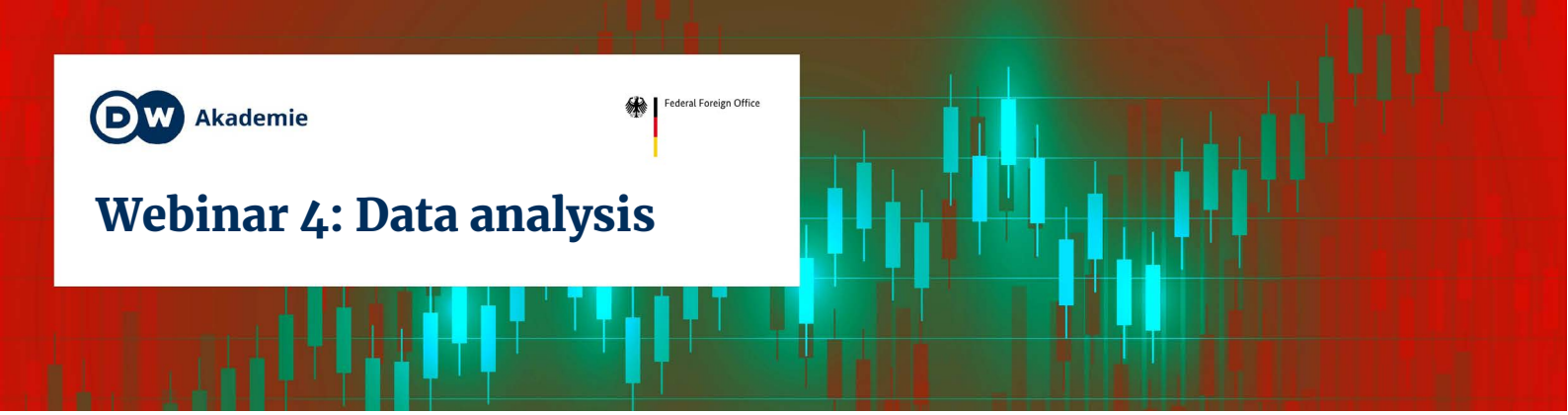


Webinar 4: Data analysis

Clean data needs to be analyzed. How to go about this? Interview your data like you would do with a person!

Data interview

Divide your main hypothesis into research questions that you can answer using your data. Be specific!

Example:

- When did the GDP last decrease by more than 3%?
- Which countries received the most exports in 2018?
- Which cities produce the most CO2?

Tip: Write down your research questions so you don't get lost in the data.

Useful tricks

MERGE DATASETS

Combine different sources to generate new insight
Example: Merge population data with geographical data
Watch out! Not all datasets are immediately comparable:

- Definitions: Geographical regions, for example, can be defined differently
- Methodology: The time of collection, for example, might be different

VLOOKUP formula in Excel etc. can merge information from two datasets

Tutorial: Excel Dashboards and Reports: The VLOOKUP Function — Dummies.com
How to Use VLOOKUP in Excel — HowToGeek.com

CALCULATE NEW COLUMNS

Get new information from the available data
Example: Calculate Covid-19 infections per 100 000 people:

- Number of infections (Column 1) **divided by** population (Column 2) **times** 100000

Tutorial: Use calculated columns in an Excel table — Microsoft.com

PIVOT TABLES

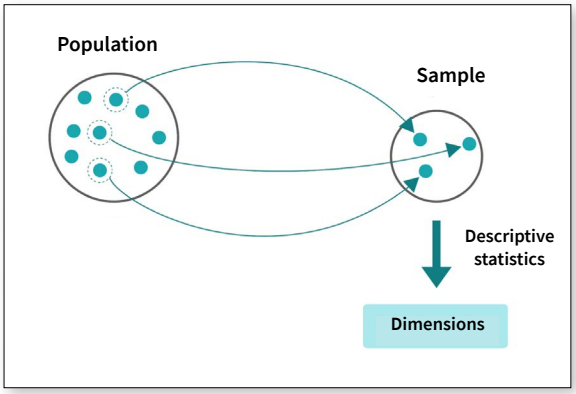
Summarize data by categories in a new table
Example: Sum Covid-19 infections per country to infections per continent
Tutorial: How to Create Pivot Tables in Excel (with Pictures) — WikiHow.com

Statistics basics

Understanding statistics helps you understand studies better and find interesting idea in data yourself.

WHAT IS STATISTICS?

- A collection of methods aimed at describing reality as best possible
- The collection, analysis, interpretation and presentation of data



POPULATION

Group that you want to draw conclusions about

Example: All inhabitants of a country, with the question “Who will they vote for?”

Watch out: The term “study” is not protected and says nothing about reliability

SAMPLE

Small part of the population that is used to draw conclusions about the whole

Example: 1000 inhabitants of a country are randomly chosen and surveyed

Watch out:

- The bigger the sample, the more meaningful the observations

  Example: Polls explained with interactive graphics – Maarten Lambrechts

- Is the sample representative of the basic population?

  Example: Results of a Facebook page poll barely allow conclusions about all followers of the page, let alone a whole country

DESCRIPTIVE STATISTICS

Summarize observations using a few meaningful measures
Example: "On average, 37 % of people vote for Party X."

Mean vs. median

There is more than one way to calculate an average!

- **Median:** Divides a sorted row of values in the middle, so that half are bigger and half are smaller

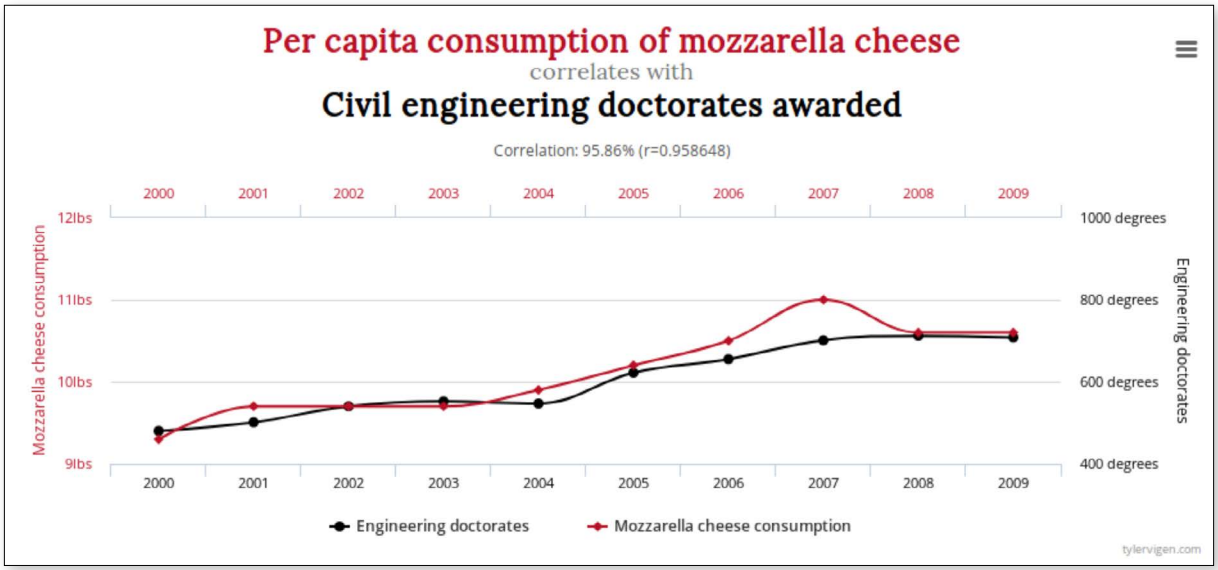
- **Mean:** The sum of all values, divided by the number of values. What's often meant when people say "average".

Watch out: Outliers – values that are unusually low or unusually high – influence the mean, but not the median.

Example: The median income of a country is often much lower than the mean income, because few very rich people distort the mean, but not the median. Note which one is being used!

CORRELATION DOES NOT EQUAL CAUSATION

Values that behave similarly (correlate), don't necessarily cause one another.



Per capita consumption of mozzarella cheese correlates with civil engineering doctorates awarded

Data shows: The more mozzarella cheese is consumed, the more civil engineering doctorates are awarded. But that doesn't necessarily mean that mozzarella cheese causes engineering doctorates, or that engineers consume huge amounts of mozzarella cheese.

Source and more examples: Spurious Correlations

EVERYTHING IS RELATIVE

Absolute numbers are hard to compare. It is often better to relate them to a common basis.
Example:

This map mostly tells us where a lot of people live

Number of unemployed people by EU country

This map tells us that Greece has the highest unemployment in the EU

Share of unemployed people in the active population by EU country

France has a lot of unemployed people. But it also has a lot of people in general. We can only compare between countries if we look at relative values – in this case, the share of unemployed people in the active population. This allows us to see that Greece has the highest unemployment in the EU.

See also: Heatmap — Xkcd.com