Data 621 Homework 5: Wine

Tommy Jenkins, Violeta Stoyanova, Todd Weigel, Peter Kowalchuk, Eleanor R-Secoquian, Anthony Pagan

12/05/2019

OVERVIEW

In this homework assignment, we will explore, analyze and model a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

Objective:

Our objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

DATA EXPLORATION

Data Summary

```
##      INDEX           TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##
##    CitricAcid      ResidualSugar        Chlorides       FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00
##  Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00
##                    NA's   :616        NA's   :638       NA's   :647
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
##  Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##  NA's   :682                         NA's   :395     NA's   :1210
```

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET	Number of Cases Purchased	None
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 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	

Figure 1:

```
##     Alcohol       LabelAppeal          AcidIndex         STARS
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##  NA's   :653                                          NA's   :3359
```

vars

n

mean

sd

median

trimmed

mad

min

max

range

skew

kurtosis

se

na_count

INDEX

1

12795

8069.9803048

4656.9051071

8110.00000

8071.0294031

5977.8432000

1.00000

16129.00000

1.6128e+04

-0.0032496

-1.2005027

41.1696565

0

TARGET

2

12795

3.0290739

1.9263682

3.00000

3.0538244

1.4826000

0.00000

8.00000

8.0000e+00

-0.3263010

-0.8772457

0.0170302

0

FixedAcidity

3

12795

7.0757171

6.3176435

6.90000

7.0736739

3.2617200

-18.10000

34.40000

5.2500e+01

-0.0225860

1.6749987

0.0558515

0

VolatileAcidity

4

12795

0.3241039

0.7840142

0.28000

0.3243890

0.4299540

-2.79000

3.68000

6.4700e+00

0.0203800

1.8322106

0.0069311

0

CitricAcid

5

12795

0.3084127

0.8620798

0.31000

0.3102520

0.4151280

-3.24000

3.86000

7.1000e+00

-0.0503070

1.8379401

0.0076213

0

ResidualSugar

6

12179

5.4187331

33.7493790

3.90000

5.5800410

15.7155600

-127.80000

141.15000

2.6895e+02

-0.0531229

1.8846917

0.3058158

616

Chlorides

7

12157

0.0548225

0.3184673

0.04600

0.0540159

0.1349166

-1.17100

1.35100

2.5220e+00

0.0304272

1.7886044

0.0028884

638

FreeSulfurDioxide

8

12148

30.8455713

148.7145577

30.00000

30.9334877

56.3388000

-555.00000

623.00000

1.1780e+03

0.0063930

1.8364966

1.3492769

647

TotalSulfurDioxide

9

12113

120.7142326

231.9132105

123.00000

120.8895367

134.9166000

-823.00000

1057.00000

1.8800e+03

-0.0071794

1.6746665

2.1071703

682

Density

10

12795

0.9942027

0.0265376

0.99449

0.9942130

0.0093552

0.88809

1.09924

2.1115e-01

-0.0186938

1.8999592

0.0002346

0

pH

11

12400

3.2076282

0.6796871

3.20000

3.2055706

0.3854760

0.48000

6.13000

5.6500e+00

0.0442880

1.6462681

0.0061038

395

Sulphates

12

11585

0.5271118

0.9321293

0.50000

0.5271453

0.4447800

-3.13000

4.24000

7.3700e+00

0.0059119

1.7525655

0.0086602

1210

Alcohol

13

12142

10.4892363

3.7278190

10.40000

10.5018255

2.3721600

-4.70000

26.50000

3.1200e+01

-0.0307158

1.5394949

0.0338306

653

LabelAppeal

14

12795

-0.0090660

0.8910892

0.00000

-0.0099639

1.4826000

-2.00000

2.00000

4.0000e+00

0.0084295

-0.2622916

0.0078777

0

AcidIndex

15

12795

7.7727237

1.3239264

8.00000

7.6431572

1.4826000

4.00000

17.00000

1.3000e+01

1.6484959

5.1900925

0.0117043

0

STARS

16

9436

2.0417550

0.9025400

2.00000

1.9711258

1.4826000

1.00000

4.00000

3.0000e+00

0.4472353

-0.6925343

0.0092912

3359

The dataset consists of two data files: training and evaluation. The training dataset contains 16 columns, and the evaluation dataset also contains 16 columns.

Missing Data

An important aspect of any dataset is to determine how much, if any, data is missing. We look at all the variables to see which if any have missing data. We look at the basic descriptive statistics as well as the missing data and percentages.

We start by looking at the dataset as a whole and determine how many complete rows, that is rows with data for all predictors we have.

```
##    Mode   FALSE    TRUE
## logical    6359    6436
```

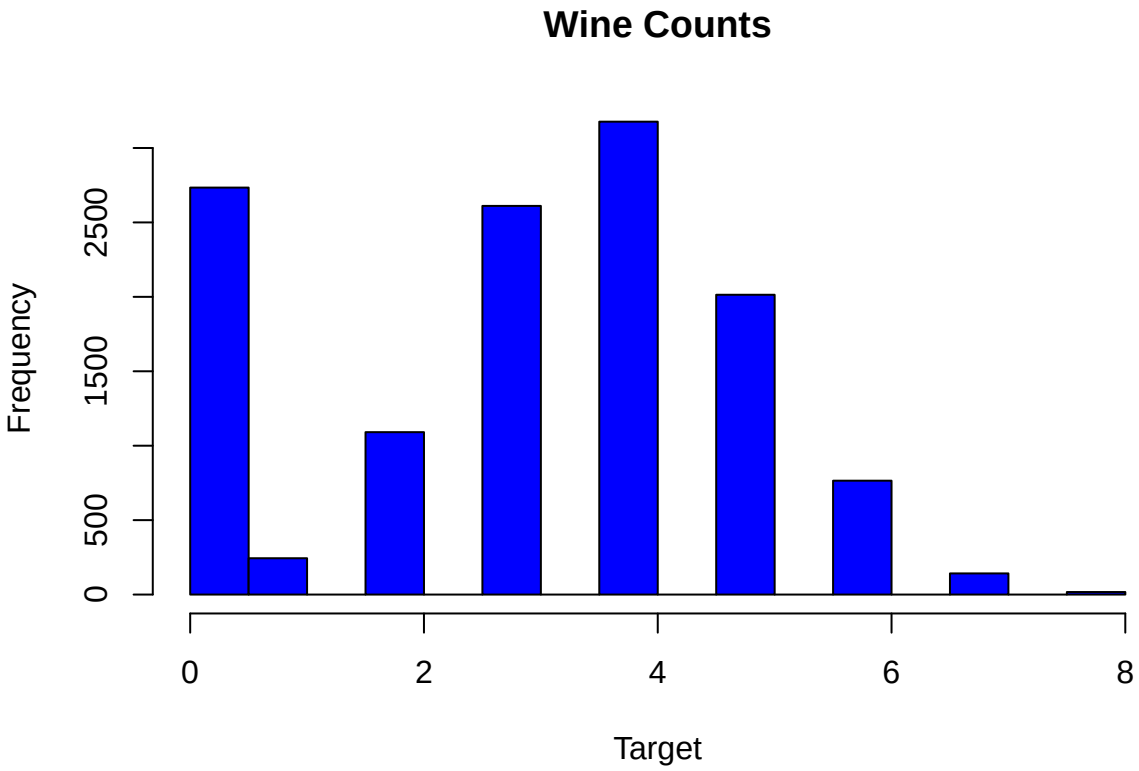
With these results, if we remove all rows with incomplete rows, there will be a total of 6436 rows out of 12795 .If we eliminate all non-complete rows and keep only rows with data for all the predictors in the dataset, our new dataset will results in 50% of the total dataset. We create a subset of data with complete cases only to use later in our analysis.

```
## Observations: 6,436
## Variables: 16
## $ INDEX              <int> 4, 5, 13, 14, 16, 23, 24, 25, 26, 27, 28, 3...
## $ TARGET             <int> 5, 3, 6, 0, 3, 4, 5, 4, 3, 2, 3, 4, 4, 3, 4...
## $ FixedAcidity       <dbl> 7.1, 5.7, 5.5, -17.2, 6.0, -1.3, 10.0, 6.8,...
```

```
## $ VolatileAcidity    <dbl> 2.640, 0.385, -0.220, 0.520, 0.330, 0.220, ...
## $ CitricAcid         <dbl> -0.88, 0.04, 0.39, 0.15, -1.06, 2.95, 0.27,...
## $ ResidualSugar      <dbl> 14.80, 18.80, 1.80, -33.80, 3.00, -53.00, 1...
## $ Chlorides          <dbl> 0.037, -0.425, -0.277, -0.022, 0.518, 0.541...
## $ FreeSulfurDioxide  <dbl> 214, 22, 62, 551, 5, -85, -188, -88, 87, 15...
## $ TotalSulfurDioxide <dbl> 142, 115, 180, 65, 378, -266, 229, 508, -28...
## $ Density            <dbl> 0.99518, 0.99640, 0.94724, 0.99340, 0.96643...
## $ pH                 <dbl> 3.12, 2.24, 3.09, 4.31, 3.55, 3.61, 3.14, 3...
## $ Sulphates          <dbl> 0.48, 1.83, 0.75, 0.56, -0.86, 0.82, 0.88, ...
## $ Alcohol            <dbl> 22.0, 6.2, 12.6, 13.1, 3.9, 10.0, 11.0, 18....
## $ LabelAppeal        <int> -1, -1, 0, 1, 1, 0, 1, -1, -1, -1, 0, 0, 1,...
## $ AcidIndex          <int> 8, 6, 8, 5, 7, 8, 11, 8, 6, 7, 8, 7, 7, 8, ...
## $ STARS              <int> 3, 1, 4, 1, 2, 3, 2, 2, 1, 1, 1, 2, 2, 1, 3...
```

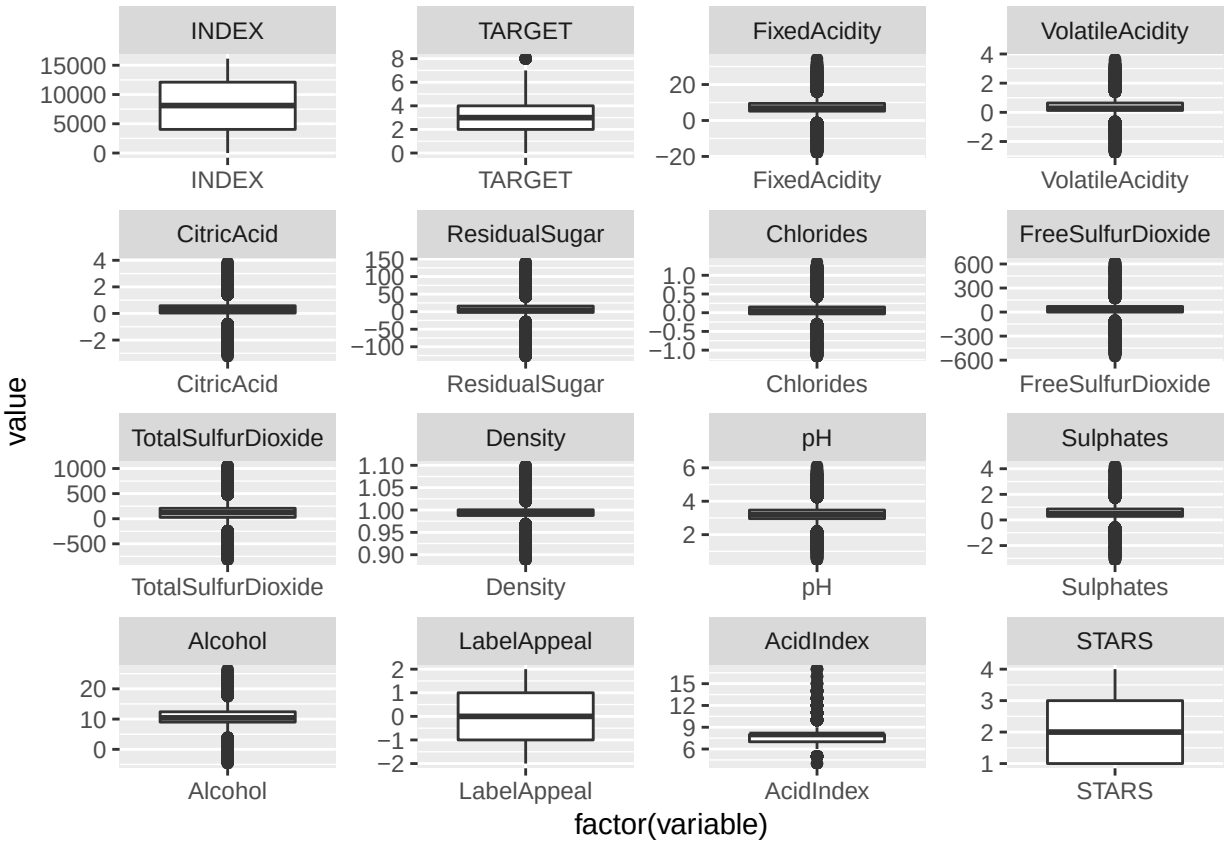
Visualization

Histogram

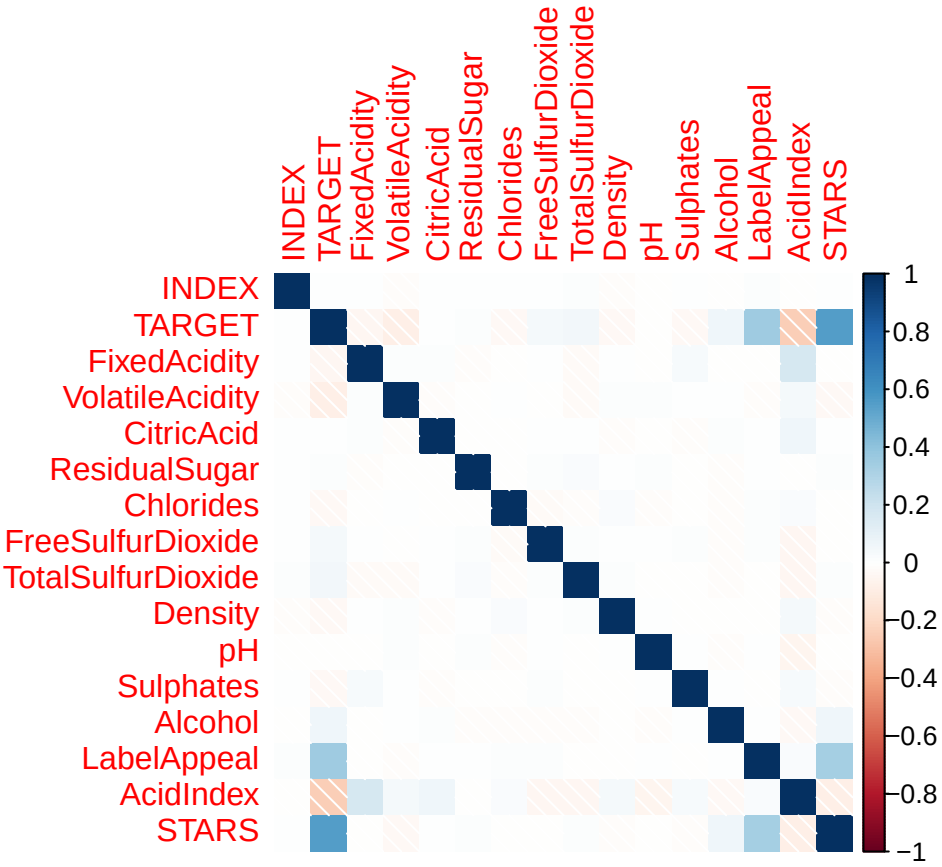


Boxplot

```
## No id variables; using all as measure variables
```



Correlation



BUILD MODEL

Model 1: Poisson Regression (all predictors)

For the first model, we used the poisson regression and all of the predictors.

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = WineTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2107  -0.2736   0.0628   0.3748   1.6983
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.578e+00  2.509e-01   6.290 3.18e-10 ***
## INDEX                1.610e-06  1.407e-06   1.144  0.25266
## FixedAcidity         3.348e-04  1.053e-03   0.318  0.75050
## VolatileAcidity     -2.563e-02  8.354e-03  -3.067  0.00216 **
## CitricAcid          -9.521e-04  7.578e-03  -0.126  0.90002
## ResidualSugar       -6.579e-05  1.941e-04  -0.339  0.73462
## Chlorides           -3.016e-02  2.056e-02  -1.467  0.14239
## FreeSulfurDioxide    6.706e-05  4.404e-05   1.523  0.12778
## TotalSulfurDioxide   2.071e-05  2.855e-05   0.725  0.46829
## Density             -3.712e-01  2.462e-01  -1.508  0.13160
```

```
## pH                -4.402e-03  9.601e-03  -0.459  0.64657
## Sulphates         -5.102e-03  7.052e-03  -0.724  0.46937
## Alcohol            3.932e-03  1.771e-03   2.221  0.02638 *
## LabelAppeal        1.769e-01  7.958e-03  22.223  < 2e-16 ***
## AcidIndex         -4.872e-02  5.903e-03  -8.254  < 2e-16 ***
## STARS              1.873e-01  7.490e-03  25.010  < 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: 5844.1  on 6435  degrees of freedom
## Residual deviance: 4007.8  on 6420  degrees of freedom
##   (6359 observations deleted due to missingness)
## AIC: 23172
##
## Number of Fisher Scoring iterations: 5
```

Model 2: Poisson Regression (reduced predictors)

For the second model, based on model 1, we reduced the number of predictors.

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS, family = poisson,
##     data = WineTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2033  -0.2749   0.0597   0.3745   1.6718
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.571e+00  2.446e-01   6.422 1.34e-10 ***
## VolatileAcidity     -2.395e-02  8.132e-03  -2.945  0.00323 **
## CitricAcid           5.444e-04  7.394e-03   0.074  0.94131
## Chlorides           -2.616e-02  2.003e-02  -1.306  0.19147
## FreeSulfurDioxide    7.887e-05  4.287e-05   1.840  0.06583 .
## TotalSulfurDioxide   2.289e-05  2.787e-05   0.821  0.41144
## Density             -3.405e-01  2.402e-01  -1.417  0.15635
## pH                  -6.192e-03  9.399e-03  -0.659  0.51000
## Sulphates           -5.530e-03  6.871e-03  -0.805  0.42088
## Alcohol              3.666e-03  1.729e-03   2.121  0.03395 *
## LabelAppeal          1.756e-01  7.742e-03  22.682  < 2e-16 ***
## AcidIndex           -4.906e-02  5.711e-03  -8.591  < 2e-16 ***
## STARS                1.886e-01  7.283e-03  25.888  < 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: 6145.7  on 6746  degrees of freedom
## Residual deviance: 4212.4  on 6734  degrees of freedom
```

```
##   (6048 observations deleted due to missingness)
## AIC: 24301
##
## Number of Fisher Scoring iterations: 5
```

Model 3: Gaussian Regression (significant predictors)

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide +
##     TotalSulfurDioxide + Chlorides + Density + pH + Sulphates +
##     LabelAppeal + AcidIndex + STARS, family = gaussian, data = WineTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0336  -0.5234   0.1254   0.7277   3.2704
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.713e+00  5.272e-01   8.941  < 2e-16 ***
## VolatileAcidity     -8.708e-02  1.750e-02  -4.977 6.62e-07 ***
## FreeSulfurDioxide    2.859e-04  9.297e-05   3.075  0.00211 **
## TotalSulfurDioxide   6.881e-05  5.987e-05   1.149  0.25040
## Chlorides           -1.088e-01  4.318e-02  -2.518  0.01181 *
## Density             -1.285e+00  5.200e-01  -2.472  0.01347 *
## pH                  -5.139e-03  2.032e-02  -0.253  0.80032
## Sulphates           -2.166e-02  1.479e-02  -1.464  0.14311
## LabelAppeal          6.403e-01  1.659e-02  38.607  < 2e-16 ***
## AcidIndex           -1.628e-01  1.160e-02 -14.031  < 2e-16 ***
## STARS                7.342e-01  1.622e-02  45.271  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.335267)
##
##     Null deviance: 17036.4  on 7102  degrees of freedom
## Residual deviance:  9469.7  on 7092  degrees of freedom
##   (5692 observations deleted due to missingness)
## AIC: 22224
##
## Number of Fisher Scoring iterations: 2
```

Model 4: Negative Binomial Regression

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + TotalSulfurDioxide +
##     pH + Sulphates + LabelAppeal + AcidIndex + STARS, family = negative.binomial(1),
##     data = WineTrain)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.90876  -0.13272   0.03009   0.17214   0.81386
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.303e+00  3.707e-02  35.142  < 2e-16 ***
## VolatileAcidity     -2.811e-02  5.336e-03  -5.268 1.41e-07 ***
## TotalSulfurDioxide   2.294e-05  1.804e-05   1.272   0.2034
## pH                  -4.773e-03  6.120e-03  -0.780   0.4354
## Sulphates           -7.543e-03  4.477e-03  -1.685   0.0921 .
## LabelAppeal          1.866e-01  5.024e-03  37.140  < 2e-16 ***
## AcidIndex           -5.547e-02  3.579e-03 -15.498  < 2e-16 ***
## STARS                1.965e-01  4.855e-03  40.479  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1) family taken to be 0.1050533)
##
##     Null deviance: 2393.6  on 7878  degrees of freedom
## Residual deviance: 1892.0  on 7871  degrees of freedom
##   (4916 observations deleted due to missingness)
## AIC: 37792
##
## Number of Fisher Scoring iterations: 5
```

SELECT MODEL

Pick the best regression model

	Model 1	Model 2	Model 3	Model 4
AIC	23172.4390791202	24301.0965825002	22224.129938863	37791.8993926271
BIC	23280.75367925	24389.7156746284	22306.5492089709	37847.675042797

With 4 models computed, we select the model with the lowest combination of AIC and BIC. From the table, we can see the model to pick is model

APPENDIX

Code used in analysis

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE)
require(knitr)
library(ggplot2)
library(tidyr)
library(MASS)
library(psych)
library(kableExtra)
library(dplyr)
library(faraway)
library(gridExtra)
library(reshape2)
library(leaps)
library(caret)
library(naniar)
library(pander)
library(pROC)
library(corrplot)
```

```r
#WineTrain <- read.csv("wine-training-data.csv",na.strings="",header=TRUE)
WineTrain <- read.csv("https://raw.githubusercontent.com/pkowalchuk/CUNY621-HW5/master/wine-training-da
WineTrain1 <- WineTrain
WineEval <- read.csv("wine-evaluation-data.csv",na.strings="",header=TRUE)
summary(WineTrain)

wine1 <- describe(WineTrain)
wine1$na_count <- sapply(WineTrain, function(y) sum(length(which(is.na(y)))))

kable(wine1, "html", escape = F) %>%
  kable_styling("striped", full_width = T) %>%
  column_spec(1, bold = T) %>%
  scroll_box(width = "100%", height = "700px")

colsTrain<-ncol(WineTrain)
colsEval<-ncol(WineEval)
missingCol<-colnames(WineTrain)[!(colnames(WineTrain) %in% colnames(WineEval))]
cc<-summary(complete.cases(WineTrain))
cWineTrain<-subset(WineTrain, complete.cases(WineTrain))
cc
glimpse(cWineTrain)
WineTrain1$INDEX <- NULL
hist(WineTrain1$TARGET, col = "blue", xlab = " Target ", main = "Wine Counts")
ggplot(melt(WineTrain), aes(x=factor(variable), y=value)) + facet_wrap(~variable, scale="free") + geom_
corrplot(as.matrix(cor(WineTrain, use = "pairwise.complete")),method = "shade")
m1 <- glm(TARGET ~ ., family = poisson, data = WineTrain)
summary(m1)
m2 <- glm(TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide
                        + TotalSulfurDioxide + Density + pH + Sulphates + Alcohol + LabelAppeal
             + AcidIndex + STARS, family = poisson, data = WineTrain)
summary(m2)
m3 <- glm(TARGET ~ VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + Chlorides + Density + pH
summary(m3)
m4 <- glm(TARGET ~ VolatileAcidity + TotalSulfurDioxide +
     pH + Sulphates + LabelAppeal + AcidIndex + STARS, family = negative.binomial(1),
     data = WineTrain)
summary(m4)
m1AIC <- AIC(m1)
m1BIC <- BIC(m1)
m2AIC <- AIC(m2)
m2BIC <- BIC(m2)
m3AIC <- AIC(m3)
m3BIC <- BIC(m3)
m4AIC <- AIC(m4)
m4BIC <- BIC(m4)

AIC <- list(m1AIC, m2AIC, m3AIC, m4AIC)
BIC <- list(m1BIC, m2BIC, m3BIC, m4BIC)
kable(rbind(AIC, BIC), col.names = c("Model 1", "Model 2", "Model 3", "Model 4"))  %>%
  kable_styling(full_width = T)
```