****

**Prepared By:** Sandeep Singh

**Student ID:** C0727422

**Subject:** 2018S-T2 BDM 1053 - Big Data Algorithms and Statistics 01

**Title:** Assignment 3

**Week:** 28 May - 3 June

**Phone:** +1 705-977-1868

**Email (Internal):** c0727422@mylambton.ca

**Email (External):** official.sandeep93@gmail.com

**Submitted To:** Mr. [Pedram Habibi](https://moodle.queenscollege.ca/moodle/user/view.php?id=370&course=1590)

**Submission Date:** 07 June, 2018

Contents

[Assignment 3 3](#_Toc516160917)

[Questions 3](#_Toc516160918)

[Answers and explanation of the code 3](#_Toc516160919)

[Question (a) 3](#_Toc516160920)

[Question (b) 4](#_Toc516160921)

[Question (c) 7](#_Toc516160922)

[Question (d) 8](#_Toc516160923)

[Question (e) 9](#_Toc516160924)

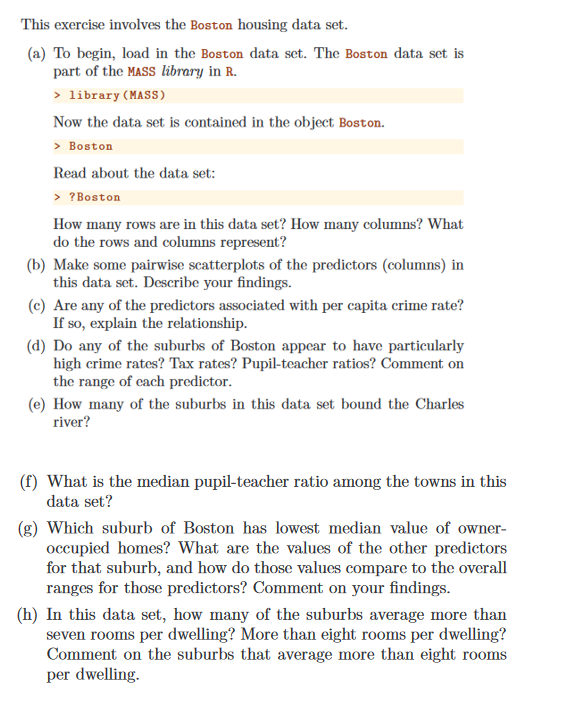
[Question (f) 9](#_Toc516160925)

[Question (g) 9](#_Toc516160926)

[Question (h) 9](#_Toc516160927)

# Assignment 3

# Questions



# Answers and explanation of the code

Initially, the current working directory has been set. Next, all relevant libraries have been imported.

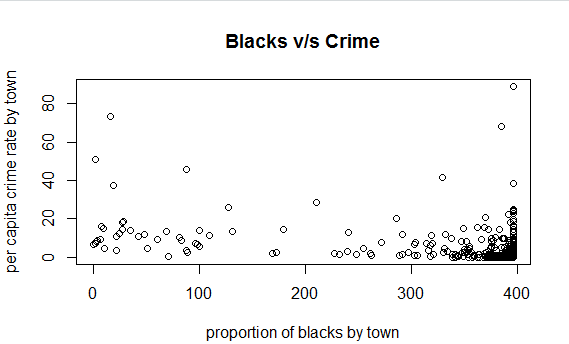
Contents of Boston dataframe were looked upon using View() and help was extracted using ?Boston

## Question (a)

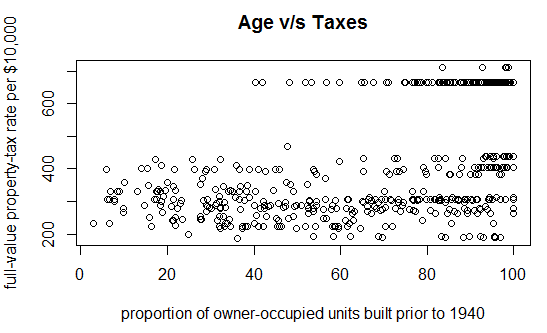
* nrow(Boston) has been used to get the number of rows in the data.
* ncol(Boston) has been used to get the number of columns.
* Rows and columns represent the housing data of various suburbs in Boston. It includes various parameters (fields) such as crime rate, distance from employment center, pupil-teacher ratio, taxes, etc.

## Question (b)

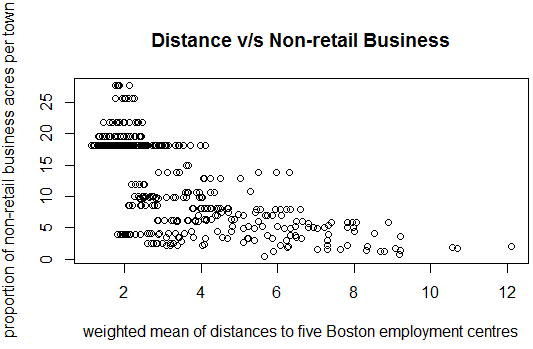
Five scatter plots have been plotted to analyse different scenarios from Boston data.

1. 

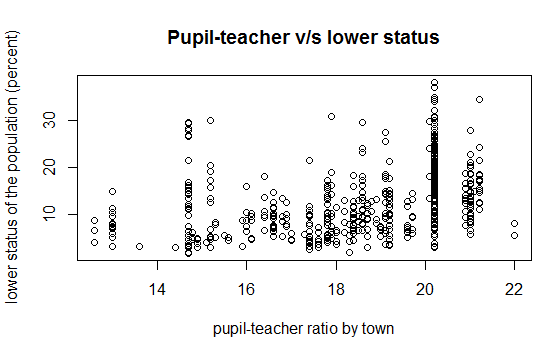
The graph is “*per capita crime rate by town*” vs. “*proportion of blacks by town*”. It is observed that the density of the plots is more towards the bottom right side of the graph. It can be deduced that the crime rate is low where the concentration of blacks is more.

1. 

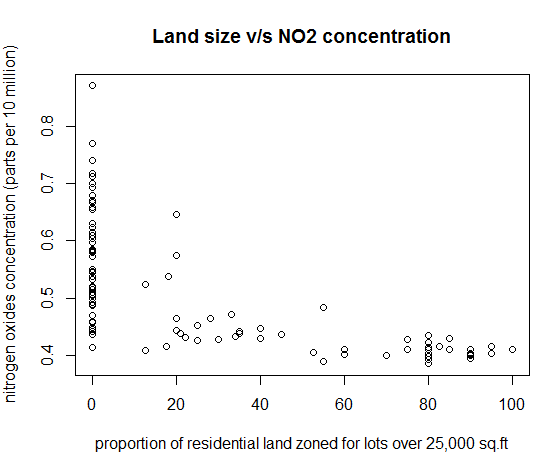
* The plot depicts two trends. The property tax randomly varies between 0 and 400 as the owner-occupied units increases. Secondly, high density of owner-occupied units pay taxes greater than 600 units.

1. 

* The plot is “*weighted mean of distances to five Boston employment centres*” vs “*proportion of non-retail business acres per town*” graph. It is observed that most of the non-retail business is located near the Boston employment centres. As this distance increases, the non-retail business is observed to decrease.

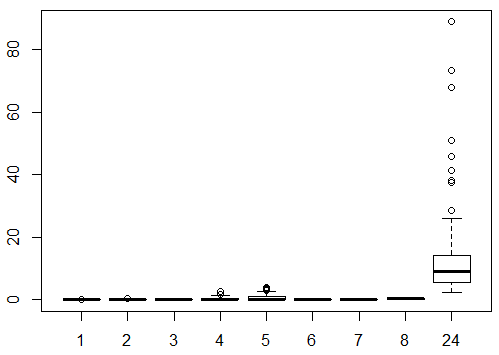
1. 

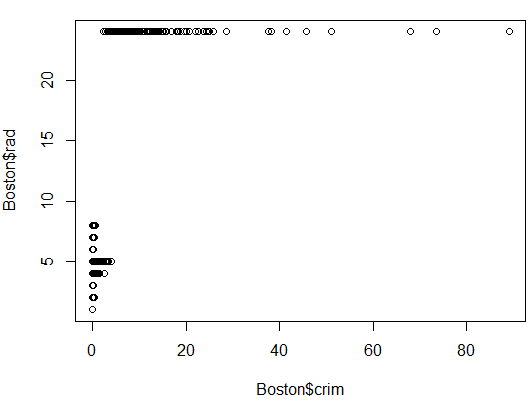
* The scatter plot is a “pupil-teacher ratio by town” vs. “lower status of the population (percent)” graph. Linear lines are observed for different pupil-teacher ratios. However, highest number is observed close to pupil-teacher ratio 21 where population as a lower status. This area can be looked upon by the govt. to take necessary actions (such as generating employment opportunities) to increase the status of the people.

1. 

* The scatter plot of “Land size fir lots over 25,000 sq.ft” and “nitrogen oxides concentration” depicts that lower amount of nitrogen oxide concentration is emitted for larger sized residential land zones.

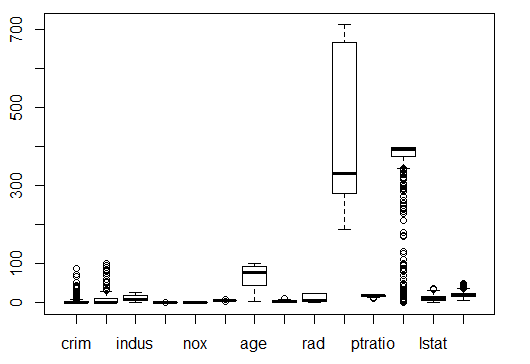
## Question (c)

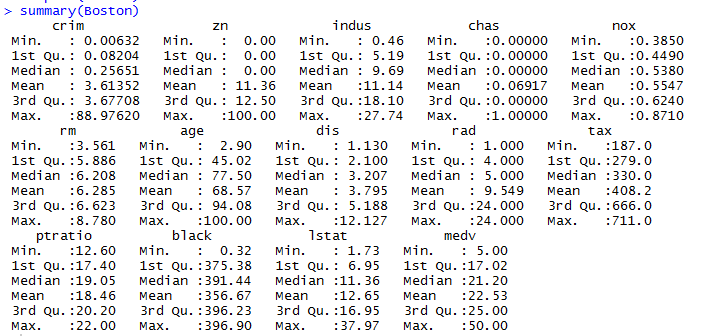




Boxplot of crime rate vs. other fields have been plotted to deduce a relationship. On plotting all the graphs, it is observed that crime rate and accessibility to highway is related. The more is the index of accessibility to the highway, the more is the crime rate. It means towns having easy accessibility to highways experience more crimes than the towns away from the highways.

## Question (d)





We see that 4 suburbs have high crime rates when we filter the data for crime rates > 50. After finding the median, we deduce that many of the suburbs have high taxes and ptratio as well.

After observing the summary and plots, we can deduce that many towns have low crime rates with no zoning. Nitrogen Oxide concentration(betwen 0 and 1) is randomly distributed. Only a few suburbs are bound to Charles river. The black population is very low for quite a number of towns. RAD also varies randomly form 0 to 24 within the suburbs.

## Question (e)

To find the number of suburbs bounding Charles river, first ‘chas’ column has been filtered where value equals 1 and then counted the records using count().

## Question (f)

Median pupil-teacher ratio has been found using median().

## Question (g)

Value for lowest median owner-occupied home was found using min() on ‘medv’ column. There are two records with lowest value which have been displayed using View(). Suburb #399 and #406 have the lowest owner-occupied homes equal to 5.

These suburbs 399 are one of the least desirable places to live in Boston as depicted from the data. Crime rate is very high compared to other Boston neighborhoods including no residential land zoned for lots over 25,000 sq.ft. Proportion of non-retail business acres per town is very high compared to most suburbs. These suburbs are not bound by the Charles river. Nitrogen oxides concentration (parts per 10 million) is high. Average number of rooms per dwelling is one of the lowest. Highest proportion of owner proportion of owner-occupied units built prior to 1940. Low weighted mean of distances to five Boston employment centres.

## Question (h)

Value for suburb average greater than 7 and 8 have been found using filter() on ‘rm’ column. For suburb average more than 8, it is observed that the crime ration is high, the proportion of residential land zoned for lots over 25,000 sq.ft is mostly 0, proportion of non-retail business acres per town is less overall. There is only one suburb that bounds river Charles, nitrogen oxide concentration does not fluctuate much. The proportion of owner-occupied units built prior to 1940 is mostly between 70 and 90. These suburbs are near to Boston employment centers and radial highway. Full-value property-tax ranges from 250 to 350. The suburbs also have a good concentration of pupil-teacher ratio and the blacks living in the area. The status of the population is low in these suburbs and the median is between 40 and 50.

\*\*\*\*\*