DATA_621_HW5

Chi Pong, Euclid Zhang, Jie Zou, Joseph Connolly, LeTicia Cancel

4/10/2022

```
train_df <- read.csv("wine-training-data.csv",fileEncoding="UTF-8-BOM")
test_df <- read.csv("wine-evaluation-data.csv",fileEncoding="UTF-8-BOM")

train_df$INDEX <- NULL
test_df$IN <- NULL
```

DATA EXPLORATION

Data Summary

```
summary(train_df)
```

```
##      TARGET       FixedAcidity      VolatileAcidity     CitricAcid
##  Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400
##  1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300
##  Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100
##  Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084
##  3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800
##  Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600
##
##  ResidualSugar        Chlorides       FreeSulfurDioxide TotalSulfurDioxide
##  Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0
##  1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0
##  Median :   3.900   Median : 0.0460   Median :  30.00   Median : 123.0
##  Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7
##  3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0
##  Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0
##  NA's   :616        NA's   :638       NA's   :647       NA's   :682
##     Density             pH          Sulphates          Alcohol
##  Min.   :0.8881   Min.   :0.480   Min.   :-3.1300   Min.   :-4.70
##  1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00
##  Median :0.9945   Median :3.200   Median : 0.5000   Median :10.40
##  Mean   :0.9942   Mean   :3.208   Mean   : 0.5271   Mean   :10.49
##  3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40
##  Max.   :1.0992   Max.   :6.130   Max.   : 4.2400   Max.   :26.50
##                   NA's   :395     NA's   :1210      NA's   :653
##   LabelAppeal          AcidIndex          STARS
##  Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
```

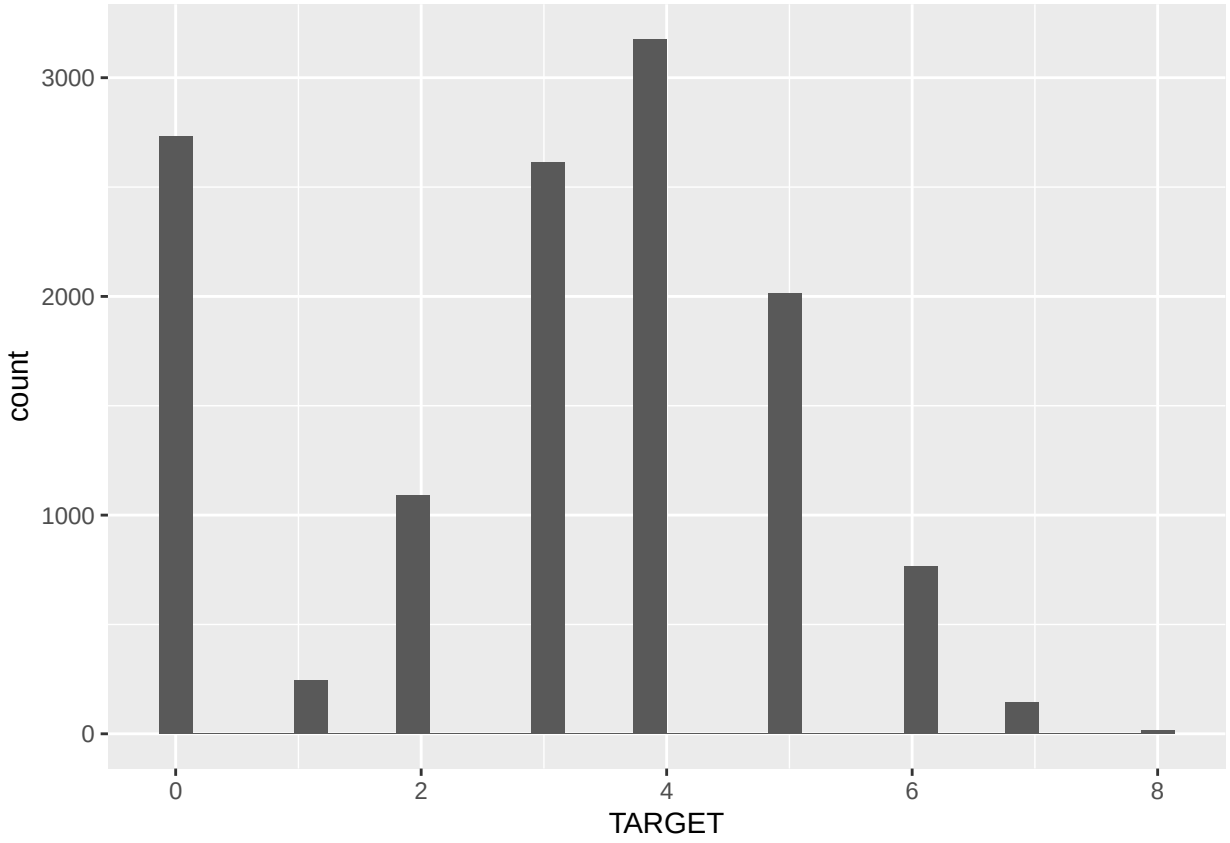
```
##  1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##                                       NA's   :3359
```

From the summary:

- We can see that most of the chemical properties range from a negative value to a positive value of similar magnitude. These predictor variables seem to be already scaled / standardized. Hence, there is no extreme outliers.
- **ResidualSugar**, **Chlorides**, **FreeSulfurDioxide**, **TotalSulfurDioxide**, **pH**, **Sulphates**, **Alcohol** have numerous missing values. We will impute the missing values using mice (multivariate imputation by chained equations).
- **STARS** also have missing values. However, the values are missing simply because they don't have a rating, not because of data collecting problems. We may consider imputing this variable differently.

Distribution plots

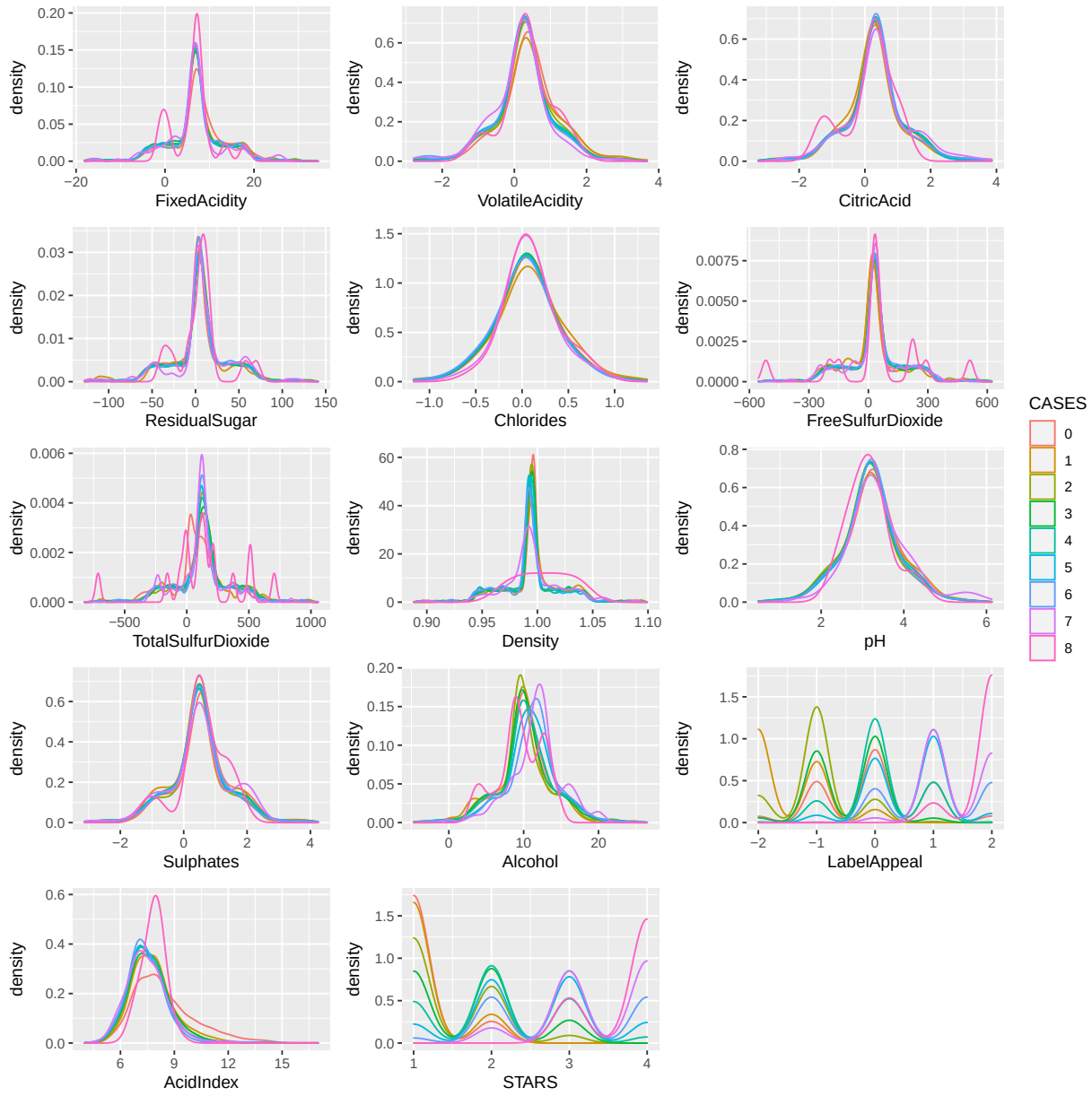
```
ggplot(train_df, aes(x=TARGET)) + geom_histogram(na.rm =TRUE)
```



From above plot, we see that the target value is **zero-inflated**, not a regular poisson distribution nor any distribution of the exponential family.

Hence, practically it is not suggested to fit the data to a poisson, negative binomial or linear model. We will not take the validity of the models seriously.

```
CASES <- as.factor(train_df$TARGET)

plot_FixedAcidity <- ggplot(train_df, aes(x=FixedAcidity, color=CASES)) + geom_density(na.rm =TRUE, bw=
plot_VolatileAcidity <- ggplot(train_df, aes(x=VolatileAcidity, color=CASES)) + geom_density(na.rm =TRU
plot_CitricAcid <- ggplot(train_df, aes(x=CitricAcid, color=CASES)) + geom_density(na.rm =TRUE, bw=0.3)
plot_ResidualSugar <- ggplot(train_df, aes(x=ResidualSugar, color=CASES)) + geom_density(na.rm =TRUE, bw
plot_Chlorides <- ggplot(train_df, aes(x=Chlorides, color=CASES)) + geom_density(na.rm =TRUE, bw=0.2)
plot_FreeSulfurDioxide <- ggplot(train_df, aes(x=FreeSulfurDioxide, color=CASES)) + geom_density(na.rm =
plot_TotalSulfurDioxide <- ggplot(train_df, aes(x=TotalSulfurDioxide, color=CASES)) + geom_density(na.rm
plot_Density <- ggplot(train_df, aes(x=Density, color=CASES)) + geom_density(na.rm =TRUE)
plot_pH <- ggplot(train_df, aes(x=pH, color=CASES)) + geom_density(na.rm =TRUE, bw=0.3)
plot_Sulphates <- ggplot(train_df, aes(x=Sulphates, color=CASES)) + geom_density(na.rm =TRUE, bw=0.3)
plot_Alcohol <- ggplot(train_df, aes(x=Alcohol, color=CASES)) + geom_density(na.rm =TRUE, bw=0.8)
plots_LabelAppeal <- ggplot(train_df, aes(x=LabelAppeal, color=CASES)) + geom_density(na.rm =TRUE, bw=0
plots_AcidIndex <- ggplot(train_df, aes(x=AcidIndex, color=CASES)) + geom_density(na.rm =TRUE, bw=0.5)
plots_STARS <- ggplot(train_df, aes(x=STARS, color=CASES)) + geom_density(na.rm =TRUE, bw=0.2)

plot_FixedAcidity+plot_VolatileAcidity+plot_CitricAcid+plot_ResidualSugar+plot_Chlorides+
  plot_FreeSulfurDioxide+plot_TotalSulfurDioxide+plot_Density+plot_pH+plot_Sulphates+
  plot_Alcohol+plots_LabelAppeal+plots_AcidIndex+plots_STARS+
  plot_layout(ncol = 3, guides = "collect")
```

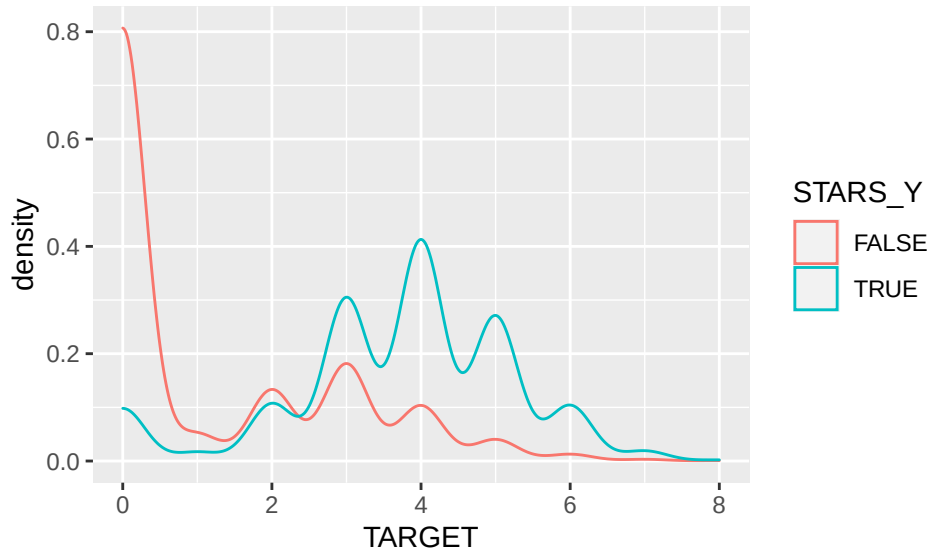


The distributions of the predictor variables show that **LabelAppeal** and **STARS** are good candidates of predicting the target variable. The distributions of other variables do not vary a lot based on the different values of the target variable.

```
ResidualSugar_Y <- !is.na(train_df$ResidualSugar)
Chlorides_Y <- !is.na(train_df$Chlorides)
FreeSulfurDioxide_Y <- !is.na(train_df$FreeSulfurDioxide)
TotalSulfurDioxide_Y <- !is.na(train_df$TotalSulfurDioxide)
pH_Y <- !is.na(train_df$pH)
Sulphates_Y <- !is.na(train_df$Sulphates)
Alcohol_Y <- !is.na(train_df$Alcohol)
STARS_Y <- !is.na(train_df$STARS)
```

Now, let's check whether **STARS** is missing or not have an effect to the cases of wine purchased.

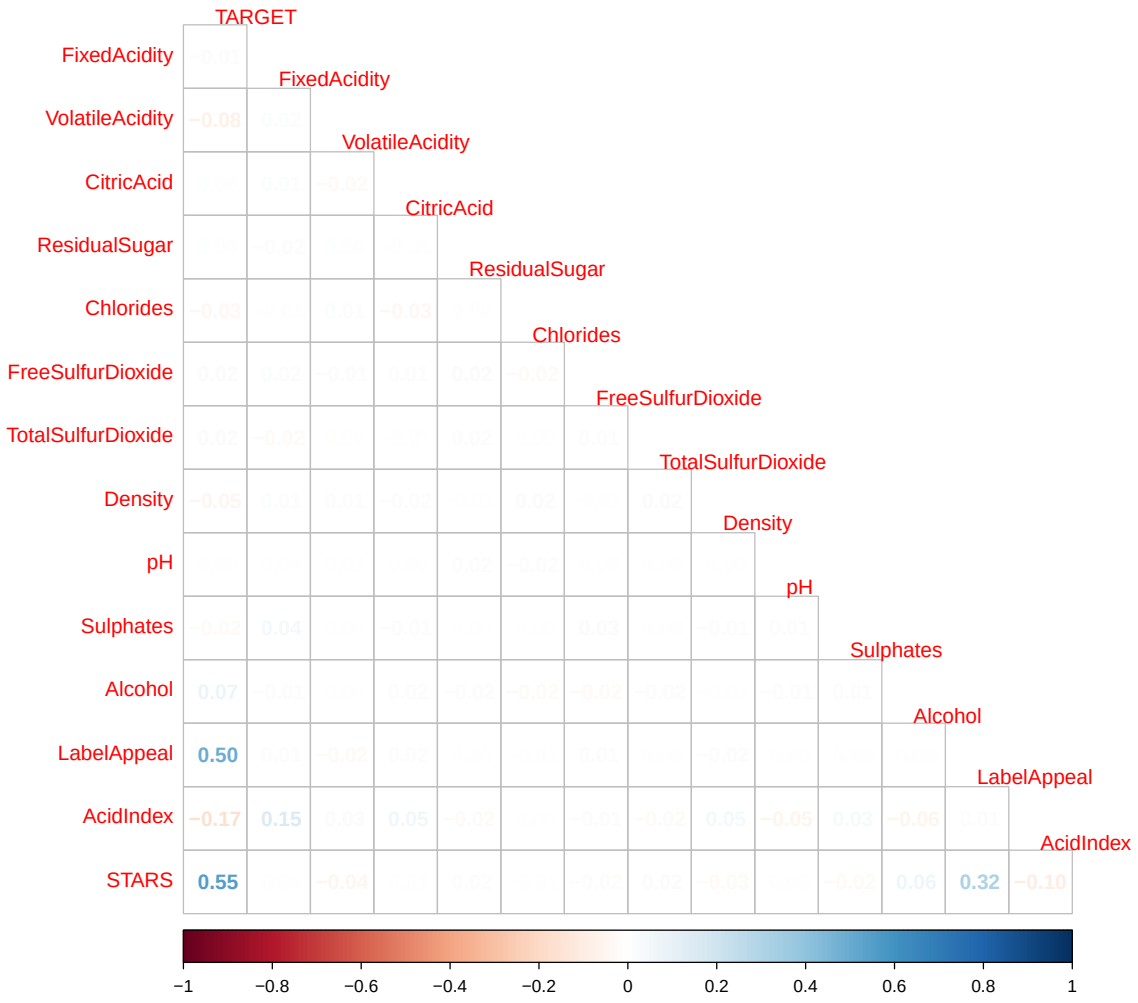
```
ggplot(train_df, aes(x=TARGET, color=STARS_Y)) + geom_density(na.rm =TRUE, bw=0.3)
```



The above distributions plot indicates that most people are willing to buy wines with **STARS** provided and not willing to buy wines with **STARS** unavailable. We may add a dummy variable, or transform the **STARS** variable to indicate **STARS** is available or not.

Correlations

```
corrplot::corrplot(cor(train_df, use = "na.or.complete"),
                   method = 'number', type = 'lower', diag = FALSE, tl.srt = 0.1)
```



Looking at the correlations of the predictor values, there is no multicollinearity problem of the data.

DATA PREPARATION

Data Imputation

For imputing the missing values of the chemical properties, the following variables are not included as predictors:

- **TARGET**: the target variable should not be used to predict the missing values of the predictors, as the objective of the models is to predict the target variables using the predictors.
- **LabelAppeal**: the label appeal of the bottle should not have anything to do with the chemical properties of the wines.

- **STARS**: More than 25% of the wines have STARS unavailable. Whether it is missing or not should not have anything to do with the chemical properties of the wines.

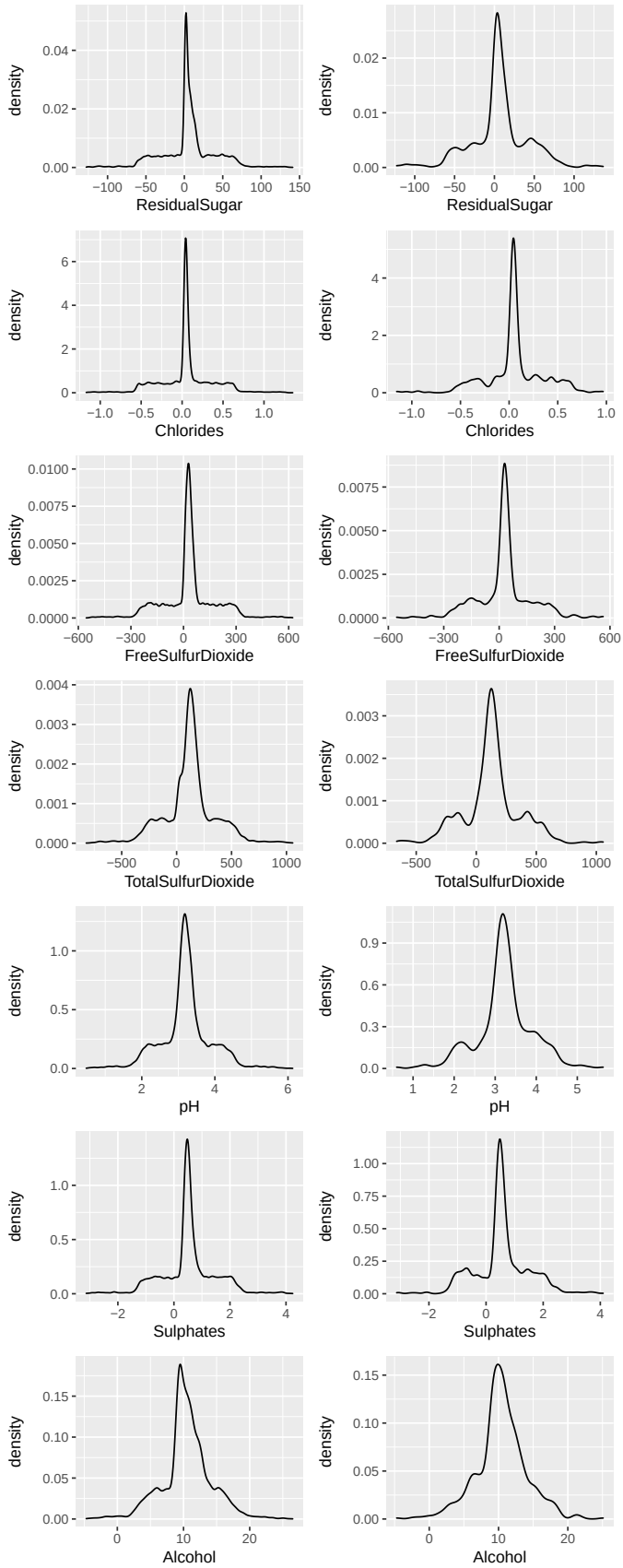
multivariate imputation by chained equations is used to impute the missing values

```
#temporary exclude TARGET, LabelAppeal, and STARS in our imputation
TARGET <- train_df$TARGET
LabelAppeal <- train_df$LabelAppeal
STARS <- train_df$STARS

train_df$TARGET <- NULL
train_df$LabelAppeal <- NULL
train_df$STARS <- NULL

#save the imputation models to impute the test data set later
mickey <- parlmice(train_df, maxit = 5, m = 1, printFlag = FALSE, seed = 2022,
                   cluster.seed = 2022)

#save the imputation result
train_df <- complete(mickey,1)

#Add TARGET, LabelAppeal, and STARS back to our dataframe
train_df$TARGET <- TARGET
train_df$LabelAppeal <- LabelAppeal
train_df$STARS <- STARS

TARGET <- NULL
LabelAppeal <- NULL
STARS <- NULL
```

We can compare the imputed data values and the original data values.
The plots on the left below show the distributions of the values from the original data.
The plots on the right below show the distributions of the imputed values.

```
plot_ResidualSugar <- ggplot(train_df[ResidualSugar_Y,], aes(x=ResidualSugar)) +
  geom_density(na.rm =TRUE)
plot_Chlorides <- ggplot(train_df[Chlorides_Y,], aes(x=Chlorides)) +
  geom_density(na.rm =TRUE)
plot_FreeSulfurDioxide <- ggplot(train_df[FreeSulfurDioxide_Y,], aes(x=FreeSulfurDioxide)) +
  geom_density(na.rm =TRUE)
plot_TotalSulfurDioxide <- ggplot(train_df[TotalSulfurDioxide_Y,], aes(x=TotalSulfurDioxide)) +
  geom_density(na.rm =TRUE)
plot_pH <- ggplot(train_df[pH_Y,], aes(x=pH)) +
  geom_density(na.rm =TRUE)
plot_Sulphates <- ggplot(train_df[Sulphates_Y,], aes(x=Sulphates)) +
  geom_density(na.rm =TRUE)
plot_Alcohol <- ggplot(train_df[Alcohol_Y,], aes(x=Alcohol)) +
  geom_density(na.rm =TRUE)

plot_ResidualSugar2 <- ggplot(train_df[!ResidualSugar_Y,], aes(x=ResidualSugar)) +
  geom_density(na.rm =TRUE)
plot_Chlorides2 <- ggplot(train_df[!Chlorides_Y,], aes(x=Chlorides)) +
  geom_density(na.rm =TRUE)
```

```
plot_FreeSulfurDioxide2 <- ggplot(train_df[!FreeSulfurDioxide_Y,], aes(x=FreeSulfurDioxide)) +
  geom_density(na.rm =TRUE)
plot_TotalSulfurDioxide2 <- ggplot(train_df[!TotalSulfurDioxide_Y,], aes(x=TotalSulfurDioxide)) +
  geom_density(na.rm =TRUE)
plot_pH2 <- ggplot(train_df[!pH_Y,], aes(x=pH)) +
  geom_density(na.rm =TRUE)
plot_Sulphates2 <- ggplot(train_df[!Sulphates_Y,], aes(x=Sulphates)) +
  geom_density(na.rm =TRUE)
plot_Alcohol2 <- ggplot(train_df[!Alcohol_Y,], aes(x=Alcohol)) +
  geom_density(na.rm =TRUE)

plot_ResidualSugar+plot_ResidualSugar2+
  plot_Chlorides+plot_Chlorides2+
  plot_FreeSulfurDioxide+plot_FreeSulfurDioxide2+
  plot_TotalSulfurDioxide+plot_TotalSulfurDioxide2+
  plot_pH+plot_pH2+
  plot_Sulphates+plot_Sulphates2+
  plot_Alcohol+plot_Alcohol2+
  plot_layout(ncol = 2, guides = "collect")
```


density
ResidualSugar
Chlorides
FreeSulfurDioxide
TotalSulfurDioxide
pH
Sulphates
Alcohol

The distributions look similar and so the imputed values are plausible.

Data Transformation

As discussed above, whether **STARS** is available or not is predictive of the target. Moreover, the marginal effect of increasing 1 star may not be equal. For example, the effect from 1 star to 2 star may not be the same as the effect from 4 star to 5 star. Hence, we will impute the missing values of **STARS** by 0 and convert **STARS** to a factor variable. The variable will then be converted to 4 dummies variables in the models.

Similarly, we will also convert **LabelAppeal** to a factor variable as the marginal effect may changes.

```
train_df$STARS[!STARS_Y] <- 0
train_df$STARS <- as.factor(train_df$STARS)
train_df$LabelAppeal <- as.factor(train_df$LabelAppeal)
```

BUILD MODELS

Poisson models

We start building our Poisson model with all predictors.

```
poisson_full <- glm(TARGET ~ ., data=train_df, family=poisson)
summary(poisson_full)
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2252  -0.6543  -0.0040   0.4508   3.7773
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          6.972e-01  1.990e-01   3.504 0.000458 ***
## FixedAcidity         2.626e-05  8.200e-04   0.032 0.974453
## VolatileAcidity     -3.051e-02  6.529e-03  -4.673 2.96e-06 ***
## CitricAcid           5.119e-03  5.897e-03   0.868 0.385411
## ResidualSugar        3.821e-05  1.503e-04   0.254 0.799361
## Chlorides           -3.924e-02  1.611e-02  -2.435 0.014894 *
## FreeSulfurDioxide    7.974e-05  3.425e-05   2.328 0.019896 *
## TotalSulfurDioxide   7.190e-05  2.210e-05   3.254 0.001139 **
## Density             -2.571e-01  1.918e-01  -1.341 0.179995
## pH                  -1.381e-02  7.529e-03  -1.835 0.066577 .
## Sulphates           -1.084e-02  5.477e-03  -1.980 0.047752 *
## Alcohol              3.543e-03  1.376e-03   2.575 0.010014 *
## AcidIndex           -7.989e-02  4.572e-03 -17.474  < 2e-16 ***
## LabelAppeal-1        2.353e-01  3.799e-02   6.193 5.89e-10 ***
## LabelAppeal0         4.254e-01  3.705e-02  11.480  < 2e-16 ***
## LabelAppeal1         5.577e-01  3.769e-02  14.794  < 2e-16 ***
```

```
## LabelAppeal2        6.958e-01  4.244e-02  16.395  < 2e-16 ***
## STARS1              7.663e-01  1.954e-02  39.214  < 2e-16 ***
## STARS2              1.085e+00  1.824e-02  59.500  < 2e-16 ***
## STARS3              1.205e+00  1.920e-02  62.753  < 2e-16 ***
## STARS4              1.325e+00  2.431e-02  54.490  < 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: 22861  on 12794  degrees of freedom
## Residual deviance: 13639  on 12774  degrees of freedom
## AIC: 45623
##
## Number of Fisher Scoring iterations: 6
```

Backward Elimination by AIC

Starting with our full model, perform backward elimination by comparing the **AIC** of the models.

```
poisson_AIC <- step(poisson_full,trace=0)
summary(poisson_AIC)
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + pH + Sulphates + Alcohol + AcidIndex +
##     LabelAppeal + STARS, family = poisson, data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2282  -0.6537  -0.0040   0.4485   3.7686
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         4.428e-01  6.105e-02   7.253 4.08e-13 ***
## VolatileAcidity    -3.071e-02  6.528e-03  -4.705 2.54e-06 ***
## Chlorides          -3.981e-02  1.611e-02  -2.471  0.01347 *
## FreeSulfurDioxide   7.976e-05  3.423e-05   2.330  0.01982 *
## TotalSulfurDioxide  7.142e-05  2.209e-05   3.234  0.00122 **
## pH                 -1.379e-02  7.527e-03  -1.832  0.06697 .
## Sulphates          -1.083e-02  5.475e-03  -1.978  0.04789 *
## Alcohol             3.576e-03  1.375e-03   2.600  0.00931 **
## AcidIndex          -7.987e-02  4.514e-03 -17.695  < 2e-16 ***
## LabelAppeal-1       2.351e-01  3.799e-02   6.190 6.03e-10 ***
## LabelAppeal0        4.254e-01  3.705e-02  11.481  < 2e-16 ***
## LabelAppeal1        5.579e-01  3.769e-02  14.800  < 2e-16 ***
## LabelAppeal2        6.954e-01  4.244e-02  16.388  < 2e-16 ***
## STARS1              7.665e-01  1.954e-02  39.229  < 2e-16 ***
## STARS2              1.086e+00  1.823e-02  59.550  < 2e-16 ***
## STARS3              1.206e+00  1.920e-02  62.789  < 2e-16 ***
## STARS4              1.325e+00  2.431e-02  54.523  < 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: 22861  on 12794  degrees of freedom
## Residual deviance: 13642  on 12778  degrees of freedom
## AIC: 45618
##
## Number of Fisher Scoring iterations: 6
```

Backward Elimination by BIC

Starting with our full model, perform backward elimination by comparing the **BIC** of the models.

```
poisson_BIC <- step(poisson_full,trace=0, k=log(nrow(train_df)))
summary(poisson_BIC)
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + TotalSulfurDioxide +
##     AcidIndex + LabelAppeal + STARS, family = poisson, data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2430  -0.6534  -0.0061   0.4548   3.8100
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          4.337e-01  5.236e-02   8.283  < 2e-16 ***
## VolatileAcidity     -3.090e-02  6.529e-03  -4.732 2.22e-06 ***
## TotalSulfurDioxide   7.142e-05  2.207e-05   3.235  0.00122 **
## AcidIndex           -8.049e-02  4.496e-03 -17.901  < 2e-16 ***
## LabelAppeal-1        2.350e-01  3.798e-02   6.187 6.12e-10 ***
## LabelAppeal0         4.253e-01  3.705e-02  11.479  < 2e-16 ***
## LabelAppeal1         5.569e-01  3.768e-02  14.778  < 2e-16 ***
## LabelAppeal2         6.951e-01  4.243e-02  16.384  < 2e-16 ***
## STARS1               7.689e-01  1.953e-02  39.362  < 2e-16 ***
## STARS2               1.089e+00  1.823e-02  59.727  < 2e-16 ***
## STARS3               1.210e+00  1.917e-02  63.136  < 2e-16 ***
## STARS4               1.330e+00  2.427e-02  54.812  < 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: 22861  on 12794  degrees of freedom
## Residual deviance: 13667  on 12783  degrees of freedom
## AIC: 45633
##
## Number of Fisher Scoring iterations: 6
```

Negative Binomial models

We start building our Negative Binomial model with all predictors.

Because the data is zero inflated. the glm.nb function is not able to find the optimal value for the additional parameter r. Since the density is highest at target = 0, we will build our model using r = 1.

```
nb_full <- glm(TARGET ~ ., data=train_df,negative.binomial(1))
summary(nb_full)
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = negative.binomial(1), data = train_df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.90250  -0.34154  -0.01238   0.21600   2.02600
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.0743616  0.2382782   4.509 6.58e-06 ***
## FixedAcidity        -0.0001696  0.0009888  -0.171  0.86384
## VolatileAcidity     -0.0419826  0.0078679  -5.336 9.67e-08 ***
## CitricAcid           0.0065673  0.0071404   0.920  0.35773
## ResidualSugar        0.0001077  0.0001815   0.593  0.55306
## Chlorides           -0.0568091  0.0194292  -2.924  0.00346 **
## FreeSulfurDioxide    0.0001182  0.0000414   2.855  0.00431 **
## TotalSulfurDioxide   0.0001229  0.0000266   4.620 3.87e-06 ***
## Density             -0.2809959  0.2317799  -1.212  0.22541
## pH                  -0.0276481  0.0090722  -3.048  0.00231 **
## Sulphates           -0.0183074  0.0066113  -2.769  0.00563 **
## Alcohol              0.0021878  0.0016561   1.321  0.18651
## AcidIndex           -0.1134576  0.0051465 -22.046  < 2e-16 ***
## LabelAppeal-1        0.2216406  0.0365876   6.058 1.42e-09 ***
## LabelAppeal0         0.3896430  0.0356763  10.922  < 2e-16 ***
## LabelAppeal1         0.4900872  0.0369138  13.277  < 2e-16 ***
## LabelAppeal2         0.6309517  0.0461819  13.662  < 2e-16 ***
## STARS1               0.7581734  0.0187883  40.353  < 2e-16 ***
## STARS2               1.0885802  0.0179498  60.646  < 2e-16 ***
## STARS3               1.2172280  0.0201044  60.545  < 2e-16 ***
## STARS4               1.3490347  0.0301766  44.705  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1) family taken to be 0.3428193)
##
##     Null deviance: 9042.5  on 12794  degrees of freedom
## Residual deviance: 6474.6  on 12774  degrees of freedom
## AIC: 55248
##
## Number of Fisher Scoring iterations: 5
```

Backward Elimination by AIC

Starting with our full model, perform backward elimination by comparing the **AIC** of the models.

```
nb_AIC <- step(nb_full,trace=0)
summary(nb_AIC)
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + pH + Sulphates + AcidIndex + LabelAppeal +
##     STARS, family = negative.binomial(1), data = train_df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.90425  -0.34172  -0.01118   0.21688   2.02832
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8.229e-01  6.248e-02  13.171  < 2e-16 ***
## VolatileAcidity     -4.229e-02  7.868e-03  -5.375 7.78e-08 ***
## Chlorides           -5.825e-02  1.942e-02  -2.999  0.00271 **
## FreeSulfurDioxide    1.174e-04  4.139e-05   2.837  0.00456 **
## TotalSulfurDioxide   1.224e-04  2.659e-05   4.603 4.21e-06 ***
## pH                  -2.772e-02  9.071e-03  -3.056  0.00225 **
## Sulphates           -1.827e-02  6.609e-03  -2.764  0.00572 **
## AcidIndex           -1.139e-01  5.066e-03 -22.472  < 2e-16 ***
## LabelAppeal-1        2.209e-01  3.659e-02   6.037 1.62e-09 ***
## LabelAppeal0         3.888e-01  3.568e-02  10.899  < 2e-16 ***
## LabelAppeal1         4.892e-01  3.691e-02  13.253  < 2e-16 ***
## LabelAppeal2         6.298e-01  4.619e-02  13.635  < 2e-16 ***
## STARS1               7.585e-01  1.879e-02  40.367  < 2e-16 ***
## STARS2               1.090e+00  1.794e-02  60.744  < 2e-16 ***
## STARS3               1.219e+00  2.008e-02  60.704  < 2e-16 ***
## STARS4               1.352e+00  3.014e-02  44.849  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1) family taken to be 0.3429827)
##
##     Null deviance: 9042.5  on 12794  degrees of freedom
## Residual deviance: 6476.1  on 12779  degrees of freedom
## AIC: 55239
##
## Number of Fisher Scoring iterations: 5
```

Backward Elimination by BIC

Starting with our full model, perform backward elimination by comparing the **BIC** of the models.

```
nb_BIC <- step(nb_full,trace=0, k=log(nrow(train_df)))
summary(nb_BIC)
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + TotalSulfurDioxide +
##     AcidIndex + LabelAppeal + STARS, family = negative.binomial(1),
##     data = train_df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.91436  -0.34141  -0.01335   0.21564   2.06009
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.226e-01  5.375e-02  13.443  < 2e-16 ***
## VolatileAcidity    -4.243e-02  7.869e-03  -5.392 7.10e-08 ***
## TotalSulfurDioxide  1.237e-04  2.659e-05   4.652 3.32e-06 ***
## AcidIndex          -1.137e-01  5.050e-03 -22.520  < 2e-16 ***
## LabelAppeal-1       2.213e-01  3.659e-02   6.050 1.49e-09 ***
## LabelAppeal0        3.892e-01  3.568e-02  10.909  < 2e-16 ***
## LabelAppeal1        4.896e-01  3.691e-02  13.266  < 2e-16 ***
## LabelAppeal2        6.282e-01  4.618e-02  13.602  < 2e-16 ***
## STARS1              7.608e-01  1.878e-02  40.512  < 2e-16 ***
## STARS2              1.091e+00  1.793e-02  60.839  < 2e-16 ***
## STARS3              1.222e+00  2.007e-02  60.868  < 2e-16 ***
## STARS4              1.352e+00  3.014e-02  44.853  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1) family taken to be 0.3431495)
##
##     Null deviance: 9042.5  on 12794  degrees of freedom
## Residual deviance: 6487.7  on 12783  degrees of freedom
## AIC: 55243
##
## Number of Fisher Scoring iterations: 5
```

Multiple Linear Regression Models

We start building our Multiple Linear Regression model with all predictors.

```
lm_full <- lm(TARGET ~ ., data=train_df)
summary(lm_full)
```

```
##
## Call:
## lm(formula = TARGET ~ ., data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9597 -0.8612  0.0221  0.8418  6.1818
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.865e+00  4.465e-01   6.416 1.45e-10 ***
```

```
## FixedAcidity        5.594e-04  1.859e-03   0.301 0.763526
## VolatileAcidity    -9.474e-02  1.478e-02  -6.411 1.50e-10 ***
## CitricAcid          1.711e-02  1.344e-02   1.273 0.202920
## ResidualSugar       1.295e-04  3.415e-04   0.379 0.704491
## Chlorides          -1.245e-01  3.646e-02  -3.414 0.000643 ***
## FreeSulfurDioxide   2.378e-04  7.798e-05   3.049 0.002299 **
## TotalSulfurDioxide  2.052e-04  4.994e-05   4.109 4.01e-05 ***
## Density            -7.945e-01  4.359e-01  -1.823 0.068381 .
## pH                 -3.598e-02  1.704e-02  -2.111 0.034753 *
## Sulphates          -2.914e-02  1.241e-02  -2.347 0.018917 *
## Alcohol             1.204e-02  3.114e-03   3.867 0.000111 ***
## AcidIndex          -2.005e-01  9.098e-03 -22.033  < 2e-16 ***
## LabelAppeal-1       3.609e-01  6.286e-02   5.741 9.62e-09 ***
## LabelAppeal0        8.278e-01  6.130e-02  13.503  < 2e-16 ***
## LabelAppeal1        1.292e+00  6.403e-02  20.177  < 2e-16 ***
## LabelAppeal2        1.882e+00  8.436e-02  22.309  < 2e-16 ***
## STARS1              1.363e+00  3.292e-02  41.411  < 2e-16 ***
## STARS2              2.398e+00  3.202e-02  74.910  < 2e-16 ***
## STARS3              2.965e+00  3.707e-02  79.982  < 2e-16 ***
## STARS4              3.650e+00  5.925e-02  61.597  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.306 on 12774 degrees of freedom
## Multiple R-squared:  0.5411, Adjusted R-squared:  0.5404
## F-statistic: 753.1 on 20 and 12774 DF,  p-value: < 2.2e-16
```

Backward Elimination by AIC

Starting with our full model, perform backward elimination by comparing the **AIC** of the models.

```
lm_AIC <- step(lm_full,trace=0)
summary(lm_AIC)
```

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     AcidIndex + LabelAppeal + STARS, data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9600 -0.8616  0.0237  0.8388  6.1758
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.873e+00  4.465e-01   6.436 1.27e-10 ***
## VolatileAcidity    -9.509e-02  1.477e-02  -6.436 1.27e-10 ***
## Chlorides          -1.249e-01  3.645e-02  -3.426 0.000614 ***
## FreeSulfurDioxide   2.394e-04  7.796e-05   3.071 0.002140 **
## TotalSulfurDioxide  2.058e-04  4.992e-05   4.123 3.77e-05 ***
## Density            -8.031e-01  4.358e-01  -1.843 0.065388 .
## pH                 -3.596e-02  1.704e-02  -2.111 0.034798 *
```

```
## Sulphates          -2.937e-02  1.241e-02  -2.367 0.017942 *
## Alcohol             1.208e-02  3.112e-03   3.882 0.000104 ***
## AcidIndex          -1.992e-01  8.940e-03 -22.282  < 2e-16 ***
## LabelAppeal-1       3.605e-01  6.286e-02   5.735 9.96e-09 ***
## LabelAppeal0        8.274e-01  6.130e-02  13.498  < 2e-16 ***
## LabelAppeal1        1.292e+00  6.402e-02  20.173  < 2e-16 ***
## LabelAppeal2        1.882e+00  8.435e-02  22.312  < 2e-16 ***
## STARS1              1.364e+00  3.292e-02  41.428  < 2e-16 ***
## STARS2              2.399e+00  3.200e-02  74.976  < 2e-16 ***
## STARS3              2.965e+00  3.706e-02  80.007  < 2e-16 ***
## STARS4              3.651e+00  5.924e-02  61.623  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.306 on 12777 degrees of freedom
## Multiple R-squared:  0.541,  Adjusted R-squared:  0.5404
## F-statistic:   886 on 17 and 12777 DF,  p-value: < 2.2e-16
```

Backward Elimination by BIC

Starting with our full model, perform backward elimination by comparing the **BIC** of the models.

```
lm_BIC <- step(lm_full,trace=0, k=log(nrow(train_df)))
summary(lm_BIC)
```

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide +
##     Alcohol + AcidIndex + LabelAppeal + STARS, data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0103 -0.8631  0.0264  0.8393  6.2004
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.956e+00  1.001e-01  19.540  < 2e-16 ***
## VolatileAcidity     -9.587e-02  1.478e-02  -6.486 9.17e-11 ***
## Chlorides           -1.262e-01  3.646e-02  -3.460 0.000541 ***
## TotalSulfurDioxide   2.074e-04  4.995e-05   4.152 3.32e-05 ***
## Alcohol              1.195e-02  3.113e-03   3.840 0.000124 ***
## AcidIndex           -2.003e-01  8.912e-03 -22.477  < 2e-16 ***
## LabelAppeal-1        3.615e-01  6.290e-02   5.748 9.26e-09 ***
## LabelAppeal0         8.301e-01  6.134e-02  13.534  < 2e-16 ***
## LabelAppeal1         1.294e+00  6.407e-02  20.198  < 2e-16 ***
## LabelAppeal2         1.882e+00  8.440e-02  22.294  < 2e-16 ***
## STARS1               1.368e+00  3.293e-02  41.544  < 2e-16 ***
## STARS2               2.404e+00  3.201e-02  75.117  < 2e-16 ***
## STARS3               2.971e+00  3.706e-02  80.167  < 2e-16 ***
## STARS4               3.653e+00  5.927e-02  61.639  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.307 on 12781 degrees of freedom
## Multiple R-squared:  0.5402, Adjusted R-squared:  0.5398
## F-statistic:  1155 on 13 and 12781 DF,  p-value: < 2.2e-16
```

Model Coefficients Comparison

Now, let's compare the result of our Poisson, Negative Binomial, and Linear models

```
poisson_full_coef <- data.frame(poisson_full=poisson_full$coefficients)
poisson_AIC_coef <- data.frame(poisson_AIC=round(poisson_AIC$coefficients,4))
poisson_BIC_coef <- data.frame(poisson_BIC=round(poisson_BIC$coefficients,4))
nb_AIC_coef <- data.frame(nb_AIC=round(nb_AIC$coefficients,4))
nb_BIC_coef <- data.frame(nb_BIC=round(nb_BIC$coefficients,4))
lm_AIC_coef <- data.frame(lm_AIC=round(lm_AIC$coefficients,4))
lm_BIC_coef <- data.frame(lm_BIC=round(lm_BIC$coefficients,4))
```

```
summary_table <- merge(x=poisson_full_coef, y=poisson_AIC_coef,
                       by="row.names", all=TRUE)
summary_table <- merge(x=summary_table, y=poisson_BIC_coef,
                       by.x="Row.names", by.y = "row.names", all=TRUE)
summary_table <- merge(x=summary_table, y=nb_AIC_coef,
                       by.x="Row.names", by.y="row.names", all=TRUE)
summary_table <- merge(x=summary_table, y=nb_BIC_coef,
                       by.x="Row.names", by.y="row.names", all=TRUE)
summary_table <- merge(x=summary_table, y=lm_AIC_coef,
                       by.x="Row.names", by.y="row.names", all=TRUE)
summary_table <- merge(x=summary_table, y=lm_BIC_coef,
                       by.x="Row.names", by.y="row.names", all=TRUE)
summary_table$poisson_full <- NULL
summary_table
```

```
##             Row.names poisson_AIC poisson_BIC  nb_AIC  nb_BIC  lm_AIC  lm_BIC
## 1         (Intercept)      0.4428      0.4337  0.8229  0.7226  2.8732  1.9558
## 2           AcidIndex     -0.0799     -0.0805 -0.1139 -0.1137 -0.1992 -0.2003
## 3             Alcohol      0.0036          NA      NA      NA  0.0121  0.0120
## 4           Chlorides     -0.0398          NA -0.0583      NA -0.1249 -0.1262
## 5          CitricAcid          NA          NA      NA      NA      NA      NA
## 6             Density          NA          NA      NA      NA -0.8031      NA
## 7        FixedAcidity          NA          NA      NA      NA      NA      NA
## 8   FreeSulfurDioxide      0.0001          NA  0.0001      NA  0.0002      NA
## 9       LabelAppeal-1      0.2351      0.2350  0.2209  0.2213  0.3605  0.3615
## 10       LabelAppeal0      0.4254      0.4253  0.3888  0.3892  0.8274  0.8301
## 11       LabelAppeal1      0.5579      0.5569  0.4892  0.4896  1.2916  1.2940
## 12       LabelAppeal2      0.6954      0.6951  0.6298  0.6282  1.8820  1.8816
## 13                 pH     -0.0138          NA -0.0277      NA -0.0360      NA
## 14      ResidualSugar          NA          NA      NA      NA      NA      NA
## 15              STARS1      0.7665      0.7689  0.7585  0.7608  1.3637  1.3680
## 16              STARS2      1.0859      1.0886  1.0899  1.0911  2.3993  2.4042
## 17              STARS3      1.2056      1.2105  1.2191  1.2215  2.9652  2.9712
## 18              STARS4      1.3253      1.3301  1.3518  1.3518  3.6505  3.6534
## 19           Sulphates     -0.0108          NA -0.0183      NA -0.0294      NA
## 20 TotalSulfurDioxide      0.0001      0.0001  0.0001  0.0001  0.0002  0.0002
## 21    VolatileAcidity     -0.0307     -0.0309 -0.0423 -0.0424 -0.0951 -0.0959
```

- Both **STARS** and **LabelAppeal** have postive effect in all models.
- The coefficients of **STARS** and **LabelAppeal** are very close in the poisson and negative binomial models.
- **TotalSulfurDioxide** has positive effect in all models. The coefficients seem small but the scale of **TotalSulfurDioxide** is more than 100 times larger than the scales of most other variables.
- **CitricAcid**, **FixedAcidity**, **ResidualSugar** are not significant in all models.
- **AcidIndex** and **VolatileAcidity** have negative effect in all models.
- **Alcohol** and **FreeSulfurDioxide** have positive or no effect in all models.
- **Chlorides**, **Density**, **pH**, and **Sulphates** have negative or no effect in all models.

As discussed above, the target variable is zero inflated. It should be better fit the data in a Hurdle model or a zero-inflated model.

Hurdle Model

```
mod_hurdle <- hurdle(TARGET~.-FixedAcidity-Density-CitricAcid-ResidualSugar-Chlorides,
                     data=train_df)
summary(mod_hurdle)
```

```
##
## Call:
## hurdle(formula = TARGET ~ . - FixedAcidity - Density - CitricAcid - ResidualSugar -
##     Chlorides, data = train_df)
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.099818 -0.442467 -0.002774  0.395516  4.566516
##
## Count model coefficients (truncated poisson with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         3.626e-01  7.045e-02   5.146 2.66e-07 ***
## VolatileAcidity    -1.057e-02  6.912e-03  -1.529 0.126222
## FreeSulfurDioxide   1.696e-05  3.556e-05   0.477 0.633501
## TotalSulfurDioxide -2.848e-05  2.260e-05  -1.260 0.207632
## pH                  7.395e-03  7.941e-03   0.931 0.351693
## Sulphates           1.532e-03  5.780e-03   0.265 0.790903
## Alcohol             7.332e-03  1.444e-03   5.077 3.84e-07 ***
## AcidIndex          -1.655e-02  4.934e-03  -3.354 0.000796 ***
## LabelAppeal-1       5.392e-01  4.973e-02  10.842  < 2e-16 ***
## LabelAppeal0        8.427e-01  4.881e-02  17.267  < 2e-16 ***
## LabelAppeal1        1.040e+00  4.937e-02  21.071  < 2e-16 ***
## LabelAppeal2        1.201e+00  5.319e-02  22.576  < 2e-16 ***
## STARS1              4.931e-02  2.142e-02   2.302 0.021317 *
## STARS2              1.635e-01  1.997e-02   8.189 2.63e-16 ***
## STARS3              2.545e-01  2.092e-02  12.164  < 2e-16 ***
## STARS4              3.576e-01  2.588e-02  13.819  < 2e-16 ***
## Zero hurdle model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         4.358e+00  2.757e-01  15.808  < 2e-16 ***
## VolatileAcidity    -1.841e-01  3.645e-02  -5.052 4.37e-07 ***
## FreeSulfurDioxide   5.546e-04  1.953e-04   2.841  0.00450 **
```

```
## TotalSulfurDioxide  8.095e-04  1.235e-04   6.555 5.57e-11 ***
## pH                 -1.914e-01  4.192e-02  -4.565 4.99e-06 ***
## Sulphates          -9.504e-02  3.060e-02  -3.106  0.00190 **
## Alcohol            -2.072e-02  7.697e-03  -2.692  0.00710 **
## AcidIndex          -3.898e-01  2.141e-02 -18.206  < 2e-16 ***
## LabelAppeal-1      -4.803e-01  1.371e-01  -3.503  0.00046 ***
## LabelAppeal0       -9.002e-01  1.339e-01  -6.724 1.76e-11 ***
## LabelAppeal1       -1.445e+00  1.434e-01 -10.074  < 2e-16 ***
## LabelAppeal2       -1.814e+00  2.219e-01  -8.175 2.95e-16 ***
## STARS1              1.830e+00  6.140e-02  29.797  < 2e-16 ***
## STARS2              4.266e+00  1.171e-01  36.426  < 2e-16 ***
## STARS3              2.024e+01  3.634e+02   0.056  0.95558
## STARS4              2.039e+01  6.942e+02   0.029  0.97657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 23
## Log-likelihood: -2.03e+04 on 32 Df
```

Zero Inflation Model

```
mod_zeroinfl <- zeroinfl(TARGET~.-FixedAcidity-Density-CitricAcid-ResidualSugar-Chlorides,
                         data=train_df)
summary(mod_zeroinfl)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ . - FixedAcidity - Density - CitricAcid -
##     ResidualSugar - Chlorides, data = train_df)
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.262008 -0.428094  0.001613  0.381012  5.354279
##
## Count model coefficients (poisson with log link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         4.991e-01  6.402e-02   7.796 6.41e-15 ***
## VolatileAcidity    -1.231e-02  6.706e-03  -1.836  0.06631 .
## FreeSulfurDioxide   1.528e-05  3.453e-05   0.443  0.65804
## TotalSulfurDioxide -1.753e-05  2.194e-05  -0.799  0.42435
## pH                  4.774e-03  7.708e-03   0.619  0.53567
## Sulphates           1.619e-03  5.614e-03   0.288  0.77303
## Alcohol             6.907e-03  1.401e-03   4.931 8.18e-07 ***
## AcidIndex          -1.921e-02  4.832e-03  -3.975 7.02e-05 ***
## LabelAppeal-1       4.401e-01  4.134e-02  10.647  < 2e-16 ***
## LabelAppeal0        7.284e-01  4.041e-02  18.024  < 2e-16 ***
## LabelAppeal1        9.185e-01  4.108e-02  22.358  < 2e-16 ***
## LabelAppeal2        1.076e+00  4.559e-02  23.601  < 2e-16 ***
## STARS1              6.121e-02  2.113e-02   2.897  0.00377 **
## STARS2              1.823e-01  1.975e-02   9.229  < 2e-16 ***
## STARS3              2.803e-01  2.068e-02  13.556  < 2e-16 ***
## STARS4              3.785e-01  2.561e-02  14.778  < 2e-16 ***
```

```
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -6.245e+00  4.481e-01 -13.937  < 2e-16 ***
## VolatileAcidity     1.865e-01  4.348e-02   4.289 1.80e-05 ***
## FreeSulfurDioxide  -7.064e-04  2.351e-04  -3.005 0.002659 **
## TotalSulfurDioxide -9.075e-04  1.471e-04  -6.168 6.92e-10 ***
## pH                  2.225e-01  5.007e-02   4.443 8.88e-06 ***
## Sulphates           1.240e-01  3.658e-02   3.390 0.000699 ***
## Alcohol             2.833e-02  9.232e-03   3.068 0.002152 **
## AcidIndex           4.318e-01  2.569e-02  16.810  < 2e-16 ***
## LabelAppeal-1       1.503e+00  3.325e-01   4.520 6.19e-06 ***
## LabelAppeal0        2.262e+00  3.300e-01   6.853 7.25e-12 ***
## LabelAppeal1        2.970e+00  3.355e-01   8.855  < 2e-16 ***
## LabelAppeal2        3.418e+00  3.866e-01   8.841  < 2e-16 ***
## STARS1             -2.089e+00  7.622e-02 -27.406  < 2e-16 ***
## STARS2             -5.747e+00  3.291e-01 -17.462  < 2e-16 ***
## STARS3             -2.024e+01  3.401e+02  -0.060 0.952541
## STARS4             -2.039e+01  6.405e+02  -0.032 0.974601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 38
## Log-likelihood: -2.034e+04 on 32 Df
```

SELECT MODELS

Root Mean Squared Error

Since we have models of different types (Poisson, Negative Binomial, Multiple Linear), it's not appropriate to use R squared, AIC or BIC to compare the models. To comparing the performance, the root mean squared error would be a good standard since it is a measurement of the distance between the predicted values and the actual values.

```
data.frame(poisson_AIC=sqrt(mean(residuals(poisson_AIC, type="response")^2)),
           poisson_BIC=sqrt(mean(residuals(poisson_BIC, type="response")^2)),
           nb_AIC=sqrt(mean(residuals(nb_AIC, type="response")^2)),
           nb_BIC=sqrt(mean(residuals(nb_BIC, type="response")^2)),
           lm_AIC=sqrt(mean(residuals(lm_AIC, type="response")^2)),
           lm_BIC=sqrt(mean(residuals(lm_BIC, type="response")^2)),
           mod_hurdle=sqrt(mean(residuals(mod_hurdle, type="response")^2)),
           mod_zeroinfl=sqrt(mean(residuals(mod_zeroinfl, type="response")^2)))
```

```
##   poisson_AIC poisson_BIC   nb_AIC   nb_BIC   lm_AIC   lm_BIC mod_hurdle
## 1    1.300983    1.302231 1.327142 1.324879 1.305009 1.306158   1.262333
##   mod_zeroinfl
## 1     1.263992
```

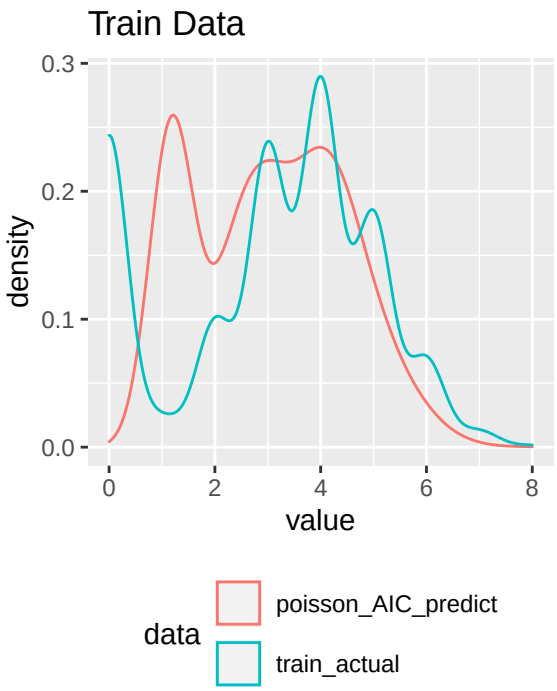
From the RMSE of all models, the hurdle model has the best performance. This is expected since hurdle model is designed for zero-inflated data.

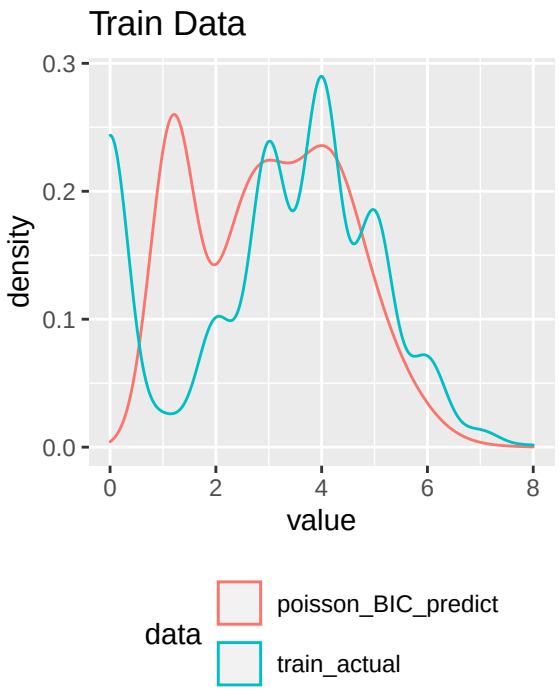
Distribution of Predicted Values (train data)

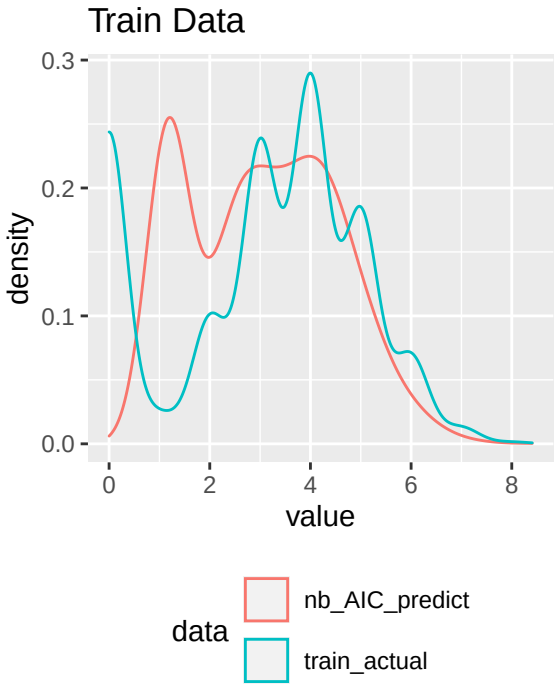
We can also look at the distributions of the model predictions of the training data.

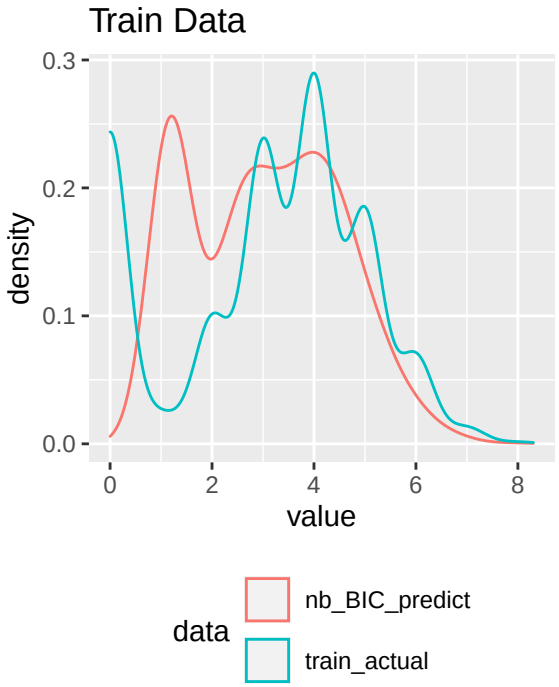
```
train_actual <- train_df$TARGET
poisson_AIC_predict <- predict(poisson_AIC,type="response")
poisson_BIC_predict <- predict(poisson_BIC,type="response")
nb_AIC_predict <- predict(nb_AIC,type="response")
nb_BIC_predict <- predict(nb_BIC,type="response")
lm_AIC_predict <- predict(lm_AIC,type="response")
lm_BIC_predict <- predict(lm_BIC,type="response")
mod_hurdle_predict <- predict(mod_hurdle,type="response")
mod_zeroinfl_predict <- predict(mod_zeroinfl,type="response")

dist_df <- data.frame(rbind(
      cbind(train_actual,"train_actual"),
      cbind(poisson_AIC_predict,"poisson_AIC_predict"),
      cbind(poisson_BIC_predict,"poisson_BIC_predict"),
      cbind(nb_AIC_predict,"nb_AIC_predict"),
      cbind(nb_BIC_predict,"nb_BIC_predict"),
      cbind(lm_AIC_predict,"lm_AIC_predict"),
      cbind(lm_BIC_predict,"lm_BIC_predict"),
      cbind(mod_hurdle_predict,"mod_hurdle_predict"),
      cbind(mod_zeroinfl_predict,"mod_zeroinfl_predict")
      ),stringsAsFactors=FALSE)
colnames(dist_df) <- c("value","data")
dist_df$value <- as.numeric(dist_df$value)
```

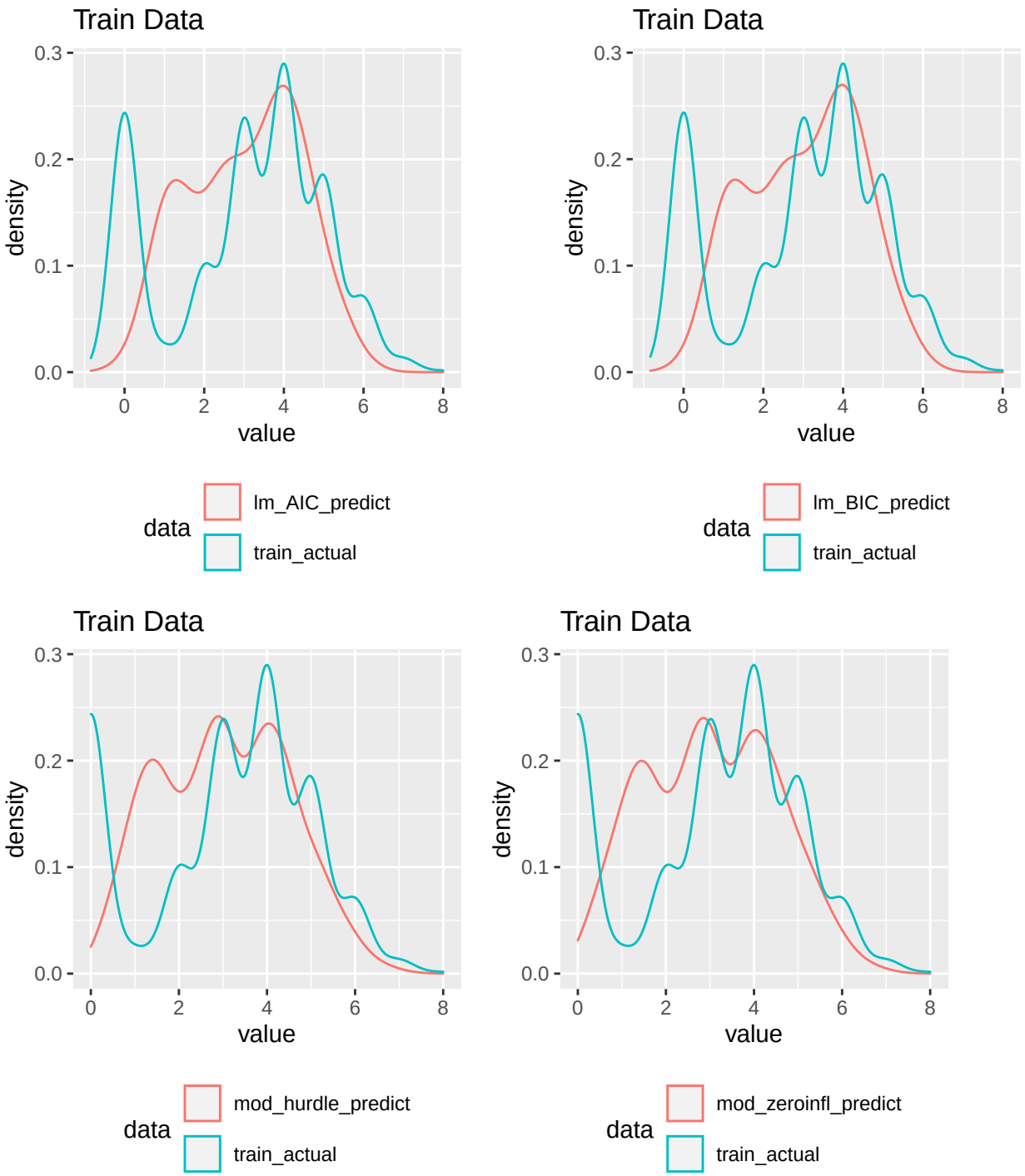
```
models <- unique(dist_df$data)[-1]
for (model in models) {
    plot<-ggplot(dist_df[dist_df$data=="train_actual" | dist_df$data==model,],
           aes(x=value, color=data))+ggtitle("Train Data")+geom_density(bw=0.35)+
           theme(legend.position="bottom")+
           guides(color=guide_legend(nrow=2, byrow=TRUE))
    print(plot)
}
```


Train Data
density
value
data
poisson_AIC_predict
train_actual


Train Data
density
value
data
poisson_BIC_predict
train_actual


Train Data
density
value
data
nb_AIC_predict
train_actual


Train Data
density
value
data
nb_BIC_predict
train_actual



The predictions of Poisson models and Negative Binomial models have similar distribution. They do well in modeling the peak near 0. However, the peak is at 1, there is nearly no prediction of target = 0.

The linear models are the worst, they do even predict some negative values since it is not bounded.

The hurdle model and the zero-inflated model do not model the peak near 0 as well as the Poisson models or Negative Binomial models do. However, they successfully predict some cases with target = 0. Moreover, the models are fitting the data better at target greater or equal to 3.

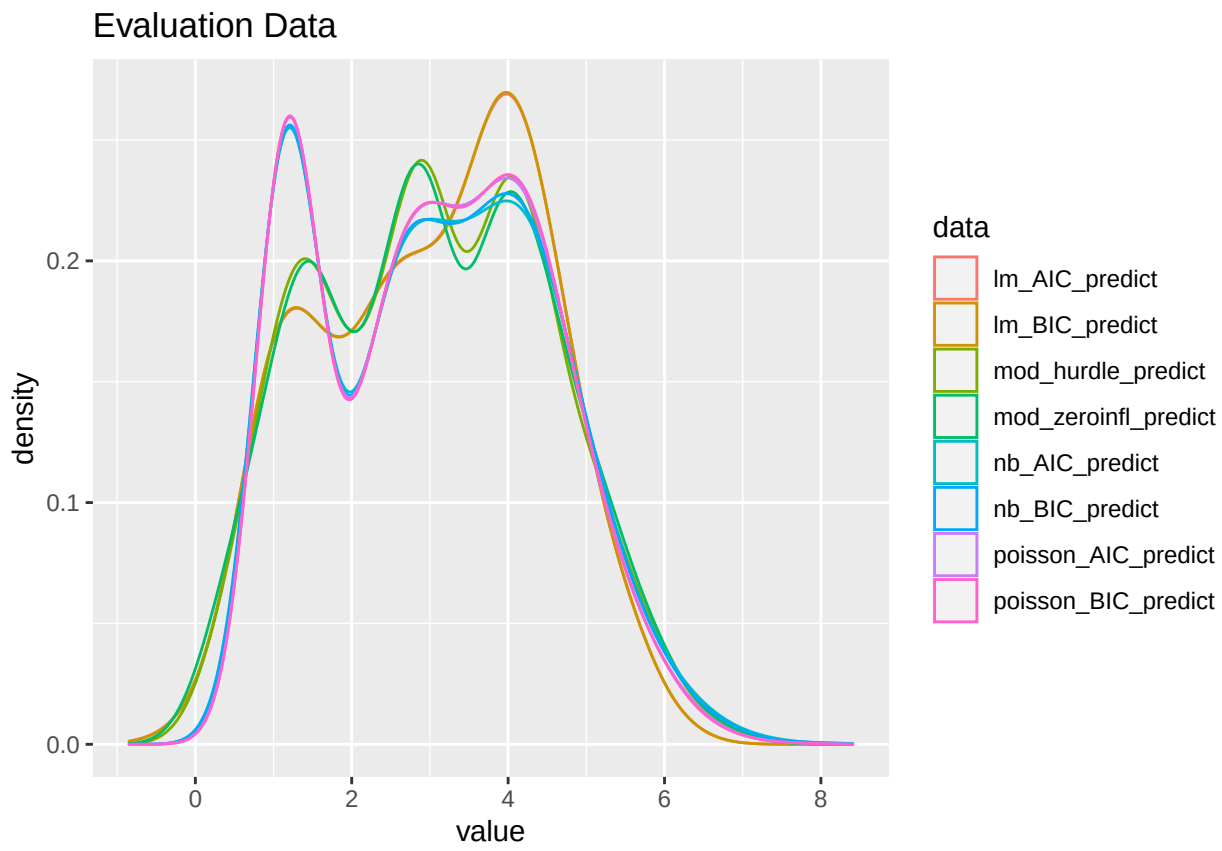
This confirms our findings above that the hurdle model and the zero-inflated model suit our data better.

Distribution of Predicted Values (test data)

```
#temporary exclude LabelAppeal and STARS in our imputation
LabelAppeal <- test_df$LabelAppeal
STARS <- test_df$STARS

test_df$TARGET <- NULL
test_df$LabelAppeal <- NULL
test_df$STARS <- NULL

#save the imputation result
test_df <- mice.reuse(mickey, test_df, maxit = 5, printFlag = FALSE, seed = 2022)[[1]]

#Add TARGET, LabelAppeal, and STARS back to our dataframe
test_df$LabelAppeal <- LabelAppeal
test_df$STARS <- STARS

LabelAppeal <- NULL
STARS <- NULL

#data transformation
STARS_Y <- !is.na(test_df$STARS)
test_df$STARS[!STARS_Y] <- 0
test_df$STARS <- as.factor(test_df$STARS)
test_df$LabelAppeal <- as.factor(test_df$LabelAppeal)
```

The following are the distributions of the predicted values of our models using the evaluation data.

```
poisson_AIC_predict <- predict(poisson_AIC,type="response",data=test_df)
poisson_BIC_predict <- predict(poisson_BIC,type="response",data=test_df)
nb_AIC_predict <- predict(nb_AIC,type="response",data=test_df)
nb_BIC_predict <- predict(nb_BIC,type="response",data=test_df)
lm_AIC_predict <- predict(lm_AIC,type="response",data=test_df)
lm_BIC_predict <- predict(lm_BIC,type="response",data=test_df)
mod_hurdle_predict <- predict(mod_hurdle,type="response",data=test_df)
mod_zeroinfl_predict <- predict(mod_zeroinfl,type="response",data=test_df)

dist_df <- data.frame(rbind(
      cbind(poisson_AIC_predict,"poisson_AIC_predict"),
      cbind(poisson_BIC_predict,"poisson_BIC_predict"),
      cbind(nb_AIC_predict,"nb_AIC_predict"),
      cbind(nb_BIC_predict,"nb_BIC_predict"),
      cbind(lm_AIC_predict,"lm_AIC_predict"),
      cbind(lm_BIC_predict,"lm_BIC_predict"),
      cbind(mod_hurdle_predict,"mod_hurdle_predict"),
      cbind(mod_zeroinfl_predict,"mod_zeroinfl_predict")
      ),stringsAsFactors=FALSE)
colnames(dist_df) <- c("value","data")
dist_df$value <- as.numeric(dist_df$value)

ggplot(dist_df, aes(x=value, color=data))+
  ggtitle("Evaluation Data")+geom_density(bw=0.35)
```



The distributions are very close to our predictions using the training data. The predictions produce plausible results.