**Student: Seif Kungulio**

**Date: 03/16/2025**

**Subject: Project 1**

**Class: DSCI 512**

**Section: 01W**

**Instructor: Dr. Nengbing Tao**

**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.

A blue screen with white text

AI-generated content may be incorrect.

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

A blue screen with white text

AI-generated content may be incorrect.

1. Number of rows: 506
2. Number of columns: 13
3. Each row represents suburbs in the Boston area, while each column corresponds to a different attribute of these suburbs.
4. Select the 1st, 100th, and 500th rows with columns tax and medv.

A blue screen with white text

AI-generated content may be incorrect.

1. 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 blue screen with white text

AI-generated content may be incorrect.

1. The correlation matrix provides valuable insight into the relationships between different predictors, particularly in relation to the per capita crime rate (crim). For example:
   1. Positive correlation: crim appears to be positively correlated with variables like indus (the proportion of non-retail business acres per town), suggesting that areas with more industrial zones tend to have higher crime rates.
   2. Negative correlation: Conversely, crim may show a negative correlation with dis (the weighted mean of distances to five Boston employment centers), indicating that suburbs farther from employment centers generally have lower crime rates.
2. Correlation values close to 1 or -1 signify strong relationships, while values near 0 indicate weak or no significant relationships. A detailed understanding of these correlations is crucial for identifying factors that influence crime rates in the Boston suburbs.
3. 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 screen

AI-generated content may be incorrect.

From the pairwise scatterplot matrix, we can examine the relationships between the predictors: crim, rad, tax, indus, and lstat. Below is a description of each pairwise relationship along with key observations:

* **crim vs. rad:**

A concentration of points at lower levels of both variables suggests that many areas have low crime rates and limited access to radial highways (rad). However, some higher-crime areas show a correlation with higher rad values, potentially indicating a relationship between highway accessibility and crime rates.

* **crim vs. tax:**

The scatterplot reveals a slight spread of points as tax values increase, suggesting that higher crime rates may coincide with higher tax rates. Still, many observations are clustered around lower tax and crime values, indicating exceptions to this trend.

* **crim vs. indus:**

A positive trend is visible, implying that areas with a higher proportion of industrial activities (non-retail business) tend to experience higher crime rates. Notably, there are several significant outliers in the high-crime and high-industry range, which warrant further investigation.

* **crim vs. lstat:**

There appears to be a moderate positive correlation between crime rates (crim) and the percentage of the population with lower socioeconomic status (lstat). This suggests that areas with a higher proportion of low-income residents tend to experience higher crime rates, potentially reflecting a socioeconomic influence on crime.

* **rad vs. tax:**

A strong positive relationship is evident between radial highway accessibility (rad) and tax rates, suggesting that areas with better highway access tend to have higher tax rates. This could be reflective of urban infrastructure and property value dynamics.

* **rad vs. indus:**

There is some clustering where areas with higher rad values are associated with higher industrial activity (indus), suggesting that areas with better access to radial highways are more industrialized.

* **rad vs. lstat:**

The relationship between rad and lstat is not immediately clear, though a few cases show higher lstat values paired with mid-range rad values.

* **tax vs. indus:**

A positive trend between tax rates and industrial activity suggests that areas with a higher proportion of industrial activities tend to have higher tax rates. This relationship appears relatively strong compared to other pairs.

* **tax vs. lstat:**

While no clear linear trend emerges between tax rates and lstat, there seems to be a mild positive association, indicating a weak correlation between higher taxes and a higher proportion of low-income residents.

* **indus vs. lstat:**

A moderate positive correlation is apparent between industrial activity and lower socioeconomic status, suggesting that areas with a higher concentration of industry tend to have higher proportions of low-income residents.

**Key Takeaways:**

* Strong relationships exist between some pairs of variables, such as rad vs. tax and indus vs. lstat, suggesting potential co-dependencies or shared influencing factors within the dataset.
* The correlations between crim and variables like lstat, indus, and rad suggest that socioeconomic factors and infrastructure may play a significant role in determining crime rates.
* Outliers are present in multiple scatterplots, indicating the presence of extreme values that could potentially skew linear models or require further analysis.

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?

A graph with a red bar

AI-generated content may be incorrect.

1. Based on the histogram, the distribution of the crime rate (crim) variable is heavily right skewed. Most suburbs have relatively low crime rates, with the highest frequency concentrated in the first bin, indicating that most suburbs in Boston experience very low crime rates.

However, there are bins beyond a crime rate of 20, particularly those exceeding 80, which suggest the presence of 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 notably higher crime rates, with at least one suburb surpassing the crime rate of 80.

**Key Observations:**

* Low crime rates dominate: Most of the data points are clustered in the first bin, indicating that approximately 452 suburbs have very low crime rates, close to zero.
* Outliers with high crime rates: The few higher bins, especially those beyond 20, highlight suburbs with significantly higher crime rates. These suburbs are clear outliers in the distribution.

In summary, while most suburbs exhibit low crime rates, a small number of suburbs with much higher crime rates stand out as distinct outliers in this dataset.

1. Range of crime: 0.00632 88.9762

A blue screen with white text

AI-generated content may be incorrect.

1. How many of the suburbs in this dataset bound the Charles River?

A blue screen with white text

AI-generated content may be incorrect.

The number of suburbs that border the Charles River is: 35

1. What is the median pupil-teach ratio among the towns in this dataset? What’s the mean?

A blue screen with white text

AI-generated content may be incorrect.

The median pupil-teach ratio: 19.05 and the mean pupil-teach ratio: 18.45553

1. 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 screenshot of a computer screen

AI-generated content may be incorrect.

1. Suburbs averaging more than 7 rooms: 64
2. Suburbs averaging more than 8 rooms: 13
3. Suburbs with more than 8 rooms per home are typically associated with greater affluence, as larger homes often correlate with higher property values and wealthier populations. These areas may represent attractive investment opportunities due to their spacious housing and potential for long-term value appreciation.
4. Convert chas to a factor. Boxplot the medv against chas. Are houses around the Charles River more expensive?
5. Boston$chas <- as.factor(Boston$chas)

A screenshot of a computer program

AI-generated content may be incorrect.

1. Boxplot

A green and blue rectangular shapes

AI-generated content may be incorrect.

1. 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.