# **Drills With R On Importing And Plotting Data, And Finding The Distribution Measures**

Suraj Eswaran

University Of The Cumberlands

2024 Spring – Statistics For Data Science (MSDS – 531 – A01) – First Bi-Term

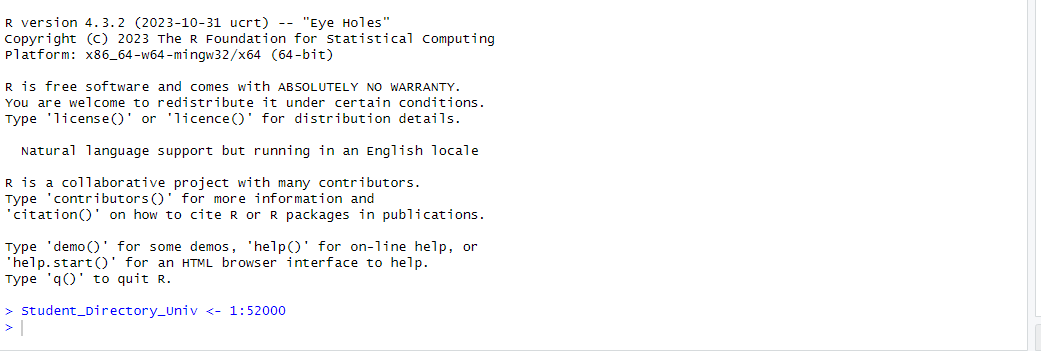
Professor Dr. Danny T. Barnes

01/14/2024

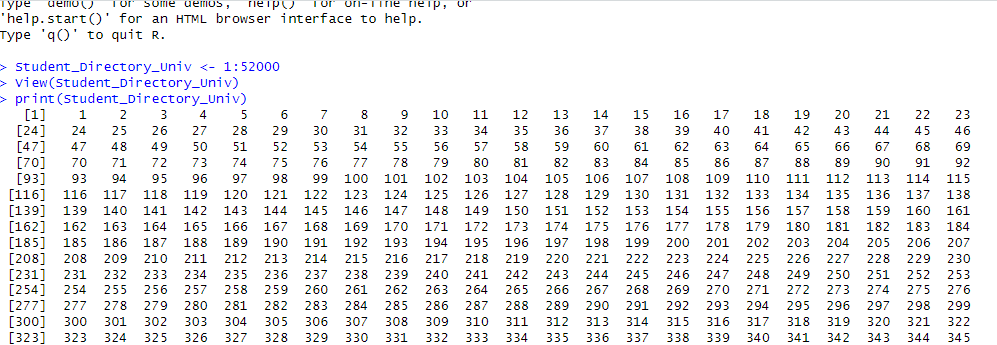
## Provide in the plain text R commands that finds/solves the following:

**The student directory for a large university has 400 pages with 130 names per page, a total of 52,000 names. Using software, show how to select a simple random sample of 10 names.**

We are given a student directory for a large university where it has total of 52,000 names in 400 pages containing 130 names per pages. So, let’s create a student directory which consists of 52,000 records.



In order to check whether it has stored, I have used “view” and “print” (Thulin, 2021, p. 38) command where you get to see values from 1 to 52000 is stored into Student\_Directory\_Univ directory.



As the main objective of this is to generate 10 random numbers, I will be utilizing the function “sample” (Thulin, 2021, p. 184) where it brings out the sample from the given data. Here our Student\_Directory\_Univ is the data that we have, and so I have used sample function with size as 10.



But here if we just sample function, it will regenerate random number, but we won’t storing those numbers which have been generated. For maintaining consistency throughout the runs, we can store those random numbers into another directory where each time we use it, will be showing the same list of values.



From environment panel in RStudio, we are able to validate that collected samples random\_ten\_samples are listing those same numbers as we had generated it. This property can be helpful in university, as they can use this directory for identifying their selected list. An example where this property can be helpful is where a professor selecting ten random different students from his/her class to perform a presentation, the professor would select random list from this and in order to keep it in track for that generated list, they can store it in another directory.



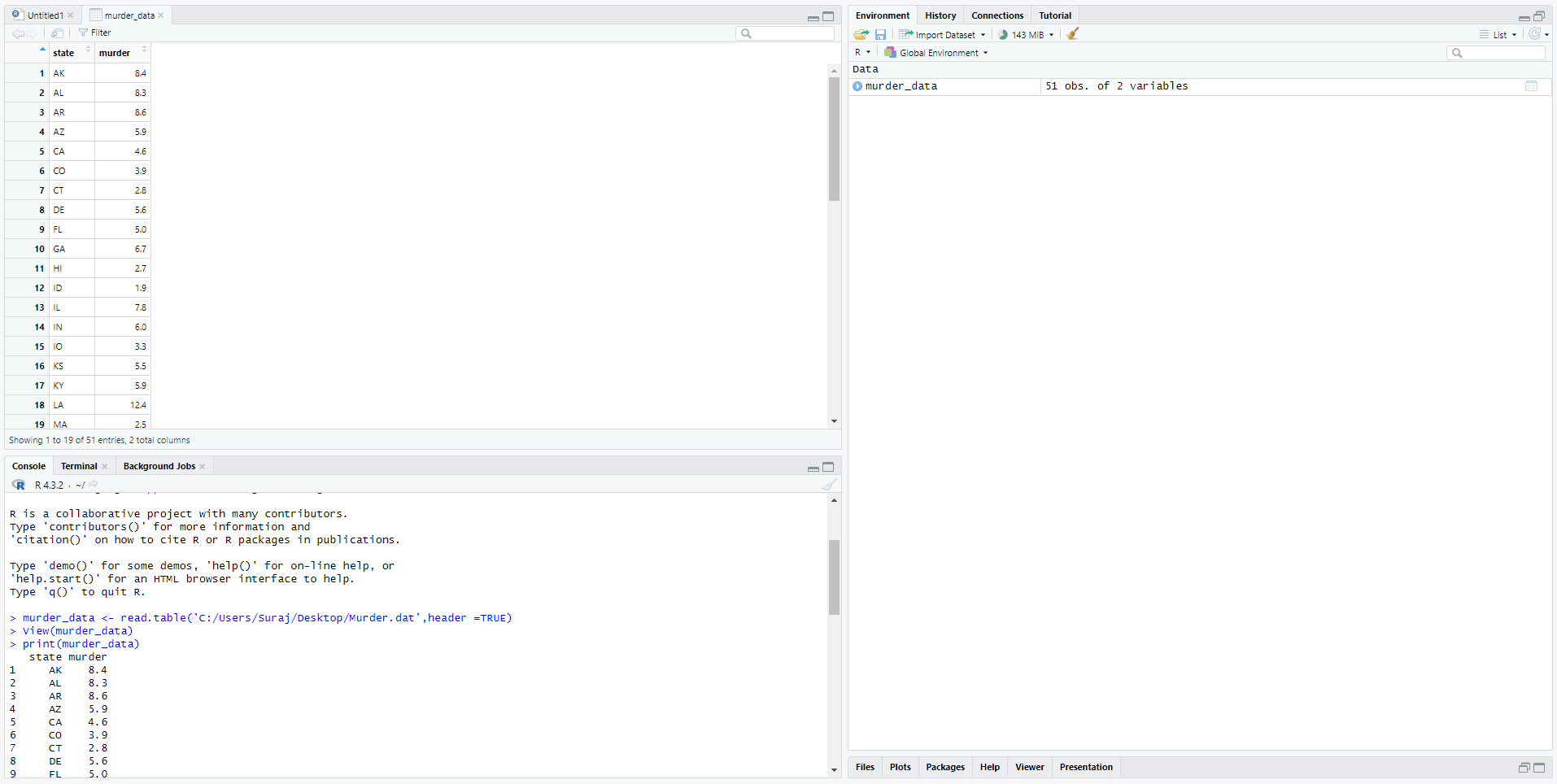
**From the Murder data file, use the variable murder, which is the murder rate (per 100,000 population) for each state in the U.S. in 2017 according to the FBI Uniform Crime Reports. At first, do not use the observation for D.C. (DC). Using software:**

**Find the mean and standard deviation and interpret their values.**

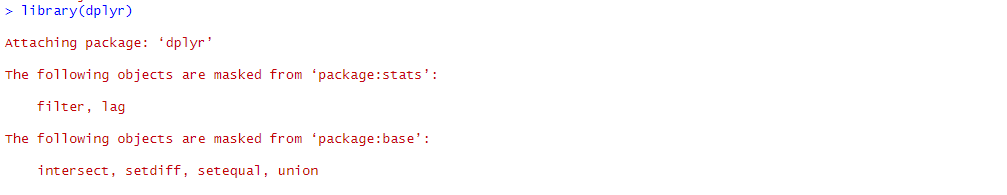
For this, I have used read.table function where I need to read Murder data that has State and Murder rate variable as headers in a tabular form. Since we have a header in our data file, we need to specify header as “TRUE” (Thulin, 2021, p. 120).

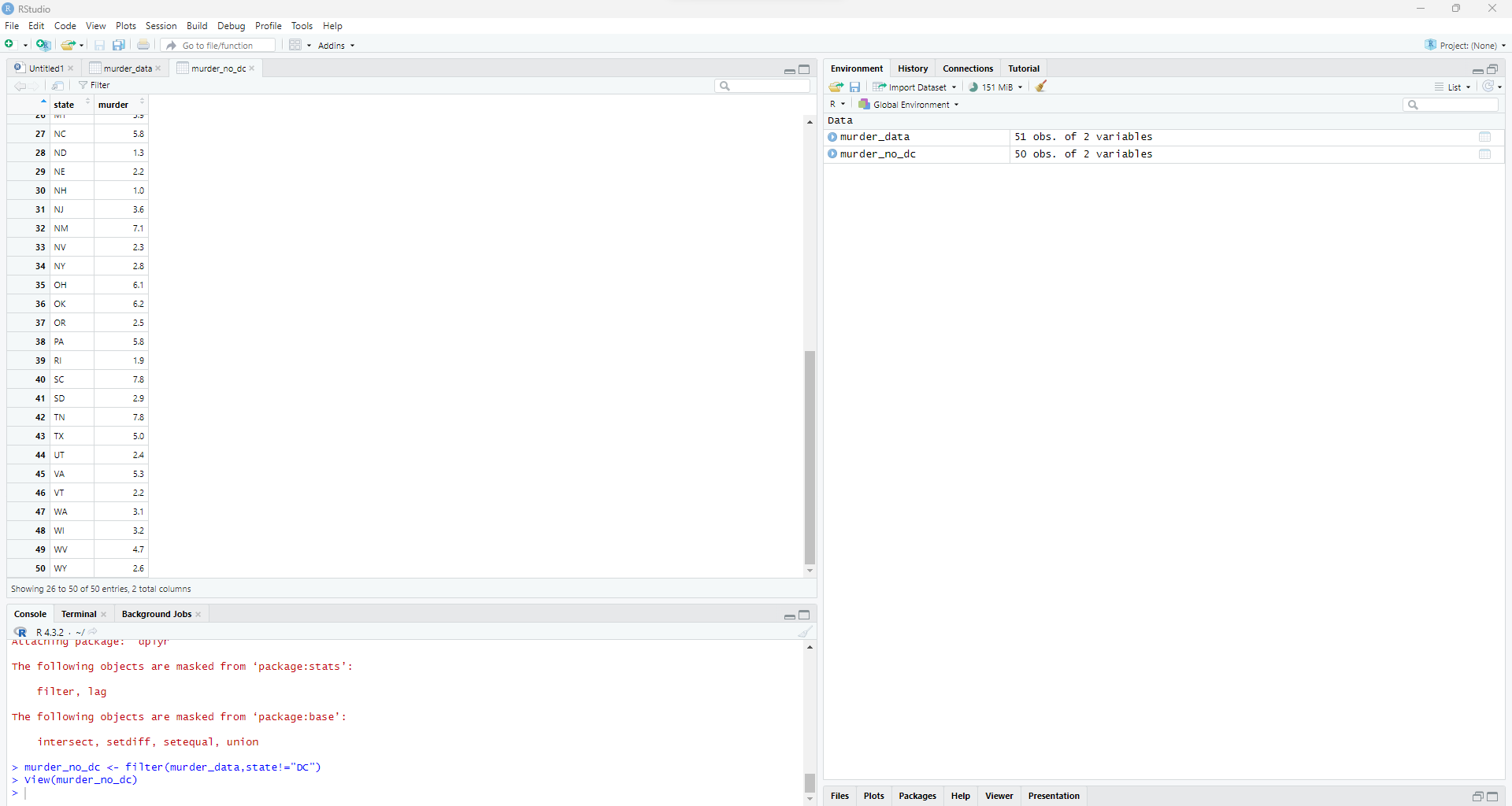


Validating the directory murder\_data, we will use “view” and “print” (Thulin, 2021, p. 38). where we will get to see data for various states with their murder rates.

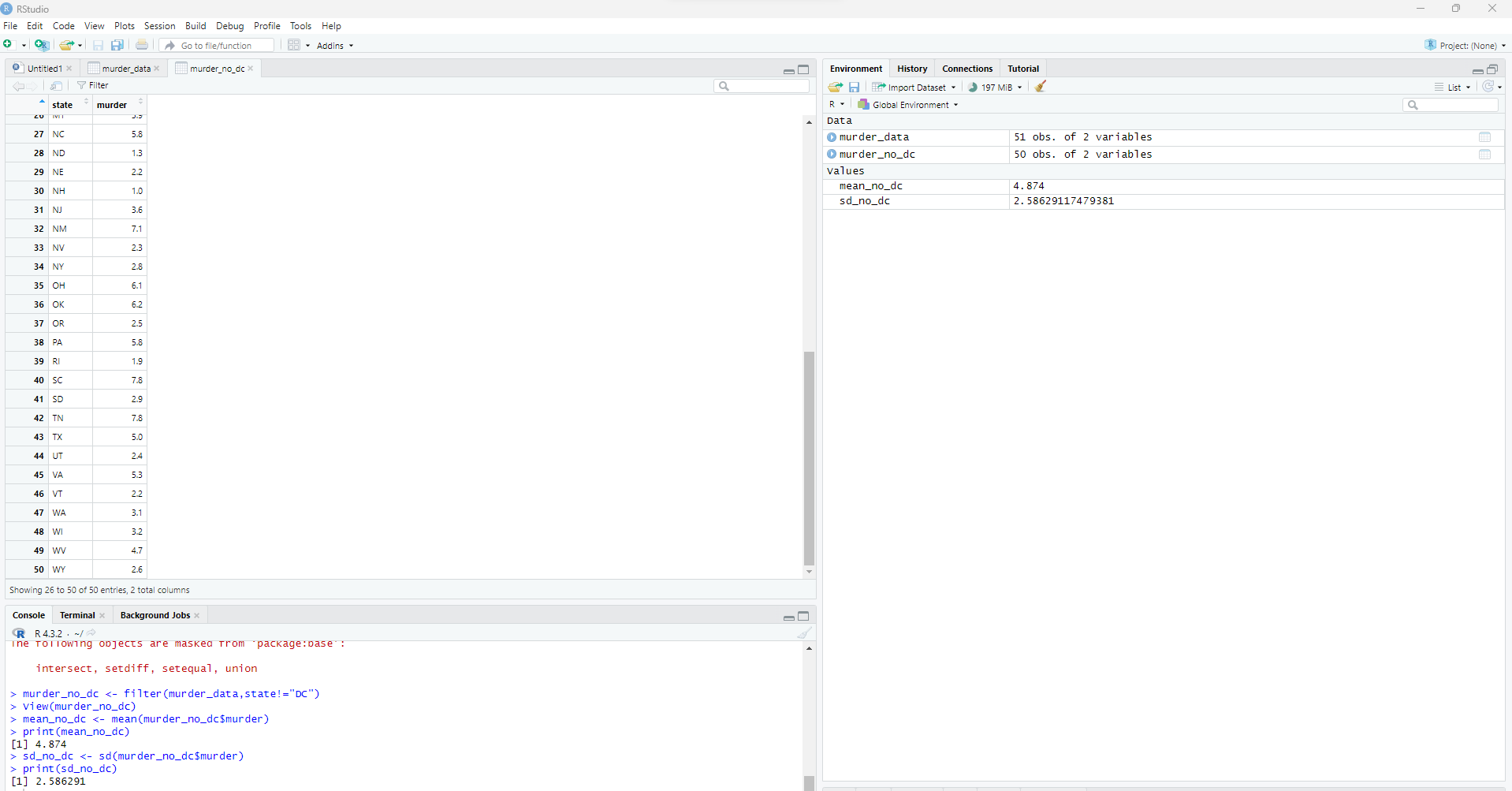


For omitting the data of “DC”, we will create another directory where we are using “filter” function, where we will add all the data except data from “DC”. In order to use that function, we need to use “dplyr” as it is helps in cleaning and filtering unstructured data. As I have installed it in our system, we will activate the library “dplyr”, we will first enter “library(dplyr)” (Thulin, 2021, p. 114), so that we can use filter option.



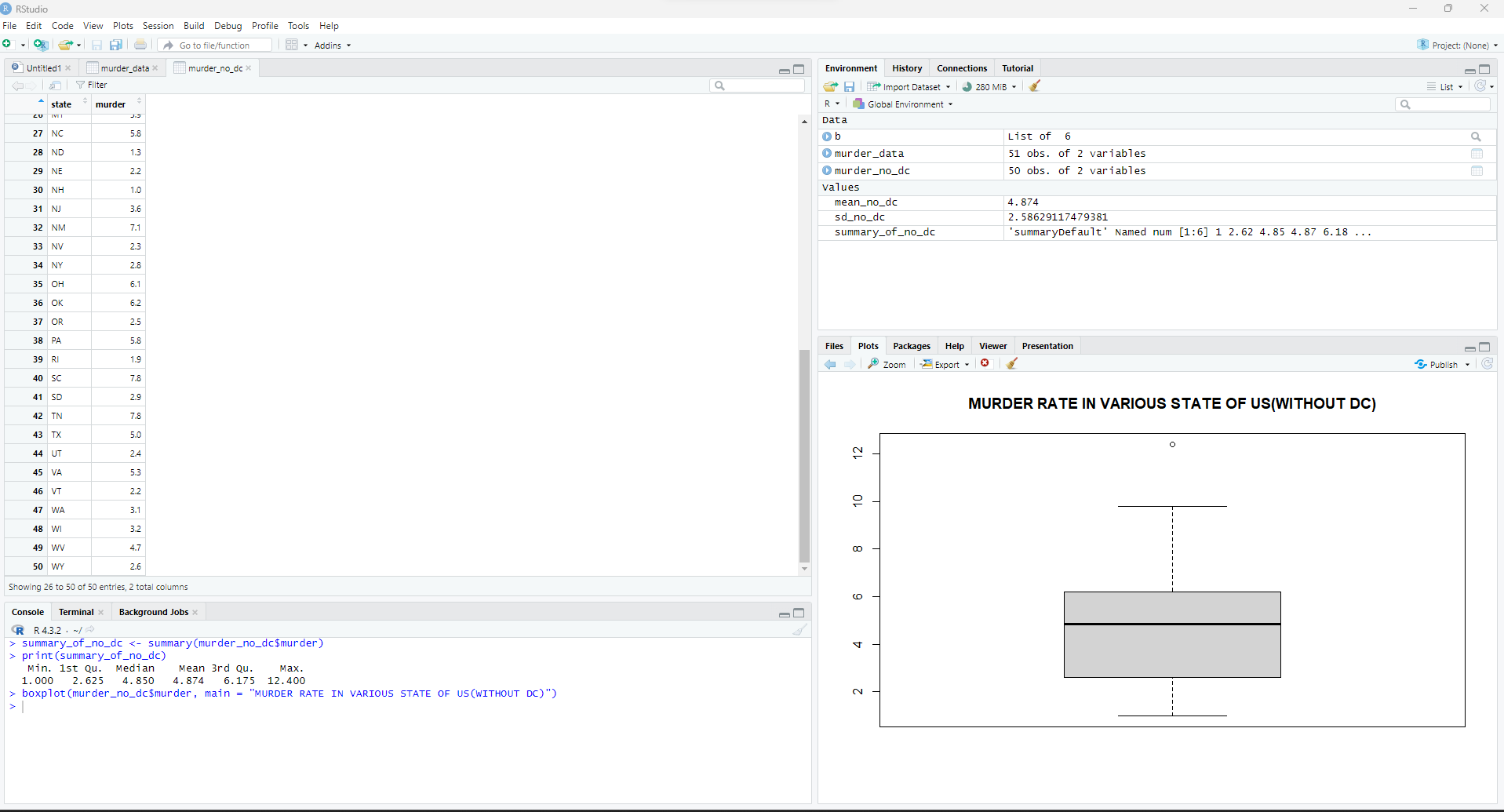
Using filter function(Thulin, 2021, p. 182), we have omitted the row “DC” and now calculating the mean and standard deviation of that directory. 

Mean can be calculated by using “mean” (Thulin, 2021, p. 33) function and standard function by “sd” function(Thulin, 2021, p. 40). Calculating those, we have obtained a value of 4.874 as mean and 2.586291 as standard deviation.



**Find the five-number summary and construct the corresponding boxplot.**

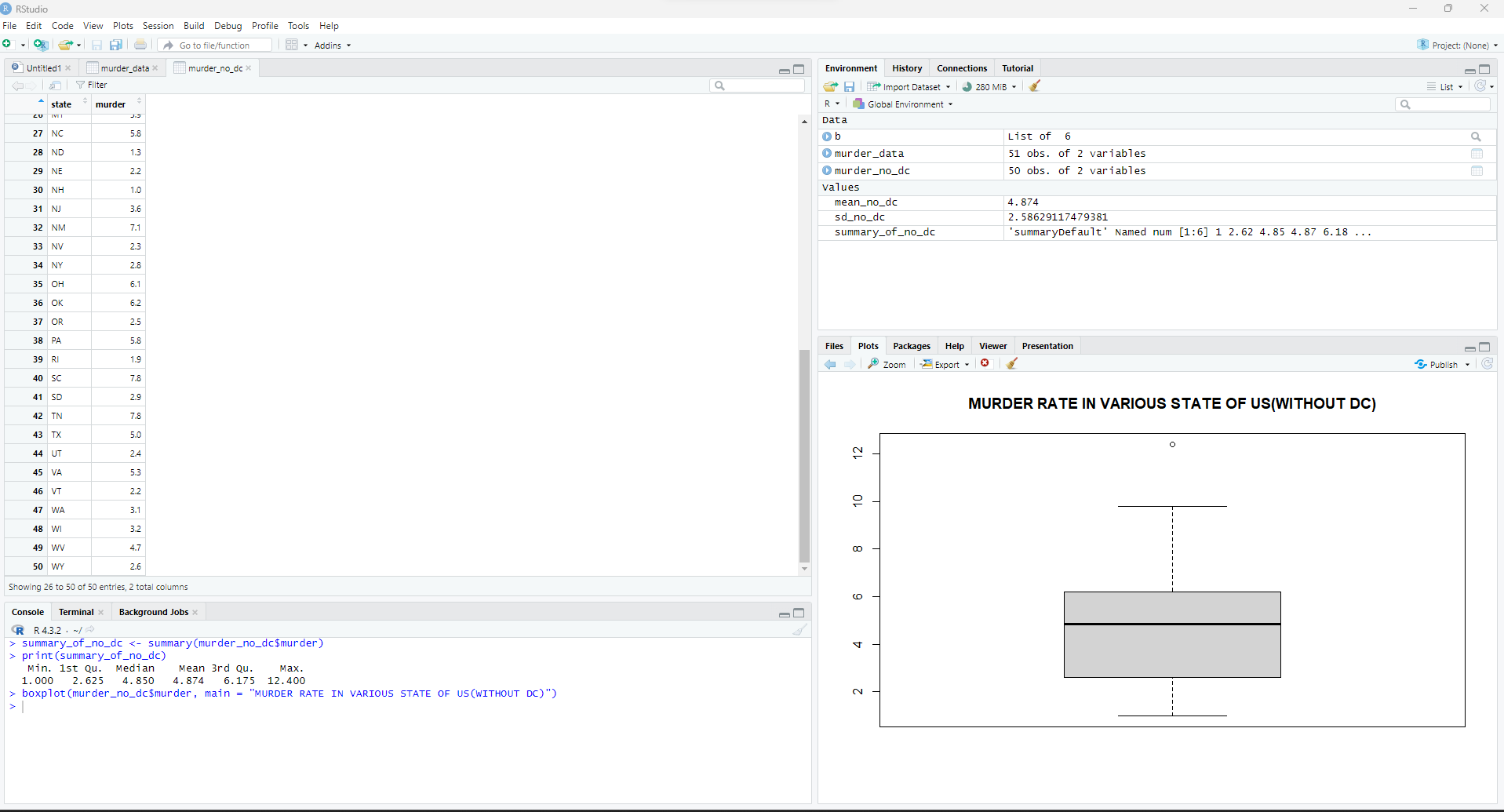
For interpretation, we are using box plot for graphical representation with the help of “boxplot” function (Thulin, 2021, p. 48) and summarize the data that we have.



From the graph and summary function, we have obtained the horizontal line(i.e.) median as 4.850 which means that 50 percent of the states below median have less murder rate while the rest of the 50 percent have higher murder rate than the median. Our minimum value is 1.00 and maximum as 12.40 with first quartile as 2.626 and third quartile as 6.175.

**Now include the observation for D.C. What is affected more by this outlier: The mean or the median?**

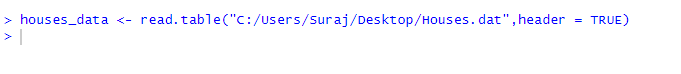
Now we will consider all the murder rate including the DC record and calculate the summary with boxplot.



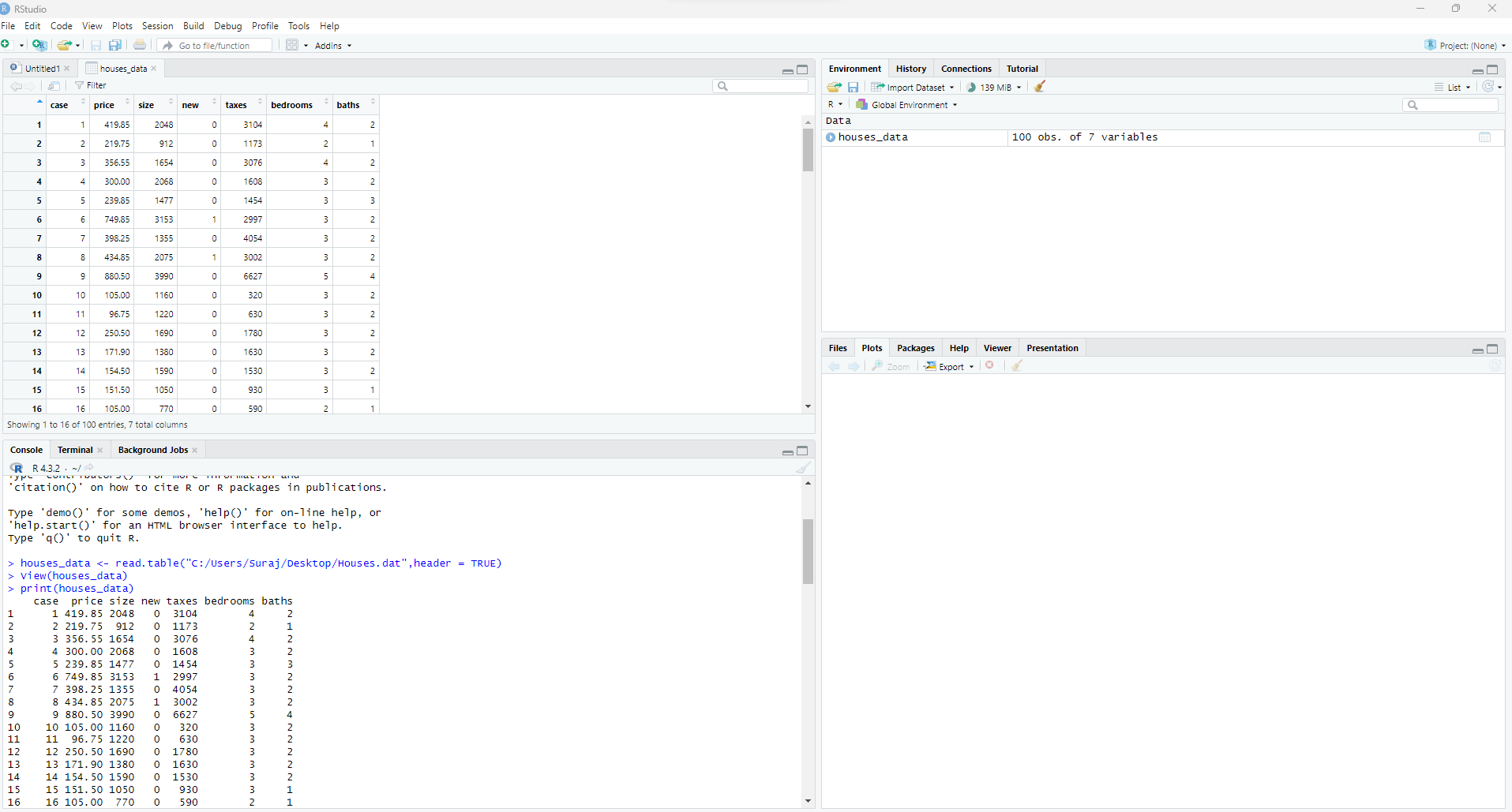
Combining with DC records, we have obtained mean of 5.252941, standard deviation of 3.725391 and median as 5.000. Comparing both the ways, adding DC records, it impacted the mean more than median although both the values have increased. As adding a higher outlier, the range between them will increase significantly followed by huge increase of mean with little increase of median with no difference on mode.

**The Houses data file lists the selling price (thousands of dollars), size (square feet), tax bill (dollars), number of bathrooms, number of bedrooms, and whether the house is new (1 = yes,0 = no) for 100 home sales in Gainesville, Florida. Let’s analyze the selling prices.**

For this scenario, we have utilized read.table function in order to read House data that has Case, Price, Size, New, Taxes, Bedrooms and Baths variable as headers in a tabular form. Since we have a header in our data file, we need to specify header as “TRUE” (Thulin, 2021, p. 120).



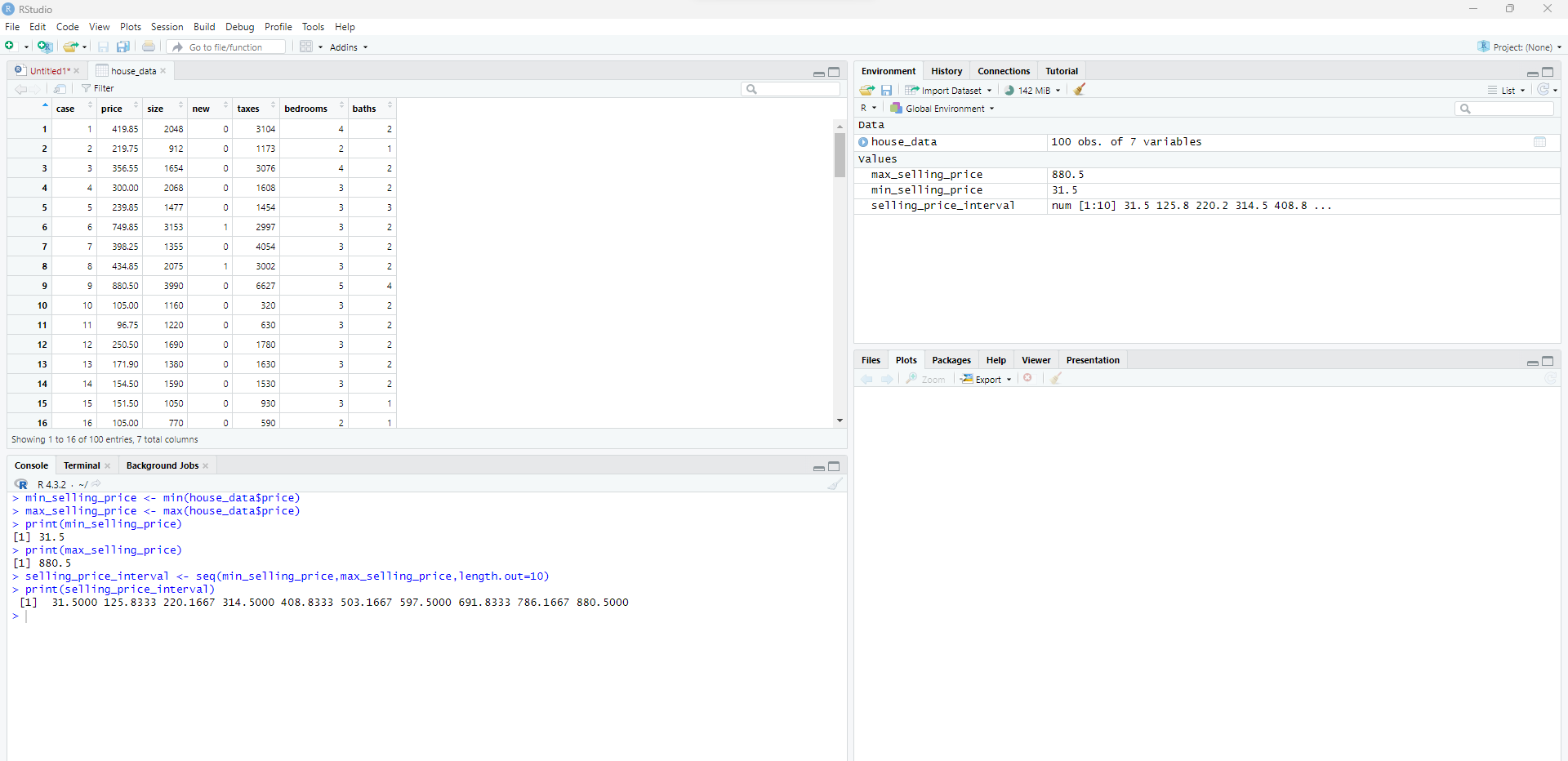
Using view and print option for seeing the house data with various prices and sizes of all kinds of houses(Thulin, 2021, p. 38).



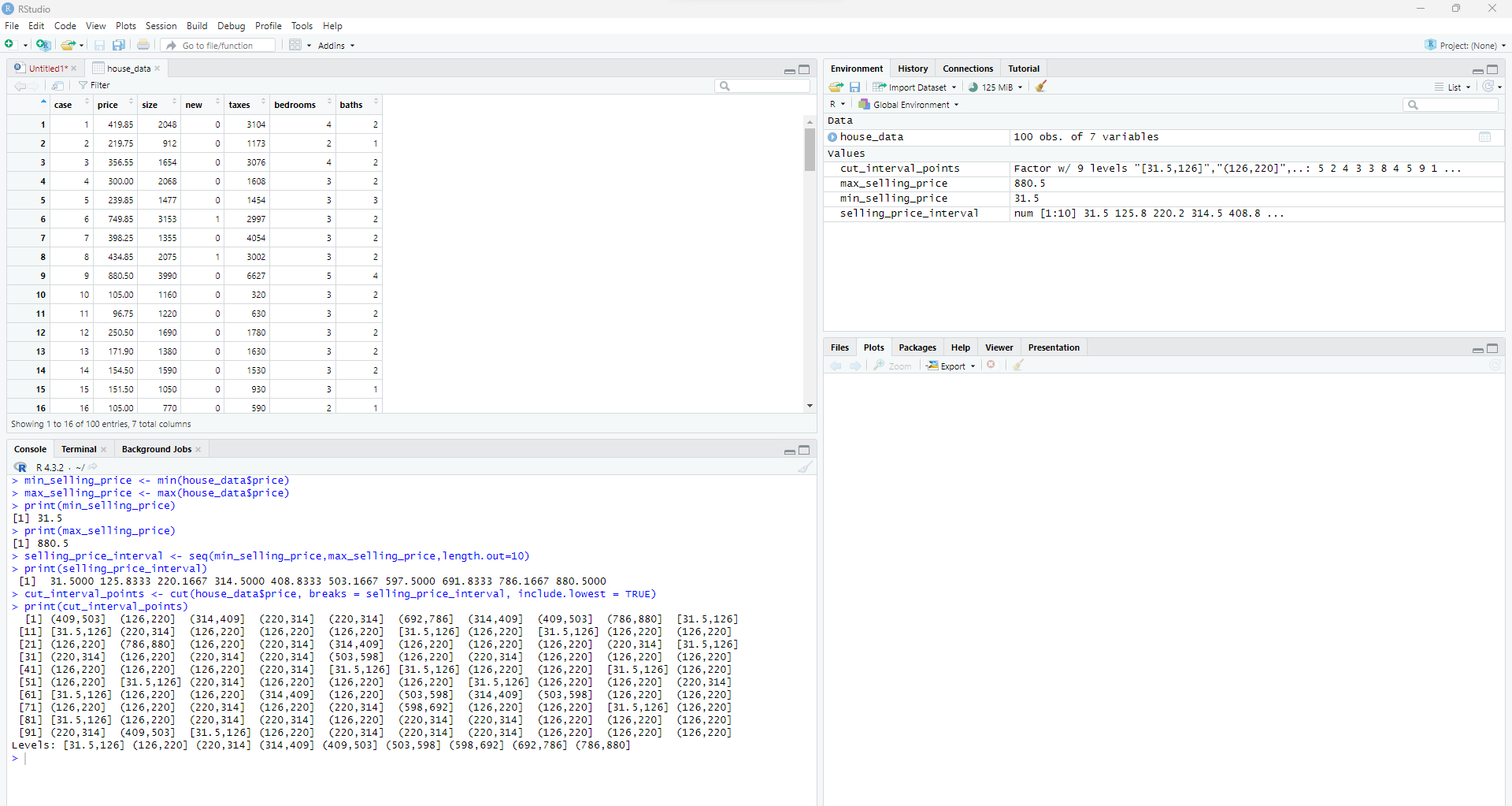
**Construct a frequency distribution and a histogram.**

As there are about 100 records and if we try to impose frequency distribution table with histogram, the table and graph might be clumsy. In order to have a proper representation of those data, we will set frequency table based on intervals and plot that in a histogram graph.

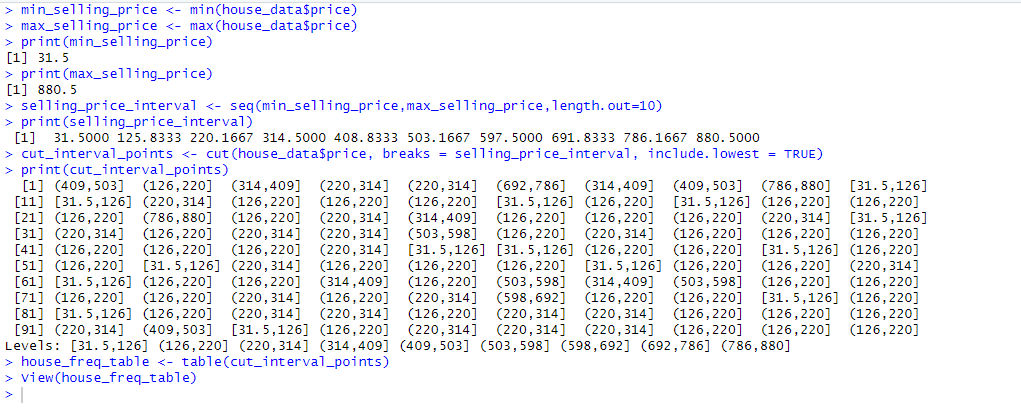
For creating intervals for frequency distribution, we can utilize a function named “seq” where we will put the lowest and highest selling price value with number of break that is required for our table (Thulin, 2021, p. 172). For this scenario, let’s use the total length of the sequence to be 10, as it might provide us equal intervals. Using “min” and “max” function(Thulin, 2021, p. 40), we have found out that lowest and highest value is 31.5 and 880.5 respectively. Using length.out as 10 with lowest and highest value (Thulin, 2021, p. 224), the interval value can be calculated as followed:

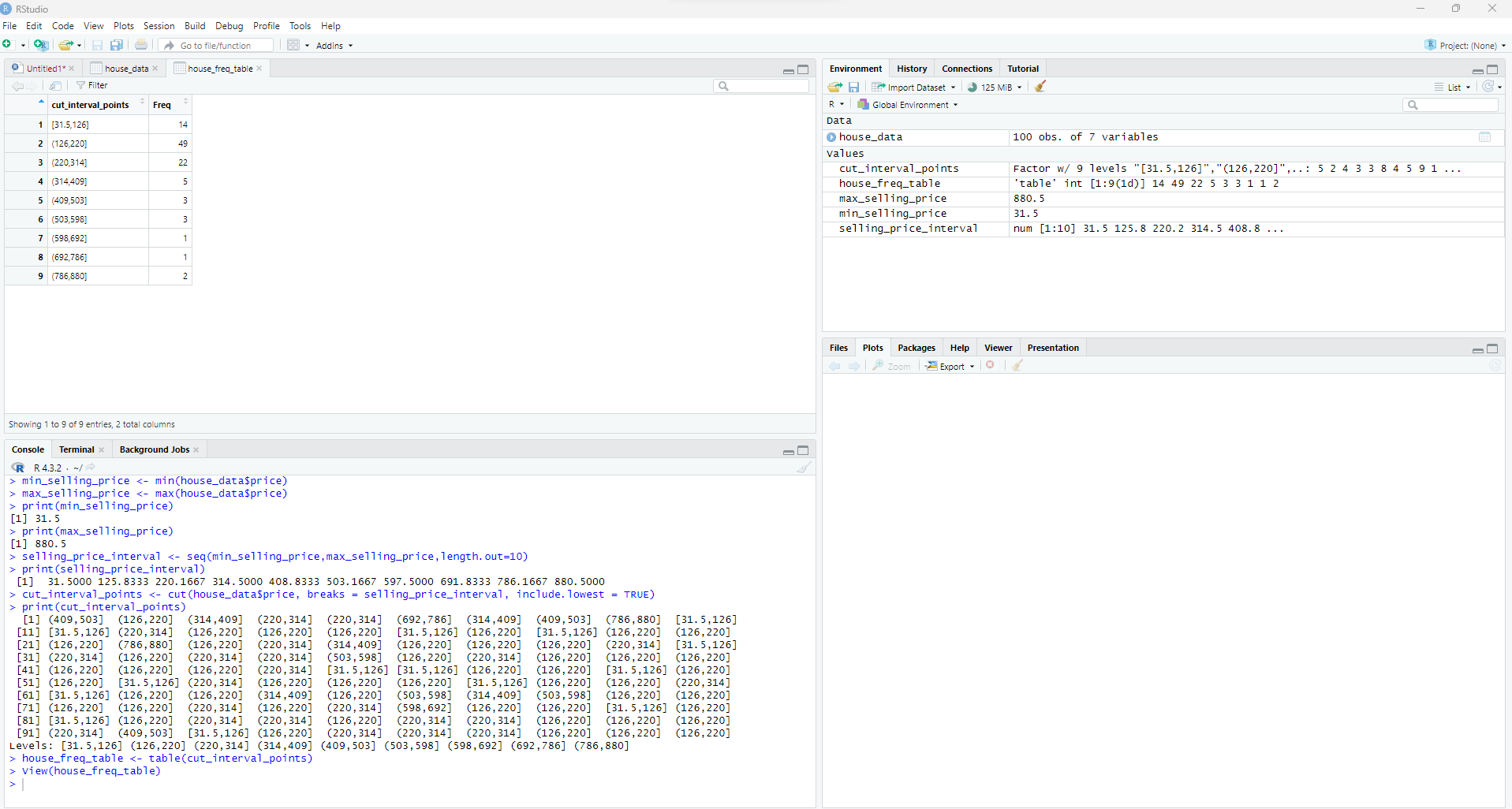


In order to group those values into our interval range, we will use “cut” function where it will categories values based on the intervals(buckets) formed earlier using “seq” function with selling\_price\_interval as a breaking point for the table, include.lowest to be mentioned as it also includes the least interval point.

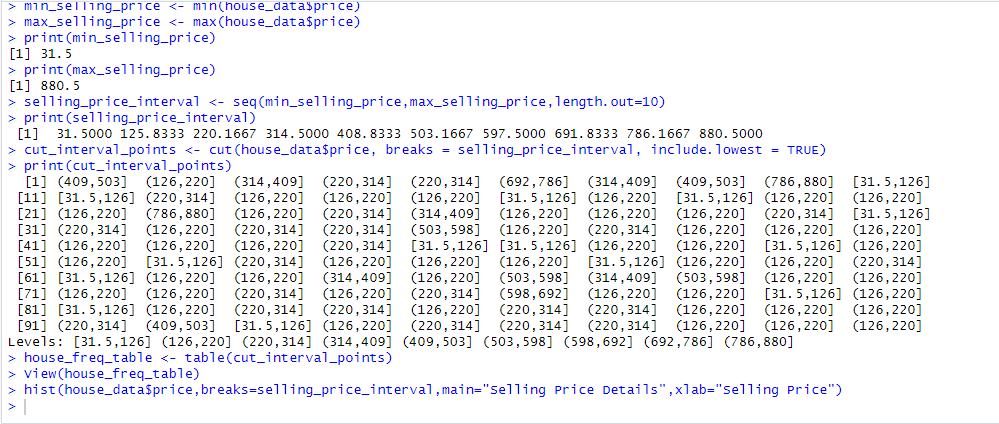


Finally, to create frequency distribution table, the method “table” is used where it computes different frequencies for all the selling prices listed in the house data.





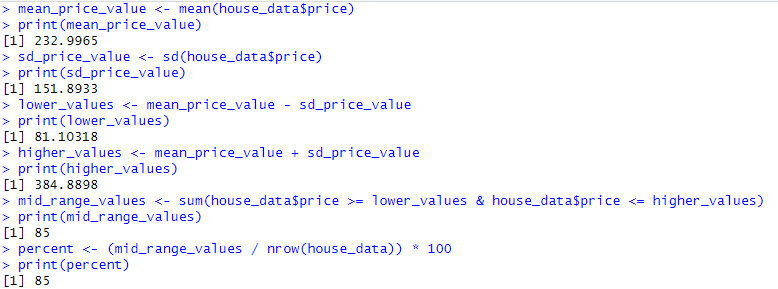
Plotting histogram using “hist” method with breaks as selling\_price\_interval, x-axis to be selling prices from house data and y-axis as frequency(Thulin, 2021, p. 50).





**Find the percentage of observations that fall within one standard deviation of the mean.**

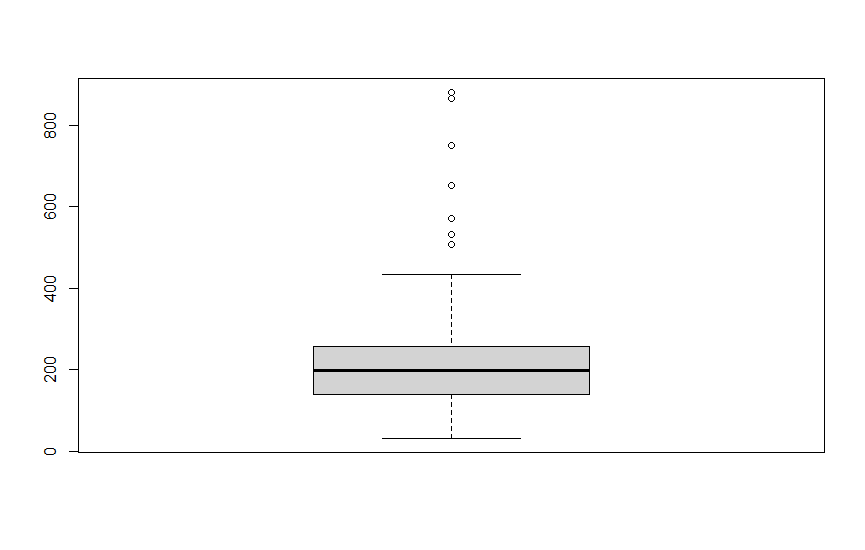
Using mean and standard deviation method, we have found out that obtained mean and standard deviation value are 232.9965 and 151.8933 respectively. Here, the main objective is to find out those values which fall into category of standard deviation of the mean. For this, we need to know lower range value and higher range value where we are subtracting mean with standard deviation for lower range value and adding mean with standard deviation for higher range value. Then adding all those price values which are greater than or equal to lower range and lesser or equal to higher range value in order to find out all the value which are within the range of the standard deviation. Thus, we are finding percentage by dividing the total number of values that are within the range with total number of records (Total =100) and multiplying with 100. By this calculation, we have found out that there is 85 percent of observations that are falling within one standard deviation of the mean.



**Construct a boxplot.**

For interpretation, we are using box plot for graphical representation with the help of “boxplot” function and summarize the data that we have.





**References**

Thulin, M. (2021). Modern Statistics with R: From wrangling and exploring data to inference and predictive modelling. BoD-Books on Demand.