DATA ANALYTICS FOR FINANCE

DATA VISUALIZATION PORTFOLIO

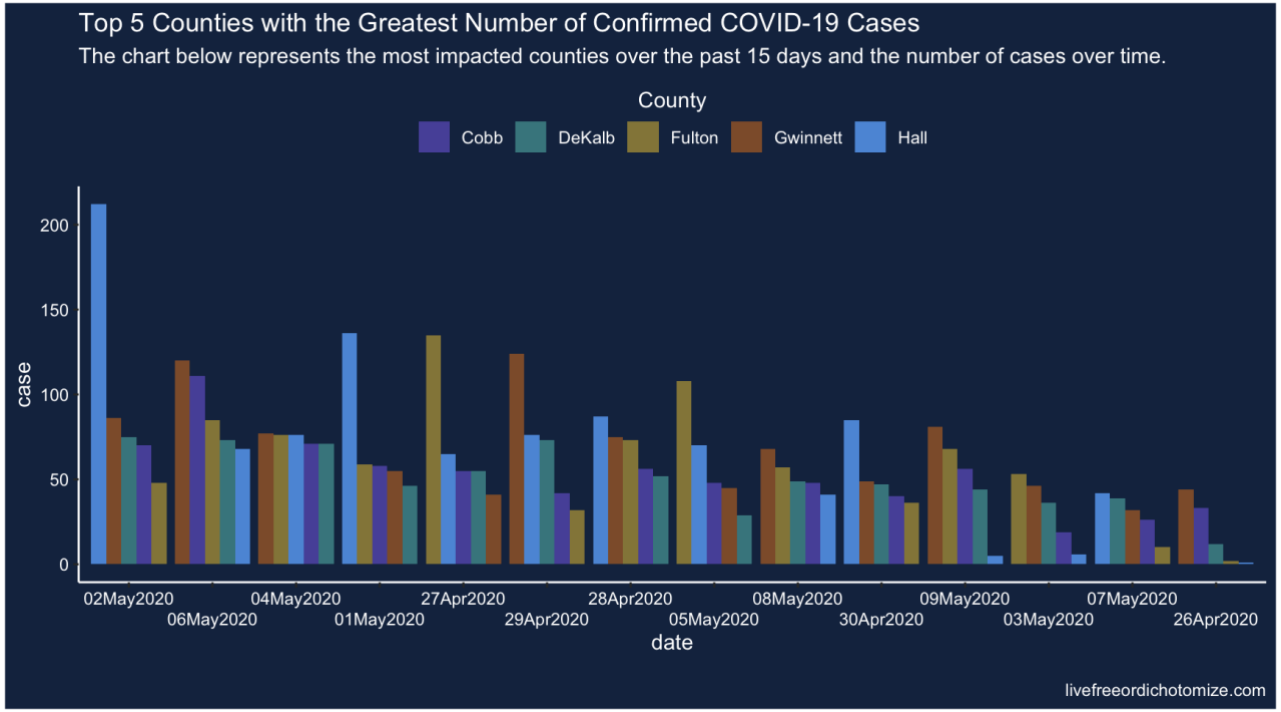
IOANNIS KOTSIAS

ioannis.kotsias@usi.ch

9-01-2024

1) A bad and/or manipulative visualization

This graph is presented by the official Georgia's Department of Health X account, displaying COVID-19 data. It may seem promising at first, but upon closer examination of the dates on the X-axis, it becomes evident that they have arranged the dates out of order to create a misleading appearance of decline.



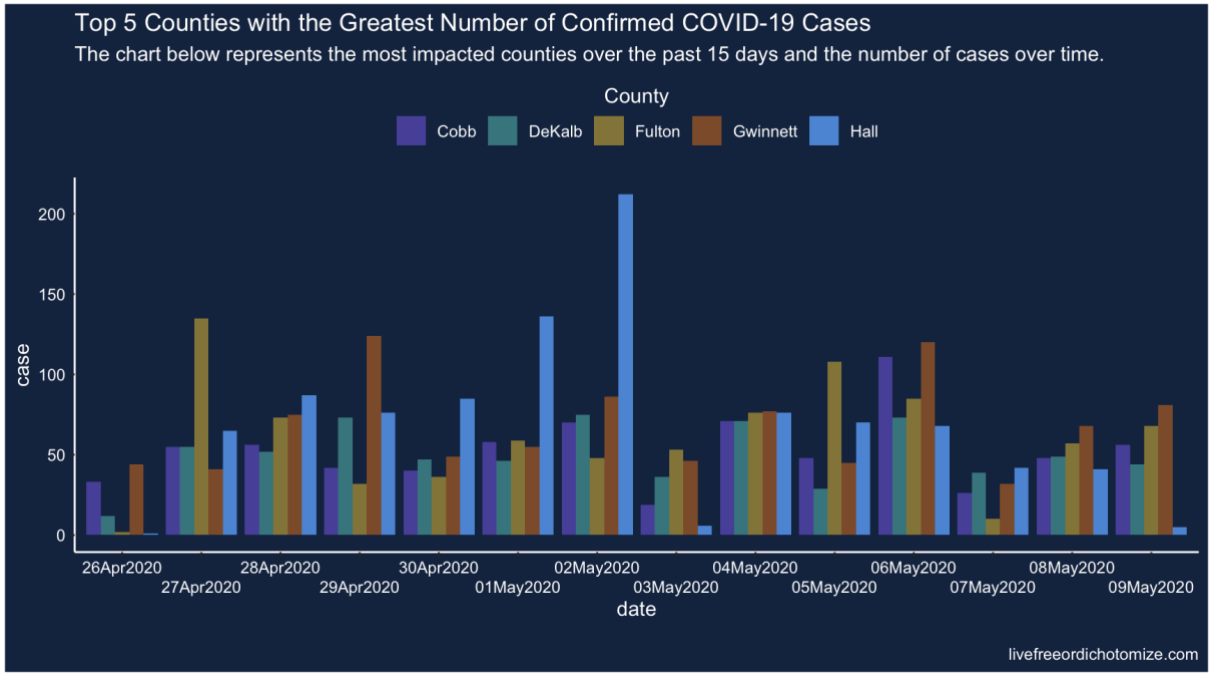
Source livefreeordichotomize

Such a fundamental error can significantly distort the perceived trends in the data. Additionally, the graph fails to specify whether the bars represent cumulative or new daily cases, a distinction that is vital for understanding the trajectory of the pandemic.

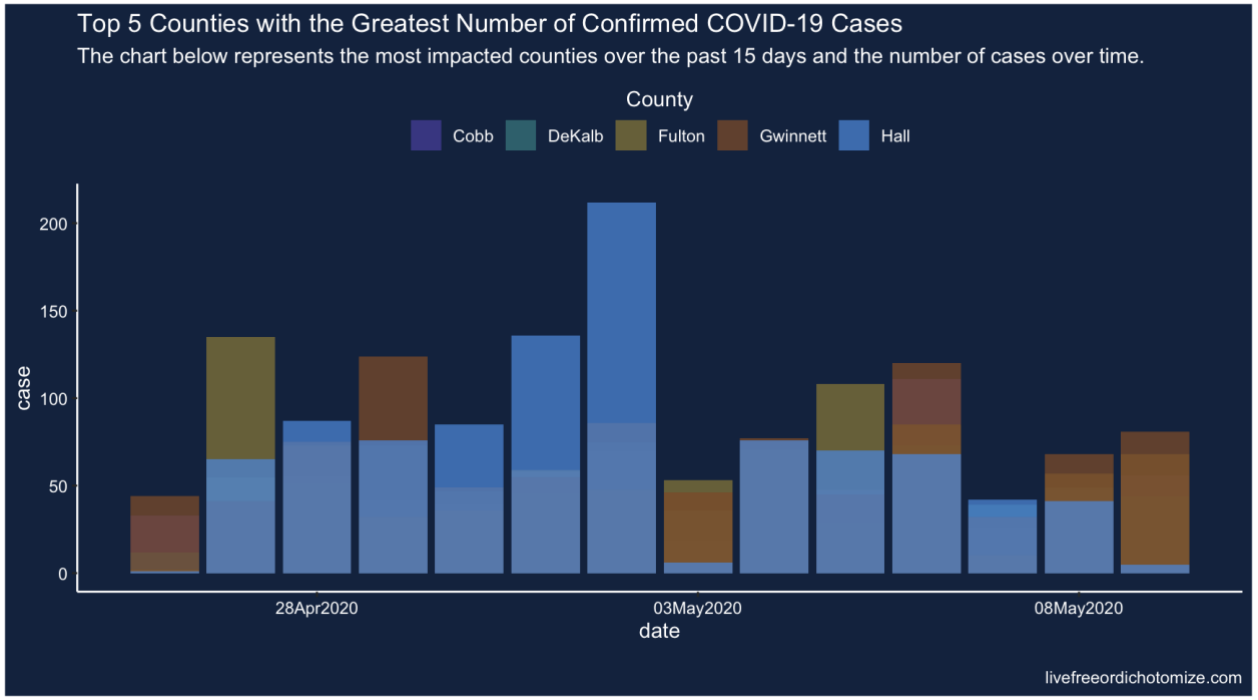
Furthermore, the labeling on the y-axis is not only incorrect grammatically ("case" should be "cases") but also the use of color is inconsistent and unclear, with some counties changing colors midway through the timeline, potentially leading to misinterpretation of which bars correspond to which counties.

2) Improved version of the "bad/manipulative" visualization

Rearranging the dates and placing them accordingly, we come up with this graph.

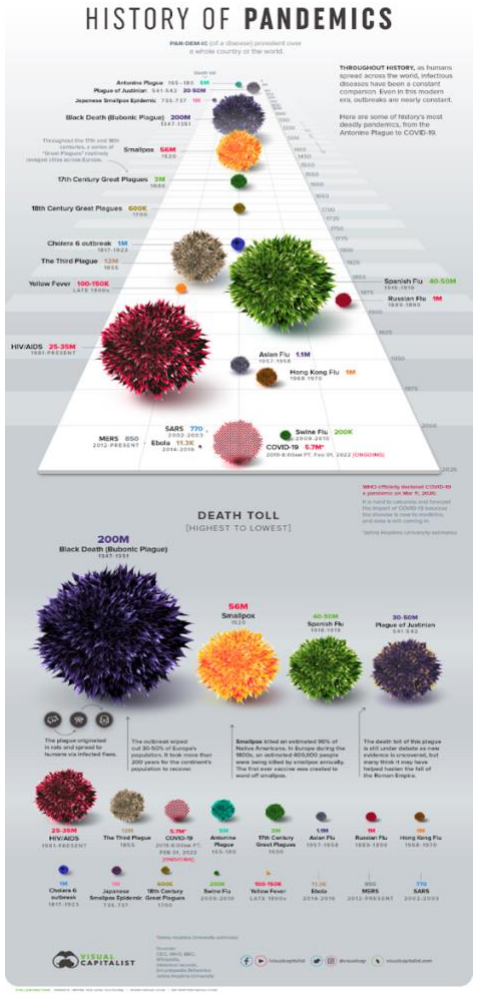


But it's still busy. By using Overlaid Histograms it can be further simplified.



Source livefreeordichotomize

3) A good visualization

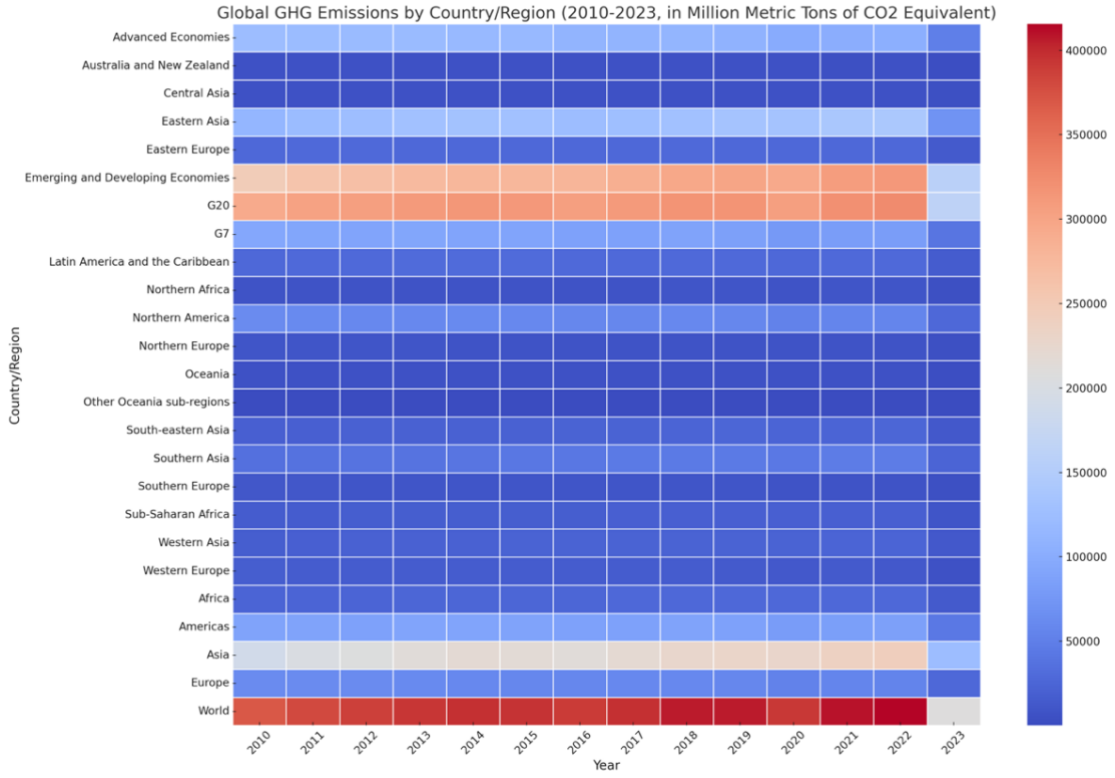


Source visualcapitalist

The "History of Pandemics" infographic by Visual Capitalist is an exemplary piece of data visualization, adeptly transforming complex epidemiological data into a format that is both accessible and engaging for a broad audience. The graphic employs a spectrum of vividly colored 3D spheres, each representing a different pandemic, cleverly scaled to convey the magnitude of the death toll associated with each event. This scaling allows for an immediate visual comparison of the impacts of various pandemics across centuries. The chronological arrangement against a timeline not only enhances the understanding of the historical context but also underscores the cyclic nature of global health crises. The design's aesthetic appeal is undeniable, with a careful selection of colors that draw the viewer’s eye while differentiating between the various diseases. Accompanying each sphere is succinct yet informative text, offering essential details without overwhelming the reader. This infographic is particularly timely, as it includes the ongoing COVID-19 pandemic, thereby linking historical pandemics with current challenges. Inclusion in a college data analysis portfolio showcases the infographic’s strength as a pedagogical tool, demonstrating how data can be effectively communicated to inform and educate on matters of significant historical and contemporary relevance.

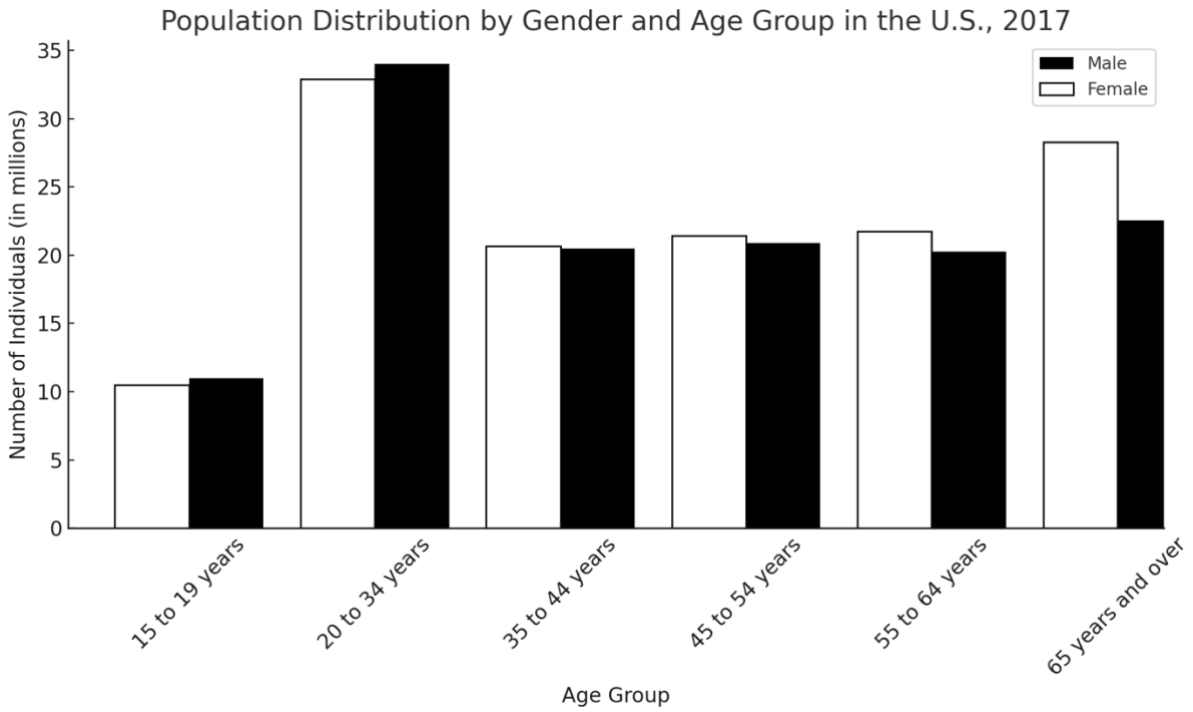
4) A visualization about climate change

The dataset was sourced from the International Monetary Fund's Climate Data portal, specifically from their "Quarterly Greenhouse Gas (GHG) Air Emissions Accounts" dataset. For more details, visit *IMF Climate Data - GHG Emissions*.



This heat map visualization represents the greenhouse gas (GHG) emissions from various countries and regions across the globe, spanning from 2010 to 2023. The data is measured in million metric tons of CO2 equivalent, providing a comprehensive view of emissions over the specified period. It effectively captures the trends in GHG emissions, showcasing whether a country's emissions have increased, decreased, or remained stable over the years. This graph, along with all following graphs were created with Python and Matplotlib.

5) A black-and-white visualization



The visualization above represents a snapshot of the United States population in 2017, segmented by gender and various age groups. This bar graph is derived from the "Throwback Data Thursday Week 30 - United States Marriage Status 2005 to 2017" dataset, originally sourced from the *U.S. Census Bureau - U.S. Marriage Status 2005 to 2017*.

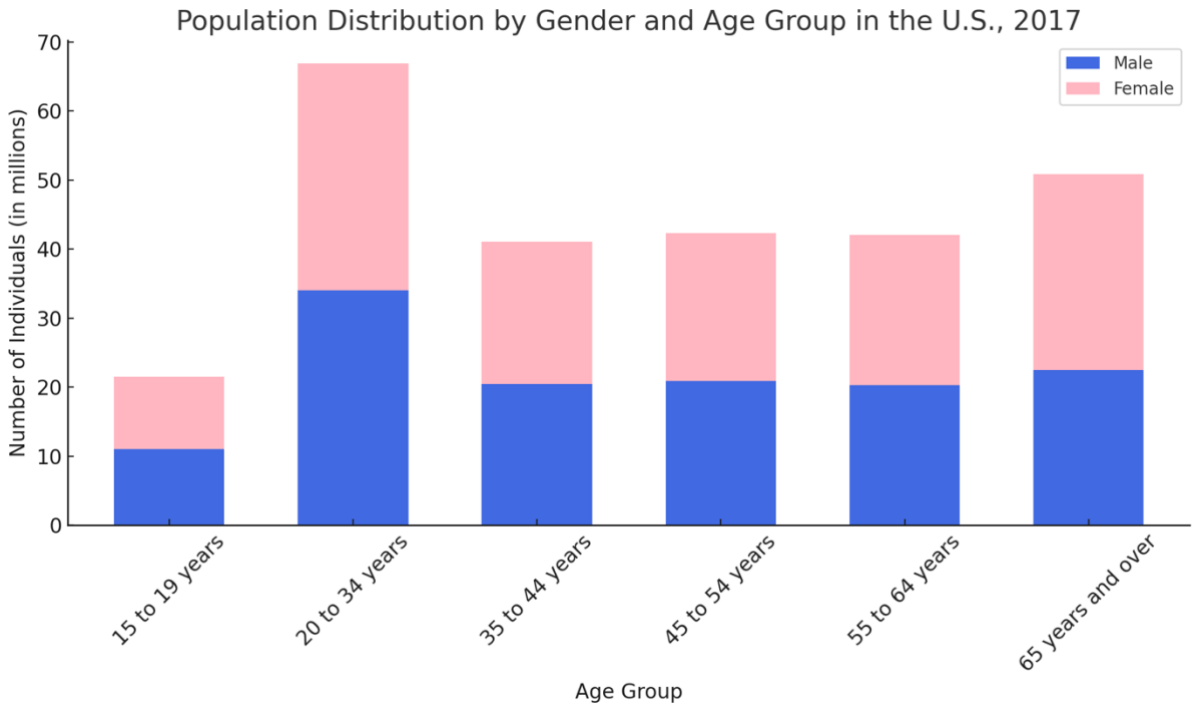
In the provided graph, the distribution of the population is clearly depicted with black bars representing the male population and white bars outlined in black representing the female population. This visual comparison allows us to discern patterns and anomalies within the population demographics. For instance, we observe a significant number of individuals in the "20 to 34 years" age bracket, which could be indicative of the prime marrying age. Moreover, the equal representation of both genders within this age group may reflect the societal norms and patterns regarding marriage or partnership.

The data also allows us to contemplate the broader implications on economic, social, and policy decisions. For example, the substantial population in the working-age groups (20 to 64 years) has direct implications for the labor market, economic growth, and dependency ratios.

This graph, while simple in its aesthetics, serves a dual purpose in the context of a data analytics portfolio. It not only demonstrates the ability to convey information clearly and effectively but also showcases the analytical thought process behind interpreting the data. The story it tells goes beyond mere numbers; it provides insights into the demographic composition of a nation at a specific point in time, which is crucial for informed decision-making in various sectors.

6) A visualization that uses color as an important aesthetics

Taking into consideration the previous uncolored visualization and applying colors to it we come up to this graph.



The graph titled "Population Distribution by Gender and Age Group in the U.S., 2017" utilizes color effectively to convey complex information in an intuitive and accessible manner. The use of royal blue for the male population and light pink for the female population provides a clear visual distinction between the two genders. This color coding is more than just an aesthetic choice; it serves several crucial functions in the interpretation of the data:

1. Quick Recognition: Colors allow viewers to easily distinguish between categories, like male and female in a graph, without text or legends.

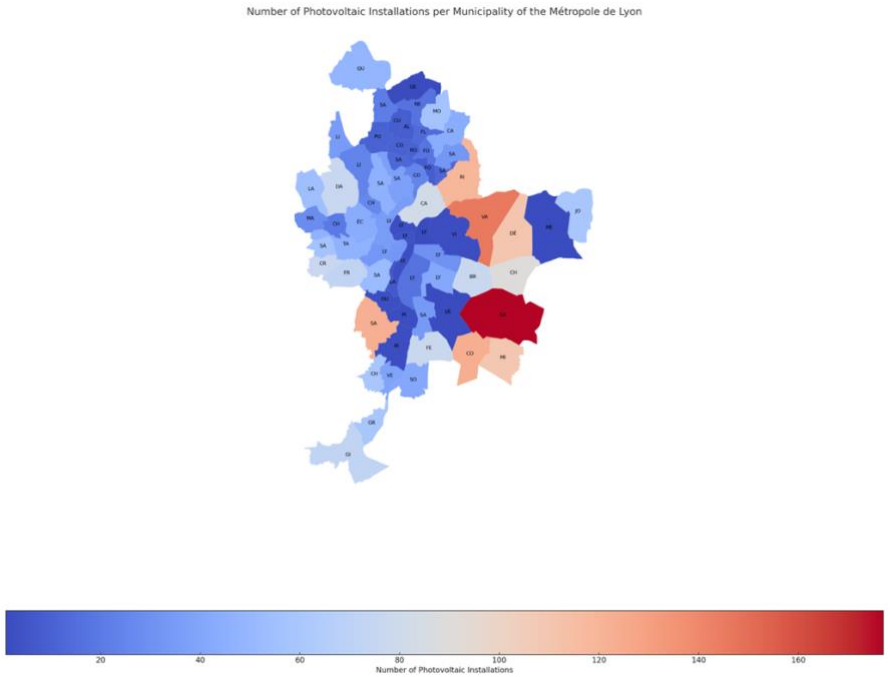
2. Improved Understanding: Color contrasts, such as royal blue and light pink, help compare and comprehend data, showing population differences in each age group clearly.

3. Cultural Impact: Colors carry emotional and cultural meanings. Using blue for males and pink for females, despite being stereotypical, makes the graph more intuitive for a broad audience.

4. Aesthetic Appeal: Strategic color use makes the graph visually attractive, engaging viewers and keeping their attention.

5. Accessibility: The specific blue and pink shades are distinguishable for those with color vision deficiencies, making the graph more inclusive.

In comparison, a black and white graph may lack these benefits. Color adds depth to data, making complex information more understandable and engaging, crucial in data analytics for clear and efficient communication. But this visualization works well even with only black and white colors. But the following one doesn’t.



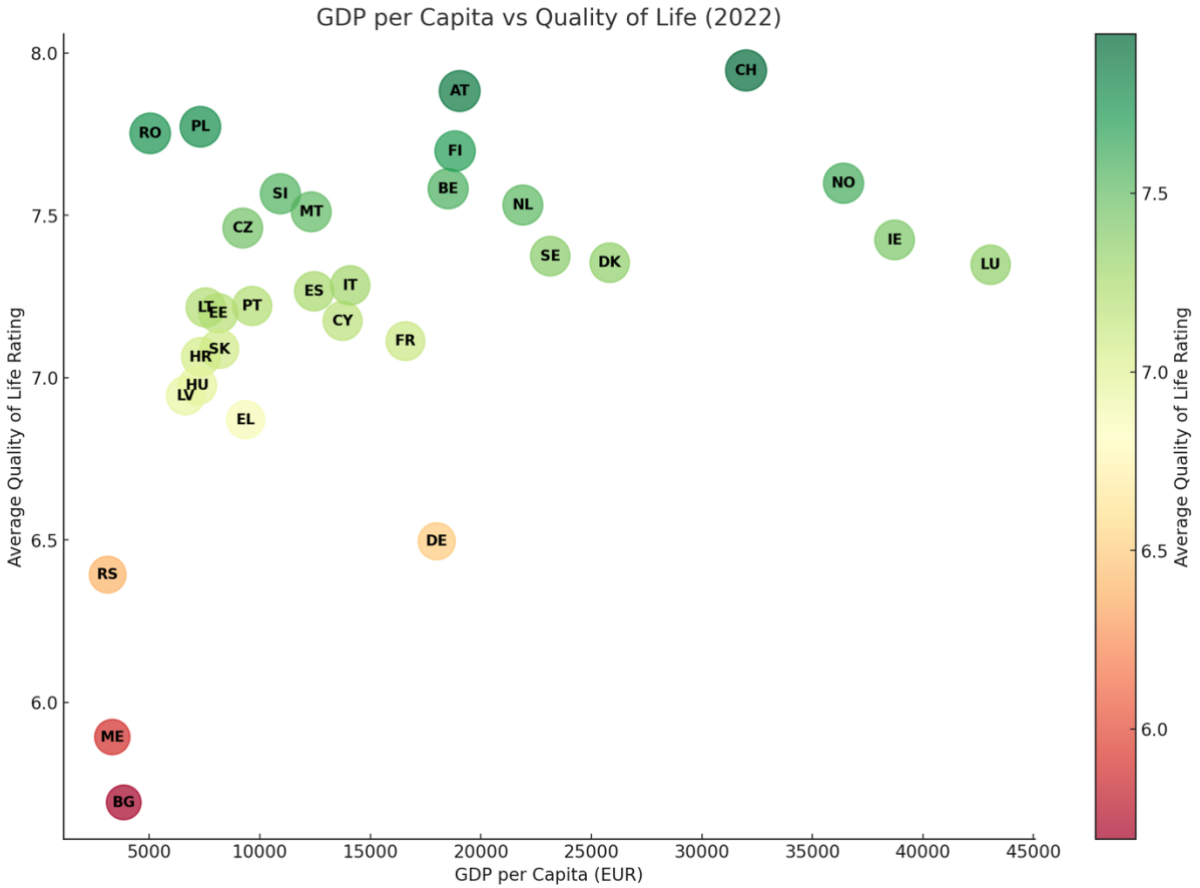
This visualization is based on a dataset obtained from Grand Lyon Data, specifically focusing on the number of photovoltaic production facilities connected to the distribution network across various municipalities. The data is a synthesis from the aggregated dataset of the Register of Electricity and Gas Production Facilities, made available by Distribution System Operators (DSOs) to RTE for open data publication. This dataset tracks the temporal evolution of photovoltaic installations since 2017, with data points for January 2018 (representing 2017), January 2019 (for 2018), January 2020 (for 2019), and March 2020 (for 2020). *Source*

The choropleth map employs color as a critical aesthetic and analytical tool. The 'coolwarm' color palette effectively distinguishes between municipalities with varying densities of photovoltaic installations. Areas with a higher number of installations are represented in warm hues, while those with fewer installations are shown in cooler tones. This color coding is not merely visually striking but also facilitates an immediate and intuitive grasp of the data distribution, enabling viewers to quickly identify regions where solar energy adoption is more prevalent.

Each municipality is labeled with the first two letters of its name, ensuring accurate and easy identification without cluttering the visual space. These annotations are strategically placed at the geographic centroids of the municipalities for precision.

This map is a powerful tool for energy sector stakeholders, policymakers, and the public, as it vividly portrays the current landscape of photovoltaic energy production in the Métropole de Lyon. It highlights areas where solar energy has been effectively harnessed and, conversely, where there is scope for increased solar energy deployment. Such insights are crucial for strategic planning in the energy sector, particularly in the context of advancing sustainable energy solutions.

7) A visualization that uses data from at least two data sources



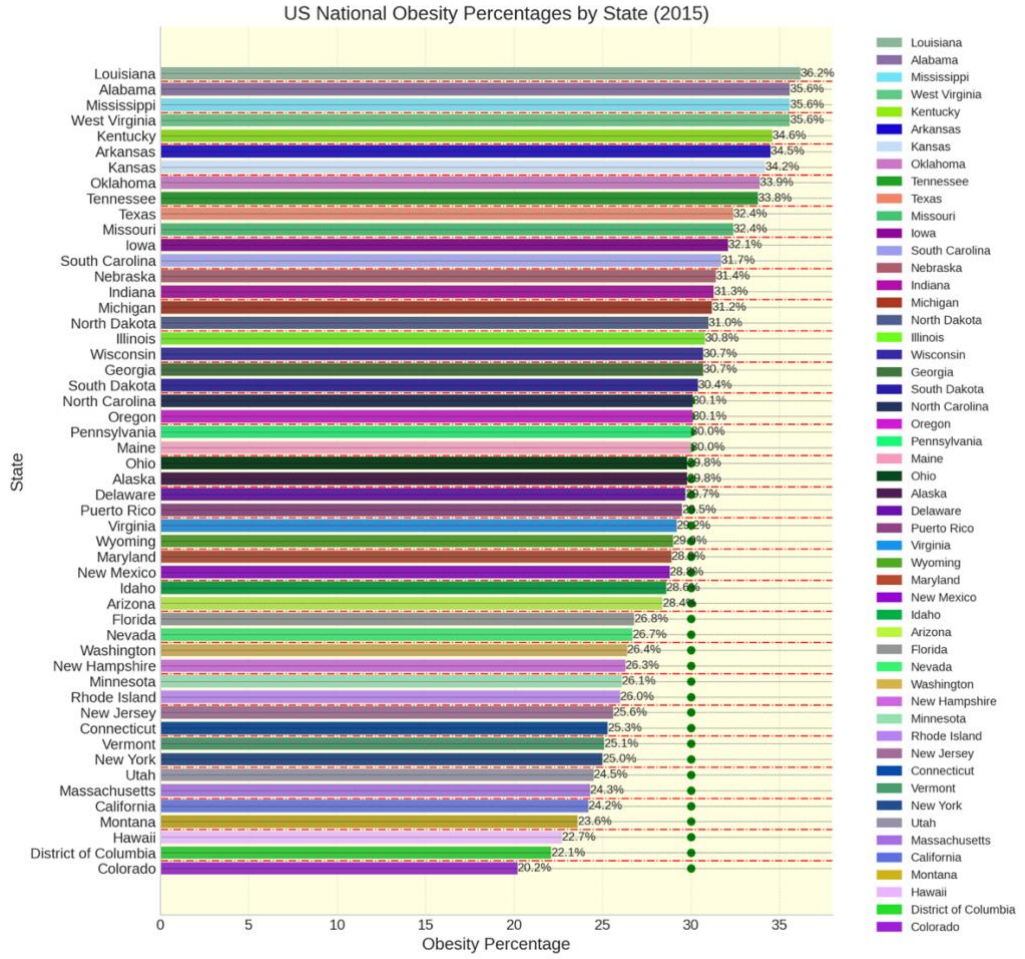
This was created by merging and analyzing data from two distinct sources. This bubble plot depicts the relationship between the economic metric of GDP per Capita and the subjective metric of Quality of Life across various countries in 2022, both sourced from Eurostat.

Analysis Process

- **Data Merging**: The first step in the analysis involved merging the two datasets based on common attributes, namely the country code and the year. This merging allowed for a direct comparison between the GDP per capita and Quality of Life ratings for each country.
- **Data Transformation**: Post-merging, the data was transformed to suit the needs of the visualization. It involved averaging the Quality of Life ratings for each country to provide a singular, comparable metric against the GDP per capita.
- **Visualization Technique**: A bubble plot was chosen for its effectiveness in showing the relationship between two quantitative variables. Each bubble represents a country, with its size proportional to the Quality of Life rating. The color of each bubble ranges from red (lower quality of life) to green (higher quality of life), providing an intuitive visual cue.
- The visualization reveals interesting patterns and correlations between GDP per capita and Quality of Life. While a higher GDP per capita often correlates with a higher Quality of Life rating, there are notable exceptions. This indicates that economic wealth is just one of the many factors influencing the perceived quality of life in a country.

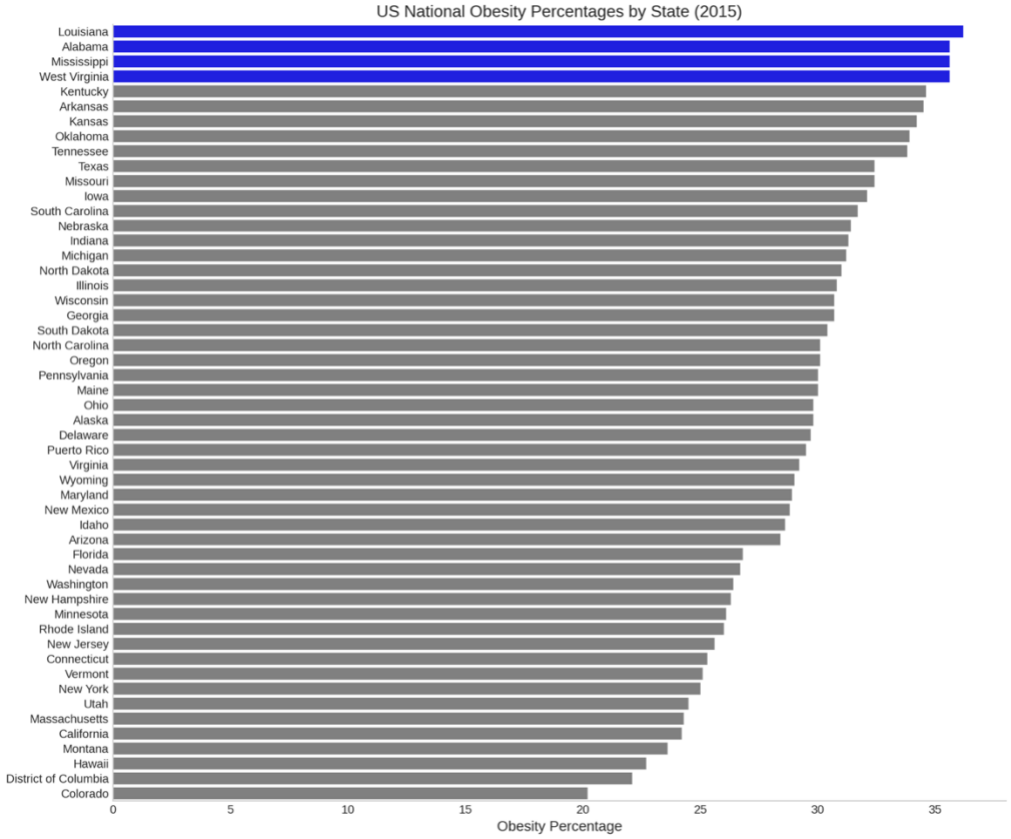
8) A visualization that rigorously maximizes Tufte's "data-ink ratio"

Let's start from the opposite. By minimizing Tufte's "data-ink ratio".



A plethora of non-essential graphical elements are introduced. The use of individual colors for each state, along with additional decorative elements like gridlines, background patterns, and annotations, create a visually overwhelming experience. The core data - the state obesity percentages - become lost amidst the visual clutter. This approach, while aesthetically intricate, hinders the viewer's ability to quickly interpret the critical data.

Let’s try to maximize Tufte’s “data-ink ratio”, but without losing important information.

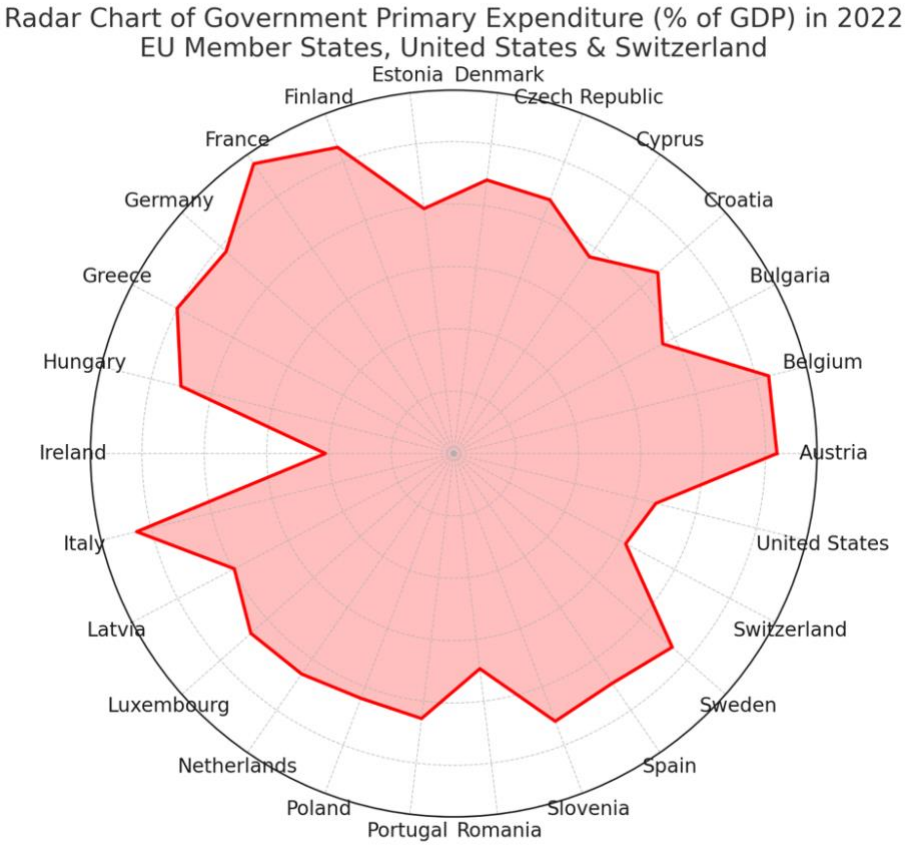


This visualization strips away all non-essential graphical elements. The result is a clean, straightforward bar chart that allows for immediate comprehension of the data. Each state is represented succinctly, and the obesity percentages are clear and unobstructed. This minimalistic design enhances the viewer's ability to quickly grasp the varying obesity rates, highlighting the most critical information without distraction.

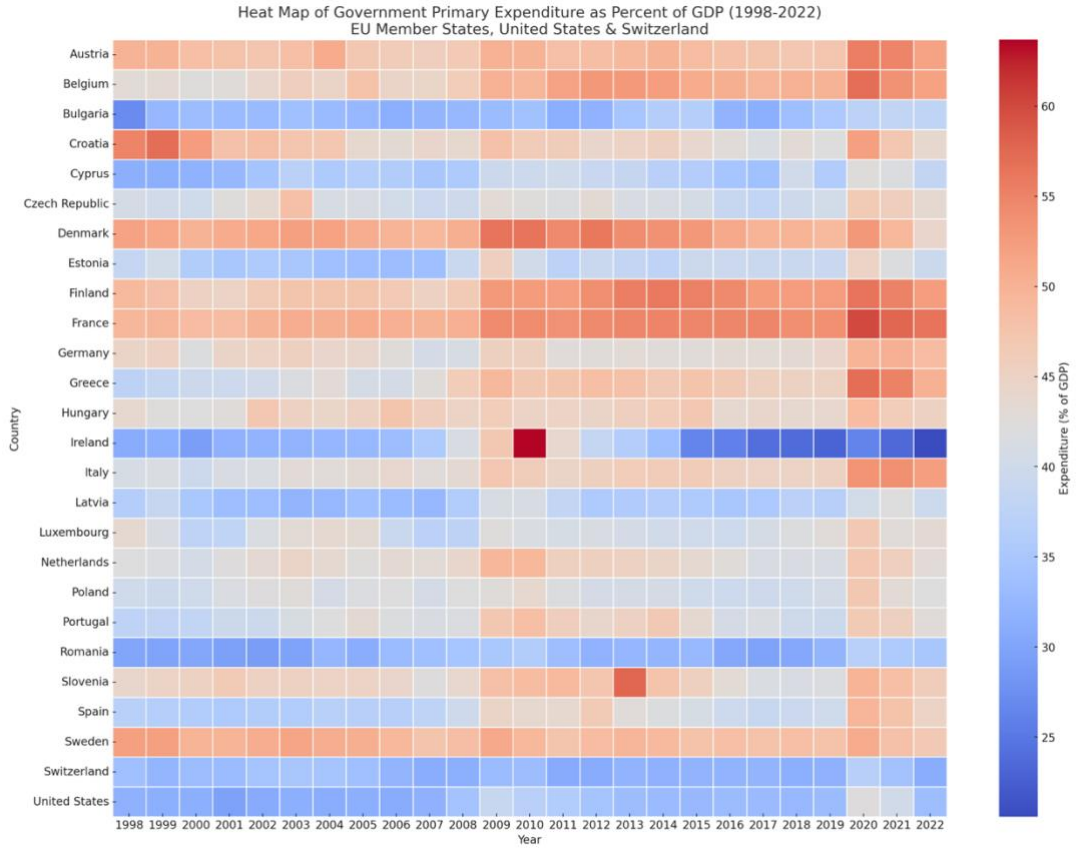
By embracing the principles of Tufte’s "data-ink ratio," analysts can create visualizations that are not only informative but also intuitively understandable, ensuring that the data’s story is heard loud and clear. The dataset used in this visualization project is the "National Obesity Percentages by State" sourced from the 2015 CDC BRFSS Survey. It provides a critical insight into the obesity rates across various states in the United States, an essential aspect of public health understanding.

9) A visualization that is none of the following: map, bar chart, scatter plot, pie chart, doughnut chart, line chart, box plot, density plot, histogram

The following radar chart encapsulates government primary expenditure as a percentage of GDP for the year 2022, focusing on the same set of countries as in the heat map. The data, again obtained from the IMF, is represented in a spider-web-like format, with each spoke standing for a different country.



The chart offers an immediate visual comparison of expenditure levels among these nations. The distance from the center to a point on a country's axis indicates the relative size of that country's government spending in relation to its GDP. This format is particularly effective in showcasing the diversity of fiscal policies and priorities across countries. It's a snapshot of how each country allocated its financial resources in a single year, providing insights into their economic strategies and responses to contemporary global challenges.



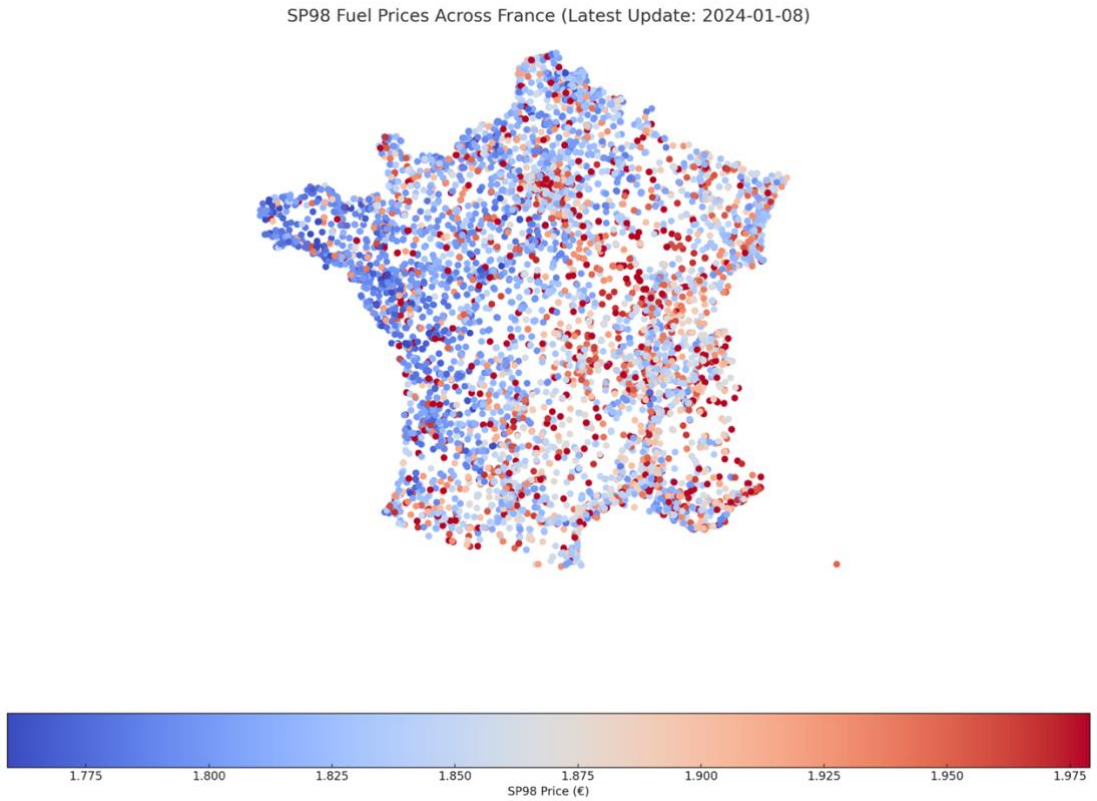
This heat map offers a comprehensive visualization of government primary expenditure as a percentage of GDP, focusing on a selection of countries including EU member states, the United States, and Switzerland, for the period from 1998 to 2022. Sourced

from the International Monetary Fund (IMF), the data reflects the yearly financial commitment of these governments in relation to their economic sizes.

Each cell in the map represents a country-year combination, with the color intensity indicating the expenditure level: warmer tones denote higher percentages, while cooler tones indicate lower percentages. This visualization effectively reveals patterns and fluctuations in government spending, highlighting how economic events like the global financial crisis and the COVID-19 pandemic have influenced fiscal policies. It provides a multi-year, cross-national perspective, essential for understanding the dynamic nature of governmental fiscal behavior in varying economic climates.

10) Data map

This project, centered on the analysis of fuel prices across France, leverages data from the Fuel Price Information System, as mandated by the Ministerial Order of December 12, 2006. The dataset, accessible via the official website *Prix-Carburants* provides a comprehensive overview of fuel prices and related information at various sales points across the country.



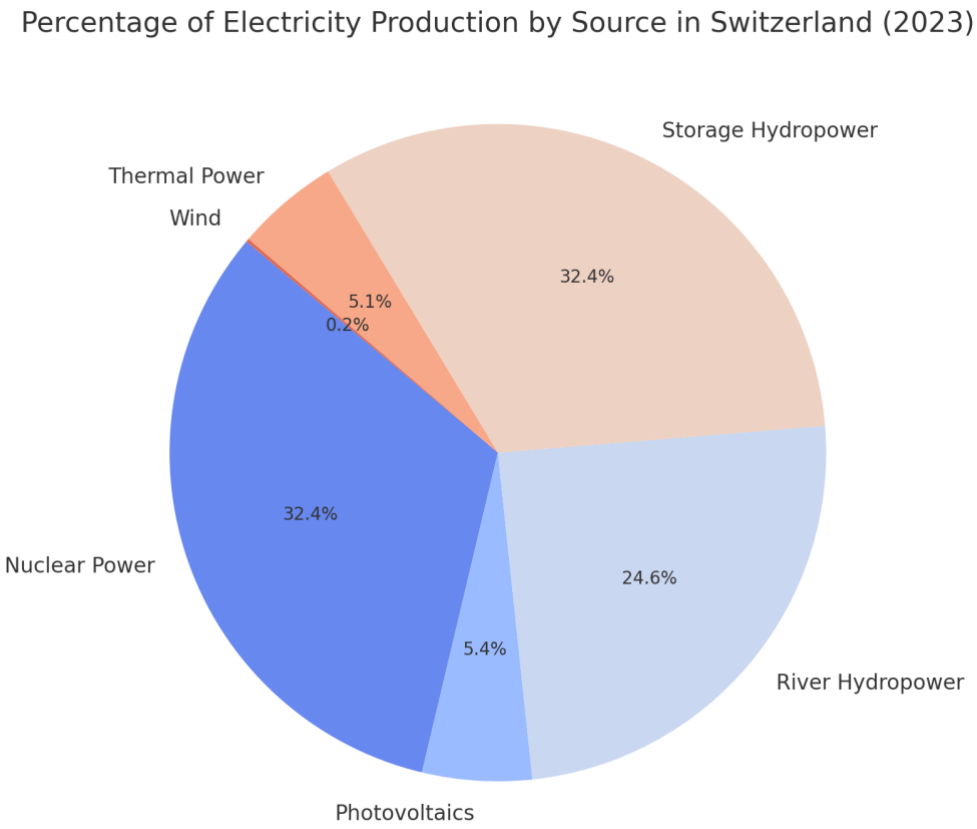
The core of this analysis involved creating a geospatial map to visualize the distribution and variation of SP98 fuel prices throughout France. The map is a result of meticulous data processing and visualization techniques, including:

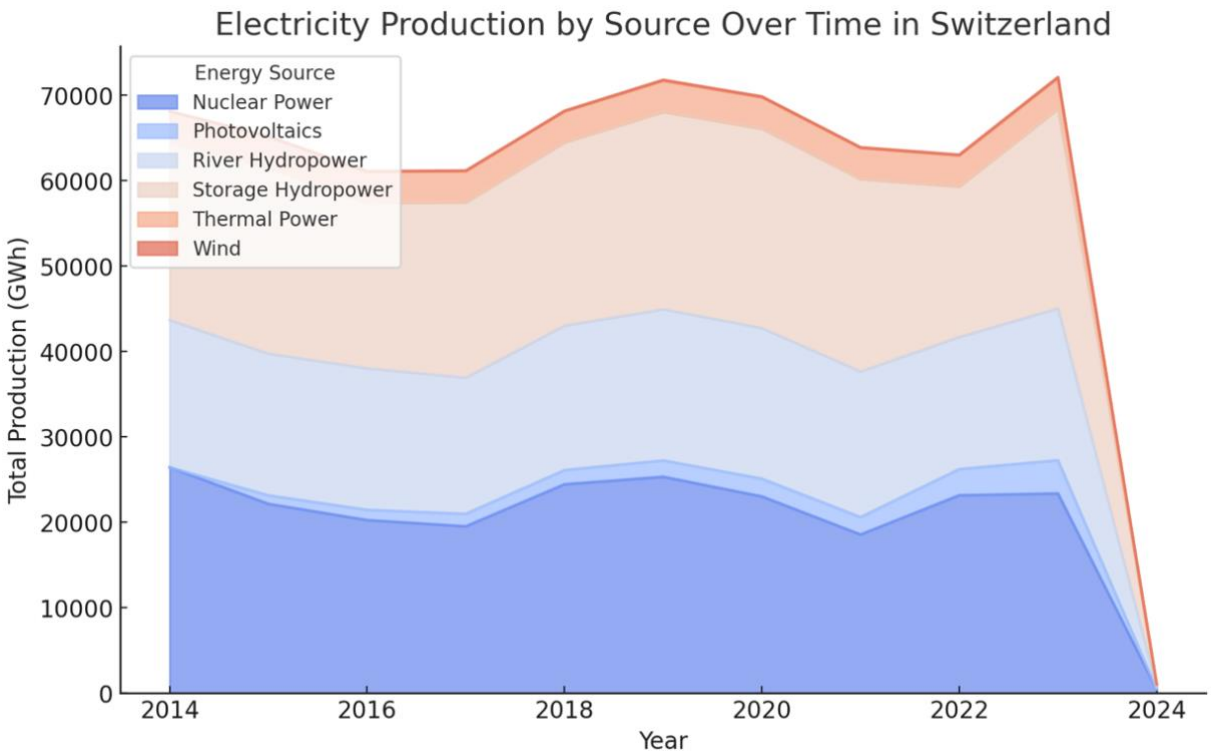
- Geographical plotting of sales points based on latitude and longitude.
- Coloring of points to reflect SP98 fuel prices, with a carefully adjusted color scale to highlight regional price variations.
- Optimization of data representation to maintain clarity and avoid information overload.

This project showcases the power of data visualization in uncovering trends and narratives hidden within complex datasets. The ability to transform raw data into a clear, informative, and visually appealing map provides not only a snapshot of the current state of fuel prices in France but also a template for similar analyses in other sectors or geographical regions. As the dataset continues to evolve, so does the potential for deeper, more nuanced insights, making this an ongoing journey in data exploration and storytelling.

11) Visualization based on Swiss Open Data

The following visualizations provide a comprehensive look at Switzerland's electricity production from various sources over time, using a dataset that distinguishes between nuclear power, thermal power, river hydropower, storage hydropower, wind power, and photovoltaics.



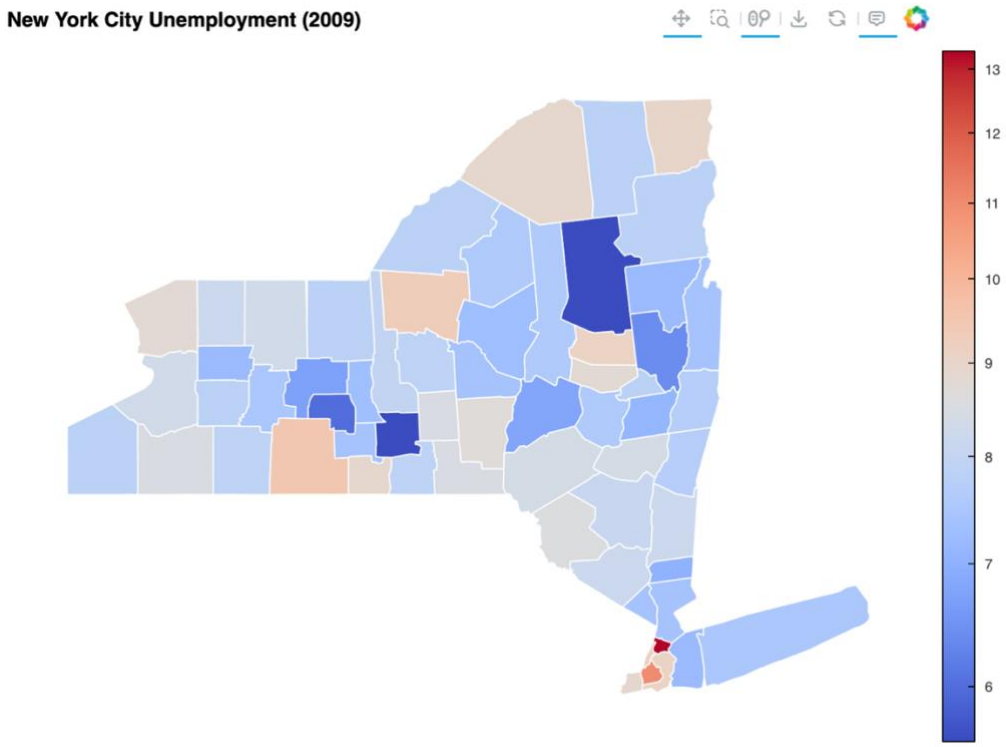


- The dominance of certain energy sources in specific periods, reflecting Switzerland's evolving energy policies and market demands.
- The impact of external factors like technology advancements and environmental policies on the production trends of different energy sources.
- A clear depiction of how renewable energy sources (like wind and solar) have contributed to the overall energy mix over the years.

12) Interactive visualization

The following visualization you've created represents the unemployment rates across different counties in New York State for the year 2009, a time which was heavily influenced by the global financial crisis of 2008, providing a narrative about the economic conditions of that time. The choropleth map, with its varying shades of blue, immediately draws the viewer's attention to the areas with higher unemployment rates, while lighter shades indicate relatively lower rates.

You can find it here: [https://white-bette-81.tiiny.site]



This map tells a story of economic disparity and the varied impact of economic downturns on different regions. The darkest shades, representing the highest unemployment rates, are likely to be areas where industries were hit hardest by the recession. These could be regions reliant on manufacturing, construction, or services that experienced significant declines during this period.

The interactive nature of the map, enhanced by the hover tool, allows viewers to engage with the data on a more personal level. By moving their cursor over each county, they can see the exact unemployment rate, providing a clearer understanding of the economic health of each area.

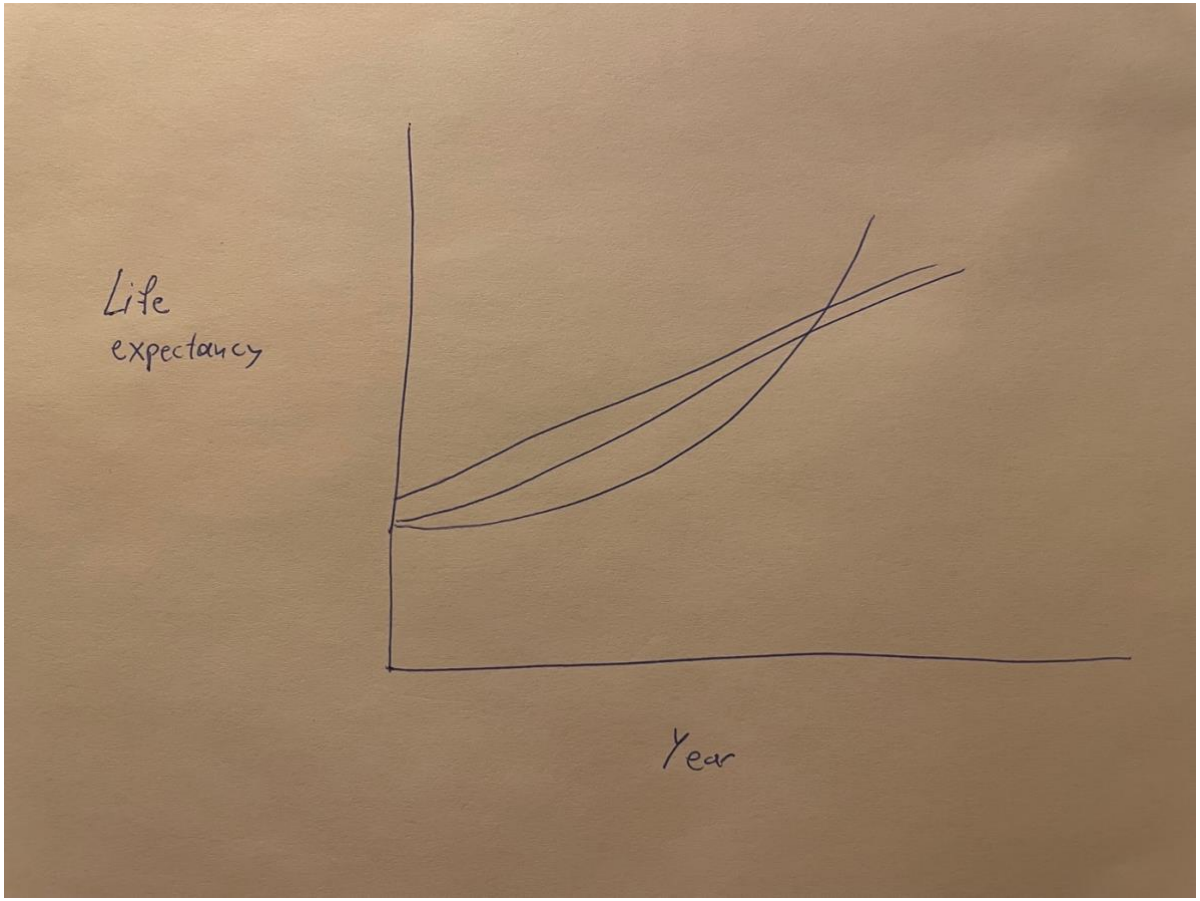
The choropleth map is a suitable choice for geographical data related to regions or demarcations such as counties. It allows for quick visual comparisons across regions.

By using verified data from the Bureau of Labor Statistics and proper percentage scaling, the map accurately reflects the unemployment rates.

This interactive graph was made through the use of HoloViews and Bokeh libraries for Python

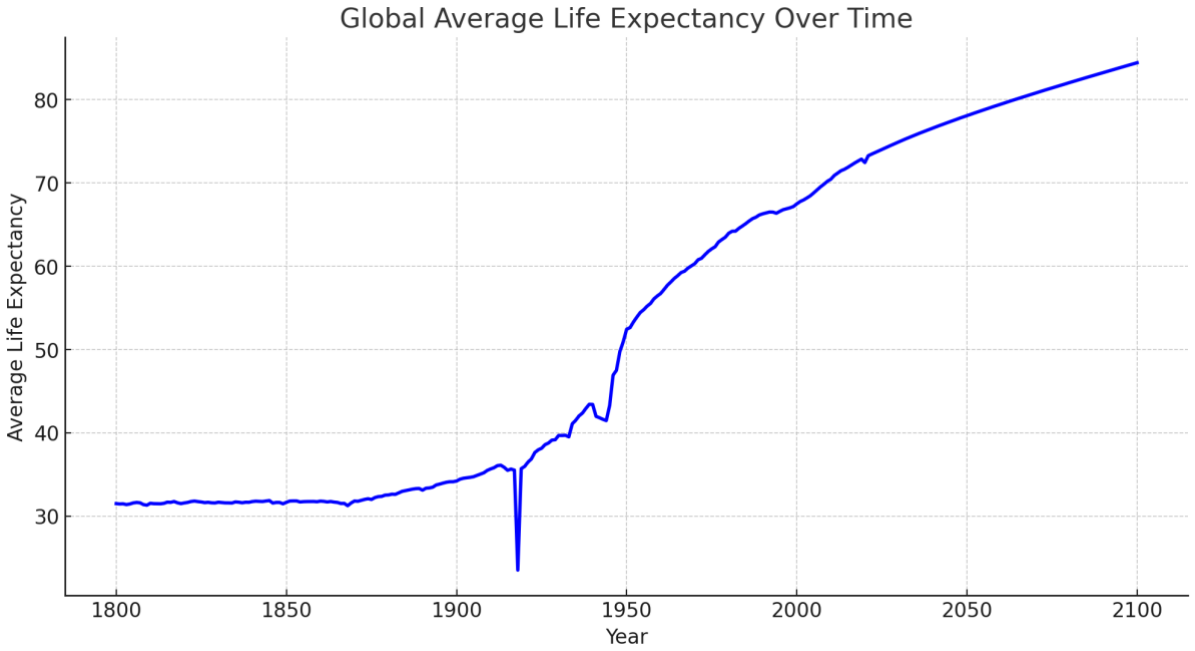
13) Creation process of one visualization to final version

Let’s start exploring the Evolution of Global Life Expectancy. It should probably look like this.

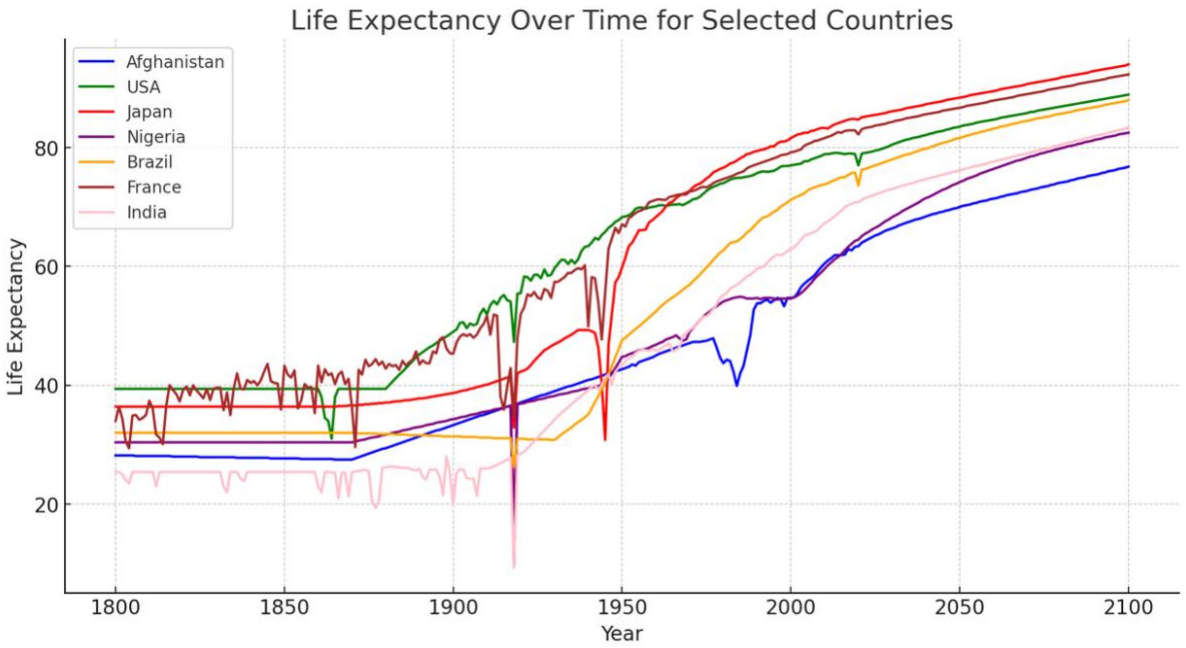


I start by sourcing the data. I found a good dataset from Gapminder. *Source*

My first attempt looked like this.

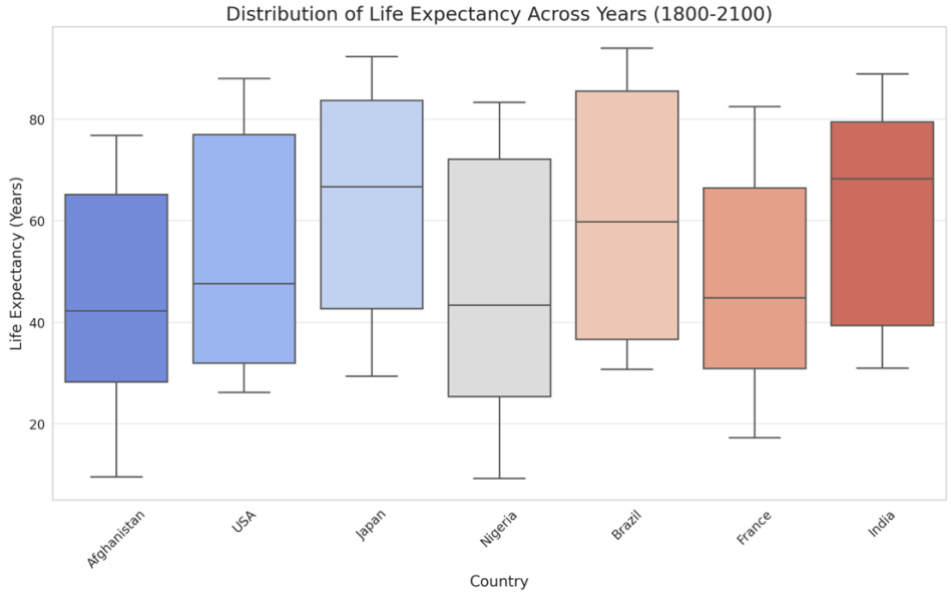


Here is the plot of the global average life expectancy over time. Now let’s add separate countries. The final line plot looks like this.

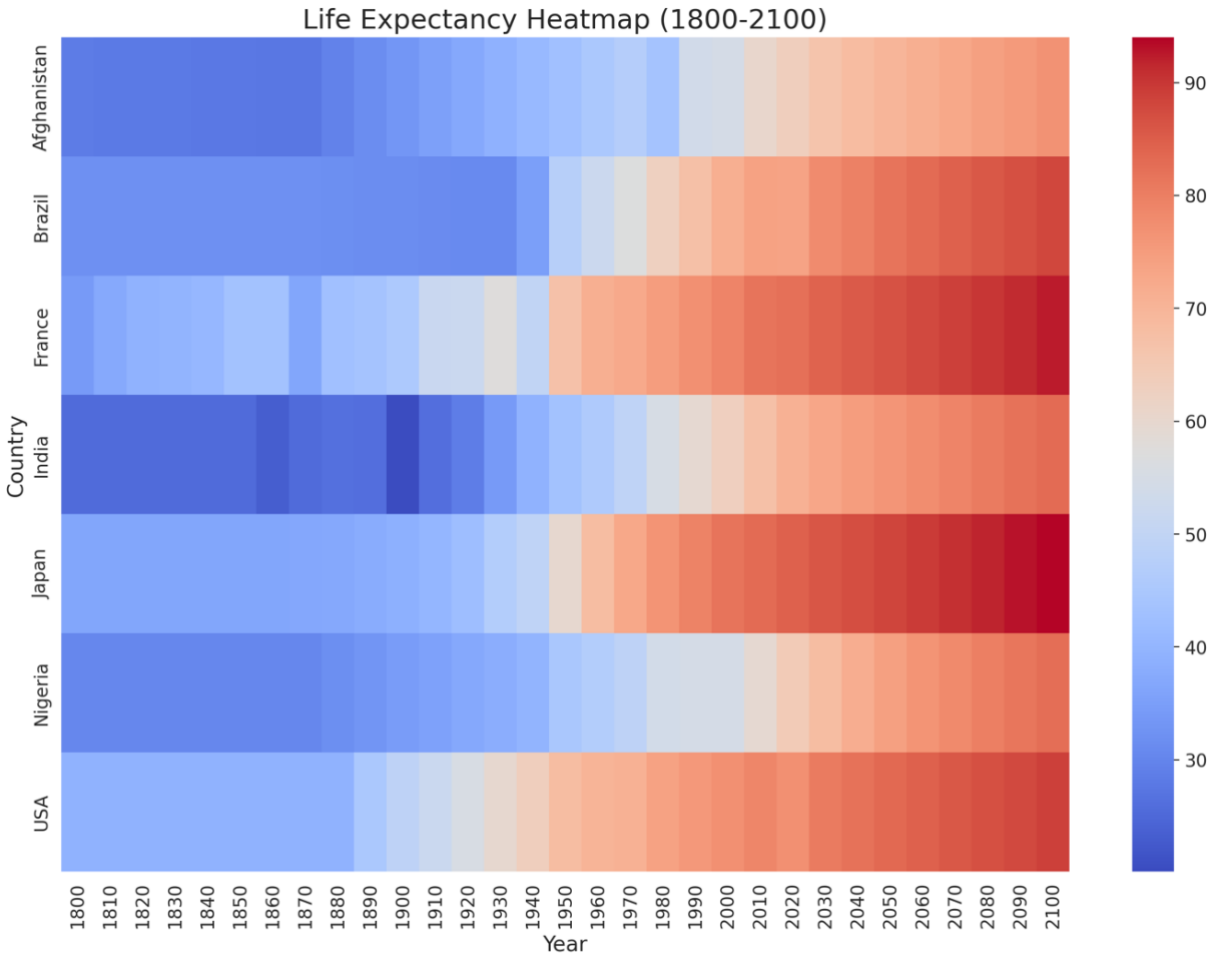


Let’s experiment with more visualizations.

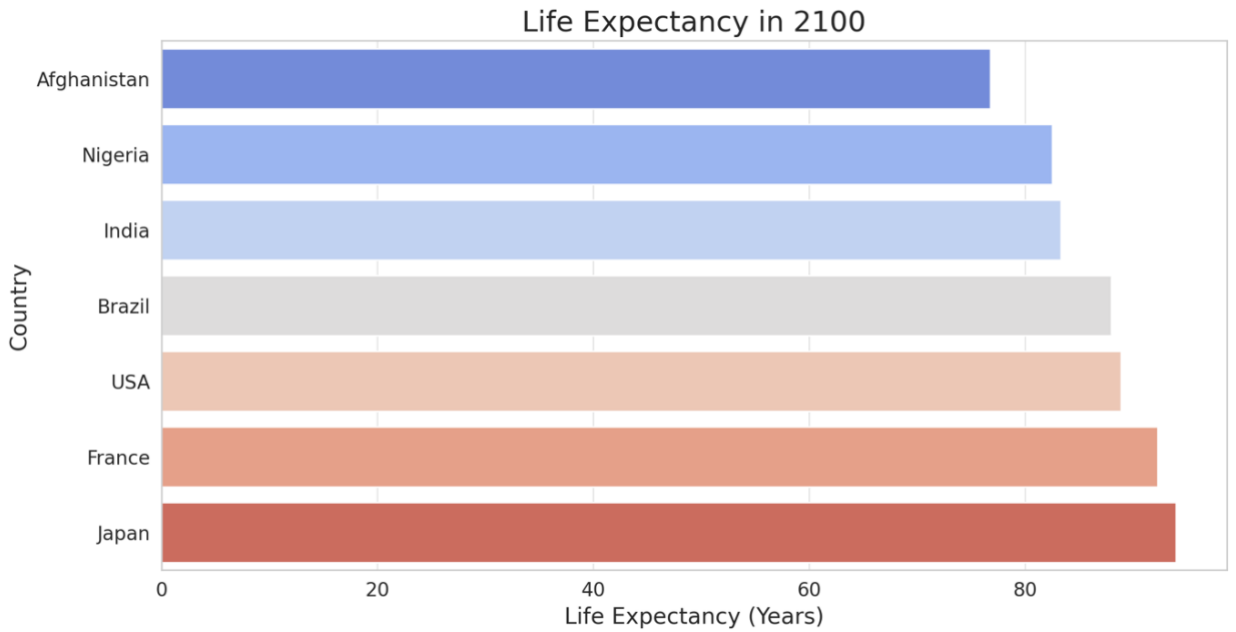

Life Expectancy Distribution Across Years (1800-2100) - Violin Plot
Life Expectancy (Years)
0
20
40
60
80
100
Afghanistan
Brazil
France
India
Japan
Nigeria
USA
Country


Distribution of Life Expectancy Across Years (1800-2100)
Life Expectancy (Years)
20
40
60
80
Afghanistan
USA
Japan
Nigeria
Brazil
France
India
Country

Focusing on the year 2100, bar plots reveal the projected life expectancies, providing a future glimpse. The violin plots add depth to this analysis, showing the distribution of life expectancy over time, highlighting the range and density of values, which reflect periods of stability and turbulence in health trends.



This heatmap shows elegantly the life expectancy per country and makes easy to conclude that developed countries have much higher life expectancies than developing ones. In the following graph, we primarily focus on year 2100. The results are as expected.



These visualizations not only highlight the remarkable progress humanity has made in extending life expectancy but also underscores the disparities and unique challenges faced by different nations. The visualizations collectively tell a story of hope, resilience, and the ongoing quest for longevity. They serve as a reminder of the importance of continued efforts in healthcare, policy-making, and international cooperation to ensure that the benefits of longer life are shared by all.

15) Tools and cheatsheets

Core Principles of Data Visualization

Audience

Always consider your audience —whether they need a short, written report, a more in-depth paper, or an online exploratory data tool.

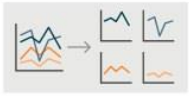
Use pie charts with care

We are not very good at discerning quantities from the slices of the pie chart. Other chart types—for example, bars, stacked bars, treemaps, or slope charts—may be a better choice.

Start bar and column charts at zero

Bar and column charts that do not start at zero overemphasize the differences between the values. For small changes in quantities, consider visualizing the difference or the change in the values.

Try small multiples



Breaking up a complicated chart into smaller chunks can be an effective way to visualize your data.

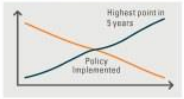
Color and font considerations

Avoid default colors and fonts—they all look the same and don't stand out.

Consider color blindness—about 10% of people (mostly men) have some form of color blindness.

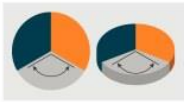
Avoid the rainbow color palette—it doesn't map to our number system and there is no logical ordering.

Include annotation



Add explanatory text to help the reader understand how to read or use the visualization (if necessary) and also to guide them through the content.

Avoid 3D



Using 3D when you don't have a third variable will usually distort the perception of the data and should thus be avoided.

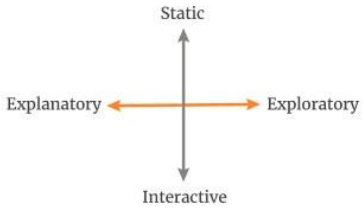
Make labels easy to read

When applicable, rotate bar and column charts to make the labels horizontal. If possible, make vertical axis labels horizontal, possibly below the title. In general, make labels clear, concise, and easy for your reader to understand.

Use maps carefully

Use maps carefully, always being sure it is the geographic point you are trying to make. Column and bar charts, for example, are often better at enabling comparisons between geographic units.

Visualization Mapping: Form and Function

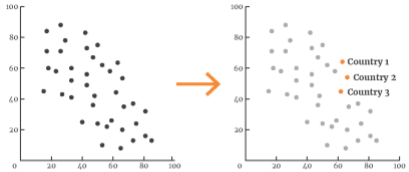


PolicyViz

Core Principles of Data Visualization

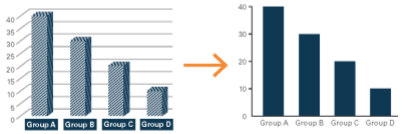
Show the data

People read graphs in a research report, article, or blog to understand the story being told. The data is the most important part of the graph and should be presented in the clearest way possible. But that does not mean that all of the data must be shown—indeed, many graphs show too much.



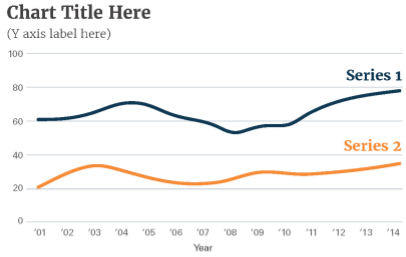
Reduce the clutter

Chart clutter, those unnecessary or distracting visual elements, will tend to reduce effectiveness. Clutter comes in the form of dark or heavy gridlines; unnecessary tick marks, labels, or text; unnecessary icons or pictures; ornamental shading and gradients; and unnecessary dimensions. Too often graphs use textured or filled gradients.



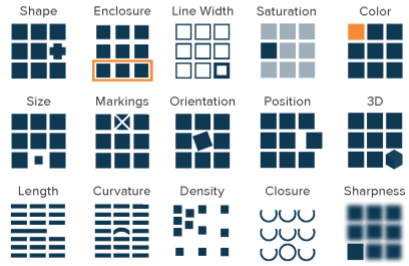
Integrate the text and the graph

Standard research reports often suffer from the **slideshow effect**, in which the writer narrates the text elements that appear in the graph. A better model is one in which visualizations are constructed to complement the text and at the same time to contain enough information to stand alone. As a simple example, legends that define or explain a line, bar, or point are often placed far from the content of the graph—off to the right or below the graph. Integrated legends—right below the title, directly on the chart, or at the end of a line—are more accessible.

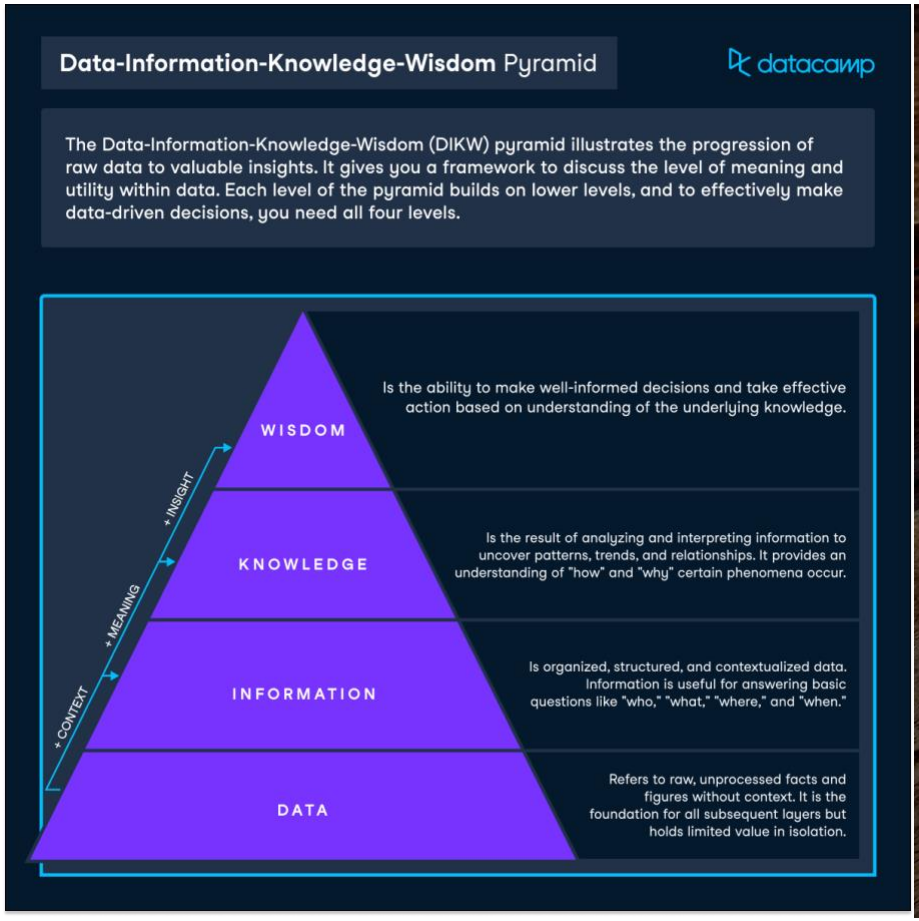


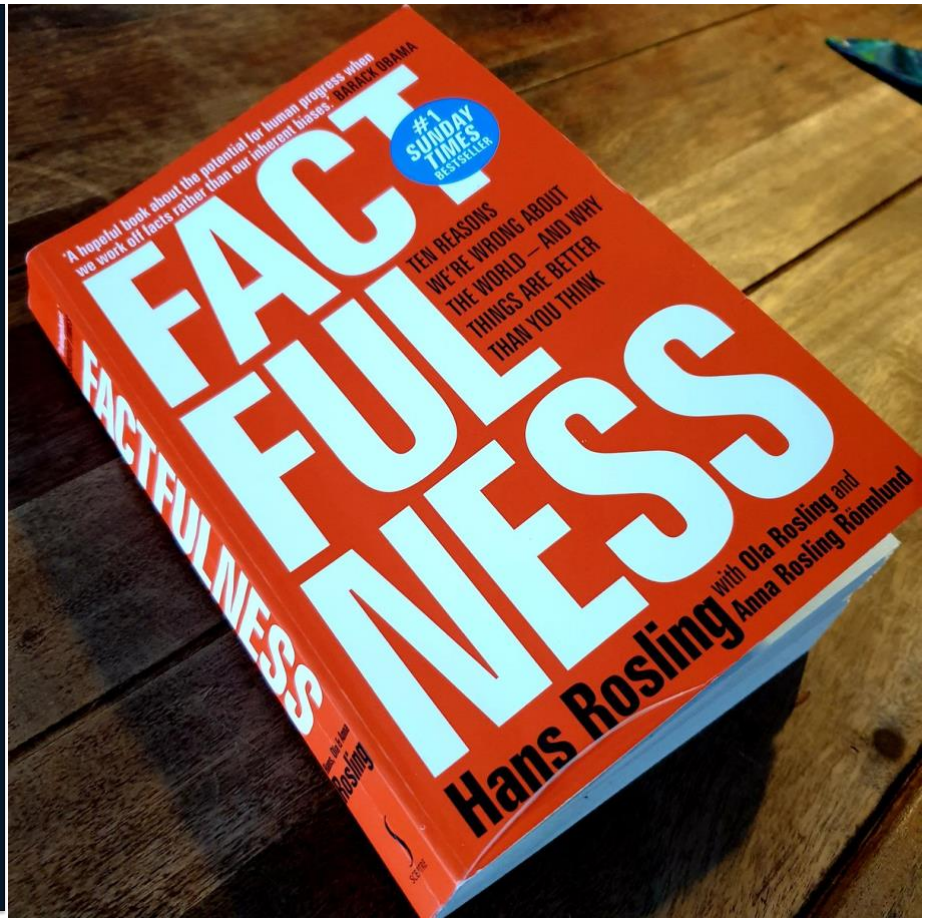
Preattentive Processing

Effective data visualization taps into the brain's **preattentive visual processing.** Because our eyes detect a limited set of visual characteristics (such as shape and contrast), we combine various characteristics of an object and unconsciously perceive them as comprising an image. Preattentive processing refers to the cognitive operations that can be performed prior to focusing attention on any particular region of an image. In other words, it's the stuff you notice right away.



PolicyViz


Data-Information-Knowledge-Wisdom Pyramid
datacamp
The Data-Information-Knowledge-Wisdom (DIKW) pyramid illustrates the progression of raw data to valuable insights. It gives you a framework to discuss the level of meaning and utility within data. Each level of the pyramid builds on lower levels, and to effectively make data-driven decisions, you need all four levels.
WISDOM
Is the ability to make well-informed decisions and take effective action based on understanding of the underlying knowledge.
KNOWLEDGE
Is the result of analyzing and interpreting information to uncover patterns, trends, and relationships. It provides an understanding of "how" and "why" certain phenomena occur.
INFORMATION
Is organized, structured, and contextualized data. Information is useful for answering basic questions like "who," "what," "where," and "when."
DATA
Refers to raw, unprocessed facts and figures without context. It is the foundation for all subsequent layers but holds limited value in isolation.
+ INSIGHT
+ MEANING
+ CONTEXT


"A hopeful book about the potential for human progress when we work off facts rather than our inherent biases." BARACK OBAMA
#1 SUNDAY TIMES BESTSELLER
FACTFULNESS
TEN REASONS WE'RE WRONG ABOUT THE WORLD — AND WHY THINGS ARE BETTER THAN YOU THINK
Hans Rosling
with Ola Rosling and Anna Rosling Rönnlund

16) Reference

https://livefreeordichotomize.com

https://climatedata.imf.org/datasets/543872e1d86c49e3a3bdf38f2b758f92_0/about

https://data.world/throwback-thurs/throwback-thursday-week-30-us-marriage-status-2005-2017

https://ec.europa.eu/eurostat/databrowser/view/sdg_08_10/default/table

https://data.europa.eu/data/datasets/rvss0ooeyeqxujdexk0g?locale=en

https://deliveringdataanalytics.com/the-data-to-ink-ratio-chart-makeover/#:~:text=In%20a%20nutshell%2C%20the%20data,without%20sacrificing%20design%20too%20much

https://www.imf.org/external/datamapper/rgc@FPP/USA/FRA/JPN/GBR/SWE/ESP/ITA/ZAF/IND

https://www.prix-carburants.gouv.fr

https://opendata.swiss/en

https://opendata.swiss/en/dataset/fahrzeiten-2018-der-vbz-im-soll-ist-vergleich-nachfuhrung-eingestellt2

https://www.gapminder.org/resources/

Generative AI was leveraged for text generation.