GCSE Maths – Statistics

## Scatter Graphs

Notes

WORKSHEET

Scatter Graphs

Scatter graphs are a way of representing **two sets of data**, allowing you to see whether they have a connection.
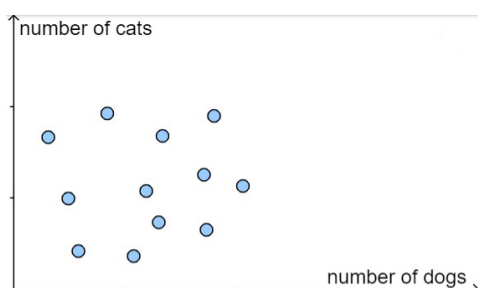
Correlation

By plotting two data sets on the same graph, you can analyse how the points align to see if the variables are connected. The **closer the points are to representing a straight line**, the **more correlated** they are said to be.
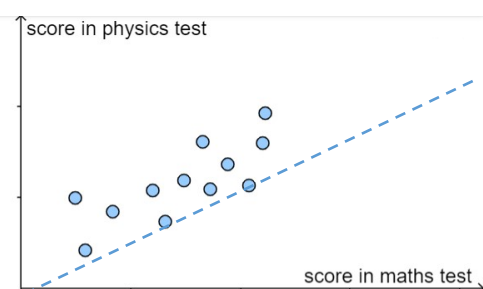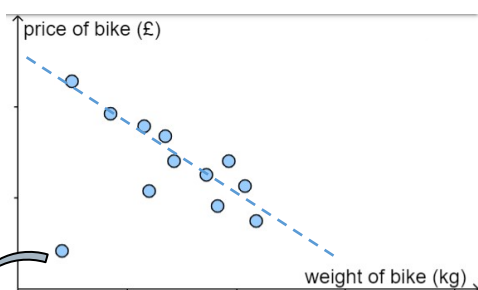

Positive Correlation

As the variable on the x axis increases, the variable on the y axis increases. The line of best fit has a positive gradient.

No correlation

There is no relationship between the two sets of data. There is no identifiable line of best fit.

Negative correlation

As the variable on the x axis increases, the variable on the y axis decreases. The line of best fit has a negative gradient.

A point on a graph which does not follow the general trend is known as **an outlier** or **an anomaly.** It is well away from the other points, and far from a line of best fit.

The **closeness of the points** to each other indicates the **strength of the correlation**:

- If all the points are very close to each other and to the line of best fit, the correlation is **strong**.
- If the points are more dispersed, but still represent a straight line, the correlation is described as **weak**.

Just because there is a correlation between two variables does not mean that the two variables are connected: **correlation does not always mean causation.**

For example, if there is a positive correlation between the number of ice creams sold and the number of traffic jams on the road, this does not mean ice creams cause traffic jams. If the weather is warm, more ice creams will be sold and more people will be travelling, which may explain the increased number of traffic jams.

Lines of best fit

A line of best fit can be drawn through the points on a scatter graph to **best represent the relationship between the points.**

The line of best fit should be drawn as close to as many of the points on the graph as possible and have roughly the same number of points lying on either side of it. We **ignore anomalies** when drawing a line of best fit.
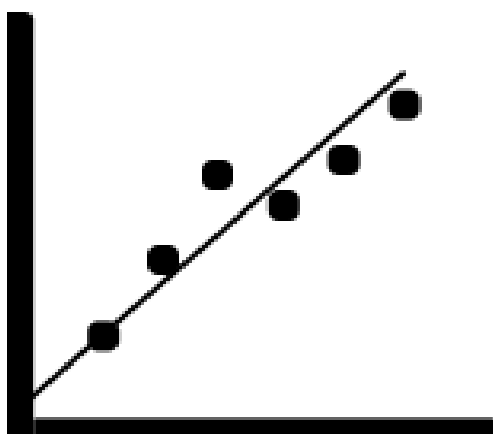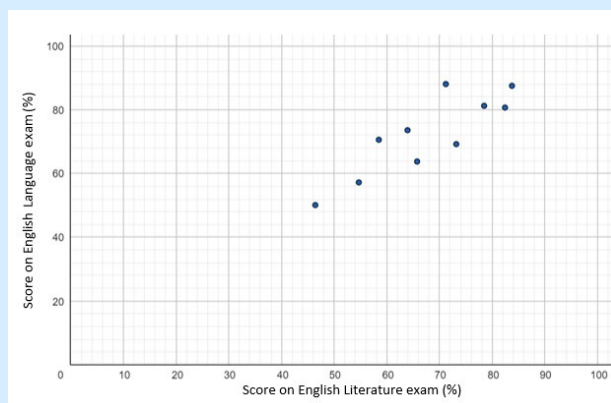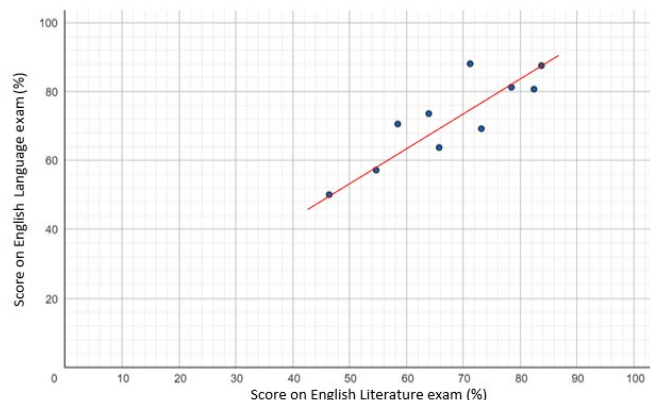

Diagram from The Noun Project CC BY 3.0

Using a line of best fit, we can **estimate** and **predict values** in and beyond the data range.

Example: Below is a graph of the scores of a class on an English Literature and English Language exam. Draw a line of best fit.

1. Draw a line between the points, so that there are roughly the same number of points either side of it.

The line of best fit should only span the data range.

Estimating and predicting values

We can use a line of best to predict values within the data range. This is called **interpolation.**

If we extend the line of best fit beyond the data set, we can predict values beyond the data range. This is called **extrapolation.**

- Interpolation is **more accurate** than extrapolation.
- **If we predict a value by extrapolation, it is less reliable** as we are assuming the trend shown by the line of best fit will continue exactly the same beyond the data range. This may not be the case.

Example: A small business has been tracking how effective its marketing strategy is. They have made a graph of the number of hours spent on marketing and the number of clients gained, which is pictured below. If the employees spend a total of two hours in a week on marketing, how many clients can they expect to gain?
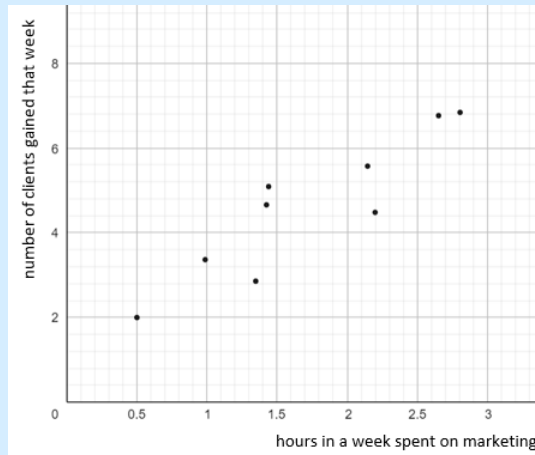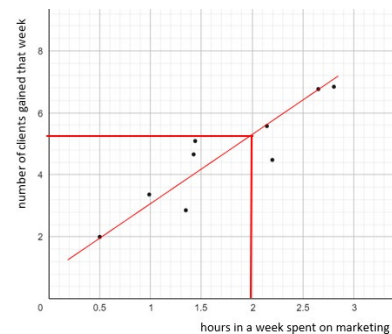

1. **Draw a line of best fit** on the graph. This should go through the middle of all the points and should have the same number of points on each side of the line.
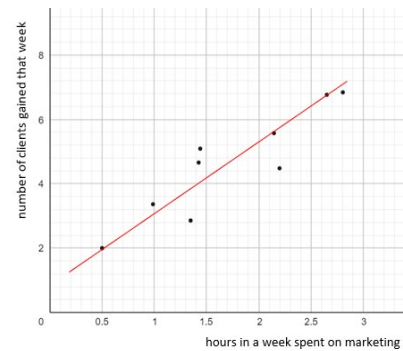
2. Draw a **vertical line** from 2 hours on the x-axis to the line of best fit.
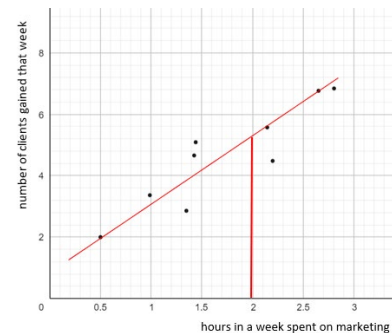
3. Draw a horizontal line across to the y-axis to **estimate** the number of clients gained from two hours of marketing.

   *The value read from the y-axis rounds to **5 clients gained in a week**.*

Example: Shown is a graph of the number of times a person has gone running in a week and the number of times they visited the doctor that year. If Jason goes running 16 times in a week, how many times can he expect to see the doctor over a year? Comment on the reliability of your estimate.
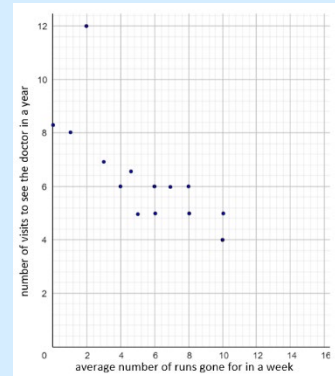
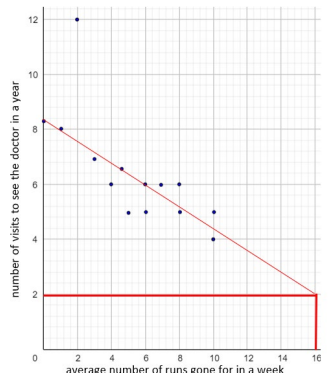1. Draw a line of best fit on the graph, ignoring the anomaly at (2,12).
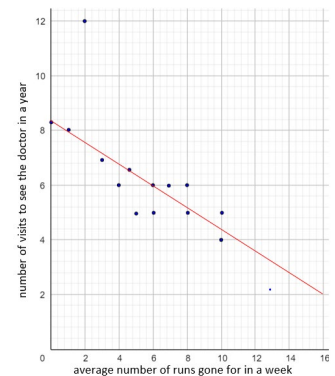
2. Draw a line from 16 runs on the x axis to the line of best fit.
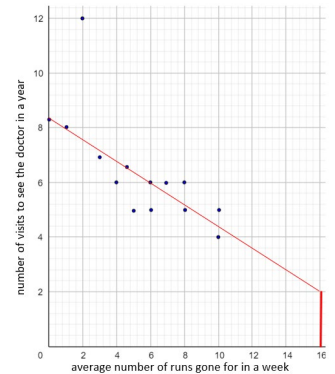
3. Draw a line across to the y axis, to estimate how many visits to the doctors I can expect.

   *If I go on 16 runs weekly, I can expect to see the doctor twice.*

4. Decide whether you have used interpolation or extrapolation to evaluate the reliability of the estimate.

   *The value of 16 runs is beyond the data range, which, as we can tell from the graph, is 0 to 10 runs. So, we have **extrapolated**, which means the **estimate is not very reliable.***
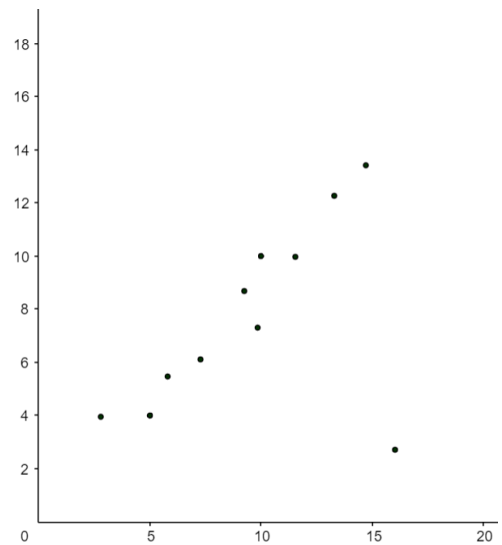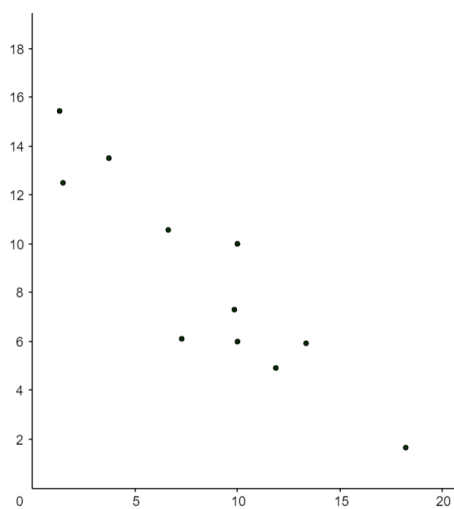
Scatter Graphs - Practice Questions

1. A small film company makes a graph of the money spent and made from each feature film it has produced. If the company spends £1000 making a film, how much can they expect to make from it?
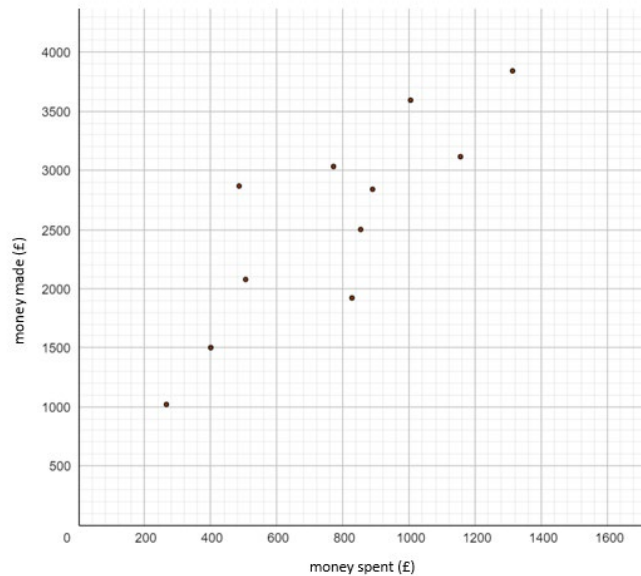
2. On the following scatter graphs, draw the line of best fit and describe the correlation. Circle any anomalies.

3. In an amateur darts competition, during the practice run, the organisers record the distance each turn is taken from, and the score achieved. If a person stands 1.5m from the dart board, estimate what they will score.
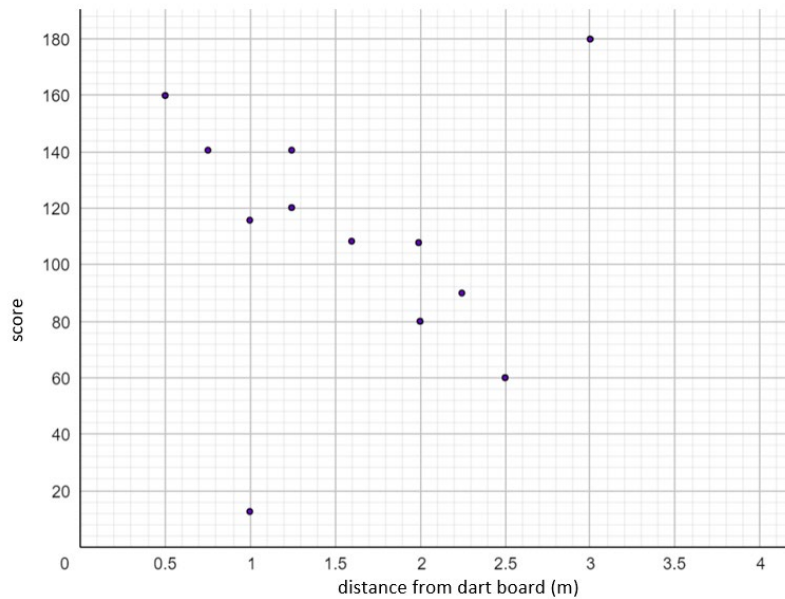
4. A scientist has devised a quantitative method of measuring tiredness, which gives subjects a score out of 5.

The scientist tests his measure out and creates a scatter graph of participants' scores against how many hours of sleep they have per night. Using the graph, predict how many hours of sleep someone with a score of 3 gets per night. Comment on the reliability of the prediction.
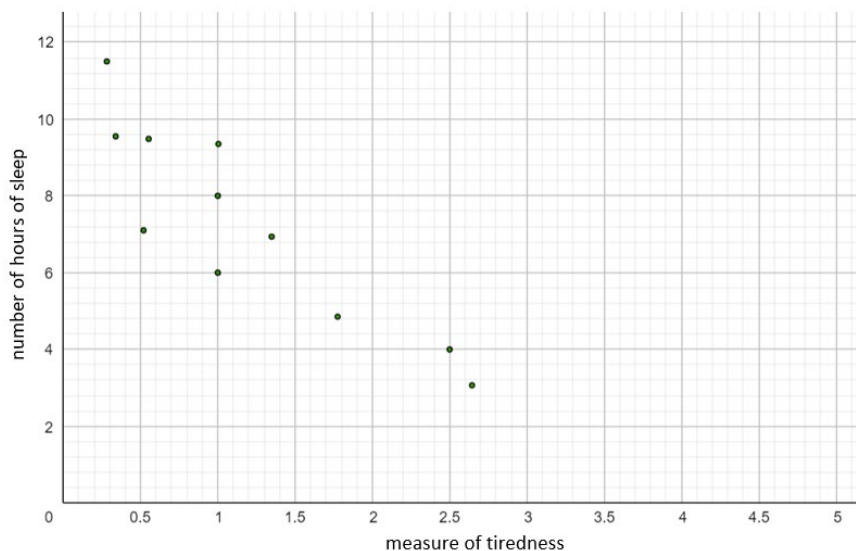
Worked solutions for the practice questions can be found amongst the worked solutions for the corresponding worksheet file.