Seif Kungulio
10/27/2024
Project 1
DATA 640
Section: 01W
Instructor: Chris Shannon
File Name: Project1_Kungulio_Seif.docx

1. Read the dataset in Boston.csv into R. Call the loaded data Boston. Make sure that you have the directory set to the correct location for the data.

```
# Set the working directory
setwd("C:/Users/shkungulio/Desktop/DATA-640 Predictive Models/Week_1")

# Load the Boston dataset
Boston <- read.csv("Boston.csv")
```

2. How many rows are in the data frame? How many columns? What do the rows and columns represent?

    a) Number of rows: 506
    b) Number of columns: 14
    c) Each row represents suburbs in the Boston area, and each columns represent different attribute of these suburbs.

3. Select the 1st, 100th, and 500th rows with columns tax and medv.

```
    tax medv
1   296 24.0
100 276 33.2
500 391 17.5
>
```

4. Look at the data using cor function. Are any of the predictors associated with per capita crime rate? If so, explain the relationship based on correlation coefficients.

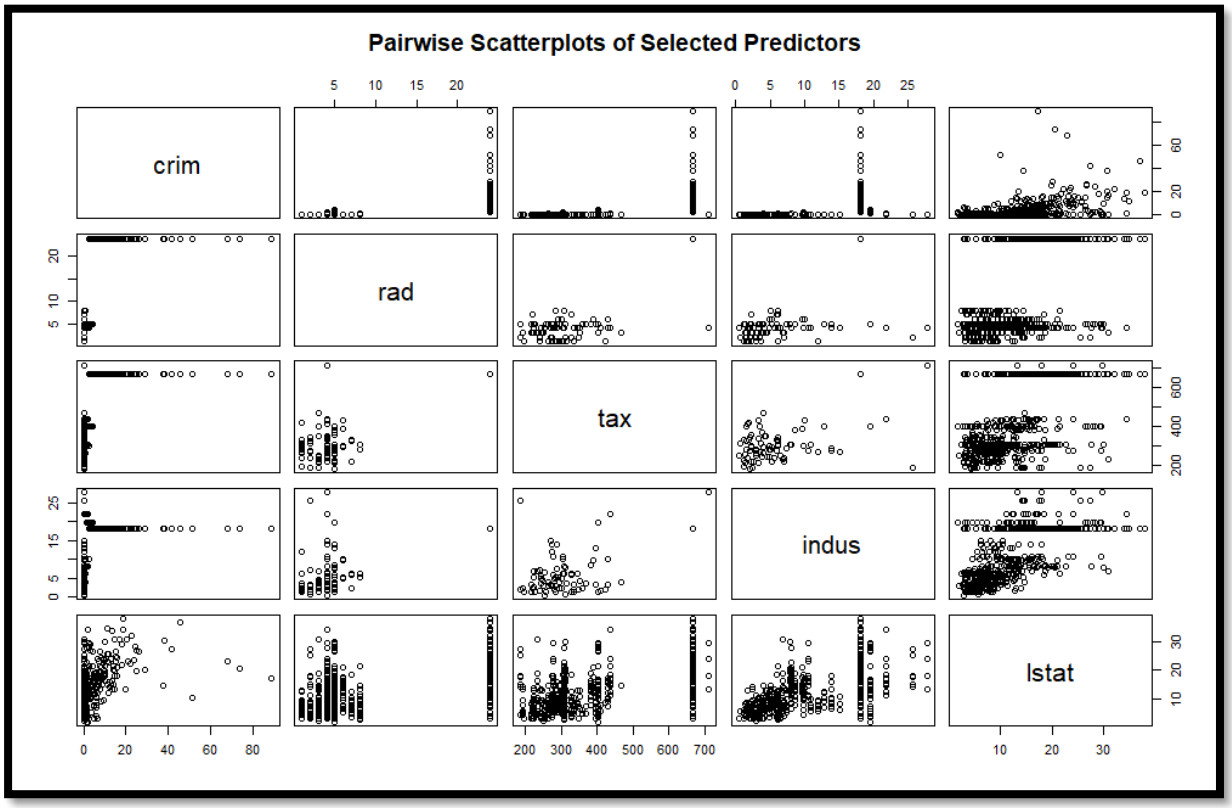
```
> print(cor_matrix["crim", ])
       crim          zn       indus        chas         nox          rm         age
 1.00000000 -0.20046922  0.40658341 -0.05589158  0.42097171 -0.21924670  0.35273425
        dis         rad         tax     ptratio       black       lstat        medv
-0.37967009  0.62550515  0.58276431  0.28994558 -0.38506394  0.45562148 -0.38830461
>
```

    1. The correlation matrix gives an insight into how different predictors relate to each other and specifically with the per capita crime rate (crim). For instance:

a. We might observe that crim is positively correlated with variables like indus (proportion of non-retail business acres per town), indicating that areas with more industrial zones tend to have higher crime rates.
b. On the other hand, crim may show a negative correlation with dis (weighted mean of distances to five Boston employment centers), implying that suburbs farther from employment centers generally have lower crime rates.

2. Correlation values close to 1 or -1 indicate stronger relationships, while values near 0 suggest weaker relationships. A detailed understanding of these relationships helps in identifying factors associated with crime rates in the Boston suburbs.

5. Make some pairwise scatterplots of the predictors, crim, rad, tax, indus, and lstat in this data set. Describe your findings.



From the pairwise scatterplot matrix, I can examine the relationships between the predictors: crim, rad, tax, indus, and lstat. Here's a description of each pairwise relationship and key observations:

- crim vs. rad:
  There seems to be a concentration of points at lower levels of both variables, indicating that many observations have low crime rates and lower accessibility to radial highways (rad). However, there are some higher-crime areas that seem correlated with higher rad values, which might suggest a relationship between highway accessibility and crime rates.

- crim vs. tax:
  There is a slight spread of points as the tax values increase, but the scatterplot suggests that higher crime rates may occur with higher tax rates. However, there are many cases clustered at lower tax values and low crime rates.

- crim vs. indus:
  A positive trend appears to exist, indicating that areas with a higher proportion of non-retail business (industrial areas) may experience higher crime rates. There are a few significant outliers in the high-crime and high-indus range, which should be investigated further.

- crim vs. lstat:
  The relationship between crim and lstat suggests a moderate positive correlation. As the percentage of the lower socioeconomic status population (lstat) increases, there appears to be a corresponding rise in crime rates. This could indicate a socio-economic impact on crime.

- rad vs. tax:
  There seems to be a strong positive relationship between rad and tax, indicating that areas with more radial highways also have higher tax rates. This could reflect urban infrastructure and property value relationships.

- rad vs. indus:
  There seems to be some clustering of points with higher rad values associated with areas of higher industrial activity (indus). This suggests that areas with greater accessibility to radial highways are more industrialized.

- rad vs. lstat:
  There isn't a very clear trend in the rad vs. lstat scatterplot, but there are a few cases with higher lstat values and mid-range rad values.

- tax vs. indus:
  A positive trend appears in the tax vs. indus plot, suggesting that areas with a higher proportion of industrial activities tend to have higher tax rates. This relationship seems relatively strong compared to other pairs.

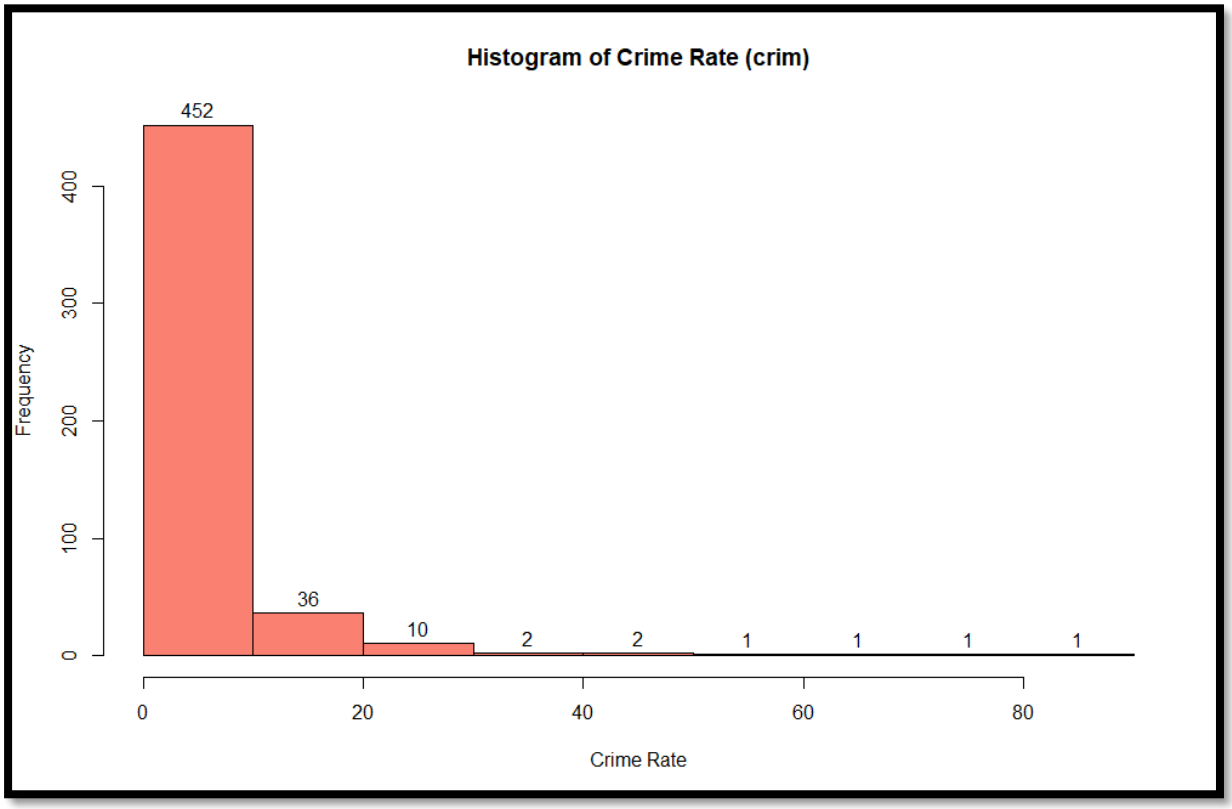
- tax vs. lstat:
  There isn't a clear linear trend between tax and lstat, but there seems to be a mild positive association between these two variables.

- indus vs. lstat:
  A moderate positive correlation appears between these two variables, indicating that areas with a higher percentage of industrial activity may also have higher proportions of lower socioeconomic status residents.

Key Takeaways:

- There are clear relationships between some pairs of variables, such as rad vs. tax and indus vs. lstat, which suggest potential co-dependencies or shared influences in the dataset.
- The relationship between crim and variables like lstat, indus, and rad indicates that socioeconomic and infrastructural factors could be contributing to crime rates.
- Outliers exist in multiple plots, indicating the presence of extreme values that could skew simple linear models.

6. Do any of the suburbs of Boston appear to have particularly high crime rates by looking at the histogram of crim? What is the range of crim by using range() function in R?



a. Based on the histogram, the distribution of the crime rate (crim) variable is heavily right-skewed. Most of the suburbs have relatively low crime rates (close to zero), with the largest frequency concentrated in the first bin. This shows that a vast number of suburbs in Boston have very low crime rates.
However, the presence of the bins beyond 20, especially those that reach above 80, indicates that there are some suburbs with unusually high crime rates. These are outliers compared to the rest of the data. Specifically, there are about five or six suburbs with significantly higher crime rates, with at least one of them having a crime rate exceeding 80.

Key Observations:

- Most suburbs have low crime rates: A large majority of the data points are clustered in the first bin, showing that around 452 suburbs have very low crime rates, close to zero.
- A few suburbs have extremely high crime rates: The presence of a few higher bins, particularly beyond 20, indicates that there are some suburbs with notably higher crime rates. These suburbs are distinct outliers compared to the rest of the distribution.

In summary, while most of the suburbs exhibit low crime rates, a few suburbs stand out with significantly higher crime rates, making them noteworthy outliers in this dataset.

b. Range of crime: 0.00632 88.9762

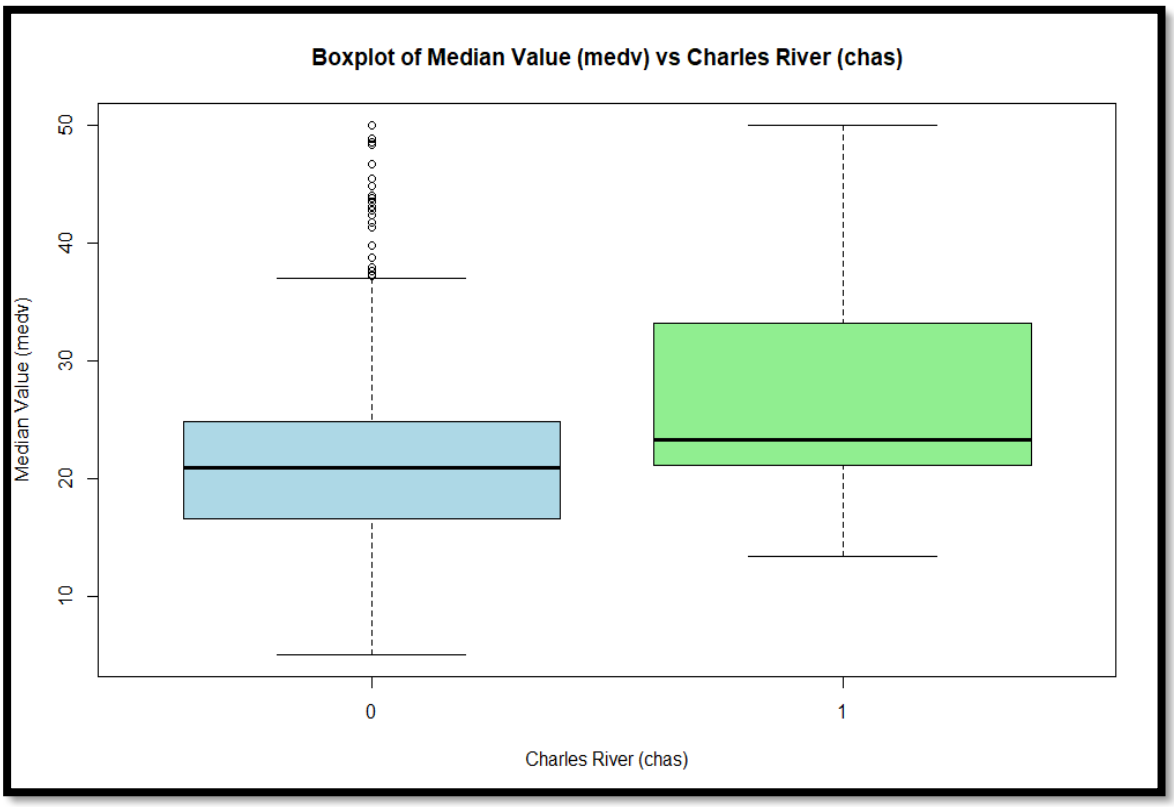
7. How many of the suburbs in this dataset bound the Charles River?
    a. The number of suburbs that border the Charles River is: 35

8. What is the median pupil-teach ratio among the towns in this dataset? What's the mean?
    a. Median pupil-teach ratio: 19.05
    b. Mean pupil-teach ratio: 18.45553

9. 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.
    a. Suburbs averaging more than 7 rooms: 64
    b. Suburbs averaging more than 8 rooms: 13
    c. Suburbs with more rooms than 8 rooms per dwelling are often indicative of greater affluence, as larger homes tend to correlate with higher property values and wealthier populations. These areas may present attractive investment opportunities due to their spacious housing options and the potential for long-term value appreciation.

10. Convert chas to a factor. Boxplot the medv against chas. Are houses around the Charles River more expensive?
    a. Boston$chas <- as.factor(Boston$chas)

b. Boxplot



c. The boxplot might show that homes near the Charles River (where chas = 1) have a higher median value compared to those farther away (chas = 0). This makes sense, as being close to a major water body like the Charles River could be perceived as an amenity, leading to higher property values in those suburbs.