Data Visualisation

European Union
European Structural
and Investment Funds

Investing
in your future

Diana Pilvar
Data Manager
University of Tartu
ELIXIR Estonia
diana.pilvar@ut.ee

Learning Outcomes

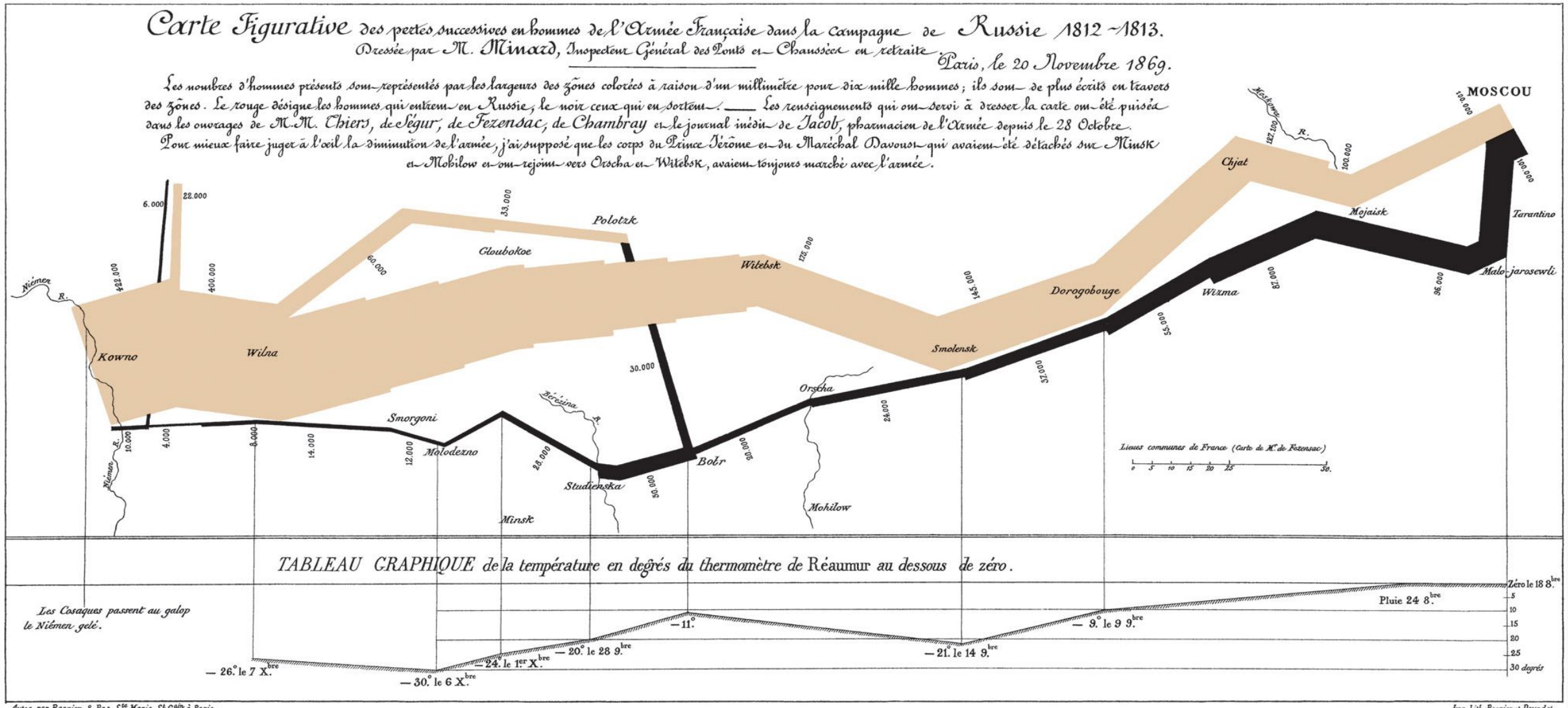
- Choose appropriate chart types
- Avoid common mistakes in data visualisation
- Maximise data-ink ratio

Overview

- Data type determines chart type (Theory → Example → Practice)
- Avoiding common mistakes (Theory → Examples)
- Figure design best practices (Theory → Examples)
- Improving the visual design of the figure

Introduction

Napoleon’s invasion of Russia mapped



https://datavizblog.files.wordpress.com/2013/05/map-full-size1.png
https://datavizblog.com/2013/05/26/dataviz-history-charles-minards-flow-map-of-napoleons-russian-campaign-of-1812-part-5/

Introduction to data visualisation

Data visualisation is an interdisciplinary field that deals with the graphic representation of data and information.

- Science - accurately convey the data
- Art - aesthetically pleasing

Must not mislead or distort

Data visualisation

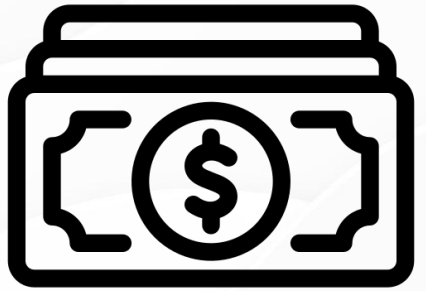
+ Makes the data easy to digest
+ Explore opportunities interactively (identify trends, patterns, outliers)
+ Image is easier to remember

- Biased or inaccurate information
- Correlations does not always mean causation
- Core messages can get lost in translation

Data type

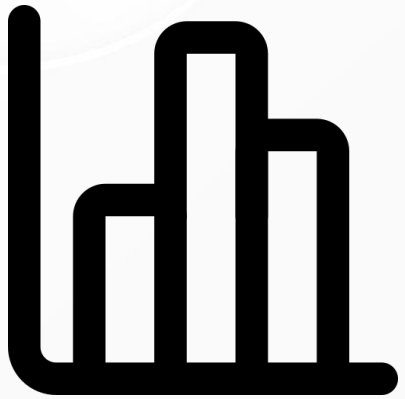
Variable type: Amount

- **Total number of a particular item or subject**
  - Quantitative/ numerical, discrete
- Examples:
  - Amount of money
  - Website traffic received each hour of the day
  - Sales for each product category
- Common traits
  - Values are non-negative
  - Absolute values are important, not relative changes

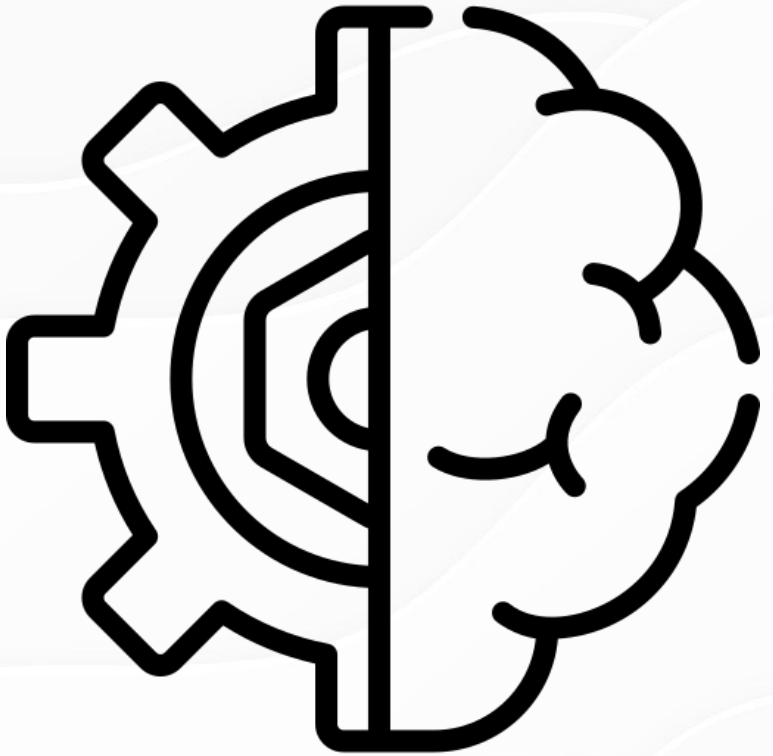


Variable type: distribution

- Represents the possible values of a variable and how often they occur
  - Quantitative/ numerical, continuous
- Examples:
  - Ages of students in a school
  - Income
  - Grades in a class



Practice



Which of the following is an amount?

1. Number of employees in UT
2. Temperature in degrees Celsius
3. Daily temperature changes in degrees Celsius
4. Duration of a movie in minutes
5. Scores in a standardized test

Which of the following is a distribution?

6. Pencil sales in September 2023
7. Popular shoe sizes in a shop
8. Traffic flow throughout the day
9. Distance traveled in kilometers
10. SARS-CoV-2 viral strains in Estonia over time

Correct answers

Which of the following is an amount?

- ✅ Number of employees in UT
- ✅ Temperature in degrees Celsius
- ❌ Daily temperature changes in degrees Celsius
- ✅ Duration of a movie in minutes
- ❌ Scores in a standardized test

Which of the following is a distribution?

- ❌ Pencil sales in September 2023
- ✅ Popular shoe sizes in a shop
- ✅ Traffic flow throughout the day
- ❌ Distance traveled in kilometers
- ✅ SARS-CoV-2 viral strains in Estonia over time

Variable type: proportion

- A part or share of the whole data set,
  often expressed as a percentage or fraction
- Examples:
  - ⅓ of market share
  - Gender ratio
  - Pass rate - how many students
    passed/failed/never showed up

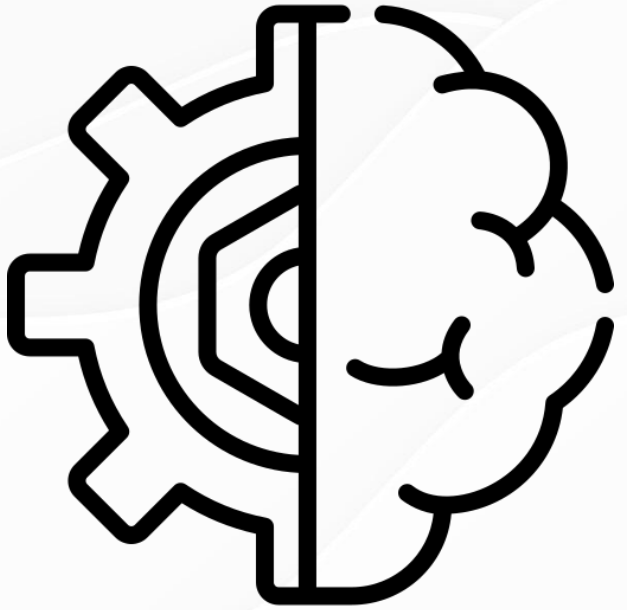
Variable type: associations

- The relationship or correlation between two or more variables in a dataset
- Examples:
  - Body mass and energy demands
  - Education and income
  - Smoking and lung cancer

Variable type: time series

- Time series: set of data points collected or recorded in a chronological order over a certain period of time
- Examples:
  - Stock prices
  - Weather data
  - Inflation rate

Practice



Which of the following is a proportion?

1. The probability of rain or snow given as a percentage
2. Days in a month when it was raining
3. UT workers who have been given flu shots in 2023
4. Ice creams sold during festival
5. Customers who are satisfied with a service

Which of the following is a time series?

6. Smartwatch measurements during exercise
7. Daily hospital admission numbers
8. Emails in a mailbox that are marked as spam
9. Hourly electricity consumption
10. Employment rate in a country

Correct answers

Which of the following is a proportion?

- ❌ The probability of rain or snow given as a percentage
- ✅ Days in a month when it was raining
- ✅ UT workers who have been given flu shots in 2023
- ❌ Ice creams sold during festival
- ✅ Customers who are satisfied with a service

Which of the following is a time series?

- ✅ Smartwatch measurements during exercise
- ✅ Daily hospital admission numbers
- ❌ Emails in a mailbox that are marked as spam
- ✅ Hourly electricity consumption
- ❌ Employment rate in a country

Variable type: geospatial

- Data that is associated with a specific geographical location or physical space.
- Examples:
  - Elevation
  - Postal codes
  - Agricultural land use

Poll

What type of data do you need to visualise recently/ the most?

- Amount
- Distribution
- Proportion
- Association
- Time series
- Geospatial

Telling a story

A picture is worth a thousand words

A **story** is a set of observations, facts, or events, true or invented, that are presented in a specific order such that they create an emotional reaction in the audience.

- Important role in our reasoning and memory
- The audience will a get a story anyways, whether we influence it or not
- Multiple visualisations
- Audience MUST UNDERSTAND the figures you are showing

Less is more.

examples from:
doi: 10.1371/journal.pone.0131606 https://pubmed.ncbi.nlm.nih.gov/26135579/
https://elixir-europe.org/sites/default/files/documents/annual-report-2022.pdf page 53
https://flowingdata.com/2023/10/24/data-on-net-worth-income-and-savings/

Know your target publication

Research article figures:

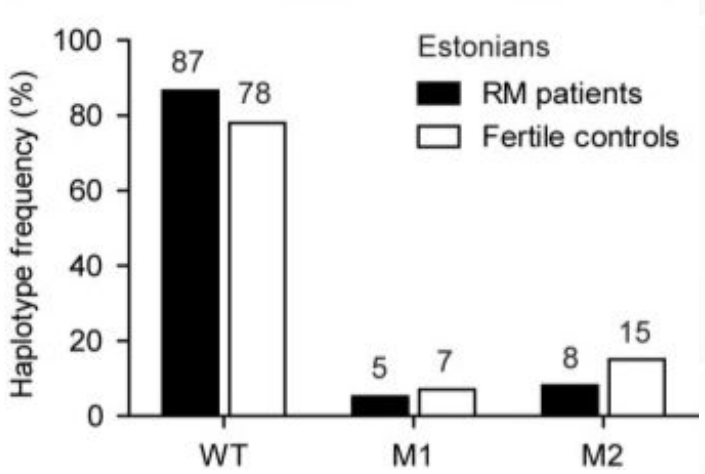
- Classical shapes and symbols
- Mostly black-and-white

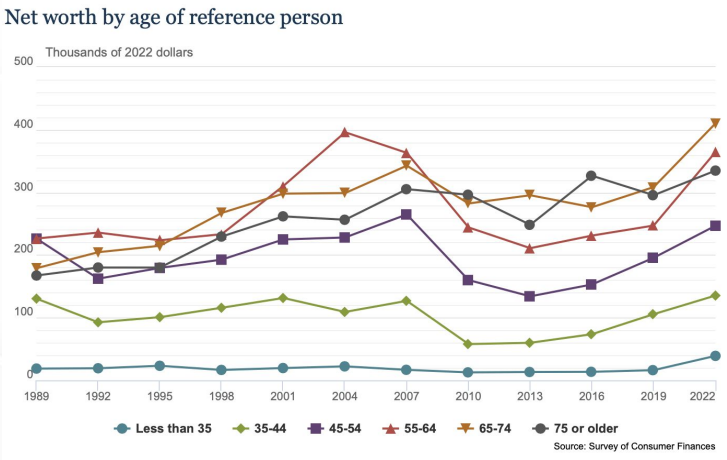
Report figures:

- Colorful
- Can play with style
- Classical shapes and symbols
- Gridlines

Presentation figures:

- Colorful
- Objects different than classical symbols
- White areas are good
- Highlight
- Can play more with visual aspect



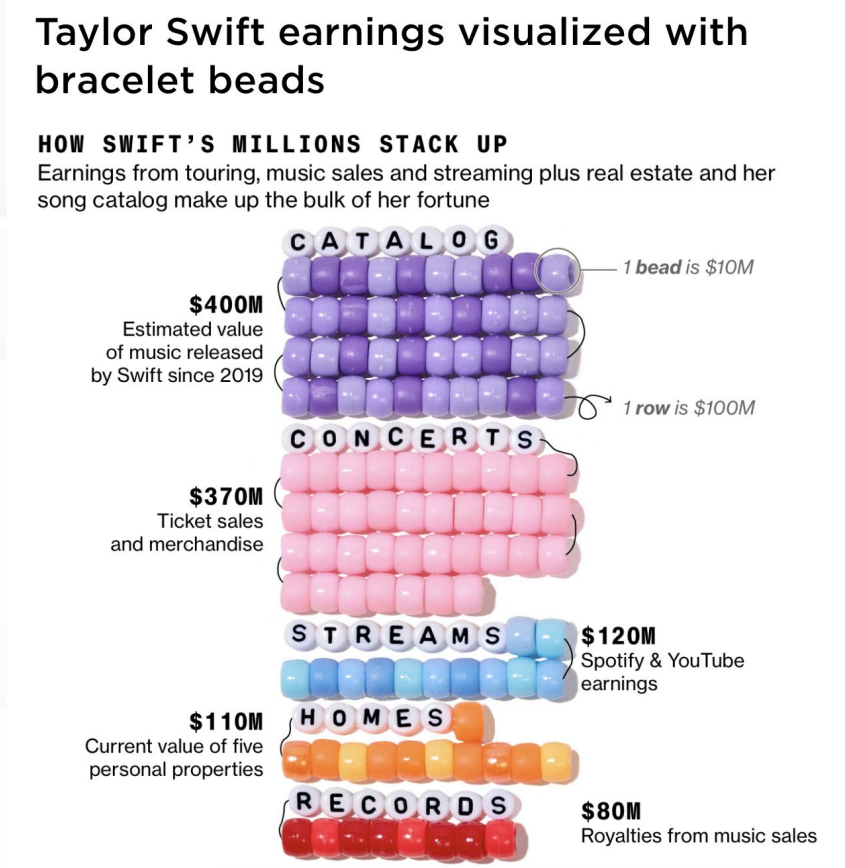


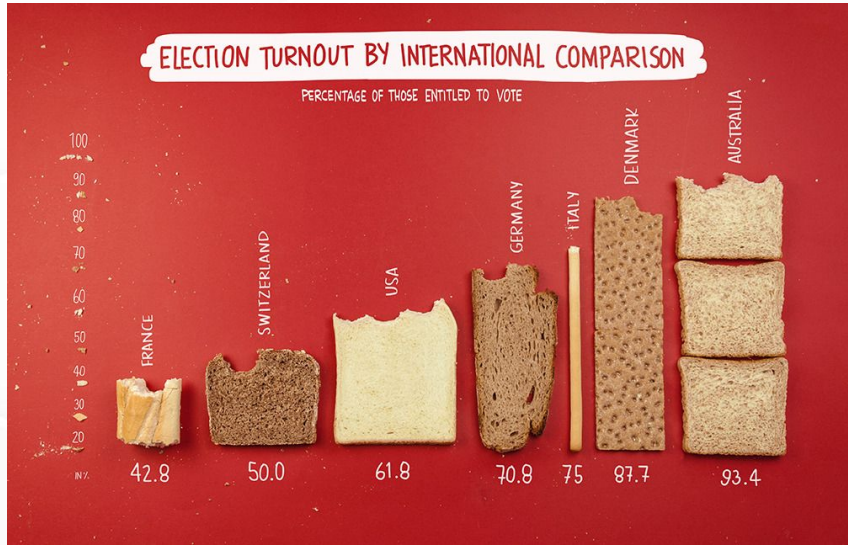
Know your target publication

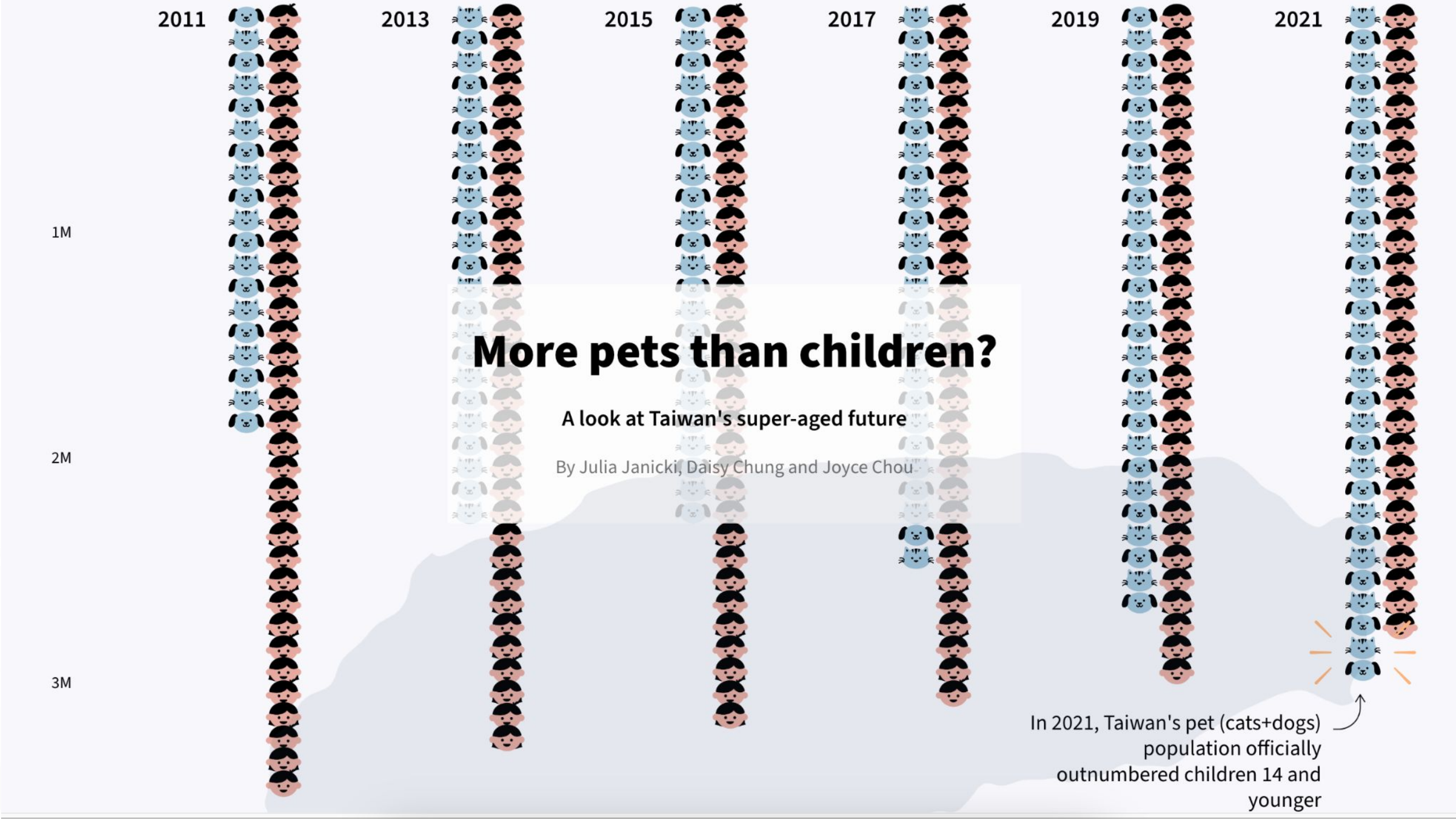
Press visualisations

- Most playful
- Attractive and creative
- Object differ from classical symbols
- Background pictures
- Highlighting parts
- Concentration of information

examples:
https://flowingdata.com/2023/10/30/taylor-swift-earnings-visualized-with-bracelet-beads/
https://bindersfullofburgers.tumblr.com/post/74961505700/amp






2011
2013
2015
2017
2019
2021
1M
2M
3M
More pets than children?
A look at Taiwan's super-aged future
By Julia Janicki, Daisy Chung and Joyce Chou
In 2021, Taiwan's pet (cats+dogs) population officially outnumbered children 14 and younger

Find out your target audience

- Academics vs General Public
- Age
- Beginners or Advanced
- Cultural differences
- Language
- Political background
- Historical background

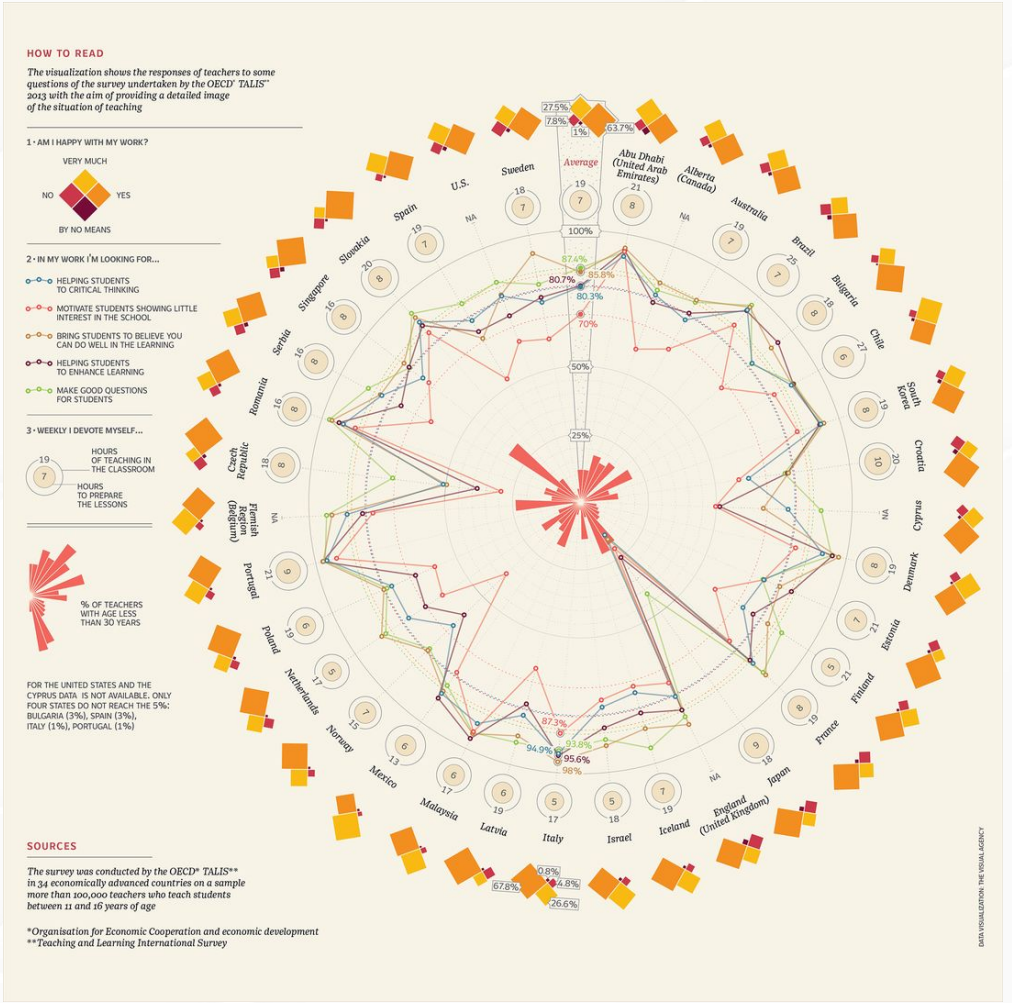


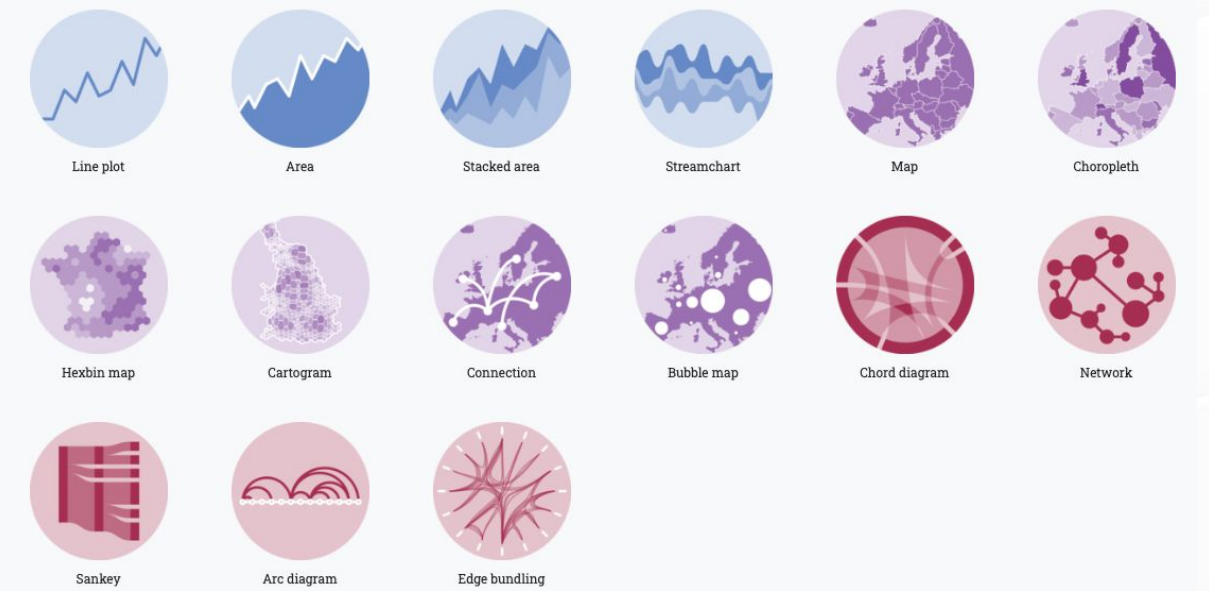
Figure out your focus

What is your main message?

Will help you choose the right graph.

Chart types

Chart types

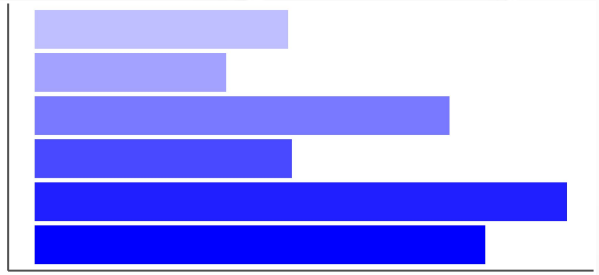


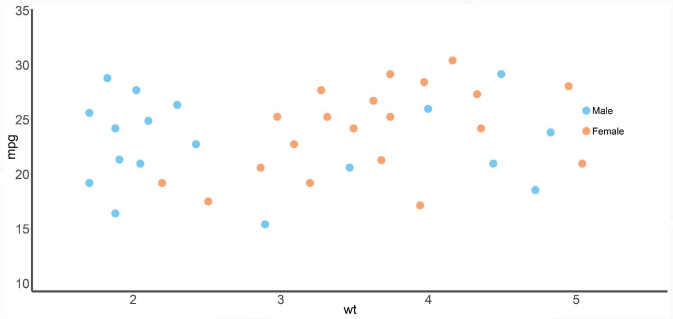
https://www.data-to-viz.com/

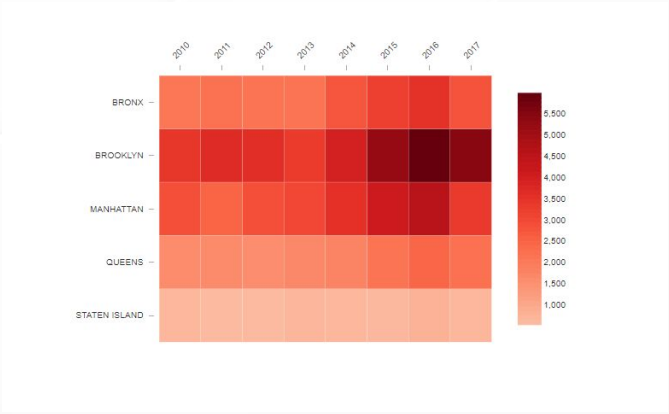
Data type: Amounts

Amounts - numerical values that correspond to specific categories

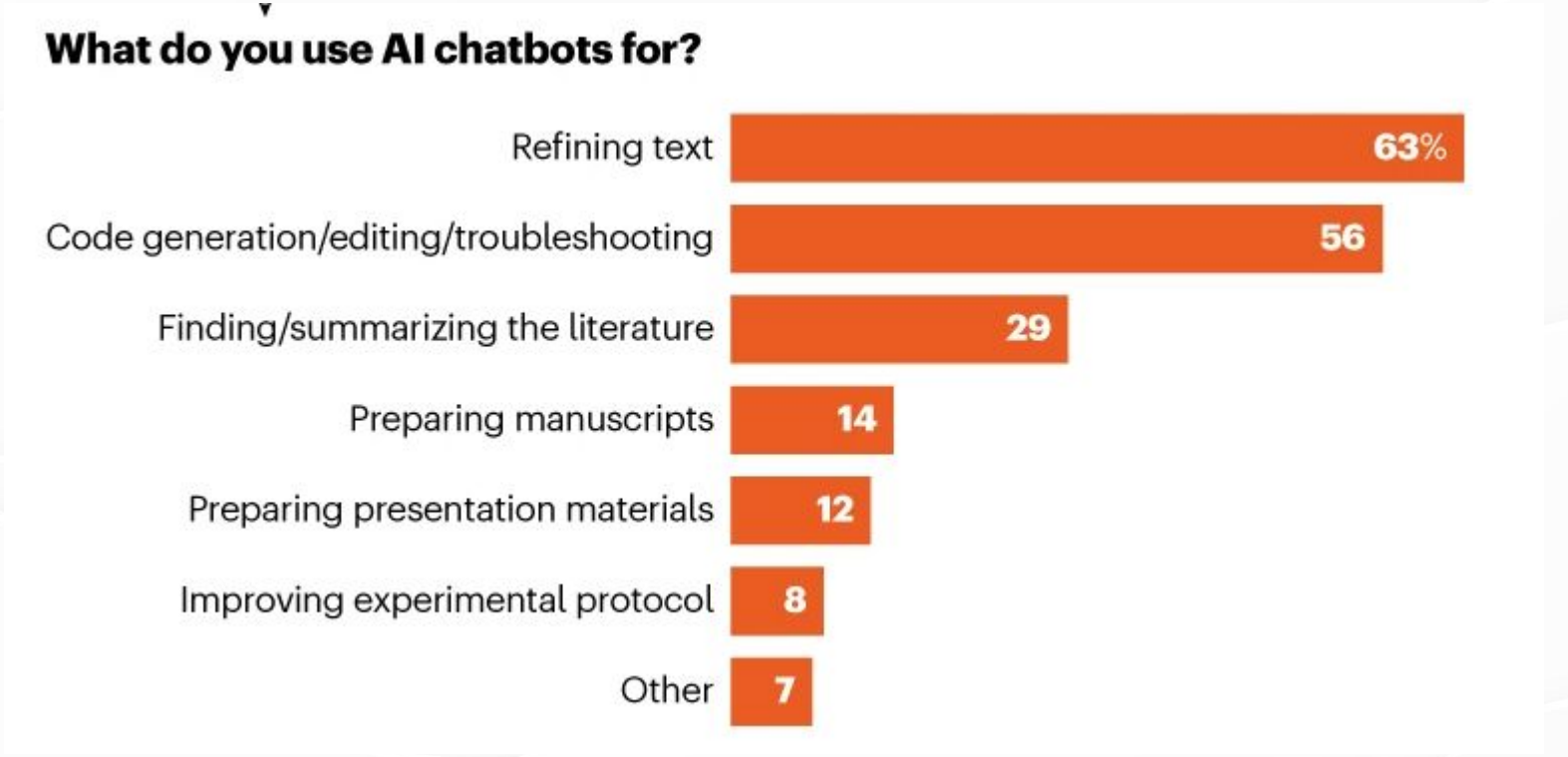
- Bar charts
- Dot plots
- Heatmap

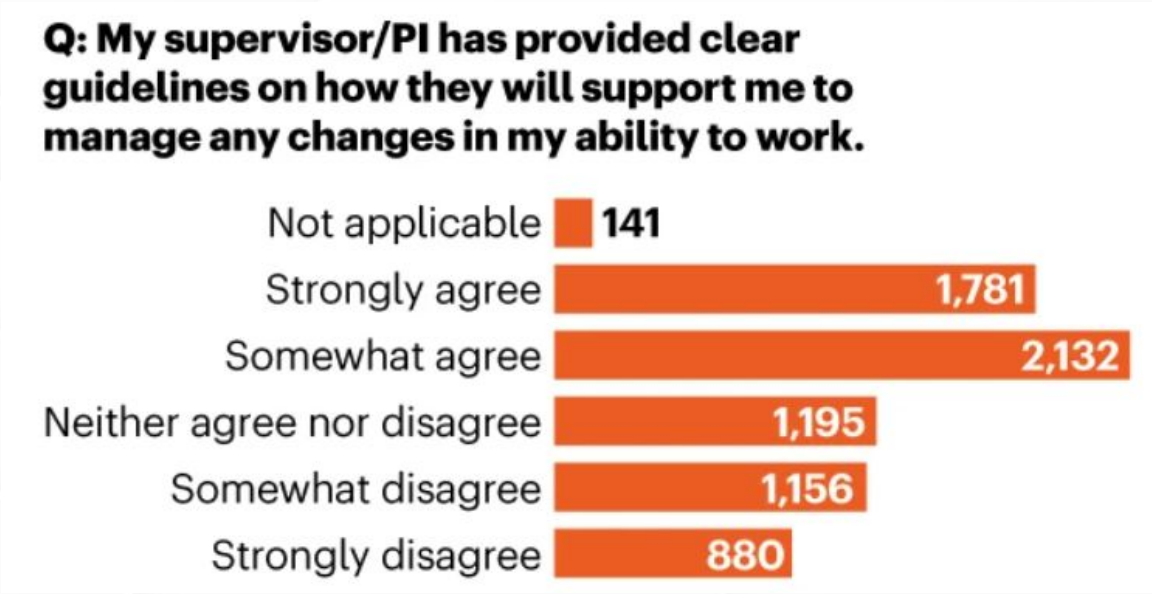






Example: Bar chart





https://www.nature.com/articles/d41586-023-03235-8
https://www.nature.com/articles/d41586-020-02548-2

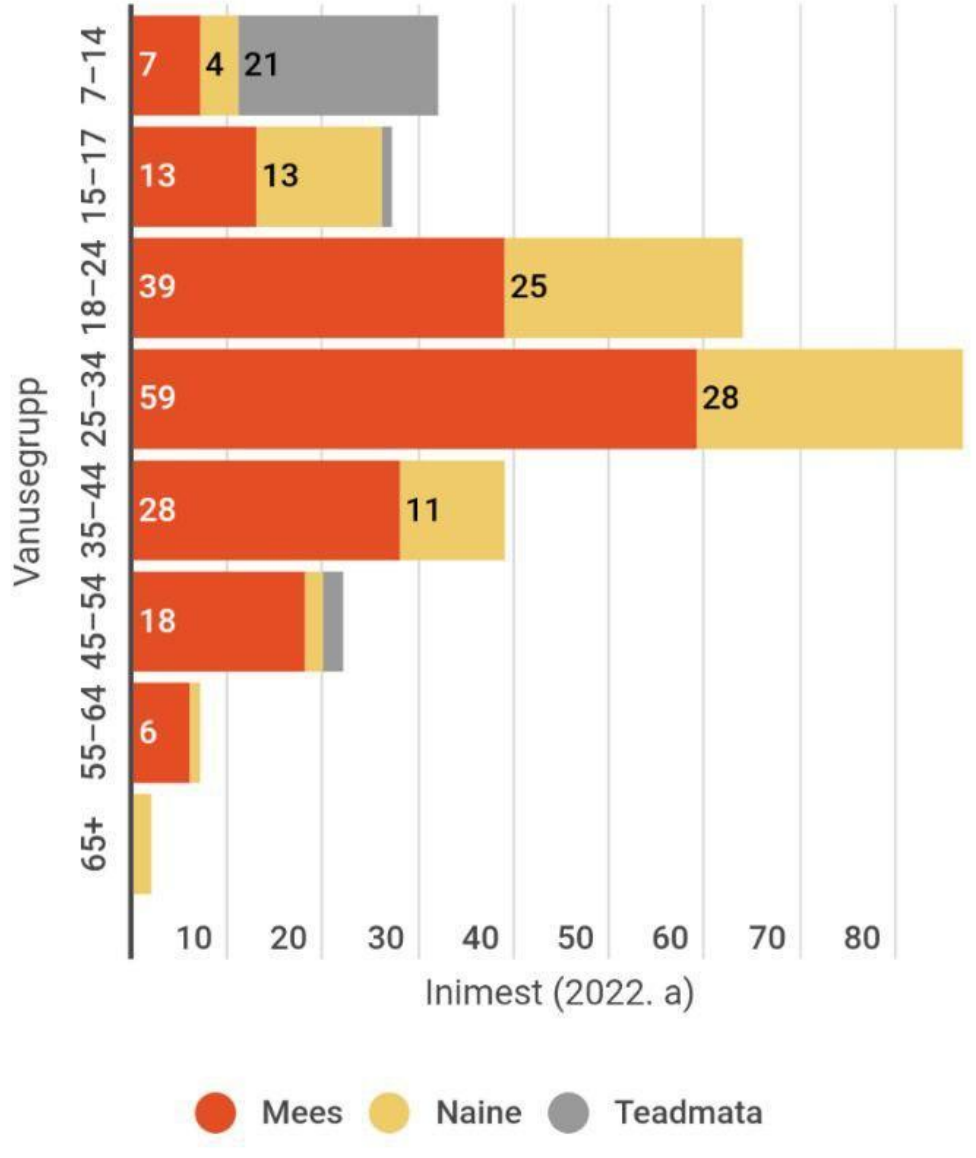
Example: Stacked bar chart

Age group comparison of people involved in traffic accidents with electric scooter rider in 2022

Men colored in red

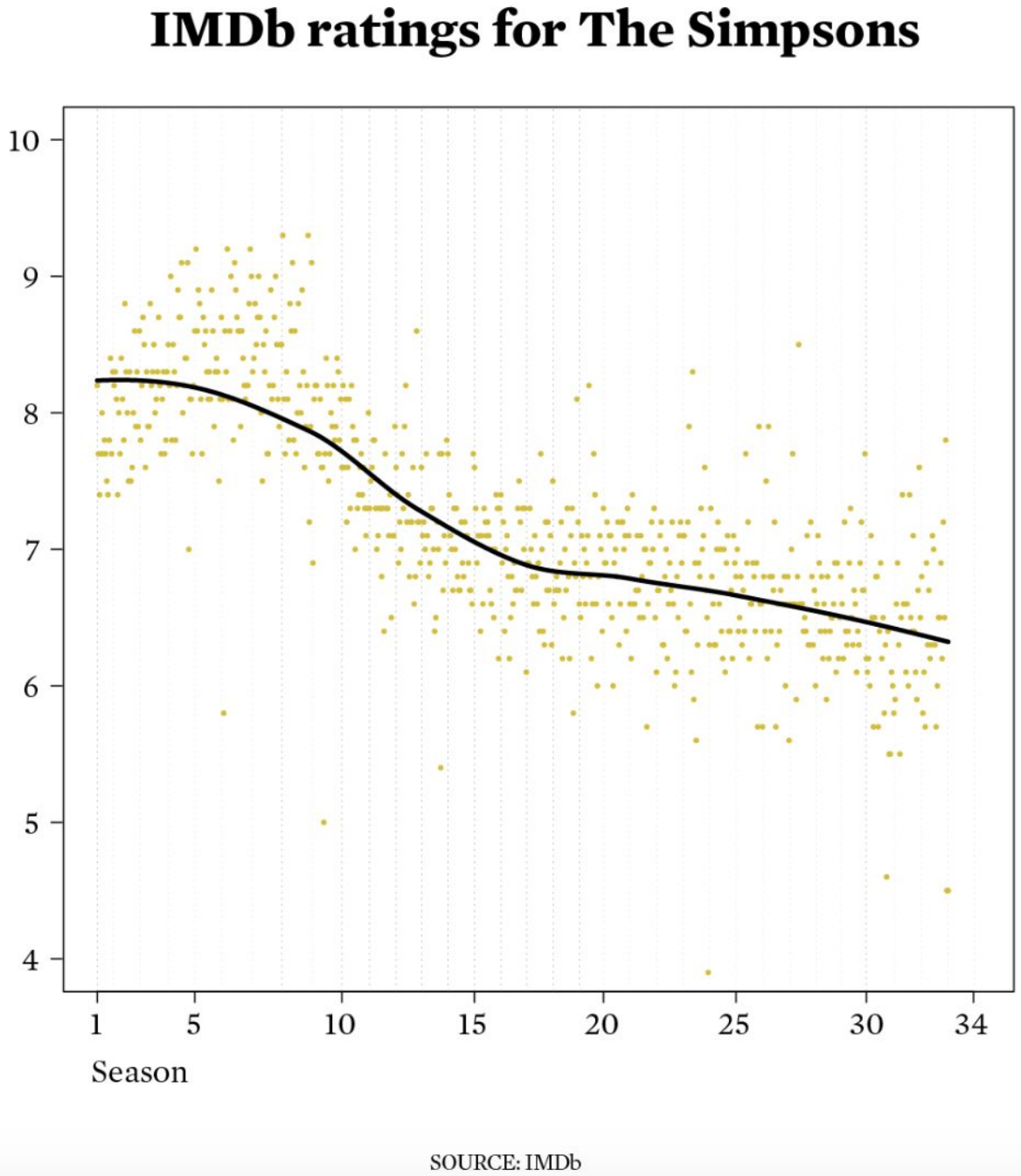
Females yellow

Grey - unknown



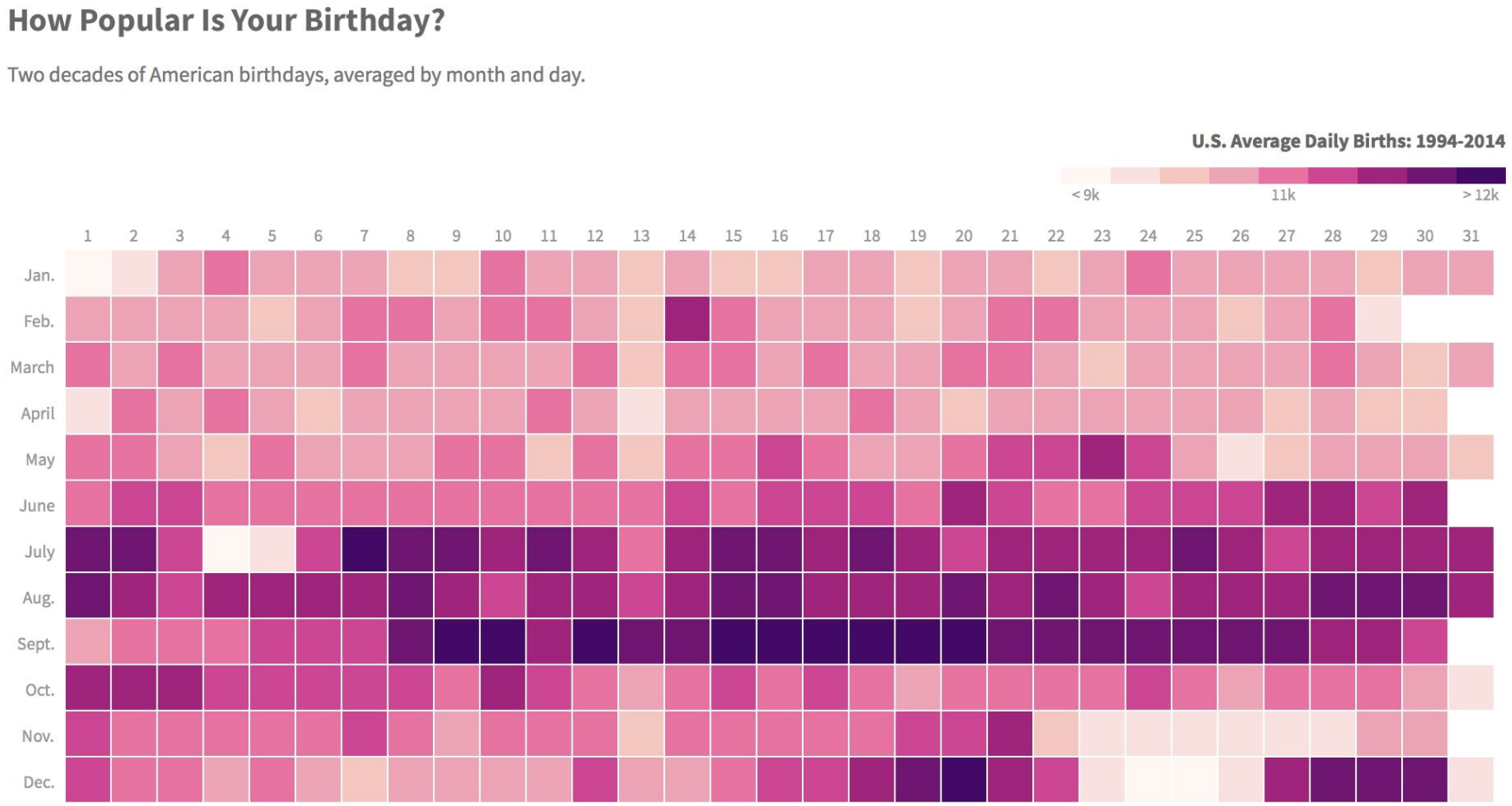
https://ekspress.delfi.ee/artikkel/120170908/metsik-laas-saab-lopu-riigid-tombavad-touksianarhiale-paitseid-pahe-ka-eesti-uljaspaid-ootavad-uued-karmimad-seadused
article (in estonian)
Data source https://public.tableau.com/app/profile/transpordiamet/viz/Kergliikuri_L/nnetusteldandmed

Example: Dot plot



https://flowingdata.com/2023/05/02/one-day-chart-challenge/

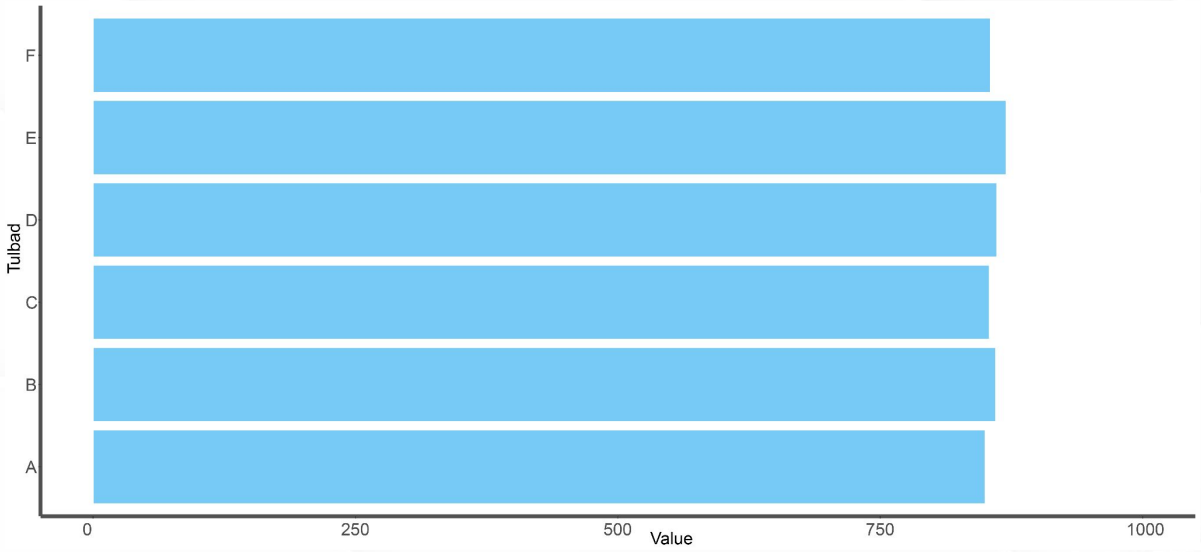
Example: Heatmap

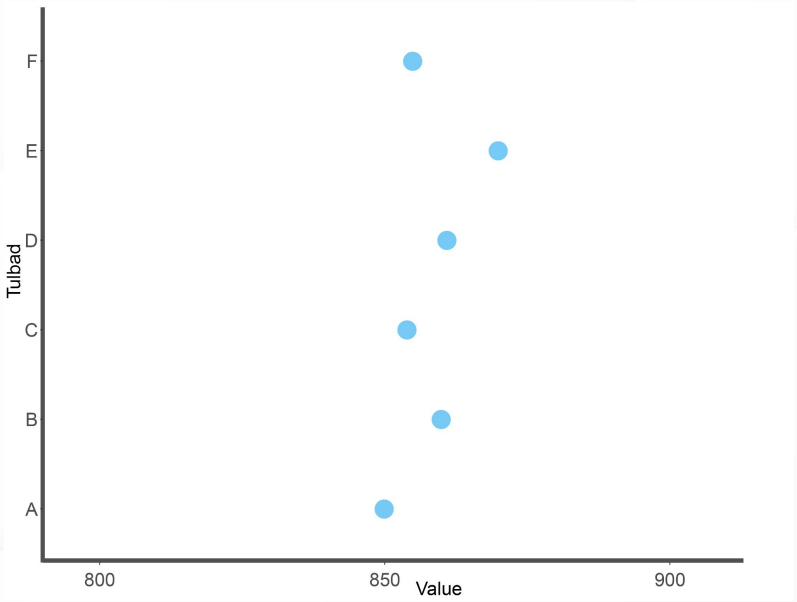


http://thedailyviz.com/2016/09/17/how-common-is-your-birthday-dailyviz/

Limitations

Bar charts: need to start at 0; Dot plots don't need to





For large datasets, prefer heatmap!

Order data (if possible) for ease of understanding!

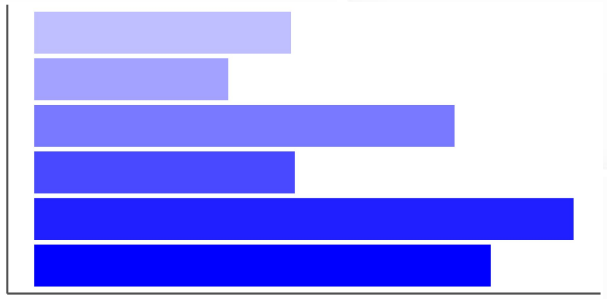
Chart types comparison for amounts

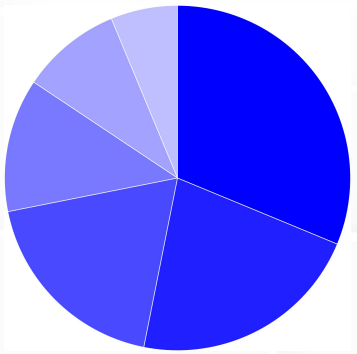
	Bar chart	Stacked bars	Grouped bars	Dot plots and heatmaps
Needs to start at 0	✅	✅	✅	❌
Quantitative amount varies with respect to one categorical variable	✅	❌	❌	❌
Quantitative amount varies with respect to 2+ categorical variables at the same time	❌	✅	✅	✅
Sum of the amounts represented is in itself a meaningful amount	❌	✅	❌	❌
Individual bars represent counts	✅	✅	✅	❌
Need to pay attention to the ordering of the data values	✅	✅	✅	✅

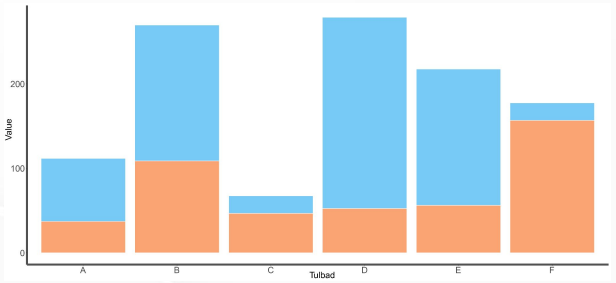
Proportions

To show how some group, entity, or amount breaks down into individual pieces that each represent a proportion of the whole

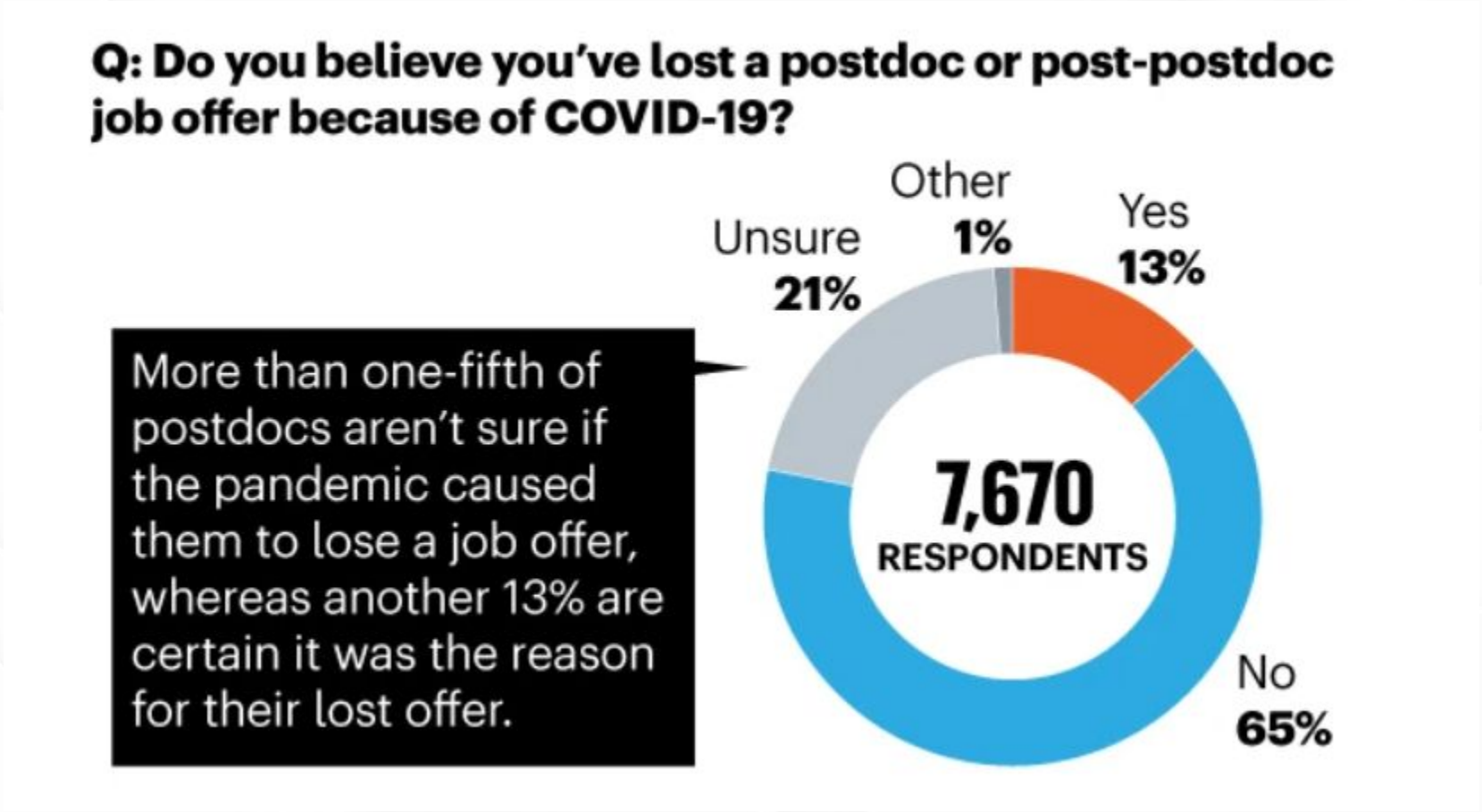
- Pie chart
- Stacked bar chart
- Bar chart





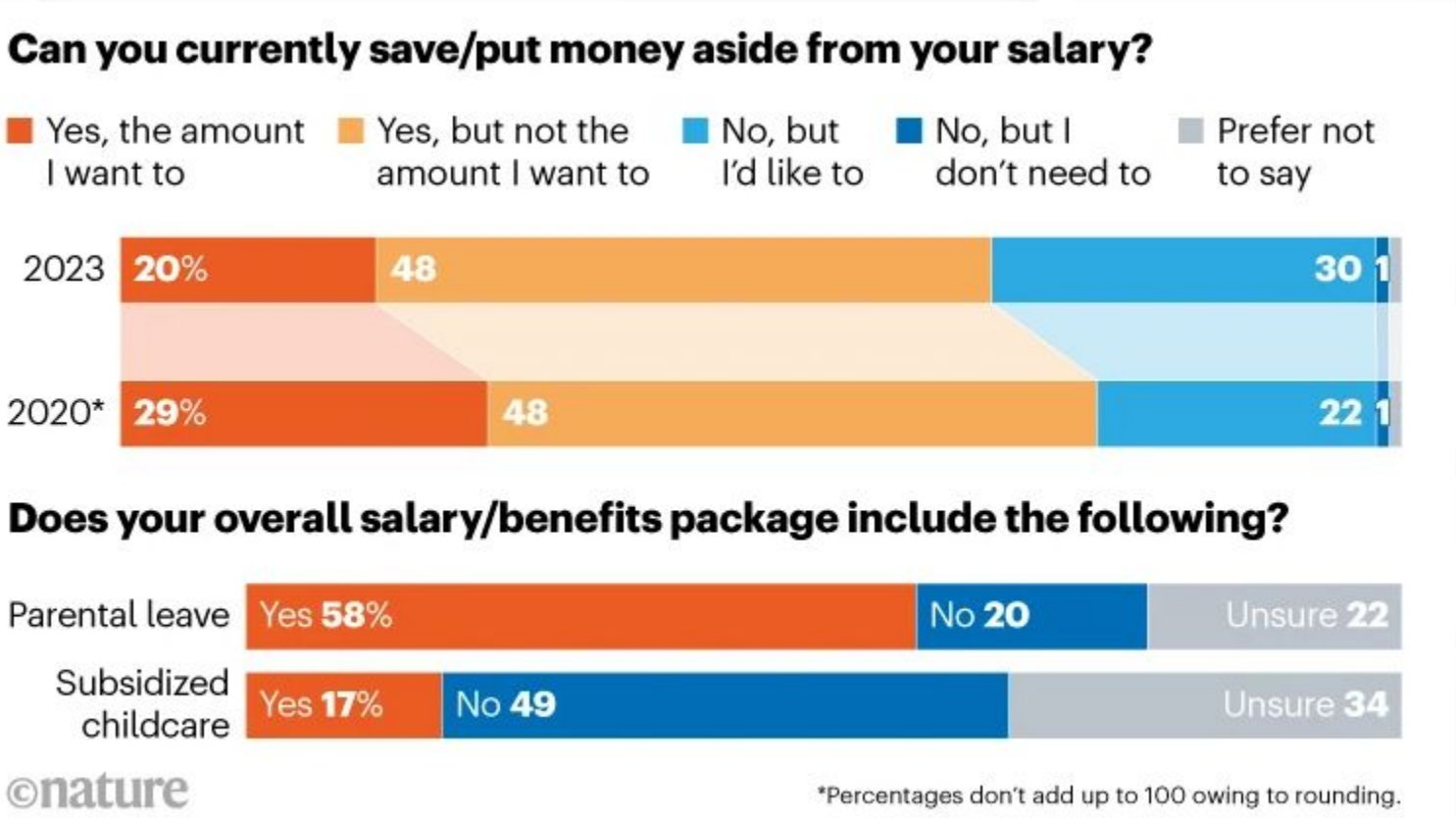


Example: Donut chart



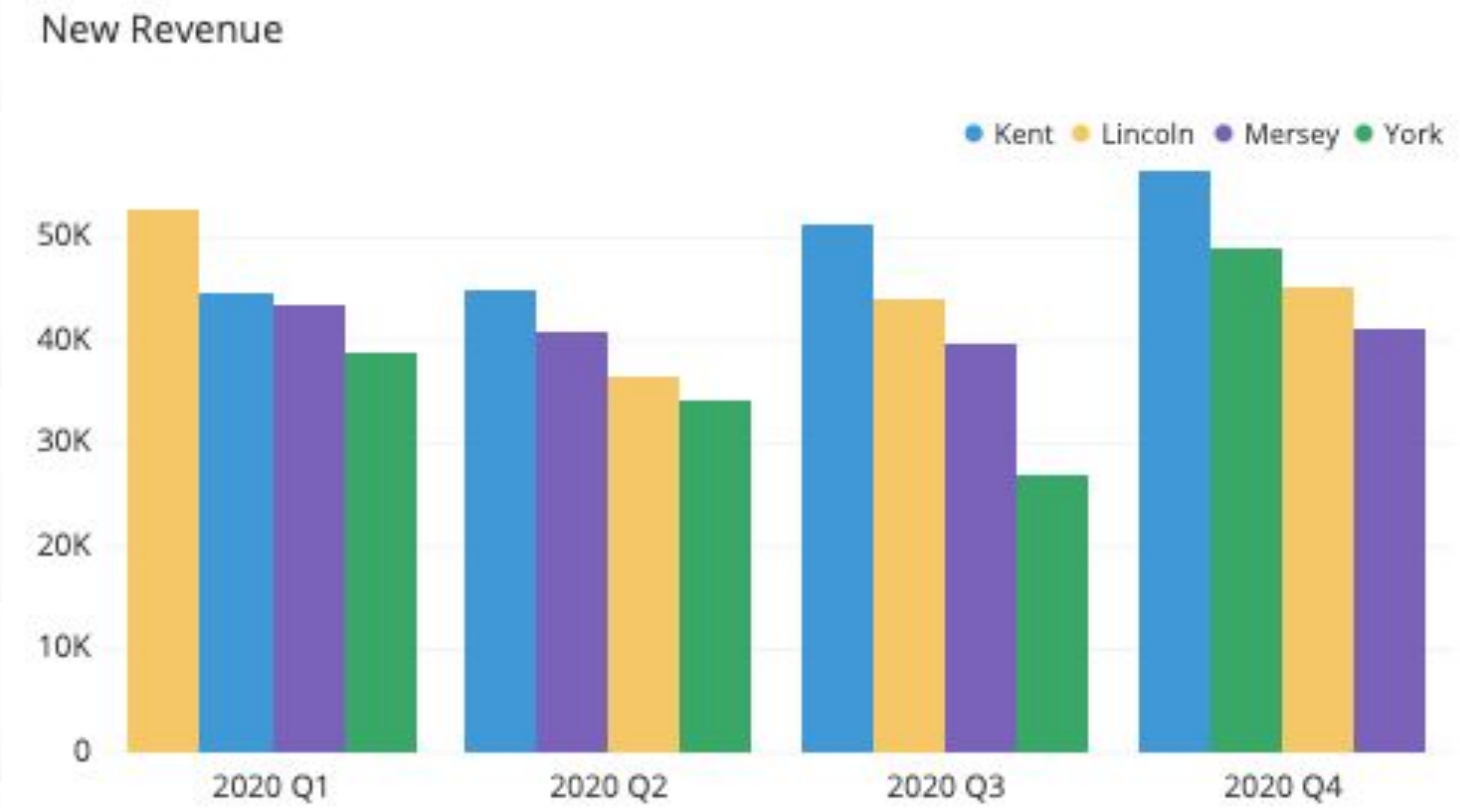
https://www.nature.com/articles/d41586-020-02548-2

Example: Stacked bars



https://www.nature.com/articles/d41586-023-03296-9

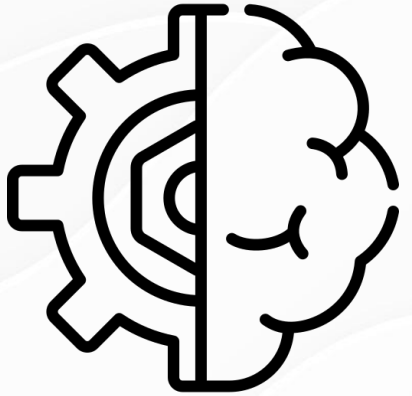
Example: grouped bars



https://chartio.com/learn/charts/grouped-bar-chart-complete-guide/

Chart types comparison for proportions

	Pie chart	Stacked bar chart	Side-by-side bar chart
Visualizes the data as proportions of a whole	✅	✅	❌
Comparison of the relative proportions	❌	❌	✅
Visually emphasizes simple fractions, such as 1/2, 1/4	✅	❌	❌
Suitable for small datasets	✅	❌	✅
Comparing parts of a bigger set of data	❌	❌	✅
Visualization of many sets of proportions or time series of proportions	❌	✅	❌



Practice

Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Number of different categories of books on the bookshelf

- Bar chart
- Stacked bar chart
- Pie chart
- Dot plot
- Heatmap

What would I do...

Prof Booknose has a massive bookshelf in his study.

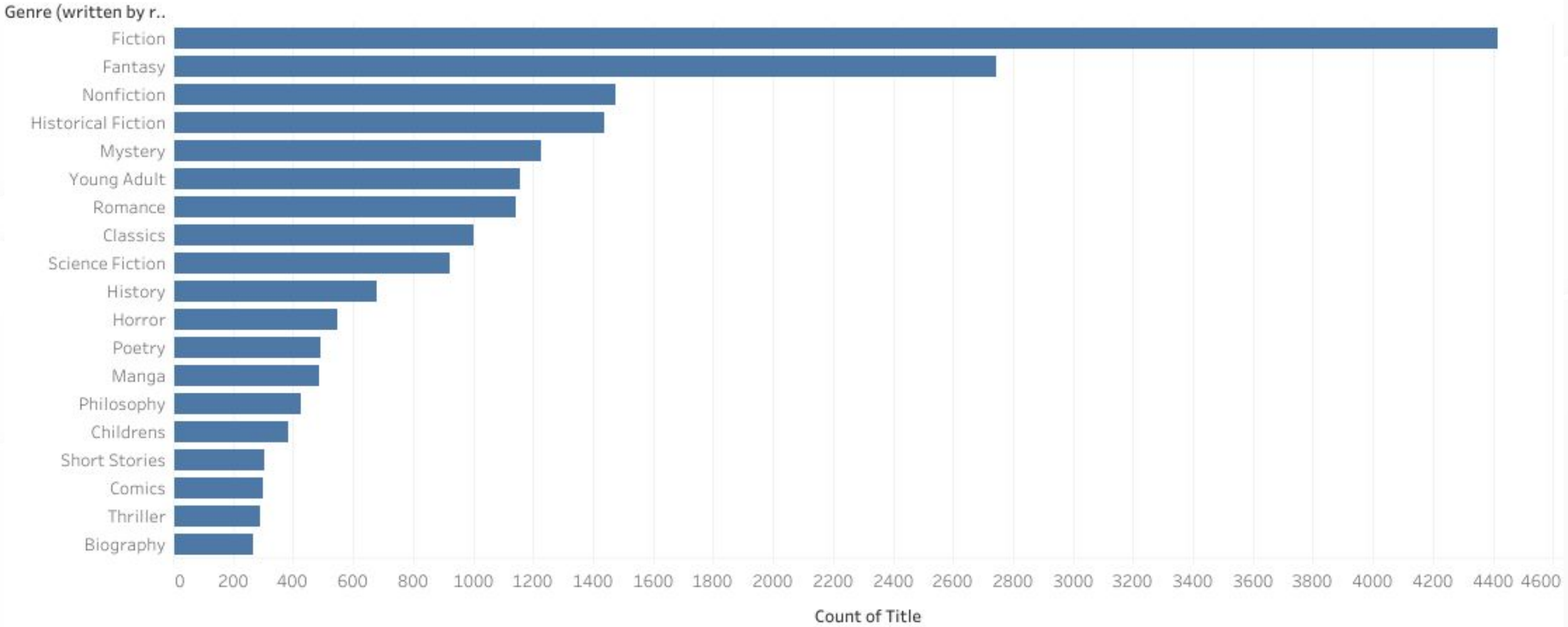
What chart would you choose to show:

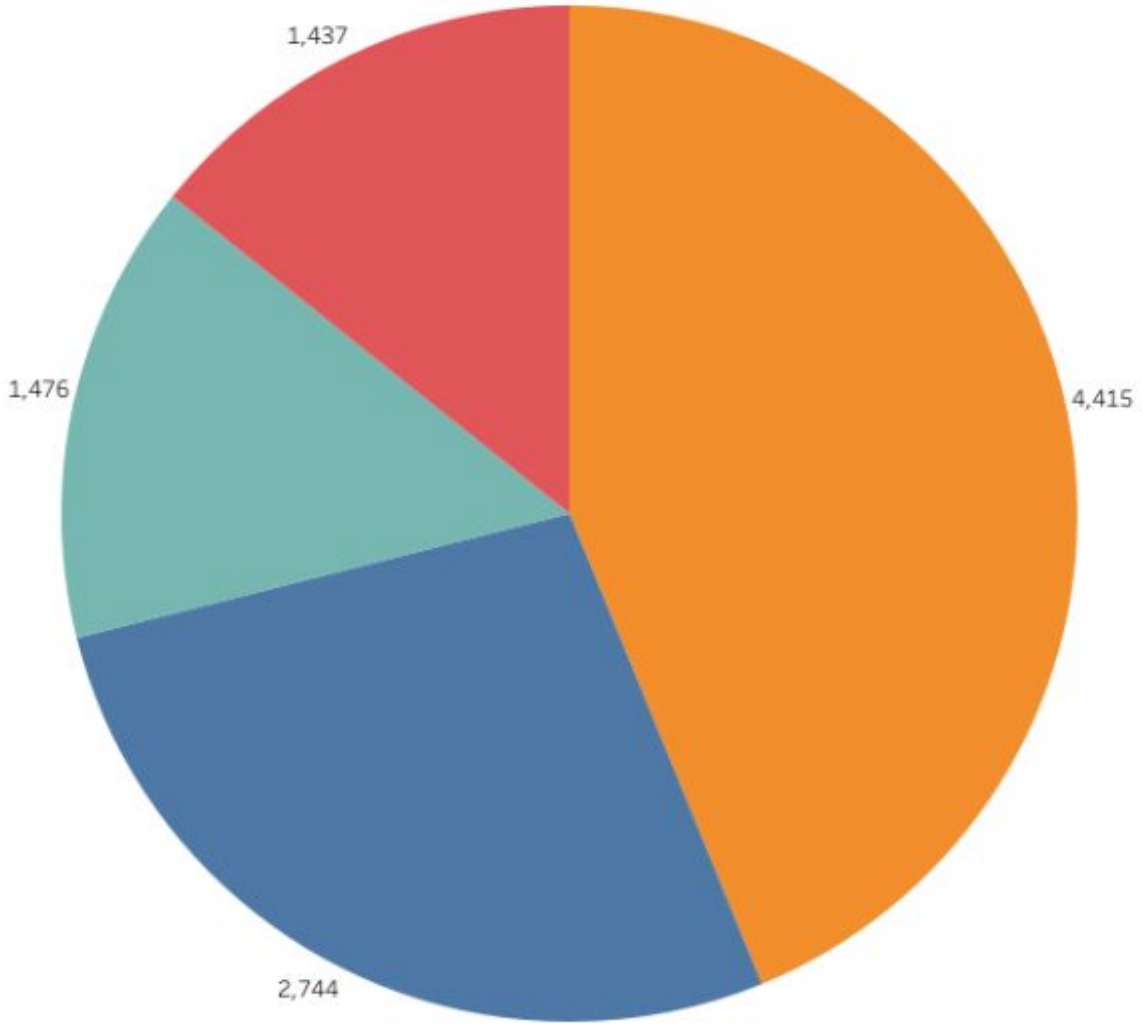
- Number of different categories of books on the bookshelf

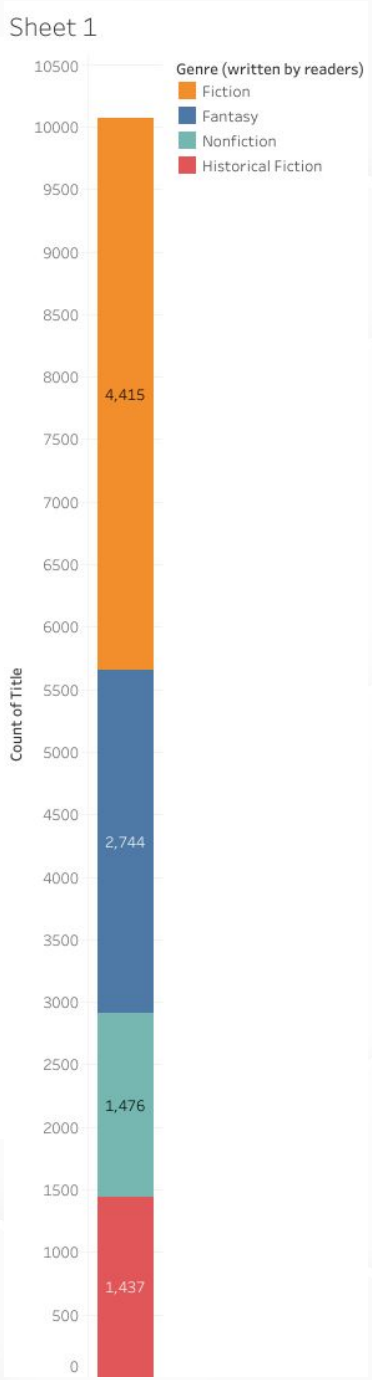
✅ bar chart

✅ pie chart if <=5 categories, stacked bar chart works too then

Sheet 1

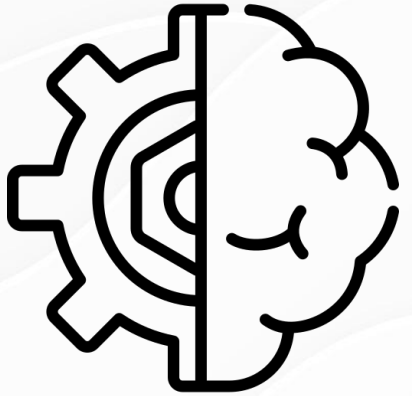

Genre (written by r..
Fiction
Fantasy
Nonfiction
Historical Fiction
Mystery
Young Adult
Romance
Classics
Science Fiction
History
Horror
Poetry
Manga
Philosophy
Childrens
Short Stories
Comics
Thriller
Biography
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
4000
4200
4400
4600
Count of Title


1,437
1,476
4,415
2,744


Sheet 1
Genre (written by readers)
Fiction
Fantasy
Nonfiction
Historical Fiction
Count of Title
10500
10000
9500
9000
8500
8000
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
4,415
2,744
1,476
1,437


Genre (written by readers)
Fiction
Fantasy
Nonfiction
Historical Fiction

Practice



Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Read/ unread % of each category of books on the bookshelf

- Bar chart
- Stacked bar chart
- Pie chart
- Dot plot
- Heatmap
- Histogram

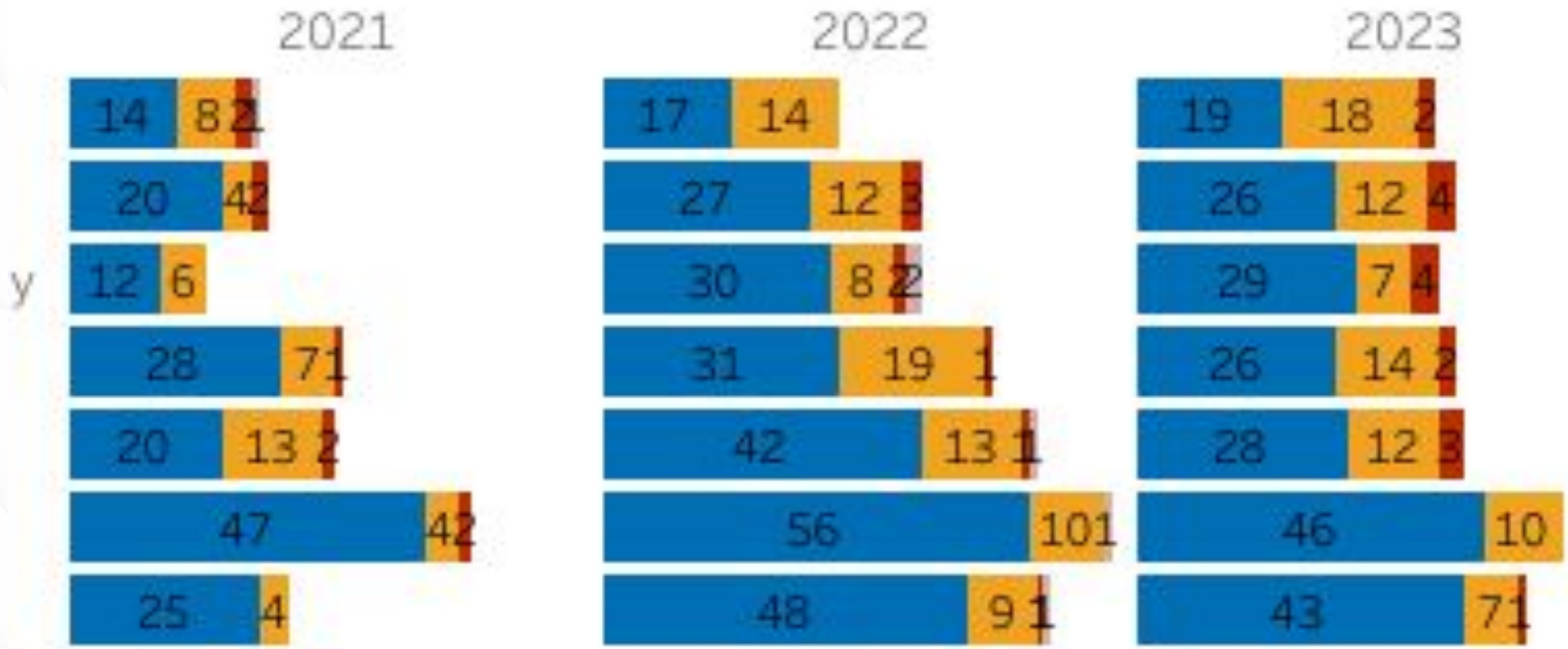
What would I do...

Prof Booknose has a massive bookshelf in his study.

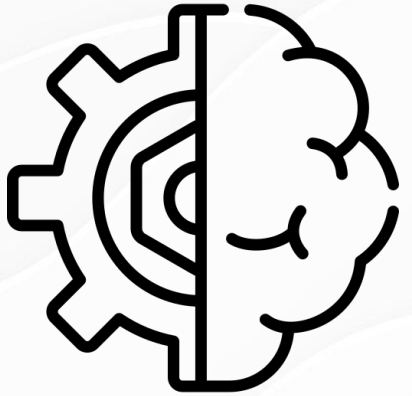
What chart would you choose to show:

Read/ unread % of each category of books on the bookshelf

✅ stacked bar chart - we are comparing % in each genre.


2021
2022
2023
y
14 8 2 1
20 4 2
12 6
28 7 1
20 13 2
47 4 2
25 4
17 14
27 12 3
30 8 2 2
31 19 1
42 13 1 1
56 10 1
48 9 1 1
19 18 2
26 12 4
29 7 4
26 14 2
28 12 3
46 10
43 7 1

Practice



Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Compare books with prof Hardback and prof Bookmark

- Bar chart
- Stacked bar chart
- Pie chart
- Grouped bar chart
- Heatmap
- Histogram

What would I do...

Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Compare books with prof Hardback and prof Bookmark

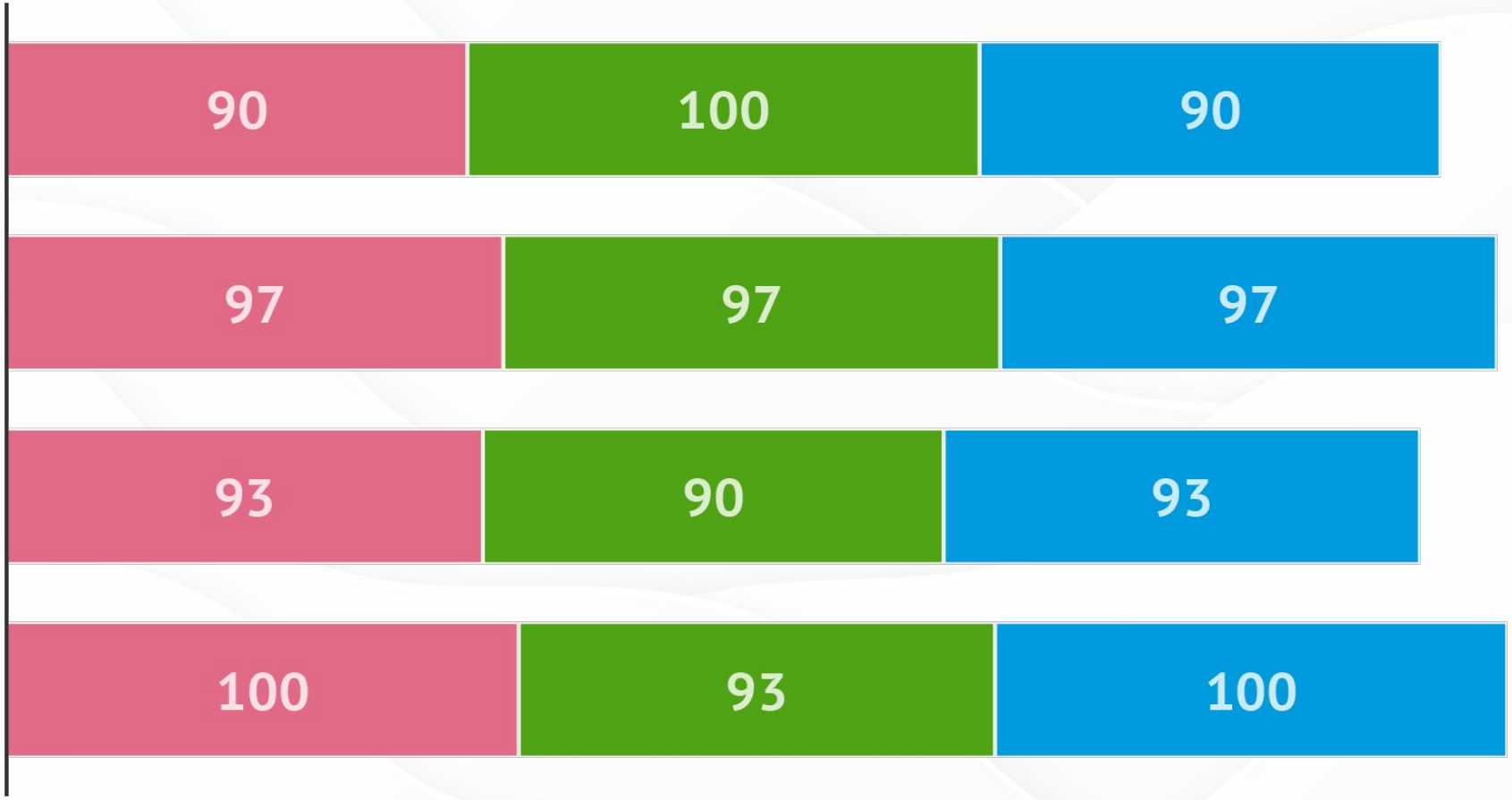
✅ grouped bar chart

Why not stacked bar chart?

Source: Maarten Lambrechts, CC BY SA 4.0

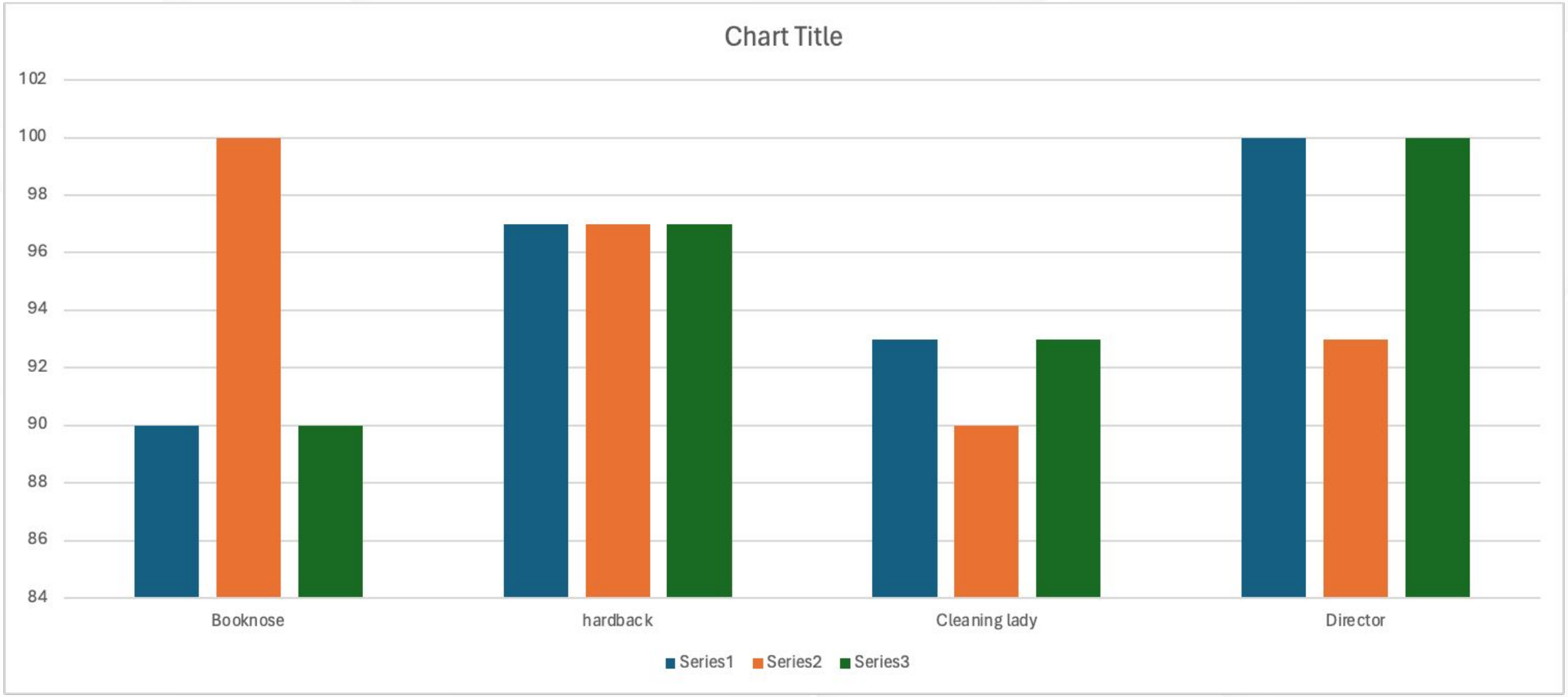
https://data.europa.eu/apps/data-visualisation-guide/stacked-charts

Why not stacked bar chart?



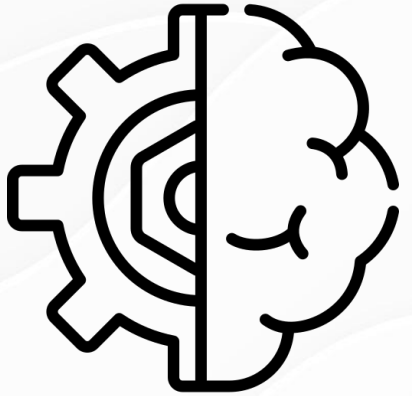
https://data.europa.eu/apps/data-visualisation-guide/stacked-charts

If done in excel, immediate graph...



Starting from 0

Practice



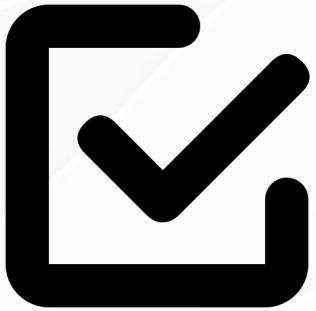
Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Book ratings (given by professors; 1-5) by genres

- Bar chart
- Stacked bar chart
- Box plot
- Pie chart
- Dot plot
- Grouped bar chart

What would I do...



Prof Booknose has a massive bookshelf in his study.

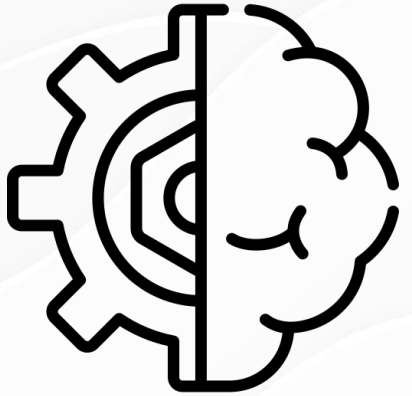
What chart would you choose to show:

Book ratings (given by professors; 1-5) by genres

✅ Dot plot, heatmap if many professors and genres

✅ Grouped bar chart if few professors

Practice



Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Book formats (Hardcover, Paperback, eBook, Audiobook)

- Bar chart
- Stacked bar chart
- Box plot
- Pie chart
- Dot plot
- Histogram

What would I do...

Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Book formats (Hardcover, Paperback, eBook, Audiobook)

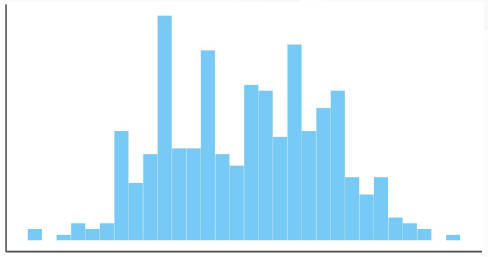
✅ Pie chart

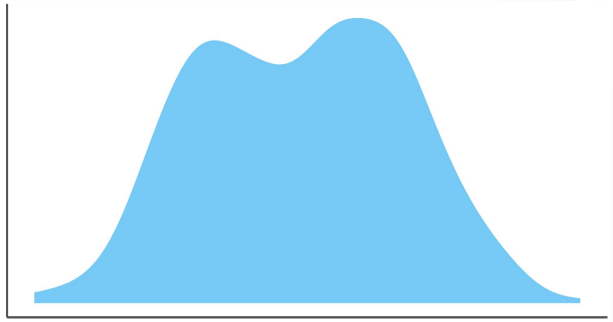
✅ Stacked bar chart

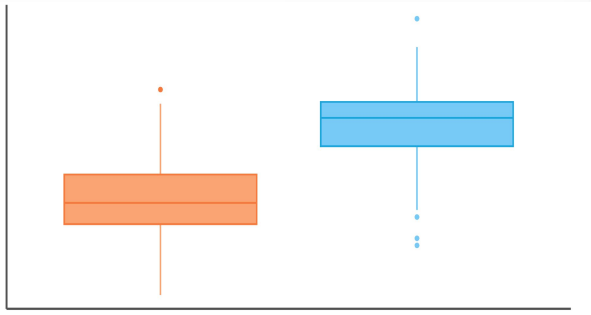
Data type: Distribution

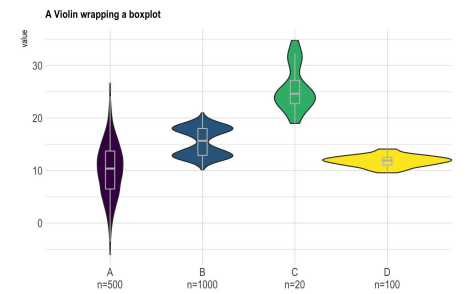
Distributions - probability of a particular value or value range of a variable

- Histogram
- Density plot
- Box plot
- Violin plot

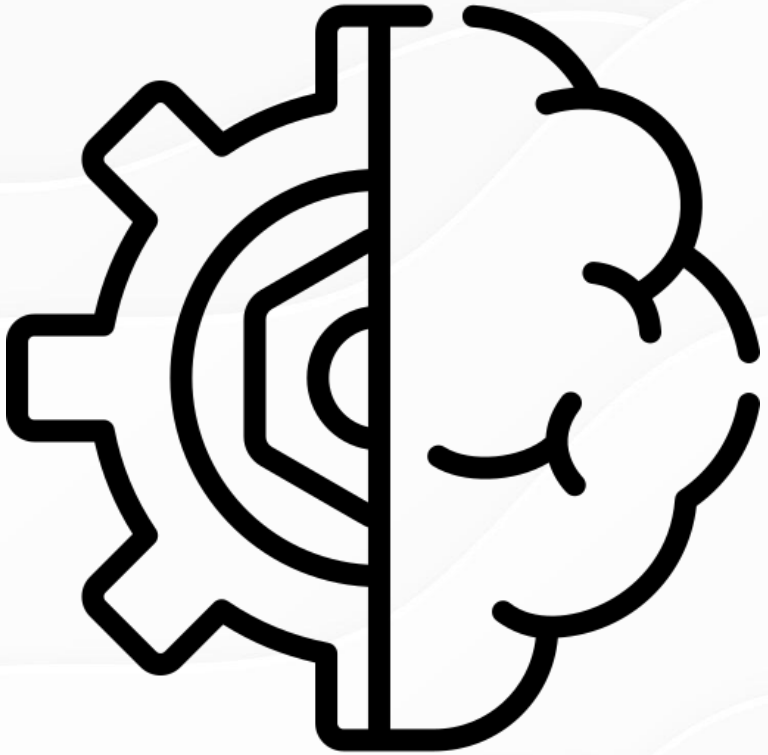








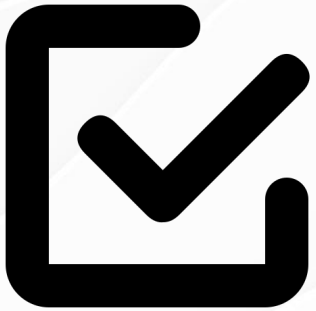
Let’s do this once more.



Which of the following is a distribution?

- ❏ IQ scores
- ❏ Color of cars in a parking lot
- ❏ Age of population
- ❏ Heights of people
- ❏ Names of people in a city
- ❏ Student ID numbers
- ❏ Coin toss

Let’s do this once more.



Which of the following is a distribution?

✅ IQ scores

❌ Color of cars in a parking lot

✅ Age of population

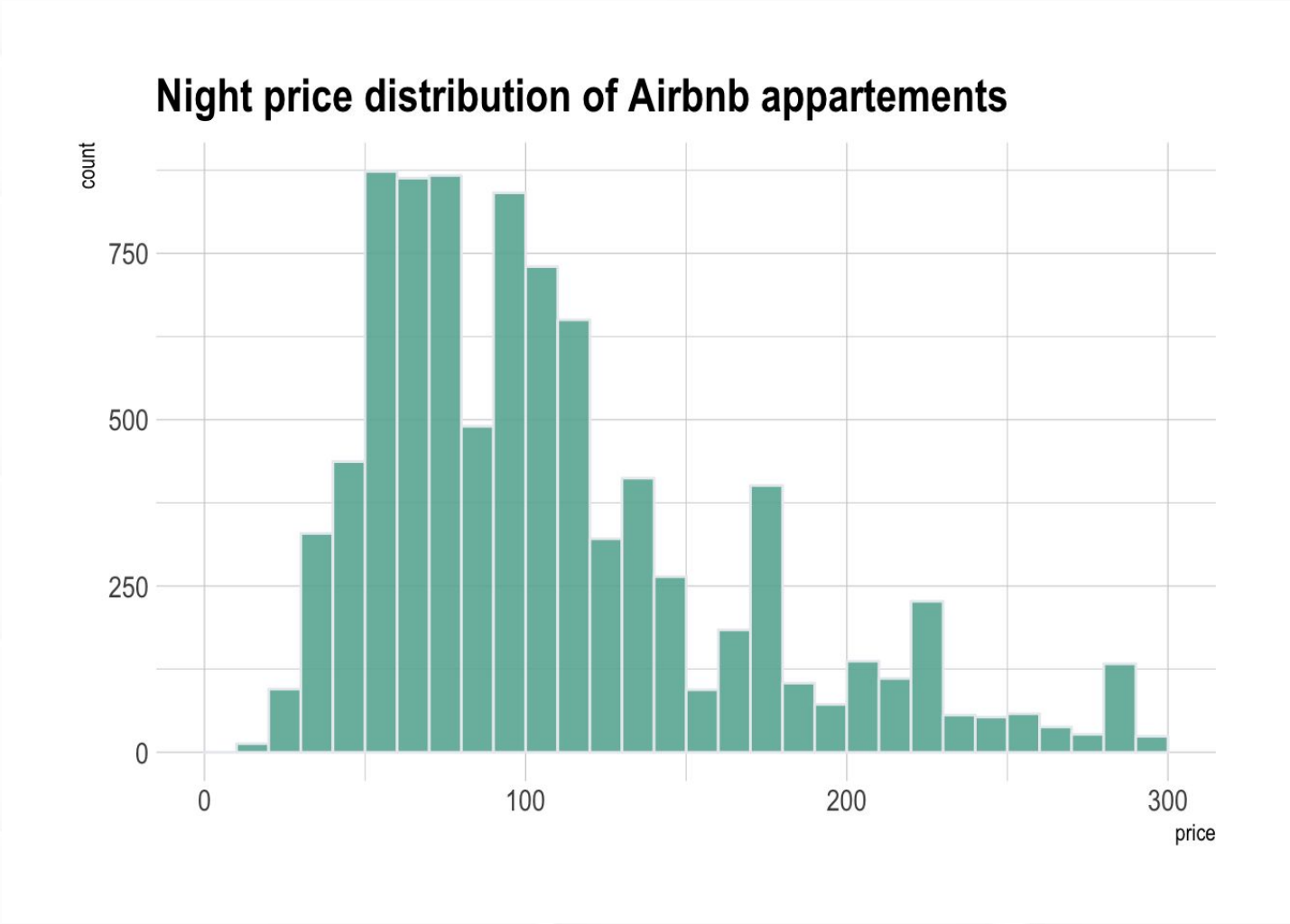
✅ Heights of people

❌ Names of people in a city

❌ Student ID numbers

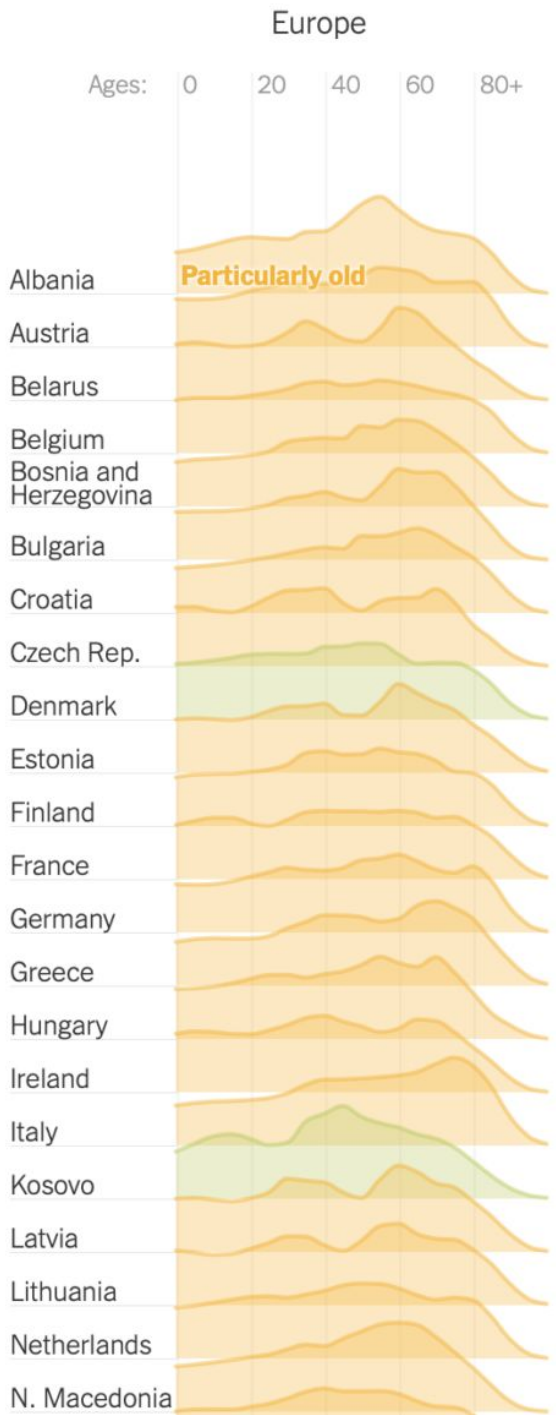
✅ Coin toss

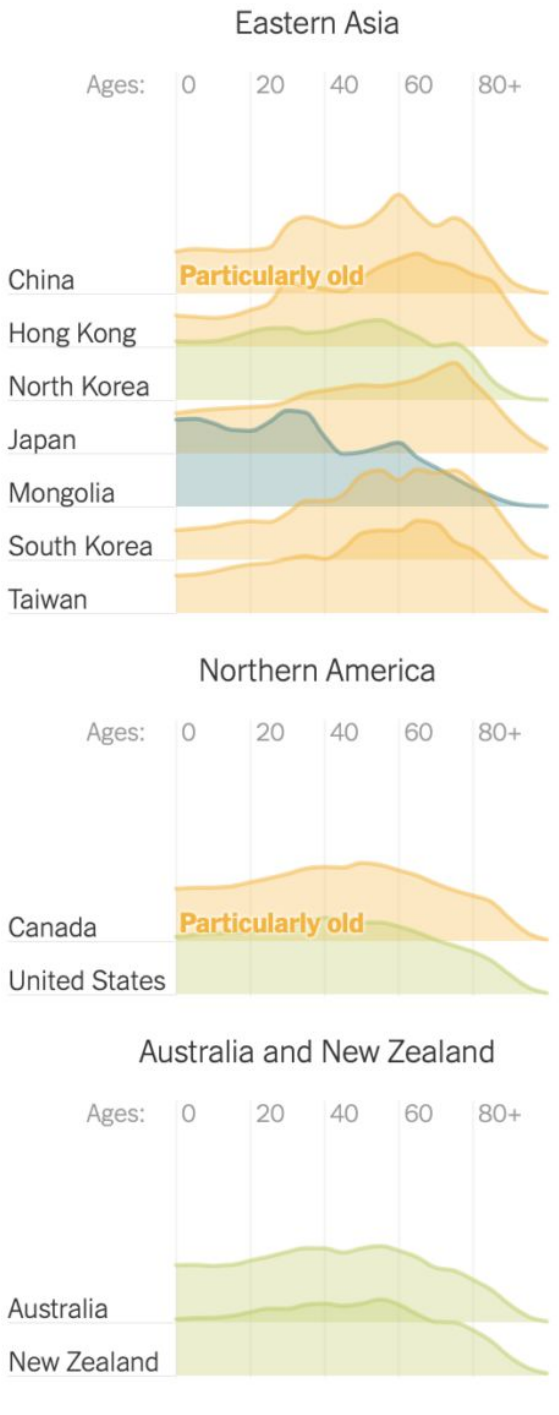
Example: Histogram

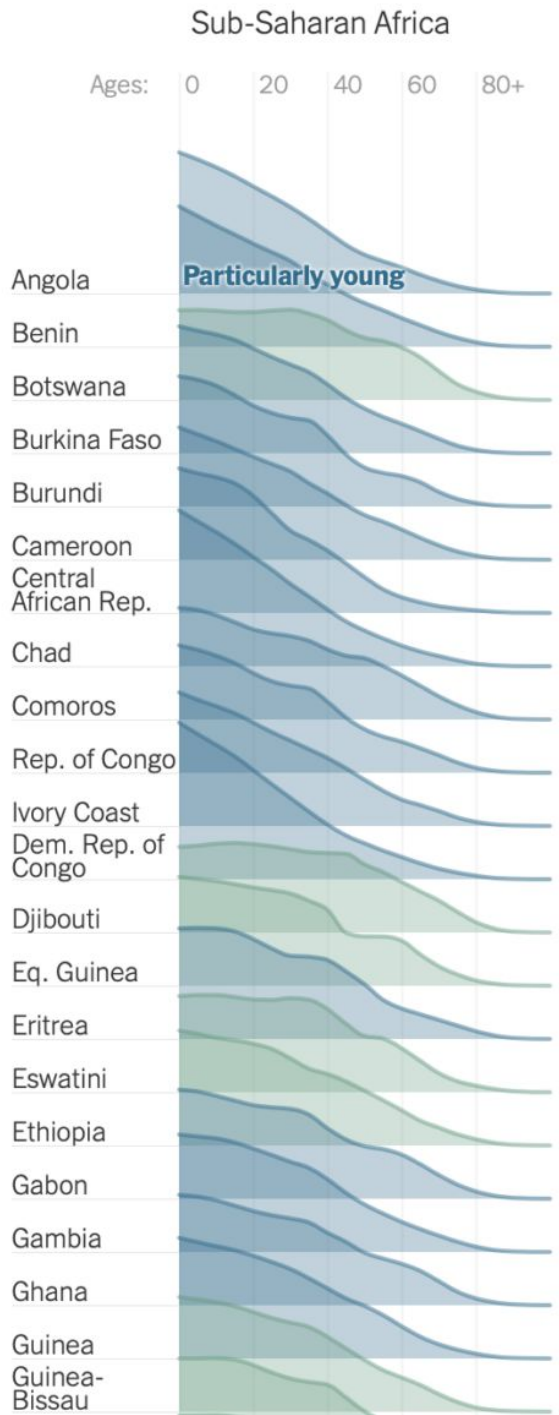


https://www.data-to-viz.com/graph/histogram.html

Example: Density

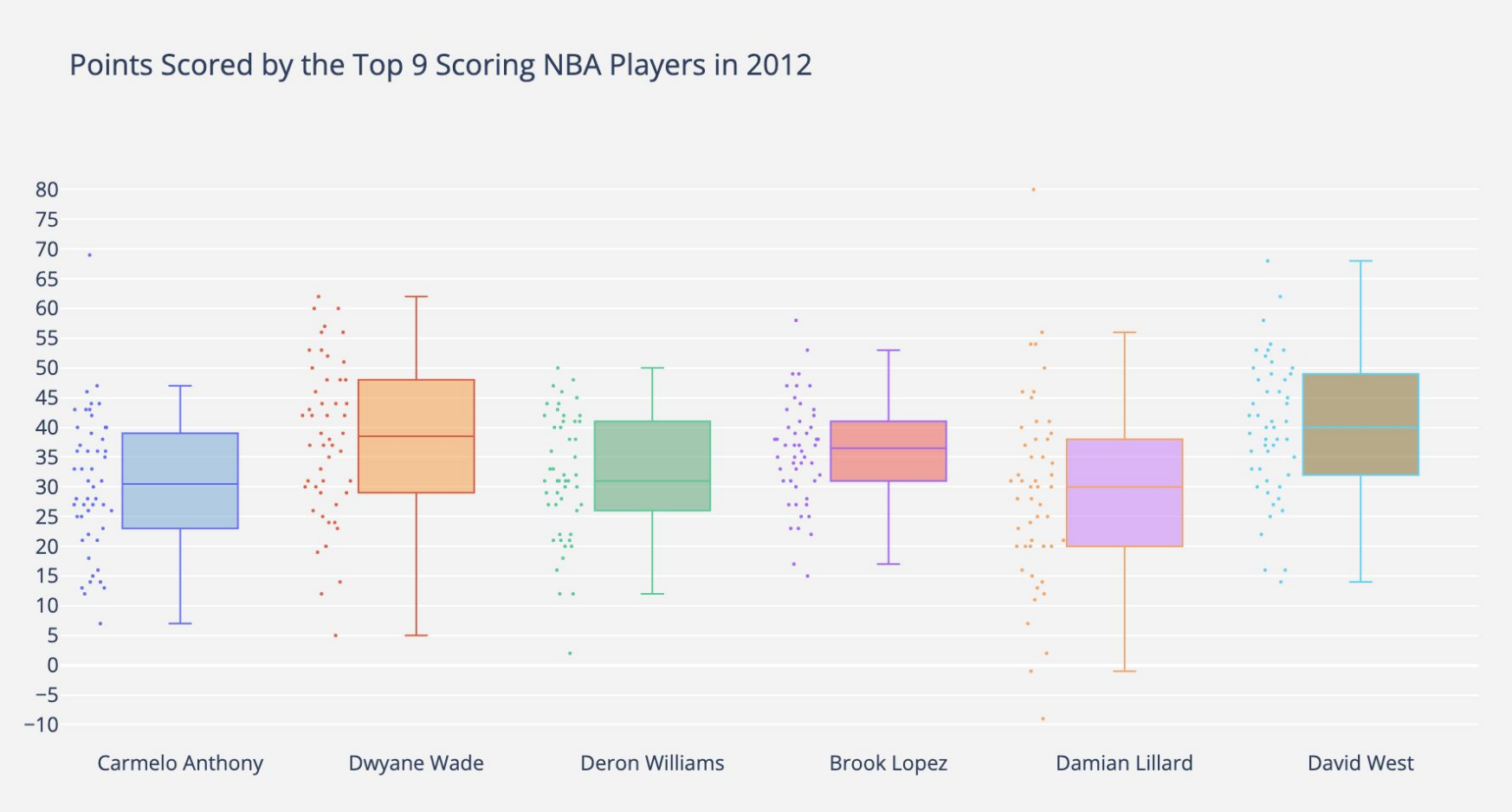






https://flowingdata.com/2023/07/17/age-shifts-around-the-world/

Example: box plot



https://plotly.com/python/box-plots/

Box plot vs violin plot

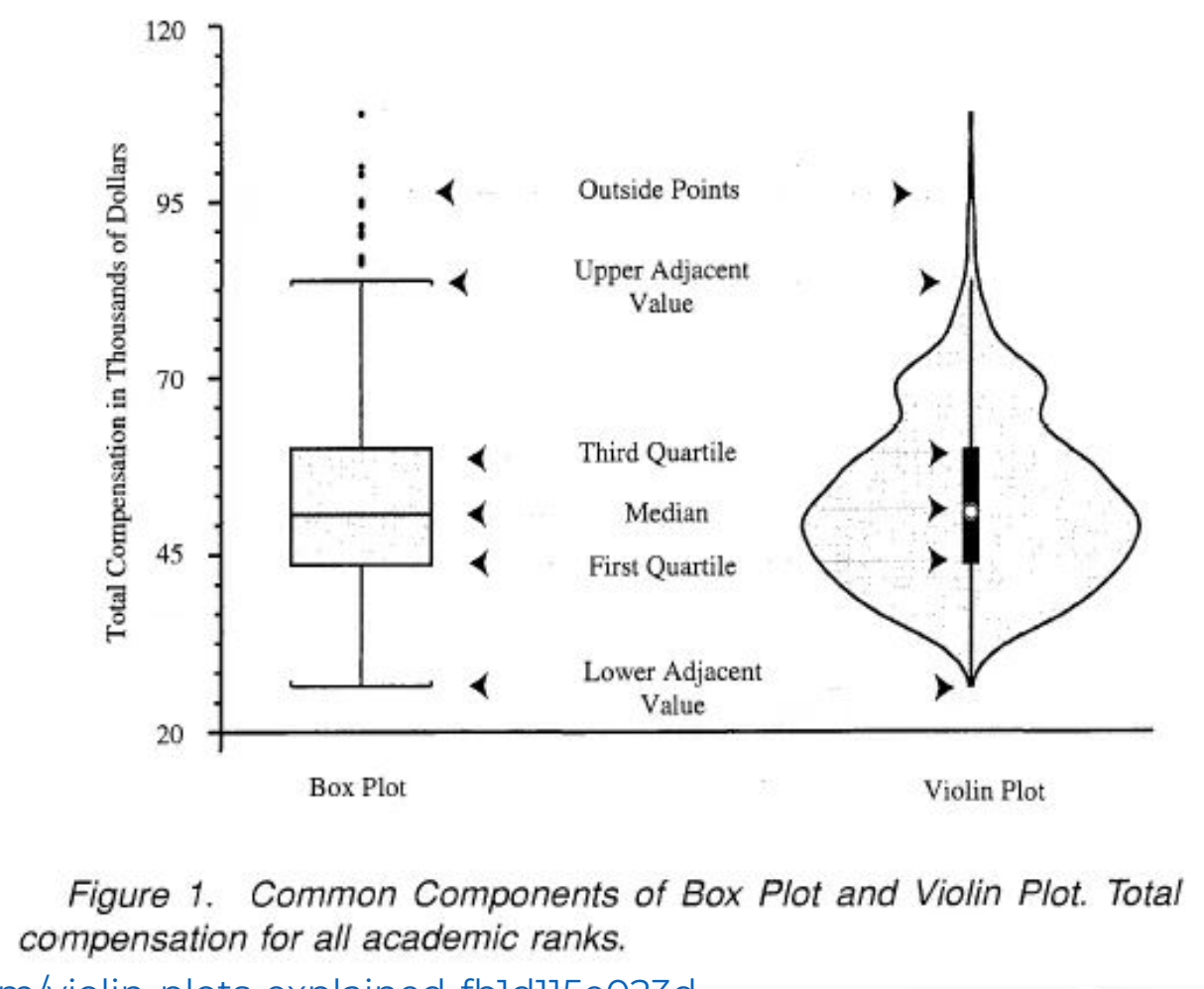
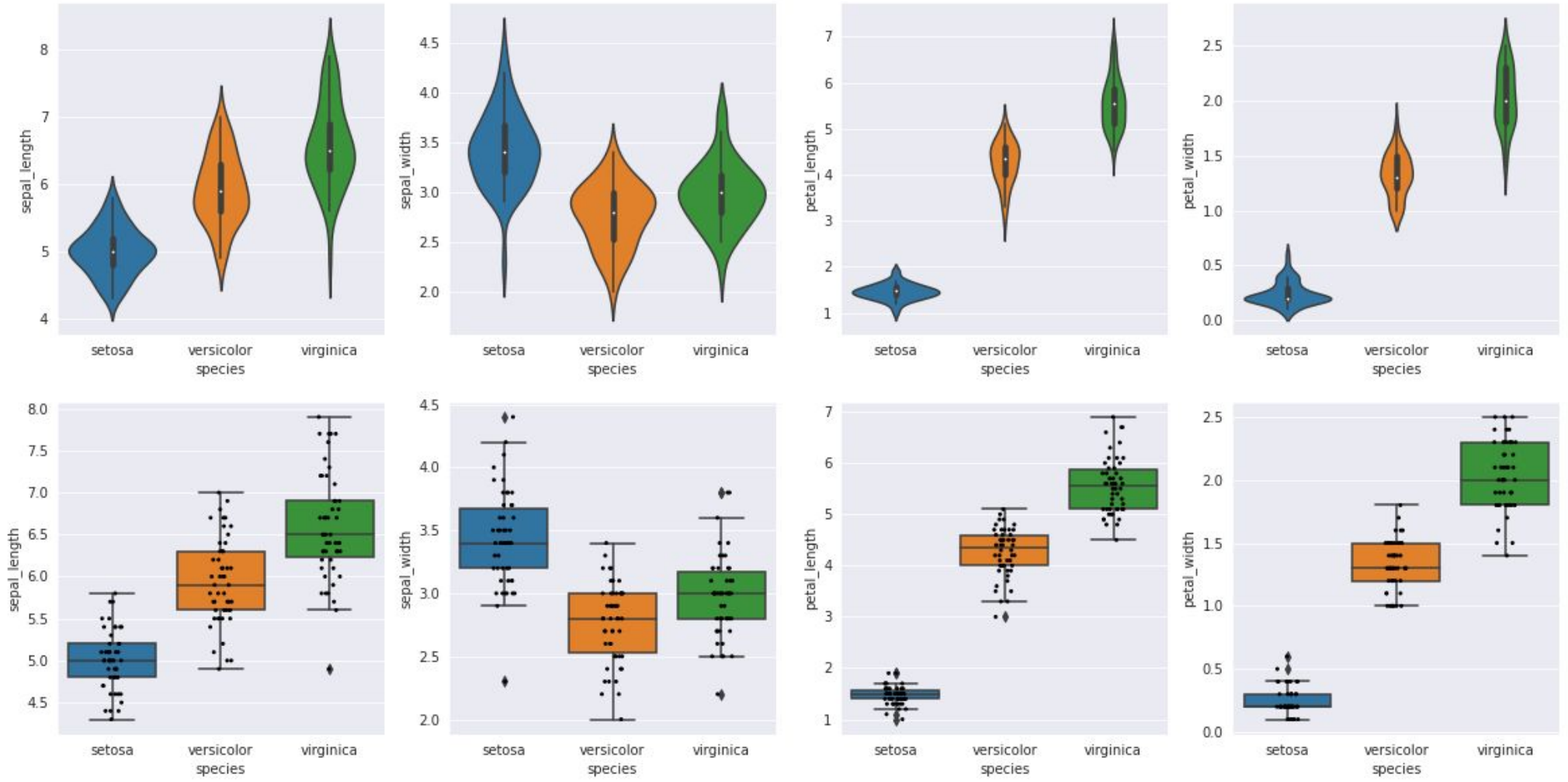


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

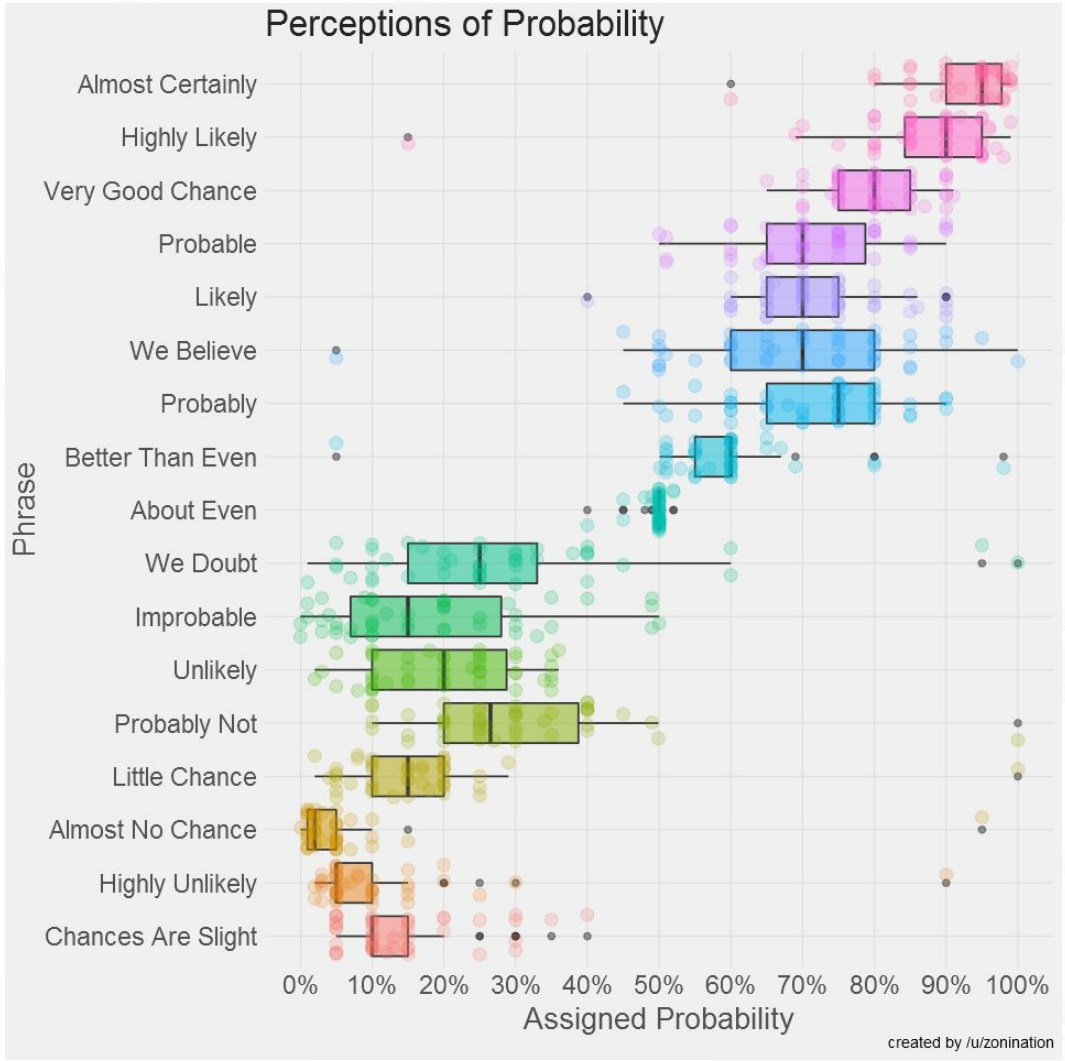
https://towardsdatascience.com/violin-plots-explained-fb1d115e023d

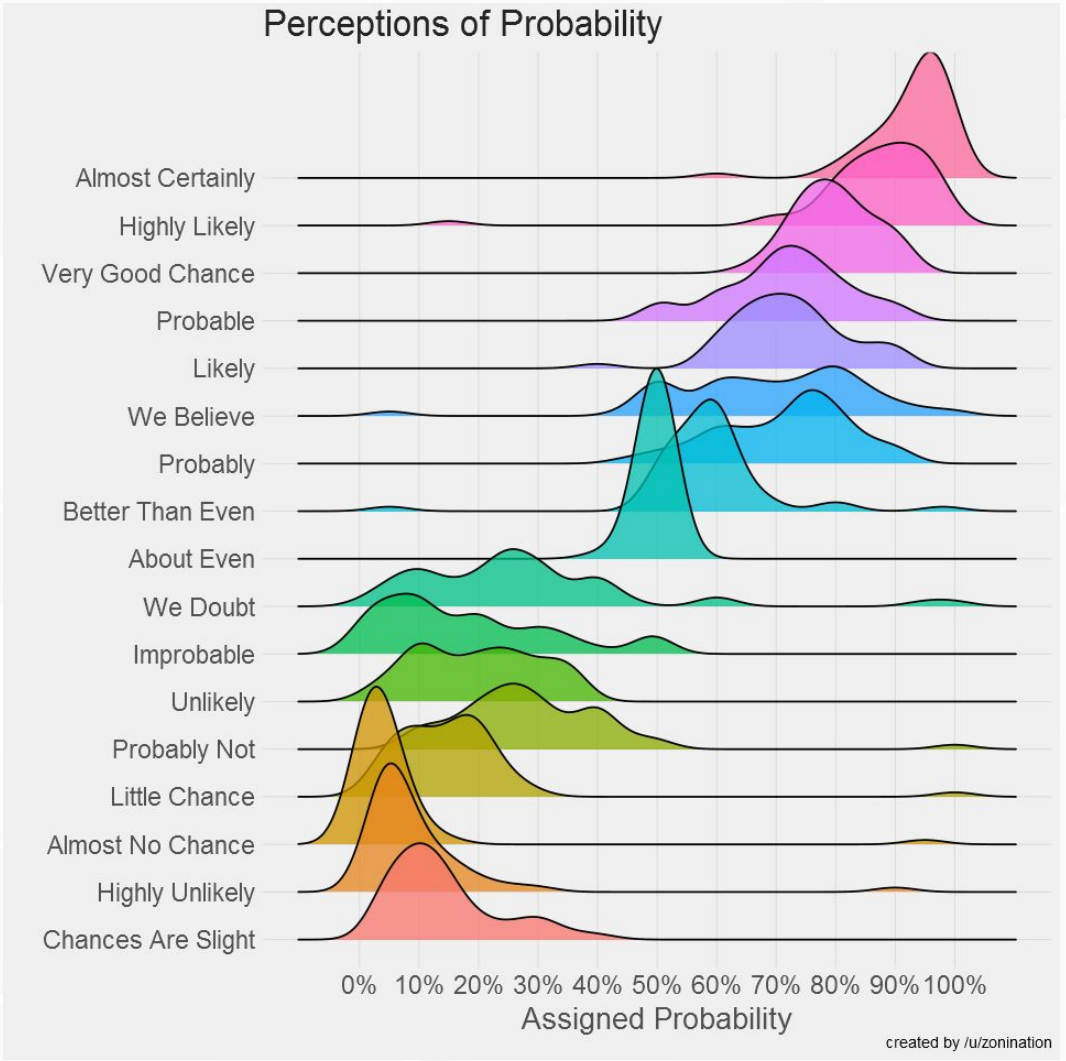
Same data, different graph



https://deepnote.com/@econdesousa/ViolinPlotvsBoxPlot-aadf0c53-53b4-4221-89b9-4388c54c68bd

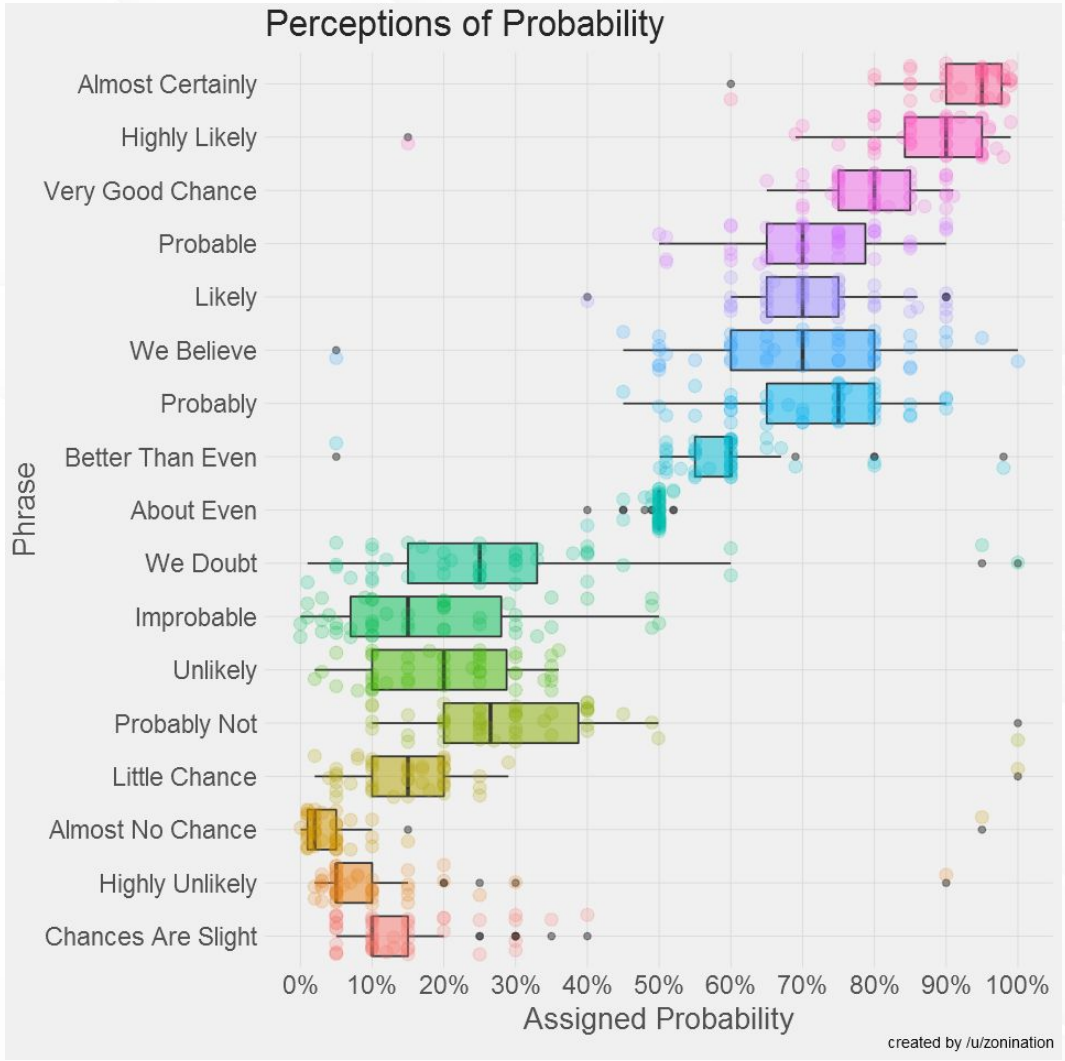
Same data, different graph

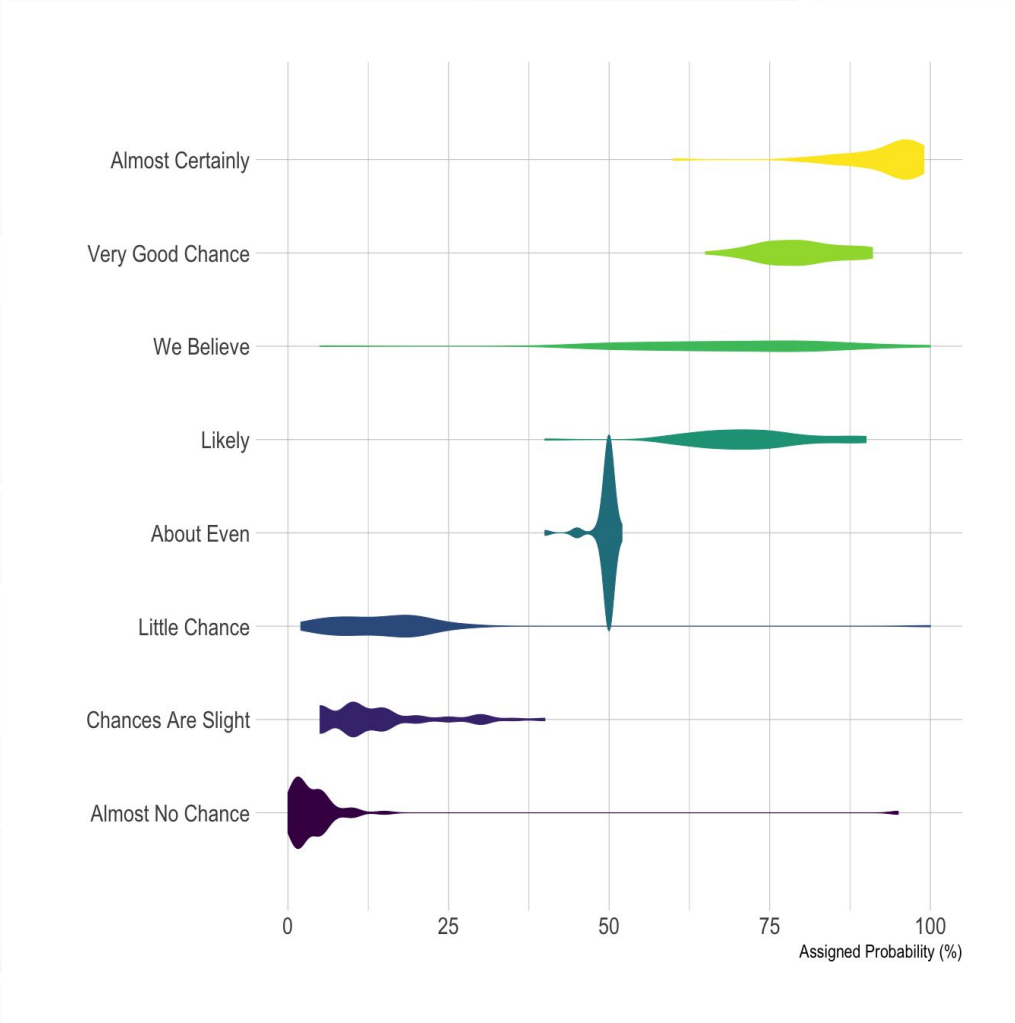




https://github.com/zonination/perceptions

Same data, different graph





https://github.com/zonination/perceptions
https://www.data-to-viz.com/graph/violin.html

How to choose

DENSITY

Very close from an histogram. Show a density of distribution instead of bars

RIDGELINE

Allows to compare the distribution of a few variables. Usefull if there is a pattern to observe

BOXPLOT

Summarize the distribution of several numeric variables using boxes. Use it if you have a limited dataset size.

VIOLIN

Show the density of several numeric variables. Use it with a large amount of data

HISTOGRAM

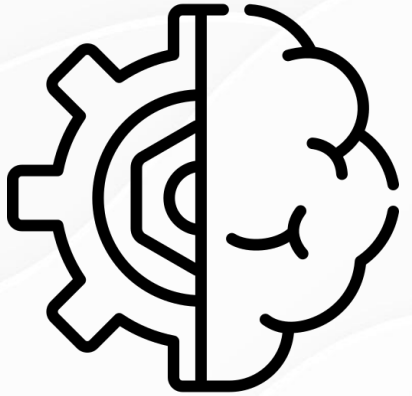
Very close from a density. Cut the data in several bins and show the number of entry per bin.

https://www.data-to-viz.com/graph/density.html

Chart types comparison for distributions

	Histogram	Density plot	Box plot	Violin plot
Visualizing a single distribution	✅	✅	✅	✅
Visualizing multiple distributions at the same time	❌	✅	✅	✅
Visualize the underlying probability distribution of the data	❌	✅	✅	✅
Bandwidth parameters change chart looks	✅	✅	❌	✅
Area under the curve equals one	❌	✅	❌	✅
Tendency to produce the appearance of data where none exists, in particular in the tails	❌	✅	❌	✅

Practice



Prof Booknose has a massive bookshelf in his study.
What chart would you choose to show:

Distribution of book publication years

- Heatmap
- Density plot
- Histogram
- Box plot
- Bar plot
- Violin plot

What would I do...

Prof Booknose has a massive bookshelf in his study.
What chart would you choose to show:

Distribution of book publication years

✅ histogram

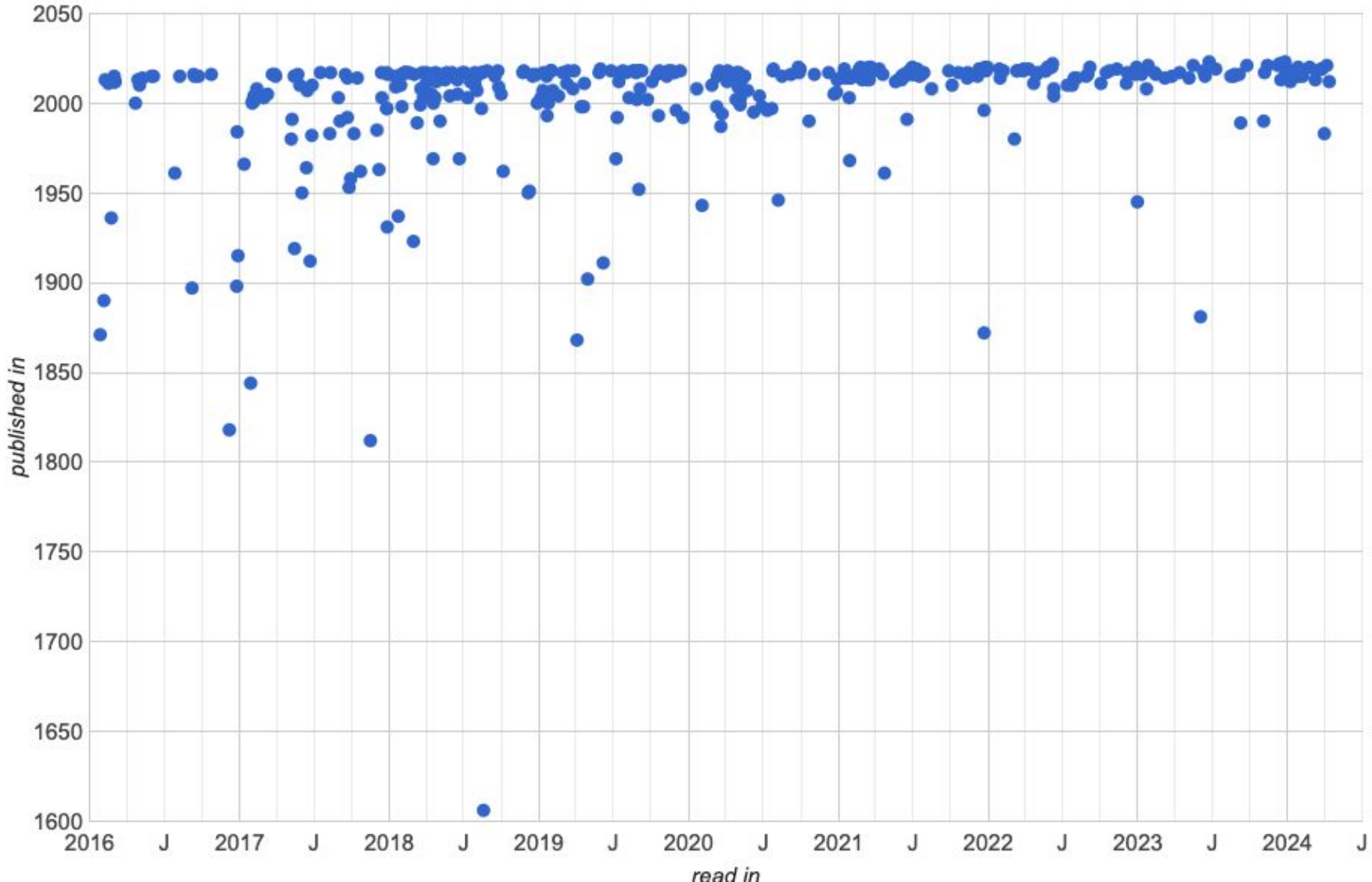
Why can’t I use dot plot instead of histogram?

- Histograms are better suited for large data sets
  - dot plot can become crowded and difficult to read
- Histograms are ideal for continuous data
- Histograms allow you to group data into intervals - simplifying. Patterns and trends more apparent
- Histogram shows easily when data is skewed

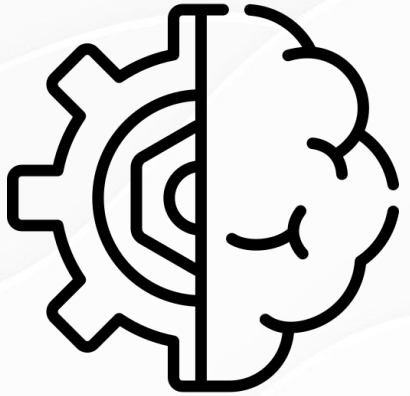
goodreads Home My Books Browse Community Search books

My Books > Stats

books pages publication year books over time pages over time



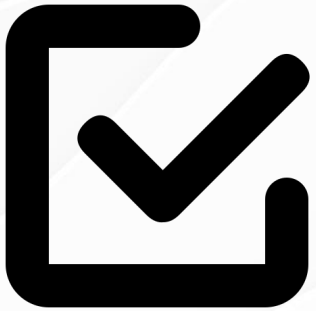
Practice



Prof Booknose has a massive bookshelf in his study.
What chart would you choose to show:
Distribution of book lengths per genre. Outliers?

- Heatmap
- Density plot
- Histogram
- Box plot
- Bar plot
- Violin plot

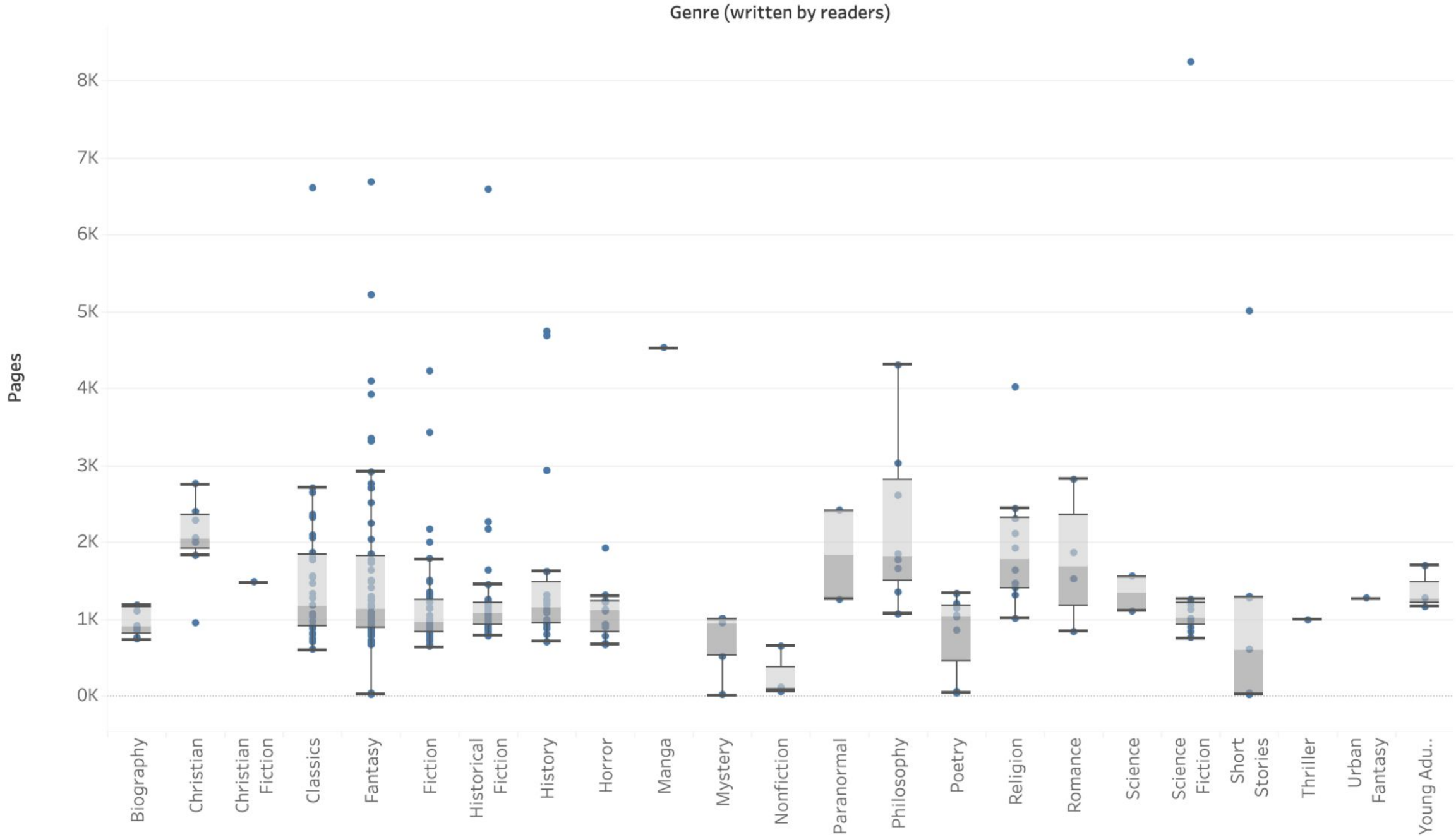
What would I do...



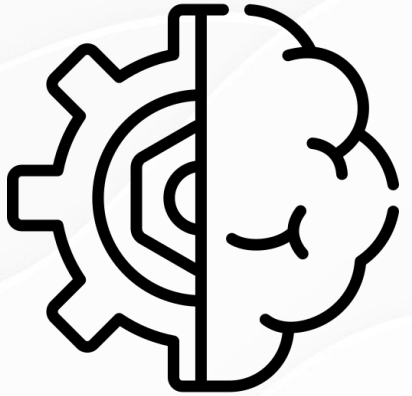
Prof Booknose has a massive bookshelf in his study.
What chart would you choose to show:
Distribution of book lengths per genre. Outliers?

✅ box plot

✅ violin plot -a lot of data


Genre (written by readers)
Pages
8K
7K
6K
5K
4K
3K
2K
1K
0K
Biography
Christian
Christian Fiction
Classics
Fantasy
Fiction
Historical Fiction
History
Horror
Manga
Mystery
Nonfiction
Paranormal
Philosophy
Poetry
Religion
Romance
Science
Science Fiction
Short Stories
Thriller
Urban Fantasy
Young Adu..

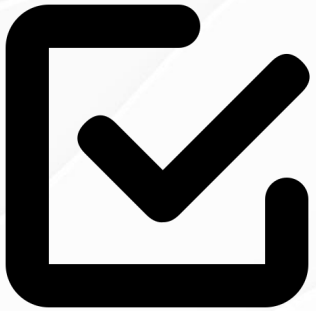
Practice



Prof Booknose has a massive bookshelf in his study.
What chart would you choose to show:
Distribution of book prices for the professor's books.

- Heatmap
- Histogram
- Box plot
- Bar plot
- Density plot
- Violin plot

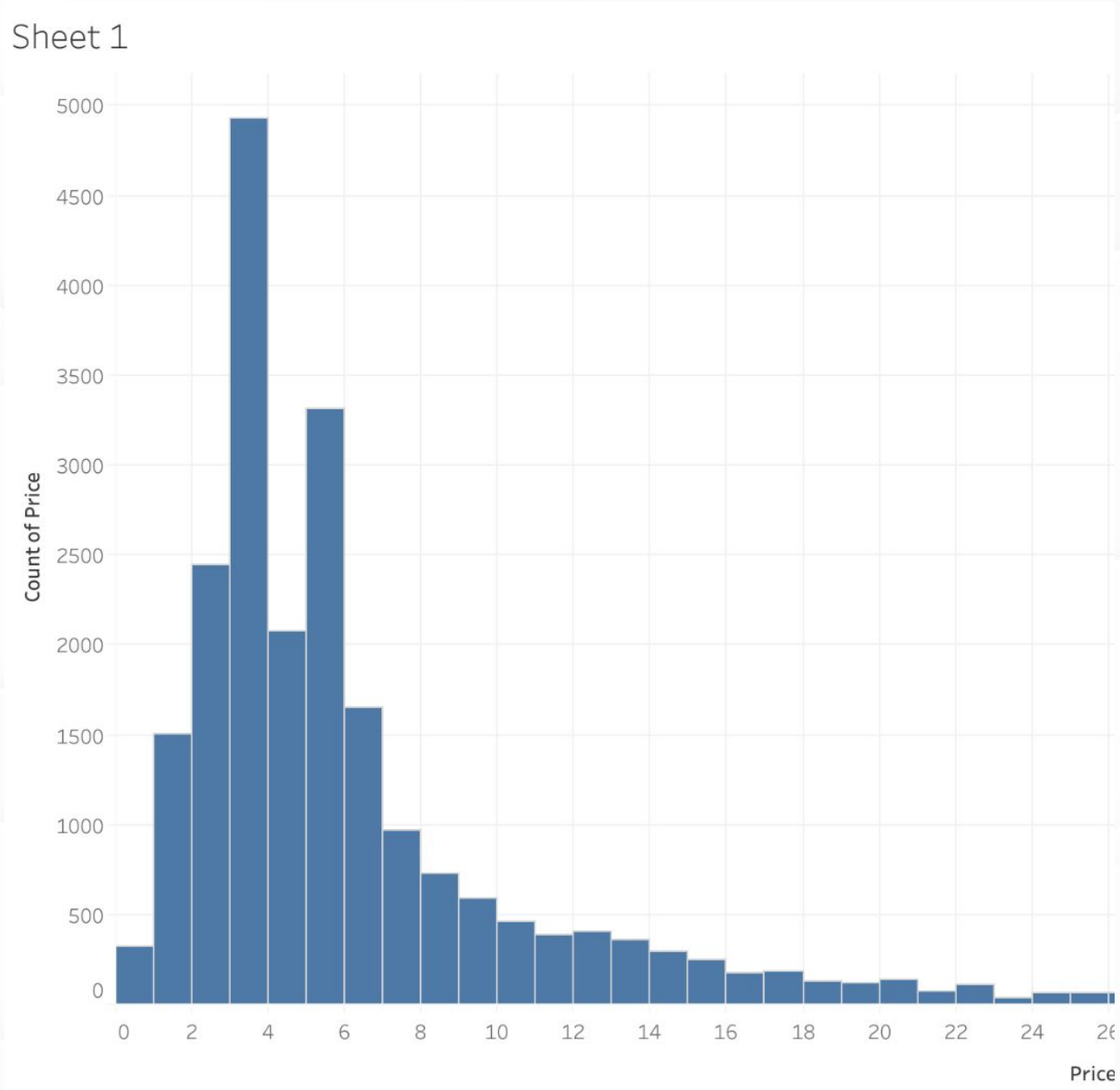
What would I do...



Prof Booknose has a massive bookshelf in his study.
What chart would you choose to show:
Distribution of book prices for the professor's books.

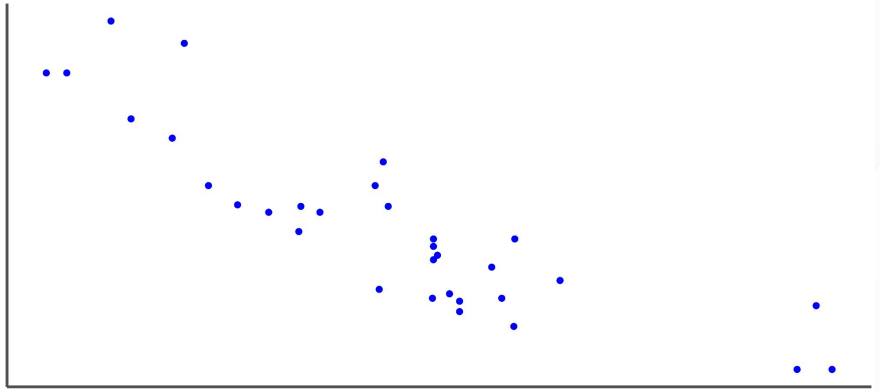
✅ histogram (price ranges)

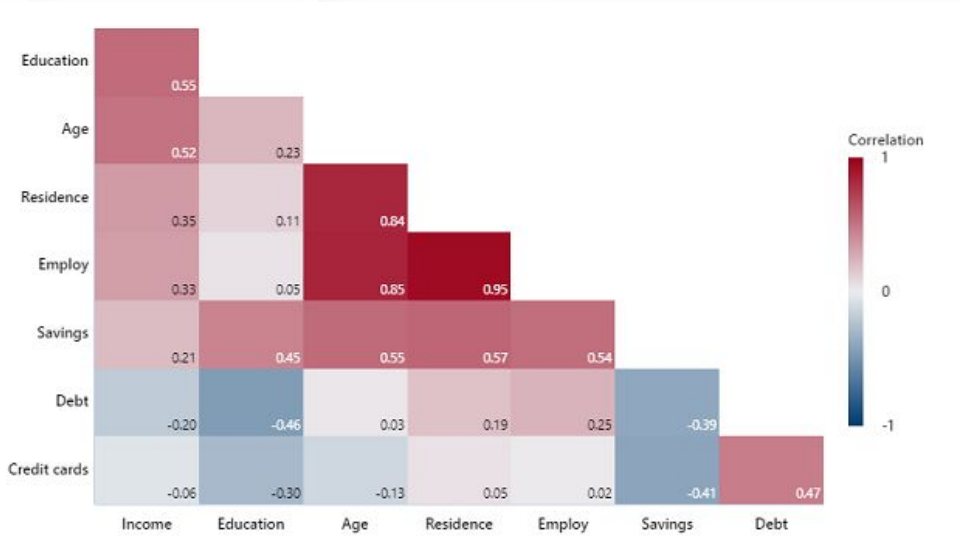
✅ density plot


Sheet 1
Count of Price
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
2
4
6
8
10
12
14
16
18
20
22
24
Price

Relationship between two quantitative variables

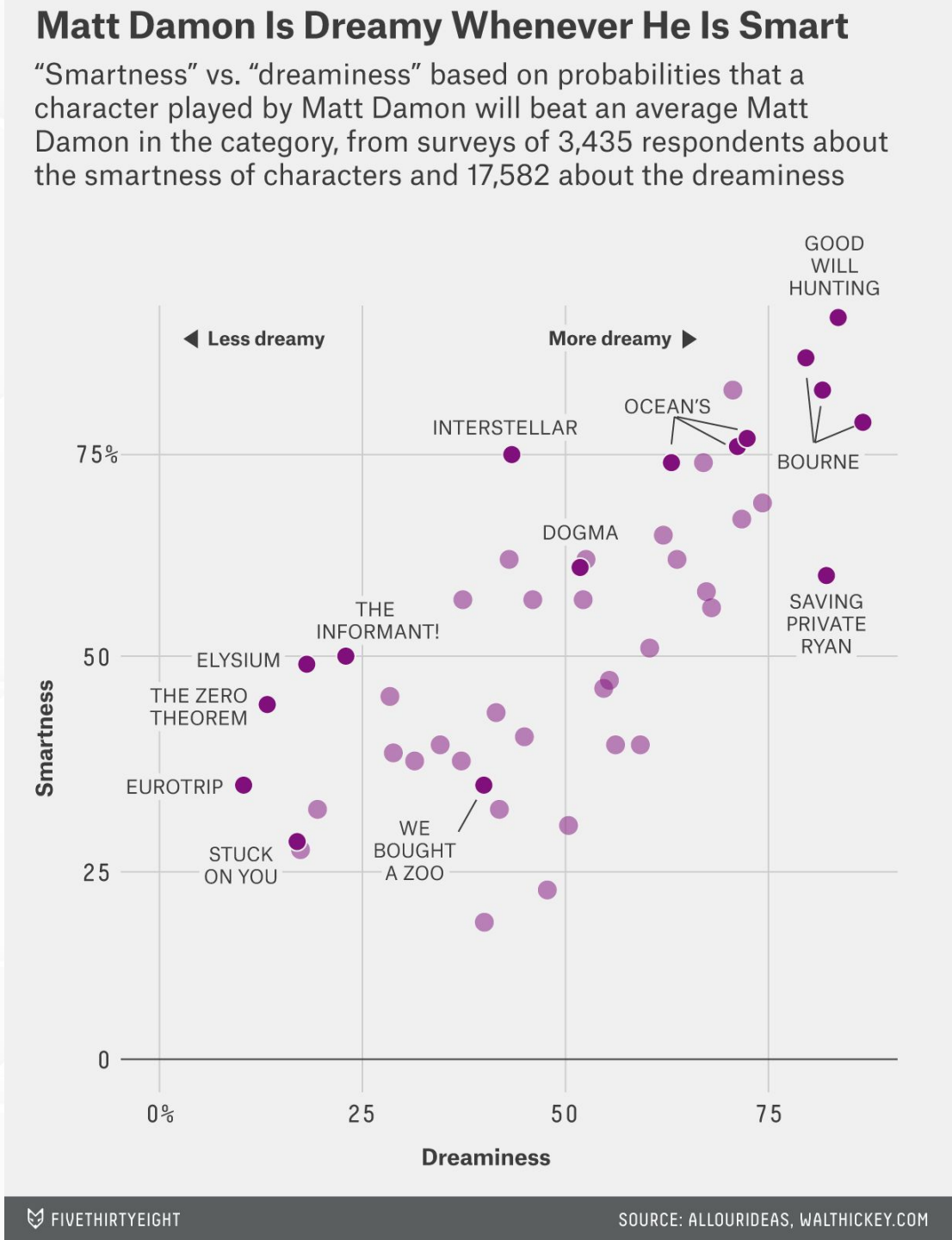
- Scatterplot
- Correlogram
- Paired data





https://support.minitab.com/en-us/minitab/21/help-and-how-to/graphs/correlogram/before-you-start/example/

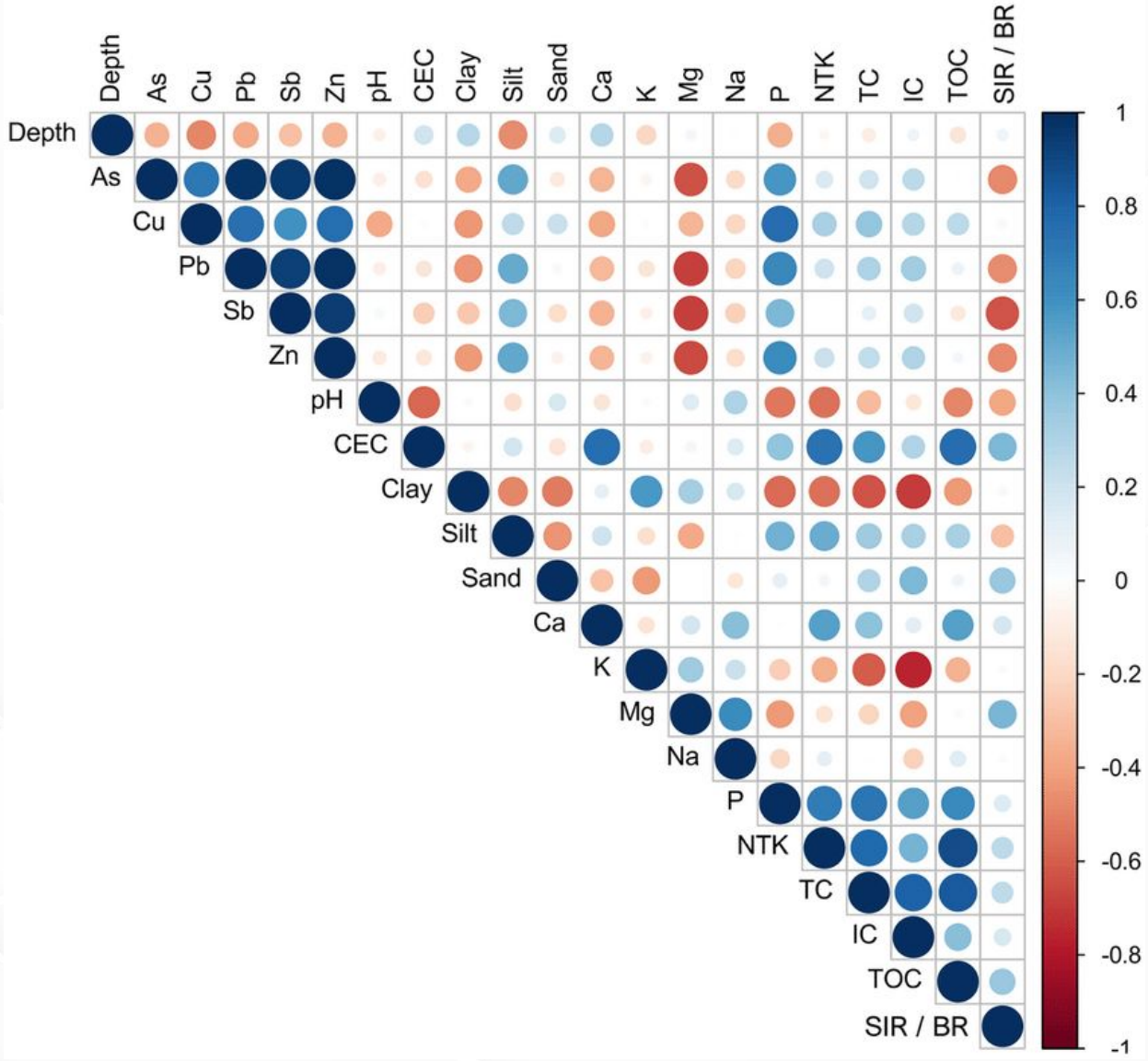
Example: Scatterplot



https://fivethirtyeight.com/features/matt-damon-the-martian-sexy-smart/

Example: Correlogram

Correlogram of Spearman correlation test performed between pedological properties and soil PTE (metal and metalloid pseudo-total concentrations in mg/kg) measured at both soil depths (0–10 cm and 10–20 cm). Blue dots correspond to the positive correlations and red dots to the negative correlations. The size and color intensity of the dots are proportional to the correlation coefficient value.



https://www.researchgate.net/publication/361325582_Coronilla_juncea_a_native_candidate_for_phytostabilization_of_potentially_toxic_elements_and_restoration_of_Mediterranean_soils

Heckenroth, Alma & Prudent, Pascale & Folzer, Hélène & Rabier, Jacques & Criquet, Stéven & Saatkamp, Arne & Salducci, Marie-Dominique & Vassalo, Laurent & Laffont-Schwob, Isabelle. (2022). Coronilla juncea, a native candidate for phytostabilization of potentially toxic elements and restoration of Mediterranean soils. Scientific Reports. 12. 10.1038/s41598-022-14139-4.

Example: paired data

Szkwara, Jaclyn & Milne, Nikki & Rathbone, Evelyne. (2020). A prospective quasi-experimental controlled study evaluating the use of dynamic elastomeric fabric orthoses to manage common postpartum ailments during postnatal care.

Women's Health. 16. 174550652092719. 10.1177/1745506520927196.

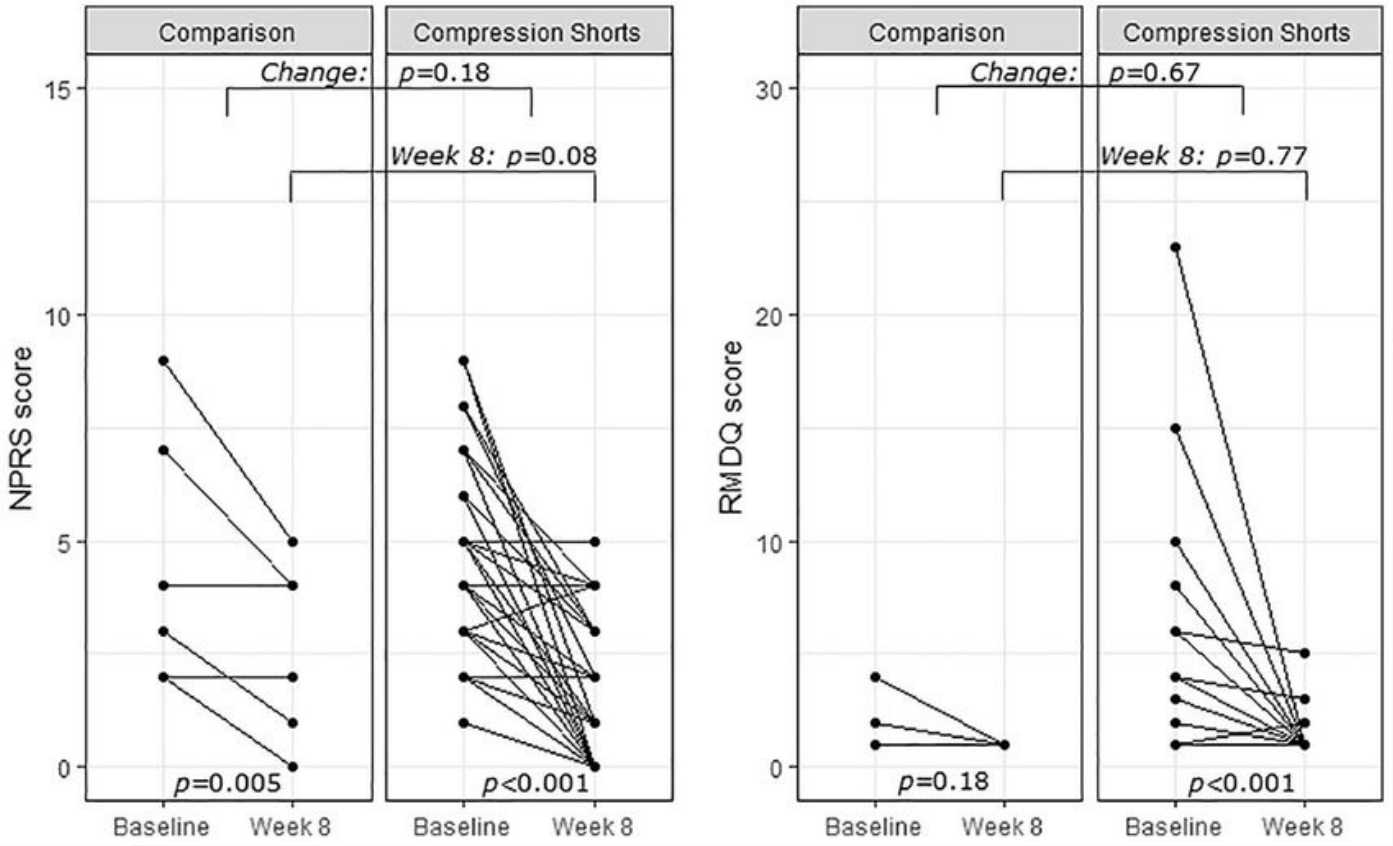
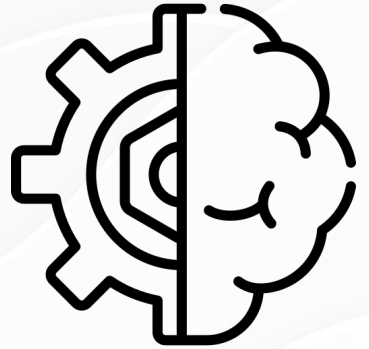


Chart types comparison for associations

	Scatter plot	Bubble chart	Correlogram	Dimension reduction (PCA)
Plot the relationship of just two variables	✅	❌	❌	❌
Plot the relationship of 2+ variables	❌	✅	✅	❌
Plot very high-dimensional datasets	❌	❌	❌	✅
Fairly abstract	❌	❌	✅	❌
Difficult to visually ascertain the strengths of associations between the various variables	❌	✅	❌	❌

Practice



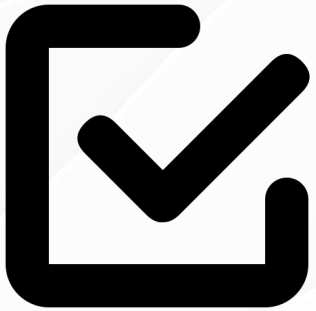
Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Book price and rating. Which books were worth buying?

- Box plot
- Density plot
- Heatmap
- Scatter plot
- Bar plot
- Histogram

What would I do...



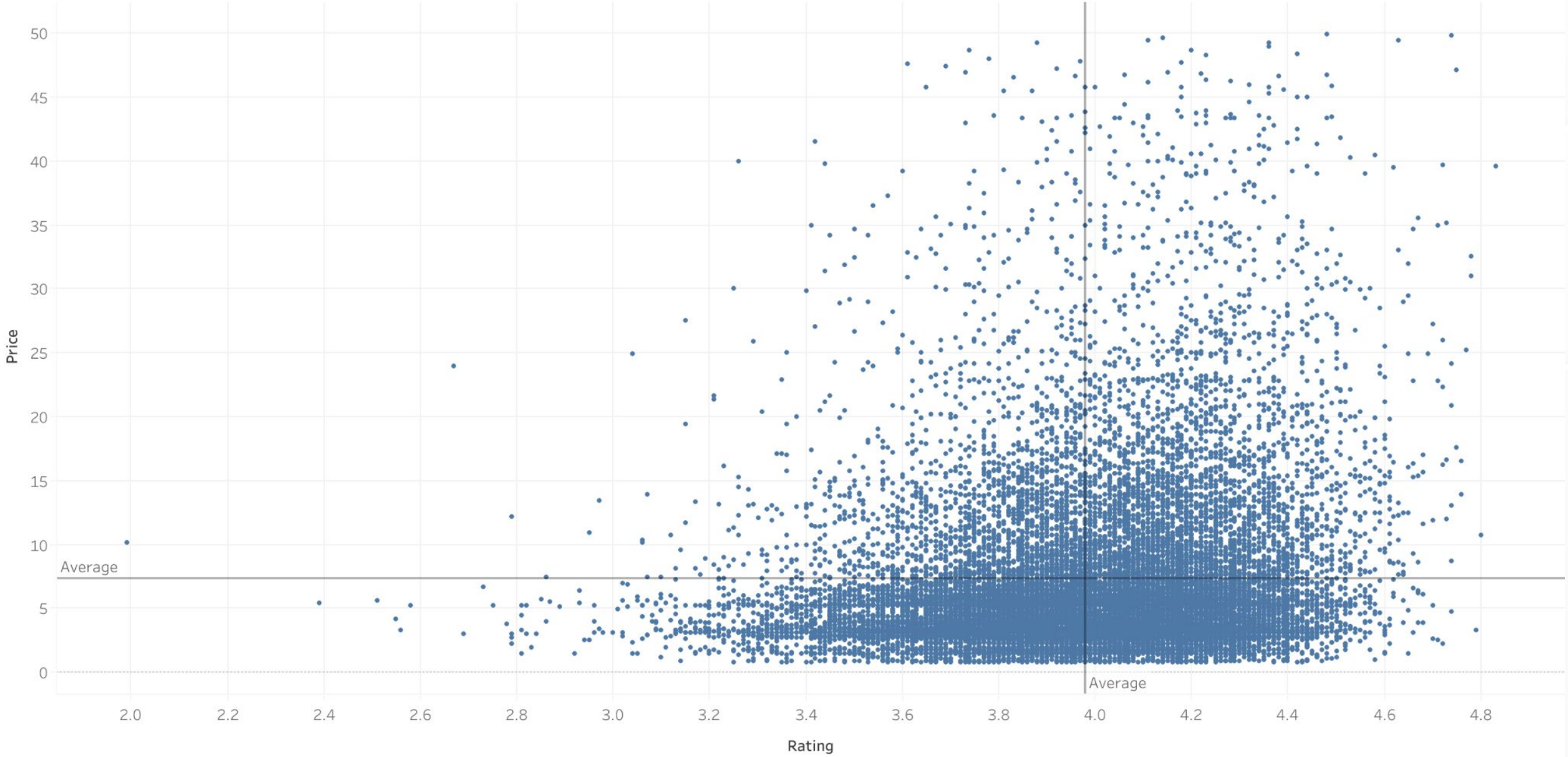
Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

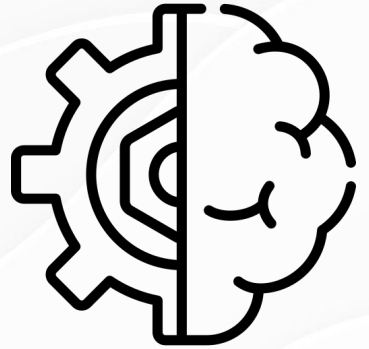
Book price and rating. Which books were worth buying?

✅ scatter plot

Sheet 1


Price
50
45
40
35
30
25
20
15
10
5
0
Average
Average
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
4.6
4.8
Rating

Practice



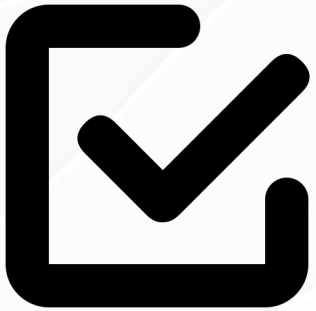
Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Book prices in different genres. Are genres priced differently?

- Box plot
- Density plot
- Heatmap
- Scatter plot
- Bar plot
- Histogram

What would I do...

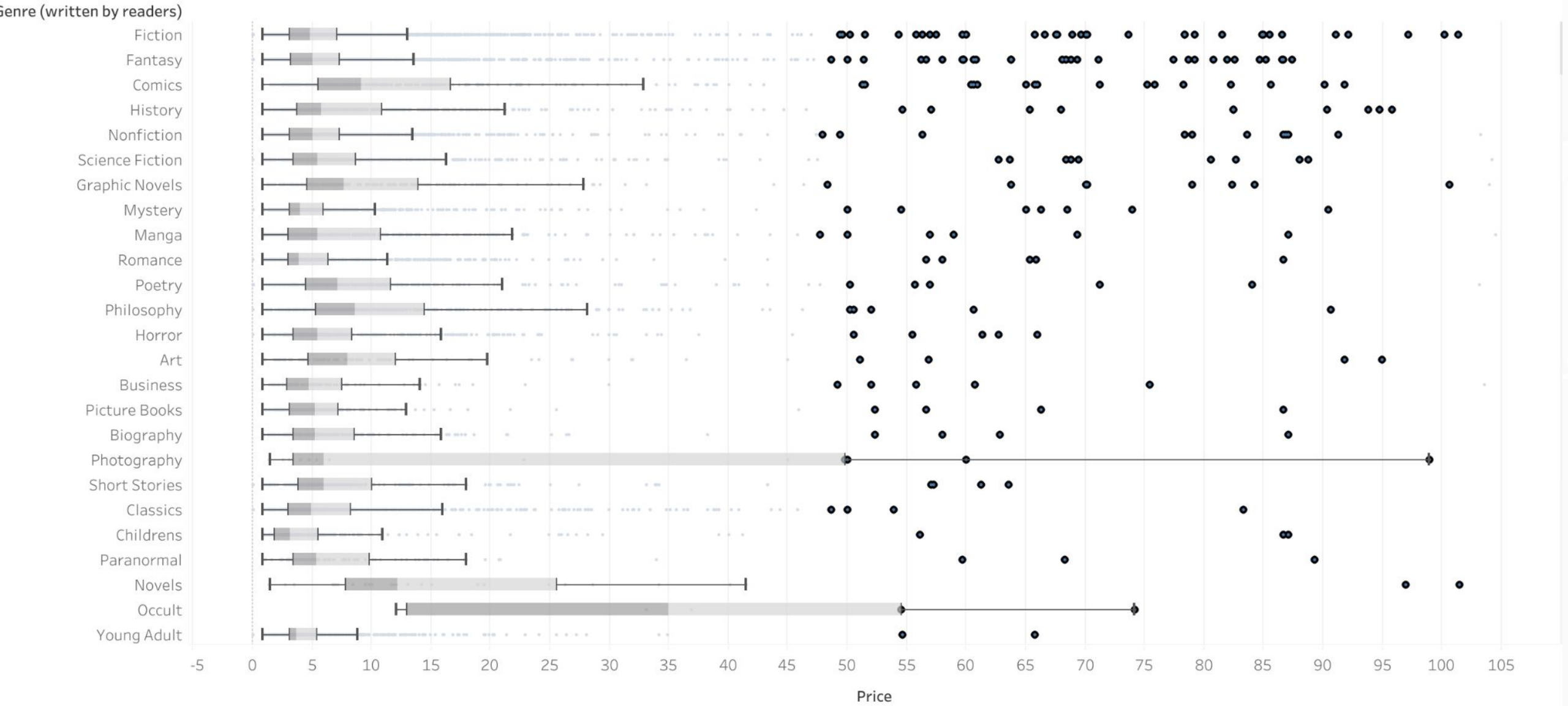


Prof Booknose has a massive bookshelf in his study.

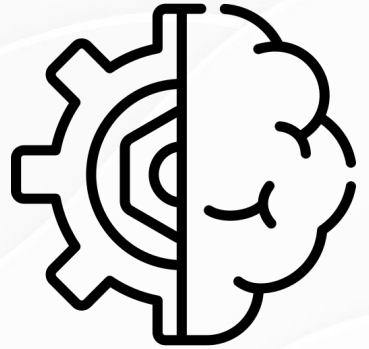
What chart would you choose to show:

Book prices in different genres. Are genres priced differently?

✅ Box plot


Genre (written by readers)
Fiction
Fantasy
Comics
History
Nonfiction
Science Fiction
Graphic Novels
Mystery
Manga
Romance
Poetry
Philosophy
Horror
Art
Business
Picture Books
Biography
Photography
Short Stories
Classics
Childrens
Paranormal
Novels
Occult
Young Adult
-5
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
Price

Practice



Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Book Genre Vs. Reading Time

- Box plot
- Density plot
- Heatmap
- Scatter plot
- Bar plot
- Histogram

What would I do...

Prof Booknose has a massive bookshelf in his study.

What chart would you choose to show:

Book Genre Vs. Reading Time

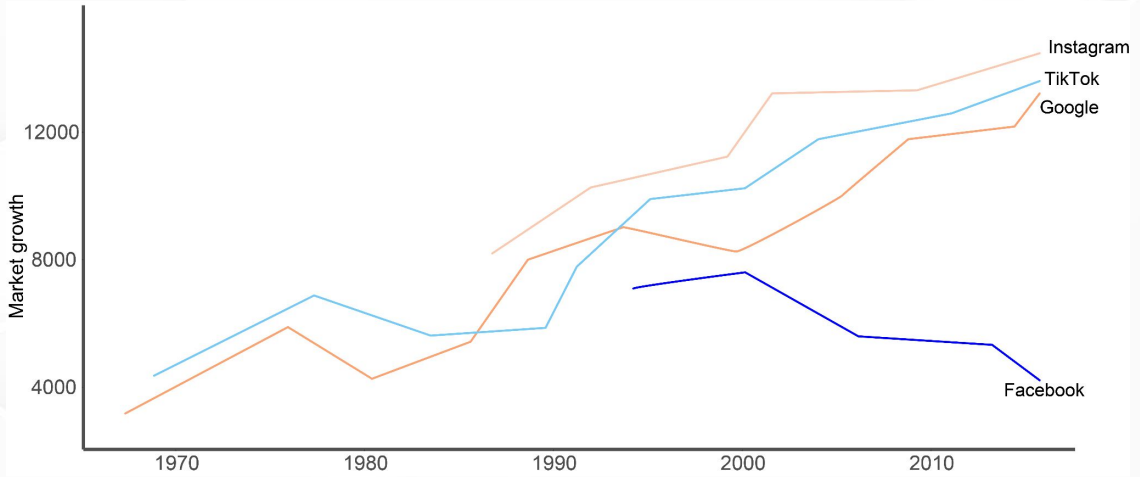
✅ box plot

✅ violin plot

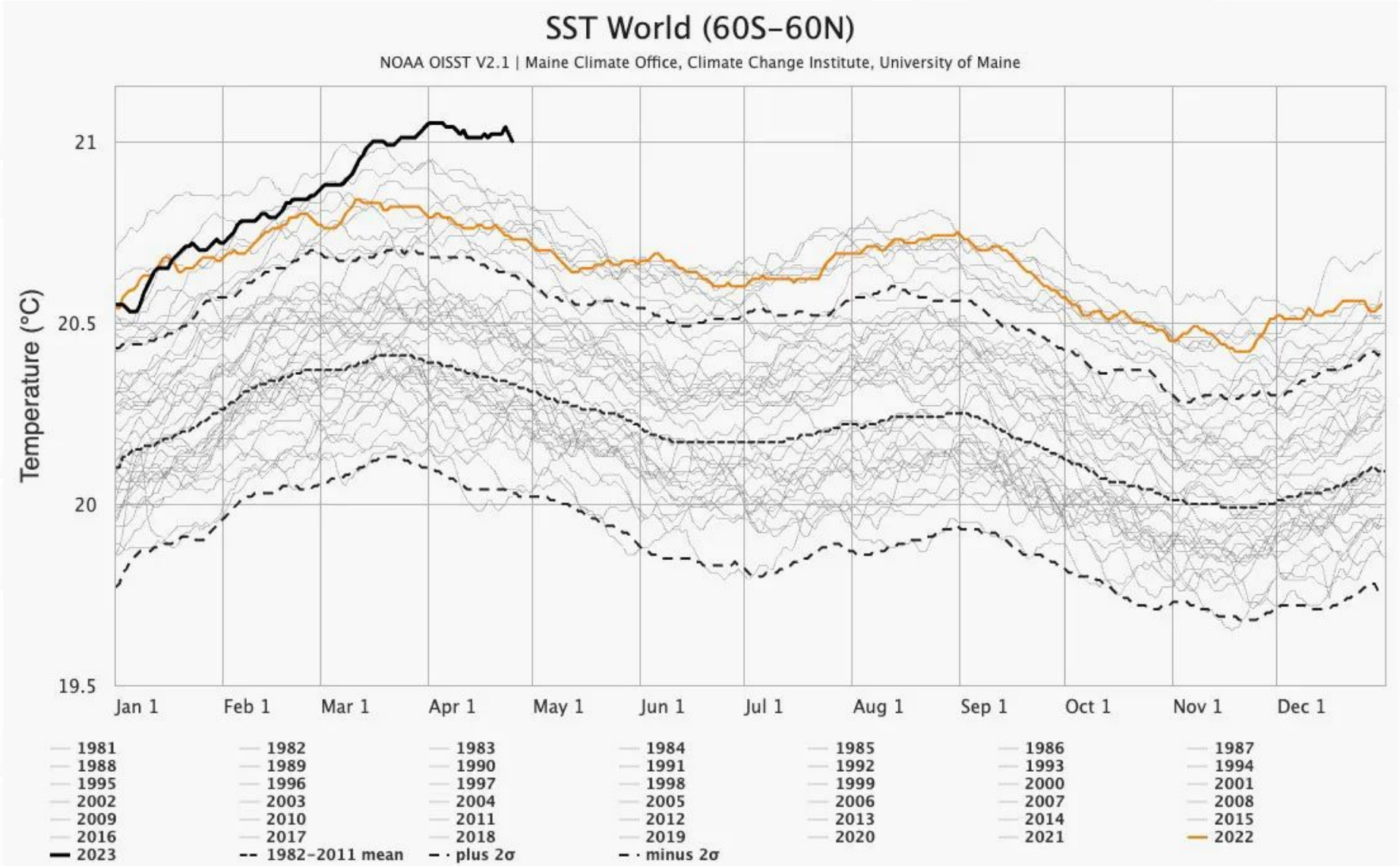
Data type: Time series

Time series: set of data points collected or recorded in a chronological order over a certain period of time

- Line graphs



Example: time series

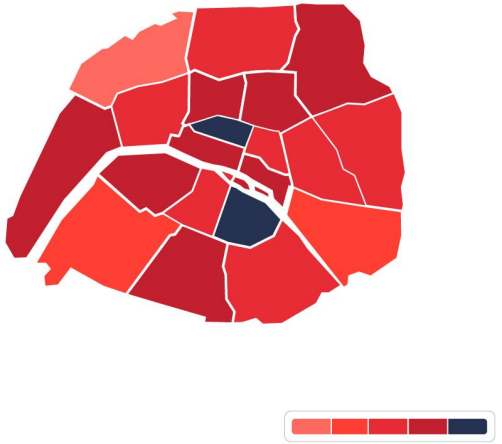


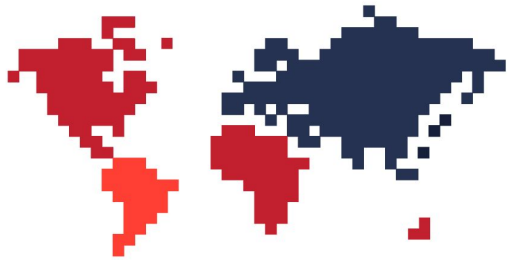
https://www.wired.com/story/an-ominous-heating-event-is-unfolding-in-the-oceans/

Geospatial

Data that is associated with a specific geographical location or physical space.

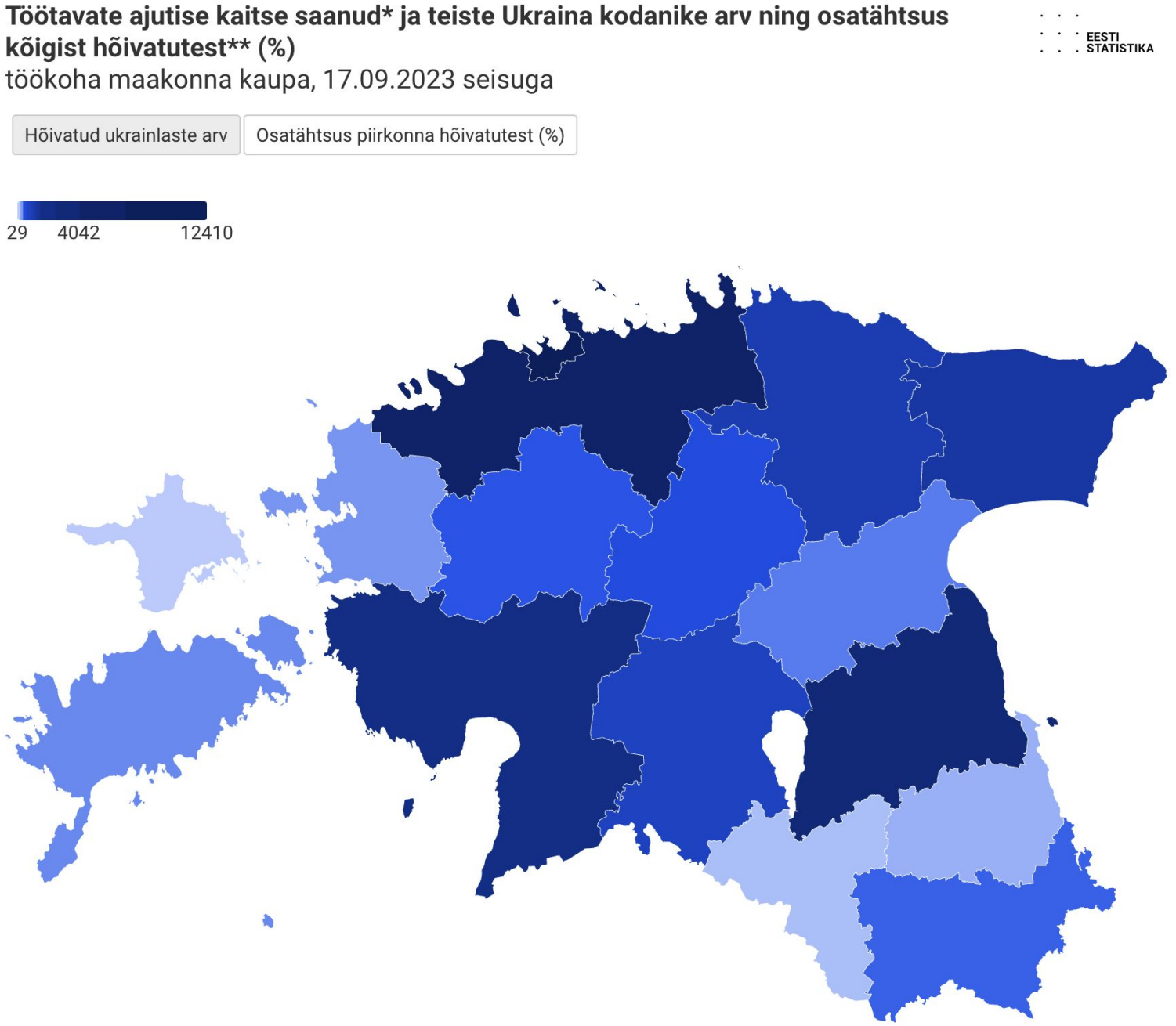
- Choropleth mapping
- Cartograms





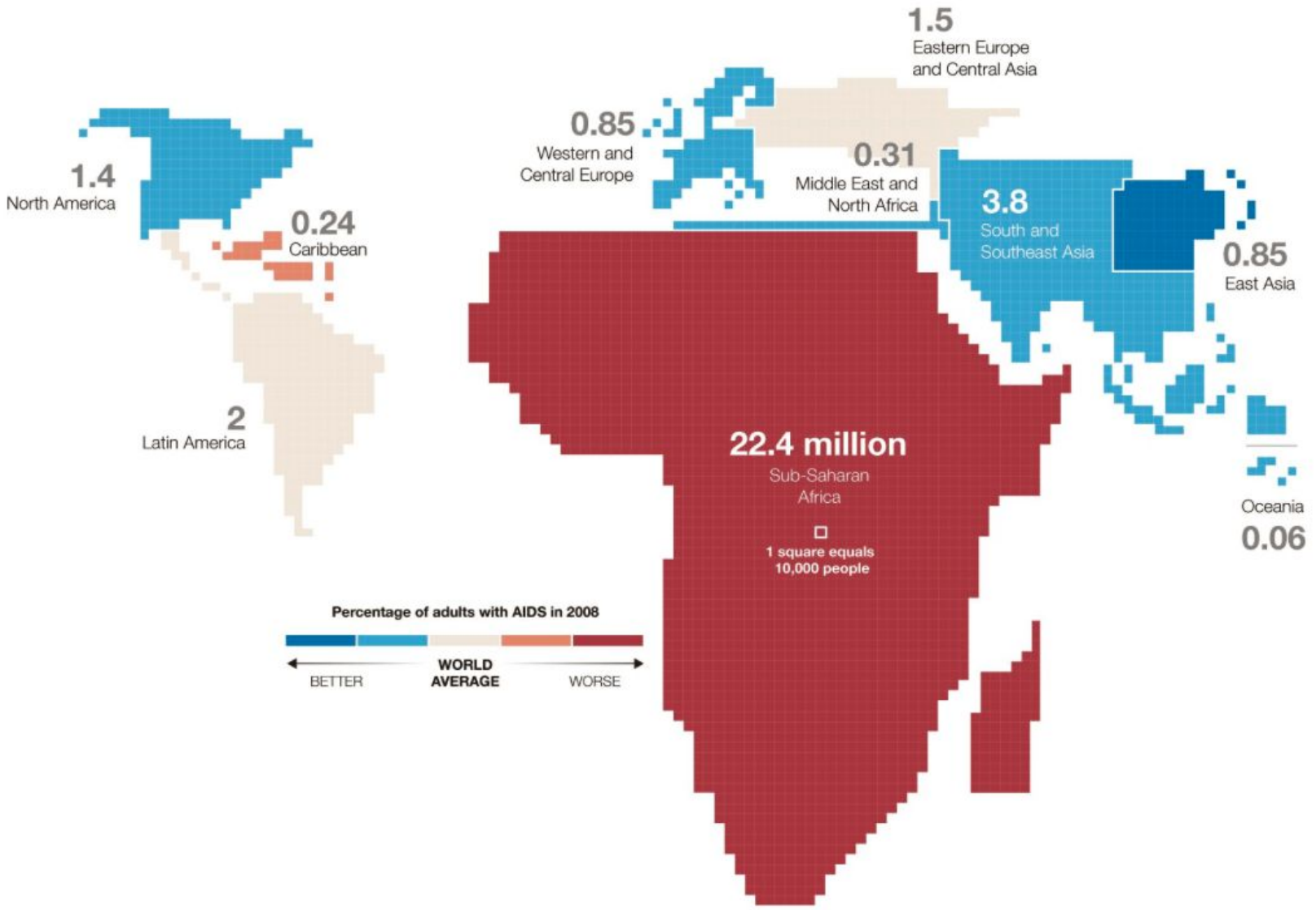
pictures from
https://datavizproject.com/data-type/cartogram/
https://datavizproject.com/data-type/choropleth-map-2/

Example: Choropleth



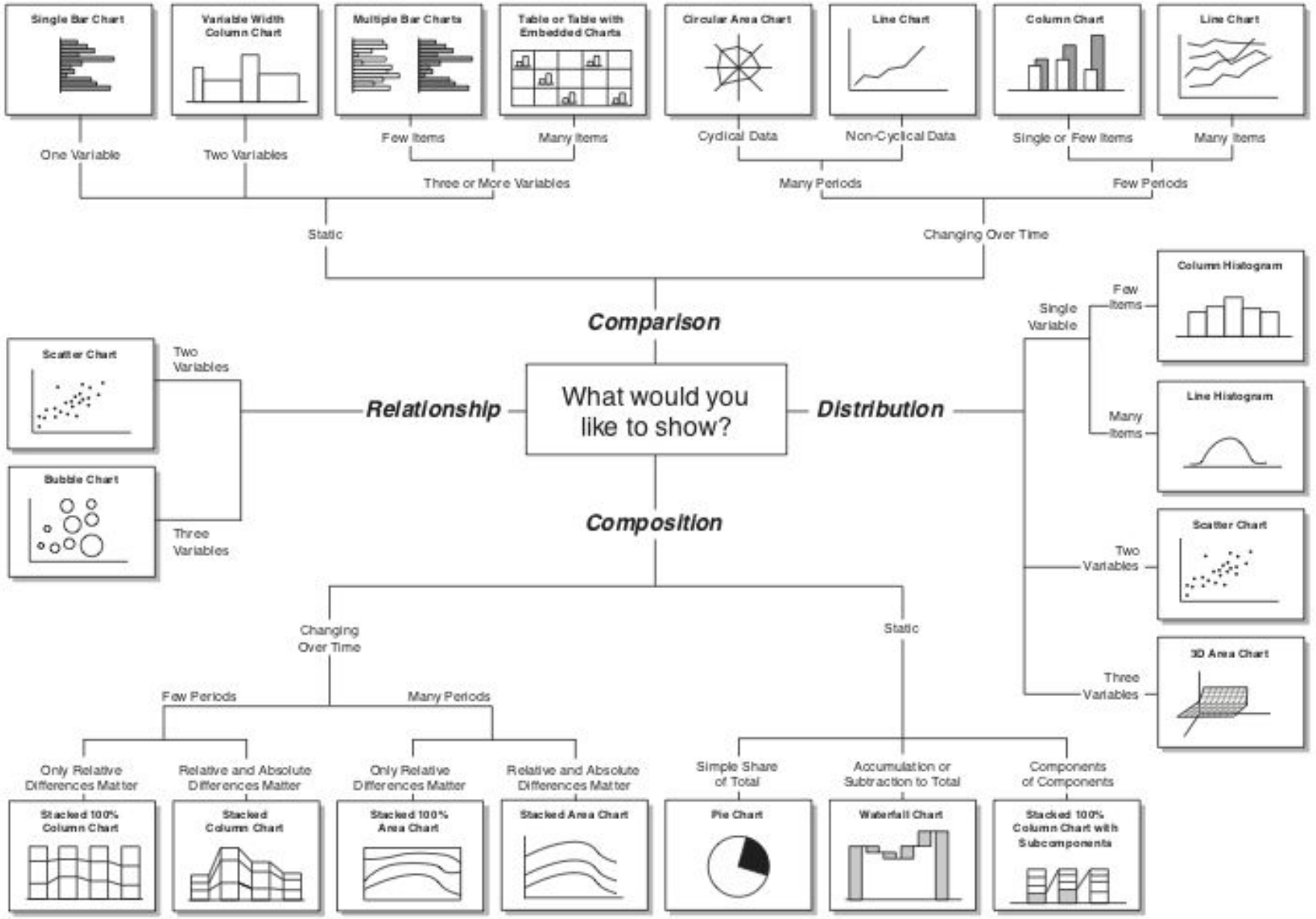
https://www.stat.ee/et/avasta-statistikat/kiirstatistika/ukrainlased-eesti-tooturul

Example: Cartogram



https://contrarian.ca/2009/12/08/a-world-map-of-aids/

Chart Chooser

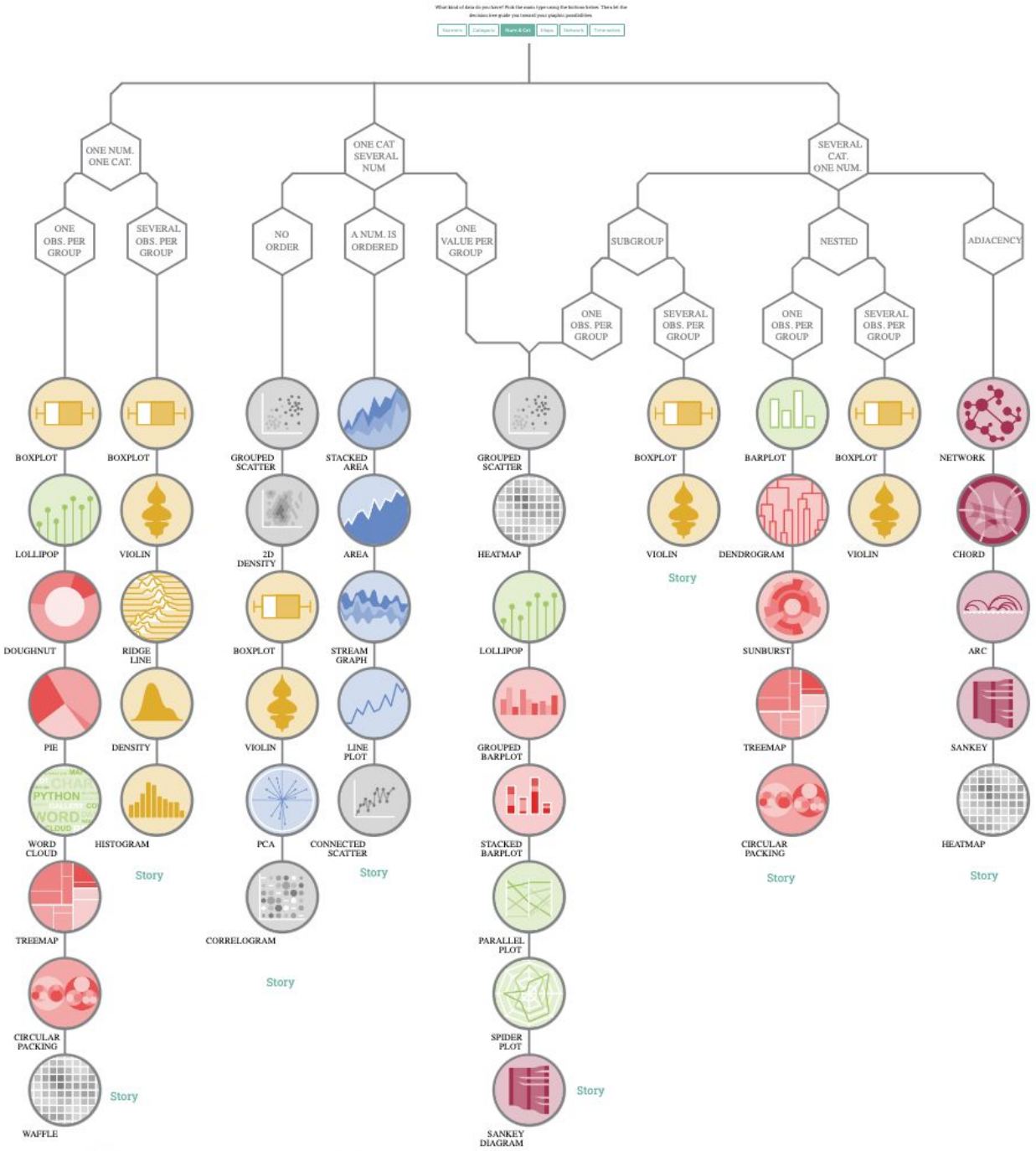


https://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

Resource

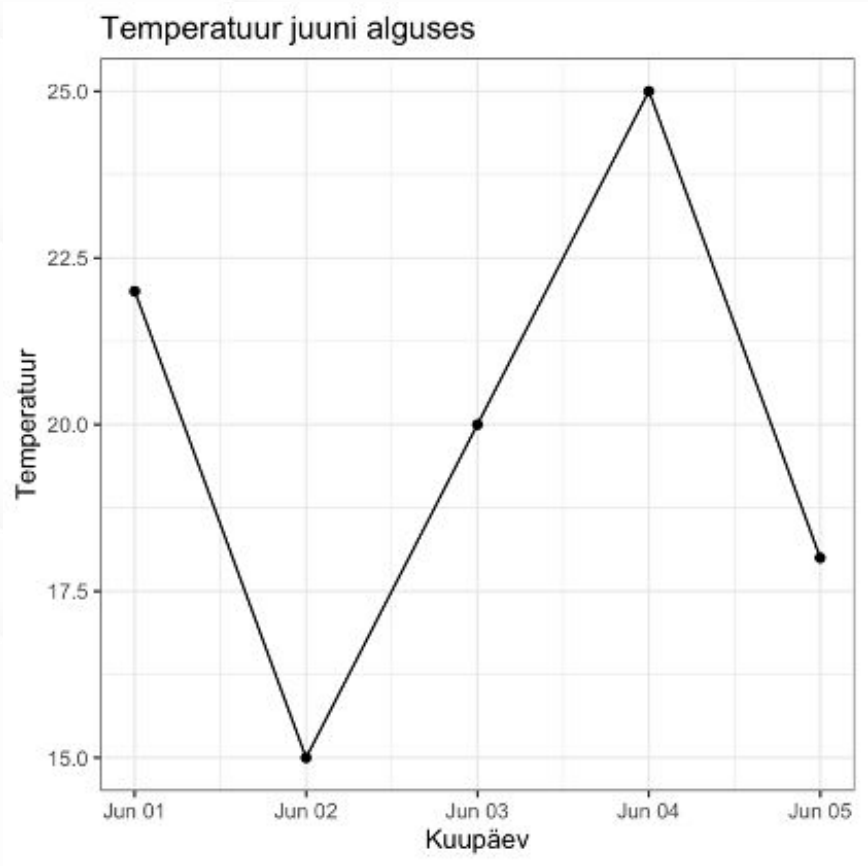
What kind of data do you have? Pick the main type using the buttons below. Then let the decision tree guide you toward your graphic possibilities.

Numeric | Categoric | Num & Cat | Maps | Network | Time series

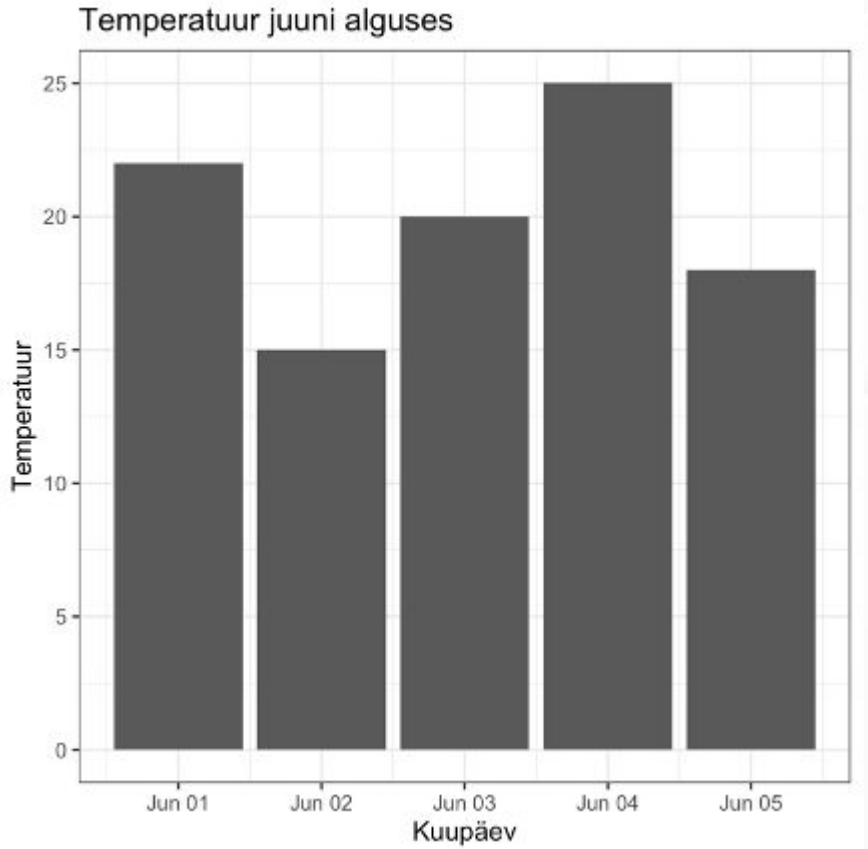


https://www.data-to-viz.com/

Which graph would you use for temperature change?

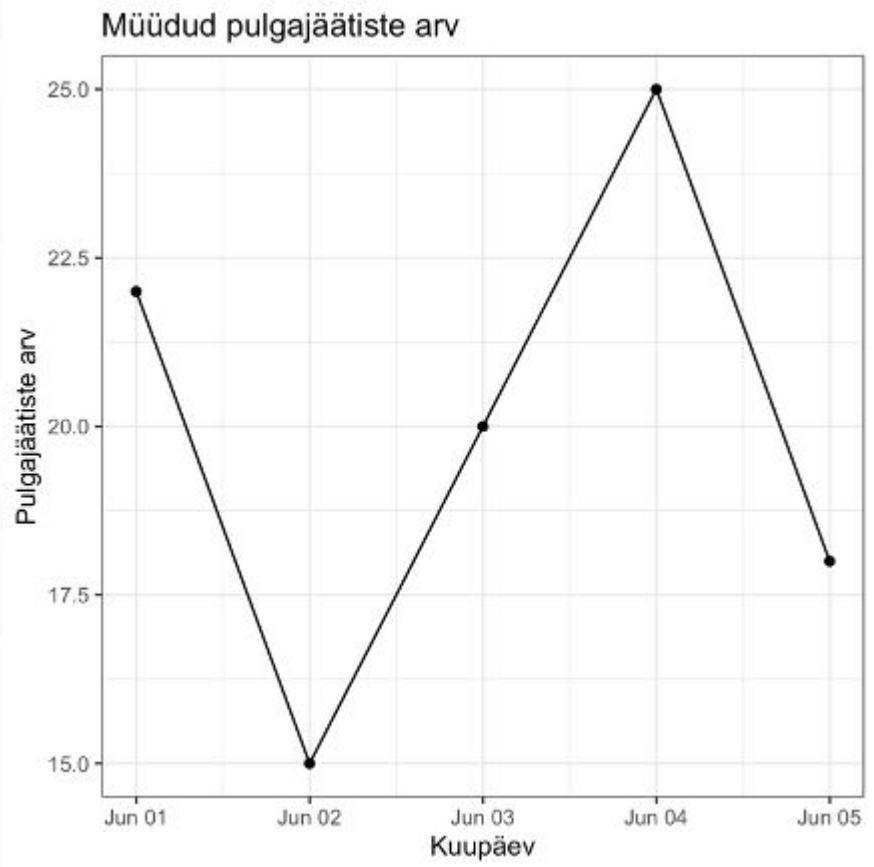


A

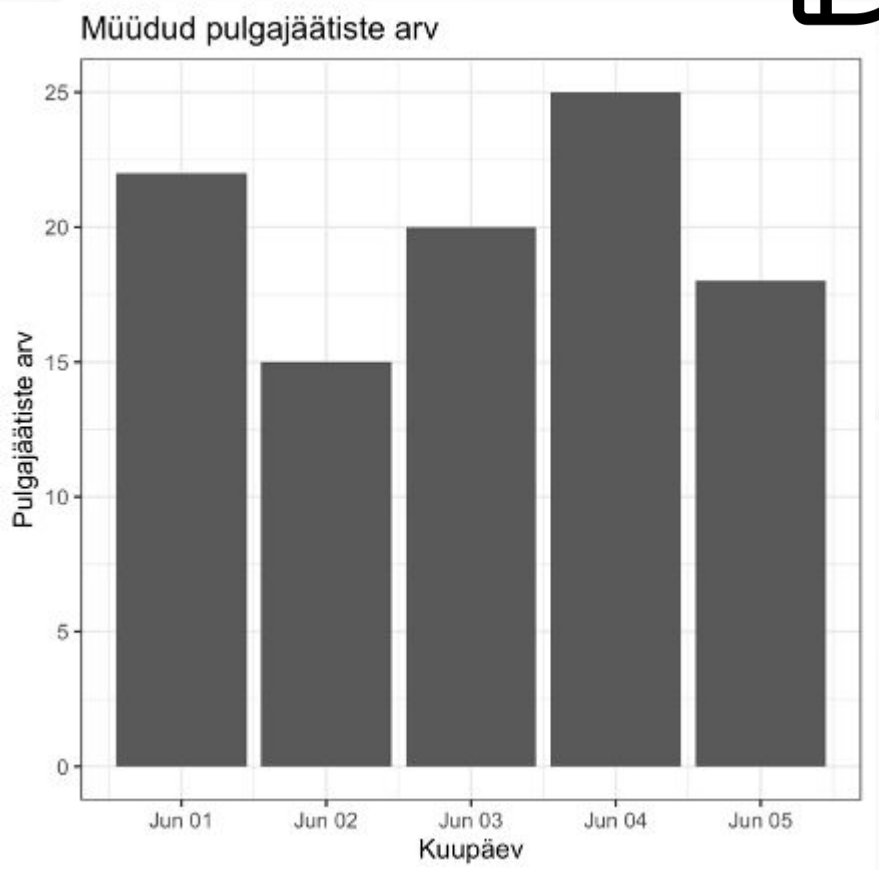


B

Which graph would you use for sold ice cream count?

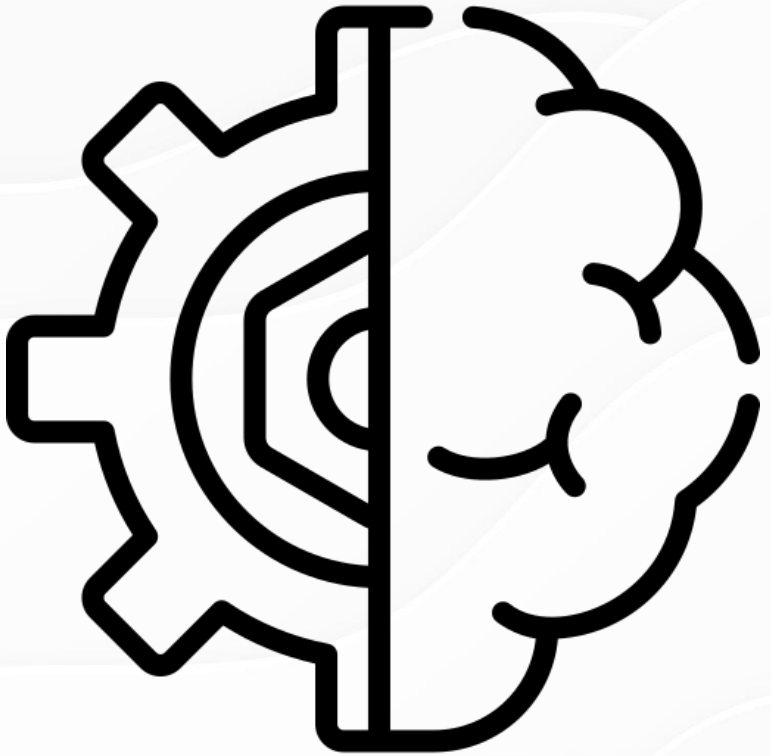


A



B

Practice



https://100.datavizproject.com/

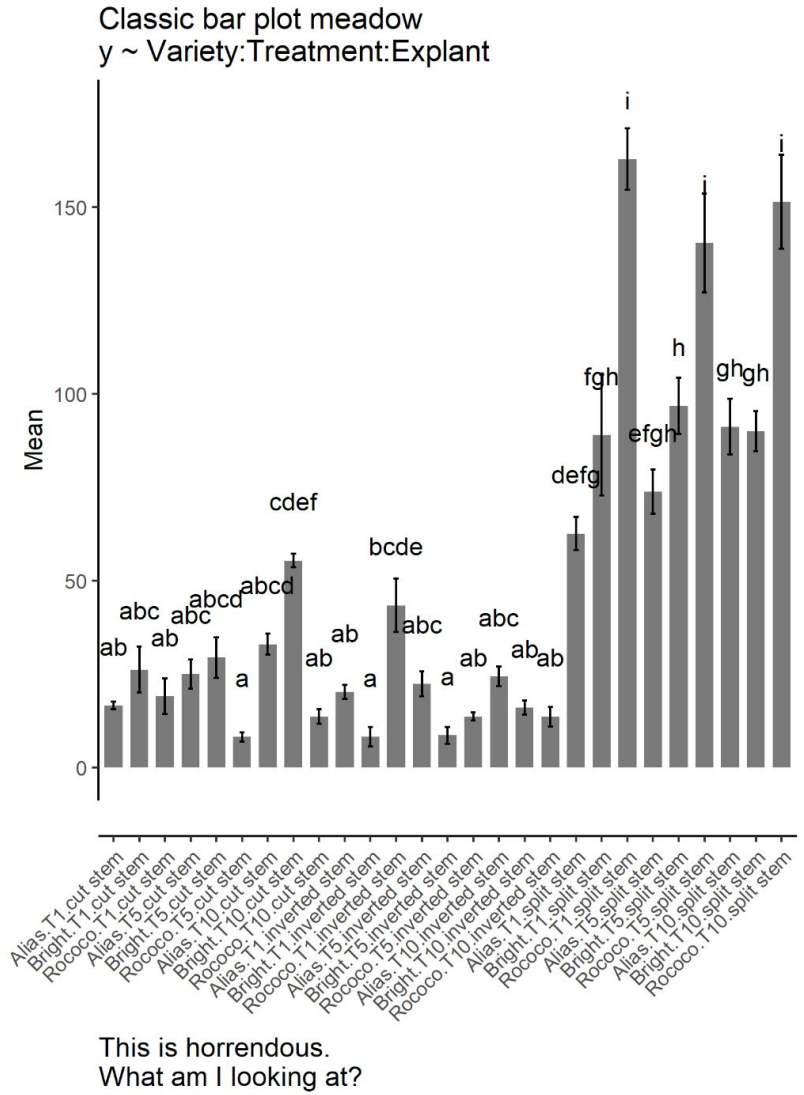
Same data, 100 data visualisations

3 questions:

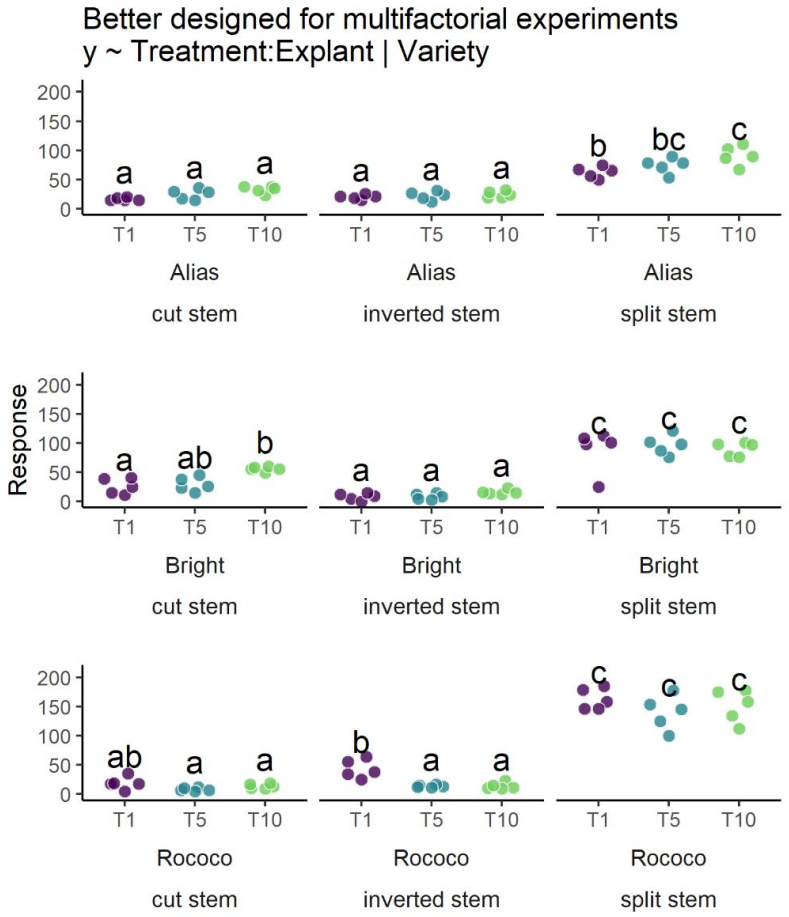
1. Most informative graph
2. most confusing graph
3. visually striking and memorable graph

Common mistakes and best practices

Common mistake: Bar plot meadow



This is horrendous.
What am I looking at?

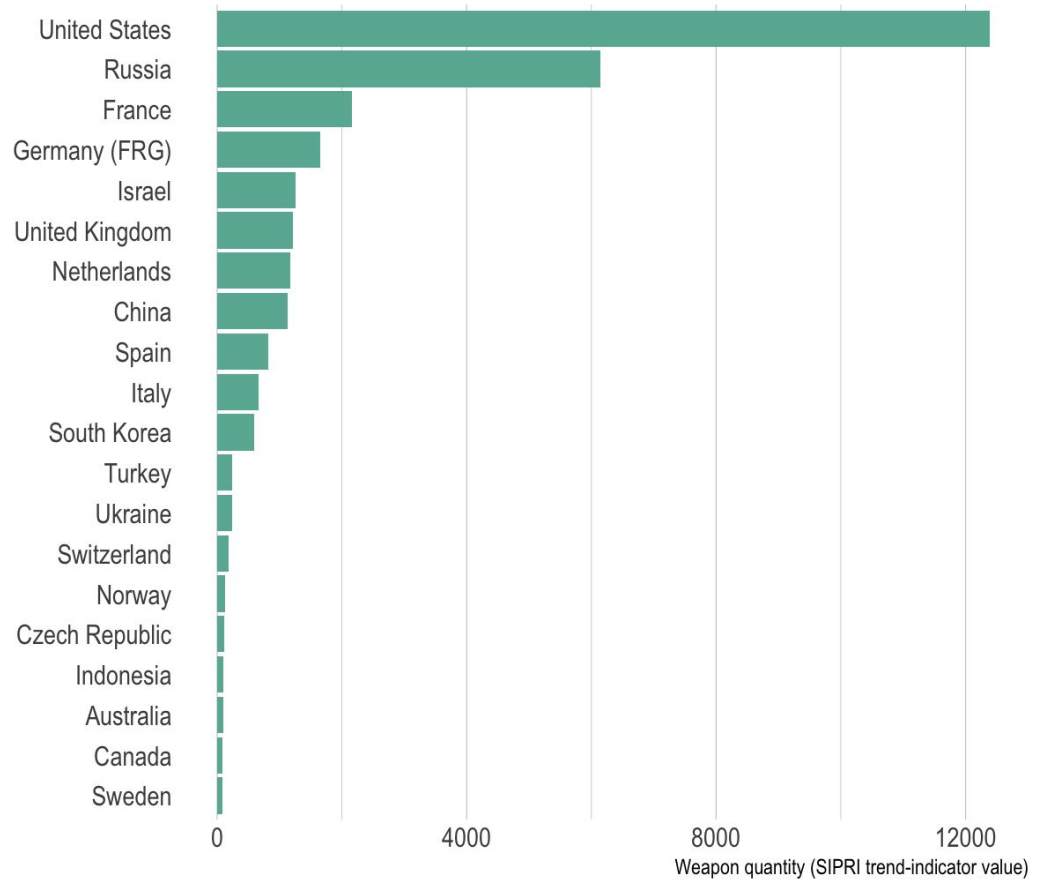


That's better.
Reader's attention is more focused.

https://github.com/cxli233/FriendsDontLetFriends
C. Li. (2023). cxli233/FriendsDontLetFriends: FriendsDontLetFriends (v3) (Version v3). Zenodo. https://doi.org/10.5281/zenodo.7097522

Best practice: Bar chart

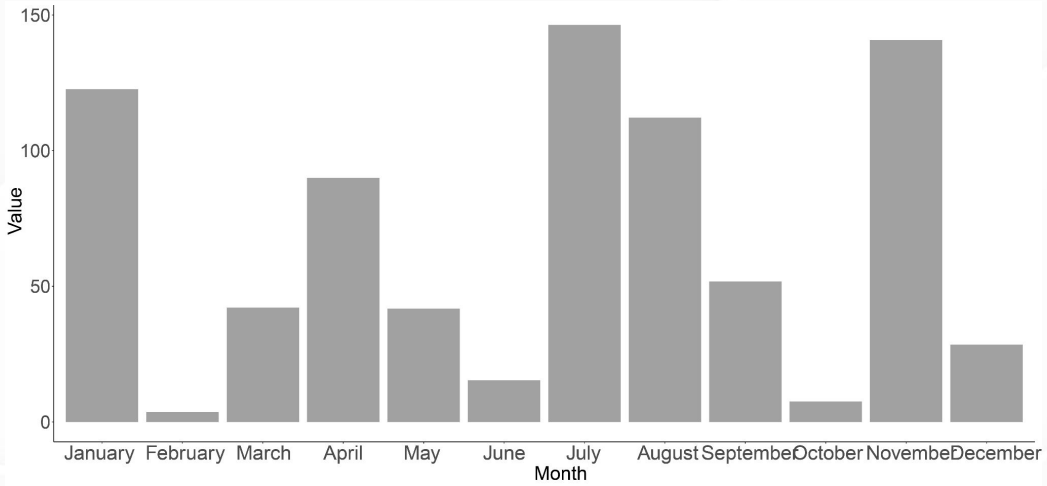
- **Values starts from 0**
- Horizontal VS vertical bar chart
  - 5-7 columns - vertical bar chart
  - Over 7 columns - horizontal bar chart
- Order the columns if possible

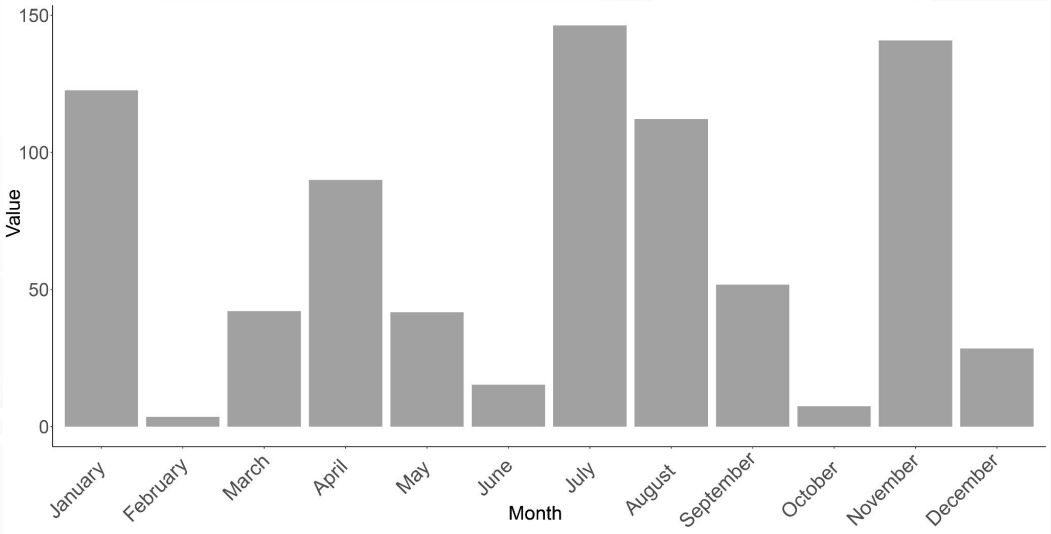


Common mistakes: Bar chart

Problem: Too many letters in labels. Each bar takes up a lot of horizontal space

Solution: Rotate the labels.

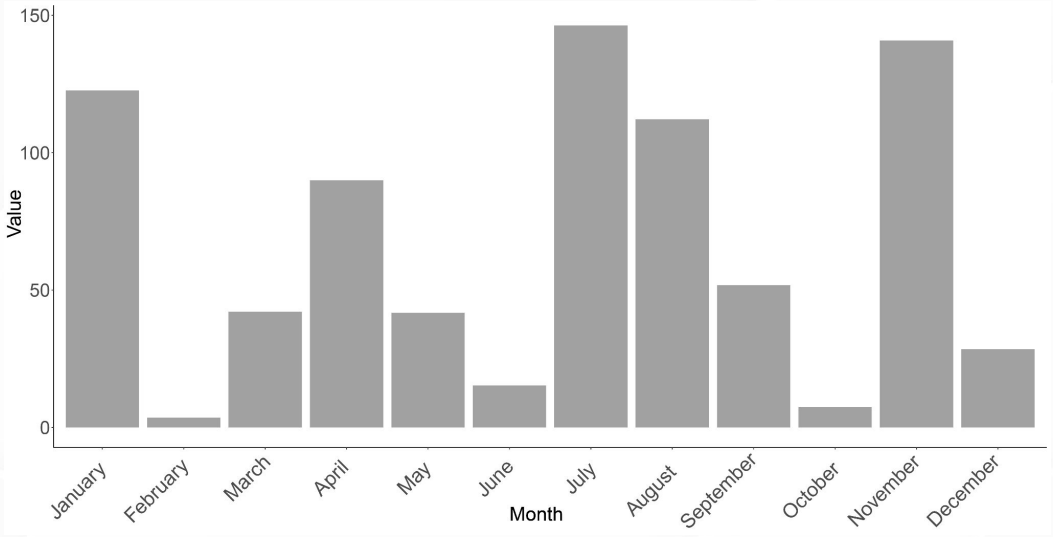




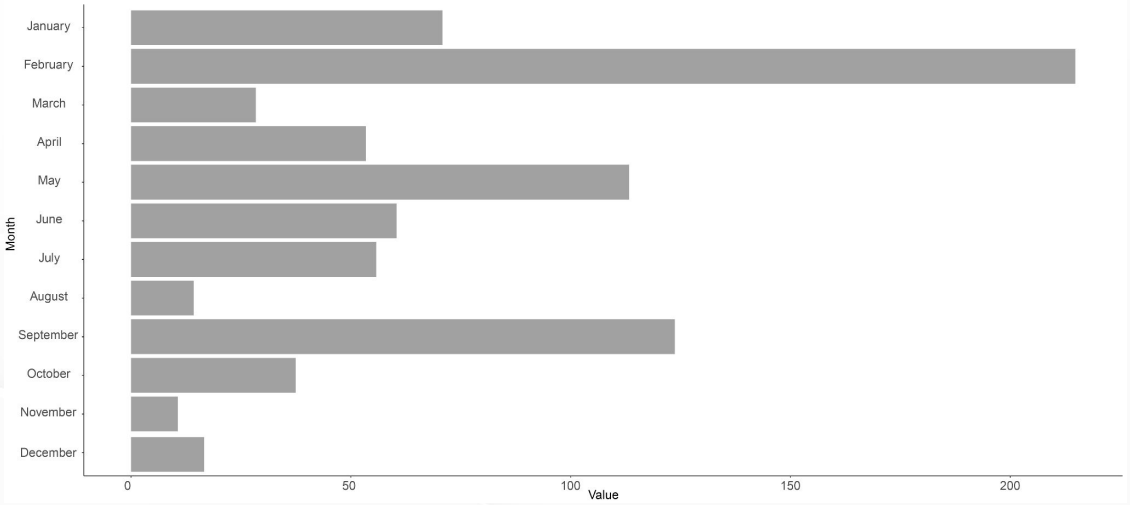
Common mistakes: Bar chart

Problem: Rotated labels are awkward and difficult to read

Solution: Swap for horizontal bars.



Maximum 5 variables in vertical bar charts

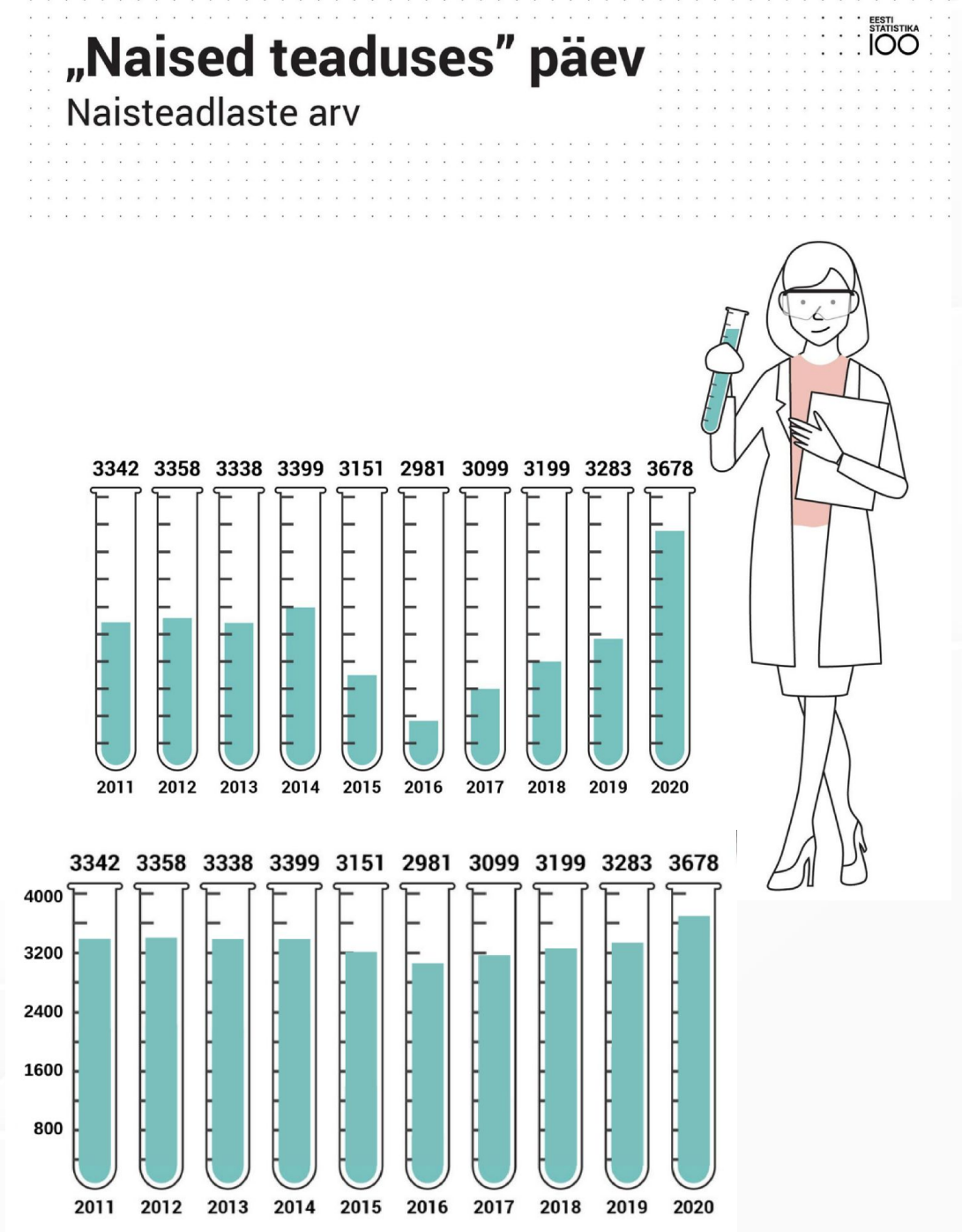


Common mistakes: Bar chart

Problem: scale doesn't start from 0. Results are misleading

Solution: either make it start from 0 or switch to a different graph.

Example:
https://www.canva.com/design/DAFlx7uruoU/khBHFqkeehI9olmGygtN6Q/watch?utm_content=DAFlx7uruoU&utm_campaign=designshare&utm_medium=link&utm_source=publishsharelink

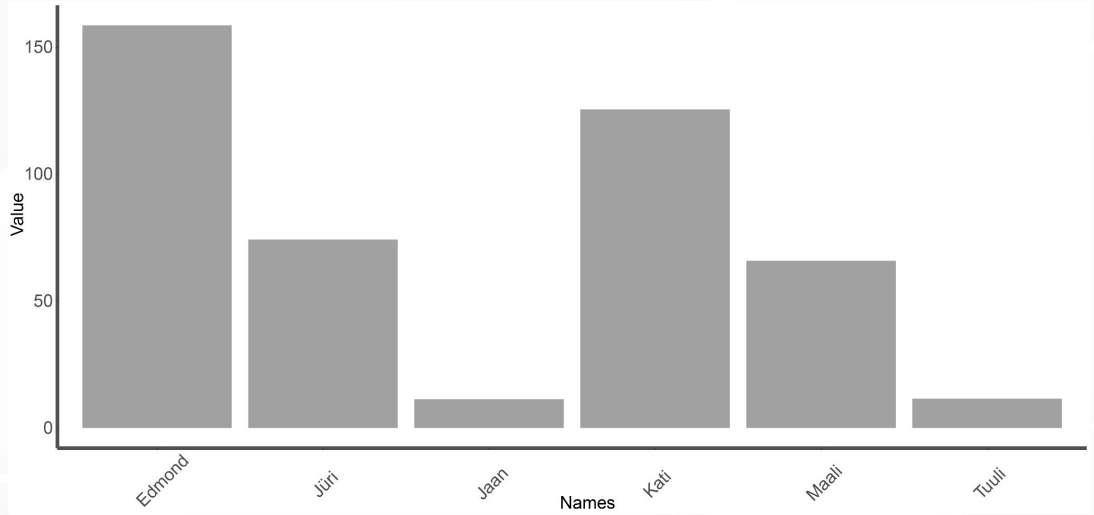


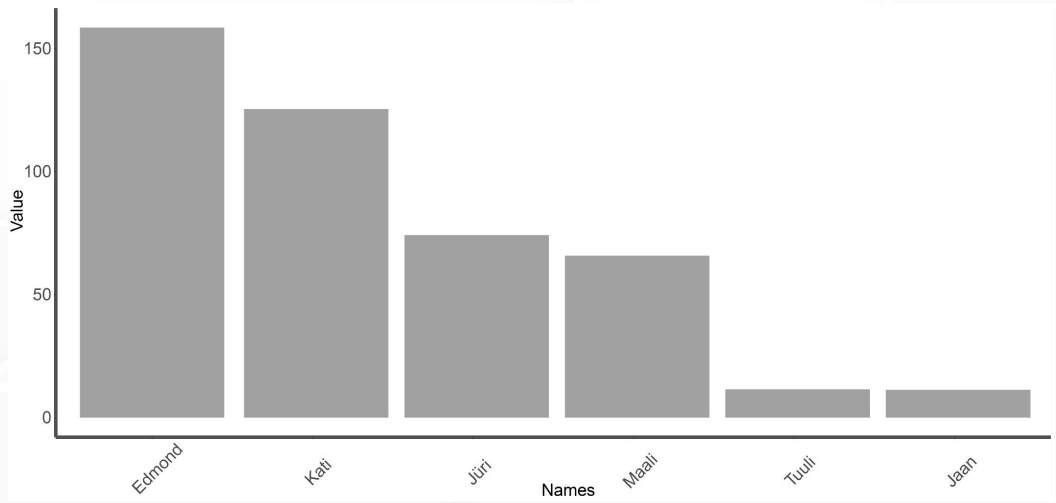
Common mistakes: Bar chart

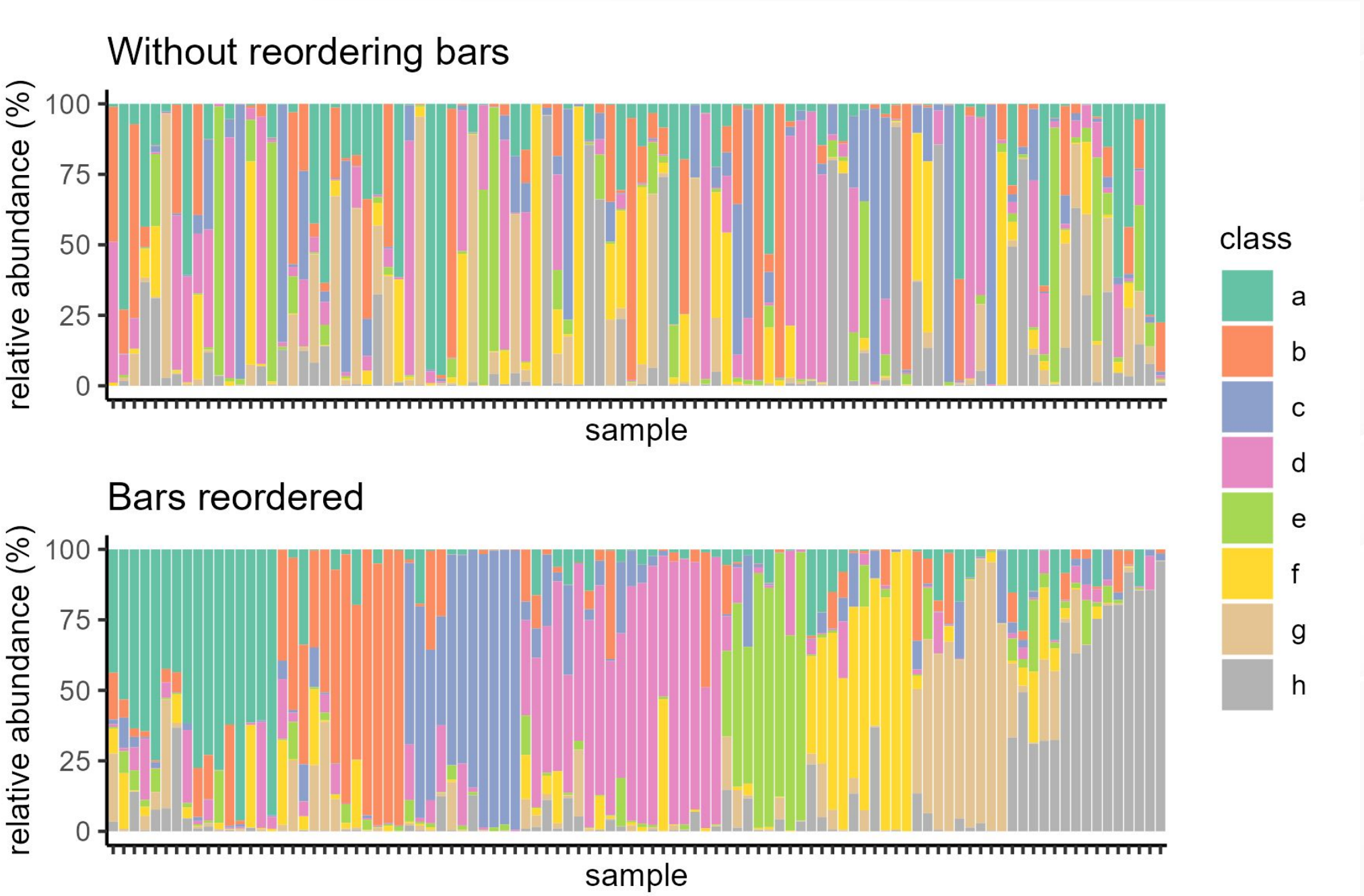
Problem: Order of the bars is messy (due to programs liking to order them by alphabet).

Solution: Arrange in order of size

NB! Be careful with data that has natural order! No size arranging! Example: Months in a year







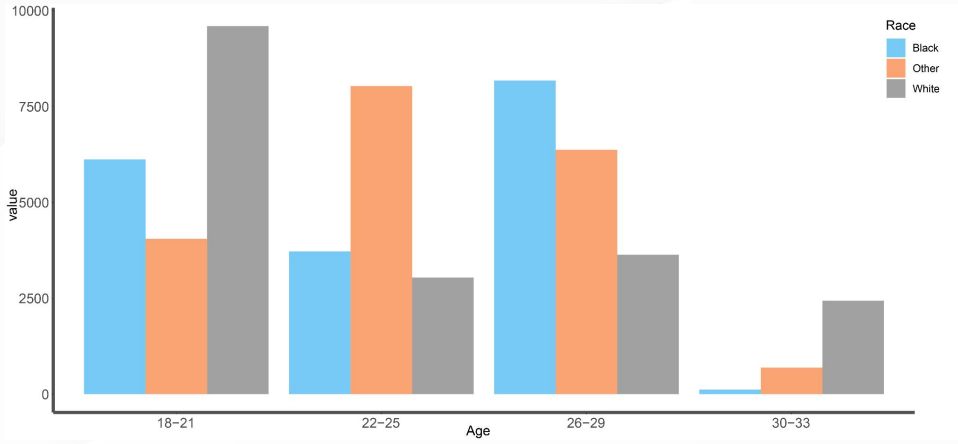
https://github.com/cxli233/FriendsDontLetFriends
C. Li. (2023). cxli233/FriendsDontLetFriends: FriendsDontLetFriends. Zenodo. https://doi.org/10.5281/zenodo.7097522

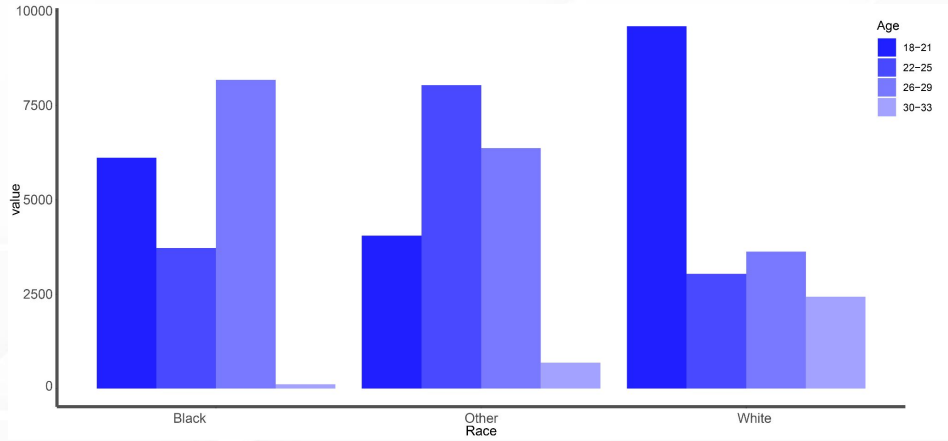
Common mistakes: Bar chart

Problem: Adding another variable might complicate the readability of the plot

Solution: switch which variable is shown with color.

Solution: just make separate graphs

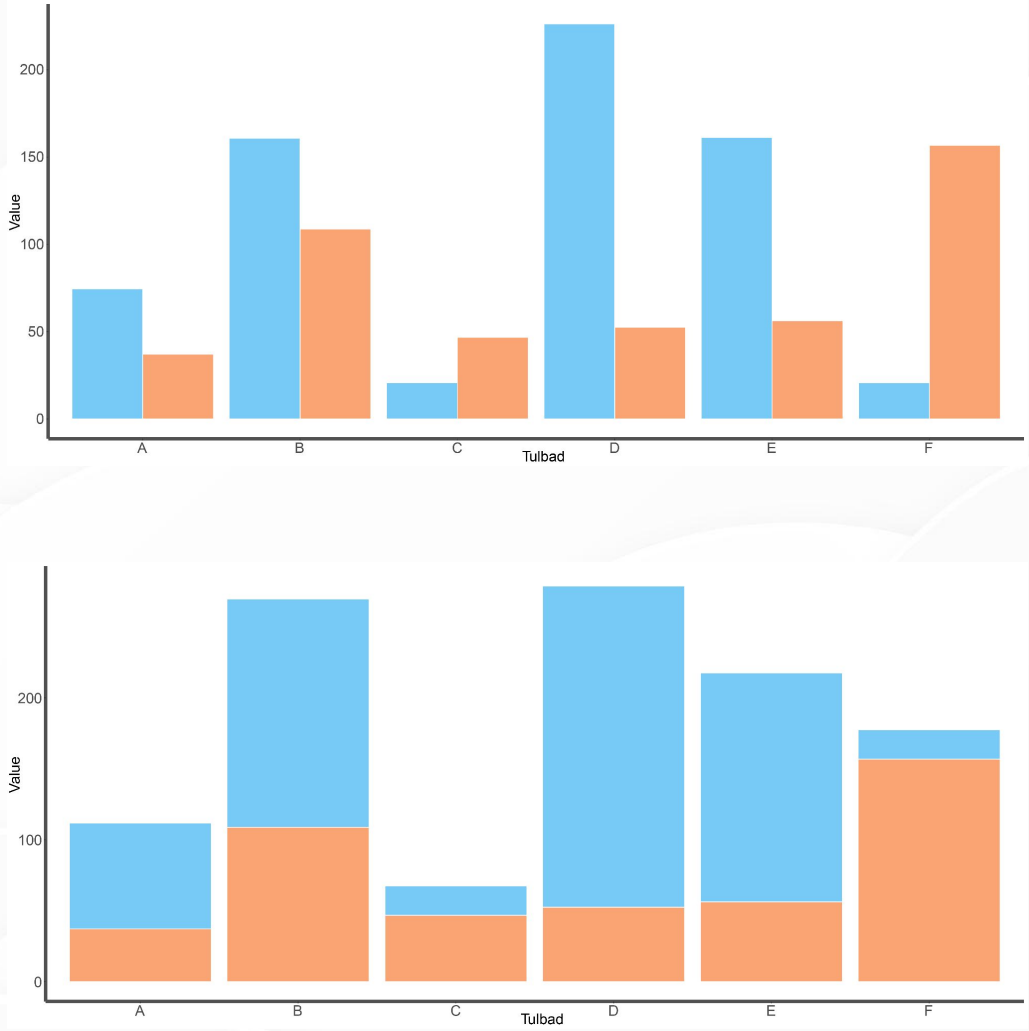




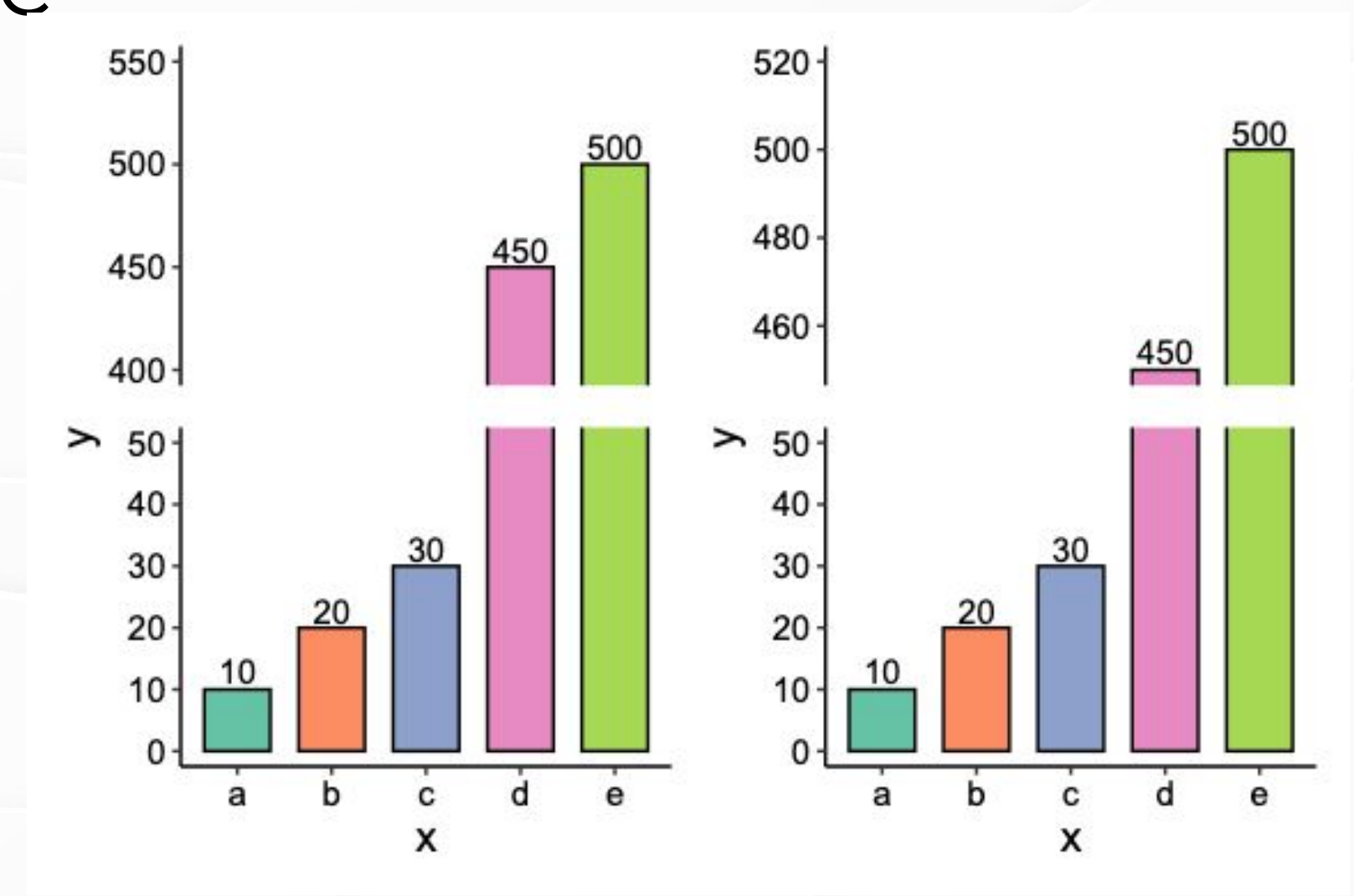
Common mistakes: Bar chart

Problem: Adding another variable might complicate the readability of the plot

Solution: Use stacked bar chart
(sum of the amounts represented by the individual stacked bars is in itself a meaningful amount,when the individual bars represent counts)

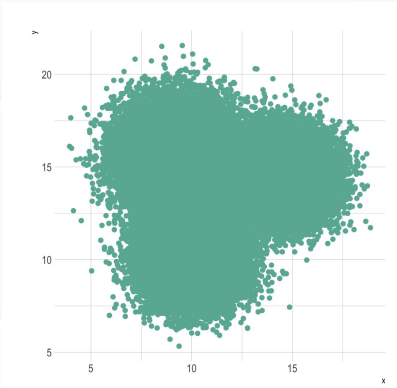


Common mistake: broken axis wrong place



https://github.com/cxli233/FriendsDontLetFriends
C. Li. (2023). cxli233/FriendsDontLetFriends: FriendsDontLetFriends. Zenodo. https://doi.org/10.5281/zenodo.7097522

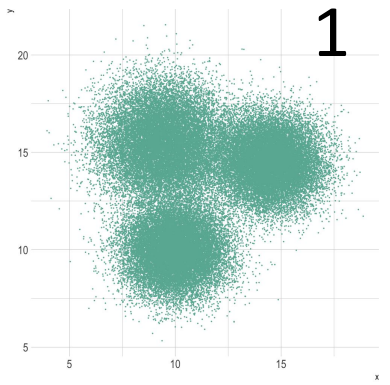
Common mistakes: Dot plot

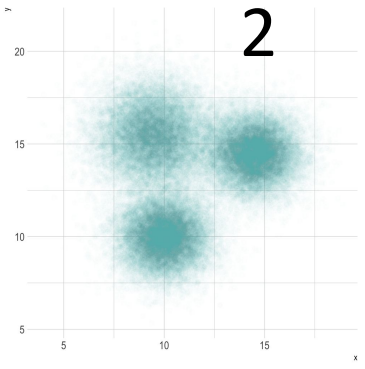


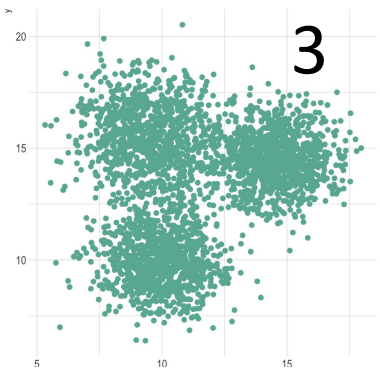
Problem: overplotting

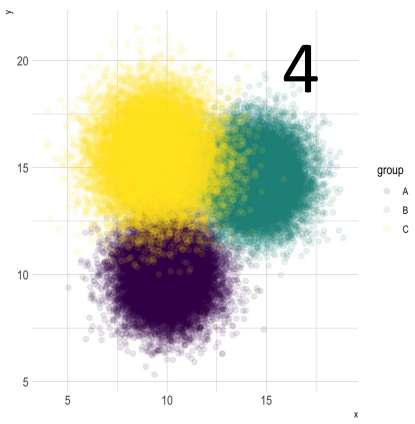
Solutions:

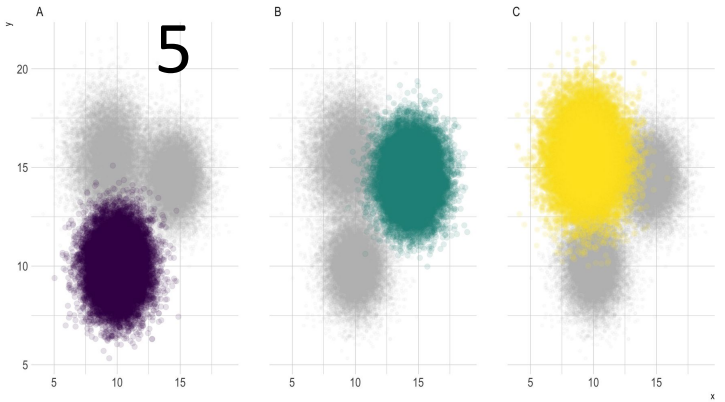
1. Decrease dot size
2. Transparency
3. Plot a fraction of your data
4. Grouping
5. Faceting
6. Jittering

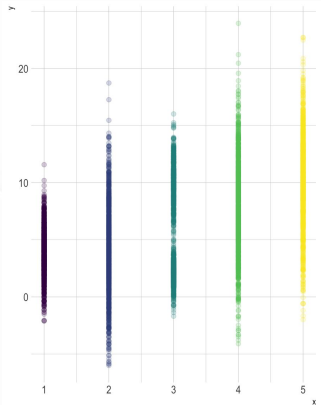


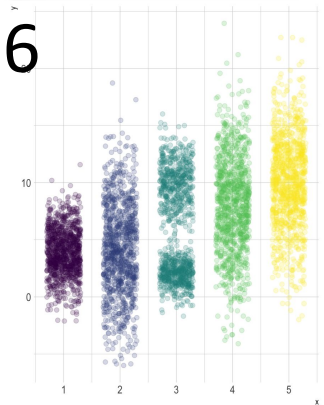




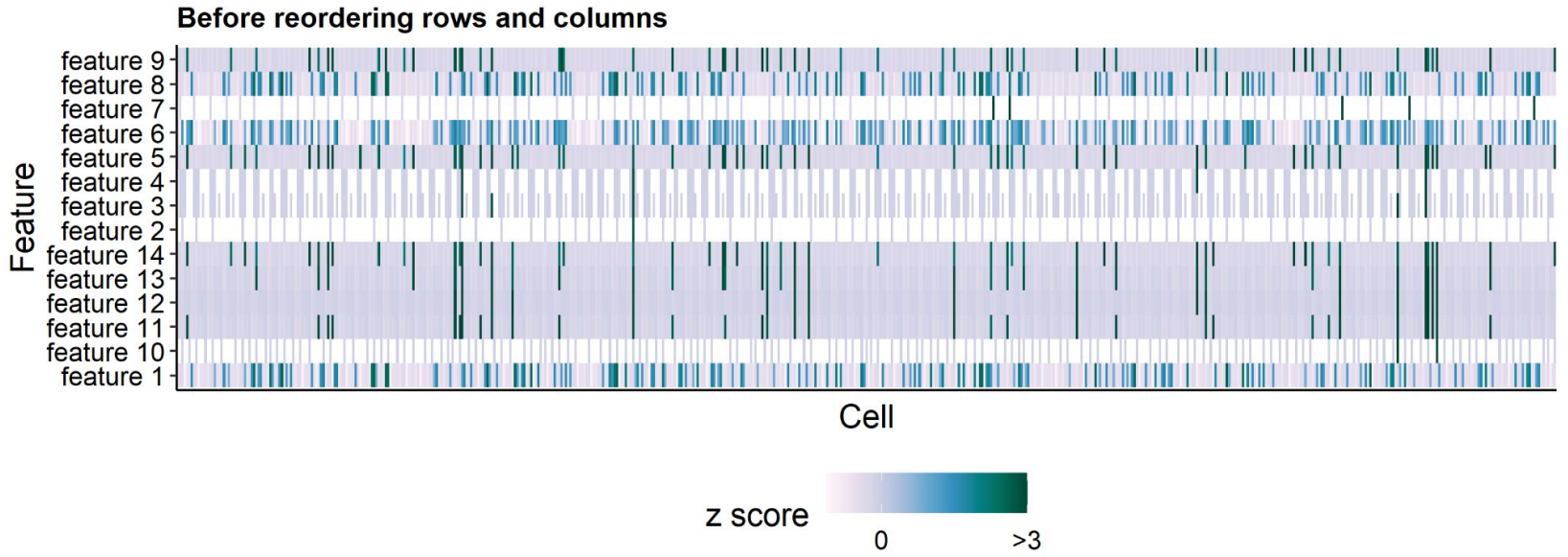




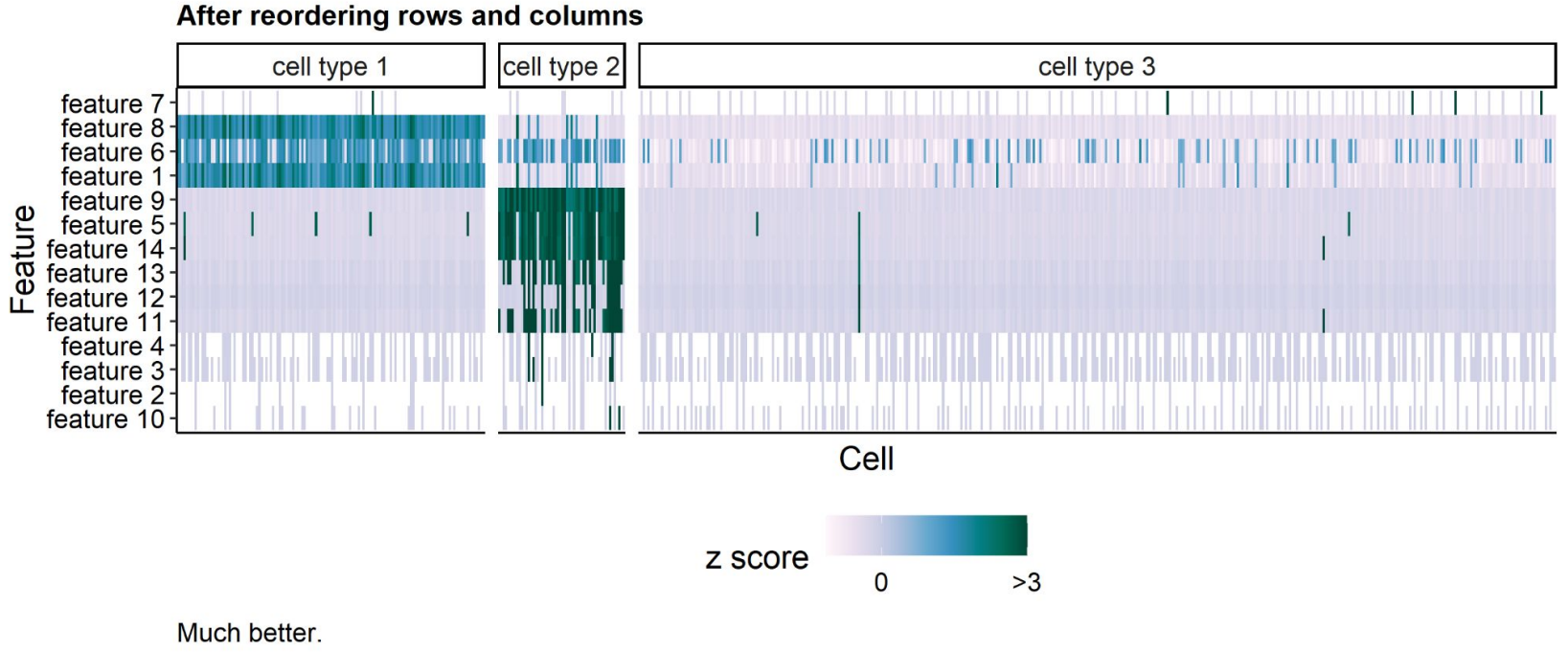




Common mistake: Heatmap without reordering



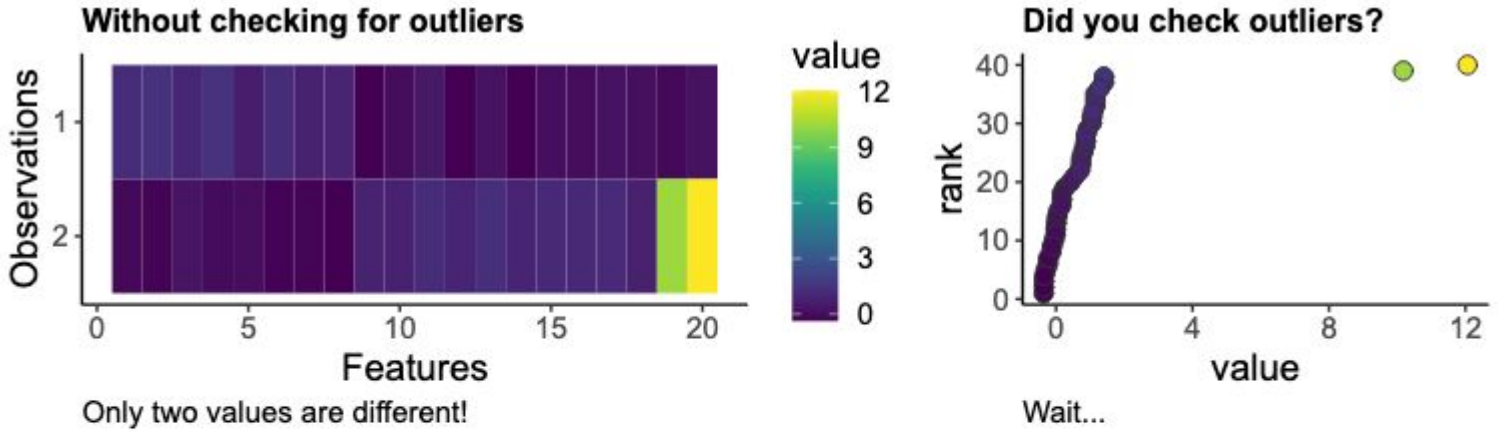
Am I looking at a glitching TV?



Much better.

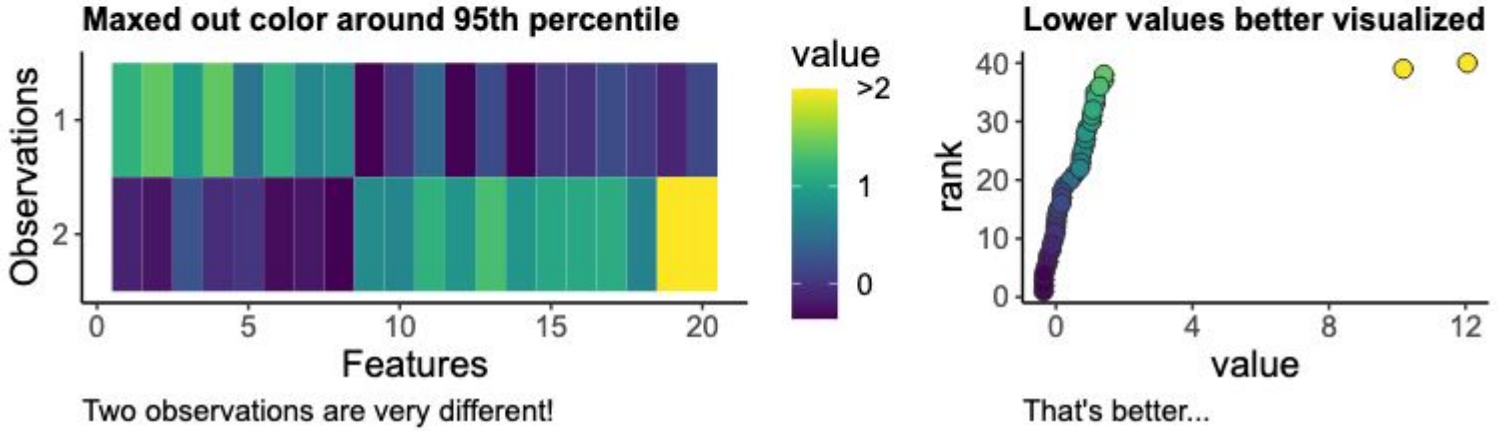
https://github.com/cxli233/FriendsDontLetFriends
C. Li. (2023). cxli233/FriendsDontLetFriends: FriendsDontLetFriends. Zenodo. https://doi.org/10.5281/zenodo.7097522

Common mistake: Heatmap and outliers



Only two values are different!

Wait...



Two observations are very different!

That's better...

https://github.com/cxli233/FriendsDontLetFriends
C. Li. (2023). cxli233/FriendsDontLetFriends: FriendsDontLetFriends. Zenodo. https://doi.org/10.5281/zenodo.7097522

Best practices: heatmap

- Normalize your data

- Use cluster analysis and thus permute the rows and the columns of the matrix to place similar values near each other according to the clustering

- Color palette is important

  choose one from here https://coolors.co/

Common mistake: confusing bar chart and histograms

A barchart shows the relationship between a numeric and a categoric variable.
Each entity of the categoric variable is represented as a bar.
The size of the bar represents its numeric value.

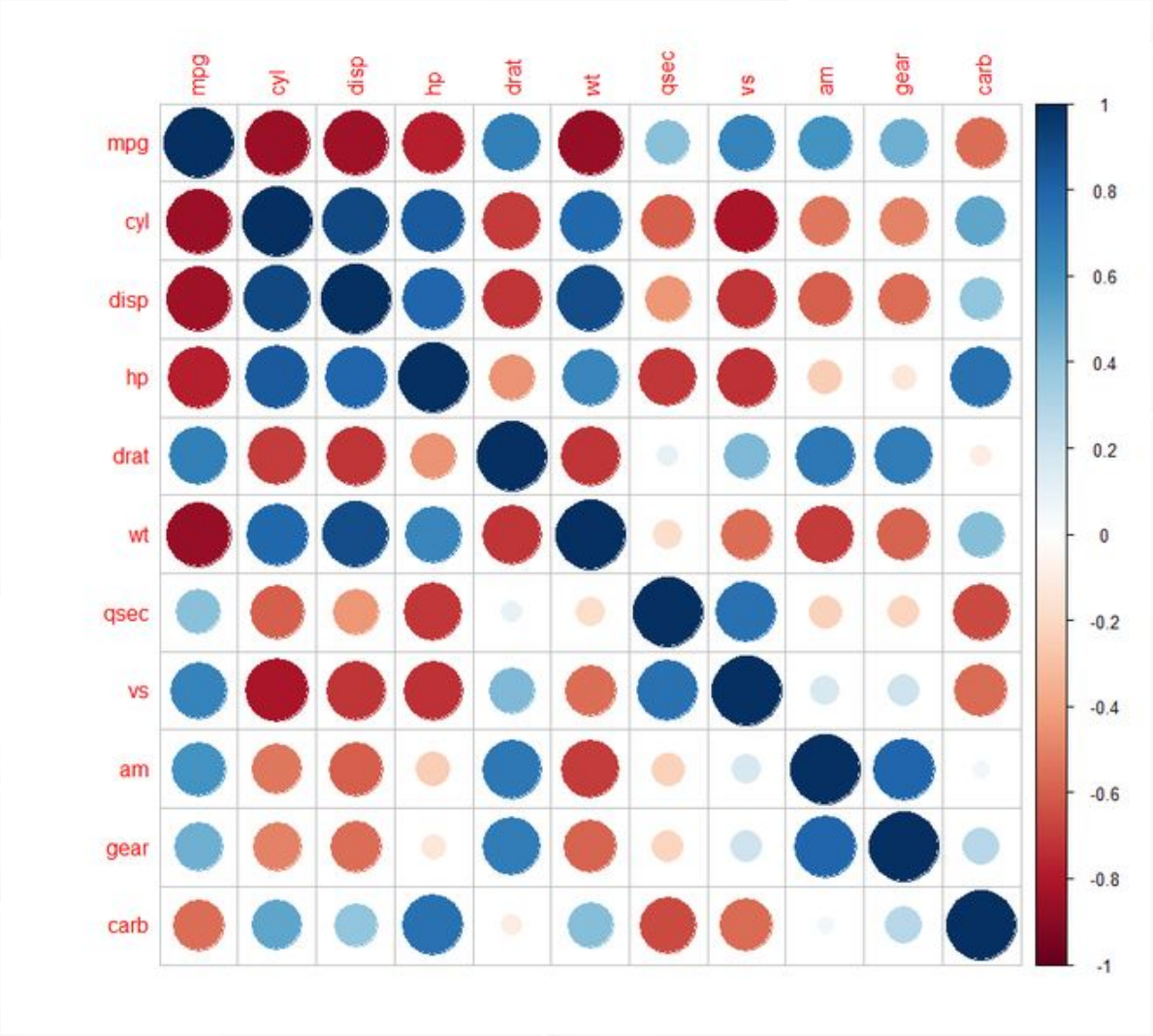
A histogram has only a numeric variable as input
and shows its distribution.

Best practice: correlogram

Try to have less than

10 variables

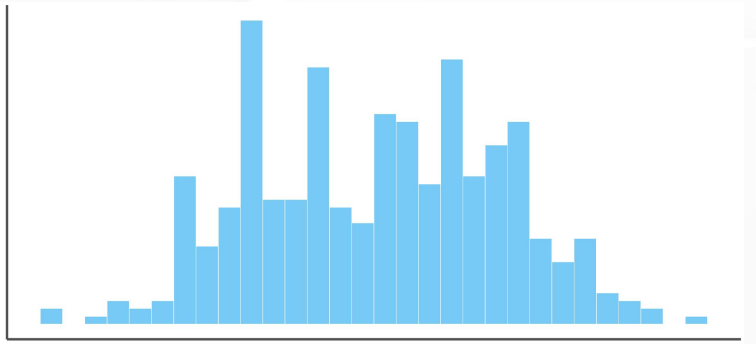
Displaying the relationship between more than ~10 variables makes the plot very hard to read

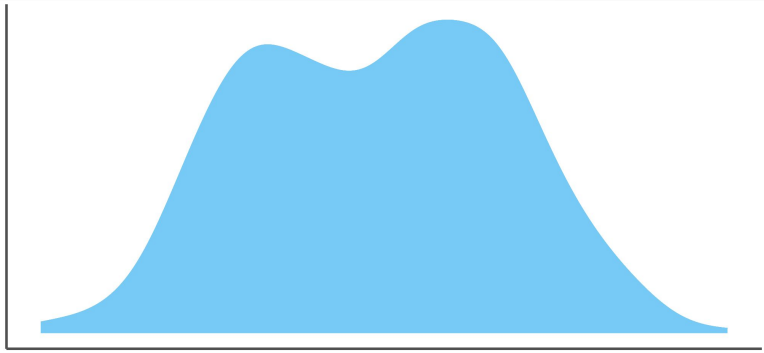


Best practices: Histograms and Density plots

- Values start from 0
- Choose an appropriate number of bins
- equal number of data points in each bin and no outliers; if not use box plot
- Choose the algorithm that fits your data
- DO NOT USE unequal bin widths
- Don’t compare more than ~3 groups in the same histogram.

    The graphic gets cluttered and difficult to understand. Instead use a violin plot, a boxplot, a ridgeline plot or use multiple small ones.

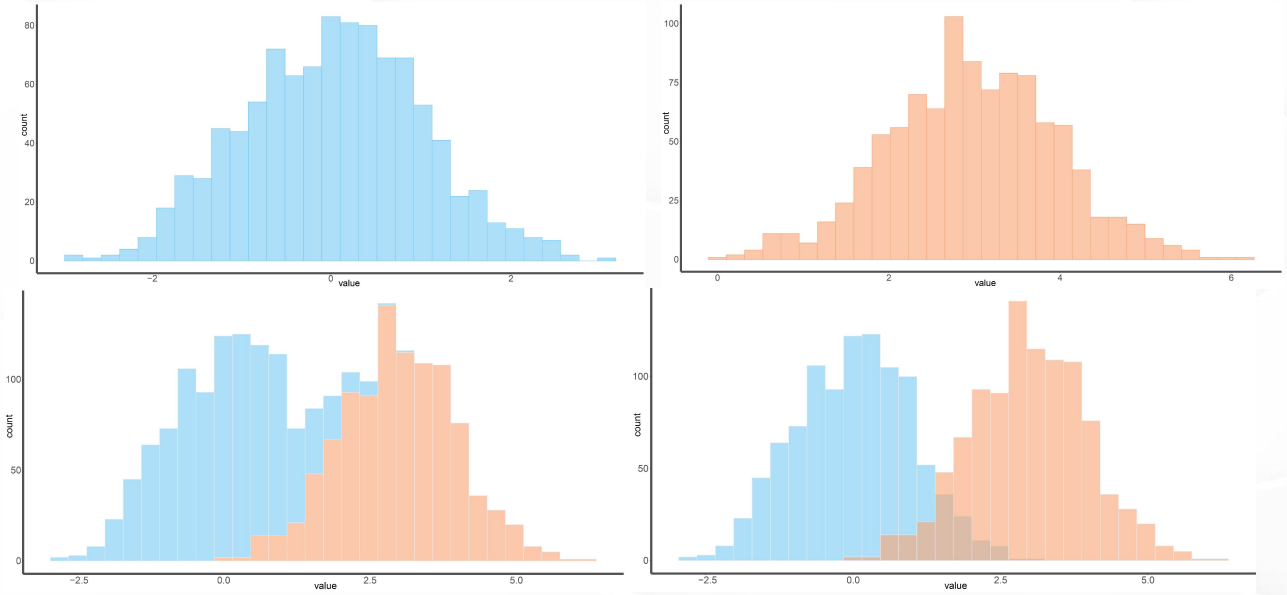


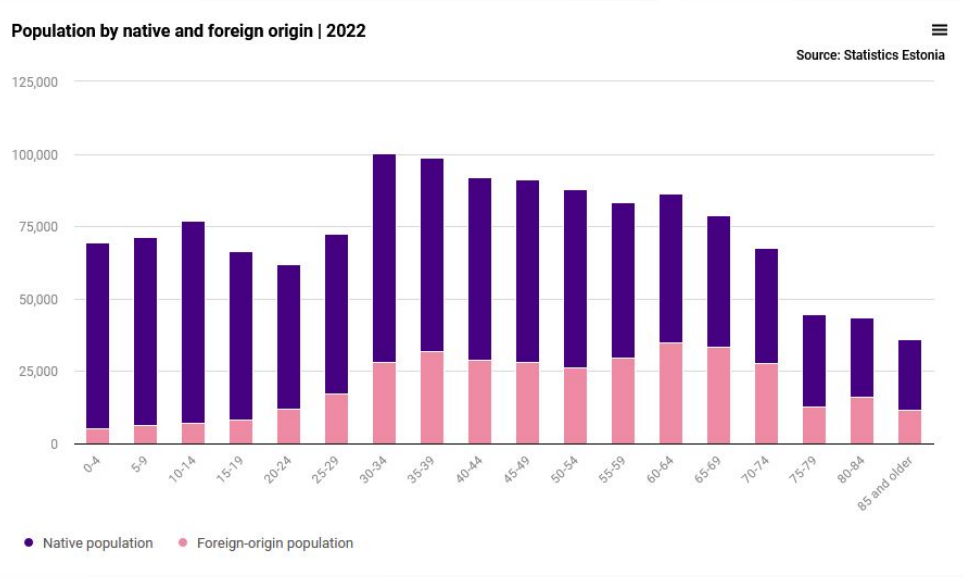


Common mistake: Stacked histogram

PROBLEM:

- Stacking - hard to read
- Partially transparent - appears three groups of data



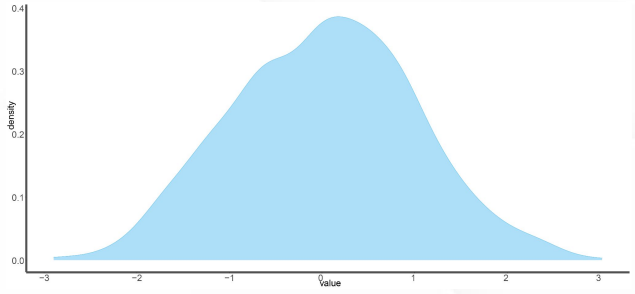


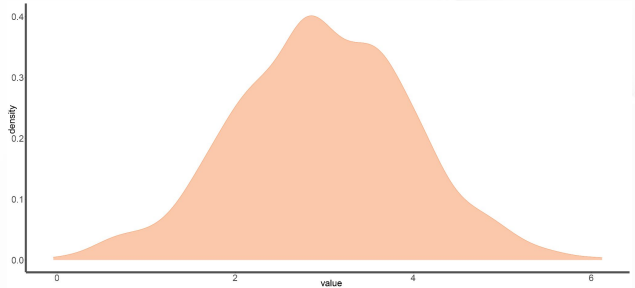
Common mistake: Stacked histogram

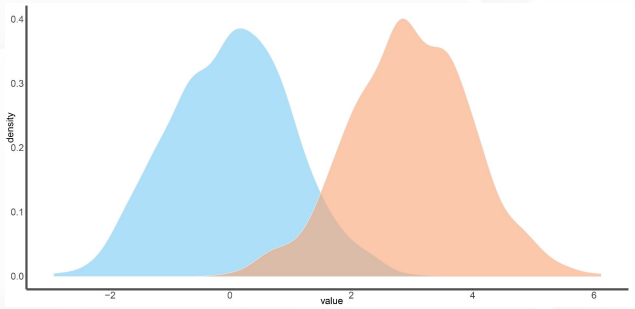
PROBLEM:

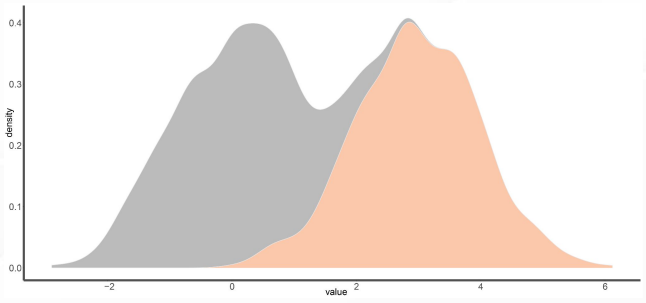
- Stacking - hard to read
- Partially transparent - appears three groups of data

Solution: Overlapping density plots
continuous density lines help the eye keep the distributions separate.









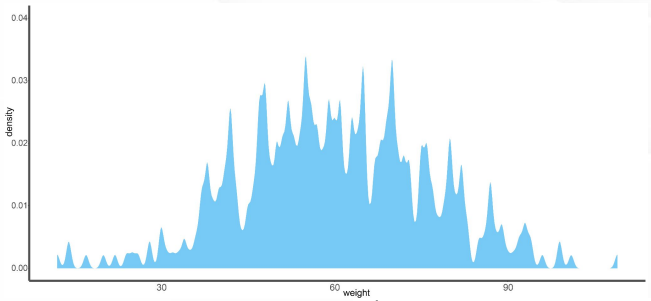
Common mistakes: Histograms and Density plots

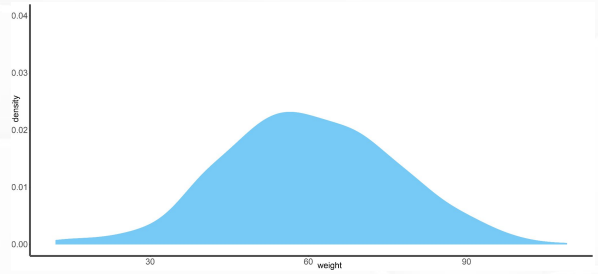
PROBLEM:

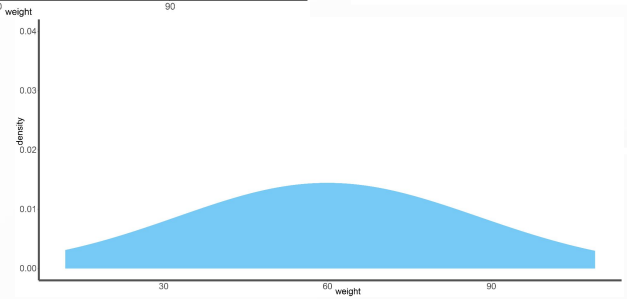
- The default bin “width” does not represent data appropriately

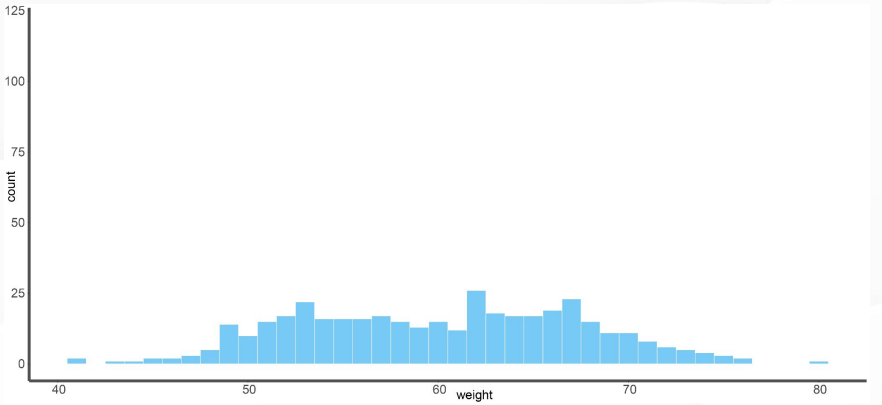
SOLUTION:

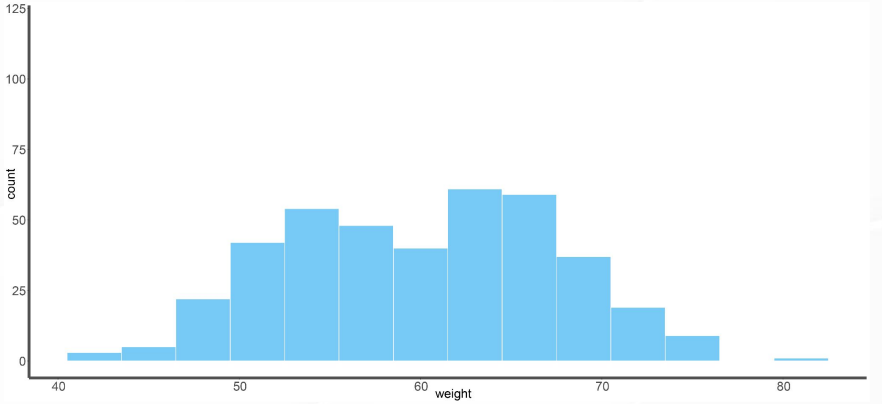
- Test many different bin widths for new insights

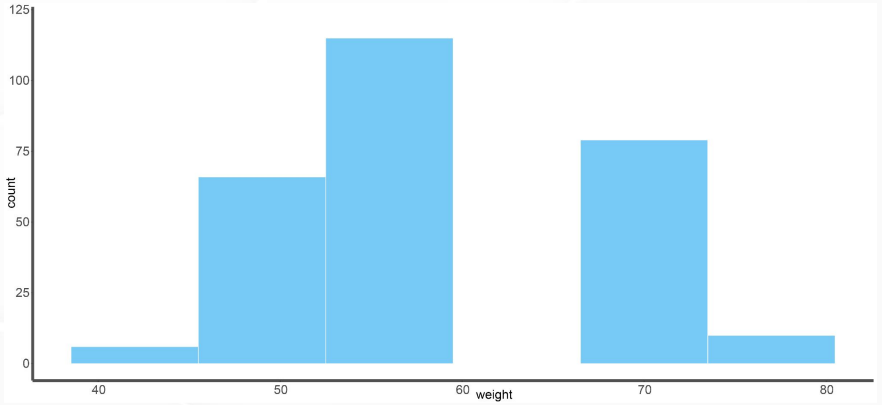






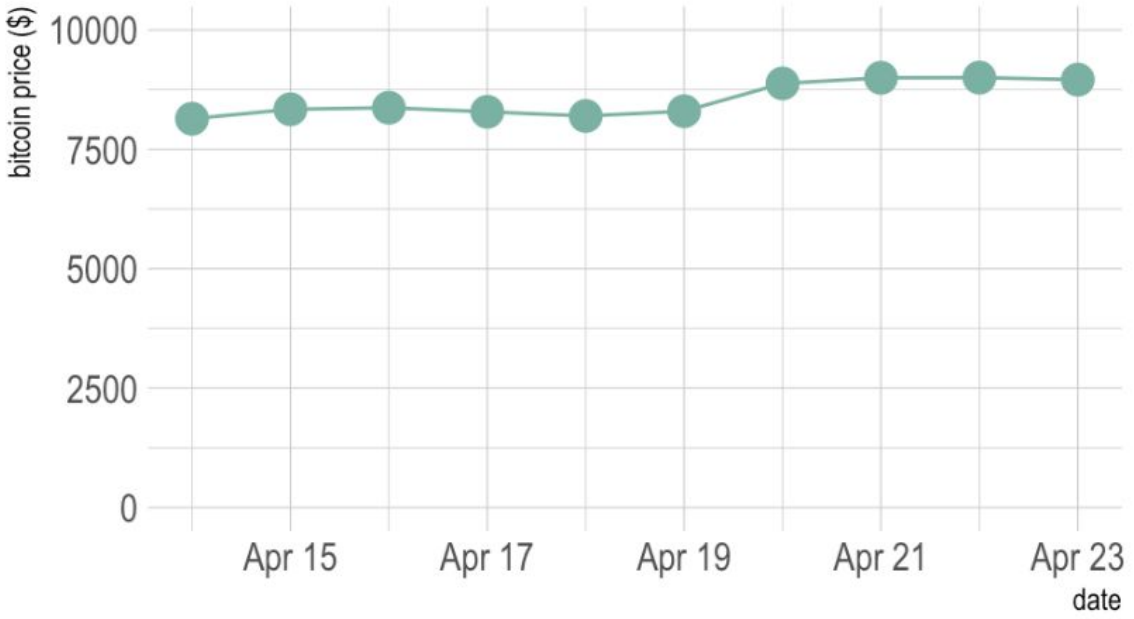


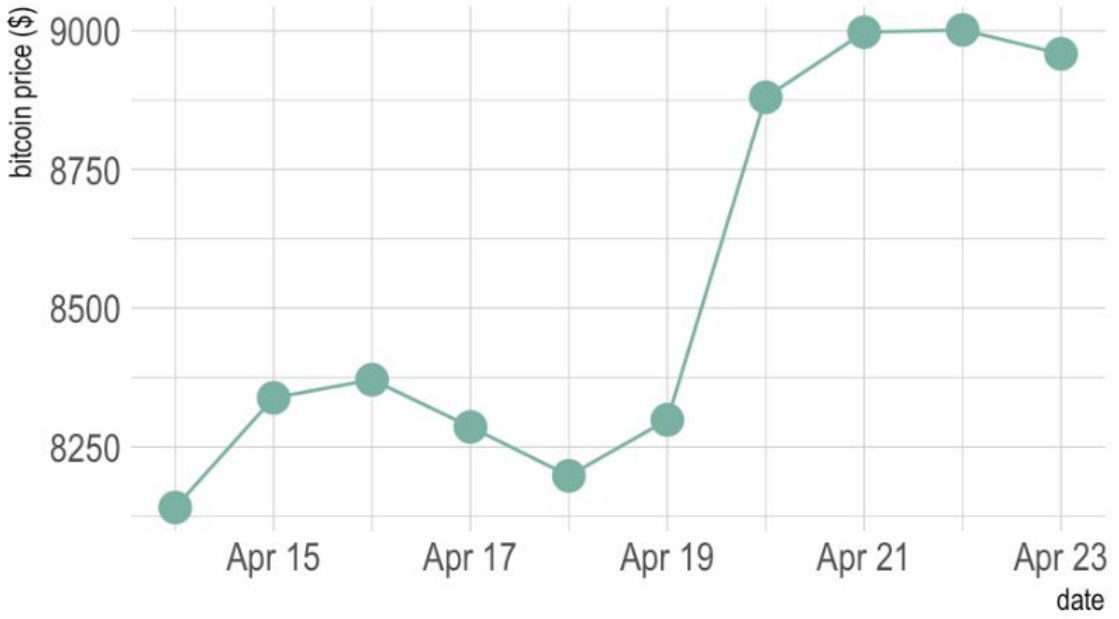




Best practice: Line graph

- If necessary, cut the Y axis
- Mind the spaghetti chart: too many lines make the chart unreadable.



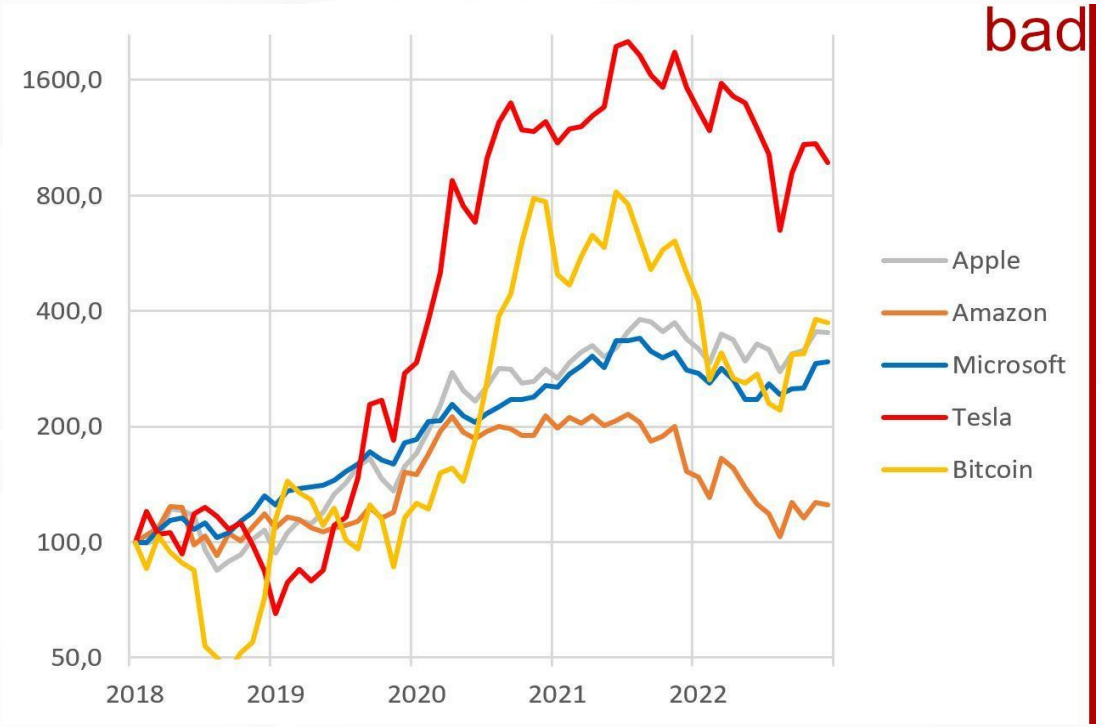


https://www.data-to-viz.com/graph/line.html

Common mistakes: line graph

Problem: y-axis is not labeled and there are no units.

Problem: Companies legend should be in the same order as trendlines (Tesla should be first).



Stock price over time for four major tech companies and Bitcoin. The price has been normalized to equal 100 in May 2018.
Data source: Yahoo Finance

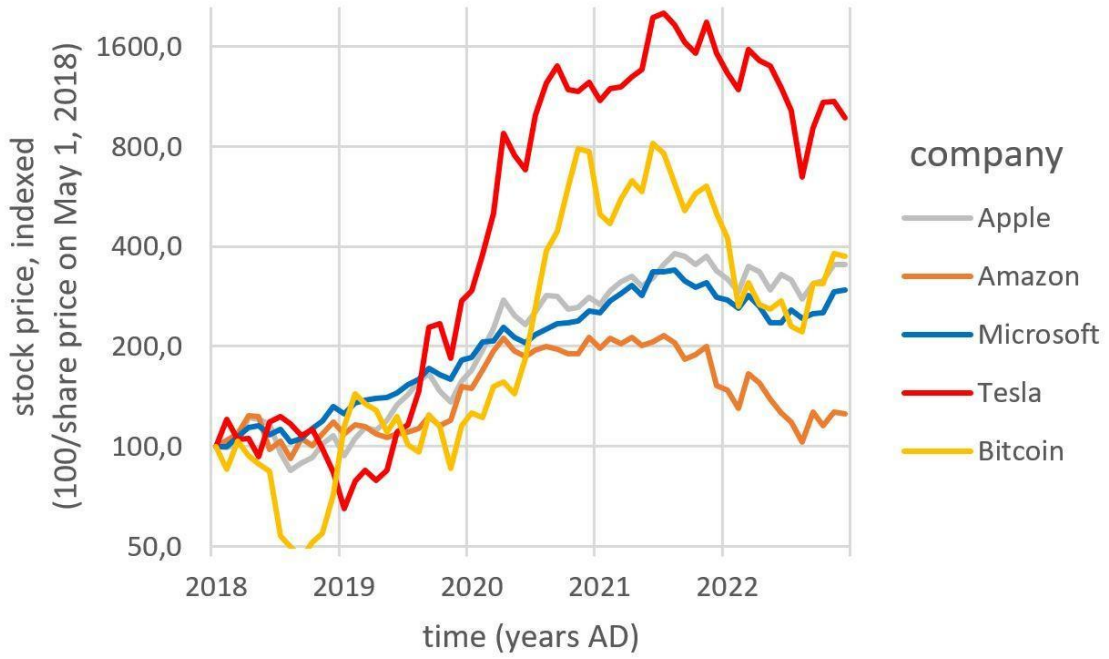
Common mistakes: line graph

Problem: overlabeling.

It's clear that we are dealing with years, no need to add "time(years AD)".

Same goes for companies. Companies legend should be in the same order as trendlines (Tesla should be first).

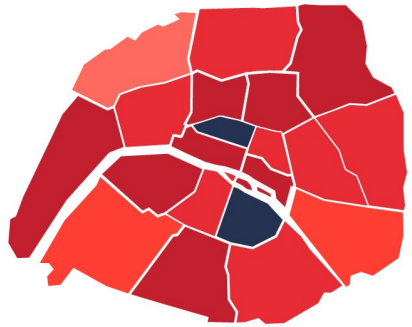
ugly



Stock price over time for four major tech companies and Bitcoin. The price has been normalized to equal 100 in May 2018.
Data source: Yahoo Finance

Best practice: choropleth

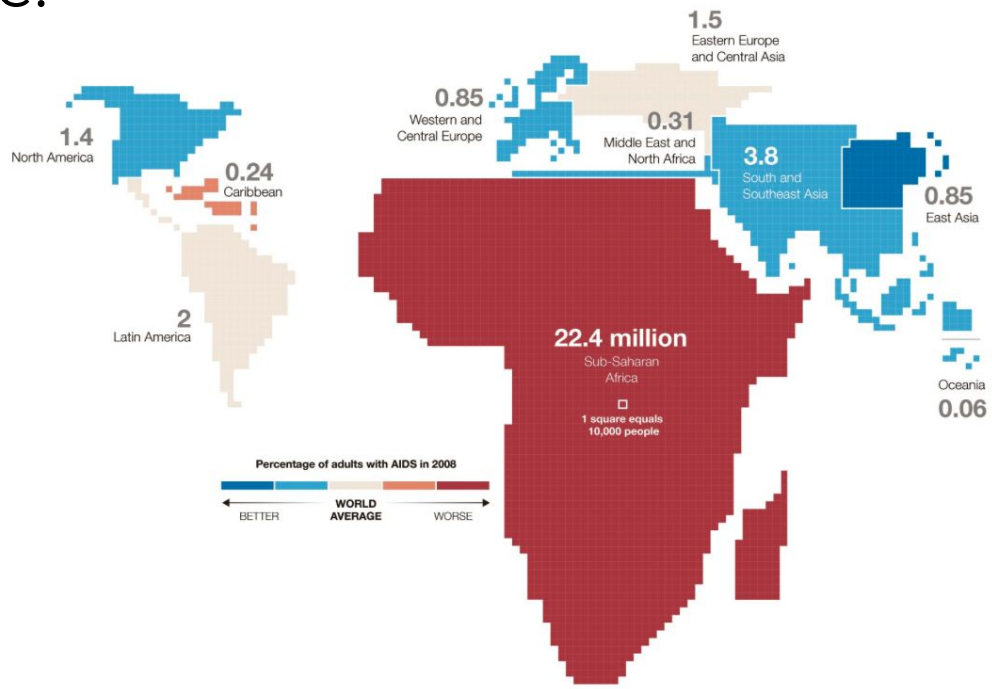
- Normalize your variable: you cannot compare raw numbers between regions of distinct size or population.
- Take care when choosing the continuous color palette.
- Don’t forget the legend.
- If your regions have a broad range of sizes it introduces a bias. You could consider using hexbin maps instead.
- Don’t call it chLoropleth map.



Common mistakes: cartogram

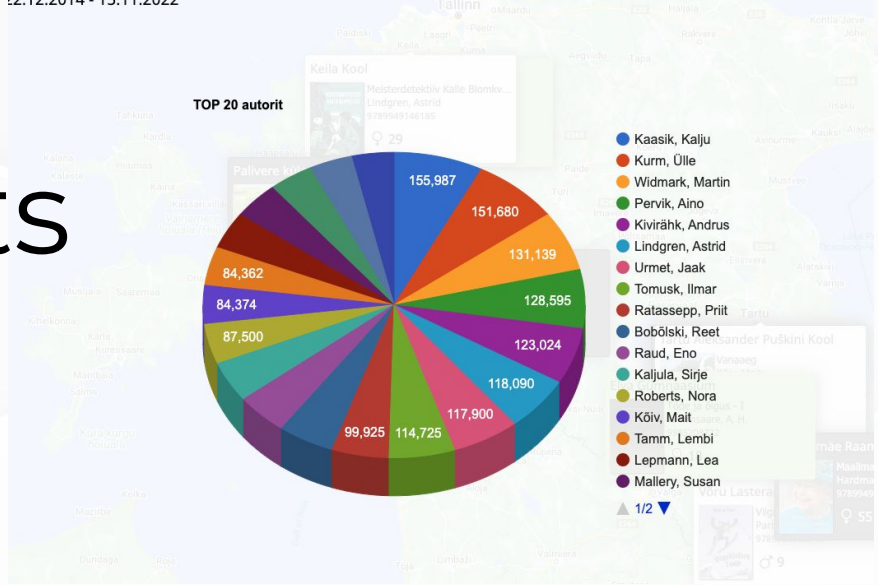
- Distorts real boundaries and thus makes the map harder to identify.

- Be careful not to confuse your audience: you need to introduce it with good explanations and showing the initial map is probably a good practice.



Best practice: Pie charts

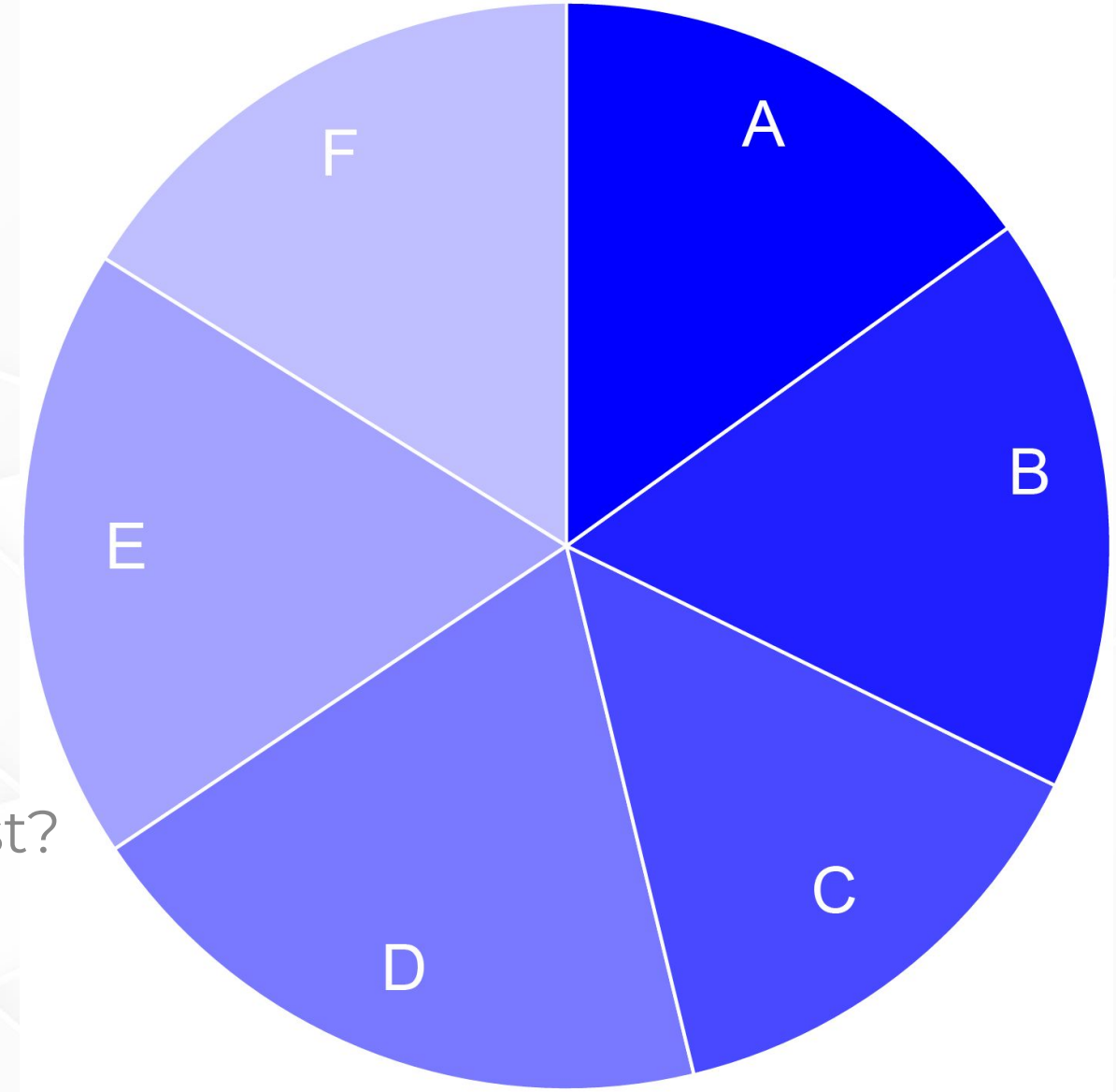
- Don’t use more than five sections
- Avoid comparing one pie to another
- Place the largest slice to the top right corner, and then go by size order
- Don’t use 3D pie



https://www.raamatukogud.ee/#

Poll

Which of the slices is the biggest?

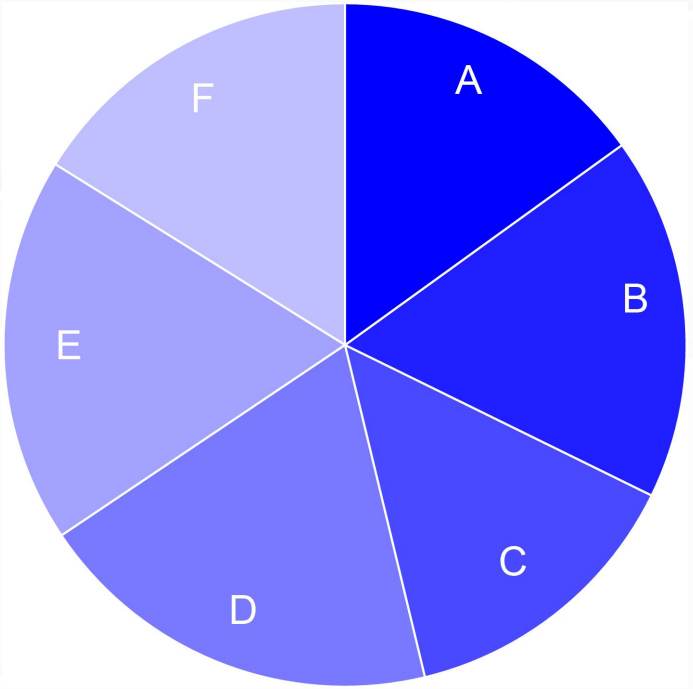


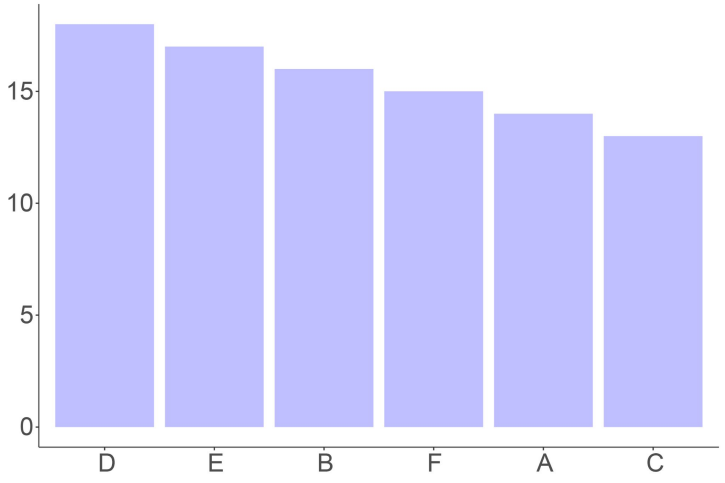
Common mistakes: Pie charts

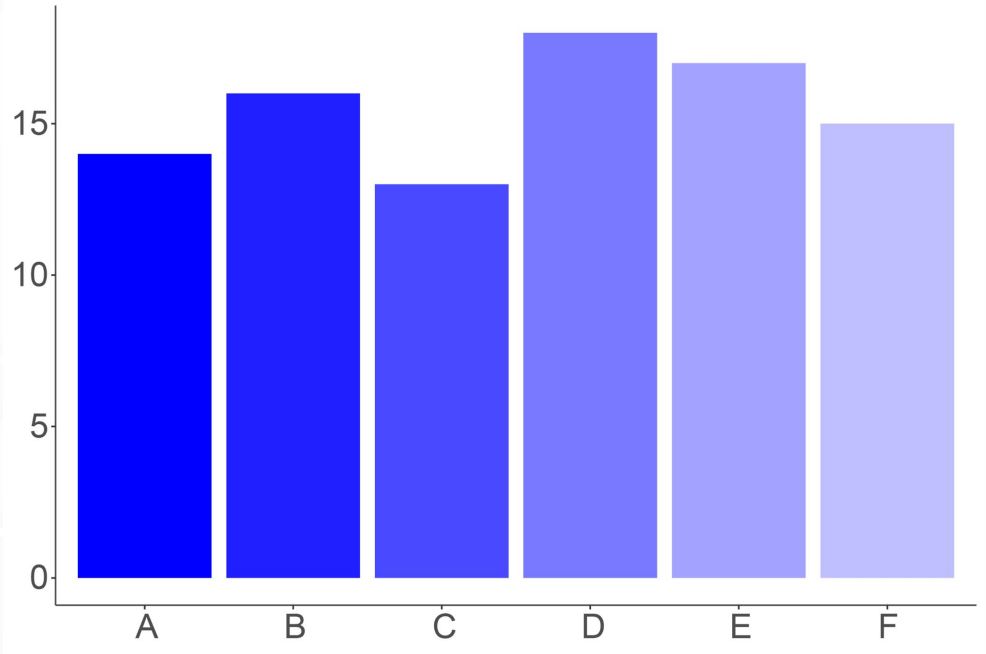
PROBLEM:
Humans are pretty bad at reading angles

Solution: Use bar chart

only use pie chart for simple fractions!
½ ; ⅓; ¼

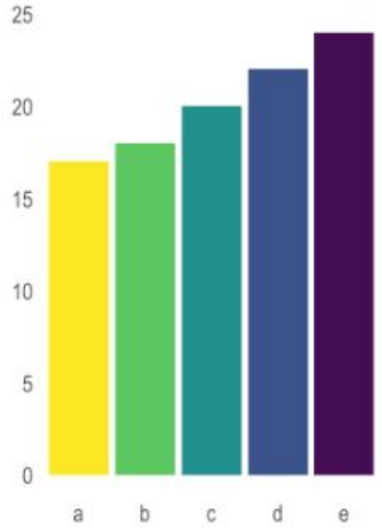


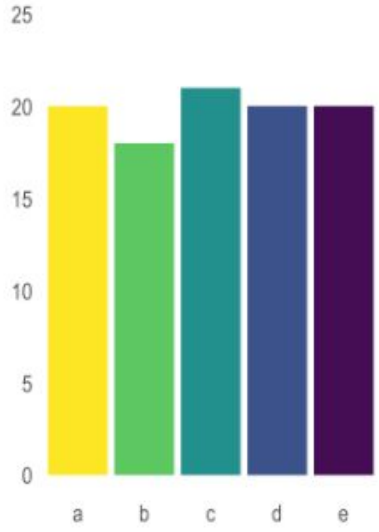


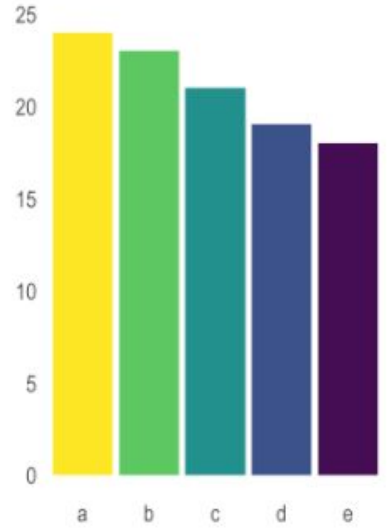


Common mistakes: pie chart

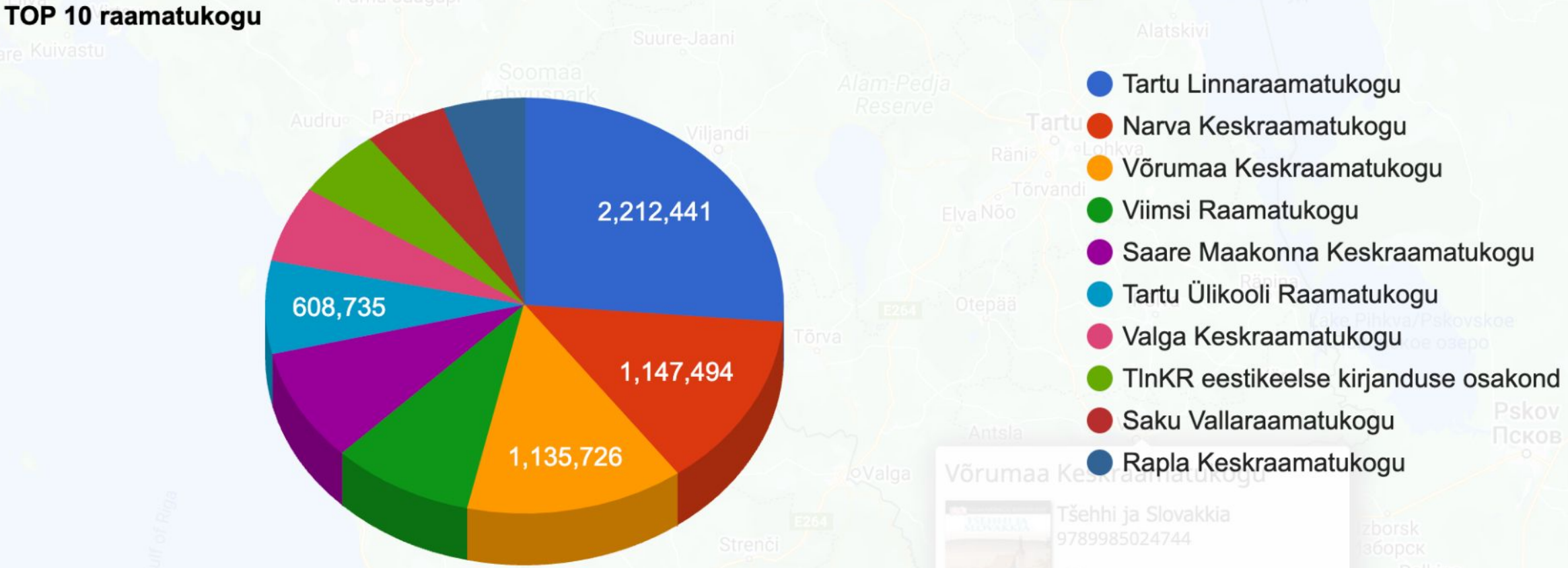
Significant difference between the three pie plots with a hidden pattern that you definitely don't want to miss when you tell your story



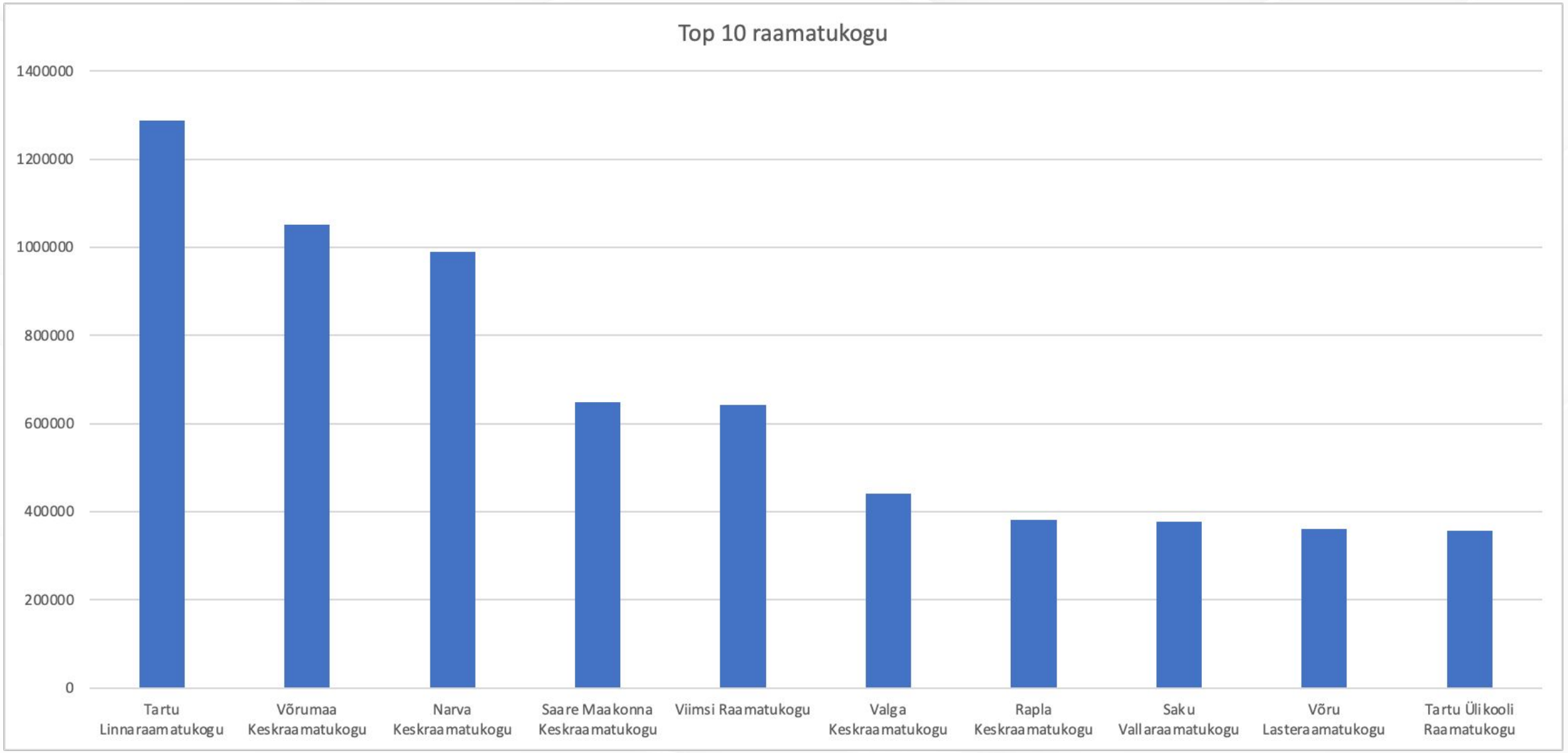




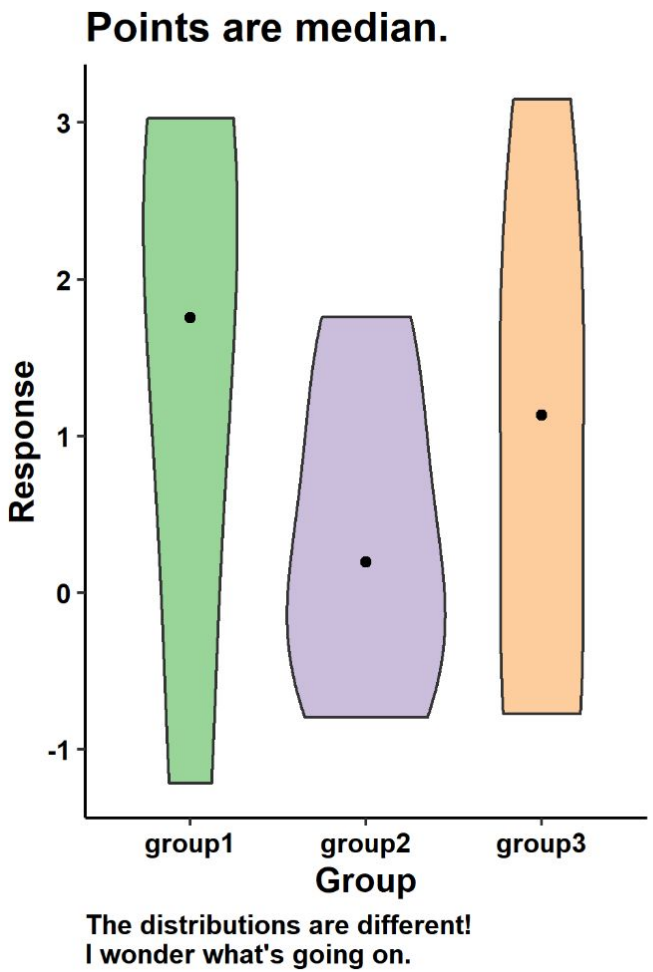
Pie chart: Real life example

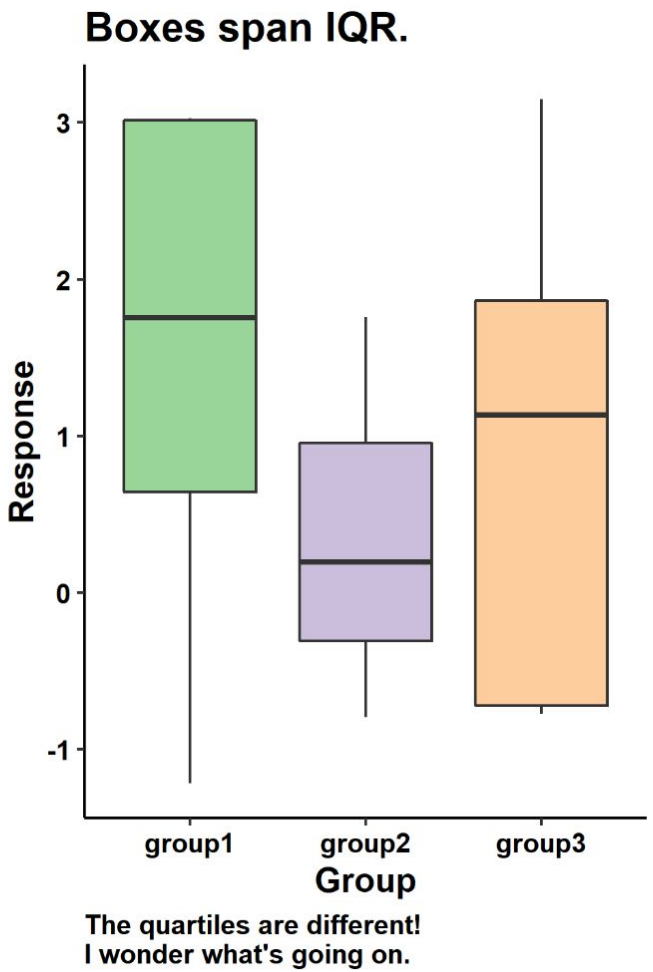


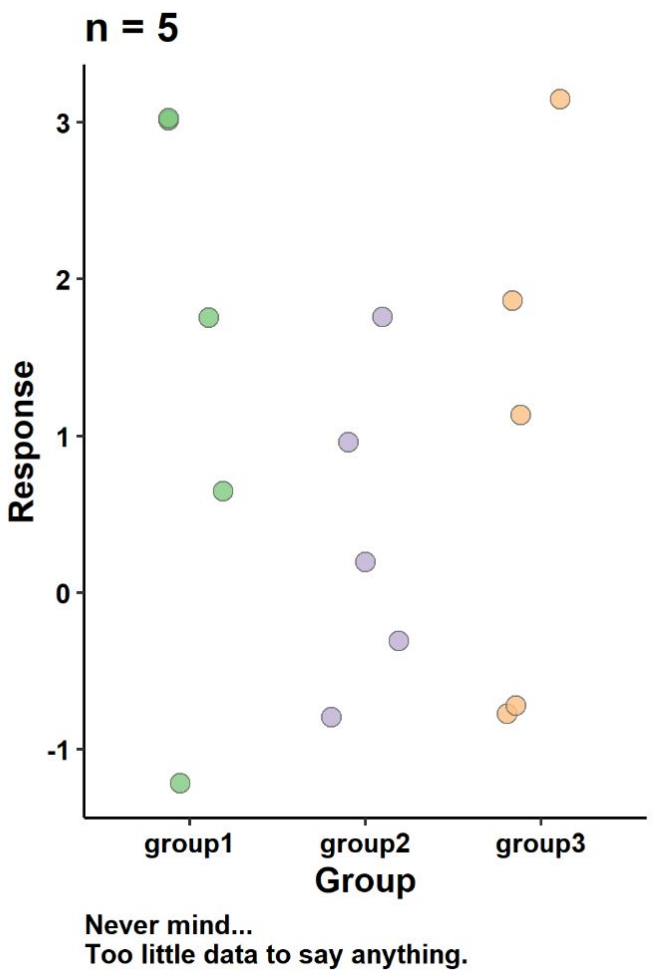
https://www.raamatukogud.ee/#


Top 10 raamatukogu
1400000
1200000
1000000
800000
600000
400000
200000
0
Tartu Linnaraamatukogu
Võrumaa Keskraamatukogu
Narva Keskraamatukogu
Saare Maakonna Keskraamatukogu
Viimsi Raamatukogu
Valga Keskraamatukogu
Rapla Keskraamatukogu
Saku Vallaraamatukogu
Võru Lasteraamatukogu
Tartu Ülikooli Raamatukogu

Common mistake: Violin plot for small sample size





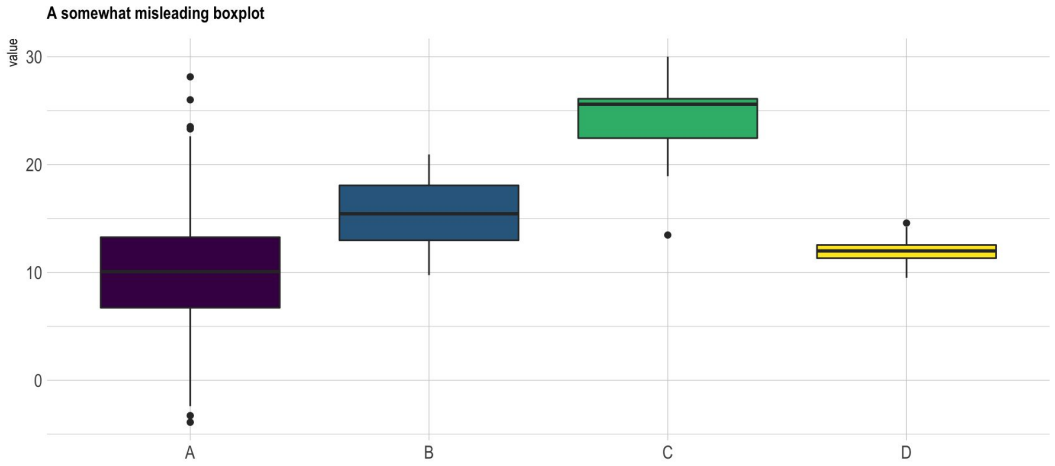


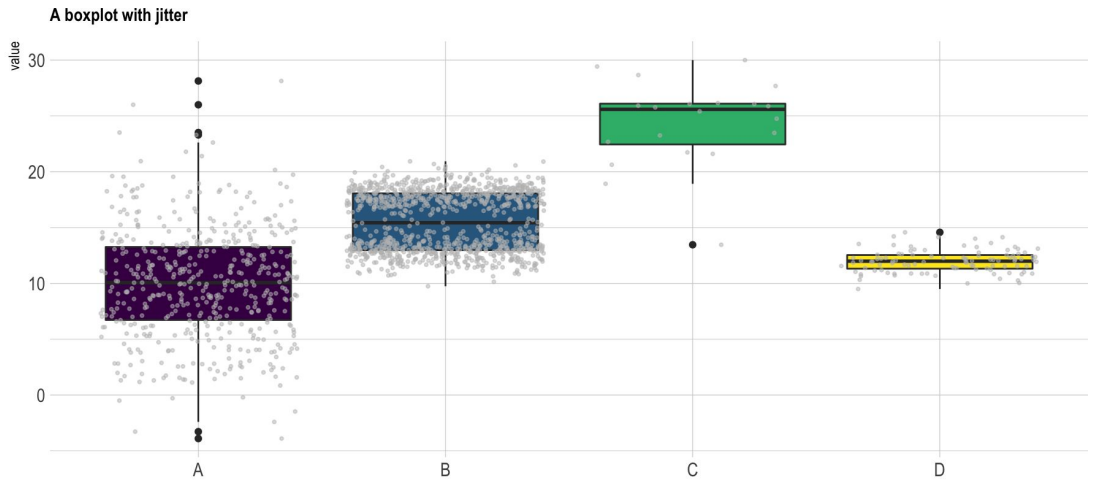
https://github.com/cxli233/FriendsDontLetFriends
C. Li. (2023). cxli233/FriendsDontLetFriends: FriendsDontLetFriends. Zenodo. https://doi.org/10.5281/zenodo.7097522

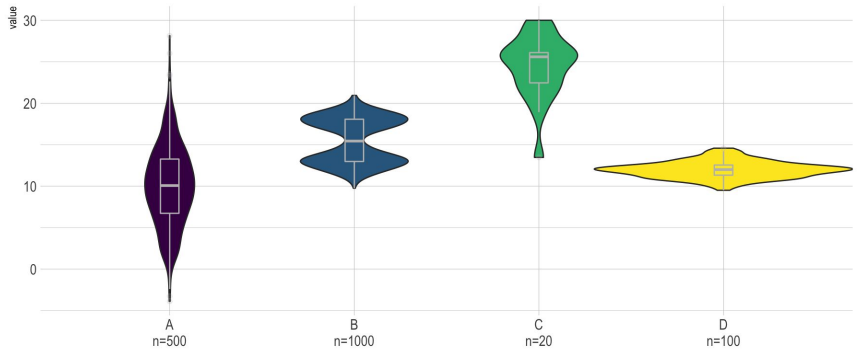
Best practices: Violin plot

- If you compare groups with very different sample size, show it.
- Ordering groups by median value makes the chart more insightful.
- If you have just a few groups, you are probably interested by ridgeline charts.

Common mistakes: uninformative graph







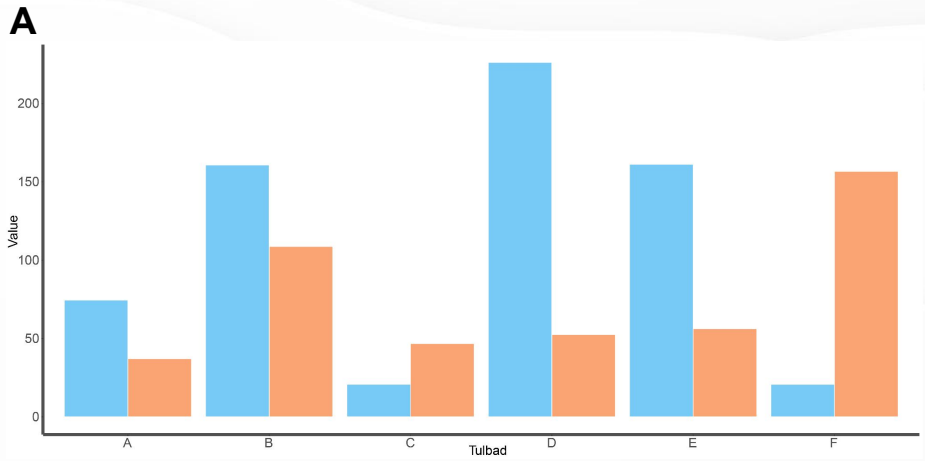
https://www.data-to-viz.com/caveat/boxplot.html

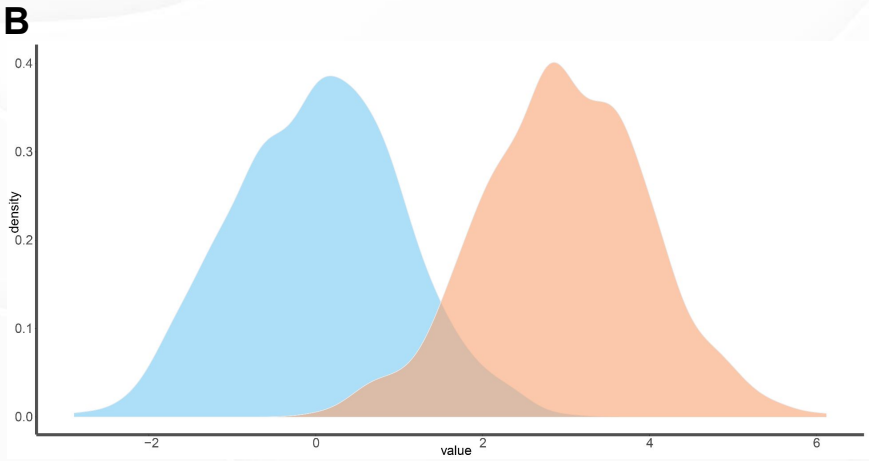
Figures belonging together

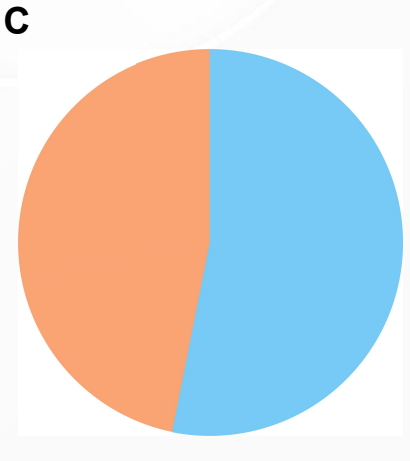
Be consistent but not repetitive.

Figures belonging together should look similar, but not exactly the same.

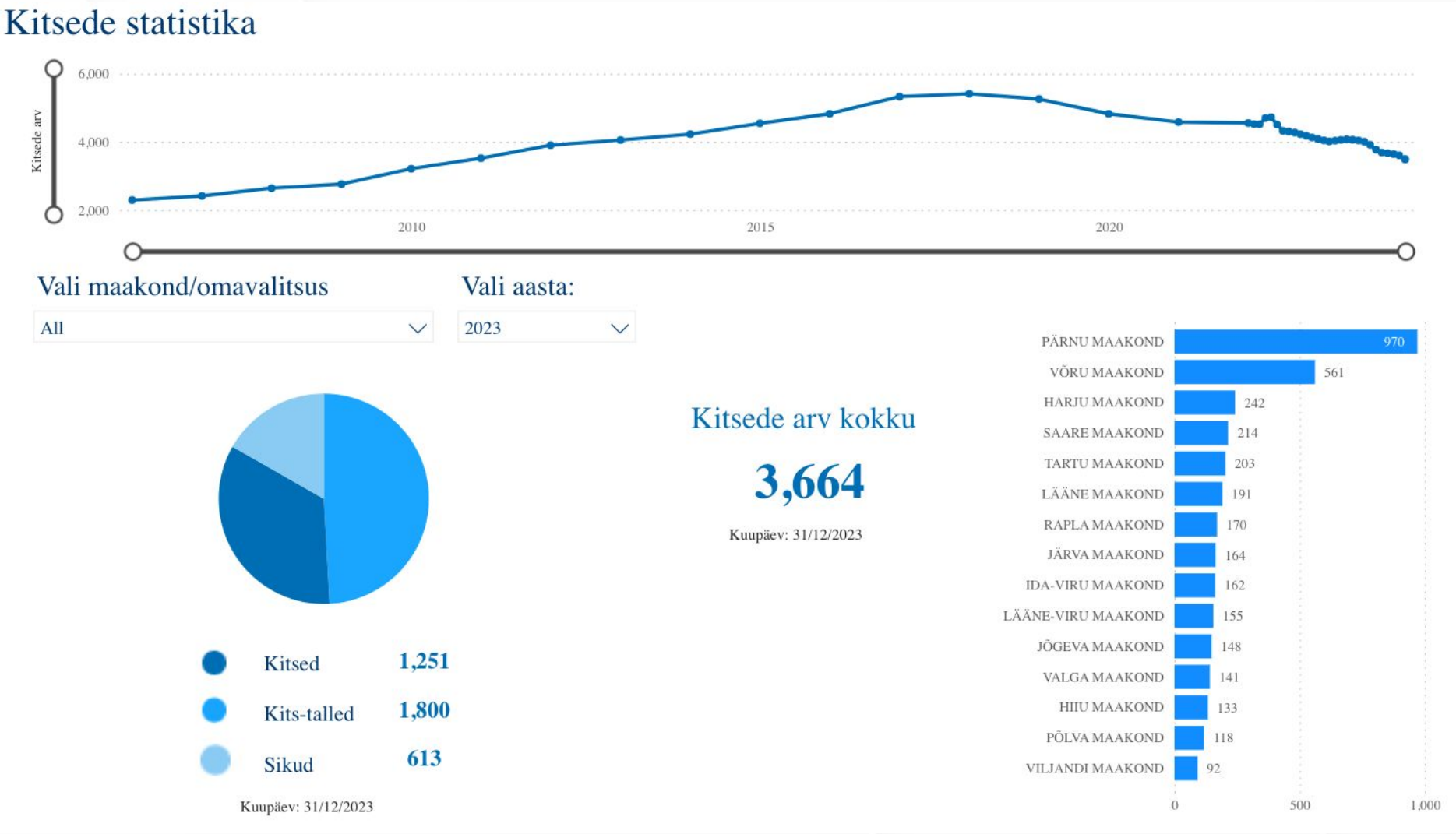
Mix up different chart types, but use similar layout, types, color choices.







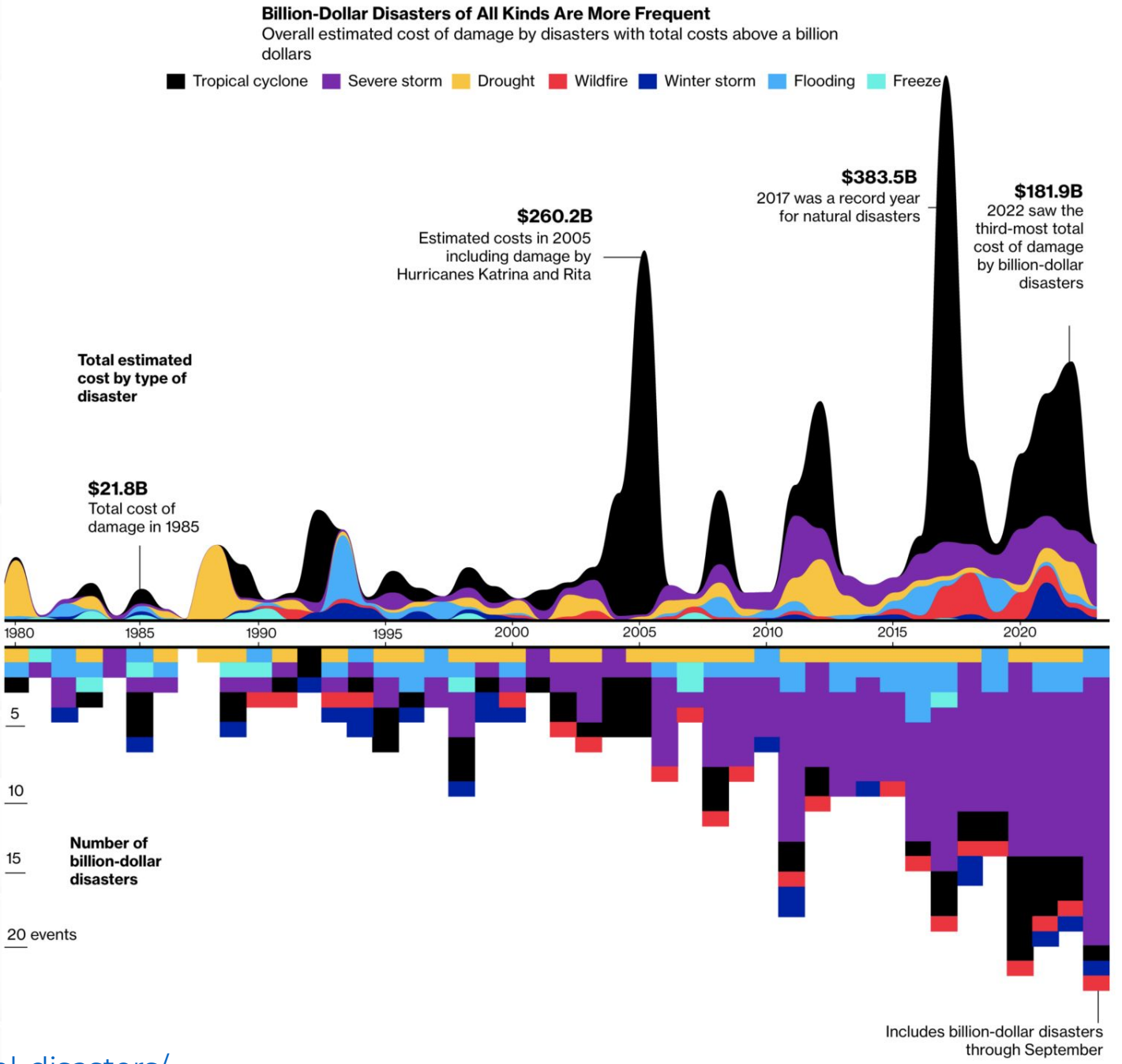
Example



https://www.pria.ee/infokeskus/statistika/loomad#kitsede-statistika

What would you change?

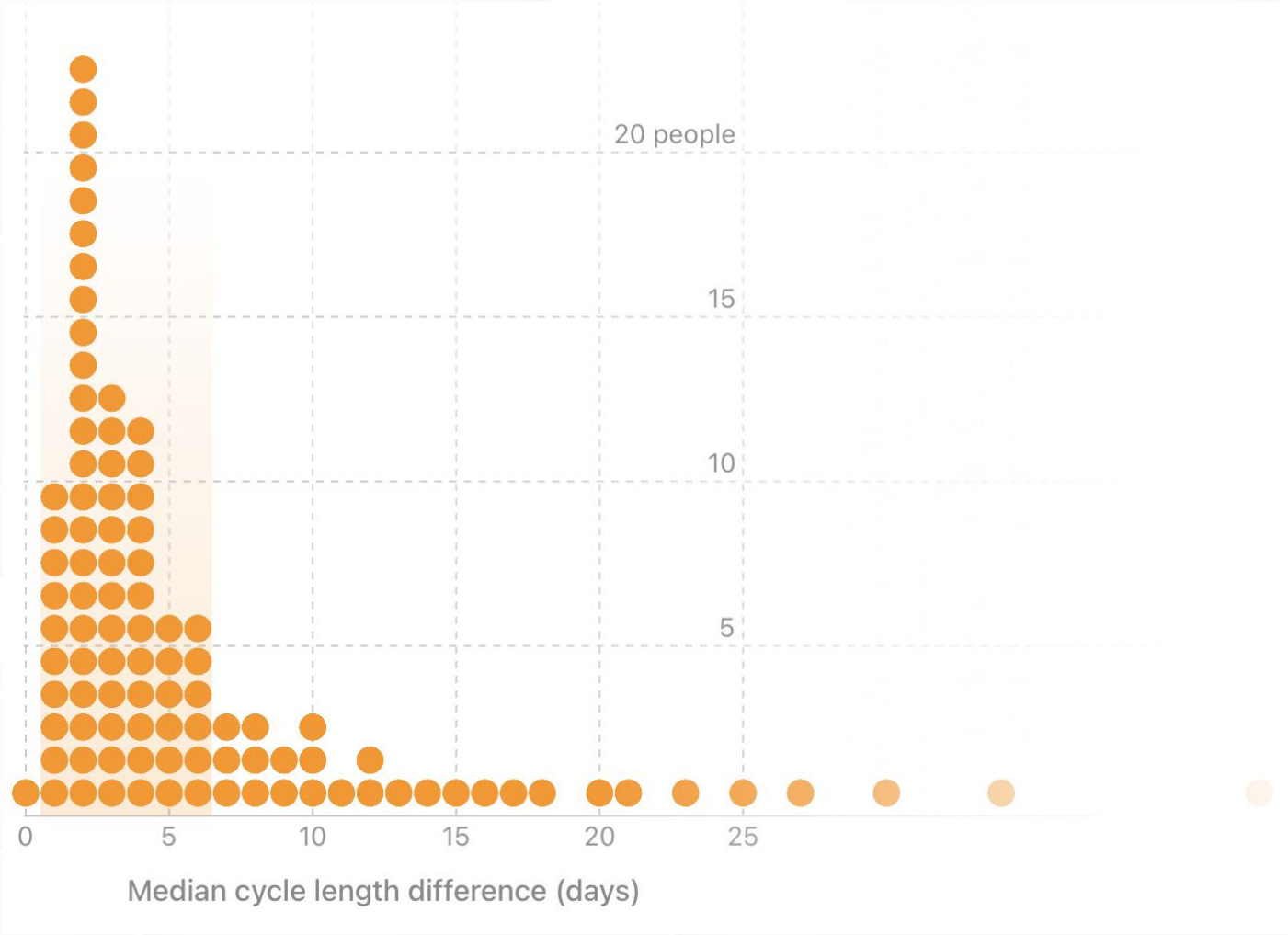
Combo stacked area chart and stacked bar chart on the top and bottom to show increased cost of billion dollar disasters and the counts over time.



https://flowingdata.com/2023/10/03/billion-dollar-natural-disasters/

What would you change?

For the Apple Women's Health Study, which uses cycle tracking data from iPhones and Apple Watches, the Harvard T.H. Chan School of Public Health provides a visual explanation of how menstrual cycles vary.

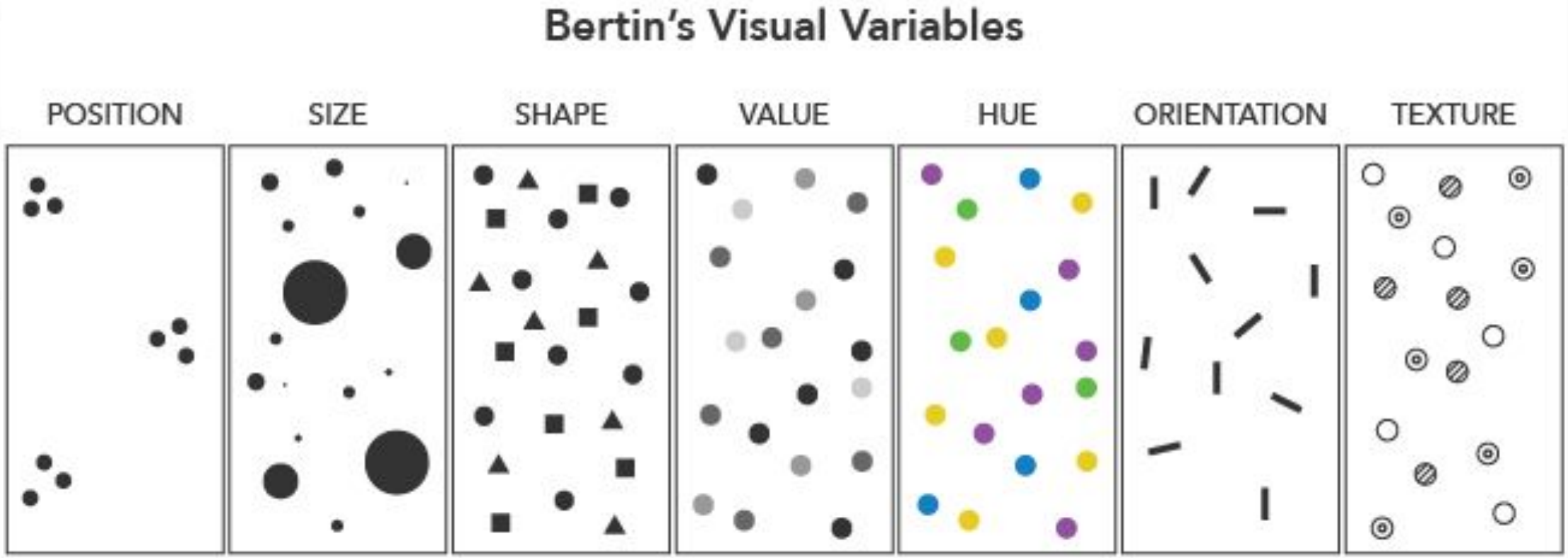


https://flowingdata.com/2023/10/02/visual-explanation-of-menstrual-cycle-length-and-variability/

Visual variables

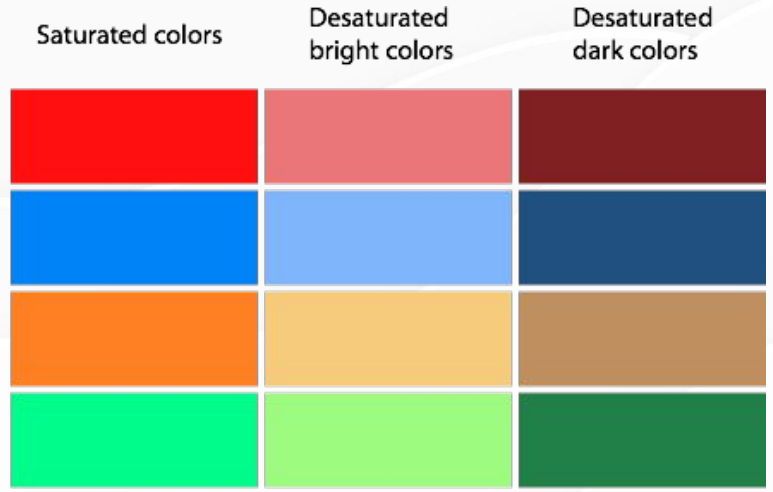
Visual Variables

Fundamental ways in which graphic symbols can be



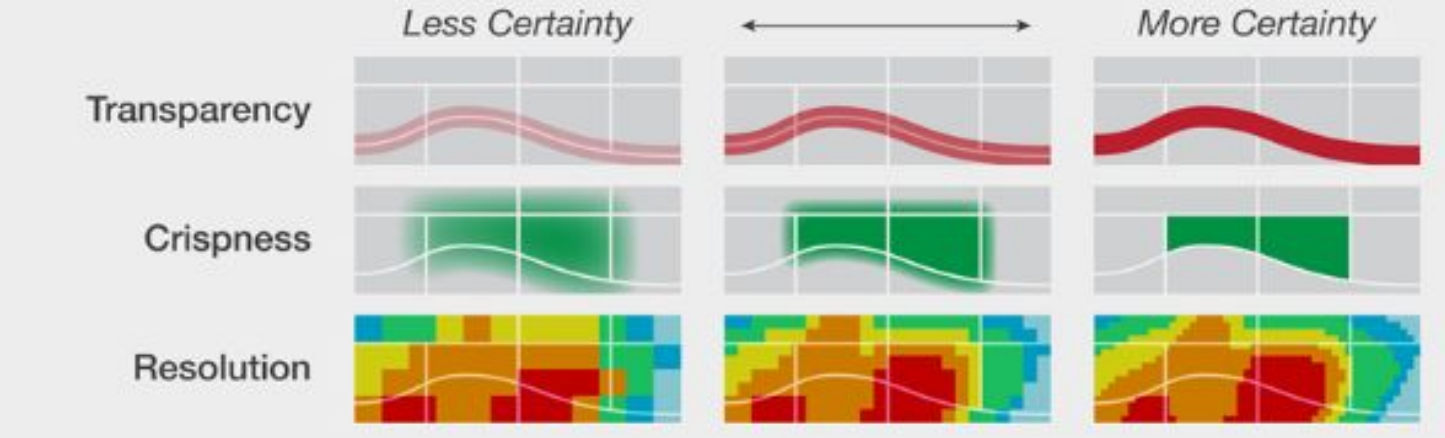
https://www.axismaps.com/guide/visual-variables

Visual Variables



The list has since been further expanded by other researchers (Roth, Joel Morrison, Alan MacEachren):

- Color saturation
- Arrangement
- Crispness
- Resolution
- Transparency

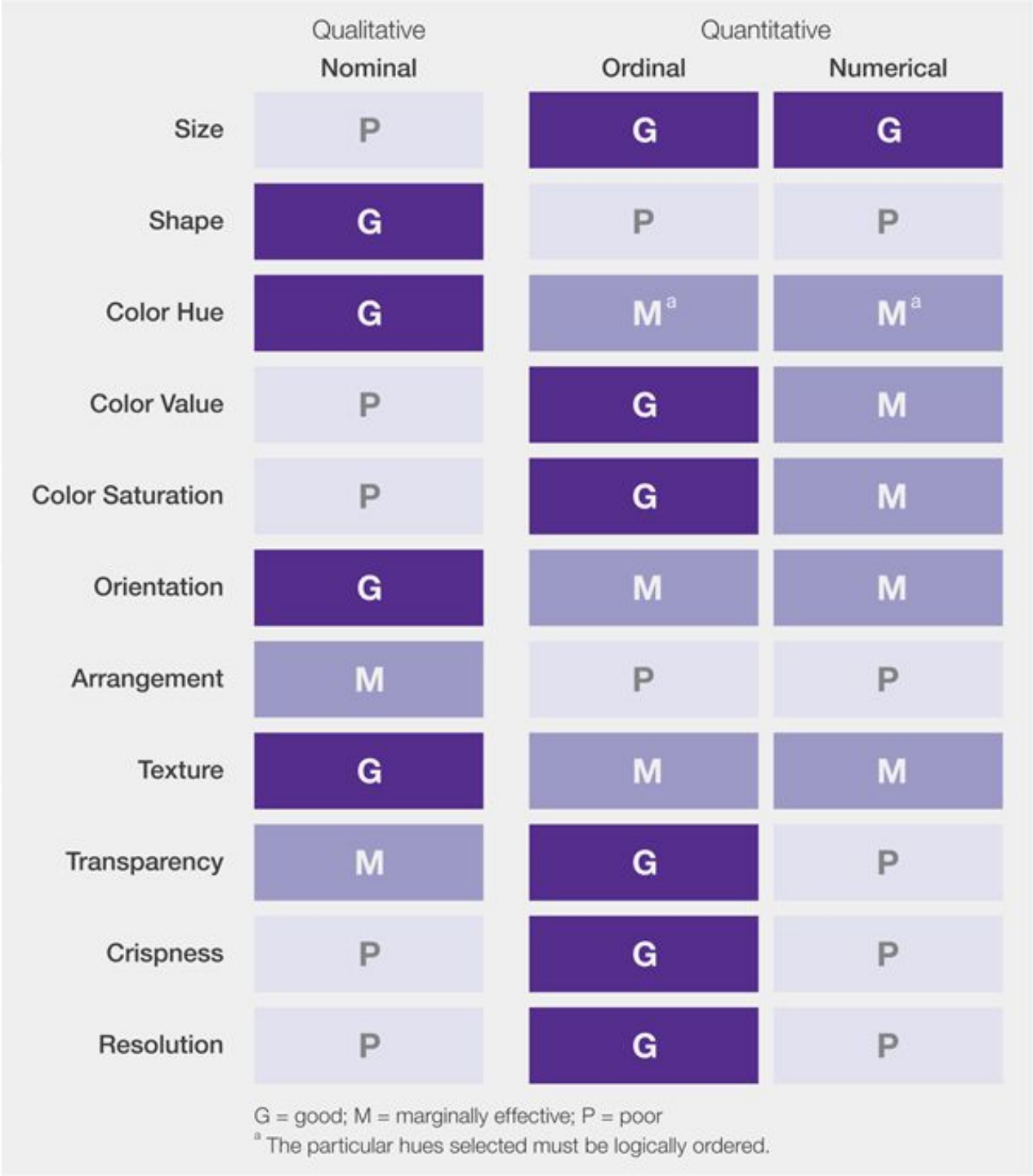


White, T. (2017). Symbolization and the Visual Variables. *The Geographic Information Science & Technology Body of Knowledge* (2nd Quarter 2017 Edition), John P. Wilson (ed.).DOI: 10.22224/gistbok/2017.2.3

https://www.axismaps.com/guide/visual-variables

Variable	Definition
Size	Variations in the length, area, or volume of a symbol (e.g., graduated circles) denote different quantities of data.
Shape	The appearance or form of a symbol; different shapes (e.g., circles, squares, triangles) denote different categories of data.
Color Hue	The dominant wavelength of visible light in the electromagnetic spectrum (red, orange, yellow, green, blue, indigo, violet); different hues can denote categorical or numerical differences in the data.
Color Value	Light or dark variations of a single hue (e.g., light red–medium red–dark red) denote different quantities of data. Sometimes called *Lightness* or *Brightness*.
Color Saturation	The intensity of a single hue. Fully saturated hues appear vivid, and fully desaturated hues appear muted. Denotes ordinal-level data or uncertainty in data. Sometimes called *Intensity* or *Chroma*.
Orientation	The direction or angle of rotation of an entire map symbol or the individual marks that make up the map symbol. Can denote categorical or numerical differences.
Arrangement	The distribution or layout of individual marks that make up a map symbol (e.g., regular or irregular). Primarily used in areal patterns to denote categorical differences in the data.
Texture	The relative fineness or coarseness of the areal fill within a map symbol. Can denote categorical or numerical differences in the data.
Transparency	The blend level between a symbol and a background layer; reliable data appear opaque, whereas transparency increases with uncertainty.
Crispness	The sharpness of boundaries; reliable data have crisp boundaries, but borders become fuzzy or imprecise as uncertainty increases. Also called *Fuzziness*.
Resolution	The level of detail or precision of a spatial data set. Analogous to the spatial resolution of raster grids. High resolution data are more detailed, and resolution decreases as uncertainty increases.

White, T. (2017). Symbolization and the Visual Variables. *The Geographic Information Science & Technology Body of Knowledge* (2nd Quarter 2017 Edition), John P. Wilson (ed.).DOI: 10.22224/gistbok/2017.2.3



	Qualitative	Quantitative	
	Nominal	Ordinal	Numerical
Size	P	G	G
Shape	G	P	P
Color Hue	G	M[a]	M[a]
Color Value	P	G	M
Color Saturation	P	G	M
Orientation	G	M	M
Arrangement	M	P	P
Texture	G	M	M
Transparency	M	G	P
Crispness	P	G	P
Resolution	P	G	P

G = good; M = marginally effective; P = poor

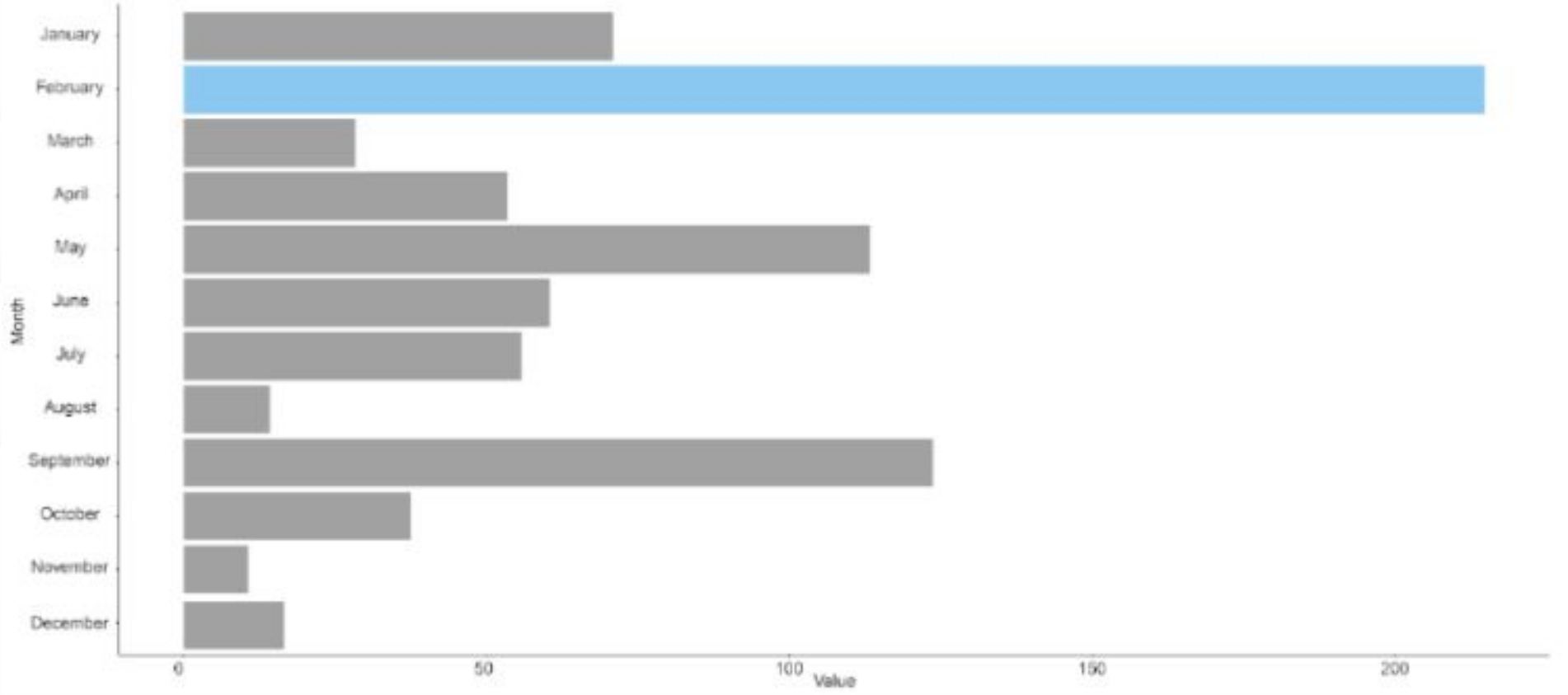
[a] The particular hues selected must be logically ordered.

- Nominal: categorical data, unordered and non-numerical
- Ordinal: ordered or ranked data with no assigned numerical values

White, T. (2017). Symbolization and the Visual Variables. *The Geographic Information Science & Technology Body of Knowledge* (2nd Quarter 2017 Edition), John P. Wilson (ed.).DOI: 10.22224/gistbok/2017.2.3

Figure design

What is the first thing you notice when seeing this figure?



What is the first thing you notice when seeing this figure?

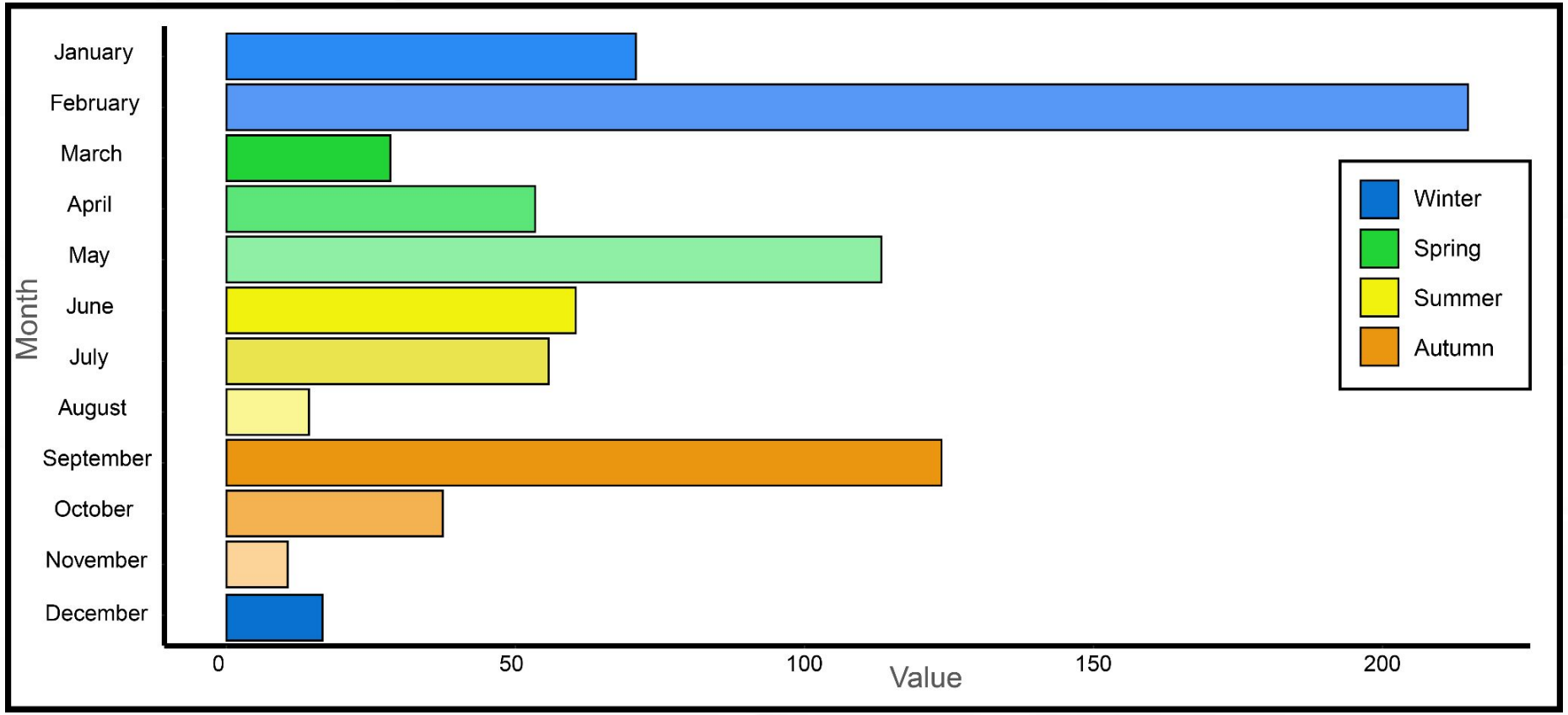


Figure design

Reading charts

- Reading a picture ≠ reading a text
- Depends a lot on colors picked, highlights, additional features

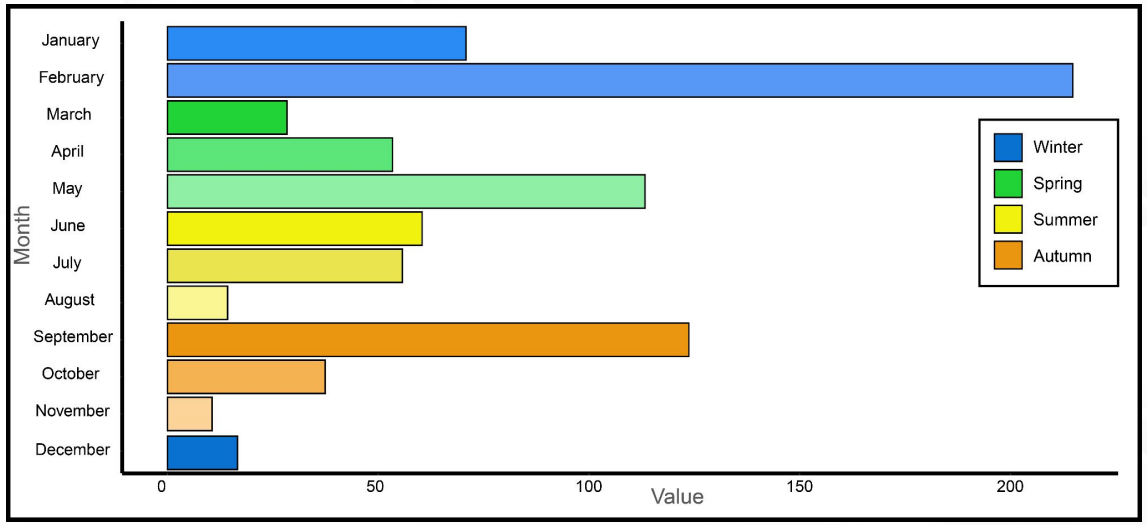
Data VS context:

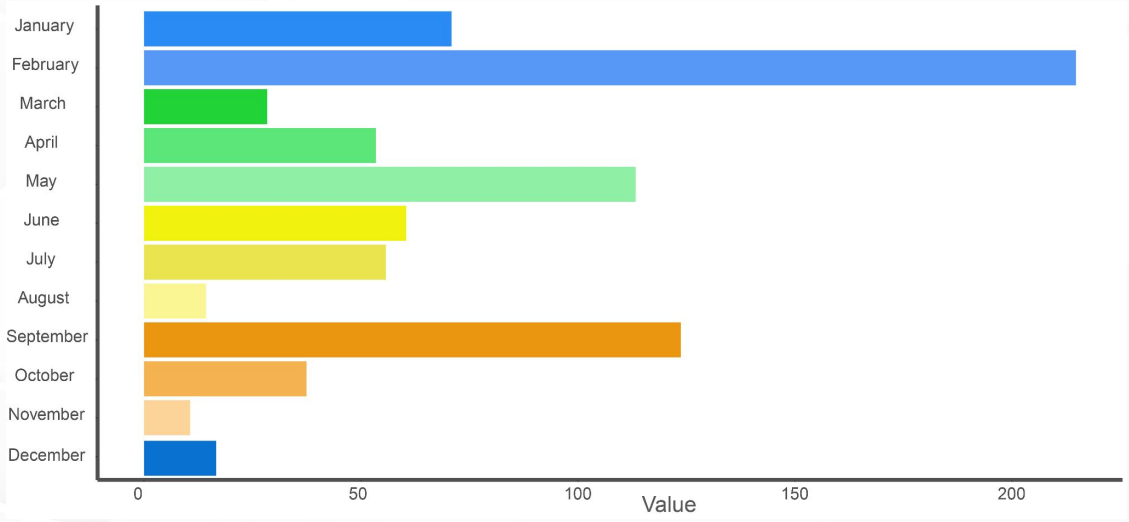
- Elements that represent the **data**
  - Points in scatter plot
  - Bars in histogram or bar chart
  - Shaded areas in heat map

- Elements that do not represent data - **non-data ink**
  - Axis and it’s labels and titles
  - Legends
  - Plot annotations

Data-ink ratio

Maximise the data-ink ratio - remove the clutter and strive for a clean and elegant design.





Font

Font choices

- 2-3 fonts maximum (title, comments, subheaders)
- Make certain fonts match well
- Size: 8-20 points

https://design.tutsplus.com/articles/15-best-fonts-to-pair-with-times-new-roman--cms-42083

Title

Accurately convey a reader what the figure is about

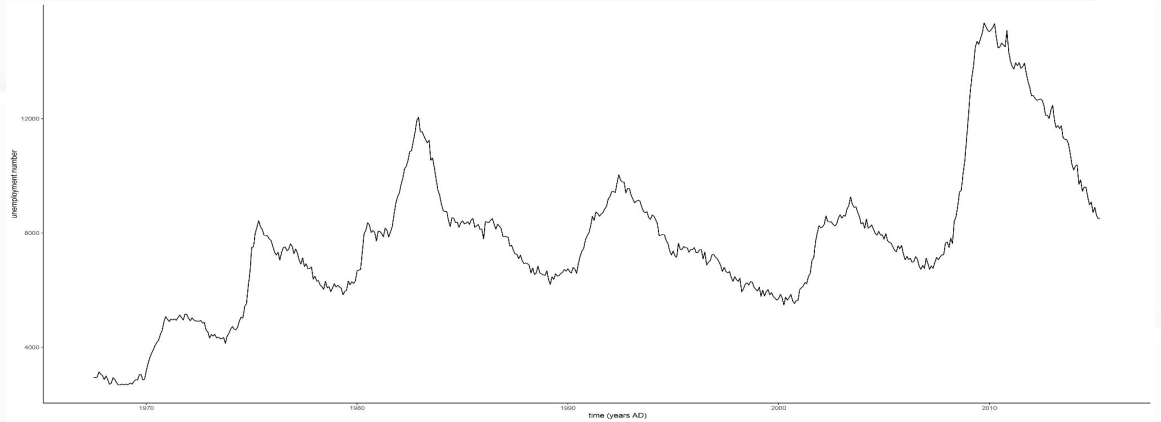
- Title as a caption block underneath the figure
  - Books
  - Articles
- Title in the display (on top of the figure, without caption text)
  - Infographics
  - Social media
  - Web page

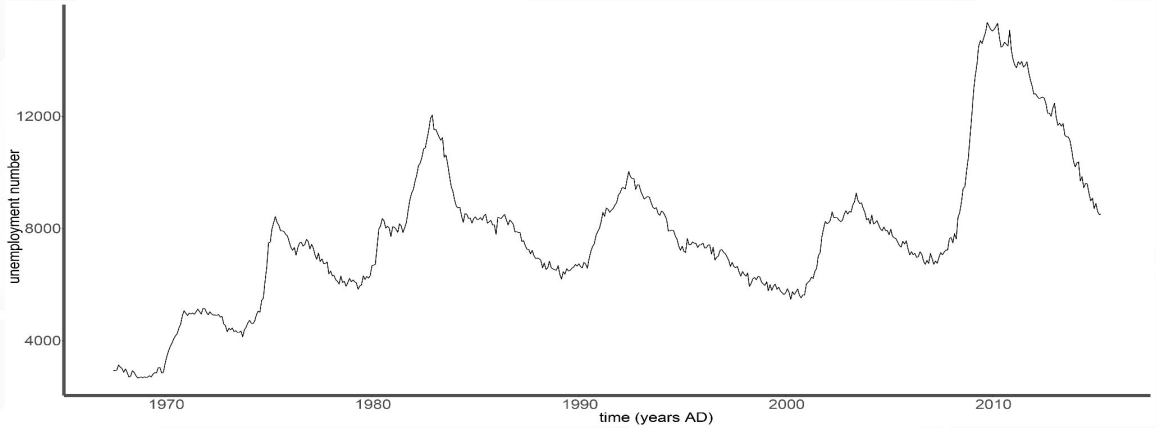
Title should NEVER be omitted. The first part of the caption is ALWAYS the title, not a description of the contents of the figure.

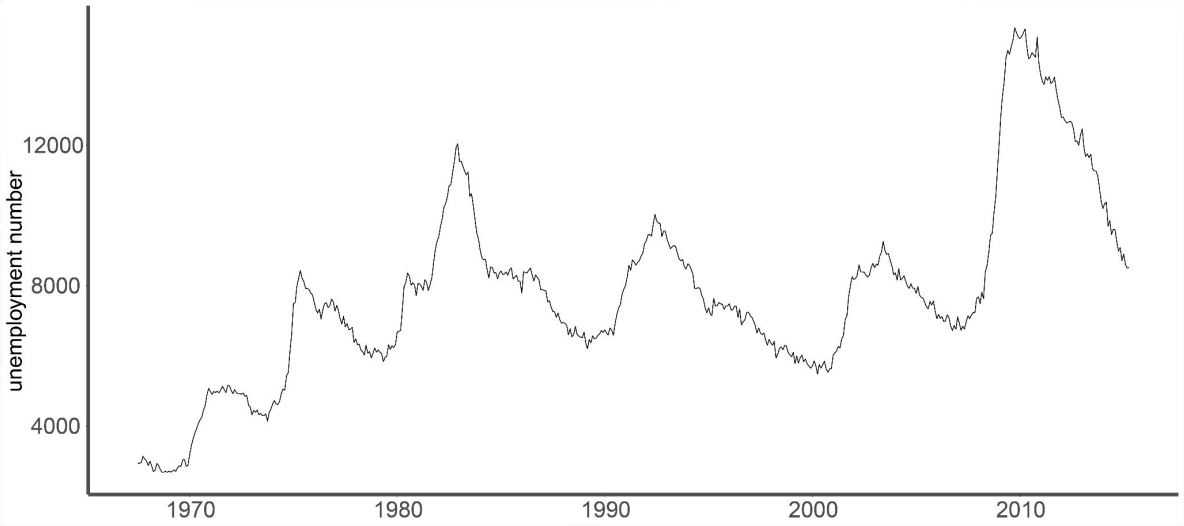
Title size

Make it bigger

Computer monitor lies







Axis and legend

Explain what the displayed data values are and how they map to plot aesthetics

Numerical variables

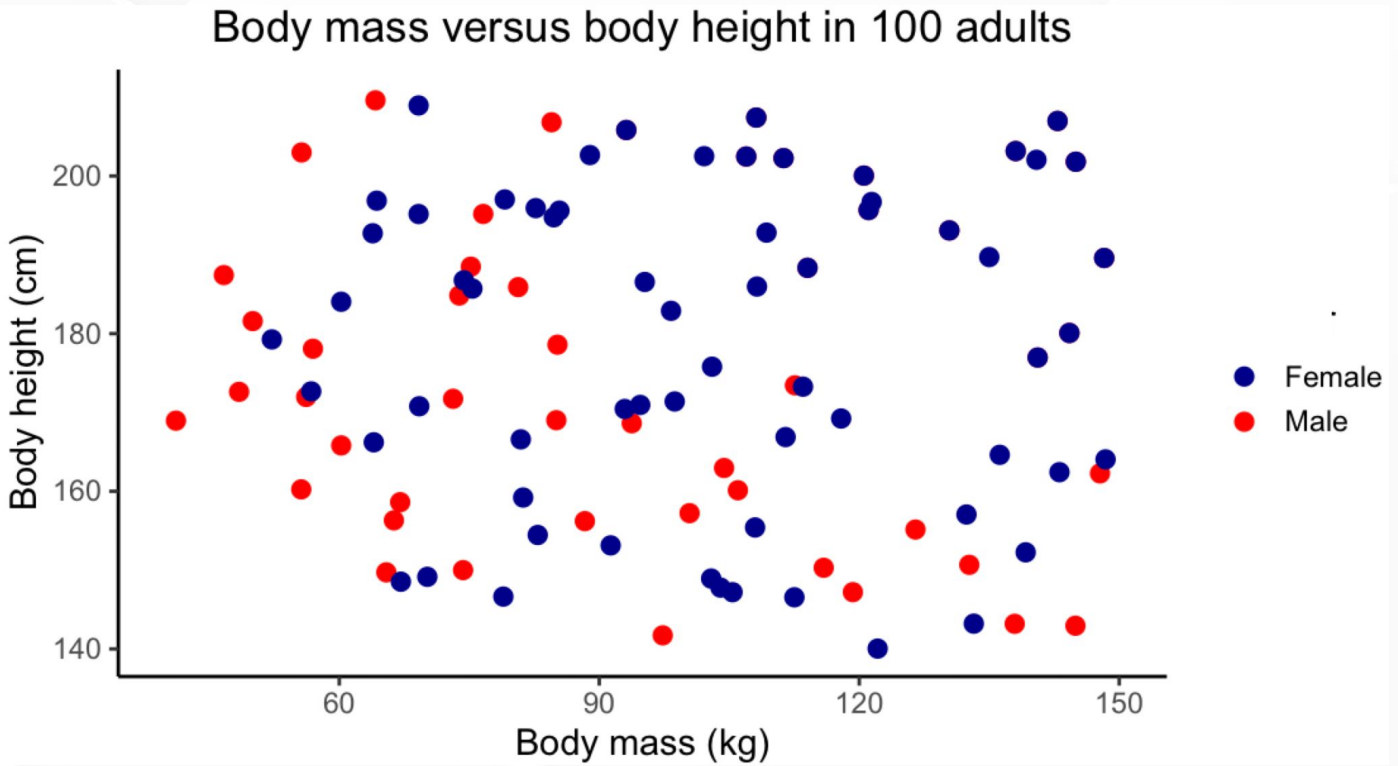
- State variables shown
- Add units of measurements

Categorical variables

- No need to add units

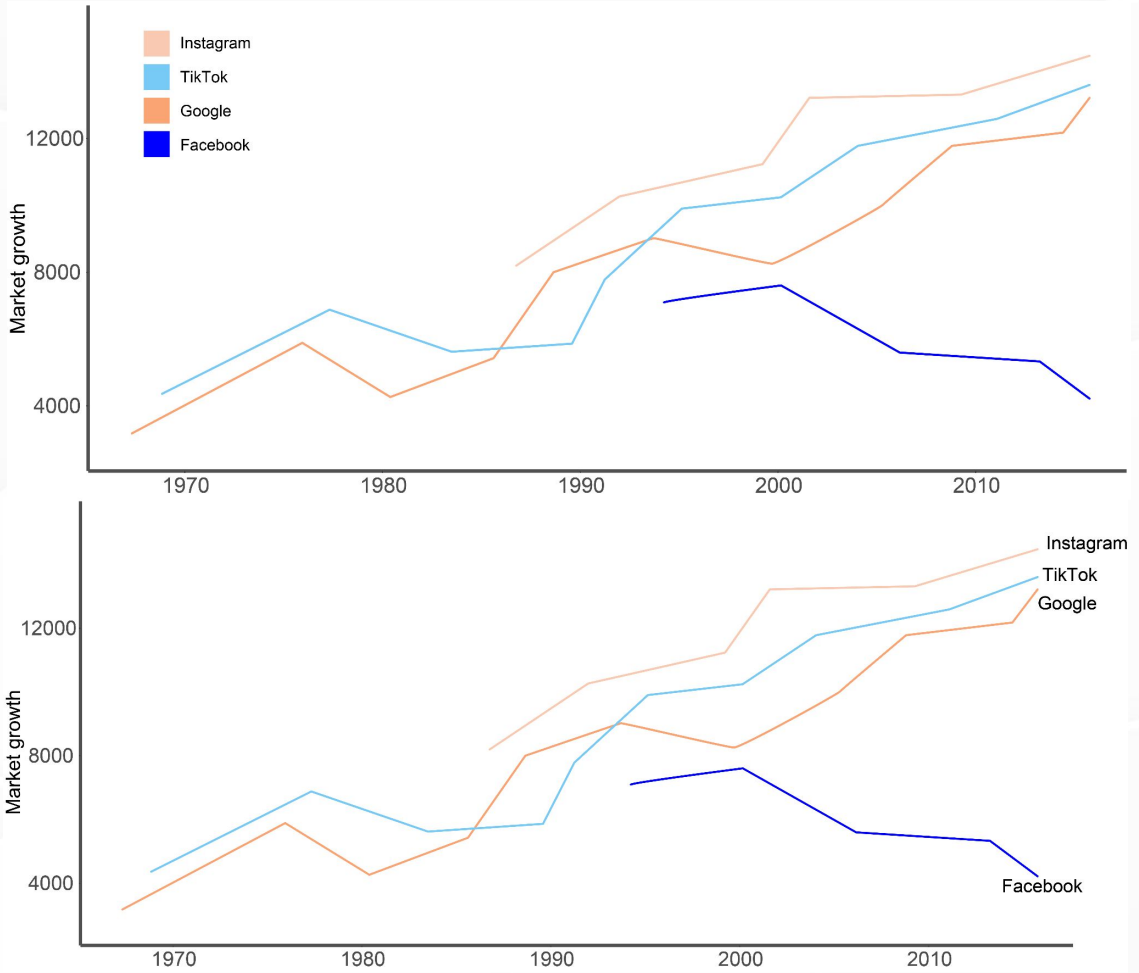
EXCEPTION:

- Labels are fully explanatory



Legend

- If there is a clear visual order in your data, match the legend with it
- If possible, design your figure so that it does not need a legend
  - Direct labeling
- Multiple chart figure
  - Single legend that covers all the charts



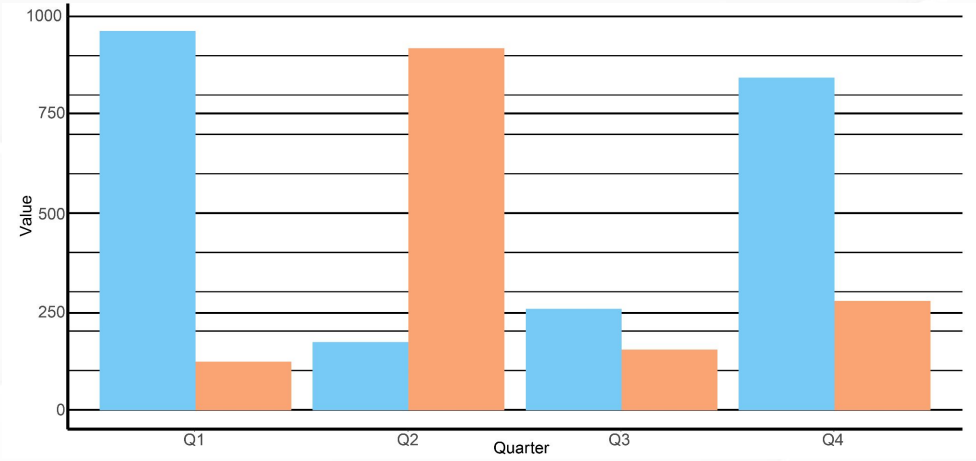
Grids

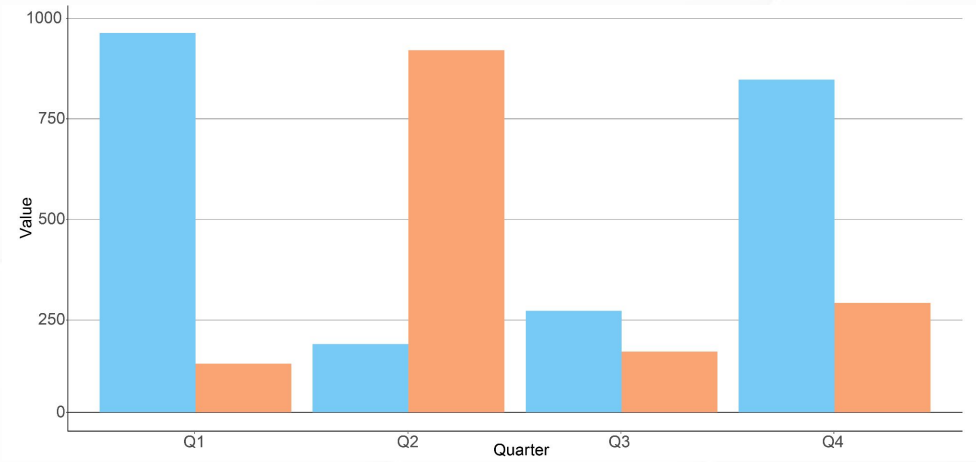
Positives

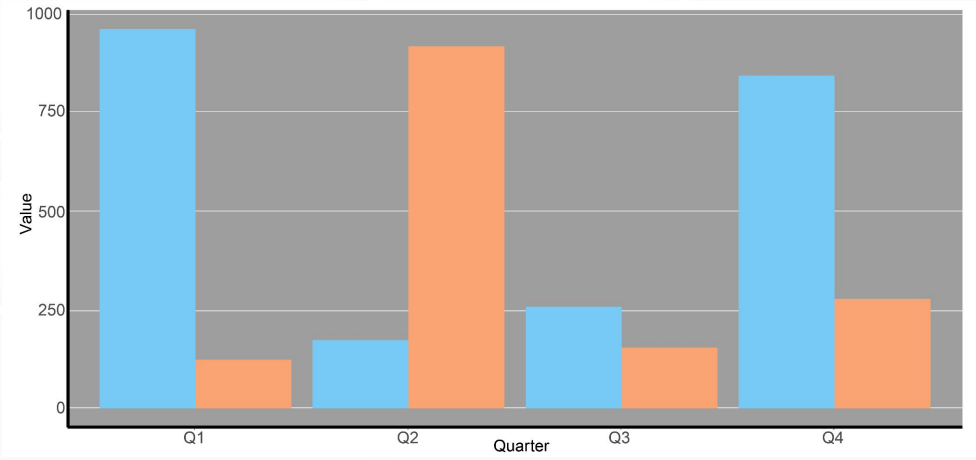
- Helps the plot to be perceived as a single visual entity
- Prevents the plot to appear as a white box in surrounding dark text

Negatives

- White-on-gray background grid isn't attractive
- The gray background can detract from the actual data
- Grid with major and minor lines can be too dense
- Gray squares in the legend confusing



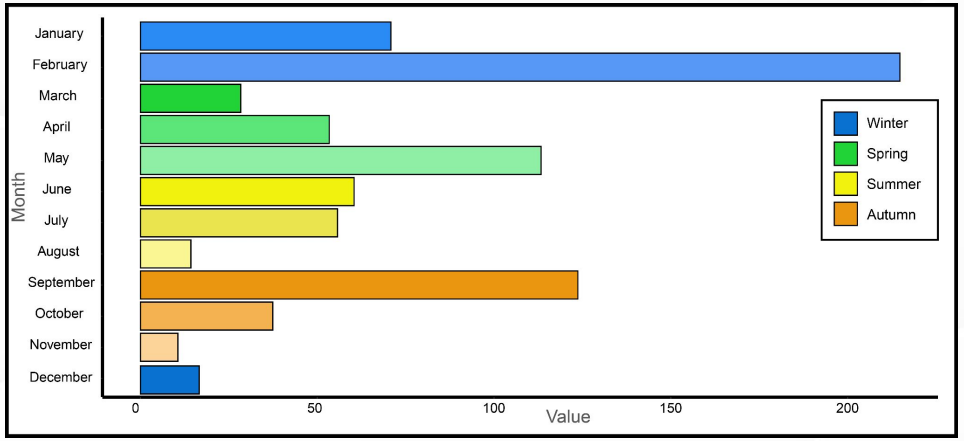


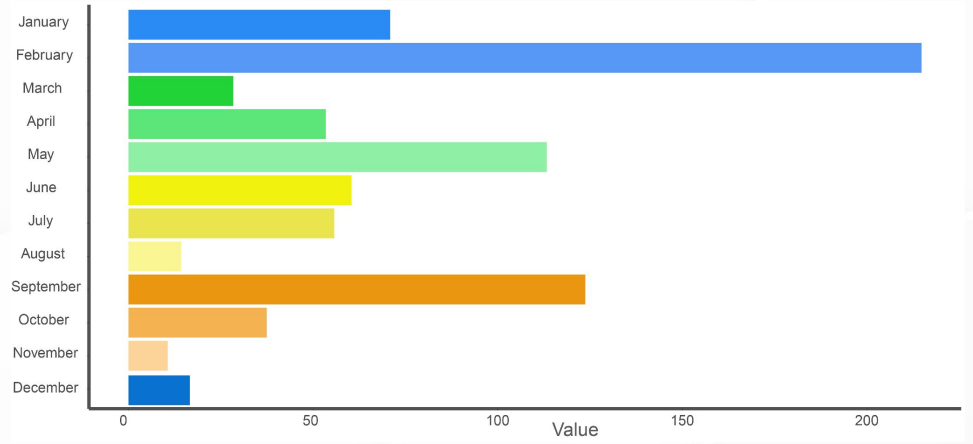


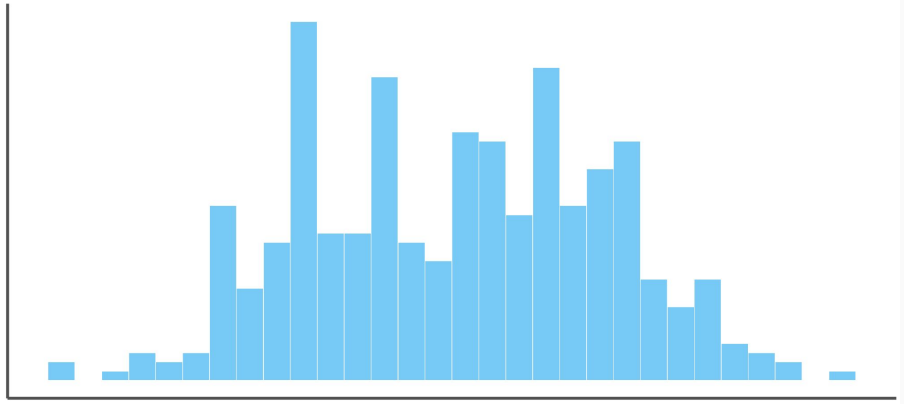
Lines in drawings

No lines to the shapes

- Perceived as one object
- Clearer indication of the size
  - Histogram and bar plots







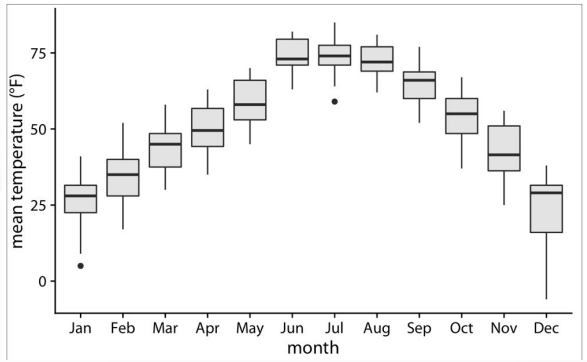
Lines in drawings

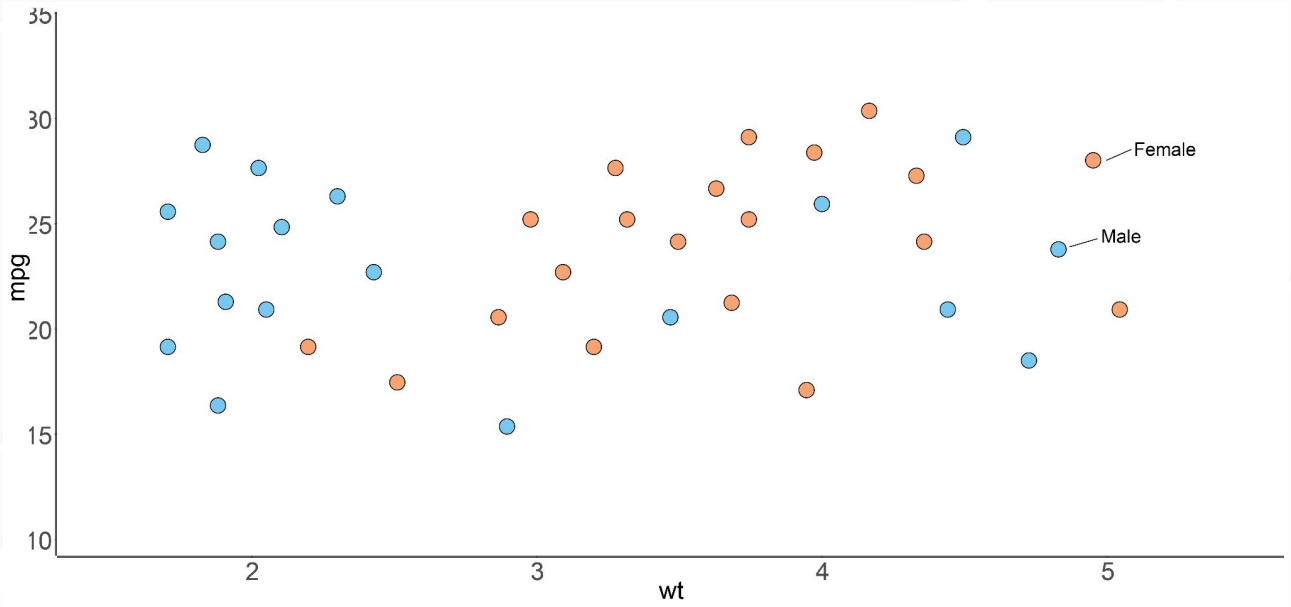
No lines to the shapes

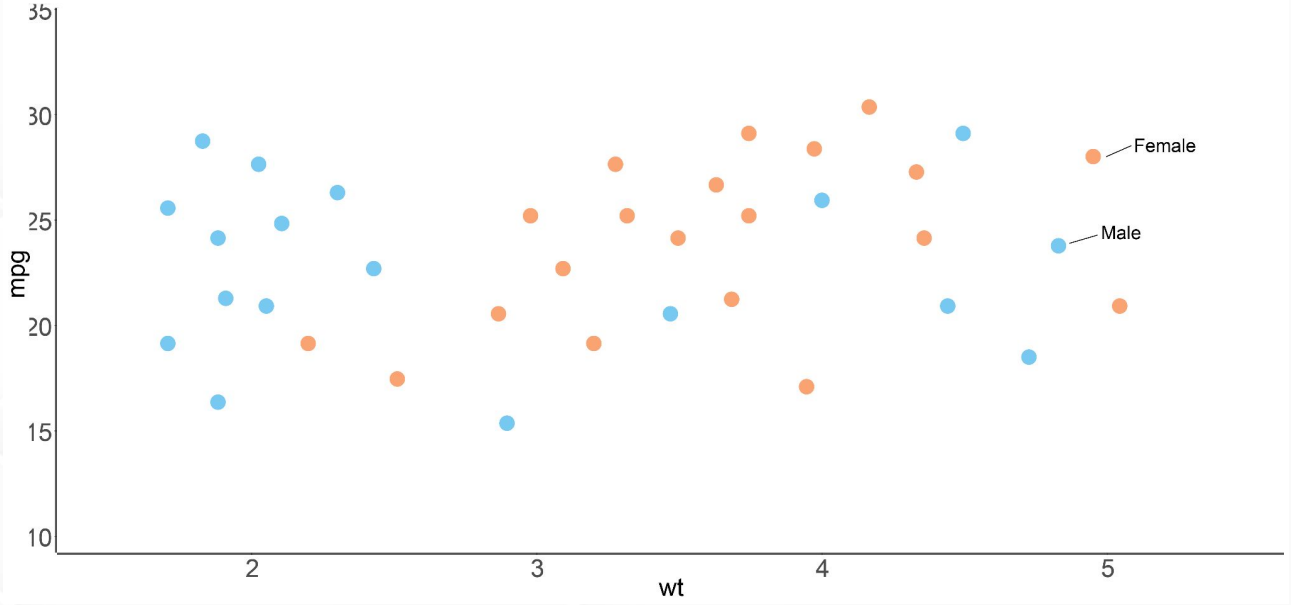
- Perceived as one object
- Clearer indication of the size
  - Histogram and bar plots
  - Symbols in dot plots
- Pleasant to look at

EXCEPTION:

- Boxplots

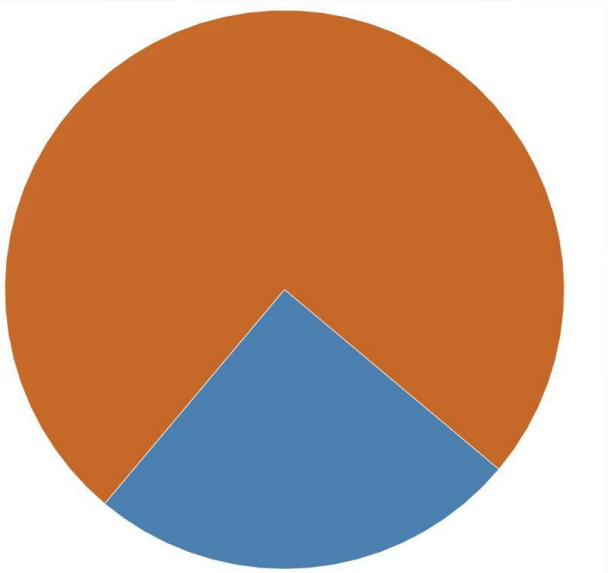


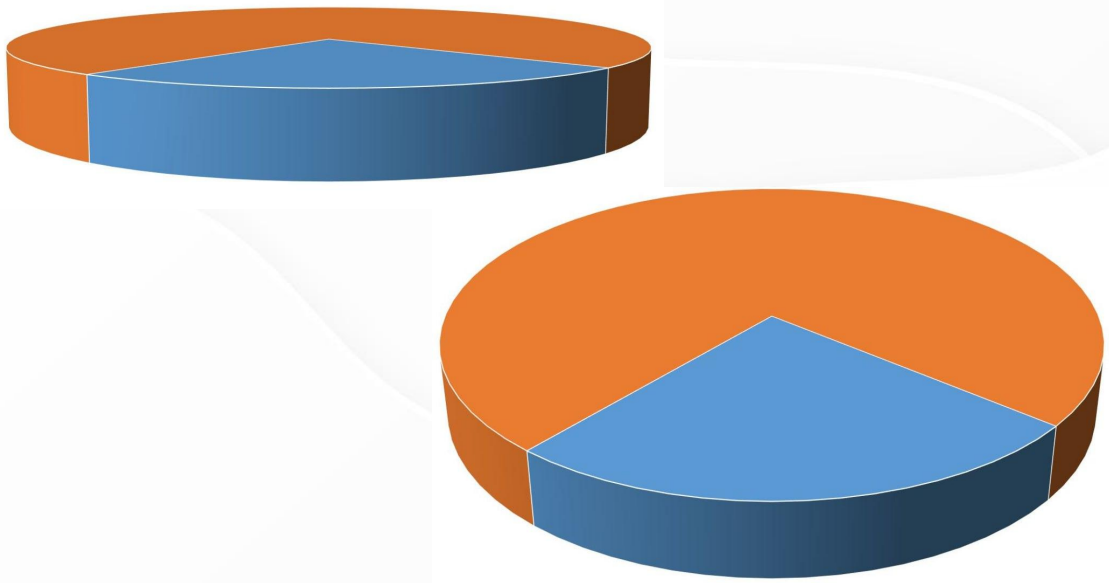


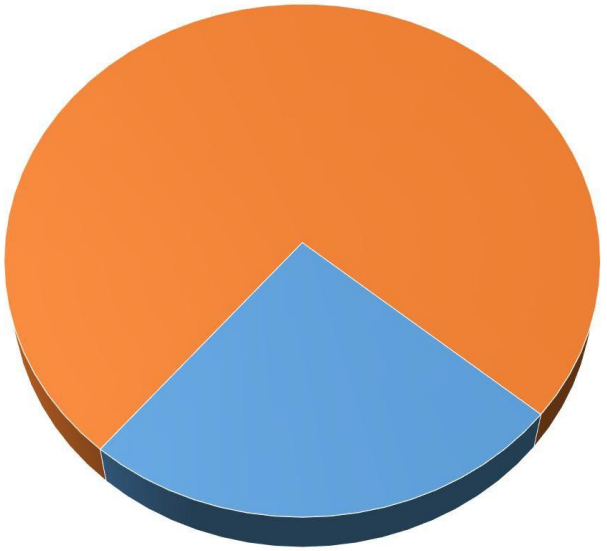


3D charts

just don’t do it! The projection of a 3D objects into two dimensions always distorts the data.







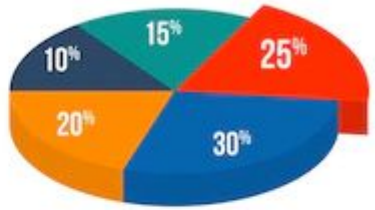
Example: Blue slice is 25% in every graph

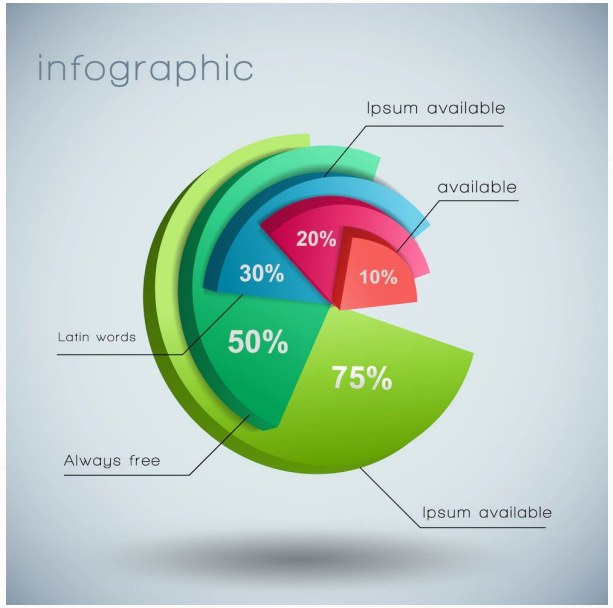
3D charts

just don’t do it!

Exceptions:

- Interactive visualisations
- VR or augmented reality environment
- Non-interactive, but showing it rotating
- Data mapped onto 3D things
  - Maps
  - Protein structure

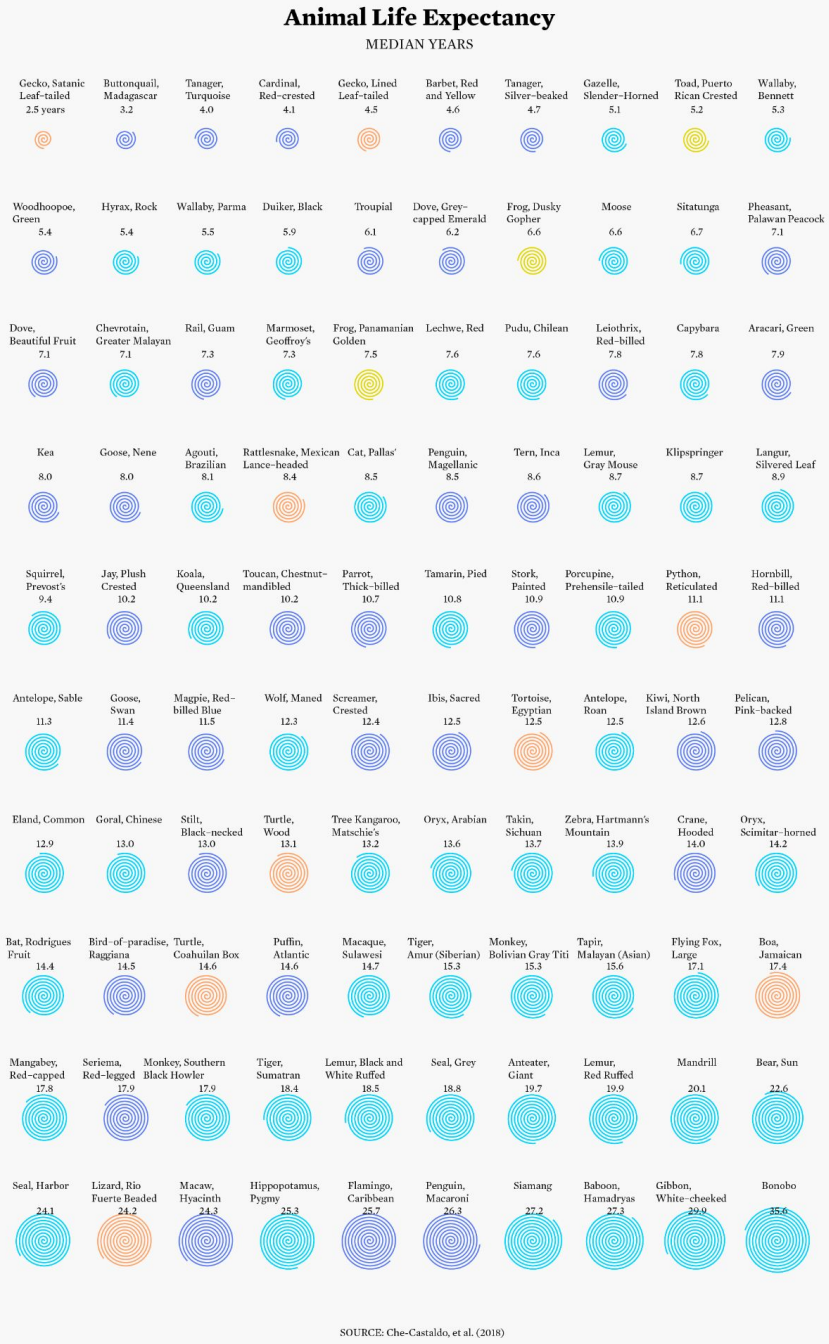




https://www.freepik.com/free-vector/charts-diagrams-graphs-modern-isometric-3d-flat-style-infographic-presentation-design-data-finance_10700767.htm
https://www.freepik.com/free-vector/3d-business-diagram-template-with-text-fields-marked-by-different-colors-chart_11408197.htm Image by macrovector on Freepik

How could this graph be better?

https://flowingdata.com/2023/05/02/one-day-chart-challenge/



Colors

Color choices

To enhance the figure and make it easier to read.

- 3-5 categories of data - qualitative color scales
- More categories - use direct labeling, matching colors becomes laborious
- Help! Hard to choose colors?
  - https://coolors.co/
  - https://www.heavy.ai/blog/12-color-palettes-for-telling-better-stories-with-your-data

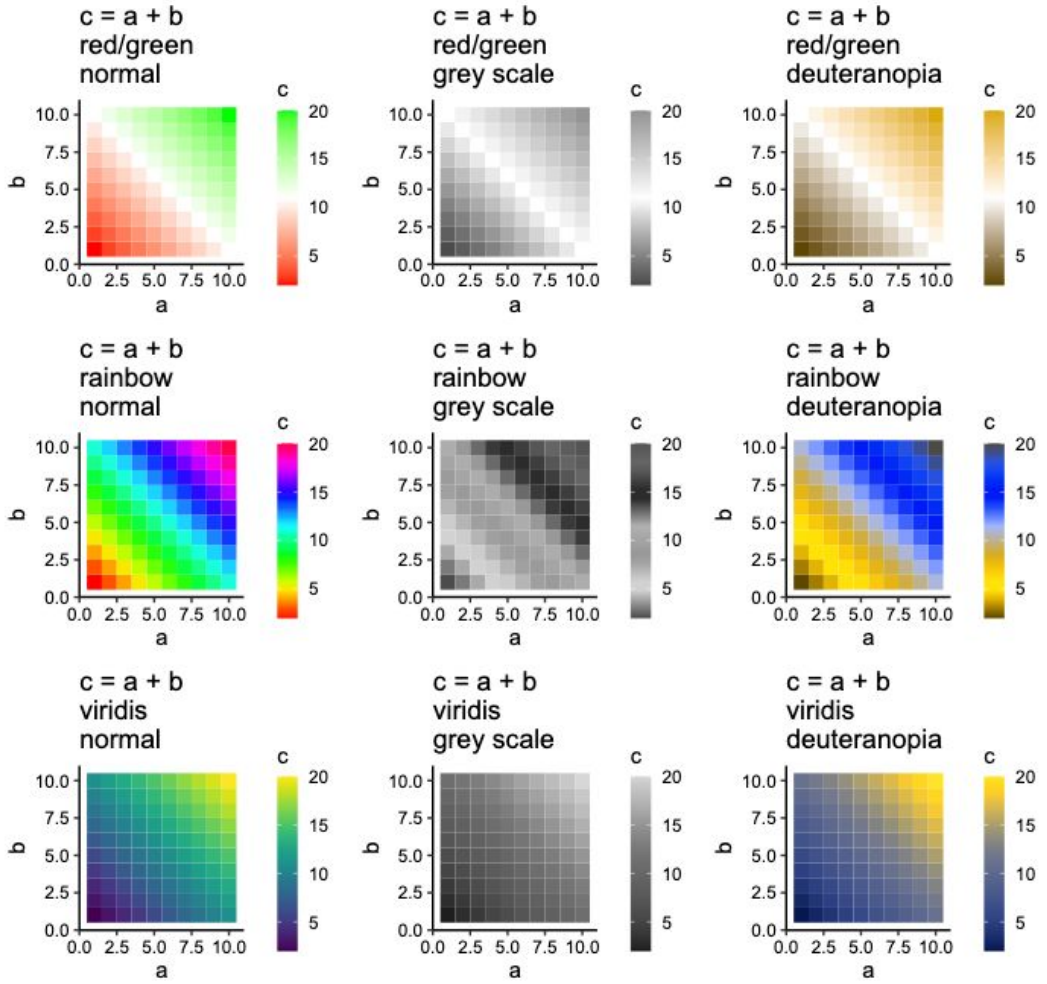
Common mistakes: Color choices

- Saturated and intense colors - hard to look at
- Using default color choices usually doesn’t work well
  - Rainbow - red-to-red
  - Scale ends needs to have maximum differences
  - Regions where colors change slowly VS rapidly

https://www.heavy.ai/blog/12-color-palettes-for-telling-better-stories-with-your-data

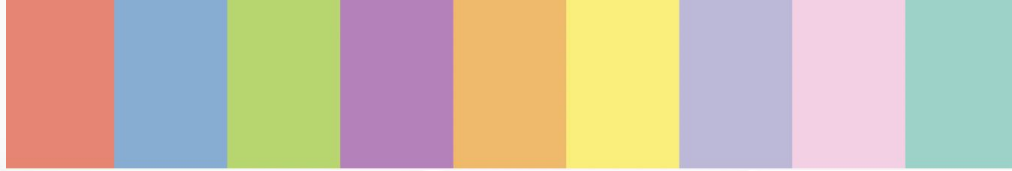
Common mistake: color blind unfriendliness

Deuteranomaly is the most common type of red/green colorblindness, occurring in 1/16 male and 1/256 female.



https://github.com/cxli233/FriendsDontLetFriends
C. Li. (2023). cxli233/FriendsDontLetFriends: FriendsDontLetFriends. Zenodo. https://doi.org/10.5281/zenodo.7097522

Color scales: Qualitative

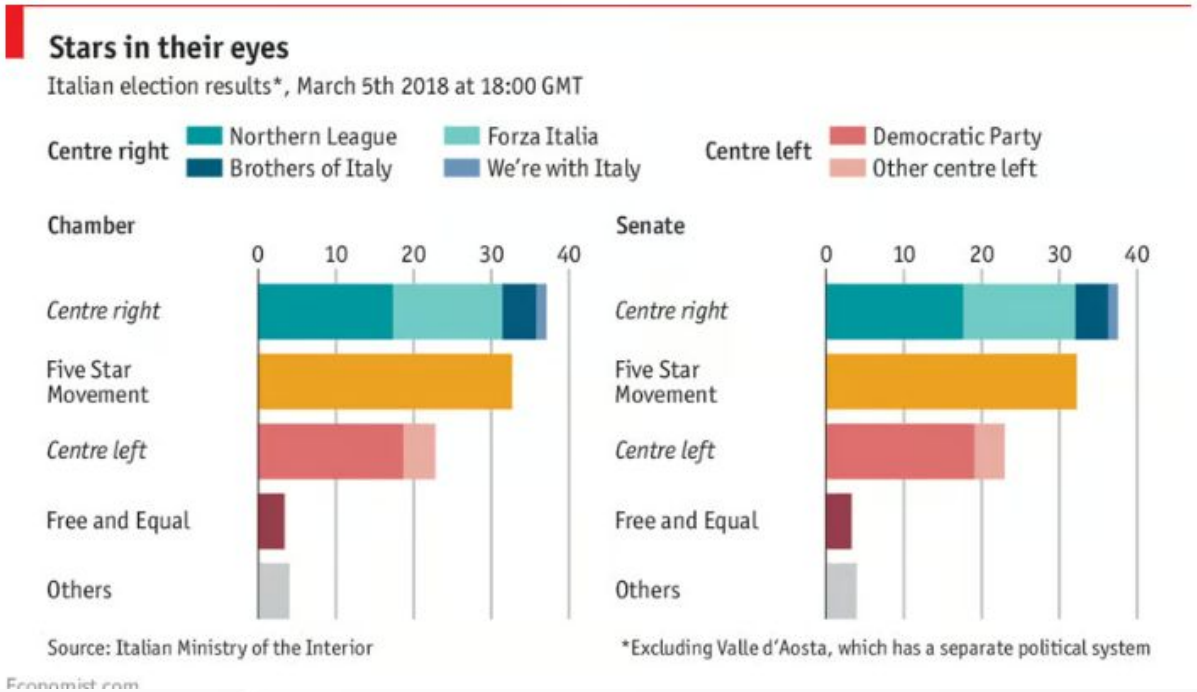


```
["#fd7f6f", "#7eb0d5", "#b2e061", "#bd7ebe", "#ffb55a", "#ffee65",
"#beb9db", "#fdcce5", "#8bd3c7"]
```

https://www.heavy.ai/blog/12-color-palettes-for-telling-better-stories-with-your-data

Distinguishing discrete items or groups that do not have intrinsic order (countries on a map)

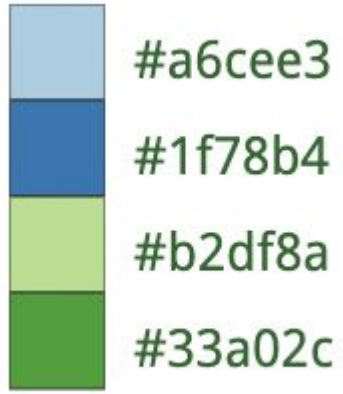
- Finite number of colors that are similar but distinct from each other.
- No color should specifically stand out from others.
- Colors should not create an apparent order.



https://www.economist.com/graphic-detail/2018/03/05/the-italian-election-in-charts

Qualitative color blind friendly

Tool:
https://davidmathlogic.com/colorblind/#%23D81B60-%231E88E5-%23FFC107-%23004D40
play around to see how different color blindness affects what we see



https://colorbrewer2.org/

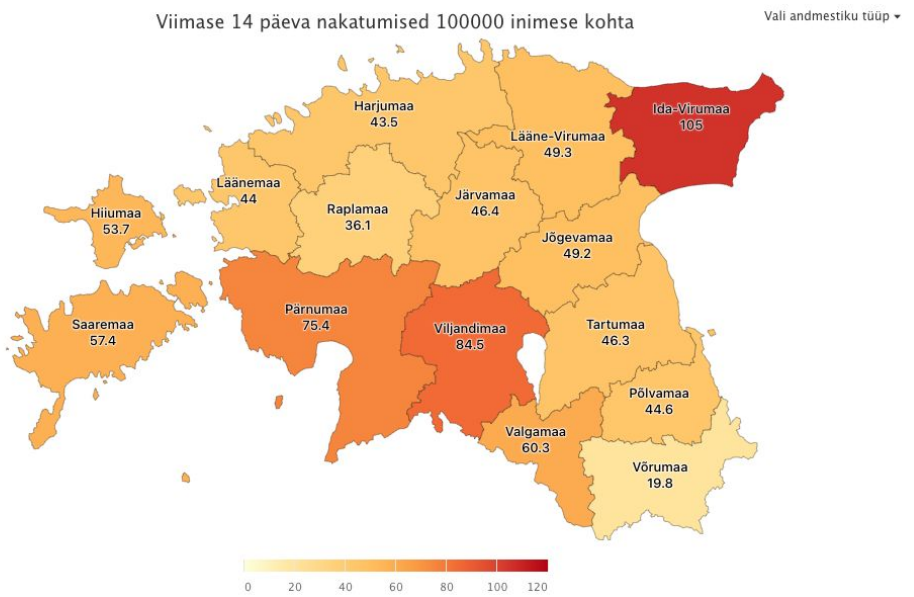
Color scales: Sequential

Representing data values

Scale of color clearly indicates order in the values of the data or how two specific values differ from each other.

Use single hue (dark blue to light blue) or multi hue scales (gradient should be seen in natural world, dark red to light yellow).

https://koroonakaart.ee/et



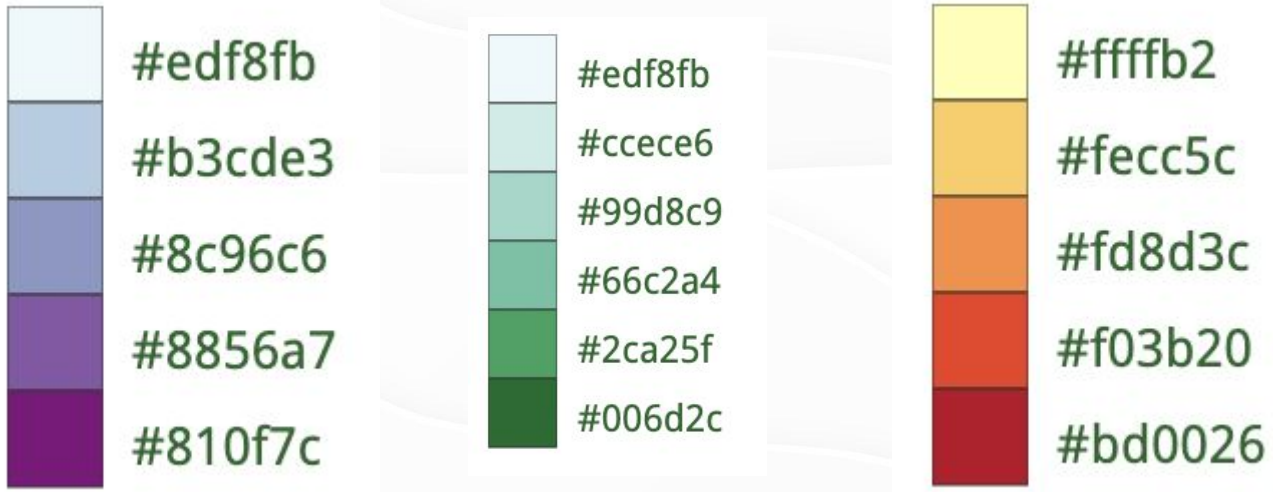
```
["#d7e1ee", "#cbd6e4", "#bfcbdb", "#b3bfd1", "#a4a2a8", "#df8879", "#c86558",
"#b04238", "#991f17"]
```

```
["#115f9a", "#1984c5", "#22a7f0", "#48b5c4", "#76c68f", "#a6d75b", "#c9e52f",
"#d0ee11", "#d0f400"]
```

https://www.heavy.ai/blog/12-color-palettes-for-telling-better-stories-with-your-data

Color scales:
Sequential color-blind friendly

sequential scale should present a continuous gradient from dark to light colors and should pose no problems.

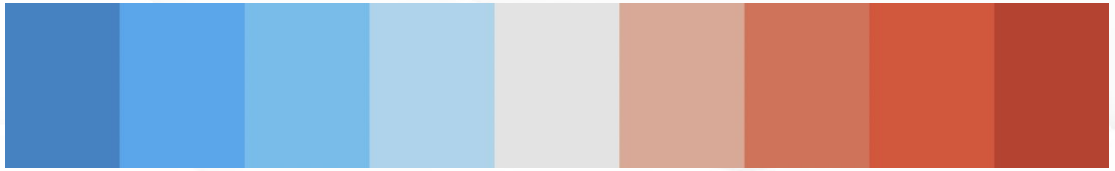


https://colorbrewer2.org/

Color scales: Diverging

Two sequential scales stitched together at a common midpoint, which usually is represented by a light color.

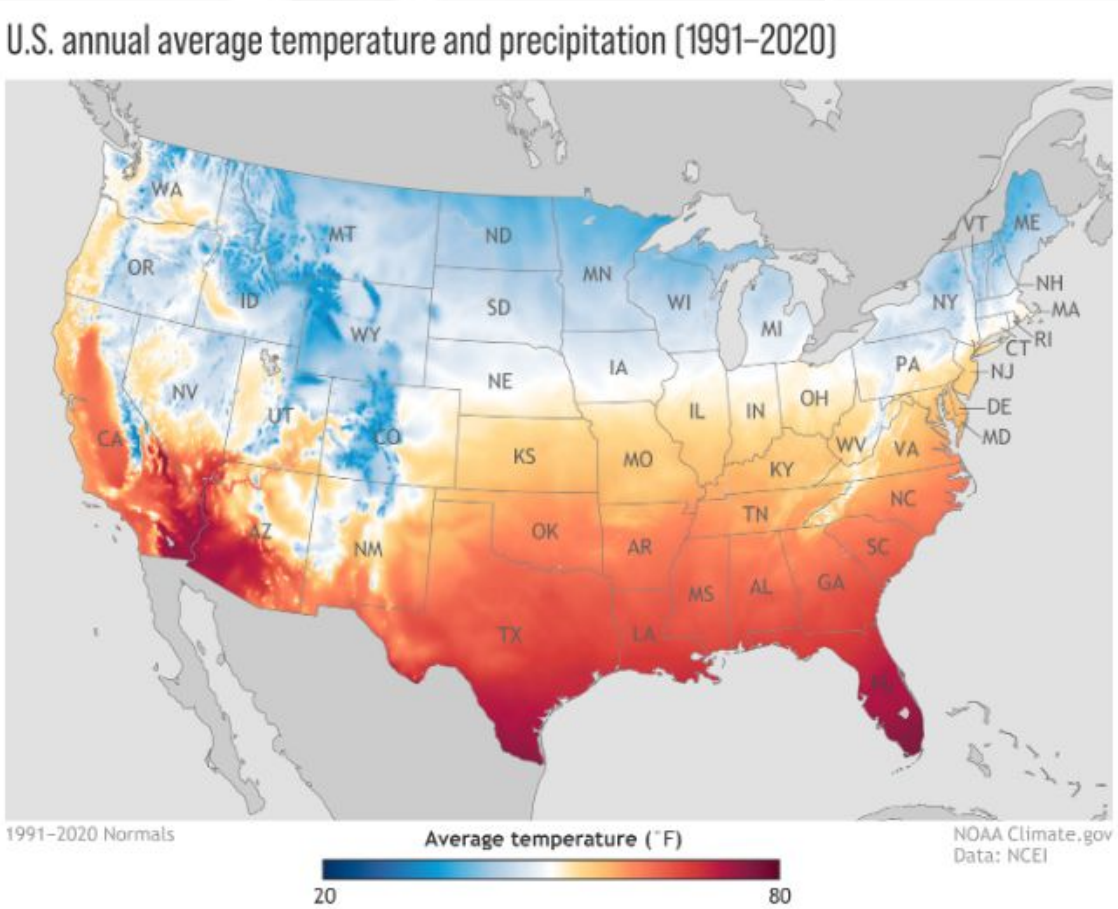
Needs to be balanced.

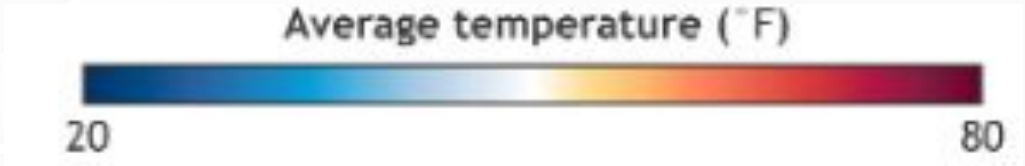


["#1984c5", "#22a7f0", "#63bff0", "#a7d5ed", "#e2e2e2", "#e1a692", "#de6e56", "#e14b31", "#c23728"]

["#54bebe", "#76c8c8", "#98d1d1", "#badbdb", "#dedad2", "#e4bcad", "#df979e", "#d7658b", "#c80064"]

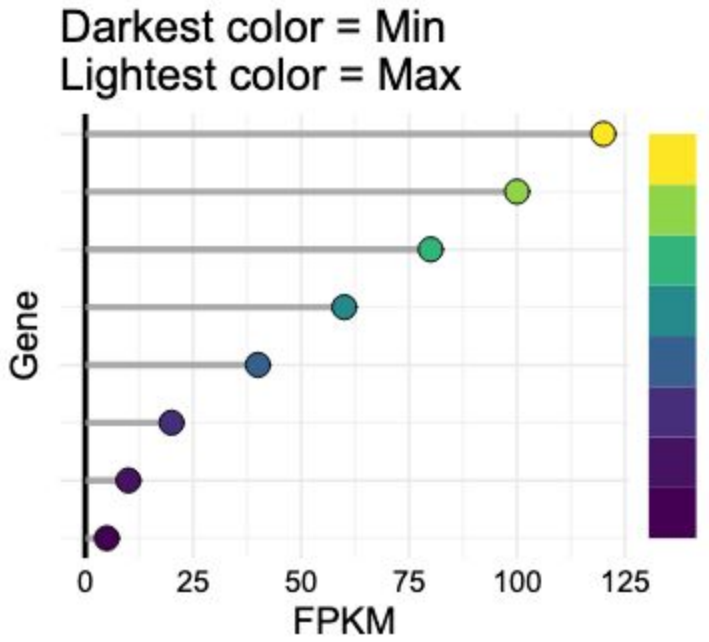
https://www.heavy.ai/blog/12-color-palettes-for-telling-better-stories-with-your-data



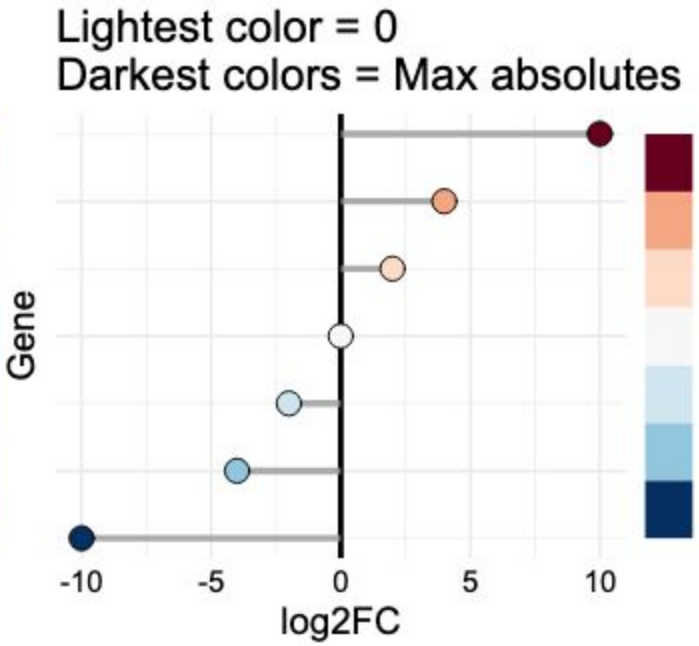


https://www.climate.gov/news-features/featured-images/new-maps-annual-average-temperature-and-precipitation-us-climate

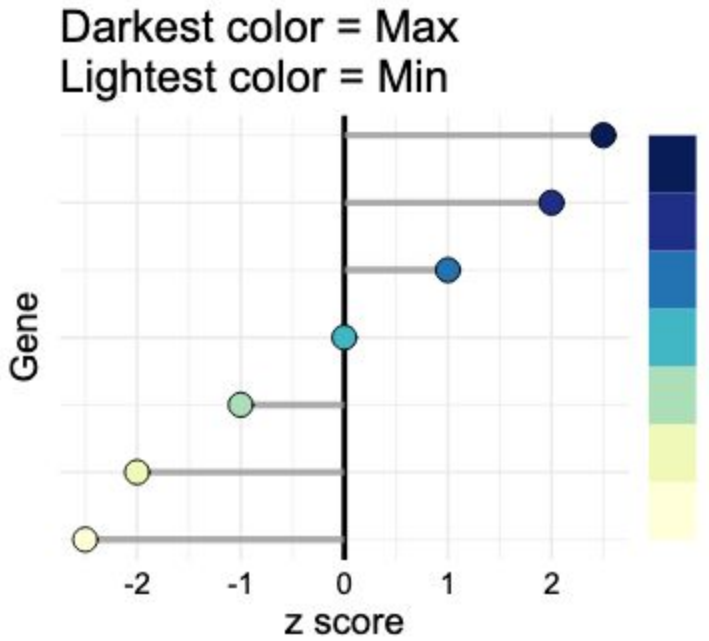
Common mistake: Diverging scale for unidirectional data



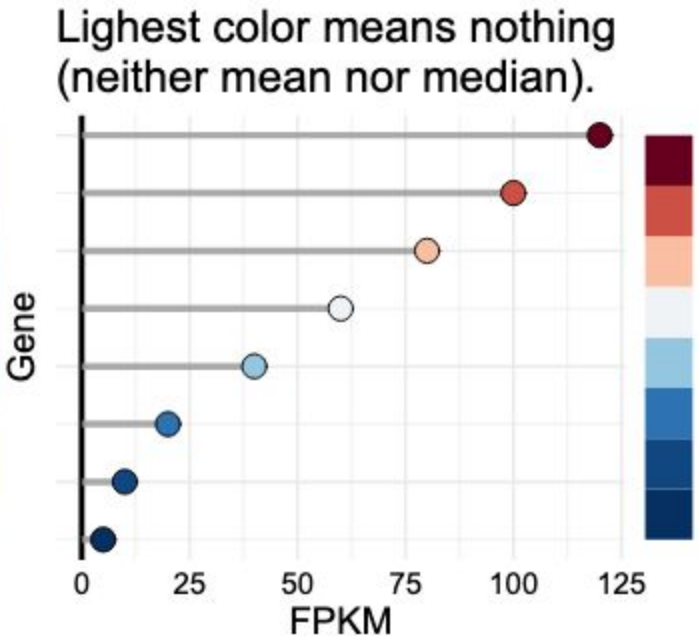
This is good.



This is good.



This is good.

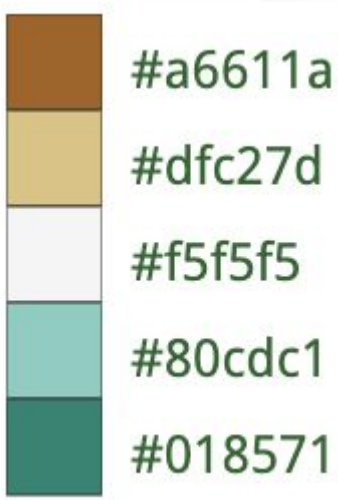


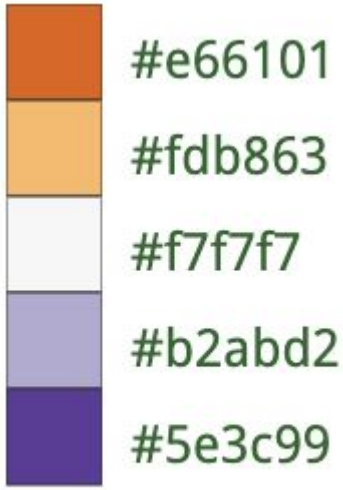
A data visualization sin

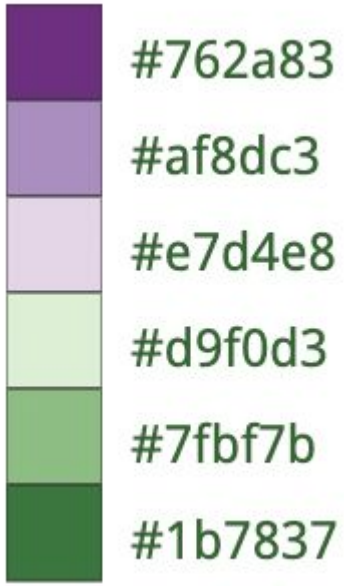
https://github.com/cxli233/FriendsDontLetFriends
C. Li. (2023). cxli233/FriendsDontLetFriends: FriendsDontLetFriends (v3) (Version v3). Zenodo. https://doi.org/10.5281/zenodo.7097522

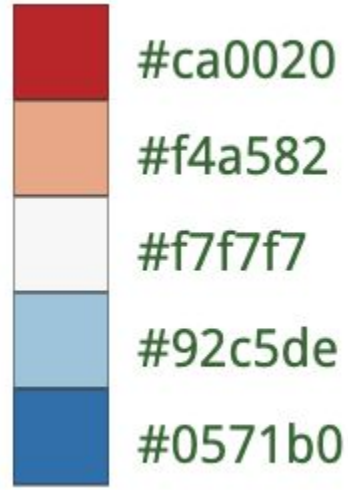
Color scales: Diverging color blind friendly

Slightly problematic, popular color contrasts can become indistinguishable (green and red).





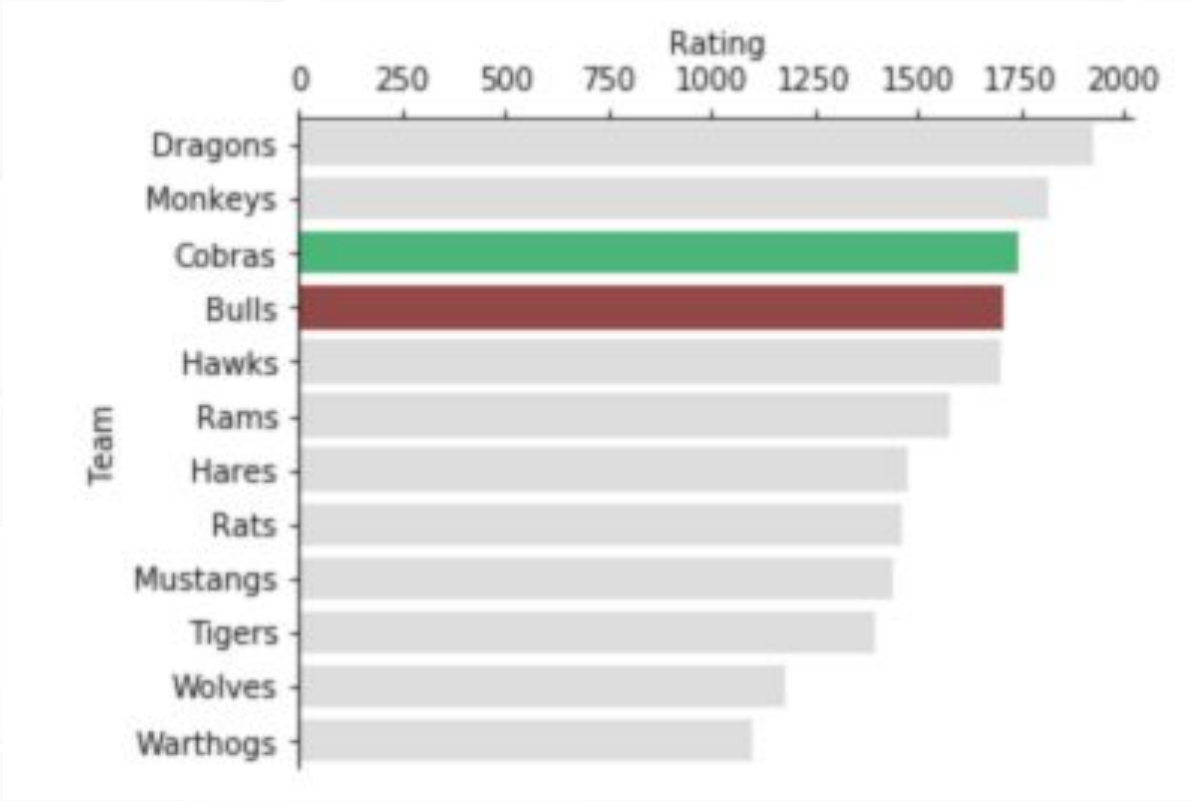




https://colorbrewer2.org/

Color scales: Accent

Highlighting a specific data point/element.



https://chartio.com/learn/charts/bar-chart-complete-guide/

Improving the visual design of the figure

Color is easier to distinguish when applies to larger areas than to small ones or thin lines.

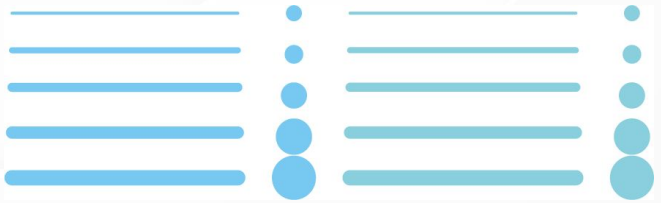


Tableau public visualisations

- https://public.tableau.com/app/profile/tervise.arengu.instituut/viz/Snnid/Sisukord Birth statistics in Estonia by National Institute for Health Development

- https://public.tableau.com/app/profile/tervise.arengu.instituut/viz/Imikuterinnapiimagatoitmine/Imikuterinnapiimagatoitmine Breastfeeding by National Institute for Health Development

- https://public.tableau.com/app/profile/transpordiamet/viz/Kergliikuri_L/nnetusteldandmed traffic accidents with electric scooter rider by Estonian Transport Administration

Tools for visualisation

Logo	Name	Free plan	Premium plans user/ month	Notes	need to code?
Microsoft | Power BI	Microsoft Power BI	+	10$- 20$	Slow with big data. Integrated with other Microsoft products. User friendly UI.	no
tableau	Tableau	+	75$	Public is free but data will be public too. Apply for Tableau plans for student, teacher or data kids to get it for free (time limited). Higher learning curve. Exceptional visualisation options	only for very specific details
X	Microsoft Excel	-	5,60$-20,60$	Trial version is available. Easy to learn.	no
RStudio	RStudio	+	87$	ggplot2 package is free and designed for data visualisation. Popular among scientists. Can handle large datasets. Reproducible. Exceptional visualisation options	always

Take away message

- Figure out your data type
- Keep in mind your publication venue and audience
- Think of your main message
- Choose the right graph (using slides, flowchart guide or https://www.data-to-viz.com/)
- Avoid common mistakes
- Use best practices
- Focus on the important - maximise data-ink ratio

Take away message

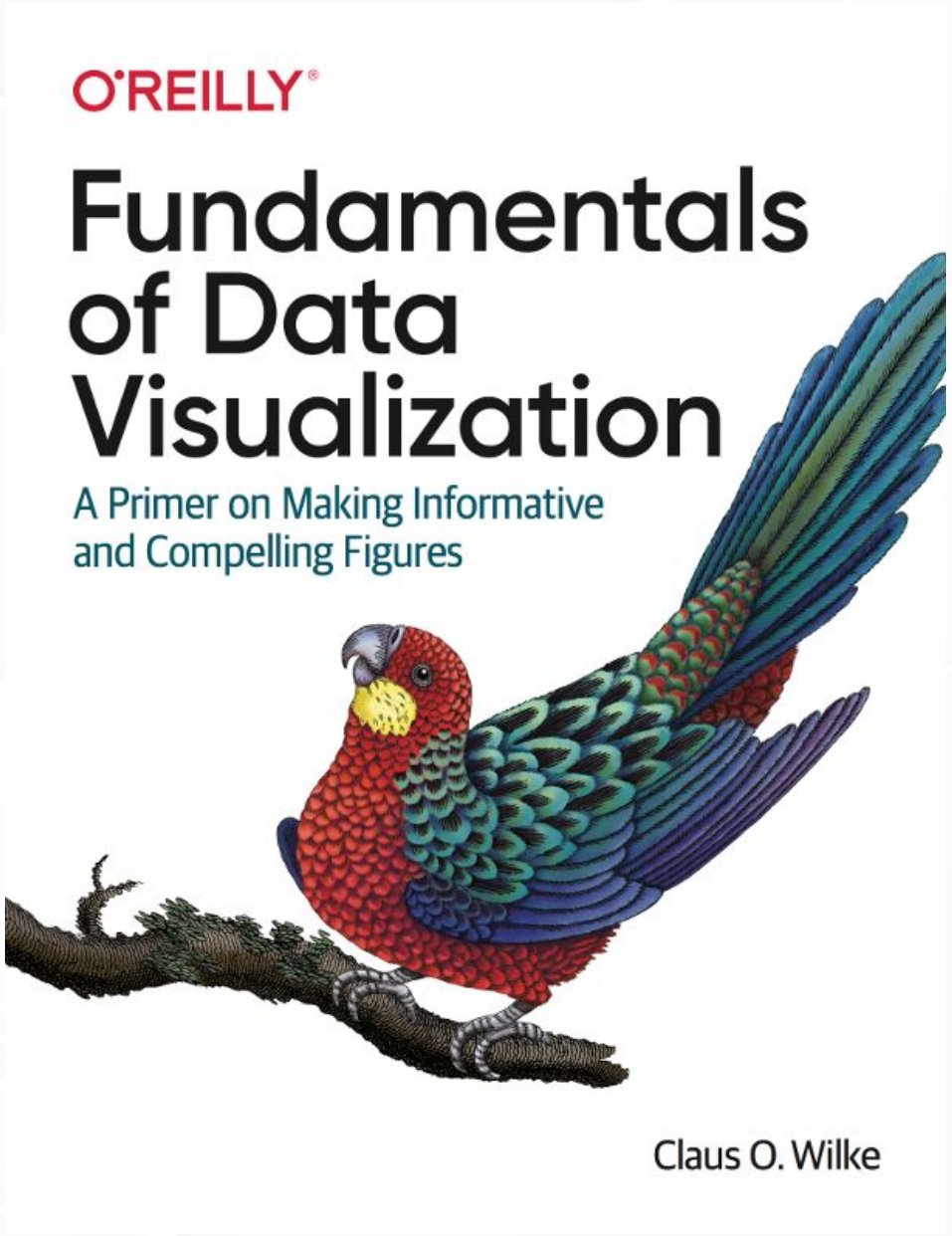
- Bar chart starts at 0
- Order the variables if possible
- No more than 5 sections for vertical bar or pie charts
- Use consistent visual language
  - 1-2 fonts (similar type)
  - Make titles always bigger
  - Colors - easy on the eye and compliment each other
    - Pay attention to color blindness!
  - Play with line thickness
  - If there is clear visual order in your data, match the legend with it

Feedback

https://forms.gle/tTAqavvdfAaJNqVf6

Reference

Materials based on
Claus O. Wilke. (2019). Fundamentals of Data Visualization. O'Reilly Media, 319 p. (NB! Draft is freely available at https://serialmentor.com/dataviz/)



Resource

- White, T. (2017). Symbolization and the Visual Variables. The Geographic Information Science & Technology Body of Knowledge (2nd Quarter 2017 Edition), John P. Wilson (ed.).
  DOI: 10.22224/gistbok/2017.2.3
  http://dx.doi.org/10.22224/gistbok/2017.2.3
- https://www.data-to-viz.com/ wonderful resource for choosing the right graph
- https://www.storytellingwithdata.com/chart-guide
  - and the book https://github.com/Saurav6789/Books-/blob/master/storytelling-with-data-cole-nussbaumer-knaflic.pdf

References

General information about data visualisation

- Using Design Techniques for Clear and Appealing Data Visualisation by nullQueries https://www.youtube.com/watch?v=0Smgm2UTUSo
- How To Choose The Right Graph by UNDATABLE https://www.youtube.com/watch?v=o7F-tbBl_hA
- How To Use COLOR in Your Data Visualisation by UNDATABLE https://www.youtube.com/watch?v=v5brQ4WTlmQ
- Data Visualisation in 2022 by Visme https://www.youtube.com/watch?v=loYuxWSsLNc
- Which is the best Chart by 365 Data Science https://www.youtube.com/watch?v=C07k0euBpr8
- Five Data Storytelling Tips to Improve Your Charts and Graphs by VIsme https://www.youtube.com/watch?v=4pymfPHQ6SA

References

- https://datavizproject.com/
- https://100.datavizproject.com/
- https://data.europa.eu/apps/data-visualisation-guide/

Color

- Color Contrast Checker https://coolors.co/contrast-checker/112a46-acc8e5
- Color Palette Generator https://coolors.co/
- Color Scheme Designer https://paletton.com/#uid=1000u0kllllaFw0g0qFqFg0w0aF
- https://color.adobe.com/create/color-contrast-analyzer - color blind safe tool here

Genomic Data Visualization

https://jef.works/genomic-data-visualization-2024/course

vabalt kättesaadavad materjalid

Photo resources

- https://www.freepik.com/free-photo/digital-graph-with-businessman-hand-overlay_15474099.htm title slide
- https://www.freepik.com/free-vector/white-elegant-texture-background-design_6764485.htm slide background

Thank you for listening!