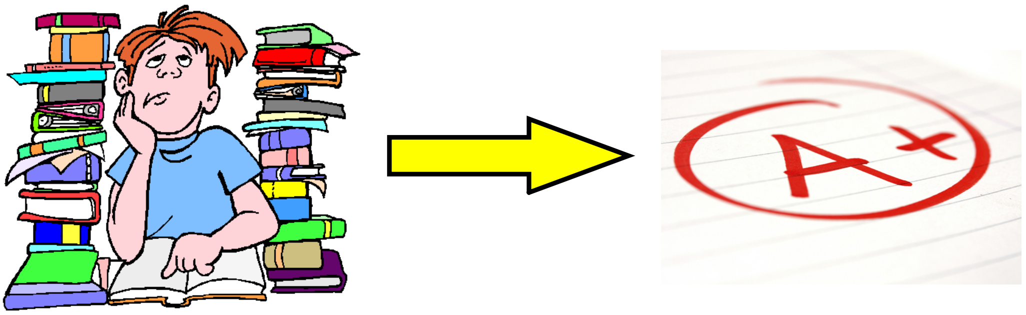
Lesson 4: Scatterplots and Correlation

May 4-1:00 AM

Sometimes we wonder if one event is related to another event

For example, if I study longer, will I get a better grade on my final math examination?



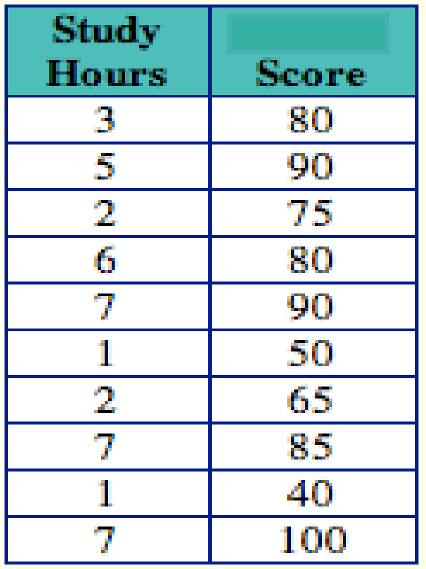
May 2-9:26 PM

Statisticians gather data to determine correlations (relationships) between such events.

Scatter plots will often show at a glance whether a relationship exists between two sets of data.

May 2-9:36 PM

Let's decide if studying longer will affect final exam grades based upon a specific set of data.

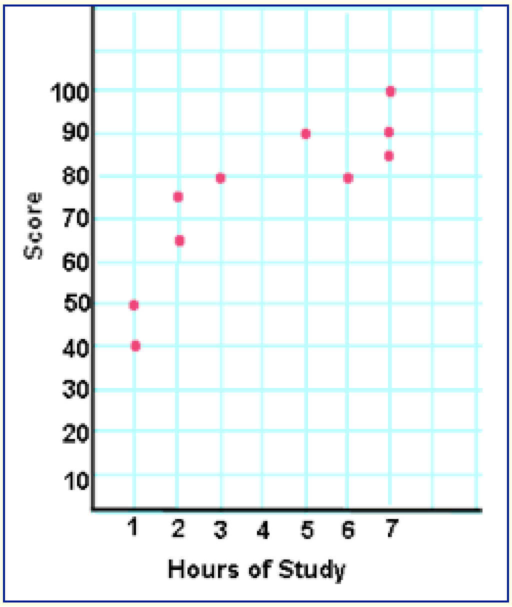


Study Hours	Score
3	80
5	90
2	75
6	80
7	90
1	50
2	65
7	85
1	40
7	100

May 2-9:37 PM

Given the data, a scatter plot has been created to represent the data.

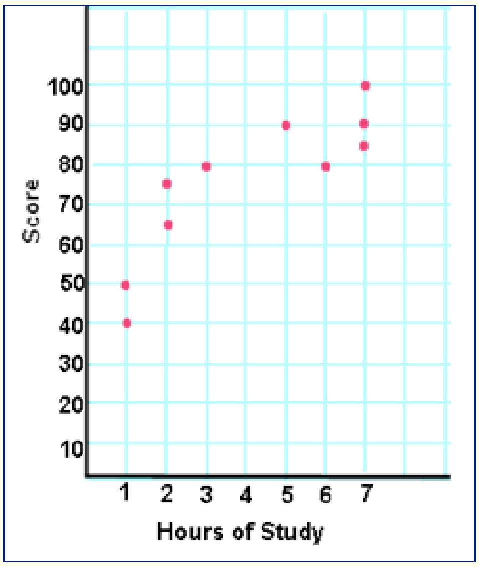
Study Hours	Score
3	80
5	90
2	75
6	80
7	90
1	50
2	65
7	85
1	40
7	100



Remember when making a scatter plot, do NOT connect the dots.

May 2-9:41 PM

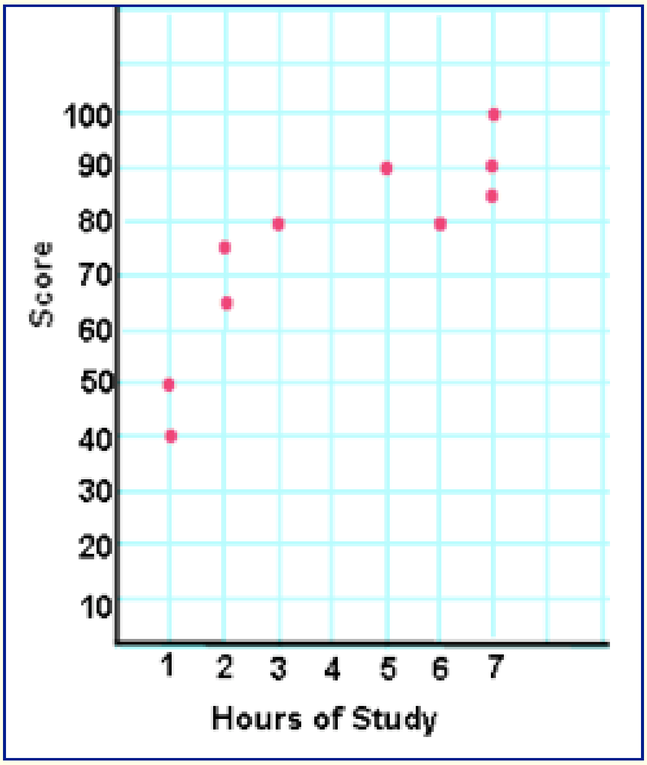
The data displayed on the graph resembles a line rising from left to right.



May 2-9:41 PM

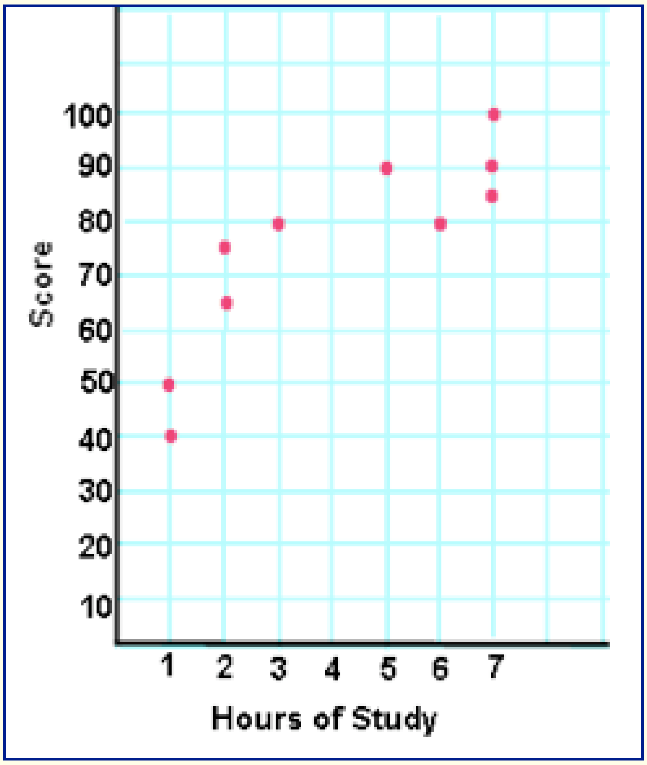
Since the slope of the line is positive, there is a positive correlation between the two sets of data.

As one variable increase, the other variable is also increasing



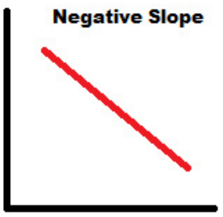
May 2-9:41 PM

This means that according to this set of data, the longer I study, the better grade I will get on my final exam.



May 2-9:41 PM

If the slope of the line had been negative (falling from left to right), a negative correlation would exist



As one variable increases the other decreases.

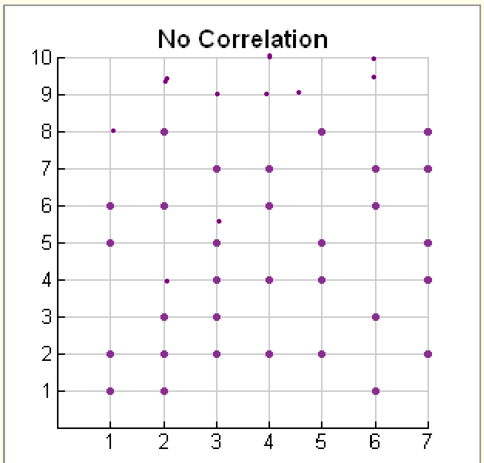
example: As the number of hours that I play x-box increases, my history mark decreases

May 2-9:49 PM

If the plot on the graph is scattered in such a way that it does not approximate a line (it does not appear to rise or fall), there is no correlation between the sets of data.

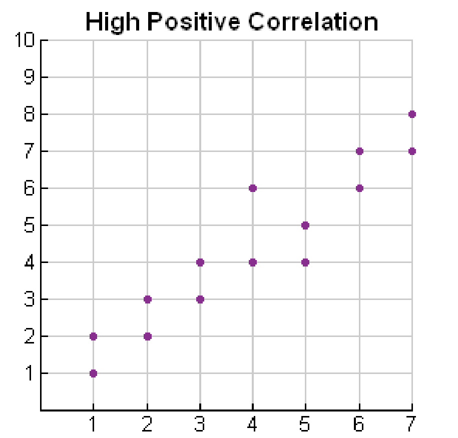
example:

"No correlation" between how many chin ups you can do and how many songs you play on the piano

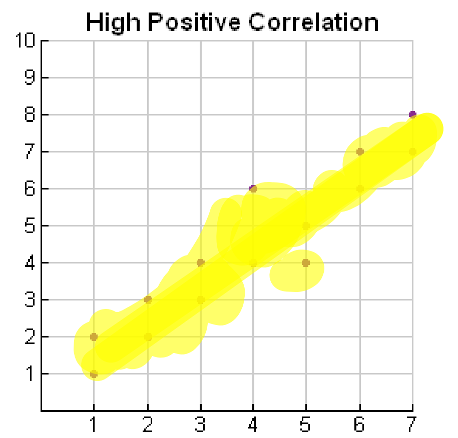


There is no way of determining from these points, if the pattern is rising or falling. There is no evidence of a straight line.

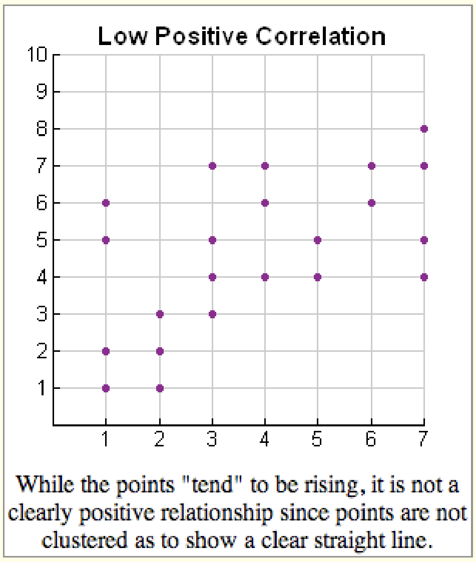
May 2-9:52 PM



The points are clustered as to resemble a rising straight line with a positive slope.

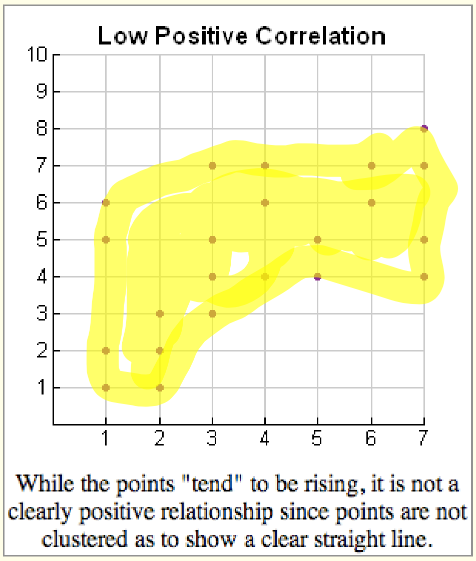


The points are clustered as to resemble a rising straight line with a positive slope.



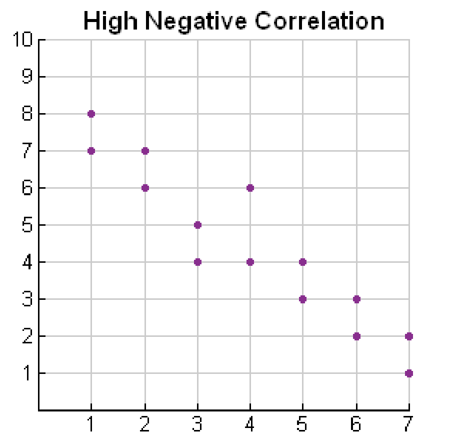
While the points "tend" to be rising, it is not a clearly positive relationship since points are not clustered as to show a clear straight line.

May 2-9:54 PM



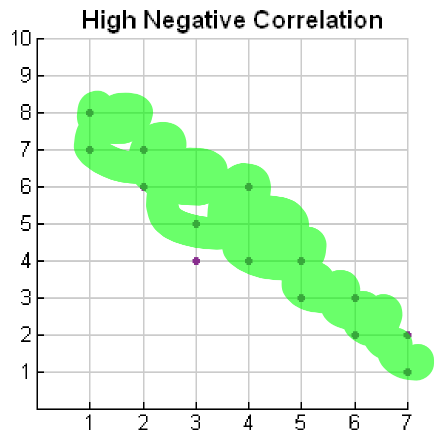
While the points "tend" to be rising, it is not a clearly positive relationship since points are not clustered as to show a clear straight line.

May 2-9:54 PM



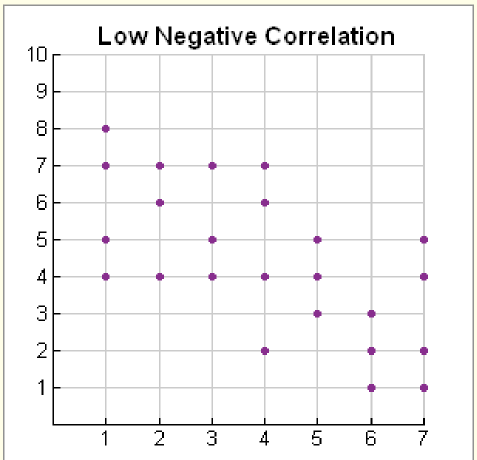
The points are clustered as to resemble a falling straight line with a negative slope.

May 2-9:55 PM



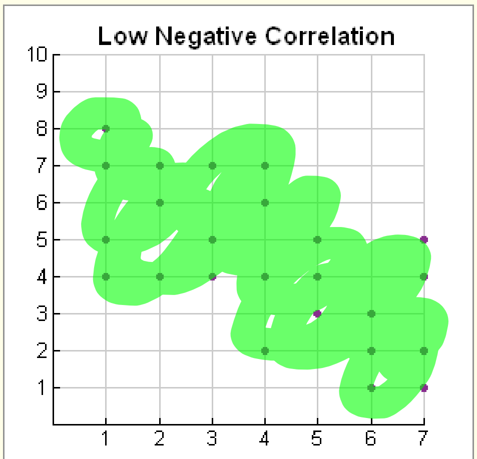
The points are clustered as to resemble a falling straight line with a negative slope.

May 2-9:55 PM



While the points "tend" to be falling, it is not a clearly negative relationship since points are not clustered as to show a clear straight line.

May 2-9:56 PM



While the points "tend" to be falling, it is not a clearly negative relationship since points are not clustered as to show a clear straight line.

May 2-9:56 PM

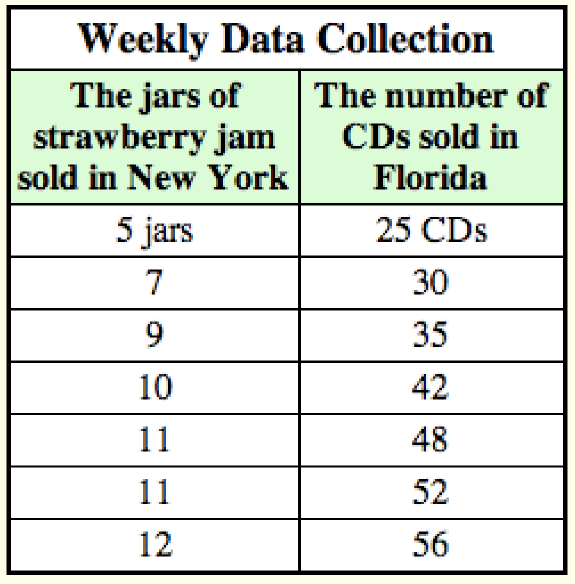
Warning!!

Correlation does not necessarily mean Causation.

Just because there is a strong correlation between data, does not necessarily mean that one set of data is causing the effect that is occurring in the other set of data.

May 2-9:56 PM

During the months of February and March, the weekly number of jars of strawberry jam sold at a local market in New York was recorded. For the same time frame, the number of copies of a popular classical music CD sold in Florida was recorded. The data was examined and was plotted

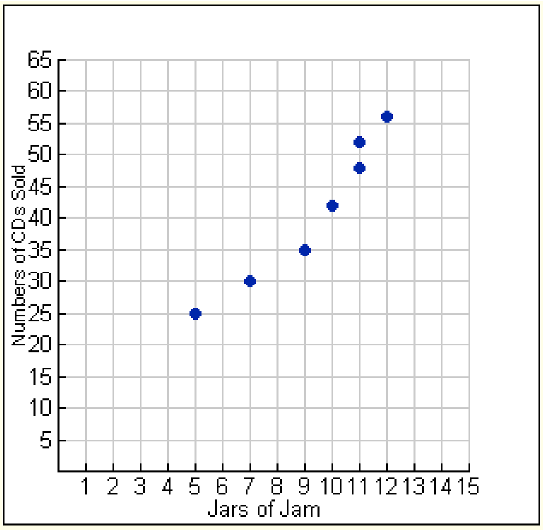


Weekly Data Collection

The jars of strawberry jam sold in New York	The number of CDs sold in Florida
5 jars	25 CDs
7	30
9	35
10	42
11	48
11	52
12	56

May 2-10:10 PM

From looking at the graph, it can be seen that there is a high positive correlation between these two sets of data.



So, do the number of jars of strawberry jam sold in New York cause an increase in the number of classical music CDs sold in Florida?

Of course this is not true!

May 2-10:12 PM

Always be careful what you infer from your statistical analyses.
Be sure the relationship makes sense. Also keep in mind that other factors may be involved in a cause-effect relationship.

May 2-10:10 PM

As ice cream sales increase, the rate of drowning deaths increases sharply.

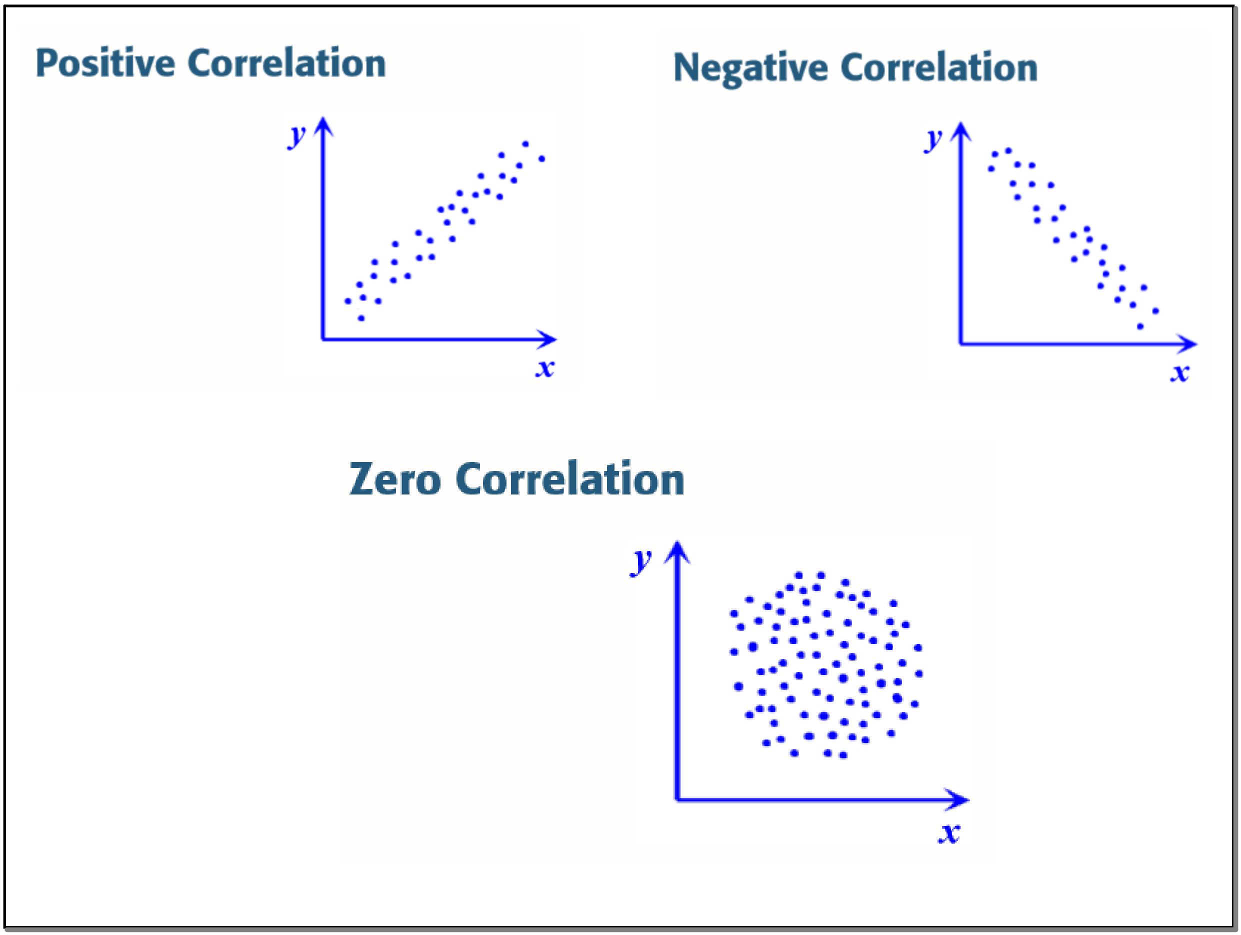
Therefore, ice cream consumption causes drowning. ????

May 2-9:59 PM

With a decrease in the number of pirates, there has been an increase in global warming over the same period.

Therefore, global warming is caused by a lack of pirates. ????

May 2-10:08 PM

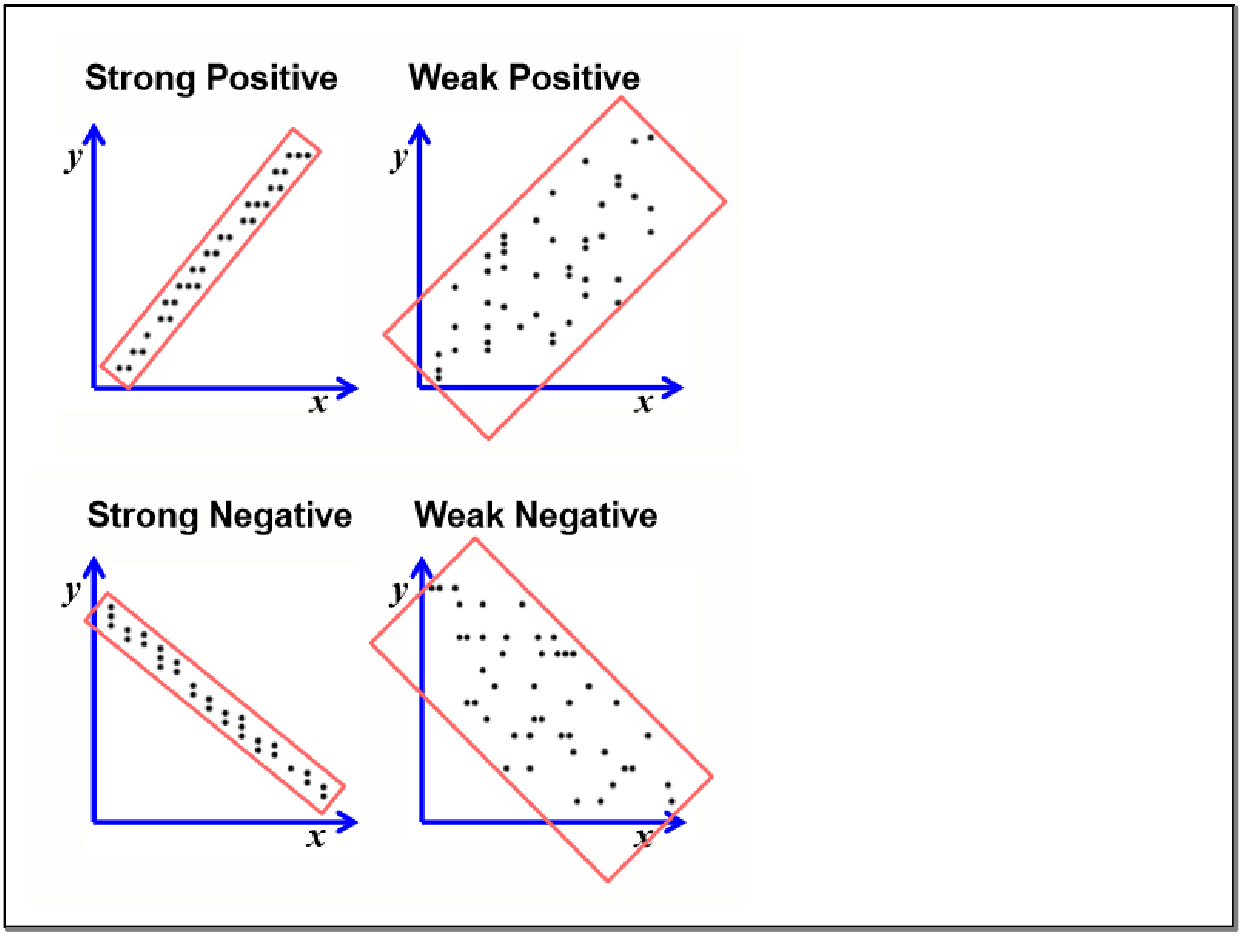


Positive Correlation

Negative Correlation

Zero Correlation

Mar 25-5:38 PM



Strong Positive

Weak Positive

Strong Negative

Weak Negative

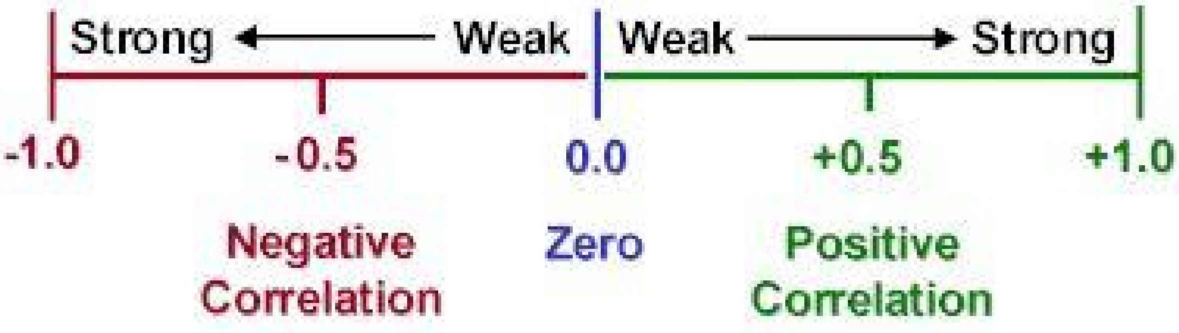
Mar 25-5:40 PM

The strength of the association between two variables is quantified by the correlation coefficient (r)

May 4-12:26 AM

Correlation Coefficient

Shows Strength & Direction of Correlation



$$-1 \le r \le 1$$

Apr 19-1:46 PM

RECAP

The **stronger** the relationship between 2 variables the closer the correlation coefficient will be to **1 or -1**

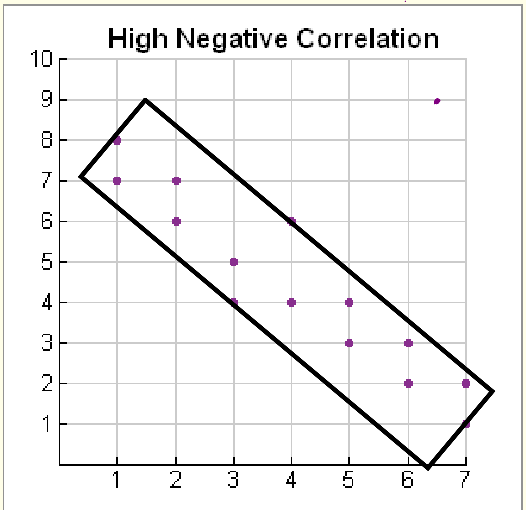
The **weaker** the relationship between 2 variables the closer the correlation coefficient will be to **0**

May 11-10:06 PM

"estimate" correlation coefficient

- draw a rectangle around the data points...ignoring serious **outliers**. The rectangle fits snuggly around all points
- measure both sides of the rectangle
- decide if the "r" is positive or negative

$r = \mp(1 - (\text{shorter side}/\text{longer side}))$



The points are clustered as to resemble a falling straight line with a negative slope.

May 4-12:38 AM

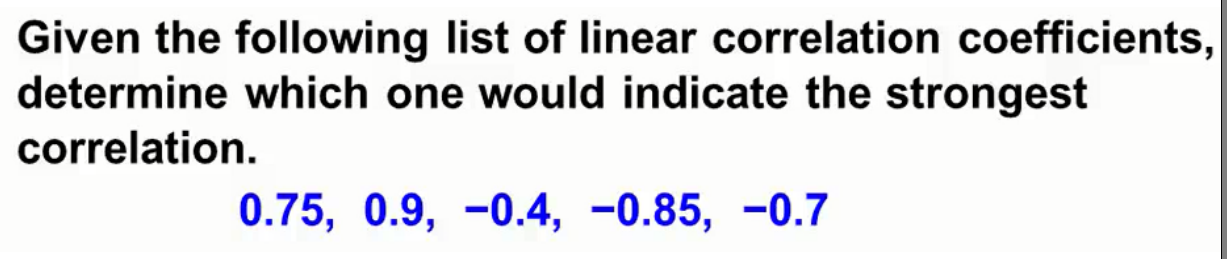
correlation greater than 0.8 → strong

correlation less than 0.5 → weak.

values of r are relative....a lower number (closer to 0) will mean a weaker correlation than a higher number (closer to 1)

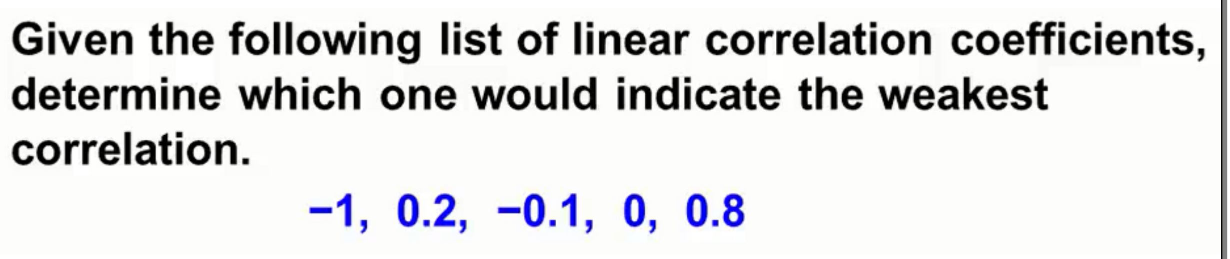
May 4-12:30 AM

example:



Given the following list of linear correlation coefficients, determine which one would indicate the strongest correlation.

0.75, 0.9, −0.4, −0.85, −0.7

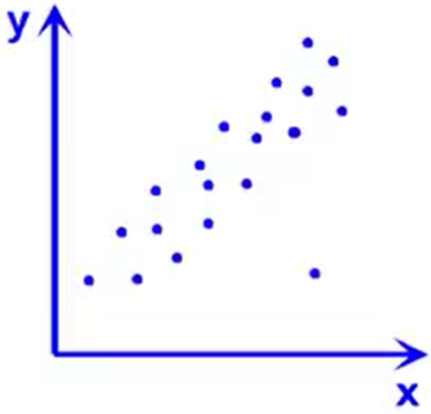


Given the following list of linear correlation coefficients, determine which one would indicate the weakest correlation.

−1, 0.2, −0.1, 0, 0.8

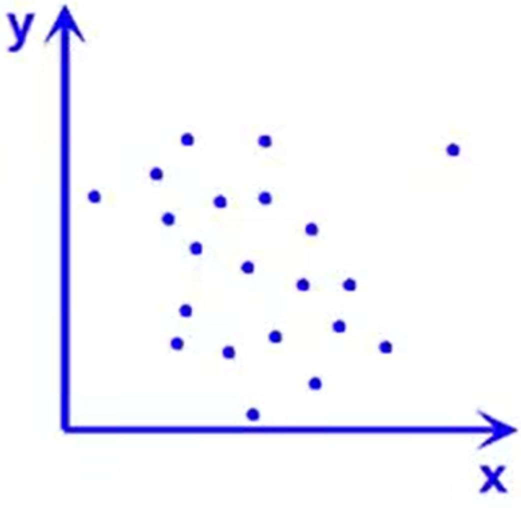
May 2-9:45 PM

example:



Mar 25-6:00 PM

example:



Mar 25-6:36 PM