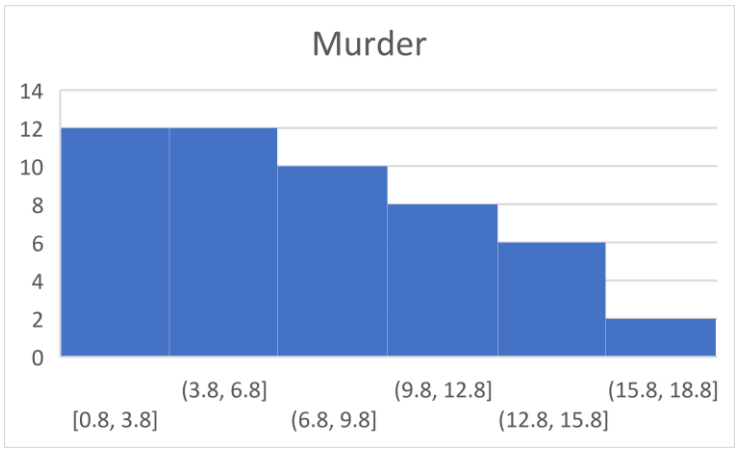
InClassAssignment1(Group of two)
CS160-02
Introduction to Data Science
Spring 2023

Working on Techniques for Analyzing Data

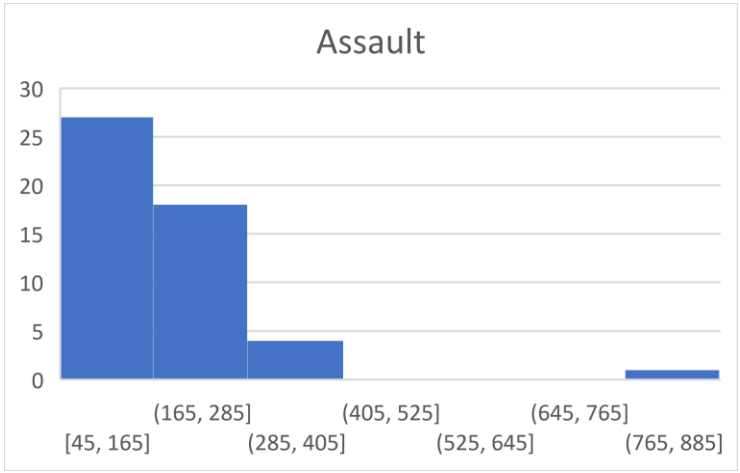
Instructions: Complete the following activities for this project.

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.

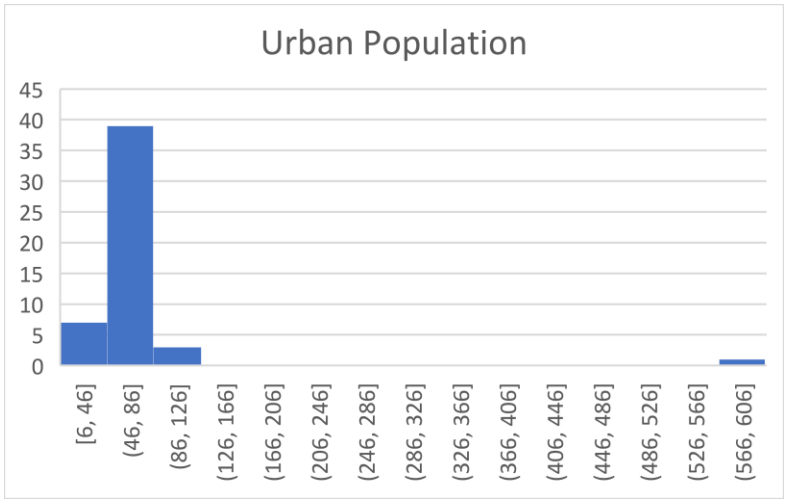
    a. What are the differences between data analysis and data analytics?
        a. **Data analysis is exploration and evaluation and looks for a cause for the effect. Data analytics tries to take into data and predict the future and includes data analysis.**
    b. Comment on variable types of Murder, Assault, and urban pop.
        a. **All are independent**
        b. **The states are categorical and nominal. Murder, assault, and urban population are continuous**
        c. **All variables are ratio.**
    c. What is the difference between interval and ratio data?
        a. **Interval data can go below zero (ex. Temperature). Ratio data has an absolute zero (ex. Number of people in a room).**
    d. What is descriptive analysis? Represent the data of Murder, Assault, and urban pop. Comment on the distribution.
        a. Descriptive analysis is the first step in data analysis. Uses concepts such as plotting the data on a histogram to find distribution and using measures such as mean, median, mode, range, etc.

b. 

c. Distribution is right skewed, with most of the data from 3.8 to 6.8. This distribution is not as right skewed as others because the mode is 13.2 but because the mean (7.79) is greater than the median (7.25), it is right skewed.

d. 

e. Distribution is right skewed with a high outlier of 879 with most of the data from 45 to 165. The distribution is right skewed because the mode (120) is less than the median (159) which is less than the mean (182.18).

f. 

g. The distribution is right skewed with a high outlier of 570. Most of the data is between 46 and 86. It is right skewed because the mean (74.2) is greater than the median (66)

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

a. Measure of dispersion is how to describe how far or close the data is. IQR of murder is 7.18. IQR of assault is 140. IQR is 24.5 for urban population.

b.

	murder	assault	urban pop
q1	0.8	45	6
q2	4.075	109	53.25
q3	7.25	159	66
q4	11.25	249	77.75
q5	17.4	879	570
IQR	7.175	140	24.5

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

a. The measure of centrality is the value that shows the center of the distribution. The mean, median, and mode are as follows, respectively:

b. Murder – 7.79, 7.25, 13.2

c. Assault – 182.18, 159, 140

d. Urban population – 74.2, 66, 80

e.

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.

a. Diagnostic analytics are used to find what kind of relationship two variables have. For example, the correlation between murder and assault is .64, indicating it has a somewhat strong, positive correlation. The correlation between murder and urban pop has a weak negative correlation. The correlation between assault and urban pop also has a negative and weak relationship.

b.

correlation murder and assault	0.649375912
correlation murder and urban pop	-0.186169557
correlation assault and urban pop	-0.140663143

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.