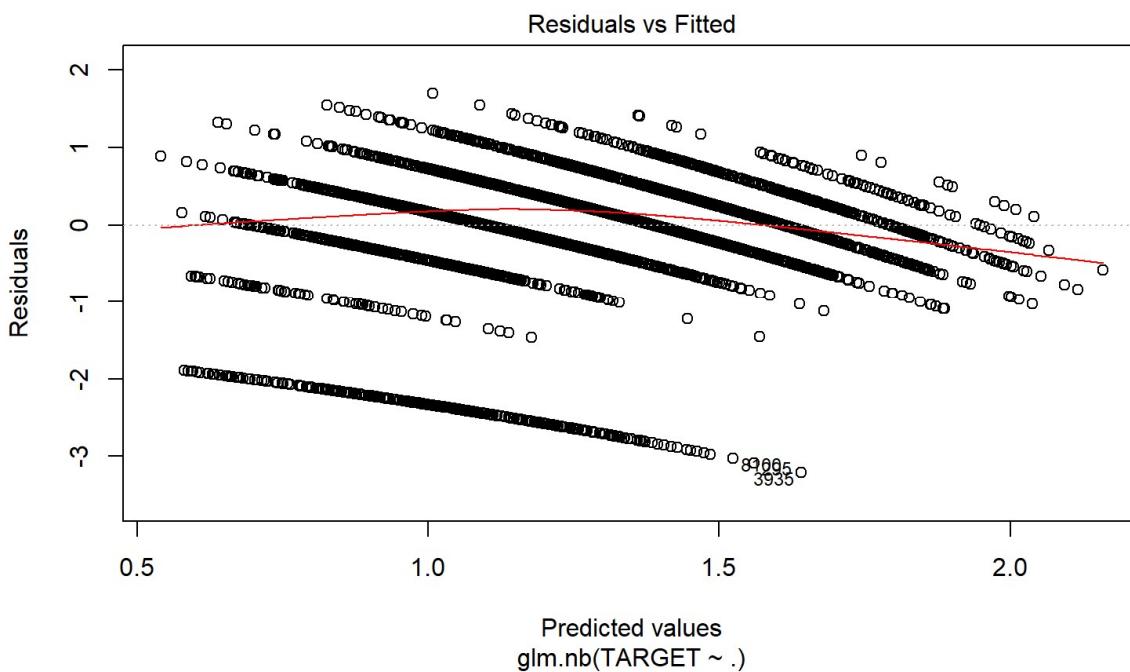
Model 5 : Negative Binomial without imputations

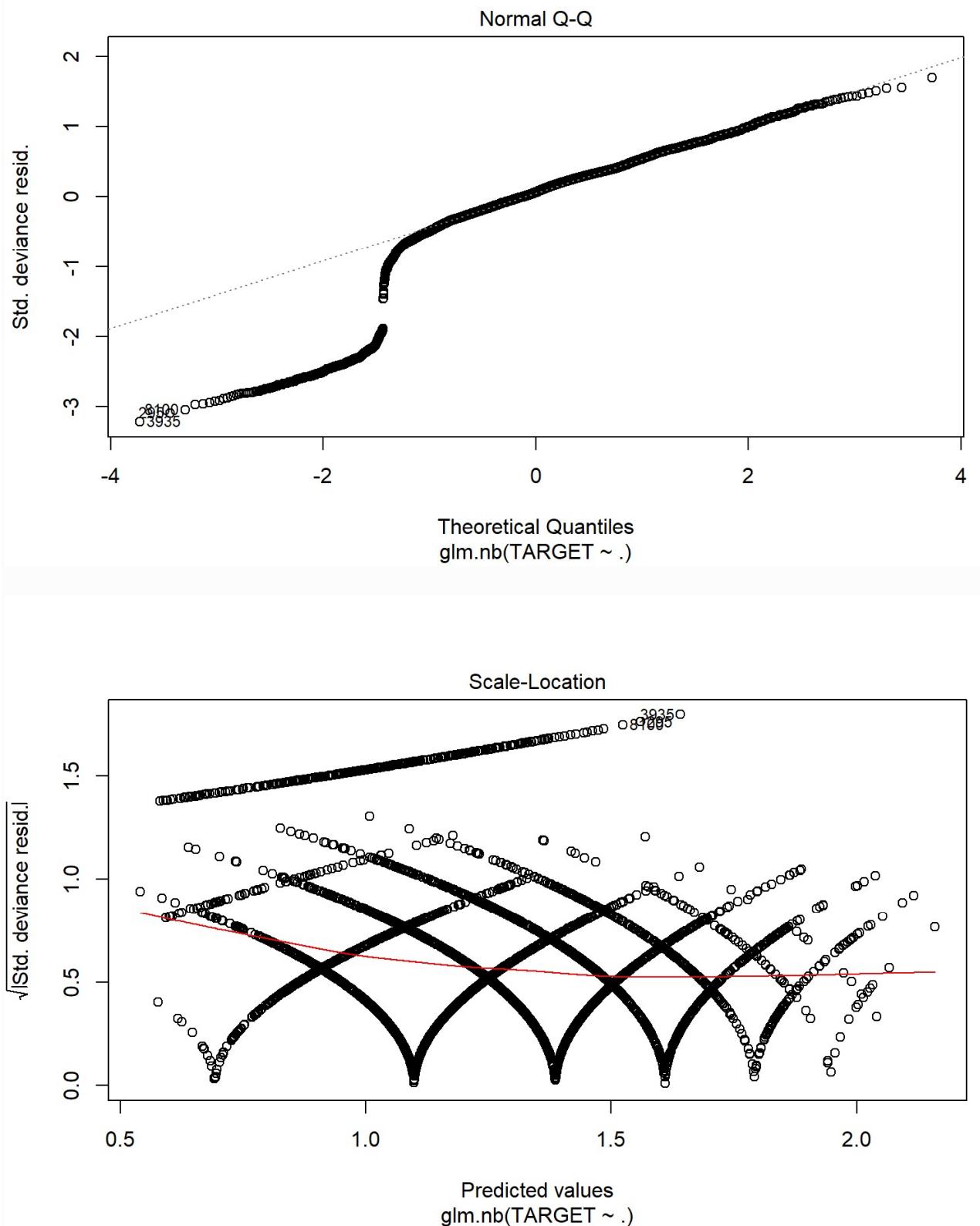
```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = wine_train1, init.theta = 138898.9107,
##   link = log)
##
## Deviance Residuals:
##   Min     1Q     Median      3Q     Max 
## -3.2127 -0.2757  0.0647  0.3766  1.6981 
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (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
```

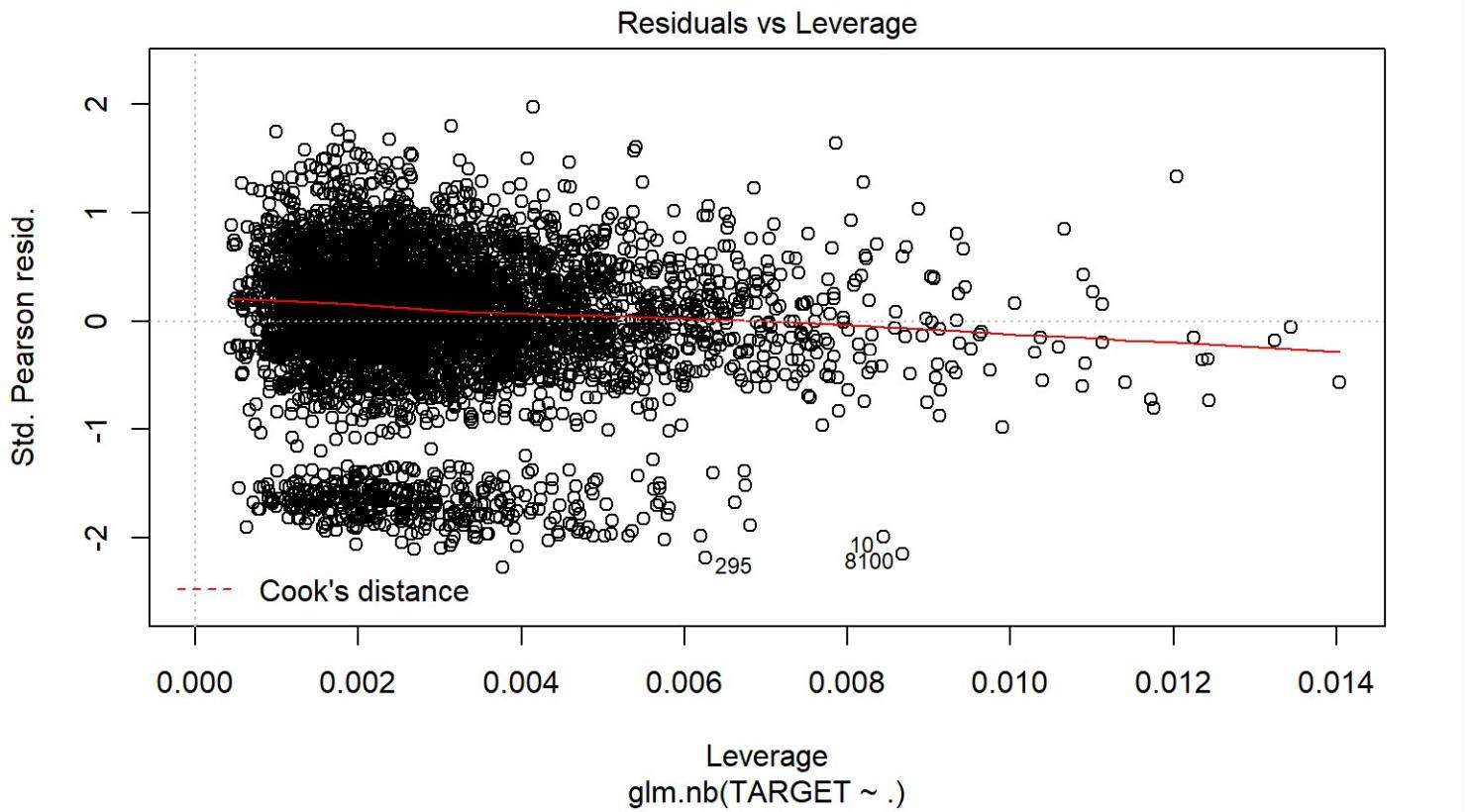
```

## 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

```







Model 6 : Negative Binomial without imputations and only significant variables

```

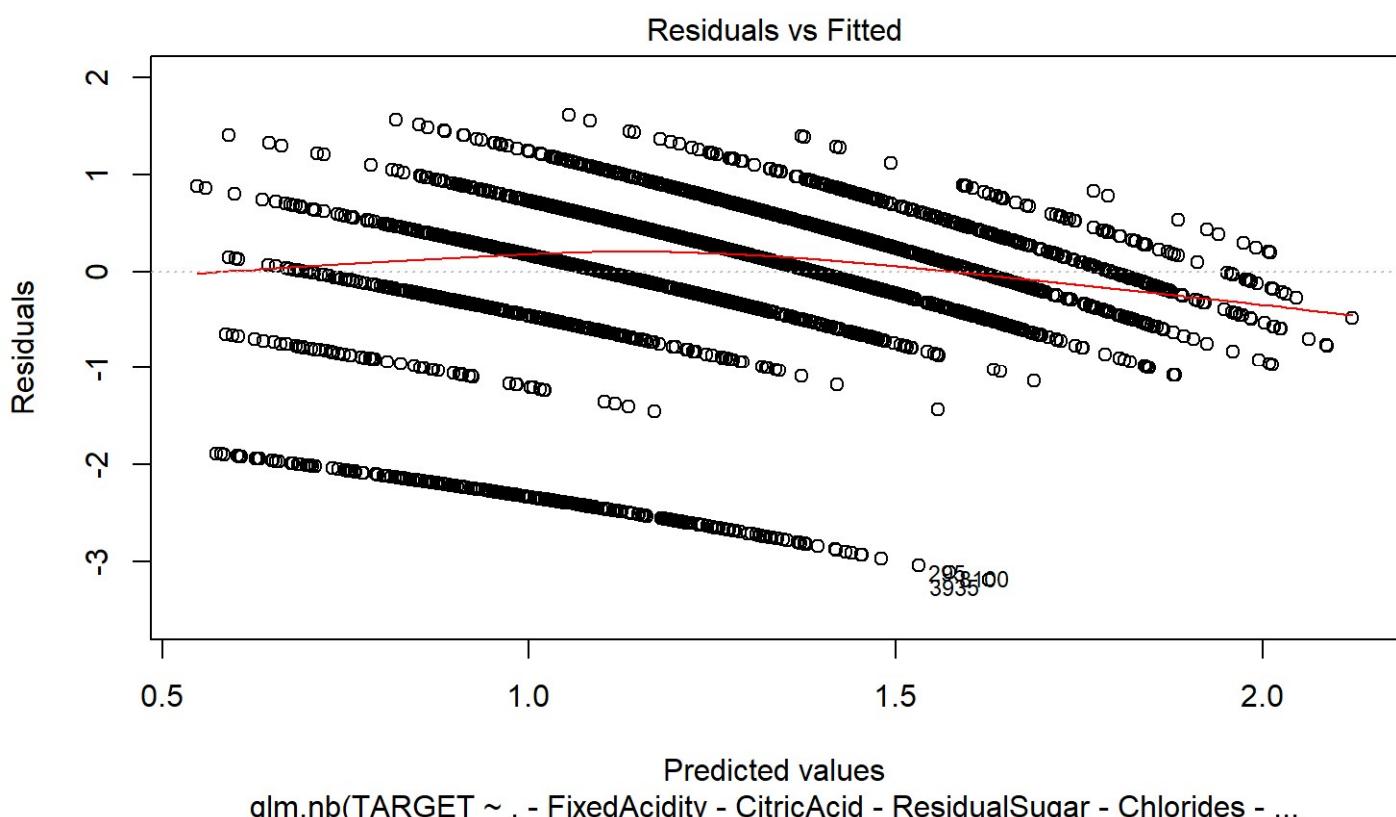
## 
## Call:
## glm.nb(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##   Chlorides - FreeSulfurDioxide - TotalSulfurDioxide - Density -
##   pH - Sulphates - Alcohol, data = wine_train1, init.theta = 138402.5261,
##   link = log)
## 
## Deviance Residuals:
##   Min     1Q     Median      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

```

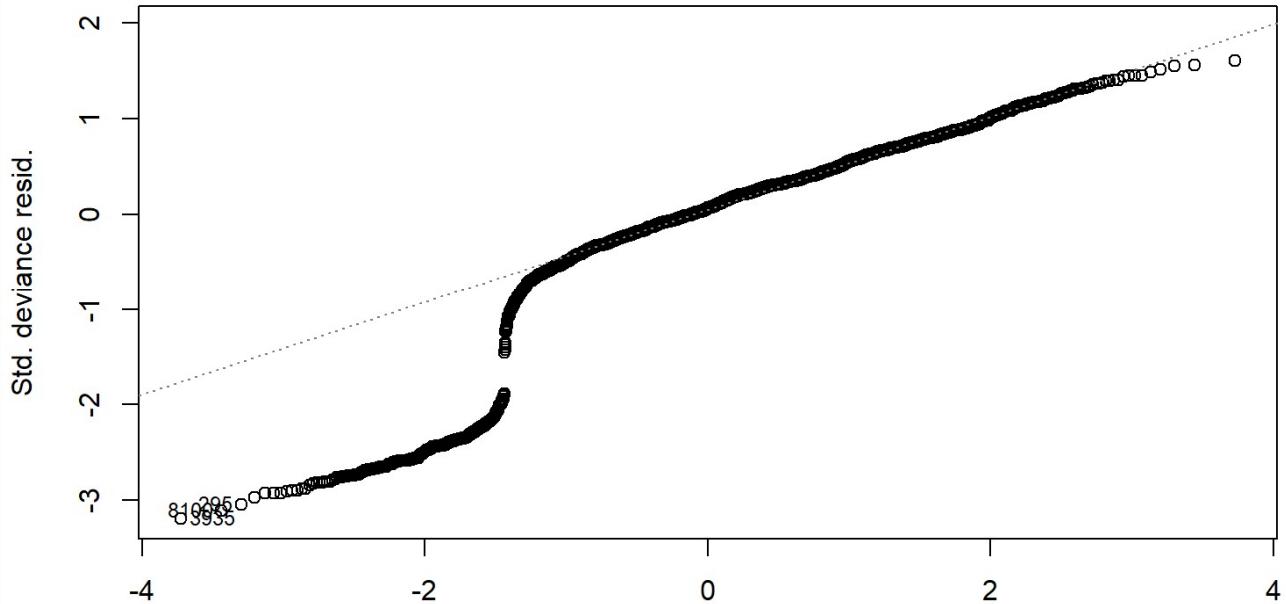
```

## 
## (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

```

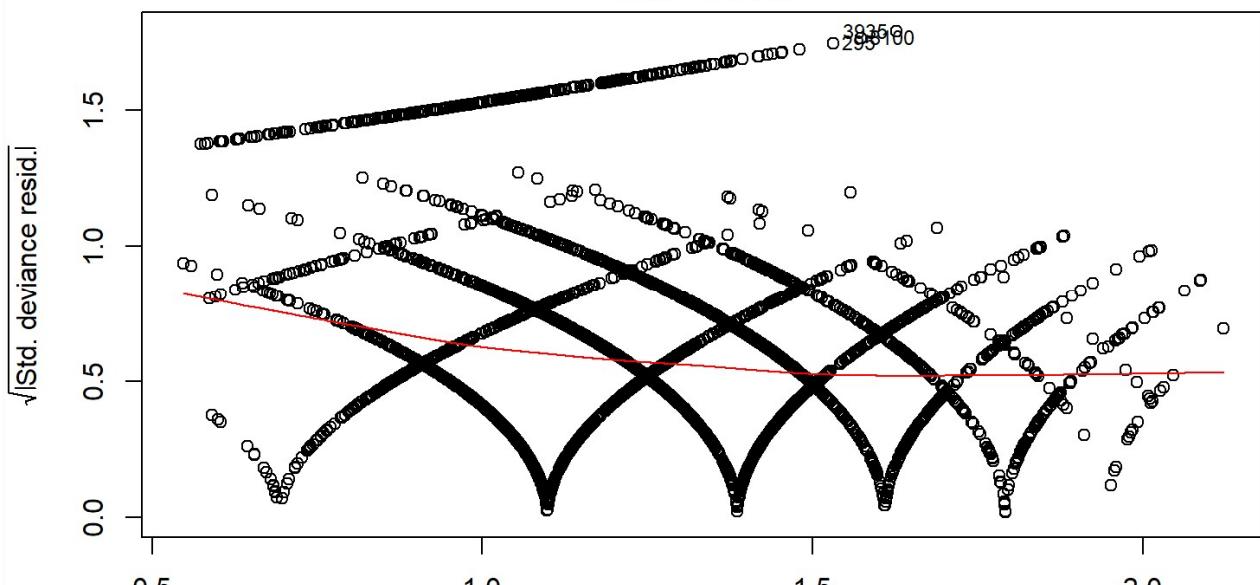


Normal Q-Q

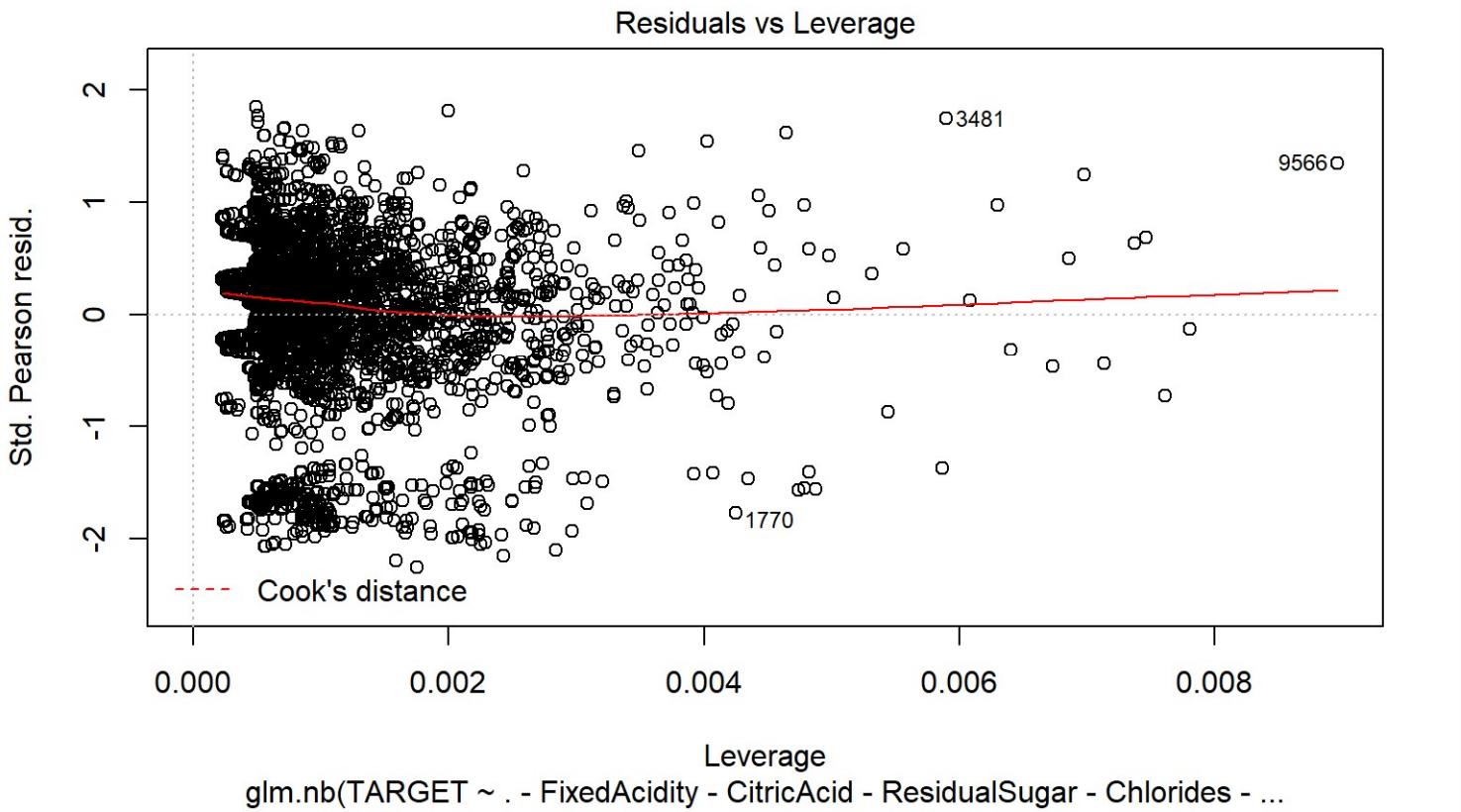


Theoretical Quantiles
glm.nb(TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar - Chlorides - ...)

Scale-Location



Predicted values
glm.nb(TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar - Chlorides - ...)



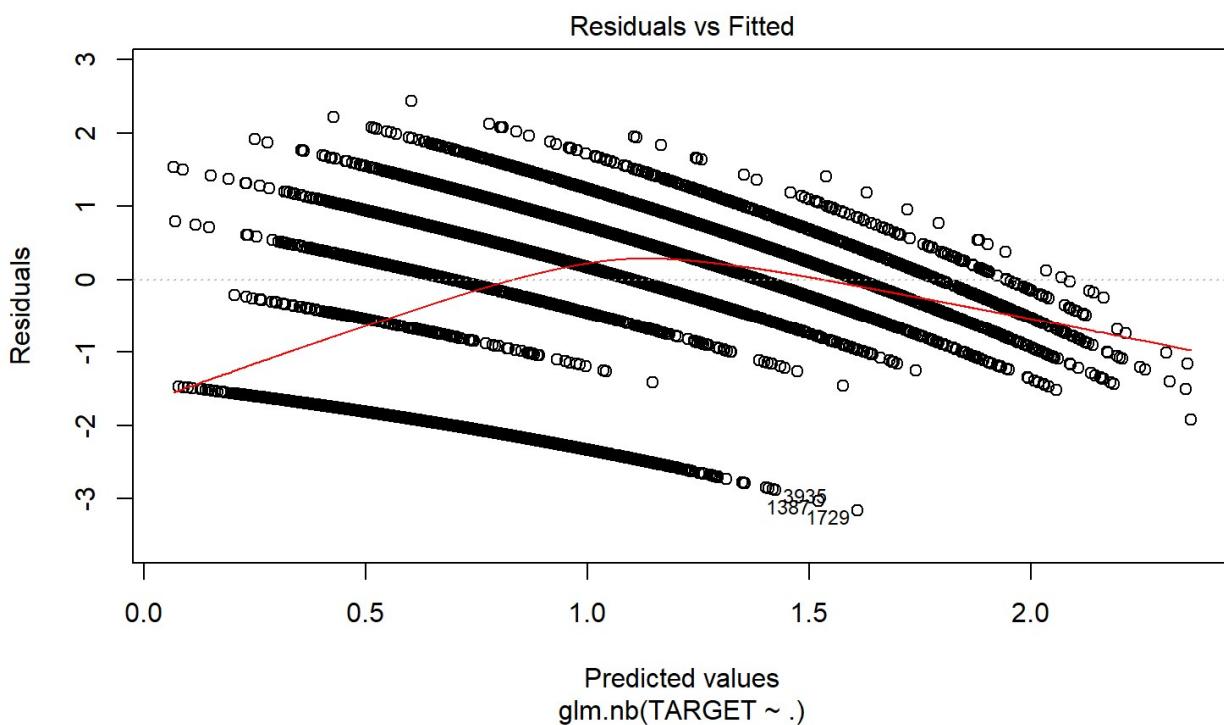
Model 7 : Negative Binomial with imputations

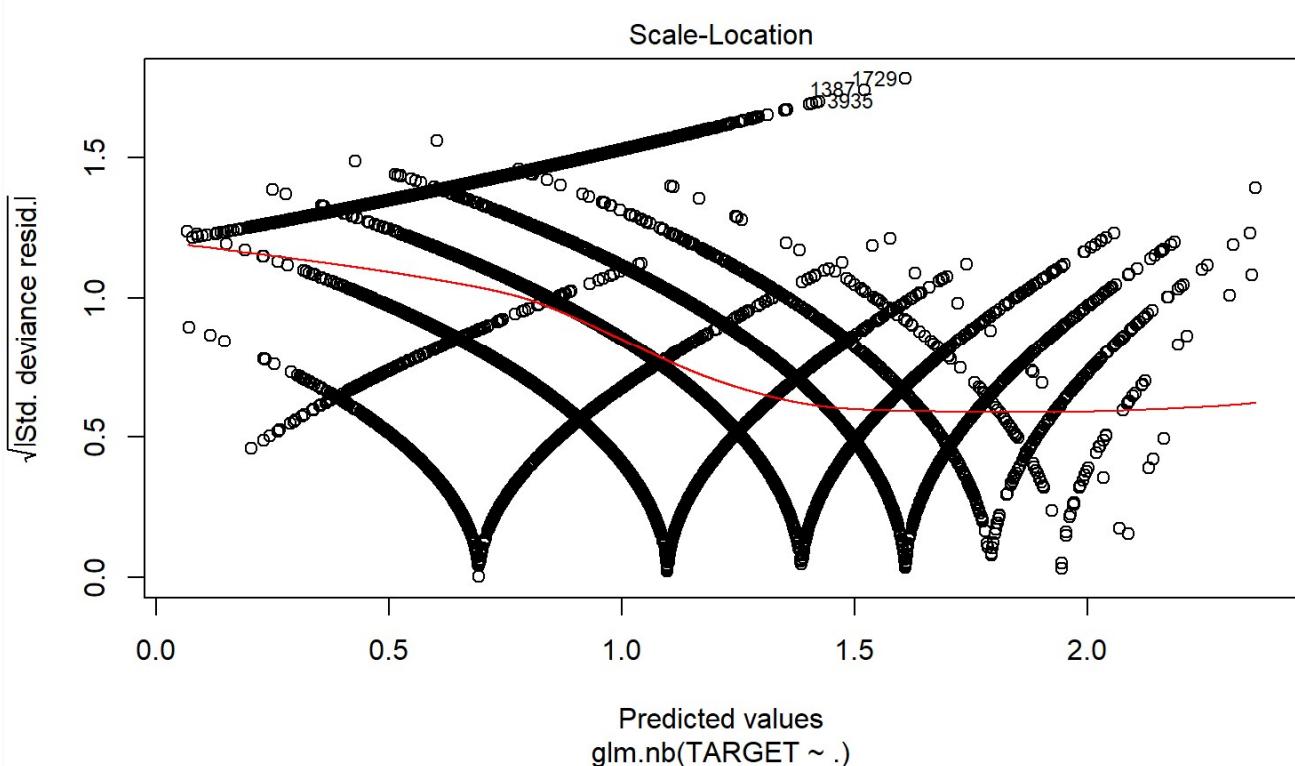
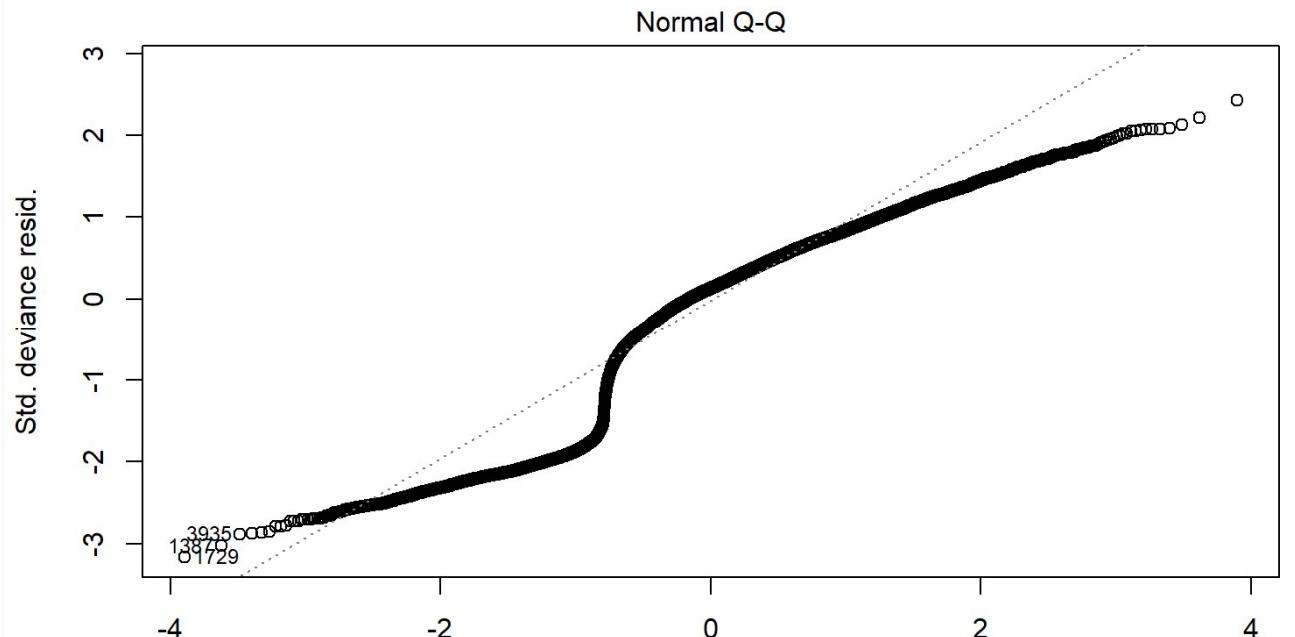
```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = wine_train2, init.theta = 49078.50992,
##   link = log)
##
## Deviance Residuals:
##      Min      1Q      Median      3Q      Max 
## -3.1629 -0.6739  0.1305  0.6337  2.4320 
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 2.337e+00 2.281e-01 10.242 < 2e-16 ***
## FixedAcidity 2.250e-04 9.190e-04 0.245 0.806608  
## VolatileAcidity -4.313e-02 7.286e-03 -5.919 3.23e-09 ***
## CitricAcid    8.534e-03 6.573e-03 1.298 0.194177  
## ResidualSugar 1.271e-04 1.675e-04 0.759 0.448021  
## Chlorides     -6.573e-02 1.790e-02 -3.673 0.000240 *** 
## FreeSulfurDioxide 1.336e-04 3.804e-05 3.512 0.000444 *** 
## TotalSulfurDioxide 9.235e-05 2.460e-05 3.754 0.000174 *** 
## Density      -3.404e-01 2.144e-01 -1.588 0.112389
```

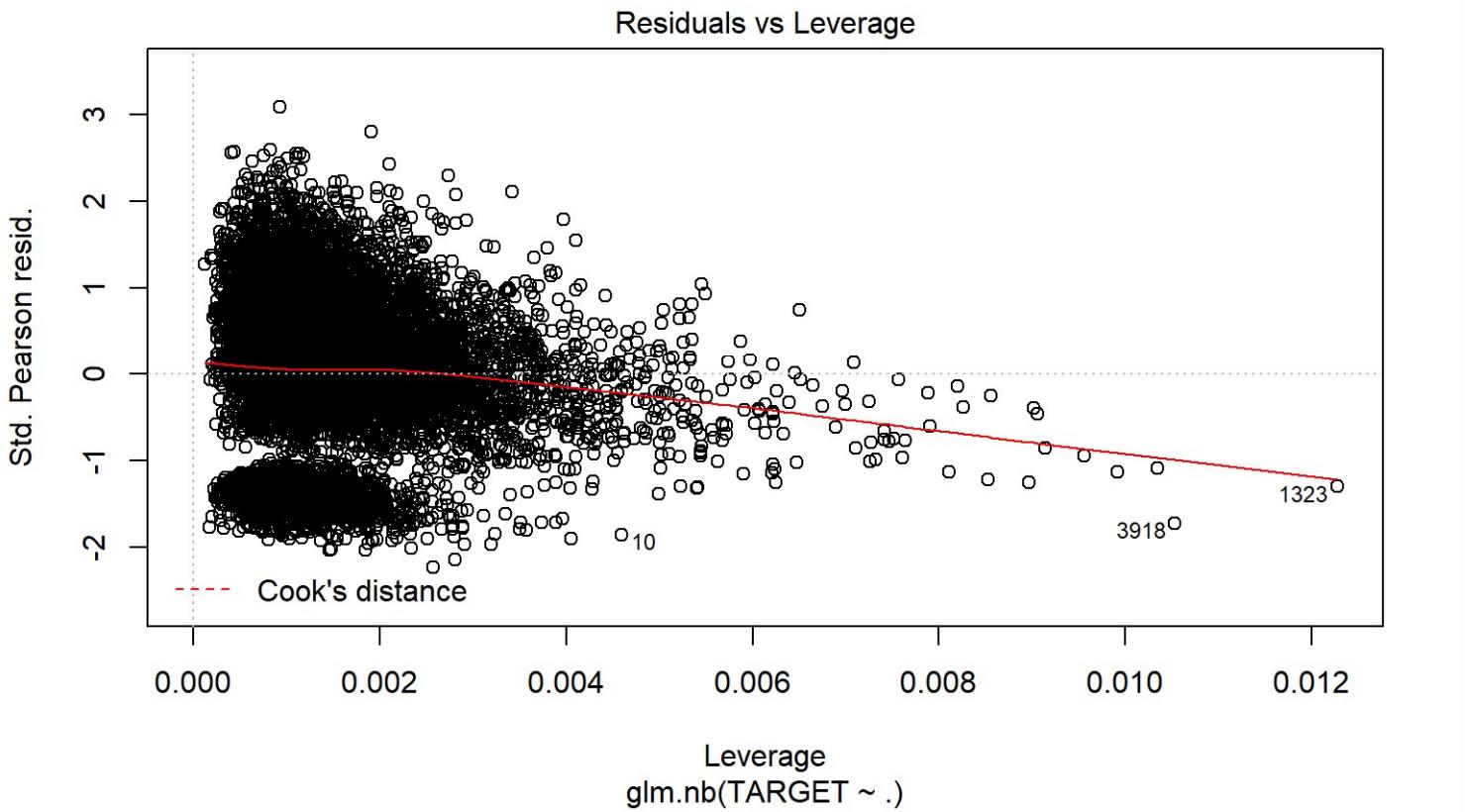
```

## pH      -1.962e-02 8.418e-03 -2.331 0.019745 *
## Sulphates -1.569e-02 6.157e-03 -2.549 0.010806 *
## Alcohol   2.951e-03 1.554e-03  1.898 0.057642 .
## LabelAppeal 1.409e-01 6.798e-03 20.723 < 2e-16 ***
## AcidIndex  -7.709e-01 3.999e-02 -19.279 < 2e-16 ***
## STARS     3.407e-01 6.270e-03 54.335 < 2e-16 ***
## ---
## Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
##
## (Dispersion parameter for Negative Binomial(49078.51) family taken to be 1)
##
## Null deviance: 18290 on 10236 degrees of freedom
## Residual deviance: 12828 on 10222 degrees of freedom
## AIC: 38419
##
## Number of Fisher Scoring iterations: 1
##
##
## Theta: 49079
## Std. Err.: 63619
## Warning while fitting theta: iteration limit reached
##
## 2 x log-likelihood: -38387.04

```







Model 8 : Negative Binomial with imputations and only significant variables

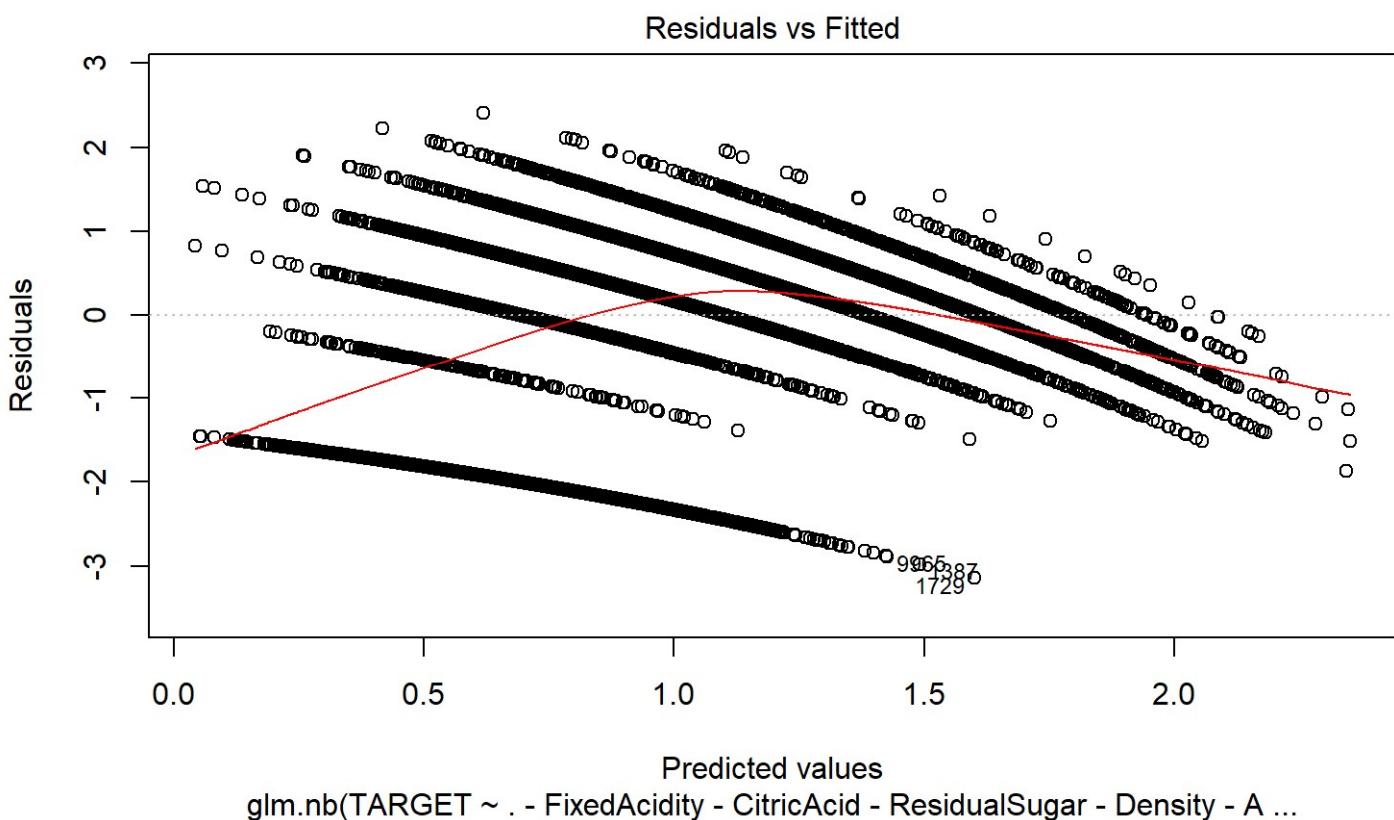
```

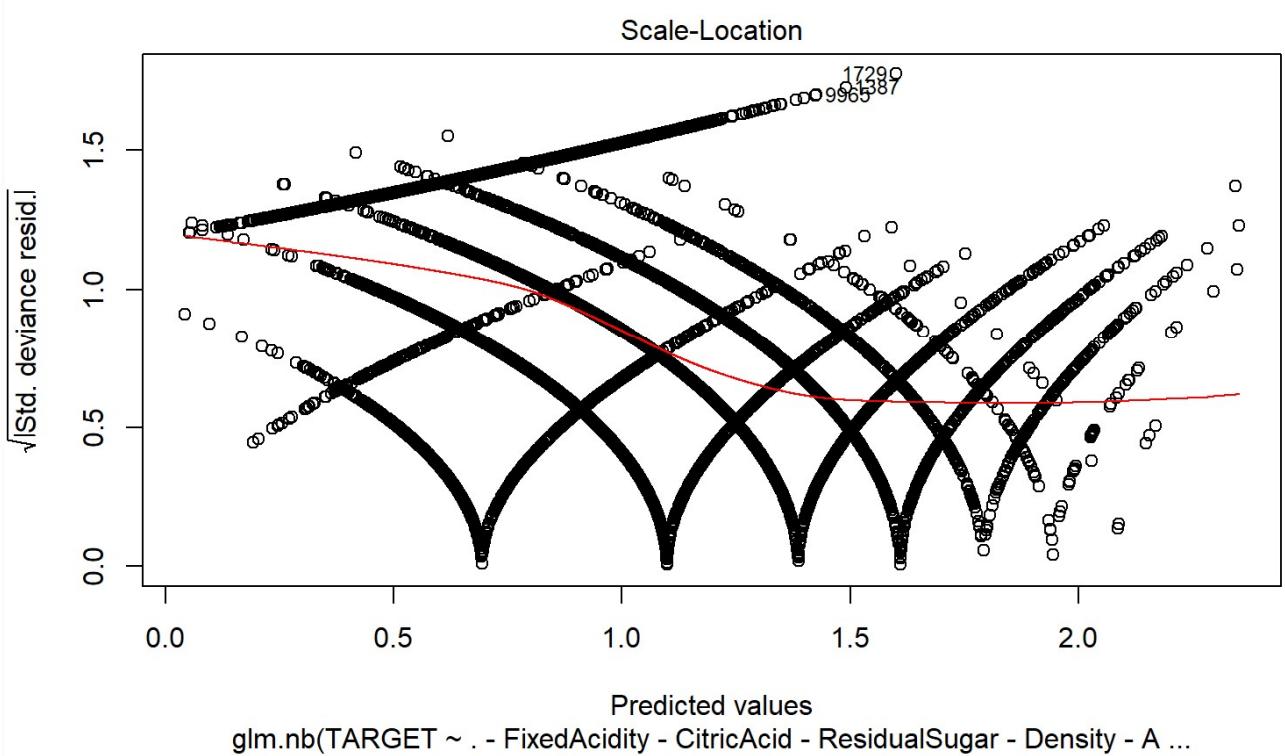
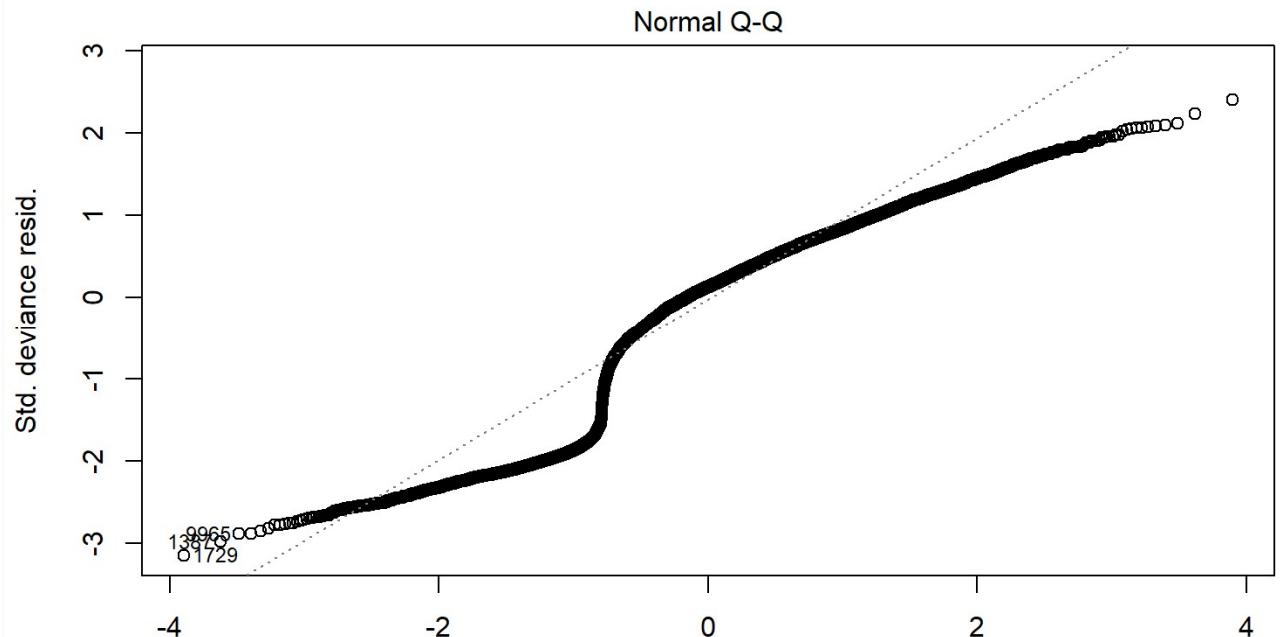
## 
## Call:
## glm.nb(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar -
##         Density - Alcohol, data = wine_train2, init.theta = 48992.35936,
##         link = log)
## 
## Deviance Residuals:
##   Min     1Q     Median      3Q     Max 
## -3.1469 -0.6828  0.1295  0.6379  2.4053 
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 2.038e+00 8.840e-02 23.052 < 2e-16 ***
## VolatileAcidity -4.348e-02 7.284e-03 -5.969 2.39e-09 ***
## Chlorides    -6.726e-02 1.789e-02 -3.760 0.000170 *** 
## FreeSulfurDioxide 1.316e-04 3.801e-05 3.461 0.000537 *** 
## TotalSulfurDioxide 9.150e-05 2.458e-05 3.723 0.000197 *** 
## pH          -1.991e-02 8.415e-03 -2.366 0.018004 *  
## Sulphates    -1.563e-02 6.153e-03 -2.540 0.011087 *  
## LabelAppeal   1.409e-01 6.798e-03 20.726 < 2e-16 *** 
## AcidIndex    -7.730e-01 3.936e-02 -19.636 < 2e-16 *** 
## 
```

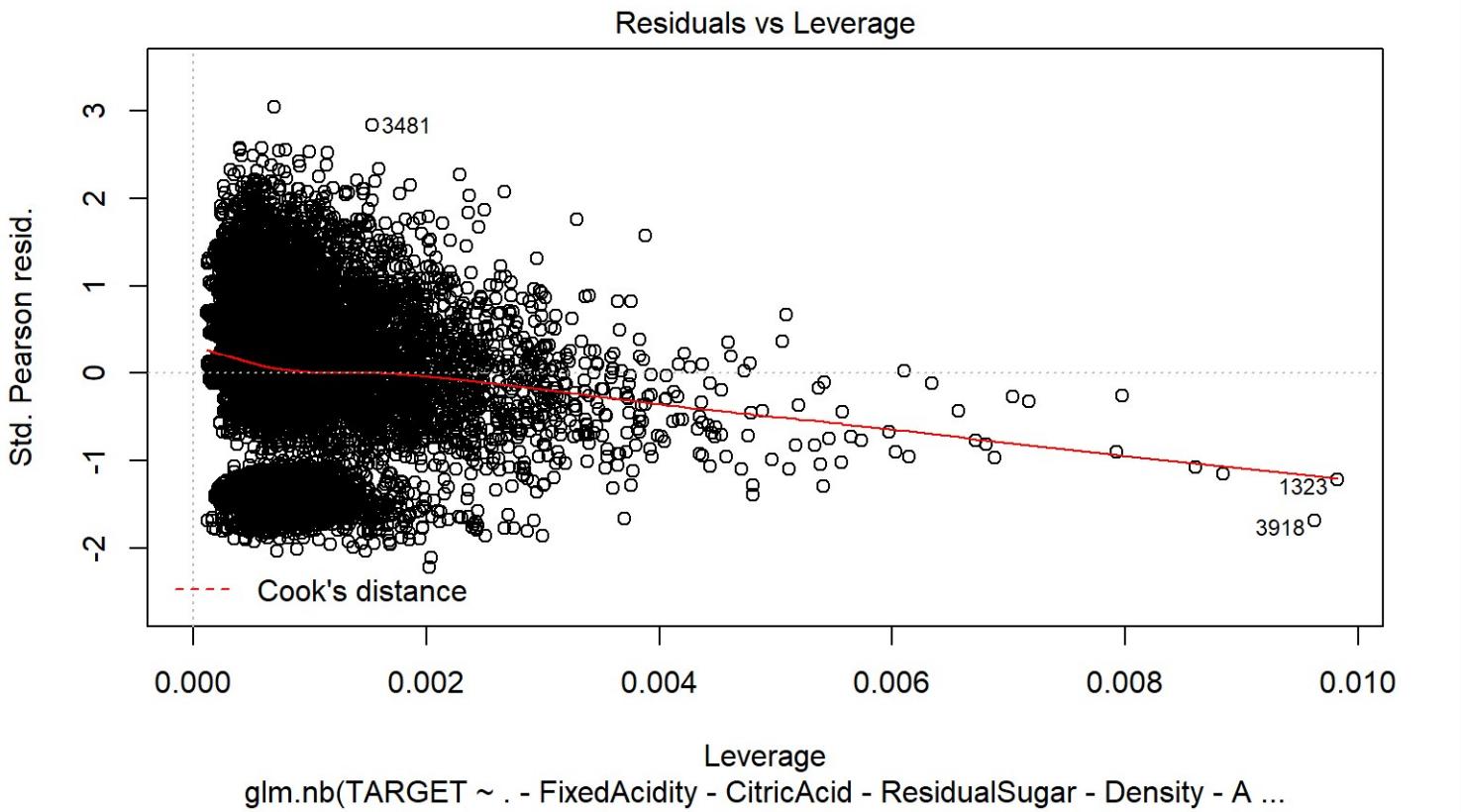
```

## STARS      3.417e-01 6.255e-03 54.632 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
##
## (Dispersion parameter for Negative Binomial(48992.36) family taken to be 1)
##
## Null deviance: 18290 on 10236 degrees of freedom
## Residual deviance: 12837 on 10227 degrees of freedom
## AIC: 38418
##
## Number of Fisher Scoring iterations: 1
##
##
## Theta: 48992
## Std. Err.: 63531
## Warning while fitting theta: iteration limit reached
##
## 2 x log-likelihood: -38395.56

```







Model III: Linear Model

Model 9 : Linear Model with imputations

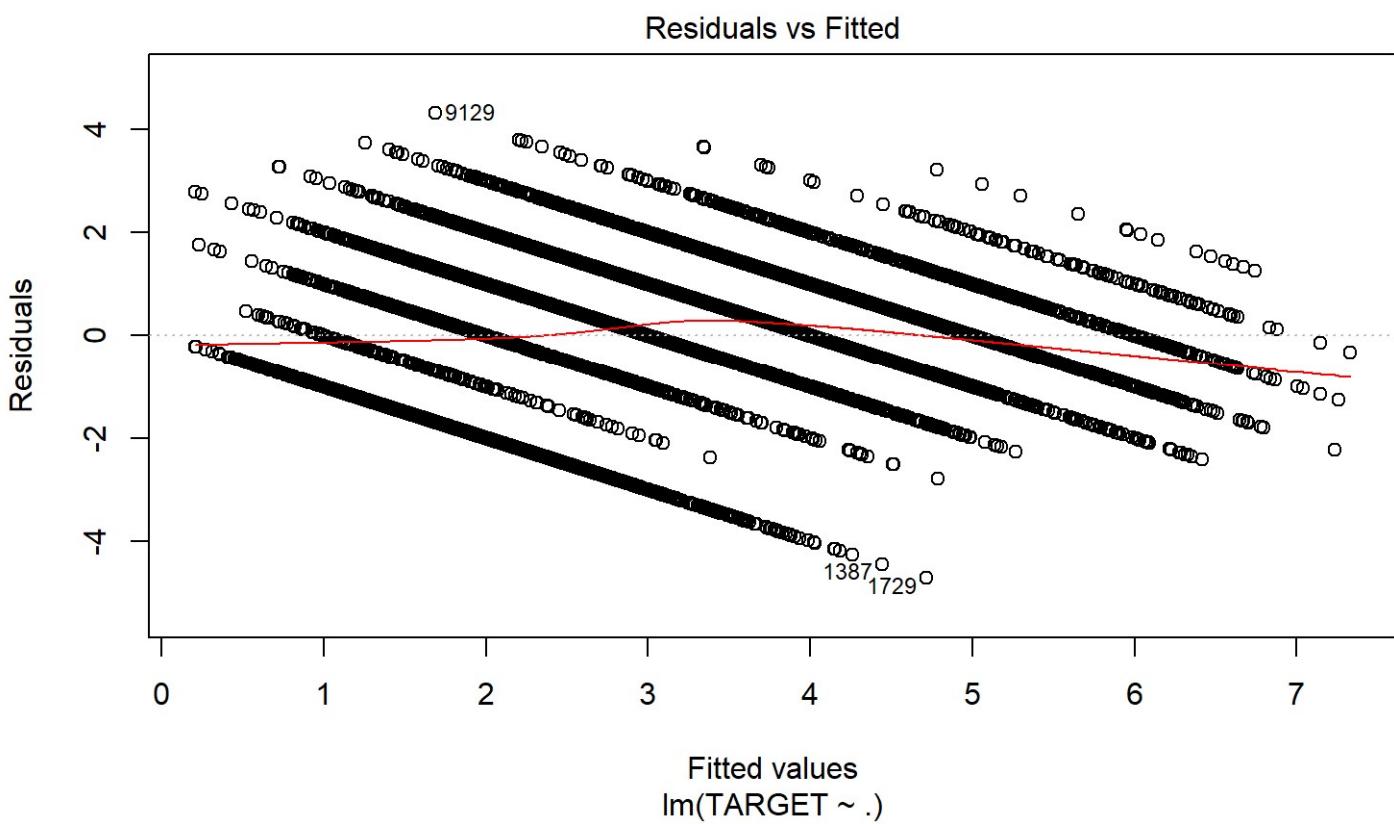
Using Linear Regression Model on imputed training data.

```
##
## Call:
## lm(formula = TARGET ~ ., data = wine_train2)
##
## Residuals:
##   Min   1Q   Median   3Q   Max
## -4.7147 -1.0144  0.1737  1.0276  4.3109
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.968e+00 5.567e-01 10.719 < 2e-16 ***
## FixedAcidity 1.297e-03 2.253e-03  0.576 0.564897
## VolatileAcidity -1.269e-01 1.791e-02 -7.085 1.48e-12 ***
## CitricAcid    2.625e-02 1.629e-02  1.611 0.107133
## ResidualSugar 4.231e-04 4.132e-04  1.024 0.305939
##
```

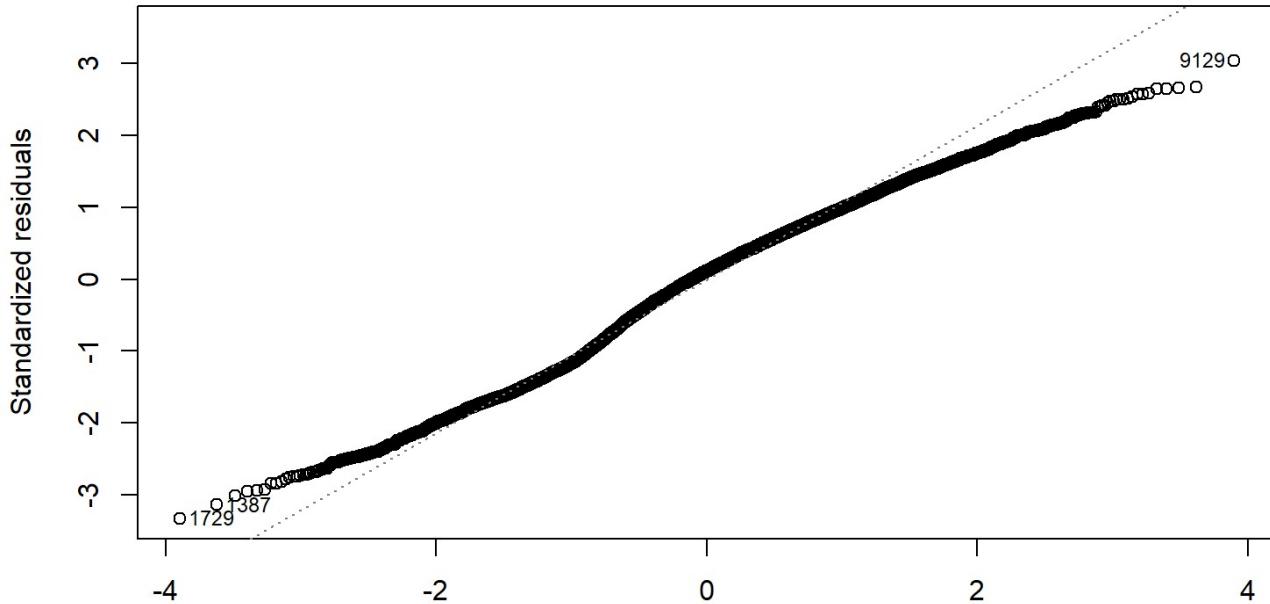
```

## Chlorides      -2.023e-01 4.391e-02 -4.606 4.15e-06 ***
## FreeSulfurDioxide 3.635e-04 9.387e-05 3.873 0.000108 ***
## TotalSulfurDioxide 2.432e-04 6.023e-05 4.038 5.42e-05 ***
## Density       -8.659e-01 5.260e-01 -1.646 0.099740 .
## pH            -4.730e-02 2.072e-02 -2.283 0.022424 *
## Sulphates     -4.212e-02 1.512e-02 -2.786 0.005346 **
## Alcohol        1.251e-02 3.807e-03 3.285 0.001025 **
## LabelAppeal    4.311e-01 1.646e-02 26.191 < 2e-16 ***
## AcidIndex     -2.068e+00 9.237e-02 -22.392 < 2e-16 ***
## STARS         1.167e+00 1.671e-02 69.805 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
##
## Residual standard error: 1.416 on 10222 degrees of freedom
## Multiple R-squared: 0.4605, Adjusted R-squared: 0.4598
## F-statistic: 623.3 on 14 and 10222 DF, p-value: < 2.2e-16

```

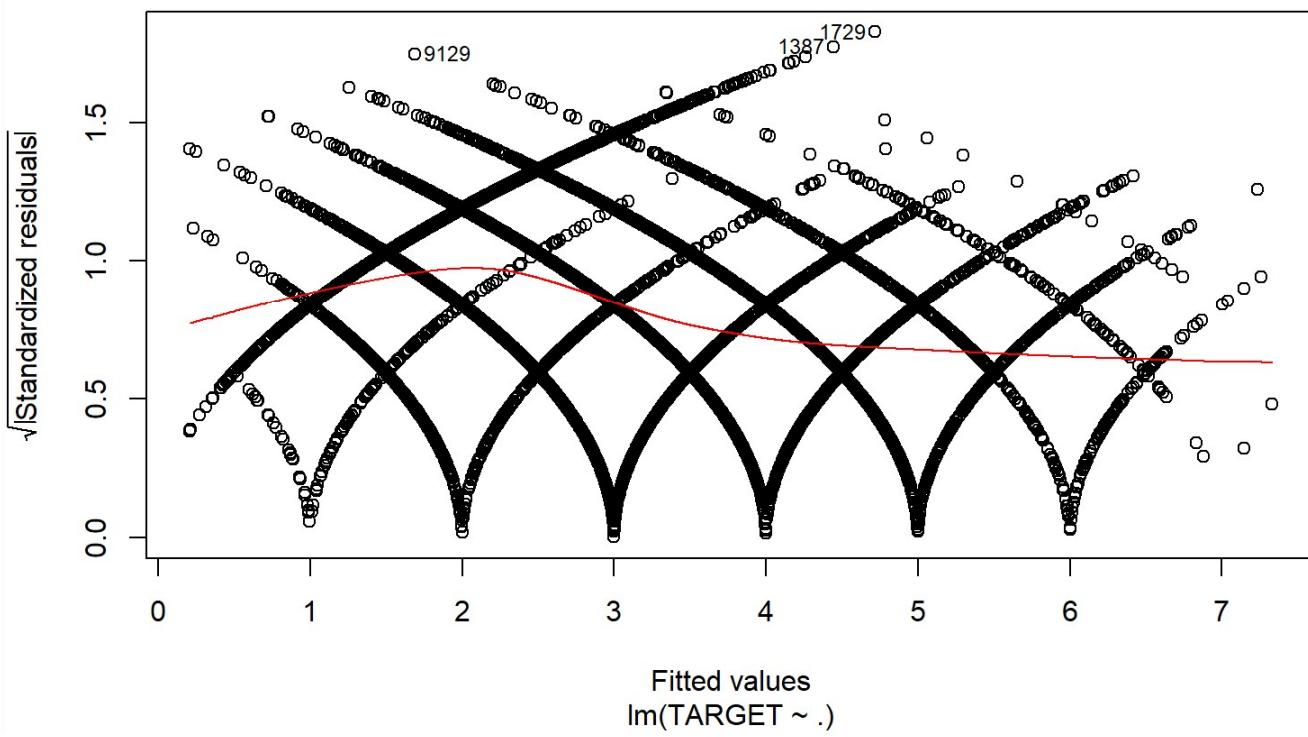


Normal Q-Q

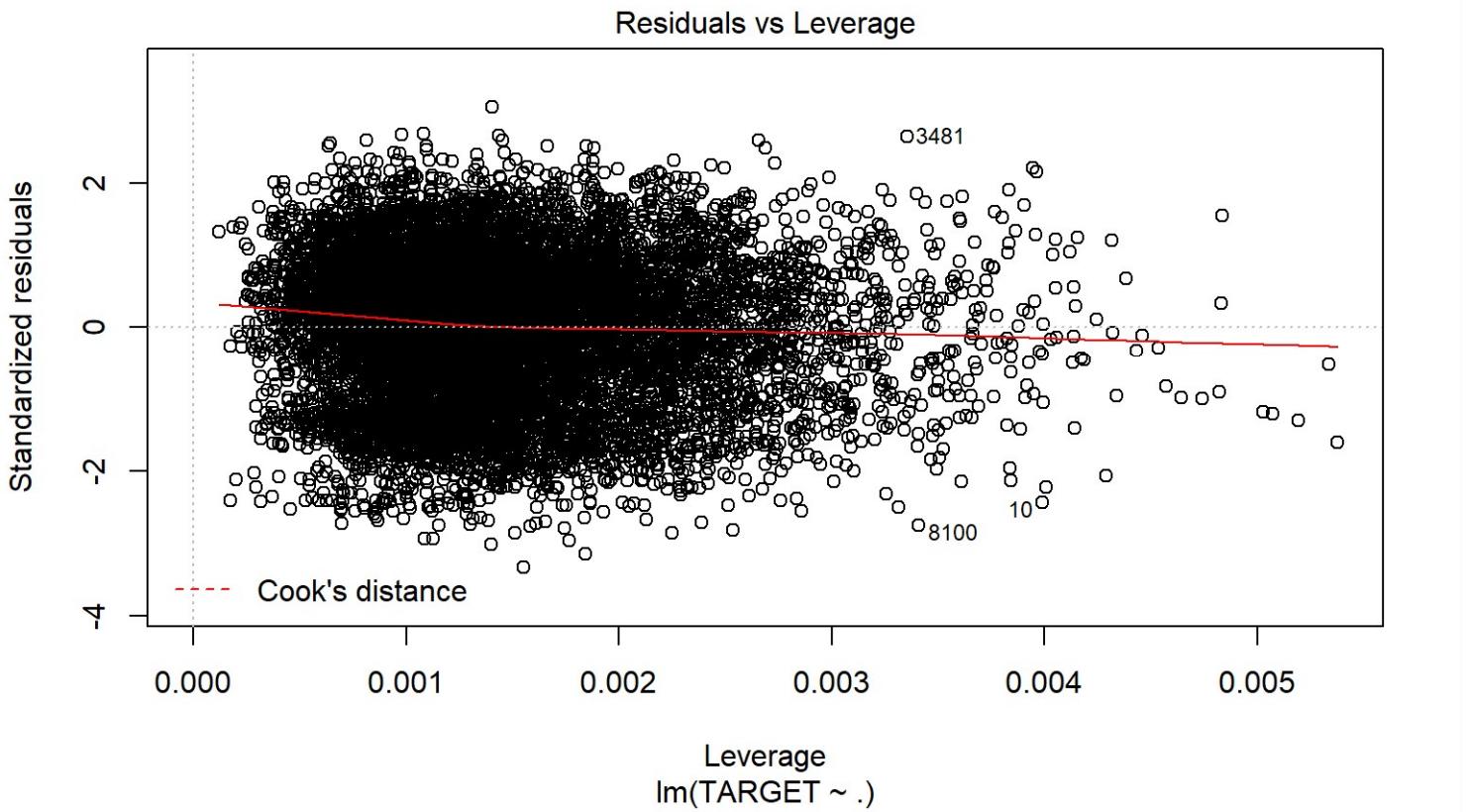


Theoretical Quantiles
 $\text{Im}(\text{TARGET} \sim .)$

Scale-Location



Fitted values
 $\text{Im}(\text{TARGET} \sim .)$



Model 10 : Linear Model with imputations and only significant variables.

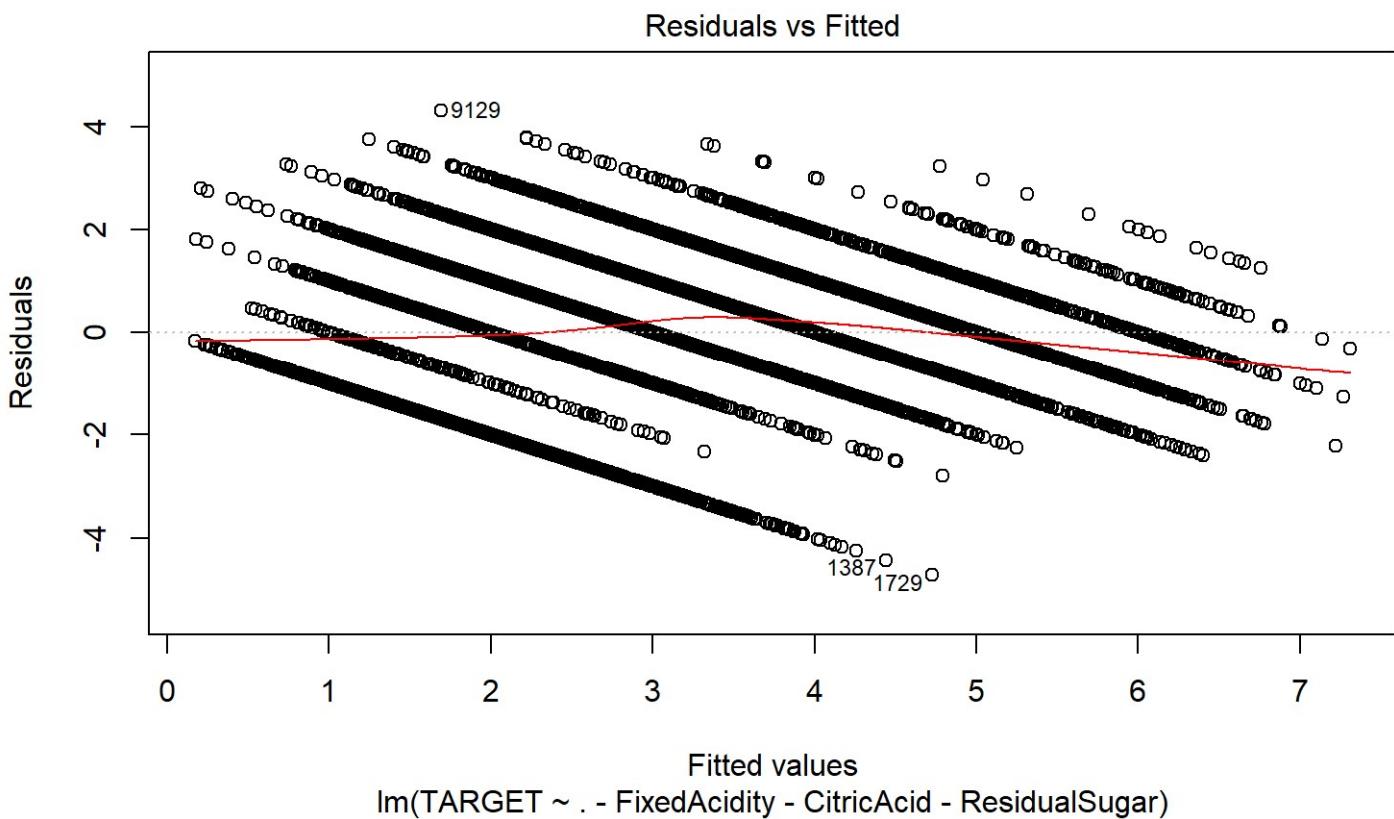
As we know the significant variables are `FixedAcidity`, `CitricAcid` and `ResidualSugar`, so using the same in the Linear Regression Model and applying the same of imputed training data.

```
##
## Call:
## lm(formula = TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar,
##     data = wine_train2)
##
## Residuals:
##   Min   1Q   Median   3Q   Max
## -4.7242 -1.0131  0.1728  1.0331  4.3050
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.951e+00 5.566e-01 10.691 < 2e-16 ***
## VolatileAcidity -1.278e-01 1.791e-02 -7.138 1.01e-12 ***
## Chlorides     -2.032e-01 4.391e-02 -4.627 3.75e-06 ***
## FreeSulfurDioxide 3.660e-04 9.386e-05 3.899 9.71e-05 ***
## TotalSulfurDioxide 2.454e-04 6.021e-05 4.075 4.63e-05 ***
```

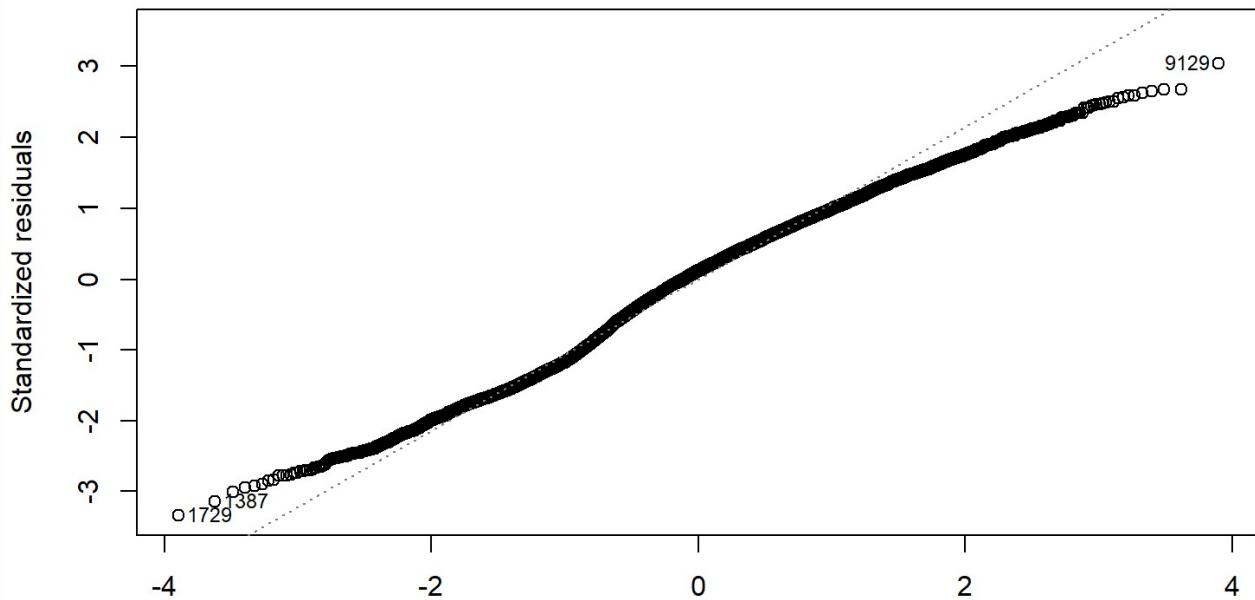
```

## Density      -8.712e-01 5.260e-01 -1.656 0.09768 .
## pH          -4.728e-02 2.071e-02 -2.282 0.02249 *
## Sulphates   -4.236e-02 1.511e-02 -2.803 0.00507 **
## Alcohol     1.249e-02 3.807e-03 3.281 0.00104 **
## LabelAppeal 4.310e-01 1.646e-02 26.186 < 2e-16 ***
## AcidIndex   -2.048e+00 9.074e-02 -22.572 < 2e-16 ***
## STARS       1.167e+00 1.671e-02 69.833 < 2e-16 ***
## ---
## Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
##
## Residual standard error: 1.416 on 10225 degrees of freedom
## Multiple R-squared: 0.4603, Adjusted R-squared: 0.4598
## F-statistic: 792.9 on 11 and 10225 DF, p-value: < 2.2e-16

```

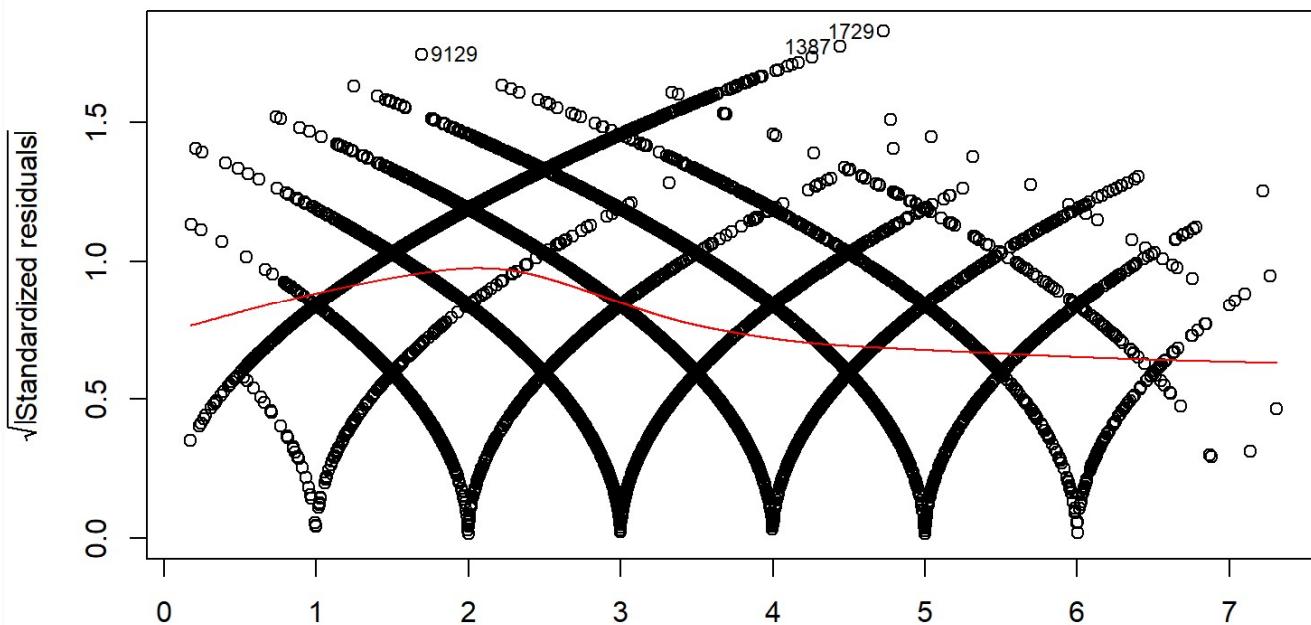


Normal Q-Q

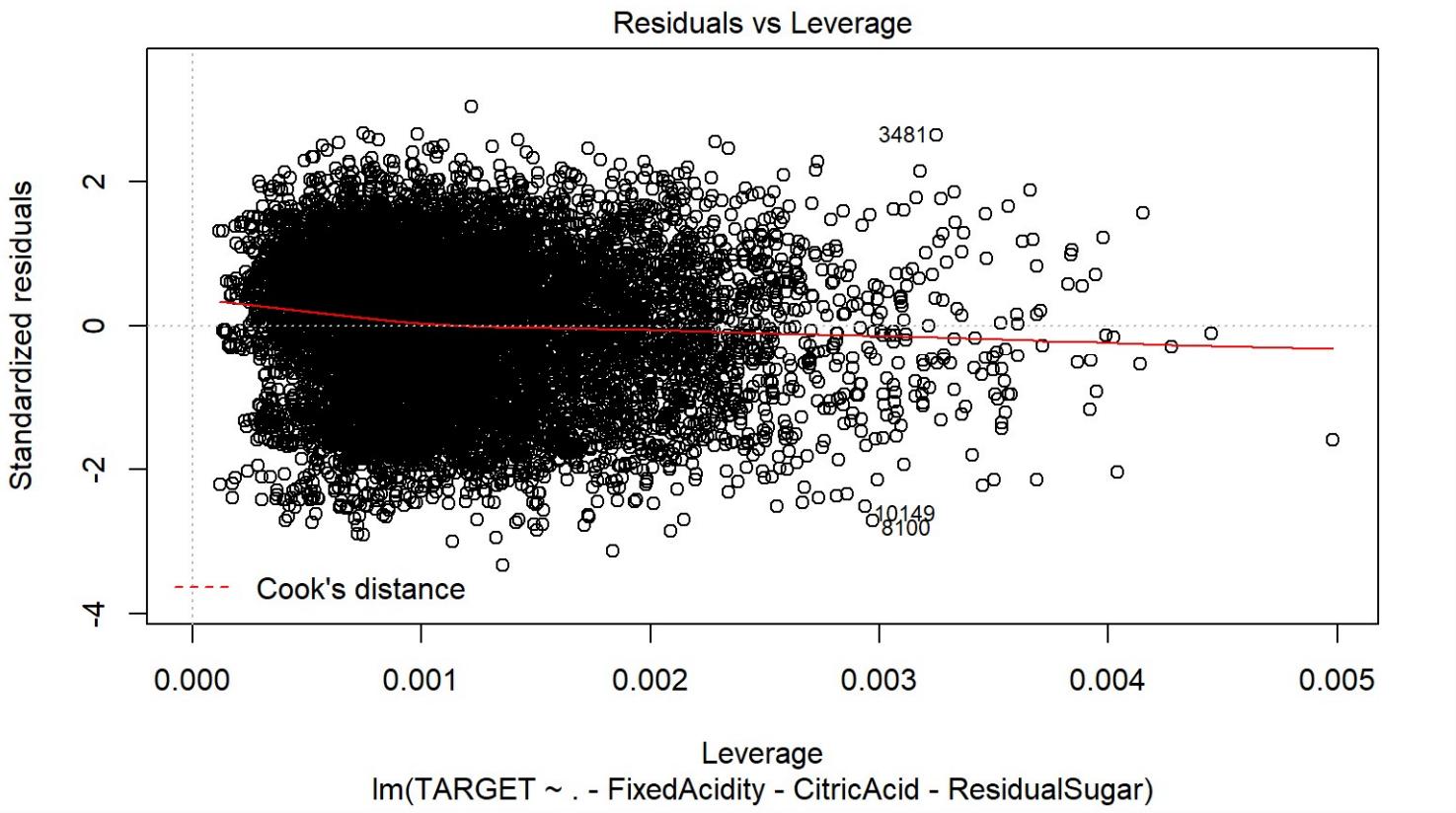


Theoretical Quantiles
lm(TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar)

Scale-Location



Fitted values
lm(TARGET ~ . - FixedAcidity - CitricAcid - ResidualSugar)



Model 11 : Ordinal Logistic Regression

This regression uses ordered factors. I would expect this to be one of the top performers.

```

## Call:
## polr(formula = TARGET ~ ., data = polrDF, Hess = TRUE)
##
## Coefficients:
##             Value Std. Error t value
## FixedAcidity 0.0021819 0.0029055 0.7510
## VolatileAcidity -0.1555960 0.0232760 -6.6848
## CitricAcid    0.0289713 0.0211212 1.3717
## ResidualSugar 0.0003196 0.0005320 0.6008
## Chlorides     -0.2627487 0.0566657 -4.6368
## FreeSulfurDioxide 0.0004607 0.0001216 3.7875
## TotalSulfurDioxide 0.0002716 0.0000783 3.4686
## Density      -1.2981402 0.1490930 -8.7069
## pH          -0.0314095 0.0268078 -1.1717
## Sulphates    -0.0339150 0.0196712 -1.7241
## Alcohol      0.0269097 0.0048969 5.4953
## LabelAppeal   0.8256163 0.0237699 34.7337

```

```

## AcidIndex      -2.6646249 0.1250905 -21.3016
## STARS        1.4684471 0.0256683 57.2086
##
## Intercepts:
##   Value Std. Error t value
## 0|1 -5.9211 0.1357 -43.6446
## 1|2 -5.7842 0.1355 -42.6743
## 2|3 -5.1811 0.1351 -38.3486
## 3|4 -3.8133 0.1350 -28.2556
## 4|5 -1.9656 0.1372 -14.3273
## 5|6  0.0034 0.1437  0.0237
## 6|7  2.2069 0.1675 13.1788
## 7|8  4.5480 0.3034 14.9895
##
## Residual Deviance: 30016.23
## AIC: 30060.23

```

Model 12 : Zero inflation

Zero inflation understands that some Poisson distributions are dominated by many zeros. As such it corrects for this. This is one of the most promising ones because as we saw in our data exploration, there were more zeros, and then normally distributed data after that.

```

##
## Call:
## zeroinfl(formula = TARGET ~ . | STARS, data = wine_train2, dist = "negbin")
##
## Pearson residuals:
##   Min    1Q Median    3Q   Max
## -2.09180 -0.49650  0.07134  0.48208  2.08565
##
## Count model coefficients (negbin with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.824e+00 2.385e-01  7.648 2.04e-14 ***
## FixedAcidity 4.523e-04 9.473e-04  0.477 0.633017
## VolatileAcidity -1.936e-02 7.560e-03 -2.561 0.010429 *
## CitricAcid   2.116e-03 6.718e-03  0.315 0.752830
## ResidualSugar -6.323e-05 1.722e-04 -0.367 0.713427
## Chlorides    -3.245e-02 1.851e-02 -1.754 0.079503 .
## FreeSulfurDioxide 4.932e-05 3.854e-05  1.280 0.200613
## TotalSulfurDioxide 4.760e-06 2.460e-05  0.194 0.846535
## Density     -2.990e-01 2.224e-01 -1.344 0.178840
## pH          -1.910e-03 8.749e-03 -0.218 0.827223
## Sulphates   -4.586e-03 6.391e-03 -0.718 0.473029
## Alcohol     5.950e-03 1.591e-03  3.741 0.000184 ***
## LabelAppeal  2.240e-01 7.112e-03 31.491 < 2e-16 ***
## AcidIndex    -2.682e-01 4.492e-02 -5.971 2.36e-09 ***
## STARS       1.227e-01 6.997e-03 17.531 < 2e-16 ***

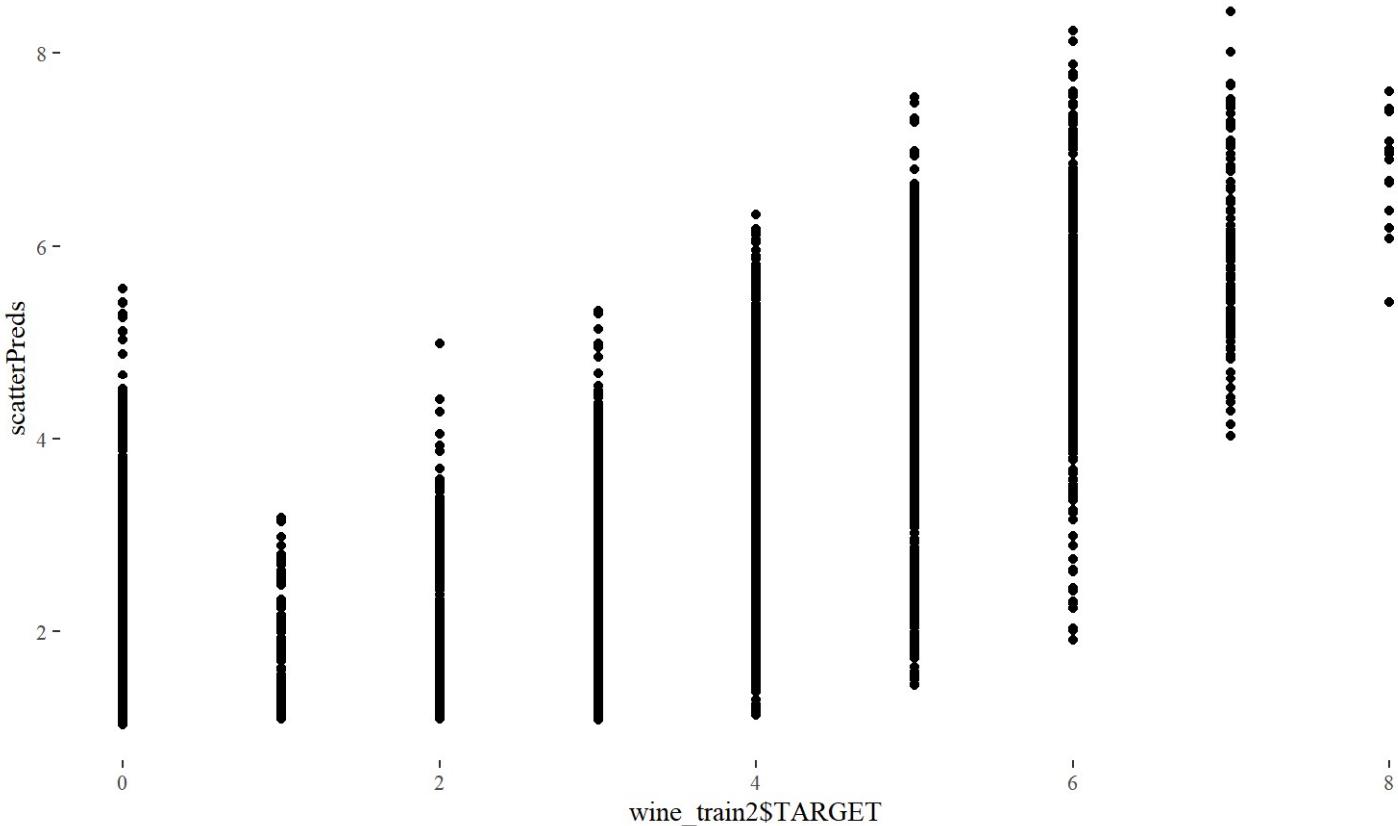
```

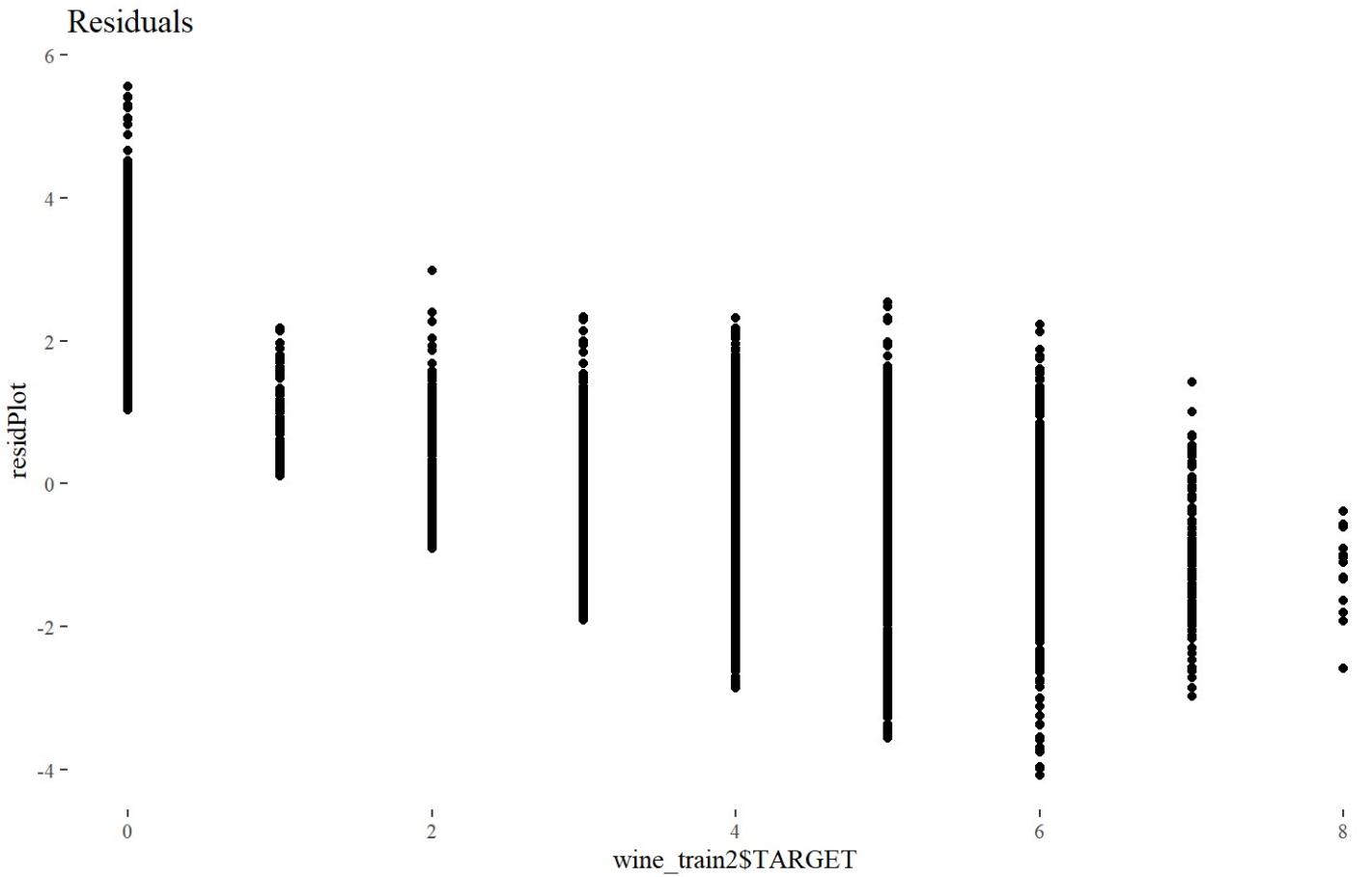
```

## Log(theta)      1.860e+01 2.758e+00 6.742 1.56e-11 ***
## 
## Zero-inflation model coefficients (binomial with logit link):
##   Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.31678  0.11092 20.89 <2e-16 ***
## STARS     -2.66899  0.09689 -27.55 <2e-16 ***
## ---
## Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
## 
## Theta = 119421399.3444
## Number of iterations in BFGS optimization: 22
## Log-likelihood: -1.725e+04 on 18 Df

```

Predicted vs Actual





MODEL SELECTION

Compare Models by MSE/AIC

MSE	AIC	
Model1	6.929787	6.926722
Model2	6.849005	6.849920
Model3	6.929788	6.926723
Model4	6.849001	6.849916
Model5	2.002207	2.002977
Model6	NA	1.988813
Model7	18544.983192	18535.285631

MSE	AIC	
Model8	38416.867732	38415.386508
Model9	18547.067808	18537.370350
Model10	38419.044066	38417.562708
Model11	6.929787	6.926722
Model12	6.849005	6.849920

Compare Models by Loss

Now lets see the output of the Models using test data

We will use the squared loss to validate the model. We will use the squared difference to select a model (MSE) from predictions on the training sets. (Lower numbers are better.)

	Loss:
Model1	5.471932
Model2	5.456474
Model3	6.827070
Model4	6.828240
Model5	5.471926
Model6	5.456468
Model7	6.827066
Model8	6.828236
Model9	2.034647
Model10	2.034185

Next

Following deductions can be made as per above values:

- A regular Poisson regression does not perform very well.
- The same can be said for the Negative Binomial.
- The linear model actually performs very well.
- Very surprisingly, Ordinal Logistic Regression does not work as well as the linear model.

Because we are not interested in gaining insight into the underlying causes of wine selection, we will use the squared loss. This will tell us how accurate our model is without caring about confidence intervals etc.

Based on this metric, Zero Poisson Inflation is the most accurate.

From the above models, Model12 - Zero Poisson Inflations has least loss as it uses less variables and is parsimonious. Also the R2 looks fine. The squared loss is also fine.

- In terms of MSE and AIC, Model 5 is best followed by Model 4 and tied Model 2 and Model 12.
- In terms of Loss, Model 12 is best followed closely by Model 9 and Model 10.

Overall, we choose Zero Poisson Inflation due to following : - least loss - good MSE score - good AIC score

6.3 Prediction on Evaluation Data

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

	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	Free
1	3.01474113581073	5.4	-0.86	0.27	-10.7	0.092	
2	3.62186861554457	12.4	0.385	-0.76	-19.7	1.169	
3	1.73831939712905	7.2	1.75	0.17	-33	0.065	
4	1.50697167978128	6.2	0.1	1.8	1	-0.179	
5	1.67686808814121	11.4	0.21	0.28	1.2	0.038	

	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	Free
6	5.66375927126166	17.6	0.04	-1.15	1.4	0.535	
7	3.38429159128796	15.5	0.53	-0.53	4.6	1.263	
8	5.17446610497721	15.9	1.19	1.14	31.9	-0.299	
9	1.54080400134772	11.6	0.32	0.55	-50.9	0.076	
10	4.28161045753364	3.8	0.22	0.31	-7.7	0.039	

Showing 1 to 10 of 3,335 entries

- **TARGET:** Number of Cases Purchased as Predicted

• ##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	StdD	Skew	Kurt
• ##	1.03	1.91	3.30	3.24	4.17	8.27	1.38	0.51	-0.32

Make Predictions

Predictions can be found in the following:

https://github.com/Rajwantmishra/DATA621_CR4/blob/master/HW5/Evaluation_Full_Data.csv

Appendix

https://github.com/Rajwantmishra/DATA621_CR4/blob/master/HW5/HomeWork5.Rmd

Thank you