DATA 621 - Homework 3

Ian Costello

11/6/2021

Overview

General Objective

For this assignment, we will be exploring, analyzing, and modeling data related to crime statistics for various areas of a major U.S. city. The primary objective is to understand how, or if, variable indicate whether crime in a particular area will be above or below the median crime rate for the entire city. The models will be binary logistic regression using combinations or constructions of the variables provided.

About the Data

- 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)

Libraries Used

We use pretty standard packages for this assignment, including the ever-useful `tidyverse`, `ggplot2`, and `caret`. New additions for this assignment include `VIM`, `DataExplorer`, and `broom`.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## Warning: package 'forcats' was built under R version 4.0.5
```

```
library(ggplot2)
library(VIM)
```

```
## Warning: package 'VIM' was built under R version 4.0.5

## Warning: package 'colorspace' was built under R version 4.0.5
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.5
```

```
library(broom)
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.0.5
```

```
library(tidymodels)
```

```
## Warning: package 'tidymodels' was built under R version 4.0.5

## Warning: package 'dials' was built under R version 4.0.5

## Warning: package 'infer' was built under R version 4.0.5

## Warning: package 'modeldata' was built under R version 4.0.5

## Warning: package 'parsnip' was built under R version 4.0.5

## Warning: package 'recipes' was built under R version 4.0.5

## Warning: package 'rsample' was built under R version 4.0.5

## Warning: package 'tune' was built under R version 4.0.5

## Warning: package 'workflows' was built under R version 4.0.5

## Warning: package 'workflowsets' was built under R version 4.0.5

## Warning: package 'yardstick' was built under R version 4.0.5
```

```
library(DataExplorer)
```

```
## Warning: package 'DataExplorer' was built under R version 4.0.5
```

```
library(psych)
```

Data Exploration

As usual, our data are stored on GitHub at our team's main repository for easy access across team members.

```
# Load data
# Training
rawTrain <- read.csv("https://raw.githubusercontent.com/MsQCompSci/Data621Group4/main/HW3/crime-training

#Testing data
rawTest <- read.csv("https://raw.githubusercontent.com/MsQCompSci/Data621Group4/main/HW3/crime-evaluati
```

Data Structure and Summary Statistics

```
str(rawTrain)
```

```
## '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 ...
```

```
summary(rawTrain, digits = 2)
```

```
##        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
```

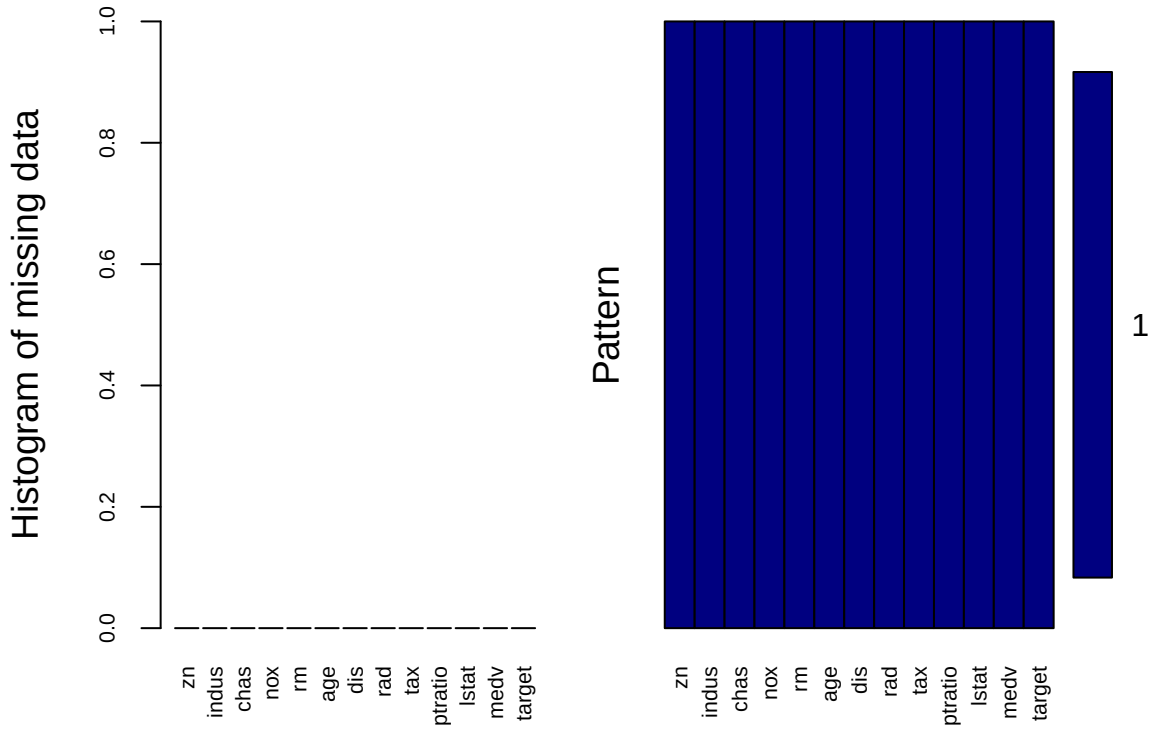
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
zn	1	466	11.5772532	23.3646511	0.00000	5.3542781	0.0000000	0.0000	100.0000	100.0000	2.1768152	3.8135765	1.0823466
indus	2	466	11.1050215	6.8458549	9.69000	10.9082353	9.3403800	0.4600	27.7400	27.2800	0.2885450	-1.2432132	0.3171281
chas	3	466	0.0708155	0.2567920	0.00000	0.0000000	0.0000000	0.0000	1.0000	1.0000	3.3354899	9.1451313	0.0118957
nox	4	466	0.5543105	0.1166667	0.53800	0.5442684	0.1334340	0.3890	0.8710	0.4820	0.7463281	-0.0357736	0.0054045
rm	5	466	6.2906738	0.7048513	6.21000	6.2570615	0.5166861	3.8630	8.7800	4.9170	0.4793202	1.5424378	0.0326516
age	6	466	68.3675966	28.3213784	77.15000	70.9553476	30.0226500	2.9000	100.0000	97.1000	-0.5777075	-1.0098814	1.3119625
dis	7	466	3.7956929	2.1069496	3.19095	3.5443647	1.9144814	1.1296	12.1265	10.9969	0.9988926	0.4719679	0.0976026
rad	8	466	9.5300429	8.6859272	5.00000	8.6978610	1.4826000	1.0000	24.0000	23.0000	1.0102788	-0.8619110	0.4023678
tax	9	466	409.5021459	167.9000887	334.50000	401.5080214	104.5233000	187.0000	711.0000	524.0000	0.6593136	-1.1480456	7.7778214
ptratio	10	466	18.3984979	2.1968447	18.90000	18.5970588	1.9273800	12.6000	22.0000	9.4000	-0.7542681	-0.4003627	0.1017669
lstat	11	466	12.6314592	7.1018907	11.35000	11.8809626	7.0720020	1.7300	37.9700	36.2400	0.9055864	0.5033688	0.3289887
medv	12	466	22.5892704	9.2396814	21.20000	21.6304813	6.0045300	5.0000	50.0000	45.0000	1.0766920	1.3737825	0.4280200
target	13	466	0.4914163	0.5004636	0.00000	0.4893048	0.0000000	0.0000	1.0000	1.0000	0.0342293	-2.0031131	0.0231835

```
##  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
```

```
kable(describe(rawTrain),booktabs = TRUE) %>%
  kable_styling(latex_options = "scale_down")
```

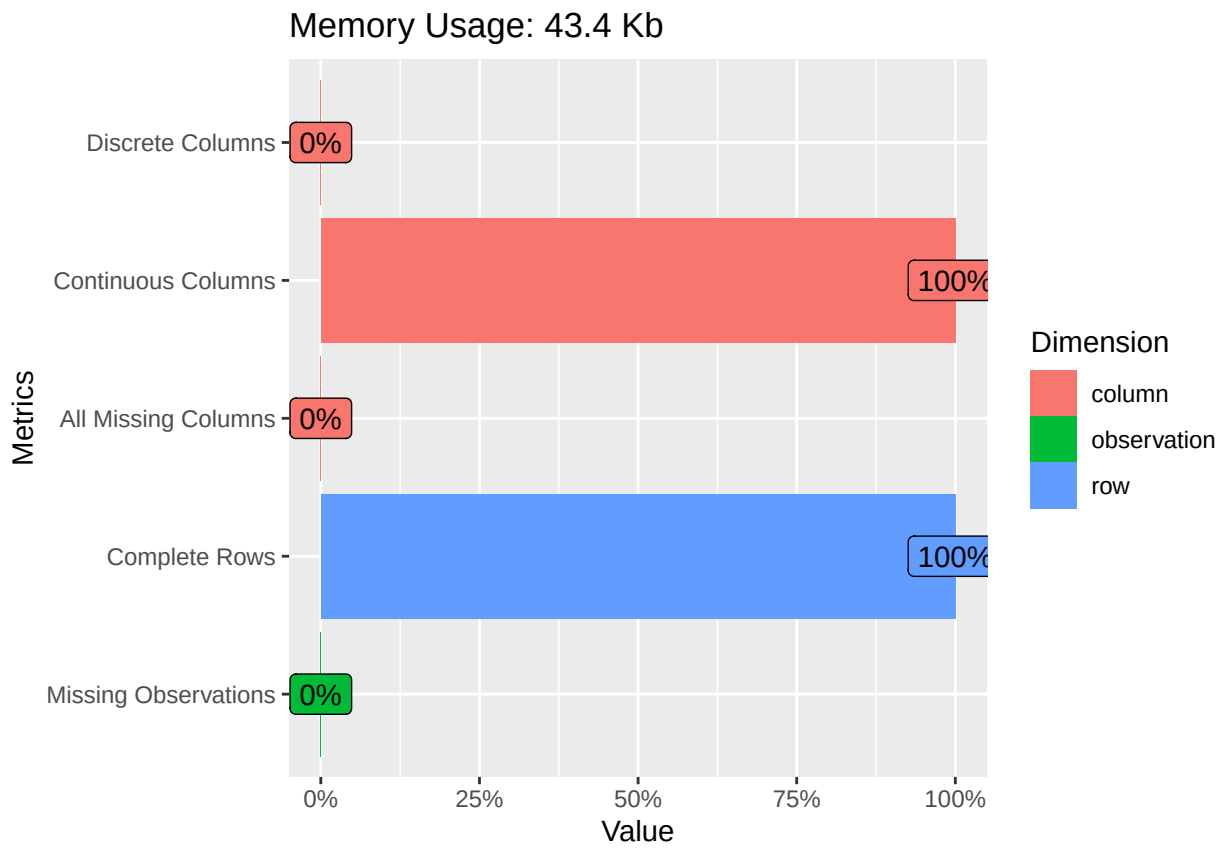
Missing Data Checks

```
#plot missing values using VIM package
aggr(rawTrain , col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(rawTrain), cex.axis=
```




```
##
##  Variables sorted by number of missings:
##  Variable Count
##        zn     0
##     indus     0
##      chas     0
##       nox     0
##        rm     0
##       age     0
##       dis     0
##       rad     0
##       tax     0
##   ptratio     0
##     lstat     0
##      medv     0
##    target     0
```

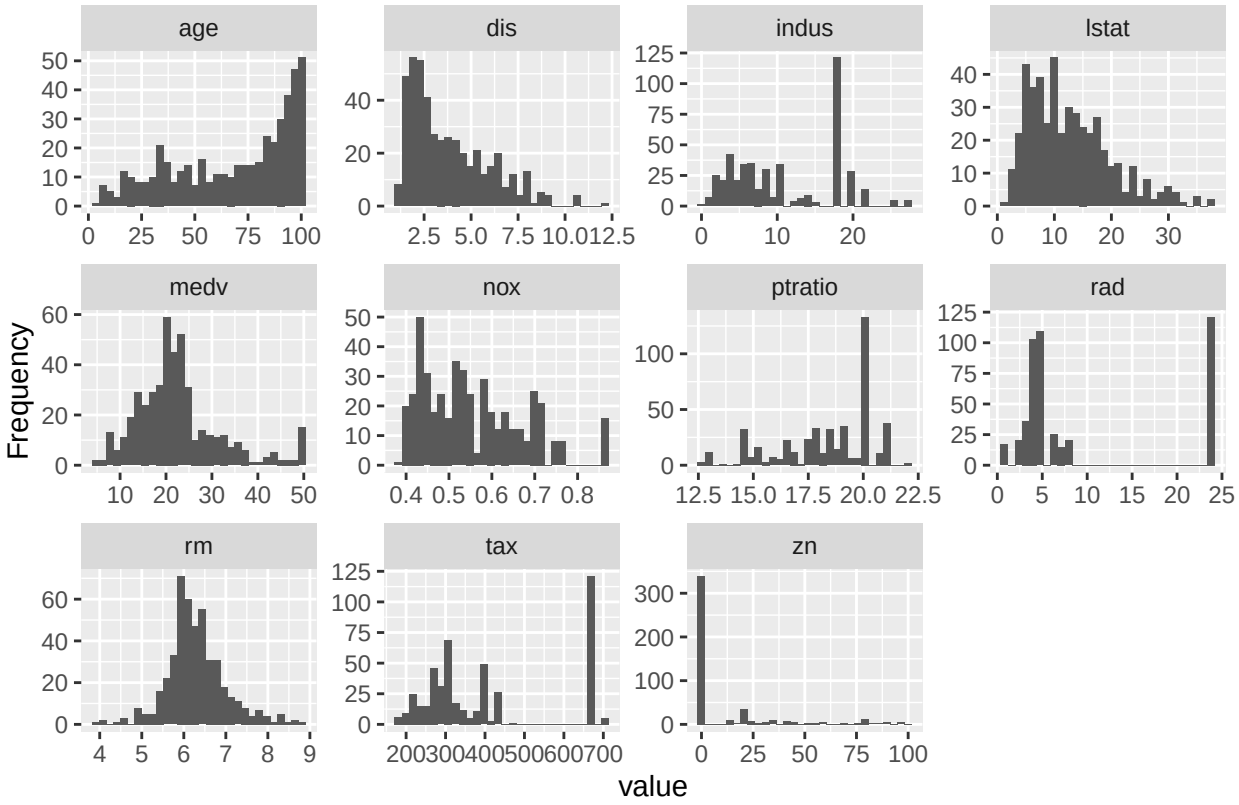
```
DataExplorer::plot_intro(rawTrain)
```



- 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.

Feature Histograms

```
DataExplorer::plot_histogram(rawTrain)
```



- 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`.

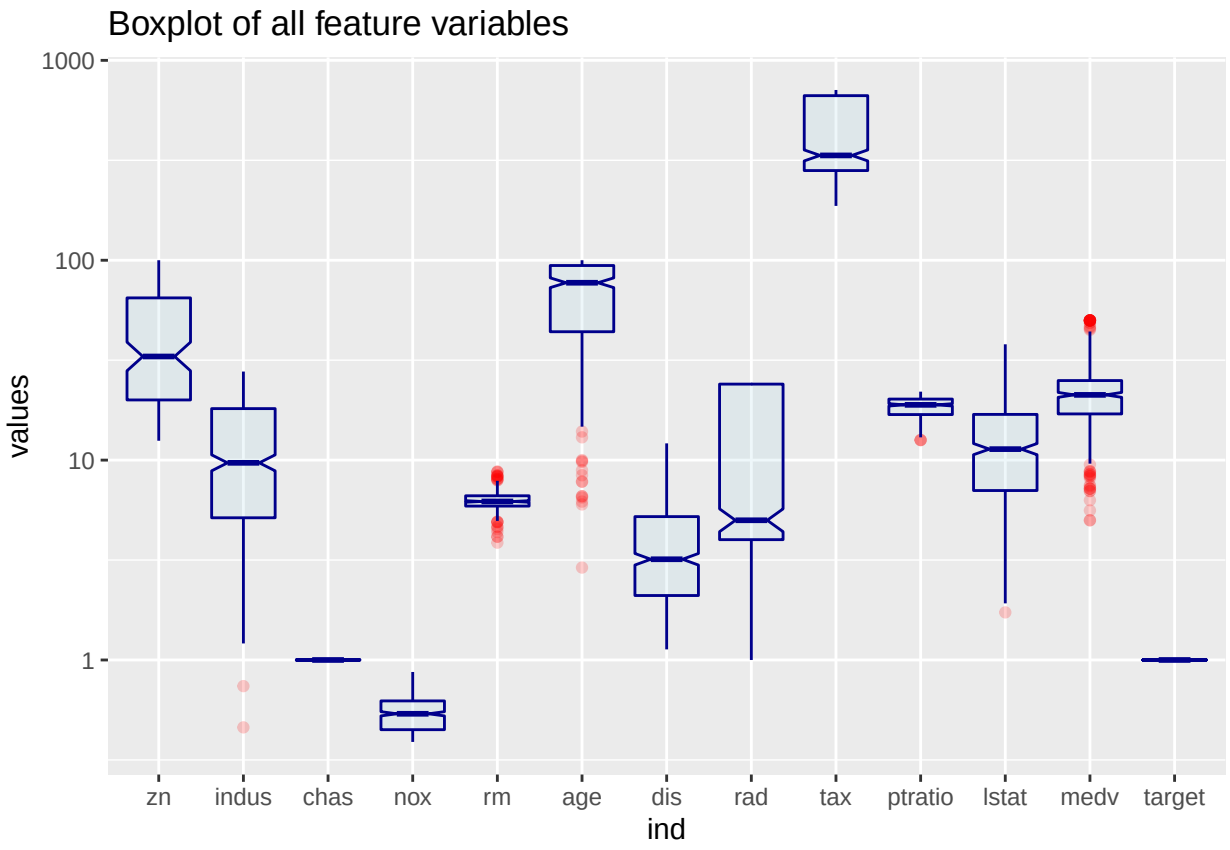
Feature Boxplots

- 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.

```
ggplot(stack(rawTrain), aes(x = ind, y = values)) +
  geom_boxplot(color = "darkblue",
               fill = "lightblue",
               alpha = 0.2,
               outlier.color = "red",
               outlier.fill = "red",
               outlier.alpha = 0.2,
               notch = TRUE) +
  labs(title = "Boxplot of all feature variables") +
  scale_y_log10()
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 1009 rows containing non-finite values (stat_boxplot).
```



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.

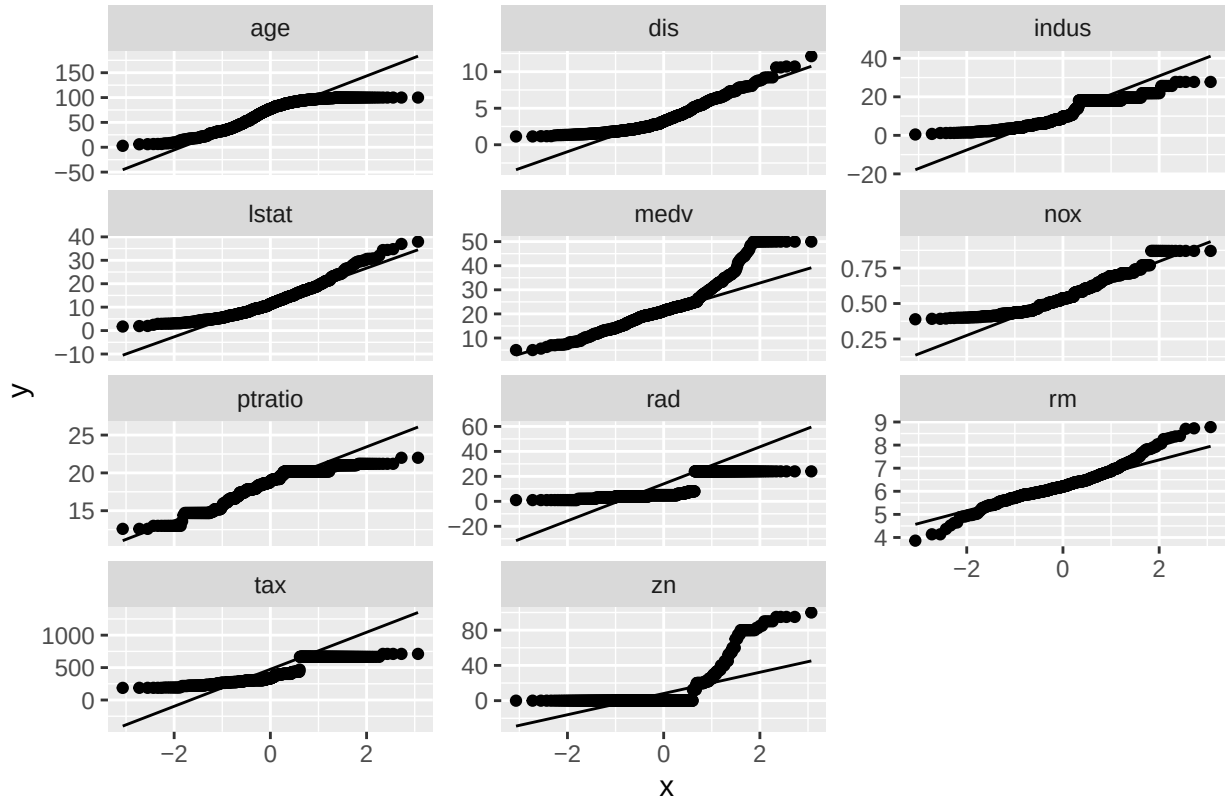
Feature QQ Plots

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

QQ Plots

```
qq_train_data <- rawTrain[, c("age", "dis", "indus", "lstat",
                              "medv", "nox", "ptratio", "rad",
                              "rm", "tax", "zn")]

DataExplorer::plot_qq(qq_train_data, nrow = 4L, ncol = 3L)
```



- 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.

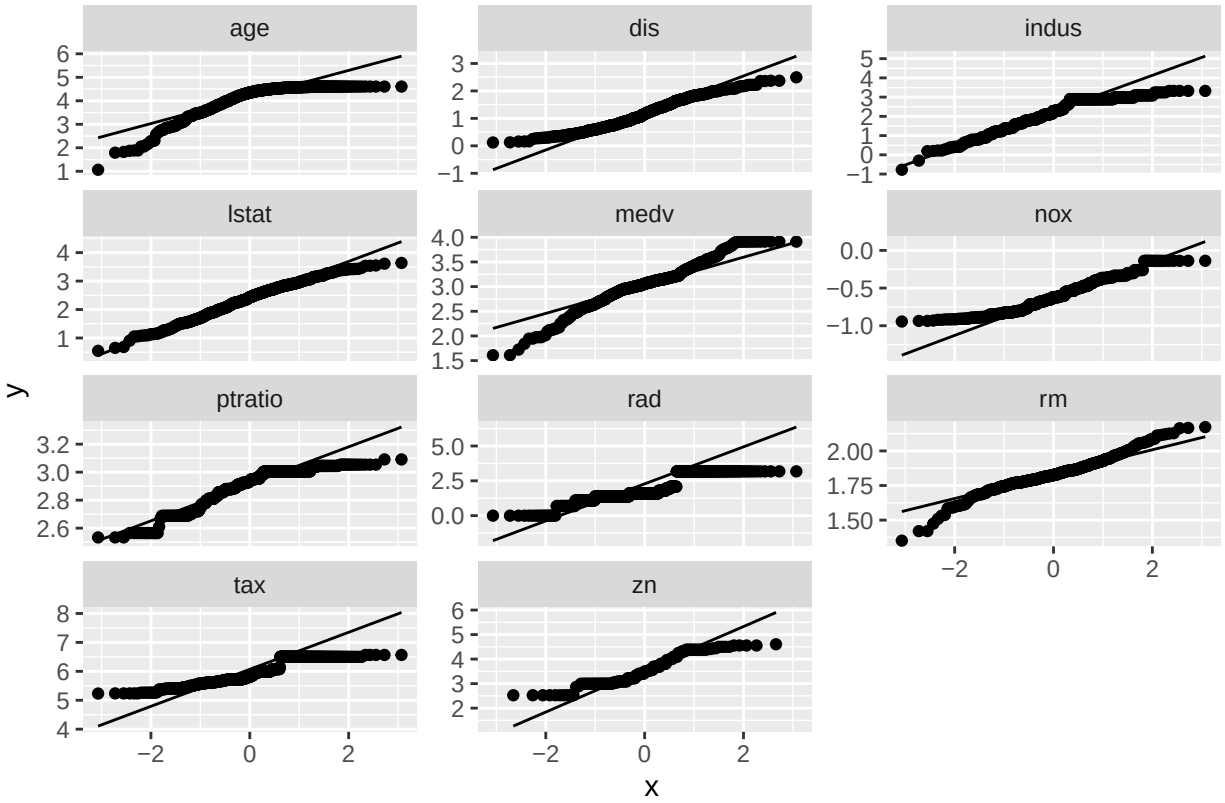
Log QQ Plots

```
log_qq_train_data <- DataExplorer::update_columns(qq_train_data,
                                                  ind = names(qq_train_data),
                                                  what = log)

DataExplorer::plot_qq(log_qq_train_data, nrow = 4L, ncol = 3L)
```

```
## Warning: Removed 339 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 339 rows containing non-finite values (stat_qq_line).
```



- 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.