Data 621: Homework 3

Binary Logistic Regression

Aaron Grzasko

April 14, 2018

Introduction

The objective of this assignment is to predict if a given neighborhood in Boston is likely to experience a high crime rate. For categorical purposes, we define a neighborhood crime rate as "high" if it exceeds the citywide median crime rate.

Because we are interested in making simple yes/no determinations, we will build various binary logistic regression models to make our predictions.

We will train our models using a prescribed data set that captures a variety of metrics for each Boston neighborhood.

Data Exploration

High Level Overview

Our training data comprises 466 observations and 13 variables.

Below is a brief description of the variables in our data set:

Variable Name	Description	Variable Type
zn	proportion of residental land zoned for large lots (>25k sq ft)	predictor
indus	proportion non-retail business acres per suburb	predictor
chas	dummy variable indicating if suburb borders Charles River (1=y,0=n)	predictor
nox	nitrogen oxide concentration (ppm)	predictor
rm	avg rooms per dwelling	predictor
age	proportion of owner occupied units built prior to 1940	predictor
dis	weighted mean distance to five Boston employemtn centers	predictor
rad	index of accessibility to radial highways	predictor
tax	full-value property tax per $10k	predictor
ptratio	pupil-teacher ratio	predictor
lstat	lower status percentage of population	predictor
medv	median value of owner-occupied homes	predictor
target	whether crime rate is above median (1=y, 0=n)	response

Here are sample observations from the training data:

```
'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 ...
```

As expected, our response variable, `target`, is binary. There is one binary variable, `chas`, among our predictor variables. Four of the predictors are expressed as proportions, with values ranging from 0 to 100:

- `indus`

- `indust`

- `age`

- `lstat`

Two of the variables are dimensionless ratios:

- `nox`: dimensionless measure, with theoretical values ranging from 0 to 10 million parts per 10 million.

- `ptratios`: pupil to teacher ratio, with a theoretical minimum of 1 and unbounded above.

The variable `rad` is desribed as an index value of accessibility to radial highways. We assume this variable is an ordinal data type. 0 The remaining predictors provide a variety of count, distance, and monetary measures.

Now, let's examine a brief descriptive summary of all of our variables:

```
       zn           indus           chas            nox             rm           age     
 Min.   :  0   Min.   : 0.5   Min.   :0.00   Min.   :0.39   Min.   :3.9   Min.   :  3  
 1st Qu.:  0   1st Qu.: 5.1   1st Qu.:0.00   1st Qu.:0.45   1st Qu.:5.9   1st Qu.: 44  
 Median :  0   Median : 9.7   Median :0.00   Median :0.54   Median :6.2   Median : 77  
 Mean   : 12   Mean   :11.1   Mean   :0.07   Mean   :0.55   Mean   :6.3   Mean   : 68  
 3rd Qu.: 16   3rd Qu.:18.1   3rd Qu.:0.00   3rd Qu.:0.62   3rd Qu.:6.6   3rd Qu.: 94  
 Max.   :100   Max.   :27.7   Max.   :1.00   Max.   :0.87   Max.   :8.8   Max.   :100  
      dis            rad            tax          ptratio         lstat          medv   
 Min.   : 1.1   Min.   : 1.0   Min.   :187   Min.   :12.6   Min.   : 2   Min.   : 5  
 1st Qu.: 2.1   1st Qu.: 4.0   1st Qu.:281   1st Qu.:16.9   1st Qu.: 7   1st Qu.:17  
 Median : 3.2   Median : 5.0   Median :334   Median :18.9   Median :11   Median :21  
 Mean   : 3.8   Mean   : 9.5   Mean   :410   Mean   :18.4   Mean   :13   Mean   :23  
 3rd Qu.: 5.2   3rd Qu.:24.0   3rd Qu.:666   3rd Qu.:20.2   3rd Qu.:17   3rd Qu.:25  
 Max.   :12.1   Max.   :24.0   Max.   :711   Max.   :22.0   Max.   :38   Max.   :50  
     target    
 Min.   :0.00  
 1st Qu.:0.00  
 Median :0.00  
 Mean   :0.49  
 3rd Qu.:1.00  
 Max.   :1.00  
```

Surpisingly, we find no missing values in a data set. Based on this high-level summary and variable descriptions, we see no immediately obvious data entry errors or suspicious observation.

Let's take a closer look at each variable individually.

Response Variable: zn

The proportion of residental land zoned for large lots, `zn` for short, is a variable with values ranging from 0 to 1. The variable has a significant positive skew, as evident in the histogram, box plot, and qq plots below. We also note that 73% of observations (339 of 466 total) have a value of 0. Furthermore, there appears to be relationship between crime rates and `zn`: 93% of high crime areas have no land zoned for large lots. On the other hand, only about half of low crime areas exclude large-lot zoning. This relationship is consistent with our intuition: large lots tend be concentrated in suburban areas with low crime rates, while areas with mostly small lots are concentrated in urban areas with higher crime rates. Finally, the boxplots by crime type indicate a much higher variance of `zn` values for low crime neighborhoods compared to high crime areas.

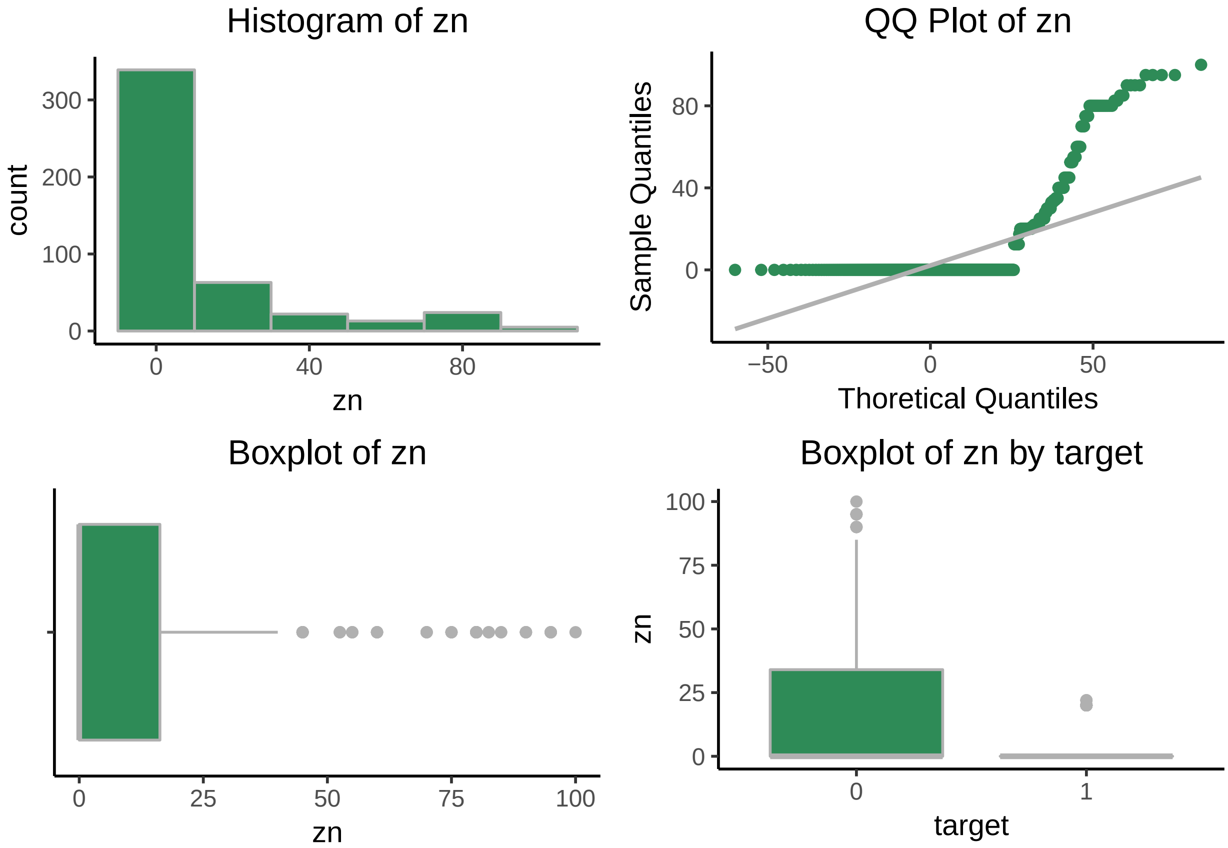
Here is a summarized table of observed `zn` observations, with values rounded to the nearest 5:

```
  0  10  20  25  30  35  40  45  50  55  60  70  75  80  85  90  95 100 Sum 
339  10  36   8   9   9   7   6   3   3   4   3   3  15   2   4   4   1 466 
```

Below are summary statistics:

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.         SD       Skew       Kurt 
  0.000000   0.000000   0.000000  11.577253  16.250000 100.000000  23.364651   2.183841   6.842914 
```

Finally, here are relevant plots:



Response Variable: indus

The variable `indus` represents the proportion of non-retail business acres per suburb. The histogram below indicates a bi-modal quality to the variable's distribution, with many values clustering in two ranges: the mid-to-upper single digits, and the upper teens through low 20s.
While the distribution has a very mild positive skew, the kurtosis is significantly below that of a normal

distribution. In the final boxplot below, we see that high crime areas tend to have a higher industry concentration compared to low crime areas. In contrast, the variances of industry concentration by crime-type are similar.

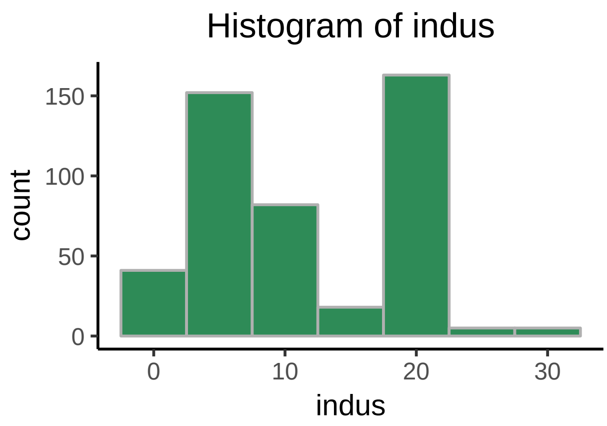
Let's look at summmary table, with values rounded to the nearest percentage:

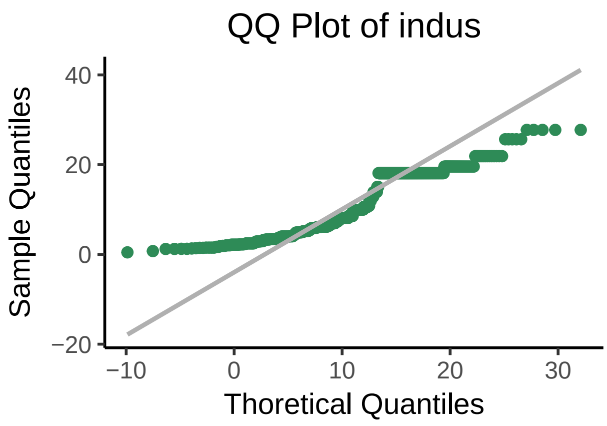
```
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  18  20  22  26  28 Sum 
  1   9  31  29  30  27  45  21  26  11  27  14   4   6   9   3 121  28  14   5   5 466 
```

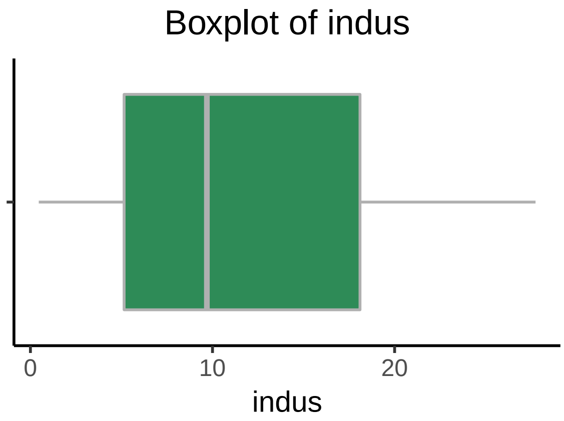
Here are the summary statistics:

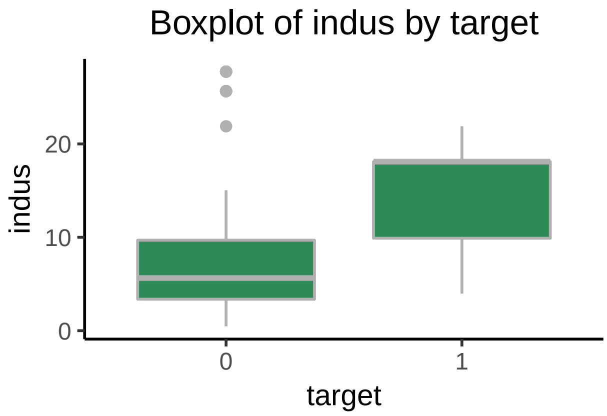
```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.         SD       Skew       Kurt 
 0.4600000  5.1450000  9.6900000 11.1050215 18.1000000 27.7400000  6.8458549  0.2894763  1.7643510 
```

Lastly, below are the plots:









Response Variable: chas

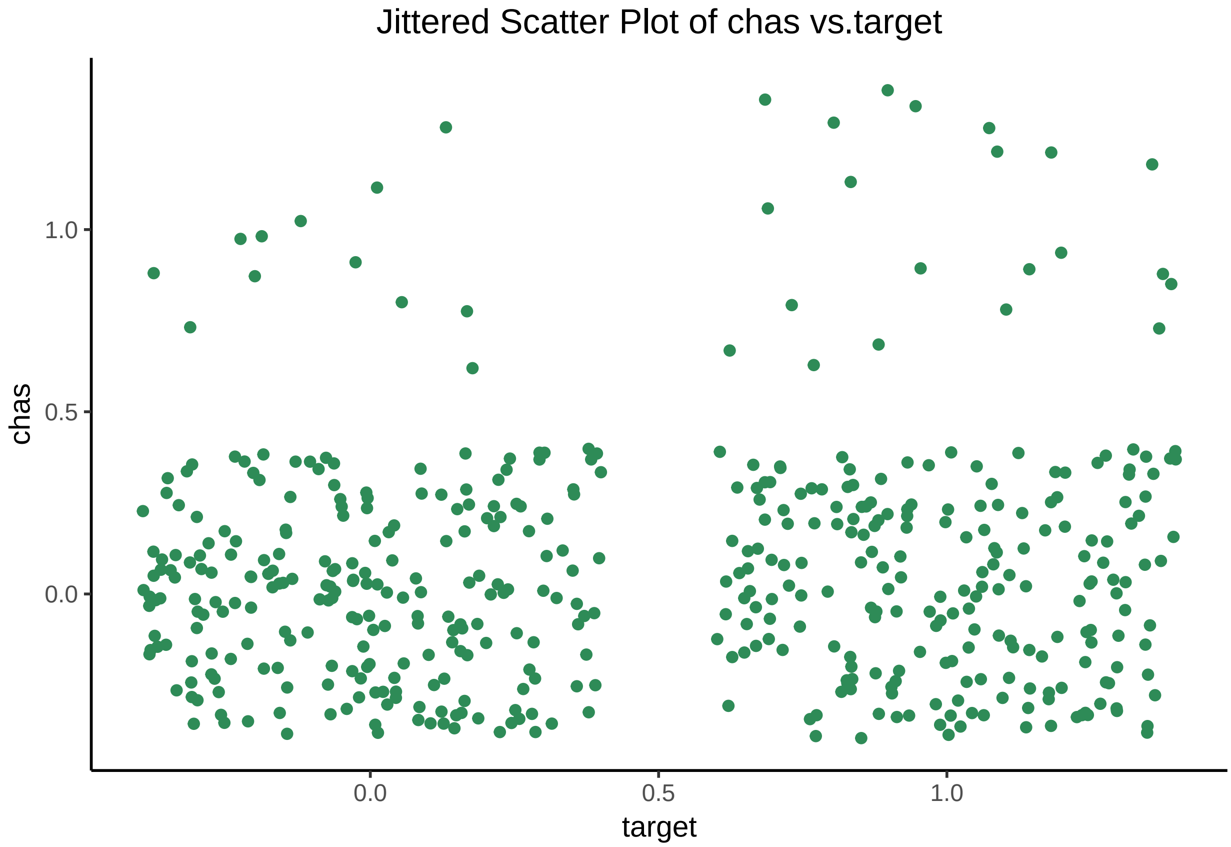
The variable `chas` is binary with the value 1 indicating that the neighborhood borders the Charles River. Only 7% of all neighborhoods (33 in total) border the river.

Let's look at the summary of observations:

```
  0   1 Sum 
433  33 466 
```

Because this variable is binary, we will not produce the standard plots and summary statistics shown for previous variables.

However, we will still visually examine the relationship between `chas` and `target`, the categorical crime level:



Here is a two-way table of the data depicted in the scatter plot:

```
        0   1 Sum
  0   225 208 433
  1    12  21  33
  Sum 237 229 466
```

Of the areas bordering the Charles river, roughly two-thirds (or 21 total) are in high crime areas. However, the differences in the proportions of high-crime areas by `chas` value do not appear appear to be statistically significant at the 95% level of significance. This non significant result is likely due to the small sample size of the Charles-River bordering observations. See below for the t-test details:

```
    2-sample test for equality of proportions with continuity correction

data:  table(crime$chas, crime$target)
X-squared = 2.394, df = 1, p-value = 0.1218
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0310513  0.3430395
sample estimates:
   prop 1    prop 2
```

```
0.5196305 0.3636364
```

Response Variable: nox

The variable `nox` represent the concentration of nitrogen oxide in each Boston area. This variable exhibits a seomewhat moderate, positive skew, with a kurtosis value similar to that of a normally distributed variable. The final boxplot below indicates higher median `nox` values in high crime areas vis-a-vis the low crime counterparts. We also see moderately higher `nox` variance in high crime areas.

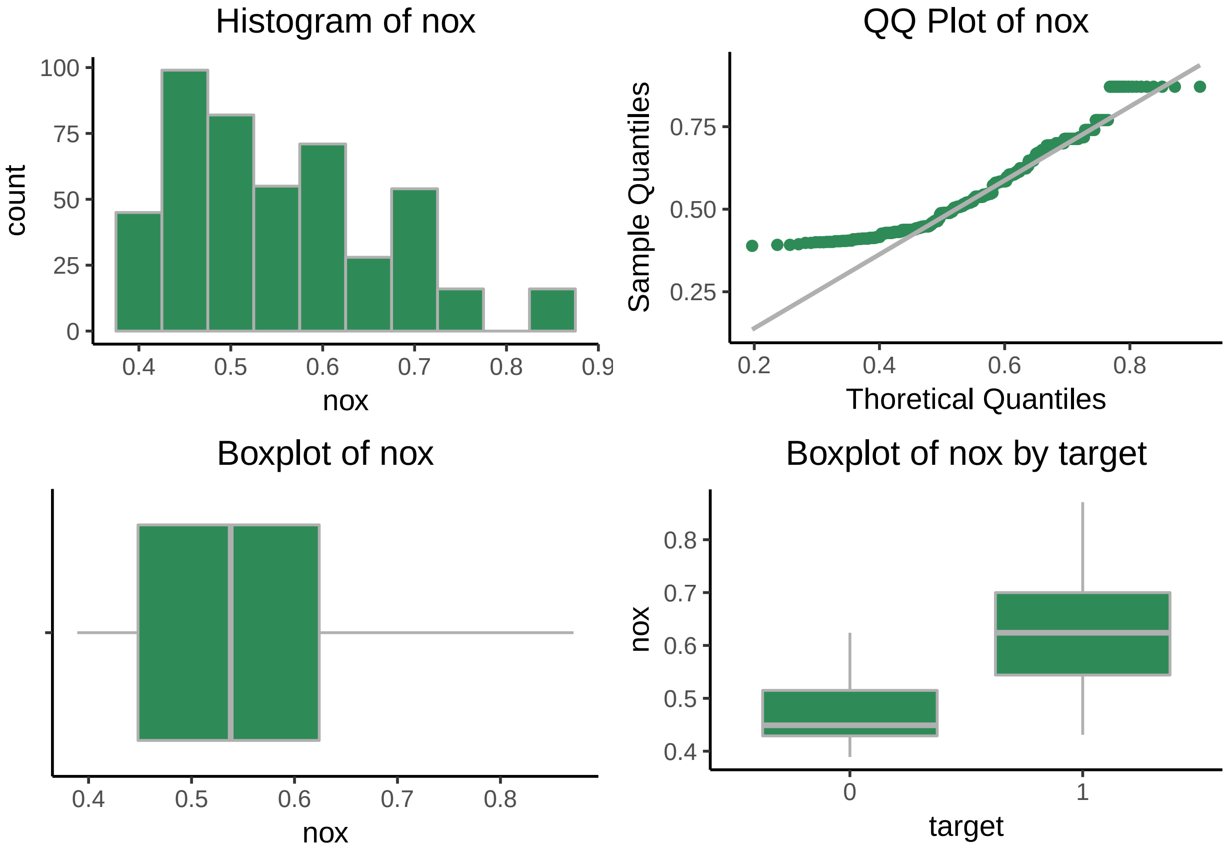
Here is a summary of observed values, rounded to the nearest 0.1 parts per 10 million:

```
0.4 0.5 0.6 0.7 0.8 0.9 Sum
122 149  95  76   8  16 466
```

Below are summary statistics:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD    Skew    Kurt
   0.39    0.45    0.54    0.55    0.62    0.87    0.12    0.75    2.98
```

Now, let's look at the plots:



Response Variable: rm

The predictor `rm` is count measure describing the average number of rooms per dwelling. The distribution is roughly bell-shaped but with somewhat heavier tails than is typical in a normally distributed variable. Lower crime areas tend be associated with a higher number of rooms per dwelling, however the room counts are fairly close between crime types. The room count variances by crime count also do not appear to be drastically different.

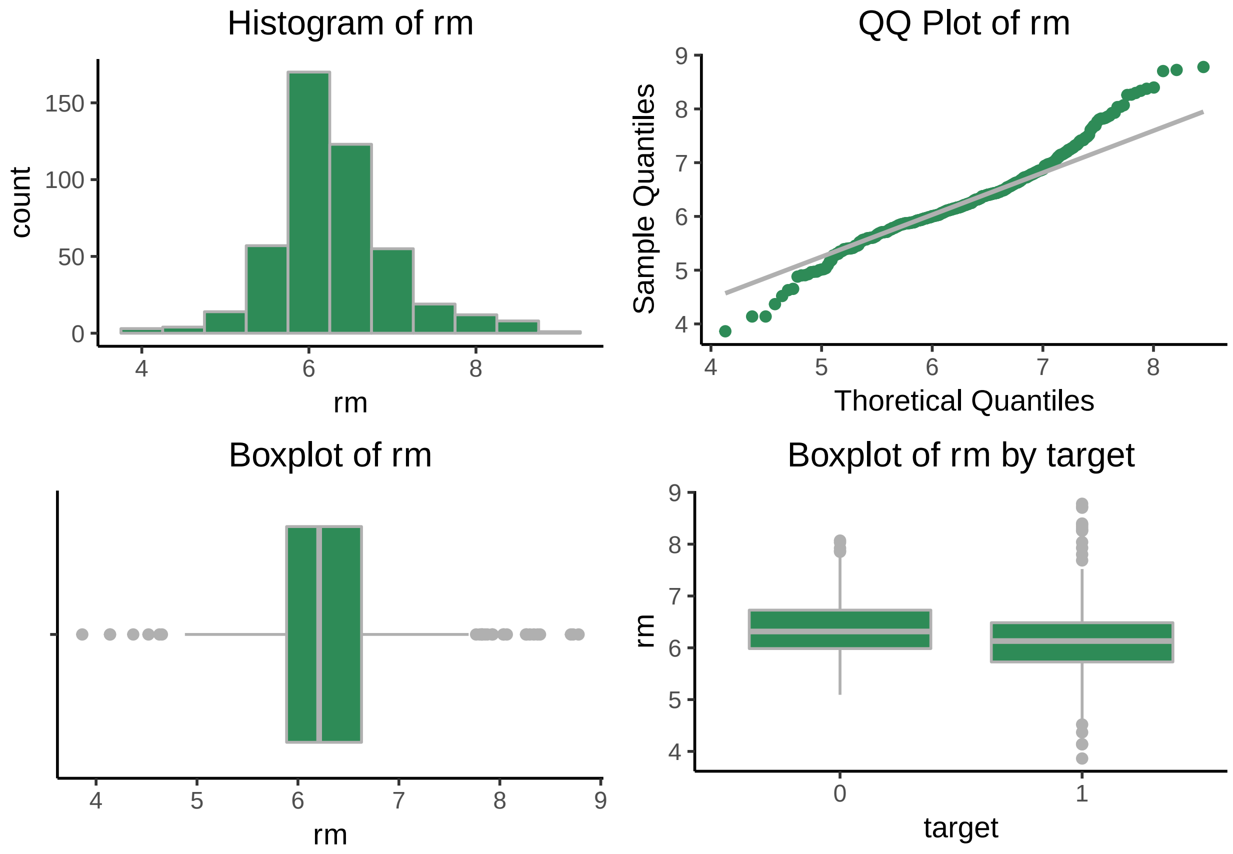
Here is a distribution of room counts, rounded to the nearest whole number:

```
  4   5   6   7   8   9 Sum
  4  37 284 115  23   3 466
```

Summary statistics are below:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD    Skew    Kurt
   3.86    5.89    6.21    6.29    6.63    8.78    0.70    0.48    4.56
```

Finally, here are the plots:



Response Variable: age

The variable `age` indicates the proportion of owner occupied units built prior to 1940. This predictor has a significant left skew. This result is not surprising, given the many historical districts within the greater Boston area. In the boxplots below, we see a significantly higher mean percentage of older homes in high crime areas. This result is expected, given that older neighborhoods tend to be located in dense urban areas.

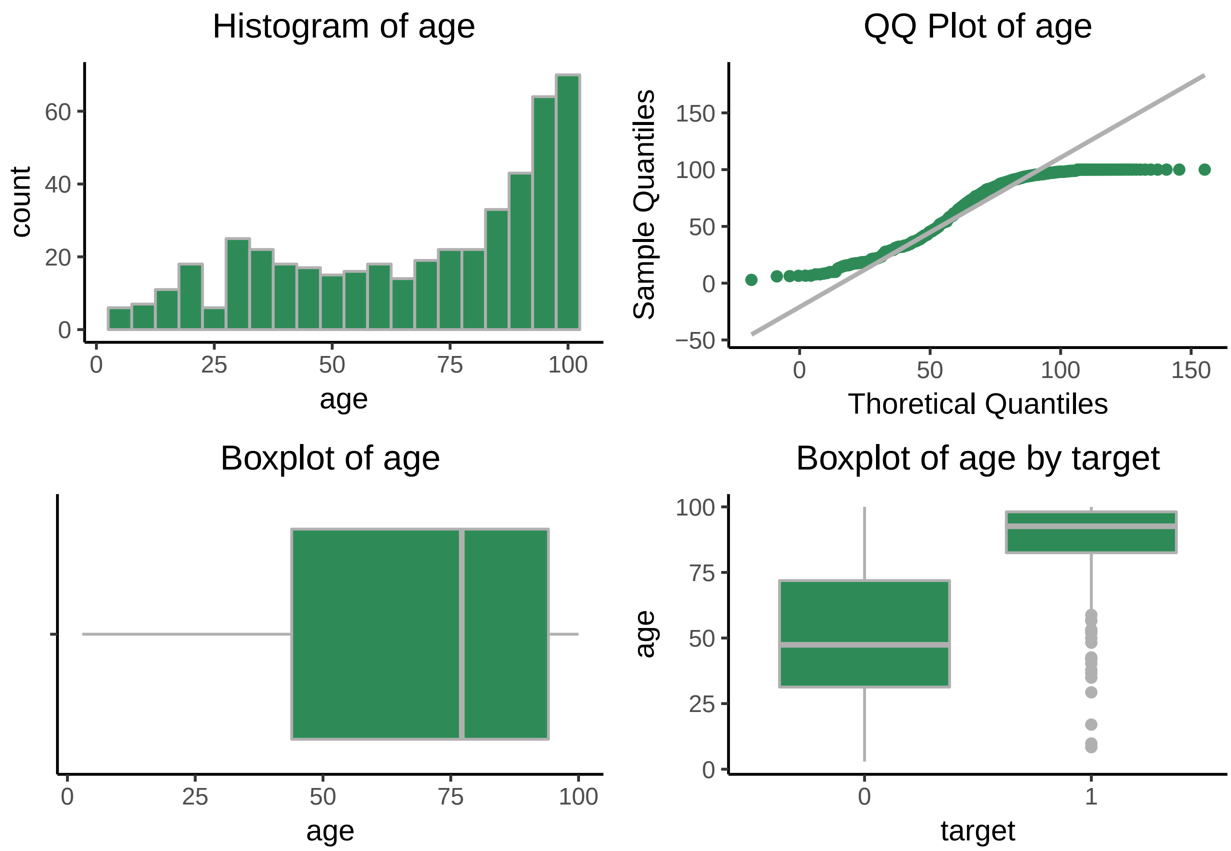
Here is a table of `age` values rounded to the nearest 5th percentage.

```
  5  10  15  20  25  30  35  40  45  50  55  60  65  70  75  80  85  90  95 100 Sum
  6   7   9  20   6  25  22  18  17  15  16  18  14  19  22  22  33  43  64  70 466
```

Let's review the summary statistics:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD    Skew    Kurt 
   2.90   43.88   77.15   68.37   94.10  100.00   28.32   -0.58    2.00 
```

Now, let's look at the plots:



Response Variable: dist

The predictor `dist` describes the average distance to Boston employment centers. The variable is moderately right skewed. Based on the boxplots below, we see see that low crime areas are associated with higher average distances to employment centers. This result is consistent with our intuition, as major employment centers tend to be in dense, urban areas.

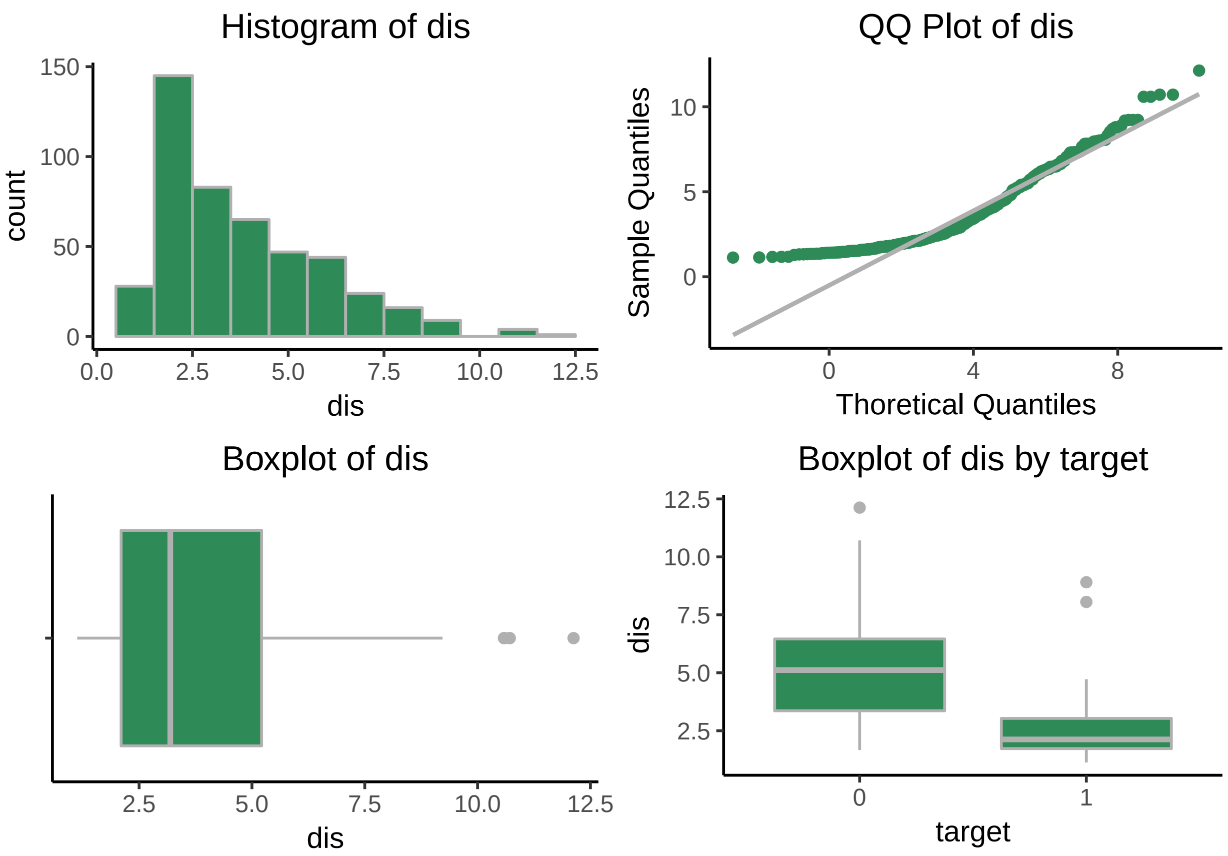
Here is a table of distance values, rounded to the nearest unit:

```
  1   2   3   4   5   6   7   8   9  11  12 Sum 
 28 145  83  65  47  44  24  16   9   4   1 466 
```

Summary statistics are as follows:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD    Skew    Kurt 
    1.1     2.1     3.2     3.8     5.2    12.1     2.1     1.0     3.5 
```

Plots are below:



Response Variable: rad

The `rad` variable is an integer-valued index measure indicating an area's accessibility to radial highways. As stated earlier, we assume this variable contains ordinal, categorical data. In the boxplots below, there appears to be a significant positive association between high crime rates and `rad` value. In fact, the high crime rate areas appear to be heavily concentrated in areas with `rad` values of 24.

The distribution of this variable is tri-modal, with values clustering around index values of 4, 5, and 24.

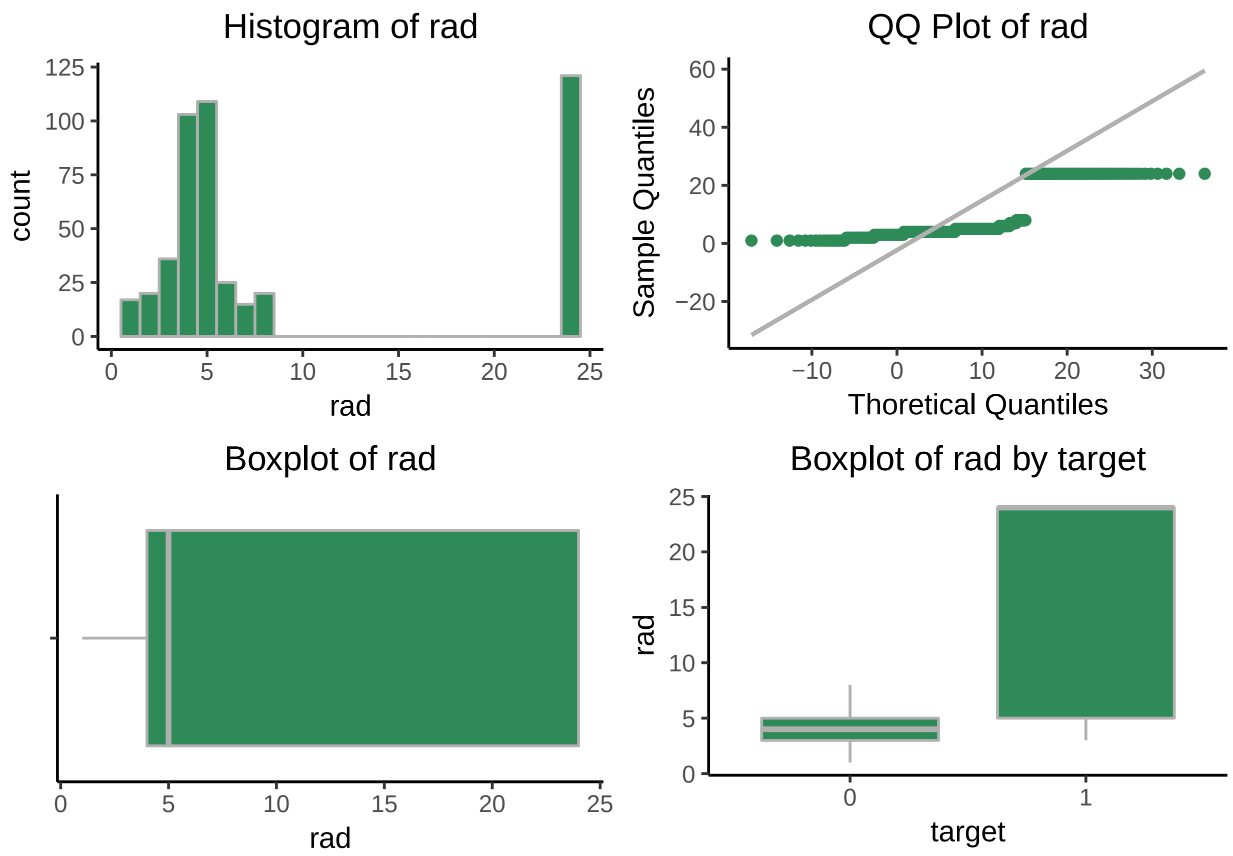
Here is a numerical distribution of the index values:

```
  1   2   3   4   5   6   7   8  24 Sum
 17  20  36 103 109  25  15  20 121 466
```

Below are summary statistics:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD    Skew    Kurt
    1.0     4.0     5.0     9.5    24.0    24.0     8.7     1.0     2.1
```

Finally, here are the plots:



Response Variable: tax

The `tax` variable refers to the the tax rate per $10k of property value. This variable is densely distributed around two of the following approximate values: 300 and 700–the latter value is close tothe mode of the distribution. High crime areas also appear to have a strong, postive association with the `tax` value–see the boxplots below.

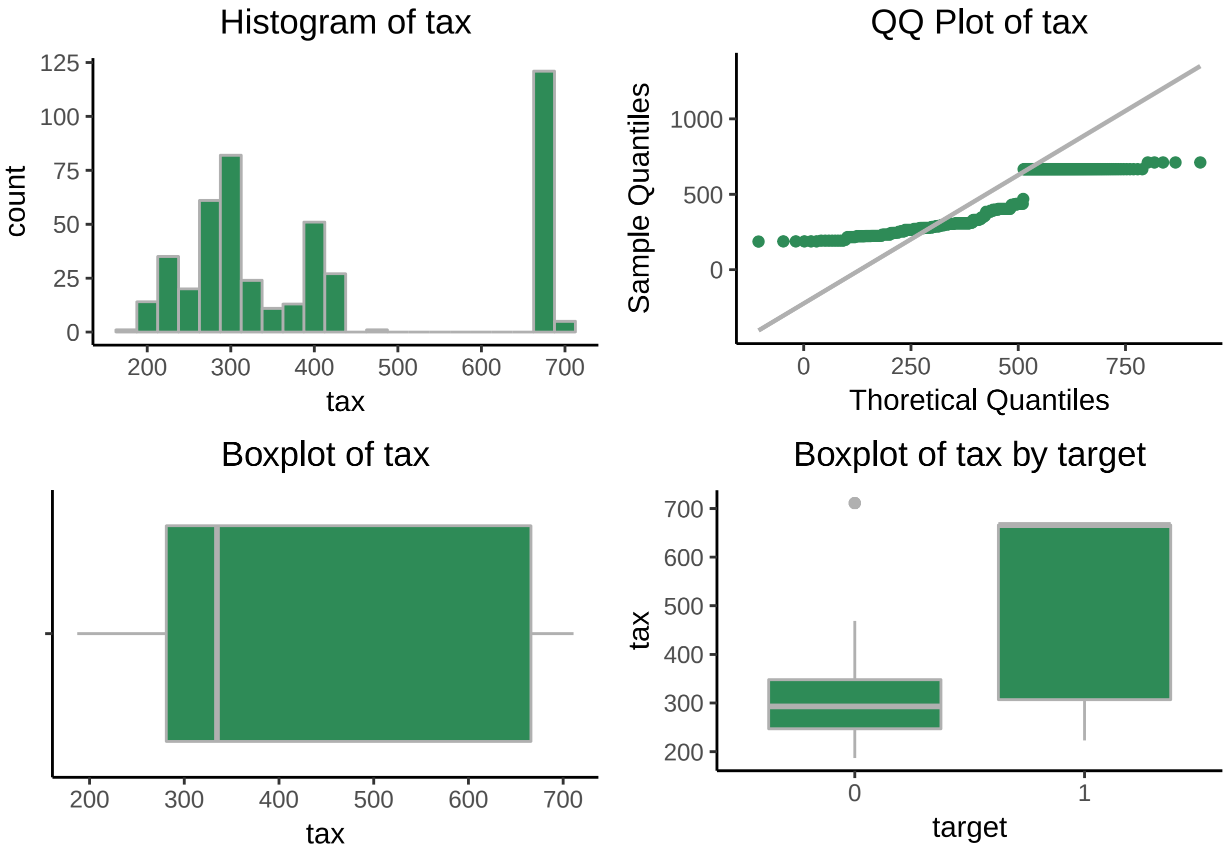
Here is distribution of `tax` values, rounded the nearest $25:

```
175 200 225 250 275 300 325 350 375 400 425 475 675 700 Sum
  1  14  35  20  61  82  24  11  13  51  27   1 121   5 466
```

Now, here are the summary statistics:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD    Skew    Kurt
 187.00  281.00  334.50  409.50  666.00  711.00  167.90    0.66    1.86
```

Plots are below:



Response Variable: ptratio

The predictor `ptratio` indicates the average school, pupil-to-student ratio, and has a right skewed distribution. The mode of the distribution is roughly 20, which is relatively close to the maximum value of 22. The boxplots below indicate a positive relationship between `ptratio` and high crime.

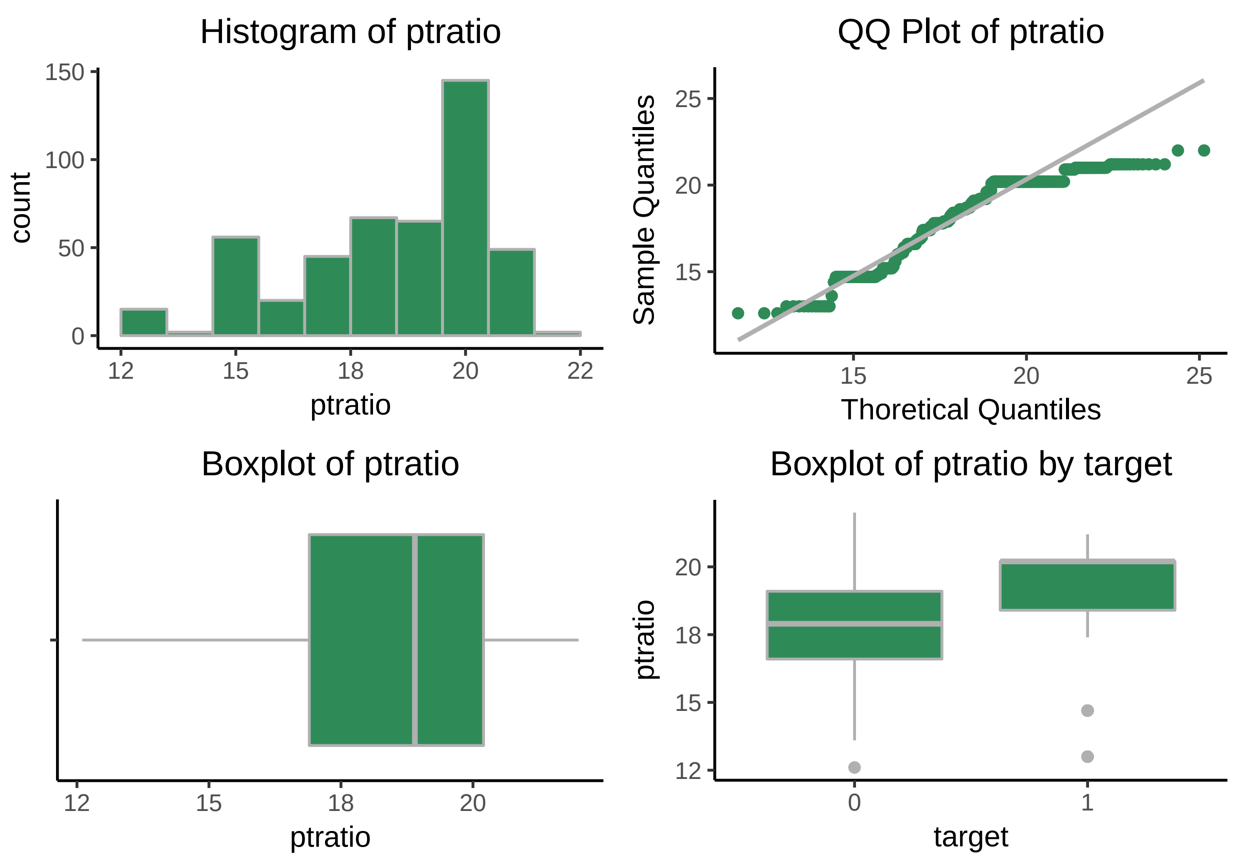
Here is a distribution of `ptratio` values, rounded to the nearest whole number:

```
 13  14  15  16  17  18  19  20  21  22 Sum 
 15   2  55  21  45  67  65 145  49   2 466 
```

Below are the summary statistics:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD    Skew    Kurt 
  12.60   16.90   18.90   18.40   20.20   22.00    2.20   -0.76    2.61 
```

Plots are below:



Response Variable: lstat

The variable `lstat` indicates the proportion of the population deemed to be of lower status. The variable is right skewed, and high crime areas tend to have be associated with larger `lstat` values.

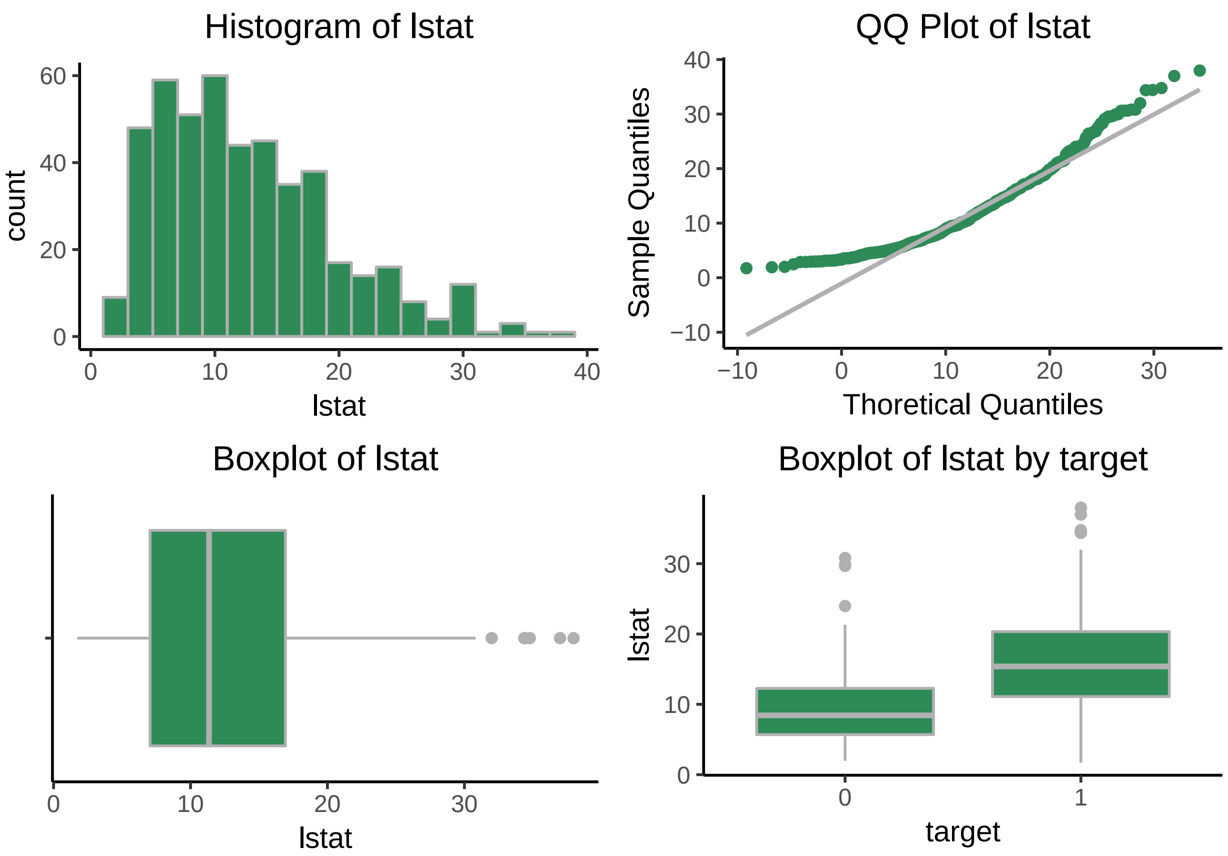
Here is a summary table of `lstat` values rounded to the nearest 2:

```
  2   4   6   8  10  12  14  16  18  20  22  24  26  28  30  32  34  36  38 Sum
  9  48  59  51  60  44  45  35  38  17  14  16   8   4  12   1   3   1   1 466
```

Summary statistics are provided below:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD    Skew    Kurt
   1.73    7.04   11.35   12.63   16.93   37.97    7.10    0.91    3.52
```

Let's look at the plots:



Response Variable: medv

The last feature variable in our data set is `medv`, which represents the median value of residential homes in a given area, in $1,000s. The variable is right skewed, and high values of `medv` appear to be associated with lower crime rates. Variances of `medv` by crime type (e.g. high or low) are similar.

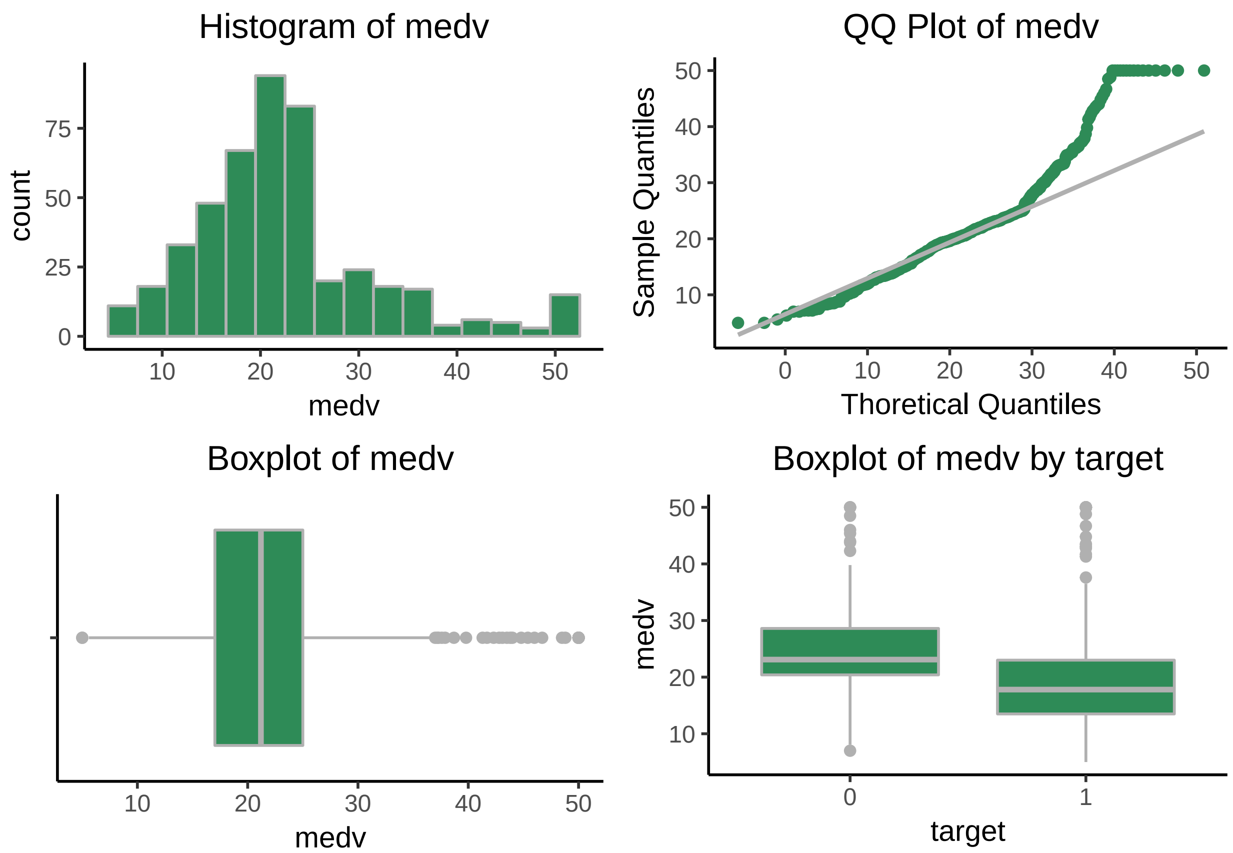
Let's look at `medv` values, rounded to the neares $3k:

```
  6   9  12  15  18  21  24  27  30  33  36  39  42  45  48  51 Sum
 11  17  34  46  69  91  86  19  25  18  17   4   6   5   3  15 466
```

Now, let's examine the summary statistics:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD    Skew    Kurt
    5.0    17.0    21.2    22.6    25.0    50.0     9.2     1.1     4.4
```

Finally, here are the plots:



Target Variable: target

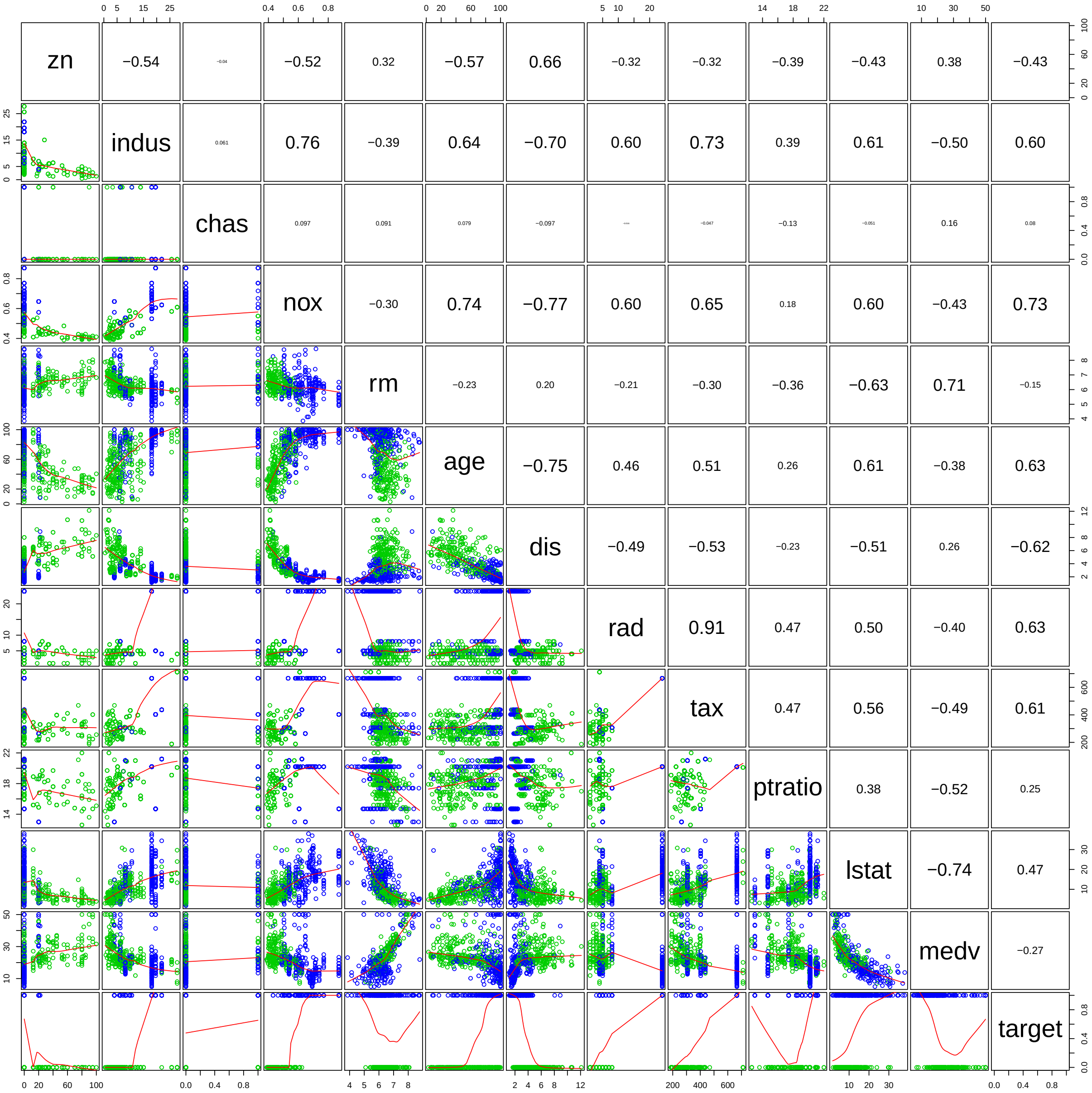
Our binary response variable, `target`, indicates whether a particular area has a crime rate above the median Boston crime rate. As such, we expect roughly half of the observations to have a value of 0, and the other half to have a value of 1. Let's check:

```
  0   1 Sum
237 229 466
```

The split is very close to 50/50.

Bivariate Relationships

Let's look at scatter plot/correlation matrix to get a better feel for the relationships between pairs of variables. In the figure below, we plotted the high crime areas in blue and the low crime areas in green. We also included a loess curve in the scatter plots to get a better feel for the relationship between variables.



Multicollinearity Concerns

There appear to be many moderate pairwise correlations between the predictor variables in our data set. However, only variables with high correlations should be problematic for model interpretation. Let's review pairwise correlations with absolute values of 0.75 or higher:

- `indus` and `nox`: correlation value of 0.76. This result makes sense, as we expect areas with dense industry concentration to have higher environmental pollutants such as NO_2

- `dist` and `nox`: correlation value of -0.77. This result is consistent with our intuition: we expect areas close to employment centers to have higher concentrations of environmental pollutants, and areas farther away to have lower concentrations.

- `rad` and `tax`: correlation value of 0.91. Access to radial highways and tax rates appear are strongly correlated values. We are particularly concerned about the multicollinearity effects of these two variables.

To assess the impact of multicollinearity, we can fit an OLS model to our full data set and compute the variance inflation factors (VIFs). Below is the VIF output from fitting this regression model:

```
      zn    indus     chas      nox       rm      age      dis      rad      tax ptratio    lstat     medv
    2.32     4.12     1.09     4.51     2.35     3.13     4.24     6.78     9.22    2.01     3.65     3.67
```

The highest VIF values are associated with the `rad` and `tax` variables. We can bring all VIFs to an acceptable level by removing one of these two variables. For instance, if we remove `tax`, our VIF factors are:

```
      zn    indus     chas      nox       rm      age      dis      rad ptratio    lstat     medv
    2.18     3.24     1.08     4.50     2.35     3.13     4.24     2.23    2.01     3.64     3.59
```

Unusual Predictor/Target Relationships

We noted a few complex loess curve shapes relating our predictor variables to the target variable. These relationships may reflect actual, highly nonlinear relationships with the target variable, or could be the result of interactions with other predictors. Alternatively, the loess fits could be the result of fitting to relatively sparse data points:

- `rm`: Our loess curve initially indicates a negative relationship between the number of rooms and the crime rate category. However, when the number of rooms exceeds approximately 7, the relationship becomes positive. This strange loess curve shape is most likely the result of the model fitting to sparse data, as there are only about 30 observations with an average room count in excess of 7. In general, we expect higher room counts to be associated with lower crime rates.
- `ptratio`: we would expect crime rates to be higher in areas with high pupil-teacher ratios. However, the loess curve initially indicates an increasing propensity for high crime rates with increases to the this `ptratio`. Because this variable is left skewed with a high density of ratios clustered around 20, we belivve this unusual curve shape is due to the loess model fitting to sparse data at low ratio values.
- `medv`: We expect high median home values to be associated with higher median values. This pattern appears to hold in our loess curve for median values below approximately $30k. The pattern then reverses for values above $30k. Once again, we believe this pattern reversal is due to sparse data. The variable `medv` is right skewed, with relatively few data observations where the median value exceeds $30k$

Data Preparation

In this section, we describe data modifications and transformations applied before fitting our regression models.

Missing Values

As mentioned previously, our data set contains no missing elements; so we do not need to use any imputation procedures.

Outliers and Unusual Observations

At this stage, we see no clear data entry errors in our data set. We will examine potential outliers and influential points once we have fit initial models to the training data.

Variable Transformations

In binary logistic regression, it's desirable to have predictor variables that are normally distributed, whenever possible. When the predictor is normally distributed with similar variances across both target variable valuees, then the log odds are a linear function of the predictor. And when the predictor is normally distributed with different variances for the seprate target response values, then the log odds are a quadratic function of the predictor.

Let's review each predictor variable and review transformation procedures, if applicable.

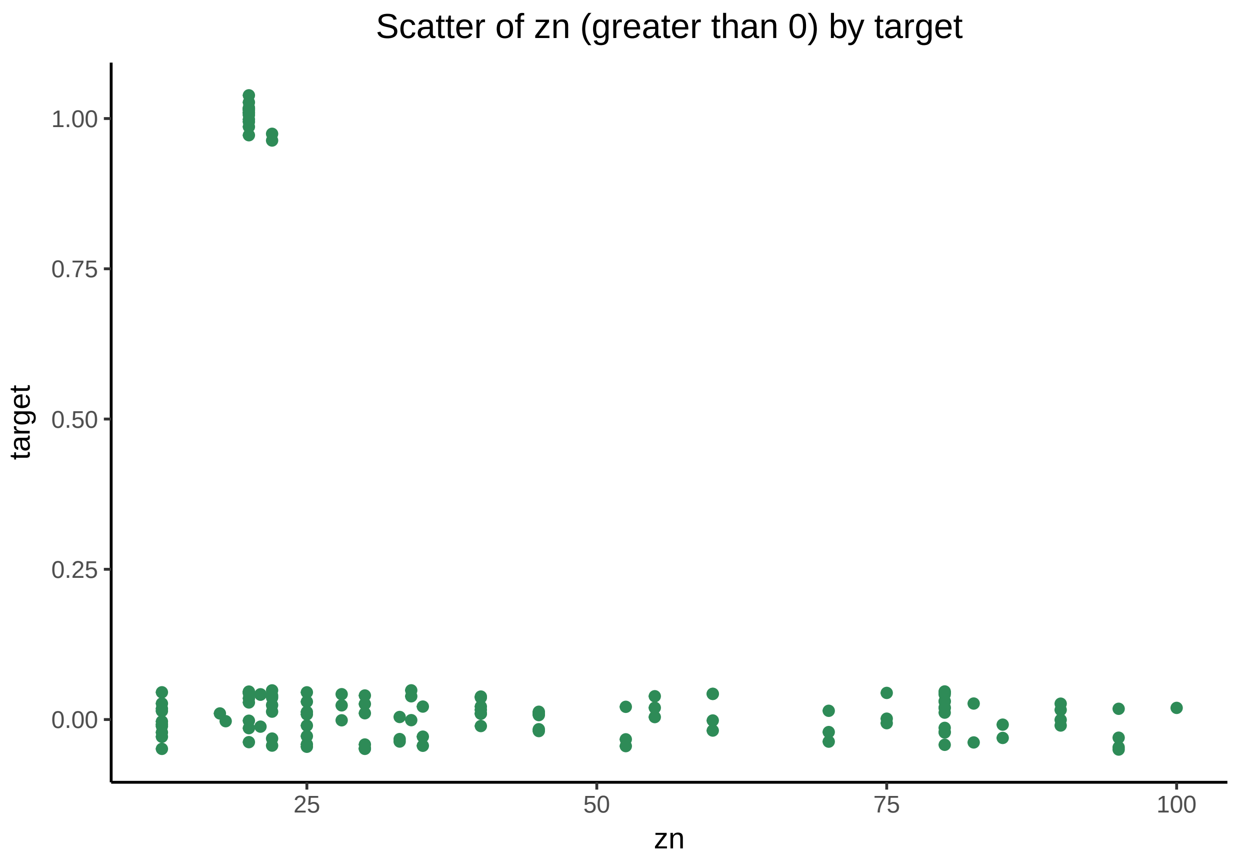
zn

The `zn` predictor is highly right skewed. The data also has many zero values–over 70% of observations. Let's compare the percentage of high crime areas for zero-valued `zn` vs. values 1 or higher:

```
     zn_zero zn_1_plus
[1,]   0.631     0.118


    2-sample test for equality of proportions with continuity correction

data:  table(crime$zn_zero, crime$target)
X-squared = 100, df = 1, p-value <2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.432 0.595
sample estimates:
prop 1 prop 2
 0.882  0.369
```

The difference in these proportions is statistically significant. Given that the vast majority of high crime areas also have a zero `zn` value, we will proceed to create a new, categorical variable, `zn_zero.` A value of 1 indicates that no land is zoned for large residential lots. A value of 0 indicates that at least some land is zoned for large residential lots.

Finally, let's look at a jittered scatter plot of nonzero `zn` by target value:



There are very few nonzero observations of `zn` with high crime values, and there does not appear to be a straightforward relationship between the target value and nonzero `zn` proportion; so we will perform no additional transformations with this variable.

indus

We noted previously that the distribution of`indus` has a bi-modal appearance, with values clustered around two ranges. Basic power transformations will not result in an approximately normal or symmeteric distribution. Also, we previously saw that high crime areas are primarily concentrated in areas with high industry concentration. Therefore, we recommend making a new categorical variable, `indus_high`, with a value of 1 if the industry proportion is close to the higher valued mode, and value of 0 if the industry value is close to the lower mode. We've assigned a 0 value to `indus_high` for all `indus` values 14 and lower, and 1 otherwise. The cutoff choice reflects an approximate halfway point between the two mode centers.

Let's do a sanity check of our proposed bifurcation by comparing the high crime rates for each value of the new `indust_high` variable:

```
     val_0 val_1
[1,] 0.231  0.92
```

The percentage of high crime areas are very different for each value of `indus_high`.

chas

The predictor `chase` is a binary, categorical variable. We will leave the variable as-is.

nox

The variable `nox` is moderately right skewed. We perform the box-cox procedure to determine an appropriate transformation:

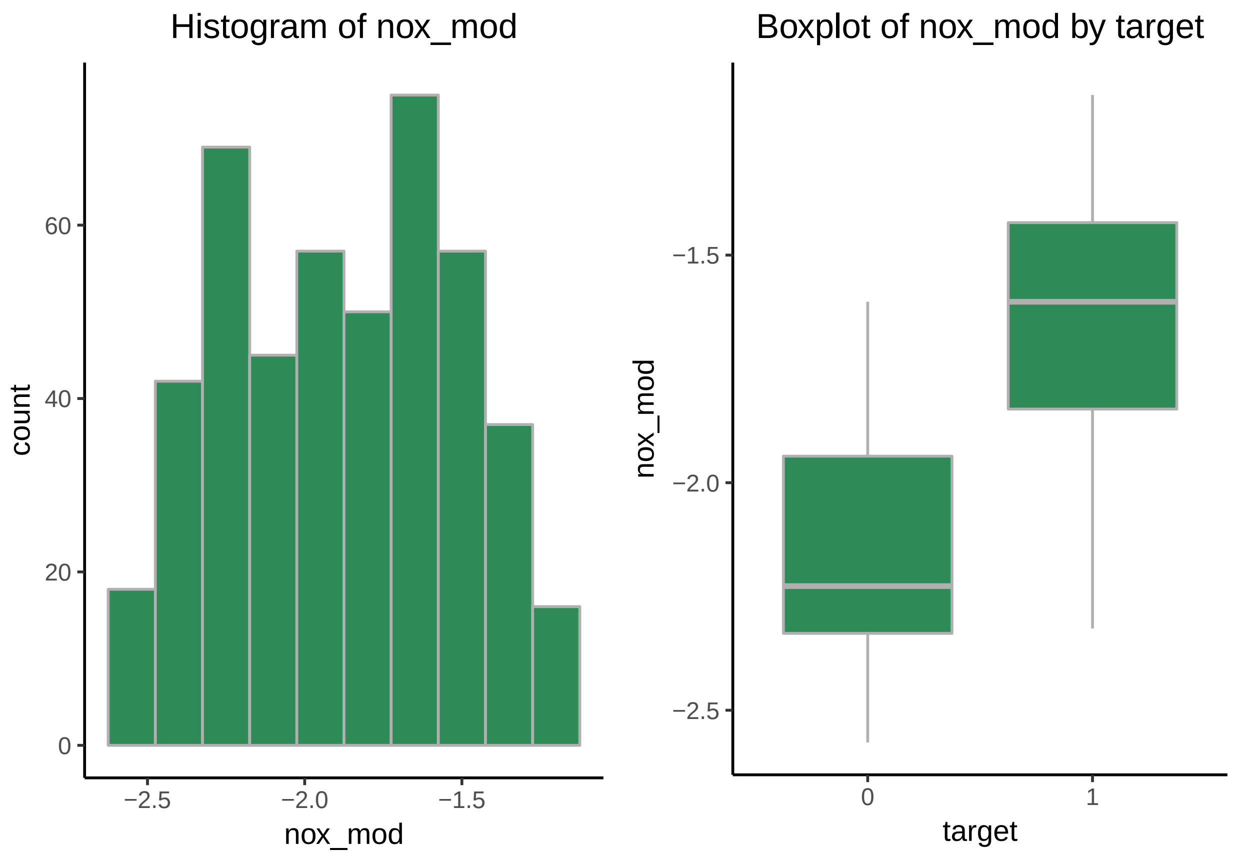
```
Fitted parameters:
 lambda    beta sigmasq 
 -0.949  -0.862   0.126 

Convergence code returned by optim: 0
```

We will create a new variable `nox_mod`, that is the reciprocal of the raw `nox` value. We then multiply the reciprocal by -1 to preserve the direction of the original relationship.

The transformed variable is more symmetrical, with a skewness value closer to zero:

```
[1] 0.0487
```

The variances for each target value are also similar–see below:



rm

The variable `rm` had a mild positive skew and high kurtosis value. Let's look at the suggested box cox transformation:

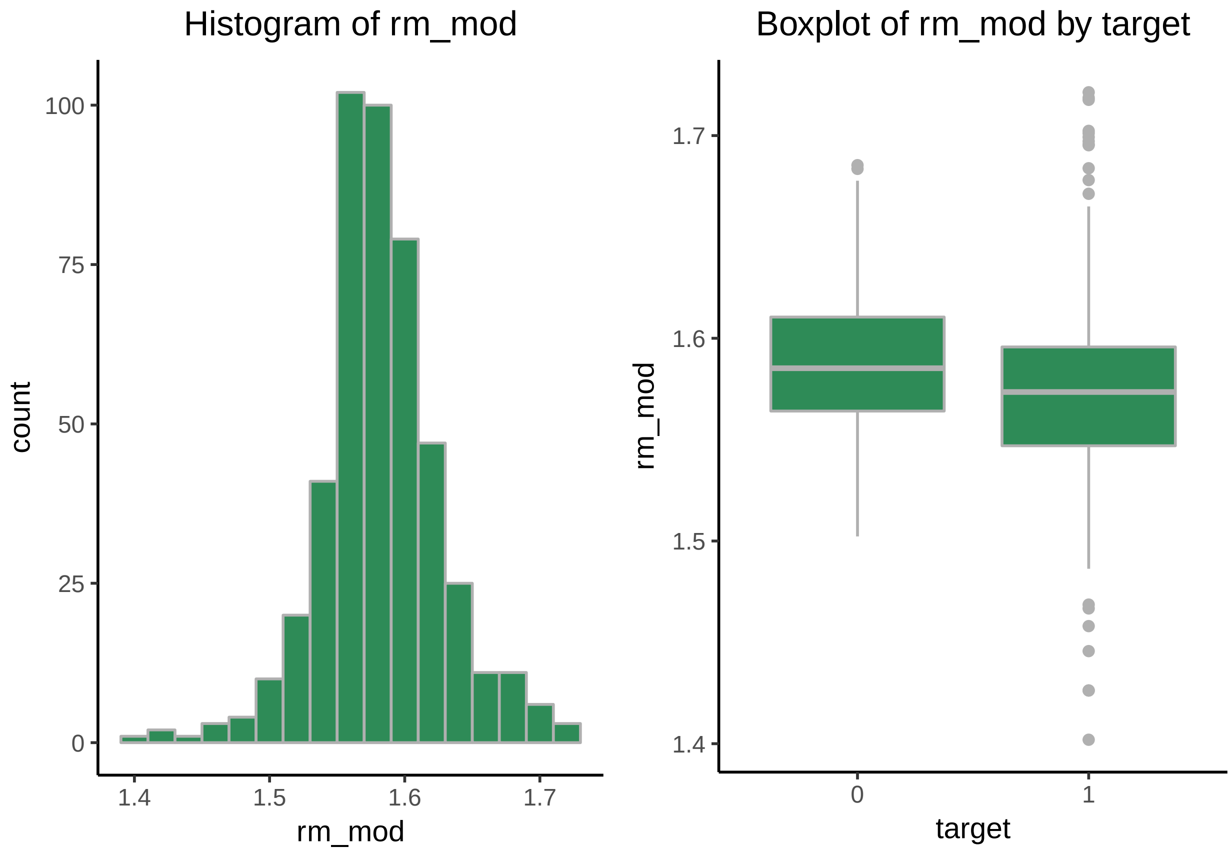
```
Fitted parameters:
 lambda    beta sigmasq 
 0.2038  2.2239  0.0263 

Convergence code returned by optim: 0
```

Based on this output, we will transform the variable by taking the quarter root of the raw value.

The transformed variable is more symmetric, with a skewness value of:

```
[1] 0.0416
```

The variances of `rm_mod` for each target value appear to be fairly similar:

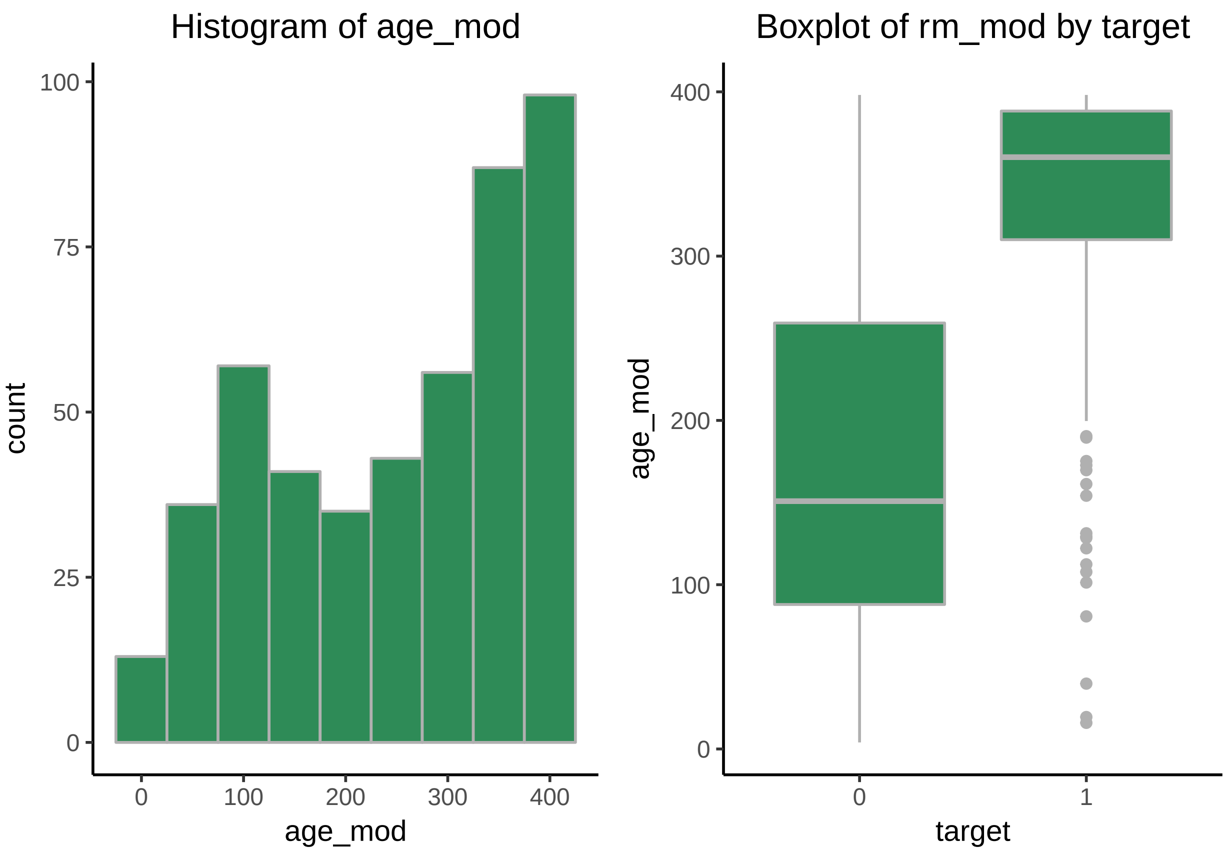


age

Age has a moderate left skew. Let’s review the suggested box-cox tranformation:

```
Fitted parameters:
  lambda     beta  sigmasq
    1.32   205.70 10492.78

Convergence code returned by optim: 0
```

We apply the suggested power transformation of 1.3 and store in a new variable, `age_mod`.



The transformed variable's skew has improved slightly, but there is still see significant negative skew. The variances are also significantly different across the two values of the target variable.

dis

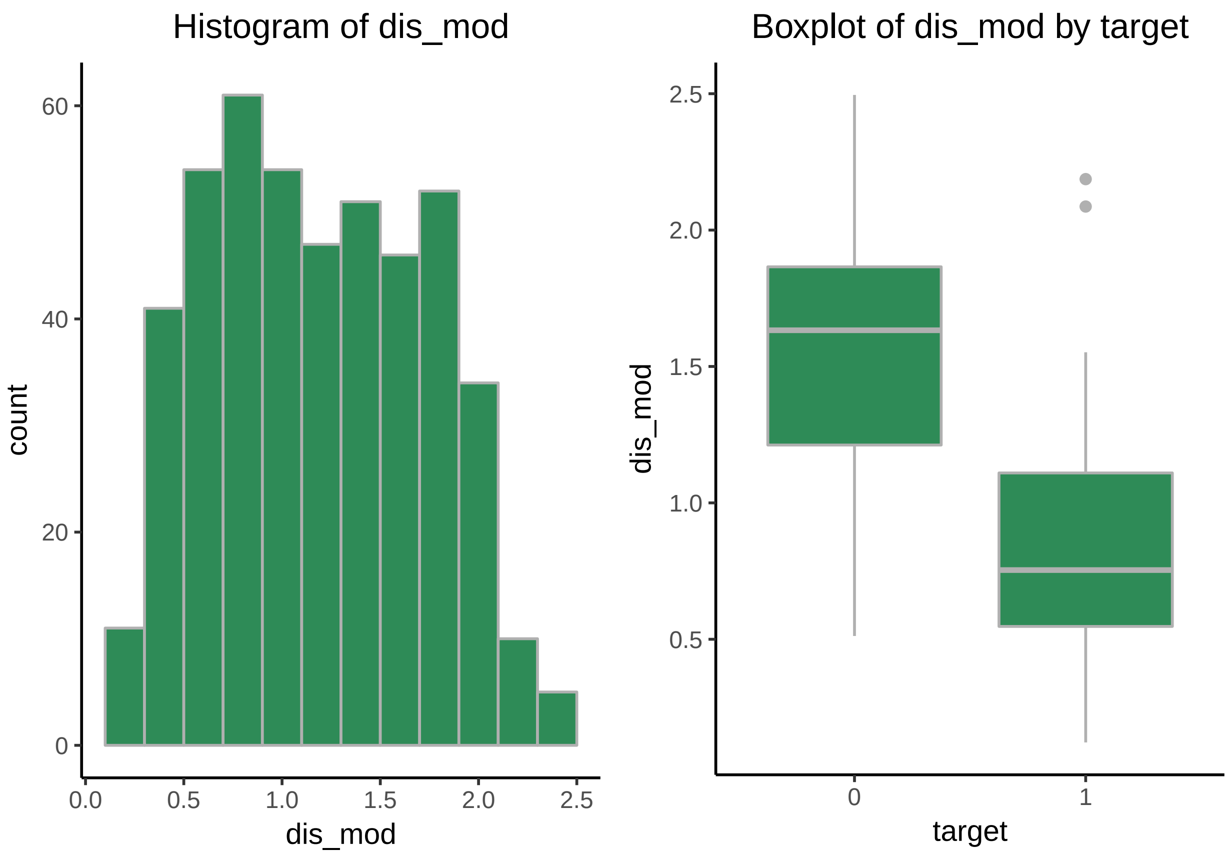
The predictor `dis` has a moderate positive skew. Let's transform using the box-cox transformation:

```
Fitted parameters:
 lambda    beta sigmasq
 -0.147   1.072   0.205

Convergence code returned by optim: 0
```

Given that the value of the lambda parameter is fairly close to 0, we will use the log transformation and save to results to a new variable, `dis_mod`.

The transformed distribution has improved skew:

```
[1] 0.143
```



The transformed variable has similar variances across each target value.

rad

The predictor `rad` has a multi-modal appearance, with values densely clustered around 4-5 and 24. Given this dense clustering–and the limited effectiveness of power tranformations to generate an approximate normal distribution–we'll create a categorical variable, `rad_high`, that assigns a value of 1 when the rad index level is 15 or higher, and 0 otherwise. The choice of 15 as a cutoff reflects an approximate halfway point between the the two cluster centers. Note: In our training data there are no `rad` measures between 9 and 23.

As a sanity check to make sure this transformation is reasonable, let's look at the percentage of high crime areas for each value of the new `rad_high` variable:

```
     val_0 val_1
[1,] 0.313     1
```

In our training data, all of the `rad_high` values of 1 were located in high crime areas, while only 31% of the 0 values were in low crime areas.

tax

The variable `tax` also has a bi-modal shape, with values densely clustered around 300 and 700–with no values recording in the training data between between 470 and 665. Because power transformations have limited effectiveness in approximating a normal distribution, we'll create a new categorical variable, `tax_high`, that assigns a value of 1 when the tax value is greater than or equal to 500, and 0 otherwise. The 500 cutoff reflects an approximate halfway point between the two modal centers.

Let's perform another sanity check to determine if there is significant relationship with our target variable:

```
     val_0 val_1
[1,] 0.318  0.96
```

We see that 96% percent of the tax_high variables with value one are in high crime areas, and only 32% of the zero values are in high crime areas; so our bifurcation makes sense.

ptratio

The predictor variable ptratio has a moderate negative skew. Let's perform box-cox transformations:

```
Fitted parameters:
  lambda     beta  sigmasq
4.14e+00 4.57e+04 3.61e+08

Convergence code returned by optim: 0
```

The suggested power transformation of 4 does not correct the the left skew, and its implementation also creates unusually large, transformed ptratio values(e.g. 200,000 and higher). Therefore, We will forgo the power transformation for the sake of simplicity. We noted earlier that there appears to be a relationship between high ptratio values and high crime rates. However,this relationship does not appear to be particularly strong; so the variable may not be statistically significant in our subsequent model building step.

lstat

The lstat variable has a moderate positive skew.

Let's look the suggested power transformation from the box-cox procedure:

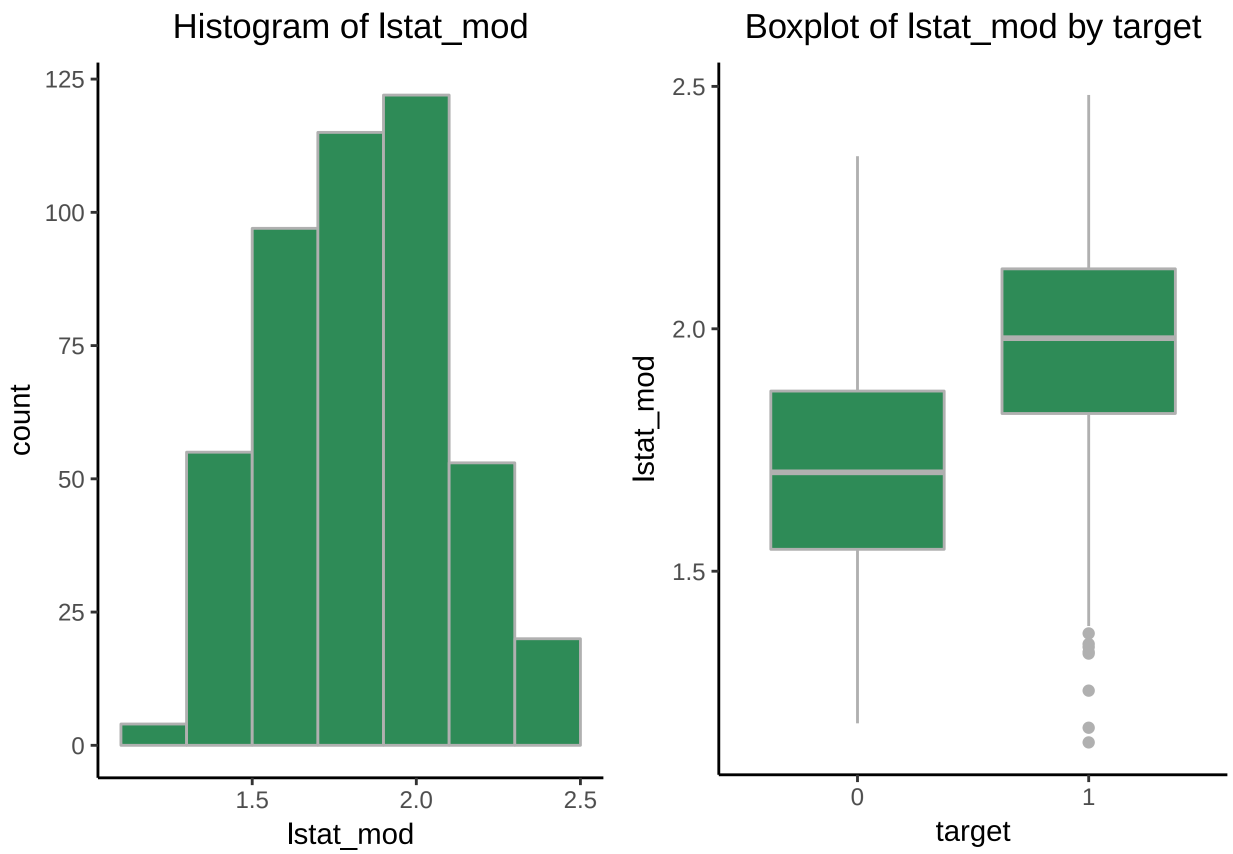
```
Fitted parameters:
 lambda    beta sigmasq
  0.233   3.235   1.055

Convergence code returned by optim: 0
```

Based on this output, we create a new variable lstat_mod, that applies a quarter root transformation to the original variable.

The transformed variable is fairly symmetric, with a skewness value of:

```
[1] -0.00564
```

The variances of the transformed variable are also similar across each target variable value.



medv

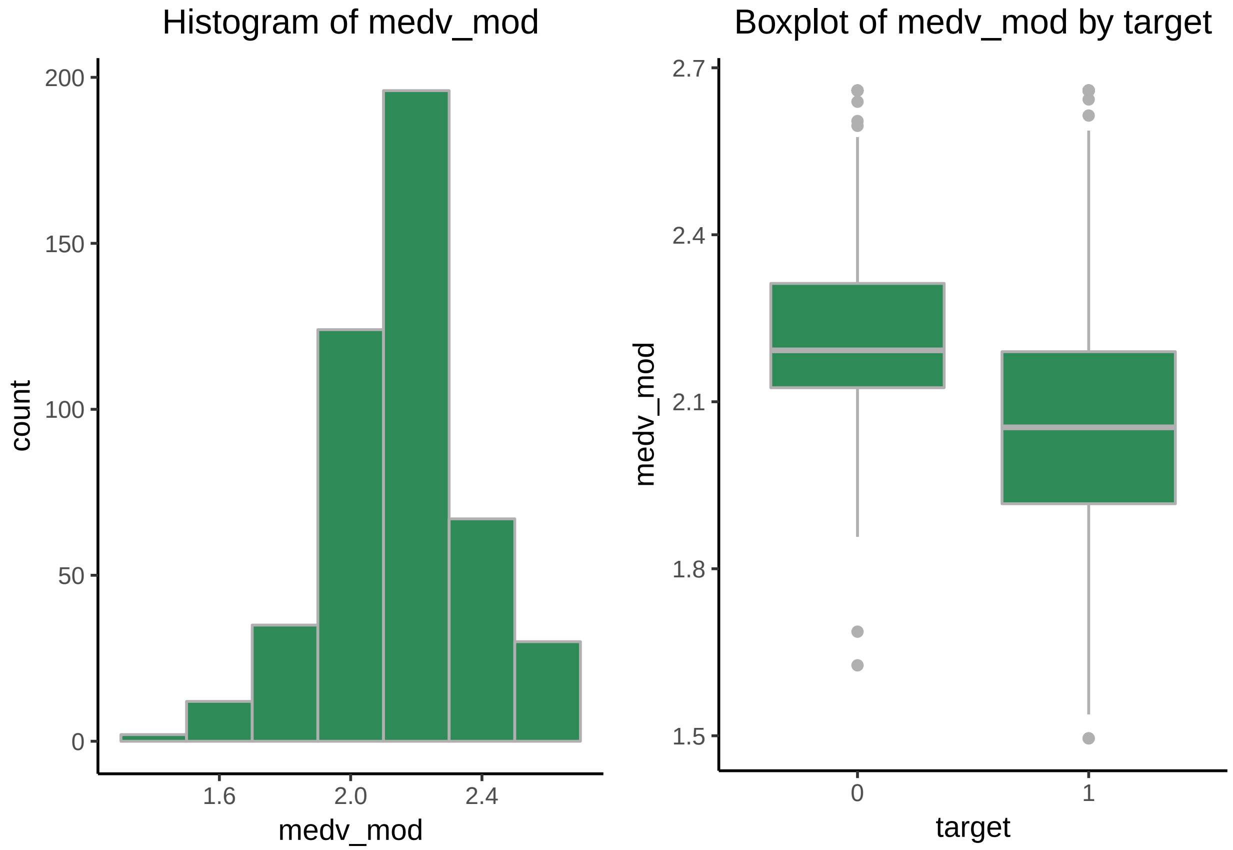
The predictor `medv` has a moderate, positive skew. Let's look a the suggested box-cox transformation:

```
Fitted parameters:
 lambda    beta sigmasq
  0.235   4.469   0.693

Convergence code returned by optim: 0
```

Based on this output, we will apply a quarter root transformation and assign to a new variable, `medv_mod`.

The new transformed variable approximately symmetric, with a skewness value of:

```
[1] 0.0402
```



While the new variable appears to be fairly symmetric for both values of the target value, the variance associated with a target value of 1 appears to be greater than the 0 value counterpart.

Build Model

Variable Overview

Let's take stock of variables that we will consider for our models:

- `zn_zero`: a transformed, binary variable

- `indus_high`: a transformed, binary variable

- `chas`: a binary variable
- `nox_mod`: a transformed, continous variable
- `rm_mod`: a transformed, continous variable
- `age_mod`: transformed, continous variable. This variable is somewhat asymmetric and has different variances for each target value. We may consider a quadratic term in our model to account for these extra challenges.

- `dis_mod`: a transformed, continous variable

- `rad_high`: a transformed, binary variable. We should be not include this variable and `tax_high` together, given the high correlation between these variables.

- `tax_high`: a transformed, binary variable. We should not include this variable in the same model as the `rad_high` variable.

- `ptratio`: This variable has moderate skew and different variances for each value of the target predictor. We may consider adding quadratic terms to our model to account for these challenges.

- `lstat`: a transformed, continous variable
- `medv`: a transformed, continous variable.