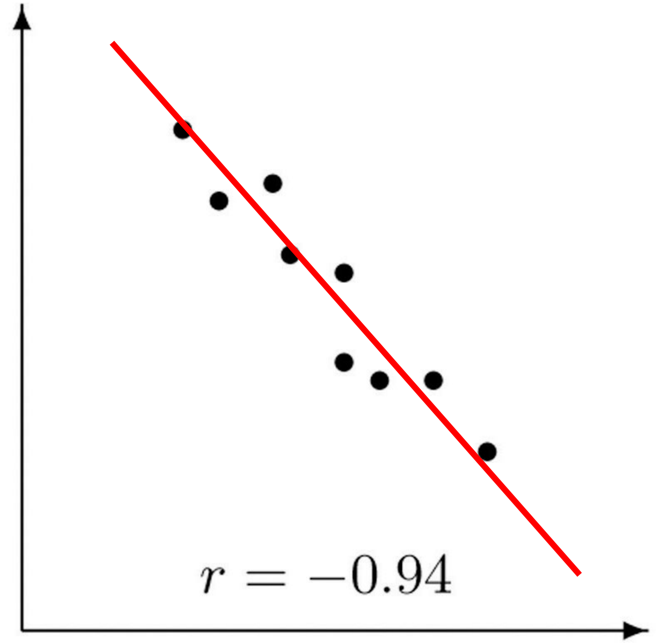
Pearsons Correlation Coefficient

YORK COLLEGE & UNIVERSITY CENTRE

What is a Correlation

- A coefficient shows the link between two variables
- Say we have variable x and variable y we can write out the relationship between these two variables as a correlation coefficient
- This can be written as “r = number”
- For example, higher temperature is correlated with higher ice cream sales.



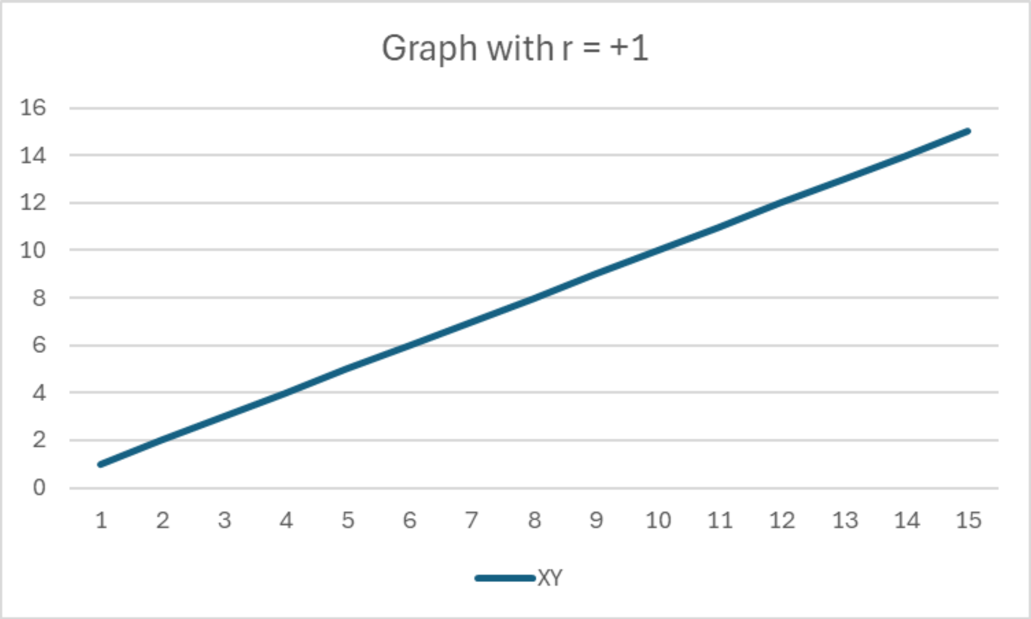
Positive and Negative Correlations

- **Positive Correlation**
- When one variable increases, the other also increases.
- When one variable decreases, the other also decreases.
- Example: More study time → Higher test scores.
- Graph trend: Upward slope (↗).

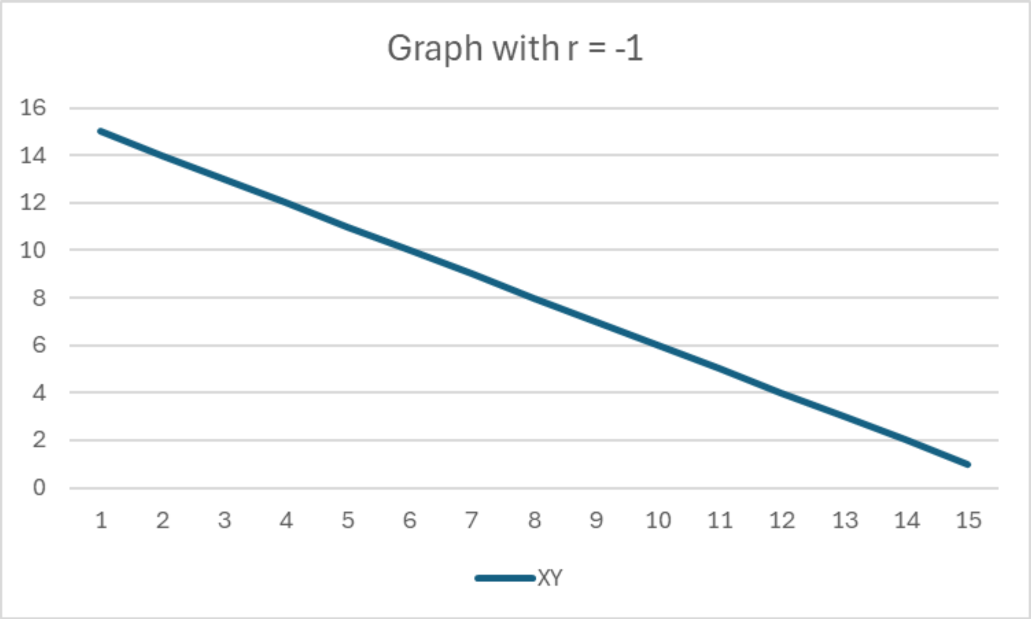
- **Negative Correlation**
- When one variable increases, the other decreases.
- When one variable decreases, the other increases.
- Example: More exercise → Lower body weight.
- Graph trend: Downward slope (↘).

Positive and Negative Correlations

- **Positive Correlation**



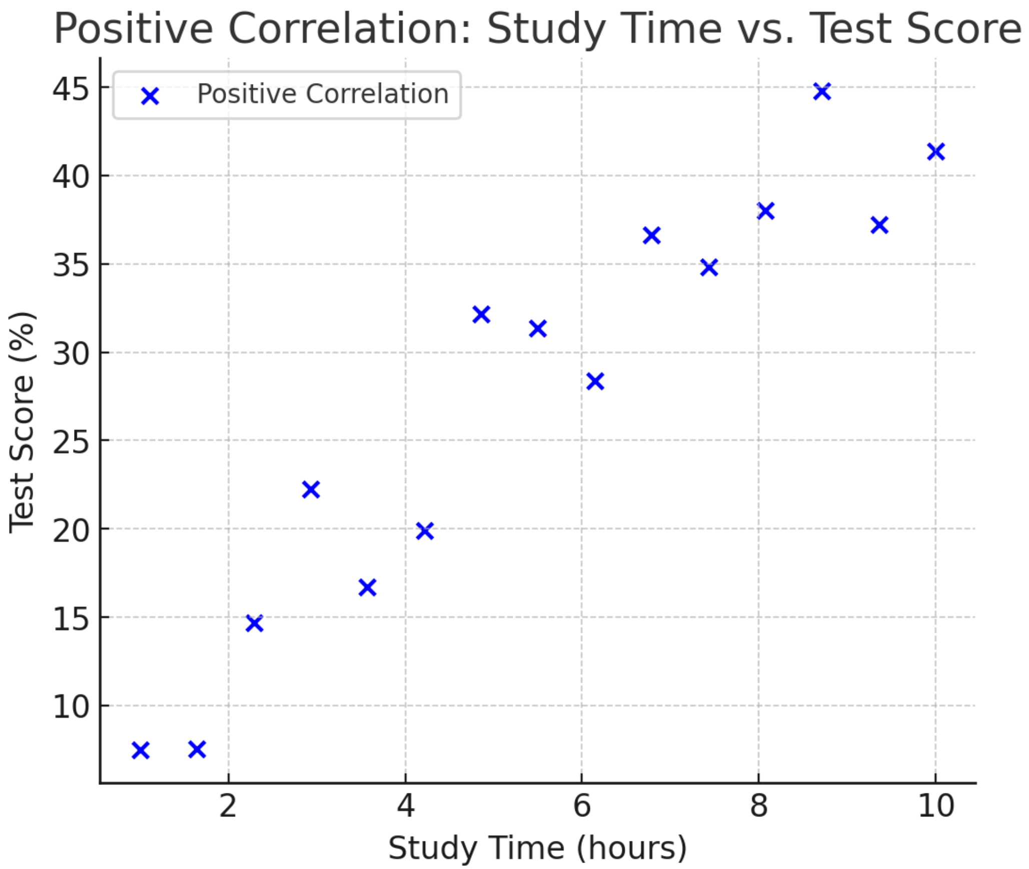
- **Negative Correlation**



The magnitude of both correlations is **“1”** as all the data points fall exactly on the line of best fit

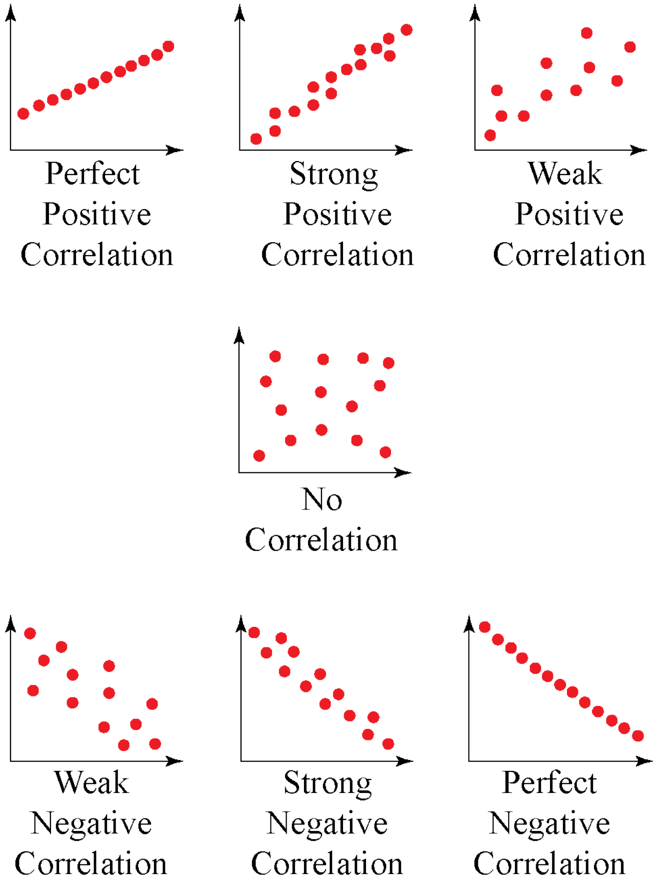
Understanding a correlation

- With the graph on the left there is a correlation as we can see the points follow a rough line
- We know that it would have a positive correlation from the way the line goes
- So, there is a positive correlation between test score and study time



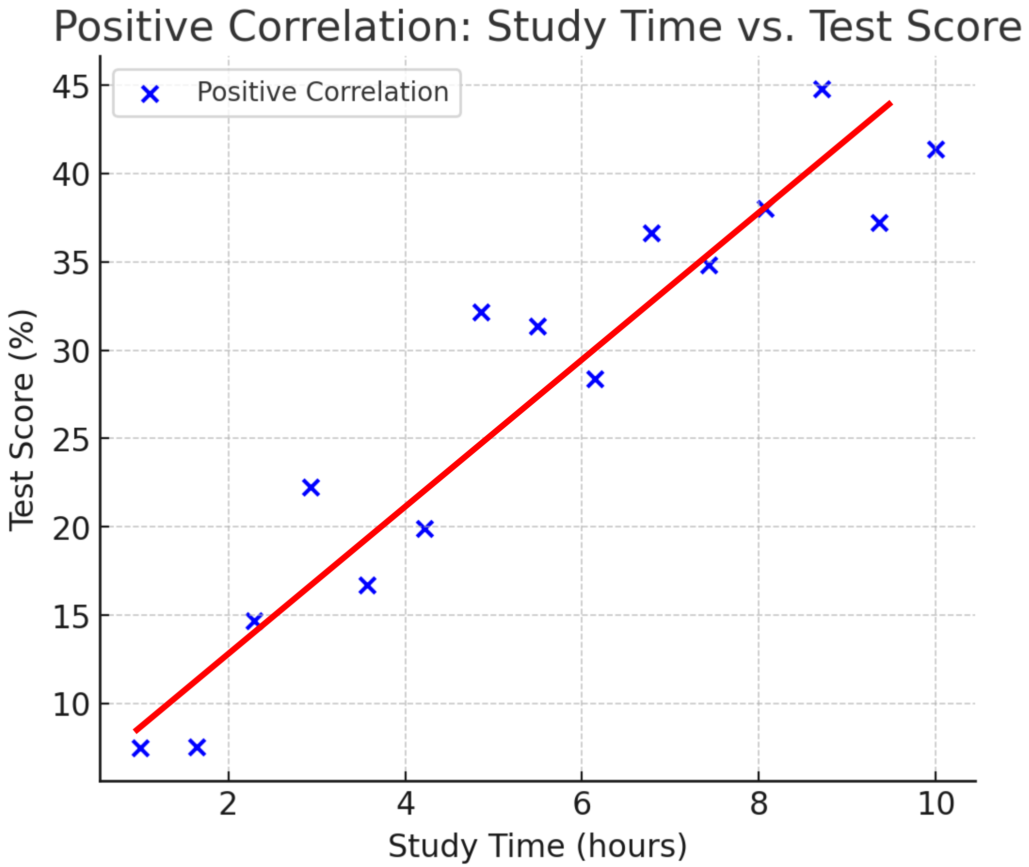
What is a correlation coefficient

- A numerical measure of the strength and direction of correlation.
- Ranges from -1 to +1:
  - +1 → Perfect positive correlation
  - 0 → No correlation
  - -1 → Perfect negative correlation
- The closer |r| is to 1, the stronger the correlation.
- Example: If r = 0.8, there is a strong positive correlation between temperature and ice cream sales.



Understanding a correlation coefficient

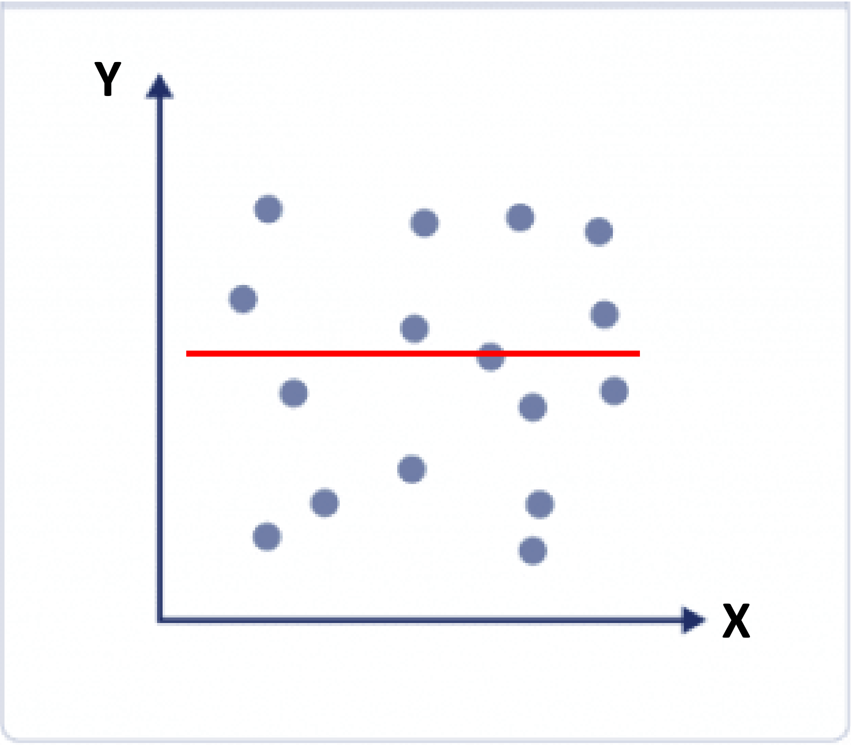
- If we drew on our line of best fit, how close the points are from that line determine the correlation coefficient.
- For this as the line has a positive correlation the coefficient would be in the range 0 < r < 1



Understanding a correlation coefficient

- If we draw our line of best fit on this data, we can see that we get a flat line, this means there is no correlation between x and y

- For this as the line has a no correlation the coefficient would be 0



Working out correlation coefficient

- We can work out a correlation coefficient using a table method
- We put into this values the x and y values from the graph/table and then we work out some other values

X	Y
1	90
2	88
3	85
4	83
5	80
6	78
7	75
8	72

Working out correlation coefficient

- We add the following to the table:
  - XY
  - X^2
  - Y^2

X	Y	XY	X^2	Y^2
1	90	90	1	8100
2	88	176	4	7744
3	85	255	9	7225
4	83	332	16	6889
5	80	400	25	6400
6	78	468	36	6084
7	75	525	49	5625
8	72	576	64	5184

Working out correlation coefficient

- Now let's add a total row to the bottom

- Column 1 total is $\sum x$
- Column 2 total is $\sum y$
- Column 3 total is $\sum xy$
- Column 4 total is $\sum x^2$
- Column 5 total is $\sum y^2$

X	Y	XY	X^2	Y^2
1	90	90	1	8100
2	88	176	4	7744
3	85	255	9	7225
4	83	332	16	6889
5	80	400	25	6400
6	78	468	36	6084
7	75	525	49	5625
8	72	576	64	5184
36	651	2822	204	53251

Working out correlation coefficient

- Now we have our values we can use our formula to work out the coefficient:
- $r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$
- We have all these values in our table all we need to know is n which is the number of rows we have (excluding total)
- So, for our example n = 8

X	Y	XY	X^2	Y^2
1	90	90	1	8100
2	88	176	4	7744
3	85	255	9	7225
4	83	332	16	6889
5	80	400	25	6400
6	78	468	36	6084
7	75	525	49	5625
8	72	576	64	5184
36	651	2822	204	53251

Working out correlation coefficient

- $r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$

- $r = \frac{8*(2822) - 36*651}{\sqrt{[8*(204) - (36)^2][8*(53251) - (651)^2]}}$

- $r = -0.9986829742$

- $r \approx -0.99$

X	Y	XY	X^2	Y^2
1	90	90	1	8100
2	88	176	4	7744
3	85	255	9	7225
4	83	332	16	6889
5	80	400	25	6400
6	78	468	36	6084
7	75	525	49	5625
8	72	576	64	5184
36	651	2822	204	53251

What can we deduce

- $r = -0.9986829742$

- There is a negative correlation between x and y (as x increases y decreases)

- The points are very close to the line of best fit meaning there is a strong correlation

X	Y	XY	X^2	Y^2
1	90	90	1	8100
2	88	176	4	7744
3	85	255	9	7225
4	83	332	16	6889
5	80	400	25	6400
6	78	468	36	6084
7	75	525	49	5625
8	72	576	64	5184
36	651	2822	204	53251

Your Turn

- Can you work out the correlation coefficient for this data
- What does this correlation coefficient tell us
- If you finish that task, please plot out the graph.

Cost of part (£)	Longevity (Hours)
10.0	32.42
16.92	30.39
23.85	63.88
30.77	99.61
37.69	69.53
44.62	83.38
51.54	142.56
58.46	136.11
65.38	119.03
72.31	158.18
79.23	146.88
86.15	160.66
93.08	192.20
100.0	152.17

Your Turn: Answers

Cost of part (£)	Longevity (Hours)	Cost * Longevity	Cost^2	Longevity^2
10.0	32.42	324.2	100	1051.0564
16.92	30.39	514.1988	286.2864	923.5521
23.85	63.88	1523.538	568.8225	4080.6544
30.77	99.61	3064.9997	946.7929	9922.1521
37.69	69.53	2620.5857	1420.5361	4834.4209
44.62	83.38	3720.4156	1990.9444	6952.2244
51.54	142.56	7347.5424	2656.3716	20323.3536
58.46	136.11	7956.9906	3417.5716	18525.9321
65.38	119.03	7782.1814	4274.5444	14168.1409
72.31	158.18	11437.9958	5228.7361	25020.9124
79.23	146.88	11637.3024	6277.3929	21573.7344
86.15	160.66	13840.859	7421.8225	25811.6356
93.08	192.20	17889.976	8663.8864	36940.84
100.0	152.17	15217	10000	23155.7089
770	1587	104677.7854	53253.7078	213284.3182

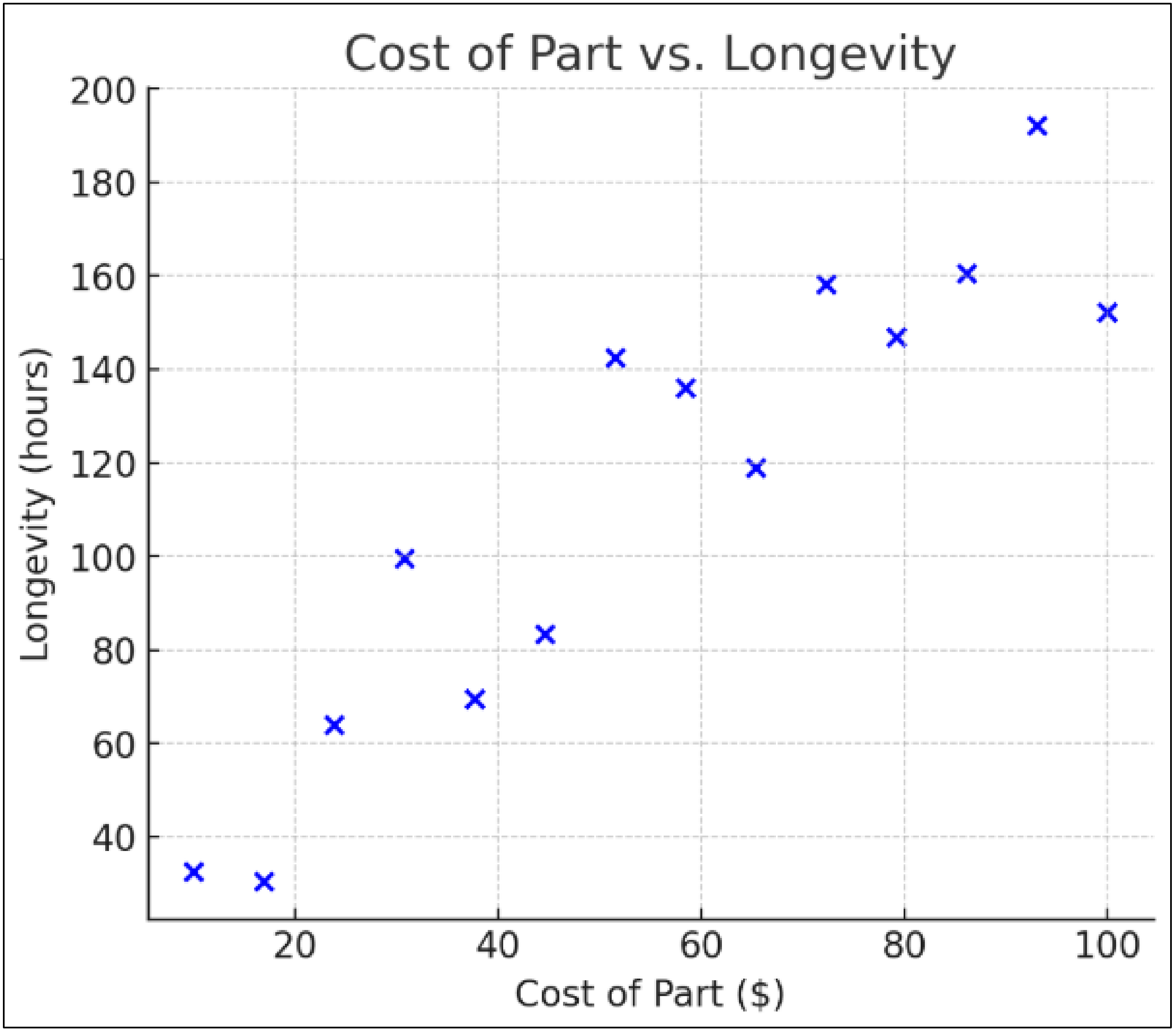
$$n = 14$$

$$r = \frac{14(104677.7854) - 770 * 1587}{\sqrt{(14(53253.7078) - (770)^2)(14(213284.3182) - (1587)^2)}}$$

$$r = 0.9115833375$$

There is a strong positive correlation between the cost of a part and how long it lasts

Your Turn: Answers



The other method

- We can also use the mean to work out the value of our correlation coefficient, again using a table helps us substantially
- $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$
- Note that $\bar{x}$ is the mean of x

X	Y
1	90
2	88
3	85
4	83
5	80
6	78
7	75
8	72

The other method

- We can work out the mean of a section by adding up all the values and dividing by n

X	Y	$\bar{x}$	$\bar{y}$
1	90	4.5	81.375
2	88		
3	85		
4	83		
5	80		
6	78		
7	75		
8	72		
36	651		

The other method

- Next let's add the $x - \bar{x}$ and $y - \bar{y}$ columns

- Then we write the totals in the totals row

X	Y	$\bar{x}$	$\bar{y}$	X - $\bar{x}$	Y - $\bar{y}$
1	90	4.5	81.375	-3.5	8.625
2	88			-2.5	6.625
3	85			-1.5	3.625
4	83			-0.5	1.625
5	80			0.5	-1.375
6	78			1.5	-3.375
7	75			2.5	-6.375
8	72			3.5	-9.375
36	651				

The other method

- Next let's add a column for:

  $(\mathbf{X} - \bar{x})(\mathbf{Y} - \bar{y})$

- Next let's add the total of that column in the total row

- Now we have the top of the equation, we just need to work out the bottom

X	Y	$\bar{x}$	$\bar{y}$	X - $\bar{x}$	Y - $\bar{y}$	(X - $\bar{x}$)(Y - $\bar{y}$)
1	90	4.5	81.375	-3.5	8.625	-30.1875
2	88			-2.5	6.625	-16.5625
3	85			-1.5	3.625	-5.4375
4	83			-0.5	1.625	-0.8125
5	80			0.5	-1.375	0.6875
6	78			1.5	-3.375	-5.0625
7	75			2.5	-6.375	-15.9375
8	72			3.5	-9.375	-32.8125
36	651					-106.125

The other method

- For the bottom we need:

  (X - $\bar{x}$)^2 and **(Y - $\bar{y}$)^2**

- Again, we can just add these in as columns in our table

- Then just add the total to the bottom again

X	Y	$\bar{x}$	$\bar{y}$	X - $\bar{x}$	Y - $\bar{y}$	(X - $\bar{x}$)(Y - $\bar{y}$)	(X - $\bar{x}$)^2	(Y - $\bar{y}$)^2
1	90	4.5	81.375	-3.5	8.625	-30.1875	12.25	74.390625
2	88			-2.5	6.625	-16.5625	6.25	43.890625
3	85			-1.5	3.625	-5.4375	2.25	13.140625
4	83			-0.5	1.625	-0.8125	0.25	2.640625
5	80			0.5	-1.375	0.6875	0.25	1.890625
6	78			1.5	-3.375	-5.0625	2.25	11.390625
7	75			2.5	-6.375	-15.9375	6.25	40.640625
8	72			3.5	-9.375	-32.8125	12.25	87.890625
36	651					-106.125	42	276

The other method

- Finally, we just put what we have in the equation

- $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$

- $r = \frac{-106.125}{\sqrt{42 * 276}} = -0.9856858385$

- $r \approx -0.99$

X	Y	$\bar{x}$	$\bar{y}$	X - $\bar{x}$	Y - $\bar{y}$	(X - $\bar{x}$)(Y - $\bar{y}$)	(X - $\bar{x}$)^2	(Y - $\bar{y}$)^2
1	90	4.5	81.375	-3.5	8.625	-30.1875	12.25	74.390625
2	88			-2.5	6.625	-16.5625	6.25	43.890625
3	85			-1.5	3.625	-5.4375	2.25	13.140625
4	83			-0.5	1.625	-0.8125	0.25	2.640625
5	80			0.5	-1.375	0.6875	0.25	1.890625
6	78			1.5	-3.375	-5.0625	2.25	11.390625
7	75			2.5	-6.375	-15.9375	6.25	40.640625
8	72			3.5	-9.375	-32.8125	12.25	87.890625
36	651					-106.125	42	276

Note this value is not the same as the value we got before but due to rounding in the process its close enough

Your Turn

- Can you work out the correlation coefficient for this data using this equation:

- $$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$$

- What does this correlation coefficient tell us

- Can you plot this scatter graph

X	Y	$\bar{x}$	$\bar{y}$	X - $\bar{x}$	Y - $\bar{y}$	(X - $\bar{x}$)(Y - $\bar{y}$)	(X - $\bar{x}$)^2	(Y - $\bar{y}$)^2
1.0	98							
1.5	96							
2.0	94							
2.5	91							
3.0	89							
3.5	85							
4.0	82							
4.5	78							
5.0	74							
5.5	70							
6.0	65							
6.5	60							
7.0	55							

X is screen time and y is eyesight rating