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 analysis is more of hands-on data exploration and evaluation of that data of what happened. Data analytics deals more with modeling the future outcomes based on analysis that is done. Analytics is used as a much broader term typically.

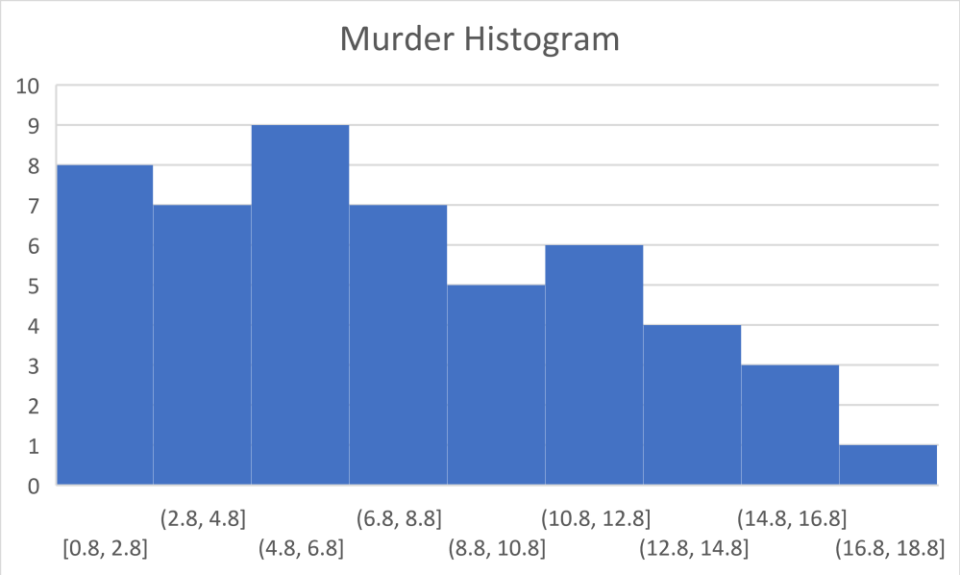
    b. Comment on variable types of Murder, Assault, and urban pop.

Murder, assault, and urban population are all independent, continuous variables representing ratio data since the measurements all have an absolute zero. The state column contains categorical nominal data.

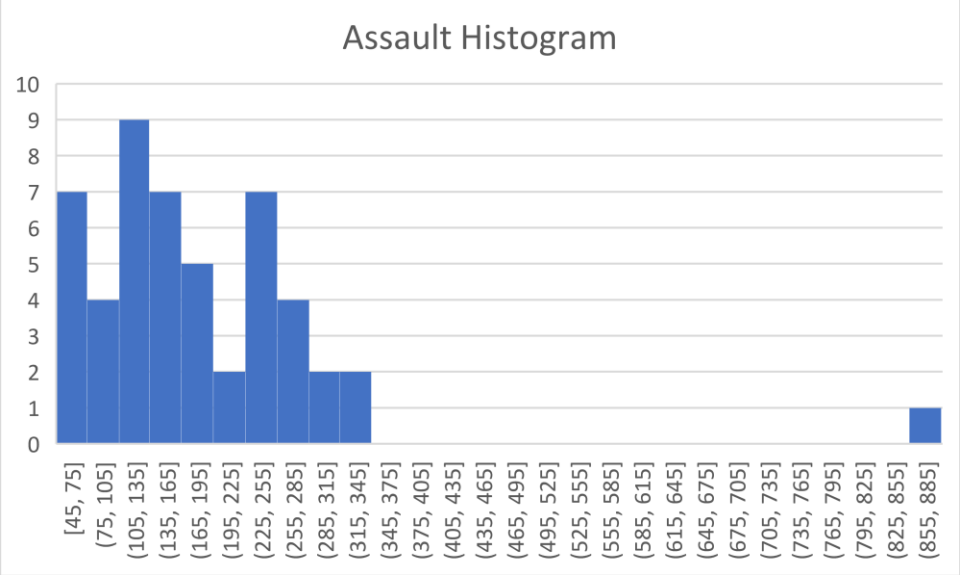
    c. What is the difference between interval and ratio data?

Interval data is when there is equal spacing between adjacent number values without an absolute zero, meaning values can go above and below the number zero. Ratio data is when numbers have units of equal magnitude and rank order on a scale with an absolute zero such as height and weight. Ratio data cannot be negative.

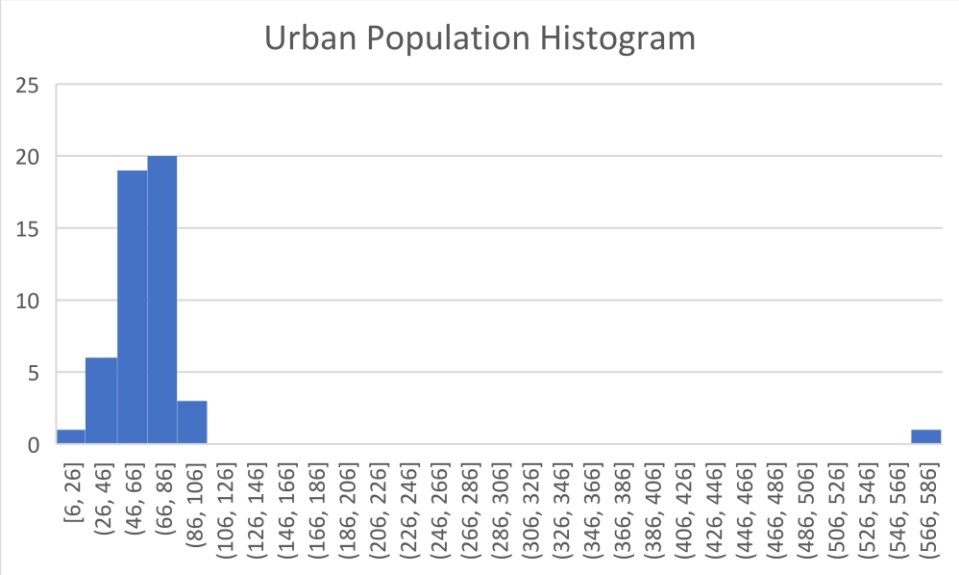
    d. What is descriptive analysis? Represent the data of Murder, Assault, and urban pop. Comment on the distribution.



In the murder histogram, the distribution appears to be right-skewed, but slightly. The mean value is 7.788, the median value is 7.25, and the mode value is 13.2. The data is positively skewed because the mean value is greater than the median. However, the mode is greater than both values.

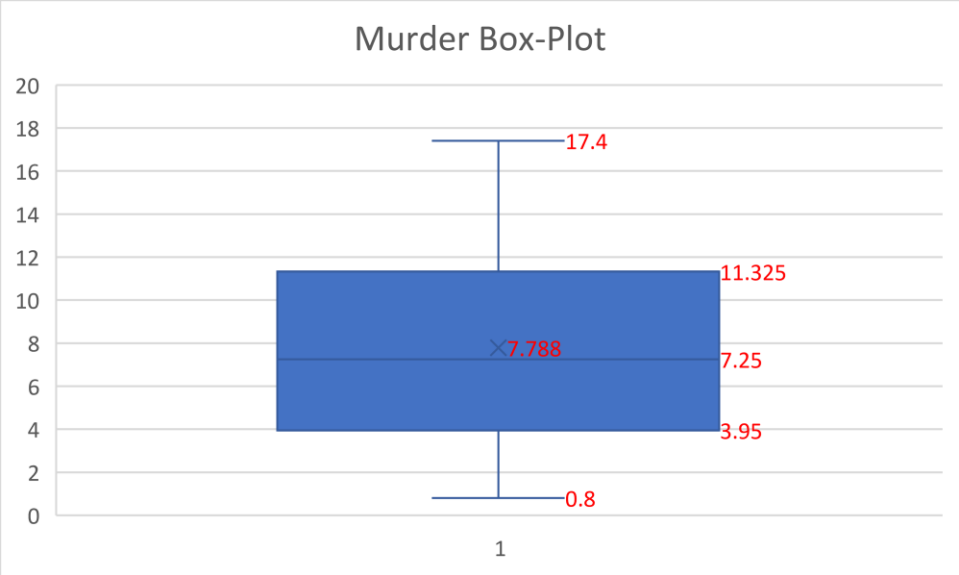


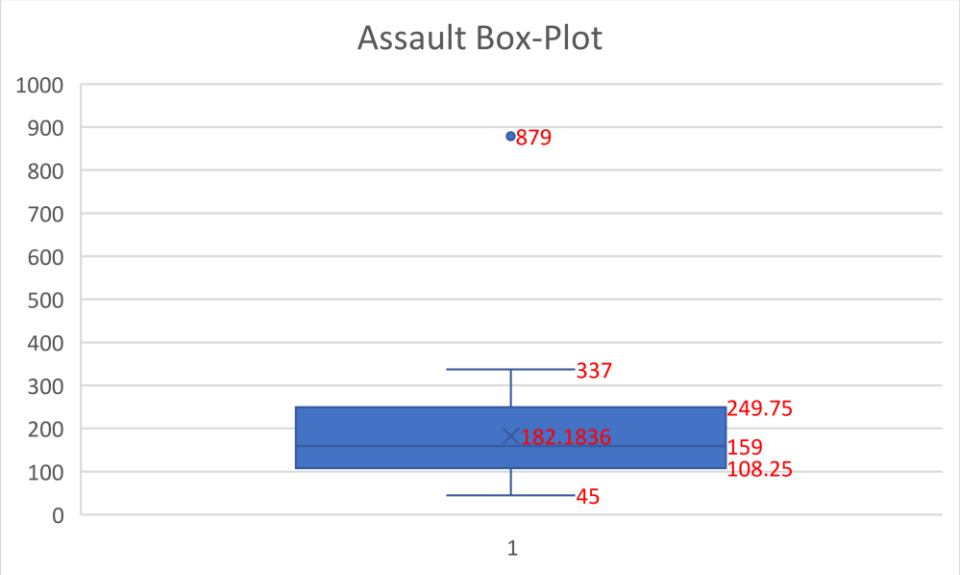
In the assault histogram, the distribution is also positively skewed. The mean value is 182.1836, the mode value is 120, and the median value is 159. The data is positively skewed because the mean is greater than the mode, the median is greater than the mode, and the mean value is greater than the median. There is an apparent outlier of 879.

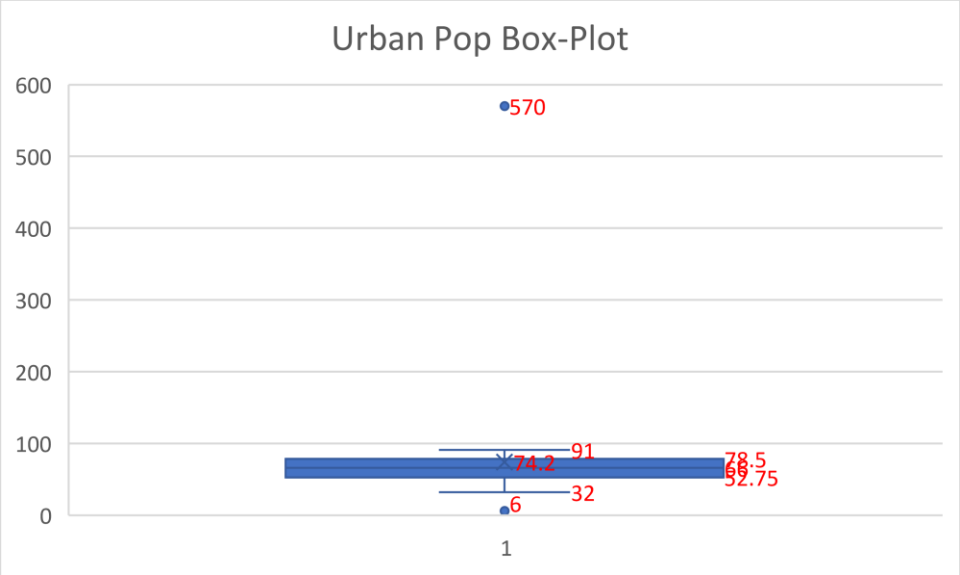


In the urban population histogram, the data appears to be very positively skewed. The median value is 66, the mode is 80, and the mean is 74.2. Like the murder histogram, the mode value is greater than both the median and the mean, but this is due to an apparent outlier value of 570.

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







	Murder	Assault	Urban Pop.
Q1	4.075	109	53.25
Q3	11.25	249	77.75
IQR	7.175	140	24.5

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

Murder	Assault	Urban Pop.

Mean	7.788	182.1836	74.2
Median	7.25	159	66
Mode	13.2	120	80

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

	Correlation
Murder & Urban Pop.	-0.18617
Murder & Assault	0.649376
Assault & Urban Pop.	-0.14066

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.