Data 621 - Homework 3

Group 2: William Aiken, Donald Butler, Michael Ippolito, Bharani Nittala, and Leticia Salazar

November 6, 2022

Contents

Overview:

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:

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).

Description:

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 radical 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)

Load Libraries:

These are the libraries used to explore, prepare, analyze and build our models

```
library(tidyverse)
library(caret)
library(pROC)
library(corrplot)
library(GGally)
library(psych)
library(car)
library(kableExtra)
library(gridExtra)
library(performance)
library(faraway)
library(jtools)
```

Load Data set:

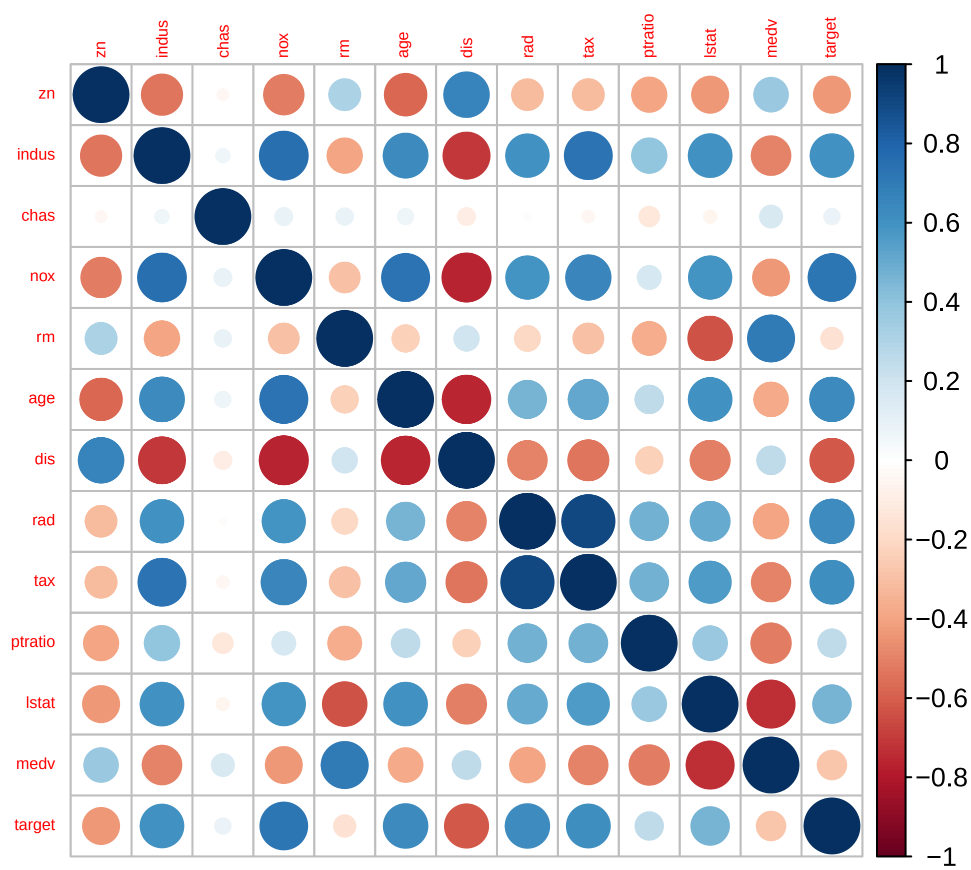
We have included the original data sets in our GitHub account and read from this location. Our data set includes 466 records and 13 variables.

```
## 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,~
## $ 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,~
```

Data Exploration:

The correlation plot below is measuring the degree of linear relationship within the training data set. The values in which this is measured falls between -1 and +1, with +1 being a stronger correlation.



To give more insight on our data set we used the `summary()` and `describe()` functions below:

`summary()`:

```
##        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
```

describe():

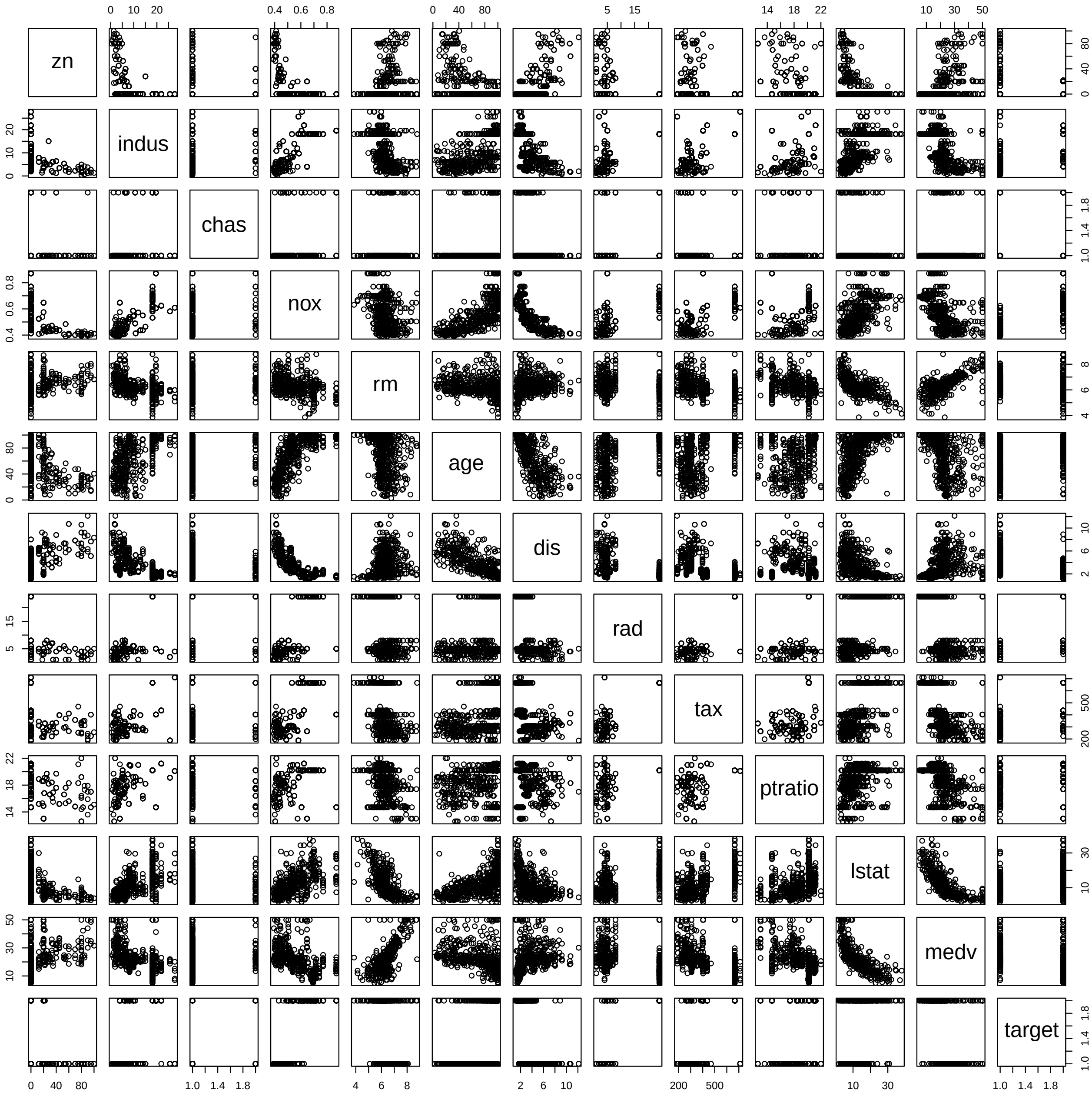
```
##         vars   n   mean     sd median trimmed    mad    min    max  range  skew
## zn         1 466  11.58  23.36   0.00    5.35   0.00   0.00 100.00 100.00  2.18
## indus      2 466  11.11   6.85   9.69   10.91   9.34   0.46  27.74  27.28  0.29
## chas       3 466   0.07   0.26   0.00    0.00   0.00   0.00   1.00   1.00  3.34
## nox        4 466   0.55   0.12   0.54    0.54   0.13   0.39   0.87   0.48  0.75
## rm         5 466   6.29   0.70   6.21    6.26   0.52   3.86   8.78   4.92  0.48
## age        6 466  68.37  28.32  77.15   70.96  30.02   2.90 100.00  97.10 -0.58
## dis        7 466   3.80   2.11   3.19    3.54   1.91   1.13  12.13  11.00  1.00
## rad        8 466   9.53   8.69   5.00    8.70   1.48   1.00  24.00  23.00  1.01
## tax        9 466 409.50 167.90 334.50  401.51 104.52 187.00 711.00 524.00  0.66
## ptratio   10 466  18.40   2.20  18.90   18.60   1.93  12.60  22.00   9.40 -0.75
## lstat     11 466  12.63   7.10  11.35   11.88   7.07   1.73  37.97  36.24  0.91
## medv      12 466  22.59   9.24  21.20   21.63   6.00   5.00  50.00  45.00  1.08
## target    13 466   0.49   0.50   0.00    0.49   0.00   0.00   1.00   1.00  0.03
##         kurtosis   se
## zn          3.81 1.08
## indus      -1.24 0.32
## chas        9.15 0.01
## nox        -0.04 0.01
## rm          1.54 0.03
## age        -1.01 1.31
## dis         0.47 0.10
## rad        -0.86 0.40
## tax        -1.15 7.78
## ptratio    -0.40 0.10
## lstat       0.50 0.33
## medv        1.37 0.43
## target     -2.00 0.02
```

Factor categorical variables:

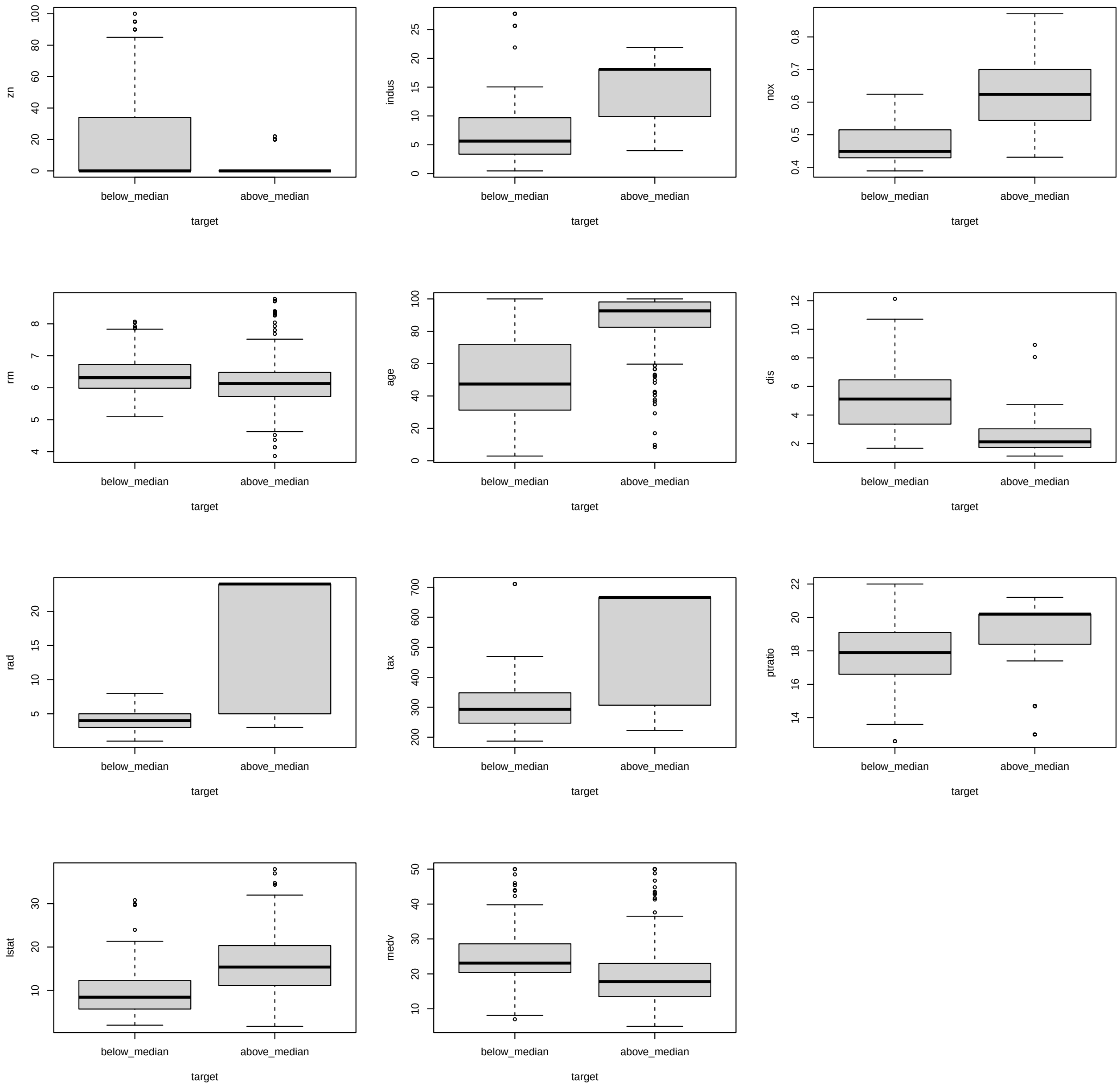
Since categorical variables enter differently into statistical model, storing them as factors insures that the functions will treat the data correctly.

```
# from the training data set; variable: target
dftrain$target <- factor(dftrain$target, levels = c(0, 1))
levels(dftrain$target) <- list(below_median = 0, above_median = 1)

# from the training data set; variable: chas
dftrain$chas <- factor(dftrain$chas)
levels(dftrain$chas) <- list(not_on_charles = 0, on_charles = 1)

# from the evaluation data set; variable: chas
dfeval$chas <- factor(dfeval$chas)
levels(dfeval$chas) <- list(not_on_charles = 0, on_charles = 1)
```

The plot matrix below consists of scatter plots corresponding to each data frame

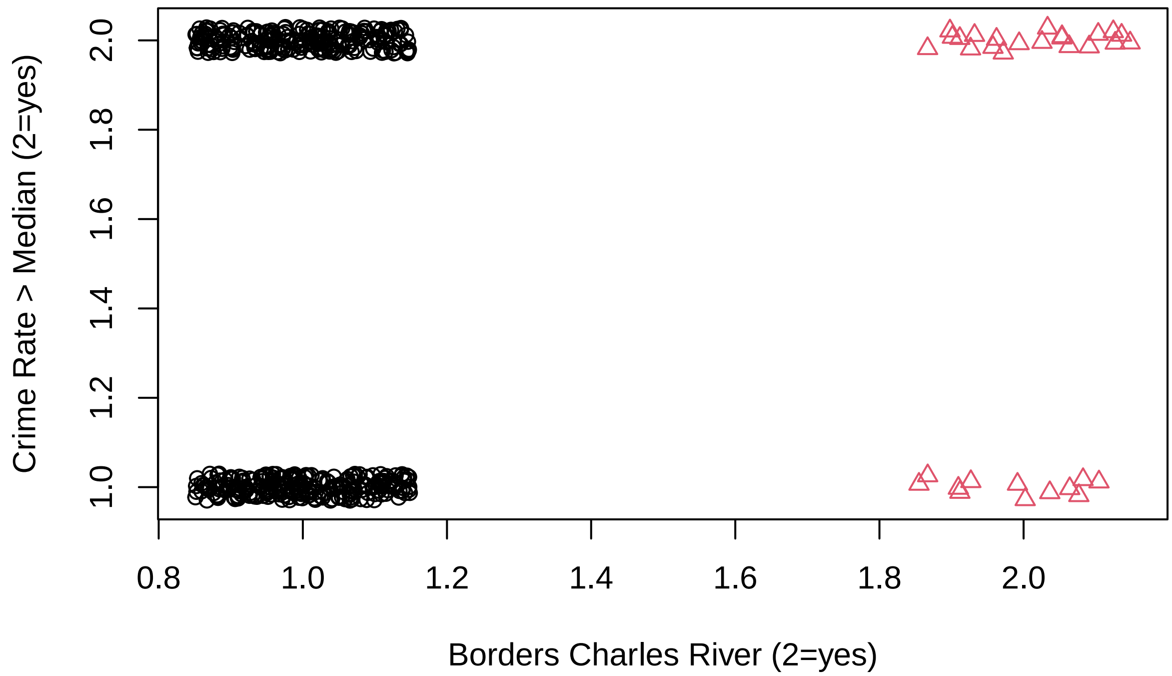


These boxplots show plenty of variables in our training data set with outliers. We also notice that variables `rad` and `tax` have a higher median for crime rate.



We created a contingency table to show the distribution of a variables `target` and `chas`. By using a jitter plot we are trying to visualize the relationship between these two variables.

```
##                 chas
## target          not_on_charles on_charles
##   below_median             225         12
##   above_median             208         21
```



Relationship of median crime rate to the following predictor variables:

Predictor	Definition	Relationship to Median Crime Rate
zn	Proportion of residential land zoned for large lots (over 25000 square feet)	negative
indus	Proportion of non-retail business acres per suburb	positive
chas	Dummy var. for whether the suburb borders the Charles River	unclear
nox	Nitrogen oxides concentration	positive
rm	Average number of rooms per dwelling	unclear
age	Proportion of owner-occupied units built prior to 1940	positive
dis	Weighted mean of distances to five Boston employment centers	negative
rad	Index of accessibility to radial highways	positive
tax	Full-value property-tax rate per $10,000	positive
ptratio	Pupil-teacher ratio by town	positive
lstat	Lower status of the population (percent)	positive
medv	Median value of owner-occupied homes in $1000s	negative

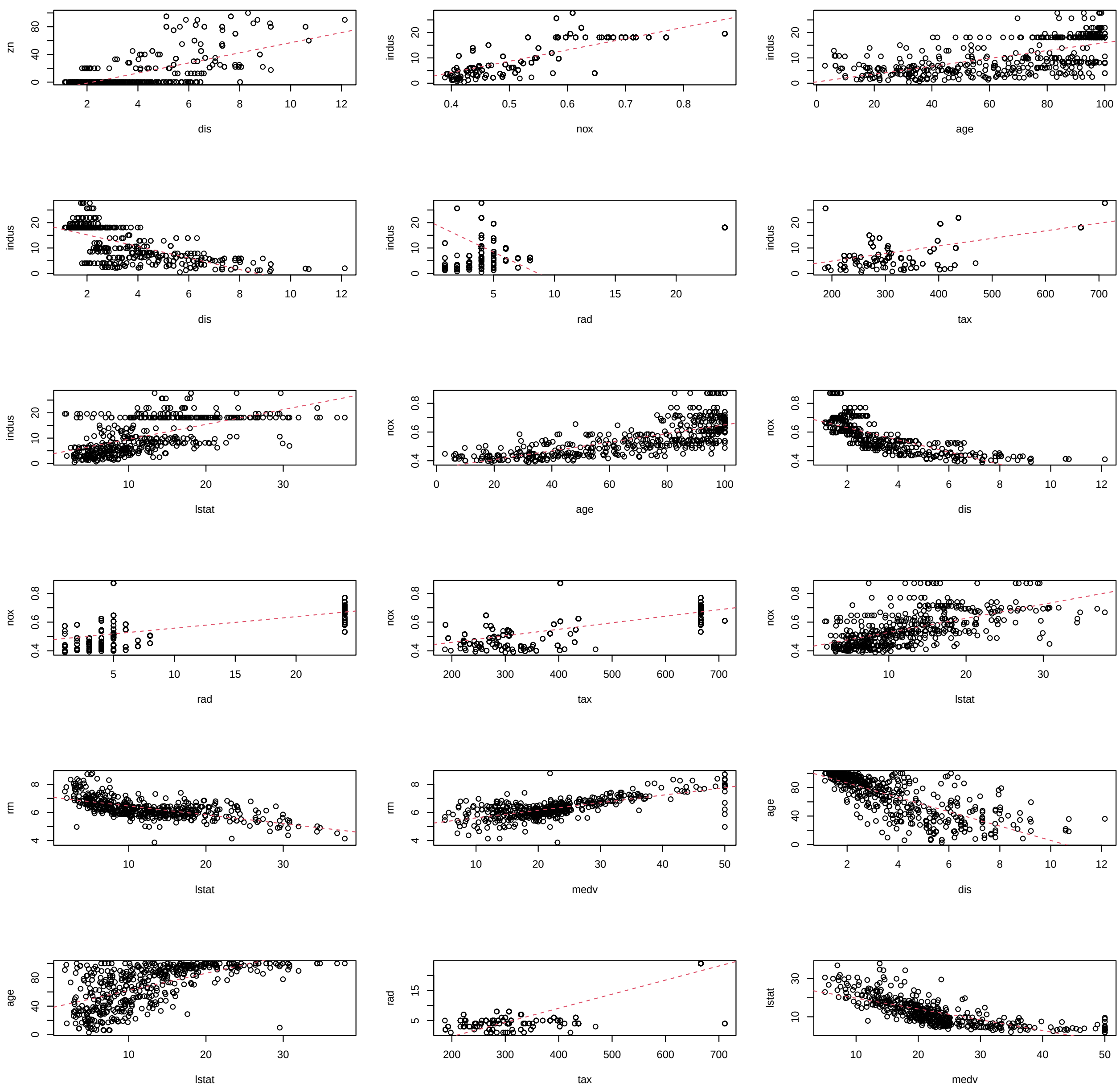
As indicated in the table, several predictors exhibit an inverse relationship with median crime rate. Based on the `zn` and `medv` variables, larger lot sizes and higher median home values correspond to a drop in crime rate, which is expected since larger lots and higher home values typically indicate higher economic status and, hence, lower crime. The same is true for the `dis` variable, which indicates that the farther a neighborhood is away from a major employment center, the lower the crime rate; this also makes sense, given that employment centers are often located in denser, more urban settings, which typically have higher rates of crime.

For the most part, variables exhibiting positive relationships with median crime rate also make intuitive sense. It would follow that neighborhoods having higher rates of industry (and, therefore, higher concentrations of pollutants like nitrogen oxides–`nox`–in the air) would also have higher crime rates. Likewise, neighborhoods with older homes (indicated by the `age` variable) located near radial highways (`rad` variable) and with a high pupil-to-teacher ratio (`ptratio` variable) could also be interpreted to have higher rates of crime.

Two variables didn't exhibit a clear relationships to crime rate: whether the neighborhood borders the Charles River (`chas`) and the average number of rooms per dwelling (`rm`). In addition, while the the `lstat` predictor exhibited a positive relationship with crime rate, the description of the variable ("lower status of the population") didn't clearly state what the data values represent.

Let's look for any significant relationships among predictor variables. We considered correlation values above 0.6 to be significant. We explore colinearity of predictor variables with the help of a correlation matrix:

```
##            zn indus   nox    rm   age   dis   rad   tax ptratio lstat  medv
## zn       1.00 -0.54 -0.52  0.32 -0.57  0.66 -0.32 -0.32   -0.39 -0.43  0.38
## indus   -0.54  1.00  0.76 -0.39  0.64 -0.70  0.60  0.73    0.39  0.61 -0.50
## nox     -0.52  0.76  1.00 -0.30  0.74 -0.77  0.60  0.65    0.18  0.60 -0.43
## rm       0.32 -0.39 -0.30  1.00 -0.23  0.20 -0.21 -0.30   -0.36 -0.63  0.71
## age     -0.57  0.64  0.74 -0.23  1.00 -0.75  0.46  0.51    0.26  0.61 -0.38
## dis      0.66 -0.70 -0.77  0.20 -0.75  1.00 -0.49 -0.53   -0.23 -0.51  0.26
## rad     -0.32  0.60  0.60 -0.21  0.46 -0.49  1.00  0.91    0.47  0.50 -0.40
## tax     -0.32  0.73  0.65 -0.30  0.51 -0.53  0.91  1.00    0.47  0.56 -0.49
## ptratio -0.39  0.39  0.18 -0.36  0.26 -0.23  0.47  0.47    1.00  0.38 -0.52
## lstat   -0.43  0.61  0.60 -0.63  0.61 -0.51  0.50  0.56    0.38  1.00 -0.74
## medv     0.38 -0.50 -0.43  0.71 -0.38  0.26 -0.40 -0.49   -0.52 -0.74  1.00
```



As shown in the graphs above, a number of significant correlations exist. Some of the stronger relationships are discussed here. First, the proportion of area zoned for large lots (`zn`) has a positive relationship with the distance to employment centers (`dis`), since it is more difficult to locate large lots close to the city center. A strong positive correlation exists between `indus` and `nox`, which is intuitively obvious. Likewise, tax rates in industrial areas are likely to be higher, as shown by the strong positive correlation of 0.73. Another strong correlation that makes obvious intuitive sense is that between median home values (`medv`) and the average number of rooms per dwelling (`rm`). The strongest positive correlation (0.91) exists between tax rate (`tax`) and the index of accessibility to radial highways (`rad`), which also corresponds to the fact that industrial areas are typically close to radial highways and also exhibit higher tax rates. The strongest negative correlation (-0.77) exists between nox and dis, indicating that the farther away from employment centers (and, hence, industrial areas), the lower the concentration of nitrogen oxide pollutants. Almost equally strong (-0.75) is the correlation between the age of dwellings (`age`) and the distance from employment centers (`dis`), indicating that the farther from urban centers, the newer the houses, which makes intuitive sense.

Data Preparation:

There are no missing values for our data sets

training data:

```
##      zn   indus    chas     nox      rm     age     dis     rad     tax ptratio
##       0       0       0       0       0       0       0       0       0       0
##   lstat    medv  target
##       0       0       0
```

evaluation data:

```
##      zn   indus    chas     nox      rm     age     dis     rad     tax ptratio
##       0       0       0       0       0       0       0       0       0       0
##   lstat    medv
##       0       0
```

The `rad` predictor is a categorical value and has some unknown meaning for values 1-8 and 24. We need to introduce dummy variables `rad1`, `rad2`, etc to indicate if the neighborhood is in which category. We will exclude `rad24` since we only need N-1 variables to represent each value.

Cleaned training data:

```
head(dftrain_clean)
```

```
##         target zn indus           chas   nox    rm   age    dis tax ptratio
## 1 above_median  0 19.58 not_on_charles 0.605 7.929  96.2 2.0459 403    14.7
## 2 above_median  0 19.58     on_charles 0.871 5.403 100.0 1.3216 403    14.7
## 3 above_median  0 18.10 not_on_charles 0.740 6.485 100.0 1.9784 666    20.2
## 4 below_median 30  4.93 not_on_charles 0.428 6.393   7.8 7.0355 300    16.6
## 5 below_median  0  2.46 not_on_charles 0.488 7.155  92.2 2.7006 193    17.8
## 6 below_median  0  8.56 not_on_charles 0.520 6.781  71.3 2.8561 384    20.9
##   lstat medv rad_1 rad_2 rad_3 rad_4 rad_5 rad_6 rad_7 rad_8
## 1  3.70 50.0     0     0     0     0     1     0     0     0
## 2 26.82 13.4     0     0     0     0     1     0     0     0
## 3 18.85 15.4     0     0     0     0     0     0     0     0
## 4  5.19 23.7     0     0     0     0     0     1     0     0
## 5  4.82 37.9     0     0     1     0     0     0     0     0
## 6  7.67 26.5     0     0     0     0     1     0     0     0
```

Cleaned evaluation data:

```
head(dfeval_clean)
```

```
##   zn indus           chas   nox    rm  age    dis tax ptratio lstat medv rad_1
## 1  0  7.07 not_on_charles 0.469 7.185 61.1 4.9671 242    17.8  4.03 34.7     0
## 2  0  8.14 not_on_charles 0.538 6.096 84.5 4.4619 307    21.0 10.26 18.2     0
```

```
## 3  0  8.14 not_on_charles 0.538 6.495 94.4 4.4547 307    21.0 12.80 18.4     0
## 4  0  8.14 not_on_charles 0.538 5.950 82.0 3.9900 307    21.0 27.71 13.2     0
## 5  0  5.96 not_on_charles 0.499 5.850 41.5 3.9342 279    19.2  8.77 21.0     0
## 6 25  5.13 not_on_charles 0.453 5.741 66.2 7.2254 284    19.7 13.15 18.7     0
##   rad_2 rad_3 rad_4 rad_5 rad_6 rad_7 rad_8
## 1     1     0     0     0     0     0     0
## 2     0     0     1     0     0     0     0
## 3     0     0     1     0     0     0     0
## 4     0     0     1     0     0     0     0
## 5     0     0     0     1     0     0     0
## 6     0     0     0     0     0     0     1
```

Model Building:

You can't calculate residuals for a factor so we created a dummy target variable for this model. Below are the results:

Observations	466
Dependent variable	target
Type	OLS linear regression

F(19,446)	55.99
R^2	0.70
Adj. R^2	0.69

We start our model building with the following models:

- **Logit Model**
- **Logit Model with Backward Elimination**

```
## Start:  AIC=156.98
## target ~ zn + indus + chas + nox + rm + age + dis + tax + ptratio +
##     lstat + medv + rad_1 + rad_2 + rad_3 + rad_4 + rad_5 + rad_6 +
##     rad_7 + rad_8
##
##           Df Deviance    AIC
## - ptratio  1   117.04 155.04
## - chas     1   117.06 155.06
## - lstat    1   118.07 156.07
## - age      1   118.44 156.44
## - rad_7    1   118.47 156.47
## - rm       1   118.50 156.50
## - indus    1   118.82 156.82
## <none>         116.98 156.98
## - rad_8    1   119.91 157.91
## - tax      1   120.42 158.42
## - dis      1   121.06 159.06
```

	Est.	S.E.	t val.	p
(Intercept)	-0.52	0.37	-1.40	0.16
zn	-0.00	0.00	-1.06	0.29
indus	-0.00	0.00	-0.55	0.58
chason_charles	-0.07	0.05	-1.38	0.17
nox	2.14	0.24	8.88	0.00
rm	0.01	0.03	0.21	0.83
age	0.00	0.00	3.73	0.00
dis	0.00	0.01	0.19	0.85
tax	-0.00	0.00	-0.55	0.58
ptratio	-0.01	0.01	-1.47	0.14
lstat	0.00	0.00	0.80	0.42
medv	0.01	0.00	2.89	0.00
rad_1	-0.55	0.11	-5.05	0.00
rad_2	-0.67	0.11	-6.06	0.00
rad_3	-0.56	0.10	-5.50	0.00
rad_4	-0.21	0.08	-2.63	0.01
rad_5	-0.50	0.08	-6.02	0.00
rad_6	-0.60	0.09	-6.79	0.00
rad_7	-0.38	0.11	-3.47	0.00
rad_8	0.06	0.10	0.56	0.58

Standard errors: OLS

Observations	466
Dependent variable	target
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(19)$	528.89
Pseudo-R^2 (Cragg-Uhler)	0.90
Pseudo-R^2 (McFadden)	0.82
AIC	156.98
BIC	239.87

```
## - medv     1   122.98 160.98
## - zn       1   125.74 163.74
## - rad_4    1   137.55 175.55
## - rad_2    1   141.93 179.93
## - rad_3    1   149.43 187.43
## - rad_1    1   156.00 194.00
## - rad_5    1   156.41 194.41
## - rad_6    1   177.89 215.89
## - nox      1   185.39 223.39
##
## Step:  AIC=155.04
## target ~ zn + indus + chas + nox + rm + age + dis + tax + lstat +
##     medv + rad_1 + rad_2 + rad_3 + rad_4 + rad_5 + rad_6 + rad_7 +
##     rad_8
##
```

	Est.	S.E.	z val.	p
(Intercept)	-11.23	1943.96	-0.01	1.00
zn	-0.16	0.07	-2.45	0.01
indus	-0.16	0.12	-1.34	0.18
chason_charles	-0.26	0.96	-0.27	0.79
nox	68.63	13.62	5.04	0.00
rm	-1.23	1.01	-1.21	0.23
age	0.02	0.02	1.19	0.23
dis	0.54	0.27	2.00	0.05
tax	-0.01	0.01	-1.74	0.08
ptratio	0.05	0.20	0.24	0.81
lstat	0.07	0.06	1.05	0.29
medv	0.22	0.10	2.20	0.03
rad_1	-44.04	5457.08	-0.01	0.99
rad_2	-44.49	5327.70	-0.01	0.99
rad_3	-26.20	1943.94	-0.01	0.99
rad_4	-21.82	1943.94	-0.01	0.99
rad_5	-24.54	1943.94	-0.01	0.99
rad_6	-26.66	1943.94	-0.01	0.99
rad_7	-17.04	1943.94	-0.01	0.99
rad_8	-18.40	1943.94	-0.01	0.99

Standard errors: MLE

```
##          Df Deviance    AIC
## - chas    1   117.11 153.11
## - lstat   1   118.14 154.15
## - age     1   118.46 154.46
## - rad_7   1   118.47 154.47
## - rm      1   118.53 154.53
## <none>        117.04 155.04
## - indus   1   119.35 155.35
## - rad_8   1   119.91 155.91
## - tax     1   120.42 156.42
## - dis     1   121.17 157.17
## - medv    1   124.02 160.02
## - zn      1   127.07 163.07
## - rad_4   1   137.67 173.67
## - rad_2   1   142.58 178.58
## - rad_3   1   149.60 185.60
## - rad_1   1   156.42 192.42
## - rad_5   1   158.34 194.34
## - rad_6   1   179.73 215.73
## - nox     1   187.89 223.89
##
## Step:  AIC=153.11
## target ~ zn + indus + nox + rm + age + dis + tax + lstat + medv +
##     rad_1 + rad_2 + rad_3 + rad_4 + rad_5 + rad_6 + rad_7 + rad_8
##
##          Df Deviance    AIC
## - lstat   1   118.17 152.17
## - age     1   118.46 152.46
```

```
## - rad_7  1   118.50 152.50
## - rm     1   118.54 152.54
## <none>       117.11 153.11
## - rad_8  1   119.94 153.94
## - indus  1   120.17 154.17
## - tax    1   120.66 154.66
## - dis    1   121.41 155.41
## - medv   1   124.07 158.07
## - zn     1   127.10 161.10
## - rad_4  1   138.03 172.03
## - rad_2  1   144.31 178.31
## - rad_3  1   152.05 186.05
## - rad_1  1   156.55 190.55
## - rad_5  1   159.20 193.20
## - rad_6  1   180.63 214.63
## - nox    1   190.43 224.43
##
## Step:  AIC=152.17
## target ~ zn + indus + nox + rm + age + dis + tax + medv + rad_1 +
##     rad_2 + rad_3 + rad_4 + rad_5 + rad_6 + rad_7 + rad_8
##
##         Df Deviance    AIC
## - rad_7  1   119.97 151.97
## <none>       118.17 152.17
## - age    1   120.74 152.74
## - indus  1   120.93 152.93
## - rm     1   121.05 153.05
## - rad_8  1   121.62 153.62
## - tax    1   121.73 153.73
## - dis    1   122.35 154.35
## - medv   1   125.18 157.18
## - zn     1   127.58 159.58
## - rad_4  1   138.44 170.44
## - rad_2  1   145.60 177.60
## - rad_3  1   152.90 184.90
## - rad_1  1   159.16 191.16
## - rad_5  1   160.76 192.76
## - rad_6  1   180.95 212.95
## - nox    1   191.60 223.60
##
## Step:  AIC=151.97
## target ~ zn + indus + nox + rm + age + dis + tax + medv + rad_1 +
##     rad_2 + rad_3 + rad_4 + rad_5 + rad_6 + rad_8
##
##         Df Deviance    AIC
## - rad_8  1   121.75 151.75
## <none>       119.97 151.97
## - tax    1   122.40 152.40
## - age    1   122.45 152.45
## - rm     1   123.24 153.24
## - dis    1   124.24 154.24
## - indus  1   125.03 155.03
## - medv   1   127.81 157.81
## - zn     1   140.31 170.31
```

```
## - rad_4  1   141.71 171.71
## - rad_2  1   150.79 180.79
## - rad_3  1   159.80 189.80
## - rad_1  1   164.34 194.34
## - rad_5  1   172.56 202.56
## - rad_6  1   186.82 216.82
## - nox    1   208.97 238.97
##
## Step:  AIC=151.75
## target ~ zn + indus + nox + rm + age + dis + tax + medv + rad_1 +
##     rad_2 + rad_3 + rad_4 + rad_5 + rad_6
##
##         Df Deviance    AIC
## - tax    1   123.70 151.70
## <none>       121.75 151.75
## - age    1   124.18 152.18
## - rm     1   124.86 152.86
## - dis    1   125.30 153.30
## - indus  1   127.23 155.23
## - medv   1   129.59 157.59
## - zn     1   140.72 168.72
## - rad_4  1   142.60 170.60
## - rad_2  1   152.11 180.11
## - rad_1  1   165.96 193.96
## - rad_3  1   165.98 193.98
## - rad_5  1   189.53 217.53
## - rad_6  1   194.20 222.20
## - nox    1   209.71 237.71
##
## Step:  AIC=151.7
## target ~ zn + indus + nox + rm + age + dis + medv + rad_1 + rad_2 +
##     rad_3 + rad_4 + rad_5 + rad_6
##
##         Df Deviance    AIC
## <none>       123.70 151.70
## - age    1   126.82 152.82
## - rm     1   128.16 154.16
## - dis    1   128.76 154.76
## - medv   1   135.26 161.26
## - zn     1   141.21 167.21
## - rad_4  1   144.35 170.35
## - indus  1   145.93 171.93
## - rad_2  1   162.33 188.33
## - rad_3  1   166.25 192.25
## - rad_1  1   168.22 194.22
## - rad_6  1   194.71 220.71
## - rad_5  1   202.57 228.57
## - nox    1   213.94 239.94
```

- **Logit Minimal Model with forward elimination**
- **Forward Elimination**

```
## Start:  AIC=647.88
```

Observations	466
Dependent variable	target
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(13)$	522.18
Pseudo-R² (Cragg-Uhler)	0.90
Pseudo-R² (McFadden)	0.81
AIC	151.70
BIC	209.72

	Est.	S.E.	z val.	p
(Intercept)	-27.05	6.14	-4.40	0.00
zn	-0.18	0.06	-3.25	0.00
indus	-0.29	0.07	-4.09	0.00
nox	69.44	12.13	5.72	0.00
rm	-1.70	0.82	-2.07	0.04
age	0.02	0.01	1.75	0.08
dis	0.58	0.27	2.18	0.03
medv	0.24	0.08	3.07	0.00
rad_1	-24.31	1917.60	-0.01	0.99
rad_2	-22.65	2049.05	-0.01	0.99
rad_3	-9.11	2.15	-4.24	0.00
rad_4	-4.43	1.42	-3.11	0.00
rad_5	-7.36	1.50	-4.89	0.00
rad_6	-10.00	2.03	-4.92	0.00

Standard errors: MLE

Observations	466
Dependent variable	target
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(0)$	-0.00
Pseudo-R² (Cragg-Uhler)	0.00
Pseudo-R² (McFadden)	0.00
AIC	647.88
BIC	652.02

	Est.	S.E.	z val.	p
(Intercept)	-0.03	0.09	-0.37	0.71

Standard errors: MLE

```
## target ~ 1
##
##            Df Deviance    AIC
```

```
## + nox      1   292.01 296.01
## + dis      1   409.50 413.50
## + age      1   424.75 428.75
## + tax      1   442.38 446.38
## + indus    1   453.23 457.23
## + zn       1   518.46 522.46
## + lstat    1   528.01 532.01
## + rad_3    1   603.67 607.67
## + medv     1   609.62 613.62
## + ptratio  1   615.64 619.64
## + rad_2    1   617.96 621.96
## + rad_1    1   622.27 626.27
## + rad_6    1   624.91 628.91
## + rm       1   634.82 638.82
## + rad_7    1   636.98 640.98
## + rad_8    1   637.41 641.41
## + rad_5    1   640.49 644.49
## + chas     1   642.86 646.86
## + rad_4    1   643.69 647.69
## <none>         645.88 647.88
##
## Step:  AIC=296.01
## target ~ nox
##
##           Df Deviance    AIC
## + rad_8    1   254.85 260.85
## + rad_6    1   268.66 274.66
## + rad_2    1   274.39 280.39
## + rad_1    1   278.70 284.70
## + rad_5    1   279.56 285.56
## + rad_4    1   284.30 290.30
## + rm       1   284.63 290.63
## + medv     1   285.86 291.86
## + indus    1   288.11 294.11
## + zn       1   288.29 294.29
## + tax      1   288.40 294.40
## + chas     1   288.47 294.47
## + rad_3    1   289.72 295.72
## <none>         292.01 296.01
## + ptratio  1   290.13 296.13
## + rad_7    1   290.53 296.53
## + age      1   290.62 296.62
## + dis      1   290.91 296.91
## + lstat    1   291.93 297.93
##
## Step:  AIC=260.85
## target ~ nox + rad_8
##
##           Df Deviance    AIC
## + rad_6    1   234.80 242.80
## + rad_4    1   235.47 243.47
## + rad_2    1   239.16 247.16
## + rad_1    1   243.22 251.22
## + rad_5    1   249.11 257.11
```

```
## + ptratio  1   250.38 258.38
## + tax      1   250.41 258.41
## + rad_7    1   251.21 259.21
## + dis      1   252.33 260.33
## + zn       1   252.34 260.33
## + indus    1   252.78 260.78
## <none>         254.85 260.85
## + lstat    1   254.24 262.24
## + medv     1   254.32 262.32
## + rad_3    1   254.38 262.38
## + rm       1   254.45 262.45
## + chas     1   254.46 262.46
## + age      1   254.51 262.51
##
## Step:  AIC=242.8
## target ~ nox + rad_8 + rad_6
##
##           Df Deviance    AIC
## + rad_2    1   215.90 225.90
## + rad_1    1   220.72 230.72
## + rad_4    1   221.90 231.90
## + rad_5    1   223.12 233.12
## + indus    1   227.51 237.51
## + tax      1   229.81 239.81
## + ptratio  1   231.20 241.20
## + rad_7    1   231.22 241.22
## + zn       1   231.72 241.72
## <none>         234.80 242.80
## + dis      1   233.69 243.69
## + lstat    1   233.81 243.81
## + rad_3    1   234.18 244.18
## + medv     1   234.48 244.48
## + chas     1   234.69 244.69
## + rm       1   234.79 244.79
## + age      1   234.79 244.79
##
## Step:  AIC=225.9
## target ~ nox + rad_8 + rad_6 + rad_2
##
##           Df Deviance    AIC
## + rad_5    1   198.67 210.67
## + rad_1    1   199.77 211.77
## + rad_4    1   206.67 218.67
## + rad_7    1   212.18 224.18
## + ptratio  1   212.31 224.31
## + zn       1   212.52 224.52
## <none>         215.90 225.90
## + lstat    1   214.53 226.53
## + indus    1   215.00 227.00
## + tax      1   215.18 227.18
## + rad_3    1   215.22 227.22
## + medv     1   215.48 227.48
## + dis      1   215.62 227.62
## + chas     1   215.88 227.88
```

```
## + age      1   215.89 227.89
## + rm       1   215.90 227.90
##
## Step:  AIC=210.67
## target ~ nox + rad_8 + rad_6 + rad_2 + rad_5
##
##           Df Deviance    AIC
## + rad_1    1   175.59 189.59
## + indus    1   195.32 209.32
## + rad_3    1   195.99 209.99
## + medv     1   196.14 210.14
## + rad_7    1   196.56 210.56
## <none>         198.67 210.67
## + zn       1   196.69 210.69
## + rad_4    1   197.98 211.98
## + rm       1   198.14 212.14
## + dis      1   198.27 212.27
## + chas     1   198.35 212.35
## + lstat    1   198.41 212.41
## + tax      1   198.62 212.62
## + ptratio  1   198.65 212.65
## + age      1   198.66 212.66
##
## Step:  AIC=189.59
## target ~ nox + rad_8 + rad_6 + rad_2 + rad_5 + rad_1
##
##           Df Deviance    AIC
## + indus    1   167.39 183.39
## + rad_3    1   172.00 188.00
## + medv     1   173.55 189.55
## <none>         175.59 189.59
## + rad_7    1   173.63 189.63
## + tax      1   173.91 189.91
## + rm       1   174.20 190.20
## + zn       1   174.46 190.46
## + rad_4    1   174.99 190.99
## + dis      1   175.05 191.05
## + lstat    1   175.34 191.34
## + chas     1   175.46 191.46
## + ptratio  1   175.51 191.51
## + age      1   175.56 191.56
##
## Step:  AIC=183.39
## target ~ nox + rad_8 + rad_6 + rad_2 + rad_5 + rad_1 + indus
##
##           Df Deviance    AIC
## + rad_3    1   159.81 177.81
## <none>         167.39 183.39
## + rad_7    1   165.68 183.68
## + rad_4    1   166.18 184.18
## + zn       1   166.19 184.19
## + chas     1   166.35 184.35
## + medv     1   166.43 184.43
## + dis      1   166.69 184.69
```

```
## + rm       1   166.72 184.72
## + tax      1   166.76 184.76
## + age      1   167.19 185.19
## + lstat    1   167.29 185.29
## + ptratio  1   167.30 185.30
##
## Step:  AIC=177.81
## target ~ nox + rad_8 + rad_6 + rad_2 + rad_5 + rad_1 + indus +
##     rad_3
##
##           Df Deviance    AIC
## + rad_4    1   148.84 168.84
## + medv     1   154.38 174.38
## + zn       1   156.86 176.86
## + rm       1   157.42 177.42
## <none>         159.81 177.81
## + chas     1   158.65 178.65
## + lstat    1   159.01 179.01
## + rad_7    1   159.43 179.43
## + ptratio  1   159.59 179.59
## + tax      1   159.75 179.75
## + dis      1   159.81 179.81
## + age      1   159.81 179.81
##
## Step:  AIC=168.85
## target ~ nox + rad_8 + rad_6 + rad_2 + rad_5 + rad_1 + indus +
##     rad_3 + rad_4
##
##           Df Deviance    AIC
## + zn       1   137.67 159.67
## + rad_7    1   143.52 165.52
## + tax      1   145.42 167.42
## + chas     1   145.82 167.82
## + medv     1   145.85 167.85
## <none>         148.84 168.84
## + rm       1   148.51 170.51
## + ptratio  1   148.56 170.56
## + dis      1   148.73 170.73
## + lstat    1   148.82 170.82
## + age      1   148.84 170.84
##
## Step:  AIC=159.67
## target ~ nox + rad_8 + rad_6 + rad_2 + rad_5 + rad_1 + indus +
##     rad_3 + rad_4 + zn
##
##           Df Deviance    AIC
## + tax      1   129.89 153.89
## + medv     1   131.67 155.67
## + ptratio  1   134.68 158.68
## + chas     1   135.07 159.07
## <none>         137.67 159.67
## + dis      1   136.16 160.16
## + rm       1   136.33 160.33
## + rad_7    1   137.13 161.13
```

```
## + lstat    1   137.66 161.66
## + age      1   137.67 161.67
##
## Step:  AIC=153.89
## target ~ nox + rad_8 + rad_6 + rad_2 + rad_5 + rad_1 + indus +
##     rad_3 + rad_4 + zn + tax
##
##           Df Deviance    AIC
## + medv     1   126.39 152.39
## + rad_7    1   127.23 153.23
## <none>         129.89 153.89
## + ptratio  1   128.52 154.52
## + rm       1   129.06 155.06
## + dis      1   129.07 155.07
## + chas     1   129.67 155.67
## + lstat    1   129.74 155.74
## + age      1   129.88 155.88
##
## Step:  AIC=152.39
## target ~ nox + rad_8 + rad_6 + rad_2 + rad_5 + rad_1 + indus +
##     rad_3 + rad_4 + zn + tax + medv
##
##           Df Deviance    AIC
## + lstat    1   123.59 151.59
## + rad_7    1   124.34 152.34
## <none>         126.39 152.39
## + dis      1   124.50 152.50
## + rm       1   125.08 153.08
## + age      1   126.03 154.03
## + ptratio  1   126.30 154.30
## + chas     1   126.33 154.33
##
## Step:  AIC=151.59
## target ~ nox + rad_8 + rad_6 + rad_2 + rad_5 + rad_1 + indus +
##     rad_3 + rad_4 + zn + tax + medv + lstat
##
##           Df Deviance    AIC
## + dis      1   120.57 150.57
## <none>         123.59 151.59
## + rad_7    1   122.05 152.05
## + rm       1   123.19 153.19
## + age      1   123.53 153.53
## + chas     1   123.56 153.56
## + ptratio  1   123.56 153.56
##
## Step:  AIC=150.57
## target ~ nox + rad_8 + rad_6 + rad_2 + rad_5 + rad_1 + indus +
##     rad_3 + rad_4 + zn + tax + medv + lstat + dis
##
##           Df Deviance    AIC
## <none>         120.57 150.57
## + rad_7    1   119.15 151.15
## + rm       1   119.78 151.78
## + age      1   120.03 152.03
```

```
## + ptratio  1   120.36 152.36
## + chas     1   120.55 152.55
```

Observations	466
Dependent variable	target
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(14)$	525.30
Pseudo-R^2 (Cragg-Uhler)	0.90
Pseudo-R^2 (McFadden)	0.81
AIC	150.57
BIC	212.74

	Est.	S.E.	z val.	p
(Intercept)	-32.69	6.57	-4.97	0.00
nox	72.84	12.05	6.05	0.00
rad_8	-2.26	2.05	-1.10	0.27
rad_6	-11.44	2.15	-5.32	0.00
rad_2	-26.34	1870.40	-0.01	0.99
rad_5	-8.77	1.69	-5.20	0.00
rad_1	-26.65	1874.33	-0.01	0.99
indus	-0.19	0.09	-2.04	0.04
rad_3	-9.88	2.14	-4.61	0.00
rad_4	-6.22	1.66	-3.75	0.00
zn	-0.20	0.05	-3.75	0.00
tax	-0.01	0.00	-1.92	0.05
medv	0.13	0.05	2.60	0.01
lstat	0.12	0.06	1.99	0.05
dis	0.42	0.24	1.72	0.08

Standard errors: MLE

Select Models:

We evaluated the performance our models to decide which model should we use:

- **Linear Model Accuracy**

```
##     pred
## true   0   1
##    0 225  12
##    1  19 210
```

Accuracy: $\frac{210+225}{466} = 93\%$

Classification Error Rate: $\frac{12+19}{466} = 7\%$

Precision: $\frac{210}{210+12} = 95\%$

Sensitivity: $\frac{210}{210+19} = 92\%$

Specificity: $\frac{225}{225+12} = 95\%$

F1 Score: $\frac{2*.95*.92}{.95+.92} = 93\%$

- **Logit Model Prediction Accuracy**

```
##               pred
## true             0   1
##   below_median 233   4
##   above_median  10 219
```

Accuracy: $\frac{219+233}{466} = 97\%$

Classification Error Rate: $\frac{4+10}{466} = 3\%$

Precision: $\frac{219}{219+4} = 98\%$

Sensitivity: $\frac{219}{219+10} = 96\%$

Specificity: $\frac{233}{233+4} = 98\%$

F1 Score: $\frac{2*.98*.96}{.98+.96} = 97\%$

- **Logit Model with Forward Elimination Prediction Accuracy**

```
##               pred
## true             0   1
##   below_median 234   3
##   above_median   9 220
```

Accuracy: $\frac{220+234}{466} = 97\%$

Classification Error Rate: $\frac{5+12}{466} = 3\%$

Precision: $\frac{220}{220+3} = 99\%$

Sensitivity: $\frac{220}{220+9} = 96\%$

Specificity: $\frac{234}{234+3} = 98\%$

F1 Score: $\frac{2*.99*.96}{.99+.96} = 97\%$

- **Logit Model with Backward Elimination Prediction Accuracy**

```
##               pred
## true             0   1
##   below_median 232   5
##   above_median  12 217
```

Accuracy: $\frac{217+232}{466} = 96\%$

Classification Error Rate: $\frac{5+12}{466} = 4\%$

Precision: $\frac{217}{217+5} = 98\%$

Sensitivity: $\frac{217}{217+12} = 95\%$

Specificity: $\frac{232}{232+5} = 98\%$

F1 Score: $\frac{2*.98*.95}{.98+.95} = 96\%$

Model AUCs

- **Linear Model**

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Area under the curve: 0.9332
```

- **Logit Model**

```
## Setting levels: control = below_median, case = above_median

## Setting direction: controls < cases

## Area under the curve: 0.9697
```

- **Logit Model with Forward Elimination**

```
## Setting levels: control = below_median, case = above_median

## Setting direction: controls < cases

## Area under the curve: 0.974
```

- **Logit Model with Backward Elimination**

```
## Setting levels: control = below_median, case = above_median

## Setting direction: controls < cases

## Area under the curve: 0.9633
```

Findings

All of our Logistic models had a better AUC than our Linear Model with 'Logit with Forward Elimination' having the best (0.974).

Now lets make predictions with the evaluation data:

- **Linear Model**

```
##  [1]  0.044471432  0.549248991  0.590432182  0.549527352  0.100429012
##  [6]  0.602991759  0.668344513  0.036611979  0.004762499 -0.072870822
## [11]  0.249882596  0.236346164  0.697204745  0.529904005  0.501715088
## [16]  0.470213603  0.694800279  0.993663602  0.016477435 -0.148806180
## [21] -0.251508242  0.272093476  0.511840154  0.103025106  0.054423115
## [26]  0.174900012  0.209310715  1.142008184  1.096606893  0.794790626
## [31]  1.129376086  1.148942218  1.125367437  1.197800888  1.145515096
## [36]  1.072375460  1.096093208  0.927138131  0.234910336  0.338470879
```

- **Logit Model**

```
##  [1] -23.1837846   1.8909140   1.7995517   1.8557650  -3.2553526  -2.0334222
##  [7]  -2.5229875  -7.1842002  -7.7530724 -26.2911541 -19.5120234 -19.5664789
## [13]   3.5376684   1.2190996   0.4821279  -0.7066933   3.2735770   6.1682811
## [19]  -2.2482983 -41.5077624 -40.2884118  -1.1204161   0.8969585  -3.5445734
## [25]  -3.6931577  -0.7112376 -15.5439099  34.5405444  28.6931706  19.9909294
## [31]  30.8838151  31.5823360  30.7894623  31.8744218  31.0194002  28.7850963
## [37]  29.6277567  26.3487159  -1.2933127 -19.4359816
```

- **Logit Model with Forward Elimination**

```
##  [1] -21.1324127   1.6471631   1.9687105   2.8502070  -3.1267587  -3.3028850
##  [7]  -3.6732098  -6.6556256  -7.2930650 -23.8504338 -17.0961819 -17.1653572
## [13]   3.6153389   0.9210658   0.4594639  -1.1097315   4.0131436   6.5538691
## [19]  -1.5678079 -43.1228507 -41.8272191  -1.0616325   1.0108975  -3.3616476
## [25]  -3.3259980  -0.7694337 -18.6851352  15.7643086  11.6898199   5.7354155
## [31]  18.1336636  17.3945880  17.0257378  17.4618888  17.2745735  15.2985175
## [37]  15.4125285  11.6671505  -1.3291711 -17.4399249
```

- **Logit Model with Backward Elimination**

```
##  [1] -18.5749014   2.2043350   1.8065924   0.8982560  -3.0322514  -1.0444374
##  [7]  -1.6745047  -7.1433557  -7.7359050 -21.1945465 -19.0579804 -19.1775642
## [13]   2.8342223   1.2115299   0.3085187  -0.9777271   5.1807736   7.9402710
## [19]  -2.0835523 -40.2270283 -39.3183668  -1.8615699   1.2132010  -3.4206755
## [25]  -3.5992473  -0.9833513 -16.1256587  21.2691328  14.8455059   4.2619536
## [31]  14.2147561  15.9542636  15.0602397  16.5703311  15.1893301  13.0524380
## [37]  14.3065313  11.1485972  -1.2497896 -17.6153601
```

Using the logit model with forward elimination and a 0.5 threshold with our `dfeval_clean` data to predict how many neighborhoods are above and below the median crime rate, we obtain the following results:

```
## [1] "21 are above median crime rate and 19 are below median crime rate."
```

Appendix:

Code used in this homework

```r
# libraries used
library(tidyverse)
library(caret)
library(pROC)
library(corrplot)
library(GGally)
library(psych)
library(car)
library(kableExtra)
library(gridExtra)
library(performance)
library(faraway)
library(jtools)

# loading data
dftrain <- read.csv("https://raw.githubusercontent.com/letisalba/Data_621/master/Homework_3/csv/crime-tr
glimpse(dftrain)

dfeval <- read.csv("https://raw.githubusercontent.com/letisalba/Data_621/master/Homework_3/csv/crime-eva

# correlation plot
corrplot(cor(dftrain, use = "complete.obs"), tl.cex = 0.5)

# summarizing data set
summary(dftrain)
describe(dftrain)

# factor categorical variables from the training data set;
# variable: target
dftrain$target <- factor(dftrain$target, levels = c(0, 1))
levels(dftrain$target) <- list(below_median = 0, above_median = 1)

# from the training data set; variable: chas
dftrain$chas <- factor(dftrain$chas)
levels(dftrain$chas) <- list(not_on_charles = 0, on_charles = 1)

# from the evaluation data set; variable: chas
dfeval$chas <- factor(dfeval$chas)
levels(dfeval$chas) <- list(not_on_charles = 0, on_charles = 1)

# matrix scatter plot
pairs(dftrain)
```

```r
# Boxplots
par(mfrow = c(4, 3))
boxplot(zn ~ target, data = dftrain)
boxplot(indus ~ target, data = dftrain)
# boxplot(chas ~ target, data=dftrain) # excluding chas
boxplot(nox ~ target, data = dftrain)
boxplot(rm ~ target, data = dftrain)
boxplot(age ~ target, data = dftrain)
boxplot(dis ~ target, data = dftrain)
boxplot(rad ~ target, data = dftrain)
boxplot(tax ~ target, data = dftrain)
boxplot(ptratio ~ target, data = dftrain)
boxplot(lstat ~ target, data = dftrain)
boxplot(medv ~ target, data = dftrain)

# Contingency table
dfconting <- data.frame(target = dftrain$target, chas = dftrain$chas)
table(dfconting)

# jittered plot for chas variable
plot(jitter(as.numeric(dftrain$chas), amount = 0.15), jitter(as.numeric(dftrain$target),
    amount = 0.03), xlab = "Borders Charles River (2=yes)", ylab = "Crime Rate > Median (2=yes)",
    col = dftrain$chas, pch = as.numeric(dftrain$chas))

# Correlation matrix
print("Correlation matrix (numerical variables):")
round(cor(dftrain[, c(1, 2, 4:12)]), 2)

# plots
par(mfrow = c(6, 3))

plot(zn ~ dis, data = dftrain)
abline(lm(zn ~ dis, data = dftrain), lt = 2, col = 2)

plot(indus ~ nox, data = dftrain)
abline(lm(indus ~ nox, dftrain), lt = 2, col = 2)

plot(indus ~ age, data = dftrain)
abline(lm(indus ~ age, dftrain), lt = 2, col = 2)

plot(indus ~ dis, data = dftrain)
abline(lm(indus ~ dis, dftrain), lt = 2, col = 2)

plot(indus ~ rad, data = dftrain)
abline(lm(indus ~ dis, dftrain), lt = 2, col = 2)

plot(indus ~ tax, data = dftrain)
abline(lm(indus ~ tax, dftrain), lt = 2, col = 2)

plot(indus ~ lstat, data = dftrain)
abline(lm(indus ~ lstat, dftrain), lt = 2, col = 2)

plot(nox ~ age, data = dftrain)
```

```r
abline(lm(nox ~ age, dftrain), lt = 2, col = 2)

plot(nox ~ dis, data = dftrain)
abline(lm(nox ~ dis, dftrain), lt = 2, col = 2)

plot(nox ~ rad, data = dftrain)
abline(lm(nox ~ rad, dftrain), lt = 2, col = 2)

plot(nox ~ tax, data = dftrain)
abline(lm(nox ~ tax, dftrain), lt = 2, col = 2)

plot(nox ~ lstat, data = dftrain)
abline(lm(nox ~ lstat, dftrain), lt = 2, col = 2)

plot(rm ~ lstat, data = dftrain)
abline(lm(rm ~ lstat, dftrain), lt = 2, col = 2)

plot(rm ~ medv, data = dftrain)
abline(lm(rm ~ medv, dftrain), lt = 2, col = 2)

plot(age ~ dis, data = dftrain)
abline(lm(age ~ dis, dftrain), lt = 2, col = 2)

plot(age ~ lstat, data = dftrain)
abline(lm(age ~ lstat, dftrain), lt = 2, col = 2)

plot(rad ~ tax, data = dftrain)
abline(lm(rad ~ tax, dftrain), lt = 2, col = 2)

plot(lstat ~ medv, data = dftrain)
abline(lm(lstat ~ medv, dftrain), lt = 2, col = 2)

# checking for missing values
round(100 * colSums(is.na(dftrain))/nrow(dftrain), 2)
round(100 * colSums(is.na(dfeval))/nrow(dfeval), 2)

# cleaning train data
clean_df <- function(df) {
    df$rad_1 <- ifelse(df$rad == 1, 1, 0)
    df$rad_2 <- ifelse(df$rad == 2, 1, 0)
    df$rad_3 <- ifelse(df$rad == 3, 1, 0)
    df$rad_4 <- ifelse(df$rad == 4, 1, 0)
    df$rad_5 <- ifelse(df$rad == 5, 1, 0)
    df$rad_6 <- ifelse(df$rad == 6, 1, 0)
    df$rad_7 <- ifelse(df$rad == 7, 1, 0)
    df$rad_8 <- ifelse(df$rad == 8, 1, 0)
    df$rad <- NULL
    return(df)
}

dftrain_clean <- clean_df(dftrain)
dftrain_clean <- dftrain_clean %>%
    select(target, everything())
```

```r
dfeval_clean <- clean_df(dfeval)
head(dftrain_clean)
head(dfeval_clean)

# Model building

# Start with dummy target variable
dftrain_clean_dummy <- dftrain_clean %>%
    mutate(target = as.numeric(target == "above_median"))
olsreg <- lm(data = dftrain_clean_dummy, formula = target ~ .)
summ(olsreg)

# logit
logit <- glm(data = dftrain_clean, formula = target ~ ., family = binomial(link = "logit"))
summ(logit)

# logit backwards elimination
lmod.back <- step(logit, data = dftrain_clean, direction = "backward")
summ(lmod.back)

# logit minimal forward elimination
lmod.min <- glm(target ~ 1, family = binomial(), data = dftrain_clean)
summ(lmod.min)

# forward elimination
lmod.fwd <- step(lmod.min, data = dftrain_clean, direction = "forward",
    scope = formula(logit))
summ(lmod.fwd)

# Model Selection Linear Model Accuracy
table(true = dftrain_clean_dummy$target, pred = round(fitted(olsreg)))
# Logit Model Prediction Accuracy
table(true = dftrain_clean$target, pred = round(fitted(logit)))
# Logit Model with Forward Elimination Prediction Accuracy
table(true = dftrain_clean$target, pred = round(fitted(lmod.back)))
# Logit Model with Backward Elimination Prediction Accuracy
table(true = dftrain_clean$target, pred = round(fitted(lmod.back)))

# Model AUCs Linear Model
pred = round(fitted(olsreg))
pROC::auc(dftrain_clean_dummy$target, pred)

# Logit Model
pred = round(fitted(logit))
pROC::auc(dftrain_clean$target, pred)

# Logit Model with Forward Elimination
pred = round(fitted(lmod.fwd))
pROC::auc(dftrain_clean$target, pred)

# Logit Model with Backward Elimination
pred = round(fitted(lmod.back))
pROC::auc(dftrain_clean$target, pred)
```

```r
# Findings Linear Model
prediction <- broom::augment(olsreg, newdata = dfeval_clean)
prediction$.fitted

# Logit Model
prediction <- broom::augment(logit, newdata = dfeval_clean)
prediction$.fitted

# Logit Model with Forward Elimination
prediction <- broom::augment(lmod.fwd, newdata = dfeval_clean)
prediction$.fitted

# Logit Model with Backward Elimination
prediction <- broom::augment(lmod.back, newdata = dfeval_clean)
prediction$.fitted

# final prediction using eval data
predict <- predict(lmod.fwd, dfeval_clean, interval = "prediction")
eval <- table(as.integer(predict > 0.5))
print(paste(eval[1], "are above median crime rate", "and", eval[2],
    "are below median crime rate."))
```