Homework 3

2022-11-01

Importing Data

Problem Statement and Goals

In this report, we generate a binary logistic regression model that is able to predict whether or not the crime rate for a neighborhood is above the median crime rate (1) or not (0). The independent and dependent variables that are used in order to generate this model use data from various neighborhoods of a major city. The analysis detailed in this report shows the testing of several models from which a best model was selected based on model performance and various metrics.

Data Exploration

The following is a summary of the variables provided within the data to generate the binary logistic regression model:

- `zn`: proportion of residential land zoned for large lots (over 25000 square feet) (predictor variable)
- `indus`: proportion of non-retail business acres per suburb (predictor variable)
- `chas`: a dummy var. for whether the suburb borders the Charles River (1) or not (0) (predictor variable)
- `nox`: nitrogen oxides concentration (parts per 10 million) (predictor variable)
- `rm`: average number of rooms per dwelling (predictor variable)
- `age`: proportion of owner-occupied units built prior to 1940 (predictor variable)
- `dis`: weighted mean of distances to five Boston employment centers (predictor variable)
- `rad`: index of accessibility to radial highways (predictor variable)
- `tax`: full-value property-tax rate per $10,000 (predictor variable)
- `ptratio`: pupil-teacher ratio by town (predictor variable)
- `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)

A summary of the variables is shown below. See that within the summary, there does not seem to be any extremely high or extremely low values relative to the medians and means for each of the continuous predictor variables. The single binary predictor variable `chas` has reasonable values as well.

```
       zn             indus        chas         nox               rm       
 Min.   :  0.00   Min.   : 0.460   0:433   Min.   :0.3890   Min.   :3.863  
 1st Qu.:  0.00   1st Qu.: 5.145   1: 33   1st Qu.:0.4480   1st Qu.:5.887  
 Median :  0.00   Median : 9.690           Median :0.5380   Median :6.210  
 Mean   : 11.58   Mean   :11.105           Mean   :0.5543   Mean   :6.291  
 3rd Qu.: 16.25   3rd Qu.:18.100           3rd Qu.:0.6240   3rd Qu.:6.630  
 Max.   :100.00   Max.   :27.740           Max.   :0.8710   Max.   :8.780  
      age              dis              rad              tax        
 Min.   :  2.90   Min.   : 1.130   Min.   : 1.00   Min.   :187.0  
 1st Qu.: 43.88   1st Qu.: 2.101   1st Qu.: 4.00   1st Qu.:281.0  
 Median : 77.15   Median : 3.191   Median : 5.00   Median :334.5  
 Mean   : 68.37   Mean   : 3.796   Mean   : 9.53   Mean   :409.5  
 3rd Qu.: 94.10   3rd Qu.: 5.215   3rd Qu.:24.00   3rd Qu.:666.0  
 Max.   :100.00   Max.   :12.127   Max.   :24.00   Max.   :711.0  
```

```
   ptratio          lstat             medv       target
Min.   :12.6   Min.   : 1.730   Min.   : 5.00   0:237
1st Qu.:16.9   1st Qu.: 7.043   1st Qu.:17.02   1:229
Median :18.9   Median :11.350   Median :21.20
Mean   :18.4   Mean   :12.631   Mean   :22.59
3rd Qu.:20.2   3rd Qu.:16.930   3rd Qu.:25.00
Max.   :22.0   Max.   :37.970   Max.   :50.00
```

Figure XX reveals that there are no missing values within the dataset. Therefore, no imputing is required for this dataset.

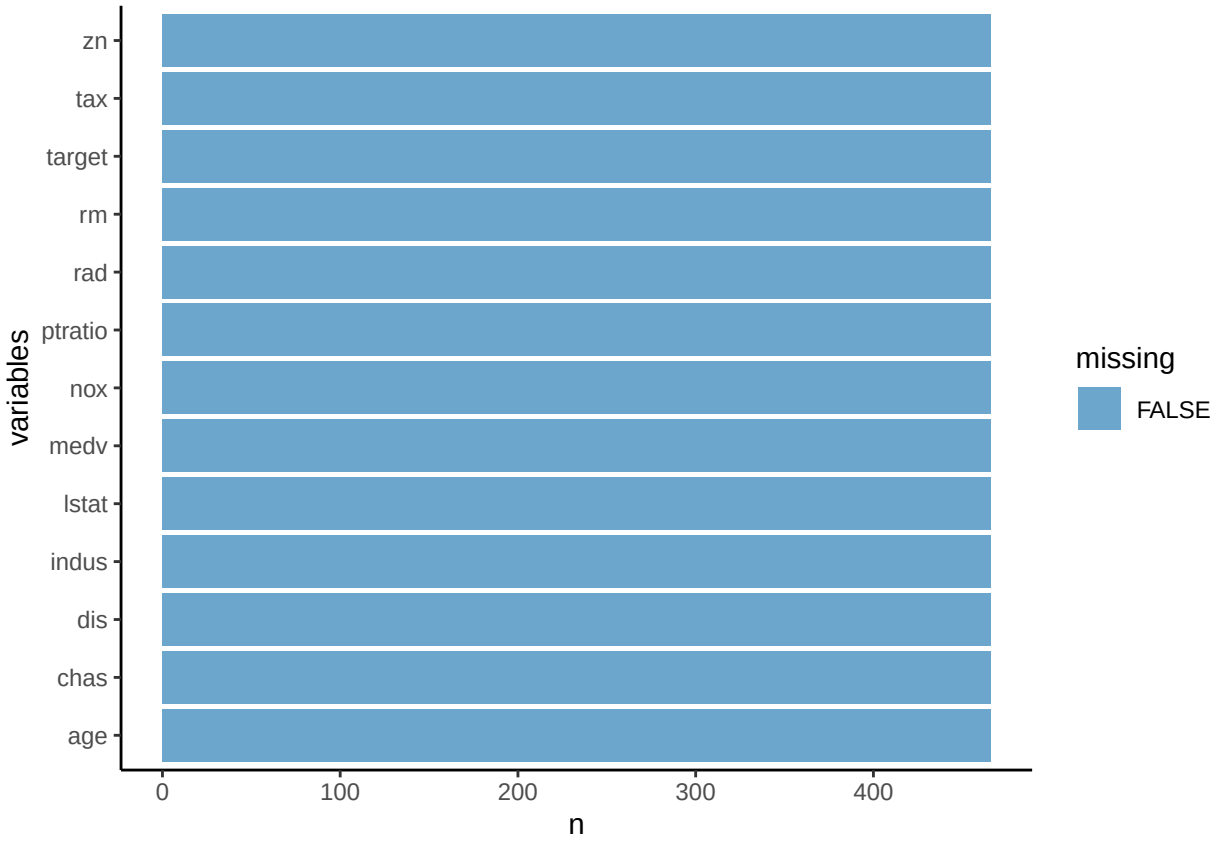


Figure XX: Chart showing the count of missing values for each of the variables in the dataset. Note that since there are no missing values, the legend only shows one item.

Outliers

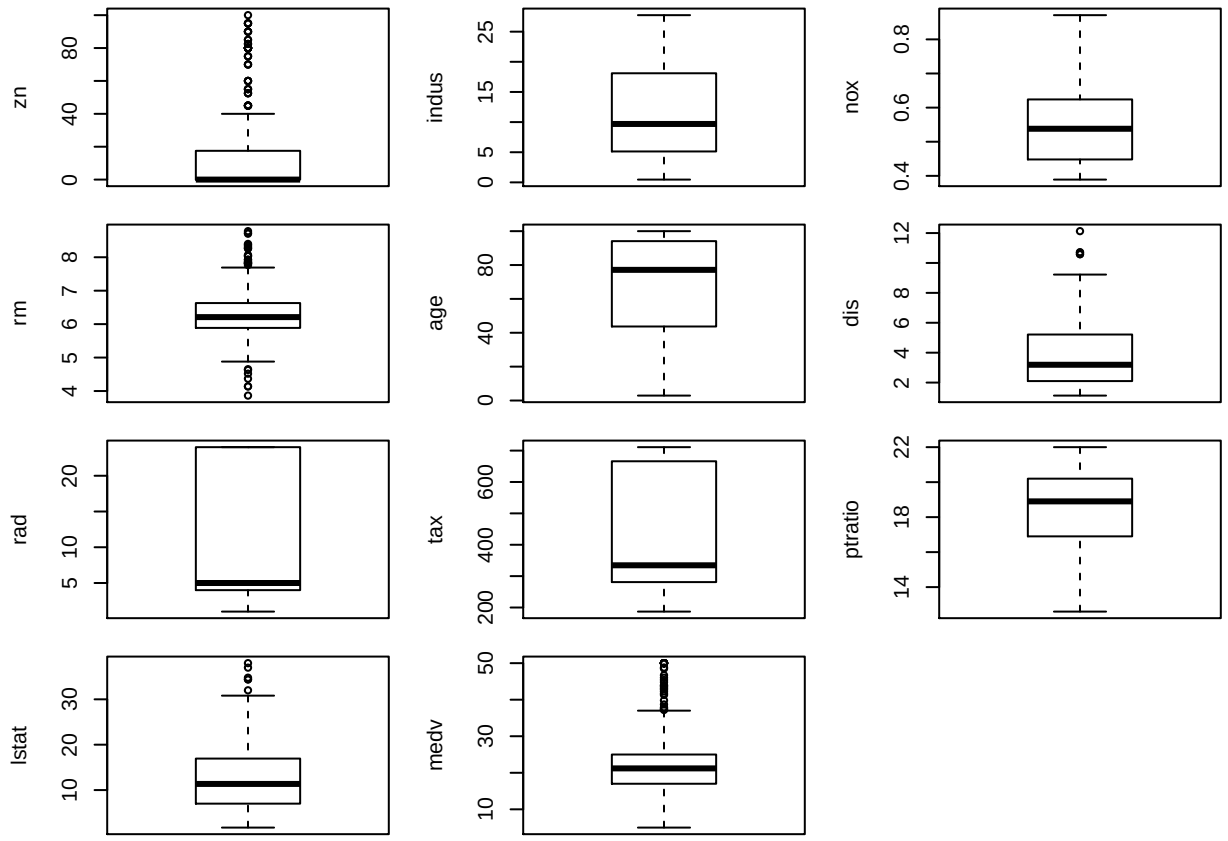


Figure XX: Box plots for each of the variables in the dataset.

Figure XX shows boxplots for the continuous variables. While `zn`, `rm`, `dis`, `lstat` and `medv` contain outliers, the outliers in general do not seem to be any significant enough to affect the model greatly. However. note that `rad` and `tax` have significantly large interquartile ranges which indicates skewness.

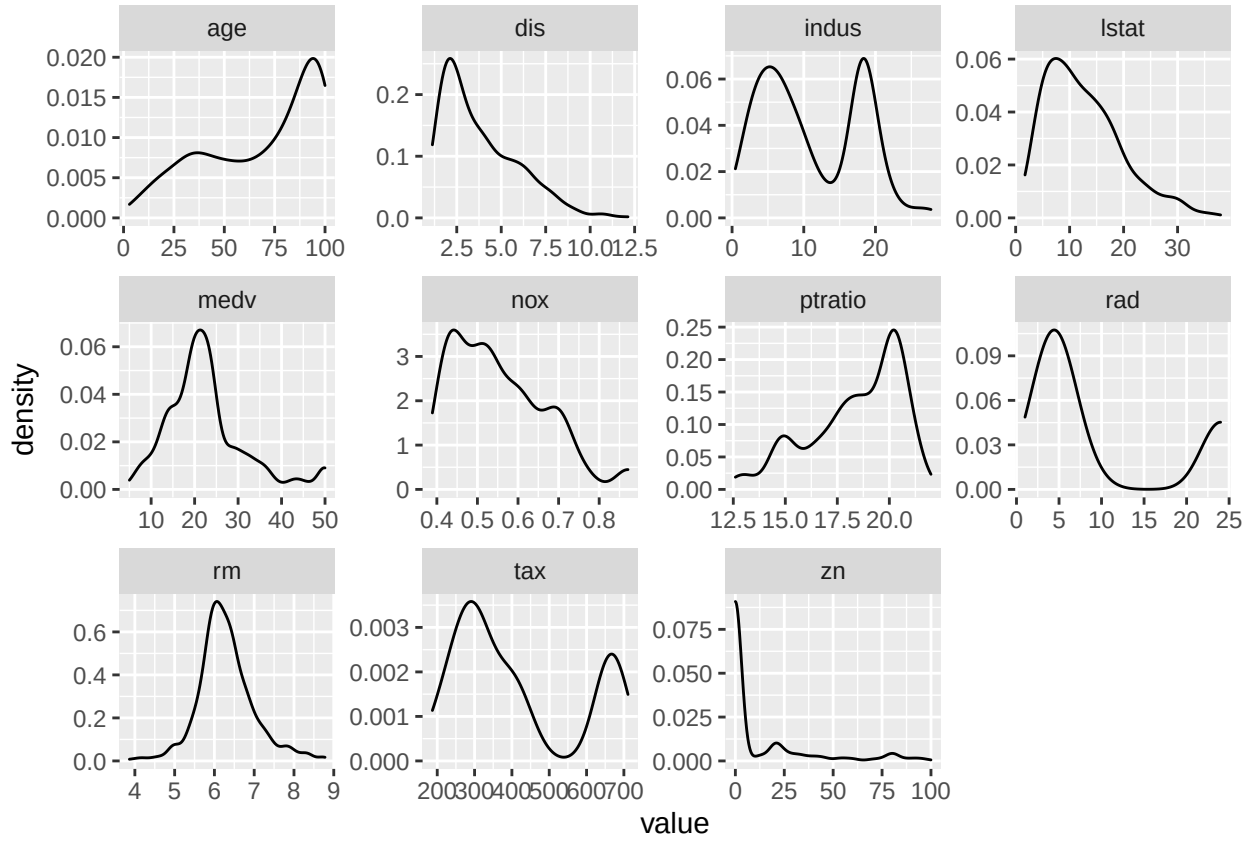


Figure XX: Density plots for continuous variables

Figure xx reveals that `tax`, `indus`, and `rad` have bimodality. `age` appears to have bimodality as well but it is not as pronounced as the others. `rm` is relatively normally distributed while all of the other variables possess skewness, with `zn` possessing extreme skewness. Dummy variables for each of the bimodal variables were created and are given an explanation in the "Dealing with Bimodal Variables" section.

Simple Model

A simple model was generated using all of the predictors and served as a baseline to which the other models were compared against.

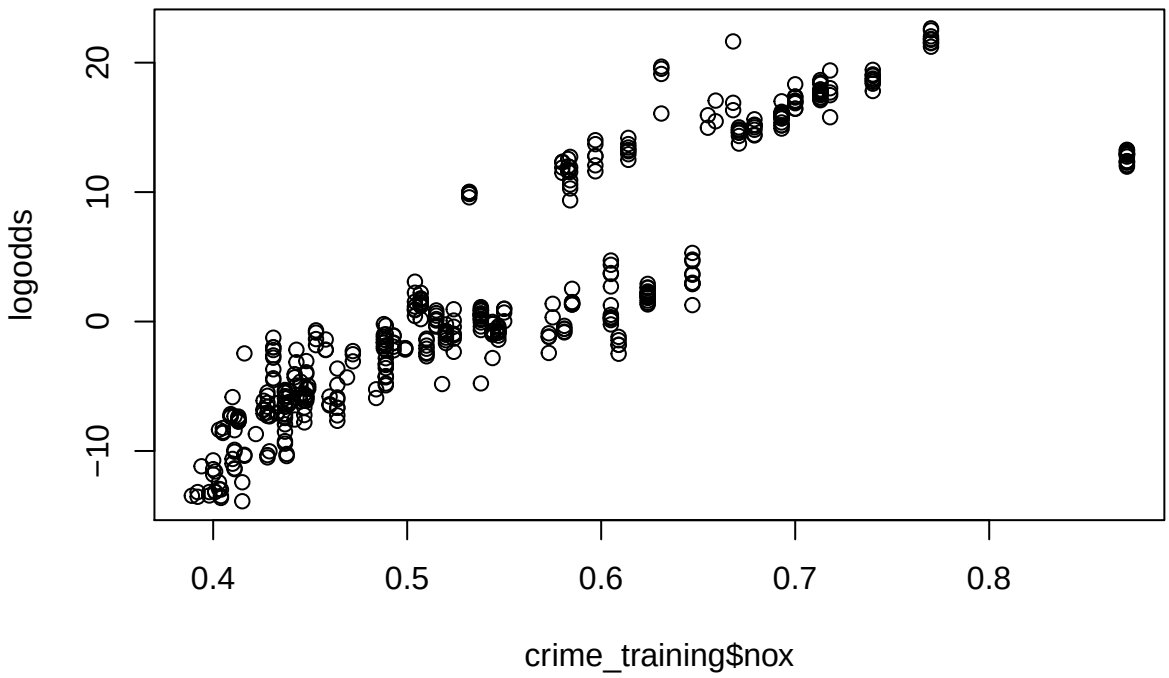
```
Call:
glm(formula = target ~ ., family = binomial(link = "logit"),
    data = crime_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8464  -0.1445  -0.0017   0.0029   3.4665

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -40.822934   6.632913  -6.155 7.53e-10 ***
zn           -0.065946   0.034656  -1.903  0.05706 .
indus        -0.064614   0.047622  -1.357  0.17485
chas1         0.910765   0.755546   1.205  0.22803
nox          49.122297   7.931706   6.193 5.90e-10 ***
```

```
rm           -0.587488   0.722847  -0.813  0.41637
age           0.034189   0.013814   2.475  0.01333 *
dis           0.738660   0.230275   3.208  0.00134 **
rad           0.666366   0.163152   4.084 4.42e-05 ***
tax          -0.006171   0.002955  -2.089  0.03674 *
ptratio       0.402566   0.126627   3.179  0.00148 **
lstat         0.045869   0.054049   0.849  0.39608
medv          0.180824   0.068294   2.648  0.00810 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 645.88  on 465  degrees of freedom
Residual deviance: 192.05  on 453  degrees of freedom
AIC: 218.05

Number of Fisher Scoring iterations: 9
```

`nox` has an extremely low P-value. The U.S. Department of Housing indicates that low-income communities are much more likely than others to experience crime. The National Institute of Environmental Health Sciences indicates that poor communities are exposed to elevated pollution levels, which probably explains why there `nox` is statistically significant. Note that the `lstat`, `indus`, `rm` variables have a p-value greater than 0.05. We reasoned that if the skewed variables were transformed to a normal distribution, than the p-values could decrease, but the p-values actually increased further, thus negating the need to transform the variables.

One of the assumptions for a logistic regression model is linearity of independent variables and log-odds, which explains that the relationship between the logit of the outcome and each continuous independent variable is linear. Using the simple model, we can determine if each of the continuous independent variables meets this assumption. For any independent variables that do not meet this assumption, transformations were performed.

```
 MLE of lambda Score Statistic (z)  Pr(>|z|)
      -0.51585             -9.0634 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

iterations =  6
```



Cook's distance was used to identify observations within the dataset that negatively affect the regression model. Figure XX reveals that for the simple model, none of the observations exceed the dashed red lines, indicating that there are no influential datapoints within the dataset when using the simple model.

Examining Feature Multicollinearity

Finally, it is imperative to understand which features are correlated with each other in order to address and avoid multicollinearity within our models. By using a correlation plot, we can visualize the relationships between certain features. The correlation plot is only able to determine the correlation for continuous variables. There are methodologies to determine correlations for categorical variables (tetrachoric correlation). However there is only one binary predictor variable which is why the multicollinearity will only be considered for the continuous variables.

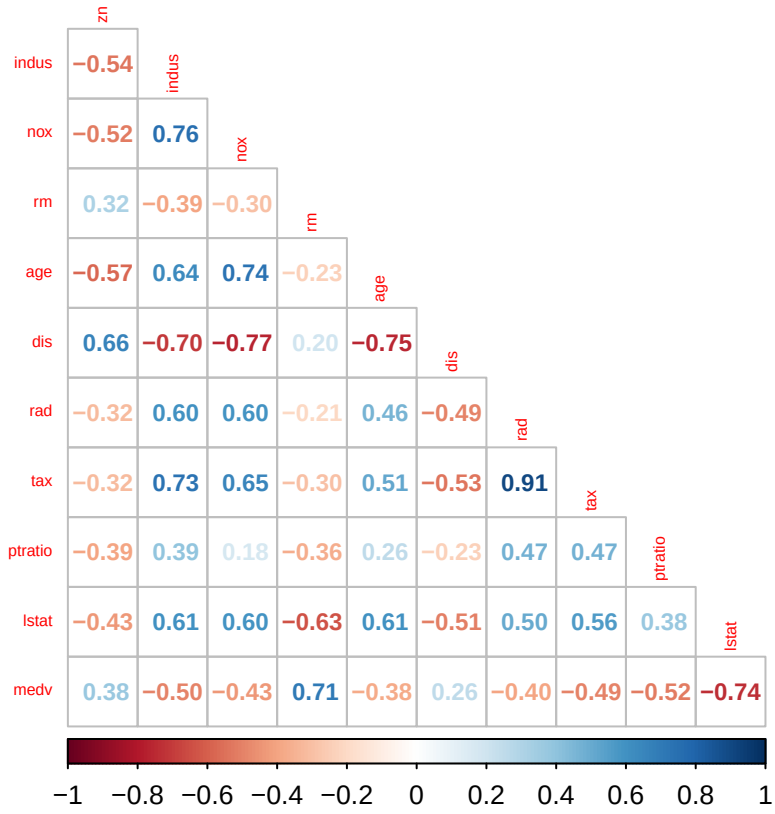


Figure xx: Multicollinearity plot for continuous predictor variables

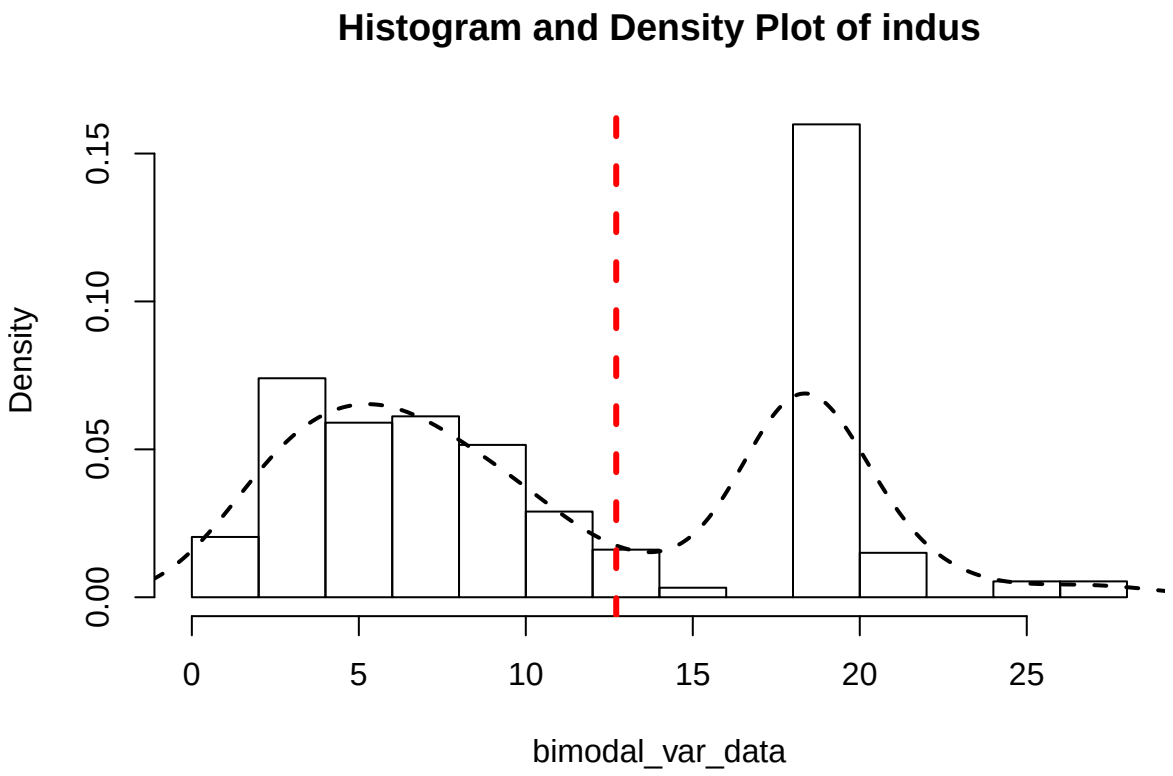
Figure XX reveals that `rad` and `tax` have an extremely high correlation of 0.91. What this indicates that there is a significant correlation between access to radial highways and property taxes. Therefore, the `tax` variable is removed from the dataset to mitigate this high degree of correlation.

Dealing with Bimodal Variables

Bimodal distributions in data are interesting, in that they represent features which actually contain multiple (2) inherent systems resulting in separated distributional peaks. Our approach to solving this is to create dummy variables representing which side of the local minimum each datapoint falls with respect to it's original bimodal distribution. Three new dummy variables were created for the three bimodal variables (`bi_indus`, `bi_tax`, and `bi_rad`). The algorithm that was written to determine the local minimum was able to determine the local minimum for `indus` and `tax` to be 12.70692 and 624.0793 respectively. However, for the `tax` variable, the density plot shows that the local minimum sits somewhere at 550, which is the value that will be used as the cutoff point. The algorithm was unable to detect a local minimum for `rad`. There is probably not enough information for the right peak for the algorithm to work properly. Nevetheless, we determined that a cutoff value of 15 for this variable would suffice. To summarize:

- `bi_indus`: 1 if `indus` is greater than 12.70692, 0 otherwise.
- `bi_tax`: 1 if `tax` is greater than 550, 0 otherwise.
- `bi_rad`: 1 if `rad` is greater than 15, 0 otherwise.

```
number of iterations= 21
[1] 12.7069
```



Works Cited

[1] What is Cook's Distance? (StatisticsHowTo): https://www.statisticshowto.com/cooks-distance/

[2] How to Calculate Correlation Between Categorical Variables (Statology): https://www.statology.org/correlation-between-categorical-variables/

[3] The 6 Assumptions of Logistic Regression (Statology): https://www.statology.org/assumptions-of-logistic-regression/

[4] Neighborhoods and Violent Crime (U.S. Department of Housing) https://www.huduser.gov/portal/periodicals/em/summer16/highlight2.html

[5] Poor Communities Exposed to Elevated Air Pollution Levels https://www.niehs.nih.gov/research/programs/geh/geh_newsletter/2016/4/spotlight/poor_communities_exposed_to_elevated_air_pollution_levels.cfm

[6] https://github.com/kennethleungty/Logistic-Regression-Assumptions/blob/main/Box-Tidwell-Test-in-R.ipynb