Student: Kabirat Nasiru

UC student ID number: 005018391

Course: 2024 Spring - Statistics for Data Science (MSDS-531-A01) - First Bi-term

Course Advisor: Dr. Danny Barnes

**Random Sampling**

In order to select a simple random sample of 10 names from the instruction provided, the below code has been derived to generate 10 numbers between 1:52000.

# Total number of names in the student directory is

total\_students\_names <- 52000

# Number of names to sample (simple random sample)

sample\_student\_size <- 10

# Generating a simple random sample with R

simple\_random\_sample <- sample(x=total\_students\_names, size = sample\_student\_size)

# Print the sampled names (numbers in this case)

print(simple\_random\_sample)

[1] 13508 32403 19856 4311 19146 24634 4906 40604 47080 15939.

The result was also printed.

**Murder Rate Analysis**

The first step was to download the data from the Index of data provided, and determine names for the variables, read and load the tables, then excluded the observation recorded for DC in the *U.S. Murder Rate in 2017 from the FBI Uniform Crime dataset*, calculated mean and standard deviation, found the five number summary, constructed a boxplot excluding DC data. After completing this steps, we were able to analyze the data as is, that is including DC input data, and e were ultimately able to complete the analysis by comparing mean and median.

> US\_murder\_data <- read.table ("C:/Users/kabir/OneDrive/Documents/Murder.dat.txt", header = TRUE)

> US\_murder\_without\_WDC <- subset(US\_murder\_data, state != "DC")

> mean\_US\_murder\_without\_WDC <- mean(US\_murder\_without\_WDC$murder)

> median\_US\_murder\_without\_WDC <- median(US\_murder\_data$murder)

> print(median\_US\_murder\_without\_WDC)

[1] 5

> print(mean\_US\_murder\_without\_WDC)

[1] 4.874

> standard\_deviation\_US\_murder\_without\_WDC <- sd(US\_murder\_without\_WDC$murder)

> print(standard\_deviation\_US\_murder\_without\_WDC)

[1] 2.586291

> five\_no\_summary\_without\_WDC <- summary(US\_murder\_without\_WDC$murder)

> print (five\_no\_summary\_without\_WDC)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 2.625 4.850 4.874 6.175 12.400

> boxplot(US\_murder\_without\_WDC$murder, main= "U.S. 2017 Murder Rates (Without Washington DC)", ylab = "Murder Rate (per 100,000 People)")

**Figure 1**

*U.S. Murder Rate in 2017 from the FBI Uniform Crime*A graph with a black line

Description automatically generated

> #Calculate with DC

> US\_murder\_with\_WDC <- rbind(US\_murder\_without\_WDC, US\_murder\_data[US\_murder\_data$state == "DC", ])

> mean\_US\_murder\_with\_WDC <- mean(US\_murder\_data$murder)

> print(mean\_US\_murder\_with\_WDC)

[1] 5.252941

> median\_US\_murder\_with\_WDC <- median(US\_murder\_data$murder)

> print(median\_US\_murder\_with\_WDC)

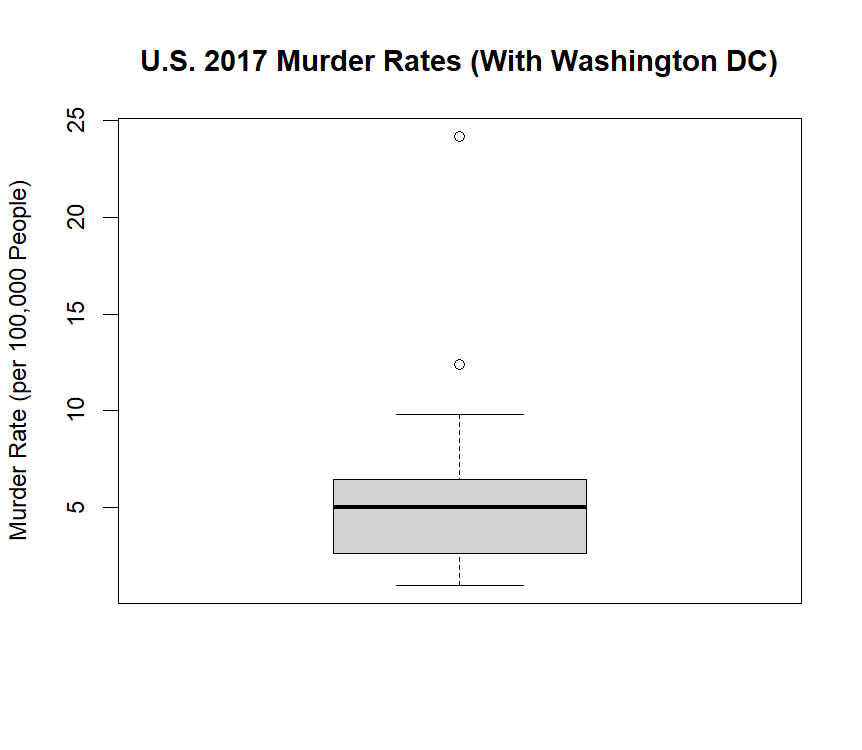
[1] 5

> boxplot(US\_murder\_with\_WDC$murder, main= "U.S. 2017 Murder Rates (With Washington DC)", ylab = "Murder Rate (per 100,000 People)")

The mean is the only distribution that is affected by the outlier (DC), without DC (4.874), with DC (5.2594).

**Figure 2**

*U.S. Murder Rate in 2017 from the FBI Uniform Crime*



**Housing Sale Analysis**

To carry out this analysis, the first step was to download the data from the Index of data provided, and determine names for the variables, read and load the tables, then proceed to analyses home sales for Gainesville Florida. Output includes Histogram and Boxplot.

Florida\_house\_data <- read.table ("C:/Users/kabir/OneDrive/Documents/Houses.dat.txt", header = TRUE)

> price\_range <- seq(min(Florida\_house\_data$price), max(Florida\_house\_data$price), length.out = 10)

> frequency\_table <- table(cut(Florida\_house\_data$price, breaks = price\_range, include.lowest = TRUE))

> hist(Florida\_house\_data$price, main = "Histogram of Florida House Selling Prices", xlab = "Selling Price (unit:thousands of Dollars)", ylab = "Frequency", breaks = 10)

> print(price\_range)

[1] 31.5000 125.8333 220.1667 314.5000 408.8333 503.1667 597.5000 691.8333

[9] 786.1667 880.5000

> print(frequency\_table)

[31.5,126] (126,220] (220,314] (314,409] (409,503] (503,598] (598,692]

14 49 22 5 3 3 1

(692,786] (786,880]

1. 2

**Figure 3**

*Histogram of Homes in Gainesville, Florida*

A graph of a house selling prices

Description automatically generated

> mean\_selling\_price <- mean(Florida\_house\_data$price)

> standard\_deviation\_selling\_price <- sd(Florida\_house\_data$price)

> within\_one\_standard\_deviation <- Florida\_house\_data$price[Florida\_house\_data$price > (mean\_selling\_price - standard\_deviation\_selling\_price) & Florida\_house\_data$price < (mean\_selling\_price + standard\_deviation\_selling\_price)]

> percentage\_within <- length(within\_one\_standard\_deviation) / length(Florida\_house\_data$price)\*100

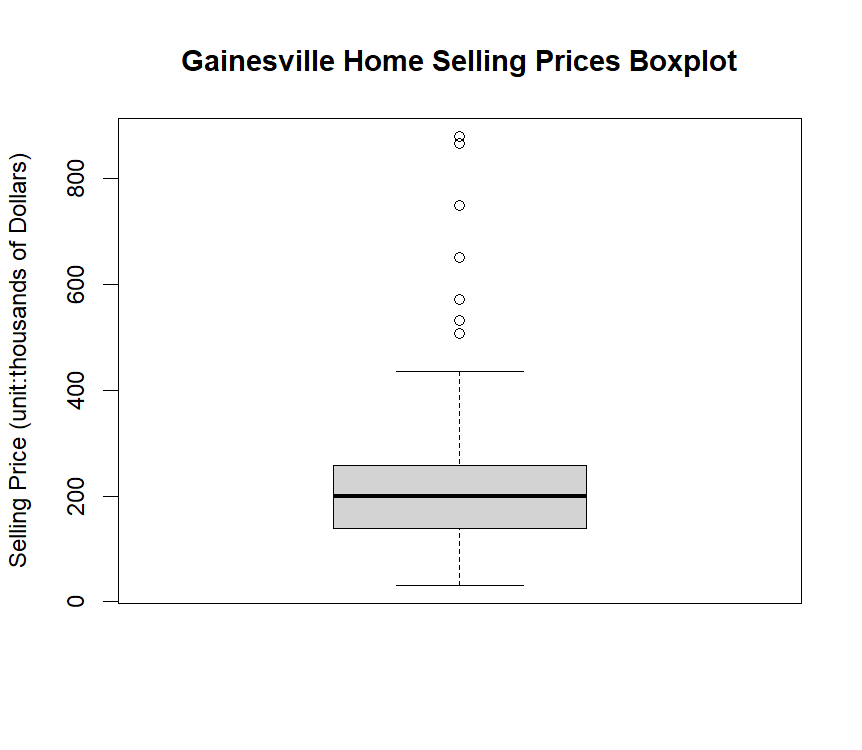
> cat("Percentage of observations of mean Gainesville Florida:", percentage\_within, "%\n")

Percentage of observations of mean Gainesville Florida: 85 %

> boxplot(Florida\_house\_data$price, main = "Florida House Selling Prices Boxplot", ylab = "Selling Price (unit:thousands of Dollars)")

**Figure 4**

*Boxplot Figure of Homes in Gainesville, Florida*



**References**

Thulin, M. (2021). Modern Statistics with R. Eos Chasma Press. ISBN 9789152701515.

University of the Cumberlands (2024) *Index of Dataset (Dr. Danny Barnes).* https://stat4ds.rwth-aachen.de/data/.