Homework 5

Group 1

Contents

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 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

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	$\widetilde{x}$	$\bar{x}$	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 Imputting 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

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
12795	0	470	1	7	7	-4	-1	5	7	10	16	18

```
lowest : -18.1 -18.0 -17.7 -17.5 -17.4, highest:  32.4  32.5  32.6  34.1  34.4
```

VolatileAcidity

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
12795	0	815	1	0.3	0.8	-1.0	-0.7	0.1	0.3	0.6	1.4	1.6

```
lowest : -2.790 -2.750 -2.745 -2.730 -2.720, highest:  3.500  3.550  3.565  3.590  3.680
```

CitricAcid

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
12795	0	602	1	0.3	0.9	-1.16	-0.84	0.03	0.31	0.58	1.43	1.79

```
lowest : -3.24 -3.16 -3.10 -3.08 -3.06, highest:  3.63  3.68  3.70  3.77  3.86
```

ResidualSugar

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
12795	0	2685	1	5	34	-52.0	-38.4	-0.5	4.1	15.0	48.9	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

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
12795	0	2285	1	0.05	0.3	-0.48	-0.36	-0.01	0.05	0.13	0.47	0.59

```
lowest : -1.171 -1.170 -1.158 -1.156 -1.155, highest:  1.260  1.261  1.270  1.275  1.351
```

FreeSulfurDioxide

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
12795	0	1626	1	31	150	-220	-165	3	30	66	223	281

```
lowest : -555 -546 -536 -535 -532, highest:  613  617  618  622  623
```

TotalSulfurDioxide

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
12795	0	2039	1	121	238	-266	-175	33	123	200	412	507

```
lowest : -823 -816 -793 -781 -779, highest: 1032 1041 1048 1054 1057
```

Density

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
12795	0	5933	1	1	0.03	0.9	1.0	1.0	1.0	1.0	1.0	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

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
12795	0	863	1	3	0.7	2	2	3	3	3	4	4

```
lowest : 0.48 0.53 0.54 0.58 0.59, highest: 5.91 5.94 6.02 6.05 6.13
```

[2] Stekhoven, 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	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

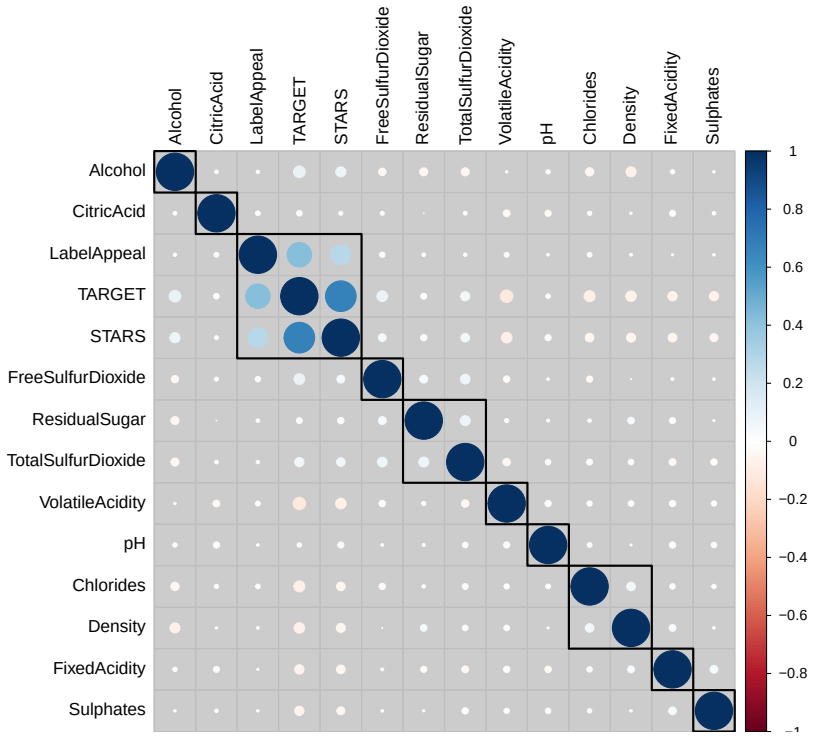
```
1 (5305, 0.415), 2 (4569, 0.357), 3 (2309, 0.180), 4 (612, 0.048)
```

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.

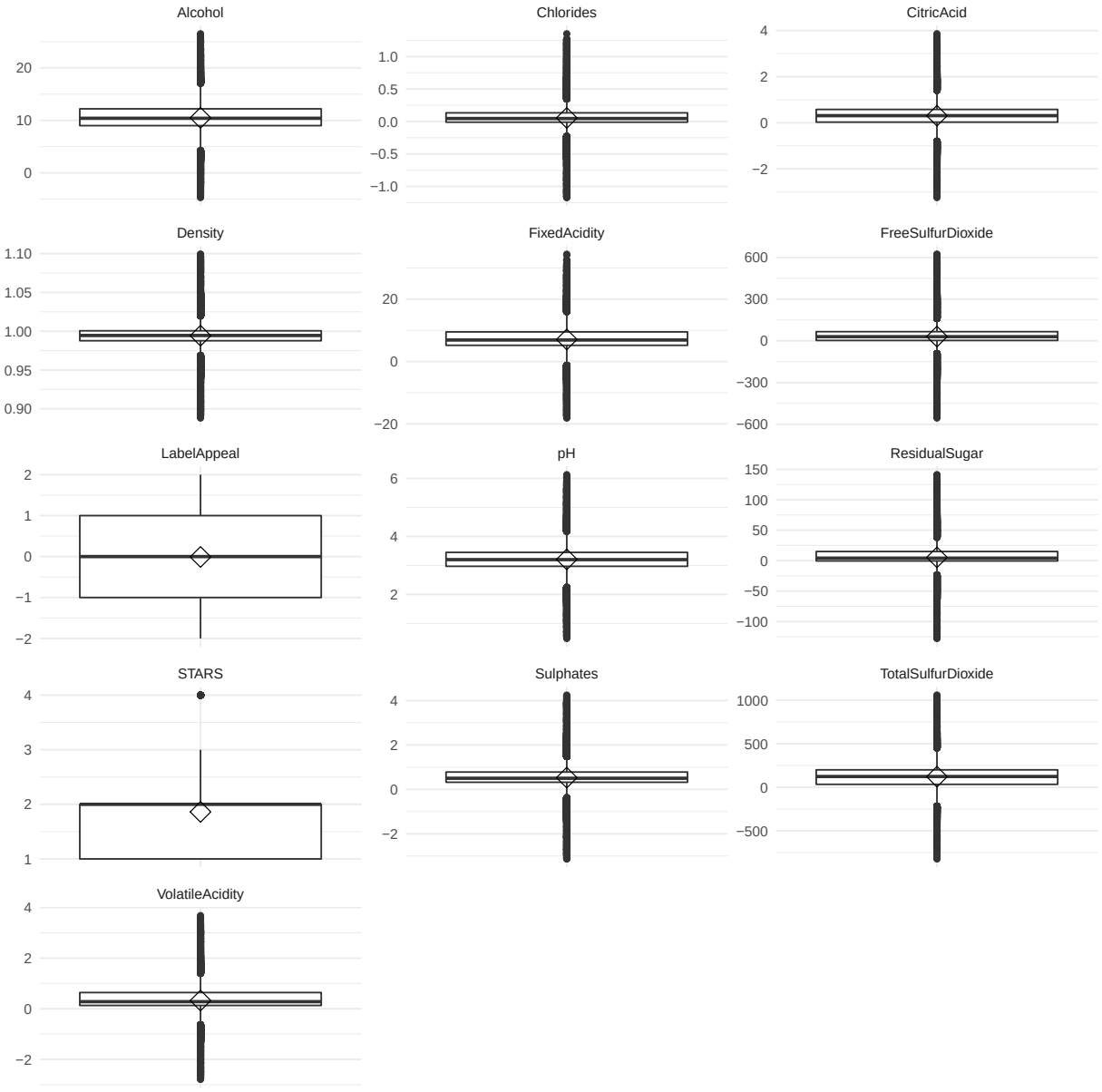


4 Data Transformation

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 are 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 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[4].

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

```
## Warning in `[<-.factor`(`*tmp*`, ri, value = structure(c(2L, 3L, 3L, 1L, :
## invalid factor level, NA generated
```

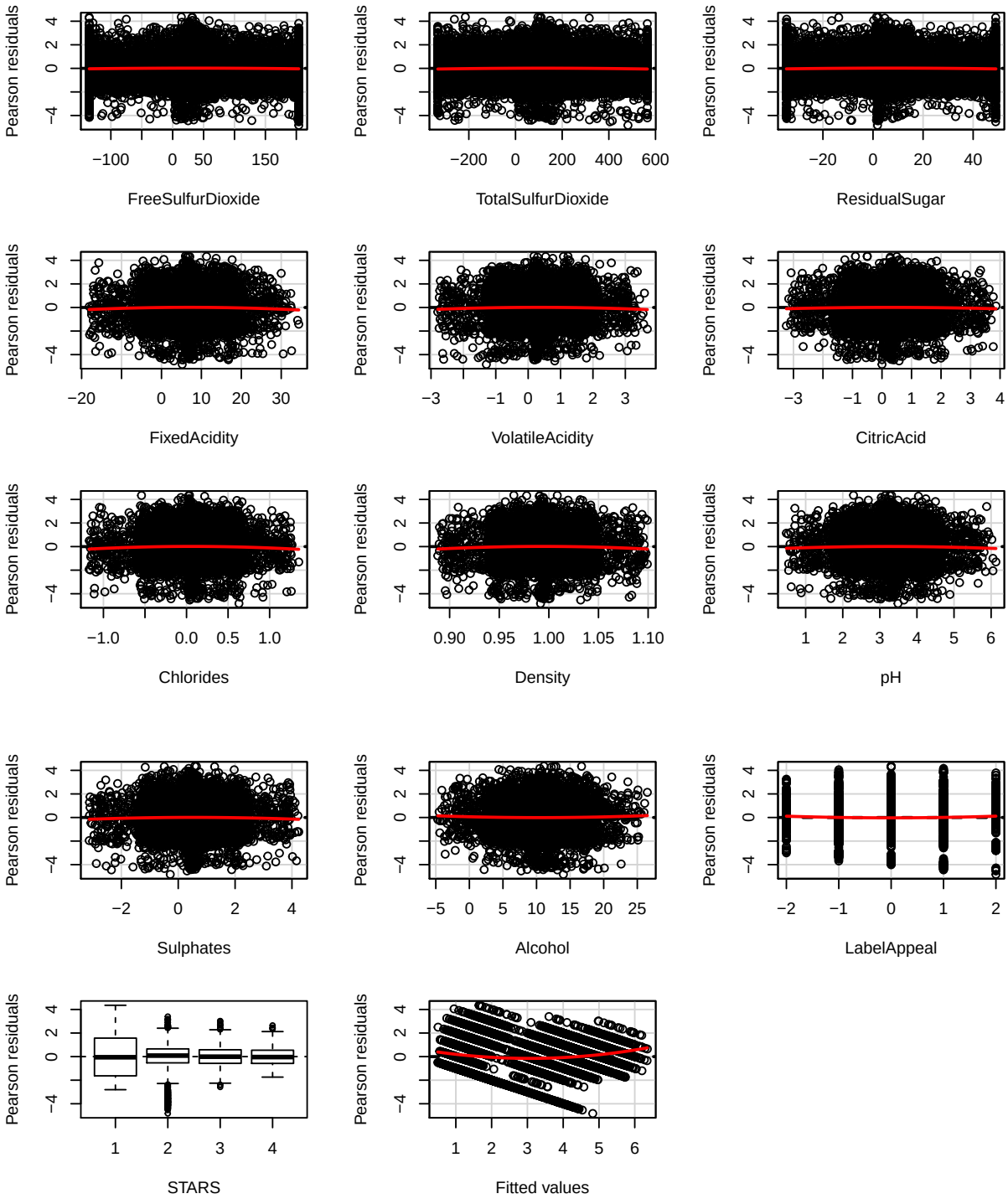
Table 4:

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

[4]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

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.


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

```
##                     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 Determing 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	$FreeSulfurDioxide^{1.25}$

[5] Osborne, 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 Multiple Linear Regression

5.1.1 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 6: Linear Model with all variables

	Dependent variable:
	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

5.1.1.1 Linear Model Metrics with all Variables

5.1.1.1.1 Multicollinearity

We square `GVIF^(1/(2*Df))` [6] 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

5.1.1.1.2 Diagnositic Plots

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

5.1.2 Linear Regression Selection using AIC

5.1.2.1 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 significiant 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
```

5.1.2.2 Model using Variable Selection

We see slight variation in our intercept and some variable coefficient 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 8: Linear Model with select variables

	Dependent variable:
	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

5.1.2.3 Linear Model Metrics with select Variables

5.1.2.3.1 Diagnositic Plots

5.1.2.3.2 Multicollinearity

We square `GVIF^(1/(2*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	GVIF^(1/(2*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

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, stats, utils
- Loaded via a namespace (and not attached): backports 1.0.4, digest 0.6.10, evaluate 0.10, htmltools 0.3.5, knitr 1.15.1, magrittr 1.5, Rcpp 0.12.8, rmarkdown 1.2, rprojroot 1.1, stringi 1.1.2, stringr 1.1.0, tools 3.3.2, 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 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

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