

Homework 5

Group 1

Contents

1	Introduction	2
2	Statement of the Problem	2
3	Data Exploration	2
3.1	Variables Explained	2
3.2	Variables Summary Statistics	3
3.3	Imputing Missing Values	4
3.4	Correlation of Variables	6
4	Data Transformation	7
4.1	Outliers Treatment	7
4.2	BoxCox Transformations	9
5	Models Built	13
5.1	Poisson Regression Models	13
5.2	Binomial Regression Models	21
5.3	Multiple Linear Regression	23
6	Selected Model	29
7	Appendix A	30
7.1	Session Info	30
7.2	Data Dictionary	30
7.3	R source code	31

Prepared for:

Dr. Nathan Bastian

City University of New York, School of Professional Studies - Data 621

Prepared by:

Group 1

Senthil Dhanapal

Yadu Chittampalli

Christophe Hunt

1 Introduction

The wine industry was valued at \$257.5 billion in 2012 and is predicted to be valued at \$303.6 billion by 2016.¹ As wine is a consumer product, accommodating consumer preference is critical to maintaining a competitive advantage. By understanding the factors involved in wine sales we can better understand consumer behavior and adjust our strategies accordingly.

2 Statement of the Problem

The purpose of this report is to develop statistical models to make inference into the factors associated with the number of cases of wine sold.

3 Data Exploration

3.1 Variables Explained

The variables provided in the Wine Training Data Set are explained below:

Variable Code	Definition
INDEX	Identification Variable (do not use)
TARGET	Number of Cases Purchased
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average
Alcohol	Alcohol Content
Chlorides	Chloride content of wine
CitricAcid	Citric Acid Content
Density	Density of Wine
FixedAcidity	Fixed Acidity of Wine
FreeSulfurDioxide	Sulfur Dioxide content of wine
LabelAppeal	Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customers don't like the design.
ResidualSugar	Residual Sugar of wine
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor
Sulphates	Sulfate content of wine
TotalSulfurDioxide	Total Sulfur Dioxide of Wine
VolatileAcidity	Volatile Acid content of wine
pH	pH of wine

¹"Research and Markets: Wine: 2012 Global Industry Almanac - The Global Wine Market Grew by 3.1% in 2011 to Reach a Value of \$257.5 Billion." Research and Markets: Wine: 2012 Global Industry Almanac - The Global Wine Market Grew by 3.1% in 2011 to Reach a Value of \$257.5 Billion | Business Wire. N.p., 21 May 2012. Web. 20 Nov. 2016.

3.2 Variables Summary Statistics

3.2.1 Discrete Variables

Interestingly, we can see some general sense of the make up of our data set. In this set, most wines sell between 3 and 5 cases, have no label appeal, and very few received 4 stars with most wines receiving 2 or 1 stars. Additionally, we should note that 21.4% of our wines had no case sales.

Table 2: Wine Training Data Table of Discrete Variables

Variable	Levels	n	%	$\sum\%$
TARGET	0	2734	21.4	21.4
	1	244	1.9	23.3
	2	1091	8.5	31.8
	3	2611	20.4	52.2
	4	3177	24.8	77.0
	5	2014	15.7	92.8
	6	765	6.0	98.8
	7	142	1.1	99.9
	8	17	0.1	100.0
all		12795	100.0	
LabelAppeal	-2	504	3.9	3.9
	-1	3136	24.5	28.5
	0	5617	43.9	72.3
	1	3048	23.8	96.2
	2	490	3.8	100.0
	all		12795	100.0
STARS	1	3042	32.2	32.2
	2	3570	37.8	70.1
	3	2212	23.4	93.5
	4	612	6.5	100.0
	all		9436	100.0

3.2.2 Continous Variables

We see that Density is a very narrow measurement, the minimum value is 0.9 and the maximum is 1.1. The remaining continuous variables appear to have a larger range of variability, with the largest being TotalSulfurDioxide which has a range from -823 to 1057. In our models, this variability will provide some insights to our coefficients and the impact to the dependent variable.

Table 3: Wine Training Data Table of Continuous Variables

Variable	n	Min	q ₁	\tilde{x}	\bar{x}	q ₃	Max	s	IQR	#NA
FixedAcidity	12795	-18.1	5.2	6.9	7.1	9.5	34.4	6.3	4.3	0
VolatileAcidity	12795	-2.8	0.1	0.3	0.3	0.6	3.7	0.8	0.5	0
CitricAcid	12795	-3.2	0.0	0.3	0.3	0.6	3.9	0.9	0.5	0
ResidualSugar	12179	-127.8	-2.0	3.9	5.4	15.9	141.2	33.7	17.9	616
Chlorides	12157	-1.2	0.0	0.0	0.1	0.2	1.4	0.3	0.2	638
FreeSulfurDioxide	12148	-555.0	0.0	30.0	30.8	70.0	623.0	148.7	70.0	647
TotalSulfurDioxide	12113	-823.0	27.0	123.0	120.7	208.0	1057.0	231.9	181.0	682
Density	12795	0.9	1.0	1.0	1.0	1.0	1.1	0.0	0.0	0
pH	12400	0.5	3.0	3.2	3.2	3.5	6.1	0.7	0.5	395
Sulphates	11585	-3.1	0.3	0.5	0.5	0.9	4.2	0.9	0.6	1210
Alcohol	12142	-4.7	9.0	10.4	10.5	12.4	26.5	3.7	3.4	653

3.3 Imputing Missing Values

In order to address the missing values in our variables we used a non-parametric imputation method (Random Forest) using the `missForest` package. The function is particularly useful in that it can handle any type of input data and it will make as few assumptions about the structure of the data as possible.²

**Table 4 : Imputed Descriptive Statistics
13 Variables 12795 Observations**

FixedAcidity																			
12795	n	missing	0	distinct	470	Info	1	Mean	7	Gmd	7	.05	.10	.25	.50	.75	.90	.95	.18
lowest : -18.1 -18.0 -17.7 -17.5 -17.4, highest: 32.4 32.5 32.6 34.1 34.4																			
VolatileAcidity																			
12795	n	missing	0	distinct	815	Info	1	Mean	0.3	Gmd	0.8	.05	.10	.25	.50	.75	.90	.95	.16
lowest : -2.790 -2.750 -2.745 -2.730 -2.720, highest: 3.500 3.550 3.565 3.590 3.680																			
CitricAcid																			
12795	n	missing	0	distinct	602	Info	1	Mean	0.3	Gmd	0.9	.05	.10	.25	.50	.75	.90	.95	.179
lowest : -3.24 -3.16 -3.10 -3.08 -3.06, highest: 3.63 3.68 3.70 3.77 3.86																			
ResidualSugar																			
12795	n	missing	0	distinct	2685	Info	1	Mean	5	Gmd	34	.05	.10	.25	.50	.75	.90	.95	.62.1
lowest : -127.80 -127.10 -126.20 -126.10 -125.70 highest: 136.50 137.60 138.00 140.65 141.15																			
Chlorides																			
12795	n	missing	0	distinct	2285	Info	1	Mean	0.05	Gmd	0.3	.05	.10	.25	.50	.75	.90	.95	.59
lowest : -1.171 -1.170 -1.158 -1.156 -1.155, highest: 1.260 1.261 1.270 1.275 1.351																			
FreeSulfurDioxide																			
12795	n	missing	0	distinct	1626	Info	1	Mean	31	Gmd	150	.05	.10	.25	.50	.75	.90	.95	.281
lowest : -555 -546 -536 -535 -532, highest: 613 617 618 622 623																			
TotalSulfurDioxide																			
12795	n	missing	0	distinct	2039	Info	1	Mean	121	Gmd	238	.05	.10	.25	.50	.75	.90	.95	.507
lowest : -823 -816 -793 -781 -779, highest: 1032 1041 1048 1054 1057																			
Density																			
12795	n	missing	0	distinct	5933	Info	1	Mean	1	Gmd	0.03	.05	.10	.25	.50	.75	.90	.95	.1.0
lowest : 0.88809 0.88949 0.88978 0.88983 0.89167 highest: 1.09658 1.09679 1.09695 1.09791 1.09924																			
pH																			
12795	n	missing	0	distinct	863	Info	1	Mean	3	Gmd	0.7	.05	.10	.25	.50	.75	.90	.95	.4
lowest : 0.48 0.53 0.54 0.58 0.59, highest: 5.91 5.94 6.02 6.05 6.13																			

²Stekhoven, Daniel J., and Peter B?hlmann. "MissForest-non-parametric missing value imputation for mixed-type data." Bioinformatics 28.1 (2012): 112-118.

Sulphates

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
12795	0	1695	1	0.5	0.9	-1.0	-0.6	0.3	0.5	0.8	1.7	2.1

lowest : -3.13 -3.12 -3.10 -3.07 -3.03, highest: 4.11 4.16 4.19 4.21 4.24

Alcohol

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
12795	0	1036	1	10	4	.05	.4	.6	.9	.10	.12	.15	.17

lowest : -4.7 -4.5 -4.4 -4.3 -4.1, highest: 25.4 25.6 26.0 26.1 26.5

LabelAppeal

n	missing	distinct	Info	Mean	Gmd
12795	0	5	0.887	-0.009	1

lowest : -2 -1 0 1 2, highest: -2 -1 0 1 2

-2 (504, 0.039), -1 (3136, 0.245), 0 (5617, 0.439), 1 (3048, 0.238), 2 (490, 0.038)

STARS

n	missing	distinct
12795	0	4

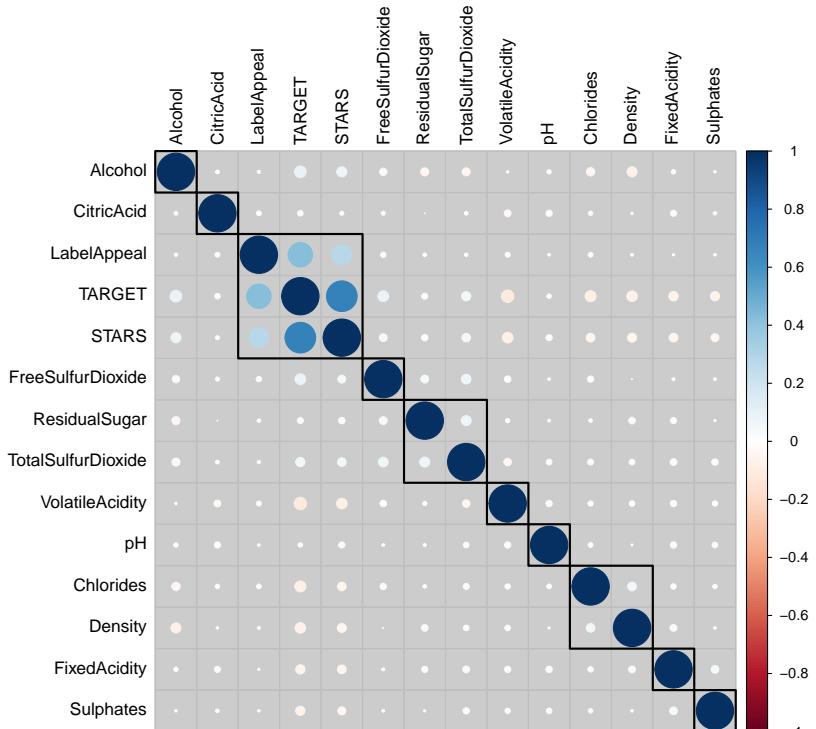
1 (5305, 0.415), 2 (4569, 0.357), 3 (2309, 0.180), 4 (612, 0.048)

3.4 Correlation of Variables

3.4.1 Correlation Matrix

If we modify our data frame to a matrix in our evaluation data set we can further plot a correlation matrix. There are surprisingly few interesting correlations in the data, but the lack of correlation in the data set is in itself interesting.

- STARS has the most positive correlation and strongest correlation with our dependent variable TARGET. It is intuitive that the greater the STARS value the more cases our wine would sell.
- LabelAppeal is the second most correlated with our dependent variable to our dependent variable. It is interesting that the two most correlated variables have less to do with wine quality and more to do with the appearance of a sophisticated wine.
- The lack of strong correlations is interesting in itself. It is concerning that most variables have nearly no correlation with our dependent variable but represent the actual quality of the wine. We see that public perception of wine is more important than the actual quality of the wine as measured by these variables.

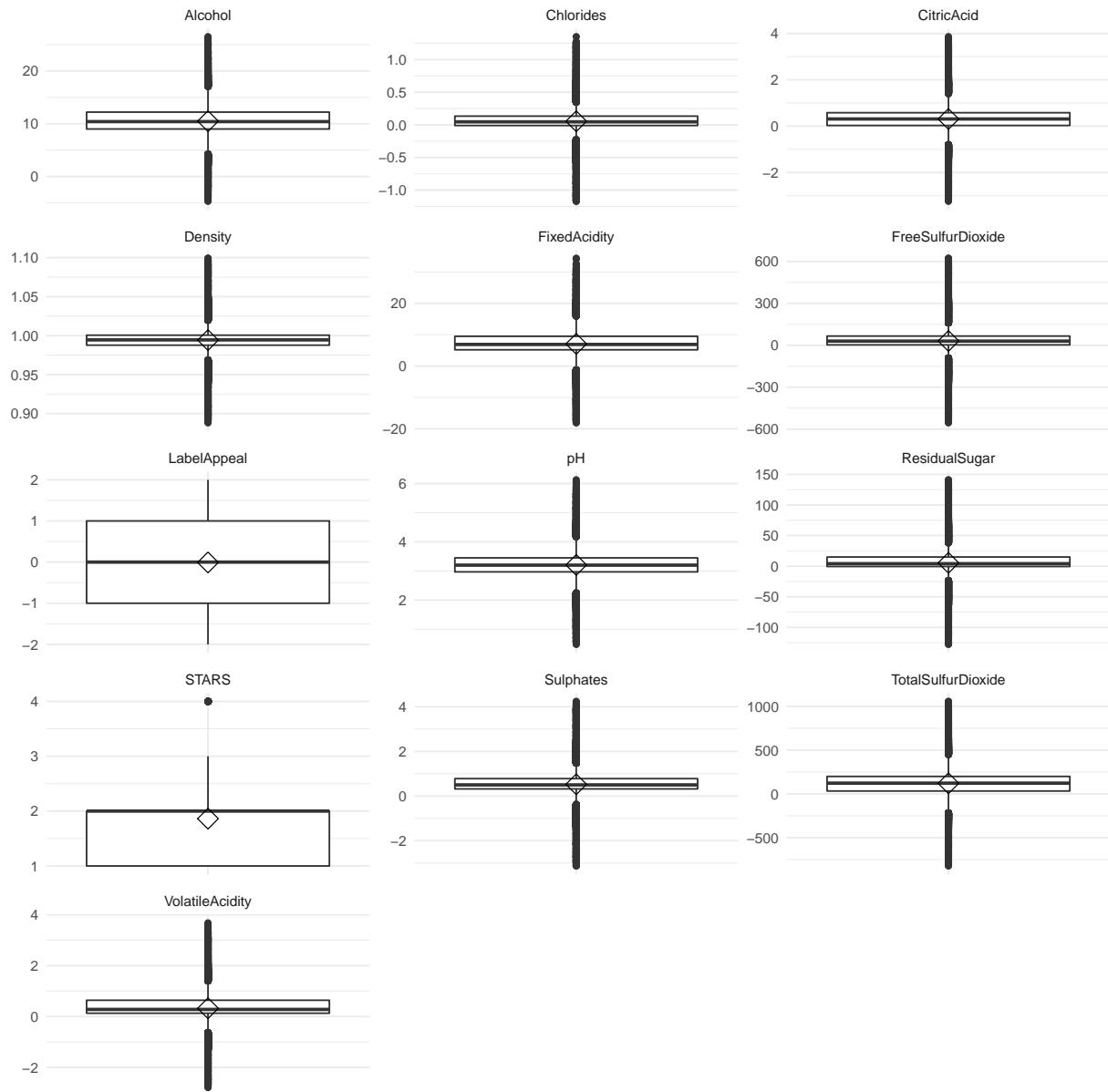


4 Data Transformation

4.1 Outliers Treatment

4.1.1 Box Plots of Variables for Winsorizing

Box Plots provide a visualization of the quartiles and outliers of our data set.³ Using the box plots, we can conclude that the variables to be winsorized are Free Sulfur Dioxide, Residual Sugar, and Total Sulfur Dioxide.



³"Box Plot." Wikipedia. Wikimedia Foundation, n.d. Web. 24 Nov. 2016.

4.1.2 Winsorizing

We chose winsorizing as the method to address outliers. Instead of trimming values, winsorizing uses the interquartile range to replace values that are above or below the interquartile range multiplied by a factor. Those values above or below the range multiplied by the factor are then replaced with max and min value of the interquartile range. Using the factor 2.2 for winsorizing outliers is a method developed by Hoaglin and Iglewicz and published Journal of American Statistical Association in 1987⁴.

The below table is the summary results of the winsorizing of the data.

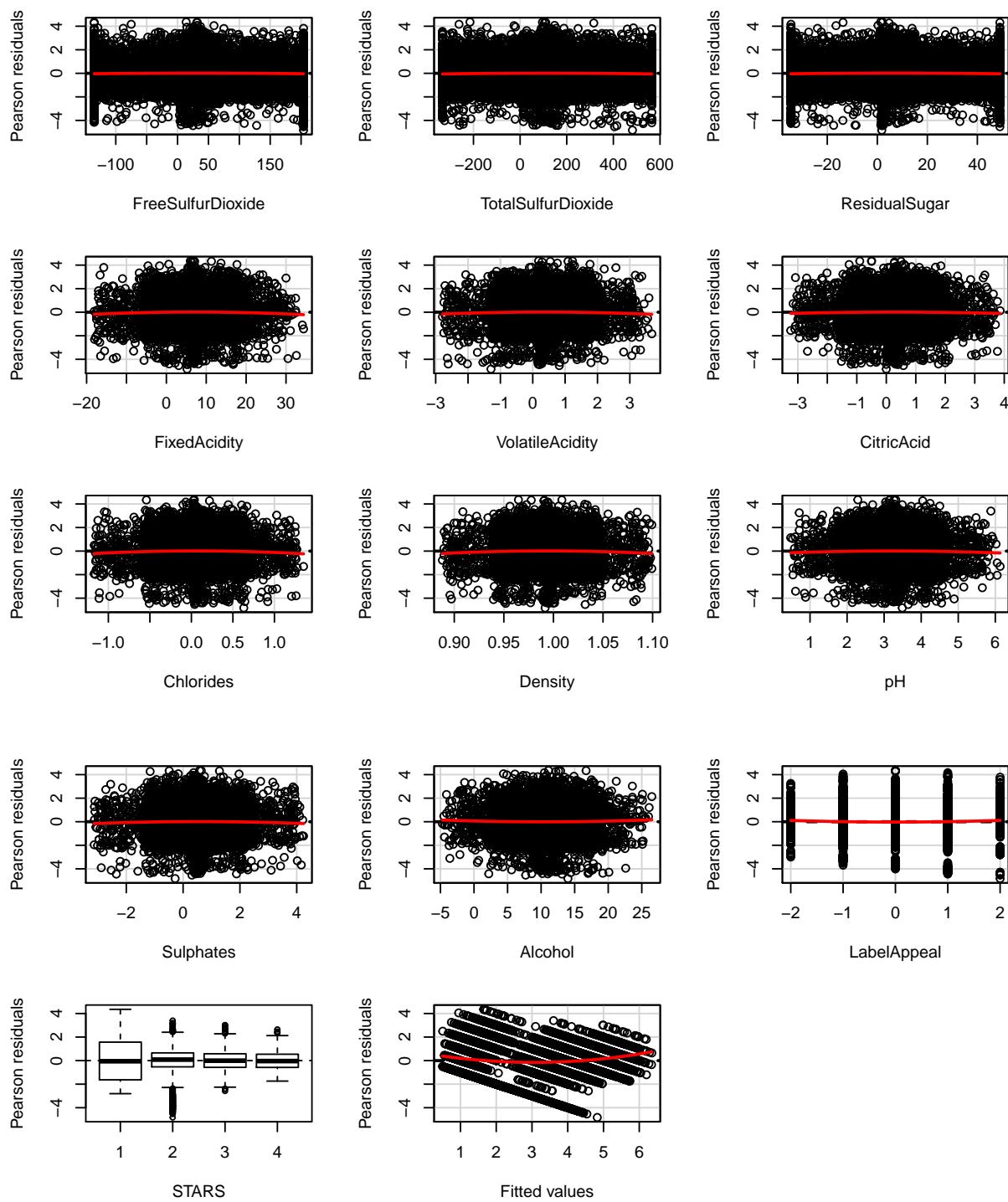
Table 4:

Statistic	N	Mean	St. Dev.	Min	Max
FreeSulfurDioxide	12,796	31.978	99.033	-135.000	204.000
TotalSulfurDioxide	12,796	120.521	203.181	-333.000	565.000
ResidualSugar	12,796	5.927	23.816	-34.600	49.100
TARGET	12,796	3.029	1.926	0	8
FixedAcidity	12,796	7.075	6.317	-18.100	34.400
VolatileAcidity	12,796	0.324	0.784	-2.790	3.680
CitricAcid	12,796	0.308	0.862	-3.240	3.860
Chlorides	12,796	0.055	0.313	-1.171	4.000
Density	12,796	0.994	0.032	0.888	3.000
pH	12,796	3.208	0.670	0.480	6.130
Sulphates	12,796	0.527	0.888	-3.130	4.240
Alcohol	12,796	10.489	3.636	-4.700	26.500
LabelAppeal	12,796	-0.009	0.891	-2	3

⁴Hoaglin, D. C., and Iglewicz, B. (1987), Fine tuning some resistant rules for outlier labeling, Journal of American Statistical Association, 82, 1147-1149.

4.2 BoxCox Transformations

Even after Winsorization we see non-constant variance in the Pearson Residuals for `FreeSulferDioxide`, `TotalSulfurDioxide`, and `ResidualSugar`. The Box-Cox evaluation was completed on these variables, based on the residual plots. In the residual plots, these three variables showed a great deal of non-constant variance because the plots were hyperbolic-shaped.



```
##               Test stat Pr(>|t|)  
## FreeSulfurDioxide      -1.892   0.058  
## TotalSulfurDioxide     -1.751   0.080  
## ResidualSugar         -2.101   0.036  
## FixedAcidity          -1.881   0.060  
## VolatileAcidity       -1.694   0.090  
## CitricAcid            -1.092   0.275  
## Chlorides              -2.370   0.018  
## Density                -2.286   0.022  
## pH                     -1.500   0.134  
## Sulphates              -1.616   0.106  
## Alcohol                1.408   0.159  
## LabelAppeal            3.071   0.002  
## STARS                  NA      NA  
## Tukey test             17.998  0.000
```

4.2.1 Determining BoxCox Transformations

Using the `BoxCox.lambda` function from the `forecast` package we are able to determine our necessary transformations to our independent variables.

λ	Variables
1.22449234379866	Free Sulfur Dioxide
1.0182875042235	Total Sulfur Dioxide
1.18389893233879	Residual Sugar

Utilizing transformations based on the lambda value of the BoxCox and rounding to the nearest tenth we further transform our independent variables for our regression models. We see that the `TotalSulfurDioxide` variable does not require further transformation

Box-Cox Transformations ⁵	
λ	Y'
0	$\log(Y)$
.25	$\sqrt[4]{Y}$
0.5	$Y^{0.5} = \sqrt{(Y)}$
1	$Y^1 = Y$
1.25	$Y^{1.25}$

variable	variable transformation
ResidualSugar	$ResidualSugar^{1.25}$
FreeSulfurDioxide	$FreeSulfurDioxide^{1.25}$

⁵Osborne, Jason W. "Improving your data transformations: Applying the Box-Cox transformation." Practical Assessment, Research & Evaluation 15.12 (2010): 1-9.

5 Models Built

5.1 Poisson Regression Models

First, we investigate the unconditional variance is slightly > unconditional mean. Which we do see in the below table so there may be some over-dispersion.

mean	var
3.029074	3.710895

5.1.1 Poisson Regression Model 1

We build out first Poisson Regression model but we need to verify the confidence levels are appropriate. After producing the Confidence level for ResidualSugar we see it runs through 1 and its P(z) value is clearly not significant. Based on both items we can remove ResidualSugar from the model.

Table 7: Poisson Regression Model 1

<i>Dependent variable:</i>	
	TARGET
Constant	0.406*** (0.018)
STARS2	0.825*** (0.013)
STARS3	1.064*** (0.015)
STARS4	1.229*** (0.021)
Alcohol	0.005*** (0.001)
ResidualSugar	0.00005 (0.0002)
Observations	12,795
Log Likelihood	-23,658.340
Akaike Inf. Crit.	47,328.690
Residual Deviance	15,374.670 (df = 12789)
Null Deviance	22,860.890 (df = 12794)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: Confidence intervals

	2.5%	97.5%
(Intercept)	1.5	1.447
STARS2	2.283	2.224
STARS3	2.898	2.816
STARS4	3.419	3.284
Alcohol	1.005	1.002
ResidualSugar	1	0.9997

We removed ResidualSugar and Model with the independent variables STARS and Alcohol. We can see how great the STARS variable has on our model, 2 STARS is 2.28 times wine case sales when compared to one STAR. Furthermore, 3 Stars will have wine case sales of 2.90 times more than one STAR, and four STARS will be 3.42 times more than one STAR. Alcohol = For one unit increase in Alcohol, Wine sales will be 1.004 times more

Table 9: Poisson Regression Model 1 without ResidualSugar

Dependent variable:	
	TARGET
Constant	0.406*** (0.018)
STARS2	0.825*** (0.013)
STARS3	1.064*** (0.015)
STARS4	1.229*** (0.021)
Alcohol	0.005*** (0.001)
Observations	12,795
Log Likelihood	-23,658.390
Akaike Inf. Crit.	47,326.790
Residual Deviance	15,374.760 (df = 12790)
Null Deviance	22,860.890 (df = 12794)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 10: Confidence intervals

	2.5%	97.5%
(Intercept)	1.501	1.448
STARS2	2.283	2.224
STARS3	2.898	2.816
STARS4	3.419	3.284
Alcohol	1.005	1.002

5.1.1.1 Poisson Regression Model 1 Metrics

5.1.1.1.1 Dispersion

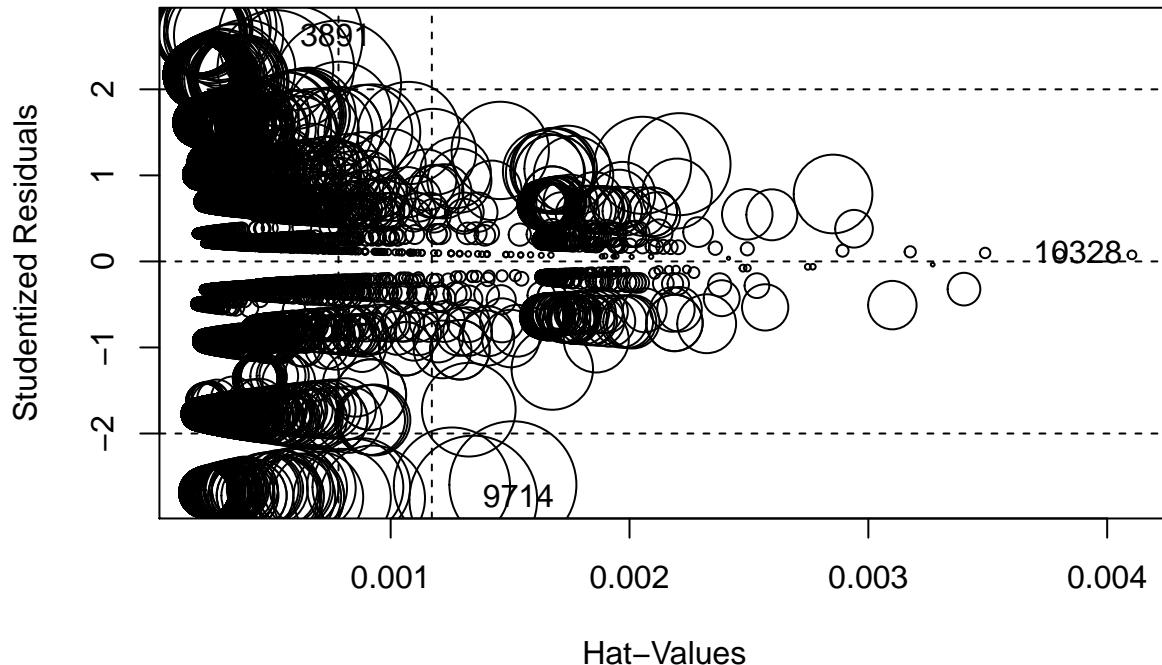
dispersion 0.9730167

Our dispersion results are very close to 1. So, we can say it is not over or under dispersed. However, we can further test for overdispersion by dividing the deviance of our model by the residuals. The result of this test is 1.2020925 and since this result is not greater than 1.5, we can claim that the data is not over-dispersed.⁶

⁶“Multiple Logistic Regression.” R Companion: Multiple Logistic Regression. N.p., n.d. Web. 04 Dec. 2016.

5.1.1.2 Influential Points

We further test for influential points in the data set. This test indicates that rows 10328, 9714, and 3891, have great influence on our model. It would be important to discuss these rows with the appropriate data steward to understand if these are accurate measurements and should be included in the analysis. Due to time limitations, we are not able to verify these rows and they have been included in this analysis.



```
##           StudRes      Hat      CookD
## 3891    2.59248567 0.0005804453 1.31210e-03
## 9714   -2.76735034 0.0013458546 1.03277e-03
## 10328   0.07597296 0.0041037245 4.80641e-06
```

5.1.1.3 Verifying Predictions

We also verify predicted values for the training dataset, in order to verify the output of our model against the training data set.

TARGET	STARS	Alcohol	Fitted
3	2	9.9	3.596
3	3	10.07	4.569
5	3	22	4.844
3	1	6.2	1.547
4	2	13.7	3.663
0	1	15.4	1.618

The predictions are close in value, we can further see the prediction quality of the model by reviewing the frequency table for observed vs predicted values.

Target	Obs	Predicted
0	2734	0
1	244	51
2	1091	5254
3	2611	262
4	3177	4574
5	2014	2553
6	765	101
7	142	0
8	17	0

Goodness of Fit Test

Goodness of Fit Test using Pearson Chi square test gives us the result of 0 shows that our model is good.

5.1.2 Poisson Regression Model 2

As in our first Poisson model, we build our second Poisson Regression model and we need to verify the confidence levels are appropriate. After producing the Confidence level for Confidence level for CitricAcid and pH, we see that they run through 1 and and their P(z) values are clearly not significant. Based on both info we can remove CitricAcid and pH from the model.

Table 13: Poisson Regression Model 2

	<i>Dependent variable:</i>
	TARGET
Constant	0.461*** (0.044)
LabelAppeal-1	0.383*** (0.038)
LabelAppeal0	0.685*** (0.037)
LabelAppeal1	0.909*** (0.037)
LabelAppeal2	1.092*** (0.042)
CitricAcid	0.004 (0.006)
pH	-0.011 (0.008)
Observations	12,795
Log Likelihood	-26,380.660
Akaike Inf. Crit.	52,775.320
Residual Deviance	20,819.300 (df = 12788)
Null Deviance	22,860.890 (df = 12794)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 14: Confidence intervals

	2.5 %	97.5 %
(Intercept)	1.585	1.455
LabelAppeal-1	1.467	1.363
LabelAppeal0	1.983	1.846
LabelAppeal1	2.481	2.309
LabelAppeal2	2.98	2.747
CitricAcid	1.004	0.9926
pH	0.9894	0.9747

We removed CitricAcid and pH then created a new Model with only LabelAppeal. The impact of LabelAppeal is very significant since this variable is explaining a great deal of variation in our dependent variable. A neutral LabelAppeal of 0 will have wine sales 1.98 times greater than a very negative LabelAppeal of -2. Also, a great LabelAppeal of 2 will have 2.98 times greater wine sales than a than a very negative LabelAppeal of -2.

Table 15: Poisson Regression Model 2 with LabelAppeal

<i>Dependent variable:</i>	
	TARGET
Constant	0.428*** (0.036)
LabelAppeal-1	0.383*** (0.038)
LabelAppeal0	0.685*** (0.037)
LabelAppeal1	0.909*** (0.037)
LabelAppeal2	1.092*** (0.042)
Observations	12,795
Log Likelihood	-26,381.890
Akaike Inf. Crit.	52,773.780
Residual Deviance	20,821.760 (df = 12790)
Null Deviance	22,860.890 (df = 12794)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16: Confidence intervals

	2.5%	97.5%
(Intercept)	1.534	1.428
LabelAppeal-1	1.467	1.363
LabelAppeal0	1.983	1.847
LabelAppeal1	2.482	2.31
LabelAppeal2	2.979	2.747

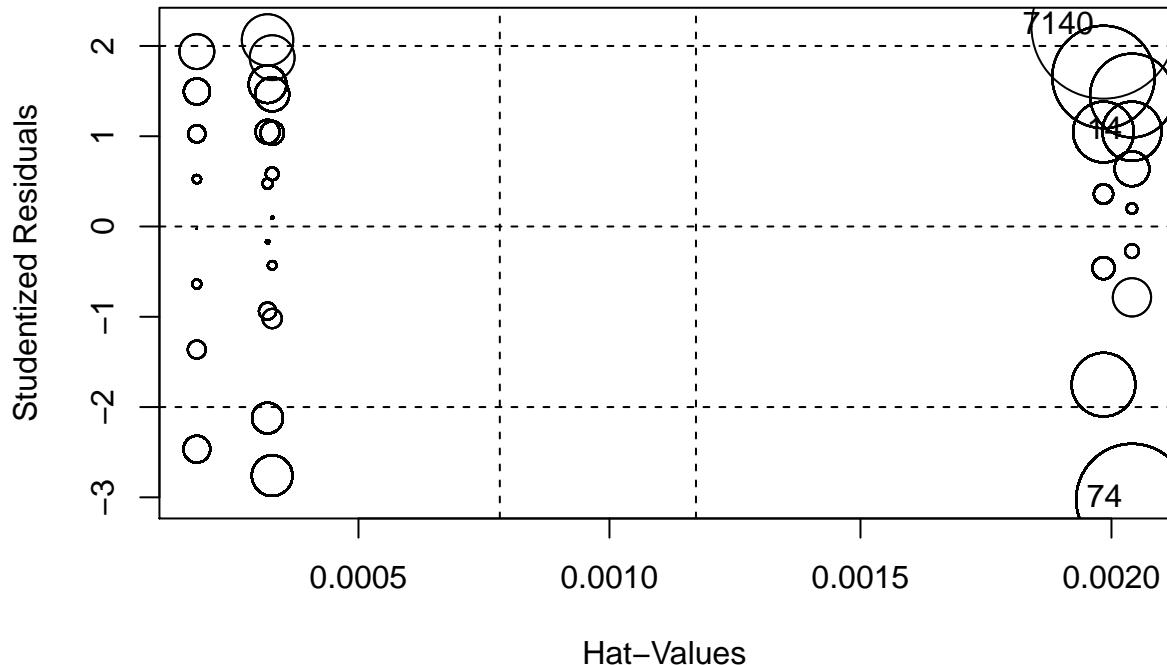
5.1.2.1 Poisson Regression Model 2 Metrics

dispersion 1.035287

Our dispersion results are very close to 1. So, we can say it is not over or under dispersed. However, we can further test for overdispersion by dividing the deviance of our model by the residuals. The result of this test is 1.6279715 and since this result is greater than 1.5, we can claim that the data is over-dispersed.

5.1.2.1.1 Influential points

We further test for influential points in the data set. This test indicates that rows 7140, 14, and 74, have great influence on our model. It would be important to discuss these rows with the appropriate data steward to understand if these are accurate measurements and should be included in the analysis. Due to time limitations, we are not able to verify these rows and they have been included in this analysis.



```
##          StudRes      Hat      CookD
## 14     1.054918 0.002040816 0.0005298851
## 74    -3.024586 0.002040816 0.0018726921
## 7140   2.213682 0.001984127 0.0031210472
```

5.1.2.1.2 Verifying Predictions

We also verify predicted values for the training dataset, in order to verify the output of our model against the training data set.

TARGET	LabelAppeal	Alcohol	Fitted
3	0	9.9	3.042
3	-1	10.07	2.25
5	-1	22	2.25
3	-1	6.2	2.25
4	0	13.7	3.042
0	0	15.4	3.042

The predictions are close in value, we can further see the prediction quality of the model by reviewing the frequency table for observed vs predicted values.

Target	Obs	Predicted
0	2734	0
1	244	0
2	1091	3640
3	2611	5617
4	3177	3048
5	2014	490
6	765	0
7	142	0
8	17	0

Goodness of fit test

The goodness of fit test using pearson Chisquare test results are 0 which shows that our model is good and statistically significant.

5.2 Binomial Regression Models

5.2.1 Binomial Regression Model 1

In the first negative binomial regression model, all of the coefficients are positive. The variable that had to be removed was wine rating, due to the fact that it led to an error in the model, stating that the iteration limit was reached. Even though this categorical variable guarantees high significance and also higher coefficients (0.4 for STARS = 2, 0.6 for STARS = 3, and 0.7 for STARS = 4), this variable would not be appropriate to use for negative binomial regression. The alcohol content is also an equally significant variable but does not have a coefficient as high as those of the wine rating. However, this variable and residual sugar can be used for negative binomial regression because the resulting overdispersion is not so high.

Table 19: Binomial Regression Model 1

	<i>Dependent variable:</i> TARGET	
	(1)	(2)
Constant	0.405*** (0.018)	0.987*** (0.019)
STARS2	0.825*** (0.013)	
STARS3	1.064*** (0.015)	
STARS4	1.229*** (0.021)	
Alcohol	0.005*** (0.001)	0.011*** (0.002)
ResidualSugar	0.0002 (0.0002)	0.001** (0.0003)
Observations	12,795	12,796
Log Likelihood	-23,659.120	-27,142.350
θ	39,834.370 (37,787.650) ($p = 0.292$)	7.521*** (0.435) ($p = 0.000$)
Akaike Inf. Crit.	47,330.240	54,290.700
Residual Deviance	15,373.240 (df = 12789)	18,163.620 (df = 12793)
Null Deviance	22,859.700 (df = 12794)	18,214.010 (df = 12795)

Note:

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

5.2.1.1 Binomial Regression Model 1 Evaluation

In the model-fitting criterion, the chi-squared p-value is 1. This implies a failure to reject the null hypothesis.

residual.deviance	residual.degrees.of.freedom	chisq.p.value
15374	12790	2.201e-52

5.2.2 Binomial Regression Model 2

In the second negative binomial regression model, all of the coefficients are positive except for that of pH. The only significant variable is the label appeal. Except for the score of 3, all of the other scores for label appeal, yield significant results. Most of the coefficients for label appeal are close to 1 or slightly greater than 1 (0.7 for Label Appeal = 0, 0.9 for Label Appeal = 1, 1.09 for Label Appeal = 2, and 0.7 for Label Appeal = 3). The only score that yeilds a coefficient that is less than 1 is -1. The coefficient for this is 0.4. The standard error is 3.5. The theta value is 23.46, guaranteeing a lower level of overdispersion.

Table 21: Binomial Regression Model 2

	<i>Dependent variable:</i>	
	TARGET	
	(1)	(2)
Constant	0.462*** (0.045)	0.428*** (0.037)
CitricAcid	0.004 (0.006)	
pH	-0.011 (0.008)	
LabelAppeal1	0.383*** (0.039)	0.383*** (0.039)
LabelAppeal0	0.685*** (0.038)	0.685*** (0.038)
LabelAppeal1	0.909*** (0.038)	0.909*** (0.038)
LabelAppeal2	1.092*** (0.044)	1.092*** (0.044)
LabelAppeal3	0.703 (0.615)	0.671 (0.614)
Observations	12,796	12,796
Log Likelihood	-26,356.870	-26,358.000
θ	23.465*** (3.501) ($p = 0.000$)	23.422*** (3.489) ($p = 0.000$)
Akaike Inf. Crit.	52,729.730	52,728.000
Residual Deviance	19,237.720 (df = 12788)	19,237.400 (df = 12790)
Null Deviance (df = 12795)	21,051.280	21,048.310

Note:

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

5.2.2.1 Binomial Regression Model 2 Evaluation

Unlike the previous model, the chi-squared p-value for this model is close to 0. This means that the null hypothesis can be rejected.

residual.deviance	residual.degrees.of.freedom	chisq.p.value
19237	12790	4.356e-269

5.3 Multiple Linear Regression

5.3.1 Linear Regression with All Variables

The first linear regression we generate includes all variables from our data set. The intercept is at 3.139 cases and Density shows a large negative impact on cases sold but with its narrow range its difficult to tell how meaningful this variable is in cases sold. The STARS variable shows an expected impact on cases sold, the difference between 1 Star and 4 Stars is an added 3.36 cases in sales.

Table 23: Linear Model with all variables

	<i>Dependent variable:</i>
	TARGET
Constant	3.139*** (0.606)
FixedAcidity	−0.008*** (0.003)
VolatileAcidity	−0.125*** (0.020)
CitricAcid	0.013 (0.019)
I(ResidualSugar^1.25)	−0.001** (0.0002)
Chlorides	−0.234*** (0.052)
I(FreeSulfurDioxide^1.25)	0.0001*** (0.00003)
TotalSulfurDioxide	0.0002** (0.0001)
Density	−1.480** (0.602)
pH	−0.003 (0.024)
Sulphates	−0.046** (0.018)
Alcohol	0.016*** (0.004)
LabelAppeal	0.427*** (0.019)
STARS2	1.905*** (0.037)
STARS3	2.697*** (0.046)
STARS4	3.355*** (0.080)
Observations	7,256
R ²	0.501
Adjusted R ²	0.500
Residual Std. Error	1.356 (df = 7240)
F Statistic	485.267*** (df = 15; 7240) (p = 0.000)

Note: *p<0.1; **p<0.05; ***p<0.01

5.3.1.1 Linear Model Metrics with all Variables

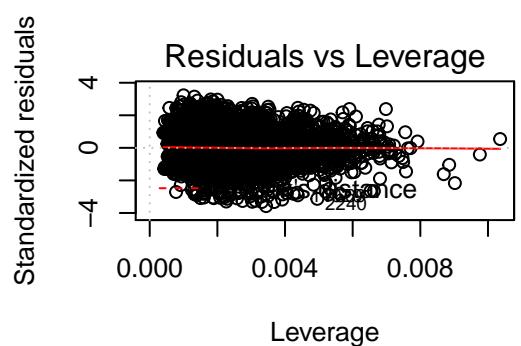
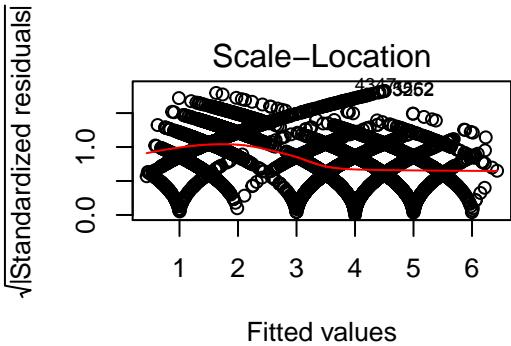
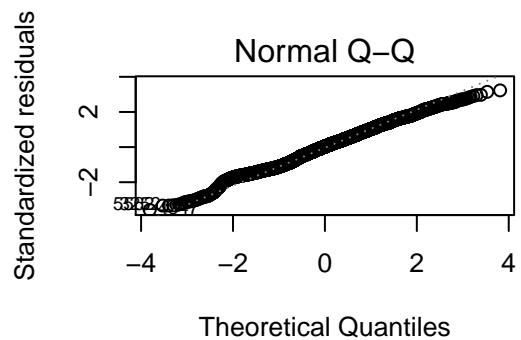
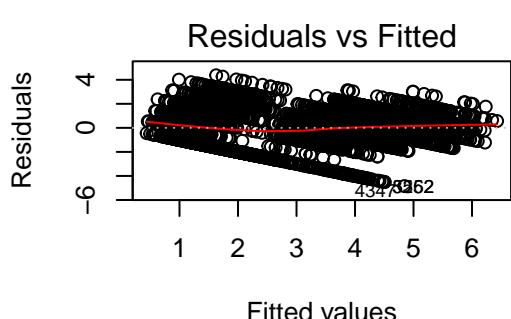
5.3.1.1.1 Multicollinearity

We square $GVIF^{(1/(2*Df))}$ ⁷ in order to use the VIF threshold of 5 for multicollinearity. Fortunately, we find that no variable exceeds our pre-established threshold of 5 for multicollinearity.

rn	GVIF	Df	$GVIF^{(1/(2*Df))}$	Adjusted_GVIF
FixedAcidity	1.003005	1	1.001501	1.003005
VolatileAcidity	1.004193	1	1.002095	1.004193
CitricAcid	1.003311	1	1.001654	1.003311
I(ResidualSugar^1.25)	1.001823	1	1.000911	1.001823
Chlorides	1.002043	1	1.001021	1.002043
I(FreeSulfurDioxide^1.25)	1.003812	1	1.001904	1.003812
TotalSulfurDioxide	1.005073	1	1.002533	1.005073
Density	1.002075	1	1.001037	1.002075
pH	1.002538	1	1.001268	1.002538
Sulphates	1.003861	1	1.001929	1.003861
Alcohol	1.006806	1	1.003397	1.006806
LabelAppeal	1.106042	1	1.051685	1.106042
STARS	1.121941	3	1.019362	1.039098

⁷"Which Variance Inflation Factor Should I Be Using: $GVIF$ or $GVIF^{1/(2*Df)}$?" R. N.p., n.d. Web. 13 Nov. 2016.

5.3.1.1.2 Diagnostic Plots



5.3.2 Linear Regression Selection using AIC

5.3.2.1 Variable Selection

Using the R package MASS we can utilize the `stepAIC` function with the parameter of `direction` set to both to select our best subset of variables for a new model.

The method effectively removed pH and CitricAcid which were both shown to be not significant in the previous linear model using all variables.

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + I(ResidualSugar^1.25) +
##       Chlorides + I(FreeSulfurDioxide^1.25) + TotalSulfurDioxide +
##       Density + pH + Sulphates + Alcohol + LabelAppeal + STARS
##
## Final Model:
## TARGET ~ FixedAcidity + VolatileAcidity + I(ResidualSugar^1.25) +
##       Chlorides + I(FreeSulfurDioxide^1.25) + TotalSulfurDioxide +
##       Density + Sulphates + Alcohol + LabelAppeal + STARS
##
##           Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                   7240    13311.93 4435.174
## 2 - pH             1  0.03552113    7241    13311.97 4433.193
## 3 - CitricAcid    1  0.94340380    7242    13312.91 4431.707
```

5.3.2.2 Model using Variable Selection

We see slight variation in our intercept and some variable coefficients which is expected with the reduced number of variables. However, we don't see any large changes, one benefit with the reduced variables is our model interpretability is improved and our F Statistic has increased with the reduced degrees of freedom.

Additionally, we see that the adjusted R^2 has not changed which is expected since we removed variables that were not considered significant.

Table 25: Linear Model with select variables

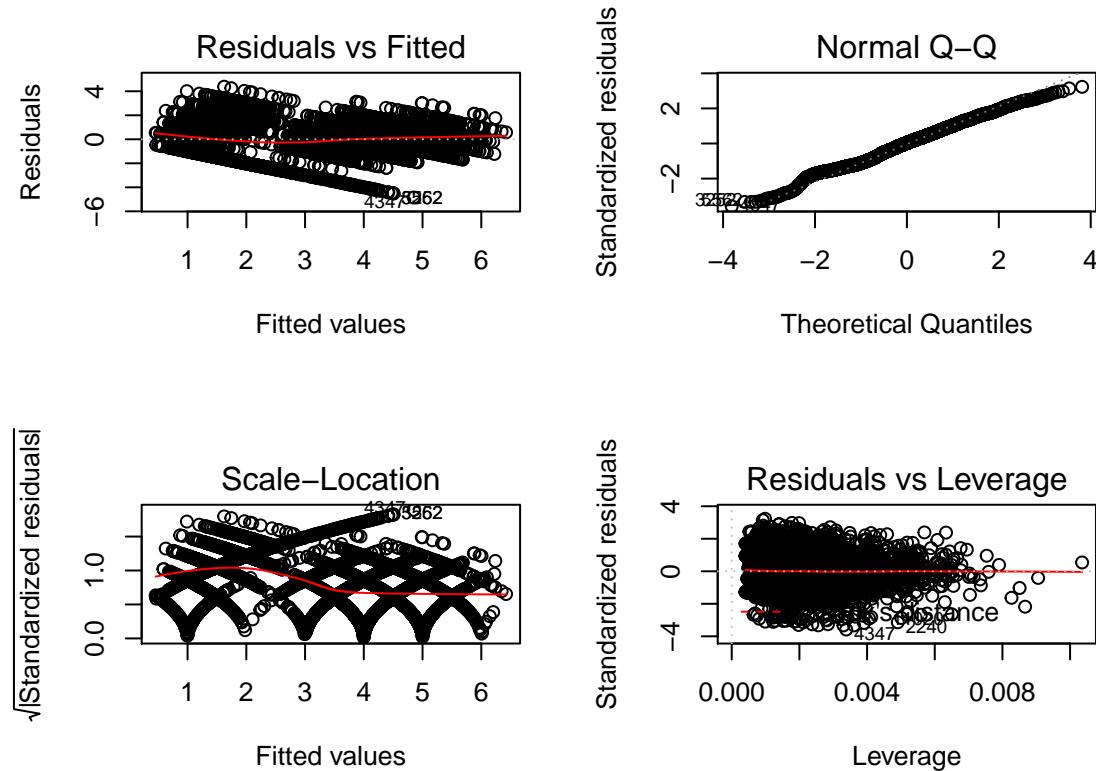
	<i>Dependent variable:</i>
	TARGET
Constant	3.134*** (0.601)
FixedAcidity	−0.008*** (0.003)
VolatileAcidity	−0.126*** (0.020)
I(ResidualSugar^1.25)	−0.001** (0.0002)
Chlorides	−0.235*** (0.052)
I(FreeSulfurDioxide^1.25)	0.0001*** (0.00003)
TotalSulfurDioxide	0.0002** (0.0001)
Density	−1.482** (0.602)
Sulphates	−0.046** (0.018)
Alcohol	0.016*** (0.004)
LabelAppeal	0.427*** (0.019)
STARS2	1.906*** (0.037)
STARS3	2.697*** (0.046)
STARS4	3.355*** (0.080)
Observations	7,256
R ²	0.501
Adjusted R ²	0.500
Residual Std. Error	1.356 (df = 7242)
F Statistic	559.996*** (df = 13; 7242) (p = 0.000)

Note:

*p<0.1; **p<0.05; ***p<0.01

5.3.2.3 Linear Model Metrics with select Variables

5.3.2.3.1 Diagnostic Plots



5.3.2.3.2 Multicollinearity

We square $\text{GVIF}^{(1/(2*\text{Df}))}$ in order to use the VIF threshold of 5 for multicollinearity. Fortunately, we find that no variable exceeds our pre-established threshold of 5 for multicollinearity.

rn	GVIF	Df	$\text{GVIF}^{(1/(2*\text{Df}))}$	Adjusted_GVIF
FixedAcidity	1.002826	1	1.001412	1.002826
VolatileAcidity	1.004095	1	1.002045	1.004095
I(ResidualSugar^1.25)	1.001601	1	1.000800	1.001601
Chlorides	1.001750	1	1.000875	1.001750
I(FreeSulfurDioxide^1.25)	1.003762	1	1.001879	1.003762
TotalSulfurDioxide	1.004725	1	1.002360	1.004725
Density	1.002057	1	1.001028	1.002057
Sulphates	1.003403	1	1.001700	1.003403
Alcohol	1.006586	1	1.003288	1.006586
LabelAppeal	1.105763	1	1.051553	1.105763
STARS	1.118645	3	1.018862	1.038080

6 Selected Model

In order to select our most appropriate model we will use the AIC as our selection criteria. This method was outlined by Snipes & Taylor for their similar research in selecting models from wine ratings and prices.⁸ As per Snipes & Taylor - "The best model is then the model with the lowest AICc (or AIC) score".

model	AIC
Poisson Regression 1	47326.78
Poisson Regression 2	52773.78
Binomial Regression 1	54290.7
Binomial Regression 2	52728
Linear Regression 1	25028.81
Linear Regression 2	25025.34

⁸Snipes, & Taylor. (2014). Model selection and Akaike Information Criteria: An example from wine ratings and prices. *Wine Economics and Policy*, 3(1), 3-9.

7 Appendix A

7.1 Session Info

- R version 3.3.2 (2016-10-31), x86_64-w64-mingw32
- Locale: LC_COLLATE=English_United States.1252, LC_CTYPE=English_United States.1252, LC_MONETARY=English_United States.1252, LC_NUMERIC=C, LC_TIME=English_United States.1252
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, utils
- Other packages: abc 2.1, abc.data 1.0, AER 1.2-4, bibtex 0.4.0, boot 1.3-18, car 2.1-4, corrplot 0.77, DAAG 1.22, data.table 1.10.0, doParallel 1.0.10, dplyr 0.5.0, e1071 1.6-7, foreach 1.4.3, forecast 7.3, Formula 1.2-1, ggplot2 2.2.0, glmulti 1.0.7, highlight 0.4.7, Hmisc 4.0-0, iterators 1.0.8, iterators 0.1-3, knitr 1.15.1, lars 1.2, lattice 0.20-34, leaps 2.9, lmtest 0.9-34, locfit 1.5-9.1, magrittr 1.5, MASS 7.3-45, matrixStats 0.51.0, missForest 1.4, nnet 7.3-12, pacman 0.4.1, pander 0.6.0, pracma 1.9.5, purrr 0.2.2, quantreg 5.29, randomForest 4.6-12, readr 1.0.0, rJava 0.9-8, sandwich 2.3-4, scales 0.4.1, SparseM 1.74, stargazer 5.2, stringr 1.1.0, survival 2.40-1, tibble 1.2, tidyverse 1.0.0, timeDate 3012.100, xlsx 0.5.7, xlsxjars 0.6.1, xtable 1.8-2, zoo 1.7-13
- Loaded via a namespace (and not attached): acepack 1.4.1, assertthat 0.1, backports 1.0.4, bitops 1.0-6, class 7.3-14, cluster 2.0.5, codetools 0.2-15, colorspace 1.3-1, DBI 0.5-1, digest 0.6.10, evaluate 0.10, foreign 0.8-67, fracdiff 1.4-2, grid 3.3.2, gridExtra 2.2.1, gtable 0.2.0, htmlTable 1.7, htmltools 0.3.5, httr 1.2.1, latticeExtra 0.6-28, lazyeval 0.2.0, lme4 1.1-12, lubridate 1.6.0, Matrix 1.2-7.1, MatrixModels 0.4-1, mgcv 1.8-16, minqa 1.2.4, munsell 0.4.3, nlme 3.1-128, nloptr 1.0.4, pbkrtest 0.4-6, plyr 1.8.4, quadprog 1.5-5, R6 2.2.0, RColorBrewer 1.1-2, Rcpp 0.12.8, RCurl 1.95-4.8, RefManager 0.13.1, RJSONIO 1.3-0, rmarkdown 1.2, rpart 4.1-10, rprojroot 1.1, splines 3.3.2, stringi 1.1.2, tools 3.3.2, tseries 0.10-35, XML 3.98-1.5, yaml 2.1.14

7.2 Data Dictionary

Variable Code	Definition
INDEX	Identification Variable (do not use)
TARGET	Number of Cases Purchased
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average
Alcohol	Alcohol Content
Chlorides	Chloride content of wine
CitricAcid	Citric Acid Content
Density	Density of Wine
FixedAcidity	Fixed Acidity of Wine
FreeSulfurDioxide	Sulfur Dioxide content of wine
LabelAppeal	Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customers don't like the design.
ResidualSugar	Residual Sugar of wine
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor
Sulphates	Sulfate content of wine
TotalSulfurDioxide	Total Sulfur Dioxide of Wine
VolatileAcidity	Volatile Acid content of wine
pH	pH of wine

7.3 R source code

Please see Homework 5.rmd on GitHub for source code.

<https://github.com/ChristopheHunt/DATA-621-Group-1/blob/master/Homework%205/Homework%205.Rmd>