LESSON 3

Check Your Assumptions

Objectives: You will be able too…

- Define the digital divide as the variation in access or use of technology by various demographic characteristics
- Identify assumptions made when drawing conclusions from data and data visualizations

GETTING STARTED:

- Survey reminder → Google Classroom
- Video: Google Flu Trends
- Prompt: What are the potential beneficial effects of using a tool like Google Flu Trends?

DISCUSSION:

- Incorrect assumptions about a dataset can lead to faulty conclusions
- Earlier prediction of flu outbreaks could limit the number of people who get sick or die from the flu each year.
- More accurate and earlier detection of flu outbreaks can ensure resources for combating outbreaks are allocated and deployed earlier (e.g., clinics could be deployed to affected neighborhoods).

Google Flu Trends Failure

- See links on msklug.weebly.com
- Thinking Prompt: Why did Google Flu Trends eventually fail? What assumptions did they make about their data or their model that ultimately proved not to be true?

DISCUSSION:

- Google Flu Trends worked well in some instances but often over-estimated, under-estimated, or entirely missed flu outbreaks. A notable example occurred when Google Flu Trends largely missed the outbreaks of the H1N1 flu virus.
- Just because someone is reading about the flu doesn't mean they actually have it.
- Some search terms like "high school basketball" might be good predictors of the flu one year but clearly shouldn't be used to measure whether someone has the flu.

DISCUSSION:

- In general, many terms may have been good predictors of the flu for a while only because, like high school basketball, they are more search in the winter when more people get the flu.
- Google began recommending searches to users, which skewed what terms people searched for. As a result, the tool was measuring Google-generated suggested searches as well, which skewed results.

SOME THOUGHTS:

- The amount of data now available makes it very tempting to draw conclusions from it.
- There are certainly many beneficial results of analyzing this data, but we need to be very careful.
- To interpret data usually means making key assumptions. If those assumptions are wrong, our entire analysis may be wrong as well
- Even when you're not conducting the analysis yourself, it's important to start thinking about what assumptions other people are making when they analyze data, too.

The Digital Divide and Checking Your Assumptions

- We are going to use the "Digital Divide and Checking Assumptions – Activity Guide"

PART 1: THE DIGITAL DIVIDE

- Access and use of the Internet differs by income, race, education, age, disability, and geography.
- As a result, some groups are over- or under-represented when looking at activity online.
- When we see behavior on the Internet, like search trends, we may be tempted to assume that access to the Internet is universal and so we are taking a representative sample of everyone.
- In reality, a "digital divide" leads to some groups being over- or under- represented. Some people may not be on the Internet at all

PART 2: CHECKING YOUR ASSUMPTIONS

WRAP UP:

- Assumptions:
  - The data collected is representative of the population at large (e.g., ignoring the "digital divide").
  - Activity online will lead to activity in the real world (e.g., people expressing interest in a candidate online means they will vote for him or her in real life).
  - Data is being collected in the manner intended (e.g., ratings are generated by actual customers, instead of business owners or robots).
  - Many other assumptions regarding data are possible.

WRAP-UP:

- Would anyone like to revise the explanation they gave for their google trends research in the previous lesson?
- Has what you've learned today changed your perspective on the "story" you thought the data was telling?
- In this course we will be looking at a lot of data, so it is important early on to get in the habit of recognizing what assumptions we are making when we interpret that data.

WRAP-UP:

- In general, it is a good idea to call out explicitly your assumptions and think critically about what assumptions other people are making when they interpret data.
- We may not become expert data analysts in this class, and even orgs like Google can make mistakes. Just be honest with yourself about assumptions you are making, correct your wrong assumptions when you can and keep an eye out for the assumptions others may make when they tell us "what the data is saying"