

## Lesson 1: Introduction to Data Science and Data Visualisation

---

### What is Data Science?

**Data science** is the process of collecting, analyzing, and interpreting large sets of data to help us make decisions, solve problems, and discover patterns. It's used in everything from business to sports, health care, and entertainment.

---

### Why is Visualising Data Important?

Raw data (numbers, tables, etc.) can be hard to understand. **Visualisation** turns that data into graphs, charts, or images that are easier to read.

This helps us to:

- ✓ Spot patterns and trends
- ✓ Make comparisons
- ✓ Discover insights we might otherwise miss

**Example:** A simple line graph of TV viewing figures can show which shows are most popular over time.

---

### Key Vocabulary

Word	Meaning
------	---------

<b>Data Science</b>	The study of data to find useful information
---------------------	--

<b>Visualisation</b>	Turning data into graphs/charts to make it easier to understand
----------------------	---

<b>Insight</b>	A new understanding or discovery from looking at the data
----------------	---

<b>Infographic</b>	A visual image (like a chart or diagram) used to represent data clearly
--------------------	---

---

### 9.3 Data Visualizations Revision Guide

## 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.

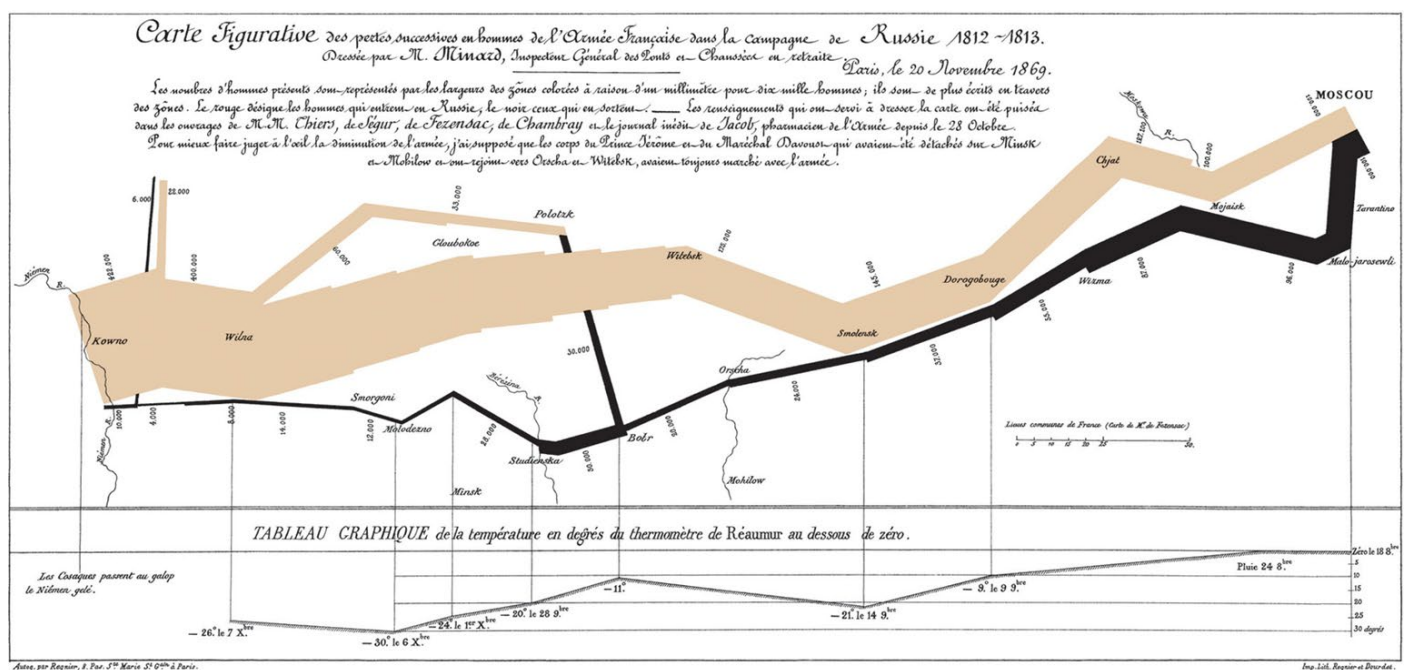
The numbers by themselves don't tell much of a story, but Joseph Minard created what is widely regarded to be the best statistical graph of all time.

```
basic
input,
(long latc city$ lont temp days date$ long latp surviv direc$ division),
(#4 #5 > #12 >> #5 #5 #5 #8 #6 #6 >> #6 >> #1 #3)
save minard
run
24.0 55.0 Kowno 37.6 0 6 Oct 18 24.0 54.9 340000 A 1
25.3 54.7 Wilna 36.0 0 6 Oct 24 24.5 55.0 340000 A 1
26.4 54.4 Smorgoni 33.2 -9 16 Nov 9 25.5 54.5 340000 A 1
26.8 54.3 Molodexno 32.0 -21 5 Nov 14 26.0 54.7 320000 A 1
27.7 55.2 Gloubockoe 29.2 -11 10 27.0 54.8 300000 A 1
27.6 53.9 Minsk 28.5 -20 4 Nov 28 28.0 54.9 280000 A 1
28.5 54.3 Studienska 27.2 -24 3 Dec 1 28.5 55.0 240000 A 1
28.7 55.5 Polotsk 26.7 -30 5 Dec 6 29.0 55.1 210000 A 1
29.2 54.4 Bobr 25.3 -26 1 Dec 7 30.0 55.2 180000 A 1
30.2 55.3 Witebak 30.3 55.3 175000 A 1
30.4 54.5 Orscha 32.0 54.8 145000 A 1
30.4 53.9 Mohilow 33.2 54.9 140000 A 1
32.0 54.8 Smolensk 34.4 55.5 127100 A 1
33.2 54.9 Dorogobouge 35.5 55.4 100000 A 1
34.3 55.2 Wixna 36.0 55.5 100000 A 1
34.4 55.5 Chjat 37.6 55.8 100000 R 1
36.0 55.5 Mojalak 37.5 55.7 98000 R 1
37.6 55.8 Moscou 37.0 55.0 97000 R 1
36.8 55.3 Tarantino 36.8 55.0 96000 R 1
36.5 55.0 Malo-jarosewli 35.4 55.3 87000 R 1
34.3 55.2 55000 R 1
33.3 54.8 37000 R 1
32.0 54.6 24000 R 1
30.4 54.4 20000 R 1
29.2 54.4 20000 R 1
28.5 54.3 20000 R 1
28.3 54.4 20000 R 1
24.0 55.1 60000 A 2
24.5 55.2 60000 A 2
25.5 54.7 60000 A 2
26.6 55.7 40000 A 2
27.4 55.6 33000 A 2
28.7 55.5 30000 R 2
29.2 54.3 30000 R 2
28.5 54.2 30000 R 2
28.3 54.3 28000 R 2
27.5 54.5 20000 R 2
26.8 54.3 12000 R 2
26.4 54.4 14000 R 2
24.6 54.5 8000 R 2
24.4 54.4 4000 R 2
24.2 54.4 4000 R 2
24.1 54.3 4000 R 2
24.0 55.2 22000 A 3
24.5 55.3 22000 A 3
24.6 55.8 6000 R 3
24.2 54.4 6000 R 3
24.1 54.3 6000 R 3
exit
```

The data above is about Napoleon's march on Russia in 1812.

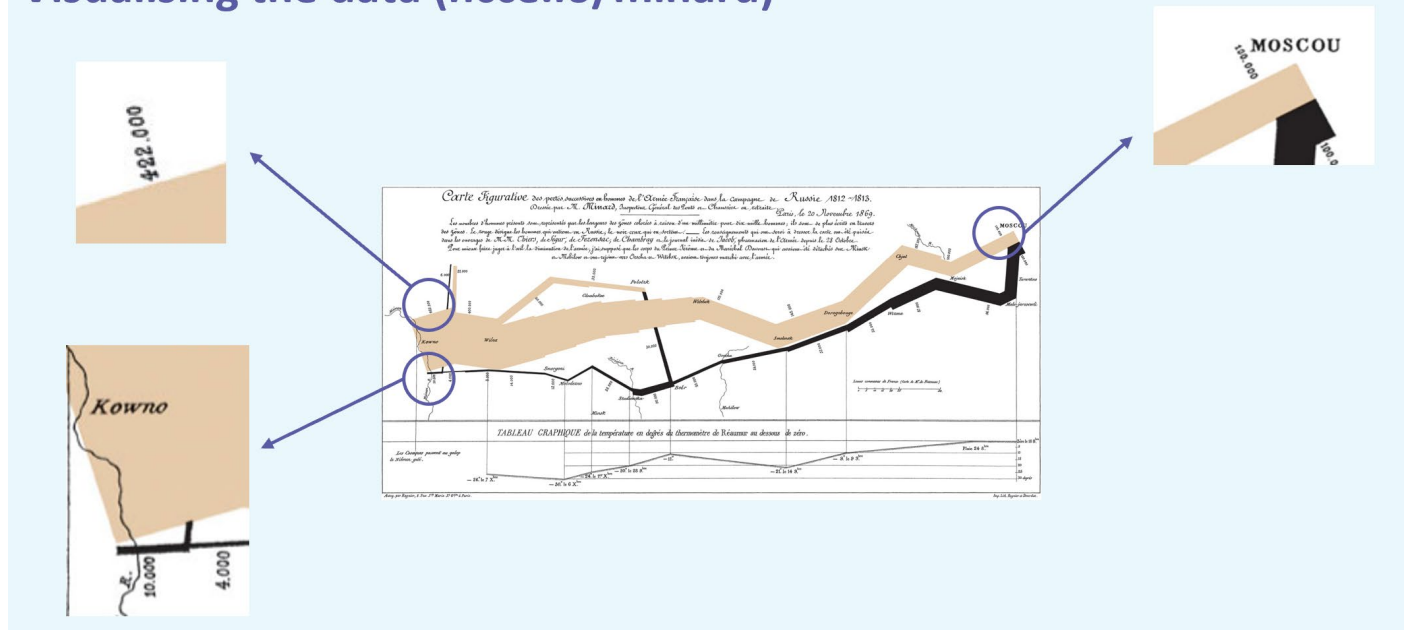
This is better visualized in the graphic below.

The width of the beige lines represents the number of soldiers marching towards Moscow, while the black line is the retreating soldiers. It is clear to see from the width of the line how the number of soldiers diminishes the further they march. Napoleon started with 422,000 troops; by the time he reached Moscow, 322,000 soldiers had died, and when he returned to France there were only 10,000 remaining.

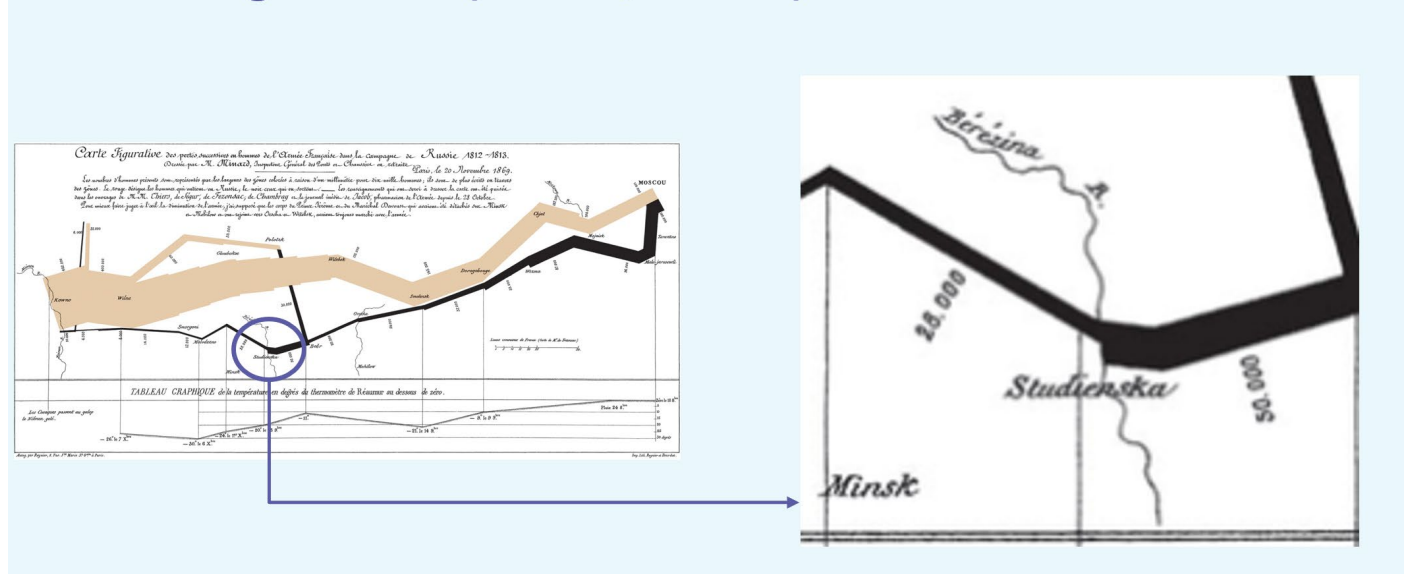


### 9.3 Data Visualizations Revision Guide

## Visualising the data (ncce.io/minard)



## Visualising the data (ncce.io/minard)



The visualisation also includes elements of geography, such as locations and the rivers the troops had to cross. On 28 September 1813, the data labels on the black line read 50,000, then 28,000. This shows that 22,000 men died crossing the Berezina river near Minsk.

Finally, the line chart at the bottom of the image represents the temperature. At some points on the journey, the temperature reaches -30 degrees Celsius.

This is simply a different representation of the numerical data but by visualising it in this way, we can draw more information from it.

## John Snow 's Data Visualisation

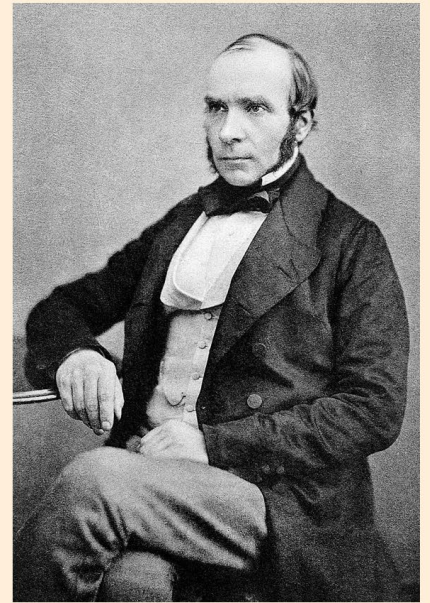
John Snow famously provided a visualisation to support his theory that cholera was not due to poor air quality, but was being transmitted orally.

### John Snow's 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.



### The Broad Street pump

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.

This map helped convince the local council to immediately remove the pump handle. Many lives were saved.



The visualisation alone doesn't provide all the information needed to draw the conclusion that the water pump was the source of the problem, but it does help us to identify patterns that can be investigated further and that might have been hard to identify without the dot map.

### Lesson 2: Global Data

---

#### **Big Data in the Real World**

Thanks to modern technology, we can now collect **huge amounts of data** from across the world — quickly and easily. This is known as **big data** and it helps governments, businesses, scientists, and even apps make smarter decisions.

---

#### **Real-Life Examples of Big Data**

You see big data in action every day!

- ✓ Weather forecasts
  - ✓ Google search trends
  - ✓ Social media analytics
  - ✓ Health data tracking (like COVID-19 stats)
  - ✓ Traffic & map services (like Google Maps)
- 

#### **Making & Testing Predictions**

In this lesson, you:

1. **Made a prediction** based on your knowledge
2. **Chose criteria** (e.g. location, time, age group) to help you test it
3. Used **data visualisations** to support or challenge your prediction

#### **Example:**

Prediction – “Countries with higher income have better internet access.”

→ Use global data to test it

→ Check for any **outliers** (countries that don’t follow the trend)

---

#### **Evaluating Findings**

After investigating your data:

- Ask yourself: Does the data support my prediction?
- Are there any **anomalies** or unusual results?
- Can I explain **why** some data points don’t fit the pattern?

This is how data science helps build strong arguments and evidence-based conclusions.

---

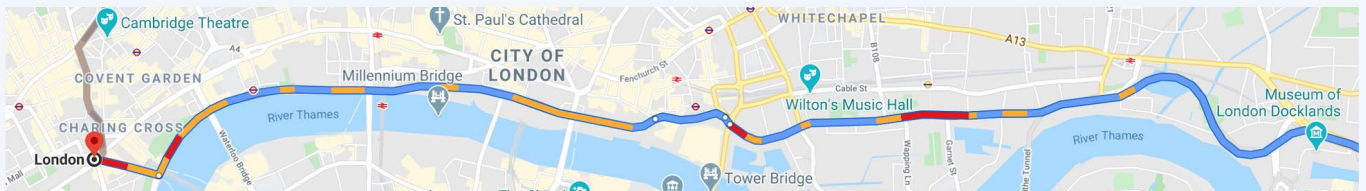


## 9.3 Data Visualizations Revision Guide

### 🌱 Key Vocabulary

Word	Meaning
<b>Data</b>	Information collected for analysis
<b>Prediction</b>	A guess about what might happen, based on knowledge
<b>Criteria</b>	The conditions or rules used to filter or focus your data
<b>Outliers</b>	Unusual data points that don't fit the trend

### Tracking the traffic



What do the different colours on this visualisation represent?

What data is needed to show you this visualisation?

How do you think this data is collected?

Think/pair/share.



### Tracking the traffic

The image shows a route map taken from Google Maps.

The blue represents the directions; the other colours represent the traffic conditions. To be able to visualise the data, the service needs to know how many cars are currently in that area and what speed they are travelling at.

The data will be collected from mobile devices in the car that are either using the Android operating system or have Google services installed. The data doesn't just come from current conditions, but also from past data that they have for that route at that time of day.

### Where is the best place to live?



### Best place to live: criteria

Most of us like the idea of living somewhere like in this picture, but would it really be the best place to spend the rest of your life?

Which of the following are most important to you?

- Life expectancy
- Average income/wealth
- Health
- CO<sub>2</sub> emissions



Anything else? 

In this lesson we used the data visualization tools at [gapminder.org](https://gapminder.org) to find out where is the best place to live based on variables of our choosing. The image above shows the island of Fiji which looks like paradise. You will notice that when you track life expectancy against income for Fiji alone, the graph spikes negatively in 1875. This is because a third of the population died that year due to a measles outbreak. Investigating outliers such as this are part of data visualizations. It is good practice to investigate these anomalies online.

### Lesson 3: Statistical State of mind

---

#### What Is the Investigative Cycle (PPDAC)?

PPDAC is a process that helps you **solve problems using data**. It stands for:

**P – Problem:** What question are we trying to answer?

**P – Plan:** How will we collect and analyse the data?

**D – Data:** Collect or find relevant data

**A – Analyse:** Look for patterns, trends, correlations, and outliers

**C – Conclusion:** What do the results show? What is the answer or recommendation?

---

#### Data in Action: The Roller Coaster Scenario

You explored what makes a “cool” roller coaster by:

- ✓ Refining the problem into clear questions
  - ✓ Visualising the roller coaster data (e.g. height, speed, length)
  - ✓ Finding **correlations** (e.g. taller coasters tend to be faster?)
  - ✓ Spotting **outliers** that didn’t follow the pattern
  - ✓ Making a final recommendation based on your findings
- 

#### Key Concepts

Term	Meaning
<b>Correlation</b>	A relationship between two things (e.g. when one goes up, the other might too)
<b>Outliers</b>	Data points that are very different from the rest
<b>PPDAC</b>	A 5-step cycle used to solve problems with data
<b>Investigative Cycle</b>	Another term for PPDAC – it’s how data scientists approach a problem

---

#### How to Use Data to Support a Recommendation

Once you’ve analysed the data:

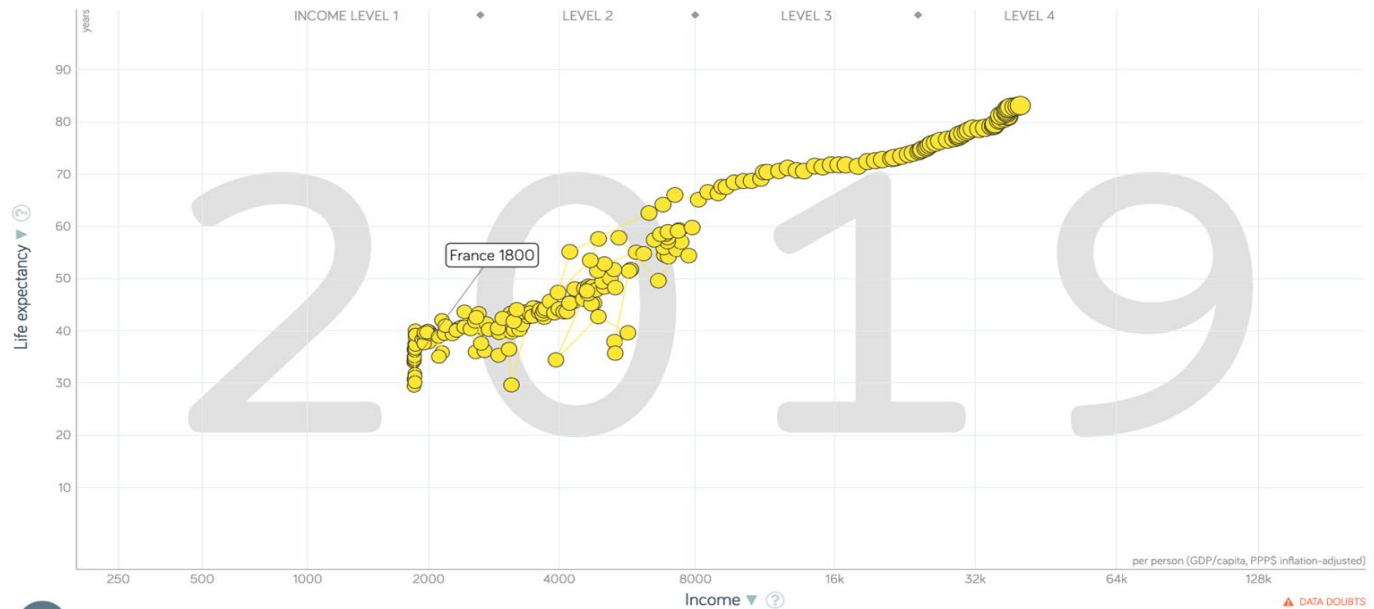
- **Summarise your findings clearly**
  - **Explain any patterns or surprises**
  - **Use visual evidence** (like graphs) to back up your conclusion
  - **Make a recommendation** based on what the data shows
-



## 9.3 Data Visualizations Revision Guide

### ✓ Key Learning Takeaways

- ◆ PPDAC helps you think like a real data scientist
- ◆ Correlations and outliers tell important stories in your data
- ◆ Your final recommendation should be **evidence-based**



The graph compares two **variables**:

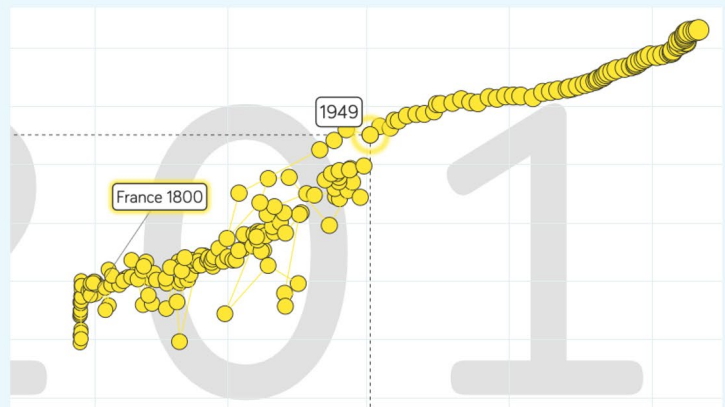
- Life expectancy
- Income

It compares the life expectancy of people living in France between 1800 and 2019.

### Does the graph show a trend?

Yes. From 1949 onwards it shows a clear upwards trend, showing that there is a relationship between the two **variables**; we call that a **correlation**.

This example has a **positive correlation**, meaning that as one variable increases, the other one increases too.

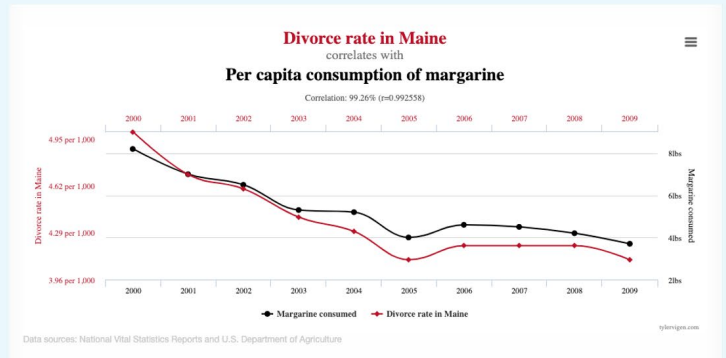


### Correlation vs Causation

#### 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?



Source: [www.tylervigen.com/spurious-correlations](http://www.tylervigen.com/spurious-correlations)

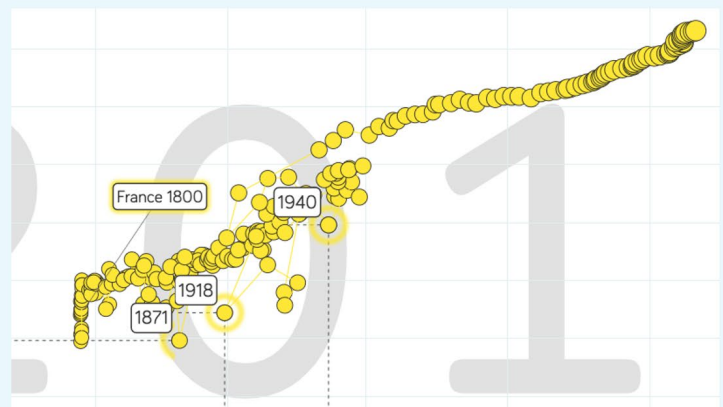
### Anomalies in data

#### Where are the anomalies in the data?

Until 1949, most of the data follows a slow upward trend, but there are a few odd blips.

Data that sits outside a trend is known as 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.



## Reasons for the outliers

1869 life expectancy = 41.1

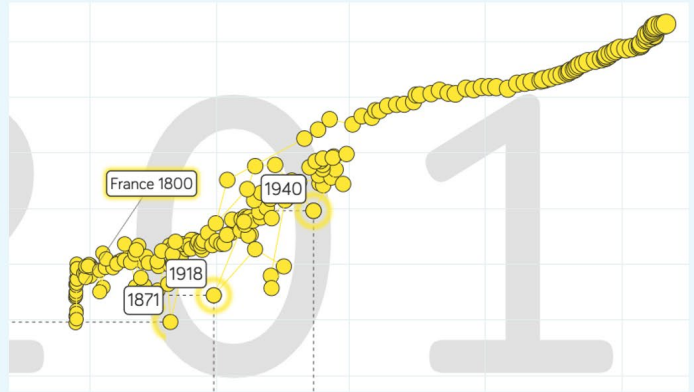
1870 life expectancy = **36.4**

1871 life expectancy = **29.6**

1872 life expectancy = 42.6

### Likely reasons:

Franco-Prussian War (19 July 1870 to 28 January 1871)



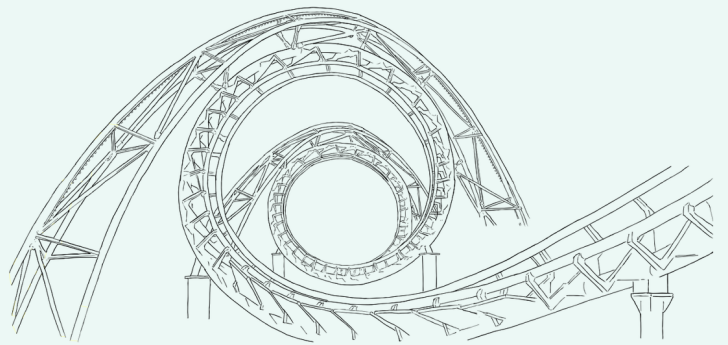
## Rollercoaster Problem

### The problem: roller coasters (River Kingdom)

“What makes a really cool roller coaster?” would be considered a poorly defined problem.

It doesn't help us to understand what we are measuring.

What variables about roller coasters could we **measure** in order to help us answer that question?



In this lessons we investigated what makes a cool rollercoaster by analysing data from the CODAP website. By graphing variables such speed, height, drop, number of inversions etc. we could start answering the question but first we had to generate more precise questions such as...

- Is there a relationship between top speed and maximum height?
- Is there a relationship between speed and seating position?
- What is the average number of inversions on roller coasters?
- Do roller coasters get faster or taller the more recently they were built?

## Lesson 4: Data for Action

---

### The Investigative Cycle: PPDAC (Recap)

You're now applying the **PPDAC cycle** to a **real-world problem** — litter in your school.

**P – Problem:** What's the issue? (e.g. Where is litter a problem?)

**P – Plan:** What questions will help solve it? What data do you need?

In this lesson, you focused on the **first two steps**:

- ✓ Defining the problem
  - ✓ Planning how to collect the right data
- 

### Asking Smart Questions

To investigate litter, you need to ask **clear, focused questions**, like:

- Where is litter most commonly found?
- What types of litter are most common?
- When is litter most likely to appear?

These questions guide the **type of data** you'll need.

---

### Collecting the Right Data

You created a **data capture form** to collect the information needed to answer your questions.

Your form might include:

- ✓ Location
- ✓ Type of litter
- ✓ Time of day
- ✓ Quantity

This makes it easier to collect **consistent and useful data**.

---

### Key Vocabulary

Word	Meaning
<b>PPDAC</b>	Problem, Plan, Data, Analyse, Conclusion – the cycle used to investigate problems
<b>Investigative cycle</b>	Another name for PPDAC



## 9.3 Data Visualizations Revision Guide

Word	Meaning
<b>Data capture</b>	The process of collecting data
<b>Data source</b>	Where your data comes from (e.g. a survey, observation, online database)

---

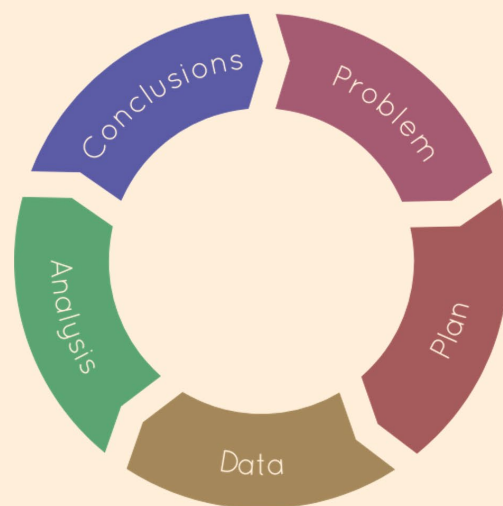
### 📌 What You've Learned

- ◆ How to **identify a problem** and plan an investigation
- ◆ How to **choose the right data** to collect
- ◆ How to **design a digital form** to capture data accurately

### The investigative cycle

So far we have spent time investigating data sets to see patterns or to extract meaning.

The **PPDAC** cycle is a framework for us to follow when asking and answering real-world problems using data.



### The litter at Disneyland

“Walt Disney wanted to know just how long a park patron would go with trash in their hand before just letting it drop to the ground. So he sat on a bench and watched the visitors of his park, counting the steps of those looking for a place to throw out their garbage. He counted 30 steps on average, and that is still the distance between each trash can in Disney, further ensuring a clean experience.”

- How did Walt Disney narrow down the larger problem into a more specific question? What was that question?
- What was the outcome?
- Did it solve the problem?



It should be clear from the story that Walt Disney had a problem that he wanted to solve. He decided he would try and solve it by posing the question, “How long will a park patron go with trash in their hand before letting it drop to the ground?” He worked out the average distance it took for people to drop litter if there was no bin.

## 9.3 Data Visualizations Revision Guide

The story by itself does not show whether the problem was solved. Further data would need to be collected to see what impact the action had had (see cycle element of the PPDAC investigative cycle.)

In this lesson we turned our attention to our own school community and asked the question, how can we use data to help us improve our school community by reducing the amount of litter?

The next step was to decide...what questions do we need answers to, to help us solve this problem?

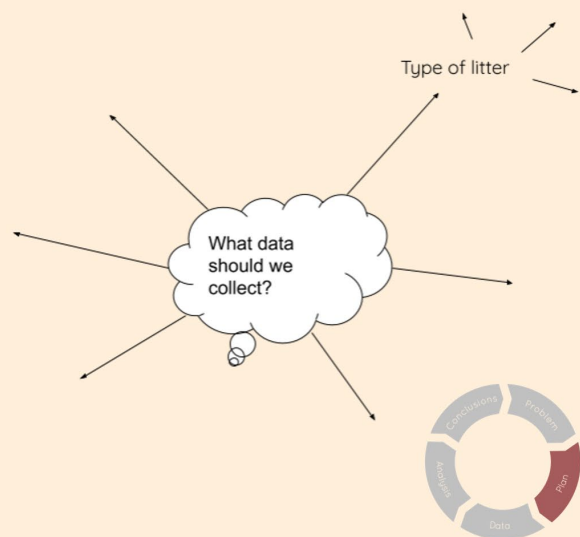
### The plan: part 2

The next part of your plan involves thinking about the data.

We don't have a data set to analyse, so we will need to collect the data ourselves.

What data could we collect about the litter that we find?

Complete Task 3 on your worksheet.



### Ideas for data to collect

- Type of litter
  - Food waste
  - Food packaging
  - Stationery
- Day/time of day
- Material
- Quantity of litter at location
- Location where litter was dropped
- Distance from bin
- Is it recyclable?
  - What category is it?
  - Is it recycled locally?



We spent time collecting data in class using MS Forms but questions on the use of MS Forms will not be part of this exam.

### Lesson 5: Clean It Up

---

#### Back to the PPDAC Cycle




In this lesson, you focused on the **D – Data** and **A – Analyse** stages of the **PPDAC cycle**. You worked with the **litter investigation data** you collected, preparing it for analysis and beginning to answer your key questions.

---

#### What Is Data Cleansing and Why Is It Important?

**Data cleansing** means checking and correcting data to make sure it's accurate and useful. Problems in your data can lead to **wrong results** when you analyse it!

Common issues include:

-  Typos or missing values
  -  Inconsistent labels (e.g. “Playground” vs “playground”)
  -  Duplicate entries
- 

#### What You Did in This Lesson

1. **Downloaded your collected data**
  2. **Cleansed the data** to fix errors and inconsistencies
  3. **Uploaded your cleaned data** to **CODAP** (a visual data tool)
  4. **Created charts and graphs** to look for patterns, trends, and insights
- 

#### Why Visualisation Matters

Once your data is clean, visualising it helps you:

- ✓ Spot patterns (e.g. where litter is most common)
  - ✓ See trends over time or location
  - ✓ Prepare to draw conclusions in the next lesson
- 

#### Key Vocabulary

Word	Meaning
<b>PPDAC</b>	Problem, Plan, Data, Analyse, Conclusion – your guide for investigating with data



## 9.3 Data Visualizations Revision Guide

Word	Meaning
<b>Data</b>	Information collected to answer a question
<b>Analysis</b>	Looking at data to find patterns, relationships, and meaning
<b>Data cleansing</b>	The process of fixing or removing inaccurate or messy data

---

### What You've Learned

- ◆ How to **identify and fix problems** in a data set
  - ◆ How to **prepare data for analysis**
  - ◆ How to **visualise data** to start answering your own investigation questions
- 

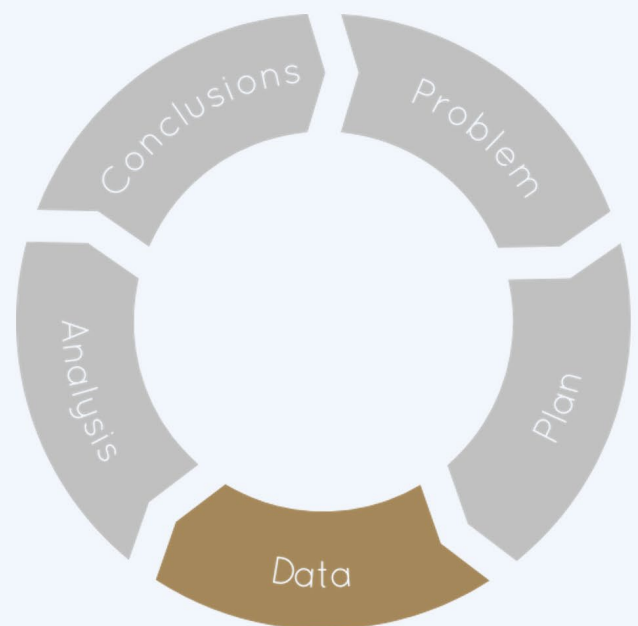
Clean data = Clear answers!  

### Clean it up

In this step, we gathered the data.

Once we have the data we need to help us answer our question, we should look through it to see if it needs **cleansing**.

Cleansing involves **detecting** and **correcting**, or **removing**, corrupt or inaccurate data.



In this lesson we learned about the importance of cleaning 'dirty data' before beginning the analysis stage. Cleaning up data by searching for missing values, duplication and invalid data is a necessary part of the process.



## 9.3 Data Visualizations Revision Guide

### Missing values

Name	Animal type	Weight (kg)	Height (m)	Age	Gender
Echo	Elephant	5900	3.3	45	Female
Stretch	Giraffe	800	5.9		Male
Yakov	Meerkat	0.73	0.32	11	Male
Maiya	Meerkat	0.67	0.33	11	Female
Vassily	Meerkat	0.69	0.32	4	Male
Bogdan	Meerkat	0.76	3	1	Male
Spot	Giraffe	820	5.1	1	Male
Batyr	Elephant	6000	3.2	32	Male
Sher	Lion		1.1	7	Female
Lavi	Lion	130	1.2	4	Female
Sarabi	Lion	124	1.3	4	Female
Drona	Elephant	0	3.1	9	Male
Alexander	Meerkat	0.71	3	4	Male
Scar	Lion	190	1.3	8	Male
Beo	Giraffe	750	4.9	5	Female
Alexander	Meerkat	0.71	3	4	Male

### Duplicate entry

Name	Animal type	Weight (kg)	Height (m)	Age	Gender
Echo	Elephant	5900	3.3	45	Female
Stretch	Giraffe	800	5.9		Male
Yakov	Meerkat	0.73	0.32	11	Male
Maiya	Meerkat	0.67	0.33	11	Female
Vassily	Meerkat	0.69	0.32	4	Male
Bogdan	Meerkat	0.76	3	1	Male
Spot	Giraffe	820	5.1	1	Male
Batyr	Elephant	6000	3.2	32	Male
Sher	Lion		1.1	7	Female
Lavi	Lion	130	1.2	4	Female
Sarabi	Lion	124	1.3	4	Female
Drona	Elephant	0	3.1	9	Male
Alexander	Meerkat	0.71	3	4	Male
Scar	Lion	190	1.3	8	Male
Beo	Giraffe	750	4.9	5	Female
Alexander	Meerkat	0.71	3	4	Male

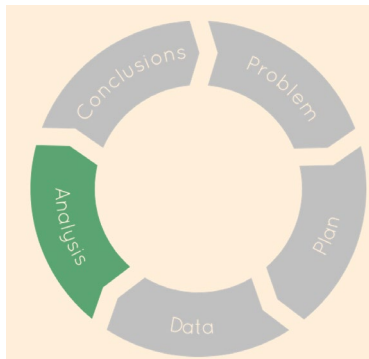
### Invalid data (outside normal range)

Name	Animal type	Weight (kg)	Height (m)	Age	Gender
Echo	Elephant	5900	3.3	45	Female
Stretch	Giraffe	800	5.9		Male
Yakov	Meerkat	0.73	0.32	11	Male
Maiya	Meerkat	0.67	0.33	11	Female
Vassily	Meerkat	0.69	0.32	4	Male
Bogdan	Meerkat	0.76	3	1	Male
Spot	Giraffe	820	5.1	1	Male
Batyr	Elephant	6000	3.2	32	Male
Sher	Lion		1.1	7	Female
Lavi	Lion	130	1.2	4	Female
Sarabi	Lion	124	1.3	4	Female
Drona	Elephant	0	3.1	9	Male
Alexander	Meerkat	0.71	3	4	Male
Scar	Lion	190	1.3	8	Male
Beo	Giraffe	750	4.9	5	Female
Alexander	Meerkat	0.71	3	4	Male

(Meerkats are typically not 3 meters tall)

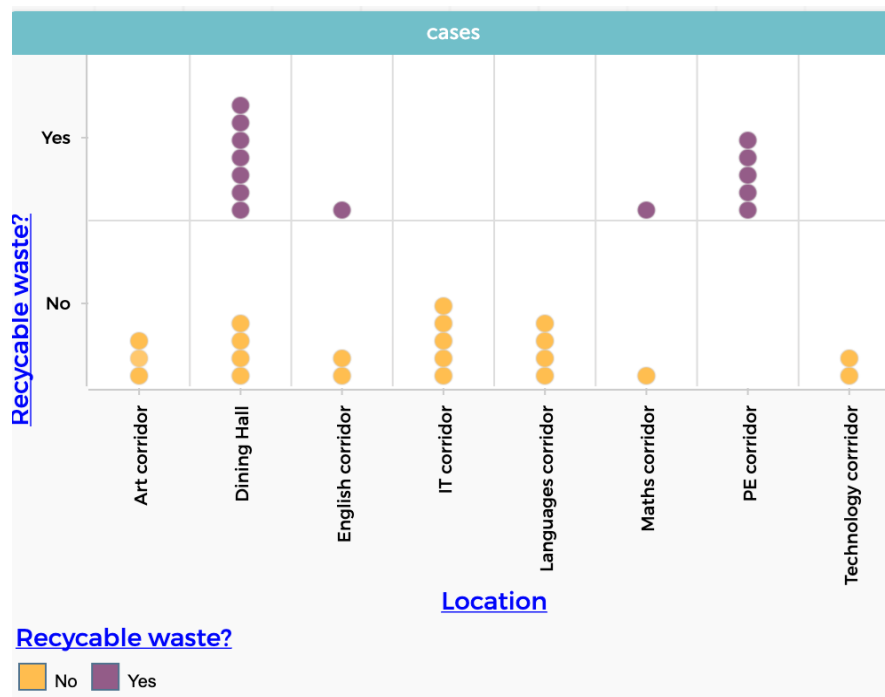
## 9.3 Data Visualizations Revision Guide

### ANALYSIS



Sample data set - Data sheet							
cases (35 cases)							
index	Description	Type of litter	Location	Purchased/canteen	Distance to nearest bin	Recyclable waste?	Compostable?
1	sandwich	food waste	Dining room	Yes	1	No	Yes
2	hot can	food waste	Dining room	Yes	1	No	Yes
3	hot can	food waste	Dining room	Yes	5	Yes	No
4	canteen	food waste	Dining room	Yes	1	Yes	No
5	flavoured	drinks	Dining room	Yes	5	Yes	No
6	flavoured	drinks	Dining room	Yes	1	Yes	No
7	sandwich	food waste	Dining room	Yes	5	No	Yes
8	hot can	food waste	Dining room	Yes	1	No	Yes
9	canteen	food waste	Dining room	Yes	1	Yes	No
10	flavoured	drinks	Dining room	Yes	5	Yes	No
11	flavoured	drinks	Dining room	Yes	5	Yes	No
12	crisp packet	food waste	Language corridor	No	15	No	No
13	crisp packet	food waste	Language corridor	No	15	No	No
14	crisp packet	food waste	Language corridor	Yes	5	No	No
15	Pen	stationery	Language corridor	No	15	No	No
16	crisp packet	food waste	Maths corridor	No	1	No	No
17	paper	Paper	Maths corridor	No	1	Yes	No
18	canteen	food waste	Art corridor	Yes	1	No	Yes

Next we analysed our data by plotting variables against each other to create visualizations using the CODAP website.



When concluding, it is important to consider...

- The original question that you posed
- The visualisation that helps you to answer the question, and which variables you have included in the visualisation
- What conclusions (if any) you can draw
- Would any further investigation or data help you to answer the question?