A PRIMER OF DATA VISUALIZATION

Patrick Boily[1,2,3], Stephen Davies[2,4]

Abstract

Why do we display evidence in a report, in a newspaper article, or online? What is the fundamental goal of our charts and graphs? Representation data properly, in a manner that will allow an audience to gain insight about the underlying situation, is, without a doubt, the most important skill a quantitative consultant must possess. In this chapter, we introduce some commonly-used charts, discuss the fundamental principles of analytical designs, and give a brief overview of dashboards.

Keywords

Data visualization, multivariate charts, analytical design, fundamental principles, dashboards.

Funding Acknowledgement

Parts of this chapter were funded by Carleton University's Centre for Quantitative Analysis and Decision Support.

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada
[2]Data Action Lab, Ottawa, Canada
[3]Idlewyld Analytics and Consulting Services, Wakefield, Canada
[4]DAVHILL Group, Ottawa, Canada
Email: pboily@uottawa.ca

Contents

1. Data and Charts

As data scientist Damian Mingle once put it, modern data analysis is a different beast: "Discovery is no longer limited by the collection and processing of data, but rather management, analysis, and visualization. [17]"

What can be done with the data, once it has been collected/processed? Two suggestions come to mind:

- **analysis** is the process by which we extract actionable insights from the data (this process is discussed in later subsections), while
- **visualization** is the process of presenting data and analysis outputs in a visual format; visualization of data *prior* to analysis can help simplify the analytical process; **post-**analysis, it allows for the results to be communicated to various stakeholders.

In this section, we focus on important visualization concepts and methods; we shall provide examples of data displays to illustrate the various possibilities that might be produced by the data presentation component of a data analysis system.

1.1 Pre-Analysis Use

Even before the analytical stage is reached, data visualization can be used to set the stage for analysis by:

- detecting **invalid entries** and **outliers**;
- shaping the **data transformations** (binning, standardization, dimension reduction, etc.);
- getting a sense for the data (data analysis as an art form, exploratory analysis), and
- identifying **hidden data structures** (clustering, associations, patterns which may inform the next stage of analysis, etc.).

1.2 Presenting Results

The crucial element of data presentations is that they need to help **convey the insight** (or the message); they should be clear, engaging, and (more importantly) readable.

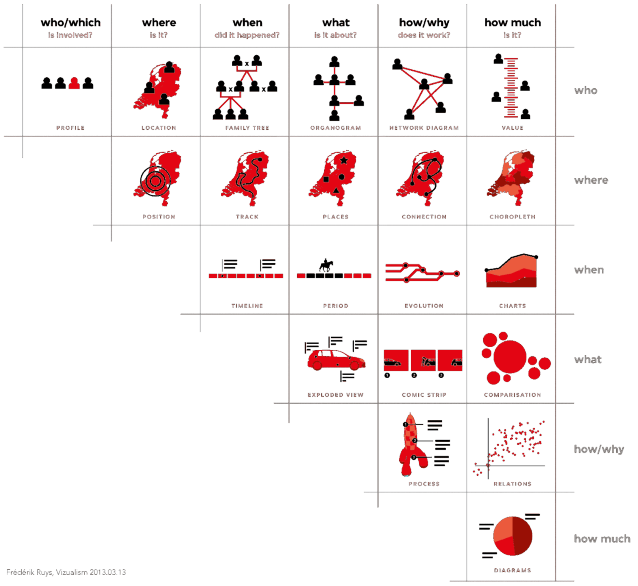


Figure 1. Data visualization suggestions, by type of question (F. Ruys, Vizualism).

Our ability to think of questions (and to answer them) is in some sense limited by what we can visualize. There is always a risk that if certain types of visualization techniques dominate in evidence presentations, the kinds of questions that are particularly well-suited to providing data for these techniques will come to dominate the landscape, which will then affect data collection techniques, data availability, future interest, and so forth.

Generating Ideas and Insights In *Beautiful Evidence* [22], E.Tufte explains that evidence is presented to assist our thinking processes. He further suggests that there is a symmetry to visual displays of evidence – that visualization consumers should be seeking exactly (and explicitly) what the visualization producers should be providing, namely:

- meaningful comparisons;
- causal networks and underlying structure;
- multivariate links;
- integrated and relevant data, and
- a primary focus on content.

More details can be found in Section 2.

Selecting a Chart Type The choice of visualization methods is strongly dependent on the analysis objective, that is, on the **questions that need to be answered**. Presentation methods should not be selected randomly (or simply from a list of easily-produced templates) [1].

In Figure 1, F. Ruys suggests various types of visual displays that can be used, depending on the objective:

- who is involved?
- where is the situation taking place?
- when is it happening?
- what is it about?
- how/why does it work?
- how much?

A general dashboard should at least be able to produce the following types of display:

- **charts** – comparison and relation (scatterplots, bubble charts, parallel coordinate charts, decision trees, cluster plots, trend plots)
- **choropleth maps** (heat maps, classification maps)
- **network diagrams** and connection maps (association rule networks, phrase nets)
- **univariate diagrams** (word clouds, box plots, histograms)

1.3 Multivariate Elements in Charts

At most two fields can be represented by position in the plane. How can we then represent other crucial elements on a flat computer screen? Potential solutions include:

- third dimension
- marker size
- marker colour
- colour intensity and value
- marker texture
- line orientation
- marker shape
- motion/movie

These elements do not always mix well – efficient design is as much art as it is science.

The following examples, along with concise descriptions of key components and lists of questions that they could help answer, highlight charts' strenghts (and limitations). Some additional diagrams showcasing the four presentation types discussed previously are provided in Figures 5 to 11.

Bubble Chart: Health and Wealth of Nations (Figure 2)

- **Data:**
  - 2012 life expectancy in years
  - 2012 inflation adjusted GDP/capita in USD
  - 2012 population for 193 UN members and 5 other countries
- **Some Questions and Comparisons:**
  - Can we predict the life expectancy of a nation given its GDP/capita?
    (*The trend is roughly linear: Expectancy* $\approx 6.8 \times \ln GDP/capita + 10.6$)
  - Are there outlier countries? *Botswana, South Africa, and Vietnam, at a glance*.
  - Are countries with a smaller population healthier? *Bubble size seems uncorrelated with the axes' variates*.

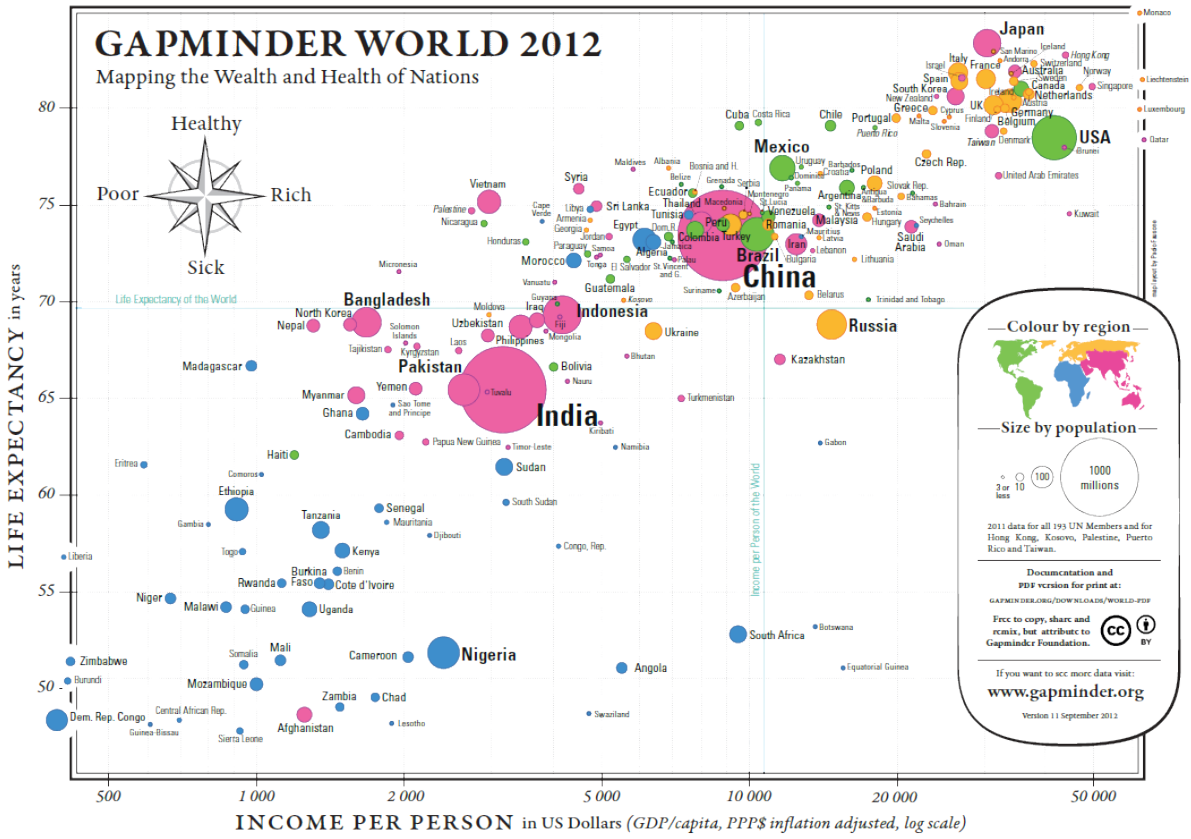


Figure 2. Gapminder's Health and Wealth of Nations (H.Rosling) [20].

- Is continental membership an indicator of health and wealth levels? *There is a clear divide between Western Nations (and Japan), most of Asia, and Africa*.
- How do countries compare against world values for life expectancy and GDP per capita? *The vast majority of countries fall in three of the quadrants. There are very few wealthy countries with low life expectancy. China sits near the world values, which is expected for life expectancy, but more surprising when it comes to GDP/capita – compare with India*.

- **Multivariate Elements:**
  - positions for health and wealth
  - bubble size for population
  - colour for continental membership
  - labels to identify the nations
- **Comments:**
  - Are life expectancy and GDP/capita appropriate proxies for health and wealth?
  - A fifth element could also be added to a screen display: the passage of time. In this case, how do we deal with countries coming into existence (and ceasing to exist as political entities)?

Choropleth Map: Mean Elevation by U.S. State (Figure 3)

- **Data:** 50 observations, ranging from sea level (0-250) to (6000+)
- **Some Questions and Comparisons:**
  - Can the mean elevation of the U.S. states tell us something about the global topography of the U.S.? *West has higher mean elevation related to the presence of the Rockies; Eastern coastal states are more likely to suffer from rising water levels, for instance*.
  - Are there any states that do not "belong" in their local neighbourhood, elevation-wise? *West Virginia and Oklahoma seem to have the "wrong" shade – is that an artifact of the colour gradient and scale in use?*
- **Multivariate Elements:** Geographical distribution and purple-blue colour gradient (as the marker for mean elevation)
- **Comments:**
  - Is the 'mean' the right measurement to use for this map? *It depends on the author's purpose.*
  - Would there be ways to include other variables in this chart? *Population density with texture, for instance*.

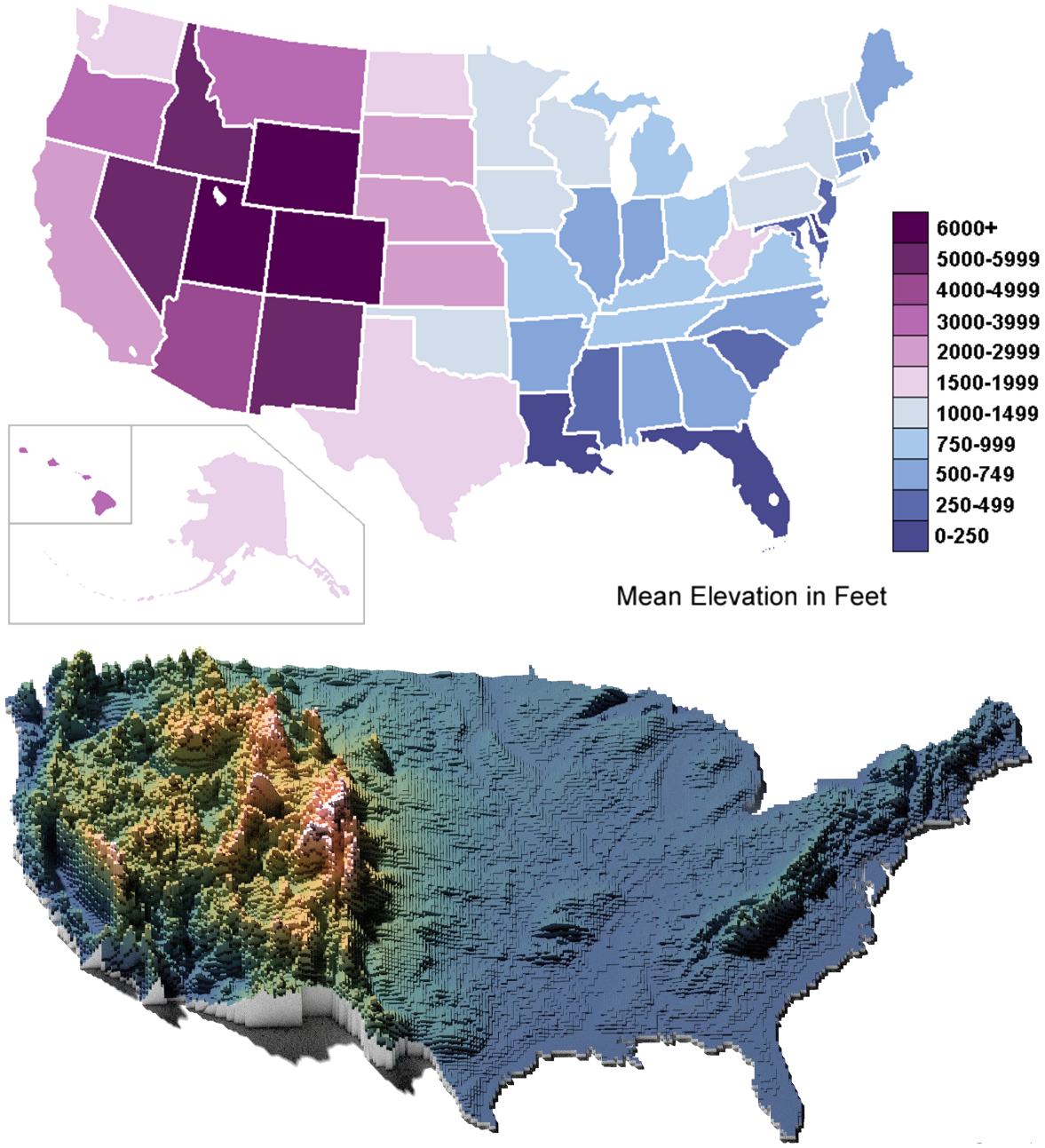


Figure 3. Heat map: mean elevation by U.S. state, in feet (top – source unknown); compare with high resolution elevation map (bottom – by Twitter user @cstats.)

Network Diagram: Lexical Distances (Figure 4)

- **Data:**
  - speakers and language groups for 43 European languages
  - lexical distances between languages
- **Some Questions and Comparisons:**
  - Are there languages that are lexically closer to languages in other lexical groups than to languages in their own groups? *French is lexically closer to English than it is to Romanian, say*.
  - Which language has the most links to other languages? *English has 10 links.*
  - Are there languages that are lexically close to multiple languages in other groups? *Greek is lexically close to 5 groups*.
  - Is there a correlation between the number of speakers and the number of languages in a language group? *Language groups with more speakers tend to have more languages*.
  - Does the bubble size refer only to European speakers? *Portuguese is as large as French?*

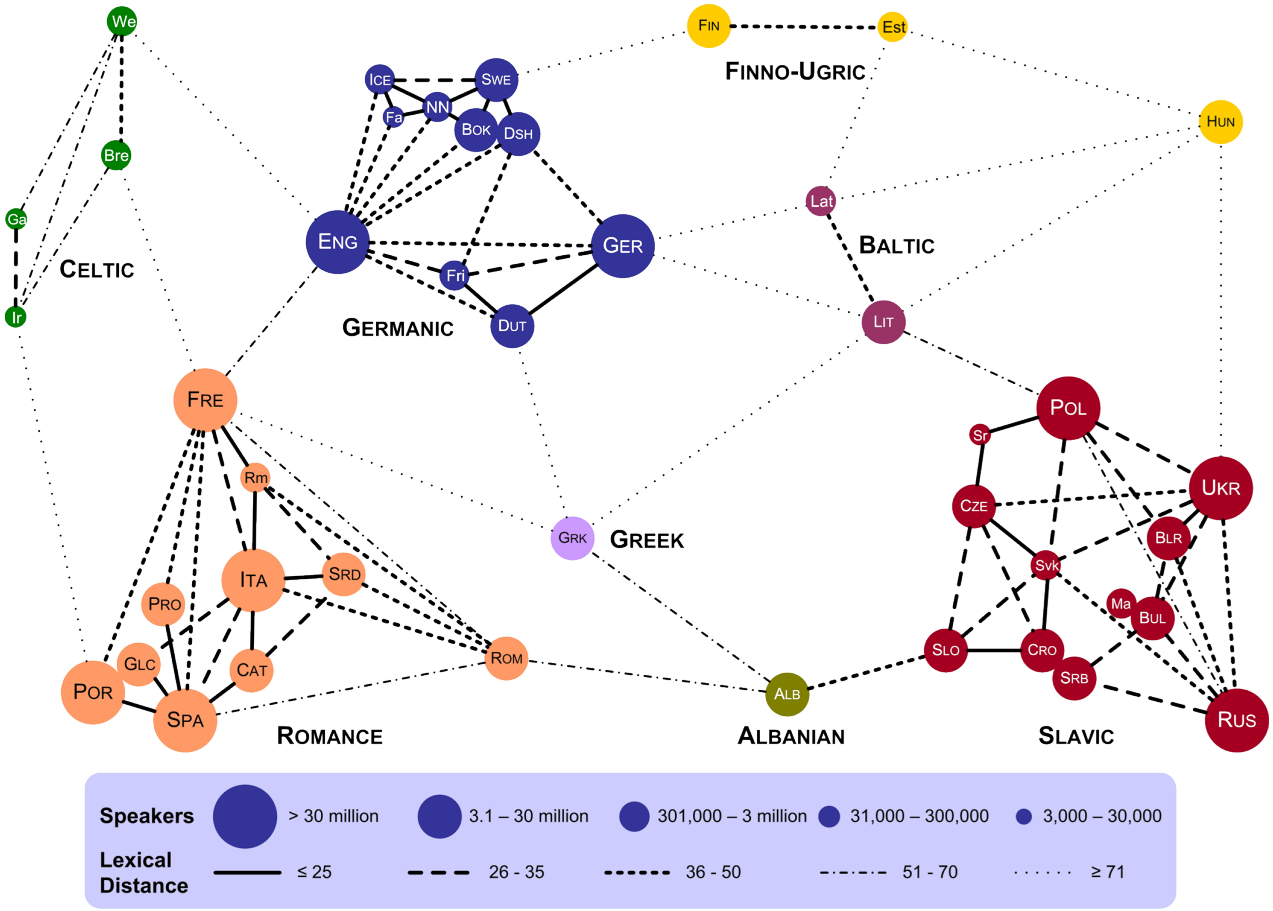


Figure 4. Network diagram: lexical distance of European languages (DVT. Elms, [8]).

- **Multivariate Elements:**
  - colour and cluster for language group
  - line style for lexical distance
  - bubble size for number of speakers
- **Comments:**
  - How is lexical distance computed?
  - Some language pairs are not joined by links – does this mean that their lexical distance is large enough not to be rendered?
  - Are the actual geometrical distances meaningful? For instance, Estonian is closer to French in the chart than it is to Portuguese – is it also lexically closer?

Visualization Catalogue In Figures 5 to 11, we show some other examples of visualizations; more comprehensive catalogues can be found in [1, 1–3, 16, 25], among others.

1.4 A Word About Accessibility

While visual displays can help provide analysts with insight, some work remains to be done in regard to visual impairment – short of describing the features/emerging structures in a visualization, graphs can at best succeed in conveying relevant information to a subset of the population.

The onus remains on the analyst to not only produce **clear** and **meaningful** visualizations (through a clever use of **contrast**, say), but also to describe them and their features in a fashion that allows all to "see" the insights. One drawback is that in order for this description to be done properly, the analyst needs to have seen all the insights, which is not always possible. Examples of "data physicalizations" can be found in [7].

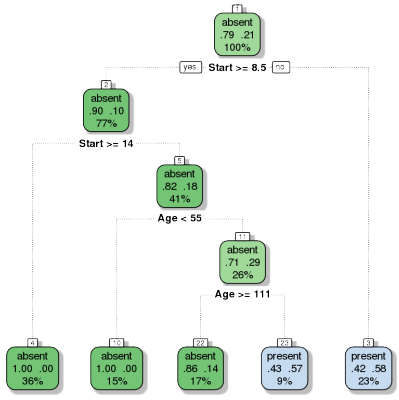


Figure 5. Decision Tree: classification scheme for the kyphosis dataset (personal file).

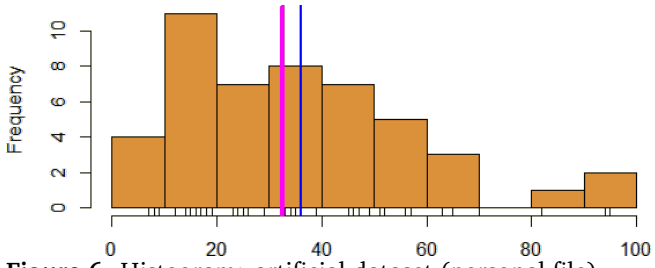


Figure 6. Histogram: artificial dataset (personal file).

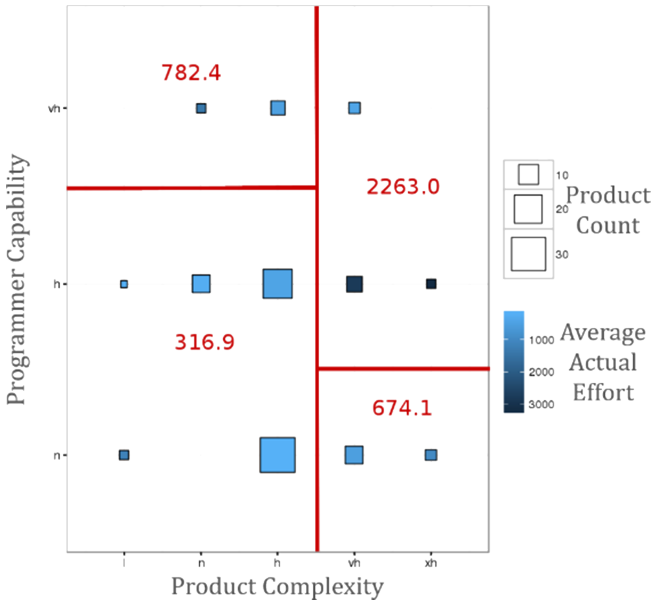


Figure 7. Decision tree bubble chart: estimated average project effort (in red) over-layed over product complexity, programmer capability, and product count in NASA's COCOMO dataset (personal file).

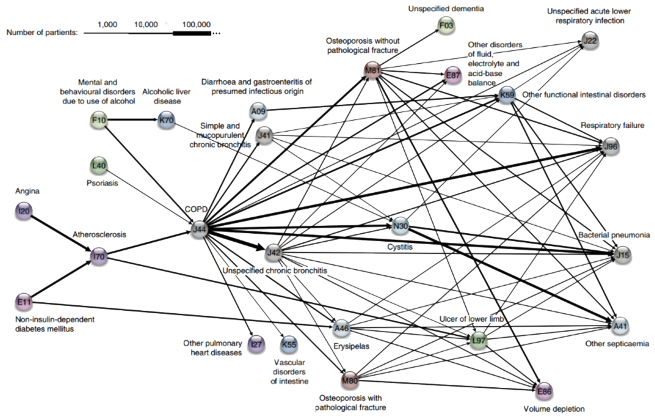


Figure 8. Association rules network: diagnosis network around COPD in the Danish Medical Dataset [14].

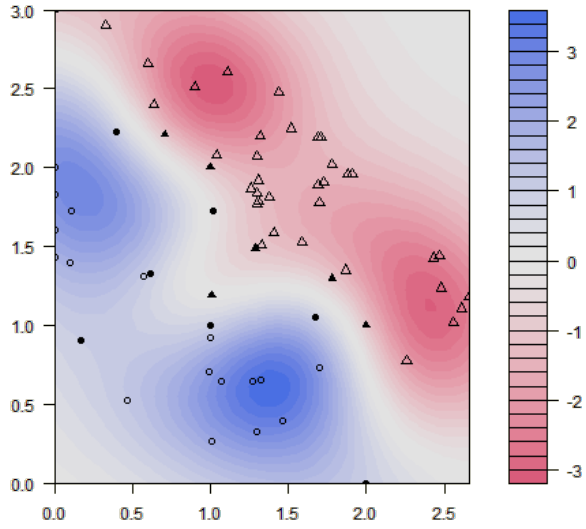


Figure 9. Classification scatterplot: artificial dataset (personal file).

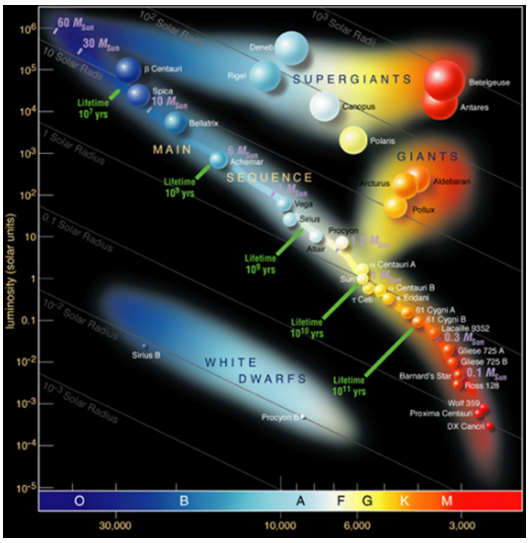


Figure 10. Classification bubble chart: Hertzsprung-Russell diagram (European Southern Observatory).

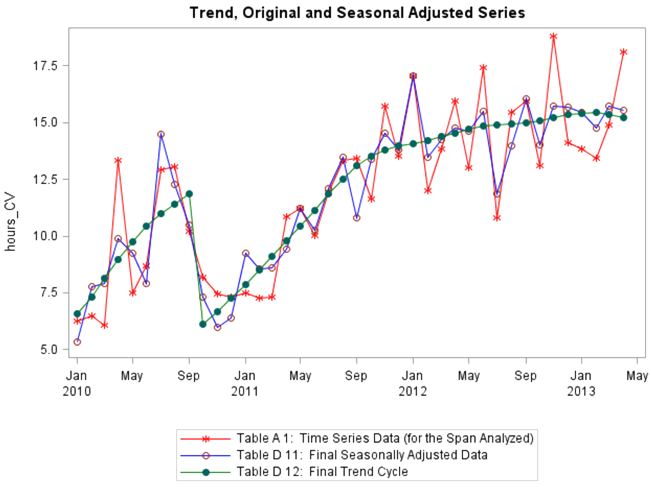


Figure 11. Time series: trend, seasonality (personal file).

2. Fundamental Principles of Analytical Design

In his 2006 offering *Beautiful Evidence*, E.Tufte highlights what he calls the **Fundamental Principles of Analytical Design** [22].

Tufte suggests that we present evidence to assist our thinking processes [22, p.137].

In this regard, his principles are universal – a strong argument can be made that they are dependent neither on technology nor on culture. Reasoning (and communicating our thoughts) is intertwined with our lives in a causal and dynamic multivariate Universe (the 4 dimensions of space-time making up only a small subset of available variates); whatever cognitive skills allow us to live and evolve can also be brought to bear on the presentation of evidence.

Tufte also highlights a particular symmetry to visual displays of evidence, being that **consumers of charts should be seeking exactly what producers of charts should be providing** (more on exactly what that is in a little bit).

Physical science displays tend to be less descriptive and verbal, and more visual and quantitative; up to now, these trends have tended to be reversed when dealing with evidence displays about human behaviour.

In spite of this, Tufte argues that his principles of analytical design can also be applied to social science and medicine. To demonstrate the universality of his principles, he describes in detail how they are applied in a visual display by Charles Joseph Minard .

His lengthy analysis of the image is well worth the read [22, pp.122-139] – it will not be repeated here (we discuss the chart in [5, 6]).

Rather, we will illustrate the principles with the help of the Gapminder's Foundation *Health and Wealth* data visualization (2012) (see Figure 2 in Section 1.3 and a larger version in Figure 12), a bubble chart that plots the 2012 life expectancy, adjusted income per person in USD (log-scaled), population, and continental membership for 193 UN members and 5 other countries, using the latest available data.[1]

Tufte identifies 6 basic properties of superior analytical charts:

- meaningful comparisons
- causal and underlying structures
- multivariate links
- integrated and relevant data
- honest documentation
- primary focus on content

[1] A high-resolution version of the image, as well as more recent charts, can be found on the Gapminder website .

2.1 Comparisons

First Principle

Show comparisons, contrasts, differences. [22, p.127]

Comparisons come in varied flavours: for instance, one could compare a:

- unit at a given time against the same unit at a later time;
- unit's component against another of its components;
- unit against another unit,
- or any number of combinations of these flavours.

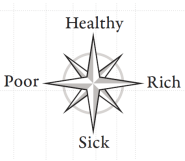
Tufte further explains that

> the fundamental analytical act in statistical reasoning is to answer the question "Compared with what?" Whether we are evaluating changes over space or time, searching big data bases, adjusting and controlling for variables, designing experiments, specifying multiple regressions, or doing just about any kind of evidence-based reasoning, **the essential point is to make intelligent and appropriate comparisons** [emphasis added]. Thus, visual displays [...] should show comparisons. [22, p.127]

Not every comparison will turn out to be insightful, but avoiding comparisons altogether is equivalent to producing a useless display, built from a single datum.

Health and Wealth of Nations First, note that each bubble represents a different country, and that the location of each bubble's centre is a precise point corresponding to the country's life expectancy and its GDP per capita. The size of the bubble correlates with the country's population and its colour is linked to continental membership.

The chart's compass provides a handy comparison tool:



- a bubble further to the right (resp. the left) represents a wealthier (resp. poorer) country;
- a bubble further above (resp. below) represents a healthier (resp. sicker) country.

For instance, a comparison between Japan, Germany and the USA shows that Japan is healthier than Germany, which is itself healthier than the USA, as determined by life expectancy, while the USA is wealthier than Germany, which is itself wealthier than Japan, as determined by GDP per capita (see p. 9, top left).

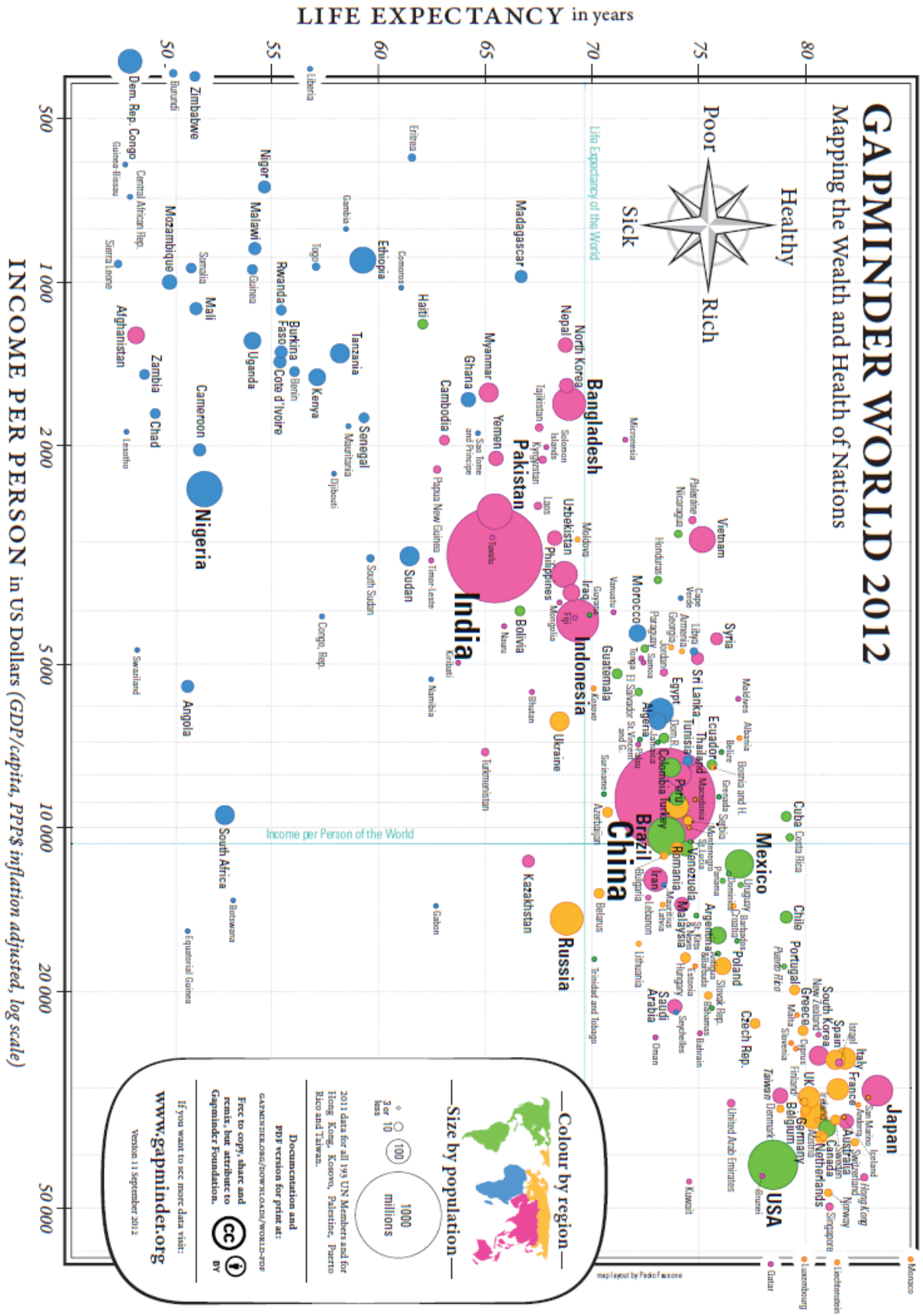
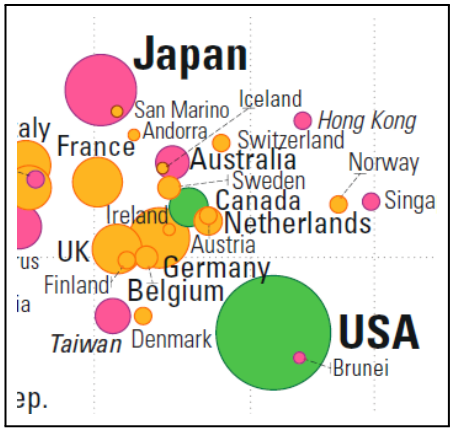
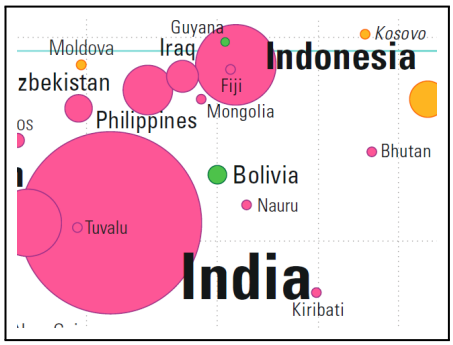


Figure 12. Life expectancy and income per capita in 2012, by nation (Gapminder Foundation) [20].

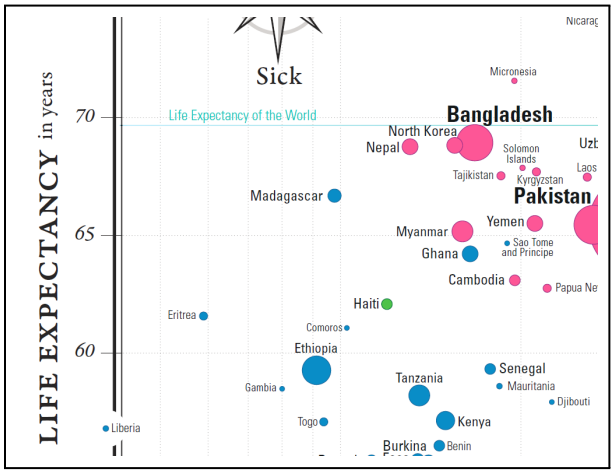


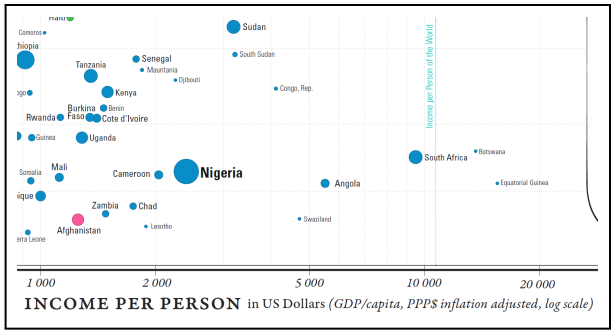
It is possible for two countries to have roughly the same health and the same wealth: consider Indonesia and Fiji, or India and Tuvalu, for instance (see below).



In each pair, the centres of both bubbles (nearly) overlap: any difference in the data must be found in the bubbles' area or in their colour.

Countries can also be compared against world values for life expectancy and GDP per capita (a shade under 70 years and in the neighbourhood of 11K\$, respectively). The world's mean life expectancy and income per person are traced in light blue (see next two images).





Wealthier, **healthier**, **poorer**, and **sicker** are relative terms, but we can also use them to classify the world's nations with respect to these mean values, "wealthier" meaning "wealthier than the average country", and so on.

2.2 Causality, Mechanism, Structure, Explanation

Second Principle

Show causality, mechanism, explanation, systematic structure [22, p.128].

In essence, this is the core principle behind data visualization: the display needs to explain *something*, it needs to provide (potential) links between cause and effect.

As Tufte points out,

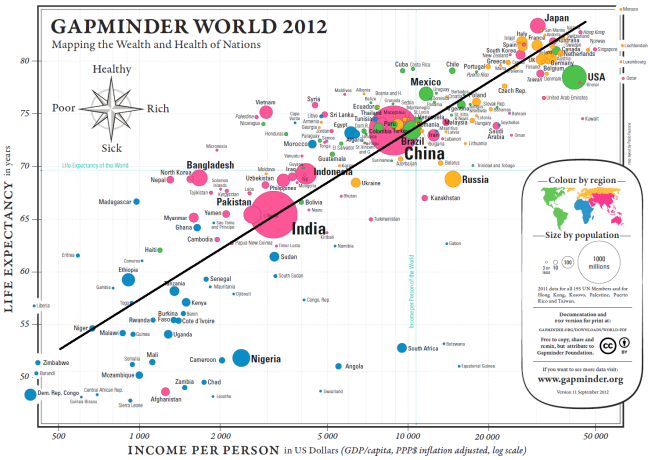
> often **the reason that we examine evidence is to understand causality, mechanism, dynamics, process, or systematic structure** [emphasis added]. Scientific research involves causal thinking, for Nature's laws are causal laws. [...] Reasoning about reforms and making decisions also demands causal logic. To produce the desired effects, we need to know about and govern the causes; thus "policy-thinking is and must be causality-thinking" [22, p.128], [4].

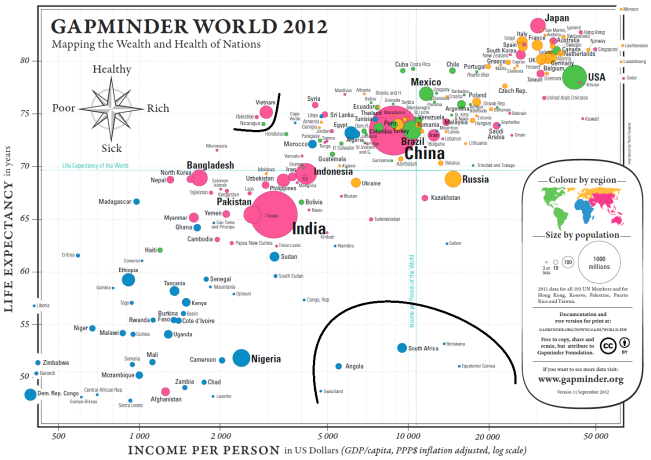
Note also that

> simply collecting data may provoke thoughts about cause and effect: measurements are inherently comparative, and comparisons promptly lead to reasoning about various sources of differences and variability [22, p.128].

Finally, if the visualization can be removed without diminishing the narrative, then that chart should in all probability be excluded from the final product, no matter how pretty and modern it looks, or how costly it was to produce.

Health and Wealth of Nations (continued) At a glance, the relation between life expectancy and the logarithm of the income per person seems to be increasing more or less linearly. Without access to the data, the exact parameter values cannot be estimated analytically, but an approximate line-of-best-fit has been added to the figure (next page).



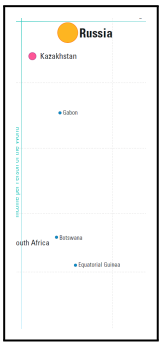


Using the points $(10K, 73.5)$ and $(50K, 84.5)$ yields a line with equation

$$\text{Life Expectancy} \approx 6.83 \times \ln(\text{Income Per Capita}) + 10.55$$

The exact form of the relationship and the numerical values of the parameters are of little significance at this stage – the key insight is that wealthier countries appear to be healthier, generally, and *vice-versa* (although whether wealth drives health, health drives wealth, or some other factor(s) [education?] drive both wealth and health cannot be answered without further analysis and access to knowledge external to the chart).

The chart also highlights an interesting feature in the data, namely that the four quadrants created by separating the data along the Earth's average life expectancy and the GDP per capita for the entire planet do not host the same patterns.



Naïvely, it might have been expected that each of the quadrants would contain about 25% of the world's countries (although the large population of China and India muddle the picture somewhat). However, one quadrant is substantially underrepresented in the visualization. Should it come as a surprise that there are so few "wealthier" yet "sicker" countries? (see image on the left).

It could even be argued that Russia and Kazakhstan are in fact too near the separators to really be considered clear-cut members of the quadrant, so that the overwhelming majority of the planet's countries are found in one of only three quadrants.

In the same vein, when we consider the data visualization as a whole, there seems to be one group of outliers below the main trend, to the right, and to a lesser extent, one group above the main trend, to the left (see top of next column).

These cry out for an explanation: South Africa, for instance, has a relatively high GDP per capita but a low life expectancy (potentially, income disparity between a poor majority and a substantially wealthier minority might help push the bubble to the right, while the lower life expectancy of the majority drives the overall life expectancy to the bottom).

This brings up a crucial point about data visualization: it seems virtually certain that the racial politics of *apartheid* played a major role in the position of the South African outlier, but the chart emphatically DOES NOT provide a proof of that assertion. Charts suggest, but proof comes from deeper domain-specific analyses.

2.3 Multivariate Analysis

Third Principle

Show multivariate data; that is, show more than 1 or 2 variables. [22, p.130]

In an age where data collection is becoming easier by the minute, this seems like a no-brainer: why waste time on uninformative univariate plots? Indeed,

> nearly all the interesting worlds (physical, biological, imaginary, human) we seek to understand are inevitably multivariate in nature. [22, p.129]

Furthermore, as Tufte suggest,

> the analysis of cause and effect, initially bivariate, quickly becomes multivariate through such necessary elaborations as the conditions under which the causal relation holds, interaction effects, multiple causes, multiple effects, causal sequences, sources of bias, spurious correlation, sources of measurement error, competing variables, and whether the alleged cause is merely a proxy or a marker variable (see for instance, [12]). [22, p.129]

While we should not dismiss low-dimensional evidence simply because it is low-dimensional, Tufte cautions that

> **reasoning about evidence should not be stuck in 2 dimensions, for the world we seek to understand is profoundly multivariate** [emphasis added]. [22, p.130]

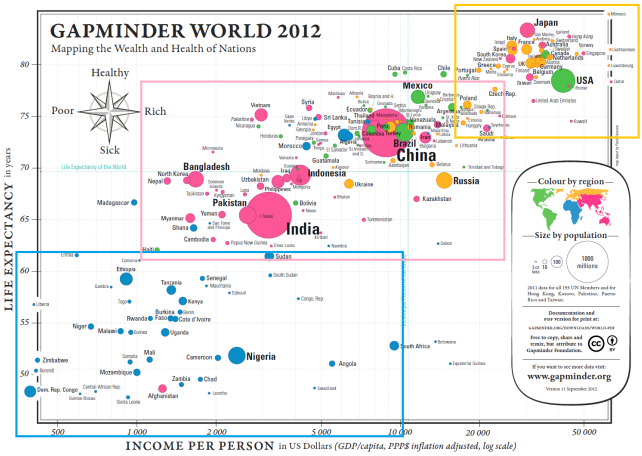
Consultants and analysts may question the ultimate validity of this principle: after all, doesn't Occam's Razor warn us that "it is futile to do with more things that which can be done with fewer"? This would seem to be a fairly strong admonition to not reject low-dimensional visualizations out of hand.

This interpretation depends, of course, on what it means to "do with fewer": are we attempting to "do with **fewer**", or to "**do** with fewer"?

If it is the former, then we can produce simple charts to represent the data (which quickly balloons into a multivariate meta-display), but any significant link between 3 and more variables is unlikely to be shown, which drastically reduces the explanatory power of the charts.

If it is the latter, the difficulty evaporates: we simply retain as many features as are necessary to maintain the desired explanatory power.

Health and Wealth of Nations (continued) Only 4 variables are represented in the display, which we could argue just barely qualifies the data as multivariate. The population size seems uncorrelated with both of the axes' variates, unlike continental membership: there is a clear divide between the West, most of Asia, and Africa (see below). This "clustering" of the world's nations certainly fits with common wisdom about the state of the planet, which provides some level of validation for the display.



Other variables could also be considered or added, notably the year, allowing for bubble movement: one would expect that life expectancy and GDP per capita have both been increasing over time. The Gapminder Foundation's online tool can build charts with other variates, leading to interesting inferences and suggestions.

2.4 Integration of Evidence

Fourth Principle

Completely integrate words, numbers, images, diagrams. [22, p.131]

Data does not live in a vacuum. Tufte's approach is clear:

> the evidence doesn't care what it is – whether word, number, image. **In reasoning about substantive problems, what matters entirely is the evidence, not particular modes of evidence** [emphasis added]. [22, p.130]

The main argument is that evidence from data is better understood when it is presented with context and accompanying meta-data.

Indeed,

> **words, numbers, pictures, diagrams, graphics, charts, tables belong together** [emphasis added]. Excellent maps, which are the heart and soul of good practices in analytical graphics, routinely integrate words, numbers, lineart, grids, measurement scales. [22, p.131]

Finally, Tufte makes the point that we should think of data visualizations and data tables as elements that provide vital evidence, and as such they should be integrated in the body of the text:

> tables of data might be thought of as paragraphs of numbers, tightly integrated with the text for convenience of reading rather than segregated at the back of a report. [...] Perhaps the number of data points may stand alone for a while, so we can get a clean look at the data, although techniques of layering and separation may simultaneously allow a clean look as well as bringing other information into the scene. [22, p.131]

There is a flip side to this, of course, and it is that charts and displays should be annotated with as much text as is required to make the context clear.

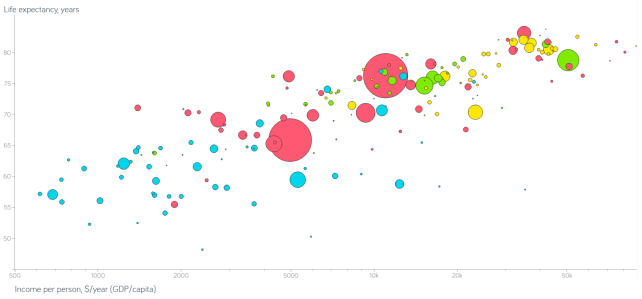
When authors and researchers select a single specific method or mode of information during the inquiries, the focus switches from "can we explain what is happening?" to "can the method we selected explain what is happening?"

There is an art to method selection, and experience can often suggest relevant methods, but remember that "when all one has is a hammer, everything looks like a nail": the goal should be to use whatever (and all) evidence is necessary to shed light on "what is happening".

If that goal is met, it makes no difference which modes of evidence were used.

Health and Wealth of Nations (continued) The various details attached to the chart (such as country names, font sizes, axes scale, grid, and world landmarks) provide substantial benefits when it comes to consuming the display. They may become lost in the background, with the consequence of being taken for granted.

Compare the display obtained from (nearly) the same data, but without integration of evidence (see below).



2.5 Documentation

Fifth Principle

Thoroughly describe the evidence. Provide a detailed title, indicate the authors and sponsors, document the data sources, show complete measurement scales, point out relevant issues. [22, p.133]

We cannot always tell at a glance whether a pretty graphic speaks the truth or presents a relevant piece of information. Documented charts may provide a hint, as

> the credibility of an evidence presentation depends significantly on the quality and integrity of the authors and their data sources. Documentation is an essential mechanism of quality control for displays of evidence. **Thus authors must be named, sponsors revealed, their interests and agenda unveiled, sources described, scales labeled, details enumerated** [emphasis added]. [22, p.132]

Depending on the context, questions and items to address could include:

- What is the title/subject of the visualization?
- Who did the analysis? Who created the visualization? (if distinct from the analyst(s))
- When was the visualization published? Which version of the visualization is rendered here?
- Where did the underlying data come from? Who sponsored the display?
- What assumptions were made during data processing and clean-up?
- What colour schemes, legends, scales are in use in the chart?

It is not obvious whether all this information can fit inside a single chart in some cases. But, keeping in mind the principle of Integration of Evidence, charts should not be presented in isolation in the first place, and some of the relevant information can be provided in the text, on the webpage, or in an accompanying document.

This is especially important when it comes to discussing the methodological assumptions used for data collection, processing, and analysis. An honest assessment may require sizable amounts of text, and it may not be reasonable to include that information with the display (in that case, a link to the accompanying documentation should be provided):

> publicly attributed authorship indicates to readers that someone is taking responsibility for the analysis; conversely, the absence of names signals an evasion of responsibility. [...] **People do things, not agencies, bureaus, departments, divisions** [emphasis added]. [22, p.132-133]

Health and Wealth of Nations (continued) The Gapminder map might just be one of the best-documented charts in the data visualization ecosystem. Let us see if we can answer the questions suggested above.

- **What is the title/subject of the visualization?**
  The health and wealth of nations in 2012, using the latest available data (2011).
- **Who did the analysis? Who sponsored the display? Who created the visualization?**
  The analysis was done by the Gapminder Foundation; the map layout was created by Paulo Fausone. No data regarding the sponsor is found on the chart or in the documentation. It seems plausible that there is no external sponsor, but that is no certainty.
- **When was the visualization published? Which version is rendered here?**
  The 11th version of this chart was published in September 2012.
- **Where did the underlying data come from? What assumptions were made during data processing and clean-up?**
  Typically, the work that goes into preparing the data is swept under the carpet in favour of the visualization itself; there are no explicit source of data on this chart, for instance. However, there is a URL in the legend box that leads to detailed information. For most countries, life expectancy data was collected from:
  - the Human Mortality database,
  - the UN Population Division World Population Prospects,
  - files from historian James C. Riley,
  - the Human Life Table database,
  - data from diverse national statistical agencies,

- the CIA World Fact book,
- the World Bank, and
- the South Sudan National Bureau of Statistics.

Benchmark 2005 GDP data was derived via regression analysis from International Comparison Program data for 144 countries, and extended to other jurisdictions using another regression against data from

- the UN Statistical Division,
- Maddison Online,
- the CIA World Fact book,and
- estimates from the World Bank.

The 2012 values were then derived from the 2005 benchmarks using long-term growth rates estimate from

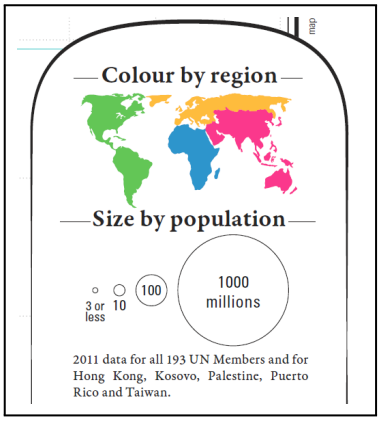
- Maddison Online,
- Barro & Ursua,
- the United Nations Statistical Division,
- the Penn World Table (mark 6.2),
- the International Monetary Fund's World Economic Outlook database,
- the World Development Indicators,
- Eurostat, and
- national statistical offices or some other specific publications.

Population estimates were collated from

- the United Nations Population Division World Population Prospects,
- Maddison Online,
- Mitchell's International Historical Statistics,
- the United Nations Statistical Division,
- the US Census Bureau,
- national sources,
- undocumented sources, and
- "guesstimates".

Exact figures for countries with a population below 3 million inhabitants were not needed as this marked the lower end of the chart resolution.

- **What colour schemes, legends, scales are in use in the chart?**
  The *Legend Inset* is fairly comprehensive (see below):



Perhaps the last item of note is that the scale of the axes differs: life expectancy is measured linearly, whereas GDP per capita is measured on a logarithmic scale.

2.6 Content Counts Most of All

Sixth Principle

Analytical presentations ultimately stand of fall depending on the quality, relevance, and integrity of their content. [22, p.136]

Any amount of time and money can be spent on graphic designers and focus groups, but

> the most effective way to improve a presentation **is to get better content** [emphasis added] [...] design devices and gimmicks cannot salvage failed content. [...] The first questions in constructing analytical displays are not "How can this presentation use the color purple?" Not "How large must the logotype be?" Not "How can the presentation use the Interactive Virtual Cyberspace Protocol Display Technology?" Not decoration, not production technology. The first question is "**What are the content-reasoning tasks that this display is supposed to help with?**" [22, p.136]

The main objective is to produce a compelling narrative, which may not necessarily be the one that was initially expected to emerge from a solid analysis of sound data. Simply speaking, the visual display should assist in explaining the situation at hand and in answering the original questions that were asked of the data.

Health and Wealth of Nations (continued) How would we answer the following questions:

- Do we observe similar patterns every year?
- Does the shape of the relationship between life expectancy and log-GDP per capita vary continuously over time?
- Do countries ever migrate large distances in the display over short periods?
- Do exceptional events affect all countries similarly?
- What are the effects of secession or annexation?

The 2012 Health and Wealth of Nations data represent a single datum in the general space of data visualizations; in this context, getting better content means getting data for other years as well as for 2012.

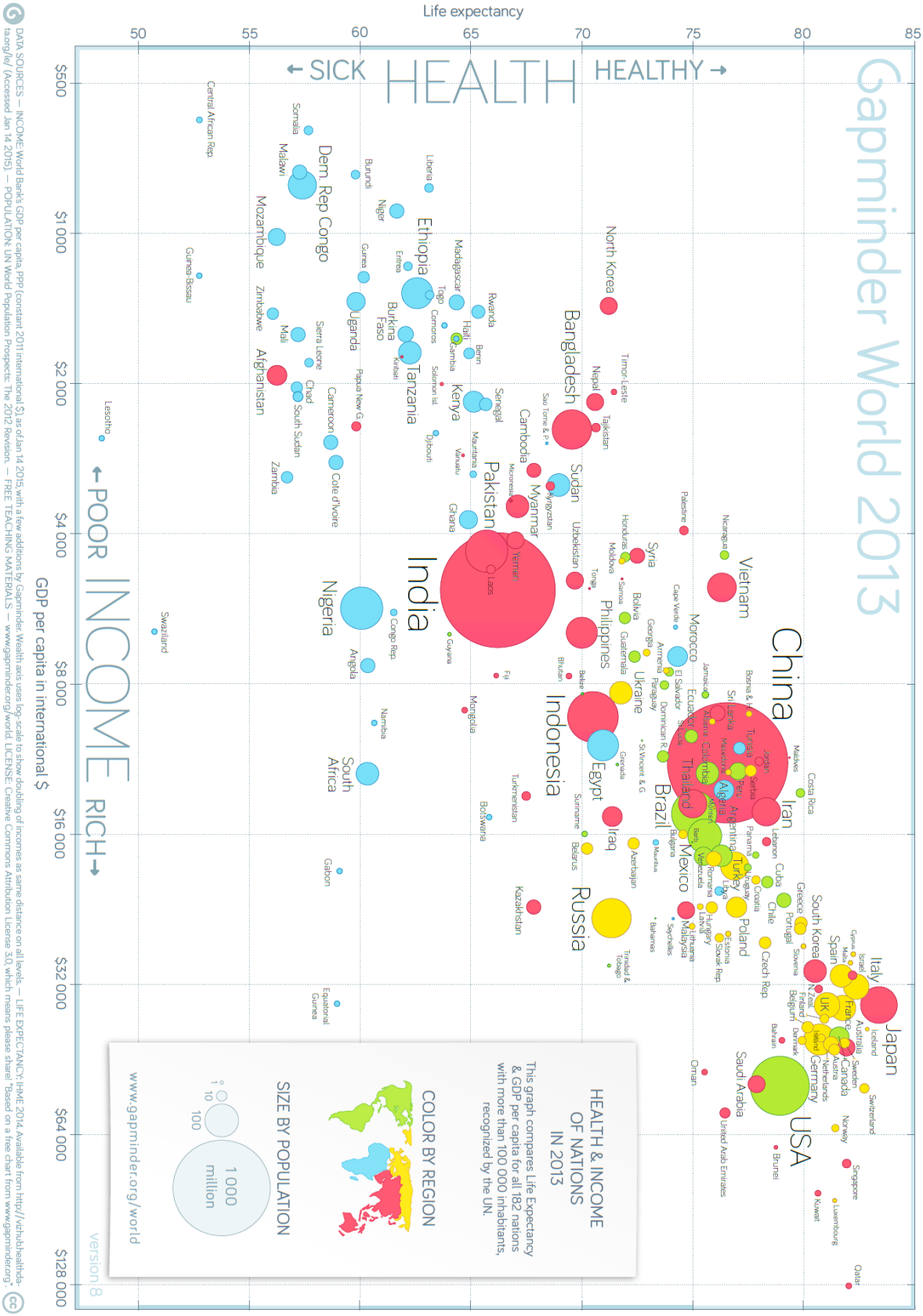


Figure 13. Life expectancy and income per capita in 2013, by nation (Gapminder Foundation) [20].

3. Dashboards

Dashboards are a helpful way to **communicate** and **report** data. They are versatile in that they support multiple types of reporting. Dashboards are predominantly used in business intelligence contexts, but they are being used more frequently to communicate data and visualize analysis for non-business services also.

Popular dashboarding platforms include Tableau, and Power BI, although there are other options, such as Excel, R + Shiny, Geckoboard, Matillion, JavaScript, etc.

These technologies aim to make creating data reports as simple and user-friendly as possible. They are intuitive and powerful; creating a dashboard with these programs is quite easy, and there are tons of how-to guides available online [10, 11, 23].

In spite of their ease of use, however, dashboards suffer from the same limitations as other forms of data communication, to wit: how can results be **conveyed effectively** and how can an **insightful data story** be relayed to the desired audience? Putting together a "good" dashboard is more complicated then simply learning to use a dashboarding application [19].

3.1 Foundation of Dashboards

Effective dashboarding requires that the designers answer questions about the planned-for display:

- who is the target audience?
- what value does the dashboard bring?
- what type of dashboard is being created?

Answering these questions can guide and inform the visualization choices that go into creating dashboards.

Selecting the **target audience** helps inform data decisions that meet the needs and abilities of the audience. When thinking of an audience, consider their **role** (what decisions do they make?), their **workflow** (will they use the dashboard on a daily basis or only once?), and **data expertise level** (what is their level of data understanding?).

When creating a dashboard, its important to understand (and keep in mind) why one is needed in the first place – does it find **value** in:

- helping managers make decisions?
- educating people?
- setting goals/expectations?
- evaluating and communicating progress?

Dashboards can be used to communicate numerous concepts, but not all of them can necessarily be displayed in the same space and at the same time so it becomes important to know where to direct the focus to meet individual dashboards goals.

Dashboard decisions should also be informed by the **scope**, the **time horizon**, the required **level of detail**, and the dashboard's **point-of-view**.

In general,

- the **scope** of the dashboard could be either broad or specific – an example of a broad score would be displaying information about an entire organization, whereas a specific scope could focus on a specific product or process;
- the **time horizon** is important for data decisions – it could be either historical, real-time, snapshot, or predictive:
  - **historical** dashboards look at past data to evaluate previous trends;
  - **real-time** dashboards refresh and monitor activity as it happens;
  - **snapshot** dashboards show data from a single time point, and
  - **predictive** dashboards use analytical results and trend-tracking to predict future performances;
- the **level of detail** in a dashboard can either be high level or drill-able – **high level** dashboards provide only the most critical numbers and data; **drill-able** dashboards provide the ability to "drill down" into the data in order to gain more context.
- the dashboard **point of view** can be prescriptive or exploratory – a **prescriptive** dashboard prescribes a solution to an identified problem by using the data as proof; an **exploratory** dashboard uses data to explore the data and find possible issues to be tackled.

The foundation of good dashboards comes down to deciding what information is most important to the audience in the context of interest; such dashboards should have a **core theme** based on either a **problem to solve** or a **data story to tell**, while removing extraneous information from the process.

3.2 Dashboard Structure

The dashboard structure is informed by four main considerations:

- **form** – format in which the dashboard is delivered;
- **layout** – physical look of the dashboard
- **design principles** – fundamental objectives to guide design
- **functionality** – capabilities of the dashboard

Dashboards can be presented on paper, in a slide deck, in an online application, over email (messaging), on a large screen, on a mobile phone screen, etc.

Selecting a **format** that suits the dashboard needs is a necessity; various formats might need to be tried before arriving at a final format decision.

The structure of the dashboard itself is important because visuals that tell similar stories (or different aspects of the same story) should be kept close together, as **physical proximity of interacting components** is expected from the viewers and consumers.

Poor structural choices can lead to important dashboard elements being undervalued. The dashboard shown in Figure 15 provides an example of **group visuals** that tell similar stories.

Knowing which visual displays to use with the "right" data helps dashboards achieve structural integrity:

- **distributions** can be displayed with **bar charts** and **scatter plots**;
- **compositions** with **pie charts**, **bar charts**, and **tree maps**;
- **comparisons** use **bubble charts** and **bullet plots**, and
- **trends** are presented with **line charts** and **area plots**.

An interesting feature of dashboard structure is that it can be used to guide **viewer attention**; critical dashboard elements can be highlighted with the help of visual cues such as use of **icons**, **colours**, and **fonts**.

Using **filters** is a good way to allow dashboard viewers of a dashboard to customize the dashboard scope (to some extent) and to investigate specific data categories more closely. The dashboard shown in Figure 16 provides an example of a dashboard that makes use of an interactive filter to analyze data from specific categories.

3.3 Dashboard Design

An understanding of design improves dashboards; **dissonant** designs typically make for poor data communication. Design principles are discussed in [1, 2, 16, 18, 21, 22].

For dashboards, the crucial principles relate to the use of **grids**, **white space**, **colour**, and **visuals**.

When laying out a dashboard, **gridding** helps direct viewer attention and makes the space easier to parse; note, in Figure 15, how the various visuals are **aligned** in a grid format to lay the data out in a clean, readable manner.

In order to help viewers avoid becoming overwhelmed by clutter or information overload, consider leaving a enough **blank space** around and within the various charts; note, in Figure 16, that while the dashboard displays a lot of information, there is a lot of blank/white space between the various visuals, which provides viewers with space to breathe, so to speak. In general, clutter shuts down the communication process (see Figure 14 for two impressive examples of data communication breakdown).

Colour provides meaning to data visualizations – bright colours, for instance, should be used as alarm indicators as they immediately draw the viewer's attention. Colour themes create cohesiveness, which improves the overall readability of a dashboard.

There are no perfect dashboards – no collection of charts will ever suit everyone who encounters it. That being said, dashboards that are **elegant** (as well as **truthful** and **functional**) will deliver a bigger bang for their buck [2, 3].

In the same vein, keep in mind that all dashboards are by necessity **incomplete**. A good dashboards may still lead to dead ends, but it should allow its users to ask: "Why? What is the root cause of the problem?"

Finally, designers and viewers alike must remember that a dashboard can **only be as good as the data it uses**; a dashboard with badly processed or unrepresentative data, or which is showing the results of poor analyses, cannot be an effective communication tool, independently of design.

3.4 Examples

Dashboards are used in varied contexts, such as:

- interactive displays that allows people to explore motor insurance claims by city, province, driver age, etc.;
- a PDF file showing key audit metrics that gets e-mailed to a Department's DG on a weekly basis;
- a wall-mounted screen that shows call centre statistics in real-time;
- a mobile app that allows hospital administrators to review wait times on an hourly- and daily-basis for the current year and the previous year; etc.

The Ugly While the previous dashboards all have some strong elements, it is a little bit harder to be generous for the two examples provided in Figures 14. Is it easy to figure out, at a glance, who their audience is meant to be? What are their strengths (do they have any)? What are their limitations? How could they be improved?

The first of these is simply "un-glanceable" and the overuse of colour makes it unpleasant to look at; the seond one features 3D visualizations (rarely a good idea), distracting borders and background, lack of filtered data, insufficient labels and context, among others.

Golden Rules and Two Examples In a (since-deleted) article, N. Smith posted his 6 Golden Rules:

- **consider the audience** (who are you trying to inform? does the DG really need to know that the servers are operating at 88% capacity?);
- **select the right type of dashboard** (operational, strategic/executive, analytical);
- **group data logically, use space wisely** (split functional areas: product, sales/marketing, finance, people, etc.);
- **make the data relevant to the audience** (scope and reach of data, different dashboards for different departments, etc.);
- **avoid cluttering the dashboard** (present the most important metrics only), and
- **refresh your data at the right frequency** (real-time, daily, weekly, monthly, etc.).

Let us see how some of these can be applied to two datasets (*Global Cities Index* [24], *2015 NHL Draft Data* [13]). Screenshots of associated (Power BI) dashboards are provided in Figures 15 and 16, respectively.

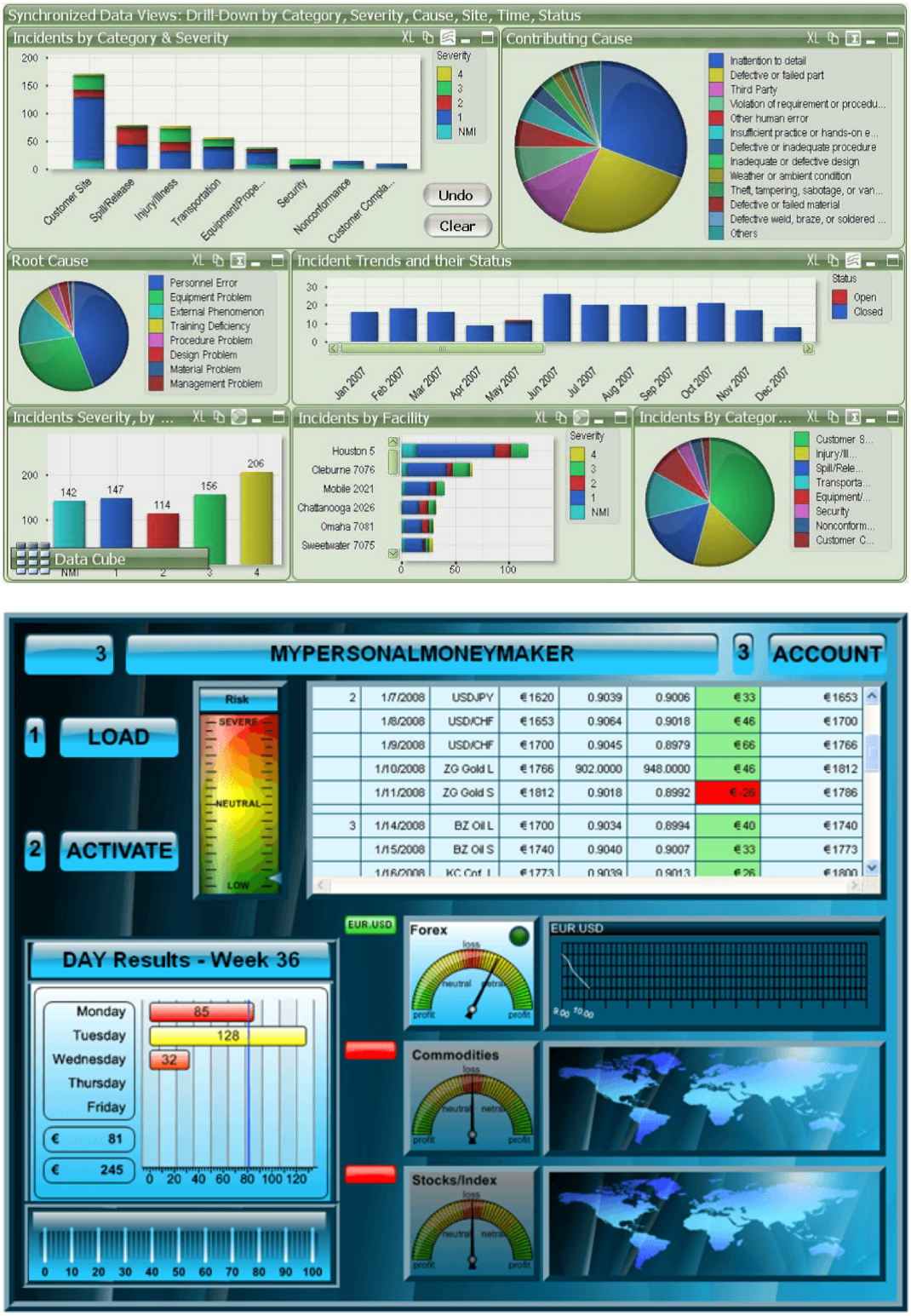


Figure 14. Anonymous "ugly" dashboards [9, 15].

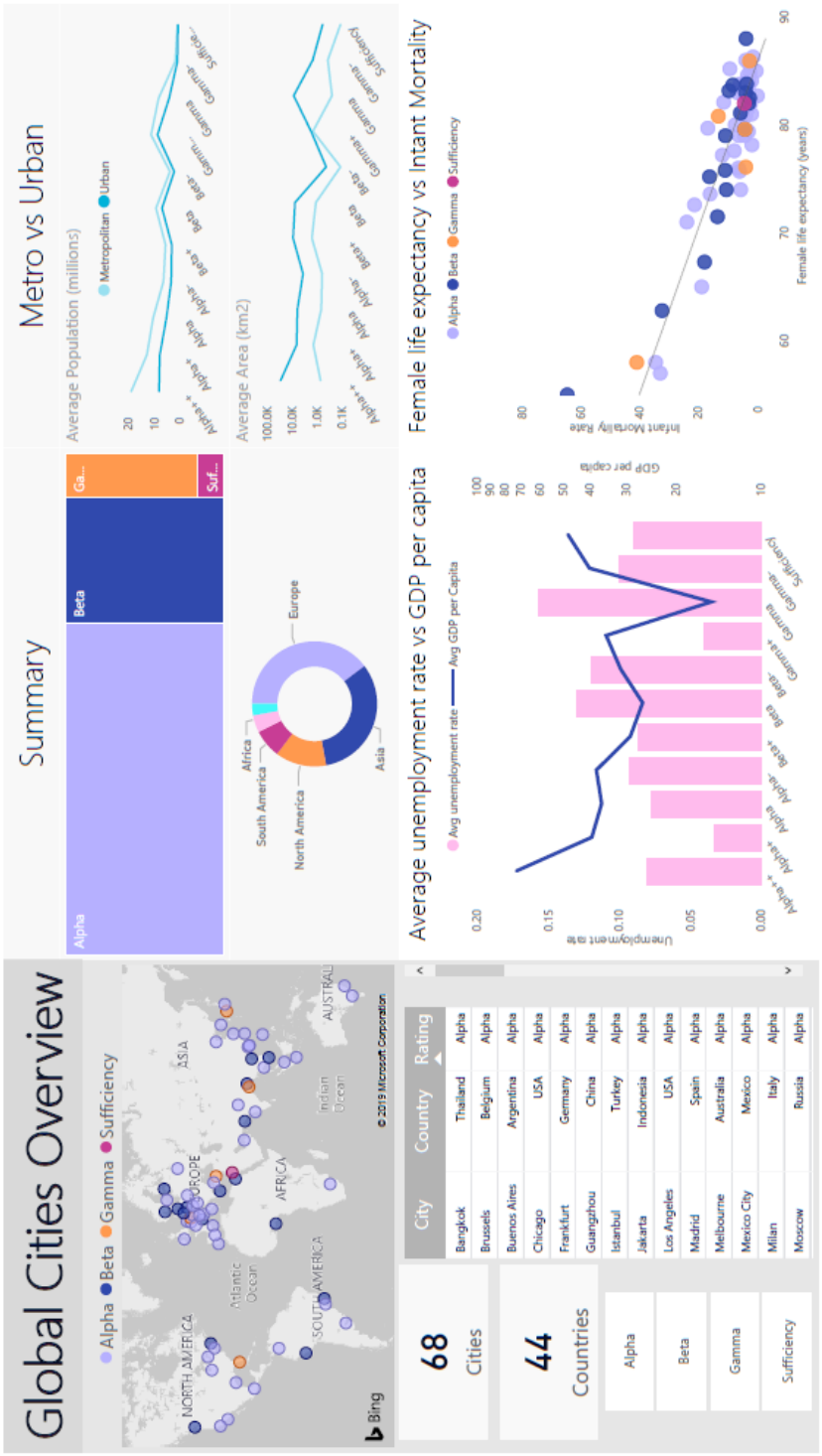


Figure 15. An exploratory dashboard showing metrics about various cities ranked on the Global Cities Index. The dashboard goal is to allow a **general audience** to **compare and contrast** the various globally ranked cities – statistics that contribute to a "higher" ranking immediately pop out. Viewers can also very easily make comparisons between high- and low-ranking cities. The background is kept neutral with a fair amount of blank space in order to keep the dashboard open and easy to read. The colours complement each other (*via* the use of a colour theme picker in Power BI) and are clearly indicative of ratings rather than comparative statistics [19].

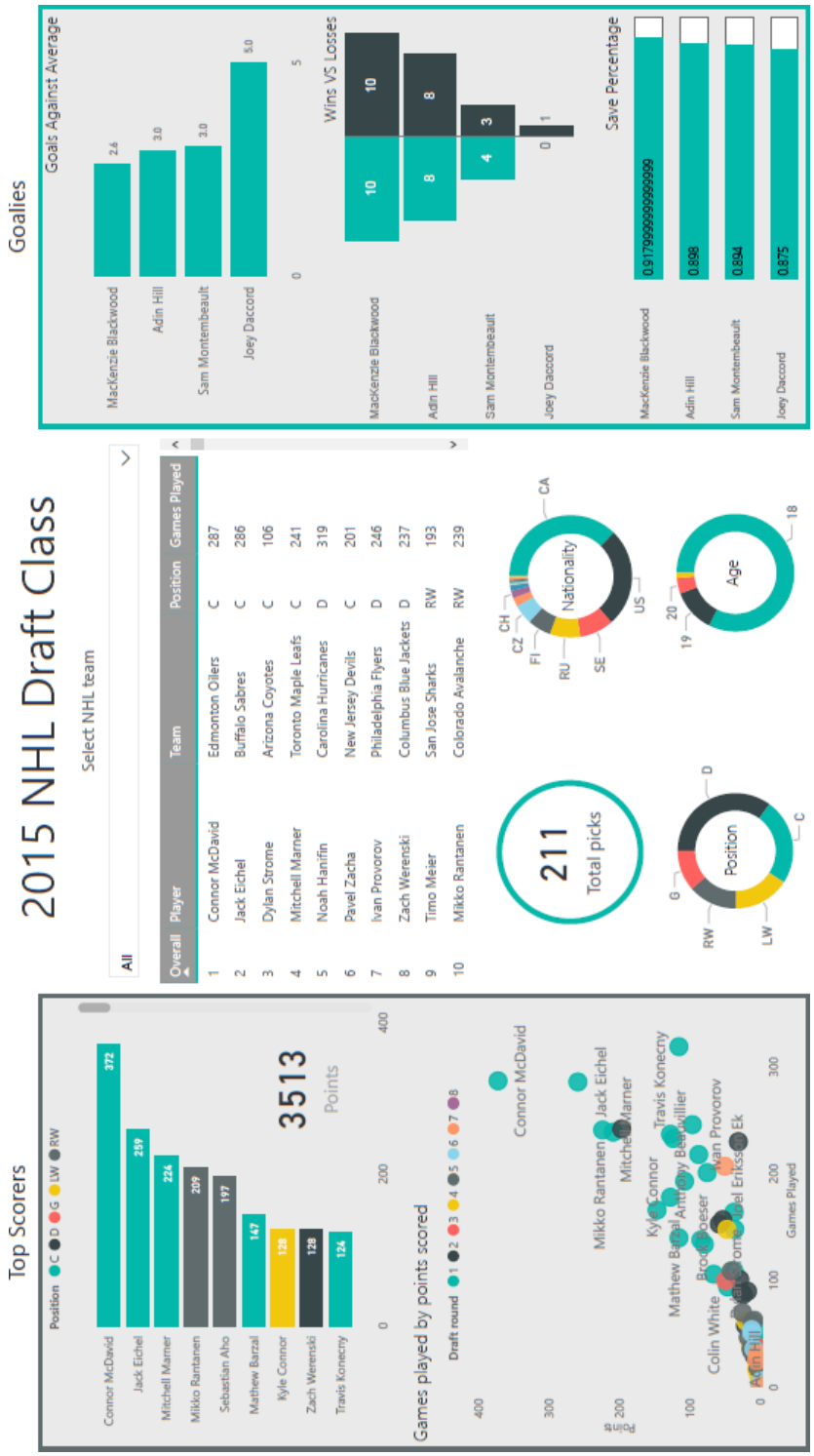


Figure 16. An exploratory dashboard showing information about the National Hockey League draft class of 2015. The dashboard displays professional statistics (as of August 2019) of hockey players drafted into the NHL in 2015, as well as their overall draft position. This dashboard allows **casual hockey fans** to **evaluate the performance** of players drafted in 2015. It provides demographic information to give context to possible market deficiencies during this draft year (i.e. defence players were drafted more frequently than any other position). This dashboard is designed to be interactive; the filter tool at the top allows dashboard viewers to drill-down on specific teams [19].

References

[1] P. Boily, S. Davies, and J. Schellinck. *Practical Data Visualization*. Data Action Lab/Quadrangle, 2021.

[2] A. Cairo. *The Functional Art*. New Riders, 2013.

[3] A. Cairo. *The Truthful Art*. New Riders, 2016.

[4] R. A. Dahl. Cause and effect in the study of politics. In D. Lerner, editor, *Cause and Effect*, pages 75–98. New York: Free Press, 1965.

[5] Data Action Lab. Data Analysis Short Course , 2020.

[6] Data Action Lab Podcast. Episode 3 - Minard's March to Moscow , 2020.

[7] P. Dragicevic and Y. Jansen. *List of Physical Visualizations and Related Artifacts* .

[8] T. Elms. *Lexical distance of European languages* . Etymologikon, 2008.

[9] Geckoboard.com. Two Terrible Dashboard Examples .

[10] Z. Gemignani and C. Gemignani. *A Guide to Creating Dashboards People Love to Use* . (ebook).

[11] Z. Gemignani and C. Gemignani. *Data Fluency: Empowering Your Organization with Effective Data Communication*. Wiley, 2014.

[12] A. Hill. The environment and disease: Association or causation? *Proc R Soc Med*, 58(5):295–300, 1965.

[13] hockey reference.com. 2015 NHL Entry Draft . 215.

[14] A. Jensen, P. Moseley, T. Oprea, S. Ellesøe, R. Eriksson, H. Schmock, P. Jensen, L. Jensen, and S. Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5, 2014.

[15] Matillion.com. Poor Use of Dashboard Software .

[16] I. Meireilles. *Design for Information*. Rockport, 2013.

[17] @DamianMingle .

[18] C. Nussbaumer Knaflic. *Storytelling with Data*. Wiley, 2015.

[19] M. Pelletier and P. Boily. Dashboard and data visualization, with examples. *Data Science Report Series*, 2019.

[20] H. Rosling. *The Health and Wealth of Nations* . Gapminder Foundation, 2012.

[21] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2001.

[22] E. Tufte. *Beautiful Evidence*. Graphics Press, 2008.

[23] S. Wexler, J. Shaffer, and A. Cotgreave. *The Big Book of Dashboards*. Wiley, 2017.

[24] Wikipedia. Globalization and World Cities Research Network .

[25] N. Yau. *FlowingData* .