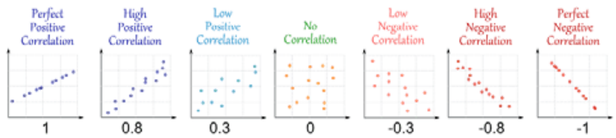
Correlation

When two sets of data are strongly linked together we say they have a **High Correlation**.

The word Correlation is made of **Co-** (meaning "together"), and **Relation**

- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases

Like this:



Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

The value shows **how good the correlation is** (not how steep the line is), and if it is positive or negative.

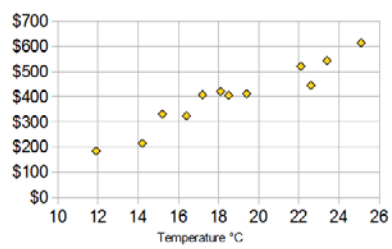
Example: Ice Cream Sales

The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days:

Ice Cream Sales vs Temperature

Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

And here is the same data as a Scatter Plot:



You can easily see that warmer weather leads to more sales, the relationship is good but not perfect.

In fact the correlation is **0.9575** ... see at the end how I calculated it.

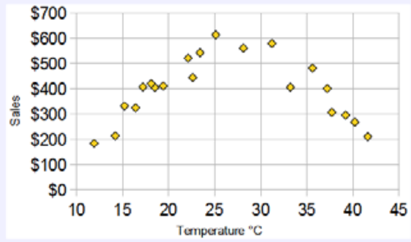
Correlation Is Not Good at Curves

The correlation calculation only works well for relationships that follow a straight line.

Our Ice Cream Example: **there has been a heat wave!**

It gets so hot that people aren't going near the shop, and **sales start dropping.**

Here is the latest graph:



The correlation is now **0**: "No Correlation" ... !

The calculated value of correlation is 0 (trust me, I worked it out), which says there is "no correlation".

But we can see the data follows a nice **curve** that reaches a peak around 25° C. But the correlation calculation is not "smart" enough to see this.

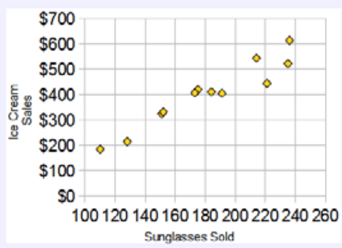
Moral of the story: make a Scatter Plot, and look at it!
You may see more than the correlation value says.

Correlation Is Not Causation

"Correlation Is Not Causation" ... by that I mean: when there is a correlation it does **not** mean that one thing causes the other

Example: Sunglasses vs Ice Cream

Our Ice Cream shop finds how many sunglasses were sold by a big store for each day and compares them to their ice cream sales:



The correlation between Sunglasses and Ice Cream sales is high

Does this mean that sunglasses make people want ice cream?

How To Calculate

How did I calculate the value **0.9575** at the top?

I used "Pearson's Correlation". There is software that can calculate it for you, such as the CORREL() function in Excel or OpenOffice Calc ...

... but here is how to calculate it yourself:

Let us call the two sets of data "x" and "y" (in our case Temperature is **x** and Ice Cream Sales is **y**):

- Step 1: Find the mean of **x**, and the mean of **y**
- Step 2: Subtract the mean of x from every x value (call them "**a**"), do the same for y (call them "**b**")
- Step 3: Calculate: $a \times b$, a^2 and b^2 for every value
- Step 4: Sum up $a \times b$, sum up a^2 and sum up b^2

- Step 5: Divide the sum of a × b by the square root of [(sum of a^2) × (sum of b^2)]

Here is how I calculated the first Ice Cream example (values rounded to 1 or 0 decimal places):

2 Subtract Mean

3 Calculate ab, a^2 and b^2

Temp °C	Sales	"a"	"b"	a×b	a^2	b^2
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

1 Calculate Means

4 Sum Up

5 $$\frac{5,325}{\sqrt{177.0 \times 174,757}} = 0.9575$$

As a formula it is:

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

As a formula it is:

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

Where:

- Σ is Sigma, the symbol for "sum up"
- $(x_i - \bar{x})$ is each x-value minus the mean of x (called "a" above)
- $(y_i - \bar{y})$ is each y-value minus the mean of y (called "b" above)

You probably won't have to calculate it like that, but at least you know it is not "magic", but simply a routine set of calculations.

Approximate Values

There are also approximate ways to calculate a correlation coefficient, such as "Spearman's rank correlation coefficient", but I prefer using a spreadsheet like above.

http://www.mathsisfun.com/data/correlation.html