Data Visualization

COSC / DATA 301

MUCH OF THE WORK IN THIS COURSE HAS BEEN DEVELOPPED BY PRECEDING COLLEAGUES

Announcements

Prepare for next week by requesting your free Tableau for Students License

Heads up:

- Live lecture cancelled next week.
- Approaching “end of Term crunch”! Stay on top of your deadlines!

Why learn Visualization?

Visualization allows people to understand and extract information faster and with more accuracy than displaying text and numbers.

A good visualization makes data more understandable and reachable to more people.

High quality visualization encourages confidence in the data analysis and inspires people to utilize the data more effectively.

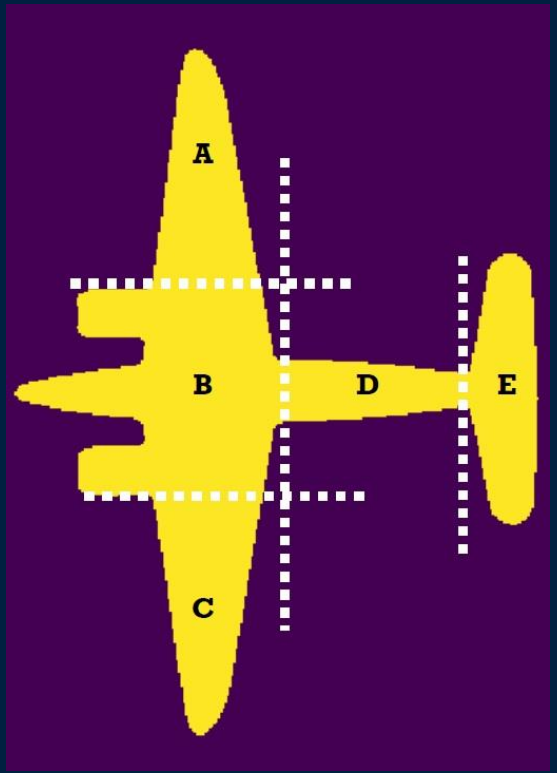
Motivating the need for EDA

Case Study: Planes in WW2

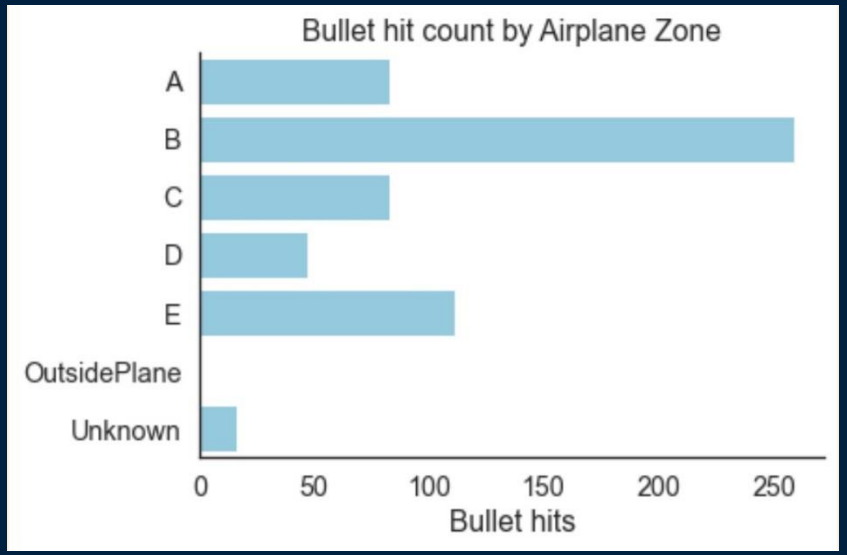
You have been given a dataset and tasked with trying to solve a problem. In WW2, expensive fighter planes were going down quite frequently due to bullet fire. The military decided to conduct an analysis and surveyed all the surviving planes in an effort to catalogue which regions of the plane should be reinforced.

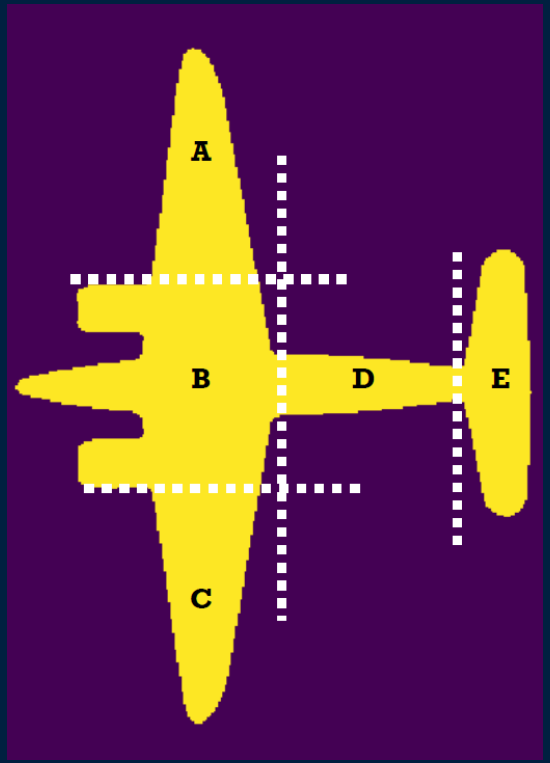
You're given a schematic of the plane. It is divided in 5 regions. A value of 1 indicates a bullet hole across all planes that returned. Areas without bullet holes are marked as 0.

You receive a csv file with the data (bullet_data.csv).



Case Study: Planes in WW2





So…

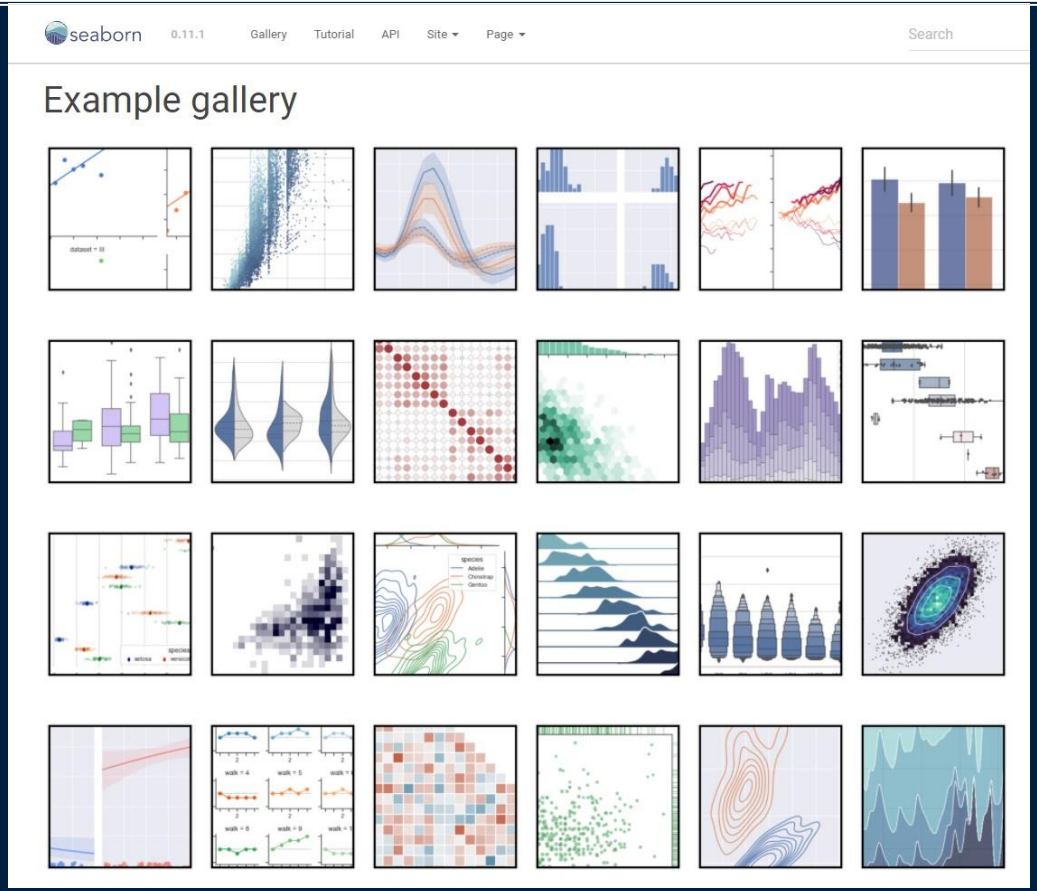
Look at your data

Talk to someone about your data

Look at your data another way

Think about your data and what it means!

Seaborn Gallery

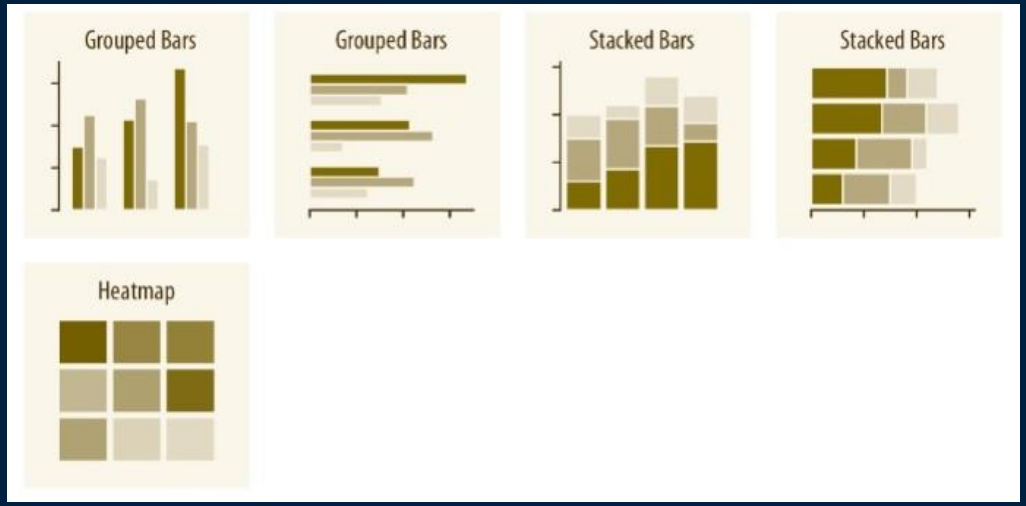


Source: https://seaborn.pydata.org/examples/index.html

Choosing an appropriate data visualization

Amounts:

- Most common approach is using bars, either vertically or horizontally arranged.
- Other option, place dots at the location where the bar would end.
- If two or more categories, we can group or stack the bars, map them on x,y axis, show amounts by colors, or via a heat map



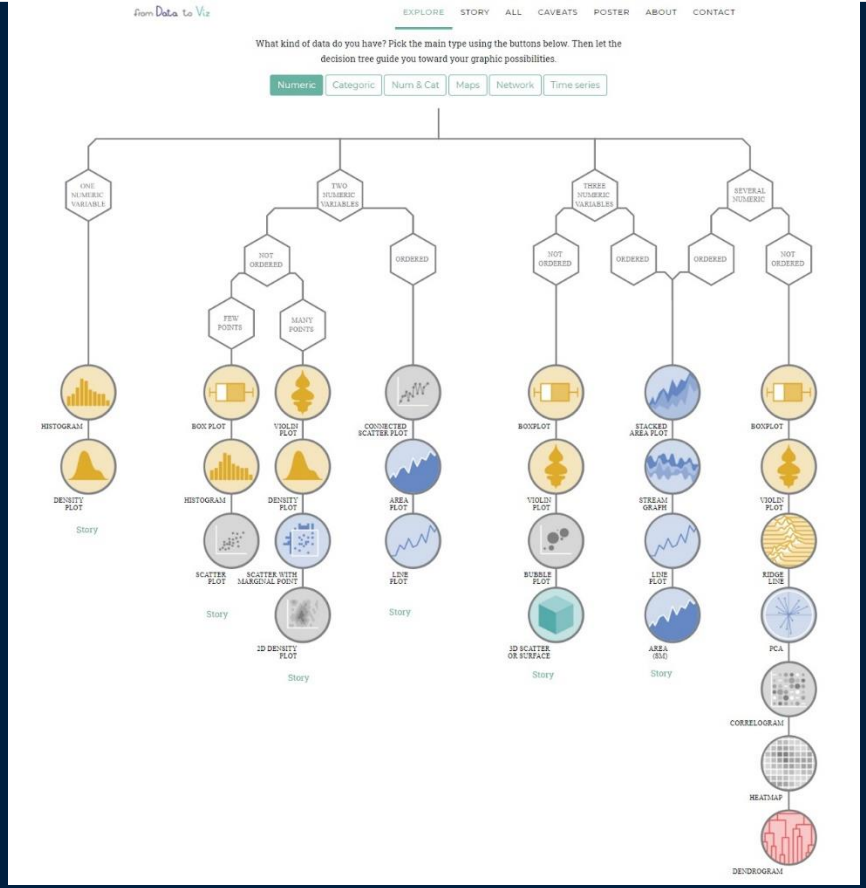
Source: https://clauswilke.com/dataviz/directory-of-visualizations.html

Directory of visualizations

From Data to Visualization:

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

It provides a decision tree based on input data format. This tree leads to twenty formats representing the most common dataset types. For each, an example of analysis based on real-life data is provided.



Principles of Effective Visualizations

Principle	Definition	Examples
Proportional Ink	The amount of ink used to indicate a value should be proportional to the value itself.	Truncating the y-axis on a bar chart exaggerates the difference between bars and violates this principle.
Data: ink ratio	Remove distracting visual elements to focus attention on the data.	Lighten line weights, remove backgrounds, avoid unnecessary/redundant labels.
Labels & Legends	Use axes labels and titles to highlight/communicate data.	Never leave your data column names and axes! Generally good to add a title
Overplotting	With large data sets, points overlap, resulting in large clouds of data.	To fix overplotting, could plot just a sample subset of the data, use alpha, and uses smaller points. Or jitter – but check if appropriate!
Visualization choice	Must be informed by the **data** you have, the **research question** being asked and the **audience** that cares.	Pick the simplest plot that best shows most/all the data needed to answer the research question. If you have summary statistics, you cannot show the distributions. Tailor the visualization to your audience (within reason) but don't dumb it down.
Colour & Accessibility	Colour can be used to encode information or for aesthetic/style/design. However, colour can also distract if used inappropriately or poorly.	Choose a perceptually uniform colour palette; can be sequential or diverging for quantitative data. Opt for color-blind friendly palettes. Categorical data can use qualitative colour schemes.

Revisiting a principle…

Principle	Definition	Examples
Visualization choice	Must be informed by: 1. The **data** you have, 2. The **research question** being asked and 3. The **audience** that cares	- Summary statistics >> do not show distribution - Pick the simplest plot that best shows most/all of the data needed to answer the research question - Tailor the visualization to your audience (within reason)

Select a plot

Use Zoom stamps to select the **“best”** ✓

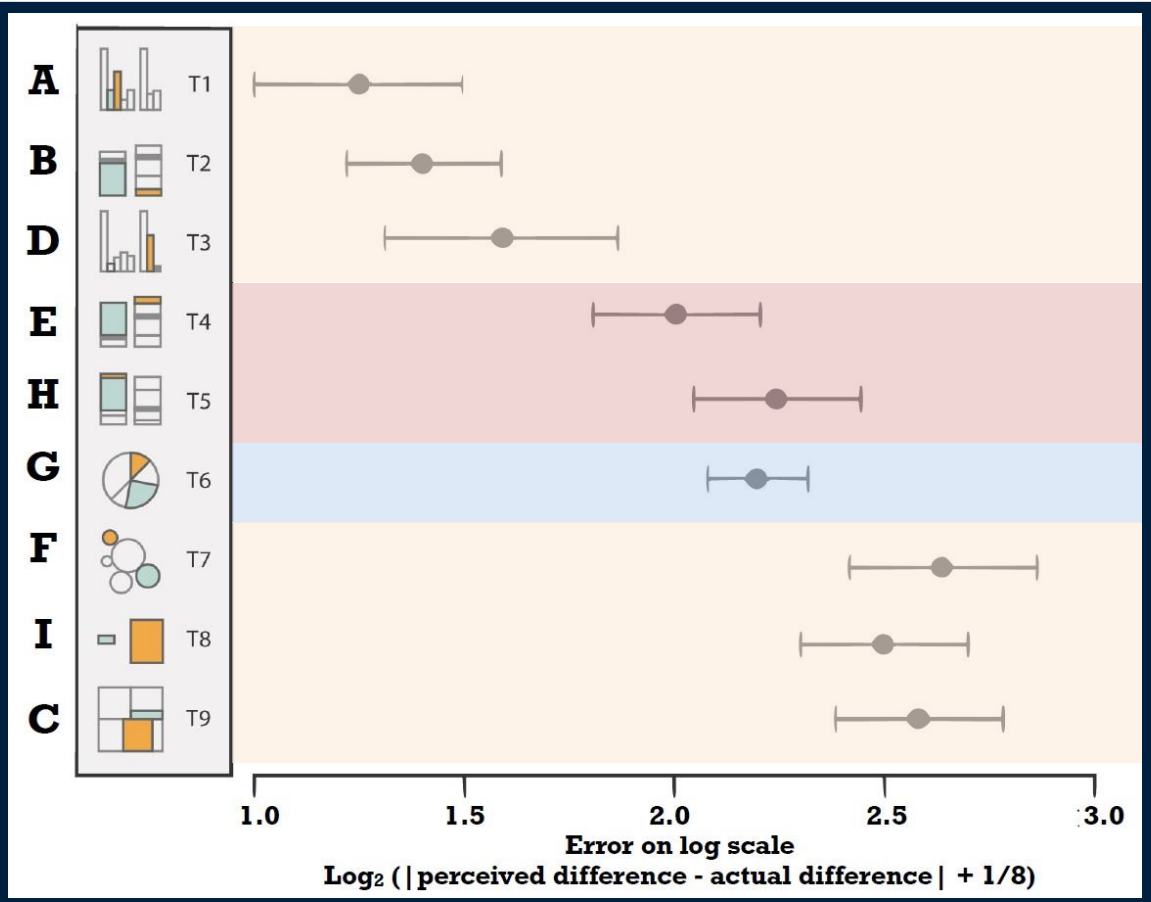
and the **“worst”** ✗

plot that :

quantify the difference between the blue and orange regions

Sources: Heer & Bostock, Cleverland & McGill, and Tamara Munzner’s textbook

Error on a log scale



Position on a scale (y-axis)

Length/Height comparison

Angles/curves

Areas

Sources: Heer & Bostock, Cleverland & McGill, and Tamara Munzner's textbook

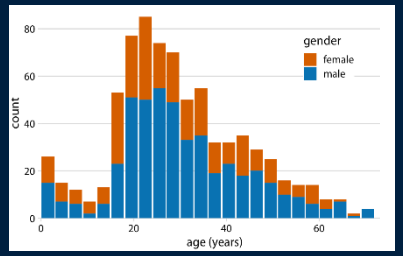
Select a plot

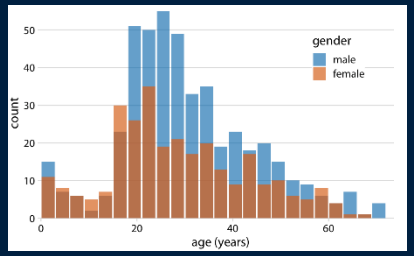
Use Zoom stamps to select the **“best”** ✓

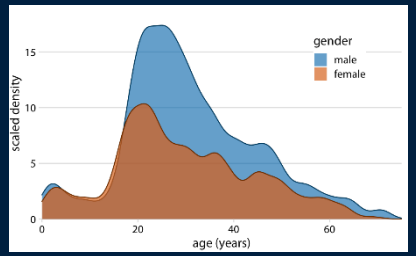
and the **“worst”** ✗

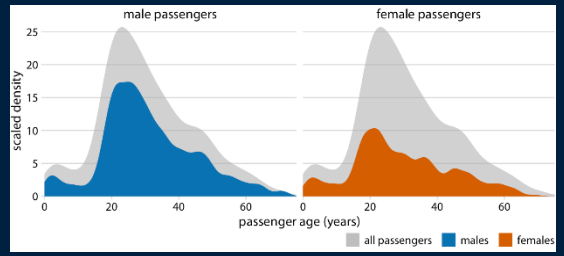
plot that :

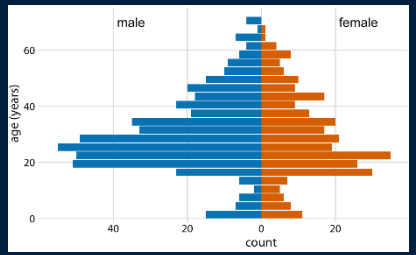
quantify the difference between the blue and orange regions

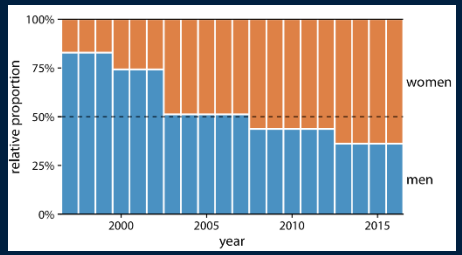








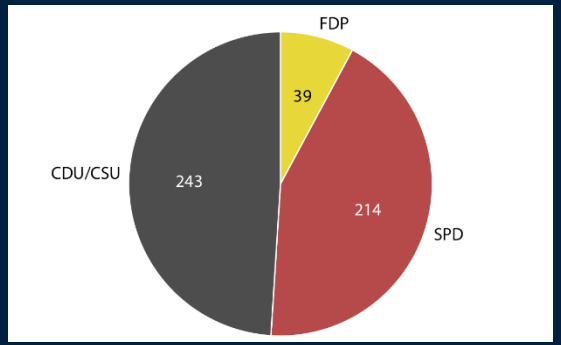


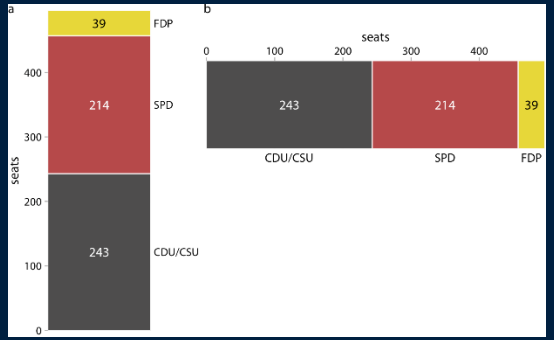


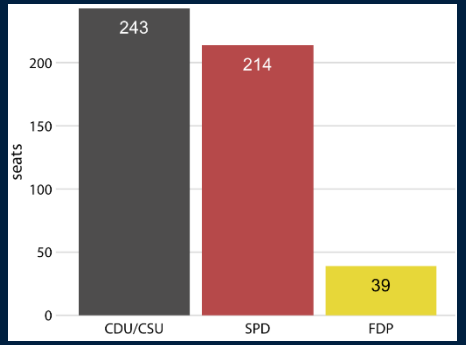
https://clauswilke.com/dataviz/histograms-density-plots.html#multiple-histograms-densities
https://clauswilke.com/dataviz/visualizing-proportions.html#stacked-densities

What does each plot highlights?

From 1961 to 1983, the German parliament (called the Bundestag) was composed of members of three different parties, CDU/CSU, SPD, and FDP. During most of this time, CDU/CSU and SPD had approximately comparable numbers of seats, while the FDP typically held only a small fraction of seats. For example, in the 8th Bundestag, from 1976–1980, the CDU/CSU held 243 seats, SPD 214, and FDP 39, for a total of 496.

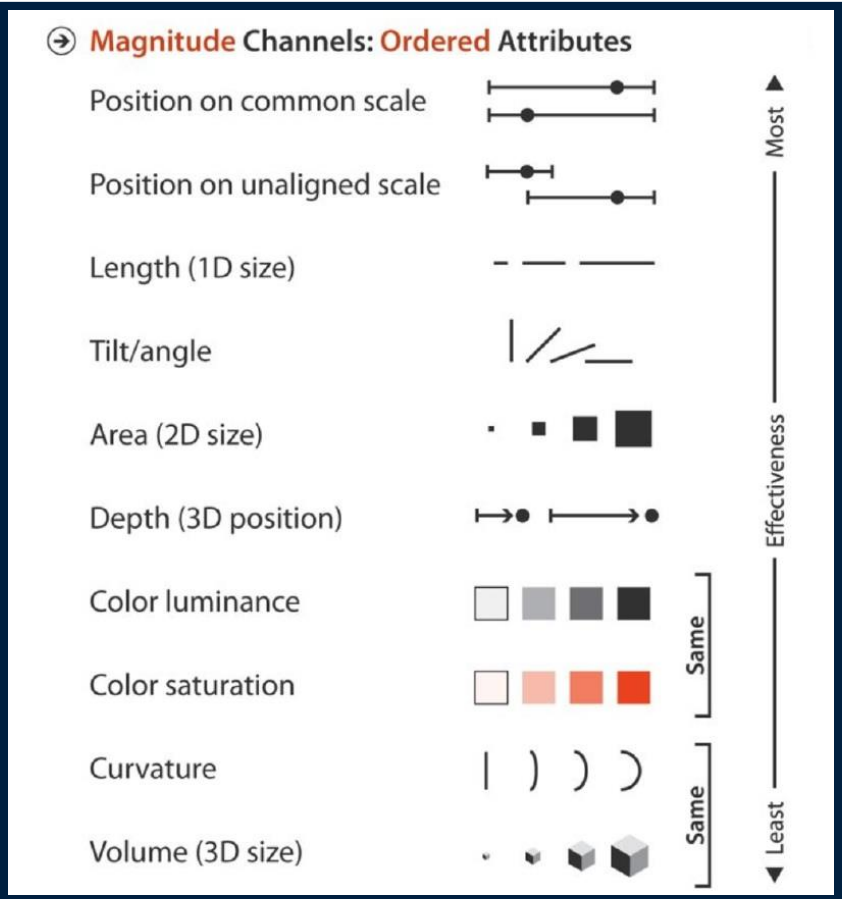






https://clauswilke.com/dataviz/visualizing-proportions.html

Channels: Expressiveness Types and Ranks



Source: chapter 5, Tamara Munzner's textbook

Judicious use of colours

Use of colours and shading is essential in most (if not all) visualizations

HOWEVER

Appropriate colours and schemes must be used to retain plot effectiveness

Accessibility: colour vision deficiencies affects 1 in 12 men and 1 in 200 women

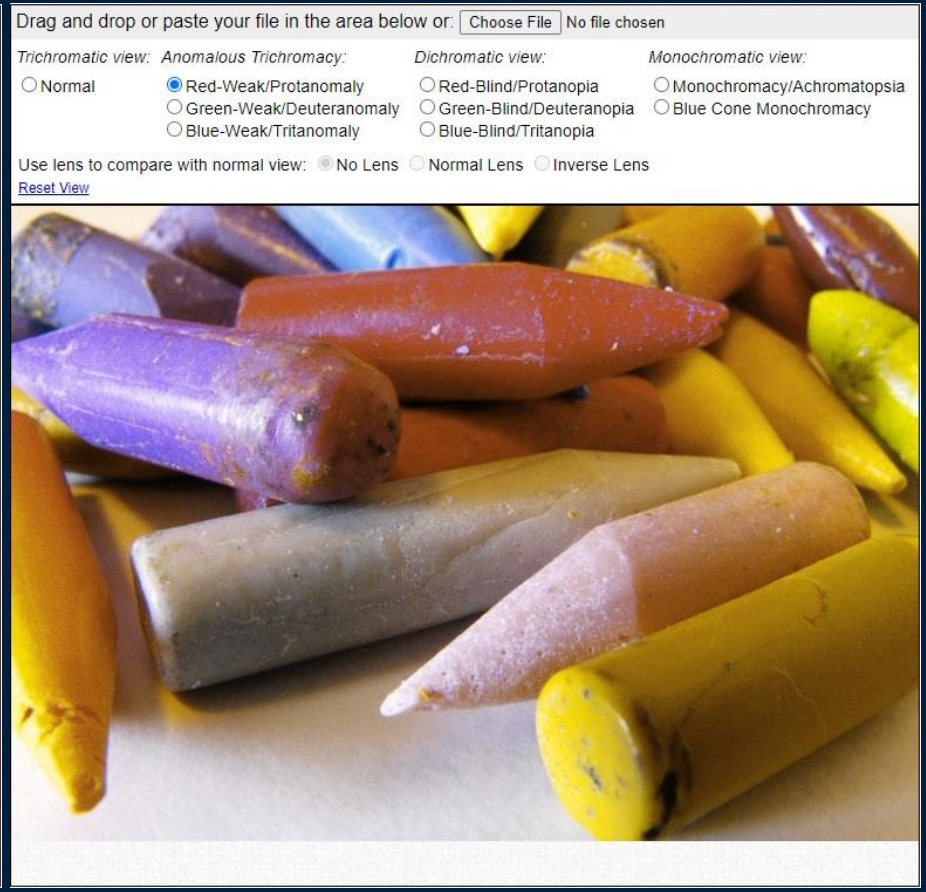
https://www.colourblindawareness.org/colour-blindness/

Colour blindness

Check your figure with a color blindness simulator

Color Blindness simulator:
https://www.color-blindness.com/coblis-color-blindness-simulator/

Colour blindness



Color as a tool to distinguish

We frequently use color as a means to distinguish discrete items or groups that do not have an intrinsic order, such as different countries on a map or different manufacturers of a certain product. In this case, we use a *qualitative* color scale. Such a scale contains a finite set of specific colors that are chosen to look clearly distinct from each other while also being equivalent to each other. The second condition requires that no one color should stand out relative to the others. And, the colors should not create the impression of an order, as would be the case with a sequence of colors that get successively lighter. Such colors would create an apparent order among the items being colored, which by definition have no order.

Many appropriate qualitative color scales are readily available. Figure 4.1 shows three representative examples. In particular, the ColorBrewer project provides a nice selection of qualitative color scales, including both fairly light and fairly dark colors (Brewer 2017).

Chapter 4 and 9 of Fundamental of data visualization