DATA 621 Homework 3

Critical Thinking Group 3 - Dominika Markowska-Desvallons & Shana Green

Due 10/31/2021

Introduction

In this homework assignment, you will explore, analyze and model a data set containing information on crime for various neighborhoods of a major city. Each record has a response variable indicating whether or not the crime rate is above the median crime rate (1) or not (0).

Objective

Your objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. You will provide classifications and probabilities for the evaluation data set using your binary logistic regression model. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

- **zn**: proportion of residential land zoned for large lots (over 25000 square feet) (predictor variable)
- **indus**: proportion of non-retail business acres per suburb (predictor variable)
- **chas**: a dummy var. for whether the suburb borders the Charles River (1) or not (0) (predictor variable)
- **nox**: nitrogen oxides concentration (parts per 10 million) (predictor variable)
- **rm**: average number of rooms per dwelling (predictor variable)
- **age**: proportion of owner-occupied units built prior to 1940 (predictor variable)
- **dis**: weighted mean of distances to five Boston employment centers (predictor variable)
- **rad**: index of accessibility to radial highways (predictor variable)
- **tax**: full-value property-tax rate per $10,000 (predictor variable)
- **ptratio**: pupil-teacher ratio by town (predictor variable)
- **black**: $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town (predictor variable)
- **lstat**: lower status of the population (percent) (predictor variable)
- **medv**: median value of owner-occupied homes in $1000s (predictor variable)
- **target**: whether the crime rate is above the median crime rate (1) or not (0) (response variable)

Dataset

```
## # A tibble: 466 x 14
##       zn indus  chas   nox    rm   age   dis   rad   tax ptratio lstat  medv
##    <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl> <int> <int>   <dbl> <dbl> <dbl>
##  1     0 19.6      0 0.605  7.93  96.2  2.05     5   403    14.7  3.7   50
##  2     0 19.6      1 0.871  5.40 100    1.32     5   403    14.7 26.8   13.4
##  3     0 18.1      0 0.74   6.48 100    1.98    24   666    20.2 18.8   15.4
##  4    30  4.93     0 0.428  6.39   7.8  7.04     6   300    16.6  5.19  23.7
```

```
##  5     0  2.46     0 0.488  7.16  92.2  2.70     3   193    17.8  4.82  37.9
##  6     0  8.56     0 0.52   6.78  71.3  2.86     5   384    20.9  7.67  26.5
##  7     0 18.1      0 0.693  5.45 100    1.49    24   666    20.2 30.6    5
##  8     0 18.1      0 0.693  4.52 100    1.66    24   666    20.2 37.0    7
##  9     0  5.19     0 0.515  6.32  38.1  6.46     5   224    20.2  5.68  22.2
## 10    80  3.64     0 0.392  5.88  19.1  9.22     1   315    16.4  9.25  20.9
## # ... with 456 more rows, and 2 more variables: target <int>, head(10) <dbl>
```

Structure of Dataset

```
## Rows: 466
## Columns: 13
## $ zn      <dbl> 0, 0, 0, 30, 0, 0, 0, 0, 0, 80, 22, 0, 0, 22, 0, 0, 100, 20, 0~
## $ indus   <dbl> 19.58, 19.58, 18.10, 4.93, 2.46, 8.56, 18.10, 18.10, 5.19, 3.6~
## $ chas    <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ nox     <dbl> 0.605, 0.871, 0.740, 0.428, 0.488, 0.520, 0.693, 0.693, 0.515,~
## $ rm      <dbl> 7.929, 5.403, 6.485, 6.393, 7.155, 6.781, 5.453, 4.519, 6.316,~
## $ age     <dbl> 96.2, 100.0, 100.0, 7.8, 92.2, 71.3, 100.0, 100.0, 38.1, 19.1,~
## $ dis     <dbl> 2.0459, 1.3216, 1.9784, 7.0355, 2.7006, 2.8561, 1.4896, 1.6582~
## $ rad     <int> 5, 5, 24, 6, 3, 5, 24, 24, 5, 1, 7, 5, 24, 7, 3, 3, 5, 5, 24, ~
## $ tax     <int> 403, 403, 666, 300, 193, 384, 666, 666, 224, 315, 330, 398, 66~
## $ ptratio <dbl> 14.7, 14.7, 20.2, 16.6, 17.8, 20.9, 20.2, 20.2, 20.2, 16.4, 19~
## $ lstat   <dbl> 3.70, 26.82, 18.85, 5.19, 4.82, 7.67, 30.59, 36.98, 5.68, 9.25~
## $ medv    <dbl> 50.0, 13.4, 15.4, 23.7, 37.9, 26.5, 5.0, 7.0, 22.2, 20.9, 24.8~
## $ target  <int> 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0,~
```

The train dataset contains 466 cases. Looking at the given variables, we can see that **chas** and **target** are dummy variables, based on the values given.

Summary Statistic

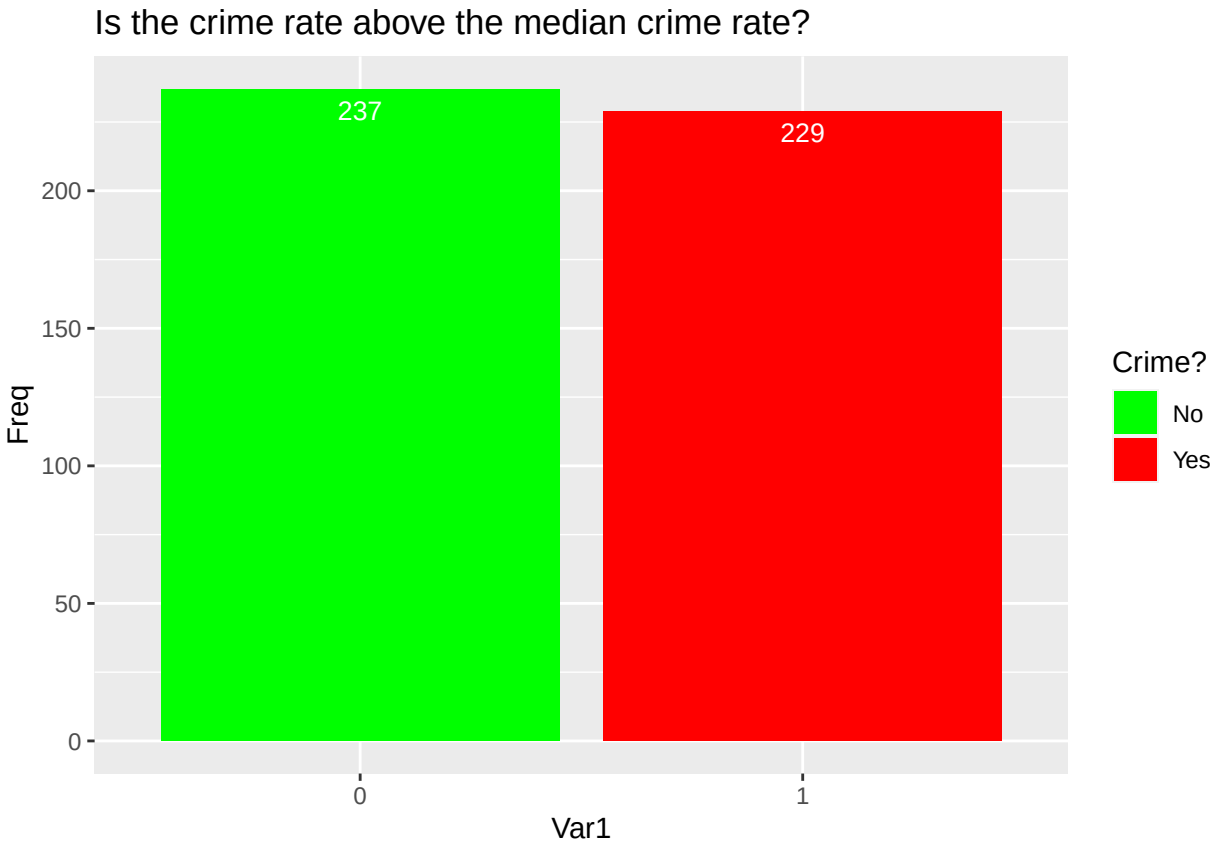
```
##        zn              indus             chas              nox        
##  Min.   :  0.00   Min.   : 0.460   Min.   :0.00000   Min.   :0.3890  
##  1st Qu.:  0.00   1st Qu.: 5.145   1st Qu.:0.00000   1st Qu.:0.4480  
##  Median :  0.00   Median : 9.690   Median :0.00000   Median :0.5380  
##  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
```

After reviewing the summary of the train dataset, we observed no missing NA values.

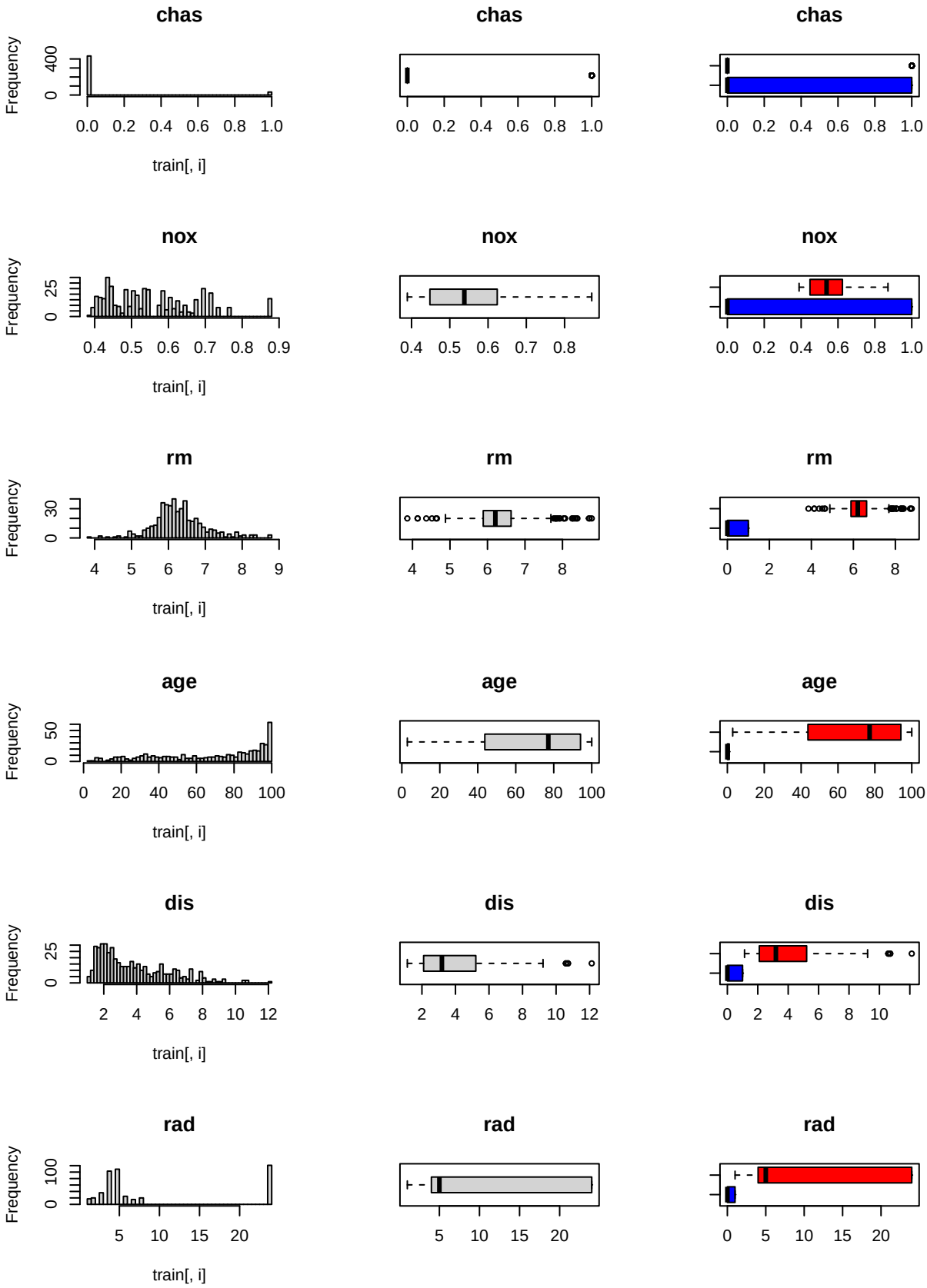
Data Exploration

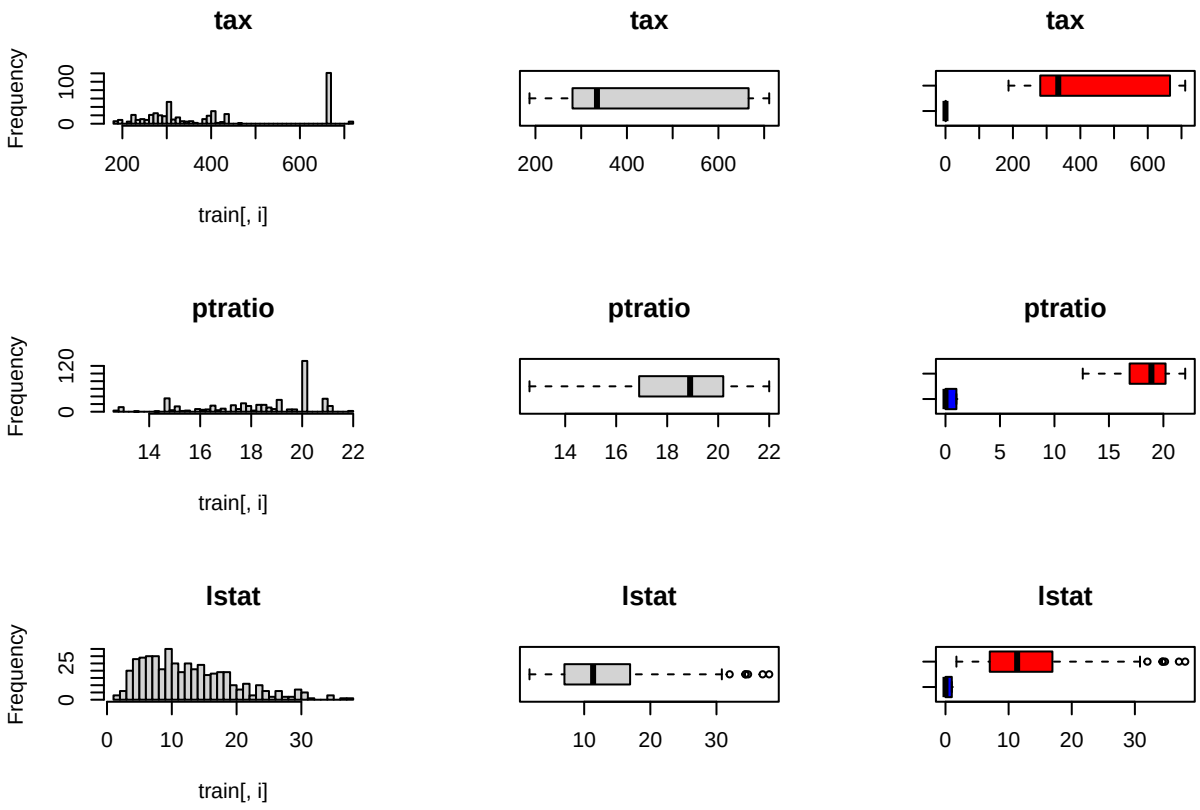
We wanted to take a closer look at the target variable to see if the crime rate was indeed above the median crime rate.

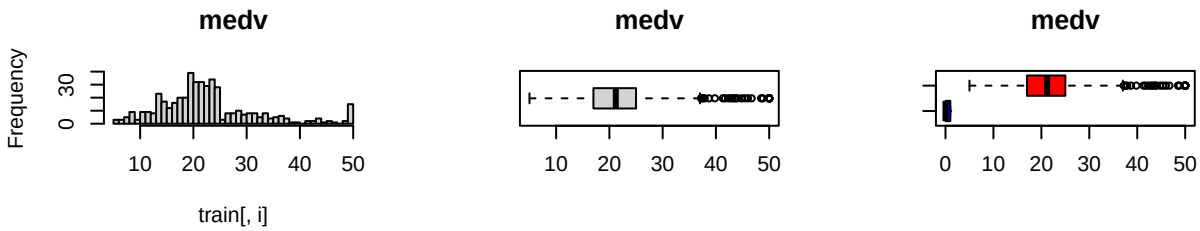


According to the histogram presented, the crime rate was not above the median crime rate.

Histogram and Box Plots



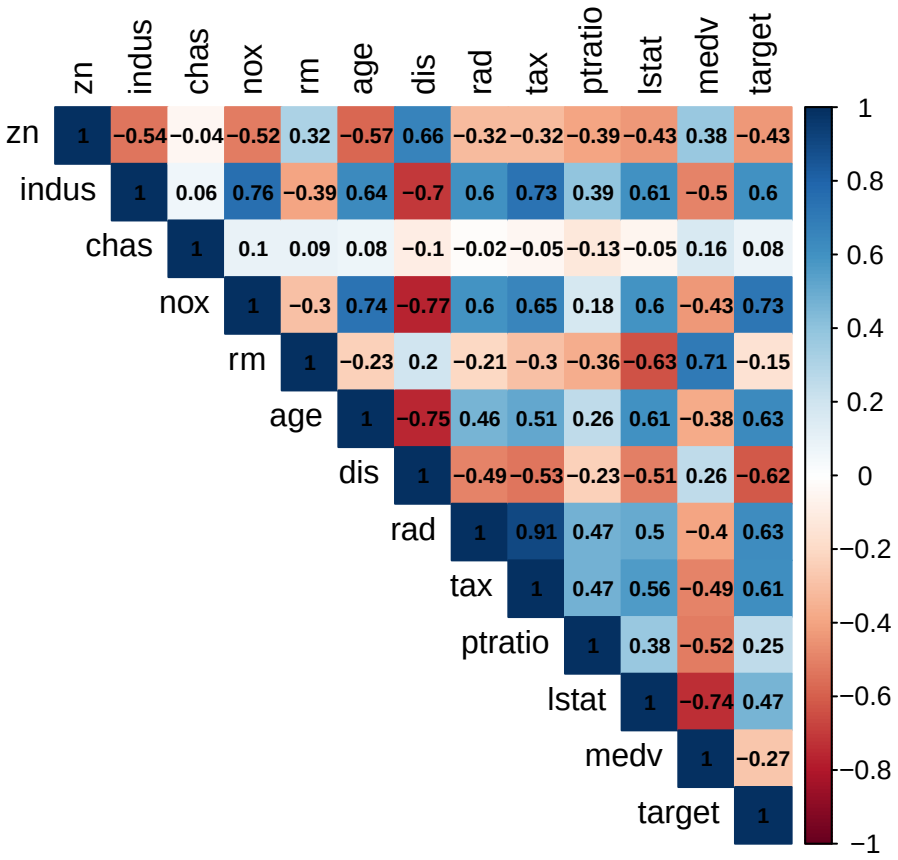

tax
Frequency
0
100
200
400
600
train[, i]
tax
200
400
600
tax
0
200
400
600
ptratio
Frequency
0
120
14
16
18
20
22
train[, i]
ptratio
14
16
18
20
22
ptratio
0
5
10
15
20
lstat
Frequency
0
25
0
10
20
30
train[, i]
lstat
10
20
30
lstat
0
10
20
30



Some of the independent variables appear to have a normal distribution, while others are skewed. The third column displayed in the box plots compared each independent variable to the **target** response variable.

Correlation Plot

We conducted a correlation plot to check for collinearity.



Certain variables relate to each other differently, and some actually correlated stronger than others. If we look at the target column, we can see how the independent variables correlate with the response variable. It appears that **indus**, **nox**, **age**, **rad**, **tax**, **ptratio**, and **lstat** have a positive correlation, whereas **zn**, **dis**, and **medv** have negative correlation.

Data Preparation

We initially thought it would be best not to perform any transformations, since there are no concentrations as a strong predictor of crime. However, we transformed the **chas** and **target** variables to factors since the columns are dummy variables.

```
levels(train$target) = make.names(levels(factor(train$target)))

train$chas = as.factor(train$chas)

eval$chas= as.factor(eval$chas)
eval$target = NULL

set.seed(12)
train1 = train %>% sample_n(., 40)
train1$chas = as.factor(train1$chas)

set.seed(30)
train2 = train %>% sample_n(., 40)
train2$chas = as.factor(train2$chas)
```

```r
set.seed(50)
train3 = train %>% sample_n(., 40)
train3$chas = as.factor(train3$chas)
```

Build Models