The Erdős Institute
Data Visualization Mini Course
Problem Set 1

Technical Practice

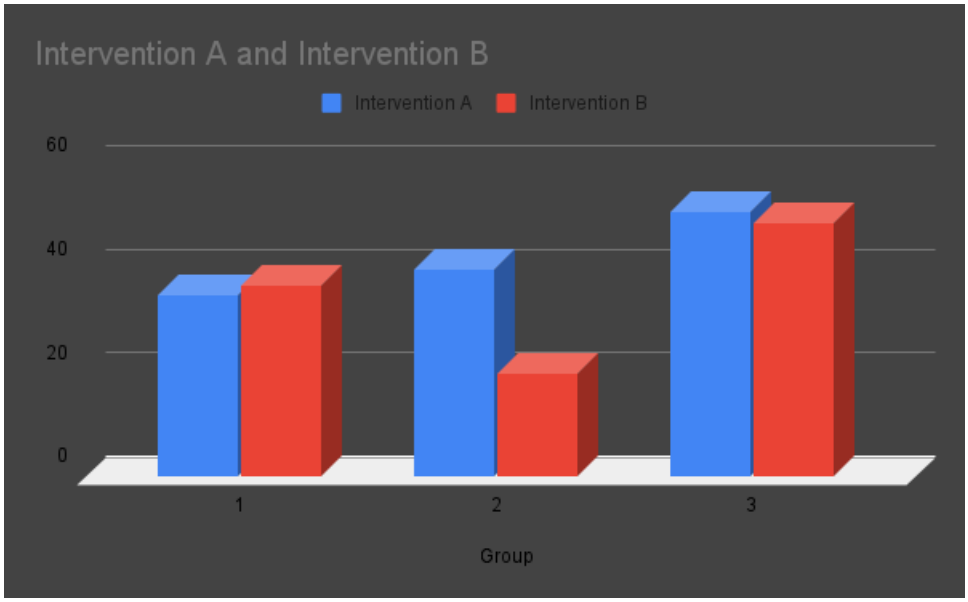
Work through the following `jupyter notebooks`:

1. `problem_sets/problem_set_1/Question 1.ipynb`
2. `problem_sets/problem_set_1/Question 2.ipynb`
3. `problem_sets/problem_set_1/Question 3.ipynb`
4. `problem_sets/problem_set_1/Question 4.ipynb`

Clutter or Chartjunk

Question 1

A key concept in data visualization concerns the idea of *clutter* or *chartjunk*. Clutter or chartjunk is any element of the plot that distracts the audience from the central data. As an example, check out the following graphic:



The dark background, three-dimensional bars, and, for some practitioners, the grid lines are examples of chartjunk. To remedy this figure you could:

- Either make the background lighter or make the text have better contrast with the dark background,
- Make the bars two-dimensional,
- Remove the grid lines and label the bars directly.

Using the file `chartjunk_bar_data.csv` in the `data` folder of the repository, remake this graph in a way that reduces the clutter. Note that you should feel free to try other chart types.

Question 2

All of the following graphics have clutter. Try to identify the clutter and make a suggestion on how you could improve the figure.

Graph 1

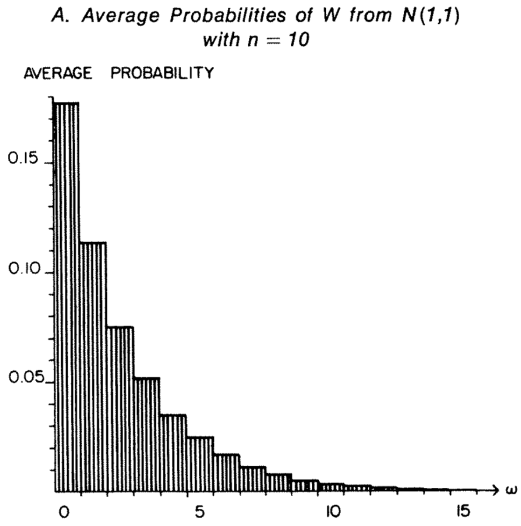


Figure 1: Source: "JASA Style Sheet." Journal of the American Statistical Association, 71 (March 1976), 260-261. Note that the visual effect caused by such close vertical hatching is known as the Moiré effect.

Graph 2

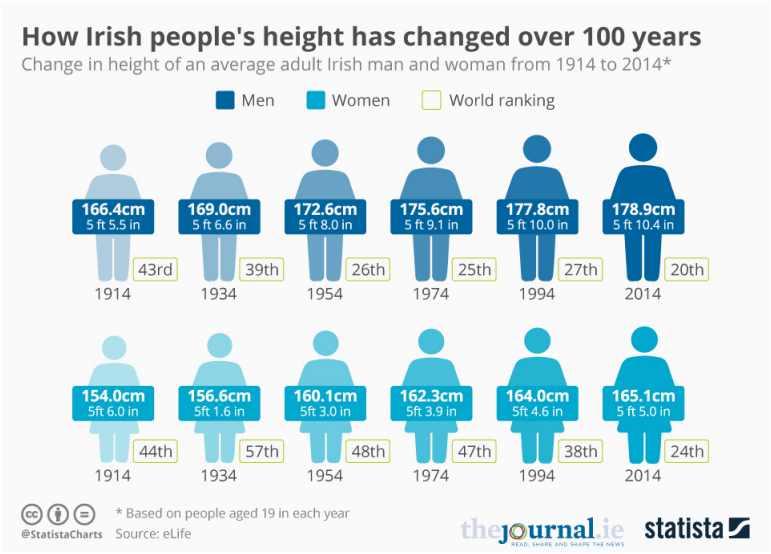


Figure 2: Source: https://www.statista.com/chart/5441/how-irish-peoples-height-has-changed-over-100-years/

Graph 3

Figure 3: Source: Bateman, Scott, et al. "Useful junk? The effects of visual embellishment on comprehension and memorability of charts." Proceedings of the SIGCHI conference on human factors in computing systems. 2010.

Think of Your Audience

One important piece of advice in data visualization is to know and think of your audience. The point of a data visualization is to convey something about the data to some kind of audience. When you make design decisions you should be mindful of who will ultimately be reading your charts and what you are trying to convey to them.

Tufte's Boxplot

Edward Tufte is a notable figure in modern data visualization. In his work, The Visual Display of Quantitative Information, he laid out a number of principles that help guide data visualization thinking. In particular, he defined the term chartjunk while discussing the concepts of *data ink* and *data ink ratio*. Data ink refers to the amount of "ink" in your graphic dedicating to conveying the data, while the data ink ratio is the ratio of data ink to all ink used. Tufte's philosophy was that as practitioners we should strive to maximize the data ink ratio of our graphs. One such way was to reduce the clutter in our charts.

However, this extreme approach may not always be the best approach (see the paper associated with Figure 3 for one such example). If, in our quest for data ink ratio maximization, we make our plots more difficult for the audience to read or interpret, then the graph is likely a failure.

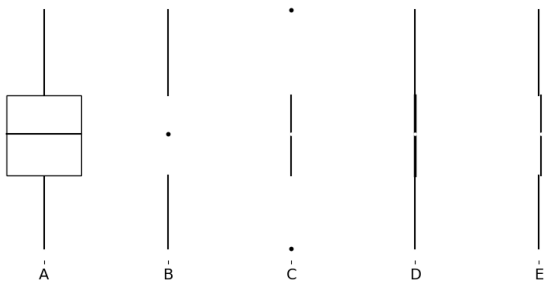


Figure 4: Five variations on a box and whisker plot.

Consider the variations on the box and whisker plot presented in Figure 4. In The Visual Display of Quantitative Information Tufte claimed that the box portion of the box and whisker plot could be mostly erased without a loss of information. He then proposed these four designs as charts that had higher data

ink ratios while still conveying the same information about the data. Compare these five designs and see whether you agree. Which of these designs do you prefer and why? Do you think any of the designs could pose a challenge to an audience in a publication or presentation?

Langren's Graph

In their book, A History of Data Visualization & Graphic Communication, Friendly and Wainer claim that Michael van Langren produced the first statistical graphic in 1644. This graph was a single number line whose origin started at 0 and marked the longitude position of Toledo, Spain. Apart from the point directly beneath Toledo, the plotted values displayed the available estimates for the longitudinal distance from Toledo to Rome, Italy.

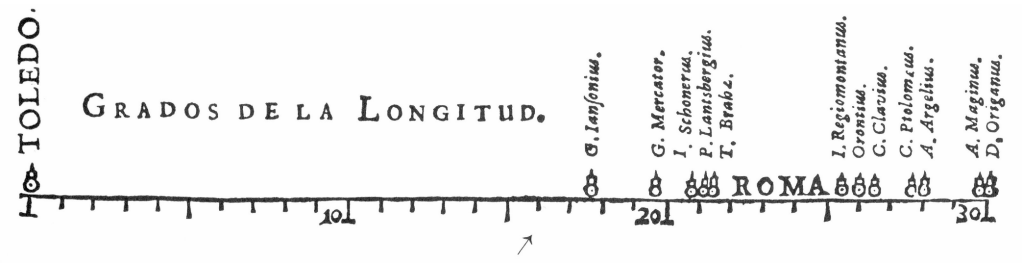


Figure 5: Source: https://www.datavis.ca/gallery/langren/. The arrow annotation has been added to mark the correct distance from Toledo to Rome and was not present in the original formulation of the graph.

At that time, determining correct longitudes while at sea was incredibly important and developing new more accurate techniques could be incredibly profitable. Langren included an early version of this graphic in what was essentially a grant proposal to the governor of the Spanish Netherlands. Pretend you are governor Isabella and try and interpret the meaning behind Langren's chart (seen in Figure 5). What do you think is being conveyed here? When you are ready proceed to the next page to read what Langren's intention was per Friendly and Wainer.

From Friendly and Wainer:
"You can plainly see in my charts that even the Longitude between Toledo and Rome is subject to large errors. 'If the Longitude between Toledo and Rome is not known with certainy, consider Your Highness, what it will be for the Western and Oriental Indies, that in comparison the former distance is almost nothing.' "

Also:
"Thus he makes explicit that the purpose of drawing a graph is to show the 'countless errors' in the determination of longitude distance between two relatively well-known locations. In statistical language, his presentation goal was to show uncertainty or variability rather than a best estimate obtained from pooling the data points."

How did this compare with your idea of Langren's take away message? Would it surprise you to know that it took Langren several attempts to secure funding for his research into the problem?

(Unintentional) Misleading Graphics

Cholera in 1800s London

While there are some groups that intentionally make misleading graphics, it is also true that we can unintentionally make a misleading graphic. One of the most famous cases of such a graphic can be found in work on cholera mortality in 1800s London.

Prior to the work of John Snow the running theory of how cholera spread throughout London followed the miasmatic theory of disease, which held that cholera was spread from noxious "bad air" emanating off of the Thames due to the dumping of untreated sewage directly into the river. William Farr, who led the collection of official medical statistics in England and Wales, used cholera mortality data to try and prove this theory true with his over 500 pg *Report on the Mortality of Cholera in England, 1848-49*. In his work he found that if he arranged London's districts in order of their elevation above the Thames there appeared to be an inverse relationship with the cholera mortality rate in that district. That is it appeared that:

$$\text{Cholera Mortality Rate} \propto \frac{1}{\text{Relative Elevation to Thames}}.$$

Farr developed his "natural law of cholera" by devising the following formula using mortality data:

$$E : E' \text{ as } C' : C \implies C' = \frac{E+a}{E'+a}C',$$

where C is the cholera mortality rate at elevation E, C' is the cholera mortality rate at elevation E', and a is a scaling constant found to be 12.8. Using this law Farr calculated estimates for the cholera mortality rate at various elevations relative to the Thames. At the time, the notion of a scatter chart had not yet been invented, so Farr instead crafted the plot seen in Figure 6 in which the vertical positional represents the elevation and the horizontal length represents the cholera mortality rate per 10,000 people according to his law.

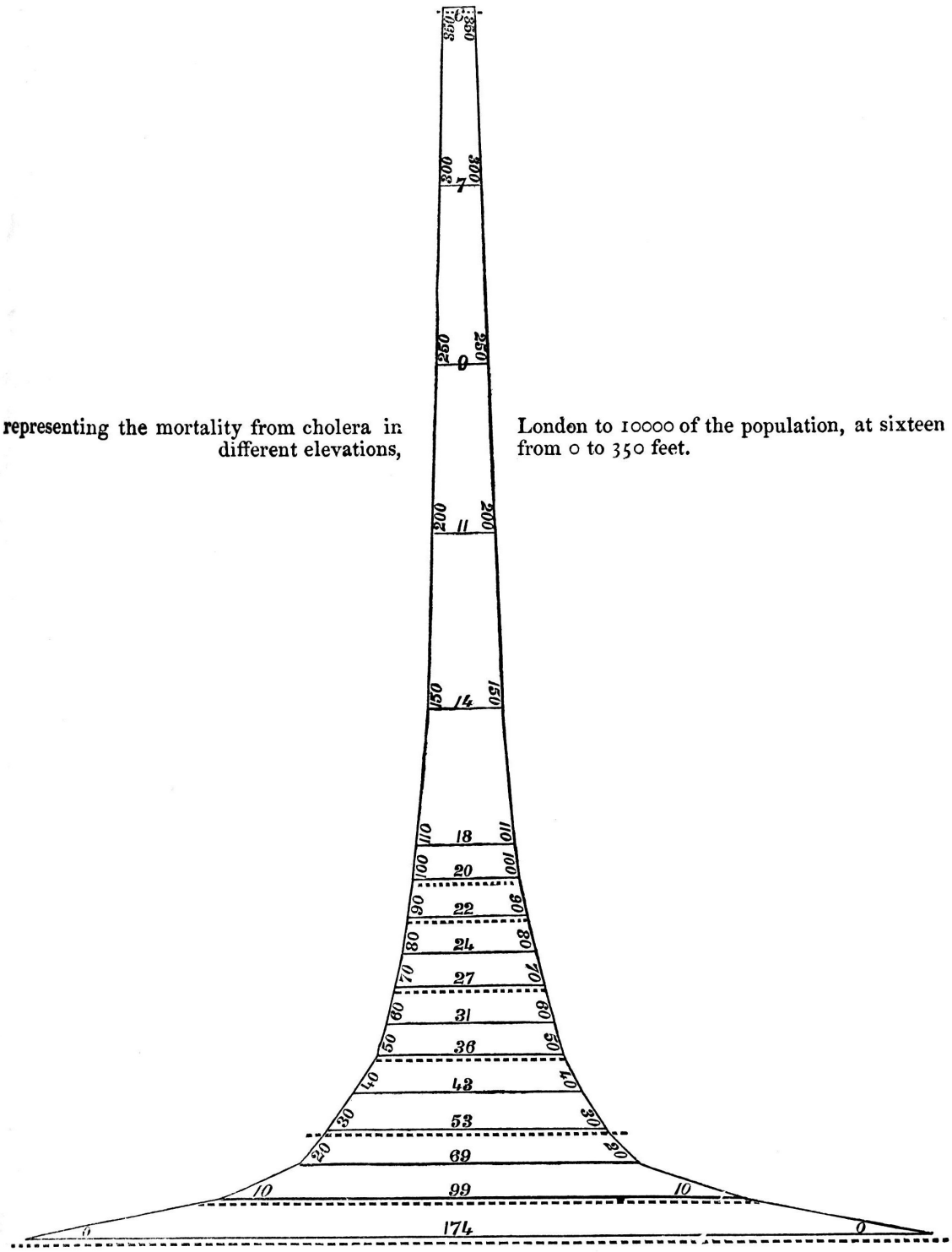


Figure 6: Source: General Register Office, Report on the Mortality of Cholera in England, 1848–49. London: Printed by W. Clowes, for HMSO, 1852. p. lxv.

In this exercise you will use scatter charts and line charts to examine Farr's natural law.

Part A

Using the `farr_cholera.csv` data in the `data` folder make a scatter plot of the `cholera_drate` column against the `elevation` column. `cholera_drate` gives the cholera mortality rate for the corresponding `district` per 10,000 people and the `elevation` gives the elevation relative to the River Thames. Then plot a line chart on top of the scatter plot using the data in `farr_law.csv`, which provides the data Farr plotted in his chart.

If you want, you can use `seaborn` to also overlay a loess regression over the data.

Does Farr's natural law appear to fit these data well?

Part B

Using the same data from Part A make a scatter plot of the actual death rate against the inverse of the elevations plus 12.8, $1/(E + 12.8)$. Calculate the correlation between these two variables. Then plot a line chart of the same variables predicted by Farr's natural law.

Again how do these compare?

Part C

Okay, now you will look at what may have be the issue with Farr's work.

Remake the previous two scatter plots, but now color the markers (or change their shapes or both) by their water supply region, contained in the `water`.

What do you see? Do you think that elevation was the driving cause of cholera mortality at this time?

Part D

While Farr's work can be found to be slightly incorrect, as you have in Part C, I do not think that we should look too harshly on him in hindsight. Even ignoring the aid of computers, the techniques and technology we used in this exercise were not yet devised, and Farr's report was quite thorough. This is not a case of malicious misuse of data (more on that in the next problem set).

John Snow is credited as demonstrated that cholera was spread via the consumption of contaminated water through his work in the 1854 London cholera outbreak. In this work John Snow enlisted the aid of the map-based graphic shown in Figure 7.

Figure 7: Source: John Snow, On the Mode of Communication of Cholera, 2nd ed. London: John Churchill, 1855 / Wikimedia Commons.

In this graphic, John Snow marked deaths from cholera with black bars at the address of the deceased's residence. Snow identified that deaths appeared to cluster near the pump indicated in the zoom-in on the

right and hypothesized that water from that pump was contaminated. The handle that allowed the pump to produce water was removed and the outbreak began to subside, although Snow thought it was possible the outbreak had already begun to subside prior to the decommissioning of the pump. To learn more about John Snow's work check out his wikipedia entry, https://en.wikipedia.org/wiki/John_Snow.

Textbook References

- "A History of Data Visualization Graphic Communication", by Michael Friendly and Howard Wainer, Harvard University Press, Cambridge MA: 308 pages including references and index. 2021.
- "The Visual Display of Quantitative Information", Tufte, Edward R., Graphics Press; 2nd edition (February 14, 2001): 200 pages. 2001.