How should we present our data to communicate better?

Datavisualisation: overview and simplicity

Christine Lagarenne[*], *Frédéric Minodier*[**] *and Odile Samson*[***]

The image is a powerful vector of transmission: messages are passing quite instantly from the eye to the brain where they become information. That's why image has been used by statisticians, starting from 19th century when maps and charts were made popular (Rendgen, 2020). Storytelling was added to spell out the intended message, before web technologies completed the scope by enabling the general public to easily access a very large amount of data and by helping to comprehend them through dynamic media. Graphic semiology and design are complementary: semiology has improved the effectiveness of illustrations in terms of comprehension; design helps to reduce the reading effort. When it comes to producing a datavisualisation or dataviz, the statistician's basic rules must be kept in mind (metadata, rigour and presentation of figures). When implementing, simple and generic solutions should drive choices to enhance user experience and enable statisticians to keep the control for technology, especially in the highly evolving environment. Using dataviz, a mean of communication, official statistics continue to enlighten public debate and to keep it well informed with large amounts of highly accessible data.

* Head of the Editorial Content Department, *Direction de la diffusion et de l'action régionale* (Dissemination and Regional Action Directorate – DDAR), INSEE.
christine.lagarenne@insee.fr

** At the time of writing, Head of the Publishing and Online Dissemination Division, DDAR, INSEE.
frederic.minodier@insee.fr

*** Head of the Graphics Department, DDAR, INSEE.
odile.samson@insee.fr

> ***Datavisualisation or "dataviz" brings together all visual representations of data.***

Datavisualisation or "dataviz" brings together all visual representations of data, from long-established bar charts to infographics and dynamic dashboards. Whatever the medium, it is a vector of information or communication, either in paper or, increasingly, digital format. It makes it possible to present an overview of a set of data, as well as to support the user in understanding the data, thereby allowing them to take the results of the study on board more easily.

In the context of massive data development, datavisualisation is an opportunity for official statistics to make use of such data to an even greater extent. It makes it possible not only to disseminate data in a more sophisticated manner, but also to disseminate the results of statistical and economic studies in a more accessible way. It plays a part in the strategy against fake news by making it possible to reach a wide audience due to the ease with which it can be understood.

The development of the various forms of datavisualisation is shown using some examples of the literature on the subject, as well as the practices of INSEE and the Ministerial Statistical Offices. The prerequisites for developing datavisualisation and staying up to date, while also meeting the dissemination criteria for official statistics, are then set out in detail.

▶ Images, a useful medium...

> ***"A picture is worth a thousand words," said the philosopher Confucius.***

"A picture is worth a thousand words," said the philosopher Confucius. Data are taken on board better when they are displayed in a visual rather than textual format. Not only does the brain process images at very high speed[1] but the recognition rate for an image after three days is 90% (Haber, 2013). It should be said that on average, only about ten characters are read per instance of visual fixation (Brysbaert, 2019). Readers don't really read initially. They forage, they scan, they read in a selective and associative manner.

The brain seeks to give form and a meaningful structure to what it perceives, so as to simplify and organise it in front of the complexity of our environment. Gestalt theory (the name of which is taken from the German term for "form" or "shape") inspired by Christian von Ehrenfels (Dortier, 2012) in the early 20th century deals with the psychology of patterns. There are six laws or principles of Gestalt theory, three of which will be of use for datavisualisation. First of all, the law of good figure, the main law from which the others derive: in everything it perceives, our brain seeks to recognise simple and stable forms that are familiar to it and the direct use of such forms in representations improves our speed of reading and comprehension. Next, the principle of proximity groups together the points closest to one another and is used in histograms, for example. Finally, the principle of similarity assigns

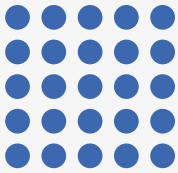
1 *https://news.mit.edu/2014/in-the-blink-of-an-eye-0116.*

> ***In everything it perceives, our brain seeks to recognise simple and stable forms that are familiar to it.***

a common character (colour, size etc.) to elements, allowing them to be grouped together visually even if they are spread apart (***figure 1***).

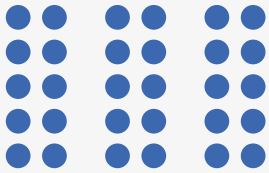
In order for the greatest number of people to take data on board, especially statistics, the graphical form is an essential and widely used resource. It makes it possible to not only provide an overview, but also and above all, to distinguish the most important elements. The simplicity of the representation (a curve or bar chart) is paramount. More technical representations have also proved their worth in accessing information, allowing quick understanding and acting in support of written material[2]. They are now part of our everyday lives (***Box 1***), such as maps.

▶ Figure 1 - Three Principles Useful for datavisualisation



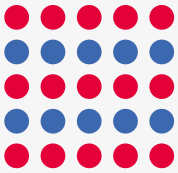
Good Figure (closure)

Our brain seeks to recognise simple and stable forms that are familiar to it.



Proximity

Nearby elements are spontaneously linked to create a group.



Similarity

Similar shapes (shape, colour, orientation etc.) are spontaneously grouped together.

0	1	2	3	2	6	9	4	5	5	7	6	1
2	0	3	0	6	4	1	0	6	3	0	4	2
1	5	6	3	0	2	5	6	9	7	4	1	0
3	6	5	1	8	0	0	6	7	0	9	1	5
0	1	2	4	8	8	1	4	1	3	6	9	4
1	5	0	2	1	0	8	7	0	6	9	5	4
1	3	5	9	7	3	2	6	9	4	5	5	7
6	1	2	0	3	7	6	4	1	0	3	9	0

How many times does the number 3 appear?

0	1	2	**3**	2	6	9	4	5	5	7	6	1
2	0	**3**	0	6	4	1	0	6	**3**	0	4	2
1	5	6	**3**	0	2	5	6	9	7	4	1	0
3	6	5	1	8	0	0	6	7	0	9	1	5
0	1	2	4	8	8	1	4	1	**3**	6	9	4
1	5	0	2	1	0	8	7	0	6	9	5	4
1	**3**	5	9	7	**3**	2	6	9	4	5	5	7
6	1	2	0	**3**	7	6	4	1	0	**3**	9	0

Thanks to the principle of similarity, by simply colouring a piece of data it becomes immediately legible.

2 See, in particular, the report entitled "*La culture statistique des Français : constats, enjeux et perspectives*" (The statistical culture of the French: findings, challenges and prospects), INSEE No 2023_14/DG75-B001, IGÉSR No 21-22 316A – February 2023, IGAC No 2023-05.
https://intranet.insee.fr/jcms/18876156_DBFileDocument/mi-2022-6-litteratie-statistique-rapport-ig-21-04-2023?details=true.

▶ Box 1. Cartographic representation, at the service of travellers

Cartographic representation has been based on an increasingly detailed degree of knowledge of our planet over the centuries. Aerial photos and satellites, including the famous Global Positioning System (GPS), have made it possible to clarify the maps of the world from the Renaissance age and the records of the various scientific expeditions of the Enlightenment period. However, ensuring the accuracy of the map is not always necessary. This is how the map of the London Underground came about, the archetype of which dates from 1931. This type of map is now the norm for subway and railway networks.

Closer to our matters, public transport schedules take the form of a tree diagram (stem and leaf plots). In order to concentrate the information, the y-axis represents the hours and the x-axis represents the minutes of the train's departure schedule.

In both cases, we distance ourselves from the thing we want to study: we are far removed from geographical accuracy or the linear representation of time. "The map is not the territory" (Korzybski, 1933).

▶ ...Which have been illustrating statistics since the 19th century

The development of the graphic form took place gradually. For example, changes over time have been shown as curves since the 10th century (Andry et al., 2022), when they were used to describe the motion of planets. Its use for statistics is more recent.

> ***The use of these simple graphical representations is intentional, to allow an understanding at a glance.***

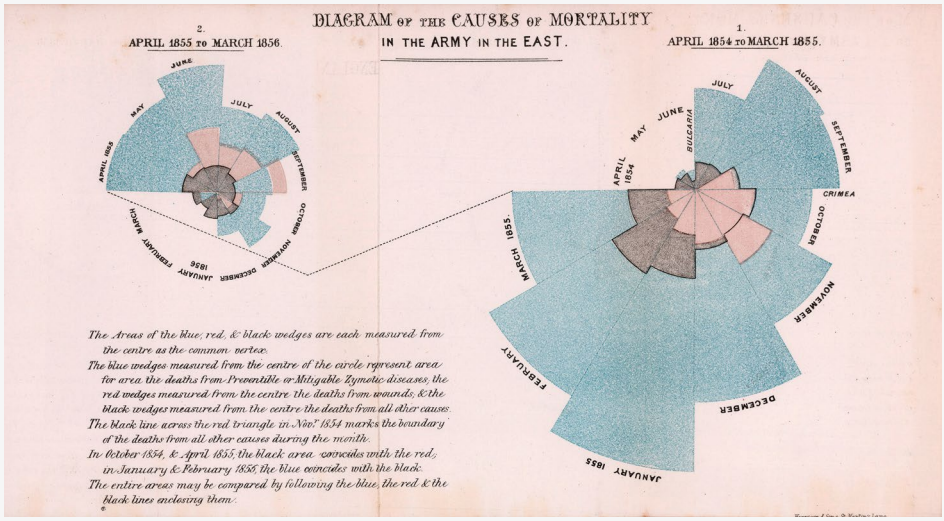
Thus, at the end of the 18th century, William Playfair used the three fundamental modes of statistical representation for the first time, namely the curve for time series, the bar chart for numbers or proportions and the pie chart for proportions. The use of these simple graphical representations is intentional, to allow an understanding at a glance. In the 19th century, the use of the first charts made it possible to identify correlations and new, more complex representations began to appear. Thus, in 1854, John Snow[3] compared deaths and water access points on a map, allowing him to find the origin of the cholera epidemic in London: an infected pump.

As a nurse during the Crimean War in the winter of 1854–1855, Florence Nightingale managed to draw the attention of politicians to the fact that one dies not only on the battlefield, but mainly from wounds due to deplorable sanitary conditions, using a coxcomb diagram (***figure 2***). It was a polar area diagram; each sector, corresponding to a given month, was divided into three concentric areas according to three causes of death (death in combat, infectious disease or other cause).

Whether dealing with proportions or changes over time, there are multiple types of graphic representations adapted to the data available to statisticians to bring meaning to numbers and highlight results, to establish facts and convince people. Hence the importance of choosing the right visualisation method to identify and carry messages.

3 "On the Mode of Communication of Cholera" by John Snow, originally published in 1854 by C.F. Cheffins, Lith, Southhampton Buildings, London, England. *https://archive.org/details/b28985266/page/n57/mode/2up*.

▶ Figure 2 - Diagram of the causes of mortality in the army in the east



Sources: Harrison and Sons, Saint Martin's Lane

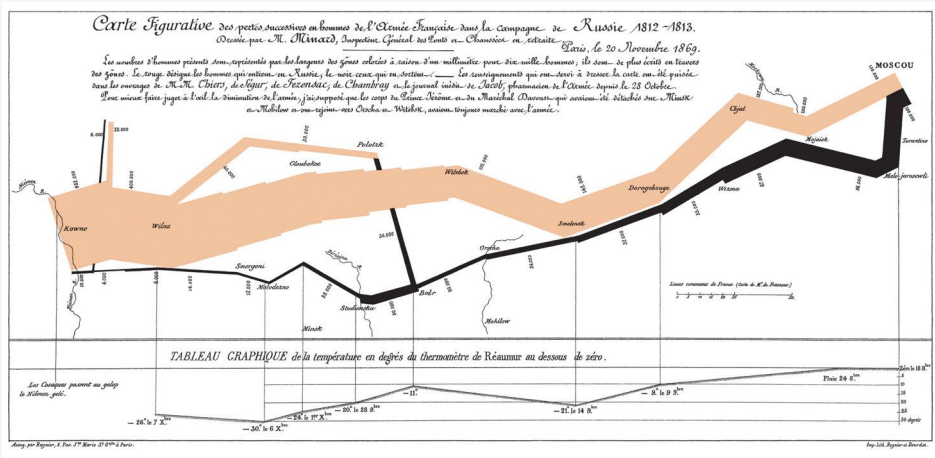
▶ Storytelling

The goals of a datavisualisation are not only to provide support and enable an understanding of the data but also to communicate, illustrate messages and make readers want to read (draw in the reader). Datavisualisation sits at the crossroads between dissemination and communication.

In order to engage readers, using visuals to depict a narrative captures their attention: in English, the term "storytelling" is used. It can take various forms: the map of the French invasion of Russia drawn up by Charles Joseph Minard (1869), one of the first achievements of multidimensional datavisualisation, was similar to storytelling before the digital era. The author succeeds in graphically depicting the movements of the French army between June and December 1812 (***figure 3***). In a single space, he simultaneously describes its itinerary, topographically indicating locations (rivers and cities) and its direction (brown for the approach towards Moscow and black for the retreat); the statistical dimension is incorporated by means of the width of the colouration, which represents the size of the army at each moment, as well as the changes in air temperature during the retreat. Minard developed an original form of cartography with the aim of "making immediately appreciable to the eye, as much as possible, the proportions of numeric results" (Palsky, 1996).

> ***The author succeeds in graphically depicting the movements of the French army between June and December 1812.***

▶ Figure 3 - Flow Map of Napoleon's Russian Campaign of 1812-1813



Sources: Charles Joseph Minard, 1869.
Reproduction from the École nationale des Ponts et Chaussées (National School of Bridges and Roads), Paris.

Put more simply, highlighting information on a complex chart is already a form of data storytelling in itself. For example, singling out one curve among several or isolating one part in a pie chart, provides readers with an implicit reading key, a message.

▶ Towards infographics and "scrollytelling", two forms of storytelling

An infographic includes a chart, a diagram or any other visual image, such as a pictogram; it is intended to present complex information in an easy-to-understand format[4] that is suitable for dissemination via digital or print media. For example, an infographic in the study entitled *Femmes et Hommes : une lente décrue des inégalités* (Women and Men: a Slow Decline in Inequalities) shines a light on the career trajectories of women after obtaining their French baccalauréat diploma. Comparisons between men and women are shown by two podiums (six numbers taken from two figures) (***figure 4***).

In the digital age, another form of storytelling is developing: "scrollytelling" which is a datavisualisation animation used on a web page. This term comes from the contraction of storytelling and scroll, which is the action of moving up and down on a screen. It includes all types of multimedia elements (sounds, texts, videos, infographics, animations, photos or drawings) in a fluid manner thanks to the parallax effect[5]. It can be interactive to a greater or lesser extent and can direct users to different scenarios.

4 According to the Collins Dictionary: *infographics, a graph, diagram, or other visual image designed to present complex information in an easily understandable form*. The Robert and Larousse dictionaries give a more generic definition of infographics (image via a computer).

5 Parallax is a visual effect related to a difference in scrolling speeds between the foreground and the background when scrolling.

▶ Figure 4 - Proportion of Women in Different Higher Education Courses in 2020–2021

Part des femmes dans les différentes formations d'enseignement supérieur en 2020-2021

en %

Formations paramédicales et sociales[1]	86,0
Universités - Langues, lettres, sciences humaines	69,7
Universités - Médecine, odontologie, pharmacie	65,6
Universités - Droit, économie, AES[2]	61,7
Ensemble des universités	59,7
Ensemble	**55,9**
Écoles de commerce, gestion et comptabilité	51,1
Sections de techniciens supérieurs (STS) et assimilés	48,9
Classes préparatoires aux grandes écoles (CGPE)	41,9
Universités - Sciences, Staps	41,4
Préparation au DUT	40,9
Formations d'ingénieurs[3]	28,9

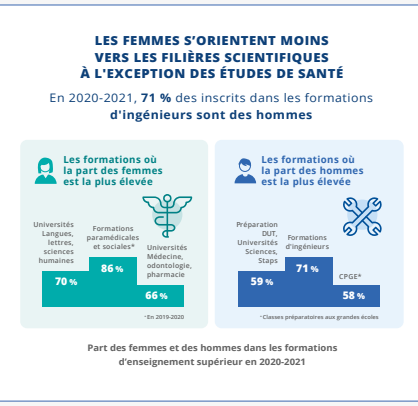
1. Les dernières données disponibles portent sur 2019-2020.
2. Administration économique et sociale.
3. Y compris les formations d'ingénieurs dépendantes des universités, du groupe des instituts nationaux polytechniques (INP), des universités de technologies et les formations d'ingénieurs en partenariat.
Lecture : en 2020-2021, 28,9 % des étudiants en formation d'ingénieurs sont des femmes.
Champ : France.
Source : Mesri-Sies - Systèmes d'information SISE et Scolarité, enquêtes sur les établissements d'enseignement supérieur, enquêtes spécifiques aux ministères en charge de l'agriculture, de la santé et des affaires sociales et de la culture.

Data table

- The data are difficult to understand;
- No stand-out elements;
- Users have to read it, which takes time.

Datavisualisation

- Highlights weak or strong elements through the length of the bars and the colour scheme;
- The data can be read quickly;
- Often accompanied by explanatory text.



Infographics

- Immediate visualisation of the elements to be remembered from a study: text + numbers;
- Requires the producer to sort the information;
- Sets of text, charts, numbers, pictograms;
- Allows easy understanding and better communication.

Collection: Insee Références
Men and Women : Equality in question. (2022 edition)

The Ministerial Statistical Office of the Ministry of Agriculture presented the results of the 2020 agricultural census (Le Grand, 2022) in this manner; similarly, the UK National Statistical Office uses this approach to disseminate the results of the population census (*How the population changed where you live: Census 2021*). Scrollytelling is engaging for the reader but requires a great effort in terms of storytelling, such as the infographics associated with the studies. The approach consists in ordering the results and presenting them in decreasing order of importance using an inverted pyramid technique, which is a journalistic editorial technique (Angel, 2009).

▶ Interactive datavisualisation

With scrollytelling, the user is active, scrolling the pages, but interactive datavisualisation goes further in its relationship with the numbers: it allows the data to be explored freely by anyone. It offers users the opportunity to explore the data, like the statisticians who examine them before disseminating them (***Box 2***).

The development of interactivity and the plethora of graphical representations available make it possible to handle the greatest number of large data sets. Interactive age pyramids were one of the first exercises of this type: by using animation, the tool allows users to scroll through the results produced over more than 70 years. A more recent example, the reference book entitled "*Tableaux de l'économie française*" (Tables on the French Economy) has been completely redesigned and reworked with a focus on users: while the annual paper version has been preserved, the product is now modernised and continuously updated with the latest information, providing a better service to Internet users with different reading levels, while preserving simplicity for data managers. The definition of this book required compromises in the selection of indicators and their use, through an assessment of their structural nature, their timeliness and their ability to be understood by the public. The nomenclature used for the topics has been modelled on the one used on the "insee.fr" website so that users can easily navigate through the various sections. Most recently, the interactive tool on *life expectancies* brings together in a single resource all of the life expectancies calculated by INSEE since 1946, broken down by sex, age, territory, qualifications, standard of living and social category. Users can take this wealth of information on board through an educational video and can amend the set of interactive figures using various criteria.

However, the user journey must be sufficiently well designed to allow users to find their way around. The user experience (UX) must be designed through testing, especially involving external users.

▶ Mass dissemination...

Statisticians retain their support role in the provision of data. First of all, decisions made in relation to cross-referencing must be relevant (a simple list of eclectic variables is insufficient for the purposes of broad dissemination), as must the way in which the data are formatted. Metadata (concepts, variables and alternatives) are essential to enable proper understanding of the data.

> ***Metadata (concepts, variables and alternatives) are essential to enable proper understanding of the data.***

The Official Statistical Service is increasingly disseminating structured data, which are necessary for datavisualisation, on a massive scale. It provides tools[6] that make it possible to both better discover and better explore data (catalogue of datasets and dynamic tables, respectively); metadata is essential for these uses. Tools for data management (also known as data stewardship[7]) make it possible to broaden the reuse of data; other parties can make use of them to propose other uses of official statistics.

▶ Box 2. Datavisualisation for quality

The exploratory aspect of the data refers to the work of statisticians, in their initial understanding of the collected data, prior to their dissemination. Thus, "Exploratory data analysis is therefore, combined together, instruments that are effective and simple to use, powerful and user-friendly software, the rehabilitation of charts as tools for analysis, but also and above all, an attitude of a statistician towards their problem and their data." (Destandau et al., 1999).

Statisticians use various representations of the data to analyse them, identify particular sub-populations and detect errors; online, the possibilities are varied, the use of datavisualisation tools contributes to improving the quality of the data by multiplying the possibilities for checks. For example, a map of journeys between home and place of study has made it possible to correct outliers in the dissemination of the population census. The first interactive datavisualisation on wages resulted from the validation tools for these data. This interactive tool (*Dataviz salaires*) allows an in-depth exploration of the wage distribution, with different focal points for analysis (age, sector of activity, geography etc.). This type of tool is a powerful data analysis resource: it makes it possible to identify, for a particular sub-population, certain results that would not have been highlighted or found their place in broad-based publications.

Made available on the INSEE website, the tool allows Internet users to access extremely detailed data in a simple and fast (user-friendly) manner. This is an example of the application of the concept of reuse contained in the acronym FAIR*.

* *The acronym FAIR is used to denote the properties of open data: Findability, Accessibility, Interoperability and Reuse.*

▶ ...While retaining the rigour of a statistician

To target as many people as possible, it is advisable to use simple words[8] without abandoning scientific rigour or the use of appropriate concepts. For example, it will be necessary to ensure that definitions are easily accessible on the dissemination site.

6 The Melodi project is developing these tools for INSEE. The Ministerial Statistical Offices of the Ministry of Higher Education, the Ministry of Health and the Ministry of Culture already have operational solutions.

7 "Data stewardship" and "data management" are two terms used to designate the processes, tools and techniques for managing data stored in a company, with a threefold objective of consistency, quality and security. *https://www.oracle.com/fr/database/definition-data-steward/*. *https://www.lebigdata.fr/data-management*.

8 In order to make some of its study results accessible to a wider public, the *Direction de la Recherche, des Études, de l'Évaluation et des Statistiques* (Directorate of Research, Studies, Evaluation and Statistics – DREES), the Ministerial Statistical Office of the Ministry of Health and Solidarity, publishes transcripts using the *Facile à lire et à comprendre* (Easy to Read and Understand – FALC) method. *https://www.culture.gouv.fr/Thematiques/Developpement-culturel/Culture-et-handicap/Facile-a-lire-et-a-comprendre-FALC-une-methode-utile*.

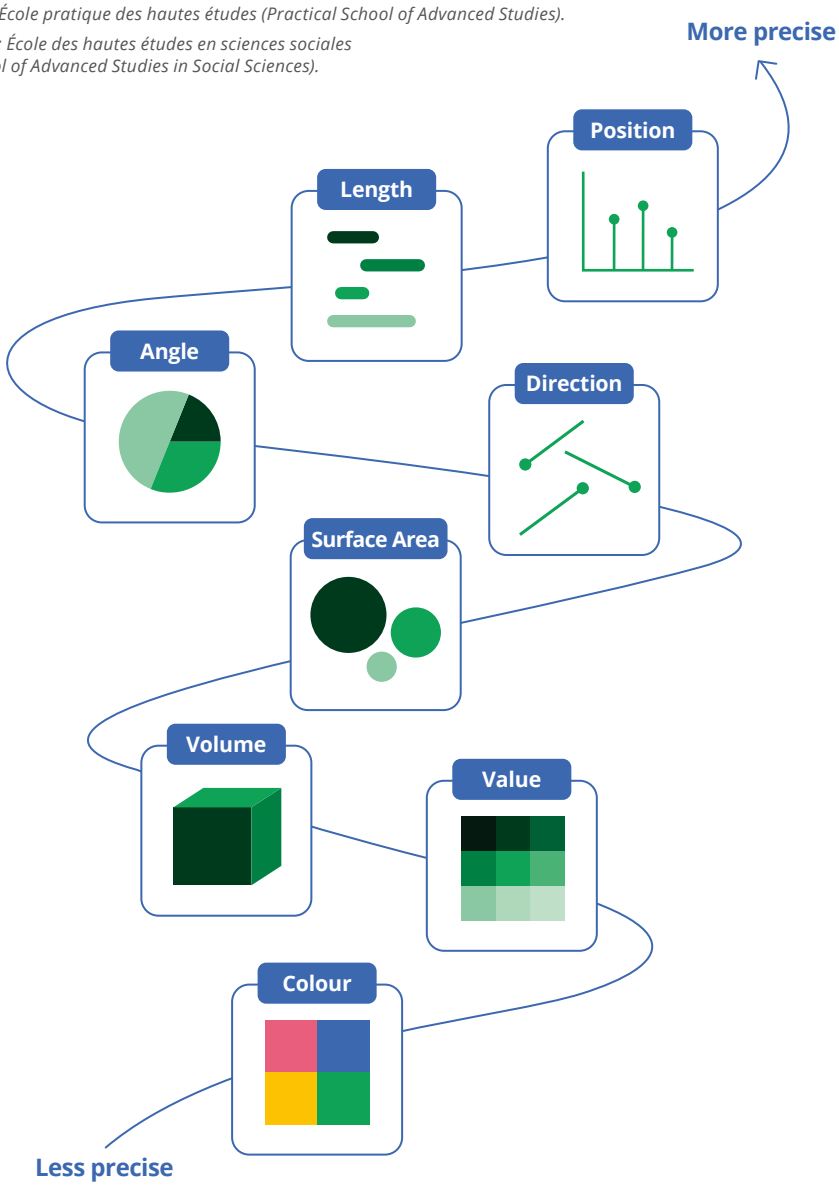
▶ Box 3. Theory around graphic semiotics

The key visual variables in a chart are size, value, texture, colour, orientation and shape (Bertin, 1967). Depending on the nature of the characteristics of the data to be represented, Jacques Bertin determined the most appropriate combination of visual variables: "*...if, in order to obtain a correct and complete answer to a given question, [...] one construction requires a shorter observation time than another construction, it will be said that it is more efficient for this question.*" This body of theoretical principles was created thanks to the experience that he accumulated during his career at the EPHE* (now EHESS**) cartography laboratory.

A few years later, the foundations of a scientific theory were laid to support the principles of semiotics (Cleveland and Mac Gill, 1984), based on tests on the perception of values through the various elements of a chart. The previous system of visual variables was refined. Ultimately, adding one or more scales is preferable to other ways of indicating a value, with shadow and colour being deemed least preferable.

* *EPHE: École pratique des hautes études (Practical School of Advanced Studies).*

** *EHESS: École des hautes études en sciences sociales (School of Advanced Studies in Social Sciences).*



More generally, to ensure that dissemination remains consistent, it is essential to respect a graphic charter and editorial rules, in particular the rules of the Official Statistical Service on datavisualisation tools: statistical unit, clearly displayed data scales, title, date, coverage for each illustration, display of sources, data available for download and respect for statistical confidentiality (Darriau, 2020). This contributes to the brand image of the Institute for Statistics.

As for a publication of four pages or more, the creation of a datavisualisation and, in particular, an infographic requires a radical selection of the information to be presented and a focus on the most significant results. It is necessary to accept that not everything can be said at once, which can be difficult for researchers. However, let us bear in mind that, in the statistical world, datavisualisation usually supplements detailed content in the form of studies, working papers etc. It does not replace data or publications and the associated methodology. Its role is to draw readers in and encourage them to read the study.

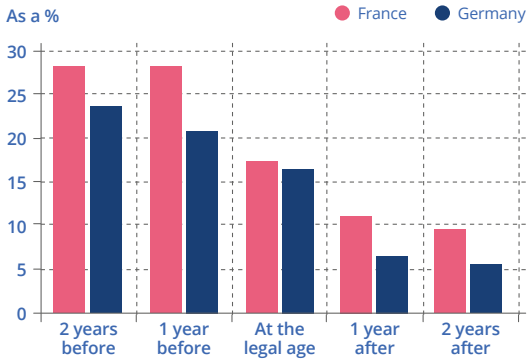
▶ With neat graphic representations...

A datavisualisation is 30 times more likely to be seen and read than a simple text or table. But how do we ensure that the datavisualisation conveys the data in the best manner? Semiotics gives us an answer to this question (***Box 3***): it is "the set of rules of a graphic system of signs for the transmission of information" (Bertin, 1967) that strives for efficiency. Efficiency is reinforced by the idea of minimalism (Tufte, 1983): we then strive for "graphic excellence". The latter is achieved when the amount of information is conveyed to the reader in a minimum amount of time and using the least amount of ink possible. The notion of a data-ink ratio is introduced (***figure 5***): this is the ratio between the – minimal and necessary – part of the chart representing data that cannot be removed without reducing the information disseminated, and the total amount of printed ink used in the chart. Thus, with a grid, it is best to avoid using a visible grid structure that relegates the data to the background or, at least, obstructs the reading of the data. The goal is, in so far as possible, to eliminate everything that can be removed from a chart without the datavisualisation losing meaning (chart junk).

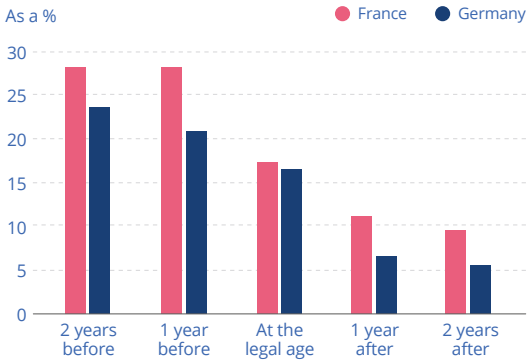
Colours also contribute to a better understanding of charts. But too many colours make a chart unreadable! They must be as clean as possible and use a limited and harmonious pallet (***figure 6***). For example, a single colour should be assigned to a single type of data. Finally, all charts must adhere to accessibility standards for different visual impairments (colour blind people, visually impaired people etc.).

▶ **Figure 5 - Data-Ink Ratios**

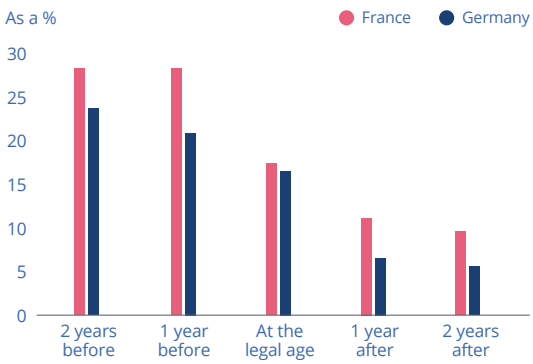
High Data-Ink Ratio



Medium data-ink ratio

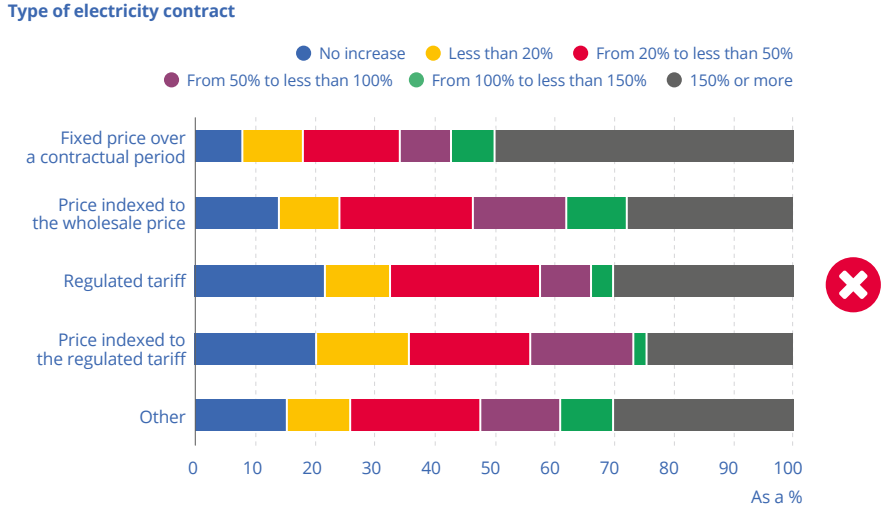


Low Data-Ink Ratio

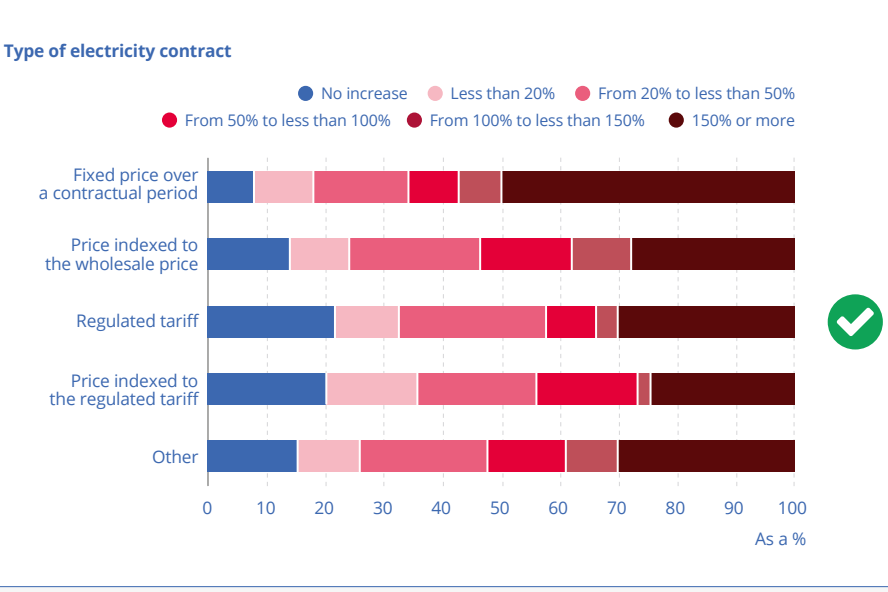


▶ Figure 6 - Example: Price of Electricity

Too Many Colours



Restricted Colour Palette

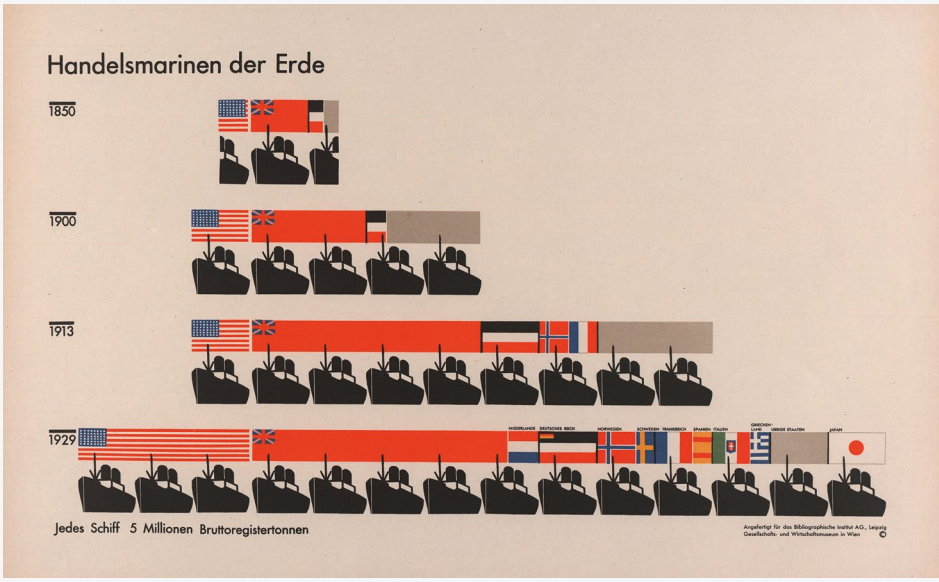


▶ ... and illustrations that speak to viewers

> ***Better access to information for the widest possible population.***

As with text, the use of commonly understood language helps to convey the message in a conventional manner: red for negative and green for positive. Thus, in Germany, isotypes (International System of Typographic Picture Education) were created (Neurath et al., 1925) establishing a simple and universal visual language influenced by the aesthetics of the *Bauhaus* movement[9], allowing better access to information for the widest possible population. These are the forebears of the pictograms (***figure 7***) that we find in datavisualisation, especially in infographics.

▶ Figure 7 - The Merchant Navies of the World



Reading note: Each ship represents 5 million gross tonnes of cargo.

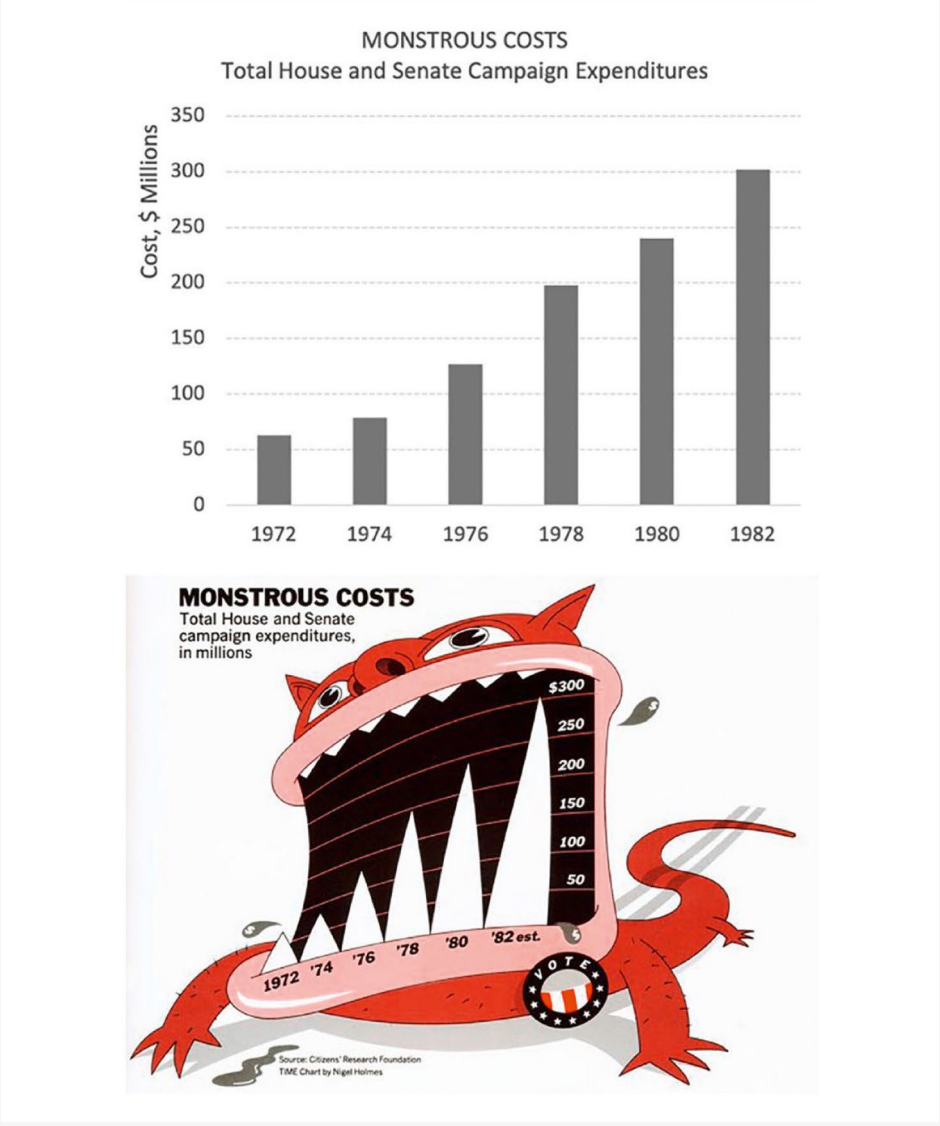
Sources: "Gesellschaft und Wirtschaft" page 55, Verlag des Bibliographischen Instituts AG. Leipzig, 1930.

Newspapers, especially those with large circulations, frequently use illustrations that go beyond the minimalist approach to communicate better, as illustrated by the depiction of costs deemed monstrous using a diagram within the mouth of a monster (***figure 8***). "Embellishments" can help to convey a message and ensure that it is remembered (Bateman et al., 2010).

9 Artistic movement that began in Weimar, Germany, in 1919 and concerned architecture and design in particular.

These authors show, through an experiment using two series of charts (with and without embellishments), that the reading of the charts is not distorted by the additions and that there is no significant difference in short-term memorisation (5 minutes), but that after two or three weeks, embellished charts are described better than others.

▶ Figure 8 - Monstrous Costs



Sources: Holmes, N. Designer's Guide to Creating Charts and Diagrams, Watson-Guptill Publications, 1984.

This is called emotional design (or cognitive science, Norman, 2012), according to which attractive images are remembered better because they stimulate greater enjoyment when used. Definitions of efficiency – focused solely on the visual and functional attributes of datavisualisations – need to be broadened, as they do not take into account the individual and their own characteristics: emotions (Kirk, 2016).

That is quite a bit for statisticians to think about, right? Finally, a fair balance must be found and, while it is beneficial for a chart to leave a mark on readers and enable them to memorise the message of an article, it is important not to sink into a purely seductive approach: the message and rigour must retain primacy.

The author will endeavour to remain neutral, to identify the true message and not to make the statistics lie, such as when the choice of scale masks the real phenomenon; the datavisualisation then conveys a significantly different message (Huff, 1954). This risk of "manipulating" statistics is not new, nor specific to their dissemination in visual format, but it remains or is even potentially amplified. This is especially the case since nuances are difficult to convey; datavisualisation is therefore not necessarily adapted to all types of data.

The foregoing principles should be applied regardless of the type of media used for dissemination and communication, the variety in which has expanded considerably in recent years.

▶ A variety of media forming an ecosystem

In addition to print or online publications, which are traditional media formats, in the 2010s official statistics, particularly in Europe through the DIGICOM programme[10], began opening up to new forms of media and channels for communication: videos with graphic animation (motion design) and social media, especially on smartphones, which rely heavily on the use of images. Datavisualisation has therefore adapted to this ecosystem. Video strengthens the scripting of statistical results, it is very similar to data storytelling since it incorporates a narrative flow (***Box 4***).

▶ Box 4. Webography

Due to the fact that datavisualisation is not a set of fixed rules but a body of rules in constant flux, with a strong degree of innovation, a large proportion of which is developed on the Internet, the following sites contain a variety of datavisualisation output. They include both general-purpose sites and examples of specific interactive applications, which can be used without moderation.*

https://informationisbeautiful.net/.

https://www.awwwards.com/.

https://www.visualcapitalist.com/.

https://www.dataviz-inspiration.com/.

https://visualisingdata.com/.

https://www.data-to-art.com/.

Examples of datascrolling:

https://www.spiegel.de/wissenschaft/zirkel-der-genies-a-90c50289-30ac-4a4b-bc49-348676ce6687.

https://vizagreste.agriculture.gouv.fr/age-et-devenir-des-exploitations-agricoles.html.

* *Websites [accessed 1 December 2023].*

10 DIGICOM was a European digital communication programme, active between 2018 and 2022.

Furthermore, images promote the viral spread of information through legacy media or social media. For example, a tweet about a publication will have all the more impact when it includes an image (a chart from the publication or a specific infographic). The main numerical results of the study will thus reach a large number of readers who will remember them.

The use of these new media formats results in additional steps for the dissemination of the study; for video, time management (to the second) and whether to use live action or motion design must be taken into consideration and for social media, a 280-character tweet must be written with text taken from the leader of the article.

▶ Using the most generic components possible...

Interactive tools, which are even more complex in technical terms, have gradually been developed on the insee.fr website. The first tool was the inflation simulator, followed by the initial age pyramids, both of which were adapted from an experiment conducted by the German Statistical Office (Destatis). These are genuine computer applications that are most often designed around a targeted theme.

Eurostat's experience in this area is interesting: the DIGICOM project incorporated a work package aimed at preparing an interactive publication covering the entirety of Europe, including translation into the various languages. On 20 October 2017, European Statistics Day, the project resulted in the publication entitled "The life of women and men in Europe", published on the INSEE website[11]. This exercise was repeated to cover multiple subject areas, with several of the resulting publications being translated and disseminated on the insee.fr website. Indeed, the defined technical architecture, based on graphical building blocks accompanied by text, was flexible enough to adapt to other subject areas and made it possible to adjust the functions of the product: it was only necessary to develop a new type of visualisation so that it could be used in other publications. This scalability has enabled Eurostat to publish a dozen tools online in just a few years, working mainly on editorial aspects and no longer on the technical building blocks.

> ***It was only necessary to develop a new type of visualisation so that it could be used in other publications.***

▶ ...that are simple and robust over time

At INSEE, the preference was for the in house development of the website: this preference was linked to the fact that dissemination is a core part of the Institute's activities. The chart library on the insee.fr website was retained as part of the Web4G project (2014-2016): designed by statisticians for statisticians, the library fulfilled its expected functions and used a grammar controlled by the various teams that provide the content for the site. Nevertheless, to overcome its obsolescence, the incorporation of new libraries is underway, making it possible to offer 2500 new state-of-the-art products each year, while preserving the historical depth (ten years) that makes the site such a valuable resource.

11 *https://www.insee.fr/fr/outil-interactif/3142332/index.html.*

For the new interactive tools, the return on investment was judged to be better for continuing products (for which the data are updated), such as the French economy dashboard, than for other *ad hoc* products. This tool has been designed to be continuously populated with data by internal teams, by integrating development capabilities for the territorial component from the outset. By proceeding in this manner, it is possible to capitalise by adapting it to different subject areas and to maintain it over time.

Indeed, it is problematic, in the context of long-term official statistics, to develop tools with a short life cycle, high maintenance cost and data that are not updated.

We are therefore a long way off having sophisticated solutions, even if nowadays, with the latest innovations, it is no longer necessary to be an experienced computer scientist to create an application and publish it online.

▶ Managing a scalable environment...

This is made possible by technological developments over the last twenty years. There have been two major areas of transformation, with the emergence of a global data ecosystem, on the one hand, and the democratisation of data processing, on the other. These transformations include:

- The explosion in the volume of data disseminated in the 2000s, with storage on CD and then DVD before widespread Internet access, with XML, CSV and JSON data formats (Dondon et al., 2023);

- The development of open data, with the INSPIRE Directive[12] which paved the way, in 2007, for the unveiling of data.gov in the UK in 2010, followed by the change in legal paradigm with the French law for a Digital Republic and the public data service, which was also rolled out at European level with the introduction of high value datasets: the provision of data free of charge, which was encouraged previously, is now the norm[13];

- The development of processing tools (R v1 in 2000 and the appearance of Rstudio in 2011) and datavisualisation tools in particular (creation of D3.js in 2011) (***Box 5***);

- The development of access to data via the Internet, with REST APIs, which allow the easy exchange of data, and only data. Web services, such as those offered on the api.insee.fr portal, have become the gateway to data, allowing users to filter the field of interest or useful variables.

> ***Rapid and profound movements that are generating challenges for official statisticians.***

These developments form part of rapid and profound movements that are generating challenges for official statisticians. They are actually sources of both constraints and opportunities. Access via an API allows for an independent update of the data and its formatting: a simple update of the data and the new version of the tool is ready; or a simple interface change and the entire history of the dataset is accessible. However, care must

12 The European Directive of 14 March 2007, known as the INSPIRE Directive, aims to establish a geographic information infrastructure to promote the protection of the environment.
13 French Law for a Digital Republic of 7 October 2016 (Article 1).

▶ Box 5. A constantly renewed range of software tools for datavisualisation

As an extension to business intelligence, the "*Tableau*"* (french word for table) tool promises to create a surge in storytelling. Like any proprietary solution for which scaling up requires significant financial investment, "*Tableau*" is currently not in widespread use within the Official Statistical Service. In France, the software is used in particular by the DGFiP**. Other solutions worth mentioning include: *Qlik*, a relatively old stakeholder, *Microsoft PowerBI* and solutions used in the journalistic world like *Datawrapper* or *Infogram*. The approach used by these products is to define a datavisualisation product as one would with an office automation tool; reproducibility is not an objective in and of itself.

Behind these turnkey products, other accessible tools are more complex to handle but offer complementary services that are of use to statisticians. Libraries (such as *Gapminder* and *Highcharts*) offer sets of simple charts for web pages. Hans Rosling's presentation*** on changes in life expectancy and income for each country over 200 years is a benchmark in terms of datavisualisation: in particular, it shows that, while technical work is important, scripting is even more important in engaging the viewer.

In addition, there is the Javascript library "D3.js", the functional richness of which is demonstrated by the home page of its website. The trade-off for this functional richness is a much greater level of complexity in its use. The import and processing of data, as well as their display and presentation, requires a configuration that is restricted to developers.

For statisticians, things are even easier: many "Javascript" libraries are ported under R. A simple "Save as..." with the html widgets package then makes it possible to have code that can be read directly by a browser.

The Official Statistical Service has experimented with various approaches. DREES has developed datavisualisations under *Rshiny*, published online through the site *shinyapps.io*.

Following the lead of the *Jupyter notebooks* used to generate reproducible sequences based on the data, the *Observable* technology****, developed by Mike Bostock, creator of D3.js, allows users to present the entirety of a study as a series of chunks of text, code and results. Each user can customise the results by modifying them to attach to a particular analytical focus and/or by executing them to complete the rest of the analysis.

* *https://www.tableau.com/fr-fr.*

** *DGFIP: Direction générale des finances publiques (Directorate-General for Public Finance) is a directorate within the French central general government, which reports to the Ministry of the Economy.*

*** *https://www.presentationzen.com/presentationzen/2010/07/hans-rosling-tips-on-presenting-data.html.*

**** *https://observablehq.com/.*

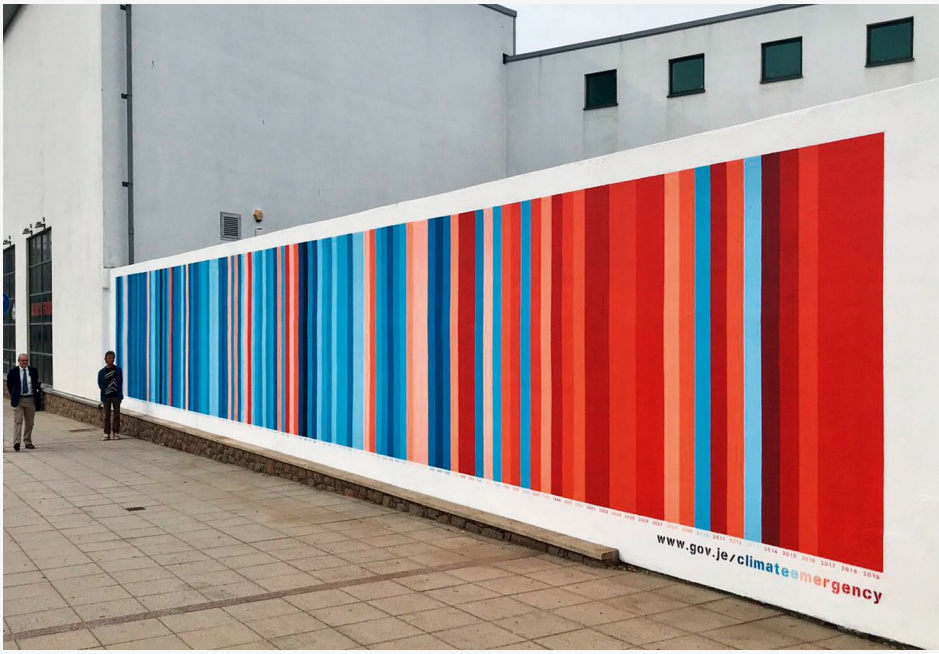
be taken as the pressure for faster delivery can affect data quality. This is the reason behind the importance of automating checks on the initial data.

The open data movement also requires an increasingly broad dissemination, which may potentially encounter constraints due to confidentiality and requires an in-depth analysis of the trade-off between utility and protection of the disseminated data; conversely, it provides or at least facilitates access to new sources of data, thus allowing greater cross-referencing, bringing more information, which must then be assessed and highlighted using datavisualisation.

▶ ...to the (measured) benefit of users?

To measure the impact of datavisualisation in official statistics, we monitor the number of times the products concerned are accessed online, the number of views of the videos, the number of tweets etc. The result of our actions becomes measurable and is measured continuously.

▶ Figure 9 - Climate Stripes



Reading note: Local artist Ian Rolls has created a new mural showing the average increase in air temperatures in Jersey, as a way of drawing attention to climate change.

Sources: https://www.bailiwickexpress.com/jsy/news/126-reasons-be-green/. © Copyright Bailiwick Publishing 2023

> ***Datavisualisation must remain an opportunity to communicate better and convey a strong brand image of official statistics in accordance with its time.***

The objective is not only to capture the attention of users and maximise the time they spend reading, but also to elicit user feedback to verify that we are reaching all audiences and that the figures speak for themselves. A good datavisualisation can be widely shared, thus increasing its impact on public debate. For example, the illustration of the change in annual global temperature since 1850 has been shared and is now well known (***figure 9***).

Ethics then plays a major role, to remain neutral and to maintain the necessary distance to ensure an objective understanding of the figures. Datavisualisation must remain an opportunity to communicate better and convey a strong brand image of official statistics in accordance with its time.

Going gurther in datavisualisation to communicate

Finally, official statistics has reached a turning point in relation to datavisualisation: it was disseminated in both a static version, allowing a better understanding of the publications through images and scripting, and a dynamic version, allowing users to see stacks of data. The problems in relation to the latter aspect are significant and relate to the identity of the statistician concerned: to what extent should we reveal the data and what is the level of quality necessary for dissemination?

Specific resources are usually needed (design and assembly): the need for a "talented designer" is identified (Bateman et al., 2010). It is necessary to combine the skills of a statistician and a graphic designer to create products that are not only aesthetically pleasing and engaging, but also rigorous and informative.

The wide availability of data allows other parties to obtain it and create their own datavisualisations. Data journalists increasingly relay the information disseminated by official statistics in the form of datavisualisations, which gives added value to the data produced.

To a lesser extent, should official statistics engage in data-art, which pushes aesthetics to the level of a work of art in the representation of data? Engaging the "reader" is an essential issue and the figures must leave an impression in the minds of readers. During the exhibition held in partnership with the SNCF for the 75th anniversary of the Institute, our work was presented in the form of large educational panels. Beyond the quality of our data, INSEE's presence outside its walls is essential for conveying its messages.

> ***An extremely effective way to share information and communicate in relation to statistics.***

In the age of the Internet, with an ever-increasing amount of data and less and less time to read it, datavisualisation is an extremely effective way to share information and communicate in relation to statistics. Datavisualisation is developing further and further, in an increasingly sophisticated manner, and the trend does not seem close to stopping. It therefore has a genuine role to play in public debate, as well as a crucial role in communication.

Bibliography

- ANDRY, Tiffany, KIEFFER, Suzanne and LAMBOTTE, François, 2022. De Boeck Supérieur. ISBN 978 2807341579.
- ANGEL, Jean-William, 2009. Plan d'un article : inversez la pyramide ! In: *Courrier des statistiques*. Insee Hors série, December 2009, pp. 21-24. [online]. [Accessed 9 August 2023]. Available at: *https://www.bnsp.insee.fr/ark:/12148/bc6p06xt18x.pdf*.
- BATEMAN, Scott, MANDRYK, Regan L., GUTWIN, Carl, GENEST, Aaron, Mc DINE, David and BROOKS, Christopher, 2010. *Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts*. [online]. [Accessed 9 August 2023]. Available at: *http://www.stat.columbia.edu/~gelman/communication/Bateman2010.pdf*.
- BERTIN, Jacques, 1967. *Sémiologie graphique : les diagrammes, les réseaux, les cartes*. Gauthier-Villars.
- BERTIN, Jacques, 1977. *La graphique et le traitement graphique de l'information*. Flammarion.
- BRYSBAERT, Marc, 2019. *How many words do we read per minute? A review and meta-analysis of reading rate*. Journal of Memory and Language, Volume 109, December 2019.
- CLEVELAND, William S. and McGILL, Robert, 1984. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, Vol. 79, No. 387. September 1984, pp. 531-554.
- CLEVELAND, William S. *How William Cleveland Turned Data Visualization Into a Science*. [online]. [Accessed 9 August 2023]. Available at: *https://priceonomics.com/how-william-cleveland-turned-data-visualization/*.
- DARRIAU, Valérie, 2020. Grid Data, Innovative Tools and Methods Used to See the Reality in the Territories. In: *Courrier des statistiques*. [online]. 31 December 2020. No N5, pp. 53-73. [Accessed 9 August 2023]. Available at: *https://www.insee.fr/en/information/6043143?sommaire=5894773*.
- DESTANDAU, Sophie, LADIRAY, Dominique and LE GUEN, Monique, 1999. In: *Courrier des statistiques*. [online]. June 1999. Insee. No 90. [Accessed 9 August 2023]. Available at: *https://www.bnsp.insee.fr/ark:/12148/bc6p06xt287/f1.pdf*.
- DONDON, Alexis and LAMARCHE, Pierre, 2023. Quels formats pour quelles données ? In: *Courrier des statistiques*. [online]. 30 June 2023. Insee. No N9, pp. 86-103. [Accessed 9 August 2023]. Available at: *https://www.insee.fr/fr/information/7635827?sommaire=7635842*.
- DORTIER, Jean-François, 2012. Une histoire des sciences humaines. pp.150-153. ISBN 978-2361061678.
- HUFF, Darrel, 1954. *How to Lie With Statistics*. Norton, New York, ISBN 0-393-31072-8.
- KIRK, Andy, 2016. *Data visualisation: A handbook for data driven design*. Londres, Royaume-Uni: Sage Publications.

- KORZYBSKI, Alfred, 1933. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. p. 58.
- LE GRAND, Hervé, 2022. The 2020 Agricultural Census – Five Innovations That Will Go Down in History. In: *Courrier des statistiques*. [online]. 20 January 2022. Insee. No N7, pp. 48-67. [Accessed 9 August 2023]. Available at: *https://www.insee.fr/en/information/7722517?sommaire=7722566*.
- NEURATH, Otto, 1939. *Modern Man in the Making*. Lars Muller Publishers. ISBN 978-3037786765.
- NIGHTINGALE, Florence, 1858. "*Diagram of the causes of mortality in the army in the East*". Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army.
- NORMAN, Donald A., 2012. *Design émotionnel : pourquoi aimons-nous (ou détestons-nous) les objets qui nous entourent ?* Bruxelles, Belgique: De Boeck.
- PALSKY, Gilles, 1996. *Des chiffres et des cartes. Naissance et développement de la cartographie quantitative au XIX*e *siècle*. Collection : Mémoires de la section de géographie physique et humaine - No 19.
- PLAYFAIR, William, 1805. *A Statistical Account of the United States of America by D. F. Donnant*. London: J. Whiting. William Playfair.
- RENDGEN, Sandra, 2020. *Le Système Minard - Anthologie des représentations statistiques de Charles-Joseph Minard* - Collection de l'École nationale des ponts et chaussées. 20 November 2020. Éditions B42. ISBN 978-2-490077-45-8.
- RODIGHIERO, Dario, 2021. *Mapping Affinities: Democratizing Data Visualizations*. Métis Presses.
- STANDING, Lionel, CONEZIO, Jerry and HABER, Ralph, 2013. *Perception and memory for pictures: Single-trial learning of 2500 visual stimuli*. [online]. [Accessed 9 August 2023]. Available at: *https://link.springer.com/article/10.3758/BF03337426*.
- TUFTE, Edward, 2001. *The Visual Display of Quantitative Information*. Graphics Press.