Data Ethics

Dav King

2/1/2022

Discuss the misleading data representations here.

Baby Boomers

The graph is trying to show the proportion of Baby Boomers who identify with certain terms - e.g., 40% of Boomers would call themselves "Leaders".

This graph is misleading because it maps these proportions onto an areal image, as if to suggest they sum to 100% in the same manner as a pie chart would (which are another type of graph that frequently get abused in this manner), but these proportions do not and are not supposed to sum to 100% - instead, they are each relative to 100% of themselves, not relative to one another.

To improve this graph, I would make each one of those terms be its own graphic of a person, with the fill determined by the statistic of each category (i.e. a graph for "Leaders" would be 40% full).

Brexit

This graph is *maybe* trying to show the proportion of people who are planning to vote on each side of Brexit. However, it is misleading enough that I have questions about whether that's actually what it's trying to show.

It seems that we have forgotten about how numbers actually scale. Votes for "No", 43% of the population, take up notably over half of the pie chart. Votes for "Yes", meanwhile, at 47% are more like 33% of the area of the chart. Additionally, placing them on the background of the UK flag makes it difficult to interpret to begin with. It also obscures the actual data being measured - I'm leaving my incorrect answer from before because that is firmly what I believed it to be showing, but in fact this data shows the proportion of people in Scotland who would vote to leave Britain after Brexit.

I wouldn't make a pie chart (never make pie charts) - I would make different bars in a chart for each relative voting outcome.

Spurious Correlation

This is showing a correlation between two unrelated variables, implying that they might be causally linked in some way even though it is highly unlikely that such a thing would be the case.

Discuss webscraping here.

Tweets

The researcher is violating the website's setup and likely its terms and condition in order to steal data for analysis that they themselves did not collect. The weather application spent money and effort to collect that

data, and is being kind enough to provide some of it for free. The researcher is gaming the system for this, when they very firmly shouldn't be.

The researcher can simply pay for the data or wait until enough time has passed that they can get it for free without making the six accounts.

Posting Data from Social Media

This is the most identified data I have seen in my life. To be honest, this is literally worse than just leaving the names in for a normal study. *Geographic location* and *IP Addresses*?? Not to mention a whole host of other issues. If one were to publish this data set publicly, lots and lots of people could be tracked down and associated with data that they do not want to be publicly associated with.

You have no need for this specific information. Use pseudonyms (i.e. PIDs) for your individual data points, do not publish specific demographic data unless it has been aggregated, and make sure your fully identified data set is heavily encrypted and stored away offline.

Once again, take away all of your specific identifiers and only publish aggregate demographic info unless a) it is absolutely impossible to do so or b) said info is broad enough that it is unlikely to be identifiable. If possible, they should also contact the participants (but this is rarely possible, especially with big data studies).

Answer the discussion questions here.

Simpson's paradox comes back in a number of places. While I'm not sure whether it was this or a different stat-based class I'm taking this semester, there was a very interesting trend in the Palmer Penguins dataset where a correlation appeared overall to be positive but, when subset by species of penguin, was negative within each species.

We should consider what the algorithm is trained on (men), what it therefore perpetuates (more men), and whether or not we should build a counter into the algorithm to ensure a more equitable display of jobs between the genders (assuming we believe that both genders are capable of these jobs, which we should).

Data lies. It can say whatever you want it to say, even moreso when you're building out data visualizations. Don't attack your data with an agenda, attack it with an open mind and a plan to explore anything that may come out of it. Know where your data is coming from, how it was collected, what it contains, what issues it may hold, and how it should be treated.