Q10

Rahul Atre

2023-10-04

Q10 [Data Visualization and Data Exploration]

This exercise involves the Boston housing dataset.

1. Load the Boston dataset (it is part of the MASS library in R).

How many rows are in this dataset? How many columns? What do the rows and columns represent?

We can use the dim() function to obtain the dimensionality of the dataframe

```
dim(Boston)
```

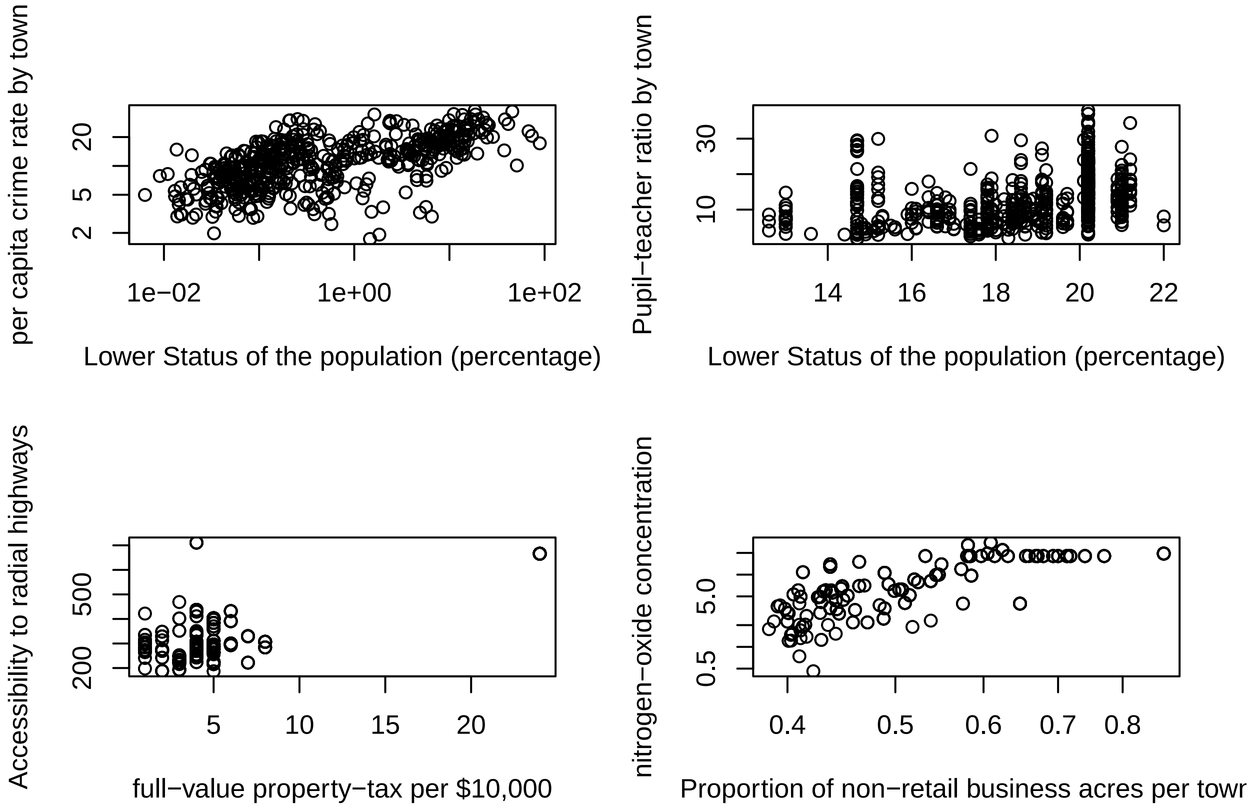
```
## [1] 506  14
```

Therefore, there are 506 rows and 14 columns for this dataset. The rows represent the individual suburbs (or towns) in Boston, and the 14 columns represent the various attributes or features of those suburbs, such as tax, pupil-teacher ratio, crime rate, etc.

2. Make some pairwise scatterplots of the predictors (columns) in this dataset. Describe your findings.

- pairs(Boston) #Output not shown due to size, however it runs properly

```
par(mfrow = c(2, 2))
plot(Boston$crim, Boston$lstat, log = 'xy', xlab = 'Lower Status of the population (percentage)', ylab =

plot(Boston$ptratio, Boston$lstat, xlab = 'Lower Status of the population (percentage)', ylab = 'Pupil-t

plot(Boston$rad, Boston$tax, xlab = 'full-value property-tax per $10,000', ylab = 'Accessibility to radi

plot(Boston$nox, Boston$indus, log = 'xy', xlab = 'Proportion of non-retail business acres per town', yl
```



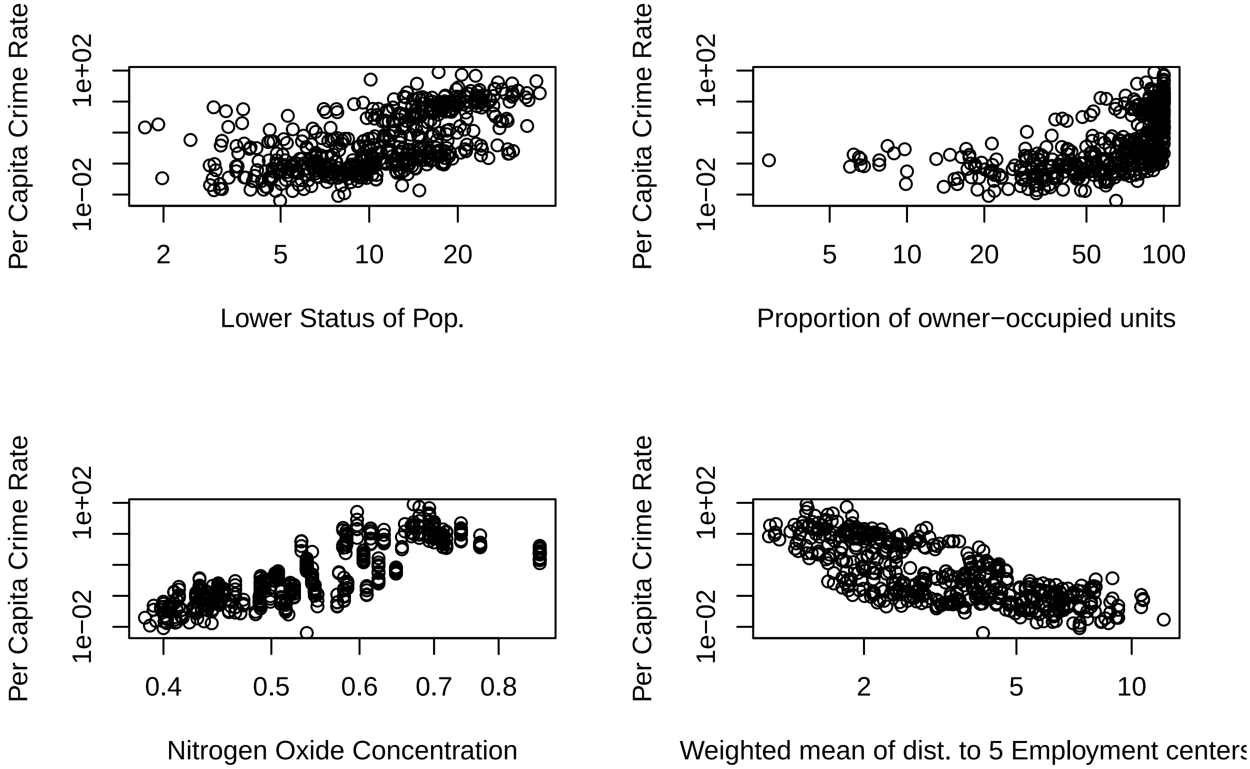
```
#We used a logarithmic scale for the y-axis due to the exponential relationship between some of the pre
```

Based on the scatterplot, we created four subplots to derive the significant findings of the data set, which are:

- There is a positive correlation between the lower status of population and overall crime. If a high percent of population are of lower status, then there is a larger volume of crime in the suburb.
- Similar to the previous case, there is also a positive correlation in lower status of population and a lower pupil-teacher ratio. Suburbs with poorer families have underfunded schools with less teachers and more students.
- There is a weak (still existing) positive correlation in property tax rate per $10,000 and the index of accessibility to radial highways.
- This is a weak (still existing) positive correlation between the proportion of non-retail business acres per town and the nitrogen-oxide concentration. Based on knowledge of science, businesses tend to generate carbon emissions depending on whether they produce goods that utilize fossil fuels. However, as data scientists, we cannot assume that this is the implication of the data. Correlation does not equal causation or in this case, assumptions of outcome through data.

3. Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

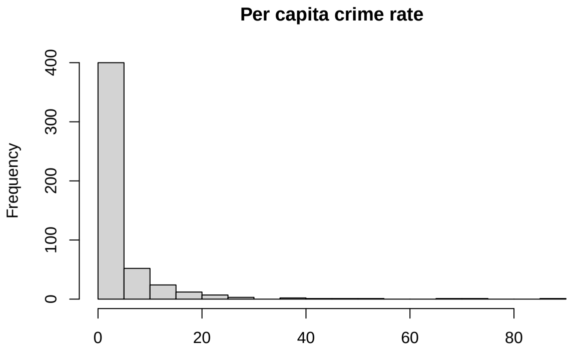
```
par(mfrow = c(2, 2))
plot(Boston$crim ~ Boston$lstat, log = 'xy', xlab = 'Lower Status of Pop.', ylab = 'Per Capita Crime Ra
plot(Boston$crim ~ Boston$age, log = 'xy', xlab = 'Proportion of owner-occupied units', ylab = 'Per Capi

plot(Boston$crim ~ Boston$nox, log = 'xy', xlab = 'Nitrogen Oxide Concentration', ylab = 'Per Capita Cr

plot(Boston$crim ~ Boston$dis, log = 'xy', xlab = 'Weighted mean of dist. to 5 Employment centers', yla
```



For predictors Lower Status, Proportion of owner-occupied units, and nitrogen-oxide concentration, there is a strong positive correlation with the overall crime rate. For the weighted mean of distance to five employment centers however, there is a negative correlation with crime, meaning that when there are more employment centers, the crime rate falls drastically (**exponential**). Also it is important to note that, all of the following correlations are exponential, meaning that even for the positive correlations, there is an **exponential increase** in crime if the predictor variables increase.

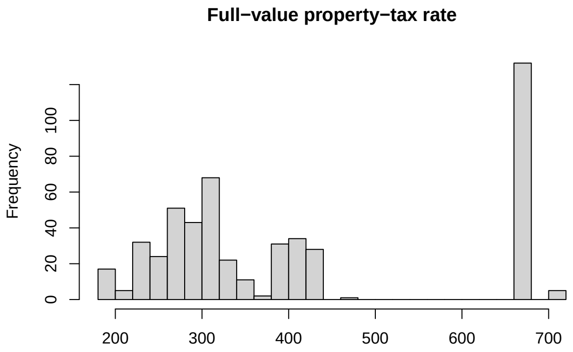
4. Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

```
hist(Boston$crim, breaks=20, main = "Per capita crime rate", xlab ='')
```



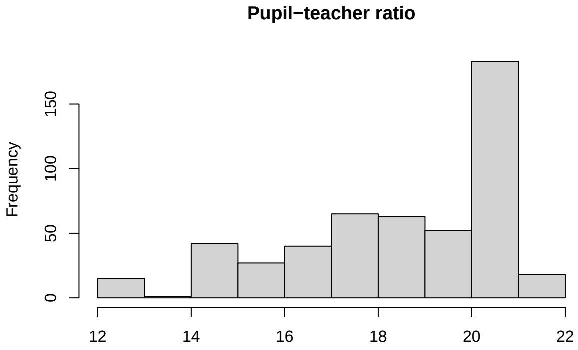
For the above histogram, we can strongly conclude that majority of the suburbs in Boston have zero per capita crime rate. There are only a few suburbs which cross a 20 per capita crime rate, and even fewer that reach till 40.

```
hist(Boston$tax, breaks=25, main = "Full-value property-tax rate", xlab ='')
```



For the full-value property-tax rate, there is a significant number of values near the 700 mark, though a large portion of the suburbs have a tax range between 200 and 400 as well.

```
hist(Boston$ptratio, breaks=10, main = "Pupil-teacher ratio", xlab ='')
```



The above histogram for pupil-teacher ratio indicates that there is an evenly distributed number of suburbs with a ratio between 14 and 22, although there is a relative peak at value 20, indicating a higher frequency of pupil-teacher ratio for that specific data point.

5. How many of the suburbs in this dataset bound the Charles river?

```
sum(Boston$chas == 1)
```

```
## [1] 35
```

6. What is the median pupil-teacher ratio among the towns in this dataset?

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

7. Which suburb of Boston has lowest median value of owner- occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
lowestSuburb = Boston[order(Boston$medv), ] #Order dataframe from lowest to highest medv values
lowestSuburb[1, ] #Select first row from dataframe
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio black lstat
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9 30.59
##     medv
## 399    5
```

From the above calculation, Suburb 399 has the lowest median value of owner-occupied homes at $5000. Provided above are the values of the other predictors for that suburb (i.e. per capita crime is 38.35, average room per dwelling is 5.453, and so forth).

```
summary(Boston)
```

We can generate a summary of all predictors (not shown due to size). Comparing each predictor to the respective quantiles of the summary, it appears that: crim: very high zn: very low indus: high nox: high rm: low age: very high dis: low rad: very high tax: very high ptratio: very high black: very high lstat: high

8. In this dataset, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling

```
sum(Boston$rm > 7)
```

```
## [1] 64
```

There are 64 suburbs that average more than seven rooms per dwelling.

```r
sum(Boston$rm > 8)
```

```
## [1] 13
```

There are 13 suburbs that average more than eight rooms per dwelling.

```r
Boston_avgEight = Boston[Boston$rm > 8, ]
summary(Boston_avgEight)
```

```
##       crim               zn            indus             chas
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000
##       nox               rm             age              dis
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907
##       rad              tax           ptratio          black
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9
##      lstat           medv
##  Min.   :2.47   Min.   :21.9
##  1st Qu.:3.32   1st Qu.:41.7
##  Median :4.14   Median :48.3
##  Mean   :4.31   Mean   :44.2
##  3rd Qu.:5.12   3rd Qu.:50.0
##  Max.   :7.44   Max.   :50.0
```

From the suburbs that average more than eight rooms per dwelling, we can see that the overall crime rate is very low, at a median of 0.52. The lower status of population is also much lower, at a median of 4.14 (median of all suburbs is 11.36). We can conclude that the average suburb with 8 or more rooms per dwelling seem to be maintained better based on a few of the predictors when compared relative to all suburbs in the dataset.