ASSESSMENT 1 – Pratical assignment

Pierce Kinghorn - 12091381

CQUniversity

Data Science - COIT12209

Table of Contents

[Question 1 2](#_Toc17003149)

[Question 2 3](#_Toc17003150)

[Question 3 4](#_Toc17003151)

[Question 4 4](#_Toc17003152)

[Question 5 4](#_Toc17003153)

[Question 6 5](#_Toc17003154)

[Question 7 6](#_Toc17003155)

[Question 8 8](#_Toc17003156)

[Question 9 9](#_Toc17003157)

[Question 10 17](#_Toc17003158)

# Question 1

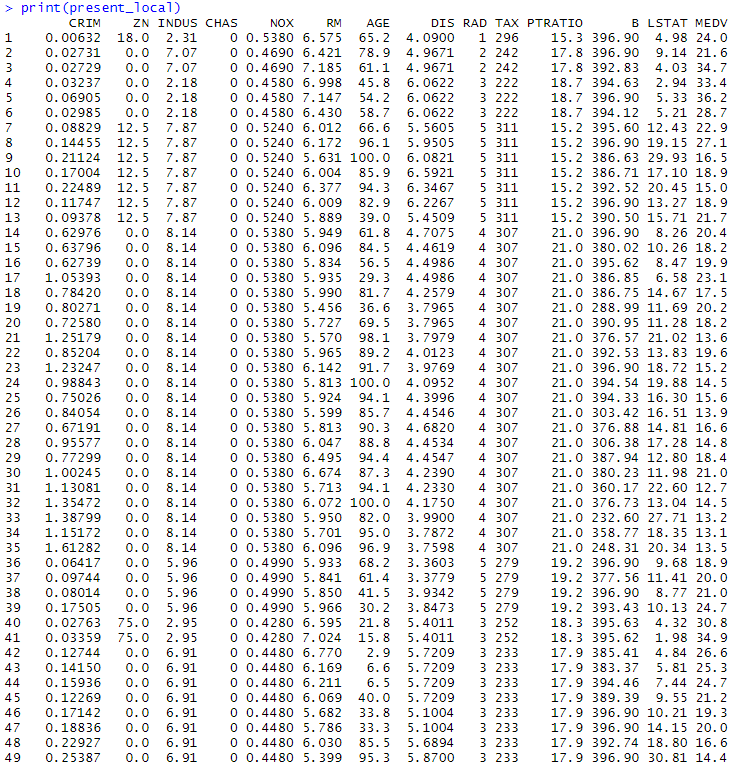
The code required to complete question one is:

setwd('E:\\CQU\\Data Science\\Assessment')

present\_local<- read.csv("housing.csv")

print(present\_local)

This code selects the location to select the file from, before creating a vector called ‘present\_local’ and storing the housing data within it. To confirm the action, it then prints all the data within the vector. The first 49 rows are as shown below (cant fit it all).



# Question 2

In order to complete question two, the code is:

dim(present\_local)

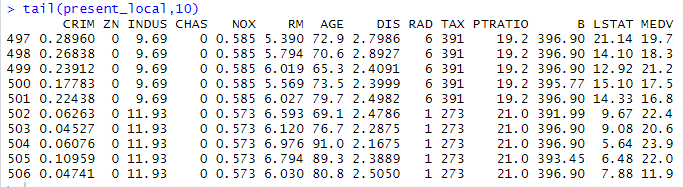
This will display the number of rows and columns within the data set.



# Question 3

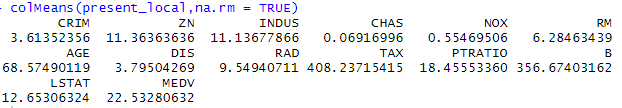
To display the last 10 rows of the data set the code is:

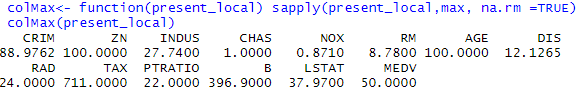
tail(present\_local,10)

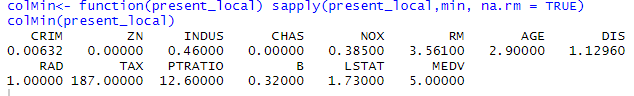


# Question 4

To display the mean, max, min and standard deviation and assuming the relevant packages are installed, the code for each of these are:







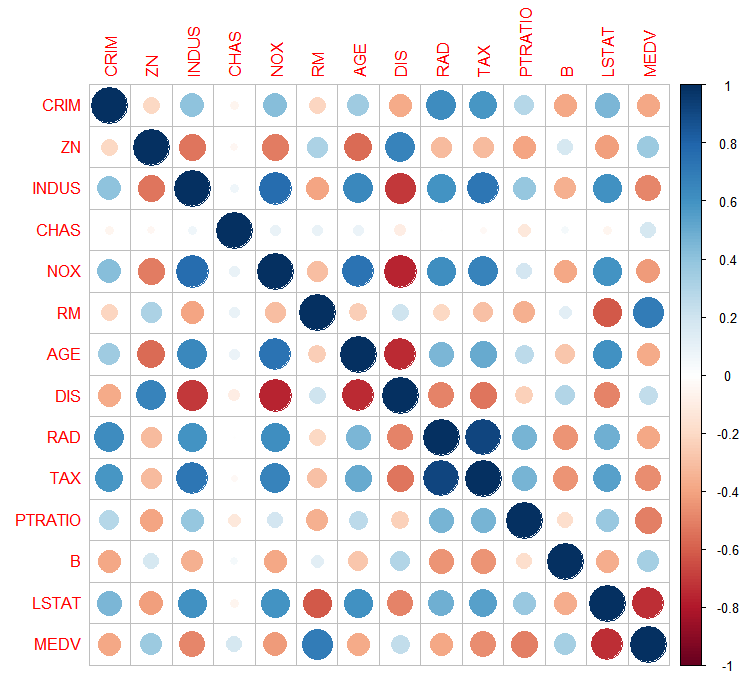


# Question 5

To create the necessary correlation matrix, the code is:

M<- cor(present\_local)

corrplot(M, method = "circle")



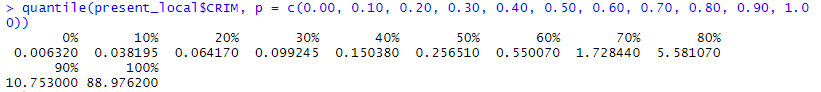
This correlation matrix shows that the capita crime rate per town rises along with median value of owner-occupied homes, distance to employment centres and proportion of blacks by town. Contrasting this, crimes are low in the case where property tax is high and there is access to radial highways.

# 

# Question 6

To see the summary of per capita crime rate by town the code used is:

quantile(present\_local$CRIM, p = c(0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00))

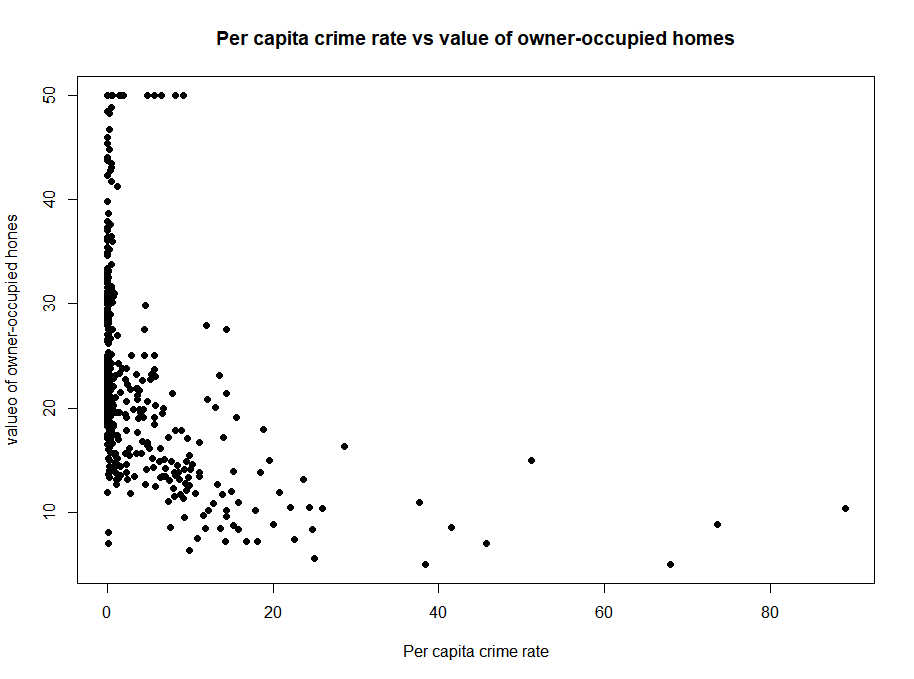


The data shows that is considered a high crime rate for a town when the capita per crimes is higher is 10.75 and makes of the remaining 78.2. low crime rate is when capita per crimes is below 0.03 a large difference compared to high crime rates.

# Question 7

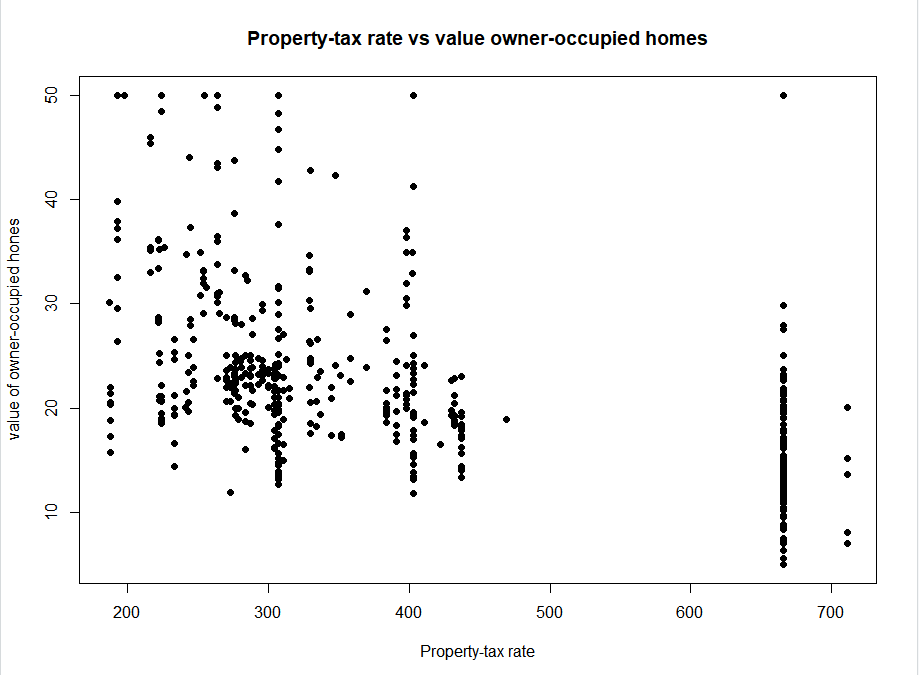
To plot the relationship between each of the variables the code is:

plot(CRIM,MEDV, main ="Per capita crime rate vs owner-occupied homes",xlab="Per capita crime rate",ylab="Owner-occupied hones", pch=19)



This scatter plot shows that crimes rates dramaticaly drop of as the value of owner-occupied houses rises. Crimes mostly occur when the per capita crime rate is low and slowly decays as the per capita increases.

plot(TAX,MEDV, main ="Property-tax rate vs owner-occupied homes", xlab="Property-tax rate",ylab="Owner-occupied hones", pch=19)



Reviewing the scatterplot above it shows that the median value of owner occupied homes clusters at around the 20 mark along with property tax rate sitting at 300 and below. It also shows that a large number of the property tax rates are sitting below 700 at around 660.

plot(LSTAT,MEDV, main ="%lower status of population vs owner-occupied homes", xlab="%lower status of population",ylab="Owner-occupied hones", pch=19)

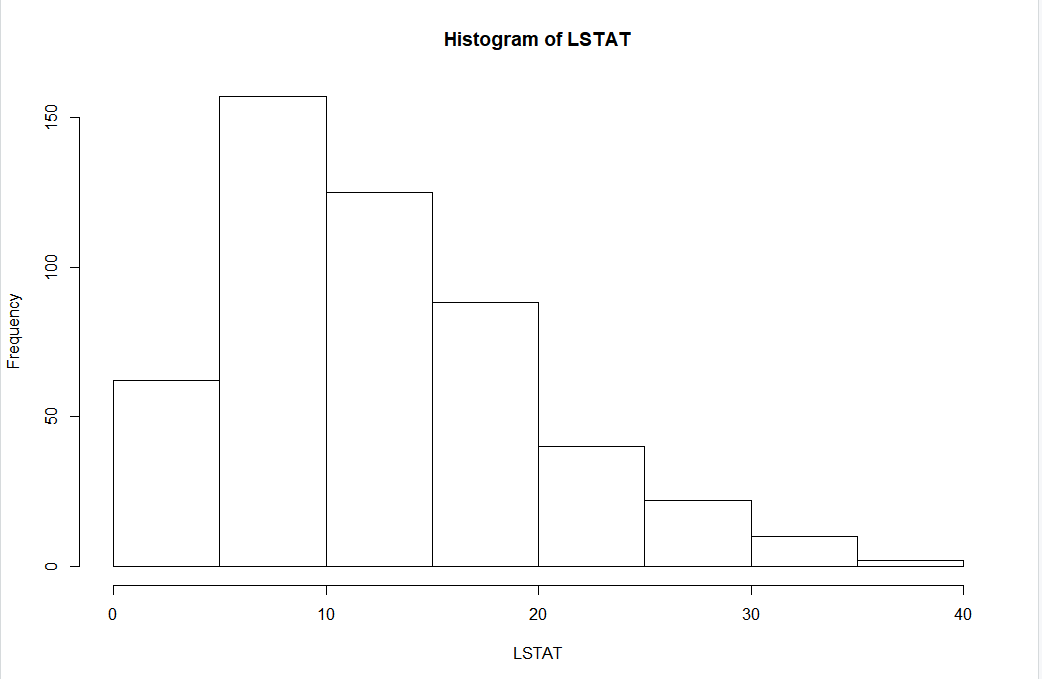


The results of this scatter plot is more spread and linear in comparison to the others. This shows that if most cases high status population owns house of high value which drops off propotionaly with population status with lower status and causing a drease in home value.

# Question 8

To display a histogram of the LSTAT variable the code is:

hist(LSTAT)

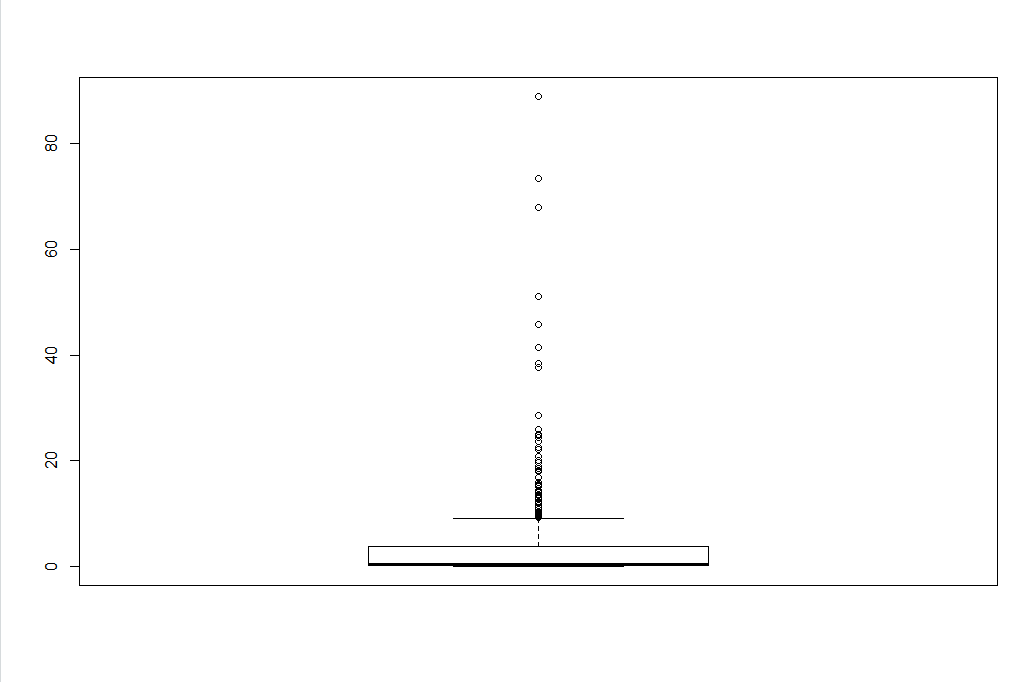


This histogram shows that the majority of population status resides between 5 and 10 having a frequency of just over 150 and then slowly declines as the lower status rises down to a frequency of around 5. The 0-5 status is and stands out in comparison to the rest having frequency of a little over 50.

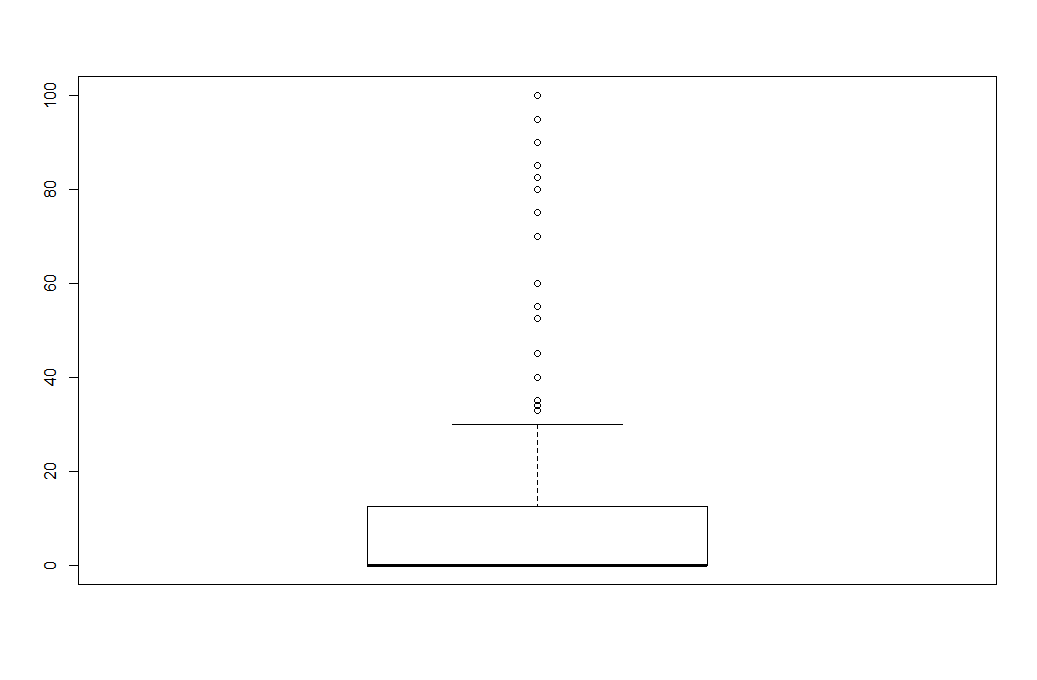
# Question 9

The boxplot is found through the code:

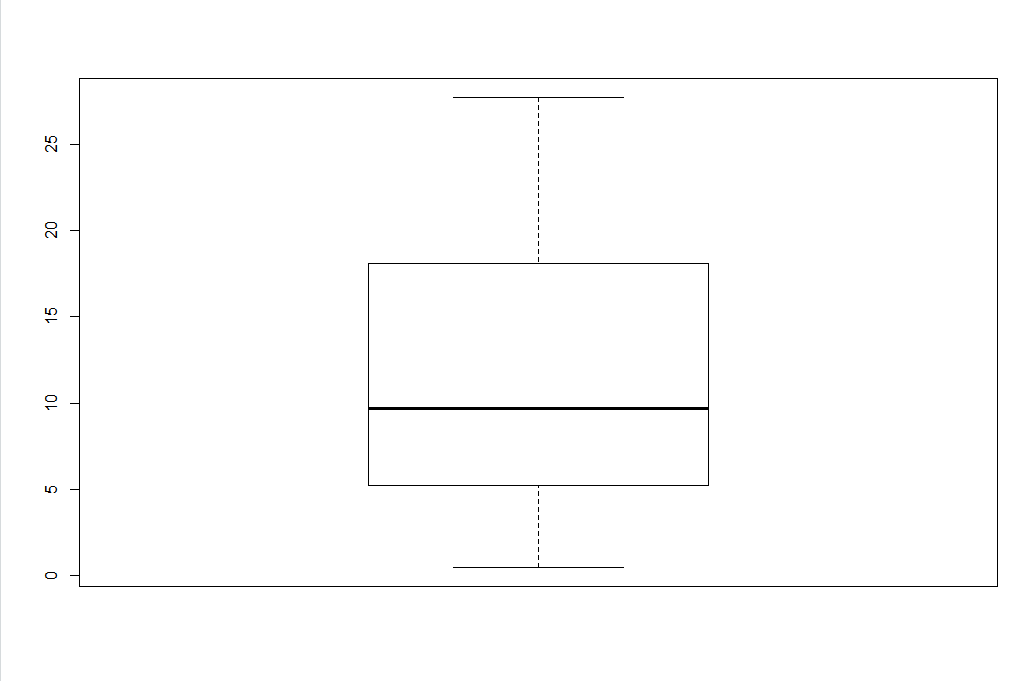
boxplot(x) where ‘x’ is the variable name e.g. boxplot(CRIM)

The CRIM boxplot

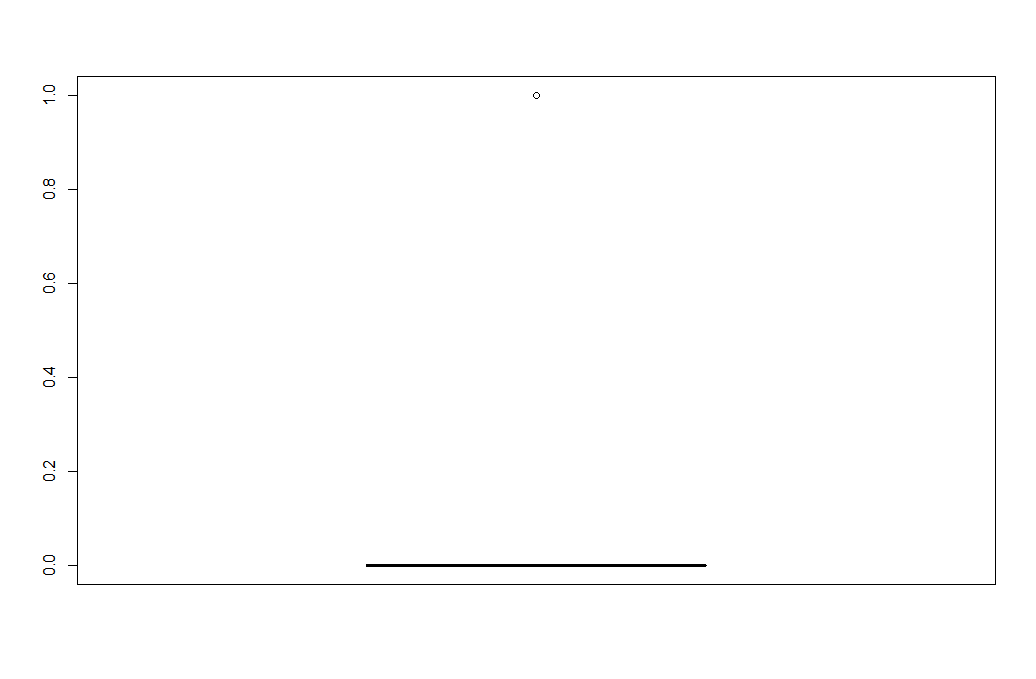
The ZN boxplot



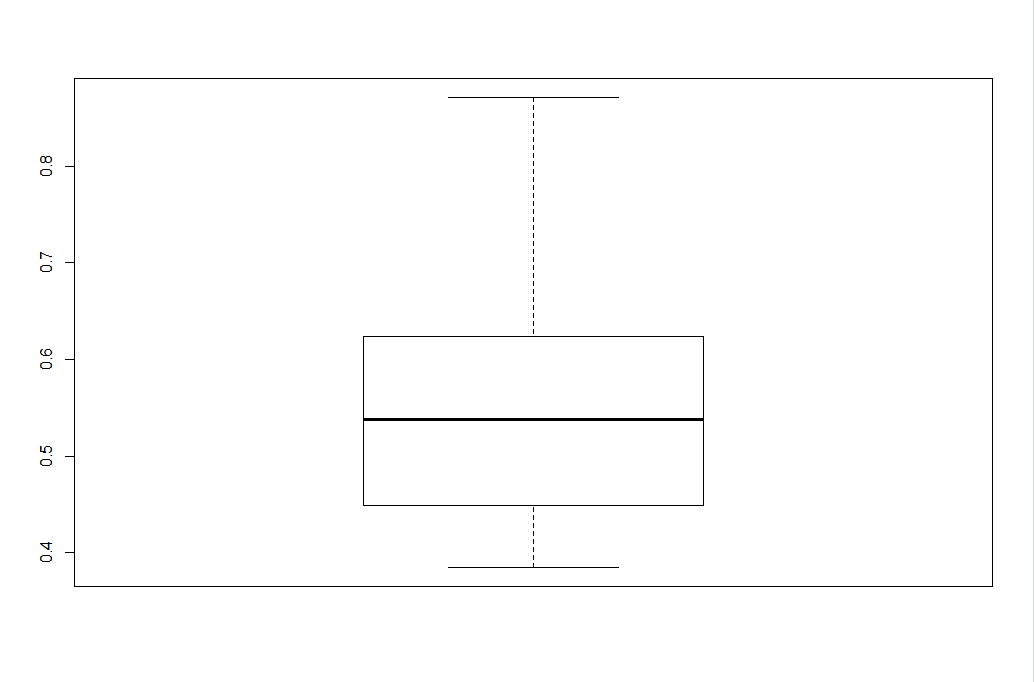
The INDUS boxplot



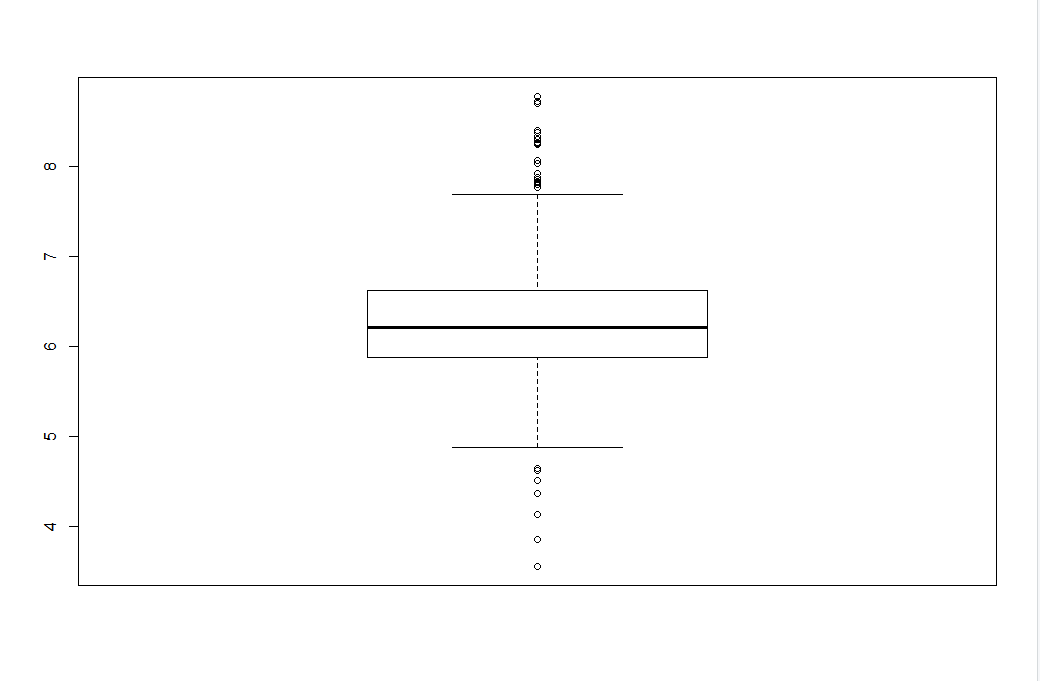
The CHAS boxplot



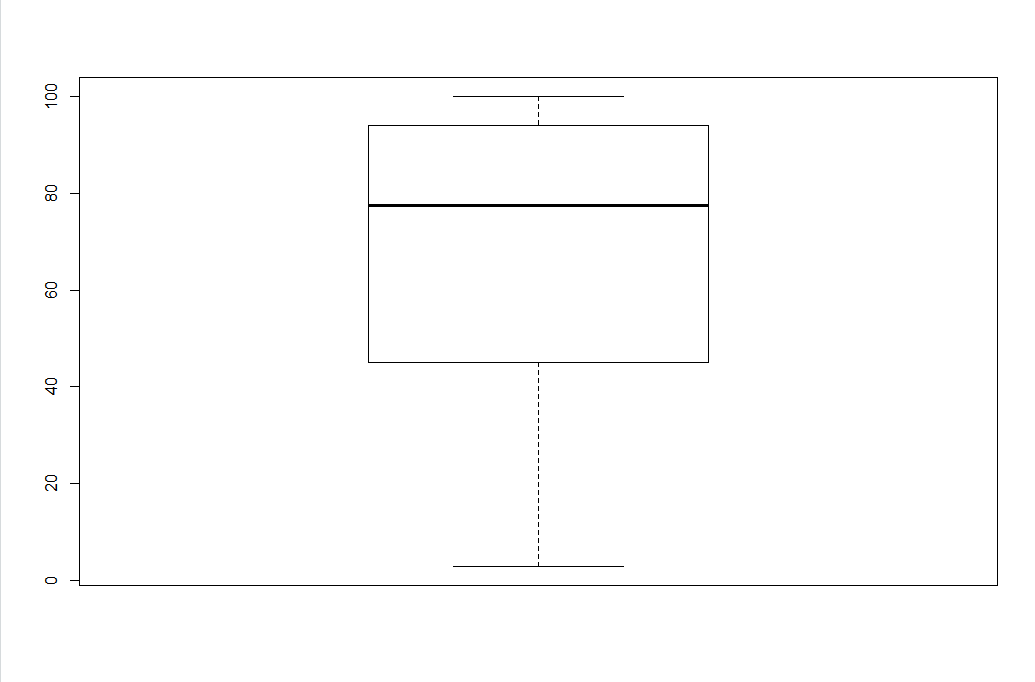
The NOX boxplot



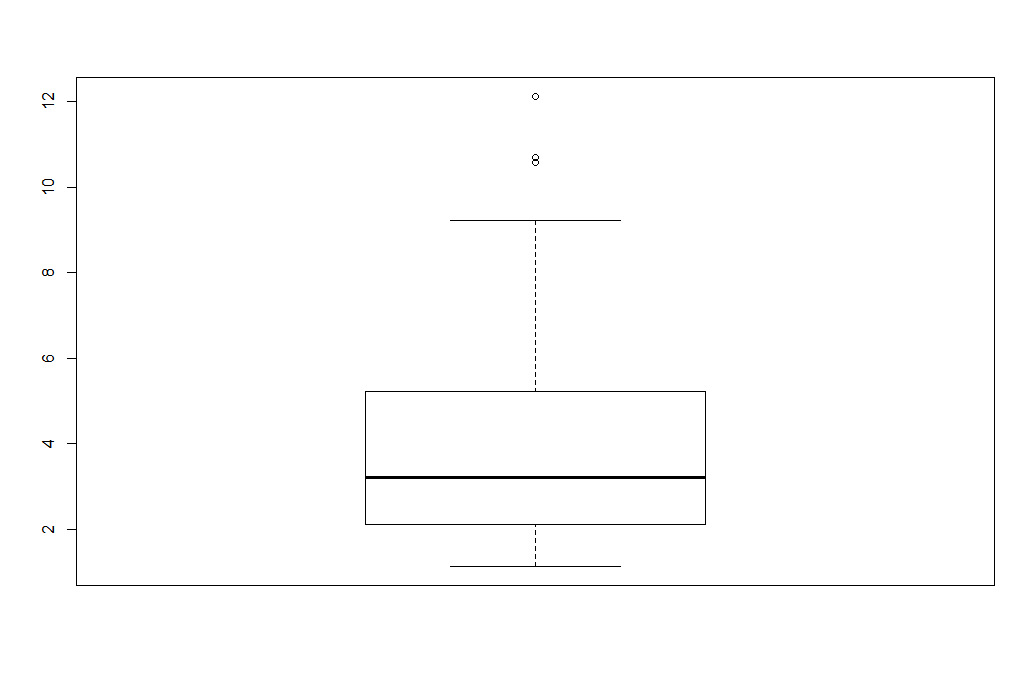
The RM boxplot



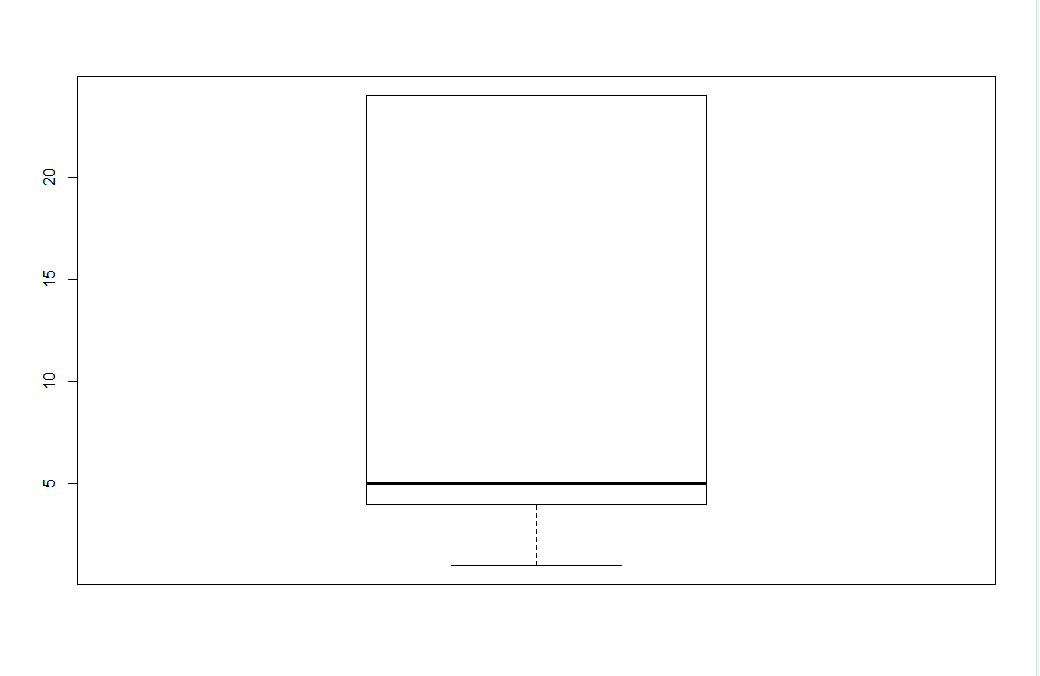
The AGE boxplot



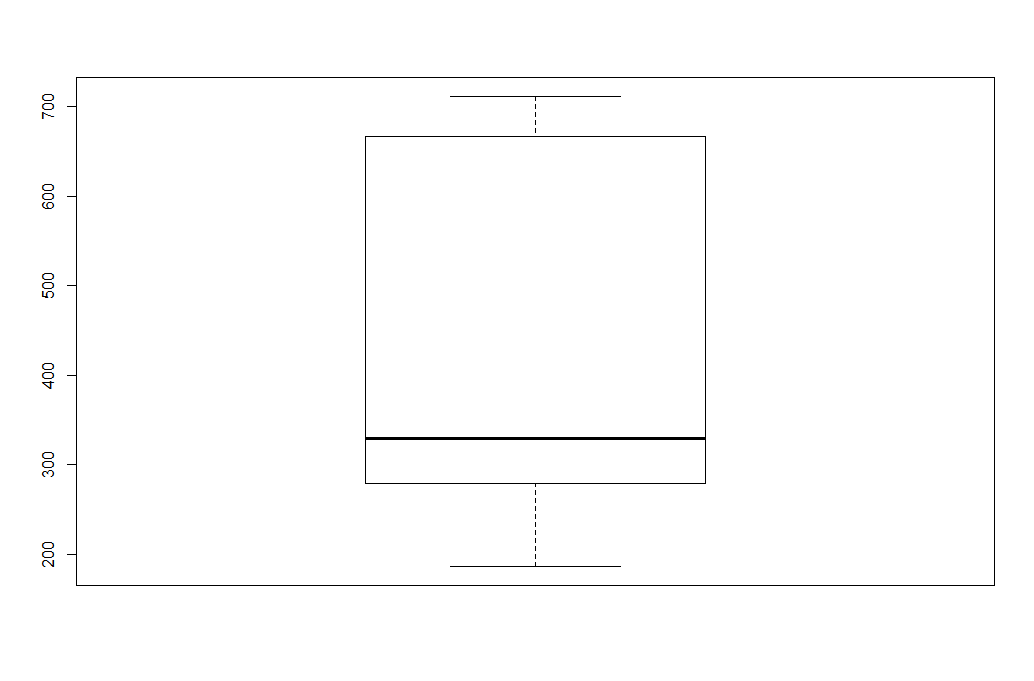
The DIS boxplot



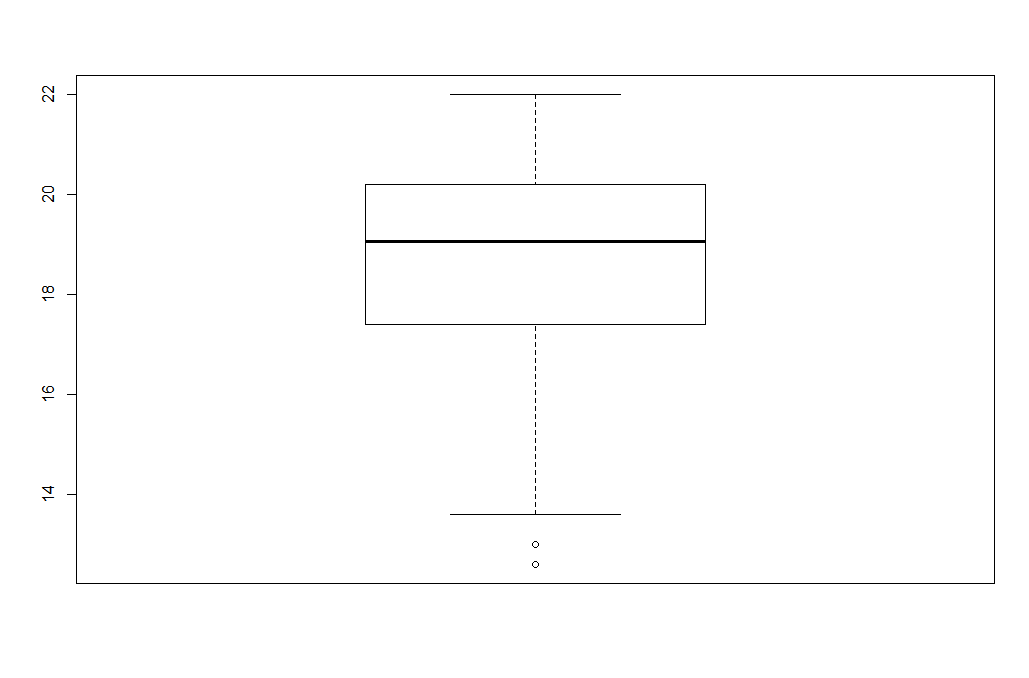
The RAD boxplot



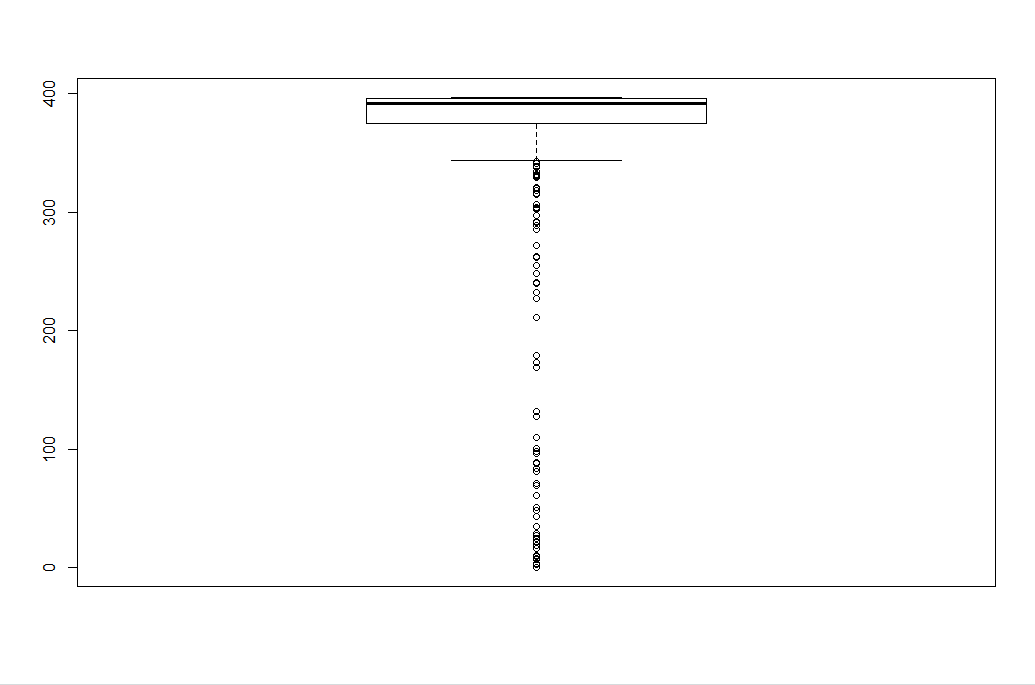
The TAX boxplot



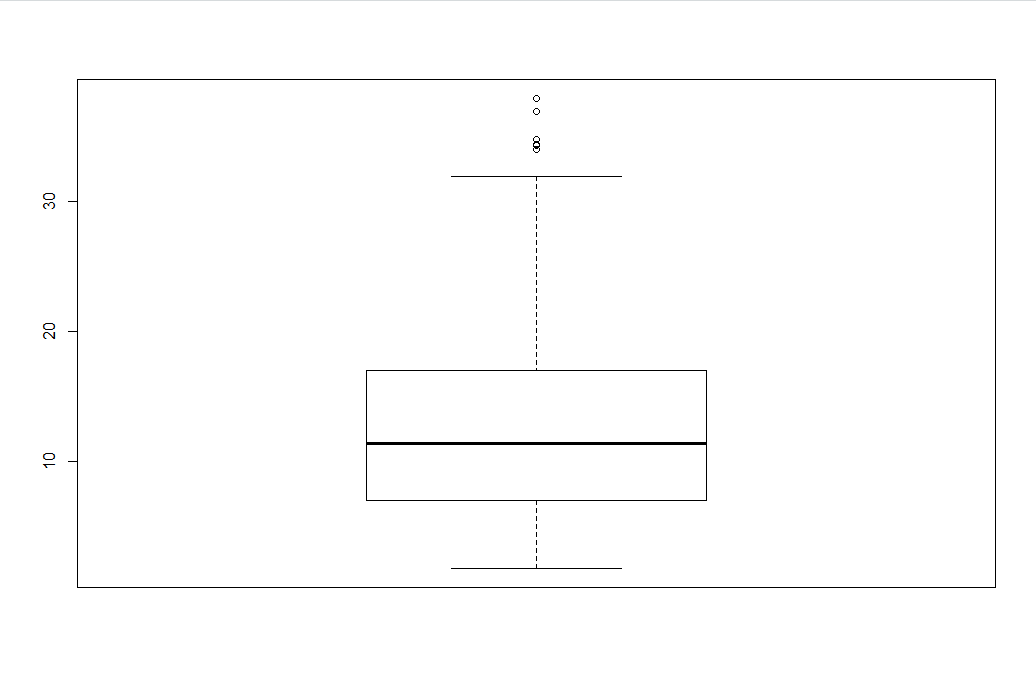
The PTRATIO boxplot



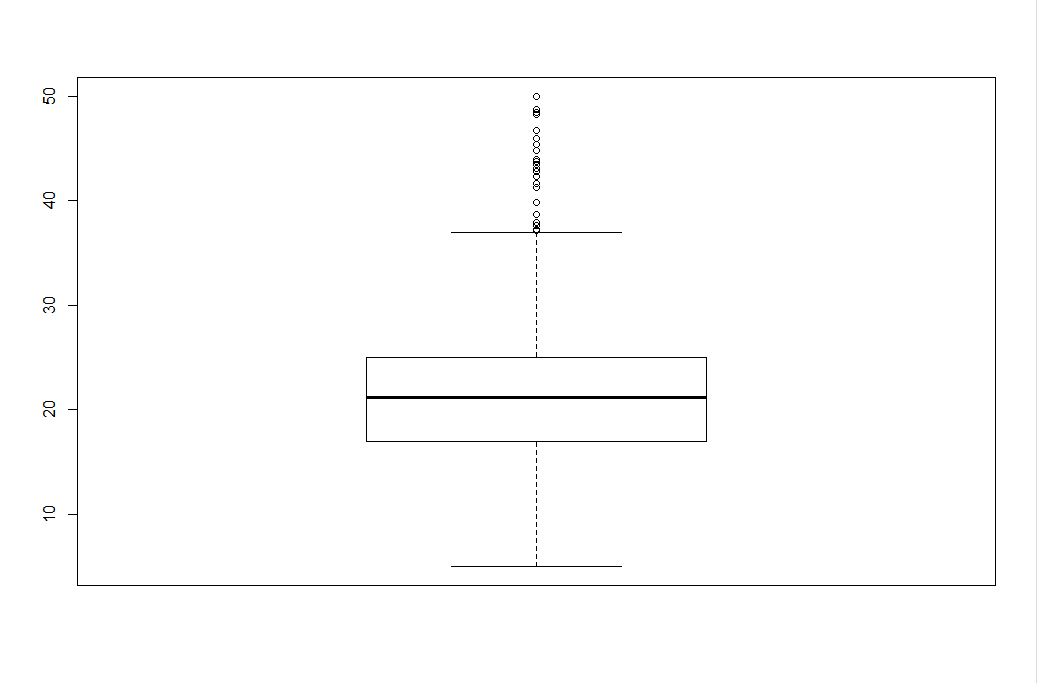
The B boxplot



The LSTAT boxplot



The MEDV boxplot



The boxplots above show that the variables with outliers are CRIM, ZN, CHAS, RM, DIS, PTRATIO, B, LSTAT, MEDV. The PTRATIO, DIS, LSTAT and CHAS variables only have a couple of outliers in comparison to the rest which have a large number. The CHAS variable only has 2 data being either 0 or 1, because most of the data is 0 this causes all the 1s to be outliers.

# Question 10

The results of analysis show the relationship between each of the variables and how they correlate and scale with each other such as scatterplots that allowed for an accurate interpretation of the relationship between certain variables. The numerous instances provided with the data allowed for more accurate analysis and comparisons within the data, with allowed for clustering to spawn within the scatter plots and for more accurate boxplots. This further led to a better understanding of what the information offered.