

Introduction to Data Science

Data Ethics - Misrepresentation and Data Privacy

Joanna Bieri DATA101

Small Group Discussion

DISCUSS

Each individual:

- ① What book or articles are you reading?
- ② What so far have been the most interesting ethical points?

As a group:

- ① Define misrepresentation and data privacy in your own words.
- ② Can you draw some examples from your individual reading that might pertain to our discussion today?

Mini Review - Misrepresentation

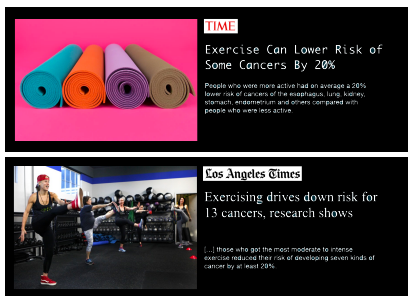
REVIEW

- Misrepresentation can happen intentionally or unintentionally.
- It can arise because of lack of knowledge or competence in data science.
- It is important to be aware of misrepresentation and be able to spot it a mile away in your work and in other peoples work!

Today: Causality, Data Visualization, Maps and areas, Visualizing uncertainty, Data Privacy.

Mini Review - Causality

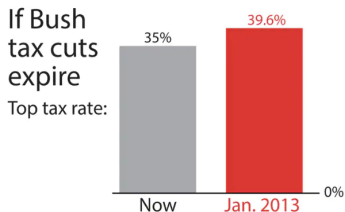
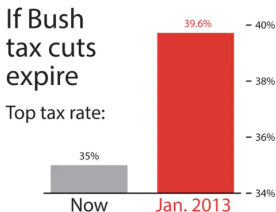
REVIEW



- **Volunteers** were **asked** about their physical activity level over the preceding year. (Survey!)
- So what can we actually say? Exercise was associated with lower cancer rates.

Mini Review - Axes and Scale

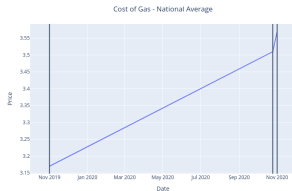
REVIEW



- The numbers at the top of the bars are the same. Notice that the minimum for the y-axis is different!
- Your axis should **usually** start at zero, unless there is a good, honest, data visualization reason.

Mini Review - Axes and Scale

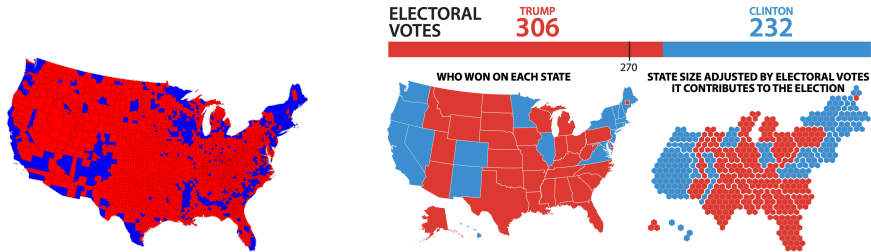
REVIEW



- The x-axis is evenly spaced even though it represents very different time steps!
- We read the data off of the plot, put it into a data frame, and made a plot of our own.
- **Your axis should always be consistent in scale!!!**

Mini Review - Maps and Areas

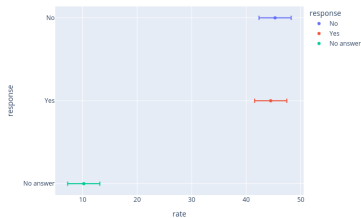
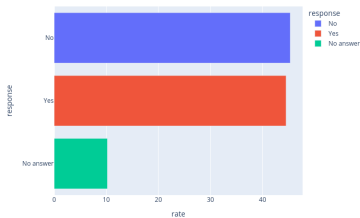
REVIEW



- Each of these counties do not have the same number of people.
- Geographic areas do not decide elections. Electoral votes do!

Mini Review - Visualizing Uncertainty

REVIEW



- “Catalan public opinion swings toward ‘no’ for independence, says survey”
- The margin of error is $\pm 2.95\%$. By plotting the data with clear error bars, it makes it more clear how close the vote actually is.

Mini Review - Data Privacy

REVIEW

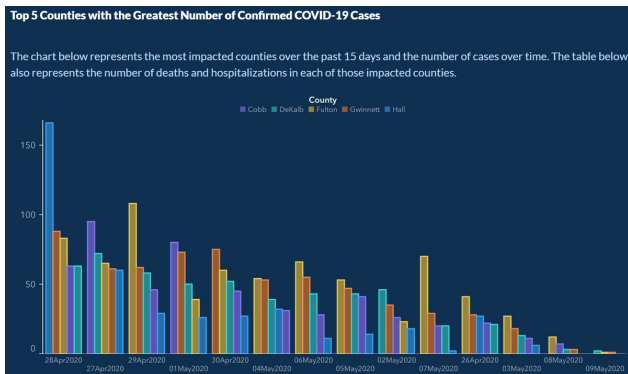
What amount of expected privacy are we entitled to when we put our data online?

- **AOL search data leak:** the names were not included in the data, but it was very easy to connect the names.
- **OkCupid:** researchers published data of 70,000 users. Usernames were often real names or reused across platforms.
- **Facebook & Cambridge Analytica:** a personality quiz also collected likes and personal information. It also grabbed their friends data.

You should be very critical of where your data is coming from. Make sure it was sourced ethically!

What Is Wrong with This Picture?

ASK



What story does this graph tell? What is the real story in the data?

What Is Wrong with This Picture?

ANSWER

- The bars get shorter from left to right, so it looks like cases are going down over time.
- Read the dates on the x-axis: 28 April, 27 April, 29 April, 1 May, 30 April... They are NOT in order.
- The days were sorted by the height of the bars, and the counties are in a different order inside each day.
- To fix it: put the dates in order and keep each county in the same place.

Who Is Winning?

PREDICT

poll

	candidate	percent	margin of error
0	A	46	3
1	B	44	3

A headline says "A pulls ahead of B". Is that what the data says?

Who Is Winning?

ANSWER

	candidate	lowest	highest
0	A	43	49
1	B	41	47

We only know each number to within 3. A could be as low as 43 and B could be as high as 47, so the two ranges overlap and we can not say who is ahead. Show the error bars!

Doing the Right Thing?

ASK

More exercise is good for you, right?

Is it okay to misrepresent your data
if you are getting people
to "do the right thing"?

Doing the Right Thing?

ANSWER

- There is not one right answer here. Listen for reasons.
- We want to make sure we are not overreaching, giving people false hope or knowledge.
- Just because we can't make a causal inference doesn't mean the study is bad, but we want to be clear what the data says.
- If people find out one result was stretched, will they believe the next one?

Already Public?

ASK

The OkCupid researchers said:
"all the data found in the dataset are
or were already publicly available"
Is that enough?

Already Public?

ANSWER

- When users gave this information was there an expectation of privacy?
- Did they think the data would be released in a more useful form for analysis?
- Real names and pictures were not released, but identities could easily be uncovered from the usernames.
- How would you want your data treated?

Should I Scrape This Data?

ASK

Does the answer to
"Should I scrape this data?"
depend on what the data is about?
Give two examples.

Should I Scrape This Data?

ANSWER

- Yes. On Tuesday we scraped countries, capitals, and book prices. Nobody can be harmed by that.
- Dating profiles, health posts, or search histories are about people, and people can be identified.
- Ask: who could be harmed, did they expect this, and do I need all of it?

Questions?

Questions?