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.

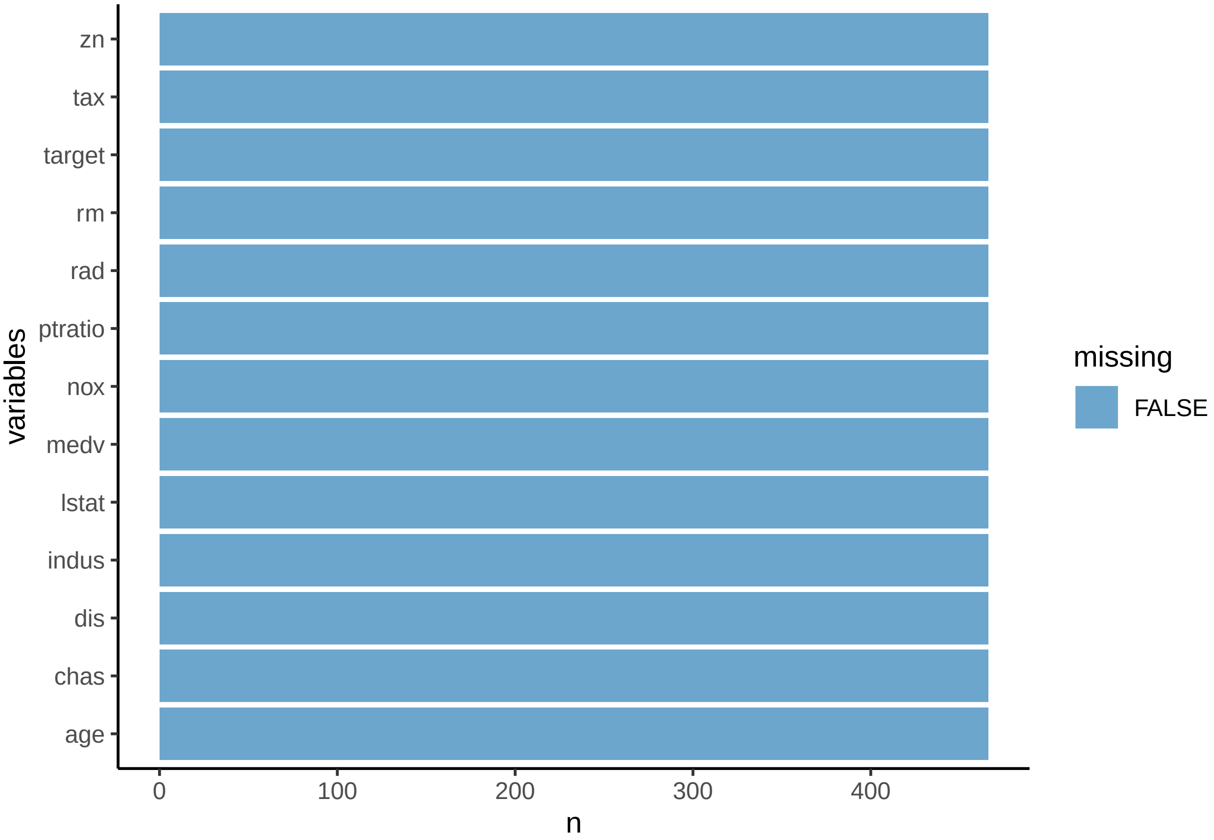


Figure XX: Chart showing the count of missing values for each of the variables in the dataset. Note that since there are no missing values, the legend only shows one item.

Outliers

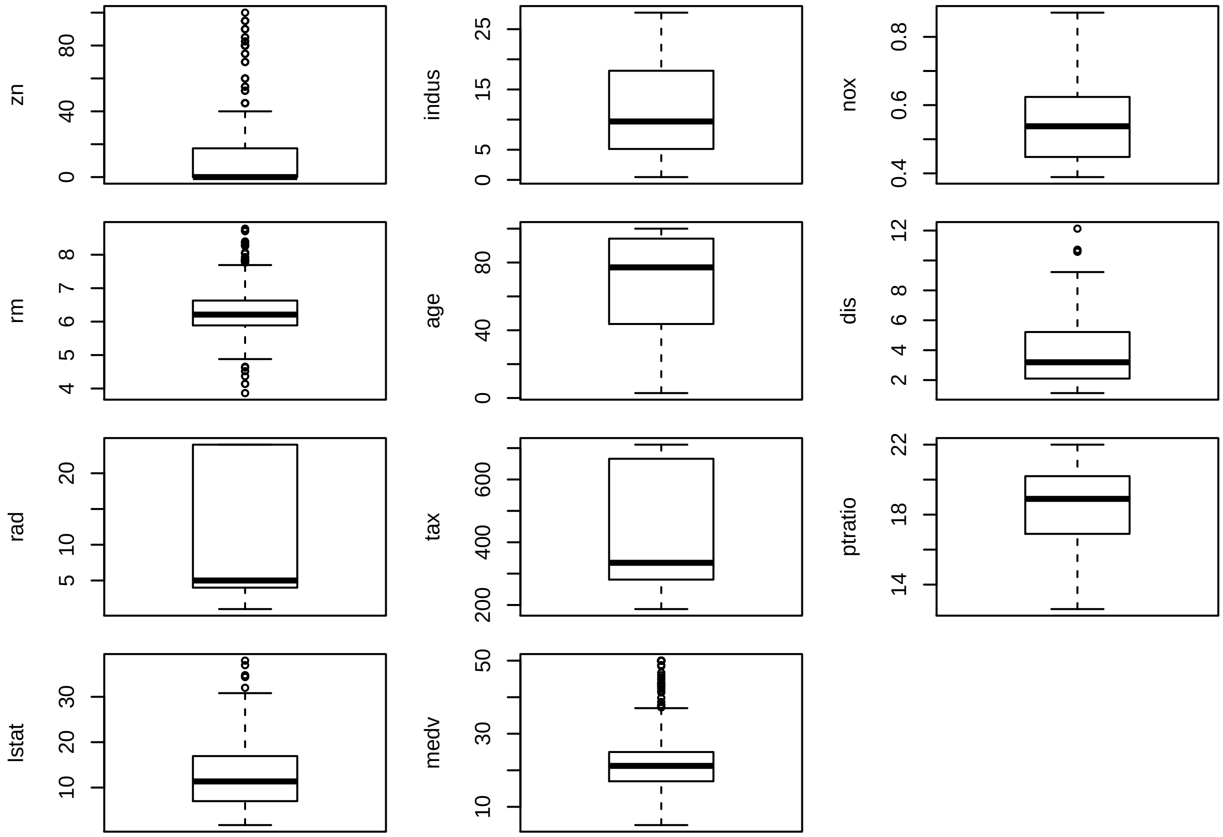


Figure XX: Box plots for each of the variables in the dataset.

Note that the boxplots shown in Figure xx show that `rad` and `tax` have significantly large interquartile ranges which indicates skewness. The density plots for these variables are provided in the "Transformations" section of this report.

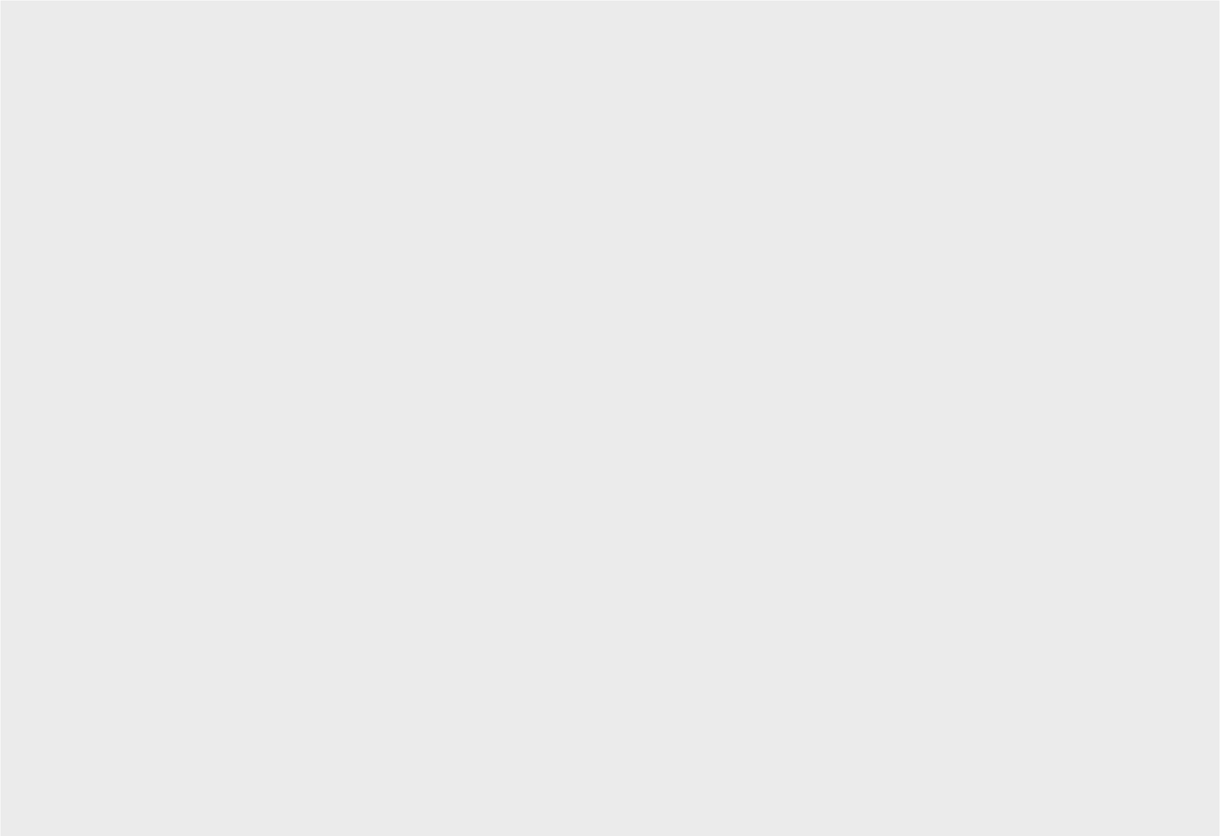
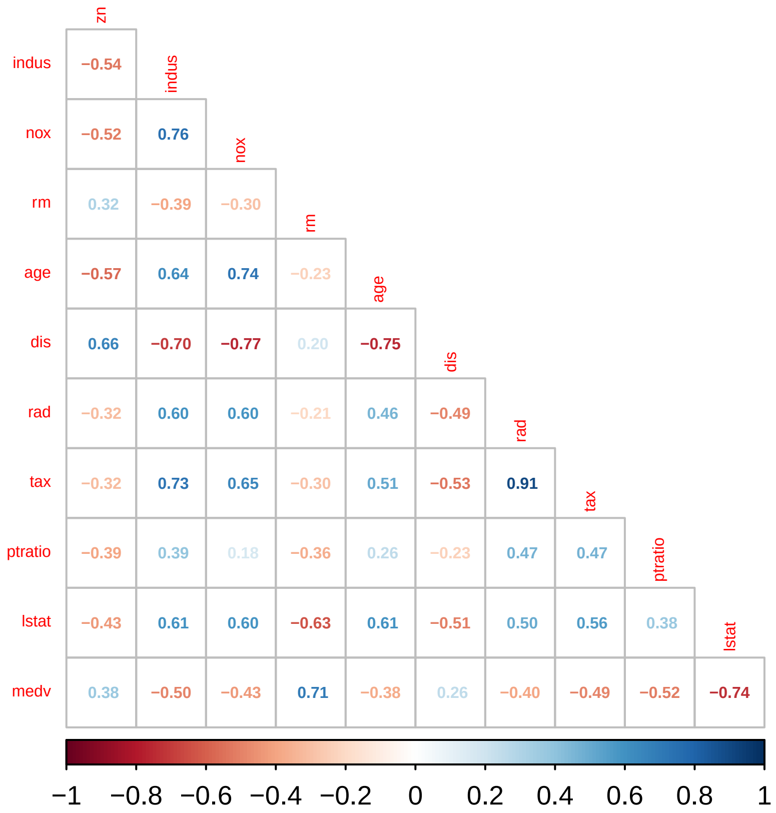


Figure XX shows boxplots for the continuous variables. While `zn`, `rm`, `dis`, `lstat` and `medv` contain outliers, the outliers in general do not seem to be any significant enough to affect the model greatly. Cook's distance can only be used after a regression model has been fit to the data. Therefore, outlier analysis was conducted after the models were generated to identify points that negatively affect the regression 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.



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

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. Property taxes

Transformations for normality

Simple Model

Fit all the independent variables here and fit the dependent variable as well.

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/