Critical Thinking Group 4 - HW3

Sreejaya, Suman, Vuthy

October 10, 2016

Overview

The purpose of this project is to predict if a neighborhood will be at risk for high crime levels using binary logistic regression models. Below is a short description of the variables in the dataset.

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

Dataset
Crime - Training data
Crime - Evaluation Data

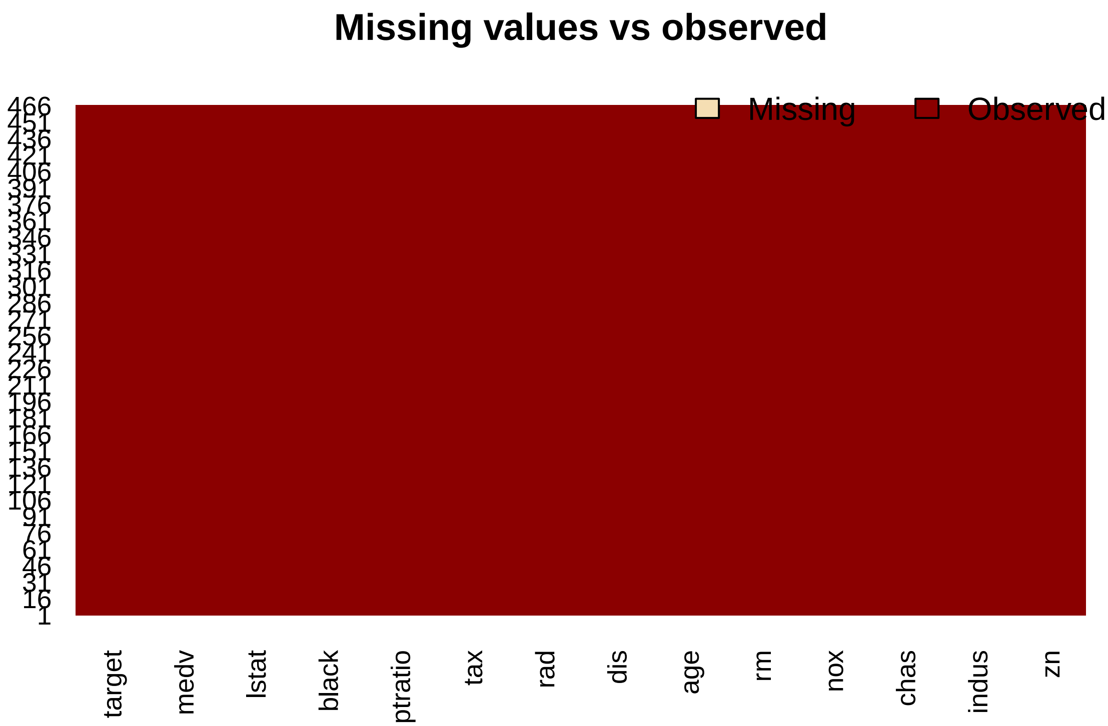
Data Exploration

The dataset contains 466 observations and 14 variables. The response variable is **target** variable. Below is a glimpse of the data.

```
## Observations: 466
## Variables: 14
## $ zn      <dbl> 0, 0, 0, 30, 0, 0, 0, 0, 0, 80, 22, 0, 0, 22, 0, 0, 10...
## $ indus   <dbl> 19.58, 19.58, 18.10, 4.93, 2.46, 8.56, 18.10, 18.10, 5...
## $ chas    <int> 0, 1, 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...
## $ rm      <dbl> 7.929, 5.403, 6.485, 6.393, 7.155, 6.781, 5.453, 4.519...
## $ age     <dbl> 96.2, 100.0, 100.0, 7.8, 92.2, 71.3, 100.0, 100.0, 38....
## $ dis     <dbl> 2.0459, 1.3216, 1.9784, 7.0355, 2.7006, 2.8561, 1.4896...
## $ rad     <int> 5, 5, 24, 6, 3, 5, 24, 24, 5, 1, 7, 5, 24, 7, 3, 3, 5,...
## $ tax     <int> 403, 403, 666, 300, 193, 384, 666, 666, 224, 315, 330,...
## $ ptratio <dbl> 14.7, 14.7, 20.2, 16.6, 17.8, 20.9, 20.2, 20.2, 20.2, ...
## $ black   <dbl> 369.30, 396.90, 386.73, 374.71, 394.12, 395.58, 396.90...
## $ lstat   <dbl> 3.70, 26.82, 18.85, 5.19, 4.82, 7.67, 30.59, 36.98, 5....
## $ medv    <dbl> 50.0, 13.4, 15.4, 23.7, 37.9, 26.5, 5.0, 7.0, 22.2, 20...
## $ target  <int> 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, ...
```

A visual take on the missing values might be helpful:

the Amelia package has a special plotting function missmap() that will plot your dataset and highlight missing



values:
There are no missing values in the dataset.

```
## [1] 466
```

```
##
##     0 12.5 17.5    18    20    21    22    25    28    30    33    34    35    40    45
##   339    10     1     1    21     4     9     8     3     6     3     3     3     7     6
## 52.5    55    60    70    75    80 82.5    85    90    95   100
##     3     3     4     3     3    13     2     2     4     4     1
```

Out of 466 values 339 are zeros. So we would like to treat zn as binary, land size over 25,000 sq.ft as 1 and below 25,000 sq.ft as 0

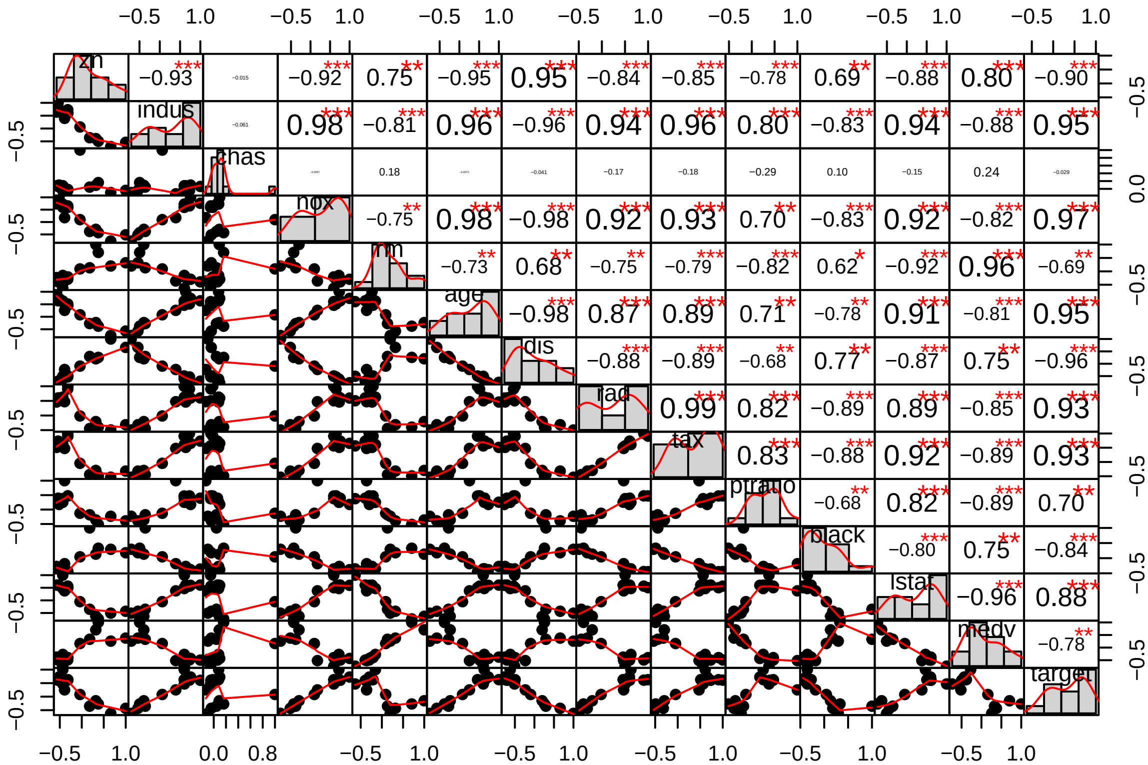
```
##
##     0 12.5 17.5    18    20    21    22    25    28    30    33    34    35    40    45
##   339    10     1     1    21     4     9     8     3     6     3     3     3     7     6
## 52.5    55    60    70    75    80 82.5    85    90    95   100
##     3     3     4     3     3    13     2     2     4     4     1
```

Lets check the summary of the given dataset, as well as check for any NA values in the data set.

```
##        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         black            lstat
##  Min.   :187.0   Min.   :12.6   Min.   :  0.32   Min.   : 1.730
##  1st Qu.:281.0   1st Qu.:16.9   1st Qu.:375.61   1st Qu.: 7.043
##  Median :334.5   Median :18.9   Median :391.34   Median :11.350
##  Mean   :409.5   Mean   :18.4   Mean   :357.12   Mean   :12.631
##  3rd Qu.:666.0   3rd Qu.:20.2   3rd Qu.:396.24   3rd Qu.:16.930
##  Max.   :711.0   Max.   :22.0   Max.   :396.90   Max.   :37.970
##       medv           target
##  Min.   : 5.00   Min.   :0.0000
##  1st Qu.:17.02   1st Qu.:0.0000
##  Median :21.20   Median :0.0000
##  Mean   :22.59   Mean   :0.4914
##  3rd Qu.:25.00   3rd Qu.:1.0000
##  Max.   :50.00   Max.   :1.0000
```

There appears to be no missing values. Lets plot the correlation between the variables.



From the above correlation matrix , the target varaible seems to have correlation with

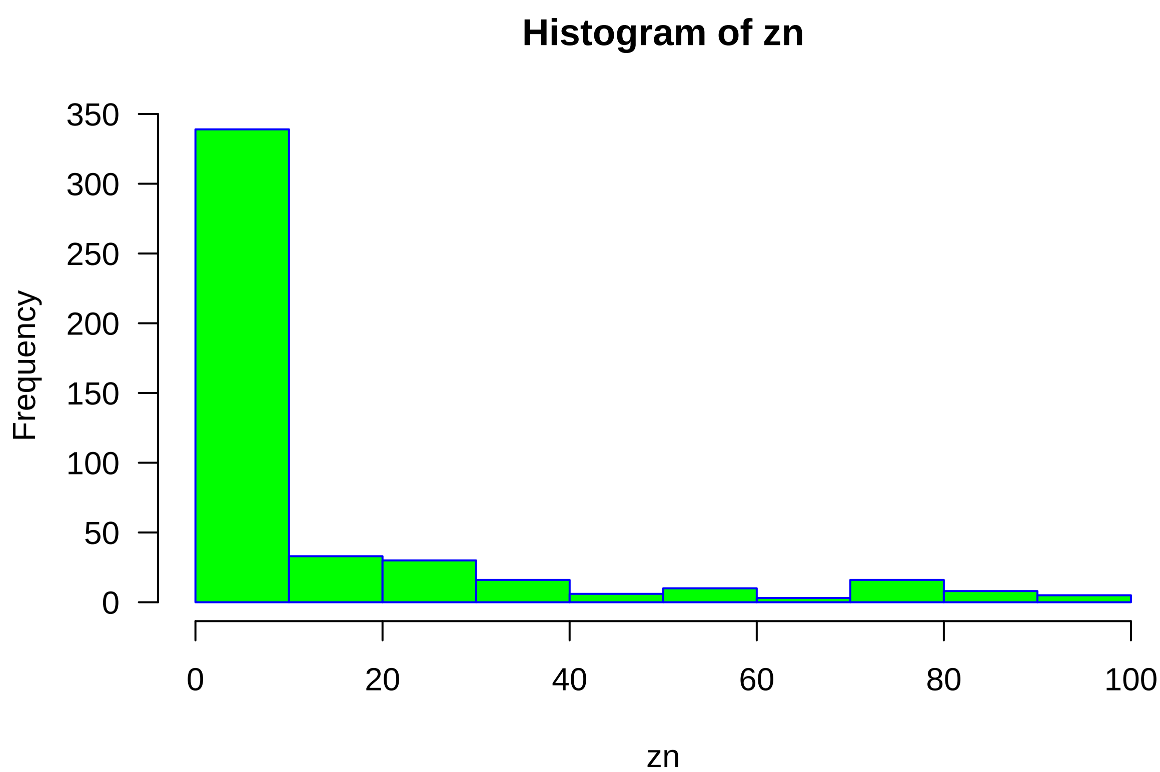
- zn - proportion of residential land zoned for large lots
- indus - proportion of non-retail business acres per suburb
- nox - nitrogen oxides concentration
- age - proportion of owner-occupied units built prior to 1940
- dis - weighted mean of distances to five Boston employment centers

- rad - index of accessibility to radial highways
- tax - full-value property-tax rate per $10,000
- lstat - lower status of the population

Data Preparation

Lets look at each of the predictor variable's data:

***zn* - proportion of residential land zoned for large lots**




```
##
##     0 12.5 17.5    18    20    21    22    25    28    30    33    34    35    40    45
##   339    10     1     1    21     4     9     8     3     6     3     3     3     7     6
## 52.5    55    60    70    75    80 82.5    85    90    95   100
##     3     3     4     3     3    13     2     2     4     4     1
```

	0	1
0	0.37	0.63
12.5	1.00	0.00
17.5	1.00	0.00
18	1.00	0.00
20	0.38	0.62
21	1.00	0.00
22	0.78	0.22
25	1.00	0.00
28	1.00	0.00
30	1.00	0.00
33	1.00	0.00
34	1.00	0.00
35	1.00	0.00
40	1.00	0.00
45	1.00	0.00

	0	1
52.5	1.00	0.00
55	1.00	0.00
60	1.00	0.00
70	1.00	0.00
75	1.00	0.00
80	1.00	0.00
82.5	1.00	0.00
85	1.00	0.00
90	1.00	0.00
95	1.00	0.00
100	1.00	0.00

From the above, it appears like majority of the neighborhoods have no residential land zoned for large lots. When we looked at the average response rates for the zn data, we have identified following categories:

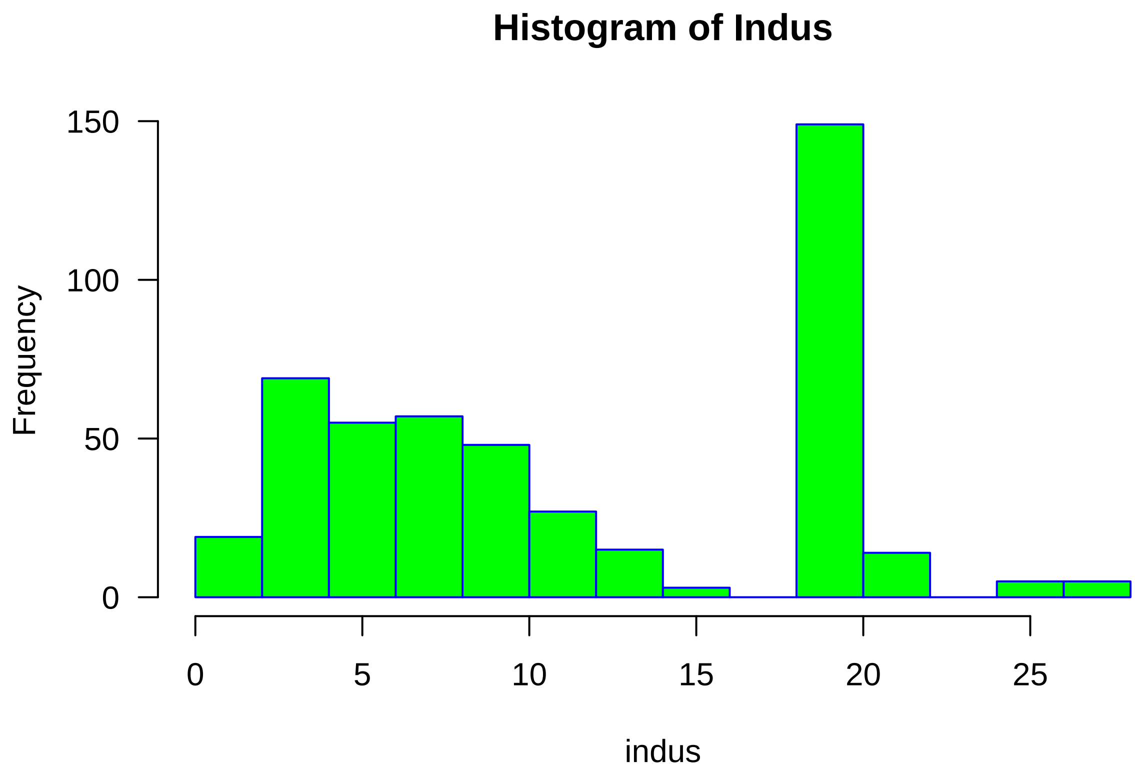
Target	
0, 1	zn
0.37,0.63	0
0.38,0.62	20
0.78,0.22	22
1.0,0.00	others.

So, we left with these 4 categories. So, by definition, we need to make 3 dummy variables.

indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv	target	zn1	zn2	zn3
19.58	0	0.605	7.929	96.2	2.0459	5	403	14.7	369.30	3.70	50.0	1	1	0	0
19.58	1	0.871	5.403	100.0	1.3216	5	403	14.7	396.90	26.82	13.4	1	1	0	0
18.10	0	0.740	6.485	100.0	1.9784	24	666	20.2	386.73	18.85	15.4	1	1	0	0
4.93	0	0.428	6.393	7.8	7.0355	6	300	16.6	374.71	5.19	23.7	0	0	0	0
2.46	0	0.488	7.155	92.2	2.7006	3	193	17.8	394.12	4.82	37.9	0	1	0	0
8.56	0	0.520	6.781	71.3	2.8561	5	384	20.9	395.58	7.67	26.5	0	1	0	0
18.10	0	0.693	5.453	100.0	1.4896	24	666	20.2	396.90	30.59	5.0	1	1	0	0
18.10	0	0.693	4.519	100.0	1.6582	24	666	20.2	88.27	36.98	7.0	1	1	0	0
5.19	0	0.515	6.316	38.1	6.4584	5	224	20.2	389.71	5.68	22.2	0	1	0	0
3.64	0	0.392	5.876	19.1	9.2203	1	315	16.4	395.18	9.25	20.9	0	0	0	0
5.86	0	0.431	6.438	8.9	7.3967	7	330	19.1	377.07	3.59	24.8	0	0	0	1
12.83	0	0.437	6.286	45.0	4.5026	5	398	18.7	383.23	8.94	21.4	0	1	0	0
18.10	0	0.532	7.061	77.0	3.4106	24	666	20.2	395.28	7.01	25.0	1	1	0	0
5.86	0	0.431	8.259	8.4	8.9067	7	330	19.1	396.90	3.54	42.8	1	0	0	1
2.46	0	0.488	6.153	68.8	3.2797	3	193	17.8	387.11	13.15	29.6	0	1	0	0

Similarly, let's proceed with others

***indus* - proportion of non-retail business acres per suburb**

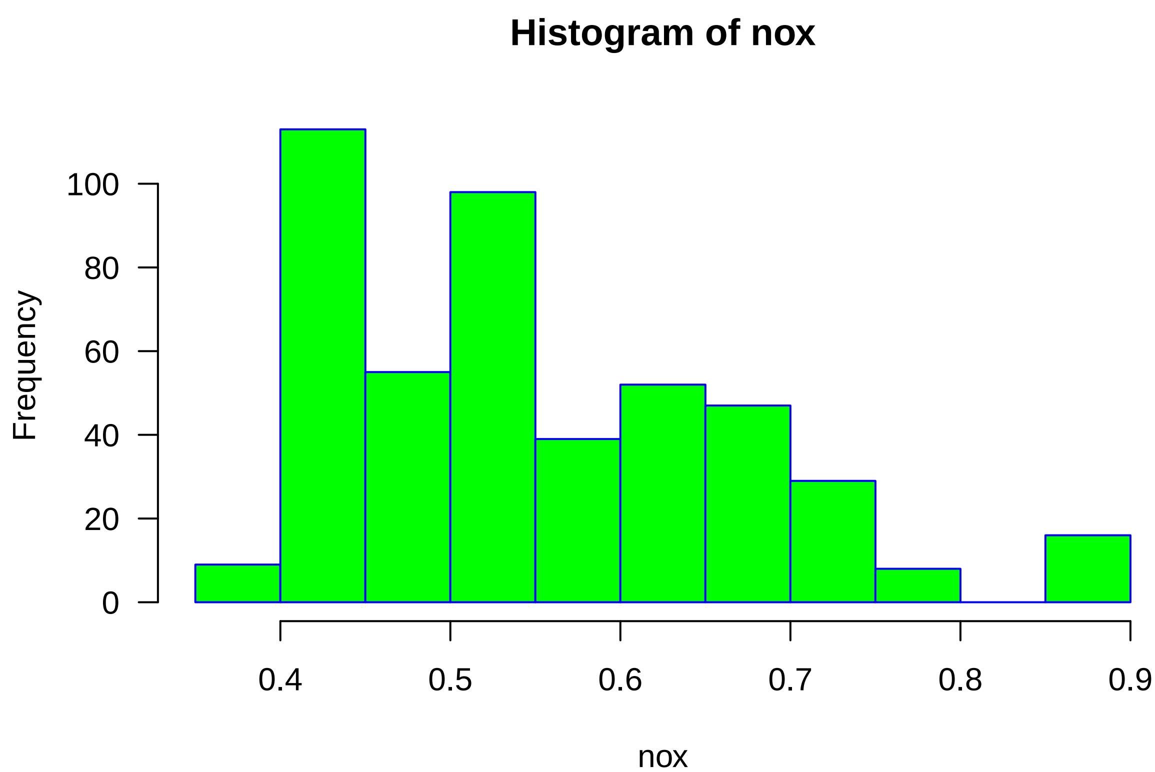


Indus	Target
2.95	1.00
3.24	1.00
3.33	1.00
3.37	1.00
3.41	1.00
3.44	1.00
3.64	1.00
3.75	1.00
3.78	1.00
4	1.00
4.05	1.00
4.15	1.00
4.39	1.00
4.49	1.00
4.86	1.00
4.93	1.00
4.95	1.00
5.13	1.00
5.19	1.00
5.32	1.00
5.64	1.00
5.96	1.00

Indus	Target
6.06	1.00
6.07	1.00
6.09	1.00
10.01	1.00
10.81	1.00
11.93	1.00
12.83	1.00
13.89	1.00
13.92	1.00
15.04	1.00
6.41	1.00
6.91	1.00
7.07	1.00
7.87	1.00
25.65	1.00
27.74	1.00
——	——
7.38	0.67
——	——
9.69	0.71
10.59	0.70
——	——
5.86	0.78
6.96	0.80
——	——
8.56	0.91
——	——
9.9	0.18
——	——
21.89	0.07
——	——
18.1	0.00
19.58	0.00
8.14	0.00
3.97	0.00
6.2	0.00
——	——

The distribution above appears some what weired, and we could not find a meaningful categorization here.

***nox* - nitrogen oxides concentration**



nox	Target
0.389	1.00
0.392	1.00
0.394	1.00
0.398	1.00
0.4	1.00
0.401	1.00
0.403	1.00
0.404	1.00
0.405	1.00
0.409	1.00
0.41	1.00
0.411	1.00
0.413	1.00
0.415	1.00
0.4161	1.00
0.422	1.00
0.426	1.00
0.428	1.00
0.429	1.00
0.433	1.00
0.437	1.00
0.4379	1.00

nox	Target
0.439	1.00
0.442	1.00
0.4429	1.00
0.445	1.00
0.447	1.00
0.448	1.00
0.449	1.00
0.453	1.00
0.458	1.00
0.46	1.00
0.469	1.00
0.472	1.00
0.484	1.00
0.488	1.00
0.499	1.00
0.51	1.00
0.515	1.00
0.518	1.00
0.524	1.00
0.547	1.00
0.55	1.00
0.573	1.00
0.581	1.00
0.609	1.00
0.52	0.91
0.493	0.67
0.585	0.71
0.431	0.78
0.489	0.79
0.464	0.88
0.544	0.18
0.624	0.07
0.538	0.05
0.504	0.00
0.507	0.00
0.532	0.00
0.575	0.00
0.58	0.00
0.583	0.00
0.584	0.00
0.597	0.00
0.605	0.00
0.614	0.00
0.631	0.00
0.647	0.00

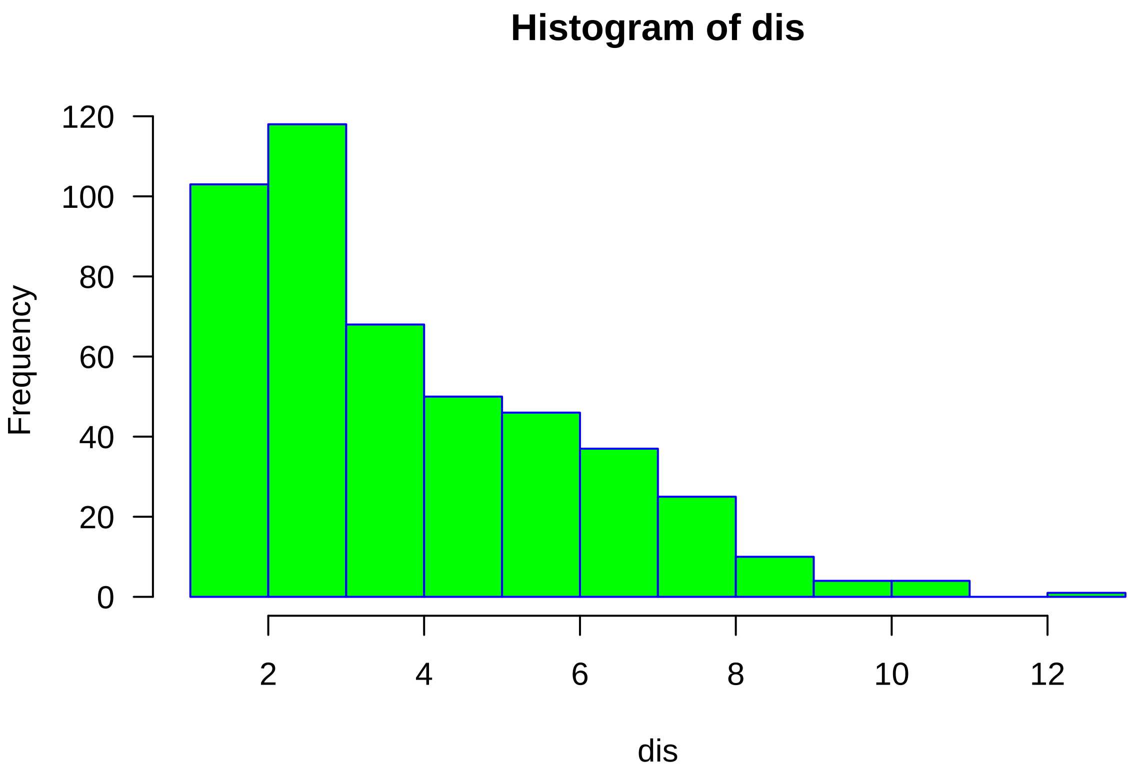
nox	Target
0.655	0.00
0.659	0.00
0.668	0.00
0.671	0.00
0.679	0.00
0.693	0.00
0.7	0.00
0.713	0.00
0.718	0.00
0.74	0.00
0.77	0.00
0.871	0.00
——-	——

There is no meaningful categorization can be concluded from the above. Let's proceed with other variables.

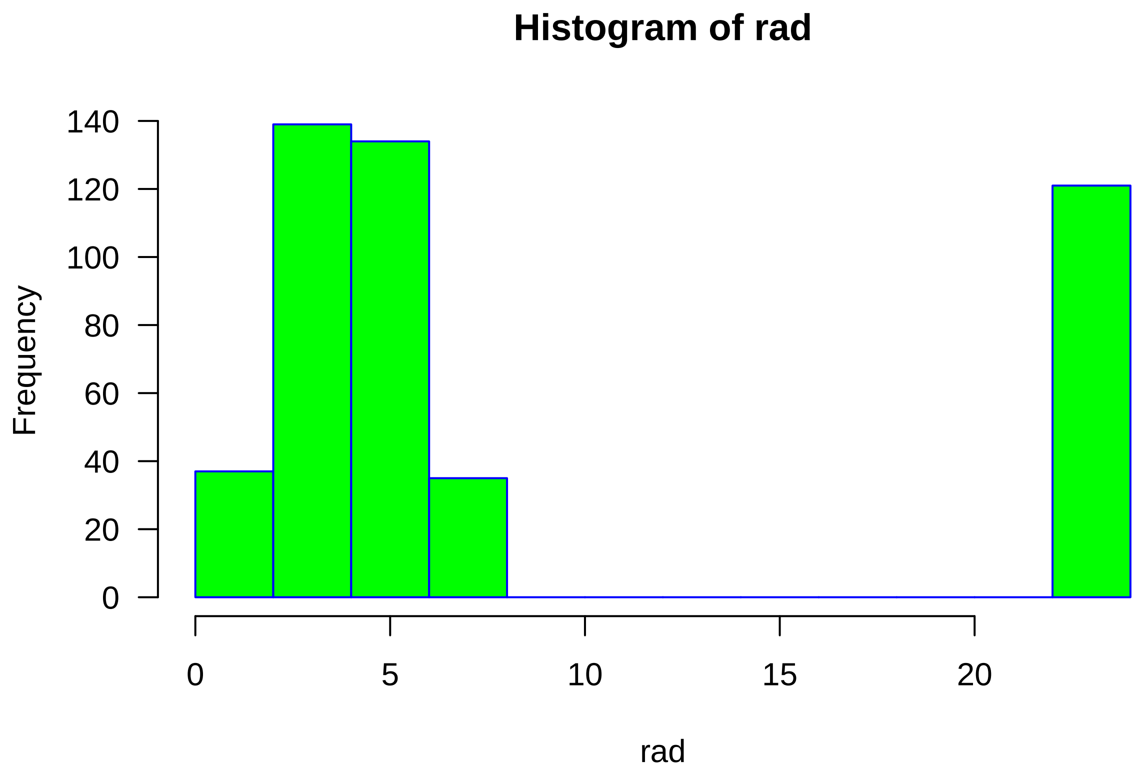
***age* - proportion of owner-occupied units built prior to 1940**

Looks like the buildings with age > 100 are mentioned as 100 in the above. We could not derive a specific categorization here, so, we leave the variable as is.

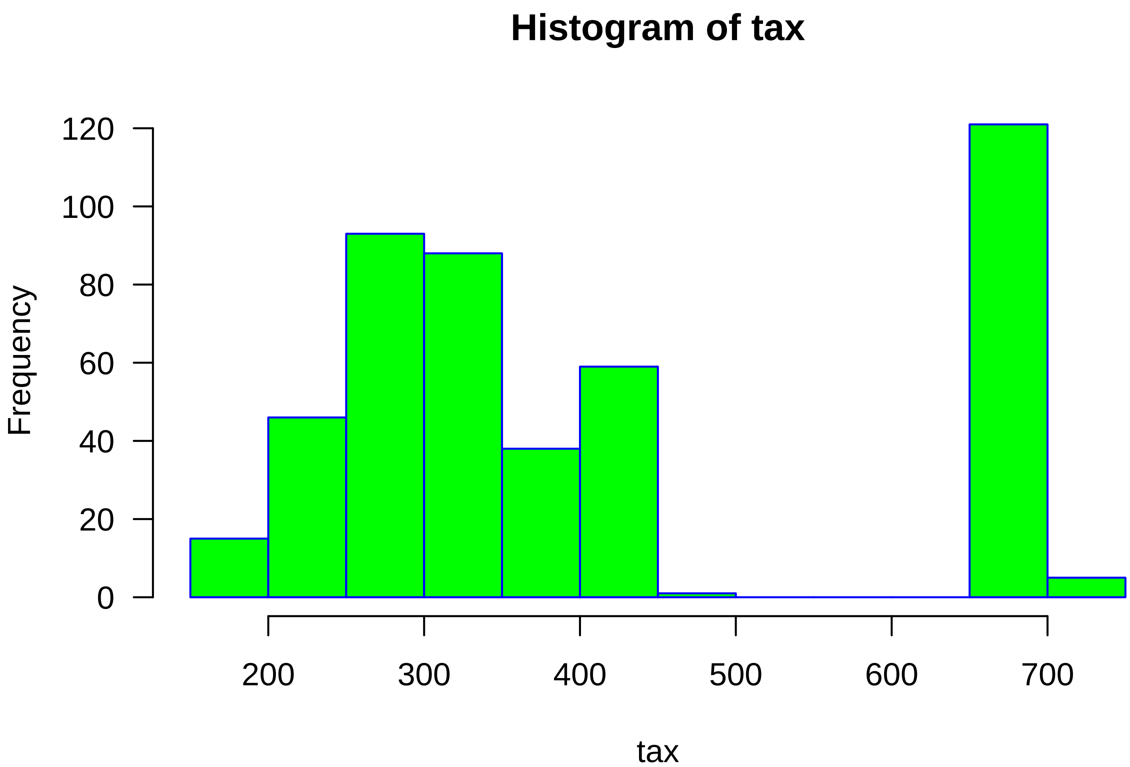
***dis* - weighted mean of distances to five Boston employment centers**



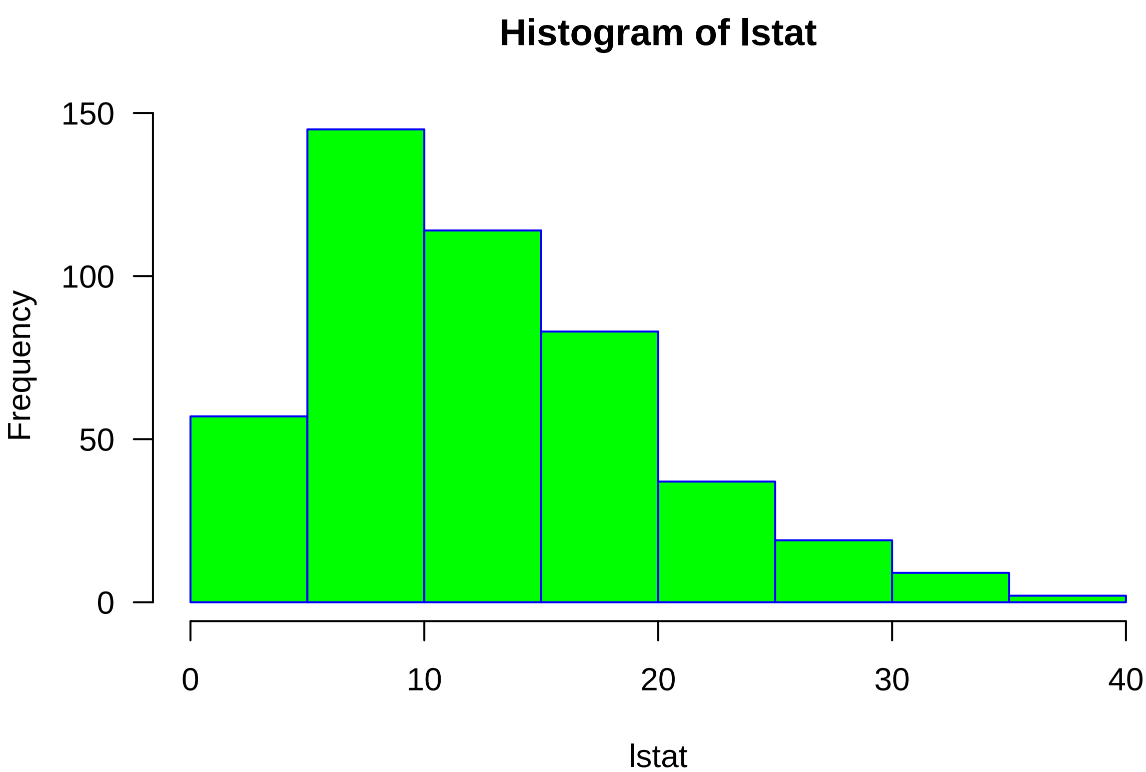
rad **- index of accessibility to radial highways**



tax **- full-value property-tax rate per $10,000**



***lstat* - lower status of the population**



Let's remove the predictors with low correlation with the target:

Also, let's split our dataset into training (80%) and test (20%).

Here's the glimpse of our training and test datasets for model building & validation:

Training dataset

```
## Observations: 372
## Variables: 9
## $ zn     <dbl> 30, 0, 80, 0, 0, 0, 0, 0, 0, 0, 0, 20, 22, 60, 0, 33, 0...
## $ indus  <dbl> 4.93, 18.10, 1.91, 18.10, 3.41, 10.59, 18.10, 18.10, 18...
## $ nox    <dbl> 0.4280, 0.6590, 0.4130, 0.6310, 0.4890, 0.4890, 0.6790,...
## $ age    <dbl> 52.9, 100.0, 21.9, 96.8, 73.9, 100.0, 78.7, 96.7, 91.2,...
## $ dis    <dbl> 7.0355, 1.1781, 10.5857, 1.3567, 3.0921, 3.8750, 1.8629...
## $ rad    <int> 6, 24, 4, 24, 2, 4, 24, 24, 24, 24, 24, 5, 7, 1, 24, 7,...
## $ tax    <int> 300, 666, 334, 666, 270, 277, 666, 666, 666, 666, 666, ...
## $ lstat  <dbl> 11.22, 23.34, 8.05, 3.73, 8.20, 23.09, 14.52, 18.03, 30...
## $ target <int> 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0...
```

Test dataset

```
## Observations: 94
## Variables: 9
```

```
## $ zn     <dbl> 0.0, 0.0, 0.0, 80.0, 20.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0...
## $ indus  <dbl> 19.58, 18.10, 8.56, 3.64, 3.97, 3.24, 6.20, 2.89, 18.10...
## $ nox    <dbl> 0.6050, 0.7400, 0.5200, 0.3920, 0.6470, 0.4600, 0.5070,...
## $ age    <dbl> 96.2, 100.0, 71.3, 19.1, 62.8, 32.2, 66.5, 62.5, 98.9, ...
## $ dis    <dbl> 2.0459, 1.9784, 2.8561, 9.2203, 1.9865, 5.8736, 3.6519,...
## $ rad    <int> 5, 24, 5, 1, 5, 4, 8, 2, 24, 5, 3, 5, 1, 5, 6, 24, 5, 4...
## $ tax    <int> 403, 666, 384, 315, 264, 430, 307, 276, 666, 403, 233, ...
## $ lstat  <dbl> 3.70, 18.85, 7.67, 9.25, 10.45, 9.09, 8.05, 6.19, 20.85...
## $ target <int> 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0...
```

Build Models

1. family=binomial in the glm() function.

1. Let us start with all the parameters

```
##
## Call:
## glm(formula = target ~ ., family = binomial, data = crime.train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.84235  -0.23828  -0.00695   0.00672   2.99187
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -23.265475   4.174649  -5.573 2.50e-08 ***
## zn           -0.060072   0.032177  -1.867   0.0619 .
## indus        -0.035322   0.049407  -0.715   0.4747
## nox          35.922207   7.139248   5.032 4.86e-07 ***
## age           0.022215   0.011691   1.900   0.0574 .
## dis           0.481844   0.210740   2.286   0.0222 *
## rad           0.604857   0.153366   3.944 8.02e-05 ***
## tax          -0.007431   0.002959  -2.512   0.0120 *
## lstat         0.027215   0.042154   0.646   0.5185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.31  on 371  degrees of freedom
## Residual deviance: 172.26  on 363  degrees of freedom
## AIC: 190.26
##
## Number of Fisher Scoring iterations: 9
```

2. without any parameter

```
##
## Call:
## glm(formula = target ~ 1, family = binomial, data = crime.train)
##
```

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.150  -1.150  -1.150   1.205   1.205
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.06454    0.10375  -0.622    0.534
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.31  on 371  degrees of freedom
## Residual deviance: 515.31  on 371  degrees of freedom
## AIC: 517.31
##
## Number of Fisher Scoring iterations: 3
```

2. Backward elimination method

```
## Start:  AIC=190.26
## target ~ zn + indus + nox + age + dis + rad + tax + lstat
##
##         Df Deviance    AIC
## - lstat  1   172.68 188.68
## - indus  1   172.78 188.78
## <none>       172.26 190.26
## - age    1   176.07 192.07
## - zn     1   176.87 192.87
## - dis    1   177.88 193.88
## - tax    1   179.21 195.21
## - rad    1   213.66 229.66
## - nox    1   216.15 232.15
##
## Step:  AIC=188.68
## target ~ zn + indus + nox + age + dis + rad + tax
##
##         Df Deviance    AIC
## - indus  1   173.05 187.05
## <none>       172.68 188.68
## - zn     1   177.93 191.93
## - age    1   178.27 192.27
## - tax    1   179.22 193.22
## - dis    1   179.24 193.24
## - rad    1   213.74 227.74
## - nox    1   216.39 230.39
##
## Step:  AIC=187.05
## target ~ zn + nox + age + dis + rad + tax
##
##        Df Deviance    AIC
## <none>      173.05 187.05
## - age   1   178.61 190.61
## - zn    1   178.65 190.65
## - dis   1   179.62 191.62
```

```
## - tax   1   183.86 195.86
## - rad   1   220.88 232.88
## - nox   1   221.68 233.68
```

```
## target ~ zn + nox + age + dis + rad + tax
```

```
##
## Call:
## glm(formula = target ~ zn + nox + age + dis + rad + tax, family = binomial,
##     data = crime.train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.75813  -0.25512  -0.00687   0.00797   3.01782
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -22.369572   3.919045  -5.708 1.14e-08 ***
## zn           -0.062721   0.031062  -2.019  0.04347 *
## nox          33.724272   6.258278   5.389 7.10e-08 ***
## age           0.024821   0.010814   2.295  0.02172 *
## dis           0.507966   0.206578   2.459  0.01393 *
## rad           0.617858   0.141604   4.363 1.28e-05 ***
## tax          -0.007693   0.002613  -2.944  0.00324 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.31  on 371  degrees of freedom
## Residual deviance: 173.05  on 365  degrees of freedom
## AIC: 187.05
##
## Number of Fisher Scoring iterations: 8
```

3. Forward elimination method

```
## Start:  AIC=517.31
## target ~ 1
##
##         Df Deviance    AIC
## + nox    1   236.88 240.88
## + rad    1   321.57 325.57
## + age    1   335.73 339.73
## + dis    1   338.08 342.08
## + tax    1   349.81 353.81
## + indus  1   358.85 362.85
## + zn     1   407.26 411.26
## + lstat  1   412.07 416.07
## <none>       515.31 517.31
##
## Step:  AIC=240.88
## target ~ nox
```

```
##
##         Df Deviance    AIC
## + rad    1   197.59 203.59
## + tax    1   233.34 239.34
## + zn     1   233.64 239.64
## + dis    1   234.37 240.37
## + indus  1   234.42 240.42
## <none>       236.88 240.88
## + age    1   235.53 241.53
## + lstat  1   236.49 242.49
##
## Step:  AIC=203.59
## target ~ nox + rad
##
##         Df Deviance    AIC
## + tax    1   186.30 194.30
## + indus  1   192.83 200.83
## + zn     1   193.76 201.76
## + age    1   194.55 202.55
## + dis    1   195.50 203.50
## <none>       197.59 203.59
## + lstat  1   197.00 205.00
##
## Step:  AIC=194.3
## target ~ nox + rad + tax
##
##         Df Deviance    AIC
## + age    1   182.21 192.21
## + zn     1   182.93 192.93
## + lstat  1   183.12 193.12
## <none>       186.30 194.30
## + dis    1   184.63 194.63
## + indus  1   185.61 195.61
##
## Step:  AIC=192.21
## target ~ nox + rad + tax + age
##
##         Df Deviance    AIC
## + dis    1   178.65 190.65
## + zn     1   179.62 191.62
## <none>       182.21 192.21
## + lstat  1   180.81 192.81
## + indus  1   181.58 193.58
##
## Step:  AIC=190.65
## target ~ nox + rad + tax + age + dis
##
##         Df Deviance    AIC
## + zn     1   173.05 187.05
## <none>       178.65 190.65
## + lstat  1   177.87 191.87
## + indus  1   177.93 191.93
##
## Step:  AIC=187.05
```

```
## target ~ nox + rad + tax + age + dis + zn
##
##          Df Deviance    AIC
## <none>        173.05 187.05
## + indus  1   172.68 188.68
## + lstat  1   172.78 188.78
```

```
## target ~ nox + rad + tax + age + dis + zn
```

```
##
## Call:
## glm(formula = target ~ nox + rad + tax + age + dis + zn, family = binomial,
##     data = crime.train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.75813  -0.25512  -0.00687   0.00797   3.01782
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -22.369572   3.919045  -5.708 1.14e-08 ***
## nox          33.724272   6.258278   5.389 7.10e-08 ***
## rad           0.617858   0.141604   4.363 1.28e-05 ***
## tax          -0.007693   0.002613  -2.944  0.00324 **
## age           0.024821   0.010814   2.295  0.02172 *
## dis           0.507966   0.206578   2.459  0.01393 *
## zn           -0.062721   0.031062  -2.019  0.04347 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.31  on 371  degrees of freedom
## Residual deviance: 173.05  on 365  degrees of freedom
## AIC: 187.05
##
## Number of Fisher Scoring iterations: 8
```

From the above two models we can see that zn,& age are not statistically significant. As for the statistically significant variables, rad & nox have a strong positive association of crime rate while tax has a negative coefficient, suggests as all other variables being equal as tax increases crime rate decreases.

4. Manual model1

We would drop out Zn and age from the above models.

```
##
## Call:
## glm(formula = target ~ nox + rad + tax + dis, family = binomial(link = "logit"),
##     data = crime.train)
##
## Deviance Residuals:
```

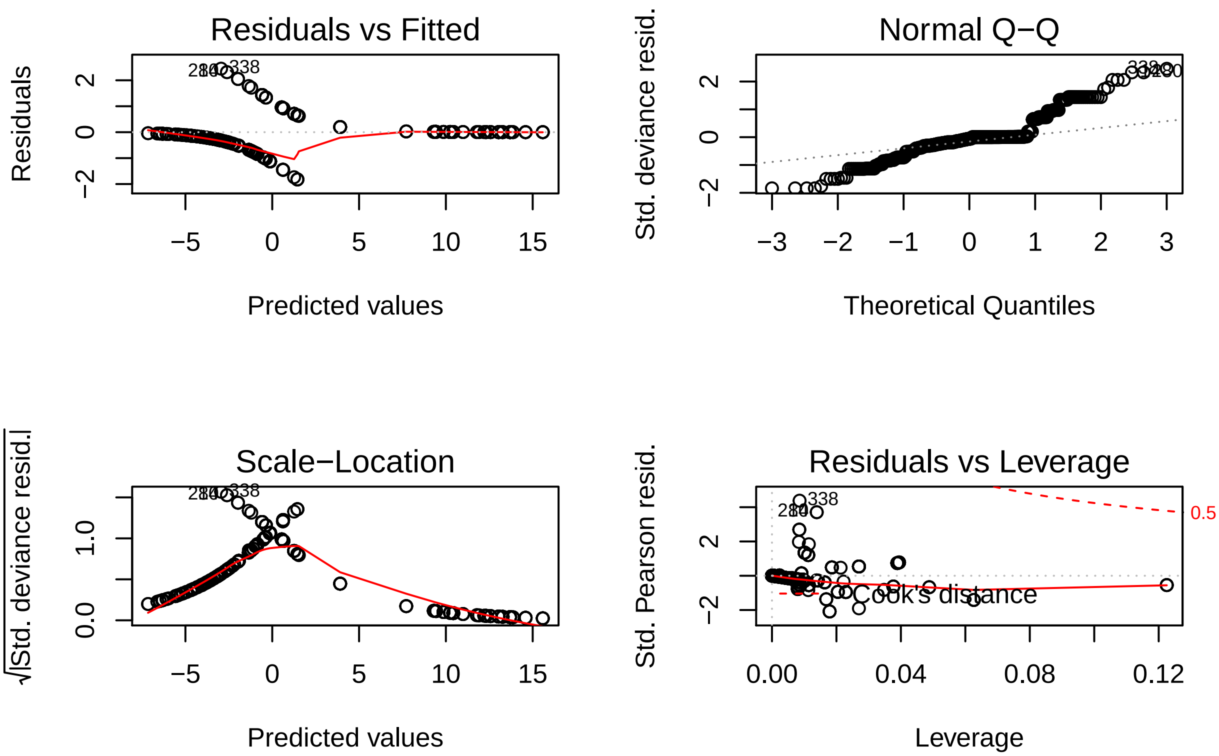
```
##      Min        1Q    Median        3Q       Max
## -1.85375  -0.31225  -0.06564   0.00749   2.51549
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -21.553260   3.652807  -5.900 3.62e-09 ***
## nox          37.708340   6.100531   6.181 6.36e-10 ***
## rad           0.562072   0.127164   4.420 9.87e-06 ***
## tax          -0.007533   0.002504  -3.009  0.00262 **
## dis           0.209893   0.161309   1.301  0.19319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.31  on 371  degrees of freedom
## Residual deviance: 184.63  on 367  degrees of freedom
## AIC: 194.63
##
## Number of Fisher Scoring iterations: 8
```

5. Manual model2

We would drop out distance from the above model since the p value is not significant. Now the new model:

```
##
## Call:
## glm(formula = target ~ nox + rad + tax, family = binomial(link = "logit"),
##     data = crime.train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.82233  -0.32010  -0.05947   0.00843   2.44822
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -18.250944   2.447132  -7.458 8.78e-14 ***
## nox          33.039818   4.740219   6.970 3.17e-12 ***
## rad           0.562869   0.127143   4.427 9.55e-06 ***
## tax          -0.007685   0.002517  -3.053  0.00227 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.31  on 371  degrees of freedom
## Residual deviance: 186.30  on 368  degrees of freedom
## AIC: 194.3
##
## Number of Fisher Scoring iterations: 8
```

A unit increase in index of accessibility to radial highways increses the log odds by 0.56. Also unit increase in nitrogen oxides concentration increases the logodds by 33.03, while increase in tax rate reduces the log odds by 0.008.



6.Bayesian Approach

```
##
## Call:
## bic.glm.formula(f = target ~ ., data = crime.train, glm.family = "binomial")
##
##
##   11  models were selected
##  Best  5  models (cumulative posterior probability =  0.7851 ):
##
##            p!=0    EV          SD         model 1     model 2     model 3
## Intercept  100    -18.468750  3.310003   -1.825e+01  -1.719e+01  -1.633e+01
## zn          22.3   -0.010981  0.024787        .           .      -4.111e-02
## indus        2.8   -0.001139  0.010645        .           .           .
## nox        100.0   32.014998  5.930401    3.304e+01   2.846e+01   2.940e+01
## age         26.8    0.005791  0.011041        .       2.014e-02       .
## dis         18.0    0.064152  0.164134        .           .           .
## rad        100.0    0.574138  0.134391    5.629e-01   5.779e-01   5.785e-01
## tax         97.4   -0.007641  0.002868   -7.685e-03  -8.246e-03  -7.441e-03
## lstat        9.7    0.006459  0.022926        .           .           .
##
## nVar                                        3           4           4
## BIC                                      -1.992e+03  -1.990e+03  -1.989e+03
## post prob                                  0.381       0.153       0.107
##            model 4     model 5
## Intercept  -1.852e+01  -2.170e+01
## zn              .      -6.279e-02
```

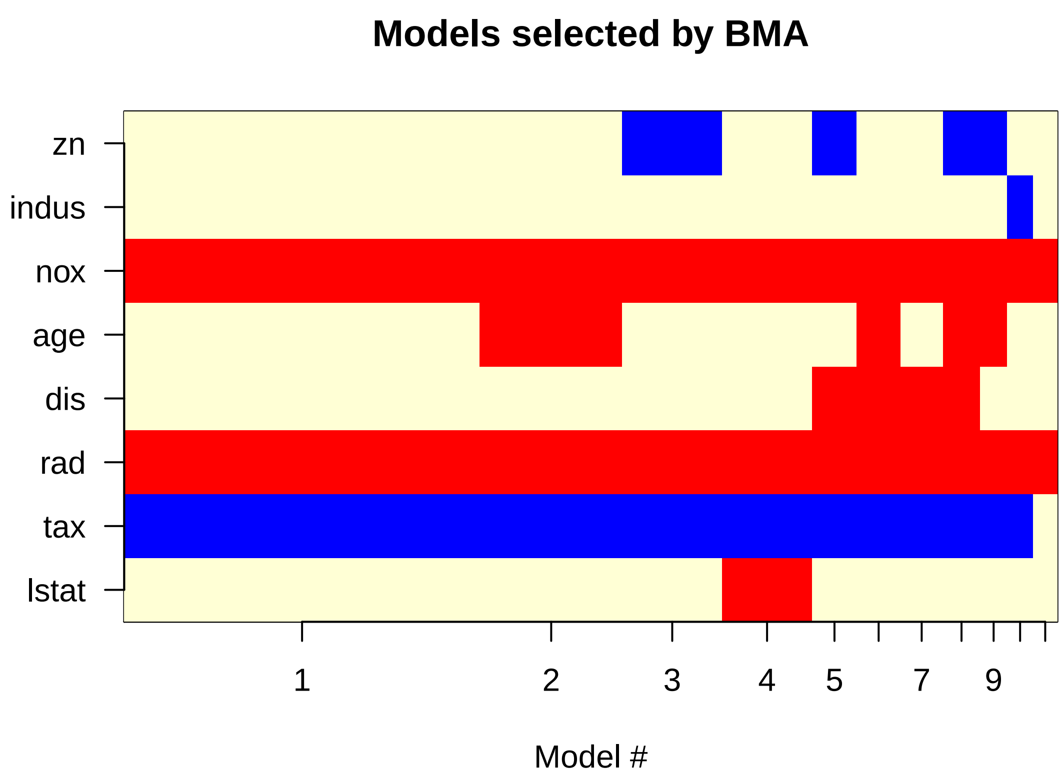
```
## indus              .            .
## nox          3.223e+01   3.646e+01
## age                .            .
## dis                .       4.025e-01
## rad          6.267e-01   5.860e-01
## tax         -9.002e-03  -6.924e-03
## lstat        6.687e-02         .
##
## nVar           4            5
## BIC         -1.989e+03  -1.988e+03
## post prob    0.097        0.048
```

```
##  [1] 0.38118926 0.15274163 0.10666531 0.09659408 0.04789106 0.04680092
##  [7] 0.04540118 0.04000225 0.02882907 0.02790397 0.02598127
```

```
##  [1] "nox,rad,tax"             "nox,age,rad,tax"
##  [3] "zn,nox,rad,tax"          "nox,rad,tax,lstat"
##  [5] "zn,nox,dis,rad,tax"      "nox,age,dis,rad,tax"
##  [7] "nox,dis,rad,tax"         "zn,nox,age,dis,rad,tax"
##  [9] "zn,nox,age,rad,tax"      "indus,nox,rad,tax"
## [11] "nox,rad"
```

```
## [1] "zn"    "indus" "nox"   "age"   "dis"   "rad"   "tax"   "lstat"
```

```
##    zn indus   nox   age   dis   rad   tax lstat
##  22.3   2.8 100.0  26.8  18.0 100.0  97.4   9.7
```



```
##   (Intercept)            zn         indus           nox           age
## -18.468749718  -0.010981198  -0.001139441  32.014997721   0.005790626
##           dis           rad           tax         lstat
##   0.064152469   0.574138329  -0.007640527   0.006459032
```

from the above resuls it is clear nitrogen oxides concentration(nox), accessibility to radial highways(rad) and property-tax rate(tax) are the 3 variables -probability they should be in the model The model is target ~ nox+rad+tax

Select Models

1. anova() function on the model to analyze the table of deviance

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: target
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                   371     515.31
## nox   1  278.438       370     236.88 < 2.2e-16 ***
## rad   1   39.290       369     197.59 3.654e-10 ***
## tax   1   11.291       368     186.30 0.0007789 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better. Nitrogen oxides concentration is the least deviation, so this variable can be dropped from the model. ### 2. Specificity and Sensitivity ### 3. AUC

```
## fitted.results
##  0  1
## 47 47
```

```
##
##  0  1
## 45 49
```

Predictions

Appendix