Data 621 - Homework 3

Group 4 Layla Quinones, Ian Costello, Dmitriy Burtsev & Esteban Aramayo

2021-10-24

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). Your objective is to 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). 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 radial highways (predictor variable)
- tax: full-value property-tax rate per $10,000 (predictor variable)
- ptratio: pupil-teacher ratio by town (predictor variable)
- black: 1000(Bk - 0.63)2 where Bk 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)

1. Data Exploration

Initial data inspection

Let's take a glance at the training data.

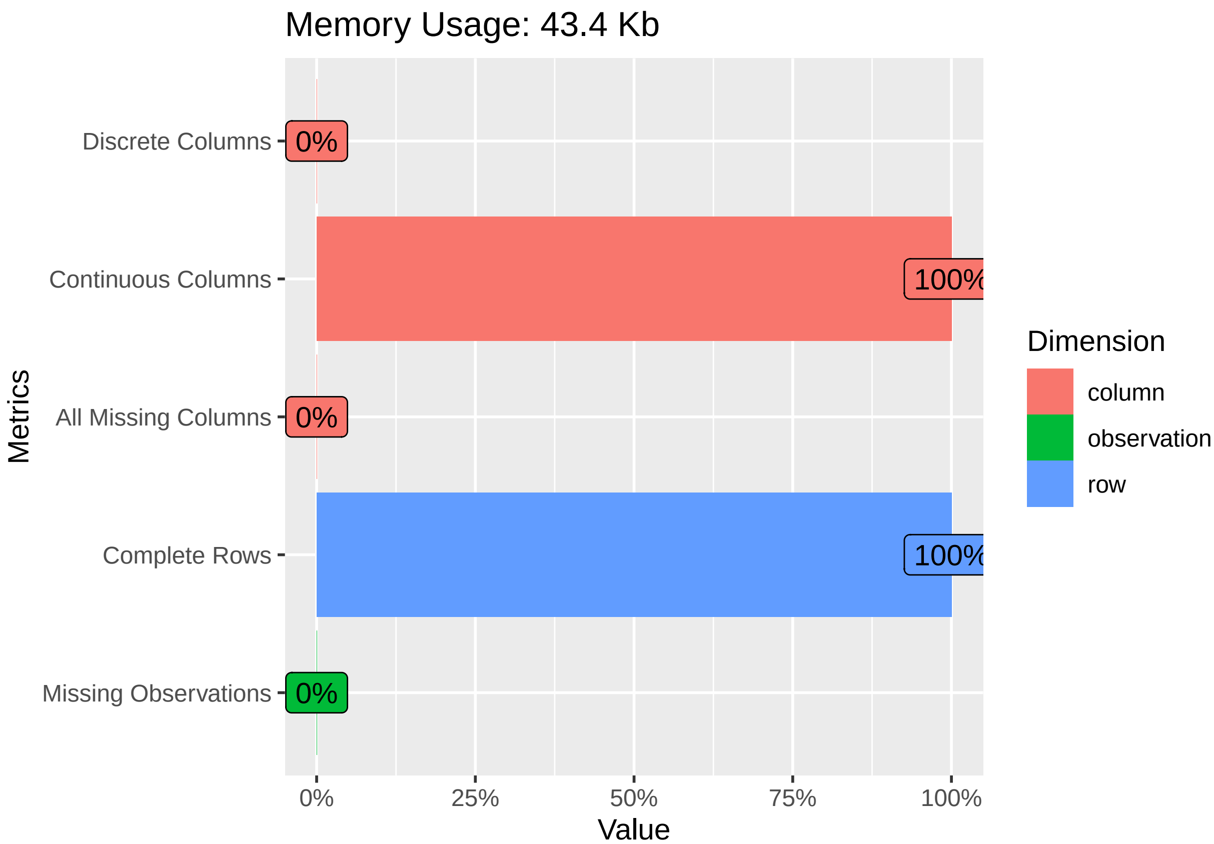
zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv	target
0	19.58	0	0.605	7.929	96.2	2.0459	5	403	14.7	3.70	50.0	1
0	19.58	1	0.871	5.403	100.0	1.3216	5	403	14.7	26.82	13.4	1
0	18.10	0	0.740	6.485	100.0	1.9784	24	666	20.2	18.85	15.4	1
30	4.93	0	0.428	6.393	7.8	7.0355	6	300	16.6	5.19	23.7	0
0	2.46	0	0.488	7.155	92.2	2.7006	3	193	17.8	4.82	37.9	0
0	8.56	0	0.520	6.781	71.3	2.8561	5	384	20.9	7.67	26.5	0
0	18.10	0	0.693	5.453	100.0	1.4896	24	666	20.2	30.59	5.0	1
0	18.10	0	0.693	4.519	100.0	1.6582	24	666	20.2	36.98	7.0	1
0	5.19	0	0.515	6.316	38.1	6.4584	5	224	20.2	5.68	22.2	0
80	3.64	0	0.392	5.876	19.1	9.2203	1	315	16.4	9.25	20.9	0

Metrics on training data set

To get acquainted with the training data set, let's get some metrics on it.

Metric	Count
rows	466
columns	13
discrete_columns	0
continuous_columns	13
all_missing_columns	0
total_missing_values	0
complete_rows	466
total_observations	6058
memory_usage	44440

Let's visualize the observed metrics on the training data set.



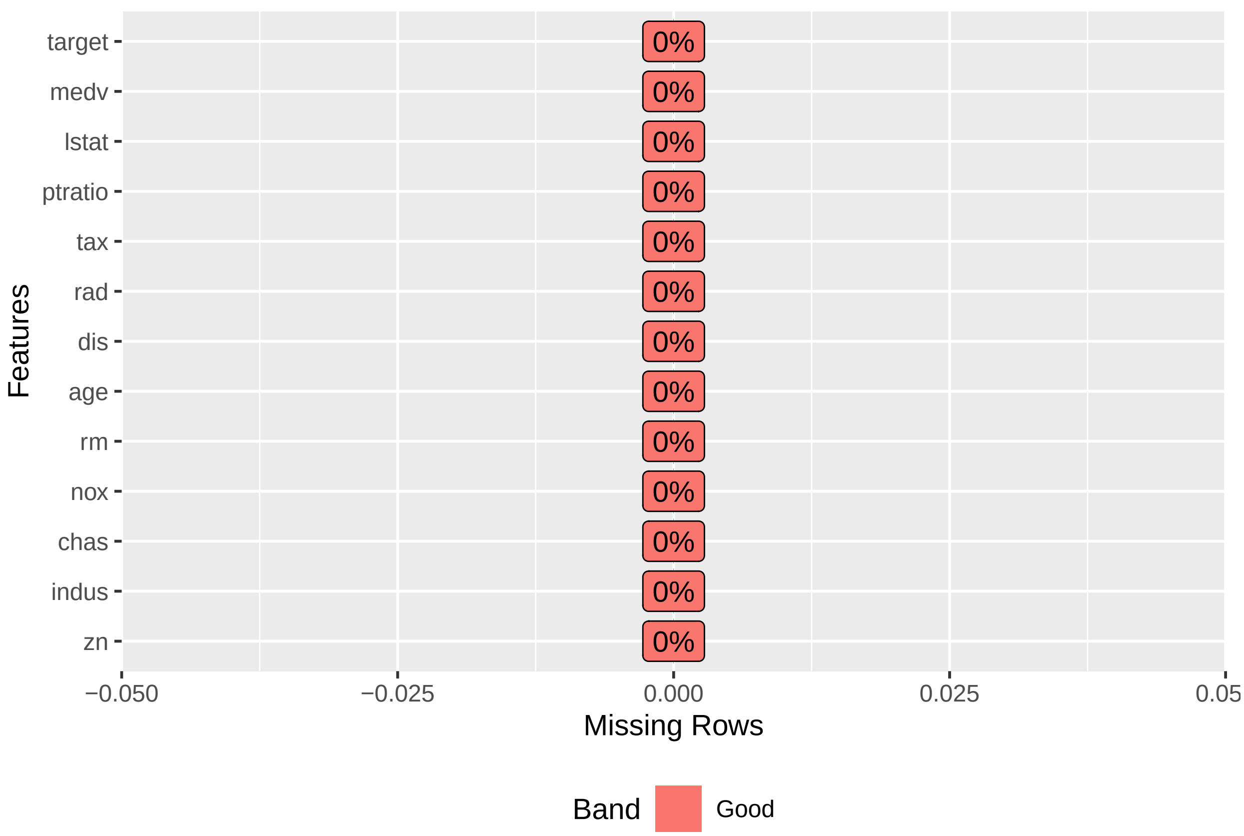
- We can see that most of the variables appear to be continuous. But, from the description of the predictors in the overview section of this document, we know that some of them can be treated as discrete and/or categorical. We will know more later when we test for value uniqueness.
- No columns with missing values were detected.
- All rows are complete.

Summary statistics per variable

Below are the summary statistics for all variables in the training 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         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  
```

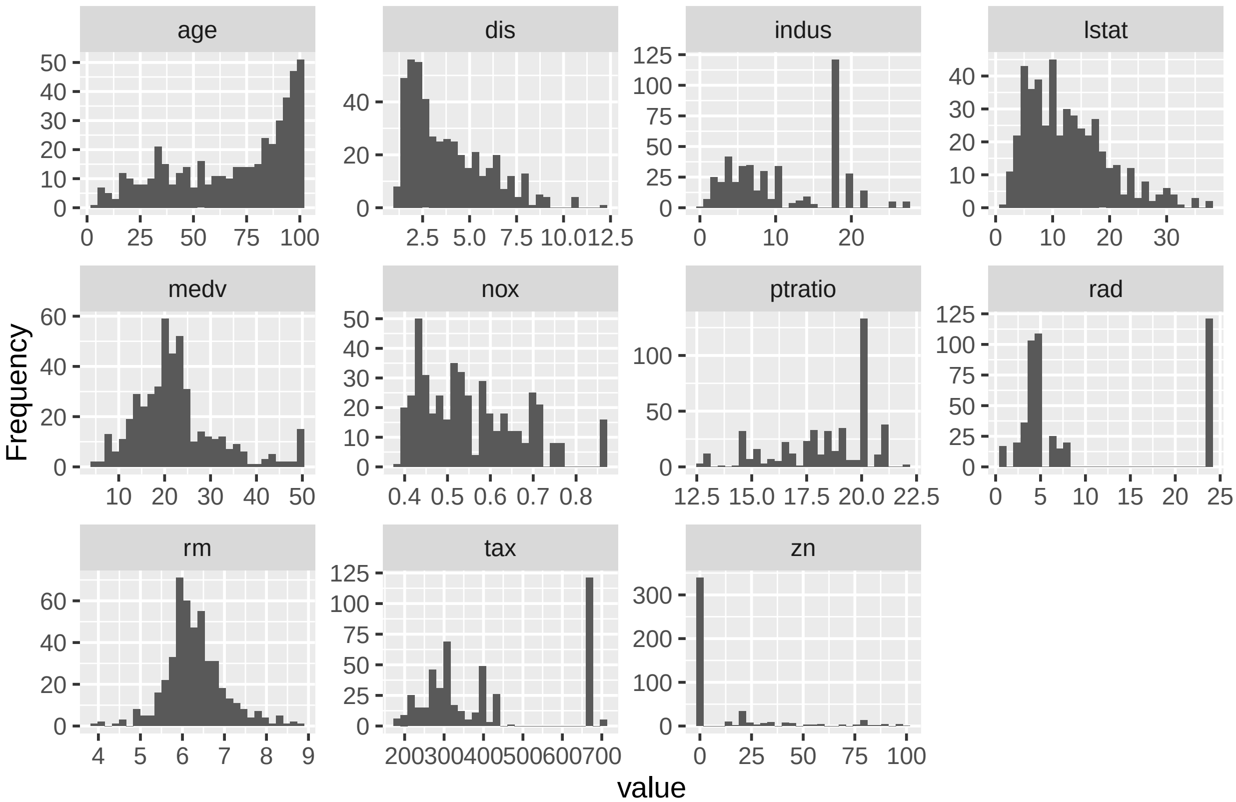
Missing values



From the chart we do not see any variable with missing values.

Histograms

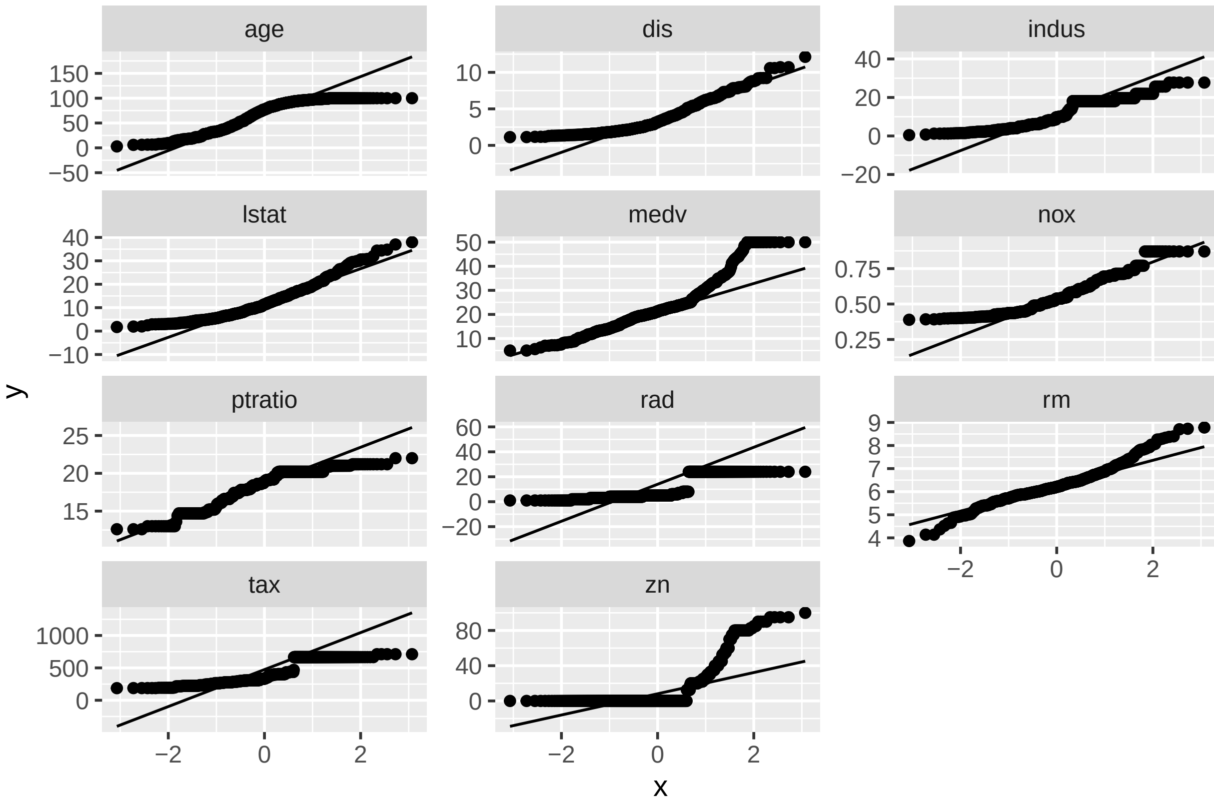
Let's visualize distributions for all continuous features:



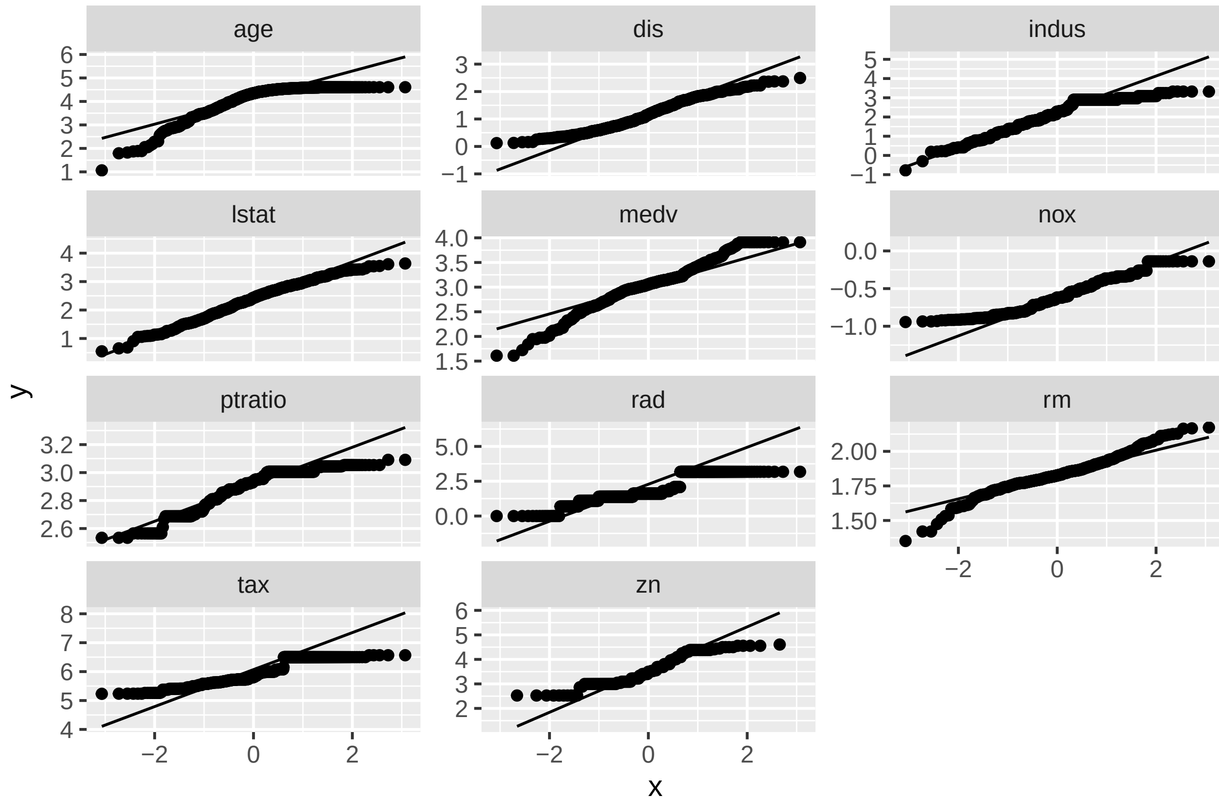
- None of the predictor variables seem to be nearly normal with exception of perhaps "rm".
- Multiple predictors appear to be skewed such as "age", "dis", "lstat", "ptratio". It will be necessary to apply transformations to these.
- Outliers can be seen for predictors "dis", "indus", "lstat", "nox", "ptratio", "rad", "rm", "tax", and "zn". Later, we will verify this using box plots.

QQ Plots

- Let's use Quantile-Quantile plots to visualize the deviation of the predictors compared to the normal distribution.



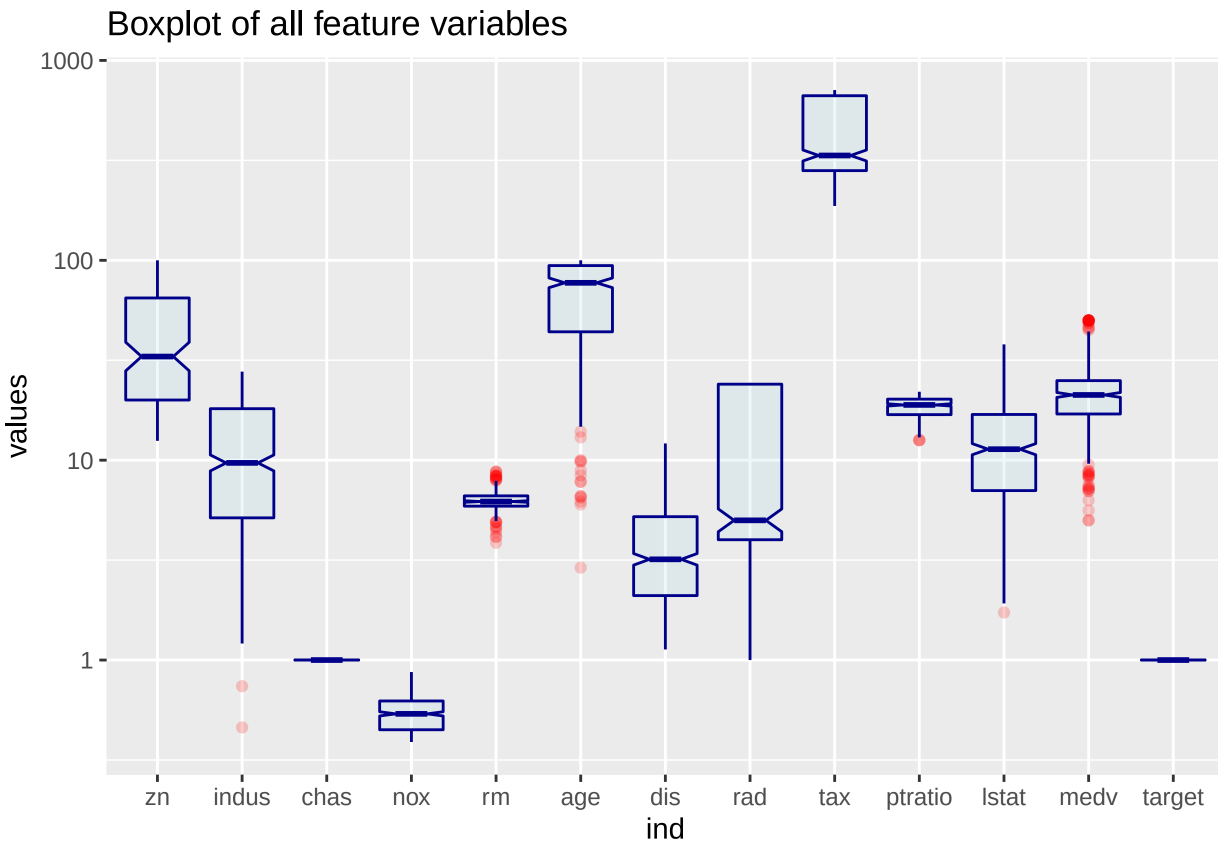
- It appears that, with exception of the “chas” predictor, all other predictors will need to be transformed for linear regression.
- Let’s apply a simple log transformation and plot them again to see any difference can be observed.



- The distributions look better now. So, as part of the data preparation we will transform the necessary predictors before we use them for the models.

Boxplot Analysis

- Let's generate box plots for all the feature variables.
- Let's also apply a log re-scaling to better compare the values across variables using a common scale.
- Let's use notches to compare groups. If the notches of two boxes do not overlap, then this suggests that the medians are significantly different.



We can see obvious outliers for variables “indus”, “rm”, “age”,“ptratio” “lstat”, and “medv”.

Most Common Values for outlier variables

indus	n	Freq
18.10	121	0.2596567
19.58	28	0.0600858
8.14	19	0.0407725
6.20	16	0.0343348
21.89	14	0.0300429

rm	n	Freq
5.713	3	0.0064378
6.127	3	0.0064378
6.167	3	0.0064378
6.229	3	0.0064378
6.405	3	0.0064378
6.417	3	0.0064378

age	n	Freq
100.0	42	0.0901288
95.4	4	0.0085837
96.0	4	0.0085837
97.9	4	0.0085837
98.2	4	0.0085837
98.8	4	0.0085837

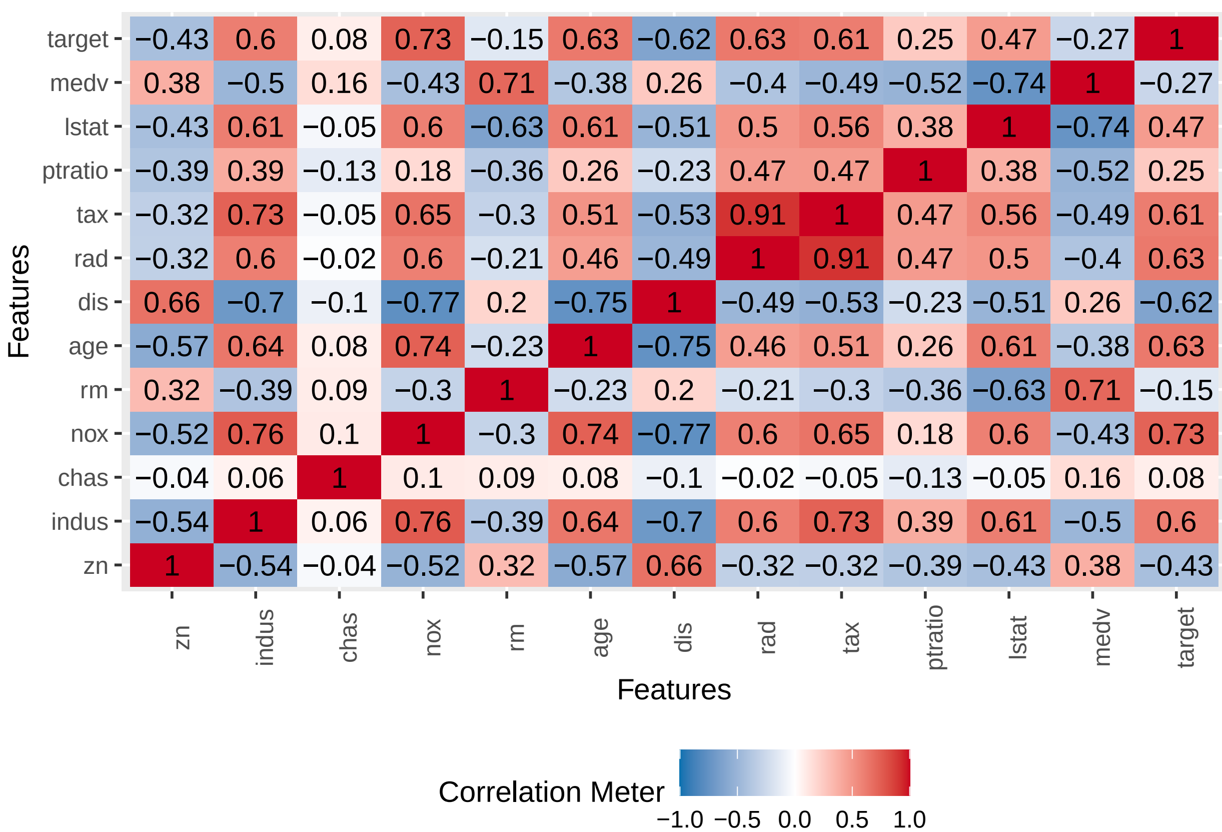
ptratio	n	Freq
20.2	128	0.2746781
14.7	32	0.0686695
21.0	23	0.0493562
17.8	22	0.0472103
19.2	17	0.0364807

lstat	n	Freq
6.36	3	0.0064378
7.79	3	0.0064378
8.05	3	0.0064378

medv	n	Freq
50.0	15	0.0321888
22.0	7	0.0150215
23.1	7	0.0150215
19.4	6	0.0128755
20.6	6	0.0128755
21.7	6	0.0128755
25.0	6	0.0128755

Correlation Analysis

Let's use a heatmap to visualize correlation for all features:



- We see significant correlation between the variables below:

Var1	Var2	Correlation
rad	tax	0.91
indus	nox	0.76
nox	age	0.74
indus	tax	0.73
nox	target	0.73*
rm	medv	0.71
age	target	0.63*
rad	target	0.03*
tax	target	0.61*

2. Data Preparation

3. Build Models

4. Select Models