

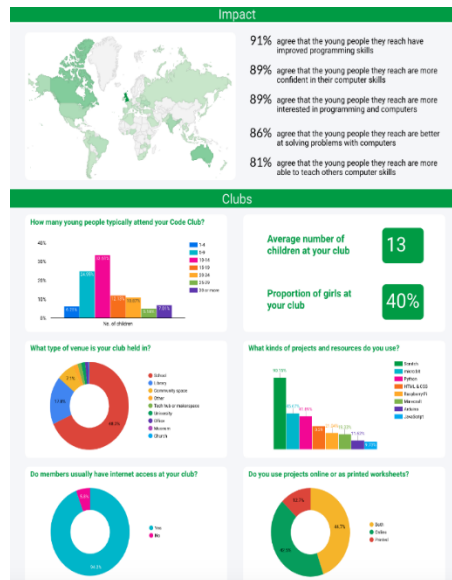
What is Data Science

Data science is extracting meaning from large data sets in order to gain insights to support decision-making

Infographics versus data visualisations

Data visualisations are visual representation of data (such as charts and graphs) intended to help an audience process the information more easily and get a clear idea about the data at a glance.

Infographics are visual representations of data, often involving pictures that reflect patterns and help tell a story. Infographics can include visualisations.



Statistics

A correlation shows that there is a relationship between two or more variables. For example data could indicate a clear upwards trend, showing that there is a relationship between the two **variables**; we call that a **correlation**. This would be an example of a **positive correlation**, meaning that as one variable increases, the other one increases too.

However correlation doesn't always mean causation

A **correlation** shows that there is a relationship between two or more variables, but that doesn't guarantee that one causes the other.

For example, there is likely to be a correlation between ice cream sales and the weather. Does that mean that ice cream sales cause hot weather? The correlation doesn't guarantee that one causes the other.

Key Vocabulary

Key Word	Definition
Analysis	A thorough study doing a careful analysis of a problem.
Causation	This is when one variable influences another.
Correlation	Shows that there is a relationship between two or more variables.
Data visualisation	Representation of data with charts and graphs to help the audience process the information easily.
Infographics	Visual representations of data, often involving pictures.
Outlier	A data point on a graph or in a set of results that is very much bigger or smaller than the next data point.
Positive correlation Variable	A relationship between two variables that tend to move in the same direction. Something that may or does vary or change.

Jon Snows Visualisation



 In 1854 there was an outbreak of cholera in the Soho area of London.

At the time it was widely believed that cholera was caused by pollution in the air.

John Snow's observation of the evidence led to him discounting this belief, but he could not prove how people did become infected.

John Snow made a dot map of Soho.

The dots (or shaded-in parts) on the map represent where a cholera-related death had occurred.

John Snow highlighted on the map the position of a water pump on Broad Street.

This data visualisation helped him to prove his theory that all the deaths had been of people who had used this water pump for drinking water.



What about anomalies?

Data that sits outside a trend is known an **outlier**.

Outliers can cause problems when working out statistics such as the mean, but they shouldn't be removed from the data set without investigating the reason for them.

Data Cleansing

Once data has been collected it should be checked through it to see if it needs **cleansing**. Cleansing involves **detecting** and **correcting**, or **removing, corrupt or inaccurate data**. There are several things that need to be checked when cleansing data, this will help to make sure the data set is accurate and robust.

Missing values

Is there some data missing on the data set, how would this affect the results of the research?

Duplicate entries

Duplicate data can happen when the same data has been entered twice in the dataset. This has often been entered mistakenly and will need to be removed or corrected during the data cleansing process.

Invalid data

This could be data outside the normal range that would be expected. For example if someone had their height recorded as over 3 meters tall then this would clearly be a mistake or inaccurate data. Also a persons age could be recorded as zero, this again isn't possible and would be inaccurate data and should be either corrected or removed.

Data Science History

The foundations of data science were laid by Joseph Minard.

Joseph Minard used these numbers in 1869 to find meaning and tell a story with the data.

The data you looked at before relates to Napoleon's march on Russia in 1812.

The numbers by themselves don't tell much of a story, but Joseph Minard created what is widely regarded to be the best statistical graph of all time.

