Wine Evaluation

Daniel Craig, John Cruz, Shaya Engelman, Noori Selina, Gavriel Steinmetz-Silber

2024-04-23

Introduction

A data set containing information on approximately 12,000 commercially available wines and their variables mostly related to chemical properties is analyzed for impact on sales and used to predict on sales to give accurate forecasts for manufacturing. 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.

Required Libraries

```
library(tidyverse)
library(janitor)
library(knitr)
library(kableExtra)
library(latex2exp)
library(psych)
library(scales)
library(stringr)
library(ggcorrplot)
library(ggmice)
library(caret)
library(mice)
library(bestNormalize)
library(e1071)
library(diptest)
library(MASS)
library(performance)
```

Data Exploration

To-Do List: 1. Check for typo's - No Typo's, all data is int 2. Check for missing 3. Show distributions 4. Determine Categorical/Continuous

Data Summary

A table below expands on the variables included in analysis with comments from domain experts on expected effects.

Table 1: 5 Number Summary

	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
TARGET	3.0290739	1.9263682	3.00000	3.0538244	1.4826000	0.00000	8.00000	8	-0.326301039267588	**-0.877245713363431**	0.0170302
FixedAcidity	7.0757171	6.3176435	6.90000	7.0736739	3.2617200	-18.10000	34.40000	**52.5**	-0.0225859613642668	1.67499867419602	0.0558515
VolatileAcidity	0.3241039	0.7840142	0.28000	0.3243890	0.4299540	-2.79000	3.68000	6.47	0.0203799652905512	1.83221063847995	0.0069311
CitricAcid	0.3084127	0.8620798	0.31000	0.3102520	0.4151280	-3.24000	3.86000	7.1	-0.0503070404378392	1.837940071767	0.0076213
ResidualSugar	5.4187331	33.7493790	3.90000	5.5800410	15.7155600	-127.80000	141.15000	**268.95**	-0.0531229052496501	1.88469166771449	0.3058158
Chlorides	0.0548225	0.3184673	0.04600	0.0540159	0.1349166	-1.17100	1.35100	2.522	0.0304271748147184	1.78860442940177	0.0028884
FreeSulfurDioxide	30.8455713	148.7145577	30.00000	30.9334877	56.3388000	-555.00000	623.00000	**1178**	0.0063930101150823	1.8364966248458	1.3492769
TotalSulfurDioxide	120.7142326	231.9132105	123.00000	120.8895367	134.9166000	-823.00000	1057.00000	**1880**	-0.00717935086303868	1.67466647637973	2.1071703
Density	0.9942027	0.0265376	0.99449	0.9942130	0.0093552	0.88809	1.09924	0.21115	-0.0186937638734045	1.89995920703906	0.0002346
pH	3.2076282	0.6796871	3.20000	3.2055706	0.3854760	0.48000	6.13000	5.65	0.0442880137342456	1.64626805925174	0.0061038
Sulphates	0.5271118	0.9321293	0.50000	0.5271453	0.4447800	-3.13000	4.24000	7.37	0.00591189546131464	1.75256551530022	0.0086602
Alcohol	10.4892363	3.7278190	10.40000	10.5018255	2.3721600	-4.70000	26.50000	31.2	-0.0307158360932212	1.53949494703394	0.0338306
LabelAppeal	-0.0090660	0.8910892	0.00000	-0.0099639	1.4826000	-2.00000	2.00000	4	0.00842945702449739	**-0.262291551256824**	0.0078777
AcidIndex	7.7727237	1.3239264	8.00000	7.6431572	1.4826000	4.00000	17.00000	13	**1.64849594529687**	**5.19009248111377**	0.0117043
STARS	2.0417550	0.9025400	2.00000	1.9711258	1.4826000	1.00000	4.00000	3	0.447235291548031	**-0.692534319319664**	0.0092912

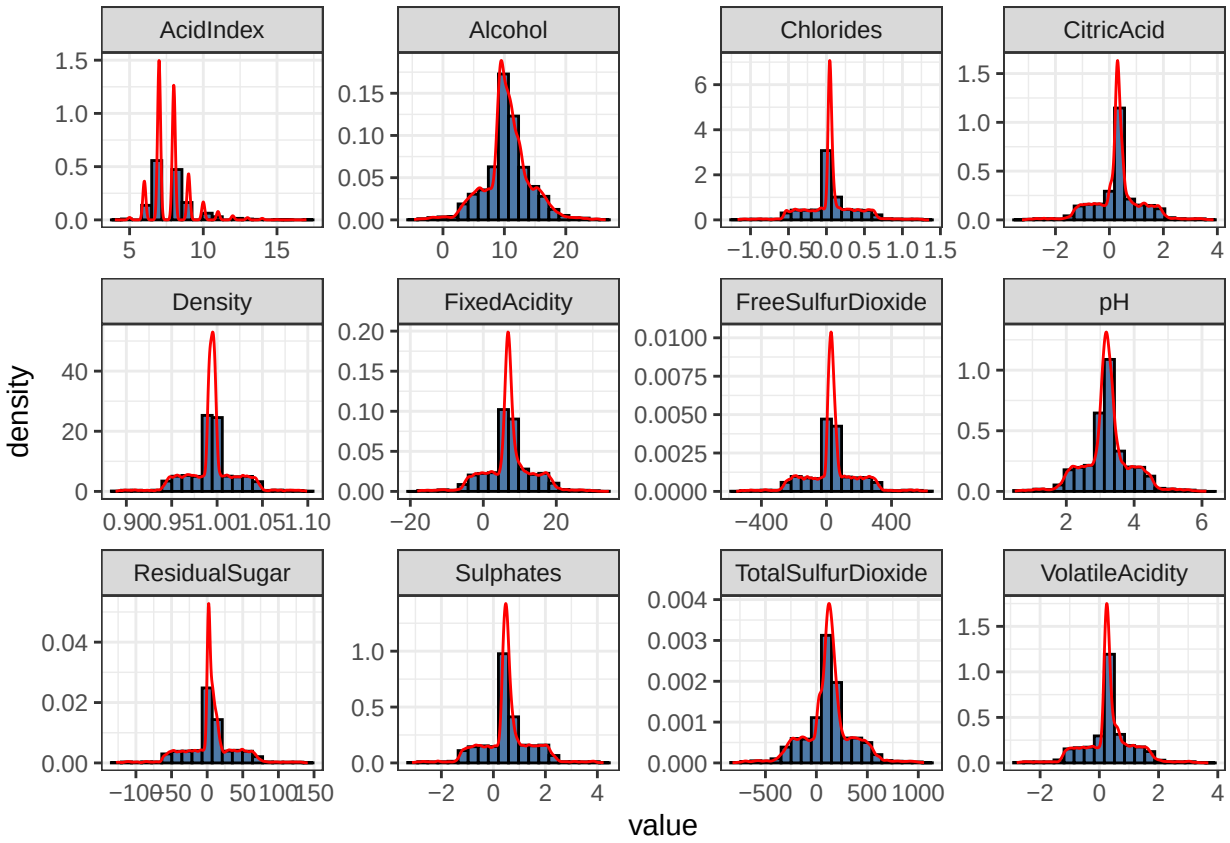
```
##
## | **VARIABLE**          | **DEFINITION**                          | **THEORETICAL EFFECT**
## |:----------------------|:----------------------------------------|:----------------------
## | `INDEX`               | ID Variable                             | None
## | `TARGET`              | Cases Purchased                         | None
## | `AcidIndex`           | Total Acidity Test                      | Unknown
## | `Alcohol`             | Alcohol Content                         | Higher alcohol, higher sales
## | `Chlorides`           | Chloride Content                        | Low levels, higher quality
## | `CitricAcid`          | Citric Acid Content                     | Suggests freshness, impacts sales
## | `Density`             | Wine Density                            | Higher suggests richer wines
## | `FixedAcidity`        | Fixed Acidity                           | Affects taste
## | `FreeSulfurDioxide`   | Free SO2 Content                        | Preserves freshness, impacts sales
## | `LabelAppeal`         | Label Appeal                            | More appealing, enhances sales
## | `ResidualSugar`       | Sugar Content                           | Sweetness impacts sales
## | `STARS`               | Expert Rating                           | Higher ratings, higher sales
## | `Sulphates`           | Sulfate Content                         | Affects preservation and taste
## | `TotalSulfurDioxide`  | Total SO2                               | Affects longevity and freshness
## | `VolatileAcidity`     | Volatile Acid                           | Lower suggests higher quality
## | `pH`                  | pH                                      | Optimal pH impacts taste and stabil
```

```
## \newpage
```

A quick look at the variables 5 number summary reveals that several variables have large ranges which when relating to their mean may suggest significantly different scales between variables, a high amount of skew, bi-modal distributions, or outliers. FixedAcidity, ResidualSugar, FreeSulfurDioxide, and TotalSulfurDioxide have fairly extreme ranges in comparison to their means. Variables with Kurtosis greater than 4 will have observations distributed into heavy or long tails and may suggest numerous outliers, less than 2 suggest distributions centered around their mean with short or thin tails. Many of the variables are just below 2 suggesting many will have sharp peaks around the mean. Only AcidIndex shows as a non-ordinal or discrete distribution with extreme values of kurtosis, suggesting it will contain many outliers.

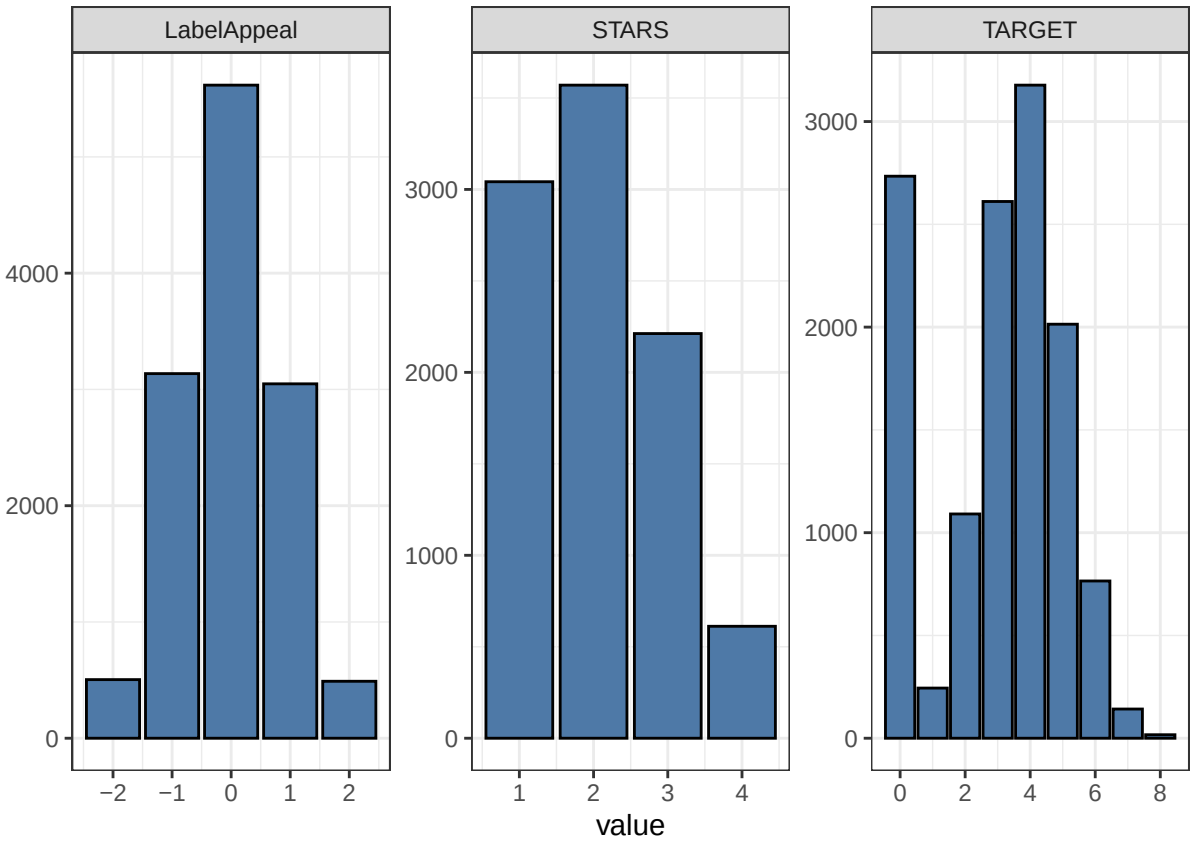
Histograms

As kurtosis foreshadowed, many of the distributions have sharp peaks at the mean with only the AcidIndex showing a bi-modal distribution. With the sharp centers around the peaks in the histograms, a high number of outliers may present themselves.



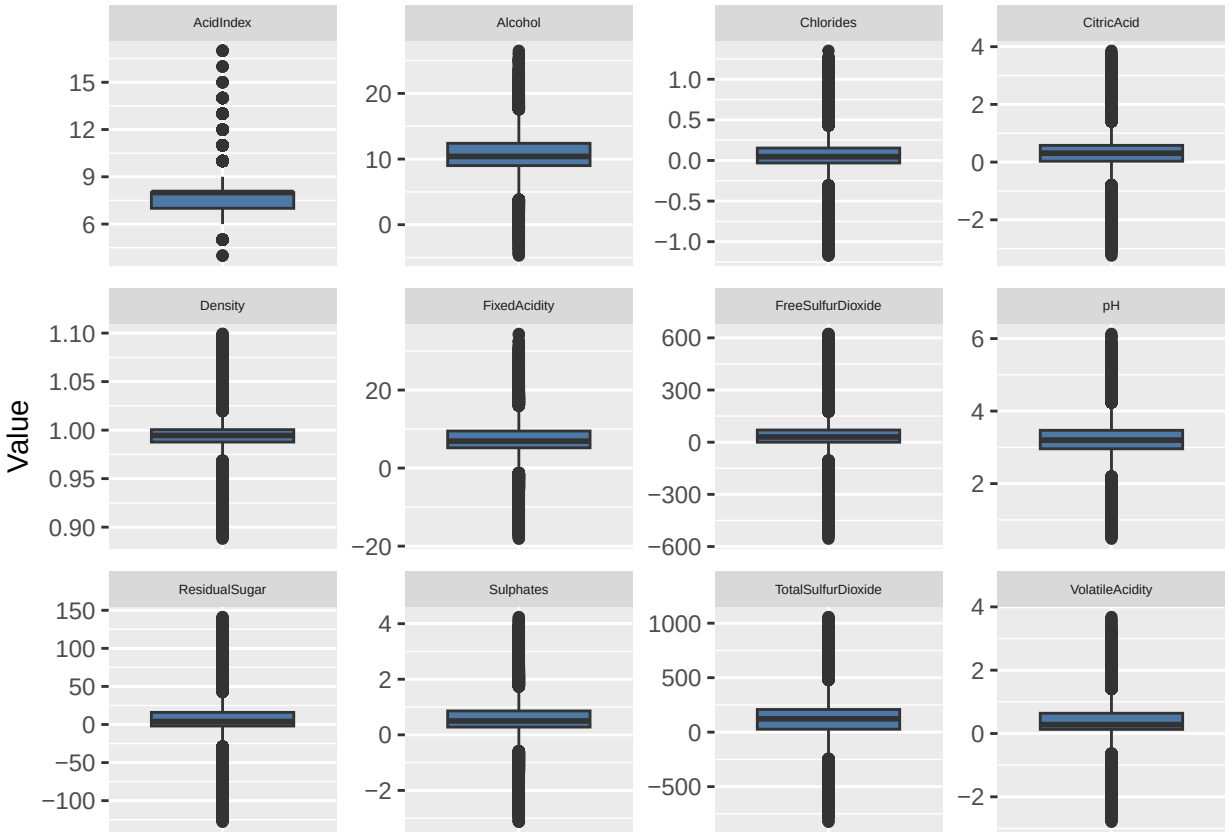
Bar Plots

There is a relatively normal distribution to LabelAppeal, but both STARS and TARGET tend to favor their lower values suggesting it's quite difficult to gain either a critic's praise or a significant amount of cases sold.



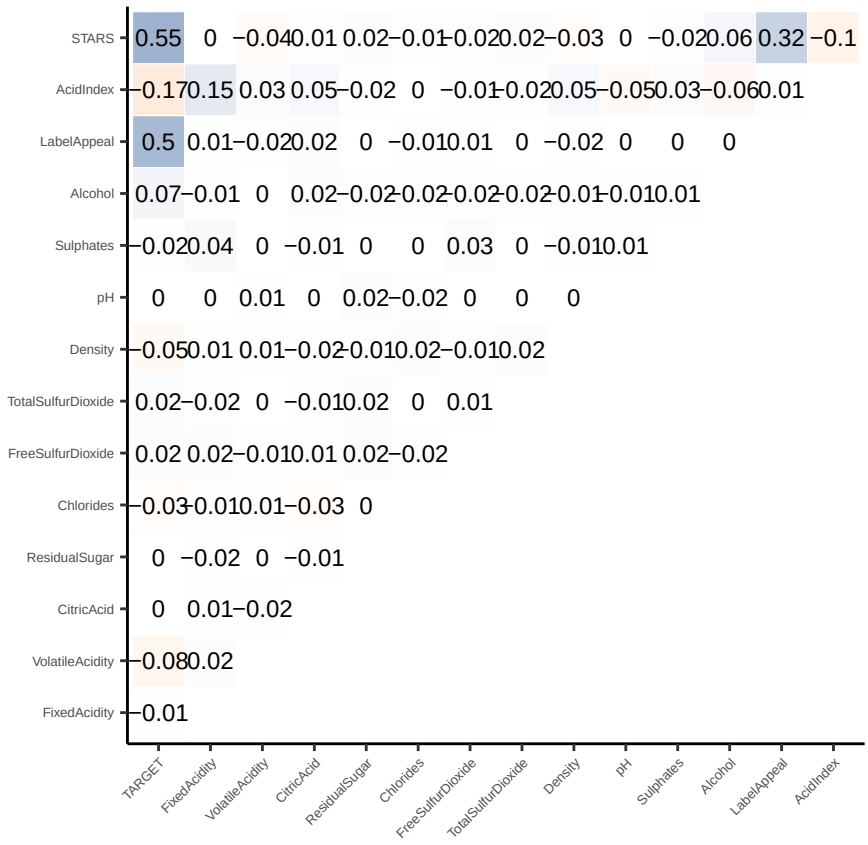
Box Plots

Boxplots reveal a significant number of residuals in all of the variables.



Correlation Matrix

The correlation matrix reveals a moderate relationship between STARS and LabelAppeal with Target. Although both STARS and LabelAppeal seem to be somewhat correlated to each other suggesting potential colinearity. The AcidIndex, being a propietary method that aggregates across Acid metrics, does show some relationship with FixedAcidity but is relatively minor.



Missing Values

While missing values may be indicative of the target, the STARS variable is missing 26% of its values. Determining the relationship it has to cases sold may be useful before removing it from the dataset. To view the relationship of the "missing" STARS ratings, NA's have been replaced with a category of "Unrated" and the bar plots are shown again. Chlorides, FreeSulfurDioxide, Alcohol, and TotalSulfurDioxide are missing around 5% or about 600 values. Sulphates is missing about 10% of its values and about 1200 values.

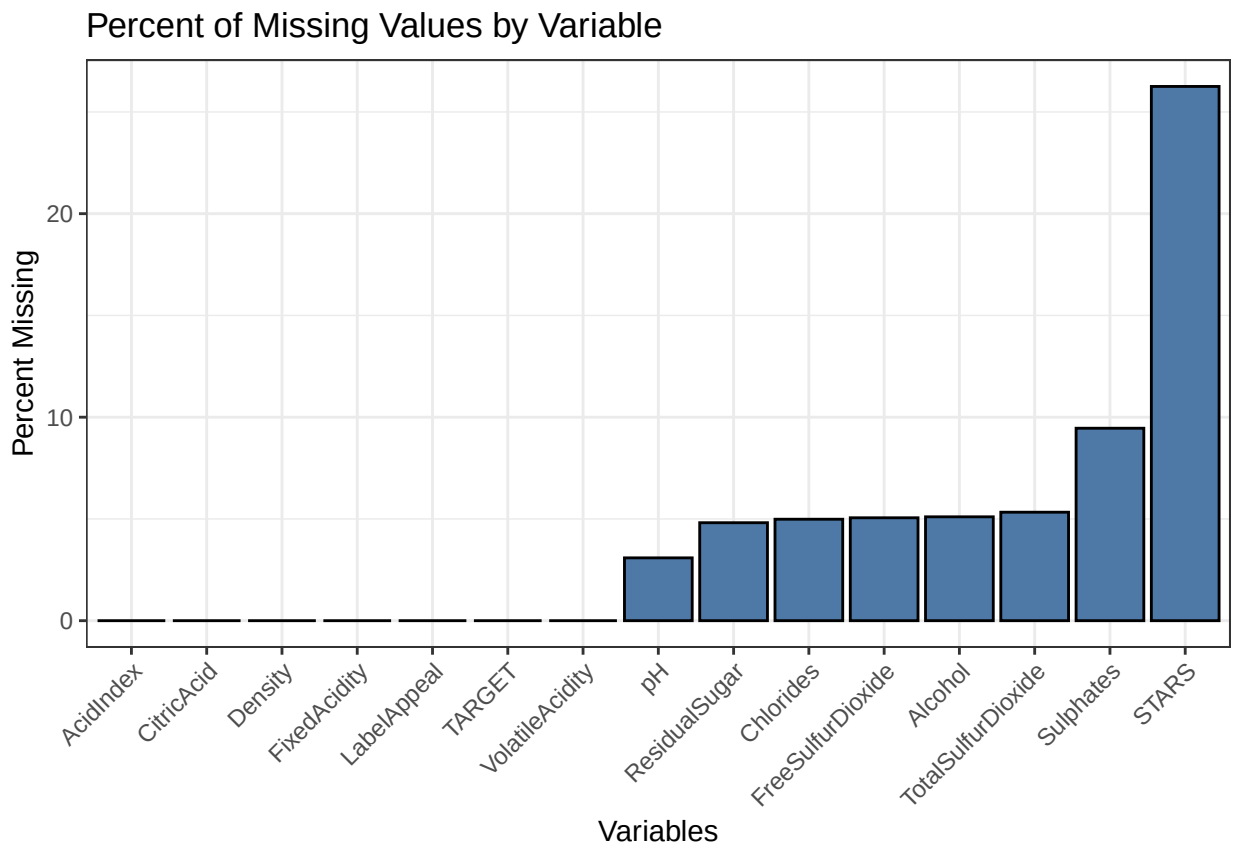
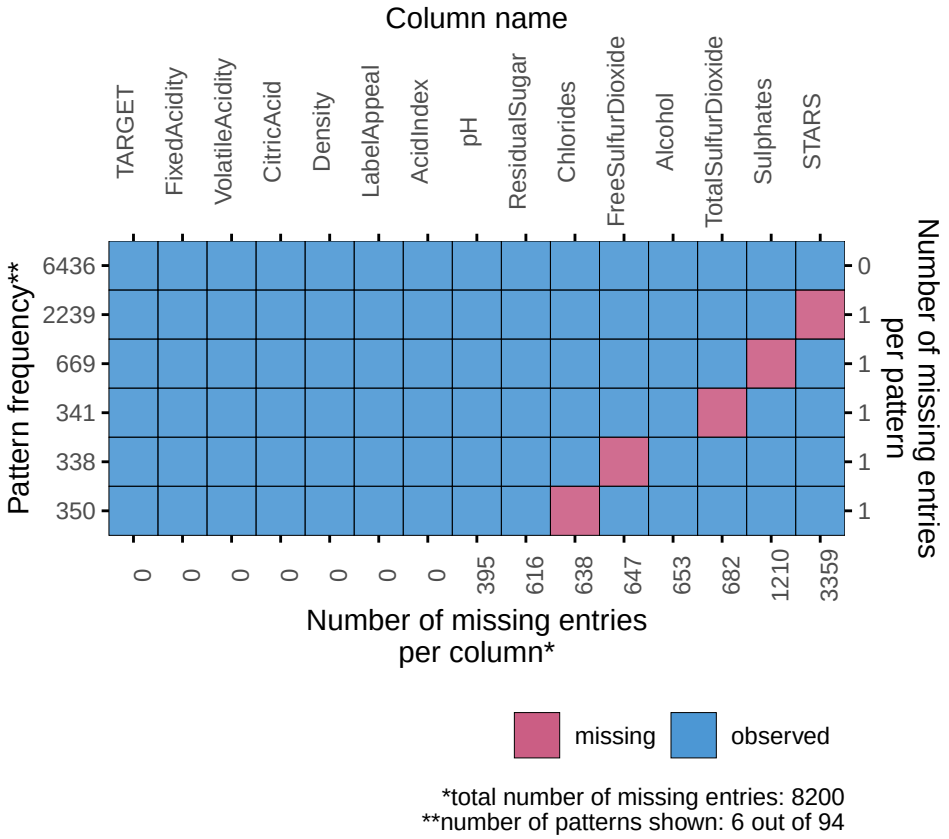


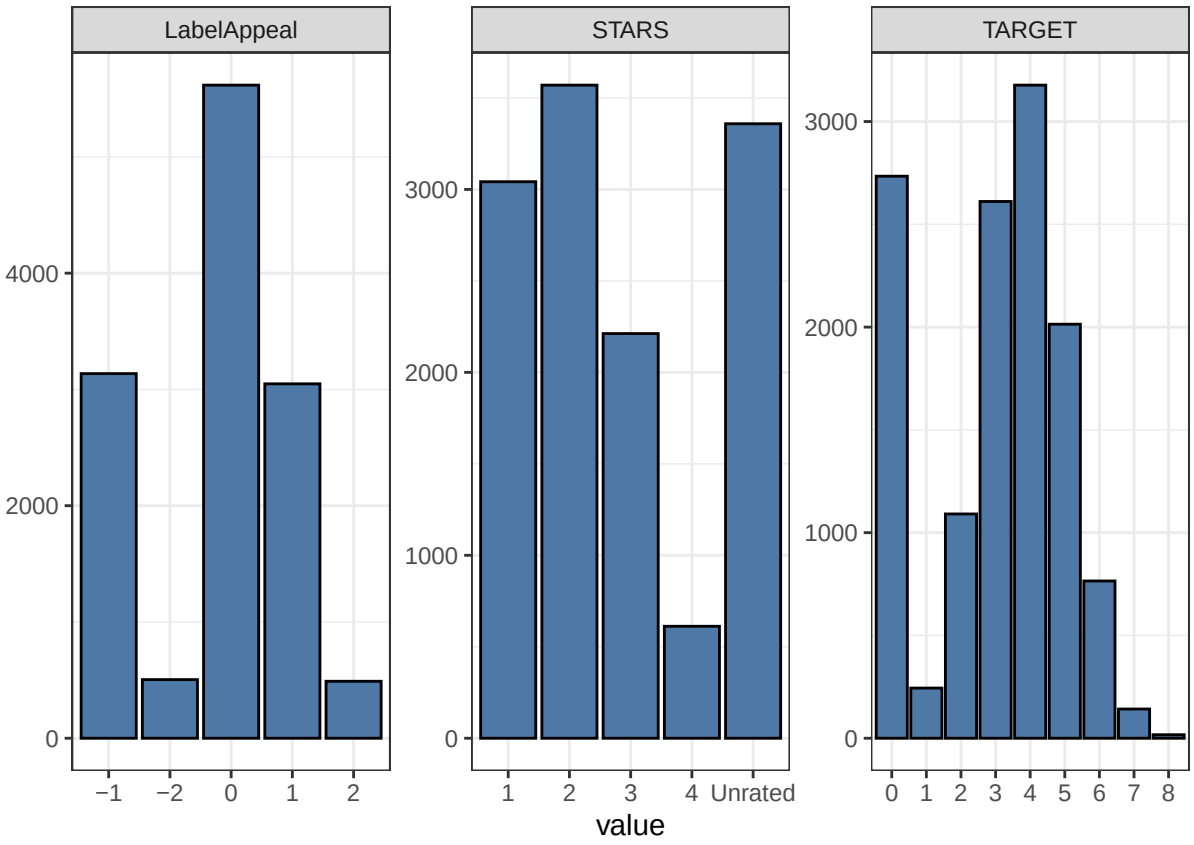
Table 2: Missing Values Count

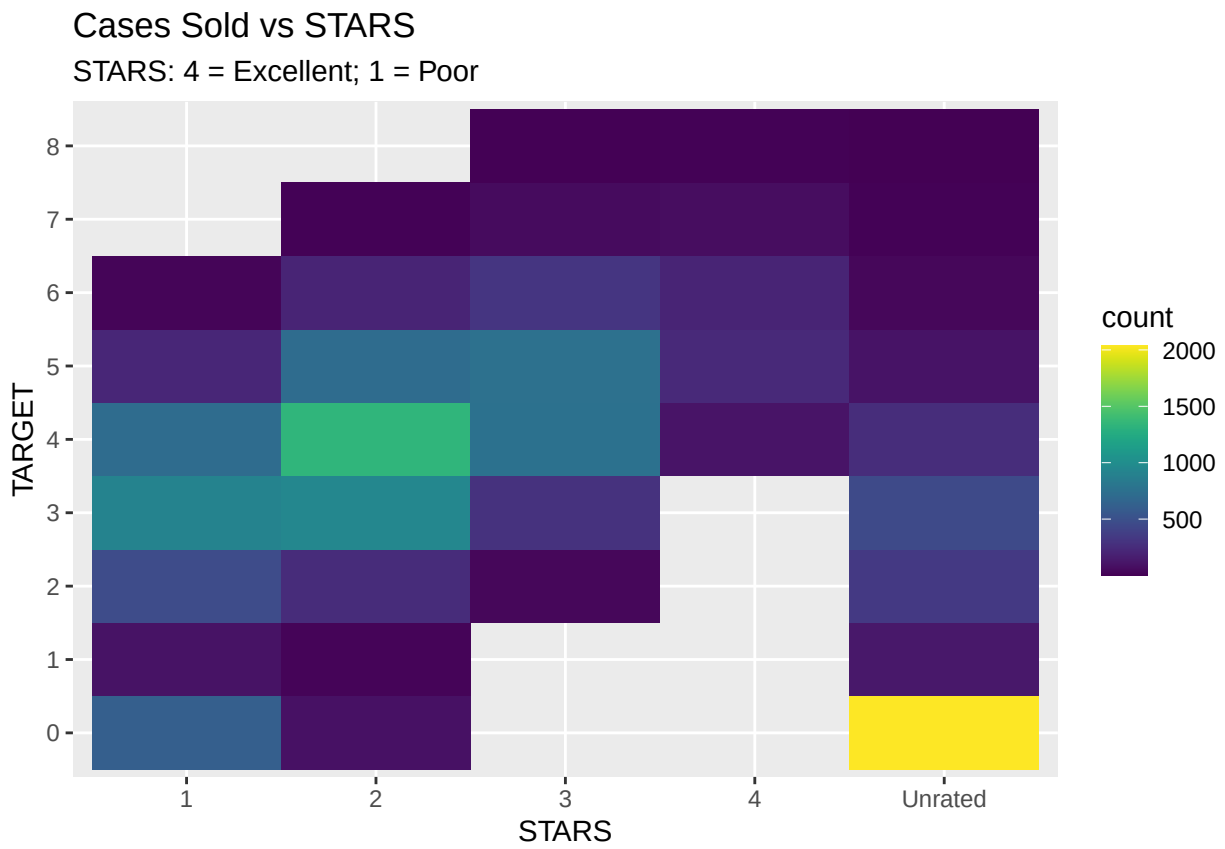
ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	pH	Sulphates	Alcohol	STARS
616	638	647	682	395	1210	653	3359

```
plot_pattern(train, square = TRUE, rotate = TRUE, npat = 6)
```

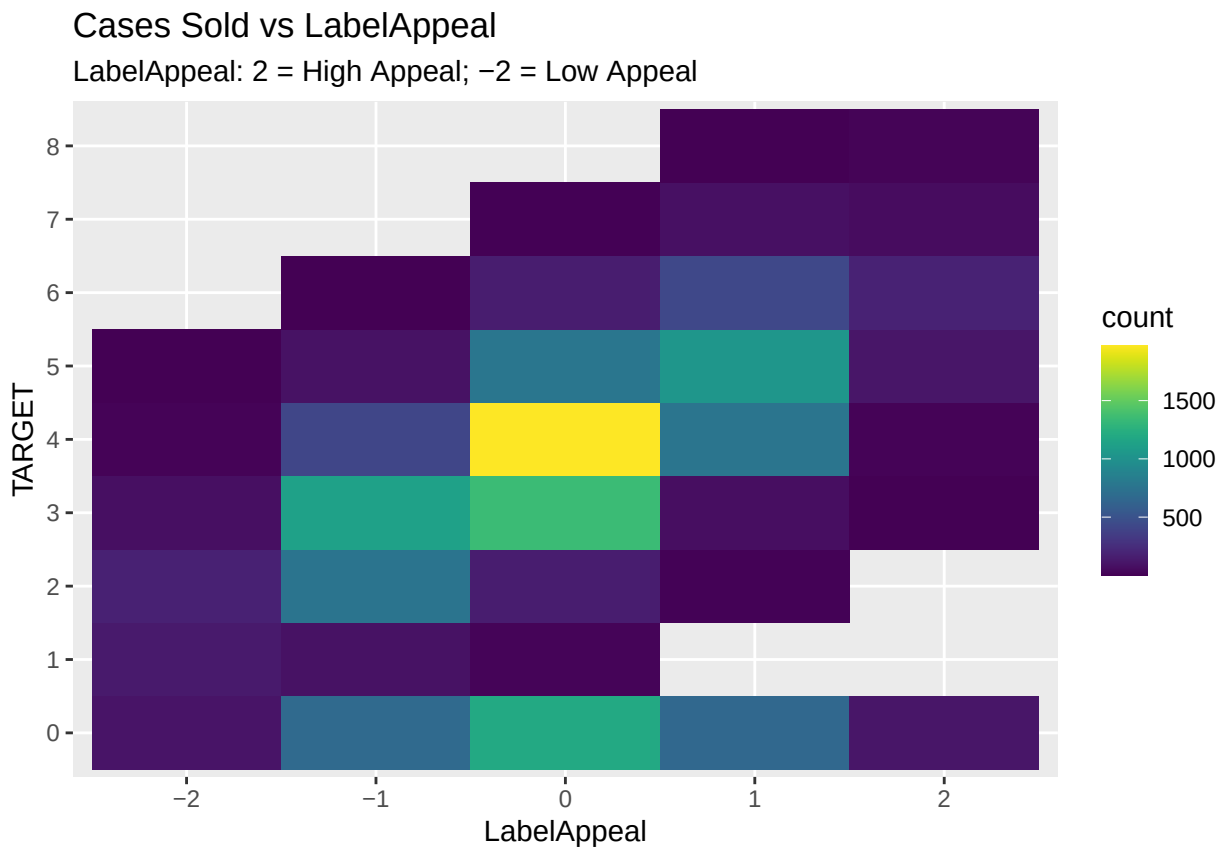


While unrated wines typically aren't purchased, there are some that sell about 3 cases. This might suggest that non-rated wines are not submitted for critic's appraisal and should be used as a feature in the modeling. This plot also reveals a heavy preference for 2 star wines.


LabelAppeal
STARS
TARGET
4000
2000
0
3000
2000
1000
0
3000
2000
1000
0
−1
−2
0
1
2
1
2
3
4
Unrated
0
1
2
3
4
5
6
7
8
value



The majority of 0 value appeals center on 4 cases sold and does tend to show a linear relationship between the two.



Data Preparation

Now that we have explored our data set, we can move on to data preparation to prep out data for modeling and analysis.

Data Wrangling

To do list: 1. Split the data into training and testing sets 2. Impute missing values 3. Normalize the data 4. Deal with outliers 5. One hot encode the categorical variables

The data has already been partially cleaned with the removal of the INDEX variable. The missing values in STARS were replaced with "Unrated" to indicate non-rated wines.

Data Imputation

Before we can impute missing values, we perform the train-test split to avoid data leakage:

Now, we can impute the missing values in the training and testing data sets. We will use the MICE package to impute the missing values. In the imputation process, we will exclude the TARGET variable from the predictors, as the target variable should not be used to predict the missing values of the predictors. All imputation will be done for all three of the training, testing, and evaluation data sets. In order to make the dataframes match we drop the INDEX column from the evaluation data set.

Now that we've imputed the missing values, we can compare the summary statistics of the original data and the imputed data. The summary statistics are calculated for the following variables: Chlorides, FreeSulfurDioxide, Alcohol, TotalSulfurDioxide, pH, and Sulphates. The summary statistics are calculated for the full training data set, the training data set after imputation, and the testing data set after imputation. The summary statistics are calculated for the minimum, 1st quartile, median, mean, 3rd quartile, and maximum values of the variables. The summary statistics are then compared across the three data sets to see how the imputation process has affected the data.

Table 3: Summary Statistics Comparison Across Datasets

Variable_Stat	Dataset (Pre-Imputations)	Train Imputed	Test Imputed
Chlorides_min	-1.17000000	-1.17100000	-1.17100000
Chlorides_q1	-0.03900000	-0.03850000	-0.00700000
Chlorides_median	0.04600000	0.04600000	0.04700000
Chlorides_mean	0.05075939	0.05138424	0.06377821
Chlorides_q3	0.14325000	0.14600000	0.17100000
Chlorides_max	1.35100000	1.35100000	1.26000000
FreeSulfurDioxide_min	-546.00000000	-546.00000000	-555.00000000
FreeSulfurDioxide_q1	1.00000000	0.00000000	-2.00000000
FreeSulfurDioxide_median	30.00000000	30.00000000	31.00000000
FreeSulfurDioxide_mean	30.88763649	30.69956468	31.09989572
FreeSulfurDioxide_q3	69.00000000	69.00000000	73.00000000
FreeSulfurDioxide_max	623.00000000	623.00000000	617.00000000
Alcohol_min	-4.70000000	-4.70000000	-4.40000000
Alcohol_q1	9.00000000	9.00000000	9.00000000
Alcohol_median	10.40000000	10.40000000	10.40000000
Alcohol_mean	10.46157651	10.46814317	10.56149635
Alcohol_q3	12.30000000	12.30000000	12.40000000
Alcohol_max	26.10000000	26.10000000	26.50000000
TotalSulfurDioxide_min	-816.00000000	-816.00000000	-823.00000000
TotalSulfurDioxide_q1	26.00000000	26.00000000	29.00000000
TotalSulfurDioxide_median	123.00000000	123.00000000	124.00000000
TotalSulfurDioxide_mean	121.37894489	121.52347360	119.30724713
TotalSulfurDioxide_q3	209.00000000	209.00000000	205.00000000
TotalSulfurDioxide_max	1057.00000000	1057.00000000	1041.00000000
pH_min	0.48000000	0.48000000	0.53000000
pH_q1	2.95000000	2.96000000	2.96000000
pH_median	3.20000000	3.20000000	3.20000000
pH_mean	3.20488704	3.20552093	3.21393691
pH_q3	3.47000000	3.46000000	3.48000000
pH_max	6.02000000	6.05000000	6.21000000
Sulphates_min	-3.13000000	-3.13000000	-3.10000000
Sulphates_q1	0.29000000	0.30000000	0.25000000
Sulphates_median	0.50000000	0.50000000	0.50000000
Sulphates_mean	0.53412426	0.53572050	0.51387226
Sulphates_q3	0.86000000	0.87000000	0.85000000
Sulphates_max	4.24000000	4.24000000	4.21000000

The above table is very encouraging. The summary statistics for the variables with missing data did not seem to change much after the imputation. Most of the discrepancies appear in the test data set, this plausibly due to the smaller sample size.

Transformations

Table 4: Best Transformations

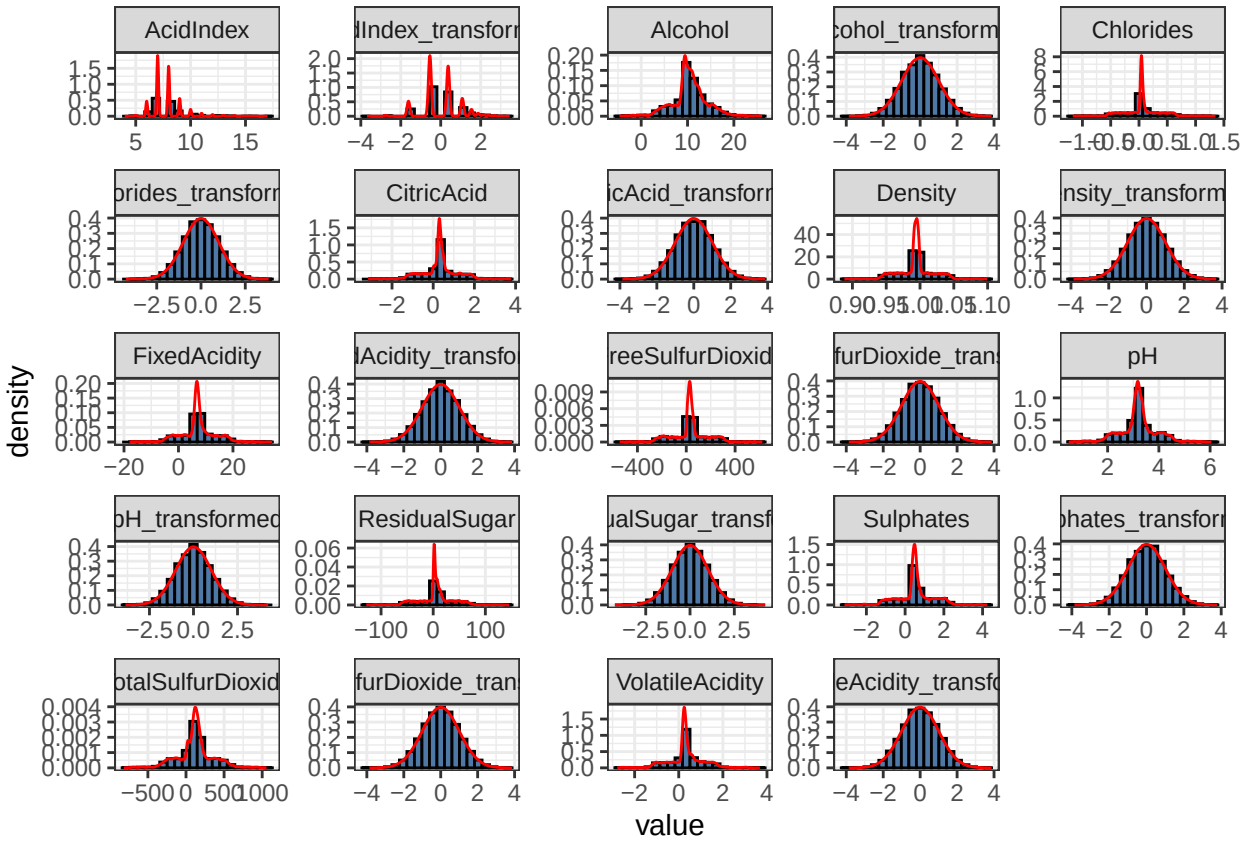
Variable	Transformation
FixedAcidity	orderNorm
VolatileAcidity	orderNorm
CitricAcid	orderNorm
ResidualSugar	orderNorm
Chlorides	orderNorm
FreeSulfurDioxide	orderNorm
TotalSulfurDioxide	orderNorm
Density	orderNorm
pH	orderNorm
Sulphates	orderNorm
Alcohol	orderNorm
AcidIndex	orderNorm

Again, we must be quite careful to avoid data leakage, calculating the parameters for these transformations using only the training set, and then applying the transformations to our other sets using the same parameters.

Table 5: Pre and Post Transformation Skewness Comparison

Variable	Pre-Transformation Skew	Post-Transformation Skew
FixedAcidity	-0.043	0.000
VolatileAcidity	-0.009	0.000
CitricAcid	-0.043	0.000
ResidualSugar	-0.085	0.000
Chlorides	0.026	0.000
FreeSulfurDioxide	0.037	0.000
TotalSulfurDioxide	0.014	0.000
Density	0.005	0.000
pH	0.012	0.000
Sulphates	-0.013	0.000
Alcohol	-0.034	0.000
AcidIndex	1.660	0.149

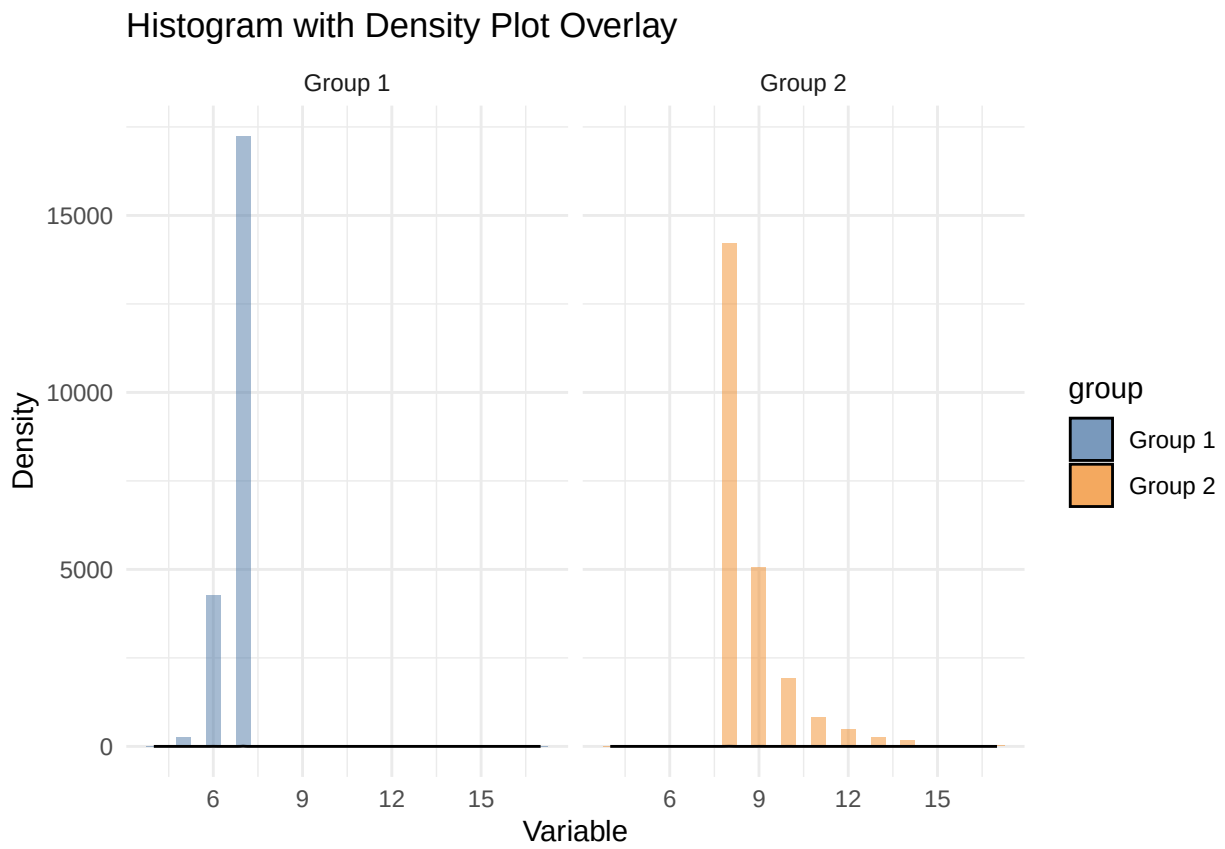
The transformations almost completely got rid of any skew in our data. We can visualize this using by recreating the histograms with the transformed data.



While the first table seems to suggest that AcidIndex_transformed had its skewness lowered to a relatively insignificant amount, the histogram reveals that the variable still seemingly is bimodal. This may suggest that the transformation was not the best choice for this variable and grouping the data may be a better choice. However, upon further investigation, the appearance of bimodality may be due to the amount of bins selected for the histograms. More bins reveal a more normal distribution. We can test whether it is bimodal using a dip test.

```
##
##  Hartigans' dip test for unimodality / multimodality
##
## data:  train_data_transformed$AcidIndex
## D = 0.15872, p-value < 2.2e-16
## alternative hypothesis: non-unimodal, i.e., at least bimodal
```

The extremely low p-value suggests that the AcidIndex variable is bimodal. We will group the data into two categories to deal with this issue.



The resulting groups are shown in the histogram above. The plot reveals that, while not evenly distributed, there really is only one group. The appearance of bimodality is likely due to the much larger amount of the non-median group. We will not group the data and move on to dealing with outliers.

Now that we've transformed our data, we can move on to dealing with outliers.

Outliers

We will use the IQR method to detect outliers in the data. The IQR method is a robust method for detecting outliers that is not sensitive to the presence of extreme values. The IQR method defines an outlier as any value that is below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. The lower and upper limits for each variable are calculated using the IQR method.

Using the IQR limits, there is a significant amount of outliers in the data. The transformation process did not impact the number of outliers in the data. Using a Box-Cox transformation might have been a better way to get rid of the outliers but it was not an option for many of the variables due to them containing negative and zero values.

Ultimately, due to the large amount of outliers, removing them would result in a significant loss of data. We will keep the outliers in the data and move on to one-hot encoding the categorical variables.

One-Hot Encoding

We have two factor columns in LabelAppeal and STARS. LabelAppeal can be converted to numeric as it is ordinal. While STARS is also ordinal, it also has an 'unrated' category. We will one-hot encode this column but also keep the original column for now.

After our data has been prepped, we can now move on to modeling.

Modeling

With the data exploration and preparation out of the way, we turn to build different types of regression models to predict the number of cases of wine ordered by distributors. Again, the response variable is the *count* of cases, and so it is appropriate to consider Poisson regression, negative binomial regression, and multiple linear regression. We build models of each type with some commentary, and then we will consider more generally how the models compare to one another.

Poisson Regression

We first consider Poisson regression models. Now, it's critical to note, despite the transformations we performed in the previous section, Poisson models do not require normally distributed data, and so leveraging transformed data is actually counter-productive. As such, the relevant dataframes are:

- train_data_imputed
- test_data_imputed
- eval_data_imputed

Poisson Model 1

We start with a rather simple model, with all the variables along with a few more sophisticated variables:

1. Alcohol:LabelAppeal in case these two variables have an especially strong combined effect
2. STARS:Alcohol since high quality wines with certain alcohol content might sell especially well
3. LabelAppeal:STARS in case people might be especially likely to buy visually appealing and highly rated wines
4. Alcohol^2 as intuitively alcohol content doesn't have a strictly linear relationship with the target variable

```
##  Factor w/ 9 levels "0","1","2","3",..: 4 4 6 4 5 1 5 7 1 5 ...
```

```
##
## Call:
## glm(formula = TARGET ~ Alcohol + Alcohol^2 + LabelAppeal + STARS +
##     AcidIndex + Chlorides + CitricAcid + Density + FixedAcidity +
##     FreeSulfurDioxide + ResidualSugar + Sulphates + TotalSulfurDioxide +
##     VolatileAcidity + pH + Alcohol:LabelAppeal + STARS:Alcohol +
##     LabelAppeal:STARS, family = poisson(), data = train_data_imputed)
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  1.344e+00  1.233e-01  10.893  < 2e-16 ***
## Alcohol                      8.200e-03  5.323e-03   1.541 0.123437
## LabelAppeal-1                3.952e-01  6.877e-02   5.747 9.10e-09 ***
## LabelAppeal0                 6.420e-01  6.722e-02   9.551  < 2e-16 ***
## LabelAppeal1                 7.300e-01  6.904e-02  10.573  < 2e-16 ***
## LabelAppeal2                 7.548e-01  8.778e-02   8.599  < 2e-16 ***
## STARS2                       2.600e-01  5.629e-02   4.618 3.87e-06 ***
## STARS3                       5.118e-01  8.568e-02   5.973 2.33e-09 ***
## STARS4                       8.040e-01  6.222e-02  12.922  < 2e-16 ***
## STARSUnrated                -5.029e-01  5.674e-02  -8.862  < 2e-16 ***
```

```
## AcidIndex                  -7.576e-02  2.453e-03 -30.881  < 2e-16 ***
## Chlorides                  -3.808e-02  8.609e-03  -4.423 9.71e-06 ***
## CitricAcid                  5.293e-03  3.182e-03   1.663 0.096234 .
## Density                    -3.336e-01  1.029e-01  -3.243 0.001183 **
## FixedAcidity                1.147e-04  4.369e-04   0.263 0.792929
## FreeSulfurDioxide           9.423e-05  1.841e-05   5.118 3.09e-07 ***
## ResidualSugar               1.471e-04  8.122e-05   1.811 0.070104 .
## Sulphates                  -9.047e-03  2.953e-03  -3.063 0.002189 **
## TotalSulfurDioxide          8.507e-05  1.180e-05   7.208 5.66e-13 ***
## VolatileAcidity            -3.508e-02  3.466e-03 -10.122  < 2e-16 ***
## pH                         -1.512e-02  3.989e-03  -3.791 0.000150 ***
## Alcohol:LabelAppeal-1      -4.535e-03  5.507e-03  -0.823 0.410277
## Alcohol:LabelAppeal0       -8.650e-03  5.367e-03  -1.612 0.107037
## Alcohol:LabelAppeal1       -6.552e-03  5.481e-03  -1.195 0.232001
## Alcohol:LabelAppeal2       -2.205e-02  6.358e-03  -3.468 0.000524 ***
## Alcohol:STARS2              7.646e-03  2.031e-03   3.764 0.000167 ***
## Alcohol:STARS3              3.722e-03  2.234e-03   1.666 0.095685 .
## Alcohol:STARS4              3.878e-03  3.233e-03   1.199 0.230435
## Alcohol:STARSUnrated        1.849e-05  2.792e-03   0.007 0.994716
## LabelAppeal-1:STARS2       -9.189e-02  5.377e-02  -1.709 0.087490 .
## LabelAppeal0:STARS2        -4.243e-02  5.275e-02  -0.804 0.421141
## LabelAppeal1:STARS2         5.258e-02  5.419e-02   0.970 0.331901
## LabelAppeal2:STARS2         3.508e-01  7.237e-02   4.847 1.25e-06 ***
## LabelAppeal-1:STARS3       -1.215e-01  8.480e-02  -1.433 0.151866
## LabelAppeal0:STARS3        -1.384e-01  8.315e-02  -1.665 0.095960 .
## LabelAppeal1:STARS3        -8.809e-02  8.397e-02  -1.049 0.294134
## LabelAppeal2:STARS3         2.314e-01  9.602e-02   2.410 0.015964 *
## LabelAppeal-1:STARS4       -1.855e-01  7.662e-02  -2.421 0.015470 *
## LabelAppeal0:STARS4        -2.962e-01  5.521e-02  -5.366 8.06e-08 ***
## LabelAppeal1:STARS4        -2.817e-01  5.527e-02  -5.098 3.44e-07 ***
## LabelAppeal2:STARS4                NA         NA      NA       NA
## LabelAppeal-1:STARSUnrated -9.569e-02  5.166e-02  -1.852 0.064001 .
## LabelAppeal0:STARSUnrated  -2.357e-01  5.051e-02  -4.667 3.05e-06 ***
## LabelAppeal1:STARSUnrated  -5.390e-01  5.437e-02  -9.912  < 2e-16 ***
## LabelAppeal2:STARSUnrated  -6.257e-02  7.974e-02  -0.785 0.432609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 80043  on 44794  degrees of freedom
## Residual deviance: 47535  on 44751  degrees of freedom
## AIC: 159458
##
## Number of Fisher Scoring iterations: 6
```

The model seems promising; for example there's major reduction in deviance from the null model to the full model. Still, there's much work to be done. We start by noting there are undefined coefficients because of singularities. Let's take a look at potential multicollinearity.

```
## Model :
## TARGET ~ Alcohol + Alcohol^2 + LabelAppeal + STARS + AcidIndex +
##     Chlorides + CitricAcid + Density + FixedAcidity + FreeSulfurDioxide +
##     ResidualSugar + Sulphates + TotalSulfurDioxide + VolatileAcidity +
```

```
##     pH + Alcohol:LabelAppeal + STARS:Alcohol + LabelAppeal:STARS
##
## Complete :
##                       (Intercept) Alcohol LabelAppeal-1 LabelAppeal0 LabelAppeal1
## LabelAppeal2:STARS4  0           0       0             0            0
##                      LabelAppeal2 STARS2 STARS3 STARS4 STARSUnrated AcidIndex
## LabelAppeal2:STARS4  0            0      0      1      0            0
##                      Chlorides CitricAcid Density FixedAcidity FreeSulfurDioxide
## LabelAppeal2:STARS4  0         0          0       0            0
##                      ResidualSugar Sulphates TotalSulfurDioxide VolatileAcidity
## LabelAppeal2:STARS4  0             0         0                   0
##                      pH Alcohol:LabelAppeal-1 Alcohol:LabelAppeal0
## LabelAppeal2:STARS4  0  0                     0
##                      Alcohol:LabelAppeal1 Alcohol:LabelAppeal2 Alcohol:STARS2
## LabelAppeal2:STARS4  0                    0                    0
##                      Alcohol:STARS3 Alcohol:STARS4 Alcohol:STARSUnrated
## LabelAppeal2:STARS4  0              0              0
##                      LabelAppeal-1:STARS2 LabelAppeal0:STARS2
## LabelAppeal2:STARS4  0                    0
##                      LabelAppeal1:STARS2 LabelAppeal2:STARS2
## LabelAppeal2:STARS4  0                   0
##                      LabelAppeal-1:STARS3 LabelAppeal0:STARS3
## LabelAppeal2:STARS4  0                    0
##                      LabelAppeal1:STARS3 LabelAppeal2:STARS3
## LabelAppeal2:STARS4  0                   0
##                      LabelAppeal-1:STARS4 LabelAppeal0:STARS4
## LabelAppeal2:STARS4 -1                   -1
##                      LabelAppeal1:STARS4 LabelAppeal-1:STARSUnrated
## LabelAppeal2:STARS4 -1                   0
##                      LabelAppeal0:STARSUnrated LabelAppeal1:STARSUnrated
## LabelAppeal2:STARS4  0                         0
##                      LabelAppeal2:STARSUnrated
## LabelAppeal2:STARS4  0


##
## Call:
## glm(formula = TARGET ~ Alcohol + I(Alcohol^2) + LabelAppeal +
##     STARS + AcidIndex + Chlorides + CitricAcid + Density + FixedAcidity +
##     FreeSulfurDioxide + ResidualSugar + Sulphates + TotalSulfurDioxide +
##     VolatileAcidity + pH + Alcohol:LabelAppeal + STARS:Alcohol,
##     family = poisson(), data = train_data_imputed)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            1.478e+00  1.211e-01  12.205  < 2e-16 ***
## Alcohol                1.814e-03  5.754e-03   0.315 0.752540
## I(Alcohol^2)           2.115e-04  1.027e-04   2.059 0.039449 *
## LabelAppeal-1          3.036e-01  6.323e-02   4.801 1.58e-06 ***
## LabelAppeal0           5.301e-01  6.170e-02   8.592  < 2e-16 ***
## LabelAppeal1           6.403e-01  6.282e-02  10.193  < 2e-16 ***
## LabelAppeal2           9.445e-01  7.224e-02  13.074  < 2e-16 ***
## STARS2                 2.554e-01  2.234e-02  11.432  < 2e-16 ***
## STARS3                 4.111e-01  2.498e-02  16.456  < 2e-16 ***
## STARS4                 5.284e-01  3.691e-02  14.319  < 2e-16 ***
```

```
## STARSUnrated          -7.541e-01  3.044e-02 -24.774  < 2e-16 ***
## AcidIndex             -7.803e-02  2.445e-03 -31.913  < 2e-16 ***
## Chlorides             -3.866e-02  8.609e-03  -4.491 7.09e-06 ***
## CitricAcid             6.276e-03  3.184e-03   1.971 0.048704 *
## Density               -3.394e-01  1.028e-01  -3.302 0.000959 ***
## FixedAcidity           1.370e-04  4.367e-04   0.314 0.753636
## FreeSulfurDioxide      9.861e-05  1.840e-05   5.359 8.38e-08 ***
## ResidualSugar          1.619e-04  8.111e-05   1.996 0.045970 *
## Sulphates             -8.979e-03  2.951e-03  -3.043 0.002345 **
## TotalSulfurDioxide     8.624e-05  1.179e-05   7.318 2.52e-13 ***
## VolatileAcidity       -3.567e-02  3.462e-03 -10.304  < 2e-16 ***
## pH                    -1.513e-02  3.988e-03  -3.793 0.000149 ***
## Alcohol:LabelAppeal-1 -3.224e-03  5.469e-03  -0.590 0.555497
## Alcohol:LabelAppeal0  -6.556e-03  5.327e-03  -1.231 0.218415
## Alcohol:LabelAppeal1  -3.875e-03  5.435e-03  -0.713 0.475845
## Alcohol:LabelAppeal2  -1.860e-02  6.303e-03  -2.951 0.003163 **
## Alcohol:STARS2         6.739e-03  2.015e-03   3.345 0.000823 ***
## Alcohol:STARS3         3.169e-03  2.222e-03   1.426 0.153860
## Alcohol:STARS4         2.983e-03  3.217e-03   0.927 0.353707
## Alcohol:STARSUnrated   9.260e-04  2.781e-03   0.333 0.739112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 80043  on 44794  degrees of freedom
## Residual deviance: 48067  on 44765  degrees of freedom
## AIC: 159962
##
## Number of Fisher Scoring iterations: 6
```

Table 6: VIF Values simple model 2

Term	VIF	VIF_CI_low	VIF_CI_high	SE_factor	Tolerance	Tolerance_CI_low	Tolerance_CI_high
Alcohol	62.616827	61.477857	63.777246	7.913080	0.0159701	0.0156796	0.0162660
I(Alcohol^2)	9.975630	9.801962	10.152725	3.158422	0.1002443	0.0984957	0.1020204
LabelAppeal	9648.902267	9471.976675	9829.132974	98.228826	0.0001036	0.0001017	0.0001056
STARS	8808.211331	8646.701743	8972.738064	93.852071	0.0001135	0.0001114	0.0001157
AcidIndex	1.063457	1.053840	1.074791	1.031241	0.9403296	0.9304135	0.9489102
Chlorides	1.005769	1.001139	1.029226	1.002880	0.9942642	0.9716036	0.9988626
CitricAcid	1.008438	1.002766	1.025737	1.004210	0.9916325	0.9749088	0.9972411
Density	1.005458	1.000984	1.030295	1.002725	0.9945712	0.9705957	0.9990175
FixedAcidity	1.024270	1.016272	1.036199	1.012062	0.9763054	0.9650659	0.9839888
FreeSulfurDioxide	1.006183	1.001359	1.028133	1.003087	0.9938548	0.9726369	0.9986429
ResidualSugar	1.003295	1.000195	1.055650	1.001646	0.9967154	0.9472837	0.9998049
Sulphates	1.004457	1.000549	1.036204	1.002226	0.9955626	0.9650609	0.9994516
TotalSulfurDioxide	1.005488	1.000998	1.030182	1.002740	0.9945418	0.9707026	0.9990031
VolatileAcidity	1.007462	1.002119	1.026271	1.003724	0.9925937	0.9744016	0.9978852
pH	1.008664	1.002923	1.025681	1.004323	0.9914106	0.9749619	0.9970857
Alcohol:LabelAppeal	61462.048045	60335.009097	62610.140075	247.915405	0.0000163	0.0000160	0.0000166
Alcohol:STARS	13147.238693	12906.163046	13392.817768	114.661409	0.0000761	0.0000747	0.0000775

There's clearly a high degree of collinearity, but it's critical to remove one column at a time and reassess

colinearity:

Term	VIF	VIF_CI_low	VIF_CI_high	SE_factor	Tolerance	Tolerance_CI_low	Tolerance_CI_high
Alcohol	13.336455	13.101136	13.576349	3.651911	0.0749824	0.0736575	0.0763293
I(Alcohol^2)	9.951011	9.777779	10.127662	3.154522	0.1004923	0.0987395	0.1022727
LabelAppeal	1.138945	1.127608	1.151289	1.067214	0.8780054	0.8685913	0.8868329
STARS	7702.157091	7560.917052	7846.035882	87.761934	0.0001298	0.0001275	0.0001323
AcidIndex	1.062630	1.053036	1.073960	1.030839	0.9410613	0.9311336	0.9496352
Chlorides	1.005394	1.000952	1.030556	1.002693	0.9946352	0.9703504	0.9990488
CitricAcid	1.007554	1.002178	1.026198	1.003770	0.9925022	0.9744705	0.9978264
Density	1.005196	1.000859	1.031420	1.002595	0.9948305	0.9695372	0.9991413
FixedAcidity	1.024144	1.016154	1.036084	1.012000	0.9764256	0.9651730	0.9841024
FreeSulfurDioxide	1.005811	1.001160	1.029103	1.002901	0.9942224	0.9717200	0.9988410
ResidualSugar	1.003246	1.000184	1.057221	1.001622	0.9967644	0.9458759	0.9998159
Sulphates	1.003738	1.000309	1.045273	1.001867	0.9962755	0.9566876	0.9996914
TotalSulfurDioxide	1.004897	1.000727	1.033014	1.002446	0.9951265	0.9680415	0.9992740
VolatileAcidity	1.006717	1.001662	1.027138	1.003353	0.9933278	0.9735790	0.9983402
pH	1.008217	1.002615	1.025816	1.004100	0.9918499	0.9748339	0.9973914
Alcohol:STARS	11232.455188	11026.473271	11442.285333	105.983278	0.0000890	0.0000874	0.0000907

Table 8: VIF Values simple model 2

Term	VIF	VIF_CI_low	VIF_CI_high	SE_factor	Tolerance	Tolerance_CI_low	Tolerance_CI_high
Alcohol	9.907412	9.734964	10.083264	3.147604	0.1009345	0.0991742	0.1027225
I(Alcohol^2)	9.889156	9.717043	10.064668	3.144703	0.1011209	0.0993575	0.1029120
LabelAppeal	1.137419	1.126113	1.149740	1.066499	0.8791833	0.8697620	0.8880108
STARS	1.170559	1.158588	1.183435	1.081924	0.8542924	0.8449979	0.8631198
AcidIndex	1.061721	1.052151	1.073046	1.030398	0.9418672	0.9319266	0.9504335
Chlorides	1.004810	1.000689	1.033565	1.002402	0.9952134	0.9675249	0.9993113
CitricAcid	1.007451	1.002112	1.026286	1.003719	0.9926038	0.9743877	0.9978922
Density	1.004941	1.000745	1.032762	1.002468	0.9950833	0.9682775	0.9992554
FixedAcidity	1.023935	1.015961	1.035893	1.011897	0.9766241	0.9653503	0.9842895
FreeSulfurDioxide	1.005755	1.001131	1.029278	1.002873	0.9942781	0.9715549	0.9988701
ResidualSugar	1.003118	1.000157	1.061810	1.001558	0.9968917	0.9417877	0.9998427
Sulphates	1.003582	1.000265	1.048346	1.001789	0.9964310	0.9538838	0.9997347
TotalSulfurDioxide	1.004041	1.000402	1.040660	1.002019	0.9959749	0.9609283	0.9995985
VolatileAcidity	1.006550	1.001565	1.027412	1.003270	0.9934923	0.9733190	0.9984372
pH	1.007897	1.002401	1.025972	1.003941	0.9921648	0.9746857	0.9976046

At this point, we've removed all high correlation variables. Now, there are still two variables with fairly high VIFs, namely Alcohol and I(Alcohol^2)–this is unsurprising to say the least. It would be odd to remove only the former term, so let's see the model summary and consider whether we ought to remove I(Alcohol^2):

```
##
## Call:
## glm(formula = TARGET ~ Alcohol + I(Alcohol^2) + LabelAppeal +
##     STARS + AcidIndex + Chlorides + CitricAcid + Density + FixedAcidity +
##     FreeSulfurDioxide + ResidualSugar + Sulphates + TotalSulfurDioxide +
##     VolatileAcidity + pH, family = poisson(), data = train_data_imputed)
##
## Coefficients:
```

```
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.505e+00  1.065e-01  14.134  < 2e-16 ***
## Alcohol            -4.225e-04  2.289e-03  -0.185 0.853539
## I(Alcohol^2)        2.183e-04  1.023e-04   2.134 0.032804 *
## LabelAppeal-1       2.688e-01  2.074e-02  12.962  < 2e-16 ***
## LabelAppeal0        4.599e-01  2.027e-02  22.689  < 2e-16 ***
## LabelAppeal1        5.983e-01  2.060e-02  29.041  < 2e-16 ***
## LabelAppeal2        7.446e-01  2.318e-02  32.127  < 2e-16 ***
## STARS2              3.258e-01  7.656e-03  42.553  < 2e-16 ***
## STARS3              4.437e-01  8.361e-03  53.075  < 2e-16 ***
## STARS4              5.586e-01  1.150e-02  48.565  < 2e-16 ***
## STARSUnrated       -7.445e-01  1.046e-02 -71.179  < 2e-16 ***
## AcidIndex          -7.794e-02  2.442e-03 -31.914  < 2e-16 ***
## Chlorides          -3.930e-02  8.603e-03  -4.568 4.91e-06 ***
## CitricAcid          6.049e-03  3.183e-03   1.900 0.057403 .
## Density            -3.418e-01  1.027e-01  -3.328 0.000875 ***
## FixedAcidity        1.468e-04  4.365e-04   0.336 0.736670
## FreeSulfurDioxide   9.913e-05  1.840e-05   5.387 7.16e-08 ***
## ResidualSugar       1.618e-04  8.110e-05   1.995 0.046018 *
## Sulphates          -8.779e-03  2.950e-03  -2.976 0.002917 **
## TotalSulfurDioxide  8.684e-05  1.178e-05   7.374 1.66e-13 ***
## VolatileAcidity    -3.535e-02  3.460e-03 -10.218  < 2e-16 ***
## pH                 -1.541e-02  3.987e-03  -3.864 0.000112 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 80043  on 44794  degrees of freedom
## Residual deviance: 48099  on 44773  degrees of freedom
## AIC: 159978
##
## Number of Fisher Scoring iterations: 6
```

Indeed, I(Alcohol^2) *is* statistically significant, so we won't remove it. However, there are a couple variables that appear less promising, and we will removes those one at a time (a manual backwards elimination process).

```
##
## Call:
## glm(formula = TARGET ~ Alcohol + I(Alcohol^2) + LabelAppeal +
##     STARS + AcidIndex + Chlorides + Density + FreeSulfurDioxide +
##     Sulphates + TotalSulfurDioxide + VolatileAcidity + pH, family = poisson(),
##     data = train_data_imputed)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.507e+00  1.064e-01  14.158  < 2e-16 ***
## Alcohol            -4.546e-04  2.289e-03  -0.199 0.842536
## I(Alcohol^2)        2.199e-04  1.022e-04   2.151 0.031464 *
## LabelAppeal-1       2.685e-01  2.073e-02  12.948  < 2e-16 ***
## LabelAppeal0        4.595e-01  2.027e-02  22.668  < 2e-16 ***
## LabelAppeal1        5.980e-01  2.060e-02  29.029  < 2e-16 ***
## LabelAppeal2        7.450e-01  2.317e-02  32.150  < 2e-16 ***
```

```
## STARS2              3.262e-01  7.654e-03  42.618  < 2e-16 ***
## STARS3              4.440e-01  8.360e-03  53.113  < 2e-16 ***
## STARS4              5.589e-01  1.150e-02  48.596  < 2e-16 ***
## STARSUnrated       -7.447e-01  1.046e-02 -71.201  < 2e-16 ***
## AcidIndex          -7.751e-02  2.412e-03 -32.128  < 2e-16 ***
## Chlorides          -3.975e-02  8.601e-03  -4.622 3.80e-06 ***
## Density            -3.439e-01  1.027e-01  -3.349 0.000811 ***
## FreeSulfurDioxide   9.995e-05  1.840e-05   5.432 5.56e-08 ***
## Sulphates          -8.900e-03  2.948e-03  -3.019 0.002539 **
## TotalSulfurDioxide  8.728e-05  1.177e-05   7.413 1.23e-13 ***
## VolatileAcidity    -3.551e-02  3.459e-03 -10.267  < 2e-16 ***
## pH                 -1.527e-02  3.986e-03  -3.830 0.000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 80043  on 44794  degrees of freedom
## Residual deviance: 48107  on 44776  degrees of freedom
## AIC: 159980
##
## Number of Fisher Scoring iterations: 6
```

There are a number of takeaways from this model summary, most of which are totally expected. First, only the quadratic term for alcohol is significant; this suggests that after a certain point, small changes in alcohol content–past a certain point–can have a large impact on the target variable. Second, the higher the label appeal level, the higher the log counts of cases ordered. Third, higher star ratings are highly associated with higher values for the target variable; being unrated significantly decreases the log count of cases ordered–we return to this point momentarily. Finally, a number of the chemical properties have effects on the target variable. For example, AcidIndex, Density, and VolatileAcidity all have negative coefficients. While I'm not a wine connoisseur myself, a highly dense wine seems unappealing at least.

Again, we note the huge reduction in deviance when going from the null model to the full model–this speaks well to our model. Now, a further question is if a Poisson model is appropriate here. A key condition for Poisson is that the mean and variance of the response variable are equal. We check this now:

```
## [1] 0.8837684
```

So there certainly isn't over-dispersion. The under-dispersion is somewhat surprising, but the value is close enough to 1, and certainly close enough for a baseline model. We turn now to construct a new model

Poisson Model 2 (Zero-Inflated)

We observed in our last model that unrated wines perform especially badly. Recall, though, we actually turned those values to unrated; they were missing at first. What if, then, these values should actually be a "0" rating? If so, we might be able to use a model that both improves accuracy and interpretability. The first step, then, is to create a new column changing the unrated values to zeros.

Table 9: STARS Value Counts

Var1	Freq
0	11585
1	10755
2	12625
3	7635
4	2195

Immediately we see that this change leads to a large number of zeroes. This is a strong indicator for considering a zero-inflated model, especially given that a different process may well have led to a zero rating than the process that led to the other ratings.

We start with creating a zero-inflated model using the same variables as the most recent Poisson model as that provides a strong baseline:

```
##
## Call:
## zeroinfl(formula = TARGET ~ Alcohol + I(Alcohol^2) + LabelAppeal + original_stars +
##     AcidIndex + Chlorides + Density + FreeSulfurDioxide + Sulphates +
##     TotalSulfurDioxide + VolatileAcidity + pH | original_stars, data = train_data_imputed,
##     dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.18614 -0.51971  0.01761  0.40950  2.87565
##
## Count model coefficients (poisson with log link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         9.261e-01  1.102e-01   8.405  < 2e-16 ***
## Alcohol             6.327e-03  2.298e-03   2.754  0.00589 **
## I(Alcohol^2)        5.421e-06  1.008e-04   0.054  0.95714
## LabelAppeal-1       3.886e-01  2.144e-02  18.123  < 2e-16 ***
## LabelAppeal0        6.625e-01  2.100e-02  31.554  < 2e-16 ***
## LabelAppeal1        8.558e-01  2.139e-02  40.000  < 2e-16 ***
## LabelAppeal2        1.019e+00  2.397e-02  42.529  < 2e-16 ***
## original_stars      9.769e-02  2.783e-03  35.106  < 2e-16 ***
## AcidIndex          -2.690e-02  2.669e-03 -10.078  < 2e-16 ***
## Chlorides          -2.682e-02  8.807e-03  -3.045  0.00232 **
## Density            -3.123e-01  1.062e-01  -2.941  0.00327 **
## FreeSulfurDioxide   3.227e-05  1.862e-05   1.733  0.08302 .
## Sulphates           1.543e-04  3.024e-03   0.051  0.95930
## TotalSulfurDioxide  1.201e-05  1.173e-05   1.024  0.30599
## VolatileAcidity    -1.917e-02  3.549e-03  -5.402 6.58e-08 ***
## pH                  2.642e-03  4.097e-03   0.645  0.51889
##
## Zero-inflation model coefficients (binomial with logit link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.38027    0.01948   19.52   <2e-16 ***
## original_stars -2.18435    0.02824  -77.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Number of iterations in BFGS optimization: 23
## Log-likelihood: -7.297e+04 on 18 Df
```

It's quite interesting how this one change changed the model fairly significantly. We will finish removing variables, again using a backward elimination process, and then add more commentary.

```
##
## Call:
## zeroinfl(formula = TARGET ~ Alcohol + LabelAppeal + original_stars +
##     AcidIndex + Chlorides + Density + VolatileAcidity | original_stars,
##     data = train_data_imputed, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.18913 -0.51720  0.01805  0.40872  2.87298
##
## Count model coefficients (poisson with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.9344984  0.1087370   8.594  < 2e-16 ***
## Alcohol          0.0064163  0.0007475   8.583  < 2e-16 ***
## LabelAppeal-1    0.3881827  0.0214393  18.106  < 2e-16 ***
## LabelAppeal0     0.6624722  0.0209931  31.557  < 2e-16 ***
## LabelAppeal1     0.8560520  0.0213906  40.020  < 2e-16 ***
## LabelAppeal2     1.0192949  0.0239632  42.536  < 2e-16 ***
## original_stars   0.0974396  0.0027784  35.071  < 2e-16 ***
## AcidIndex       -0.0271256  0.0026596 -10.199  < 2e-16 ***
## Chlorides       -0.0273112  0.0087975  -3.104  0.00191 **
## Density         -0.3075116  0.1061497  -2.897  0.00377 **
## VolatileAcidity -0.0191426  0.0035470  -5.397 6.78e-08 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.38060    0.01947   19.55   <2e-16 ***
## original_stars -2.18398    0.02821  -77.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 19
## Log-likelihood: -7.297e+04 on 13 Df
```

There are many observations to be made. The first is that we used the p-value to eliminate predictors that were not significant, and it is striking that we were able to eliminate five variables once we switched to a zero-inflated model. Second, the quadratic alcohol term was one of those terms that was no longer significant. We were also able to eliminate Sulphates, pH, and the SulfurDioxide variables. As for the variables that persisted, the effects are not all that different: Label Appeal and Stars have a positive effect, chemical properties have negative effects. The key difference is that Alcohol now has a positive effect, but that's intuitive now that the previously positively impacting quadratic term is now removed.

As for the Stars variable (here called original_stars), again higher star ratings are associated with a higher log count of cases ordered. It's also the case that wines with no star ratings are more likely to have zero cases ordered.

Let's now generate predictions for the two Poisson models:

```
## MAE Poisson Model:  1.016386
```

```
## RMSE Poisson Model:  1.275739
```

```
## MAE ZIP Model:  0.9956513
```

```
## RMSE ZIP Model:  1.284232
```

Again, we will compare all models once all models are built, although it's worth noting that both MAE and RMSE values are pretty close to 1, which might be acceptable. But before we get ahead of ourselves, let's build the next two models.

Negative Binomial

There is reason to believe that switching to a negative binomial model will yield better results. Specifically, the negative binomial is appropriate when we are working with count data that has over-dispersion (the variance is greater than the mean). Now, it is true that earlier we saw under-dispersion relative to what one Poisson model expects. However, the truth is that it is really worthwhile to more get a direct measure of the dispersion in the outcome variable, before even modelling:

```
observed_variance <- var(train_data_imputed$TARGET)
expected_mean <- mean(train_data_imputed$TARGET)
print(observed_variance / expected_mean)
```

```
## [1] 1.225267
```

We see that this dispersion statistic is greater than 1. This suggests that we ought to try a negative binomial model.

Negative Binomial Model 1

Much like earlier, we will start with a relatively simple model, at first using all the variables as well as the interaction terms attempted earlier, and then engaging in variable selection.

As a reminder, those additional variables are:

1. Alcohol:LabelAppeal
2. STARS:Alcohol
3. Alcohol^2

(We omit LabelAppeal:STARS for the reason discussed earlier)

While it is true that most or all of these additional variables were not significant in the previous two models, it does not follow that they'll be insignificant in the negative binomial models.

```
##
## Call:
## glm.nb(formula = TARGET ~ Alcohol + I(Alcohol^2) + LabelAppeal +
##     STARS + AcidIndex + Chlorides + CitricAcid + Density + FixedAcidity +
##     FreeSulfurDioxide + ResidualSugar + Sulphates + TotalSulfurDioxide +
##     VolatileAcidity + pH + Alcohol:LabelAppeal + STARS:Alcohol,
##     data = train_data_imputed, init.theta = 41158.99707, link = log)
##
```

```
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.478e+00  1.211e-01  12.204  < 2e-16 ***
## Alcohol                 1.814e-03  5.755e-03   0.315 0.752556
## I(Alcohol^2)            2.115e-04  1.027e-04   2.059 0.039449 *
## LabelAppeal-1           3.036e-01  6.324e-02   4.801 1.58e-06 ***
## LabelAppeal0            5.301e-01  6.170e-02   8.592  < 2e-16 ***
## LabelAppeal1            6.403e-01  6.282e-02  10.192  < 2e-16 ***
## LabelAppeal2            9.446e-01  7.225e-02  13.074  < 2e-16 ***
## STARS2                  2.554e-01  2.234e-02  11.431  < 2e-16 ***
## STARS3                  4.111e-01  2.498e-02  16.455  < 2e-16 ***
## STARS4                  5.284e-01  3.691e-02  14.318  < 2e-16 ***
## STARSUnrated           -7.541e-01  3.044e-02 -24.773  < 2e-16 ***
## AcidIndex              -7.804e-02  2.445e-03 -31.912  < 2e-16 ***
## Chlorides              -3.866e-02  8.609e-03  -4.491 7.09e-06 ***
## CitricAcid              6.276e-03  3.184e-03   1.971 0.048712 *
## Density                -3.394e-01  1.028e-01  -3.302 0.000959 ***
## FixedAcidity            1.370e-04  4.367e-04   0.314 0.753627
## FreeSulfurDioxide       9.861e-05  1.840e-05   5.359 8.38e-08 ***
## ResidualSugar           1.619e-04  8.112e-05   1.996 0.045968 *
## Sulphates              -8.979e-03  2.951e-03  -3.043 0.002344 **
## TotalSulfurDioxide      8.624e-05  1.179e-05   7.318 2.52e-13 ***
## VolatileAcidity        -3.567e-02  3.462e-03 -10.304  < 2e-16 ***
## pH                     -1.513e-02  3.988e-03  -3.793 0.000149 ***
## Alcohol:LabelAppeal-1  -3.225e-03  5.470e-03  -0.590 0.555504
## Alcohol:LabelAppeal0   -6.556e-03  5.327e-03  -1.231 0.218412
## Alcohol:LabelAppeal1   -3.875e-03  5.435e-03  -0.713 0.475857
## Alcohol:LabelAppeal2   -1.860e-02  6.303e-03  -2.951 0.003163 **
## Alcohol:STARS2          6.739e-03  2.015e-03   3.345 0.000824 ***
## Alcohol:STARS3          3.168e-03  2.222e-03   1.426 0.153880
## Alcohol:STARS4          2.983e-03  3.217e-03   0.927 0.353757
## Alcohol:STARSUnrated    9.260e-04  2.781e-03   0.333 0.739116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(41159) family taken to be 1)
##
##     Null deviance: 80039  on 44794  degrees of freedom
## Residual deviance: 48065  on 44765  degrees of freedom
## AIC: 159965
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  41159
##           Std. Err.:  18751
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -159903.1
```

If the previous model was any indication, there's likely high collinearity. Let's check:

Table 10: VIF Values for NB Model

Term	VIF	VIF_CI_low	VIF_CI_high	SE_factor	Tolerance	Tolerance_CI_low	Tolerance_CI_high
Alcohol	62.614401	61.475476	63.774775	7.912926	0.0159708	0.0156802	0.0162666
I(Alcohol^2)	9.975600	9.801932	10.152694	3.158417	0.1002446	0.0984960	0.1020207
LabelAppeal	9648.864165	9471.939272	9829.094160	98.228632	0.0001036	0.0001017	0.0001056
STARS	8808.209861	8646.700301	8972.736567	93.852064	0.0001135	0.0001114	0.0001157
AcidIndex	1.063458	1.053841	1.074792	1.031241	0.9403290	0.9304130	0.9489096
Chlorides	1.005769	1.001139	1.029227	1.002880	0.9942643	0.9716033	0.9988627
CitricAcid	1.008438	1.002766	1.025737	1.004210	0.9916325	0.9749088	0.9972411
Density	1.005458	1.000983	1.030296	1.002725	0.9945714	0.9705952	0.9990176
FixedAcidity	1.024270	1.016272	1.036199	1.012062	0.9763050	0.9650655	0.9839884
FreeSulfurDioxide	1.006183	1.001359	1.028133	1.003087	0.9938549	0.9726368	0.9986429
ResidualSugar	1.003295	1.000195	1.055652	1.001646	0.9967154	0.9472821	0.9998049
Sulphates	1.004457	1.000549	1.036204	1.002226	0.9955626	0.9650611	0.9994516
TotalSulfurDioxide	1.005488	1.000998	1.030182	1.002740	0.9945418	0.9707025	0.9990031
VolatileAcidity	1.007461	1.002119	1.026271	1.003724	0.9925939	0.9744014	0.9978853
pH	1.008664	1.002923	1.025681	1.004323	0.9914107	0.9749619	0.9970857
Alcohol:LabelAppeal	61459.298561	60332.310032	62607.339231	247.909860	0.0000163	0.0000160	0.0000166
Alcohol:STARS	13147.088803	12906.015905	13392.665078	114.660755	0.0000761	0.0000747	0.0000775

Yet again, there are some extraordinarily high VIF values, likely because of the interaction terms. Again, then, we consult the VIF, remodel, and repeat until there are no variables with high collinearity.

```
##
## Call:
## glm.nb(formula = TARGET ~ Alcohol + I(Alcohol^2) + LabelAppeal +
##     STARS + AcidIndex + Chlorides + CitricAcid + Density + FixedAcidity +
##     FreeSulfurDioxide + ResidualSugar + Sulphates + TotalSulfurDioxide +
##     VolatileAcidity + pH, data = train_data_imputed, init.theta = 41131.90506,
##     link = log)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.505e+00  1.065e-01  14.134  < 2e-16 ***
## Alcohol             -4.228e-04  2.289e-03  -0.185 0.853454
## I(Alcohol^2)         2.183e-04  1.023e-04   2.134 0.032803 *
## LabelAppeal-1        2.688e-01  2.074e-02  12.961  < 2e-16 ***
## LabelAppeal0         4.599e-01  2.027e-02  22.689  < 2e-16 ***
## LabelAppeal1         5.983e-01  2.060e-02  29.040  < 2e-16 ***
## LabelAppeal2         7.446e-01  2.318e-02  32.126  < 2e-16 ***
## STARS2               3.258e-01  7.656e-03  42.552  < 2e-16 ***
## STARS3               4.437e-01  8.361e-03  53.073  < 2e-16 ***
## STARS4               5.586e-01  1.150e-02  48.562  < 2e-16 ***
## STARSUnrated        -7.445e-01  1.046e-02 -71.178  < 2e-16 ***
## AcidIndex           -7.794e-02  2.442e-03 -31.914  < 2e-16 ***
## Chlorides           -3.930e-02  8.604e-03  -4.568 4.92e-06 ***
## CitricAcid           6.049e-03  3.183e-03   1.900 0.057412 .
## Density             -3.418e-01  1.027e-01  -3.328 0.000875 ***
## FixedAcidity         1.468e-04  4.366e-04   0.336 0.736665
## FreeSulfurDioxide    9.914e-05  1.840e-05   5.387 7.16e-08 ***
## ResidualSugar        1.618e-04  8.110e-05   1.995 0.046016 *
## Sulphates           -8.780e-03  2.950e-03  -2.976 0.002916 **
```

```
## TotalSulfurDioxide  8.684e-05  1.178e-05   7.374 1.66e-13 ***
## VolatileAcidity    -3.535e-02  3.460e-03 -10.218  < 2e-16 ***
## pH                 -1.541e-02  3.987e-03  -3.864 0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(41131.91) family taken to be 1)
##
##     Null deviance: 80039  on 44794  degrees of freedom
## Residual deviance: 48097  on 44773  degrees of freedom
## AIC: 159982
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  41132
##           Std. Err.:  18745
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -159935.6
```

There are still a few columns that don't have significant predictors, and so again we consult the p-values to remove them one at a time:

```
##
## Call:
## glm.nb(formula = TARGET ~ Alcohol + I(Alcohol^2) + LabelAppeal +
##     STARS + AcidIndex + Chlorides + CitricAcid + Density + FreeSulfurDioxide +
##     ResidualSugar + Sulphates + TotalSulfurDioxide + VolatileAcidity +
##     pH, data = train_data_imputed, init.theta = 41133.97167,
##     link = log)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.504e+00  1.065e-01  14.131  < 2e-16 ***
## Alcohol            -4.237e-04  2.289e-03  -0.185 0.853166
## I(Alcohol^2)        2.184e-04  1.023e-04   2.135 0.032731 *
## LabelAppeal-1       2.687e-01  2.074e-02  12.959  < 2e-16 ***
## LabelAppeal0        4.599e-01  2.027e-02  22.687  < 2e-16 ***
## LabelAppeal1        5.982e-01  2.060e-02  29.038  < 2e-16 ***
## LabelAppeal2        7.445e-01  2.318e-02  32.124  < 2e-16 ***
## STARS2              3.258e-01  7.656e-03  42.551  < 2e-16 ***
## STARS3              4.438e-01  8.360e-03  53.080  < 2e-16 ***
## STARS4              5.586e-01  1.150e-02  48.562  < 2e-16 ***
## STARSUnrated       -7.445e-01  1.046e-02 -71.178  < 2e-16 ***
## AcidIndex          -7.783e-02  2.418e-03 -32.182  < 2e-16 ***
## Chlorides          -3.934e-02  8.603e-03  -4.573 4.81e-06 ***
## CitricAcid          6.072e-03  3.183e-03   1.908 0.056420 .
## Density            -3.413e-01  1.027e-01  -3.323 0.000890 ***
## FreeSulfurDioxide   9.916e-05  1.840e-05   5.388 7.11e-08 ***
## ResidualSugar       1.616e-04  8.110e-05   1.993 0.046291 *
## Sulphates          -8.753e-03  2.949e-03  -2.969 0.002992 **
## TotalSulfurDioxide  8.678e-05  1.178e-05   7.369 1.72e-13 ***
## VolatileAcidity    -3.534e-02  3.460e-03 -10.215  < 2e-16 ***
```

```
## pH                -1.539e-02  3.987e-03  -3.861 0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(41133.97) family taken to be 1)
##
##     Null deviance: 80039  on 44794  degrees of freedom
## Residual deviance: 48097  on 44774  degrees of freedom
## AIC: 159980
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  41134
##           Std. Err.:  18747
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -159935.7

##
## Call:
## glm.nb(formula = TARGET ~ Alcohol + I(Alcohol^2) + LabelAppeal +
##     STARS + AcidIndex + Chlorides + Density + FreeSulfurDioxide +
##     ResidualSugar + Sulphates + TotalSulfurDioxide + VolatileAcidity +
##     pH, data = train_data_imputed, init.theta = 41133.49204,
##     link = log)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.507e+00  1.064e-01  14.154  < 2e-16 ***
## Alcohol            -4.679e-04  2.289e-03  -0.204 0.838026
## I(Alcohol^2)        2.214e-04  1.023e-04   2.165 0.030391 *
## LabelAppeal-1       2.683e-01  2.073e-02  12.938  < 2e-16 ***
## LabelAppeal0        4.595e-01  2.027e-02  22.667  < 2e-16 ***
## LabelAppeal1        5.979e-01  2.060e-02  29.022  < 2e-16 ***
## LabelAppeal2        7.445e-01  2.318e-02  32.125  < 2e-16 ***
## STARS2              3.260e-01  7.655e-03  42.579  < 2e-16 ***
## STARS3              4.438e-01  8.360e-03  53.085  < 2e-16 ***
## STARS4              5.589e-01  1.150e-02  48.592  < 2e-16 ***
## STARSUnrated       -7.447e-01  1.046e-02 -71.196  < 2e-16 ***
## AcidIndex          -7.752e-02  2.413e-03 -32.128  < 2e-16 ***
## Chlorides          -3.967e-02  8.601e-03  -4.612 4.00e-06 ***
## Density            -3.436e-01  1.027e-01  -3.345 0.000821 ***
## FreeSulfurDioxide   9.935e-05  1.840e-05   5.398 6.72e-08 ***
## ResidualSugar       1.585e-04  8.107e-05   1.955 0.050578 .
## Sulphates          -8.817e-03  2.949e-03  -2.990 0.002789 **
## TotalSulfurDioxide  8.687e-05  1.177e-05   7.378 1.61e-13 ***
## VolatileAcidity    -3.541e-02  3.459e-03 -10.235  < 2e-16 ***
## pH                 -1.541e-02  3.987e-03  -3.865 0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(41133.49) family taken to be 1)
##
```

```
##     Null deviance: 80039  on 44794  degrees of freedom
## Residual deviance: 48101  on 44775  degrees of freedom
## AIC: 159981
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  41133
##           Std. Err.:  18748
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -159939.4
```

```
##
## Call:
## glm.nb(formula = TARGET ~ Alcohol + I(Alcohol^2) + LabelAppeal +
##     STARS + AcidIndex + Chlorides + Density + FreeSulfurDioxide +
##     Sulphates + TotalSulfurDioxide + VolatileAcidity + pH, data = train_data_imputed,
##     init.theta = 41129.92159, link = log)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.507e+00  1.065e-01  14.158  < 2e-16 ***
## Alcohol            -4.549e-04  2.289e-03  -0.199 0.842451
## I(Alcohol^2)        2.199e-04  1.022e-04   2.151 0.031464 *
## LabelAppeal-1       2.685e-01  2.073e-02  12.948  < 2e-16 ***
## LabelAppeal0        4.595e-01  2.027e-02  22.668  < 2e-16 ***
## LabelAppeal1        5.980e-01  2.060e-02  29.028  < 2e-16 ***
## LabelAppeal2        7.450e-01  2.317e-02  32.149  < 2e-16 ***
## STARS2              3.262e-01  7.654e-03  42.617  < 2e-16 ***
## STARS3              4.440e-01  8.360e-03  53.111  < 2e-16 ***
## STARS4              5.589e-01  1.150e-02  48.594  < 2e-16 ***
## STARSUnrated       -7.447e-01  1.046e-02 -71.200  < 2e-16 ***
## AcidIndex          -7.751e-02  2.413e-03 -32.128  < 2e-16 ***
## Chlorides          -3.975e-02  8.601e-03  -4.622 3.80e-06 ***
## Density            -3.440e-01  1.027e-01  -3.349 0.000811 ***
## FreeSulfurDioxide   9.996e-05  1.840e-05   5.432 5.56e-08 ***
## Sulphates          -8.901e-03  2.949e-03  -3.019 0.002539 **
## TotalSulfurDioxide  8.728e-05  1.177e-05   7.413 1.23e-13 ***
## VolatileAcidity    -3.551e-02  3.459e-03 -10.266  < 2e-16 ***
## pH                 -1.527e-02  3.986e-03  -3.830 0.000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(41129.92) family taken to be 1)
##
##     Null deviance: 80039  on 44794  degrees of freedom
## Residual deviance: 48105  on 44776  degrees of freedom
## AIC: 159983
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  41130
```

```
##           Std. Err.:  18745
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -159943.2
```

And so we arrive at a negative binomial model with some very interesting results. In particular, the estimates and standard errors are nearly identical as with the simple Poisson model! In fact, even the AIC values are quite similar.

Table 11: Comparison of Simple Poisson and Negatiive Binomial Models

Term	Poisson_Estimate	NB_Estimate	Poisson_Std_Error	NB_Std_Error
(Intercept)	1.5071	1.5071	0.1064	0.1065
Alcohol	-0.0005	-0.0005	0.0023	0.0023
I(Alcohol^2)	0.0002	0.0002	0.0001	0.0001
LabelAppeal-1	0.2685	0.2685	0.0207	0.0207
LabelAppeal0	0.4595	0.4595	0.0203	0.0203
LabelAppeal1	0.5980	0.5980	0.0206	0.0206
LabelAppeal2	0.7450	0.7450	0.0232	0.0232
STARS2	0.3262	0.3262	0.0077	0.0077
STARS3	0.4440	0.4440	0.0084	0.0084
STARS4	0.5589	0.5589	0.0115	0.0115
STARSUnrated	-0.7447	-0.7447	0.0105	0.0105
AcidIndex	-0.0775	-0.0775	0.0024	0.0024
Chlorides	-0.0398	-0.0398	0.0086	0.0086
Density	-0.3439	-0.3440	0.1027	0.1027
FreeSulfurDioxide	0.0001	0.0001	0.0000	0.0000
Sulphates	-0.0089	-0.0089	0.0029	0.0029
TotalSulfurDioxide	0.0001	0.0001	0.0000	0.0000
VolatileAcidity	-0.0355	-0.0355	0.0035	0.0035
pH	-0.0153	-0.0153	0.0040	0.0040

This is striking at first, and it is *possible* that it's due to convergence issues with the negative binomial model. It also seems that the dispersion is quite close to Poisson Assumptions (i.e. mean approximately equal to the variance), which is supported by the very high value of theta. Again, model selection will occur after all the models are built, but it certainly seems that there is no reason to accept this model over the simple Poisson one.

Before we give up entirely on a negative binomial model, though, let's try a zero-inflated model as we did earlier:

Negative Binomial Model 2 (Zero-Inflated)

We mimic the approach from earlier, starting by creating a zero-inflated model using the same variables as the most recent Negative Binomial model as that provides a strong baseline:

```
##
## Call:
## zeroinfl(formula = TARGET ~ Alcohol + I(Alcohol^2) + LabelAppeal + original_stars +
##     AcidIndex + Chlorides + Density + FreeSulfurDioxide + Sulphates +
##     TotalSulfurDioxide + VolatileAcidity + pH | original_stars, data = train_data_imputed,
##     dist = "negbin")
```

```
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.18618 -0.51967  0.01759  0.40947  2.87559
##
## Count model coefficients (negbin with log link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         9.255e-01  1.102e-01   8.400  < 2e-16 ***
## Alcohol             6.329e-03  2.298e-03   2.755  0.00588 **
## I(Alcohol^2)        5.381e-06  1.008e-04   0.053  0.95745
## LabelAppeal-1       3.887e-01  2.144e-02  18.126  < 2e-16 ***
## LabelAppeal0        6.626e-01  2.100e-02  31.558  < 2e-16 ***
## LabelAppeal1        8.559e-01  2.139e-02  40.004  < 2e-16 ***
## LabelAppeal2        1.019e+00  2.397e-02  42.532  < 2e-16 ***
## original_stars      9.768e-02  2.783e-03  35.103  < 2e-16 ***
## AcidIndex          -2.690e-02  2.669e-03 -10.080  < 2e-16 ***
## Chlorides          -2.682e-02  8.807e-03  -3.045  0.00233 **
## Density            -3.118e-01  1.062e-01  -2.937  0.00332 **
## FreeSulfurDioxide   3.227e-05  1.862e-05   1.733  0.08304 .
## Sulphates           1.622e-04  3.024e-03   0.054  0.95724
## TotalSulfurDioxide  1.201e-05  1.173e-05   1.024  0.30569
## VolatileAcidity    -1.917e-02  3.549e-03  -5.402 6.58e-08 ***
## pH                  2.646e-03  4.097e-03   0.646  0.51831
## Log(theta)          1.795e+01        NaN     NaN      NaN
##
## Zero-inflation model coefficients (binomial with logit link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.38026    0.01948   19.52   <2e-16 ***
## original_stars -2.18434    0.02824  -77.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 62194291.0775
## Number of iterations in BFGS optimization: 66
## Log-likelihood: -7.297e+04 on 19 Df
```

Notice again, the extremely high theta value suggests that the variance is very close to that of a Poisson distribution. The similarites in estimates and errors are thus predictable. Still, we'll complete the backward elimination before doing a proper comparison with the Poisson zero-inflated model.

```
##
## Call:
## zeroinfl(formula = TARGET ~ Alcohol + LabelAppeal + original_stars +
##     AcidIndex + Chlorides + Density + VolatileAcidity | original_stars,
##     data = train_data_imputed, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.1894 -0.5172  0.0179  0.4088  2.8725
##
## Count model coefficients (negbin with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.9341047  0.1087374   8.590  < 2e-16 ***
## Alcohol          0.0064161  0.0007475   8.583  < 2e-16 ***
```

```
## LabelAppeal-1    0.3882117  0.0214397  18.107  < 2e-16 ***
## LabelAppeal0     0.6625338  0.0209936  31.559  < 2e-16 ***
## LabelAppeal1     0.8562690  0.0213910  40.029  < 2e-16 ***
## LabelAppeal2     1.0194905  0.0239634  42.544  < 2e-16 ***
## original_stars   0.0974036  0.0027784  35.058  < 2e-16 ***
## AcidIndex       -0.0271432  0.0026596 -10.206  < 2e-16 ***
## Chlorides       -0.0273333  0.0087975  -3.107  0.00189 **
## Density         -0.3069899  0.1061500  -2.892  0.00383 **
## VolatileAcidity -0.0191950  0.0035470  -5.412 6.25e-08 ***
## Log(theta)      12.1996101  1.9602807   6.223 4.86e-10 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.38079    0.01947   19.56   <2e-16 ***
## original_stars -2.18488    0.02823  -77.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 198711.6575
## Number of iterations in BFGS optimization: 30
## Log-likelihood: -7.297e+04 on 14 Df
```

Note, these are *again* the same coefficients as in the zero-inflated Poisson–this despite the fact that we only used the p-values to guide our variable selection at this latter phase.

Table 12: Comparison of Zero-Inflated Poisson and Negatiive Binomial Models

Term	ZIP_Estimate	ZINB_Estimate	ZIP_Std_Error	ZINB_Std_Error
(Intercept)	0.9261	0.9255	0.1102	0.1102
Alcohol	0.0063	0.0063	0.0023	0.0023
I(Alcohol^2)	0.0000	0.0000	0.0001	0.0001
LabelAppeal-1	0.3886	0.3887	0.0214	0.0214
LabelAppeal0	0.6625	0.6626	0.0210	0.0210
LabelAppeal1	0.8558	0.8559	0.0214	0.0214
LabelAppeal2	1.0194	1.0195	0.0240	0.0240
original_stars	0.0977	0.0977	0.0028	0.0028
AcidIndex	-0.0269	-0.0269	0.0027	0.0027
Chlorides	-0.0268	-0.0268	0.0088	0.0088
Density	-0.3123	-0.3118	0.1062	0.1062
FreeSulfurDioxide	0.0000	0.0000	0.0000	0.0000
Sulphates	0.0002	0.0002	0.0030	0.0030
TotalSulfurDioxide	0.0000	0.0000	0.0000	0.0000
VolatileAcidity	-0.0192	-0.0192	0.0035	0.0035
pH	0.0026	0.0026	0.0041	0.0041
Log(theta)	NA	17.9458	NA	NaN

So again, we are looking at nearly identical statistics from the zero-inflated Poisson to the zero-inflated negative binomial. Also again, we should prefer the zero-inflated Poisson to the zero-inflated negative binomial, since the distrubtion appars to b eclose neough to a Poisson.

And what of the comparison beetwene the two ngative binomial models? Of course, it is going to look extremeely similar to the comparison btween the two Poisson models. Still, we add it below for sake of completeness:

```
## MAE Negative Binomial Model:  1.016387
```

```
## RMSE Negative Binomial Model:  1.27574
```

```
## MAE Zero-Inflated Negative Binomial Model:  0.9956453
```

```
## RMSE Zero-Inflated Negative Binomial Model:  1.28426
```

And indeed, the comparison is extremely similar as to earlier.

Multiple Linear Regression

We will look at a more direct approach with multiple linear regression using our normalized, transformed variables.

Looking at each predictor variable, we see that each predictor besides `FixedAcidity_transformed` are statistically significant within a 95% confidence level We also see that our $Adj.R^2 = 0.5405$, is where our model accounts on average for 54% of the variation of the `TARGET` variable.

```
##
## Call:
## lm(formula = as.numeric(as.character(TARGET)) ~ AcidIndex + FixedAcidity_transformed +
##     VolatileAcidity_transformed + CitricAcid_transformed + ResidualSugar_transformed +
##     Chlorides_transformed + FreeSulfurDioxide_transformed + TotalSulfurDioxide_transformed +
##     Density_transformed + pH_transformed + Sulphates_transformed +
##     Alcohol_transformed + STARS.1 + STARS.2 + STARS.3 + STARS.4 +
##     STARS.Unrated + LabelAppeal, data = train_data_prepped)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7226 -0.8468  0.0138  0.8399  6.1074
##
## Coefficients: (1 not defined because of singularities)
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.968728   0.051007  38.597  < 2e-16 ***
## AcidIndex                      -0.190274   0.004985 -38.169  < 2e-16 ***
## FixedAcidity_transformed        0.009494   0.006384   1.487 0.136974
## VolatileAcidity_transformed    -0.105261   0.006209 -16.954  < 2e-16 ***
## CitricAcid_transformed          0.032272   0.006216   5.192 2.09e-07 ***
## ResidualSugar_transformed       0.020849   0.006188   3.369 0.000754 ***
## Chlorides_transformed          -0.062869   0.006205 -10.132  < 2e-16 ***
## FreeSulfurDioxide_transformed   0.062972   0.006206  10.147  < 2e-16 ***
## TotalSulfurDioxide_transformed  0.066956   0.006205  10.790  < 2e-16 ***
## Density_transformed            -0.051618   0.006209  -8.314  < 2e-16 ***
## pH_transformed                 -0.040865   0.006203  -6.588 4.52e-11 ***
## Sulphates_transformed          -0.028562   0.006191  -4.613 3.97e-06 ***
## Alcohol_transformed             0.061323   0.006214   9.868  < 2e-16 ***
## STARS.1                         1.311245   0.017628  74.384  < 2e-16 ***
## STARS.2                         2.354115   0.017173 137.080  < 2e-16 ***
## STARS.3                         2.906541   0.020015 145.219  < 2e-16 ***
## STARS.4                         3.564148   0.031432 113.393  < 2e-16 ***
## STARS.Unrated                         NA         NA      NA       NA
```

```
## LabelAppeal-1                   0.413529   0.033736  12.258  < 2e-16 ***
## LabelAppeal0                    0.886686   0.032975  26.890  < 2e-16 ***
## LabelAppeal1                    1.375899   0.034432  39.960  < 2e-16 ***
## LabelAppeal2                    1.983104   0.045480  43.604  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.306 on 44774 degrees of freedom
## Multiple R-squared:  0.5407, Adjusted R-squared:  0.5405
## F-statistic:  2635 on 20 and 44774 DF,  p-value: < 2.2e-16
```

Calculating the RMSE for our training multiple regression model, we obtain an $RMSE = 1.3057$

```
## [1] 1.305704
```

Stepwise Regression

We will create a stepwise regression model, to perform forward and backward elimination on our multiple linear regression attempt.

The results show that we strictly removed the `STARS.Unrated` column which makes sense, as it does not provide additional information as the other STARS columns accounts for it. The $Adj.R^2 = 0.5405$ which stayed the same as before, and the RMSE stayed the same as well.

```
## Start:  AIC=23939.46
## as.numeric(as.character(TARGET)) ~ AcidIndex + FixedAcidity_transformed +
##     VolatileAcidity_transformed + CitricAcid_transformed + ResidualSugar_transformed +
##     Chlorides_transformed + FreeSulfurDioxide_transformed + TotalSulfurDioxide_transformed +
##     Density_transformed + pH_transformed + Sulphates_transformed +
##     Alcohol_transformed + STARS.1 + STARS.2 + STARS.3 + STARS.4 +
##     STARS.Unrated + LabelAppeal
##
##
## Step:  AIC=23939.46
## as.numeric(as.character(TARGET)) ~ AcidIndex + FixedAcidity_transformed +
##     VolatileAcidity_transformed + CitricAcid_transformed + ResidualSugar_transformed +
##     Chlorides_transformed + FreeSulfurDioxide_transformed + TotalSulfurDioxide_transformed +
##     Density_transformed + pH_transformed + Sulphates_transformed +
##     Alcohol_transformed + STARS.1 + STARS.2 + STARS.3 + STARS.4 +
##     LabelAppeal
##
##                                 Df Sum of Sq    RSS   AIC
## <none>                                       76369 23939
## - FixedAcidity_transformed       1         4  76373 23940
## - ResidualSugar_transformed      1        19  76389 23949
## - Sulphates_transformed          1        36  76406 23959
## - CitricAcid_transformed         1        46  76415 23964
## - pH_transformed                 1        74  76443 23981
## - Density_transformed            1       118  76487 24007
## - Alcohol_transformed            1       166  76535 24035
## - Chlorides_transformed          1       175  76544 24040
## - FreeSulfurDioxide_transformed  1       176  76545 24040
## - TotalSulfurDioxide_transformed 1       199  76568 24054
```

```
## - VolatileAcidity_transformed     1       490  76860 24224
## - AcidIndex                       1      2485  78854 25372
## - LabelAppeal                     4      7629  83999 28197
## - STARS.1                         1      9437  85807 29157
## - STARS.4                         1     21931  98301 35246
## - STARS.2                         1     32051 108420 39635
## - STARS.3                         1     35970 112339 41226

##
## Call:
## lm(formula = as.numeric(as.character(TARGET)) ~ AcidIndex + FixedAcidity_transformed +
##     VolatileAcidity_transformed + CitricAcid_transformed + ResidualSugar_transformed +
##     Chlorides_transformed + FreeSulfurDioxide_transformed + TotalSulfurDioxide_transformed +
##     Density_transformed + pH_transformed + Sulphates_transformed +
##     Alcohol_transformed + STARS.1 + STARS.2 + STARS.3 + STARS.4 +
##     LabelAppeal, data = train_data_prepped)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7226 -0.8468  0.0138  0.8399  6.1074
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.968728   0.051007  38.597  < 2e-16 ***
## AcidIndex                      -0.190274   0.004985 -38.169  < 2e-16 ***
## FixedAcidity_transformed        0.009494   0.006384   1.487 0.136974
## VolatileAcidity_transformed    -0.105261   0.006209 -16.954  < 2e-16 ***
## CitricAcid_transformed          0.032272   0.006216   5.192 2.09e-07 ***
## ResidualSugar_transformed       0.020849   0.006188   3.369 0.000754 ***
## Chlorides_transformed          -0.062869   0.006205 -10.132  < 2e-16 ***
## FreeSulfurDioxide_transformed   0.062972   0.006206  10.147  < 2e-16 ***
## TotalSulfurDioxide_transformed  0.066956   0.006205  10.790  < 2e-16 ***
## Density_transformed            -0.051618   0.006209  -8.314  < 2e-16 ***
## pH_transformed                 -0.040865   0.006203  -6.588 4.52e-11 ***
## Sulphates_transformed          -0.028562   0.006191  -4.613 3.97e-06 ***
## Alcohol_transformed             0.061323   0.006214   9.868  < 2e-16 ***
## STARS.1                         1.311245   0.017628  74.384  < 2e-16 ***
## STARS.2                         2.354115   0.017173 137.080  < 2e-16 ***
## STARS.3                         2.906541   0.020015 145.219  < 2e-16 ***
## STARS.4                         3.564148   0.031432 113.393  < 2e-16 ***
## LabelAppeal-1                   0.413529   0.033736  12.258  < 2e-16 ***
## LabelAppeal0                    0.886686   0.032975  26.890  < 2e-16 ***
## LabelAppeal1                    1.375899   0.034432  39.960  < 2e-16 ***
## LabelAppeal2                    1.983104   0.045480  43.604  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.306 on 44774 degrees of freedom
## Multiple R-squared:  0.5407, Adjusted R-squared:  0.5405
## F-statistic:  2635 on 20 and 44774 DF,  p-value: < 2.2e-16

## [1] 1.305704
```

Model Selection

Let's now look at all the models compared to each other based on their results with using the test data. The best model in comparison to the others is the original theory of using Poisson with the best $R^2 = 0.5611089$ and the lowest $RMSE = 1.2757393$. We could go with the basis of the AIC metric, however, our goal is to predict the best results with our evaluation set. Even with all the transformations, interactions and trying to handle zero inflation factors, a Poisson model is our best choice to predict on the evaluation set.

Again AIC would be a better choice if we wanted to have more of an inference between the models.

Table 13: Model Performance

	Poisson	ZI_Poisson	Neg_Binom	ZI_Neg_Binom	Stepwise
RMSE	1.2757393	1.2842321	1.2757404	1.2842601	1.2889856
Rsquared	0.5611089	0.5554209	0.5611082	0.5554045	0.5519412
MAE	1.0163856	0.9956513	1.0163873	0.9956453	1.0208232
AIC	159979.7702125	145969.2867016	159983.1967400	145971.8130167	151064.1672811

Predictions

Predictions

As we have now selected our models, we are ready to make predictions on the evaluation set. This is a slightly complicated process because our second model is dependent on our first one. We complete this process below:

Table 14: Preview: Predictions for Evaluation Dataset

TARGET
3
4
2
2
2
5
3
5
1
3

Conclusion:

In this project, we aimed to develop predictive models for wine sales using statistical techniques and machine learning algorithms. We started by exploring the data, handling missing values, and transforming variables. Initially, we experimented with Poisson regression and zero-inflated Poisson models, as well as negative binomial regression to address over-dispersion. However, the simple Poisson model emerged as the best performer. We refined our models through stepwise regression but found no significant improvement over the simple Poisson model. Overall, our models offer valuable insights for wine producers and distributors, aiding in resource allocation and marketing strategy adjustments to optimize sales and enhance profitability in the wine industry.