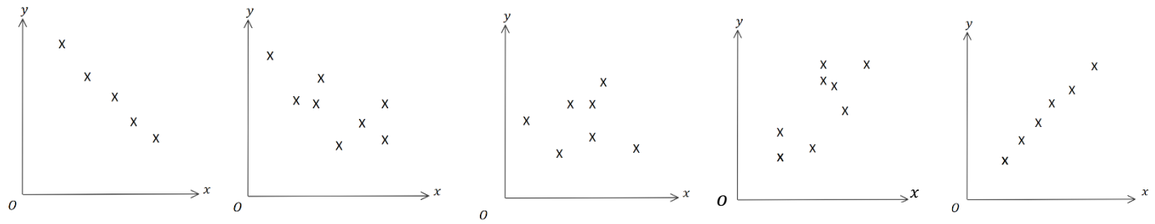
Correlation Cheat Sheet

Bivariate data is data which has pairs of values for two variables. You can represent bivariate data on a scatter diagram. The independent or explanatory variable is something which the researcher can control and is usually plotted on the x-axis. The dependent or response variable, which is measured by the researcher, is usually plotted on the y-axis.

Correlation describes the nature of linear relationships between two variables.



Strong negative correlation | Weak negative correlation | No linear correlation | Weak positive correlation | Strong positive correlation

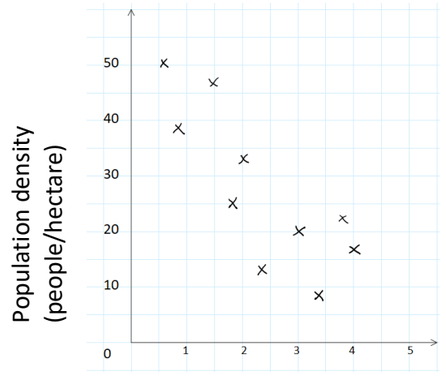
A negative correlation means that one variable decreases when the other increases. Positive correlation means that one variable increases with the increase of the other variable.

Two variables have a causal relationship if a change in one variable causes a change in the other. If two variables are correlated, you need to look at the context of the question to determine if they have a causal relationship.

Example 1: In the study of a city, the population density, in people/hectare, and the distance from the city centre, in km, was investigated by picking a number of sample areas with the following results.

Area	A	B	C	D	E	F	G	H	I	J
Distance (km)	0.6	3.8	2.4	3.0	2.0	1.5	1.8	3.4	4.0	0.9
Population density (people/hectare)	50	22	14	20	33	47	25	8	16	38

a. Draw a scatter diagram to represent this data.



Remember to label your axis and include units

b. Describe the correlation between distance and population density.

There is a weak negative correlation.

Describe the strength of correlation and whether it is positive or negative

c. Interpret your answer to part b.

As distance from the centre increases, the population density decreases.

Interpret results in context to the question

Linear regression

The least squares regression line, or regression line, is a line of best fit which can be drawn on a scatter plot. This is the straight line that minimises the sum of the squares of the distances of each datapoint from the line.

The regression line of y on x is written in the form:

$$y = a + bx$$

The coefficient b tells you the change in y for each unit change in x.

- For positively correlated data, b is positive
- For negatively correlated data, b is negative

You can substitute a known value of the independent variable into x and use the regression line to estimate the corresponding value of the dependent variable. This should only be done within the range of data given and is known as interpolation. Extrapolation out of the data range gives a much less reliable estimate.

If you need to predict a value of x for a given value of y, you will need to use the regression line of x on y.

Example 2: The daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 15 days in May in Camborne. The data was plotted on a scattered diagram. The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$.

a. Give an interpretation of the value of the gradient of this regression line.
The daily maximum gust is expected to increase by approximately 1.8 knots when the daily mean windspeed increases by 1 knot.

b. Predict the daily maximum gust when the daily mean speed is 16 knots.
Substitute 16 into w in the regression equation:
$g = 7.23 + 1.82(16)$
$= 36.35$ knots