Homework 5

Hector Santana, Zachary Safir, Mario Pena

Contents

1 Introduction

In this homework assignment, we 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.

Our 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.

2 Data Exploration

2.1 Summary Statistics

Below we can view both our data as well as a table with the summary statistics of our 14 predictor variables.

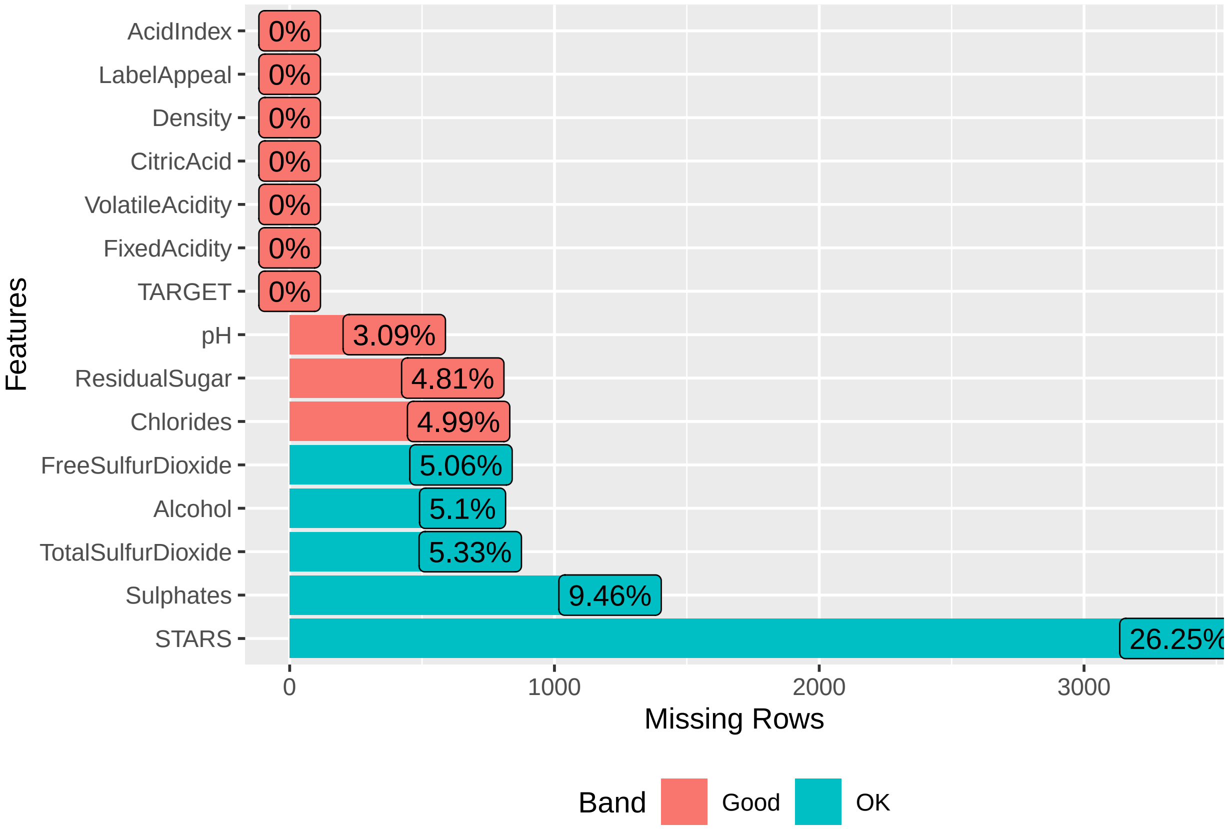
TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
3	3.2	1.160	-0.98	54.2	-0.567	NA	268	0.99280	3.33	-0.59	9.9	0	8	2
3	4.5	0.160	-0.81	26.1	-0.425	15	-327	1.02792	3.38	0.70	NA	-1	7	3
5	7.1	2.640	-0.88	14.8	0.037	214	142	0.99518	3.12	0.48	22.0	-1	8	3
3	5.7	0.385	0.04	18.8	-0.425	22	115	0.99640	2.24	1.83	6.2	-1	6	1
4	8.0	0.330	-1.26	9.4	NA	-167	108	0.99457	3.12	1.77	13.7	0	9	2
0	11.3	0.320	0.59	2.2	0.556	-37	15	0.99940	3.20	1.29	15.4	0	11	NA

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
FixedAcidity	1	12795	7.0757171	6.3176435	6.90000	7.0736739	3.2617200	-18.10000	34.40000	52.50000	-0.0225860	1.6749987	0.0558515
VolatileAcidity	2	12795	0.3241039	0.7840142	0.28000	0.3243890	0.4299540	-2.79000	3.68000	6.47000	0.0203800	1.8322106	0.0069311
CitricAcid	3	12795	0.3084127	0.8620798	0.31000	0.3102520	0.4151280	-3.24000	3.86000	7.10000	-0.0503070	1.8379401	0.0076213
ResidualSugar	4	12179	5.4187331	33.7493790	3.90000	5.5800410	15.7155600	-127.80000	141.15000	268.95000	-0.0531229	1.8846917	0.3058158
Chlorides	5	12157	0.0548225	0.3184673	0.04600	0.0540159	0.1349166	-1.17100	1.35100	2.52200	0.0304272	1.7886044	0.0028884
FreeSulfurDioxide	6	12148	30.8455713	148.7145577	30.00000	30.9334877	56.3388000	-555.00000	623.00000	1178.00000	0.0063930	1.8364966	1.3492769
TotalSulfurDioxide	7	12113	120.7142326	231.9132105	123.00000	120.8895367	134.9166000	-823.00000	1057.00000	1880.00000	-0.0071794	1.6746665	2.1071703
Density	8	12795	0.9942027	0.0265376	0.99449	0.9942130	0.0093552	0.88809	1.09924	0.21115	-0.0186938	1.8999592	0.0002346
pH	9	12400	3.2076282	0.6796871	3.20000	3.2055706	0.3854760	0.48000	6.13000	5.65000	0.0442880	1.6462681	0.0061038
Sulphates	10	11585	0.5271118	0.9321293	0.50000	0.5271453	0.4447800	-3.13000	4.24000	7.37000	0.0059119	1.7525655	0.0086602
Alcohol	11	12142	10.4892363	3.7278190	10.40000	10.5018255	2.3721600	-4.70000	26.50000	31.20000	-0.0307158	1.5394949	0.0338306
LabelAppeal	12	12795	-0.0090660	0.8910892	0.00000	-0.0099639	1.4826000	-2.00000	2.00000	4.00000	0.0084295	-0.2622916	0.0078777
AcidIndex	13	12795	7.7727237	1.3239264	8.00000	7.6431572	1.4826000	4.00000	17.00000	13.00000	1.6484959	5.1900925	0.0117043
STARS	14	9436	2.0417550	0.9025400	2.00000	1.9711258	1.4826000	1.00000	4.00000	3.00000	0.4472353	-0.6925343	0.0092912

2.2 Missing Variables

According to our summary statistics above, and the visual shown below, about half of our predictor variables have missing values. From the least amount of missing values to the most, these include “pH”, “ResidualSugar”, “Chlorides”, “FreeSulfurDioxide”, “Alcohol”, “TotalSulfurDioxide”, “Sulphates”, and “STARS”.

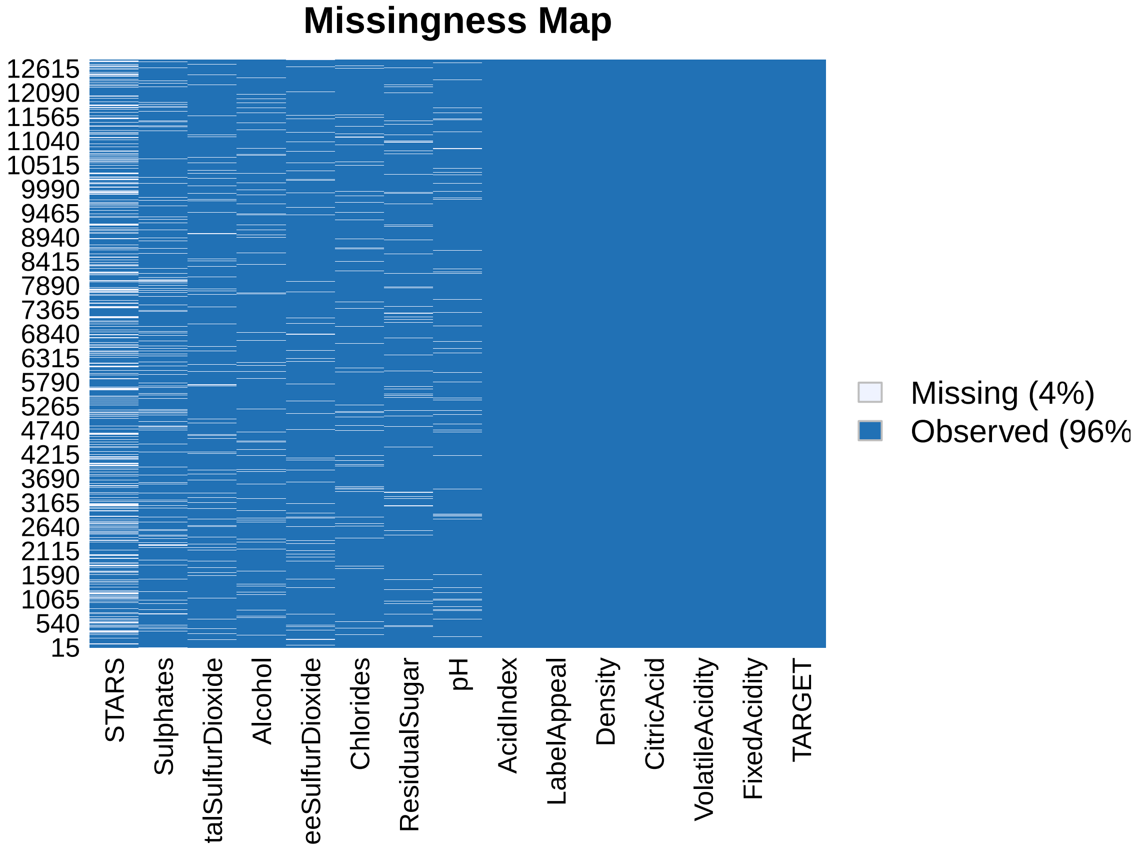
We also observe that there are quite a few variables with negative values, and we will look into this to verify that they are legitimate.



We can also observe that the "TARGET" variable, which is the number of cases purchased, ranges from 0 to 8 and it does not seem to be evenly distributed throughout the data. We can observe that about 44% of the wines in the data have been purchased in batches of 3 to 4 cases.

```
          0           1           2           3           4           5 
0.213677218 0.019069949 0.085267683 0.204064088 0.248300117 0.157405236 
          6           7           8 
0.059788980 0.011098085 0.001328644 
```

We can observe below a map of the missing values. It appears that the missing values for "STARS" is spread out through the data, indicating that our missing values do no represent a singular group of wines.



Below is the result of the Little's test statistic to assess if data is missing completely at random (MCAR). The null hypothesis in this test is that the data is MCAR, and the test statistic is a chi-squared value. We can conclude from this test that our missing values are Missing at Random (MAR).

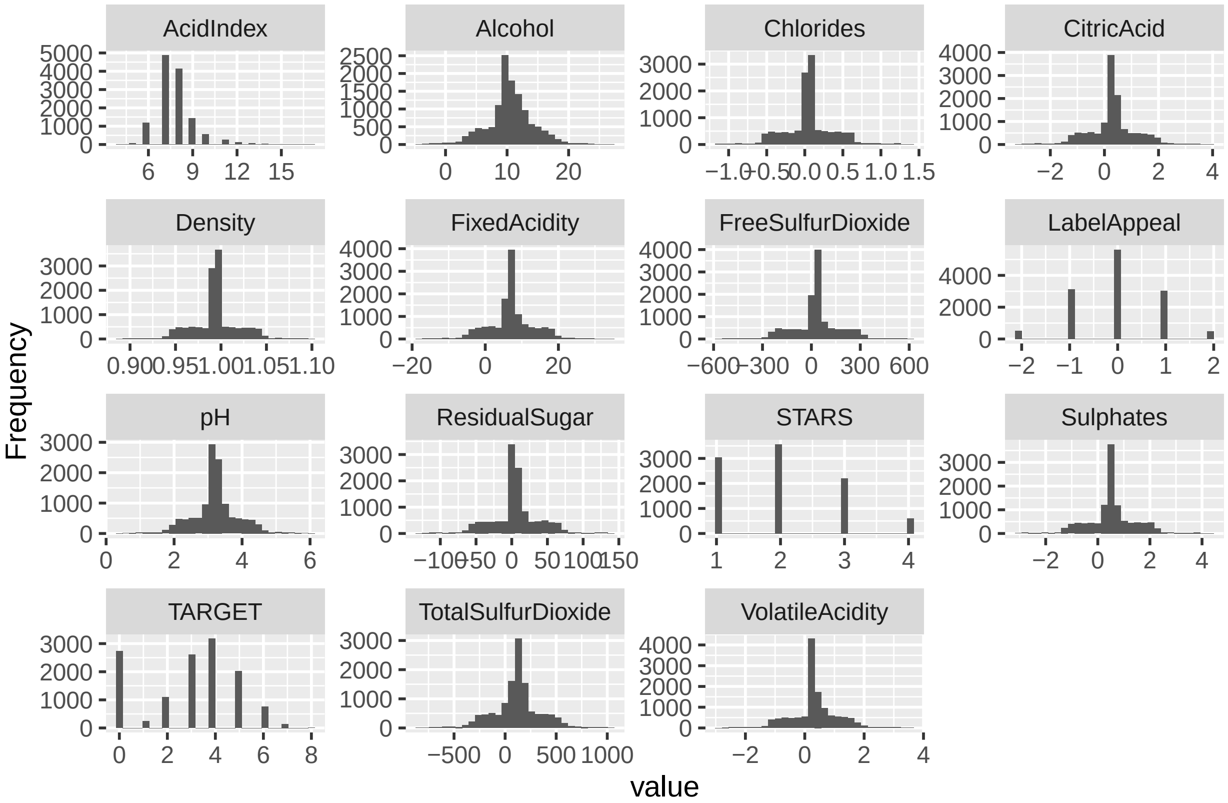
Table 1: Little's Test

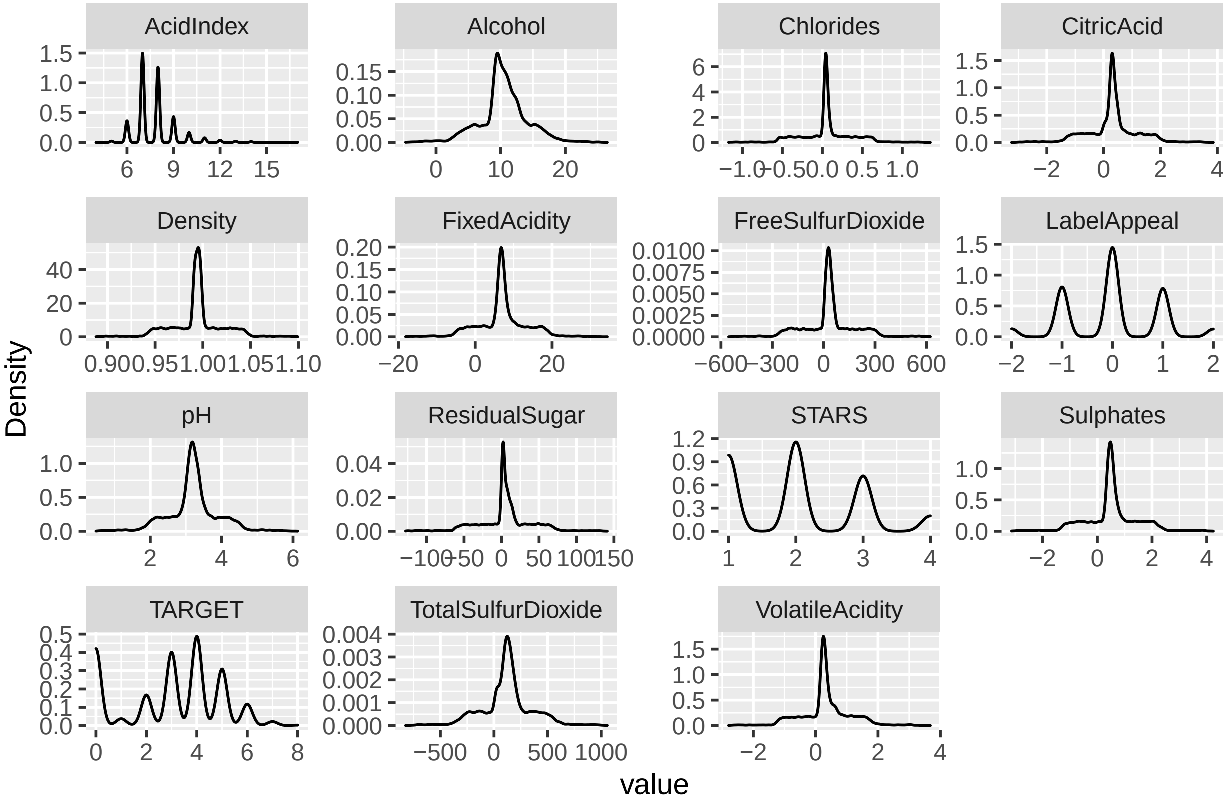
statistic	df	p.value	missing.patterns
5337.974	1142	0	94

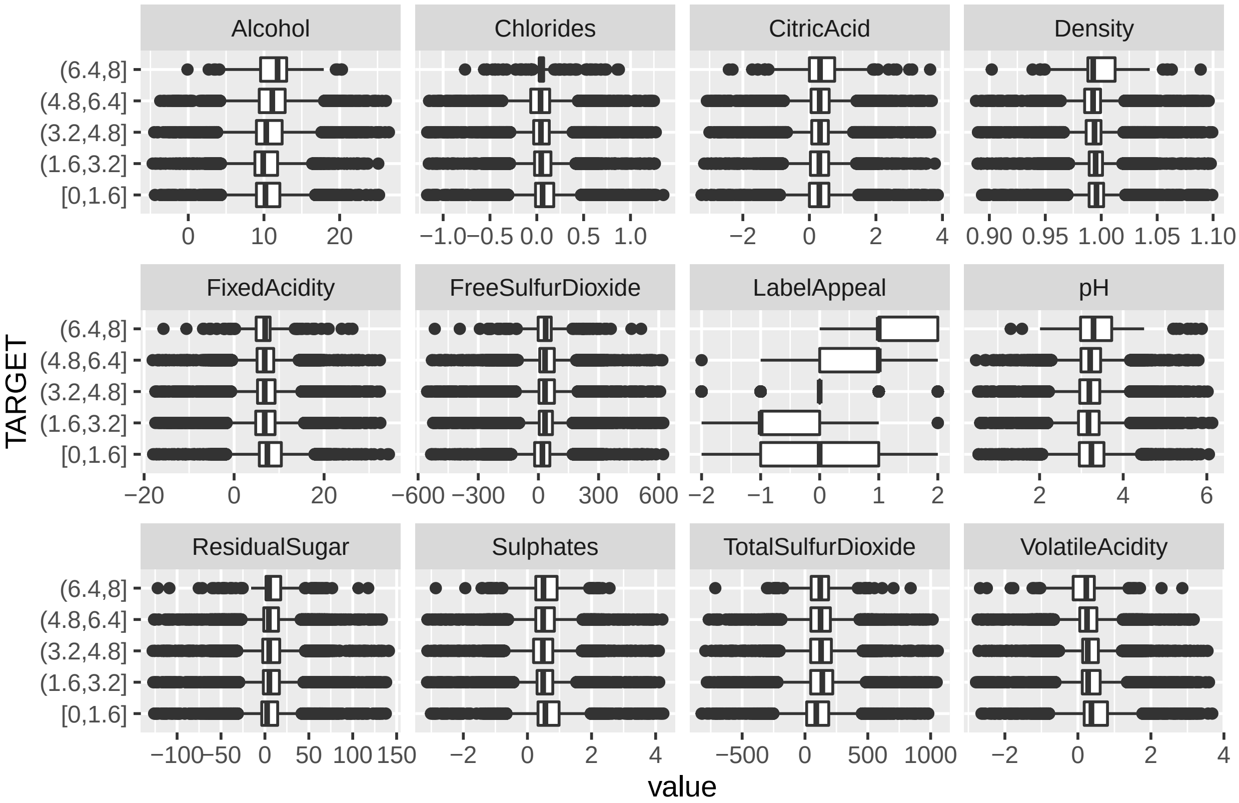
2.3 Visual Analysis of Data Structure

The insight gained from the statistical analysis permitted us to make note of further data of interest that needed to be analyzed in depth prior to the creation of our models. To confirm these irregularities we then constructed visual representations consisting of density plots, histograms, and boxplots.

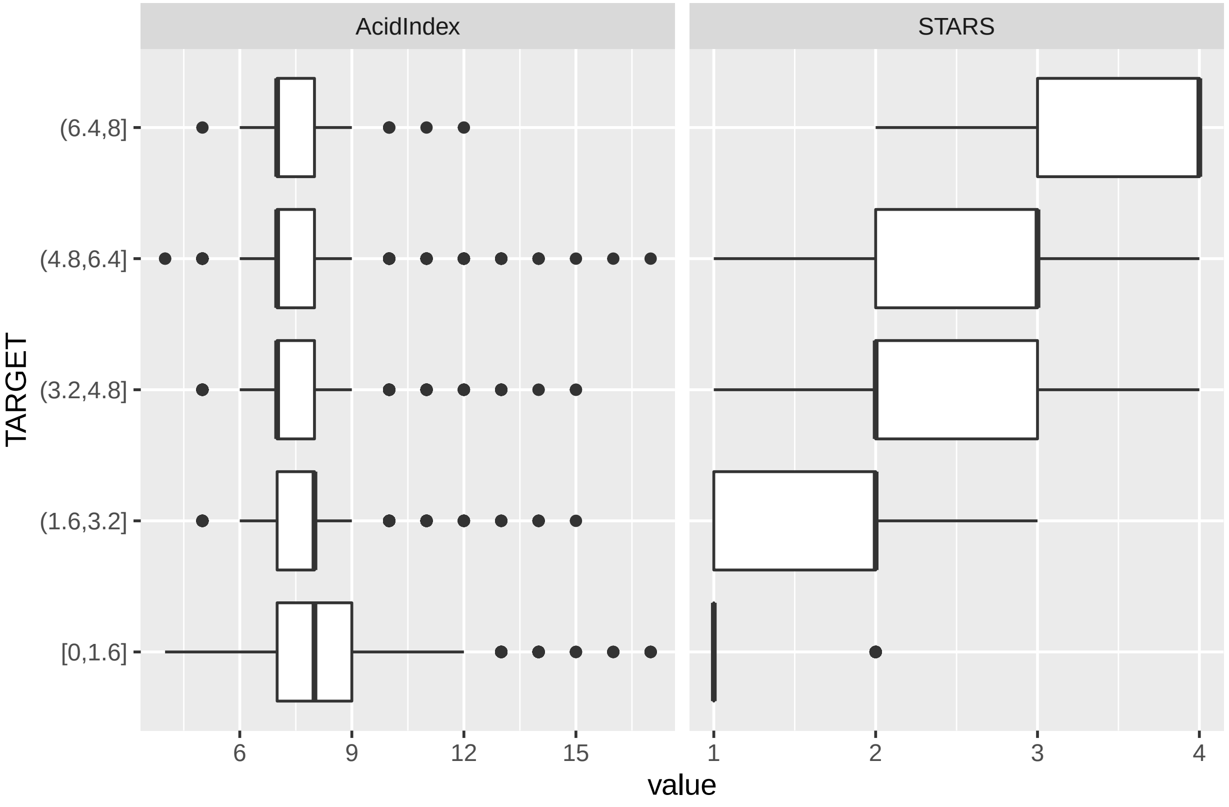
We can observe from the histograms below that most of our variables appear to be following a close to normal distribution, as it is also evident in the density plots. The only variables that do not show a close to normal distribution seem to be ordinal variables. "AcidIndex", "LabelAppeal" and "STARS" are based on a rating scale about an attribute of the wine.


AcidIndex
Alcohol
Chlorides
CitricAcid
Density
FixedAcidity
FreeSulfurDioxide
LabelAppeal
pH
ResidualSugar
STARS
Sulphates
TARGET
TotalSulfurDioxide
VolatileAcidity
Frequency
value
0
1000
2000
3000
4000
5000
500
1500
2500
6
9
12
15
0
10
20
−1.0
−0.5
0.0
0.5
1.0
1.5
−2
0
2
4
0.90
0.95
1.00
1.05
1.10
−20
−600
−300
300
600
−1
1
3
6
−100
−50
50
100
150
8
−500
1000


AcidIndex
Alcohol
Chlorides
CitricAcid
Density
FixedAcidity
FreeSulfurDioxide
LabelAppeal
pH
ResidualSugar
STARS
Sulphates
TARGET
TotalSulfurDioxide
VolatileAcidity
Density
value
0.0
0.5
1.0
1.5
0.00
0.05
0.10
0.15
0.20
0
2
4
6
20
40
0.0000
0.0025
0.0050
0.0075
0.0100
0.02
0.04
0.3
0.6
0.9
1.2
0.1
0.2
0.4
0.000
0.001
0.002
0.003
0.004
9
12
15
10
−1.0
−0.50
−2
0.90
0.95
1.00
1.05
1.10
−20
−600
−300
300
600
−1
1
−100
−50
50
100
150
3
8
−500
500
1000


Alcohol
Chlorides
CitricAcid
Density
FixedAcidity
FreeSulfurDioxide
LabelAppeal
pH
ResidualSugar
Sulphates
TotalSulfurDioxide
VolatileAcidity
TARGET
(6.4,8]
(4.8,6.4]
(3.2,4.8]
(1.6,3.2]
[0,1.6]
0 10 20
−1.0 −0.5 0.0 0.5 1.0
−2 0 2 4
0.90 0.95 1.00 1.05 1.10
−20 0 20
−600 −300 0 300 600
−2 −1 0 1 2
2 4 6
−100 −50 0 50 100 150
−2 0 2 4
−500 0 500 1000
−2 0 2 4
value

Page 1


AcidIndex
STARS
TARGET
(6.4,8]
(4.8,6.4]
(3.2,4.8]
(1.6,3.2]
[0,1.6]
6 9 12 15
1 2 3 4
value

Page 2

3 Data Preparation

3.1 Removing Eroneous Values

In our data set we discovered many negative values. After some investigation, we concluded that the chemical properties of wine may not be measured in negative values. Only positive values describe if certain chemicals exist in wine. Zero indicating the chemical is not present or a positive value otherwise. We will use take absolute value of all our variables to fix this problem.

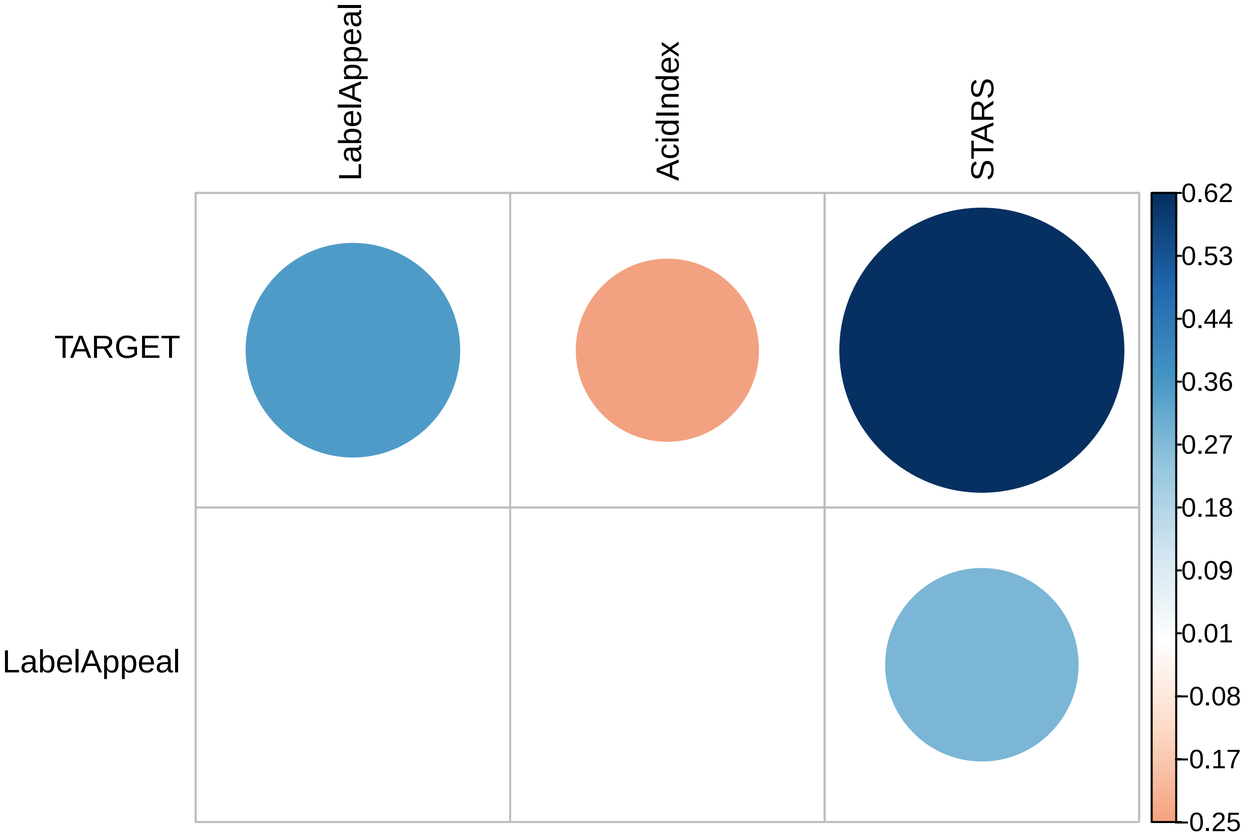
The variable "LabelAppeal" has approximately 28% negative values. It provides a rating of wine bottle label and may be treated as a categorical variable. It is safe to assume the value of -2 is worst, and 2 is best as the variable only has five values.

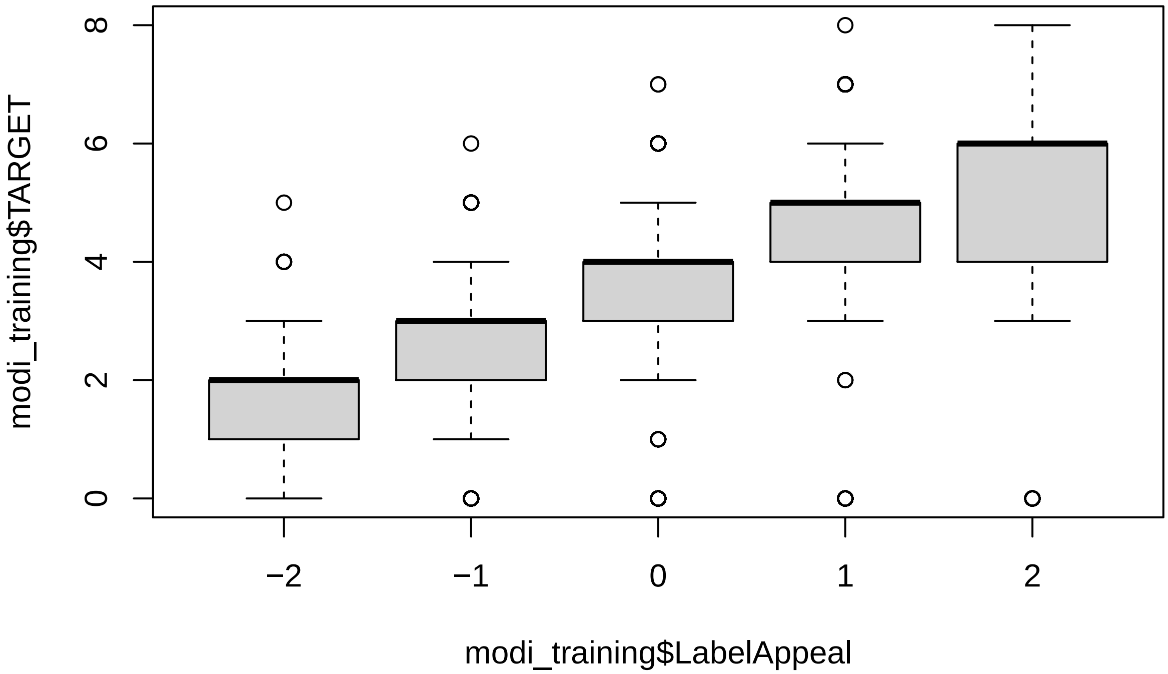
In addition, we also discovered many wines that were either more acidic than stomach acid or, on the opposite end, had a pH that was almost the same as water. After some research, we decided to cut off all variables with a pH below 3, and above a pH of 4.19.

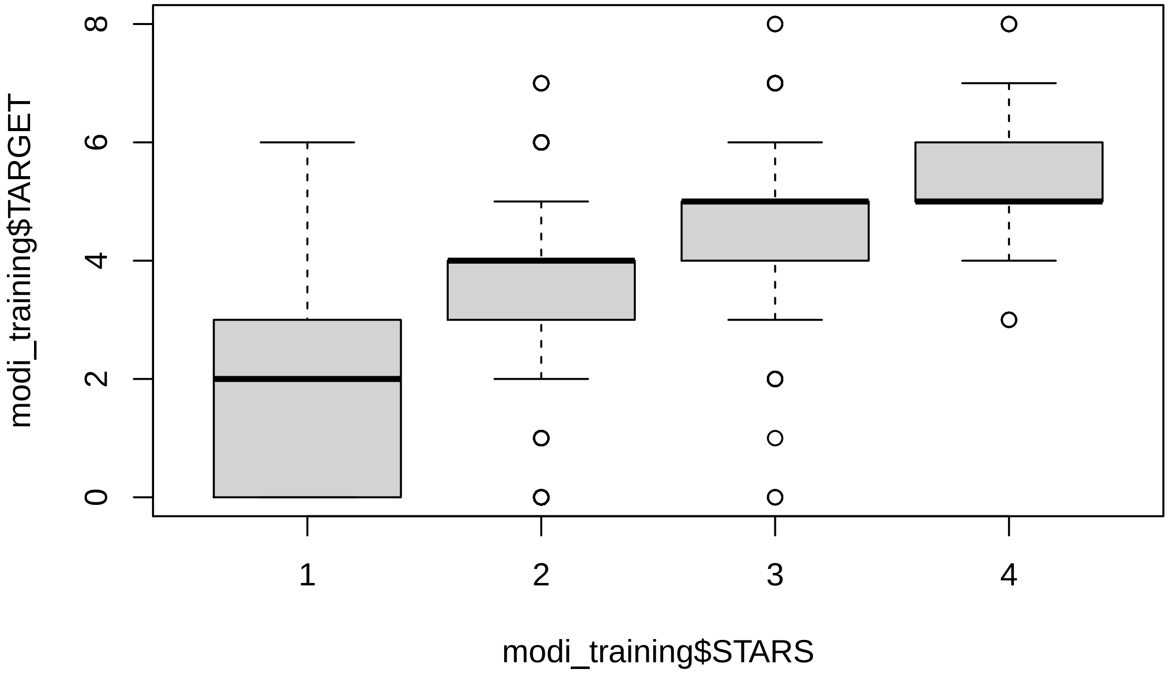
3.2 Missing Value Handling

As we mentioned earlier, more than half of our predictor variables have missing values. We will be using the MICE package to impute the missing values.

Below is a look at the largest correlations of our newly created data set, using the imputed values from MICE







	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
FixedAcidity	1	8417	8.0602649	5.0061176	7.0000	7.5574165	2.9652000	0.00000	34.40000	34.40000	1.2081660	2.0829102	0.0545660
VolatileAcidity	2	8417	0.6433652	0.5523736	0.4200	0.5532056	0.3261720	0.00000	3.59000	3.59000	1.6381020	3.0453792	0.0060208
CitricAcid	3	8417	0.6905833	0.6136274	0.4400	0.5913318	0.3261720	0.00000	3.77000	3.77000	1.6633638	3.0226787	0.0066885
ResidualSugar	4	8417	23.4817690	25.2079033	12.8000	19.4533482	16.3086000	0.00000	141.15000	141.15000	1.4777199	2.2347400	0.2747629
Chlorides	5	8417	0.2250494	0.2342352	0.1040	0.1869823	0.1082298	0.00000	1.27000	1.27000	1.4476904	2.0417846	0.0025531
FreeSulfurDioxide	6	8417	106.7542474	107.5453967	57.0000	89.5755011	60.7866000	0.00000	623.00000	623.00000	1.5295155	2.4431525	1.1722309
TotalSulfurDioxide	7	8417	205.8036117	164.4355210	155.0000	181.3768374	102.2994000	0.00000	1054.00000	1054.00000	1.5871142	2.8725116	1.7923259
Density	8	8417	0.9942839	0.0265410	0.9944	0.9942537	0.0091921	0.88809	1.09924	0.21115	0.0076901	1.8670983	0.0002893
pH	9	8417	3.3621979	0.2947639	3.2700	3.3193838	0.2223900	3.00000	4.19000	1.19000	1.1655411	0.4900929	0.0032129
Sulphates	10	8417	0.8478662	0.6529773	0.5900	0.7444618	0.3261720	0.00000	4.21000	4.21000	1.6893964	3.1962519	0.0071174
Alcohol	11	8417	10.5370203	3.6011478	10.4000	10.5093096	2.3721600	0.00000	26.00000	26.00000	0.1945216	1.0599462	0.0392520
AcidIndex	12	8417	7.7497921	1.3213956	8.0000	7.6190052	1.4826000	4.00000	17.00000	13.00000	1.7216166	5.6365915	0.0144030

4 Model Building

4.1 Poisson Models

Following the data preparation phase, we brainstormed how best to construct an appropriate model design process. We split our data into an additional training and test set in order to use 70% of it in the models and then evaluate their performance with the predictions against the remaining 30%.

Using the partition, we constructed a saturated count regression model which contained all variables of the dataset. This gave us a starting point to analyze the statistical significance of each variable and their associated correlations to the dependent variable.

```
Call:
glm(formula = TARGET ~ ., family = poisson, data = additional_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2377  -0.6520   0.0612   0.5669   2.7866

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.983e+00  3.008e-01   6.591 4.37e-11 ***
FixedAcidity        1.530e-03  1.530e-03   1.000 0.317178
VolatileAcidity    -2.930e-02  1.379e-02  -2.124 0.033642 *
CitricAcid          6.404e-03  1.226e-02   0.523 0.601303
ResidualSugar      -5.493e-05  2.977e-04  -0.185 0.853611
Chlorides          -5.949e-02  3.255e-02  -1.827 0.067641 .
FreeSulfurDioxide  -2.849e-05  6.960e-05  -0.409 0.682231
TotalSulfurDioxide  1.128e-04  4.494e-05   2.509 0.012091 *
Density            -6.425e-01  2.789e-01  -2.304 0.021228 *
pH                 -1.091e-01  2.622e-02  -4.160 3.19e-05 ***
Sulphates          -2.516e-02  1.176e-02  -2.140 0.032370 *
Alcohol             4.346e-03  2.101e-03   2.069 0.038563 *
LabelAppeal-1       2.012e-01  5.715e-02   3.521 0.000429 ***
LabelAppeal0        3.795e-01  5.571e-02   6.812 9.64e-12 ***
LabelAppeal1        5.093e-01  5.655e-02   9.007  < 2e-16 ***
LabelAppeal2        6.103e-01  6.366e-02   9.588  < 2e-16 ***
AcidIndex          -1.004e-01  6.727e-03 -14.929  < 2e-16 ***
STARS2              6.291e-01  1.964e-02  32.026  < 2e-16 ***
STARS3              8.221e-01  2.153e-02  38.178  < 2e-16 ***
STARS4              9.590e-01  3.132e-02  30.614  < 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: 10642.2  on 5894  degrees of freedom
Residual deviance:  7212.9  on 5875  degrees of freedom
AIC: 21931

Number of Fisher Scoring iterations: 5
```

Our second model is based on only the predictor variables that had statistical significance from our previous model.

```
Call:
glm(formula = TARGET ~ pH + LabelAppeal + AcidIndex + STARS, 
    family = poisson, data = additional_train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.2122  -0.6570   0.0776   0.5692   2.8556  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)    1.393013   0.118316  11.774  < 2e-16 ***
pH            -0.110982   0.026168  -4.241 2.22e-05 ***
LabelAppeal-1  0.198700   0.057092   3.480 0.000501 ***
LabelAppeal0   0.375394   0.055670   6.743 1.55e-11 ***
LabelAppeal1   0.506666   0.056478   8.971  < 2e-16 ***
LabelAppeal2   0.607604   0.063616   9.551  < 2e-16 ***
AcidIndex     -0.101990   0.006629 -15.386  < 2e-16 ***
STARS2         0.630608   0.019589  32.193  < 2e-16 ***
STARS3         0.825874   0.021495  38.421  < 2e-16 ***
STARS4         0.964784   0.031235  30.888  < 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: 10642.2  on 5894  degrees of freedom
Residual deviance:  7242.9  on 5885  degrees of freedom
AIC: 21941

Number of Fisher Scoring iterations: 5
```

4.2 Negative Binomial

```
Call:
glm.nb(formula = TARGET ~ ., data = additional_train, init.theta = 45573.1326, 
    link = log)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.2377  -0.6520   0.0612   0.5669   2.7866  

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)         1.983e+00  3.008e-01   6.591 4.38e-11 ***
FixedAcidity        1.530e-03  1.530e-03   1.000 0.317189    
VolatileAcidity    -2.930e-02  1.379e-02  -2.124 0.033644 *  
CitricAcid          6.404e-03  1.226e-02   0.523 0.601321    
ResidualSugar      -5.493e-05  2.977e-04  -0.185 0.853610    
Chlorides          -5.949e-02  3.256e-02  -1.827 0.067645 .  
FreeSulfurDioxide  -2.849e-05  6.960e-05  -0.409 0.682249    
TotalSulfurDioxide  1.128e-04  4.494e-05   2.509 0.012091 *  
Density            -6.425e-01  2.789e-01  -2.304 0.021231 *  
pH                 -1.091e-01  2.622e-02  -4.159 3.19e-05 ***
Sulphates          -2.516e-02  1.176e-02  -2.140 0.032371 *  
Alcohol             4.346e-03  2.101e-03   2.069 0.038576 *  
LabelAppeal-1       2.012e-01  5.715e-02   3.521 0.000429 ***
LabelAppeal0        3.795e-01  5.571e-02   6.812 9.65e-12 ***
LabelAppeal1        5.093e-01  5.655e-02   9.007  < 2e-16 ***
LabelAppeal2        6.103e-01  6.366e-02   9.587  < 2e-16 ***
AcidIndex          -1.004e-01  6.727e-03 -14.929  < 2e-16 ***
STARS2              6.291e-01  1.964e-02  32.025  < 2e-16 ***
STARS3              8.221e-01  2.153e-02  38.177  < 2e-16 ***
STARS4              9.590e-01  3.133e-02  30.613  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(45573.13) family taken to be 1)

    Null deviance: 10641.7  on 5894  degrees of freedom
Residual deviance:  7212.6  on 5875  degrees of freedom
AIC: 21933

Number of Fisher Scoring iterations: 1


              Theta:  45573 
          Std. Err.:  68711 
Warning while fitting theta: iteration limit reached 

 2 x log-likelihood:  -21890.93 
```

We can then generate a new model using the stepAIC function

```
Call:
glm.nb(formula = TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide +
    Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex +
    STARS, data = additional_train, init.theta = 45557.52805,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2430  -0.6511   0.0621   0.5709   2.7792

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.991e+00  3.006e-01   6.624  3.5e-11 ***
VolatileAcidity    -2.933e-02  1.379e-02  -2.127 0.033424 *
Chlorides          -5.914e-02  3.256e-02  -1.816 0.069302 .
TotalSulfurDioxide  1.128e-04  4.492e-05   2.510 0.012071 *
Density            -6.471e-01  2.788e-01  -2.321 0.020291 *
pH                 -1.086e-01  2.620e-02  -4.145  3.4e-05 ***
Sulphates          -2.483e-02  1.175e-02  -2.114 0.034531 *
Alcohol             4.318e-03  2.100e-03   2.056 0.039779 *
LabelAppeal-1       1.996e-01  5.712e-02   3.495 0.000474 ***
LabelAppeal0        3.784e-01  5.570e-02   6.793  1.1e-11 ***
LabelAppeal1        5.080e-01  5.652e-02   8.987  < 2e-16 ***
LabelAppeal2        6.091e-01  6.365e-02   9.570  < 2e-16 ***
AcidIndex          -9.934e-02  6.653e-03 -14.932  < 2e-16 ***
STARS2              6.287e-01  1.964e-02  32.016  < 2e-16 ***
STARS3              8.217e-01  2.153e-02  38.168  < 2e-16 ***
STARS4              9.592e-01  3.129e-02  30.652  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(45557.53) family taken to be 1)

    Null deviance: 10641.7  on 5894  degrees of freedom
Residual deviance:  7214.1  on 5879  degrees of freedom
AIC: 21926

Number of Fisher Scoring iterations: 1

              Theta:  45558
          Std. Err.:  68696
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -21892.41
```

Finally, we explore using the transformed data

```
Call:
glm.nb(formula = TARGET ~ ., data = additional_train_tran, init.theta = 44593.82149,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0837  -0.6528   0.0463   0.5564   3.3097

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.219e+05  2.547e+05  -0.871 0.383658
FixedAcidity                1.310e-02  2.236e-02   0.586 0.558104
VolatileAcidity            -3.190e-01  2.354e-01  -1.355 0.175409
CitricAcid                  6.671e-01  1.954e-01   3.415 0.000639 ***
ResidualSugar              -4.118e-04  1.842e-03  -0.224 0.823108
Chlorides                  -3.029e+00  1.114e+00  -2.718 0.006570 **
FreeSulfurDioxide           1.113e-03  3.747e-04   2.969 0.002985 **
TotalSulfurDioxide         -3.972e-04  3.185e-04  -1.247 0.212259
Density                    -2.270e+05  2.605e+05  -0.871 0.383583
pH                          1.311e+00  5.807e+00   0.226 0.821322
Sulphates                   1.294e-02  1.809e-01   0.071 0.943003
Alcohol                    -3.826e-02  3.514e-01  -0.109 0.913293
LabelAppeal-1               2.047e-01  5.720e-02   3.580 0.000344 ***
LabelAppeal0                3.852e-01  5.579e-02   6.905 5.03e-12 ***
LabelAppeal1                5.150e-01  5.663e-02   9.095  < 2e-16 ***
LabelAppeal2                6.196e-01  6.378e-02   9.715  < 2e-16 ***
AcidIndex                  -1.603e-01  2.567e-01  -0.624 0.532419
STARS2                      6.072e-01  1.971e-02  30.806  < 2e-16 ***
STARS3                      7.887e-01  2.171e-02  36.329  < 2e-16 ***
STARS4                      9.311e-01  3.151e-02  29.550  < 2e-16 ***
FixedAcidity_lam_one       -3.324e-01  5.414e-01  -0.614 0.539279
FixedAcidity_log            1.816e-01  2.622e-01   0.692 0.488636
VolatileAcidity_lam_one     8.598e-01  4.732e-01   1.817 0.069240 .
VolatileAcidity_log         1.136e+00  7.526e-01   1.509 0.131193
CitricAcid_lam_one         -2.228e+00  4.925e-01  -4.524 6.05e-06 ***
CitricAcid_log             -2.694e+00  6.815e-01  -3.952 7.74e-05 ***
ResidualSugar_lam_one      -4.822e-01  1.760e+01  -0.027 0.978148
ResidualSugar_log           2.954e-02  8.019e-01   0.037 0.970614
Chlorides_lam_one           1.132e+00  3.514e-01   3.222 0.001271 **
Chlorides_log               5.713e+00  2.017e+00   2.832 0.004624 **
FreeSulfurDioxide_lam_one  -9.271e+00  2.213e+00  -4.190 2.80e-05 ***
FreeSulfurDioxide_log       1.216e+00  2.827e-01   4.302 1.69e-05 ***
TotalSulfurDioxide_lam_one  4.845e-02  1.379e-01   0.351 0.725375
TotalSulfurDioxide_log      5.308e-02  1.086e-01   0.489 0.624950
Density_lam_one             2.217e+05  2.545e+05   0.871 0.383636
Density_log                 3.952e+03  4.481e+03   0.882 0.377713
pH_lam_one                  2.552e+02  1.882e+03   0.136 0.892150
pH_log                     -5.593e+00  3.146e+01  -0.178 0.858878
Sulphates_lam_one           7.579e-02  5.869e-01   0.129 0.897261
Sulphates_log              -3.045e-02  6.941e-01  -0.044 0.965001
Alcohol_lam_one             6.118e-02  4.824e-01   0.127 0.899077
Alcohol_log                -4.261e-02  2.376e-01  -0.179 0.857637
```

```
AcidIndex_lam_one          -3.815e+01  2.969e+01  -1.285 0.198821
AcidIndex_log              -1.430e+00  3.842e+00  -0.372 0.709665
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(44593.82) family taken to be 1)

    Null deviance: 10641.7  on 5894  degrees of freedom
Residual deviance:  7023.5  on 5851  degrees of freedom
AIC: 21792

Number of Fisher Scoring iterations: 1

              Theta:  44594
          Std. Err.:  64650
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -21701.84
```

4.3 Multiple Linear Regression

First we create a model with all our variables

```
Call:
lm(formula = TARGET ~ ., data = additional_train)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6601 -0.9911  0.1351  1.0093  4.4194

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.9057520  0.7341614   8.044 1.04e-15 ***
FixedAcidity        0.0049486  0.0037431   1.322  0.18621
VolatileAcidity    -0.0851012  0.0336152  -2.532  0.01138 *
CitricAcid          0.0192510  0.0305535   0.630  0.52867
ResidualSugar      -0.0002968  0.0007389  -0.402  0.68794
Chlorides          -0.1695195  0.0799180  -2.121  0.03395 *
FreeSulfurDioxide  -0.0001070  0.0001721  -0.622  0.53411
TotalSulfurDioxide  0.0003413  0.0001121   3.043  0.00235 **
Density            -1.8598406  0.6904561  -2.694  0.00709 **
pH                 -0.3158038  0.0632950  -4.989 6.23e-07 ***
Sulphates          -0.0750711  0.0286615  -2.619  0.00884 **
Alcohol             0.0151218  0.0051906   2.913  0.00359 **
LabelAppeal-1       0.3135850  0.1050276   2.986  0.00284 **
LabelAppeal0        0.7596630  0.1022733   7.428 1.26e-13 ***
LabelAppeal1        1.2249225  0.1062727  11.526  < 2e-16 ***
LabelAppeal2        1.6489818  0.1381271  11.938  < 2e-16 ***
AcidIndex          -0.2556911  0.0145751 -17.543  < 2e-16 ***
STARS2              1.5606139  0.0434429  35.923  < 2e-16 ***
STARS3              2.3946959  0.0531663  45.042  < 2e-16 ***
STARS4              3.1481024  0.0934119  33.701  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

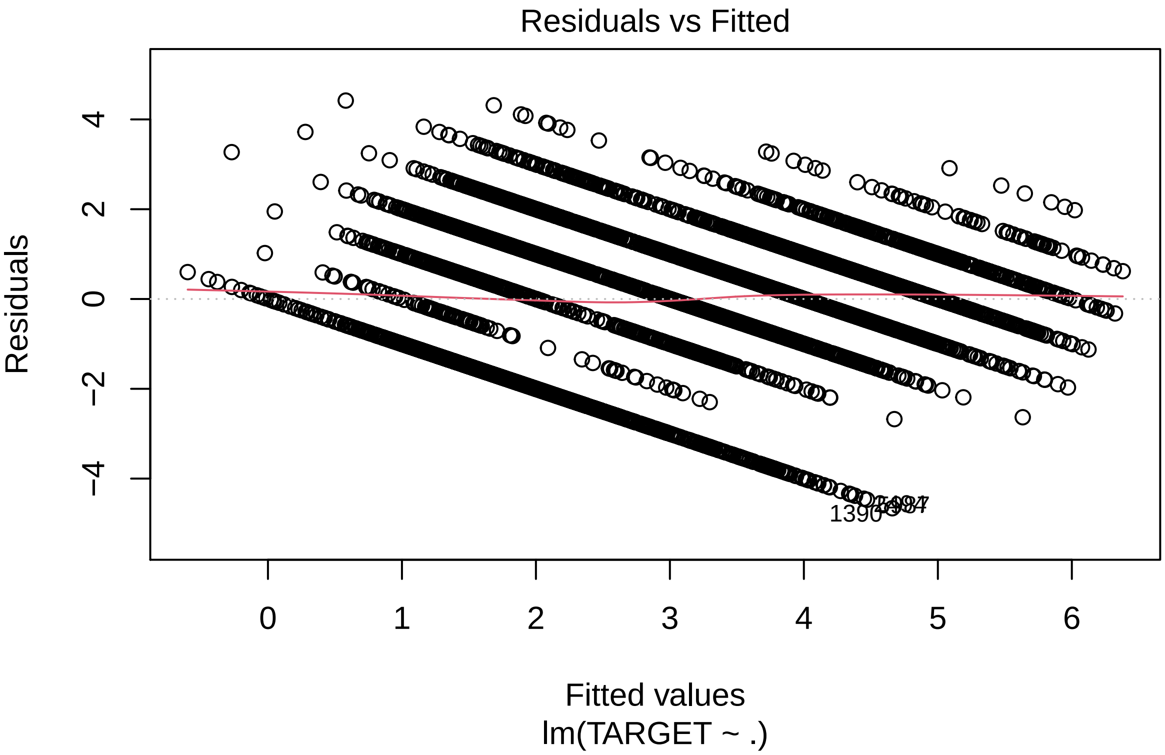
```
Residual standard error: 1.422 on 5875 degrees of freedom
Multiple R-squared:  0.4606,    Adjusted R-squared:  0.4589
F-statistic: 264.1 on 19 and 5875 DF,  p-value: < 2.2e-16
```

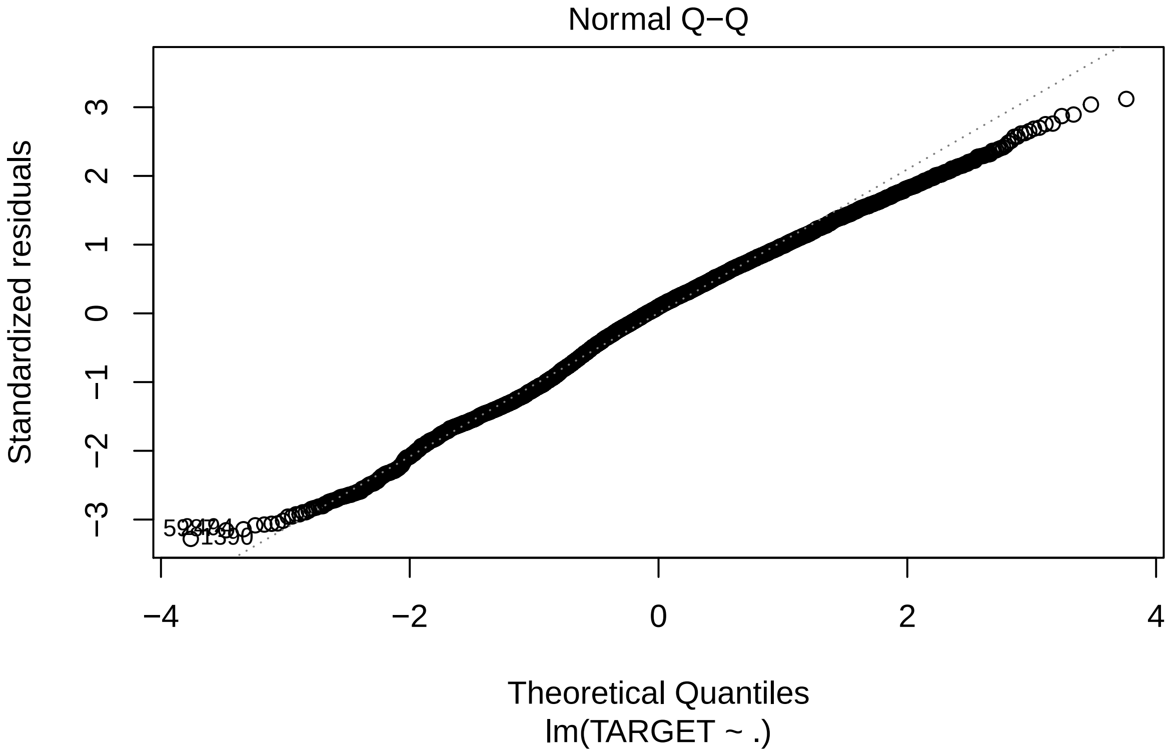
Now we look at the results of using stepAIC

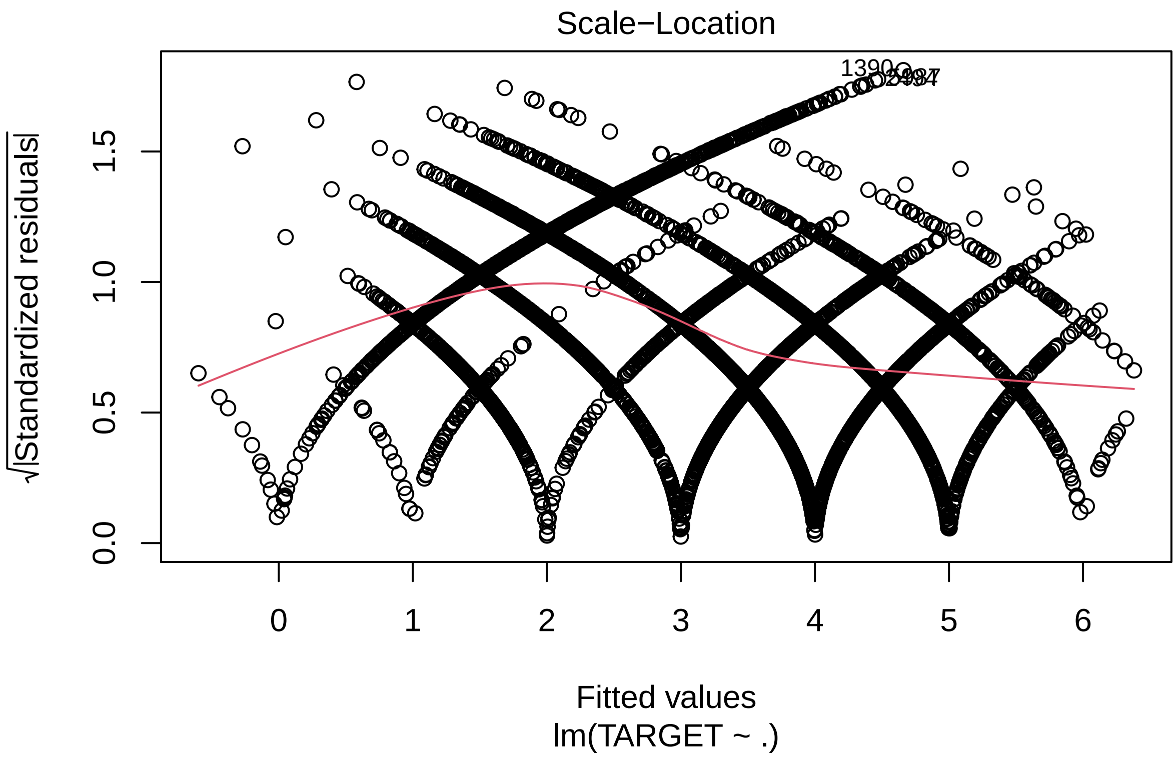
```
Call:
lm(formula = TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide +
    Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex +
    STARS, data = additional_train)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6714 -0.9855  0.1313  1.0131  4.4000

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.9260127  0.7330700   8.084 7.56e-16 ***
VolatileAcidity    -0.0849476  0.0336050  -2.528  0.01150 *
Chlorides          -0.1688963  0.0799054  -2.114  0.03458 *
TotalSulfurDioxide  0.0003396  0.0001121   3.031  0.00245 **
Density            -1.8756677  0.6901792  -2.718  0.00659 **
pH                 -0.3150481  0.0632686  -4.980 6.56e-07 ***
Sulphates          -0.0736993  0.0286299  -2.574  0.01007 *
Alcohol             0.0151022  0.0051893   2.910  0.00362 **
LabelAppeal-1       0.3093298  0.1049597   2.947  0.00322 **
LabelAppeal0        0.7564492  0.1022156   7.401 1.55e-13 ***
LabelAppeal1        1.2212962  0.1061991  11.500  < 2e-16 ***
LabelAppeal2        1.6467358  0.1380937  11.925  < 2e-16 ***
AcidIndex          -0.2517406  0.0143290 -17.569  < 2e-16 ***
STARS2              1.5592938  0.0434220  35.910  < 2e-16 ***
STARS3              2.3934054  0.0531460  45.034  < 2e-16 ***
STARS4              3.1487967  0.0933184  33.743  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.422 on 5879 degrees of freedom
Multiple R-squared:  0.4604,    Adjusted R-squared:  0.459
F-statistic: 334.4 on 15 and 5879 DF,  p-value: < 2.2e-16
```

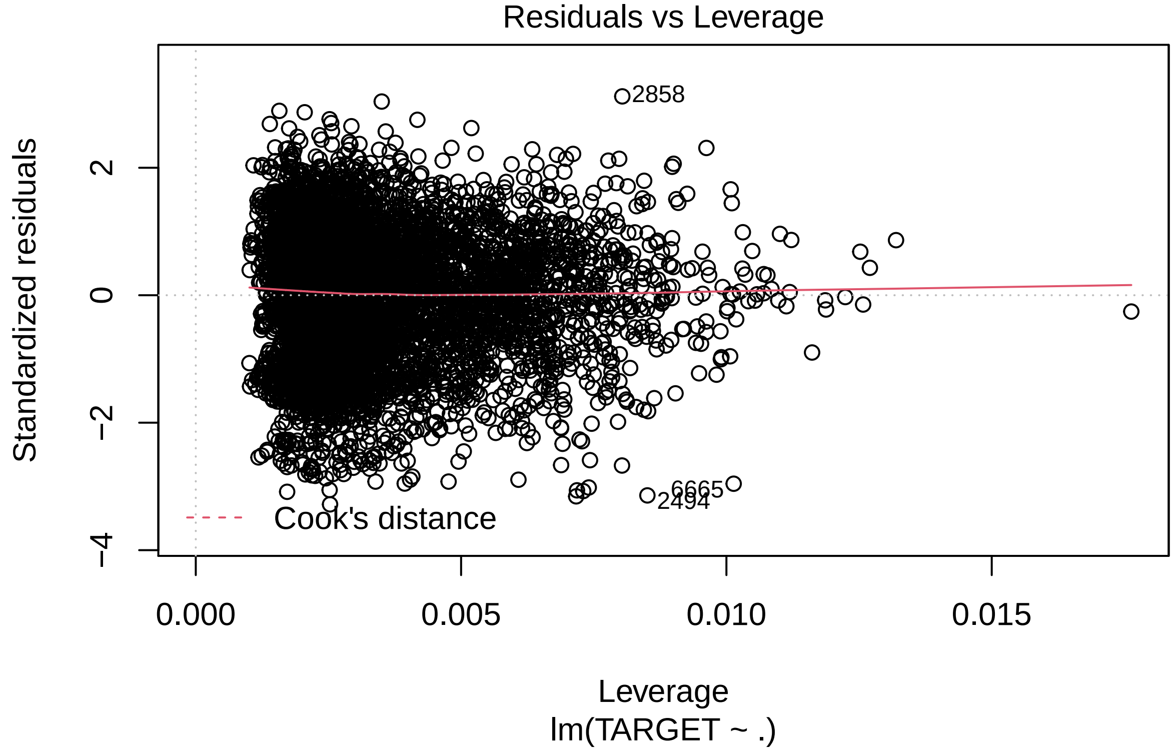
5 Model Selection

We begin by automatically disqualifying our multiple linear regression models. As the predictor variable is not a continuous numeric set of values with a linear relationship, we cannot pass the normality assumption required for linear regression and therefore must rule out these models as being valid.

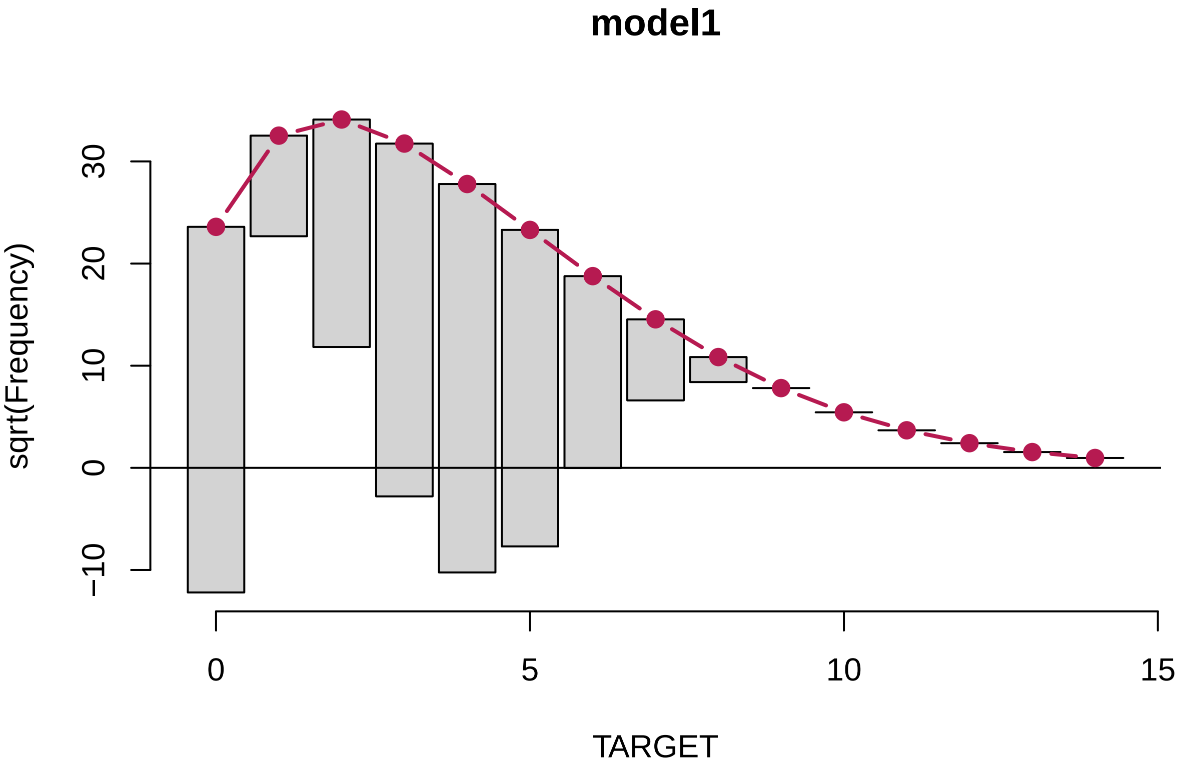


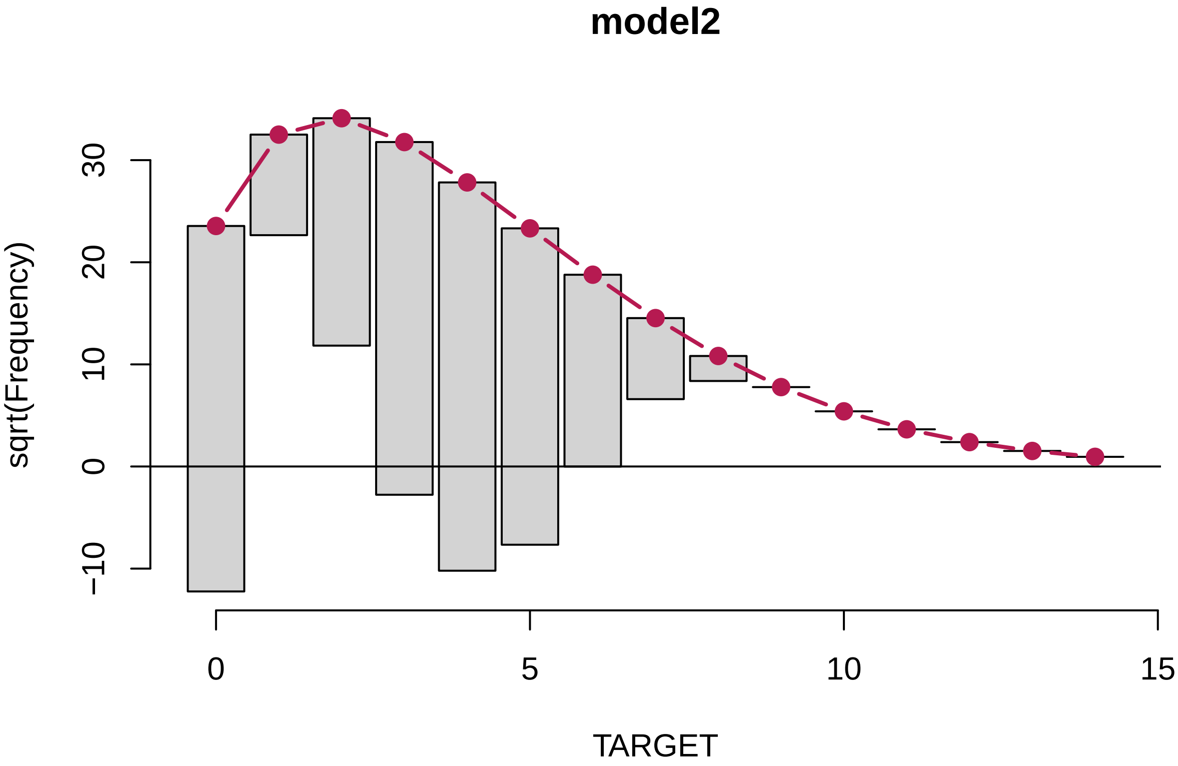


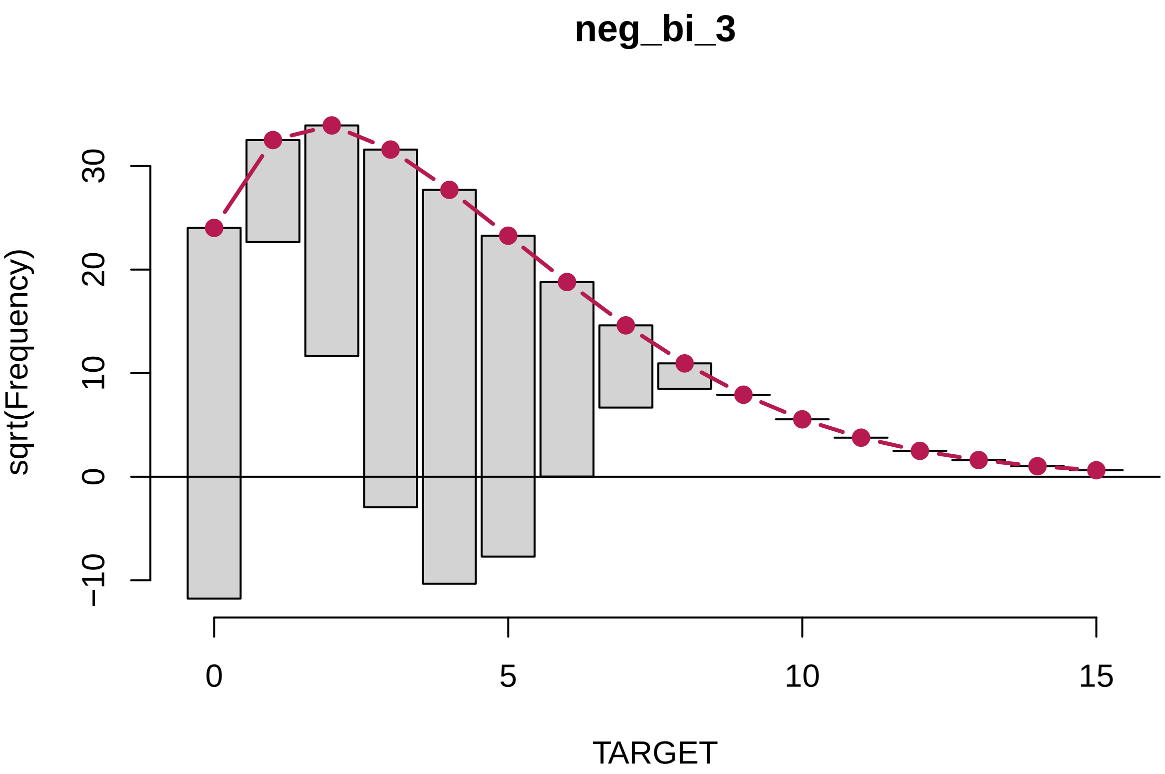




We notice that both the poisson and negative binomal models are hardly any different. All our models appear to struggle in the same areas. There are counts shown below that are severely overfitted as well as other counts that are severely underffiting. We see that all our models also have a very similar Rsquared value when predicting the TARGET value on the evaluation data we created. Overall, it appears that our third negative binomal model, which uses all variables plus some transformed values, had the best RSquared score, but it's only by a tiny margin. We shall use this model for now and look to find ways to improve it in the future.


model1

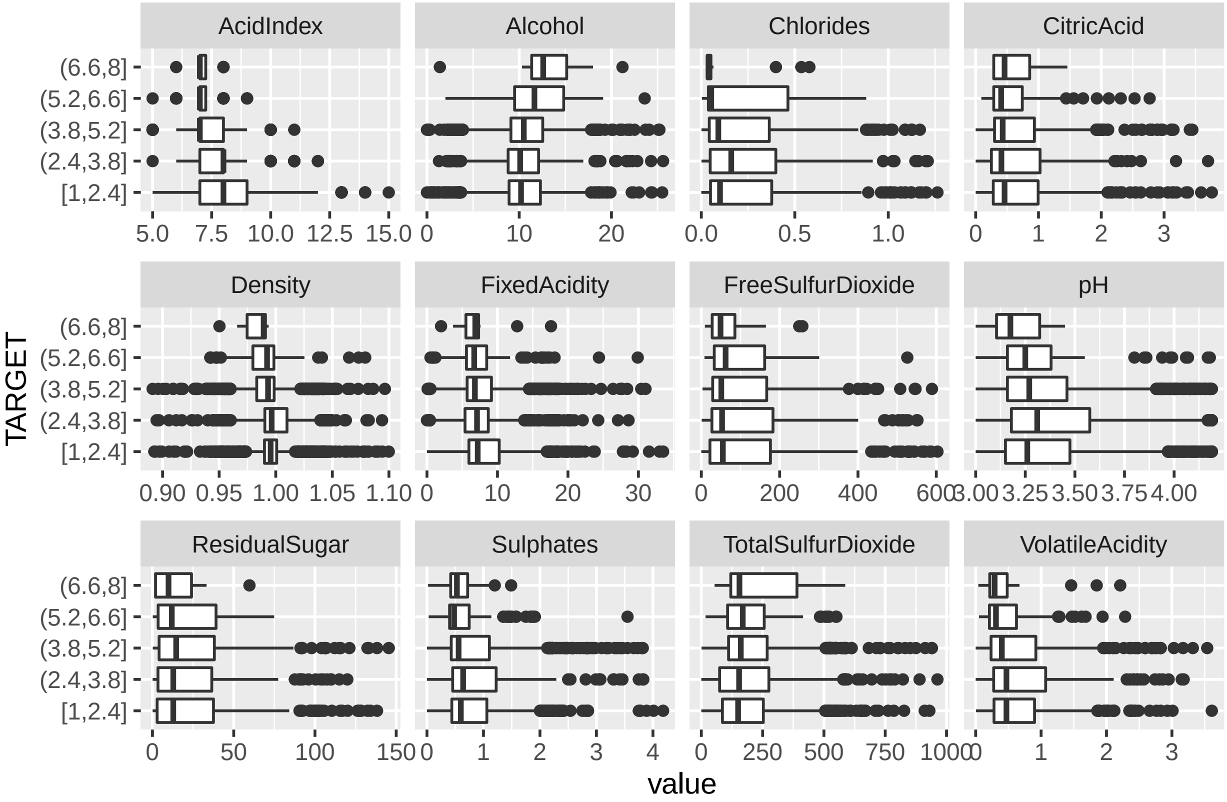

model2



```
     RMSE  Rsquared       MAE 
2.6239727 0.4588123 2.2946415 

     RMSE  Rsquared       MAE 
1.4353639 0.4504921 1.1197462 

     RMSE  Rsquared       MAE 
1.4379859 0.4482697 1.1256939 

     RMSE  Rsquared       MAE 
1.4241321 0.4603157 1.0915940 
```

5.1 Predicting

We compute our 'TARGET' variable on the evaluation data and look at the boxplot produced using it below.



6 Conclusion

This assignment had us work with some truly dirty data. Negative values where there should only be positive values, variables that were clearly impossible with pH values that would either melt your stomach or taste like flavored water. Topping it off, we had an overwhelming number of missing values. We managed to deal with these issues however, it would definitely be worth investigating alternative approaches to the ones used here.

In building our models, we were not able to get the most accurate of scores. It seems apparent from our variable exploration that there were not many significant correlations in our data, and it proved to undermine our ability to accurately predict our target variable. Given more time, we would love to explore alternative modeling techniques that could be used to improve our predictive capability. For now, we are overall very happy with the results we had. Thank you very much for reading our assignment.