III. The Golden Rules of Data Visualization

→ Recommendations to acheive excellence in graphics

① Show Data
- The goal is to inform about data
- Don't hide it, don't overprocess it

② Induce the viewer to think about the substance not graphic, technology of graphic ....
(Don't use elements that steal the attention away from info)

③ Avoid distorting what data has to say
(Never allow design choice to misrepresent the data)

④ Present many data in a small space
(Don't waste a big space for very few data)

⑤ Encourage the eye to compare different pieces of data
(put comparision in the same graph)

→ Golden Rule

Maximize the number of ideas while minimizing resources spent

Ideas ↑↑
we want to transmit as much as info. as we can

- Time ↓
Have the attention of the viewer for a moment
- Ink ↓
Ink is related to info. When we use ink user will try to decode info.
- Space ↓

التشويه

→ Graphical Distortion

- A graphic doesn't distort if the visual representation is consistent with the numerical representation.
- Relationship between the numerical measure reported and the perceived measure

⇒ perceived area = (real area)$^{\alpha}$

$\alpha = 1$	$\alpha < 1$	$\alpha > 1$
perceived = real (no perception mistake)	perceived < really (perceived smaller area)	perceived > reality (perceived bigger area)

→ Graphical Integrity (Telling the Truth)

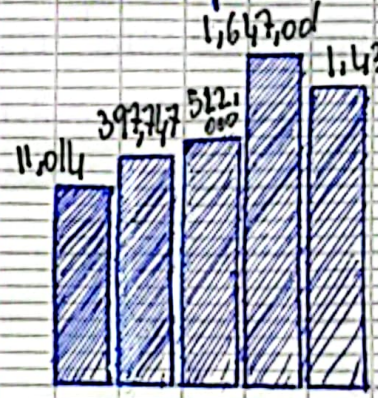
① Representation of numbers on a graph should be (The physical) directly proportional to the actual numerical values they represent.

⇒ How much the graphs lie :

$$\text{Lie Factor} = \frac{\text{Size of effect shown in the graphic}}{\text{real size of effect in the data}}$$

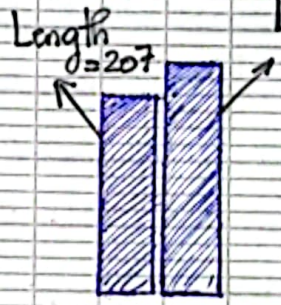
< 1	$= 1$	> 1
rep. of data underpresenting the real effect of the data	rep. of data is true to the real quantitative data	rep. of data is over-representing the real effect of the data

. Example:



Original Plot.

choose a pair of bars ⟹



⇒ Effect shown in the graphic

$$\text{Relative change} = \frac{\text{length}(2) - \text{length}(1)}{\text{length}(1)} = \frac{227 - 207}{207} = 0.0966$$

⇒ Effect in the data

Data bar 1 = 11,014

Data bar 2 = 397,747

$$\text{Relative change} = \frac{397{,}747 - 11{,}814}{11{,}014} = 35{,}113$$

$$\Rightarrow \text{Lie Factor} = \frac{\text{Graphic Effect}}{\text{Data Effect}} = \frac{0.0966}{35.113} = 0.00275$$

In this example, the lie factor is much smaller than one. This is telling us that the plot is really underestimating the effect of the data

. Notes about calculating the lie factor

→ The lie factor might not be constant along the plot

→ For a graph not to lie: the lie factor should be 1 if we compare all pairs

→ It doesn't matter if we take measurement in CM, IN, or pixels. (Use the same unit)

→ Only proportion matter

→ ~~Bar/area~~, measure magnitude that need to be encoded (area, volume, length)

↓ bubble plot

↓ bar chart or similar

② Clear, Detailed, and Thorough labeling should be used to defeat graphical distortion and ambiguity
- Label important events, with explanation
- Annotation highlight important events in data (help understand)
- But too many annotations lead to distortion
- Show how to read the graph

③ Data Variation Not Design Variation
↳ make sure to use the same scale (info-carrying)

④ Number of / variable dimension should not exceed the number of dimensions in the data
↳ Nb of dimensions in the graphical representation
=
Nb of dimensions in the data

- Each dimension should be for a reason
- Don't use extra color until it means something
- Don't use weird perspective
- Never use 3d plots unless strictly necessary

Do you think you really need it?

Is your data more than 2d?

Is it really 3d or 2d with multiple instances

Is it really important to show rlt. between 3 or more var. or just want to save space?

There is no other way to represent this data?

→ Data-Ink Ratio

. Data graphics should draw the viewer's attention to sense and substance of data., not something else

. Data-Ink vs. Non-Data-Ink

↙ non-erasable core of the graphic non-redundant ink

↘ The rest of the ink in the graph

. We have to maximize the data-ink ratio:

Data-Ink Ratio = Data-Ink / Total Ink in the graph

↙ Should be close to 1 as possible (maximize)

how?
. erase non-data-ink, within reason
. erase redundant data-ink, //

→ Data Density

. Display highest amount of data in smallest space possible

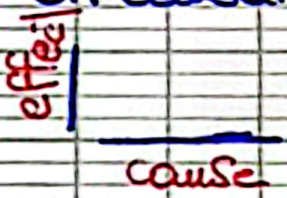
. Data Density = Nb. of entries in the data matrix / area of the plot
↓ (unit/cm²)

. maximize data density

very ! Low density ⇒ don't draw a graph

→ Proportion and Scale

- Graphics should be drawn in landscape mode (x longer than y)
  - easier to understand and read
  - allow to put emphasis on causality



- For time series: (banking to 45°)
  all segment of time series should have an angle close to 45°
  (maximize the discriminability of the orientations of the line segments in the chart)

golden rectangle
1:1.6