DATA621-HW5-SmoothOperators

Rob Hodde, Matt Farris, Jeffrey Burmood, Bin Lin

5/11/2017

Problem Description

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.

The 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.

Data Exploration

Data Exploration

There are numerous NAs in certain variables, and variables with negative values. Variables with negative values have nearly normal distributions so it is possible some previous data adjustments have been made. The variable data with negative values in stable, normal distributions will be used as-is. Below is a summary of variables by type, followed by their basic statistical summaries:

VAR	TYPE
TARGET	integer
FixedAcidity	double
VolatileAcidity	double
CitricAcid	double
ResidualSugar	double
Chlorides	double
FreeSulfurDioxide	double
TotalSulfurDioxide	double
Density	double
pH	double
Sulphates	double
Alcohol	double
LabelAppeal	integer
AcidIndex	integer

VAR	TYPE
STARS	integer

TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides
Min. :0.000	Min. :-18.100	Min. :-2.7900	Min. :-3.2400	Min. :-127.800	Min. :-1.1710
1st Qu.:2.000	1st Qu.: 5.200	1st Qu.: 0.1300	1st Qu.: 0.0300	1st Qu.: -2.000	1st Qu.:-0.0310
Median :3.000	Median : 6.900	Median : 0.2800	Median : 0.3100	Median : 3.900	Median : 0.0460
Mean :3.029	Mean : 7.076	Mean : 0.3241	Mean : 0.3084	Mean : 5.419	Mean : 0.0548
3rd Qu.:4.000	3rd Qu.: 9.500	3rd Qu.: 0.6400	3rd Qu.: 0.5800	3rd Qu.: 15.900	3rd Qu.: 0.1530
Max. :8.000	Max. : 34.400	Max. : 3.6800	Max. : 3.8600	Max. : 141.150	Max. : 1.3510
NA	NA	NA	NA	NA's :616	NA's :638

FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol
Min. :-555.00	Min. :-823.0	Min. :0.8881	Min. :0.480	Min. :-3.1300	Min. :-4.70
1st Qu.: 0.00	1st Qu.: 27.0	1st Qu.:0.9877	1st Qu.:2.960	1st Qu.: 0.2800	1st Qu.: 9.00
Median : 30.00	Median : 123.0	Median :0.9945	Median :3.200	Median : 0.5000	Median :10.40
Mean : 30.85	Mean : 120.7	Mean :0.9942	Mean :3.208	Mean : 0.5271	Mean :10.49
3rd Qu.: 70.00	3rd Qu.: 208.0	3rd Qu.:1.0005	3rd Qu.:3.470	3rd Qu.: 0.8600	3rd Qu.:12.40
Max. : 623.00	Max. :1057.0	Max. :1.0992	Max. :6.130	Max. : 4.2400	Max. :26.50
NA's :647	NA's :682	NA	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	NA	NA's :3359

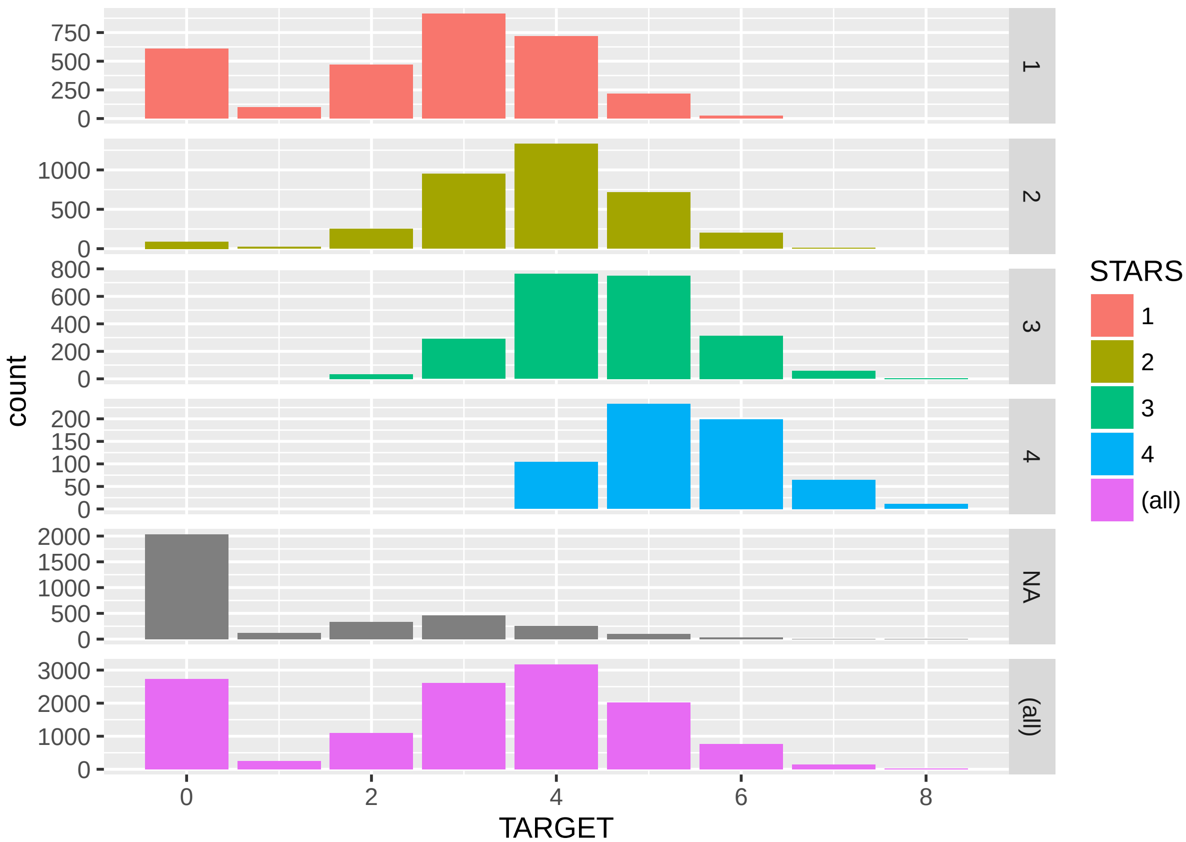
NEED VERBIAGE - DATA EXPLORATION: There are numerous NAs in certain variables, and variables with negative values. Variables with negative values have apparently normal distributions so it's possible some previous data adjustments have been made. The variable data with negative values in stable, normal distributions will be used as-is.

NEED TO EXPLORE MEAN AND VARIANCE OF OUTCOME VARIABLE. POISSON PROBABLITY MASS FUNCTION

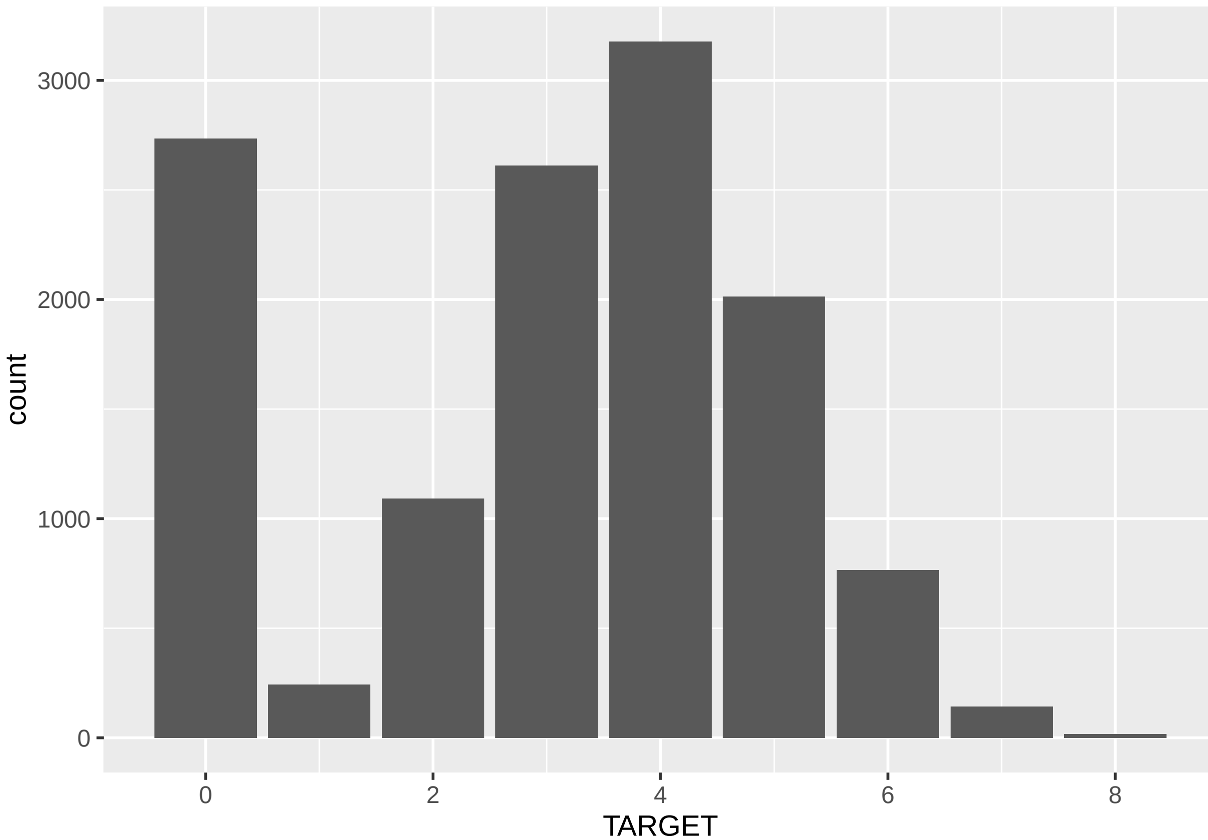
NEED TARGET COUNT HISTOGRAM, NOT ORGANIZED BY STARS.

Clean data by removing unnecessary columns, replacing NA's, and setting the unrated wines (no stars) to zero stars, so they can be analyzed.

```r
ggplot(wine, aes(TARGET, fill = STARS)) + geom_bar(stat = "count") + facet_grid(STARS ~
    ., margins = TRUE, scales = "free")
```



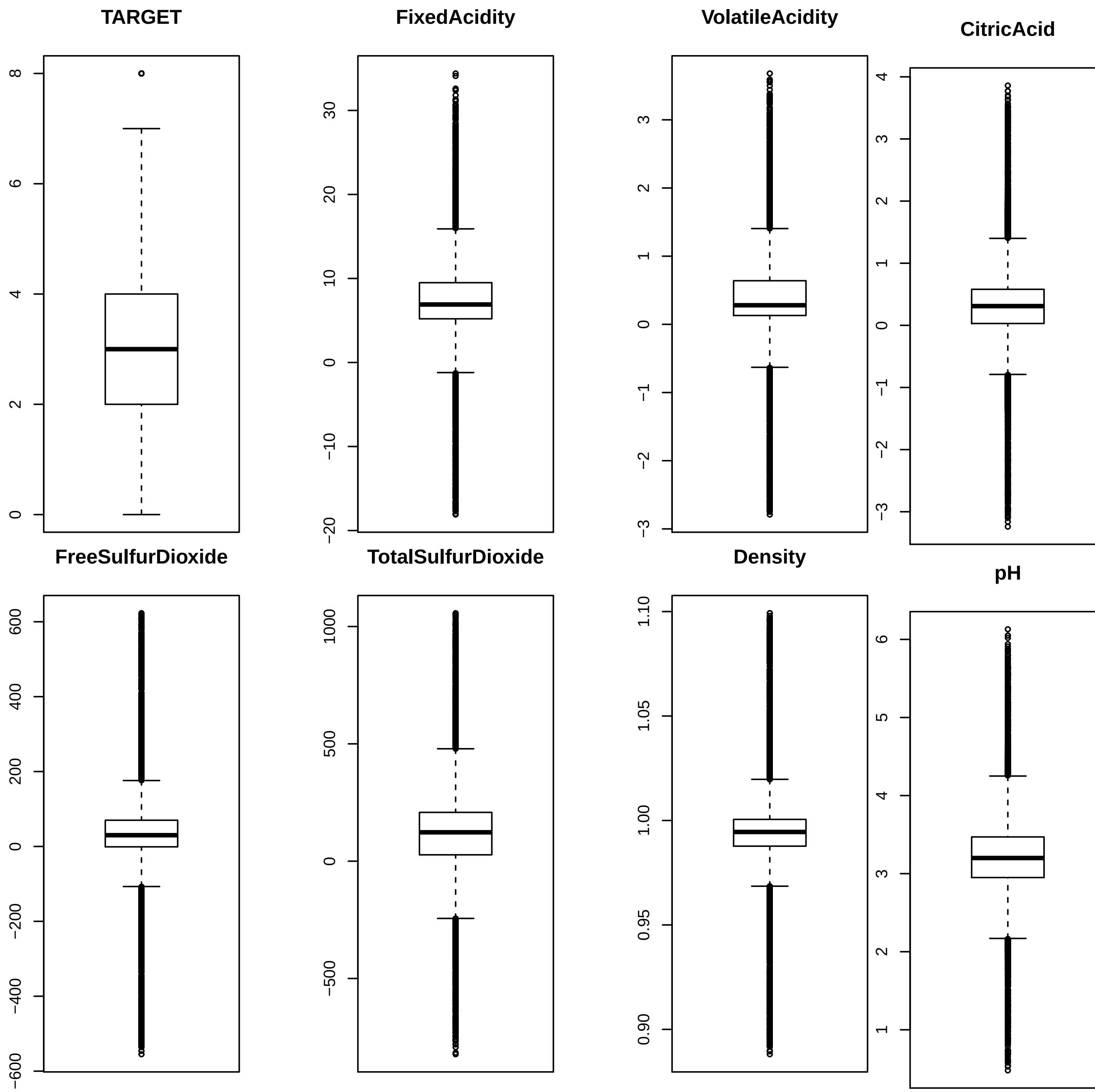
Lastly, we'll look at the whole distribution of counts for the TARGET variable.

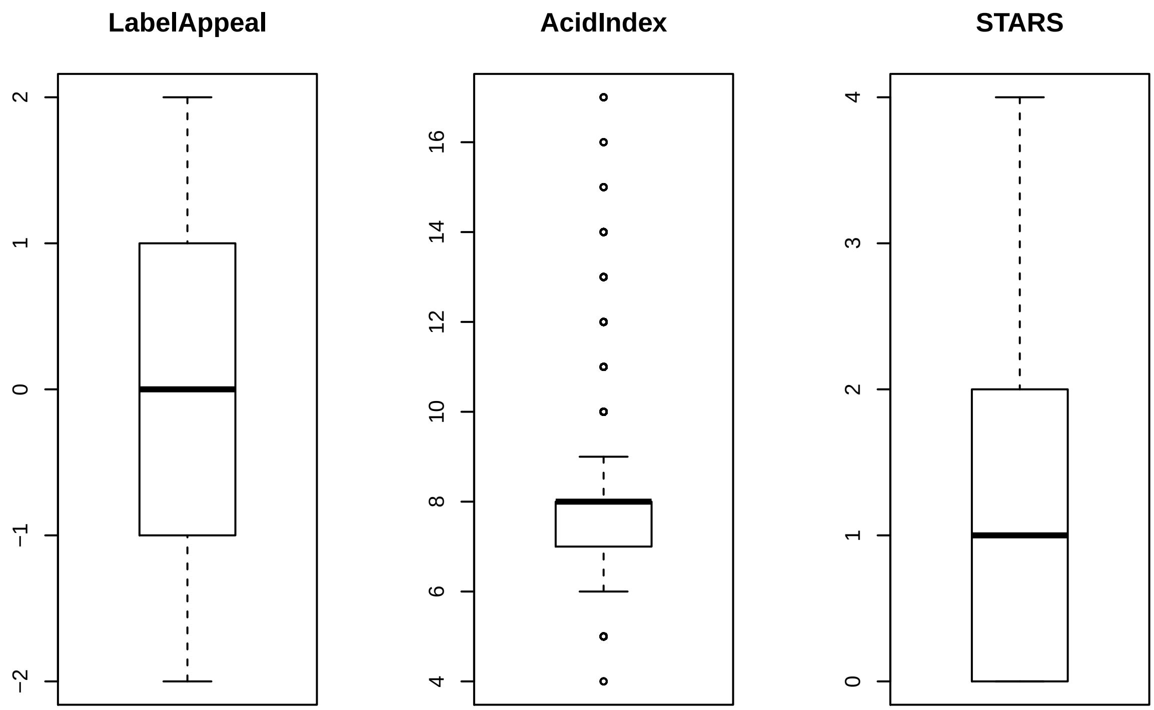


Data Preparation

Data Preparation

We will cleanse the data by removing the index column, using the MICE package to replace NA's with meaningful values, and setting the unrated wines (no stars) to zero stars, so they can be analyzed quantitatively.


TARGET
FixedAcidity
VolatileAcidity
CitricAcid
FreeSulfurDioxide
TotalSulfurDioxide
Density
pH



The final data preparation step is to split the training data into two portions, Train and Test. We will use 80% of the data for training the model, and 20% for evaluation.

Build Models

Build Models

By looking at these models we suspect there may be two forces at work. The first we will call Perception. The two Perception variables are Stars and Label Appeal. Based on the high coefficients and high significance, Perception seems to impact the outcome much more than anything else. The second force we will call Chemistry. All the other variables could belong to this group. The pattern we see here is that the best outcome (highest number of cases purchased) tends to occur when the Chemistry variables are close to the mean.

Linear Regression Models

NEED VERBIAGE - LINEAR MODELS

Regular Poisson Model

Next we will create a generalized linear model, Poisson family, that combines all the variables:

```
# create generalized linear model, poisson distribution.  this is for analyzing count data
pm <- glm(as.formula(paste(colnames(train)[1], "~", paste(colnames(train)[-1], collapse = "+"), sep = "
summary(pm)
```

```
##
## Call:
```

```
## glm(formula = as.formula(paste(colnames(train)[1], "~", paste(colnames(train)[-1],
##     collapse = "+"), sep = "")), family = poisson(), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9754  -0.7224   0.0655   0.5811   3.2366
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.658e+00  2.183e-01   7.596 3.06e-14 ***
## FixedAcidity        -2.671e-04  9.124e-04  -0.293 0.769745
## VolatileAcidity     -3.583e-02  7.324e-03  -4.892 9.99e-07 ***
## CitricAcid           4.493e-03  6.566e-03   0.684 0.493822
## ResidualSugar        2.188e-05  1.688e-04   0.130 0.896901
## Chlorides           -4.339e-02  1.781e-02  -2.437 0.014826 *
## FreeSulfurDioxide    1.262e-04  3.831e-05   3.293 0.000992 ***
## TotalSulfurDioxide  7.892e-05  2.464e-05   3.203 0.001359 **
## Density             -4.335e-01  2.145e-01  -2.022 0.043223 *
## pH                  -1.336e-02  8.397e-03  -1.591 0.111635
## Sulphates           -1.486e-02  6.131e-03  -2.424 0.015339 *
## Alcohol              2.316e-03  1.534e-03   1.510 0.131058
## LabelAppeal          1.330e-01  6.800e-03  19.553  < 2e-16 ***
## AcidIndex           -8.589e-02  5.125e-03 -16.758  < 2e-16 ***
## STARS                3.138e-01  5.076e-03  61.817  < 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: 18225  on 10235  degrees of freedom
## Residual deviance: 11716  on 10221  degrees of freedom
## AIC: 37311
##
## Number of Fisher Scoring iterations: 5
```

Here we see that the Perception variables have an outsize impact on the outcome.

Let's create a Poisson model using only the two Perception variables:

```
pm2 <- glm(TARGET ~ STARS + LabelAppeal,data = train,family=poisson())
summary(pm2)
```

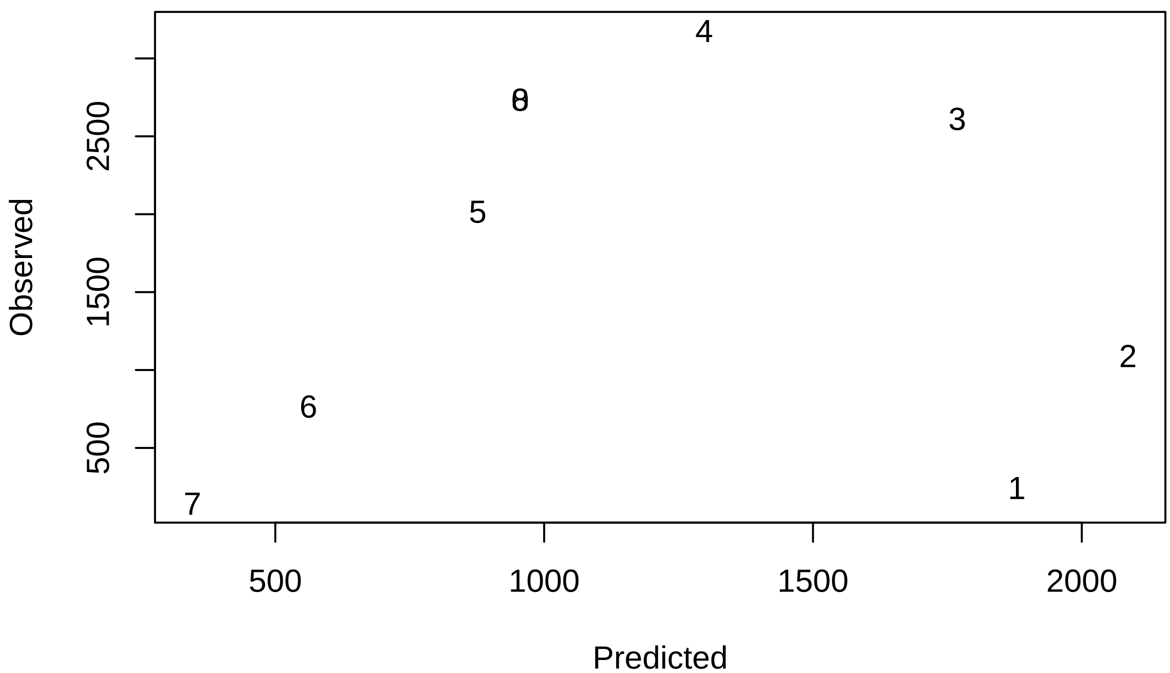
```
##
## Call:
## glm(formula = TARGET ~ STARS + LabelAppeal, family = poisson(),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8887  -0.7644   0.0787   0.6151   3.2902
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.514677   0.011250   45.75   <2e-16 ***
```

```
## STARS       0.331690   0.004969   66.75   <2e-16 ***
## LabelAppeal 0.125219   0.006773   18.49   <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: 18225  on 10235  degrees of freedom
## Residual deviance: 12107  on 10233  degrees of freedom
## AIC: 37677
##
## Number of Fisher Scoring iterations: 5
```

NEED VERBIAGE - REGULAR POISSION MODEL

Zero-inflated Poisson Model

We next explore the seemingly high number of zero cases in the TARGET count as seen in the previous histrogram. We can easily see if the number of zeros observed is in line with the number of zeros predicted by the poission model alone.



The number of observed zero cases and the predicted zero cases do not match up well so we'll move to look at the influence of the zero counts on the model by separating out the modeling of zero counts and the modeling of the non-zero counts.

Staying with our concepts of Perception and Chemistry, we will look treating the high number of zero counts using the Perception variables of STARS and LabelAppeal, and the non-zero counts will use all other variables as the Chemistry variables.

```
##
## Call:
## zeroinfl(formula = TARGET ~ . - (STARS + LabelAppeal) | STARS +
```

```
##     LabelAppeal, data = wine, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -1.95738 -0.49271  0.04324  0.52506  4.77668
##
## Count model coefficients (poisson with log link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.856e+00  2.013e-01   9.222  < 2e-16 ***
## FixedAcidity        4.833e-05  8.347e-04   0.058 0.953823
## VolatileAcidity    -2.338e-02  6.678e-03  -3.502 0.000462 ***
## CitricAcid          4.259e-03  6.025e-03   0.707 0.479684
## ResidualSugar       5.580e-05  1.536e-04   0.363 0.716495
## Chlorides          -2.190e-02  1.637e-02  -1.337 0.181063
## FreeSulfurDioxide   3.729e-05  3.457e-05   1.079 0.280806
## TotalSulfurDioxide -1.892e-05  2.189e-05  -0.864 0.387504
## Density            -4.170e-01  1.978e-01  -2.108 0.034994 *
## pH                  9.340e-03  7.683e-03   1.216 0.224114
## Sulphates          -3.423e-03  5.633e-03  -0.608 0.543438
## Alcohol             8.943e-03  1.382e-03   6.470 9.80e-11 ***
## AcidIndex          -2.973e-02  5.042e-03  -5.896 3.72e-09 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.57356    0.03747   15.31   <2e-16 ***
## STARS       -2.28586    0.05256  -43.49   <2e-16 ***
## LabelAppeal  0.55872    0.03628   15.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 21
## Log-likelihood: -2.196e+04 on 16 Df
```

```
## [1] 3104.853
```

```
## [1] "Chi-Square Test =  0.492307100075339"
```

Given the large p-value from the chi-square test, we conclude our model approach for Chemsitry vs Perception is valid.

After analyzing the p-values for the Chemistry portion of the zero-inflated model, there are only 4 statistically significant variables: VolatileAcidity, Density, Alcohol, and AcidIndex. We'll re-reun the zero-inflated poission model with just these variables in the poission portion.

```
##
## Call:
## zeroinfl(formula = TARGET ~ (VolatileAcidity + Density + Alcohol +
##     AcidIndex) - (STARS + LabelAppeal) | STARS + LabelAppeal, data = wine,
##     dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -1.95787 -0.49209  0.04346  0.52825  4.79571
##
```

```
## Count model coefficients (poisson with log link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.893616   0.199157   9.508  < 2e-16 ***
## VolatileAcidity -0.023471   0.006676  -3.516 0.000438 ***
## Density         -0.424586   0.197596  -2.149 0.031653 *
## Alcohol          0.008989   0.001381   6.509 7.57e-11 ***
## AcidIndex       -0.030056   0.004978  -6.038 1.56e-09 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.57363    0.03747   15.31   <2e-16 ***
## STARS       -2.28583    0.05254  -43.51   <2e-16 ***
## LabelAppeal  0.55895    0.03627   15.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 13
## Log-likelihood: -2.196e+04 on 8 Df
```

We have reduced the degrees-of-freedom from 16 down to 8 which is as far as we'll go with the zero-inflated poission model.

Regular Negative Binomial Model

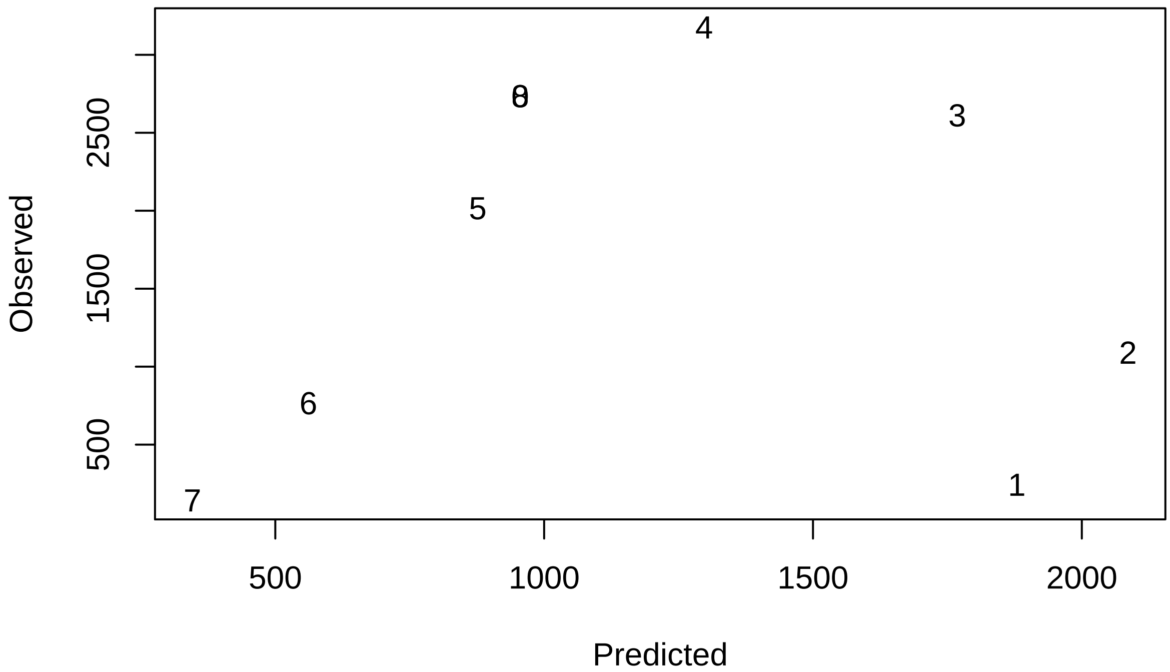
```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached

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

NEED VERBIAGE - REGULAR NEGATIVE BINOMIAL MODEL

Zero-inflated Negative Regession Model

We'll continue our exploration of the seemingly high number of zero cases in the TARGET count as seen in the previous histrogram. In this case, we'll see if the number of zeros observed is in line with the number of zeros predicted by the negative binomial model alone.



The number of observed zero cases and the predicted zero cases do not match up well so we'll move to look at the influence of the zero counts on the model by separating out the modeling of zero counts and the modeling of the non-zero counts.

Staying with our concepts of Perception and Chemistry, we will look treating the high number of zero counts using the Perception variables of STARS and LabelAppeal, and the non-zero counts will use all other variables as the Chemistry variables.

```
##
## Call:
## zeroinfl(formula = TARGET ~ . - (STARS + LabelAppeal) | (STARS +
##     LabelAppeal), data = wine, dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -1.95734 -0.49272  0.04329  0.52502  4.77668
##
## Count model coefficients (negbin with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.855e+00  2.013e-01   9.217  < 2e-16 ***
## FixedAcidity         4.985e-05  8.347e-04   0.060 0.952381
## VolatileAcidity     -2.335e-02  6.678e-03  -3.497 0.000471 ***
## CitricAcid           4.259e-03  6.026e-03   0.707 0.479683
## ResidualSugar        5.551e-05  1.536e-04   0.361 0.717898
## Chlorides           -2.190e-02  1.637e-02  -1.338 0.181055
## FreeSulfurDioxide    3.732e-05  3.457e-05   1.080 0.280320
## TotalSulfurDioxide -1.889e-05  2.189e-05  -0.863 0.388130
## Density             -4.164e-01  1.978e-01  -2.106 0.035243 *
## pH                   9.325e-03  7.683e-03   1.214 0.224864
## Sulphates           -3.453e-03  5.633e-03  -0.613 0.539892
## Alcohol              8.943e-03  1.382e-03   6.470 9.83e-11 ***
## AcidIndex           -2.969e-02  5.042e-03  -5.888 3.92e-09 ***
## Log(theta)           1.166e+01  3.296e+00   3.536 0.000406 ***
```

```
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.57359    0.03747   15.31   <2e-16 ***
## STARS       -2.28592    0.05256  -43.49   <2e-16 ***
## LabelAppeal  0.55870    0.03628   15.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 115399.9042
## Number of iterations in BFGS optimization: 36
## Log-likelihood: -2.196e+04 on 17 Df

## Warning in sqrt(diag(vc)[np]): NaNs produced

## [1] 3105.036

## [1] "Chi-Square Test =  0.49644118505298"
```

Given the large p-value from the chi-square test, we conclude our model approach for Chemsitry vs Perception is valid.

After analyzing the p-values for the Chemistry portion of the zero-inflated model, there are only 4 statistically significant variables: VolatileAcidity, Density, Alcohol, and AcidIndex. We'll re-reun the zero-inflated poission model with just these variables in the negative binomial portion.

```
##
## Call:
## zeroinfl(formula = TARGET ~ (VolatileAcidity + Density + Alcohol +
##     AcidIndex) - (STARS + LabelAppeal) | STARS + LabelAppeal, data = wine,
##     dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -1.95787 -0.49210  0.04347  0.52824  4.79557
##
## Count model coefficients (negbin with log link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.893733   0.199157   9.509  < 2e-16 ***
## VolatileAcidity -0.023471   0.006676  -3.516 0.000438 ***
## Density         -0.424669   0.197596  -2.149 0.031620 *
## Alcohol          0.008988   0.001381   6.509 7.58e-11 ***
## AcidIndex       -0.030060   0.004978  -6.038 1.56e-09 ***
## Log(theta)      15.157583  12.636907   1.199 0.230345
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.57361    0.03747   15.31   <2e-16 ***
## STARS       -2.28573    0.05254  -43.51   <2e-16 ***
## LabelAppeal  0.55893    0.03627   15.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Theta = 3826965.5283
## Number of iterations in BFGS optimization: 39
## Log-likelihood: -2.196e+04 on 9 Df
```

We have reduced the degrees-of-freedom from 17 down to 9 which is as far as we'll go with the zero-inflated negative binomial model.

Select Models

Select Models

NEED VERBIAGE - SELECT MODELS

Smooth Operators - All Done!