

Data

Textbook

Data



Data is all around us. In today's world, people can track their heart rate throughout the day, their sleep cycles, the miles per gallon of gas on their car, their social media use, and their temperature preferences for their homes. All of this information can be helpful for people to know. When data is understood, it can help people save money, live healthier, and more efficient lives.

Understanding what data means is important to be able to help us make decisions. Imagine that you wanted to know how many sleep cycles you have at night. You wear a device that tracks your sleep and the next morning it just shows you a list of numbers. Without some context, you don't know what the numbers mean! Identifying patterns in data is important to be able to extract meaning from data sets. Information is the collection of facts and patterns extracted from data.

Identifying Patterns in Data

Florida's High School Graduation Rates



Let's take a look at this set of data generated by the Department of Education in Florida. They have tracked high school graduation rates for the state of Florida over the years. Let's take a look at the percentage of students who graduated from high school each year.

Year	Florida High School Graduation Rates by Year
2010-2011	70.6%
2011-2012	74.5%
2012-2013	75.6%
2013-2014	76.1%
2014-2015	77.9%
2015-2016	80.7%
2016-2017	82.3%
2017-2018	86.1%
2018-2019	86.9%

Learn more about this data set [here](#).

What pattern do you see in this data set? If you noticed the percentage got larger each year, you are correct! This is an upward trend.

The table helps organize data in a way that reveals trends. Programs such as spreadsheets help efficiently organize and find trends in information.

Number of Telephone Landlines

Now let's look at another data set. This set of data shows the number of landline phones in the United States. Landline phones were used as the main phone line for households before cell phones.

Year	Number of Landlines in the United States
2013	133.23 million
2014	128.5 million
2015	124.85 million
2016	121.34 million
2017	116.3 million
2018	110.33 million
2019	106.43 million
2020	101.8 million
2021	97.22 million
2022	93.83 million

Learn more about this data set [here](#).

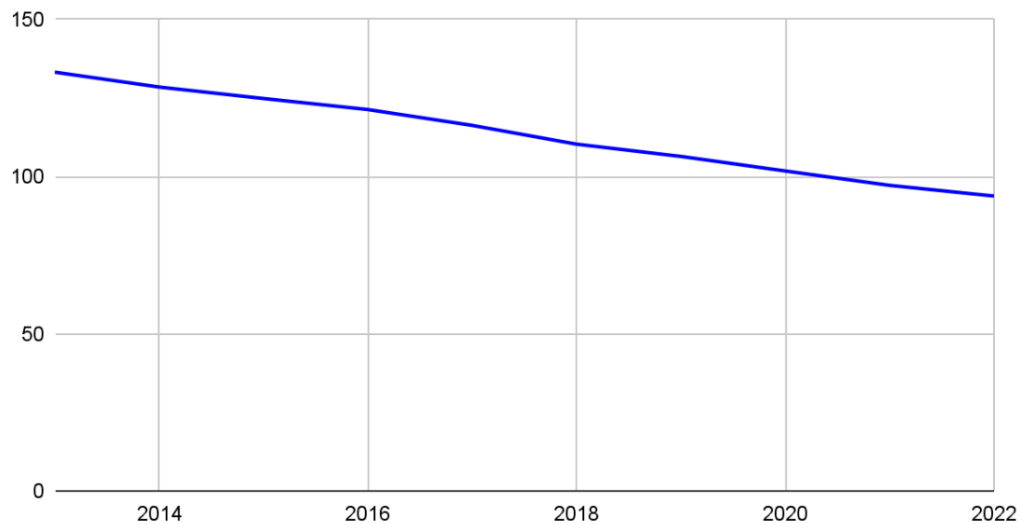
If you look at the number of households that have landline phones, it's slowly going down every year. This is a decreasing trend, so we are seeing fewer houses with landline phones. Why do you think that is? What do you think will happen in the future?

Sometimes a chart helps to understand data more clearly with visualizations.

Visualizing Data

Sometimes a chart helps to understand data more clearly with visualizations.

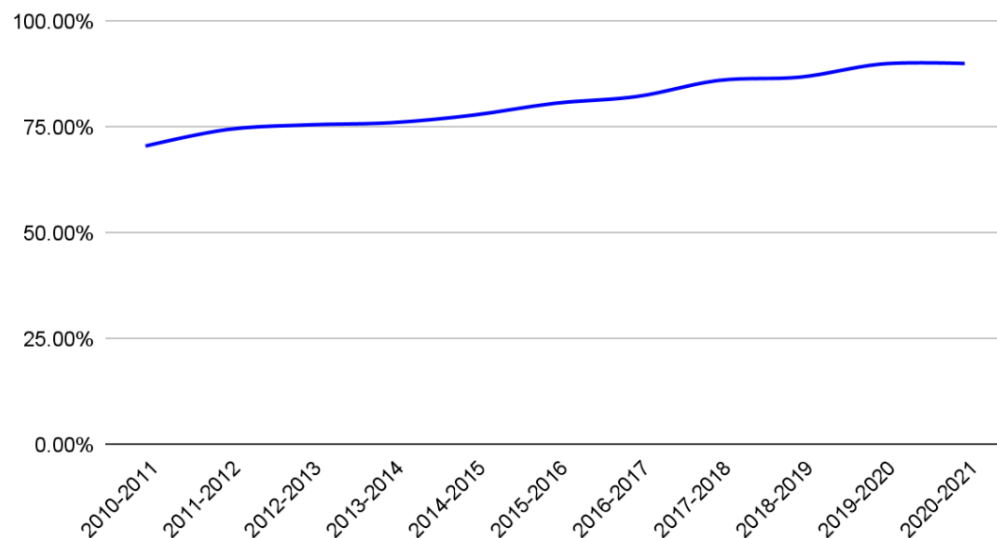
Number of Landlines in the US (by million)



Visualizing Data

Let's see a chart of the high school graduation rates in Florida.

Florida High School Graduation Rates



As you can see, the number of high school graduates in Florida clearly has an upward trend. Data visualizations such as charts help us to understand overall trends in data.

Graphs help us to analyze topics by diving into their data in a visual way. Data can help us to improve quality of life for people all over the world as we understand different situations.

What kinds of predictions can we make about the high school graduation rate in Florida or the number of landline phones in 2029? What about 2030 or even 2050?

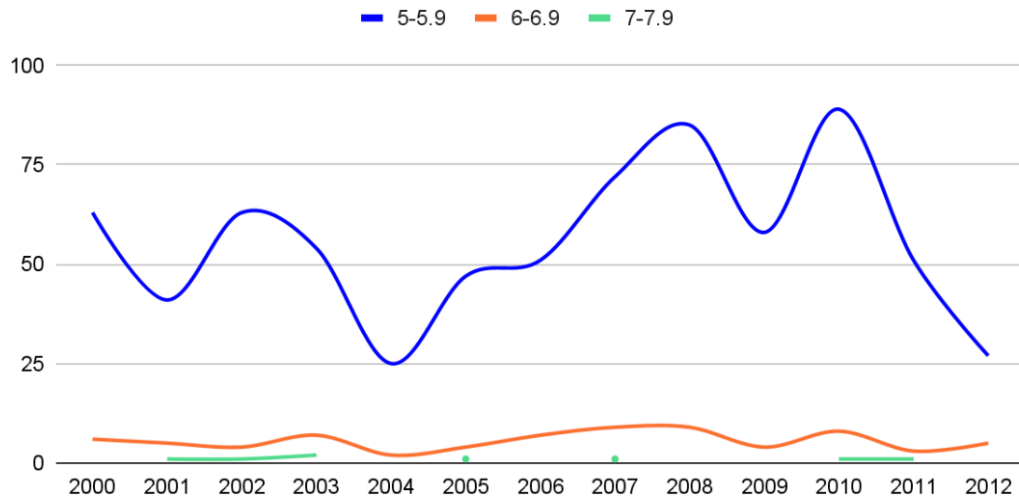
Data can help us to make predictions about the future.

Complex Charts

Often graphs are more complicated than the above two charts. Here is a graph of the number of earthquakes in the United States. Earthquakes are measured in magnitude. Magnitude is the size of the earthquake.

Earthquakes in the United States

By Magnitude



Practice understanding different sets of data at [Google Trends](#). Learn more about this data set [here](#).

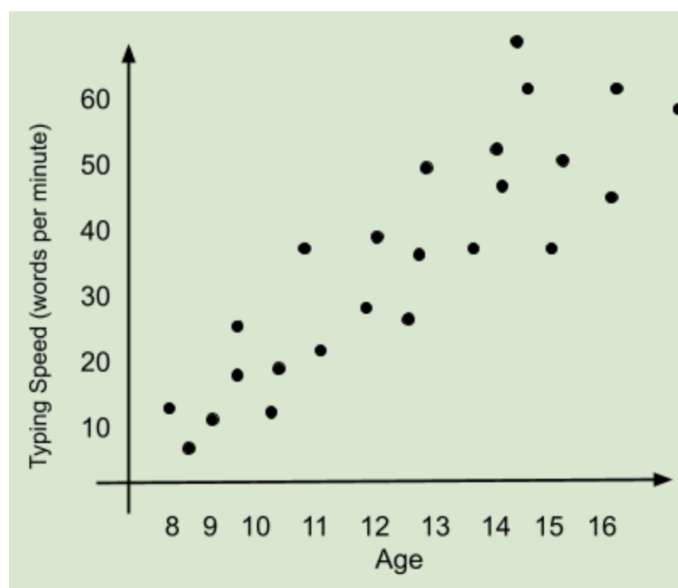
In the above chart we can see that there were more earthquakes that had a 5-5.9 magnitude than anything else consistently. What else can we learn from this chart?

Data filtering systems are important tools for finding information and recognizing patterns in data.

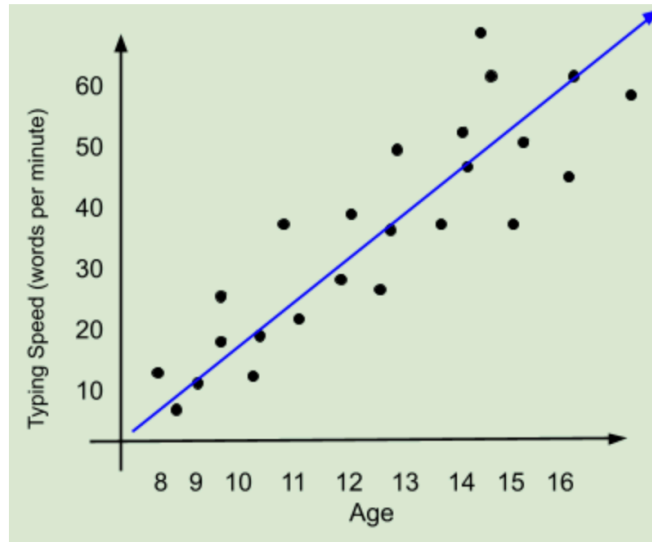
Looking for Correlations

Sometimes correlations are harder to see right away. Here are some examples of a scatter plot. Let's see if we can find a pattern in these charts.

This chart shows typing speed as age increases. Can you see a trend in this data?

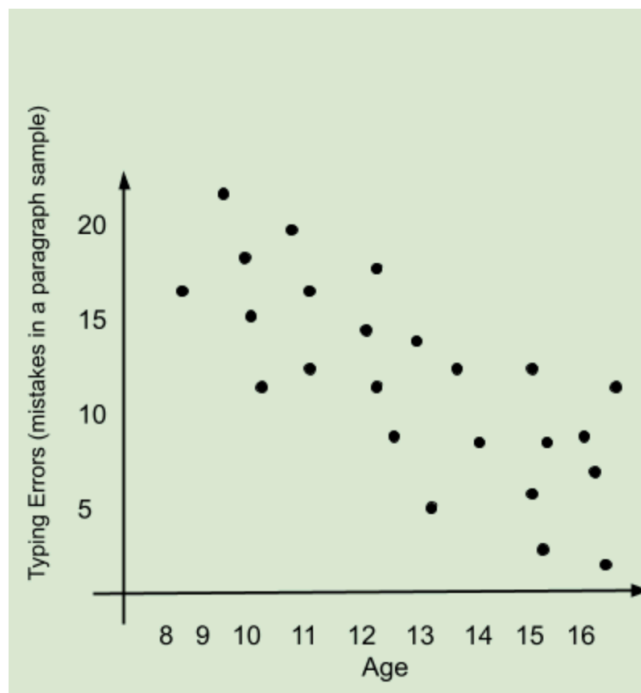


Let's add a trend line to see where the dots tend to go.



Looking at this graph, we can see that typing speed tends to get faster with age between the ages of 8 and 16.

Let's look at another scatter plot. Let's look at the comparison of age and typing errors in a paragraph sample.



This chart has a downward trend. We can tell that as age increases, the number of typing errors decreases.

It's important to note that just because something is correlated, does not show causation. For example, getting older doesn't *cause* faster typing speed. Practice leads to faster typing speeds. Additional research is needed to understand the exact nature of the relationship. Keep this in mind as you take a look at different data sets.

Often, a single source does not contain the data needed to draw a conclusion. It may be necessary to combine data from a variety of sources to formulate a conclusion.

Data Analysis

Data analysis is crucial for truly understanding complex systems, whether they're natural or human-made. By carefully examining large sets of data, scientists and researchers can find patterns, trends, and relationships that aren't obvious at first glance. For example, analyzing climate data helps us understand global warming, while studying economic data reveals insights into market behavior. This process allows us to build better models of these systems, make more accurate predictions, and ultimately develop more effective solutions to real-world challenges.

Data Collection

Now let's generate a data set of your own! Large data sets have many methods of collecting data—many of which require the use of computers! Some data sets have information that covers years worth of material. The earthquake example above comes from data that was collected over twelve years! That is a lot of data!

Let's start on a smaller scale. Ask 10 different classmates how much sleep they got last night. Write down their answers on a piece of paper.

Here's an example of the data you might have collected.

Classmate	Hours of Sleep
1	5
2	8
3	6
4	7
5	7
6	8
7	6
8	9
9	4
10	8

And just like that! You've generated a set of data!

How can we do this using code?

Data Collection with Python Code

If you haven't learned Python yet, don't worry about this section. If you have, see if you can create the code before checking to see our examples.

Create a python program that asks the user how many hours of sleep they got.



[Show answer/example](#)

Add their answer to an empty list.

[Show answer/example](#)

Can you think of a way to use a while loop that will keep asking if you have another person to answer the question?

[Show answer/example](#)

Now you have a list of answers in the variable named `data`. Use what you have learned in Python to sort the list, find the total number of hours slept, and the average number of hours slept.

[Show answer/example](#)

The data about hours slept is a simple example, but the principles are the same for large data sets. Computers can be a very powerful tool to analyze large data sets very quickly. Using this program we just created, we can sort, organize, and analyze a list with 10 items in it, or a list with thousands of items in it. The program is the same.

Data Visualizations



Take the data set you've just collected about hours of sleep and go to chartgo.com.

Select a bar graph.

In the chart data section, add the numbers 1-10 for your classmates in the X Data section. In the Y Data section, add the corresponding hours slept.

Select "create chart" to see a visualization of your data!

Big Data

When we talk about "massive" data, often called Big Data, we mean large amounts of data, created at high speeds, come in many varieties, have some uncertain qualities, and offer great value. Regular data analysis tools can't handle this, so we use special techniques.

Here are key ways to analyze huge amounts of data:

- **Distributed Processing:** Instead of one computer doing all the work, the data is split up and processed by many computers at the same time.
 - **Hadoop:** This is like a big storage and processing system. It stores huge files across many machines (HDFS) and processes them in large batches (MapReduce). Good for looking at past data.
 - **Apache Spark:** A faster, more flexible alternative. It processes data in memory, making it much quicker for complex tasks, machine learning, and even real-time data.
- **NoSQL Databases:** Unlike typical databases that use strict tables, NoSQL databases are designed for many different kinds of data, including messy or unstructured data (like social media posts). They are flexible and can grow very large. Examples are MongoDB and Cassandra.
- **Smart Analysis and Machine Learning:**
 - **Descriptive Analysis:** Just tells "What happened?" (e.g., last month's sales).
 - **Diagnostic Analysis:** Explains "Why did it happen?" (e.g., why sales dropped).
 - **Predictive Analysis:** Uses old data to guess "What will happen?" (e.g., next month's sales). This uses algorithms that learn from patterns.
 - **Prescriptive Analysis:** Suggests "What should we do?" (e.g., best price to set).
 - **Machine Learning (ML) and Artificial Intelligence (AI):** These are core to advanced analysis. ML helps computers find patterns, make predictions, and learn on their own. Techniques like **clustering** (grouping similar data) and **natural language processing (NLP)** (understanding human language) are used.
- **Real-time Processing:** For super-fast data (like live stock prices or fraud detection), data is analyzed as it arrives, giving instant insights for quick decisions.
- **Data Warehouses and Data Lakes:** These are special storage places. A **data warehouse** holds neatly organized data for reports. A **data lake** holds all data in its original, raw form, which is more flexible for new types of analysis and machine learning.

In short, for smaller, organized data, regular methods work. But for Big Data's huge, fast, and varied nature, you need powerful tools like distributed processing and machine learning to find valuable insights. The right method depends on the data and what you want to learn from it.

States of Data

Data exists in three states: **data at rest**, **data in transit**, and **data in use**. **Data at rest** is stored on devices like hard drives or cloud storage and can be threatened by hackers or unauthorized access. **Data in transit** is moving between devices or networks, such as emails or online messages, and is vulnerable to interception if not properly encrypted. **Data in use** is actively being processed by a computer, like when editing a document, and can be at risk from malware or unauthorized changes. Protecting data in all three states with strong passwords, encryption, and security software helps keep it safe from threats.

Data Scientists

With the increasing availability of large sets of data, there's an increasing demand for data scientists. A [data scientist](#) is someone who can compile, analyze, and present complex data in understandable ways. People who understand computer programming and data have many opportunities available to them, as

many different companies are interested in understanding data. Data scientists are in demand, and often have high salaries.

Thought Question: Would you be interested in becoming a data scientist? What about it sounds appealing to you? What might you enjoy? What might you not enjoy?

Data vs Metadata

Data is the actual information you want to know or use. For example, if you have a list of people's names, the names themselves are the data.

Metadata, on the other hand, is information **about** the data. It describes things like when the data was created, who created it, or how big the file is. So, if the names list was stored in a file, the metadata would tell you things like the file's size, its creation date, or the format it's saved in.

In short:

- **Data** = the actual content (e.g., names, numbers, pictures)
- **Metadata** = information about the content (e.g., size, date, creator)

Critical Thinking Questions

1. **Interpreting Data Patterns:** Reflect on the significance of identifying patterns in data sets. How does recognizing trends in data, such as upward or downward trends, help us understand real-world phenomena better?

2. **Role of Data Scientists:** Consider the role of data scientists in today's society. Why do you think there is a growing demand for data scientists? What skills and qualities do you think are essential for someone interested in pursuing a career in data science?

Summary

Data is a large part of virtually everything we do. Analyzing and understanding sets of data is becoming increasingly important as computers make large sets of data more readily available. Computer programs help us quickly and efficiently compile and analyze large data sets and better understand complex problems about the world around us. Creating charts helps to visualize the data in ways that can easily be understood. [Data scientists](#) specialize in compiling and organizing data into understandable formats.

Questions (3)

1. True or False: Correlation indicates causation.

MULTIPLE CHOICE

Choose the correct answer:

- A. True
- B. False

2. True or False: Data can help us to improve quality of life for people all over the world as we understand complicated situations.

MULTIPLE CHOICE

Choose the correct answer:

- A. True
- B. False

3. True or False: It may be necessary to combine data from a variety of sources to formulate a conclusion.

MULTIPLE CHOICE

Choose the correct answer:

- A. True
- B. False

Answer Keys & Solutions

Questions

1. True or False: Correlation indicates causation.

MULTIPLE CHOICE

Correct Answer:

A. True

✗ Incorrect

B. False

✓ Correct

Explanation:

Just because something is correlated, doesn't mean it shows causation.

2. True or False: Data can help us to improve quality of life for people all over the world as we understand complicated situations.

MULTIPLE CHOICE

Correct Answer:

A. True

✓ Correct

B. False

✗ Incorrect

Explanation:

Data can help us make more informed decisions

3. True or False: It may be necessary to combine data from a variety of sources to formulate a conclusion.

MULTIPLE CHOICE

Correct Answer:

A. True

✓ Correct

B. False

✗ Incorrect

Explanation:

Often, double checking research is necessary