Data 621 - Homework 3

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

2021-10-31

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.

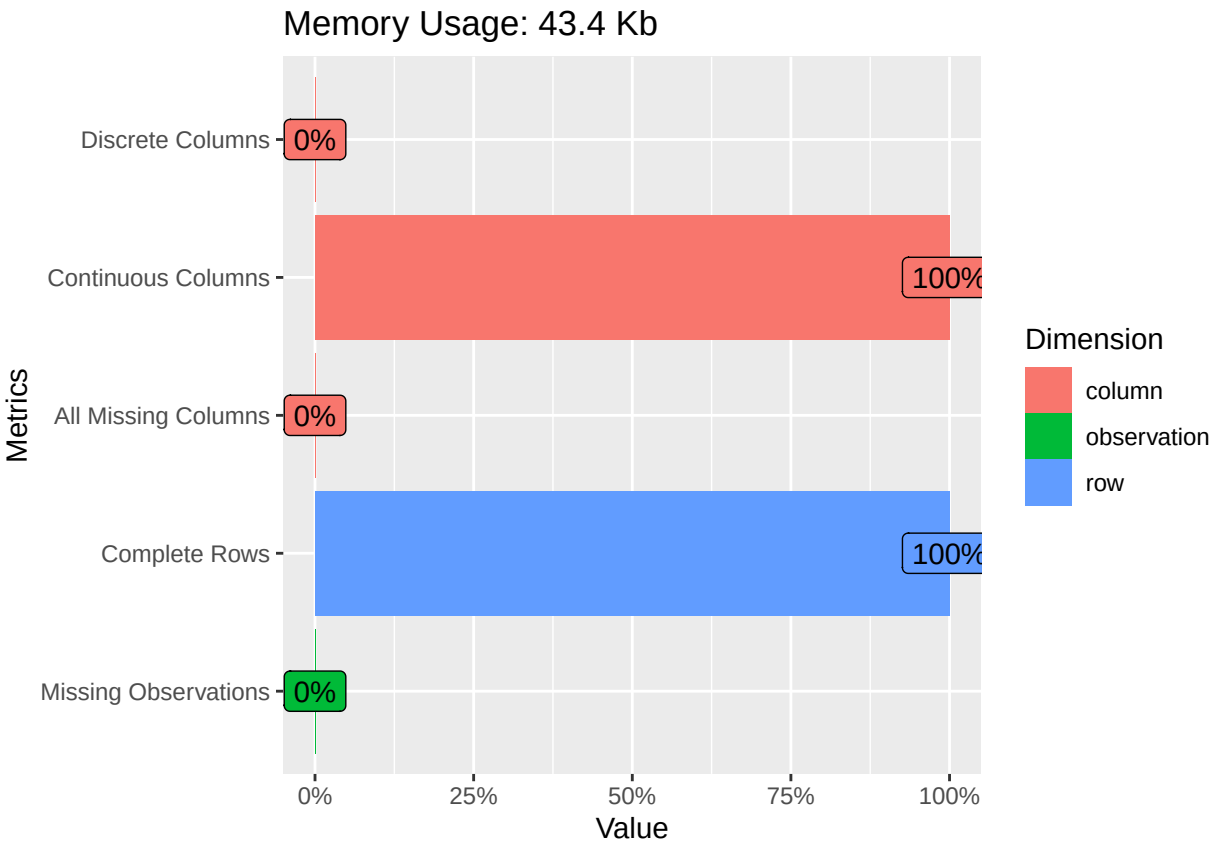
zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv	target
0	19.58	0	0.605	7.929	96.2	2.0459	5	403	14.7	3.70	50.0	1
0	19.58	1	0.871	5.403	100.0	1.3216	5	403	14.7	26.82	13.4	1
0	18.10	0	0.740	6.485	100.0	1.9784	24	666	20.2	18.85	15.4	1
30	4.93	0	0.428	6.393	7.8	7.0355	6	300	16.6	5.19	23.7	0
0	2.46	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.



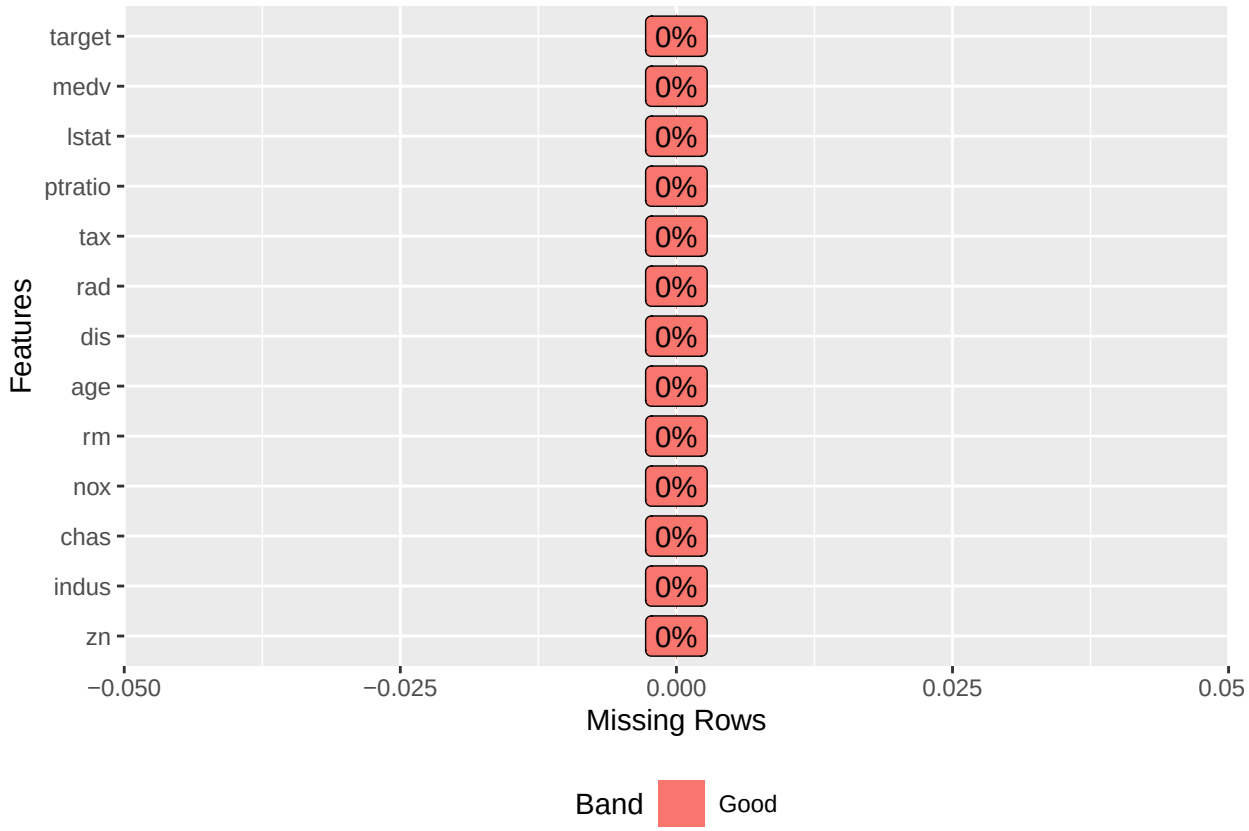
- We can see that most of the variables appear to be continuous. But, from the description of the predictors in the overview section of this document, we know that some of them can be treated as discrete and/or categorical. We will know more later when we test for value uniqueness.
- No columns with 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              rm
##  Min.   :  0   Min.   : 0.46   Min.   :0.000   Min.   :0.39   Min.   :3.9
##  1st Qu.:  0   1st Qu.: 5.14   1st Qu.:0.000   1st Qu.:0.45   1st Qu.:5.9
##  Median :  0   Median : 9.69   Median :0.000   Median :0.54   Median :6.2
##  Mean   : 12   Mean   :11.11   Mean   :0.071   Mean   :0.55   Mean   :6.3
##  3rd Qu.: 16   3rd Qu.:18.10   3rd Qu.:0.000   3rd Qu.:0.62   3rd Qu.:6.6
##  Max.   :100   Max.   :27.74   Max.   :1.000   Max.   :0.87   Max.   :8.8
##       age             dis              rad             tax         ptratio
##  Min.   :  2.9   Min.   : 1.1   Min.   : 1.0   Min.   :187   Min.   :13
##  1st Qu.: 43.9   1st Qu.: 2.1   1st Qu.: 4.0   1st Qu.:281   1st Qu.:17
##  Median : 77.2   Median : 3.2   Median : 5.0   Median :334   Median :19
##  Mean   : 68.4   Mean   : 3.8   Mean   : 9.5   Mean   :410   Mean   :18
##  3rd Qu.: 94.1   3rd Qu.: 5.2   3rd Qu.:24.0   3rd Qu.:666   3rd Qu.:20
##  Max.   :100.0   Max.   :12.1   Max.   :24.0   Max.   :711   Max.   :22
##      lstat            medv        target
##  Min.   : 1.7   Min.   : 5   Min.   :0.00
##  1st Qu.: 7.0   1st Qu.:17   1st Qu.:0.00
##  Median :11.3   Median :21   Median :0.00
##  Mean   :12.6   Mean   :23   Mean   :0.49
##  3rd Qu.:16.9   3rd Qu.:25   3rd Qu.:1.00
##  Max.   :38.0   Max.   :50   Max.   :1.00
```

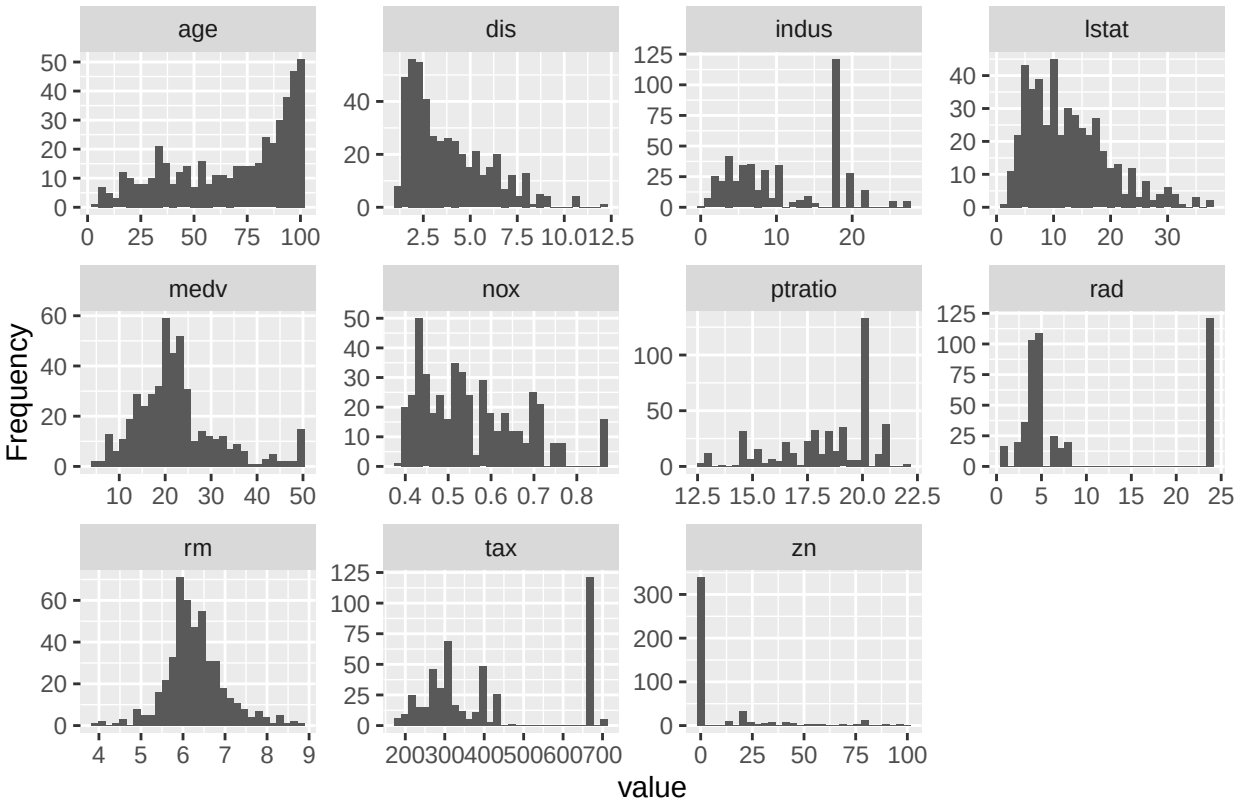
Missing values



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

Histograms

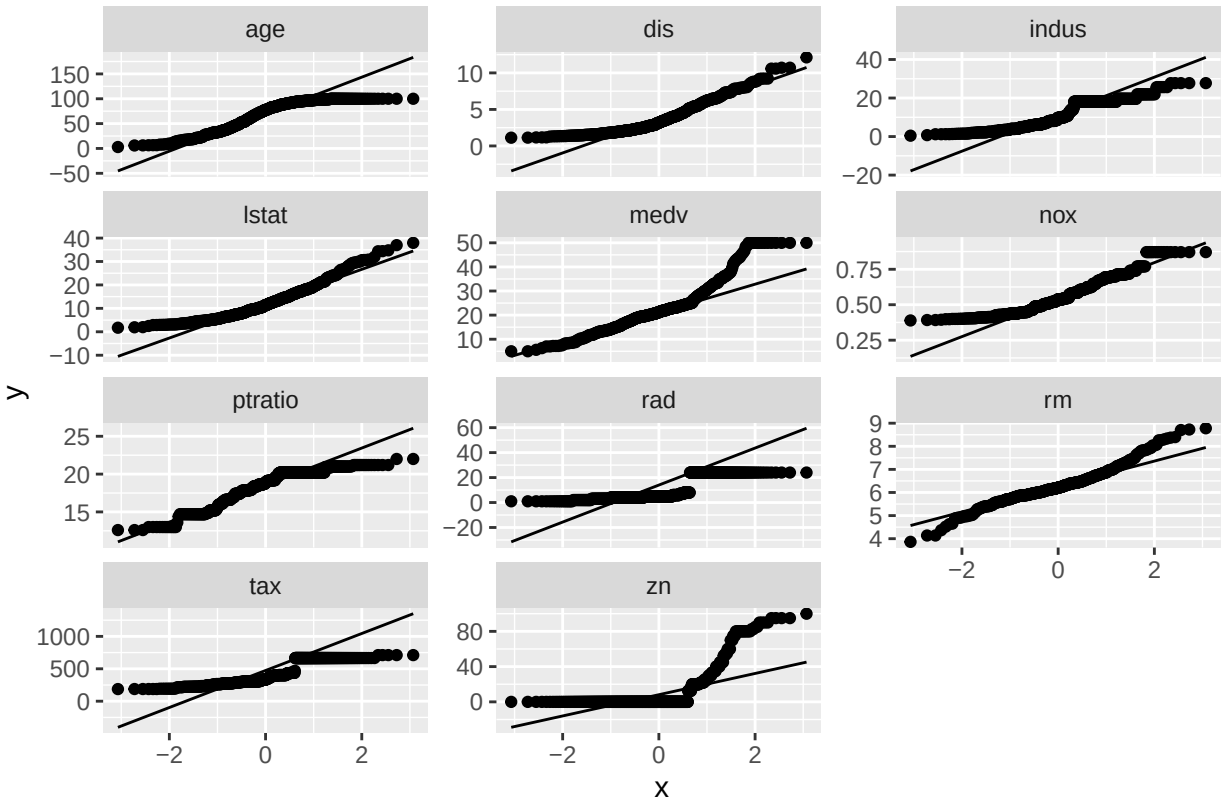
Let's visualize distributions for all continuous features:



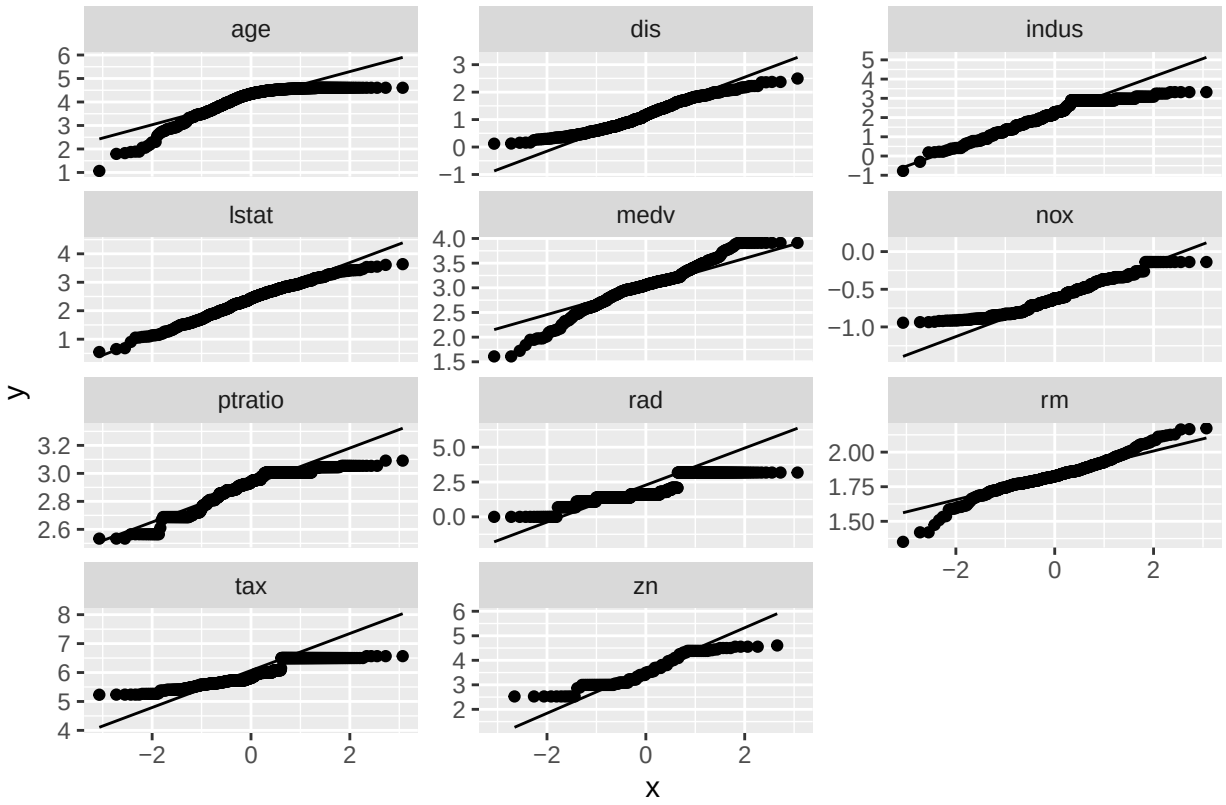
- None of the predictor variables seem to be nearly normal with exception of perhaps `rm`.
- Multiple predictors appear to be skewed such as `age`, `dis`, `lstat`, `ptratio`. It will be necessary to apply transformations to these.
- Possible outliers can be seen for predictors `dis`, `indus`, `lstat`,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

QQ Plots

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



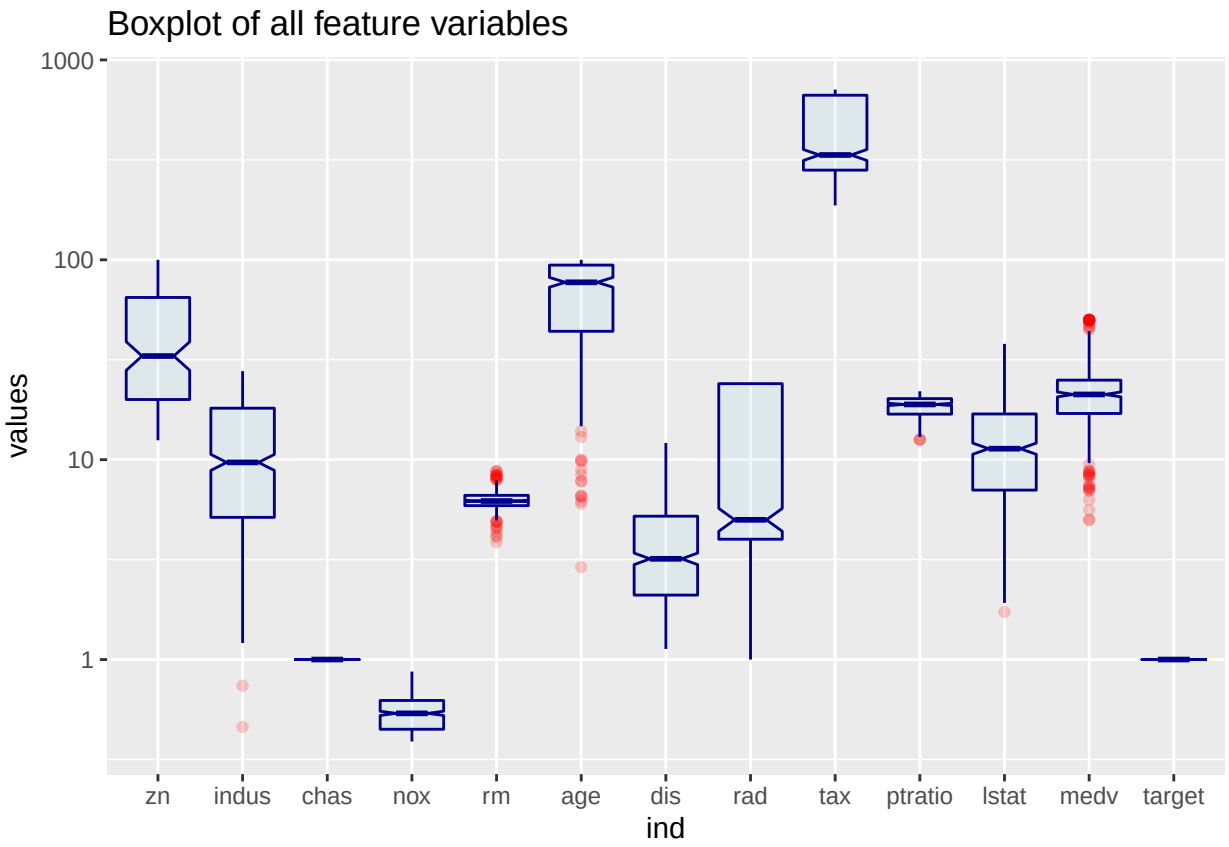
- It appears that, with exception of the "chas" predictor, all other predictors will need to be transformed for linear regression.
- Let's apply a simple log transformation and plot them again to see any difference can be observed.



- 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.



The boxplots confirm that there are obvious outliers for variables `age`, `indus`, `lstat`, `medv`, `ptratio`, and `rm`. These outliers will need to be imputed to prevent them from skewing the results of the modeling.

Unique Value Counts per Variable

Let's count unique values per variable to see which variables might need to be converted to factors if they have small number of unique values.

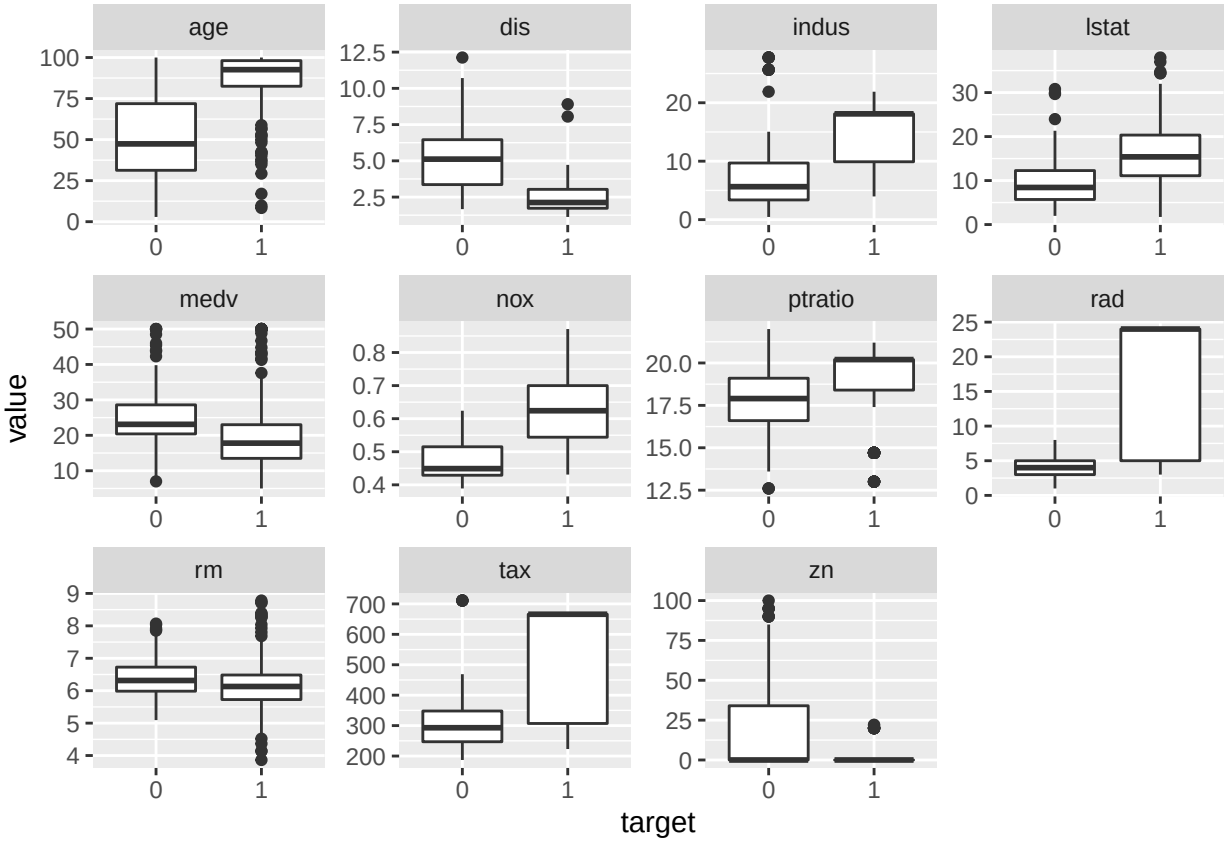
variable	unique.value.count
age	333
chas	2
dis	380
indus	73
lstat	424
medv	218
nox	79
ptratio	46
rad	9
rm	419
target	2
tax	63
zn	26

Convert Categorical variables to factors

From the unique counts above, we can see that the variables `chas` and `target` can be considered as categorical variables due to their low number of unique values. so, we are converting them to factor data type.

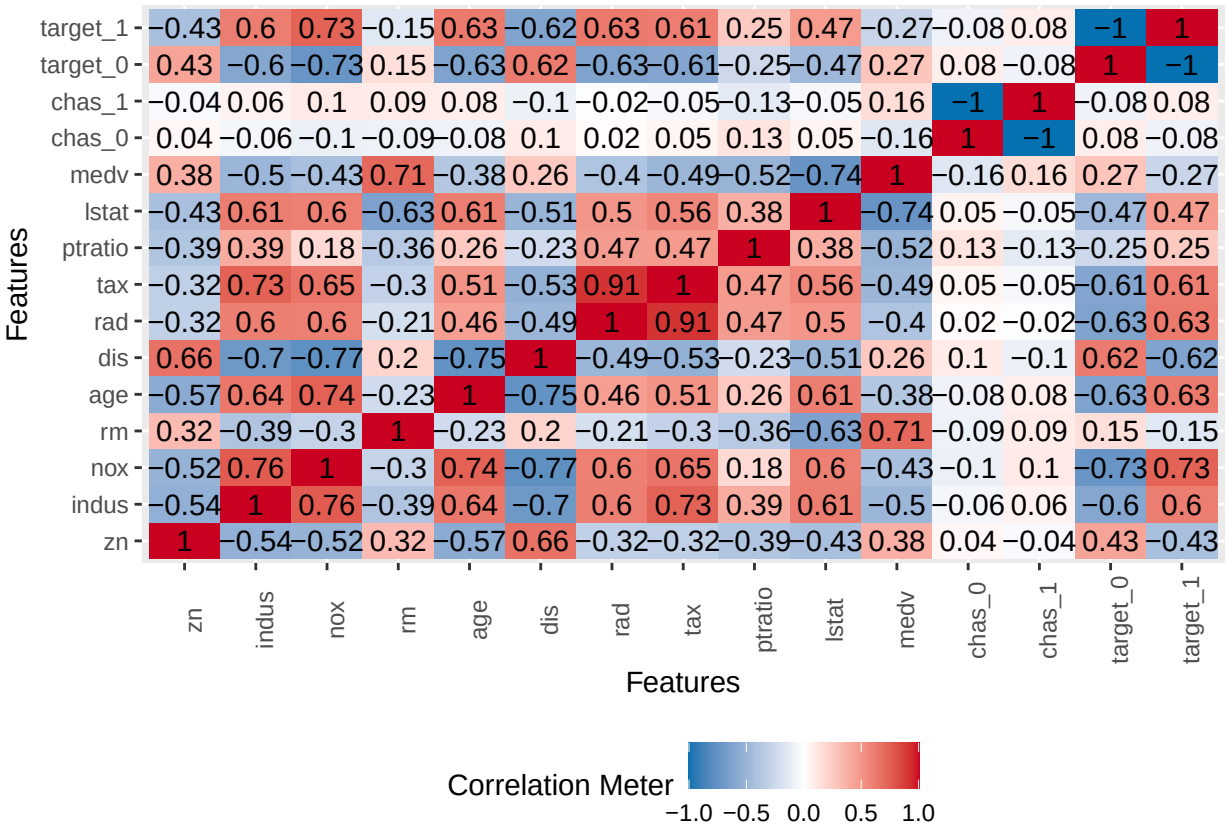
Outlier Value Analysis

Let's analyze the variables with outliers and their relation to the `target` variable.



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*

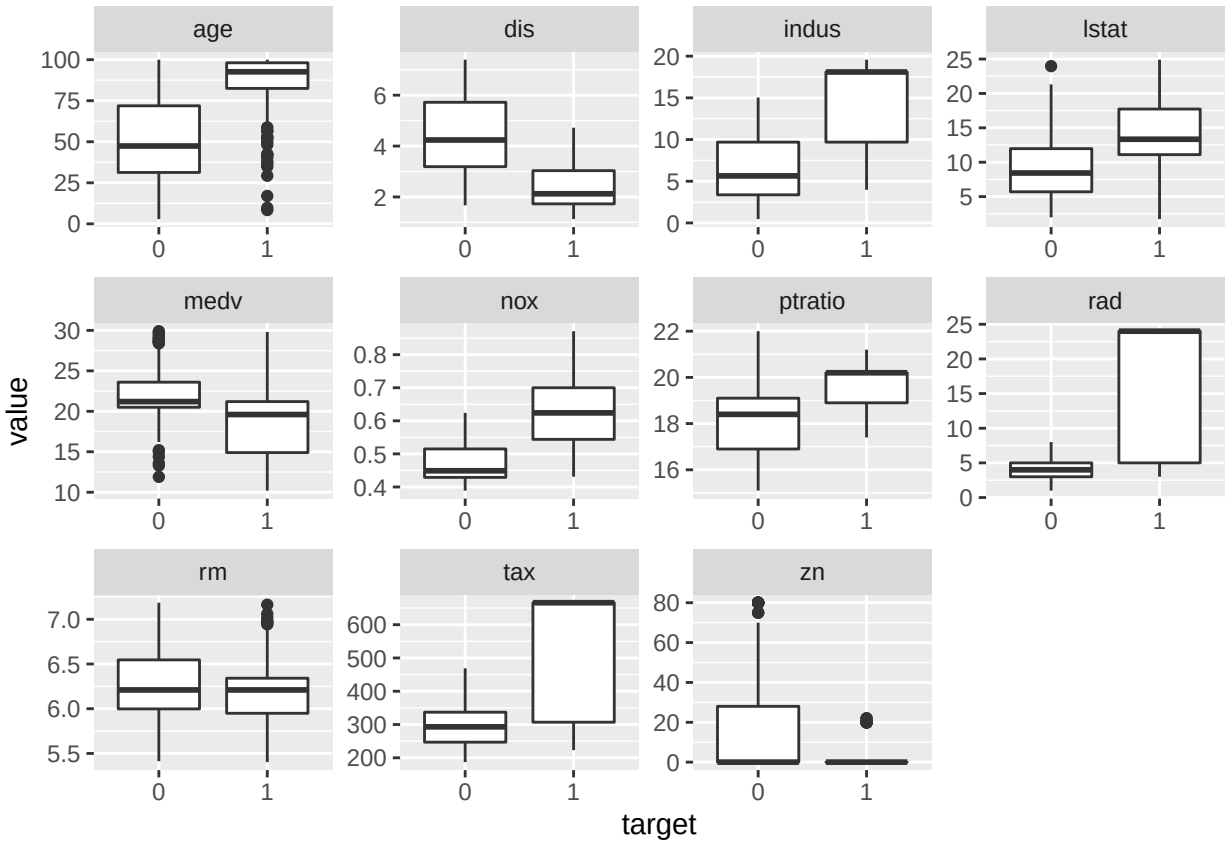
2. Data Preparation

Outlier imputation

The boxplots above confirm that there are obvious outliers for variables `dis`, `indus`, `lstat`, `medv`, `ptratio`, `rm`, `tax`, and `zn`. These outliers need to be imputed to prevent them from skewing the results of the modeling.

Let's impute outliers with median values.

After imputing the variables with outliers, let's analyze them again with respect to their relation to the `target` variable. This is to ensure that outliers have been eliminated or at least minimized as much as possible.



3. Build Models

Using the training data, build at least three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Be sure to explain how you can make inferences from the model, as well as discuss other relevant model output. Discuss the coefficients in the models, do they make sense? Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

Model 1: Generalized Linear model

Let's begin with building a initial model including all the variables.

```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = crime_train_preped)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.90615  -0.21082  -0.00464   0.00049   3.03334
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -25.523464   6.339099  -4.026 5.66e-05 ***
## zn           -0.030047   0.023867  -1.259 0.208061
## indus         0.117949   0.082514   1.429 0.152878
```

```
## chas1         1.295912   0.762003   1.701 0.089006 .
## nox          24.877600   5.103650   4.874 1.09e-06 ***
## rm            0.048545   0.700099   0.069 0.944719
## age           0.016641   0.012067   1.379 0.167888
## dis           0.177180   0.209031   0.848 0.396646
## rad           0.896944   0.178940   5.013 5.37e-07 ***
## tax          -0.008987   0.004248  -2.116 0.034368 *
## ptratio       0.556289   0.157687   3.528 0.000419 ***
## lstat        -0.084323   0.052915  -1.594 0.111037
## medv         -0.120249   0.082825  -1.452 0.146546
## ---
## 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: 203.78  on 453  degrees of freedom
## AIC: 229.78
##
## Number of Fisher Scoring iterations: 9
```

Model 2: Forward Selection model

Let's build a forward step-wise selection based on AIC.

Let's use the variables that we found to have the highest correlation

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*

4. Select Models