Data Visualisation

Class 1

Camille DUQUESNE

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

Structure of the course

Class 1: Core data visualisation

Class 2: Data visualization pitfalls and good practices

Class 3: Case Study: analyzing data journalism articles

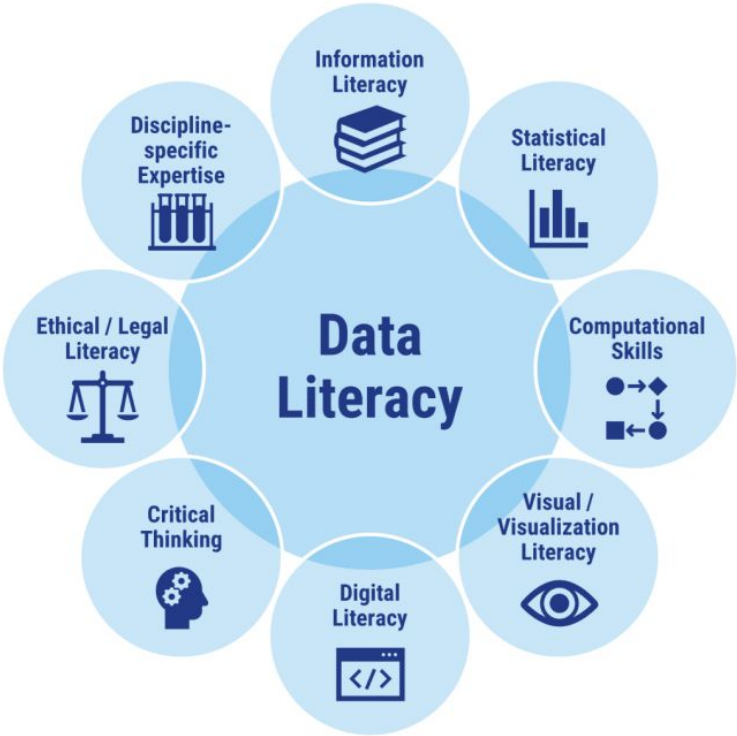
Class 4: Interactive data visualization + choosing a data set for your final project

Class 5: Peer feedback on your final project + wrapping up your project

Data literacy

One of the biggest goal of this class is for you to acquire data literacy.

Data literacy can be defined by **the ability to explore, understand, and communicate with data in a meaningful way**. [source]



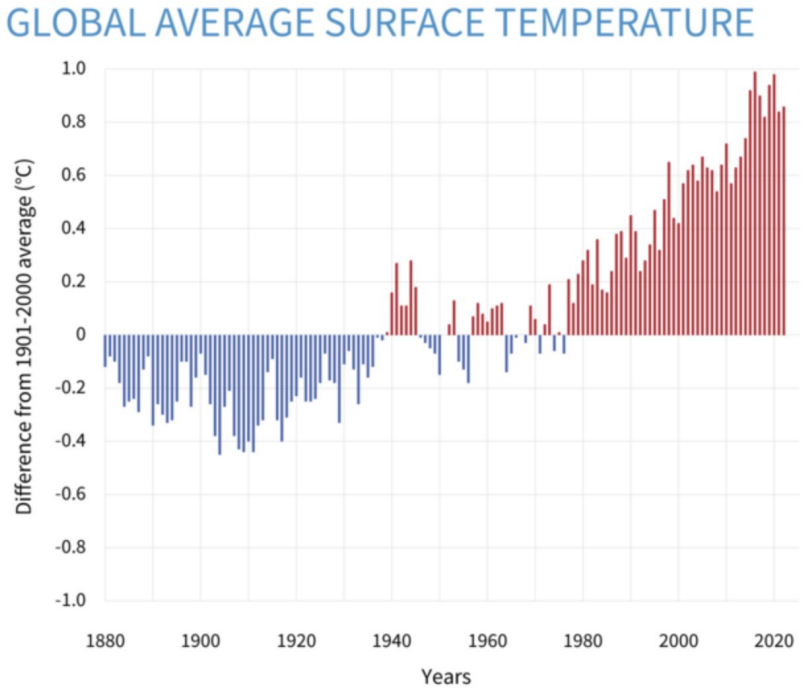
Structure of class 1

- Introduction to the course and objectives.
- Why data visualization is crucial in data analysis and communication.
- Key principles and goals of effective data visualization.
- Statistics refresher
- Why background information is important in data visualization
- An overview of different types of data visualizations: line charts, bar charts, scatter plots, histograms, box plot, pie chart
- When to use each type based on the data and the message you want to convey.
- Data cleaning and preprocessing for visualization
- Case study.

Learning objectives of class 1

- Describe the importance of data visualization in simplifying complex datasets and its role in facilitating insight discovery.
- Design data visualizations that adhere to key principles and goals, demonstrating an understanding of their significance.
- Explain the significance of understanding the context, sources, and quality of data in the data analysis process.
- Analyze potential consequences of neglecting the background of data, including misinterpretation and bias.
- Evaluate the ethical implications of using data without proper background knowledge.
- Recall fundamental statistical concepts, such as mean, median, standard deviation, and correlation.
- Apply statistical techniques to analyze and summarize data effectively.
- Analyze and interpret statistical results in the context of specific research questions or problems
- Evaluate the appropriateness of different statistical methods in various data analysis scenarios.
- Recall various types of data visualizations, including line charts, bar charts, scatter plots, histograms, box plots, and pie charts.
- Explain the characteristics and uses of each data visualization type.
- Create basic examples of each visualization type to demonstrate understanding.
- Analyze real-world visualizations to identify the type used and its appropriateness.
- Evaluate the effectiveness of different visualization types in conveying specific data and messages.
- Understand the importance of data cleaning and preprocessing in data analysis and visualization.
- Apply data cleaning techniques to identify and handle missing values, outliers, and other data quality issues.
- Create a report outlining key informations from a dataset, supported by visualisations.
- Analyse visualisation in a scientific and rigorous manner to extract key insights.

Why is data analysis important?

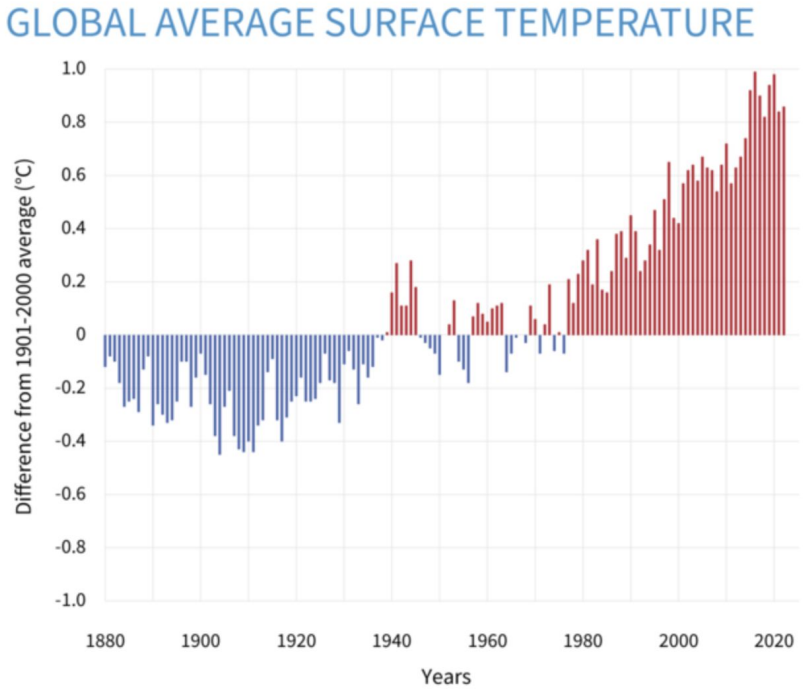


Yearly surface temperature compared to the 20th-century average from 1880–2022. Blue bars indicate cooler-than-average years; red bars show warmer-than-average years. NOAA Climate.gov graph, based on data from the National Centers for Environmental Information.

- Simplifies data
- Facilitates insight discovery
- Enhances decision making
- Supports communication
- Helps detect anomalies and outlier

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

Key Principles of data Visualization



Yearly surface temperature compared to the 20th-century average from 1880–2022. Blue bars indicate cooler-than-average years; red bars show warmer-than-average years. NOAA Climate.gov graph, based on data from the National Centers for Environmental Information.

- Simplicity
- Clarity
- Accuracy
- Relevance
- Consistency
- Audience-Centric
- Ethical

Statistics and Data refresher

What are the different types of data?

Quantitative data

Quantitative data is data that can be measured and represented as numbers. Quantitative data is either discrete or continuous.

Ex: the number of chairs in this room, the time spent revising your lessons, the number of people in this room, the temperature, ...

Qualitative data

Qualitative data is descriptive data that cannot be measured. Qualitative data is either nominal or ordinal (categories with an rank).

Ex: the shopping list, your rank in a video game (bronze, diamond, etc.), the weather (sun, cloud, rain), your eye color, your level of satisfaction, ...

The percentages

Percentage is the **ratio of a part of a system to the total of the system**. It can be written as a fraction or as a decimal number.

Often we write percentages with a ratio of 100. In a closed system we can also consider that the minimum and maximum values are 0% and 100% respectively.

The percentages

Among my 10 friends, 3 have blue eyes, what is the percentage of my friends who have blue eyes?

Out of a population of 2456 people, 15% have curly hair, how many people does this represent?

Transform 0.15 into %

Transform 0.05 into %

Transform 1.01 into %

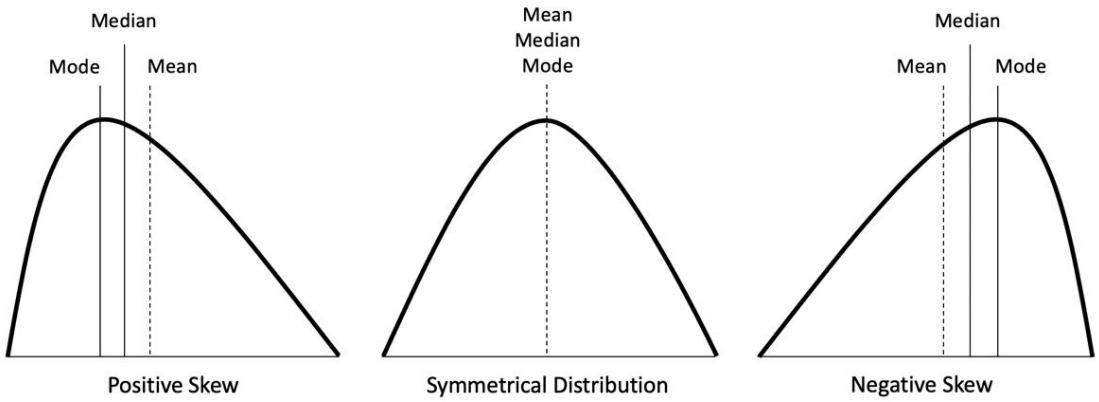
Transform 14/178 into %

Mean, median and mode

Mean: average value computed by adding up all values in a dataset and then dividing the sum by the number of values

Median: The middle value in a set of ordered data.

Mode: The most frequent value in the dataset.



Mean, median and mode

You have a list with the following values:

[1, 2, 2, 2, 2, 2, 3, 4, 6, 7, 8, 9, 9]

What is the mean?

What is the median

What is the mode?

How do these values change if we replace the last 9 with the number 100?

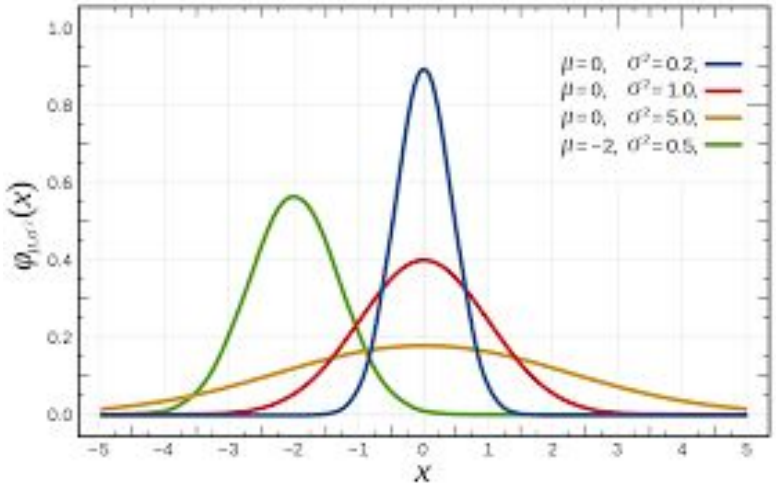
What about the mean, median and mode for the following list:

[1, 2, 3, 4, 4, 5, 5, 5, 6, 6, 7, 8, 9]

The standard deviation

Standard deviation: The square root, of the mean squared difference of values from the mean, to measure the dispersion of a data set from the mean.

The lower the standard deviation, the more the data is gathered around the mean. The larger the standard deviation, the more dispersed and far from the average the data is.



$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \bar{X})^2}{N}}$$

The standard deviation

We can calculate the standard deviation from our previous list:

[1, 2, 2, 2, 2, 2, 3, 4, 6, 7, 8, 9, 9]

First, let's calculate the sum of the squared difference of each point from the mean:

(1 - 4.38)2 + ((2 - 4.38)2 x 5) + (3 - 4.38)2 + (4 - 4.38)2 + (6 - 4.38)2 + (7 - 4.38)2 + (8 - 4.38)2 + ((9 - 4.38)2 x 2) = 107

Finally we can take the square root of our previous result divided by the number of elements in the list: $\sqrt{107/13} = 2.86$

It is also possible to automatically calculate this in python using the numpy module: numpy.std([1, 2, 2, 2, 2, 2, 3, 4, 6, 7, 8, 9, 9])

Probability

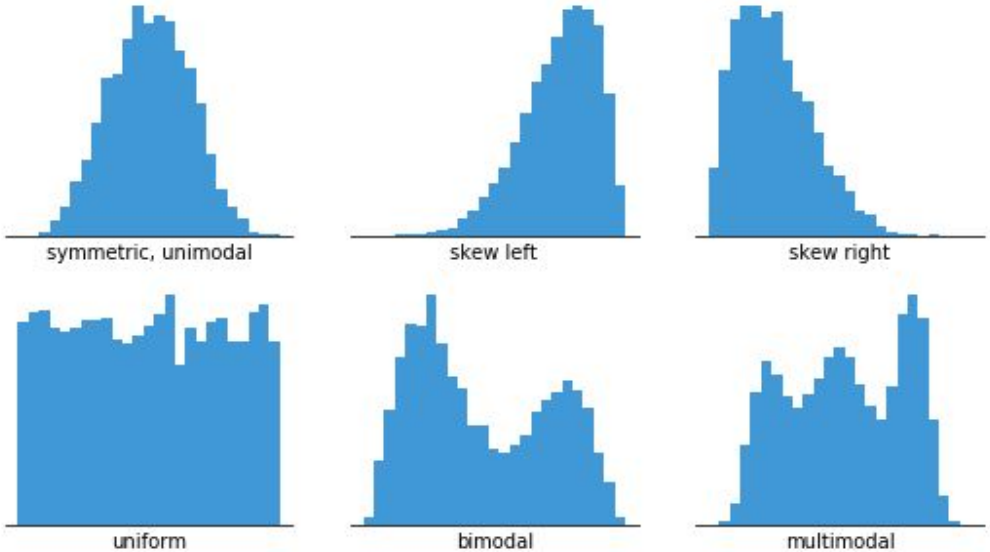
Probability is the measure of the possibility of an event occurring in a random experiment.

The probability of an event can therefore be measured as the number of favorable outcomes divided by the total number of possible outcomes.

Let's look at this link:
https://seeing-theory.brown.edu/basic-probability/index.html

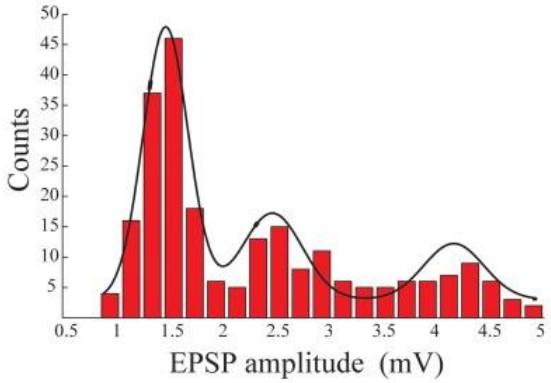
Distributions

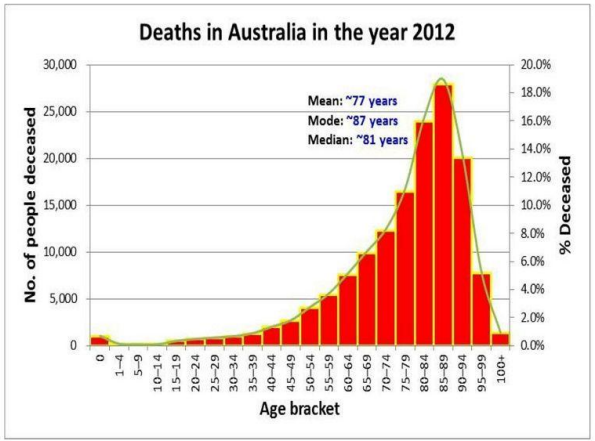
The distribution of observed probabilities can take several forms. The most common form is the normal distribution, which is symmetrical and where the mean, median and mode are the same. This is for example the case for the distribution of human weight, height, the diameter of tree trunks, etc.

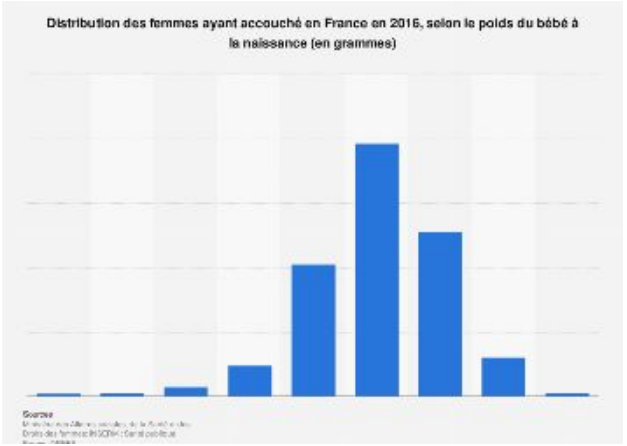


Distribution types

What are the distributions in the following graphs

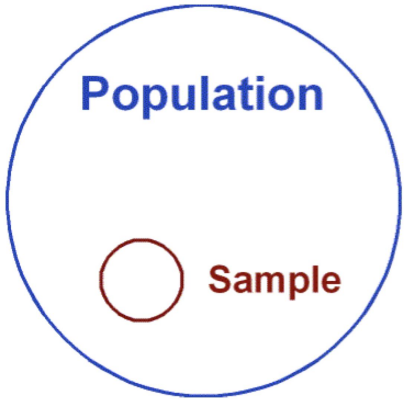






Sample vs population

It is crucial to understand that any data study is based on a **sample of the population** and not on the entire actual population. We certainly try to have the most representative sample of the population, but it is essential to apply a critical mindset to our conclusions because they will never be based on 100% of the population.



Quick data background checklist

One of the most crucial aspect in data analysis is **understanding the background of your data.** Without this knowledge, it's challenging to make informed decisions, detect potential biases, or effectively communicate findings, hindering the value and trustworthiness of data-driven analyses.

- ❑ What data was collected ? What is the data type of each column ?
- ❑ When was the data collected ?
- ❑ Who collected the data ? Who financed the collection of the data
- ❑ For which purpose was the data collected ?
- ❑ What is the licence of the data set ?
- ❑ How much data was collected ?
- ❑ Are there any missing data ?
- ❑ Are there any duplicates ?

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

Let’s switch to jupyter notebook !