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 a part of data analytics; the two are not the same, though the terms sound alike. Data analytics is a broad field that makes decisions based on data using tools. It models the future or predicts a result – this is only possible because of data analysis, the process of hands-on data exploration and evaluation. Data analysis is a necessary step in data analytics.

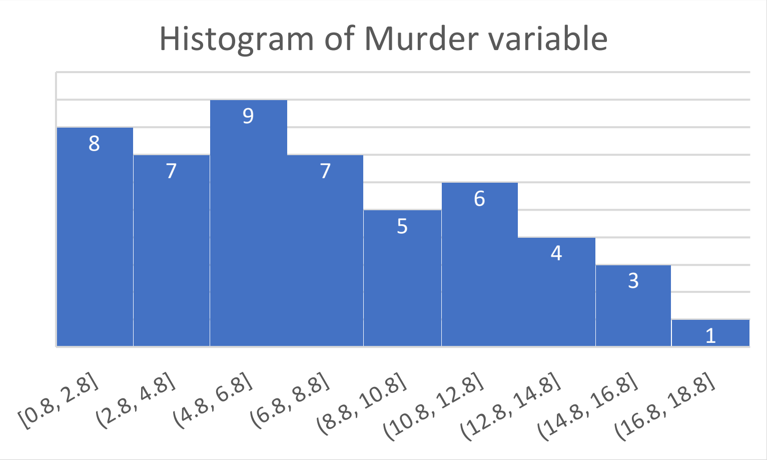
   b. Comment on variable types of Murder, Assault, and urban pop.

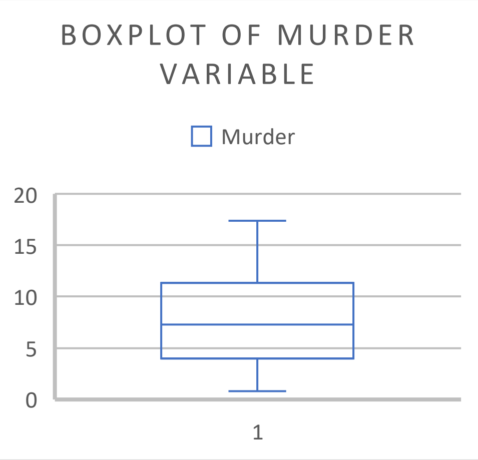
   Each of these variables are independent – none of these variables are influenced by any others. All these variables are numerical (quantitative) and continuous. They are continuous because they include decimal points – between any two values there is an infinite number of values. Additionally, these data are ratio data; they each have an absolute zero meaning numbers are always positive. There is one other variable in the dataset: the names of the states. This variable is categorical (qualitative) and nominal, meaning there is no order or ranking in the data.

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

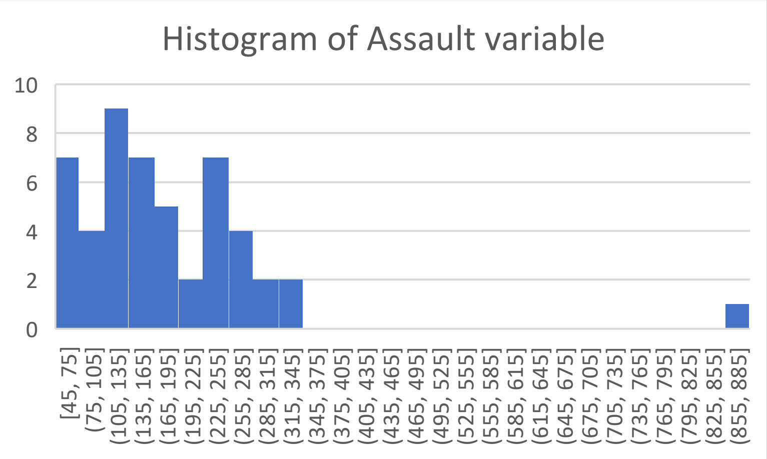
   Interval data is a set of numbers that have equal spacing between adjacent values without an absolute zero – this means the data can be negative values, there is no floor at zero. For example, data on subjects' changes in weight over 2 months can reasonably be positive values (if weight is gained), negative values (if weight is lost), or zero (if weight doesn't change). Ratio data is when numbers have units of equal magnitude and rank order on a scale with an absolute zero – meaning the data cannot be negative, like current height and current weight for example.

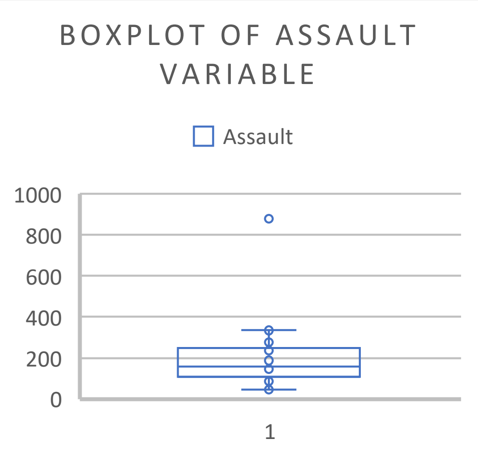
d. What is descriptive analysis? Represent the data of Murder, Assault, and urban pop. Comment on the distribution.



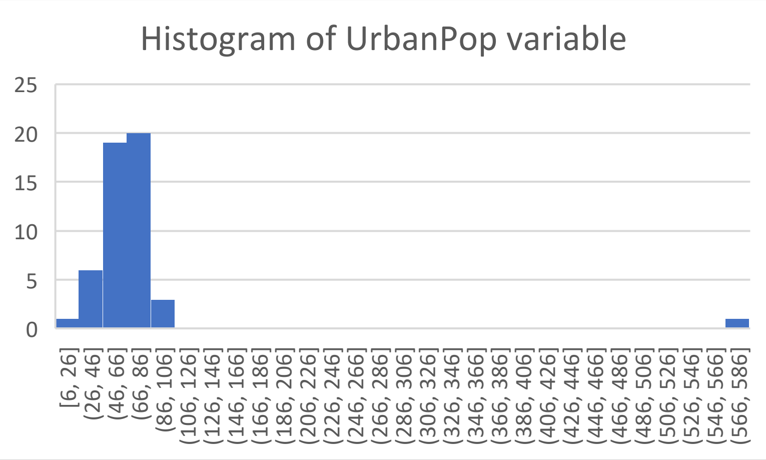


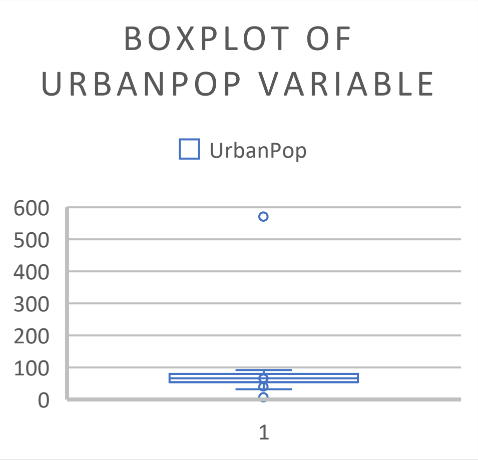
This histogram shows the distribution of the Murder variable. We can see skew in both the histogram and the boxplot. The median is smaller than the mean – in a Normal distribution, the median should be the same as the mean as the data would be evenly distributed on both sides of the median. Since the mean is larger than the median in this case, we know that the skew is positive.





This histogram of the Assault variable shows most of the observations clustered between 45 and 345 with one very large outlier at 879. With the outlier removed, the distribution is not obviously skewed positively or negatively. This isn't to say the distribution is Normal, it appears bimodal and does not show the Normal bell curve. The boxplot shows the same and the large outlier is clearly shown by the small circle far above the box and whiskers which indicates an outlier. For this variable, the mean is the largest, followed by median, then mode. This relationship - where mean is larger than median and mode, and median is larger than mode – points to a positively skewed distribution (largely because of the strong outlier).





Like the distribution of the Assault variable, the UrbanPop variable contains a very strong outlier. This causes the distribution to be right (or positively) skewed – however, when the outlier is removed, the distribution appears approximately Normal. With the outlier included, the mean is higher than the median (although the mode is higher than both median and mean), which confirms the positive skew of the distribution.

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

   Dispersion is the how the data are spread throughout the distribution. We measure this with range, interquartile range, variance, and standard deviation. For each of these variables, the interquartile ranges are Murder = 7.175, Assault = 140, and UrbanPop = 24.5.

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

   To describe the center of the distribution, we observe the mean, median, and mode. For the Murder variable, mean is 7.78, median is 7.25, mode is 13.2. For Assault, mean is 189.184, median is 159, and mode is 120. For UrbanPop, mean is 74.2, median is 66, mode is 80.

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

   Diagnostic analytics are used to determine why something happened. Most frequently, the Pearson correlation technique is used to understand relationships between variables. The correlation value for Murder and Assault is 0.649 – this tells us there is a somewhat strong positive relationship and that states with high murder rates are likely to have high assault rates as well, and vice versa. For Murder and UrbanPop, the correlation coefficient is –0.186 which indicates a weakly negative relationship between the two, higher murder rates are connected to smaller urban populations, but not very strongly. Similarly, the correlation value for Assault and UrbanPop is –0.141.

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.