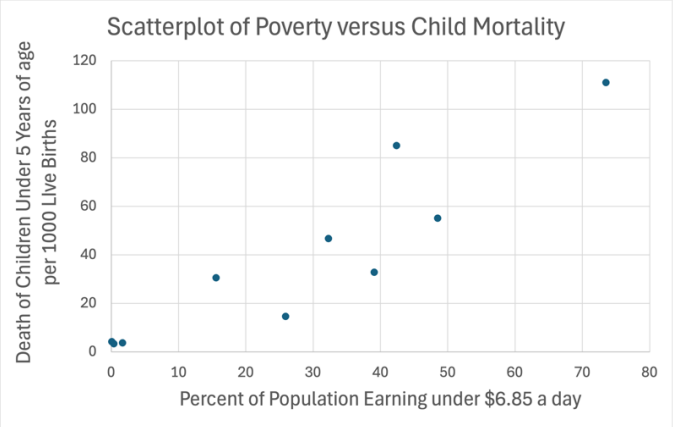
ACTIVITY 2: Worksheet Answers

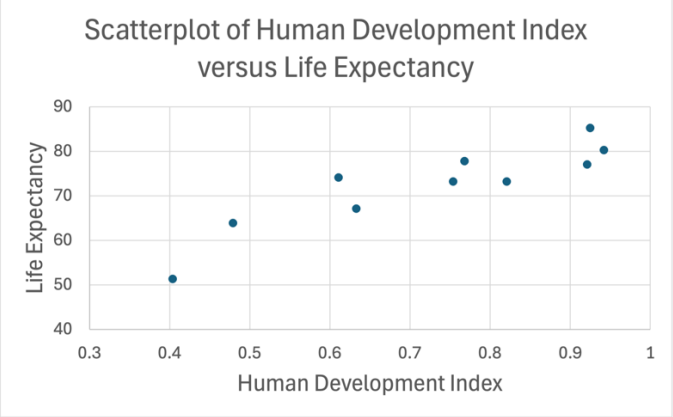
In this activity, you will be learning about scatterplots and the correlation coefficient. Recall that correlation is not causation, and in a later activity, we will distinguish between the two.

A **scatterplot** is a graph that plots, as a point, the numerical value of two different variables for each statistical unit on a graph. Scatterplots must include a title and labels with units on each axis.

Note: A statistical unit is an individual or object being measured or observed in a study whereas the mathematical unit refers to the standard of measurement used to quantify the values of a variable and is included when labeling the axes.

Below are some examples of scatterplots.





Examine the scatterplots to the left.

1. For the scatterplot of poverty versus child mortality, which variable is the independent variable and which variable is the dependent variable?
   The independent variable is the percent of the population earning under $6.85 a day and the dependent variable is the death of children under 5 years of age per 1000 live births.

2. For the scatterplot of human development index versus life expectancy, which variable is the independent variable, and which is the dependent variable?
   The independent variable is the human development index, and the dependent variable is life expectancy.

3. For both graphs, the independent variable was on which axis?
   The independent variables for both graphs were on the x-axis.

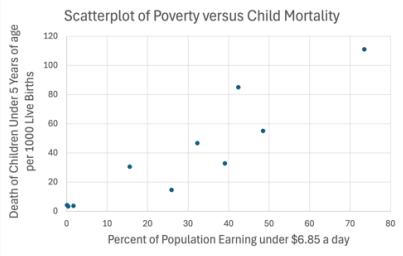
4. For both graphs, the dependent variable was on which axis?
   The dependent variables for both graphs were on the y-axis.

Based on your answers for #3 and #4, create a rule for how you need to set up your scatterplots when you have an independent and dependent variable.
The independent variable should always be on the x-axis and the dependent variable on the y-axis.

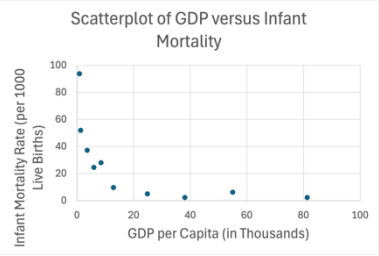
Scatterplots can show the <u>form</u>, <u>direction</u>, and <u>strength</u> of the relationship between two quantitative variables.

Form: Is the relationship linear or non-linear?

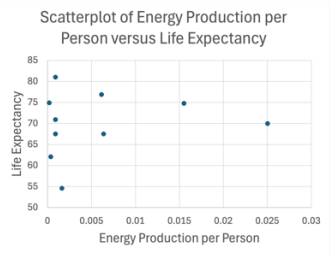
5. For the scatterplots below, classify the variables as having a linear or non-linear relationship. Justify your answers. If the data looks like it can be modeled by a line, use the appropriate phrase of either "constant rate of change" or "varying rate of change."



Linear – There appears to be a linear relationship between the percentage of the population earning under $6.85 per day and the deaths of children under 5 years of age per 1000 live births because the data can be modeled with a line which has a constant rate of change.



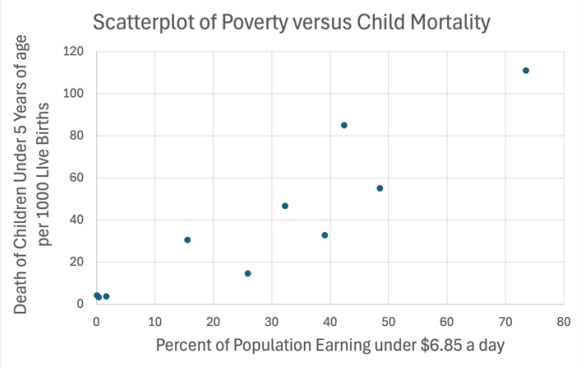
Non-Linear – There is a varying rate of change between the GDPs per capita (in thousands) and the Infant mortality rate (per 1000). More specifically, the rate of change approaches zero as GDP increases.



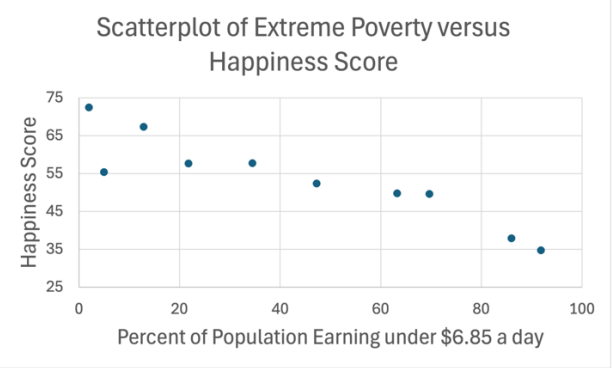
Non-Linear – There is no relationship between these two variables as the points are just randomly scattered.

Direction: Does the relationship show a positive or negative slope?

6. For the scatterplots below, does an increase in the independent variable correspond to a positive change (positive relationship), a negative change (negative relationship), or no change (no relationship) in the dependent variable? Justify your answers.



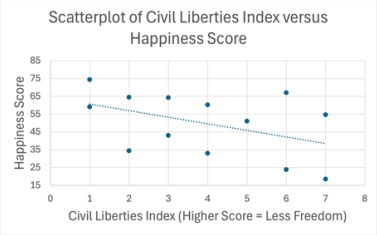
Positive – As the percentage of the population earning under $6.85 per day increases, the death of children under 5 years of age per 1000 live births also increases.



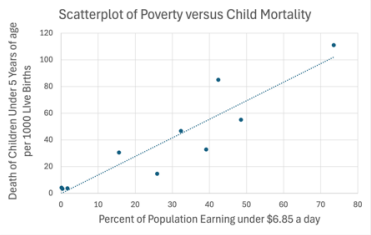
Negative – As the percentage of the population earning under $6.85 per day increases, the happiness score decreases.

Strength: For a strong relationship the points will closely follow a pattern. More specifically, linear relationships are stronger when points hug the regression line, whereas linear relationships are weaker when the points are more spread out and away from the regression line.

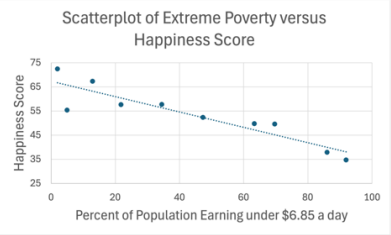
7. For the scatterplots below, classify the variables as having a weak, moderate, or strong linear relationship. Justify your answers.



Weak – The points vary considerably around the line, so the relationship would be considered weak.



Moderate – The points vary moderately around the line, so the relationship would be considered moderate.



Strong – The points closely hug the line, so the relationship would be considered strong.

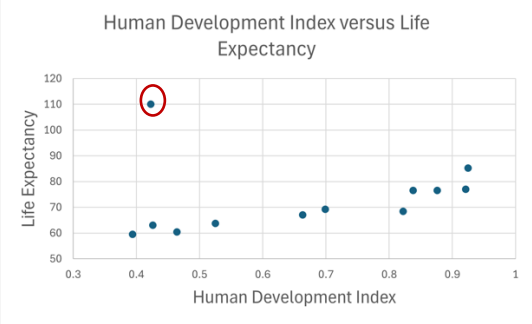
Scatterplots can also show outliers. Below are two types of outliers.

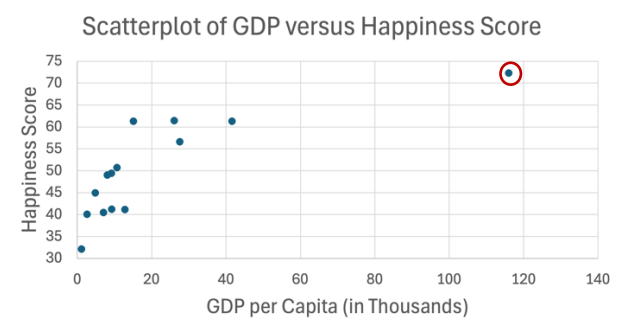
An outlier in the **y-direction** is a point that is significantly higher or lower than the general trend of the rest of the data that includes values with similar x-values. Note that outliers in the y-direction have considerably higher or lower values for the dependent variable compared to other points *with similar values* for the independent variable.

An outlier in the **x-direction** is a point that is significantly to the left or right compared to the rest of the data. Note that outliers in the x-direction will have values for the independent variable that are considerably different to *all other* values of the independent variable.

Important: When identifying outliers, it's crucial to assess their validity. Since errors can occur, it's important to verify whether the outlier is reasonable within the context of your data.

8. **A.** Circle the outliers below and state whether they are in the x-direction or y-direction.





y-direction

x-direction

B. For the graph that compares the Human Development Index and life expectancy, do you think the outlier makes sense in the context of the data or do you think it could have been a typo or mistake? Explain.

The outlier does not make sense. It is feasible for a person to live to 110 years old, but it is not feasible for the average age of everyone in an entire country to be 110 years old.

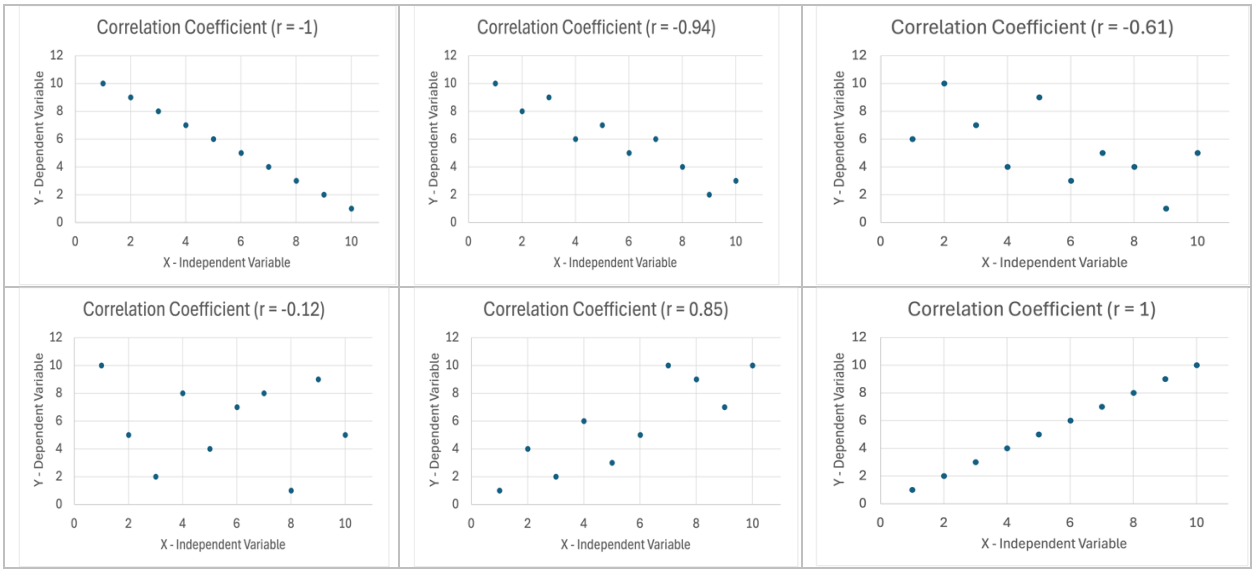
C. For the graph that compares GDP per capita (in thousands) to happiness score, the outlier belongs to Qatar. Why do you think Qatar might have such a high GDP compared to the other countries in the sample?

Qatar is a big oil producer, so it has an extremely high GDP compared to many other countries.

As you most likely noticed, trying to determine the strength of a relationship from just the scatterplot is a bit subjective. For linear relationships, one can calculate what is called the correlation coefficient which measures the strength of the linear relationship. The correlation coefficient, *r*, satisfies the following rule:

$$-1 \leq r \leq 1.$$

Based on the pictures and value of the corresponding correlation coefficients below, answer the questions that follow.



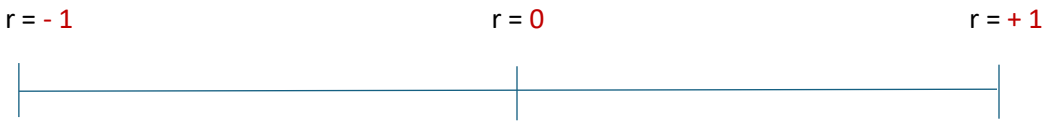
9. How does the dependent variable change as the independent variable increases when the correlation coefficient is negative? What would be the direction of the relationship?
The dependent variable decreases as the independent variable increases; therefore, the direction is negative.

10. How does the dependent variable change as the independent variable increases when the correlation coefficient is positive? What would be the direction of the relationship?
The dependent variable increases as the independent variable increases; therefore, the direction is positive.

11. What is the strength of the linear relationship when the correlation coefficient is close to zero?
A correlation coefficient close to zero indicates a weak relationship.

12. What is the strength of the linear relationship when the correlation coefficient is close to -1 or 1?
A correlation coefficient close to -1 or 1 indicates a strong relationship.

Give the correlation coefficients that fit with the number line below:

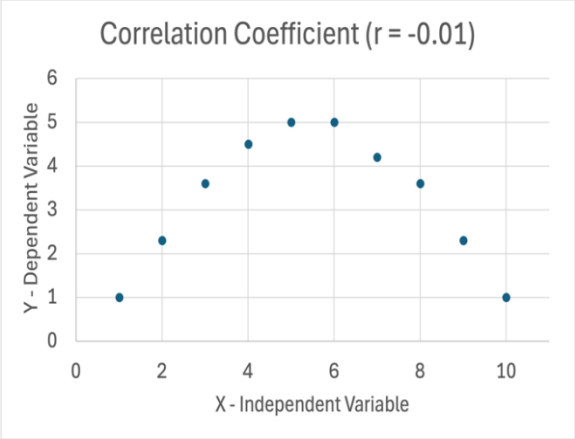


Strong negative relationship:
As the independent variable increases, the dependent variable decreases at a near-perfect constant rate of change.

No relationship:
As the independent variable increases, the dependent variable varies without pattern.

Strong positive relationship:
As the independent variable increases, the dependent variable increases at a near-perfect constant rate of change.

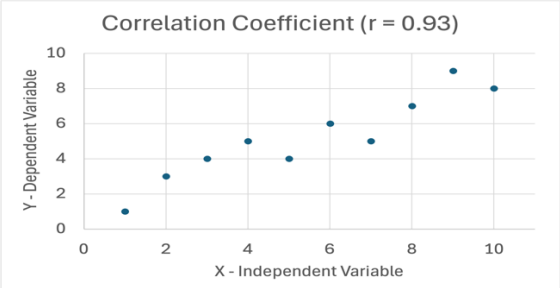
It is important to understand when it is appropriate to calculate the correlation coefficient and how outliers impact its value. Keeping in mind that a strong relationship occurs between variables when the points in a scatterplot closely follow a pattern, answer the questions below.



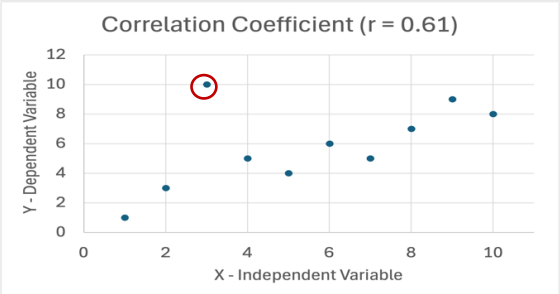
13. How would you describe the strength and form of the relationship based solely on the scatterplot to the left?
The relationship is nonlinear and extremely strong.

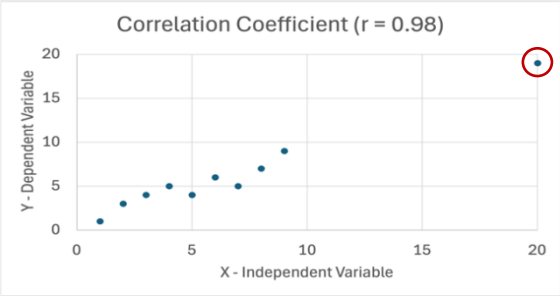
14. The correlation coefficient is only r = -0.01. Is this a good indicator of the strength of the relationship?
No, because the relationship is very strong.

15. Do you think it is appropriate to use the correlation coefficient for nonlinear relationships?
No, it does not accurately reflect the strength of a nonlinear relationship.



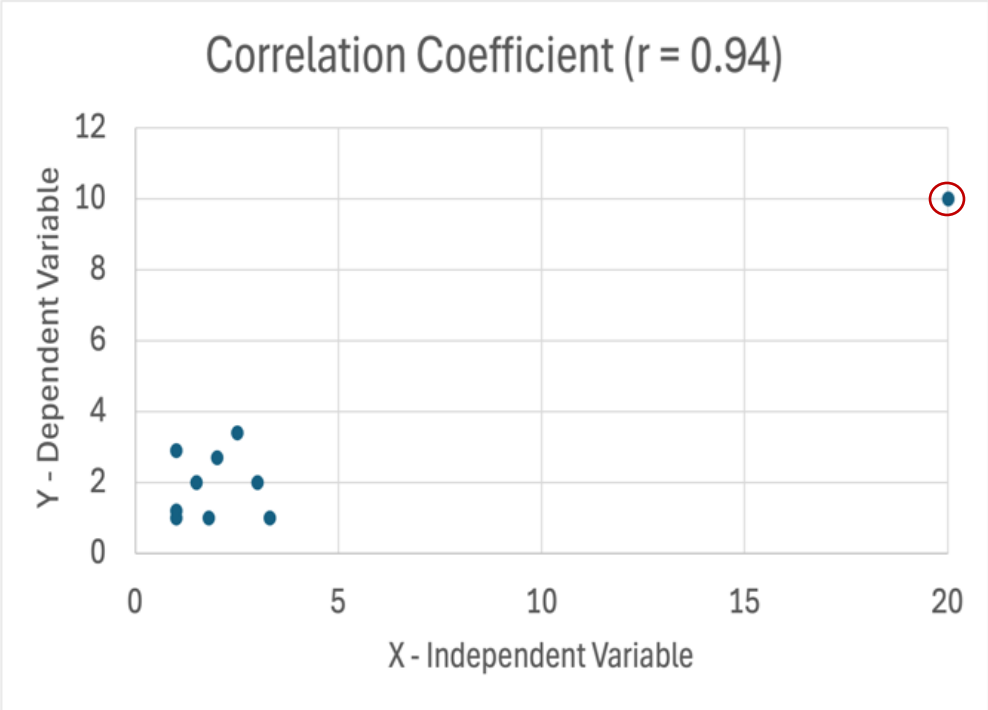
16. The scatterplot to the left does not have any outliers. The scatterplot below it, however, shows the same data with an outlier replacing one of the non-outliers. In what direction is this outlier and how does it change the correlation coefficient?
The outlier is in the y-direction, and it has decreased the value of the correlation coefficient.





17. The last scatterplot shows the same data with a different outlier replacing one of the non-outliers. In what direction is this outlier and how does it change the correlation coefficient?
The outlier is in the x-direction, and it has increased the value of the correlation coefficient.

We will see in the next activity that outliers in the x-direction can be considered influential values, which means they have an extremely strong influence on the correlation coefficient and the regression line. In such cases, they should be removed from your data. On the other hand, outliers in the y-direction can affect the correlation coefficient, but their impact on the regression line is typically less significant; this is especially true for larger data sets.
Answer the questions below based on the graphic and the correlation coefficient.



18. The scatterplot above has an extreme outlier in which direction?
Note: You may need to revisit the definitions of outliers.
The scatterplot above shows an outlier in the x-direction.
Note: The outlier is not in the y-direction as there are no data values with x-values "near" x=20.

19. The scatterplot above has a correlation coefficient equal to 0.94. What type of relationship does a correlation coefficient of 0.94 indicate?
A correlation coefficient of 0.94 indicates a strong, positive, and linear relationship.

20. Do you think that the correlation coefficient accurately reflects the relationship?
No, the points on the left do not look linearly related.

21. If the outlier is removed, the correlation coefficient will be 0.06. What type of relationship does a correlation coefficient of 0.06 indicate?
A correlation coefficient of 0.06 indicates a very weak almost non-existent relationship.

22. What does the answer to #20 and #21 tell you about calculating correlation coefficients without looking at the scatterplot?
It is not advised as the correlation coefficient is greatly impacted by outliers in the x-direction.

Use the given Excel instructions to create a scatterplot and calculate the correlation coefficient for the data you collected in Activity 1.

Excel Instructions for obtaining the scatterplot and calculating the Correlation Coefficient.

1. Type your data into two columns with the independent variable in the left column and the dependent variable in the right column.
2. Highlight both columns of data, click "Insert", select the graph that looks like a scatterplot, and choose the upper left option.
3. Title your scatterplot appropriately by clicking on the chart title and typing in your title.
4. Title your x- and y-axes by clicking on your graph and selecting "Chart Design". You can then select "Add Chart Elements" (on the upper left); followed by "Axis Titles"; and then followed by "Primary Horizontal". Type in the appropriate label for the independent variable in the box that appears on the graph. Then select "Add Chart Elements" (on the upper left); followed by "Axis Titles"; and then followed by "Primary Vertical". Type in the appropriate label for the dependent variable in the box that appears on the graph.
5. **Note:** Typically, you want the points to be spread out across all values of your variables. You may need to adjust the bounds of one (or both) of your axes. To do so, double-click on one of the numbers on the axis you want to adjust. Under the "Format Axis" option, click on the bar graph icon. You'll then notice that you can adjust the bounds under "Axis Options."
6. If your data is linearly related, you can put in a trend line by selecting "Add Chart Elements"; followed by "Trend Line"; and then followed by "Linear".
7. To calculate the correlation, click on an empty box in Excel and type in "=CORREL(". Then highlight the first column of data, type a "," highlight the second column of data, and type a ")". You can then hit enter.
   Note: You will only type the information that is between the quotation marks, not the quotation marks themselves.
8. **Note:** Your final equation will look similar to "=CORREL(A2:A26,B2:B26)".

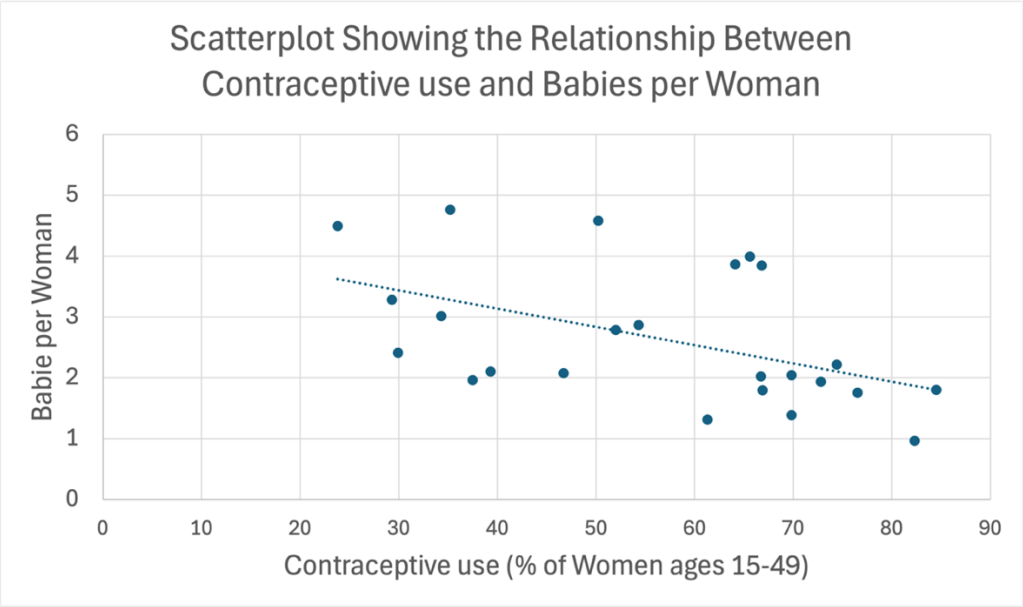
Once you have your graph and correlation coefficient, you are ready to create your next two PowerPoint slides.

On Slide 2, paste the scatterplot and describe the relationship of your data. Make sure you use complete sentences and mention the variables in your description. Remember that you are describing the form, direction, and strength of the relationship and whether or not you have any outliers.

On Slide 3, give the correlation coefficient and state whether it was appropriate to find this value and explain why.

For the example data collected in Activity 1, the following would be the PowerPoint slides.

Slide 2:



The relationship between contraceptive use (% of women ages 15-49) and babies per woman is linear, negative, and weak.

Slide 3 :

The correlation coefficient is r = -0.49.

It would be appropriate to find the correlation coefficient for this data since the relationship, although weak, is linear with no outliers.