Assignment/HW5

Group 2

April 19, 2018

Overview In this homework assignment we explore, analyze, and model a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant.

A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then it will be able to adjust their wine offering to maximize sales.

Our objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine.

# 1. Data Exploration

The data were loaded into a Github repository for better reproducability. After taking a quick look at the head of the data, there are a number of NA’s we will have to deal with. We will also remove the INDEX column as it will not be useful in our analysis.

## Data Exploration and Statistic Measures

The purpose of the data exploration and statistic measures phase is to understand the data to determine how to process the dataset for modelling.

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Variable Type |
| Target | Number of cases purchased | Response |
| AcidIndex | Proprietary method of testing total acidity of wine by using a weighted average | Predictor |
| Alcohol | Alcohol content | Predictor |
| Chlorides | Chloride content of wine | Predictor |
| CitricAcid | Citric acid content | Predictor |
| Density | Density of wine | Predictor |
| FixedAcidity | Fixed acidity of wine | Predictor |
| FreeSulfurDioxide | Sulfur dioxide content of wine | Predictor |
| LabelAppeal | Marketing score indidcating appeal of label design for consumers. Positive is like, negative is dislike | Predictor |
| ResidualSugar | Resdiual sugar of wine | Predictor |
| Stars | Wine rating by a team of experts. 4 stars = Excellent, 1 star = Poor | Predictor |
| Sulphates | Sulfate content of wine | Predictor |
| TotalSulfurDioxide | Total sulfur dioxide of wine | Predictor |
| VolatileAcidity | Volatile acid content of wine | Predictor |
| pH | pH of wine | Predictor |

The training set has 12,795 observations on 1 response and 14 predictor variables, representing certain charachteristics about a bottle of wine.

## Observations: 12,795  
## Variables: 15  
## $ TARGET <int> 3, 3, 5, 3, 4, 0, 0, 4, 3, 6, 0, 4, 3, 7, 4...  
## $ FixedAcidity <dbl> 3.2, 4.5, 7.1, 5.7, 8.0, 11.3, 7.7, 6.5, 14...  
## $ VolatileAcidity <dbl> 1.160, 0.160, 2.640, 0.385, 0.330, 0.320, 0...  
## $ CitricAcid <dbl> -0.98, -0.81, -0.88, 0.04, -1.26, 0.59, -0....  
## $ ResidualSugar <dbl> 54.20, 26.10, 14.80, 18.80, 9.40, 2.20, 21....  
## $ Chlorides <dbl> -0.567, -0.425, 0.037, -0.425, NA, 0.556, 0...  
## $ FreeSulfurDioxide <dbl> NA, 15, 214, 22, -167, -37, 287, 523, -213,...  
## $ TotalSulfurDioxide <dbl> 268, -327, 142, 115, 108, 15, 156, 551, NA,...  
## $ Density <dbl> 0.99280, 1.02792, 0.99518, 0.99640, 0.99457...  
## $ pH <dbl> 3.33, 3.38, 3.12, 2.24, 3.12, 3.20, 3.49, 3...  
## $ Sulphates <dbl> -0.59, 0.70, 0.48, 1.83, 1.77, 1.29, 1.21, ...  
## $ Alcohol <dbl> 9.9, NA, 22.0, 6.2, 13.7, 15.4, 10.3, 11.6,...  
## $ LabelAppeal <int> 0, -1, -1, -1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 2...  
## $ AcidIndex <int> 8, 7, 8, 6, 9, 11, 8, 7, 6, 8, 5, 10, 7, 8,...  
## $ STARS <int> 2, 3, 3, 1, 2, NA, NA, 3, NA, 4, 1, 2, 2, 3...

### Descriptive Statistics

Descriptive statistics were performed for all predictor and response variables to explore the data. Some of our predictors are numerical continuous (such as VolatileAcidity or Density), while others are numerical discrete (such as LabelAppeal or TARGET). There are no categorical variables in this data set.

We will now look at the summary of descriptive statistics:

## vars n mean sd median trimmed mad min  
## TARGET 1 12795 3.03 1.93 3.00 3.05 1.48 0.00  
## FixedAcidity 2 12795 7.08 6.32 6.90 7.07 3.26 -18.10  
## VolatileAcidity 3 12795 0.32 0.78 0.28 0.32 0.43 -2.79  
## CitricAcid 4 12795 0.31 0.86 0.31 0.31 0.42 -3.24  
## ResidualSugar 5 12179 5.42 33.75 3.90 5.58 15.72 -127.80  
## Chlorides 6 12157 0.05 0.32 0.05 0.05 0.13 -1.17  
## FreeSulfurDioxide 7 12148 30.85 148.71 30.00 30.93 56.34 -555.00  
## TotalSulfurDioxide 8 12113 120.71 231.91 123.00 120.89 134.92 -823.00  
## Density 9 12795 0.99 0.03 0.99 0.99 0.01 0.89  
## pH 10 12400 3.21 0.68 3.20 3.21 0.39 0.48  
## Sulphates 11 11585 0.53 0.93 0.50 0.53 0.44 -3.13  
## Alcohol 12 12142 10.49 3.73 10.40 10.50 2.37 -4.70  
## LabelAppeal 13 12795 -0.01 0.89 0.00 -0.01 1.48 -2.00  
## AcidIndex 14 12795 7.77 1.32 8.00 7.64 1.48 4.00  
## STARS 15 9436 2.04 0.90 2.00 1.97 1.48 1.00  
## max range skew kurtosis se IQR Q0.25 Q0.75  
## TARGET 8.00 8.00 -0.33 -0.88 0.02 2.00 2.00 4.00  
## FixedAcidity 34.40 52.50 -0.02 1.67 0.06 4.30 5.20 9.50  
## VolatileAcidity 3.68 6.47 0.02 1.83 0.01 0.51 0.13 0.64  
## CitricAcid 3.86 7.10 -0.05 1.84 0.01 0.55 0.03 0.58  
## ResidualSugar 141.15 268.95 -0.05 1.88 0.31 17.90 -2.00 15.90  
## Chlorides 1.35 2.52 0.03 1.79 0.00 0.18 -0.03 0.15  
## FreeSulfurDioxide 623.00 1178.00 0.01 1.84 1.35 70.00 0.00 70.00  
## TotalSulfurDioxide 1057.00 1880.00 -0.01 1.67 2.11 181.00 27.00 208.00  
## Density 1.10 0.21 -0.02 1.90 0.00 0.01 0.99 1.00  
## pH 6.13 5.65 0.04 1.65 0.01 0.51 2.96 3.47  
## Sulphates 4.24 7.37 0.01 1.75 0.01 0.58 0.28 0.86  
## Alcohol 26.50 31.20 -0.03 1.54 0.03 3.40 9.00 12.40  
## LabelAppeal 2.00 4.00 0.01 -0.26 0.01 2.00 -1.00 1.00  
## AcidIndex 17.00 13.00 1.65 5.19 0.01 1.00 7.00 8.00  
## STARS 4.00 3.00 0.45 -0.69 0.01 2.00 1.00 3.00

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | n | missing | miss\_ratio | mean | sd | min | max | skew | kurtosis | median | IQR | Q0.25 | Q0.75 |
| TARGET | 12795 | 0 | 0.0000 | 3.0290739 | 1.9263682 | 0.00000 | 8.00000 | -0.3263010 | -0.8772457 | 3.00000 | 2.000000 | 2.00000 | 4.000000 |
| FixedAcidity | 12795 | 0 | 0.0000 | 7.0757171 | 6.3176435 | -18.10000 | 34.40000 | -0.0225860 | 1.6749987 | 6.90000 | 4.300000 | 5.20000 | 9.500000 |
| VolatileAcidity | 12795 | 0 | 0.0000 | 0.3241039 | 0.7840142 | -2.79000 | 3.68000 | 0.0203800 | 1.8322106 | 0.28000 | 0.510000 | 0.13000 | 0.640000 |
| CitricAcid | 12795 | 0 | 0.0000 | 0.3084127 | 0.8620798 | -3.24000 | 3.86000 | -0.0503070 | 1.8379401 | 0.31000 | 0.550000 | 0.03000 | 0.580000 |
| ResidualSugar | 12179 | 616 | 4.8144 | 5.4187331 | 33.7493790 | -127.80000 | 141.15000 | -0.0531229 | 1.8846917 | 3.90000 | 17.900000 | -2.00000 | 15.900000 |
| Chlorides | 12157 | 638 | 4.9863 | 0.0548225 | 0.3184673 | -1.17100 | 1.35100 | 0.0304272 | 1.7886044 | 0.04600 | 0.184000 | -0.03100 | 0.153000 |
| FreeSulfurDioxide | 12148 | 647 | 5.0567 | 30.8455713 | 148.7145577 | -555.00000 | 623.00000 | 0.0063930 | 1.8364966 | 30.00000 | 70.000000 | 0.00000 | 70.000000 |
| TotalSulfurDioxide | 12113 | 682 | 5.3302 | 120.7142326 | 231.9132105 | -823.00000 | 1057.00000 | -0.0071794 | 1.6746665 | 123.00000 | 181.000000 | 27.00000 | 208.000000 |
| Density | 12795 | 0 | 0.0000 | 0.9942027 | 0.0265376 | 0.88809 | 1.09924 | -0.0186938 | 1.8999592 | 0.99449 | 0.012795 | 0.98772 | 1.000515 |
| pH | 12400 | 395 | 3.0871 | 3.2076282 | 0.6796871 | 0.48000 | 6.13000 | 0.0442880 | 1.6462681 | 3.20000 | 0.510000 | 2.96000 | 3.470000 |
| Sulphates | 11585 | 1210 | 9.4568 | 0.5271118 | 0.9321293 | -3.13000 | 4.24000 | 0.0059119 | 1.7525655 | 0.50000 | 0.580000 | 0.28000 | 0.860000 |
| Alcohol | 12142 | 653 | 5.1036 | 10.4892363 | 3.7278190 | -4.70000 | 26.50000 | -0.0307158 | 1.5394949 | 10.40000 | 3.400000 | 9.00000 | 12.400000 |
| LabelAppeal | 12795 | 0 | 0.0000 | -0.0090660 | 0.8910892 | -2.00000 | 2.00000 | 0.0084295 | -0.2622916 | 0.00000 | 2.000000 | -1.00000 | 1.000000 |
| AcidIndex | 12795 | 0 | 0.0000 | 7.7727237 | 1.3239264 | 4.00000 | 17.00000 | 1.6484959 | 5.1900925 | 8.00000 | 1.000000 | 7.00000 | 8.000000 |
| STARS | 9436 | 3359 | 26.2524 | 2.0417550 | 0.9025400 | 1.00000 | 4.00000 | 0.4472353 | -0.6925343 | 2.00000 | 2.000000 | 1.00000 | 3.000000 |

### Quality of the Data

From the summary of the data statistic, we can see that we may have a problem with the quality of the data. Not only we have missing values for variables: ResiudalSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, and STARS, but we have range of values that do not make sense for the type of data we have. Chemical component of wine such as Alcohol level, ResidualSugar, CitricAcid, FixedAcidity, VolatileAcidity, Chlorides, Sulphates, FreeSulphurDioxide, TotalSulfurDioxide cannot have negative values.

Based on some research and in consultation with a wine connoisseur, we compile the information on the predictors.

***Acid Index:*** Acid Index is a proprietary method of testing total acidity of wine by using a weighted average.  
Range: Unkown

***Fixed Acidity:*** Fixed Acidity in wine cannot be evaporated (as opposed to Volatile Acidity) is most constituated of Tartaric acid and Malic acid. Fixed Acidity will provide the tartness to the wine which is part of its characteristic. Tartaric and Malic acid are produced by the grapes and are highly dependent on the weather; grapes grown in cool weather will produce high level of acidity (and low level of sugar) while grapes produced in warm level with have low level of acidity and high level of sugar.  
Typically measured in g/dm^3  
Tartaric Acid: 1 - 4 g/dm^3 Malic Acid: 0 - 8 g/dm^3

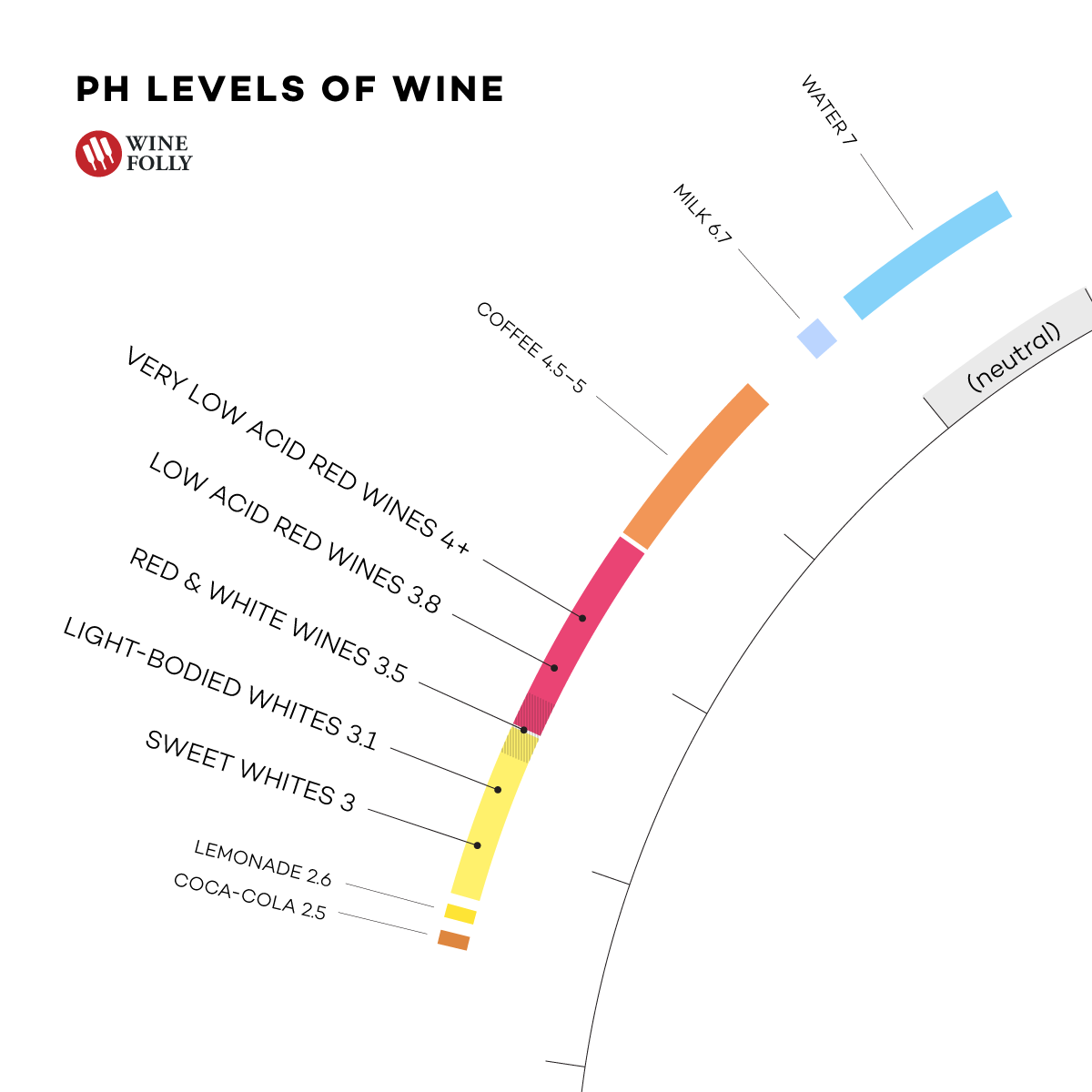
In addition, wine could have level of succinic acid, which is produced by yeast during the fermentation process, range: 0.5 - 2 g/dm^3.

***Citric Acid:*** Citric Acid is considered a fixed Acid (see above). Grape contains very low level of citric acid and most, if not all of the citric acid naturally present in the grapes is consumed by bacteria during fermentation. Many winemakers use citric acid to acidify wines that are too basic and as a flavor additive. Citric Acid is what would give wine a “freshness”. Typically measured in g/dm^3  
Range: 0 - 0.5 g/dm^3

***Volatile Acidity:*** Volatile acidity in wine is due predominently to acetic acid (Vinegar), it evaporates hence volatile and is produced as part of the fermenation process.  
Typically measured in g/dm^3  
Range: negligeable to 3 g/dm^3

U.S. legal limits of Volatile Acidity:  
Red Table Wine 1.2 g/dm^3  
White Table Wine 1.1 g/dm^3

***pH:*** pH is a measure of a solution’s acidity and is measured on a logarithm scale. Hence a wine with a pH of 3 is 10 times more acidic than a wine with a pH of 4. It is a scale from 0 - 14 with the lower the number the more acidic the compound. pH of 7 is neutral.

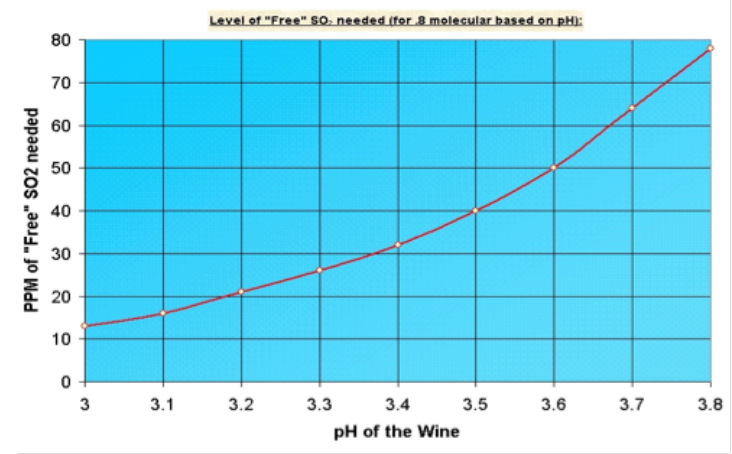


pH Scale for drinks

***Chlorides:*** Presence of Sodium Chloride in wine is largely dependent on where the grapes where grown as the grapes will absorbe the minerals from the soil. It is responsible for saltiness of the wine.  
Typically measured in g/dm^3  
Range: 2 - 4 g/dm^3

***Sulphates:*** Usually due to potassium sulphates.  
Typically measured in g/dm^3  
Range: not known, possibly less than 1 g/dm^3

***FreeSulfurDioxide:*** Free is the portion of sufphite not chemically bounded to other chemical compound of wine such as aldehydes, acids, furfural, sugars, solids, yeast/bacteria, etc that are naturally present in wine. Free protects wine from oxidation and microbial contamination during ageing/storage. This allows the wine to continue developing safely throughout its entire lifetime. The free levels required to protect wine from oxidation and spoiliage is pH dependent; the higher the pH the higher the level of required.  
Typically measured in mg/dm^3  
Range: varies on pH



SO2 vs pH

***TotalSulfurDioxide:*** Total = Free + bound .  
Typically measured in mg/dm^3  
Range: There is a legal limit of 350 mg/dm^3 Total . From a sensory point of view, Total < 100 mg/dm^3.

***Alcohol:*** Alcohol in wine is the direct result of fermentation. As the sugar is consumed by yeast, alcohol, most predominantly ethanol is produced. Alcohol is measured in Alcohol by volume or ABV  
Range for wine: 5 - 22 ABV

***Residual Sugars:*** Residual Sugars are the remaining sugars that have not been transformed as part of fremenation process. The higher the level of alcohol, the lower the residual sugars.  
Typically measured in g/dm^3 Range: 2 - 150

***Density:*** Density is defined by the mass or weight per volume of a material. In the case of liquids, density is often measured in units of g/mL, which is the weight in grams (g) of each cm^3 of liquid. Density of water is 1. Density of wine is very closed to water.  
Typically measured in g/cm^3

***Label Apeal:*** 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.

***STARS:*** Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor

***TARGET:*** Number of cases of wine purchased. Our response variable.

Now that we have a better understanding of each predictor and response variable and typical range, let us examine the negative values in our data set.

We have 89.67 of rows in our data set that have at least one negative values.

We have several possibilities on how to deal with these negative entries.

1. Remove any rows from data set with a negative entry
2. Set these negative entries to NA and impute the data
3. For the variables with negatvie values, we could offset all the values by set number that would displaced the whole distribution to the rigth.
4. Switch the sign of every negative values

We will examine each method and determine which is the best approach for this data set.

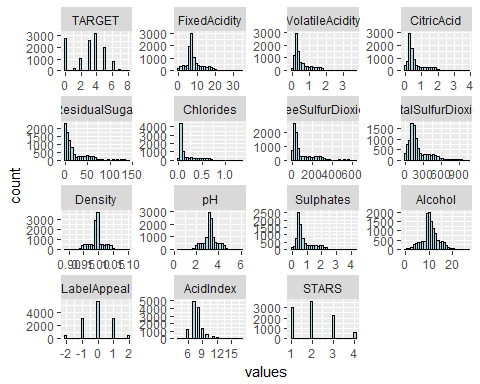
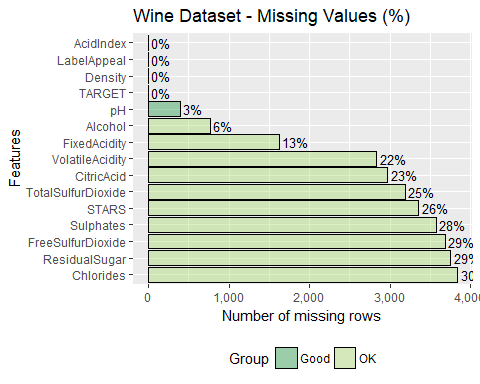
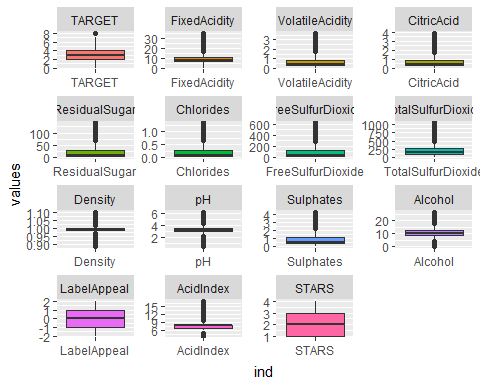
#### 1.Remove any rows with a negative entry

Since we have over 89 percent of our data with at least one negative entries, this approach is not a viable one.

#### 2. Set negative entries to NA and Impute

As we could see with already have quite a bit of data with NA. We will attempt this approach.

|  |  |
| --- | --- |
|  | x |
| TARGET | 0 |
| FixedAcidity | 1621 |
| VolatileAcidity | 2827 |
| CitricAcid | 2966 |
| ResidualSugar | 3752 |
| Chlorides | 3835 |
| FreeSulfurDioxide | 3683 |
| TotalSulfurDioxide | 3186 |
| Density | 0 |
| pH | 395 |
| Sulphates | 3571 |
| Alcohol | 771 |
| LabelAppeal | 0 |
| AcidIndex | 0 |
| STARS | 3359 |

The residual histograms indicate that a few of the distributions (without the negative and NA values) are skewed; Fixed Acidity, Volatile Acidity, Free SulfurDioxide, Total SulfurDioxide, Citric Acid, Residual Sugar, Chlorides, and Sulphates have a skeweed distribution. This may prove a problem if we choose to take this approach and impute the data with Mice package since its algorithms are sensitive to skewness of the data.

#### 3. Add an Offset to shift distribution to the right

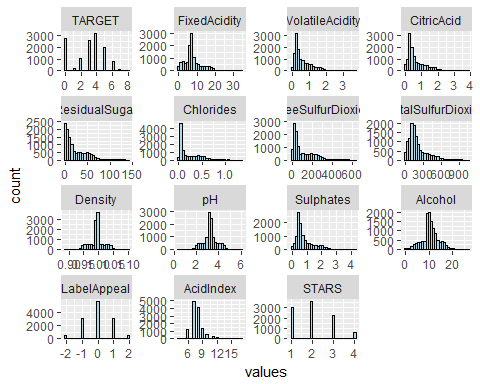
In this approach, we would offset each values for the variables with negative by a common number to shift the distribution into a valid range for this variable. Let us consider the variable Alcohol. The lowest negative value for this variable is .

To bring this value in the lower range of the Alcohol level for wine (5), we would need to offset it by about 9, or at the very least 8. This would force the highest value for this variable, currently - to 34.5, which is very much outside the valid range for this variable.

Similarly for other variables, hence offsetting each variables is not a vaviable solution.

#### 4. Switching Negative Entries to positive

In this approach, we would just switch each negative entries by their corresponding positive by multiplying each negative entries by -1. This is a very arbitrary solution and since we have absolutely no idea how the corruption of the data set took place; Invalid measuring of various chemical compound, misscalibration of the measuring isntruments, sabotage… and we were not able to gain additional information on how the data set was collected. This appraoch seemed to be riskier than the imputation approach.



The resulting histograms indicate that a few of the distributions mainly Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulphates are right skweed but seemed consistent with the distributions without negative values and NA.

Although we are concerns on imputing about 30% of the data for some variables, since we have no insight on how the data was corrupted. There is no indication of calibration errors or equipment failure and although there does not seemed to be a patern the the negative data, the percentage of incorrect data is too high to be only the result of data entry error. We will therefore pursue imputation of the negative data after replacing them with NA.

We need to consider the case of missing data prior to imputation and devise a strategy for these was well.

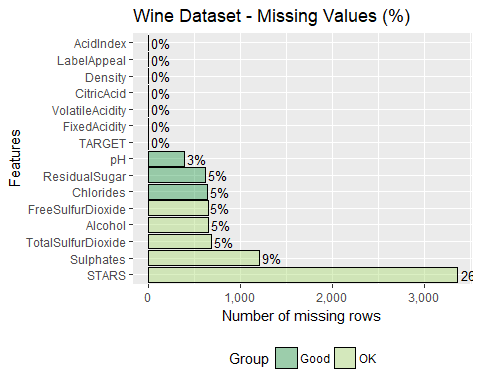
### Missing Values

8 out of 15 variables contain missing values. The predictor ResidualSugarcontains 616 (5%) missing values, Chlorides with 638 (5%), FreeSulfurDioxide with 647 (5%), TotalSulfurDioxide with 682 (5%), pH with 395 (3%), Sulphates with 1,210 (9%), Alcohol with 653 (5%), and STARS with 3,359 (26%) missing values .

There are a total of 8,200 NA values within the 12,795 observations and 15 columns, a rate of **4.27%** of the entire data set.

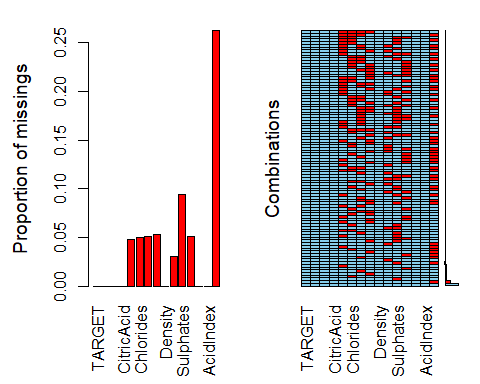
## [1] 4.27

|  |  |
| --- | --- |
|  | x |
| TARGET | 0 |
| FixedAcidity | 0 |
| VolatileAcidity | 0 |
| CitricAcid | 0 |
| ResidualSugar | 616 |
| Chlorides | 638 |
| FreeSulfurDioxide | 647 |
| TotalSulfurDioxide | 682 |
| Density | 0 |
| pH | 395 |
| Sulphates | 1210 |
| Alcohol | 653 |
| LabelAppeal | 0 |
| AcidIndex | 0 |
| STARS | 3359 |

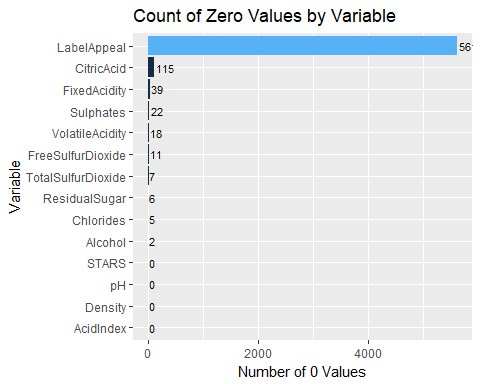


***Stars*** variable is missing 26% of data. However, if we consider that an absence of rating is significant, we consider ‘NA’ entries for Stars variables has “No Rating” and assign a value of zero. This will preserve the ordinal charecteristic of the variable.

For the remaining missing values, we will impute the data. This will be done in addition to imputing the previously negative entries.



Before we impute the missing data, let us consider the zero values. In this data set, a value of zero is significant.



After Label Appeal, the variable with the most 0 values is Citric Acid, it is possible for some wine not to have any level of citric acid. It seems improbable that some wines will not 0 g/md^3 level of fixed acidity but since we have no insight on how the data was captured it is difficult to make any assumption about the rest of the 0 values. They will be identified as outliers for each variables.

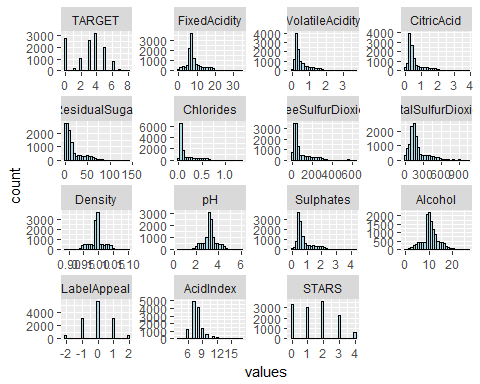
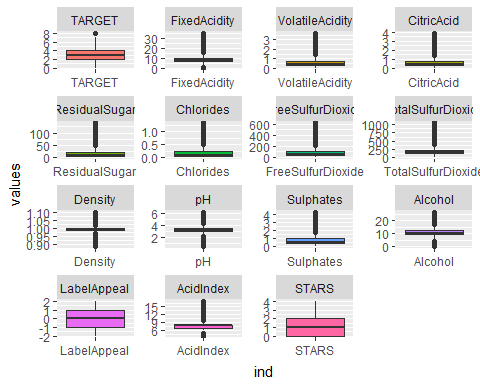
## Missing data and Negative data strategy

* We will convert the missing data for variable STARS to zero value
* We will then convert all negative data to NA and impute all NA’s using K-nearest-algorithm. We are concerned that since some of the distributions for the variables to be inputed are skewed the imputation method from MICE package will not be optimum.

## Imput Recoded dataset

We will now impute the missing data on the recoded data set. None of the data to be imputed are categorical.

Let us examine the results of the imputation.

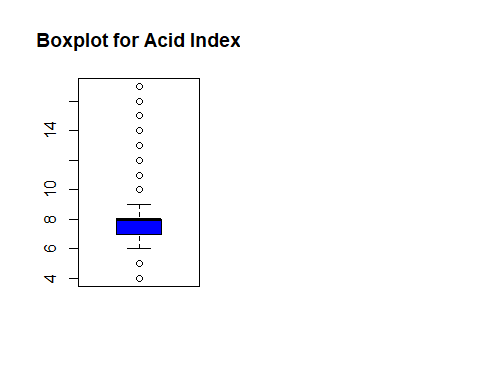
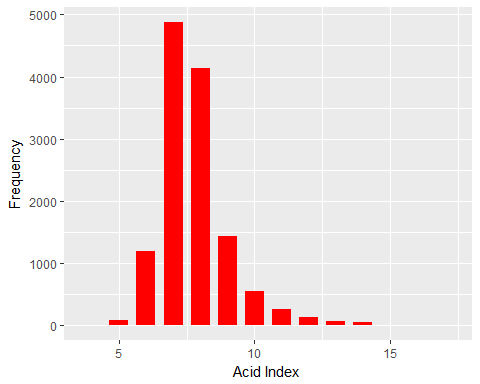
 

### Analysis of individual variables

We will now examined each predictor now that the problems with the data has been addressed to understand their distribution and determine whether any transformation is required.

#### Acid Index

The Acid Index variable is skewed to the right. There are a high number of outliers. This variable is derived by a proprietary method of testing total acidity of wine by using a weighted average. We will treat this variable as a discreet variable. We would expect this variable to be highly corelated with the other acidity variables and negatively corelated with pH.

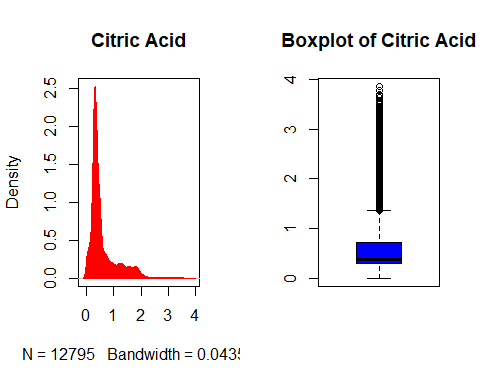


***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | 5, 4 |
| Highest | 17, 16, 15, 14, 13, 12, 11, 10 |

#### Citric Acid

Citric Acid is found only in very minute quantities in wine grapes but is sometimes added as part of wine making process. It provides a “freshness” to the wine. The distribution for this variable is right skewed with a high number of outliers. We would expect this variable to be corelated to the Acidity Index and negatively corelated to pH.

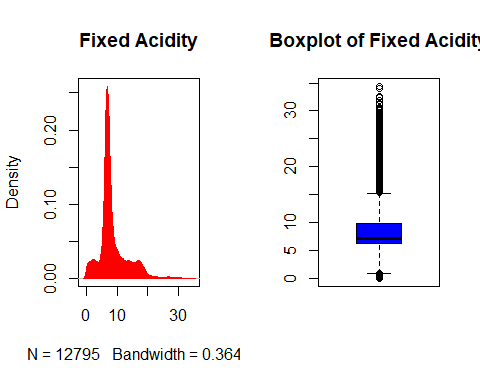


***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | None |
| Highest | 3.86, 3.77, 3.7, 3.68, 3.63, 3.57, 3.56, 3.53, 3.52, 3.5 |

#### Fixed Acidity

Acids are major wine constituents and contribute greatly to its taste. They impart the sourness or tartness that is a fundamental feature in wine taste. The distribution for this variable is right skewed with a high level of outliers. We would expect this variable to be highly corelated with the Acidity Index and negatively corelated with pH.

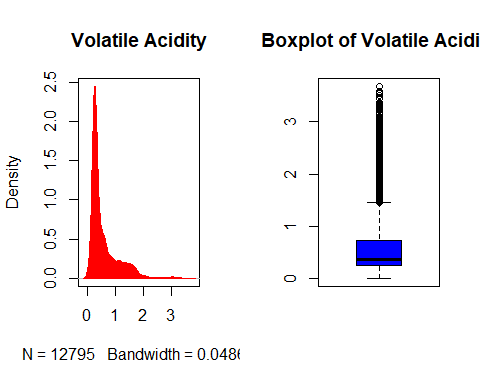


***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0 |
| Highest | 34.4, 34.1, 32.6, 32.5, 32.4, 31.8, 31.3, 31.2, 31.1, 30.7 |

#### Volatile Acidity

Volatile Acidity is a measure of the level of acetic acid in the wine. It is basically the process of wine turning into vinegar. This variable has a rightl skewed distribution with a high number of outliers. We would expect this variable to be corelated with the Acidity Index and negatively corelated wtih pH.

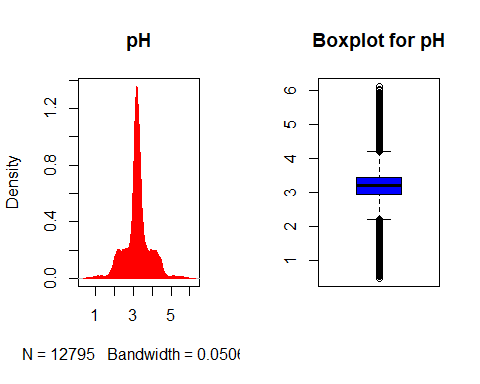


***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | None |
| Highest | 3.68, 3.59, 3.565, 3.55, 3.5, 3.44, 3.38, 3.37, 3.35, 3.33 |

#### pH

The pH of wine is a measure of acidity. PH varies between 0 and 14 with a value of 7 to be neutral and a value less than 7 to be acidic, the lower the pH, the higher content of acid present. we would expect this variable to be negatively corelated with acidity linked variables. Wines tends to have a pH around 3 and we can see from the distribution that our data reflect this fact. This variable has a symmetrical distribution with a high number of outliers.

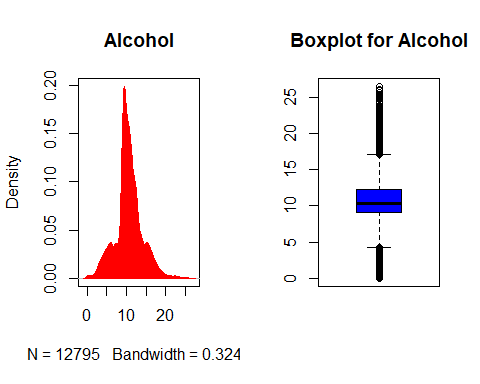


***Extreme Variables***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | 2.2, 2.19, 2.18, 2.17, 2.16, 2.15, 2.14, 2.13, 2.12, 2.11 |
| Highest | 6.13, 6.05, 6.02, 5.94, 5.91, 5.88, 5.87, 5.85, 5.84, 5.81 |

#### Alcohol

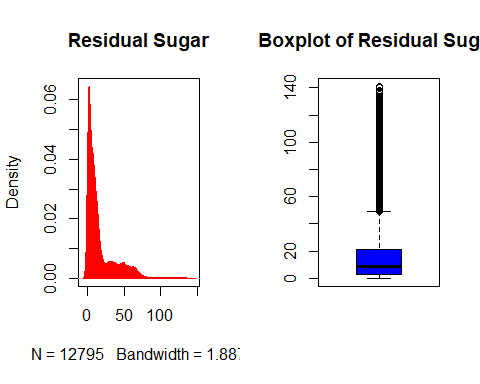
There does not appear to be any skewness and seems to have a symetrical distribution. This variable indicates the level of alcohol in the wine.

 ***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | 4.2, 4.15, 4.1333333, 4.1, 4, 3.9, 3.85, 3.8, 3.7, 3.6333333 |
| Highest | 26.5, 26.1, 26, 25.6, 25.4, 25.2, 25.1, 25, 24.8, 24.7 |

#### Residual Sugar

Residual Sugar measure the content of sugar remaining in the wine after the fermentation process took place or sugar that had been added to the wine making process to make it sweetter. A dry wine should have no perceptible sweetness. A sweet wine - such as a late-harvest type wine or icewine or desert wine - can have any level of residual sugar. This distribution is highly skweed to the right with a high number of outliers.

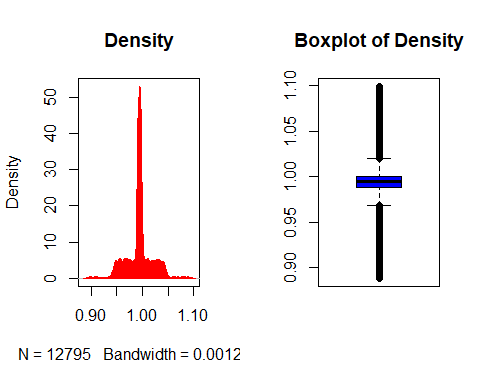


***Extreme Variables***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | None |
| Highest | 141.15, 140.65, 138, 137.6, 136.5, 135, 134.9, 134, 133.2, 133.1 |

#### Density

Density of wine is very close to density of water. The distribution for this variable symmetrical with a high numbers of outliers.

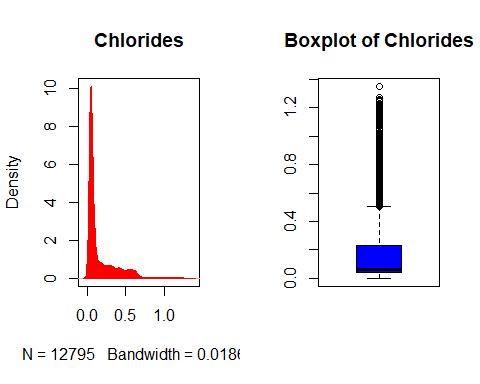


***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | 0.96851, 0.96849, 0.96847, 0.96845, 0.96844, 0.96843, 0.96841, 0.96836, 0.96835, 0.96834 |
| Highest | 1.09924, 1.09791, 1.09695, 1.09679, 1.09658, 1.09608, 1.09593, 1.09576, 1.0955, 1.09512 |

#### Chlorides

Chlorides level contributes to saltiness of a wine and is related to the “terroir” where the grape was grown. The distribution for this variable is highly right skweed with a high presence of outliers.

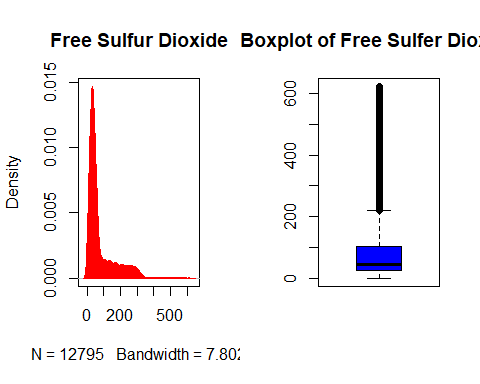


***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | None |
| Highest | 1.351, 1.275, 1.27, 1.261, 1.26, 1.254, 1.252, 1.251, 1.25, 1.249 |

#### Free Sulfur Dioxide

Free Sulfur Dioxide () (not associated with wine molecules) is effectively the buffer against microbes and oxidation. It occurs naturally in wine but is also usually added in the wine process. The distribution for this variable is right skwed with a high number of outliers.

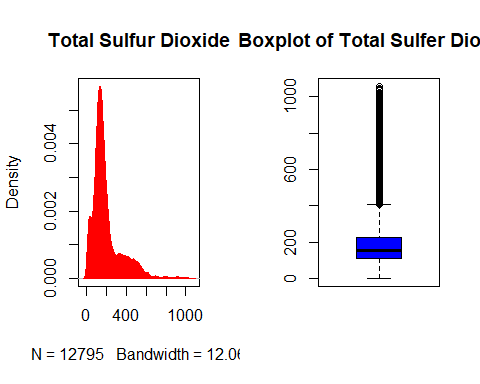


\*\*\* Extreme Values\*\*\*

|  |  |
| --- | --- |
| Range | Values |
| Lowest | None |
| Highest | 623, 622, 618, 617, 613, 611, 608, 607, 605, 604 |

#### Total Sulfur Dioxide

This variabale is indicative of the total level of Sulfur Dioxide (). It represents the bind and the free . We would expect this variable to be highly corelated to “Free Sulfur Dioxide”. The distribution for this variable is right skwed with a high level of outliers.

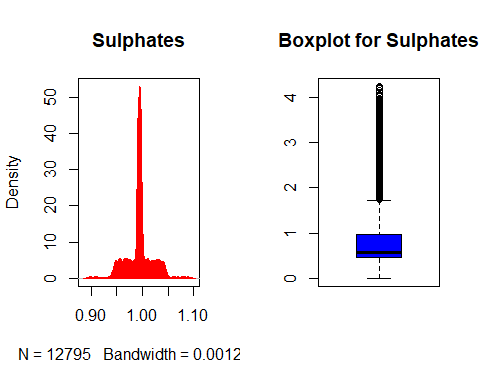


***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | None |
| Highest | 1057, 1054, 1048, 1041, 1032, 1022, 1016, 1011, 1009, 1006 |

#### Sulphates

Level of Sulphates in wine, possibly potassium sulphates. The distribution for this variable is symmetrical with a high number of outliers.

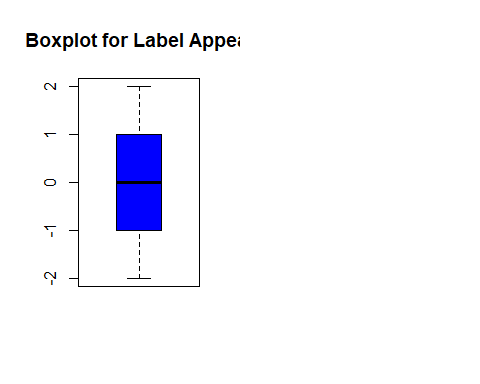


***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | None |
| Highest | 4.24, 4.21, 4.19, 4.16, 4.11, 4.1, 4.05, 4.04, 4.02, 3.98 |

#### Label Appeal

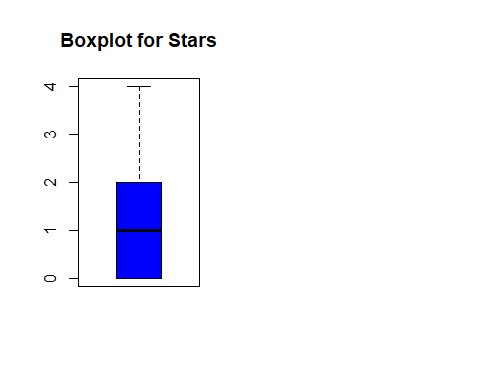
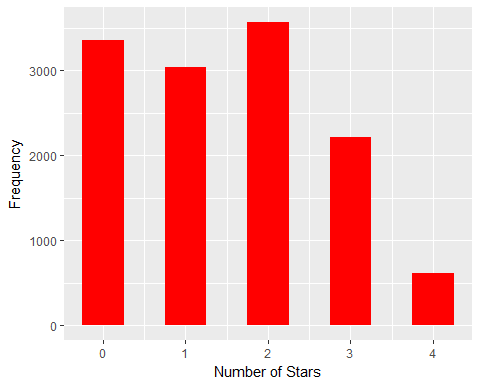
The variable is a measure of how appealing to customer the label is. High numbers suggest customers like the label design and negative number suggest they do not. It is not rare to see customer buy a wine based on the label appeal. We will treat this variable as a ordered categorical variable. The distribution for this variable is fairly symmetrical without presence of outliers.

 ***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | None |
| Highest | None |

#### Stars

This variable is indicative of a rating given to a wine by a team of expert, a rating of 1 = poor and rating of 4 is excellent. We notice a high level of missing values for this variables that will be addressed in the next section. We will treat this variable as ordered categorical. The distribution for this variable is almost uniform without presence of outliers.

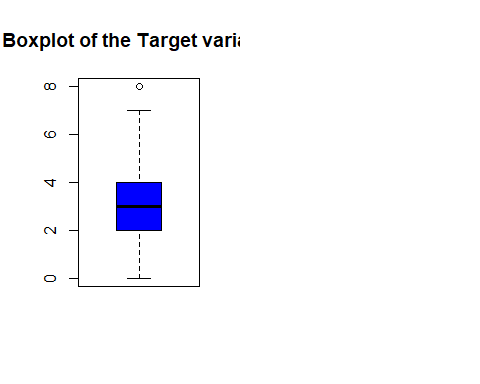
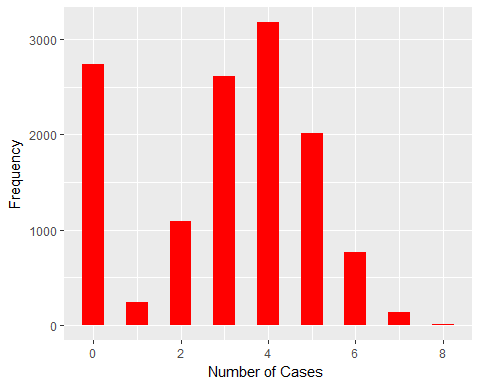


***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | None |
| Highest | None |

#### TARGET

This represents our response variables. It indicates the number of cases purchased. This is a numerical count variable. The distribution for this variables is multimodals with peaks at value = 0 and value = 4. There is outliers.



***Extreme Values***

|  |  |
| --- | --- |
| Range | Values |
| Lowest | None |
| Highest | 8 |

We would surmize that the data for the chemical content variables (Fixed Acidity, Volatile Acidity, Citric Acid, Redidual Sugar, Chlorides, and Sulphates) has been transformed. We would expect these variables to have positive values mostly measured in mg/l (mili-grams per liter) or in ppm (part-per-million). They appear to have been centered.

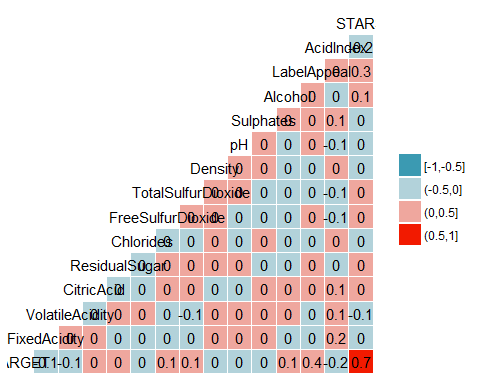
We will now explore the corlation between the variables.

## Variable to Variable Analysis

Now that we have completed our analysis of individual variables with will explore possible relation between variables.

### Correlation Analysis

The correlation matrix shown below highlights correlations among several predictor variables. Correlation between between Target and Stars is high at 0.7. The second highest correlation is between the Target variable and Label Appeal at 0.4. Note, both these values increased from correlation numbers prior to addressing the data issues.



The predictor Stars and Label Appear have the highest positive corelation with the response vaiable. As expected, the various acidity variables show some corelation with the Acidity Index and negative corelation with the pH.

Let us now look at the correlation between each response variable and the predictors.

## [1] -0.06210303

## [1] -0.116243

## [1] 0.01871322

## [1] 0.02284846

## [1] -0.04833133

## [1] 0.05919466

## [1] 0.06414211

## [1] -0.0355175

## [1] -0.009046558

## [1] -0.04668931

## [1] 0.06651698

## [1] 0.3565005

## [1] -0.2460494

## [1] 0.6853815

### Multicollinearity

This section will test the predictor variables to determine if there is correlation among them. Variance inflaction factor (VIF) is used to detect multicollinearity, specifically among the entire set of predictors versus within pairs of variables.

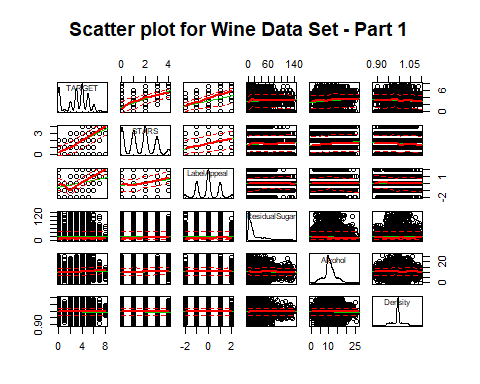
Testing for collinearity among the predictor variables, we see that none of the numeric predictor variables appear to have a problem with collinearity based on their low VIF scores.

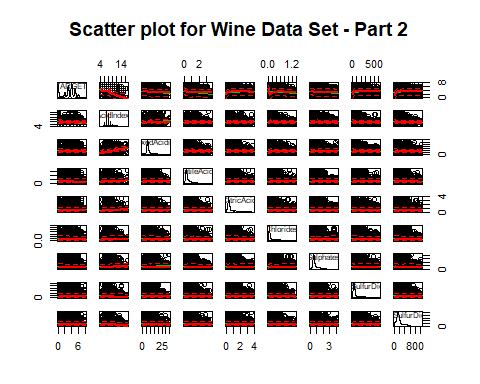
## No variable from the 14 input variables has collinearity problem.   
##   
## The linear correlation coefficients ranges between:   
## min correlation ( Alcohol ~ Chlorides ): 0.0009825581   
## max correlation ( STARS ~ LabelAppeal ): 0.2510298   
##   
## ---------- VIFs of the remained variables --------   
## Variables VIF  
## 1 FixedAcidity 1.041208  
## 2 VolatileAcidity 1.015287  
## 3 CitricAcid 1.008204  
## 4 ResidualSugar 1.003513  
## 5 Chlorides 1.002006  
## 6 FreeSulfurDioxide 1.006917  
## 7 TotalSulfurDioxide 1.015643  
## 8 Density 1.005174  
## 9 pH 1.008648  
## 10 Sulphates 1.005934  
## 11 Alcohol 1.004677  
## 12 LabelAppeal 1.073064  
## 13 AcidIndex 1.086549  
## 14 STARS 1.110568

Since the VIF scores are closed to 1, there is no multicollinearity among factors.

### Relations between variables

We will now explore the relationships between the various predictors and between the response variables and each predictors. We will do so graphically with the air of a scatter plot.





From the plots, it is difficult to identify a relationship between the TARGET response and some of the predictors. Also, we have very low correlation between response and the predictors and between predictors even when would expect some. For example, we would have expected a negative corelation between Alcohol and Residual Sugars, a negative corelation betwen the pH and other various acidity variables (Fixedacidity, VolatileAcidity, and Citric Acid).

# 2. Data Preparation

## Data Transformation

As indicated earlier, it appears that some of the predictors, mostly the chemical elements such as; Residual Sugar, Chloride, Free Sulur Dioxide, Total Sulfur Dioxide, Suphates, Fixed Acidity, and Volatile Acidity have skewed distribution. We will not consider further transformation on the data set now what missing values have been imputed except for changing some of predictors to categorical values; Stars and Label Apeal.

# 3.Model Buildings

We will now split our data set into a trainning data set and development evaluation data set. This will allow us to evaluate the accuracy of our models on a non-training data set to aliviate the risk of overfitting. For reproducibility we will upload our data sets to github.

## 3.1 Poisson Models

### Poisson Base model

The variance of the repsonse variable, the number of cases of wine sold, is larger than the mean - 3.710895 and 3.029074 respectively. Since the variace and mean are not equal, this is an indication the poisson condition has not been met. This is an indication of over dispersion.

Over dispersion in the data means the MLE is correct but standard errors are wrong. Can use MLE but will need to correct the standard error.

Examiniation of the out of the poisson model confirms over dispersion in the data. The residual deviance and the degrees of freedom indicates variance greater than 1. Due to the dispersion in the data, the standard error values cannot be used. The quasi poisson model will be calculated to correct for the incorrect standard error values.

##   
## Call:  
## glm(formula = TARGET ~ ., family = poisson, data = wine\_data\_dev\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9210 -0.7357 0.0698 0.5853 3.2610   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.405e+00 2.328e-01 6.037 1.57e-09 \*\*\*  
## FixedAcidity -5.983e-04 1.366e-03 -0.438 0.661279   
## VolatileAcidity -7.122e-02 1.237e-02 -5.760 8.43e-09 \*\*\*  
## CitricAcid 2.186e-02 1.083e-02 2.019 0.043520 \*   
## ResidualSugar 1.797e-04 2.824e-04 0.636 0.524634   
## Chlorides -8.554e-02 3.040e-02 -2.814 0.004895 \*\*   
## FreeSulfurDioxide 2.252e-04 6.379e-05 3.531 0.000415 \*\*\*  
## TotalSulfurDioxide 1.426e-04 4.001e-05 3.565 0.000364 \*\*\*  
## Density -1.285e-01 2.281e-01 -0.564 0.573082   
## pH -2.048e-02 9.006e-03 -2.274 0.022963 \*   
## Sulphates -2.867e-02 1.021e-02 -2.809 0.004976 \*\*   
## Alcohol 2.739e-03 1.776e-03 1.542 0.123002   
## LabelAppeal 1.322e-01 7.326e-03 18.051 < 2e-16 \*\*\*  
## AcidIndex -8.824e-02 5.509e-03 -16.016 < 2e-16 \*\*\*  
## STARS 3.099e-01 5.424e-03 57.140 < 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: 16145 on 8957 degrees of freedom  
## Residual deviance: 10332 on 8943 degrees of freedom  
## AIC: 32679  
##   
## Number of Fisher Scoring iterations: 5

### Quasi Poisson Model

The Quasi Poission model corrects the standard where there is dispersion in the data. The standard error of the Quasi Poission model is lorger than that of the Poisson model.

##   
## Call:  
## glm(formula = TARGET ~ ., family = quasipoisson(), data = wine\_data\_dev\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9210 -0.7357 0.0698 0.5853 3.2610   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.405e+00 2.154e-01 6.524 7.22e-11 \*\*\*  
## FixedAcidity -5.983e-04 1.264e-03 -0.474 0.635859   
## VolatileAcidity -7.122e-02 1.144e-02 -6.225 5.04e-10 \*\*\*  
## CitricAcid 2.186e-02 1.002e-02 2.182 0.029163 \*   
## ResidualSugar 1.797e-04 2.613e-04 0.688 0.491741   
## Chlorides -8.554e-02 2.813e-02 -3.041 0.002365 \*\*   
## FreeSulfurDioxide 2.252e-04 5.903e-05 3.816 0.000137 \*\*\*  
## TotalSulfurDioxide 1.426e-04 3.702e-05 3.853 0.000118 \*\*\*  
## Density -1.285e-01 2.111e-01 -0.609 0.542538   
## pH -2.048e-02 8.334e-03 -2.458 0.014006 \*   
## Sulphates -2.867e-02 9.446e-03 -3.035 0.002410 \*\*   
## Alcohol 2.739e-03 1.643e-03 1.667 0.095593 .   
## LabelAppeal 1.322e-01 6.779e-03 19.508 < 2e-16 \*\*\*  
## AcidIndex -8.824e-02 5.098e-03 -17.309 < 2e-16 \*\*\*  
## STARS 3.099e-01 5.019e-03 61.751 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for quasipoisson family taken to be 0.8562059)  
##   
## Null deviance: 16145 on 8957 degrees of freedom  
## Residual deviance: 10332 on 8943 degrees of freedom  
## AIC: NA  
##   
## Number of Fisher Scoring iterations: 5

## Poisson\_coef PoissonSE Quasi\_coef QuasiSE  
## (Intercept) 1.4053018055 2.327950e-01 1.4053018055 2.154085e-01  
## FixedAcidity -0.0005983070 1.365535e-03 -0.0005983070 1.263549e-03  
## VolatileAcidity -0.0712223493 1.236560e-02 -0.0712223493 1.144206e-02  
## CitricAcid 0.0218601627 1.082890e-02 0.0218601627 1.002013e-02  
## ResidualSugar 0.0001796580 2.823843e-04 0.0001796580 2.612941e-04  
## Chlorides -0.0855421036 3.040003e-02 -0.0855421036 2.812957e-02  
## FreeSulfurDioxide 0.0002252252 6.379126e-05 0.0002252252 5.902694e-05  
## TotalSulfurDioxide 0.0001426277 4.000838e-05 0.0001426277 3.702031e-05  
## Density -0.1285354536 2.280944e-01 -0.1285354536 2.110589e-01  
## pH -0.0204809746 9.006390e-03 -0.0204809746 8.333738e-03  
## Sulphates -0.0286714981 1.020852e-02 -0.0286714981 9.446089e-03  
## Alcohol 0.0027393432 1.776146e-03 0.0027393432 1.643493e-03  
## LabelAppeal 0.1322477447 7.326435e-03 0.1322477447 6.779252e-03  
## AcidIndex -0.0882364399 5.509198e-03 -0.0882364399 5.097737e-03  
## STARS 0.3099191606 5.423902e-03 0.3099191606 5.018812e-03

## Poisson Model - Significant Predictors

This model is based on the predictor variables that were significant from the Quasi Poisson model above.

##   
## Call:  
## glm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +   
## FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +   
## Alcohol + LabelAppeal + AcidIndex + STARS, family = poisson,   
## data = wine\_data\_dev\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9200 -0.7339 0.0696 0.5858 3.2601   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.407e+00 2.327e-01 6.045 1.49e-09 \*\*\*  
## VolatileAcidity -7.119e-02 1.237e-02 -5.757 8.56e-09 \*\*\*  
## CitricAcid 2.185e-02 1.083e-02 2.018 0.043631 \*   
## Chlorides -8.560e-02 3.040e-02 -2.816 0.004863 \*\*   
## FreeSulfurDioxide 2.248e-04 6.379e-05 3.525 0.000424 \*\*\*  
## TotalSulfurDioxide 1.434e-04 3.999e-05 3.585 0.000337 \*\*\*  
## Density -1.288e-01 2.281e-01 -0.565 0.572222   
## pH -2.044e-02 9.000e-03 -2.272 0.023116 \*   
## Sulphates -2.883e-02 1.021e-02 -2.825 0.004731 \*\*   
## Alcohol 2.714e-03 1.776e-03 1.528 0.126498   
## LabelAppeal 1.323e-01 7.326e-03 18.065 < 2e-16 \*\*\*  
## AcidIndex -8.863e-02 5.441e-03 -16.290 < 2e-16 \*\*\*  
## STARS 3.100e-01 5.424e-03 57.151 < 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: 16145 on 8957 degrees of freedom  
## Residual deviance: 10332 on 8945 degrees of freedom  
## AIC: 32676  
##   
## Number of Fisher Scoring iterations: 5

## Goodness of Fit of Poisson Model - Significant Predictors

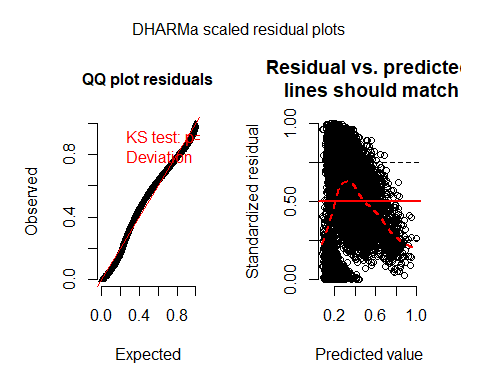
The Pearson Chi Squared and deviance tests were used to test the goodness of fit. The residual difference is the difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed.

Since this goodness of fit test is not signifcant this is an indication that the model fits the data - all the variablility in the data are explained.

## res.deviance df   
## [1,] 10332.38 8945 2.413491e-23

## Poisson Model Diagnosis

The QQ plot residulas show there is little deviation from the expected distribution. The Resiual vs, Predicted shows there is little despersion in tehd data.



## 3.2 Negaive Binomial Models

### Negative binomial base model

Negative binomial regression can be used for over-dispersed count data, that is when the conditional variance exceeds the conditional mean. It can be considered as a generalization of Poisson regression since it has the same mean structure as Poisson regression and it has an extra parameter to model the over-dispersion. If the conditional distribution of the outcome variable is over-dispersed, the confidence intervals for the Negative binomial regression are likely to be narrower as compared to those from a Poisson regression model.

First we performed a base negative binomial regression model. We also calculated and compared the mean and variance of the data set, and saw the variance aws over 20% larer than the mean.

## [1] 3.028243

## [1] 3.742755

##   
## Call:  
## glm.nb(formula = TARGET ~ ., data = wine\_data\_dev\_train, init.theta = 48453.5149,   
## link = log)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9209 -0.7357 0.0698 0.5853 3.2609   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.405e+00 2.154e-01 6.524 7.23e-11 \*\*\*  
## FixedAcidity -5.983e-04 1.264e-03 -0.473 0.635880   
## VolatileAcidity -7.122e-02 1.144e-02 -6.225 5.04e-10 \*\*\*  
## CitricAcid 2.186e-02 1.002e-02 2.182 0.029163 \*   
## ResidualSugar 1.797e-04 2.613e-04 0.688 0.491728   
## Chlorides -8.554e-02 2.813e-02 -3.041 0.002364 \*\*   
## FreeSulfurDioxide 2.252e-04 5.903e-05 3.816 0.000137 \*\*\*  
## TotalSulfurDioxide 1.426e-04 3.702e-05 3.853 0.000118 \*\*\*  
## Density -1.285e-01 2.111e-01 -0.609 0.542551   
## pH -2.048e-02 8.334e-03 -2.458 0.014003 \*   
## Sulphates -2.867e-02 9.446e-03 -3.035 0.002410 \*\*   
## Alcohol 2.739e-03 1.644e-03 1.667 0.095599 .   
## LabelAppeal 1.322e-01 6.779e-03 19.507 < 2e-16 \*\*\*  
## AcidIndex -8.824e-02 5.098e-03 -17.309 < 2e-16 \*\*\*  
## STARS 3.099e-01 5.019e-03 61.751 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(48453.51) family taken to be 0.856166)  
##   
## Null deviance: 16144 on 8957 degrees of freedom  
## Residual deviance: 10331 on 8943 degrees of freedom  
## AIC: 32681  
##   
## Number of Fisher Scoring iterations: 1

The second negative binomial regression model included backward + forward selection

## Start: AIC=32679.22  
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +   
## Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density +   
## pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS  
##   
## Df AIC  
## - FixedAcidity 1 32677  
## - Density 1 32678  
## - ResidualSugar 1 32678  
## <none> 32679  
## - Alcohol 1 32680  
## - CitricAcid 1 32681  
## - pH 1 32682  
## - Sulphates 1 32685  
## - Chlorides 1 32685  
## - FreeSulfurDioxide 1 32690  
## - TotalSulfurDioxide 1 32690  
## - VolatileAcidity 1 32711  
## - AcidIndex 1 32945  
## - LabelAppeal 1 33003  
## - STARS 1 36010  
##   
## Step: AIC=32677.41  
## TARGET ~ VolatileAcidity + CitricAcid + ResidualSugar + Chlorides +   
## FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +   
## Alcohol + LabelAppeal + AcidIndex + STARS  
##   
## Df AIC  
## - Density 1 32676  
## - ResidualSugar 1 32676  
## <none> 32677  
## - Alcohol 1 32678  
## + FixedAcidity 1 32679  
## - CitricAcid 1 32679  
## - pH 1 32681  
## - Chlorides 1 32683  
## - Sulphates 1 32683  
## - FreeSulfurDioxide 1 32688  
## - TotalSulfurDioxide 1 32688  
## - VolatileAcidity 1 32709  
## - AcidIndex 1 32953  
## - LabelAppeal 1 33001  
## - STARS 1 36009  
##   
## Step: AIC=32675.73  
## TARGET ~ VolatileAcidity + CitricAcid + ResidualSugar + Chlorides +   
## FreeSulfurDioxide + TotalSulfurDioxide + pH + Sulphates +   
## Alcohol + LabelAppeal + AcidIndex + STARS  
##   
## Df AIC  
## - ResidualSugar 1 32674  
## <none> 32676  
## - Alcohol 1 32676  
## + Density 1 32677  
## + FixedAcidity 1 32678  
## - CitricAcid 1 32678  
## - pH 1 32679  
## - Chlorides 1 32682  
## - Sulphates 1 32682  
## - FreeSulfurDioxide 1 32686  
## - TotalSulfurDioxide 1 32686  
## - VolatileAcidity 1 32708  
## - AcidIndex 1 32953  
## - LabelAppeal 1 33000  
## - STARS 1 36010  
##   
## Step: AIC=32674.15  
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +   
## TotalSulfurDioxide + pH + Sulphates + Alcohol + LabelAppeal +   
## AcidIndex + STARS  
##   
## Df AIC  
## <none> 32674  
## - Alcohol 1 32674  
## + ResidualSugar 1 32676  
## + Density 1 32676  
## + FixedAcidity 1 32676  
## - CitricAcid 1 32676  
## - pH 1 32677  
## - Chlorides 1 32680  
## - Sulphates 1 32680  
## - FreeSulfurDioxide 1 32684  
## - TotalSulfurDioxide 1 32685  
## - VolatileAcidity 1 32706  
## - AcidIndex 1 32952  
## - LabelAppeal 1 32999  
## - STARS 1 36009

The bi-directional stepwise regression model that chose the variables VolatileAcidity, CitricAcid, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, LabelAppeal, AcidIndex, and STARS reduced the AIC value to 32674, which is slightly lower than the base model.

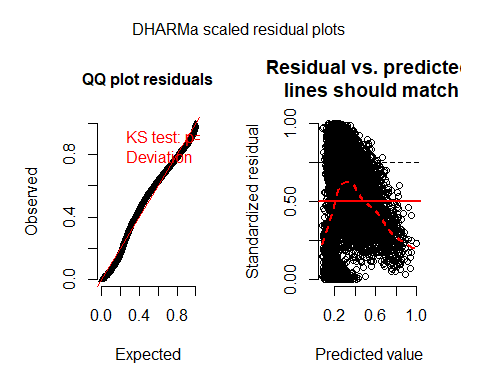
## Goodness of Fit of Negative Binomial Model - Significant Predictors

Since this goodness of fit test is not signifcant this is an indication that the model fits the data - all the variablility in the data are explained.

## res.deviance df   
## [1,] 10332.34 8946 2.604065e-23

## Negative Binomia Model Diagnosis

## Likelihood ratio tests of Negative Binomial Models  
##   
## Response: TARGET  
## Model  
## 1 VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS  
## 2 FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS  
## theta Resid. df 2 x log-lik. Test df LR stat. Pr(Chi)  
## 1 48457.40 8946 -32650.15   
## 2 48453.51 8943 -32649.22 1 vs 2 3 0.9250991 0.819367



## 3.3 Multilinear Models

One concern with using multilinear model to predict count data could lead to negative result unlike when using Negative Binomial or Poisson model. We will explore these models in later sections.

The first model we will build is a multilinear models. We will build this first model manually to keep control on the building process. We will use the anova testing as we continue to add variables in our model to confirm that the variables we are adding increase contribution to our models.

### Model a - Manual Multilinear model

We select to build our model with predictors STARS and LABELAPPEAL; these variables have the highest correlation with our response variable, TARGET.

##   
## Call:  
## lm(formula = TARGET ~ STARS + LabelAppeal, data = wine\_dev\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.9085 -0.7848 0.0915 0.7685 6.3393   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.51798 0.07467 6.937 4.28e-12 \*\*\*  
## STARS1 1.47315 0.04000 36.829 < 2e-16 \*\*\*  
## STARS2 2.57075 0.03877 66.305 < 2e-16 \*\*\*  
## STARS3 3.13782 0.04471 70.183 < 2e-16 \*\*\*  
## STARS4 3.86259 0.07077 54.581 < 2e-16 \*\*\*  
## LabelAppeal-1 0.26555 0.07775 3.416 0.000639 \*\*\*  
## LabelAppeal0 0.69607 0.07576 9.188 < 2e-16 \*\*\*  
## LabelAppeal1 1.14273 0.07899 14.467 < 2e-16 \*\*\*  
## LabelAppeal2 1.81977 0.10453 17.409 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.338 on 8949 degrees of freedom  
## Multiple R-squared: 0.5224, Adjusted R-squared: 0.522   
## F-statistic: 1224 on 8 and 8949 DF, p-value: < 2.2e-16

The F-statistic of 1224 and p-value < 2.2e-16, clearly subject that we should reject the null hypothesis that predictors STARS and LabelAppeal have collectively no affect on response variable TARGET. However, our adjusted R^2 is only of 0.522.

Let us consider adding the variable AcidityIndex and determine whether this improve the model.

##   
## Call:  
## lm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex, data = wine\_dev\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.8642 -0.8898 0.0485 0.8309 6.2794   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.17238 0.11269 19.278 < 2e-16 \*\*\*  
## STARS1 1.39845 0.03939 35.507 < 2e-16 \*\*\*  
## STARS2 2.44848 0.03852 63.572 < 2e-16 \*\*\*  
## STARS3 2.99304 0.04445 67.341 < 2e-16 \*\*\*  
## STARS4 3.70275 0.06984 53.021 < 2e-16 \*\*\*  
## LabelAppeal-1 0.29016 0.07619 3.808 0.000141 \*\*\*  
## LabelAppeal0 0.73810 0.07426 9.939 < 2e-16 \*\*\*  
## LabelAppeal1 1.20998 0.07748 15.617 < 2e-16 \*\*\*  
## LabelAppeal2 1.90504 0.10252 18.582 < 2e-16 \*\*\*  
## AcidIndex -0.20772 0.01076 -19.303 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.311 on 8948 degrees of freedom  
## Multiple R-squared: 0.5415, Adjusted R-squared: 0.541   
## F-statistic: 1174 on 9 and 8948 DF, p-value: < 2.2e-16

Adjusted R^2 is better. We will perform anova test to compare the 2.

## Analysis of Variance Table  
##   
## Model 1: TARGET ~ STARS + LabelAppeal  
## Model 2: TARGET ~ STARS + LabelAppeal + AcidIndex  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 8949 16011   
## 2 8948 15371 1 640.05 372.59 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Based on the F value and p-value, we are rejecting the null hypothesis that coefficient AcidIndex = 0. We will continue by adding Alcohol. We would expect Alcohol to have a positive relationship with TARGET and again we will test our model.

##   
## Call:  
## lm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol,   
## data = wine\_dev\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.9613 -0.8593 0.0411 0.8297 6.2469   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.005699 0.121381 16.524 < 2e-16 \*\*\*  
## STARS1 1.397857 0.039358 35.517 < 2e-16 \*\*\*  
## STARS2 2.446515 0.038492 63.559 < 2e-16 \*\*\*  
## STARS3 2.985818 0.044459 67.159 < 2e-16 \*\*\*  
## STARS4 3.687661 0.069907 52.751 < 2e-16 \*\*\*  
## LabelAppeal-1 0.294094 0.076145 3.862 0.000113 \*\*\*  
## LabelAppeal0 0.741419 0.074216 9.990 < 2e-16 \*\*\*  
## LabelAppeal1 1.216142 0.077443 15.704 < 2e-16 \*\*\*  
## LabelAppeal2 1.909748 0.102459 18.639 < 2e-16 \*\*\*  
## AcidIndex -0.206727 0.010757 -19.218 < 2e-16 \*\*\*  
## Alcohol 0.014802 0.004023 3.680 0.000235 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.31 on 8947 degrees of freedom  
## Multiple R-squared: 0.5422, Adjusted R-squared: 0.5417   
## F-statistic: 1060 on 10 and 8947 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Model 1: TARGET ~ STARS + LabelAppeal + AcidIndex  
## Model 2: TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 8948 15371   
## 2 8947 15348 1 23.225 13.539 0.0002351 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Again from the F-value and p-value, we will reject the hypothesis that the coefficient of Alcohol = 0, Adjusted R^2 has increase slightly. We will now consider Residual Sugar, although Residual Sugar is negatively correlated to Alcohol, sweet wines are raising in popularity.

##   
## Call:  
## lm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol +   
## ResidualSugar, data = wine\_dev\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.9569 -0.8563 0.0398 0.8283 6.2504   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.9971256 0.1220449 16.364 < 2e-16 \*\*\*  
## STARS1 1.3979722 0.0393594 35.518 < 2e-16 \*\*\*  
## STARS2 2.4462761 0.0384948 63.548 < 2e-16 \*\*\*  
## STARS3 2.9856354 0.0444608 67.152 < 2e-16 \*\*\*  
## STARS4 3.6873942 0.0699102 52.745 < 2e-16 \*\*\*  
## LabelAppeal-1 0.2939865 0.0761471 3.861 0.000114 \*\*\*  
## LabelAppeal0 0.7412591 0.0742182 9.988 < 2e-16 \*\*\*  
## LabelAppeal1 1.2157937 0.0774473 15.698 < 2e-16 \*\*\*  
## LabelAppeal2 1.9090138 0.1024675 18.630 < 2e-16 \*\*\*  
## AcidIndex -0.2066414 0.0107581 -19.208 < 2e-16 \*\*\*  
## Alcohol 0.0148560 0.0040237 3.692 0.000224 \*\*\*  
## ResidualSugar 0.0004361 0.0006448 0.676 0.498835   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.31 on 8946 degrees of freedom  
## Multiple R-squared: 0.5422, Adjusted R-squared: 0.5416   
## F-statistic: 963.2 on 11 and 8946 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Model 1: TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol  
## Model 2: TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol + ResidualSugar  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 8947 15348   
## 2 8946 15347 1 0.78478 0.4575 0.4988

Adjusted R^2 has not improved and and based on the F and p-value we would failed to regject to null hypothesis that the coefficint of Residual Sugar is 0. Let us adding Citric Acid; Citric Acid in wine gives it freshness. This may be a desirable feature of a wine.

##   
## Call:  
## lm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol +   
## CitricAcid, data = wine\_dev\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0347 -0.8582 0.0325 0.8270 6.2702   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.987485 0.121643 16.339 < 2e-16 \*\*\*  
## STARS1 1.396235 0.039357 35.476 < 2e-16 \*\*\*  
## STARS2 2.444634 0.038494 63.507 < 2e-16 \*\*\*  
## STARS3 2.985417 0.044450 67.164 < 2e-16 \*\*\*  
## STARS4 3.686994 0.069893 52.752 < 2e-16 \*\*\*  
## LabelAppeal-1 0.295646 0.076132 3.883 0.000104 \*\*\*  
## LabelAppeal0 0.742260 0.074201 10.003 < 2e-16 \*\*\*  
## LabelAppeal1 1.214878 0.077429 15.690 < 2e-16 \*\*\*  
## LabelAppeal2 1.909302 0.102437 18.639 < 2e-16 \*\*\*  
## AcidIndex -0.208556 0.010787 -19.333 < 2e-16 \*\*\*  
## Alcohol 0.014801 0.004022 3.680 0.000235 \*\*\*  
## CitricAcid 0.054335 0.024928 2.180 0.029307 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.309 on 8946 degrees of freedom  
## Multiple R-squared: 0.5424, Adjusted R-squared: 0.5419   
## F-statistic: 964.1 on 11 and 8946 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Model 1: TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol  
## Model 2: TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol + CitricAcid  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 8947 15348   
## 2 8946 15340 1 8.1467 4.751 0.02931 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We have a very slight increase in adjusted R^2. The F value and p-statistic support the rejection of the null hypothesis that coefficient of Citri Acid is null.

We will now consider adding Volatil Acidity, we assume that all acidity indicators were contains in the Acidity Index but Volatil Acidity correspond to Acetic Acid in wine (vinegar) that is highly undesirable.

##   
## Call:  
## lm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol +   
## CitricAcid + VolatileAcidity, data = wine\_dev\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0907 -0.8575 0.0225 0.8314 6.1925   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.083724 0.122029 17.076 < 2e-16 \*\*\*  
## STARS1 1.388966 0.039258 35.381 < 2e-16 \*\*\*  
## STARS2 2.427311 0.038459 63.113 < 2e-16 \*\*\*  
## STARS3 2.969229 0.044380 66.904 < 2e-16 \*\*\*  
## STARS4 3.664751 0.069763 52.532 < 2e-16 \*\*\*  
## LabelAppeal-1 0.291589 0.075918 3.841 0.000123 \*\*\*  
## LabelAppeal0 0.739897 0.073991 10.000 < 2e-16 \*\*\*  
## LabelAppeal1 1.208997 0.077214 15.658 < 2e-16 \*\*\*  
## LabelAppeal2 1.907820 0.102147 18.677 < 2e-16 \*\*\*  
## AcidIndex -0.205091 0.010768 -19.047 < 2e-16 \*\*\*  
## Alcohol 0.015113 0.004011 3.768 0.000166 \*\*\*  
## CitricAcid 0.053948 0.024857 2.170 0.030007 \*   
## VolatileAcidity -0.192020 0.026625 -7.212 5.96e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.306 on 8945 degrees of freedom  
## Multiple R-squared: 0.5451, Adjusted R-squared: 0.5445   
## F-statistic: 893.1 on 12 and 8945 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Model 1: TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol + CitricAcid  
## Model 2: TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol + CitricAcid +   
## VolatileAcidity  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 8946 15340   
## 2 8945 15251 1 88.685 52.014 5.957e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The adjusted R^2 value increase and the F and p-value support rejecting the null hypothesis that the coefficient of Volatil Acidity is zero. Finally we will consider to add the level of Sulphites to our model. Free Sulfur Dioxide prevents oxidation in wine and helps preserve it.

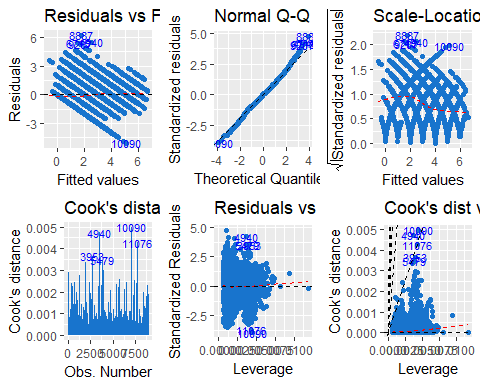
##   
## Call:  
## lm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol +   
## CitricAcid + VolatileAcidity + FreeSulfurDioxide, data = wine\_dev\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1029 -0.8542 0.0247 0.8325 6.2181   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.0287399 0.1230177 16.491 < 2e-16 \*\*\*  
## STARS1 1.3850861 0.0392510 35.288 < 2e-16 \*\*\*  
## STARS2 2.4226283 0.0384611 62.989 < 2e-16 \*\*\*  
## STARS3 2.9673876 0.0443572 66.898 < 2e-16 \*\*\*  
## STARS4 3.6639049 0.0697217 52.550 < 2e-16 \*\*\*  
## LabelAppeal-1 0.2925143 0.0758733 3.855 0.000116 \*\*\*  
## LabelAppeal0 0.7371326 0.0739516 9.968 < 2e-16 \*\*\*  
## LabelAppeal1 1.2069552 0.0771701 15.640 < 2e-16 \*\*\*  
## LabelAppeal2 1.9009281 0.1021061 18.617 < 2e-16 \*\*\*  
## AcidIndex -0.2033942 0.0107726 -18.881 < 2e-16 \*\*\*  
## Alcohol 0.0154285 0.0040094 3.848 0.000120 \*\*\*  
## CitricAcid 0.0536707 0.0248424 2.160 0.030764 \*   
## VolatileAcidity -0.1904647 0.0266127 -7.157 8.9e-13 \*\*\*  
## FreeSulfurDioxide 0.0005036 0.0001476 3.411 0.000650 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.305 on 8944 degrees of freedom  
## Multiple R-squared: 0.5457, Adjusted R-squared: 0.545   
## F-statistic: 826.3 on 13 and 8944 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Model 1: TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol + CitricAcid +   
## VolatileAcidity  
## Model 2: TARGET ~ STARS + LabelAppeal + AcidIndex + Alcohol + CitricAcid +   
## VolatileAcidity + FreeSulfurDioxide  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 8945 15251   
## 2 8944 15231 1 19.814 11.635 0.0006502 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The Adjusted R^2 increase and based on the F-value and p-value, we will reject the null hypothesis that the coefficient for Free Sulfur Dioxide is zero.

We will consider this our first multiple linear model. We would not look at the residual plots to evaluate how good fit is the model.

#### Plots Diagnostics



***Residual Plot***  
From this plot we observed that we have residuals shows some pattern It appears that some relationship have not been accounted for in the model.

***Normal Q-Q Plot***  
From this plot we observed that distribution of the residuals is fairly normal.

***Scale-Location Plot***  
Fom this plot we observed that the residuals are not quite equally spread along the range of predictors. Thus, the variance is not equal.

***Residuals vs Leverage points***  
From the bottom 3 plots, we observed that we have some influencial point beyond Cook’s distances.

#### Model Interpretation

We will now look at the coefficients of our model;

|  |  |
| --- | --- |
|  | Values |
| (Intercept) | 2.0287399 |
| STARS1 | 1.3850861 |
| STARS2 | 2.4226283 |
| STARS3 | 2.9673876 |
| STARS4 | 3.6639049 |
| LabelAppeal-1 | 0.2925143 |
| LabelAppeal0 | 0.7371326 |
| LabelAppeal1 | 1.2069552 |
| LabelAppeal2 | 1.9009281 |
| AcidIndex | -0.2033942 |
| Alcohol | 0.0154285 |
| CitricAcid | 0.0536707 |
| VolatileAcidity | -0.1904647 |
| FreeSulfurDioxide | 0.0005036 |

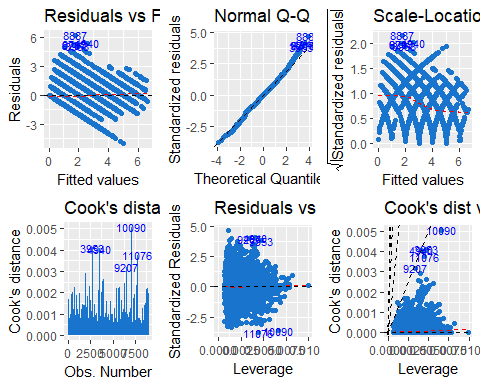
* **STARS:** Wines with better rating (higher stars) will most likely sell higher number of cases than wines with lower rating. This is as we would expect.
* **LabelAppearl:** Wines with higher label appeal will most likely sell higher number of cases than wines with lower label appear. Although the label design has no bearing on the quality of a wine this result is not surprising since some customers will select a wine based on the label.
* **AcidIndex**: Wines with higher Acid Index will most likely sell lower number of cases than wines with a lower Acid Index.
* **Alcool**: Wines with higher Alcohol content will most likely sell higher number of cases than wines with less Alcohol.
* **CitricAcid**: Wines with higher levels of Citric Acid will sell higher number of cases than wines with lower levels.
* **VolatileAcidity**: Wines with higher levels of VolatileAcidity will sell lower number of cases than wines with lower levels.
* **FreeSulfurDioxide**: Wines with higher levels of Free Sulfur Dioxide will sell higher number of cases than wines with lower levels.

### Model b - Backward Multiple Linear Model using AIC

We will now use the Backward method leveraging AIC to find another multiple linear model.

##   
## Call:  
## lm(formula = TARGET ~ STARS + LabelAppeal + Density + Alcohol +   
## FixedAcidity + VolatileAcidity + Sulphates + FreeSulfurDioxide +   
## TotalSulfurDioxide + Chlorides, data = wine\_dev\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.9928 -0.8689 0.0322 0.8442 6.2107   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.2878216 0.5320528 2.420 0.015520 \*   
## STARS1 1.4397496 0.0397466 36.223 < 2e-16 \*\*\*  
## STARS2 2.5180611 0.0386628 65.129 < 2e-16 \*\*\*  
## STARS3 3.0782484 0.0445598 69.081 < 2e-16 \*\*\*  
## STARS4 3.7862768 0.0704435 53.749 < 2e-16 \*\*\*  
## LabelAppeal-1 0.2716755 0.0771018 3.524 0.000428 \*\*\*  
## LabelAppeal0 0.7002731 0.0751300 9.321 < 2e-16 \*\*\*  
## LabelAppeal1 1.1498921 0.0783653 14.673 < 2e-16 \*\*\*  
## LabelAppeal2 1.8335980 0.1036810 17.685 < 2e-16 \*\*\*  
## Density -0.7457409 0.5259157 -1.418 0.156230   
## Alcohol 0.0179070 0.0040739 4.396 1.12e-05 \*\*\*  
## FixedAcidity -0.0100877 0.0030549 -3.302 0.000963 \*\*\*  
## VolatileAcidity -0.2061453 0.0270336 -7.626 2.68e-14 \*\*\*  
## Sulphates -0.0930770 0.0230476 -4.038 5.43e-05 \*\*\*  
## FreeSulfurDioxide 0.0006018 0.0001499 4.015 5.99e-05 \*\*\*  
## TotalSulfurDioxide 0.0004600 0.0000938 4.904 9.57e-07 \*\*\*  
## Chlorides -0.2686644 0.0684182 -3.927 8.67e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.326 on 8941 degrees of freedom  
## Multiple R-squared: 0.5314, Adjusted R-squared: 0.5305   
## F-statistic: 633.6 on 16 and 8941 DF, p-value: < 2.2e-16

#### Plots Diagnostics



***Residual Plot***  
From this plot we observed that we have residuals shows some pattern It appears that some relationship have not been accounted for in the model.

***Normal Q-Q Plot***  
From this plot we observed that distribution of the residuals is fairly normal.

***Scale-Location Plot***  
Fom this plot we observed that the residuals are not quite equally spread along the range of predictors. Thus, the variance is not equal.

***Residuals vs Leverage points***  
From the bottom 3 plots, we observed that we have some influencial point beyond Cook’s distances.

|  |  |
| --- | --- |
|  | Values |
| (Intercept) | 1.2878216 |
| STARS1 | 1.4397496 |
| STARS2 | 2.5180611 |
| STARS3 | 3.0782484 |
| STARS4 | 3.7862768 |
| LabelAppeal-1 | 0.2716755 |
| LabelAppeal0 | 0.7002731 |
| LabelAppeal1 | 1.1498921 |
| LabelAppeal2 | 1.8335980 |
| Density | -0.7457409 |
| Alcohol | 0.0179070 |
| FixedAcidity | -0.0100877 |
| VolatileAcidity | -0.2061453 |
| Sulphates | -0.0930770 |
| FreeSulfurDioxide | 0.0006018 |
| TotalSulfurDioxide | 0.0004600 |
| Chlorides | -0.2686644 |

* **STARS:** Wines with better rating (higher stars) will most likely sell higher number of cases than wines with lower rating. This is as we would expect.
* **LabelAppearl:** Wines with higher label appeal will most likely sell higher number of cases than wines with lower label appear. Although the label design has no bearing on the quality of a wine this result is not surprising since some customers will select a wine based on the label.
* **Density**: Wines with lower density will sell higher number of cases than wines with higher density.
* **Alcohol**: Wines with higher Alcohol content will most likely sell higher number of cases than wines with less Alcohol.
* **FixedAcidity**: Wines with higher levels of FixedAcidity will sell lower number of cases than wines with lower levels.
* **VolatileAcidity**: Wines with higher levels of VolatileAcidity will sell lower number of cases than wines with lower levels.
* **Sulphates**: Wines with higher Sulphates levels will most likely sell lower number of cases than wines with a lower Sulphates levels.
* **FreeSulfurDioxide**: Wines with higher levels of Free Sulfur Dioxide will sell higher number of cases than wines with lower levels.
* **TotalSulfurDioxide**: Wines with higher levels of Total Sulfur Dioxide will sell higher number of cases than wines with lower levels.
* **Chlorides**: Wines with higher levels of Chlorides will sell lower number of cases than wines with lower levels.

#### Multilinear Models Comparison

To compare our 2 multilinear models, we will use Adjusted R^2, AIC, and RMSE on prediction for our dev-Evaluation data set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | RMSE | SSE | AdjRSquared | ACI | BIC |
| LM.Modela | 1.33457070548686 | 6834 | 0.544994605994852 | 30206.715 | 30313.22 |
| LM.Modelb | 1.36278394788422 | 7126 | 0.530513247956123 | 30490.371 | 30618.177 |

From the table of measures, we would select our linear model a although we may have prediction to be -1. As we indicated prior, there is no guarantee that negative value will not be predicted.

# 4.Model Selection and Evaluation

## Model Selection

We will consider the AIC measure to determine which model to select of the 3 we have selected from the poisson models, negative binomial model, and multiple linear regression models.

|  |  |  |
| --- | --- | --- |
|  | Model | ACI |
| result1 | Poisson | 32675.63432312 |
| result2 | Neg Binomial | 32676.1473442188 |
| result3 | Multi Linear | 30206.715 |

From the AIC measures alone we would surprisingly select the multilinear regression model. However, we have observed that we cannot guarantee that the prediction result will be a number >=0, since a negative number has no meaning in this context, we would select the quazi poisson model.

## Evaluation

We first need to perform the same data cleansing steps on the evaluation data set.

# Conclusion

This was a challenging problem due to the state of the data. The high level of corrupted observations in the data set made for difficult modeling. Surprisingly, it appears that the “manual” multiple linear model is a better model however since this model could and would predict potentially negative value, we felt it was not possible to select it. The next best one from a quazi poisson model.

# References:

***Wines Component Analysis***  
<http://winefolly.com/tutorial/the-lightest-to-the-strongest-wine/> <http://winefolly.com/review/understanding-acidity-in-wine/> <https://www.wineperspective.com/wine-acidity/> <http://waterhouse.ucdavis.edu/> <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-20612015000100095> <http://morewinemaking.com/public/pdf/so2.pdf> <http://winefolly.com/tutorial/alcohol-content-in-wine/> <https://winemakermag.com/501-measuring-residual-sugar-techniques>

***Technical References*** <https://www.youtube.com/watch?v=u8XvfhBdbMw>  
<http://data.library.virginia.edu/diagnostic-plots/>  
<http://www.stat.wisc.edu/~ane/st572/notes/lec05.pdf>

Poisson:  
#<https://www.youtube.com/watch?v=mgIyPRhrVTY>