Correlation coefficient

We saw earlier how to calculate the covariance of two variables, and we learned that the covariance calculation is limited in its usefulness because of the fact that it's very sensitive to units of measure.

In other words, we can find drastically different values for covariance, just by changing the units of measure in a data set, even though we aren't actually changing the data itself.

Pearson correlation coefficient

To fix this problem with covariance, we'll prefer instead to use the **Pearson correlation coefficient**, which we calculate by dividing covariance by the product of the standard deviations of the variables.

Correlation coefficient

Population

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}}{\sigma_x \sigma_y}$$

Sample

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{s_x s_y}$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \qquad s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

The sample correlation coefficient estimates the population correlation coefficient.

The Pearson correlation coefficient is one of several correlation coefficients that can be used to measure correlation. It makes sense to calculate the Pearson correlation coefficient if all of the following conditions are true:

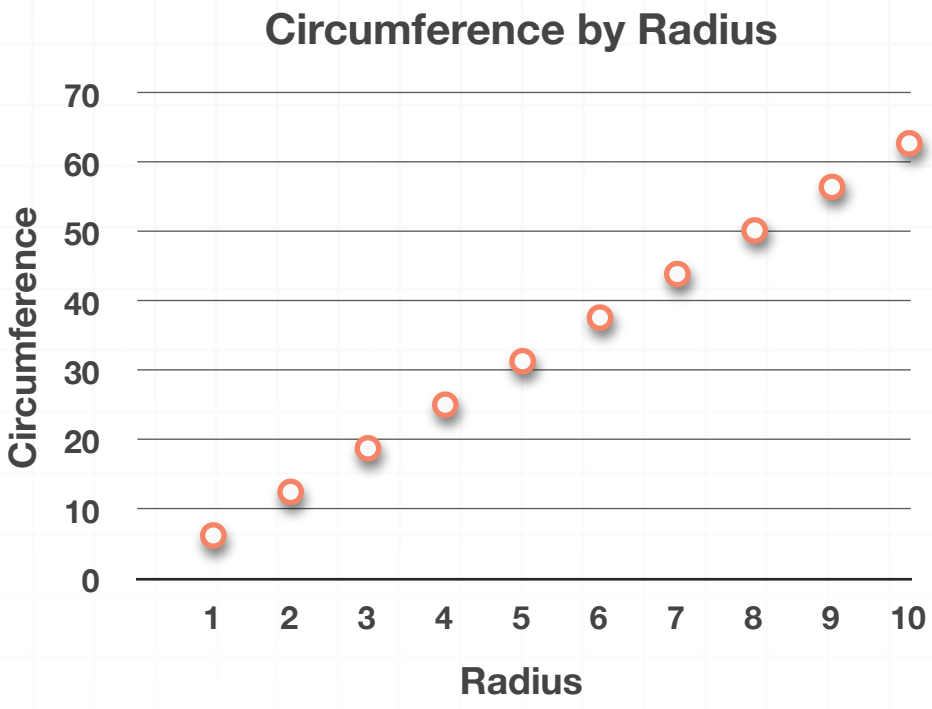
- Both variables are quantitative.
- The variables are normally distributed or close to normally distributed.
- There are no outliers. The presence of outliers may significantly skew the data and the resulting Pearson correlation coefficient will not accurately reflect the correlation of the two variables.
- The relationship between the variables is linear. Pearson correlation is best for data sets that with a reasonably straight trend line.

Interpreting correlation

We can think about the **correlation** of two variables as the degree of the relationship between them. We talk about correlation when change in one variable results in a change in the other variable.

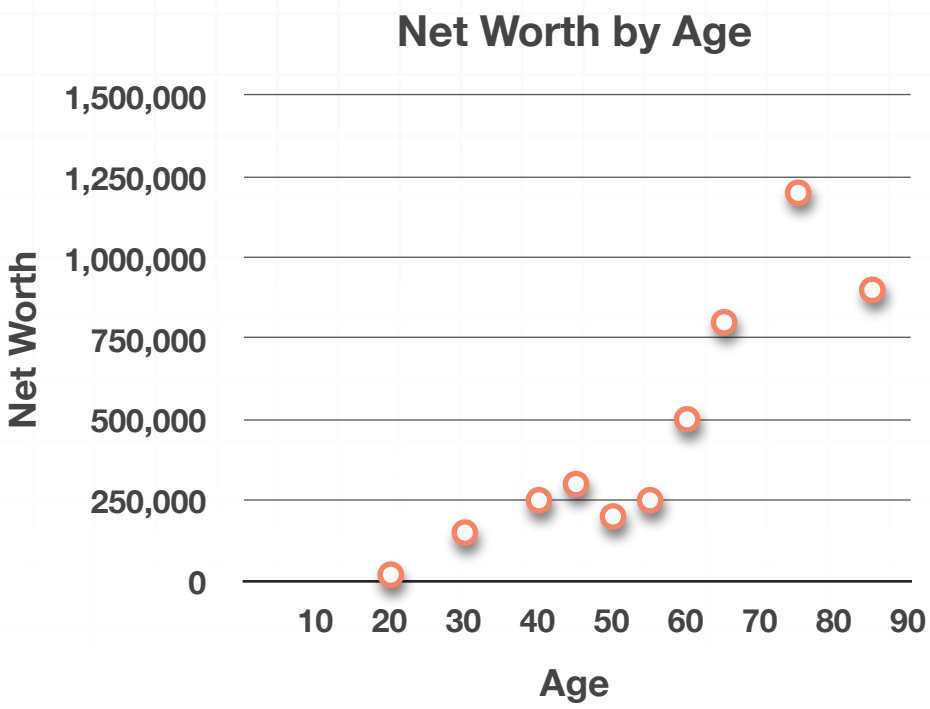
The circumference and radius of a circle are **perfectly correlated** because those values will always satisfy the equation $C = 2\pi r$ exactly. There's perfect correlation between C and r because the length of the radius

perfectly predicts the circumference of the circle, and vice versa. So a data set of radius and circumference measurements might look like this:



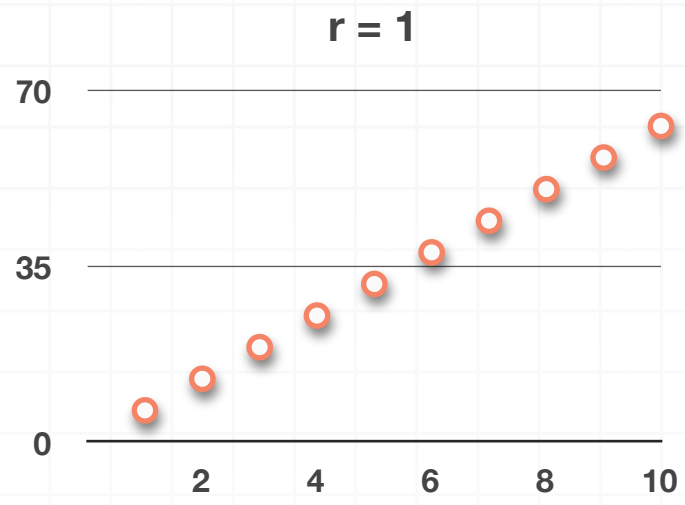
On the other hand, there's no correlation between the values on a pair of dice. No matter how many times we toss the dice, the value we get on one die is completely **uncorrelated** with the value we get on the other die. So a data set of dice rolls might look like this:

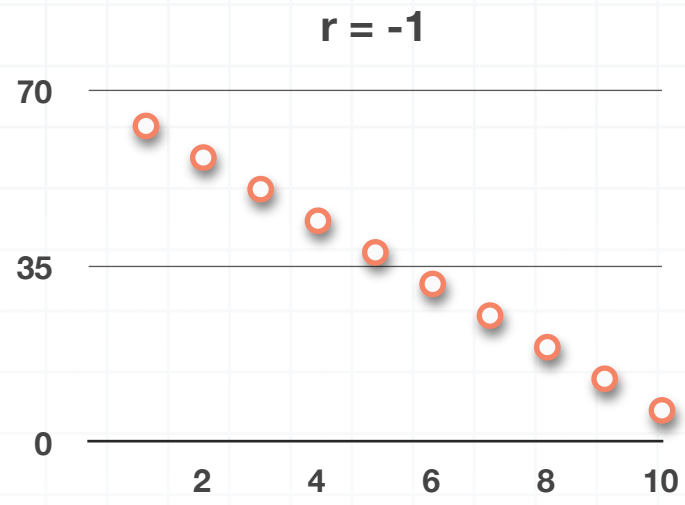
In between no correlation and perfect correlation, some variables are **somewhat correlated**, like age and net worth. Typically, people tend to start out with very little money when they're young, accumulate more money and assets as they get older, and then start spending down those assets in older age. So while we can observe a general trend between these variables (some correlation), we can't predict exact net worth solely by age. And therefore a data set of net worth by age might look like this:



Range of the correlation coefficient

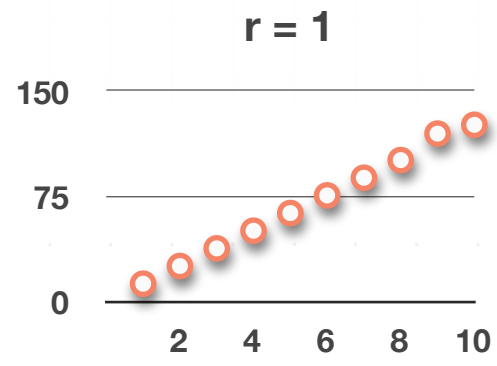
If the points in the data set all fall on a line with a positive slope, the correlation coefficient of the data set will be $r = 1$, and if the points all fall on a line with a negative slope, the correlation coefficient will be $r = -1$.

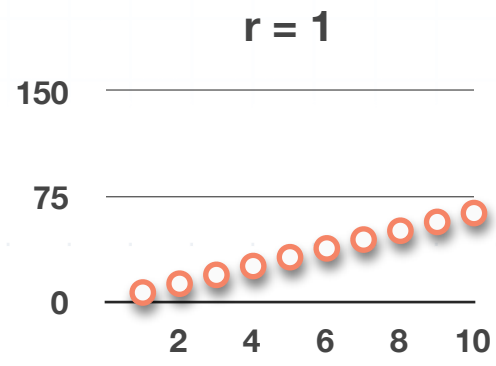


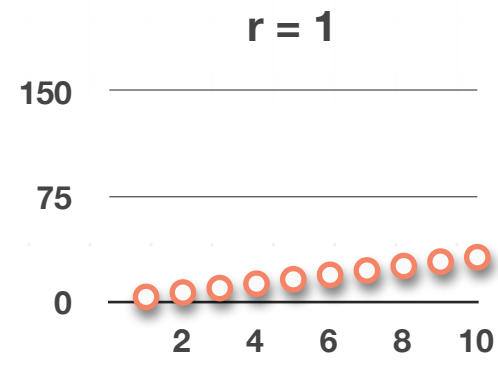


This is not the same as saying that the data all fall on a line with a slope of $m = 1$, or that they all fall on a line with a slope of $m = -1$. The exact slope of the line isn't relevant. All the matters is whether the line is positively sloped or negatively sloped, regardless of exactly how steep it is in either direction.

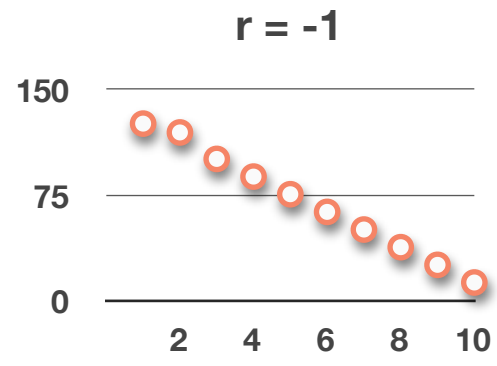
So $r = 1$ indicates a perfect positive, or direct, linear relationship,

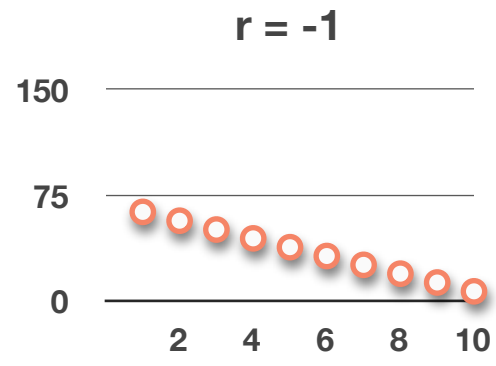


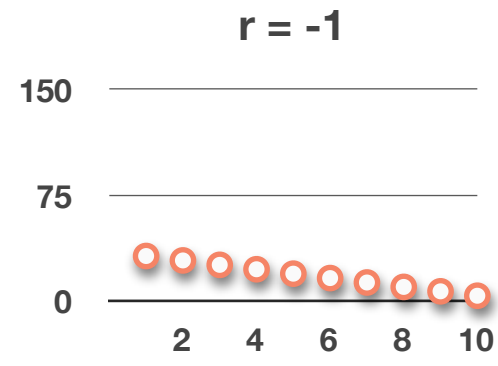




while $r = -1$ indicates a perfect negative, or inverse, linear relationship,

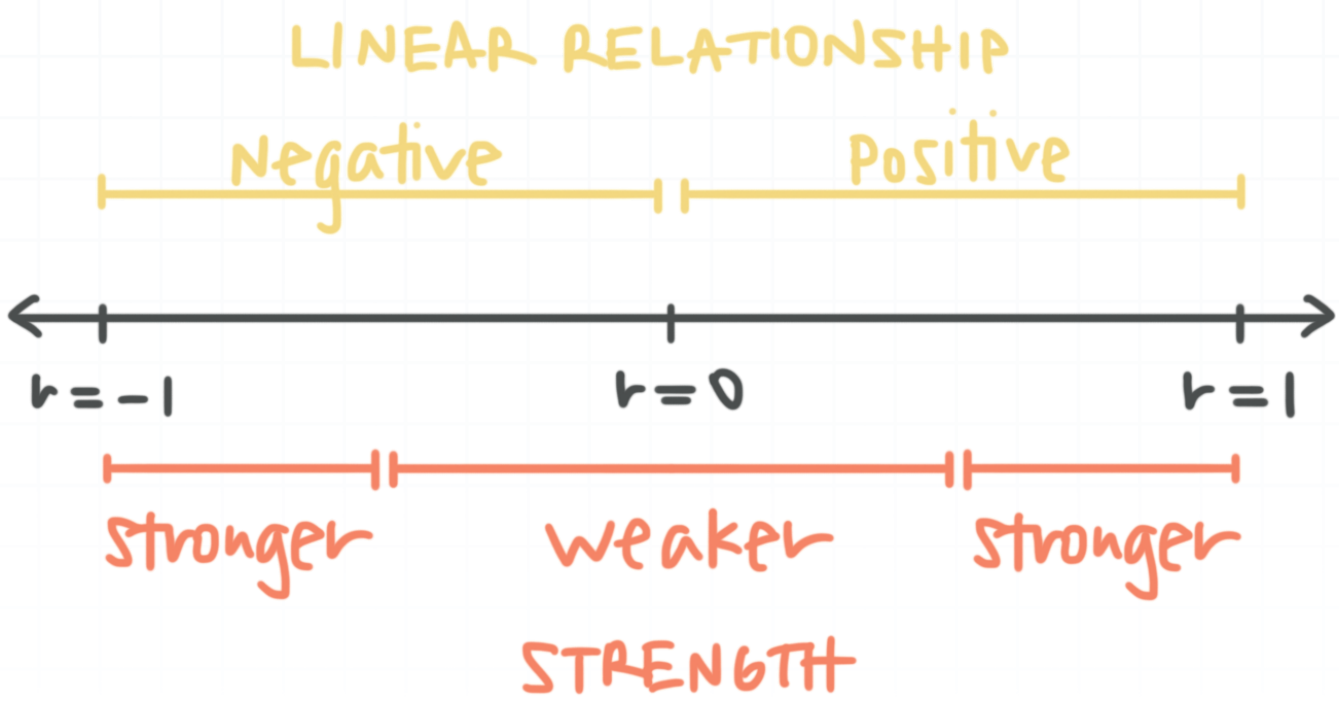






A value of r between 0 and 1 indicates a positive linear relationship, with stronger positive relationships having a correlation coefficient closer to 1 and weaker ones having a correlation coefficient closer to 0.

A value of r between 0 and -1 indicates a negative linear relationship, with stronger negative relationships having a correlation coefficient closer to -1 and weaker ones having a correlation coefficient closer to 0.



In other words, the correlation coefficient indicates both the strength and direction of the relationship between two variables. In general,

- Any value of the correlation coefficient between -0.70 and -1 (or between 0.70 and 1) indicates a strong negative (or positive) correlation.
- Any value of the correlation coefficient between -0.30 and -0.70 (or between 0.30 and 0.70) indicates a moderate negative (or positive) correlation.

- Any value of the correlation coefficient between 0 and -0.30 (or between 0 and 0.30) indicates a weak negative (or positive) correlation.

Correlation is not causation

Despite the fact that the correlation coefficient can be a useful measure, we have to remember that any correlation we find between variables is not equivalent to causation between the variables.

In other words, correlation coefficients close to $r = -1$ or $r = 1$ don't necessarily indicate that a change in one variable specifically causes a change in the other variable.

Let's look at an example where we calculate the Pearson correlation coefficient for a data set.

Example

Calculate the correlation coefficient of the sample, then interpret the result.

Age	Net Worth
20	$20,000
30	$150,000
40	$250,000
45	$300,000
50	$200,000
55	$250,000
60	$500,000
65	$800,000
75	$1,200,000
85	$900,000

Start by calculating mean age,

$$\bar{x} = \frac{20 + 30 + 40 + 45 + 50 + 55 + 60 + 65 + 75 + 85}{10}$$

$$\bar{x} = \frac{525}{10}$$

$$\bar{x} = 52.5$$

and mean net worth.

$$\bar{y} = \frac{10{,}000(2 + 15 + 25 + 30 + 20 + 25 + 50 + 80 + 120 + 90)}{10}$$

$$\bar{y} = \frac{4{,}250{,}000}{10}$$

$$\bar{y} = 425{,}000$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (20 - 52.5)(20{,}000 - 425{,}000)$$

$$+(30 - 52.5)(150{,}000 - 425{,}000) + (40 - 52.5)(250{,}000 - 425{,}000)$$

$$+(45 - 52.5)(300{,}000 - 425{,}000) + (50 - 52.5)(200{,}000 - 425{,}000)$$

$$+(55 - 52.5)(250{,}000 - 425{,}000) + (60 - 52.5)(500{,}000 - 425{,}000)$$

$$+(65 - 52.5)(800{,}000 - 425{,}000) + (75 - 52.5)(1{,}200{,}000 - 425{,}000)$$

$$+(85 - 52.5)(900{,}000 - 425{,}000)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (-32.5)(-405{,}000) + (-22.5)(-275{,}000)$$

$$+(-12.5)(-175{,}000) + (-7.5)(-125{,}000) + (-2.5)(-225{,}000)$$

$$+2.5(-175{,}000) + 7.5(75{,}000) + 12.5(375{,}000) + 22.5(775{,}000)$$

$$+32.5(475{,}000)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 60{,}725{,}000$$

$$s_{xy} = \frac{60{,}725{,}000}{10 - 1}$$

$$s_{xy} \approx 6{,}747{,}222.22$$

Next we'll need the standard deviation for age,

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (20-52.5)^2 + (30-52.5)^2 + (40-52.5)^2$$

$$+(45-52.5)^2 + (50-52.5)^2 + (55-52.5)^2 + (60-52.5)^2$$

$$+(65-52.5)^2 + (75-52.5)^2 + (85-52.5)^2$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (-32.5)^2 + (-22.5)^2 + (-12.5)^2 + (-7.5)^2$$

$$+(-2.5)^2 + 2.5^2 + 7.5^2 + 12.5^2 + 22.5^2 + 32.5^2$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 1{,}056.25 + 506.25 + 156.25 + 56.25$$

$$+6.25 + 6.2556.25 + 156.25 + 506.25 + 1{,}056.25$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 3{,}562.5$$

$$s_x = \sqrt{\frac{3{,}562.5}{10-1}}$$

$$s_x \approx 19.8956$$

and the standard deviation for net worth.

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (20{,}000 - 425{,}000)^2 + (150{,}000 - 425{,}000)^2$$

$$+(250{,}000 - 425{,}000)^2 + (300{,}000 - 425{,}000)^2$$

$$+(200{,}000 - 425{,}000)^2 + (250{,}000 - 425{,}000)^2$$

$$+(500{,}000 - 425{,}000)^2 + (800{,}000 - 425{,}000)^2$$

$$+(1{,}200{,}000 - 425{,}000)^2 + (900{,}000 - 425{,}000)^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (-405{,}000)^2 + (-275{,}000)^2 + (-175{,}000)^2$$

$$+(-125{,}000)^2 + (-225{,}000)^2 + (-175{,}000)^2 + (75{,}000)^2$$

$$+(375{,}000)^2 + (775{,}000)^2 + (475{,}000)^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 1{,}339{,}650{,}000{,}000$$

$$s_y = \sqrt{\frac{1{,}339{,}650{,}000{,}000}{10-1}}$$

$$s_y \approx 385{,}810.83$$

Now we can plug the covariance into the formula for correlation.

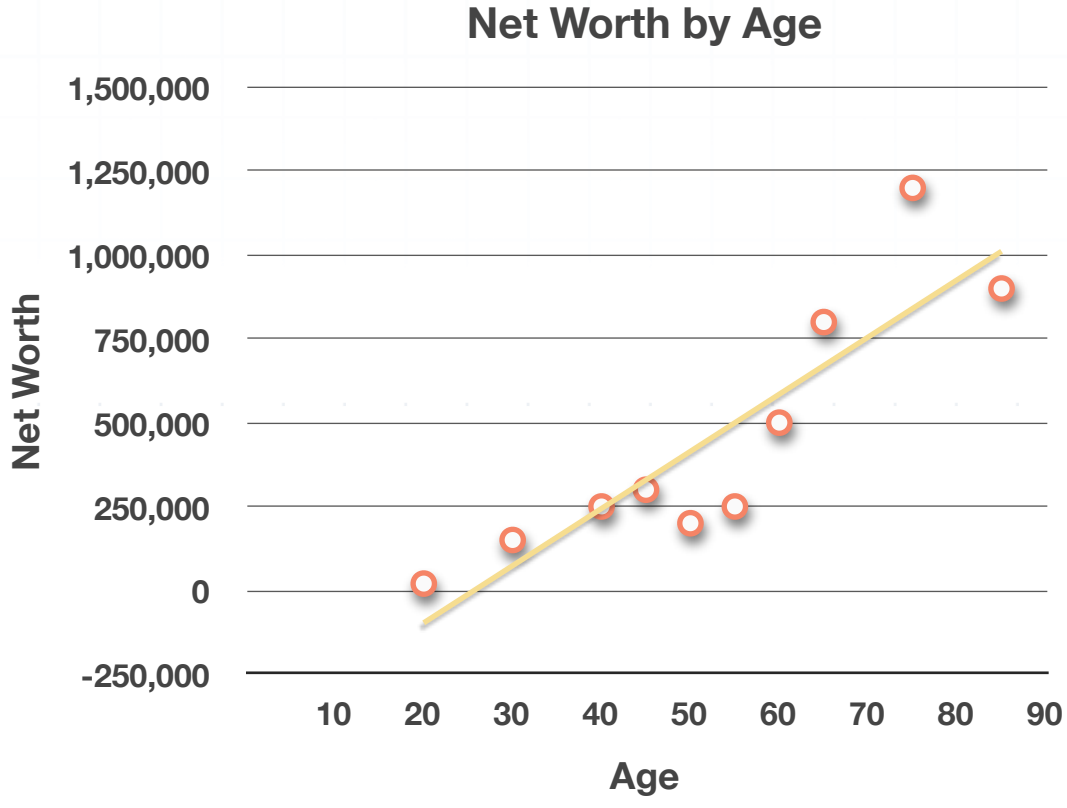
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{6{,}747{,}222.22}{(19.8956)(385{,}810.83)}$$

$$r_{xy} \approx 0.8790$$

Any Pearson correlation coefficient between $r = 0$ and $r = 1$ indicates a positive linear relationship between the variables, and since $r \approx 0.879$ is quite close to $r = 1$, it indicates a strong positive linear relationship. Which means that when one variable increases, the other variable tends to increase as well.

We see the strong positive linear relationship indicated by $r \approx 0.879$ in a plot of the data set.



However, as we learned earlier, correlation doesn't necessarily imply causation. Even if two variables are highly correlated, it doesn't prove that one causes the other. There may be other variables or factors that are influencing the relationship.