4.3 Sales forecasting (HL)

Tool: Simple linear regression

What is linear regression analysis?

Imagine that you are the national sales manager of a clothing business and you are trying to predict next month's sales figures. There might be several factors that can impact your predictions, from changes in the weather to competitors' marketing strategies, or unforeseen events such as COVID-19.

Simple linear regression analysis provides a mathematical way to sort out the possible factors that might impact future sales, or other elements of a business. Regression models describe the relationship between two variables by fitting a line of best fit. The simple linear regression tool allows businesses to estimate how a dependent variable changes as the independent variable changes.

Simple linear regression can also be used to analyse how effective the marketing strategies of some businesses have been. For example, the analysis can show to what extent the spending on marketing has been successful in generating sales.

Features of simple linear regression include:

- The dependent variable: the main factor that the business is trying to predict. For example, the dependent variable could be monthly sales.
- The independent variable: the factor that the business suspects has an impact on its dependent variable (for example, monthly sales).

Simple linear regression involves the following steps:

1. Creating scatter diagrams to plot data from two variables.
2. Sketching a line of best fit.
3. Extrapolating the data to make predictions.

Scatter diagrams

A scatter diagram is a special type of graph designed to show the relationship between two variables. With simple regression analysis, you can use a scatter diagram to see if the data given in terms of X and Y are linearly related.

Example: Suppose Business A wants to know the relationship between its online advertising costs (spending) and its e-commerce sales. The business has been able to get the survey results from its seven online stores for the last year. **Table 1** represents the survey results from the seven online stores.

Table 1. Business A's online advertising costs versus monthly e-commerce sales (in thousands of $), showing a positive relationship.

Online stores for Business A	Online advertising costs (in thousands of $)	Monthly e-commerce sales (in thousands of $)
1	1.9	379
2	1.6	335
3	2.4	595
4	4.5	785
5	1.5	350
6	2.7	525
7	1.1	310

From the table above, it can be seen that there is a positive correlation between the online advertising costs and monthly e-commerce sales. In simple terms this means that, as the value of the independent variable (advertising costs) increases, the value of the dependent variable (e-commerce sales) also increases. With this data, the business can predict that, as the advertising costs increase, so do the monthly e-commerce sales.

Using the information from **Table 1**, the following scatter diagram can be made with advertising costs plotted on the *x*-axis and monthly e-commerce sales plotted on the *y*-axis.

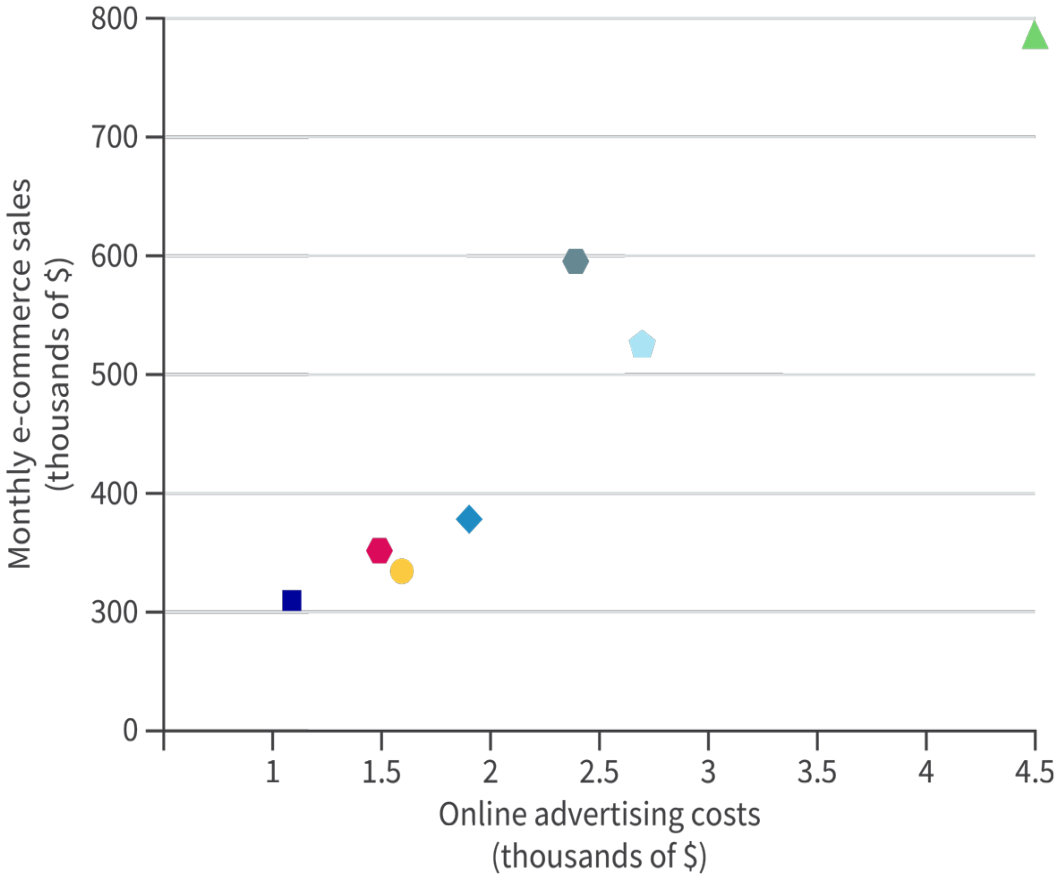

Figure 1. A scatter diagram showing data from advertising costs and e-commerce sales.

A scatter diagram has both benefits and limitations, as shown in **Table 2**.

Table 2. The benefits and limitations of scatter diagrams.

Benefits of a scatter diagram	Limitations of a scatter diagram

Benefits of a scatter diagram	Limitations of a scatter diagram
Scatter diagrams are easy to plot. A scatter diagram depicts the relationship between two variables, which is good for visual learners. Scatter diagrams show non-linear patterns with ease. It is easy to observe and interpret the pattern depicted in a scatter diagram. Maximum and minimum values are easily determined in a scatter diagram.	Scatter diagrams cannot give you the exact extent of correlation. A scatter diagram cannot take more than two variables into account. Only relationships between two variables can be illustrated. A scatter diagram only depicts quantitative data and cannot reflect qualitative data.

Line of best fit

A scatter diagram shows all the relationships between individual pieces of data for the independent and dependent variables. However, to be useful, a business needs to find a general relationship between the variables that can be used for predictions. A line of best fit will express this general relationship.

The line of best fit is a line through a scatter plot of data that captures the relationship between the independent and dependent variables. The line of best fit should be sketched in a way that is closest to the most number of points in the scatter diagram. It goes roughly through the middle of all the points on the scatter diagram.

To make the line of best fit even more accurate, it is important that you draw the line through a point that represents the mean of the independent data and the mean of the dependent data. Also, if possible, a roughly equal number of points should be above and below the line of best fit.

In **Figure 2** below, the same data from **Table 1** has been used. The diagonal line is the line of best fit, which is also called a line of regression. It illustrates the predicted relationship between each possible value of advertising costs and e-commerce sales.

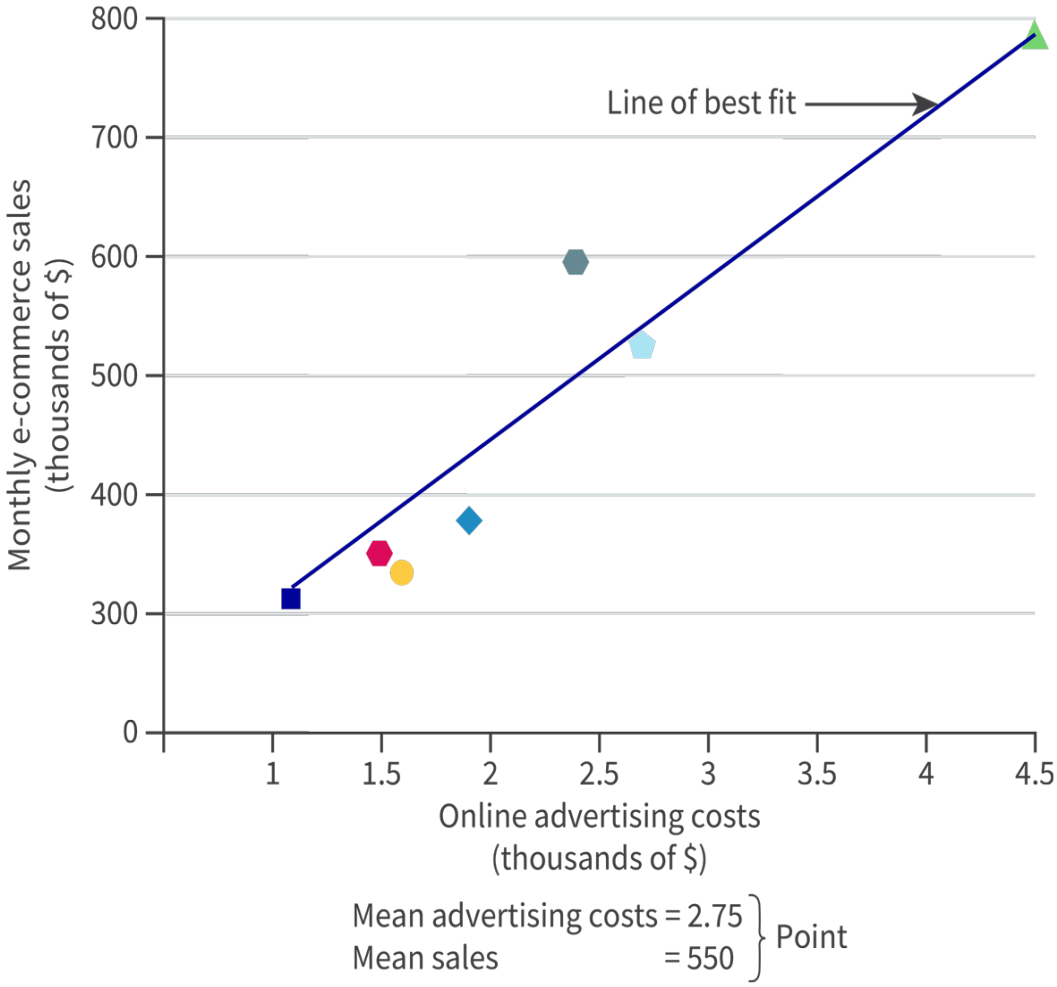

Figure 2. Line of best fit showing the relationship between advertising costs and e-commerce sales.

Now take the same example of advertising and e-commerce sales, but with data that shows a negative relationship between the two variables. This could be because the advertising is ineffective or is turning customers off. **Table 3** represents the survey results from the seven online stores.

Table 3. Online advertising costs versus monthly e-commerce sales (in thousands of $), showing a negative relationship.

Online stores for business A	Online advertising costs (in thousands of $)	Monthly e-commerce sales (in thousands of $)
1	1.9	468
2	2.1	450

Online stores for business A	Online advertising costs (in thousands of $)	Monthly e-commerce sales (in thousands of $)
3	2.9	375
4	3.2	355
5	3.7	300
6	4.2	285
7	4.5	250

From the table above, it can be seen that there is a negative correlation between the online advertising costs (x-axis) and monthly e-commerce sales (y-axis). In simple terms, this means that as the value of the independent variable (advertising costs) increases, the value of the dependent variable (e-commerce sales) decreases. With this data, the business can predict that, as the advertising costs increase, the predicted monthly e-commerce sales decrease.

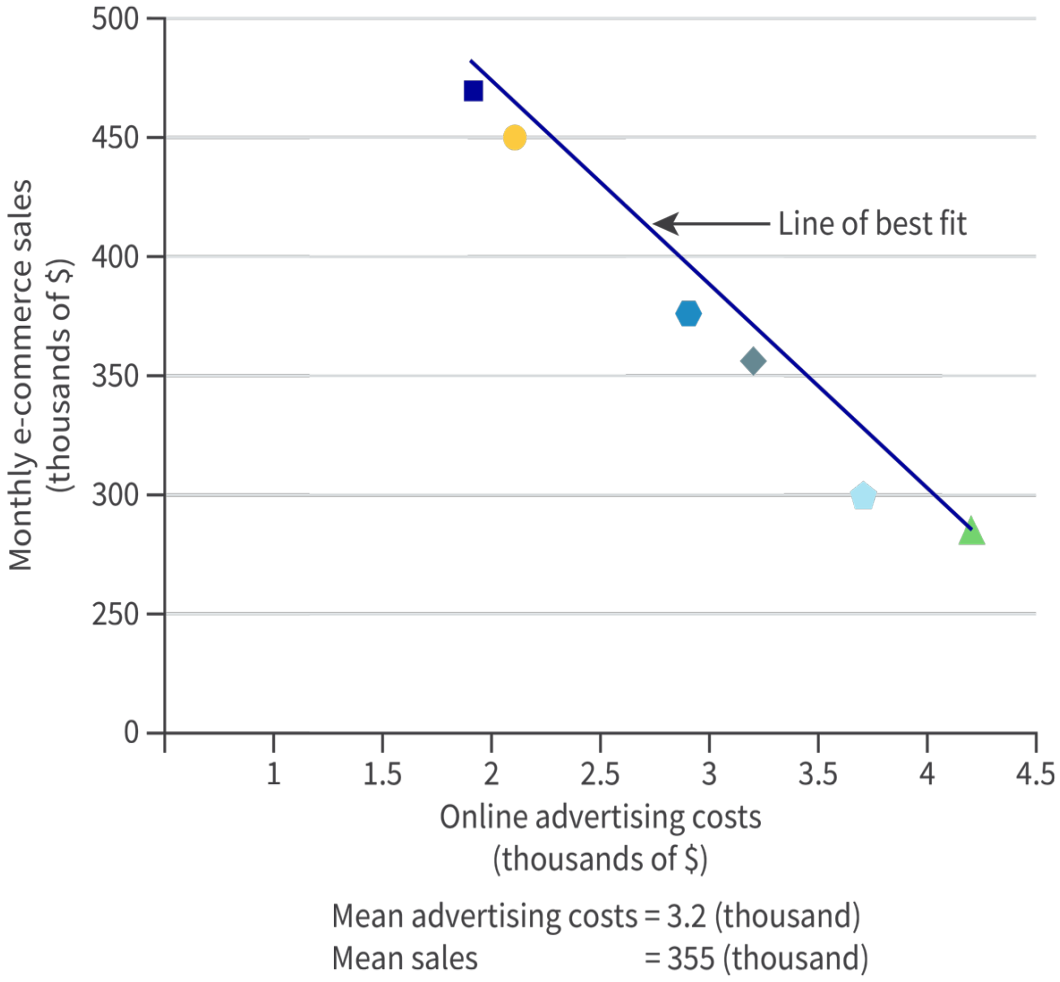

Figure 3. Line of best fit diagram.

Activity

SaniaR is a fast fashion clothing retailer. It delivers a new range of clothing every two weeks in its over 1500 stores worldwide. SaniaR's unique selling point is the rapid changes to its clothing lines, achieved by launching new styles every two weeks. The company's sales depend on the number of different styles they launch every two weeks – the greater the number of styles, the greater the sales. Of course, introducing new lines of clothing affects costs of production.

Table 4 shows the relationship between SaniaR's production and marketing costs and their monthly sales. (The table uses fictional data for the purpose of understanding.)

1. Draw a scatter diagram with the line of best fit to illustrate the information given in **Table 4**. [2 marks]
2. Explain whether there is a positive or negative correlation between the two variables. [2 marks]

Table 4. SaniaR's production and marketing costs versus mean monthly sales per year.

Year	Production and marketing costs (in thousands of $)	Monthly sales (in thousands of $)
1	105	195
2	110	205
3	106	204
4	125	250
5	135	300
6	145	350
7	165	450

Exam tip

In an exam question, your sketch of a line of best fit will not be precise. Thus, examiners will be instructed to accept reasonable attempts at drawing the line of best fit.

However, you must make sure to find the mean of the data for the independent and dependent variables, and make sure your line goes through that point. This is required to access full marks in exam questions asking for the line of best fit.

Video 1. Using scatter plots to describe relationships between two variables, and using lines of best fit to make predictions.

Time series analysis: moving averages

In Section 4.3.1 (/study/app/y12-business-management-a-hl-may-2024/sid-351-cid-174702/book/sales-forecasting-id-38738), you learned that sales data can vary over time quite significantly. When this happens, it can be difficult to see the overall trend and to draw the line of best fit. So it is important to smooth out the data by finding the mean of groups of data. Calculating a moving average can help you do that.

Analysing data using the trend analysis of time series data enables a business to understand a number of things. A trend is a pattern over time. Firstly, the business can know the trend of the sales it is making – in other words – whether this is rising or falling over time. Secondly, the business can understand any seasonal fluctuations. This is important for businesses that sell seasonal products such as ice creams, holidays or clothing. Thirdly, the business can pay attention to any cyclical fluctuations. This means those fluctuations that are the result of economic growth or recession in the broader economy.

The easiest way to understand this concept is by using an example. Imagine a business that specialises in selling second-hand cars. The business has a number of loyal customers, who on average replace their cars once every three years.

Step 1: Calculate the three-year moving average

The second-hand car business is thinking of expanding. However, it will only be profitable for it to do so if forecast sales for 2022 are above 100 cars a year. **Table 5** shows the company's sales figures and moving average for the last nine years. A moving average attempts to 'smooth out' any peaks or troughs in sales data so that underlying trends in data can be seen. Sales data goes through a three-year cycle. A three-part moving average can therefore be used. The three-year moving average is calculated in **Table 5**.

Table 5. Annual car sales from 2011 to 2022.

Year	Car sales	Three-part moving average (trend)
2011	34	
2012	110	80
2013	96	83
2014	43	85
2015	116	86
2016	99	88
2017	49	90
2018	122	91
2019	102	**93**
2020	?	**94**
2021	?	**96**
2022	?	**98**

The three-part moving average is calculated using a mean. For example, the first figure of 80 was calculated by taking the average of the sales figures from 2011, 2012 and 2013 as follows:

$$\text{Three-part moving average} = \frac{(34+110+96)}{3} = 80$$

Each figure for the three-part moving average is then graphed. This shows the overall trend, smoothing out the extreme variations in the data. This is shown graphically in **Figure 4**.

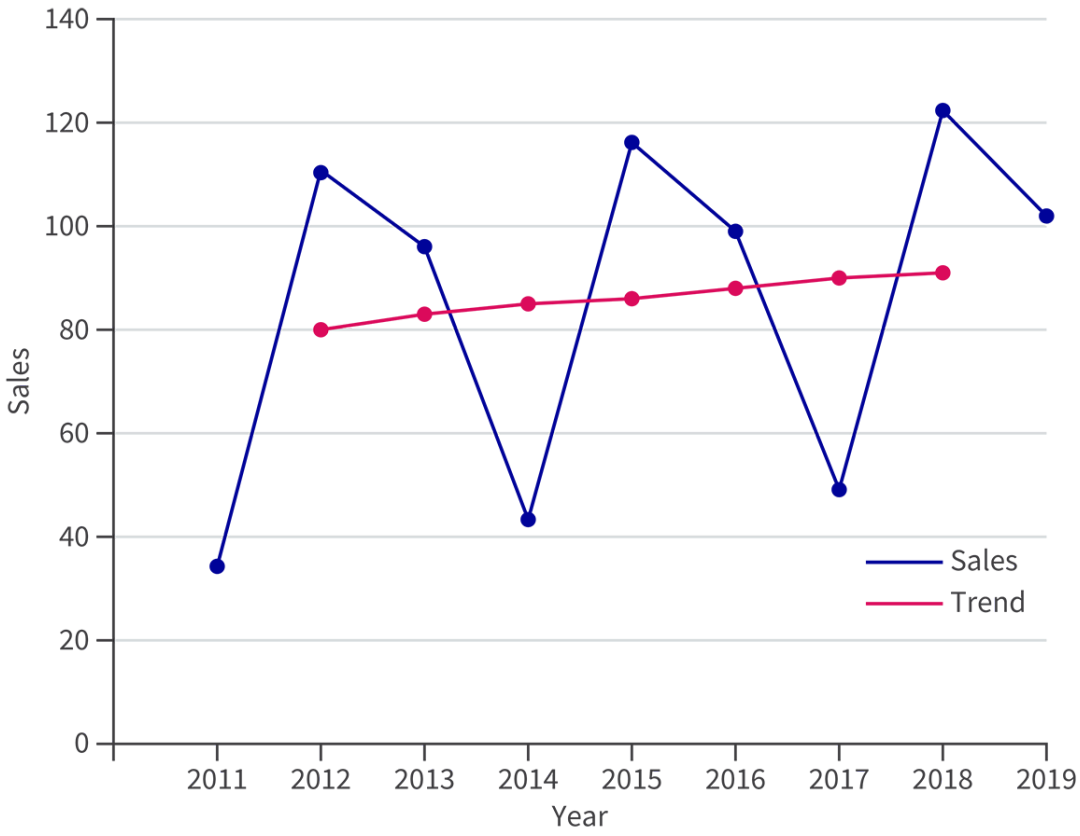

Figure 4. Three-part moving average $= \frac{(34+110+96)}{3} = 80$

Step 2: Extrapolate the trend

Extrapolation of the trend is a forecasting method used by businesses to identify trends using past data, and extending this information and trend to be able to predict what future sales might look like.

Extrapolation is useful if the correlation between the two sets of data is clear, such as sales revenue over a period of time. Smoothing out the data using the three-point moving average (as above) helps make the overall trend in the data clearer. However,

the data representing the three-point moving average is still not linear. To make a prediction into the future, a line of best fit needs to be added, as was done with the scatter diagram earlier in this section.

From the example in **Figure 4**, a line of best fit can be added and extended to predict future sales. In **Figure 5**, the line of best fit has been added and extrapolated (extended) out as far as 2022.

As was mentioned earlier in this section, it is important that the line of best fit goes through the mean values of both the time data and the sales data.

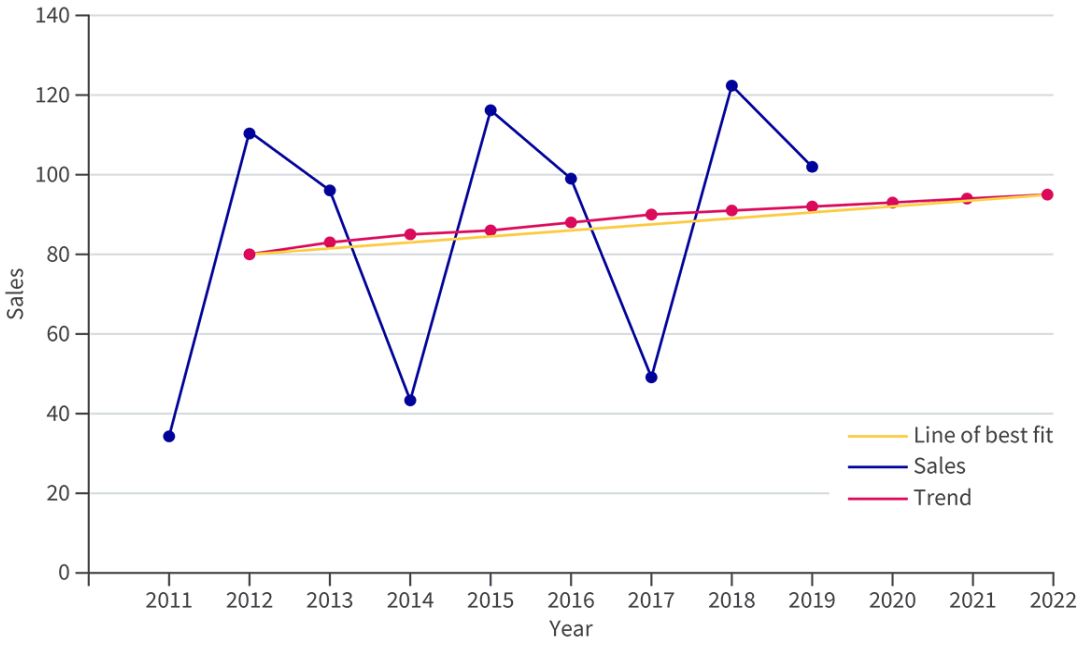

Figure 5. The moving average or trend line extrapolated (extended) out as far as 2022.

You can now clearly see the difference between the blue sales line, and the red line that have been plotted with the three-period moving average data. The blue line is very jagged and moves up and down frequently. Every third year, sales drop quite dramatically. However the general trend for the business is still upwards over the period of time. Using the extrapolated line of best fit, the forecast sales figure for 2011 to 2022 have now been added to the graph above.

You are now in a position to answer the original question: should the business expand? Based on the analysis, the business can predict that it will have sales of just 98 cars in 2022. This is less than the desired figure of 100. Therefore, based on the analysis, expansion would not be recommended at this time.

The same calculations can be made to work out the four-part moving average, where a mean of four years is taken into account, and a line of best fit is plotted and extrapolated.

Exam tip

In the exam, you will not be expected to calculate moving averages. The information above has been outlined so that you understand moving average data if it is given to you in a table in the exam.

However, you are expected to be able to:

- graph sales data
- graph given trend data, which may include moving averages
- sketch a line of best fit that goes through mean values for the independent and dependent variables
- extrapolate the line of best fit to make a sales forecast

Activity

Learner profile: Knowledgeable
Approaches to learning: Thinking skills (critical thinking); Communication skills

Marix produces a variety of sports shoes, such as running and walking shoes. Marix is a market leader that has dominated the sports shoe industry for a long time. Marix manufactures shoes in batches of different ranges.

As the sports shoe market is growing, so is the demand for Marix's sports shoes. Celebrity endorsements have helped increase sales. Recently, however, costs of production are increasing as resources are becoming more expensive. The business is experiencing diseconomies of scale.

The mean sales per month for Marix for the years 2015 to 2021 are given in **Table 6**.

Table 6. Marix's mean sales of shoes per month from 2015 to 2021.

Year	Mean sales of shoes per month
2015	185
2016	250
2017	400

Year	Mean sales of shoes per month
2018	510
2019	700
2020	925
2021	950

Questions

1. Calculate the mean of the mean sales per month for Marix. (You have studied mean in Section 4.4.6 (/study/app/y12-business-management-a-hl-may-2024/sid-351-cid-174702/book/tool-descriptive-statistics-id-39001).)
2. Calculate the mean year (Section 4.4.6 (/study/app/y12-business-management-a-hl-may-2024/sid-351-cid-174702/book/tool-descriptive-statistics-id-39001)).
3. Using graph paper, plot the mean sales of Marix per year from 2015 to 2021. Label your graph clearly.
4. On the graph, construct a line of best fit through the mean sales data obtained from question 2.
5. On the graph, extrapolate a value for mean sales in 2022 and 2023 from the line of best fit.