

BASIC DATA ANALYSIS AND VISUALISATION



CENTRAL TENDENCY MEASURES

An average is a single 'typical' value that is used to represent a set of values.

There are three main types of average.

They are called the **Mean, Median** and **Mode**.

MEAN

- The mean is the most used average.
- To find the mean of a list of numbers, add them all together and divide by how many numbers there are:

$$\text{Mean} = \frac{\text{Sum of Numbers}}{\text{Amount of Numbers}}$$

MEAN

- Find the mean of 1.46, 1.70, 1.58, 1.73, 1.23

$$\text{Mean} = \frac{1.46 + 1.70 + 1.58 + 1.73 + 1.23}{5} = 1.54$$

MEDIAN

- The median is the middle number in a set of data, when the data has been written in ascending or descending size order.
- If there is an odd number of items of data, there will be one number in the middle.
- If there is an even number of items of data, there will be two numbers in the middle. The median is the number that is halfway between these two numbers.

MEDIAN

- Find the median of 28, 10, 11, 37, 11, 19, 16, 13, 14

Sort them into ascending order:

10, 11, 11, 13, 14, 16, 19, 28, 37

Cross off the lowest and highest value repeatedly in pairs until the middle number is left.

~~10, 11, 11, 13, 14, 16, 19, 28, 37~~

The median is 14.

MEDIAN

- Find the median of 31, 28, 19, 14, 11, 30, 27, 20

Sort them into ascending order:

11, 14, 19, 20, 27, 28, 30, 31

Cross off the lowest and highest value repeatedly in pairs until the middle number is left.

~~11~~, ~~14~~, ~~19~~, **20**, **27**, ~~28~~, ~~30~~, ~~31~~

The median is the number in the middle of these two values

$$\text{Median} = \frac{20 + 27}{2} = 23.5$$

MODE

- The mode is the number, or item, which occurs most often in a set of data.

MODE

- Find the mode of 5, 3, 4, 7, 7, 5, 6, 5, 4

Sort them into ascending order:

3, 4, 4, **5, 5, 5**, 6, 7, 7

The most common number, or mode is 5.

It is possible for a set of values to have no mode if each value is listed the same number of times. It is also possible to have more than one mode, if there is more than one value that appears most often.

DISPERSION MEASURES

Dispersion is a measure that shows how much the values in a data set differ from each other or from the average.

It helps us understand how spread out or close together the data values are.

There are three main types of dispersion.

They are called the **Range**, **Variance**, and **Standard Deviation**.

RANGE

- In statistics, a range shows how spread out a set of data is.
- The bigger the range, the more spread out the data.
- If the range is small, the data is closer together or more consistent

RANGE

- Find the range of 52, 16, 22, 23, 57, 17, 36

Sort them into ascending order:

16, 17, 22, 23, 36, 52, 57

To find the range, subtract the lowest value from the highest value.

$$\text{Range} = 57 - 16 = 41$$

VARIANCE

- For example, the following two data sets are significantly different in nature and yet have the same mean, median and range.
- Some sort of numerical measure which distinguishes between them would be useful.

1, 7, 12, 15, 20, 22, 28

1, 15, 15, 15, 15, 16, 28

VARIANCE

- Shows how far each value is from the mean on average, by measuring the average of the squared differences from the mean.
- The bigger the variance, the more spread out the data.
- If the variance is small, the data is closer together or more consistent

VARIANCE

- Variance of a population is denoted by the Greek letter sigma squared and is written as:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{n}$$

Σ is the summation function

x is an observation in the dataset

μ is the population mean

n is the population size

VARIANCE

- Find the variance of the following population: 4, 7, 9, 11, 13, 15, 18

First calculate the mean

$$\text{Mean} = \mu = \frac{4 + 7 + 9 + 11 + 13 + 15 + 18}{7} = 11$$

VARIANCE

- Find variance of the following: 4, 7, 9, 11, 13, 15, 18

x	$x-\bar{x}$	$(x-\bar{x})^2$
4		
7		
9		
11		
13		
15		
18		

$$\sigma^2 = \frac{138}{7}$$

$$\sigma^2 = 19.7$$

VARIANCE

- Find variance of the following: 4, 7, 9, 11, 13, 15, 18

x	$x - \bar{x}$	$(x - \bar{x})^2$
4	-7	
7	-4	
9	-2	
11	0	
13	2	
15	4	
18	7	

$$\sigma^2 = \frac{138}{7}$$

$$\sigma^2 = 19.7$$

VARIANCE

- Find variance of the following: 4, 7, 9, 11, 13, 15, 18

x	$x - \bar{x}$	$(x - \bar{x})^2$
4	-7	49
7	-4	16
9	-2	4
11	0	0
13	2	4
15	4	16
18	7	49
		138

$$\sigma^2 = \frac{138}{7}$$

$$\sigma^2 = 19.7$$

STANDARD DEVIATION

- Standard deviation is the square root of the variance.
- It is in the same units as the data, which makes it easier to interpret than variance.

STANDARD DEVIATION

- Standard deviation of a population is denoted by the Greek letter sigma and is written as:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

Σ is the summation function

x is an observation in the dataset

μ is the population mean

n is the population size

STANDARD DEVIATION

- Find the standard deviation of the following population: 4, 7, 9, 11, 13, 15, 18
- First calculate the mean

$$\text{Mean} = \mu = \frac{4 + 7 + 9 + 11 + 13 + 15 + 18}{7} = 11$$

STANDARD DEVIATION

- Find the standard deviation of the following: 4, 7, 9, 11, 13, 15, 18

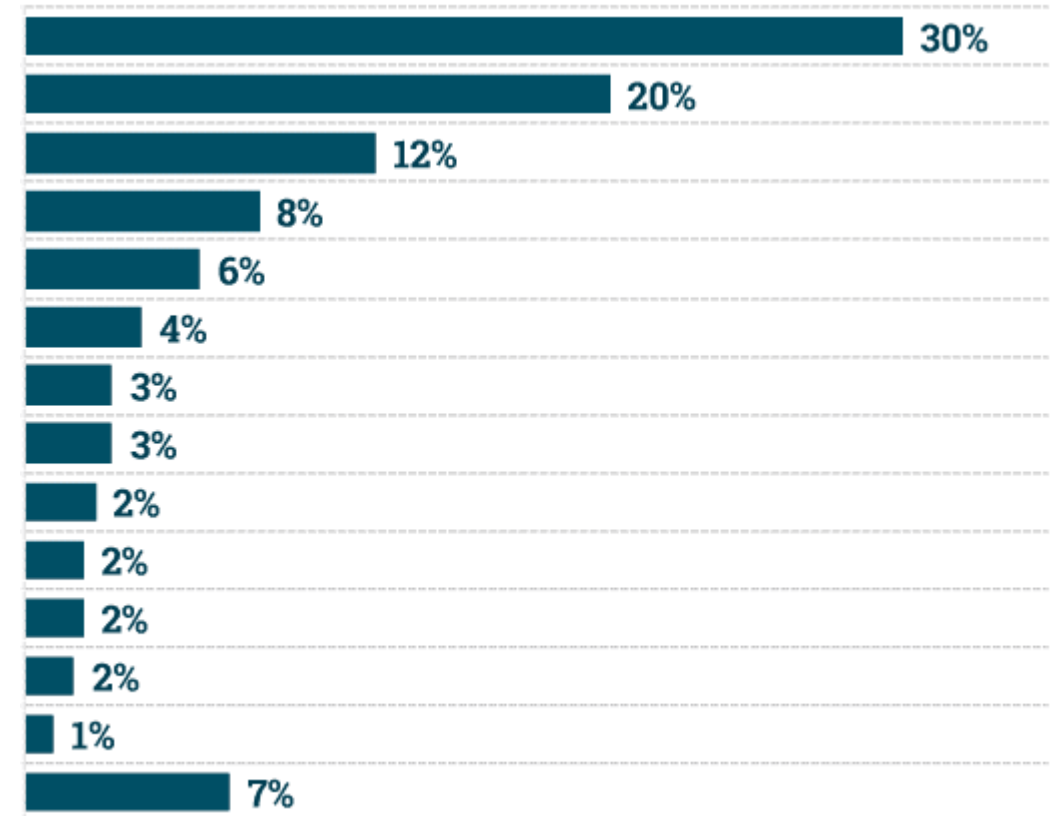
x	$x - \bar{x}$	$(x - \bar{x})^2$
4	-7	49
7	-4	16
9	-2	4
11	0	0
13	2	4
15	4	16
18	7	49
		138

$$\sigma = \sqrt{\frac{138}{7}}$$

$$\sigma = \sqrt{\frac{138}{7}} = 4.440$$

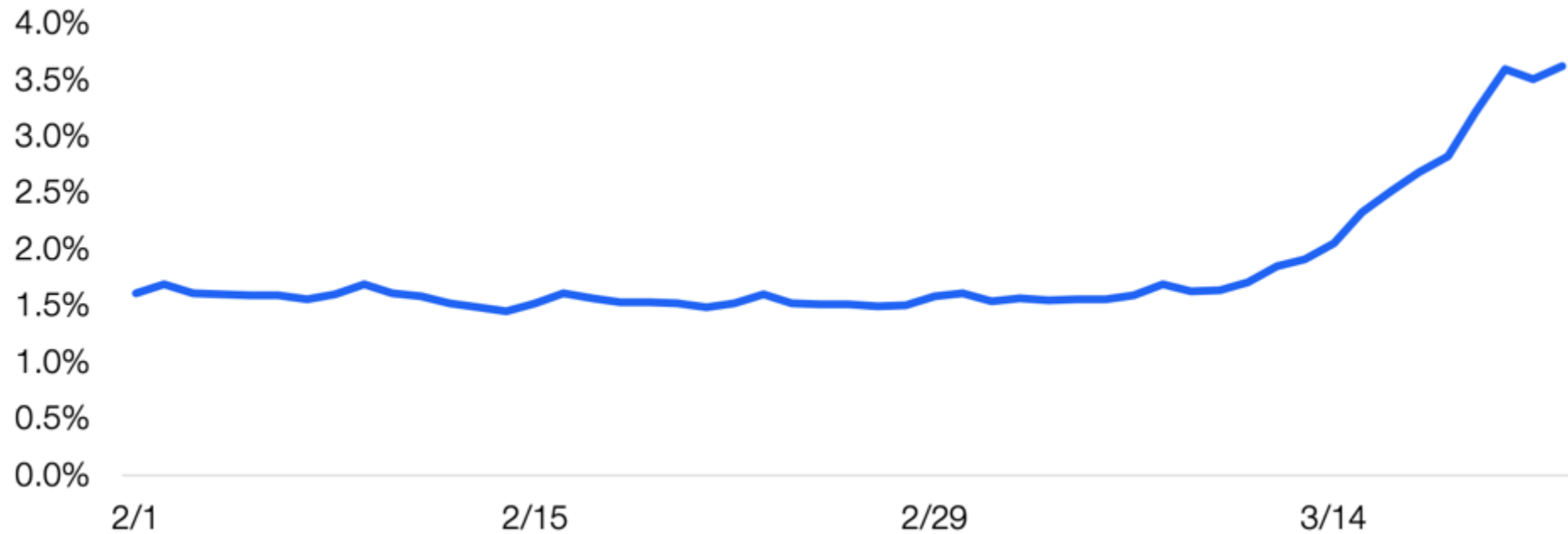
BAR CHART

- A bar chart is a visual representation of data where rectangular bars are used to show the size or frequency of categories.
- Each bar's length or height is proportional to the value it represents, allowing for easy comparison across categories.
- Bar charts can be oriented horizontally or vertically and are useful for displaying data in discrete categories, such as survey responses, sales data by region, or monthly expenses.
- They're particularly helpful when comparing quantities across different groups.



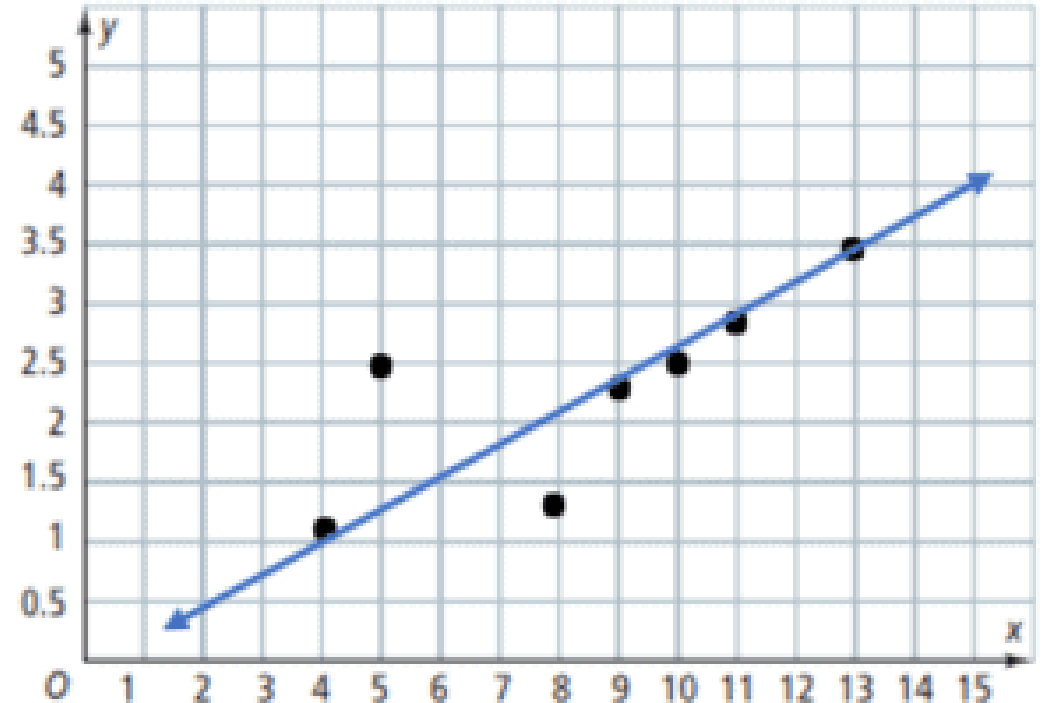
LINE CHART

- You can use a line chart to plot continuous data or data with infinite values.



SCATTER PLOT

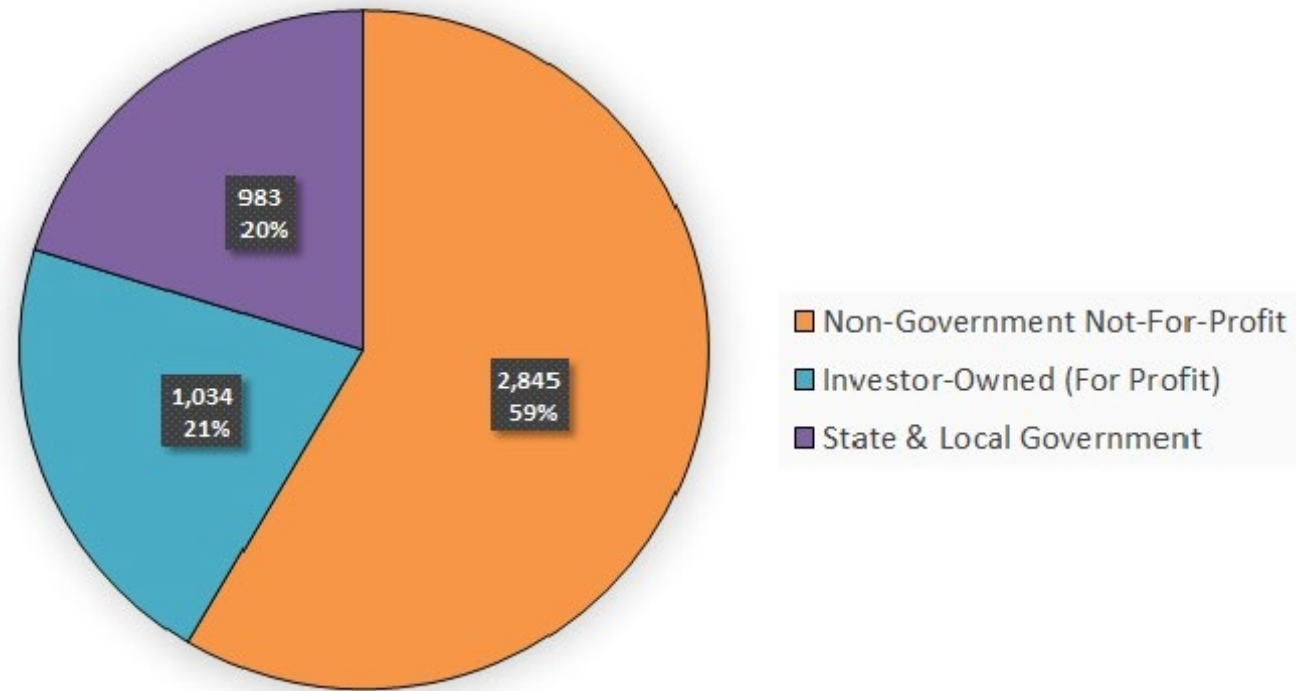
- A scatter plot or a scatter chart helps show the relationship between items based on two different variables and data sets.
- Dots (or plot data) are plotted in an x-y coordinate system.
- In some scatter plots, a trend line is added to a scatter plot



PIE CHART

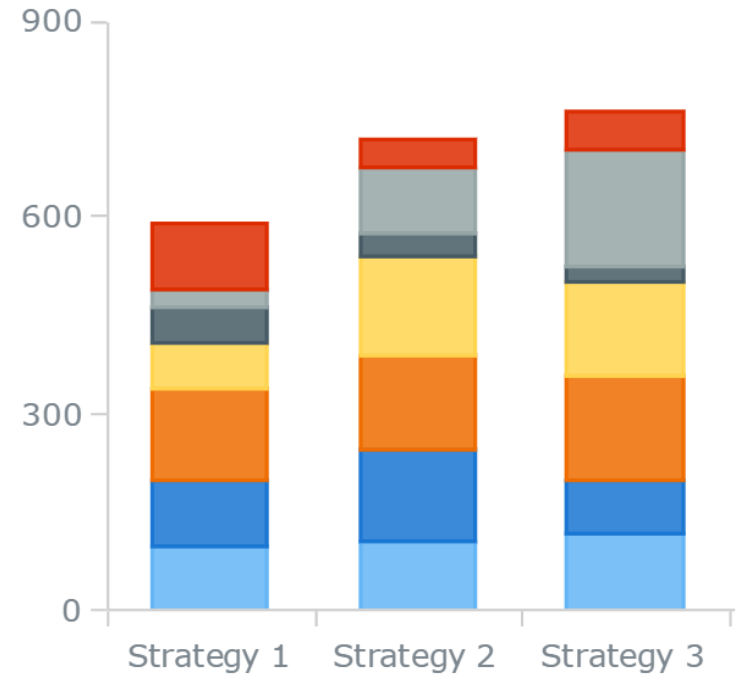
- Pie charts highlight data and statistics in pie-slice format.
- A pie chart represents numbers in percentages, and the total sum of all pies should equal 100 percent.
- When considering charts and graphs to employ to visualize data, pie charts are most impactful to your audience if you have a small data set.

Community Hospitals by Ownership Type (Total 4,862)



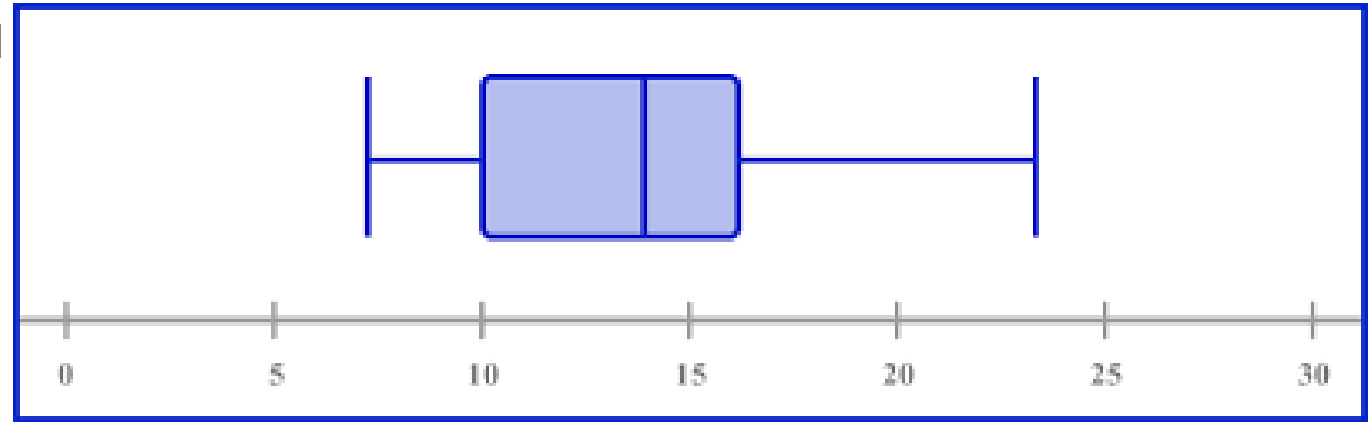
STACKED BAR CHART

- A stacked bar graph breaks down and compares parts of a whole.
- Each bar represents a total, and segments in the bar represent different categories or parts of that total.



BOX PLOT

- A box plot, also known as a whisker plot, displays a summary of the range and statistical distribution of a dataset based on a five-number summary: minimum, first quartile, median, third quartile, and maximum.



PRINCIPLES OF EFFECTIVE DATA VISUALISATION

CHOOSING THE RIGHT CHART OR GRAPH



Bar Chart Usage

Bar charts are ideal for comparing categorical data and showing differences among groups clearly.

Line Graph Application

Line graphs effectively display trends over time, showing continuous data relationships.

Pie Chart Purpose

Pie charts represent parts of a whole, illustrating proportions within data sets visually.

Scatter Plot Function

Scatter plots reveal correlations between variables and distribution patterns in data.



BEST PRACTICES FOR CLEAR VISUALS

Simplicity in Design

Using simple design elements helps viewers quickly grasp the main message without unnecessary complexity.

Proper Labelling

Clear and accurate labels guide the audience to understand data and visuals easily and effectively.

Consistent Scales

Maintaining consistent scales in visuals ensures accurate comparisons and prevents misinterpretation of data.

Thoughtful Use of Colour

Careful selection of colours enhances comprehension and highlights key information without causing distraction.