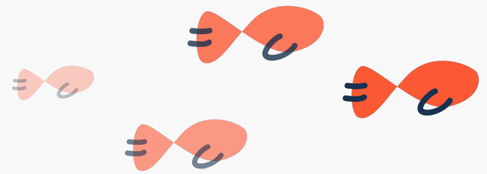
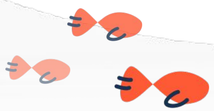
neue fische

School and Pool for Digital Talent

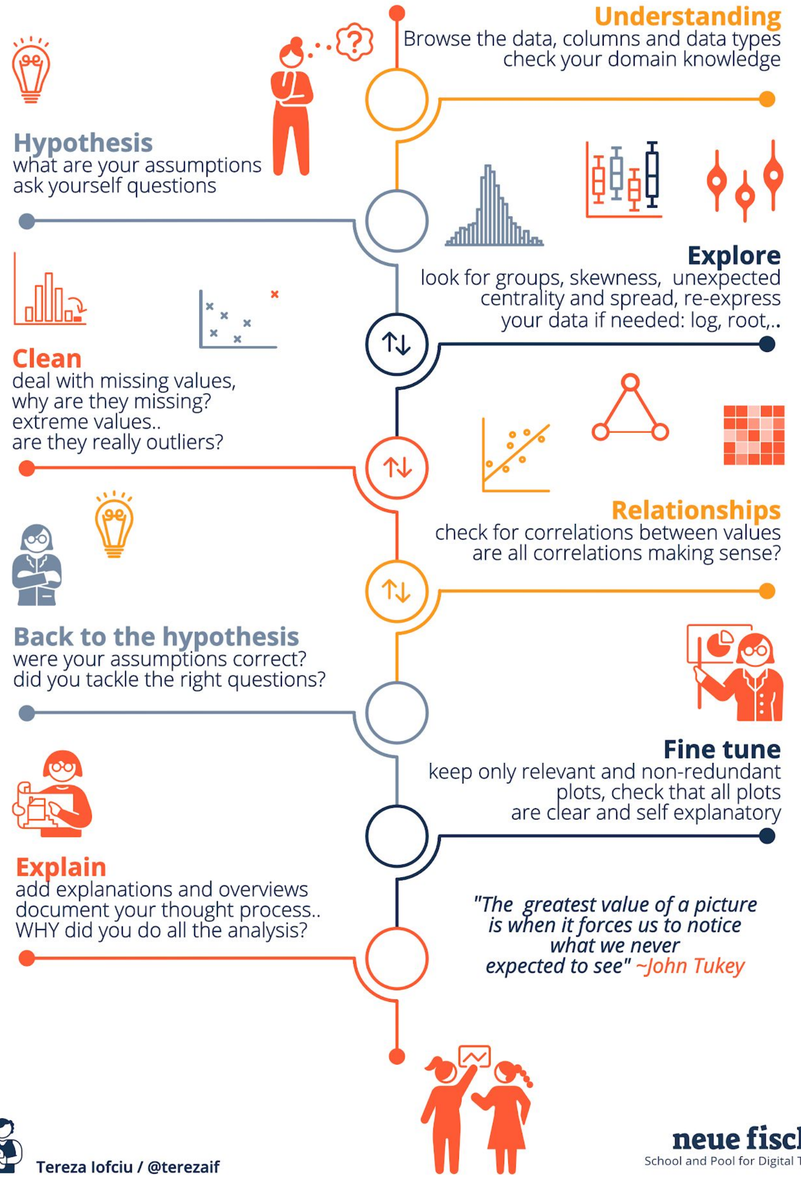
EDA & Presenting your Results



THE EDA PROCESS - recap



EXPLORATORY DATA ANALYSIS
= check list =



Tereza Iofciu / @terezaif

neue fische
School and Pool for Digital Talent

https://github.com/neuefische/datascience-infographics/blob/main/EDA_Checklist.md

EDA

Jupyter notebook

- overview and goals at the beginning
- description of data
- general stats about the data
- hypothesis about the data (n = 3)
- data cleaning
- analysis
- findings

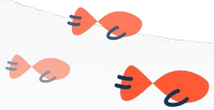
EDA Proof of Concept - Workflow

Timeboxed work! so use an iterative process

- make a draft
- do simple plots
- answer main questions (from hypothesis generation)
- iterate: go deeper, go prettier, go better
- clean up and document

Start early with the slides!

Dos and Don’ts of EDA



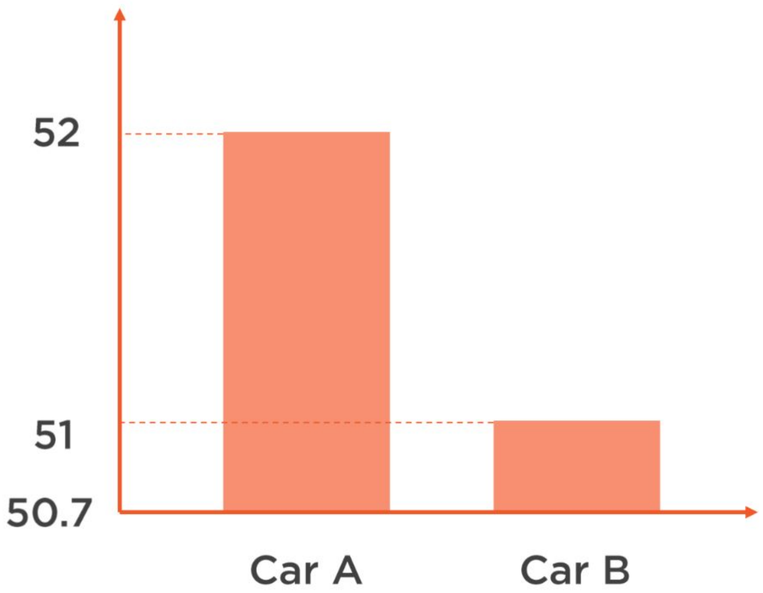
Dos

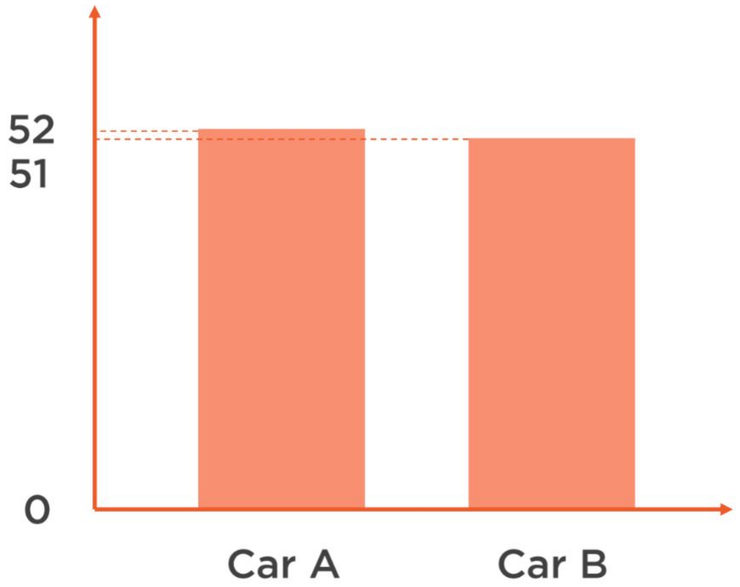
- **Be truthful, based on real data -** you might be lying without knowing
- **Be accurate and avoid ambiguity**
- **Easy interpretation for your audience -** don’t make them work at trying to decipher a chart or computation
- **Elegant and aesthetically pleasing for better understanding -** it’s not about doing pretty charts, but about better understanding

Common mistakes

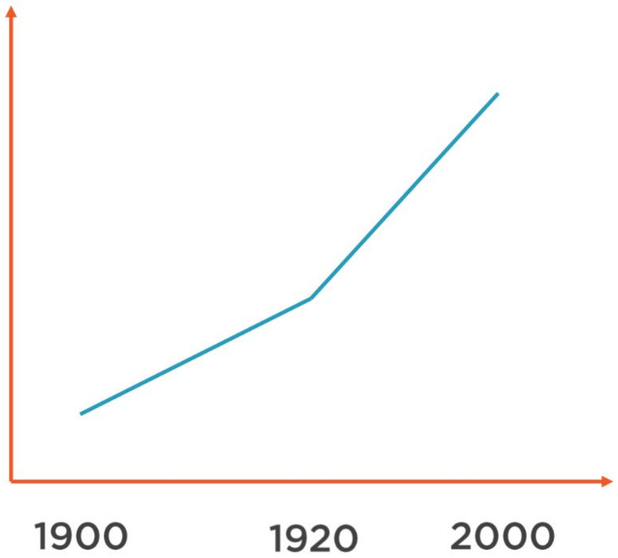
- **Not cleaning the data**
- **Cherry picking -** focusing on a metric that proves your assumptions
- **Focusing on (or ignoring) outliers -** outliers should be considered as a factor and not as an indicator
- **Chart junk -** less is more
- **Missing or ignoring data patterns -** seasonality, holiday, weekends ...
- **Lacking action -** recommendations, conclusion, hypothesis

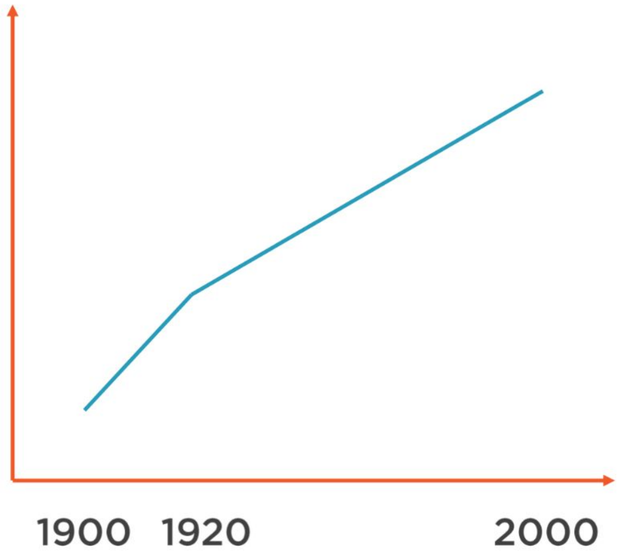
How not to lie… don’t truncate



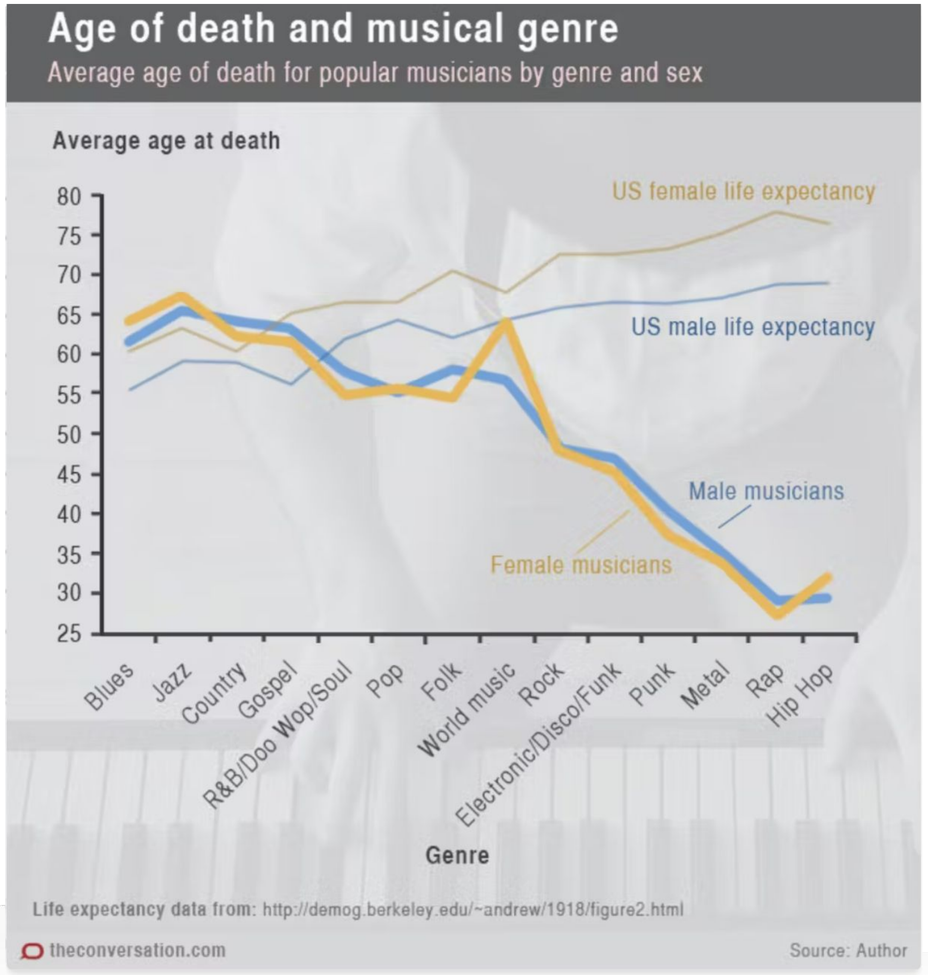


How not to lie… use an appropriate scale

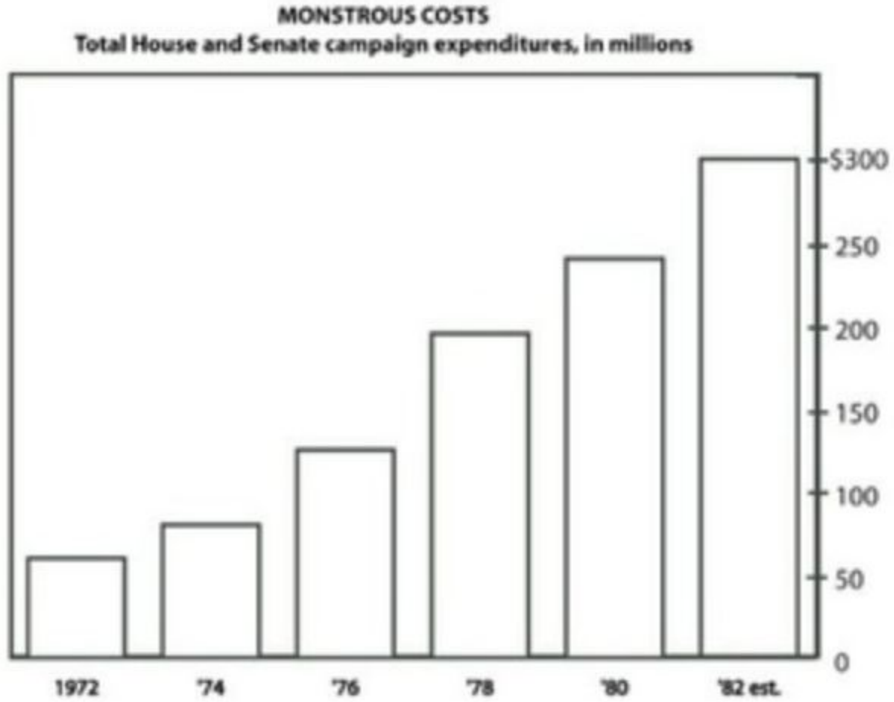




How not to lie… apply domain knowledge

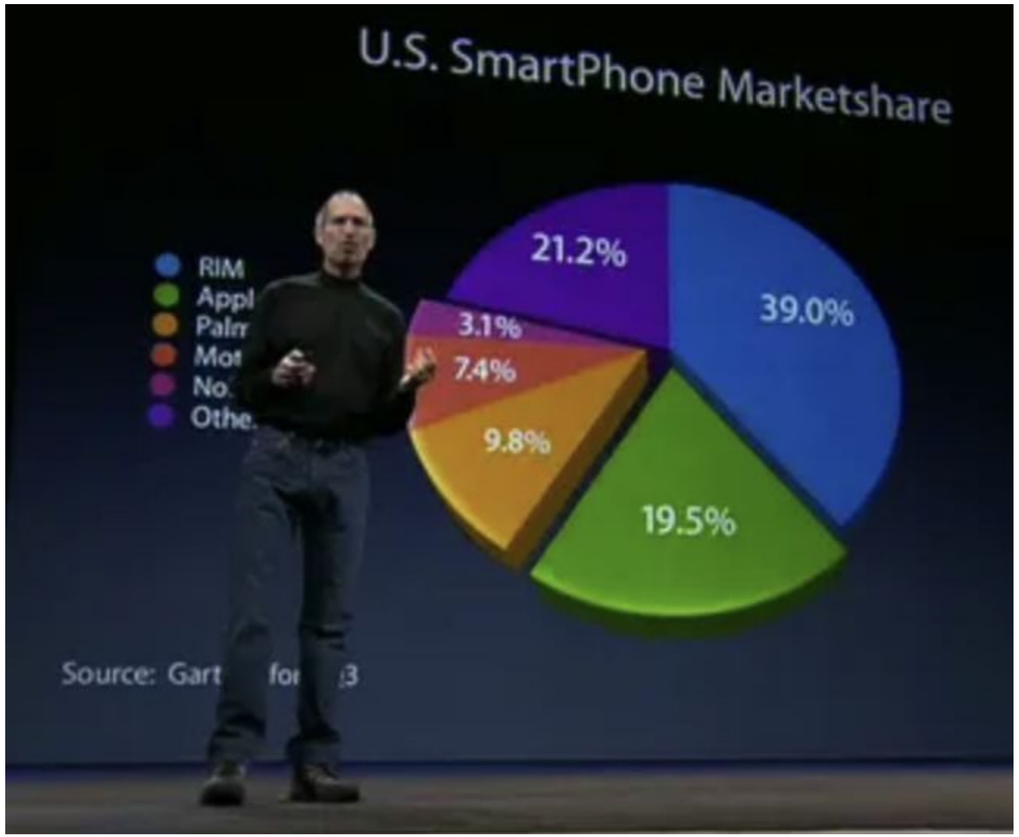


Avoid chart junk

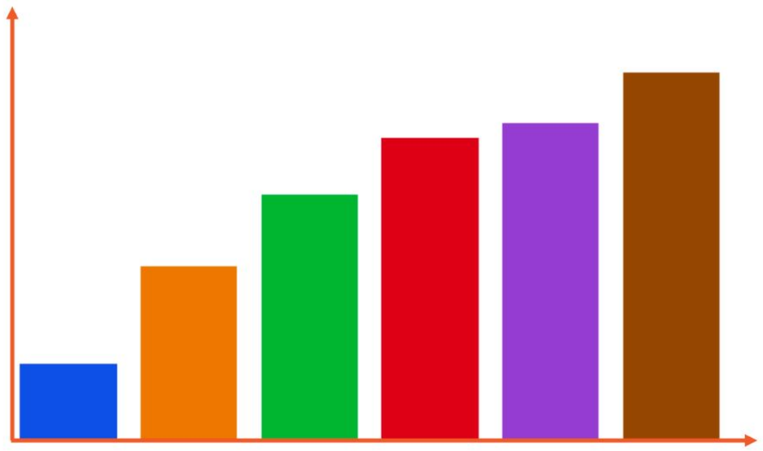


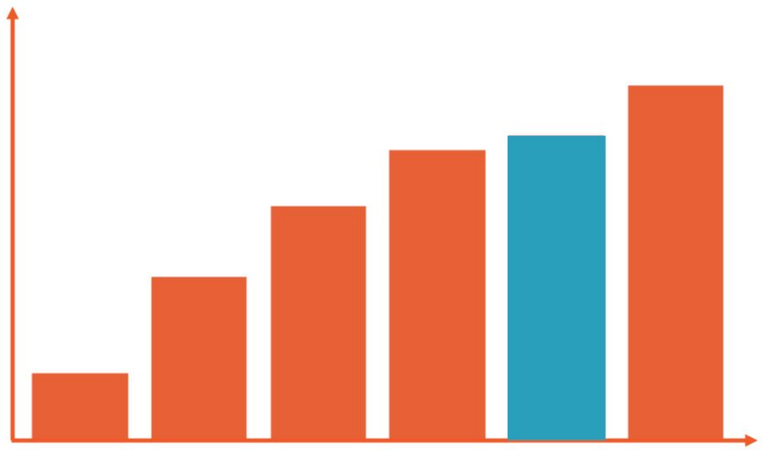
https://nigelholmes.com/

No pie charts!



If you use colors be intentional

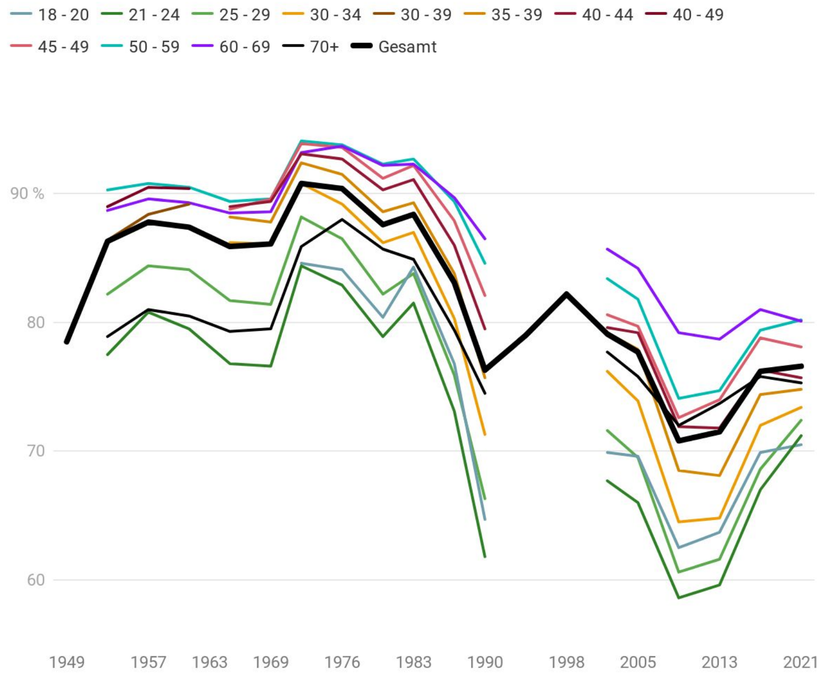




Choose a distinguishable color scheme

Wahlbeteiligung nach Altersgruppen

Bundestagswahlen 1949 bis 2021

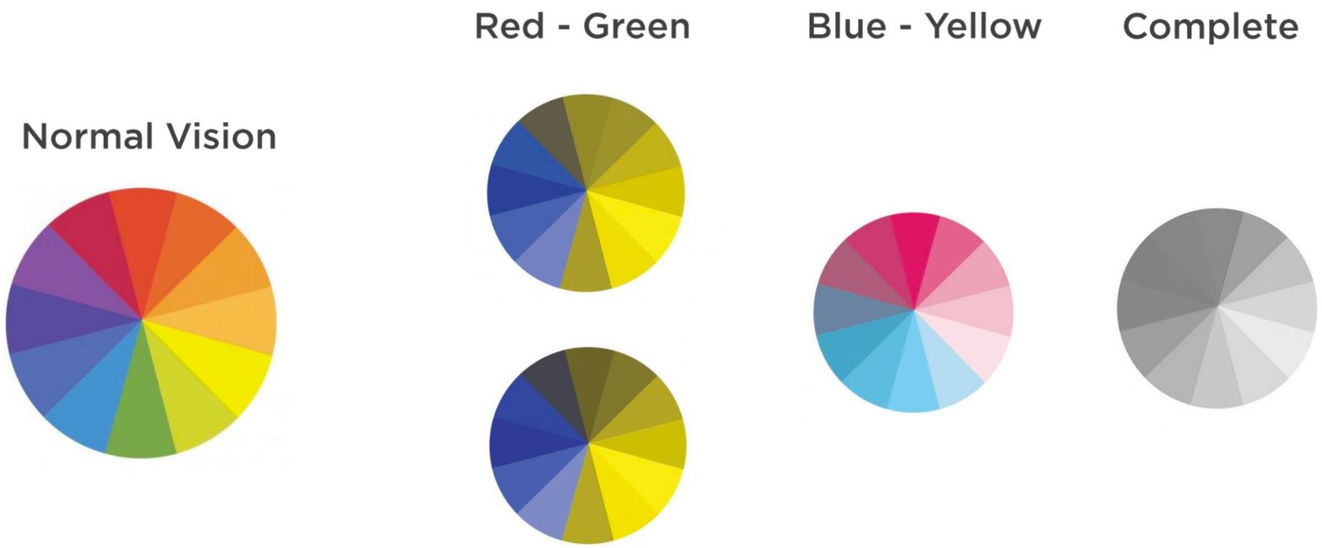


Die Angaben über die Wahlbeteiligung nach Altersgruppen stammen aus der repräsentativen Wahlstatistik. 1949, 1994 und 1998 wurde keine repräsentative Wahlstatistik durchgeführt. 1953 nur ohne die Beteiligung der Länder Rheinland-Pfalz, Bayern und Saarland.
Die Daten stehen unter der Datenlizenz Deutschland – Namensnennung – Version 2.0.

Grafik: bpb • Quelle: Der Bundeswahlleiter

Color blindness - avoid mixing green and red at least

color blindness friendly color palette: https://davidmathlogic.com/colorblind/

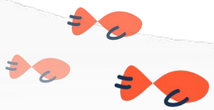


Presenting your results

Presentation of EDA (10 min)

Use Slides!

- setting the scene
  - **intro to the dataset**
  - **intro about client**
  - **intro about the quality of the data**
- focus on hypothesis: whys
- methodology: hows
- describing the findings and changes in approach... in context
- generated knowledge: insights
- future work
- if possible: show impact and applications



how to make your ds presentation great

Don’t lose your audience…

- Keep it simple and stupid (KISS principle)
- Be concise and accurate, no extra information if not needed
- Join explanation blocks with “whys” and “hows”
- Be clear, don’t be ambiguous - clarity inspires trust in your results