Practical AI

Emanuele Fabbiani

From the Last Lecture...

1. Business Understanding

- Workshop
- Evaluation metrics
- Literature review

2. Data Gathering

- Source discovery
- Data preparation
- Quality assessment

3. Data Exploration

- Explorative analysis
- Data cleaning

4. Modelling

- Feature engineering
- Model design
- Model implementation

5. Evaluation

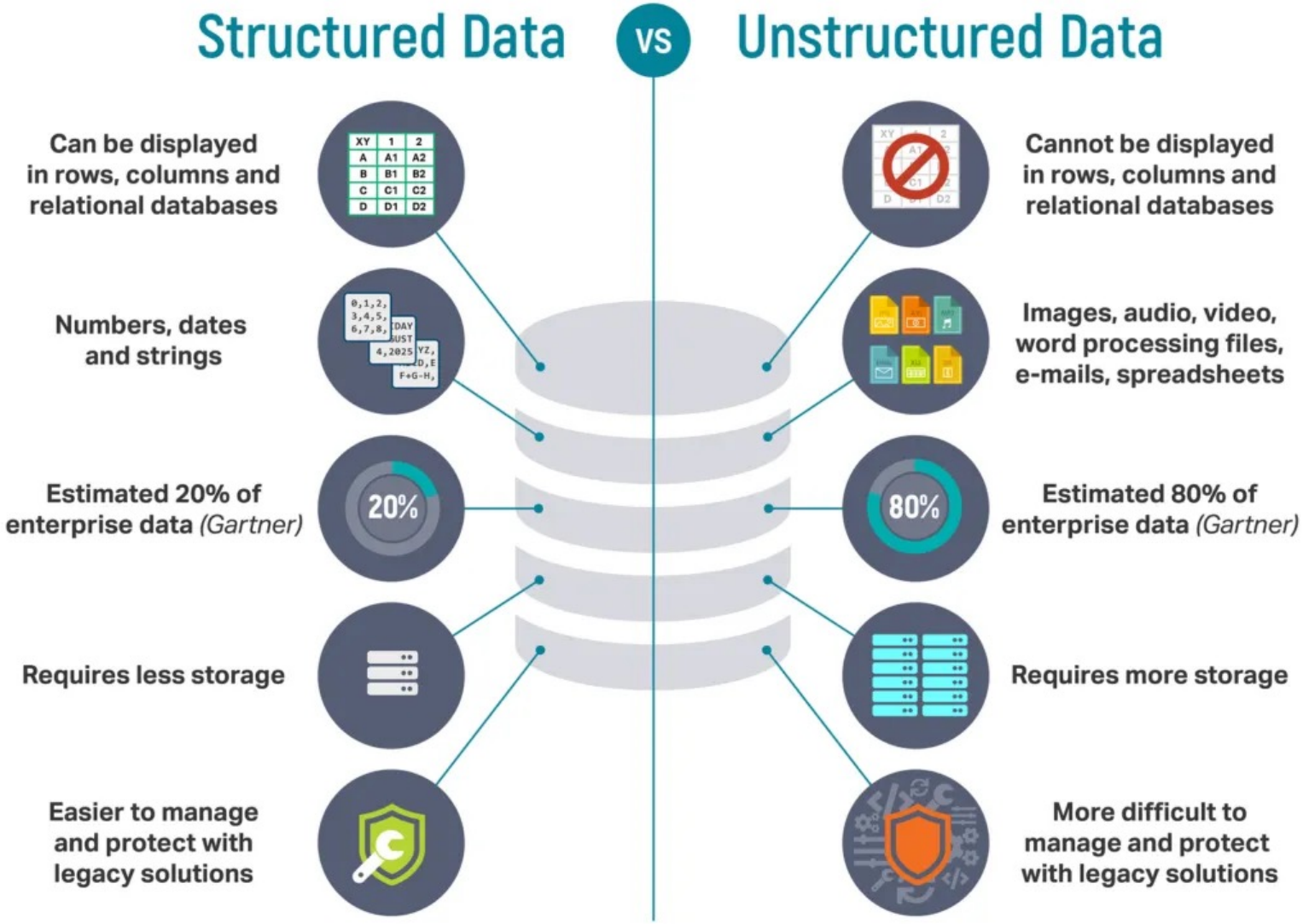
- Performance evaluation
- Time evaluation

6. Deployment

- Execution scheduling
- Performance tuning
- Integration

Problem Statement Checklist

1. Who will use your system?
2. Why will they use it?
3. What is the goal the company wants to achieve?
4. How can we measure such progress towards the goal?
5. What data do we need?
6. When do we need the data to be available?
7. What outcome should we produce?
8. When should the outcome become available?
9. What constraints should we comply with (e.g. regulation, business processes, ...)?
10. (Bonus) What is the budget and the resources we can use to build and run the system?



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

Rescheduling

We need to reschedule 1 lecture and 2 exercise sessions, each 4 hours long.

There will be no difference between lectures and exercise sessions.

Please send me an email by next Friday with **your preferences**.

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.

Data Exploration

The Grammar of Datasets

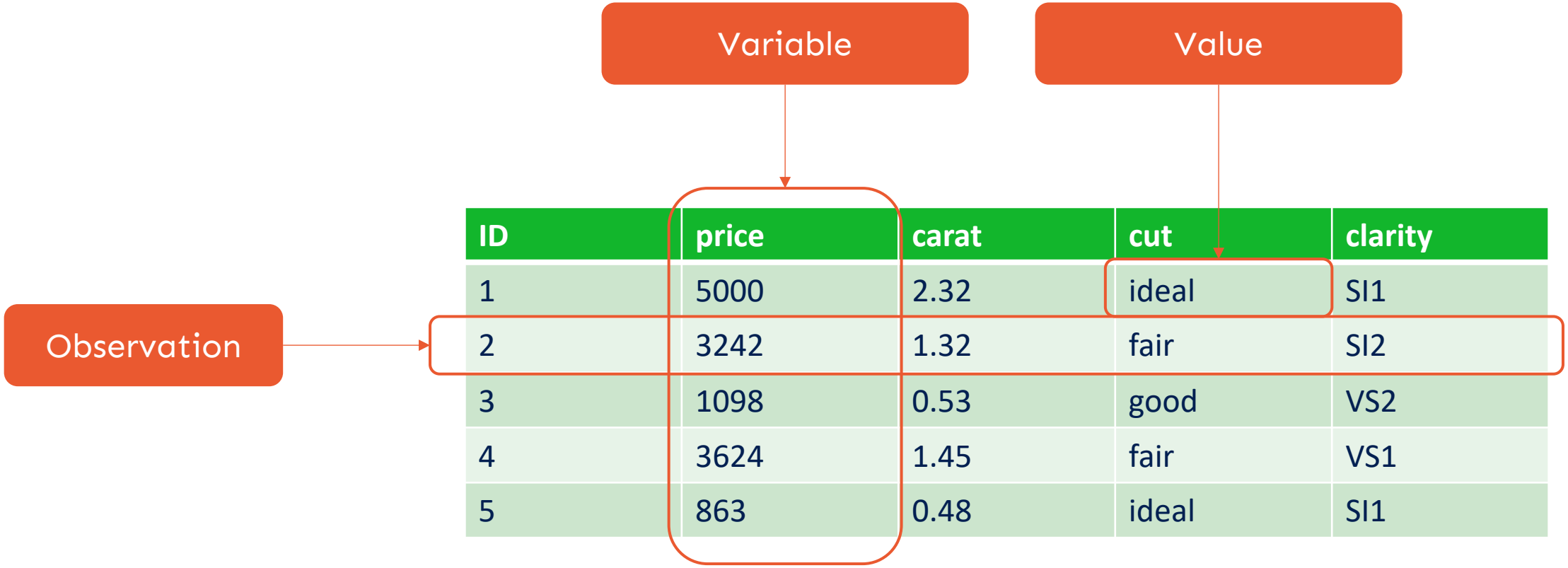
Variable: a property that can be measured.

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

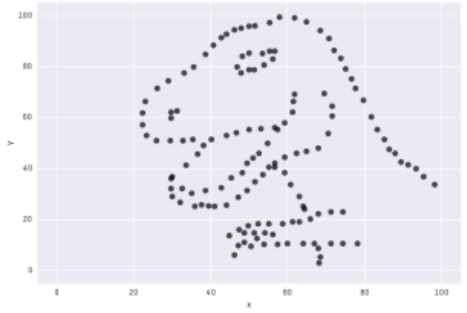
The most common summary statistics for **interval and ratio** variables are:

- **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)

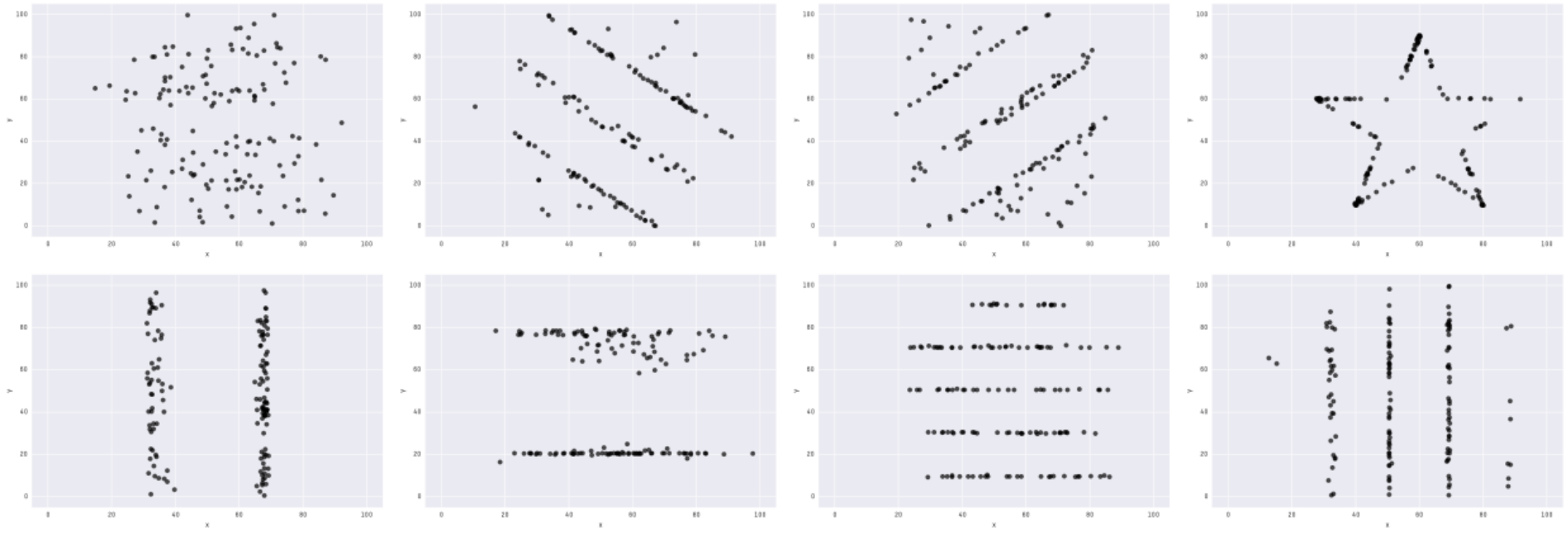
Again, the most informative quantity is the full distribution of the data.


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:

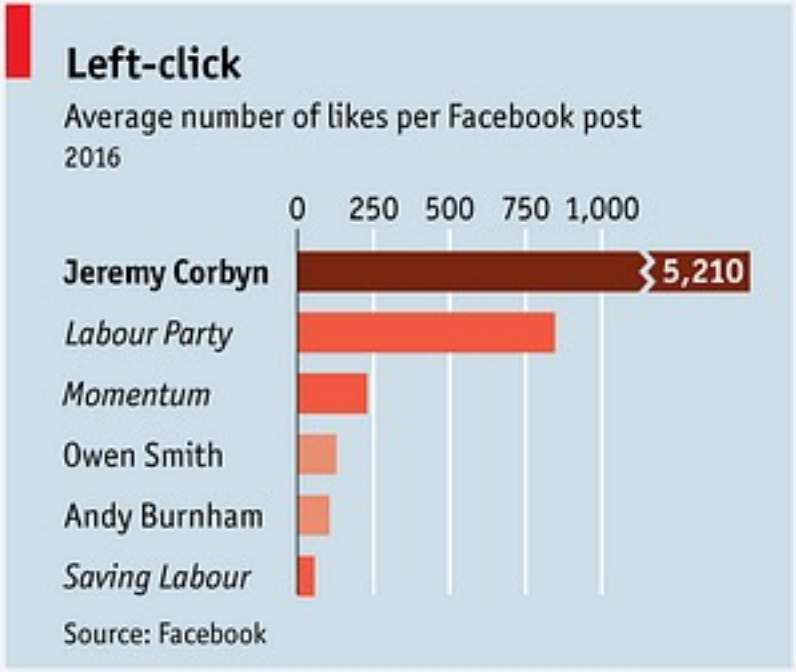
- S. Leo, Mistakes, we've drawn a few
- E. Fabbiani, The traps of data visualization
- C. O. Wilke, Fundamentals of data visualization [chap. 1]

Discussion

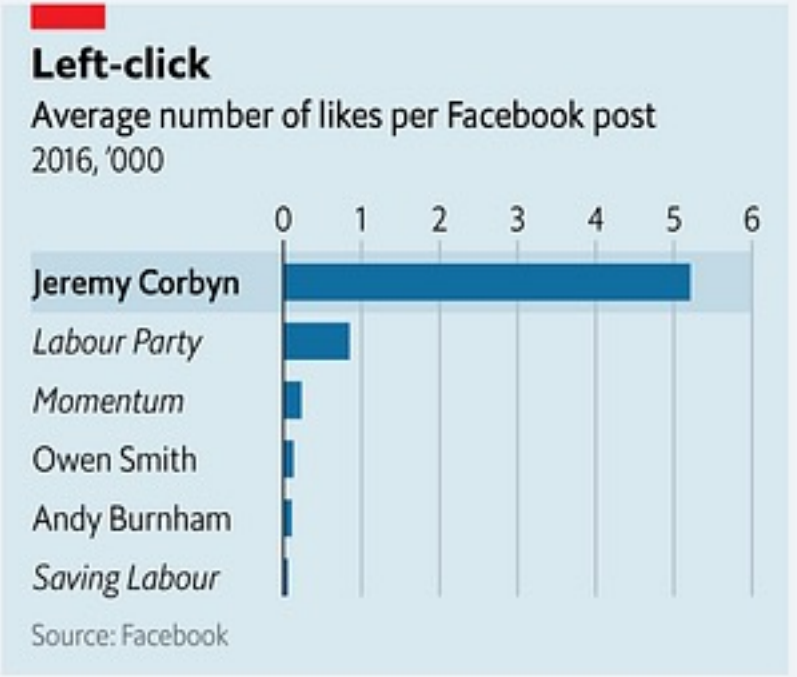
Let’s discuss and criticise some visualizations!

From Sarah Leo, The Economist

Original



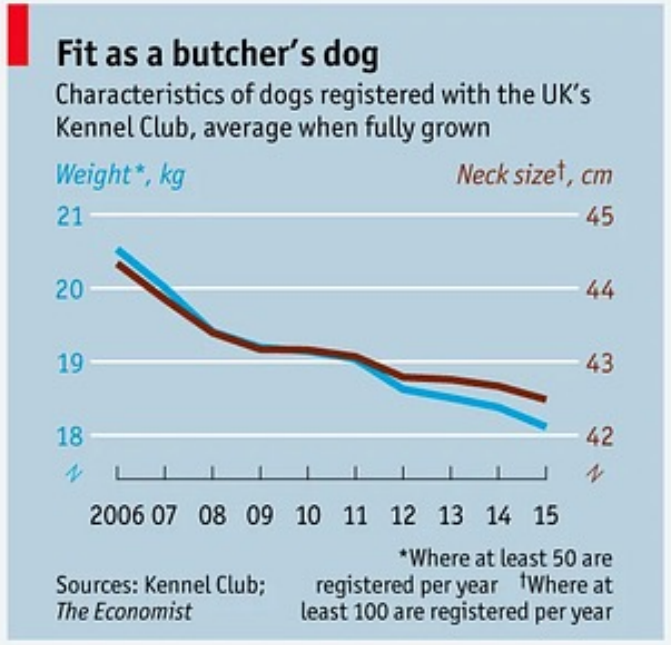
Better



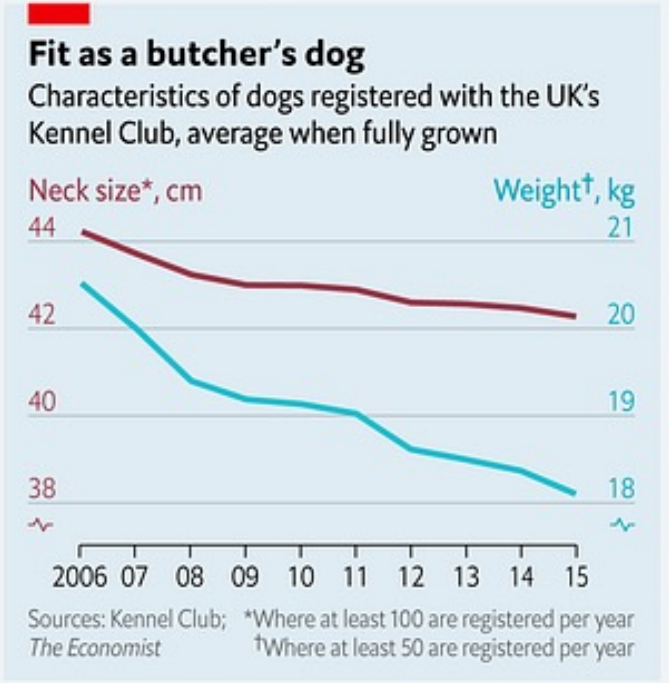
https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368

From Sarah Leo, The Economist

Original



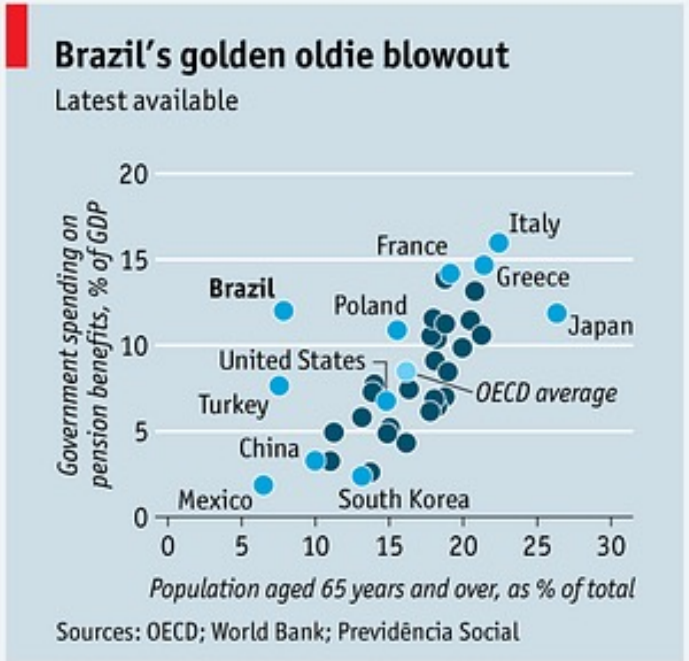
Better



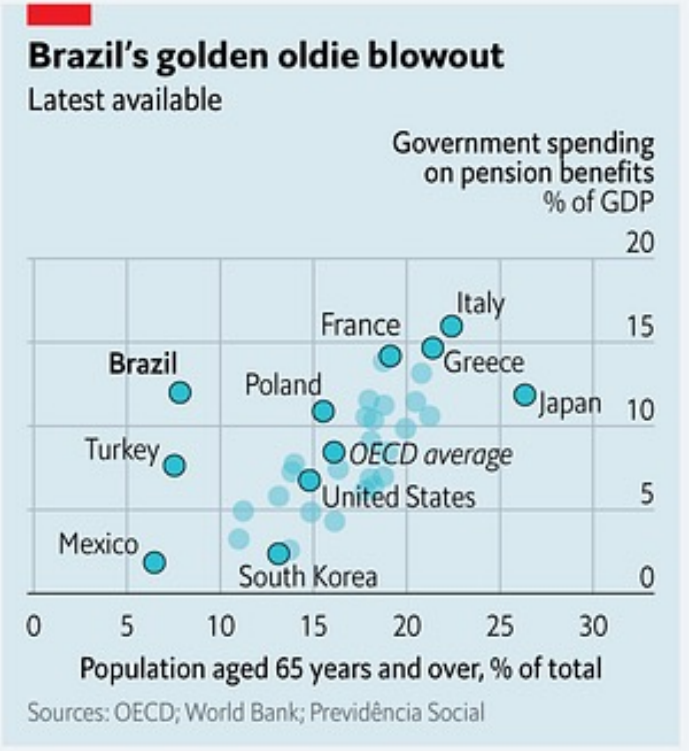
https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368

From Sarah Leo, The Economist

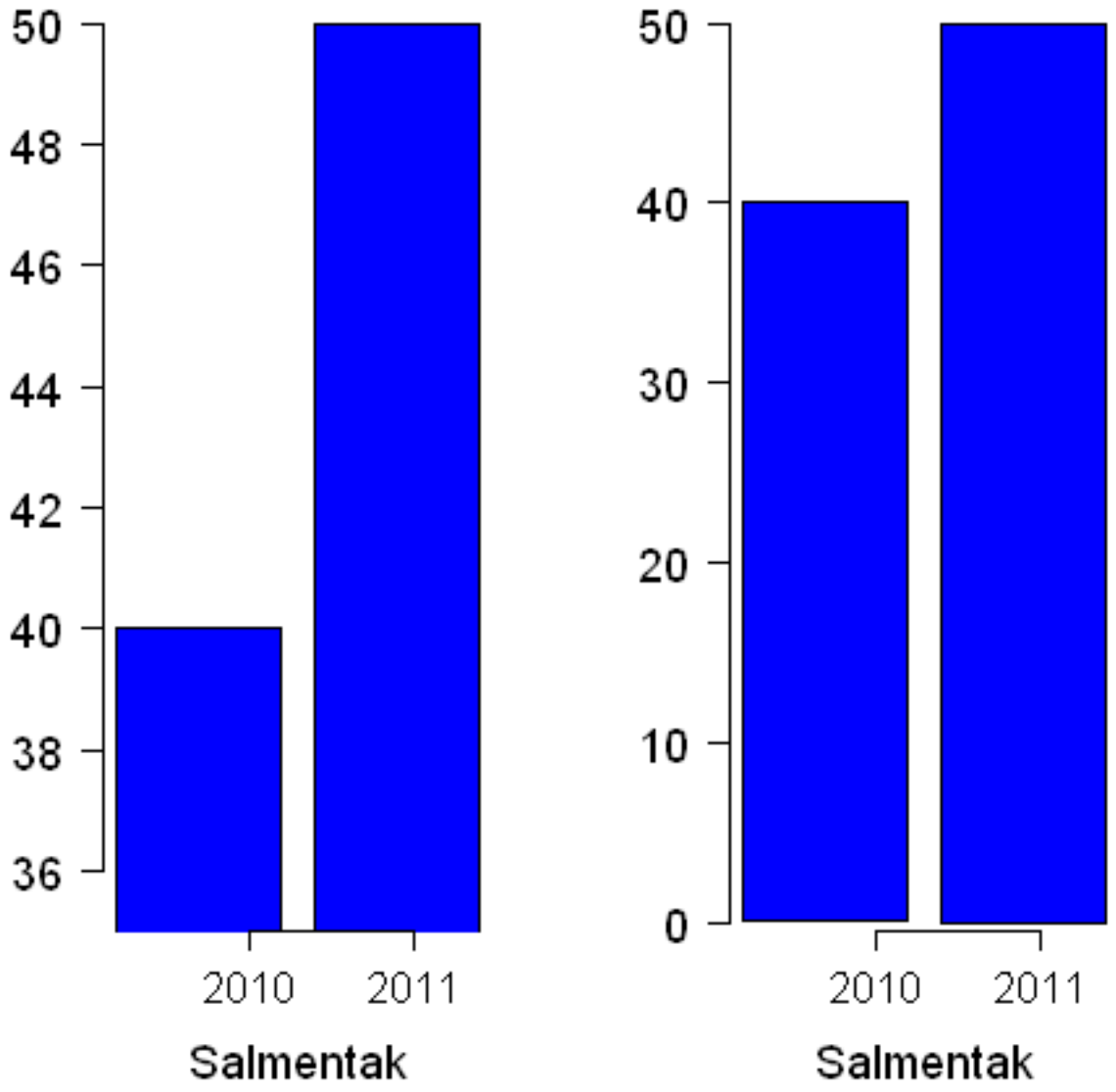
Original



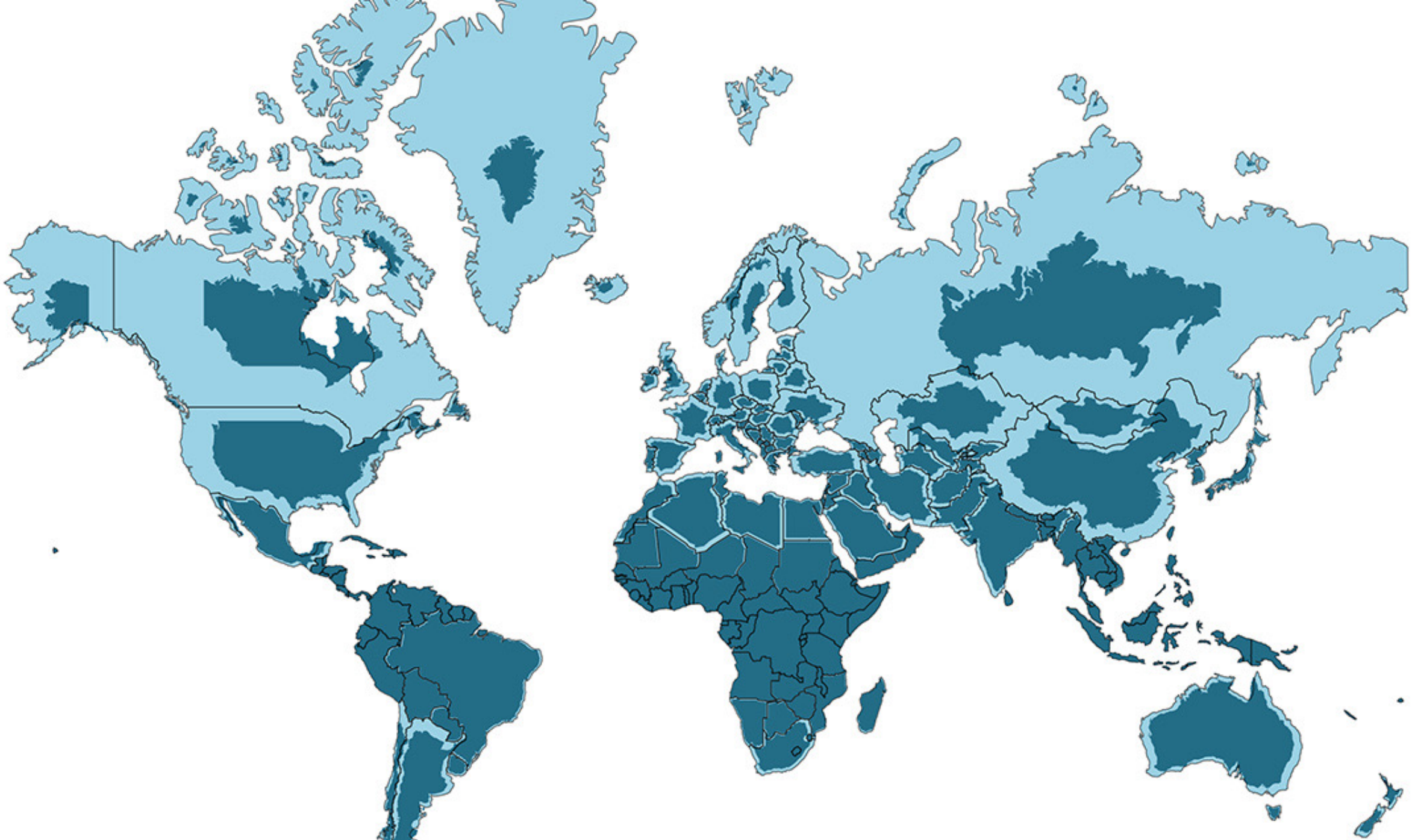
Better



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



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.

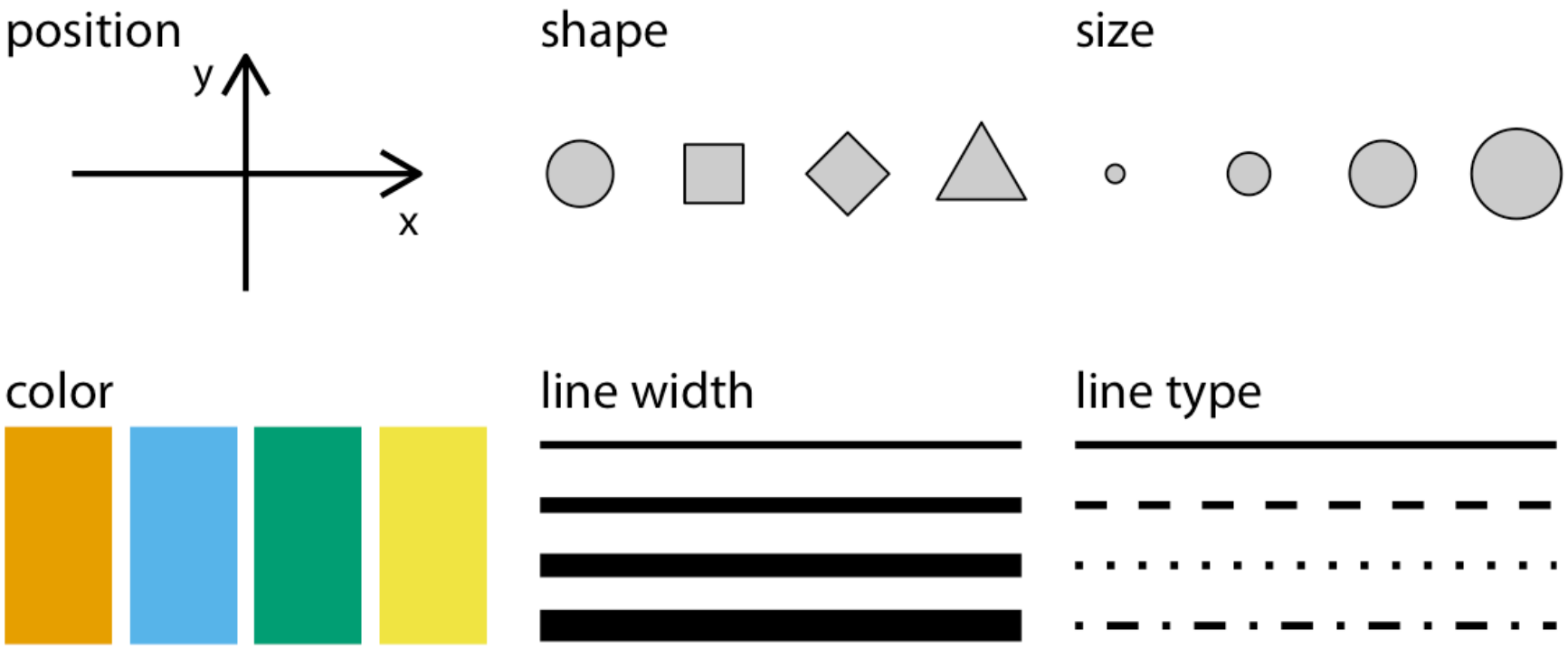
By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

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

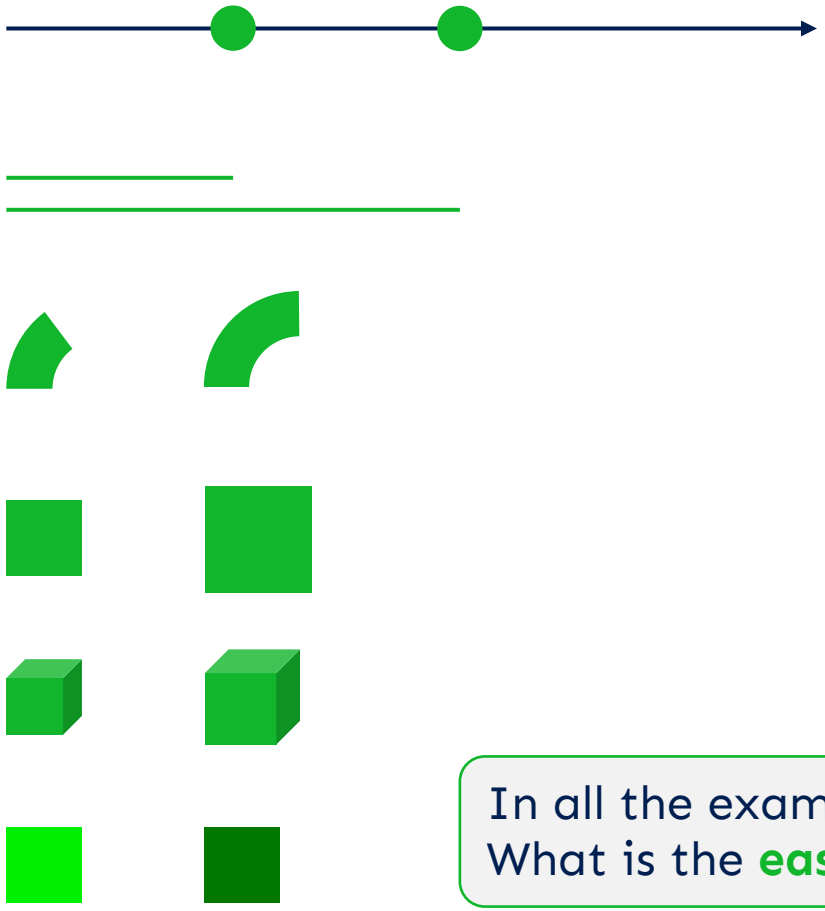
5. Volume

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

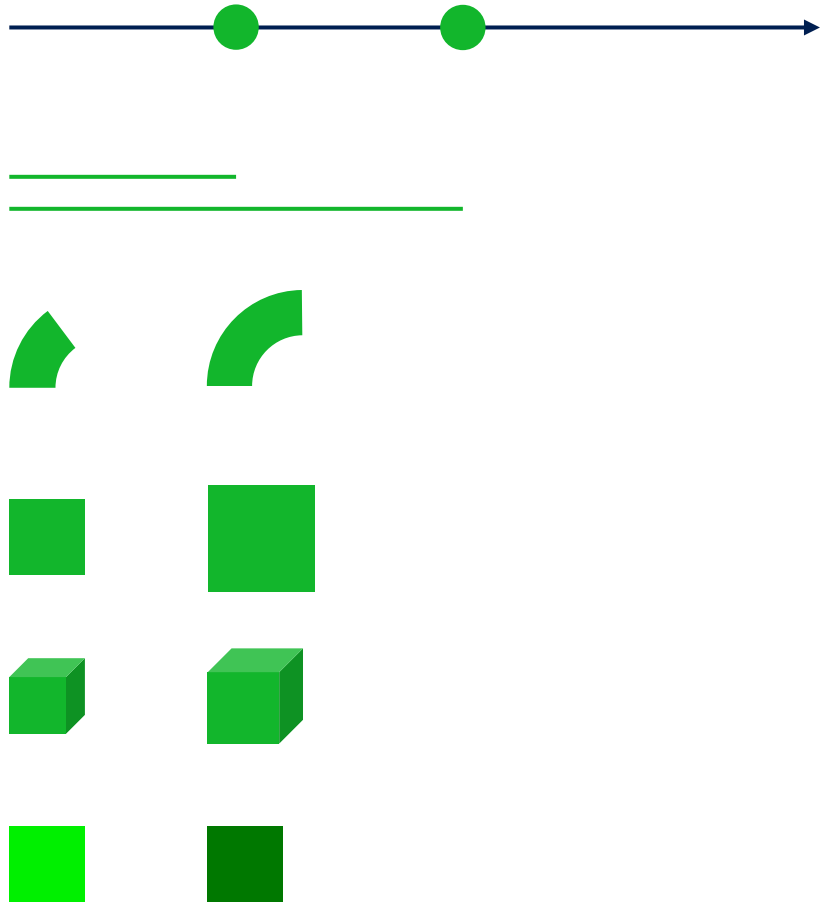


In all the examples, the ratio is 2:1.
What is the **easiest** representation to understand?

Mapping Numerical Variables

Perception accuracy

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



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

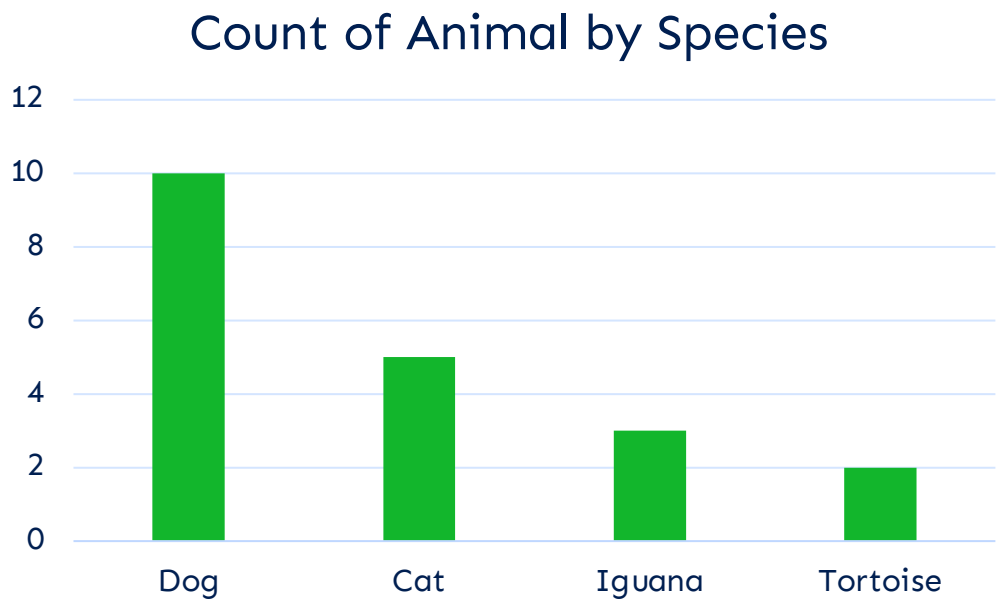
Plotting 1D Data

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

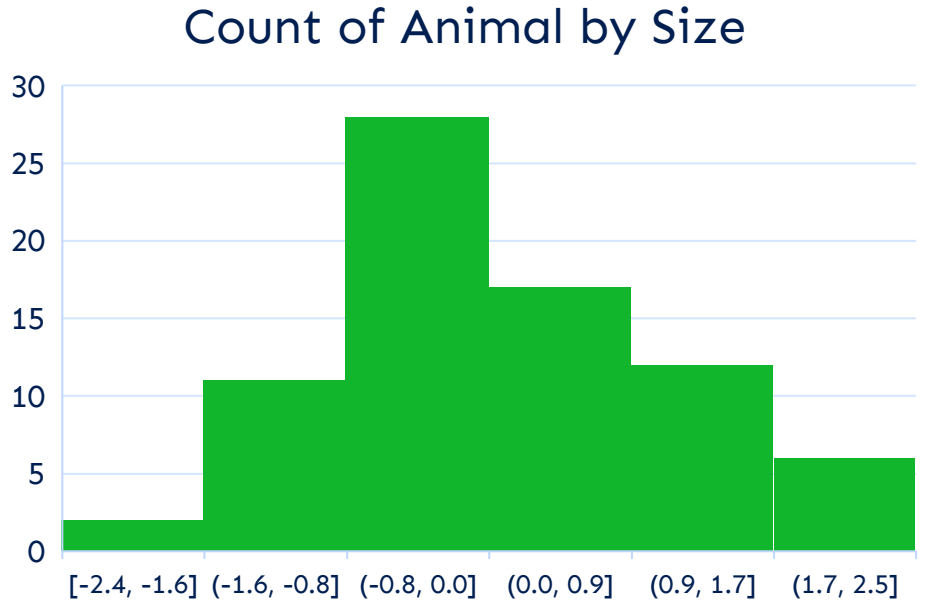
For categorical data, the distribution represents the count or proportion of samples in each category.

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.

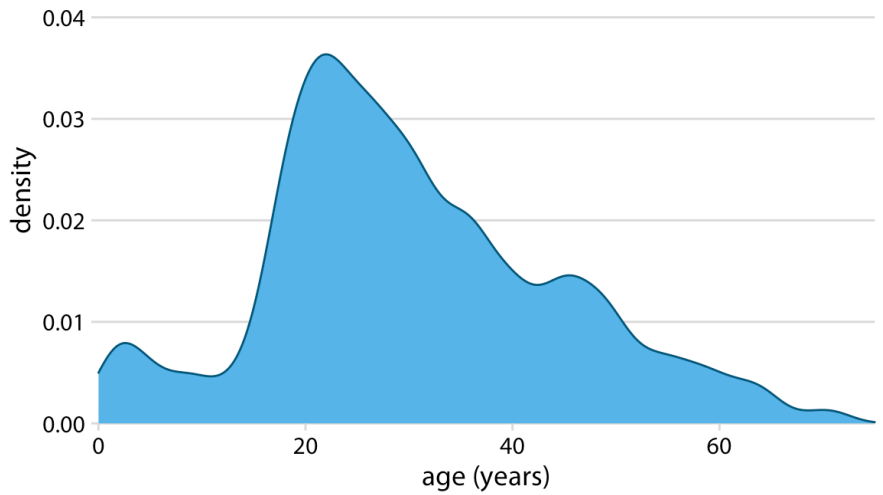


Histogram for numerical variables. Uses both length and position as aesthetics.

Area Charts and KDE

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

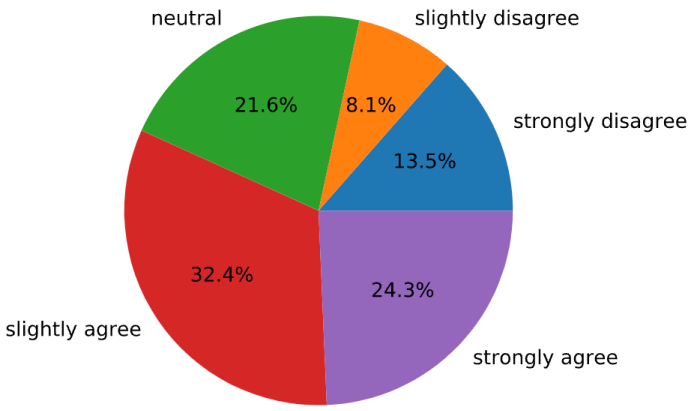
They use the same aesthetics and convey the same information as histograms and are as effective as histograms.

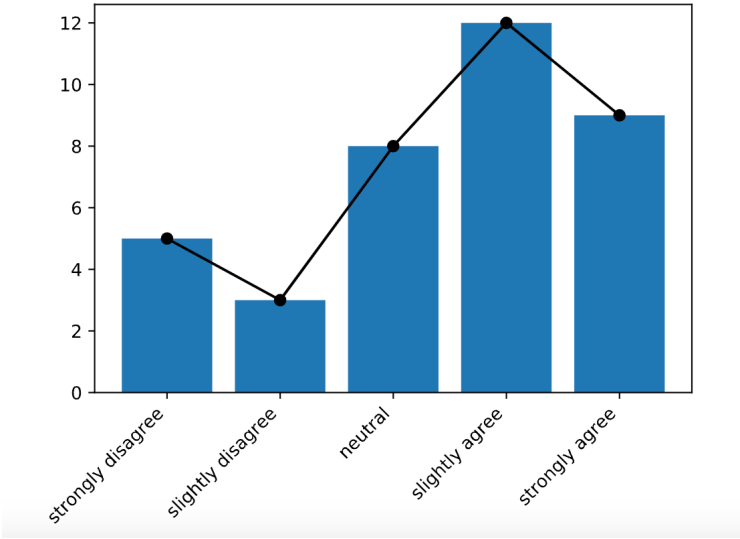


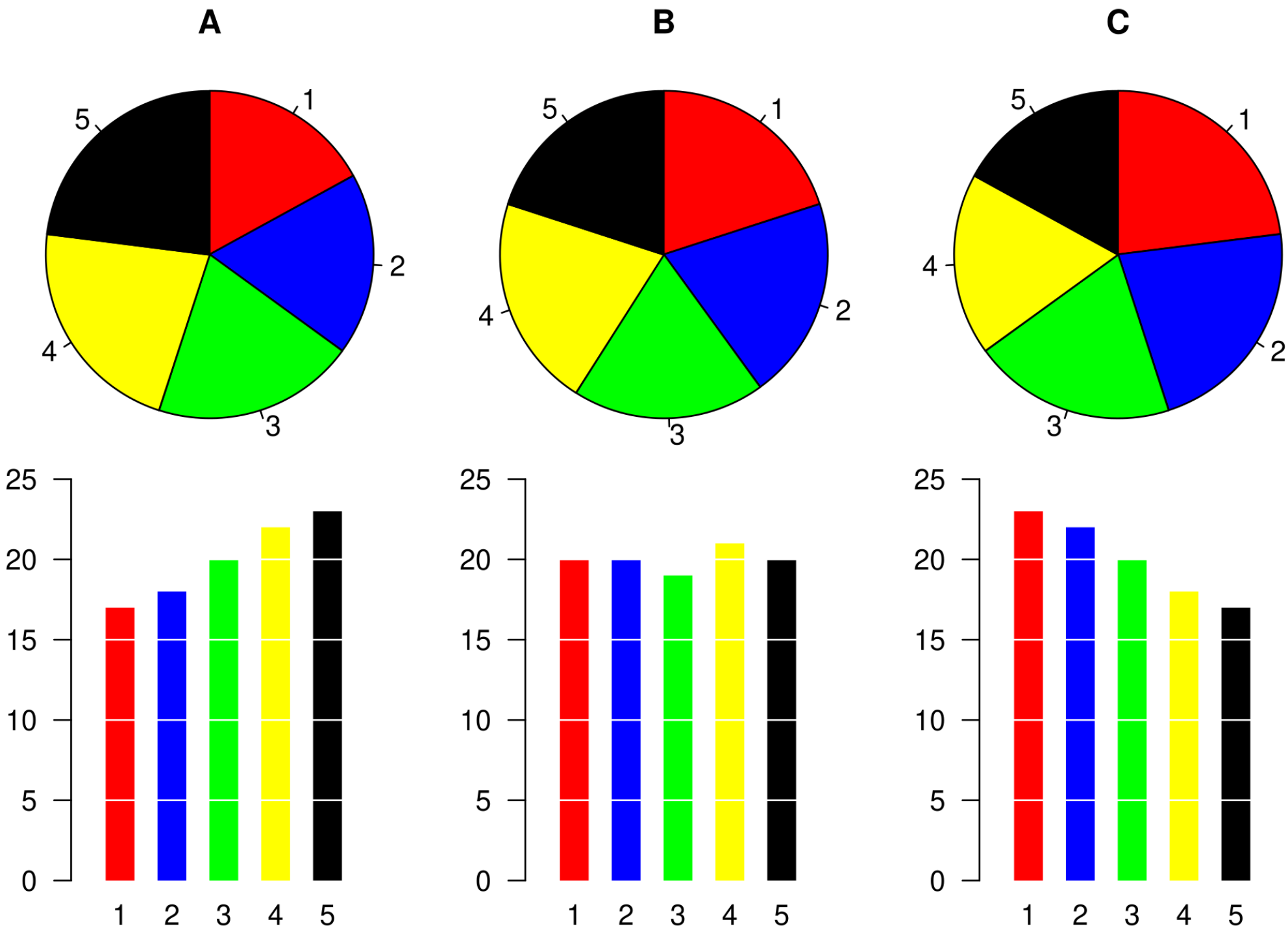
Do not...

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.






A
B
C
1
2
3
4
5
0
5
10
15
20
25

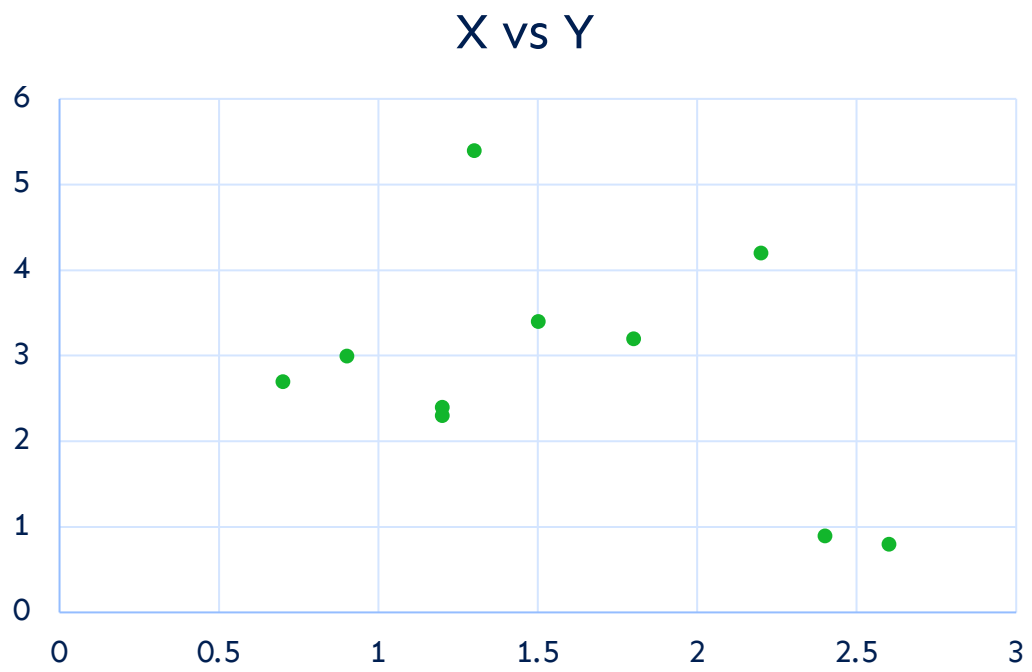
Plotting 2D Data

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.

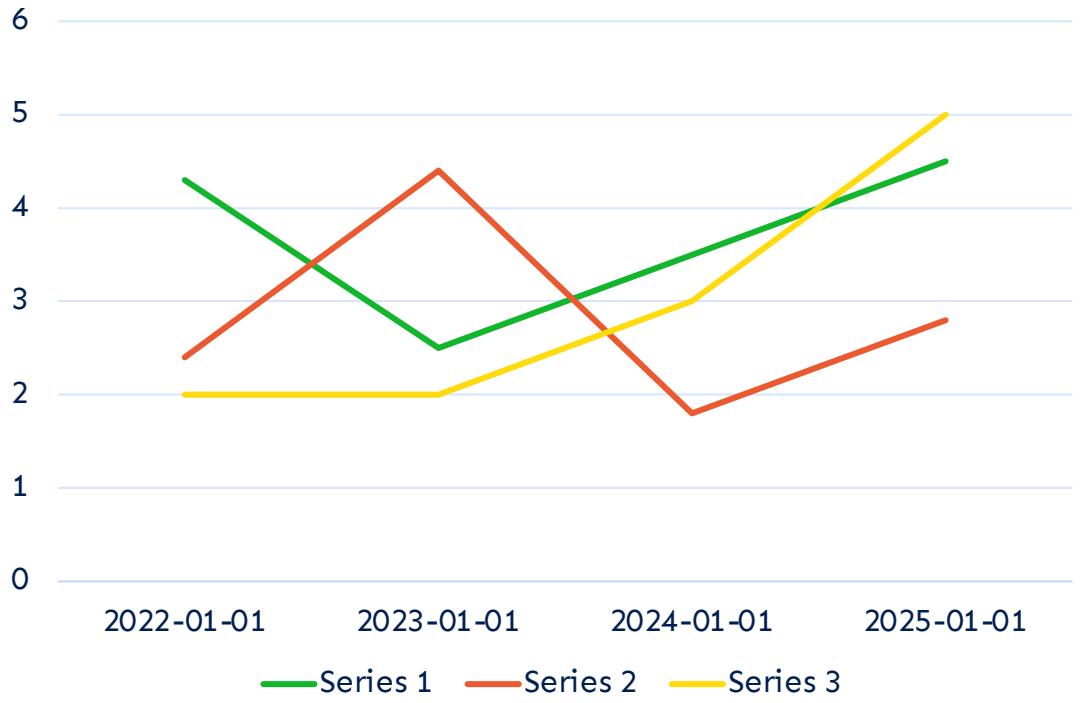
Scatter Plots

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.



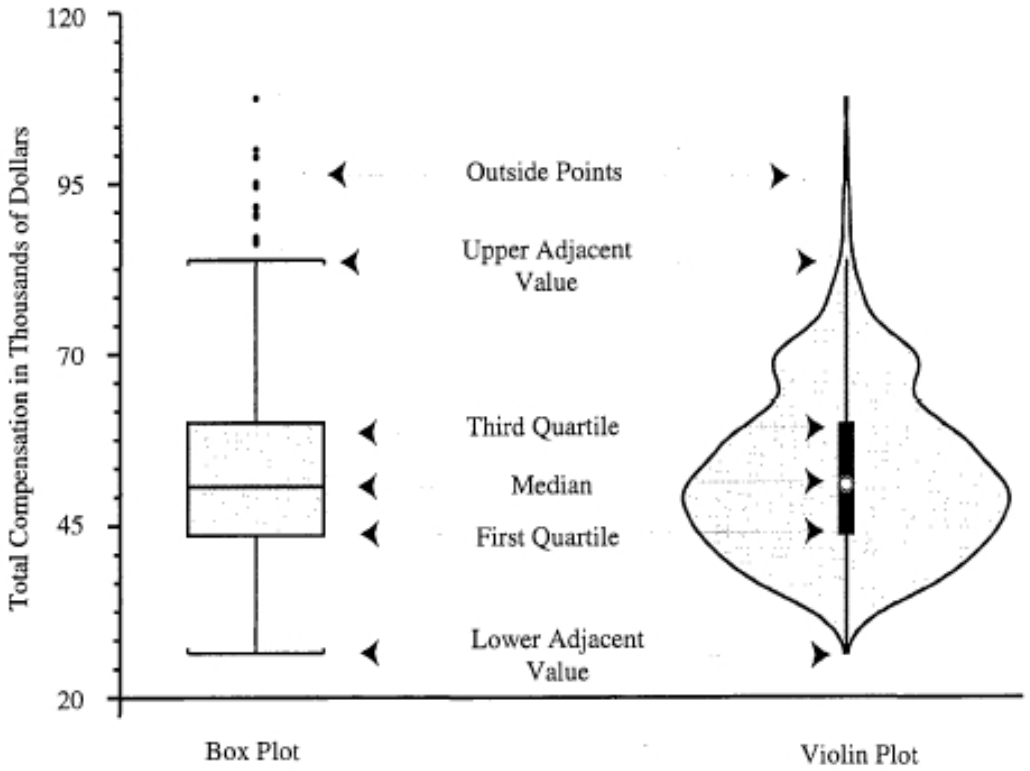
Line Charts

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.



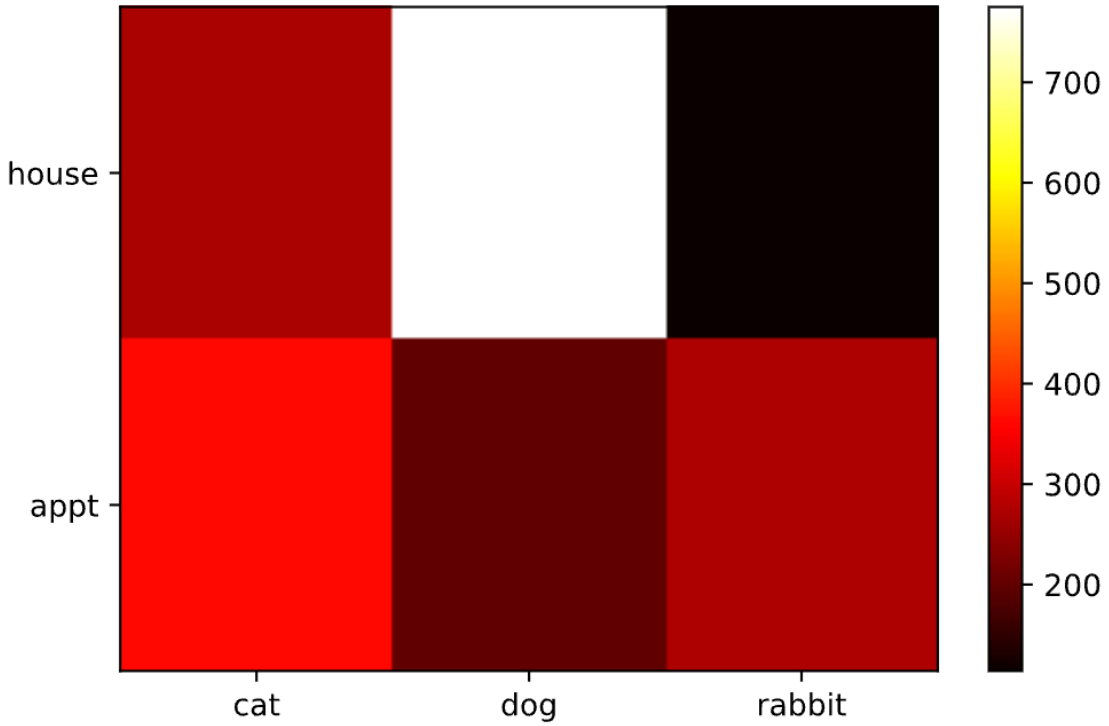
Box and Violin Plots

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.

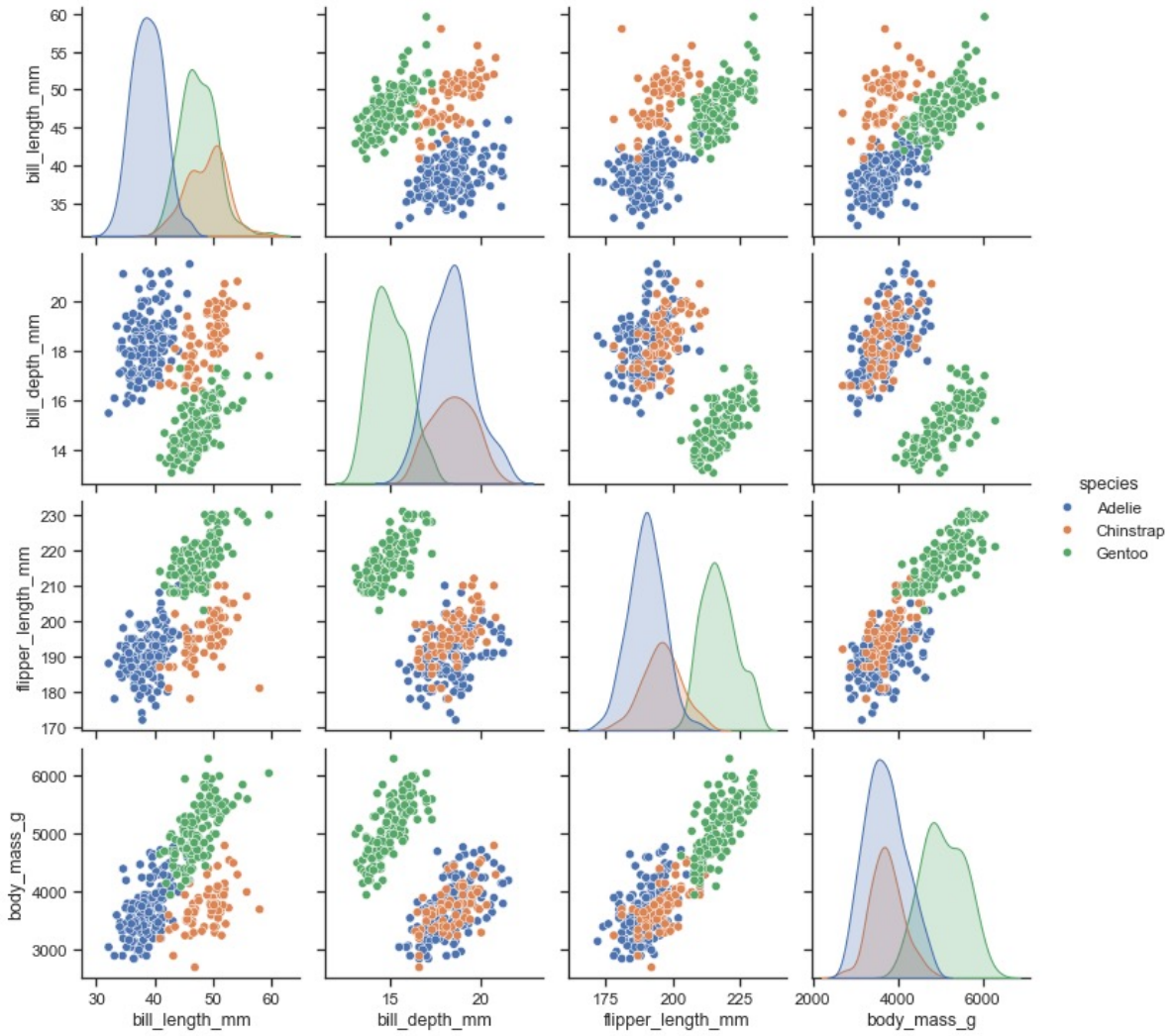


Plotting Higher-Dimensional Data

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**.

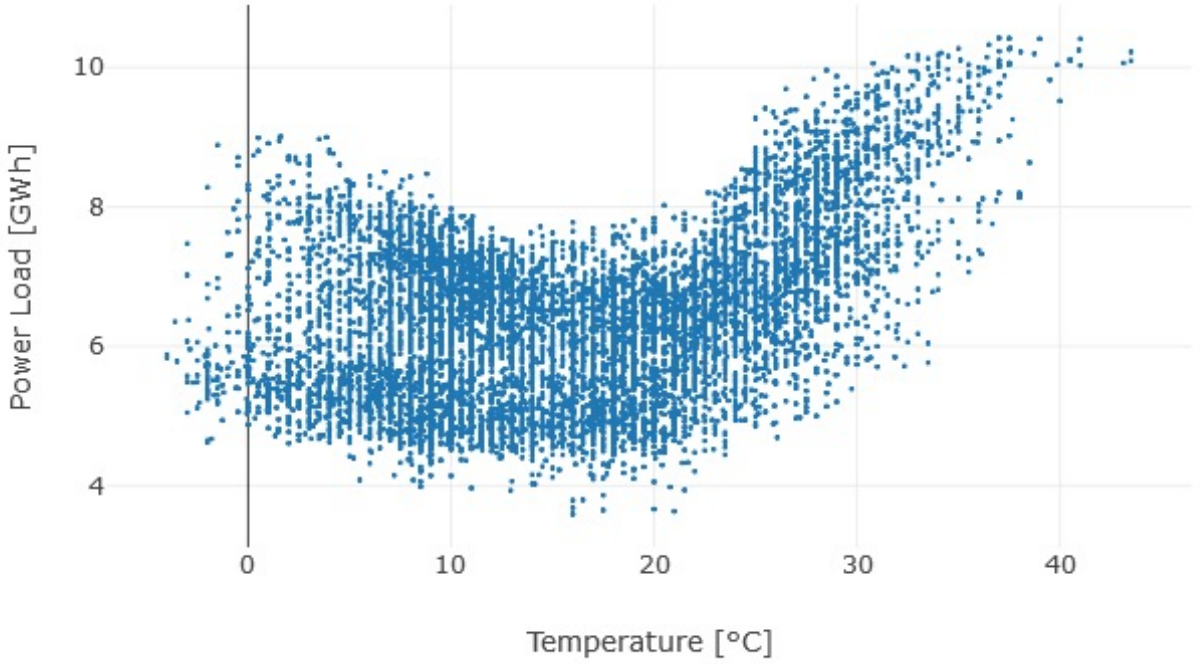
Alternatively, multiple charts can be created and arranged spatially, such as with scatterplot matrices.



Discussion

Let’s draw some insights from a chart!

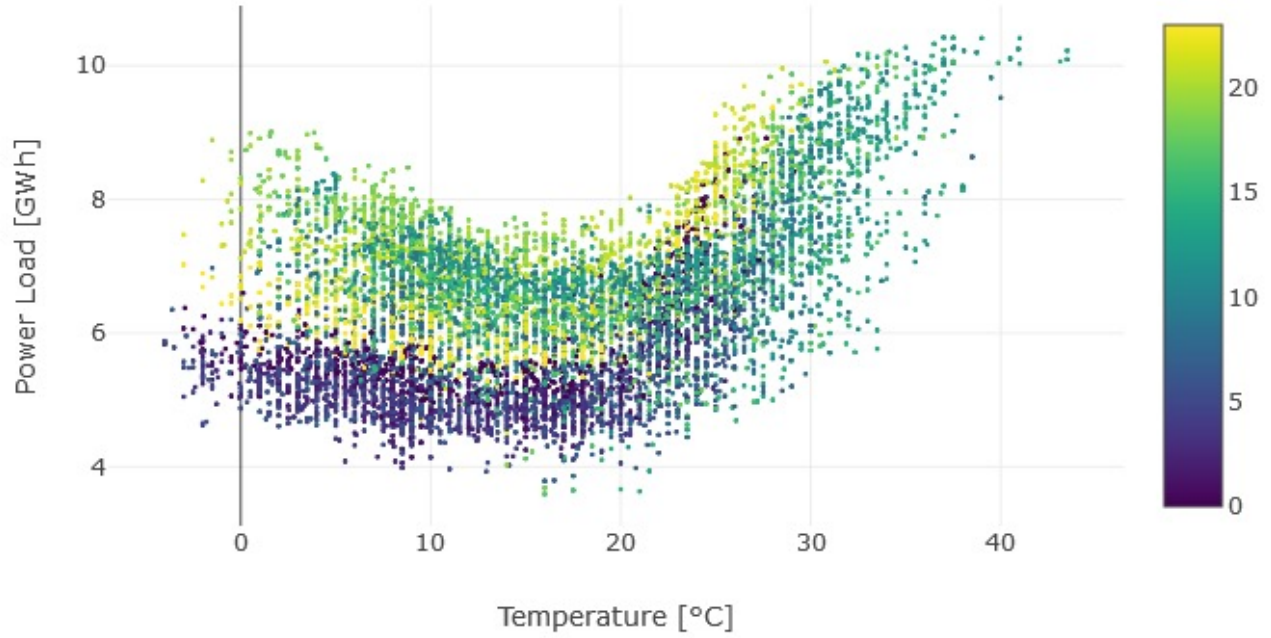
Power Load vs Temperature



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

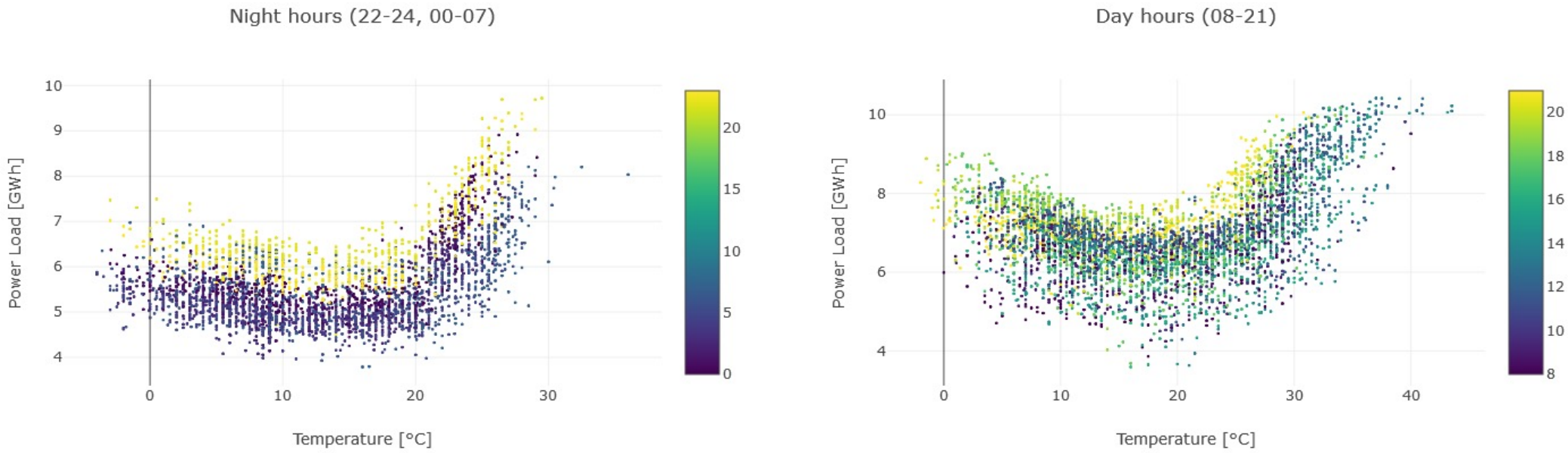
Power Load vs Temperature



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


Data
Cleaning

Tidy datasets are all alike, but every messy dataset is messy in its own way.

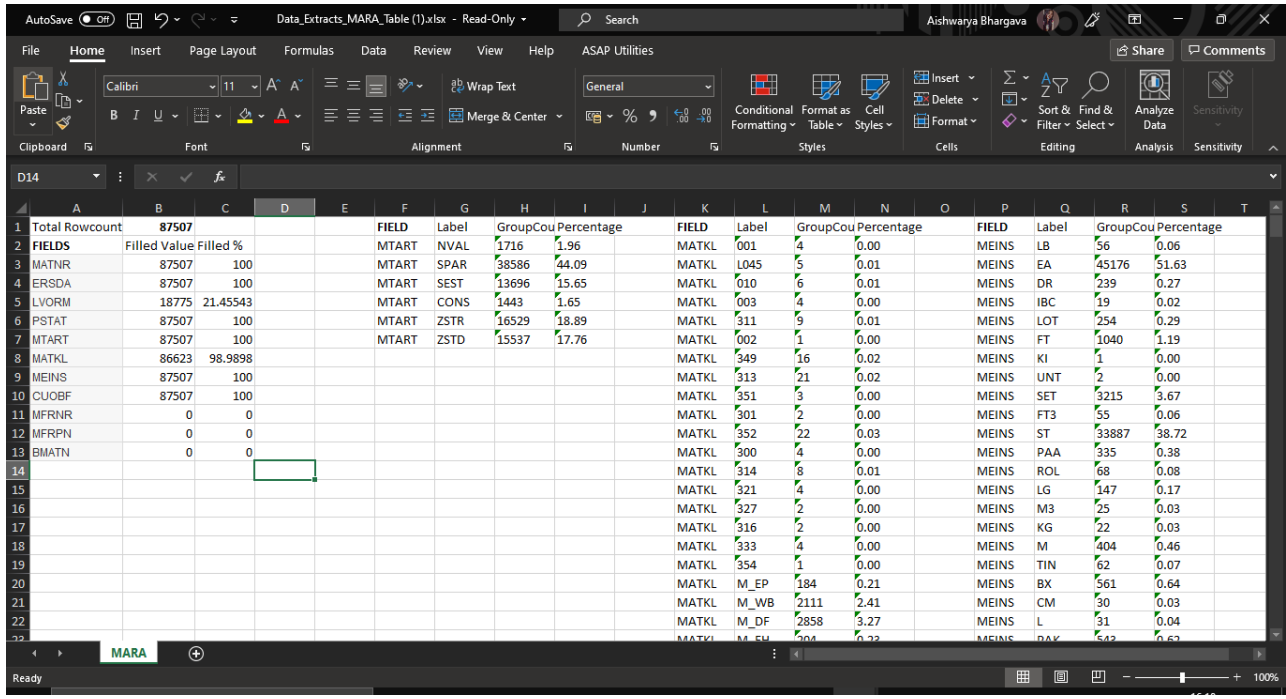
Hadley Wickham

Issues to Address

- Bad formatting
- Multiple variables in a single column
- Duplicated values
- Unrealistic values
- Wrong strings for categorical variables
- Outliers
- Missing values (and their placeholders)

Bad Formatting

Managing multiple tables in an Excel workbook can be challenging, as most libraries are designed to read only one table per sheet.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Total Rowcount	87507				FIELD	Label	GroupCou	Percentage		FIELD	Label	GroupCou	Percentage		FIELD	Label	GroupCou	Percentage	
2	FIELDS	Filled Value	Filled %			MTART	NVAL	1716	1.96		MATKL	001	4	0.00		MEINS	LB	56	0.06	
3	MATNR	87507	100			MTART	SPAR	38586	44.09		MATKL	L045	5	0.01		MEINS	EA	45176	51.63	
4	ERSDA	87507	100			MTART	SEST	13696	15.65		MATKL	010	6	0.01		MEINS	DR	239	0.27	
5	LVORM	18775	21.45543			MTART	CONS	1443	1.65		MATKL	003	4	0.00		MEINS	IBC	19	0.02	
6	PSTAT	87507	100			MTART	ZSTR	16529	18.89		MATKL	311	9	0.01		MEINS	LOT	254	0.29	
7	MTART	87507	100			MTART	ZSTD	15537	17.76		MATKL	002	1	0.00		MEINS	FT	1040	1.19	
8	MATKL	86623	98.9898								MATKL	349	16	0.02		MEINS	KI	1	0.00	
9	MEINS	87507	100								MATKL	313	21	0.02		MEINS	UNT	2	0.00	
10	CUOBF	87507	100								MATKL	351	3	0.00		MEINS	SET	3215	3.67	
11	MFRNR	0	0								MATKL	301	2	0.00		MEINS	FT3	55	0.06	
12	MFRPN	0	0								MATKL	352	22	0.03		MEINS	ST	33887	38.72	
13	BMATN	0	0								MATKL	300	4	0.00		MEINS	PAA	335	0.38	
14											MATKL	314	8	0.01		MEINS	ROL	68	0.08	
15											MATKL	321	4	0.00		MEINS	LG	147	0.17	
16											MATKL	327	2	0.00		MEINS	M3	25	0.03	
17											MATKL	316	2	0.00		MEINS	KG	22	0.03	
18											MATKL	333	4	0.00		MEINS	M	404	0.46	
19											MATKL	354	1	0.00		MEINS	TIN	62	0.07	
20											MATKL	M_EP	184	0.21		MEINS	BX	561	0.64	
21											MATKL	M_WB	2111	2.41		MEINS	CM	30	0.03	
22											MATKL	M_DF	2858	3.27		MEINS	L	31	0.04	

Solution

To simplify this process, consider reformatting the workbook to ensure each sheet contains only one table.

Alternatively, you can just avoid Excel altogether.

Multiple Variables in a Column

Condensing multiple variables into a single column can complicate processing and lead to information loss.

Solution

Process the value to map each variable to a single column.

	A	B
1	Name	Address
2	Paolo Rossi	Piazza Affari 1, Milano, Italy
3	Mario Bianchi	Corso Venezia 1, Milano, Italy
4	Bruno Neri	Corso Buenos Aires 1, Milano, Italy
5		

Duplicated Values

Datasets often include a primary key, a variable designed to uniquely identify each sample, such as a person's tax code, a user's email, or a timestamp in a time series.

Errors and poor validation can result in duplicate keys.

	A	B	C
1	Company VAT	Revenue	EBITDA
2	01234567891	€ 1,000,000.00	2%
3	01234567892	€ 1,500,000.00	3%
4	01234567893	€ 2,000,000.00	6%
5	01234567891	€ 1,100,000.00	10%

Solution

Remove one of the duplicate samples, using a context-dependent rule to determine which to keep.

Unrealistic Values

By "unrealistic" values, we refer to those that cannot be correct, as recording them would be impossible.

These values often result from errors in the recording or data collection process.

	A	B	C
1	Date	City	Temperature
2	2024-07-01	Palermo	33.00
3	2024-07-02	Palermo	35.00
4	2024-07-03	Palermo	-3.00
5	2024-07-04	Palermo	24.00

Solution

Treat the value as missing and use one of the techniques outlined in the following slides.

Wrong Strings for Categorical

The data collection process often stores categorical values as strings, which can lead to the same category being represented by different strings.

Solution
Identify all strings representing the same category and replace them with the correct one.

	A	B	C
1	Dog ID	Race	Weight
2	1	chihuahua	3.50
3	2	chiwawa	3.65
4	3	chihuaua	3.44
5	4	chihuahua	3.23
6	5	chihuahua	3.32

Missing Data

Handling missing data is a fascinating and challenging topic. We will only discuss the most basic techniques.

More advanced methods are available both in the literature and in python packages.

Missing Data

Types of missing data:

- **Missing Completely at Random**: there is no pattern in missing value (e.g. measurement errors due to random noise in a sensor)
- **Missing at Random**: the pattern of missing value only depends on known variables (e.g. in a survey, sex is recorded, and men are more likely to refuse to answer another question)
- **Missing not at Random**: the pattern of missing value depends on the variable where the missing data appear (e.g. people with higher income are more likely to refuse to report their income)

It is usually very hard, if not impossible, to prove that values are missing at random.

Dropping Missing Data?

Missing Completely at Random: can be safely dropped.

Missing at Random: can be dropped if all the variables related to missing values are considered.

Missing not at Random: informative, should not be dropped.

The ratio of missing data in each variable is an important factor to consider when deciding whether to discard missing data. If the dataset is large, dropping the missing data is often the easiest and safest approach. However, if data is scarce, this method could be risky.

Imputing Missing Data

- **Mean / Mode imputation**: replacing the missing values with the mean or mode. It is easy, but it modifies the distribution of the variable and the joint distribution with other variables.

- **Similarity-based imputation**: replacing the missing value with a value from a similar sample. This method is particularly suitable for data missing at random, but it relies on a well-defined measure of similarity, which can often be challenging to determine.

- **Model-based imputation**: replacing missing values with predictions generated by a model using the available variables. This approach is well-suited for data missing at random but depends on the quality and accuracy of the model used for imputation.

Not Imputing Missing Data

For categorical variables, a **new category** labelled "MISSING" can be created. For numerical variables, a new Boolean variable indicating whether the original variable is missing (1 if missing) can be added.

However, many modern training algorithms, particularly tree-based models, **can handle missing values directly**.

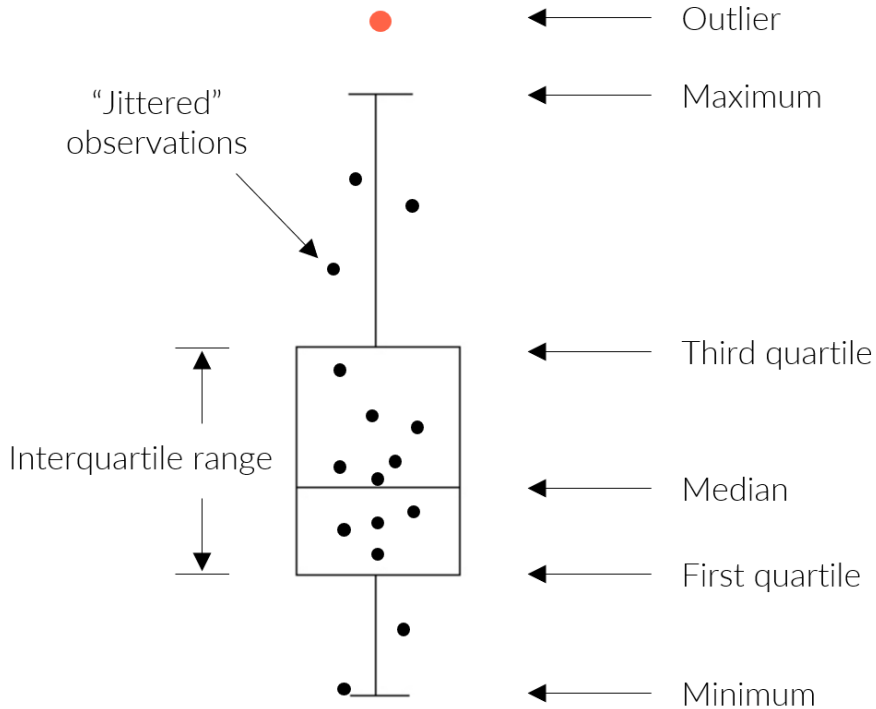
When the model supports missing values, they should not be imputed.

Outliers

An outlier is a value that falls outside the distribution of the rest of the data.

Outliers may result from data collection errors, but they can also represent legitimate values.

For example, rogue waves are outliers in wave height, and market crashes are outliers in stock returns. Such values **should not be excluded** from your analysis.



Handling Outliers

If outliers result from errors in the data collection process, they should be **treated as unrealistic or missing values**.

If they are legitimate values, the following approaches can be considered:

- **Exclude** them from the dataset if the goal is to prevent the model from learning from such samples.
- **Clip** them to a maximum or minimum value, with thresholds determined by the variable's distribution.
- **Leave them unchanged** if the intention is for the model to learn from them. In some cases, it may be more effective to develop specialized models to handle extreme situations.

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/JL2U9UBRZMUHeVP4A


The End