Student Name

Mathematics Stage 4

DATA ANALYSIS

Book 2 Interpret the effect individual data points have on measures of centre and range

Version: 260207

Report mistakes & suggest improvements:
https://MrDingMaths.com

CONTENTS

OVERVIEW

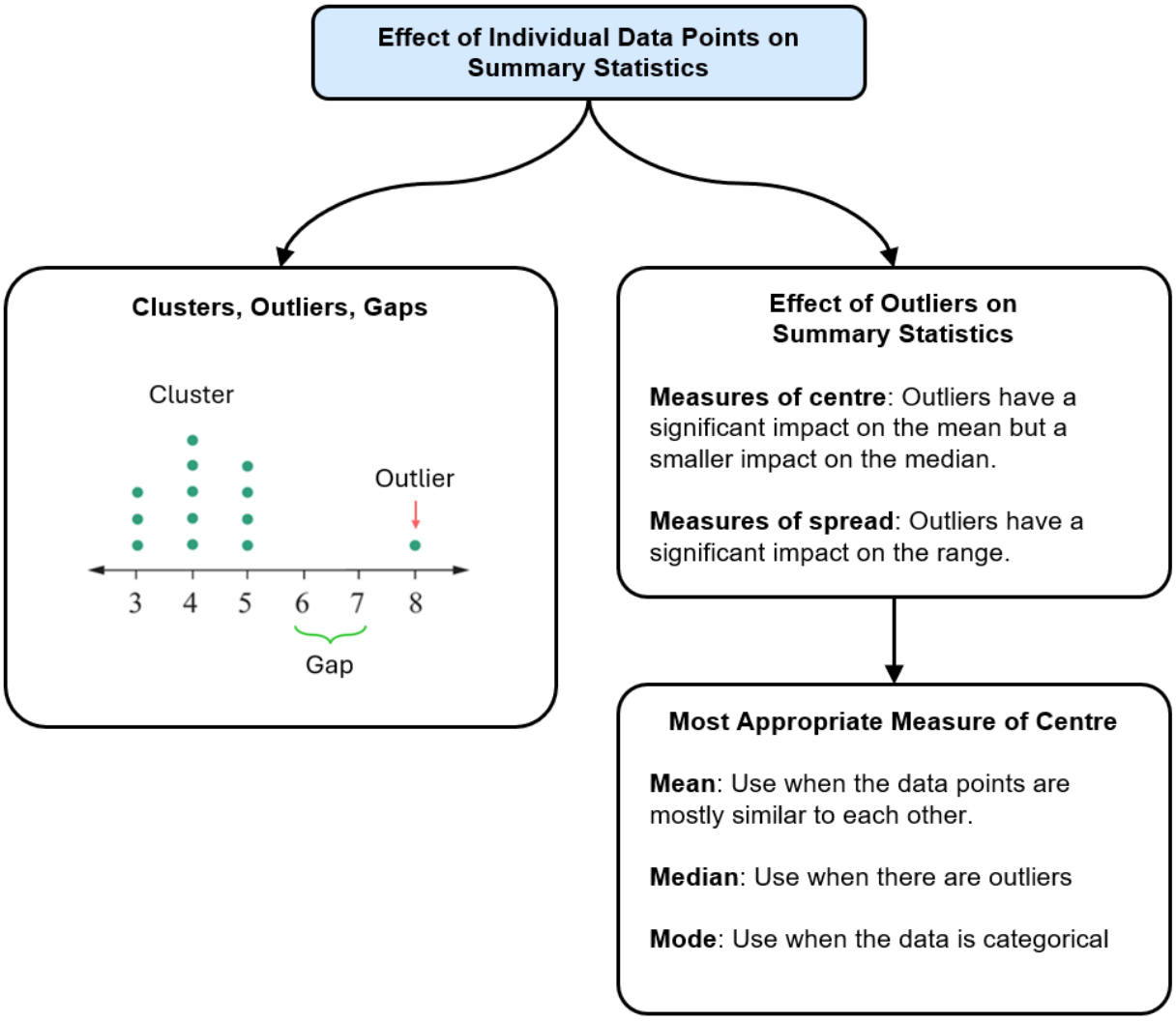
SYLLABUS CONTENT

MA4-DAT-C-02 analyses simple datasets using measures of centre, range and shape of the data

Interpret the effect individual data points have on measures of centre and range

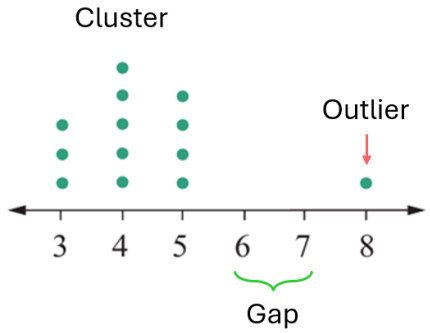
- Informally identify clusters, gaps and outliers in datasets and give reasons for their occurrence in the context of the data
- Identify and explain the impact of adding or removing data values that are clustered at one end of a dataset on the measures of centre
- Identify and explain the impact of outliers on the measures of centre and range
- Determine and justify the most appropriate measure of centre to summarise the data in its context

SUMMARY



CLUSTERS, GAPS, OUTLIERS

- A **cluster** is data **grouped** around specific values.
- An **outlier** is a data value **much higher or lower** than other data.
- A **gap** is an interval with no data.



1 Circle the outlier in these datasets.

a

172 102 93 89 100 105

b

2 2.7 7.9 3.12 1.8 2.2

c

3 0 −2 −1 9 −1

d

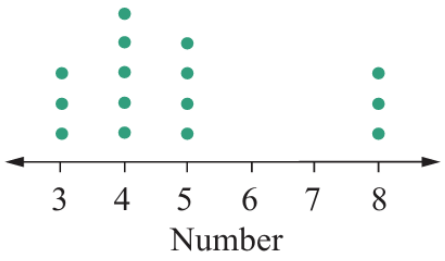
$\frac{8}{10}$ $\frac{41}{50}$ $\frac{7}{20}$ 0.77 $\frac{4}{5}$ $\frac{79}{100}$

2 Identify any clusters or gaps in the following sets of data.

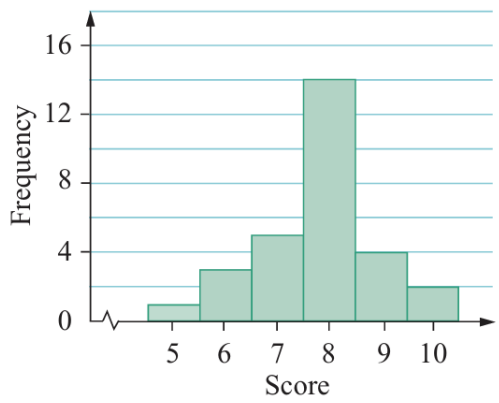
a

Stem	Leaves
4	8
5	3 6
6	0 1 1 3 4 4 4 6 8 8 8 9 9
7	5
8	1

b



c



d

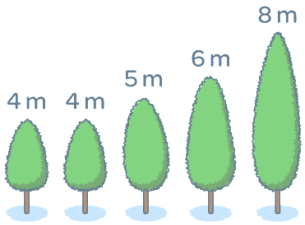
Stem	Leaves
5	2 4 4 7 9
6	0 5 6 6
7	
8	3 8
9	1 1 6 8 1

IMPACT OF ADDING OR REMOVING DATA VALUES

- Adding or removing data at the edges of a dataset predictably affects its summary statistics.

Example Describe the impact of adding or removing data values.

Consider the heights of these five trees. The summary statistics are:



Mean: 5.4 m

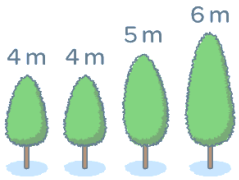
Median: 5 m

Mode: 4 m

Range: 4 m

One of the trees is cut down.

Calculate the summary statistics and describe how they've been affected.



Mean: $\bar{x} = \frac{4+4+5+6}{4} = 4.75$ m. The mean decreased.

Median: $\frac{5+6}{2} = 4.5$ m. The median decreased.

Mode: 4 m. The mode didn't change.

Range: $6 - 4 = 2$ m. The range decreased.

Which tree was cut down? How high was this tree compared to the others?

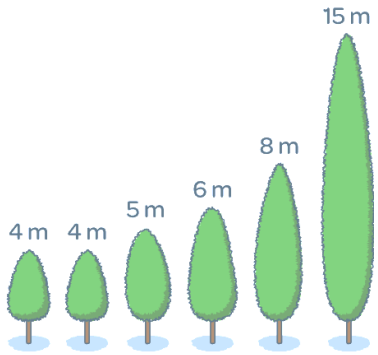
The tree of height .....m was cut down. This was the ..... tree.

What effect did this have on the mean and median? Which changed more?

By removing the ..... tree, the mean and median both .....

The ..... changed more.

Imagine that instead of a tree being cut down, a tree was planted.
Describe how the summary statistics change.



Mean:

The mean increased from 5.4 m to 7 m

Median:

The median increased from 5 m to 5.5 m

Mode:

The mode is unchanged from 4 m.

Range:

The range increased from 4 m to 11 m

Investigation Describe the impact of outliers on summary statistics.

Jericho and Jason organise their maths test marks.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	
Jericho	58	91	50	96	60	94	100	96	89	
Jason	95	84	82	70	76	70	72	71	85	

Who has the higher mean?

……

Who has the higher median?

……

Who has the smaller range?

……

Who do you think is better at tests? Explain why.

……

……

In Test 10, Jericho had a headache and scored 2, while Jason scored 90.
What happened to Jericho's mean? Does this represent how he usually performs?

……

……

……

IMPACT OF OUTLIERS

- **Outliers** in a dataset **strongly affect the mean** but have **less impact on the median**.

Data	Mean	Median
8 9 9 10 11	9.4	9
8 9 9 10 11 **100**	24.5	9.5

Example Describe the impact of outliers.

An egg farmer recorded the numbers of eggs laid by six hens over one year:

$$256, 267, 299, 304, 304, 394$$

The outlier is 394

Calculate the summary statistics with and without the outlier.

	Mean	Median	Mode	Range
With Outlier $256, 267, 299, 304, 304, 394$	304	301.5	304	138
Without Outlier $256, 267, 299, 304, 304$	286	299	304	48

The outlier increases the mean and range significantly.

The mode is the same and the median did not change by much.

The median is the better measure of centre.

In a community of 7 households, the yearly income is:

$$\$0,\ \$50\,000,\ \$50\,000,\ \$53\,325,\ \$58\,620,\ \$59\,950,\ \$61\,032,$$

The outlier is ……

Calculate the summary statistics with and without the outlier.

	Mean	Median	Mode	Range
With Outlier				
Without Outlier				

Comment on how the summary statistics have changed.

……

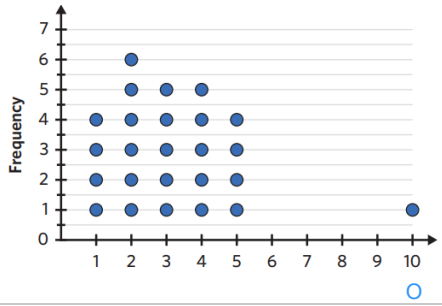
……

Key ideas

1. Outliers are data values that are very …… to the rest of the dataset.
2. When there are outliers, the …… is the better measure of centre as it is not greatly impacted by outliers.

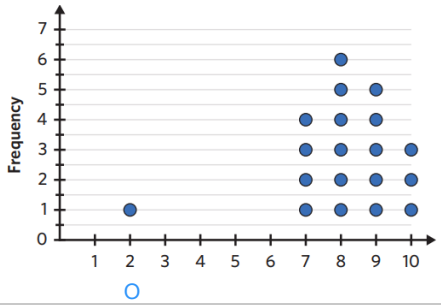
1 Circle the outlier in these datasets.

a



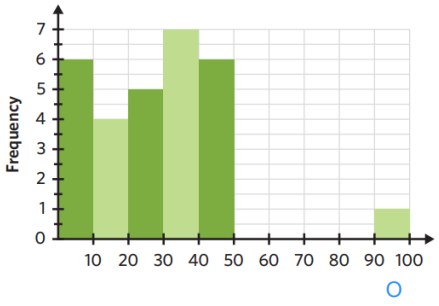
O

b



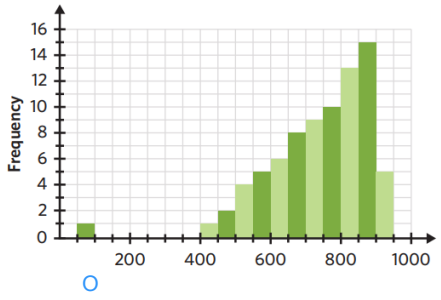
O

c



O

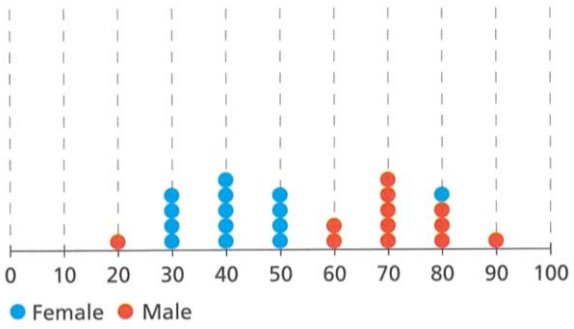
d



O

2 The chart shows the masses, in kg, of some leopards.

Charlie says there are no outliers.
Explain why he is incorrect.



There appears to be no outliers when all leopards are but together, but there are outliers when you consider females and males separately.

3 The mean mass of a team of rugby players is 95 kg.

a Would the mean increase or decrease if a new player of mass 80 kg joined the team?

decrease

b Describe how the mean would change if a player of mass 95 kg joins the team.

No change

c Describe how the mean would change if a player of mass 105 kg joins the team.

increase

d How would the mean change if the smallest item is removed from a data set?

increase

e How would the mean change if the largest item is removed from a data set?

decrease

4 Annie collected data on the length of time, in minutes, that people spent in a coffee shop one morning. The times are shown below:

1, 2, 4, 6, 7, 7, 12, 14, 25, 72

The range is 71 minutes. Explain why the range is not a good measure to choose to represent the spread of this dataset.

The range is significantly affected by the outliers.

5 Read this information:

- When there are outliers, we should either remove the outlier or use the median instead.
- We should only remove an outlier if:
  - The outlier is obviously due to a measurement mistake.
    E.g., measuring height of a person as 17.3 cm instead of 173 cm.
  - The outlier is due to a sampling error.
    E.g., surveying school students but accidentally asking a teacher.

When should we remove an outlier from a dataset?

When the outlier is obviously due to a measurement mistake or sampling error.

6 The heights of students in a school netball team were measured and recorded as 166 cm, 170 cm, 168 cm, 67 cm, 170 cm, 169 cm.

a Calculate the mean, median and mode of this data.

Mean: 151.67 cm, median: 168.6 cm, mode: 170 cm

b Identify the outlier in this set of data.

67 cm

c Ignore the outlier and calculate the mean, median and mode of the remaining 5 scores. Comment on how the summary statistics have changed.

Mean: 168.6 cm, median: 169 cm, mode: 170 cm

The mean has changed a lot, the median less so, and the mode is identical

d Why should the outlier be removed from the dataset in this case?

67 cm is a recording error or measurement mistake, as it represents the height of a young child rather than a secondary school netball player.

7 One metal rod was measured by 6 students and the results were recorded as 112 mm, 111 mm, 110 mm, 13 mm, 112 mm, 112 mm.

a Calculate the mean, median and mode of this data.

Mean: 95 mm, median: 111.5 mm, mode: 112 mm

b Identify the outlier in this set of data.

13 mm

c Ignore the outlier and calculate the mean, median and mode of the remaining 5 scores. Should the outlier by included in reporting the mean, median and mode for this data? Why or why not?

The outlier should not be included. The value 13 mm is a measurement error (likely 113 mm misrecorded). The true length of the rod is approximately 111-112 mm, so including the error distorts all three measures, with the mean affected most significantly.

8 The times (in minutes) taken to travel to work in a 5-day week were recorded as

17, 15, 16, 18, 55

a Calculate the mean, median and mode of this data.

Mean: 24.2, median: 17, mode: none

b Identify the outlier in this set of data.

55 minutes

c 55 minutes was one of the actual times taken to get to work. Explain why the outlier shouldn't be removed from the dataset in this case.

The value 55 minutes represents a genuine travel time, not a recording error.
Removing it would misrepresent the actual variation in commute times experienced during the week.

9 The following shows the total number of red cards of 20 professional footballers.

Stem	Leaf										
1	4	4	4	5	6	7	7	7	8	8	8	9
2	0	0	1	1	2	7	7					
3												
4	6											

Key

$2 \mid 0 = 20$ red cards

a Determine the value of the outlier in the dataset.

46

b What are the values of the three modes in the distribution?

14, 17, and 18

c Calculate the mean, median, and range of the data, including the outlier.

Mean: 19.1, median: 18, range: 32

d Calculate the mean, median, and range of the data, not including the outlier. Compare your answer to part **c** and comment on any changes in the statistics.

Mean: 17.6, median: 18, range: 13

The outlier significantly affects the mean and range.

10

a In a class of five students, the mean of four students in a test is 90.
The 5th student scores 2. Calculate the new mean.

72.4

b In a class of thirty students, the mean of twenty-nine students in a test is 90.
The 30th student scores 2. Calculate the new mean.

87.06

c By considering parts **a** and **b**, explain why scientists always want to repeat experiments.

A single outlier affects the mean less when the sample size is larger.
Repeating experiments increases sample size and reduces the impact of outliers, giving more reliable conclusions.

11

A class of 30 students guess how many sweets there are in a jar.
The mean guess is 648. One student's guess was 9000. If this outlier is removed:

a Find the new mean of the guesses.

360

b Work out the percentage change in the mean.

-44.4%

12 State how the summary statistics change using “increase, decrease, no change”

a Original dataset

New dataset

Impact on summary statistics

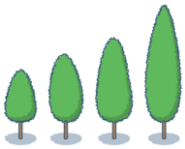
Mean: ..... increase

Median: ..... increase

Range: ..... increase

b Original dataset

New dataset



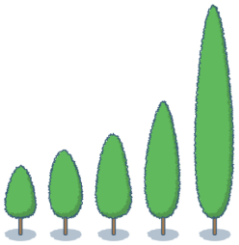
Impact on summary statistics

Mean: ..... increase

Median: ..... decrease

Range: ..... increase

c Original dataset



New dataset

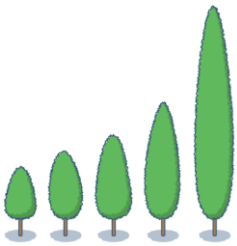
Impact on summary statistics

Mean: ..... increase

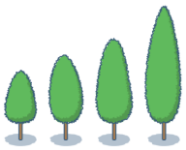
Median: ..... increase

Range: ..... decrease

d Original dataset



New dataset



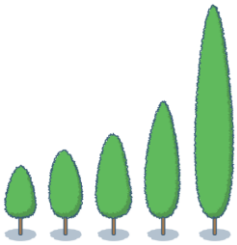
Impact on summary statistics

Mean: ..... increase

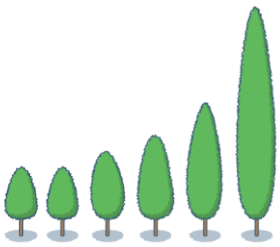
Median: ..... increase

Range: ..... decrease

e Original dataset



New dataset



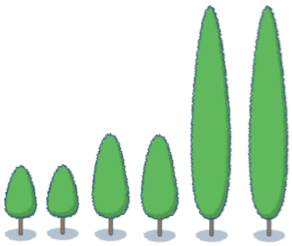
Impact on summary statistics

Mean: ..... decrease

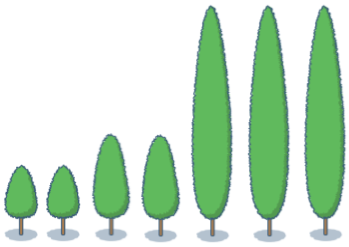
Median: ..... decrease

Range: ..... No change

f Original dataset



New dataset



Impact on summary statistics

Mean: ..... increase

Median: ..... No change

Range: ..... no change

13 Complete the table. Write:

"**will** increase/decrease" "**will** stay the same" "**may** increase/decrease"

	Effect on Range
Add a new data value larger than current maximum	Will increase
Add a new data value smaller than current minimum	Will increase
Remove all the maximum data value(s)	Will decrease
Remove all the minimum data value(s)	Will decrease
Add a data value equal to the current maximum	Will stay the same
Add a data value equal to the current minimum	Will stay the same
Remove a data value equal to the current maximum	May decrease
Remove a data value equal to the current minimum	May decrease

	Effect on Mean
Add a data value greater than the current mean	Will increase
Add a data value less than the current mean	Will decrease
Remove a data value greater than the current mean	Will decrease
Remove a data value less than the current mean	Will increase
Add a data value equal to the current mean	Will stay the same
Remove a data value equal to the current mean	Will stay the same

	Effect on Median
Add a data value below the current median	May decrease
Add a data value above the current median	May increase
Remove a data value below the current median	May increase
Remove a data value above the current median	May decrease
Add a data value above the median and a data value below the median.	Will stay the same

	Effect on Mode
Add a data value so that it appears more frequently than the current mode	Will change
Remove a data value that is the current mode	May change
Add a data value equal to the current mode	Will stay the same

RELATIVE MERITS OF MEAN, MEDIAN, MODE

	Advantages	Disadvantages
Mean	Value depends on the value of every score in the dataset.	Distorted by outliers. Not suitable for categorical data.
Median	Not as distorted by outliers.	May not be central value, e.g., 0 1 1 **5** 5 5 5 Not suitable for categorical data.
Mode	Not affected by outliers. Suitable for categorical data.	May not be a central value.

- For measures of spread, we have only learned the range. The range is affected by outliers. We will learn more measures of range in Stage 5.

Check your understanding

1 Consider the dataset.

$$-3, -1, 2, 3, 3, 5, 6, 7, 7, 7, 8, 9, 10, 11, 12, 46$$

Why might the range **not** be a good measure of spread?

a Because there are negative numbers.

b Because some of the numbers are repeated.

c Because there are an even number of data values.

d Because 46 would make the range very large.

2 The mean of each dataset is shown below.

For which dataset is the mean not a good measure of centre?

a 3, 4, 5, 23, 24, 25 $\bar{x} = 14$

b 3, 3, 3, 25, 25, 25 $\bar{x} = 14$

c 3, 4, 5, 6, 7, 25 $\bar{x} = 8.\dot{3}$

d $-3, -4, -5, -23, -24, -25$ $\bar{x} = -14$

3 Which of the following is a limitation of the median as a measure of centre?

a It does not take into account the precise value of each piece of data and so does not use all the information available in the data.

b It is distorted by outliers.

c It cannot be used with data sets that contain negative numbers.

d It cannot be used with datasets that have an even number of values.

4 For which dataset is the mode not a good average to use?

a 8, 9, 10, 10, 11, 12

b 4, 1, 2, 7, 4, 6

c 1, 1, 7, 8, 9

d $-2, -2, -2, -5, 1$

1 Find the median of these scores: 2, 3, 3, 3, 5, 5, 5, 5, 5.

a Calculate the mode. Why is the mode not a good measure of centre for this dataset?

5. The mode is the largest data value, it doesn't represent the centre.

b Calculate the median. Why is the median not a good measure of centre for this dataset?

5. The median is the largest data value, it doesn't represent the centre.

c Calculate the mean. Is the mean a good measure of centre for these scores?

4. Yes

2 Consider these two datasets.

Dataset 1: 8, 8, 9, 10, 11 | Dataset 2: 8, 8, 9, 10, 110

a Identify the mode of each dataset. Why is the mode not a suitable measure of centre?

Mode is 8 for both datasets, not suitable as the mode is a central value.

b Find the mean of each of the sets of scores.

Dataset 1: 9.2, Dataset 2: 29

c Explain why the mean is a good measure of centre for the first dataset, but not the second.

9.2 represents the central value of dataset 1, but not dataset 2 because the outlier has pulled the mean away from where most values lie.

d Would the median or mode be a better measure of centre for dataset 2?

The median would be better. Dataset 2 median = 9, which represents the centre of most values without being affected by the outlier 110.

3 Geoff scored these marks in a series of topic tests: 6 3 6 3 6 2 6 1

a Find the mean, mode and median for Geoff's marks.

Mean: 4.125, Mode: 6, Median: 4.5

b Which of these measures of centre would Geoff prefer to tell his parents?

The mode (6), as it makes his performance appear best.

4 The eye colours of a group of 10 students are recorded below.

brown blue brown green blue grey brown brown blue green

Which, if any, of the mean, mode and median can be used with this categorical data?

Only the mode can be used.

Mean and median require numerical data that can be ordered meaningfully, categorical data like colours cannot be averaged or placed in a meaningful numerical order.

5 Which is the only suitable measure of centre for categorical data?

Mode.

6 One day the number of minutes that trains were late to arrive at a station was recorded. The times are listed below.

0	7	0	0	1	2	5	0	0	0
6	0	1	52	0	10	1	1	8	22

a Calculate the mean, median, and mode of this data.

Mean: 5.9 minutes, median: 1 minute, mode: 0 minutes

b Explain which value would be the best to use to argue that the trains arrive late too often.

The mean (5.9 minutes) is best to argue trains are late too often, as it accounts for the severe delays which increase the average waiting time.

c Explain who might use the mode and why it might be an advantage to them.

The train operator might use the mode (0 minutes) to claim most trains arrive on time.

7 The prices of land sold by a real-estate agent in a particular suburb in one month were:

$510 000 $630 000 $960 000 $540 000 $560 000

a The real-estate agent claimed that, for this month, the 'average' price of houses in the suburb was $640 000. To which 'average' is the estate agent referring?

The mean

b Is the 'average' from part **a** an appropriate 'average' to use? Give reasons for your answer

No. The $960,000 property is an outlier that pulls the mean up to $640,000, whilst four of the five properties sold for $560,000 or less. The median ($560,000) better represents the typical house price.

8 A shoe store had a discount sale of women's running shoes and sold shoes of the following sizes in one day.

6 10 4 7 8 7 6 5 7 8 7 5 6 4 3 7

a Find the mean, mode, and median.

Mean: 6.25, mode: 7, median: 7

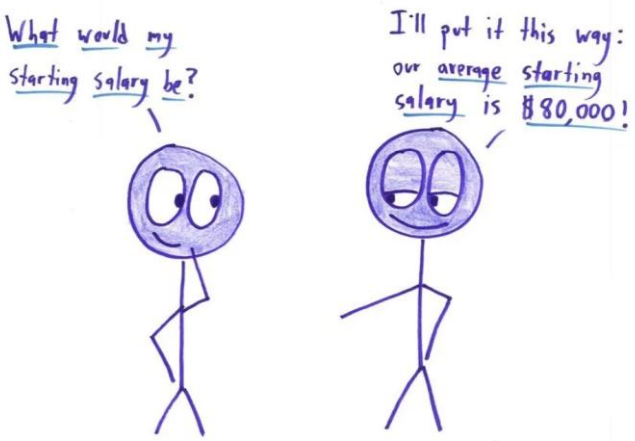
b Explain why the mean is not useful for the shoe store owner.

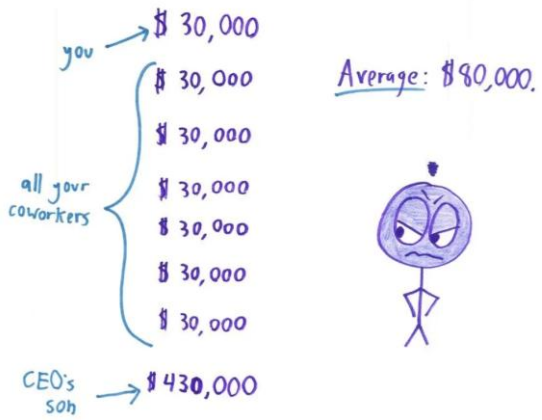
Size 6.25 does not exist.

c Which measure of centre would be most useful? Why?

The mode is most useful as it tells the owner which size sold most frequently, helping them decide which sizes to stock more of in future.

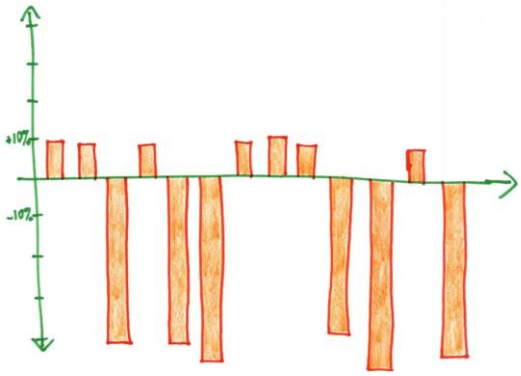
9 For each of these scenarios, explain why the summary statistic is misleading.



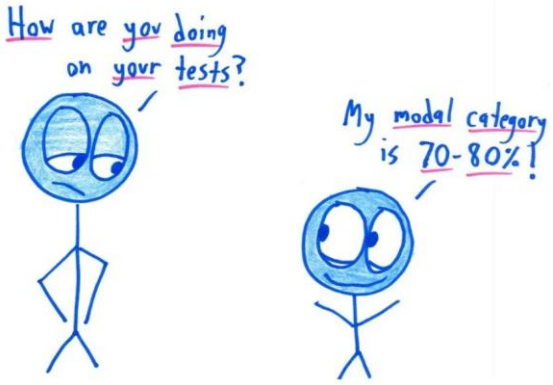


The colleague uses mean (average) of $80 000, but the CEO's son earns $430,000 which skews the data.

Most starting salaries are $30 000.

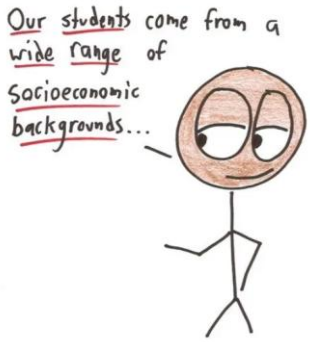


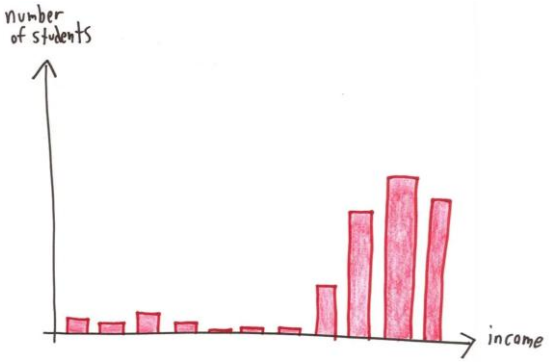
The median does not account for the size of the gains and losses.



Score Category	Number of Tests
90s	0
80s	0
70s	2
60s	1
50s	1
40s	1
30s	1
20s	1

The mode does not represent the centre of the data.





The range does not tell you the actual distribution of the students.

Most students come from high income backgrounds.