Homework Assignment - 05

Critical Thinking Group 5
Arindam Barman
Mohamed Elmoudni
Shazia Khan
Kishore Prasad

Contents

1. Overview

In this homework assignment, we will explore, analyze and model a data set containing information on approximately 12795 commercially available wines using 16 variables. 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.

2. Objective

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. Using the training data set, we will 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 (or the same variables with different transformations).

To attain our objective, we will be following the below best practice steps and guidelines:

1 -Data Exploration
2 -Data Preparation
3 -Build Models
4 -Select Models

3 Data Exploration Analysis

In section we will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:
-Variable identification
-Variable Relationships
-Data summary analysis
-Outliers and Missing Values Identification

3.1 Variable identification

First we look the variables' datatypes and their roles.

Variable	Datatype	Role
INDEX	int	none
TARGET	int	response
FixedAcidity	num	predictor
VolatileAcidity	num	predictor
CitricAcid	num	predictor
ResidualSugar	num	predictor

Variable	Datatype	Role
Chlorides	num	predictor
FreeSulfurDioxide	num	predictor
TotalSulfurDioxide	num	predictor
Density	num	predictor
pH	num	predictor
Sulphates	num	predictor
Alcohol	num	predictor
LabelAppeal	int	predictor
AcidIndex	int	predictor
STARS	int	predictor

From the Table 1 above, we see that that all variables are quantitative mainly of numeric and integer datatype. Also, we will ignore the INDEX variable as it is just a unique identifier for each row. However, we will use the TARTGET variable as response variable and the remaining variables as predictors.

3.2 Variable Relationships

Next let's display and examine the variable relationships as shown in table 2.

Table 2: Variable Description

VARIABLE	DEFINITION	THEORETICAL.EFFECT
INDEX	Identification Variable (do not use) None	None
TARGET	Number of Cases Purchased None	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 customes 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 conten of wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	

At first glance, we can easily deduce that that the FreeSulfurDioxide (Sulfur Dioxide content of wine) can be derived from the TotalSulfurDioxide (Total Sulfur Dioxide of Wine). However, looking closer at the role of the sulfur dioxide SO_2, as it is used as a preservative because of its anti-oxidative and anti-microbial properties

in wine and also as a cleaning agent for barrels and winery facilities, we realize that when a winemaker says his/her wine has 100 ppm (part per million) of SO_2, he/she is most probably referring to the total amount of SO_2 in his wine, and that means:
total SO2 = free SO_2 + bound SO_2.
free SO_2: molecular SO_2 + bisulfites + sulfites
bound SO_2: sulfites attached to either sugars, acetaldehyde or phenolic compounds
In this case the free SO_2 portion (not associated with wine molecules) is effectively the buffer against microbes and oxidation... Hence without knowing the bound SO_2, we won't be able to derive FreeSulfurDioxide from TotalSulfurDioxide.

Also, looking breifly at the VolatileAcidity (Volatile Acid content of wine) and FixedAcidity (Fixed Acidity of Wine), we can easily deduce AcidIndex as the Acid index = Total acid (g/L) - pH. where Total acidity = Volatile Acid + Fixed Acidity. However, in our case the index is weighted average and we don't know the weighted average of either Volatile Acid or Fixed Acidity. Hence we will assume these variable do not have strict arithmetic relationships.

3.3 Data summary analysis

In this section, we will create summary data to better understand the initial relationship variables have with our dependent variable using correlation, central tendency, and dispersion As shown in table 3.

```
## '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             : int  2 3 3 1 2 NA NA 3 NA 4 ...
```

Table 3: Data Summary

	mean	sd	median	trimmed
TARGET	3.0290739	1.9263682	3.00000	3.0538244
FixedAcidity	7.0757171	6.3176435	6.90000	7.0736739
VolatileAcidity	0.3241039	0.7840142	0.28000	0.3243890
CitricAcid	0.3084127	0.8620798	0.31000	0.3102520
ResidualSugar	5.4187331	33.7493790	3.90000	5.5800410
Chlorides	0.0548225	0.3184673	0.04600	0.0540159
FreeSulfurDioxide	30.8455713	148.7145577	30.00000	30.9334877
TotalSulfurDioxide	120.7142326	231.9132105	123.00000	120.8895367
Density	0.9942027	0.0265376	0.99449	0.9942130
pH	3.2076282	0.6796871	3.20000	3.2055706
Sulphates	0.5271118	0.9321293	0.50000	0.5271453

	mean	sd	median	trimmed
Alcohol	10.4892363	3.7278190	10.40000	10.5018255
LabelAppeal	-0.0090660	0.8910892	0.00000	-0.0099639
AcidIndex	7.7727237	1.3239264	8.00000	7.6431572
STARS	2.0417550	0.9025400	2.00000	1.9711258

Below is the missing values and correlation table of the predictor variables to the response variables.

Table 4: Missing Data and Data Correlation

	Missing	Correlation
TARGET	0	1.0000000
FixedAcidity	0	-0.0490109
VolatileAcidity	0	-0.0887932
CitricAcid	0	0.0086846
ResidualSugar	616	0.0164913
Chlorides	638	-0.0382631
FreeSulfurDioxide	647	0.0438241
TotalSulfurDioxide	682	0.0514784
Density	0	-0.0355175
pH	395	-0.0094448
Sulphates	1210	-0.0388496
Alcohol	653	0.0620616
LabelAppeal	0	0.3565005
AcidIndex	0	-0.2460494
STARS	3359	0.5587938

Missing Values and Correlation Interpretation

From tables 3 and 4 above, we observe the followings:

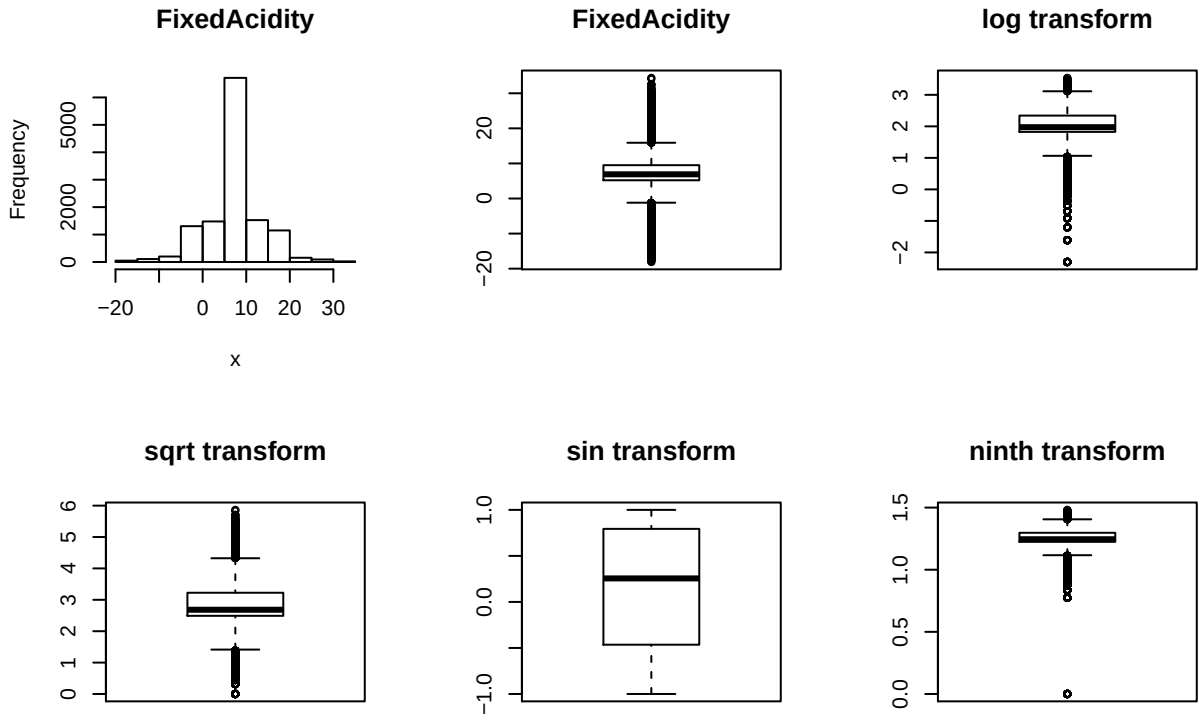
- Variable ResidualSugar has 616 and 0.0164913 correlation. Given the low correlation we will try try some imputation techniques to handle the missing the values and replace missing values with their respective value.
- variable Chlorides 638 -0.0382631 correlation. . Given the low negative correlation we will try we would replace missing values with their respective value
- Variable FreeSulfurDioxide 647 0.0438241. Given the low correlation we will impute the missing values with their respective value
- Variable TotalSulfurDioxide has 682 missing values with 0.0514784 correlation. Given the low correlation we will impute the missing values with their respective value.
- Variable Alcohol has 682 missing values with 0.0620616 correlation. Given the low correlation we will impute the missing values with their respective value.

Please note that ResidualSugar, Chlorides, FreeSulfurDioxide, Alcohol, and TotalSulfurDioxide variables have similar number of missing values. They are chemically related. However, we don't think they are arithmetically related.

- In addition, variable pH has 395 missing values with negative correlation of -0.0094448. Again we may just ignore these missing values especially that it has very low negative correlation to the target variable.
- Variable Sulphates has much higher missing values of 1210 with low negative correlation of -0.0388496. We will be imputing this values with their respective value
- Now, variable STARS has the highest missing values of 3359 and highest correlation of 0.5587938. This is very important variable and it drives sales and consequently heavily impacts our response variable. We have to be careful in fixing the missing values as this variable STARS is rating score variable with 1 being the lowest and 4 the highest

3.4 Outliers Identification

In this section we look at boxplots to determine the outliers in variables and decide on whether to act on the outliers. Lets do some univariate analysis. We will look at the Histogram and Boxplot for each variable to detect outliers if any and treat it accordingly.



***Please note that we generated the above plots for all other variables. However we hid the results for ease of streamlining our report.

4. Data Preparation

Now that we have completed the preliminary analysis, we will be cleaning and consolidating data into one dataset for use in analysis and modeling. We will be puring the belwo steps as guidlines:
- Missing Flags
- Missing values treatment
- Outliers treatment
- Dummy Variables

4.1 Missing Flags

We create flag variables to indicate whether some of the fields are missing any values. If the value is missing, we code it with 1 and if the value is present we code it with 0. The following are the variables that are created:

- ResidualSugar_MISS

- Chlorides_MISS
- FreeSulfurDioxide_MISS
- TotalSulfurDioxide_MISS
- pH_MISS
- Sulphates_MISS
- Alcohol_MISS
- STARS_MISS

4.2 Missing values treatment

Next we impute missing values. We can go ahead and use the mean as impute values. We will replace the missing values in the original variables. However, for STARS, we will code the missing value as a '0' instead of a mean. The following are the variables that are impacted:

- ResidualSugar
- Chlorides
- FreeSulfurDioxide
- TotalSulfurDioxide
- pH
- Sulphates
- Alcohol
- STARS

4.3 Outliers treatment

For outliers, we will use the capping method. In this method, we will replace all outliers that lie outside the 1.5 times of IQR limits. We will cap it by replacing those observations less than the lower limit with the value of 5th %ile and those that lie above the upper limit with the value of 95th %ile.

Accordingly we create the following new variables while retaining the original variables.

- FixedAcidity_CAP
- VolatileAcidity_CAP
- CitricAcid_CAP
- ResidualSugar_CAP
- Chlorides_CAP
- FreeSulfurDioxide_CAP
- TotalSulfurDioxide_CAP
- Density_CAP
- pH_CAP
- Sulphates_CAP
- Alcohol_CAP
- AcidIndex_CAP

4.4 Dummy Variables

Finally, we will also create dummy variables for the following variables:

- LabelAppeal : For this variable, we create a dummy variable to indicate if the value is Zero / Positive or Negative.
- STARS - We create a Dummy Variable for each of the star ratings - 1,2,3,4. The value is 1 in the respective variable based on the STARS value. A Zero value in all of the STARS dummy vars indicate that the value was missing in the original variable.

4.5 Correlation for new variables

Lets see how the new variables stack up against the TARGET.

Table 5: Correlation between TARGET and predictor variables

	Correlation
STARS_3	0.3597277
STARS_4	0.2783731
STARS_2	0.2484240
Alcohol_CAP	0.0634633
TotalSulfurDioxide_CAP	0.0503492
FreeSulfurDioxide_CAP	0.0417585
LabelAppeal_Positive	0.0206261
ResidualSugar_CAP	0.0204409
CitricAcid_CAP	0.0120351
ResidualSugar_MISS	0.0111995
TotalSulfurDioxide_MISS	0.0061720
Chlorides_MISS	0.0026937
Alcohol_MISS	0.0014776
FreeSulfurDioxide_MISS	-0.0001501
pH_MISS	-0.0099654
pH_CAP	-0.0102565
Sulphates_MISS	-0.0125039
Chlorides_CAP	-0.0304686
Density_CAP	-0.0315375
Sulphates_CAP	-0.0359312
FixedAcidity_CAP	-0.0510757
VolatileAcidity_CAP	-0.0891214
STARS_1	-0.1300422
AcidIndex_CAP	-0.2353997
STARS_MISS	-0.5715792

From the above Correlations, we can make the following observations:

- The following variables have a positive correlation with TARGET: STARS_3, STARS_4, STARS_2, Alcohol_CAP, TotalSulfurDioxide_CAP, FreeSulfurDioxide_CAP, LabelAppeal_Positive, ResidualSugar_CAP, CitricAcid_CAP, ResidualSugar_MISS, TotalSulfurDioxide_MISS, Chlorides_MISS, Alcohol_MISS.
- The following variables have a negative correlation with TARGET: FreeSulfurDioxide_MISS, pH_MISS,

pH_CAP, Sulphates_MISS, Chlorides_CAP, Density_CAP, Sulphates_CAP, FixedAcidity_CAP, VolatileAcidity_CAP, STARS_1, AcidIndex_CAP, STARS_MISS.

- Not all variable have a strong correlation in either direction. However, the following stand out for having a stronger correlation: STARS_MISS, STARS_3, STARS_4, STARS_2, AcidIndex_CAP, STARS_1, VolatileAcidity_CAP, Alcohol_CAP, FixedAcidity_CAP, TotalSulfurDioxide_CAP.

5. Build Models

Since we are dealing with count variables, our modeling technique will mainly focus on using variation of the Generalized Linear Model (GLM) family functions. We will start with the classical Poisson regression; then we will enhance it using model Negative binominal model.
In addition, we will also create models using linear regression.

Using original and transformed datasets, we will build at least twelve models as follow:
- Two Poisson models
- Two Quasi-Poisson models
- Two Zero-inflated Poisson models
- Two Negative binomial models
- Two Zero-inflated Negative Binomial models
- Two Linear regression models

Below is a summary table showing models' variables.

Table 6: Model Variables

Variable	Original	Transformed	Comments
TARGET	Y	Y	The TARGET variable
FixedAcidity	Y		Imputed with Mean
VolatileAcidity	Y		Imputed with Mean
CitricAcid	Y		Imputed with Mean
ResidualSugar	Y		Imputed with Mean
Chlorides	Y		Imputed with Mean
FreeSulfurDioxide	Y		Imputed with Mean
TotalSulfurDioxide	Y		Imputed with Mean
Density	Y		Imputed with Mean
pH	Y		Imputed with Mean
Sulphates	Y		Imputed with Mean
Alcohol	Y		Imputed with Mean
LabelAppeal	Y		Original Variable
AcidIndex	Y		Imputed with Mean
STARS	Y		Original Variable
ResidualSugar_MISS		Y	Missing Flag
Chlorides_MISS		Y	Missing Flag
FreeSulfurDioxide_MISS		Y	Missing Flag
TotalSulfurDioxide_MISS		Y	Missing Flag
pH_MISS		Y	Missing Flag
Sulphates_MISS		Y	Missing Flag
Alcohol_MISS		Y	Missing Flag
STARS_MISS		Y	Missing Flag
FixedAcidity_CAP		Y	Imputed with Mean and Outliers capped
VolatileAcidity_CAP		Y	Imputed with Mean and Outliers capped
CitricAcid_CAP		Y	Imputed with Mean and Outliers capped

Variable	Original	Transformed	Comments
ResidualSugar_CAP		Y	Imputed with Mean and Outliers capped
Chlorides_CAP		Y	Imputed with Mean and Outliers capped
FreeSulfurDioxide_CAP		Y	Imputed with Mean and Outliers capped
TotalSulfurDioxide_CAP		Y	Imputed with Mean and Outliers capped
Density_CAP		Y	Imputed with Mean and Outliers capped
pH_CAP		Y	Imputed with Mean and Outliers capped
Sulphates_CAP		Y	Imputed with Mean and Outliers capped
Alcohol_CAP		Y	Imputed with Mean and Outliers capped
AcidIndex_CAP		Y	Imputed with Mean and Outliers capped
LabelAppeal_Positive		Y	Positive or Negative Dummy Variable
STARS_1		Y	Dummy Variable
STARS_2		Y	Dummy Variable
STARS_3		Y	Dummy Variable
STARS_4		Y	Dummy Variable

5.1 Poisson models

Our first attempt to capture the relationship between the wine chemical properties and number of cases of the wine being sold in a parametric regression model, we fit the basic Poisson regression model

5.1.1 Poisson Model 1

We will explore the Poisson regression model Using original data with replacing all missing data with the means.

Table 7: Model 1 Poisson Original Data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5259824	0.1954718	7.8066616	0.0000000
FixedAcidity	-0.0003045	0.0008205	-0.3711814	0.7105024
VolatileAcidity	-0.0334329	0.0065161	-5.1308519	0.0000003
CitricAcid	0.0077726	0.0058922	1.3191354	0.1871238
ResidualSugar	0.0000568	0.0001546	0.3670421	0.7135876
Chlorides	-0.0414139	0.0164498	-2.5175957	0.0118159
FreeSulfurDioxide	0.0001254	0.0000351	3.5705960	0.0003562
TotalSulfurDioxide	0.0000830	0.0000227	3.6466783	0.0002657
Density	-0.2823481	0.1919703	-1.4707905	0.1413478
pH	-0.0157219	0.0076380	-2.0583793	0.0395537
Sulphates	-0.0126738	0.0057487	-2.2046321	0.0274799
Alcohol	0.0022014	0.0014100	1.5613311	0.1184457
LabelAppeal	0.1331963	0.0060633	21.9676836	0.0000000
AcidIndex	-0.0870512	0.0045483	-19.1391650	0.0000000
STARS	0.3112869	0.0045311	68.6999887	0.0000000

5.1.1.2 Interpretation Poisson Model 1

From this output, we have the following estimated model:

$$\hat{y} = e^{B_0x_0+B_1x_1+B_2x_2+B_3x_3+B_4x_4+B_5x_5+B_6x_6+B_7x_7+B_8x_8+B_9x_9+B_{10}x_{10}+B_{11}x_{11}+B_{12}x_{12}+B_{13}x_{13}+B_{14}x_{14}}$$

where
$B_0 = 1.526$
$B_1 = -3.045e-04$
$B_2 = -3.343e-02$
$B_3 = 7.773e-03$
$B_4 = 5.676e-05$
$B_5 = -4.141e-02$
$B_6 = 1.254e-04$
$B_7 = 8.296e-05$
$B_8 = -2.823e-01$
$B_9 = -1.572e-02$
$B_10 = -1.267e-02$
$B_11 = 2.201e-03$
$B_12 = 1.332e-01$
$B_13 = -8.705e-02$
$B_14 = 3.113e-0$

and

$x_0 = 1$
$x_1 = FixedAcidity$
$x_2 = VolatileAcidity$
$x_3 = CitricAcid$
$x_4 = ResidualSugar$
$x_5 = Chlorides$
$x_6 = FreeSulfurDioxide$
$x_7 = TotalSulfurDioxide$
$x_8 = Density$
$x_9 = pH$
$x_10 = Sulphates$
$x_11 = Alcohol$
$x_12 = LabelAppeal$
$x_13 = AcidIndex$
$x_14 = STARS$

5.1.1.3 Coefficient Analysis

In addition, the coefficient for VolatileAcidity, FreeSulfurDioxide, TotalSulfurDioxide, LabelAppeal, AcidIndex, and STARS are highly significant.

Unlike the linear model, in order to interpret the slope coefficient in a Poisson regression, it makes better sense to look at the ratio of predicted responses (instead of the difference) for a unit increase in x. for instance:

$$\frac{e^{b_0+B_1(x+1)}}{e^{b_0+B_1x}} = e^{B_1}$$

For instance, for with $B_1 = -(.0003045)$, we have $e^{B_1} = e^{-(.0003045)} = 0.999695$

Thus, for a unit increase in the FixedAcidity, we would expect to see the number of cases of wine that will be sold given certain properties of the wine to decrease by a factor of = 0.999695.

Hence, for a unit increase in our highly significant variables:
- VolatileAcidity, we expect a decrease of $e^{-(0.0343)} = 0.9662816$ the number of cases of wine that will be sold
- FreeSulfurDioxide, we expect an increase of $e^{0.0000829} = 1.000083$ the number of cases of wine that will be sold
- TotalSulfurDioxide, we expect a decrease of $e^{-(0.2823)} = 0.7540474$ the number of cases of wine that will be sold
- LabelAppeal, we expect a increase of $e^{(.1332)} = 1.142478$ the number of cases of wine that will be sold
- AcidIndex,we expect a decrease of $e^{-(08705)} = 0.9166313$ the number of cases of wine that will be sold
- STARS,we expect a increase of $e^{(3.113)} = 22.48841$ the number of cases of wine that will be sold

5.1.1.4 Overdisperson Analysis:

Another common problem with Poisson regression is that the response is more variable than what is expected by the model; this is called overdisperson. Thus checking for overdispersion, we will examine if the residual deviance greatly exceeds the residual degrees of freedom, then that is an indication of an overdispersion problem.
For our model(1), we see that our Residual deviance is 14728 and degrees of freedom is 12780; our Residual deviance 1.15 greater than our Residual degrees of freedom. Hence, the response is little more variable than

what is expected by model (1). However, we won't address this issue as the Residual deviance does not greatly exceed residual degrees of freedom.

Sine we see that we have over dispersion, let's find out the dispersion parameter ϕ. Since the variance in the Poisson model is identical to the mean, the expectations are to have $\phi = 1$.

```
## [1] 0.851513
```

Our dispersion parameter is 0.851513; obviously it is not 1.

5.1.2 Quasi-Poisson model (Model 2)

We will explore the Quasi-Poisson regression model Using original data with replacing all missing data with the means.

Another way of dealing with over-dispersion is to use Quasi-Poisson model which uses the mean regression function and the variance function from the Poisson GLM but to leave the dispersion parameter ϕ unrestricted. Thus, ϕ is not assumed to be fixed at 1 but is estimated from the data. This strategy leads to the same coefficient estimates as the standard Poisson model but inference is adjusted for over-dispersion.

Table 8: Model 2 Quasi-Poisson Original Data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5259824	0.1803772	8.4599527	0.0000000
FixedAcidity	-0.0003045	0.0007571	-0.4022433	0.6875117
VolatileAcidity	-0.0334329	0.0060129	-5.5602211	0.0000000
CitricAcid	0.0077726	0.0054372	1.4295257	0.1528776
ResidualSugar	0.0000568	0.0001427	0.3977576	0.6908155
Chlorides	-0.0414139	0.0151795	-2.7282776	0.0063753
FreeSulfurDioxide	0.0001254	0.0000324	3.8693971	0.0001096
TotalSulfurDioxide	0.0000830	0.0000210	3.9518462	0.0000780
Density	-0.2823481	0.1771460	-1.5938719	0.1109895
pH	-0.0157219	0.0070482	-2.2306323	0.0257228
Sulphates	-0.0126738	0.0053048	-2.3891241	0.0169030
Alcohol	0.0022014	0.0013011	1.6919892	0.0906724
LabelAppeal	0.1331963	0.0055951	23.8060229	0.0000000
AcidIndex	-0.0870512	0.0041971	-20.7408030	0.0000000
STARS	0.3112869	0.0041812	74.4490647	0.0000000

5.1.2.1 Interpretation Quasi-Poisson model

From this output, we have the following estimated model:

$$\hat{y} = e^{B_0x_0+B_1x_1+B_2x_2+B_3x_3+B_4x_4+B_5x_5+B_6x_6+B_7x_7+B_8x_8+B_9x_9+B_{10}x_{10}+B_{11}x_{11}+B_{12}x_{12}+B_{13}x_{13}+B_{14}x_{14}}$$

where

$B_0 = 1.526$
$B_1 = -0.0003$
$B_2 = -0.03343$
$B_3 = 0.00777$
$B_4 = 0.00006$

$B_5 = -0.04141$
$B_6 = 0.00013$
$B_7 = 0.00008$
$B_8 = -0.2823$
$B_9 = -0.01572$
$B_10 = -0.01267$
$B_11 = 0.0022$
$B_12 = 0.1332$
$B_13 = -0.08705$
$B_14 = 0.3113$

and

$x_0 = 1$
$x_1 = FixedAcidity$
$x_2 = VolatileAcidity$
$x_3 = CitricAcid$
$x_4 = ResidualSugar$
$x_5 = Chlorides$
$x_6 = FreeSulfurDioxide$
$x_7 = TotalSulfurDioxide$
$x_8 = Density$
$x_9 = pH$
$x_10 = Sulphates$
$x_11 = Alcohol$
$x_12 = LabelAppeal$
$x_13 = AcidIndex$
$x_14 = STARS$

5.1.2.2 Coefficient Analysis

The coefficient for VolatileAcidity, FreeSulfurDioxide, TotalSulfurDioxide, LabelAppeal, AcidIndex, STARS are highly significant. For a unit increase in our highly significant variables:

- VolatileAcidity, we expect a decrease of $e^{(-0.03343)} = 0.967123$ in the number of cases of wine that will be sold
- FreeSulfurDioxide, we expect an increase of $e^{(0.0001254)} = 1.000125$ in the number of cases of wine that will be sold
- TotalSulfurDioxide, we expect an increase of $e^{(0.00008296)} = 1.000083$ in the number of cases of wine that will be sold
- LabelAppeal, we expect an increase of $e^{(0.1332)} = 1.142478$ in the number of cases of wine that will be sold
- AcidIndex, we expect a decrease of $e^{(-0.08705)} = 0.916631$ in the number of cases of wine that will be sold
- STARS, we expect an increase of $e^{(0.3113)} = 1.365199$ in the number of cases of wine that will be sold

Please note that the Quasi-Poisson model leads to the same coefficient estimates as the standard Poisson model but inference is adjusted for over-dispersion. Hence please refer to Poison model Coefficient Analysis for details.

Please note that dispersion parameter in the Quasi-Poisson model is 0.851513; which is similar to that of the classical Poisson Model (1)

5.1.3 zero-inflation model (Model 3)

We will explore the zero-inflationregression model Using original data with replacing all missing data with the means.

Next we will proceed with zero-inflation model as another very common occurrence when working with count data is that there will be an overabundance of zero counts which is not consistent with the Poisson model.

5.1.3.1 Coefficient Analysis (Model 3)

Table 9: Model 3, Zero Inflation Poisson Original Data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.4434026	0.2019769	7.1463741	0.0000000
FixedAcidity	0.0003383	0.0008420	0.4017674	0.6878552
VolatileAcidity	-0.0121070	0.0067206	-1.8014925	0.0716253
CitricAcid	0.0004926	0.0060239	0.0817791	0.9348224
ResidualSugar	-0.0000770	0.0001586	-0.4854804	0.6273356
Chlorides	-0.0224087	0.0169086	-1.3252855	0.1850765
FreeSulfurDioxide	0.0000255	0.0000355	0.7178064	0.4728767
TotalSulfurDioxide	-0.0000178	0.0000226	-0.7874563	0.4310148
Density	-0.2845413	0.1982976	-1.4349207	0.1513097
pH	0.0059315	0.0078586	0.7547699	0.4503870
Sulphates	0.0001726	0.0059190	0.0291624	0.9767350
Alcohol	0.0068863	0.0014396	4.7834823	0.0000017
LabelAppeal	0.2329532	0.0063025	36.9618412	0.0000000
AcidIndex	-0.0185821	0.0048975	-3.7941732	0.0001481
STARS	0.1009199	0.0052013	19.4029966	0.0000000
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.4483881	1.3374162	-3.3261061	0.0008807
FixedAcidity	0.0007591	0.0055469	0.1368541	0.8911461
VolatileAcidity	0.1937198	0.0438512	4.4176612	0.0000100
CitricAcid	-0.0296037	0.0399713	-0.7406225	0.4589224
ResidualSugar	-0.0011765	0.0010429	-1.1280299	0.2593073
Chlorides	0.0921158	0.1093491	0.8424010	0.3995636
FreeSulfurDioxide	-0.0007419	0.0002422	-3.0632312	0.0021896
TotalSulfurDioxide	-0.0009866	0.0001523	-6.4761642	0.0000000
Density	0.4900517	1.3159510	0.3723936	0.7095998
pH	0.2160935	0.0512207	4.2188696	0.0000246
Sulphates	0.1323441	0.0387670	3.4138357	0.0006406
Alcohol	0.0279120	0.0095782	2.9141240	0.0035669
LabelAppeal	0.7229711	0.0429468	16.8340974	0.0000000
AcidIndex	0.4347418	0.0258387	16.8251979	0.0000000
STARS	-2.3768721	0.0603161	-39.4069034	0.0000000

"From this output, we have the following estimated model:

$$\hat{y} = e^{B_0x_0+B_1x_1+B_2x_2+B_3x_3+B_4x_4+B_5x_5+B_6x_6+B_7x_7+B_8x_8+B_9x_9+B_{10}x_{10}+B_{11}x_{11}+B_{12}x_{12}+B_{13}x_{13}+B_{14}x_{14}}$$

where

$B_0 = 1.443$
$B_1 = 0.00034$
$B_2 = -0.01211$
$B_3 = 0.00049$
$B_4 = -0.00008$
$B_5 = -0.02241$
$B_6 = 0.00003$
$B_7 = -0.00002$
$B_8 = -0.2845$
$B_9 = 0.00593$
$B_10 = 0.00017$
$B_11 = 0.00689$

$B_12 = 0.233$
$B_13 = -0.01858$
$B_14 = 0.1009$

and

$x_0 = 1\$$
$x_1 = FixedAcidity$
$x_2 = VolatileAcidity$
$x_3 = CitricAcid$
$x_4 = ResidualSugar$
$x_5 = Chlorides$
$x_6 = FreeSulfurDioxide$
$x_7 = TotalSulfurDioxide$
$x_8 = Density$
$x_9 = pH$
$x_10 = Sulphates$
$x_11 = Alcohol$
$x_12 = LabelAppeal$
$x_13 = AcidIndex$
$x_14 = STARS$

5.1.3.2 Coefficient Analysis

The coefficient for Alcohol, LabelAppeal, AcidIndex, STARS are highly significant. For a unit increase in our highly significant variables:
- Alcohol, we expect an increase of $e^{(0.006886)} = 1.00691$ in the number of cases of wine that will be sold
- LabelAppeal, we expect an increase of $e^{(0.233)} = 1.262381$ in the number of cases of wine that will be sold
- AcidIndex, we expect a decrease of $e^{(-0.01858)} = 0.981592$ in the number of cases of wine that will be sold
- STARS, we expect an increase of $e^{(0.1009)} = 1.106166$ in the number of cases of wine that will be sold

We noticed that some variables have their coefficient sign changed from negative to positive and vice versa. For instance;

FixedAcidity changed from -3.045e-04 in model 1 to 3.383e-04 in the zip model ResidualSugar changed from 5.676e-05 in model 1 to -7.702e-05 in the zip model TotalSulfurDioxide changed from 8.296e-05 in model 1 to -1.783e-05 in the zip model. pH changed from -1.572e-02 in model 1 to pH 5.931e-03 in the zip model. Sulphates changed from -1.267e-02 in model 1 to 1.726e-04 in the zip model.

5.1.3.2 Overdisperson Analysis

Please note that dispersion parameter in the zero-inflation modelis 0.4636815; which is lower than of the classical Poisson Model (1)

```
## [1] 0.4636815
```

Note that the zip model output above does not indicate in any way if our zero-inflated model is an improvement over a standard Poisson regression. We can determine this by running the corresponding standard negative Poisson model and then performing a Vuong test of the two models.

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                   47.98330 model1 > model2 < 2.22e-16
## AIC-corrected         47.73759 model1 > model2 < 2.22e-16
## BIC-corrected         46.82150 model1 > model2 < 2.22e-16
```

The Vuong test suggests that the zero-inflated Poisson model is slight improvement over a standard Poisson model.

5.2 Poisson Model (Model 4)

In this model we will be using the basic Poisson regression model; however using transformed data.

Table 10: Model 4, Poisson Transformed Data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.5701252	0.2001466	12.8412129	0.0000000
ResidualSugar_MISS	0.0228341	0.0234038	0.9756567	0.3292346
Chlorides_MISS	0.0030173	0.0232957	0.1295225	0.8969442
FreeSulfurDioxide_MISS	0.0230001	0.0236607	0.9720801	0.3310107
TotalSulfurDioxide_MISS	0.0188307	0.0224578	0.8384906	0.4017552
pH_MISS	-0.0349529	0.0299113	-1.1685516	0.2425843
Sulphates_MISS	-0.0067580	0.0175716	-0.3845970	0.7005360
Alcohol_MISS	0.0213581	0.0230597	0.9262075	0.3543381
STARS_MISS	-1.4710696	0.0237121	-62.0387249	0.0000000
FixedAcidity_CAP	-0.0005712	0.0009179	-0.6223390	0.5337190
VolatileAcidity_CAP	-0.0355011	0.0072476	-4.8983557	0.0000010
CitricAcid_CAP	0.0074304	0.0065266	1.1384863	0.2549175
ResidualSugar_CAP	0.0001348	0.0001538	0.8762370	0.3809012
Chlorides_CAP	-0.0266371	0.0161831	-1.6459779	0.0997683
FreeSulfurDioxide_CAP	0.0001600	0.0000527	3.0392789	0.0023715
TotalSulfurDioxide_CAP	0.0000838	0.0000260	3.2244078	0.0012623
Density_CAP	-0.2847644	0.1945730	-1.4635349	0.1433211
pH_CAP	-0.0136064	0.0086724	-1.5689265	0.1166651
Sulphates_CAP	-0.0119359	0.0059076	-2.0204432	0.0433374
Alcohol_CAP	0.0039558	0.0016456	2.4038658	0.0162227
AcidIndex_CAP	-0.0780062	0.0052584	-14.8345268	0.0000000
LabelAppeal_Positive	-0.0255998	0.0185449	-1.3804212	0.1674570
STARS_1	-0.7179018	0.0208066	-34.5035486	0.0000000
STARS_2	-0.3426734	0.0194390	-17.6281016	0.0000000
STARS_3	-0.1733976	0.0200561	-8.6456244	0.0000000

5.2.1 Interpretation Poisson Model 4

From this output, we have the following estimated model:

$$\hat{y} = e^{B_0x_0+B_1x_1+B_2x_2+B_3x_3+B_4x_4+B_5x_5+B_6x_6+B_7x_7+B_8x_8+B_9x_9+B_{10}x_{10}+B_{11}x_{11}+B_{12}x_{12}+B_{13}x_{13}+B_{14}x_{14}+B_{15}x_{15}+B_{16}x_{16}+B_{17}x_{17}+B_{18}x_{18}}$$

where

$B_0 = 2.57$
$B_1 = 0.02283$
$B_2 = 0.00302$
$B_3 = 0.023$
$B_4 = 0.01883$
$B_5 = -0.03495$
$B_6 = -0.00676$
$B_7 = 0.02136$
$B_8 = -1.471$
$B_9 = -0.00057$
$B_10 = -0.0355$
$B_11 = 0.00743$
$B_12 = 0.00013$
$B_13 = -0.02664$
$B_14 = 0.00016$
$B_15 = 0.00008$
$B_16 = -0.2848$
$B_17 = -0.01361$
$B_18 = -0.01194$
$B_19 = 0.00396$
$B_20 = -0.07801$
$B_21 = -0.0256$
$B_22 = -0.7179$
$B_23 = -0.3427$
$B_24 = -0.1734$

and

$x_0 = 1$
$x_1 = ResidualSugar_MISS$
$x_2 = Chlorides_MISS$
$x_3 = FreeSulfurDioxide_MISS$
$x_4 = TotalSulfurDioxide_MISS$
$x_5 = pH_MISS$
$x_6 = Sulphates_MISS$
$x_7 = Alcohol_MISS$
$x_8 = STARS_MISS$
$x_9 = FixedAcidity_CAP$
$x_10 = VolatileAcidity_CAP$
$x_11 = CitricAcid_CAP$
$x_12 = ResidualSugar_CAP$
$x_13 = Chlorides_CAP$
$x_14 = FreeSulfurDioxide_CAP$
$x_15 = TotalSulfurDioxide_CAP$
$x_16 = Density_CAP$
$x_17 = pH_CAP$
$x_18 = Sulphates_CAP$
$x_19 = Alcohol_CAP$
$x_20 = AcidIndex_CAP$
$x_21 = LabelAppeal_Positive$
$x_22 = STARS_1$
$x_23 = STARS_2$
$x_24 = STARS_3$

5.2.1.1 Coefficient Analysis

The coefficient for STARS_MISS, VolatileAcidity_CAP, AcidIndex_CAP, STARS_1, STARS_2, STARS_3 are highly significant. For a unit increase in our highly significant variables:
- STARS_MISS, we expect a decrease of $e^{(-1.471)} = 0.229696$ in the number of cases of wine that will be sold
- VolatileAcidity_CAP, we expect a decrease of $e^{(-0.0355)} = 0.965123$ in the number of cases of wine that will be sold
- AcidIndex_CAP, we expect a decrease of $e^{(-0.07801)} = 0.924955$ in the number of cases of wine that will be sold
- STARS_1, we expect a decrease of $e^{(-0.7179)} = 0.487776$ in the number of cases of wine that will be sold
- STARS_2, we expect a decrease of $e^{(-0.3427)} = 0.709851$ in the number of cases of wine that will be sold
- STARS_3, we expect a decrease of $e^{(-0.1734)} = 0.840801$ in the number of cases of wine that will be sold

Most of the coefficients stayed still significant in the model. However, some variables experienced a decrease in p values especially the ones that have capped; which was expected as in the original they had untreated outliers. For instance FixedAcidity p-value went from 0.710502 to 0.53372. The same for ResidualSugar variable went from 0.713588 to 0.38090. Again this is due to outliers' treatment.

In addition, the Poisson model with transformed data has a slight improved as its AIC, 46368, is slightly lower than the model 1 AIC (46700.); which was run against the original data.

5.2.1.2 Overdisperson Analysis

For our model(2), we see that our Residual deviance is 14376 and degrees of freedom is 12770; our Residual deviance 1.12 greater than our Residual degrees of freedom. Hence, the response is little more variable than what is expected by model (2). Please note that this is a slight improvement from model 1 with original data which was 1.15.

Sine we see that we have over dispersion, let's find out the dispersion parameter ϕ. Since the variance in the Poisson model is identical to the mean, the expectations are to have $\phi = 1$.

```
## [1] 0.9667917
```

Our dispersion parameter for Modle (2) is 0.9667917 which is much closer to 1 than the dispersion parameter of our Modle (1).

5.2.2 Quasi-Poisson with transformed data (model 5)

In this model we will be using the Quasi-Poisson regression model; however using transformed data

Table 11: Model 5 Quasi-Poisson Transformed Data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5701252	0.1967953	13.0598920	0.0000000
ResidualSugar_MISS	0.0228341	0.0230120	0.9922716	0.3210839
Chlorides_MISS	0.0030173	0.0229056	0.1317282	0.8952014
FreeSulfurDioxide_MISS	0.0230001	0.0232645	0.9886341	0.3228609
TotalSulfurDioxide_MISS	0.0188307	0.0220818	0.8527697	0.3938030
pH_MISS	-0.0349529	0.0294105	-1.1884514	0.2346777
Sulphates_MISS	-0.0067580	0.0172774	-0.3911465	0.6956955
Alcohol_MISS	0.0213581	0.0226736	0.9419804	0.3462205

	Estimate	Std. Error	t value	Pr(>|t|)
STARS_MISS	-1.4710696	0.0233151	-63.0952117	0.0000000
FixedAcidity_CAP	-0.0005712	0.0009025	-0.6329371	0.5267860
VolatileAcidity_CAP	-0.0355011	0.0071262	-4.9817721	0.0000006
CitricAcid_CAP	0.0074304	0.0064173	1.1578742	0.2469370
ResidualSugar_CAP	0.0001348	0.0001512	0.8911588	0.3728607
Chlorides_CAP	-0.0266371	0.0159122	-1.6740081	0.0941535
FreeSulfurDioxide_CAP	0.0001600	0.0000518	3.0910362	0.0019989
TotalSulfurDioxide_CAP	0.0000838	0.0000256	3.2793177	0.0010434
Density_CAP	-0.2847644	0.1913150	-1.4884581	0.1366548
pH_CAP	-0.0136064	0.0085272	-1.5956445	0.1105929
Sulphates_CAP	-0.0119359	0.0058086	-2.0548503	0.0399138
Alcohol_CAP	0.0039558	0.0016180	2.4448023	0.0145066
AcidIndex_CAP	-0.0780062	0.0051704	-15.0871510	0.0000000
LabelAppeal_Positive	-0.0255998	0.0182344	-1.4039291	0.1603643
STARS_1	-0.7179018	0.0204582	-35.0911259	0.0000000
STARS_2	-0.3426734	0.0191135	-17.9282989	0.0000000
STARS_3	-0.1733976	0.0197203	-8.7928549	0.0000000

5.2.2.1 Interpretation Quasi-Poisson model 5

"From this output, we have the following estimated model:

$\hat{y} = e^{B_0x_0+B_1x_1+B_2x_2+B_3x_3+B_4x_4+B_5x_5+B_6x_6+B_7x_7+B_8x_8+B_9x_9+B_{10}x_{10}+B_{11}x_{11}+B_{12}x_{12}+B_{13}x_{13}+B_{14}x_{14}+B_{15}x_{15}+B_{16}x_{16}+B_{17}x_{17}+B_{18}x_{18}}$

where

$B_0 = 2.57$
$B_1 = 0.02283$
$B_2 = 0.00302$
$B_3 = 0.023$
$B_4 = 0.01883$
$B_5 = -0.03495$
$B_6 = -0.00676$
$B_7 = 0.02136$
$B_8 = -1.471$
$B_9 = -0.00057$
$B_10 = -0.0355$
$B_11 = 0.00743$
$B_12 = 0.00013$
$B_13 = -0.02664$
$B_14 = 0.00016$
$B_15 = 0.00008$
$B_16 = -0.2848$
$B_17 = -0.01361$
$B_18 = -0.01194$
$B_19 = 0.00396$
$B_20 = -0.07801$
$B_21 = -0.0256$
$B_22 = -0.7179$
$B_23 = -0.3427$
$B_24 = -0.1734$

and

$x_0 = 1$
$x_1 = ResidualSugar_MISS$
$x_2 = Chlorides_MISS$
$x_3 = FreeSulfurDioxide_MISS$
$x_4 = TotalSulfurDioxide_MISS$
$x_5 = pH_MISS$
$x_6 = Sulphates_MISS$
$x_7 = Alcohol_MISS$
$x_8 = STARS_MISS$
$x_9 = FixedAcidity_CAP$
$x_10 = VolatileAcidity_CAP$
$x_11 = CitricAcid_CAP$
$x_12 = ResidualSugar_CAP$
$x_13 = Chlorides_CAP$
$x_14 = FreeSulfurDioxide_CAP$
$x_15 = TotalSulfurDioxide_CAP$
$x_16 = Density_CAP$
$x_17 = pH_CAP$
$x_18 = Sulphates_CAP$
$x_19 = Alcohol_CAP$
$x_20 = AcidIndex_CAP$
$x_21 = LabelAppeal_Positive$
$x_22 = STARS_1$
$x_23 = STARS_2$
$x_24 = STARS_3$

5.2.2.2 Coefficient Analysis

The coefficient for STARS_MISS, VolatileAcidity_CAP, AcidIndex_CAP, STARS_1, STARS_2, STARS_3 are highly significant. For a unit increase in our highly significant variables:
- STARS_MISS, we expect a decrease of $e^{(-1.471)} = 0.229696$ in the number of cases of wine that will be sold
- VolatileAcidity_CAP, we expect a decrease of $e^{(-0.0355)} = 0.965123$ in the number of cases of wine that will be sold
- AcidIndex_CAP, we expect a decrease of $e^{(-0.07801)} = 0.924955$ in the number of cases of wine that will be sold
- STARS_1, we expect a decrease of $e^{(-0.7179)} = 0.487776$ in the number of cases of wine that will be sold
- STARS_2, we expect a decrease of $e^{(-0.3427)} = 0.709851$ in the number of cases of wine that will be sold
- STARS_3, we expect a decrease of $e^{(-0.1734)} = 0.840801$ in the number of cases of wine that will be sold

Please note that the Quasi-Poisson model leads to the same coefficient estimates as the standard Poisson model but inference is adjusted for over-dispersion. Hence please refer to Poison model Coefficient Analysis for details.

Also, please note that dispersion parameter in the Quasi-Poisson model is 0.9667917; which is similar to that of the classical Poisson Model (2)

5.2.3 zero-inflation with transformed data (Model 6)

In this model we will be using the zero-inflation regression model; however using transformed data

Next we will proceed with zero-inflation model as another very common occurrence when working with count data is that there will be an overabundance of zero counts which is not consistent with the Poisson model.

Table 12: Model 6, Zero Inflation Poisson Transformed Data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.4735603	0.2058858	12.0142362	0.0000000
ResidualSugar_MISS	0.0218628	0.0240009	0.9109161	0.3623396
Chlorides_MISS	0.0073886	0.0239380	0.3086549	0.7575841
FreeSulfurDioxide_MISS	0.0201412	0.0242146	0.8317805	0.4055328
TotalSulfurDioxide_MISS	0.0235291	0.0230599	1.0203471	0.3075639
pH_MISS	-0.0277018	0.0307602	-0.9005739	0.3678149
Sulphates_MISS	-0.0061845	0.0180537	-0.3425597	0.7319297
Alcohol_MISS	0.0169631	0.0236318	0.7178104	0.4728742
STARS_MISS	-1.3604443	0.0262705	-51.7860663	0.0000000
FixedAcidity_CAP	-0.0004496	0.0009424	-0.4771128	0.6332818
VolatileAcidity_CAP	-0.0303171	0.0074597	-4.0641403	0.0000482
CitricAcid_CAP	0.0055882	0.0066993	0.8341443	0.4041997
ResidualSugar_CAP	0.0000792	0.0001577	0.5023424	0.6154267
Chlorides_CAP	-0.0211808	0.0165904	-1.2766951	0.2017099
FreeSulfurDioxide_CAP	0.0001529	0.0000538	2.8413276	0.0044926
TotalSulfurDioxide_CAP	0.0000593	0.0000264	2.2442272	0.0248178
Density_CAP	-0.2950753	0.2000327	-1.4751356	0.1401761
pH_CAP	-0.0080552	0.0089139	-0.9036631	0.3661740
Sulphates_CAP	-0.0093968	0.0060705	-1.5479456	0.1216354
Alcohol_CAP	0.0047754	0.0016869	2.8309200	0.0046414
AcidIndex_CAP	-0.0670232	0.0055600	-12.0546263	0.0000000
LabelAppeal_Positive	-0.0272156	0.0190315	-1.4300328	0.1527076
STARS_1	-0.6211646	0.0219120	-28.3481490	0.0000000
STARS_2	-0.3266830	0.0194768	-16.7728901	0.0000000
STARS_3	-0.1730168	0.0200575	-8.6260317	0.0000000

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.806828	0.0846577	-33.15502	0

5.2.3.1 Interpretation for Zero Inflation Model (Model 6)

From this output, we have the following estimated model:

$\hat{y} = e^{B_0x_0+B_1x_1+B_2x_2+B_3x_3+B_4x_4+B_5x_5+B_6x_6+B_7x_7+B_8x_8+B_9x_9+B_{10}x_{10}+B_{11}x_{11}+B_{12}x_{12}+B_{13}x_{13}+B_{14}x_{14}+B_{15}x_{15}+B_{16}x_{16}+B_{17}x_{17}+B_{18}x_{18}}$

where

$B_0 = 2.474$
$B_1 = 0.02186$
$B_2 = 0.00739$
$B_3 = 0.02014$
$B_4 = 0.02353$
$B_5 = -0.0277$
$B_6 = -0.00618$
$B_7 = 0.01696$
$B_8 = -1.36$
$B_9 = -0.00045$
$B_10 = -0.03032$
$B_11 = 0.00559$
$B_12 = 0.00008$
$B_13 = -0.02118$

$B_14 = 0.00015$
$B_15 = 0.00006$
$B_16 = -0.2951$
$B_17 = -0.00806$
$B_18 = -0.0094$
$B_19 = 0.00478$
$B_20 = -0.06702$
$B_21 = -0.02722$
$B_22 = -0.6212$
$B_23 = -0.3267$
$B_24 = -0.173$

and

$x_0 = 1$
$x_1 = ResidualSugar_MISS$
$x_2 = Chlorides_MISS$
$x_3 = FreeSulfurDioxide_MISS$
$x_4 = TotalSulfurDioxide_MISS$
$x_5 = pH_MISS$
$x_6 = Sulphates_MISS$
$x_7 = Alcohol_MISS$
$x_8 = STARS_MISS$
$x_9 = FixedAcidity_CAP$
$x_10 = VolatileAcidity_CAP$
$x_11 = CitricAcid_CAP$
$x_12 = ResidualSugar_CAP$
$x_13 = Chlorides_CAP$
$x_14 = FreeSulfurDioxide_CAP$
$x_15 = TotalSulfurDioxide_CAP$
$x_16 = Density_CAP$
$x_17 = pH_CAP$
$x_18 = Sulphates_CAP$
$x_19 = Alcohol_CAP$
$x_20 = AcidIndex_CAP$
$x_21 = LabelAppeal_Positive$
$x_22 = STARS_1$
$x_23 = STARS_2$
$x_24 = STARS_3$

5.2.3.2 Coefficient Analysis

The coefficient for STARS_MISS, VolatileAcidity_CAP, AcidIndex_CAP, STARS_1, STARS_2, STARS_3 are highly significant. For a unit increase in our highly significant variables:
- STARS_MISS, we expect a decrease of $e^{(-1.36)} = 0.256661$ in the number of cases of wine that will be sold
- VolatileAcidity_CAP, we expect a decrease of $e^{(-0.03032)} = 0.970135$ in the number of cases of wine that will be sold
- AcidIndex_CAP, we expect a decrease of $e^{(-0.06702)} = 0.935176$ in the number of cases of wine that will be sold
- STARS_1, we expect a decrease of $e^{(-0.6212)} = 0.537299$ in the number of cases of wine that will be sold
- STARS_2, we expect a decrease of $e^{(-0.3267)} = 0.7213$ in the number of cases of wine that will be sold

- STARS_3, we expect a decrease of $e^{(-0.173)} = 0.841138$ in the number of cases of wine that will be sold

Sine we see that we have over dispersion, let's find out the dispersion parameter ϕ. Since the variance in the Poisson model is identical to the mean, the expectations are to have $\phi = 1$.

```
## [1] 0.8386535
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                    6.151478 model1 > model2 3.8382e-10
## AIC-corrected          6.151478 model1 > model2 3.8382e-10
## BIC-corrected          6.151478 model1 > model2 3.8382e-10
```

The Vuong test suggests that the zero-inflated Poisson model is a slight improvement over a standard Poisson model using transformed data.

5.3 Negative Binomial models

A more formal way to accommodate over-dispersion in a count data regression model is to use a negative binomial model. Hence we will explore the negative binomial model both in original data as well as transformed data.

5.3.1 Negative Binomial with original data (Model 7)

We will explore the Negative Binomial model Using original data with replacing all missing data with the means.

5.3.1.1 Negative Binomial vs Poisson Coefficients

As per the below table, it is worth noting that the classical Poisson Coefficients are similar to that of the Negative Binomial's.
One possible explanation is that if all we care about is fitting separate means to disjoint subsets of our sample, then GLMs will always yield $\hat{\mu}_j$=$\hat{y}_j$ for each subset j, so the actual error structure and parametrization of the density both become irrelevant to the estimation. In other words, Fitting orthogonal categorical factors by maximum likelihood is equivalent to fitting separate means to disjoint subsets of our sample, so this explains why Poisson and negative binomial GLMs yield the same parameter estimates

	Poisson.Coeff	Negative.Binom.Coeffi
(Intercept)	1.5259824	1.5259982
FixedAcidity	-0.0003045	-0.0003045
VolatileAcidity	-0.0334329	-0.0334338
CitricAcid	0.0077726	0.0077727
ResidualSugar	0.0000568	0.0000568
Chlorides	-0.0414139	-0.0414151
FreeSulfurDioxide	0.0001254	0.0001254
TotalSulfurDioxide	0.0000830	0.0000830
Density	-0.2823481	-0.2823537

	Poisson.Coeff	Negative.Binom.Coeffi
pH	-0.0157219	-0.0157226
Sulphates	-0.0126738	-0.0126742
Alcohol	0.0022014	0.0022014
LabelAppeal	0.1331963	0.1331958
AcidIndex	-0.0870512	-0.0870531
STARS	0.3112869	0.3112910

5.3.1.2 Interpretation Negative Binomial Model 7

Table 14: Model 7, Negative binomial Original Data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5259982	0.1954796	7.8064333	0.0000000
FixedAcidity	-0.0003045	0.0008205	-0.3711789	0.7105043
VolatileAcidity	-0.0334338	0.0065163	-5.1307821	0.0000003
CitricAcid	0.0077727	0.0058924	1.3190989	0.1871361
ResidualSugar	0.0000568	0.0001546	0.3670612	0.7135733
Chlorides	-0.0414151	0.0164504	-2.5175707	0.0118167
FreeSulfurDioxide	0.0001254	0.0000351	3.5705137	0.0003563
TotalSulfurDioxide	0.0000830	0.0000228	3.6466598	0.0002657
Density	-0.2823537	0.1919779	-1.4707613	0.1413557
pH	-0.0157226	0.0076383	-2.0583947	0.0395523
Sulphates	-0.0126742	0.0057489	-2.2046245	0.0274805
Alcohol	0.0022014	0.0014100	1.5612389	0.1184674
LabelAppeal	0.1331958	0.0060635	21.9667507	0.0000000
AcidIndex	-0.0870531	0.0045485	-19.1388952	0.0000000
STARS	0.3112910	0.0045313	68.6981875	0.0000000

From the summary output, we have the following estimated model:

$$\hat{y} = e^{B_0x_0+B_1x_1+B_2x_2+B_3x_3+B_4x_4+B_5x_5+B_6x_6+B_7x_7+B_8x_8+B_9x_9+B_{10}x_{10}+B_{11}x_{11}+B_{12}x_{12}+B_{13}x_{13}+B_{14}x_{14}}$$

where

$B_0 = 1.526$
$B_1 = -0.0003$
$B_2 = -0.03343$
$B_3 = 0.00777$
$B_4 = 0.00006$
$B_5 = -0.04142$
$B_6 = 0.00013$
$B_7 = 0.00008$
$B_8 = -0.2824$
$B_9 = -0.01572$
$B_10 = -0.01267$
$B_11 = 0.0022$
$B_12 = 0.1332$
$B_13 = -0.08705$
$B_14 = 0.3113$

and

$x_0 = 1$
$x_1 = FixedAcidity$
$x_2 = VolatileAcidity$
$x_3 = CitricAcid$
$x_4 = ResidualSugar$
$x_5 = Chlorides$
$x_6 = FreeSulfurDioxide$
$x_7 = TotalSulfurDioxide$
$x_8 = Density$
$x_9 = pH$
$x_10 = Sulphates$
$x_11 = Alcohol$
$x_12 = LabelAppeal$
$x_13 = AcidIndex$
$x_14 = STARS$

5.3.1.2 Coefficient Analysis

The coefficient for VolatileAcidity, FreeSulfurDioxide, TotalSulfurDioxide, LabelAppeal, AcidIndex, STARS are highly significant. For a unit increase in our highly significant variables:
- VolatileAcidity, we expect a decrease of $e^{(-0.03343)} = 0.967123$ in the number of cases of wine that will be sold
- FreeSulfurDioxide, we expect an increase of $e^{(0.0001254)} = 1.000125$ in the number of cases of wine that will be sold
- TotalSulfurDioxide, we expect an increase of $e^{(0.00008296)} = 1.000083$ in the number of cases of wine that will be sold
- LabelAppeal, we expect an increase of $e^{(0.1332)} = 1.142478$ in the number of cases of wine that will be sold
- AcidIndex, we expect a decrease of $e^{(-0.08705)} = 0.916631$ in the number of cases of wine that will be sold
- STARS, we expect an increase of $e^{(0.3113)} = 1.365199$ in the number of cases of wine that will be sold

In addition, Negative Binomial Model with original data has an AIC value, 46703, is slightly higher than of model 1 AIC (46700.); which was run against the original data.

5.3.1.3 Overdisperson Analysis Negative Binomial

For our model(3), we see that our Residual deviance is 14728 and degrees of freedom is 12780; our Residual deviance 1.15 greater than our Residual degrees of freedom, which similar to that of classical Poisson model (1) with original data which was also 1.15.

Sine we see that we have over dispersion, let's find out the dispersion parameter ϕ.

```
## [1] 0.851477
```

The Negative Binomial dispersion parameter for Modle (3) is 0.851477 which is similar to that of the classical Poisson Model (1). Hence theta value of the of the Negative binomial has not had much impact in improving in having the variance approximates to the mean.

5.3.2 zero-inflation model Negative Binomial (Model 8)

We will explore the Negative Binomial zero-inflation model Using original data with replacing all missing data with the means.

Next we will proceed with the Negative Binomial zero-inflation model as it is another very common occurrence when working with count data using original data.

Table 15: Model 8, Zero Inflation Negative binomial Original Data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.4435403	0.2019767	7.1470649	0.0000000
FixedAcidity	0.0003379	0.0008420	0.4013219	0.6881832
VolatileAcidity	-0.0121070	0.0067206	-1.8014845	0.0716265
CitricAcid	0.0004923	0.0060239	0.0817271	0.9348637
ResidualSugar	-0.0000770	0.0001586	-0.4855461	0.6272890
Chlorides	-0.0224070	0.0169086	-1.3251819	0.1851108
FreeSulfurDioxide	0.0000255	0.0000355	0.7177781	0.4728941
TotalSulfurDioxide	-0.0000178	0.0000226	-0.7874315	0.4310293
Density	-0.2846701	0.1982974	-1.4355717	0.1511243
pH	0.0059288	0.0078586	0.7544352	0.4505879
Sulphates	0.0001728	0.0059190	0.0291890	0.9767139
Alcohol	0.0068864	0.0014396	4.7835176	0.0000017
LabelAppeal	0.2329534	0.0063025	36.9618850	0.0000000
AcidIndex	-0.0185817	0.0048975	-3.7941060	0.0001482
STARS	0.1009197	0.0052012	19.4029779	0.0000000
Log(theta)	16.9618128	2.7238674	6.2271066	0.0000000
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.4381583	1.3373702	-3.3185712	0.0009048
FixedAcidity	0.0007553	0.0055468	0.1361773	0.8916811
VolatileAcidity	0.1937171	0.0438506	4.4176593	0.0000100
CitricAcid	-0.0296094	0.0399708	-0.7407764	0.4588290
ResidualSugar	-0.0011762	0.0010429	-1.1278339	0.2593901
Chlorides	0.0921622	0.1093477	0.8428360	0.3993202
FreeSulfurDioxide	-0.0007420	0.0002422	-3.0633958	0.0021884
TotalSulfurDioxide	-0.0009866	0.0001523	-6.4764392	0.0000000
Density	0.4801296	1.3159245	0.3648610	0.7152151
pH	0.2160267	0.0512199	4.2176339	0.0000247
Sulphates	0.1323368	0.0387665	3.4136904	0.0006409
Alcohol	0.0279102	0.0095780	2.9139894	0.0035684
LabelAppeal	0.7229464	0.0429458	16.8339107	0.0000000
AcidIndex	0.4347283	0.0258382	16.8250287	0.0000000
STARS	-2.3767989	0.0603130	-39.4077649	0.0000000

5.3.2.1 Interpretation Zero Inflation Negative Binomial Model

From this output, we have the following estimated model:

$$\hat{y} = e^{B_0x_0+B_1x_1+B_2x_2+B_3x_3+B_4x_4+B_5x_5+B_6x_6+B_7x_7+B_8x_8+B_9x_9+B_{10}x_{10}+B_{11}x_{11}+B_{12}x_{12}+B_{13}x_{13}+B_{14}x_{14}}$$

where
$B_0 = 1.444$
$B_1 = 0.00034$
$B_2 = -0.01211$
$B_3 = 0.00049$
$B_4 = -0.00008$
$B_5 = -0.02241$
$B_6 = 0.00003$
$B_7 = -0.00002$
$B_8 = -0.2847$

$B_9 = 0.00593$
$B_10 = 0.00017$
$B_11 = 0.00689$
$B_12 = 0.233$
$B_13 = -0.01858$
$B_14 = 0.1009$

and

$x_0 = 1$
$x_1 = FixedAcidity$
$x_2 = VolatileAcidity$
$x_3 = CitricAcid$
$x_4 = ResidualSugar$
$x_5 = Chlorides$
$x_6 = FreeSulfurDioxide$
$x_7 = TotalSulfurDioxide$
$x_8 = Density$
$x_9 = pH$
$x_10 = Sulphates$
$x_11 = Alcohol$
$x_12 = LabelAppeal$
$x_13 = AcidIndex$
$x_14 = STARS$

5.3.2.2 Coefficient Analysis

The coefficient for Alcohol, LabelAppeal, AcidIndex, STARS, Log(theta) are highly significant. For a unit increase in our highly significant variables:
- Alcohol, we expect an increase of $e^{(0.006886)} = 1.00691$ in the number of cases of wine that will be sold
- LabelAppeal, we expect an increase of $e^{(0.233)} = 1.262381$ in the number of cases of wine that will be sold
- AcidIndex, we expect a decrease of $e^{(-0.01858)} = 0.981592$ in the number of cases of wine that will be sold
- STARS, we expect an increase of $e^{(0.1009)} = 1.106166$ in the number of cases of wine that will be sold
- Log(theta), we expect an increase of $e^{(16.96)} = 23207823.508859$ in the number of cases of wine that will be sold
let's find out the dispersion parameter ϕ.

```
## [1] 0.4637071
```

Note that the zero inflation model output above does not indicate in any way if our zero-inflated model is an improvement over a standard Negative Binomial regression. We can determine this by running the corresponding standard Negative Binomial model and then performing a Vuong test of the two models.

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                    47.98803 model1 > model2 < 2.22e-16
## AIC-corrected          47.74231 model1 > model2 < 2.22e-16
## BIC-corrected          46.82618 model1 > model2 < 2.22e-16
```

The Vuong test suggests that the zero-inflated Negative Binomial model is slight improvement over a standard Negative Binomial model. Please note that The model1 from the vuong() function output in this case refers to the first argument in our vuong(mod3zip,nbmod3) function which is the zero-inflation model Negative Binomial Model (3)

5.3.3 Negative Binomial with transformed data. (Model 9)

In this model we will be using the basic Negative Binomial model; however using transformed data.

Table 16: Model 9, Negative binomial Transformed Data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.5701601	0.2001567	12.8407373	0.0000000
ResidualSugar_MISS	0.0228344	0.0234051	0.9756172	0.3292542
Chlorides_MISS	0.0030168	0.0232968	0.1294918	0.8969685
FreeSulfurDioxide_MISS	0.0230007	0.0236619	0.9720538	0.3310238
TotalSulfurDioxide_MISS	0.0188313	0.0224590	0.8384745	0.4017642
pH_MISS	-0.0349554	0.0299128	-1.1685787	0.2425734
Sulphates_MISS	-0.0067590	0.0175725	-0.3846332	0.7005092
Alcohol_MISS	0.0213583	0.0230609	0.9261689	0.3543582
STARS_MISS	-1.4710700	0.0237133	-62.0357383	0.0000000
FixedAcidity_CAP	-0.0005713	0.0009179	-0.6223505	0.5337114
VolatileAcidity_CAP	-0.0355022	0.0072479	-4.8982542	0.0000010
CitricAcid_CAP	0.0074305	0.0065269	1.1384492	0.2549329
ResidualSugar_CAP	0.0001348	0.0001538	0.8762319	0.3809040
Chlorides_CAP	-0.0266378	0.0161840	-1.6459356	0.0997770
FreeSulfurDioxide_CAP	0.0001600	0.0000527	3.0392257	0.0023719
TotalSulfurDioxide_CAP	0.0000838	0.0000260	3.2244581	0.0012621
Density_CAP	-0.2847684	0.1945828	-1.4634817	0.1433356
pH_CAP	-0.0136077	0.0086729	-1.5690014	0.1166476
Sulphates_CAP	-0.0119366	0.0059079	-2.0204682	0.0433348
Alcohol_CAP	0.0039557	0.0016457	2.4036559	0.0162320
AcidIndex_CAP	-0.0780093	0.0052587	-14.8344338	0.0000000
LabelAppeal_Positive	-0.0256008	0.0185458	-1.3804059	0.1674617
STARS_1	-0.7179026	0.0208079	-34.5013990	0.0000000
STARS_2	-0.3426738	0.0194404	-17.6268758	0.0000000
STARS_3	-0.1733981	0.0200576	-8.6450228	0.0000000

5.3.3.1 Interpretation Negative Binomial Model 9

Note As per the below table, even for transformed data, it is worth noting that the classical Poisson Coefficients are similar to that of the Negative Binomial's for the same reason as was the case for original data. Please refer to Section: 5.3.1.1 "Negative Binomial vs Poisson Coefficients" for more details.

In addition, the Negative Binomial model with transformed data has an improved AIC of 46370, as it is lower than the Negative Binomial model 3 AIC (46703); which was run against the original data.

	Poisson.Coeff	Negative.Binom.Coeffi
(Intercept)	2.5701252	2.5701601
ResidualSugar_MISS	0.0228341	0.0228344

	Poisson.Coeff	Negative.Binom.Coeffi
Chlorides_MISS	0.0030173	0.0030168
FreeSulfurDioxide_MISS	0.0230001	0.0230007
TotalSulfurDioxide_MISS	0.0188307	0.0188313
pH_MISS	-0.0349529	-0.0349554
Sulphates_MISS	-0.0067580	-0.0067590
Alcohol_MISS	0.0213581	0.0213583
STARS_MISS	-1.4710696	-1.4710700
FixedAcidity_CAP	-0.0005712	-0.0005713
VolatileAcidity_CAP	-0.0355011	-0.0355022
CitricAcid_CAP	0.0074304	0.0074305
ResidualSugar_CAP	0.0001348	0.0001348
Chlorides_CAP	-0.0266371	-0.0266378
FreeSulfurDioxide_CAP	0.0001600	0.0001600
TotalSulfurDioxide_CAP	0.0000838	0.0000838
Density_CAP	-0.2847644	-0.2847684
pH_CAP	-0.0136064	-0.0136077
Sulphates_CAP	-0.0119359	-0.0119366
Alcohol_CAP	0.0039558	0.0039557
AcidIndex_CAP	-0.0780062	-0.0780093
LabelAppeal_Positive	-0.0255998	-0.0256008
STARS_1	-0.7179018	-0.7179026
STARS_2	-0.3426734	-0.3426738
STARS_3	-0.1733976	-0.1733981
STARS_4	NA	NA

5.3.3.2 Interpretation Negative Binomial Model 9

From this output, we have the following estimated model:

$$\hat{y} = e^{B_0x_0+B_1x_1+B_2x_2+B_3x_3+B_4x_4+B_5x_5+B_6x_6+B_7x_7+B_8x_8+B_9x_9+B_{10}x_{10}+B_{11}x_{11}+B_{12}x_{12}+B_{13}x_{13}+B_{14}x_{14}+B_{15}x_{15}+B_{16}x_{16}+B_{17}x_{17}+B_{18}x_{18}}$$

where

$B_0 = 2.57$
$B_1 = 0.02283$
$B_2 = 0.00302$
$B_3 = 0.023$
$B_4 = 0.01883$
$B_5 = -0.03496$
$B_6 = -0.00676$
$B_7 = 0.02136$
$B_8 = -1.471$
$B_9 = -0.00057$
$B_10 = -0.0355$
$B_11 = 0.00743$
$B_12 = 0.00013$
$B_13 = -0.02664$
$B_14 = 0.00016$
$B_15 = 0.00008$
$B_16 = -0.2848$
$B_17 = -0.01361$
$B_18 = -0.01194$

$B_19 = 0.00396$
$B_20 = -0.07801$
$B_21 = -0.0256$
$B_22 = -0.7179$
$B_23 = -0.3427$
$B_24 = -0.1734$

and

$x_0 = 1$
$x_1 = ResidualSugar_MISS$
$x_2 = Chlorides_MISS$
$x_3 = FreeSulfurDioxide_MISS$
$x_4 = TotalSulfurDioxide_MISS$
$x_5 = pH_MISS$
$x_6 = Sulphates_MISS$
$x_7 = Alcohol_MISS$
$x_8 = STARS_MISS$
$x_9 = FixedAcidity_CAP$
$x_10 = VolatileAcidity_CAP$
$x_11 = CitricAcid_CAP$
$x_12 = ResidualSugar_CAP$
$x_13 = Chlorides_CAP$
$x_14 = FreeSulfurDioxide_CAP$
$x_15 = TotalSulfurDioxide_CAP$
$x_16 = Density_CAP$
$x_17 = pH_CAP$
$x_18 = Sulphates_CAP$
$x_19 = Alcohol_CAP$
$x_20 = AcidIndex_CAP$
$x_21 = LabelAppeal_Positive$
$x_22 = STARS_1$
$x_23 = STARS_2$
$x_24 = STARS_3$

5.3.3.3 Coefficient Analysis

The coefficient for STARS_MISS, VolatileAcidity_CAP, AcidIndex_CAP, STARS_1, STARS_2, STARS_3 are highly significant. For a unit increase in our highly significant variables:
- STARS_MISS, we expect a decrease of $e^{(-1.471)} = 0.229696$ in the number of cases of wine that will be sold
- VolatileAcidity_CAP, we expect a decrease of $e^{(-0.0355)} = 0.965123$ in the number of cases of wine that will be sold
- AcidIndex_CAP, we expect a decrease of $e^{(-0.07801)} = 0.924955$ in the number of cases of wine that will be sold
- STARS_1, we expect a decrease of $e^{(-0.7179)} = 0.487776$ in the number of cases of wine that will be sold
- STARS_2, we expect a decrease of $e^{(-0.3427)} = 0.709851$ in the number of cases of wine that will be sold
- STARS_3, we expect a decrease of $e^{(-0.1734)} = 0.840801$ in the number of cases of wine that will be sold

5.3.3.4 Overdisperson Analysis Negative Binomial Model 9

For our model(4), we see that our Residual deviance is 14375 and degrees of freedom is 12770; our Residual deviance 1.12 greater than our Residual degrees of freedom, which is similar to that of classical Poisson model (1) with transformed data which was also 1.12.

Sine we see that we have over dispersion, let's find out the dispersion parameter ϕ.

```
## [1] 0.9667395
```

Our dispersion parameter for Modle (4) is 0.9667395 which is much closer to 1 than the dispersion parameter of our Modle (3). However, it is slightly lower than of the classical Poisson model using transformed data.

5.3.4 zero-inflation model NB with transformed data (Model 10)

In this model we will be using the Negative Binomial zero-inflation model; however using transformed data. Next we will proceed with the Negative Binomial zero-inflation model as it is another very common occurrence when working with count data using transformed data.

Table 18: Model 10, Zero Inflation Negative binomial Transformed Data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.4735840	0.2058856	12.0143603	0.0000000
ResidualSugar_MISS	0.0218495	0.0240010	0.9103577	0.3626339
Chlorides_MISS	0.0073860	0.0239380	0.3085476	0.7576657
FreeSulfurDioxide_MISS	0.0200984	0.0242150	0.8299999	0.4065388
TotalSulfurDioxide_MISS	0.0235234	0.0230599	1.0200968	0.3076826
pH_MISS	-0.0276855	0.0307600	-0.9000500	0.3680936
Sulphates_MISS	-0.0061929	0.0180538	-0.3430229	0.7315812
Alcohol_MISS	0.0170016	0.0236314	0.7194477	0.4718651
STARS_MISS	-1.3604209	0.0262706	-51.7849095	0.0000000
FixedAcidity_CAP	-0.0004503	0.0009424	-0.4778542	0.6327539
VolatileAcidity_CAP	-0.0303169	0.0074597	-4.0641159	0.0000482
CitricAcid_CAP	0.0055859	0.0066993	0.8338070	0.4043897
ResidualSugar_CAP	0.0000794	0.0001577	0.5033167	0.6147416
Chlorides_CAP	-0.0211686	0.0165903	-1.2759606	0.2019695
FreeSulfurDioxide_CAP	0.0001529	0.0000538	2.8414076	0.0044915
TotalSulfurDioxide_CAP	0.0000593	0.0000264	2.2440430	0.0248296
Density_CAP	-0.2951144	0.2000325	-1.4753324	0.1401232
pH_CAP	-0.0080612	0.0089139	-0.9043340	0.3658183
Sulphates_CAP	-0.0093975	0.0060705	-1.5480693	0.1216056
Alcohol_CAP	0.0047761	0.0016869	2.8313310	0.0046355
AcidIndex_CAP	-0.0670194	0.0055600	-12.0539382	0.0000000
LabelAppeal_Positive	-0.0272136	0.0190314	-1.4299307	0.1527369
STARS_1	-0.6211366	0.0219117	-28.3472611	0.0000000
STARS_2	-0.3266777	0.0194769	-16.7726002	0.0000000
STARS_3	-0.1729904	0.0200575	-8.6247291	0.0000000
Log(theta)	17.2548910	NaN	NaN	NaN

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.806541	0.0846307	-33.16222	0

5.3.4.1 Interpretation Zero Inflation Negative Binomial Model 10

From this output, we have the following estimated model:

$$\hat{y} = e^{B_0x_0+B_1x_1+B_2x_2+B_3x_3+B_4x_4+B_5x_5+B_6x_6+B_7x_7+B_8x_8+B_9x_9+B_{10}x_{10}+B_{11}x_{11}+B_{12}x_{12}+B_{13}x_{13}+B_{14}x_{14}+B_{15}x_{15}+B_{16}x_{16}+B_{17}x_{17}+B_{18}x_{18}}$$

where

$B_0 = 2.474$ $
$B_1 = 0.02185$ $
$B_2 = 0.00739$
$B_3 = 0.0201$
$B_4 = 0.02352$
$B_5 = -0.02769$
$B_6 = -0.00619$
$B_7 = 0.017$
$B_8 = -1.36$
$B_9 = -0.00045$
$B_10 = -0.03032$
$B_11 = 0.00559$
$B_12 = 0.00008$

$B_13 = -0.02117$
$B_14 = 0.00015$
$B_15 = 0.00006$
$B_16 = -0.2951$
$B_17 = -0.00806$
$B_18 = -0.0094$
$B_19 = 0.00478$
$B_20 = -0.06702$
$B_21 = -0.02721$
$B_22 = -0.6211$
$B_23 = -0.3267$
$B_24 = -0.173$

and

$x_0 = 1$
$x_1 = ResidualSugar_MISS$
$x_2 = Chlorides_MISS$
$x_3 = FreeSulfurDioxide_MISS$
$x_4 = TotalSulfurDioxide_MISS$
$x_5 = pH_MISS$
$x_6 = Sulphates_MISS$
$x_7 = Alcohol_MISS$
$x_8 = STARS_MISS$
$x_9 = FixedAcidity_CAP$
$x_10 = VolatileAcidity_CAP$
$x_11 = CitricAcid_CAP$
$x_12 = ResidualSugar_CAP$
$x_13 = Chlorides_CAP$
$x_14 = FreeSulfurDioxide_CAP$
$x_15 = TotalSulfurDioxide_CAP$
$x_16 = Density_CAP$
$x_17 = pH_CAP$
$x_18 = Sulphates_CAP$
$x_19 = Alcohol_CAP$
$x_20 = AcidIndex_CAP$
$x_21 = LabelAppeal_Positive$
$x_22 = STARS_1$
$x_23 = STARS_2$
$x_24 = STARS_3$

5.3.4.2 Coefficient Analysis

The coefficient for STARS_MISS, VolatileAcidity_CAP, AcidIndex_CAP, STARS_1, STARS_2, STARS_3 are highly significant. For a unit increase in our highly significant variables:
- STARS_MISS, we expect a decrease of $e^{(-1.36)} = 0.256661$ in the number of cases of wine that will be sold
- VolatileAcidity_CAP, we expect a decrease of $e^{(-0.03032)} = 0.970135$ in the number of cases of wine that will be sold
- AcidIndex_CAP, we expect a decrease of $e^{(-0.06702)} = 0.935176$ in the number of cases of wine that will be sold
- STARS_1, we expect a decrease of $e^{(-0.6211)} = 0.537353$ in the number of cases of wine that will be sold
- STARS_2, we expect a decrease of $e^{(-0.3267)} = 0.7213$ in the number of cases of wine that will be sold

- STARS_3, we expect a decrease of $e^{(-0.173)} = 0.841138$ in the number of cases of wine that will be sold

```
## [1] 0.8386927
```

Again, Please note that the zero inflation model output above does not indicate in any way if our zero-inflated model is an improvement over a standard Negative Binomial regression. We can determine this by running the corresponding standard Negative Binomial model and then performing a Vuong test of the two models against the transformed data.

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                   6.163416 model1 > model2 3.5596e-10
## AIC-corrected         6.163416 model1 > model2 3.5596e-10
## BIC-corrected         6.163416 model1 > model2 3.5596e-10
```

The Vuong test suggests that the zero-inflated Negative Binomial model is slight improvement over a standard Negative Binomial model uing the transformed data. Please note that The model1 from the vuong() function output in this case refers to the first argument in our vuong(mod4zip,nbmod4) function which is the zero-inflation model Negative Binomial Model (4)

5.4 Linear Regression models

Although it is highly recommended for continuous variables instead of count variables, we will also create two linear regression models.

5.4.1 Linear Regression Model with original data (Model 11)

We will explore the Linear models Using original data with replacing all missing data with the means.

Table 19: Model 11, Linear Model Original Data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.9860606	0.4487066	8.8834462	0.0000000
FixedAcidity	0.0000016	0.0018845	0.0008531	0.9993193
VolatileAcidity	-0.0992321	0.0149784	-6.6250066	0.0000000
CitricAcid	0.0208544	0.0136217	1.5309619	0.1258035
ResidualSugar	0.0002012	0.0003559	0.5653283	0.5718604
Chlorides	-0.1242663	0.0377662	-3.2904154	0.0010031
FreeSulfurDioxide	0.0003153	0.0000809	3.8966240	0.0000980
TotalSulfurDioxide	0.0002264	0.0000520	4.3532925	0.0000135
Density	-0.8011986	0.4418769	-1.8131718	0.0698288
pH	-0.0345267	0.0175380	-1.9686775	0.0490117
Sulphates	-0.0327067	0.0132170	-2.4745892	0.0133518
Alcohol	0.0109425	0.0032338	3.3837384	0.0007172
LabelAppeal	0.4326069	0.0136669	31.6536498	0.0000000
AcidIndex	-0.2083706	0.0092123	-22.6187866	0.0000000
STARS	0.9767209	0.0104537	93.4330525	0.0000000

5.4.1.1 Interpretation of Linear Model 11

Based on the summary for Linear Model 5, below are the characteristics :

- The Residual standard error is 1.3242
- Multiple R-squared: 0.528
- Adjusted R-squared: 0.5275
- F-statistic: 1021 on 14 and 12780 DF
- p-value: < 2.2e-16

Based on the available coefficients, we can make the following observations:

- Positive Impact - The following variables have a positive impact on TARGET, meaning an increase in the values of these variables leads to an increase in the number of cases sold: STARS, LabelAppeal, Alcohol, TotalSulfurDioxide, FreeSulfurDioxide, ResidualSugar, CitricAcid, FixedAcidity
- Negative Impact - The following variables have a negative impact on TARGET, meaning an increase in the values of these variables leads to an decrease in the number of cases sold: AcidIndex, Sulphates, pH, Density, Chlorides, VolatileAcidity
- The following variables have a'significant' impact. These are the more important predictors for TARGET: STARS, AcidIndex, LabelAppeal, Alcohol, Sulphates, pH, TotalSulfurDioxide, FreeSulfurDioxide, Chlorides, VolatileAcidity
- Finally, the Linear Model equation is given by the following:

3.9861 + 2e-06 * FixedAcidity - 0.099232 * VolatileAcidity + 0.020854 * CitricAcid + 0.000201 * ResidualSugar - 0.124266 * Chlorides + 0.000315 * FreeSulfurDioxide + 0.000226 * TotalSulfurDioxide - 0.801199 * Density - 0.034527 * pH - 0.032707 * Sulphates + 0.010942 * Alcohol + 0.432607 * LabelAppeal - 0.208371 * AcidIndex + 0.976721 * STARS

5.4.2 Linear Regression Model with transformed data (Model 12)

In this model we will be using the Linear Regression model; however using transformed data.

Table 20: Model 11, Linear Model Transformed Data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.9379986	0.4778719	16.6111449	0.0000000
ResidualSugar_MISS	0.0629311	0.0564913	1.1139977	0.2653011
Chlorides_MISS	0.0062078	0.0555999	0.1116507	0.9111021
FreeSulfurDioxide_MISS	0.0644372	0.0562398	1.1457581	0.2519167
TotalSulfurDioxide_MISS	0.0499841	0.0538330	0.9285033	0.3531641
pH_MISS	-0.0856250	0.0699034	-1.2249033	0.2206342
Sulphates_MISS	-0.0232490	0.0413228	-0.5626187	0.5737044
Alcohol_MISS	0.0614022	0.0549354	1.1177175	0.2637087
STARS_MISS	-4.0920344	0.0605148	-67.6204176	0.0000000
FixedAcidity_CAP	-0.0011901	0.0021754	-0.5470873	0.5843283
VolatileAcidity_CAP	-0.1065427	0.0172030	-6.1932513	0.0000000
CitricAcid_CAP	0.0222018	0.0155378	1.4288832	0.1530623
ResidualSugar_CAP	0.0003782	0.0003646	1.0375108	0.2995175
Chlorides_CAP	-0.0775422	0.0383951	-2.0195853	0.0434473

	Estimate	Std. Error	t value	Pr(>|t|)
FreeSulfurDioxide_CAP	0.0004803	0.0001261	3.8085988	0.0001404
TotalSulfurDioxide_CAP	0.0002303	0.0000616	3.7368527	0.0001872
Density_CAP	-0.9170533	0.4641765	-1.9756563	0.0482152
pH_CAP	-0.0381391	0.0206118	-1.8503547	0.0642855
Sulphates_CAP	-0.0337154	0.0140648	-2.3971470	0.0165376
Alcohol_CAP	0.0131115	0.0038937	3.3673345	0.0007612
AcidIndex_CAP	-0.2108365	0.0117680	-17.9161506	0.0000000
LabelAppeal_Positive	-0.0769502	0.0439408	-1.7512241	0.0799313
STARS_1	-2.7703255	0.0607380	-45.6110805	0.0000000
STARS_2	-1.5855052	0.0599159	-26.4621882	0.0000000
STARS_3	-0.8683451	0.0624815	-13.8976314	0.0000000

5.4.2.1 Interpretation of Linear Model 12

Based on the summary for Linear Model 6, below are the characteristics :

- The Residual standard error is 1.3667
- Multiple R-squared: 0.4976
- Adjusted R-squared: 0.4967
- F-statistic: 527 on 24 and 12770 DF
- p-value: < 2.2e-16

Based on the available coefficients, we can make the following observations:

- Positive Impact - The following variables have a positive impact on TARGET, meaning an increase in the values of these variables leads to an increase in the number of cases sold: Alcohol_CAP, TotalSulfurDioxide_CAP, FreeSulfurDioxide_CAP, ResidualSugar_CAP, CitricAcid_CAP, Alcohol_MISS, TotalSulfurDioxide_MISS, FreeSulfurDioxide_MISS, Chlorides_MISS, ResidualSugar_MISS
- Negative Impact - The following variables have a negative impact on TARGET, meaning an increase in the values of these variables leads to an decrease in the number of cases sold: STARS_3, STARS_2, STARS_1, LabelAppeal_Positive, AcidIndex_CAP, Sulphates_CAP, pH_CAP, Density_CAP, Chlorides_CAP, VolatileAcidity_CAP, FixedAcidity_CAP, STARS_MISS, Sulphates_MISS, pH_MISS
- The following variables have a'significant' impact. These are the more important predictors for TARGET: STARS_3, STARS_2, STARS_1, AcidIndex_CAP, Alcohol_CAP, Sulphates_CAP, Density_CAP, TotalSulfurDioxide_CAP, FreeSulfurDioxide_CAP, Chlorides_CAP, VolatileAcidity_CAP, STARS_MISS
- Finally, the Linear Model equation is given by the following:

7.938 + 0.062931 * ResidualSugar_MISS + 0.006208 * Chlorides_MISS + 0.064437 * FreeSulfurDioxide_MISS + 0.049984 * TotalSulfurDioxide_MISS - 0.085625 * pH_MISS - 0.023249 * Sulphates_MISS + 0.061402 * Alcohol_MISS - 4.092034 * STARS_MISS - 0.00119 * FixedAcidity_CAP - 0.106543 * VolatileAcidity_CAP + 0.022202 * CitricAcid_CAP + 0.000378 * ResidualSugar_CAP - 0.077542 * Chlorides_CAP + 0.00048 * FreeSulfurDioxide_CAP + 0.00023 * TotalSulfurDioxide_CAP - 0.917053 * Density_CAP - 0.038139 * pH_CAP - 0.033715 * Sulphates_CAP + 0.013111 * Alcohol_CAP - 0.210836 * AcidIndex_CAP - 0.07695 * LabelAppeal_Positive - 2.770326 * STARS_1 - 1.585505 * STARS_2 - 0.868345 * STARS_3

6 Model Selection

Before we proceed with our model selection, let take a quick look at our models inventory. We have 12 models using a combination of three different type distributions. First we created our models using GLM distribution; then we created few using the zero Augmented distribution, and finally the Linear distribution.

6.1 Model Selection Strategy

Our models selection will be based on the best AIC/ phi =Dispersion parameter for the GLM, AIC for Linear regression; and Vuong test for the zero Augmented distribution.

Below is summary table of model selection strategy:

Table 21: Model Selection Strategy

Distribution.Type	Model.Description	Comparaison.KPI
Classical Poisson	Poisson using original data	AIC
	Poisson using Transformed data	AIC
Quasi-Poisson	Quasi Poisson using original data	phi =Dispersion parameter
	Quasi Poisson using transformed data	phi =Dispersion parameter
Negative Binomial	NB using original data	AIC
	NB using transformed data	AIC
zero-inflation Poisson	zero inflated Pois using original data	Vuong test
	zero inflated Pois using Transforemed data	Vuong test
zero-inflation NB	zero inflated NB using original data	Vuong test
	zero inflated NB using transformed data	Vuong test
LM	linear regression using original data	AIC
	linear regression using transformed data	AIC

Below is a Model Selection KPI table. It is a summary of the major indicators we will use to select the best fit. To selefct the best model we will be using a combination of the AIC, Dispersion parameter, as well as the Vuong closeness test which is specifically for the zero inflation distributions.
However, since our data is count data and the problem of dispersion occurs more frequently in count data set, we will be using Dispersion parameter first in our process elimination, followed by AIC, and Voung test.
Hence, the "Model Selection KPI" table below is sorted using the Dispersion parameter.

Table 22: Model Selection KPI

Model.Type	Dispersion.parameter	AIC	Vuong.Selected
Linear model with transformed data	1.8678630	44321.76	
Linear model with original data	1.7533830	43508.94	
Pois with transformed data	0.9667917	46368	
Quasi-Poisson with transformed data	0.9667917	Undefined	
Negative binomial /transformed data	0.9667395	46370	
Quasi-Poisson with Original data	0.8515200	Undefined	
Pois with original data	0.8515130	46700	
Negative binomial /original data	0.8514770	46703	
zero inflation NB with transformed data	0.8386927	Undefined	zero inflation NB with transformed data
zero inflation Poisson with transformed data	0.8386535	Undefined	zero inflation Poisson with transformed
zero inflation NB with orig data	0.4637071	Undefined	zero inflation NB with orig data
zero inflation Poisson with orig data	0.4636815	Undefined	zero inflation Poisson with orig data

Therefore, from the above table, we can easily eliminate the Linear models both for in the original and transformed data as they respectively have a dispersion parameter of 1.867863 and 1.753383 which are much higher than 1.

Next we will eliminate the zero inflation Negative Binomial and Poisson for the original as they respectively have a dispersion parameter of 0.4637071 and 0.4636815which are much lower than 1.

We will also eliminate the zero inflation Negative Binomial and Poisson for the transformed data as they respectively have a dispersion parameter of 0.8386927 and 0.8386535 which are not close to 1 compared to the rest of the models.

Also, based on dispersion parameter, we will eliminate the Poission, Quasi-Poisson, and Negative binomial with original data as they respectively have a dispersion parameter of 0.851513, 0.85152, and 0.851477 which are not close to 1 compared to the rest of the models.

Finally we are left with the following 3 models:

Poisson with transformed data, with Dispersion parameter = 0.9667917 Quasi-Poisson with transformed data with Dispersion parameter = 0.9667917 Negative binomial /transformed data Dispersion parameter = 0.9667395

Since we have a virtual tie in the remaining 3 models from dispersion parameter perspective, we will use the second metric, AIC, as defining factor for our remaining 3 model selection. Hence, the Poisson model with transformed data as it has an AIC of 46368 compared to the Negative Binomial which is 46370.

7 Prediction Using Evaluation Data

Now that we have selected the final model, we will go ahead and use this model to predict the results for the evaluation dataset. After transforming the data to meet the needs of the trained model, we will apply the model.

7.1 Tranformation of Evaluation Data

First we need to transform the evaluation dataset to account for all the predictors that were used in the model.

7.2 Model Output

For ease of display we will display, in transposed format, only the first six rows as we have 42 variables.

First six Records from output

Table 23: Transposed Model Output / Results

	1	2	3	4	5	6
IN	3.00000	21.000000	37.000000	39.0000	47.00000	62.00000

	1	2	3	4	5	6
TARGET	1.00000	1.000000	1.000000	1.0000	1.00000	1.00000
FixedAcidity	5.40000	11.400000	15.900000	11.6000	3.80000	9.00000
VolatileAcidity	-0.86000	0.210000	1.190000	0.3200	0.22000	-0.21000
CitricAcid	0.27000	0.280000	1.140000	0.5500	0.31000	0.04000
ResidualSugar	-10.70000	1.200000	31.900000	-50.9000	-7.70000	51.40000
Chlorides	0.09200	0.038000	-0.299000	0.0760	0.03900	0.23700
FreeSulfurDioxide	23.00000	70.000000	115.000000	35.0000	40.00000	-213.00000
TotalSulfurDioxide	398.00000	53.000000	381.000000	83.0000	129.00000	-527.00000
Density	0.98527	1.028990	1.034160	1.0002	0.90610	0.99516
pH	5.02000	2.540000	2.990000	3.3200	4.72000	3.16000
Sulphates	0.64000	-0.070000	0.310000	2.1800	-0.64000	0.70000
Alcohol	12.30000	4.800000	11.400000	-0.5000	10.90000	14.70000
LabelAppeal	-1.00000	0.000000	1.000000	0.0000	0.00000	1.00000
AcidIndex	6.00000	10.000000	7.000000	12.0000	7.00000	10.00000
STARS	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
ResidualSugar_MISS	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
Chlorides_MISS	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
FreeSulfurDioxide_MISS	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
TotalSulfurDioxide_MISS	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
pH_MISS	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
Sulphates_MISS	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
Alcohol_MISS	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
STARS_MISS	1.00000	1.000000	1.000000	1.0000	1.00000	1.00000
FixedAcidity_CAP	5.40000	11.400000	17.500000	11.6000	3.80000	9.00000
VolatileAcidity_CAP	-1.04600	0.210000	1.190000	0.3200	0.22000	-0.21000
CitricAcid_CAP	0.27000	0.280000	1.140000	0.5500	0.31000	0.04000
ResidualSugar_CAP	-10.70000	1.200000	31.900000	-51.9000	-7.70000	61.56500
Chlorides_CAP	0.09200	0.038000	-0.479300	0.0760	0.03900	0.23700
FreeSulfurDioxide_CAP	23.00000	70.000000	115.000000	35.0000	40.00000	-216.30000
TotalSulfurDioxide_CAP	398.00000	53.000000	381.000000	83.0000	129.00000	-253.00000
Density_CAP	0.98527	1.040107	1.040107	1.0002	0.95028	0.99516
pH_CAP	4.37300	2.540000	2.990000	3.3200	4.37300	3.16000
Sulphates_CAP	0.64000	-0.070000	0.310000	2.0300	-0.99000	0.70000
Alcohol_CAP	12.30000	4.800000	11.400000	4.3000	10.90000	14.70000
AcidIndex_CAP	6.00000	10.000000	7.000000	10.0000	7.00000	10.00000
LabelAppeal_Positive	1.00000	1.000000	1.000000	1.0000	1.00000	0.00000
STARS_1	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
STARS_2	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
STARS_3	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000
STARS_4	0.00000	0.000000	0.000000	0.0000	0.00000	0.00000

8 Conclusion

After fitting multiple models using the classical Linear, classical Poisson, and the Binomial distributions using original data and transformed data, we think that the Poisson model has performed well once we have treated the outliers and missing data.

We also felt confident that the Negative Binomial would perform good as well as it has the same dispersion parameter as classical Poisson. However, the NB AIC was bit higher by .000043 which could be negligible.

In addition we felt confident that Quasi-Poisson would perform well as its dispersion parameter was .96 close to 1. However, we were not comfortable selecting the Quasi-Poisson as we could not generate the AIC value.

The zero inflation models for both Poisson and Negative yielded to promising results especially when using the Voung test. However, lack of AIC and its lower dispersion parameter had made us reconsider our decision in favor of the Poisson.

Over all, we were little bit overwhelmed with analyzing about 12 models. However, we are very satisfied with our Poisson model selection especially that it had leveraged our data preparation and transformation efforts.

Appendix A: DATA621 Homework 05 R Code

```
# changed lib path...

if (!require("ggplot2",character.only = TRUE)) (install.packages("ggplot2",repos = "http://cran.us.r-pro
if (!require("MASS",character.only = TRUE)) (install.packages("MASS",repos = "http://cran.us.r-project.
if (!require("knitr",character.only = TRUE)) (install.packages("knitr",repos = "http://cran.us.r-projec
if (!require("xtable",character.only = TRUE)) (install.packages("xtable",repos = "http://cran.us.r-proj
if (!require("dplyr",character.only = TRUE)) (install.packages("dplyr",repos = "http://cran.us.r-projec
if (!require("psych",character.only = TRUE)) (install.packages("psych",repos = "http://cran.us.r-projec
if (!require("stringr",character.only = TRUE)) (install.packages("stringr",repos = "http://cran.us.r-pro
if (!require("car",character.only = TRUE)) (install.packages("car",repos = "http://cran.us.r-project.org
if (!require("faraway",character.only = TRUE)) (install.packages("faraway",repos = "http://cran.us.r-pro
if (!require("aod",character.only = TRUE)) (install.packages("aod",repos = "http://cran.us.r-project.org
if (!require("ISLR",character.only = TRUE)) (install.packages("ISLR",repos = "http://cran.us.r-project.
if (!require("AUC",character.only = TRUE)) (install.packages("AUC",repos = "http://cran.us.r-project.org
if (!require("ROCR",character.only = TRUE)) (install.packages("ROCR",repos = "http://cran.us.r-project.
if (!require("leaps",character.only = TRUE)) (install.packages("leaps",repos = "http://cran.us.r-projec
if (!require("pander",character.only = TRUE)) (install.packages("pander",repos = "http://cran.us.r-proj

library(ggplot2)
library(MASS)
library(knitr)
library(xtable)
library(dplyr)
library(psych)
library(stringr)
library(car)
library(faraway)
library(aod)
library(Rcpp)
library(leaps)
library(ISLR)
library(AUC)
library(ROCR)
library(pander)

vartypes<- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW5/vartypes.csv")
kable(vartypes)

winedata <- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW5/wine-training-da
variables<- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW5/vars.csv")
#kable(variables, caption = "Variable Description")

pander::pander(variables, split.cells = c(20, 60, 40), split.table = Inf, justify = 'left', caption = "


library(dplyr)

winedata<- select(winedata, -(ï..INDEX))
str(winedata)
ds_stats <- psych::describe(winedata, skew = FALSE, na.rm = TRUE)[c(3:6)]
```

```
ds_stats0<- ds_stats
ds_stats <- cbind(VARIABLE_NAME = rownames(ds_stats), ds_stats)
kable(ds_stats0, caption = "Data Summary")

Variable<- rownames(ds_stats)
fun <- function(x) sum(!complete.cases(x))
Missing <- sapply(winedata[Variable], FUN = fun)
fun <- function(x, y) cor(y, x, use = "na.or.complete")
Correlation <- sapply(winedata[Variable], FUN = fun, y=winedata$TARGET)
ds_stats2 <- data.frame(cbind( Missing, Correlation))
kable(ds_stats2, caption = "Missing Data and Data Correlation")

show_charts <- function(x, ...) {

    par(mfrow=c(2,3))

    xlabel <- unlist(str_split(deparse(substitute(x)), pattern = "\\$"))[2]
#    ylabel <- unlist(str_split(deparse(substitute(y)), pattern = "\\$"))[2]

    hist(x,main=xlabel)
    boxplot(x,main=xlabel)

    y<-log(x)
    boxplot(y,main='log transform')
    y<-sqrt(x)
    boxplot(y,main='sqrt transform')
    y<-sin(x)
    boxplot(y,main='sin transform')
    y<-(x)^(1/9)
    boxplot(y,main='ninth transform')
}

show_charts(winedata$FixedAcidity)

#show_charts(winedata$FixedAcidity)
#show_charts(winedata$VolatileAcidity)
#show_charts(winedata$CitricAcid)
#show_charts(winedata$ResidualSugar)
#show_charts(winedata$Chlorides)
#show_charts(winedata$FreeSulfurDioxide)
#show_charts(winedata$TotalSulfurDioxide)
#show_charts(winedata$Density)
#show_charts(winedata$pH)
#show_charts(winedata$Sulphates)
#show_charts(winedata$Alcohol)
#show_charts(winedata$LabelAppeal)
#show_charts(winedata$AcidIndex)
#show_charts(winedata$STARS)


winedata$ResidualSugar_MISS <- ifelse(is.na(winedata$ResidualSugar), 1, 0)
winedata$Chlorides_MISS <- ifelse(is.na(winedata$Chlorides), 1, 0)
winedata$FreeSulfurDioxide_MISS <- ifelse(is.na(winedata$FreeSulfurDioxide), 1, 0)
```

```
winedata$TotalSulfurDioxide_MISS <- ifelse(is.na(winedata$TotalSulfurDioxide), 1, 0)
winedata$pH_MISS <- ifelse(is.na(winedata$pH), 1, 0)
winedata$Sulphates_MISS <- ifelse(is.na(winedata$Sulphates), 1, 0)
winedata$Alcohol_MISS <- ifelse(is.na(winedata$Alcohol), 1, 0)
winedata$STARS_MISS <- ifelse(is.na(winedata$STARS), 1, 0)



winedata$ResidualSugar[is.na(winedata$ResidualSugar)] <- mean(winedata$ResidualSugar, na.rm = T)
winedata$Chlorides[is.na(winedata$Chlorides)] <- mean(winedata$Chlorides, na.rm = T)
winedata$FreeSulfurDioxide[is.na(winedata$FreeSulfurDioxide)] <- mean(winedata$FreeSulfurDioxide, na.rm
winedata$TotalSulfurDioxide[is.na(winedata$TotalSulfurDioxide)] <- mean(winedata$TotalSulfurDioxide, na
winedata$pH[is.na(winedata$pH)] <- mean(winedata$pH, na.rm = T)
winedata$Sulphates[is.na(winedata$Sulphates)] <- mean(winedata$Sulphates, na.rm = T)
winedata$Alcohol[is.na(winedata$Alcohol)] <- mean(winedata$Alcohol, na.rm = T)
winedata$STARS[is.na(winedata$STARS)] <- 0


treat_outliers <- function(x) {
qnt <- quantile(x, probs=c(.25, .75), na.rm = T)
caps <- quantile(x, probs=c(.05, .95), na.rm = T)
H <- 1.5 * IQR(x, na.rm = T)
x[x < (qnt[1] - H)] <- caps[1]
x[x > (qnt[2] + H)] <- caps[2]

return(x)
}

winedata$FixedAcidity_CAP <- treat_outliers(winedata$FixedAcidity)
winedata$VolatileAcidity_CAP <- treat_outliers(winedata$VolatileAcidity)
winedata$CitricAcid_CAP <- treat_outliers(winedata$CitricAcid)
winedata$ResidualSugar_CAP <- treat_outliers(winedata$ResidualSugar)
winedata$Chlorides_CAP <- treat_outliers(winedata$Chlorides)
winedata$FreeSulfurDioxide_CAP <- treat_outliers(winedata$FreeSulfurDioxide)
winedata$TotalSulfurDioxide_CAP <- treat_outliers(winedata$TotalSulfurDioxide)
winedata$Density_CAP <- treat_outliers(winedata$Density)
winedata$pH_CAP <- treat_outliers(winedata$pH)
winedata$Sulphates_CAP <- treat_outliers(winedata$Sulphates)
winedata$Alcohol_CAP <- treat_outliers(winedata$Alcohol)
winedata$AcidIndex_CAP <- treat_outliers(winedata$AcidIndex)


#Lets see how the new variables look in boxplots.

# par(mfrow=c(2,3))
#
# boxplot(winedata$FixedAcidity_CAP)
# boxplot(winedata$VolatileAcidity_CAP)
# boxplot(winedata$ResidualSugar_CAP)
# boxplot(winedata$Chlorides_CAP)
# boxplot(winedata$FreeSulfurDioxide_CAP)
# boxplot(winedata$TotalSulfurDioxide_CAP)
# boxplot(winedata$Density_CAP)
```

```
# boxplot(winedata$pH_CAP)
# boxplot(winedata$Sulphates_CAP)
# boxplot(winedata$Alcohol_CAP)
# boxplot(winedata$AcidIndex_CAP)


# In the second set, we will use the sin transformation as identified in the data exploration and creat
#
# - FixedAcidity_SIN
# - VolatileAcidity_SIN
# - CitricAcid_SIN
# - ResidualSugar_SIN
# - Chlorides_SIN
# - FreeSulfurDioxide_SIN
# - TotalSulfurDioxide_SIN
# - Density_SIN
# - pH_SIN
# - Sulphates_SIN
# - Alcohol_SIN
# - AcidIndex_SIN
#
# winedata$FixedAcidity_SIN <- sin(winedata$FixedAcidity)
# winedata$VolatileAcidity_SIN <- sin(winedata$VolatileAcidity)
# winedata$CitricAcid_SIN <- sin(winedata$CitricAcid)
# winedata$ResidualSugar_SIN <- sin(winedata$ResidualSugar)
# winedata$Chlorides_SIN <- sin(winedata$Chlorides)
# winedata$FreeSulfurDioxide_SIN <- sin(winedata$FreeSulfurDioxide)
# winedata$TotalSulfurDioxide_SIN <- sin(winedata$TotalSulfurDioxide)
# winedata$Density_SIN <- sin(winedata$Density)
# winedata$pH_SIN <- sin(winedata$pH)
# winedata$Sulphates_SIN <- sin(winedata$Sulphates)
# winedata$Alcohol_SIN <- sin(winedata$Alcohol)
# winedata$AcidIndex_SIN <- sin(winedata$AcidIndex)


# par(mfrow=c(2,3))
#
# boxplot(winedata$FixedAcidity_SIN)
# boxplot(winedata$VolatileAcidity_SIN)
# boxplot(winedata$ResidualSugar_SIN)
# boxplot(winedata$Chlorides_SIN)
# boxplot(winedata$FreeSulfurDioxide_SIN)
# boxplot(winedata$TotalSulfurDioxide_SIN)
# boxplot(winedata$Density_SIN)
# boxplot(winedata$pH_SIN)
# boxplot(winedata$Sulphates_SIN)
# boxplot(winedata$Alcohol_SIN)
# boxplot(winedata$AcidIndex_SIN)


winedata$LabelAppeal_Positive <- ifelse(winedata$FreeSulfurDioxide >=0, 1, 0)

winedata$STARS_1 <- ifelse(winedata$STARS == 1, 1, 0)
```

```
winedata$STARS_2 <- ifelse(winedata$STARS == 2, 1, 0)
winedata$STARS_3 <- ifelse(winedata$STARS == 3, 1, 0)
winedata$STARS_4 <- ifelse(winedata$STARS == 4, 1, 0)


#write.csv(winedata, file = "D:/CUNY/Courses/Business Analytics and Data Mining/Assignments/data621-ctg


fun <- function(x, y) cor(y, x, use = "na.or.complete")
Correlation <- sapply(winedata[, 16:ncol(winedata)], FUN = fun, y=winedata$TARGET)
Correlation <- sort(Correlation, decreasing = TRUE)

#colnames(Correlation) <- c("Correlation")
#Correlation <- cbind(variable = rownames(Correlation), Correlation)
#rownames(Correlation) <- NULL

kable(data.frame(Correlation), caption = "Correlation between TARGET and predictor variables")

#kable(Correlation, caption = "New variables Correlation ")
#Correlation


modelvars <- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW5/ModelVars.csv")
kable(modelvars, caption = 'Model Variables')

winedata_orig <- winedata[,c(1:15)]
winedata_trans <- winedata[,c(1, 16:40)]


#winedata <- read.csv("C:/CUNY/Courses/IS621/Assignment602/Assignment05/wine-training-data.csv")
#dim(winedata)
#winedata <- select(winedata, (-(INDEX)))
#str(winedata)
## all zeros
# winedata[is.na(winedata)] <- 0

poismod1 <- glm(TARGET ~ ., data=winedata_orig, family=poisson)

c1<- summary(poismod1)
kable(c1$coefficients, caption = 'Model 1 Poisson Original Data')
pr <- residuals(poismod1,"pearson")
phi <- sum(pr^2)/df.residual(poismod1)
phi
fm_qpois <- glm(TARGET ~ ., data = winedata_orig, family = quasipoisson)
c2<- summary(fm_qpois)
kable(c2$coefficients, caption = 'Model 2 Quasi-Poisson Original Data')
qpr <- residuals(fm_qpois,"pearson")
qphi <- sum(qpr^2)/df.residual(fm_qpois)
qphi
library(sandwich)
library(msm)
library(pscl)
mod1zip <- zeroinfl(TARGET~ ., data = winedata_orig, dist = "poisson")
```

```
c3<- summary(mod1zip)


kable(c3$coefficients, caption = 'Model 3, Zero Inflation Poisson Original Data')

zippr <- residuals(mod1zip,"pearson")
zipphi <- sum(zippr^2)/df.residual(mod1zip)
zipphi
vuong(mod1zip,poismod1)
# transformed data. Poisson Model 2

poismod2 <- glm(TARGET ~ ., data=winedata_trans, family=poisson)
c4<- summary(poismod2)

kable(c4$coefficients, caption = 'Model 4, Poisson Transformed Data')

pr2 <- residuals(poismod2,"pearson")
phi2 <- sum(pr2^2)/df.residual(poismod2)
phi2

mod2qpois <- glm(TARGET ~ ., data = winedata_trans, family = quasipoisson)
c5<- summary(mod2qpois)


kable(c5$coefficients, caption = 'Model 5 Quasi-Poisson Transformed Data')
library(sandwich)
library(msm)
library(pscl)
#mod2zip <- zeroinfl(TARGET~ ., data = winedata_trans, dist = "poisson")
#summary(mod2zip)

#####

quine3 <- as.data.frame(model.matrix(poismod2)) ## all regressors
quine3 <- quine3[, !is.na(coef(poismod2))]      ## only identified
quine3 <- quine3[, -1]                       ## omit intercept
quine3$TARGET <- winedata_trans$TARGET                ## add response

## re-fit glm.nb()
fm1a <- glm(TARGET ~ ., data = quine3, family="poisson")
## equivalent to previous fit
logLik(fm1a) - logLik(poismod2)
coef(fm1a) - na.omit(coef(poismod2))

## fit zeroinfl(), now works
mod2zip<- zeroinfl(TARGET ~ . | 1, data = quine3, dist = "poisson")

c6<- summary(mod2zip)



kable(c6$coefficients, caption = 'Model 6, Zero Inflation Poisson Transformed Data')
zippr2 <- residuals(mod2zip,"pearson")
```

```r
zipphi2 <- sum(zippr2^2)/df.residual(mod2zip)
zipphi2


vuong(mod2zip,poismod2)
nbmod3 = glm.nb(TARGET ~ ., data = winedata_orig)
kable(rbind(data.frame("Poisson Coeff"= poismod1$coefficients,"Negative Binom Coeffi" = nbmod3$coefficie



c7<- summary(nbmod3)
kable(c7$coefficients, caption = 'Model 7, Negative binomial Original Data')

#library(vcd)
#distplot(winedata_orig$TARGET, type = "nbinomial")
#distplot(winedata_orig$TARGET, type = "poisson")



nbpr3 <- residuals(nbmod3,"pearson")
nbphi3 <- sum(nbpr3^2)/df.residual(nbmod3)
nbphi3

library(sandwich)
library(msm)
library(pscl)
nbmod3zip <- zeroinfl(TARGET~ ., data = winedata_orig, dist = "negbin")
c8<- summary(nbmod3zip)

kable(c8$coefficients, caption = 'Model 8, Zero Inflation Negative binomial Original Data')

nbzpr3 <- residuals(nbmod3zip,"pearson")
nbzphi3 <- sum(nbzpr3^2)/df.residual(nbmod3zip)
nbzphi3

vuong(nbmod3zip,nbmod3)

#transformed data. Negative Binomial model 4

nbmod4 = glm.nb(TARGET ~ ., data = winedata_trans)
c9<- summary(nbmod4)

kable(c9$coefficients, caption = 'Model 9, Negative binomial Transformed Data')


kable(rbind(data.frame("Poisson Coeff"= poismod2$coefficients,"Negative Binom Coeffi" = nbmod4$coefficie


nbpr4 <- residuals(nbmod4,"pearson")
nbphi4 <- sum(nbpr4^2)/df.residual(nbmod4)
nbphi4
```

```
library(sandwich)
library(msm)
library(pscl)

quine4 <- as.data.frame(model.matrix(nbmod4)) ## all regressors
quine4 <- quine4[, !is.na(coef(nbmod4))]      ## only identified
quine4 <- quine4[, -1]                      ## omit intercept
quine4$TARGET <- winedata_trans$TARGET                ## add response

## re-fit glm.nb()
fm1a <- glm.nb(TARGET ~ ., data = quine4)
## equivalent to previous fit
logLik(fm1a) - logLik(nbmod4)
coef(fm1a) - na.omit(coef(nbmod4))

## fit zeroinfl(), now works
nbmod4zip<- zeroinfl(TARGET ~ . | 1, data = quine4, dist = "negbin")

c10<- summary(nbmod4zip)



##########################



kable(c10$coefficients, caption = 'Model 10, Zero Inflation Negative binomial Transformed Data')
nbzpr4 <- residuals(nbmod4zip,"pearson")
nbzphi4 <- sum(nbzpr4^2)/df.residual(nbmod4zip)
nbzphi4
vuong(nbmod4zip,nbmod4)

#step1 <- step(fm_qpois,direction="backward",test="F")
#summary(step1)
library(AICcmodavg)
#AICc(list(fm_qpois))
AICc(fm_qpois, return.K = FALSE, second.ord = TRUE,nobs = NULL, c.hat = 1)


lmod5 = lm(TARGET ~ ., data = winedata_orig)
c11<- summary(lmod5)

kable(c11$coefficients, caption = 'Model 11, Linear Model Original Data')


x <- summary(lmod5)

rse <- round(x$sigma, 4)
r2 <- round(x$r.squared, 4)
ar2 <- round(x$adj.r.squared,4)
fstat <- paste0(round(x$fstatistic[1],0), " on ", x$fstatistic[2], " and ", x$fstatistic[3]," DF")
```

```
y<- as.data.frame(coef(x))

PositiveImpact <- ''
NegativeImpact <- ''
Lin_eq <- as.character(round(y[1,1],4))
Sig_Impact <- ''

for(i in 2:nrow(y)) {

    if(y[i,1] >= 0)

        PositiveImpact <- paste0(rownames(y)[i], ", ", PositiveImpact)
    else
        NegativeImpact <- paste0(rownames(y)[i], ", ", NegativeImpact)


    if(y[i,4] <= 0.05 )
        Sig_Impact <- paste0(rownames(y)[i], ", ", Sig_Impact)

    Lin_eq <- paste0(Lin_eq, ifelse(y[i,1]>=0, " + ", " - "), as.character(abs(round(y[i,1],6))), " * "
}

PositiveImpact <- substr(PositiveImpact, 1, str_length(PositiveImpact)-2)
NegativeImpact <- substr(NegativeImpact, 1, str_length(NegativeImpact)-2)
Sig_Impact <- substr(Sig_Impact, 1, str_length(Sig_Impact)-2)

lmod6 = lm(TARGET ~ ., data = winedata_trans)
c12<- summary(lmod6)


kable(c12$coefficients, caption = 'Model 11, Linear Model Transformed Data')

x <- summary(lmod6)


rse <- round(x$sigma, 4)
r2 <- round(x$r.squared, 4)
ar2 <- round(x$adj.r.squared,4)
fstat <- paste0(round(x$fstatistic[1],0), " on ", x$fstatistic[2], " and ", x$fstatistic[3]," DF")


y<- as.data.frame(coef(x))

PositiveImpact <- ''
NegativeImpact <- ''
Lin_eq <- as.character(round(y[1,1],4))
Sig_Impact <- ''

for(i in 2:nrow(y)) {

    if(y[i,1] >= 0)

        PositiveImpact <- paste0(rownames(y)[i], ", ", PositiveImpact)
```

```
    else
        NegativeImpact <- paste0(rownames(y)[i], ", ", NegativeImpact)


    if(y[i,4] <= 0.05 )
        Sig_Impact <- paste0(rownames(y)[i], ", ", Sig_Impact)

    Lin_eq <- paste0(Lin_eq, ifelse(y[i,1]>=0, " + ", " - "), as.character(abs(round(y[i,1],6))), " * "
}

PositiveImpact <- substr(PositiveImpact, 1, str_length(PositiveImpact)-2)
NegativeImpact <- substr(NegativeImpact, 1, str_length(NegativeImpact)-2)
Sig_Impact <- substr(Sig_Impact, 1, str_length(Sig_Impact)-2)


# PositiveImpact
# NegativeImpact
# Lin_eq
# Sig_Impact

step5 <- step(lmod5,direction="backward",test="F")
AIC(step5)
step6 <- step(lmod6,direction="backward",test="F")
AIC(step5)
AIC(step6)
lmpr5 <- residuals(lmod5,"pearson")
lmphi5 <- sum(lmpr5^2)/df.residual(lmod5)
lmphi5
lmpr6 <- residuals(lmod6,"pearson")
lmphi6 <- sum(lmpr6^2)/df.residual(lmod6)
lmphi6
modselect<- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW5/modelselection2.
kable(modselect, caption = "Model Selection Strategy")


modmetrics<- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW5/modelmetrics2.cs
kable(modmetrics, caption = "Model Selection KPI")

winedata_eval<- read.csv("https://raw.githubusercontent.com/kishkp/data621-ctg5/master/HW5/wine-evaluati


winedata_eval$ResidualSugar_MISS <- ifelse(is.na(winedata_eval$ResidualSugar), 1, 0)
winedata_eval$Chlorides_MISS <- ifelse(is.na(winedata_eval$Chlorides), 1, 0)
winedata_eval$FreeSulfurDioxide_MISS <- ifelse(is.na(winedata_eval$FreeSulfurDioxide), 1, 0)
winedata_eval$TotalSulfurDioxide_MISS <- ifelse(is.na(winedata_eval$TotalSulfurDioxide), 1, 0)
winedata_eval$pH_MISS <- ifelse(is.na(winedata_eval$pH), 1, 0)
winedata_eval$Sulphates_MISS <- ifelse(is.na(winedata_eval$Sulphates), 1, 0)
winedata_eval$Alcohol_MISS <- ifelse(is.na(winedata_eval$Alcohol), 1, 0)
winedata_eval$STARS_MISS <- ifelse(is.na(winedata_eval$STARS), 1, 0)

winedata_eval$ResidualSugar[is.na(winedata_eval$ResidualSugar)] <- mean(winedata_eval$ResidualSugar, na
winedata_eval$Chlorides[is.na(winedata_eval$Chlorides)] <- mean(winedata_eval$Chlorides, na.rm = T)
winedata_eval$FreeSulfurDioxide[is.na(winedata_eval$FreeSulfurDioxide)] <- mean(winedata_eval$FreeSulfu
```

```r
winedata_eval$TotalSulfurDioxide[is.na(winedata_eval$TotalSulfurDioxide)] <- mean(winedata_eval$TotalSul
winedata_eval$pH[is.na(winedata_eval$pH)] <- mean(winedata_eval$pH, na.rm = T)
winedata_eval$Sulphates[is.na(winedata_eval$Sulphates)] <- mean(winedata_eval$Sulphates, na.rm = T)
winedata_eval$Alcohol[is.na(winedata_eval$Alcohol)] <- mean(winedata_eval$Alcohol, na.rm = T)
winedata_eval$STARS[is.na(winedata_eval$STARS)] <- 0

treat_outliers <- function(x) {
qnt <- quantile(x, probs=c(.25, .75), na.rm = T)
caps <- quantile(x, probs=c(.05, .95), na.rm = T)
H <- 1.5 * IQR(x, na.rm = T)
x[x < (qnt[1] - H)] <- caps[1]
x[x > (qnt[2] + H)] <- caps[2]

return(x)
}

winedata_eval$FixedAcidity_CAP <- treat_outliers(winedata_eval$FixedAcidity)
winedata_eval$VolatileAcidity_CAP <- treat_outliers(winedata_eval$VolatileAcidity)
winedata_eval$CitricAcid_CAP <- treat_outliers(winedata_eval$CitricAcid)
winedata_eval$ResidualSugar_CAP <- treat_outliers(winedata_eval$ResidualSugar)
winedata_eval$Chlorides_CAP <- treat_outliers(winedata_eval$Chlorides)
winedata_eval$FreeSulfurDioxide_CAP <- treat_outliers(winedata_eval$FreeSulfurDioxide)
winedata_eval$TotalSulfurDioxide_CAP <- treat_outliers(winedata_eval$TotalSulfurDioxide)
winedata_eval$Density_CAP <- treat_outliers(winedata_eval$Density)
winedata_eval$pH_CAP <- treat_outliers(winedata_eval$pH)
winedata_eval$Sulphates_CAP <- treat_outliers(winedata_eval$Sulphates)
winedata_eval$Alcohol_CAP <- treat_outliers(winedata_eval$Alcohol)
winedata_eval$AcidIndex_CAP <- treat_outliers(winedata_eval$AcidIndex)

winedata_eval$LabelAppeal_Positive <- ifelse(winedata_eval$FreeSulfurDioxide >=0, 1, 0)

winedata_eval$STARS_1 <- ifelse(winedata_eval$STARS == 1, 1, 0)
winedata_eval$STARS_2 <- ifelse(winedata_eval$STARS == 2, 1, 0)
winedata_eval$STARS_3 <- ifelse(winedata_eval$STARS == 3, 1, 0)
winedata_eval$STARS_4 <- ifelse(winedata_eval$STARS == 4, 1, 0)


winedata_eval$TARGET <- 0

winedata_eval$TARGET <- predict(poismod2, type="response", newdata=winedata_eval)

winedata_eval$TARGET<- round(winedata_eval$TARGET)
x<-arrange(winedata_eval, (TARGET))
#x<-x[-c(1:2),]
t<- x[1:6,]
t2<- t(t)
kable(t2,  caption="Transposed Model Output / Results")
```