Data visualisation and exploration

Thomas de Jaeger

thomas.de-jaeger@epita.fr

Fall 2025

Class details

- 15h: 2h*6 +3h presentation (AIS)

- Grade: 30% Labs
  70% Project/presentation

- For anonymous feedback: Feedback - Google Docs

Class organisation

- Lectures
- Labs during class
- Project

Syllabus

- **Session 1:** Introduction to Data Visualization & Data Sources
- **Session 2:** Good Practices & Common Pitfalls
- **Session 3:** Matplotlib & Seaborn: Visualization Essentials
- **Session 4:** Cartography & Geospatial Visualization
- **Session 5:** Case Studies: Data Journalism
- **Session 6:** Visualizing Multidimensional Data & Dimensionality Reduction
- (Work on your project)
- **Session 7**: Project presentation

Project

The project is due in jupyter notebook format + PPT presentation

You will choose a dataset and analyse it. You are free to select the data you want and answer the questions you want too. However, your notebook must include:

- **Introduction**

- What is the subject and context of the project?
- What is the source of the data?
- Who collected the data ?
- When was the data collected ?
- How was the data collected ?
- In what context/for what purpose was the data collected?
- What does the dataset contain? Describe the data, its type (continuous, categorical…)
- What is the size of the dataset?
- What is the dataset's license?

Project

- **What do you want to answer with those data?**

- Cite your research question and your 2 subquestions
- Do your sub questions allow you to answer your main research question ?
- Do you have the data to answer your subquestions ?
- What are your hypotheses?

- **FOR EACH SUB QUESTIONS**

Explain its relevance to your main question.
- What is your methodology for analysis?
- Produce at least one graphical representation (chart) per sub-question: - The code is clear and readable
- The chosen graphical representation is consistent with the type of data and the analysis approach.
- Graphic representation is complete
- The graphical representation respects best practices
- Graphical representation minimizes potential bias

- What analysis can you draw from the graphs? A detailed analysis is expected.
- What are your results?
- What are the limitations?

Project

- **Conclusions**

Summarize your main findings

- Does your analysis enable you to answer your research question? Why or why not?
- What are the limitations and biases of your methodology?
- Do you have any sources that support or contradict your conclusions?
- What would be the next steps in your analysis if you had had more time?
- Can you think of another data analysis method to answer your question? Expand.

- **Presentations (for AIS)**

10-15 min of presentation. You can use only your notebook but a ppt will be more professional. The grade will be defined as:

- 5pts: Overall presentation: clarity, slides
- 5pts: Work: what you have done
- 3pts: Answering the questions
- 2pts: Asking questions to others
- 5pts: Critical thinking

Introduction to Data Visualization & Data Sources

Session 1

What is data visualisation?

Data visualisation is the process of **representing data** and information through visual elements like **charts, graphs, and maps**. It transforms complex datasets into clear, accessible visuals, making it easier **to identify patterns, trends, and outliers**. This helps people **make faster, more informed decisions** across various fields.

Data to graph

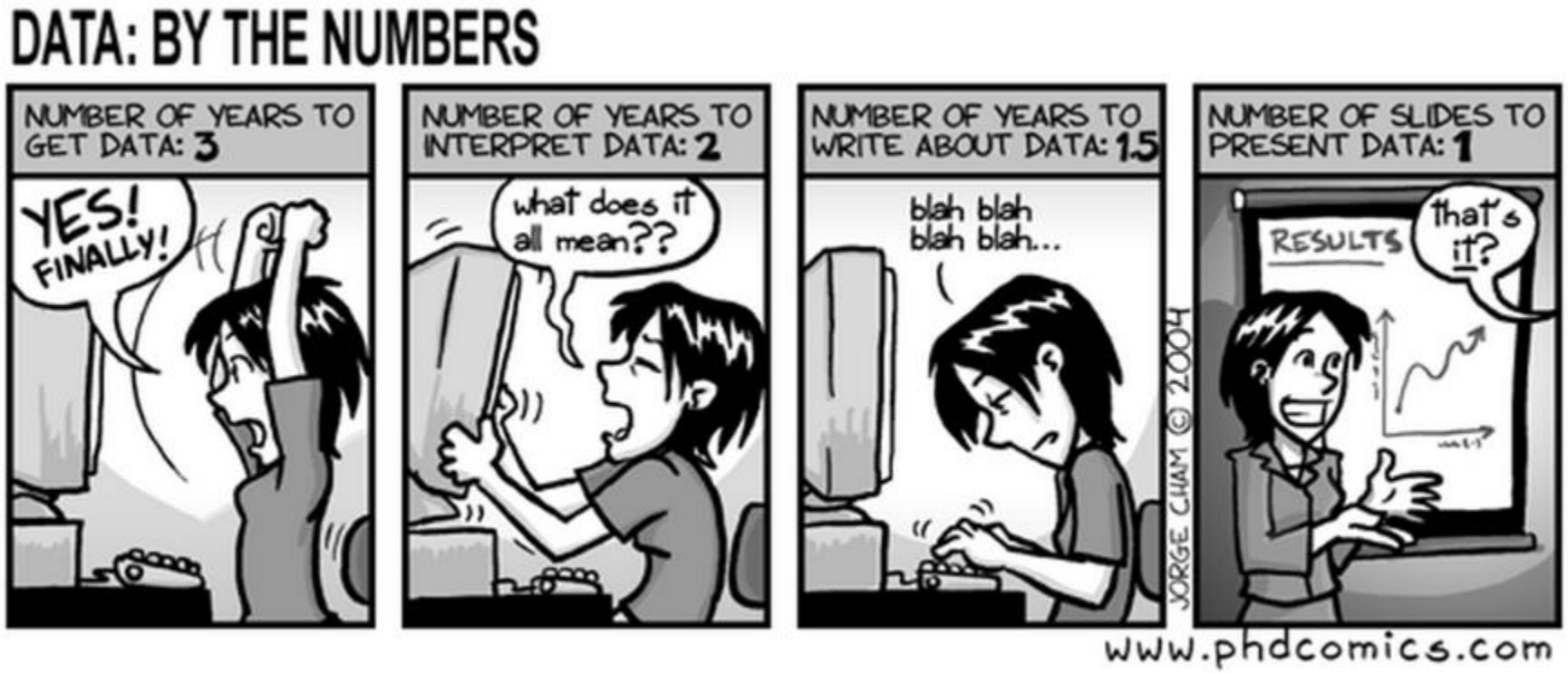


Figure 2.1: Data By the Numbers (by Jorge Cham)

Why data visualization is needed?

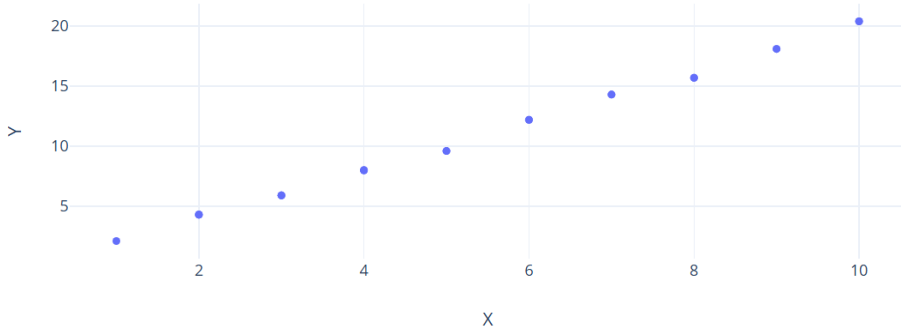
Our brains are wired for visuals:

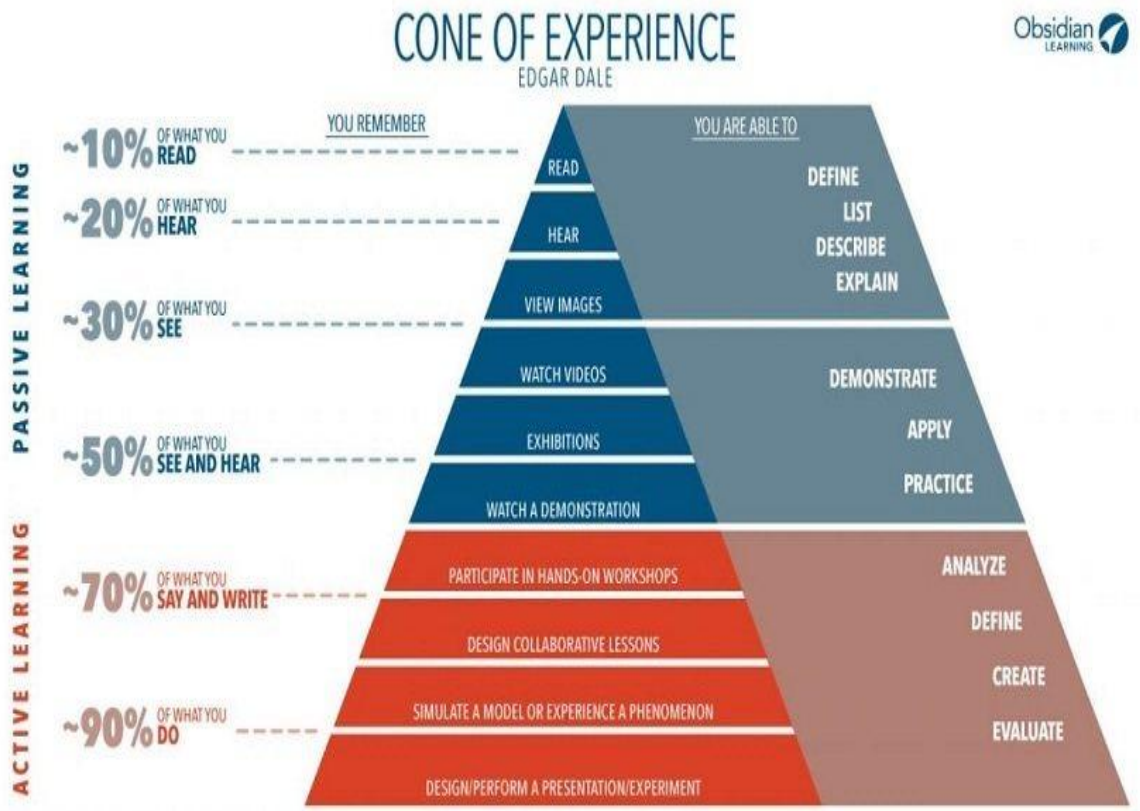
- We process images **60,000× faster** than text

Trends, outliers, clusters, and gaps stand out instantly in a chart

Tables may hide insights that a simple plot can reveal

1	2.1
2	4.3
3	5.9
4	8.0
5	9.6
6	12.2
7	14.3
8	15.7
9	18.1



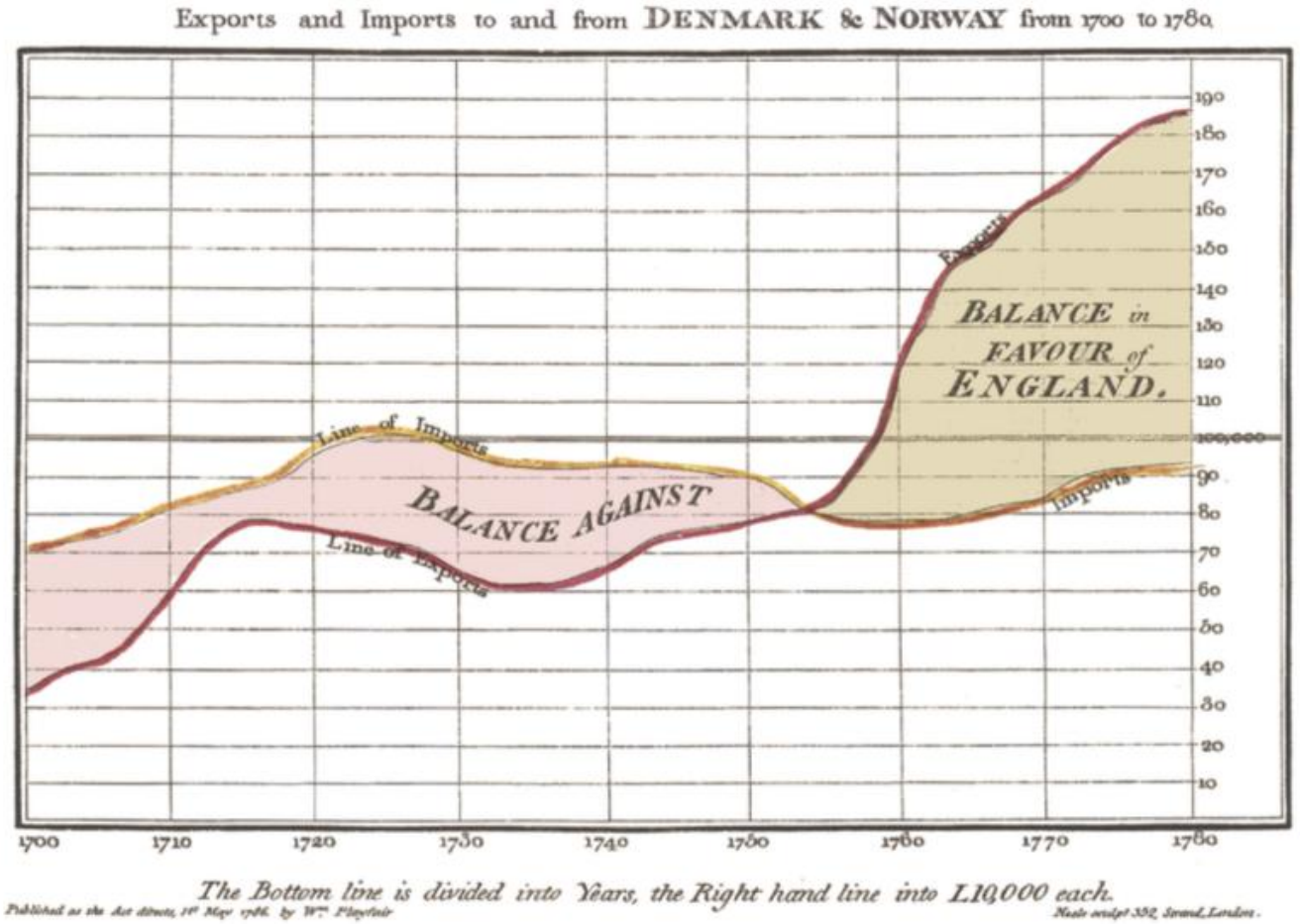


Is it something new?

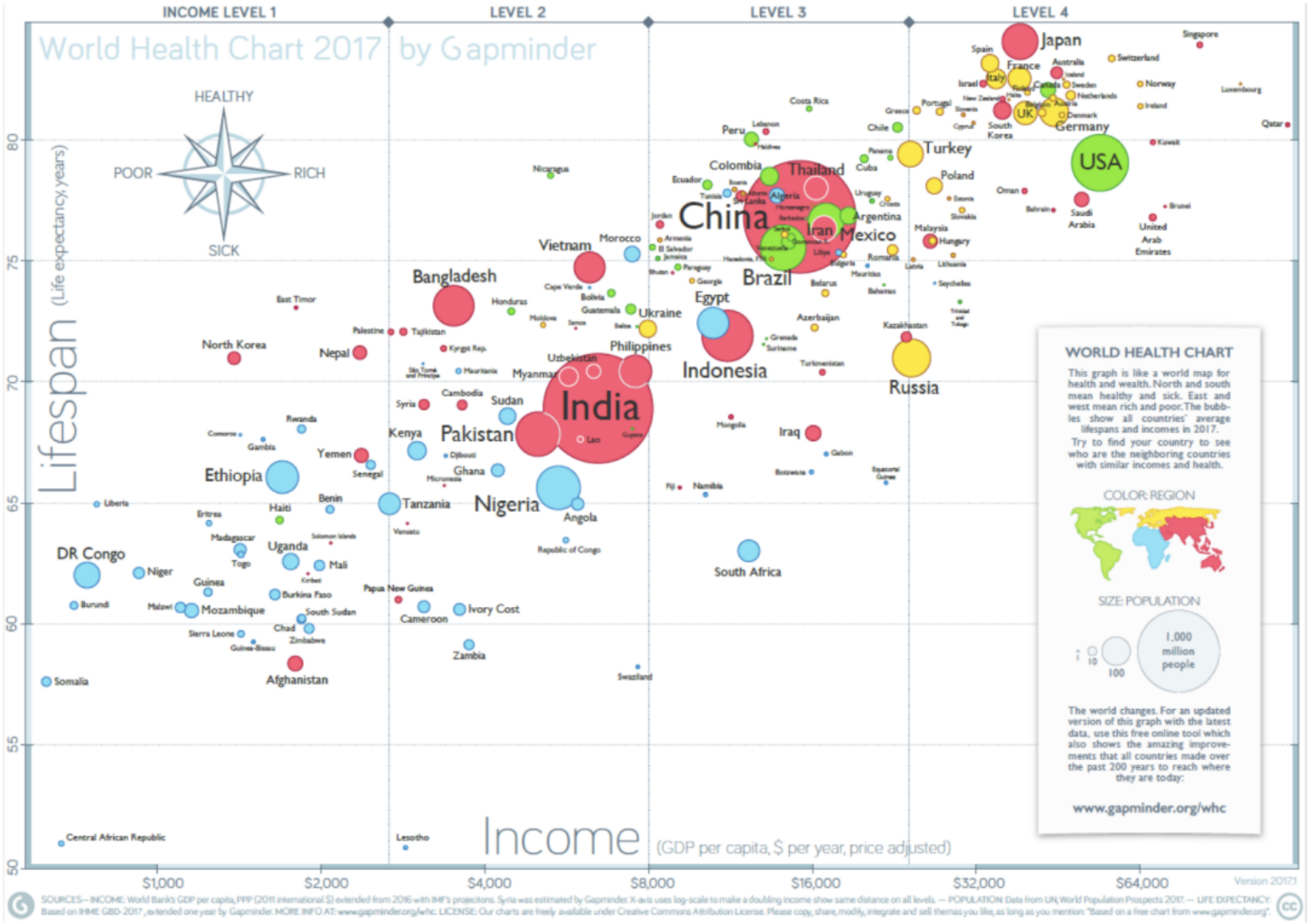
Given the recent explosion in data availability and visualization tools, one might assume that data visualization is a modern development.

However, this is not the case — data visualization has evolved over time, building on earlier methods to shape the tools and trends we use today.

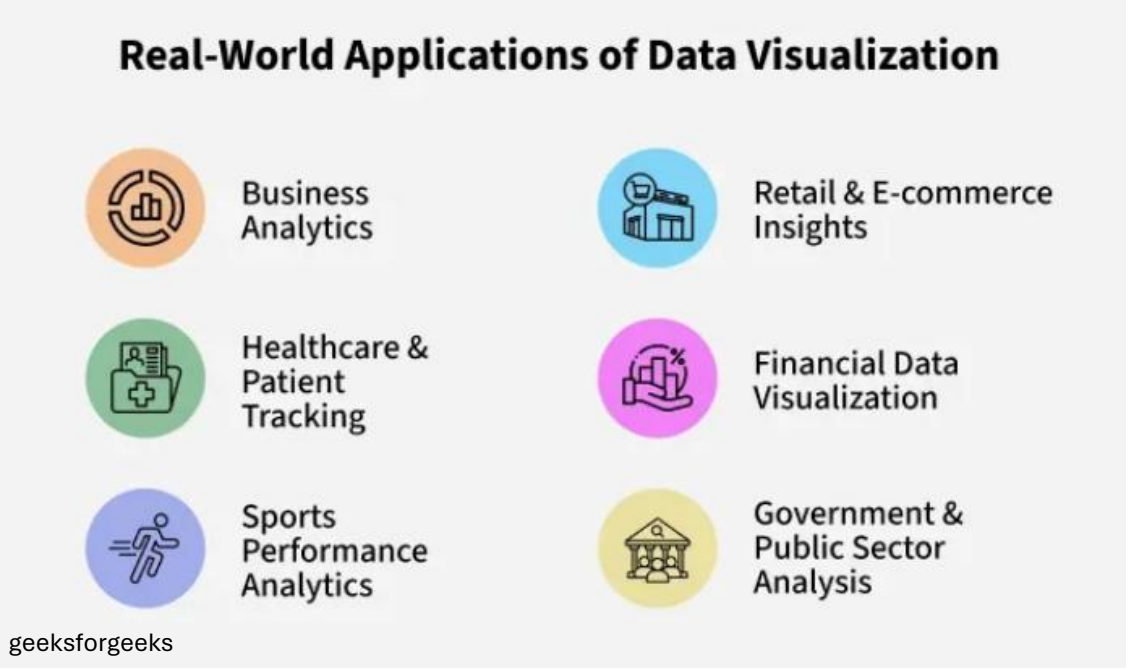
For example, as early as the 18th century.



Now


INCOME LEVEL 1
LEVEL 2
LEVEL 3
LEVEL 4
World Health Chart 2017 by Gapminder
HEALTHY
POOR
RICH
SICK
Lifespan (Life expectancy, years)
80
75
70
65
60
55
50
Japan
Singapore
Switzerland
France
Spain
Italy
Australia
Norway
Sweden
Netherlands
Germany
UK
South Korea
Qatar
Kuwait
Turkey
USA
Poland
Oman
Saudi Arabia
Brunei
United Arab Emirates
Malaysia
Hungary
Peru
Costa Rica
Chile
Colombia
Thailand
Cuba
Ecuador
China
Argentina
Iran
Mexico
Brazil
Vietnam
Morocco
Bangladesh
Egypt
Ukraine
Philippines
Indonesia
Russia
Kazakhstan
Azerbaijan
Belarus
Nicaragua
Honduras
Bolivia
Guatemala
Palestine
Tajikistan
North Korea
Nepal
East Timor
Uzbekistan
Myanmar
India
Cambodia
Syria
Sudan
Kenya
Pakistan
Yemen
Ethiopia
Senegal
Ghana
Rwanda
Comoros
Gambia
Liberia
Haiti
Benin
Tanzania
Nigeria
Angola
Eritrea
Madagascar
Uganda
Togo
Mali
DR Congo
Niger
Guinea
Burkina Faso
Burundi
Malawi
Mozambique
South Sudan
Sierra Leone
Chad
Zimbabwe
Guinea-Bissau
Afghanistan
Somalia
Papua New Guinea
Cameroon
Ivory Cost
Zambia
Republic of Congo
Swaziland
Namibia
South Africa
Iraq
Mongolia
Gabon
Botswana
Equatorial Guinea
Fiji
Lesotho
Central African Republic
Income
(GDP per capita, $ per year, price adjusted)
$1,000
$2,000
$4,000
$8,000
$16,000
$32,000
$64,000
Version 20171
WORLD HEALTH CHART
This graph is like a world map for health and wealth. North and south mean healthy and sick. East and west mean rich and poor. The bubbles show all countries' average lifespans and incomes in 2017.
Try to find your country to see who are the neighboring countries with similar incomes and health.
COLOR: REGION
SIZE: POPULATION
1
10
100
1,000 million people
The world changes. For an updated version of this graph with the latest data, use this free online tool which also shows the amazing improvements that all countries made over the past 200 years to reach where they are today:
www.gapminder.org/whc
SOURCES— INCOME: World Bank's GDP per capita, PPP (2011 international $) extended from 2016 with IMF's projections. Syria was estimated by Gapminder. X-axis uses log-scale to make a doubling income show same distance on all levels. — POPULATION: Data from UN, World Population Prospects 2017. — LIFE EXPECTANCY: Based on IHME GBD-2017, extended one year by Gapminder. MORE INFO AT: www.gapminder.org/whc. LICENSE: Our charts are freely available under Creative Commons Attribution License. Please copy, share, modify, integrate and sell themas you like, as long as you mention: "Based on a free chart from www.gapminder.org".

Real world application

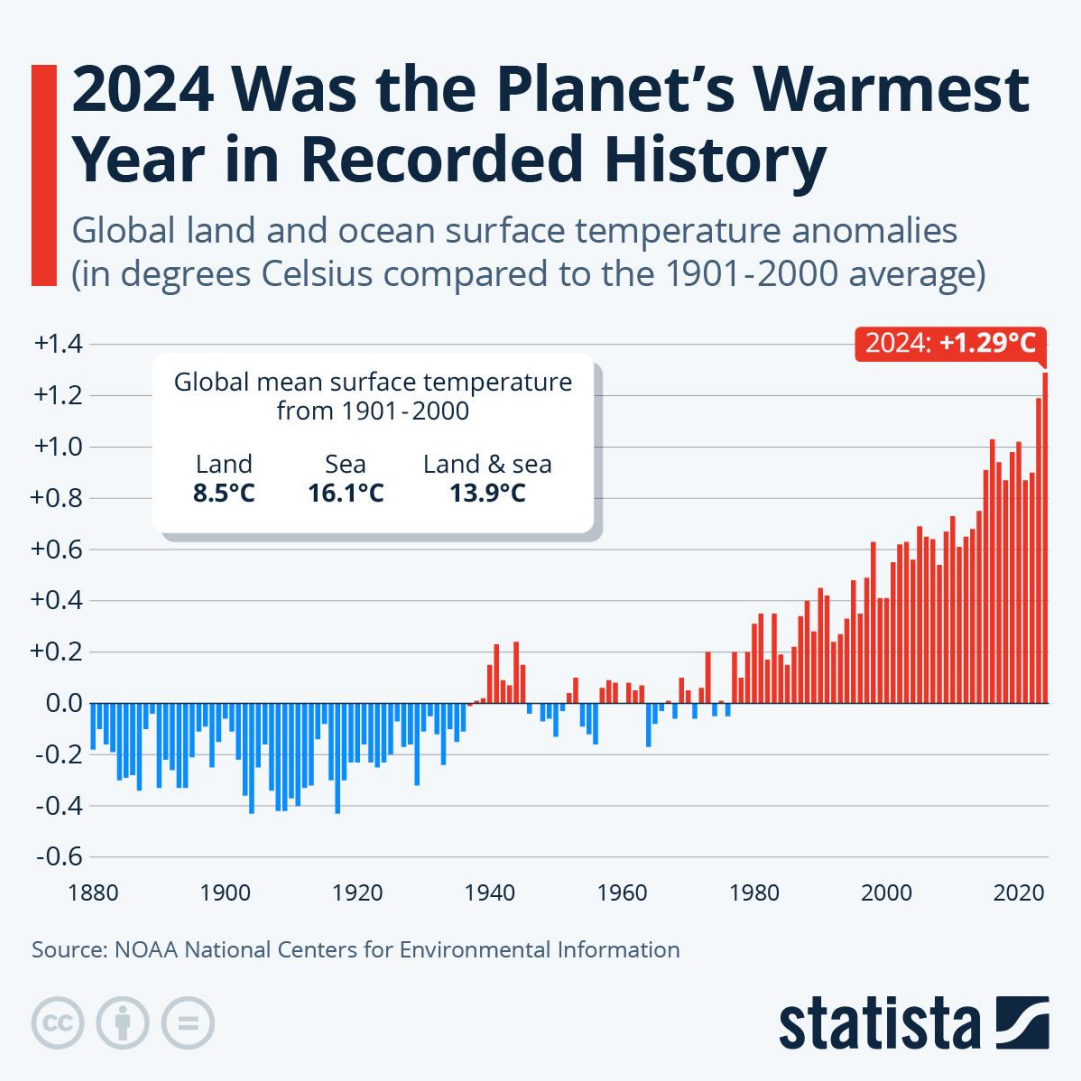


🌍 Public Policy	Track climate change trends and carbon emissions
💼 Business	Sales dashboards, customer behavior
🏥 Health	Epidemic tracking, patient diagnostics, drug effectiveness
📈 Finance	Stock market trends, risk visualization, fraud detection
⚖️ Human Rights	Visualizing refugee flows or conflict zones
⚙️ Engineering	Sensor data monitoring, anomaly detection
🧠 Research	Experimental results, network analysis
📰 Journalism	Data stories (e.g., elections, education access)

Importance of data visualisation

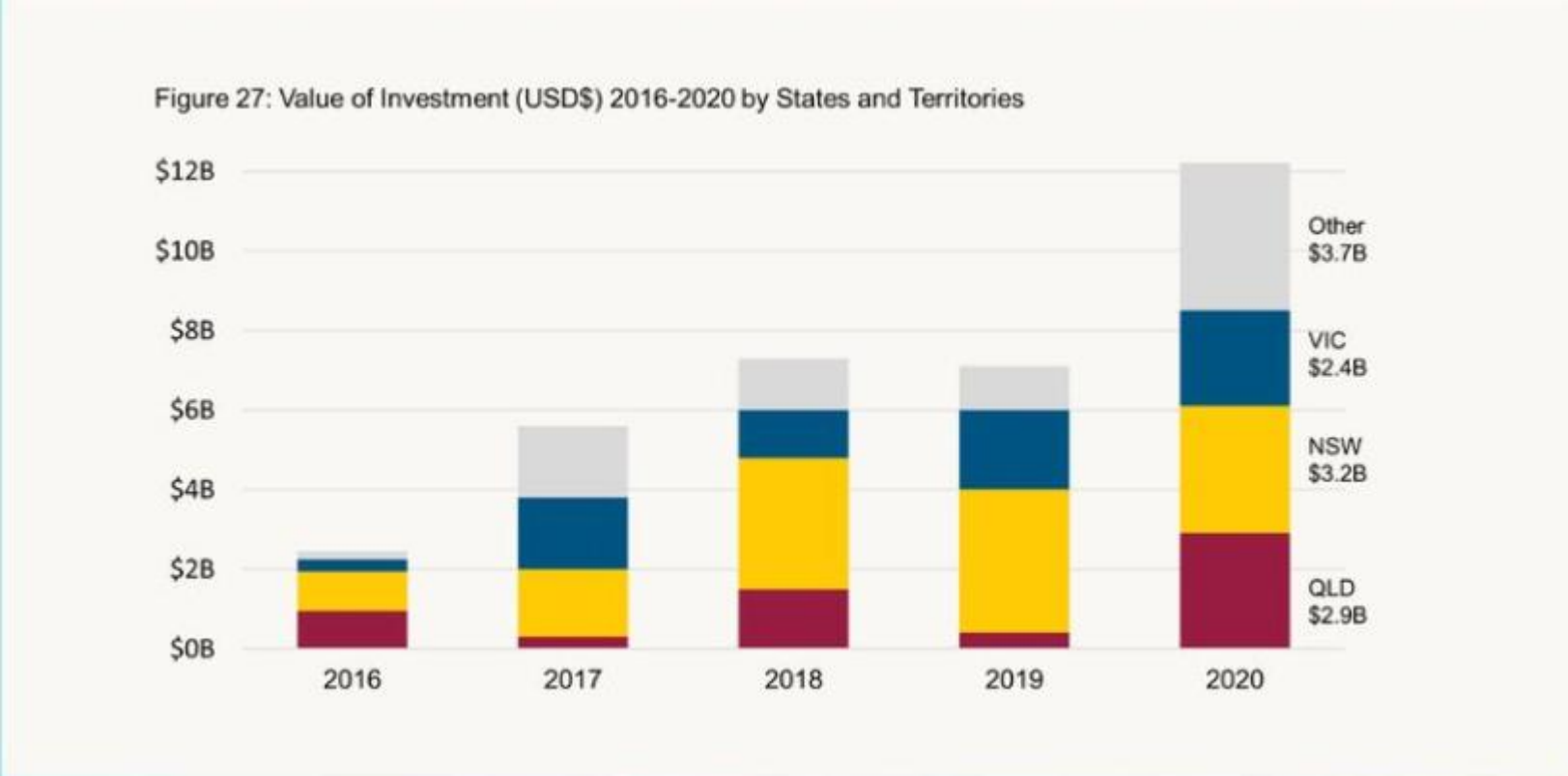
- **Simplifies Complex Data:** Converts large and complicated datasets into charts and graphs that are easier to understand.
- **Highlights Patterns and Trends:** Makes it easier to spot relationships and trends that are hard to see in raw numbers.
- **Saves Time:** Enables quick interpretation of data, reducing the need to scan through lengthy tables.
- **Enhances Communication:** Helps share insights clearly, even with people who aren't data experts.
- **Tells a Clear Story:** Guides the viewer through the data logically, supporting better understanding and decisions..

Importance of data visualisation



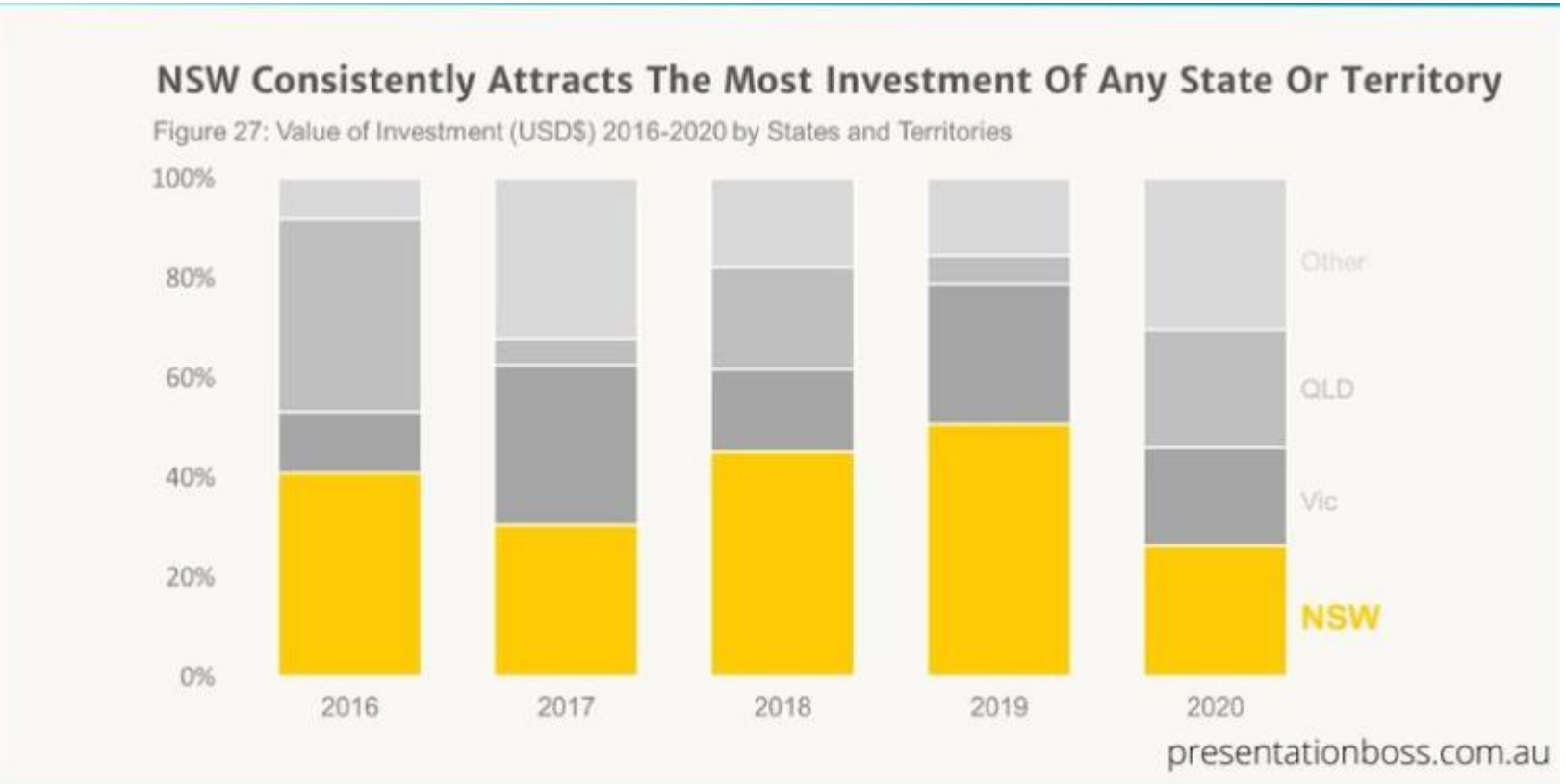
Importance of data visualisation

What is the message?

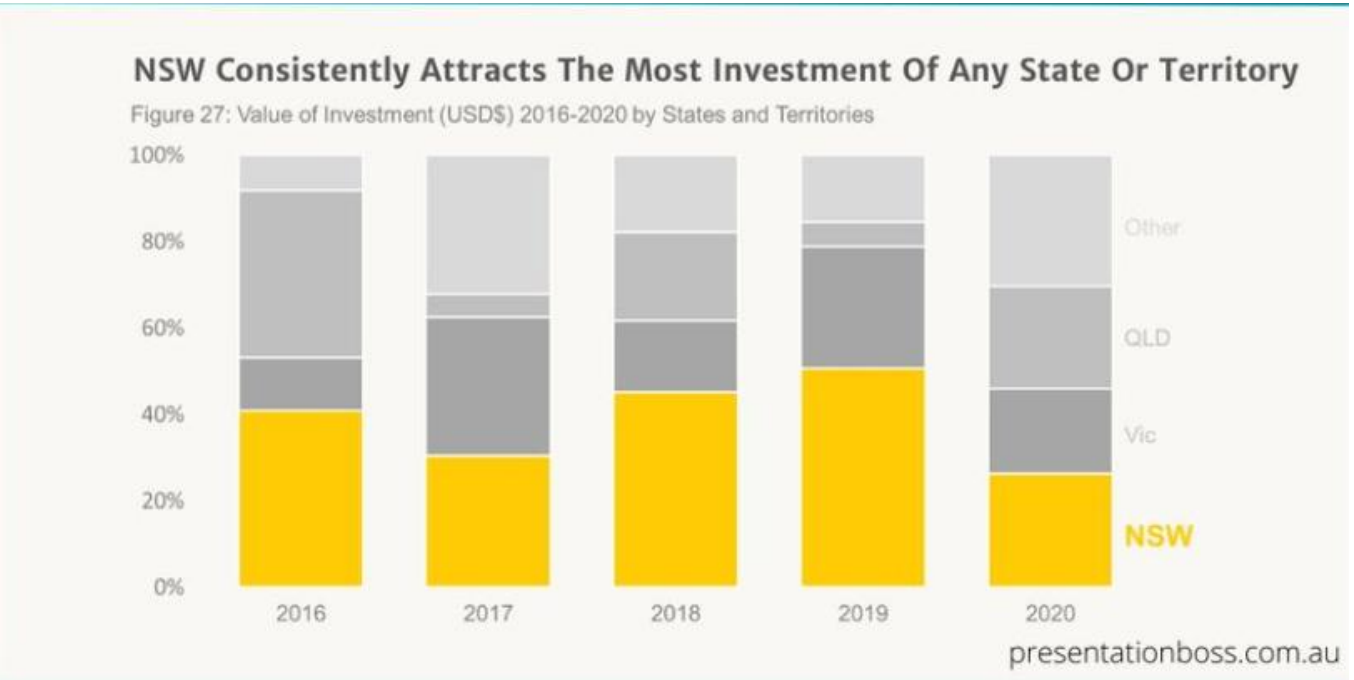


Importance of data visualisation

What is the message?



Importance of data visualisation

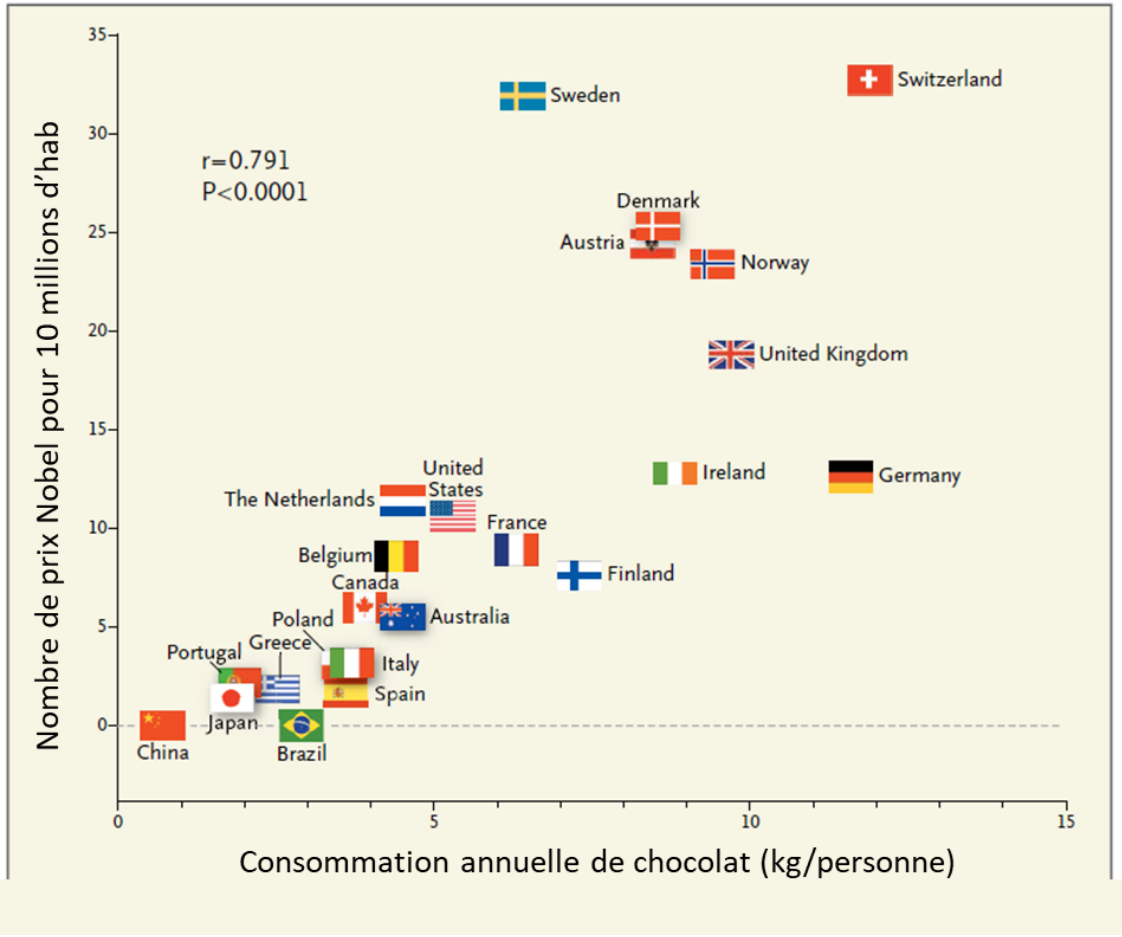


- **Informative** The visualization should be able to convey the desired information from the data to the reader.
- **Efficient** The visualization should not be ambiguous.
- **Appealing** The visualization should be captivating and visually pleasing.

Cons

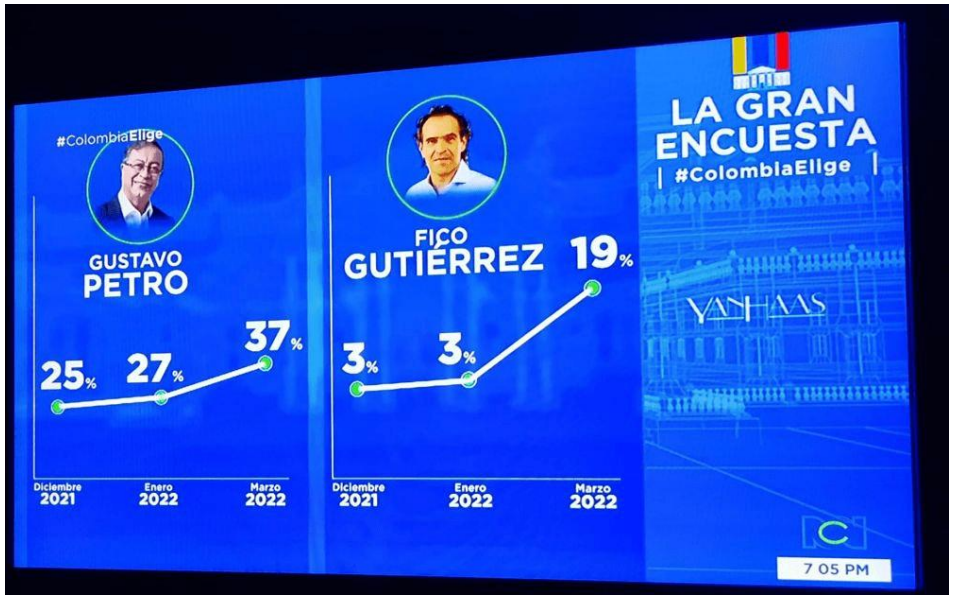
Visualization has real risks:

- **Misleading design**: truncated axes, cherry-picked data, 3D effects
- **Oversimplification**: hides uncertainty or nuance
- **Cognitive bias**: viewers may interpret visuals emotionally
- **False authority**: “it looks professional, so it must be true…”
- **Accessibility issues**: poor color choices exclude color-blind users
- **Correlation shown as causation**

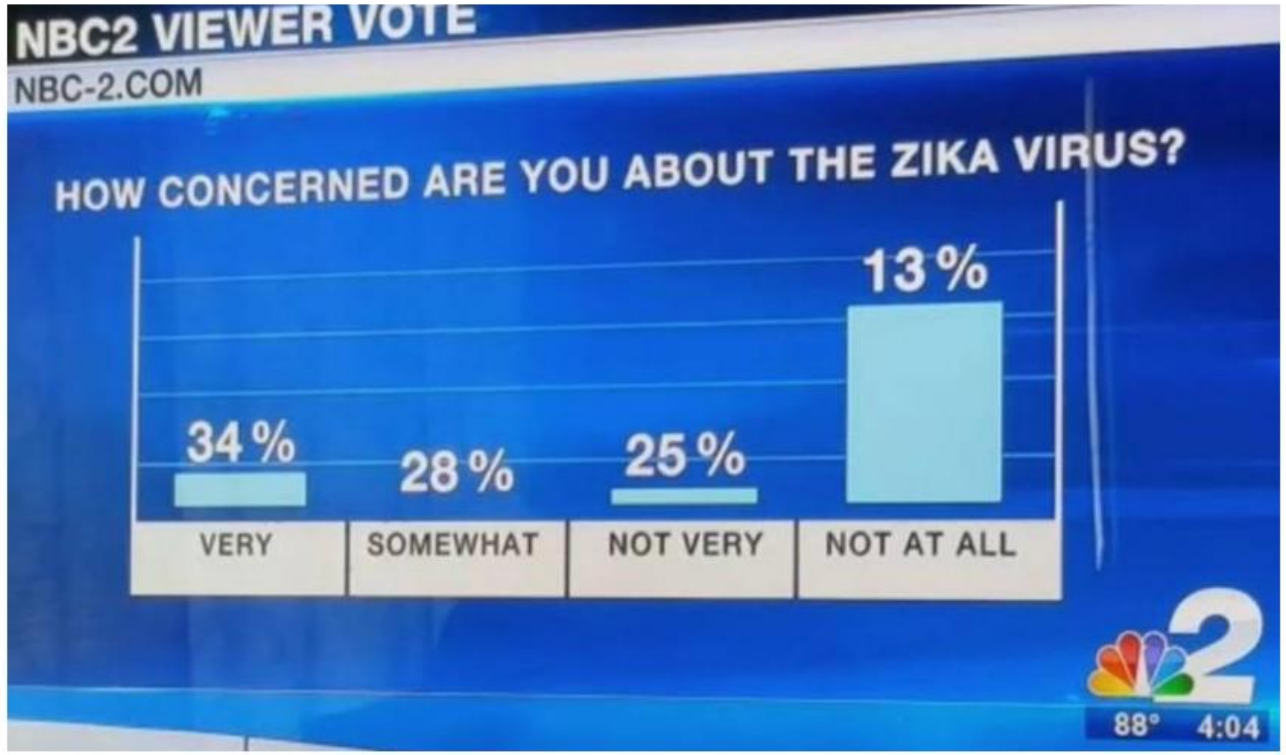


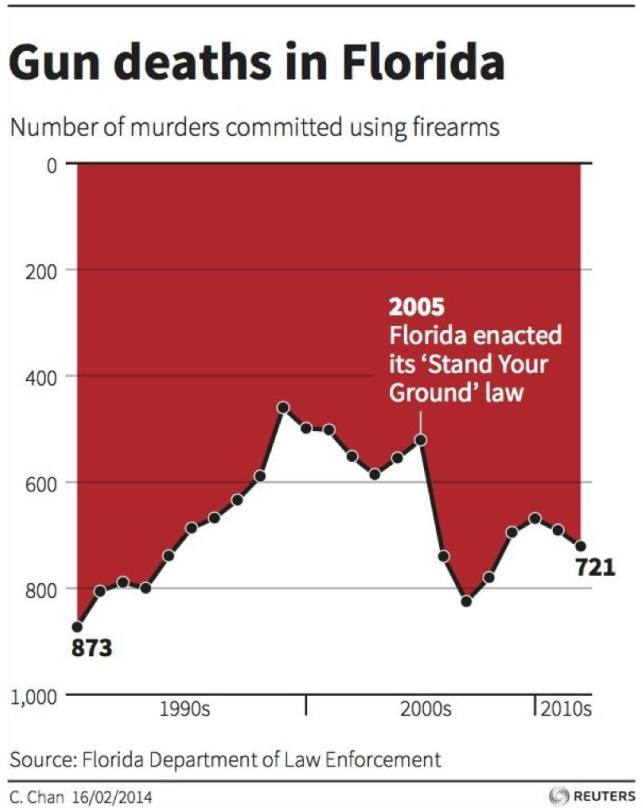
When Visuals Misled

When Visuals Mislead



When Visuals Mislead





Before You Plot — Structure Your Data

You can’t visualize what you haven’t cleaned!

What is Data Structuring?
(Also called *data wrangling*, *data preparation*, or *data engineering*)

- Remove duplicates and errors (e.g. misspelling)
- Handle missing values (drop, fill, interpolate)
- Convert formats (e.g. dates, categories)
- Normalize or scale values
- Reshape data (pivot, melt) for plotting

Types of Variables – What Kind of Data Are We Dealing With?

Two Main Types of Variables:

1. Categorical (Qualitative)

- Describes qualities or categories
- Cannot be meaningfully measured or averaged

Examples:

- Gender
- Eye color

Subtypes:

- Nominal: no order (e.g., eye color)
- Ordinal: with order (e.g., satisfaction level)

2. Numerical (Quantitative)

- Represents measurable quantities
- Can be used in mathematical operations

Examples:

- Age
- Height

Subtypes:

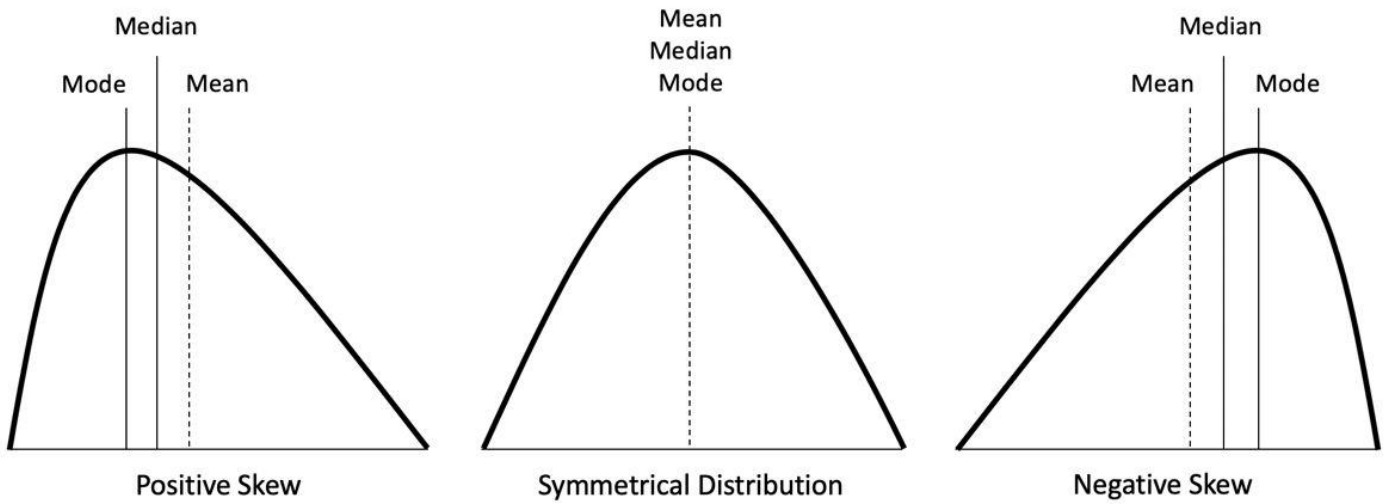
- Discrete: countable (e.g., exams passed)
- Continuous: measurable (e.g., weight, time)

Mean, median 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 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?

Mean, median 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? **4.38**
- What is the median? **3**
- What is the mode? **2**

How do these values change if we replace the last 9 with the number 100? **Mean changes to 11.38, mode and median the same**

Percentages, Percentiles & Quartiles – What's the Difference?

Percentage (%):

- A way to express a portion of a whole (out of 100)
- Example: 60% of survey respondents chose Option A
- Used in Pie Chart

Percentile

- A value below which a given percentage of data falls
- Example: 90th percentile = better than 90% of the group
- Used in standardized tests, income distribution, etc.

Quartiles

Special percentiles that divide data into four equal parts:

- Q1 (25th percentile) – 25% of values are below this
- Q2 (50th percentile) – the median
- Q3 (75th percentile) – 75% of values are below this
- Used to describe spread in boxplots

Percentages, Percentiles & Quartiles – What's the Difference?

Percentage (%):

- A way to express a portion of a whole (out of 100)
- Example: 60% of survey respondents chose Option A
- Used in Pie Chart

Percentile

- A value below which a given percentage of data falls
- Example: 90th percentile = better than 90% of the group
- Used in standardized tests, income distribution, etc.

Quartiles

Special percentiles that divide data into four equal parts:

- Q1 (25th percentile) – 25% of values are below this
- Q2 (50th percentile) – the median
- Q3 (75th percentile) – 75% of values are below this
- Used to describe spread in boxplots

Ex: weights in the family: 25, 17, 32, 11, 40, 35, 13, 5, and 46.

- Calculate the percentages of people with a weights greater than the average
- Calculate the percentages of people with a weights greater than the median
- Calculate the lower and upper quartiles
- Calculate the 90th percentile

Percentages, Percentiles & Quartiles – What's the Difference?

Percentage (%):

- A way to express a portion of a whole (out of 100)
- Example: 60% of survey respondents chose Option A
- Used in Pie Chart

Percentile

- A value below which a given percentage of data falls
- Example: 90th percentile = better than 90% of the group
- Used in standardized tests, income distribution, etc.

Quartiles

Special percentiles that divide data into four equal parts:

- Q1 (25th percentile) – 25% of values are below this
- Q2 (50th percentile) – the median
- Q3 (75th percentile) – 75% of values are below this
- Used to describe spread in boxplots

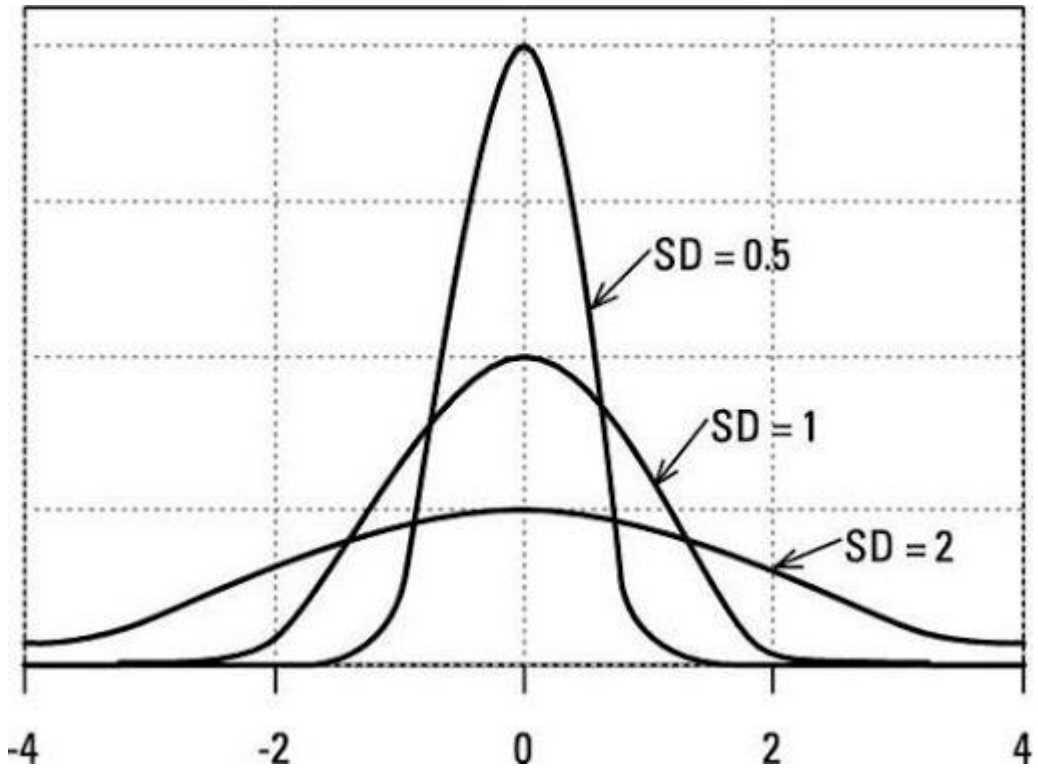
Ex: weights in the family: 25, 17, 32, 11, 40, 35, 13, 5, and 46.

- Calculate the percentages of people with a weights greater than the average: **55%**
- Calculate the percentages of people with a weights greater than the median: **44%**
- Calculate the lower and upper quartiles of the following: **12,37.5**
- Calculate the 90th percentile: **Pk=k(N+1)/100 -> P90=90*10/100=9th -> 46**

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.



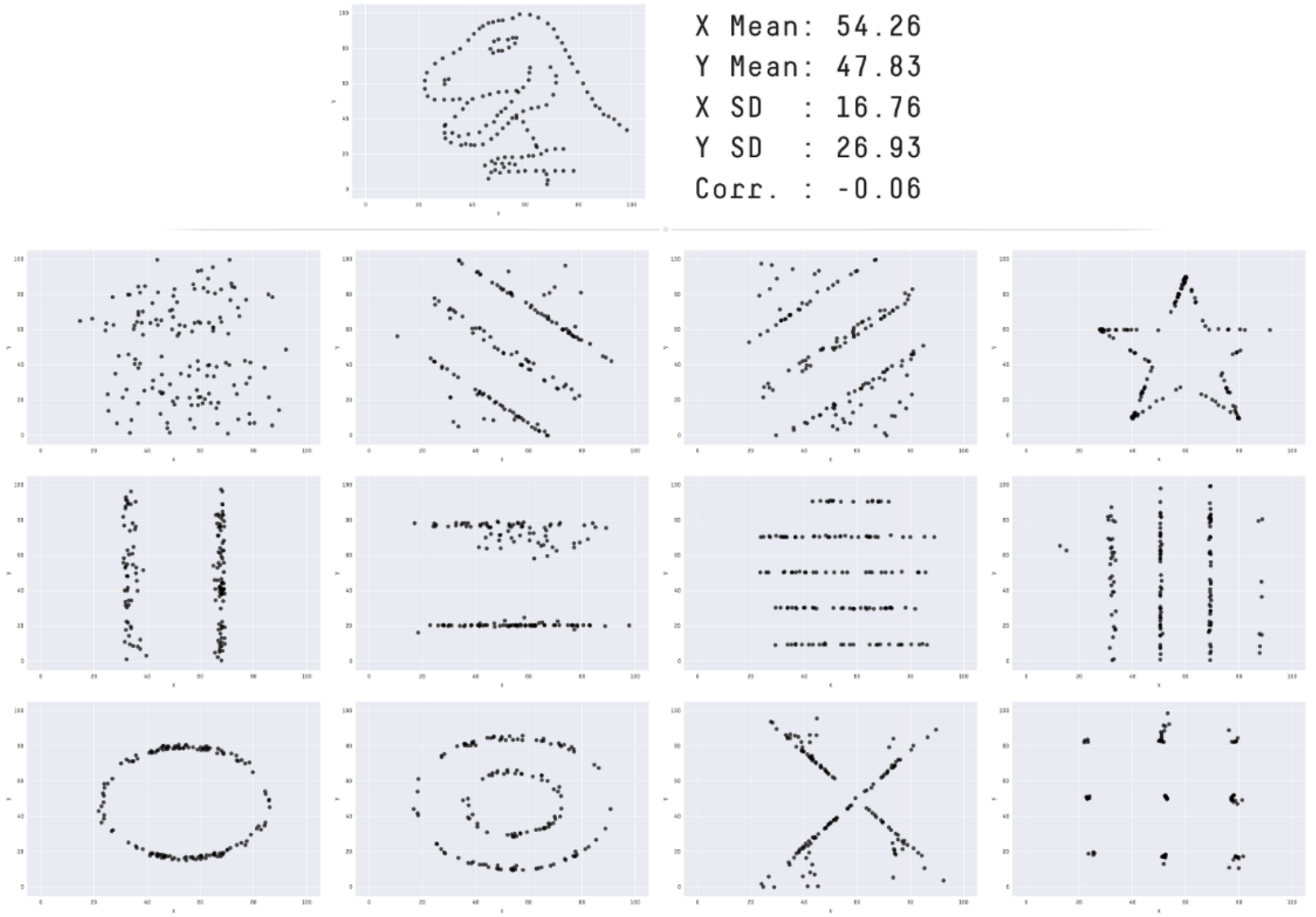
$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

n is the number of persons

x_i is the size of the individual

$\bar{x}$ is the mean value of all persons

Attention



Matejka et al. 2023

Probabilities

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

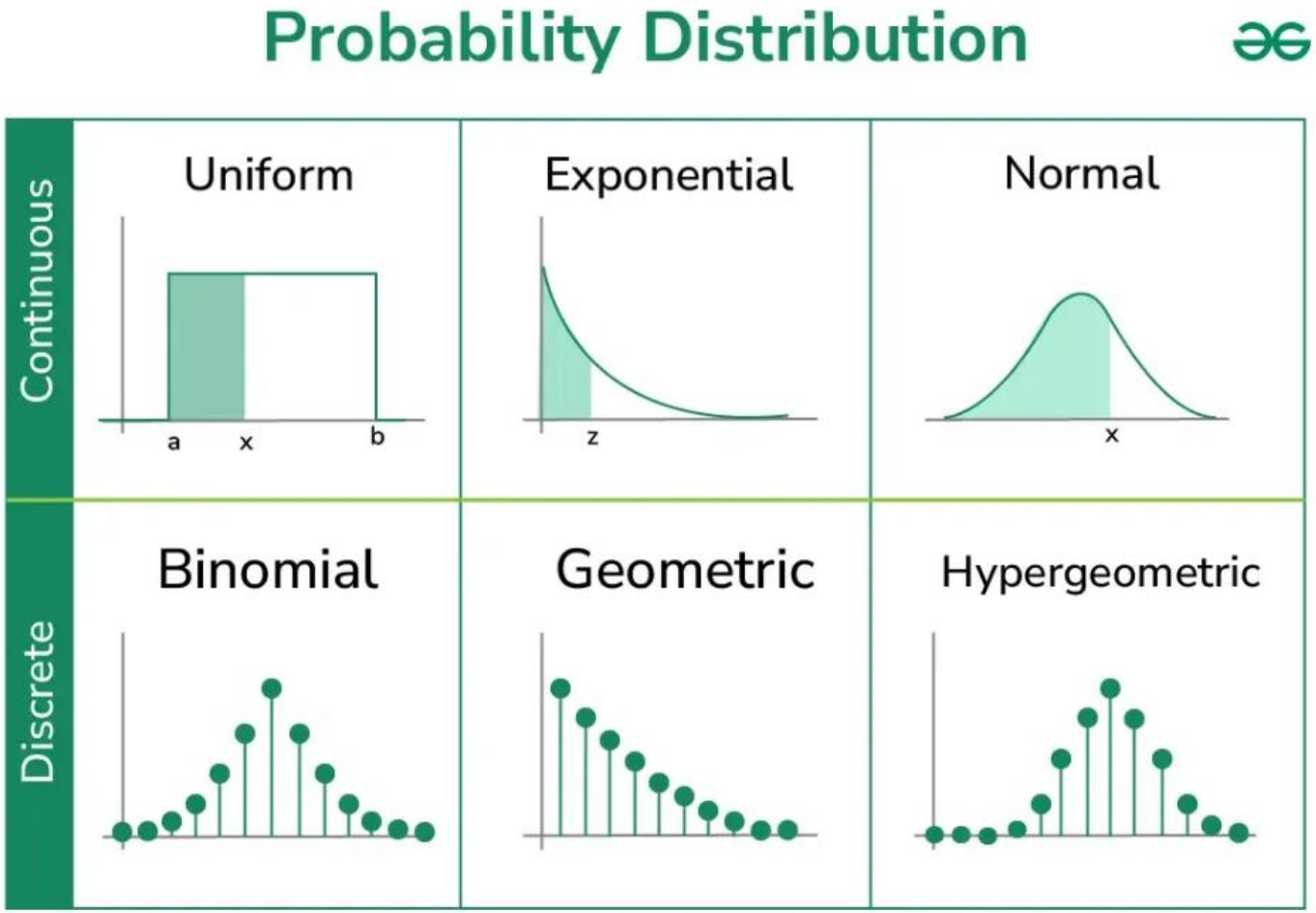
Probability is a number between **0 and 1**
0 = impossible, **1 = certain**

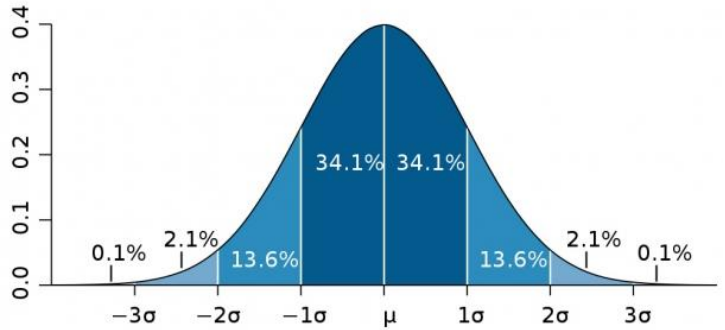
The probability of an event can therefore be measured as the number of favorable outcomes divided by the total number of possible outcomes.

Event	Probability
Flipping a coin – Heads	0.5
Rolling a 6 on a die	$1/6 \approx 0.167$
Drawing an ace from a deck	$4/52 \approx 0.077$

Distributions

The distribution of observed probabilities can take several forms. It provides a way of **modeling the likelihood of each outcome in a random experiment.**





68% of values fall within ±1σ of the mean

~95% within ±2σ

~99.7% within ±3σ

Population vs Sample – Who Are We Studying?

Definition

In statistics, a population is the set on which a statistical study is conducted. A sample is a (usually known) subset of the population

Population
→ The **entire group** we are interested in studying
Examples: all French voters, all EPITA students, all galaxies in the observable universe

Sample
→ A **subset** of the population that we actually collect data from
Examples: 1,000 survey respondents, 50 students, 10,000 galaxies from SDSS

Why use a sample?
- **Faster**
- **Cheaper**
- Often **more practical or ethical**

Goal: Make **valid conclusions** about the population **from** the sample

Project: where to get data?

Explore the following websites to find one dataset that you would like to work on:

- https://data.europa.eu/en
- https://data.worldbank.org/
- http://data.un.org/
- https://datasetsearch.research.google.com/
- https://github.com/datasets
- https://opendata.paris.fr/pages/home/
- https://www.data.gouv.fr/fr/

You can also identify a topic you would be interested to work on

Project: what is a licence?

“Licences tell you **what you can do** with the content or data that you access.

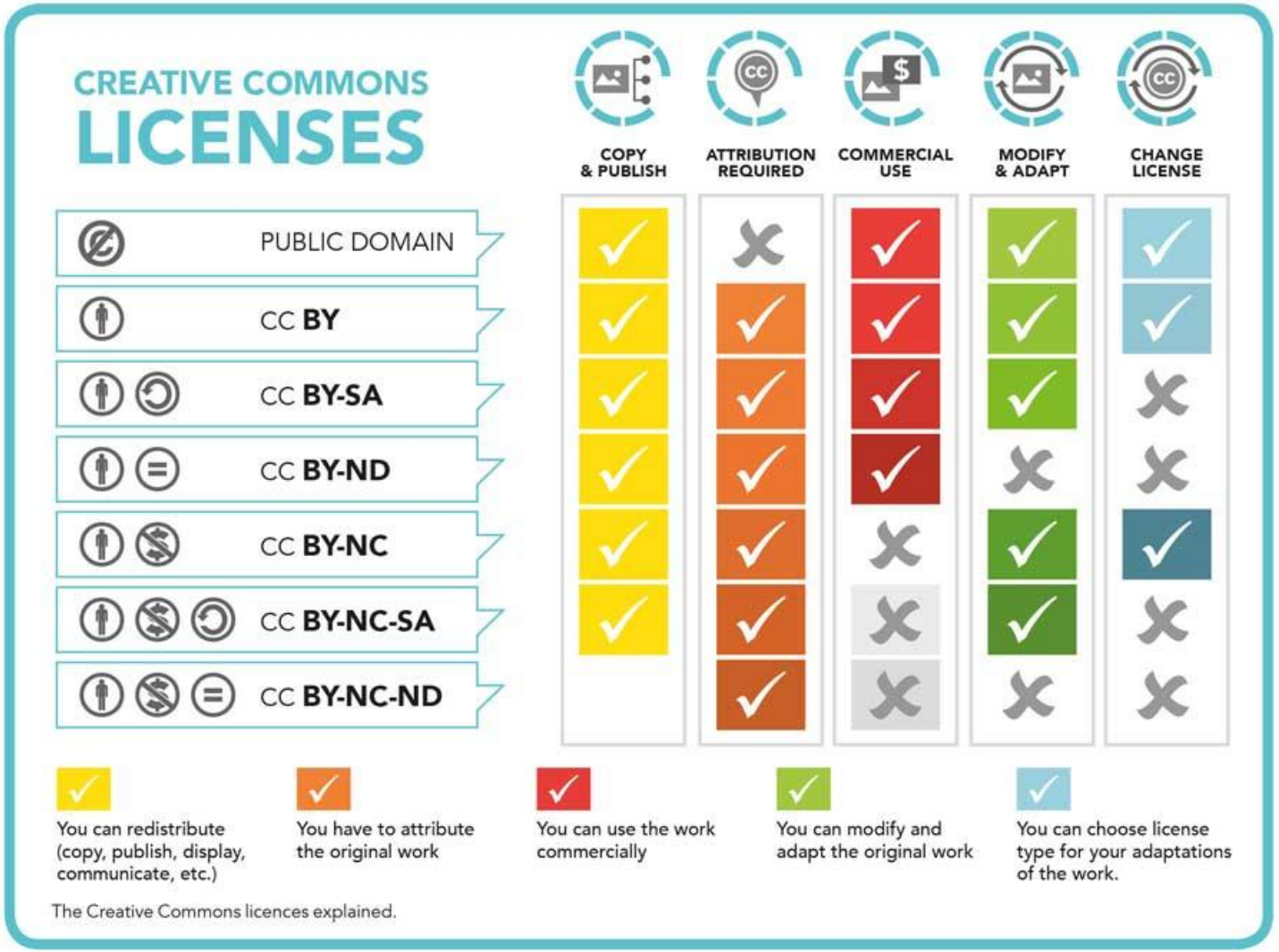
A licence will tell you whether you can:

- republish the content or data on your own website
- derive new content or data from it
- make money by selling products that use it
- republish it while charging a fee for access

If you break the terms of the licence, the owner of the content or data can take you to court.”

ttps://data.europa.eu/sites/default/files/d2.1.2_training_module_2.5_data_and_metadata_licensing_en_edp.pdf

Project: type of licence



More infos at: https://creativecommons.org/

Project: General Data Protection Regulation

All collected, stored and processed data of european citizen need to adhere to the GDPR.
Key aspects of the GDPR are:

- **Data Minimization:** Collect only the data necessary for the intended purpose.
- **Lawful Processing:** Data processing must have a legal basis (consent, contract, legal obligation, vital interests, public task, legitimate interests).
- **Transparency:** Inform individuals about data processing activities.
- **Data Security:** Implement measures to protect data from breaches. Encourage privacy by design.
- **Individual Rights:** Individuals have the right to access, rectify, erase, and object to the processing of their data.

Non-compliance can result in significant fines, with fines depending on the severity of the violation.

Project: what to do after gathering data

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 ?

Project: Research question

Every project should have a research question. It is a question that you are curious about and that you want to answer. Simply put, a research question gives you a North star, a sense of direction

It should be:

- Focused on **a single problem or issue**
- **More about explanation than description**
- **Researchable**: using primary and/or secondary sources
- **Feasible** to answer within the timeframe (project deadline is in a few weeks)
- Specific enough to answer thoroughly
- Complex enough to develop the answer: a simple Google search should not answer the question!!

Example: Subject: Men's Long Jump World Records

- What does the progression of world records look like over time?
- Is the progression steady, or are improvements linked to external factors like better training, nutrition, or technology?
- How does the men's long jump world record progression compare to the women's progression?

Project: Research question

Useful Guiding Questions

1. What are the main features of X?
2. In what ways has X evolved over time?
3. What factors contribute to the existence of X?
4. How does X connect with Y?
5. What effects does X have on Y?
6. What are the strengths and weaknesses of X?
7. What steps could make X better?

AVOID:

- **not be based on value judgements** (Is X better than Y?)

- **vague language, jargon, and too-broad ideas.**

Bad: What effect does social media have on people's minds?
Good: What effect does daily use of tiktok have on the attention span of 16-year-olds at high school?

- **not be answered with yes or no**

Project: Research question

Once you've defined your research question, the next step is to organize your analysis.

In most cases, a single overarching question is too broad to tackle all at once. Breaking it down into smaller, focused subquestions allows you to approach the problem methodically and build up to a comprehensive answer step by step.

Helpful Tips

Clear sub-questions should be:

1. Easier to address than the main question
2. Focused on a single aspect
3. Organized in a sensible sequence

Project: Research question

YOUR TURN

- **Select a dataset**
- **Explore the dataset**
- **Write a research question**
- **Define sub questions**