4

CHAPTER 4

Investigating and modelling time series data

LESSONS

KEY KNOWLEDGE

- qualitative features of time series plots; recognition of features such as trend (long-term direction), seasonality (systematic, calendar related movements) and irregular fluctuations (unsystematic, short-term fluctuations); possible outliers and their sources, including one-off real-world events, and signs of structural change such as a discontinuity in the time series
- numerical smoothing of time series data using moving means with consideration of the number of terms required (using centring when appropriate) to help identify trends in time series plot with large fluctuations
- graphical smoothing of time series plots using moving medians (involving an odd number of points only) to help identify long-term trends in time series with large fluctuations
- seasonal adjustment including the use and interpretation of seasonal indices and their calculation using seasonal and yearly means
- modelling trend by fitting a least squares line to a time series with time as the explanatory variable (data de-seasonalised where necessary), and the use of the model to make forecasts (with re-seasonalisation where necessary) including consideration of the possible limitations of fitting a linear model and the limitations of extending into the future.

4A Time series data and their graphs

STUDY DESIGN DOT POINT

- qualitative features of time series plots; recognition of features such as trend (long-term direction), seasonality (systematic, calendar related movements) and irregular fluctuations (unsystematic, short-term fluctuations); possible outliers and their sources, including one-off real-world events, and signs of structural change such as a discontinuity in the time series

KEY SKILLS

During this lesson, you will be:

- constructing time series plots
- identifying characteristics of time series data.

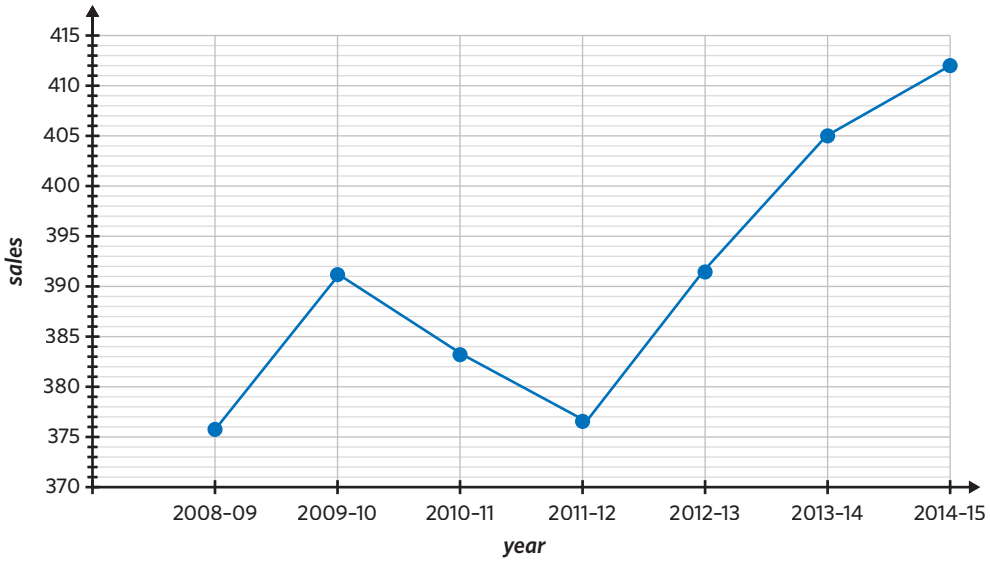
KEY TERMS

- Time series
- Coded time
- Trend
- Period
- Cyclical variation
- Seasonality
- Structural change
- Irregular fluctuations

The changes in a variable over time can be represented by a time series graph. Time can be expressed in many ways – as hours, days, weeks, months, years, seasons, etc. By graphing time series data, any patterns or changes in the response variable over time can be observed.

Constructing time series plots

Time series data is a subset of bivariate data where the explanatory variable is always time. Since time is the explanatory variable, it is always plotted on the horizontal axis. When plotting a time series, the known data points are connected by straight lines. These lines are useful in identifying any trends or changes in the data.



It can sometimes be inefficient to use the full dates when plotting time series data. Instead, simple numerical representations such as 1, 2, 3, 4... can be used. The simplified numerical representation of time is referred to as the **coded time**. Calculations to the time series can also be applied. For example, summer 2022, autumn 2022, winter 2022, and spring 2022 can be represented by the codes 1, 2, 3, and 4.

Worked example 1

The number of *deaths and serious injuries* caused by road accidents in the UK for each month from October 1982 until the end of 1983 are shown in a table.

month	Oct 1982	Nov 1982	Dec 1982	Jan 1983	Feb 1983	Mar 1983	Apr 1983	May 1983	Jun 1983	Jul 1983	Aug 1983	Sep 1983	Oct 1983	Nov 1983	Dec 1983
deaths and serious injuries	1850	1998	2079	1494	1057	1218	1168	1236	1076	1174	1139	1427	1487	1483	1513

a. Use the table to construct a time series plot.

Explanation – Method 1: By hand

Step 1: Create a time code.

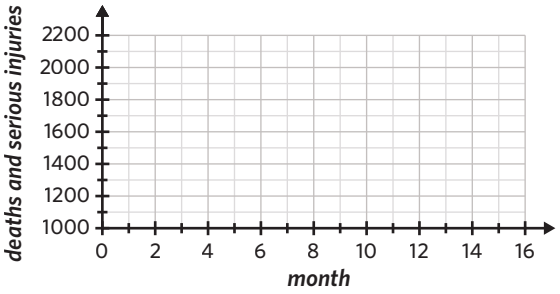
month	***coded month***
Oct 1982	1
Nov 1982	2
Dec 1982	3
Jan 1983	4
Feb 1983	5
Mar 1983	6
Apr 1983	7
May 1983	8
Jun 1983	9
Jul 1983	10
Aug 1983	11
Sep 1983	12
Oct 1983	13
Nov 1983	14
Dec 1983	15

Step 2: Draw a set of axes and label the horizontal axis '*month*' and the vertical axis '*deaths and serious injuries*'.

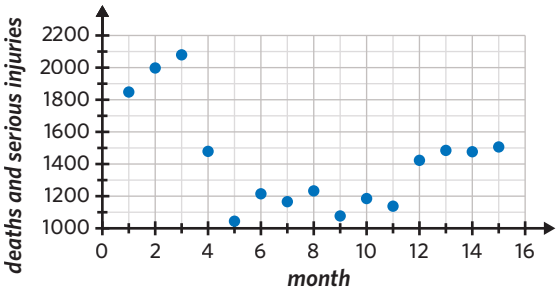
Step 3: Construct an appropriate scale for each axis.

The horizontal axis should range from at least 0 to 15.

The vertical axis should range from at least 1000 to 2100.

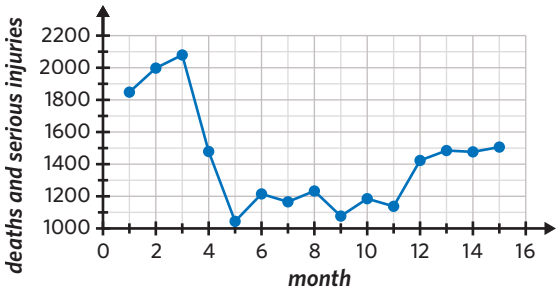


Step 4: Plot each data point on the graph.



Step 5: Connect the data points in order with straight lines.

Answer



Continues →

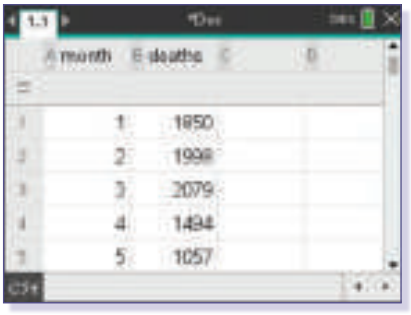
Explanation – Method 2: TI-Nspire

Step 1: Create a time code.

month	*coded month*
Oct 1982	1
Nov 1982	2
Dec 1982	3
Jan 1983	4
Feb 1983	5
Mar 1983	6
Apr 1983	7
May 1983	8
Jun 1983	9
Jul 1983	10
Aug 1983	11
Sep 1983	12
Oct 1983	13
Nov 1983	14
Dec 1983	15

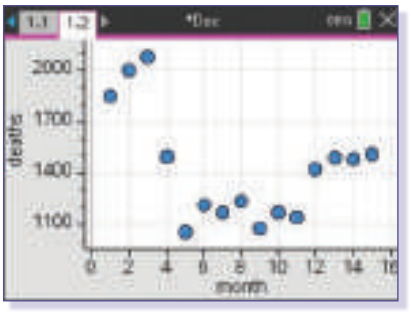
Step 2: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 3: Name column A 'month' and enter the data values starting from row 1 into the column below. Name column B 'deaths' and enter the data values starting from row 1 into the column below.



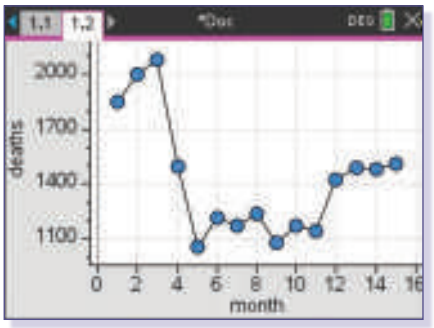
Step 4: Press ctrl + doc and select '5: Add Data & Statistics'.

Step 5: Move to the horizontal axis and select 'Click to add variable' → 'month'. Move to the vertical axis and select 'Click to add variable' → 'deaths'.



Step 6: Press menu. Select '2: Plot Properties' → '1: Connect Data Points'.

Answer



Continues →

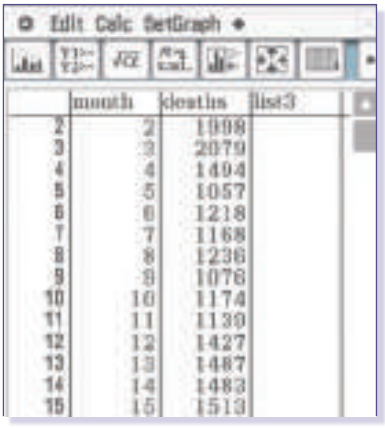
Explanation – Method 3: Casio ClassPad

Step 1: Create a time code.

month	*coded month*
Oct 1982	1
Nov 1982	2
Dec 1982	3
Jan 1983	4
Feb 1983	5
Mar 1983	6
Apr 1983	7
May 1983	8
Jun 1983	9
Jul 1983	10
Aug 1983	11
Sep 1983	12
Oct 1983	13
Nov 1983	14
Dec 1983	15

Step 2: From the main menu, tap Statistics.

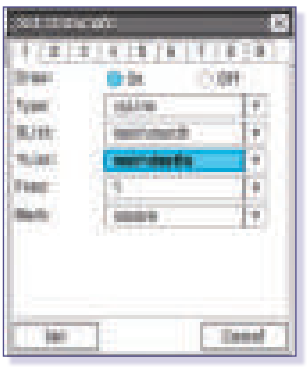
Step 3: Name list1 'month' and enter the data values starting from row 1 into the column below.
Name list2 'deaths' and enter the data values starting from row 1 into the column below.



Step 4: Configure the settings of the graph by tapping .

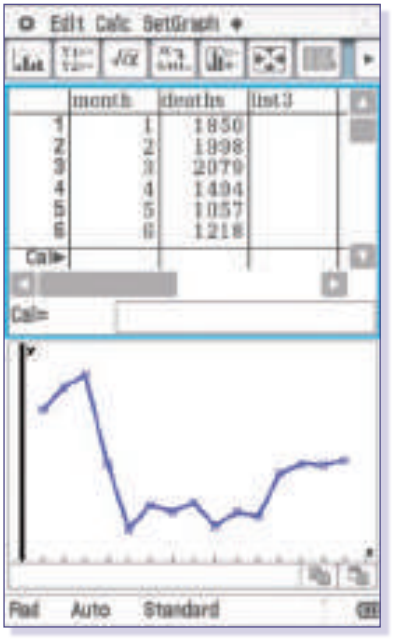
Step 5: Create a time series plot by changing 'Type' to 'xyLine'.

Step 6: Specify the data set by changing 'XList:' to 'main\month' and 'YList:' to 'main\deaths'.



Step 7: Tap 'Set' to confirm and then to plot the graph.

Answer



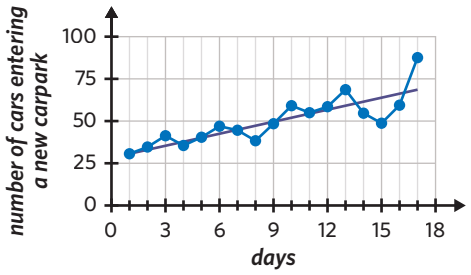
Identifying characteristics of time series data

There are many types of fluctuations or patterns that can be observed in time series data.

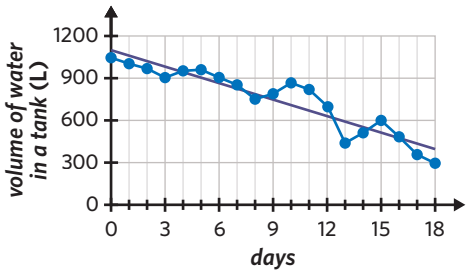
Trends

A **trend** is a general upwards (increasing) or downwards (decreasing) movement over time. This movement can be represented using a trend line.

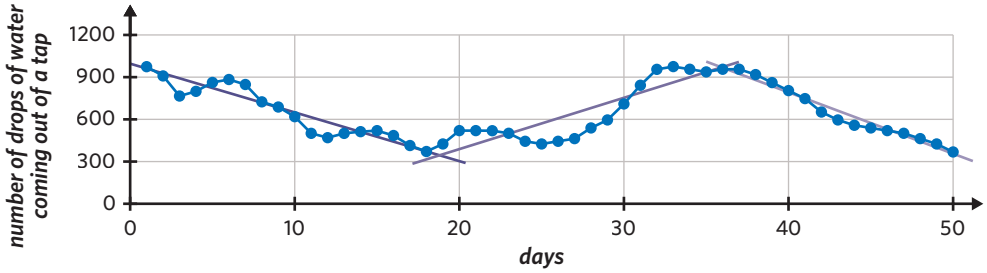
Increasing trend:



Decreasing trend:



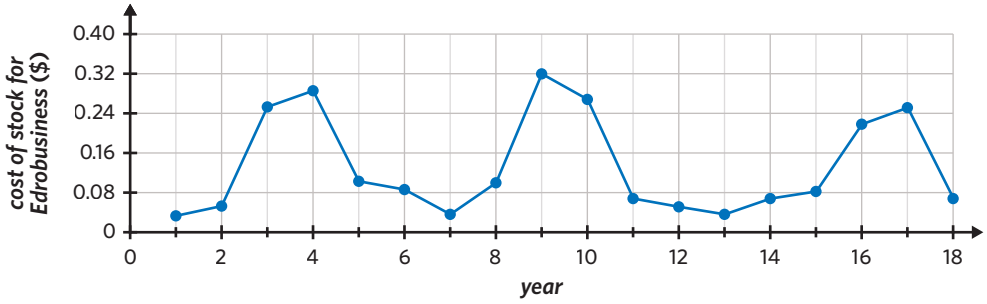
Sometimes, a time series may have multiple trends that change over time.



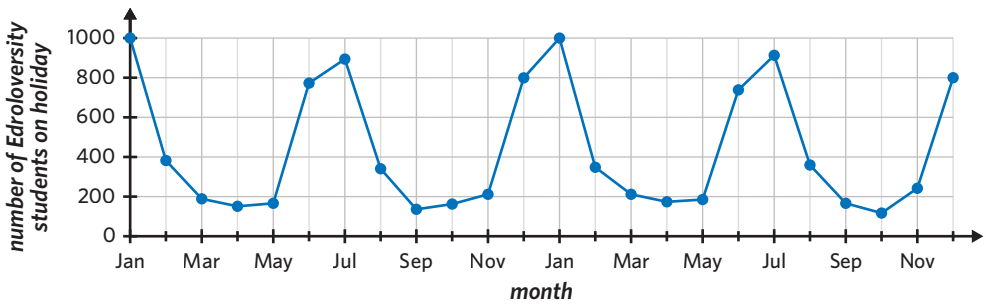
Seasonality & cyclical variation

When a graph shows numerous fluctuations, or peaks and troughs, the **period** is defined as the distance between two adjacent peaks. A period may be regular or irregular. The length of a period can be useful in identifying whether a graph demonstrates cyclical variation or seasonality.

When a time series plot rises and falls with a regular pattern over some time period, it demonstrates **cyclical variation**, or cycles. Whilst the peaks of cycles occur at approximately the same intervals, cycles can have a period which changes slightly between peaks.

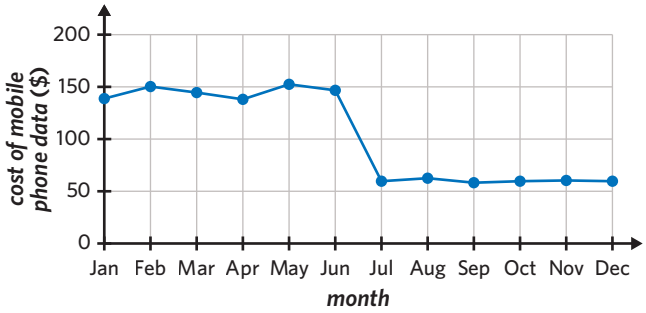


When a time series plot shows cyclical variation within a calendar-related period (e.g. week, month, quarter), the plot demonstrates **seasonality**. Note that the name seasonal does not have to mean the seasons of the year. A seasonal time series plot has regular peaks and troughs that occur at the same time each period, and the length of the period must be a year or less.

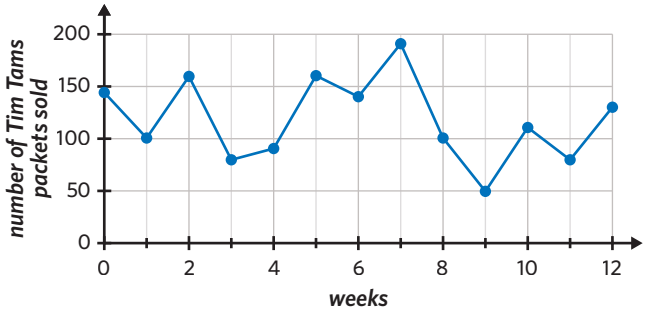


Structural change, irregular fluctuations & outliers

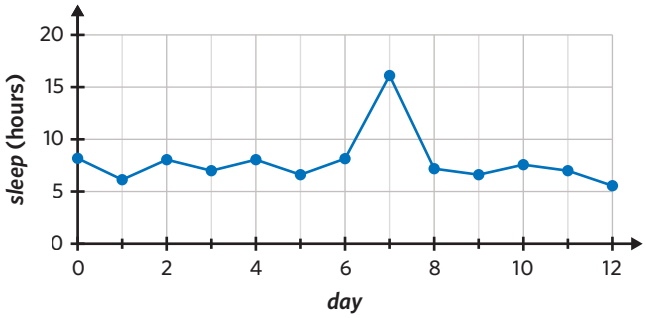
A **structural change** occurs when the established pattern of a time series plot is suddenly altered significantly for some reason.



Irregular fluctuations are random variations in a time series plot that cannot be explained by trend, seasonality, cycles or structural change. Irregular fluctuations cannot be predicted.



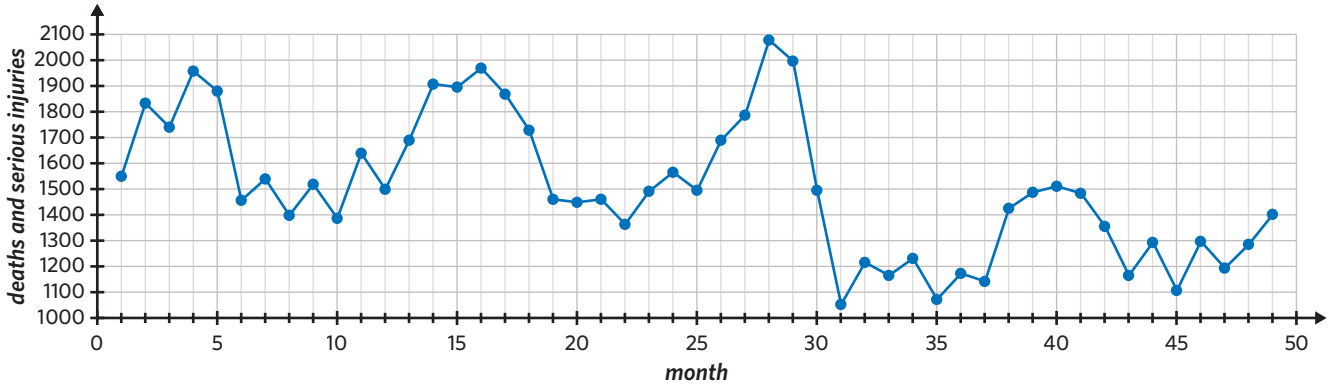
Outliers are values which fall outside of what looks normal or reasonable. There may be a situational explanation for the presence of an outlier.



Worked example 2

The number of *deaths and serious injuries* caused by road accidents in the UK for each month over a four-year period, from September 1980 until September 1984, are shown in the time series plot.

Note that seat belt use became mandatory by law in the UK in February 1983.



Describe any patterns present in the time series plot.

Continues →

Explanation

Step 1: Look for a general trend in the data.

There is no general trend in this time series plot. The peaks and troughs in the data show no clear increasing or decreasing trend.

Step 2: Look for any patterns in the data.

This time series plot has a clear pattern of regular peaks and troughs. The peaks occur every 12 months and are generally at their highest point in December (note that the time series plot begins in September). The fluctuations follow a calendar-related period suggesting that the pattern shows seasonality and not cyclical variation.

The time series plot also has a clear structural change shown by a decrease in *deaths and serious injuries* from month 29 to month 31. This structural change can be explained by the introduction of seatbelt laws in February 1983 (month 30).

No notable irregular fluctuations or outliers can be found in the time series.

Step 3: Present your findings as a brief report.

Answer

The time series plot has no obvious trend. The plot shows seasonality with a period of 12 months peaking every December. There is also a structural change in the plot that occurs between month 29 and month 31. This structural change can be explained by the introduction of mandatory seat belt usage laws in February 1983.

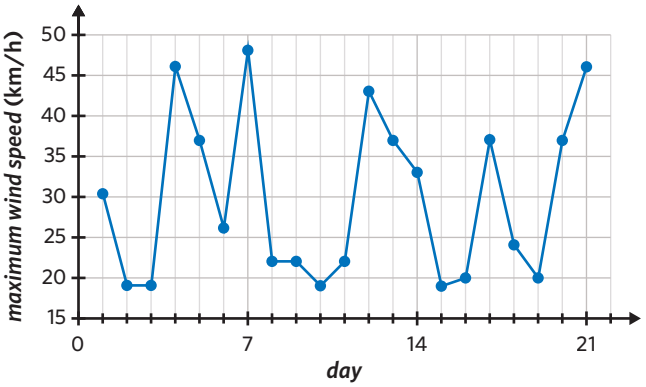
Exam question breakdown

VCAA 2017 Exam 1 Data analysis Q13

The wind speed at a city location is measured throughout the day. The time series plot shows the daily *maximum wind speed*, in kilometres per hour, over a three-week period.

The time series is best described as having

A. seasonality only.
B. irregular fluctuations only.
C. seasonality with irregular fluctuations.
D. a decreasing trend with irregular fluctuations.
E. an increasing trend with irregular fluctuations.



Explanation

Step 1: Look for a general trend in the data.

There is no general trend in this time series plot. The peaks and troughs in the data show no clear pattern of increase or decrease.

Step 2: Look for any seasonality.

Since we have been given data spanning 21 days, the only feasible calendar-related period we could identify would be weekly. In order for there to be weekly seasonality, the peaks and troughs would need to occur at the same time each week (for example, day 1, day 8, and day 15). They do not occur at the same time each week, so there is no seasonality present.

Step 3: Describe the time series.

The time series has no general trend or seasonality. It also has no clear outliers, structural change, or cyclical variation. The time series can be described as having irregular fluctuations only.

Answer

B

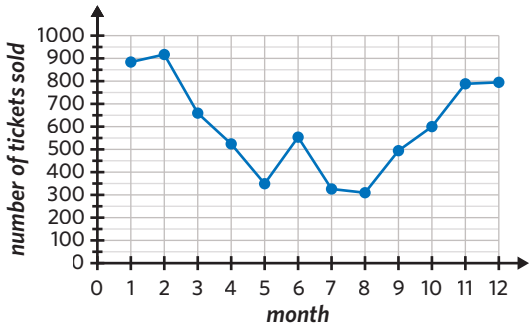
51% of students answered this question correctly.

32% of students incorrectly chose option C. These students incorrectly identified the time series graph as showing seasonality. Although there are peaks and troughs in the time series, they do not exist with regular calendar-related intervals of time between them, so seasonality cannot be concluded.

4A Questions

Constructing time series plots

1. Consider the following time series showing the number of tickets sold to Melbourne Zoo's hippopotamus show 'Hypnotic Hippos' each month this year. The *month* data has been represented by a time code, so that January is 1, February is 2, etc.



How many tickets were sold in June?

A. 320 tickets

B. 345 tickets

C. 550 tickets

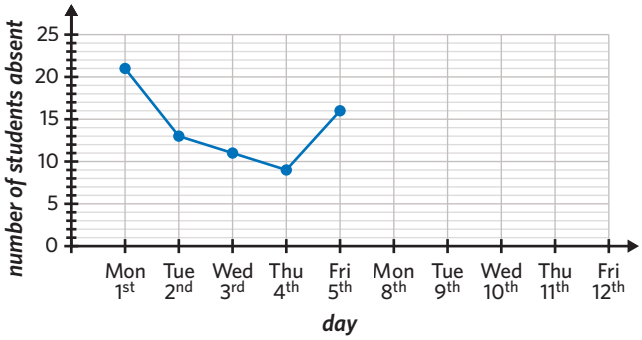
D. 910 tickets

---

2. The *number of students absent* from school each day was recorded over two weeks in April.

day	Mon 1st	Tue 2nd	Wed 3rd	Thu 4th	Fri 5th	Mon 8th	Tue 9th	Wed 10th	Thu 11th	Fri 12th
number of students absent	21	13	11	9	16	20	8	14	11	18

The data from the first week has been plotted onto the following graph. Complete the time series by plotting the data from the second week of April.



---

3. The number of *climate change sceptics* (%), as a percentage of the total population, each year from 2000–2013 is displayed in the table. Use the table to construct a time series plot using a calculator.

year	***coded year***	***climate change sceptics* (%)**
2000	1	62
2001	2	58
2002	3	57
2003	4	54
2004	5	51
2005	6	52
2006	7	40
2007	8	39
2008	9	37
2009	10	36
2010	11	37
2011	12	34
2012	13	35
2013	14	32

4. The *earnings* ($000's) of a casino were recorded for each month in 1983–84 and are given in the table. Construct a time series plot displaying this data by hand.

month	*earnings* ($000's)
Jan 1983	2400
Feb 1983	1600
Mar 1983	3200
Apr 1983	2100
May 1983	1400
Jun 1983	2200
Jul 1983	2700
Aug 1983	3200
Sep 1983	2500
Oct 1983	1900
Nov 1983	2400
Dec 1983	1600

month (cont'd)	*earnings* ($000's)
Jan 1984	1300
Feb 1984	2100
Mar 1984	2300
Apr 1984	2000
May 1984	1500
Jun 1984	1700
Jul 1984	2600
Aug 1984	3100
Sep 1984	2700
Oct 1984	2900
Nov 1984	3200
Dec 1984	2400

Identifying characteristics of time series data

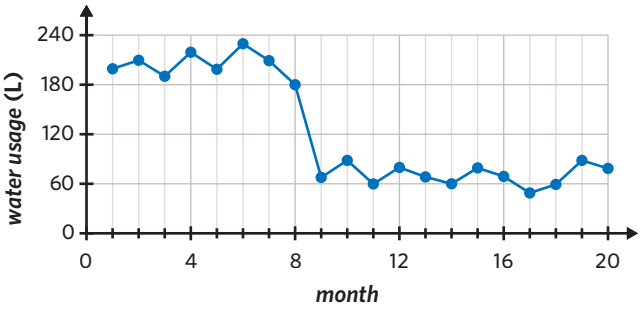
5. Identify which one of the following features is present in the time series plot.

A. Structural change

B. Outlier

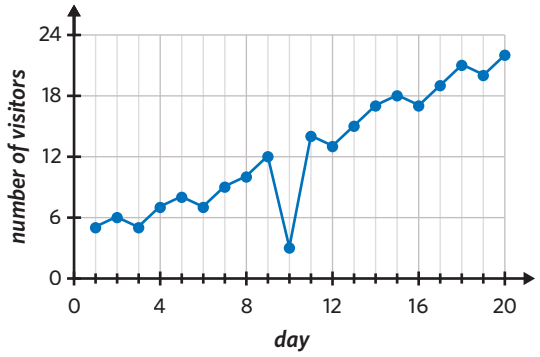
C. Seasonality

D. A decreasing trend

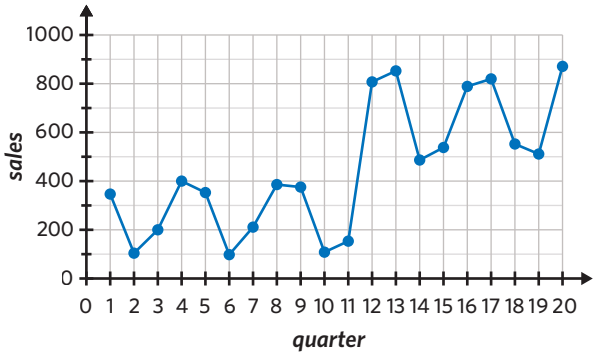


6. Identify two features that are present in each of the following time series plots.

a.

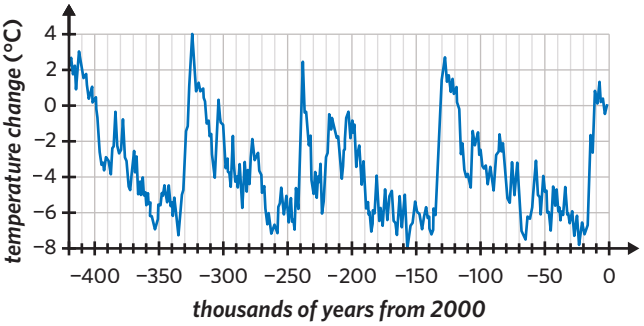


b.

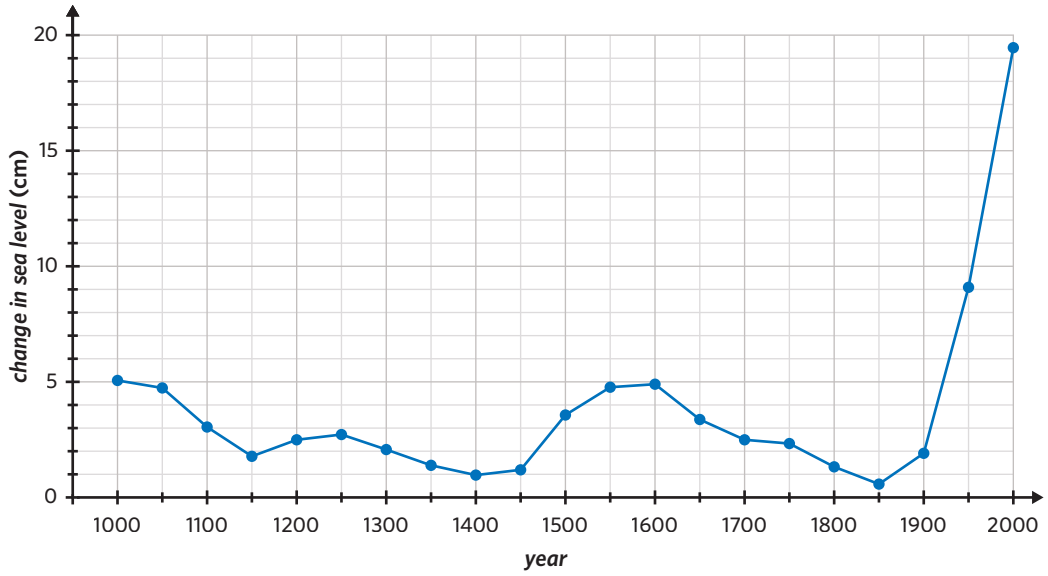


7. The following graph shows the change in global temperature (°C) since approximately 400 000 years before the year 2000.

Identify any trends or patterns in the data.



8. The following graph shows the *change in sea level* (cm) since the year 1000.



Identify any trends or patterns in the data.

Joining it all together

9. The following table shows the yearly *revenue* of a restaurant from 2012 to 2022.

year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
***revenue* ($000's)**	950	890	935	830	850	760	810	740	350	710	690

a. Construct a time series plot by hand displaying the restaurant's *revenue* between 2012 and 2022.

b. Briefly describe any patterns or characteristics in the time series plot.

---

10. The profits between 2018 and 2021 of a buy-now pay-later business are shown in the following table.

a. Complete the time code by filling in the 'coded season' column.

b. Construct a time series plot using a calculator to display the data.

c. Describe any trends or patterns in the time series plot between Summer 2018 and Summer 2020, inclusive.

d. Describe any characteristics present from Summer 2020 to Autumn 2020.

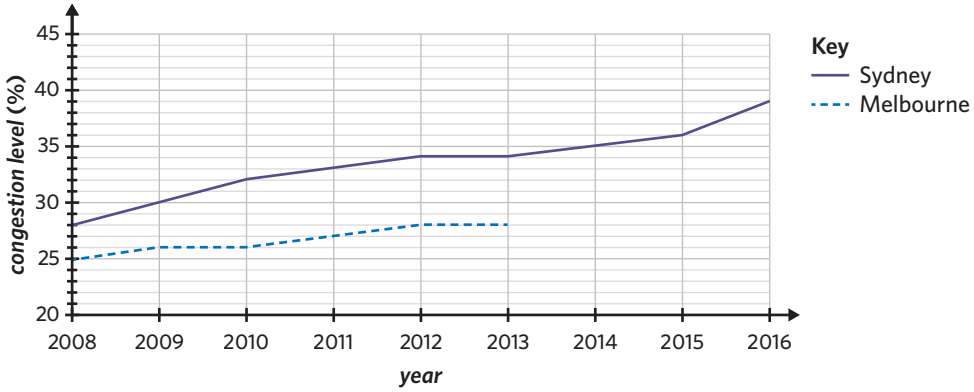
e. Describe any trends or patterns in the time series plot after Autumn 2020.

season	***coded season***	***profit* ($000 000's)**
Summer 2018		18
Autumn 2018		12
Winter 2018		13
Spring 2018		17
Summer 2019		22
Autumn 2019		16
Winter 2019		15
Spring 2019		21
Summer 2020		27
Autumn 2020		47
Winter 2020		50
Spring 2020		48
Summer 2021		47
Autumn 2021		43
Winter 2021		41
Spring 2021		40

Exam practice

11. The following table shows the yearly average traffic congestion levels in two cities, Melbourne and Sydney, during the period 2008 to 2016. Also shown is a time series plot of the same data. The time series plot for Melbourne is incomplete.

	***congestion level* (%)**								
year	2008	2009	2010	2011	2012	2013	2014	2015	2016
Melbourne	25	26	26	27	28	28	28	29	33
Sydney	28	30	32	33	34	34	35	36	39



Use the data in the table to complete the time series plot for Melbourne. (1 MARK)

VCAA 2018 Exam 2 Data analysis Q3a

88% of students answered this question correctly.

12. Consider the time series plot.

The pattern in the time series plot shown is best described as having

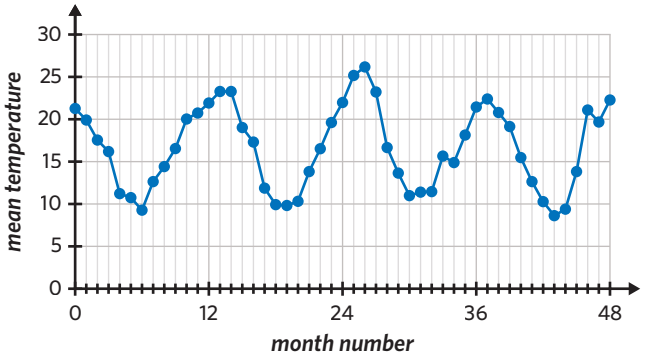
A. irregular fluctuations only.

B. an increasing trend with irregular fluctuations.

C. seasonality with irregular fluctuations.

D. seasonality with an increasing trend and irregular fluctuations.

E. seasonality with a decreasing trend and irregular fluctuations.



VCAA 2016 Exam 1 Data analysis Q13

72% of students answered this question correctly.

13. The time series plot shows the quarterly *sales*, in thousands of dollars, of a small business for the years 2010 to 2020.

The time series plot is best described as having

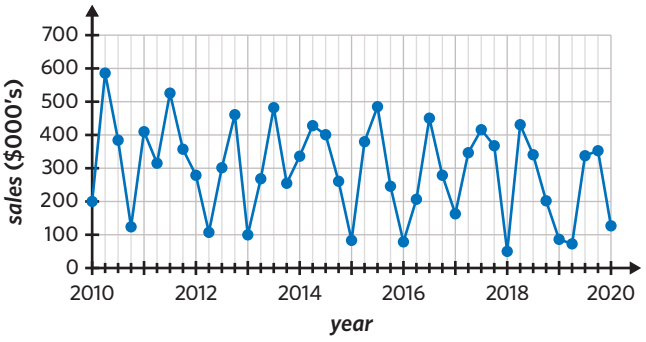
A. seasonality only.

B. irregular fluctuations only.

C. seasonality with irregular fluctuations.

D. a decreasing trend with irregular fluctuations.

E. a decreasing trend with seasonality and irregular fluctuations.



VCAA 2021 Exam 1 Data analysis Q12

34% of students answered this question correctly.

Questions from multiple lessons

Data analysis *Year 11 content*

14. There is found to be a strong negative association between the time spent on social media and average test score of a sample of students. Which of the following conclusions can be drawn from this information?

A. A decrease in the average test score of a student will cause a decrease in their time spent on social media.

B. An increase in the average test score of a student will cause a decrease in their time spent on social media.

C. Students who spend more time on social media tend to have lower average test scores.

D. Students who spend less time on social media tend to have lower average test scores.

E. The association must be due to coincidence.

Adapted from VCAA 2016 Exam 1 Data analysis Q12

Recursion and financial modelling *Year 11 content*

15. Alessandro deposited \$4400 into a savings account with an interest rate of 3.5% per annum, compounding annually. Which of the following recurrence relations models the balance of the savings account, B_n, after n years?

A. $B_0 = 4400, \quad B_{n+1} = 3.5 \times B_n$

B. $B_0 = 4400, \quad B_{n+1} = B_n + 3.5 \times 154$

C. $B_0 = 4400, \quad B_{n+1} = B_n + 154$

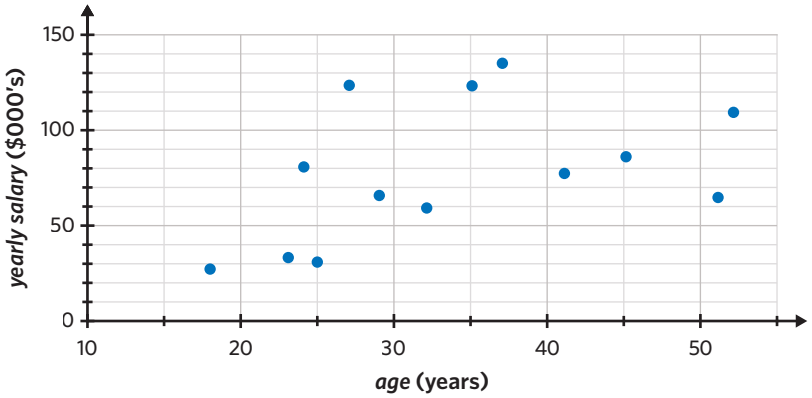
D. $B_0 = 4400, \quad B_{n+1} = 154 \times B_n$

E. $B_0 = 4400, \quad B_{n+1} = 1.035 \times B_n$

Adapted from VCAA 2017NH Exam 1 Recursion and financial modelling Q18

Data analysis

16. The *age*, in years, and *yearly salary*, in dollars, of 13 office workers are shown in the following scatterplot and table.



age (years)	*yearly salary* (\$)
18	27 000
23	33 500
24	81 000
25	31 000
27	123 000
29	66 000
32	59 000
35	122 500
37	134 500
41	77 000
45	86 000
51	64 500
52	109 000

In this sample, the relationship between *age* and *yearly salary* is non-linear.

In order to linearise the data, a log transformation can be applied to the variable *age*.

a. Apply the log transformation and calculate the equation of the least squares regression line that predicts *yearly salary* from $\log_{10}(age)$. Give values correct to one decimal place. (1 MARK)

b. Using this rounded regression equation, predict the *yearly salary* of someone 33 years of age. Give your answer correct to the nearest dollar. (1 MARK)

Adapted from VCAA 2014 Exam 2 Data analysis Q3

4B Smoothing – moving means

STUDY DESIGN DOT POINT

- numerical smoothing of time series data using moving means with consideration of the number of terms required (using centring when appropriate) to help identify trends in time series plot with large fluctuations

4A 4B 4C 4D 4E

KEY SKILLS

During this lesson, you will be:

- smoothing over an odd number of data points using moving means
- smoothing over an even number of data points using moving means
- plotting and interpreting a mean smoothed time series.

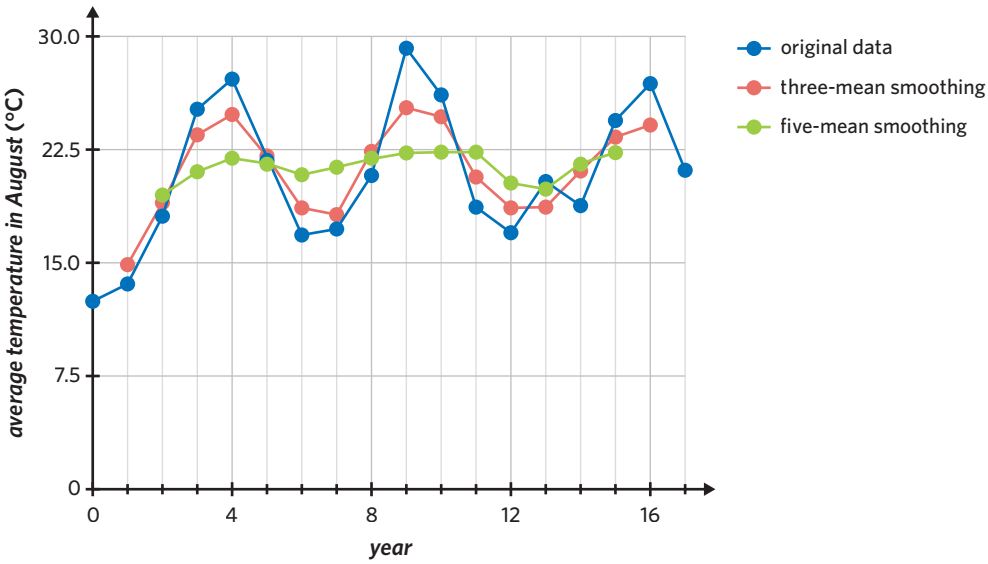
KEY TERMS

- Smoothing
- Moving mean smoothing
- Centring

When a time series contains many fluctuations, it is often difficult to identify the underlying trend. In order to reveal the trend, smoothing can be implemented to remove some of the fluctuations. Moving mean smoothing is one method used to smooth time series data.

Smoothing over an odd number of data points using moving means

Smoothing is the process of removing fluctuations in time series data to reveal any underlying trends. It is possible to smooth using either moving means or moving medians. Smoothing using means is known as **moving mean smoothing**.



Three-mean smoothing involves replacing each data value with the mean of itself and each adjacent value.

The smoothed value of y_2:

$$\textit{smoothed } y_2 = \frac{y_1 + y_2 + y_3}{3}$$

There are no smoothed values for the first and last data entries because they do not have a value on either side.

day	***temp.* (°C)**	***calculation***	**three-mean smoothed *temperature* (°C)**
Mon	24	–	–
Tue	27	$\frac{24 + 27 + 21}{3}$	24
Wed	21	$\frac{27 + 21 + 18}{3}$	22
Thu	18	$\frac{21 + 18 + 15}{3}$	18
Fri	15	$\frac{18 + 15 + 15}{3}$	16
Sat	15	$\frac{15 + 15 + 12}{3}$	14
Sun	12	–	–

Five-mean smoothing is similar to three-mean smoothing but uses five values. Each data value is replaced with the mean of itself and the two values on each side of it.

The smoothed value of y_3:

$$\textit{smoothed } y_3 = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$$

There are no smoothed values for the first two and last two data entries because they do not have two values on either side.

day	***temp.* (°C)**	***calculation***	**five-mean smoothed *temperature* (°C)**
Mon	24	–	–
Tue	27	–	–
Wed	21	$\frac{24 + 27 + 21 + 18 + 15}{5}$	21
Thu	18	$\frac{27 + 21 + 18 + 15 + 15}{5}$	19.2
Fri	15	$\frac{21 + 18 + 15 + 15 + 12}{3}$	16.2
Sat	15	–	–
Sun	12	–	–

Moving mean smoothing can be extended over any number of data points. Smoothing over any odd number of data points is done in the same way as three-mean and five-mean smoothing.

Worked example 1

The following table shows the daily *rainfall* in Dunedin for a week.

day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
***rainfall* (mm)**	2	3	5	5	1	0	2
three-mean smoothed *rainfall*	–						–
five-mean smoothed *rainfall*	–	–				–	–

Continues →

a. Find the three-mean smoothed *rainfall* for Wednesday, correct to one decimal place.

Explanation

Step 1: Find the *rainfall* for Wednesday as well as its adjacent values, and write them in the order they appear in the time series.

3 5 5

Step 2: Find the mean of these three values.

$$mean = \frac{3 + 5 + 5}{3}$$
$$= 4.333...$$

Answer

4.3 mm

---

b. Find the three-mean smoothed values, correct to one decimal place, for the entire time series and fill in the table.

Explanation – Method 1: TI-Nspire

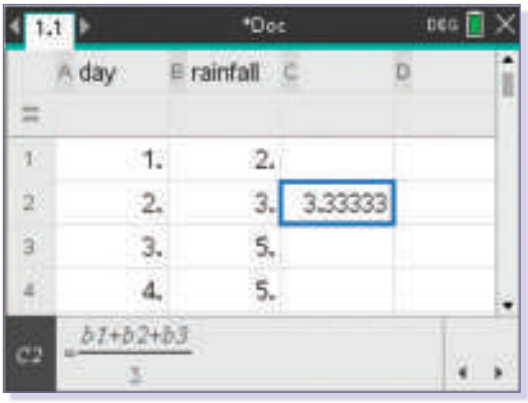
Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Enter the numbers from 1 to 7 in the first column. Each number will represent a day of the week. Name the column 'day'.

Step 3: Enter the *rainfall* values into the second column and name the column 'rainfall'.

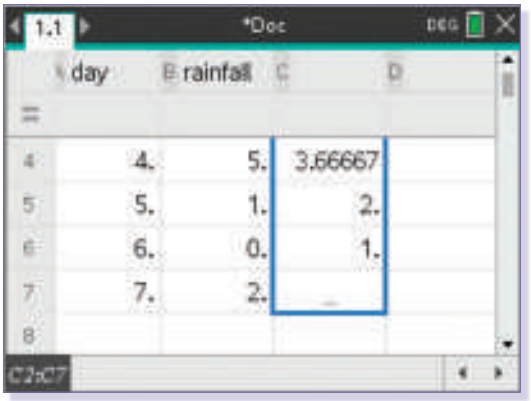
Step 4: Select cell C2. Type '=(b1+b2+b3)/3' to find the smoothed *rainfall* for day 2 (Tuesday).

Press enter.



Note: C1 is left blank as there is no three-mean smoothed value for the first data point.

Step 5: With cell C2 still selected, move the cursor to the bottom right corner of the cell and click and drag downwards to apply the formula to the remaining rows.



Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Spreadsheet.

Step 2: Enter the numbers from 1 to 7 in the first column. Each number will represent a day of the week.

Step 3: Enter the *rainfall* values into the second column.

Continues →

Step 4: Select cell C2. Type '=(b1+b2+b3)/3' to find the smoothed *rainfall* for day 2 (Tuesday). Press EXE.

Note: C1 is left blank as there is no three-mean smoothed value for the first data point.

Step 5: With cell C2 still selected, drag from C2 down to C3. This will copy the formula down to the next cell. Repeat this until cell C6.

Answer – Method 1 and 2

day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
***rainfall* (mm)**	2	3	5	5	1	0	2
three-mean smoothed *rainfall*	–	3.3	4.3	3.7	2.0	1.0	–

Note: Alternatively, this question could be solved manually using the technique from part **a**.

---

c. Find the five-mean smoothed *rainfall* for Wednesday.

Explanation

Step 1: Find the *rainfall* for Wednesday as well as the two values on each side of it, and write them in the order they appear in the time series.

2 3 5 5 1

Step 2: Find the mean of these five values.

$$mean = \frac{2 + 3 + 5 + 5 + 1}{5} = 3.2$$

Answer

3.2 mm

---

d. Find the five-mean smoothed values, correct to one decimal place, for the entire time series and fill in the table.

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

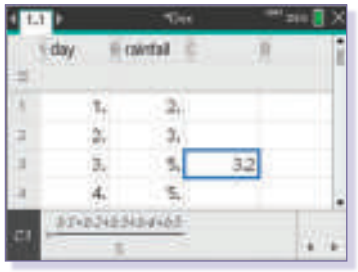
Step 2: Enter the numbers 1 to 7 in the first column. Each number will represent a day of the week. Name the column 'day'.

Step 3: Enter the *rainfall* values into the second column and name the column 'rainfall'.

Continues →

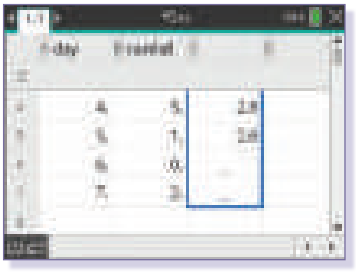
Step 4: Select cell C3. Type '=(b1+b2+b3+b4+b5)/5' to find the smoothed *rainfall* for day 3 (Wednesday).

Press enter.



Note: C1 and C2 are left blank as there are no five-mean smoothed values for the first two data points.

Step 5: With cell C3 still selected, move the cursor to the bottom right corner of the cell and click and drag downwards to apply the formula to the remaining rows.



Explanation – Method 2: Casio ClassPad

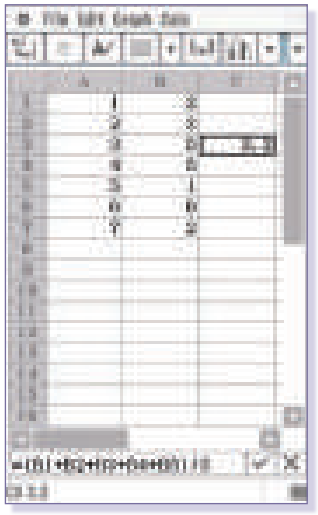
Step 1: From the main menu, tap Spreadsheet.

Step 2: Enter the numbers 1 to 7 in the first column. Each number will represent a day of the week.

Step 3: Enter the *rainfall* values into the second column.

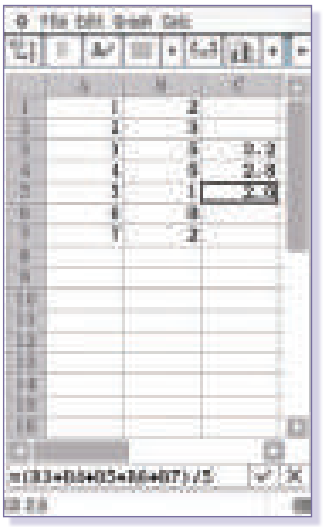
Step 4: Select cell C3. Type '=(b1+b2+b3+b4+b5)/5' to find the smoothed *rainfall* for day 3 (Wednesday).

Press EXE.



Note: C1 and C2 are left blank as there are no five-mean smoothed values for the first two data points.

Step 5: With cell C3 still selected, drag from C3 down to C4. This will copy the formula down to the next cell. Repeat this for cell C5.



Answer – Method 1 and 2

day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
***rainfall* (mm)**	2	3	5	5	1	0	2
five-mean smoothed *rainfall*	–	–	3.2	2.8	2.6	–	–

Note: Alternatively, this question could be solved manually using the technique from part **c**.

Smoothing over an even number of data points using moving means

Moving mean smoothing over an even number of data points is slightly more complicated because the centre of each set of points is halfway between two data points, and can therefore not be plotted on the original time series. A process called **centring** must be applied in order to align the data with the time series.

Two-mean smoothing with centring involves first calculating two non-centred means. These are the mean of the original data point and each of its adjacent values, in two separate calculations. The centred mean of the original data point is then found from the mean of the two non-centred means.

To find the smoothed value of y_2:

$$\textit{smoothed } y_2 = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2}}{2}$$

This formula can be simplified to:

$$\textit{smoothed } y_2 = \frac{y_1 + 2y_2 + y_3}{4}$$

There are no smoothed values for the first and last data entries because they do not have a value on either side.

day	*temp.* (°C)	*before centring*	*after centring*
Mon	24		–
		$\frac{24 + 27}{2} = 25.5$	
Tue	27		$\frac{25.5 + 24}{2} = 24.75$
		$\frac{27 + 21}{2} = 24$	
Wed	21		$\frac{24 + 19.5}{2} = 21.75$
		$\frac{21 + 18}{2} = 19.5$	
Thu	18		–

Four-mean smoothing with centring is similar to two-mean smoothing but uses four values in the mean calculations. The non-centred means are each calculated from four adjacent values surrounding the original data point. The centred mean of the original data point is then found from the mean of the two non-centred means.

To find the smoothed value of y_3:

$$\textit{smoothed } y_3 = \frac{\frac{y_1 + y_2 + y_3 + y_4}{4} + \frac{y_2 + y_3 + y_4 + y_5}{4}}{2}$$

This formula can be simplified to:

$$\textit{smoothed } y_3 = \frac{y_1 + 2y_2 + 2y_3 + 2y_4 + y_5}{8}$$

There are no smoothed values for the first two and last two data entries because they do not have two values on either side.

day	*temp.* (°C)	*before centring*	*after centring*
Mon	24		–
Tue	27		–
		$\frac{24 + 27 + 21 + 18}{4} = 22.5$	
Wed	21		$\frac{22.5 + 20.25}{2} = 21.375$
		$\frac{27 + 21 + 18 + 15}{4} = 20.25$	
Thu	18		$\frac{20.5 + 17.25}{2} = 18.75$
		$\frac{21 + 18 + 15 + 15}{4} = 17.25$	
Fri	15		$\frac{17.25 + 15}{2} = 16.125$
		$\frac{18 + 15 + 15 + 12}{4} = 15$	
Sat	15		–
Sun	12		–

Smoothing over any even number of data points is done in the same way as two-mean and four-mean smoothing.

Worked example 2

Amelia recorded the distance, in km, she ran each month for a year.

month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
***distance run* (km)**	42	47	39	41	37	35	36	28	31	28	35	41
two-mean smoothed *distance run* with centring (km)	–											–
four-mean smoothed *distance run* with centring (km)	–	–									–	–

a. Find the two-mean smoothed *distance run* for July.

Explanation

Step 1: Find the *distance run* for July as well as its adjacent values and write them in the order they appear in the time series.

35 36 28

Step 2: Find the mean of the first two values.

$mean_1 = \frac{35 + 36}{2} = 35.5$

Step 3: Find the mean of the last two values.

$mean_2 = \frac{36 + 28}{2} = 32$

Step 4: Calculate the centred mean.

$mean_{centred} = \frac{35.5 + 32}{2} = 33.75$

Answer

33.75 km

Continues →

b. Find the two-mean smoothed values, correct to one decimal place, for the entire time series and fill in the table.

Explanation – Method 1: TI-Nspire

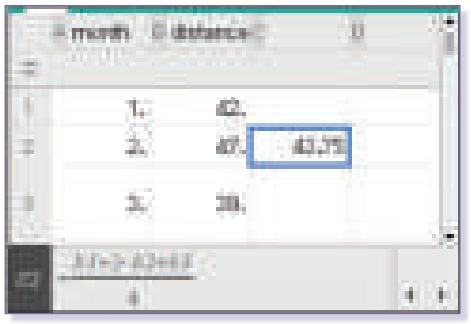
Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Enter the numbers from 1 to 12 in the first column. Each number will represent a month. Name the column 'month'.

Step 3: Enter the *distance run* values into the second column and name the column 'distance'.

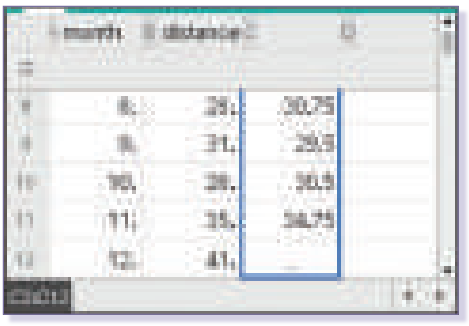
Step 4: Select cell C2. Type '=(b1+2b2+b3)/4' to find the smoothed *distance run* for month 2 (February).

Press enter.



Note: C1 is left blank as there is no two-mean smoothed value for the first data point.

Step 5: With cell C2 still selected, move the cursor to the bottom right corner of the cell and click and drag downwards to apply the formula to the remaining rows.



Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Spreadsheet.

Step 2: Enter the numbers from 1 to 12 in the first column. Each number will represent a month.

Step 3: Enter the *distance run* values into the second column.

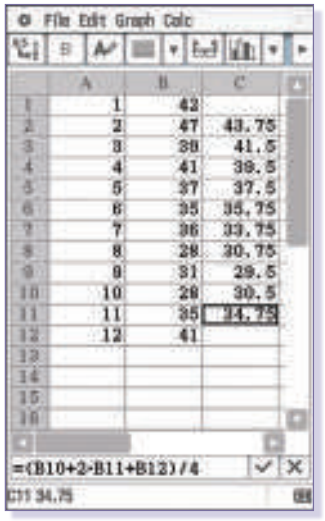
Step 4: Select cell C2. Type '=(b1+2b2+b3)/4' to find the smoothed *distance run* for month 2 (February).

Press EXE.

Note: C1 is left blank as there is no two-mean smoothed value for the first data point.

Continues →

Step 5: With cell C2 still selected, drag from C2 down to C3. This will copy the formula down to the next cell. Repeat this until cell C11.



Answer – Method 1 and 2

month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
***distance run* (km)**	42	47	39	41	37	35	36	28	31	28	35	41
two-mean smoothed *distance run* with centring (km)	–	43.8	41.5	39.5	37.5	35.8	33.8	30.8	29.5	30.5	34.8	–

Note: Alternatively, this question could be solved manually using the technique from part **a**.

---

c. Find the four-mean smoothed *distance run* for July.

Explanation

Step 1: Find the *distance run* for July as well as the two values on either side and write them in the order they appear in the time series.

37 35 36 28 31

Step 2: Find the mean of the first four values.

$$mean_1 = \frac{37 + 35 + 36 + 28}{4} = 34$$

Step 3: Find the mean of the last four values.

$$mean_2 = \frac{35 + 36 + 28 + 31}{4} = 32.5$$

Step 4: Calculate the centred mean.

$$mean_{centred} = \frac{34 + 32.5}{2} = 33.25$$

Answer

33.25 km

---

d. Find the four-mean smoothed values, correct to one decimal place, for the entire time series and fill in the table.

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

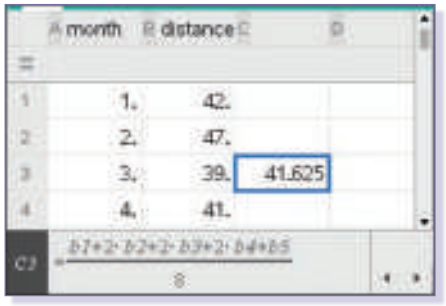
Step 2: Enter the numbers from 1 to 12 in the first column. Each number will represent a month. Name the column 'month'.

Step 3: Enter the *distance run* values into the second column and name the column 'distance'.

Continues →

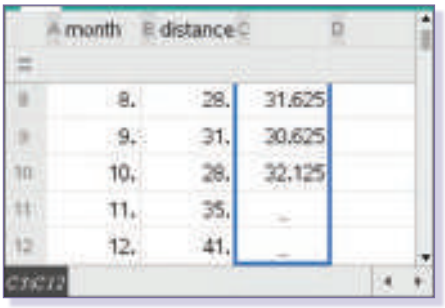
Step 4: Select cell C3. Type '=(b1+2b2+2b3+2b4+b5)/8' to find the smoothed *distance run* for month 3 (March).

Press enter.



Note: C1 and C2 are left blank as there are no four-mean smoothed values for the first two data points.

Step 5: With cell C3 still selected, move the cursor to the bottom right corner of the cell and click and drag downwards to apply the formula to the remaining rows.



Explanation – Method 2: Casio ClassPad

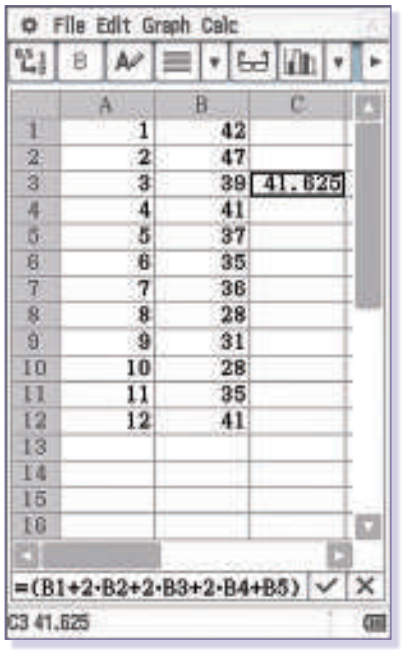
Step 1: From the main menu, tap Spreadsheet.

Step 2: Enter the numbers from 1 to 12 in the first column. Each number will represent a month.

Step 3: Enter the *distance run* values into the second column.

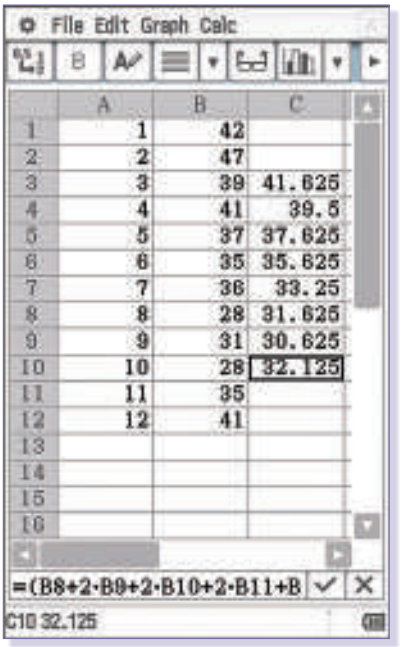
Step 4: Select cell C3. Type '=(b1+2b2+2b3+2b4+b5)/8' to find the smoothed *distance run* for month 3 (March).

Press EXE.



Note: C1 and C2 are left blank as there are no four-mean smoothed values for the first two data points.

Step 5: With cell C3 still selected, drag from C3 down to C4. This will copy the formula down to the next cell. Repeat this until cell C10.



Continues →

Answer – Method 1 and 2

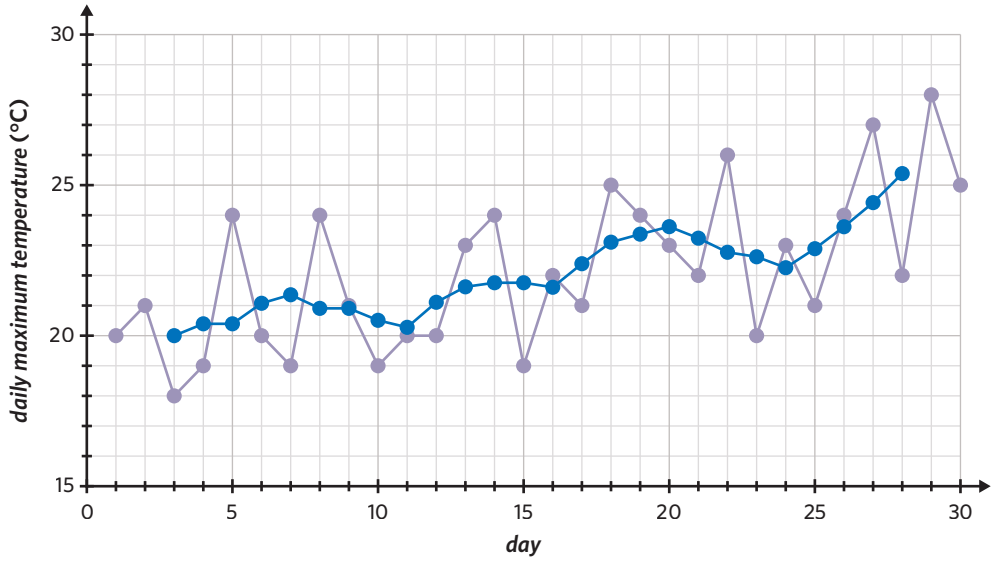
month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
***distance run* (km)**	42	47	39	41	37	35	36	28	31	28	35	41
four-mean smoothed *distance run* with centring (km)	–	–	41.6	39.5	37.6	35.6	33.3	31.6	30.6	32.1	–	–

Note: Alternatively, this question could be solved manually using the technique from part **c**.

Plotting and interpreting a mean smoothed time series

Once data values have been smoothed, they can then be plotted against the original time series. By doing so, any underlying trends become more visible and conclusions can be made about the data.

In the following time series, the *daily maximum temperature* (°C) in November was recorded and then smoothed using four-mean smoothing. The smoothed values show an increasing trend in the data.



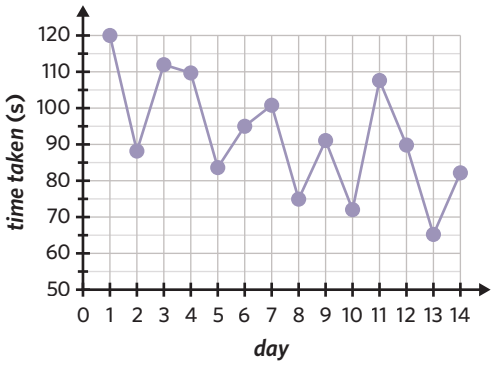
Worked example 3

Tobias wants to get faster at solving his Rubik's Cube. He solves it once a day and records the *time taken* in seconds. The data taken over two weeks is shown in the following table. Tobias has also calculated the five-mean smoothed values for the data set, correct to one decimal place.

day	1	2	3	4	5	6	7	8	9	10	11	12	13	14
***time taken* (s)**	120	88	112	110	83	95	101	75	91	72	108	90	65	82
five-mean smoothed *time taken* (s)	–	–	102.6	97.6	100.2	92.8	89.0	86.8	89.4	87.2	85.2	83.4	–	–

Continues →

Tobias has constructed the following graph to display the *time taken*.

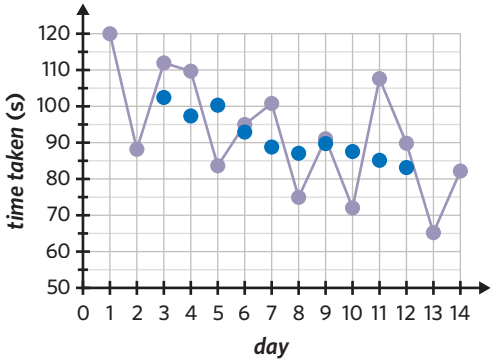


a. Plot the five-mean smoothed values onto the same graph.

Explanation

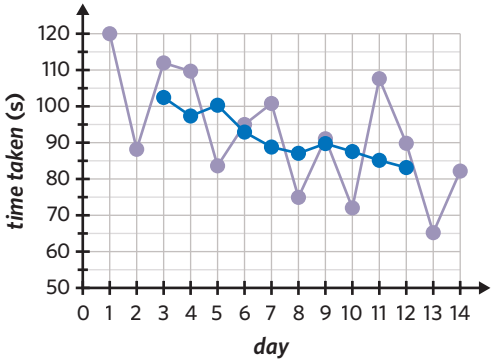
Step 1: Plot each five-mean smoothed data point on the graph.

There are no data values for the first two or last two days.



Step 2: Connect the data points from left to right.

Answer



b. What, if any, trend can be seen from the five-mean smoothed data?

Explanation

From the graph, it can be observed that the five-mean smoothed data values gradually decrease from left to right.

Answer

The five-mean smoothed data shows a decreasing trend in *time taken* (s).

Exam question breakdown

VCAA 2020 Exam 1 Data analysis Q18

The following table shows the *monthly rainfall* for 2019, in millimetres, recorded at a weather station, and the associated long-term seasonal indices for each month of the year.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
***monthly rainfall* (mm)**	18.4	17.6	46.8	23.6	92.6	77.2	80.0	86.8	93.8	55.2	97.3	69.4
seasonal index	0.728	0.734	0.741	0.934	1.222	0.973	1.024	1.121	1.159	1.156	1.138	1.072

The six-mean smoothed *monthly rainfall* with centring for August 2019 is closest to

A. 67.8 mm **B.** 75.9 mm **C.** 81.3 mm **D.** 83.4 mm **E.** 86.4 mm

Explanation

Step 1: Find the *monthly rainfall* for August 2019 as well as the three values on either side and write them in the order they appear in the time series.

92.6 77.2 80.0 86.8 93.8 55.2 97.3

Step 2: Find the mean of the first six values.

$$mean_1 = \frac{92.6 + 77.2 + 80.0 + 86.8 + 93.8 + 55.2}{6}$$
$$= 80.93...$$

Step 3: Find the mean of the last six values.

$$mean_2 = \frac{77.2 + 80.0 + 86.8 + 93.8 + 55.2 + 97.3}{6}$$
$$= 81.71...$$

Step 4: Calculate the centred mean.

$$mean_{centred} = \frac{80.93... + 81.71...}{2} = 81.325$$

Answer

C

73% of students answered this question correctly.

4B Questions

Smoothing over an odd number of data points using moving means

1. Consider the following table.

month	1	2	3	4	5	6
value	75	83	104	97	71	64

The three-mean smoothed value for month 4 is closest to

A. 91 **B.** 92 **C.** 93 **D.** 94

2. Alexia recorded the *time taken*, in minutes, to travel to school over a period of two weeks.

She displayed the data in the following table.

day	Mon 1st Oct	Tue 2nd Oct	Wed 3rd Oct	Thu 4th Oct	Fri 5th Oct	Mon 8th Oct	Tue 9th Oct	Wed 10th Oct	Thu 11th Oct	Fri 12th Oct
***time taken* (mins)**	37	33	44	40	45	31	38	41	37	50

a. Calculate the three-mean smoothed *time taken* for Tuesday 2nd October.

b. Calculate the five-mean smoothed *time taken* for Wednesday 10th October.

3. Each day, Mrs Stinton asks her third grade class if they have eaten fruit. The results for the last ten days are given.

day	1	2	3	4	5	6	7	8	9	10
number of students	14	12	19	22	17	12	19	15	12	21
three-mean smoothed *number of students*	–									–
five-mean smoothed *number of students*	–	–							–	–

a. Fill in the three-mean smoothed values for the entire time series, correct to one decimal place.

b. Fill in the five-mean smoothed values for the entire time series, correct to one decimal place.

---

4. Crime rates in the city of Ashton are on the rise. The figures for the *number of thefts* recorded in the past year are shown in the table.

month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
number of thefts	3	8	7	12	10	17	18	22	26	25	32	35
three-mean smoothed *number of thefts*	–											–
five-mean smoothed *number of thefts*	–	–									–	–

a. Find the three-mean smoothed *number of thefts* for July correct to the nearest whole number.

b. Find the three-mean smoothed values for the entire time series and fill in the table correct to one decimal place.

c. Find the five-mean smoothed values for the entire time series and fill in the table correct to one decimal place.

---

5. Rachel is a leading financial analyst based in the US. Upon analysing exchange rate data for 2003 and 2004, she noticed that for one particular USD exchange rate, the data for the first quarter of 2004 was missing.

year	2003				2004			
quarter	1	2	3	4	1	2	3	4
***exchange rate* (USD)**	0.97	0.99	1.04	1.01		0.89	0.86	0.88
three-mean smoothed *exchange rate* (USD)	–							–
five-mean smoothed *exchange rate* (USD)	–	–					–	–

a. After applying her financial skills, Rachel discovers that the three-mean smoothed *exchange rate* for the first quarter of 2004 is 0.94 USD.

Use this information to help Rachel find the exchange rate for the first quarter of 2004 correct to two decimal places.

b. Find the seven-mean smoothed *exchange rate* for the fourth quarter of 2003, correct to two decimal places.

c. Find the three-mean smoothed *exchange rate* for the entire time series, correct to two decimal places.

d. Find the five-mean smoothed *exchange rate* for the entire time series, correct to two decimal places.

Smoothing over an even number of data points using moving means

6. Consider the following table.

year	2016	2017	2018	2019	2020	2021	2022
sales	203	226	241	188	203	261	250

The two-mean smoothed *sales* for 2020 is closest to

A. 214 **B.** 217 **C.** 224 **D.** 232

7. The following table shows the *number of points* scored by the winning team of the AFL grand final between 2010 and 2021.

year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
number of points	108	119	91	77	137	107	89	108	79	114	81	140

a. Calculate the two-mean smoothed *number of points* for 2020.

b. Calculate the four-mean smoothed *number of points* for 2013.

c. Calculate the six-mean smoothed *number of points* for 2015.

8. The following table shows the number of people who visited a popular tourist attraction over the course of a year.

month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
***number of tourists* (000's)**	1.9	2.1	1.8	1.7	1.4	1.1	1.2	1.4	1.7	1.6	1.8	2.0
two-mean smoothed *number of tourists* (000's)	–											–
four-mean smoothed *number of tourists* (000's)	–	–									–	–

a. Find the two-mean smoothed *number of tourists* (000's) centred at July, correct to one decimal place.

b. Find the four-mean smoothed *number of tourists* (000's) centred at July, correct to one decimal place.

c. Find the two-mean smoothed and centred *number of tourists* (000's), correct to one decimal place, for the entire time series and fill in the table.

d. Find the four-mean smoothed and centred *number of tourists* (000's), correct to one decimal place, for the entire time series and fill in the table.

9. Brie recently opened a fromagerie called 'Really Grate Cheese'. She recorded the total *weight*, in kg, of cheese sold each week since the opening of the store.

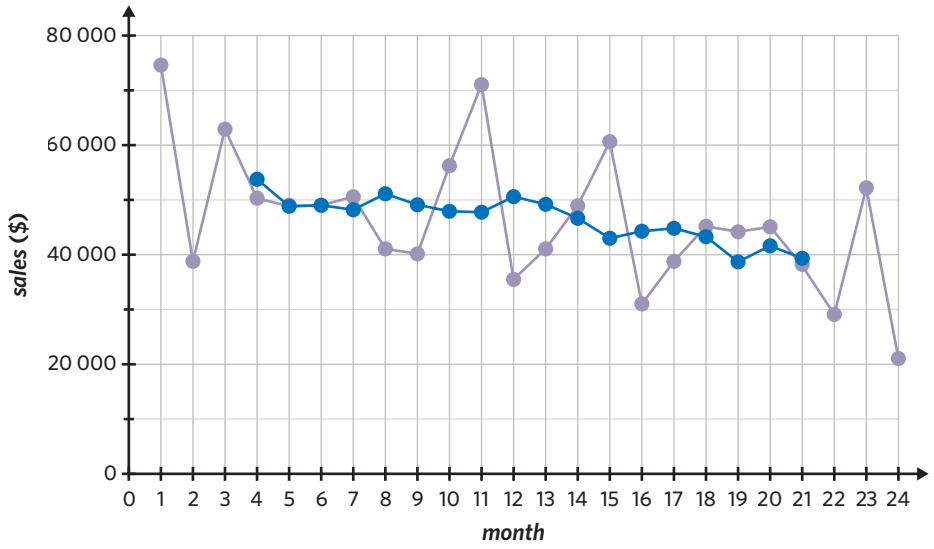
week	1	2	3	4	5	6	7	8	9	10	11
***weight* (kg)**	18.2	21.5	23.3	22.0	25.4	24.9	26.6		25.8	22.2	24.7
four-mean smoothed *weight* (kg)											

a. Brie accidentally lost her sales records from week 8. Brie knows that the four-mean smoothed *weight* of cheese sold, centred on week 8, is 24.7 kg.
Using this information, find the exact actual *weight* of cheese sold in week 8.

b. Find the four-mean smoothed and centred *weight* for the entire time series, correct to one decimal place.

Plotting and interpreting a mean smoothed time series

10. The following graph shows a retail shop's *sales* and their seven-mean smoothed *sales* over 24 months.



What trend is visible in the smoothed values?

A. Increasing

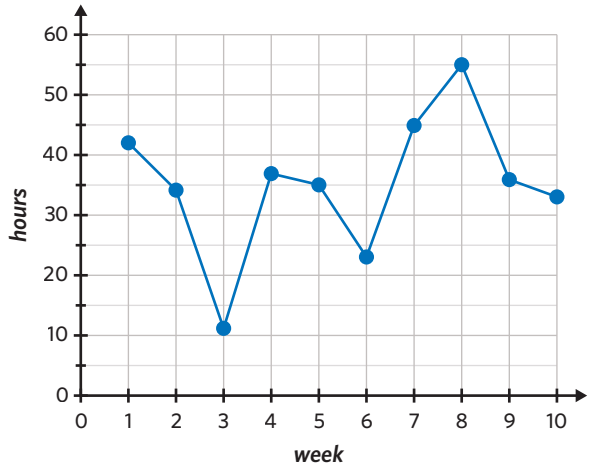
B. Decreasing

C. No trend

D. Structural change

---

11. Winston is a freelance photographer and recorded the number of *hours* he worked each week for 10 weeks, and plotted the data onto the following graph.



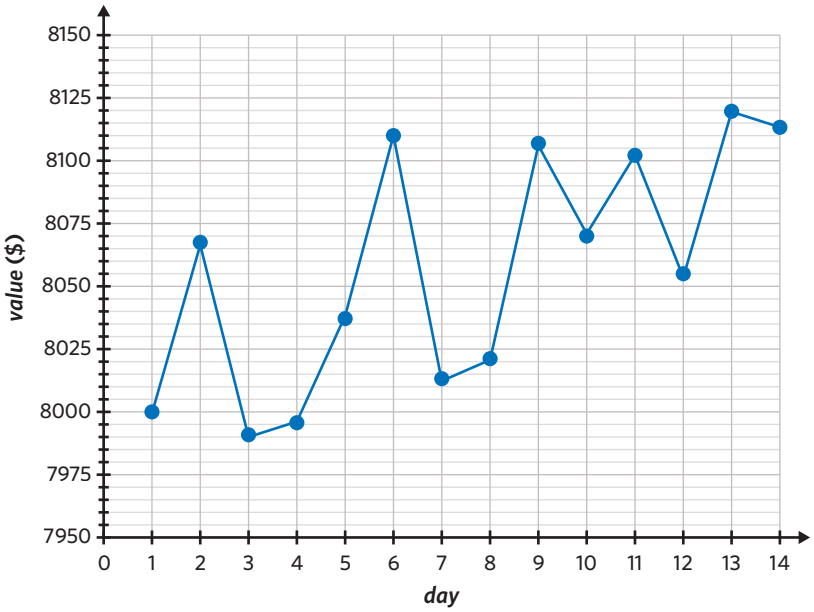
He also calculated the two-mean and five-mean smoothed values for the data, correct to one decimal place.

week	1	2	3	4	5	6	7	8	9	10
hours	42	34	11	37	35	23	45	55	36	33
two-mean smoothed *hours*	–	30.3	23.3	30.0	32.5	31.5	42.0	47.8	40.0	–
five-mean smoothed *hours*	–	–	31.8	28.0	30.2	39.0	38.8	38.4	–	–

a. In one colour, plot the two-mean smoothed values onto the graph.

b. In another colour, plot the five-mean smoothed values onto the same graph.

12. Two weeks ago, Amira invested $8000 in various cryptocurrencies and has since monitored its *value*, in dollars, daily.



Amira also calculated the four-mean smoothed *value* of her crypto portfolio with centring, as shown in the following table. Values are given correct to the nearest whole number.

day	1	2	3	4	5	6	7	8	9	10	11	12	13	14
***value* ($)**	8000	8067	7991	7996	8037	8110	8013	8021	8106	8070	8102	8055	8119	8113
four-mean smoothed *value* ($)	–	–	8018	8028	8036	8042	8054	8058	8064	8079	8085	8092	–	–

a. Plot the smoothed values onto the graph.

b. What trend is visible in the smoothed values? Interpret this in terms of the context given.

Joining it all together

13. An ice cream shop recorded the *sales* for their brand new bubblegum flavoured ice cream over the course of a year. Their findings are shown in the following table.

month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
***sales* ($)**	450	510	370	360	290	100	110	90	150	230	310	420

a. Find the three-mean smoothed *sales* figure for May.

b. Find the five-mean smoothed *sales* figure for August.

c. Find the two-mean smoothed *sales* figure with centring for September.

d. Find the four-mean smoothed *sales* figure with centring for March.

e. Find the seven-mean smoothed *sales* figure for July.

f. Find the six-mean smoothed *sales* figure with centring for June, correct to one decimal place.

14. Alfredo is a passionate yogi and recorded the number of *yoga classes* that he has attended each year since 2010.

year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
yoga classes	52	56	48	61	66	59	69	73	65	81	53	88	79

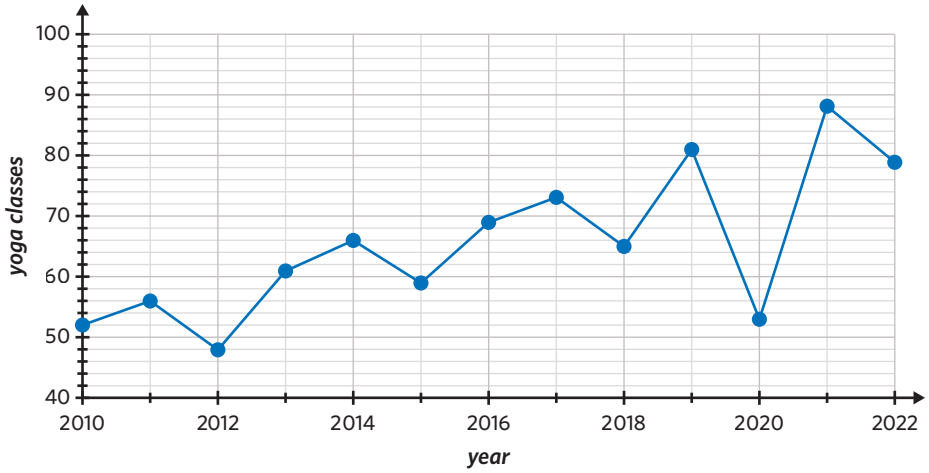
a. Alfredo considers using nine-mean smoothing to smooth the data. If he were to do this, which year would be the first to have a smoothed value?

b. Calculate the nine-mean smoothed value for the year specified in part **a**.

c. Alfredo decides that it would be more useful to use two-mean smoothing so he retains more data points. Complete the following table with the two-mean smoothed and centred values, to the nearest whole number.

year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
yoga classes	52	56	48	61	66	59	69	73	65	81	53	88	79
two-mean smoothed *yoga classes*													

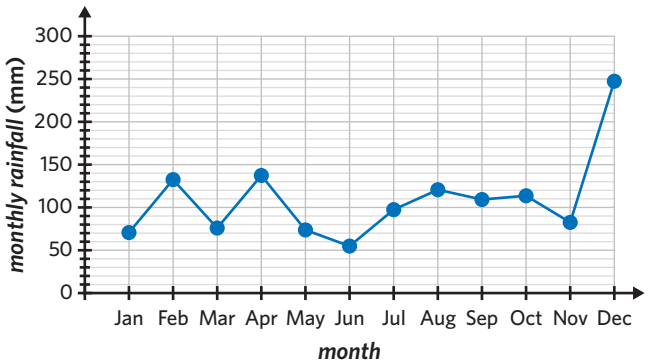
d. Alfredo has already plotted the original data onto the following time series. Plot the two-mean smoothed data onto the same graph.



e. What trend, if any, is visible in the smoothed data? Interpret this with reference to the given variables.

Exam practice

15. The following time series plot shows the *monthly rainfall* at a weather station, in millimetres, for each month in 2017.



If seven-mean smoothing is used to smooth this time series plot, the number of smoothed data points would be

A. 3

B. 5

C. 6

D. 8

E. 10

VCAA 2019 Exam 1 Data analysis Q16

74% of students answered this question correctly.

16. The wind speed at a city location is measured throughout the day over a three-week period. The following table shows the daily *maximum wind speed*, in kilometres per hour, for the days in week 2.

day	8	9	10	11	12	13	14
***maximum wind speed* (km/h)**	22	22	19	22	43	37	33

A four-point moving mean with centring is used to smooth the time series data. The smoothed *maximum wind speed*, in kilometres per hour, for day 11 is closest to

A. 22
B. 24
C. 26
D. 28
E. 30

VCAA 2017 Exam 1 Data analysis Q15

64% of students answered this question correctly.

---

17. A garden centre sells garden soil. The following table shows the daily *quantity* of garden soil sold, in cubic metres, over a one-week period.

day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
***quantity* (m^3)**	234	186				346	346

The *quantity* of garden soil sold on Wednesday, Thursday and Friday is not shown. The five-mean smoothed *quantity* of garden soil sold on Thursday is 206 m^3. The three-mean smoothed *quantity* of garden soil sold on Thursday, in cubic metres, is

A. 143
B. 166
C. 206
D. 239
E. 403

VCAA 2021 Exam 1 Data analysis Q14

48% of students answered this question correctly.

---

18. The maximum daily rainfall each month was recorded at a weather station. The following table shows the *maximum daily rainfall* each month for a period of one year.

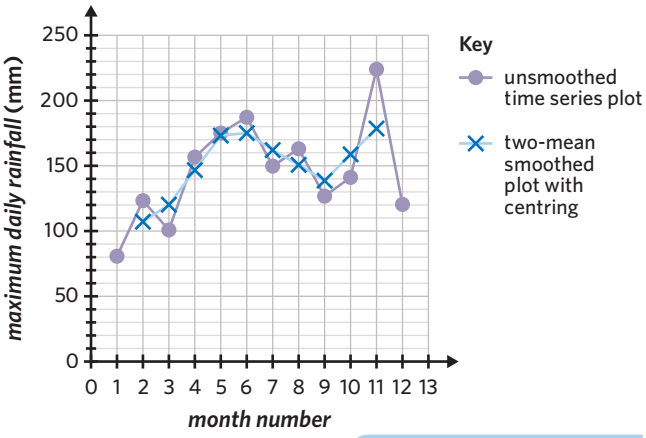
month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
month number	1	2	3	4	5	6	7	8	9	10	11	12
***maximum daily rainfall* (mm)**	79	123	100	156	174	186	149	162	124	140	225	119

The data in the table has been used to plot *maximum daily rainfall* against *month number* in the following time series plot.

Two-mean smoothing with centring has been used to smooth the time series plot. The smoothed values are marked with crosses (×).

Using the data given in the table, show that the two-mean smoothed rainfall centred on October is 157.25 mm. (2 MARKS)

VCAA 2016 Exam 2 Data analysis Q4b



The average mark on this question was **1**.

Questions from multiple lessons

Data analysis

19. The following histogram shows the distribution of the *number* of skyscrapers per 1000 m^2 for 61 different cities plotted on a $\log_{10}$ scale.

How many cities have at least one skyscraper per 1000 m^2?

A. 6

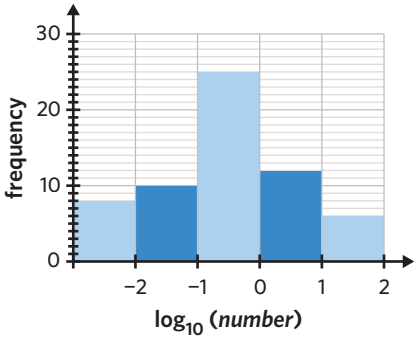
B. 12

C. 18

D. 25

E. 43

Adapted from VCAA 2016 Exam 1 Data analysis Q7



Data analysis

20. The following scatterplot shows the number of bubble teas sold versus the temperature in degrees celsius on 12 days. A least squares regression line has been fitted to the scatterplot.

The least squares line is used to predict the number of bubble teas sold on a 15 degree day. The residual is closest to

A. −38

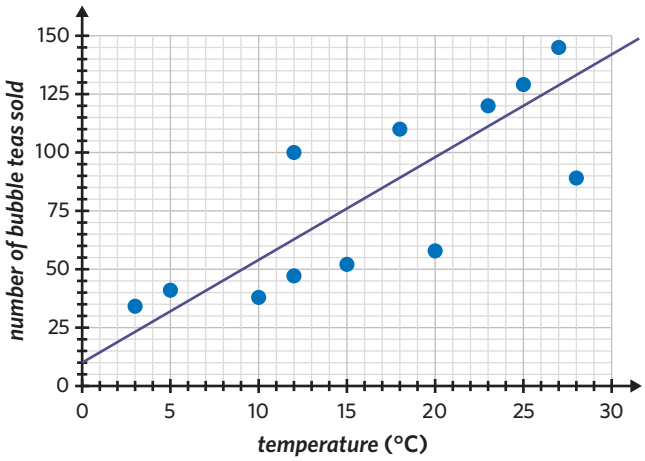
B. −23

C. −12

D. 23

E. 38

Adapted from VCAA 2017 Exam 1 Data analysis Q9

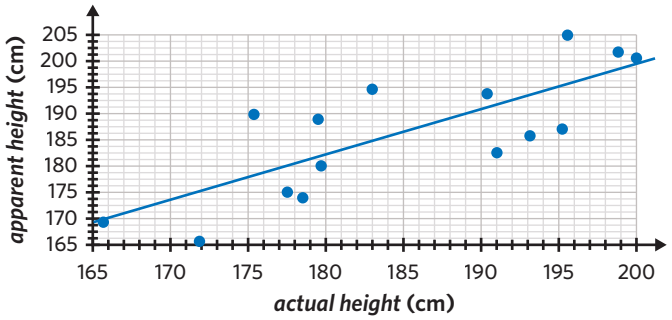


Data analysis

21. The data in the following table shows a sample of *apparent height* and *actual height* recorded from a group of Year 12 students, where *apparent height* is the height that the students claim to be.

A scatterplot of the data is also shown.

This data will be used to investigate the association between *apparent height* and *actual height*.



$n = 15$

$r^2 = 0.64$

apparent height (cm)	*actual height* (cm)
200.9	199.7
182.7	190.9
194.2	183.0
193.2	190.3
202.1	198.8
166.8	171.9
189.1	175.5
204.9	195.6
187.6	195.2
179.9	179.5
173.0	178.6
168.4	166.5
185.7	193.2
174.6	177.8
188.0	179.6

a. Use the scatterplot to describe the association between *apparent height* and *actual height* in terms of strength, form and direction. (1 MARK)

b. A least squares regression line can be used to predict the *apparent height* from the *actual height* in the form *apparent height* $= a + b \times$ *actual height*. Determine the values of a and b. Round your answers to two decimal places. (2 MARKS)

Adapted from VCAA 2016 Exam 2 Data analysis Q3a-bi

4C Smoothing – moving medians

STUDY DESIGN DOT POINT

- graphical smoothing of time series plots using moving medians (involving an odd number of points only) to help identify long-term trends in time series with large fluctuations

4A 4B **4C** 4D 4E

KEY SKILLS

During this lesson, you will be:

- smoothing using three-moving medians
- smoothing using five (or more)-moving medians.

KEY TERMS

- Moving median smoothing

Similar to moving mean smoothing, moving median smoothing is used to remove large fluctuations in time series data in order to identify long-term trends. However, unlike moving mean smoothing, moving median smoothing can be done directly onto the graph. Additionally, moving median smoothing is not impacted by outlier values. In General Mathematics, only moving median smoothing for an odd number of points is explored.

Smoothing using three-moving medians

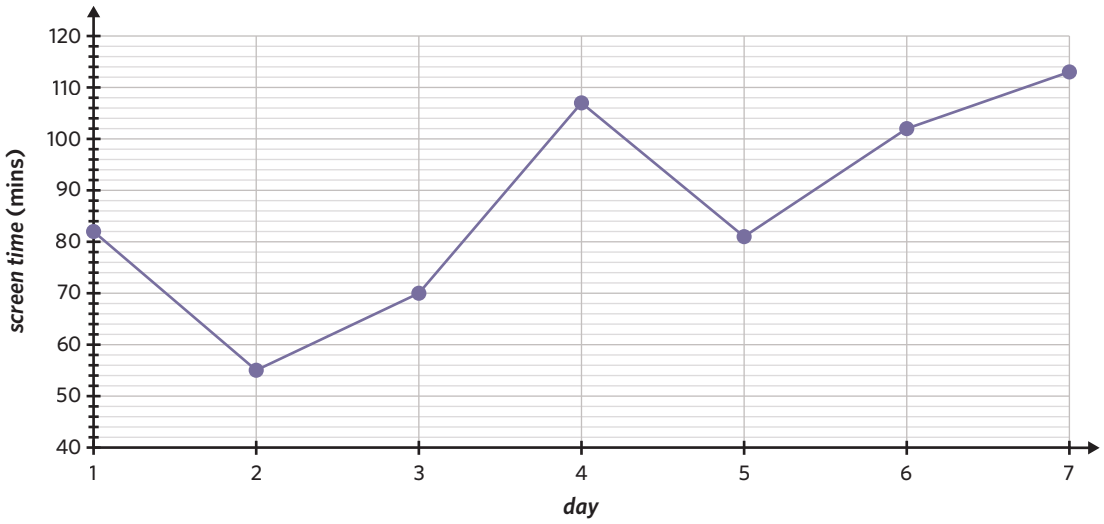
Smoothing using medians is known as **moving median smoothing**. Three-median smoothing involves replacing each data value with the median of itself and each adjacent value.

For example, the three-median smoothed value for day 2 is the median of the values for days 1, 2, and 3.

There are no smoothed values for the first and last data points because they do not have a value on both sides.

Worked example 1

Bailey recorded their *screen time*, in minutes, on their phone over the past week. Their data is presented in the following time series plot.



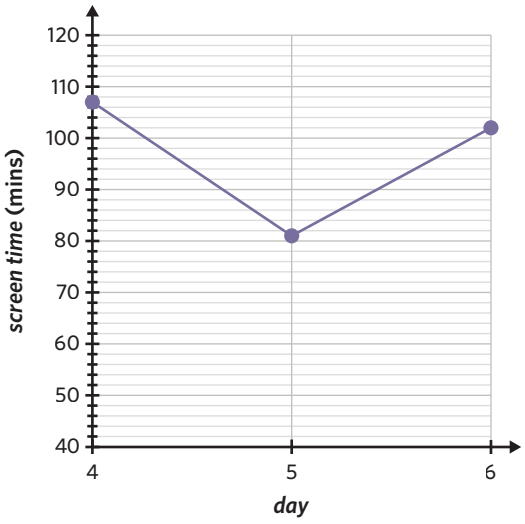
Continues →

a. Determine the three-median smoothed *screen time* for day 5.

Explanation

Step 1: Identify the necessary data points.

The three-median smoothed value for day 5 is the median of day 5 and the values on each side of it (days 4 and 6).

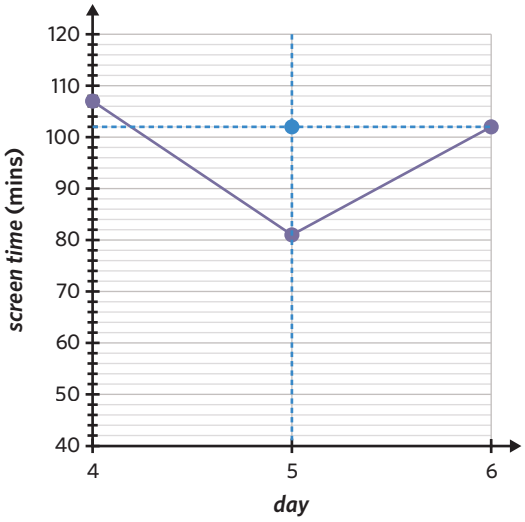


Step 2: Identify the median *screen time* from days 4 to 6.

The variable *screen time* is represented by the vertical height of the points.

As there are three values, the median is the second value from smallest to largest.

The median *screen time* is 102 minutes.



Answer

102 minutes

b. Construct a three-median smoothed plot for the entire time series.

Explanation

Step 1: Identify the necessary data points for the first smoothed value.

The first smoothed value will be for day 2 since it has a value on each side.

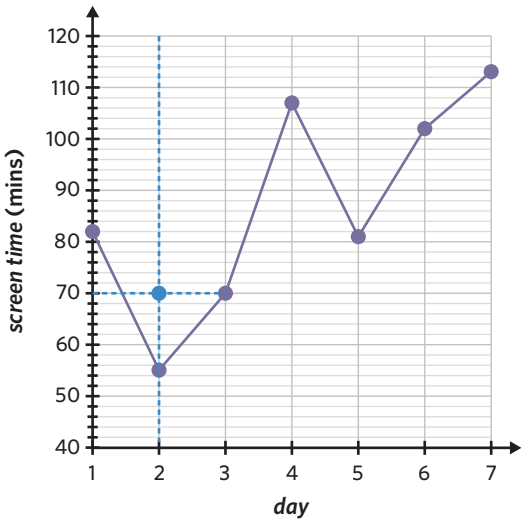
The three-median smoothed value for day 2 is the median of days 1 to 3.

Step 2: Identify the median *screen time* for days 1 to 3.

The *screen time* values for days 1, 2 and 3 are 82, 55, and 70 minutes.

The median *screen time* is 70 minutes.

Step 3: Mark the smoothed value for day 2 on the graph.



Continues →

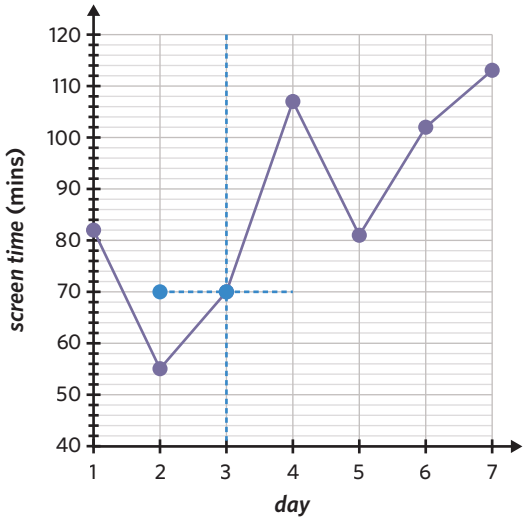
Step 4: Repeat steps 1 to 3 for day 3.

The smoothed value for day 3 is the median *screen time* for days 2 to 4.

The *screen time* values for days 2 to 4 are 55, 70, and 107 minutes.

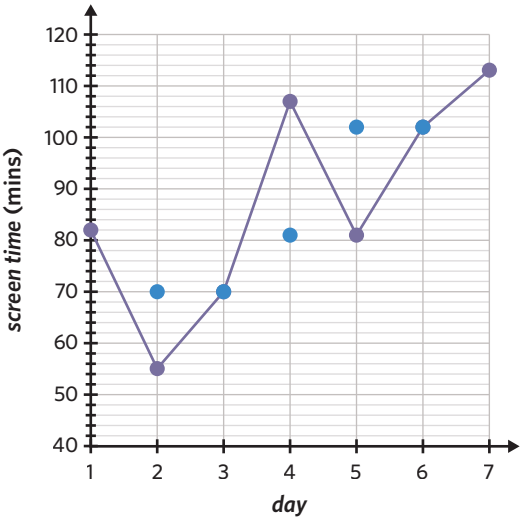
The median *screen time* is 70 minutes.

Make sure the original data points are used and not the previously found smoothed points when finding the three-moving median.



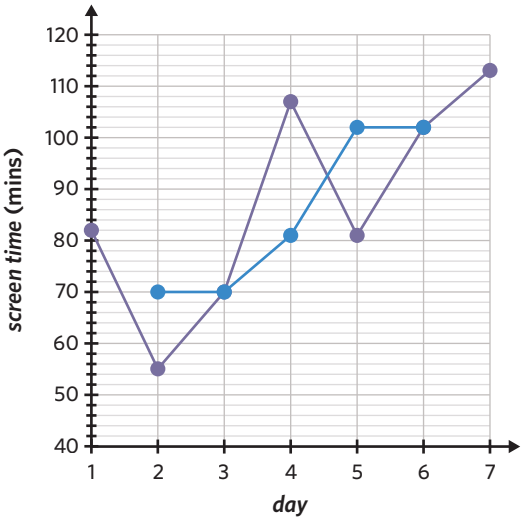
Step 5: Apply this technique for the rest of the time series.

The smoothing will stop at day 6, as this is the last day with a data value on each side.



Step 6: Connect the smoothed points with straight lines.

Answer



c. What, if any, trend can be seen from the three-median smoothed time series?

Explanation

From the graph, it can be observed that the three-median smoothed data values increase from left to right.

Answer

The three-median smoothed data shows an increasing trend in Bailey's *screen time* over the week.

Smoothing using five (or more)-moving medians

Five-median smoothing is similar to three-median smoothing but uses five values. Each data value is replaced with the median of itself and the two values on each side of it.

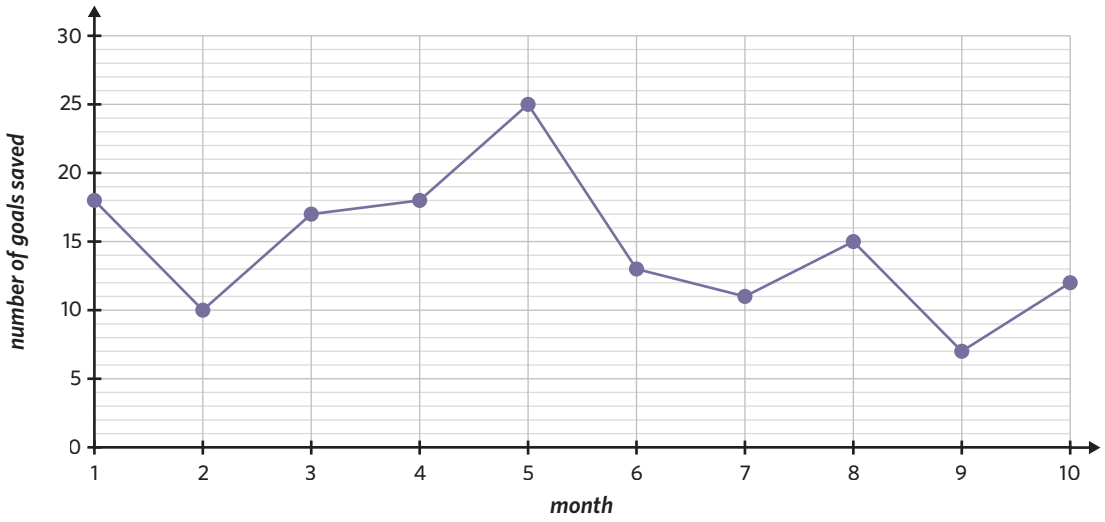
For example, the five-median smoothed value for day 3 is the median of the values for days 1, 2, 3, 4, and 5.

There are no smoothed values for the first two and last two data entries because they do not have two values on both sides.

Moving median smoothing can be done over any odd number of values, in the same way as three and five-median smoothing.

Worked example 2

Emanuel is the goalkeeper for his soccer team and has recorded the *number of goals saved* each month for the past ten months. His data has been presented in the following time series plot.

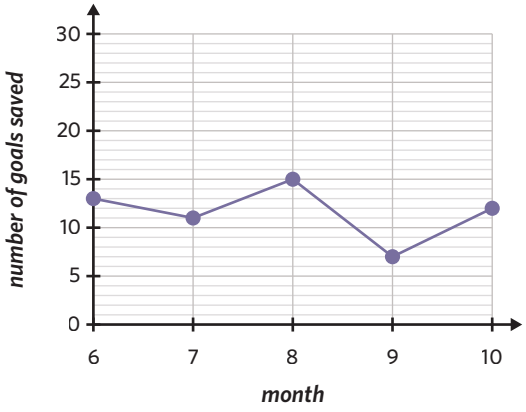


a. Determine the five-median smoothed *number of goals saved* for month 8.

Explanation

Step 1: Identify the necessary data points.

The five-median smoothed value for month 8 is the median of month 8 and the two values on each side of it (months 6, 7, 9 and 10).

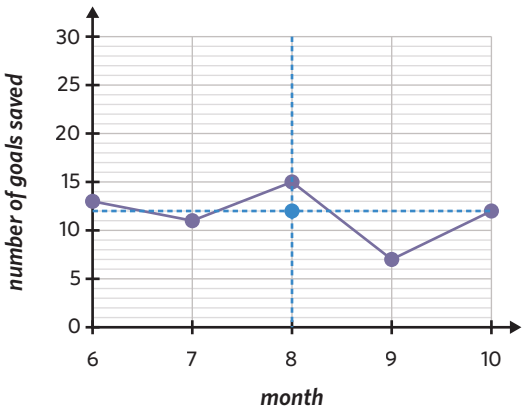


Step 2: Identify the median *number of goals saved* for months 6 to 10.

The variable *number of goals saved* is represented by the vertical height of the points.

As there are five values, the median is the third value from smallest to largest.

The median *number of goals saved* is 12.



Continues →

Answer

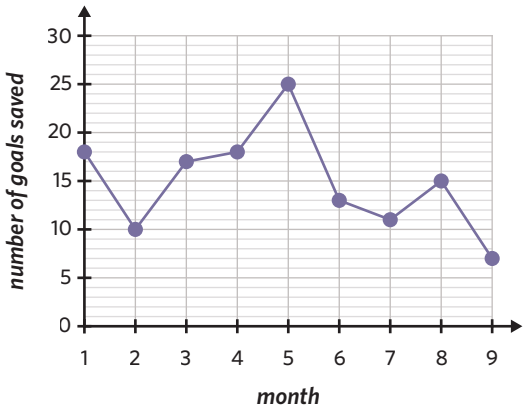
12 goals saved

b. Determine the nine-median smoothed *number of goals saved* for month 5.

Explanation

Step 1: Identify the necessary data points.

The nine-median smoothed value for month 5 is the median of month 5 and the four values on each side of it (months 1 to 4 and 6 to 9).

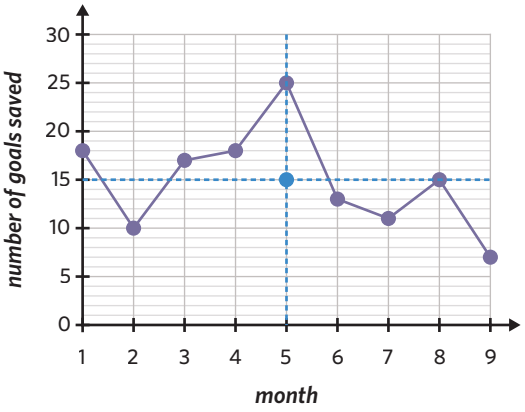


Step 2: Identify the median *number of goals saved* for months 1 to 9.

The variable *number of goals saved* is represented by the vertical height of the points.

As there are nine values, the median is the fifth value from smallest to largest.

The median *number of goals saved* is 15.



Answer

15 goals saved

c. Construct a five-median smoothed plot for the entire time series.

Explanation

Step 1: Identify the necessary data points for the first smoothed value.

The first smoothed value will be for month 3 since it has two values on each side.

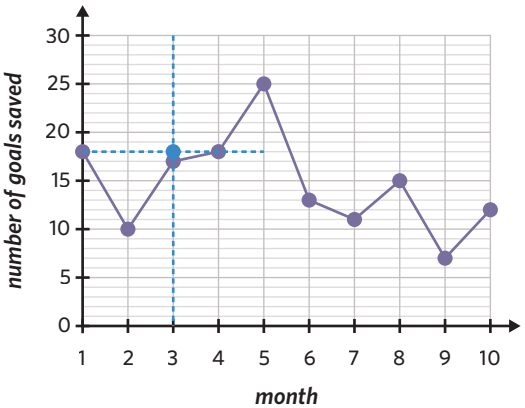
The five-median smoothed value for month 3 is the median of months 1 to 5.

Step 2: Identify the median *number of goals saved* for months 1 to 5.

The *number of goals saved* for months 1 to 5 are 18, 10, 17, 18, and 25.

The median *number of goals saved* is 18.

Step 3: Mark the smoothed value for month 3 on the graph.



Continues →

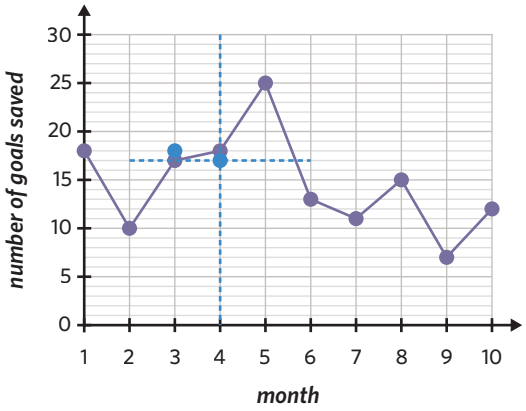
Step 4: Repeat steps 1 to 3 for month 4.

The smoothed value for month 4 is the median *number of goals saved* from months 2 to 6.

The *number of goals saved* for months 2 to 6 are 10, 17, 18, 25 and 13.

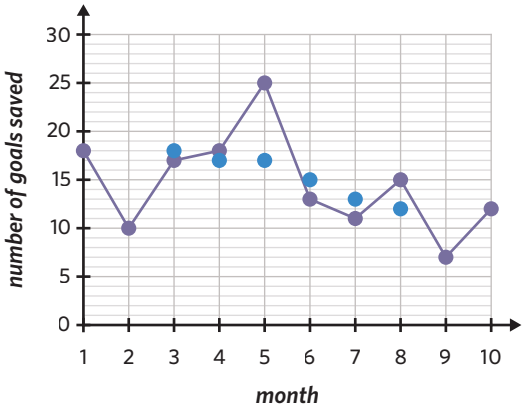
The median *number of goals saved* is 17.

Make sure the original data points are used and not the previously found smoothed points when finding the five-moving median.



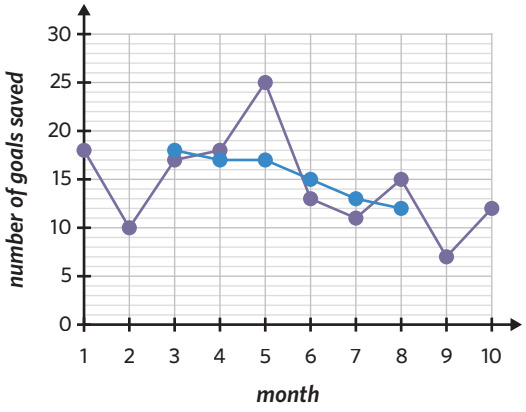
Step 5: Apply this technique for the rest of the time series.

The smoothing will stop at month 8, as this is the last month with two values on each side.



Step 6: Connect the smoothed points with straight lines.

Answer



d. What, if any, trend can be seen from the five-median smoothed time series?

Explanation

From the graph, it can be observed that the five-median smoothed data values gradually decrease from left to right.

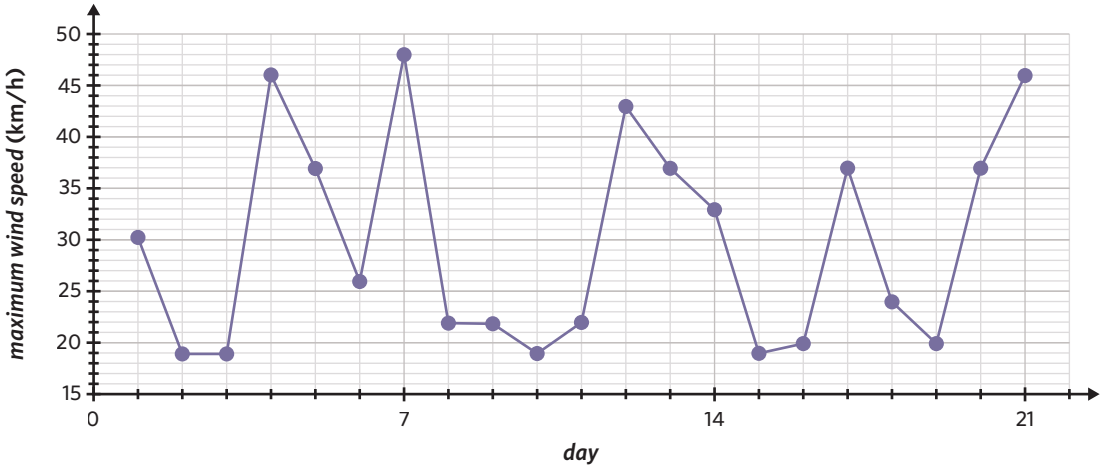
Answer

The five-median smoothed data shows a decreasing trend in *number of goals saved*.

Exam question breakdown

VCAA 2017 Exam 1 Data analysis Q14

The wind speed at a city location is measured throughout the day. The time series plot shows the daily *maximum wind speed*, in kilometres per hour, over a three-week period.



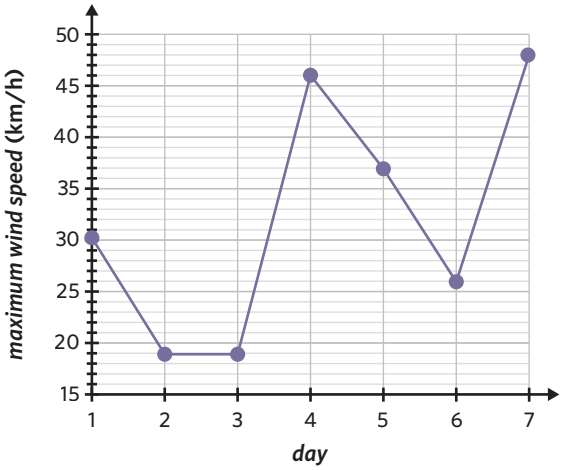
The seven-median smoothed *maximum wind speed*, in kilometres per hour, for day 4 is closest to

A. 22
B. 26
C. 27
D. 30
E. 32

Explanation

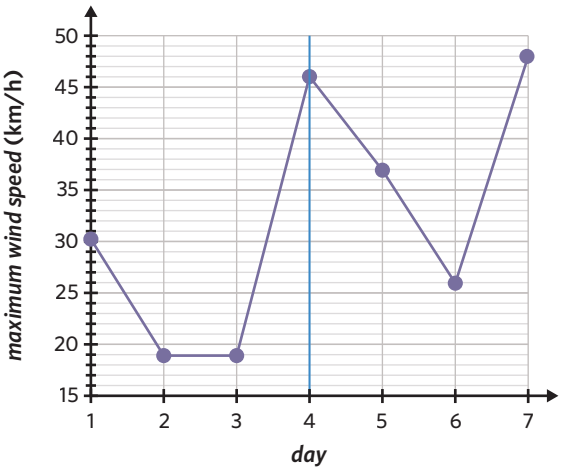
Step 1: Identify the necessary data points.

The seven-median smoothed value for day 4 is the median of day 4 and the three values on each side of it (days 1 to 3 and 5 to 7).



Step 2: Draw a vertical line at day 4.

The smoothed value for day 4 will fall along this line.



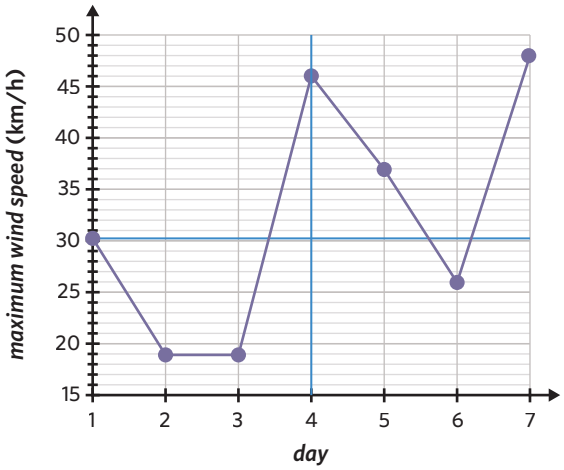
Continues →

Step 3: Identify the median value for *maximum wind speed* and draw a horizontal line.

The variable *maximum wind speed* is represented by the vertical height of the points.

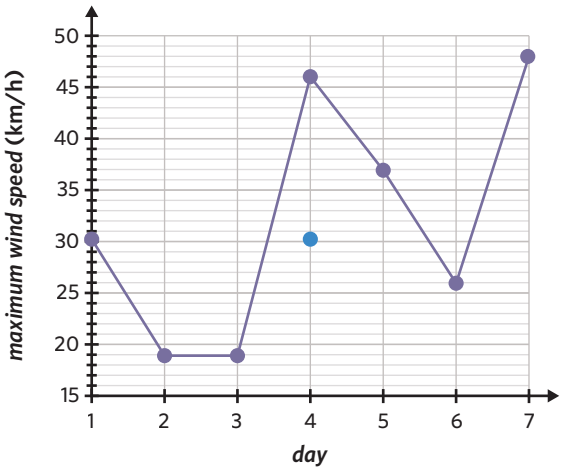
As there are seven values, the median is the fourth value from smallest to largest.

The median *maximum wind speed* is just over 30 km/h on day 1. Draw a horizontal line from this point. There should be an equal number of points above and below the line.



Step 4: Mark the smoothed value.

The seven-median smoothed value for day 4 is at the intersection of these two lines.



The seven-median smoothed *maximum wind speed* for day 4 is closest to 30 km/h.

Answer

D

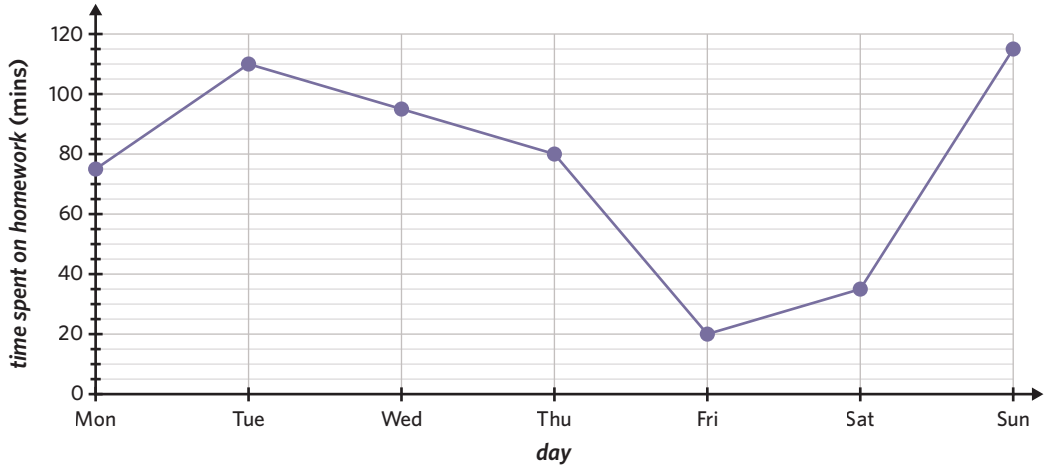
53% of students answered this question correctly.

26% of students incorrectly chose option E. This is likely because they calculated the seven-mean smoothed value, rather than the seven-median smoothed value. The seven-mean smoothed value for day 4 is closest to 32 km/h.

4C Questions

Smoothing using three-moving medians

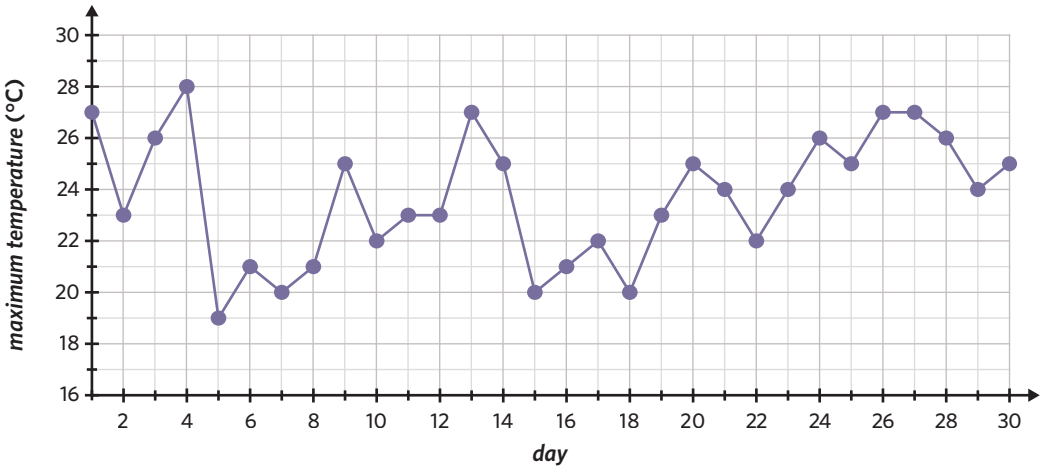
1. Consider the following time series.



The three-median smoothed *time spent on homework* for Friday is

A. 20 minutes **B.** 35 minutes **C.** 45 minutes **D.** 80 minutes

2. The daily *maximum temperature*, in °C, for November is shown.



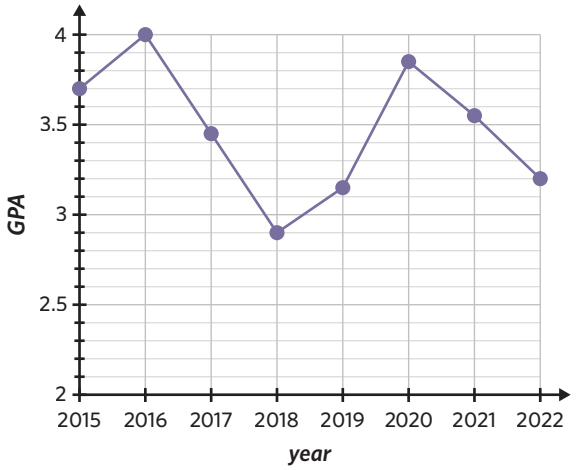
Determine the three-median smoothed *maximum temperature* for the

a. 2nd of November. **b.** 7th of November. **c.** 13th of November. **d.** 26th of November.

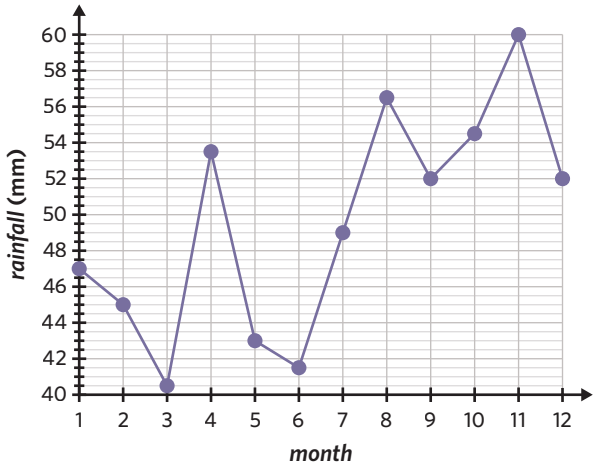
---

3. Smooth each of the following time series using three-median smoothing.

a.



b.



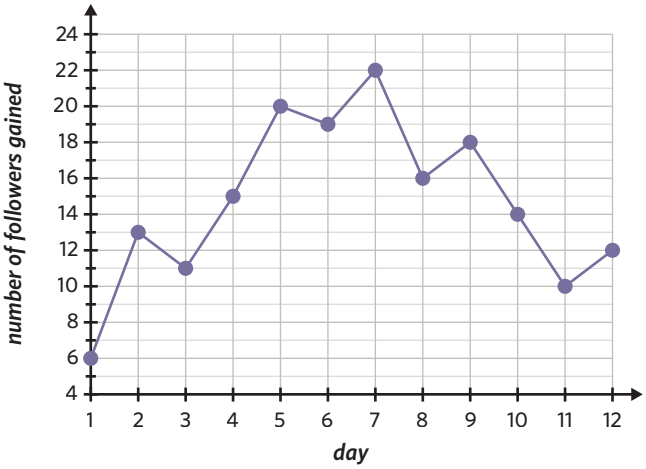
---

4. The time series plot shows the number of followers Annabel gained on Instagram each day over a 12-day period.

a. Determine the three-median smoothed *number of followers gained* for day 9.

b. Smooth the time-series graph using three-median smoothing.

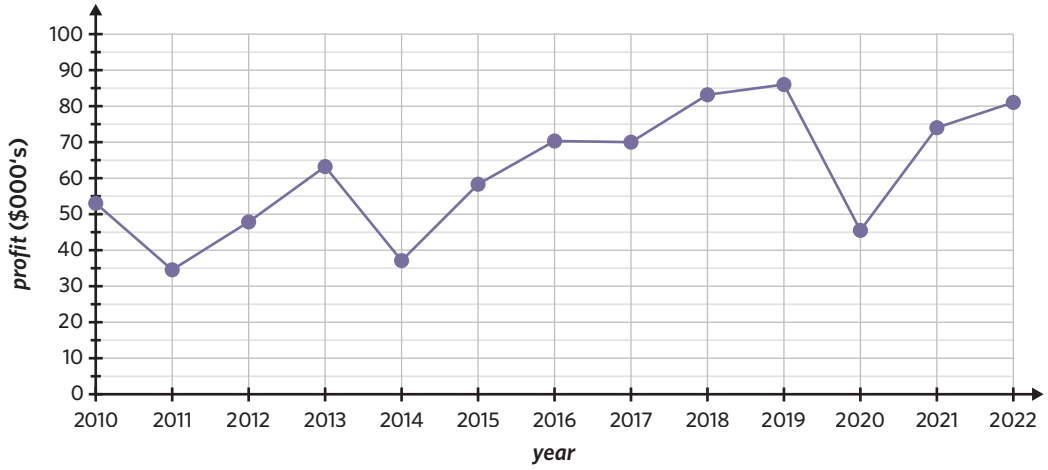
c. Describe any trends in the smoothed data.



Smoothing using five (or more)-moving medians

5. Consider the following time series.

The five-median smoothed *profit* for 2019 is closest to



A. $70 000 **B.** $74 000 **C.** $83 000 **D.** $86 000

---

6. Last year, Nikki decided to plant a large crop of potatoes in her veggie garden. This year, Nikki recorded the number of potatoes she harvested each month, as shown.

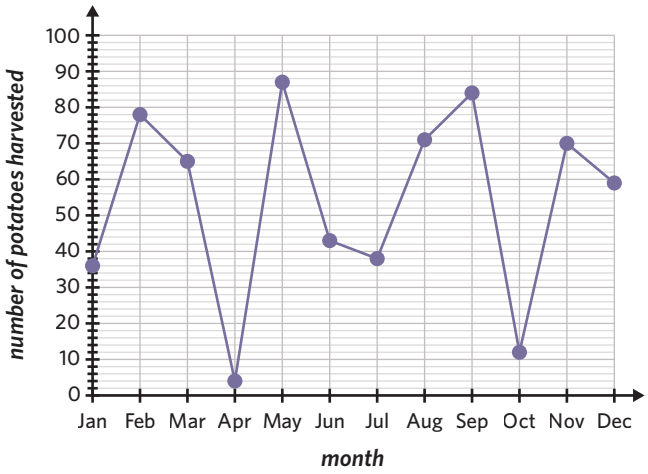
Determine

a. the five-median smoothed *number of potatoes harvested* for March.

b. the seven-median smoothed *number of potatoes harvested* for July.

c. the nine-median smoothed *number of potatoes harvested* for August.

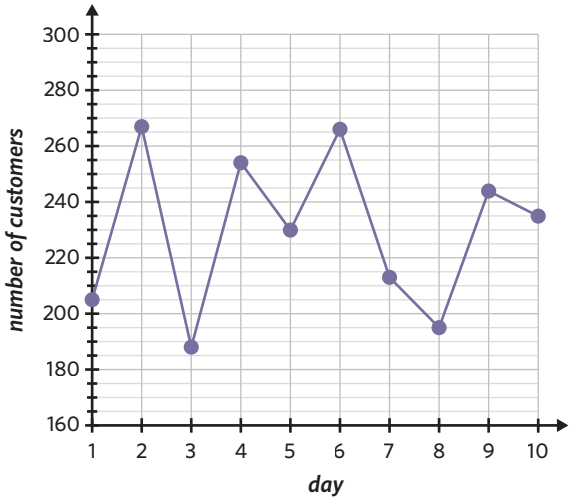
d. the eleven-median smoothed *number of potatoes harvested* for June.



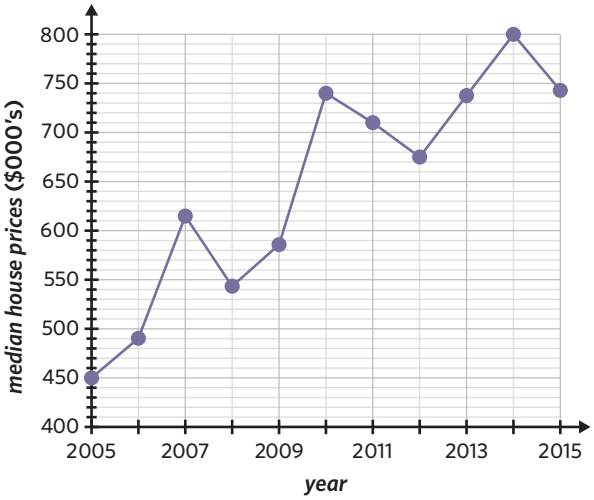
---

7. Smooth each of the following time series using five-median smoothing.

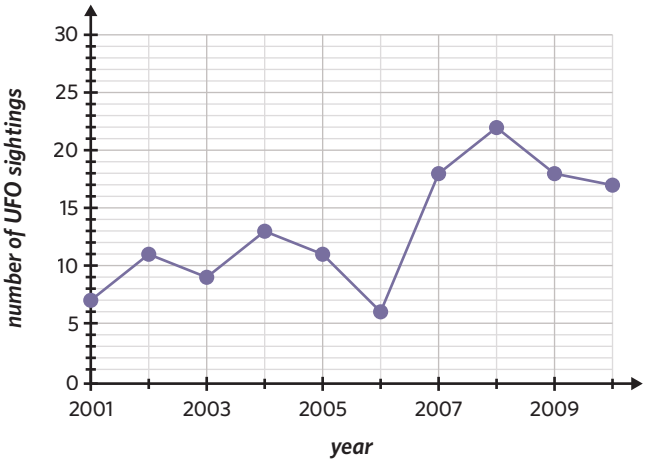
a.



b.



8. The *number of UFO sightings* in a small town was recorded over a ten-year period. The results are shown in the time series plot.



a. Determine the nine-median smoothed *number of UFO sightings* for 2006.

b. Smooth the entire time series using five-median smoothing.

c. What, if any, trend in the *number of UFO sightings* can be seen from the five-median smoothed time series?

Joining it all together

9. Jae is a small business owner. She recorded her *daily income* over the 28 days of February. Jae then smoothed the data using moving median smoothing in order to see any underlying trends. The smoothed data consisted of 22 smoothed data values.

The smoothing Jae has used is

A. three-moving median smoothing.

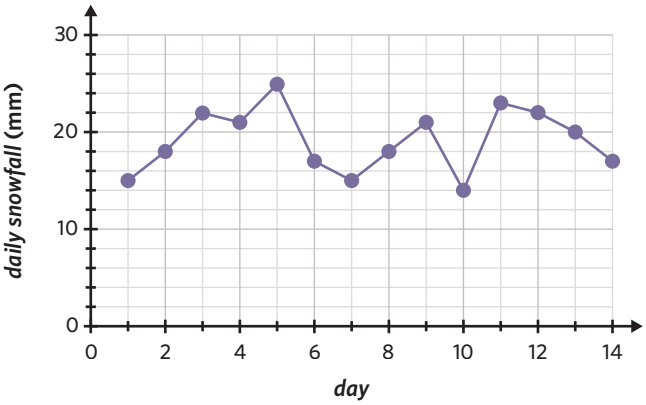
B. five-moving median smoothing.

C. seven-moving median smoothing.

D. nine-moving median smoothing.

---

10. The *daily snowfall* (mm) on Mount Buller was recorded over a two-week period. The results are shown in the following time series plot.



a. Using three-median smoothing, find the smoothed *daily snowfall* for day 4.

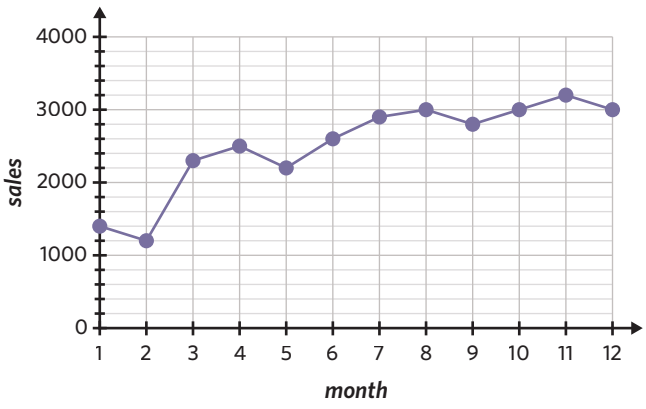
b. Using five-median smoothing, find the smoothed *daily snowfall* for day 7.

c. Using seven-median smoothing, find the smoothed *daily snowfall* for day 11.

d. Using nine-median smoothing, find the smoothed *daily snowfall* for day 8.

---

11. The *sales* figures for a new type of tennis racquet were recorded over a 12-month period. The results are shown in a time series plot.

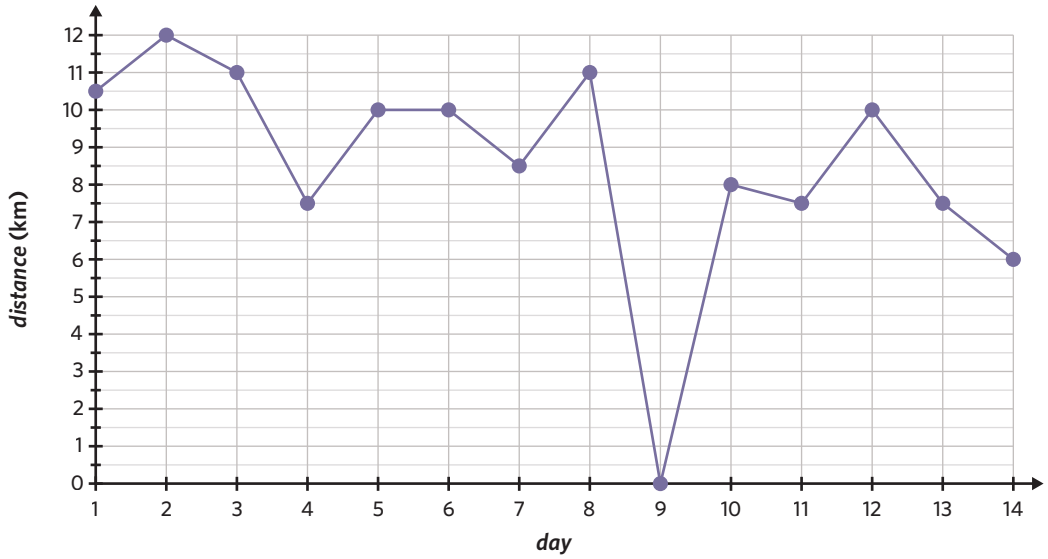


a. Using three-median smoothing, find the smoothed *sales* for month 2.

b. Smooth the time series using five-median smoothing.

c. Describe the general pattern in *sales* that is displayed by the five-median smoothed plot.

12. Amelia just finished a 14-day hike in Nepal and measured the *distance*, in km, she hiked each day. She displayed her data in the following time series.



a. Amelia had altitude sickness on day 9 so didn't hike at all. To smooth over this outlier, determine the five-median smoothed *distance* for day 9.

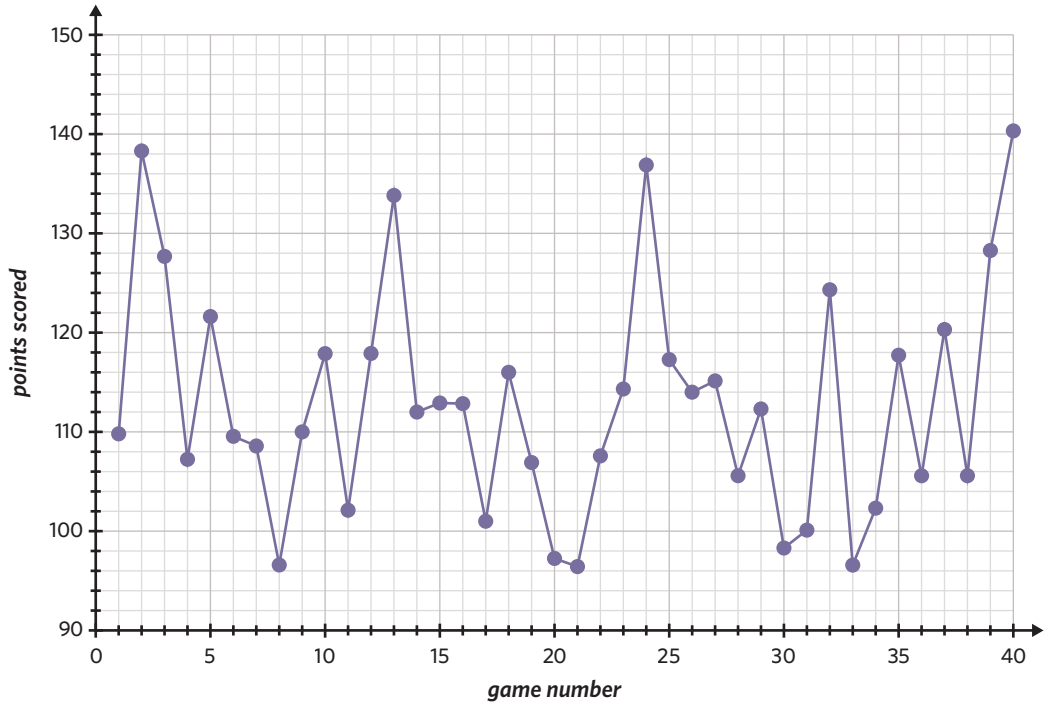
b. Smooth the entire time series using three-median smoothing.

c. Smooth the entire time series using seven-median smoothing.

d. What, if any, trend can be seen from the smoothed time series?

Exam practice

13. The time series plot shows the *points scored* by a basketball team over 40 games.



The nine-median smoothed *points scored* for game number 10 is closest to

A. 102 **B.** 108 **C.** 110
D. 112 **E.** 117

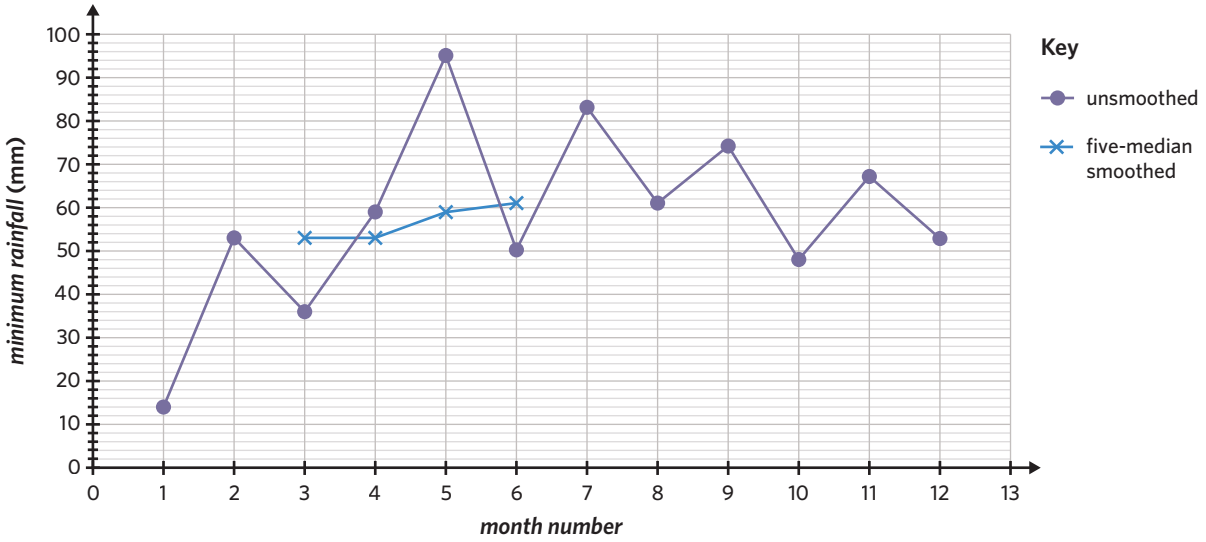
VCAA 2021 Exam 1 Data analysis Q13

53% of students answered this question correctly.

14. The time series plot shows the *minimum rainfall* recorded at the weather station each month plotted against the *month number* (1 = January, 2 = February, and so on).

Rainfall is recorded in millimetres.

The data was collected over a period of one year.



Five-median smoothing has been used to smooth the time series plot.

The first four smoothed points are shown as crosses (×).

Complete the five-median smoothing by marking smoothed values with crosses (×) on the time series plot. (2 MARKS)

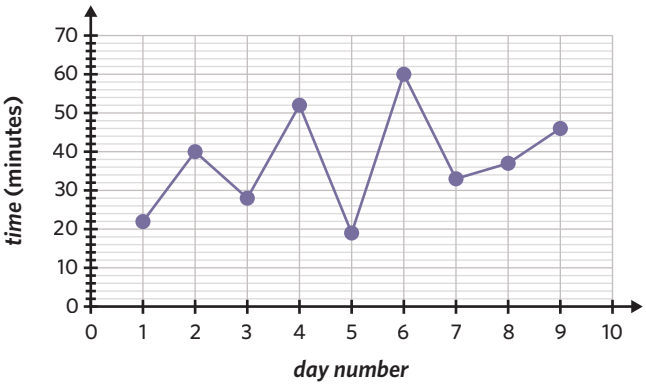
VCAA 2016 Exam 2 Data analysis Q4a

The average mark on this question was **1**.

---

15. The *time*, in minutes, that Liv ran each day was recorded for nine days.

The following time series plot was generated from this data.



Both three-median smoothing and five-median smoothing are being considered for this data. Both of these methods result in the same smoothed value on *day number*

A. 3
B. 4
C. 5
D. 6
E. 7

VCAA 2019 Exam 1 Data analysis Q13

45% of students answered this question correctly.

Questions from multiple lessons

Data analysis

16. The relationship between *vaccination rate* (%) and *illness rate* (%) for a particular illness can be modelled by a least squares regression line. The equation of the line is

$illness\ rate = 66.41 - 0.61 \times vaccination\ rate$

The coefficient of determination for this relationship is 0.7012.

The Pearson correlation coefficient, r, is closest to

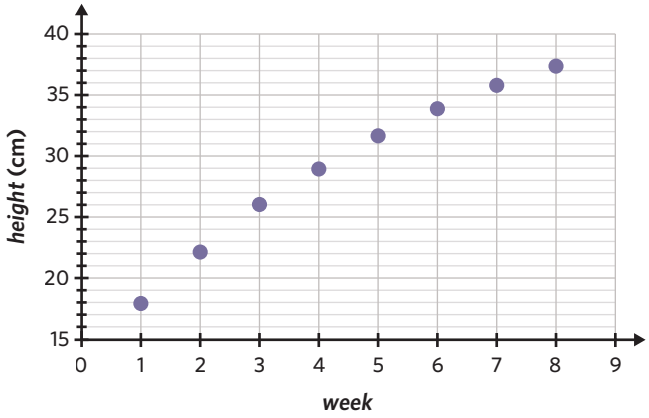
A. 0.8884 **B.** 0.8374 **C.** 0.4917 **D.** −0.4917 **E.** −0.8374

Adapted from VCAA 2018NH Exam 1 Data analysis Q13

Data analysis

17. The following table and scatterplot show the *height* of a plant after a number of weeks.

week	*height* (cm)
1	18.0
2	22.2
3	26.1
4	29.0
5	31.7
6	33.9
7	35.8
8	37.4



The scatterplot shows the relationship is non-linear.

A squared transformation is applied to the variable *height* to linearise the data.

A least squares regression line is then fitted to the data with *week* as the explanatory variable.

The equation of this line is closest to

A. $height^2 = 16.9 + 2.7 \times week$

B. $height^2 = 22.1 + 0.3 \times week$

C. $height^2 = 198.4 + 155.2 \times week$

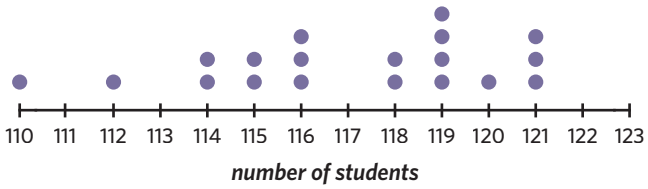
D. $week = -1.3 + 0.006 \times height^2$

E. $week = -64.8 + 3.1 \times height^2$

Adapted from VCAA 2017NH Exam 1 Data analysis Q12

Data analysis

18. The number of prep students at a school over the last 19 years was recorded. The data is displayed in the dot plot.



a. Determine the number of years in which there were more than 118 prep students. (1 MARK)

b. Determine if there is an outlier for the data, and if so, what is it? (2 MARKS)

Adapted from VCAA 2017NH Exam 2 Data analysis Q1

4D Seasonal adjustments

STUDY DESIGN DOT POINT

- seasonal adjustment including the use and interpretation of seasonal indices and their calculation using seasonal and yearly means

4A 4B 4C **4D** 4E

KEY SKILLS

During this lesson, you will be:

- calculating and interpreting seasonal indices
- deseasonalising a time series
- reseasonalising a time series
- plotting and interpreting a deseasonalised time series.

KEY TERMS

- Seasonal index
- Deseasonalising
- Reseasonalising

Some time series are affected by seasonality, where there is predictable variation in the data over regular intervals, such as days, months, or quarters. Seasonality can be distracting and make it difficult to observe the overall trends. By deseasonalising a data set, it is possible to minimise the effect of seasonality.

Calculating and interpreting seasonal indices

A **seasonal index** is a measure of how a particular season compares to the average season.

See worked example 1

When only one cycle of seasonal data is given, a seasonal index can be calculated using the formula:

$$\textit{seasonal index} = \frac{\textit{season value}}{\textit{mean of all seasons}}$$

Once the seasonal index has been calculated, it can then be interpreted in relation to the average season. For example, a seasonal index of 1.25 indicates that the value recorded for that season is typically 125% of, or 25% greater than, the average season. A seasonal index of 0.65 indicates that the value recorded for that season is typically 65% of, or 35% less than, the average season.

Because a seasonal index is the comparison of one season against an average season, the sum of the seasonal indices is equal to the number of seasons, and their average is 1.

If the data spans multiple cycles, the seasonal index for each season is the average of its seasonal indices in each cycle.

See worked example 2

Worked example 1

Consider the following table.

	summer	autumn	winter	spring
umbrella sales	205	377	528	442

a. Use the table to calculate the seasonal index for *umbrella sales* in summer, correct to two decimal places.

Explanation

Step 1: Calculate the mean of all seasons.

$$\frac{205 + 377 + 528 + 442}{4} = 388$$

Continues →

Step 2: Calculate the seasonal index for summer.

$$\text{seasonal index} = \frac{\text{season value}}{\text{mean of all seasons}}$$

$$= \frac{205}{388}$$

$$= 0.5283...$$

Answer

0.53

---

b. What does the rounded seasonal index for summer mean in terms of the *umbrella sales* in summer compared to the average season?

Explanation

Convert the seasonal index to a percentage.

$0.53 = 53\%$

Answer

On average, the number of umbrella sales in summer is 53% of, or 47% less than, the average season.

Worked example 2

For the following table, calculate the seasonal indices for summer, autumn, winter, and spring, correct to two decimal places.

	season			
year	summer	autumn	winter	spring
2019	45	30	22	51
2020	53	29	26	60
2021	54	41	29	72

Explanation

Step 1: Calculate the seasonal indices for each season in 2019.

$$\text{mean of all seasons} = \frac{45 + 30 + 22 + 51}{4}$$

$$= 37$$

$$\text{seasonal index} = \frac{\text{season value}}{\text{mean of all seasons}}$$

Summer 2019: $\text{seasonal index} = \frac{45}{37} = 1.216...$

Autumn 2019: $\text{seasonal index} = \frac{30}{37} = 0.810...$

Winter 2019: $\text{seasonal index} = \frac{22}{37} = 0.594...$

Spring 2019: $\text{seasonal index} = \frac{51}{37} = 1.378...$

Step 2: Calculate the seasonal indices for each season in 2020.

$$\text{mean of all seasons} = \frac{53 + 29 + 26 + 60}{4}$$

$$= 42$$

$$\text{seasonal index} = \frac{\text{season value}}{\text{mean of all seasons}}$$

Summer 2020: $\text{seasonal index} = \frac{53}{42} = 1.261...$

Autumn 2020: $\text{seasonal index} = \frac{29}{42} = 0.690...$

Winter 2020: $\text{seasonal index} = \frac{26}{42} = 0.619...$

Spring 2020: $\text{seasonal index} = \frac{60}{42} = 1.428...$

Continues →

Step 3: Calculate the seasonal indices for each season in 2021.

$$\text{mean of all seasons} = \frac{54 + 41 + 29 + 72}{4}$$
$$= 49$$
$$\text{seasonal index} = \frac{\text{season value}}{\text{mean of all seasons}}$$

Summer 2021: $\text{seasonal index} = \frac{54}{49} = 1.102...$

Autumn 2021: $\text{seasonal index} = \frac{41}{49} = 0.836...$

Winter 2021: $\text{seasonal index} = \frac{29}{49} = 0.591...$

Spring 2021: $\text{seasonal index} = \frac{72}{49} = 1.469...$

Step 4: Find the average seasonal indices by calculating the mean for each season.

Summer:
$\text{seasonal index} = \frac{1.216... + 1.261... + 1.102...}{3} \approx 1.19$

Autumn:
$\text{seasonal index} = \frac{0.810.... + 0.690... + 0.836...}{3} \approx 0.78$

Winter:
$\text{seasonal index} = \frac{0.594... + 0.619... + 0.591...}{3} \approx 0.60$

Spring:
$\text{seasonal index} = \frac{1.378... + 1.428... + 1.469...}{3} \approx 1.43$

Answer

summer: 1.19, autumn: 0.78, winter: 0.60, spring: 1.43

Deseasonalising a time series

Deseasonalising a time series removes the seasonal variation from the data. Any underlying trends can then be observed.

See worked example 3

Deseasonalised values can be calculated using the formula:

$$\text{deseasonalised value} = \frac{\text{actual value}}{\text{seasonal index}}$$

The concept of deseasonalisation can also be applied when analysing how data can be adjusted to correct for seasonality.

See worked example 4

Worked example 3

Use the given seasonal indices to deseasonalise the following time series data. Give values correct to the nearest whole number.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
seasonal index	0.70	1.35	0.80	0.75	1.20	1.70	0.50
number of steps	7500	14 000	6000	8000	13 000	17 000	4000
deseasonalised *number of steps*							

Explanation

Step 1: Identify the relevant values for Monday.

$\text{seasonal index} = 0.70$

$\text{actual value} = 7500$

Step 2: Substitute the values into the formula.

$$\text{deseasonalised value} = \frac{\text{actual value}}{\text{seasonal index}}$$
$$= \frac{7500}{0.70}$$
$$\approx 10\,714$$

Step 3: Repeat the process for the remaining days and fill in the table.

Answer

Mon: 10 714, Tue: 10 370, Wed: 7500, Thu: 10 667, Fri: 10 833, Sat: 10 000, Sun: 8000

Worked example 4

The seasonal index for January is 0.8. How should data for this season be adjusted to correct for seasonality?

Explanation

Step 1: Identify the seasonal index.

seasonal index $= 0.8$

Step 2: Find the reciprocal of the seasonal index.

Divide 1 by the seasonal index.

$1 \div 0.8 = 1.25$

Step 3: Interpret the value of the reciprocal.

The reciprocal value of 1.25 suggests that deseasonalised values are 125% of the actual value. This means that values for the season need to be increased by 25%.

Answer

The data from January should be increased by 25% to correct for seasonality.

Reseasonalising a time series

Reseasonalising a time series restores the normal seasonal fluctuations of previously deseasonalised data. This results in data that is representative of actual values in the data set.

Reseasonalised values can be calculated using the formula:

actual value $=$ *seasonal index* $\times$ *deseasonalised value*

Worked example 5

The seasonal index for Australian *Oodie sales* in January is 0.65.

Last January, a store in Sydney sold a deseasonalised number of 540 Oodies.

How many Oodies did the store actually sell in January?

Explanation

Step 1: Identify the relevant values.

seasonal index $= 0.65$

deseasonalised value $= 540$

Step 2: Substitute the values into the formula.

$$\begin{aligned} \textit{actual value} &= \textit{seasonal index} \times \textit{deseasonalised value} \\ &= 0.65 \times 540 \\ &= 351 \end{aligned}$$

Answer

351 Oodies

Plotting and interpreting a deseasonalised time series

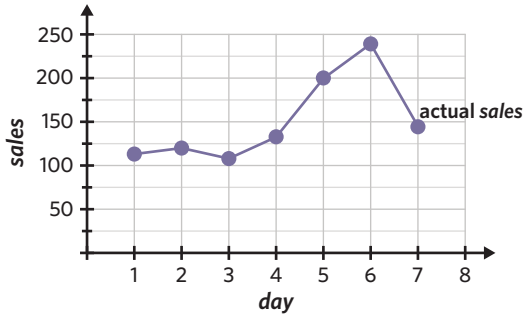
A deseasonalised time series can be plotted on a graph to visualise the underlying trends in seasonal data. It can be beneficial to deseasonalise a time series before fitting a regression line or equation.

Worked example 6

Consider the following table.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
coded *day*	1	2	3	4	5	6	7
sales	113	120	108	133	201	240	145
seasonal index	0.62	0.67	0.71	0.85	1.42	1.63	1.1
deseasonalised *sales*	182	179	152	156	142	147	132

The *sales* have been plotted on the following graph.

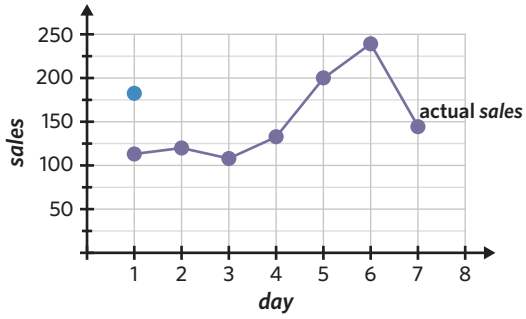


a. Plot the deseasonalised data on the graph.

Explanation

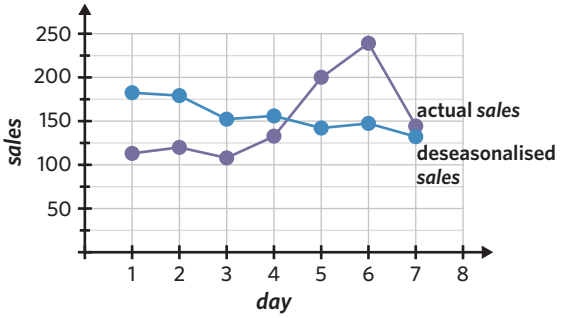
Step 1: Plot the deseasonalised sales for day 1.

Day 1 will have a horizontal axis value of 1 and a vertical axis value of 182.



Step 2: Plot the deseasonalised data for the rest of the days and join the points with a line.

Answer



Continues →

b. Identify any underlying trends using the plotted deseasonalised data.

Explanation

The data is decreasing from left to right.

Answer

There is a decreasing trend in *sales* over time.

Exam question breakdown

VCAA 2018 Exam 1 Data analysis Q16

The quarterly sales figures for a large suburban garden centre, in millions of dollars, for 2016 and 2017 are displayed in the following table.

year	quarter 1	quarter 2	quarter 3	quarter 4
2016	1.73	2.87	3.34	1.23
2017	1.03	2.45	2.05	0.78

Using these sales figures, the seasonal index for quarter 3 is closest to

A. 1.28 **B.** 1.30 **C.** 1.38 **D.** 1.46 **E.** 1.48

Explanation

Step 1: Calculate the 2016 seasonal index for quarter 3.

$$\frac{1.73 + 2.87 + 3.34 + 1.23}{4} = 2.2925$$

$$\textit{seasonal index} = \frac{3.34}{2.2925}$$

$$= 1.456...$$

Step 2: Calculate the 2017 seasonal index for quarter 3.

$$\frac{1.03 + 2.45 + 2.05 + 0.78}{4} = 1.5775$$

$$\textit{seasonal index} = \frac{2.05}{1.5775}$$

$$= 1.299...$$

Step 3: Calculate the average seasonal index for quarter 3.

$$\frac{1.456... + 1.299...}{2} = 1.378...$$

Answer

C

51% of students answered this question correctly.

19% of students incorrectly answered option D. These students found the seasonal index for 2016 only, rather than the average for both 2016 and 2017.

4D Questions

Calculating and interpreting seasonal indices

1. The seasonal index for *sales* in June is 0.73.

Which statement is correct?

A. On average, *sales* in June are 73% less than the monthly average.

B. On average, *sales* in June are 27% less than the monthly average.

C. On average, *sales* in June are 27% more than the monthly average.

D. On average, *sales* in June are 73% more than the monthly average.

2. The quarterly *sales* ($000's) of a canine clothing shop are displayed in the following table.

	quarter 1	quarter 2	quarter 3	quarter 4
***sales* ($000's)**	33.5	61.3	52.0	43.4

a. Calculate the seasonal index for quarter 1, correct to two decimal places.

b. What does the rounded seasonal index for quarter 1 mean in terms of the amount of *sales* in quarter 1 compared to the average quarter?

---

3. Consider the following table.

	summer	autumn	winter	spring
seasonal index	0.65	1.15		0.80

What is the seasonal index for winter?

A. 0.87 **B.** 1.20 **C.** 1.25 **D.** 1.40

---

4. Calculate the seasonal indices for the following time series data. Give values correct to two decimal places.

a.

	summer	autumn	winter	spring
customers	55	78	96	75

b.

	day						
week	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	11	22	25	32	22	15	6
2	10	25	24	30	24	14	8

c.

	quarter			
year	1	2	3	4
2001	10	15	20	15
2002	11	17	19	15
2003	12	18	24	17

Deseasonalising a time series

5. Consider the following table.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
seasonal index	0.87	0.94	0.91	0.85	0.96	1.26	1.21
value	603	665	509	670	702	931	950

The deseasonalised value for Thursday is closest to

A. 570 **B.** 583 **C.** 771 **D.** 788

---

6. The *number of weddings* held in the Botanical Gardens is seasonal, with the *number of weddings* in spring having a seasonal index of 1.32. If there were 58 weddings held in the Botanical Gardens this spring, the deseasonalised *number of weddings* is closest to

A. 39 **B.** 43 **C.** 44 **D.** 77

7. Use the seasonal indices to deseasonalise the following time series data. Give values correct to one decimal place.

a.

	summer	autumn	winter	spring
seasonal index	1.20	0.75	1.01	1.04
number of thunderstorms	37	30	29	33
deseasonalised *number of thunderstorms*				

b.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
seasonal index	1.25	0.80	1.23	0.72	0.86	1.45	0.69
kilometres cycled	55	26	47	26	32	63	15
deseasonalised *kilometres cycled*							

8. The seasonal index for winter is 0.8. How should data from this season be adjusted to correct for seasonality?

A. Data from winter should be increased by 20%.

B. Data from winter should be increased by 25%.

C. Data from winter should be decreased by 20%.

D. Data from winter should be decreased by 25%.

9. The *number of holidays* booked by Essendon supporters in September has a seasonal index of 2.5.

a. To correct for seasonality, by what percentage does the *number of holidays* need to be decreased?

b. Calculate the deseasonalised *number of holidays* booked if a total of 28 400 holidays were booked by Essendon supporters in September 2022.

Reseasonalising a time series

10. The deseasonalised value for a particular season is 927. If the seasonal index is 0.75, the actual value is closest to

A. 695 **B.** 864 **C.** 927 **D.** 1236

11. Consider the following table.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
seasonal index	0.65	0.55	0.60	0.80	1.00	1.15	1.50	1.45	1.40	1.25	0.90	0.75
deseasonalised *value*	86	79	80	75	78	72	66	71	72	65	66	61

The actual value for July is

A. 44 **B.** 66 **C.** 83 **D.** 99

12. Baxter has collected the following data regarding the deseasonalised number of *reservations* at his restaurant during the week. The original data that contained the actual number of *reservations* on each day is no longer available. Use the given seasonal indices to reseasonalise the time series data. Give values correct to the nearest whole number where necessary.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
seasonal index	0.65	0.60	0.54	0.75	1.40	1.85	1.21
deseasonalised *reservations*	64.62	81.67	53.70	74.67	60.00	83.24	104.13
actual *reservations*							

Plotting and interpreting a deseasonalised time series

13. The following graph shows the actual *sales* and deseasonalised *sales*.

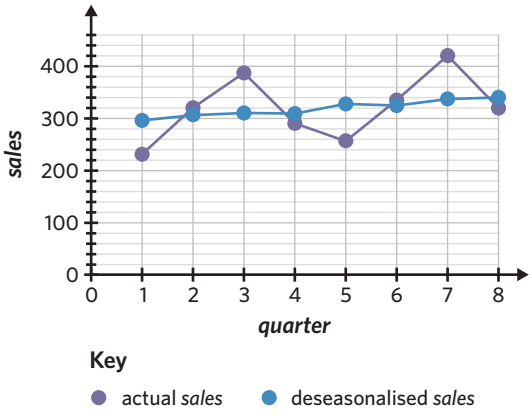
Which of the following statements is true?

A. When adjusted for seasonality, *sales* are increasing over time.

B. When adjusted for seasonality, *sales* are decreasing over time.

C. When adjusted for seasonality, *sales* are increasing over time with two large fluctuations.

D. When adjusted for seasonality, there is no underlying trend in *sales*.



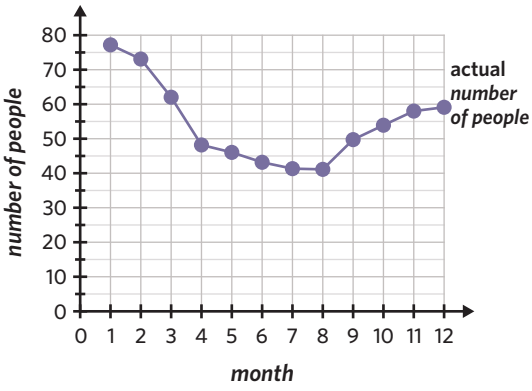
14. Hilaria is a fashion trend forecaster and for years has been researching the popularity of her least favourite sandal, the Birkenstock. The *number of people* wearing Birkenstocks at her workplace displays monthly seasonality, and her data from last year is shown in the following table.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
coded *month*	1	2	3	4	5	6	7	8	9	10	11	12
long-term seasonal index	1.25	1.21	1.06	0.86	0.82	0.81	0.79	0.77	0.96	1.10	1.16	1.21
actual *number of people*	77	73	62	48	46	43	41	41	50	54	58	59
deseasonalised *number of people*	61.6	60.3	58.5	55.8	56.1	53.1	51.9	53.2	52.1	49.1	50.0	48.8

Hilaria has constructed a graph to display the actual *number of people* seen wearing Birkenstocks each month last year.

a. Plot the deseasonalised values.

b. Identify any underlying trend visible in the deseasonalised data.



Joining it all together

15. The *sales* ($000's) for a churro company are seasonal, with the sales for 2020, 2021, and 2022 shown in the following table. The season index for autumn is 1.06 and the seasonal index for winter is 1.32.

	***sales* ($000's)**			
year	summer	autumn	winter	spring
2020	2050	3650	4300	3125
2021	1900	3075	4150	3300
2022	1625	3300	4050	3250

a. Calculate the average seasonal indices of *sales* for summer and spring. Give values correct to two decimal places.

b. What does the seasonal index for autumn tell us about the amount of *sales* in autumn compared to the average season?

c. Fill in the gap in the following sentence, giving the value correct to the nearest percent.

To correct for seasonality, the amount of *sales* in winter should be decreased by _____%.

---

16. The *number of tourists* visiting a small patisserie in Paris each month is seasonal. The *number of tourists* who visited in 2022 is shown in the following table.

	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
coded *month*	1	2	3	4	5	6	7	8	9	10	11	12
long-term seasonal index	1.05	0.83	0.76	0.81	0.92	1.03	1.21	1.25	1.10		0.97	1.12
actual *number of tourists*	830	644		703	795	911	1123	1160	1045	910	934	1095
deseasonalised *number of tourists*	790.5	775.9	790.8	867.9	864.1	884.5	928.1	928.0	950.0	957.9	962.9	

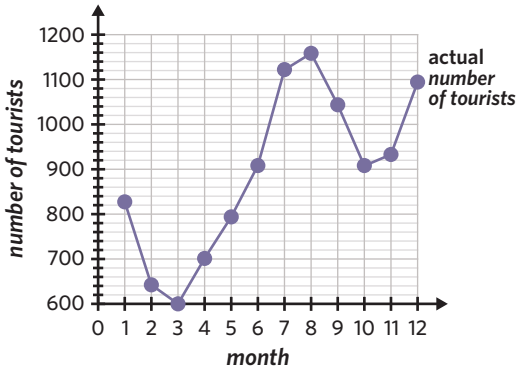
a. What is the long-term seasonal index for October?

b. Calculate the actual *number of tourists* who visited the patisserie in March, correct to the nearest whole number.

c. Calculate the deseasonalised *number of tourists* who visited the patisserie in December, correct to one decimal place.

d. The actual *number of tourists* who visited the patisserie in 2022 are shown in the following graph.

Plot the deseasonalised *number of tourists*.



e. Describe any underlying trend in the data, as shown by the deseasonalised plot.

Exam practice

17. The following table shows the monthly *rainfall* for 2019, in millimetres, recorded at a weather station, and the associated long-term seasonal indices for each month of the year.

	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
***rainfall* (mm)**	18.4	17.6	46.8	23.6	92.6	77.2	80.0	86.8	93.8	55.2	97.3	69.4
seasonal index	0.728	0.734	0.741	0.934	1.222	0.973	1.024	1.121	1.159	1.156	1.138	1.072

The deseasonalised *rainfall* for May 2019 is closest to

A. 71.3 mm **B.** 75.8 mm **C.** 86.1 mm
D. 88.1 mm **E.** 113.0 mm

VCAA 2020 Exam 1 Data analysis Q17

88% of students answered this question correctly.

18. The *total rainfall*, in millimetres, for each of the four seasons in 2015 and 2016 is shown in Table 1.

Table 1

	***total rainfall* (mm)**			
year	summer	autumn	winter	spring
2015	142	156	222	120
2016	135	153	216	96

a. The seasonal index for winter is 1.41.

Use the values in Table 1 to find the seasonal indices for summer, autumn, and spring. (2 MARKS)

b. The *total rainfall* for each of the four seasons in 2017 is shown in Table 2.

Table 2

	***total rainfall* (mm)**			
year	summer	autumn	winter	spring
2017	141	156	262	120

Use the appropriate seasonal index to deseasonalise the *total rainfall* for winter in 2017. Round your answer to the nearest whole number. (1 MARK)

VCAA 2019 Exam 2 Data analysis Q6a,b

Part **a**: The average mark for this question was **0.9**.
Part **b**: **58%** of students answered this question correctly.

---

19. The seasonal index for the *sales* of cold drinks in a shop in January is 1.6.

To correct the January *sales* of cold drinks for seasonality, the actual *sales* should be

A. reduced by 37.5%.
B. reduced by 40%.
C. reduced by 62.5%.
D. increased by 60%.
E. increased by 62.5%.

VCAA 2017 Exam 1 Data analysis Q16

32% of students answered this question correctly.

Questions from multiple lessons

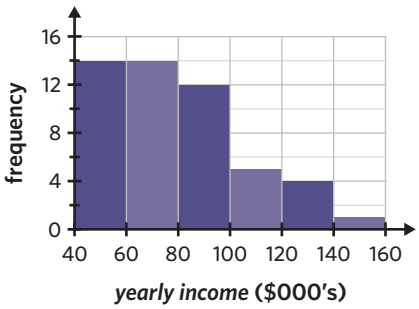
Data analysis *Year 11 content*

20. The following histogram displays the distribution of *yearly income* for 50 adults.

The shape of this distribution can be described as

A. symmetric.
B. approximately symmetric.
C. approximately normal.
D. negatively skewed.
E. positively skewed.

Adapted from VCAA 2017NH Exam 1 Data analysis Q1



Data analysis *Year 11 content*

21. Displayed in the following table is the *length*, in centimetres, and *weight*, in kilograms, of 10 ferrets.

***length* (cm)**	38	32	35	39	30	29	37	33	34	36
***weight* (kg)**	1.1	0.8	1.5	1.4	1.1	0.9	1.6	1.6	1.2	1.3

Find the value of the Pearson correlation coefficient, *r*, between *height* and *weight*, correct to two decimal places.

A. -0.53 **B.** -0.28 **C.** 0.27 **D.** 0.28 **E.** 0.53

Adapted from VCAA 2018NH Exam 1 Data analysis Q9

Data analysis

22. The *height* and *width*, in centimetres, of 15 houseplants were recorded and a least squares regression line was fitted to the data. The equation of the line is $width = 4.8 + 0.484 \times height$.

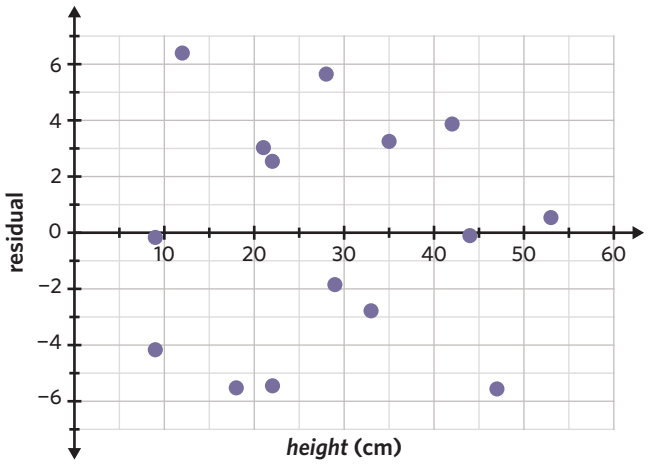
a. The regression equation is used to predict the *width* of a plant with a *height* of 47 cm. The residual value is calculated to be -5.55.

What is the actual *width* of this plant, correct to one decimal place? (1 MARK)

b. Pearson's correlation coefficient, *r*, for this data is 0.8549.

What percentage of variation in *width* can **not** be explained by the variation in *height*? Give your answer correct to one decimal place. (1 MARK)

c. When the regression line is fitted to the data, the following residual plot is acquired.



What information can be deduced from the residual plot about the association between *height* and *width* for these houseplants? (1 MARK)

Adapted from VCAA 2018NH Exam 2 Data analysis Q5d-f

4E Time series data and least squares regression modelling

STUDY DESIGN DOT POINT

- modelling trend by fitting a least squares line to a time series with time as the explanatory variable (data de-seasonalised where necessary), and the use of the model to make forecasts (with re-seasonalisation where necessary) including consideration of the possible limitations of fitting a linear model and the limitations of extending into the future

KEY SKILLS

During this lesson, you will be:

- modelling time series data
- modelling seasonal data.

It can be helpful to fit trend lines to time series data in order to make predictions about the future. If the data is seasonal, it must first be deseasonalised before fitting a trend line, and then reseasonalised after making predictions.

Modelling time series data

Trend lines for time series data are treated the same as least squares regression lines. After a trend line has been fitted, it can be used to make predictions outside the range of the data set. However, the limitations of extrapolation are still present. When extrapolating, it is assumed that the shape of the relationship between the variables will continue outside of the range of the data set. This assumption has limited reliability.

Worked example 1

A public library introduced a record collection to their catalogue. The *number of visitors* to the library each *day* following its introduction was recorded.

The day that the record collection was introduced is noted as day 0 and the first day after the introduction is day 1.

day	1	2	3	4	5	6	7
number of visitors	94	114	120	153	178	191	221

a. Determine the equation of the least squares regression line for the data. Give values correct to two decimal places where necessary.

Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

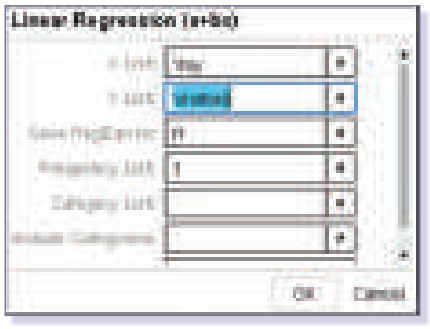
Step 2: Name column A 'day' and column B 'visitors'.

Enter the *day* values into column A, starting from row 1.

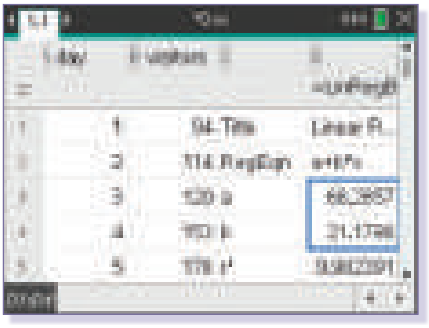
Enter the *number of visitors* values into column B, starting from row 1.

Continues →

Step 3: Press menu and select '4: Statistics' → '1: Stat Calculations' → '4: Linear Regression (a+bx)'. Select 'day' in 'X List:' and 'visitors' in 'Y List:'. Select 'OK'.



Step 4: Write down the equation for the least squares regression line, rounding the values of a and b to two decimal places.



$y = 68.29 + 21.18 \times x$

Step 5: Rewrite the equation in terms of the variables in the question.

The explanatory variable is *day*.

The response variable is *number of visitors*.

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Step 2: Name the first list 'day' and the second list 'visitors'.

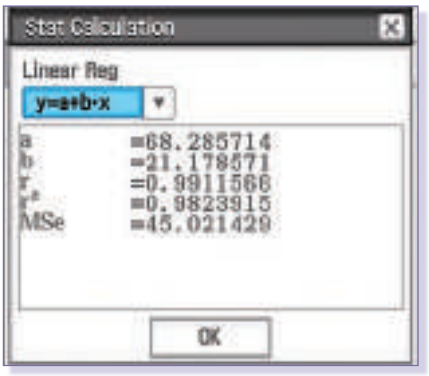
Enter the *day* values into list 'day', starting from row 1.

Enter the *number of visitors* values into list 'visitors', starting from row 1.

Step 3: Tap 'Calc' → 'Regression' → 'Linear Reg'.

Specify the data set by changing 'XList:' to 'main\day' and 'YList:' to 'main\visitors'. Tap 'OK'.

Step 4: Change the form of the equation in the drop down box to 'y=a+b·x'. Write down the equation for the least squares regression line, rounding the values of a and b to two decimal places.



$y = 68.29 + 21.18 \times x$

Step 5: Rewrite the equation in terms of the variables in the question.

The explanatory variable is *day*.

The response variable is *number of visitors*.

Answer – Method 1 and 2

number of visitors $= 68.29 + 21.18 \times$ *day*

Continues →

b. Use the rounded regression equation to estimate the *number of visitors* on the 8th day after the introduction. Round to the nearest whole number.

Explanation

Substitute $day = 8$ into the regression equation and evaluate.

$$\begin{aligned} number\ of\ visitors &= 68.29 + 21.18 \times day \\ &= 68.29 + 21.18 \times 8 \\ &= 237.73 \end{aligned}$$

Answer

238 visitors

---

c. Use the regression equation to estimate the *number of visitors* to the library on the day that the record collection was introduced. Round to the nearest whole number.

Explanation

The *number of visitors* on day 0 is equal to the y-intercept of the regression line.

The y-intercept is 68.29.

Answer

68 visitors

Modelling seasonal data

A trend line can be fitted to seasonal data in a similar way to regular time series data. However, additional steps are needed to improve the accuracy of predictions.

As seasonal data has many fluctuations, it is important to deseasonalise it before fitting a trend line. Recall that data can be deseasonalised using the formula:

$$deseasonalised\ value = \frac{actual\ value}{seasonal\ index}$$

This means that any predictions made using the trend line will result in a deseasonalised value. Predictions must be reseasonalised afterwards using the formula:

$$actual\ value = seasonal\ index \times deseasonalised\ value$$

Worked example 2

The *sales* of NRL jerseys at a merchandise store are seasonal. The *sales* for the first six months of 2022 are shown in the following table.

month	Jan	Feb	Mar	Apr	May	Jun
sales	479	513	1127	894	800	802

The long-term *seasonal index* for each month is also provided.

month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
seasonal index	0.55	0.63	1.40	1.21	1.11	1.19	1.23	1.19	1.25	0.58	0.52	1.14

Continues →

a. Determine the equation of the least squares regression line to predict the deseasonalised *sales* for a month. Give values correct to one decimal place.

Note: When deseasonalising *sales*, round values correct to one decimal place.

Explanation – Method 1: TI-Nspire

Step 1: Deseasonalise the *sales* values from January to June.

Divide each month's *sales* values by its corresponding seasonal index. Give values correct to one decimal place.

Jan: $\frac{479}{0.55} = 870.9$

Feb: $\frac{513}{0.63} = 814.3$

Mar: $\frac{1127}{1.40} = 805.0$

Apr: $\frac{894}{1.21} = 738.8$

May: $\frac{800}{1.11} = 720.7$

Jun: $\frac{802}{1.19} = 673.9$

Step 2: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

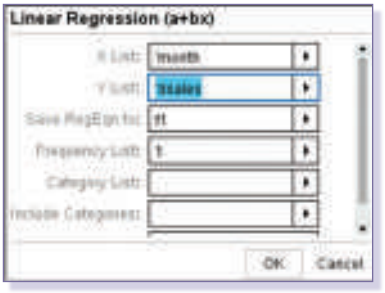
Step 3: Name column A 'month' and column B 'dsales'.

Enter the *month* values into column A, starting from row 1.

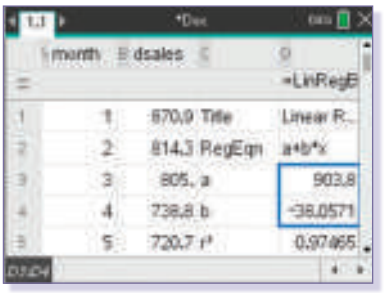
Note: Assign a value of 1 to January, 2 to February, and so on.

Enter the *deseasonalised sales* values into column B, starting from row 1.

Step 4: Press menu and select '4: Statistics' → '1: Stat Calculations' → '4: Linear Regression (a+bx)'. Select 'month' in 'X List:' and 'dsales' in 'Y List:'. Select 'OK'.



Step 5: Write down the equation for the least squares regression line, rounding the values of a and b correct to one decimal place.



$y = 903.8 - 38.1 \times x$

Step 6: Rewrite the equation in terms of the variables in the question.

The explanatory variable is *month*.

The response variable is *deseasonalised sales*.

Explanation – Method 2: Casio ClassPad

Step 1: Deseasonalise the *sales* values from January to June.

Divide each month's *sales* values by its corresponding seasonal index. Give values correct to one decimal place.

Jan: $\frac{479}{0.55} = 870.9$

Feb: $\frac{513}{0.63} = 814.3$

Mar: $\frac{1127}{1.40} = 805.0$

Apr: $\frac{894}{1.21} = 738.8$

May: $\frac{800}{1.11} = 720.7$

Jun: $\frac{802}{1.19} = 673.9$

Step 2: From the main menu, tap Statistics.

Step 3: Name the first list 'month' and the second list 'dsales'.

Enter the *month* values into list 'month', starting from row 1.

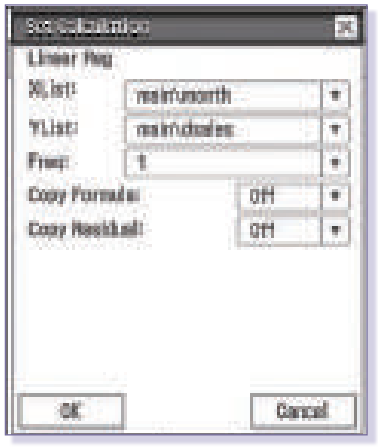
Note: Assign a value of 1 to January, 2 to February, and so on.

Enter the *deseasonalised sales* values into list 'dsales', starting from row 1.

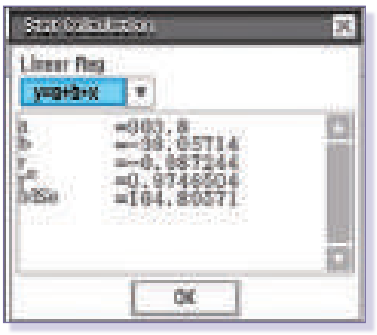
Continues →

Step 4: Tap 'Calc' → 'Regression' → 'Linear Reg'.

Specify the data set by changing 'XList:' to 'main\month' and 'YList:' to 'main\dsales'. Tap 'OK'.



Step 5: Change the form of the equation in the drop down box to 'y=a+b·x'. Write down the equation for the least squares regression line, rounding the values of a and b correct to one decimal place.



$y = 903.8 - 38.1 \times x$

Step 6: Rewrite the equation in terms of the variables in the question.

The explanatory variable is *month*.

The response variable is *deseasonalised sales*.

Answer – Method 1 and 2

deseasonalised sales $= 903.8 - 38.1 \times$ *month*

---

b. Use the rounded regression equation to predict the *sales*, correct to the nearest whole number, for October 2022.

Explanation

Step 1: Calculate the *deseasonalised sales* for October.

October is the 10^{th} month.

Substitute *month* $= 10$ into the regression equation and evaluate.

$$\begin{aligned} \textit{deseasonalised sales} &= 903.8 - 38.1 \times \textit{month} \\ &= 903.8 - 38.1 \times 10 \\ &= 522.8 \end{aligned}$$

Step 2: Reseasonalise the value.

$$\begin{aligned} \textit{actual value} &= \textit{seasonal index} \times \textit{deseasonalised value} \\ &= 522.8 \times 0.58 \\ &= 303.224 \end{aligned}$$

Answer

303 sales

---

c. Interpret the slope of the rounded regression line in terms of the change in *deseasonalised sales* each month.

Explanation

Step 1: Identify the slope.

The slope of the regression line is 38.1.

Step 2: Interpret the slope.

The slope indicates the average change in the response variable for every one-unit increase in the explanatory variable.

Answer

On average, deseasonalised sales decrease by 38.1 each month.

Exam question breakdown

VCAA 2019 Exam 1 Data analysis Q14

The *time*, in minutes, that Liv ran each day was recorded for nine days. These times are shown in the table.

day number	1	2	3	4	5	6	7	8	9
***time* (mins)**	22	40	28	51	19	60	33	37	46

The following time series plot was generated from this data.

A least squares line is to be fitted to the time series plot shown.

The equation of this least squares line, with *day number* as the explanatory variable, is closest to

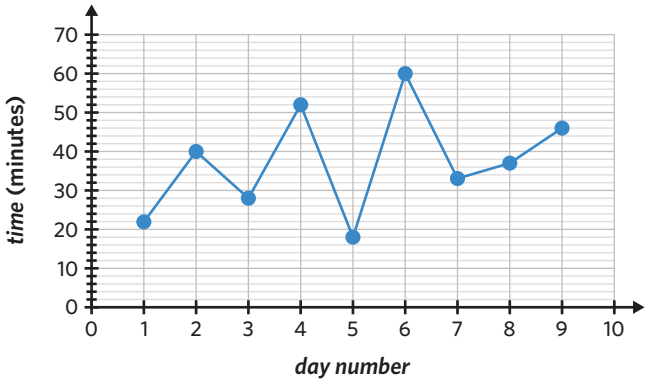
A. $day\ number = 23.8 + 2.29 \times time$

B. $day\ number = 28.5 + 1.77 \times time$

C. $time = 23.8 + 1.77 \times day\ number$

D. $time = 23.8 + 2.29 \times day\ number$

E. $time = 28.5 + 1.77 \times day\ number$



Explanation – Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

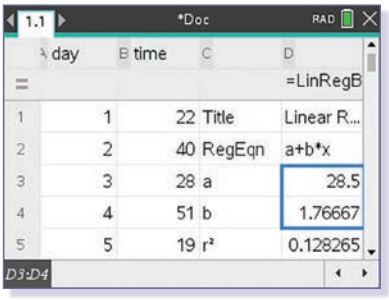
Step 2: Name column A 'day' and column B 'time'.

Enter the *day number* values into column A, starting from row 1.

Enter the *time* values into column B, starting from row 1.

Step 3: Press menu and select '4: Statistics' → '1: Stat Calculations' → '4: Linear Regression (a+bx)'. Select 'day' in 'X List:' and 'time' in 'Y List:'. Select 'OK'.

Step 4: Write down the equation for the least squares regression line, and round the values of a and b to the required number of decimal places.



$y = 28.5 + 1.77 \times x$

Step 5: Rewrite the equation in terms of the variables in the question.

The explanatory variable is *day number*.

The response variable is *time*.

$time = 28.5 + 1.77 \times day\ number$

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Step 2: Name the first list 'day' and the second list 'time'.

Enter the *day number* values into list 'day', starting from row 1.

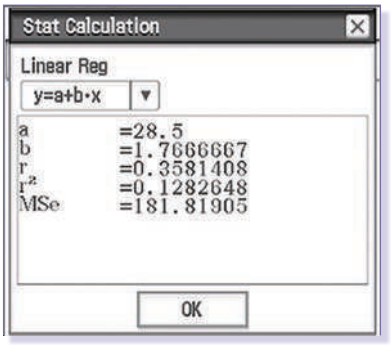
Enter the *time* values into list 'time', starting from row 1.

Step 3: Tap 'Calc' → 'Regression' → 'Linear Reg'.

Specify the data set by changing 'XList:' to 'main\day' and 'YList:' to 'main\time'. Tap 'OK'.

Continues →

Step 4: Change the form of the equation in the drop down box to 'y=a+b·x'. Write down the equation for the least squares regression line, and round the values of a and b to the required number of decimal places.



$y = 28.5 + 1.77 \times x$

Answer – Method 1 and 2

E

Step 5: Rewrite the equation in terms of the variables in the question.

The explanatory variable is *day number*.

The response variable is *time*.

$time = 28.5 + 1.77 \times day\ number$

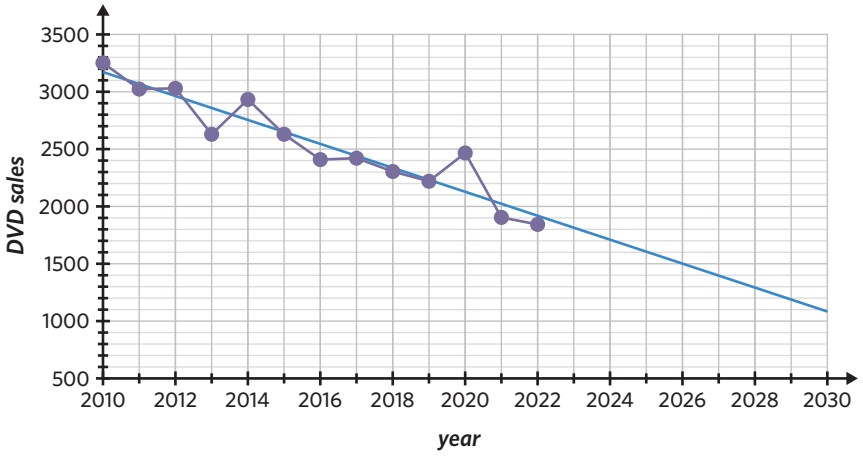
69% of students answered this question correctly.

12% of students incorrectly chose option B. This is likely because they mixed up the placement of the explanatory and response variables when replacing x and y with the actual variables.

4E Questions

Modelling time series data

1. An electronics store recorded their *DVD sales* between 2010 and 2022. They then fitted a regression line to the data, as shown.



From the regression line, *DVD sales* in 2030 are predicted to be closest to

A. 900 **B.** 1100 **C.** 1900 **D.** 2100

2. A new escape room has determined a regression equation to estimate their predicted *number of customers* each month for the upcoming year.

$number\ of\ customers = 316 + 32.4 \times month$

Note: In January, $month = 1$.

Predict the *number of customers*, correct to the nearest whole number, in

a. March. **b.** June. **c.** September. **d.** December.

3. Rafaella and Jeremiah recorded the daily *number of steps* they each walked for a week.

For each of them:

- find the regression equation used to predict the *number of steps* from the *day*. Round values correct to one decimal place.
- use the rounded regression equation to predict their *number of steps*, correct to the nearest whole number, on day 14.

a. Rafaella:

day	1	2	3	4	5	6	7
number of steps	7700	6550	9100	8600	8950	10 200	9850

b. Jeremiah:

day	1	2	3	4	5	6	7
number of steps	7450	8600	7000	6350	7150	6100	5200

---

4. Leonard conducted research into the number of *alien sightings* from 1975 to 2015 in the northern and southern hemispheres. He then used this data to construct regression equations to predict the number of *alien sightings* in each hemisphere for any given year.

Note: Leonard has used $year = 0$ to represent 1975.

The least squares regression equations are:

Southern hemisphere: $alien\ sightings = 208 + 23.4 \times year$

Northern hemisphere: $alien\ sightings = 361 + 41.2 \times year$

a. Estimate the difference between the number of *alien sightings* in each hemisphere in the year 2030, correct to the nearest whole number.

b. Explain why this prediction may be of limited reliability, even if Leonard's original data was correct.

c. Interpret the slope of the regression line in terms of the change in the number of *alien sightings* in the northern hemisphere each year.

---

5. Daniel is currently 6 weeks into a 16-week exchange program in the Netherlands.

He recorded his *bank balance* at the end of each *week* for the first 6 weeks, as shown in the table.

week	***bank balance* ($)**
1	9058
2	8624
3	7580
4	7305
5	6617
6	6303

a. Daniel wants to keep track of his finances by calculating the least squares regression equation to predict his *bank balance* from *week*. Determine the regression equation, giving values correct to two decimal places.

b. Daniel is unsure whether he has enough money to fund the entire exchange. Use the rounded regression equation to predict Daniel's bank balance at the end of his exchange, correct to the nearest cent.

c. After how many full weeks on exchange will Daniel first have a *bank balance* under $3000?

d. On average, how much money does the rounded regression equation estimate that Daniel spends each week?

e. Using the rounded regression equation, estimate Daniel's bank balance at the start of his exchange.

Modelling seasonal data

6. A regression equation has predicted the *deseasonalised value* in September to be 3750.

If September has a long-term seasonal index of 1.09, the actual predicted *value* for September is closest to

A. 3413 **B.** 3440 **C.** 4088 **D.** 4121

7. The following equation can be used to forecast the *deseasonalised sales* ($) of a small business.

deseasonalised sales $= 20\,005 + 15\,000 \times$ *quarter number*

quarter number $= 1$ in quarter 1 of 2014 and *quarter number* $= 2$ in quarter 2 of 2014.

The seasonal index for the third quarter of each year is 0.65.

What is the actual value of *sales* in the third quarter of 2015?

---

8. The *average price*, in dollars, for a punnet of strawberries at Preston market was recorded for each season over 2021 and 2022. This is shown in the following table.

season	summer 2021	autumn 2021	winter 2021	spring 2021	summer 2022	autumn 2022	winter 2022	spring 2022
coded season	1	2	3	4	5	6	7	8
***average price* ($)**	2.15	2.80	4.15	3.95	2.60	3.65	5.20	5.25

The long-term seasonal indices for the seasons are also shown.

season	summer	autumn	winter	spring
seasonal index	0.68	0.9	1.26	1.16

The data was deseasonalised and a least squares regression line was fitted to the deseasonalised data.

Calculate the equation of the regression line for *deseasonalised average price* and *coded season*, giving values correct to two decimal places.

---

9. Calliope recorded the monthly *profit* ($) for her small business in 2022. She noticed that her *profit* was seasonal. She then calculated the *deseasonalised profit* ($) for each month in 2022, as shown in the following table.

month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
coded month	1	2	3	4	5	6	7	8	9	10	11	12
***deseasonalised profit* ($)**	8703	7105	7132	6810	7304	5923	6670	6543	5539	6435	6118	5605

a. Calculate the equation of the least squares regression line to predict *deseasonalised profit* ($) from *coded month*. Give values correct to one decimal place.

b. Calliope has calculated the long-term seasonal index for August to be 1.19. Predict Calliope's *profit*, correct to the nearest dollar, in August 2023 using the rounded regression equation from part **a**.

c. Is this prediction completely reliable? Explain briefly.

---

10. Eleanor is a freelance dog-walker. She recorded her quarterly *earnings*, in dollars, for 2021 and 2022.

quarter	Q1 2021	Q2 2021	Q3 2021	Q4 2021	Q1 2022	Q2 2022	Q3 2022	Q4 2022
coded quarter	1	2	3	4	5	6	7	8
***earnings* ($)**	8540	7515	7020	8890	9625	8145	8020	9950

Over the years, Eleanor has noticed that her *earnings* are seasonal.

The seasonal index for each quarter is shown in the following table.

quarter	Q1	Q2	Q3	Q4
seasonal index	1.18	0.92	0.89	1.06

a. Fill in the following table with Eleanor's *deseasonalised earnings*, correct to the nearest dollar.

quarter	Q1 2021	Q2 2021	Q3 2021	Q4 2021	Q1 2022	Q2 2022	Q3 2022	Q4 2022
coded quarter	1	2	3	4	5	6	7	8
***deseasonalised earnings* ($)**								

b. Plot the *deseasonalised earnings* against *coded quarter*. Comment on any visible trends and state whether a least squares regression line would be suitable.

c. Calculate the equation of the least squares regression line for *deseasonalised earnings* and *coded quarter*. Give values correct to the nearest whole number.

d. Use the rounded regression equation to predict Eleanor's *earnings* in Q2 2025.

e. In what quarter will Eleanor's *deseasonalised earnings* first be over $10 500?

Joining it all together

11. Fatima is studying mathematics at university, and decided to record the *number of students* that attended her university library each day for the first two weeks of the semester.

	week 1							**week 2**						
day	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
coded day	1	2	3	4	5	6	7	8	9	10	11	12	13	14
number of students	65	73	66	71	57	49	70	81	76	101	82	61	65	87

a. Fatima then fitted a least squares regression line to this data, to predict the *number of students* from *coded day*. Write down the equation of this line, giving values correct to two decimal places.

b. Each semester is 12 weeks long. Use the rounded regression to predict the *number of students*, correct to the nearest whole number, in the library on the last teaching day of the semester (Friday of Week 12).

c. Fatima later did some further research and discovered that the *number of students* in the library each day typically displays weekly seasonality. She calculates the long-term seasonal index for each day of the week, as shown.

day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
seasonal index	1.14	1.07	1.12	1.06	0.81	0.77	1.03

Use the given seasonal indices to deseasonalise the original data. Fill in the following table, giving values correct to two decimal places.

	week 1							**week 2**						
day	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
coded day	1	2	3	4	5	6	7	8	9	10	11	12	13	14
deseasonalised number of students														

d. Calculate the equation of the least squares regression line for the deseasonalised data, to predict the *deseasonalised number of students* from *coded day*. Give values correct to two decimal places.

e. Using the rounded regression equation from part **d**, calculate the predicted actual *number of students*, correct to the nearest whole number, in the library on the last teaching day of the semester.

f. Interpret the slope of the rounded regression line in terms of the change in *deseasonalised number of students* in the library each day of the semester.

Exam practice

12. The following table shows the yearly average traffic *congestion levels* in two cities, Melbourne and Sydney, during the period 2008 to 2016.

	***congestion level* (%)**								
year	2008	2009	2010	2011	2012	2013	2014	2015	2016
Melbourne	25	26	26	27	28	28	28	29	33
Sydney	28	30	32	33	34	34	35	36	39

A least squares line is used to model the trend in the time series plot for Sydney. The equation is

$congestion\ level = -2280 + 1.15 \times year$

a. i. Use the equation of the least squares line to determine the average rate of increase in percentage congestion level for the period 2008 to 2016 in Sydney.

Write in the box provided. (1 MARK)

[ ] % per year

ii. Use the least squares line to predict when the percentage congestion level in Sydney will be 43%. (1 MARK)

b. Use the data in the table to determine the equation of the least squares line that can be used to model the trend in the data for Melbourne. The variable *year* is the explanatory variable.

Write the values of the intercept and the slope of this least squares line in the appropriate boxes provided.

Round both values to four significant figures. (2 MARKS)

$congestion\ level =$ [ ] $+$ [ ] $\times\ year$

VCAA 2018 Exam 2 Data analysis Q3bii,biii,d

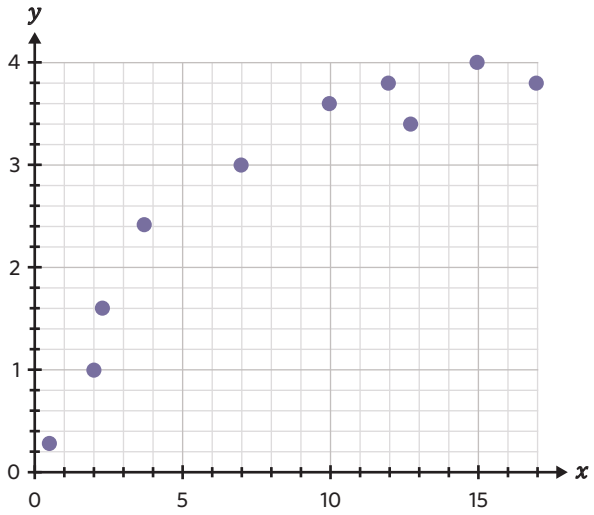
Part **ai**: **36%** of students answered this question correctly.
Part **aii**: **74%** of students answered this question correctly.
Part **b**: The average mark on this question was **1.4**.

Questions from multiple lessons

Data analysis

13. Annabelle uses the following data to generate the scatterplot shown.

x	y
15	4
7	3
3.7	2.4
0.5	0.3
17	3.8
2.3	1.6
12.7	3.4
12	3.8
2	1
10	3.6



The scatterplot demonstrates that the data is not linear, so to linearise the data, Annabelle performs a log transformation to the variable x.

Subsequently, she fits a least squares regression line to the transformed data.

With y as the response variable, the equation of this least squares regression line is closest to

A. $y = 0.78 + 2.59 \times \log(x)$

B. $y = -0.26 + 0.37 \times \log(x)$

C. $y = 1.03 + 0.20 \times \log(x)$

D. $\log(y) = 0.96 + 0.98 \times x$

E. $\log(y) = 1.75 + 3.49 \times x$

Adapted from VCAA 2018 Exam 1 Data analysis Q11

Data analysis

14. Disneyland records its long-term average number of visitors each day of the week.

The seasonal index for Tuesday is 0.79.

This means that, on average, the number of visitors to Disneyland on Tuesday is

A. 79% less than the daily average.

B. 21% less than the daily average.

C. the same as the daily average.

D. 21% more than the daily average.

E. 79% more than the daily average.

Adapted from VCAA 2016 Exam 1 Data analysis Q14

Data analysis

15. The least squares regression line that can predict a company's revenue, in dollars, based on its number of employees is

$revenue = 51.3 + 48.6 \times number\ of\ employees$

The correlation coefficient, r, for the relationship is 0.792.

a. Fizz Wizz is considered a small start-up company that produces confectionery for children. The company makes a total of \$25 000 in revenue and has 400 employees. Calculate the residual when the least squares regression line is used to predict the revenue of Fizz Wizz from its number of employees. (1 MARK)

b. What percentage of variation in the amount of revenue can be explained by the variation in the number of employees? Round to one decimal place. (1 MARK)

Adapted from VCAA 2014 Exam 2 Data analysis Q2di,ii