Advanced Statistical Methods Homework 2

Brandon Hosley

University of Illinois - Springfield

Advanced Statistical Methods Homework 2

Introduction to Statistical Learning
Chapter 4: Problem 10

This exercise involves the *Boston* housing data set.

(a) *How many rows are in this data set?*

```
nrow(Boston)
> 506
```

How many columns?

```
ncol(Boston)
> 14
```

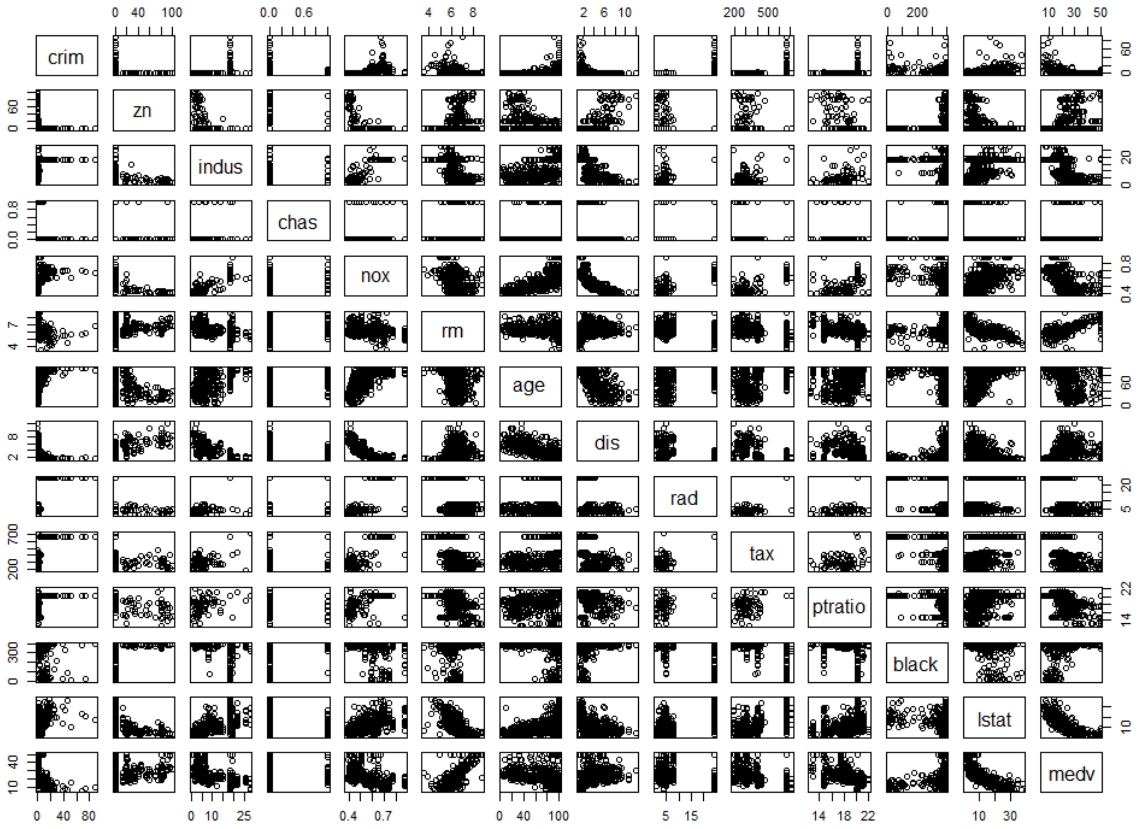
What do the rows and columns represent?

The rows represent variables in the dataset.

The columns represent the attributes of the variables.

(b) *Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.*

```
pairs(Boston)
```



Distance from business center is inversely proportional to Nitrous Oxide concentration. Median value of owner occupied home is proportional to number of rooms but not proportional to the property tax rate.

(c) *Are any of the predictors associated with per capita crime rate? If so, explain the relationship.*

Crime rate appears to be associated with:

- low percentage of houses zoned over 25,000 square feet
- between 15% and 20% non-retail business acres
- older owner-occupied houses
- higher access to radial highways
- higher full-value property tax
- higher pupil to teacher ratio
- (Dummy variable) homes bound by the Charles river

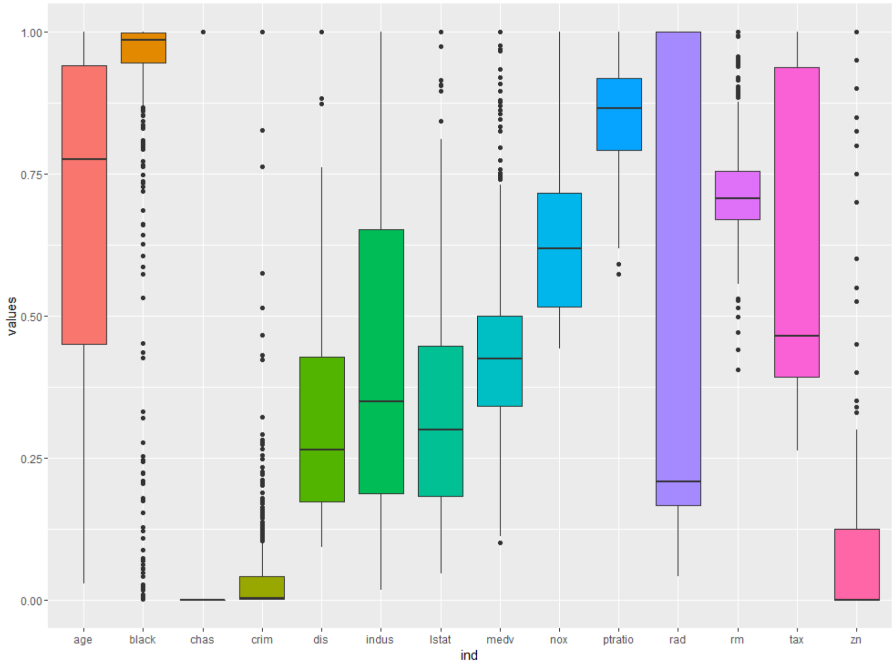
(d) *Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.*

A simple box-plot may demonstrate outliers for each of the factors. Although this method does not correlate individual variables as outliers across factors it does give us the opportunity to observe distribution of factors and in a manner somewhat more clearly than the matrix of scatter-plots was able to. Relationship between outlier data would likely be a worthy next step in investigation.

```r
library(ggplot2)
require(reshape2)

b <- Boston
# Scale data to be relative
b[,-1] = apply(b[,-1],2,function(x){x/max(x)})
b$crim <- b$crim/max(b$crim)

ggplot(stack(b), aes(x = ind, y = values)) +
        geom_boxplot(aes(fill= ind)) +
        theme(legend.position = "none")
```



(e) *How many of the suburbs in this data set bound the Charles river?*

```
library(plyr)
count(Boston$chas, vars = 1)
>   x freq
> 1 0  471
> 2 1   35
```