UNITED STATES MILITARY ACADEMY

HOMEWORK 3

MA478: GENERALIZED LINEAR MODELS

HOUR H2

COL NICHOLAS CLARK

By

CADET JOSHUA WONG ‘24, CO I3

WEST POINT, NEW YORK

07 MARCH 2024

JW______ I CERTIFY THAT I HAVE COMPLETELY DOCUMENTED ALL SOURCES THAT I USED TO COMPLETE THIS ASSIGNMENT AND THAT I ACKNOWLEDGED ALL ASSISTANCE I RECEIVED IN THE COMPLETION OF THIS ASSIGNMENT.

______ I CERTIFY THAT I DID NOT USE ANY SOURCES OR RECEIVE ANY ASSISTANCE REQUIRING DOCUMENTATION WHILE COMPLETING THIS ASSIGNMENT.

SIGNATURE: ____

MA478 Homework 3: Predictive Wine Models

CDT Joshua Wong

1 Introduction

In the dynamic landscape of the wine industry, where consumer preferences and market trends constantly evolve, the ability to predict and strategically respond to sales demands becomes paramount for wine manufacturers. This study aims to develop a count regression model to predict the number of wine cases that will be sold, given various information about the wine by leveraging a dataset encompassing approximately 12,000 commercially available wines. The motivation behind this endeavor lies in the need for large wine manufacturers to adapt their offerings to maximize sales and profit, particularly in the context of high-end restaurant placements. The response variable of interest, TARGET, is the number of sample cases purchased by distribution companies after sampling a wine, serving as a crucial indicator of a wine's market success. With a comprehensive set of variables, including chemical properties and sensory attributes, this paper delves into the intricacies of constructing a predictive model. The ensuing analysis not only explores the specificities of wine characteristics but also addresses the challenge posed by missing data, considering its potential predictive value in enhancing the accuracy of sales predictions.

2 Data Exploration

The dataset we used contained 12795 observations and 15 different variables. In addition to the response variable TARGET, which represents the number of cases of wine sold, 14 other potential explanatory variables could be used to predict the response. We provide a list of all of the variables in the dataset along with a definition for some a theoretical effect in Table 1. Additionally, Table 2 provides the summary statistics for the data to include the minimum, maximum, 1st, and 3rd quartiles, median, mean, and the number of null values found concerning each variable in the data. In the dataset, almost all variables are represented as numerical values, except for the variables TARGET, LabelAppeal, AcidIndex, and STARS, which are represented as integers. We decided to preserve the original data representations as all the numeric values were correctly represented as continuous numeric variables. At the same time, we kept the integer variables as integers to preserve their order while preserving their discrete nature.

Our first step in exploring the data involved looking at the distribution of the response variable. The response variable distribution shown in Figure 1 revealed a significant number of zero values; however, excluding the zeros, the data looked approximately normally distributed. We plotted each potential explanatory variable against our response to see any initial trends or relationships. We found only two potentially significant relationships by looking at the scatter plots. The first relationship was between the response and the LabelAppeal variable, illustrated in Figure 2. The second relationship is illustrated in Figure 3, which depicts the relationship between the variable

STARS and the response. Both cases demonstrated a positive relationship, meaning that the number of wine cases sold was positively associated with increased label appeal and star rating.

Finally, we constructed a correlation matrix to assess additional potential associations between all the dataset variables and potentially address any co-linearity. The correlation matrix in Figure 4 has similar positive associations between the response and the variables LabelAppeal and STARS. Additionally, the correlation matrix reveals a negative correlation between the variable AcidIndex and the response. There also seems to be a positive relationship between variables LabelAppeal and STARS. However, the colinearity between LabelAppeal and STARS variables is not addressed in our current models. We encourage such exploration in future work.

3 Data Preparation

In preparing and transforming the dataset, we applied a systematic approach to enhance the quality of the data and the potential for meaningful analysis. The primary transformation involved imputing the missing values in the dataset. Table 2 depicts the variables that contained missing values. As seen in Table 2, the only integer data type with missing values was the variable STARS; all the other missing values were for numerical data types. Therefore, for all numeric data types, we imputed each missing value in a column with the mean of the column. We chose this method to maintain the integrity of the dataset while addressing the absence of specific data points, ensuring a comprehensive representation of numerical features while avoiding increasing bias.

We adopted a strategy of imputing zeros to address missing values in the variable STARS. We grounded our decision in the assumption that wines not rated by a team of experts are likely perceived as having lower quality or significance than those that underwent evaluation. The rationale stems from the premise that wines deserving of a rating would naturally attract expert attention, and the absence of a rating may indicate that the wine does not meet the criteria for consideration or recognition. Consequently, assigning a zero rating to unrated wines aligns with the notion that these wines are likely perceived to have lesser appeal or merit in the absence of a formal assessment. This imputation strategy acknowledges the inherent bias introduced by missing values. It seeks to pragmatically incorporate this information into the predictive model, aligning with the underlying wine quality and market reception assumptions.

4 Build Models

This study explores two Poisson regression models, two Negative Binomial models, two Zero-Inflated Poisson models, and two Linear regression models. It compares them to each other to find the best model to predict the target response variable. The variable selection process involves an initial consideration of explanatory variables based on domain knowledge, followed by refinement using ChatGPT. Each model, including the most influential variables, is prioritized, drawing on domain expertise and ChatGPT's language model capabilities to identify top predictors for the response variable. The variables we identified as undoubtedly significant and must be included in the model were the variables LabelAppeal, AcidIndex, and STARS. After consulting ChatGPT, the second version of each of our models additionally considered the variables Alcohol, CitricAcid, FreeSulfurDioxide, and ResidualSugar. To test model robustness and generalizability, we partitioned the dataset into training and test sets, with a random 20% assigned to the test set and the remaining 80% to the training set. This approach facilitates comprehensive model evaluation on unseen data, aiming to advance count regression modeling insights and enhance the accuracy of predicting wine sales based on key characteristics.

4.1 Poisson Regression Models

We began by creating our two Poisson regression models with the above-mentioned variables. Our first and simpler Poisson regression model is formulated in Equation 1. The second larger Poisson regression model is formulated in Equation 2. The coefficients for the models can be found in Tables 3 and 4, respectively. We compared the two models using a Chi-squared ANOVA test, where we tested if the difference in deviance between the models justified adding the extra variables. The Chi-squared ANOVA test provided a p-value < 0.01, meaning we reject the null hypothesis and prefer the more complex model over the simpler one. Additionally, we conducted a Chi-squared goodness-of-fit test on both our Poisson regression models. For both cases, we found a small p-value indicating that we rejected the null hypothesis and determined that both our models are a bad fit for the data, failing the Chi-squared goodness-of-fit test. After looking at the half-norm plot of the residuals and determining that there were no significant outliers in the data, we estimate the value of ϕ to check for over-dispersion of our residuals. We estimate the value of ϕ by taking the deviance of each model and dividing it by their residual degrees of freedom. In both cases, we get ϕ value of approximately 1.16. Therefore, we concluded that because our estimates for ϕ are not far from 1, our models are most likely not over-dispersed.

4.2 Negative Binomial Regression Models

We also try to use two negative binomial regression models. However, because of the lack of over-dispersion indicated in the Poisson regression, it was unlikely that a negative binomial would be a significantly better fit than the Poisson regression from the start. Nevertheless, we create two negative binomial regression models with the same covariates in the first and second models as the first and second Poisson regression models. As with the Poisson regression models we used a Chi-squared ANOVA test to determine which model we prefer. As with the Poisson models, we get a small p-value < 0.01, indicating that we prefer the larger model, model 2, to the smaller model, model 1. The mathematical formulation of the first negative binomial regression model is in Equation 3, and the second model is in Equation 4 with coefficients in Tables 5 and 6 respectively. Within a binomial regression model, we estimate a parameter κ. Estimating κ means we cannot use deviance to conduct a goodness-of-fit test on these negative binomial models as we did with our Poisson regression model.

4.3 Zero Inflated Poisson Regression Models

During the initial data exploration phase, we noted that the response variable contained many zeros. Without the zeros, the response seemed to take on a different normal distribution. Separating a larger number of zeros from another distribution within the response variable indicates a zero-inflated Poisson (ZIP) model. The intuition behind a ZIP model is that two different processes are at play: one that makes zeros and one that follows a separate Poisson distribution. In this case, the distribution of the response looked like it could have come from two different distributions. Moreover, when simulating data using the Poisson regression models and comparing the count of simulated zeros to actual zeros, there was a significant difference above a factor of two. We hypothesize a reasonable conceptualization that could explain the ZIP mechanism in the data.

We hypothesize that the prevalence of zeros in the response variable, representing the number of cases of wine sold, could be attributed to market dynamics and consumer behavior. It is plausible that a substantial portion of the dataset comprises wines that are less commercially popular or niche products, leading to instances where these wines might not have been purchased or sampled by wine distribution companies. Factors such as varying consumer tastes, regional preferences, or specific varietal characteristics may contribute to certain wines having limited market appeal. Additionally, the dataset may include newly introduced or experimental wines yet to gain traction in the market, resulting in many wines with zero cases sold. This hypothesis aligns with the notion that consumer demand and market trends play a pivotal role in the sales performance of wines, with some products experiencing lower commercial success than others.

Based on our hypothesis and the observations gained from the distribution of our response variable, we decided to fit two ZIP models to the data, using the same variables as with the negative binomial and the Poisson regression. The formulation for our first ZIP model is in Equation 5, and the formulation for our second ZIP model is in Equation 6 with coefficients in Tables 7 and 6 respectively. Again, we compare our two ZIP models using a Chi-squared test of deviance by calculating the log-likelihood of each model and again find that we prefer the larger model over the smaller model based on p-value < 0.01. It is also noteworthy that the ZIP model was the only model that showed pH did not have a significant negative relationship with the response. This could be because having a high pH level is popular among the experimental wines that sell less than the well-known wines.

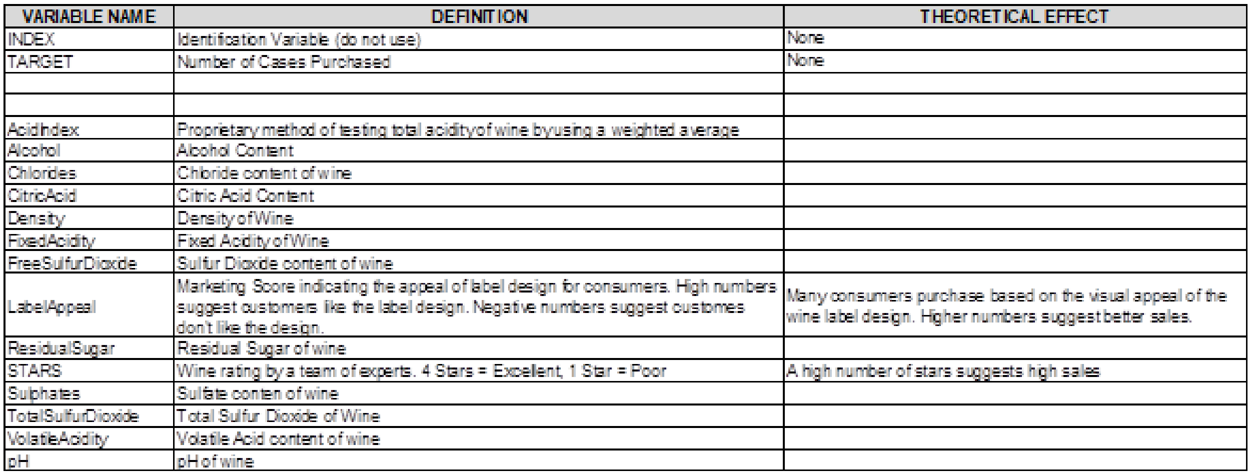
4.4 Linear Regression Models

We compared our linear regression models using an F-test, finding again that our second model is preferred over our first simpler model with a p-value < 0.01. The results and associations found in the linear regression are consistent with those of the Poisson regression and negative binomial, meaning that the variables AcidIndex and pH were the only values with a negative association with the response. The formulations for the linear regression models are in Equations 7 and 8, with coefficient values in Tables 9 and 10.

5 Model Selection

We looked at each model's mean squared error (MSE) when predicting our test and training sets to select our preferred model. We also used the AIC and Log likelihood of our models to compare them to each other. Table 11 we compare the metrics of our four different models. In all cases, we used the more complex model that was preferred to compare different types of non-nested models. Based on the information in Table 11, we prefer the ZIP model over the other three due to its lower AIC and MSE for the training and testing sets and the higher log-likelihood compared to the other three models. We should also not discount the Liner Regression as it does not account for count data, which is the response variable in this case. We use the ZIP model to subsequently predict the response variable fo the evaluation data set. The head of the eval dataset with the predicted TARGET values is in Table 12.

Appendix A: Figures, Tables and Equations



VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET	Number of Cases Purchased	None
AcidIndex	Proprietary method of testing total acidityof wine byusing a weighted average	
Alcohol	Alcohol Content	
Chlorides	Chloride content of wine	
CitricAcid	Citric Acid Content	
Density	Density of Wine	
FixedAcidity	Fixed Acidity of Wine	
FreeSulfurDioxide	Sulfur Dioxide content of wine	
LabelAppeal	Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customes don't like the design.	Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales.
ResidualSugar	Residual Sugar of wine	
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor	A high number of stars suggests high sales
Sulphates	Sulfate conten of wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	

Table 1: Dataset Variables

Variable	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**NAs**
`TARGET`	0.000	2.000	3.000	3.029	4.000	8.000	-
`FixedAcidity`	-18.100	5.200	6.900	7.076	9.500	34.400	-
`VolatileAcidity`	-2.7900	0.1300	0.2800	0.3241	0.6400	3.6800	-
`CitricAcid`	-3.2400	0.0300	0.3100	0.3084	0.5800	3.8600	-
`ResidualSugar`	-127.800	-2.000	3.900	5.419	15.900	141.150	616
`Chlorides`	-1.1710	-0.0310	0.0460	0.0548	0.1530	1.3510	638
`FreeSulfurDioxide`	-555.00	0.00	30.00	30.85	70.00	623.00	647
`TotalSulfurDioxide`	-823.0	27.0	123.0	120.7	208.0	1057.0	682
`Density`	0.8881	0.9877	0.9945	0.9942	1.0005	1.0992	-
`pH`	0.480	2.960	3.200	3.208	3.470	6.130	395
`Sulphates`	-3.1300	0.2800	0.5000	0.5271	0.8600	4.2400	1210
`Alcohol`	-4.70	9.00	10.40	10.49	12.40	26.50	653
`LabelAppeal`	-2.000	-1.000	0.000	-0.0091	1.000	2.000	-
`AcidIndex`	4.000	7.000	8.000	7.773	8.000	17.000	-
`STARS`	1.000	1.000	2.000	2.042	3.000	4.000	3359

Table 2: Summary Statistics

$$\begin{aligned}\log(TARGET_i) = \beta_0 + \beta_{\text{LabelAppeal}} \cdot \text{LabelAppeal}_i + \\ \beta_{\text{AcidIndex}} \cdot \text{AcidIndex}_i + \beta_{\text{STARS}} \cdot \text{STARS}_i\end{aligned} \tag{1}$$

$$\begin{aligned}\log(TARGET_i) = \beta_0 + \beta_{\text{AcidIndex}} \cdot \text{AcidIndex}_i + \beta_{\text{Alcohol}} \cdot \text{Alcohol}_i + \\ \beta_{\text{CitricAcid}} \cdot \text{CitricAcid}_i + \beta_{\text{FreeSulfurDioxide}} \cdot \text{FreeSulfurDioxide}_i + \\ \beta_{\text{ResidualSugar}} \cdot \text{ResidualSugar}_i + \beta_{\text{STARS}} \cdot \text{STARS}_i + \beta_{\text{pH}} \cdot \text{pH}_i + \\ \beta_{\text{LabelAppeal}} \cdot \text{LabelAppeal}_i\end{aligned} \tag{2}$$

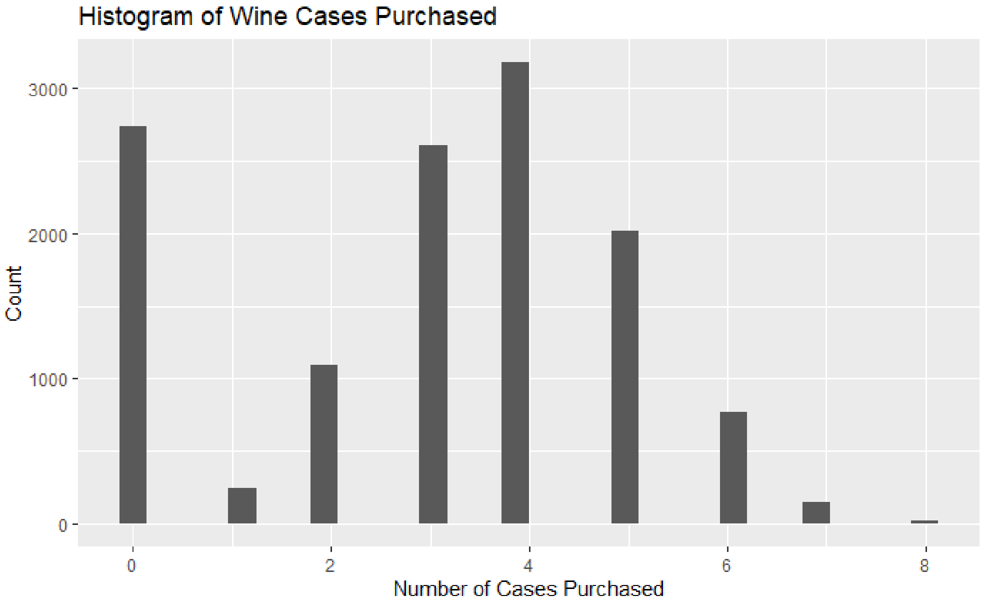


Figure 1: Histogram of Number of Wine Cases Purchased

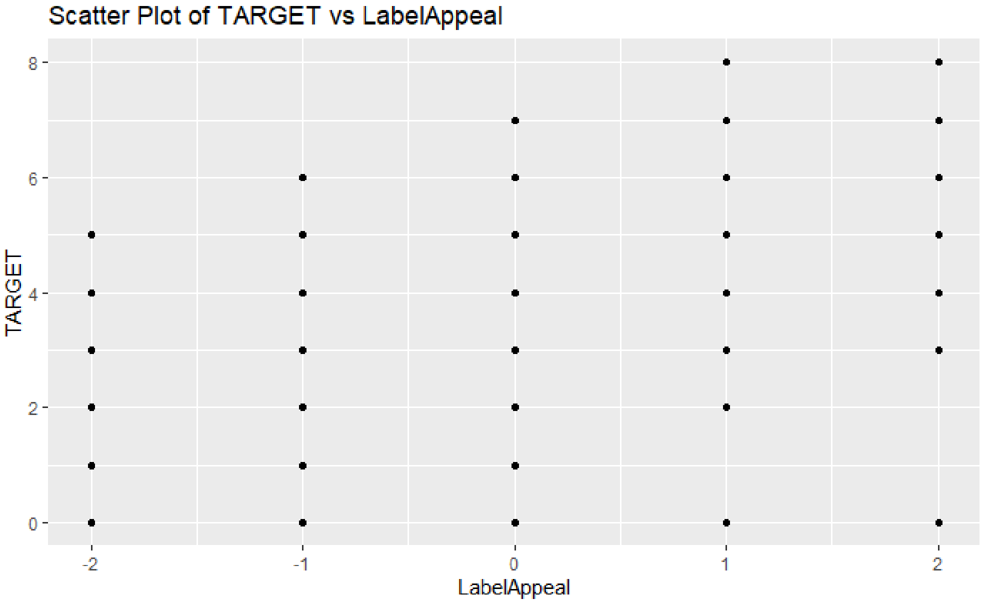


Figure 2: Scatter Plot of Wine Cases Purchased VS Label Appeal Rating

$$\log(TARGET_i) = \beta_0 + \beta_1 \cdot \text{LabelAppeal}_i + \beta_2 \cdot \text{AcidIndex}_i + \beta_3 \cdot \text{STARS}_i \tag{3}$$

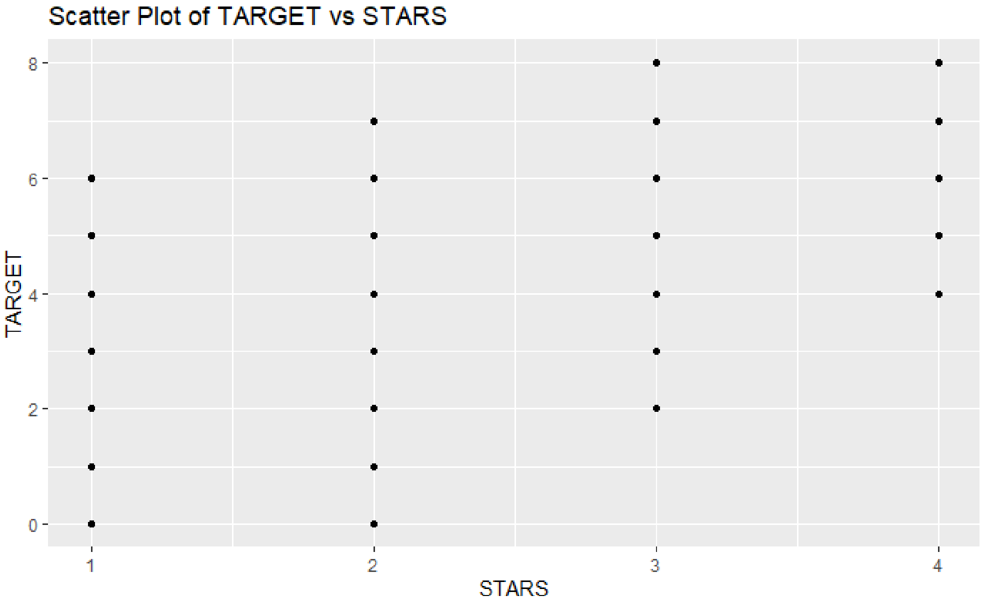


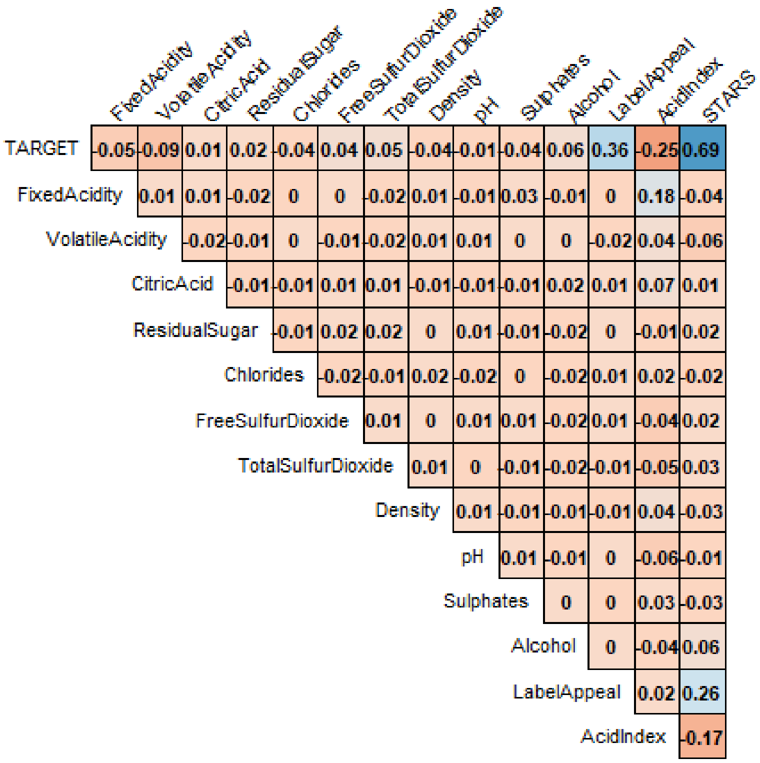
Figure 3: Scatter Plot of Wine Cases Purchased VS Star Rating

$$\begin{aligned}\log(TARGET_i) = &\beta_0 + \beta_1 \cdot \text{AcidIndex}_i + \beta_2 \cdot \text{Alcohol}_i \\ &+ \beta_3 \cdot \text{CitricAcid}_i + \beta_4 \cdot \text{FreeSulfurDioxide}_i \\ &+ \beta_5 \cdot \text{ResidualSugar}_i + \beta_6 \cdot \text{STARS}_i \\ &+ \beta_7 \cdot \text{pH}_i + \beta_8 \cdot \text{LabelAppeal}_i\end{aligned} \tag{4}$$

$$Y_i = \begin{cases} \text{logit}(TARGET_i) & \text{with probability } \phi \\ \log(TARGET_i) & \text{with probability } 1 - \phi \end{cases} \tag{5a}$$

$$\text{logit}(TARGET_i) = \beta_0 + \beta_{\text{LabelAppeal}} \cdot \text{LabelAppeal}_i + \beta_{\text{AcidIndex}} \cdot \text{AcidIndex}_i + \beta_{\text{STARS}} \cdot \text{STARS}_i \tag{5b}$$

$$\log(TARGET_i) = \beta_0 + \beta_{\text{LabelAppeal}} \cdot \text{LabelAppeal}_i + \beta_{\text{AcidIndex}} \cdot \text{AcidIndex}_i + \beta_{\text{STARS}} \cdot \text{STARS}_i \tag{5c}$$



	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
TARGET	-0.05	-0.09	0.01	0.02	-0.04	0.04	0.05	-0.04	-0.01	-0.04	0.06	0.36	-0.25	0.69
FixedAcidity		0.01	0.01	-0.02	0	0	-0.02	0.01	-0.01	0.03	-0.01	0	0.18	-0.04
VolatileAcidity			-0.02	-0.01	0	-0.01	-0.02	0.01	0.01	0	0	-0.02	0.04	-0.06
CitricAcid				-0.01	-0.01	0.01	0.01	-0.01	-0.01	-0.01	0.02	0.01	0.07	0.01
ResidualSugar					-0.01	0.02	0.02	0	0.01	-0.01	-0.02	0	-0.01	0.02
Chlorides						-0.02	-0.01	0.02	-0.02	0	-0.02	0.01	0.02	-0.02
FreeSulfurDioxide							0.01	0	0.01	0.01	-0.02	0.01	-0.04	0.02
TotalSulfurDioxide								0.01	0	-0.01	-0.02	-0.01	-0.05	0.03
Density									0.01	-0.01	-0.01	-0.01	0.04	-0.03
pH										0.01	-0.01	0	-0.06	-0.01
Sulphates											0	0	0.03	-0.03
Alcohol												0	-0.04	0.06
LabelAppeal													0.02	0.26
AcidIndex														-0.17

Figure 4: Correlation Matrix (All Data)

$$Y_i = \begin{cases} \text{logit}(TARGET_i) & \text{with probability } \phi \\ \log(TARGET_i) & \text{with probability } 1-\phi \end{cases} \tag{6a}$$

$$\text{logit}(TARGET_i) = \beta_0 + \beta_{\text{AcidIndex}} \cdot \text{AcidIndex}_i + \beta_{\text{Alcohol}} \cdot \text{Alcohol}_i + \beta_{\text{CitricAcid}} \cdot \text{CitricAcid}_i \tag{6b}$$

$$+ \beta_{\text{FreeSulfurDioxide}} \cdot \text{FreeSulfurDioxide}_i + \beta_{\text{ResidualSugar}} \cdot \text{ResidualSugar}_i \tag{6c}$$

$$+ \beta_{\text{STARS}} \cdot \text{STARS}_i + \beta_{\text{pH}} \cdot \text{pH}_i + \beta_{\text{LabelAppeal}} \cdot \text{LabelAppeal}_i \tag{6d}$$

$$\log(TARGET_i) = \beta_0 + \beta_{\text{AcidIndex}} \cdot \text{AcidIndex}_i + \beta_{\text{Alcohol}} \cdot \text{Alcohol}_i + \beta_{\text{CitricAcid}} \cdot \text{CitricAcid}_i \tag{6e}$$

$$+ \beta_{\text{FreeSulfurDioxide}} \cdot \text{FreeSulfurDioxide}_i + \beta_{\text{ResidualSugar}} \cdot \text{ResidualSugar}_i \tag{6f}$$

$$+ \beta_{\text{STARS}} \cdot \text{STARS}_i + \beta_{\text{pH}} \cdot \text{pH}_i + \beta_{\text{LabelAppeal}} \cdot \text{LabelAppeal}_i \tag{6g}$$

$$\begin{aligned} TARGET_i = {} & \beta_0 + \beta_{\text{LabelAppeal}} \times \text{LabelAppeal}_i + \beta_{\text{AcidIndex}} \times \text{AcidIndex}_i \\ & + \beta_{\text{STARS}} \times \text{STARS}_i + \epsilon_i \end{aligned} \tag{7}$$

$$\begin{aligned} TARGET_i = {} & \beta_0 + \beta_{\text{AcidIndex}} \times \text{AcidIndex}_i + \beta_{\text{Alcohol}} \times \text{Alcohol}_i + \beta_{\text{CitricAcid}} \times \text{CitricAcid}_i \\ & + \beta_{\text{FreeSulfurDioxide}} \times \text{FreeSulfurDioxide}_i + \beta_{\text{ResidualSugar}} \times \text{ResidualSugar}_i \\ & + \beta_{\text{STARS}} \times \text{STARS}_i + \beta_{\text{pH}} \times \text{pH}_i + \beta_{\text{LabelAppeal}} \times \text{LabelAppeal}_i + \epsilon_i \end{aligned} \tag{8}$$

Variable	**Estimate**	**Std. Error**	**z value**	**Pr$(>\lvert z\rvert)$**
(Intercept)	1.216941	0.040878	29.77	$< 2 \times 10^{-16}$ ***
LabelAppeal	0.131000	0.006772	19.34	$< 2 \times 10^{-16}$ ***
AcidIndex	-0.088417	0.004995	-17.70	$< 2 \times 10^{-16}$ ***
STARS	0.314971	0.005050	62.37	$< 2 \times 10^{-16}$ ***

Table 3: Poisson Model 1 Coefficients

Variable	**Estimate**	**Std. Error**	**z value**	**Pr$(>\lvert z\rvert)$**
(Intercept)	1.284e+00	5.399e-02	23.777	$< 2 \times 10^{-16}$ ***
AcidIndex	-8.932e-02	5.025e-03	-17.774	$< 2 \times 10^{-16}$ ***
Alcohol	7.831e-04	1.585e-03	0.494	0.62127
CitricAcid	1.027e-02	6.612e-03	1.554	0.12023
FreeSulfurDioxide	1.250e-04	3.935e-05	3.177	0.00149 **
ResidualSugar	1.750e-04	1.723e-04	1.016	0.30975
STARS	3.144e-01	5.066e-03	62.060	$< 2 \times 10^{-16}$ ***
pH	-2.356e-02	8.502e-03	-2.771	0.00559 **
LabelAppeal	1.310e-01	6.776e-03	19.327	$< 2 \times 10^{-16}$ ***

Table 4: Poisson Model 2 Coefficients

Variable	**Estimate**	**Std. Error**	**z value**	**Pr$(>\lvert z\rvert)$**
(Intercept)	1.216949	0.040879	29.77	$< 2 \times 10^{-16}$ ***
LabelAppeal	0.130999	0.006772	19.34	$< 2 \times 10^{-16}$ ***
AcidIndex	-0.088419	0.004995	-17.70	$< 2 \times 10^{-16}$ ***
STARS	0.314975	0.005050	62.37	$< 2 \times 10^{-16}$ ***

Table 5: Negative Binomial Model 1 Coefficients

Variable	**Estimate**	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	1.284e+00	5.399e-02	23.776	$< 2 \times 10^{-16}$ ***
AcidIndex	-8.932e-02	5.025e-03	-17.774	$< 2 \times 10^{-16}$ ***
Alcohol	7.830e-04	1.585e-03	0.494	0.62132
CitricAcid	1.027e-02	6.612e-03	1.554	0.12024
FreeSulfurDioxide	1.250e-04	3.936e-05	3.176	0.00149 **
ResidualSugar	1.750e-04	1.723e-04	1.016	0.30975
STARS	3.144e-01	5.066e-03	62.058	$< 2 \times 10^{-16}$ ***
pH	-2.356e-02	8.502e-03	-2.771	0.00559 **
LabelAppeal	1.310e-01	6.776e-03	19.326	$< 2 \times 10^{-16}$ ***

Table 6: Negative Binomial Model 2 Coefficients

Count Model (Poisson with Log Link)				
Variable	**Estimate**	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	1.251355	0.043502	28.766	$< 2 \times 10^{-16}$ ***
LabelAppeal	0.233554	0.007045	33.151	$< 2 \times 10^{-16}$ ***
AcidIndex	-0.019135	0.005422	-3.529	0.000417 ***
STARS	0.101446	0.005812	17.454	$< 2 \times 10^{-16}$ ***

Zero-Inflation Model (Binomial with Logit Link)				
Variable	**Estimate**	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	-2.89787	0.22328	-12.98	$< 2 \times 10^{-16}$ ***
LabelAppeal	0.72412	0.04709	15.38	$< 2 \times 10^{-16}$ ***
AcidIndex	0.42667	0.02766	15.42	$< 2 \times 10^{-16}$ ***
STARS	-2.34647	0.06500	-36.10	$< 2 \times 10^{-16}$ ***

Table 7: ZIP Model 1 Coefficients

Count Model (Poisson with Log Link)				
Variable	**Estimate**	**Std. Error**	**z value**	**Pr($> \|z\|$)**
(Intercept)	1.172e+00	5.681e-02	20.634	$< 2 \times 10^{-16}$***
AcidIndex	-1.778e-02	5.447e-03	-3.264	0.0011**
Alcohol	6.750e-03	1.623e-03	4.159	3.19e-05***
CitricAcid	-1.464e-04	6.755e-03	-0.022	0.9827
FreeSulfurDioxide	2.938e-05	3.953e-05	0.743	0.4574
ResidualSugar	-2.547e-05	1.769e-04	-0.144	0.8855
STARS	9.945e-02	5.831e-03	17.054	$< 2 \times 10^{-16}$***
pH	2.559e-04	8.757e-03	0.029	0.9767
LabelAppeal	2.339e-01	7.050e-03	33.181	$< 2 \times 10^{-16}$***

Zero-Inflation Model (Binomial with Logit Link)				
Variable	**Estimate**	**Std. Error**	**z value**	**Pr($> \|z\|$)**
(Intercept)	-4.1525	0.3274	-12.685	$< 2 \times 10^{-16}$***
AcidIndex	0.4404	0.0280	15.725	$< 2 \times 10^{-16}$***
Alcohol	0.0355	0.0105	3.373	0.000742***
CitricAcid	-0.0608	0.0438	-1.389	0.164949
FreeSulfurDioxide	-0.0006	0.0003	-2.268	0.023305*
ResidualSugar	-0.0018	0.0011	-1.549	0.121473
STARS	-2.3536	0.0652	-36.077	$< 2 \times 10^{-16}$***
pH	0.2565	0.0561	4.571	4.85×10^{-6}***
LabelAppeal	0.7292	0.0473	15.429	$< 2 \times 10^{-16}$***

Table 8: ZIP Model 2 Coefficients

Variable	**Estimate**	**Std. Error**	**t value**	**Pr($> \|t\|$)**
(Intercept)	3.18602	0.08469	37.62	$< 2 \times 10^{-16}$ ***
LabelAppeal	0.42516	0.01536	27.68	$< 2 \times 10^{-16}$ ***
AcidIndex	-0.21174	0.01011	-20.95	$< 2 \times 10^{-16}$ ***
STARS	0.98821	0.01174	84.20	$< 2 \times 10^{-16}$ ***

Table 9: Multiple Linear Regression Model 1 Coefficients

Variable	**Estimate**	**Std. Error**	**t value**	**Pr($> \|t\|$)**
(Intercept)	3.302e+00	1.163e-01	28.386	$< 2 \times 10^{-16}$ ***
AcidIndex	-2.133e-01	1.014e-02	-21.025	$< 2 \times 10^{-16}$ ***
Alcohol	6.788e-03	3.643e-03	1.863	0.062460 .
CitricAcid	2.754e-02	1.530e-02	1.799	0.071986 .
FreeSulfurDioxide	3.273e-04	9.104e-05	3.595	0.000326 ***
ResidualSugar	5.535e-04	3.975e-04	1.392	0.163874
STARS	9.852e-01	1.175e-02	83.880	$< 2 \times 10^{-16}$ ***
pH	-5.994e-02	1.963e-02	-3.053	0.002268 **
LabelAppeal	4.255e-01	1.534e-02	27.730	$< 2 \times 10^{-16}$ ***

Table 10: Multiple Linear Regression Model 2 Coefficients

Table 11: Model Comparison Metrics

Metric	**Poisson**	**Negative Binomial**	**ZIP**	**Linear Regression**
Train MSE	6.74474	6.74473	1.64248	1.77162
Test MSE	6.83054	6.83054	1.61981	1.73104
LogLik	-18704.06 (df=9)	-18704.17 (df=10)	-16335.56 (df=18)	N/A
AIC	37426.12	37428.34	32707.12	34922.45

	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide
1	1.5723395	5.4	-0.860	0.27	-10.7	0.092	23
2	3.9103667	12.4	0.385	-0.76	-19.7	1.169	-37
3	2.6631347	7.2	1.750	0.17	-33.0	0.065	9
4	2.4903425	6.2	0.100	1.80	1.0	-0.179	104
5	0.7461585	11.4	0.210	0.28	1.2	0.038	70
6	5.6513043	17.6	0.040	-1.15	1.4	0.535	-250

Table 12: Head of Eval Data Predictions

MA478 HW3

CDT Joshua Wong

2024-03-03

```
# Load Packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the ▯]8;;http://conflicted.r-lib.org/▯conflicted package▯]8;;▯ to force all conflicts t
o become errors
```

```
library(faraway)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

Data Exploration

```
# Load Data
wine_train <- read.csv("wine-training-data.csv")
wine_eval <- read.csv("wine-evaluation-data.csv")

# Remove Index Column
wine_train$ï..INDEX <- NULL
wine_eval$IN <- NULL

# Note: We will keep all
summary(wine_train)
```

```
##      TARGET       FixedAcidity     VolatileAcidity     CitricAcid
##  Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400
##  1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300
##  Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100
##  Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084
##  3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800
##  Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600
##
##  ResidualSugar        Chlorides       FreeSulfurDioxide TotalSulfurDioxide
##  Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0
##  1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0
##  Median :   3.900   Median : 0.0460   Median :  30.00   Median : 123.0
##  Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7
##  3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0
##  Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0
##  NA's   :616        NA's   :638       NA's   :647       NA's   :682
##     Density             pH          Sulphates          Alcohol
##  Min.   :0.8881   Min.   :0.480   Min.   :-3.1300   Min.   :-4.70
##  1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00
##  Median :0.9945   Median :3.200   Median : 0.5000   Median :10.40
##  Mean   :0.9942   Mean   :3.208   Mean   : 0.5271   Mean   :10.49
##  3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40
##  Max.   :1.0992   Max.   :6.130   Max.   : 4.2400   Max.   :26.50
##                   NA's   :395     NA's   :1210      NA's   :653
##   LabelAppeal          AcidIndex          STARS
##  Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##                                       NA's   :3359
```

```
# No Categorical Variables, Keep all as "numeric" or "integer"
sapply(wine_train, class)
```

```
##             TARGET       FixedAcidity    VolatileAcidity         CitricAcid 
##          "integer"          "numeric"          "numeric"          "numeric" 
##      ResidualSugar          Chlorides  FreeSulfurDioxide TotalSulfurDioxide 
##          "numeric"          "numeric"          "numeric"          "numeric" 
##            Density                 pH          Sulphates            Alcohol 
##          "numeric"          "numeric"          "numeric"          "numeric" 
##        LabelAppeal          AcidIndex              STARS 
##          "integer"          "integer"          "integer"
```

```
names(wine_train)
```

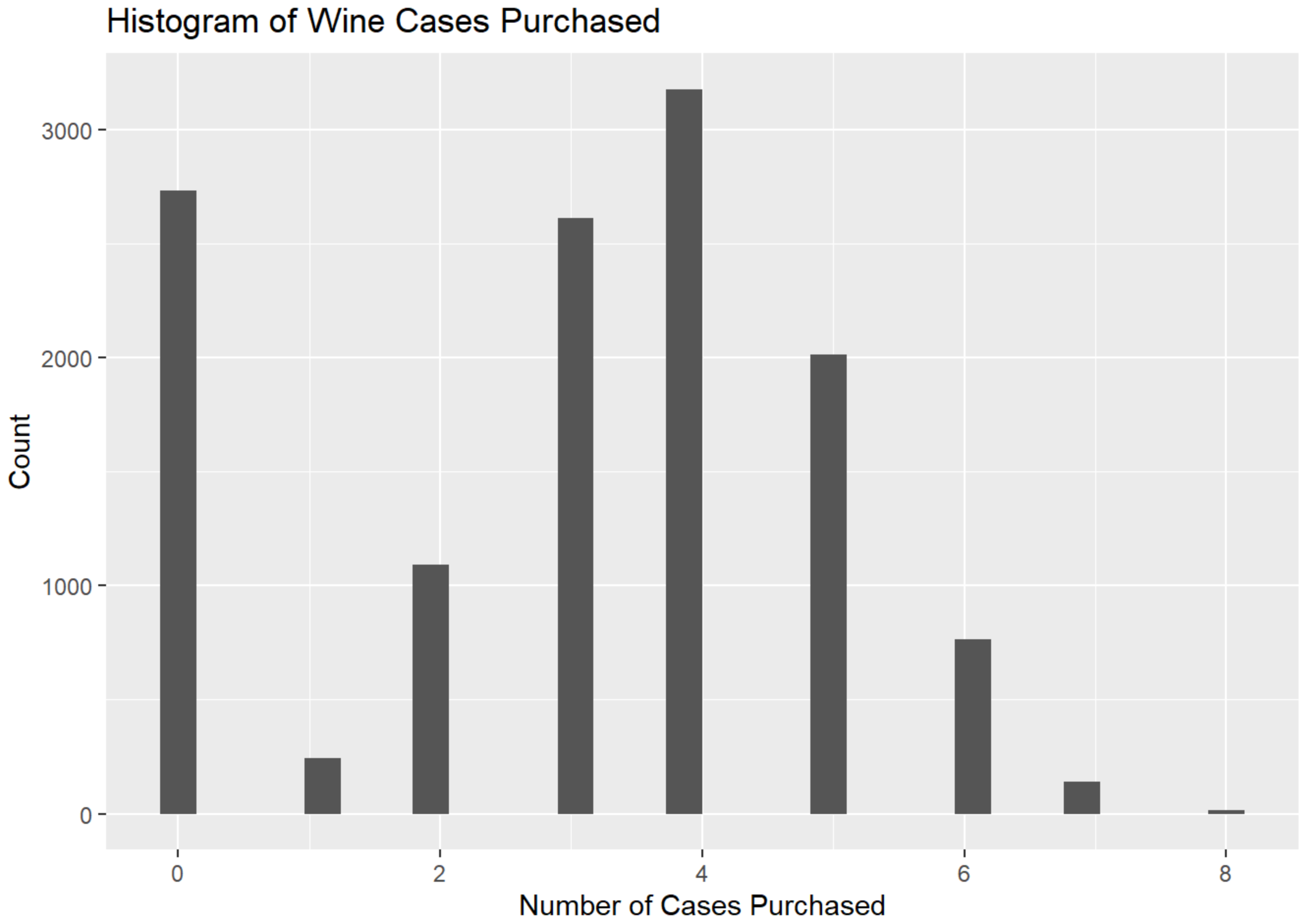
```
##  [1] "TARGET"             "FixedAcidity"       "VolatileAcidity"   
##  [4] "CitricAcid"         "ResidualSugar"      "Chlorides"         
##  [7] "FreeSulfurDioxide"  "TotalSulfurDioxide" "Density"           
## [10] "pH"                 "Sulphates"          "Alcohol"           
## [13] "LabelAppeal"        "AcidIndex"          "STARS"
```

```
dim(wine_train)
```

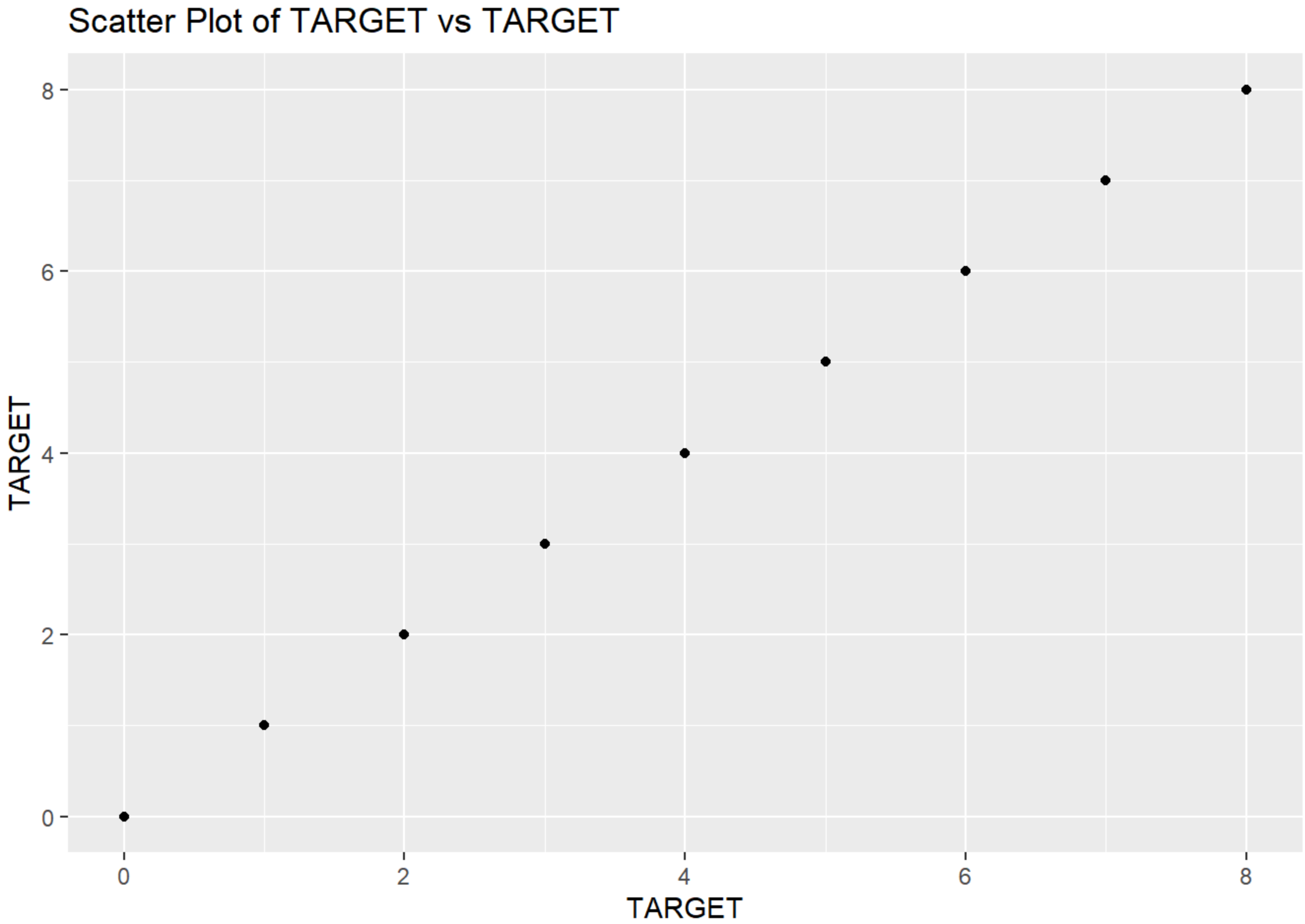
```
## [1] 12795    15
```

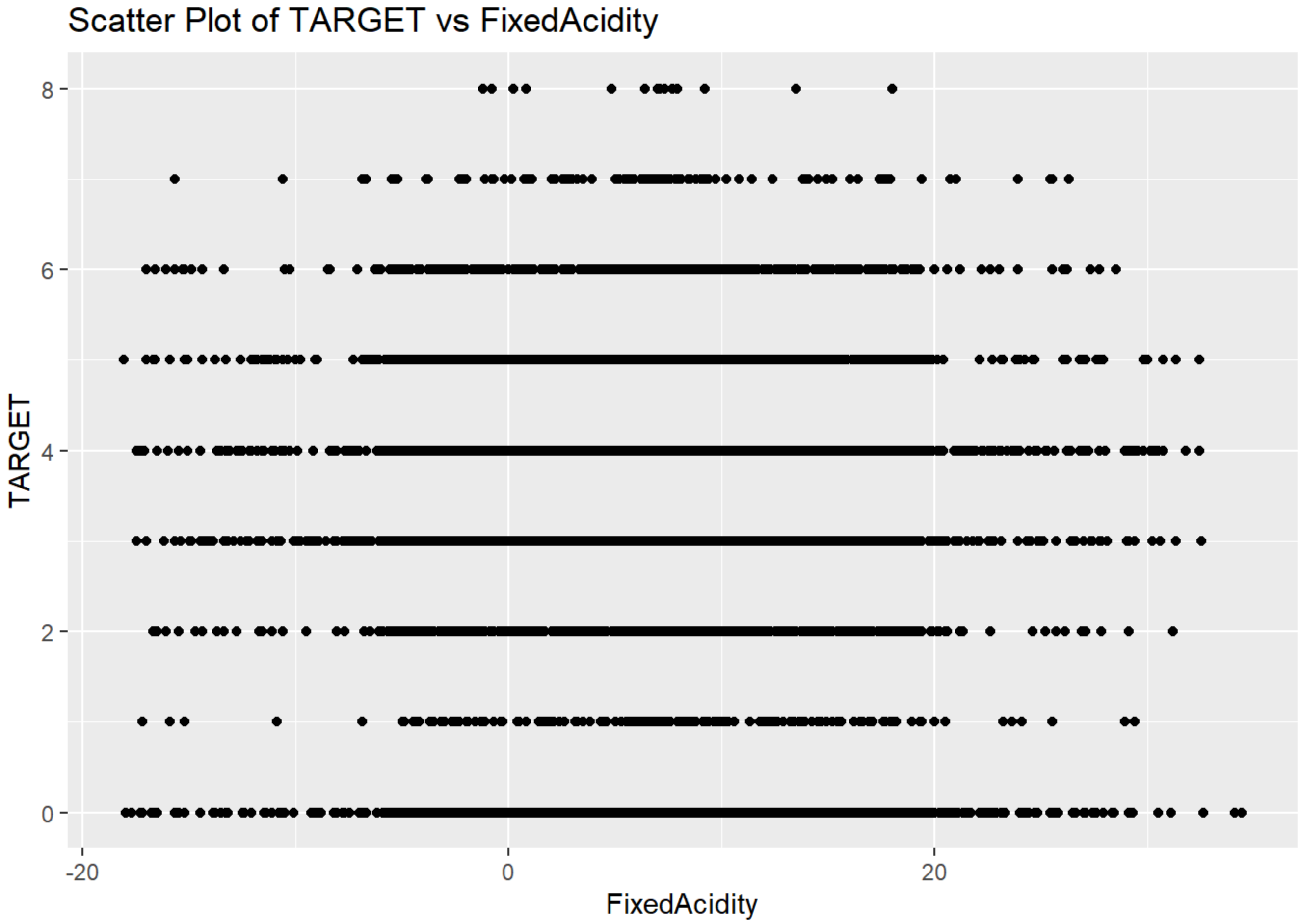
```
wine_train %>% ggplot(aes(x=TARGET)) +
  geom_histogram() +
  labs(title="Histogram of Wine Cases Purchased",
       x="Number of Cases Purchased",
       y="Count")
```

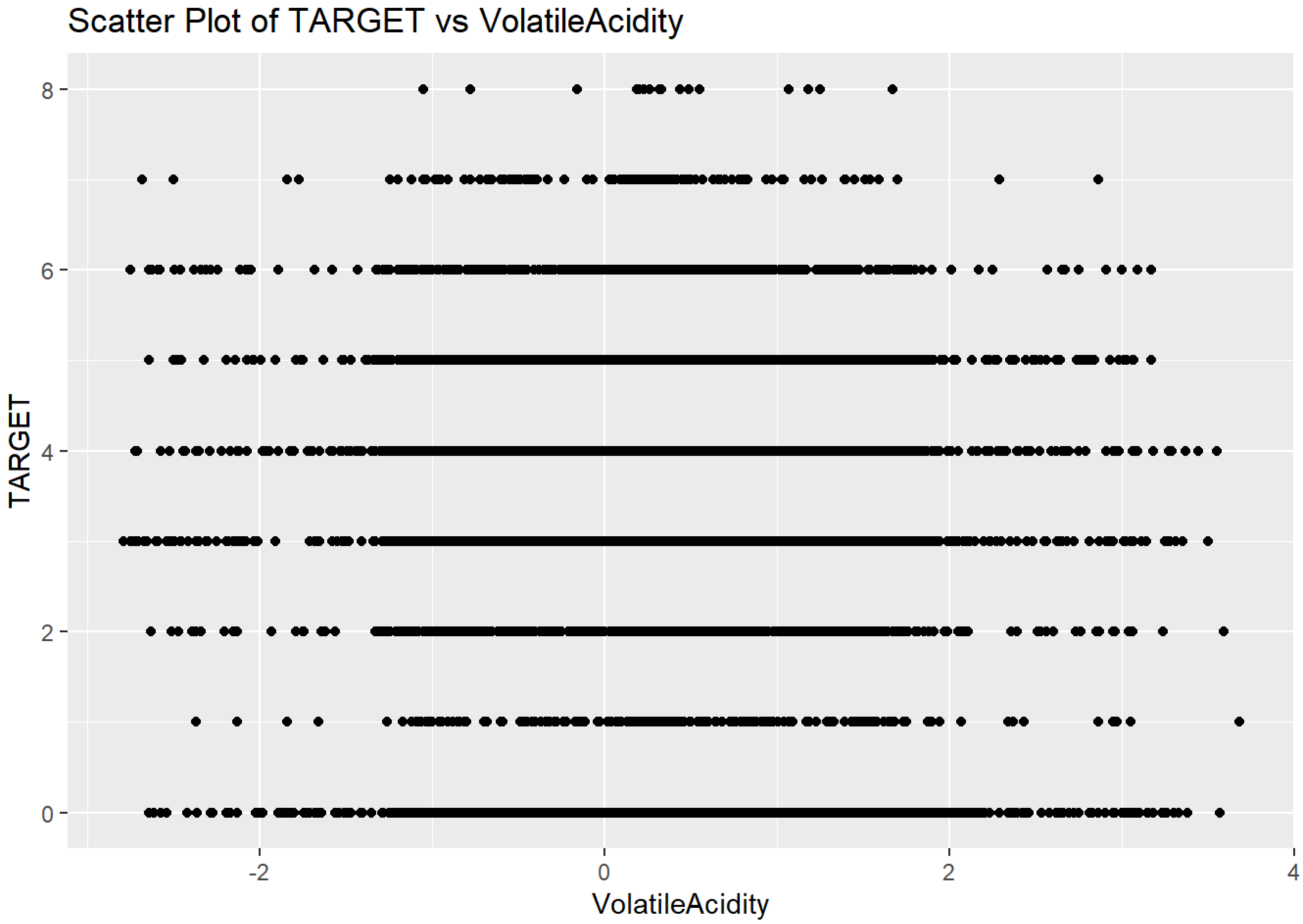
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

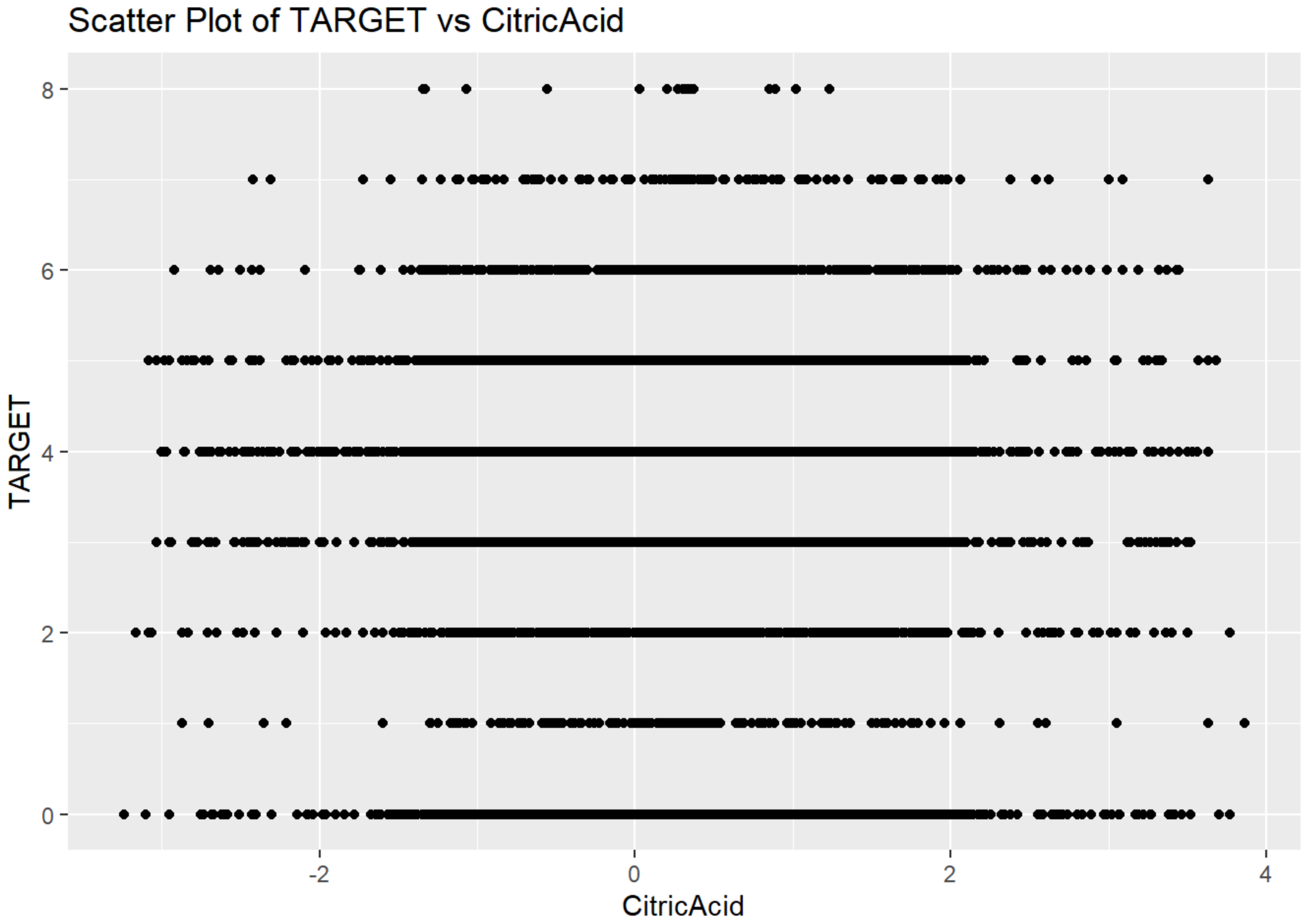


```
# Assuming you have loaded the necessary libraries and the wine_train dataframe

# Get column names
col_names <- names(wine_train)

# Create a scatter plot for each variable
for (col_name in col_names) {
  # Plot scatter plot of TARGET vs Variable
  print(
    wine_train %>%
      ggplot(aes(y = TARGET, x = .data[[col_name]])) +
      geom_point() +
      ggtitle(paste("Scatter Plot of TARGET vs", col_name))
  )
}
```

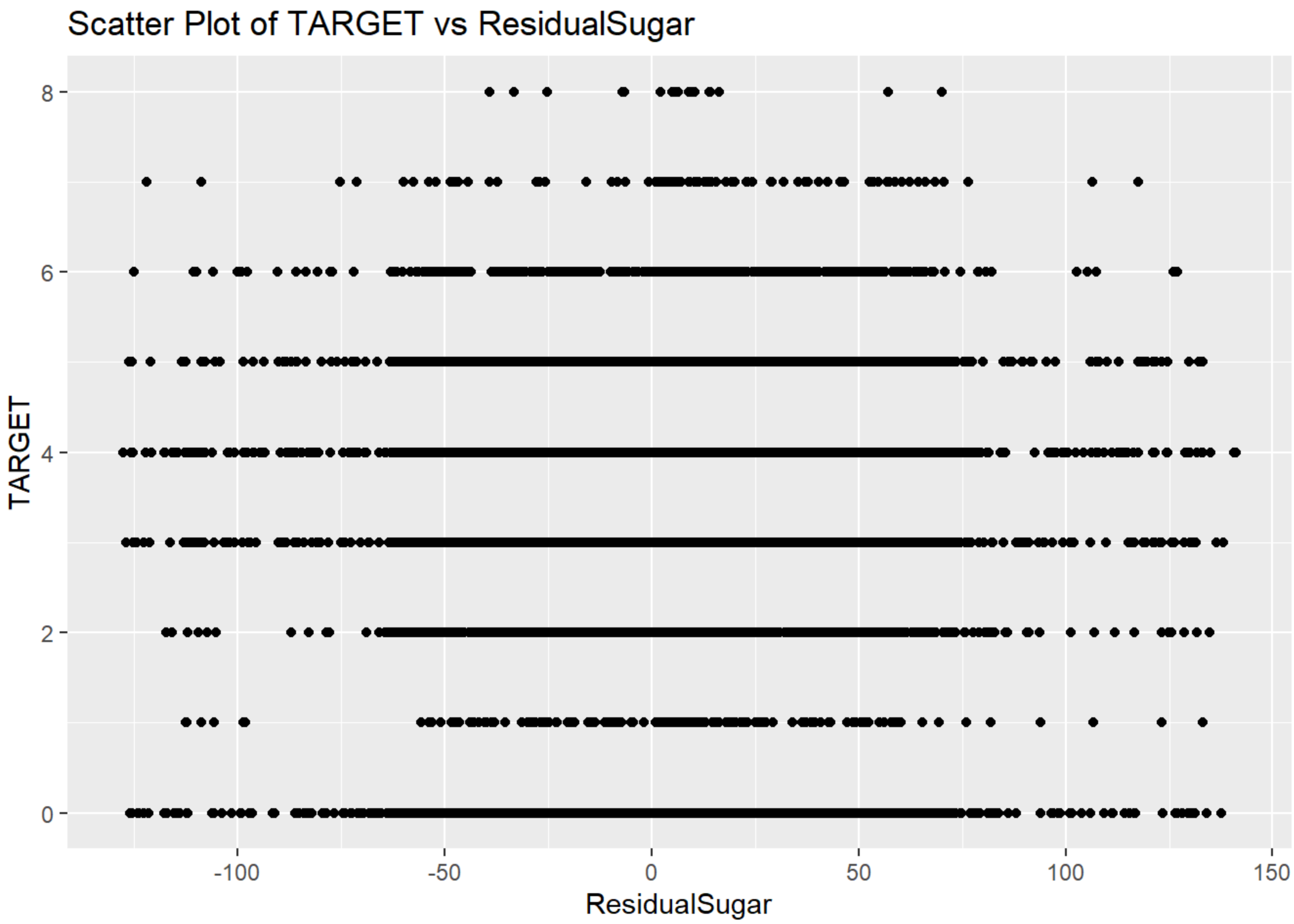

Scatter Plot of TARGET vs TARGET
TARGET
TARGET
0
2
4
6
8
0
2
4
6
8


Scatter Plot of TARGET vs FixedAcidity
TARGET
FixedAcidity
0
2
4
6
8
-20
0
20

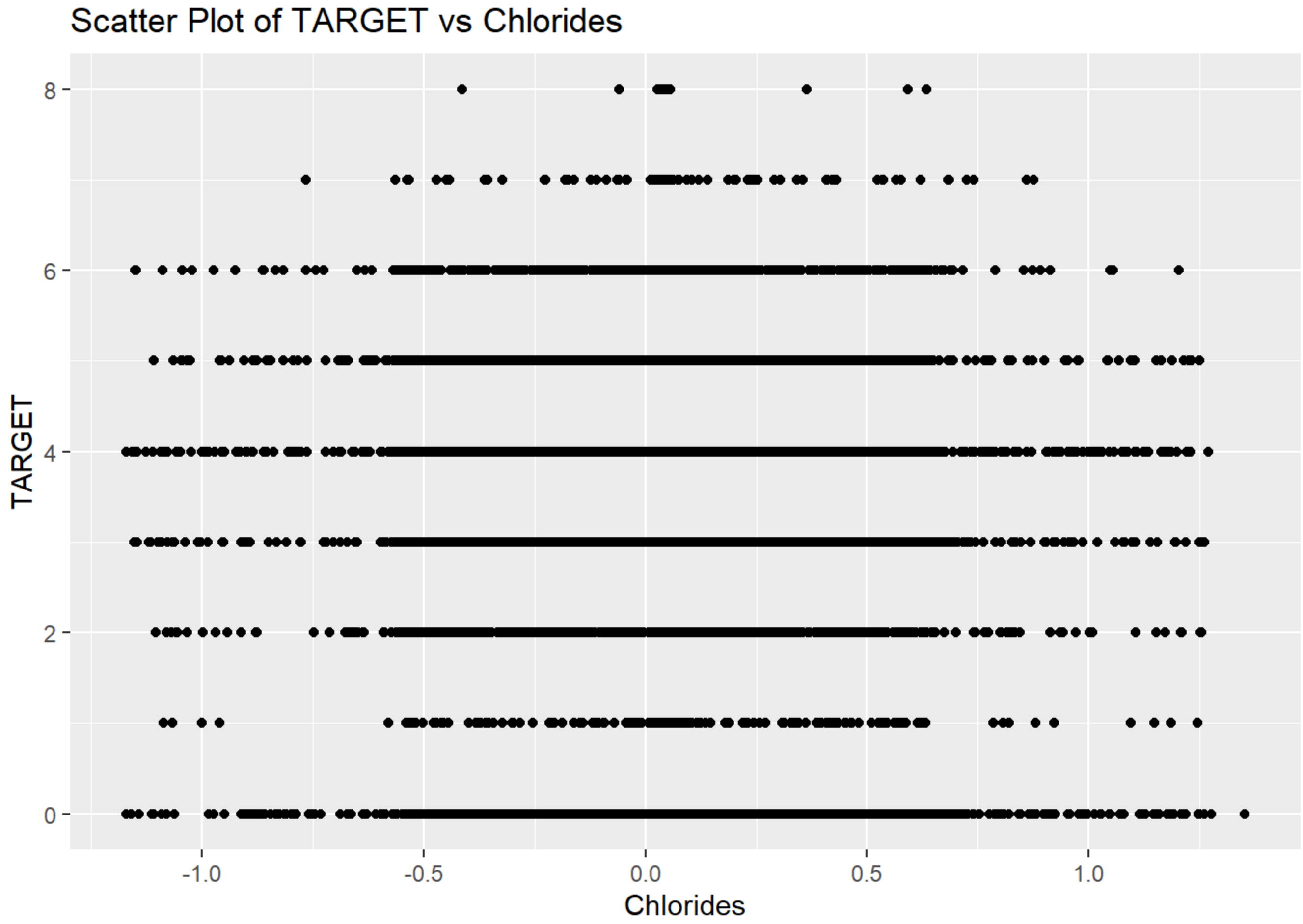

Scatter Plot of TARGET vs VolatileAcidity
TARGET
8
6
4
2
0
-2
0
2
4
VolatileAcidity


Scatter Plot of TARGET vs CitricAcid
TARGET
8
6
4
2
0
-2
0
2
4
CitricAcid

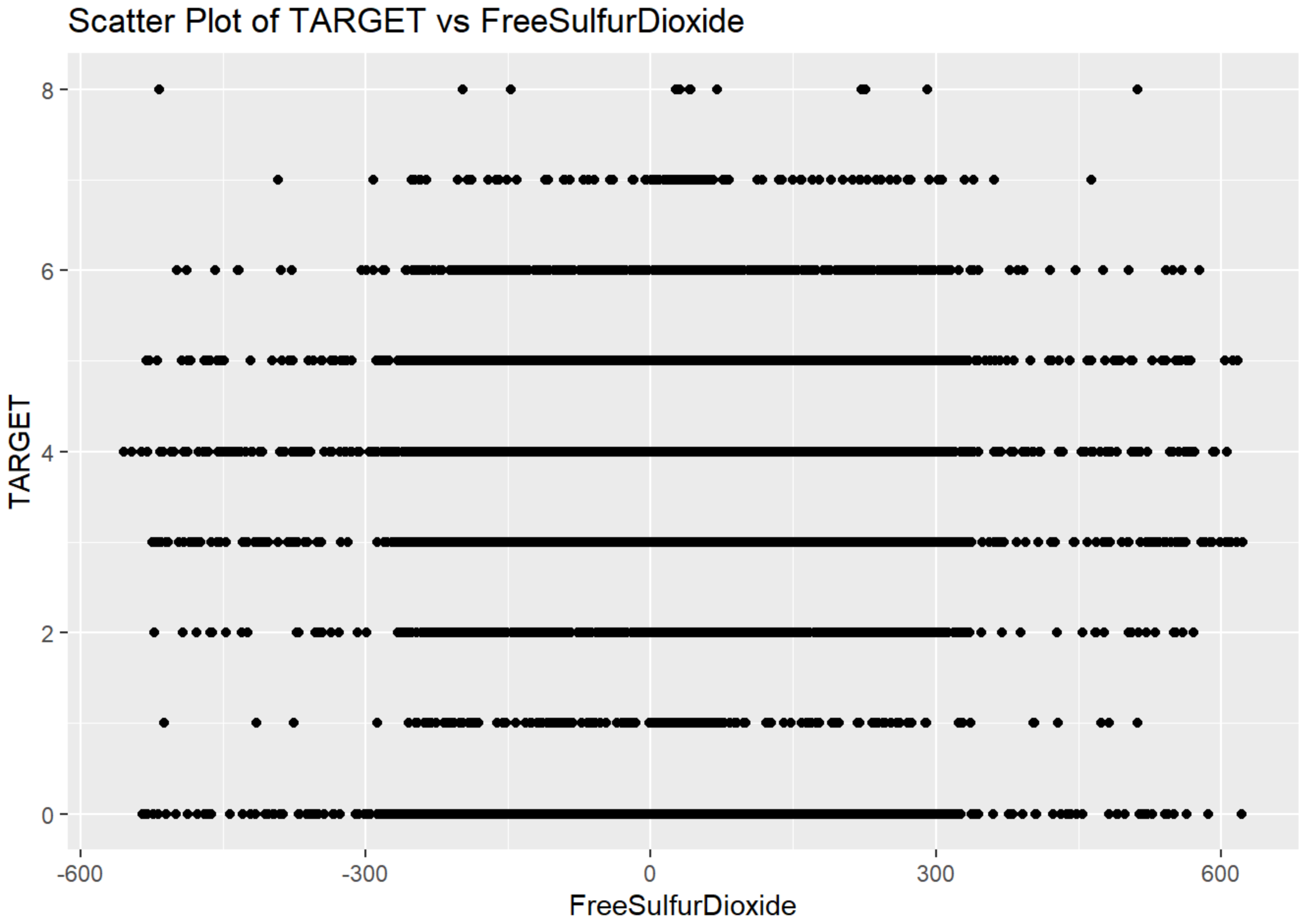
```
## Warning: Removed 616 rows containing missing values (`geom_point()`).
```



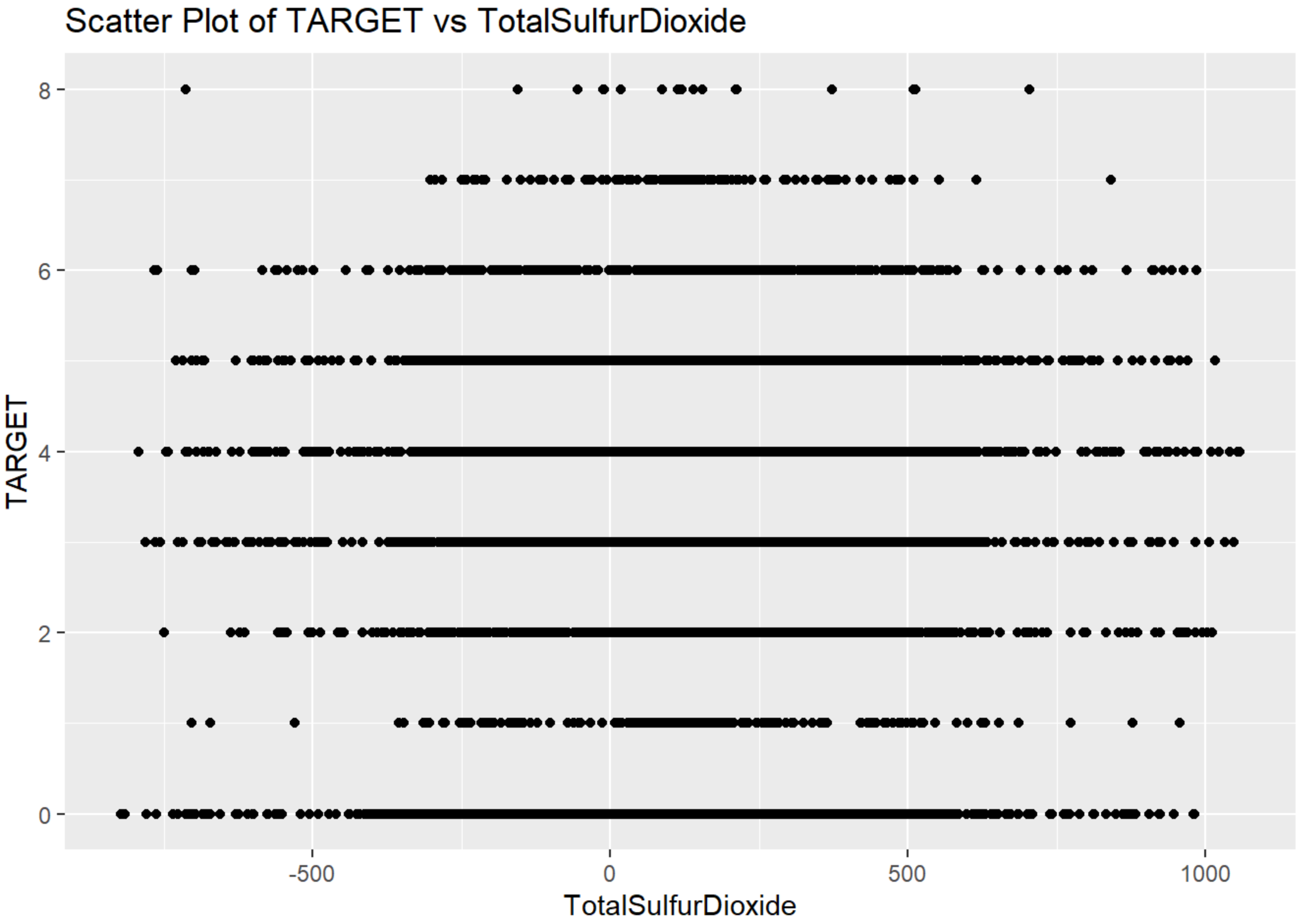
```
## Warning: Removed 638 rows containing missing values (`geom_point()`).
```

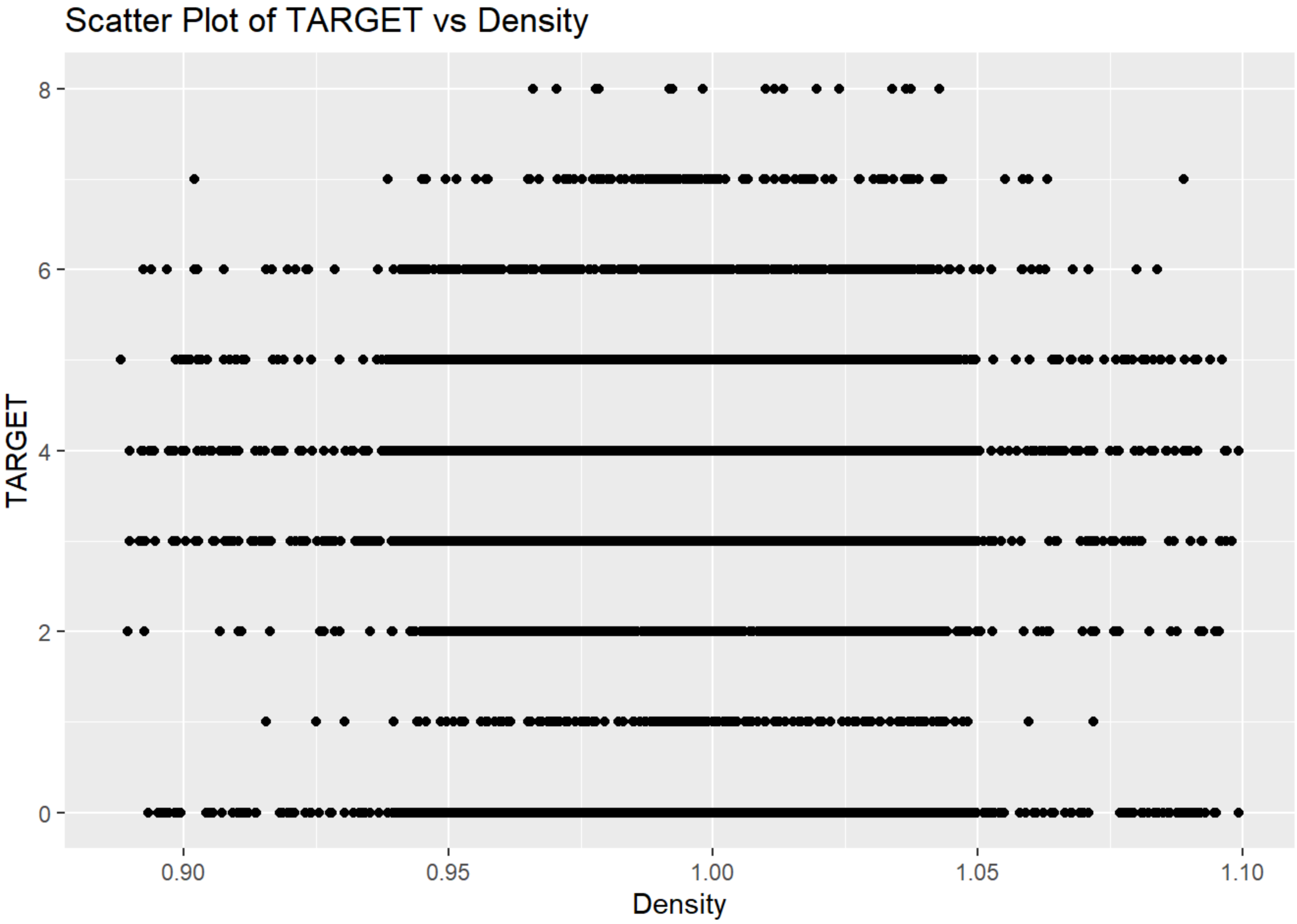


```
## Warning: Removed 647 rows containing missing values (`geom_point()`).
```

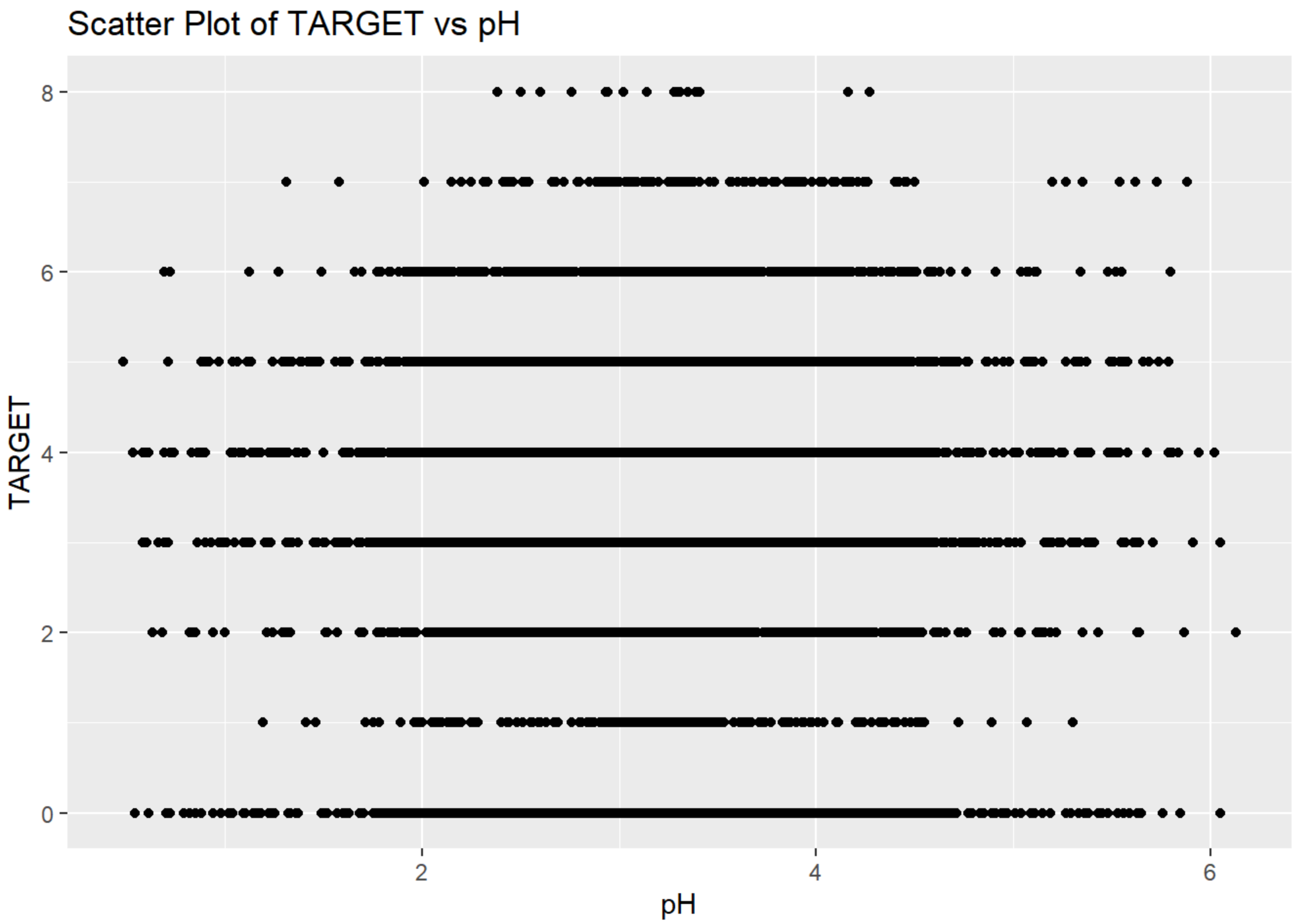


```
## Warning: Removed 682 rows containing missing values (`geom_point()`).
```

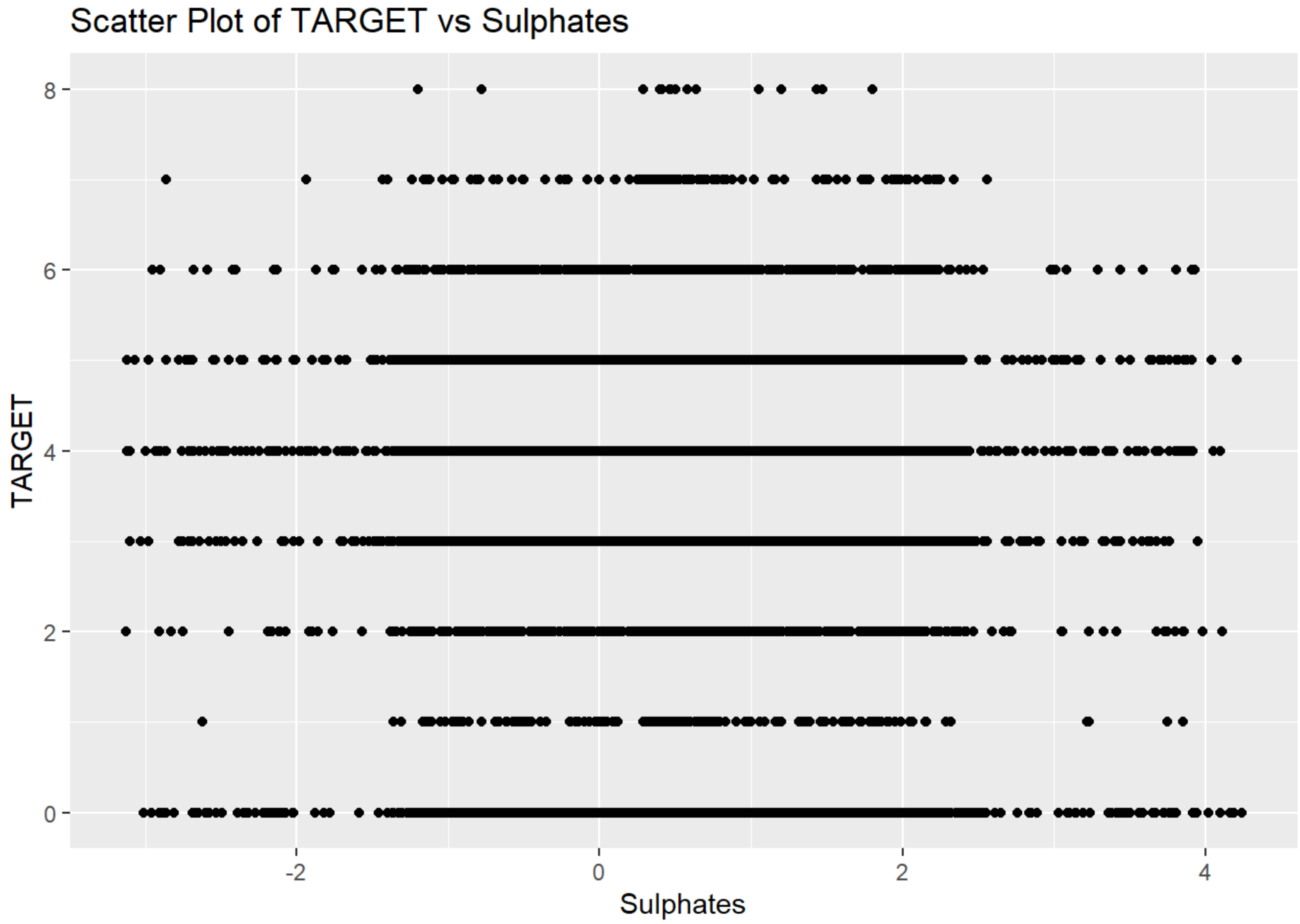

Scatter Plot of TARGET vs TotalSulfurDioxide
TARGET
0
2
4
6
8
-500
0
500
1000
TotalSulfurDioxide


Scatter Plot of TARGET vs Density
TARGET
0
2
4
6
8
0.90
0.95
1.00
1.05
1.10
Density

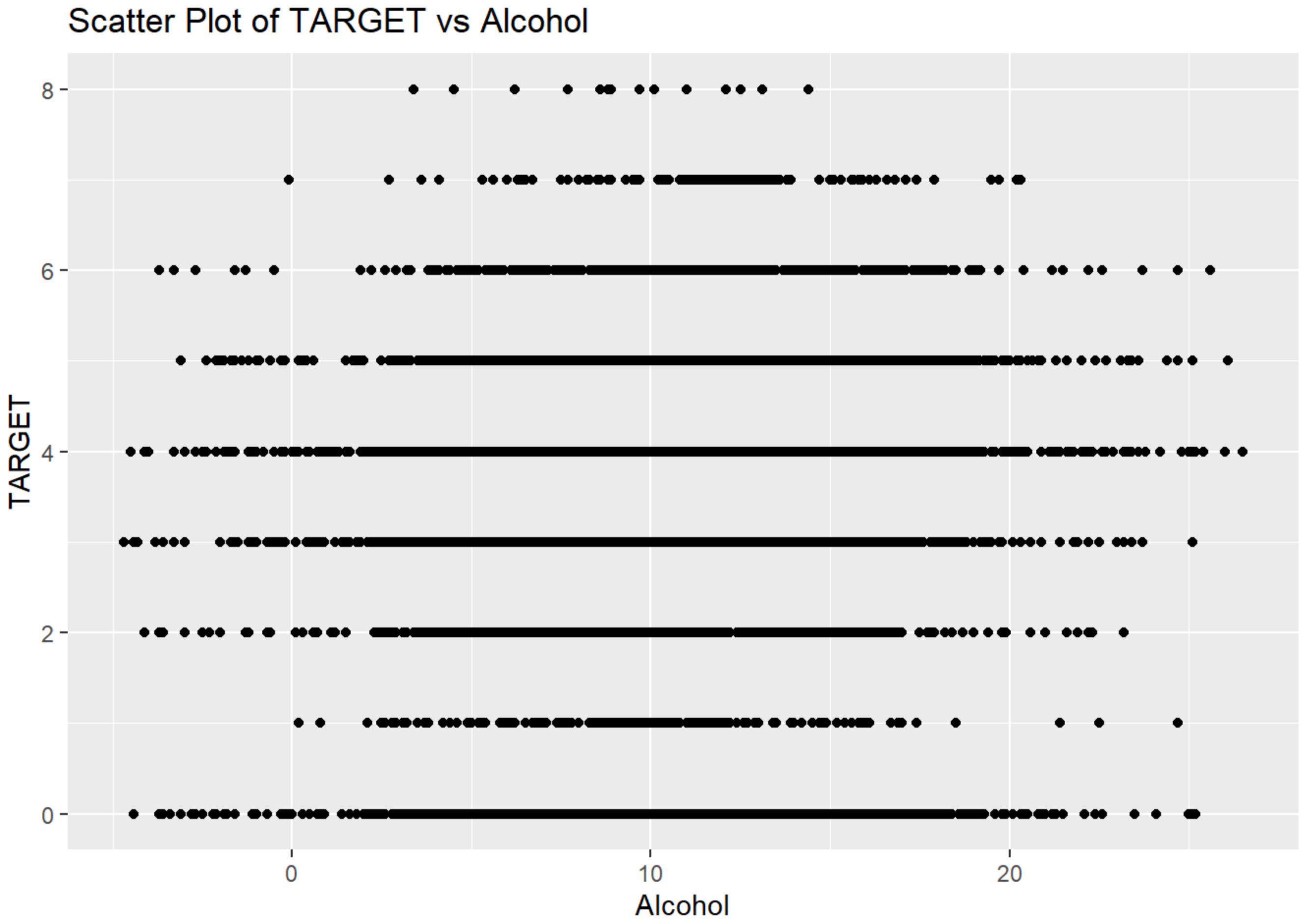
```
## Warning: Removed 395 rows containing missing values (`geom_point()`).
```

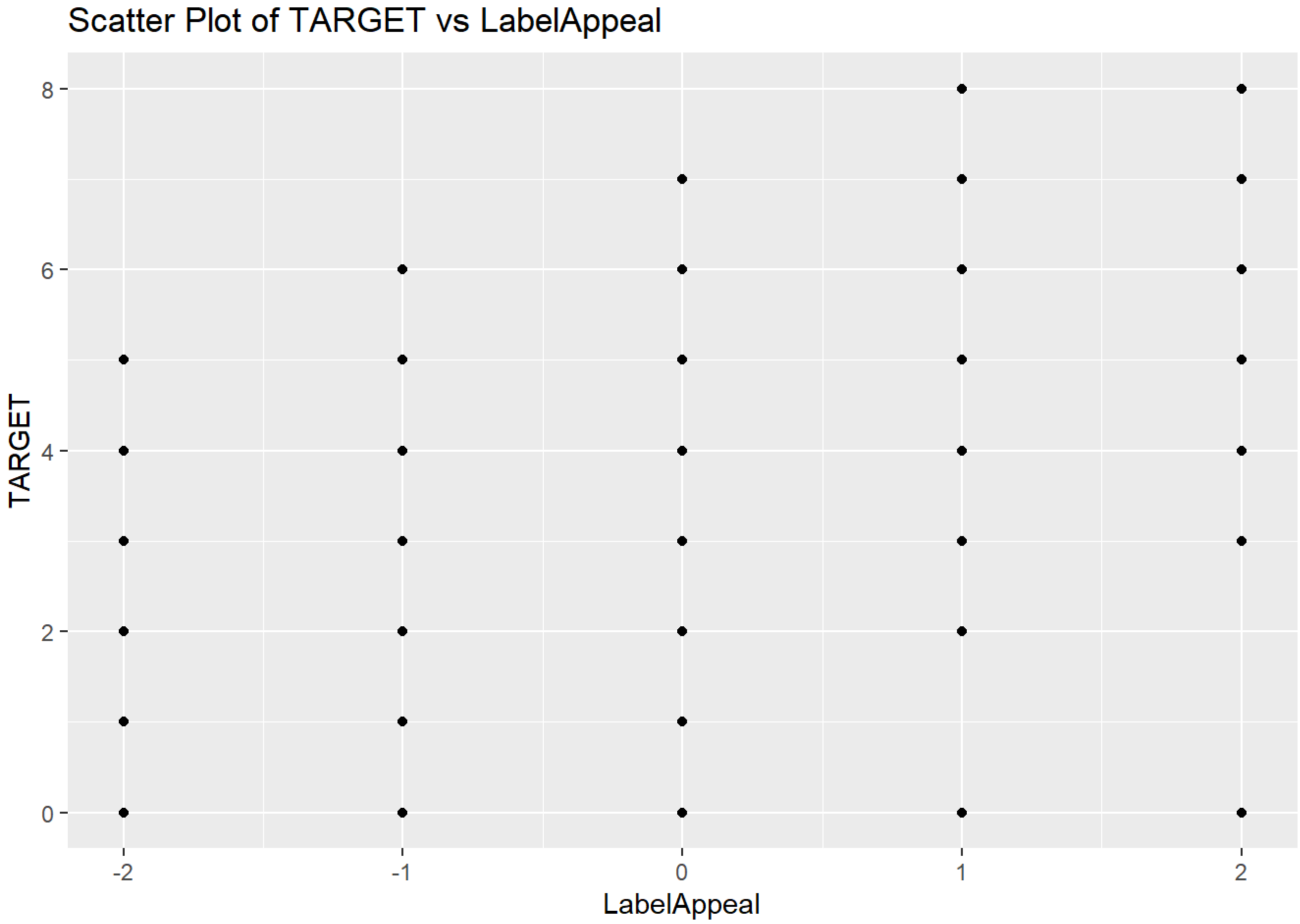


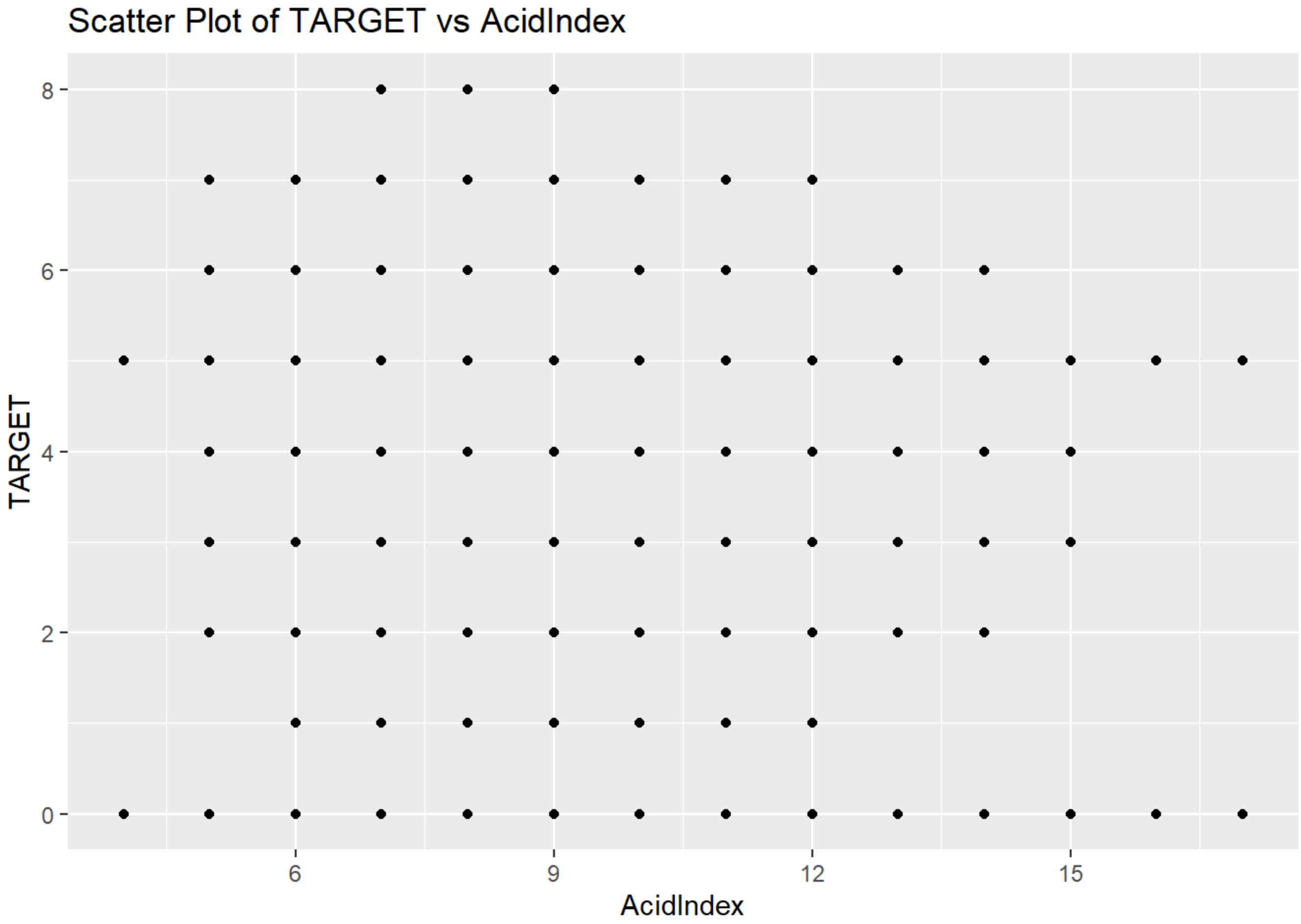
```
## Warning: Removed 1210 rows containing missing values (`geom_point()`).
```



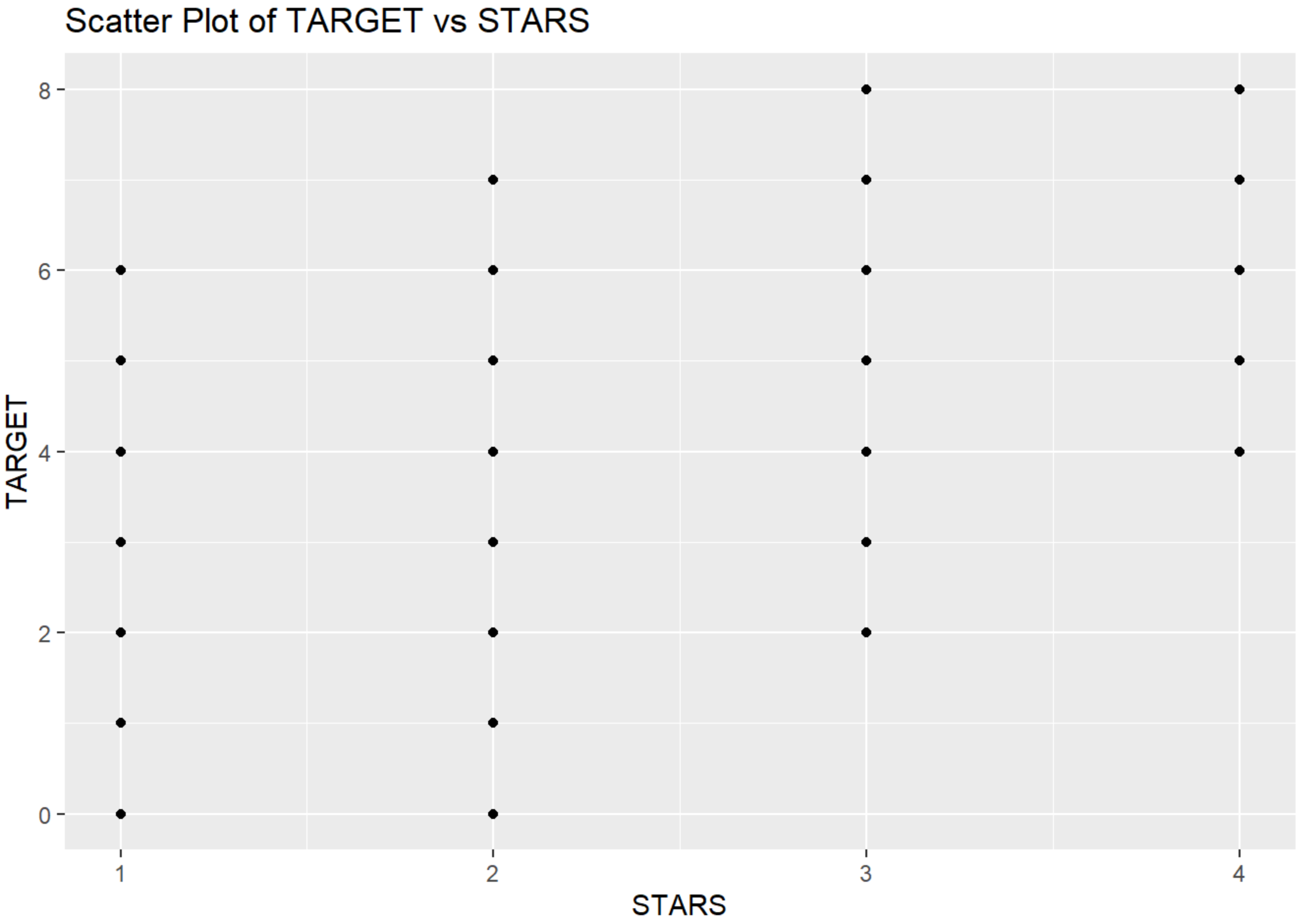
```
## Warning: Removed 653 rows containing missing values (`geom_point()`).
```


Scatter Plot of TARGET vs Alcohol
TARGET
Alcohol
0
2
4
6
8
0
10
20

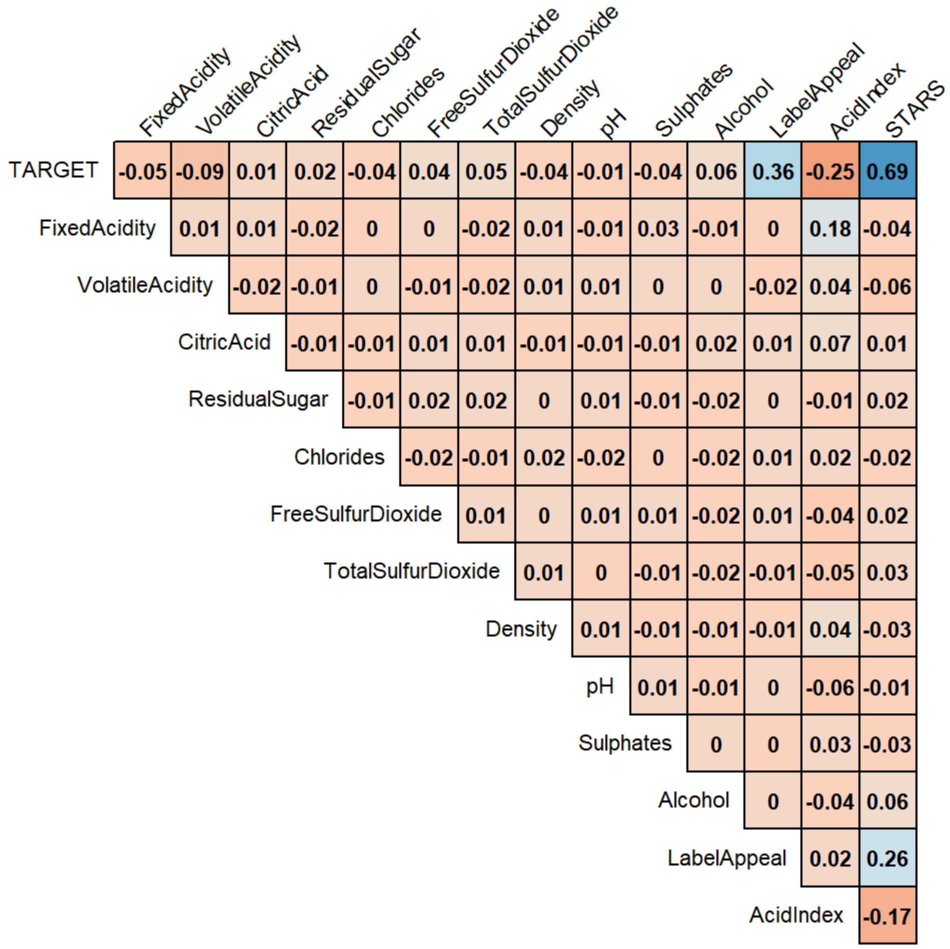

Scatter Plot of TARGET vs LabelAppeal
TARGET
LabelAppeal
0
2
4
6
8
-2
-1
0
1
2



```
## Warning: Removed 3359 rows containing missing values (`geom_point()`).
```


Scatter Plot of TARGET vs STARS
TARGET
0
2
4
6
8
STARS
1
2
3
4

```
# Correlation matrix For Wine Train Data

# impute data for correlation matrix (mean for numeric data, for integer data STARS, 0 for not r
ated)

df_imputed <- wine_train %>%
  mutate(STARS = ifelse(is.na(STARS), 0, STARS),
         across(everything(), ~replace(., is.na(.), mean(., na.rm = TRUE))))



cor_matrix <- cor(df_imputed)

# Set up a color palette (use "RdYlBu" for a blue-red color scheme)
col_palette <- colorRampPalette(c("#67001f", "#b2182b", "#d6604d", "#f4a582", "#fddbc7", "#d1e5f
0", "#92c5de", "#4393c3", "#2166ac"))(50)

# Plot the colored correlation matrix
corrplot(cor_matrix,
         method = "color",
         col = col_palette,
         type = "upper",
         tl.col = "black",
         tl.srt = 45,
         tl.cex = 0.7,
         addCoef.col = "black",
         diag = FALSE,  # Exclude correlation values on the diagonal
         cl.pos = "n",  # Do not show color legend
         outline = TRUE,
         number.cex = 0.7# Add white outlines around the boxes for better readability
)
```



	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
TARGET	-0.05	-0.09	0.01	0.02	-0.04	0.04	0.05	-0.04	-0.01	-0.04	0.06	0.36	-0.25	0.69
FixedAcidity		0.01	0.01	-0.02	0	0	-0.02	0.01	-0.01	0.03	-0.01	0	0.18	-0.04
VolatileAcidity			-0.02	-0.01	0	-0.01	-0.02	0.01	0.01	0	0	-0.02	0.04	-0.06
CitricAcid				-0.01	-0.01	0.01	0.01	-0.01	-0.01	-0.01	0.02	0.01	0.07	0.01
ResidualSugar					-0.01	0.02	0.02	0	0.01	-0.01	-0.02	0	-0.01	0.02
Chlorides						-0.02	-0.01	0.02	-0.02	0	-0.02	0.01	0.02	-0.02
FreeSulfurDioxide							0.01	0	0.01	0.01	-0.02	0.01	-0.04	0.02
TotalSulfurDioxide								0.01	0	-0.01	-0.02	-0.01	-0.05	0.03
Density									0.01	-0.01	-0.01	-0.01	0.04	-0.03
pH										0.01	-0.01	0	-0.06	-0.01
Sulphates											0	0	0.03	-0.03
Alcohol												0	-0.04	0.06
LabelAppeal													0.02	0.26
AcidIndex														-0.17

###

Data Preparation

```
# Impute the data same as for correlation matrix

# For STARS, all NA values make 0
# For all other NA values for numeric variables, impute with the mean
wine_train_imputed <- wine_train %>%
  mutate(STARS = ifelse(is.na(STARS), 0, STARS),
         across(everything(), ~replace(., is.na(.), mean(., na.rm = TRUE))))

# Note: Do not make any of the variables categorical to preserve their order. Represent ordered discete data through integers.
```

```
# Split into Train and Test Sets
set.seed(1)

test_size <- floor(dim(wine_train_imputed)[1]*0.2)
test_index <- sample(1:nrow(wine_train_imputed), test_size, replace=FALSE)
wine_train2 <- wine_train_imputed[-test_index,]
wine_test2 <- wine_train_imputed[test_index,]
```

Build Models

Try Poisson Regression Models

```
# Poisson Regression Models
pois_glm1 <- glm(TARGET~LabelAppeal+AcidIndex+STARS,data=wine_train2, family=poisson)
pois_glm2 <- glm(TARGET~AcidIndex+Alcohol+CitricAcid+FreeSulfurDioxide+ResidualSugar+STARS+pH+La
belAppeal, data=wine_train2, family=poisson)

anova(pois_glm1, pois_glm2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: TARGET ~ LabelAppeal + AcidIndex + STARS
## Model 2: TARGET ~ AcidIndex + Alcohol + CitricAcid + FreeSulfurDioxide +
##     ResidualSugar + STARS + pH + LabelAppeal
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     10232      11930
## 2     10227      11908  5   21.605 0.0006222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Goodness of fit test for poisson glm 1 (our model compared to saturated)
1-pchisq(deviance(pois_glm1), df.residual(pois_glm1))
```

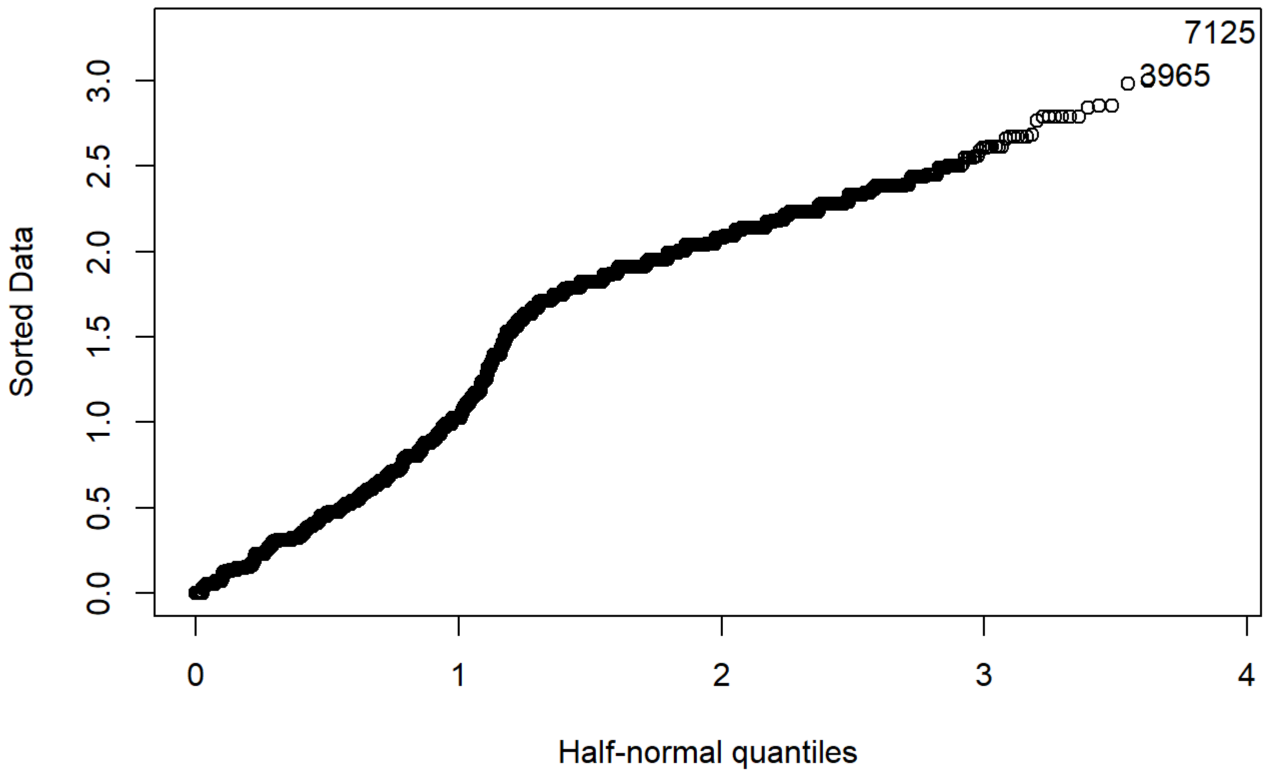
```
## [1] 0
```

```
# Goodness of fit test for poisson glm 2 (our model compared to saturated)
1-pchisq(deviance(pois_glm2), df.residual(pois_glm2))
```

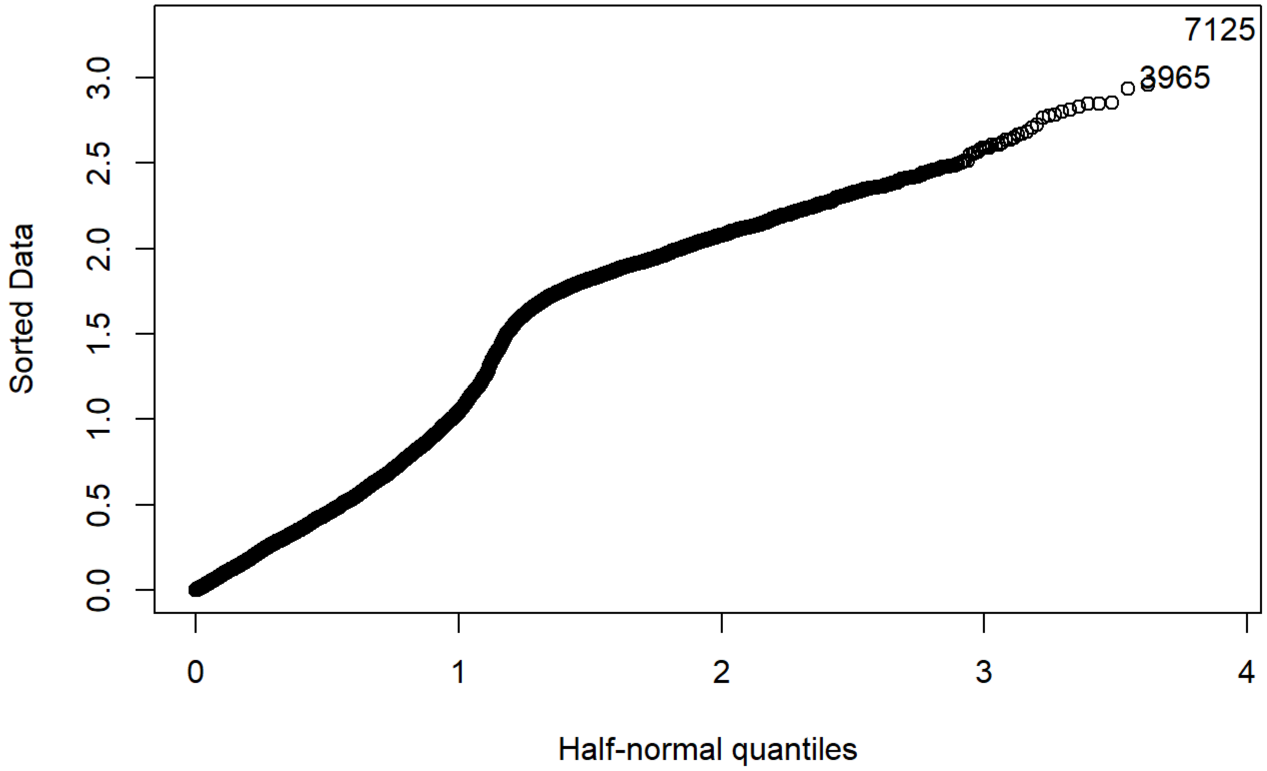
```
## [1] 0
```

Note: Low p-value means we reject the null hypothesis, that our model is not a bad fit for the data. Therefore both models are a bad fit for the data and do not pass the chi-squared goodness of fit test.

```
halfnorm(residuals(pois_glm1))
```



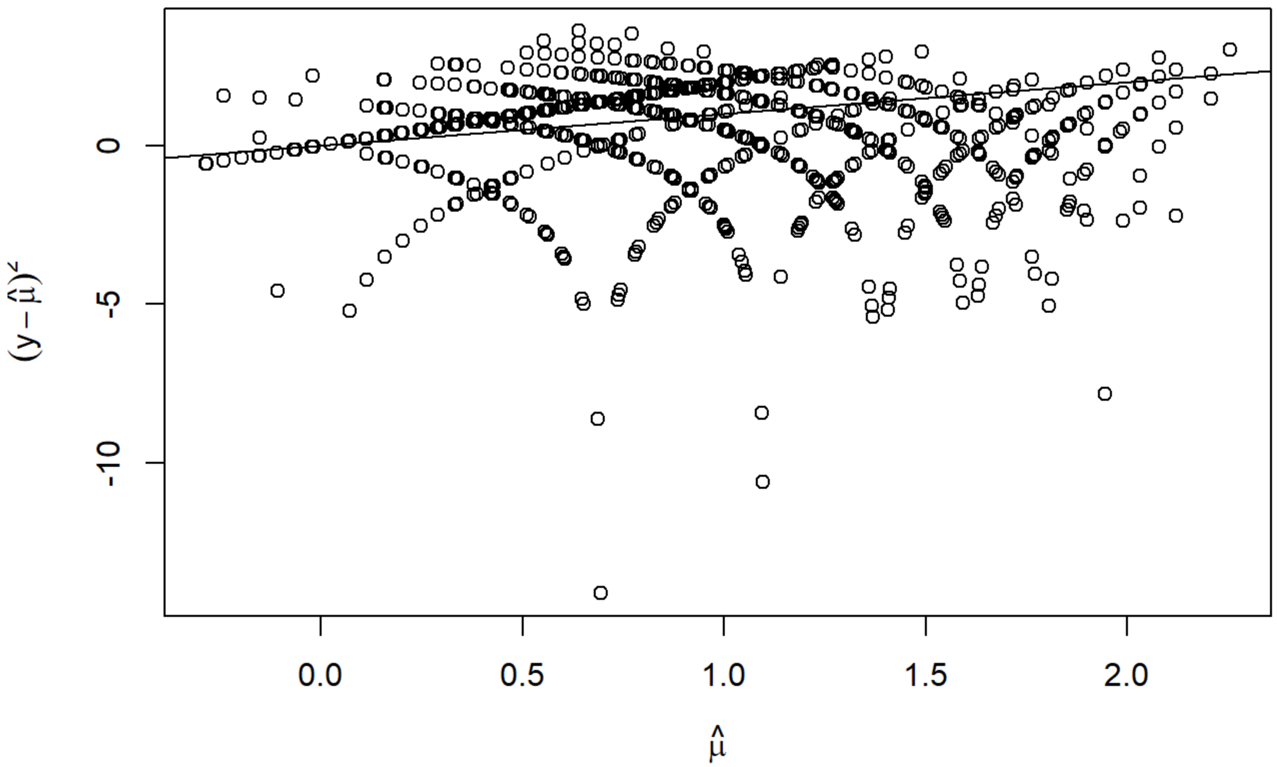
```
halfnorm(residuals(pois_glm2))
```



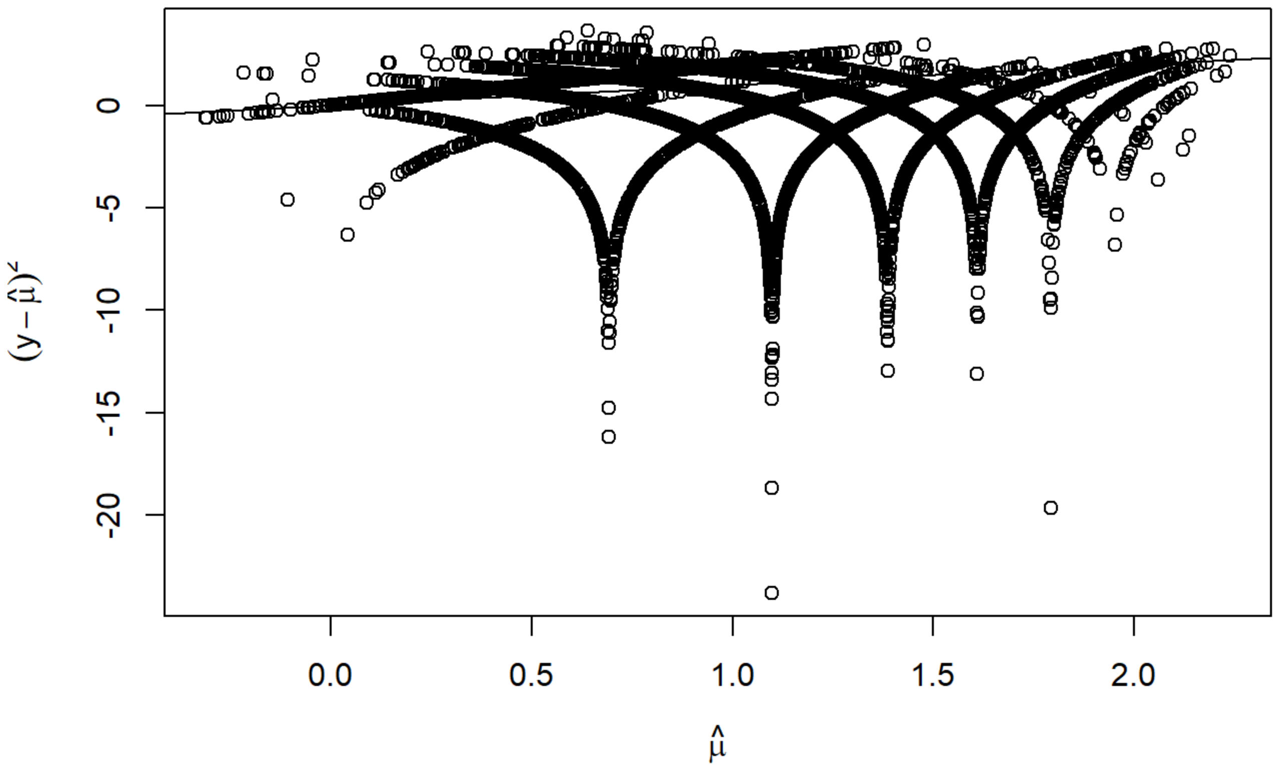
Half

Norm Plot does not indentify any significant outliers in the data.

```
plot(log(fitted(pois_glm1)), log((wine_train2$TARGET-fitted(pois_glm1))^2),
     xlab=expression(hat(mu)), ylab=expression((y-hat(mu))^2))
abline(0, 1)
```



```
plot(log(fitted(pois_glm2)), log((wine_train2$TARGET-fitted(pois_glm2))^2),
     xlab=expression(hat(mu)), ylab=expression((y-hat(mu))^2))
abline(0, 1)
```



Based on the Mean and Variance plot, there does not seem to be significant over-dispersion, possibly just a little under dispersion.

```
# Estimate phi
deviance(pois_glm1) / df.residual(pois_glm1)
```

```
## [1] 1.16592
```

```
sum(residuals(pois_glm1, type="pearson")^2)/pois_glm1$df.res
```

```
## [1] 0.862302
```

```
deviance(pois_glm2) / df.residual(pois_glm2)
```

```
## [1] 1.164378
```

```
sum(residuals(pois_glm2, type="pearson")^2)/pois_glm2$df.res
```

```
## [1] 0.8615467
```

Note: Both estimated phi values are actually pretty close to 1, therefore models are not necessarily over-dispersed or under-dispersed

```
# Take model and look at lambda (mean) values and simulate data with model
# (generate data from fitted model and see if it looks like actual data)

lambda.fitted <- predict(pois_glm1, type="response")
sim_data <- rpois(nrow(wine_train2), lambda.fitted)

sum(sim_data==0)
```

```
## [1] 935
```

```
sum(wine_train2$TARGET==0)
```

```
## [1] 2206
```

```
lambda.fitted <- predict(pois_glm2, type="response")
sim_data <- rpois(nrow(wine_train2), lambda.fitted)

sum(sim_data==0)
```

```
## [1] 963
```

```
sum(wine_train2$TARGET==0)
```

```
## [1] 2206
```

Note: Number of simulated zeros much smaller than the sum of the actual zero values, may consider ZIP model

Try ZIP models (Zero Inflated Poisson)

```
zip_glm1 <- zeroinfl(TARGET~LabelAppeal+AcidIndex+STARS,data=wine_train2, dist="poisson")
zip_glm2 <- zeroinfl(TARGET~AcidIndex+Alcohol+CitricAcid+FreeSulfurDioxide+ResidualSugar+STARS+p
H+LabelAppeal, data=wine_train2, dist="poisson")

logLik(zip_glm1) # 8 df for simple model
```

```
## 'log Lik.' -16364.19 (df=8)
```

```
logLik(zip_glm2) # 30 df for model with all variables
```

```
## 'log Lik.' -16335.56 (df=18)
```

```
# use difference of 22 for df to compare nested models

1-pchisq(2 * (logLik(zip_glm2) - logLik(zip_glm1)), df = 22)
```

```
## 'log Lik.' 5.623378e-05 (df=18)
```

P-value for nested models is small so we reject null hypothesis and prefer larger model

Try Negative Binomial Regression Models

```
nb_glm1 <- glm.nb(TARGET~LabelAppeal+AcidIndex+STARS,data=wine_train2)
```

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

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

```
nb_glm2 <- glm.nb(TARGET~AcidIndex+Alcohol+CitricAcid+FreeSulfurDioxide+ResidualSugar+STARS+pH+L
abelAppeal, data=wine_train2)
```

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

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

```
anova(nb_glm1, nb_glm2, test="Chisq")
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: TARGET
##                                                                                            M
odel
## 1                                                                 LabelAppeal + AcidIndex + S
TARS
## 2 AcidIndex + Alcohol + CitricAcid + FreeSulfurDioxide + ResidualSugar + STARS + pH + LabelAp
peal
##      theta Resid. df    2 x log-lik.   Test    df LR stat.      Pr(Chi)
## 1 48194.46     10232       -37429.95
## 2 48228.29     10227       -37408.34 1 vs 2     5 21.60494 0.0006223128
```

```
# We will not use NB because does not make sence (data not over-disperssed)
```

Try Multiple Linear Regression Models

```
lm1 <- lm(TARGET~LabelAppeal+AcidIndex+STARS,data=wine_train2)
lm2 <- lm(TARGET~AcidIndex+Alcohol+CitricAcid+FreeSulfurDioxide+ResidualSugar+STARS+pH+LabelAppe
al, data=wine_train2)

anova(lm1, lm2, test="F")
```

```
## Analysis of Variance Table
##
## Model 1: TARGET ~ LabelAppeal + AcidIndex + STARS
## Model 2: TARGET ~ AcidIndex + Alcohol + CitricAcid + FreeSulfurDioxide +
##     ResidualSugar + STARS + pH + LabelAppeal
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  10232 18190
## 2  10227 18134  5    55.353 6.2434 8.655e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = TARGET ~ AcidIndex + Alcohol + CitricAcid + FreeSulfurDioxide +
##     ResidualSugar + STARS + pH + LabelAppeal, data = wine_train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5011 -0.9656  0.0538  0.9105  6.0712
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.302e+00  1.163e-01  28.386  < 2e-16 ***
## AcidIndex         -2.133e-01  1.014e-02 -21.025  < 2e-16 ***
## Alcohol            6.788e-03  3.643e-03   1.863 0.062460 .
## CitricAcid         2.754e-02  1.530e-02   1.799 0.071986 .
## FreeSulfurDioxide  3.273e-04  9.104e-05   3.595 0.000326 ***
## ResidualSugar      5.535e-04  3.975e-04   1.392 0.163874
## STARS              9.852e-01  1.175e-02  83.880  < 2e-16 ***
## pH                -5.994e-02  1.963e-02  -3.053 0.002268 **
## LabelAppeal        4.255e-01  1.534e-02  27.730  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.332 on 10227 degrees of freedom
## Multiple R-squared:  0.5233, Adjusted R-squared:  0.5229
## F-statistic:  1403 on 8 and 10227 DF,  p-value: < 2.2e-16
```

Select Models

```
confint(pois_glm2)
```

```
## Waiting for profiling to be done...
```

```
##                          2.5 %        97.5 %
## (Intercept)        1.177865e+00  1.3894883561
## AcidIndex         -9.919222e-02 -0.0794943364
## Alcohol           -2.323465e-03  0.0038899393
## CitricAcid        -2.685497e-03  0.0232313633
## FreeSulfurDioxide  4.787949e-05  0.0002021429
## ResidualSugar     -1.626302e-04  0.0005126397
## STARS              3.044786e-01  0.3243375863
## pH                -4.022209e-02 -0.0068951814
## LabelAppeal        1.176753e-01  0.1442355075
```

```
confint(zip_glm2)
```

```
##                                2.5 %        97.5 %
## count_(Intercept)        1.060900e+00  1.283597e+00
## count_AcidIndex         -2.845609e-02 -7.102384e-03
## count_Alcohol            3.569344e-03  9.931053e-03
## count_CitricAcid        -1.338654e-02  1.309368e-02
## count_FreeSulfurDioxide -4.810373e-05  1.068625e-04
## count_ResidualSugar     -3.722364e-04  3.213018e-04
## count_STARS              8.801739e-02  1.108755e-01
## count_pH                -1.690812e-02  1.741982e-02
## count_LabelAppeal        2.201134e-01  2.477499e-01
## zero_(Intercept)        -4.794127e+00 -3.510925e+00
## zero_AcidIndex           3.855369e-01  4.953273e-01
## zero_Alcohol             1.488490e-02  5.616388e-02
## zero_CitricAcid         -1.465645e-01  2.500727e-02
## zero_FreeSulfurDioxide  -1.126306e-03 -8.215747e-05
## zero_ResidualSugar      -3.990436e-03  4.678375e-04
## zero_STARS              -2.481451e+00 -2.225725e+00
## zero_pH                  1.465089e-01  3.664600e-01
## zero_LabelAppeal         6.365868e-01  8.218598e-01
```

```
confint(nb_glm2)
```

```
## Waiting for profiling to be done...
##                           2.5 %        97.5 %
## (Intercept)        1.177874e+00  1.3895043462
## AcidIndex         -9.919460e-02 -0.0794960586
## Alcohol           -2.323663e-03  0.0038899767
## CitricAcid        -2.685859e-03  0.0232319925
## FreeSulfurDioxide  4.787832e-05  0.0002021476
## ResidualSugar     -1.626340e-04  0.0005126616
## STARS              3.044824e-01  0.3243421858
## pH                -4.022374e-02 -0.0068955389
## LabelAppeal        1.176742e-01  0.1442353746
```

```
logLik(pois_glm2)
```

```
## 'log Lik.' -18704.06 (df=9)
```

```
logLik(nb_glm2)
```

```
## 'log Lik.' -18704.17 (df=10)
```

```
logLik(zip_glm2)
```

```
## 'log Lik.' -16335.56 (df=18)
```

```
AIC(pois_glm2)
```

```
## [1] 37426.12
```

```
AIC(nb_glm2)
```

```
## [1] 37428.34
```

```
AIC(zip_glm2)
```

```
## [1] 32707.12
```

```
AIC(lm2)
```

```
## [1] 34922.45
```

```
summary(pois_glm2)
```

```
##
## Call:
## glm(formula = TARGET ~ AcidIndex + Alcohol + CitricAcid + FreeSulfurDioxide +
##     ResidualSugar + STARS + pH + LabelAppeal, family = poisson,
##     data = wine_train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9598  -0.7340   0.0679   0.5704   3.2906
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.284e+00  5.399e-02  23.777  < 2e-16 ***
## AcidIndex         -8.932e-02  5.025e-03 -17.774  < 2e-16 ***
## Alcohol            7.831e-04  1.585e-03   0.494  0.62127
## CitricAcid         1.027e-02  6.612e-03   1.554  0.12023
## FreeSulfurDioxide  1.250e-04  3.935e-05   3.177  0.00149 **
## ResidualSugar      1.750e-04  1.723e-04   1.016  0.30975
## STARS              3.144e-01  5.066e-03  62.060  < 2e-16 ***
## pH                -2.356e-02  8.502e-03  -2.771  0.00559 **
## LabelAppeal        1.310e-01  6.776e-03  19.327  < 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: 18373  on 10235  degrees of freedom
## Residual deviance: 11908  on 10227  degrees of freedom
## AIC: 37426
##
## Number of Fisher Scoring iterations: 5
```

```
summary(zip_glm2)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ AcidIndex + Alcohol + CitricAcid + FreeSulfurDioxide +
##     ResidualSugar + STARS + pH + LabelAppeal, data = wine_train2, dist = "poisson")
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.097565 -0.408948 -0.009874  0.369919  6.390181
##
## Count model coefficients (poisson with log link):
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.172e+00  5.681e-02  20.634  < 2e-16 ***
## AcidIndex         -1.778e-02  5.447e-03  -3.264   0.0011 **
## Alcohol            6.750e-03  1.623e-03   4.159 3.19e-05 ***
## CitricAcid        -1.464e-04  6.755e-03  -0.022   0.9827
## FreeSulfurDioxide  2.938e-05  3.953e-05   0.743   0.4574
## ResidualSugar     -2.547e-05  1.769e-04  -0.144   0.8855
## STARS              9.945e-02  5.831e-03  17.054  < 2e-16 ***
## pH                 2.559e-04  8.757e-03   0.029   0.9767
## LabelAppeal        2.339e-01  7.050e-03  33.181  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.1525260  0.3273535 -12.685  < 2e-16 ***
## AcidIndex          0.4404321  0.0280083  15.725  < 2e-16 ***
## Alcohol            0.0355244  0.0105305   3.373 0.000742 ***
## CitricAcid        -0.0607786  0.0437691  -1.389 0.164949
## FreeSulfurDioxide -0.0006042  0.0002664  -2.268 0.023305 *
## ResidualSugar     -0.0017613  0.0011373  -1.549 0.121473
## STARS             -2.3535880  0.0652372 -36.077  < 2e-16 ***
## pH                 0.2564845  0.0561110   4.571 4.85e-06 ***
## LabelAppeal        0.7292233  0.0472644  15.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 23
## Log-likelihood: -1.634e+04 on 18 Df
```

```
# Split Training Data into X, Y (Covariates and Response)
y_wine_train2 <- wine_train2$TARGET
x_wine_train2 <- wine_train2[, !colnames(wine_train2) %in% c("TARGET")]

# Make Predictions On Test Data
pois_pred <- predict(pois_glm2, newdata = x_wine_train2)
nb_pred <- predict(nb_glm2, newdata = x_wine_train2)
zip_pred <- predict(zip_glm2, newdata = x_wine_train2)
lm_pred <- predict(lm2, newdata = x_wine_train2)

# Calculate SSE
pois_sse <- sum((y_wine_train2 - pois_pred)^2)
nb_sse <- sum((y_wine_train2 - nb_pred)^2)
zip_sse <- sum((y_wine_train2 - zip_pred)^2)
lm_sse <- sum((y_wine_train2 - lm_pred)^2)

# Calculate MSE
pois_mse <- pois_sse / nrow(x_wine_train2)
nb_mse <- nb_sse / nrow(x_wine_train2)
zip_mse <- zip_sse / nrow(x_wine_train2)
lm_mse <- lm_sse / nrow(x_wine_train2)

print("Test Data:")
```

```
## [1] "Test Data:"
```

```
print(paste("Poisson Model MSE: ", pois_mse))
```

```
## [1] "Poisson Model MSE:  6.74473646791468"
```

```
print(paste("Negative Binomial Model MSE: ", nb_mse))
```

```
## [1] "Negative Binomial Model MSE:  6.7447317249851"
```

```
print(paste("ZIP Model MSE: ", zip_mse))
```

```
## [1] "ZIP Model MSE:  1.64247826189251"
```

```
print(paste("Linear Regression MSE: ", lm_mse))
```

```
## [1] "Linear Regression MSE:  1.77162427303882"
```

```
# Split Test Data into X, Y (Covariates and Response)
y_wine_test2 <- wine_test2$TARGET
x_wine_test2 <- wine_test2[, !colnames(wine_test2) %in% c("TARGET")]

# Make Predictions On Test Data
pois_pred <- predict(pois_glm2, newdata = x_wine_test2)
nb_pred <- predict(nb_glm2, newdata = x_wine_test2)
zip_pred <- predict(zip_glm2, newdata = x_wine_test2)
lm_pred <- predict(lm2, newdata = x_wine_test2)

# Calculate SSE
pois_sse <- sum((y_wine_test2 - pois_pred)^2)
nb_sse <- sum((y_wine_test2 - nb_pred)^2)
zip_sse <- sum((y_wine_test2 - zip_pred)^2)
lm_sse <- sum((y_wine_test2 - lm_pred)^2)

# Calculate MSE
pois_mse <- pois_sse / nrow(x_wine_test2)
nb_mse <- nb_sse / nrow(x_wine_test2)
zip_mse <- zip_sse / nrow(x_wine_test2)
lm_mse <- lm_sse / nrow(x_wine_test2)

print("Test Data:")
```

```
## [1] "Test Data:"
```

```
print(paste("Poisson Model MSE: ", pois_mse))
```

```
## [1] "Poisson Model MSE:  6.830540908078"
```

```
print(paste("Negative Binomial Model MSE: ", nb_mse))
```

```
## [1] "Negative Binomial Model MSE:  6.83053627489332"
```

```
print(paste("ZIP Model MSE: ", zip_mse))
```

```
## [1] "ZIP Model MSE:  1.61981088592182"
```

```
print(paste("Linear Regression MSE: ", lm_mse))
```

```
## [1] "Linear Regression MSE:  1.73103532879662"
```

```
summary(pois_glm2)
```

```
##
## Call:
## glm(formula = TARGET ~ AcidIndex + Alcohol + CitricAcid + FreeSulfurDioxide +
##     ResidualSugar + STARS + pH + LabelAppeal, family = poisson,
##     data = wine_train2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9598  -0.7340   0.0679   0.5704   3.2906
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.284e+00  5.399e-02  23.777  < 2e-16 ***
## AcidIndex         -8.932e-02  5.025e-03 -17.774  < 2e-16 ***
## Alcohol            7.831e-04  1.585e-03   0.494  0.62127
## CitricAcid         1.027e-02  6.612e-03   1.554  0.12023
## FreeSulfurDioxide  1.250e-04  3.935e-05   3.177  0.00149 **
## ResidualSugar      1.750e-04  1.723e-04   1.016  0.30975
## STARS              3.144e-01  5.066e-03  62.060  < 2e-16 ***
## pH                -2.356e-02  8.502e-03  -2.771  0.00559 **
## LabelAppeal        1.310e-01  6.776e-03  19.327  < 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: 18373  on 10235  degrees of freedom
## Residual deviance: 11908  on 10227  degrees of freedom
## AIC: 37426
##
## Number of Fisher Scoring iterations: 5
```

```
summary(nb_glm2)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ AcidIndex + Alcohol + CitricAcid +
##     FreeSulfurDioxide + ResidualSugar + STARS + pH + LabelAppeal,
##     data = wine_train2, init.theta = 48228.28505, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9597  -0.7340   0.0679   0.5704   3.2905
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.284e+00  5.399e-02  23.776  < 2e-16 ***
## AcidIndex         -8.932e-02  5.025e-03 -17.774  < 2e-16 ***
## Alcohol            7.830e-04  1.585e-03   0.494  0.62132
## CitricAcid         1.027e-02  6.612e-03   1.554  0.12024
## FreeSulfurDioxide  1.250e-04  3.936e-05   3.176  0.00149 **
## ResidualSugar      1.750e-04  1.723e-04   1.016  0.30975
## STARS              3.144e-01  5.066e-03  62.058  < 2e-16 ***
## pH                -2.356e-02  8.502e-03  -2.771  0.00559 **
## LabelAppeal        1.310e-01  6.776e-03  19.326  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48228.29) family taken to be 1)
##
##     Null deviance: 18372  on 10235  degrees of freedom
## Residual deviance: 11908  on 10227  degrees of freedom
## AIC: 37428
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48228
##           Std. Err.:  56006
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -37408.34
```

```
summary(zip_glm2)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ AcidIndex + Alcohol + CitricAcid + FreeSulfurDioxide +
##     ResidualSugar + STARS + pH + LabelAppeal, data = wine_train2, dist = "poisson")
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.097565 -0.408948 -0.009874  0.369919  6.390181
##
## Count model coefficients (poisson with log link):
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.172e+00  5.681e-02  20.634  < 2e-16 ***
## AcidIndex         -1.778e-02  5.447e-03  -3.264   0.0011 **
## Alcohol            6.750e-03  1.623e-03   4.159 3.19e-05 ***
## CitricAcid        -1.464e-04  6.755e-03  -0.022   0.9827
## FreeSulfurDioxide  2.938e-05  3.953e-05   0.743   0.4574
## ResidualSugar     -2.547e-05  1.769e-04  -0.144   0.8855
## STARS              9.945e-02  5.831e-03  17.054  < 2e-16 ***
## pH                 2.559e-04  8.757e-03   0.029   0.9767
## LabelAppeal        2.339e-01  7.050e-03  33.181  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.1525260  0.3273535 -12.685  < 2e-16 ***
## AcidIndex          0.4404321  0.0280083  15.725  < 2e-16 ***
## Alcohol            0.0355244  0.0105305   3.373 0.000742 ***
## CitricAcid        -0.0607786  0.0437691  -1.389 0.164949
## FreeSulfurDioxide -0.0006042  0.0002664  -2.268 0.023305 *
## ResidualSugar     -0.0017613  0.0011373  -1.549 0.121473
## STARS             -2.3535880  0.0652372 -36.077  < 2e-16 ***
## pH                 0.2564845  0.0561110   4.571 4.85e-06 ***
## LabelAppeal        0.7292233  0.0472644  15.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 23
## Log-likelihood: -1.634e+04 on 18 Df
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = TARGET ~ AcidIndex + Alcohol + CitricAcid + FreeSulfurDioxide +
##     ResidualSugar + STARS + pH + LabelAppeal, data = wine_train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5011 -0.9656  0.0538  0.9105  6.0712
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.302e+00  1.163e-01  28.386  < 2e-16 ***
## AcidIndex         -2.133e-01  1.014e-02 -21.025  < 2e-16 ***
## Alcohol            6.788e-03  3.643e-03   1.863 0.062460 .
## CitricAcid         2.754e-02  1.530e-02   1.799 0.071986 .
## FreeSulfurDioxide  3.273e-04  9.104e-05   3.595 0.000326 ***
## ResidualSugar      5.535e-04  3.975e-04   1.392 0.163874
## STARS              9.852e-01  1.175e-02  83.880  < 2e-16 ***
## pH                -5.994e-02  1.963e-02  -3.053 0.002268 **
## LabelAppeal        4.255e-01  1.534e-02  27.730  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.332 on 10227 degrees of freedom
## Multiple R-squared:  0.5233, Adjusted R-squared:  0.5229
## F-statistic:  1403 on 8 and 10227 DF,  p-value: < 2.2e-16
```

```
summary(wine_eval)
```

```
##   TARGET         FixedAcidity     VolatileAcidity     CitricAcid
##  Mode:logical   Min.   :-18.200   Min.   :-2.8300   Min.   :-3.1200
##  NA's:3335      1st Qu.:  5.200   1st Qu.: 0.0800   1st Qu.: 0.0000
##                 Median :  6.900   Median : 0.2800   Median : 0.3100
##                 Mean   :  6.864   Mean   : 0.3103   Mean   : 0.3124
##                 3rd Qu.:  9.000   3rd Qu.: 0.6300   3rd Qu.: 0.6050
##                 Max.   : 33.500   Max.   : 3.6100   Max.   : 3.7600
##
##  ResidualSugar        Chlorides        FreeSulfurDioxide TotalSulfurDioxide
##  Min.   :-128.300   Min.   :-1.15000   Min.   :-563.00   Min.   :-769.00
##  1st Qu.:  -2.600   1st Qu.: 0.01600   1st Qu.:   3.00   1st Qu.:  27.25
##  Median :   3.600   Median : 0.04700   Median :  30.00   Median : 124.00
##  Mean   :   5.319   Mean   : 0.06143   Mean   :  34.95   Mean   : 123.41
##  3rd Qu.:  17.200   3rd Qu.: 0.17100   3rd Qu.:  79.25   3rd Qu.: 210.00
##  Max.   : 145.400   Max.   : 1.26300   Max.   : 617.00   Max.   :1004.00
##  NA's   :168        NA's   :138        NA's   :152       NA's   :157
##     Density             pH          Sulphates          Alcohol
##  Min.   :0.8898   Min.   :0.600   Min.   :-3.0700   Min.   :-4.20
##  1st Qu.:0.9883   1st Qu.:2.980   1st Qu.: 0.3300   1st Qu.: 9.00
##  Median :0.9946   Median :3.210   Median : 0.5000   Median :10.40
##  Mean   :0.9947   Mean   :3.237   Mean   : 0.5346   Mean   :10.58
##  3rd Qu.:1.0005   3rd Qu.:3.490   3rd Qu.: 0.8200   3rd Qu.:12.50
##  Max.   :1.0998   Max.   :6.210   Max.   : 4.1800   Max.   :25.60
##                   NA's   :104     NA's   :310       NA's   :185
##   LabelAppeal         AcidIndex          STARS
##  Min.   :-2.00000   Min.   : 5.000   Min.   :1.00
##  1st Qu.:-1.00000   1st Qu.: 7.000   1st Qu.:1.00
##  Median : 0.00000   Median : 8.000   Median :2.00
##  Mean   : 0.01349   Mean   : 7.748   Mean   :2.04
##  3rd Qu.: 1.00000   3rd Qu.: 8.000   3rd Qu.:3.00
##  Max.   : 2.00000   Max.   :17.000   Max.   :4.00
##                                       NA's   :841
```

```
# Impute eval data
wine_eval_imputed <- wine_eval %>%
  mutate(STARS = ifelse(is.na(STARS), 0, STARS),
         across(everything(), ~replace(., is.na(.), mean(., na.rm = TRUE))))



# Split Test Data into X, Y (Covariates and Response)
x_wine <- wine_eval_imputed[, !colnames(wine_eval_imputed) %in% c("TARGET")]

# Make Predictions On Eval Data
zip_pred <- predict(zip_glm2, newdata = x_wine)

wine_eval2 <- wine_eval %>%
  mutate(TARGET = zip_pred)

head(wine_eval2)
```

```
##      TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
## 1 1.5723395          5.4          -0.860       0.27         -10.7     0.092
## 2 3.9103667         12.4           0.385      -0.76         -19.7     1.169
## 3 2.6631347          7.2           1.750       0.17         -33.0     0.065
## 4 2.4903425          6.2           0.100       1.80           1.0    -0.179
## 5 0.7461585         11.4           0.210       0.28           1.2     0.038
## 6 5.6513043         17.6           0.040      -1.15           1.4     0.535
##   FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates Alcohol
## 1                23                398 0.98527 5.02      0.64   12.30
## 2               -37                 68 0.99048 3.37      1.09   16.00
## 3                 9                 76 1.04641 4.61      0.68    8.55
## 4               104                 89 0.98877 3.20      2.11   12.30
## 5                70                 53 1.02899 2.54     -0.07    4.80
## 6              -250                140 0.95028 3.06     -0.02   11.40
##   LabelAppeal AcidIndex STARS
## 1          -1         6    NA
## 2           0         6     2
## 3           0         8     1
## 4          -1         8     1
## 5           0        10    NA
## 6           1         8     4
```

1 Works Cited

CDT Pimentel, Adrian H3 25'. Assistance Given to Author. Oral Assistance Givn, CDT Pimentel, and I were talking, and I was talking about my tendency to choose too many variables for my models, and he suggested that I use ChatGPT for variable selection, so I did.

Chat GPT. Assistance was given to the author. Written Assistance Given. I provided the following input to ChatGPT to request that it provide me with variables to use in my model best to predict wine purchases. The portion labeled ChatGPT Output was provided. I used ChatGPT's suggestions for variable selection in the second version of all of my models.

Chat GPT Input:

I am building a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine. The wine properties that I have available to me are as follows: (1) AcidIndex, a proprietary method of testing the total acidity of wine by using a weighted average. (2) Alcohol Content. (3) Chloride content of wine. (4) Citric Acid Content of wine (5) Density of Wine (6) Fized Acidity of Wine (7) Sulfur Dioxide content of wine (8) Label Appel (9) Residual Sugar of the wine (10) STARS, wine rating by a team of 4 experts, 4=Excellent 1=Poor, 0= Not rated. (11) Sulfate content of wine, (12) Total Sulfer Dioxide of Wine, (13) Volatile Acidity content of wine, and (14) PH of wine. Please choose the best properties that I should add as variables into my count regression model to best predict the target.

Chat GPT Output:

Building a count regression model for predicting the number of wine cases sold involves selecting relevant features that are likely to have an impact on the outcome. In the context of count regression, Poisson regression is a commonly used model. Here are some considerations for selecting variables:

AcidIndex: This variable represents the acidity of the wine, which might influence the consumer's preference. Including it in the model could be beneficial.

Alcohol Content: The alcohol content of wine can significantly affect its taste and popularity. It is a good candidate for inclusion in your model.

Citric Acid Content: Citric acid contributes to the overall acidity and flavor profile of the wine. Including this variable may capture its impact on sales.

Sulfur Dioxide Content: Both sulfur dioxide variables (Sulfur Dioxide content and Total Sulfur Dioxide) may be important, as they can affect the wine's stability and shelf life.

Residual Sugar: This variable reflects the sweetness of the wine, which is a crucial factor in consumer preferences. Including it could enhance the model.

STARS (Wine Rating): Consumer preferences often align with expert ratings. The STARS variable could provide valuable information about the perceived quality of the wine.

pH of Wine: The pH level influences the taste and chemical stability of the wine. Including pH in your model might capture its impact on sales.

Label Appeal: The label's appeal can influence consumer choices. If this variable is a good indicator of marketability, consider including it.

Before finalizing your variables, it's essential to check for multicollinearity among them, as highly correlated predictors can impact the model's performance. Additionally, consider testing different combinations and subsets of variables to see which set provides the best predictive power for your specific dataset. Always validate your model using appropriate evaluation metrics and ensure it generalizes well to new data.