1 Data Visualization Essentials

There is little doubt about it: we live in a golden age for data visualization. Rendering and dashboarding tools are everywhere – `R`, `ggplot2`, `seaborn`, plotly, Power BI, Tableau, among others – and it seems as though we cannot turn around without bumping into another amazing book on the topic. Notable examples include [2–16].

1.1 Historical Perspectives

If we take the position that data visualization is simply another way for us to take in our environment (and its stimulii) and represent things in a way that will allow us to make decisions in order to attempt to better grasp it and to hopefully control it, then we have been conducting data visualization for millions of years.

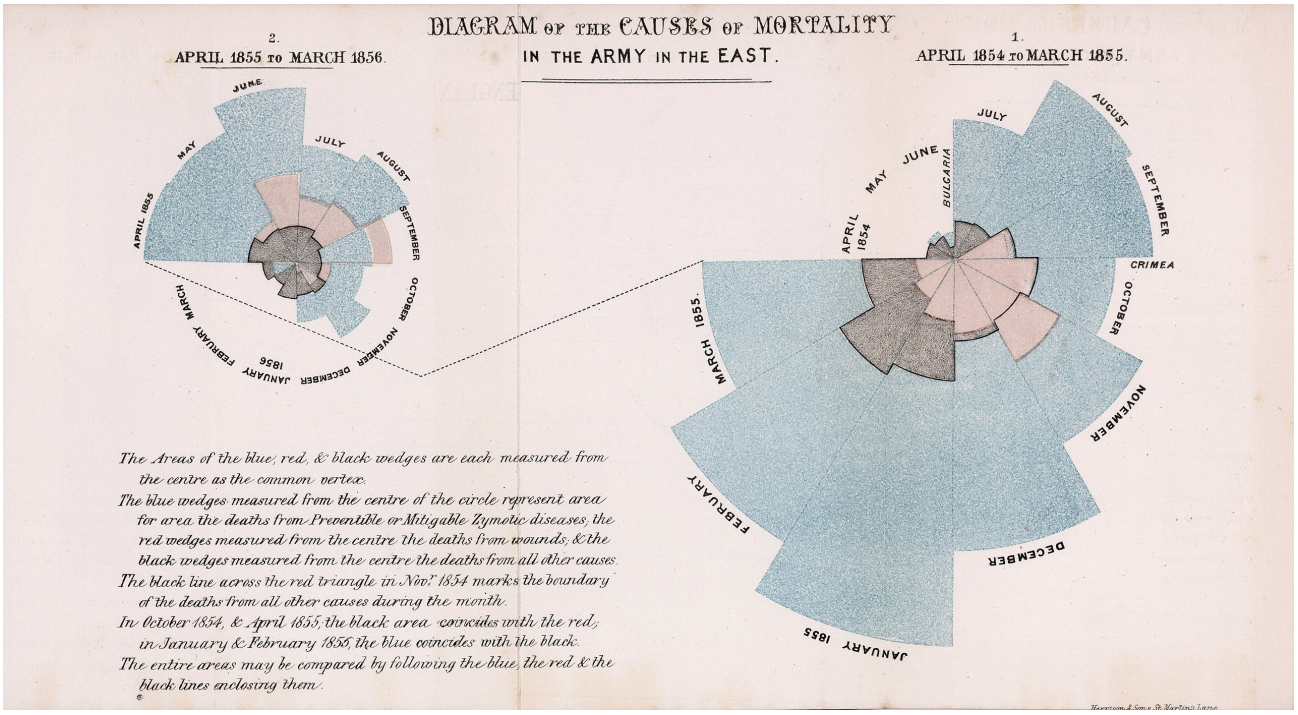
Of course, data visualization as we understand it today is not usually viewed under this all-encompassing lens: instead, we are dealing with **datasets** which have been **collected**, **transformed**, and **processed** with specific **analytical goals** in mind, and **results** conveyed to an **interested audience** using a common vocabulary which relies on **visual tropes** and **storytelling conventions**.

It is traditional, at this stage, for authors to introduce historical data visualization and to spend some time discussing their strengths, weaknesses, and "first-to-market" claims. Examples are provided in Figures 1.1 and 1.2.[1]

We have come a long way over the last 250 years or so when it comes to visualizing data insights, of course, but in a very real sense, we are still more or less following our progenitors' lead: **exploring**, **describing**, **explaining**, and **persuading**.

In one major advance for the field, however, the current consensus is that data visualization has become **analytical method** in its own right (a topic we will discuss further in Chapter 2). Other improvements include the use of **sophisticated data rendering tools** (see Chapters 11 to 13) and the concerted use of **storytelling techniques** (see Chapters 7 and 8).

1: Frequently discussed charts include William Playfair's *The Commercial and Political Atlas* [1786], John Snow's *Map of the London Cholera Outbreak of 1854*, Charles Minard's *March to Moscow* [1869] (see Section 1.4), Florence Nightingale's *Diagram of the Causes of Mortality in the Army in the East* [1858], William DuBois' *The Exhibit of American Negroes at the 1900 Paris World Exposition*, and/or Charles de Fourcroy's *Tableau Poléométrique* [1784], which have all been covered extensively in other sources.



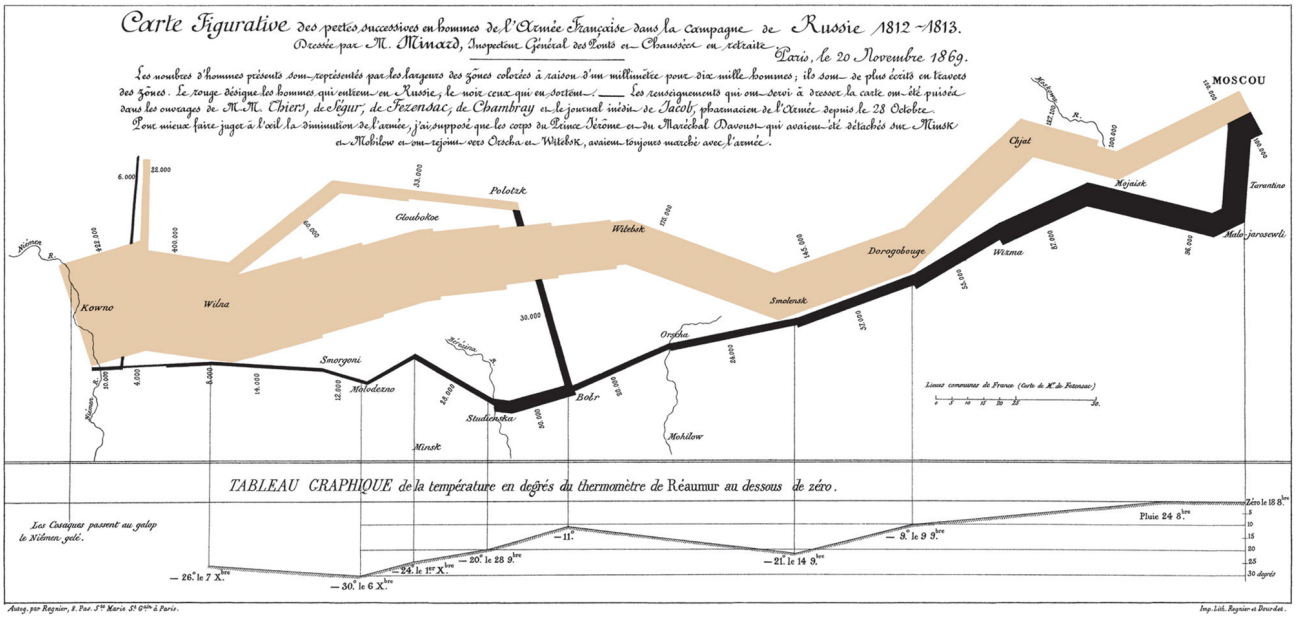


Figure 1.1: A helping of historically significant and meaningful data visualizations: Nighthingale (top row), Minard (bottom row).

Figure 1.2: A helping of historically significant and meaningful data visualizations: Playfair (top row), Du Bois (middle row), Snow (bottom row, left), and de Fourcroy (bottom row, right).

1.2 Infographics vs Visualizations

Among a subset of practitioners, it has become fashionable to differentiate between **data visualizations**, used internally in the exploration phase of data analysis, and **infographics**, used to communicate analytical results to a general audience.

In this view of the data presentation landscape, **infographics**:

- are created for story-telling purposes (subjective);
- are intended for a broad audience;
- are self-contained;
- rely on graphic design to get their message across;
- cannot usually be re-used with other data, and
- often incorporate unquantifiable information.

Data visualizations, on the other hand:

- are methods as well as items (objective);
- typically focus solely on the quantifiable;
- are used to get a sense for the data and/or to make it accessible (raw datasets can be massive and unwieldy);
- may be generated automatically, and
- do not typically rely on look-and-feel considerations (insight over aesthetic).

Prototypical examples can be found in Figure 1.3.

We favour a slightly different definition: **data visualizations** are data presentations that could in theory be prepared by analysts with a minimal amount of external design work (with simple tools) and **infographics** are presentations where the design takes centre stage and the data is an afterthought (and might be difficult to identify in the first place); between these two extremes, we find **data stories** (see Chapters 7 and 8).

Another important distinction is that data visualizations can be directly **"questioned"**. If the charts are not compatible with our data understanding, then either:

- the charts were not prepared properly, and/or
- our understanding of the situation needs to be revisited, and/or
- there is something wrong with the data that was used to build the charts.

In the data visualization example above, does the colouring of the slices match common wisdom about the wealth of each of the U.S, states? Is this more likely to be due to a data encoding error or to contextual misunderstanding?

Infographics, being designed with storytelling in mind, are not typically as easy to "question" – we understand that a fair bit of information has to be omitted from the final picture for simplicity's sake: if something seems out of place, it could just be that the crucial data and analytical elements were swept under the carpet in the interest of conveying a **compelling story**.

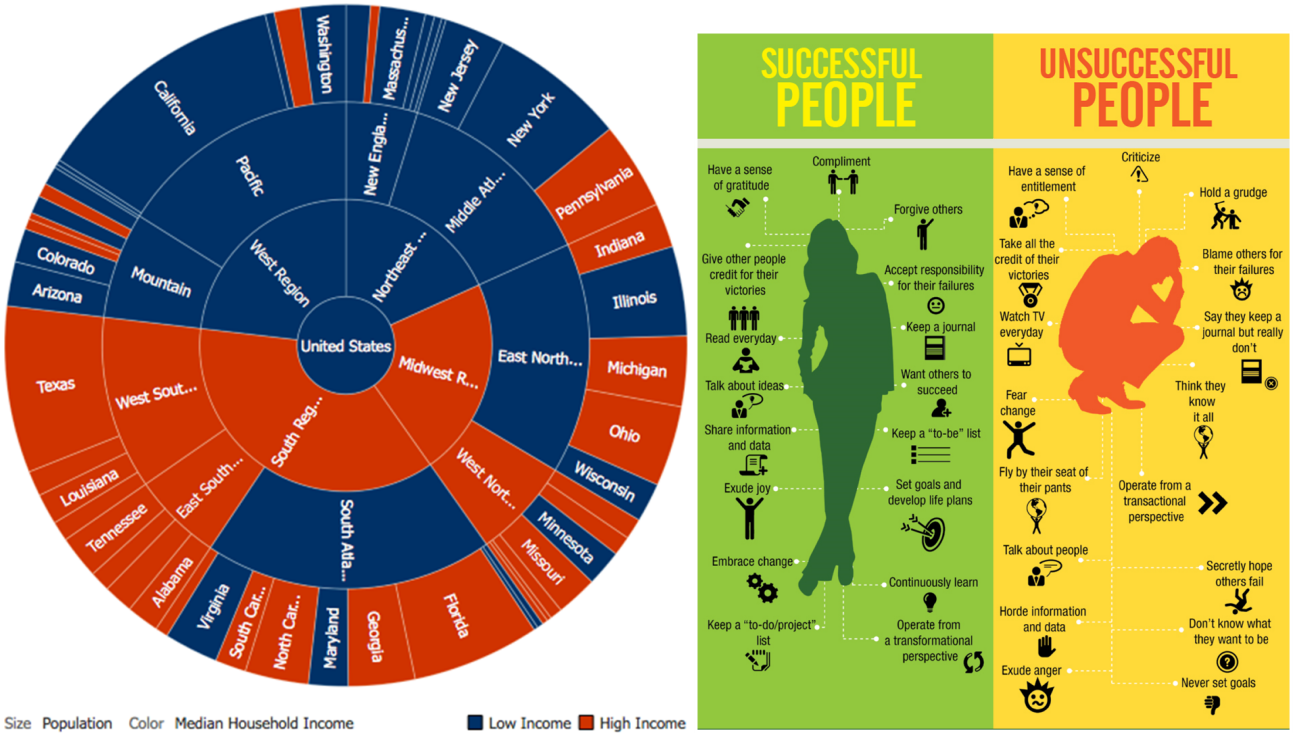


Figure 1.3: Illustration of a data visualization (left; author unknown) and of an infographic (right; Daily Infographic).

1.3 Principles of Analytical Design

In his 2006 offering *Beautiful Evidence*, E. Tufte highlights what he calls the **Fundamental Principles of Analytical Design** [3]. Tufte suggests that we present evidence to assist our thinking processes [3, p.137].

In this regard, his principles seem 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 to Minard 's celebrated *March to Moscow*.

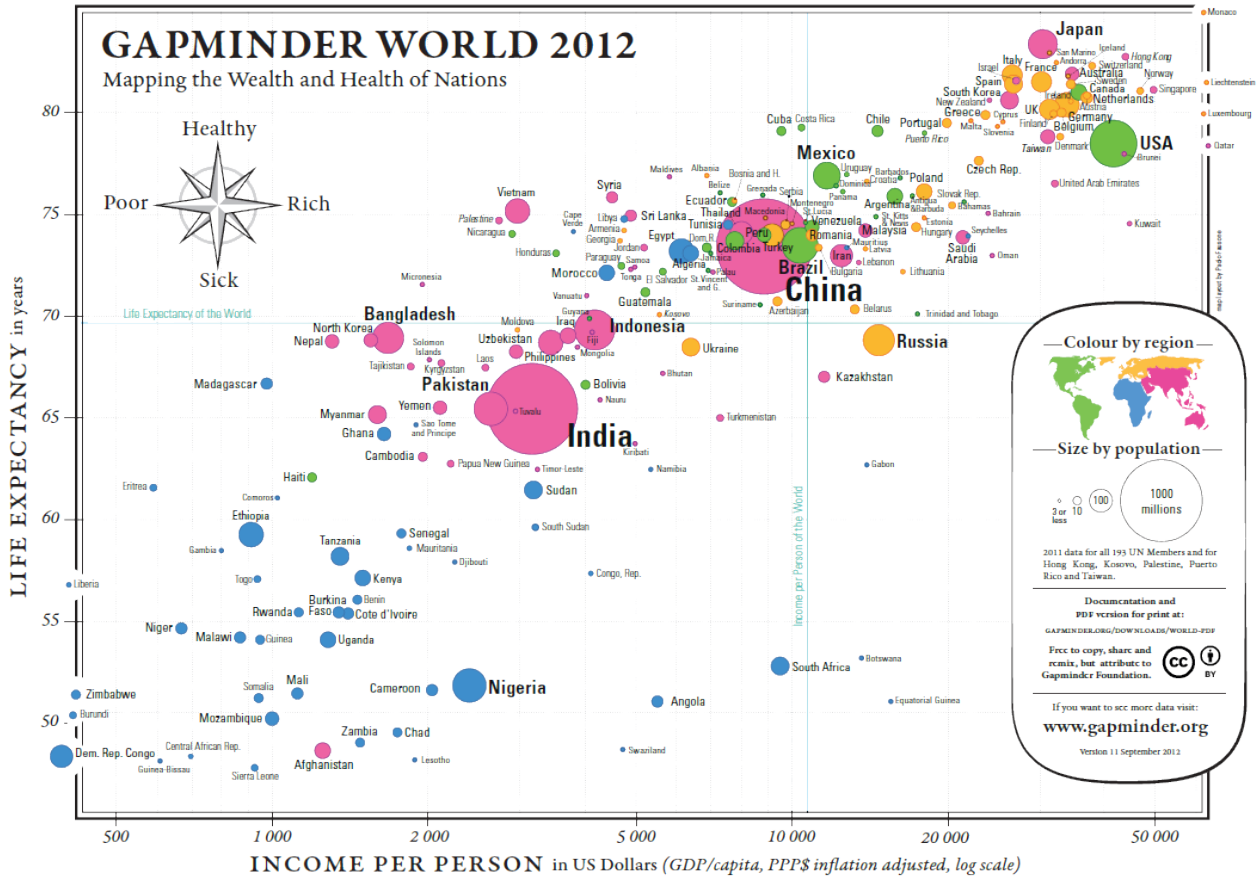


Figure 1.4: Gapminder's Health and Wealth of Nation (2012)

His lengthy analysis of the image is well worth the read [3, pp.122-139] – it will not be repeated here (although we do discuss other aspects of the chart in Section 1.4).

Rather, we will illustrate the principles with the help of the Gapminder's Foundation *2012 Health and Wealth of Nations* data visualization (see Figure 1.4), a bubble chart plotting 2012 life expectancy, adjusted income per person in USD (log-scaled), population, and continental membership for 193 UN members and 5 other countries (a high-resolution version of the image is available on the Gapminder website .)

Tufte identifies 6 basic properties of superior analytical charts:

- their ability to conduct **meaningful comparisons**;
- their ability to identify **underlying structures** and **potential causal avenues**;
- their incorporation of **multivariate links**;
- their ease of **integration** and use of **relevant data**;
- their **honest and complete documentation**, and
- their primary focus on content

Comparisons

> "Show comparisons, contrasts, differences." [3, 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. [3, p.127]

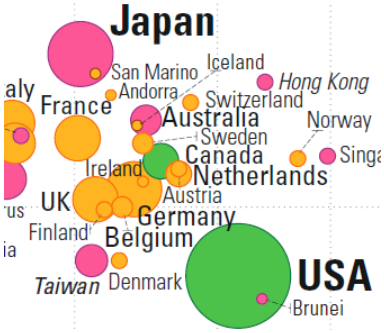
Not every comparison will be insightful, but avoiding comparisons altogether is equivalent to producing a useless display, built from a single datum.

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.

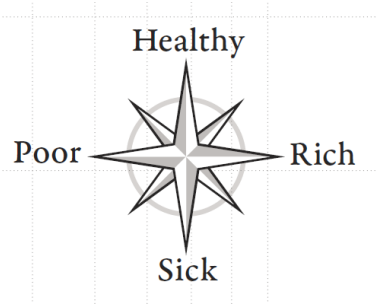
The chart's compass[2] provides a handy comparison tool:

- 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.

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 Figure 1.5).[3]



2: Top left:



3: Health and wealth are such complicated concepts that they cannot simply be represented by a single measure (and perhaps not even by multiple measures). Nevertheless, we use them as proxies in this example.

Figure 1.5: Comparisons in the Gapminder chart: country-to-country.

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 Figure 1.6).

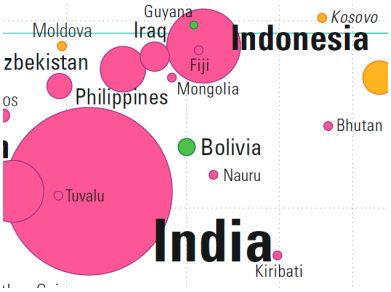


Figure 1.6: Comparisons in the Gapminder chart: country-to-country overlap.

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 $11*K*, respectively). The world's mean life expectancy and income per person are traced in light blue (see Figure 1.7).[4]

4: We see for instance, that Nepal's life expectancy is a just a hair below the life expectancy on the planet, and that Botswana's GDP per capita is above the global income per person.

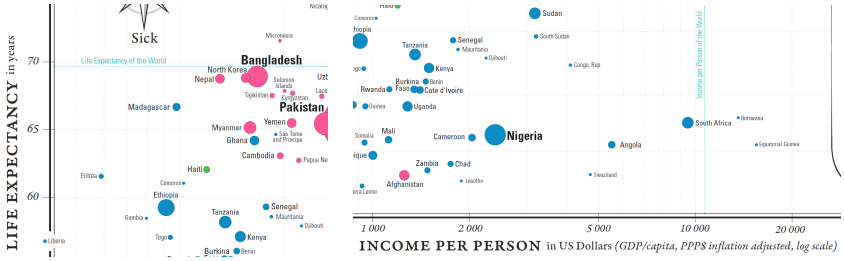


Figure 1.7: Comparisons in the Gapminder chart: country-to-world-life-expectancy (left) and country-to-world-income-per-person (right).

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.

Mechanism, Structure, Explanation

> "Show causality, mechanism, explanation, systematic structure." [3, 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*] [...] Reasoning about reforms and making decisions also demands causal logic. To produce the desired effects, we need to know and govern the causes; thus "policy-thinking is and must be causality-thinking". [3, p.128], [17]

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. [3, 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 At a glance, the relation between life expectancy and the logarithm of 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 chart in Figure 1.8.

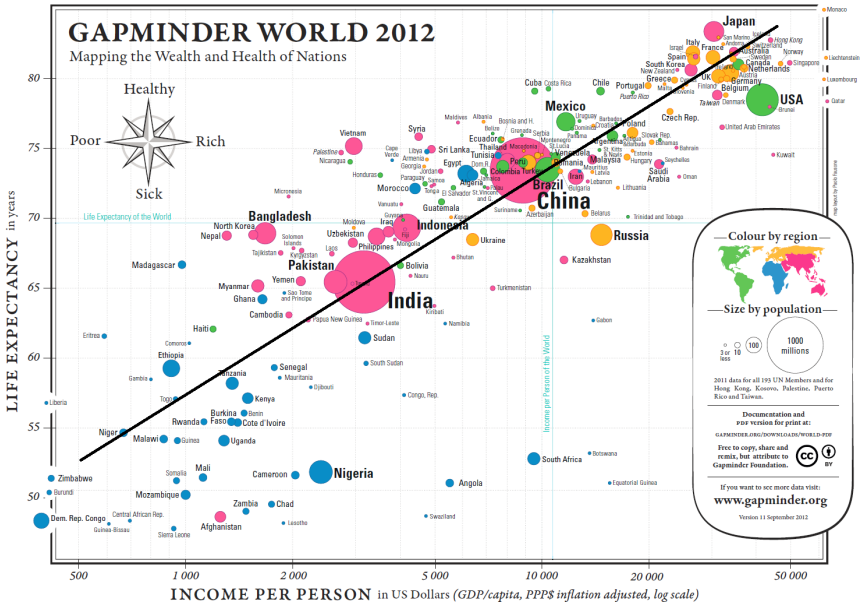


Figure 1.8: Approximate line of best fit for the Gapminder chart.

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

$$\text{Life Expectancy} \approx 6.83 \times \ln(\text{Income Per Person}) + 57.76.$$

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*.[5]

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 GDP per capita do not all host similar 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

5: 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.

and India muddle the picture somewhat). However, one quadrant is substantially under-represented in the visualization. Should it come as a surprise that there are so few "wealthier" yet "sicker" countries? (see Figure 1.9).

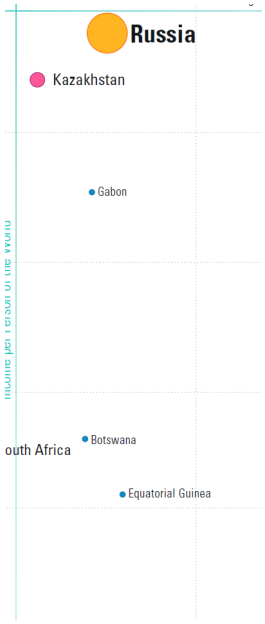


Figure 1.9: Close-up, bottom right quadrant.

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 Figure 1.10).

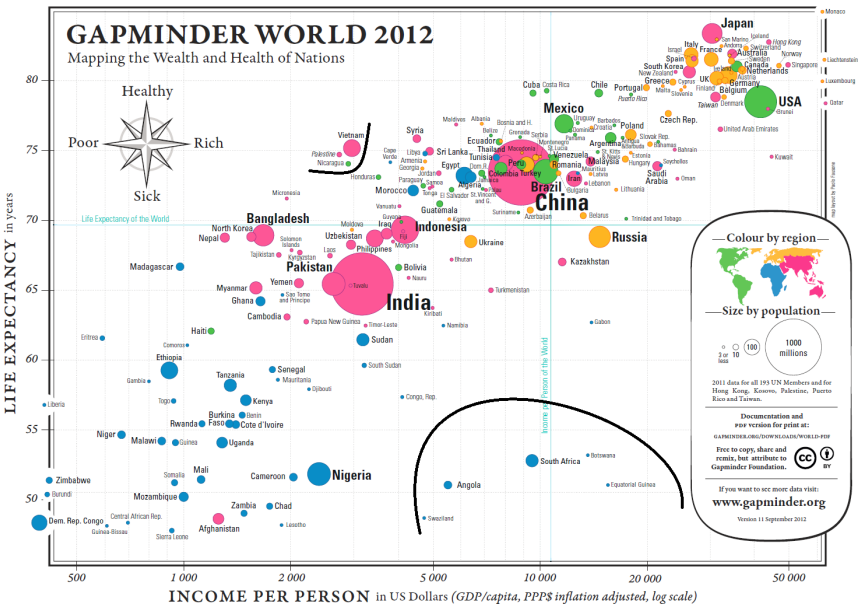


Figure 1.10: Potential outliers in the Gapminder chart.

These cry out for an explanation: South Africa, for instance, has a relatively high GDP per capita but a low life expectancy.[6] 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.[7]

Charts suggest, but "proofs" come from deeper **domain-specific analyses**.

6: Could income disparity between a poorer majority and a wealthier minority push the bubble to the right, while lower life expectancy of the majority drives the overall life expectancy downward?

7: We discuss the topic further in the exercises of Chapter 29.

Multivariate Analysis

> "Show multivariate data; that is, show more than 1 or 2 variables." [3, 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. [3, 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, [18]). [3, 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*]. [3, p.130]

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 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 Figure 1.11.

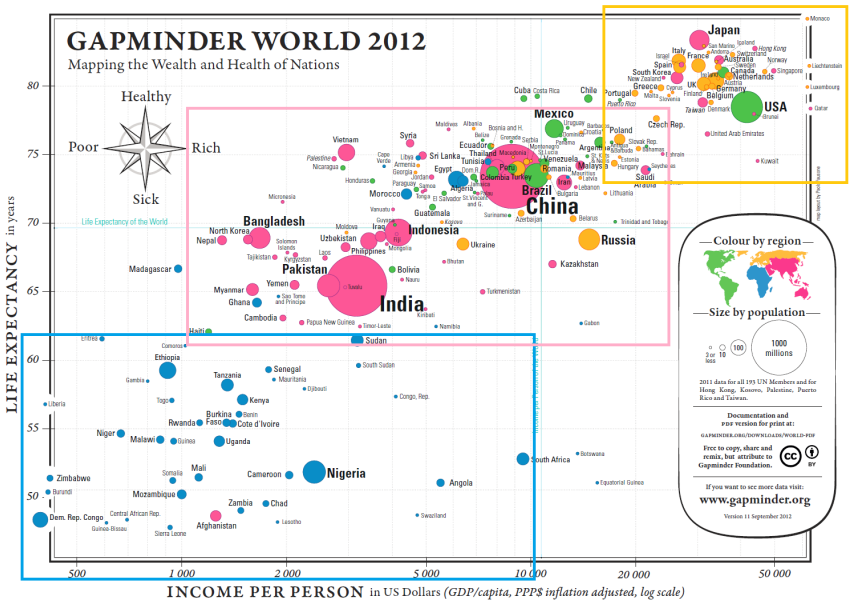


Figure 1.11: Potential outliers in the Gapminder chart.

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, but it not (by far) the **only way** to cluster the observations.

Another multivariate interpretation is afforded by Figure 1.8: countries in the upper right corner (whose location is given by 2 variables) are coloured differently (according to a third variable) than those in the lower left corner.

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.

Integration of Evidence

> "Completely integrate words, numbers, images, diagrams." [3, 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*]. [3, 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, line-art, grids, measurement scales. [3, 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. [3, p.131][8]

8: 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 their **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 the necessary evidence to shed light on "what is happening". If that goal is met, what modes of evidence were used is irrelevant.

Health and Wealth of Nations 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. ay become lost in the background, with the consequence of being taken for granted, but their presence is still crucial.

Case in point, compare the display obtained from the same data, but without integration of evidence in Figure 1.12 – who could reasonably guess what it is about without context?

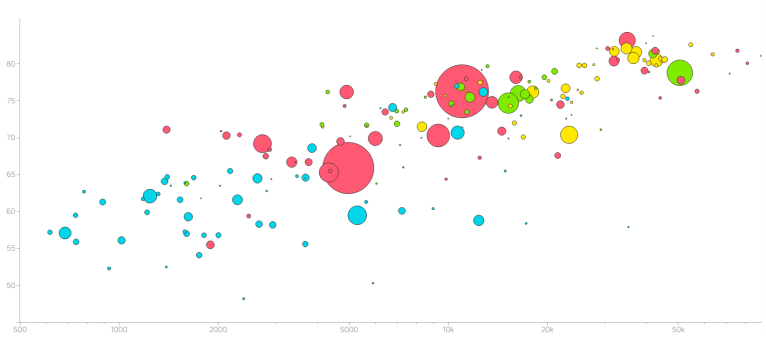


Figure 1.12: Non-integrated Gapminder chart.

Documentation

> "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." [3, 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*]. [3, 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 a 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:[9]

9: 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*]. [3, pp.132-133]

Health and Wealth of Nations 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 conducted by the Gapminder Foundation; the map layout was created by Paulo Fausone. No data regarding the sponsors is found on the chart or in the documentation. It seems plausible that there were none.
- **When was the visualization published? Which version is rendered here?** The 11th version of the 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 UN Statistical Division;
  - the Penn World Table (mark 6.2);
  - the IMF'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 UN Population Division World Population Prospects;
  - Maddison Online;
  - Mitchell's International Historical Statistics;
  - the UN Statistical Division;
  - the US Census Bureau;
  - national sources, and
  - 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 Figure 1.13).

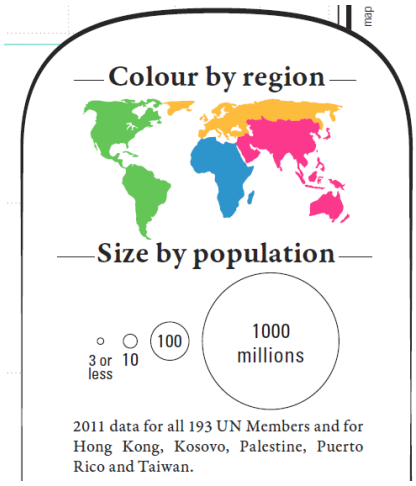


Figure 1.13: Legend inset for the Gapminder chart.

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

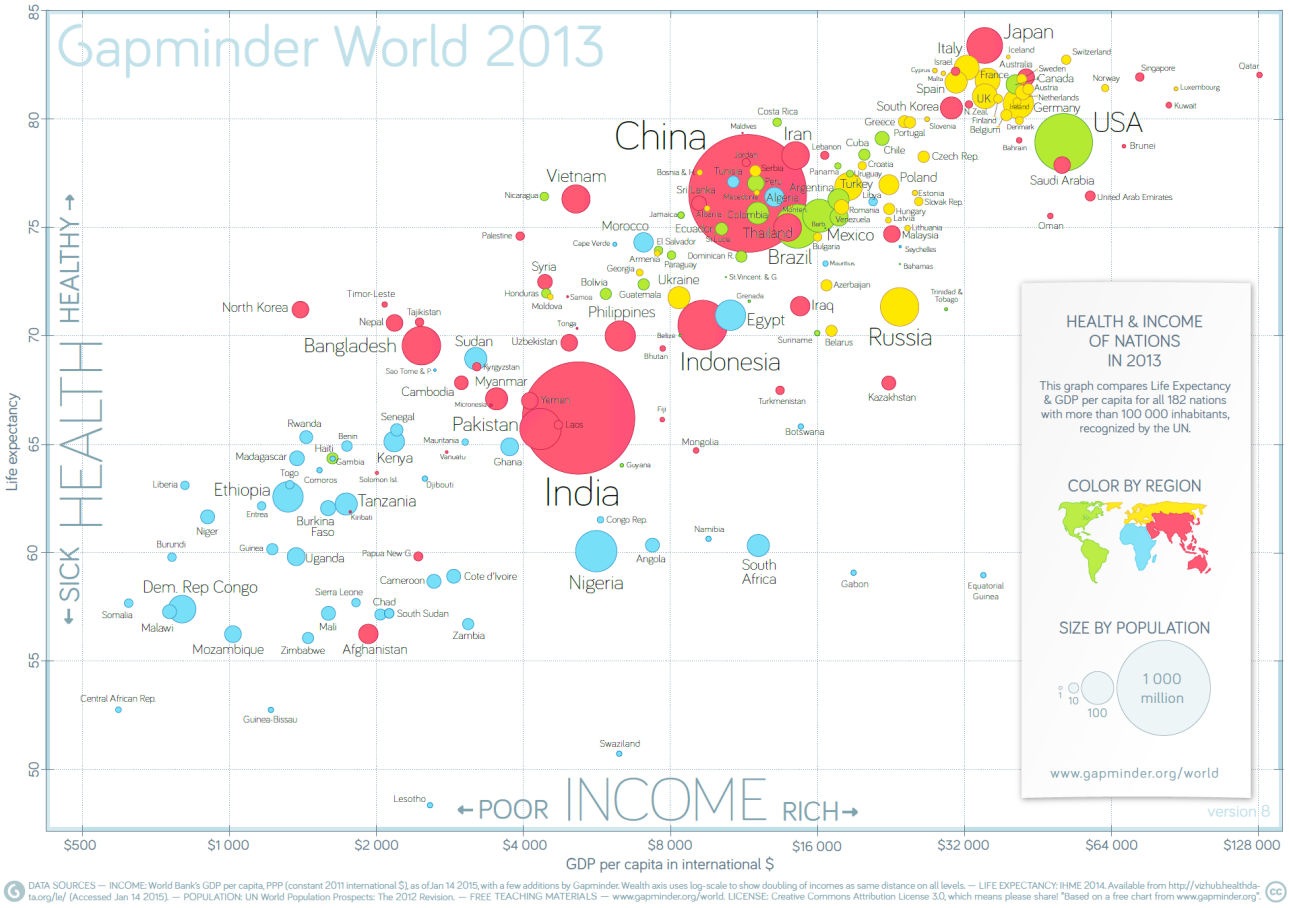
Content First and Foremost

> "Analytical presentations ultimately stand of fall depending on the quality, relevance, and integrity of their content." [3, 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?**" [*emphasis added*] [3, 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.



DATA SOURCES — INCOME: World Bank's GDP per capita, PPP (constant 2011 international $), as of Jan 14 2015, with a few additions by Gapminder. Wealth axis uses log-scale to show doubling of incomes as same distance on all levels. — LIFE EXPECTANCY: IHME 2014. Available from http://vizhub.healthdata.org/le/ (Accessed Jan 14 2015). — POPULATION: UN World Population Prospects: The 2012 Revision. — FREE TEACHING MATERIALS — www.gapminder.org/world. LICENSE: Creative Commons Attribution License 3.0, which means please share! "Based on a free chart from www.gapminder.org".

Figure 1.14: Life expectancy and income per capita in 2013, by nation (Gapminder Foundation).

Health and Wealth of Nations 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 (such as in Figure 1.14 and the Gapminder Tools .)

Are the same countries still outlying observations? Is the relationship between life expectancy and GPD per capita still linear? Are the country groupings the same? Is the bottom right quadrant still empty?

Is that story **more or less** similar? Is it **significantly** different?

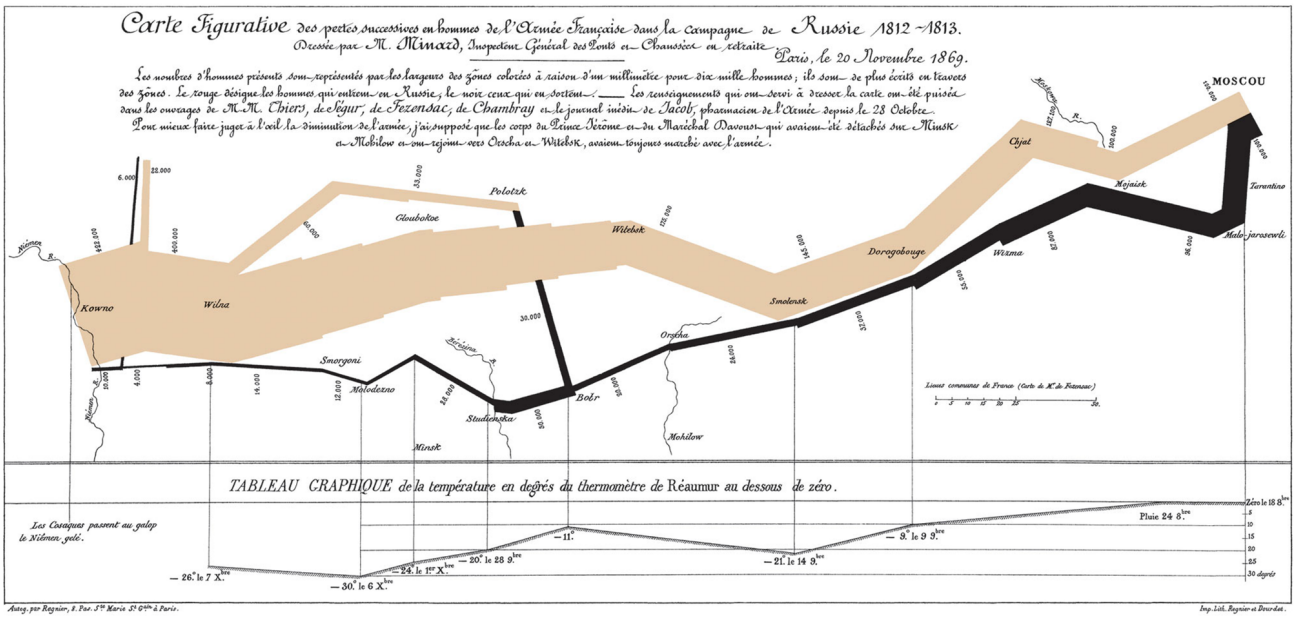


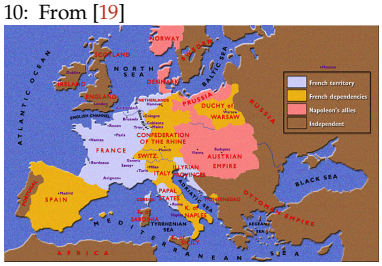
Figure 1.15: Minard's map of Napoleon's 1812 Russian campaign [13].

1.4 Minard's March to Moscow

Charles Joseph Minard was a French civil engineer who pioneered information graphics. Minard created many statistical illustrations but his most well-known is that of Napoleon's disastrous Russian campaign of 1812, the (in)famous *March to Moscow* (see Figure 1.15).

Here is some context for this chart. In 1812, Napoleon controlled a majority of Europe; only the United Kingdom and a few independents were still resisting his armies.[10] He attempted to weaken his enemy by forcing the other European nations to stop trading with the British Isles.

10: From [19]



The Russian czar Alexander refused to go along with the embargo, which earned him Napoleon's wrath. The latter gathered an army of 500,000 soldiers to invade Russia in June of 1812. The Russian troops being weaker than France's, they kept retreating before *la Grande Armée*'s advance to Moscow, but practicing a scorched earth policy of burning everything they passed doing so, which ensured that the French forces had to keep the logistical lines open to supply their campaign.

When French troops reached Moscow in October, they were diminished, famished, and ill-prepared for the Russian winter. After a failed siege of the city, they hastily retreated back to allied territory, which proved a disastrous endeavour: only 10% of the soliders made it back to their starting point. This ruinous turn of events eventually led to Napoleon's defeat and heralded the end of the First French Empire [20].[11]

11: Or so the story goes...

The chart was produced in 1864, from the relative safety provided by the Second French Empire (not so coincidentally) ruled by Napoleon's nephew, 50 years after the campaign.

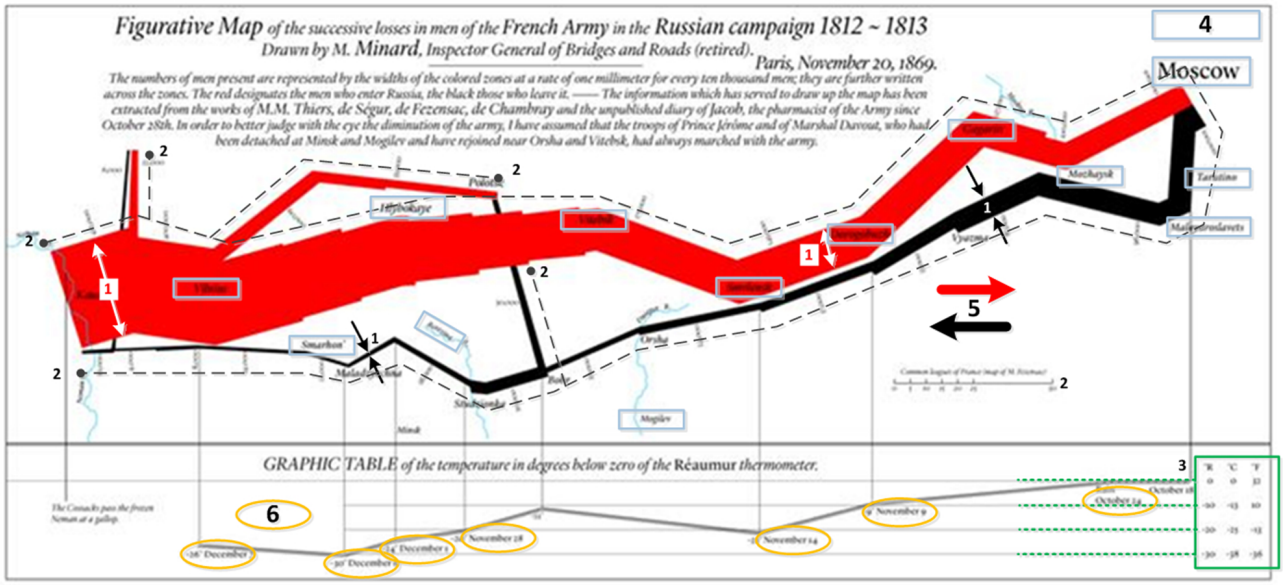


Figure 1.16: Layering of six variables on an English version of Minard's *March to Moscow* (translation originally provided in [3]).

In Minard's own words:

> The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red (or beige depending on which version you are looking at) designates the men who enter Russia, the black those who leave it. [13]

Data visualization guru E. Tufte pays the chart the ultimate compliment, commenting that it is one of the "best statistical drawings ever created." [2]

Taking a Closer Look There have been many analysis of this visualization (see [3, 13, 21], say); these are our key take-aways.

1. **Data layering:** Minard managed to represent six variables on the chart. As can be seen in Figure 1.16, he specifically overlaid the:
   a) number of troops;
   b) distance;
   c) temperature;
   d) location;
   e) direction of travel, and
   f) location relative to specific dates.
2. **Change over time without a linear time axis:** Minard is able to represent change over time (e.g., loss of troops, temperature drops, troop locations) in a non-linear but intuitive way, as in Figure 1.17.
3. **Efficient use of colour:** only using two colours has a number of advantages; in particular, it allows the reader to quickly identify a key inflection point in the story – specifically the location where *la Grande Armée* turns back to march home (see Figure 1.18).

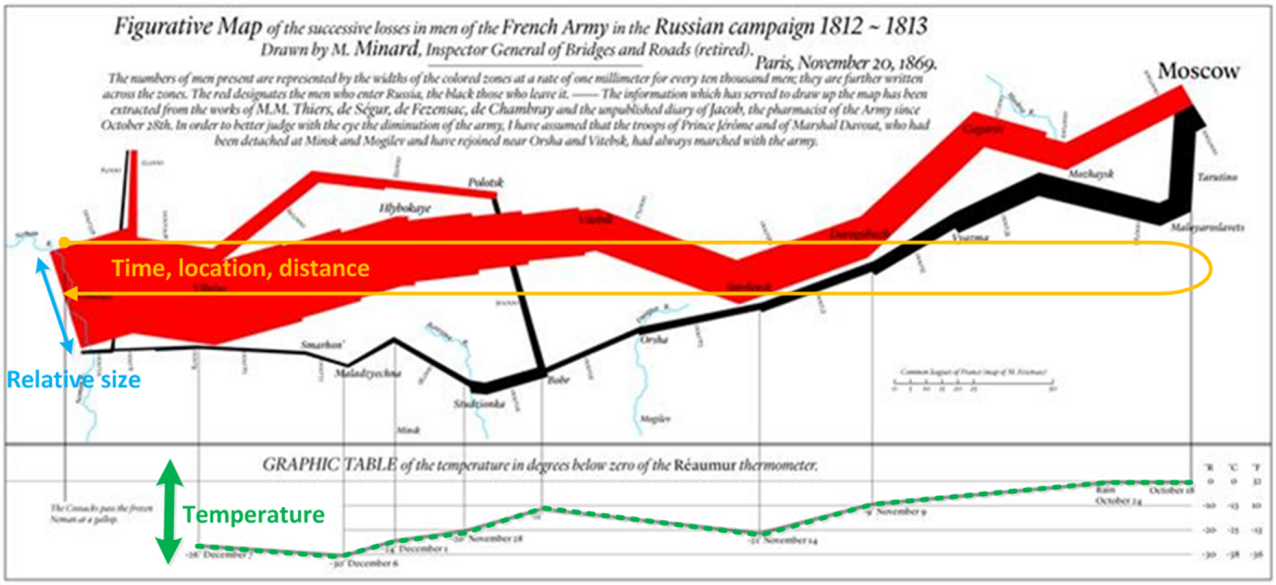


Figure 1.17: Change over time in Minard's *March to Moscow*.

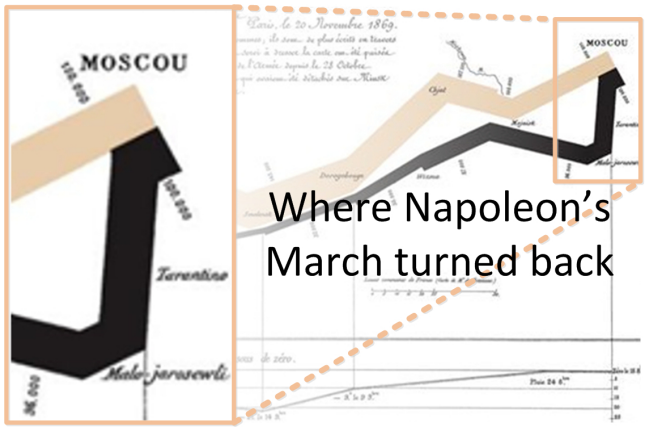


Figure 1.18: Use of colour to emphasize the inflection point in Minard's *March to Moscow*.

4. **Abstraction of location:** although the map is geographically accurate, Minard was able to clearly show movement around the geography without requiring an obvious map layer, which would have been quite distracting; this way, the reader's eye is allowed to focus on the story and on a few key pieces of geographical information (see Figure 1.19).
5. **Intuitive visual hierarchy:** Minard created an obvious visual hierarchy that allows the reader to absorb information at different levels very quickly; the eye is immediately drawn to the advance and retreat shape and colour, it naturally drops down to the temperature and finally labels and narrative come into focus (see Figure 1.20).

There are many take-away points from this visualization but we believe the following to be key:

1. **good charts tell a story** – we often create graphs, charts, diagrams, and visualizations without thinking about the story we want to tell;[12]

12: This is acceptable at the exploration stage, of course, but stories are required at the communication stage – we will revisit this topic in Chapters 7 and 8.

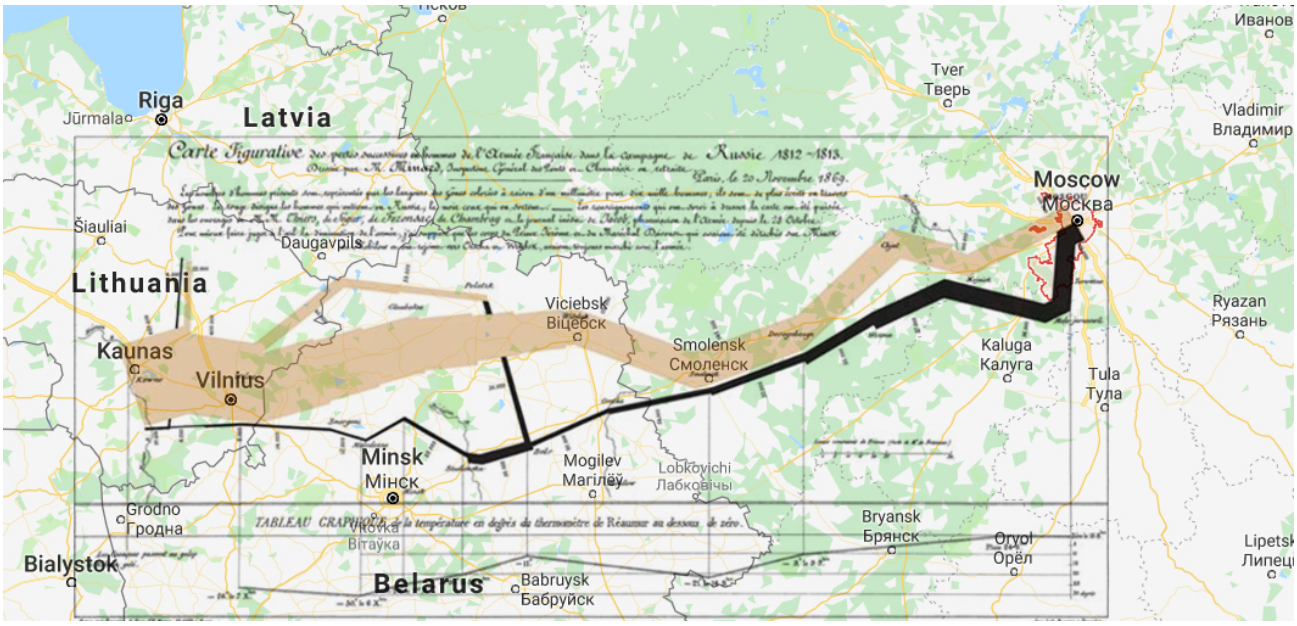


Figure 1.19: Overlay of Minard's *March to Moscow* on the geographical map of the region.

2. **the abstraction of magnitudes** — Minard manages to convey changes in army size without using numerical quantities; labels are indeed present in for temperature, dates, and locations but these are **supporting elements** in the visual hierarchy;[13]
3. **projecting multiple dimensions onto a single entity** – we usually create visualizations one dimension at a time; this is often forced on users because of the tools at our disposal.[14]

13: Readers can make sense of the story elements **without** having to mentally process numbers.

14: Get creative and figure different ways of layering information.

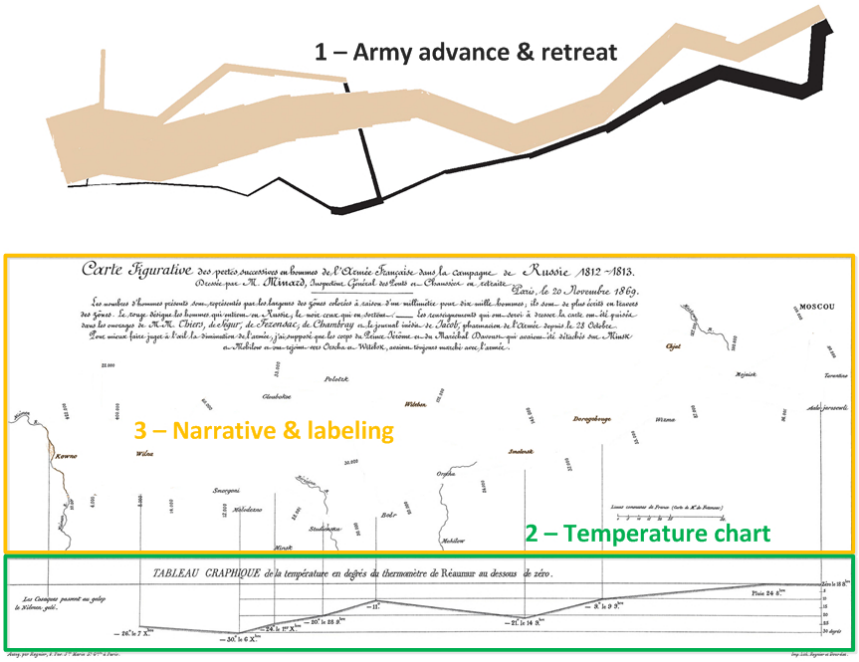


Figure 1.20: Visual hierarchy in Minard's *March to Moscow*.

1.5 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.[15]

15: 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 [11, 22, 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.

Dashboard Fundamentals

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;[16]
- 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.

16: **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 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 (see Chapters 7, *Stories and Storytelling*, and 8, *Effective Storytelling Visuals*).

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 of Figure 1.21 provides an example of **group visuals** that tell similar stories.[17]

17: The corresponding Power BI file can be found on the Data Action Lab website).

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**.

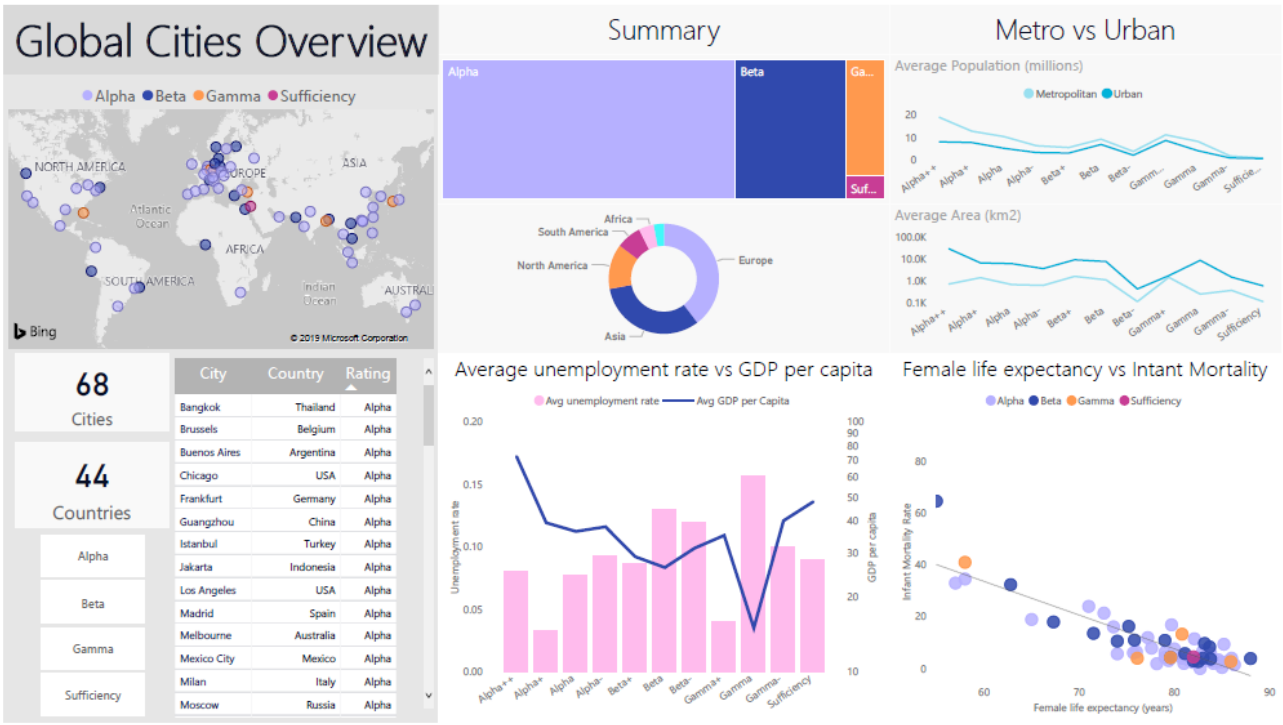


Figure 1.21: An exploratory dashboard showing metrics relating to various cities ranked on the Global Cities Index. The dashboard goal is to allow a general audience to **compare and contrast** the various globally ranked cities – statistics that contribute to a 'higher' ranking immediately pop out. Viewers can 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 (selected *via* the use of a colour theme picker in Power BI) and are clearly indicative of ratings rather than comparative statistics (produced by Maia Pelletier).

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**.[18]

The dashboard of Figure 1.22 provides an example of a dashboard that makes use of an interactive filter to analyze data from specific categories.[19]

18: Using **filters** is a good way to allow dashboard viewers of a dashboard to customize the dashboard scope and to investigate specific data categories more closely.

19: Available at Data Action Lab ↗.

Dashboard Design

An understanding of design improves dashboards; **dissonant** designs typically make for poor data communication. Design principles are discussed in [2–4, 7, 9]. 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.[20]

In order to help viewers avoid becoming overwhelmed by clutter or information overload, consider leaving enough **blank space** around and within the various charts.[21] In general, clutter shuts down the communication process – Figure 1.23 provides two impressive breakdown examples.

Colour provides meaning to data visualizations – bright colours, for instance, should be used as alarms as they draw the viewer's attention. Colour themes create cohesiveness, improving the overall readability of a dashboard.[22]

20: The various visuals are **aligned** in a grid format to lay the data out in a clean, readable manner in Figure 1.21.

21: While the dashboard of Figure 1.22 displays a lot of information, there is a lot of blank/white space between the various visuals, which provides viewers with space to breathe.

22: There are no perfect dashboards – no collection of charts will ever suit everyone who encounters it.

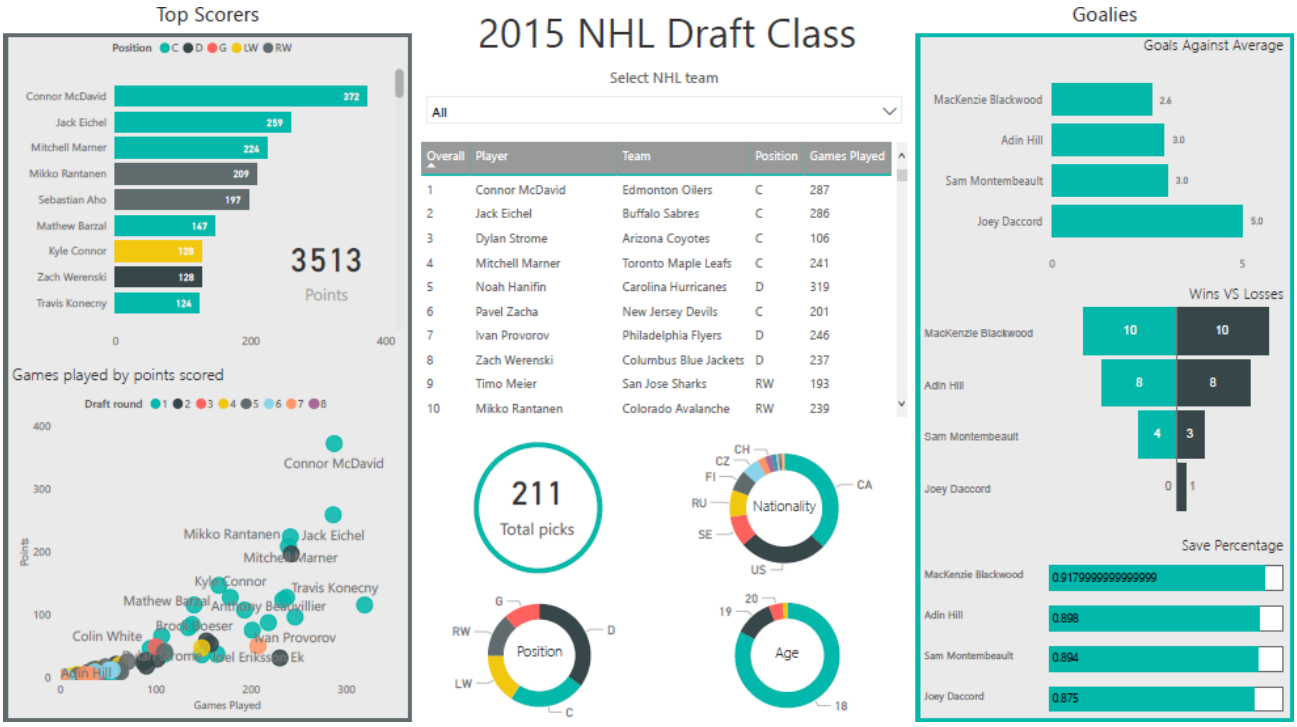


Figure 1.22: An exploratory dashboard showing information about the National Hockey League draft class of 2015. The dashboard displays professional statistics (as of August 2019) of hockey players drafted into the NHL in 2015, as well as their overall draft position. This dashboard allows **casual hockey 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 (produced by Maia Pelletier).

That being said, dashboards that are **elegant** (as well as **truthful** and **functional**) will deliver a bigger bang for their buck [7, 8]. In the same vein, keep in mind that all dashboards are by necessity **incomplete**.[23]

23: A good dashboard 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.

Examples and Final Comments

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.

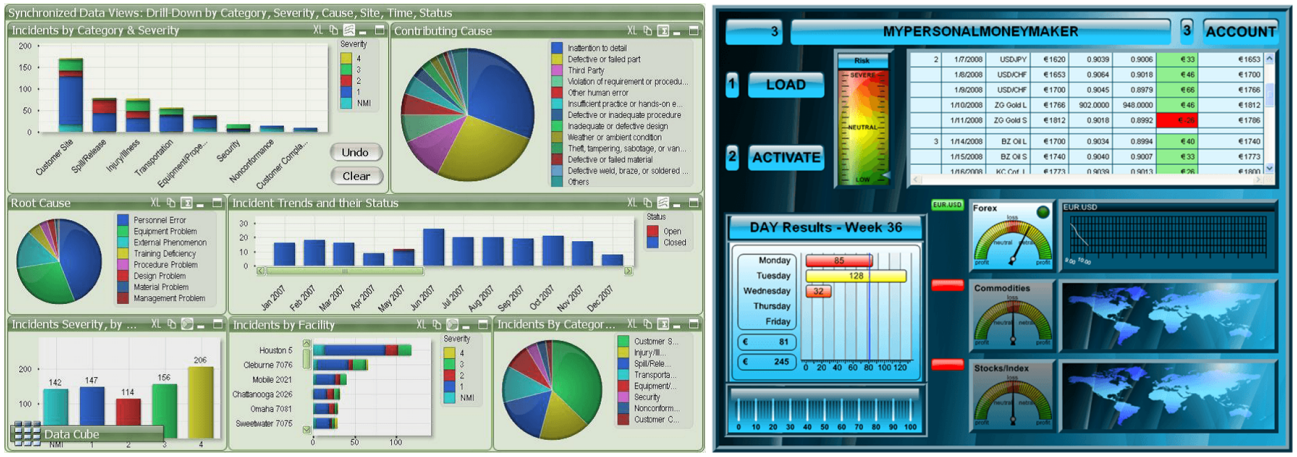


Figure 1.23: Anonymous 'ugly' dashboards [24, 25]; how about these data communication breakdowns?

24: 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? What can they be used for?

The Ugly While the previous dashboards all have some strong elements, it is a harder to be generous for the two examples provided in Figure 1.23.[24] The first of these is simply "un-glanceable" and the overuse of colour makes it unpleasant to look at; the second one features 3D visualizations (rarely a good idea), distracting borders and background, lack of filtered data, insufficient labels and context, among others.

The Good Good dashboards, on the other hand, simply breathe. The number of charts on each page is small, boxes are eschewed, simple colour schemes are preferred, and the canvas is quiet (see Figures 1.24 and 1.25, for instance). We will further discuss topics relating to data storytelling and design principles in Chapters 5, 7, and 8.

Golden Rules In a (deleted) blog article, N. Smith posted his 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.).

With dashboards, as with data analysis and data visualization in general, there is no substitute for **practice**: the best way to become a proficient builder of dashboards is to . . . well, to go out and build dashboards, try things out, and, frequently, to stumble and learn from the mistakes.

We will revisit dashboards in Chapter 13.

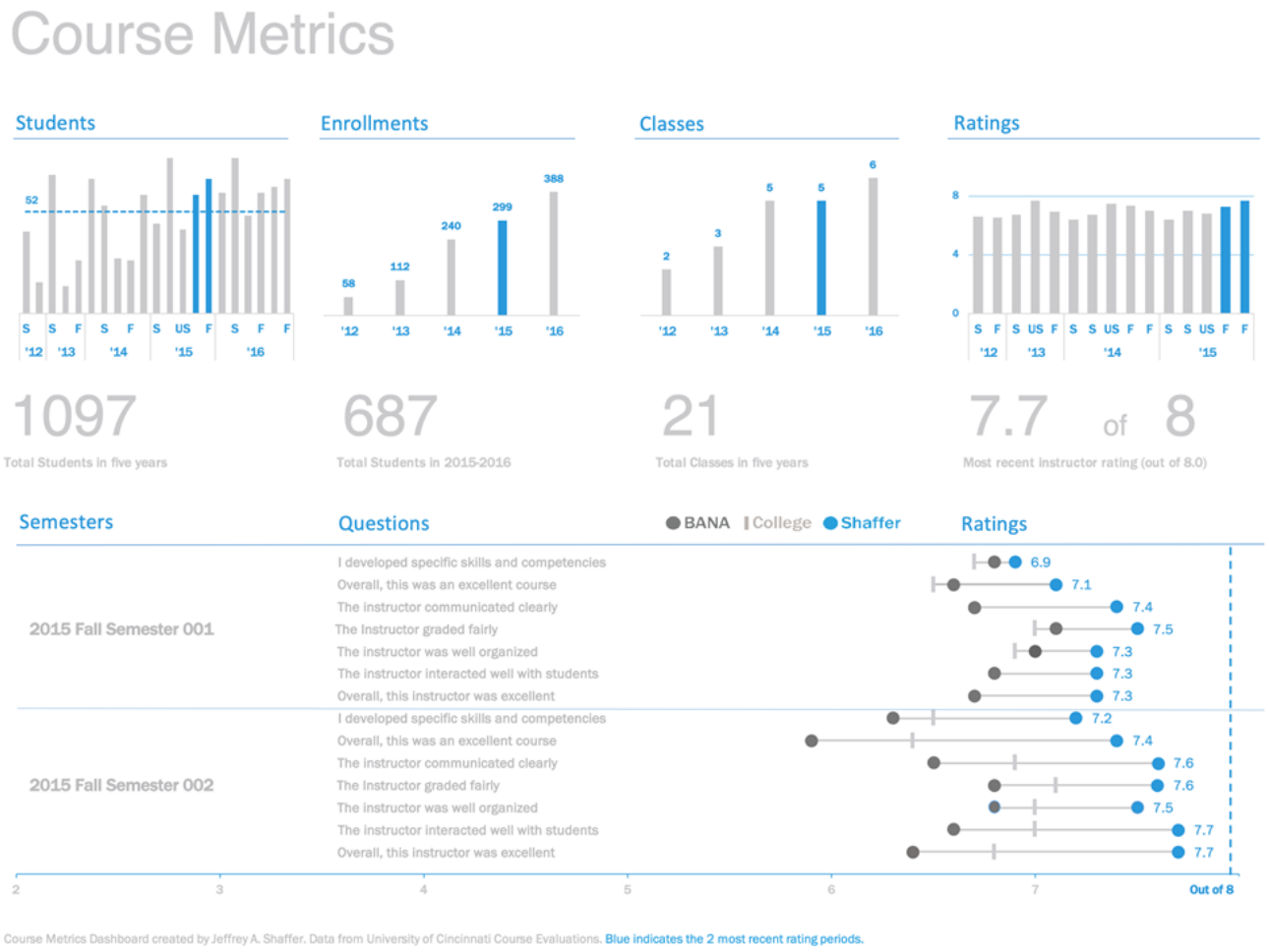


Figure 1.24: 'Zen' dashboard: course evaluations at the University of Cincinnati [11].

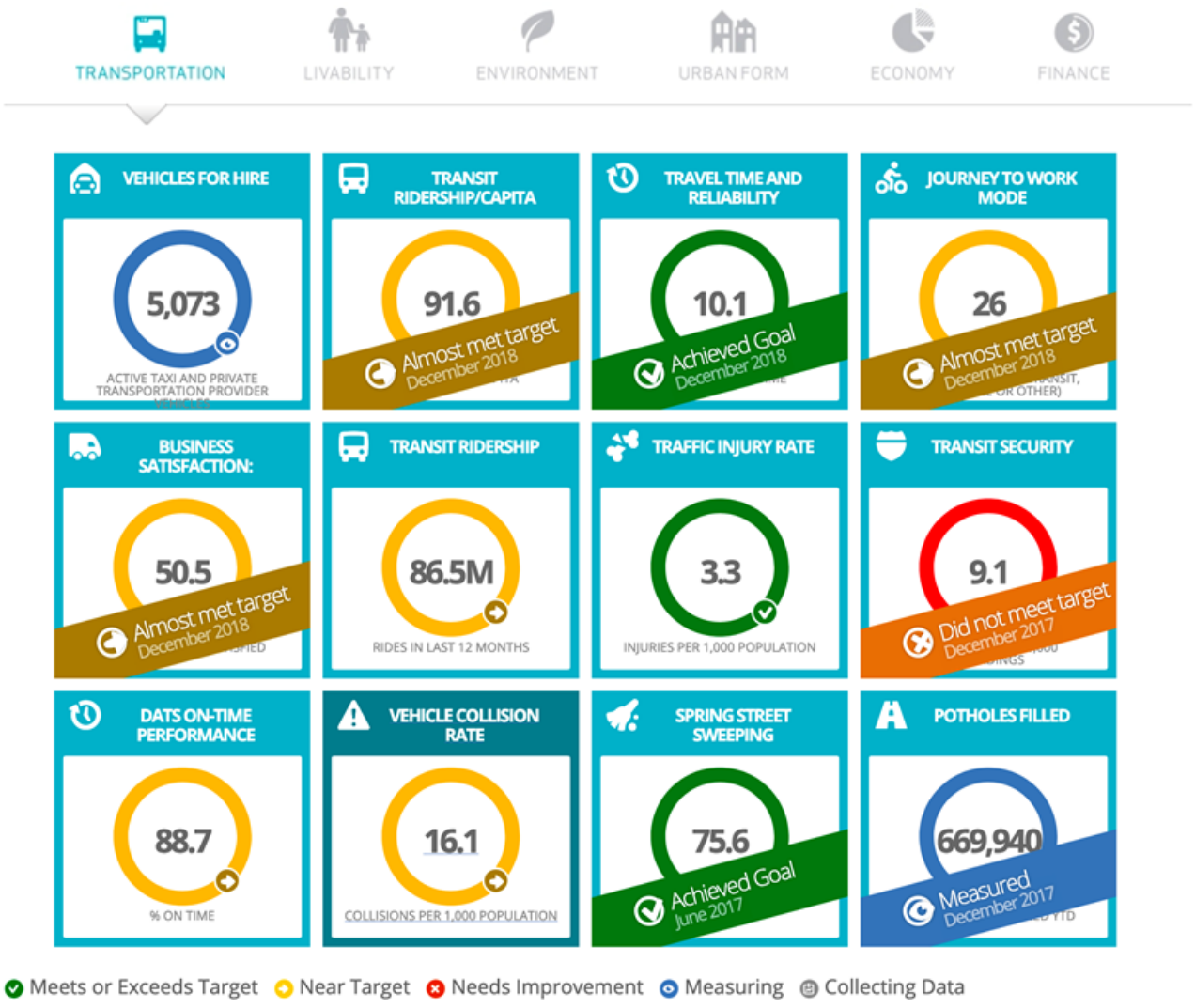


Figure 1.25: 'Zen' dashboard: former dashboard of the City of Edmonton.