JDSSV

Journal of Data Science, Statistics, and Visualisation

MMMMMM YYYY, Volume VV, Issue II. *doi: XX.XXXXX/jdssv.v000.i00*

Visual narratives of the Covid-19 pandemic

Susan Vanderplas
University of Nebraska–Lincoln

Adalbert F.X. Wilhelm
Jacobs University Bremen

Abstract

Covid-19 has sparked a worldwide interest in understanding the dynamic evolution of a pandemic and tracking the effectiveness of preventive measures and rules. For this reason, numerous media and research groups have produced comprehensive data visualisations to illustrate the relevant trends and figures. In this paper, we will look at a selection of Covid 19 data visualisations to evaluate and discuss the currently established visualisation tools in terms of their ability to provide a communication channel both within the data science team and between data analysts, domain experts and a general interested audience. Although there is no set catalogue of evaluation criteria for data visualisations, we will try to give an overview of the different core aspects of visualisation evaluation and their competing principles.

Keywords: exploratory data visualisation, logarithmic scales, visual comparisons, R.

1. Introduction

Over the past two years, several waves of Covid-19 infections with different mutants of the SARS-CoV-2 virus have swept across the globe, claiming many lives, causing numerous health damages and affecting our personal lives in many ways. According to the WHO, over 304 million confirmed cases and over 5.4 million deaths have been reported as of early January 2022 [1]. The pandemic has generated enormous interest

[1] https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---11-january

in epidemiological data, its analysis and visualisation. From the beginning of the pandemic, data on the number of infections and covid-related deaths has been published daily and made available to the public. Media, politicians and individuals use this data to build their narratives about the pandemic, discuss its evolution, justify the measures taken and discuss different prevention strategies against the spread of the virus. So there are several goals to be achieved by visualising Covid-19 data. These goals and their priorities were adapted as dynamically as the virus mutated and the pandemic changed pace. But as with any other data visualisation two general principles remain the same: ensuring clear vision by optimising the data-to-ink ratio (Tufte 2001) (or signal-to noise ratio) and ensuring clear understanding by organising the graphics in such a way that the story of the data is told most effectively.

In recent decades, many media organisations have established data teams that received unprecedented attention and wide-ranging opportunities during the pandemic as they have been showcasing their skills and abilities, not only in visualising data, but also in explaining their data collection and data analysis strategies and methods. Data journalism will certainly be one of the beneficiaries of the Covid-19 pandemic and it has become an innovative part of news publishing, with COVID-19 delivering many excellent applications, often presented in an interactive visual format on the web, such as dashboards.

At the same time we still see a lot of defective graphics disseminated and shared: some that violate fundamental visualisation and statistical reporting principles such as accuracy, relevance, timeliness, clarity, coherence, and reproducibility. These principles have been laid out in numerous standards for statistical reporting in the application areas, such as the ESS standard for quality reporting, the CONSORT, PRISMA, CHEERs guidelines, and others (see https://equator-network.org). Numerous publications, initiatives, and ideas to improve the communication of quantitative and statistical information have been prepared, see for example Hoffrage et al. (2000); Tufte (2001); Rosling and Zhang (2011); **?**.

The Covid-19 pandemic has provided ample evidence that policy measures are only effectively implemented and adhered to by the people if they are accepted by a vast majority of the population. To achieve this goal, effective communication of quantitative evidence and statistical results among scientists, governments, media, and the citizens is essential to justify the adequacy, usefulness, and relevance of the measures.

A well proven method to effective communication lies in representing information in accessible ways (Gigerenzer et al., 2007; Gigerenzer and Edwards, 2011) by saying, for instance, “one in 10” instead of 10%. Using absolute in place of relative numbers, representing information in appealing graphic forms, and summarizing most relevant facts in “fact boxes” are some of the methods developed and advocated for increasing transparency in communication between stakeholders such as health providers and patients (See, for example: https:// www. hardingcenter.de) Fact boxes combined with icon arrays are recommended for the presentation of test results. Both representations are based on natural frequencies (**??**) and present case numbers as simply and concretely as possible. Many scientific studies show that icon arrays help people understand numbers and risks more easily (e.g. **?**). The Harding Center for Risk Literacy shows many

other examples of transparent communication of risks, including COVID-19[2].

Nevertheless, both the complexity of the phenomena and the "bipolarity" of statistical thinking remain challenges.

While human thinking tends towards pattern simplification and political communication also prefers a simple cause-effect relationship, real phenomena are often multivariate. Thus, when studying COVID-19 and predicting its spread, it is not only important to consider its symptomatology, the incidence and geographic distribution of diseases, population behavior patterns, government policies and impacts on the economy, on schools, on people in nursing homes and on social life as a whole, but to integrate these into the data analyses and the communication of results. Associations observed in the data can often be caused by third-party variables (confounders). In addition, much of the data comes from observational studies, which usually makes a robust causal attribution problematic. Statisticians calling out these limitations, however, face the danger that their statements might be pulled out one side in a polarized debate (McConway and Spiegelhalter 2021).

2. The global perspective

On 11 March 2020, WHO declared the outbreak of the novel coronavirus disease (COVID-19) a pandemic, and since that date at the latest, the global perspective of the disease has been in the public eye. The spatial spread of the virus and the resulting cases and deaths are commonly visualized by choropleth maps, see for example Figure 1 showing the total number of infections reported in each country as of January 14, 2022. For the pandemic perspective a central element of the narrative is the ubiquity of the disease and the accompanying global impact. Choropleth maps based on raw numbers of cases might look convincing and fit to the purpose, but neglect a number of well-know caveats for statistical reporting and visualisation:

1. Absolute value unsuitability: As explained in (Monmonier 2005; Slocum et al. 2008; Speckmann and Verbeek 2010) among others, choropleth maps are fundamentally unsuitable for the representation of absolute numbers. Especially in the case of similarly coloured areas of the regions, viewers tend to integrate them unconsciously and perceive choropleths as representations of density. They also do not help to convey the desired message as the absolute numbers of Covid-19 cases are strongly influenced by the population size of the country, but also by the number of tests performed and the accuracy of the recording and reporting system.

2. The area-bias: The visual impression is determined more by the colour and the geographical area of the individual countries than by the number of Covid cases. Since the countries of the world differ extremely in area, the visual assessment is distorted, especially in the case of neighbouring countries with similar numbers but different areas.

3. Color-scheme obstructions: Much research in visualisation is concerned with the

[2] https://www.hardingcenter.de/de/mrna-schutzimpfung-gegen-covid-19-fuer-aeltere-menschen

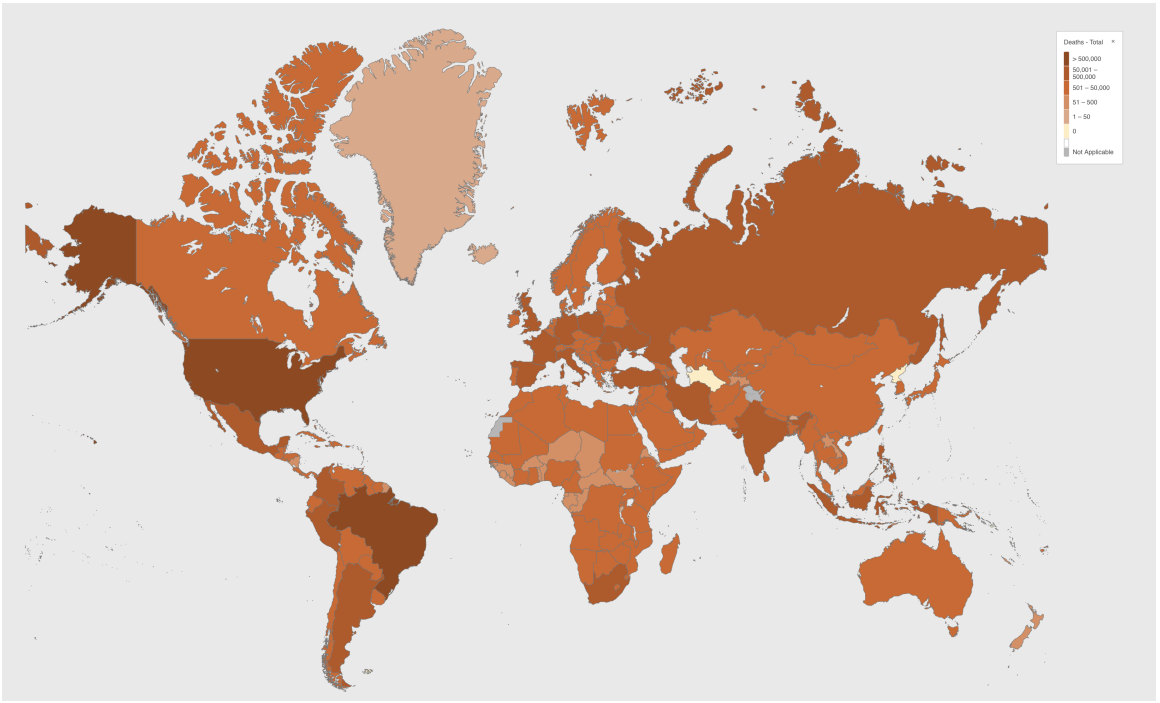

Figure 1: Choropleth map of the Covid-19 cases world-wide by country. Source: WHO https://covid19.who.int

appropriate choice of colour schemes, see (Brewer et al. 1997). The choice of a continuous scale or a categorical scale, the choice of a scale that promotes the recognition of patterns or a scale that supports the filtering out of specific map details, influences the quality of choropleth maps.

Simulations were used particularly illustratively in the media in the course of the pandemic. An inspiring example illustrating the spread of the epidemic appeared as early as 14 March 2020 in the Washington Post[3] with the title "Why outbreaks like coronavirus spread exponentially, and how to flatten the curve". The Washington Post made this simulation available free of charge and in all major languages, which led to it being distributed worldwide, including repeatedly on German television[4]. The New York Times[5] published a dynamic graphic entitled "How the Virus Won", which maps the spread of COVID-19 cases from February to June 2020 in the USA. It shows how an analysis of the associations between different COVID-19 strains and travel patterns can help understand the spread of the disease.

Another illustrative example is a simulation from ZEIT Online[6], which - based on models developed by a group of researchers at the Max Planck Institute for Chemistry - estimates the probability of an infected person infecting other people in closed rooms in various scenarios. While the visualisations present the simulated infection processes in a catchy way, the dependence of the simulations on parameter assumptions and

[3] https://www.washingtonpost.com/graphics/2020/world/corona-simulator/
[4] https://web.br.de/interaktiv/corona-simulation/
[5] https://www.nytimes.com/interactive/2020/us/coronavirus-spread.html
[6] https://www.zeit.de/wissen/gesundheit/2020-11/coronavirus-aerosole-ansteckungsgefahr-infektion-

settings is usually not addressed. Simulations should also always make transparent on which model assumptions and which data basis the simulations were created.

3. Time series

4. Comparisons and rankings

5. To log or not to log

As COVID cases grow quasi-exponentially while there are susceptible members of the population (subject to the effectiveness of mitigation measures and testing availability), it seems natural to use log scales to allow for more effective comparisons of slight changes in case counts over time. In addition, log scales make it possible to compare regions with different populations or infection rates in the same chart. As noted previously, however, interpreting log scales requires levels of numerical sophistication that may not be appropriate for the general public. Even researchers do not always read and interpret log scales correctly(Menge et al. 2018); expecting the general public to do so is difficult under normal circumstances(Heckler et al. 2013) is difficult. When panic, fear, uncertainty, and doubt about the situation are added to the mix,

One issue with assessing the use of log scales is that their effectiveness changes with the stage of the pandemic and the amount (and varieties) of data shown. Initially, log scales were incredibly useful at showing case counts, because minimal mitigation measures were in place and the growth of case counts (or presumptive positive cases, in absence of available testing) was fairly close to exponential. In addition, the use of log scales allowed for the comparison of nominal cases across entities with large population differences: in the US, we could compare cases in New York and California with cases in Michigan and Washington, even though the population of Michigan and Washington are much lower than the population of either New York or California.

While log scales are not necessarily intuitive, many outlets tried to make the graphs more intuitive by adding reference lines, as shown in Figure 3.

However, after the first wave of COVID, the issues with log scales became more apparent: it was difficult to detect slight increases in case counts that indicated the beginning of a new wave amid a background level of spread, as demonstrated in Figure 4. Diagonal reference lines from the origin were also less helpful, as the growth of cases or deaths was no longer approximately exponential and varied over time; for these reference lines to be effective there would need to be a clear idea of when the case counts started to increase exponentially, which is difficult to determine whilst in the thick of a potential COVID wave.

While log scales have their problems, linear scales are not immune from issues either. It can be very difficult to adequately compare to past situations when looking at the full time series of case counts. For example, in Figure 5, it is difficult to tell whether the first wave of COVID cases in March 2020 had an increase as fast as that in January of 2021; it is even more difficult to compare the order-of-magnitude of change in case rate

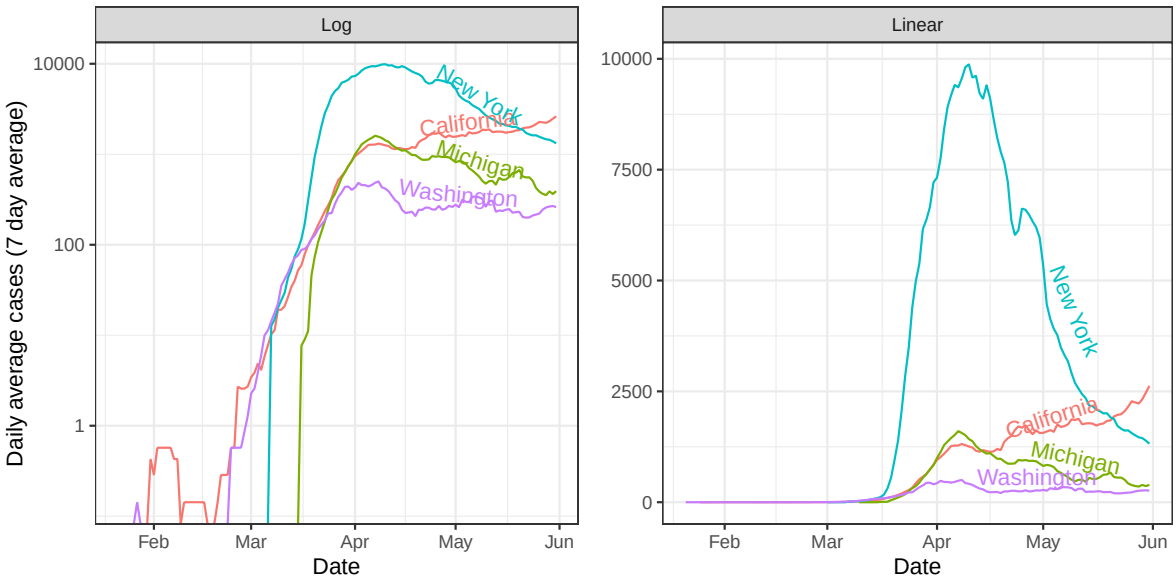

Figure 2: In the early stages of the pandemic, log scales allowed the comparison of raw case counts in locations with vastly different population and case counts.

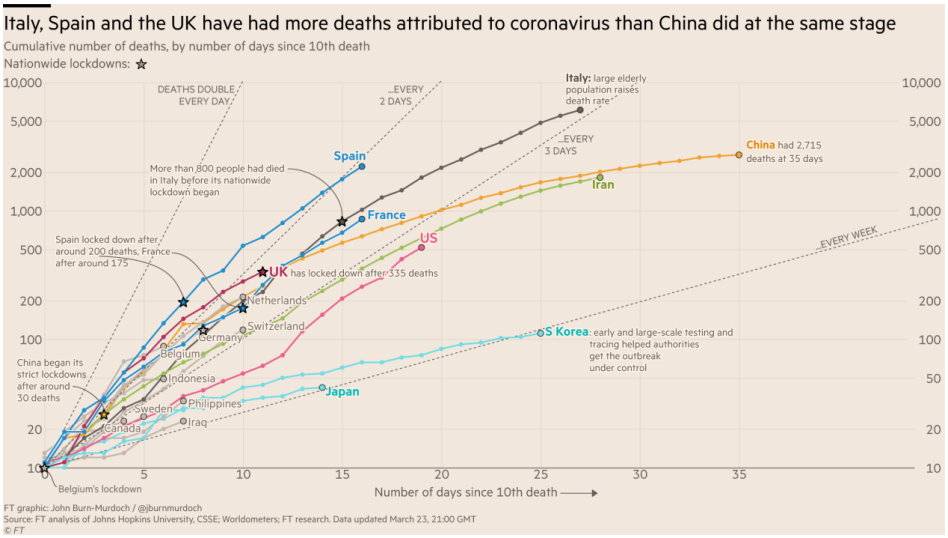

Figure 3: Reference lines to compare exponential growth rates of deaths in different countries. This provides some additional context that may help individuals use log scale data more successfully. This approach was first featured in the Financial Times, but was quickly adopted by the New York Times, 91-DIVOC, and other outlets. Graph from the Financial Times (March 23, 2020), image from Kosara (2020).

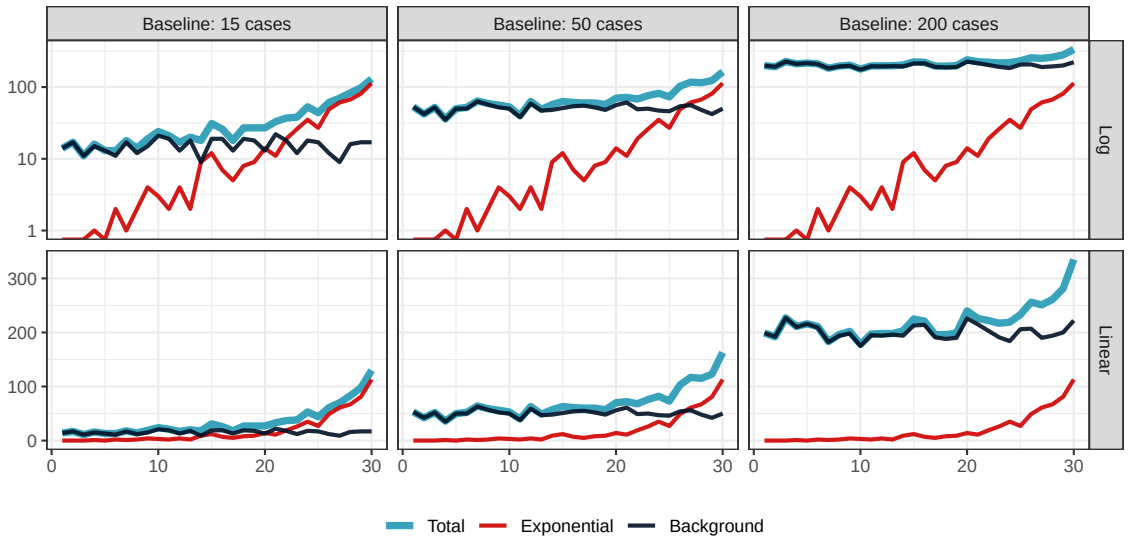

Figure 4: One problem with log scales is that if there is a background level of spread, it can be hard to notice the introduction of an additional source of exponential spread. Linear scales do not have this problem - the exponential source is noticeable very quickly in the total line, but on the log scale it is much harder to discern when the exponential source causes the total line to diverge from the background. In the top-right corner, it is difficult to identify that there is an exponential increase in cases amid the baseline, even though the exponential source makes up approximately 50% of the cases at the end of the time period shown.

growth of January 2021 relative to January 2022 when the more contagious omicron variant became prevalent.

It is not clear that the use of log or linear scales during the COVID-19 pandemic had a large effect on public opinion. Several studies were conducted in the early stages of the pandemic (Romano et al. 2020; Sevi et al. 2020; Ryan and Evers 2020) and results seem to suggest that while individuals have difficulty understanding log scale graphs, these issues do not tend to affect their support for intervention measures (perhaps in part because COVID-related news saturated the news and opinions were set outside of information provided in the experiments).

6. Summary and discussion

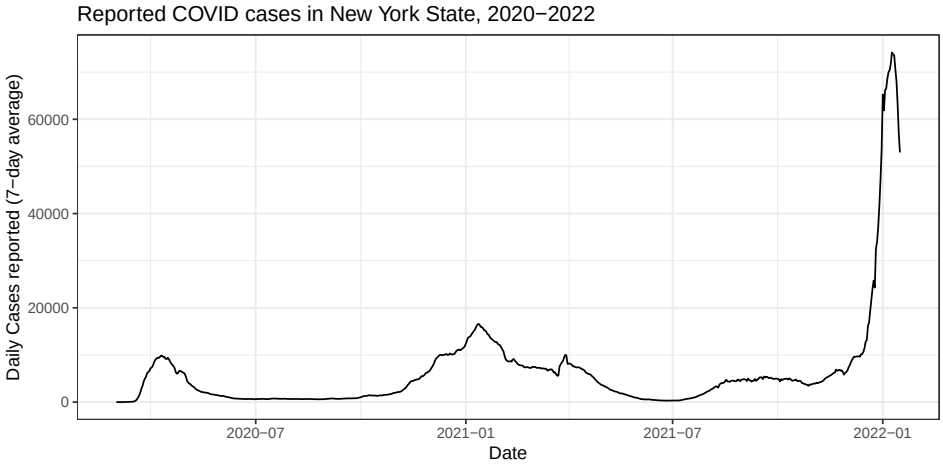

Figure 5: Reported COVID cases in New York State, 2020-2022. The linear scale makes it difficult to compare the trajectory of different waves to determine how severe the current status is relative to the past, because the primary contrast is the height of the relative peaks, rather than the growth *rate*. A similar graph on the log scale would have the peaks at much more similar heights (though there would still be a difference), allowing the reader to focus on other information.

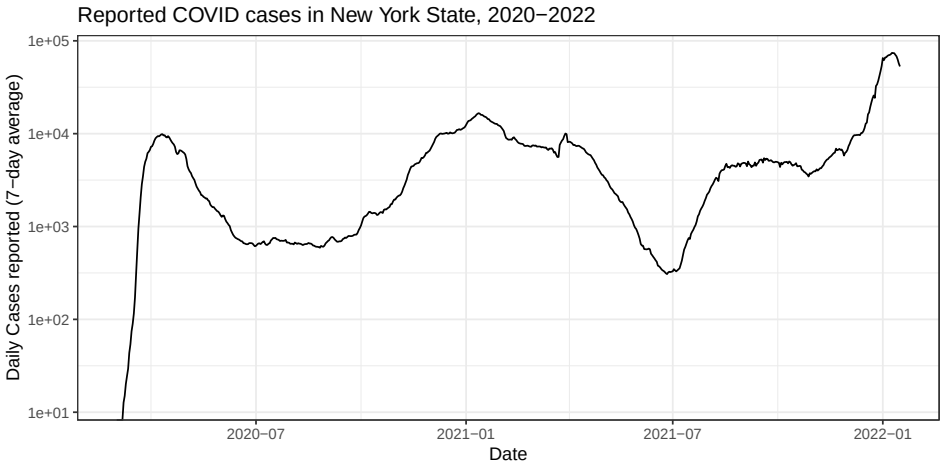

Figure 6: Reported COVID cases in New York State, 2020-2022. The linear scale makes it difficult to compare the trajectory of different waves to determine how severe the current status is relative to the past.

Computational Details

If necessary or useful, information about certain computational details such as version numbers, operating systems, or compilers could be included in an unnumbered section. Also, auxiliary packages (say, for visualizations, maps, tables, ...) that are not cited in the main text can be credited here.

The results in this paper were obtained using R 3.5.1. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

Acknowledgments

All acknowledgments should be collected in this unnumbered section before the references. It may contain the usual information about funding and feedback from colleagues/reviewers/etc. Furthermore, information such as relative contributions of the authors may be added here (if any).

References

Brewer, C. A., MacEachren, A. M., Pickle, L. W., and Herrmann, D. (1997). Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers*, 87(3):411–438, ISSN: 00045608, 14678306, http://www.jstor.org/stable/2564061.

Heckler, A. F., Mikula, B., and Rosenblatt, R. (2013). Student accuracy in reading logarithmic plots: The problem and how to fix it. In *2013 IEEE Frontiers in Education Conference (FIE)*, pages 1066–1071. DOI: 10.1109/FIE.2013.6684990.

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science*, 290(5500):2261–2262, ISSN: 0036-8075, DOI: 10.1126/science.290.5500.2261, https://science.sciencemag.org/content/290/5500/2261.

Kosara, R. (2020). In Praise of the Diagonal Reference Line. https://eagereyes.org/blog/2020/in-praise-of-the-diagonal-reference-line.

McConway, K. and Spiegelhalter, D. (2021). Sound human, steer clear of jargon, and be prepared. *Significance*, 18(2):32–34, DOI: 10.1111/1740-9713.01508.

Menge, D. N. L., MacPherson, A. C., Bytnerowicz, T. A., Quebbeman, A. W., Schwartz, N. B., Taylor, B. N., and Wolf, A. A. (2018). Logarithmic scales in ecological data presentation may cause misinterpretation. *Nature Ecology & Evolution*, 2(9):1393–1402, ISSN: 2397-334X, DOI: 10.1038/s41559-018-0610-7, http://www.nature.com/articles/s41559-018-0610-7.

Monmonier, M. (2005). Lying with maps. *Statistical Science*, 20(3):215–222, ISSN: 08834237, http://www.jstor.org/stable/20061176.

Romano, A., Sotis, C., Dominioni, G., and Guidi, S. (2020). The scale of COVID-19 graphs affects understanding, attitudes, and policy preferences. *Health Economics*, 29(11):1482–1494, ISSN: 1099-1050, DOI: 10.1002/hec.4143, https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.4143.

Rosling, H. and Zhang, Z. (2011). Health advocacy with gapminder animated statistics. *Journal of Epidemiology and Global Health*, 1:11–14, ISSN: 2210-6014, DOI: https://doi.org/10.1016/j.jegh.2011.07.001, https://doi.org/10.1016/j.jegh.2011.07.001.

Ryan, W. H. and Evers, E. R. K. (2020). Graphs with logarithmic axes distort lay judgments. *Behavioral Science & Policy*, 6(2):13–23, ISSN: 2379-4615, DOI: 10.1353/bsp.2020.0011, https://muse.jhu.edu/article/799814.

Sevi, S., Aviña, M. M., Péloquin-Skulski, G., Heisbourg, E., Vegas, P., Coulombe, M., Arel-Bundock, V., Loewen, P. J., and Blais, A. (2020). Logarithmic versus Linear Visualizations of COVID-19 Cases Do Not Affect Citizens' Support for Confinement. *Canadian Journal of Political Science*, 53(2):385–390, ISSN: 0008-4239, 1744-9324, DOI: 10.1017/S000842392000030X.

Slocum, T., McMaster, R., Kessler, F., Howard, H., and Mc Master, R. (2008). *Thematic Cartography and Geographic Visualization*. Prentice Hall, 3rd edition.

Speckmann, B. and Verbeek, K. (2010). Necklace maps. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):881–889, DOI: 10.1109/TVCG.2010.180.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 2 edition.

A. More Technical Details

Appendices can be included after the bibliography (with a page break). Each section within the appendix should have a proper section title (rather than just *Appendix*).

B. Using BibTeX

References need to be provided in a BibTeX file (`.bib`). All references should be made with `\cite`, `\citet`, `\citep`, `\citealp` etc. (and never hard-coded). This commands yield different formats of author-year citations and allow to include additional details (e.g., pages, chapters, ...) in brackets.

Cleaning up BibTeX files is a somewhat tedious task – especially when acquiring the entries automatically from mixed online sources. However, it is important that informations are complete and presented in a consistent style to avoid confusions. JDSSV requires the following format.

- Specific markup (`\proglang`, `\pkg`, `\code`) should be used in the references.
- Titles should be inserted in title case.
- Journal titles should not be abbreviated and in title case.
- DOIs should be included where available.
- Software should be properly cited as well. For R packages `citation("pkgname")` typically provides a good starting point.

Affiliation:

email: susan.vanderplas@unl.edu Susan Vanderplas
Department of Statistics
University of Nebraska–Lincoln
349A Hardin Hall
Lincoln, NE 68583-0963, USA
E-mail: susan.vanderplas@unl.edu
URL: https://statistics.unl.edu/susan-vanderplas

Adalbert F.X. Wilhelm
Department of Psychology and Methods
Jacobs University Bremen gGmbH
Campus Ring 1
28759 Bremen, Germany
E-mail: a.wilhelm@jacobs-university.de
URL: http://www.jacobs-university.de/directory/wilhelm

https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:XX.XXXXX/jdssv.v000.i00 *Accepted:* yyyy-mm-dd