InClassAssignment1(Group of two)
CS160-02
Introduction to Data Science
Spring 2023

Working on Techniques for Analyzing Data

Instructions: Complete the following activities for this project.

1. Create a new GitHub repository named Assignment1_XXX, where XXX are your initials.
2. Using excel (to generate the result) and word documents (type answers and paste the results) work on the following questions and submit your work using **pdf** format.

a. What are the differences between data analysis and data analytics?

Data analytics is a broader term that includes data analysis. The aim of data analytics is to make data understandable for businesses and individuals.

Data analysis is a specialized type of data analytics which is used to evaluate data and identify relationships between variables.

b. Comment on variable types of Murder, Assault, and urban pop.

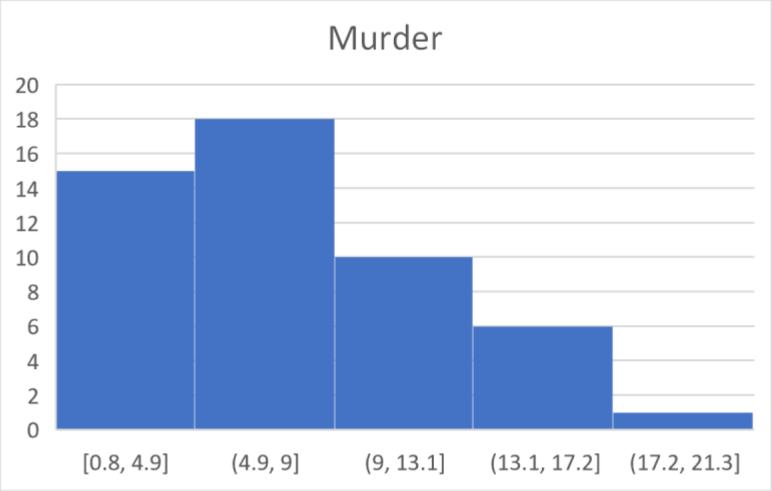
Murder, assault, and urban population are independent variables since they do not depend on each other. Furthermore, all variables are continuous as some numerical inputs are in the form of a decimal. It is also important to note that the first column which lists the states of America are categorical and nominal data. This is because this is data that has no numerical value. In addition, murder, assault, and urban pop are ratio data since it is numerical and there are no negative values, but is measured on a scale with an absolute zero.

c. What is the difference between interval and ratio data?

Both interval and ratio data is measures on a scale with an absolute zero. However, interval data can have negative values and ratio data cannot.

d. What is descriptive analysis? Represent the data of Murder, Assault, and urban pop. Comment on the distribution.

Descriptive analysis is a type of analysis performed on large volumes of data. It can represent data through histograms, box plots, and through calculating the measures of centrality and dispersion of distribution.

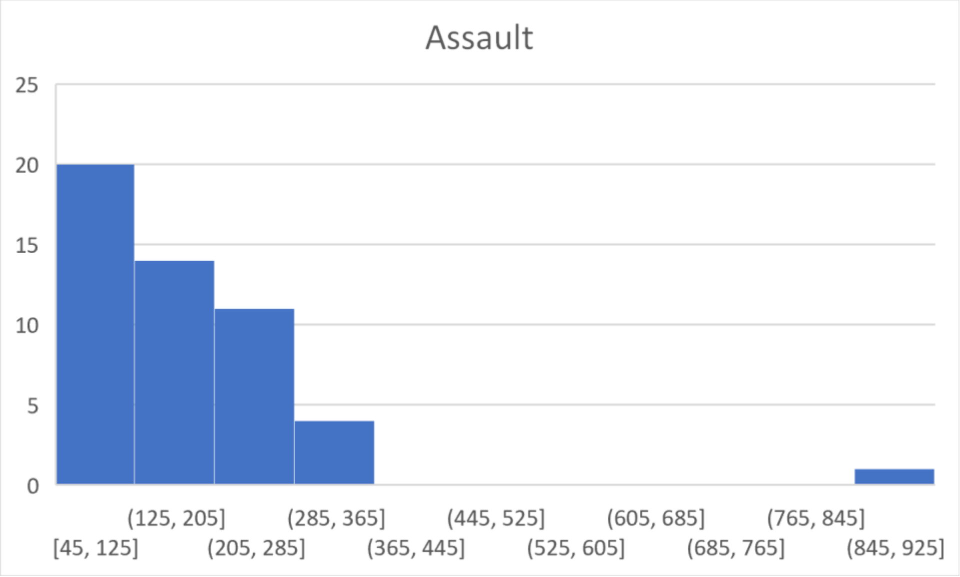


The murder rates in all states are right-skewed. This is because

Mean =AVERAGE(B:B) = 7.788

Median =MEDIAN(B:B) = 7.25

Thus, the mean is greater than the median. Therefore, the murder rates are right-skewed.

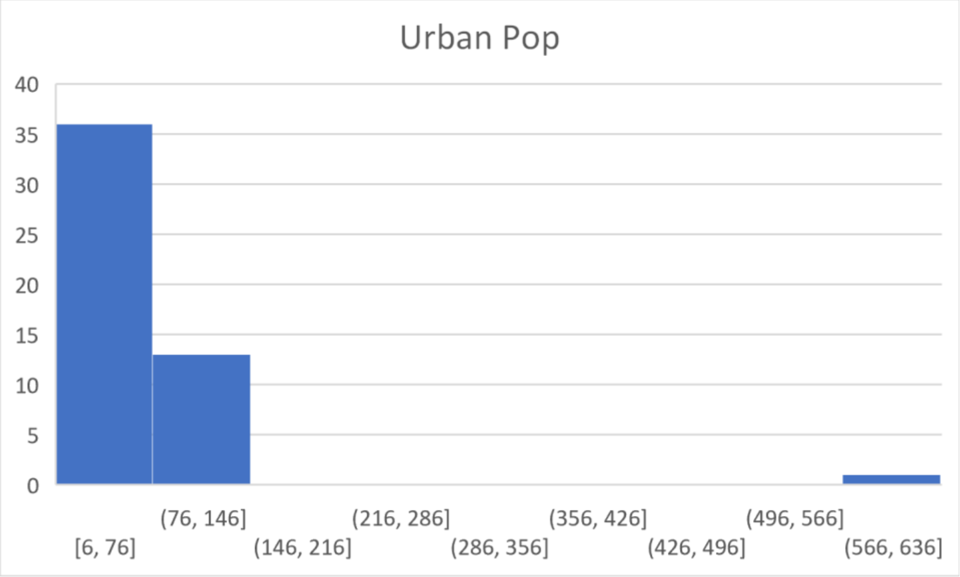


The assault rates in all states are right-skewed. This is because

Mean =AVERAGE(C:C) = 182.1837

Median =MEDIAN(C:C) = 159

Therefore, we can see again that the mean is greater than the median. Thus, the assault rates are right-skewed.



The urban pop rates are right-skewed. This is because

Mean =AVERAGE(D:D) = 74.2

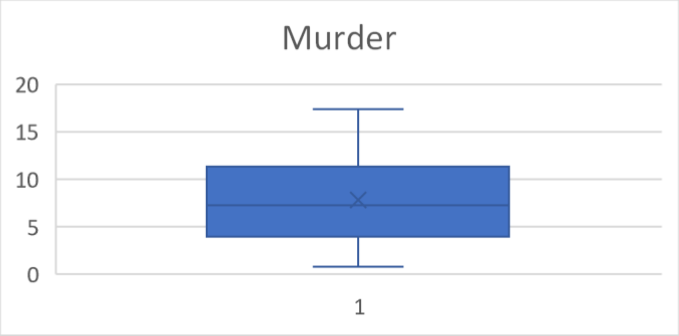
Median =MEDIAN(D:D) = 66

Hence, since the mean is greater than the median, the urban pop data is also right-skewed.

e. What is a measure of dispersion? Calculate the interquartile range of those three variables.

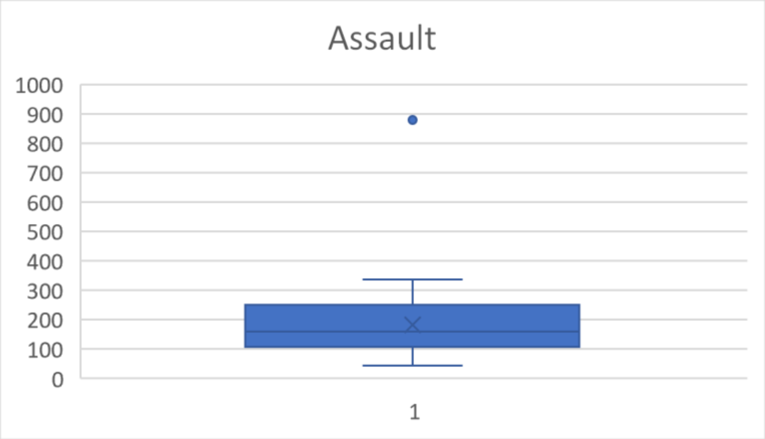
A measure of dispersion is a value that represents the spread of data.

This can be measured using the interquartile range to determine the spread of data.



Murder IQR = 7.175

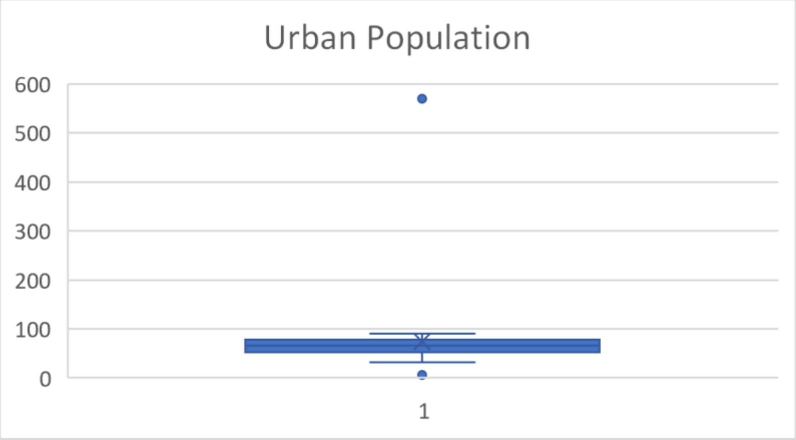
I calculated this using =QUARTILE(B:B,3)-QUARTILE(B:B,1)



There is an outlier located above the boxplot displayed above. This is data outside of where the other points are gathered.

Assault IQR = 140

I calculated this using =QUARTILE(C:C,3)-QUARTILE(C:C,1)



There is also an outlier located above the boxplot displayed above. This is data outside of where the other points are gathered.

Urban Population IQR = 24.5

I calculated this using =QUARTILE(D:D,3)-QUARTILE(D:D,1)

f. What is the measure of centrality? Find the measurement of centrality: mean, median, mode

A measure of centrality is a value that measures the spread of data at the center of the distribution.

Murder: Mean = 7.788, Median = 7.25, Mode = 13.2

Assault: Mean = 182.1837, Median = 159, Mode = 120

Urban Pop: Mean = 74.2, Median = 66, Mode = 80

g. What are diagnostic analytics? Find diagnostic analysis for pairs of variables.

Diagnostic analytics is a method of analyzing data in which determines the cause of an event i.e., why something occurred. This can be determined by finding its correlation.

The correlation coefficient for murder and assault is 0.649377, which I calculated using =CORREL(B:B,C:C)

The correlation coefficient for murder and urban pop is -0.18617, which I calculated using =CORREL(B:B,D:D)

The correlation coefficient for assault and urban pop is -0.140663, which I calculated using =CORREL(C:C,D:D)

3. Using the instructions provided by GitHub, create a git repository named DS160**InClassAssignment**, and push your pdf file to it. Each of you needs to submit your work.

Submission:

Paste a link to your GitHub repository in the area provided for this assignment and submit it by class time.