Practical AI
02 – Data & Data Visualization

Emanuele Fabbiani


Data

What is data?

Information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer. [1]

[1] https://dictionary.cambridge.org/it/dizionario/inglese/data

HISTORY OF DATA

19,000 BC

The Ishango bone holds the first evidence of data collection and storage.

1600s

John Graunt introduces the concept of data analysis in 1663.

1800s

Herman Hollerith designs a machine that helped complete the US census in 1890.

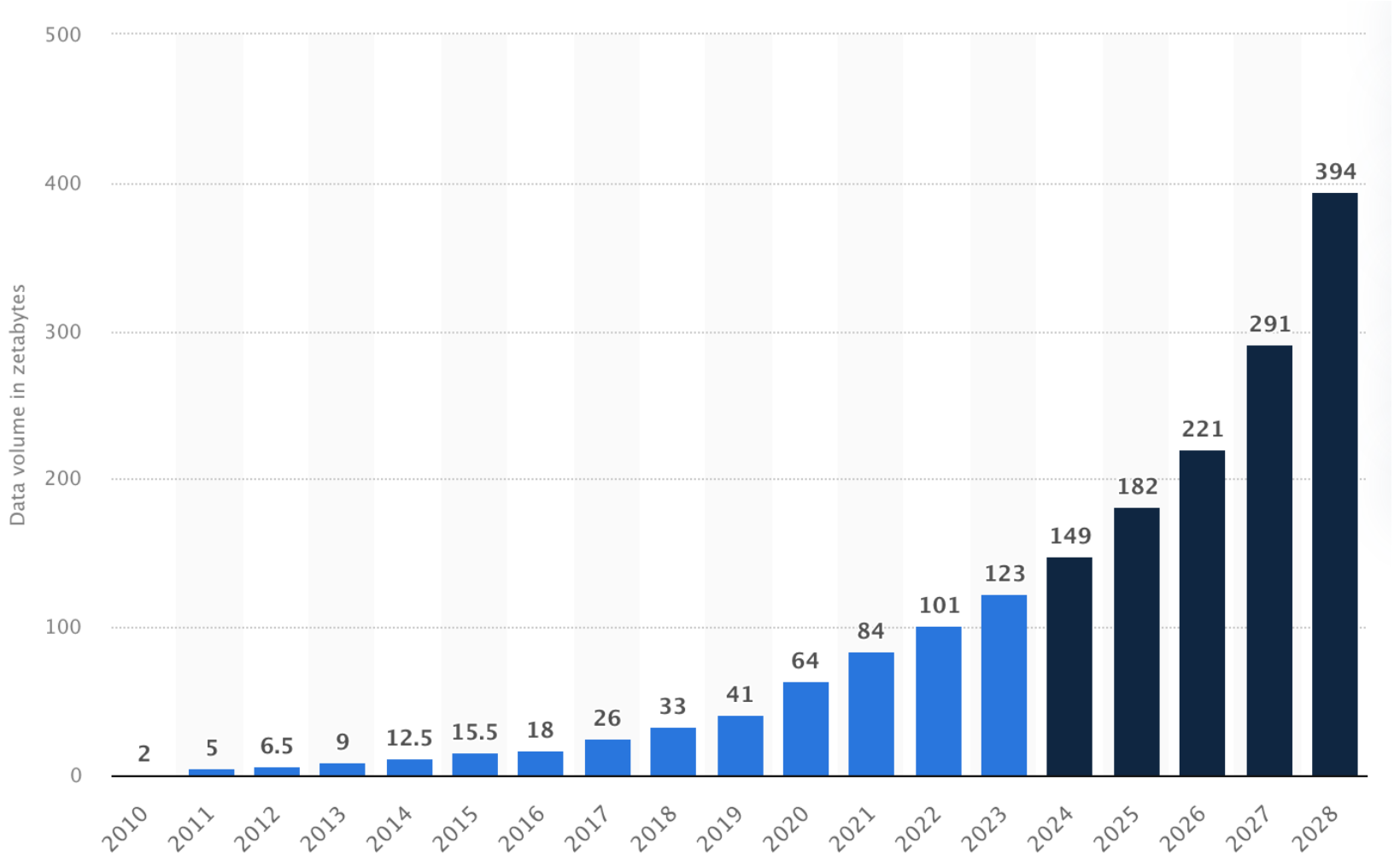
1900s

Fritz Pfleumer invents the magnetic tape which later inspired the invention of floppy disks and hard disk drives.

1990s

Sir Tim Berners Lee invents the World Wide Web.

How Much Data Do We Create?



https://www.statista.com/statistics/871513/worldwide-data-created/


The Economist
Crunch time in France
Ten years on: banking after the crisis
South Korea's unfinished revolution
Biology, but without the cells
MAY 6TH–12TH 2017
The world's most valuable resource
Google
amazon
Google
TESLA
TESLA
Data and the new rules of competition

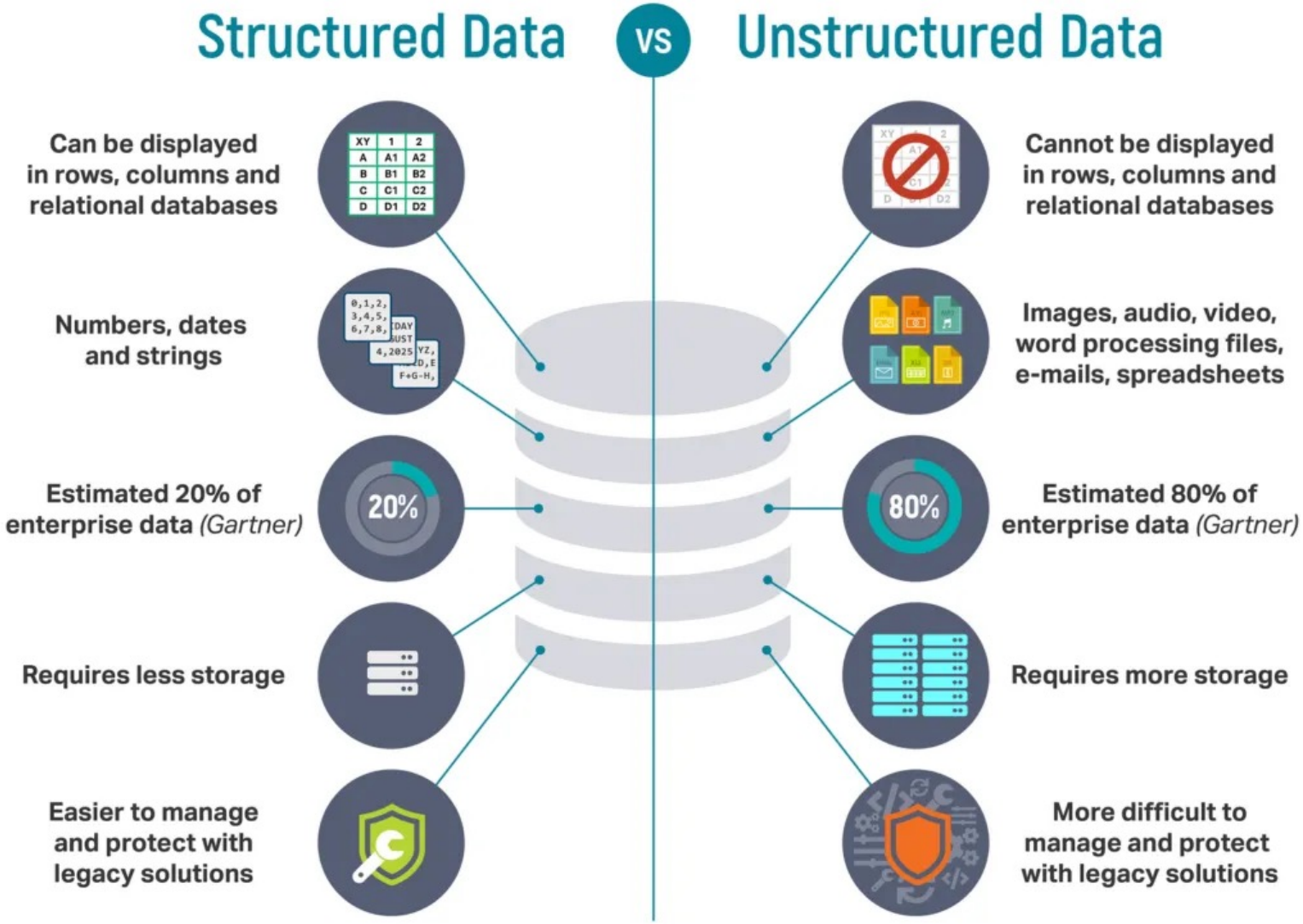
Discussion

How does Meta profit from its data?

And Revolut?

How Did Data Become Valuable?

- In the early 2000s, many businesses transitioned to digital platforms, gaining the ability to track user behaviour.
- This data became an asset, enabling companies to make informed business decisions and boost profits.
- Data-driven business models transformed industries:
  - Companies like Google and Meta **rely entirely** on data for their operations.
  - Businesses such as Amazon, Netflix, and Uber became larger and much **more profitable** due to data utilization.
  - Traditional industries like banking leveraged data to **decrease expenses**.
- The availability of hardware and strong economic incentives to harness data spurred rapid advancements in software and analytical methods.

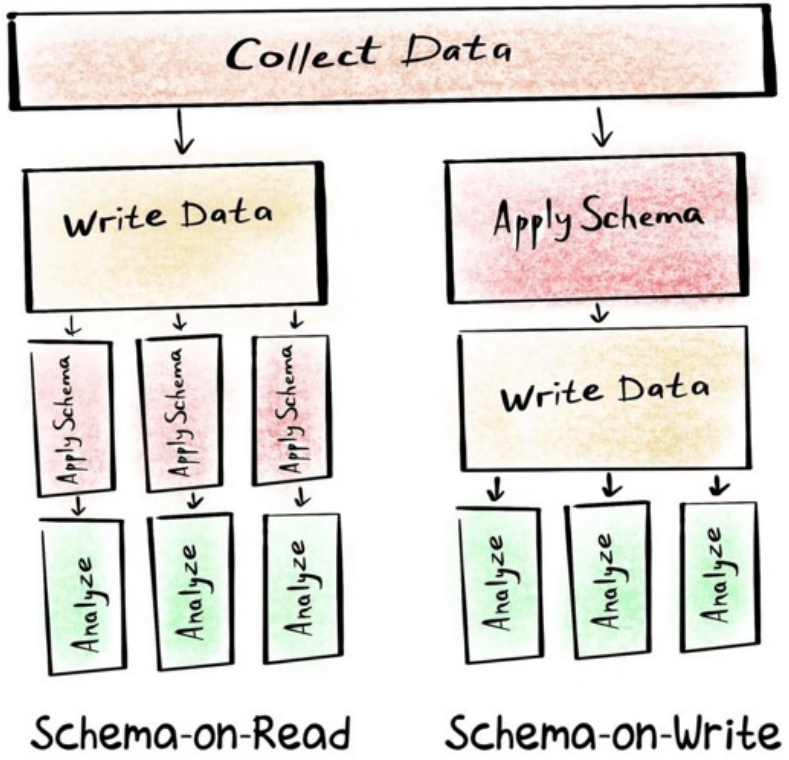


https://www.mongodb.com/resources/basics/unstructured-data/structured-vs-unstructured

Data Schemas

Structured data follow a **schema-on-write** approach, requiring data to conform to a predefined schema before it can be written.

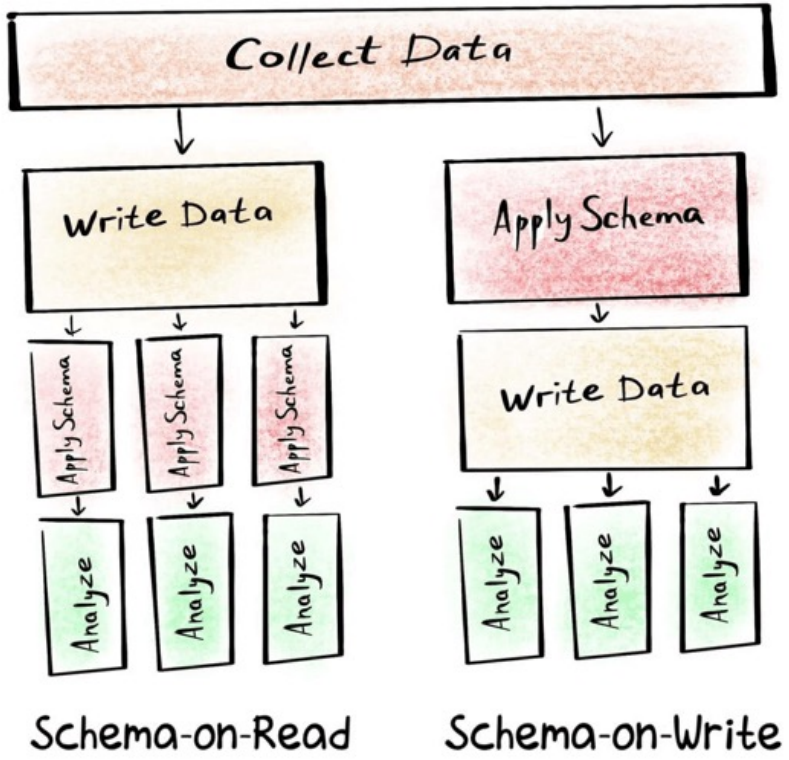
Unstructured data follow a **schema-on-read** approach, allowing data to be written in its raw form, with a suitable schema applied during reading.



Pros and Cons

Structured data are:

- **Harder to collect**, as they must fit into pre-defined schemas
- **Easier to analyse**, for the same reason
- **Easier to share** with non-technical stakeholders, as everyone can use Excel
- Historically **easier to manage**, as unstructured data only became popular in the 2010s



Data for GenAI

- Traditional machine learning (ML) models typically require relatively **small amounts** of high-quality, **structured** data.
- Generative AI (GenAI) models, by contrast, need **massive** volumes of **unstructured** data—mainly text, and more recently images, audio, and video. To collect such large datasets, companies like Google, OpenAI, and Meta have scraped large portions of the internet.
- While some of this data is open source, leading developers of state-of-the-art models (such as OpenAI, Google, and Anthropic) do not disclose the exact composition of their training datasets.
- Once a foundation model is trained, a corporate GenAI system built on a pre-trained large language model (LLM) requires **much less data**. However, this data must be of very high quality. More on this topic will be covered in the next lectures.

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao **Stella Biderman** **Sid Black** **Laurence Golding**

Travis Hoppe **Charles Foster** **Jason Phang** **Horace He**

Anish Thite **Noa Nabeshima** **Shawn Presser** **Connor Leahy**

EleutherAI
contact@eleuther.ai

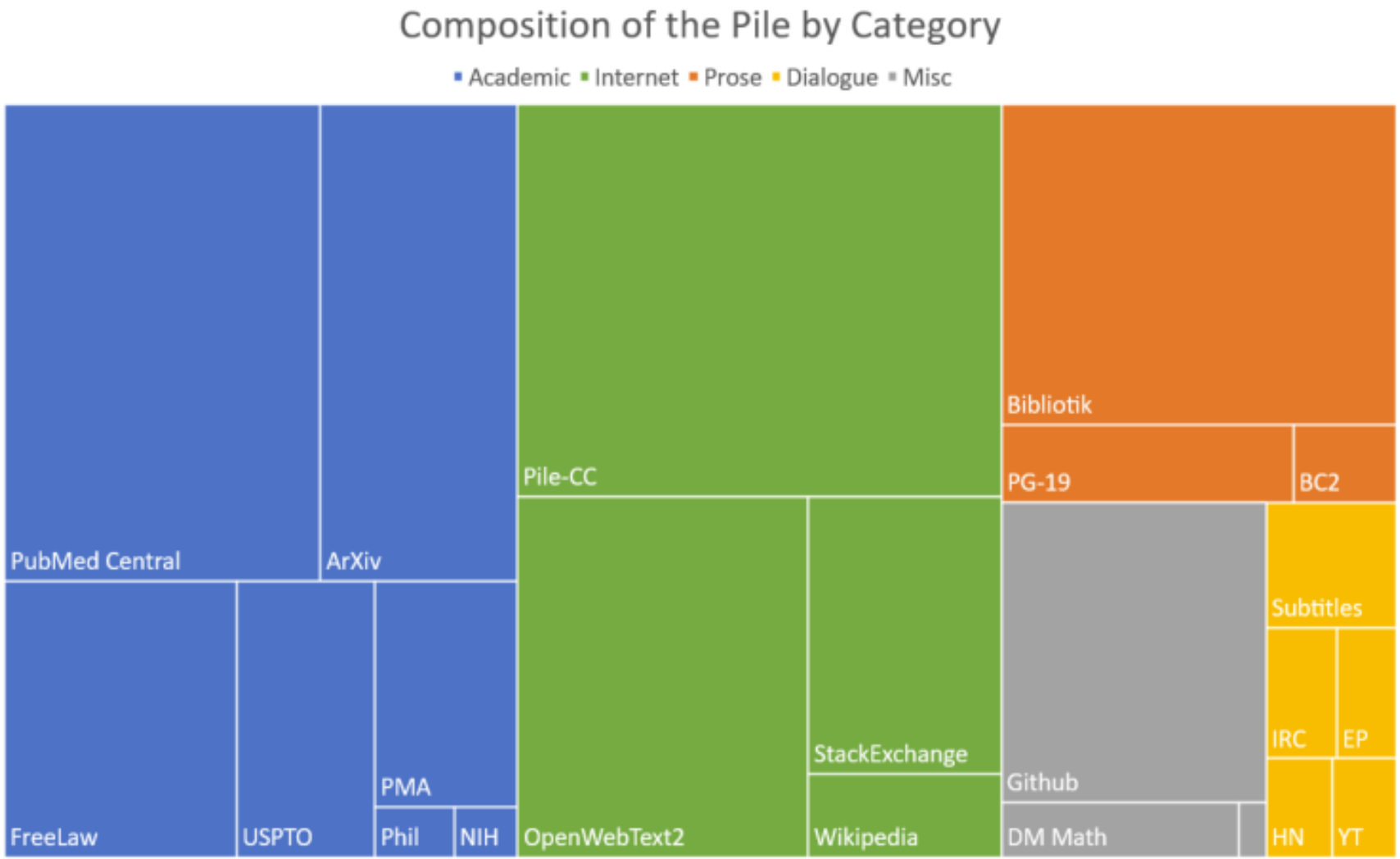
Abstract

Recent work has demonstrated that increased training dataset diversity improves general cross-domain knowledge and downstream generalization capability for large-scale language models. With this in mind, we present *the Pile*: an 825 GiB English text corpus targeted at training large-scale language models. The Pile is constructed from 22 diverse high-quality subsets—both existing and newly constructed—many of which derive from academic or professional sources. Our evaluation of the untuned performance of GPT-2 and GPT-3 on the Pile shows that these models struggle on many of its components, such as academic writing. Conversely, models trained on the Pile improve significantly over both Raw CC and CC-100 on all components of the Pile, while improving performance on downstream evaluations. Through an in-depth exploratory analysis, we document potentially concerning aspects of the data for prospective users. We make publicly available the code used in its construction.[1]

versity leads to better downstream generalization capability (Rosset, 2019). Additionally, large-scale language models have been shown to effectively acquire knowledge in a novel domain with only relatively small amounts of training data from that domain (Rosset, 2019; Brown et al., 2020; Carlini et al., 2020). These results suggest that by mixing together a large number of smaller, high quality, diverse datasets, we can improve the general cross-domain knowledge and downstream generalization capabilities of the model compared to models trained on only a handful of data sources.

To address this need, we introduce the Pile: a 825.18 GiB English text dataset designed for training large scale language models. The Pile is composed of 22 diverse and high-quality datasets, including both established natural language processing datasets and several newly introduced ones. In addition to its utility in training large language models, the Pile can also serve as a broad-coverage benchmark for cross-domain knowledge and generalization ability of language models.

https://arxiv.org/pdf/2101.00027



https://arxiv.org/pdf/2101.00027

Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training

Pierre-Carl Langlais Carlos Rosas Hinostroza Mattia Nee

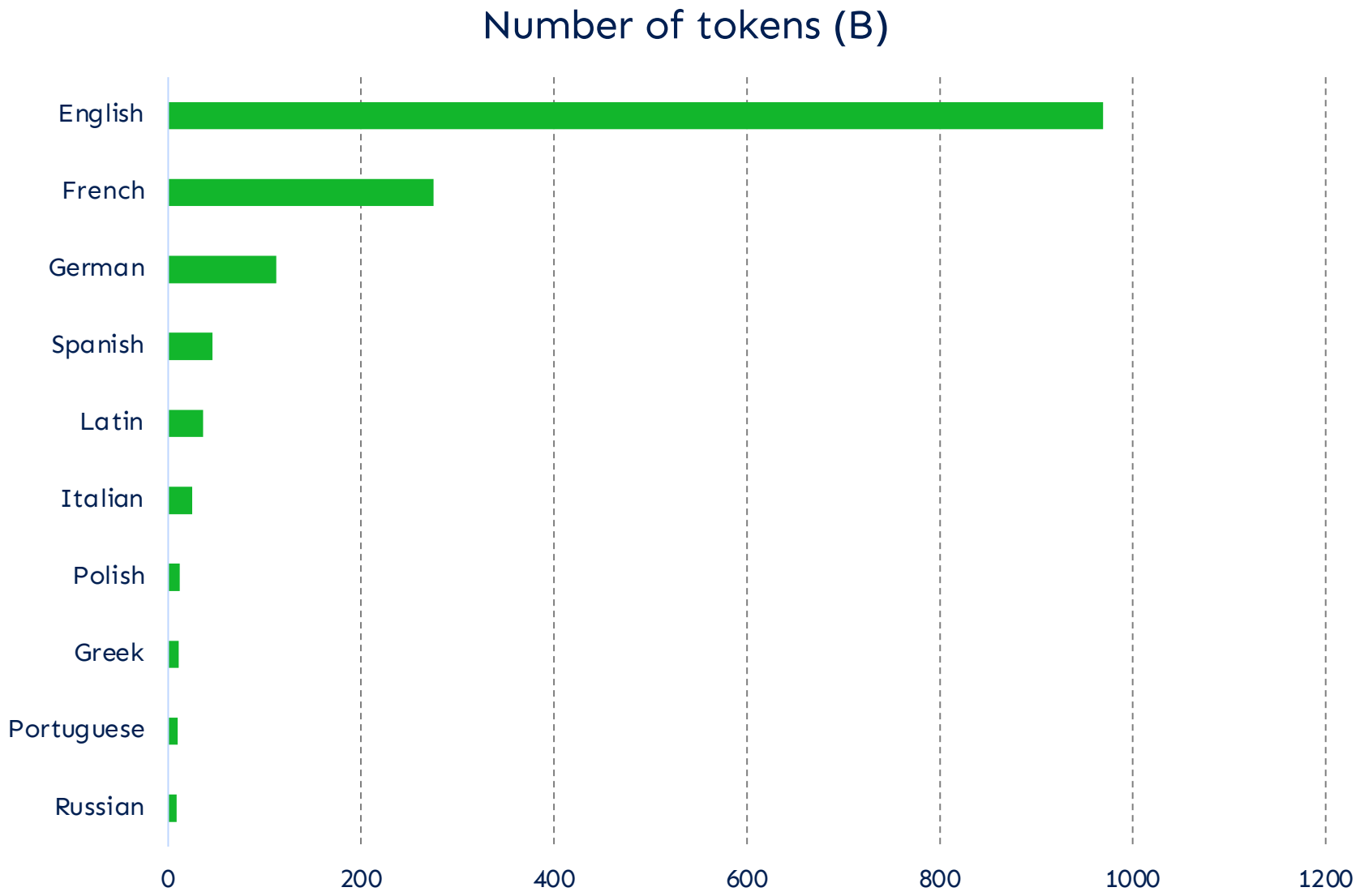
Catherine Arnett Pavel Chizhov Eliot Krzystof Jones Irène Girard

David Mach Anastasia Stasenko Ivan P. Yamshchikov

PleIAs, Paris, France https://pleias.fr/

Abstract

Large Language Models (LLMs) are pre-trained on large amounts of data from different sources and domains. These data most often contain trillions of tokens with large portions of copyrighted or proprietary content, which hinders the usage of such models under AI legislation. This raises the need for truly open pre-training data that is compliant with the data security regulations. In this paper, we introduce Common Corpus[1], the largest open dataset for language model pre-training. The data assembled in Common Corpus are either uncopyrighted or under permissible licenses and amount to about two trillion tokens. The dataset contains a wide variety of languages, ranging from the main European languages to low-resource ones rarely present in pre-training datasets; in addition, it includes a large portion of code data. The diversity of data sources in terms of covered domains and time periods opens up the paths for both research and entrepreneurial needs in diverse areas of knowledge. In this technical report, we present the detailed provenance of data assembling and the details of dataset filtering and curation. Being already used by such industry leaders as Anthropic and multiple LLM training projects, we believe that Common Corpus will become a critical infrastructure for open science research in LLMs.



https://arxiv.org/pdf/2506.01732

Databases & Data Storage

What is a Database?

A database is an **organised** collection of data or a type of data store based on the use of a **database management system** (DBMS), the software that interacts with end users, applications, and the database itself to capture and analyse the data.

https://en.wikipedia.org/wiki/Database

So, Excel...

Spreadsheets are great tools, but their **improper and indiscriminate** use causes all sorts of pain. Data gets lost, updates cannot be tracked, and accountability is impossible.

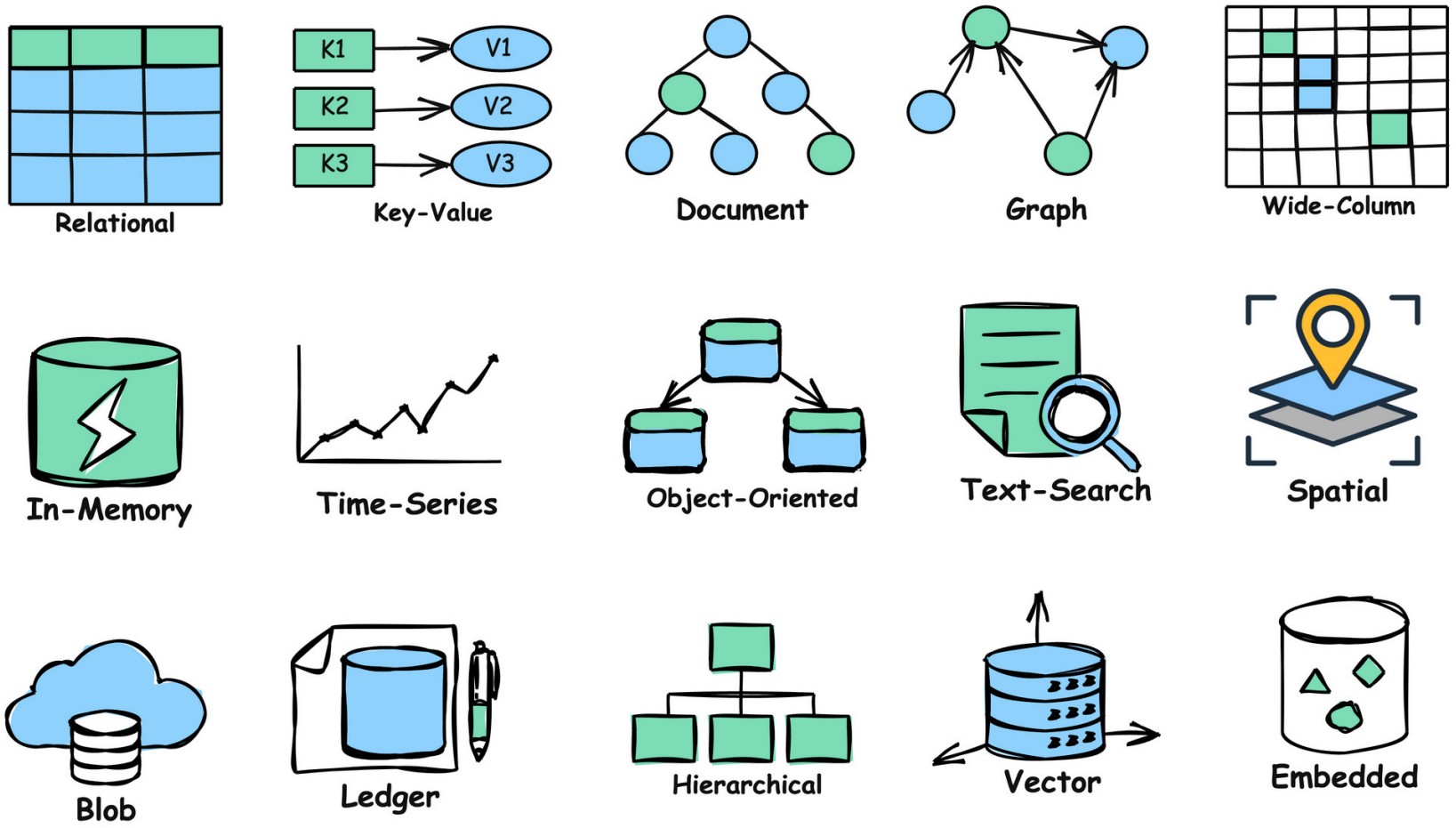
Use Excel for:

- **Individual** analysis
- Ephemeral data

Never use Excel for:

- Business-critical data
- **Sharing**
- **Persistent** storage

Database Types



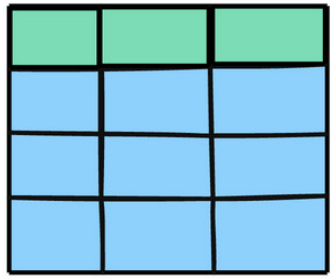
https://blog.algomaster.io/p/15-types-of-databases

Database Types

We will focus solely on **relational databases** and **blob storage**.

Blob storage is not a true database but a method for storing unstructured data.

While you may never need to select a database yourself, it's important to understand that **many options exist**, each with its advantages and disadvantages.



Relational

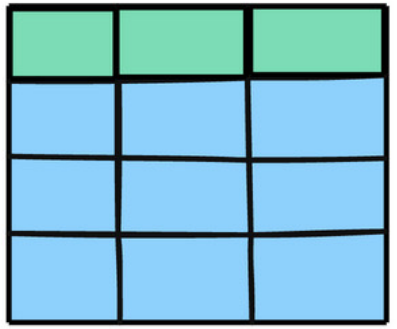
Blob

Relational Databases (1)

A relational database is a type of database that organizes data into **rows and columns**, which collectively form a **table** where the records are related to each other.

Data is typically structured across multiple tables, which can be **joined** together.

Analysts use **SQL queries** to combine different data points and summarize business performance, allowing organizations to gain insights, optimize workflows, and identify new opportunities.



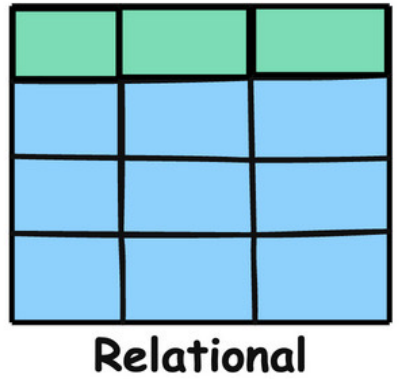
https://www.ibm.com/think/topics/relational-databases

Relational Databases (2)

The relational database management system (RDBMS) is the **database software** that allows users to create, update, insert, or delete data in the system and provides:

- Data structure
- Multi-user access
- Privilege control
- Network access

Examples of popular RDBMS systems include MySQL and PostgreSQL.

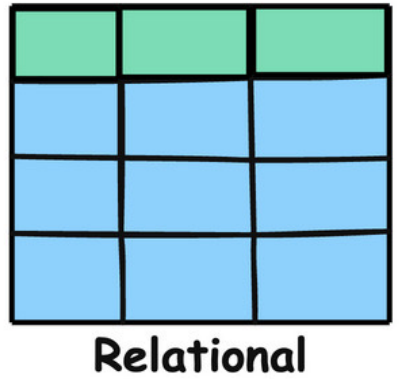


https://www.ibm.com/think/topics/relational-databases

SQL (1)

Structured Query Language (SQL) is the standard programming language for interacting with relational database management systems.

```
SELECT COMPANY_NAME, SUM(TRANSACTION_AMOUNT)
FROM TRANSACTION_TABLE A
LEFT JOIN CUSTOMER_TABLE B
ON A.CUSTOMER_ID = B.CUSTOMER_ID
WHERE YEAR(DATE) = 2022
GROUP BY COMPANY_NAME
ORDER BY SUM(TRANSACTION_AMOUNT) DESC
LIMIT 10
```



https://www.ibm.com/think/topics/relational-databases

SQL (2)

Different from Python, SQL is a **declarative** language.

You tell the RDBMS what to do, **it decides** what is the best sequence of steps (algorithm) to accomplish your task.

SQL is **much more limited** than Python: it only serves to query data.

```
SELECT COMPANY_NAME, SUM(TRANSACTION_AMOUNT)
FROM TRANSACTION_TABLE A
LEFT JOIN CUSTOMER_TABLE B
ON A.CUSTOMER_ID = B.CUSTOMER_ID
WHERE YEAR(DATE) = 2022
GROUP BY 1
ORDER BY 2 DESC
LIMIT 10
```

Blob Storage (1)

Blob storage is a type of storage for **unstructured data**.

A "blob", which is short for Binary Large Object, is a mass of data in binary form that **does not necessarily conform** to any file format.

Blob storage keeps these masses of data in non-hierarchical storage areas called **data lakes**.

https://www.cloudflare.com/learning/cloud/what-is-blob-storage/

Blob Storage (2)

Blob storage is a cloud-native technology designed to support **unstructured data.**

However, due to its cost-effectiveness and scalability compared to traditional databases, it is increasingly being used to **store structured data for analytical purposes.**

This is made possible by tools that enable **queries on collections of CSV or Parquet** files.

Resources

IBM, *What is a relational database?:* a quick introduction to relational databases.

W3School, *SQL Tutorial:* a complete tutorial to learn SQL from zero (not required for this class).

Live Coding

- Setting up docker and VS Code
- Cloning the repository
- Inspecting a relational database
- Learning about primary and foreign keys
- Playing with SQL
- Inspecting a blob storage
- Playing with Google Cloud Storage

Discussion

Let’s try to discuss a couple of real-world datasets.

Case 1

You are a manager in a challenger bank (say Aidexa or CF+).

You want to build a new credit rating model to evaluate loan applications from small and medium enterprises.

You have numeric and categorical data (revenues, category, employees of each company) as well as documents (income statements, ID cards, ...)

Case 2

You are a manager in a utility (say A2A or ENEL).

You want to build a new forecasting model to predict the power load demand in Italy.

You have time series data of power demand, generation, and price, as well as meteorological data.

Data Exploration

Disclaimer

We will focus on **structured data**.

Unstructured data, like text, images and videos, are usually handled by specialised models.

We will discuss them in future lectures.

The Grammar of Datasets

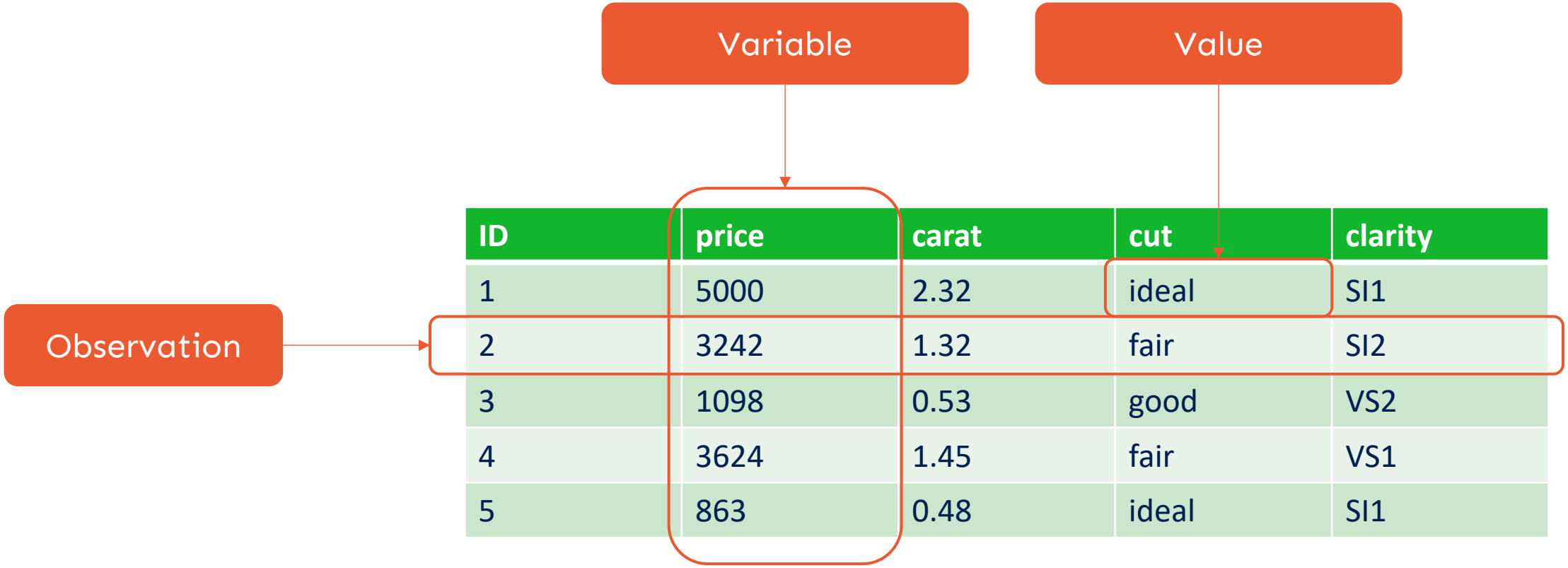
Variable: a property that can be measured.

Value: the state of a variable when the measurement happens.

Observation (or sample or data point): the set of values of several variables measured in similar conditions.

Dataset: a set of observations about a process.

A Dataset



ID	price	carat	cut	clarity
1	5000	2.32	ideal	SI1
2	3242	1.32	fair	SI2
3	1098	0.53	good	VS2
4	3624	1.45	fair	VS1
5	863	0.48	ideal	SI1

Samples in a Dataset

In most datasets, samples are generally assumed to be **independent**, although this is **rarely entirely true** in real-world scenarios. Consequently, the order of the samples does not typically matter.

When the order of the samples is important and each sample is associated with a timestamp, the data constitutes a **time series**.

Time series will not be discussed in detail in this class.

Time Series

Time series are both fascinating and important for businesses.

If you're interested in learning more, a great book to start with is Forecasting: Principles and Practice (3rd ed) by Rob J. Hyndman and George Athanasopoulos.

Types of Variables

Nominal (Categorical): categorical data with no natural ordering (e.g. hair colour).

Operations: $=, \neq$

Ordinal (Categorical): categorical data with a natural ordering (e.g. grades).

Operations: $=, \neq, >, <$

Interval (Numerical): numerical data with an arbitrary position of zero (e.g. Celsius degrees).

Operations: $=, \neq, >, <, -$

Ratio (Numerical): numerical data with a meaningful zero (e.g. Kelvin).

Operations: $=, \neq, >, <, -, \div$

Discussion

You have a dataset, like the one shown in the previous slide.

What would you do?

You Have a Dataset. Now What?

1. **Look** at the data.
2. **Understand** what each variable means.
3. Look for **missing data** / unreasonable values.
4. Perform **univariate** analysis – look at each variable alone with statistics and charts.
5. Perform **multivariate** analysis – look at the interaction between variables with statistics and charts.
6. Understand how to **handle missing data** and outliers.
7. Document and **report** your findings – or you will forget and regret it.

Data Exploration vs Presentation

Data Exploration: you search for insights to understand the process you are analysing.

Data Presentation: you want to prove a point to your audience.

Why Data Exploration?

People are very good at **detecting patterns**.

Data exploration alone can **solve most problem statements**, saving the time, effort, and resources to build and maintain expensive models.

Even when a model must be built, insights from data exploration can suggest:

- Which variables to include
- How to transform variables
- Which aspect to investigate with subject matter experts

Summary Statistics

The most common summary statistics for nominal and ordinal variables are:

- **Mode**: the most frequently occurring category
- **Gini Index**: $G = 1 - \sum_i f_i^2$, where f_i is defined below. The Gini index reaches its minimum value when all frequencies are identical, indicating maximum homogeneity. Conversely, it equals 1 when one category has a frequency of 1, while all others are 0.

The most informative computation we can perform is determining the **distribution**, specifically the frequency f_i of each category i.

Yet, the distribution is not a scalar.

Summary Statistics

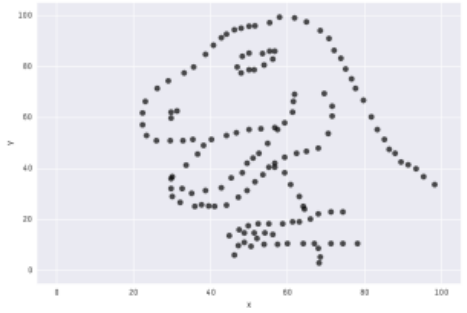
The most common summary statistics for **interval and ratio** variables are:

- **Mean**: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ the arithmetic average of the data values.
- **Median**: the middle value when the data is sorted in order. It is less sensitive to outliers than the mean.
- **Standard Deviation**: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$, the square root of the variance. It provides a measure of spread in the same units as the data.
- **Quantiles**: values below which a certain fraction of data falls (e.g., 0.25, 0.50, and 0.75)

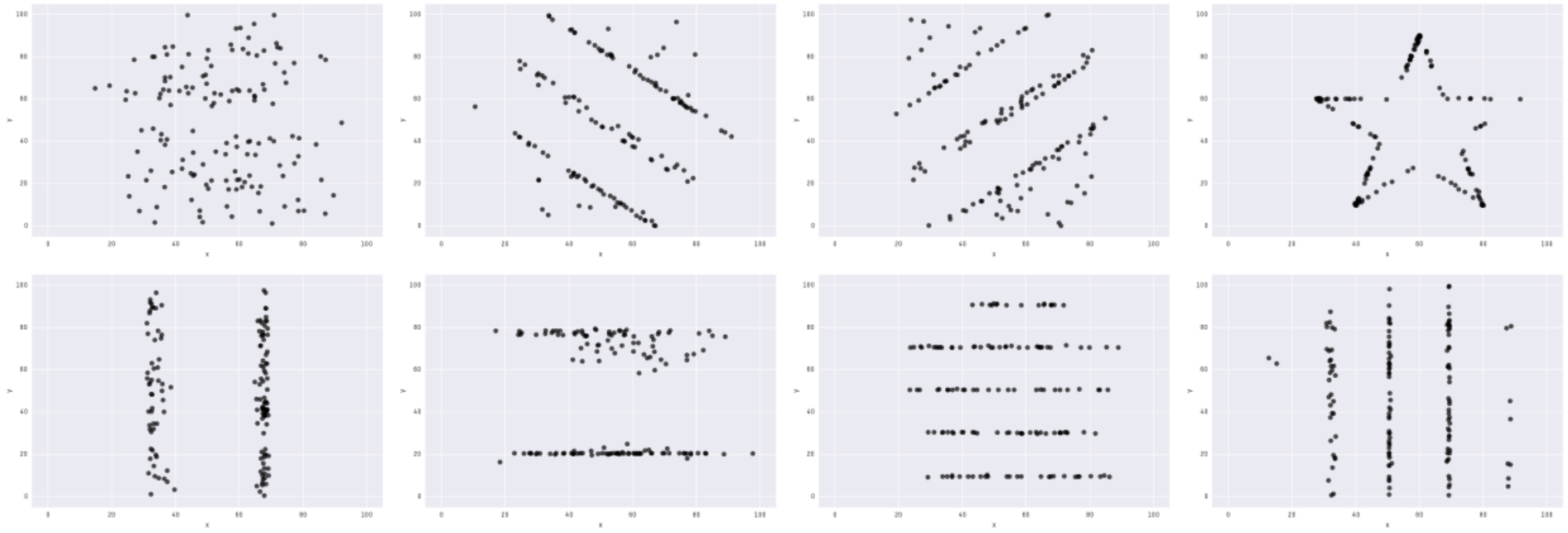
Again, the most informative quantity is the full distribution of the data.


I am once again asking
should you trust summary stats?
imgflip.com

Should You Trust Stats?



X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06



https://en.wikipedia.org/wiki/Datasaurus_dozen


Summary
Stats
Plots
imgflip.com

Data Visualization

Data Visualization is a critical skill in Data Science and Machine Learning.
Arguably, it is becoming a critical skill for every professional.

However, data visualization is not easy.

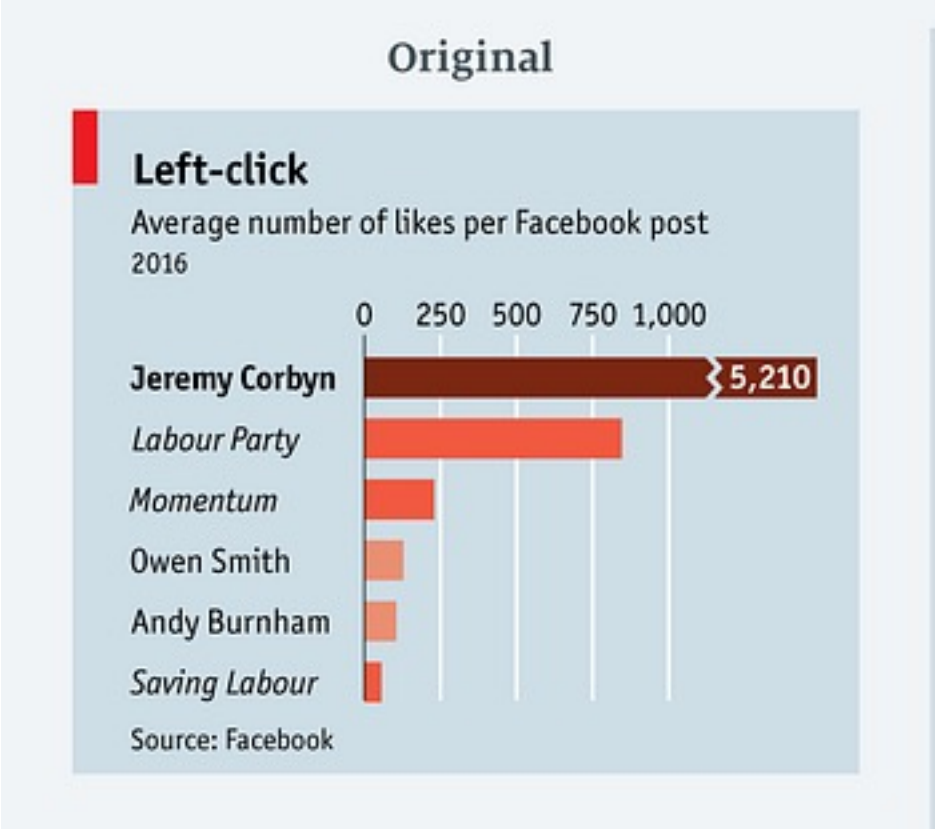
It is highly recommended to read these articles:

- S. Leo, Mistakes, we've drawn a few
- E. Fabbiani, The traps of data visualization
- C. O. Wilke, Fundamentals of data visualization [chap. 1]

Discussion

Let’s discuss and criticise some visualizations!

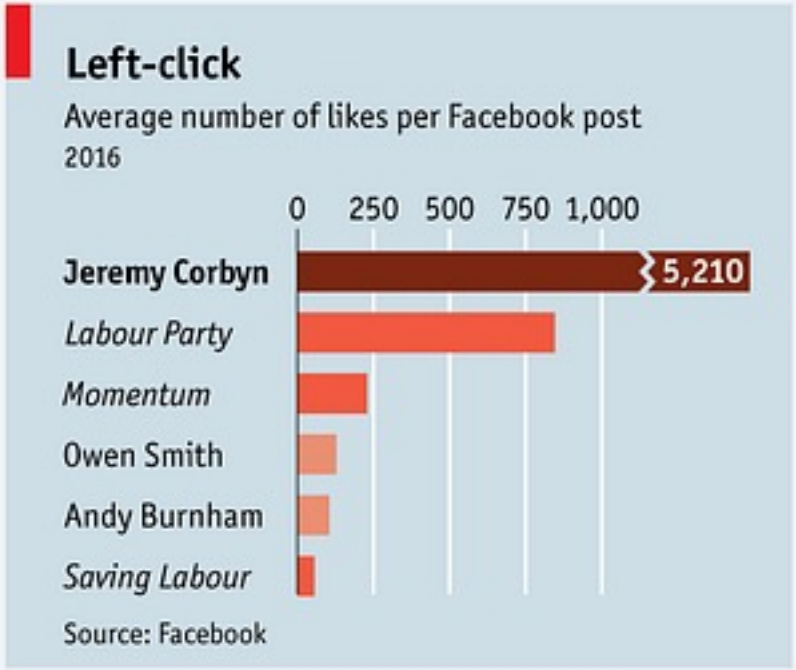
What's Wrong with This Chart?



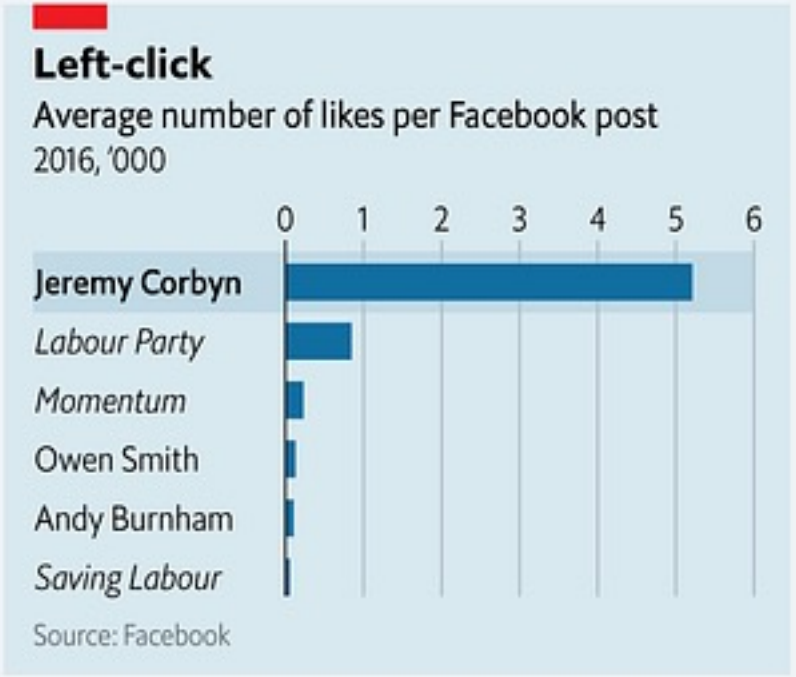
https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368

Answer

Original



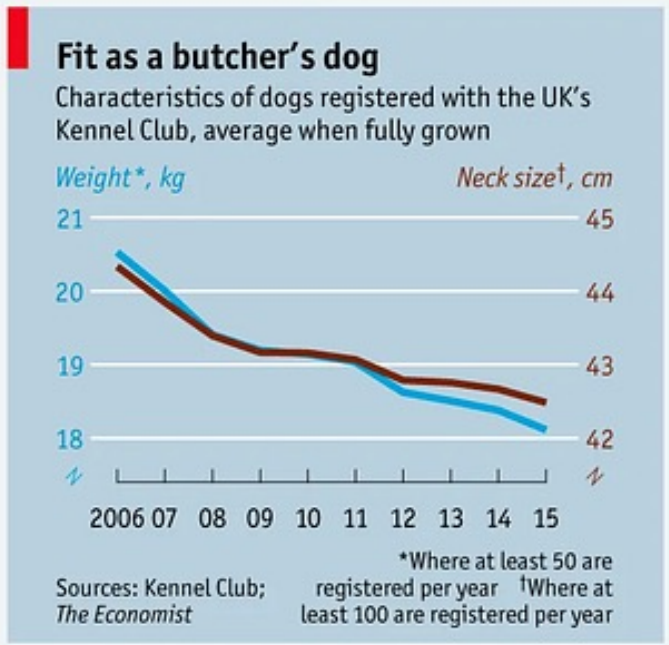
Better



https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368

What's Wrong with This Chart?

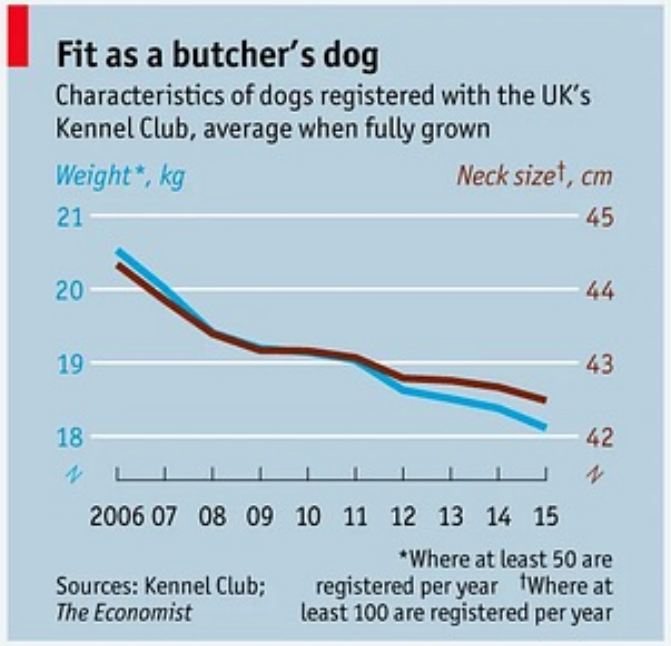
Original



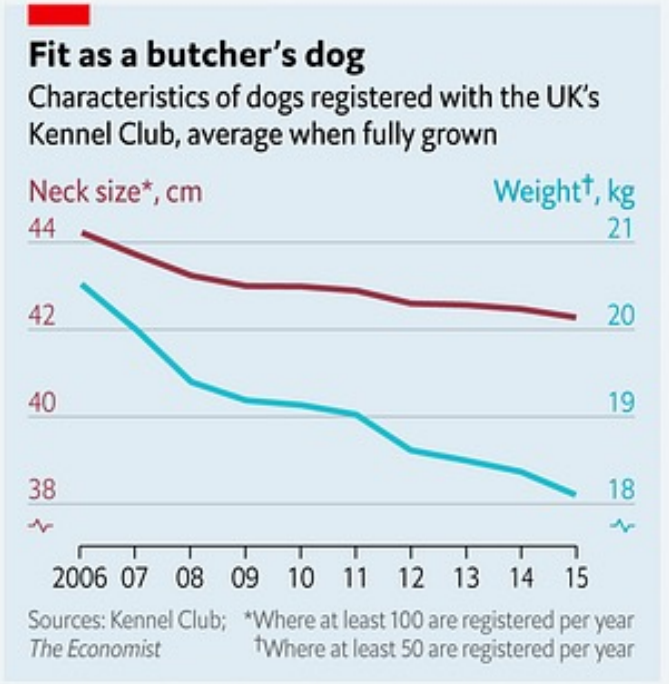
https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368

Answer

Original



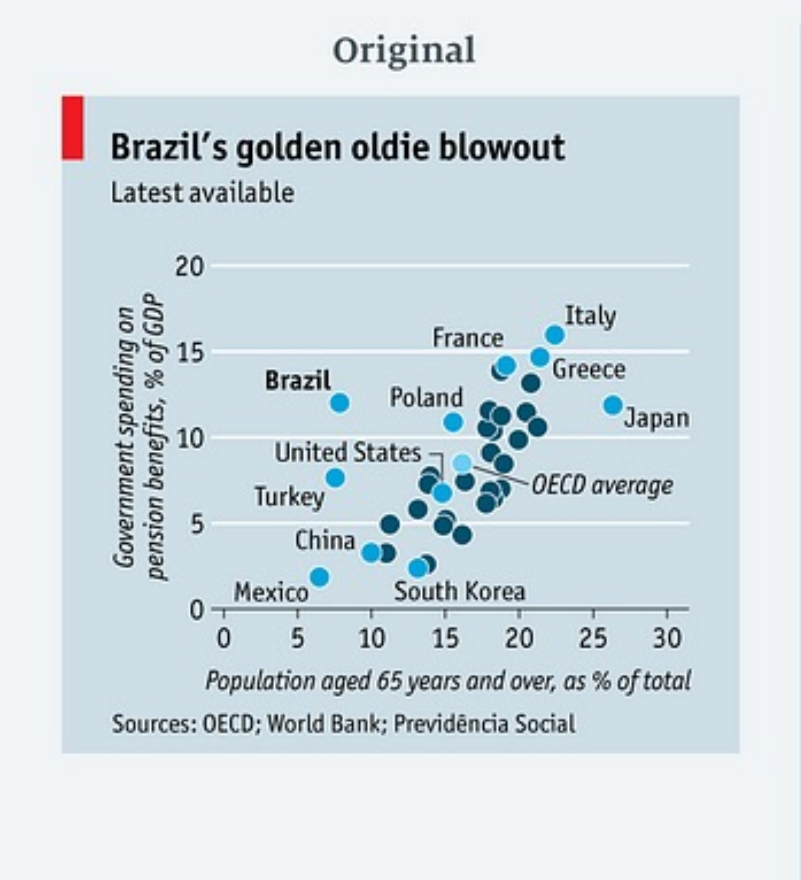
Better



https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368

What's Wrong with This Chart?

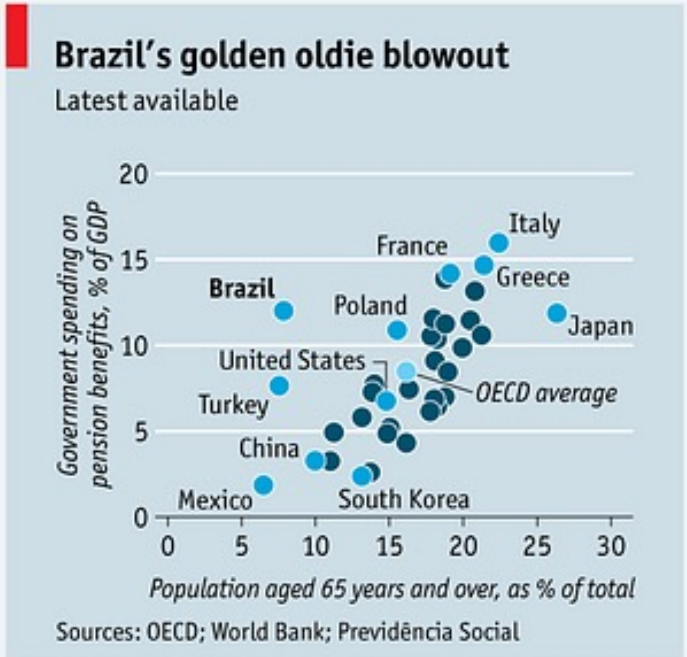
Original



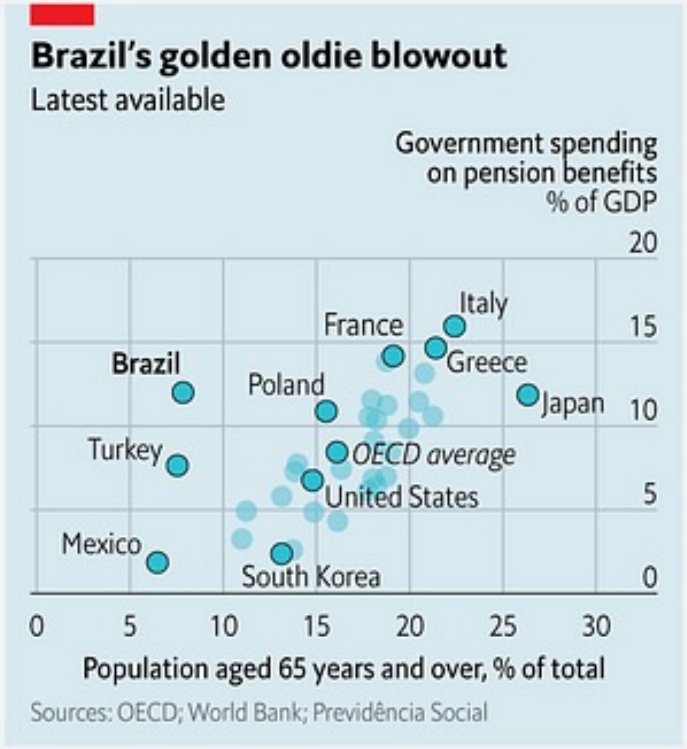
https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368

Answer

Original

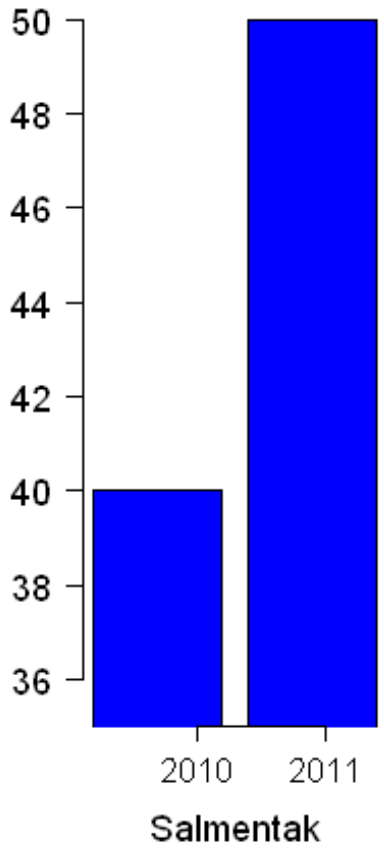


Better



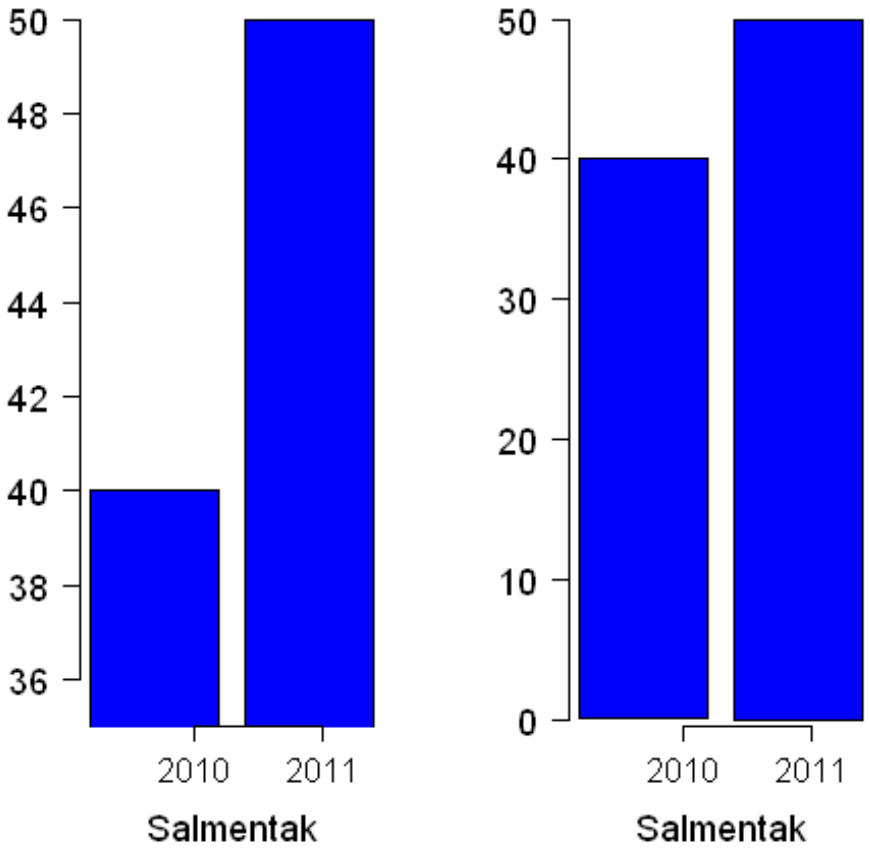
https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368

What's Wrong with This Chart?

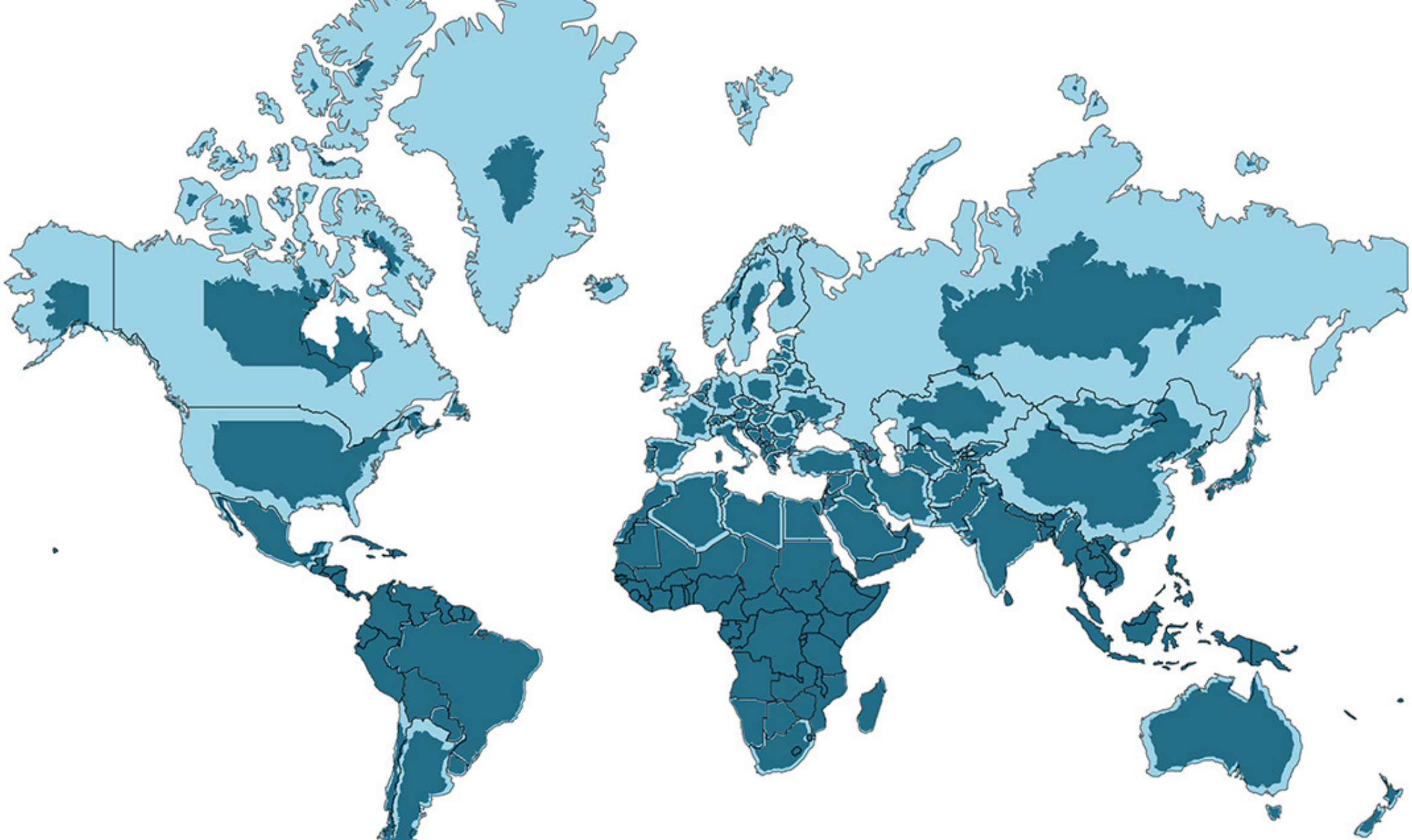


https://en.wikipedia.org/wiki/Misleading_graph

Answer



https://en.wikipedia.org/wiki/Misleading_graph



https://www.thetruesize.com/

What is Data Visualization?

Data visualization is the **graphical representation** of information and data.

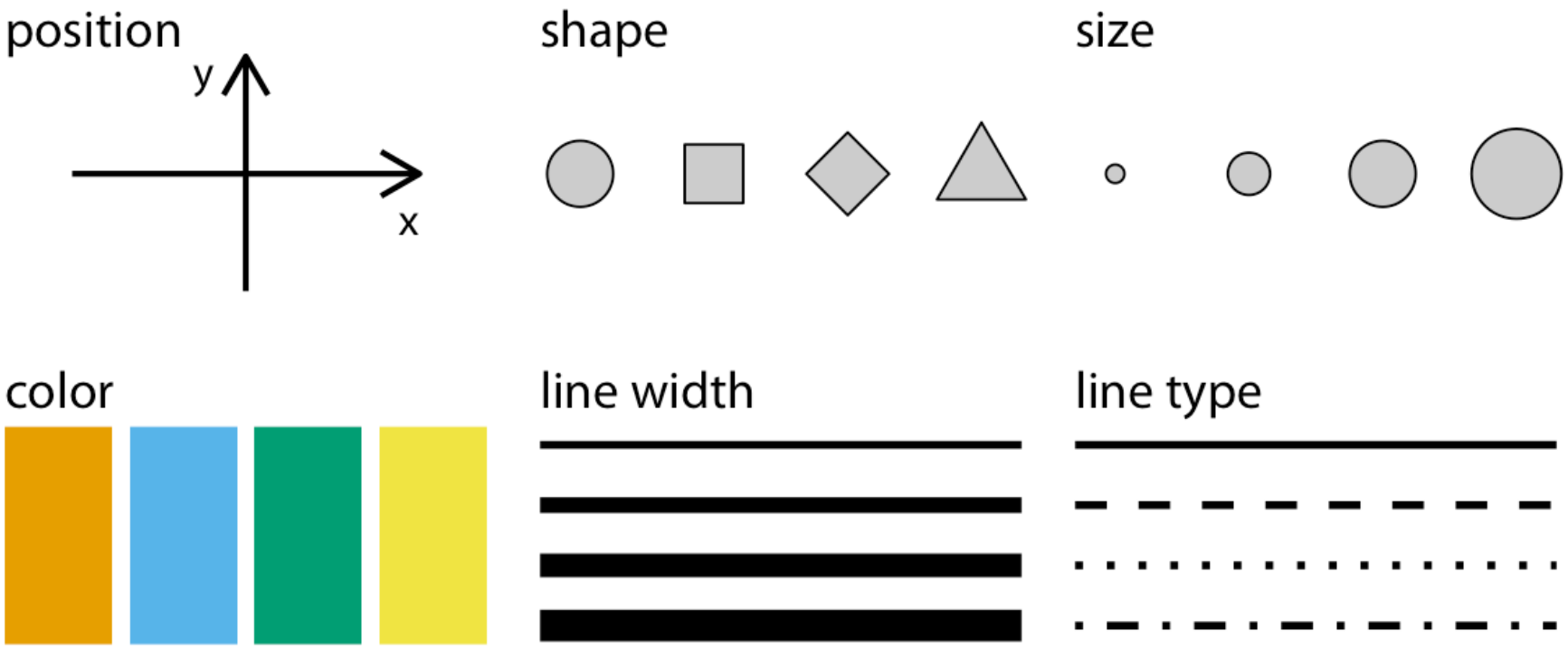
By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Additionally, it provides an excellent way for employees or managers to present data to non-technical audiences without confusion.

https://www.tableau.com/visualization/what-is-data-visualization

What is Data Visualization?

All data visualizations map data values into quantifiable features of the resulting graphic. We refer to these features as **aesthetics**.



https://clauswilke.com/dataviz/aesthetic-mapping.html

Mapping Numerical Variables

1. Position

2. Length

3. Angle/Slope

4. Area

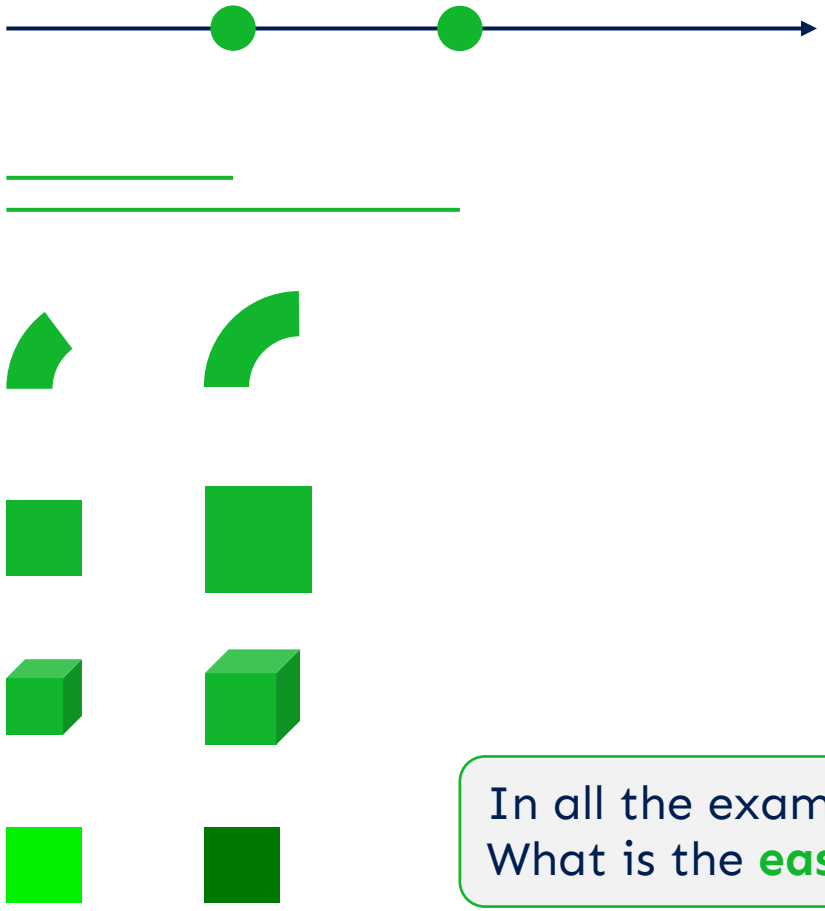
5. Volume

6. Colour/Density

https://www.coursera.org/learn/datavisualization

Mapping Numerical Variables

1. Position
2. Length
3. Angle/Slope
4. Area
5. Volume
6. Colour/Density

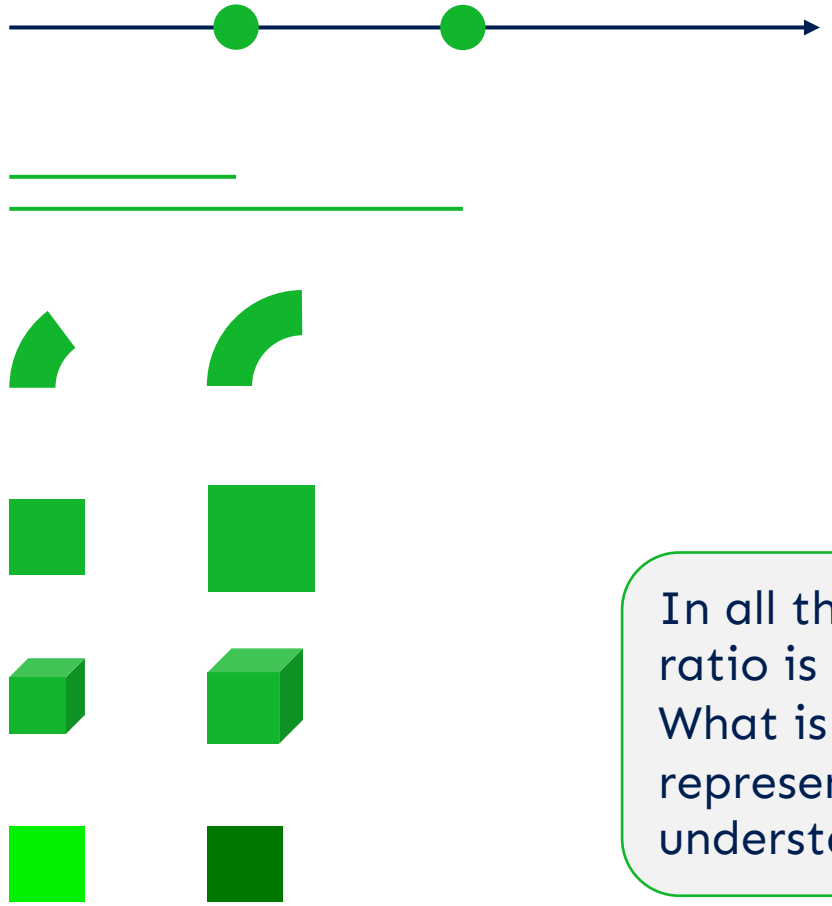


In all the examples, the ratio is 2:1.
What is the **easiest** representation to understand?

Mapping Numerical Variables

Perception accuracy

1. Position
2. Length
3. Angle/Slope
4. Area
5. Volume
6. Colour/Density



In all the examples, the ratio is 2:1.
What is the **easiest** representation to understand?

https://www.coursera.org/learn/datavisualization

Mapping Variables

Perception accuracy	Quantitative	Ordinal	Nominal
1	Position	Position	Position
2	Length	Density	Hue
3	Angle	Saturation	Texture
4	Slope	Hue	Connection
5	Area	Texture	Containment
6	Volume	Connection	Density
7	Saturation	Containment	Saturation
8	Hue	Length	Shape
9		Angle	Length
10		Slope	Angle
11		Area	Slope
12		Volume	Area
13			Volume

https://www.coursera.org/learn/datavisualization

Mapping Variables

Perception accuracy	Quantitative	Ordinal	Nominal
1	Position	Position	Position
2	Length	Density	Hue
3	Angle	Saturation	Texture
4	Slope	Hue	Connection
5	Area	Texture	Containment
6	Volume	Connection	Density
7	Saturation	Containment	Saturation
8	Hue	Length	Shape
9		Angle	Length
10		Slope	Angle
11		Area	Slope
12		Volume	Area
13			Volume

https://www.coursera.org/learn/datavisualization

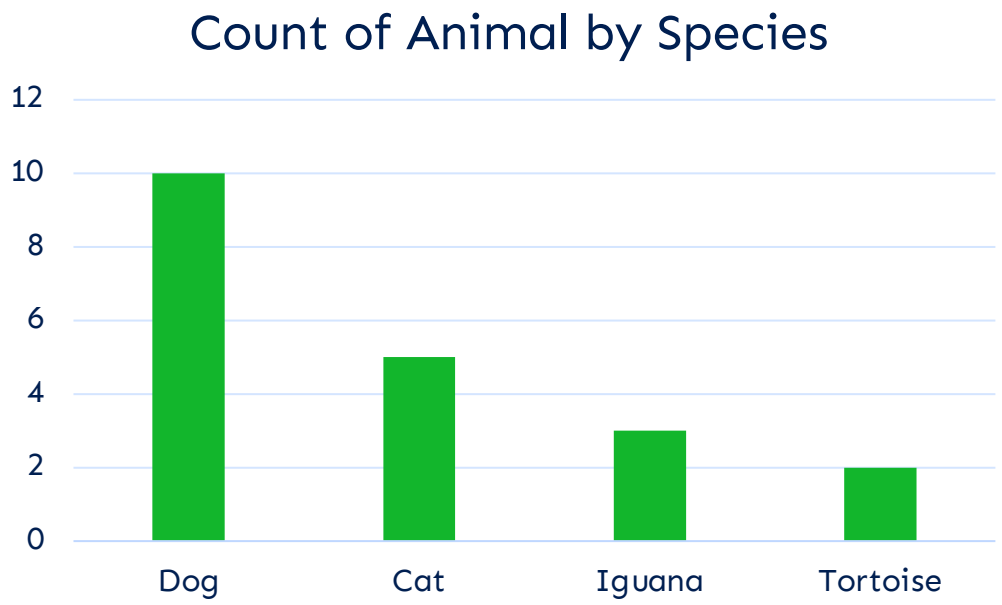
Plotting 1D Data

With one-dimensional data (single variables), the **distribution** is the most informative visualization.

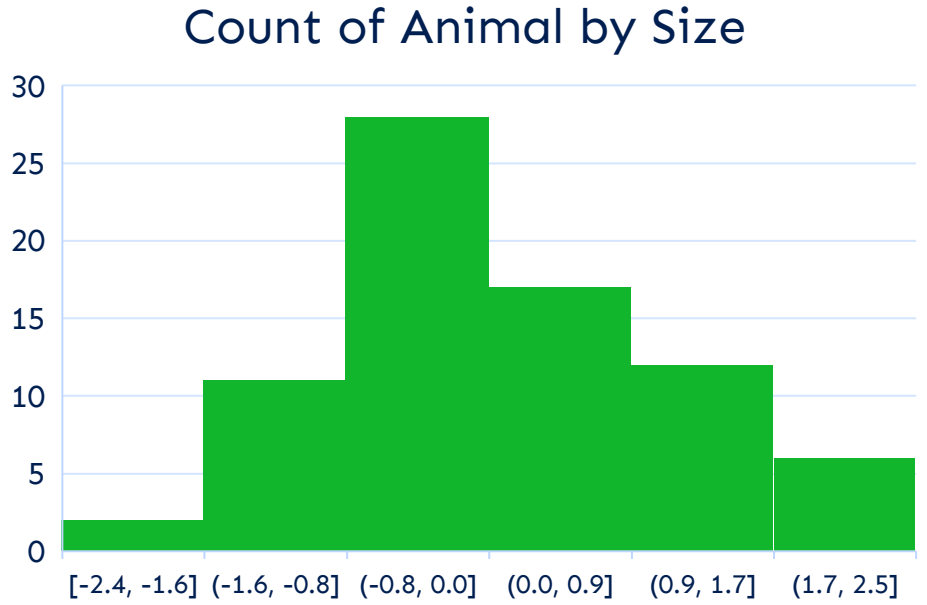
For categorical data, the distribution represents the count or proportion of samples in each category.

For numerical data, the distribution represents the count or proportion of samples within specific intervals.

Bar Charts and Histograms



Bar chart for categorical variables. Uses both length and position as aesthetics.

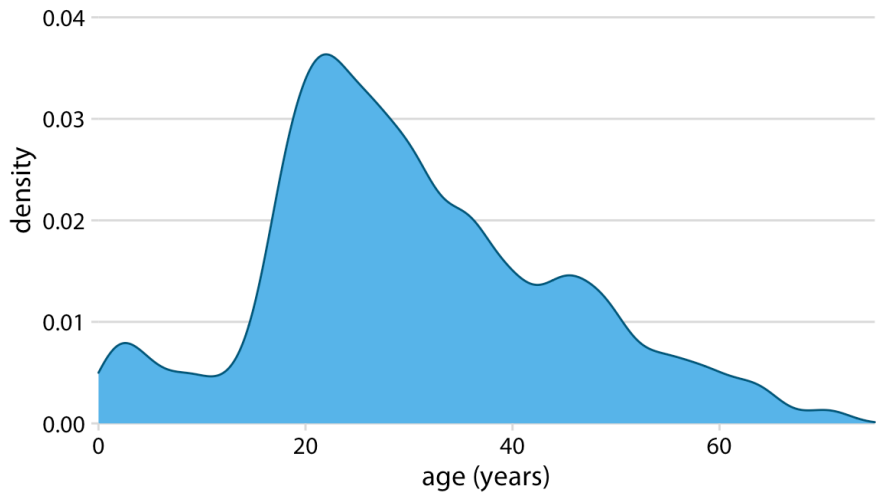


Histogram for numerical variables. Uses both length and position as aesthetics.

Area Charts and KDE

You can use **area charts** with Kernel Density Estimation for 1D numerical data.

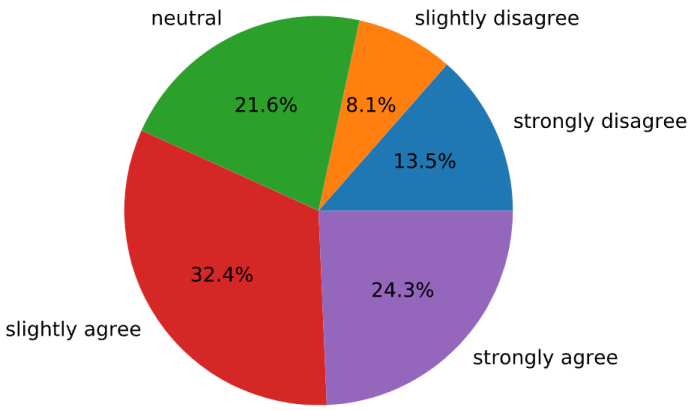
They use the same aesthetics and convey the same information as histograms and are as effective as histograms.

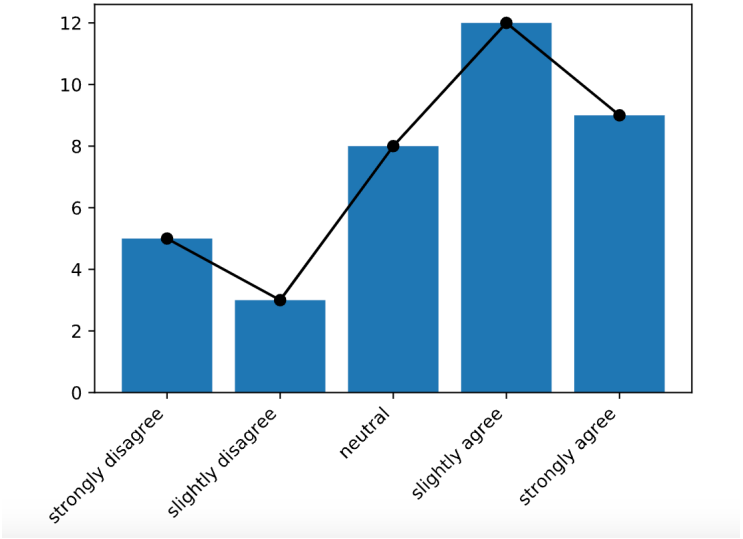


Do not...

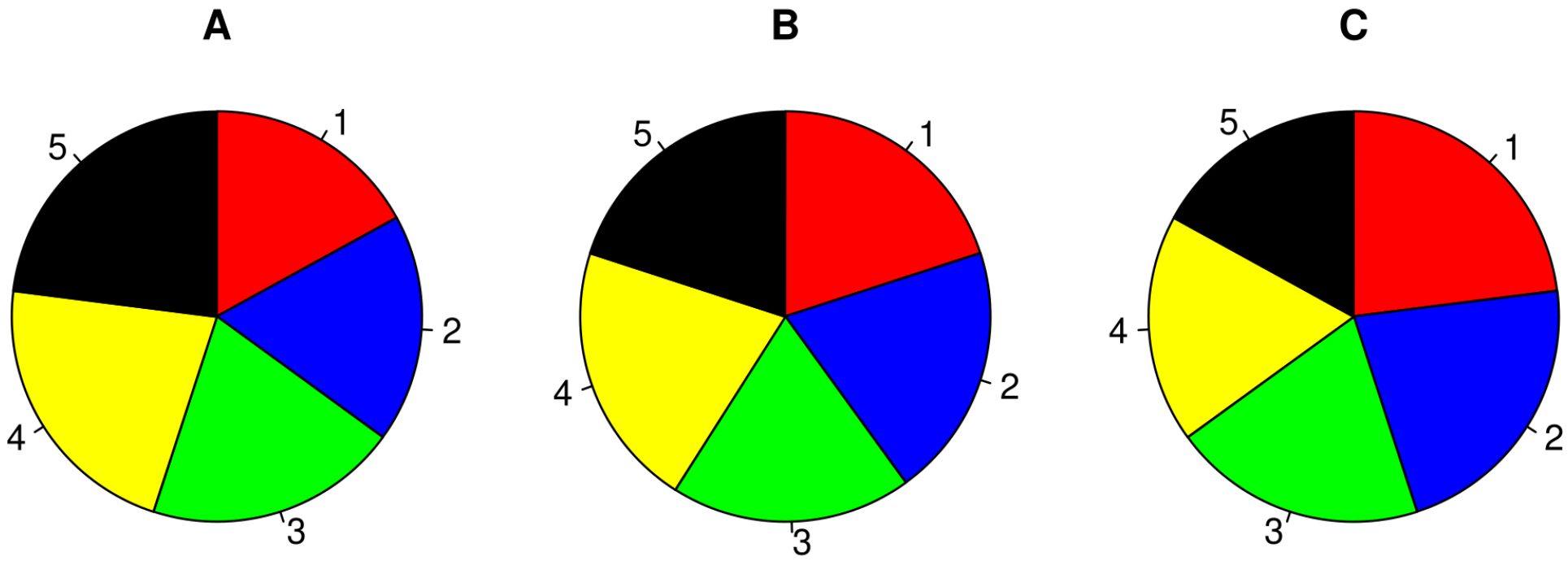
Use pie charts. They are always worse than bar charts in conveying the same information.

Add lines on top of histograms or bar charts. They are useless at best, when not plain misleading.

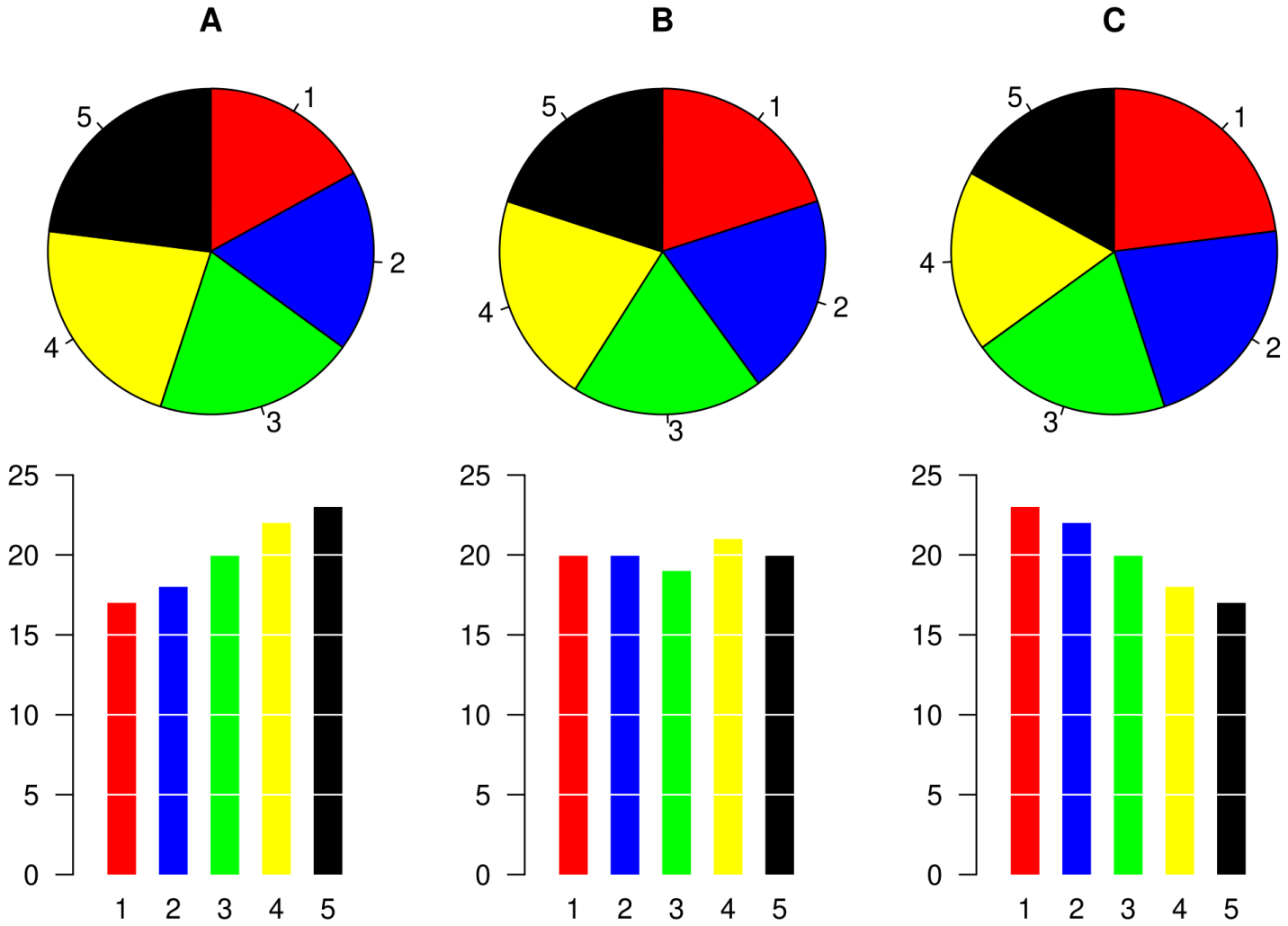




Why Pie Chats Are Bad?



That's Why Pie Chats Are Bad!



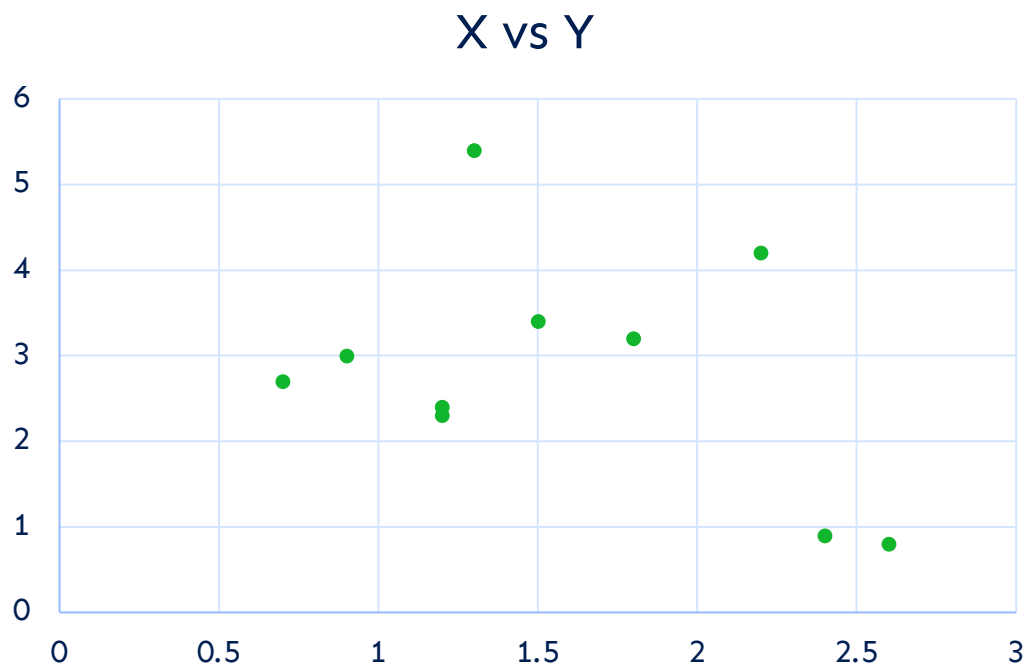
Plotting 2D Data

With two-dimensional data, various plots are available, depending on the types of variables:

- **Numerical vs. Numerical**: To visualize joint distributions, a scatter plot is typically the best choice.
- **Numerical or Ordinal vs. Date**: To show trends over time, a line or area chart is ideal.
- **Numerical vs. Categorical**: To examine the distribution of the numerical variable conditioned on the categorical variable, use a box plot or violin plot.
- **Categorical vs. Categorical**: To visualize the joint distribution, a heatmap is most effective.

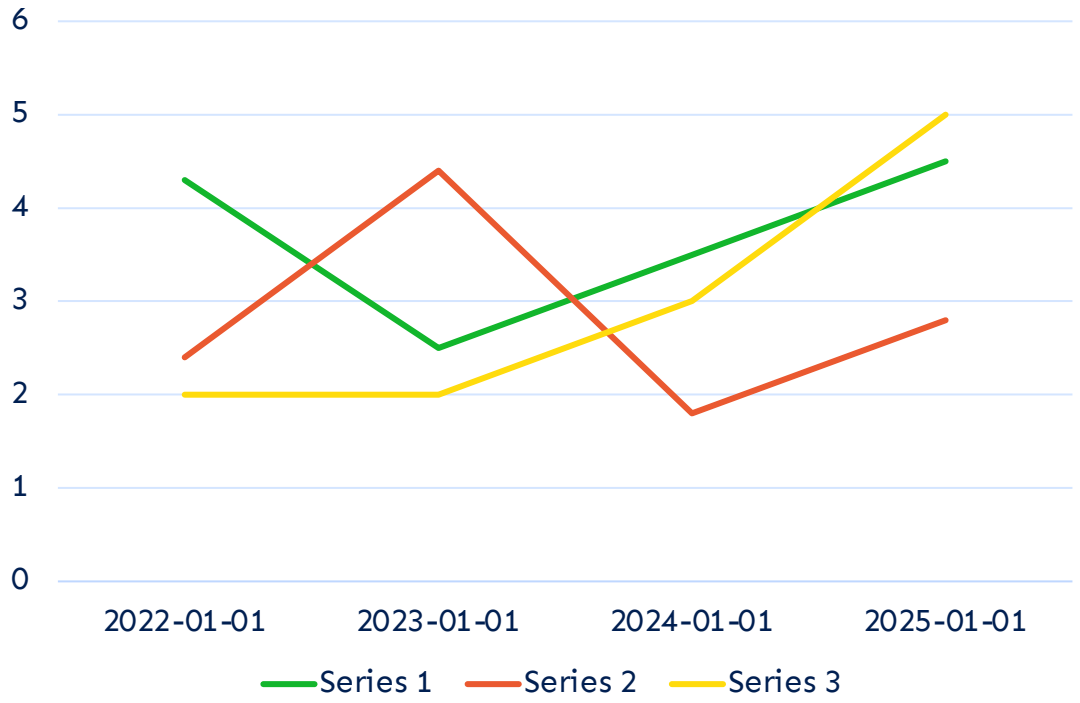
Scatter Plots

Scatter plots use position as the aesthetic for both variables. They are highly effective for exploring **joint patterns** and can incorporate **additional aesthetics**, such as colour, to represent more variables, including categorical ones.



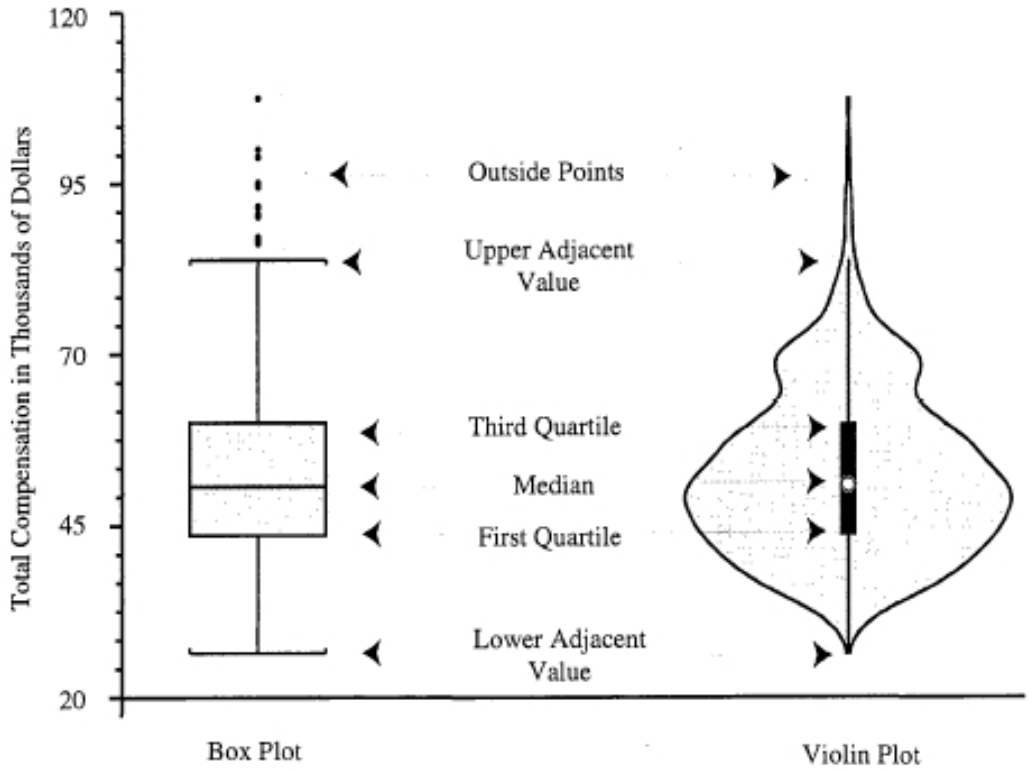
Line Charts

Line charts should only be used when a variable represents **dates or another sequence**. An additional categorical variable can be included using colour or line type.



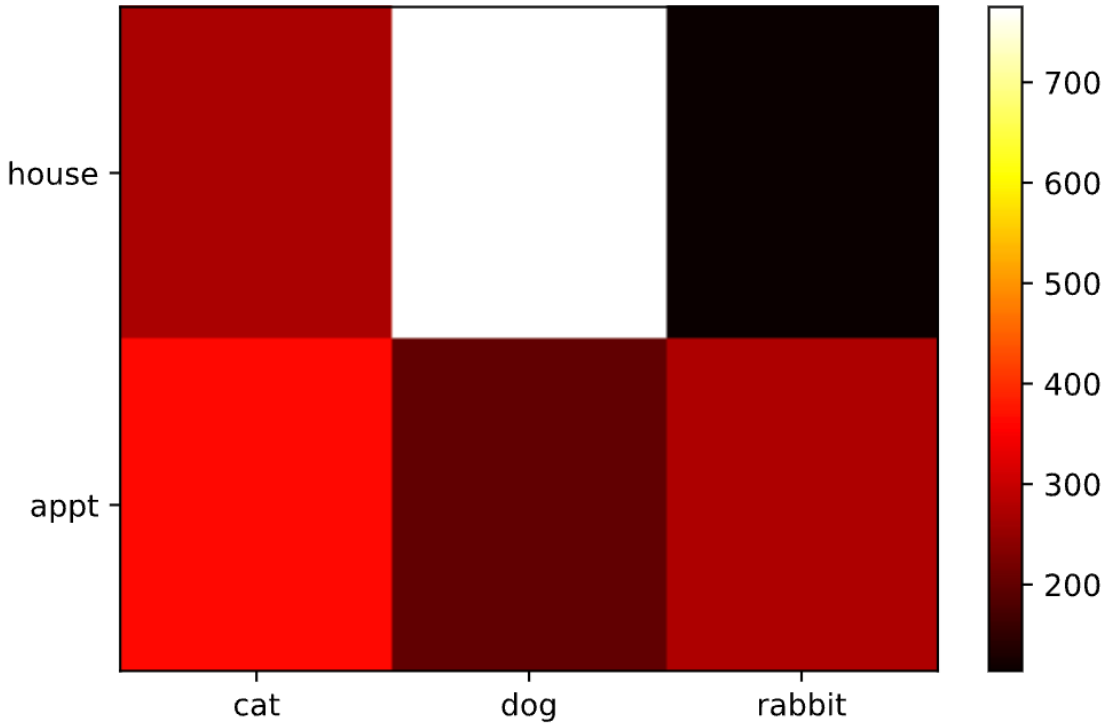
Box and Violin Plots

Both box plots and violin plots depict the distribution of a numeric variable across categories of a categorical variable. However, **violin plots are typically more informative** as they display the full distribution.



Heatmaps

Heatmaps display the **joint distribution** of two categorical variables, using colour to represent the count or frequency of each category pair. Colour is utilized because the position is already assigned to the variables.

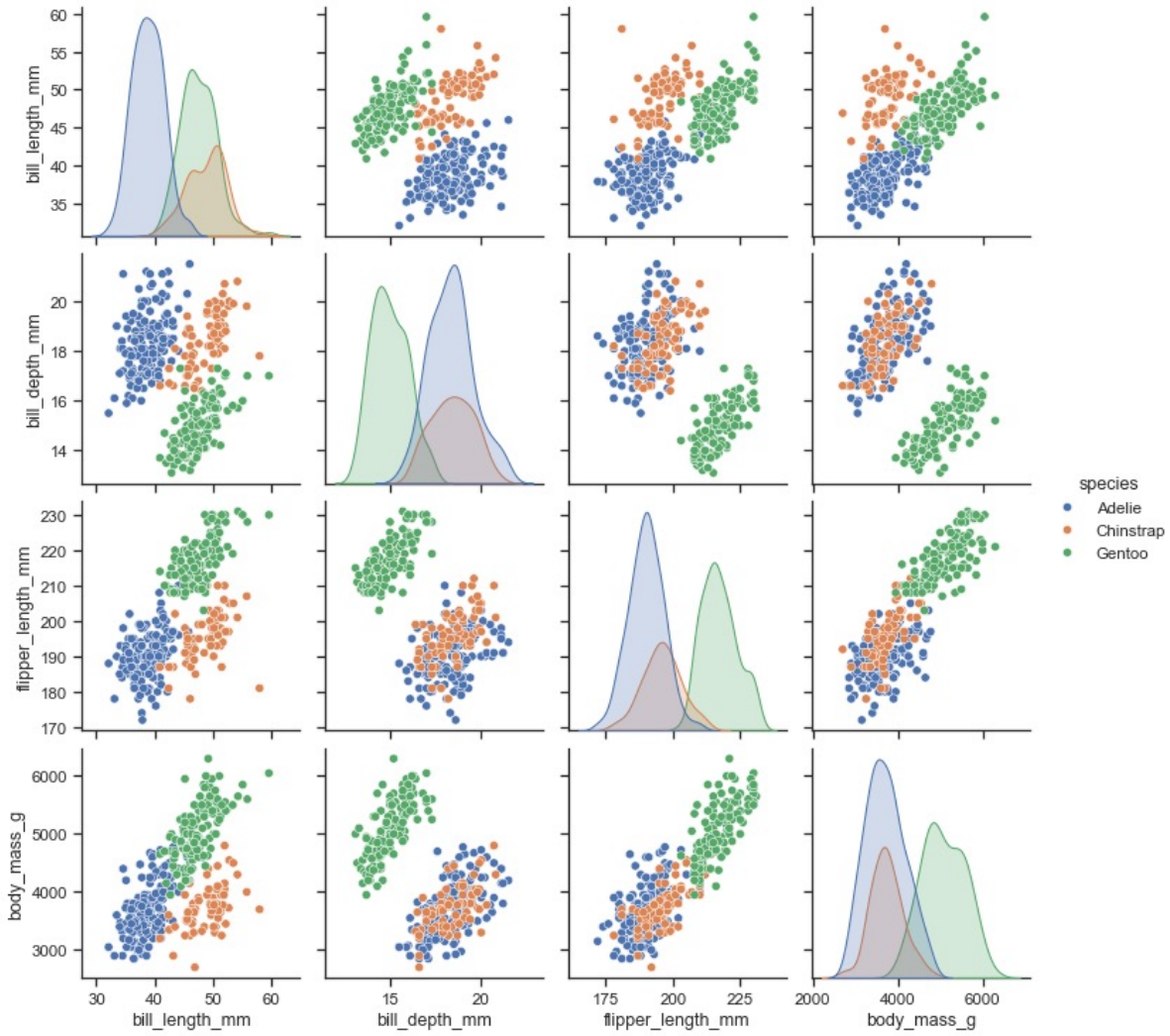


Plotting Higher-Dimensional Data

3D scatter plots, line charts, and surfaces are often difficult to interpret and **should be avoided**.

Additional variables can be incorporated into 2D charts using colour or **other visual aesthetics**.

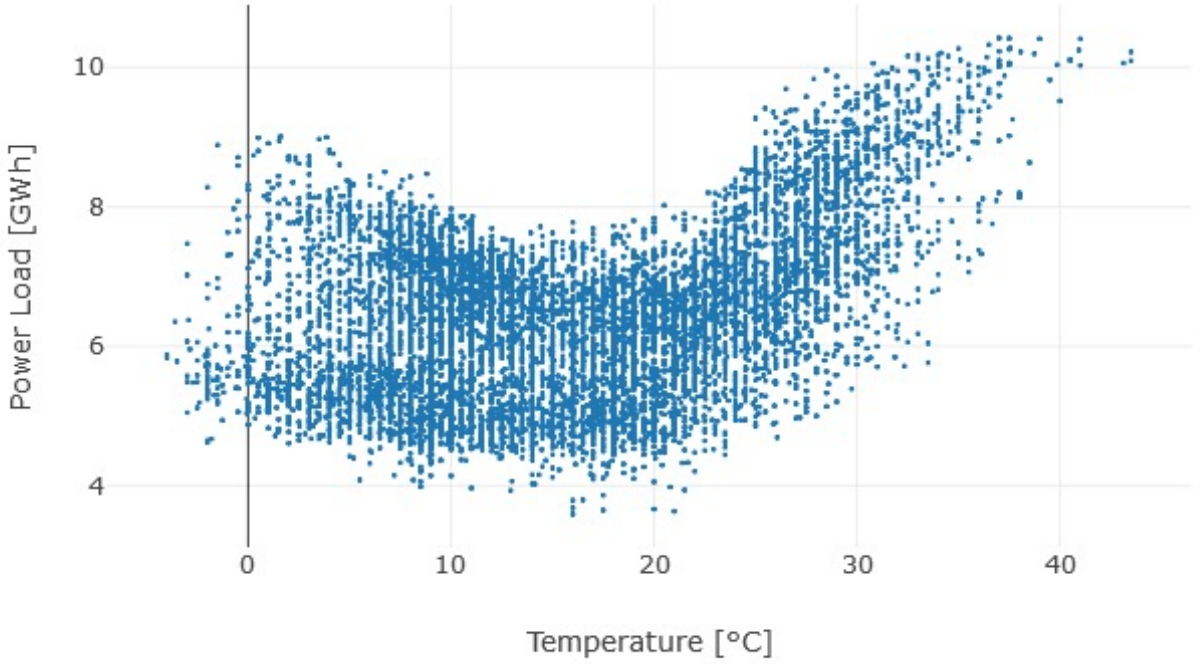
Alternatively, multiple charts can be created and arranged spatially, such as with scatterplot matrices.



Discussion

Let’s draw some
insights from a chart!

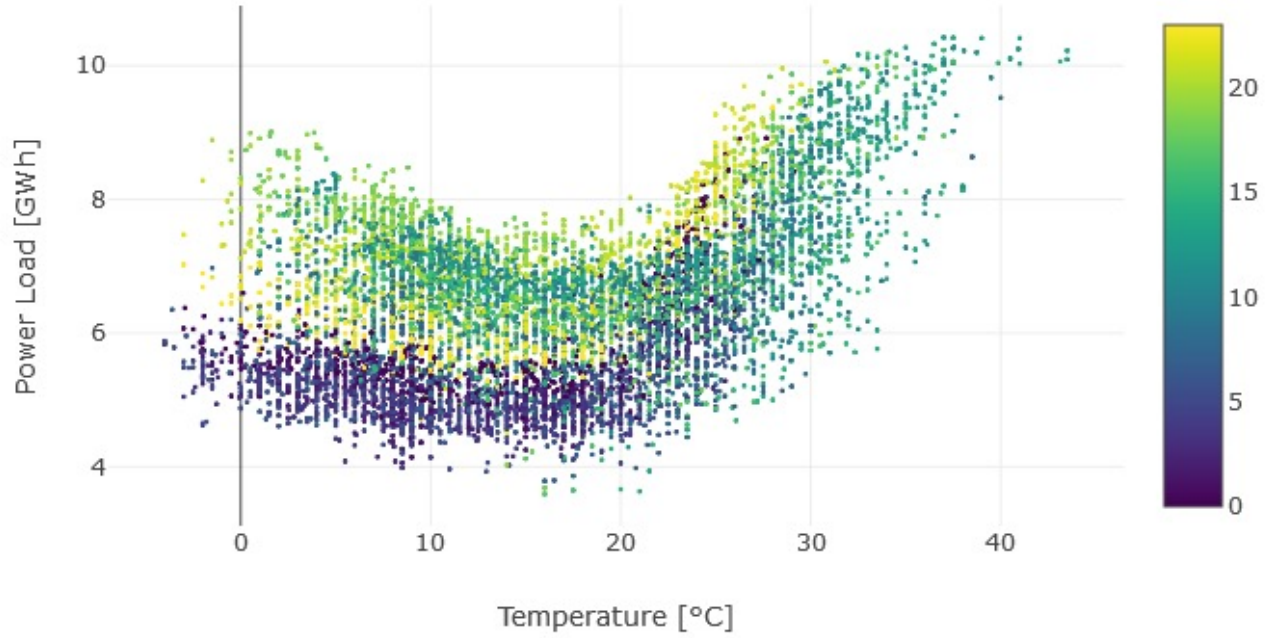
Power Load vs Temperature



Hourly power load in Greece vs average temperature.

https://towardsdatascience.com/lessons-from-a-real-machine-learning-project-part-2-the-traps-of-data-exploration-e0061ace84aa

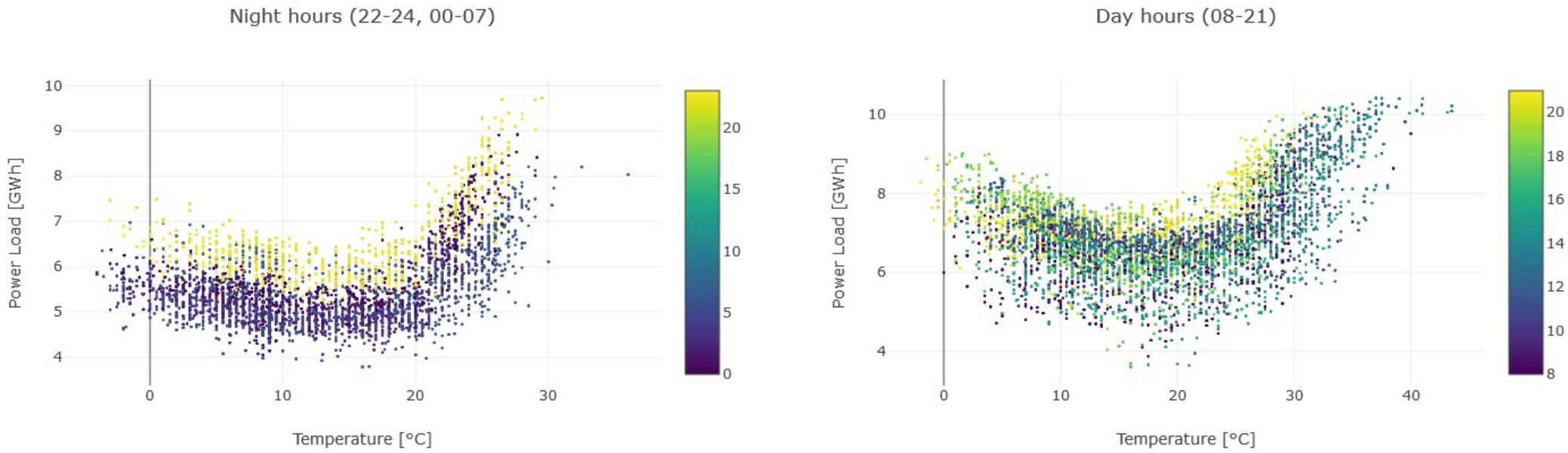
Power Load vs Temperature



Power load in Greece vs average temperature – the colour represents the hour of the day.

https://towardsdatascience.com/lessons-from-a-real-machine-learning-project-part-2-the-traps-of-data-exploration-e0061ace84aa

Power Load vs Temperature



Power load in Greece vs average temperature – the colour represents the hour of the day.

https://towardsdatascience.com/lessons-from-a-real-machine-learning-project-part-2-the-traps-of-data-exploration-e0061ace84aa

Open notebook in Colab

Exam Simulation!

This session **will not be graded**, and the difficulty level may not be comparable to that of the exam.

Take the Simulation:

https://forms.gle/Xv4zmX5CiRxR5ny1A


The End