Correlation & Regression

Project 2

An important aspect of data science is using data to make *predictions* about the future based on the information that we currently have. A question one might ask would be "Given the amount of time a student studied for an exam, how can we predict their grade?" In order to answer this question, we will investigate a method of using one variable to predict another by looking at the *correlation* between two variables.

1. Standard Units and Correlation

a. When calculating the correlation coefficient, why do we convert data to standard units?

b. Write a function called `convert_su` which takes in an array of elements called `data` and returns an array of the values represented in standard units.

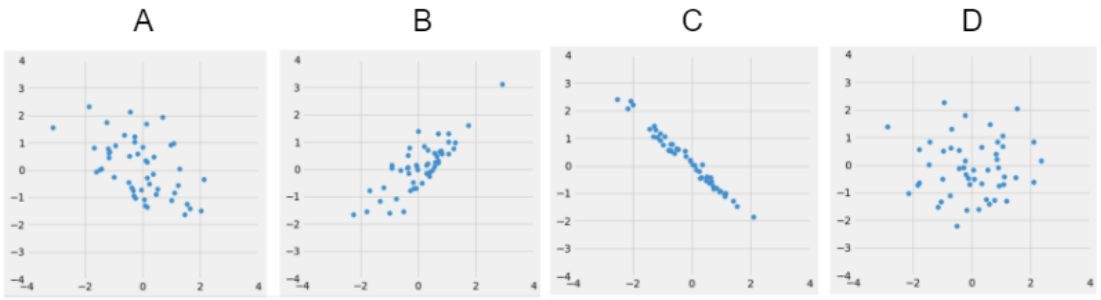
```
def convert_su(data):
    sd = _____
    mean = _____
    return _____
```

c. Write a function called `calculate_correlation` which takes in a table of data containing the columns `x` and `y` and returns the correlation coefficient.

```
def calculate_correlation(tbl, x, y):
    x_su = _____
    y_su = _____
    return _____
```

2. Comparing Correlation

Look at the following four datasets. Rank them from weakest correlation to strongest correlation. Remember that a strong correlation has $|r|$ close to 1.



We have introduced correlation as a way of quantifying the *strength* and *direction* of a linear relationship between two variables. However, the correlation coefficient can do more than just tell us about how clustered the points in a scatter plot are about a straight line. It can also help us define the straight line about which the points (in original units) are clustered, also known as the *regression line*.

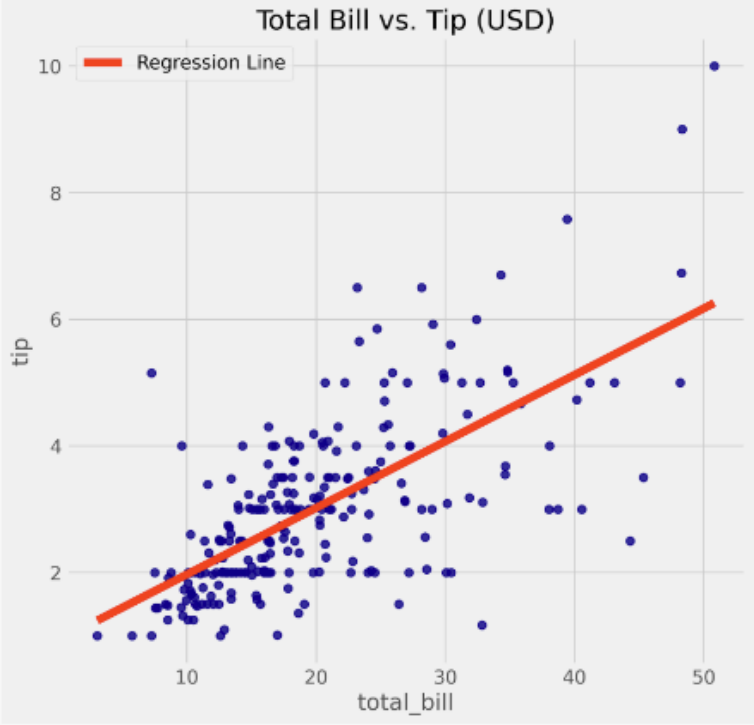
The formulae for the *slope* and *intercept* for the regression line are shown below. In fact, by a remarkable fact in mathematics, the line uniquely defined by the slope and intercept below is *always* the best possible straight line we could use.

$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} = \text{average of y} - \text{slope} \cdot \text{average of x}$$

3. Restaurants

Edwin gives you the scatter diagram shown below that shows the relationship between the total bill versus tip at American restaurants. You have calculated the line of best fit (shown in red).



a. Suppose your friend Gamy goes out to dinner and tells you his total bill was $35. Based on the regression line, what would we predict Gamy's tip to be?

b. Gamy purchased pizzas for a Data 8 pizza party. Would it be a good idea to use the regression line if Gamy's total bill was $200?

c. Later that night, Edwin tells you that the correlation between total bill and tip is 0.8. Knowing this, can you assume that the two variables have a linear association? Circle the correct statement.

A. Yes, r tells us the strength of a linear association and a high value of r always proves that the two variables have a linear association.

B. Yes, because if we can compute the value of r, the X and Y values must have a linear association.

C. No. A high value of r does not necessarily imply that the relationship between the variables is linear.

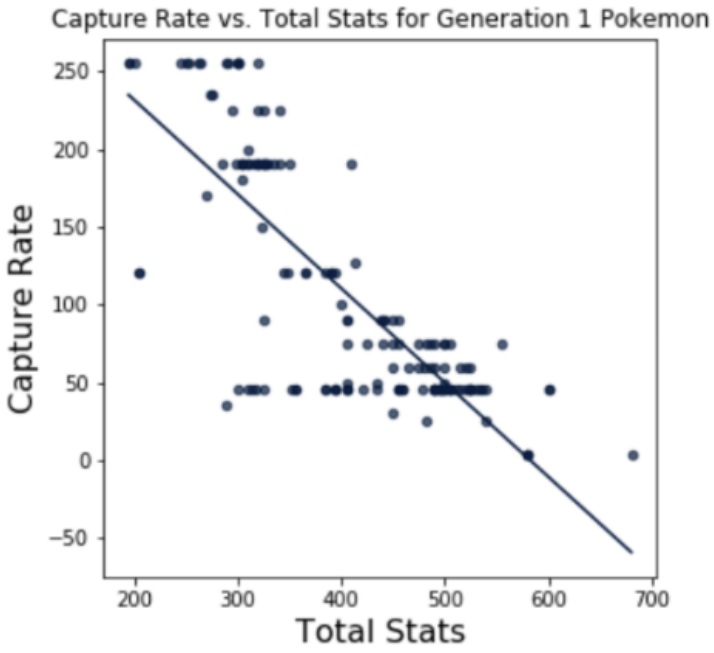
D. No, the value of r = 0.8 is not high enough to demonstrate a linear association.

4. Fa17 Final Q10 Modified

This scatter plot shows a population of pokemon. For each pokemon, we plot its `total stats` (a measure of its power) and `capture rate` (which affects how likely you are to catch the pokemon when you throw a pokeball at it). The plot also shows the regression line through this data.

- The mean of total stats is 407.1, and the standard deviation of total stats is 99.4.
- The mean of capture rate is 106.2, and the standard deviation is 76.9.
- The correlation coefficient is -0.78.

In the parts below, it is OK to write your answer as a Python expression that evaluates to the correct answer.



The summary information about the pokemon are reproduced here for convenience.

- The mean of total stats is 407.1, and the standard deviation of total stats is 99.4.
- The mean of capture rate is 106.2, and the standard deviation is 76.9.
- The correlation coefficient is -0.78.

a. Circle the correct statement:

There appears to be a positive association.
There appears to be a negative association.

b. Pikachu has a `total stats` of 320. What is Pikachu's total stats, in standard units?

c. Charmander has a `capture rate` of 45. What is Charmander's capture rate, in standard units?

d. Circle the correct statement:

The slope of the regression line is greater than zero.
The slope of the regression line is smaller than zero.

e. Calculate the slope of the regression line, in standard units.

f. Calculate the slope of the regression line, in original units.

g. Suppose we encounter a new pokemon with `total stats` 1600.3. Based on the scatter plot and statistics given above, would using the same regression line be appropriate to predict the `capture rate` of this pokemon? Explain.