Eric_Hirsch_621_Assignment_3

Predicting Town Crime Rates

Eric Hirsch

4/7/2022

Contents

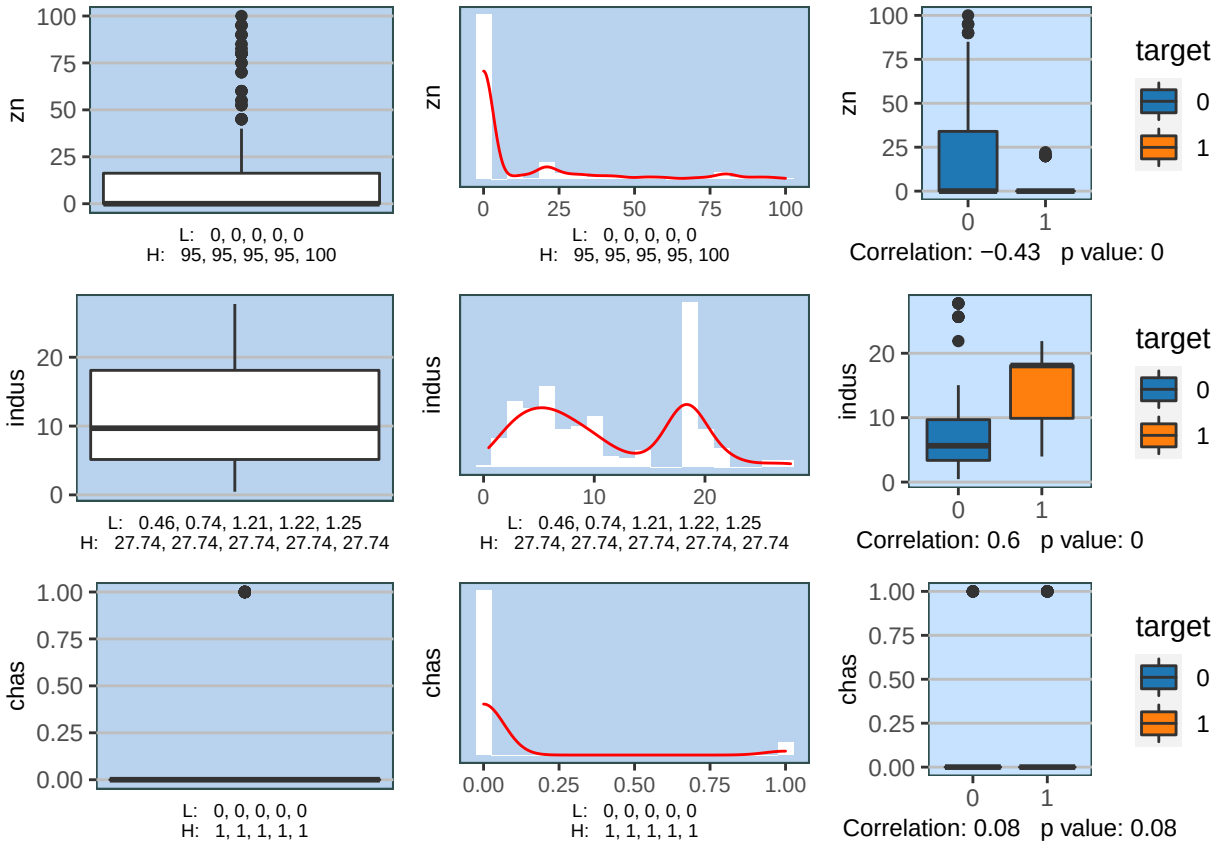
1. Data Exploration 1
A. Summary Statistics 1
B. Multicollinearity 7
2. Data Preparation 8
A. Interaction terms 8
B. Transformations 9
3. Build Models 9
A. Base Model 9
B. Enhanced Model with Dummies and Transformations 11
C. Enhanced Model with Interaction Terms 12
4. Select Model 14
5. Conclusion 22

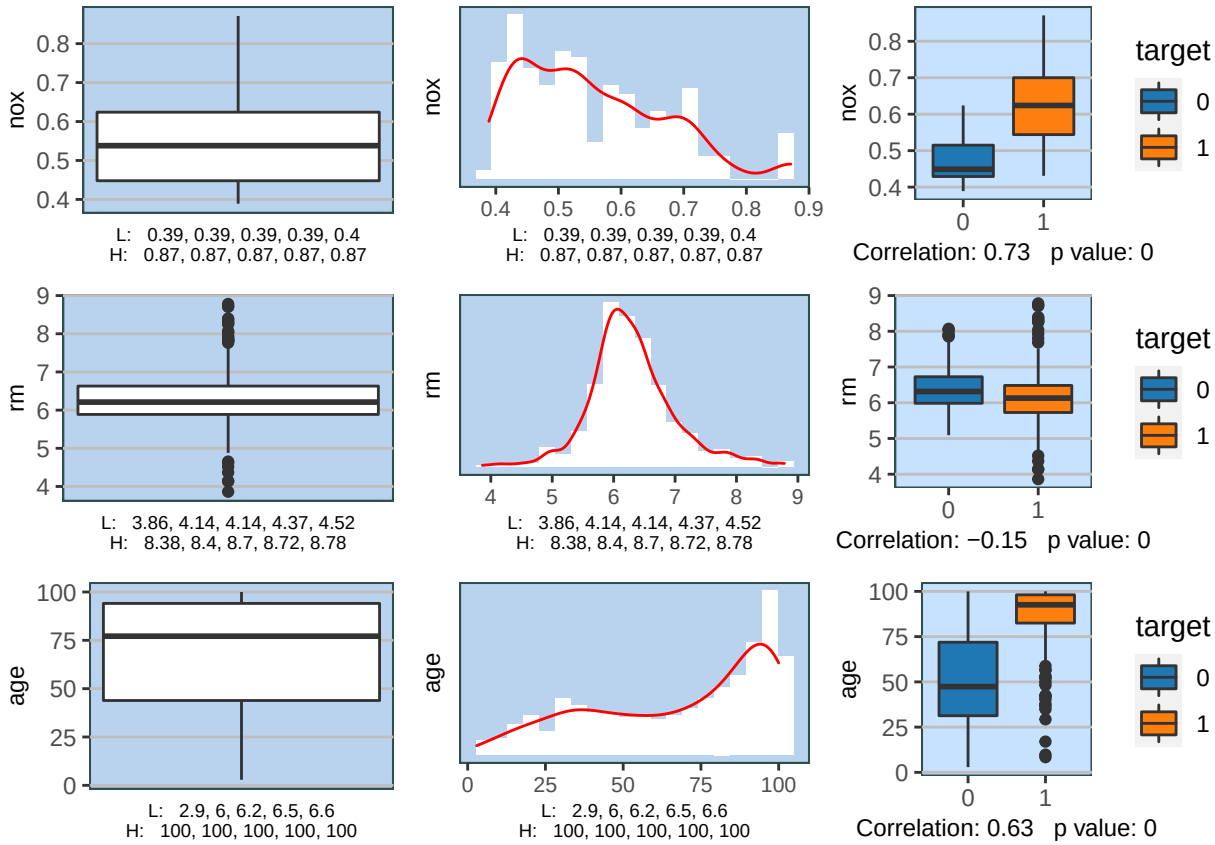
1. Data Exploration

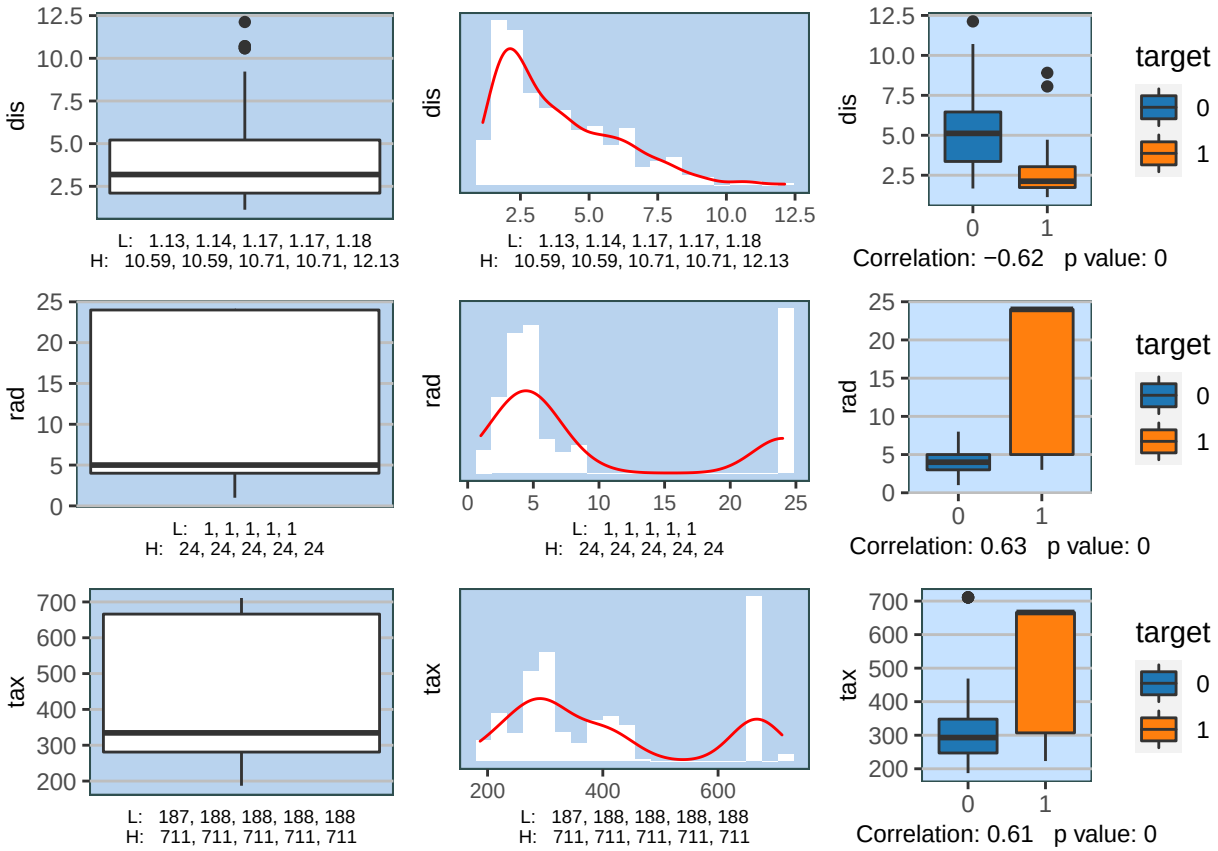
A. Summary Statistics We first examine the data. The dataset consists of 466 observations and 13 variables, all numeric. Two are binary, including the target. There are no missing values. The target appears to be relatively balanced (which makes sense, as it is an indicator of being above or below the median.)

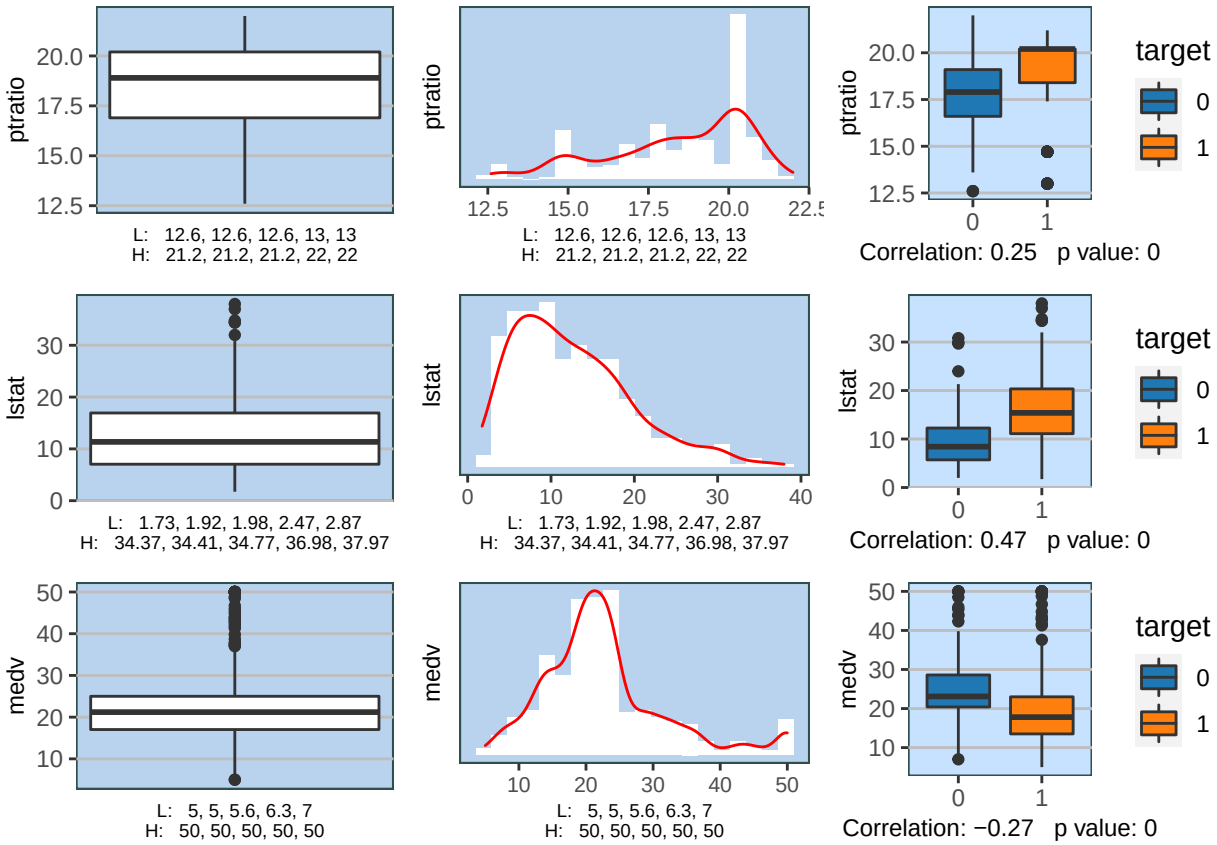
```
##        zn             indus             chas              nox        
##  Min.   :  0.00   Min.   : 0.460   Min.   :0.00000   Min.   :0.3890  
##  1st Qu.:  0.00   1st Qu.: 5.145   1st Qu.:0.00000   1st Qu.:0.4480  
##  Median :  0.00   Median : 9.690   Median :0.00000   Median :0.5380  
##  Mean   : 11.58   Mean   :11.105   Mean   :0.07082   Mean   :0.5543  
##  3rd Qu.: 16.25   3rd Qu.:18.100   3rd Qu.:0.00000   3rd Qu.:0.6240  
##  Max.   :100.00   Max.   :27.740   Max.   :1.00000   Max.   :0.8710  
##        rm              age              dis              rad        
##  Min.   :3.863   Min.   :  2.90   Min.   : 1.130   Min.   : 1.00  
##  1st Qu.:5.887   1st Qu.: 43.88   1st Qu.: 2.101   1st Qu.: 4.00  
##  Median :6.210   Median : 77.15   Median : 3.191   Median : 5.00  
```

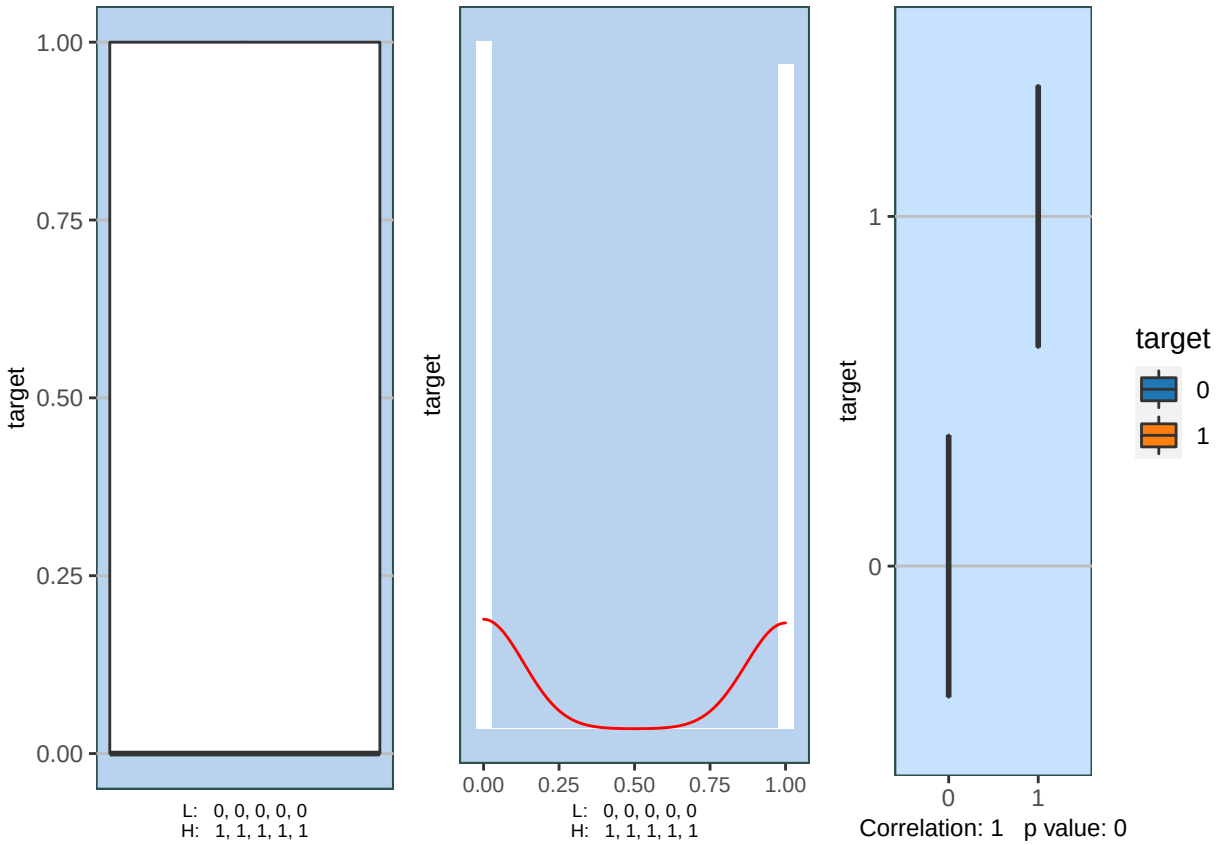
```
##  Mean   :6.291   Mean   : 68.37   Mean   : 3.796   Mean   : 9.53
##  3rd Qu.:6.630   3rd Qu.: 94.10   3rd Qu.: 5.215   3rd Qu.:24.00
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.00
##       tax           ptratio         lstat             medv
##  Min.   :187.0   Min.   :12.6   Min.   : 1.730   Min.   : 5.00
##  1st Qu.:281.0   1st Qu.:16.9   1st Qu.: 7.043   1st Qu.:17.02
##  Median :334.5   Median :18.9   Median :11.350   Median :21.20
##  Mean   :409.5   Mean   :18.4   Mean   :12.631   Mean   :22.59
##  3rd Qu.:666.0   3rd Qu.:20.2   3rd Qu.:16.930   3rd Qu.:25.00
##  Max.   :711.0   Max.   :22.0   Max.   :37.970   Max.   :50.00
##      target
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4914
##  3rd Qu.:1.0000
##  Max.   :1.0000

## 'data.frame':    466 obs. of  13 variables:
##  $ zn     : num  0 0 0 30 0 0 0 0 0 80 ...
##  $ indus  : num  19.58 19.58 18.1 4.93 2.46 ...
##  $ chas   : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.605 0.871 0.74 0.428 0.488 0.52 0.693 0.693 0.515 0.392 ...
##  $ rm     : num  7.93 5.4 6.49 6.39 7.16 ...
##  $ age    : num  96.2 100 100 7.8 92.2 71.3 100 100 38.1 19.1 ...
##  $ dis    : num  2.05 1.32 1.98 7.04 2.7 ...
##  $ rad    : int  5 5 24 6 3 5 24 24 5 1 ...
##  $ tax    : int  403 403 666 300 193 384 666 666 224 315 ...
##  $ ptratio: num  14.7 14.7 20.2 16.6 17.8 20.9 20.2 20.2 20.2 16.4 ...
##  $ lstat  : num  3.7 26.82 18.85 5.19 4.82 ...
##  $ medv   : num  50 13.4 15.4 23.7 37.9 26.5 5 7 22.2 20.9 ...
##  $ target : int  1 1 1 0 0 0 1 1 0 0 ...
```


zn
100
75
50
25
0
L: 0, 0, 0, 0, 0
H: 95, 95, 95, 95, 100
zn
0
25
50
75
100
L: 0, 0, 0, 0, 0
H: 95, 95, 95, 95, 100
zn
100
75
50
25
0
0
1
target
0
1
Correlation: −0.43 p value: 0
indus
20
10
0
L: 0.46, 0.74, 1.21, 1.22, 1.25
H: 27.74, 27.74, 27.74, 27.74, 27.74
indus
0
10
20
L: 0.46, 0.74, 1.21, 1.22, 1.25
H: 27.74, 27.74, 27.74, 27.74, 27.74
indus
20
10
0
0
1
target
0
1
Correlation: 0.6 p value: 0
chas
1.00
0.75
0.50
0.25
0.00
L: 0, 0, 0, 0, 0
H: 1, 1, 1, 1, 1
chas
0.00
0.25
0.50
0.75
1.00
L: 0, 0, 0, 0, 0
H: 1, 1, 1, 1, 1
chas
1.00
0.75
0.50
0.25
0.00
0
1
target
0
1
Correlation: 0.08 p value: 0.08


nox
0.8
0.7
0.6
0.5
0.4
L: 0.39, 0.39, 0.39, 0.39, 0.4
H: 0.87, 0.87, 0.87, 0.87, 0.87
nox
0.4 0.5 0.6 0.7 0.8 0.9
L: 0.39, 0.39, 0.39, 0.39, 0.4
H: 0.87, 0.87, 0.87, 0.87, 0.87
nox
0.8
0.7
0.6
0.5
0.4
0 1
target
0
1
Correlation: 0.73 p value: 0
rm
9
8
7
6
5
4
L: 3.86, 4.14, 4.14, 4.37, 4.52
H: 8.38, 8.4, 8.7, 8.72, 8.78
rm
4 5 6 7 8 9
L: 3.86, 4.14, 4.14, 4.37, 4.52
H: 8.38, 8.4, 8.7, 8.72, 8.78
rm
9
8
7
6
5
4
0 1
target
0
1
Correlation: −0.15 p value: 0
age
100
75
50
25
0
L: 2.9, 6, 6.2, 6.5, 6.6
H: 100, 100, 100, 100, 100
age
0 25 50 75 100
L: 2.9, 6, 6.2, 6.5, 6.6
H: 100, 100, 100, 100, 100
age
100
75
50
25
0
0 1
target
0
1
Correlation: 0.63 p value: 0


dis
12.5
10.0
7.5
5.0
2.5
L: 1.13, 1.14, 1.17, 1.17, 1.18
H: 10.59, 10.59, 10.71, 10.71, 12.13
dis
2.5
5.0
7.5
10.0
12.5
L: 1.13, 1.14, 1.17, 1.17, 1.18
H: 10.59, 10.59, 10.71, 10.71, 12.13
dis
target
0
1
Correlation: −0.62 p value: 0
rad
25
20
15
10
5
0
L: 1, 1, 1, 1, 1
H: 24, 24, 24, 24, 24
rad
0
5
10
15
20
25
L: 1, 1, 1, 1, 1
H: 24, 24, 24, 24, 24
rad
target
0
1
Correlation: 0.63 p value: 0
tax
700
600
500
400
300
200
L: 187, 188, 188, 188, 188
H: 711, 711, 711, 711, 711
tax
200
400
600
L: 187, 188, 188, 188, 188
H: 711, 711, 711, 711, 711
tax
target
0
1
Correlation: 0.61 p value: 0

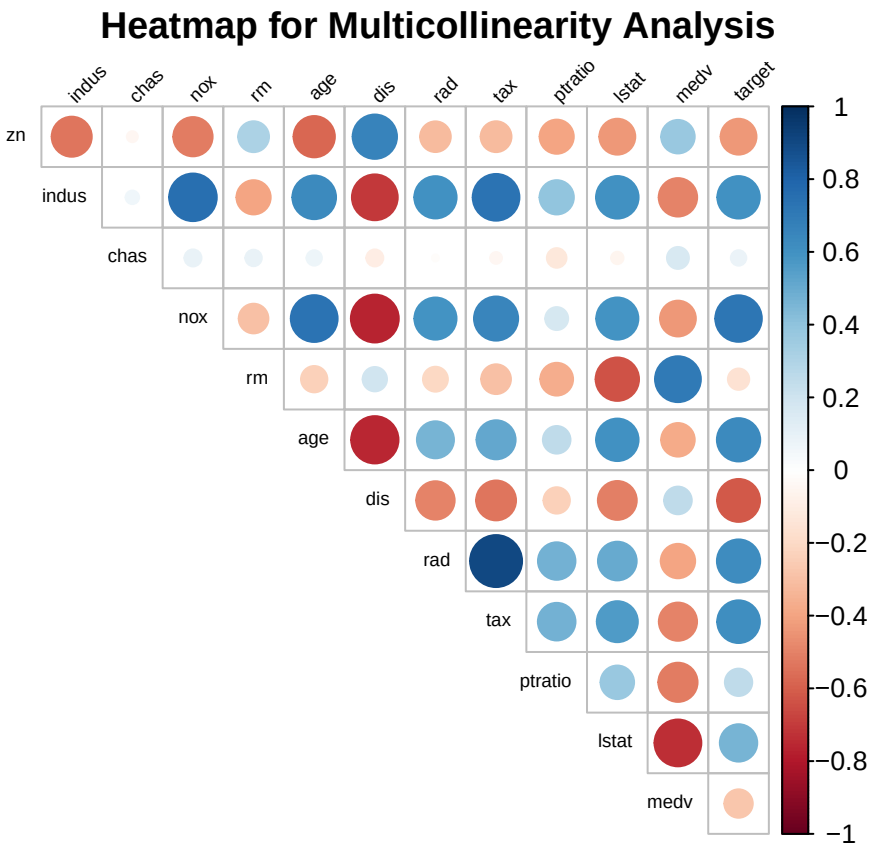

ptratio
20.0
17.5
15.0
12.5
L: 12.6, 12.6, 12.6, 13, 13
H: 21.2, 21.2, 21.2, 22, 22
ptratio
12.5
15.0
17.5
20.0
L: 12.6, 12.6, 12.6, 13, 13
H: 21.2, 21.2, 21.2, 22, 22
ptratio
target
0
1
Correlation: 0.25 p value: 0
lstat
30
20
10
0
L: 1.73, 1.92, 1.98, 2.47, 2.87
H: 34.37, 34.41, 34.77, 36.98, 37.97
lstat
0
10
20
30
40
L: 1.73, 1.92, 1.98, 2.47, 2.87
H: 34.37, 34.41, 34.77, 36.98, 37.97
lstat
target
0
1
Correlation: 0.47 p value: 0
medv
50
40
30
20
10
L: 5, 5, 5.6, 6.3, 7
H: 50, 50, 50, 50, 50
medv
10
20
30
40
50
L: 5, 5, 5.6, 6.3, 7
H: 50, 50, 50, 50, 50
medv
target
0
1
Correlation: −0.27 p value: 0



Looking at boxplots, histograms, and boxplots against the target variable, we see some areas of interest. A number of distributions are broken (e.g. zn, indus, nox and rad), suggesting there may be hidden grouping within the variables. For example, zn = 0 may include areas that are different in their makeup from zn>0. In fact, there may be a common phenomenon among all of them which identifies certain areas as highly industrial with a different crime logic, as opposed to the mixed industrial and residential areas for the rest of the observations.

Most of the correlations are unsurprising, with the exception of tax rate, which increases with increase in crime. Again, this may reflect some interaction, since residential and industrial areas may both be high tax areas, but residential areas may see more crime as they become industrial, whereas industrial areas may see less (as the # of residents approaches 0).

B. Multicollinearity Mutlicollinearity is highly evident in the database - not surprisingly, given the discussion above. The correlation between rad and tax is over 90%. We also see here the correlation between tax rate and level of industrialization (which is also correlated with high crime rates), which explains the counterintuitive result above.

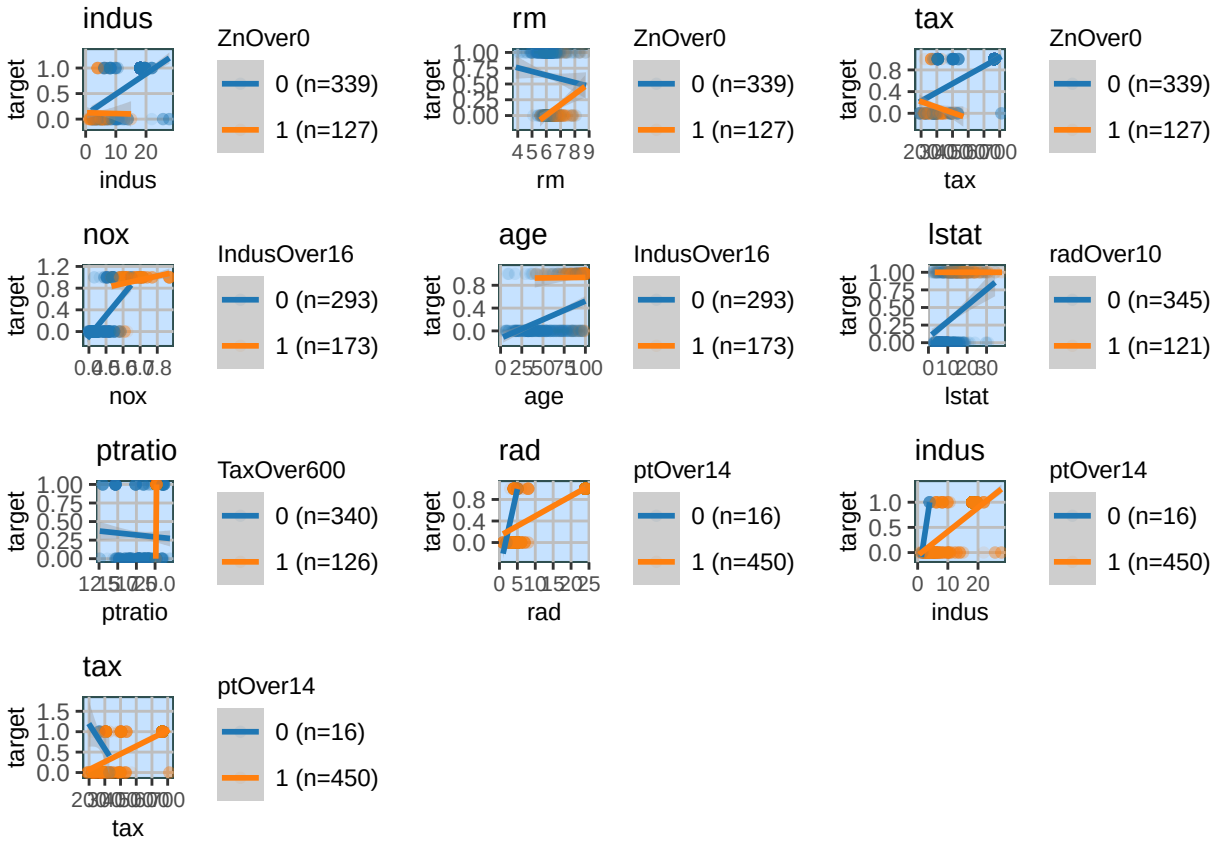


2. Data Preparation

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).

```
##
## Call:
## glm(formula = fla, family = "binomial", data = df)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8464  -0.1445  -0.0017   0.0029   3.4665
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -40.822934   6.632913  -6.155 7.53e-10 ***
## zn           -0.065946   0.034656  -1.903  0.05706 .
## indus        -0.064614   0.047622  -1.357  0.17485
## chas          0.910765   0.755546   1.205  0.22803
## nox          49.122297   7.931706   6.193 5.90e-10 ***
## rm           -0.587488   0.722847  -0.813  0.41637
## age           0.034189   0.013814   2.475  0.01333 *
## dis           0.738660   0.230275   3.208  0.00134 **
## rad           0.666366   0.163152   4.084 4.42e-05 ***
## tax          -0.006171   0.002955  -2.089  0.03674 *
## ptratio       0.402566   0.126627   3.179  0.00148 **
## lstat         0.045869   0.054049   0.849  0.39608
## medv          0.180824   0.068294   2.648  0.00810 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 192.05  on 453  degrees of freedom
## AIC: 218.05
##
## Number of Fisher Scoring iterations: 9

## [[1]]
##   (Intercept)            zn         indus          chas           nox
## -40.822933572  -0.065945977  -0.064613849   0.910765481  49.122296648
##            rm           age           dis           rad           tax
##  -0.587488460   0.034188978   0.738660343   0.666365942  -0.006171394
##       ptratio         lstat          medv
##   0.402565647   0.045868544   0.180824004
##
## [[2]]
## [1] 0
##
## [[3]]
## [1] 0
```

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.

```
## [1] "Accuracy:  0.907627820732475"
```

```
## [1] "AIC:  170.33052882748"
```

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.

```
##
## Call:
## glm(formula = fla, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2233  -0.1234  -0.0002   0.0114   3.6303
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 110.857039  23.834130   4.651 3.30e-06 ***
## rm          -34.381951   7.268697  -4.730 2.24e-06 ***
## age           0.117686   0.029434   3.998 6.38e-05 ***
## dis          -5.463106   1.212821  -4.504 6.65e-06 ***
## rad           1.004539   0.202405   4.963 6.94e-07 ***
## tax           0.005968   0.004581   1.303 0.192717
## ptratio       0.544714   0.146585   3.716 0.000202 ***
## TaxOver600  -12.678775   5.455535  -2.324 0.020124 *
## ptOver14     -7.985525   2.364753  -3.377 0.000733 ***
## lstatOver12  -1.267043   0.626465  -2.023 0.043122 *
## IndusOver16   5.786874   1.509176   3.834 0.000126 ***
## rmSquared     2.619120   0.549819   4.764 1.90e-06 ***
## nox_log      20.568245   4.372626   4.704 2.55e-06 ***
## age_log      -3.867305   1.134733  -3.408 0.000654 ***
## dis_log      26.063405   5.407276   4.820 1.44e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 132.76  on 451  degrees of freedom
## AIC: 162.76
##
## Number of Fisher Scoring iterations: 11

## [[1]]
##   (Intercept)            rm           age           dis           rad
## 110.857039007 -34.381951421   0.117686280  -5.463105933   1.004539022
##           tax       ptratio    TaxOver600      ptOver14   lstatOver12
```

```
##   0.005967664   0.544713788 -12.678774877  -7.985524796  -1.267043185
##   IndusOver16     rmSquared       nox_log       age_log       dis_log
##   5.786874133   2.619120119  20.568244518  -3.867304509  26.063405107
##
## [[2]]
## [1] 0
##
## [[3]]
## [1] 0
```

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

```
## [1] "Accuracy:  0.936337222603523"
```

```
## [1] "AIC:  131.533840800871"
```

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

```
##
## Call:
## glm(formula = fla, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2111  -0.1316   0.0000   0.0000   3.4644
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      8.841e+01  1.209e+04   0.007 0.994165
## zn              -2.701e-03  7.522e-02  -0.036 0.971355
## indus           -2.693e-02  1.313e-01  -0.205 0.837545
## chas             1.295e+00  9.155e-01   1.415 0.157073
## nox             -2.425e+01  1.640e+02  -0.148 0.882414
## rm              -3.360e+01  8.830e+00  -3.806 0.000141 ***
## age              1.255e-01  3.786e-02   3.315 0.000916 ***
## dis             -5.341e+00  1.388e+00  -3.848 0.000119 ***
## rad              1.110e+01  2.764e+03   0.004 0.996795
## tax              9.307e-03  5.846e-03   1.592 0.111410
## ptratio          7.847e-01  2.275e-01   3.449 0.000563 ***
## lstat            3.639e-02  1.859e-01   0.196 0.844794
## medv             1.650e-01  1.188e-01   1.389 0.164916
```

```
## TaxOver600      -2.797e+01  7.115e+03  -0.004 0.996863
## radOver10        2.761e+01  8.445e+03   0.003 0.997392
## ptOver14         2.501e+01  1.209e+04   0.002 0.998349
## lstatOver12     -7.499e-01  8.873e-01  -0.845 0.398049
## IndusOver16     -8.763e-01  1.143e+01  -0.077 0.938901
## ZnOver0          4.966e+00  7.872e+00   0.631 0.528152
## NoxOverPoint8    1.853e+01  3.389e+03   0.005 0.995638
## MedvBelow50      3.087e+00  4.705e+00   0.656 0.511766
## rmSquared        2.464e+00  6.843e-01   3.600 0.000318 ***
## nox_log          3.237e+01  8.350e+01   0.388 0.698218
## age_log         -3.794e+00  1.387e+00  -2.736 0.006228 **
## dis_log          2.635e+01  5.852e+00   4.503 6.71e-06 ***
## lstat_log       -1.159e+00  2.546e+00  -0.455 0.648934
## inter_z_rm      -8.088e-01  1.219e+00  -0.664 0.506867
## inter_age_indus  7.831e-02  1.217e-01   0.644 0.519869
## inter_rad_lstat  7.544e-02  2.337e+02   0.000 0.999742
## inter_pt_rad    -1.024e+01  2.764e+03  -0.004 0.997044
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.88  on 465  degrees of freedom
## Residual deviance: 125.13  on 436  degrees of freedom
## AIC: 185.13
##
## Number of Fisher Scoring iterations: 19

## [[1]]
##     (Intercept)              zn           indus            chas             nox
##    88.405179380    -0.002701214    -0.026928233     1.295405094   -24.250858011
##              rm             age             dis             rad             tax
##   -33.603875571     0.125524602    -5.340688660    11.099719226     0.009306557
##         ptratio           lstat            medv      TaxOver600       radOver10
##     0.784663220     0.036393900     0.164997414   -27.972319462    27.605765639
##        ptOver14     lstatOver12     IndusOver16         ZnOver0   NoxOverPoint8
##    25.014533258    -0.749901656    -0.876264754     4.965787827    18.530465388
##     MedvBelow50       rmSquared         nox_log         age_log         dis_log
##     3.087200855     2.463624825    32.373592763    -3.794441325    26.349177856
##       lstat_log      inter_z_rm inter_age_indus inter_rad_lstat    inter_pt_rad
##    -1.159185712    -0.808789459     0.078313317     0.075435312   -10.236961583
##
## [[2]]
## [1] 0
##
## [[3]]
## [1] 0
```

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.

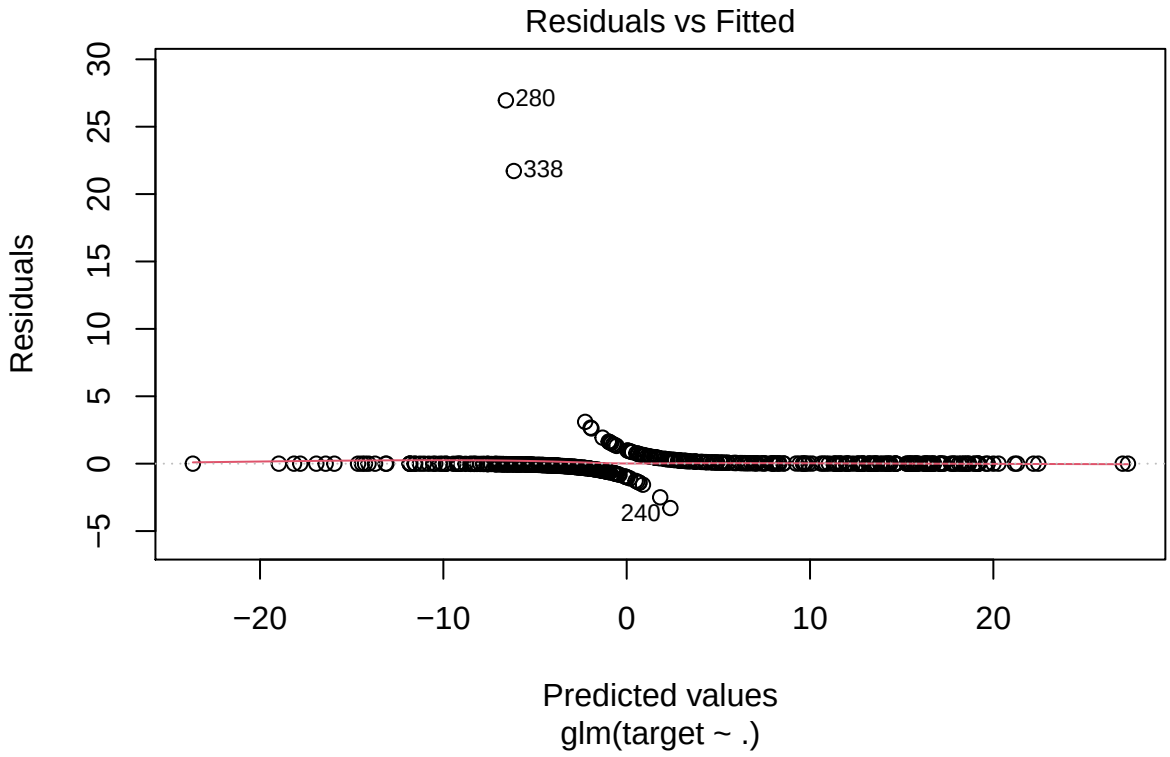
```
## [1] "Accuracy:  0.925644988680115"
```

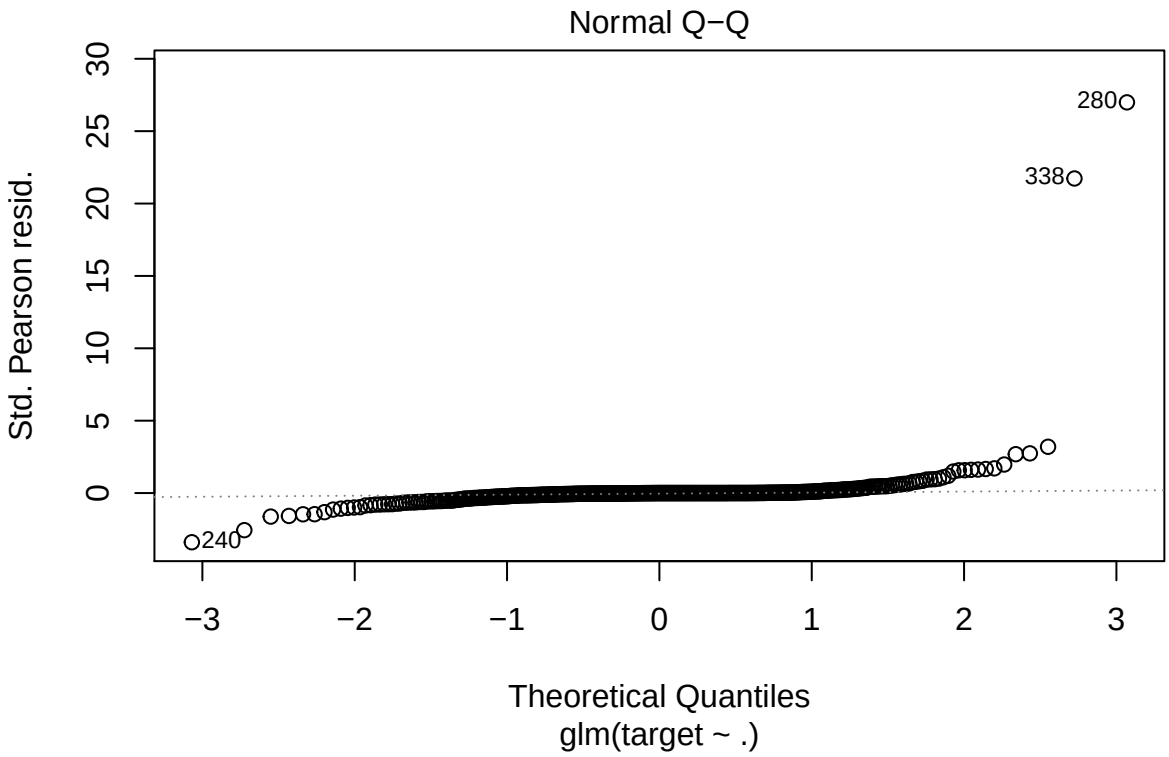
```
## [1] "AIC:  152.650548790424"
```

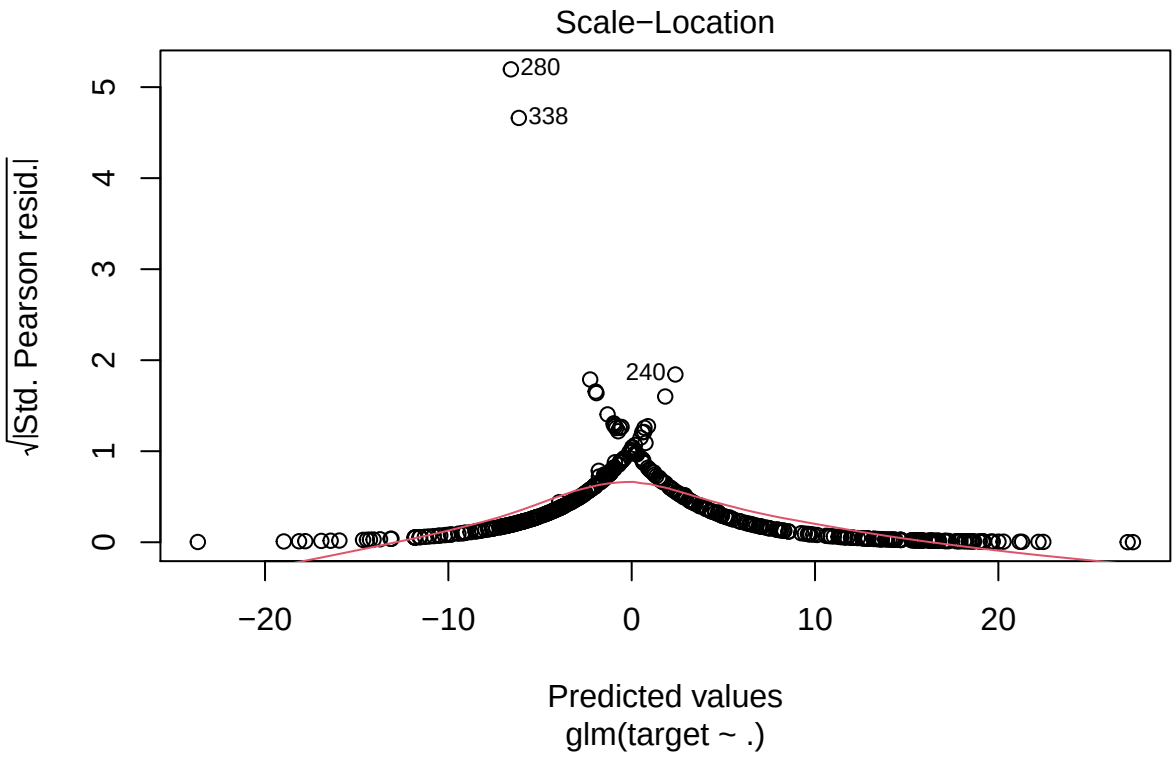
4. Select Model

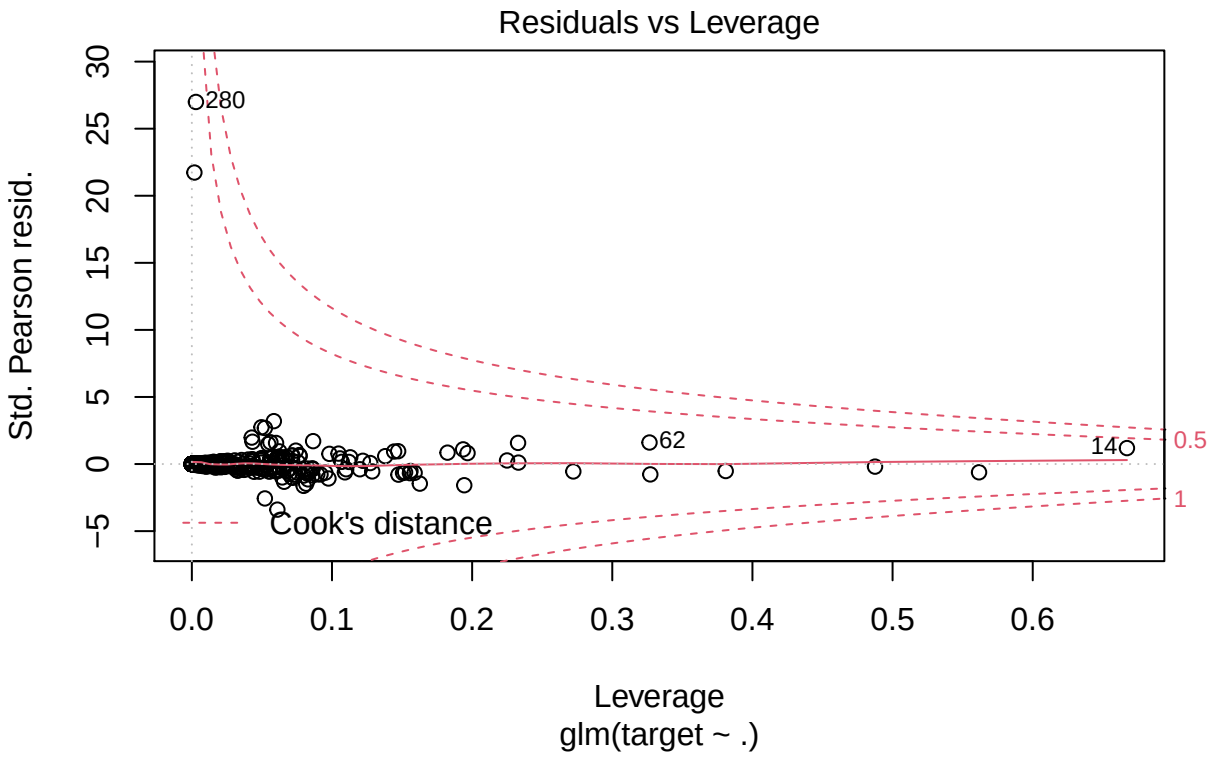
We choose model B as the model with both best accuracy and lowest AIC. First we run some diagnostics on the model.

```
##
## Call:  glm(formula = target ~ ., family = "binomial", data = dfInt5)
##
## Coefficients:
## (Intercept)           rm          age          dis          rad          tax
##  110.857039   -34.381951     0.117686    -5.463106     1.004539     0.005968
##     ptratio   TaxOver600     ptOver14  lstatOver12  IndusOver16    rmSquared
##    0.544714   -12.678775    -7.985525    -1.267043     5.786874     2.619120
##     nox_log      age_log      dis_log
##   20.568245    -3.867305    26.063405
##
## Degrees of Freedom: 465 Total (i.e. Null);  451 Residual
## Null Deviance:       645.9
## Residual Deviance: 132.8     AIC: 162.8
```

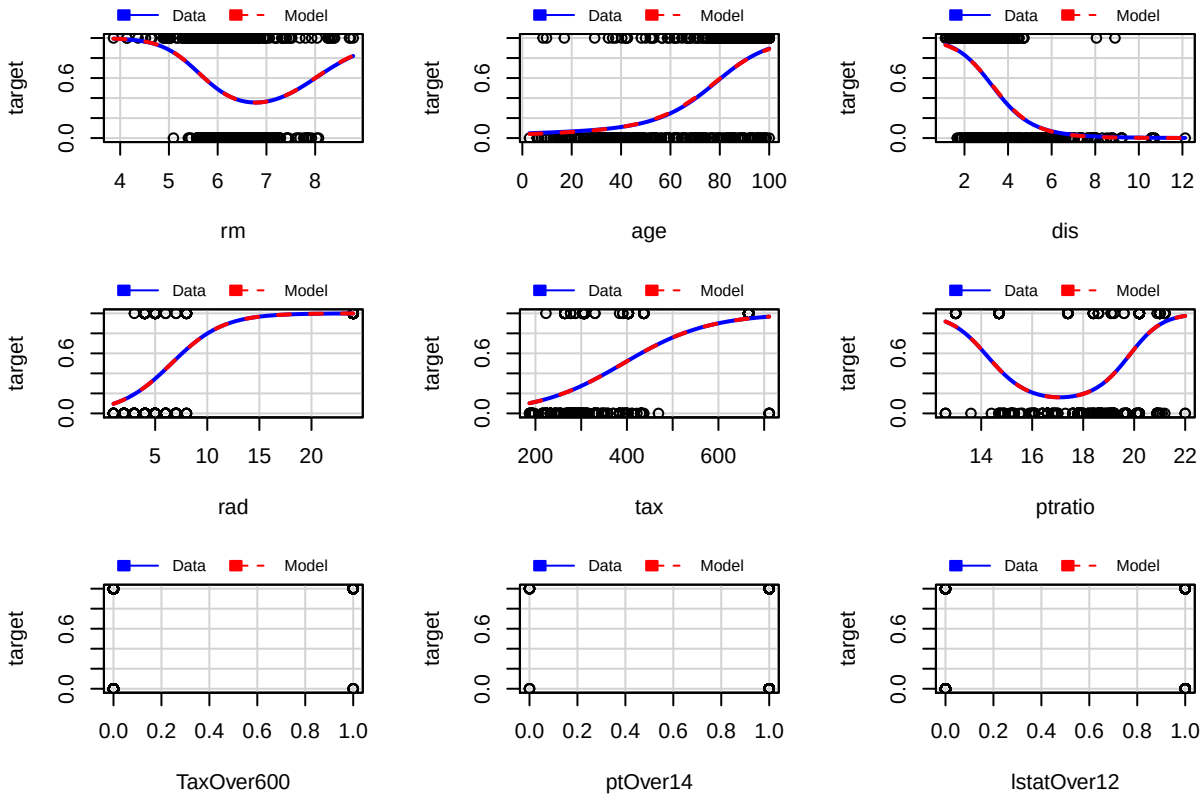



Normal Q−Q
Std. Pearson resid.
0
5
10
15
20
25
30
280
338
240
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(target ~ .)

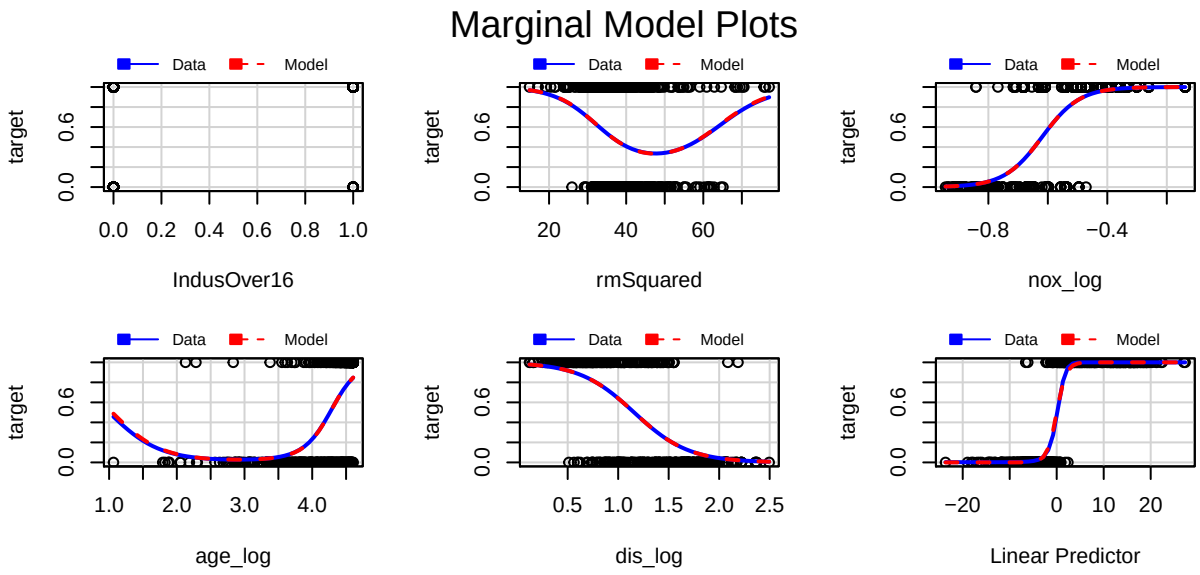

Scale−Location
280
338
240
√|Std. Pearson resid.|
0
1
2
3
4
5
−20
−10
0
10
20
Predicted values
glm(target ~ .)




```
## Error in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots) :
##   A term has fewer unique covariate combinations than specified maximum degrees of freedom
## Error in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots) :
##   A term has fewer unique covariate combinations than specified maximum degrees of freedom

## Error in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots) :
##   A term has fewer unique covariate combinations than specified maximum degrees of freedom
## Error in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots) :
##   A term has fewer unique covariate combinations than specified maximum degrees of freedom
```




```
## Error in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots) :
##   A term has fewer unique covariate combinations than specified maximum degrees of freedom
## Error in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots) :
##   A term has fewer unique covariate combinations than specified maximum degrees of freedom

## Error in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots) :
##   A term has fewer unique covariate combinations than specified maximum degrees of freedom
## Error in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots) :
##   A term has fewer unique covariate combinations than specified maximum degrees of freedom
```



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:

```
## [1] "Accuracy:  0.938131136044881"
```

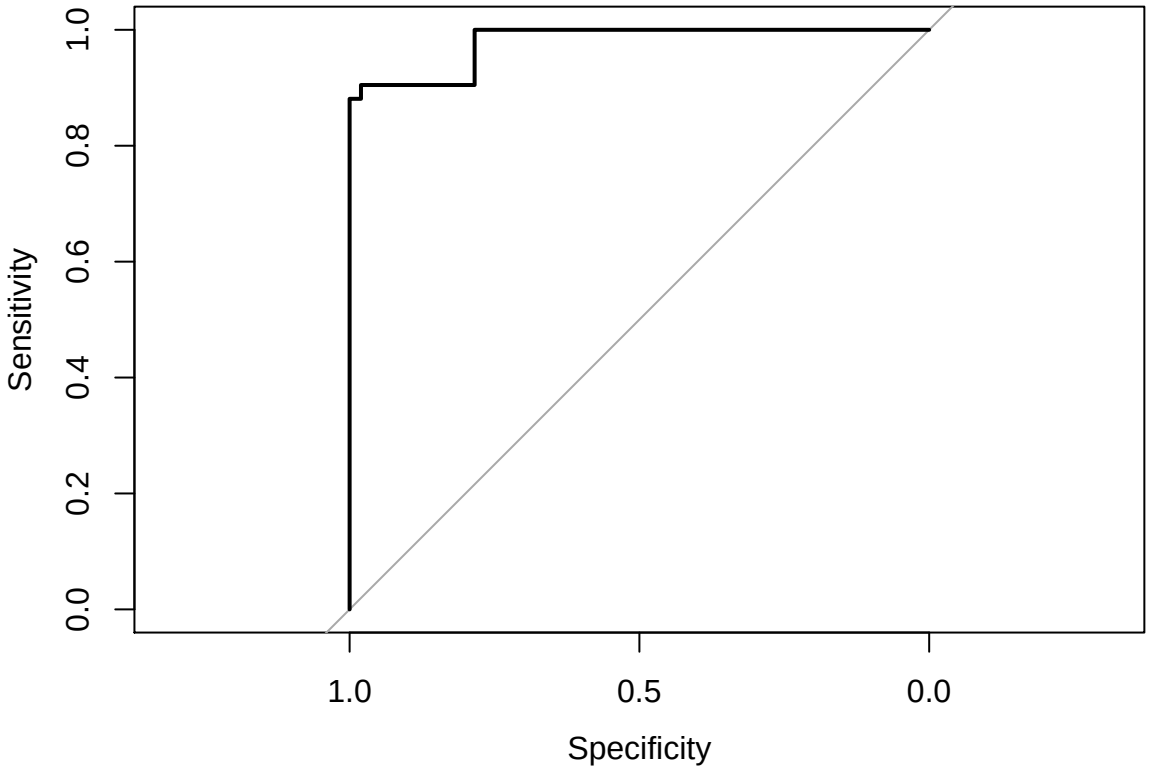
```
## [1] "AIC:  107.771064839552"
```

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.

```
##
## Call:
## glm(formula = fla, family = "binomial", data = train_reg)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.28851  -0.04629   0.00000   0.00246   2.89545
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 136.309426  37.120401   3.672 0.000241 ***
## rm          -42.040770  10.850297  -3.875 0.000107 ***
## age           0.182861   0.045894   3.984 6.76e-05 ***
## dis          -7.410362   1.956245  -3.788 0.000152 ***
```

```
## rad           1.331397   0.305130   4.363 1.28e-05 ***
## tax           0.006818   0.006790   1.004 0.315340
## ptratio       0.705146   0.215524   3.272 0.001069 **
## TaxOver600  -16.047916  15.655322  -1.025 0.305327
## ptOver14    -10.122579   4.520236  -2.239 0.025130 *
## lstatOver12  -2.404216   0.885392  -2.715 0.006619 **
## IndusOver16   7.390201   2.363650   3.127 0.001768 **
## rmSquared     3.240237   0.823325   3.936 8.30e-05 ***
## nox_log      30.859824   7.602695   4.059 4.93e-05 ***
## age_log      -5.694287   1.739125  -3.274 0.001060 **
## dis_log      35.677419   8.876253   4.019 5.83e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 514.313  on 370  degrees of freedom
## Residual deviance:  74.278  on 356  degrees of freedom
## AIC: 104.28
##
## Number of Fisher Scoring iterations: 13
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 50  5
##          1  1 37
##
##                Accuracy : 0.9355
##                  95% CI : (0.8648, 0.976)
##     No Information Rate : 0.5484
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8686
##
##  Mcnemar's Test P-Value : 0.2207
##
##             Sensitivity : 0.9804
##             Specificity : 0.8810
##          Pos Pred Value : 0.9091
##          Neg Pred Value : 0.9737
##              Prevalence : 0.5484
##          Detection Rate : 0.5376
##    Detection Prevalence : 0.5914
##       Balanced Accuracy : 0.9307
##
##        'Positive' Class : 0
##
```




```
## [1] "AUC:  0.978991596638655"
##
## Call:
## roc.default(response = dfPred_raw$class, predictor = dfPred_raw$predict_reg,     plot = TRUE)
##
## Data: dfPred_raw$predict_reg in 51 controls (dfPred_raw$class 0) < 42 cases (dfPred_raw$class 1).
## Area under the curve: 0.979

## [[1]]
##   (Intercept)            rm           age           dis           rad
## 136.309426211 -42.040769748   0.182860820  -7.410361996   1.331397197
##           tax       ptratio    TaxOver600      ptOver14   lstatOver12
##   0.006817627   0.705146353 -16.047916117 -10.122578863  -2.404216298
##   IndusOver16     rmSquared       nox_log       age_log       dis_log
##   7.390200997   3.240236654  30.859823620  -5.694286558  35.677419379
##
## [[2]]
## [1] 0.9354839
##
## [[3]]
## [1] 104.2782
```

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

```
##   predict(m, newdata = dfEval, type = "response")
```

```
## 1                                    0.999038618
## 2                                    0.990850719
## 3                                    0.999999851
## 4                                    0.006086738
## 5                                    0.004284532
## 6                                    0.278352083
```

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 94% accurate.