Best Practices in Data Visualization
Report

Anouk Paulino

Contents

BEST PRACTICES

Data visualization is an important aspect of scientific research and analysis. Its effectiveness lies not only in its visual appeal but also in its ability to communicate complex information in a clear and concise manner.

As societies developed and data became more complex, we began to use increasingly sophisticated visualization techniques. In the 20th century, advancements in technology and computing power led to new forms of data visualization, including interactive and animated graphics. Today, data visualization has become an essential tool for conveying complex information in a clear and concise manner, with applications in fields such as science, marketing, and journalism.

When designing a chart to convey information, we must therefore pay attention to different mechanisms in order for the result to be clear to our audience. Inspired by Scott Berinato's book [1].

Human perception

Innate

People's attention is often drawn to specific elements within a visualization, such as labels, titles, legends, and data points [2].

The visual hierarchy of a visualization can greatly affect how people interpret and understand the data. People tend to look at larger and more prominent elements first, such as titles and legends (useful to get the context of the graph), before moving on to smaller details like data points and labels [3]. Clean and simple visualisations are pleasing to the eye and make the reader feel confident about the conveyed message and their ability to understand it. In constrast, cluttered and overly complex visualizations can be overwhelming and difficult to interpret. It can confuse the viewer, obscure important trends and patterns.

There are different levels of information processing that happen when we view data visualizations. The "blurry level" or "fast thinking" is an almost subconscious level that allows us to quickly pick out patterns. This level is important for identifying general trends and patterns that are not necessarily specific values or points in the data. It occurs in dozens of milliseconds after exposure and can be described as reactive, automatic, instinctive and emotional. In contrast, the "high road" or "slow thinking" level of processing is a more deliberate parsing of information. This level allows us to identify specific data points and making precise comparisons and is more analytical, effortful and logical.

A great example of how this can impact visualizations are the Gestalt principles [4, 5]. There are 4 of them, and are all completely subconscious:

> Closure: if some shapes are incomplete, our brain automatically fills them in

> Common Region: if some elements are placed in a region separated from other elements, they are considered as part of a same group

> Proximity: if some elements are close to each other, they are considered as a part of a same group, as opposed to points further away

> Figure / Ground: unless the image is ambiguous, our brain recognizes the foreground first

Hereby lies **saliency**. In the context of data visualization, saliency refers to the visual elements that draw the viewer's attention and stand out from the rest of the information. It can be achieved through various visual design techniques such as color, size, contrast, and shape. By emphasizing certain parts of the data, it can help guide the viewer's attention and focus on the most important information.

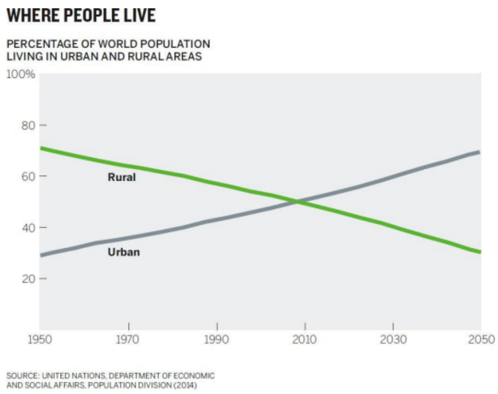

Figure 1: Example of obvious saliency
When first seeing this graph, our brain sees the meeting point of the two lines. This is what saliency is about.

Saliency helps to make the information more understandable and memorable. By highlighting the most important information, the reader is better able to grasp the main points of the data and remember them. In contrast, if all the visual elements are of equal weight and importance, the viewer may struggle to understand the main points.

It is important to ensure that the salient elements don't overwhelm the viewer though, or distract from the main message of the visualization. Effective data visualization strikes a balance between saliency and coherence.

Consequently, we naturally seek meaning and try to make connections in everything we see. This can lead to us over-analysing and jumping to conclusions based on previous experiences. This is called the confirmation bias.

Several other biases may be encountered and should be kept in mind when designing a visualization [6]. Here are some of the most relevant:

> Availability bias: the tendency to rely on information that is easily available in memory, rather than seeking out additional information or considering less salient but equally important factors.

- Anchoring bias: the tendency to rely too heavily on the first piece of information encountered when making a decision, and to be insufficiently influenced by subsequent information. This is also heavily linked to saliency.
- Framing bias: the tendency to be influenced by the way information is presented, such as the order of presentation or the wording of a question. For instance, removing numerical information from a graph can also significantly influence decision-making and cause biases in the interpretation of the data.

Acquired

As cultural and educated beings, we also rely on conventions and metaphors. Here are some examples of what can universally be assumed:

- connected data point imply that there is a relationship between the different values ;
- like colors mean like items ;
- color saturation hints value, lighter values meaning less dense or emptier ;
- categories are arranged and plotted from one extreme to another
- time is plotted from left to right ;
- maps are displayed with the north "up".

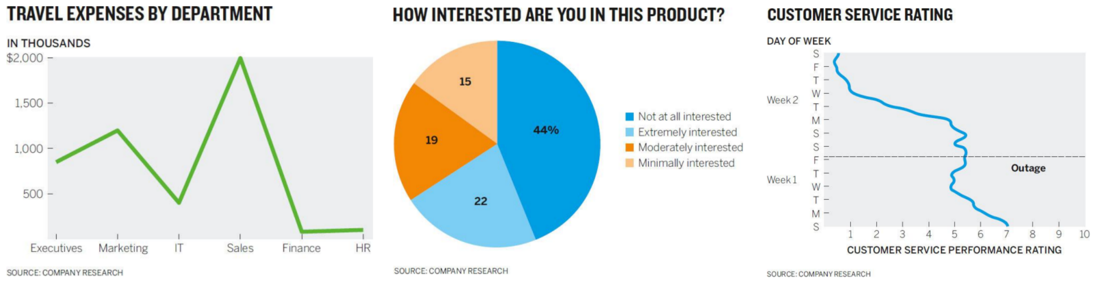

Figure 2: Examples of confusing graphs
Left: there is no relationship whatsoever between the connected points
Middle: different hues of the same color do not describe like patterns
Right: time is displayed from top to bottom

Some of those examples are tightly linked to the Gestalt Principles mentioned in the previous section.

In addition some other acquired beliefs may be influenced by our environment. Culture, education, personal experience can greatly affect our interpretation in several cases.

A well-known cultural bias is about the meaning of colors [7]. For instance, in western countries white represents innocence and fertility, while in eastern countries it is all about death and supernatural creatures.

Colors also have a great impact on behavior and reaction to a graph as the color palette can impact the overall feeling of the visualization [7].

Examples of poor cognition consideration

Taking the previous points into account, it is impossible to obtain an objective visualization, because our brains are wired to interpret what they see based on previously acquired knowledge. It means that a same chart can have as many interpretations as there are readers. Also the designer also being a human being, their work is necessarily subjective. The goal here is the become aware of those biases and try to reduce them as much as possible.

As a result there is a blurred line between visual persuasion and visual dishonesty (c.f. example of the planned parenthood graph [8]). Other good examples are websites that use what are called dark patterns [9]. These website purposely hide some type of information for the customer to give up before finding it. A well-known example is Amazon, which requires many confusing steps before allowing the user to delete their account.

Some practices are known to be misleading are must therefore be avoided:

- Truncated y-axis
  The viewer will assume the axes are complete and will not necessarily pay attention to this change of representation.
- Double y-axis
  This practice forces the viewer to compare two things that can not necessarily be compared.
- Geographical representation

Figure 3: Left: truncated y-axis, right: full y-axis
Notice how the left chart emphasizes patterns that are not necessarily visible on the right-hand side. If the viewer does not notice the truncated axis, they might misinterpret the shape of the line as a more dramatic event than it actually is.

Figure 4: Example of double y-axis misleading, displaying the price of both gold and silver on the same graph, with respect to time
Both axes do not display comparable units. When reading the graph to quickly one could assume that there is a correlation between the two, even though no link can be established when thinking about it.

When the data is not relevant in terms of geography and is it represented as a map, it can lead to a real deception of the viewer.

We notice that there are many practices to avoid and it can be overwhelming to take all those points into account. Here is a list of honesty challenging (and interesting) questions to ask oneself to make sure the produced visualization is not unintentionally biased:

- Does my chart make it easier to see the idea, or is it actively changing the idea?
- If it is changing the idea, does the new idea contradict or fight with the one in the less persuasive chart?
- Does eliminating information hide something that would rightfully challenge the idea I am showing?
- Would I feel duped if someone else presented me with a chart like this?

Having thought those questions through, we can assume that we are off to creating an honest and useful data visualization.

Figure 5: Map of the votes for the scottish referendum for independence.
In reality, while less than 5% of land mass voted "yes", actually 38% of voters voted "yes".

How to actually do it

We have reviewed how we perceive things as human beings and depending on our culture and education. The thinking behind the construct of the chart and knowing the audience is therefore a must to make sure the visualization will achieve its goals.

The importance of brainstorming

One can separate charts into four categories.

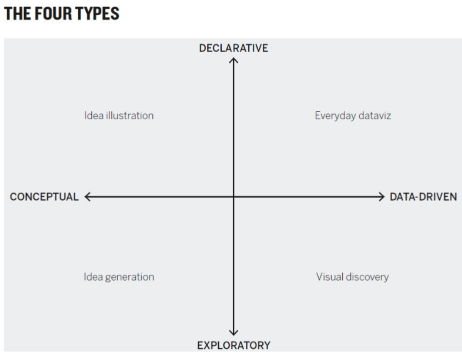

Figure 6: The four types of visualizations, according to Scott Berinato

Each of these categories represents a purpose for the chart:

- Everyday dataviz is to express a given idea to audience. The designer knows how the data is going to talk when making the graph.
- Visual discovery is a way of confirming hypotheses or feeling about the data. One does not necessarily know what they are looking for in the first place when doing visual discovery.

- *Visual confirmation* is declarative.
- *Visual exploration* uses raw data to see patterns and trends emerge.

⋗ Idea illustration is conceptual, uses metaphors, trees, etc. It is the dataless version of everyday dataviz.

⋗ Idea generation is conceptual and most commonly used in brainstorming sessions in the form of sketches. It does not necessarily end up as a proper graph. It is used for reflection and is the dataless version of visual discovery.

Since our work is based on data, we will focus on the first two.

Now, how can we actually get started with making a chart? We have a data set, an idea (or not) of what story we want to tell with this data set, and most importantly we have our audience. As described above, we have to take into account the future graph's public in order to lessen the possibles biases. To make a chart as useful and functional as possible, we have to go through the four following steps (with an advice on how much time to spend on each task represented as a percentage of the overall time spent brainstorming).

PREPARATION (no more than 5%)

As a starting point, put the data aside. Focus on the goal of the chart. A basic data visualization requires:

⋗ a purpose (a story to tell) ;

⋗ knowing its audience ;

⋗ knowing the setting in which it will be used in ;

⋗ knowing which of the four types it is ;

⋗ finding the balance between context and design .

Having those elements in mind helps determine which way to go.

TALK AND LISTEN (15%)

⋗ What am I working on ?

⋗ What am I trying to say or show (declarative)/prove (confirmatory)/learn (exploratory) ?

⋗ Why ? (most useful for declarative data viz)

Talk this over with a colleague, either external to the project if you want to set ground for further thoughts, or internal to refine the ideas. This step is primordial. If you find yourself unable to answer questions from a colleague, then the concept has not been properly thought through and requires some refining.

SKETCH (55%)

Match keywords to approaches [2] :

- Line charts compare values overtime
- Bar charts compare quantitative data from different categories
- Scatter plots display relationships between different aspects of the data
- Pie charts cannot display time evolution, they rather show parts of a whole

However do not limit yourself to the above. Trying out new ways of representing the data can lead to better understand how to convey a message.

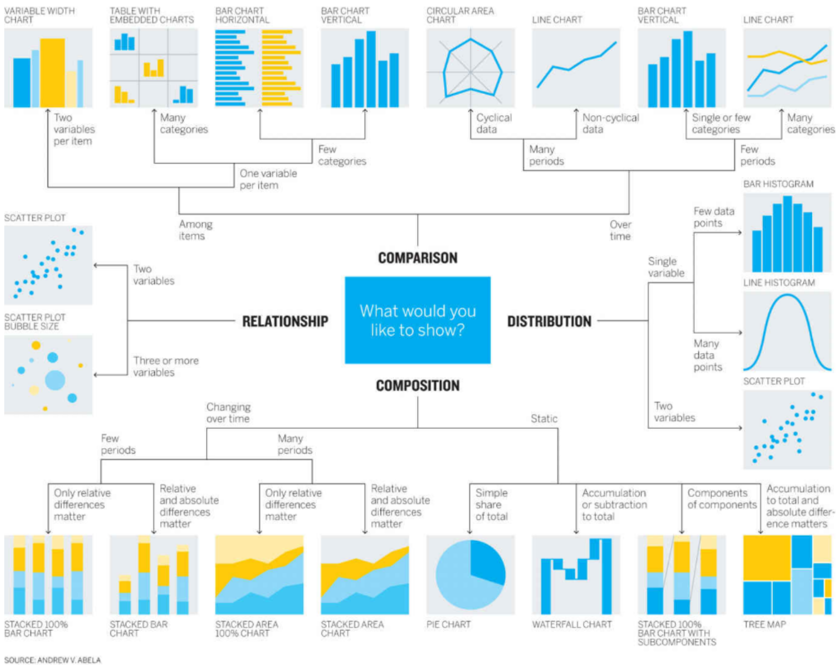

Figure 7: Andrew Abela's chart of which tool to use

Most importantly, think of your audience and make the chart inclusive. For instance, colorblindness is rather common. Several tools exist to work around it (e.g. on the Adobe Suite tools, or using alternative color palettes). Also, since color is so important, take some time to thoroughly think through the overall look of the chart:

- High contrast makes a chart easy to read
- Pleasing colors make it less painful to look at
- Complementary colors help with making links between different elements on the image

Make wise use of texts and icons to label their graph. Such additions may help the reader understand the chart, but may also confuse them if poorly arranged. We mention this in greater detail further in the document.

PROTOTYPE (25%)

Progressively incorporate the real data into the graph. At this step you can start using some online tools to produce your graphs.

An important thing to note is that the above steps are overlapping and the overall process is iterative. Trying to get a chart done in one shot will most of the time not give out great results.

Going back and forth between the steps is important to get feedback and think outside the box.

The execution

Our brain consumes lots of energy for visual recognition [2]. To increase clarity and therefore be more straight-forward, here are the main concerns to have in mind:

- Hone the main idea
  Do so by making wise use of colors, contrast, highlights, etc. Do not overuse those elements though. For clarity a title and subtitle should differ in no more than two attributes. [10]

- Make the design easy to identify
  Simple designs are more comprehensible than overcrowded ones [11]. Deleting unnecessary elements can be rather hard if you have spend a lot of time working on your chart. Use the following guide to make it simpler.

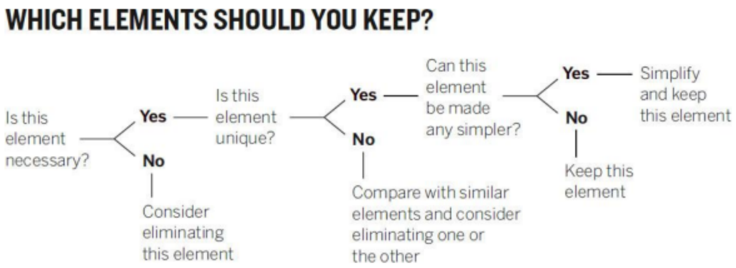

Figure 8: Guide to keeping only necessary elements

  This step is all the more important if you think everything in your plan is mandatory. You have spent quite some time reflecting on your future graph, it is normal for you to be reluctant.

- Minimize eye movements
  Some patterns are easier on the eyes. In website design, the F-pattern for instance feels rather natural [2]. In charts however there is no such contract. The goal is to minimize eye movements to make the chart pleasing to the eye.

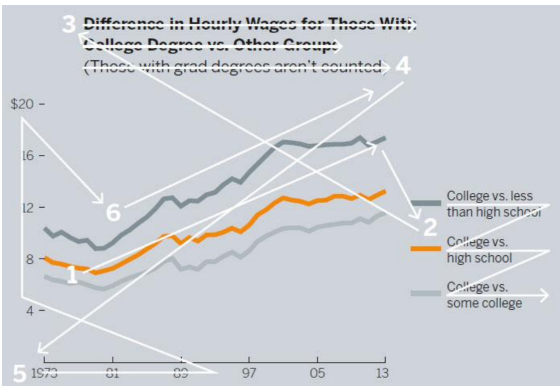

Figure 9: Example of typical eye movements when reading a chart (1 being the first step)

The more complex the pattern, the more lost will be the viewer.

⋗ Indicate what is going on
Information you do not mean to communicate can still impact interpretation. If your chart is ambiguous, a message that you did not intend on telling may be conveyed.

Another important part about the design is colors. Colors allow us to distinguish elements. As mentioned earlier, there is a huge bias in color interpretation based on culture. It is therefore necessary to pay great attention to the color palette. Also reduce the amount of different colors used when applicable, it makes the chart less overwhelming. Prefer using different hues (gray scaling) to inform about hierarchy.

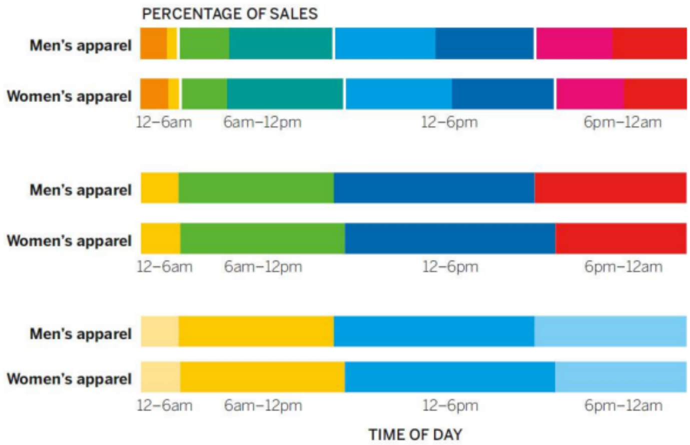

Figure 10: Comparison of several color palettes displaying sales with respect to gender and time of day.
Notice how the first one's palette seems sloppy compared to the other ones. It contains too many different tones. The last one successfully distinguishes the morning and afternoon sales for each gender with different hues of the same color for early and late day periods.

There exist several online tools to come up with a meaningful color palettes (`paletton.com`, `https://coolors.co/` are the ones I personally use).

As a(n) (almost) final step, it is always interesting to test one's design on a sample public. You can mandate the same colleague as before, having a completely new point of view may be relevant. This step allows us to spot the eventual weaknesses of the chart.

PRACTICE

Emissions

Subject description (food emissions, focus on several matters, etc.) Describe the five criterion we have data for :

- Land use
- Water use (+ scarcity-weighted version)
- Carbon emissions
- GHG emissions
- Eutrophication

What the data can tell us :

- Agricultural land use (CH) + land use per product (CH, world)
- Average intake per product for a swiss citizen
- Carbon emissions per product (and per production stage) (world)
- Water use per product + scarcity-weighted as well (world)
- GHG emissions related to agriculture (CH)
- Eutrophication per product (world)

Induced questions :

- Which products produce the most emissions (one sub-questions for each criterion) ?
- How to reduce food-related emissions of the average (swiss or not ?) consumer ?
- Which stages and/or products should we focus on to get the greatest impact on emission reduction with minimal modifications to the actual system ?

Geodata

An insteresting point that we wanted to look into was geodata. It is much graphical and allows a wide variety of visual interpretation.

We have chosen to use a basic set of population density in Switzerland (with the highest precision possible), among the many provided by Kontur [12]. We decided to turn it into a <u>beautiful</u> map. In this data set, the data is organised in hexagons. It appears to be the most optimal way of representing a surface. This way of organising geodata has been imagined by Uber to analyze traffic [13].

KeplerGl

Following this idea, we investigated the KeplerGl tool [14] (also developped by Uber to make use of those hexmaps) and wanted to give it a try.

The online interface allows the user to upload a data set and tweak some parameters to alter the look of it.

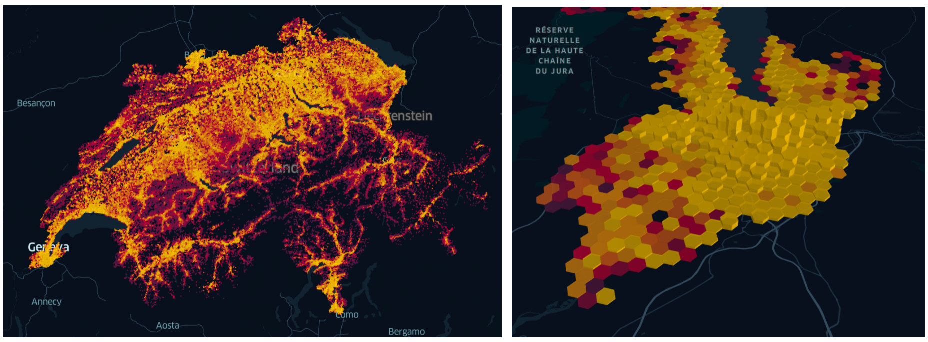

Figure 11: Visuals of Kontur data in Kepler Gl tool
Left: default settings when inserting the data
Right: focus on Geneva, 3D visualization with height proportional to population density

However it seemed rather limited. Since there exists a python package for KeplerGl, I tried to integrate it to a notebook. Once again though I was disappointed. I did not manage to make the tool it work with the Kontur data set. I apparently am not the only one having trouble getting the package in hand, due to the fact that it is rather new and does not have extended support.

I had to give up on this aspect because I lacked time.

Rayshader and Rayrender (RStudio)

Many map visualizations appear to be made using R packages (Rayshader and Rayrender mainly), so we decided to give it a try.

To add a bit of context, Rayshader is a tool that allows us to control the light sources within the 3D render. Rayrender is used by Rayshader to define the way an object interacts with light. Combined together these packages allow us to create complex 3D visualizations. I decided for RStudio as an IDE, since it is widely used and provides a useful visualization interface.

I followed a tutorial that rendered a map of Florida in RStudio which also used a Kontur data set [15]. The same author also wrote an article about it [16].

I had to adapt the content of the video to adapt it the Swiss data and map. I therefore looked for a hexmap displaying the border of Switzerland [17]. This map serves as a ground map for the data to be displayed.

I played around with the camera angle and lights sources a bit. Changing the color palette also impacts the visual. I also tried to change the material of the spikes (i.e. change the way it interacts with light using Rayrender) but it was not really convincing.

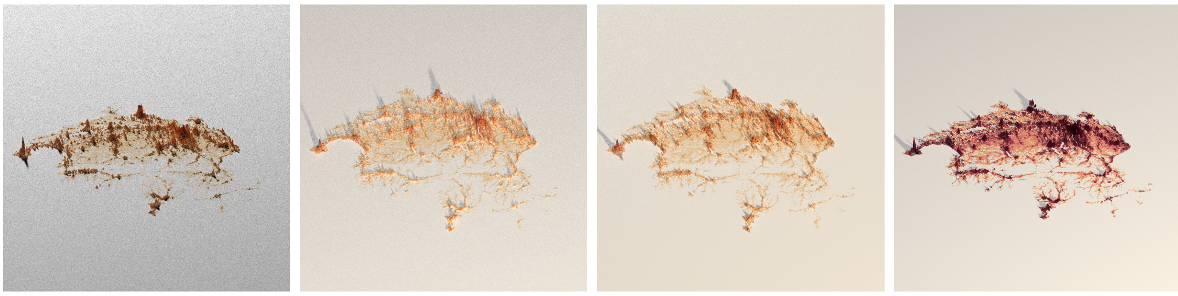

Figure 12: Results following the Rayshader tutorial, with several trials with repsect to lights and camera angles
From left to right: default light sources ; two light sources ; also two light sources but with a different angle ; glass-like render

R provides a package called Magick that allows us to write on images. It is mostly used to annotate maps such as the one I have generated above. I therefore completed with some city names to make the chart more readable.

Since all of this is done programmatically, I spent a lot of time writing the coordinates for each segment and text. I believe it would have been all the more efficient had I used a graphic design software to do this part (e.g. Inkscape [18]).

I must say I am quite satisfied with the result though. However, this map is purely qualitative, as we had decided in the first place that what mattered was the look of it. It is virtually impossible to deduce any of the data from the visuals (for instance there is no scale whatsoever). It is more of a piece of art than a graph. Such maps could be combined with more practical ones generated with matplotlib, say, to get a better grasp of the data.

Sugar

I was interested in the matter of sugar, sugar processing, sugar production, etc. To my mind, sugar production looked like a widely used process that involved a lot of waste. I wanted to see if my gut feeling was right.

While looking for interesting data about this process, I stumbled upon a paper that described how sugar production wastes could be reused in lots of ways [19]. I completed the knowledge from the paper with a bit of more practical information from a short documentary about sugar production [20].

To apply the theory of a good data visualization, I chose to turn this data into a poster which clearly conveyed the idea I wanted to give out: sugar processing has a very poor yield.

To make this poster, I first thoroughly looked into the paper to extract the more information I could get

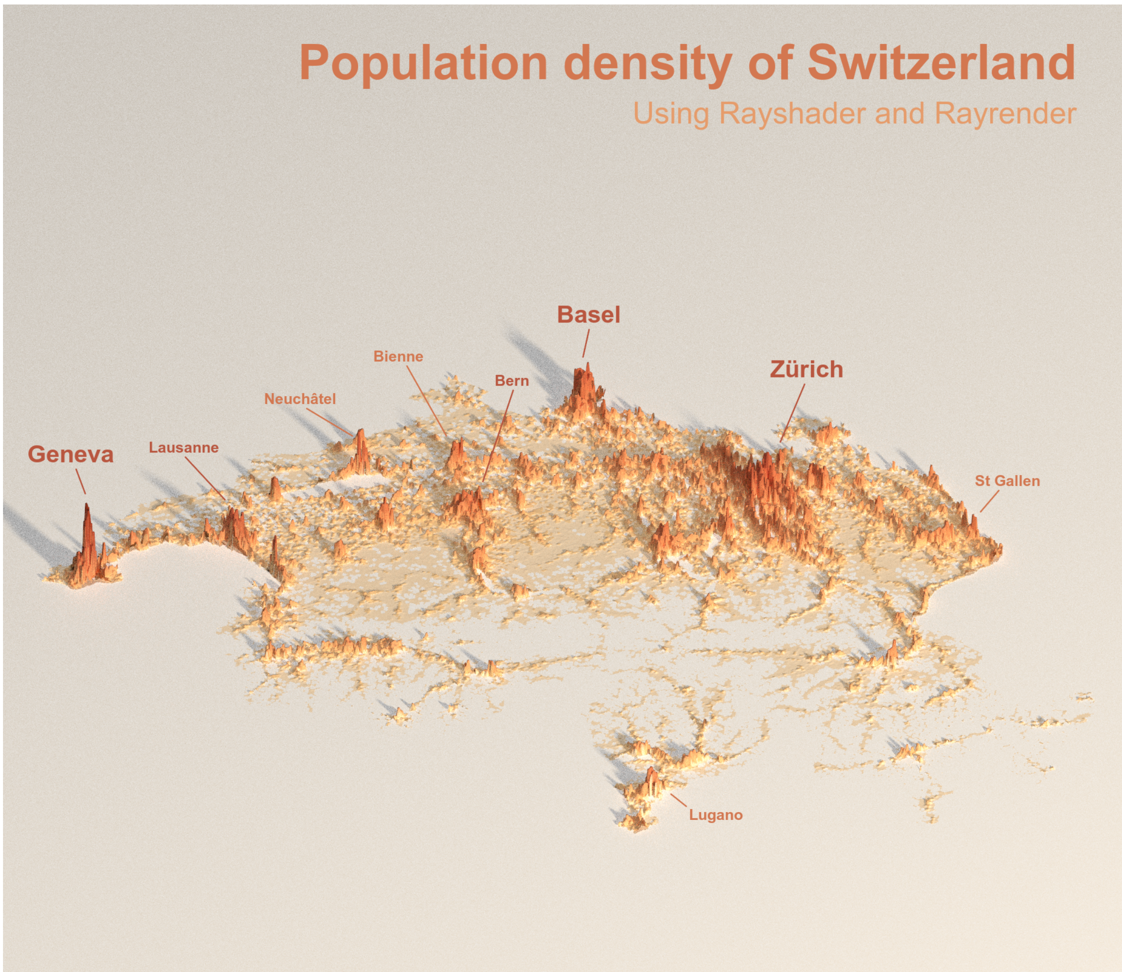

Figure 13: Map of population density in Switzerland, annotated.

I then followed the PREPARATION, TALK AND LISTEN, SKETCH, and PROTOTYPE iterative process. As I expected, a lot of time went into trying to figure out how to organize the poster, which information to add, and how to make it look appealing.

I had decided to go with a Sankey plot, which seemed to be a good way of representing the flow of the different products.

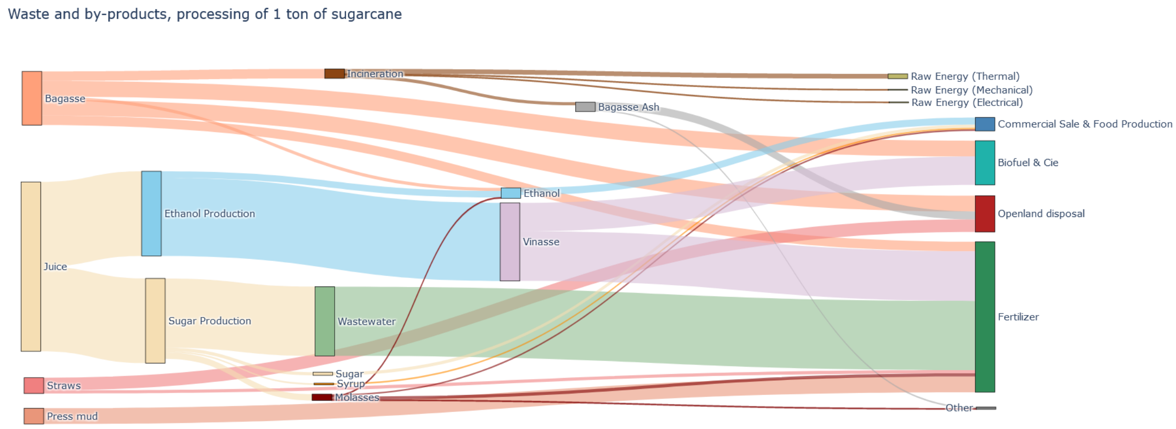

Figure 14: Sankey plot with the data found in the paper [19]

I was not convinced. The graph was ugly and most importantly it did not make real sense. As a result, even though the proposals for waste recycling were interesting, I decided to focus on the actual generated wastes nowadays. After many conversations with different people, I ended up with the final version of the poster.

I opted for a rather pastel color palette and decided to focus on the look of the poster. If the poster looks simple and harmonious, people will more likely want to read its content.

Figure 15: Sugar Production and Byproducts, realized with Inkscape

FURTHER READING

Here are also some additional articles, papers, etc. that I did not find relevant enough to mention (they either did not fit scope of this project or were simply redundant with other sources). However I found them very enlightening on a personal level.

[21] A forum where people debate on visualizations practices.

[22] Yet another article with data viz tips.

[23] An article about color palettes in geovisualizations.

[24] A psychological approach to UI.

[25] How to avoid lies in data vizualisation.

[26] Recommendations of the best data visualization books in 2022.

[27] An article about bullet graphs.

[28] An AI to analyse how efficient a design is with respect to UX.

[29] A list of recommended tools which can be used to make visualizations.

[30] A wide panel of visualizations.

[31] Some templates for infographics.

Figure 16: Color meaning by culture

CONCLUSION

References

[1] Scott Berinato. *Good charts : the HBR guide to making smarter, more persuasive data visualizations*. 2016. ISBN: 978-1-63369-070-7.

[2] Various Toptal authors. *The Psychology of Design, A Collection of our Top Articles on Behavioral Design*.

[3] Yolanda Gomez Jose Vila. *Extracting business information from graphs: An eye tracking experiment*. 2015. URL: https://www.sciencedirect.com/science/article/pii/S0148296315004713.

[4] James Enns Christopher Healey. *Attention and Visual Memory in Visualization and Computer Graphics*. 2012. URL: https://ieeexplore.ieee.org/abstract/document/5963660.

[5] Interaction Design Foundation. *The Gestalt Principles*. URL: https://www.interaction-design.org/literature/topics/gestalt-principles.

[6] Michel Fabre Jean-Pierre Caverni Jean-Francois Gonzalez. *Cognitive Bias*. Ed. by North Holland. 2004.

[7] Cameron Chapman. *Cause and Effect: Exploring Color Psychology*. URL: https://www.toptal.com/designers/ux/color-psychology.

[8] Timothy B. Lee. *Whatever you think of Planned Parenthood, this is a terrible and dishonest chart*. September 2015. URL: https://www.vox.com/2015/9/29/9417845/planned-parenthood-terrible-chart.

[9] Nerdwriter1. *How Dark Patterns Trick You Online*. March 2018. URL: https://www.youtube.com/watch?v=kxkrdLI6e6M.

[10] Tony Hammond. *9 best practices and tips to follow for effective data visualization*. January 2023. URL: https://www.thoughtspot.com/data-trends/data-visualization/best-practices-and-tips-for-effective-data-visualization.

[11] Tommy Walker. *Why Simple Website Design is The Best: The Scientific Reasons*. May 2022. URL: https://cxl.com/blog/why-simple-websites-are-scientifically-better/.

[12] *Geospatial Data and real-time risk management for business, government and humanitarian organizations*. URL: https://www.kontur.io/portfolio/population-dataset/.

[13] Uber Engineering. *[Uber Open Summit 2018] Tiling the Earth with Hexagons*. November 2018. URL: https://www.youtube.com/watch?v=_-265mfMzl4.

[14] Uber Engineering. *Kepler Gl*. URL: https://kepler.gl/.

[15] Spencer Schien. *Rayshader Tutorial—Florida Population Density*. December 2022. URL: https://www.youtube.com/watch?v=zgFXVhmKNbU.

[16] Spencer Schien. *Creating High-Quality 3D Visuals with Rayshader*. January 2023. URL: https://spencerschien.info/post/data_viz_how_to/high_quality_rayshader_visuals/.

[17] Cartography Vectors. *Switzerland Detailed Boundary*. URL: https://cartographyvectors.com/map/1043-switzerland-detailed-boundary.

[18] *Inkscape*. URL: https://inkscape.org/.

[19] Valentin Vlădut Nicoleta Ungureanu and Sorin-Stefan Biris. *Sustainable Valorization of Waste and By-Products from Sugarcane Processing*. 2022.

[20] Discovery UK. *How Sugar Is Made — How It's Made*. July 2022. URL: https://www.youtube.com/watch?v=wvuiavj0pIo.

[21] URL: https://chandoo.org/forum/forums/discuss-data-visualizations-and-charting/.

[22] *Data Visualization Tips For More Effective And Engaging Design*. URL: https://www.tableau.com/learn/articles/data-visualization-tips.

[23] Peter Kovesi. *Perceptually Uniform Colour Maps*. URL: https://colorcet.com/.

[24] Jakob Nielsen. *Mental Models*. October 2010. URL: https://www.nngroup.com/articles/mental-models/.

[25] Jakob Nielsen. *Mental Models*. October 2010. URL: https://pubs.aip.org/aip/cip/article-abstract/10/3/268/136498/How-Not-to-Lie-with-Visualization?redirectedFrom=fulltext.

[26] Evelina Judeikyte. *2022 in books*. December 2022. URL: https://www.theplot.media/p/2022-books.

[27] *Bullet Graphs*. URL: https://www.tableau.com/data-insights/reference-library/visual-analytics/charts/bullet-graphs.

[28] *Data driven design*. URL: https://www.eyequant.com/.

[29] Simplylearn. *23 Best Data Visualization Tools for 2023*. URL: https://www.simplilearn.com/data-visualization-tools-article.

[30] Anna Glivinska. *The 30 Best Data Visualizations of 2023*. May 2023. URL: https://visme.co/blog/best-data-visualizations/.

[31] *Infographic Templates*. URL: https://www.visme.co/templates/infographics/.