Home Work Assignment - 01

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

Contents

Overview **2**

Objective **2**

1 Data Exploration Analysis **2**

1.1 Variable identification 2

1.2 Variable Relationships 3

1.3 Data summary analysis 4

1.4 Outliers Identification 7

3. Build Models **19**

3.1 Poisson models 19

3.2 Negative Binomial models 21

3.3 Linear Regression models 22

3.3.1 Linear Regression Model 5 22

3.3.1 Linear Regression Model 6 23

3.3.2 Linear Regression Model 6 23

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.

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

1 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

1.1 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.

1.2 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.

1.3 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

	mean	sd	median	trimmed
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
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:

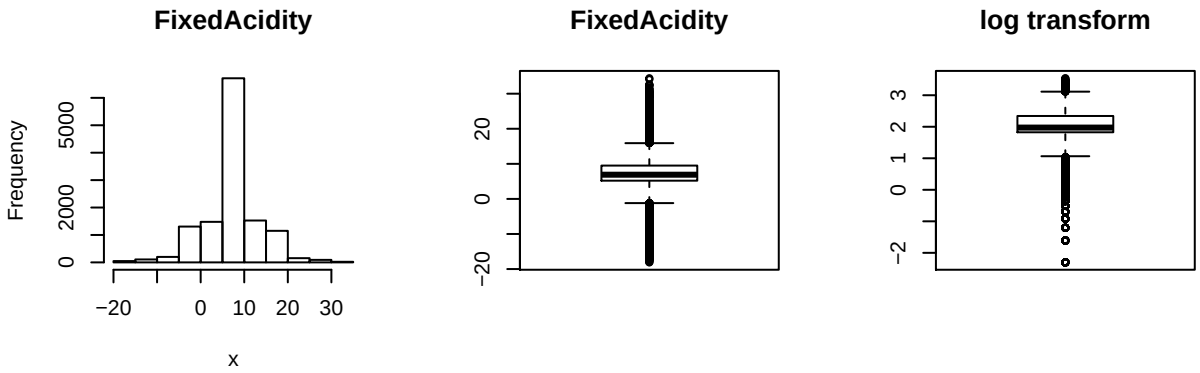
- Variable ResidualSugar has 616 and 0.0164913 correlation. Given the low correlation we will try try some imputation techniques to handle the missing the values and replace missing values with their respective value.
- variable Chlorides 638 -0.0382631 correlation. . Given the low negative correlation 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

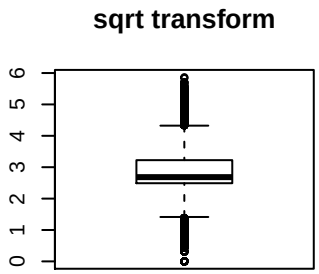
1.4 Outliers Identification

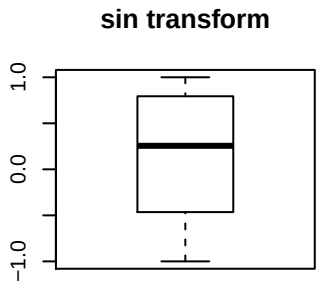
In this section we look at boxplots to determine the outliers in variables and decide on whether to act on the outliers. Lets do some univariate analysis. We will look at the Histogram and Boxplot for each variable to detect outliers if any and treat it accordingly.

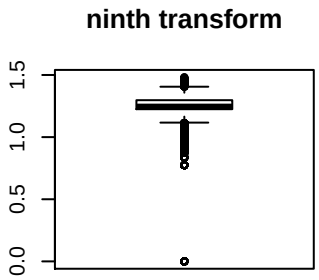

FixedAcidity
Frequency
0
2000
5000
−20
0
10
20
30
x

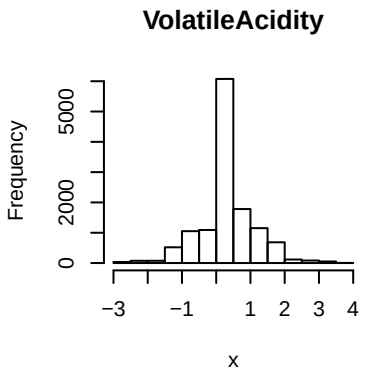
FixedAcidity
−20
0
20

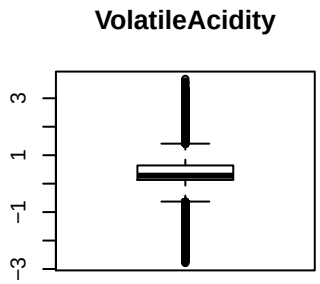
log transform
−2
0
1
2
3

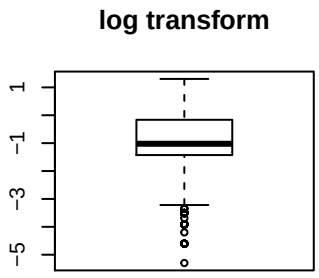

sqrt transform
0
1
2
3
4
5
6

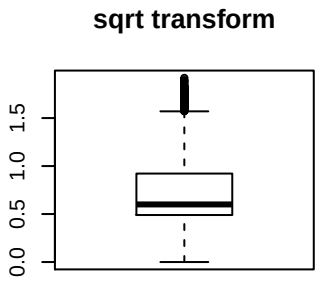

sin transform
−1.0
0.0
1.0

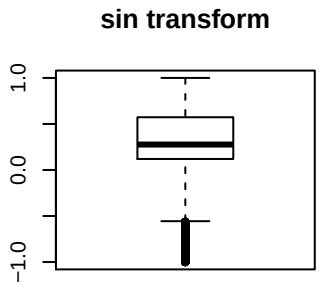

ninth transform
0.0
0.5
1.0
1.5

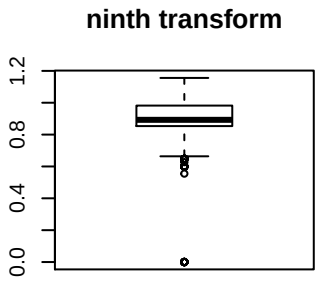

VolatileAcidity
Frequency
0
2000
5000
−3
−1
1
2
3
4
x

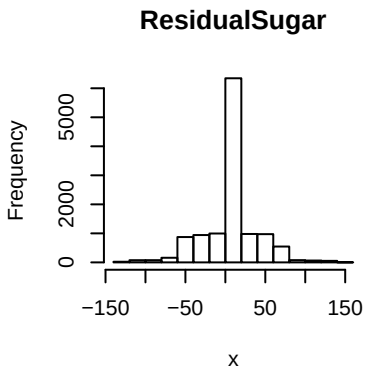

VolatileAcidity
−3
−1
1
3

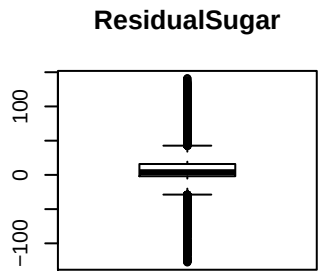

log transform
−5
−3
−1
1

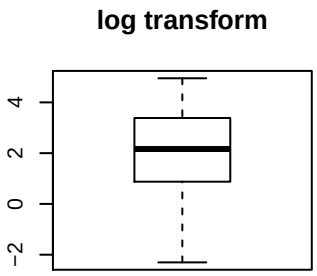

sqrt transform
0.0
0.5
1.0
1.5

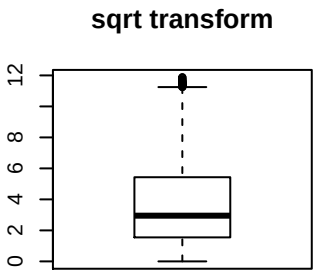

sin transform
−1.0
0.0
1.0

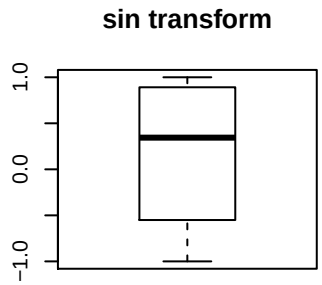

ninth transform
0.0
0.4
0.8
1.2

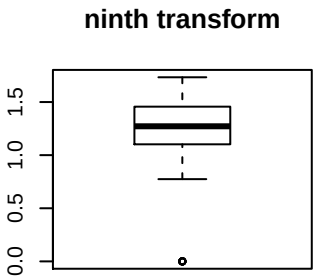

ResidualSugar
Frequency
0
2000
5000
−150
−50
50
150
x

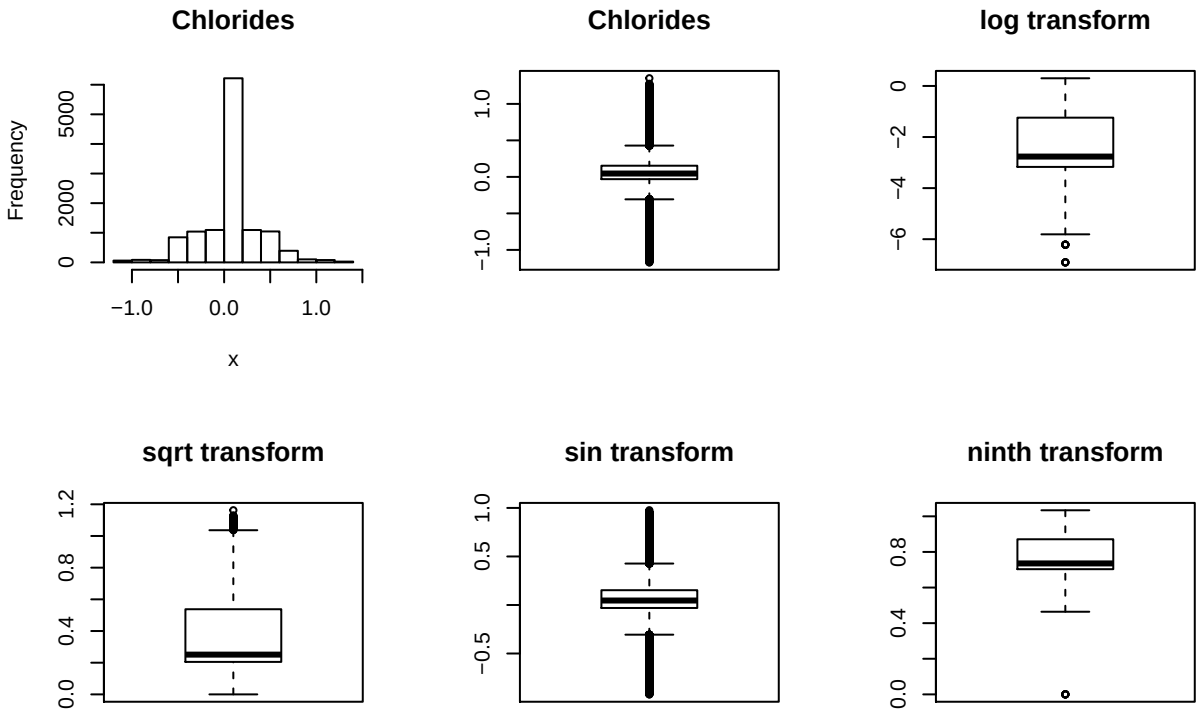

ResidualSugar
−100
0
100

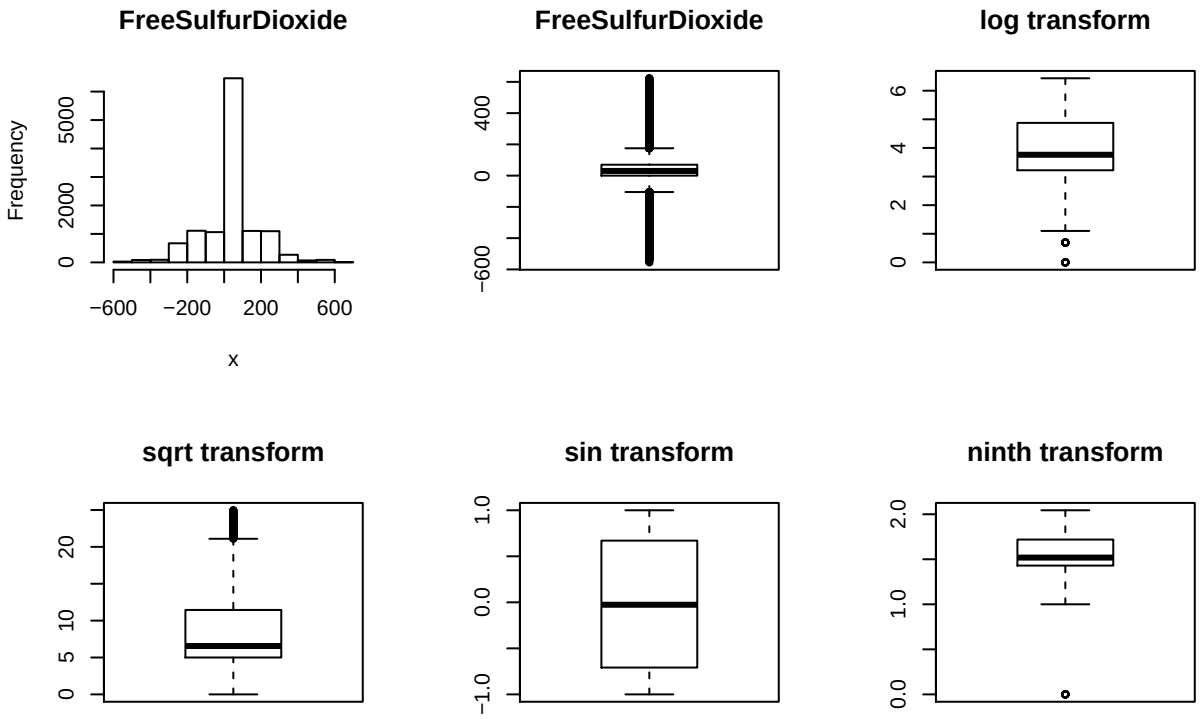

log transform
−2
0
2
4

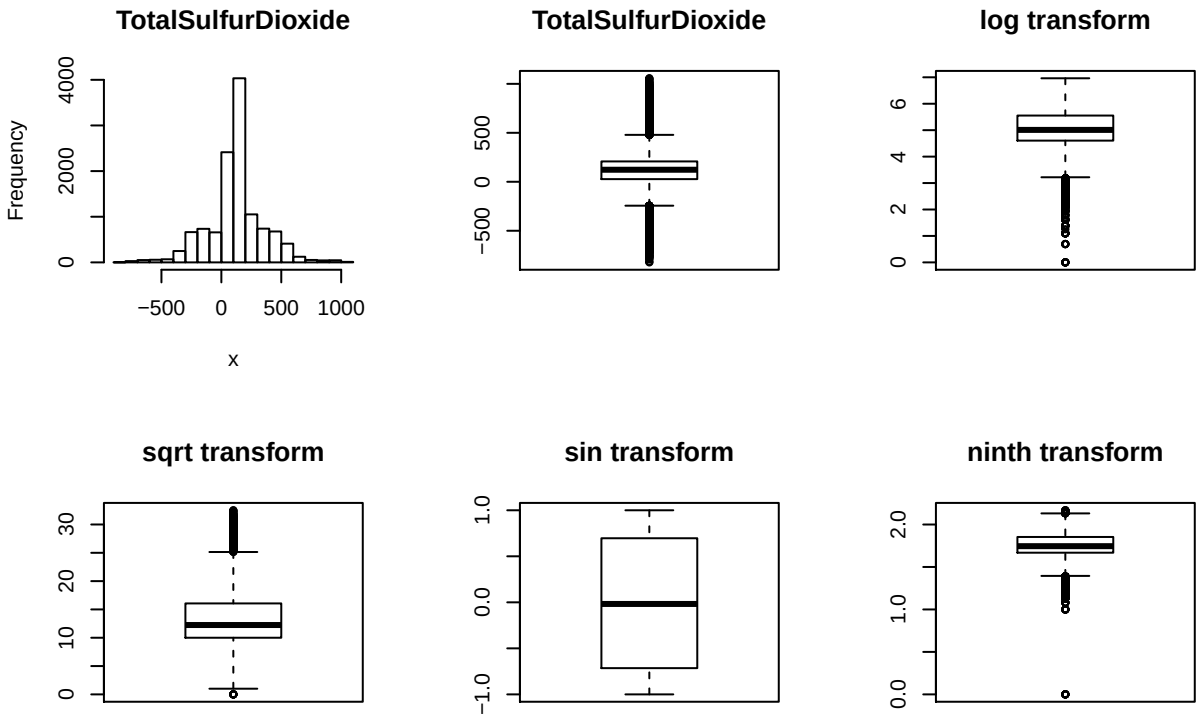

sqrt transform
0
2
4
6
8
12

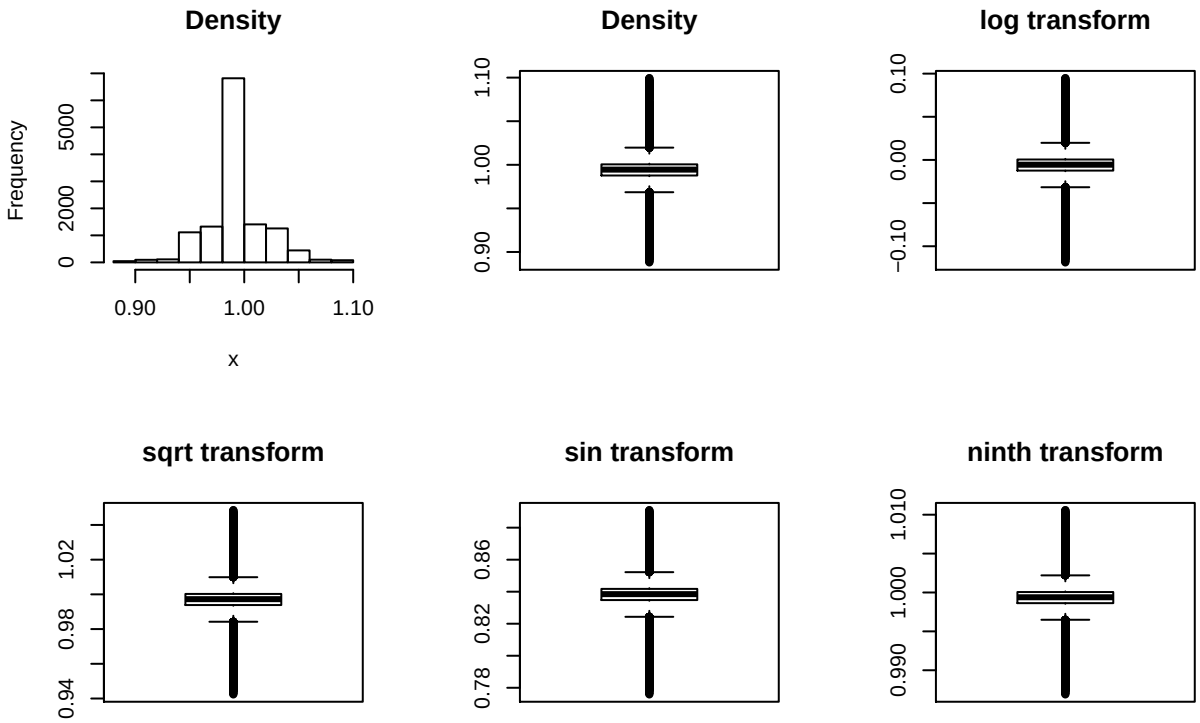

sin transform
−1.0
0.0
1.0

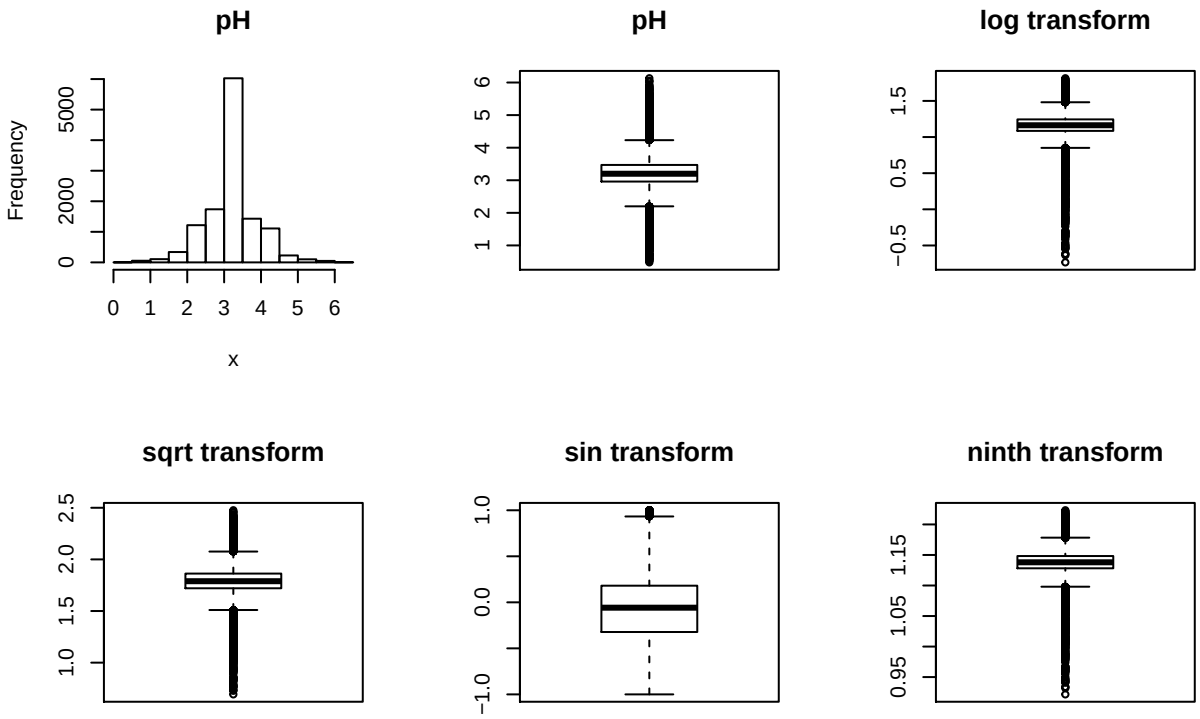

ninth transform
0.0
0.5
1.0
1.5

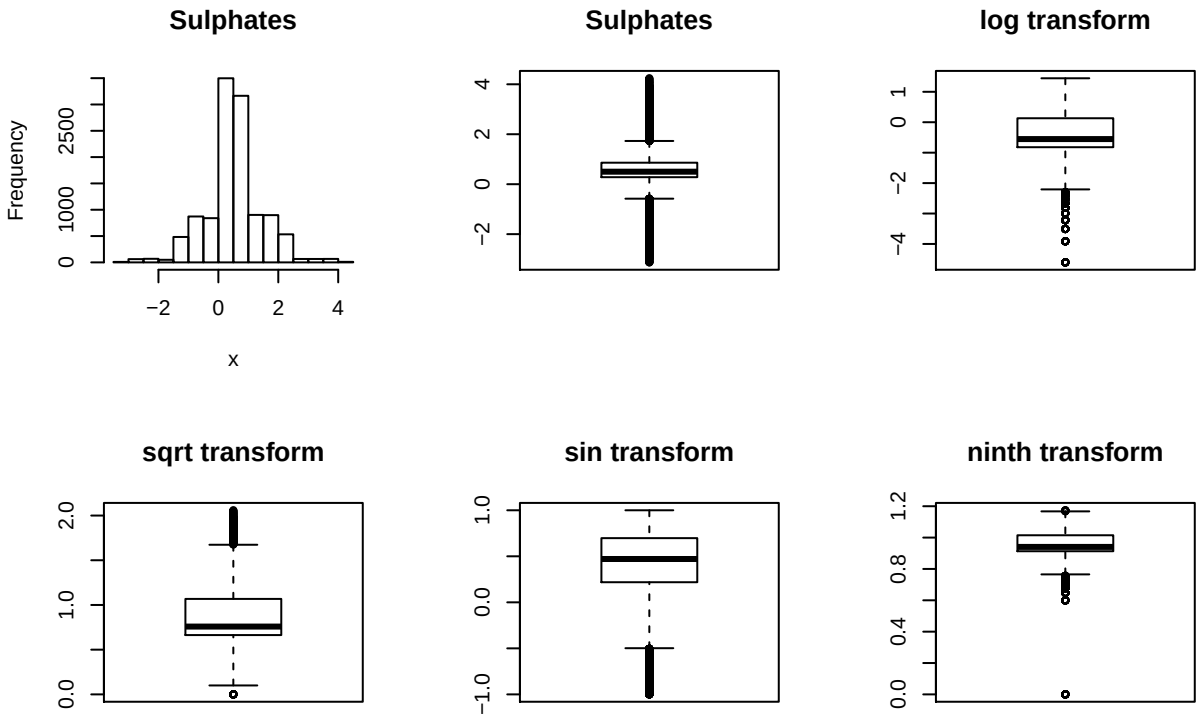

Chlorides
Frequency
0
2000
5000
−1.0
0.0
1.0
x
Chlorides
−1.0
0.0
1.0
log transform
−6
−4
−2
0
sqrt transform
0.0
0.4
0.8
1.2
sin transform
−0.5
0.5
1.0
ninth transform
0.0
0.4
0.8

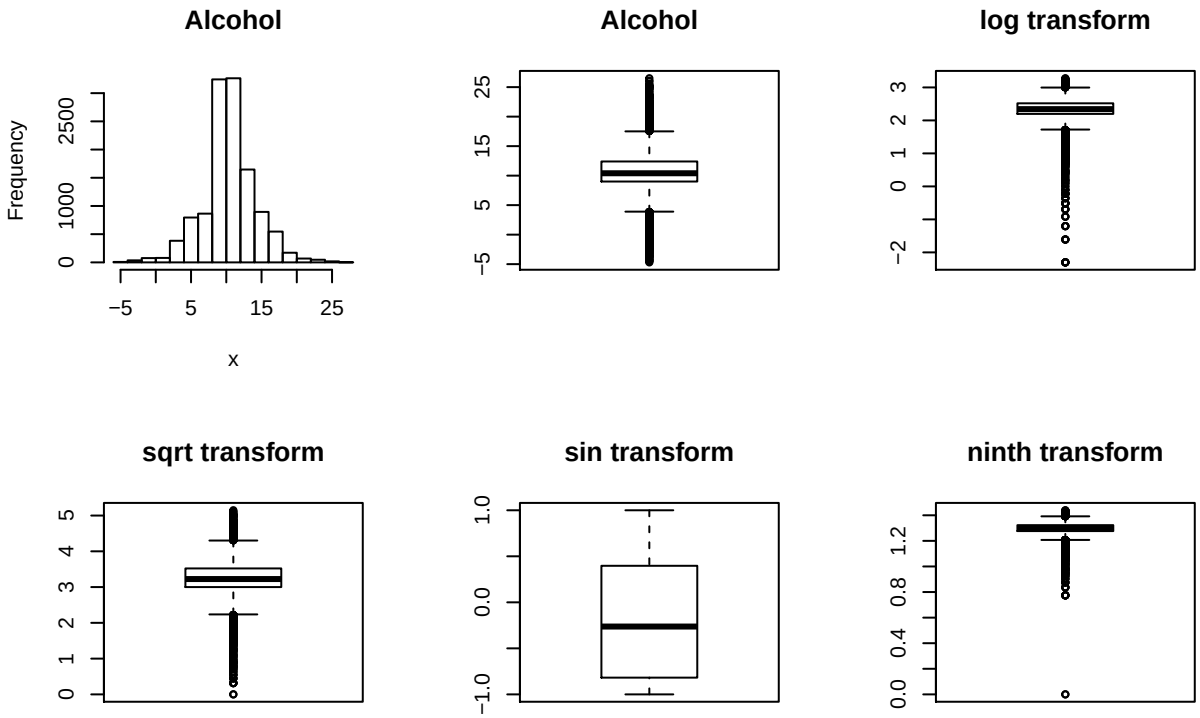

FreeSulfurDioxide
Frequency
0
2000
5000
−600
−200
200
600
x
FreeSulfurDioxide
−600
0
400
log transform
0
2
4
6
sqrt transform
0
5
10
20
sin transform
−1.0
0.0
1.0
ninth transform
0.0
1.0
2.0

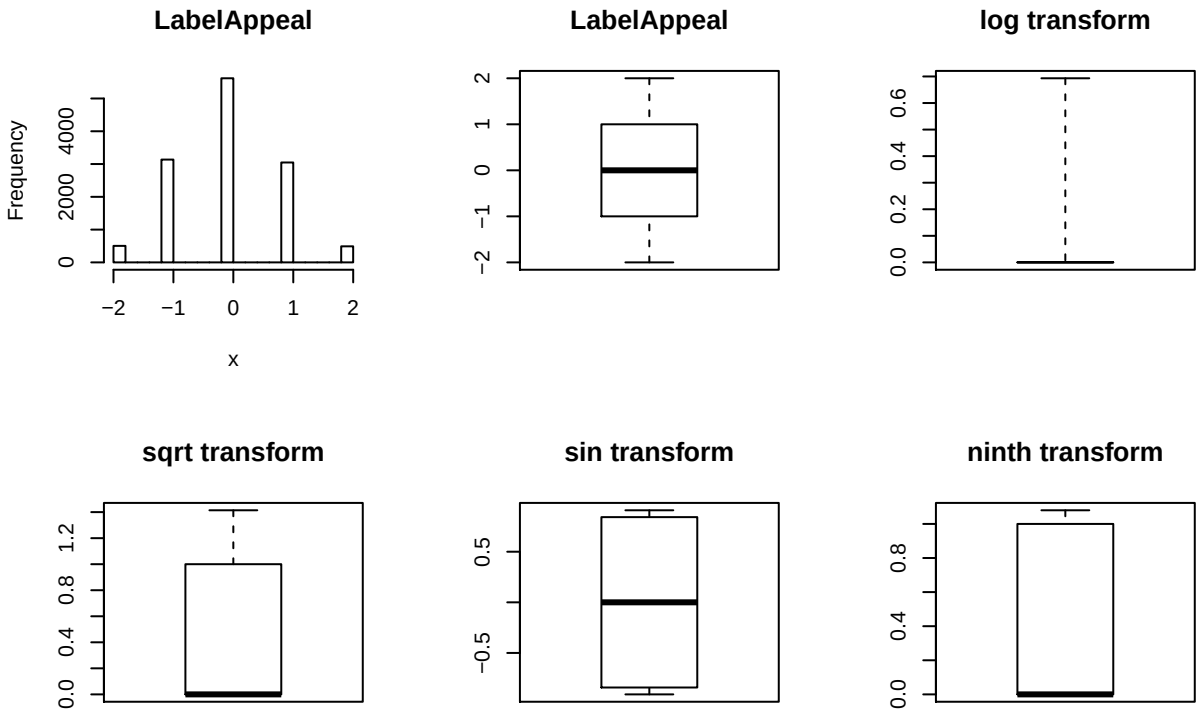

TotalSulfurDioxide
Frequency
0
2000
4000
−500
0
500
1000
x
TotalSulfurDioxide
−500
0
500
log transform
0
2
4
6
sqrt transform
0
10
20
30
sin transform
−1.0
0.0
1.0
ninth transform
0.0
1.0
2.0

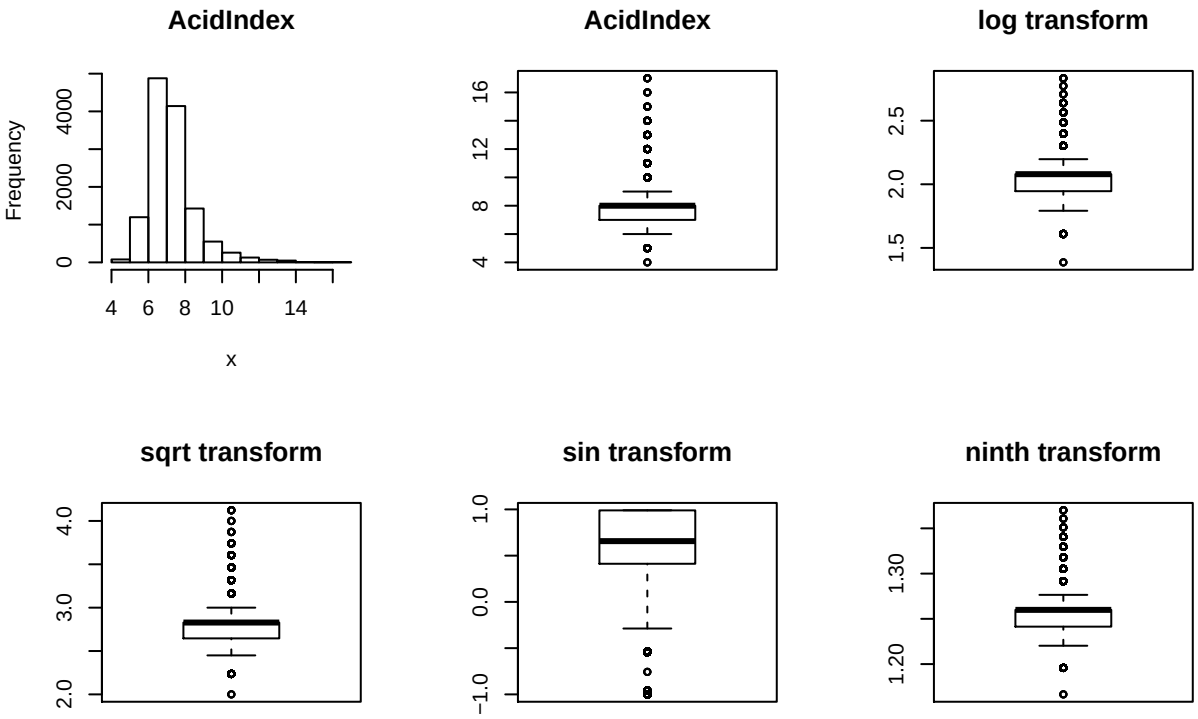

Density
Frequency
0
2000
5000
0.90
1.00
1.10
x
Density
0.90
1.00
1.10
log transform
−0.10
0.00
0.10
sqrt transform
0.94
0.98
1.02
sin transform
0.78
0.82
0.86
ninth transform
0.990
1.000
1.010

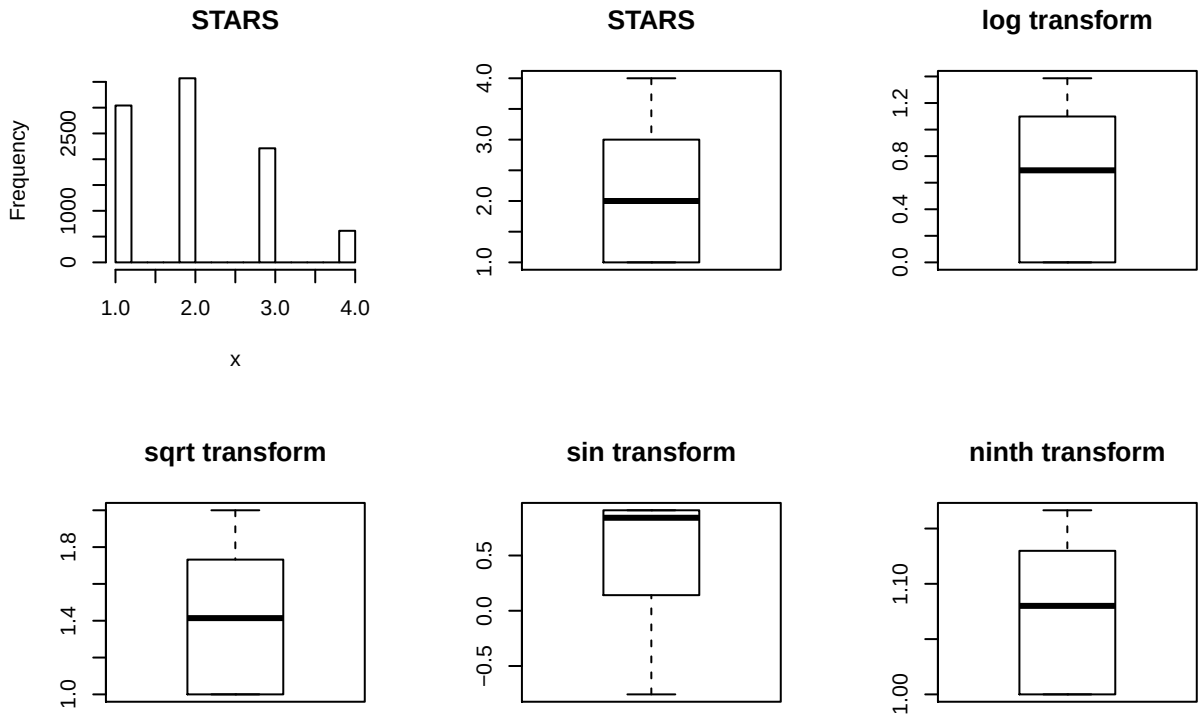

pH
Frequency
0
2000
5000
0 1 2 3 4 5 6
x
pH
1 2 3 4 5 6
log transform
−0.5 0.5 1.5
sqrt transform
1.0 1.5 2.0 2.5
sin transform
−1.0 0.0 1.0
ninth transform
0.95 1.05 1.15


Sulphates
Frequency
0
1000
2500
−2
0
2
4
x
Sulphates
−2
0
2
4
log transform
−4
−2
0
1
sqrt transform
0.0
1.0
2.0
sin transform
−1.0
0.0
1.0
ninth transform
0.0
0.4
0.8
1.2


Alcohol
Frequency
0
1000
2500
−5
5
15
25
x
Alcohol
−5
5
15
25
log transform
−2
0
1
2
3
sqrt transform
0
1
2
3
4
5
sin transform
−1.0
0.0
1.0
ninth transform
0.0
0.4
0.8
1.2


LabelAppeal
Frequency
0
2000
4000
−2
−1
0
1
2
x
LabelAppeal
−2
−1
0
1
2
log transform
0.0
0.2
0.4
0.6
sqrt transform
0.0
0.4
0.8
1.2
sin transform
−0.5
0.5
ninth transform
0.0
0.4
0.8


AcidIndex
Frequency
0
2000
4000
4
6
8
10
14
x
AcidIndex
4
8
12
16
log transform
1.5
2.0
2.5
sqrt transform
2.0
3.0
4.0
sin transform
−1.0
0.0
1.0
ninth transform
1.20
1.30



3. 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 ten models as follow:
- Two Poisson models
- Two Quasi-Poisson models
- Two Negative binomial models
- Two Zero-inflated models
- Two Linear regression models

3.1 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

3.1.1 Poisson Model 1

We will explore the Poisson regression model Using original data with replacing all missing data with 0's.

```r
#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, family=poisson)
summary(poismod1)
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = winedata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9909  -0.7242   0.0681   0.5796   3.2365
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.494e+00  1.945e-01   7.684 1.54e-14 ***
## FixedAcidity        -3.148e-04  8.202e-04  -0.384 0.701119
## VolatileAcidity     -3.359e-02  6.518e-03  -5.154 2.55e-07 ***
## CitricAcid           7.719e-03  5.892e-03   1.310 0.190181
## ResidualSugar        4.677e-05  1.545e-04   0.303 0.762128
## Chlorides           -4.125e-02  1.644e-02  -2.509 0.012105 *
## FreeSulfurDioxide    1.248e-04  3.507e-05   3.559 0.000372 ***
## TotalSulfurDioxide   7.969e-05  2.255e-05   3.533 0.000410 ***
## Density             -2.833e-01  1.919e-01  -1.476 0.139959
## pH                  -3.647e-03  5.842e-03  -0.624 0.532474
## Sulphates           -1.151e-02  5.672e-03  -2.028 0.042521 *
## Alcohol              1.320e-03  1.186e-03   1.113 0.265804
## LabelAppeal          1.330e-01  6.063e-03  21.932  < 2e-16 ***
## AcidIndex           -8.669e-02  4.542e-03 -19.088  < 2e-16 ***
## STARS                3.117e-01  4.527e-03  68.848  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 14735  on 12780  degrees of freedom
## AIC: 46707
##
## Number of Fisher Scoring iterations: 5
```

*** Interpretation Poisson Model 1***

3.1.2 Poisson Model 2

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

```
# transformed data. Poisson Model 2
```

*** Interpretation Poisson Model 2***

3.2 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.

3.2.1 Negative Binomial model 3

We will explore the Negative Binomial model Using original data with replacing all missing data with 0's.

```
nbmod3 = glm.nb(TARGET ~ ., data = winedata)
```

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

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

```
summary(nbmod3)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = winedata, init.theta = 48963.3547,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9908  -0.7242   0.0681   0.5796   3.2364
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.494e+00  1.945e-01   7.684 1.54e-14 ***
## FixedAcidity        -3.148e-04  8.203e-04  -0.384 0.701121
## VolatileAcidity     -3.359e-02  6.518e-03  -5.154 2.55e-07 ***
## CitricAcid           7.719e-03  5.892e-03   1.310 0.190193
## ResidualSugar        4.677e-05  1.545e-04   0.303 0.762112
## Chlorides           -4.125e-02  1.644e-02  -2.509 0.012106 *
## FreeSulfurDioxide    1.248e-04  3.507e-05   3.559 0.000373 ***
## TotalSulfurDioxide   7.969e-05  2.255e-05   3.533 0.000410 ***
## Density             -2.833e-01  1.919e-01  -1.476 0.139967
## pH                  -3.647e-03  5.842e-03  -0.624 0.532466
## Sulphates           -1.151e-02  5.673e-03  -2.028 0.042521 *
## Alcohol              1.319e-03  1.186e-03   1.113 0.265840
## LabelAppeal          1.330e-01  6.063e-03  21.931  < 2e-16 ***
## AcidIndex           -8.669e-02  4.542e-03 -19.088  < 2e-16 ***
## STARS                3.117e-01  4.527e-03  68.847  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48963.35) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 14734  on 12780  degrees of freedom
## AIC: 46709
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48963
##           Std. Err.:  50701
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -46677.17
```

*** Interpretation Negative Binomial Model 3***

3.2.1 Negative Binomial model 4

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

```
#transformed data. Negative Binomial model 4
```

*** Interpretation Negative Binomial Model 4***

3.3 Linear Regression models

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

3.3.1 Linear Regression Model 5

We will explore the Negative Binomial model Using original data with replacing all missing data with 0's.

```
lmod5 = lm(TARGET ~ ., data = winedata)
summary(lmod5)
```

```
##
## Call:
## lm(formula = TARGET ~ ., data = winedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5749 -0.9502  0.0661  0.9091  6.0017
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)         3.947e+00  4.465e-01   8.840  < 2e-16 ***
## FixedAcidity       -2.265e-05  1.885e-03  -0.012  0.99042
## VolatileAcidity    -9.956e-02  1.499e-02  -6.643 3.19e-11 ***
## CitricAcid          2.119e-02  1.363e-02   1.555  0.12001
## ResidualSugar       1.666e-04  3.558e-04   0.468  0.63964
## Chlorides          -1.245e-01  3.775e-02  -3.297  0.00098 ***
## FreeSulfurDioxide   3.096e-04  8.087e-05   3.829  0.00013 ***
## TotalSulfurDioxide  2.168e-04  5.165e-05   4.197 2.72e-05 ***
## Density            -8.033e-01  4.421e-01  -1.817  0.06923 .
## pH                 -7.181e-03  1.350e-02  -0.532  0.59465
## Sulphates          -2.965e-02  1.303e-02  -2.276  0.02287 *
## Alcohol             6.271e-03  2.728e-03   2.298  0.02156 *
## LabelAppeal         4.319e-01  1.367e-02  31.592  < 2e-16 ***
## AcidIndex          -2.079e-01  9.208e-03 -22.581  < 2e-16 ***
## STARS               9.780e-01  1.045e-02  93.572  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.325 on 12780 degrees of freedom
## Multiple R-squared:  0.5276, Adjusted R-squared:  0.5271
## F-statistic:  1019 on 14 and 12780 DF,  p-value: < 2.2e-16
```

3.3.1 Linear Regression Model 6

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

```
#transformed data. Linear Regression
```

3.3.2 Linear Regression Model 6