621 Assignment5

Raghu

Nov 28, 2018

Contents

1 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:

2 Data Exploration

Data set contains 15 numerical variables and 12,795 observations.

```
## 'data.frame':    12795 obs. of  15 variables:
##  $ TARGET            : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num  -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
##  $ ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num  -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num  NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num  268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num  0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##  $ Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num  9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##  $ LabelAppeal       : int  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : num  2 3 3 1 2 0 0 3 0 4 ...
```

2.1 Data Dictionary

Based on the descriptions, we would expect higher LabelAppeal and STARS vlues correspond with more number of cases purchased. The below variables could be correlated.

-AcidIndex, CitricAcid, FixedAcidity & VolatileAcidity

-FreeSulfurDioxide & TotalSulfurDioxide

-FreeSulfurDioxide, Sulphates & TotalSulfurDioxide

2.2 Summary Statistics

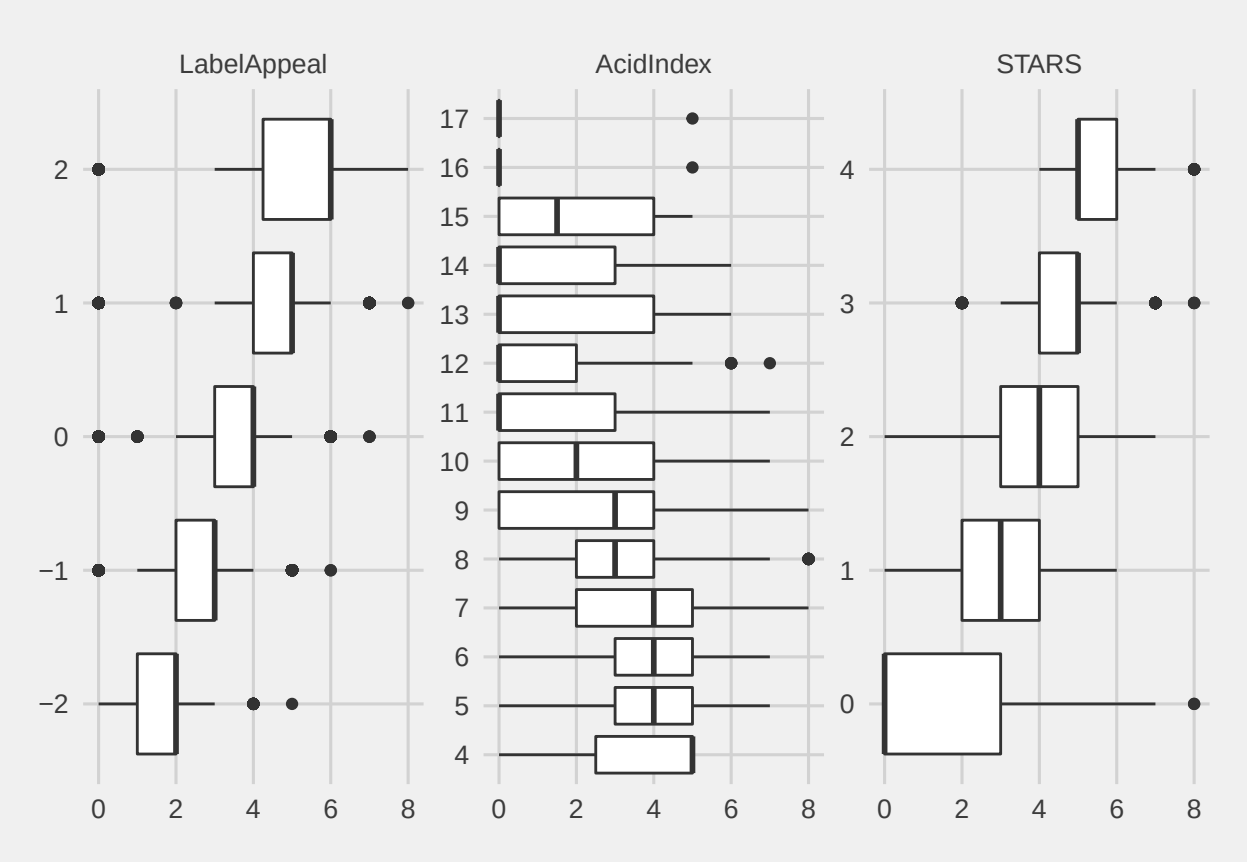
Based on the statistics below, there are missing values, variables have negative values. AcidIndex has more kurtosis than a normal distribution. Along with AcidIndex, LabelAppeal & STARS, the response variable TARGET is discrete.

	n	unique_val	min	Q.1st	median	mean	Q.3rd	max	range	sd	kurtosis
TARGET	12,795	9	0	2	3	3	4	8	8	1.9	-0.9
FixedAcidity	12,795	470	-18.1	5.2	6.9	7.1	9.5	34.4	52.5	6.3	1.7
VolatileAcidity	12,795	815	-2.8	0.1	0.3	0.3	0.6	3.7	6.5	0.8	1.8
CitricAcid	12,795	602	-3.2	0	0.3	0.3	0.6	3.9	7.1	0.9	1.8
ResidualSugar	12,179	2,078	-127.8	-2	3.9	5.4	15.9	141.2	268.9	33.7	1.9
Chlorides	12,157	1,664	-1.2	0	0	0.1	0.2	1.4	2.5	0.3	1.8
FreeSulfurDioxide	12,148	1,000	-555	0	30	30.8	70	623	1,178	148.7	1.8
TotalSulfurDioxide	12,113	1,371	-823	27	123	120.7	208	1,057	1,880	231.9	1.7
Density	12,795	5,933	0.9	1	1	1	1	1.1	0.2	0	1.9
pH	12,400	498	0.5	3	3.2	3.2	3.5	6.1	5.7	0.7	1.6
Sulphates	11,585	631	-3.1	0.3	0.5	0.5	0.9	4.2	7.4	0.9	1.8
Alcohol	12,142	402	-4.7	9	10.4	10.5	12.4	26.5	31.2	3.7	1.5
LabelAppeal	12,795	5	-2	-1	0	0	1	2	4	0.9	-0.3
AcidIndex	12,795	14	4	7	8	7.8	8	17	13	1.3	5.2
STARS	12,795	5	0	0	1	1.5	2	4	4	1.2	-0.9

2.3 Visualization

2.3.1 Boxplots

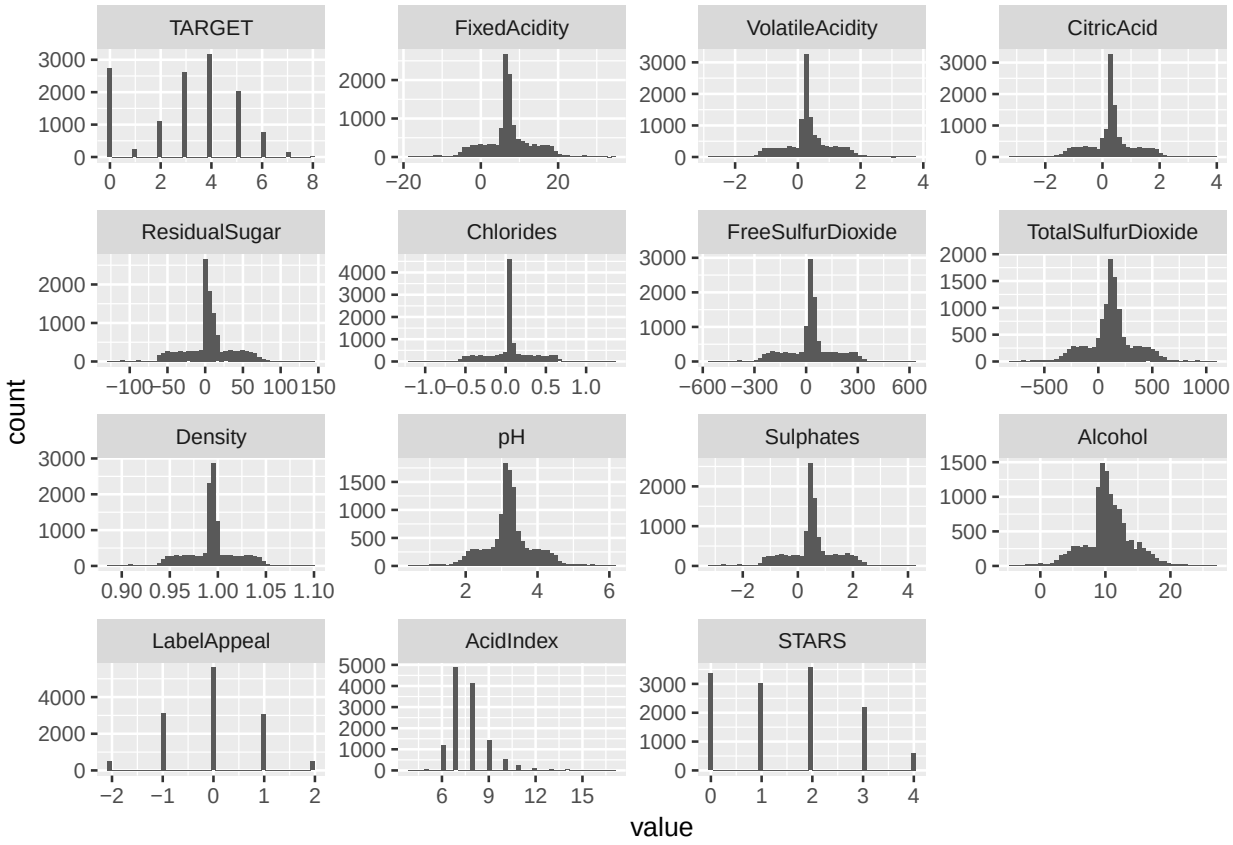
The box plots contain the TARGET distributions for each of the discrete variable values. As we expected, higher values of LabelAppeal and STARS are associated with more wine being purchased. Additionally, smaller AcidIndex values appear to be associated with more wine purchases.



2.3.2 Histogram

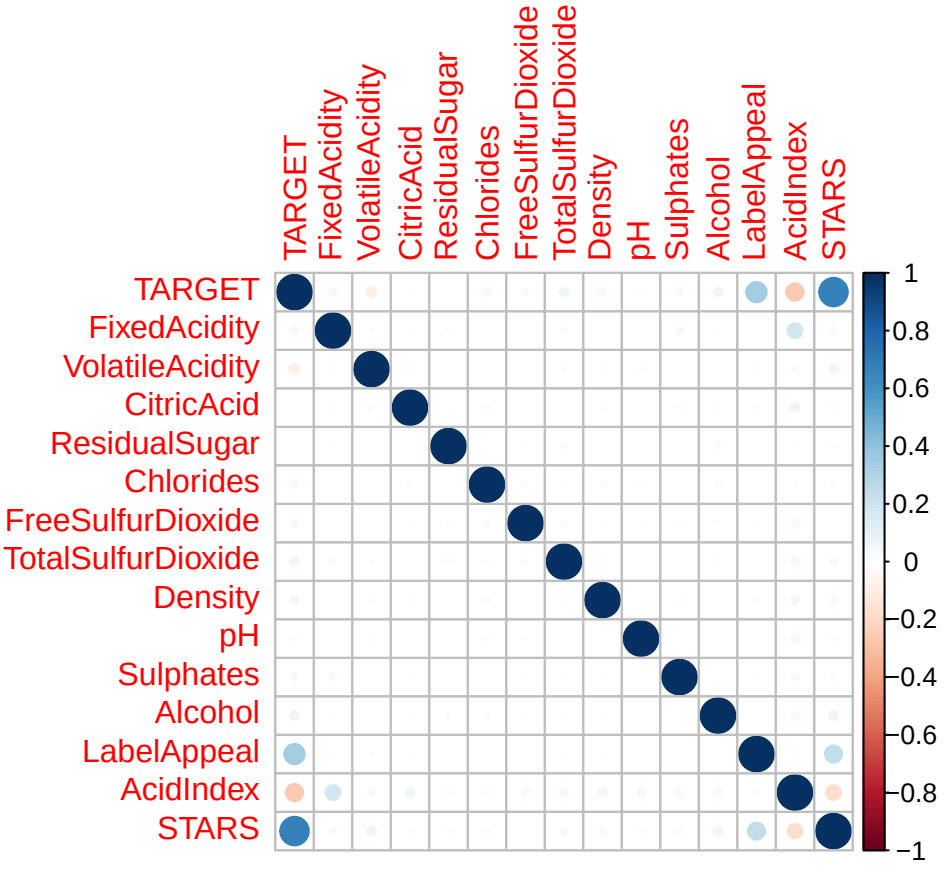
The distributions of the continuous predictor variables have smaller tails and great peaks.

```
## No id variables; using all as measure variables
```


TARGET
FixedAcidity
VolatileAcidity
CitricAcid
ResidualSugar
Chlorides
FreeSulfurDioxide
TotalSulfurDioxide
Density
pH
Sulphates
Alcohol
LabelAppeal
AcidIndex
STARS
count
value

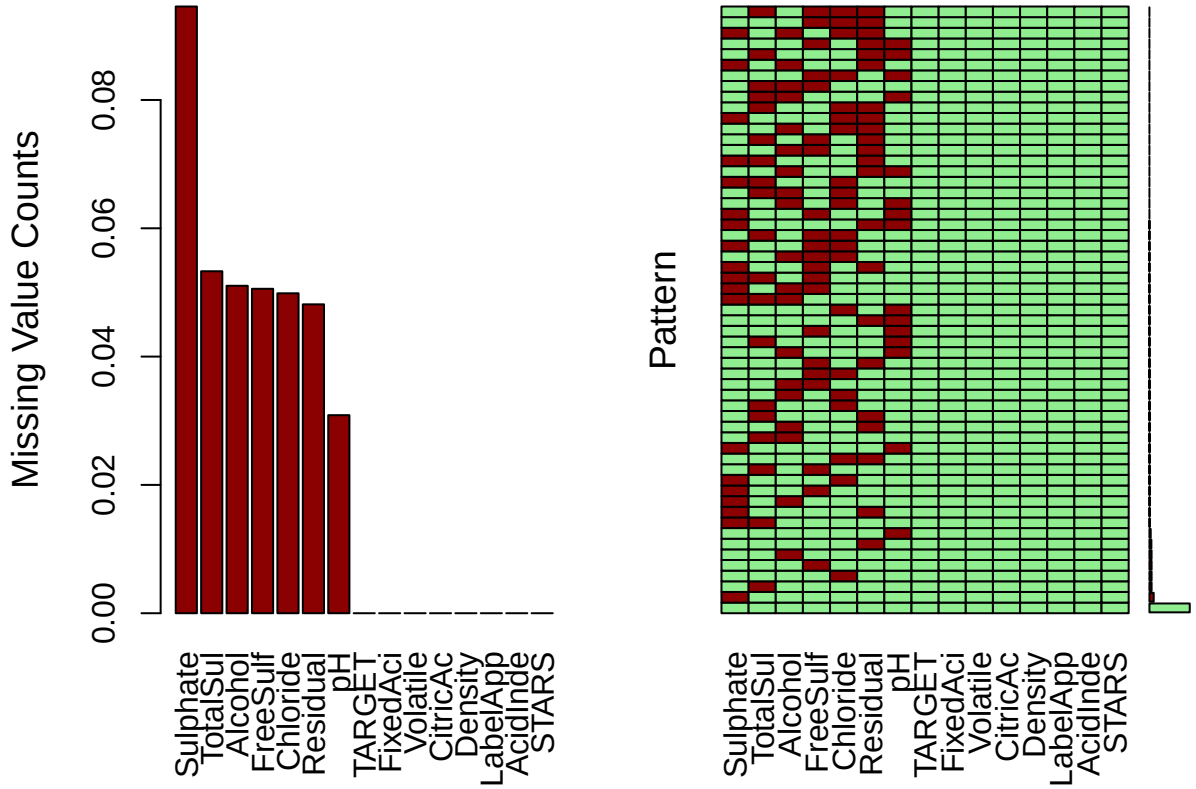
2.3.3 Correlations

There are very few predictor variables are correlated positively. STARS and LabelAppeal have moderate positive correlations with TARGET, and AcidIndex has a slight negative correlation.



2.3.4 Missing Values

In the plots and table below, we can see that 7 variables are missing values and that only 70% of the observations are complete. Since most of the predictors are not correlated with each other, it may be difficult to accurately impute the missing values.




```
##
##  Variables sorted by number of missings:
##  Variable      Count
##  Sulphate 0.09456819
##  TotalSul 0.05330207
##   Alcohol 0.05103556
##  FreeSulf 0.05056663
##  Chloride 0.04986323
##  Residual 0.04814381
##        pH 0.03087143
##    TARGET 0.00000000
##  FixedAci 0.00000000
##  Volatile 0.00000000
##  CitricAc 0.00000000
##   Density 0.00000000
##  LabelApp 0.00000000
##  AcidInde 0.00000000
##     STARS 0.00000000
```

Table 1: Variables Missing Values

	Variable	Count	pct_missing
1	Sulphates	1210	0.095
2	TotalSulfurDioxide	682	0.053
3	Alcohol	653	0.051
4	FreeSulfurDioxide	647	0.051
5	Chlorides	638	0.050
6	ResidualSugar	616	0.048
7	pH	395	0.031

3 DATA PREPARATION

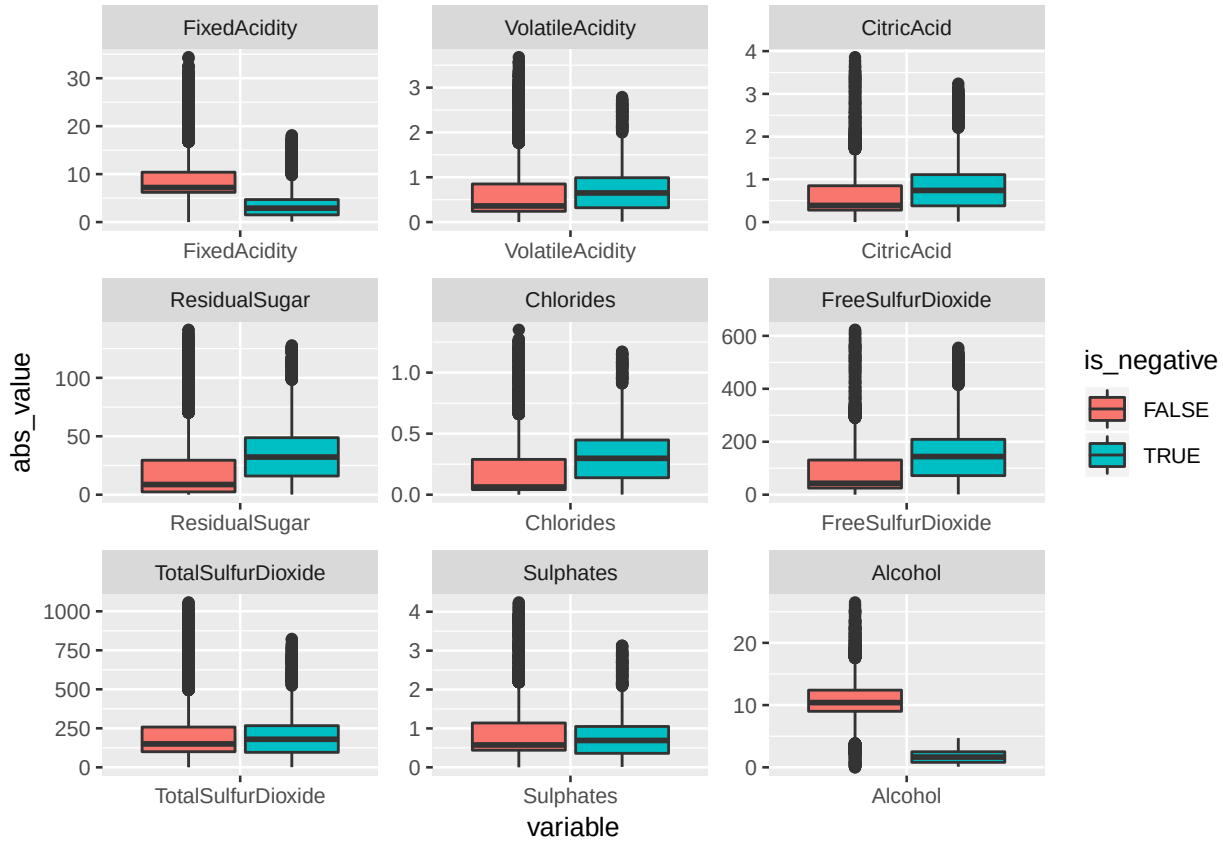
3.1 Variable Transformations

Based on summary statistics, following 9 varibles have negative values, and the table below shows that 8 variables contain more than 10% negative values. This many invalid variable values certainly raises concerns about the study's quality of the data collection and measurement.

Var	is_negative
Chlorides	0.26
ResidualSugar	0.26
FreeSulfurDioxide	0.25
CitricAcid	0.23
VolatileAcidity	0.22
TotalSulfurDioxide	0.21
Sulphates	0.20
FixedAcidity	0.13
Alcohol	0.01

In the side-by-side boxplots below, we see that if we were to take the absolute value of the negative numbers, the distributions of the transformed negative values are mostly similar to the positive value distributions. The 2 variables with dissimilar distributions FixedAcidity and Alcohol have the fewest negative values. Consequently, we will take the absolute values of these variables.

```
## Using  as id variables
```


FixedAcidity
VolatileAcidity
CitricAcid
ResidualSugar
Chlorides
FreeSulfurDioxide
TotalSulfurDioxide
Sulphates
Alcohol
abs_value
variable
is_negative
FALSE
TRUE

```
## [1] 16000
```

3.1.1 Imputing the Missing Values

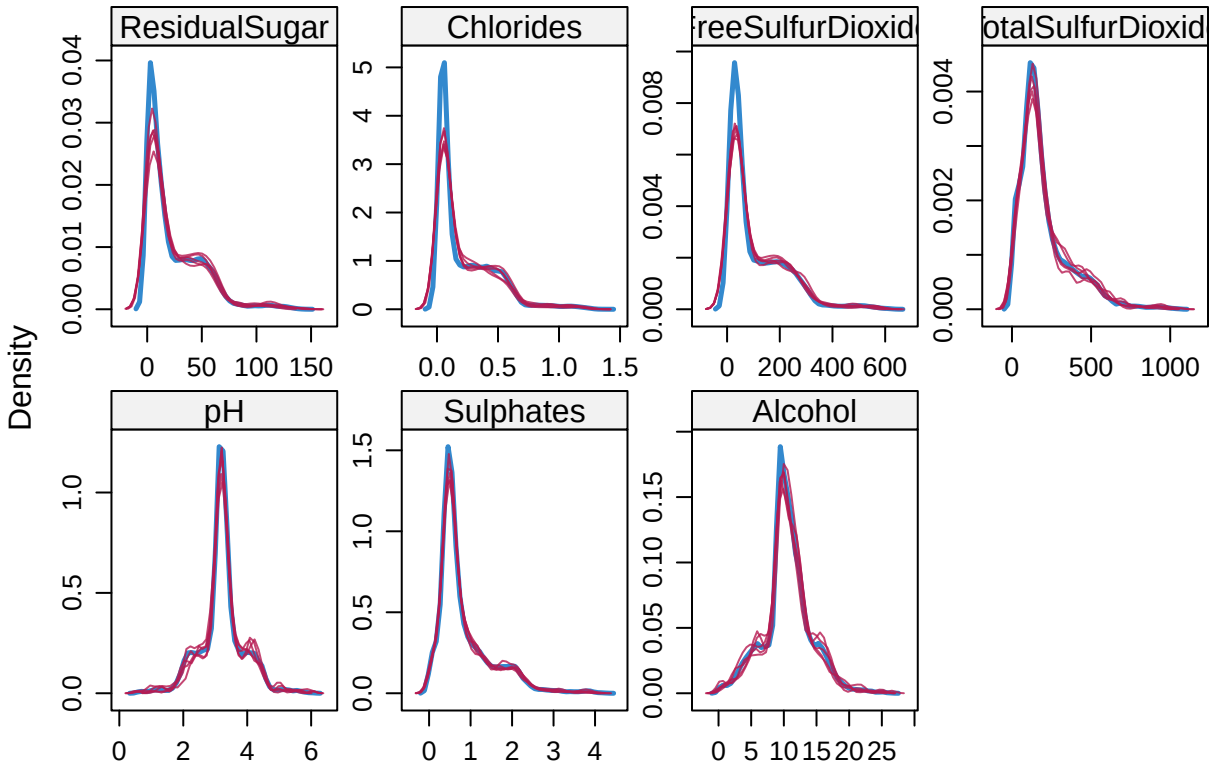
The mice() function takes care of the imputing process.m=5 refers to the number of imputed datasets. Five is the default value. meth='pmm' refers to the imputation method. The missing values have been replaced with the imputed values in the first of the five datasets.

```
## Multiply imputed data set
## Call:
## mice(data = train_transformed, m = 5, maxit = 5, printFlag = FALSE,
##     seed = 2525)
## Number of multiple imputations:  5
## Missing cells per column:
##             TARGET       FixedAcidity    VolatileAcidity
##                  0                  0                  0
##         CitricAcid      ResidualSugar          Chlorides
##                  0                616                638
##  FreeSulfurDioxide TotalSulfurDioxide            Density
##                647                682                  0
##                 pH          Sulphates            Alcohol
##                395               1210                653
##        LabelAppeal          AcidIndex              STARS
##                  0                  0                  0
## Imputation methods:
##             TARGET       FixedAcidity    VolatileAcidity
##                 ""                 ""                 ""
##         CitricAcid      ResidualSugar          Chlorides
##                 ""              "pmm"              "pmm"
##  FreeSulfurDioxide TotalSulfurDioxide            Density
##              "pmm"              "pmm"                 ""
##                 pH          Sulphates            Alcohol
##              "pmm"              "pmm"              "pmm"
##        LabelAppeal          AcidIndex              STARS
##                 ""                 ""                 ""
## VisitSequence:
##      ResidualSugar          Chlorides  FreeSulfurDioxide
##                  5                  6                  7
## TotalSulfurDioxide                 pH          Sulphates
##                  8                 10                 11
##            Alcohol
##                 12
## PredictorMatrix:
##                    TARGET FixedAcidity VolatileAcidity CitricAcid
## TARGET                  0            0                0          0
## FixedAcidity            0            0                0          0
## VolatileAcidity         0            0                0          0
## CitricAcid              0            0                0          0
## ResidualSugar           1            1                1          1
## Chlorides               1            1                1          1
## FreeSulfurDioxide       1            1                1          1
## TotalSulfurDioxide      1            1                1          1
## Density                 0            0                0          0
## pH                      1            1                1          1
## Sulphates               1            1                1          1
## Alcohol                 1            1                1          1
```

```
## LabelAppeal             0            0               0          0
## AcidIndex               0            0               0          0
## STARS                   0            0               0          0
##                    ResidualSugar Chlorides FreeSulfurDioxide
## TARGET                         0         0                 0
## FixedAcidity                   0         0                 0
## VolatileAcidity                0         0                 0
## CitricAcid                     0         0                 0
## ResidualSugar                  0         1                 1
## Chlorides                      1         0                 1
## FreeSulfurDioxide              1         1                 0
## TotalSulfurDioxide             1         1                 1
## Density                        0         0                 0
## pH                             1         1                 1
## Sulphates                      1         1                 1
## Alcohol                        1         1                 1
## LabelAppeal                    0         0                 0
## AcidIndex                      0         0                 0
## STARS                          0         0                 0
##                    TotalSulfurDioxide Density pH Sulphates Alcohol
## TARGET                              0       0  0         0       0
## FixedAcidity                        0       0  0         0       0
## VolatileAcidity                     0       0  0         0       0
## CitricAcid                          0       0  0         0       0
## ResidualSugar                       1       1  1         1       1
## Chlorides                           1       1  1         1       1
## FreeSulfurDioxide                   1       1  1         1       1
## TotalSulfurDioxide                  0       1  1         1       1
## Density                             0       0  0         0       0
## pH                                  1       1  0         1       1
## Sulphates                           1       1  1         0       1
## Alcohol                             1       1  1         1       0
## LabelAppeal                         0       0  0         0       0
## AcidIndex                           0       0  0         0       0
## STARS                               0       0  0         0       0
##                    LabelAppeal AcidIndex STARS
## TARGET                       0         0     0
## FixedAcidity                 0         0     0
## VolatileAcidity              0         0     0
## CitricAcid                   0         0     0
## ResidualSugar                1         1     1
## Chlorides                    1         1     1
## FreeSulfurDioxide            1         1     1
## TotalSulfurDioxide           1         1     1
## Density                      0         0     0
## pH                           1         1     1
## Sulphates                    1         1     1
## Alcohol                      1         1     1
## LabelAppeal                  0         0     0
## AcidIndex                    0         0     0
## STARS                        0         0     0
## Random generator seed value:  2525
```

3.1.2 Density Plot

The density of the imputed data for each imputed dataset is showed in magenta while the density of the observed data is showed in blue.



4 BUILD MODELS

Using the training data set, build at least two different poisson regression models, at least two different negative binomial regression models, and at least two multiple linear regression models, using different variables

4.1 Linear Models

4.1.1 Backward Elimination

For our first model, let's use all variables in our imputed data set with a backward elimination process that removes the predictor with the highest p-value until all of the remaining p-values are statistically significant at a .05 level. `LabelAppeal`, has the most practical significance in the model. With the other variables held constant, for every 1 point increase in the expert wine rating, we would expect an increase of 9.8 wine cases purchased.

```
##
## \begin{tabular}{l|r}
## \hline
## removed\_vars & removed\_pvalues\\
## \hline
## FixedAcidity & 0.997\\
## \hline
## ResidualSugar & 0.907\\
## \hline
## FreeSulfurDioxide & 0.080\\
## \hline
## CitricAcid & 0.073\\
## \hline
## Chlorides & 0.068\\
## \hline
## pH & 0.054\\
## \hline
## \end{tabular}
```

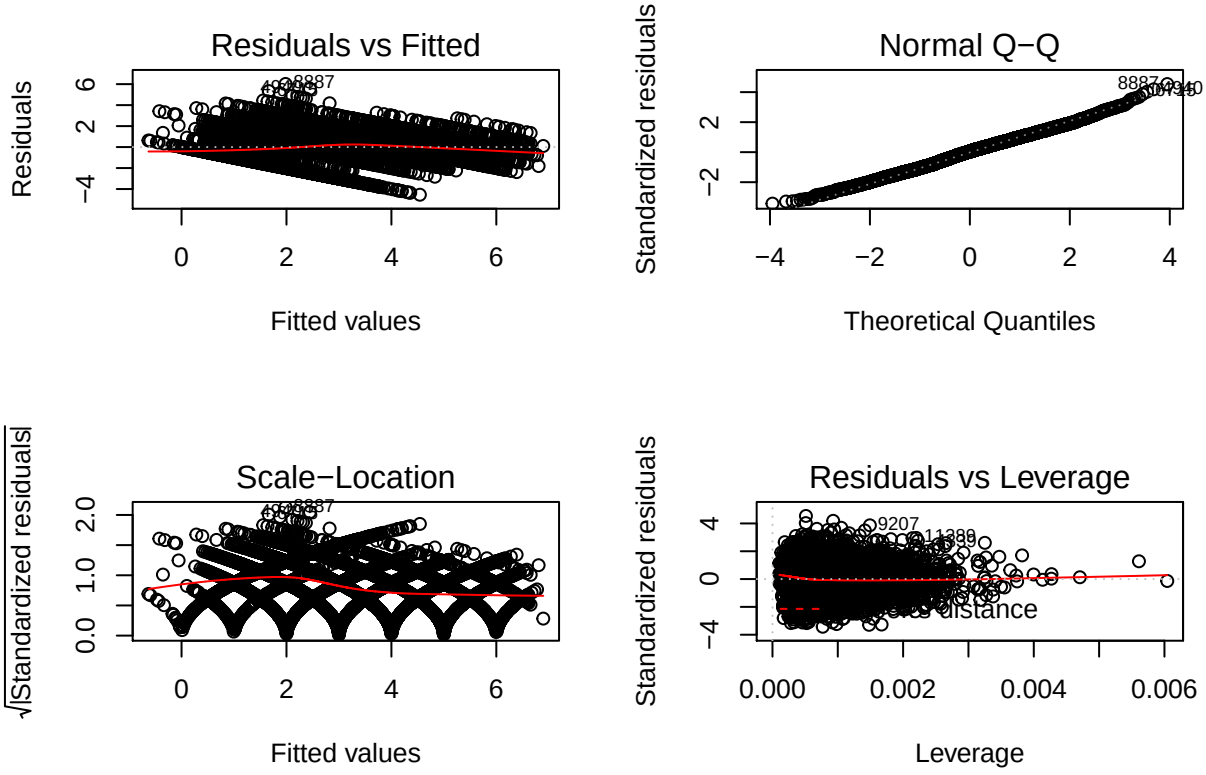
```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + TotalSulfurDioxide +
##     Density + Sulphates + Alcohol + LabelAppeal + AcidIndex +
##     STARS, data = train_imputed, x = T, y = T)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5430 -0.9615  0.0619  0.9142  6.0177
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.978e+00  4.460e-01   8.921  < 2e-16 ***
## VolatileAcidity     -1.165e-01  2.115e-02  -5.505 3.77e-08 ***
## TotalSulfurDioxide   2.846e-04  7.207e-05   3.949 7.90e-05 ***
## Density             -8.708e-01  4.424e-01  -1.968 0.049071 *
## Sulphates           -4.541e-02  1.798e-02  -2.526 0.011563 *
## Alcohol              1.195e-02  3.240e-03   3.688 0.000227 ***
## LabelAppeal          4.326e-01  1.369e-02  31.600  < 2e-16 ***
## AcidIndex           -2.093e-01  9.036e-03 -23.164  < 2e-16 ***
## STARS                9.804e-01  1.046e-02  93.771  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.326 on 12786 degrees of freedom
## Multiple R-squared:  0.5262, Adjusted R-squared:  0.5259
## F-statistic:  1775 on 8 and 12786 DF,  p-value: < 2.2e-16
```

Table 2: Model Summary Statistics

model_name	n_vars	numdf	fstat	p.value	adj.r.squared	pre.r.squared	CV_RMSE
bkwd_elim_lmod	8	8	1774.984	0.00e+00	0.526	0.526	1.324

Table 3: Model Diagnostic Statistics

DW.test	NCV.test	AD.test	VIF_gt_4
0.838	2.97e-133	2.49e-21	0



4.1.2 BIC Selection

Now let's use the original variables of the imputed data set with a BIC selection. Due to its high predictor penalty, this process removed 2 more variables than the backward elimination process, leaving the model with 6 statistically significant variables.

```
## Start:  AIC=7347.84
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Sum of Sq   RSS     AIC
## - FixedAcidity         1       0.0 22471  7338.4
## - ResidualSugar        1       0.0 22471  7338.4
## - FreeSulfurDioxide    1       5.4 22477  7341.5
## - CitricAcid           1       5.6 22477  7341.6
## - Chlorides            1       5.7 22477  7341.6
## - Density              1       6.4 22478  7342.0
## - pH                   1       6.5 22478  7342.1
## - Sulphates            1      11.0 22482  7344.6
## <none>                             22471  7347.8
## - Alcohol              1      24.0 22495  7352.0
## - TotalSulfurDioxide   1      26.6 22498  7353.5
## - VolatileAcidity      1      52.0 22523  7368.0
## - AcidIndex            1     917.6 23389  7850.5
## - LabelAppeal          1    1752.1 24223  8299.0
## - STARS                1   15431.9 37903 14027.6
##
## Step:  AIC=7338.38
## TARGET ~ VolatileAcidity + CitricAcid + ResidualSugar + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Sum of Sq   RSS     AIC
## - ResidualSugar        1       0.0 22471  7328.9
## - FreeSulfurDioxide    1       5.4 22477  7332.0
## - CitricAcid           1       5.6 22477  7332.1
## - Chlorides            1       5.7 22477  7332.2
## - Density              1       6.4 22478  7332.6
## - pH                   1       6.5 22478  7332.6
## - Sulphates            1      11.0 22482  7335.2
## <none>                             22471  7338.4
## - Alcohol              1      24.0 22495  7342.6
## - TotalSulfurDioxide   1      26.6 22498  7344.0
## - VolatileAcidity      1      52.0 22523  7358.5
## - AcidIndex            1     946.6 23418  7856.9
## - LabelAppeal          1    1752.1 24223  8289.6
## - STARS                1   15433.3 37905 14018.6
##
## Step:  AIC=7328.94
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS
##
```

```
##                       Df Sum of Sq   RSS     AIC
## - FreeSulfurDioxide    1       5.4 22477  7322.6
## - CitricAcid           1       5.6 22477  7322.7
## - Chlorides            1       5.7 22477  7322.7
## - Density              1       6.4 22478  7323.1
## - pH                   1       6.5 22478  7323.2
## - Sulphates            1      11.0 22482  7325.7
## <none>                             22471  7328.9
## - Alcohol              1      24.0 22495  7333.1
## - TotalSulfurDioxide   1      26.6 22498  7334.6
## - VolatileAcidity      1      52.0 22523  7349.1
## - AcidIndex            1     946.8 23418  7847.5
## - LabelAppeal          1    1752.1 24224  8280.1
## - STARS                1   15433.4 37905 14009.1
##
## Step:  AIC=7322.56
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + TotalSulfurDioxide +
##     Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex +
##     STARS
##
##                       Df Sum of Sq   RSS     AIC
## - CitricAcid           1       5.6 22482  7316.3
## - Chlorides            1       5.8 22483  7316.4
## - Density              1       6.3 22483  7316.7
## - pH                   1       6.5 22483  7316.8
## - Sulphates            1      10.9 22488  7319.3
## <none>                             22477  7322.6
## - Alcohol              1      23.7 22500  7326.6
## - TotalSulfurDioxide   1      26.9 22504  7328.4
## - VolatileAcidity      1      52.3 22529  7342.9
## - AcidIndex            1     950.0 23427  7842.8
## - LabelAppeal          1    1754.5 24231  8274.8
## - STARS                1   15438.6 37915 14003.3
##
## Step:  AIC=7316.31
## TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Sum of Sq   RSS     AIC
## - Chlorides            1       5.9 22488  7310.2
## - pH                   1       6.5 22489  7310.5
## - Density              1       6.5 22489  7310.5
## - Sulphates            1      10.6 22493  7312.9
## <none>                             22482  7316.3
## - Alcohol              1      23.6 22506  7320.3
## - TotalSulfurDioxide   1      27.2 22510  7322.3
## - VolatileAcidity      1      52.4 22535  7336.7
## - AcidIndex            1     945.8 23428  7834.1
## - LabelAppeal          1    1757.4 24240  8269.8
## - STARS                1   15446.4 37929 13998.4
##
## Step:  AIC=7310.19
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + Density + pH +
##     Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
```

```
##
##                       Df Sum of Sq   RSS     AIC
## - pH                   1       6.5 22495  7304.5
## - Density              1       6.7 22495  7304.5
## - Sulphates            1      11.0 22499  7307.0
## <none>                             22488  7310.2
## - Alcohol              1      23.6 22512  7314.2
## - TotalSulfurDioxide   1      27.6 22516  7316.4
## - VolatileAcidity      1      52.6 22541  7330.7
## - AcidIndex            1     950.1 23438  7830.2
## - LabelAppeal          1    1759.1 24247  8264.4
## - STARS                1   15449.8 37938 13992.0
##
## Step:  AIC=7304.46
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + Density + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Sum of Sq   RSS     AIC
## - Density              1       6.8 22502  7298.9
## - Sulphates            1      11.2 22506  7301.4
## <none>                             22495  7304.5
## - Alcohol              1      23.9 22519  7308.6
## - TotalSulfurDioxide   1      27.4 22522  7310.6
## - VolatileAcidity      1      53.3 22548  7325.3
## - AcidIndex            1     944.0 23439  7821.0
## - LabelAppeal          1    1756.8 24252  8257.2
## - STARS                1   15469.7 37965 13991.5
##
## Step:  AIC=7298.88
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS
##
##                       Df Sum of Sq   RSS     AIC
## - Sulphates            1      11.4 22513  7295.9
## <none>                             22502  7298.9
## - Alcohol              1      24.0 22526  7303.1
## - TotalSulfurDioxide   1      27.0 22529  7304.7
## - VolatileAcidity      1      53.1 22555  7319.6
## - AcidIndex            1     951.3 23453  7819.2
## - LabelAppeal          1    1757.9 24260  8251.9
## - STARS                1   15486.3 37988 13989.9
##
## Step:  AIC=7295.89
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + Alcohol + LabelAppeal +
##     AcidIndex + STARS
##
##                       Df Sum of Sq   RSS     AIC
## <none>                             22513  7295.9
## - Alcohol              1      24.0 22537  7300.1
## - TotalSulfurDioxide   1      27.3 22540  7301.9
## - VolatileAcidity      1      53.5 22567  7316.8
## - AcidIndex            1     957.8 23471  7819.5
## - LabelAppeal          1    1756.2 24269  8247.5
## - STARS                1   15503.3 38016 13990.0
```

```
## [1] "removed variable(s): 2"
## [1] "Density"   "Sulphates"

##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + TotalSulfurDioxide +
##     Alcohol + LabelAppeal + AcidIndex + STARS, data = train_imputed,
##     x = T, y = T)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5925 -0.9613  0.0648  0.9111  6.0069
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.083e+00  8.603e-02  35.839  < 2e-16 ***
## VolatileAcidity    -1.166e-01  2.116e-02  -5.512 3.61e-08 ***
## TotalSulfurDioxide  2.836e-04  7.208e-05   3.935 8.37e-05 ***
## Alcohol             1.197e-02  3.241e-03   3.694 0.000222 ***
## LabelAppeal         4.325e-01  1.369e-02  31.585  < 2e-16 ***
## AcidIndex          -2.106e-01  9.029e-03 -23.325  < 2e-16 ***
## STARS               9.812e-01  1.046e-02  93.842  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.327 on 12788 degrees of freedom
## Multiple R-squared:  0.5258, Adjusted R-squared:  0.5256
## F-statistic:  2363 on 6 and 12788 DF,  p-value: < 2.2e-16

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.6305  1.8870  2.9810  3.0290  4.0380  6.8950
```

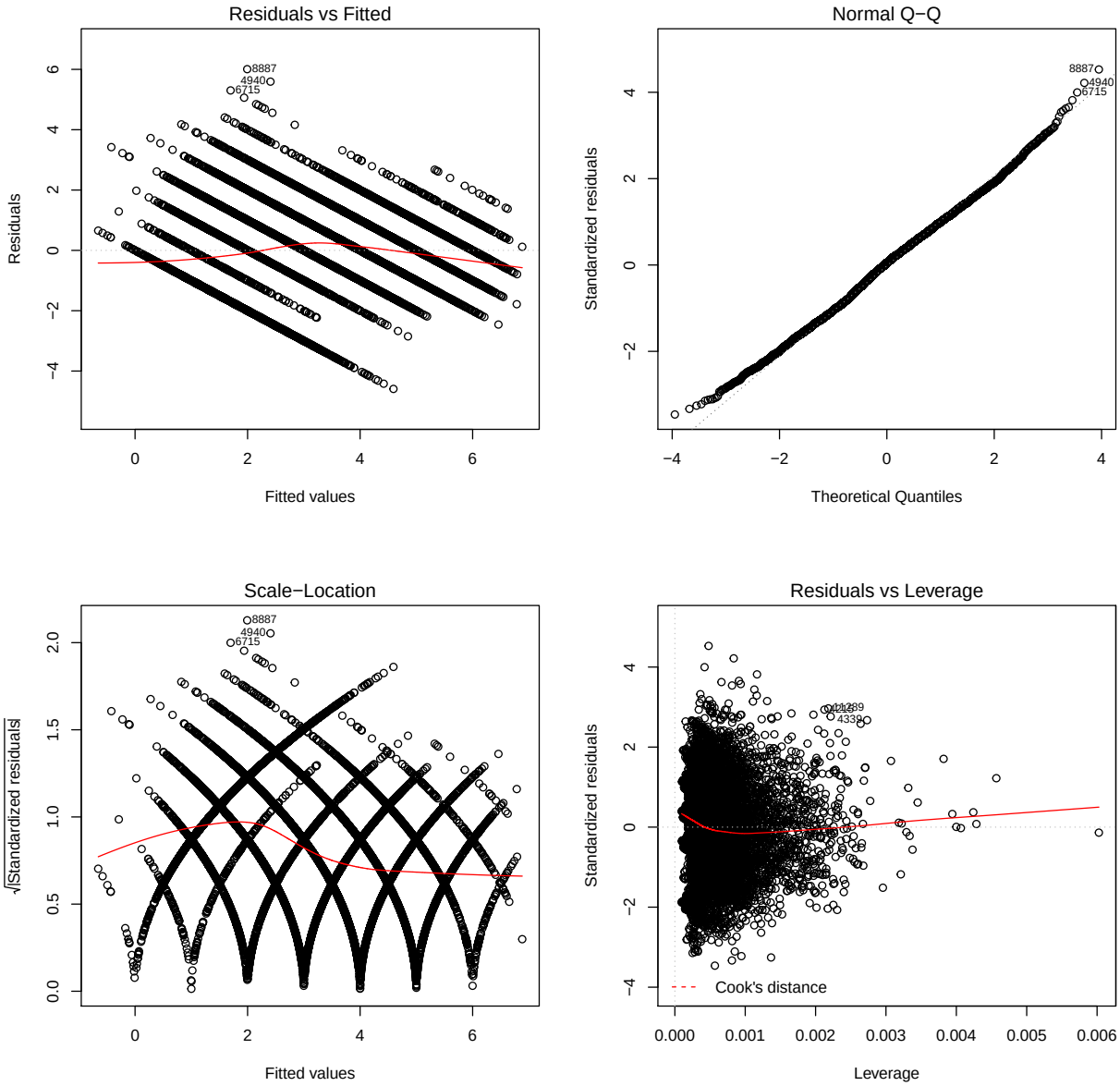
The BIC model's diagnostic plots closely resemble the plots from the backward elimination mode. While the Residuals-vs-Fitted plot displays what appears to be constant variance given the discrete response variable, the standardized-residual plot reveals some nonconstant variance along the fitted residuals. The Normal Q-Q plot shows that the standardized residuals are close to normality. The Leverage plot shows that we do not have any influential points.

Table 4: Model Summary Statistics

model_name	n_vars	numdf	fstat	p.value	adj.r.squared	pre.r.squared	CV_RMSE
bkwd_elim_lmod	8	8	1774.984	0.00e+00	0.526	0.526	1.324
BIC_lmod	6	6	2363.382	0.00e+00	0.526	0.525	1.324

Table 5: Model Diagnostic Statistics

DW.test	NCV.test	AD.test	VIF_gt_4
0.838	2.97e-133	2.49e-21	0
0.878	5.39e-134	1.01e-21	0



4.2 Poission Regression

4.2.1 Regular Poisson Model with BIC Selection

Let's use Poisson regression with the BIC variable selection process. This process removed 9 variables creating the most parsimonious model so far with 5 statistically significant features.

```
## Start:  AIC=46833.44
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - FixedAcidity         1    14750 46824
## - ResidualSugar        1    14750 46824
## - CitricAcid           1    14752 46826
## - FreeSulfurDioxide    1    14752 46826
## - Density              1    14752 46826
## - Chlorides            1    14752 46826
## - Alcohol              1    14753 46827
## - pH                   1    14754 46828
## - Sulphates            1    14755 46829
## <none>                      14750 46833
## - TotalSulfurDioxide   1    14760 46835
## - VolatileAcidity      1    14769 46844
## - AcidIndex            1    15140 47214
## - LabelAppeal          1    15232 47306
## - STARS                1    19615 51689
##
## Step:  AIC=46824.17
## TARGET ~ VolatileAcidity + CitricAcid + ResidualSugar + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - ResidualSugar        1    14750 46815
## - CitricAcid           1    14752 46817
## - FreeSulfurDioxide    1    14752 46817
## - Density              1    14752 46817
## - Chlorides            1    14752 46817
## - Alcohol              1    14753 46818
## - pH                   1    14754 46819
## - Sulphates            1    14755 46820
## <none>                      14750 46824
## - TotalSulfurDioxide   1    14761 46826
## - VolatileAcidity      1    14770 46835
## - AcidIndex            1    15152 47217
## - LabelAppeal          1    15232 47297
## - STARS                1    19615 51680
##
## Step:  AIC=46814.9
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS
```

```
##
##                       Df Deviance   AIC
## - CitricAcid           1    14752 46808
## - FreeSulfurDioxide    1    14752 46808
## - Density              1    14752 46808
## - Chlorides            1    14752 46808
## - Alcohol              1    14753 46809
## - pH                   1    14754 46810
## - Sulphates            1    14755 46811
## <none>                      14750 46815
## - TotalSulfurDioxide   1    14761 46816
## - VolatileAcidity      1    14770 46825
## - AcidIndex            1    15152 47207
## - LabelAppeal          1    15232 47288
## - STARS                1    19615 51671
##
## Step:  AIC=46807.52
## TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex +
##     STARS
##
##                       Df Deviance   AIC
## - FreeSulfurDioxide    1    14754 46800
## - Density              1    14754 46801
## - Chlorides            1    14755 46801
## - Alcohol              1    14755 46801
## - pH                   1    14756 46802
## - Sulphates            1    14757 46803
## <none>                      14752 46808
## - TotalSulfurDioxide   1    14763 46809
## - VolatileAcidity      1    14772 46818
## - AcidIndex            1    15152 47199
## - LabelAppeal          1    15235 47281
## - STARS                1    19620 51666
##
## Step:  AIC=46800.42
## TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - Density              1    14757 46793
## - Chlorides            1    14757 46794
## - Alcohol              1    14758 46794
## - pH                   1    14758 46795
## - Sulphates            1    14760 46796
## <none>                      14754 46800
## - TotalSulfurDioxide   1    14766 46802
## - VolatileAcidity      1    14775 46811
## - AcidIndex            1    15156 47192
## - LabelAppeal          1    15238 47275
## - STARS                1    19624 51660
##
## Step:  AIC=46793.34
## TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide + pH +
```

```
##     Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - Chlorides            1    14760 46787
## - Alcohol              1    14760 46787
## - pH                   1    14761 46788
## - Sulphates            1    14762 46789
## <none>                      14757 46793
## - TotalSulfurDioxide   1    14768 46795
## - VolatileAcidity      1    14777 46804
## - AcidIndex            1    15162 47189
## - LabelAppeal          1    15241 47268
## - STARS                1    19631 51658
##
## Step:  AIC=46786.58
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - Alcohol              1    14763 46780
## - pH                   1    14764 46781
## - Sulphates            1    14765 46782
## <none>                      14760 46787
## - TotalSulfurDioxide   1    14771 46788
## - VolatileAcidity      1    14780 46797
## - AcidIndex            1    15166 47184
## - LabelAppeal          1    15244 47262
## - STARS                1    19634 51652
##
## Step:  AIC=46780.38
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + pH + Sulphates +
##     LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - pH                   1    14767 46775
## - Sulphates            1    14768 46776
## <none>                      14763 46780
## - TotalSulfurDioxide   1    14774 46782
## - VolatileAcidity      1    14782 46791
## - AcidIndex            1    15173 47181
## - LabelAppeal          1    15247 47255
## - STARS                1    19676 51684
##
## Step:  AIC=46775.22
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + Sulphates + LabelAppeal +
##     AcidIndex + STARS
##
##                       Df Deviance   AIC
## - Sulphates            1    14772 46771
## <none>                      14767 46775
## - TotalSulfurDioxide   1    14778 46776
## - VolatileAcidity      1    14787 46786
## - AcidIndex            1    15174 47172
## - LabelAppeal          1    15250 47248
```

```
## - STARS               1    19688 51686
##
## Step:  AIC=46771.21
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + LabelAppeal +
##     AcidIndex + STARS
##
##                       Df Deviance   AIC
## <none>                      14772 46771
## - TotalSulfurDioxide  1    14783 46773
## - VolatileAcidity     1    14792 46782
## - AcidIndex           1    15181 47170
## - LabelAppeal         1    15254 47243
## - STARS               1    19698 51688

## [1] "removed variable(s): 9"
## [1] "FixedAcidity"      "CitricAcid"        "ResidualSugar"
## [4] "Chlorides"         "FreeSulfurDioxide" "Density"
## [7] "pH"                "Sulphates"         "Alcohol"
```

Let's exponentiate the model's coefficients in order to make them interpretable in terms of wine cases. With the other variables held constant, for every 1 point increase in the expert wine rating STARS, we would expect on average an increase of 1.37 wine cases to be purchased.

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + TotalSulfurDioxide +
##     LabelAppeal + AcidIndex + STARS, family = "poisson", data = train_imputed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0074  -0.7183   0.0628   0.5755   3.2581
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.224e+00  3.773e-02  32.448  < 2e-16 ***
## VolatileAcidity     -4.176e-02  9.393e-03  -4.447 8.72e-06 ***
## TotalSulfurDioxide   1.027e-04  3.103e-05   3.311 0.000931 ***
## LabelAppeal          1.331e-01  6.062e-03  21.953  < 2e-16 ***
## AcidIndex           -8.813e-02  4.469e-03 -19.718  < 2e-16 ***
## STARS                3.132e-01  4.513e-03  69.402  < 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: 22861  on 12794  degrees of freedom
## Residual deviance: 14772  on 12789  degrees of freedom
## AIC: 46726
##
## Number of Fisher Scoring iterations: 5

## Waiting for profiling to be done...

##                                  2.5 %    97.5 %
## (Intercept)         3.4021742 3.1598519 3.6635793
## VolatileAcidity     0.9590952 0.9415513 0.9768637
```

```
## TotalSulfurDioxide 1.0001027 1.0000417 1.0001634
## LabelAppeal        1.1423455 1.1288527 1.1559992
## AcidIndex          0.9156448 0.9076408 0.9236824
## STARS              1.3677924 1.3557528 1.3799493
```

With a p-value near zero, this 5-variable model is statistically significant when compared to the null hypothesis, but it only explains 35% of the deviance. The p-value for the goodness-of-fit chi-squared test is near zero, which indicates that the model's deviance is not small enough for a good fit. At .85, the model has underdispersion, which indicates that the data exhibit less variation than the Poisson distribution expects. None of the variables are exhibiting collinearity with a variance inflation factor greater than 4 (VIF_gt_4).

model_name	n_vars	pvalue	devianceExpl	GoFtest	dispersion_parameter	VIF_gt_4	CV_RMSE
BIC_pois_mod	5	0.00e+00	0.354	1.59e-32	0.854	0	1.406

4.2.2 Quasi-Poisson Model with BIC Selection

Since the previous Poisson model was underdispersed, let's try applying the quasi-Poisson generalized linear model to those 5 variables.

```
##
## Call:
## glm(formula = BIC_pois_mod$formula, family = "quasipoisson",
##     data = train_imputed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0074  -0.7183   0.0628   0.5755   3.2581
##
## Coefficients:
##                       Estimate  Std. Error t value   Pr(>|t|)
## (Intercept)         1.22441468  0.03486403  35.120    < 2e-16 ***
## VolatileAcidity    -0.04176494  0.00867809  -4.813 0.00000151 ***
## TotalSulfurDioxide  0.00010273  0.00002867   3.583   0.000341 ***
## LabelAppeal         0.13308357  0.00560105  23.760    < 2e-16 ***
## AcidIndex          -0.08812677  0.00412940 -21.341    < 2e-16 ***
## STARS               0.31319806  0.00416953  75.116    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.8536579)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 14772  on 12789  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

model_name	n_vars	pvalue	devianceExpl	GoFtest	dispersion_parameter	VIF_gt_4	CV_RMSE
BIC_pois_mod	5	0.00e+00	0.354	1.59e-32	0.854	0	1.406
quasi_pois_mod	5	0.00e+00	0.354	1.59e-32	0.854	0	1.406

In fact, the Poisson and quasi-Poisson model stats are exactly the same.

4.3 Negative Binomial Regression

4.3.1 BIC selection with Dispersion Parameter of 1

Now let's try negative binomial regression, which can arise out of a generalized Poisson regression. We will start with a disperson parameter of 1, which corresponds to the geometric distribution and use imputed original variables with a BIC selection process. The result of this process is a 7-feature model with the addition of the Sulphates and pH variables.

```
## Start:  AIC=55638.66
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - ResidualSugar        1   6765.6 55629
## - FixedAcidity         1   6765.6 55629
## - Alcohol              1   6765.8 55630
## - Density              1   6766.3 55632
## - CitricAcid           1   6766.3 55632
## - Chlorides            1   6766.5 55632
## - FreeSulfurDioxide    1   6766.5 55632
## - pH                   1   6768.4 55638
## <none>                     6765.6 55639
## - Sulphates            1   6768.9 55640
## - TotalSulfurDioxide   1   6771.1 55647
## - VolatileAcidity      1   6774.1 55657
## - LabelAppeal          1   6855.3 55918
## - AcidIndex            1   6942.5 56199
## - STARS                1   8188.8 60213
##
## Step:  AIC=55629.21
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - FixedAcidity         1   6765.6 55620
## - Alcohol              1   6765.9 55621
## - Density              1   6766.3 55622
## - CitricAcid           1   6766.3 55622
## - Chlorides            1   6766.5 55623
## - FreeSulfurDioxide    1   6766.5 55623
## - pH                   1   6768.4 55629
## <none>                     6765.6 55629
## - Sulphates            1   6768.9 55630
## - TotalSulfurDioxide   1   6771.2 55638
## - VolatileAcidity      1   6774.1 55647
## - LabelAppeal          1   6855.4 55909
## - AcidIndex            1   6942.5 56190
## - STARS                1   8188.9 60204
##
## Step:  AIC=55619.81
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
```

```
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - Alcohol              1   6765.9 55611
## - Density              1   6766.4 55613
## - CitricAcid           1   6766.4 55613
## - Chlorides            1   6766.5 55613
## - FreeSulfurDioxide    1   6766.6 55613
## - pH                   1   6768.5 55620
## <none>                     6765.6 55620
## - Sulphates            1   6769.0 55621
## - TotalSulfurDioxide   1   6771.2 55628
## - VolatileAcidity      1   6774.1 55638
## - LabelAppeal          1   6855.5 55900
## - AcidIndex            1   6947.7 56197
## - STARS                1   8189.1 60196
##
## Step:  AIC=55610.65
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + LabelAppeal +
##     AcidIndex + STARS
##
##                       Df Deviance   AIC
## - Density              1   6766.7 55604
## - CitricAcid           1   6766.7 55604
## - Chlorides            1   6766.8 55604
## - FreeSulfurDioxide    1   6766.8 55604
## - pH                   1   6768.8 55611
## <none>                     6765.9 55611
## - Sulphates            1   6769.3 55612
## - TotalSulfurDioxide   1   6771.4 55619
## - VolatileAcidity      1   6774.4 55628
## - LabelAppeal          1   6855.6 55890
## - AcidIndex            1   6948.9 56190
## - STARS                1   8198.6 60214
##
## Step:  AIC=55601.93
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + pH + Sulphates + LabelAppeal + AcidIndex +
##     STARS
##
##                       Df Deviance   AIC
## - CitricAcid           1   6767.4 55595
## - FreeSulfurDioxide    1   6767.5 55595
## - Chlorides            1   6767.6 55595
## - pH                   1   6769.6 55602
## <none>                     6766.7 55602
## - Sulphates            1   6770.0 55603
## - TotalSulfurDioxide   1   6772.1 55610
## - VolatileAcidity      1   6775.1 55620
## - LabelAppeal          1   6856.3 55881
## - AcidIndex            1   6951.0 56186
## - STARS                1   8200.7 60208
```

```
##
## Step:  AIC=55593.23
## TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     pH + Sulphates + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - FreeSulfurDioxide    1   6768.3 55587
## - Chlorides            1   6768.4 55587
## - pH                   1   6770.3 55593
## <none>                     6767.4 55593
## - Sulphates            1   6770.7 55594
## - TotalSulfurDioxide   1   6772.9 55601
## - VolatileAcidity      1   6775.9 55611
## - LabelAppeal          1   6857.4 55874
## - AcidIndex            1   6951.1 56175
## - STARS                1   8201.5 60202
##
## Step:  AIC=55584.67
## TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide + pH +
##     Sulphates + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - Chlorides            1   6769.3 55578
## - pH                   1   6771.2 55585
## <none>                     6768.3 55585
## - Sulphates            1   6771.6 55586
## - TotalSulfurDioxide   1   6773.9 55593
## - VolatileAcidity      1   6776.9 55603
## - LabelAppeal          1   6858.6 55866
## - AcidIndex            1   6952.7 56169
## - STARS                1   8202.2 60194
##
## Step:  AIC=55576.19
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + pH + Sulphates +
##     LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## <none>                     6769.3 55576
## - pH                   1   6772.2 55576
## - Sulphates            1   6772.7 55578
## - TotalSulfurDioxide   1   6774.9 55585
## - VolatileAcidity      1   6777.8 55594
## - LabelAppeal          1   6859.8 55858
## - AcidIndex            1   6954.5 56163
## - STARS                1   8203.0 60185

##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + TotalSulfurDioxide +
##     pH + Sulphates + LabelAppeal + AcidIndex + STARS, family = negative.binomial(1),
##     data = train_imputed)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
```

```
## -1.82874  -0.39407   0.00033   0.29966   1.75886
##
## Coefficients:
##                        Estimate  Std. Error t value   Pr(>|t|)
## (Intercept)         1.46774939  0.05087278  28.851    < 2e-16 ***
## VolatileAcidity    -0.05518415  0.01058871  -5.212 0.00000019 ***
## TotalSulfurDioxide  0.00015296  0.00003563   4.292 0.00001780 ***
## pH                 -0.02644351  0.00858502  -3.080   0.002073 **
## Sulphates          -0.02956499  0.00898175  -3.292   0.000999 ***
## LabelAppeal         0.11873190  0.00683775  17.364    < 2e-16 ***
## AcidIndex          -0.11799808  0.00477066 -24.734    < 2e-16 ***
## STARS               0.36669016  0.00516451  71.002    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1) family taken to be 0.3104549)
##
##     Null deviance: 9042.5  on 12794  degrees of freedom
## Residual deviance: 6769.3  on 12787  degrees of freedom
## AIC: 55517
##
## Number of Fisher Scoring iterations: 6
```

With a p-value near zero, this 7-variable model is statistically significant when compared to the null hypothesis, but it only explains 25% of the deviance. The p-value for the goodness-of-fit chi-squared test appears to be near 1, which indicates a good fit. None of the variables are exhibiting collinearity with a variance inflation factor greater than 4 (VIF_gt_4).

model_name	n_vars	pvalue	devianceExpl	GoFtest	dispersion_parameter	VIF_gt_4	CV_RMSE
BIC_pois_mod	5	0.00e+00	0.354	1.59e-32	0.854	0	1.406
quasi_pois_mod	5	0.00e+00	0.354	1.59e-32	0.854	0	1.406
BIC_nb_k1_mod	7	0.00e+00	0.251	1.00e+00	0.310	0	1.486

4.3.2 BIC selection with Varying Dispersion Parameter

Finally, let's run a BIC selection process on a negative binomial model where the dispersion parameter is allowed to vary. It will be estimated using the maximum likelihood. The result of this process is the 5-variable model below.

```
## Start:  AIC=46833.72
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - FixedAcidity         1    14749 46824
## - ResidualSugar        1    14749 46824
## - CitricAcid           1    14751 46826
## - FreeSulfurDioxide    1    14751 46827
## - Density              1    14752 46827
## - Chlorides            1    14752 46827
## - Alcohol              1    14752 46828
## - pH                   1    14753 46828
## - Sulphates            1    14754 46830
## <none>                      14749 46834
## - TotalSulfurDioxide   1    14760 46835
## - VolatileAcidity      1    14769 46844
## - AcidIndex            1    15139 47214
## - LabelAppeal          1    15231 47306
## - STARS                1    19614 51689
##
## Step:  AIC=46824.44
## TARGET ~ VolatileAcidity + CitricAcid + ResidualSugar + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - ResidualSugar        1    14749 46815
## - CitricAcid           1    14751 46817
## - FreeSulfurDioxide    1    14752 46817
## - Density              1    14752 46817
## - Chlorides            1    14752 46817
## - Alcohol              1    14752 46818
## - pH                   1    14753 46819
## - Sulphates            1    14754 46820
## <none>                      14749 46824
## - TotalSulfurDioxide   1    14760 46826
## - VolatileAcidity      1    14769 46835
## - AcidIndex            1    15151 47217
## - LabelAppeal          1    15232 47297
## - STARS                1    19615 51680
##
## Step:  AIC=46815.17
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS
##
```

```
##                       Df Deviance   AIC
## - CitricAcid           1    14752 46808
## - FreeSulfurDioxide    1    14752 46808
## - Density              1    14752 46808
## - Chlorides            1    14752 46808
## - Alcohol              1    14753 46809
## - pH                   1    14754 46810
## - Sulphates            1    14755 46811
## <none>                      14749 46815
## - TotalSulfurDioxide   1    14760 46817
## - VolatileAcidity      1    14769 46826
## - AcidIndex            1    15151 47208
## - LabelAppeal          1    15232 47288
## - STARS                1    19615 51671
##
## Step:  AIC=46807.79
## TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex +
##     STARS
##
##                       Df Deviance   AIC
## - FreeSulfurDioxide    1    14754 46801
## - Density              1    14754 46801
## - Chlorides            1    14754 46801
## - Alcohol              1    14755 46802
## - pH                   1    14756 46802
## - Sulphates            1    14757 46803
## <none>                      14752 46808
## - TotalSulfurDioxide   1    14762 46809
## - VolatileAcidity      1    14771 46818
## - AcidIndex            1    15152 47199
## - LabelAppeal          1    15234 47281
## - STARS                1    19619 51666
##
## Step:  AIC=46800.7
## TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - Density              1    14756 46794
## - Chlorides            1    14756 46794
## - Alcohol              1    14757 46794
## - pH                   1    14758 46795
## - Sulphates            1    14759 46796
## <none>                      14754 46801
## - TotalSulfurDioxide   1    14765 46802
## - VolatileAcidity      1    14774 46811
## - AcidIndex            1    15155 47193
## - LabelAppeal          1    15238 47275
## - STARS                1    19623 51660
##
## Step:  AIC=46793.62
## TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide + pH +
##     Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
```

```
##
##                       Df Deviance   AIC
## - Chlorides            1    14759 46787
## - Alcohol              1    14759 46787
## - pH                   1    14760 46788
## - Sulphates            1    14761 46789
## <none>                      14756 46794
## - TotalSulfurDioxide   1    14767 46795
## - VolatileAcidity      1    14776 46804
## - AcidIndex            1    15161 47189
## - LabelAppeal          1    15240 47268
## - STARS                1    19631 51658
##
## Step:  AIC=46786.86
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - Alcohol              1    14762 46781
## - pH                   1    14763 46782
## - Sulphates            1    14764 46783
## <none>                      14759 46787
## - TotalSulfurDioxide   1    14770 46789
## - VolatileAcidity      1    14779 46797
## - AcidIndex            1    15165 47184
## - LabelAppeal          1    15244 47262
## - STARS                1    19634 51652
##
## Step:  AIC=46780.65
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + pH + Sulphates +
##     LabelAppeal + AcidIndex + STARS
##
##                       Df Deviance   AIC
## - pH                   1    14766 46775
## - Sulphates            1    14768 46777
## <none>                      14762 46781
## - TotalSulfurDioxide   1    14773 46782
## - VolatileAcidity      1    14782 46791
## - AcidIndex            1    15172 47181
## - LabelAppeal          1    15246 47255
## - STARS                1    19675 51684
##
## Step:  AIC=46775.49
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + Sulphates + LabelAppeal +
##     AcidIndex + STARS
##
##                       Df Deviance   AIC
## - Sulphates            1    14772 46771
## <none>                      14766 46775
## - TotalSulfurDioxide   1    14777 46777
## - VolatileAcidity      1    14786 46786
## - AcidIndex            1    15173 47172
## - LabelAppeal          1    15249 47249
## - STARS                1    19687 51686
```

```
##
## Step:  AIC=46771.49
## TARGET ~ VolatileAcidity + TotalSulfurDioxide + LabelAppeal +
##     AcidIndex + STARS
##
##                       Df Deviance   AIC
## <none>                      14772 46771
## - TotalSulfurDioxide  1    14783 46773
## - VolatileAcidity     1    14792 46782
## - AcidIndex           1    15180 47170
## - LabelAppeal         1    15254 47244
## - STARS               1    19698 51688
```

With a p-value near zero, this 5-variable model is statistically significant when compared to the null hypothesis, but it only explains 35% of the deviance. The p-value for the goodness-of-fit chi-squared test is near zero, which indicates that the model's deviance is not small enough for a good fit. None of the variables are exhibiting collinearity with a variance inflation factor greater than 4 (VIF_gt_4).

model_name	n_vars	pvalue	devianceExpl	GoFtest	dispersion_parameter	VIF_gt_4	CV_RMSE
BIC_pois_mod	5	0.00e+00	0.354	1.59e-32	0.854	0	1.406
quasi_pois_mod	5	0.00e+00	0.354	1.59e-32	0.854	0	1.406
BIC_nb_k1_mod	7	0.00e+00	0.251	1.00e+00	0.310	0	1.486
BIC_nb_mod	5	0.00e+00	0.354	1.65e-32	0.854	0	1.406

5 SELECT MODELS

5.1 Coefficient Comparison

Let's take a look at the coefficients from our models. In the table below, we see that the intercepts are all approximately 3 to 4 wines cases. The linear coefficients largely align, and the generalized linear coefficients largely align. Among the linear models, there are small differences, but not practical differences. Among the generalized linear models, the Poisson (BIC_pois_mod), quasi-Poisson (quasi_pois_mod) & negative binomial with the varying dispersion parameter (BIC_nb_mod) are nearly identical. This exhibits how closely the Poisson and negative binomial distributions can approximate each other. Only the negative binomial model with a dispersion parameter of 1 has some small differences. Interestingly, while the variables VolatileAcidity & AcidIndex had negative effects on the response variable for the linear models, they had positive ones in the generalized linear models.

```
## Joining, by = "var"
## Joining, by = "var"
## Joining, by = "var"
## Joining, by = "var"
## Joining, by = "var"
## Joining, by = "var"
```

var	bkwd_elim_lmod	BIC_lmod	BIC_pois_mod	quasi_pois_mod	BIC_nb_k1_mod	BIC_r
(Intercept)	3.978369	3.083253	3.402174	3.402174	4.339458	3
STARS	0.980418	0.981204	1.367792	1.367792	1.442951	1
LabelAppeal	0.432602	0.432512	1.142345	1.142345	1.126068	1
Alcohol	0.011948	0.011972				
TotalSulfurDioxide	0.000285	0.000284	1.000103	1.000103	1.000153	1
Sulphates	-0.045414				0.970868	
VolatileAcidity	-0.116450	-0.116642	0.959095	0.959095	0.946311	0
AcidIndex	-0.209313	-0.210612	0.915645	0.915645	0.888698	0
Density	-0.870797					
pH					0.973903	

Table 6: The number of negative fitted values in the linear models

bkwd_elim_lmod	BIC_lmod
20	21

Table 7: Final Model Comparison

model_name	n_vars	CV_RMSE
bkwd_elim_lmod	8	1.324
BIC_lmod	6	1.324
BIC_pois_mod	5	1.406
quasi_pois_mod	5	1.406
BIC_nb_k1_mod	7	1.486
BIC_nb_mod	5	1.406

5.2 Best Model

Based on the 2 tables("The number of negative fitted values in the linear models" and "Final Model Comparison"), we can see that the 2 linear models actually have the smallest cross-validated root mean square error, these models also have a small number of negative fitted values, which demonstrates the limited application of linear models to count response variables. Among the remaining generalized linear models, while the negative binomial model with a dispersion parameter of 1 (BIC_nb_k1_mod) has the largest cross-validated root mean square error, the difference between 1.406 and 1.486 may not be practically significant. Additionally, it is the only model that had a good fit under the chi-squared distribution. Consequently, it is the best model.

6 Evaluation Data Set Predictions

We used the negative binomial model with a dispersion parameter of 1 and made predictions on the evaluation data set. The following is statistical summary of the predicted responses.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4853  1.7580  2.6440  3.0850  3.9250 11.4700
```