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

4 Data Transformation **4**
4.1 Outliers Treatment 4
4.2 BoxCox Transformations 6

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, adjusting and accomodating consumer prefence is critical to a competitive advantage. By understanding the cases sold

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 averag**
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.
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

Variable	$\mathbf{n}$	Min	$\mathbf{q_1}$	$\widetilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\mathbf{q_3}$	Max	$\mathbf{s}$	IQR	#NA
TARGET	12795	0.0	2.0	3.0	3.0	4.0	8.0	1.9	2.0	0
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
LabelAppeal	12795	-2.0	-1.0	0.0	0.0	1.0	2.0	0.9	2.0	0
STARS	9436	1.0	1.0	2.0	2.0	3.0	4.0	0.9	2.0	3359

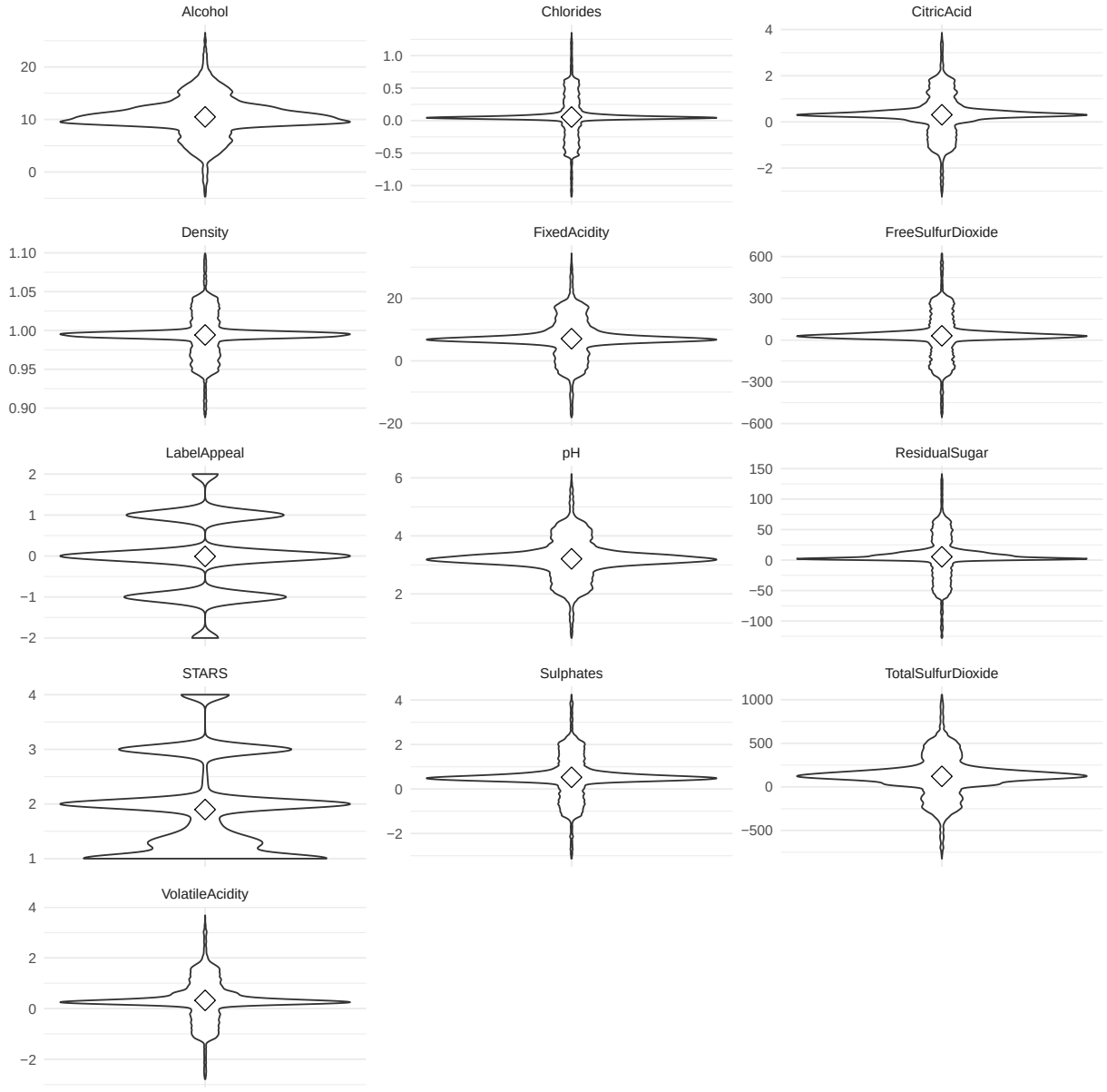
Table 2: Wine Training Data Set Summary Statistics

4 Data Transformation

4.1 Outliers Treatment

4.1.1 Violin Plots of Variables for Winsorizing

Violin Plots provide a visualization of the density plots and are capable of showing bi-modal or other non-normal characteristics of data.[2] Using the violin plots, we are can conclude that the variables to be winsorized are Free Sulfur Dioxide, Residual Sugar, and Total Sulfur Dioxide.



[2] Ridgway, Gerard (2015): A simple comparison of box plots and violin plots. figshare. Retrieved: 01 33, Nov 21, 2016 (GMT)

4.1.2 Winsorizing

We chose winsorizing as the method to address outliers. Instead of trimming values, winsorizing uses the interquantile range to replace values that are above or below the interquantile 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 interquantile range. Using the factor 2.2 for winsorizing outliers is a method developed my Hoaglin and Iglewicz and published Journal of American Statistical Association in 1987[3].

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

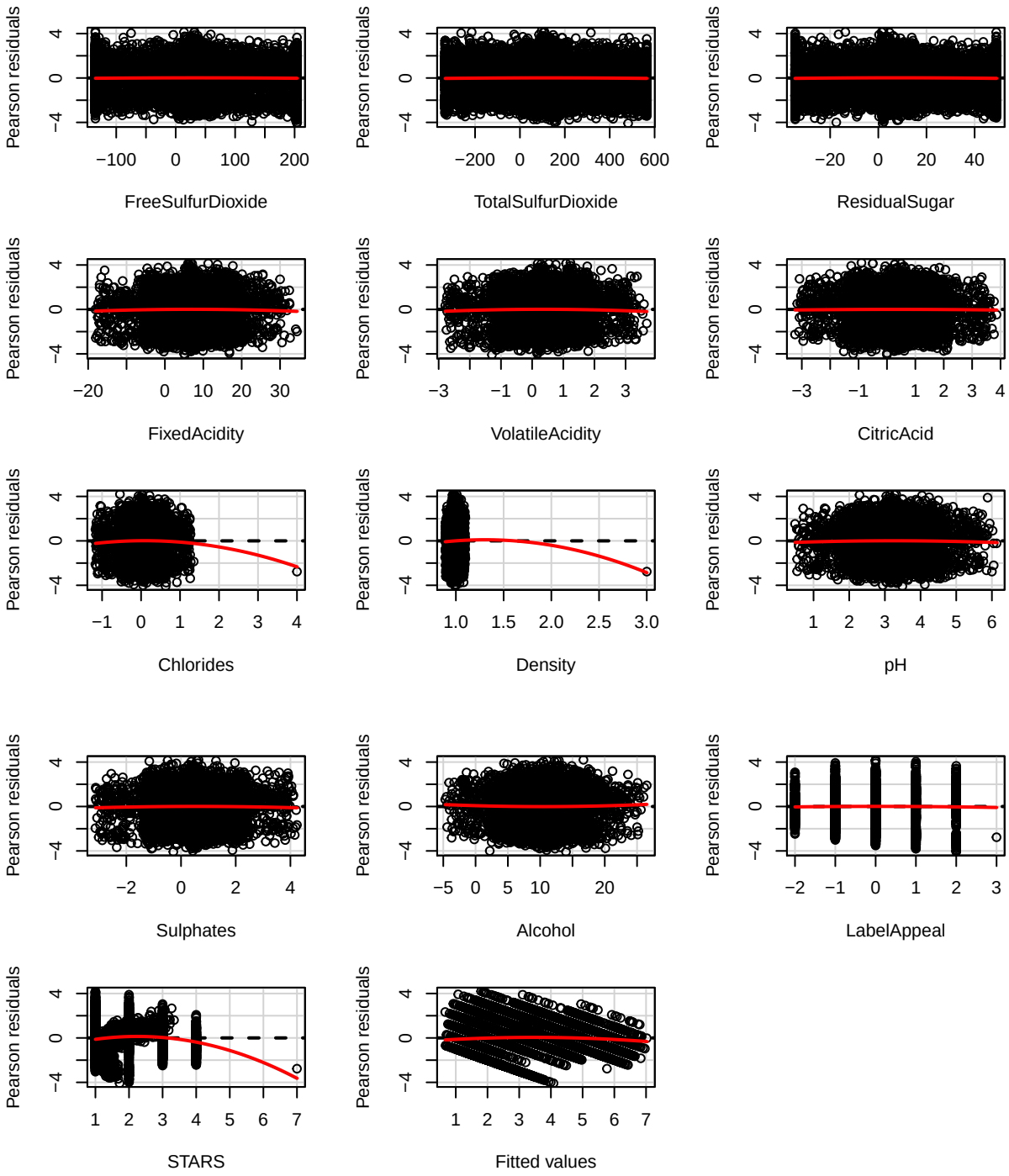
Table 3:

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
STARS	12,796	1.899	0.833	1.000	7.000

[3] Hoaglin, 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

The Box-Cox transformations were done on . These transformations were done based on the residual plots. In the residual plots, these three variables showed a great deal of non-constant variance because the plots were funnel-shaped.


Pearson residuals
FreeSulfurDioxide
TotalSulfurDioxide
ResidualSugar
FixedAcidity
VolatileAcidity
CitricAcid
Chlorides
Density
pH
Sulphates
Alcohol
LabelAppeal
STARS
Fitted values

Test stat Pr(>|t|) FreeSulfurDioxide -1.202 0.229 TotalSulfurDioxide -1.071 0.284 ResidualSugar -1.578 0.115 FixedAcidity -1.455 0.146 VolatileAcidity -1.663 0.096 CitricAcid -0.488 0.625 Chlorides -2.947 0.003 Density -2.356 0.018 pH -1.328 0.184 Sulphates -1.134 0.257 Alcohol 1.532 0.125 LabelAppeal -0.876 0.381 STARS -10.214 0.000 Tukey test -4.164 0.000