Nicholas Allen  
03/17/2024  
Project 1  
DSCI 512  
Section: 1W 24/SP2  
Instructor: Nengbing Tao  
File Name: Project1\_Allen\_Nicholas.docx

# Answers

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.
   1. A screenshot of a computer screen

      Description automatically generated
2. How many rows are in the data frame? How many columns? What do the rows and columns represent?
   1. A white background with blue text

      Description automatically generated
   2. 506 rows each with 13 columns
   3. Each row represents a single observation. They appear to be unique neighborhoods of Boston. Each column is an attribute of the neighborhood, for instance crime rate or median home value.
3. Select the 1st, 100th, and 500th rows with columns tax and medv.
   1. A close-up of a computer code

      Description automatically generated
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.

A group of numbers on a white background

Description automatically generated

* 1. A coefficient is returned for all predictors so on some level they are all associated with per capita crime rate. The strongest positive correlations seem to be with rad and tax. This would suggest same direction moves between the 3 values. If rad and tax grow larger, per capita crime rate would as well. If both rad and tax fall, we would expect per capita crime rate to as well, though not as quickly as the coefficients are not a perfect 1. Similarly, the strongest negative relationships are dis and medv. This means as the distance to employment centers or the median home price moves, the crime rate typically moves in the opposite direction. If home prices increase, crime would be expected to decrease, though at a slower pace due to the less than 1 coefficient (0.38).

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

A screenshot of a computer

Description automatically generated

* 1. In general, the plots confirm the correlation coefficients, however visually we can see more than correlation. We can see the verticality of many variables, reinforcing their discrete rather than continuous nature. Additionally, we can see outliers such as some very high tax assessed value homes. Lastly, we can also see some clustering as well, for instance, indus versus lstat looks almost like two separate plots with a relative dearth of data in the middle.

1. 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?
   1. A graph of a number of people

      Description automatically generated with medium confidence
   2. 
   3. By only looking at the histogram it is difficult to discern if any suburbs have a high crime rate as the frequency/myriad of low crime rates drowns out any visual cues of high crime rates. The range shows slightly different results. It proves that at least one suburb has a particularly high crime rate.
2. How many of the suburbs in this dataset bound the Charles River?
   1. 
3. What is the median pupil-teach ratio among the towns in this dataset? What’s the mean?
   1. A white background with blue text

      Description automatically generated
4. 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.
   1. A screenshot of a computer

      Description automatically generated
   2. A screenshot of a computer

      Description automatically generated
   3. With the exception of od 1 suburb, all suburbs where houses average more than 8 rooms per dwelling have median home values well above the median of the whole dataset.
5. Convert chas to a factor. Boxplot the medv against chas. Are houses around the Charles River more expensive?
   1. A graph of a home value

      Description automatically generated
   2. Houses around the Charles River do appear more expensive as each attribute of the box plot appears higher for houses near the river. The minimum value of the right boxplot is higher than the left. Similarly, the entire box is higher (Q1, median and Q3). The maximum value appears higher for the right box despite some outliers for the non-river homes.