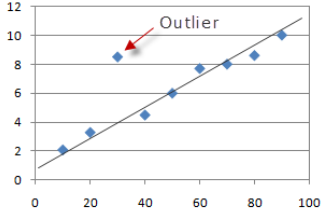
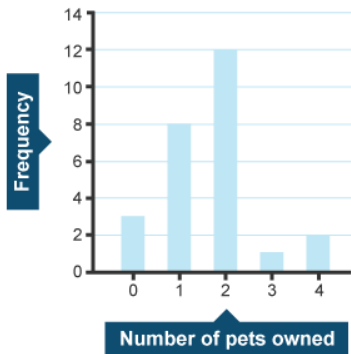
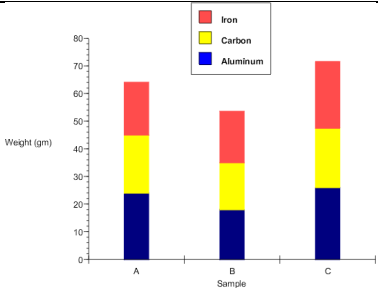
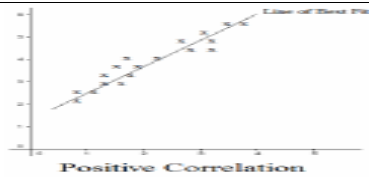
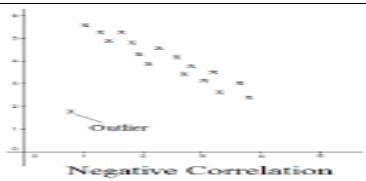
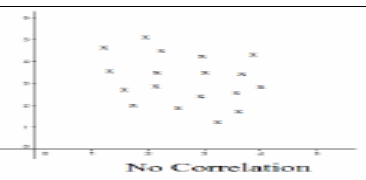
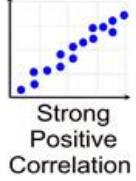
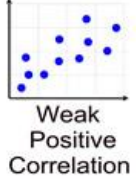
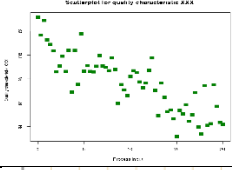
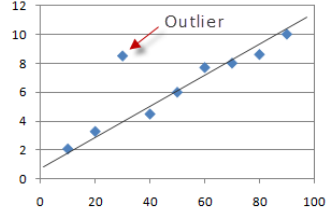


Topic: Data Handling

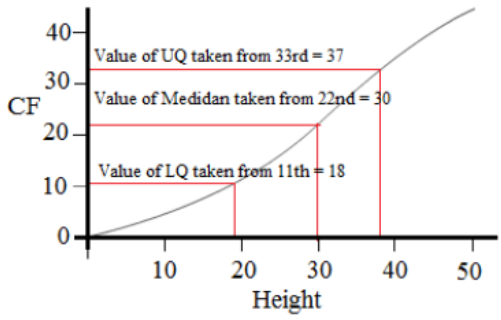
Topic/Skill	Definition/Tips	Example																				
Types of Data	Qualitative Data – non-numerical data Quantitative Data – numerical data Continuous Data – data that can take any numerical value within a given range. Discrete Data – data that can take only specific values within a given range.	Qualitative Data – eye colour, gender etc. Continuous Data – weight, voltage etc. Discrete Data – number of children, shoe size etc.																				
Grouped Data	Data that has been bundled in to categories . Seen in grouped frequency tables, histograms, cumulative frequency etc.	<table><tr><th>Foot length, l, (cm)</th><th>Number of children</th></tr><tr><td>$10 \leq l < 12$</td><td>5</td></tr><tr><td>$12 \leq l < 17$</td><td>53</td></tr></table>	Foot length, l , (cm)	Number of children	$10 \leq l < 12$	5	$12 \leq l < 17$	53														
Foot length, l , (cm)	Number of children																					
$10 \leq l < 12$	5																					
$12 \leq l < 17$	53																					
Primary/ Secondary Data	Primary Data – collected yourself for a specific purpose. Secondary Data – collected by someone else for another purpose.	Primary Data – data collected by a student for their own research project. Secondary Data – Census data used to analyse link between education and earnings.																				
Mean	Add up the values and divide by how many values there are.	The mean of 3, 4, 7, 6, 0, 4, 6 is $\frac{3 + 4 + 7 + 6 + 0 + 4 + 6}{7} = 5$																				
Mean from a Table	1. Find the midpoints (if necessary) 2. Multiply Frequency by values or midpoints 3. Add up these values 4. Divide this total by the Total Frequency If grouped data is used, the answer will be an estimate .	<table><tr><th>Height in cm</th><th>Frequency</th><th>Midpoint</th><th>F × M</th></tr><tr><td>$0 < h \leq 10$</td><td>8</td><td>5</td><td>$8 \times 5 = 40$</td></tr><tr><td>$10 < h \leq 30$</td><td>10</td><td>20</td><td>$10 \times 20 = 200$</td></tr><tr><td>$30 < h \leq 40$</td><td>6</td><td>35</td><td>$6 \times 35 = 210$</td></tr><tr><td>Total</td><td>24</td><td>Ignore!</td><td>450</td></tr></table> Estimated Mean height: $450 \div 24 = 18.75\text{cm}$	Height in cm	Frequency	Midpoint	F × M	$0 < h \leq 10$	8	5	$8 \times 5 = 40$	$10 < h \leq 30$	10	20	$10 \times 20 = 200$	$30 < h \leq 40$	6	35	$6 \times 35 = 210$	Total	24	Ignore!	450
Height in cm	Frequency	Midpoint	F × M																			
$0 < h \leq 10$	8	5	$8 \times 5 = 40$																			
$10 < h \leq 30$	10	20	$10 \times 20 = 200$																			
$30 < h \leq 40$	6	35	$6 \times 35 = 210$																			
Total	24	Ignore!	450																			
Median Value	The middle value. Put the data in order and find the middle one. If there are two middle values , find the number half way between them by adding them together and dividing by 2 .	Find the median of: 4, 5, 2, 3, 6, 7, 6 Ordered: 2, 3, 4, 5 , 6, 6, 7 Median = 5																				
Median from a Table	Use the formula $\frac{(n+1)}{2}$ to find the position of the median. n is the total frequency.	If the total frequency is 15, the median will be the $\left(\frac{15+1}{2}\right) = 8\text{th}$ position																				
Mode/Modal Value	Most frequent/common. Can have more than one mode (called bi-modal or multi-modal) or no mode (if all values appear once)	Find the mode: 4, 5, 2, 3, 6, 4, 7, 8, 4 Mode = 4																				
Range	Highest value subtract the Smallest value Range is a ‘measure of spread’. The smaller the range the more consistent the data.	Find the range: 3, 31, 26, 102, 37, 97. Range = $102 - 3 = 99$																				

Outlier	A value that 'lies outside' most of the other values in a set of data. An outlier is much smaller or much larger than the other values in a set of data.																						
Lower Quartile	Divides the bottom half of the data into two halves. $LQ = Q_1 = \frac{(n+1)}{4} \text{th value}$	Find the lower quartile of: 2, <u>3</u>, 4, 5, 6, 6, 7 $Q_1 = \frac{(7+1)}{4} = 2\text{nd value} \rightarrow 3$																					
Upper Quartile	Divides the top half of the data into two halves. $UQ = Q_3 = \frac{3(n+1)}{4} \text{th value}$	Find the upper quartile of: 2, 3, 4, 5, 6, <u>6</u>, 7 $Q_3 = \frac{3(7+1)}{4} = 6\text{th value} \rightarrow 6$																					
Interquartile Range	The difference between the upper quartile and lower quartile. $IQR = Q_3 - Q_1$ The smaller the interquartile range, the more consistent the data.	Find the IQR of: 2, 3, 4, 5, 6, 6, 7 $IQR = Q_3 - Q_1 = 6 - 3 = 3$																					
Frequency Table	A record of how often each value in a set of data occurs .	<table border="1"> <thead> <tr> <th>Number of marks</th><th>Tally marks</th><th>Frequency</th></tr> </thead> <tbody> <tr> <td>1</td><td> </td><td>7</td></tr> <tr> <td>2</td><td> </td><td>5</td></tr> <tr> <td>3</td><td> </td><td>6</td></tr> <tr> <td>4</td><td> </td><td>5</td></tr> <tr> <td>5</td><td> </td><td>3</td></tr> <tr> <td>Total</td><td></td><td>26</td></tr> </tbody> </table>	Number of marks	Tally marks	Frequency	1		7	2		5	3		6	4		5	5		3	Total		26
Number of marks	Tally marks	Frequency																					
1		7																					
2		5																					
3		6																					
4		5																					
5		3																					
Total		26																					
Bar Chart	Represents data as vertical blocks. x-axis shows the type of data y-axis shows the frequency for each type of data Each bar should be the same width There should be gaps between each bar Remember to label each axis.																						
Types of Bar Chart	Compound/Composite Bar Charts show data stacked on top of each other. Comparative/Dual Bar Charts show data side by side.																						

		Dual Bar Chart
Pie Chart	Used for showing how data breaks down into its constituent parts. When drawing a pie chart, divide 360° by the total frequency. This will tell you how many degrees to use for the frequency of each category. Remember to label the category that each sector in the pie chart represents.	If there are 40 people in a survey, then each person will be worth $360 \div 40 = 9^\circ$ of the pie chart.
Pictogram	Uses pictures or symbols to show the value of the data. A pictogram must have a key.	Black Red Green Others = 4 cars
Line Graph	A graph that uses points connected by straight lines to show how data changes in values. This can be used for time series data, which is a series of data points spaced over uniform time intervals in time order.	
Box Plots	The minimum, lower quartile, median, upper quartile and maximum are shown on a box plot. A box plot can be drawn independently or from a cumulative frequency diagram.	Students sit a maths test. The highest score is 19, the lowest score is 8, the median is 14, the lower quartile is 10 and the upper quartile is 17. Draw a box plot to represent this information.
Comparing Box Plots	Write two sentences. 1. Compare the averages using the medians for two sets of data. 2. Compare the spread of the data using the range or IQR for two sets of data. The <u>smaller</u> the range/IQR, the <u>more consistent</u> the data.	‘On average, students in class A were more successful on the test than class B because their median score was higher.’ ‘Students in class B were more consistent than class A in their test scores as their IQR was smaller.’

	You must compare box plots in the context of the problem .	
Correlation	Correlation between two sets of data means they are connected in some way.	There is correlation between temperature and the number of ice creams sold.
Causality	When one variable influences another variable.	The more hours you work at a particular job (paid hourly), the higher your income from that <u>job</u> will be.
Positive Correlation	As one value increases the other value increases .	 A scatter plot with a positive linear trend. The x-axis is labeled from 0 to 6, and the y-axis from 0 to 40. A line of best fit is drawn through the data points, starting at approximately (0, 10) and ending at (6, 35). The text 'Line of Best Fit' is written at the end of the line. Below the plot, the text 'Positive Correlation' is written.
Negative Correlation	As one value increases the other value decreases .	 A scatter plot showing a negative linear trend. The x-axis is labeled from 0 to 6, and the y-axis from 0 to 40. A line of best fit is drawn through the data points, starting at approximately (0, 35) and ending at (6, 10). One data point at approximately (1, 10) is labeled 'Outlier' with an arrow. Below the plot, the text 'Negative Correlation' is written.
No Correlation	There is no linear relationship between the two.	 A scatter plot with data points scattered randomly across the grid. The x-axis is labeled from 0 to 6, and the y-axis from 0 to 40. Below the plot, the text 'No Correlation' is written.
Strong Correlation	When two sets of data are closely linked .	 A scatter plot with data points tightly clustered along a diagonal line from bottom-left to top-right. Below the plot, the text 'Strong Positive Correlation' is written.
Weak Correlation	When two sets of data have correlation, but are not closely linked .	 A scatter plot with data points loosely clustered along a diagonal line from bottom-left to top-right. Below the plot, the text 'Weak Positive Correlation' is written.
Scatter Graph	A graph in which values of two variables are plotted along two axes to compare them and see if there is any connection between them.	 A scatter plot with green data points. The x-axis is labeled 'Factor A' and the y-axis is labeled 'Factor B'. The title 'Scatterplot for quality characteristic A&B' is at the top. Below the plot, the text 'Scatter Graph' is written.
Line of Best Fit	A straight line that best represents the data on a scatter graph.	 A scatter plot with 'x' marks for data points. A red line of best fit is drawn through the points, showing a positive correlation. Below the plot, the text 'Line of Best Fit' is written.
Outlier	A value that 'lies outside' most of the other values in a set of data. An outlier is much smaller or much larger than the other values in a set of data.	 A scatter plot with blue data points and a line of best fit. One point at approximately (25, 9) is significantly above the line and is labeled 'Outlier' with a red arrow. The x-axis is labeled from 0 to 100, and the y-axis from 0 to 12. Below the plot, the text 'Outlier' is written.

Histograms	A visual way to display frequency data using bars. Bars can be unequal in width. Histograms show frequency density on the y-axis, not frequency. $\text{Frequency Density} = \frac{\text{Frequency}}{\text{Class Width}}$ <table><tr><th>Height(cm)</th><th>Frequency</th></tr><tr><td>$0 < h \leq 10$</td><td>8</td></tr><tr><td>$10 < h \leq 30$</td><td>6</td></tr><tr><td>$30 < h \leq 45$</td><td>15</td></tr><tr><td>$45 < h \leq 70$</td><td>5</td></tr></table>	Height(cm)	Frequency	$0 < h \leq 10$	8	$10 < h \leq 30$	6	$30 < h \leq 45$	15	$45 < h \leq 70$	5	<table><tr><th>Frequency Density (FD)</th></tr><tr><td>$8 \div 5 = 1.6$</td></tr><tr><td>$6 \div 20 = 0.3$</td></tr><tr><td>$15 \div 15 = 1$</td></tr><tr><td>$5 \div 25 = 0.2$</td></tr></table>	Frequency Density (FD)	$8 \div 5 = 1.6$	$6 \div 20 = 0.3$	$15 \div 15 = 1$	$5 \div 25 = 0.2$
Height(cm)	Frequency																
$0 < h \leq 10$	8																
$10 < h \leq 30$	6																
$30 < h \leq 45$	15																
$45 < h \leq 70$	5																
Frequency Density (FD)																	
$8 \div 5 = 1.6$																	
$6 \div 20 = 0.3$																	
$15 \div 15 = 1$																	
$5 \div 25 = 0.2$																	
Interpreting Histograms	The area of the bar is proportional to the frequency of that class interval. $\text{Frequency} = \text{Freq Density} \times \text{Class Width}$	A histogram shows information about the heights of a number of plants. 4 plants were less than 5cm tall. Find the number of plants more than 5cm tall.															
Cumulative Frequency	Cumulative Frequency is a running total. <table><tr><th>Age</th><th>Frequency</th></tr><tr><td>$0 < a \leq 10$</td><td>15</td></tr><tr><td>$10 < a \leq 40$</td><td>35</td></tr><tr><td>$40 < a \leq 50$</td><td>10</td></tr></table>	Age	Frequency	$0 < a \leq 10$	15	$10 < a \leq 40$	35	$40 < a \leq 50$	10	<table><tr><th>Cumulative Frequency</th></tr><tr><td>15</td></tr><tr><td>$15 + 35 = 50$</td></tr><tr><td>$50 + 10 = 60$</td></tr></table>	Cumulative Frequency	15	$15 + 35 = 50$	$50 + 10 = 60$			
Age	Frequency																
$0 < a \leq 10$	15																
$10 < a \leq 40$	35																
$40 < a \leq 50$	10																
Cumulative Frequency																	
15																	
$15 + 35 = 50$																	
$50 + 10 = 60$																	
Cumulative Frequency Diagram	A cumulative frequency diagram is a curve that goes up. It looks a little like a stretched-out S shape. Plot the cumulative frequencies at the end-point of each interval.																

Quartiles from Cumulative Frequency Diagram	Lower Quartile (Q1): 25% of the data is less than the lower quartile. Median (Q2): 50% of the data is less than the median. Upper Quartile (Q3): 75% of the data is less than the upper quartile. Interquartile Range (IQR): represents the middle 50% of the data.	 $IQR = 37 - 18 = 19$
Hypothesis	A statement that might be true, which can be tested.	Hypothesis: 'Large dogs are better at catching tennis balls than small dogs'. We can test this hypothesis by having hundreds of different sized dogs try to catch tennis balls.

Where else is the Data Handling topic used: Graphs Probability Ratio and proportion	Links to other subjects: Sciences (Physics, Biology and Chemistry) Humanities (Geography and History)
--	---