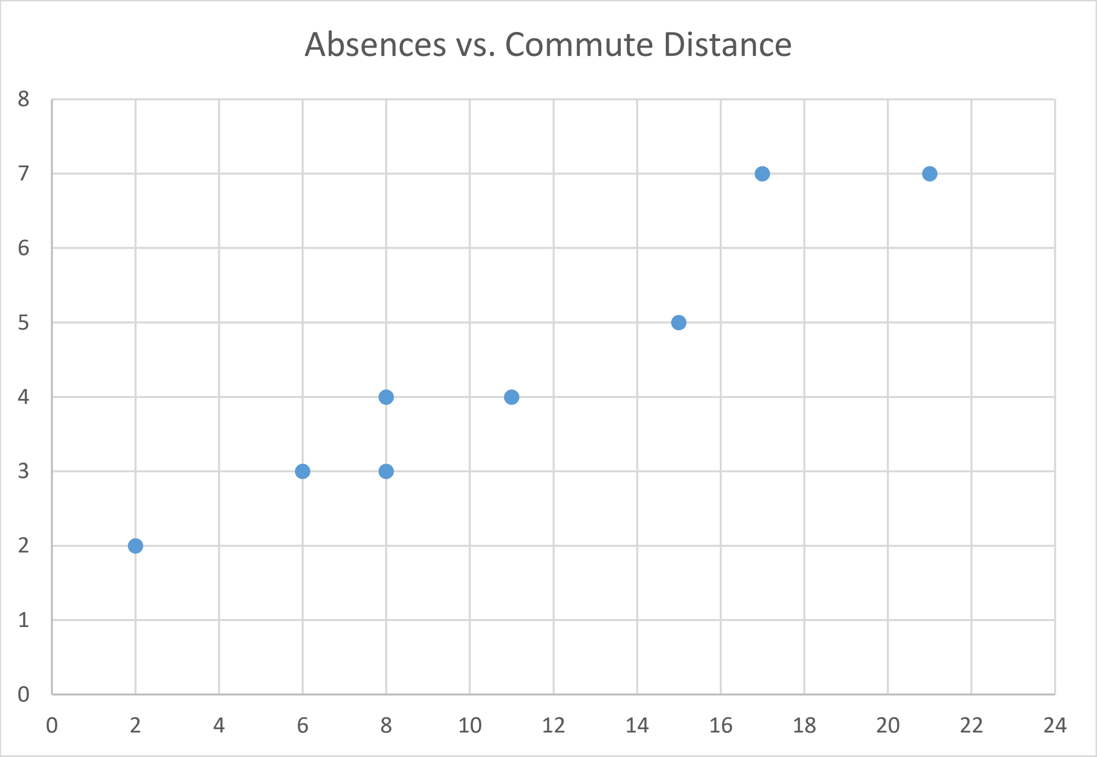
Paired Data and Scatterplots:

When data consists of pairs of values, it’s sometimes useful to plot them as points called a scatterplot.

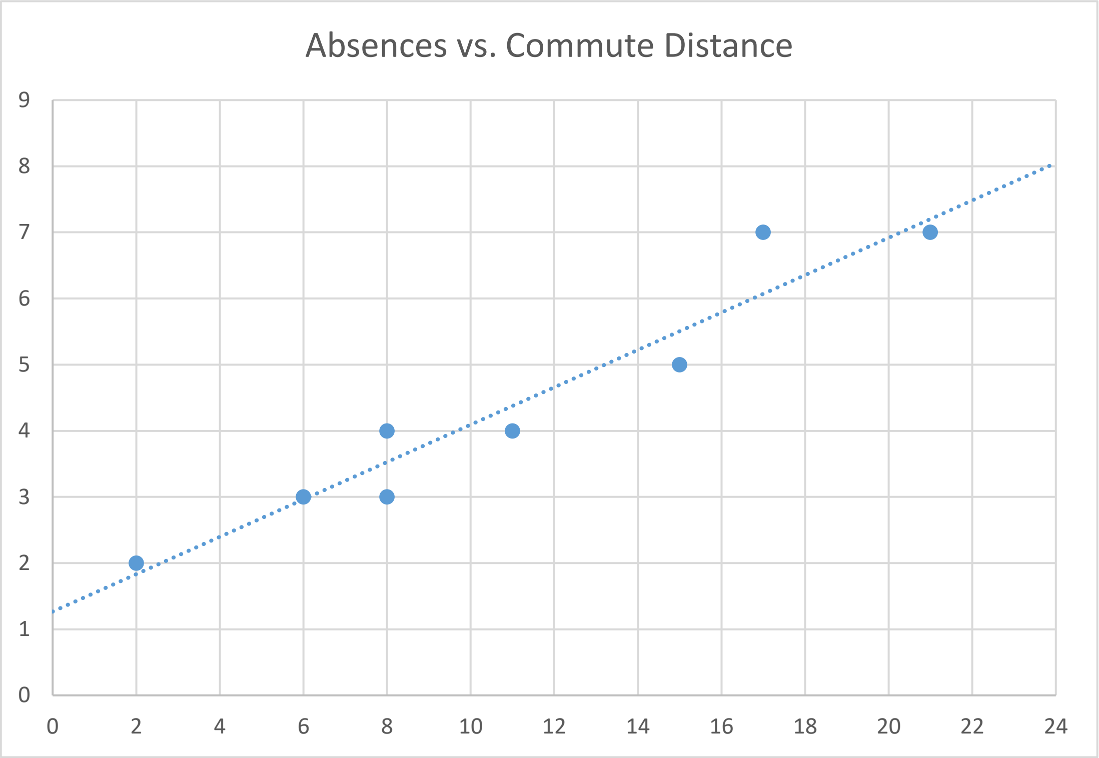
A company recorded the commuting distance in miles and number of absences in days for a group of its employees over the course of a year.

Commuting Distance	Number of Absences
8	4
21	7
6	3
8	3
2	2
15	5
17	7
11	4

Here’s the scatterplot of number of absences vs. commuting distance.



Scatterplots can help reveal relationships between the variables being measured in the paired data. The most commonly searched for relationship is a linear relationship. In this example the data values do appear to cluster on a line.



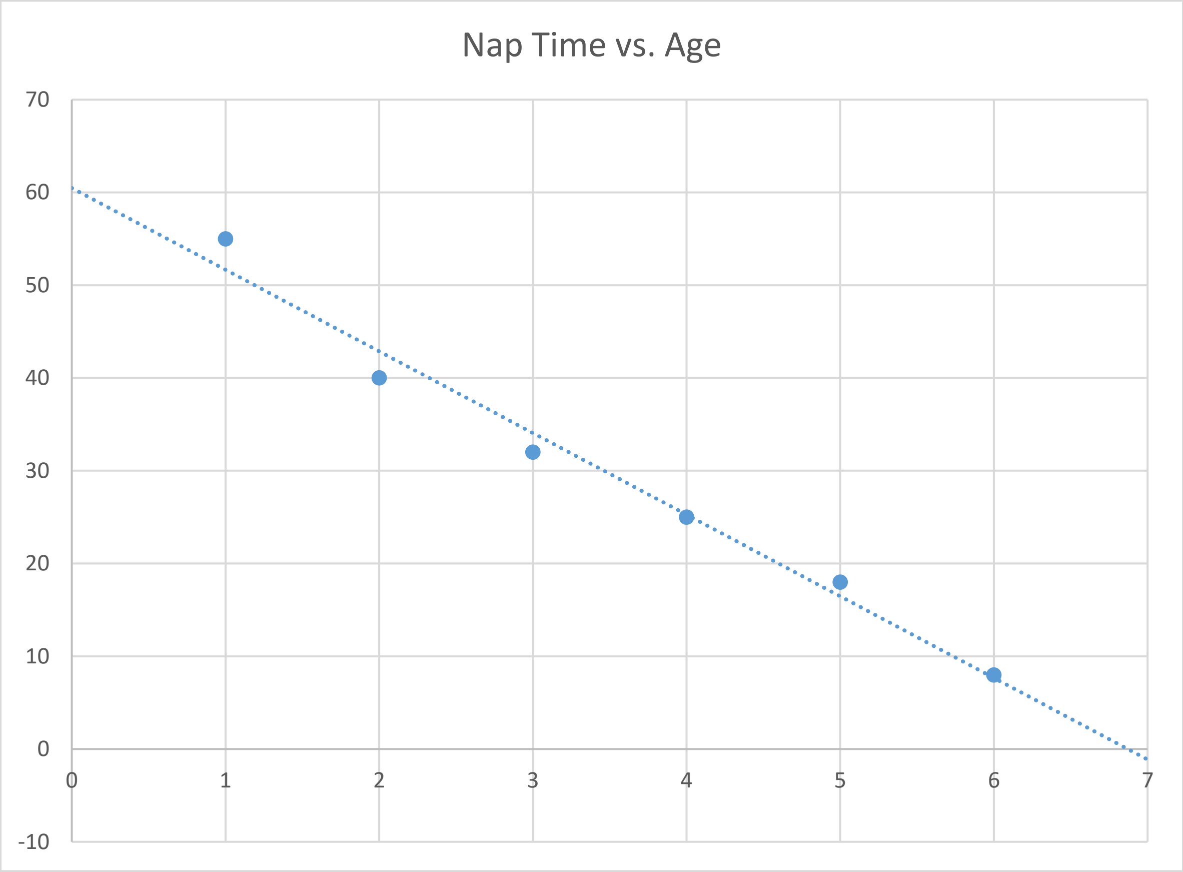
When the points cluster on a line, the variables being measured are said to be linearly correlated. If the line they cluster on has a positive slope, then the variables are said to be positively correlated. If the cluster line has a negative slope, then the variables are said to be negatively correlated. If the points don't cluster on a non-vertical, non-horizontal line, then the variables are said to be uncorrelated.

When variables are positively correlated, larger values of one variable are associated with larger values of the other variable. When variables are negatively correlated, larger values of one variable are associated with smaller values of the other variable.

Here is another paired data set of Nap time in minutes along with Age of child in years.

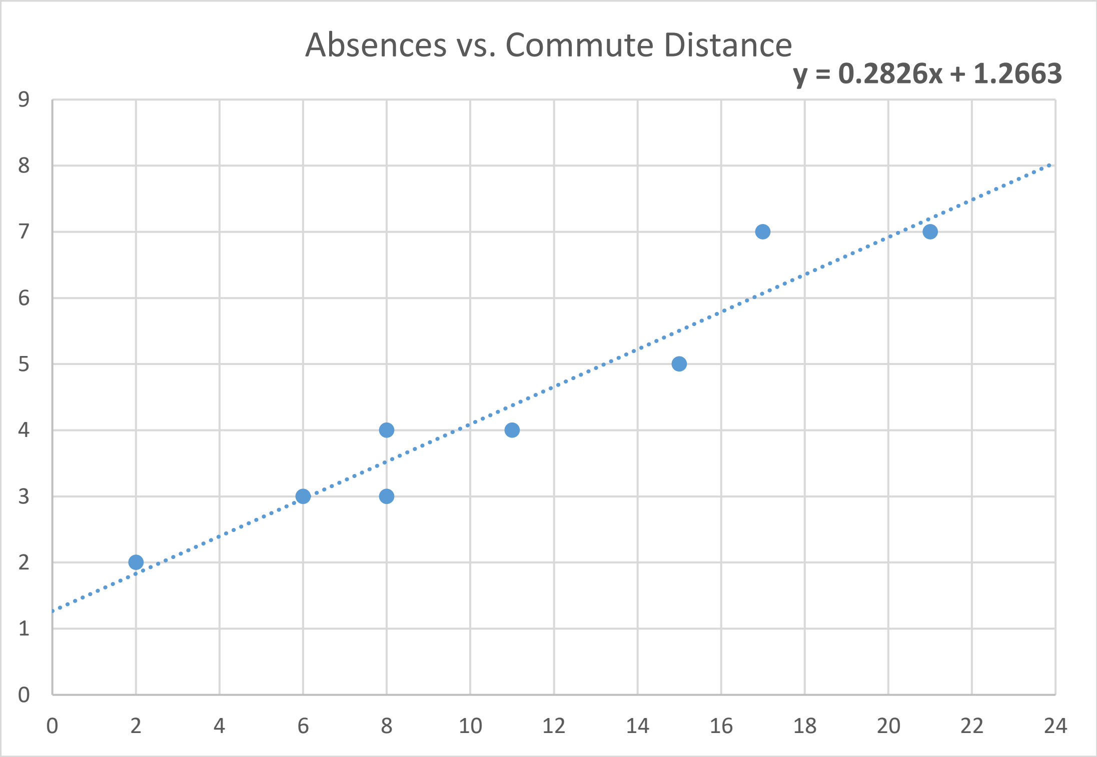
Age(years)	Nap time(minutes)
1	55
2	40
3	32
4	25
5	18
6	8

Here’s its scatterplot:



Nap time and age appear to be negatively(linearly) correlated for this group of children.

The official name of the line of best fit(cluster line) is regression line. It’s equation can be determined using statistical software.



The equation of this regression line is $y = .2826x + 1.2663$**, where** ***x*** **represents the commuting distance, and** ***y*** **represents the number of absences.**

Sometimes the equation of the regression line or its graph is used to make predictions about a value of the variables that wasn't measured.

Use the regression line equation to answer the following:

What's the predicted number of absences for an employee with a 10 mile commute?

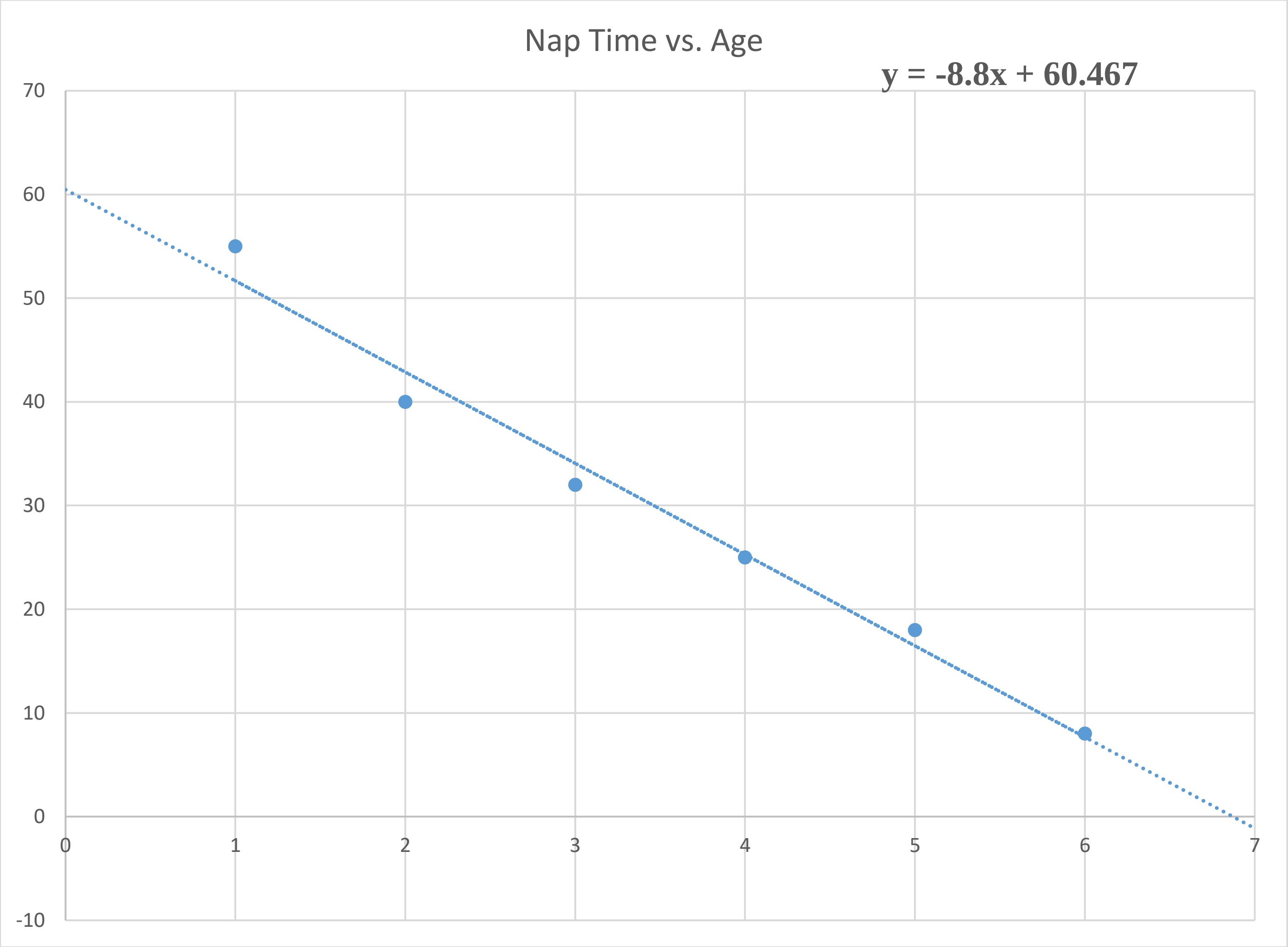
$y = .2826(10) + 1.2663 = 4.0923 \approx 4 days$ ***(interpolation, safe)***

This prediction is called an interpolation because the 10 miles falls within the distance values recorded in the data set. Interpolation predictions are considered to be safe and generally reliable.

What's the predicted number of absences for an employee with a 24 mile commute?

$y = .2826(24) + 1.2663 = 8.0487 \approx 8 days$ ***(extrapolation, dangerous)***

This prediction is called an extrapolation because the 24 miles falls outside of the distance values recorded in the data set. Extrapolation predictions are considered to risky and generally unreliable.


Nap Time vs. Age
y = -8.8x + 60.467
70
60
50
40
30
20
10
0
-10
0
1
2
3
4
5
6
7

Use the regression line equation to answer the following:

What's the predicted nap time for a 1½ year-old?

$$y = -8.8(1.5) + 60.467 = 53.167 \approx 53 \text{ minutes}$$

Is this prediction an interpolation or extrapolation?

Interpolation

What's the predicted nap time for a 7 year-old?

$$y = -8.8(7) + 60.467 = -1.133 \text{ minutes}$$

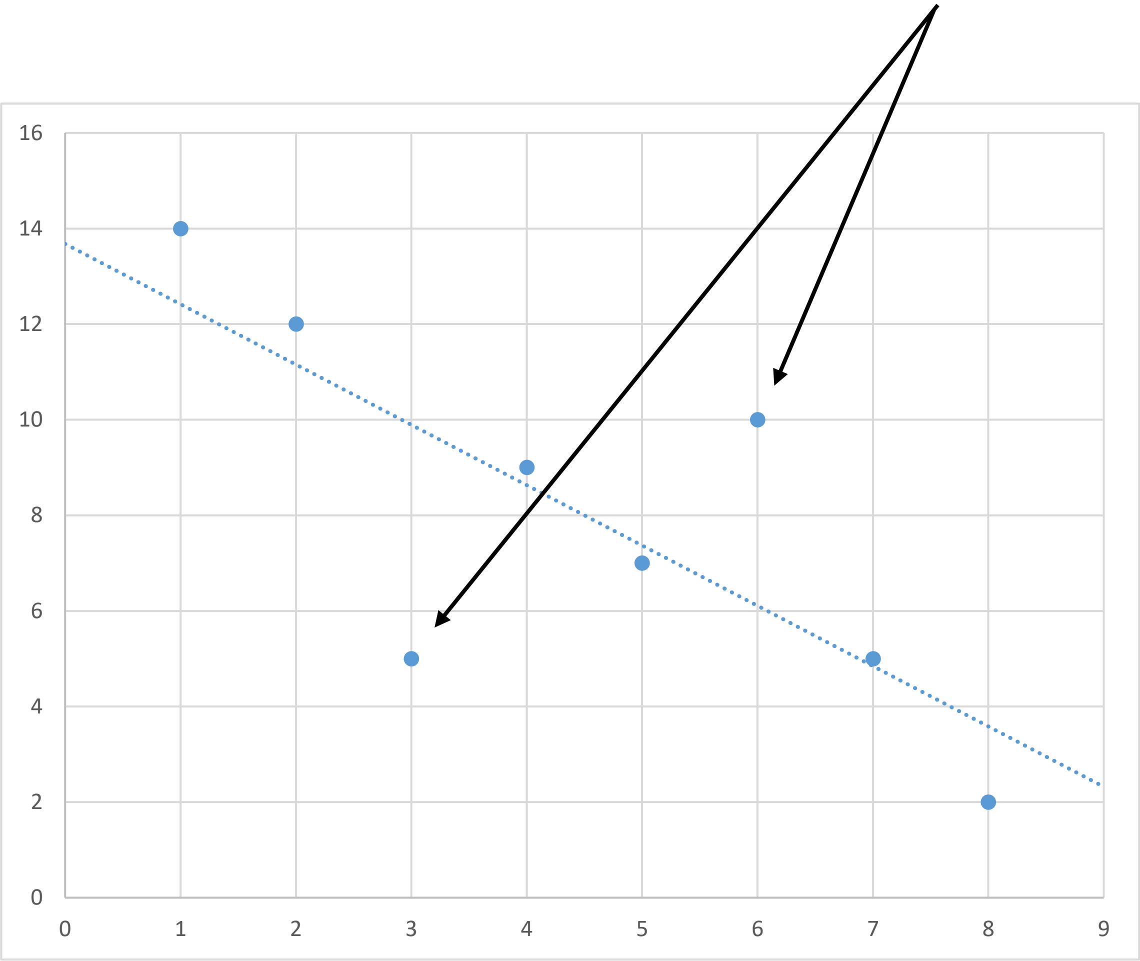
Is this prediction an interpolation or extrapolation?

Extrapolation

Is something wrong with this prediction? Explain.

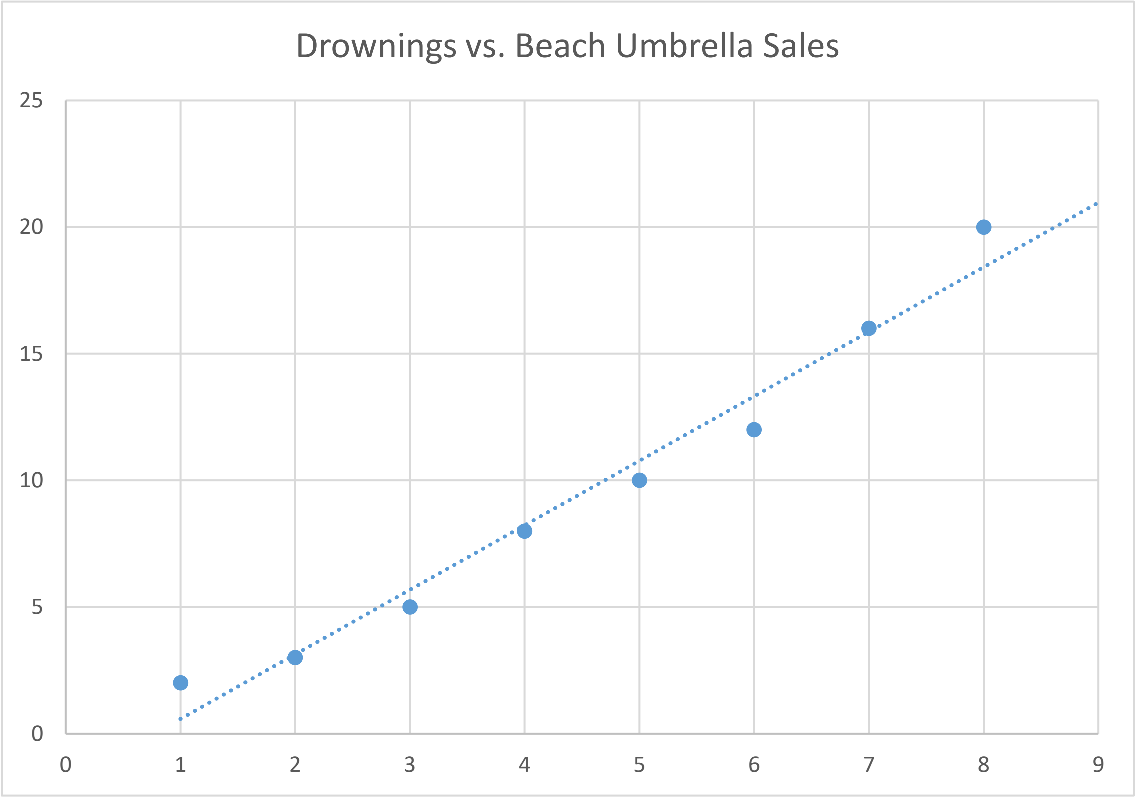
Yes, negative nap times make no sense.

Sometimes most of the points cluster on a line while a few seem to resist the linear trend. The points that resist clustering are referred to as outliers.



Sometimes outliers are attributed to measurement error, but not always.

<u>Just because variables are correlated doesn’t mean that they are causally related!</u>

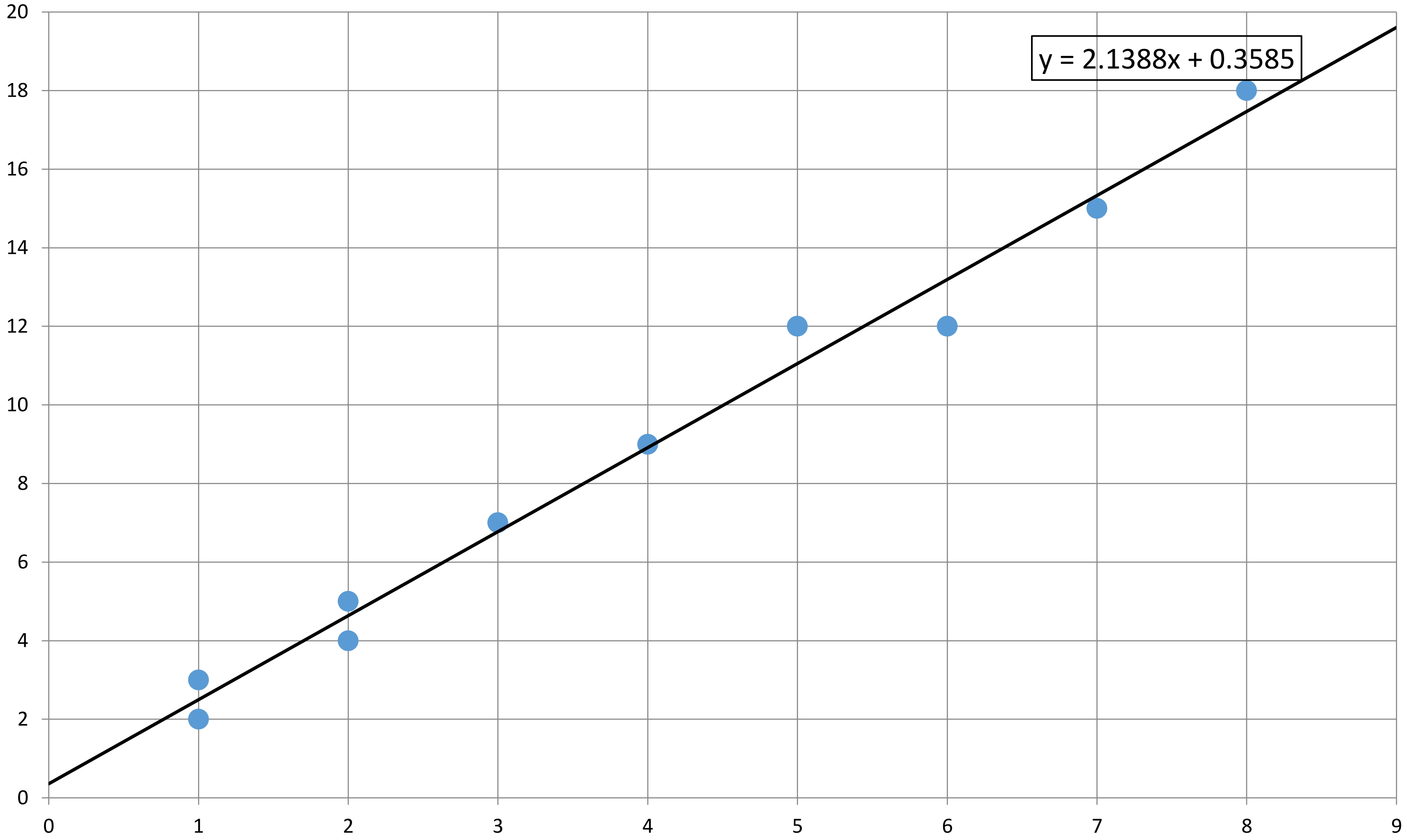


I wouldn’t conclude that beach umbrellas cause drowning, but both are correlated to temperature/season.

Here’s a set of paired data.

X	Y
1	3
2	5
2	4
1	2
4	9
5	12
6	12
3	7
7	15
8	18

Here's its scatterplot along with a regression line from Excel.

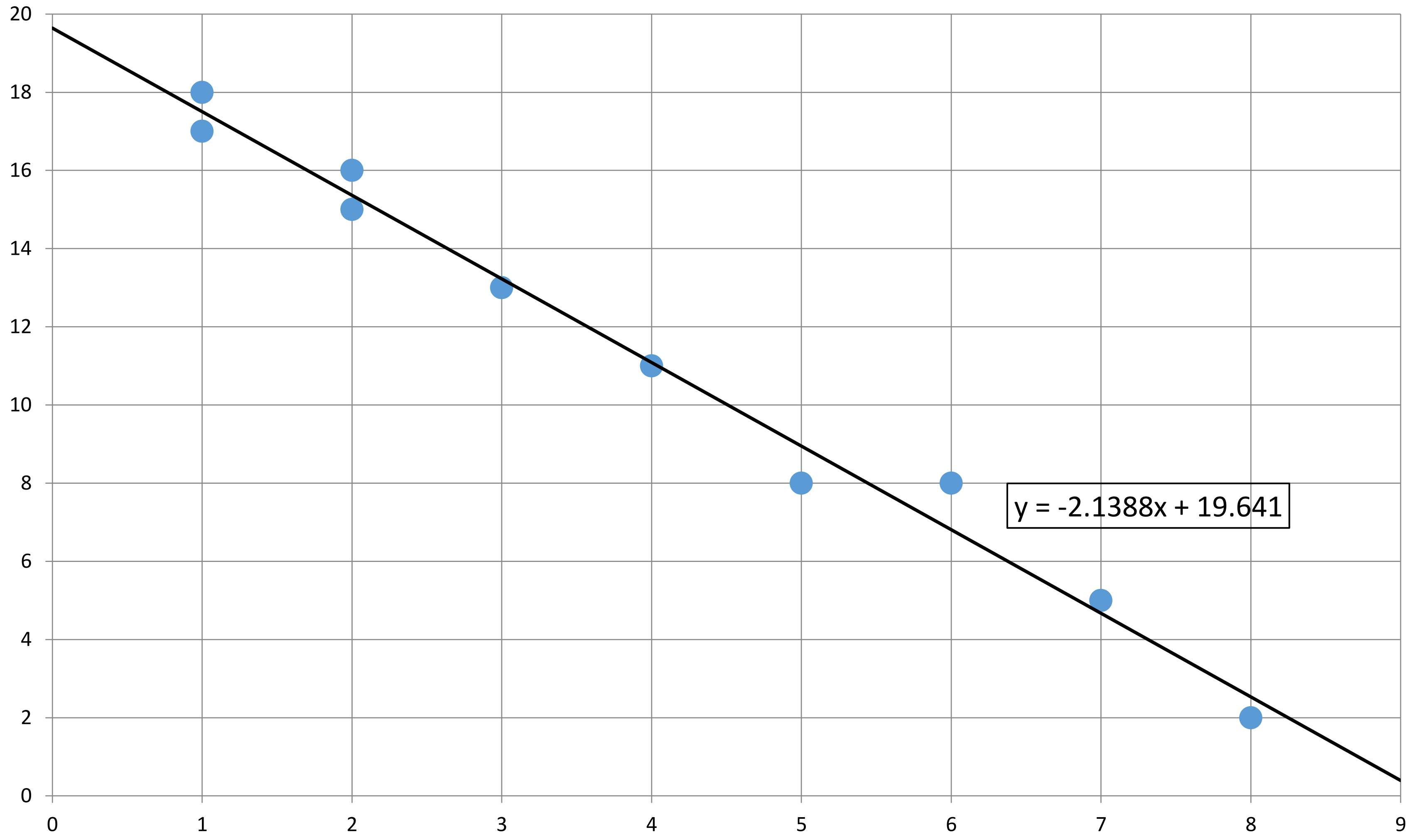


Would you say that the variables are positively correlated, negatively correlated, or uncorrelated? They cluster on a positively sloped line, so positively correlated.

Here’s a set of paired data.

X	Y
1	17
2	15
2	16
1	18
4	11
5	8
6	8
3	13
7	5
8	2

Here's its scatterplot along with a regression line from Excel.

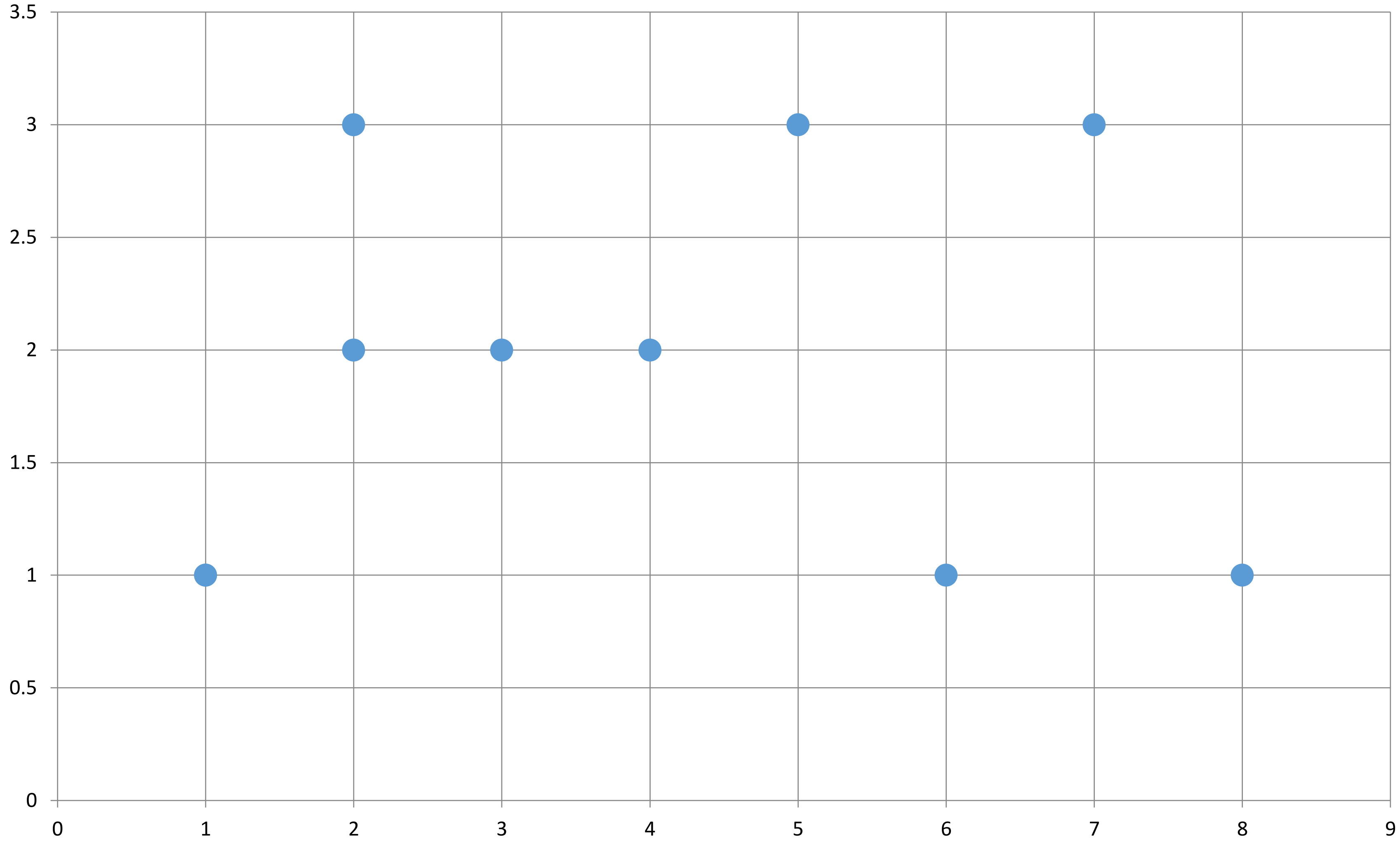


Would you say that the variables are positively correlated, negatively correlated, or uncorrelated? They cluster on a negatively sloped line, so negatively correlated.

Here’s a set of paired data.

X	Y
1	1
2	2
2	3
1	1
4	2
5	3
6	1
3	2
7	3
8	1

Here’s its scatterplot from Excel.

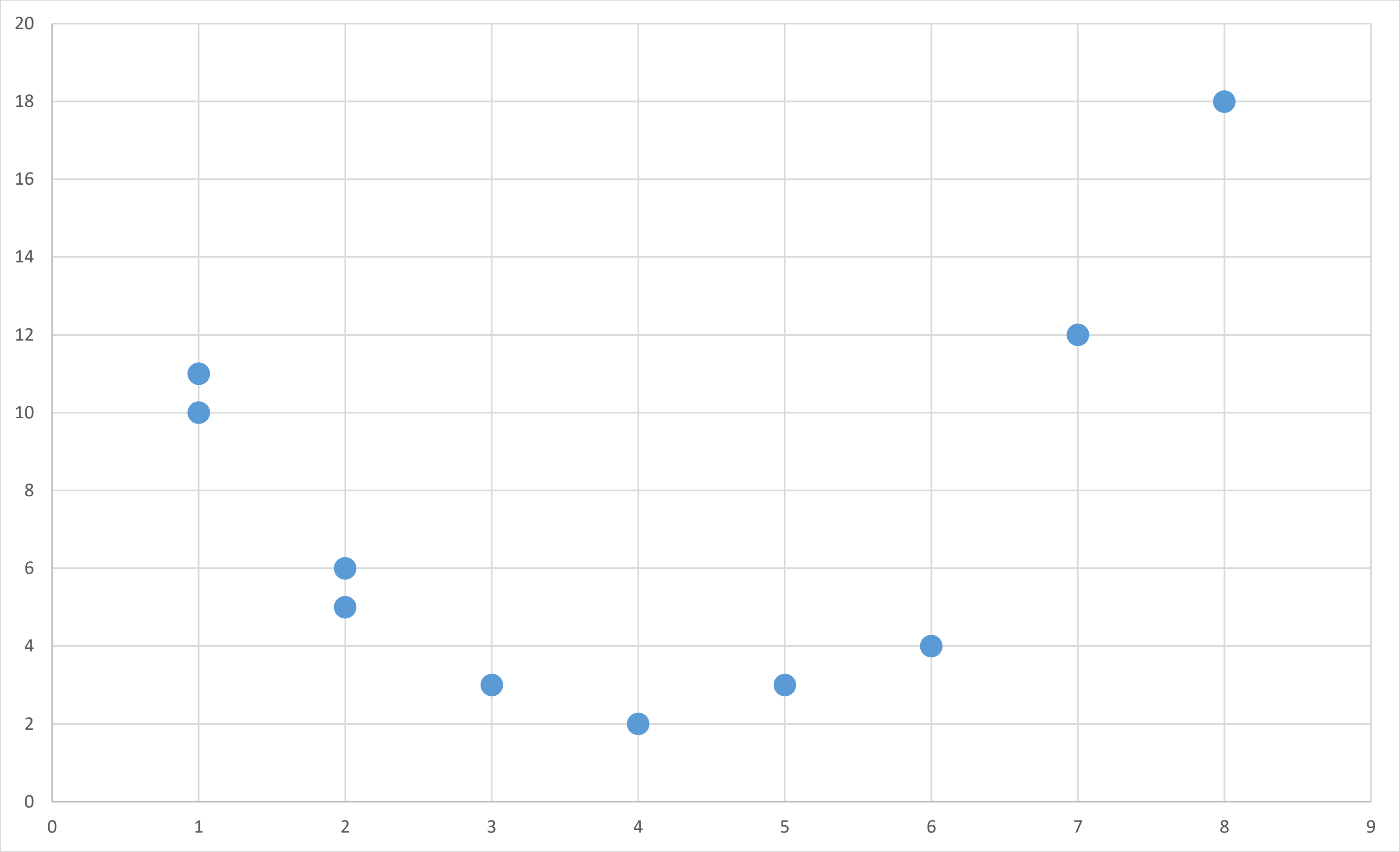


Would you say that the variables are positively correlated, negatively correlated, or uncorrelated? The points don’t cluster on a non-vertical, non-horizontal line, so uncorrelated.

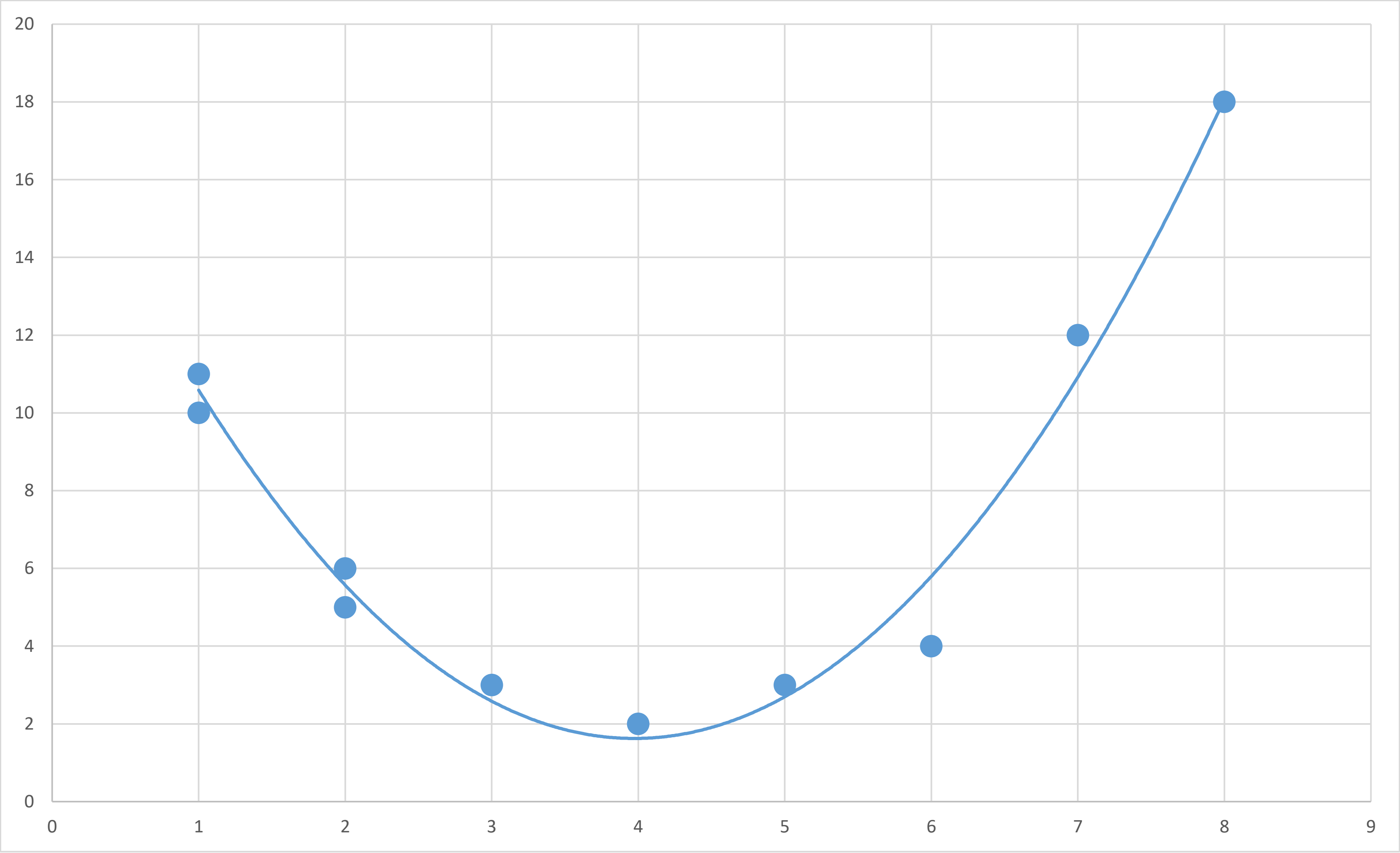
Here’s a set of paired data.

X	Y
1	11
2	6
2	5
1	10
4	2
5	3
6	4
3	3
7	12
8	18

Here’s its scatterplot from Excel.



Would you say that the variables are positively correlated, negatively correlated, or uncorrelated? The points don’t cluster on a non-vertical, non-horizontal line, so uncorrelated.



The variables don’t have a linear relationship, but they do appear to have a strong nonlinear relationship.