Data Visualization

A Practical Introduction

Nith Kosal

Future Forum

2021 YRP Seminar, May 18, 2021

Goal of Today's Seminar

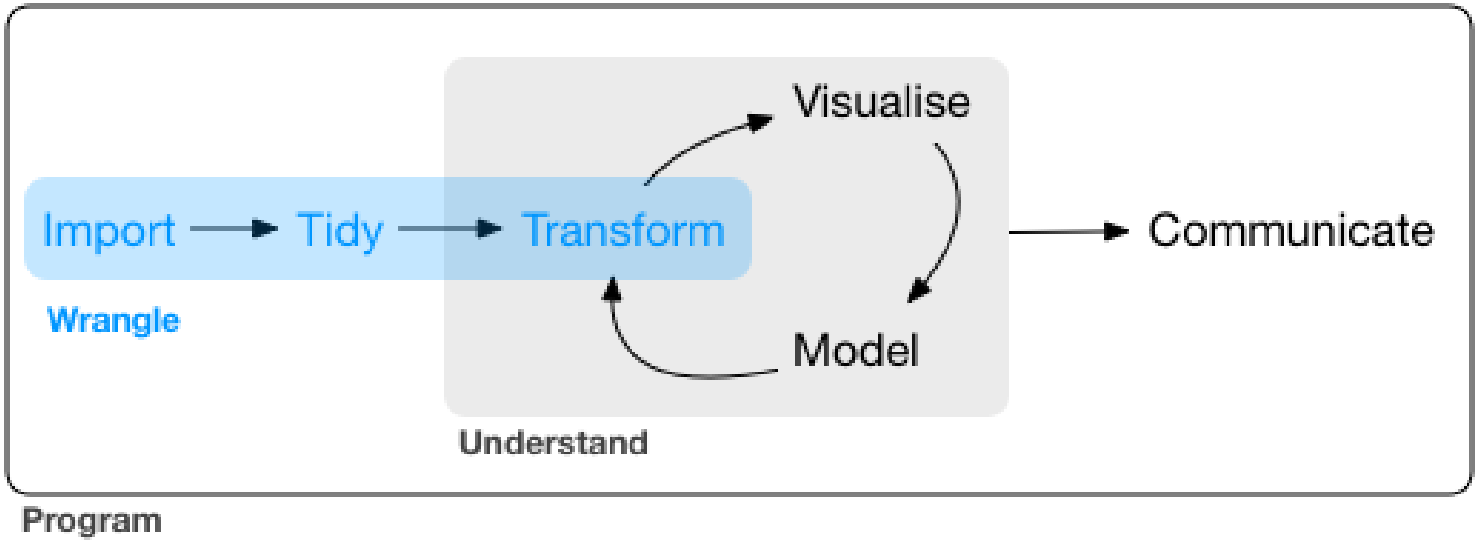
- Know the types of graphs we should avoid
- Learn about RStudio IDE (integrated development environment)
- Data visualization in R
- Data import and transformation
- Data export
- Discussion about your questions or any concerns about R or datasets

Remember: It will not work in a flow and structure, because we practice in the real example as a researcher have a dataset and applied it for a specific research study.

Seminar Materials

- You can code, data and slide from my repository at GitHub:
  `https://github.com/nithkosal/DataVisualization`
- Intermediate level: R for Data Science by Hadley Wickham and Garrett Grolemund. Available online:
  `https://r4ds.had.co.nz/`
- Advanced R by Hadley Wickham. Also available here:
  `https://adv-r.hadley.nz/`
- R Blog: `https://www.r-bloggers.com/`
- Ask questions or share your knowledge about programming:
  `https://stackoverflow.com/questions/tagged/r`

The Thing to know when you work with data



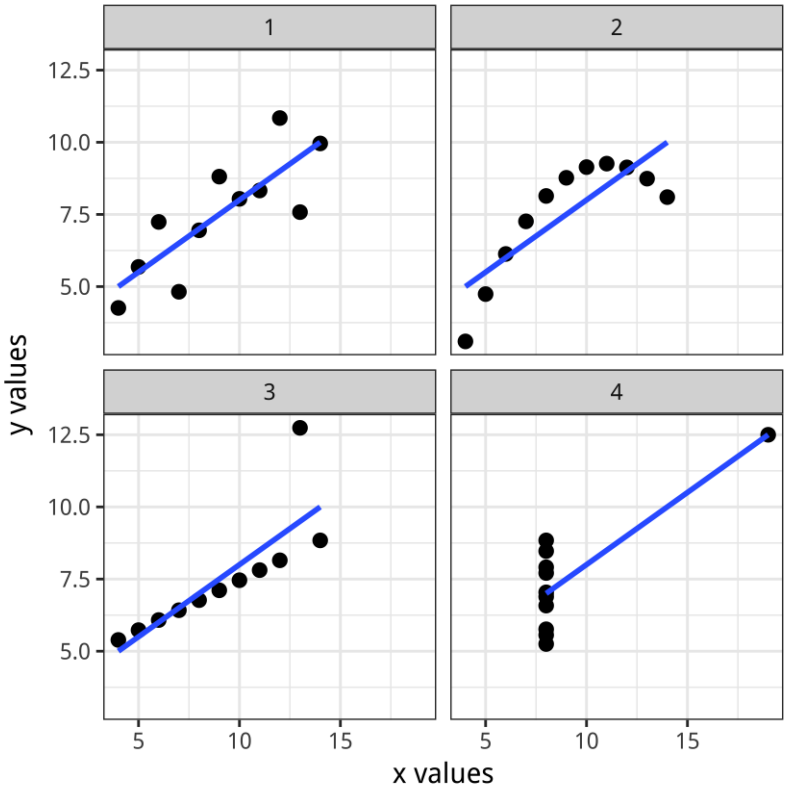
Source: Grolemund and Wickham (2016)

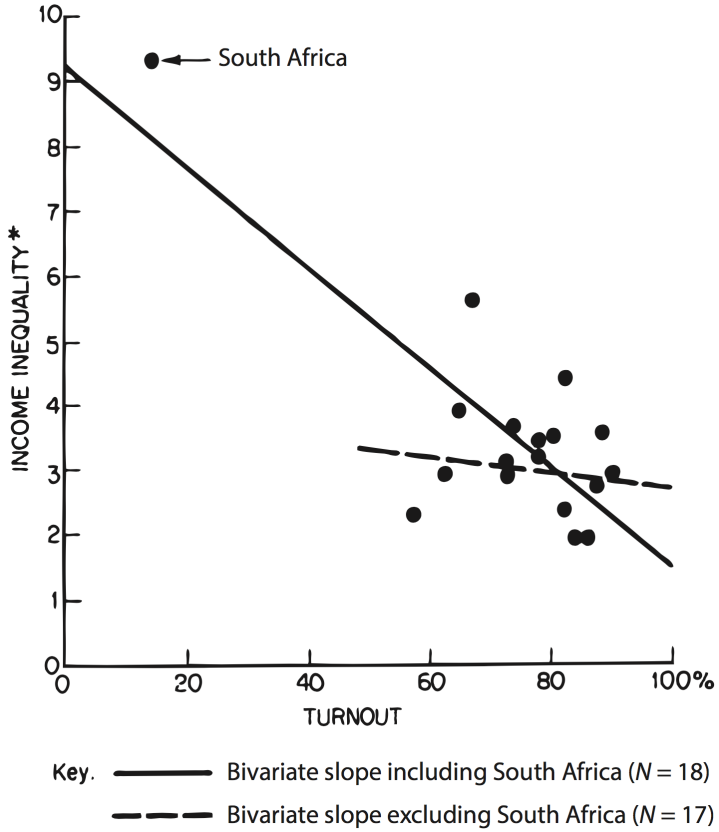
- Tidying your data means storing it in a consistent form that matches the semantics of the dataset with the way it is stored.
- Together, tidying and transforming, we called *wrangling*

What does data visualization mean?

- A reporting tool: the representation of information in the form of a chart, diagram, picture, etc.
- It is a part of *hypothesis generation* or *data exploration*.
  - A precise mathematical model to generate falsifiable predictions.
  - Only use an observation once to confirm a hypothesis.
- Modelling is an important part of the exploratory process.

Now, We Look at Some Data Visualizations

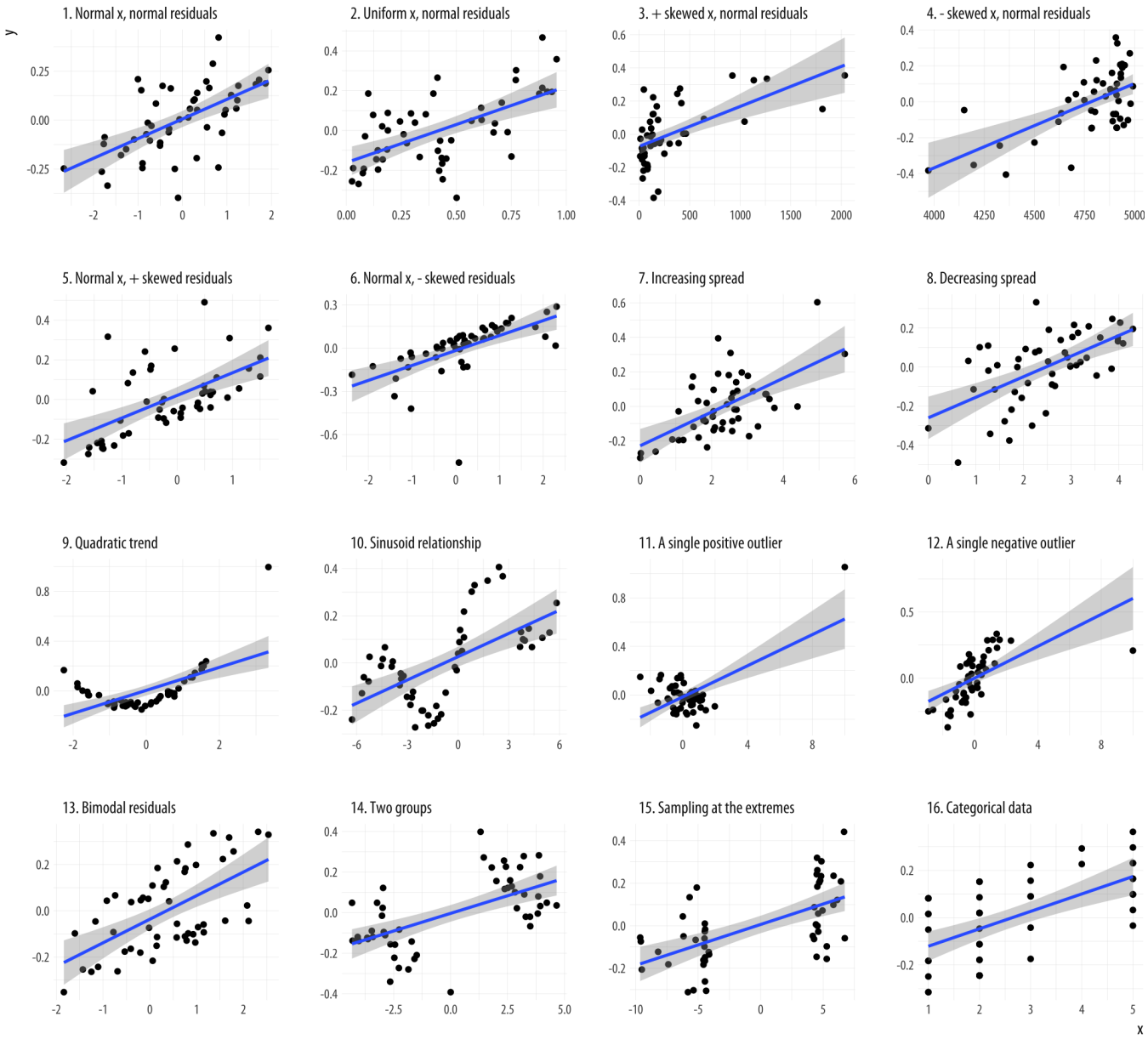




Source: Chatterjee and Firat (2007), Jackman (1980)

- Scatterplots are the workhorse of data visualization in social science.
- A scatterplot shows the relationship between two quantities, such as height and weight, age and income, or time and unemployment.

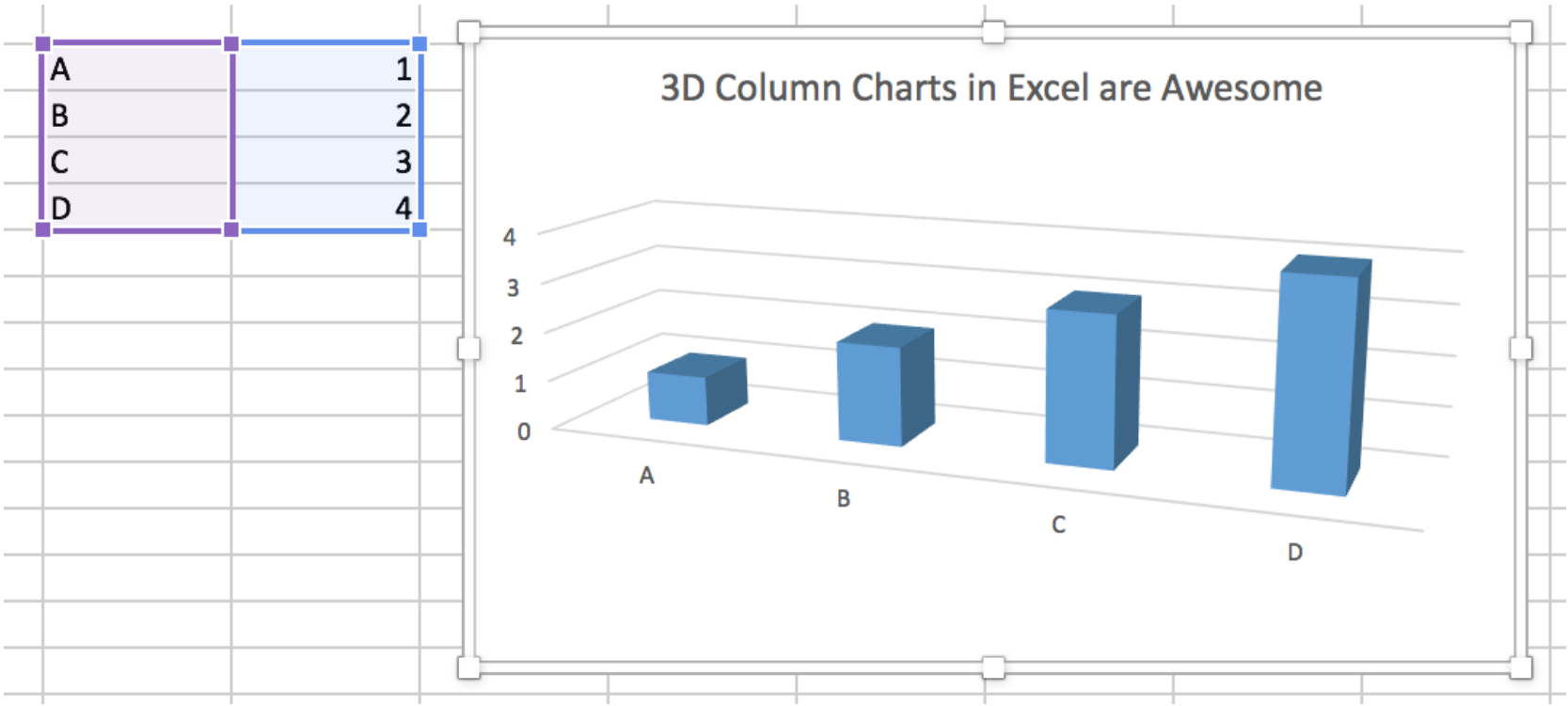
Within each panel, the correlation between the `x` and `y` variables is set to be `0.6`, a pretty good degree of association. But the actual distribution of points is created by a different process in each case.



Source: Jan Vanhove (2016)

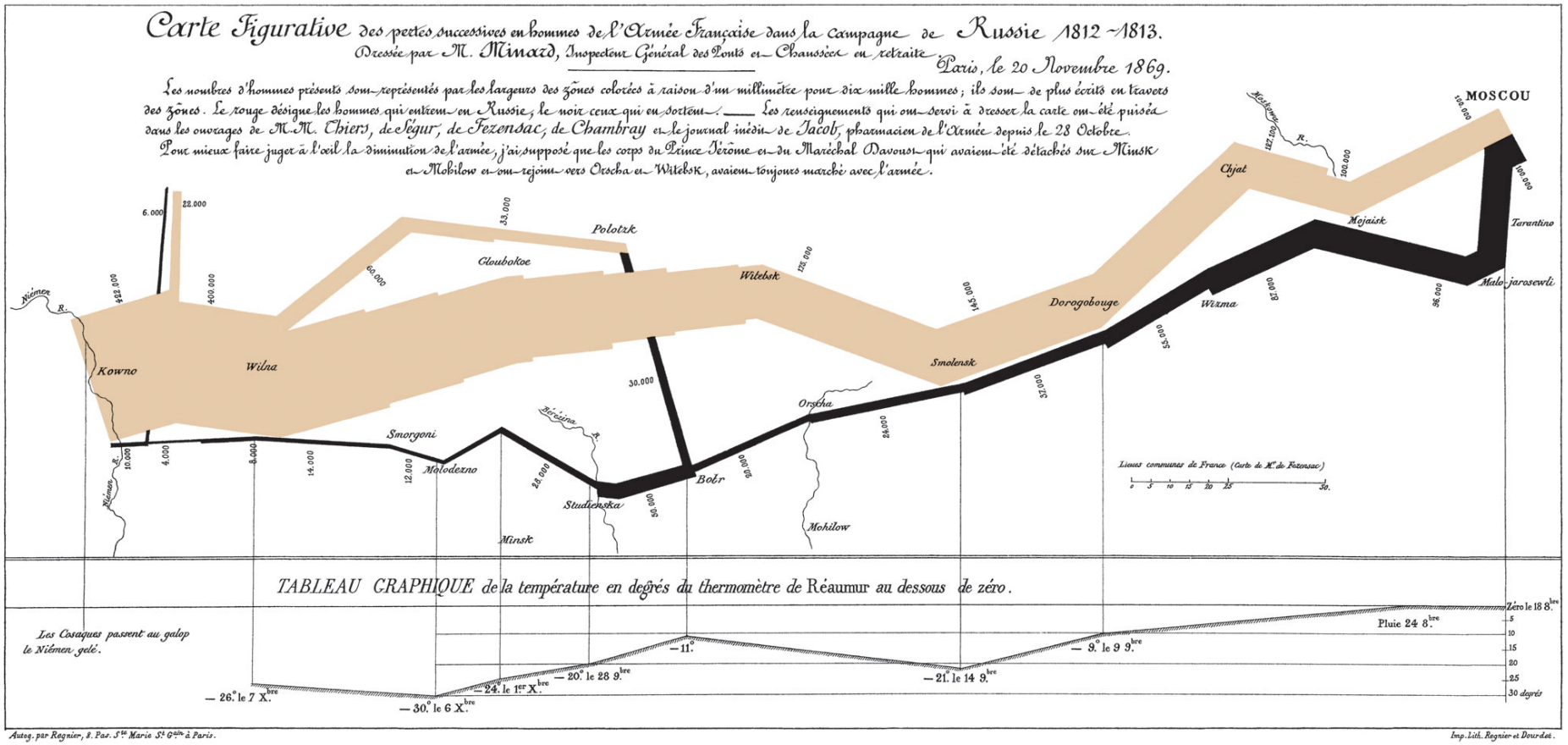
Source: Kieran Healy (2018)

The bars are hard to read and compare.



A 3-D Column Chart created in Microsoft Excel

- Charts like this are common in business presentations and popular journalism, and are also seen in academic journal articles. Here we seek to avoid too much junk by using Excel's default settings.

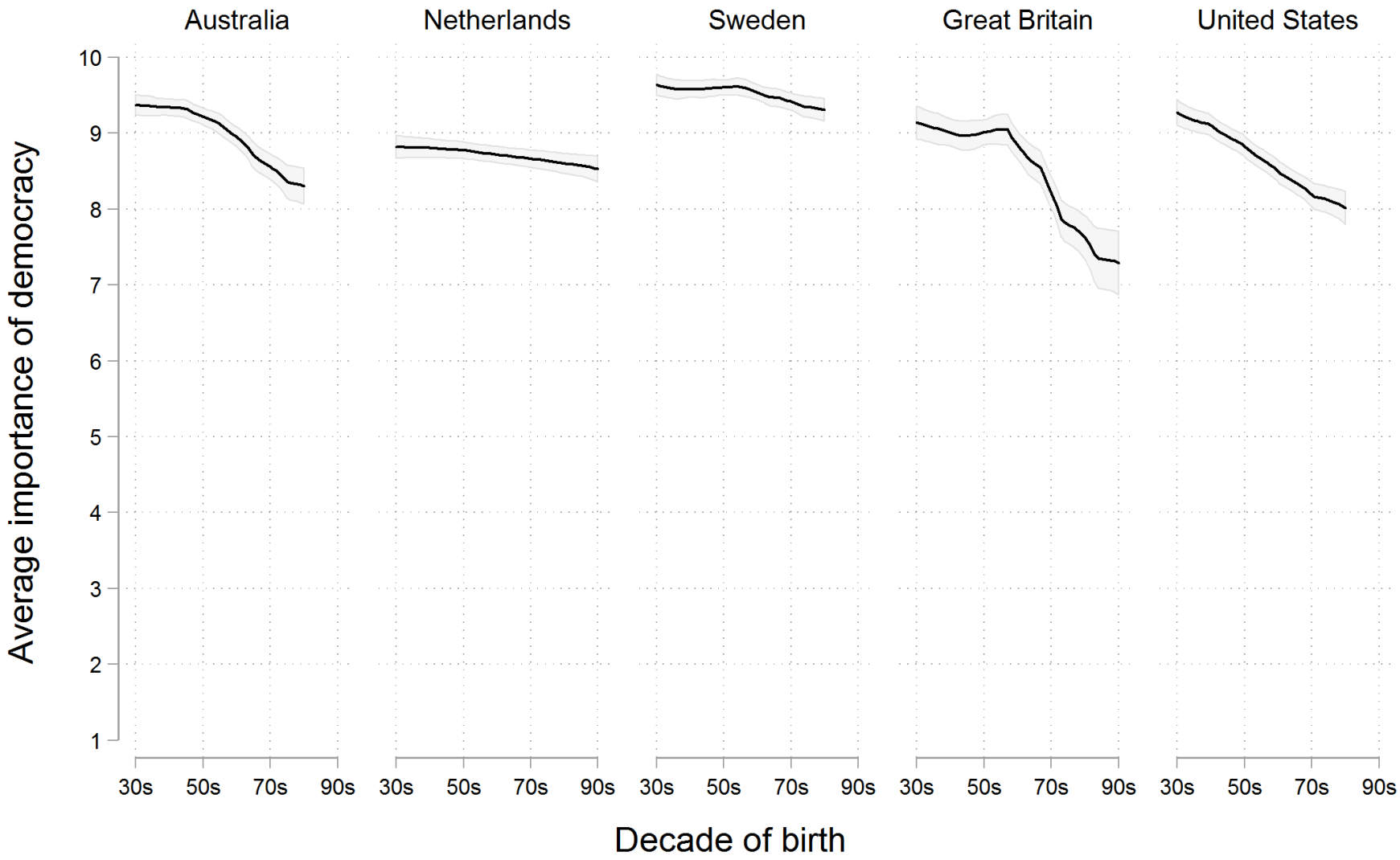


Source: Charles Joseph Minard — Minard's visualization of Napoleon's retreat from Moscow.

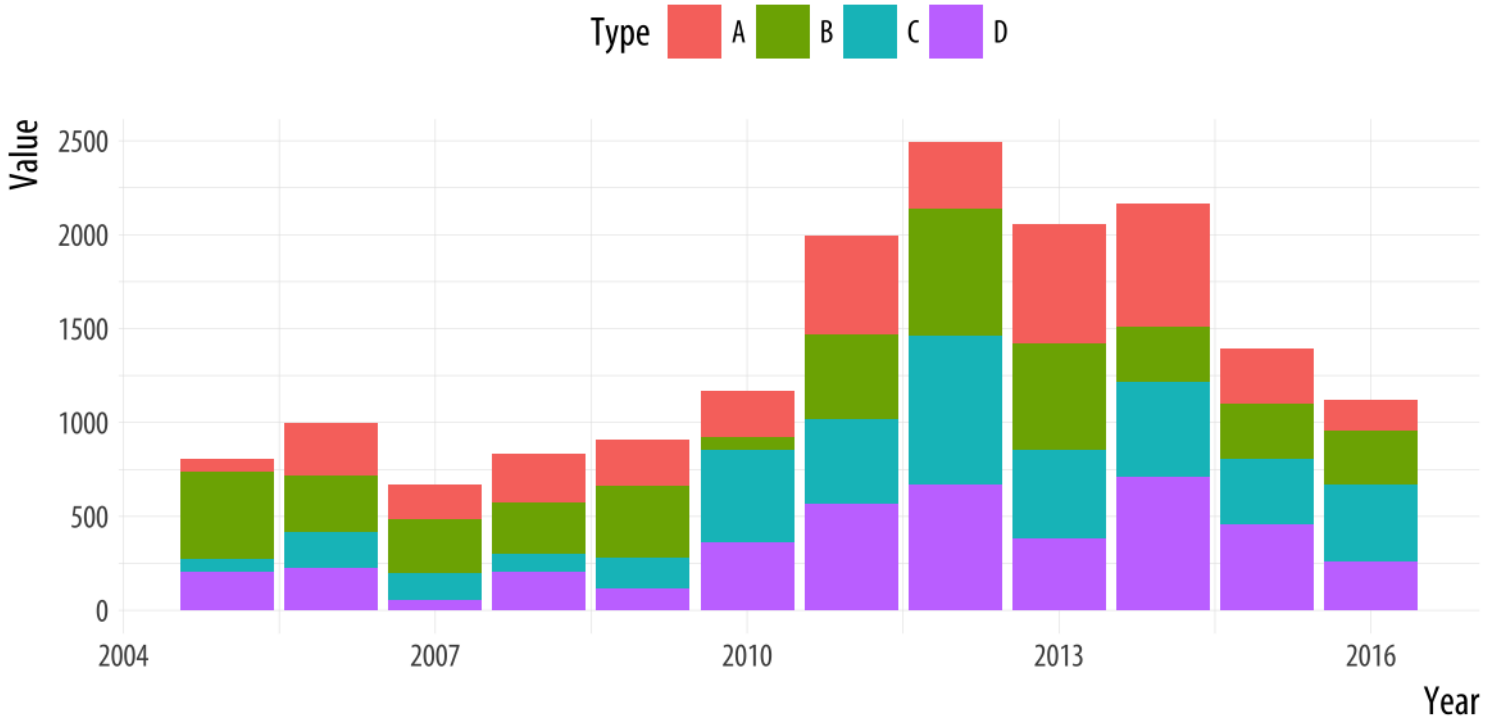
Edward Tufte's comment: "*Graphical excellence is the well-designed presentation of interesting data—a matter of substance, of statistics, and of design, it consists of complex ideas communicated with clarity, precision, and efficiency.*"

Source: 'Monstrous Costs' by Nigel Holmes

- Viewers do not find them more easily interpretable, but they do remember them more easily and also seem to find them more enjoyable to look at.
- Borkin et al. (2013) found that "*Infographic*" style graphs were more memorable than more standard statistical visualizations.

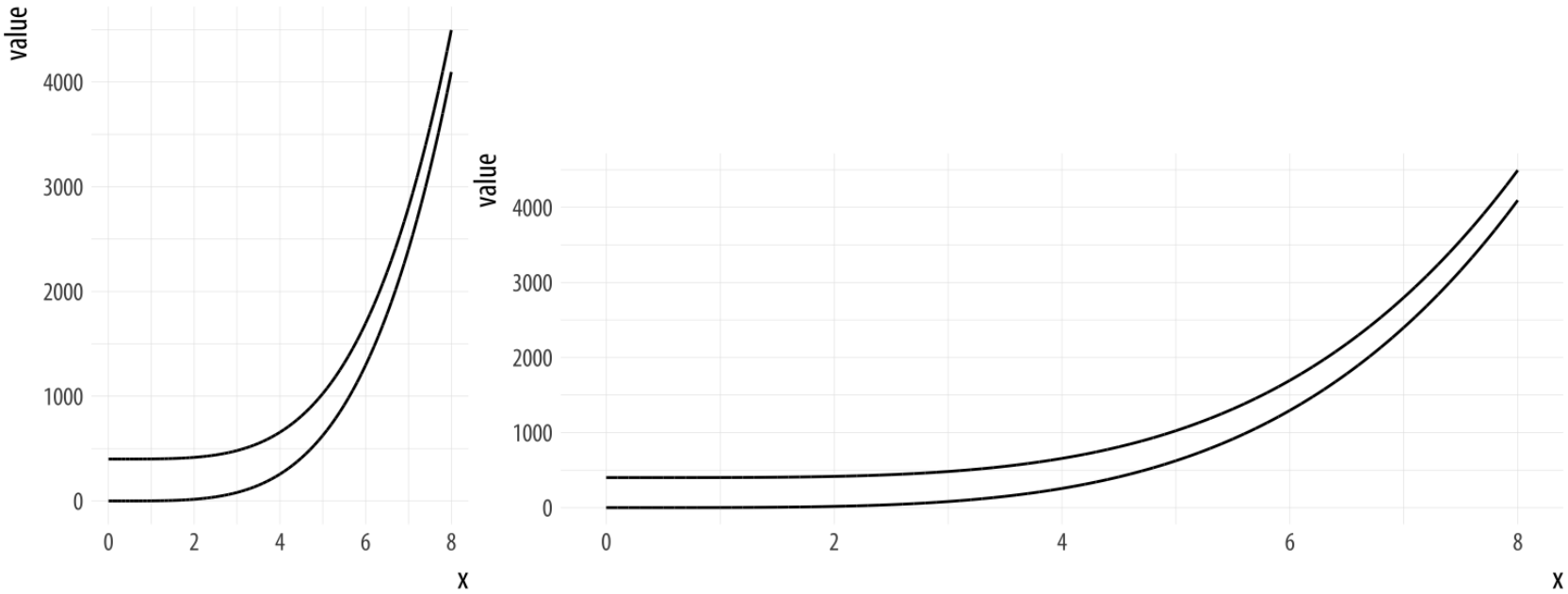


Graph by Erik Voeten, based on WVS 5



A junk-free plot that remains hard to interpret

- The bars show the total value, with subdivisions by the relative contribution of different categories to each year's observation.
- Charts like this are common when showing the absolute contribution of various products to total sales over time, or the number of different groups of people in a changing population.



Source: An example by William S. Cleveland

- In the left-side panel, the lines appear at first glance to be converging as the value of x increases.
- The data plotted in each panel are the same, however. The apparent convergence in the left panel is just a result of the aspect ratio of the figure.

In Short

These problems are not easily solved by the application of good taste, or by following a general rule to maximize the data-to-ink ratio, even though that is a good rule to follow. Instead, we need to know a little more about the role of perception in the interpretation of graphs.

Variable Types in the Statistical Software

1. `int` stands for intergers
2. `dbl` stands for doubles, or we call real numbers
3. `chr` stands for character vectors, and strings
4. `dttm` stands for date-times (a date + a time)

At the same time, there are three other common types of variables

1. `lgl` stands for logical, vectors that contain only `TRUE` or `FALSE`
2. `fctr` stands for factors, which R uses to represent categorical variables with fixed possible values.
3. `date` stands for dates.

Let us go practice!