DSA2101

Essential Data Analytics Tools: Data Visualization

Yuting Huang

Week 9 Principles of Visualization

Announcements on group project

The full guideline for the project is on Canvas → Assignments.

- Create your own group (up to five persons) on Canvas.
- Choose a TidyTuesday topic from the ones provided:
  - Wealth and income (2021-02-09)
  - UN Votes (2021-03-23)
  - Deforestation (2021-04-06)
  - Billboard Top 100 (2021-09-14)
- By the end of **this Friday March 17, 11:59pm**. Changing topic/group member after this point of time would likely be counterproductive.

TidyTuesday data

For each topic,

- A link to an original article (for context) and to the data set.
- Code to download the data.
  - From the `tidytuesdayR` package (a daily limit on the number of requests).
  - From Github directly.
- There may be some, but not all, code to clean the data.

Your task is to come up with **two** distinct and meaningful questions, answer each of them with two visualizations using `ggplot2`, and write-up your method and findings.

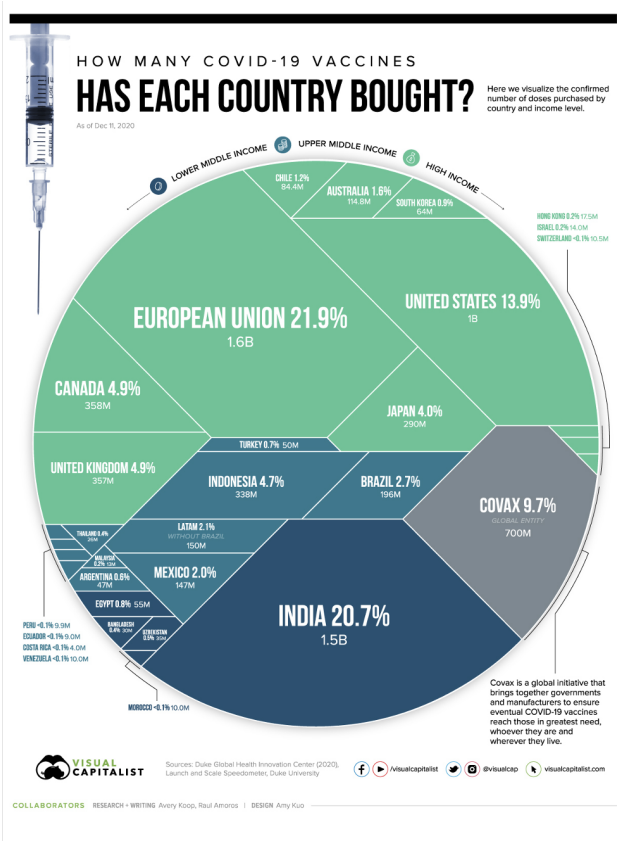
- The project itself is due on **Friday April 14, 11:59pm**.

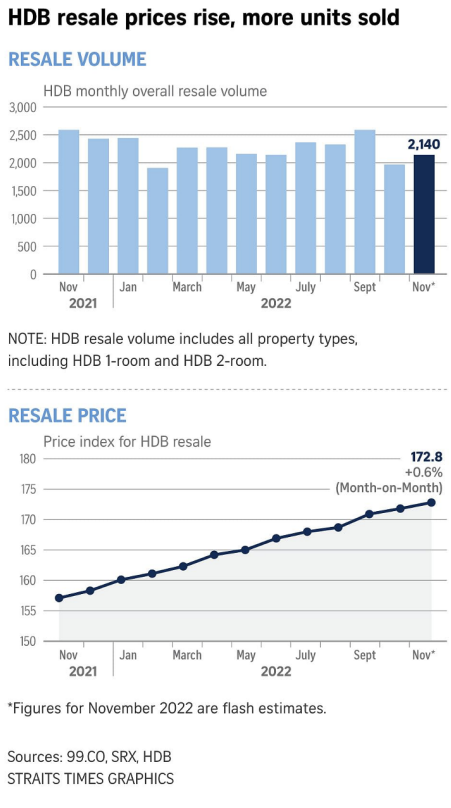
Workflow

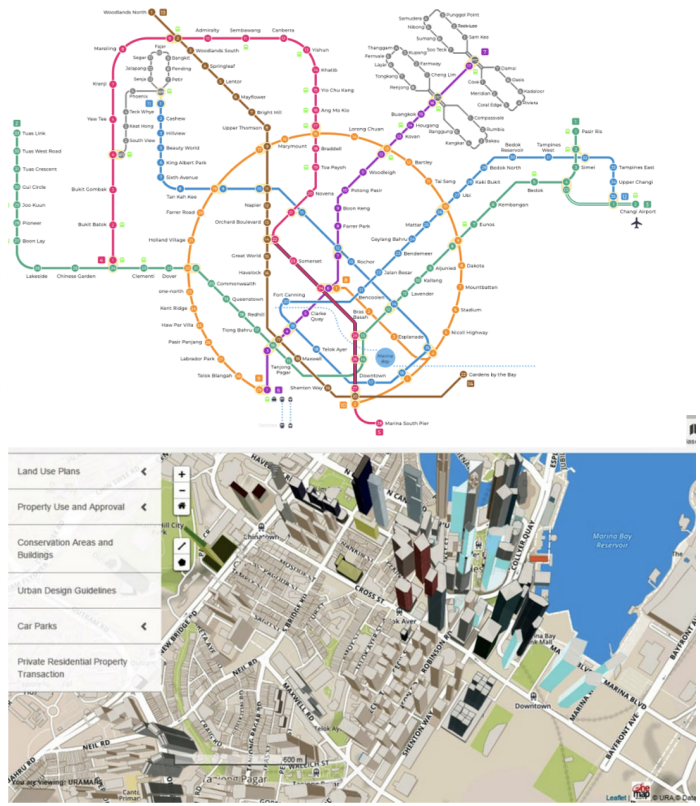
- **Create your group** and indicate your **topic choice**.
- **Prepare your data**: While most data sets on TidyTuesday are "tamed", they are not always tidy and clean! You may need to apply your data wrangling skills to turn it into a true tidy format.
  - **Always**, always manipulate your data in `R`. Document the transformations you have applied to your data in the write-up.
- **Explore your data**: Come up with interesting questions to learn from data.
  - Discuss your findings in the write-up.
  - If you choose to recreate a graph from others, comment on the quality of the plot and/or potential problems. Remember to quote your sources.
- **Finalize and submit your write-up**.

Learning to visualize

Big data, small data... Data are everywhere.







Introduction

Despite the inherent subjectivity of beauty, there are several recurring patterns and characteristics we find beautiful and pleasant. Among these recurring patterns:

- Symmetry
- Proportions (e.g., the golden ratio `1:1.618`)

Subjectivity notwithstanding, there are some techniques and principles we can follow in order to make our visualizations more effective.

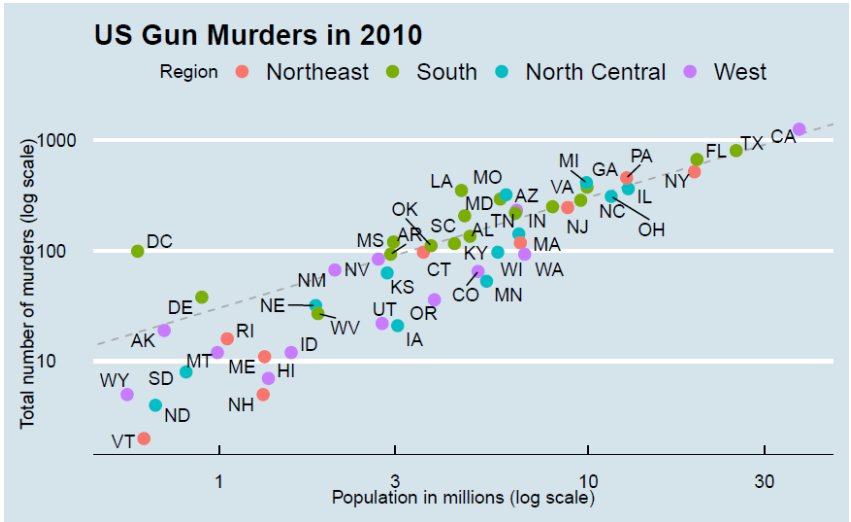
Introduction

This week, we outline a few principles of data visualization. This is by no means an exhaustive list, but rather meant as the starting point for your own list of things to keep in mind when designing visuals.

What we learn about in this class will not cover all possible types of graphics. Instead, we will try to

- Show some good graphics
- Show some bad graphics
- List some principles of data graphic design

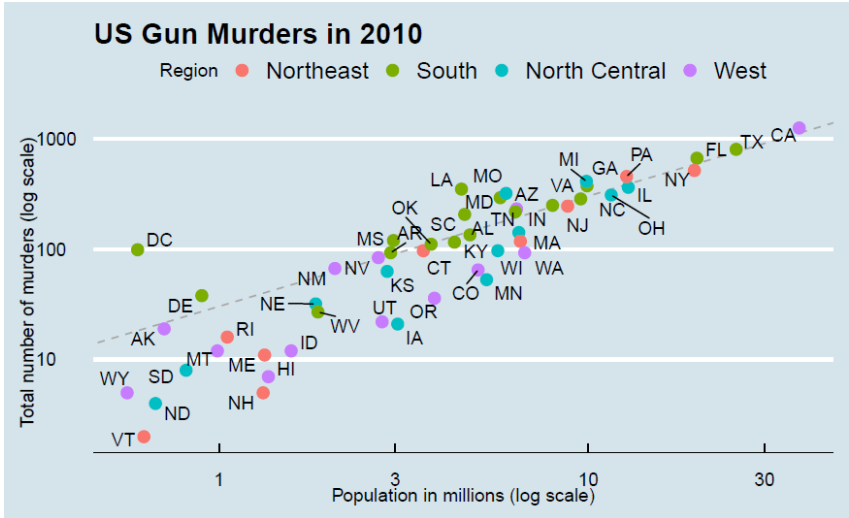
Components of a graph



What do you learn from staring at this table?

- How quickly can you determine which states have the largest/smallest populations?
- Any relationship between population size and total murders?

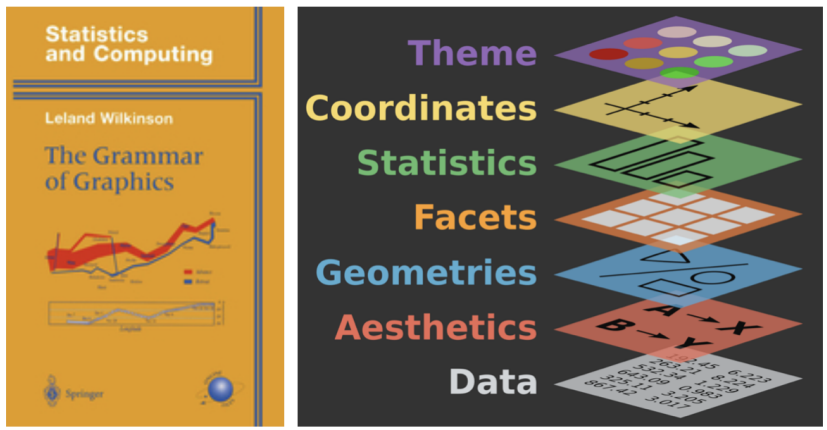
Components of a graph



- Each **point** indicates the value of population size and total murders.
- A state falling on the dashed grey **line** has the same murder rate as the US average.
- Four regions are denoted with **color**, which depicts how most southern states have murder rates above the average.

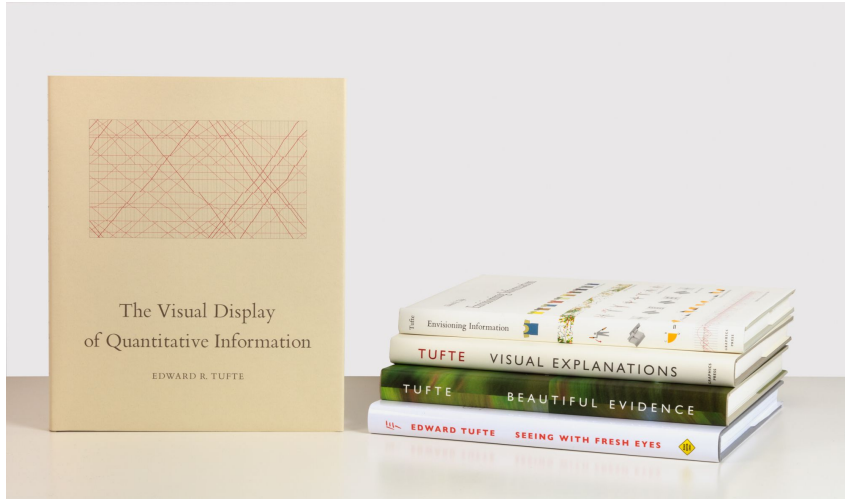
The Grammar of Graphics

Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color.



The `ggplot2` package draws on the grammar of graphics, a concept developed by Leland Wilkinson. Today we walk through the main ideas behind this complex book.

Edward Tufte



- Most of our material is taken from books by Edward Tufte.
- You might not agree with some/most/all of them, but principles in these books are closely adhered to by the `ggplot2` authors.
- The principles serve as a good guide when we need to draw up our own graphics.

Some loose definitions

- A **visualization** is any kind of visual representation of information designed to enable communication, analysis, discovery, and exploration.
- A **graph** or a **chart** is a display in which data are encoded with symbols that have different shapes, colors, or proportions.
- A **map** is a depiction of a geographical area or a representation of data that pertains to that data.
- An **infographic** is a multi-section visual representation of information intended to communicate one or more specific messages.

What makes a good graph?

We should have certain expectations of a graph.

Graphical displays should:

- **Show** the **data**.
- Induce the viewer to **think** about the **substance** rather than the methodology, graphic design, or the software used, etc.
- **Avoid distorting** what the data have to say.
- Present **many numbers** in a small space.

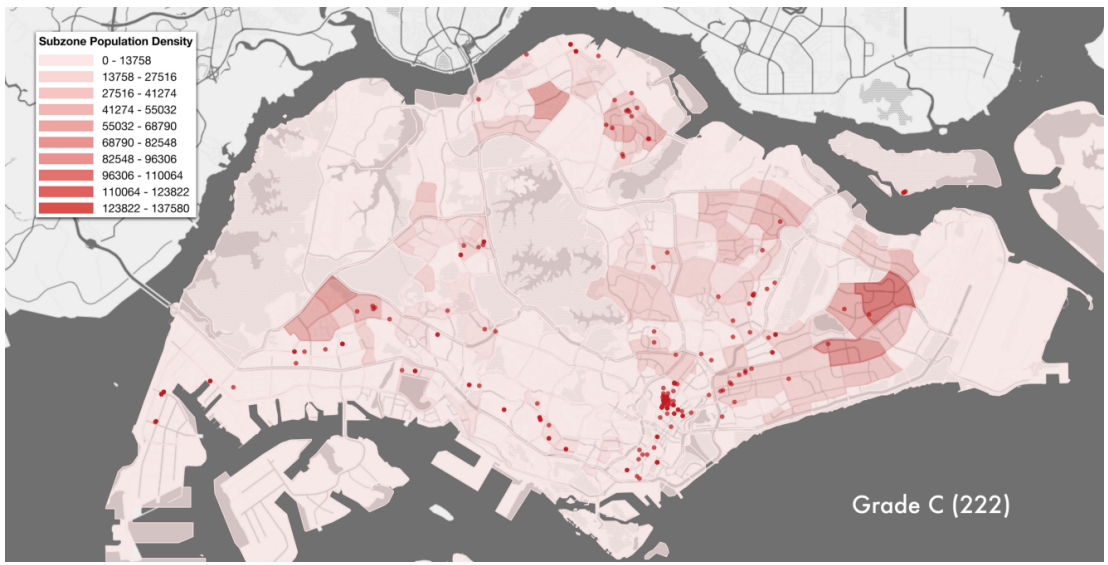
What makes a good graph?

- Make large data sets **coherent**.
- Encourage the eye to **compare** different pieces of data.
- Reveal the data at **several levels of detail**, from a broad overview to the fine structure.
- Serve a reasonably clear **purpose**: description, exploration, tabulation, or decoration.
- Be closely **integrated** with the statistical and **verbal descriptions** of a data set.

Types of graphical displays

We can broadly categorize graphics into the following groups:

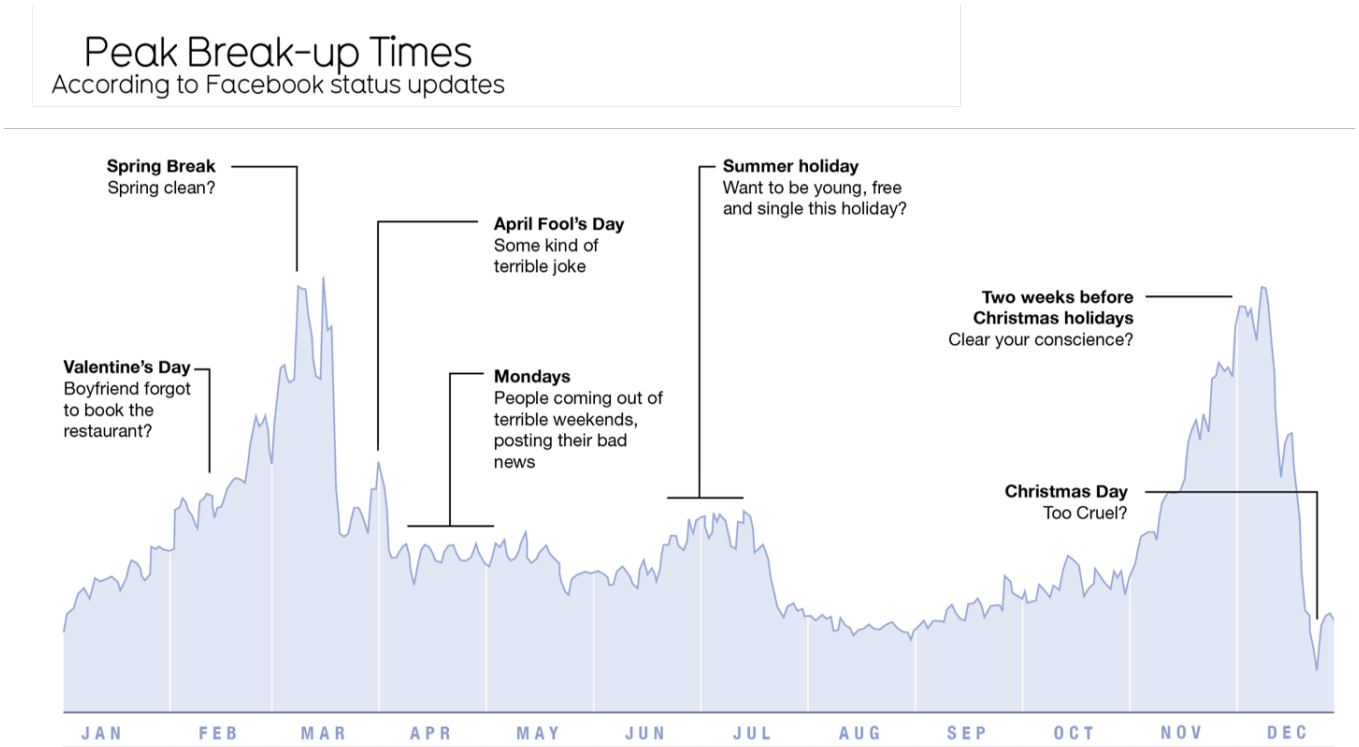
1. **Data maps**: superimpose attributes onto geographical data, allowing us to study clusters, hotspots, etc.



Source: National Environment Agency (NEA)

Types of graphical displays

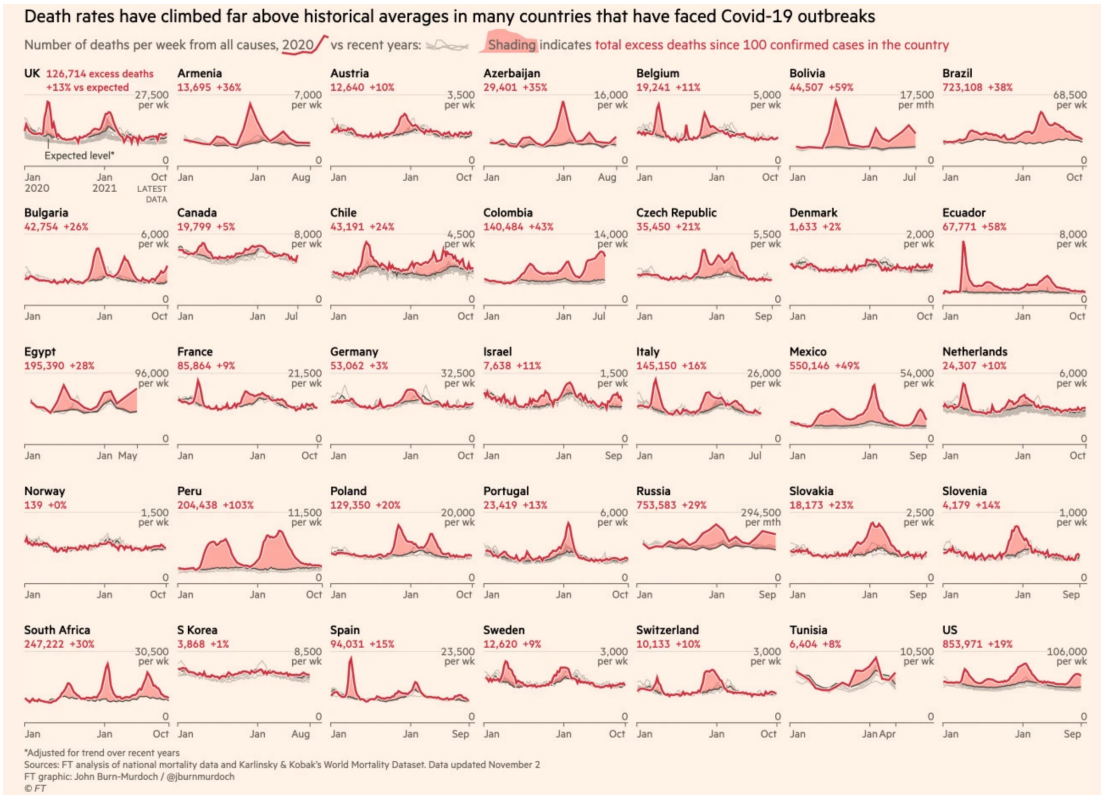
2. **Time series**: probably the most frequently used form of graphic design.



Source: David McCandless and Lee Bryon

Types of graphical displays

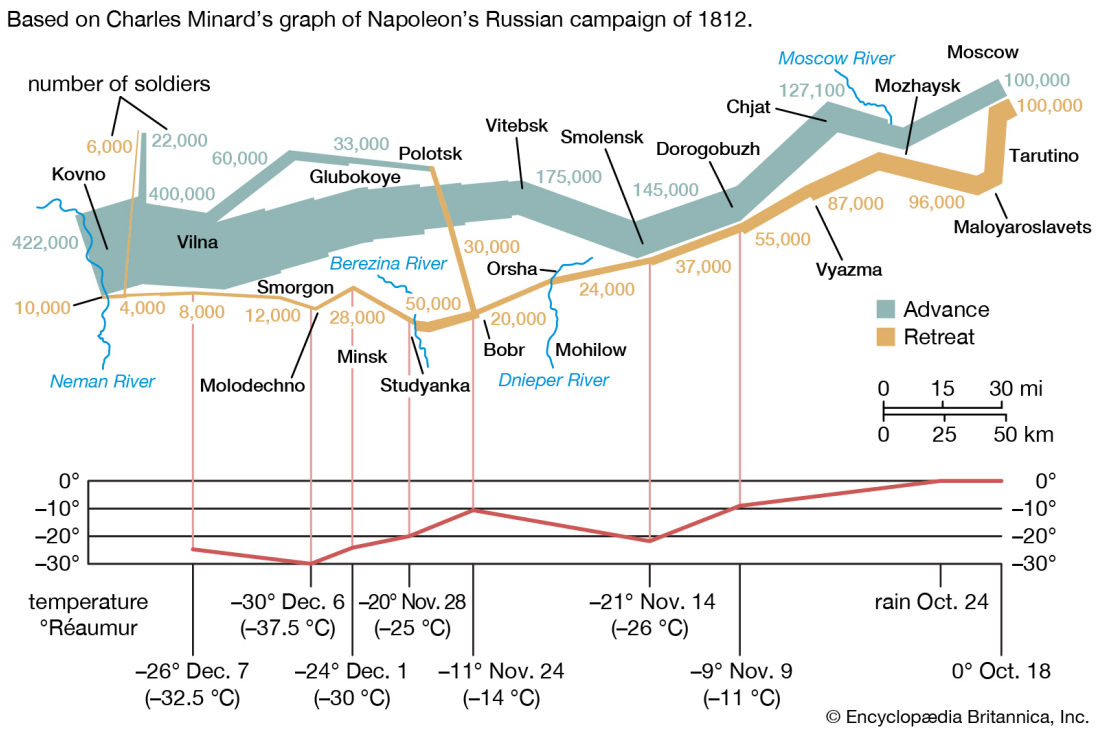
2. Multiple **time series**: co-current series across categories.



Source: Financial Times, Dec 20, 2021.

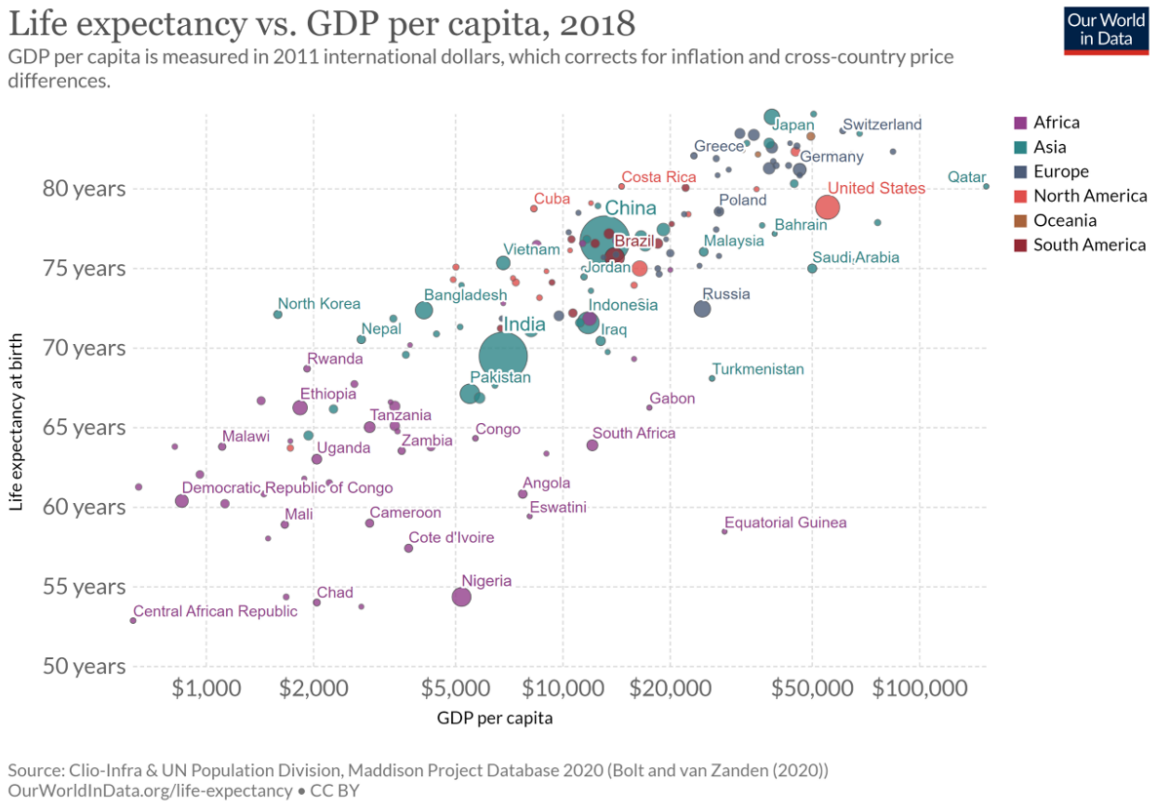
Types of graphical displays

3. **Narrative graphics of space and time**: add spatial dimensions to time series displays.



Types of graphical displays

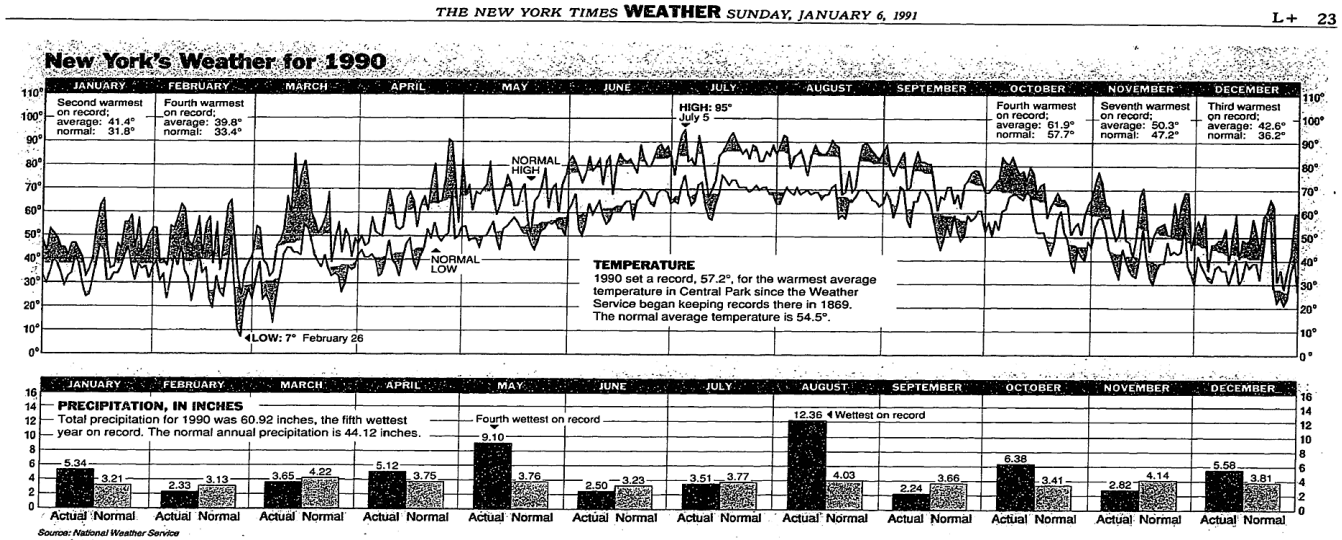
4. **Relational data**: put one one variable on the x-axis, and another on the y-axis.



Good graphics

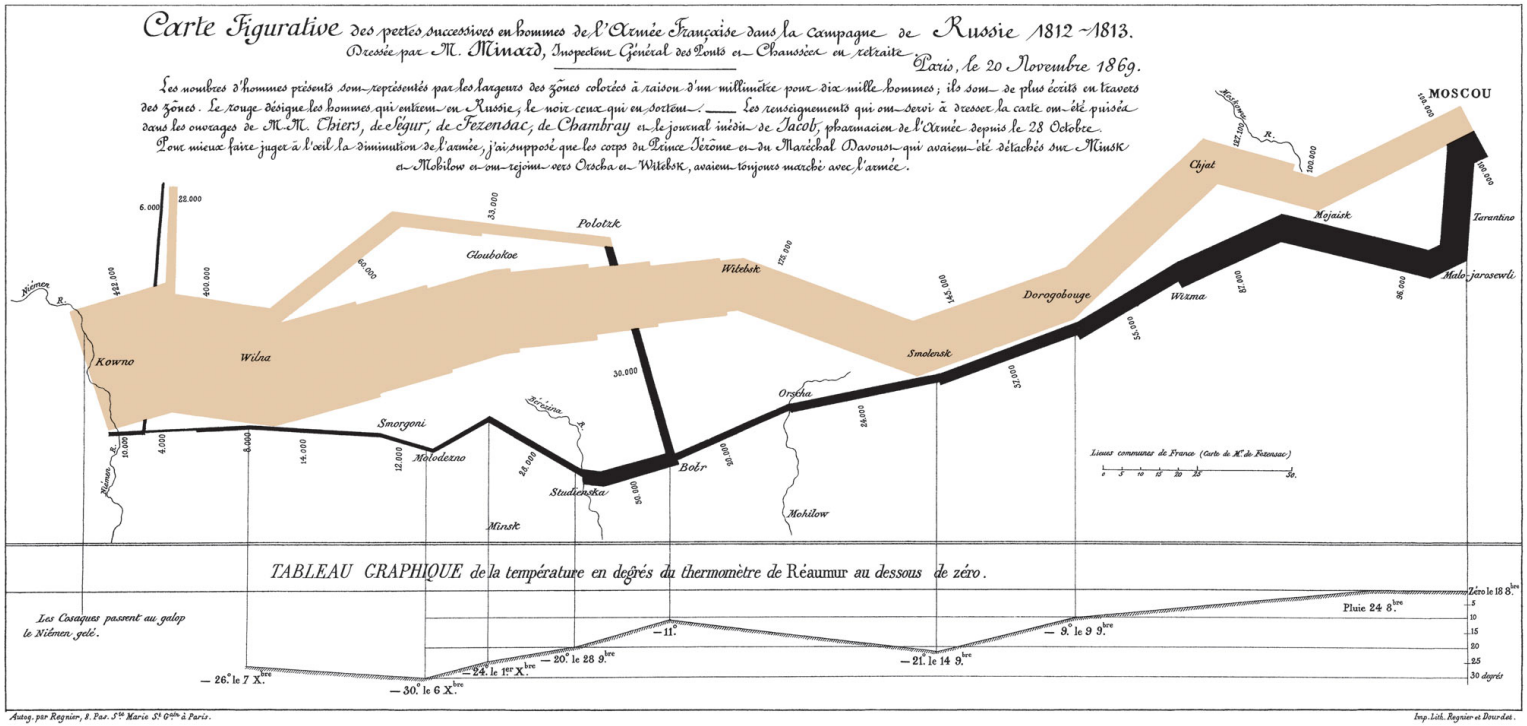
Let's take a look at some examples of good graphics

New York weather



- **Time series** graphic conveys different scales of information
  - How the temperature varies over the year
  - How rainfall varies from month to month
- One any day, we can compare the weather in 1990 with typical weather in New York, as well as the extreme weather on that day.
- Encourages comparisons within and between the time series presented.

Napoleon's army in Russia

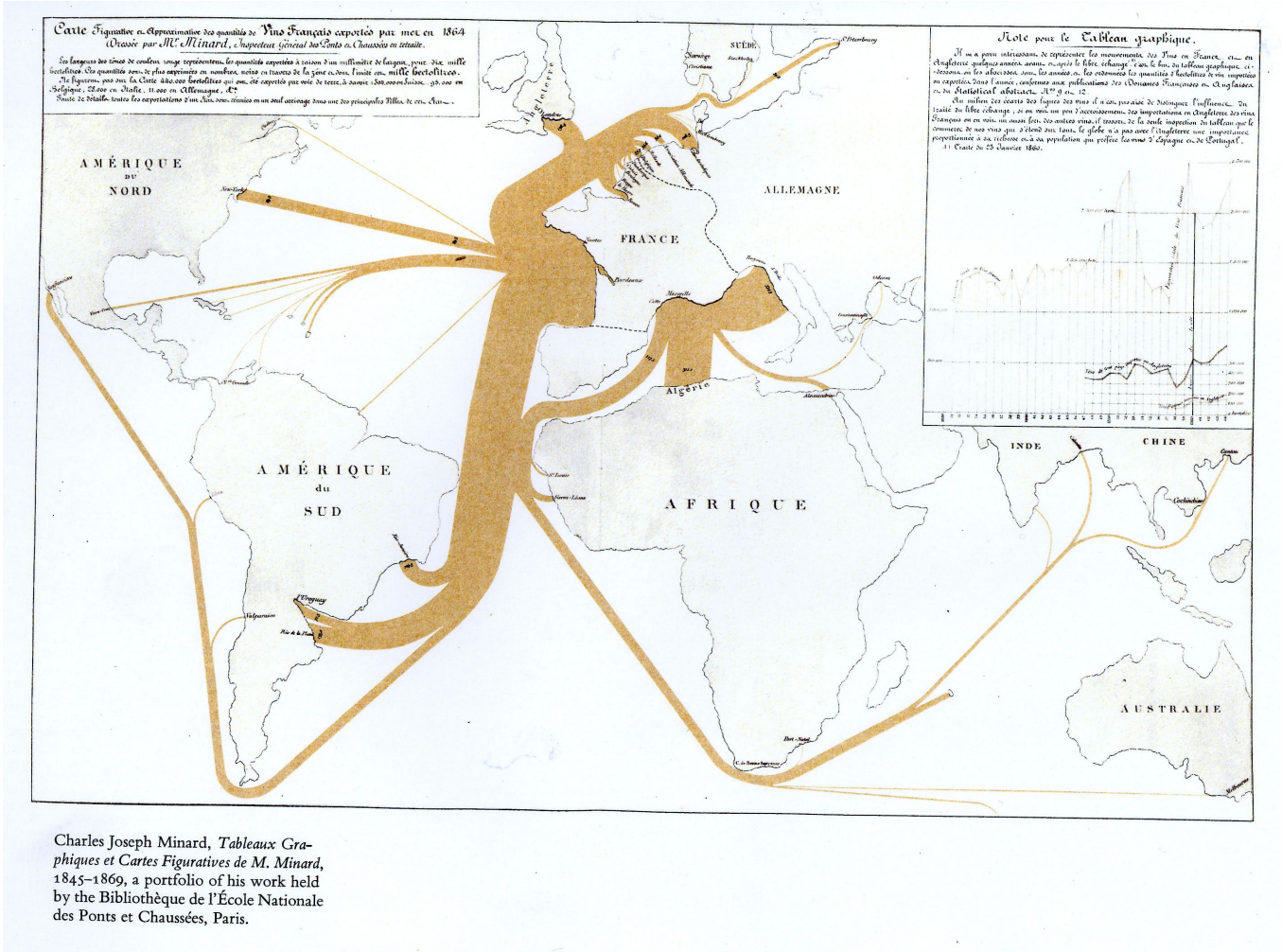


- Six variables: size of army, location on a map, direction of movement, temperature, and date.

Napoleon's army in Russia

- The graphic shows the losses suffered by Napoloan's Russian campaign in 1812.
- At the beginning, he has 422,000 soldiers, indicated by the band.
- When he reaches Moscow, there are only 100,000.
- On the way back, the numbers dwindle even more. Close to home, crossing one particular river wipes out almost half of what remains.

Wine exports from France in 1864



Charles Joseph Minard, *Tableaux Graphiques et Cartes Figuratives de M. Minard*, 1845–1869, a portfolio of his work held by the Bibliothèque de l'École Nationale des Ponts et Chaussées, Paris.

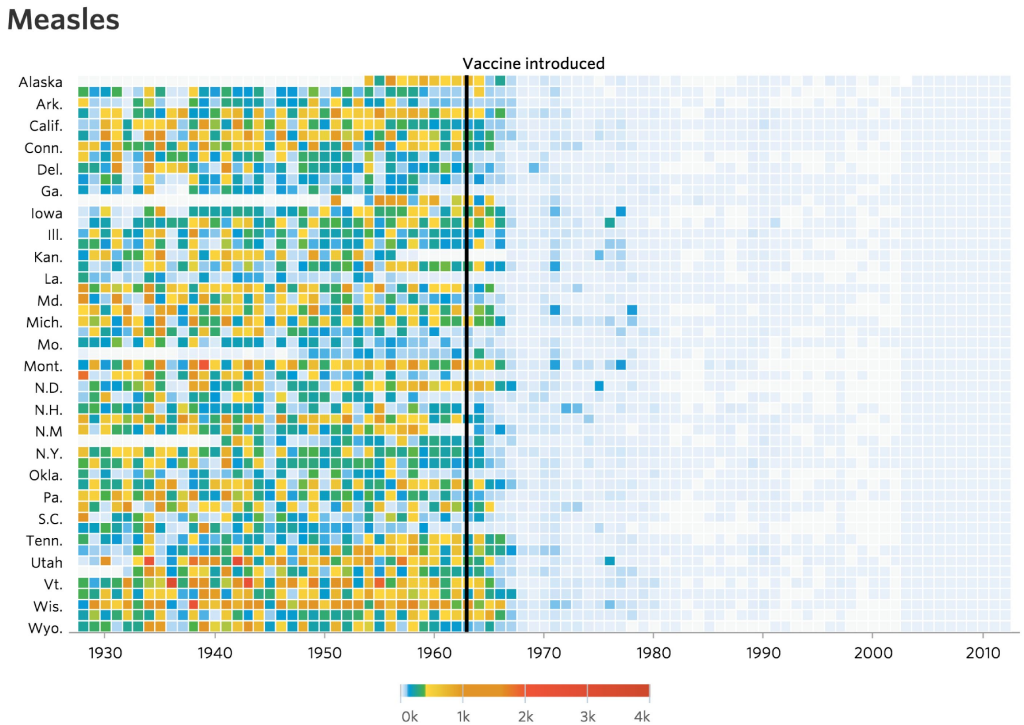
Wine exports from France in 1864

Pay attention to:

- The muted colors for the country maps and the darker color for the passage of wine.
- The width of data lines are proportional to the amount of wine sold to the country.
- The larger bars are annotated with exact amount. So are some prominent cities such as Calcutta.

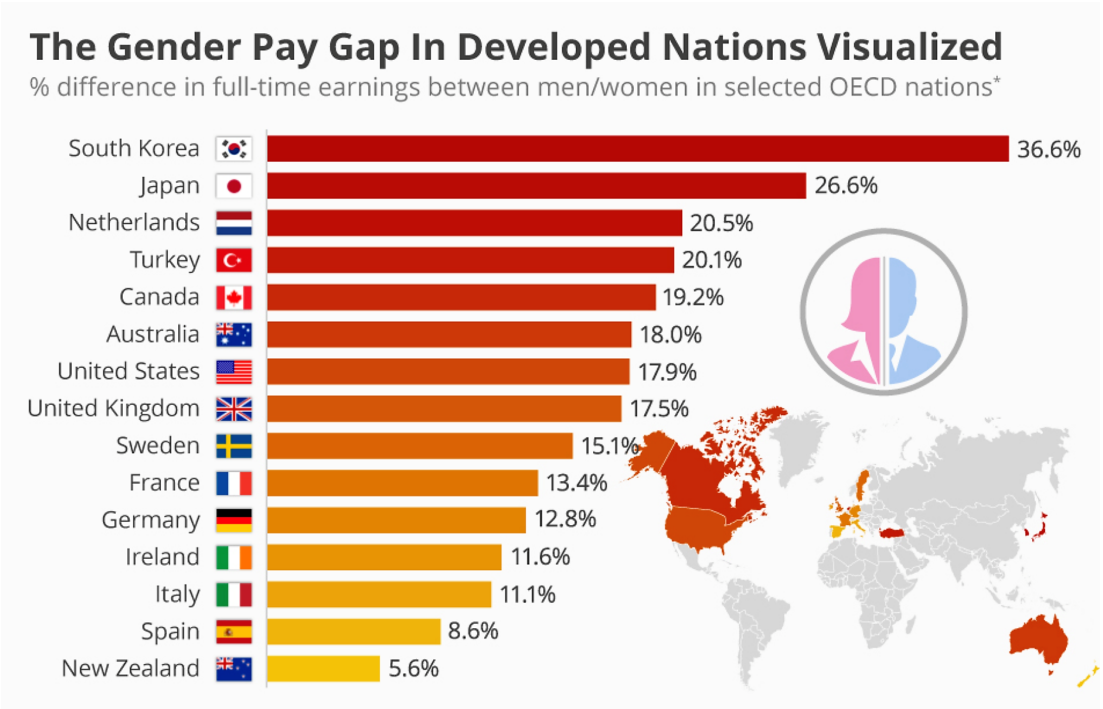
Measles

The number of infected people, measured over 70 years across 50 states and the District of Columbia, declined after vaccines were introduced.



Source: Wall Street Journal, February 11, 2015

Gender gap in earnings



Source: OECD

Principles of graphical excellence

- Well-designed presentation of interesting data
- Consist of complex ideas communicated with clarity, precision, and efficiency
- Give the viewer the greatest number of ideas in the shortest time
- Nearly always multivariate
- Require telling the truth about the data

Bad graphics

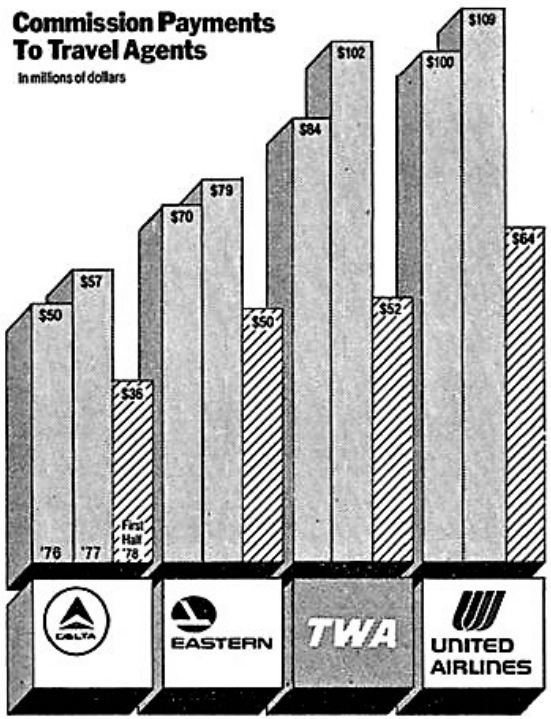
Until recently, statistical graphics were regarded with great suspicion.

- This is because many graphic makers worked with the assumption that graphics had to be "alive" or to be "communicatively dynamic".
- If not, the belief was that the audience would fall asleep.
- This belief led to deceptive graphics, that in turn led to more suspicion of graphics.

Bad graphics

Let us take a look at some bad examples.

Inconsistent basis of comparison



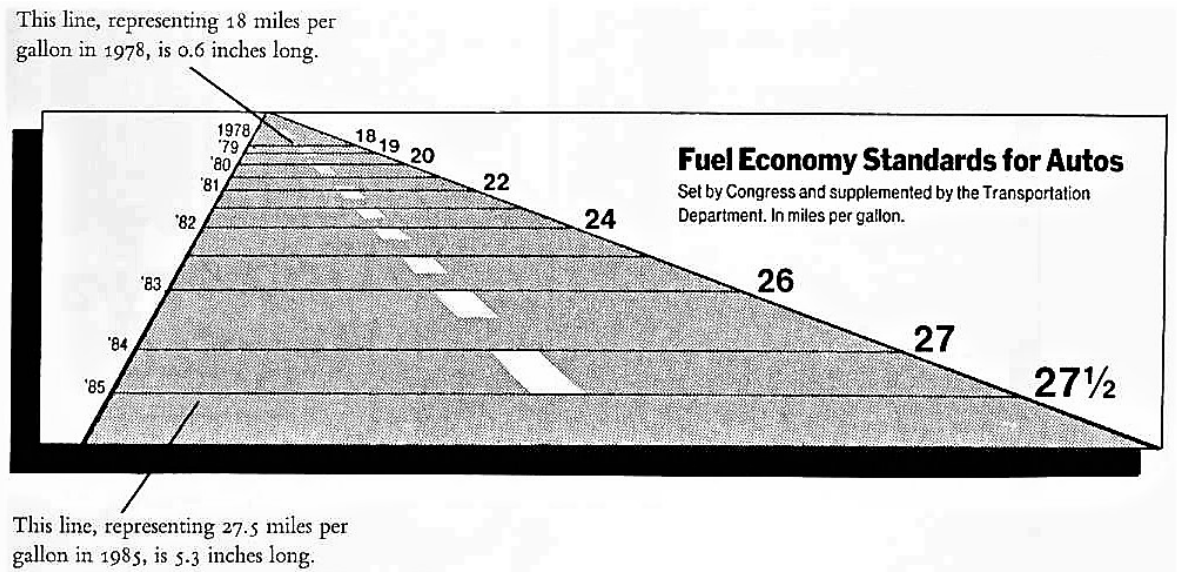
Source: The New York Times, August 8, 1978

- What is your first impression of the airlines' relative success in 1978?
- A pseudo-decline was created by comparing six months' worth of payments in 1978 to a full year's worth in 1976 and 1977, with the lie repeated four times.

Distortion in graphics

When we present a graphic, it is our duty to ensure that the visual representation of the data is **consistent** with the numerical representation.

The following graphic contains distorted data.



Distortion in graphics

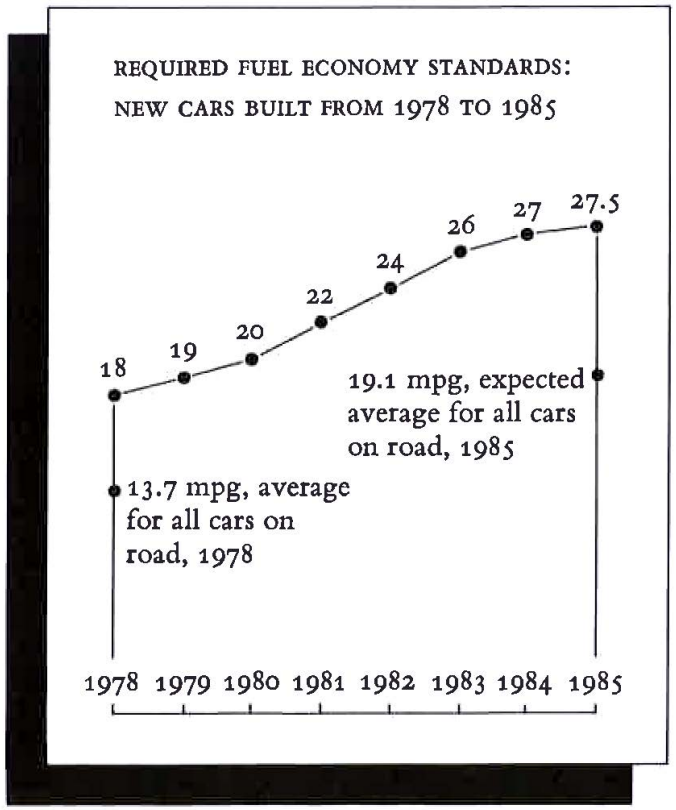
Distortions:

- Numeric representation: 18 miles per gallon (mpg) in 1978 $\Rightarrow$ 27.5 mpg in 1985 (an increase of 53%).
- Visual representation: the length of the lines represents an increase of 783%.
- The distortion gives the appearance that fuel efficiency has improved more dramatically than it actually has.

Additional confounding factors:

- Usually the future is in front of us
- Dates remain same font size while the fuel factors increase

Distortion in graphics (honest portrayal)



- The graph on the left is accurate.
- It even provides the context of actual cars' efficiency.

Distortion in graphics

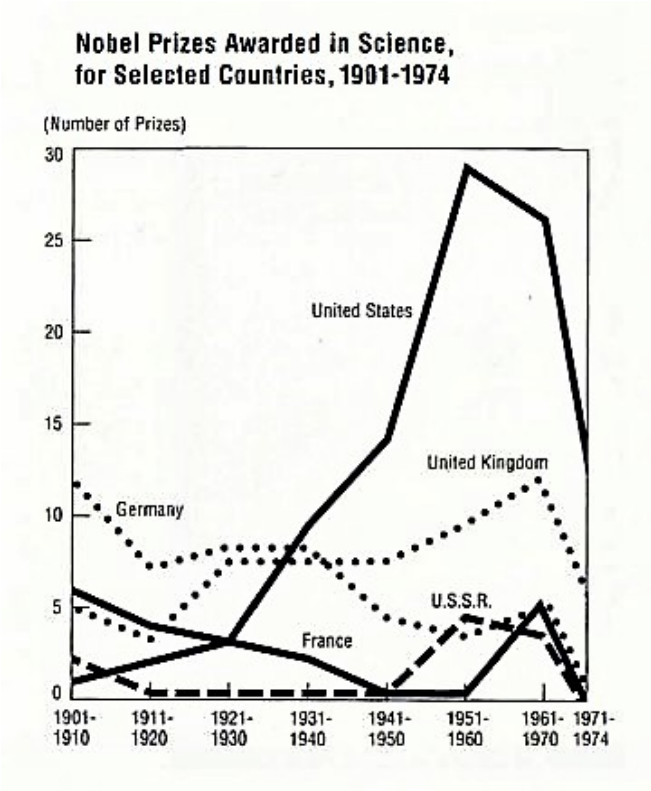
Each part of a graph generates visual expectations about its other parts.

- That is, a length of 1cm on the axis on the left side of the graph should represent the same distance on the right side.
- If it does not, then we have design variation.
- This is bad, since the viewers can be easily deceived if they do not catch the variation.

Extrapolation

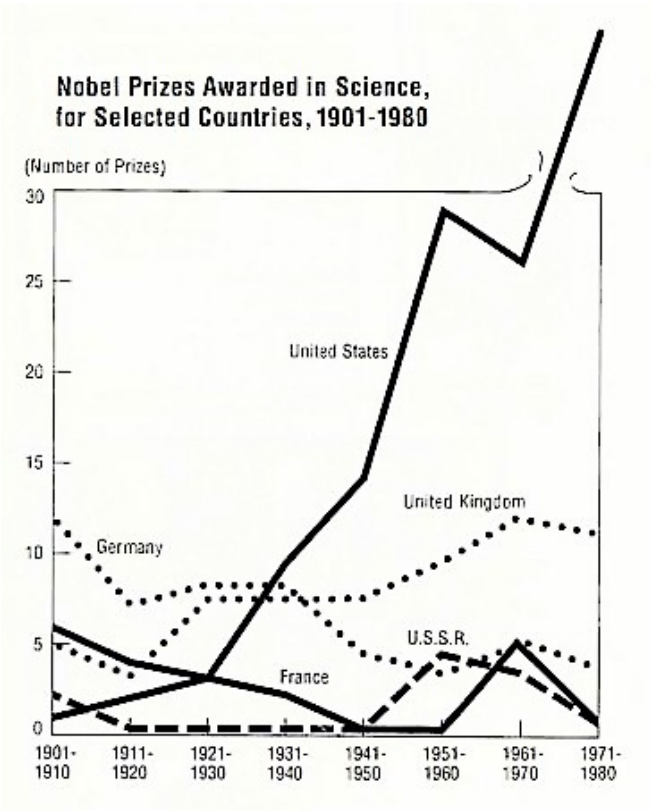
A graphic generates visual expectations – deception can result from incorrect extrapolation of visual expectations.

- The graph depicts the number of Nobel Prizes awarded to selected countries.
- Notice the decline in prizes won by United States in the most recent period.
- Is it true?



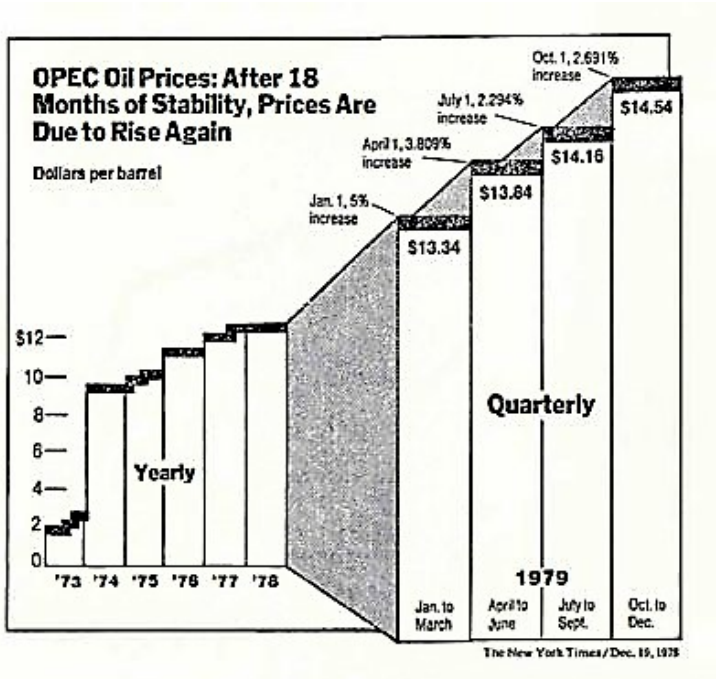
Extrapolation (revised)

- Look closer to the x-axis.
- Each data point had been representing a decade until the final one, which only represents 4 years.
- It shouldn't be surprising that the count for the incomplete decade is low, but graph gives the impression that there is a sharp drop in prizes.

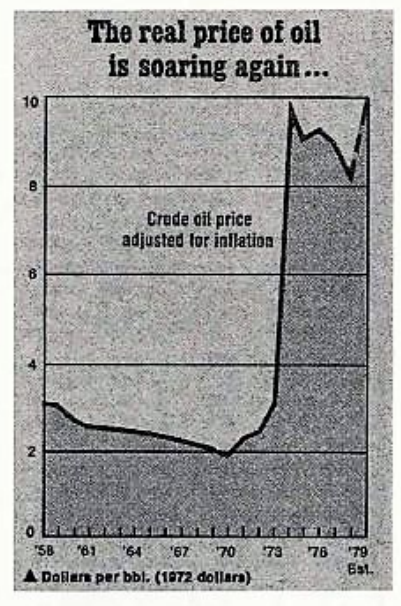


Design variation vs. data variation

- Left: With both x and y-axis shifting in scales, the distortion is multiplicative.
- Right: Revised and adjusted for inflation.

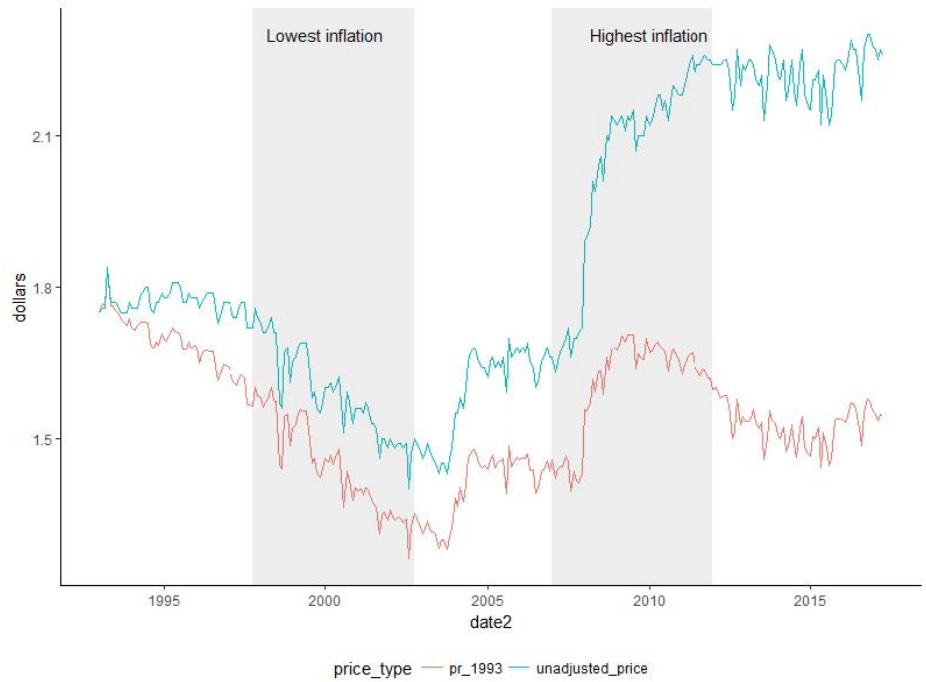


Business Week, April 9, 1979, p. 99.



Time-series monetary measurements

- On the right is a chart from a previous class.
- We used data from MTI to adjust for price of instant noodles in Singapore.
- Failure to account for the high inflation in the last 10 years would have depicted a worrying trend of increasing price of cup noodles.



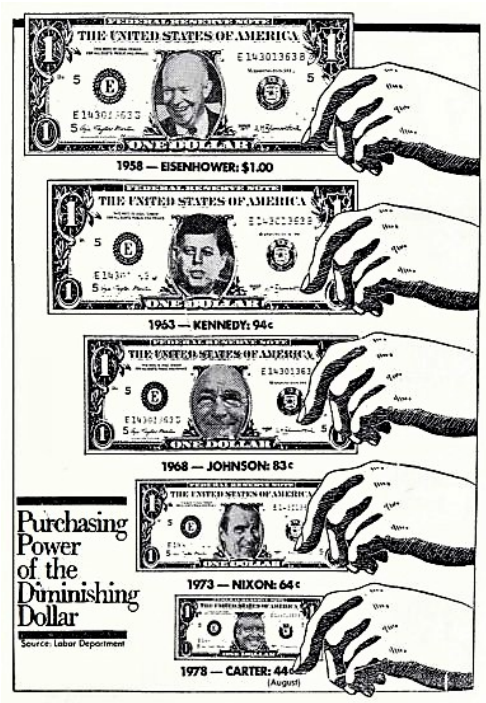
Visual area and numerical measure

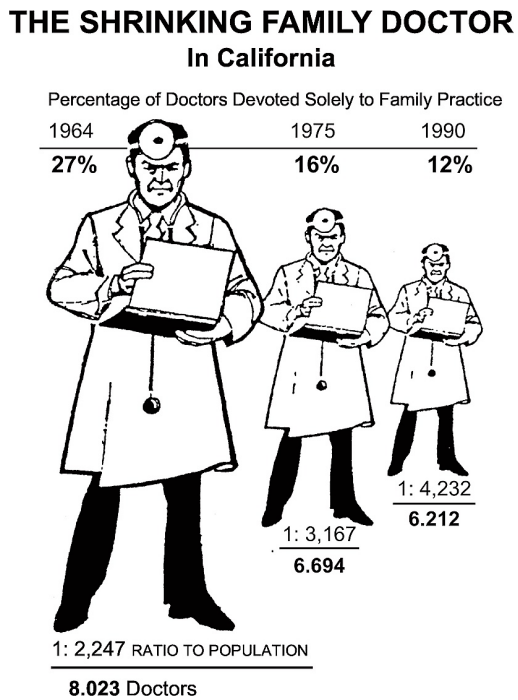
The use of two or three varying dimensions to show one-dimensional data is a weak and inefficient technique.

- Often has errors in design and ambiguity in perception.
- These designs cause many problems that should be avoided.
- Principle: The number of information-carrying dimensions depicted should not exceed the number of dimensions in the data.

Visual area and numerical measure

Should we compare their areas, lengths, or width?

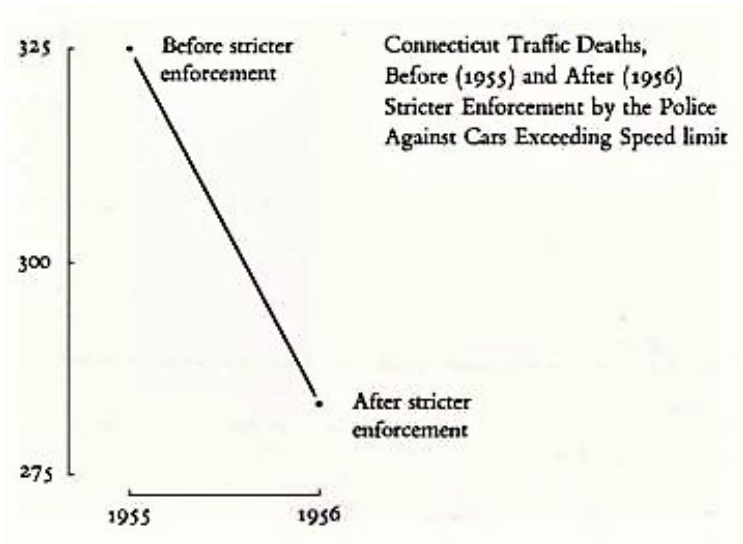




Context of data

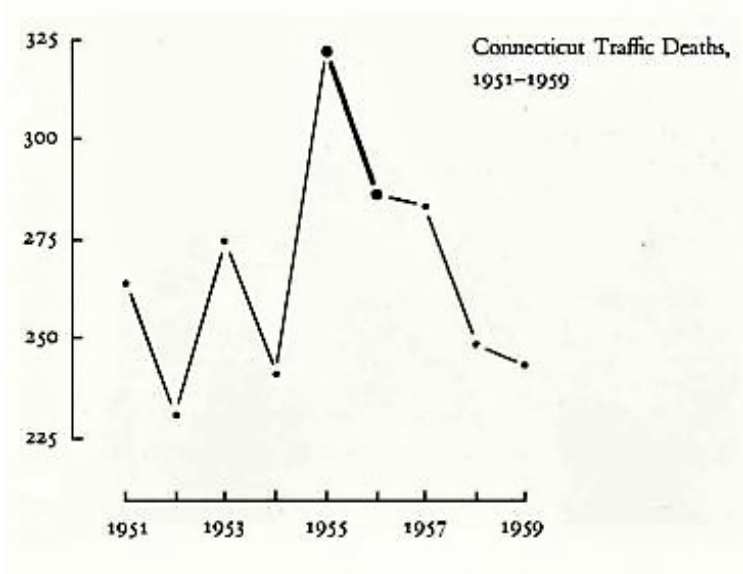
To be truthful and revealing, data graphics must bear on the question at the heart of quantitative thinking: "compared to what?"

- Data-sparse graphics should provoke suspicion. Graphics often lie by omission.
- Nearly all important questions are left unanswered by this display.



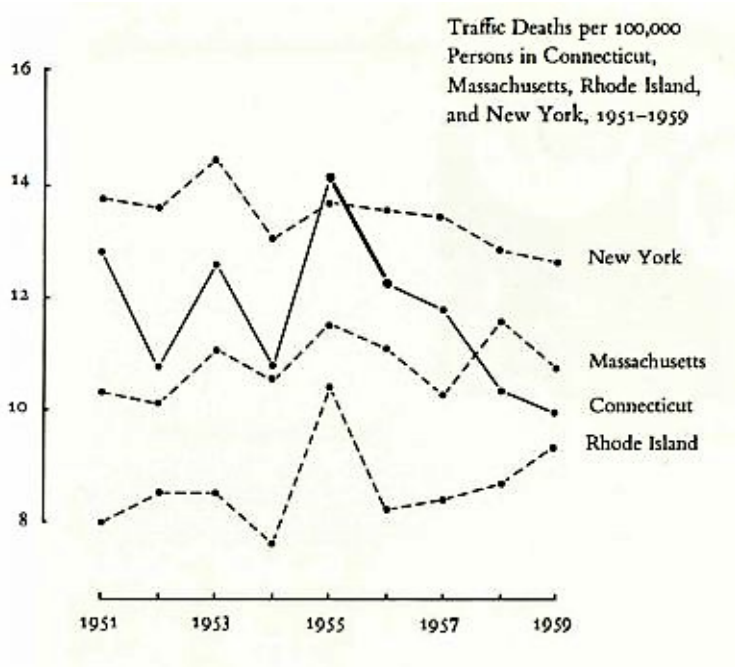
Context of data (revised)

- A few more data points add immensely to telling a more complete story.



Context of data (revised)

- Comparison with adjacent states give an even better context.
- It was not only Connecticut that enjoyed a decline in traffic fatalities in the year of the crackdown on speed driving.



Reasons for incompetent graphics

- A possible lack of quantitative skills among graphic designers.
- The doctrine that statistical data are boring – this leads to the exaggeration of evidence that we saw earlier. Here are some quotes:
  - *The challenge is to present statistics as a visual idea rather than a tedious parade of numbers.*
- The doctrine that graphics are only for the unsophisticated reader.
  - *(We placed more emphasis) on graphics than on information. We had feared children might be overwhelemed by too many facts.*
- This contempt for graphics and their audience leads to unsatisfactory displays.

Graphical integrity

Here are some guides to ensure the integrity of the graphics:

1. The graphical representation of numbers should be directly proportional to the numerical quantities.
2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. If necessary,
   - Use annotations to explain the data
   - Label important events
3. Show data variation, not design variation.

Graphical integrity (continued)

4. In time-series displays of money, deflated and standardized units of monetary measurements are nearly always better than nominal units.
5. The number of information-carrying dimensions depicted should not exceed the number of dimensions in the data.
6. Graphics must quote data in context.

Chart junks

Most of the time, interior decoration of graphics is all non-data ink or redundant data ink. It is referred to by Edward Tufte as **chart junk**.

> *Graphical decoration . . . comes cheaper than the hard work required to produce intriguing numebrs and secure evidence.*

- When a graphic is taken over by decorative forms, Tufte calls it a *duck*, in honor of this structure.
- The whole building itself is a decoration.

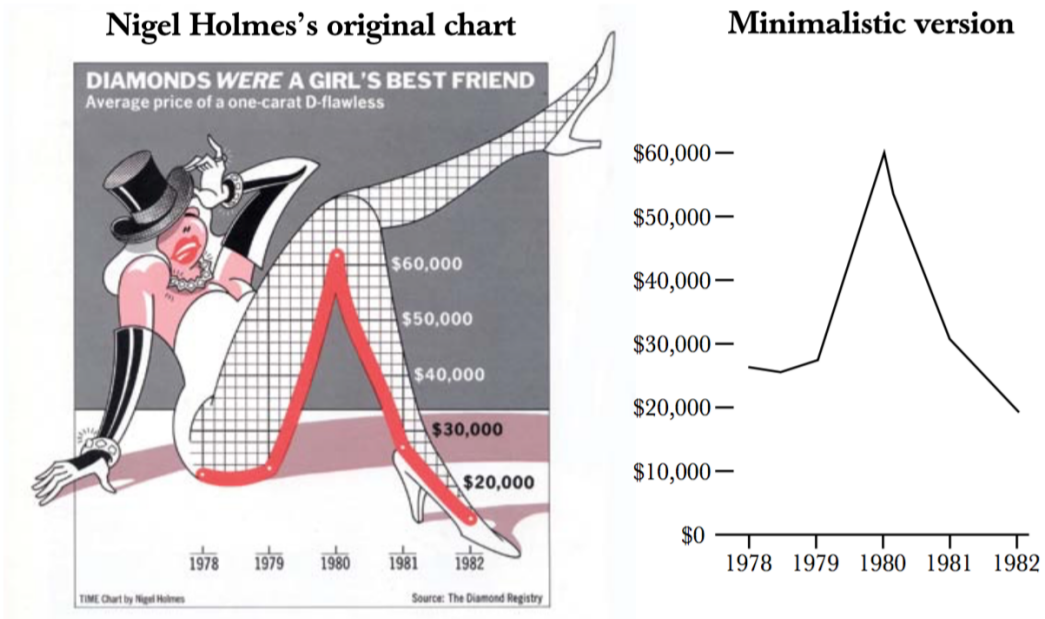
Chart junks

The most common types of chart junk are:

- Redundant representations of the simplest data
- Distracting patterns from computer software - Moire patterns
- Over-busy grid lines
- Excess ticks

Unnecessary dimension of data

The figure on the left is completely unnecessary and distracting.

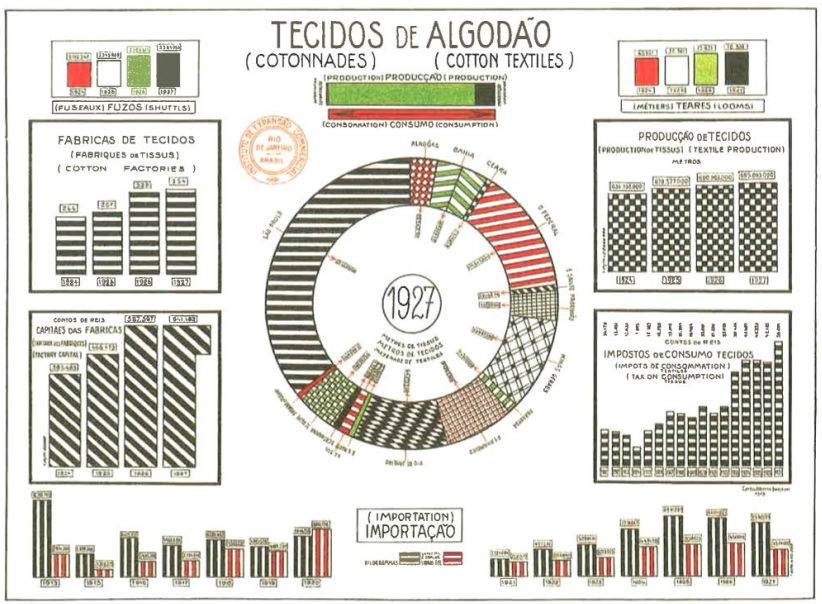


Source: Bateman et al. 2010

Unintentional optical art

Contemporary optical art relies on *moire effects* to produce the appearance of vibration and movement.

- However, when these effects are applied to statistical presentations, they are distracting and add clutter.
- The noise clouds the flow of information.



Hatching pattern in 3D

In this example, what should have been simple tables are turned into bad graphs published in major scientific journals.

- One of the variables has no axis; some pyramids hide others. It would be better to display a boxplot for each group after operation.

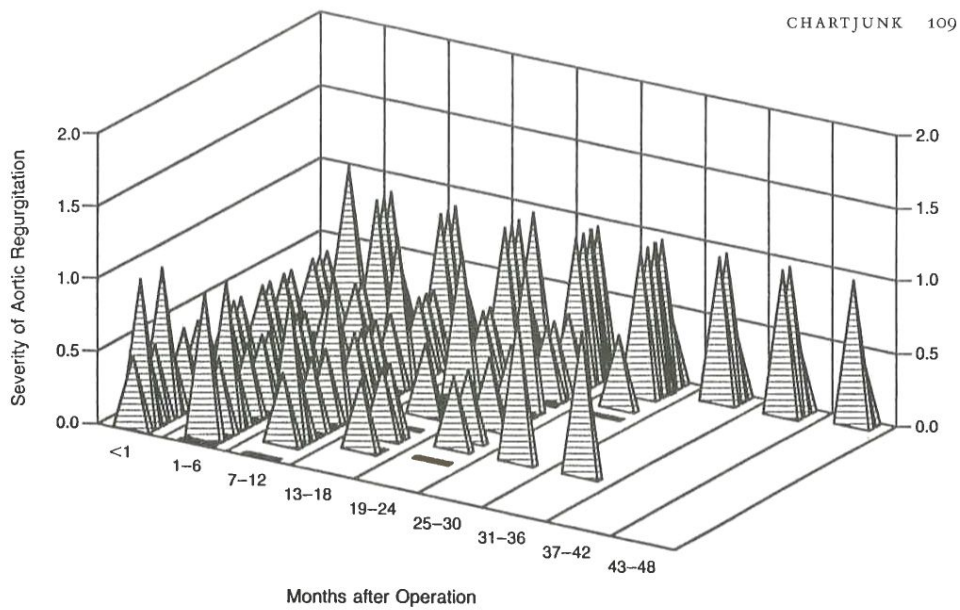
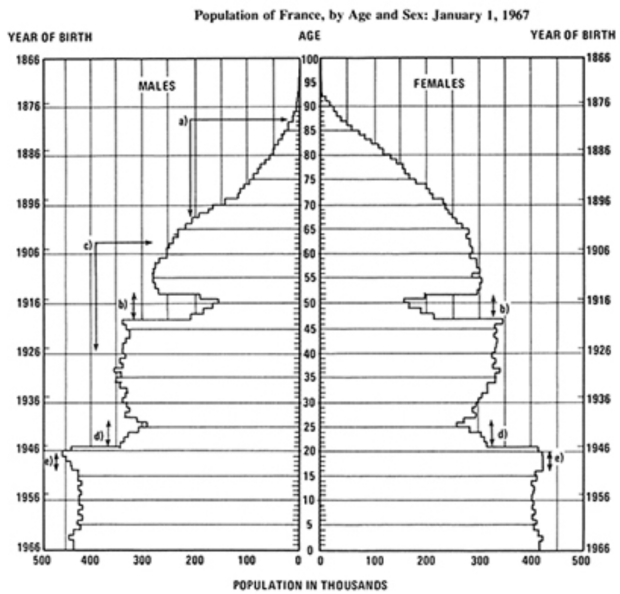


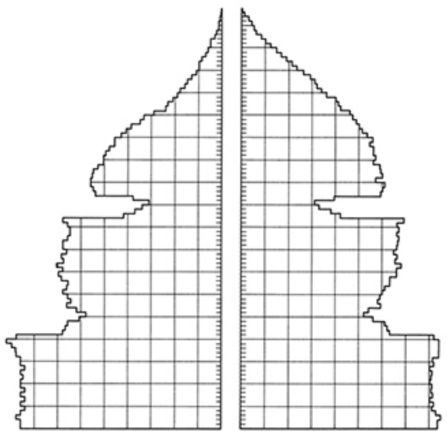
Figure 2. Serial Echocardiographic Assessments of the Severity of Regurgitation in the Pulmonary Autograft in 31 Patients. The numerical grades were assigned according to the severity of regurgitation, as follows: 0, none; 0.5, trivial; 1.0 to 1.5, mild; 2.0, moderate; and 3.0, severe.

The grid

One of the more sedate graphical elements, the grid, should usually be muted or completely suppressed so that its presence is only implicit.

- The revised graph on the right hand side quites the grid and gives the emphasis to the data.





Presenting tables

When presenting tables, we could reduce over-bearing grid lines by removing them, allowing the formatting to present the structure of the table.

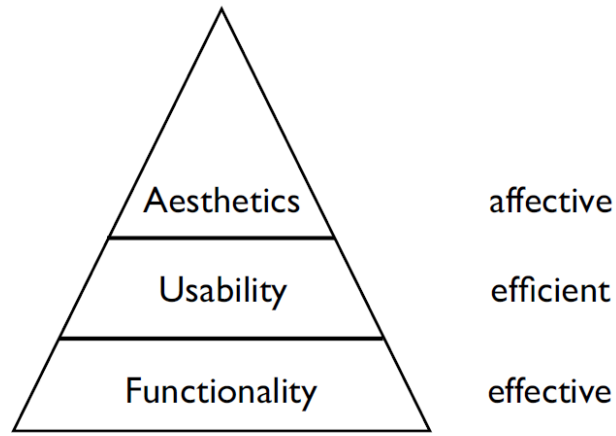
- The grid-lines seem to imprison the values in jail cells.
- In most situations, we can do away with the vertical lines, unless the cells are too narrow to distinguish themselves from one another.

Train No.	3701	3301	3801	3542	3765
New York	12:10	1:30	3:45	7:30	4:33
Newark, N. J.	1:43	10:30	5:21	8:50	11:45
North Elizabeth	....	.....	.....	.....	6:45
Elizabeth	3:33	2:05	.....	.....	7:05
Peekskill	5:34	6:40	.....	7:20	8:50
Ediison, N. J.	4:45	5:20	4:40	2:10	11:05
Princeton, N. J.	1:30	.....	.....	3:30	7:30

New York	12:10	1:30	3:45	7:30	4:33
Newark, N. J.	1:43	10:30	5:21	8:50	11:45
North Elizabeth	....	.....	.....	.....	6:45
Elizabeth	3:33	2:05	.....	.....	7:05
Peekskill	5:34	6:40	.....	7:20	8:50
Ediison, N. J.	4:45	5:20	4:40	2:10	11:05
Princeton, N. J.	1:30	.....	.....	3:30	7:30
Train No.	3701	3301	3801	3542	3765

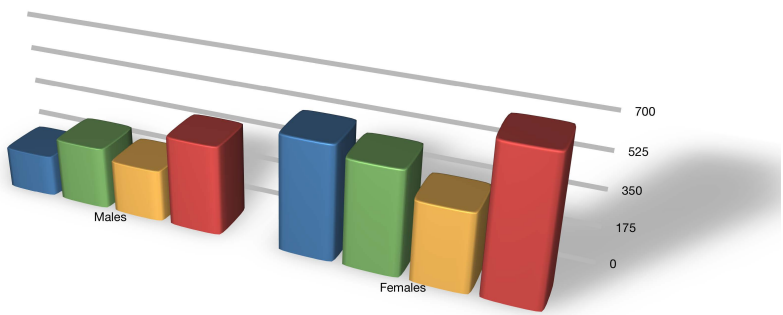
Tufte's graphical excellence

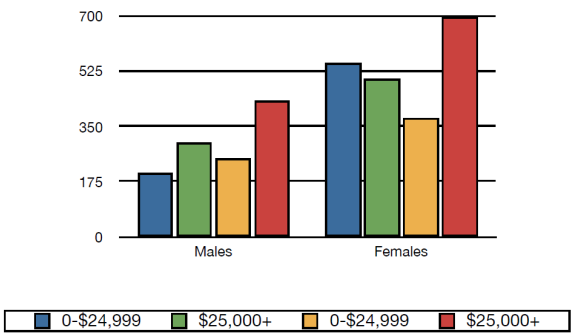
- Interesting data
  - Complex ideas, multivariate data
- Clear, precise, and concise presentation
  - Maximize data-ink ratio
- Accurate communication



Maximize data-ink ratio

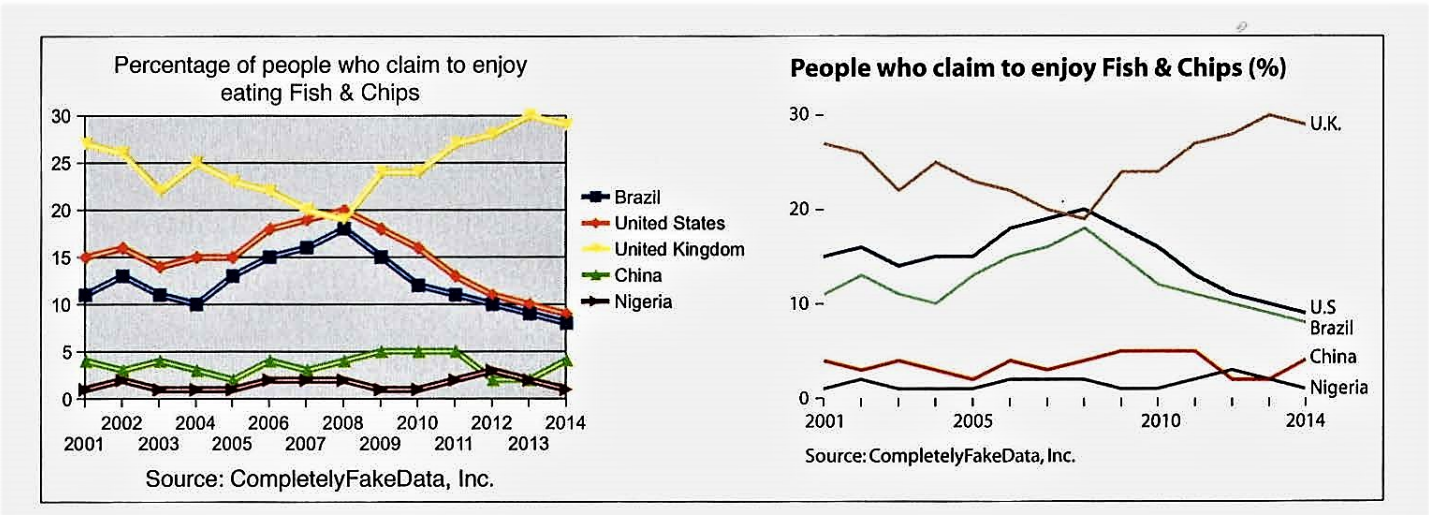
$$\text{Data-ink Ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graph}}$$
$$= 1 - \text{proportion of the graph that}$$
$$\text{can be erased without loss of information}$$





Maximize data-ink ratio

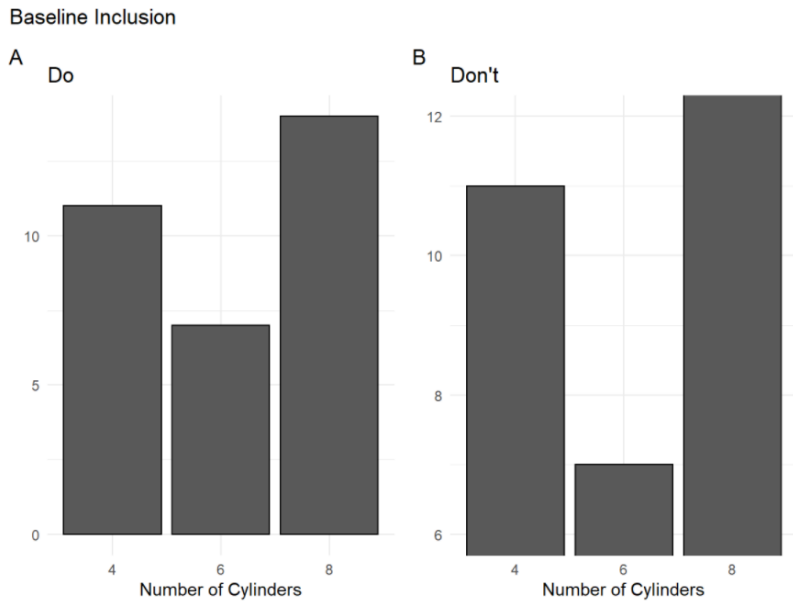
The chart on the right does not contain redundant elements. It is certainly the choice that is more aesthetically pleasing.



Other considerations

Include the baseline

- Omitting baseline values (typically 0) can mislead the audience into thinking that patterns are larger (or smaller) than they really are.



Other considerations

Use inclusive colors

Source: David Nicholes

- Choose the color schemes that can be easily identified by people with all types of color vision.
  - Read more on the viridis color palette
- When communicating the graph, clearly state the color names.
- It is a good practice to, where possible, avoid conveying information purely through color.

Other considerations

Consider the audience

- Your audience will have different levels of technical expertise or familiarity with the subject, so keep this in mind!
- Avoid going against conventions within your field.

Avoid cherry-picking

- This is a common way of attempting to deceive audiences into thinking a trend exists (or not).
- If you need to drop values or observations, make sure this is mentioned and justified.

Other considerations

The scientific process:

- Treat each "finding" as a conjecture.
- Do not believe it until you can find multiple ways of confirming it.
- Ask your team mates/ colleagues to review your work.
- Always follow-up with more questions, investigations, and plots.

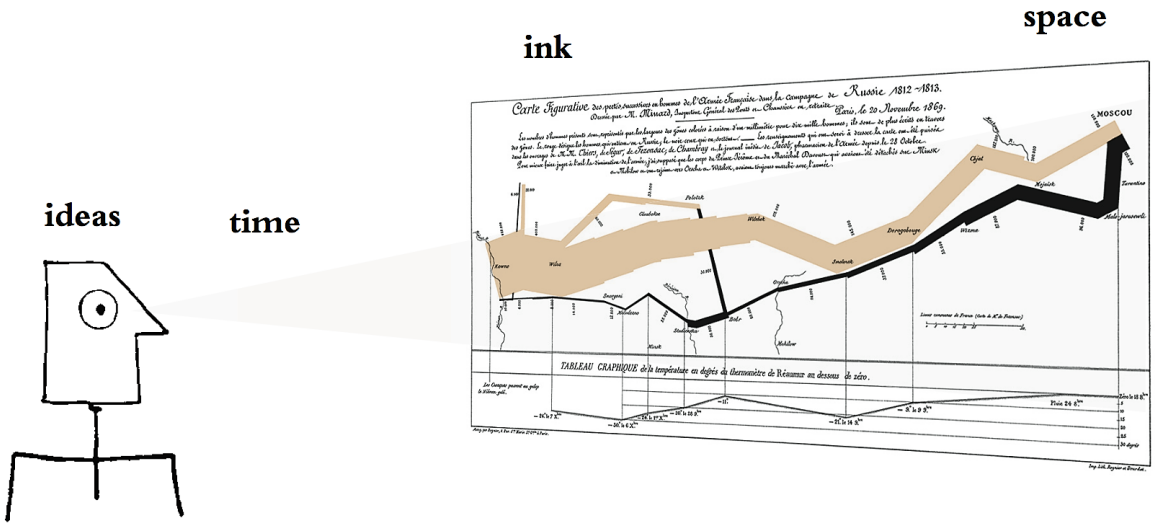
Summary

We cover the basics of visualization techniques and principles, which builds the foundation for our understanding of the `ggplot2` package.

Of course, there is much more to data visualization than we cover here.

- We do not cover interactive graphics, as it is a topic that is too advanced for our course. Some useful resources for those interested in learning more can be found below:
  - https://shiny.rstudio.com/

Closing remarks



A figure presented by Tufte, which describes graphical excellence as

> *that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.*

Closing remarks

Data graphics are paragraphs about data and should be treated as such.

Graphical elegance is often found in simplicity of design and complexity of data.