DATA621 HW5

Misha Kollontai

12/9/2020

Overview

In this homework assignment, you 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.

Your 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).

```
##   ï..INDEX TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar
## 1        1      3          3.2           1.160      -0.98          54.2
## 2        2      3          4.5           0.160      -0.81          26.1
## 3        4      5          7.1           2.640      -0.88          14.8
## 4        5      3          5.7           0.385       0.04          18.8
## 5        6      4          8.0           0.330      -1.26           9.4
## 6        7      0         11.3           0.320       0.59           2.2
##   Chlorides FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates
## 1    -0.567                NA                268 0.99280 3.33     -0.59
## 2    -0.425                15               -327 1.02792 3.38      0.70
## 3     0.037               214                142 0.99518 3.12      0.48
## 4    -0.425                22                115 0.99640 2.24      1.83
## 5        NA              -167                108 0.99457 3.12      1.77
## 6     0.556               -37                 15 0.99940 3.20      1.29
##   Alcohol LabelAppeal AcidIndex STARS
## 1     9.9           0         8     2
## 2      NA          -1         7     3
## 3    22.0          -1         8     3
## 4     6.2          -1         6     1
## 5    13.7           0         9     2
## 6    15.4           0        11    NA
```

Data Exploration

Let's calculate summary statistics and generate a box plots for further review.

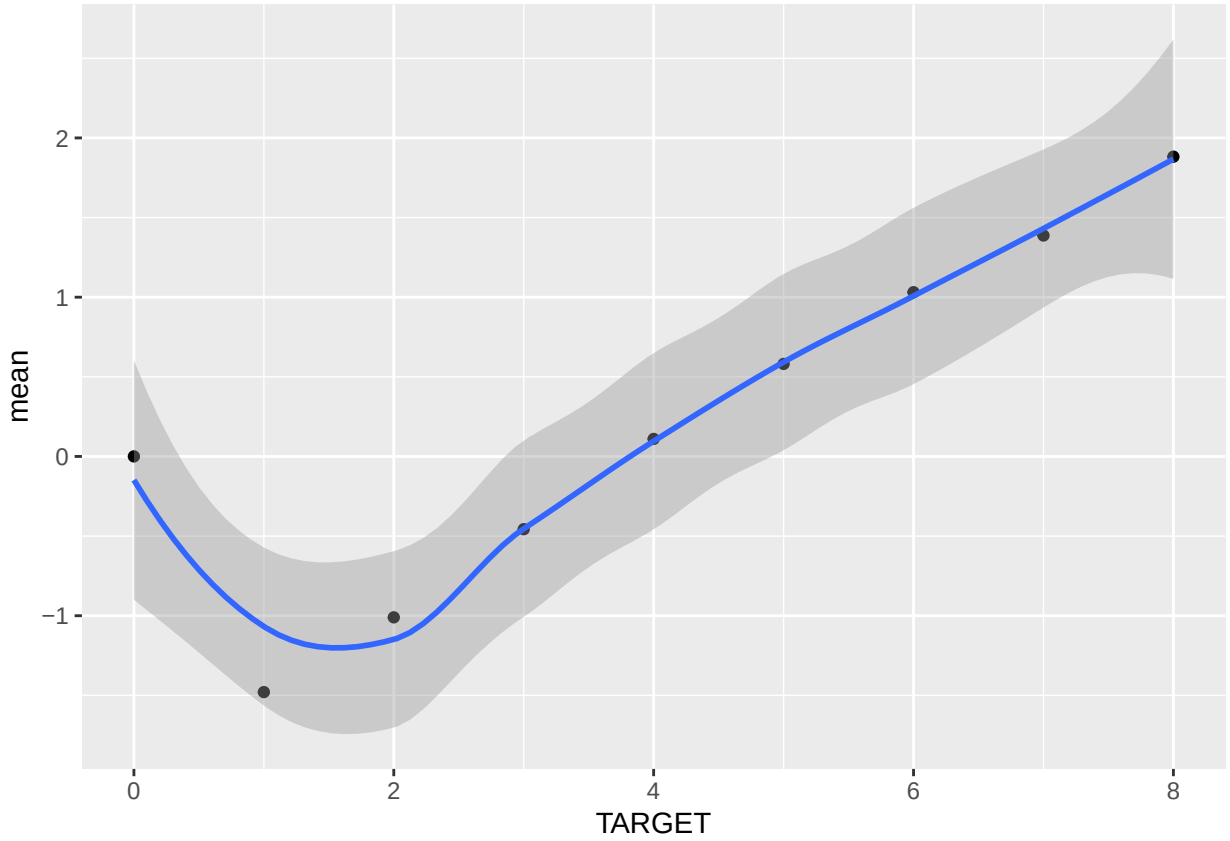
```
##     ï..INDEX         TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   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
##     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
```

There appear to be a significant amount of missing (NA) data. In order to see what effect each of our variables may have on our predictive model, let's take a look and see how the variables relate to the number of cases sold (our target variable). We will look at a spattering of the available variables.

Label Appeal

In order to evaluate the impact of the label appel, let's take a look at how many cases each "score" of label appeal sold per wine. Conventional knowledge suggests that more appealing labels will sell more cases.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

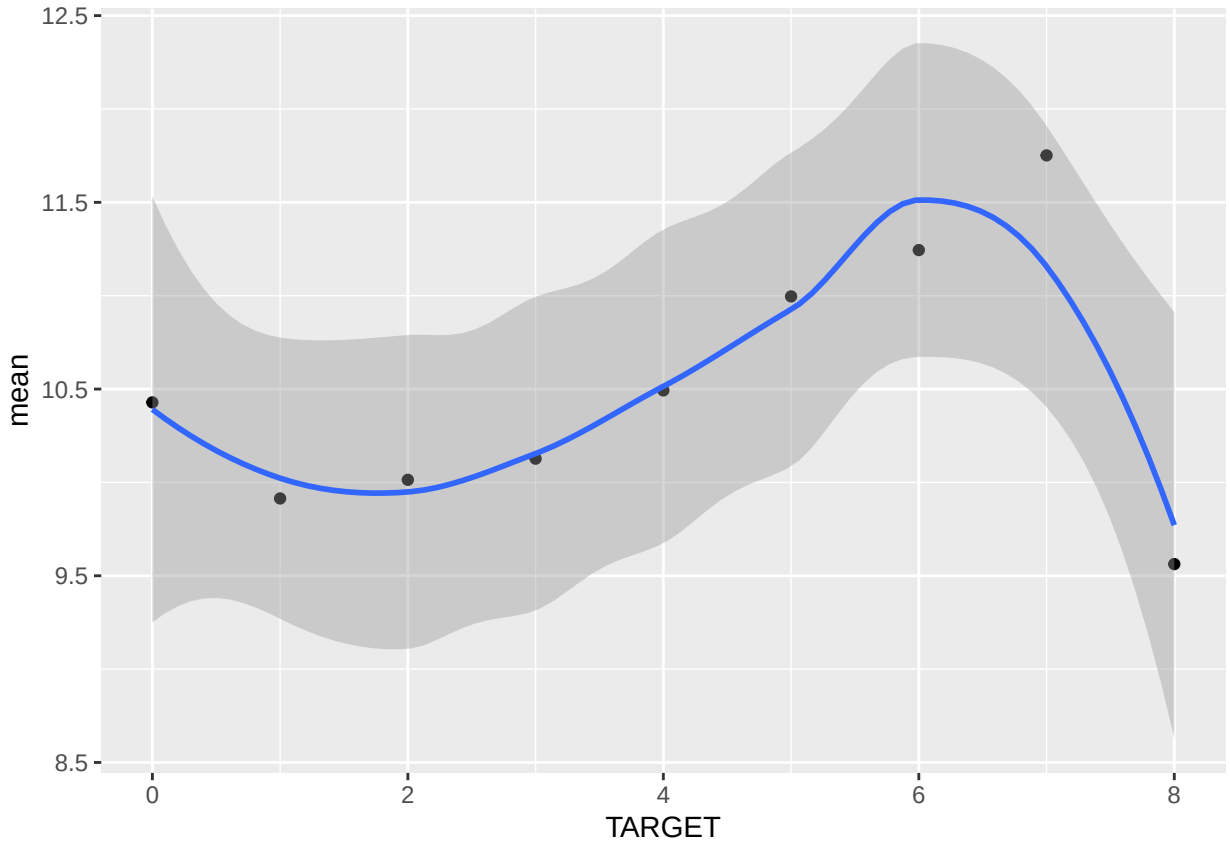


We can clearly see that the more appealing bottles sell more cases on average. This looks to be a very strong predictor of sales numbers.

Alcohol

Alcohol content is another variable we have at our disposal. Some peopl emay be looking for wine with a lower alcohol content, while others may prefer a stronger wine. Let's take a look at our data and see what trends present themselves. We can look at the average alcohol content for wines that sold a particular number of cases to identify possible relationships. We also know from above that we have over 650 NAs in the data that will need to be accounted for.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



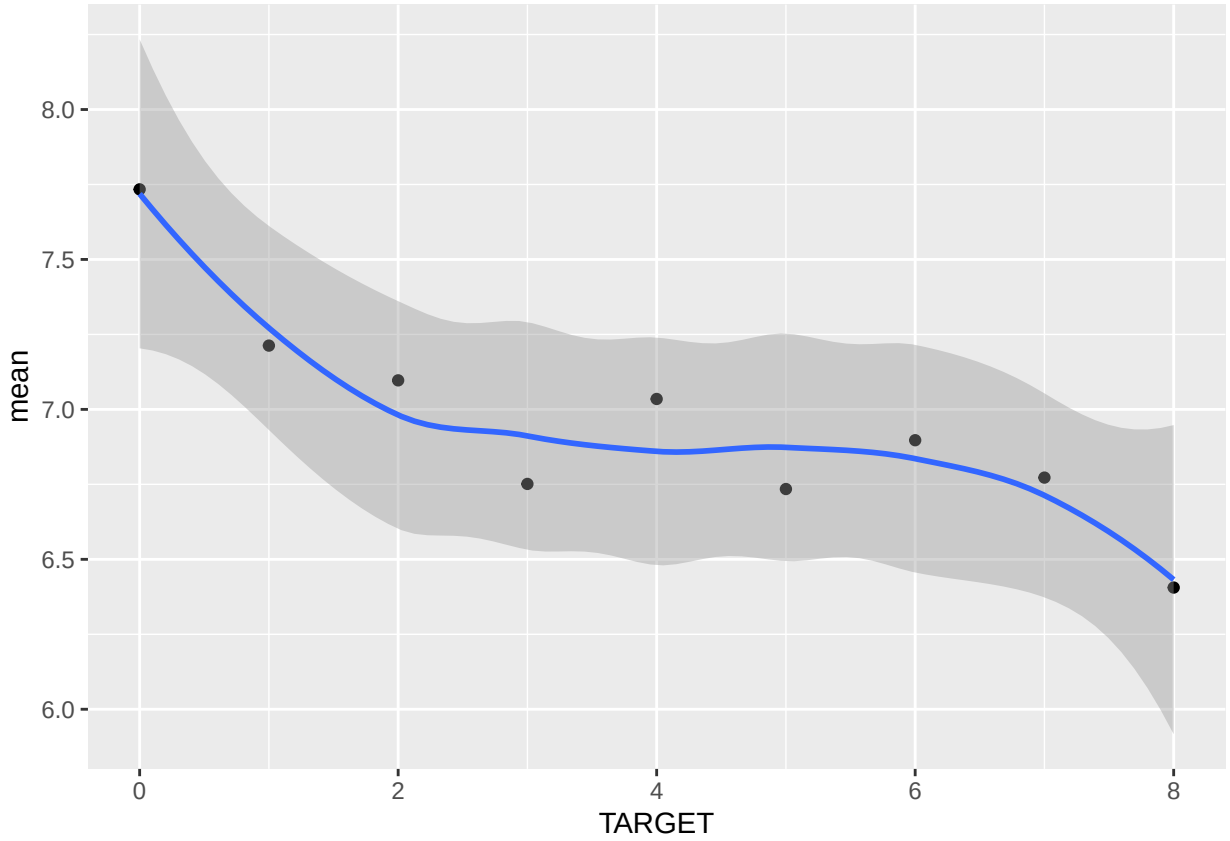
Here we can see that as the number of cases sold increases, so does the average alcool content of the wines, though it must be noted that there is a sharp dropoff at 8 cases sold to the lowest average alcohol content in the set - perhaps errors in the data coupled with a small sample size?

Acidity

Now let's take a look into some of the acidity variables.

Fixed Acidity

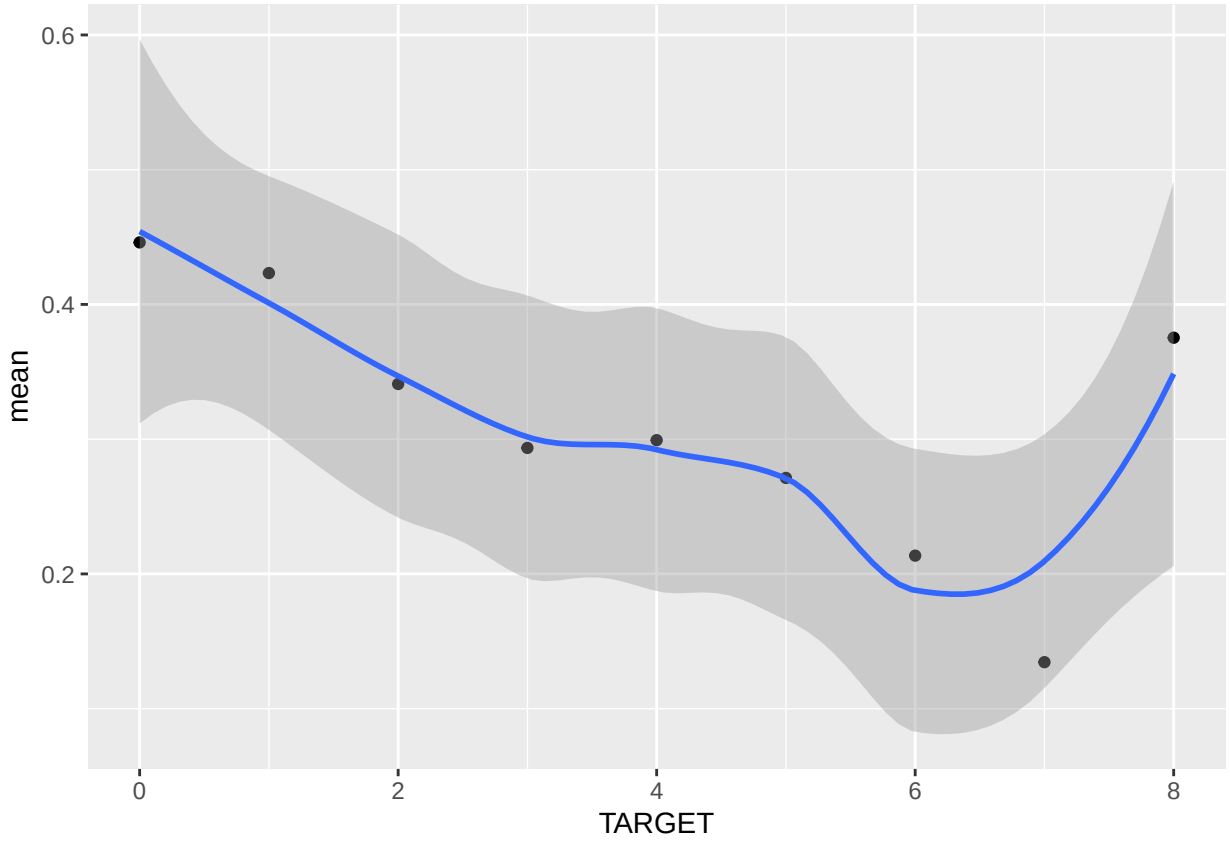
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Lower fixed acidity seems to correlate with higher cases sold.

Volatile Acidity

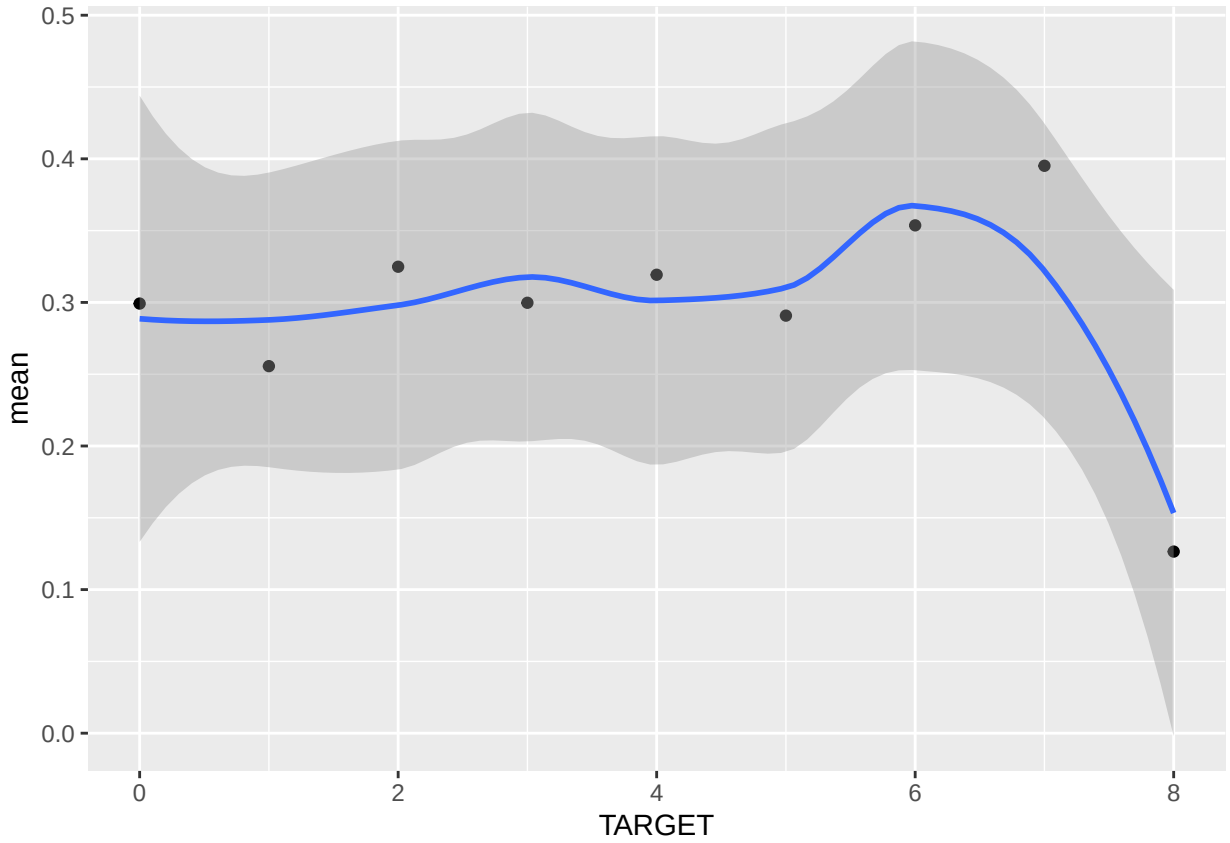
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Volatile acidity seems to follow a similar trend as the fixed acidity, though for some reason there is a spike in sales again at 8 cases. Perhaps the low sample size for 8 cases sold is skewing our data.

Citric Acidity

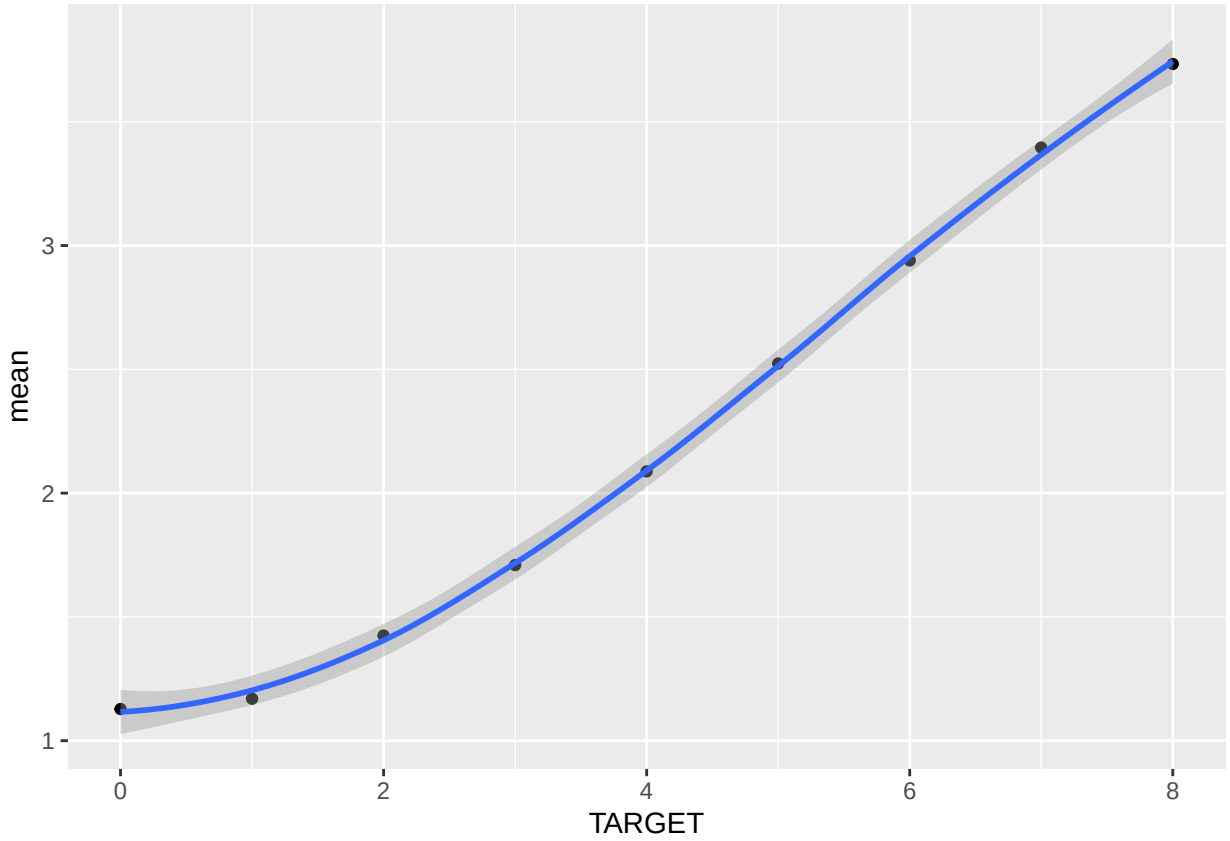
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Stars

The number of stars assigned to a bottle is likely to influence the sales numbers associated with it.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



This relationship is disturbingly obvious - to the point where it is difficult to believe the two variables are truly independent.

Data Preparation

Imputation

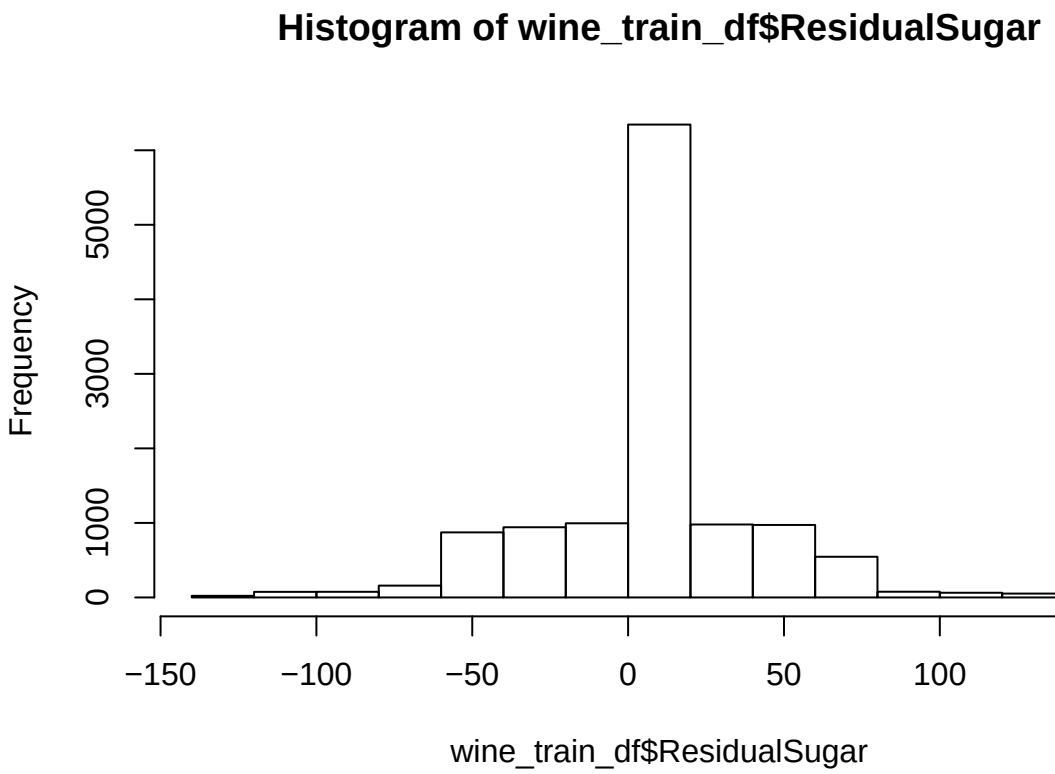
What columns are missing data?

```
##           ï..INDEX             TARGET       FixedAcidity
##                  0                  0                  0
##    VolatileAcidity        CitricAcid      ResidualSugar
##                  0                  0                616
##          Chlorides  FreeSulfurDioxide TotalSulfurDioxide
##                638                647                682
##            Density                 pH          Sulphates
##                  0                395               1210
##            Alcohol       LabelAppeal          AcidIndex
##                653                  0                  0
##              STARS
##               3359
```

As we can see, there is a large amount of data missing for one of our potentially strongest predictors - STARS. All in all, 8 of the 14 predictor variables have missing data. For the variable `STARS` there is no information

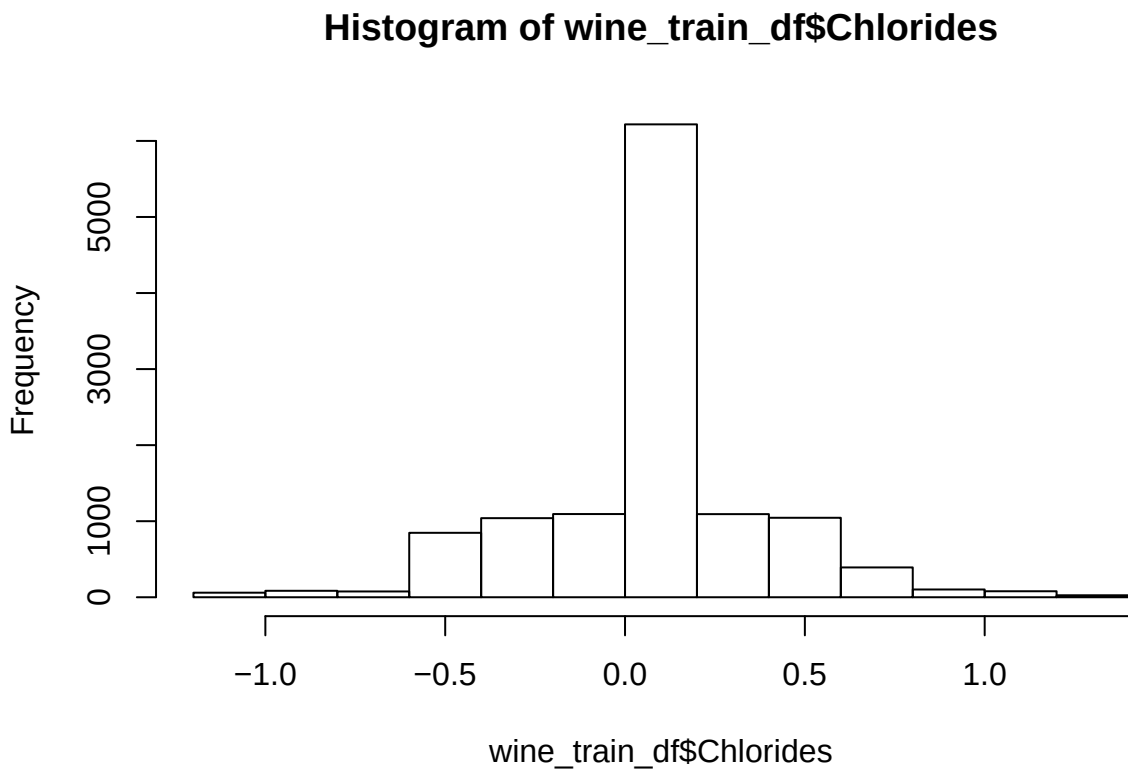
for more than 25% of the entries. Chances are we will need to come up with some strong predictors for when the STARS rating isn't available.

Let's take a look at each variable with missing data in turn to determine the best path forward for each.



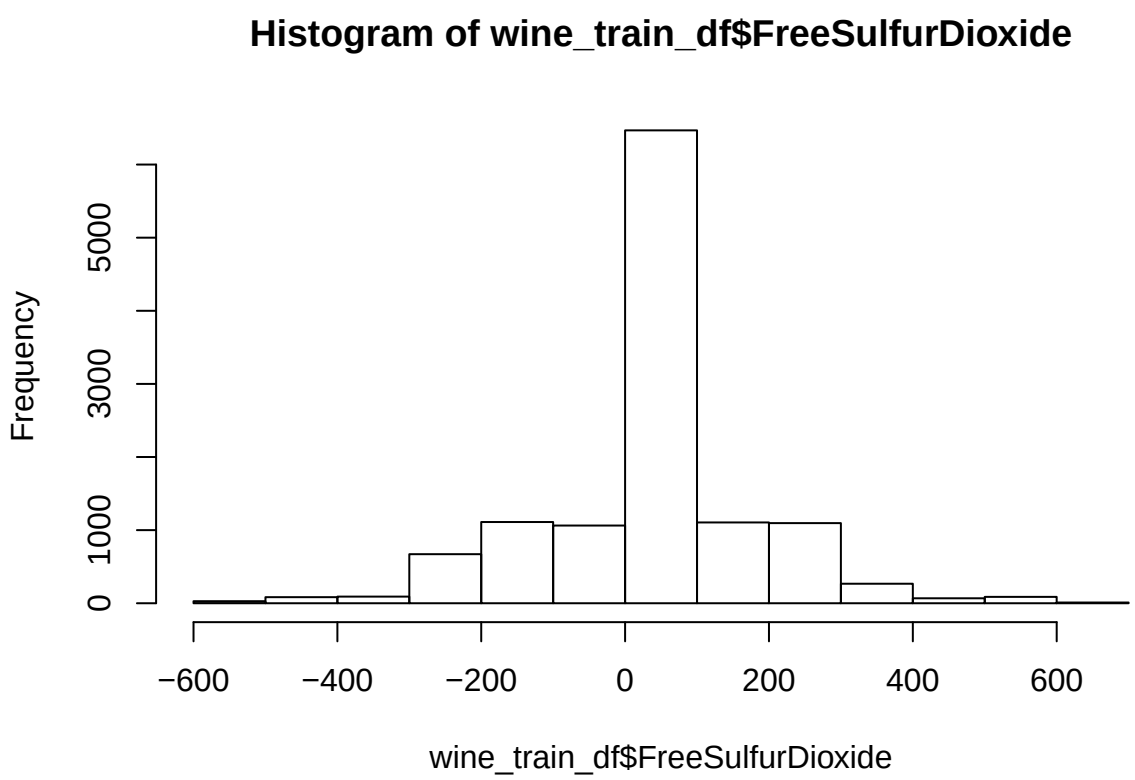
Residual Sugar Imputation

Here we see an intereseting picture - a somewhat normal distribution cenetered at 0, meaning a large portion of the data is negative. There is no such thing as a negative Residual Sugar levl, since it is measured in grams per Liter. One possibility is that the data reflects the delta from the mean or median of wines. If this assumption is correct, then assigning a value of 0 to NAs for this variable is a logcal way to go.

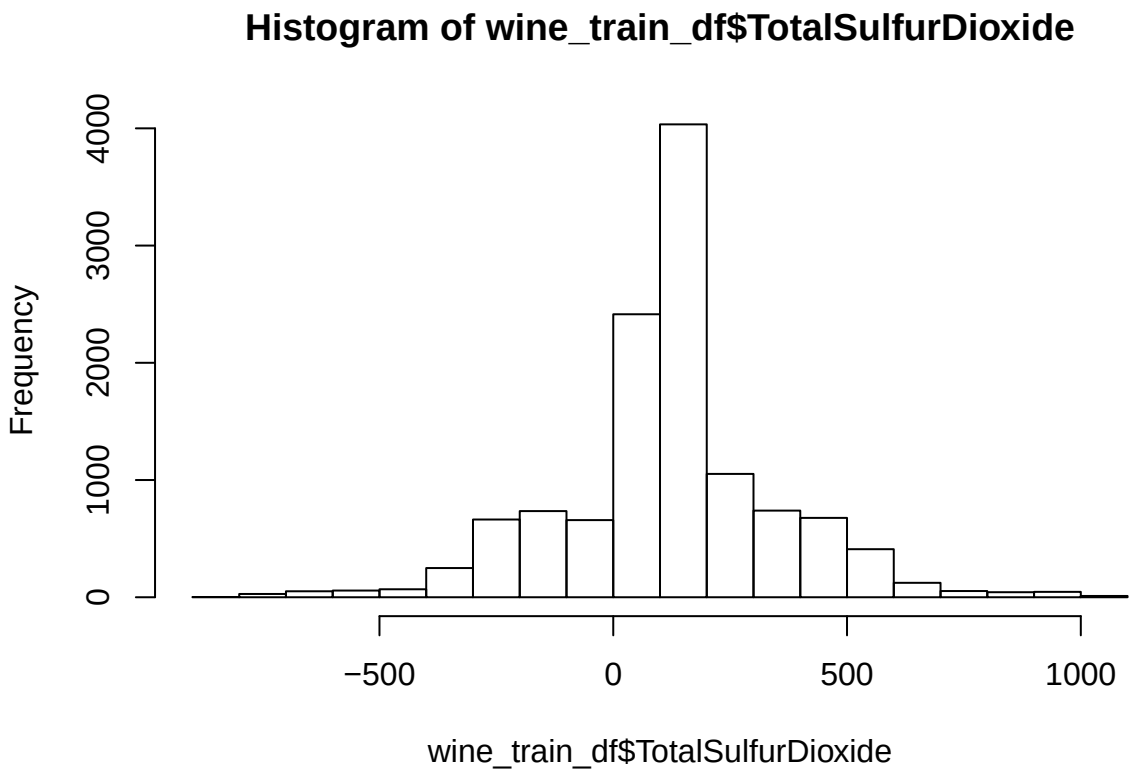


Chlorides Imputation

The data for Chlorides follows a similar distribution to that of the Residal Sugars. We will impute in the same way.

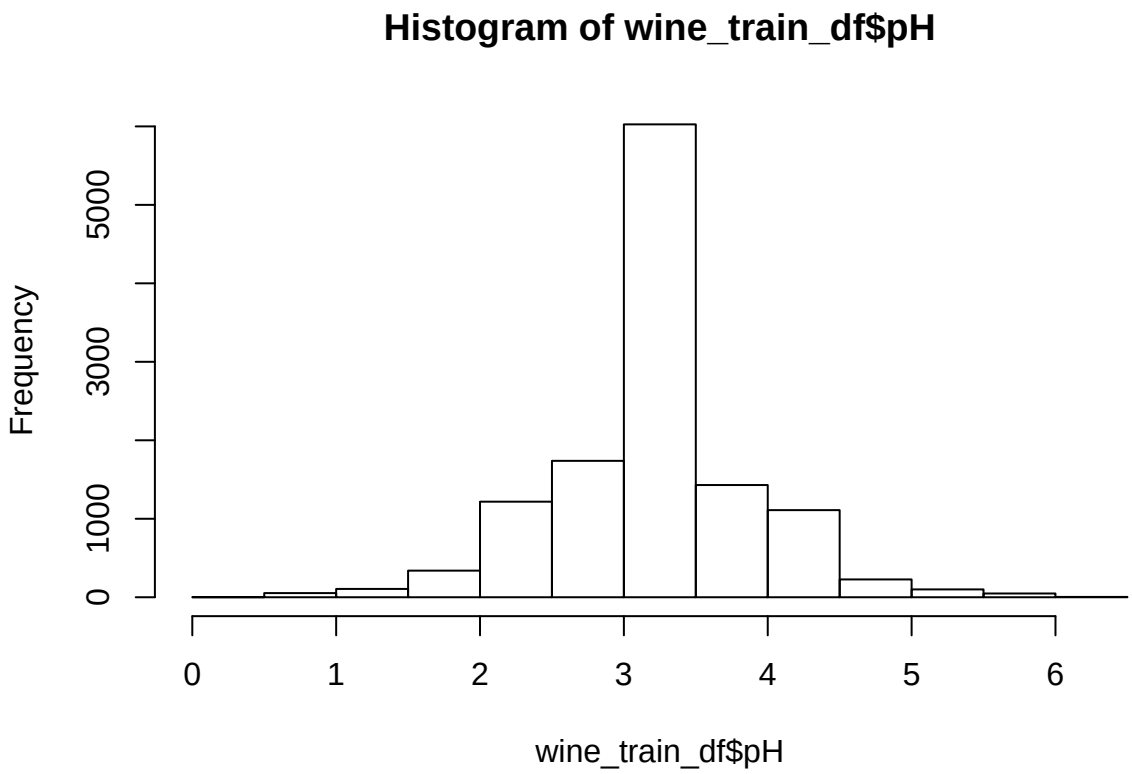


Free Sulfur Dioxide



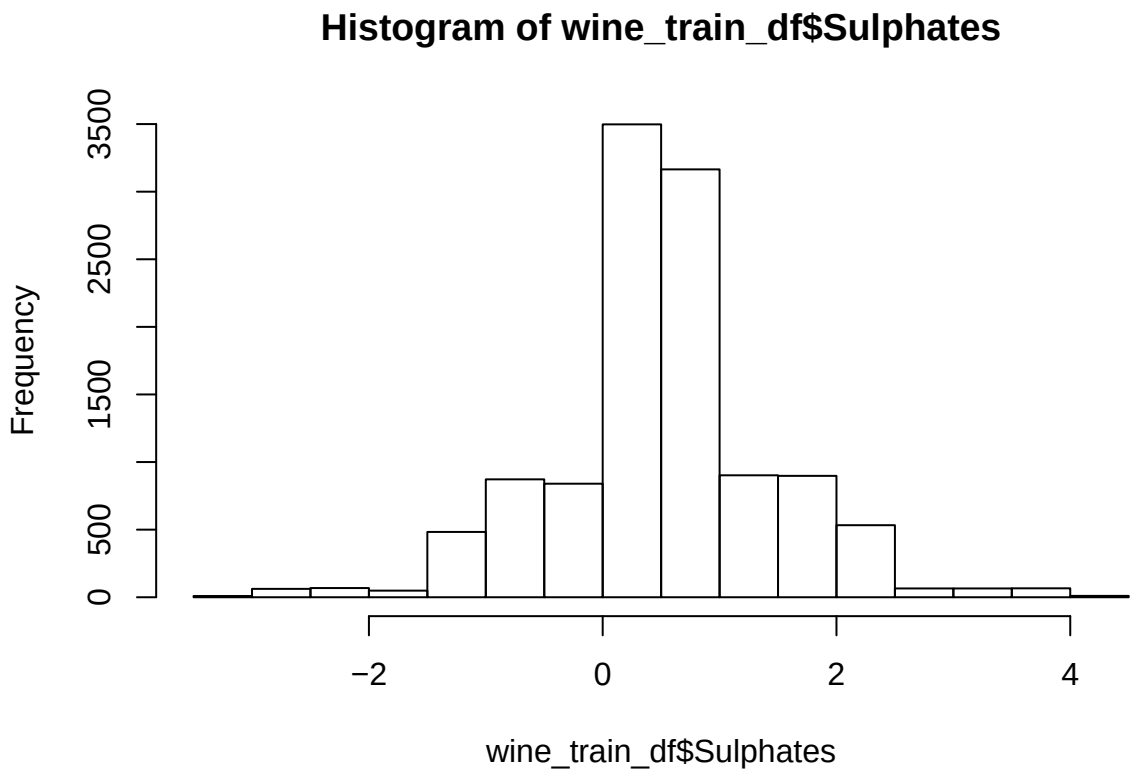
Total Sulfur Dioxide

In this case the data doesn't appear to be centered at 0, so we will impute with the median. The fact that negative "Total Sulfur Dioxide" is being reported does raise concerns about the accuracy of the provided data.



pH

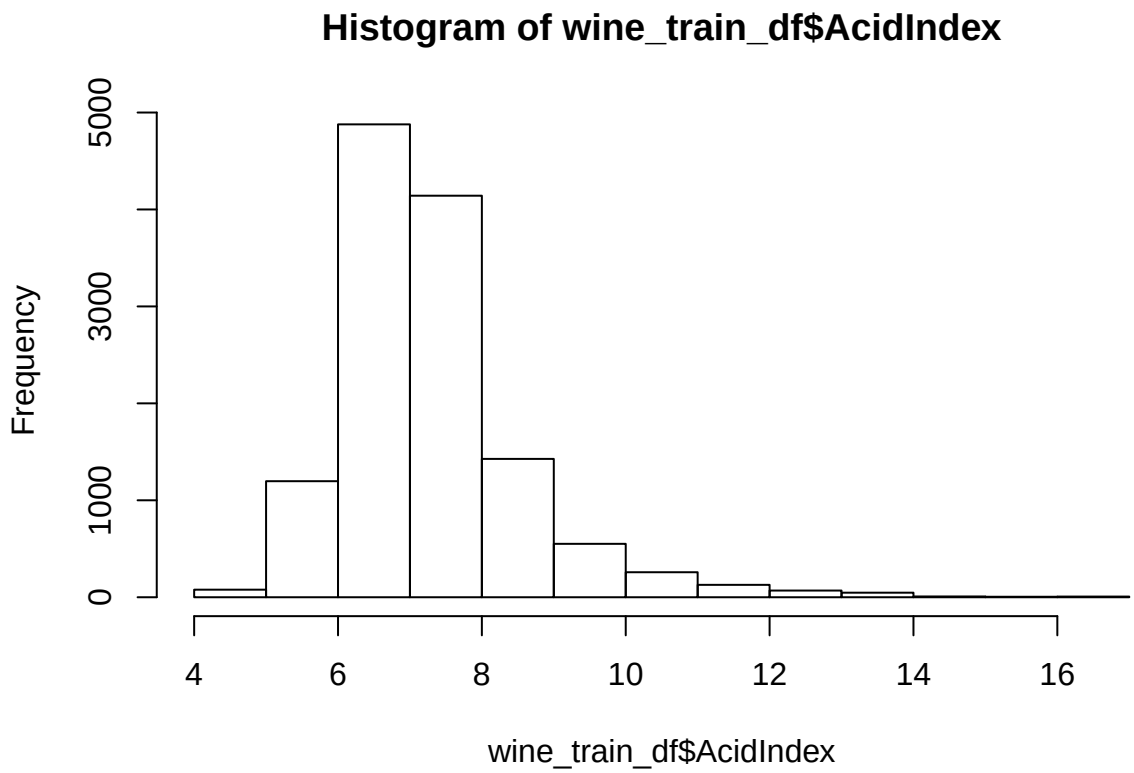
Similar to the total sulfur dioxide, the data is centered around a value other than 0 - we will impute accordingly.


Histogram of wine_train_df$Sulphates
Frequency
0
500
1500
2500
3500
−2
0
2
4
wine_train_df$Sulphates

Sulphates

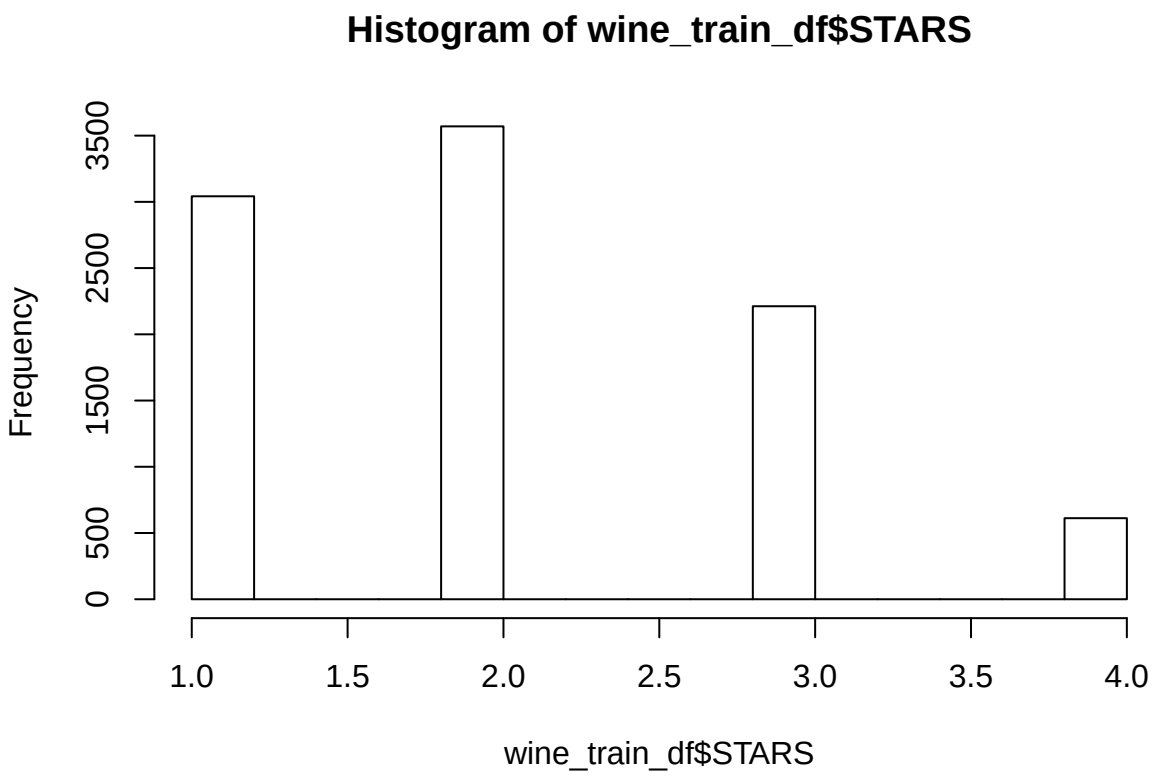

Histogram of wine_train_df$Alcohol
Frequency
0
500
1500
2500
−5
0
5
10
15
20
25
wine_train_df$Alcohol

Alcohol



Acid Index

This veriable isn't missing data, but it is interesting to see that is skewed right, with a majority of the wines having a lower index between 5 and 10. The hghest Acidity Index present in the data is 17



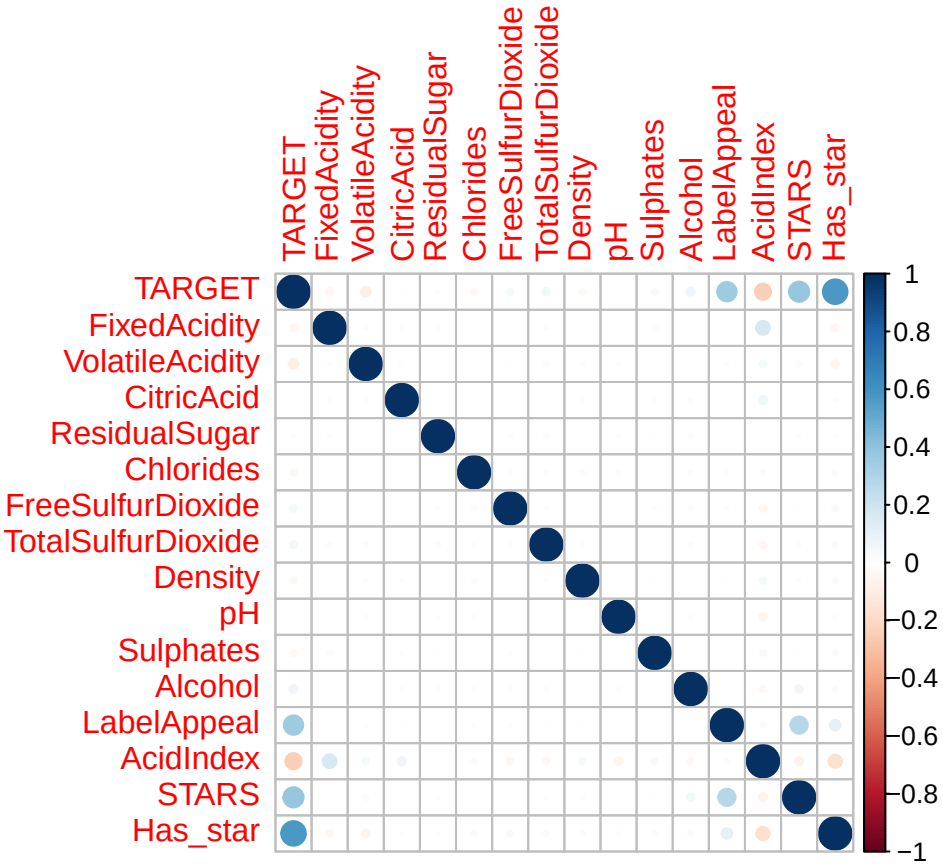
Stars

This is a case of a variable where there are discrete values available and the data is not normally distributed. This makes it difficult to backfill missing data. What we will do is create a new column that tracks whether or not a STARS variable was present. If it was not, we will assign the mean value to the STARS column.

```
##           ï..INDEX             TARGET       FixedAcidity
##                  0                  0                  0
##    VolatileAcidity         CitricAcid      ResidualSugar
##                  0                  0                  0
##          Chlorides  FreeSulfurDioxide TotalSulfurDioxide
##                  0                  0                  0
##            Density                 pH          Sulphates
##                  0                  0                  0
##            Alcohol        LabelAppeal          AcidIndex
##                  0                  0                  0
##              STARS           Has_star
##                  0                  0
```

Correlation

Do our variables correlate to the target variable at all?



The variables appear fairly independent of one another with only LabelAppeal, AcidIndex, STARS and whether it has a STAR or not showing a strong correlation with the TARGET variable. Interestingly, LabelAppeal and STARS show correlation, suggesting that one of those variables may influence the other.

Transforming Data

We created a new variable above, tracking whether or not there was STARS data available for each wine. If any other quirks in the data present themselves, we will create new variables.

Build Models

To start, let's create some Poisson Regression models to

Poisson Regressions

Model 1 - First Poisson Regression

```
##
## Call:
## glm(formula = TARGET ~ ., family = "poisson", data = wine_train_data)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.1719  -0.6515   0.0074   0.4528   3.7687
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          7.424e-01  1.966e-01   3.776 0.000159 ***
## FixedAcidity         7.837e-06  8.195e-04   0.010 0.992370
## VolatileAcidity     -3.099e-02  6.519e-03  -4.754 1.99e-06 ***
## CitricAcid           5.604e-03  5.894e-03   0.951 0.341684
## ResidualSugar        6.407e-05  1.546e-04   0.414 0.678668
## Chlorides           -3.675e-02  1.647e-02  -2.232 0.025630 *
## FreeSulfurDioxide    9.665e-05  3.503e-05   2.759 0.005797 **
## TotalSulfurDioxide   8.032e-05  2.275e-05   3.530 0.000415 ***
## Density             -2.776e-01  1.918e-01  -1.447 0.147891
## pH                  -1.304e-02  7.646e-03  -1.705 0.088207 .
## Sulphates           -1.079e-02  5.678e-03  -1.901 0.057313 .
## Alcohol              3.426e-03  1.408e-03   2.433 0.014961 *
## LabelAppeal          1.589e-01  6.128e-03  25.934  < 2e-16 ***
## AcidIndex           -8.079e-02  4.570e-03 -17.676  < 2e-16 ***
## STARS                1.878e-01  6.092e-03  30.828  < 2e-16 ***
## Has_star             1.031e+00  1.697e-02  60.748  < 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: 13767  on 12779  degrees of freedom
## AIC: 45741
##
## Number of Fisher Scoring iterations: 6
```

As expected, we see that the strongest predictors are the LabelAppeal, AcidIndex, STARS and Has_Star variables. AcidIndex is the only of these 4 that has a negative coefficient, suggesting that a lower acid index is appealing to consumers.

For our second model, let's reduce the number of less significant variables and trim the model somewhat by stepwise removing variables that have insignificant p-values.

Model 2 - Trimmed Poisson Regression

```
##
## Call:
## glm(formula = TARGET ~ . - FixedAcidity - ResidualSugar - CitricAcid -
##     Density, family = poisson, data = wine_train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1761  -0.6490   0.0074   0.4543   3.7599
##
## Coefficients:
```

```
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          4.675e-01  5.264e-02   8.881  < 2e-16 ***
## VolatileAcidity     -3.122e-02  6.518e-03  -4.789 1.67e-06 ***
## Chlorides           -3.740e-02  1.646e-02  -2.272 0.023092 *
## FreeSulfurDioxide    9.668e-05  3.502e-05   2.761 0.005763 **
## TotalSulfurDioxide   7.980e-05  2.274e-05   3.510 0.000448 ***
## pH                  -1.300e-02  7.645e-03  -1.700 0.089152 .
## Sulphates           -1.080e-02  5.675e-03  -1.904 0.056936 .
## Alcohol              3.456e-03  1.407e-03   2.456 0.014051 *
## LabelAppeal          1.590e-01  6.127e-03  25.945  < 2e-16 ***
## AcidIndex           -8.077e-02  4.512e-03 -17.903  < 2e-16 ***
## STARS                1.879e-01  6.091e-03  30.857  < 2e-16 ***
## Has_star             1.032e+00  1.697e-02  60.791  < 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: 13771  on 12783  degrees of freedom
## AIC: 45737
##
## Number of Fisher Scoring iterations: 6
```

Negative Binomial Regressions

Model 3 - First Negative Binomial Regression

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = wine_train_data, init.theta = 40611.66211,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1718  -0.6515   0.0074   0.4528   3.7686
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          7.424e-01  1.966e-01   3.776 0.000159 ***
## FixedAcidity         7.829e-06  8.196e-04   0.010 0.992379
## VolatileAcidity     -3.099e-02  6.520e-03  -4.754 1.99e-06 ***
## CitricAcid           5.604e-03  5.894e-03   0.951 0.341696
## ResidualSugar        6.407e-05  1.547e-04   0.414 0.678647
## Chlorides           -3.675e-02  1.647e-02  -2.232 0.025632 *
## FreeSulfurDioxide    9.665e-05  3.503e-05   2.759 0.005798 **
## TotalSulfurDioxide   8.032e-05  2.275e-05   3.530 0.000415 ***
## Density             -2.776e-01  1.919e-01  -1.447 0.147902
## pH                  -1.304e-02  7.646e-03  -1.705 0.088199 .
## Sulphates           -1.079e-02  5.678e-03  -1.901 0.057314 .
## Alcohol              3.426e-03  1.408e-03   2.433 0.014968 *
## LabelAppeal          1.589e-01  6.128e-03  25.933  < 2e-16 ***
## AcidIndex           -8.079e-02  4.571e-03 -17.676  < 2e-16 ***
```

```
## STARS               1.878e-01  6.092e-03  30.827  < 2e-16 ***
## Has_star            1.031e+00  1.697e-02  60.747  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(40611.66) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 13767  on 12779  degrees of freedom
## AIC: 45744
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  40612
##           Std. Err.:  34572
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -45709.79
```

Model 4 - Second Negative Binomial Regression

```
##
## Call:
## glm.nb(formula = TARGET ~ . - FixedAcidity - ResidualSugar -
##     CitricAcid - Density, data = wine_train_data, init.theta = 40601.70745,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1760  -0.6489   0.0074   0.4542   3.7598
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         4.675e-01  5.264e-02   8.881  < 2e-16 ***
## VolatileAcidity    -3.122e-02  6.519e-03  -4.789 1.67e-06 ***
## Chlorides          -3.740e-02  1.646e-02  -2.272 0.023094 *
## FreeSulfurDioxide   9.669e-05  3.502e-05   2.761 0.005764 **
## TotalSulfurDioxide  7.981e-05  2.274e-05   3.510 0.000448 ***
## pH                 -1.300e-02  7.645e-03  -1.700 0.089145 .
## Sulphates          -1.080e-02  5.675e-03  -1.904 0.056937 .
## Alcohol             3.456e-03  1.407e-03   2.456 0.014057 *
## LabelAppeal         1.590e-01  6.127e-03  25.943  < 2e-16 ***
## AcidIndex          -8.078e-02  4.512e-03 -17.903  < 2e-16 ***
## STARS               1.879e-01  6.091e-03  30.855  < 2e-16 ***
## Has_star            1.032e+00  1.697e-02  60.790  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(40601.71) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 13770  on 12783  degrees of freedom
```

```
## AIC: 45739
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  40602
##           Std. Err.:  34565
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -45712.99
```

We can see here that our Poisson and Negative Binomial models yield nearly identical results - this is a result of the fact that the Poisson regression is in fact a subset of negative binomial regressions - one that assumes (the logarithm of its expected value can be modeled by a linear combination of unknown parameters)[https://en.wikipedia.org/wiki/Poisson_regression#:~:text=Poisson%20regression%20assumes%20the%20response,used%20to%20mo

Multiple Linear Regressions

Now let's take a look at simple multiple linear regressions models using our available variables:

Model 5

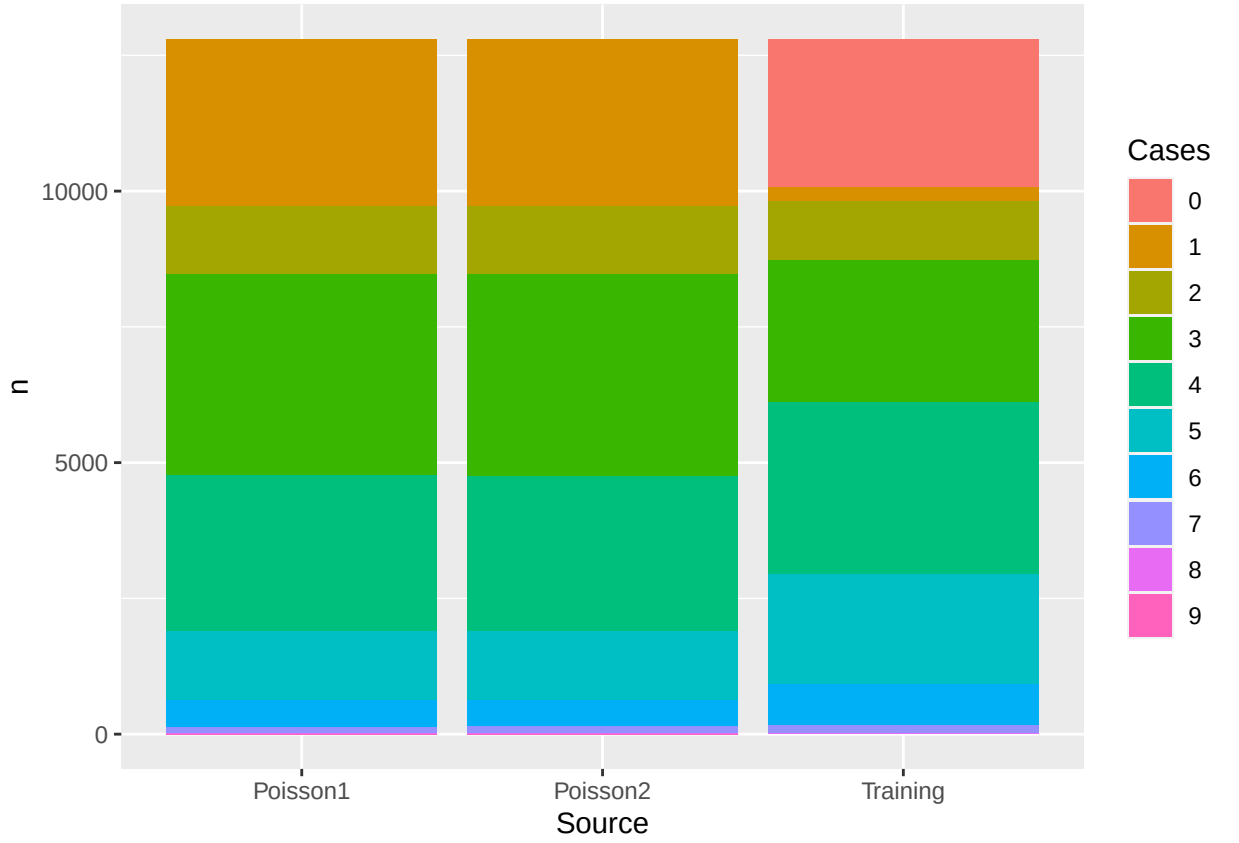
```
##
## Call:
## lm(formula = TARGET ~ ., data = wine_train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6992 -0.8524  0.0300  0.8525  6.1700
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.097e+00  4.461e-01   4.701 2.61e-06 ***
## FixedAcidity        4.308e-04  1.864e-03   0.231 0.817262
## VolatileAcidity    -9.613e-02  1.482e-02  -6.487 9.07e-11 ***
## CitricAcid          1.843e-02  1.348e-02   1.367 0.171543
## ResidualSugar       2.194e-04  3.518e-04   0.624 0.532889
## Chlorides          -1.169e-01  3.734e-02  -3.131 0.001748 **
## FreeSulfurDioxide   2.785e-04  7.999e-05   3.481 0.000501 ***
## TotalSulfurDioxide  2.236e-04  5.145e-05   4.347 1.39e-05 ***
## Density            -7.968e-01  4.371e-01  -1.823 0.068344 .
## pH                 -3.160e-02  1.735e-02  -1.821 0.068579 .
## Sulphates          -2.831e-02  1.288e-02  -2.198 0.027976 *
## Alcohol             1.233e-02  3.200e-03   3.853 0.000117 ***
## LabelAppeal         4.664e-01  1.367e-02  34.117  < 2e-16 ***
## AcidIndex          -2.012e-01  9.124e-03 -22.047  < 2e-16 ***
## STARS               7.793e-01  1.568e-02  49.710  < 2e-16 ***
## Has_star            2.276e+00  2.698e-02  84.360  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.31 on 12779 degrees of freedom
## Multiple R-squared:  0.5381, Adjusted R-squared:  0.5376
```

```
## F-statistic: 992.6 on 15 and 12779 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET ~ . - FixedAcidity - ResidualSugar - CitricAcid -
##     Density - pH - Sulphates, data = wine_train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7710 -0.8518  0.0300  0.8471  6.1872
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.186e+00  9.217e-02  12.872  < 2e-16 ***
## VolatileAcidity     -9.711e-02  1.482e-02  -6.553 5.84e-11 ***
## Chlorides           -1.173e-01  3.733e-02  -3.143 0.001677 **
## FreeSulfurDioxide    2.770e-04  7.999e-05   3.463 0.000537 ***
## TotalSulfurDioxide   2.244e-04  5.144e-05   4.361 1.30e-05 ***
## Alcohol              1.242e-02  3.200e-03   3.883 0.000104 ***
## LabelAppeal          4.663e-01  1.367e-02  34.105  < 2e-16 ***
## AcidIndex           -2.001e-01  8.941e-03 -22.383  < 2e-16 ***
## STARS                7.802e-01  1.568e-02  49.767  < 2e-16 ***
## Has_star             2.280e+00  2.697e-02  84.545  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.31 on 12785 degrees of freedom
## Multiple R-squared:  0.5376, Adjusted R-squared:  0.5373
## F-statistic:  1652 on 9 and 12785 DF,  p-value: < 2.2e-16
```

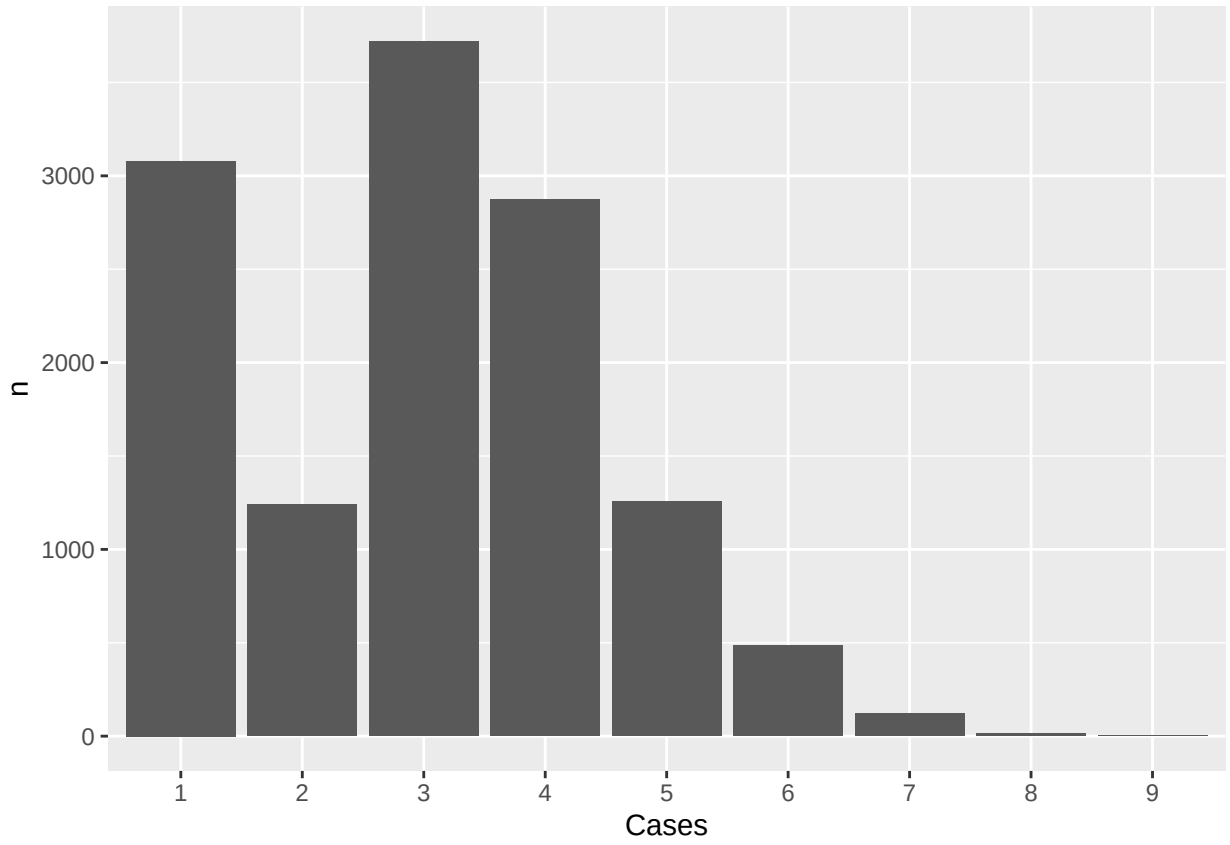
Select Models

Let's compare the distributions created by some of our models to the training data in order to evaluate which we will select as a final model.



We can see here that our models yield nearly no predictions where 0 cases are sold - this differs greatly from our traning dataset where a significant portion of wines sold 0 cases. There appears to be nearly no difference in the prediction values between our two Poisson models, showing that the impact of the removed variables is negligible. This supports our decision to do so.

If we select the second, simplified model with an AIC of 45737, the distribution of predictions we see are:


3000
2000
1000
0
n
1
2
3
4
5
6
7
8
9
Cases