Eric_Hirsch_621_Assignment_3

Predicting Wine Cases Bought

Eric Hirsch

4/7/2022

Contents

1. Data Exploration . . . 1
    - A. Summary Statistics . . . 1
    - B. Multicollinearity . . . 7
2. Data Preparation . . . 9
    - B. Address Missing Values . . . 9
    - A. Interaction terms . . . 20
    - B. Transformations . . . 20
3. Build Models . . . 20
    - A. Base Model . . . 20
    - B. Enhanced Model with Dummies and Transformations . . . 21
    - C. Enhanced Model with Interaction Terms . . . 21
4. Select Model . . . 21
5. Conclusion . . . 21


```
##          checking for file 'C:\Users\erico\AppData\Local\Temp\RtmpW4Z8cv\remotes332c3b306a74\ericons
##       -  preparing 'EHData':
##    checking DESCRIPTION meta-information ...     checking DESCRIPTION meta-information ...   v  check
##       -  checking for LF line-endings in source and make files and shell scripts
##   -  checking for empty or unneeded directories
##   -  creating default NAMESPACE file
##   -  building 'EHData_0.1.0.tar.gz'
##
##
```

1. Data Exploration

A. Summary Statistics We are modeling 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, "TARGET", is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine.

We first examine the data. The dataset consists of 12795 observations and 15 variables, all numeric. Some of the variables are count data, including the dependent varable "TARGET." The target variable has a minimum of 0, a maximum of 8 and a median of 3. Because of the small number of counts, this dataset is likely to be best modeled with a poisson or negative binomial count model.

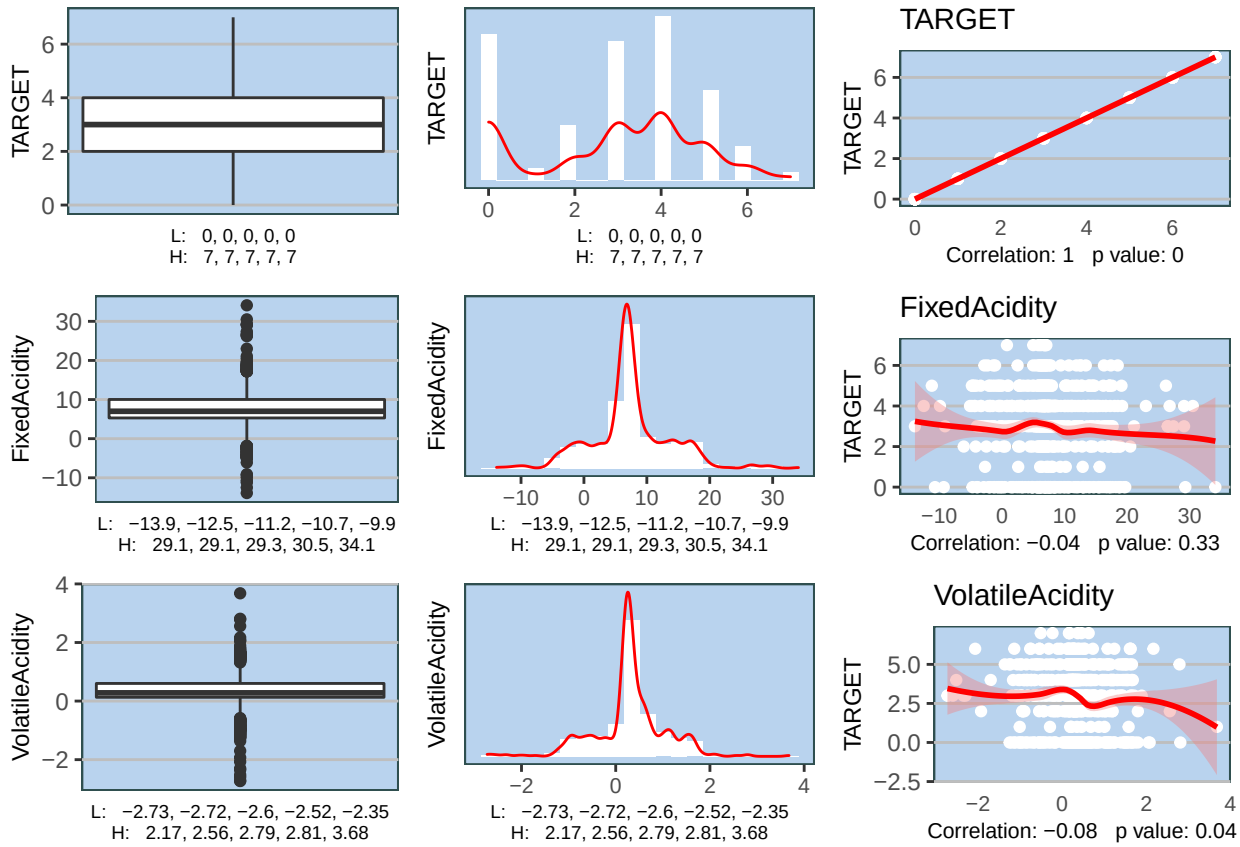
The mssing values should also be noted. In the case of STARS, e.g., 26% of the values (3359) are missing. We will need to make some decisions about these missing values.

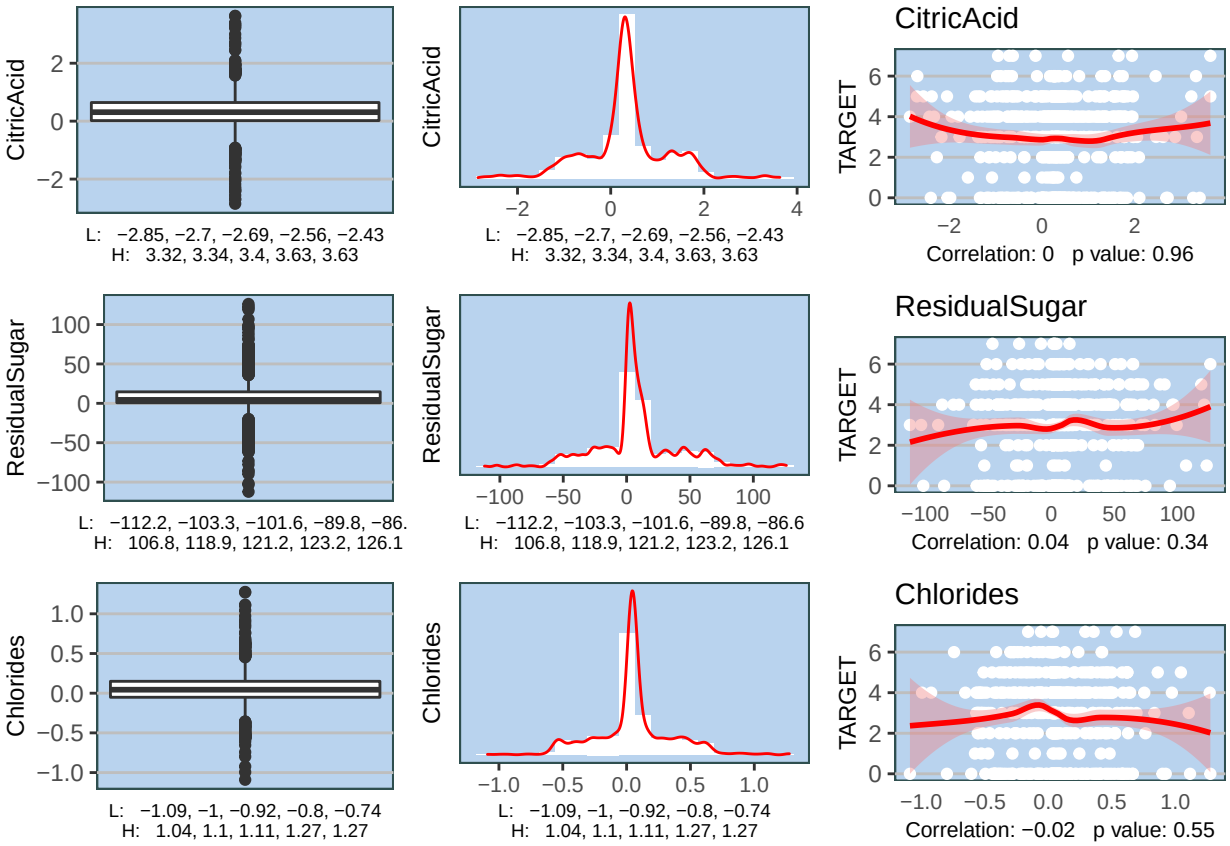
```
##      TARGET       FixedAcidity     VolatileAcidity     CitricAcid
##  Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400
##  1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300
##  Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100
##  Mean   :3.029   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
```

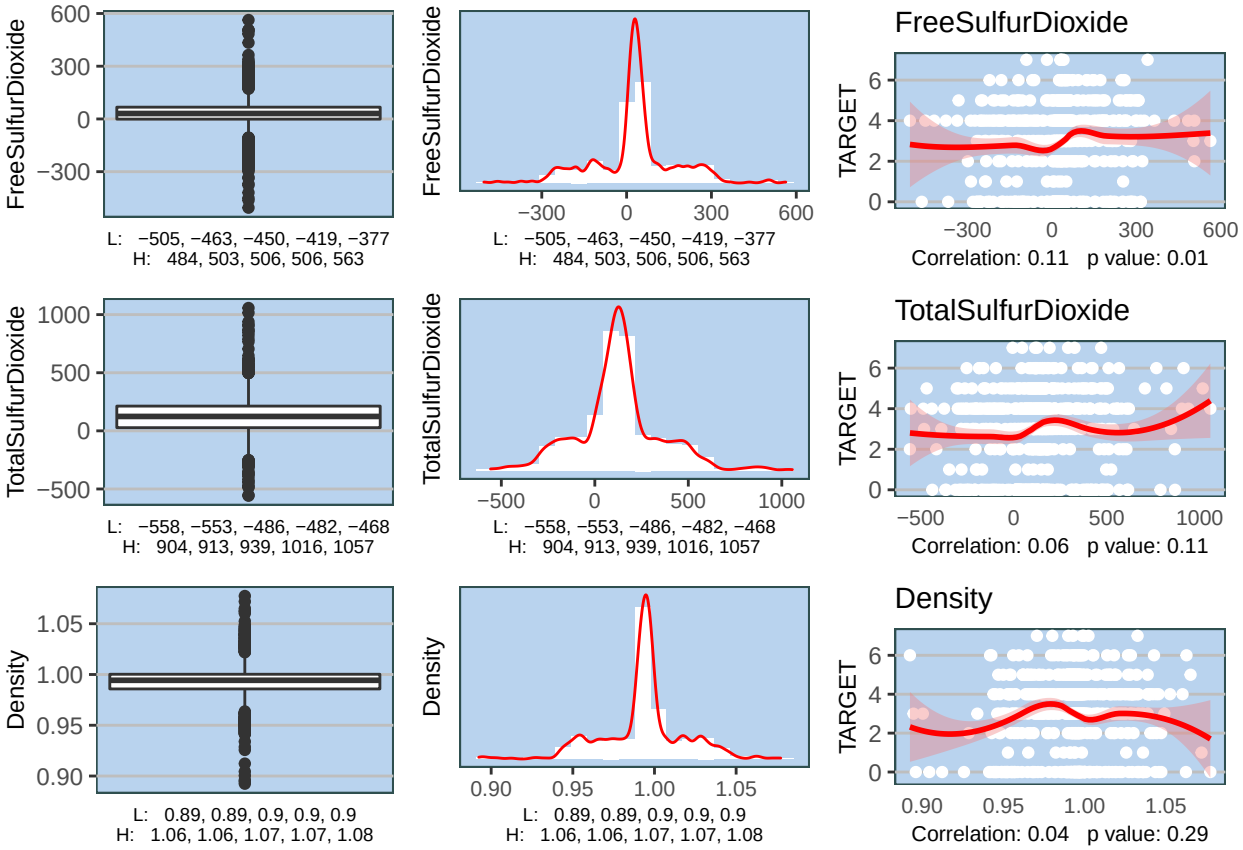
```
## 'data.frame':    12795 obs. of  15 variables:
##  $ TARGET            : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num  -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
##  $ ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num  -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num  NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num  268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num  0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##  $ Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num  9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
```

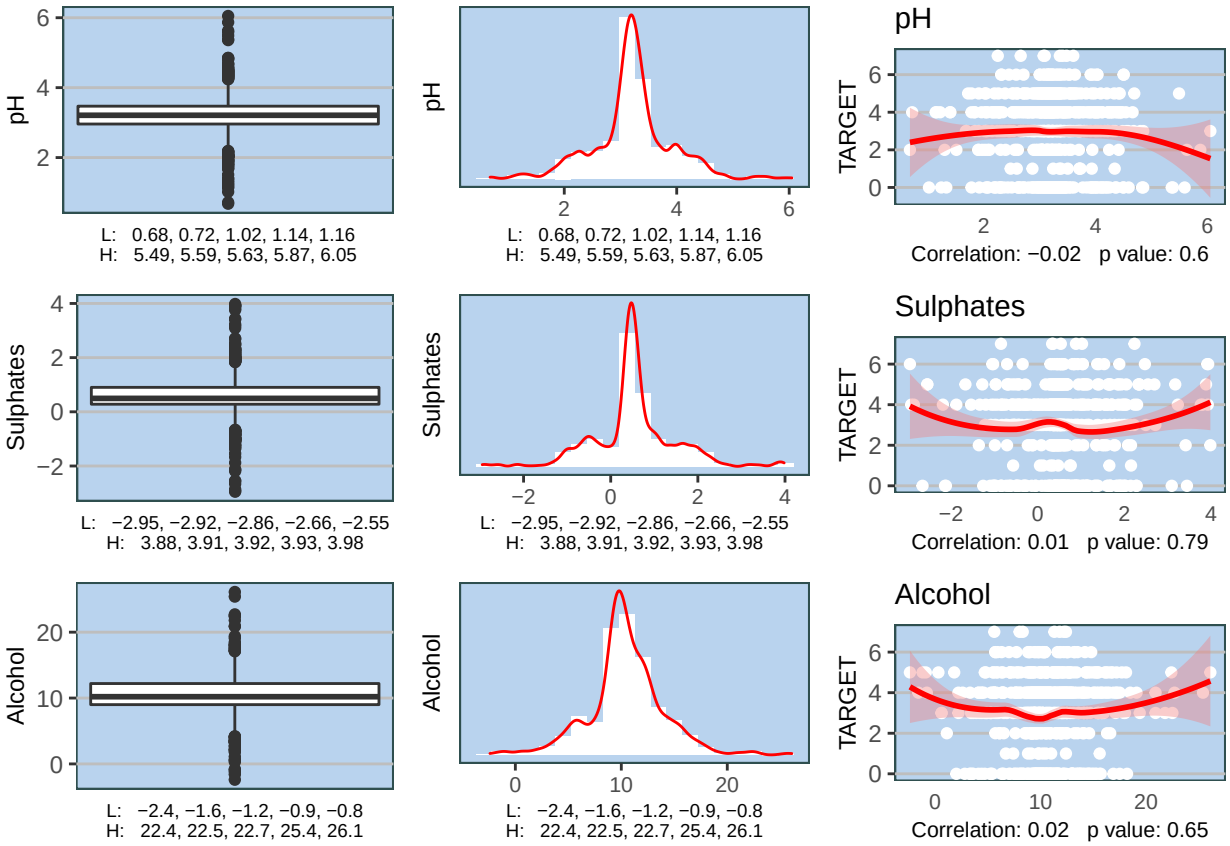
```
##  $ LabelAppeal       : int  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : int  2 3 3 1 2 NA NA 3 NA 4 ...
```

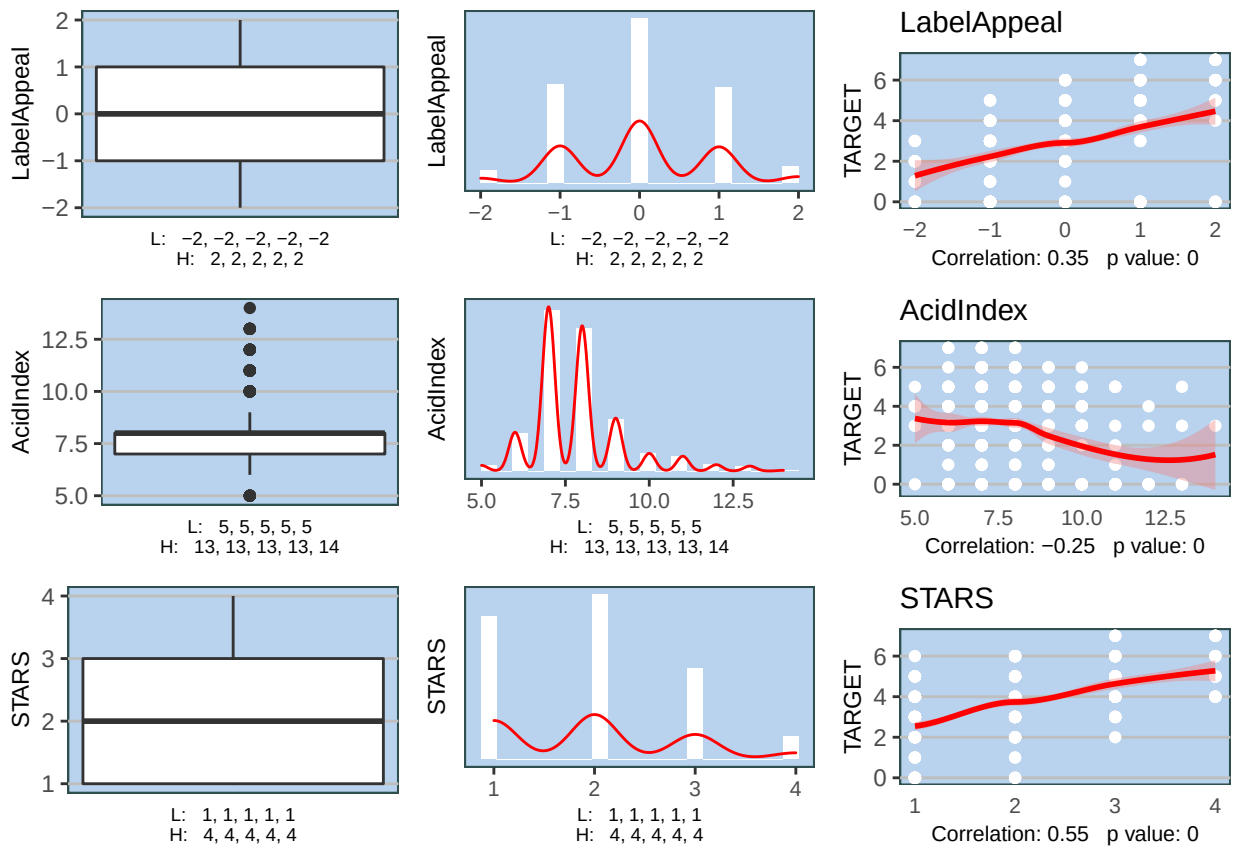
Now we examine distributions:




CitricAcid
L: −2.85, −2.7, −2.69, −2.56, −2.43
H: 3.32, 3.34, 3.4, 3.63, 3.63
CitricAcid
L: −2.85, −2.7, −2.69, −2.56, −2.43
H: 3.32, 3.34, 3.4, 3.63, 3.63
CitricAcid
TARGET
Correlation: 0 p value: 0.96
ResidualSugar
L: −112.2, −103.3, −101.6, −89.8, −86.
H: 106.8, 118.9, 121.2, 123.2, 126.1
ResidualSugar
L: −112.2, −103.3, −101.6, −89.8, −86.6
H: 106.8, 118.9, 121.2, 123.2, 126.1
ResidualSugar
TARGET
Correlation: 0.04 p value: 0.34
Chlorides
L: −1.09, −1, −0.92, −0.8, −0.74
H: 1.04, 1.1, 1.11, 1.27, 1.27
Chlorides
L: −1.09, −1, −0.92, −0.8, −0.74
H: 1.04, 1.1, 1.11, 1.27, 1.27
Chlorides
TARGET
Correlation: −0.02 p value: 0.55

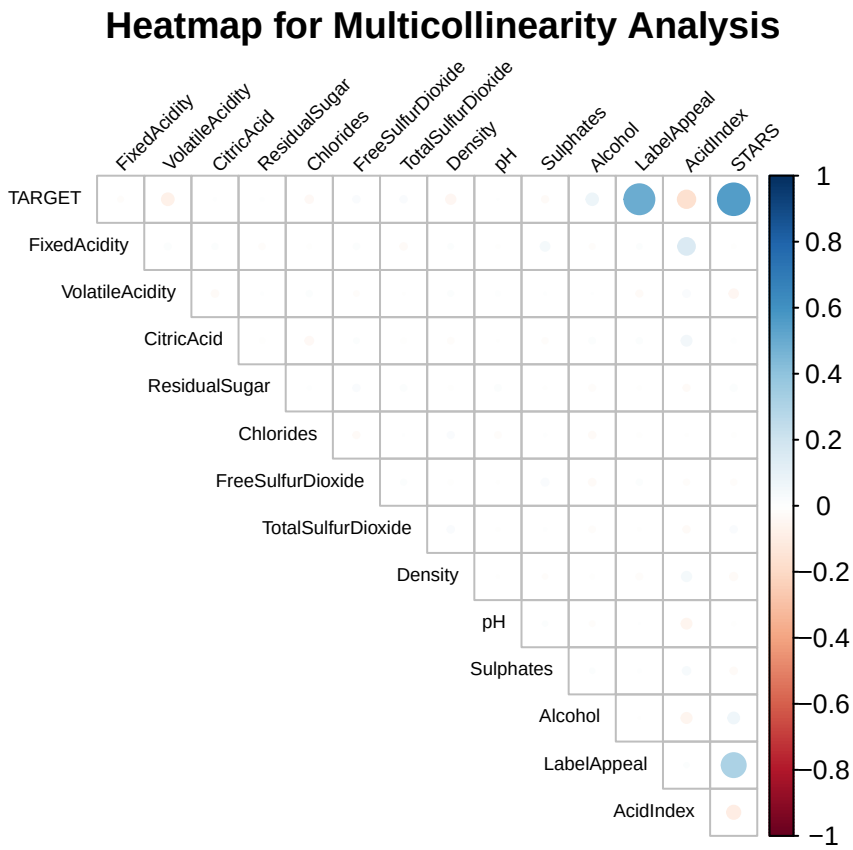

FreeSulfurDioxide
600
300
0
−300
L: −505, −463, −450, −419, −377
H: 484, 503, 506, 506, 563
FreeSulfurDioxide
−300
0
300
600
L: −505, −463, −450, −419, −377
H: 484, 503, 506, 506, 563
FreeSulfurDioxide
TARGET
6
4
2
0
−300
0
300
600
Correlation: 0.11 p value: 0.01
TotalSulfurDioxide
1000
500
0
−500
L: −558, −553, −486, −482, −468
H: 904, 913, 939, 1016, 1057
TotalSulfurDioxide
−500
0
500
1000
L: −558, −553, −486, −482, −468
H: 904, 913, 939, 1016, 1057
TotalSulfurDioxide
TARGET
6
4
2
0
−500
0
500
1000
Correlation: 0.06 p value: 0.11
Density
1.05
1.00
0.95
0.90
L: 0.89, 0.89, 0.9, 0.9, 0.9
H: 1.06, 1.06, 1.07, 1.07, 1.08
Density
0.90
0.95
1.00
1.05
L: 0.89, 0.89, 0.9, 0.9, 0.9
H: 1.06, 1.06, 1.07, 1.07, 1.08
Density
TARGET
6
4
2
0
0.90
0.95
1.00
1.05
Correlation: 0.04 p value: 0.29


pH
L: 0.68, 0.72, 1.02, 1.14, 1.16
H: 5.49, 5.59, 5.63, 5.87, 6.05
pH
L: 0.68, 0.72, 1.02, 1.14, 1.16
H: 5.49, 5.59, 5.63, 5.87, 6.05
pH
TARGET
Correlation: −0.02 p value: 0.6
Sulphates
L: −2.95, −2.92, −2.86, −2.66, −2.55
H: 3.88, 3.91, 3.92, 3.93, 3.98
Sulphates
L: −2.95, −2.92, −2.86, −2.66, −2.55
H: 3.88, 3.91, 3.92, 3.93, 3.98
Sulphates
TARGET
Correlation: 0.01 p value: 0.79
Alcohol
L: −2.4, −1.6, −1.2, −0.9, −0.8
H: 22.4, 22.5, 22.7, 25.4, 26.1
Alcohol
L: −2.4, −1.6, −1.2, −0.9, −0.8
H: 22.4, 22.5, 22.7, 25.4, 26.1
Alcohol
TARGET
Correlation: 0.02 p value: 0.65



Looking at boxplots, histograms, and scatterplots against the target variable for each variable in the dataset, we see some areas of interest. First, the target variable appears to be normally distributed for wines that had at least one purchase - but a large percentage of have zero purchases. This suggests the need for a zero-inflated model. Second, most of the variables exhibit the same pattern - a high peak at the mean of the distribution with long tails on either end. Finally, AcidIndex, LabelAppeal and Stars have, by far, the highest direct correlations with TARGET.

B. Multicollinearity Looked at from a strictly pairwise view, there is very little mutlicollinearity in this database. This is, frankly, very surprising given pairings like "FreeSulfurDixide" and "TotalSulfurDioxide", and "FixedAcidity", "VolatileAcidity" and "Ph". If I were in a workplace I would definitely explore these findings more with a chemist to understand them better and verify them.



Whether we choose a poisson or negative binomial model will depend on the residuals of our poisson model. Once we account for missing values, add columns and trasnform columns, our model will change. It is therefore a good exercise to run a poisson model at this point to get someinsight into the data.

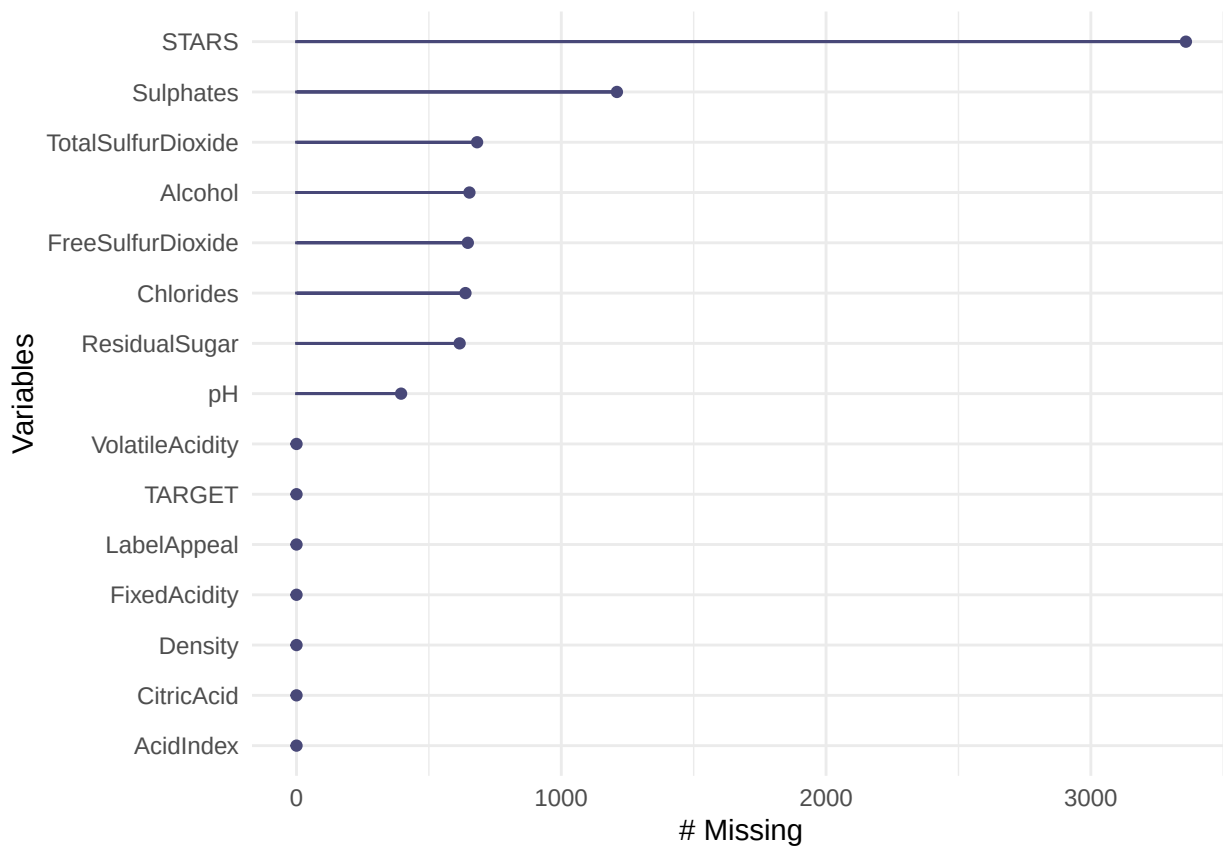
```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2158  -0.2734   0.0616   0.3732   1.6830
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.593e+00  2.506e-01   6.359 2.03e-10 ***
## FixedAcidity        3.293e-04  1.053e-03   0.313  0.75447
## VolatileAcidity    -2.560e-02  8.353e-03  -3.065  0.00218 **
## CitricAcid         -7.259e-04  7.575e-03  -0.096  0.92365
## ResidualSugar      -6.141e-05  1.941e-04  -0.316  0.75165
## Chlorides          -3.007e-02  2.056e-02  -1.463  0.14346
## FreeSulfurDioxide   6.734e-05  4.404e-05   1.529  0.12620
## TotalSulfurDioxide  2.081e-05  2.855e-05   0.729  0.46618
## Density            -3.725e-01  2.462e-01  -1.513  0.13026
## pH                 -4.661e-03  9.598e-03  -0.486  0.62722
## Sulphates          -5.164e-03  7.051e-03  -0.732  0.46398
## Alcohol             3.948e-03  1.771e-03   2.229  0.02579 *
```

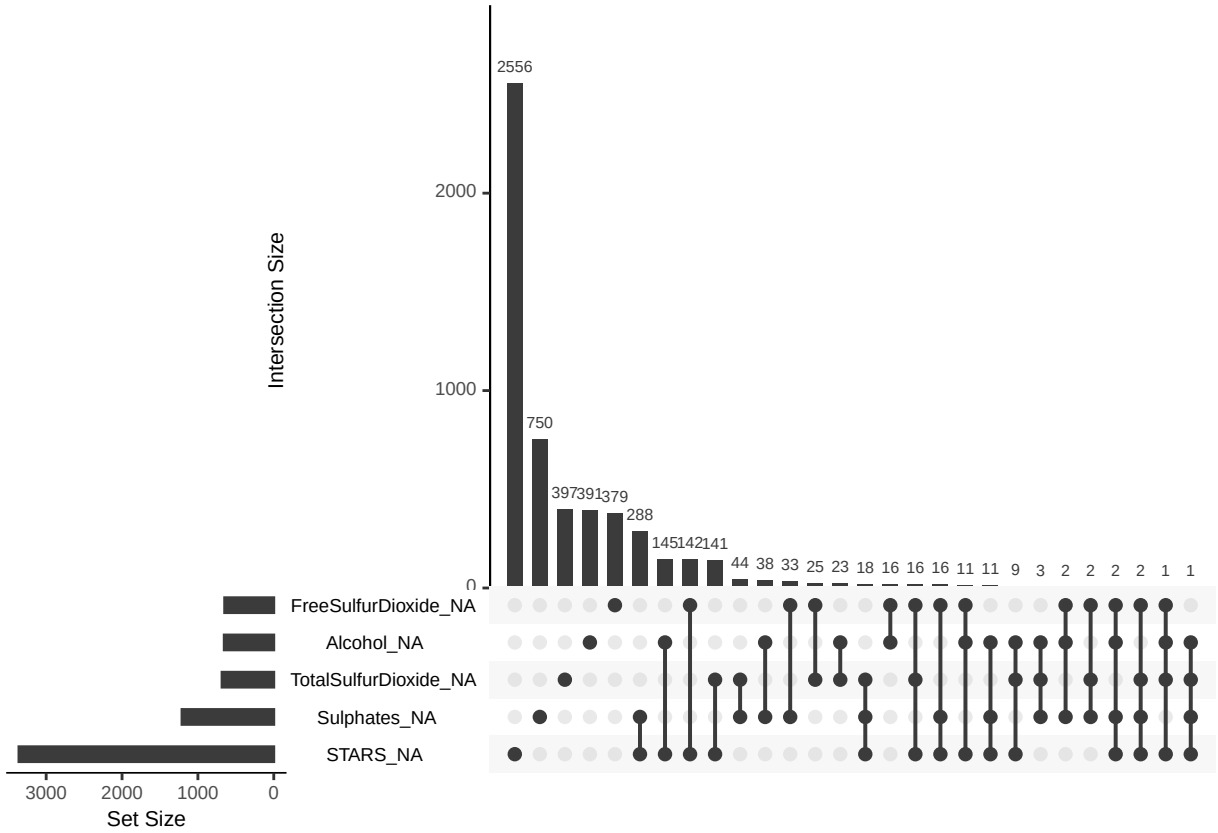
```
## LabelAppeal         1.771e-01  7.954e-03  22.271  < 2e-16 ***
## AcidIndex          -4.870e-02  5.903e-03  -8.251  < 2e-16 ***
## STARS               1.871e-01  7.487e-03  24.993  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5844.1  on 6435  degrees of freedom
## Residual deviance: 4009.1  on 6421  degrees of freedom
##   (6359 observations deleted due to missingness)
## AIC: 23172
##
## Number of Fisher Scoring iterations: 5
```

We can see that The residual deviance falls far short of the degrees of freedom, suggesting signifcant overdispersion. However, improvement in the model may help to account for more of the variance. Also, we are missing almost half the data due to missing values.

2. Data Preparation

B. Address Missing Values We consider the missing values.

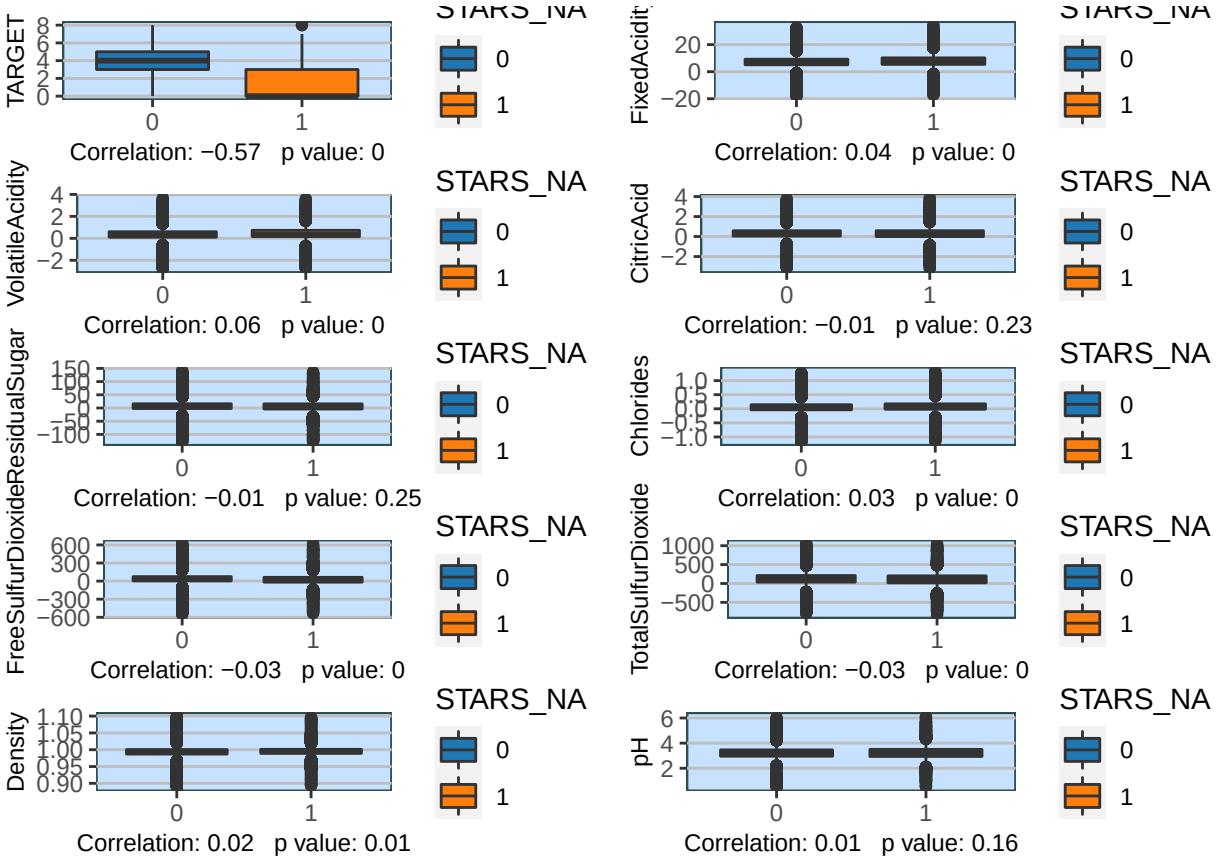


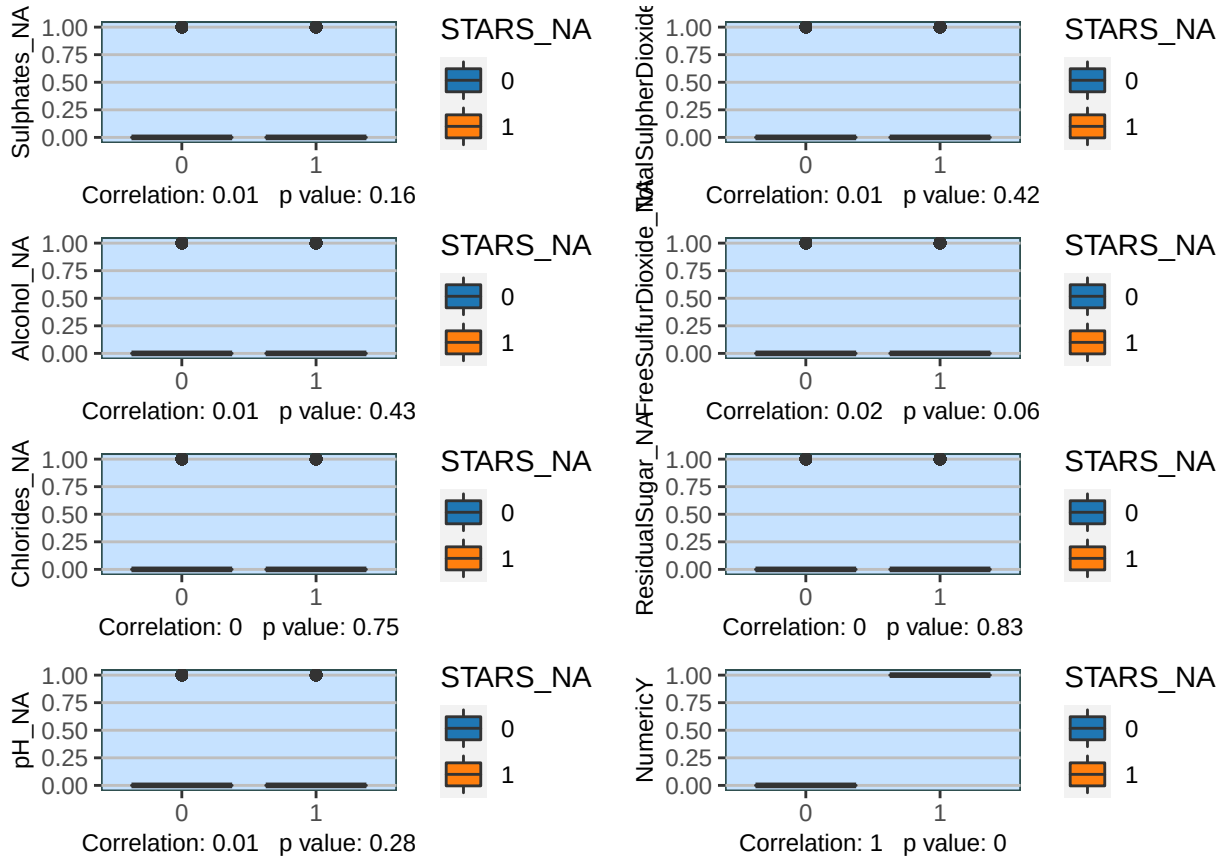


Over half of the records have missing values. Missings are confined to 8 variables. Interestingly, there is not a lot of overlap among the missings.

With so many missing values, it's prudent to create flags to track what is missing.

We investigate STARS_NA further:

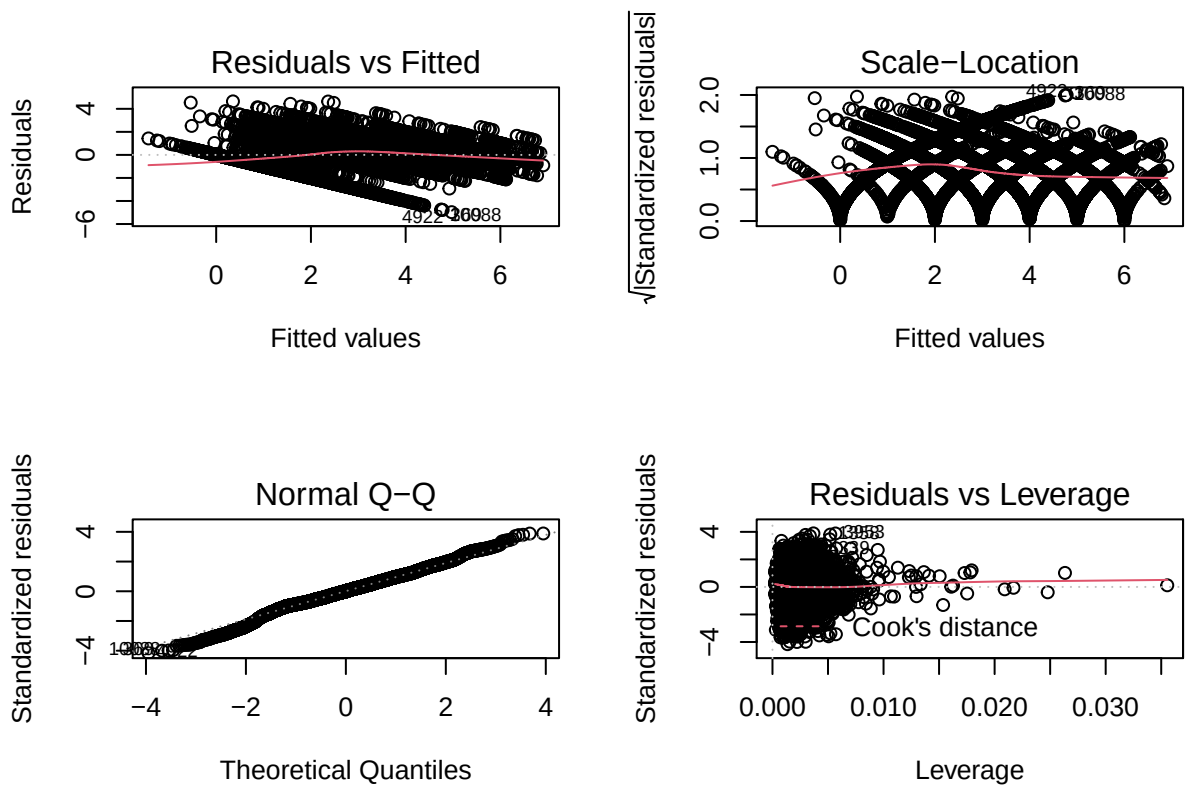

TARGET
Correlation: −0.57 p value: 0
FixedAcidity
Correlation: 0.04 p value: 0
VolatileAcidity
Correlation: 0.06 p value: 0
CitricAcid
Correlation: −0.01 p value: 0.23
ResidualSugar
Correlation: −0.01 p value: 0.25
Chlorides
Correlation: 0.03 p value: 0
FreeSulfurDioxide
Correlation: −0.03 p value: 0
TotalSulfurDioxide
Correlation: −0.03 p value: 0
Density
Correlation: 0.02 p value: 0.01
pH
Correlation: 0.01 p value: 0.16
STARS_NA
0
1



we remove the outliers

```
##
## Call:
## lm(formula = fla, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9716 -0.7529  0.0057  0.7739  4.6458
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.028297   0.010581 286.199  < 2e-16 ***
## FixedAcidity               0.002065   0.010767   0.192 0.847938
## VolatileAcidity           -0.066613   0.010624  -6.270 3.73e-10 ***
## CitricAcid                 0.010310   0.010617   0.971 0.331511
## ResidualSugar              0.003580   0.010599   0.338 0.735560
## Chlorides                 -0.029875   0.010604  -2.817 0.004849 **
## FreeSulfurDioxide          0.031869   0.010609   3.004 0.002669 **
## TotalSulfurDioxide         0.039612   0.010616   3.731 0.000191 ***
## Density                   -0.023006   0.010604  -2.169 0.030063 *
## pH                        -0.011223   0.010617  -1.057 0.290495
## Sulphates                 -0.017219   0.010603  -1.624 0.104417
## Alcohol                    0.044450   0.010632   4.181 2.93e-05 ***
## LabelAppeal                0.550764   0.013161  41.849  < 2e-16 ***
## AcidIndex                 -0.442571   0.073327  -6.036 1.63e-09 ***
## STARS                      0.229538   0.069229   3.316 0.000917 ***
```

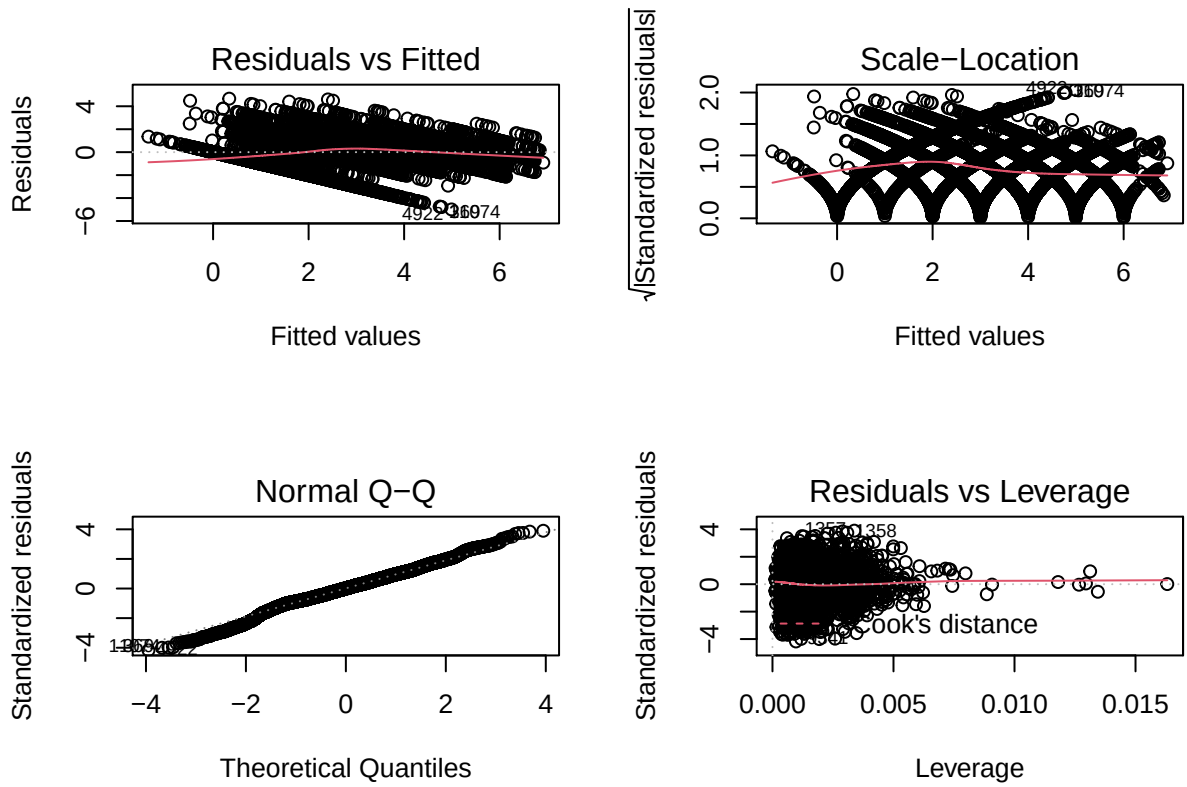
```
## STARS_NA                  -0.801752   0.011447 -70.039  < 2e-16 ***
## Sulphates_NA              -0.008752   0.010589  -0.827 0.408510
## TotalSulpherDioxide_NA     0.008232   0.010591   0.777 0.437042
## Alcohol_NA                 0.008241   0.010589   0.778 0.436410
## FreeSulfurDioxide_NA       0.014488   0.010593   1.368 0.171432
## Chlorides_NA               0.001623   0.010597   0.153 0.878252
## ResidualSugar_NA           0.006247   0.010592   0.590 0.555358
## pH_NA                     -0.015214   0.010590  -1.437 0.150819
## inter_Starsna_labelappeal -0.315352   0.012747 -24.739  < 2e-16 ***
## inter_stars_acidindex      0.547227   0.069601   7.862 4.07e-15 ***
## logAcidIndex               0.048283   0.074601   0.647 0.517504
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.197 on 12767 degrees of freedom
## Multiple R-squared:  0.6144, Adjusted R-squared:  0.6137
## F-statistic: 813.8 on 25 and 12767 DF,  p-value: < 2.2e-16
```




```
## NULL
## [1] "AIC:  40929.22494332"
```

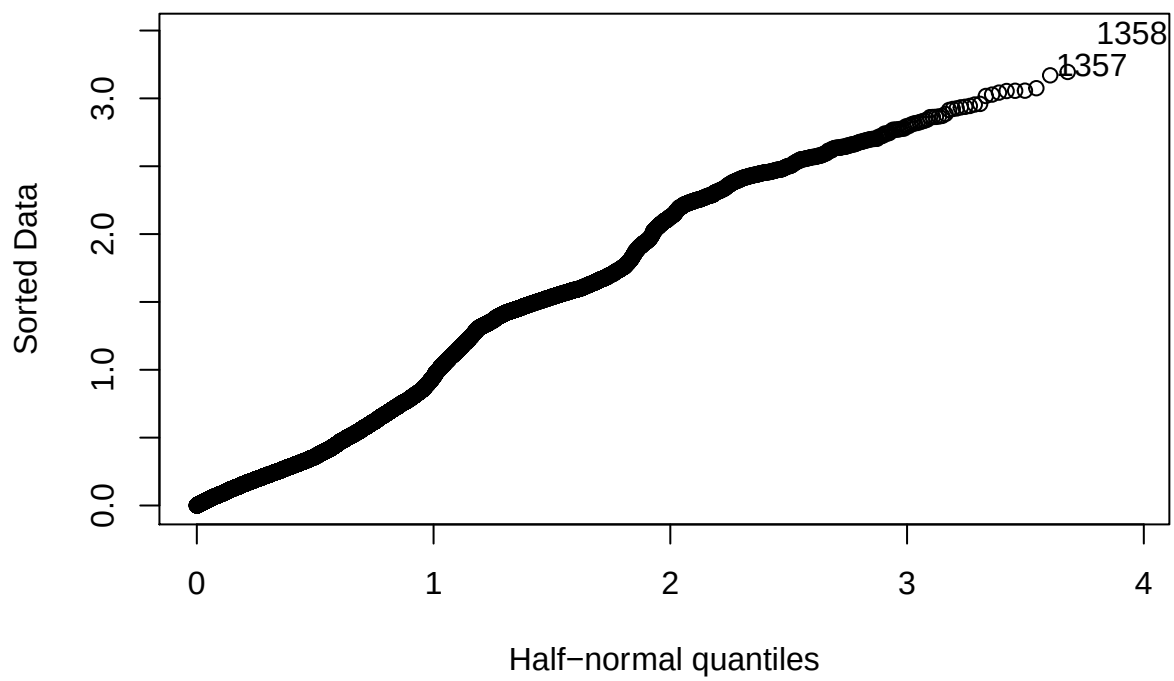
```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + Sulphates + Alcohol + LabelAppeal +
```

```
##     AcidIndex + STARS + STARS_NA + pH_NA + inter_Starsna_labelappeal +
##     inter_stars_acidindex, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9928 -0.7503  0.0113  0.7779  4.6620
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.02830    0.01058 286.245  < 2e-16 ***
## VolatileAcidity           -0.06734    0.01061  -6.346 2.29e-10 ***
## Chlorides                 -0.02964    0.01060  -2.797 0.005158 **
## FreeSulfurDioxide          0.03163    0.01060   2.985 0.002846 **
## TotalSulfurDioxide         0.04016    0.01060   3.787 0.000153 ***
## Density                   -0.02297    0.01060  -2.168 0.030214 *
## Sulphates                 -0.01709    0.01059  -1.614 0.106602
## Alcohol                    0.04472    0.01062   4.212 2.54e-05 ***
## LabelAppeal                0.55062    0.01315  41.856  < 2e-16 ***
## AcidIndex                 -0.39622    0.02414 -16.415  < 2e-16 ***
## STARS                      0.22028    0.06765   3.256 0.001132 **
## STARS_NA                  -0.80197    0.01143 -70.137  < 2e-16 ***
## pH_NA                     -0.01499    0.01058  -1.416 0.156672
## inter_Starsna_labelappeal -0.31600    0.01273 -24.816  < 2e-16 ***
## inter_stars_acidindex      0.55739    0.06797   8.200 2.62e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.197 on 12778 degrees of freedom
## Multiple R-squared:  0.6142, Adjusted R-squared:  0.6138
## F-statistic:  1453 on 14 and 12778 DF,  p-value: < 2.2e-16
```



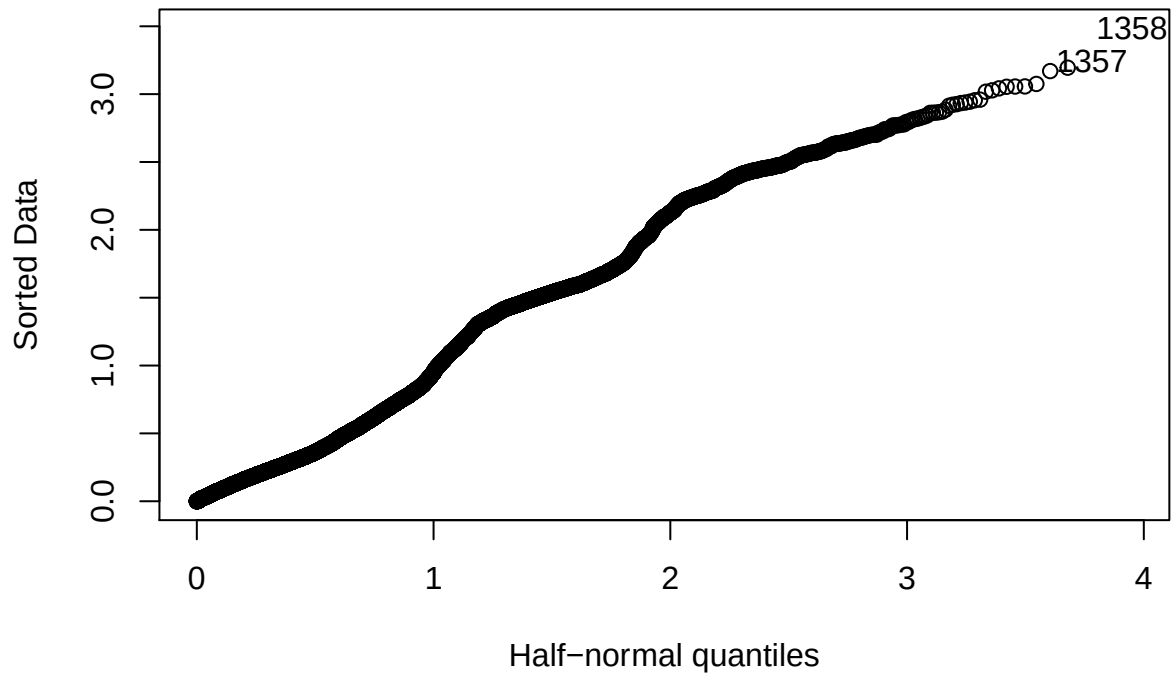

```
## NULL
## [1] "AIC:  40914.0743713364"

##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = df7a)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1699  -0.7026  -0.0227   0.4564   3.4848
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               0.9522553  0.0060073 158.518  < 2e-16 ***
## FixedAcidity              0.0007419  0.0051777   0.143  0.88606
## VolatileAcidity          -0.0216240  0.0051272  -4.217 2.47e-05 ***
## CitricAcid                0.0044275  0.0050784   0.872  0.38330
## ResidualSugar             0.0006577  0.0050970   0.129  0.89733
## Chlorides                -0.0102350  0.0051116  -2.002  0.04525 *
## FreeSulfurDioxide         0.0142354  0.0050841   2.800  0.00511 **
## TotalSulfurDioxide        0.0153250  0.0051457   2.978  0.00290 **
## Density                  -0.0083693  0.0050964  -1.642  0.10055
## pH                       -0.0062816  0.0051257  -1.226  0.22039
## Sulphates                -0.0059506  0.0050992  -1.167  0.24323
## Alcohol                   0.0111155  0.0051236   2.169  0.03005 *
```

```
## LabelAppeal                 0.1420916  0.0058149  24.436  < 2e-16 ***
## AcidIndex                  -0.4786429  0.0471301 -10.156  < 2e-16 ***
## STARS                      -0.2072496  0.0328705  -6.305 2.88e-10 ***
## STARS_NA                   -0.3928183  0.0077882 -50.437  < 2e-16 ***
## Sulphates_NA               -0.0027562  0.0051428  -0.536  0.59200
## TotalSulpherDioxide_NA      0.0029429  0.0050451   0.583  0.55968
## Alcohol_NA                  0.0018273  0.0050765   0.360  0.71889
## FreeSulfurDioxide_NA        0.0035693  0.0050859   0.702  0.48280
## Chlorides_NA               -0.0007822  0.0050704  -0.154  0.87740
## ResidualSugar_NA            0.0030013  0.0050110   0.599  0.54921
## pH_NA                      -0.0063146  0.0051744  -1.220  0.22233
## inter_Starsna_labelappeal -0.0929729  0.0088066 -10.557  < 2e-16 ***
## inter_stars_acidindex       0.4409397  0.0333852  13.208  < 2e-16 ***
## logAcidIndex                0.1976807  0.0433734   4.558 5.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22850  on 12792  degrees of freedom
## Residual deviance: 12503  on 12767  degrees of freedom
## AIC: 44489
##
## Number of Fisher Scoring iterations: 6
```




```
##
```

```
## Call:
## glm(formula = TARGET ~ ., family = quasipoisson, data = df7a)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1699  -0.7026  -0.0227   0.4564   3.4848
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.9522553  0.0053503 177.980  < 2e-16 ***
## FixedAcidity               0.0007419  0.0046115   0.161 0.872186
## VolatileAcidity           -0.0216240  0.0045666  -4.735 2.21e-06 ***
## CitricAcid                 0.0044275  0.0045231   0.979 0.327659
## ResidualSugar              0.0006577  0.0045396   0.145 0.884804
## Chlorides                 -0.0102350  0.0045526  -2.248 0.024584 *
## FreeSulfurDioxide          0.0142354  0.0045281   3.144 0.001672 **
## TotalSulfurDioxide         0.0153250  0.0045830   3.344 0.000828 ***
## Density                   -0.0083693  0.0045391  -1.844 0.065230 .
## pH                        -0.0062816  0.0045652  -1.376 0.168855
## Sulphates                 -0.0059506  0.0045416  -1.310 0.190138
## Alcohol                    0.0111155  0.0045633   2.436 0.014871 *
## LabelAppeal                0.1420916  0.0051790  27.436  < 2e-16 ***
## AcidIndex                 -0.4786429  0.0419763 -11.403  < 2e-16 ***
## STARS                     -0.2072496  0.0292760  -7.079 1.53e-12 ***
## STARS_NA                  -0.3928183  0.0069366 -56.630  < 2e-16 ***
## Sulphates_NA              -0.0027562  0.0045804  -0.602 0.547355
## TotalSulpherDioxide_NA     0.0029429  0.0044934   0.655 0.512519
## Alcohol_NA                 0.0018273  0.0045214   0.404 0.686116
## FreeSulfurDioxide_NA       0.0035693  0.0045297   0.788 0.430731
## Chlorides_NA              -0.0007822  0.0045159  -0.173 0.862497
## ResidualSugar_NA           0.0030013  0.0044630   0.672 0.501285
## pH_NA                     -0.0063146  0.0046085  -1.370 0.170647
## inter_Starsna_labelappeal -0.0929729  0.0078436 -11.853  < 2e-16 ***
## inter_stars_acidindex      0.4409397  0.0297344  14.829  < 2e-16 ***
## logAcidIndex               0.1976807  0.0386304   5.117 3.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.7932511)
##
##     Null deviance: 22850  on 12792  degrees of freedom
## Residual deviance: 12503  on 12767  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```




```
##
## Call:
## zeroinfl(formula = TARGET ~ ., data = df7a, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.17491 -0.33327 -0.02902  0.34411 12.61735
##
## Count model coefficients (poisson with log link):
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.2376360  0.0066724 185.487  < 2e-16 ***
## FixedAcidity              0.0020994  0.0053170   0.395   0.6930
## VolatileAcidity          -0.0098539  0.0052715  -1.869   0.0616 .
## CitricAcid                0.0008576  0.0051928   0.165   0.8688
## ResidualSugar            -0.0022722  0.0052221  -0.435   0.6635
## Chlorides                -0.0077965  0.0052477  -1.486   0.1374
## FreeSulfurDioxide         0.0047842  0.0051969   0.921   0.3573
## TotalSulfurDioxide       -0.0031260  0.0052617  -0.594   0.5524
## Density                  -0.0080349  0.0052578  -1.528   0.1265
## pH                        0.0035487  0.0052618   0.674   0.5000
## Sulphates                 0.0001897  0.0052394   0.036   0.9711
## Alcohol                   0.0239290  0.0052205   4.584 4.57e-06 ***
## LabelAppeal               0.1884934  0.0059421  31.721  < 2e-16 ***
## AcidIndex                -0.0122862  0.0485668  -0.253   0.8003
## STARS                     0.0425595  0.0363911   1.170   0.2422
## STARS_NA                 -0.0755743  0.0081145  -9.313  < 2e-16 ***
```

```
## Sulphates_NA              -0.0012285  0.0052722  -0.233   0.8158
## TotalSulpherDioxide_NA     0.0001752  0.0051881   0.034   0.9731
## Alcohol_NA                -0.0003047  0.0052091  -0.058   0.9534
## FreeSulfurDioxide_NA       0.0020566  0.0051975   0.396   0.6923
## Chlorides_NA               0.0007475  0.0052122   0.143   0.8860
## ResidualSugar_NA           0.0047959  0.0051138   0.938   0.3483
## pH_NA                     -0.0006054  0.0052984  -0.114   0.9090
## inter_Starsna_labelappeal  0.0635867  0.0093347   6.812 9.63e-12 ***
## inter_stars_acidindex      0.0607218  0.0371336   1.635   0.1020
## logAcidIndex              -0.0353694  0.0433339  -0.816   0.4144
##
## Zero-inflation model coefficients (binomial with logit link):
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -3.37729    0.09363 -36.071  < 2e-16 ***
## FixedAcidity               0.03464    0.04029   0.860  0.38997
## VolatileAcidity            0.15524    0.03931   3.949 7.84e-05 ***
## CitricAcid                -0.02831    0.03968  -0.713  0.47556
## ResidualSugar             -0.05868    0.03932  -1.492  0.13558
## Chlorides                  0.03360    0.03950   0.851  0.39500
## FreeSulfurDioxide         -0.05625    0.04078  -1.379  0.16776
## TotalSulfurDioxide        -0.20865    0.03933  -5.305 1.13e-07 ***
## Density                    0.01942    0.04059   0.478  0.63236
## pH                         0.11943    0.03879   3.079  0.00208 **
## Sulphates                  0.08219    0.04002   2.054  0.04002 *
## Alcohol                    0.10767    0.03979   2.706  0.00682 **
## LabelAppeal                0.56291    0.06447   8.731  < 2e-16 ***
## AcidIndex                  1.09346    0.29192   3.746  0.00018 ***
## STARS                     -2.85962    0.40047  -7.141 9.29e-13 ***
## STARS_NA                   1.43326    0.03838  37.343  < 2e-16 ***
## Sulphates_NA               0.03079    0.03833   0.803  0.42184
## TotalSulpherDioxide_NA    -0.04237    0.04103  -1.033  0.30171
## Alcohol_NA                -0.04081    0.03961  -1.030  0.30292
## FreeSulfurDioxide_NA      -0.05169    0.04053  -1.275  0.20227
## Chlorides_NA               0.02848    0.04096   0.695  0.48685
## ResidualSugar_NA           0.05278    0.03798   1.389  0.16469
## pH_NA                      0.06502    0.03686   1.764  0.07777 .
## inter_Starsna_labelappeal  0.35487    0.04878   7.275 3.46e-13 ***
## inter_stars_acidindex      0.39915    0.36702   1.088  0.27679
## logAcidIndex              -0.65661    0.26617  -2.467  0.01363 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 57
## Log-likelihood: -1.987e+04 on 52 Df
```

```
## [1] 39841.19
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## ------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                   -44.01727 model2 > model1 < 2.22e-16
## AIC-corrected         -43.53028 model2 > model1 < 2.22e-16
```

```
## BIC-corrected         -41.71462 model2 > model1 < 2.22e-16
```

A. Interaction terms As stated before, the dataset appears to hold the potential for many interaction terms, as many distributions suggest areas of very low industrialization and very high industrialization, which may affect the slope of other variables. We create some dummy variables and look for interactions. These are the dummy variables we've chosen, based on the histograms above:

```
TaxOver600
radOver10
ptOver14
lstatOver12
IndusOver16
ZnOver0
NoxOverPoint8
MedvBelow50
```

The following are just some of the possible interactions we discover affecting the dataset:

From these plots, we might draw a number of conclusions. First, in the most highly industrialized areas the crime rate appears to be high no matter the other factors. Zone 0 areas behave differently from zone 1 areas. In the few schools where the pt ratio is very small, crime rates are high, even though in general crime rates increase as ptratio increases. There may be other conclusions we can draw as well.

These interactions may or may not prove useful. We don't want to overfit the data, so it may be enough to simply include the dummy variables.

B. Transformations When the predictor variable is normally distributed but with different variance for the two values of y, we may try a quadratic function of x. rm may be considered a reasonable candidate for this.

For skewed distributions, we may include both x and log(x). nox, age, dis and lstat may be candidates for this.

Therefore we add the following transformations:

rmSquared, nox_log, age_log, dis_log, lstat_log

3. Build Models

A. Base Model We begin with the base model (all original variables from the original dataset).

In the base model, only 4 of the 12 predictors are not significant. AIC is 218. Through backward elimination we remove rm, lstat, chas and indus and arrive at a model with an AIC of 215.

Now we split the dataset 80/20 and perform 100 training iterations to test model predictive power.

The base model alone is an excellent predictor of crime rate. Accuracy is 91%. AIC of the smaller model is 171 (AIC will drop with fewer observations.)

The final results for this model, averaging 100 rounds on an 80/20 split are 91% accuracy and an AIC of 215 for the full model.

B. Enhanced Model with Dummies and Transformations Now we try a model with all of our dummy variables from above and as well as our interactions:

Despite the larger number of variables, many of which are not significant, our AIC improves substantially to 178. Through backward elimination we remove a number of variables and reduce the AIC to 163. The new model can be seen below. Despite the fact that the tax variable does not seem significant, we leave it in, as taking it out increases AIC substantially.

We split the dataset 80/20 and perform 100 training iterations to test model predictive power.

The model does a better job predicting outcomes (accuracy = .94) than the base model, and the AIC for the smaller model falls from 171 to 134.

The final results for this model, averaging 100 rounds on an 80/20 split are 94% accuracy and an AIC of 163 for the full model.

C. Enhanced Model with Interaction Terms For this model we add interaction terms in addition to dummies and transformations. We choose the following interactions as they seem the most promising:

inter_z_rm = ZnOver0∗rm
inter_age_indus = IndusOver16∗age
inter_rad_lstat = radOver10∗lstat
inter_pt_rad = ptOver14∗rad

None of the interaction terms are significant. AIC for the larger model climbs to 185, and the for the smaller it rises as well. Accuracy falls below 93%. We will therefore reject the use of interaction terms.

4. Select Model

We choose model B as the model with both best accuracy and lowest AIC. First we run some diagnostics on the model.

Observations 338 and 280 appear to be outliers and possibly influential points (especially 280). Since they are only two points, we will eliminate them to see their impact on accuracy and AIC:

This improves both accuracy and AIC and so (with the risk of overfitting) we accept the change.

Our final model is as follows. Results may differ since there as an 80/20 split - however, what is shown below is typical.

The final step is to make predictions on the evaluation set:

5. Conclusion

We examined 466 records of town statistics to create a predictive model of whether crime rates were above the median or not. We used a logistic regression to do this, testing our models on an 80/20 split 100 times and taking the average accuracy and AIC.

Several enhancements to the model increased accuracy and lowered AIC. First, some predictors were transformed with the log or square to improve fit. Second, dummy variables were introduced to capture the fact that highly industrial areas appeared to operate by a different logic than mixed use areas. interaction terms to model this phenomenon did not improve the model. The final model was 94% accurate on the holdout set.