COVID-19 data in the classroom

Maria Tackett and Mine Çetinkaya-Rundel

July 2020

COVID-19 has undoubtedly generated a plethora of data, from daily counts on tests, positive results, hospitalizations, deaths, etc. to data from scientific studies on the virus and treatment to data on the impact of the pandemic on the economy, education, mental health, etc. Engaging statistics and data science students with these data seems obvious, but the decision of whether to bring COVID-19 data into the statistics and data science classrooms is not simple. On one hand, the answer might seem like an obvious *yes* – what better way to engage students with real data than data about the pandemic that has turned all of our lives upside down? On the other hand, at least at the time of writing this column, we're still in the midst of a pandemic that has taken many lives and, which might be the very reason to say *no* to requiring students to engage with data about the pandemic that might have taken the life of a loved one or caused economic hardship or affected their lives in some other way that is causing stress and grief.

The goal of this column is not to provide a definite answer for whether or not educators *should* bring COVID-19 data into their classrooms, but instead to showcase a few approaches for *how* they can do so responsibly, and highlight resources that might help educators decide whether to do so in the first place.

DataFest: COVID-19 Virtual Data Challenge

The American Statistical Association (ASA) DataFest is a data analysis competition where teams of up to five students analyze a real and complex dataset over the course of one weekend. In 2019, DataFest was held at over 40 locations in the United States and internationally with more than 2,000 students participating in the event. In a typical DataFest a surprise data set is revealed to participants at a kick-off event on Friday afternoon, and students work throughout the weekend analyzing the data and deriving insights. On Sunday afternoon, groups present their work to a panel of judges made up of instructors and statistics and data science professionals in industry. By the end of the DataFest weekend, students have not only gained experience analyzing real data, they have also practiced presentation skills, all while connecting with other students, faculty members, and industry professionals.[1]

[1]For more information on ASA DataFest, see amstat.org/education/datafest.

In March 2020 as many colleges and universities transitioned to a remote format, the ASA DataFest steering committee considered alternatives for this year's competition. The goal was to adapt DataFest to the new remote environment, while still maintaining the parts of the event that make it an inviting and valuable experience for students with a wide range of data analysis experience. The 2020 DataFest was held as a virtual data challenge where students worked in teams to explore an impact of the COVID-19 global pandemic. Given the variety of potential topics, part of what made this year's challenge unique was that it involved participants finding a data set for their analysis.

DataFest events were held in April through June, a time when data and modeling about the direct health outcomes of the pandemic were rapidly changing and unreliable (see *Why It's So Freaking Hard To Make A Good COVID-19 Model*). Building models and drawing reliable conclusions about infection, mortality, or recovery rates would require participants to understand the nuances and limitations of the COVID-19 health data at a level that would likely not be feasible in the short span of the DataFest competition. Therefore, participants were advised to "tell us about something affected by the COVID-19 pandemic other than its direct health outcomes", to discourage them from presenting conclusions that could be potentially misleading or harmful.

A few suggested analysis questions included

- How has the pandemic affected the airline industry and what are some potential downstream effects of this other than economic strain on the industry?
- As a student, how would you quantify the effect of the pandemic on your education?
- With shelter-in-place / lockdown orders, many workers have started working from home, which requires internet access. How prepared was the nation / your local area for this shift?
- How has the spread of the pandemic affected people's opinion on government tracking and privacy?
- What is the effect of the social distancing / shelter-in-place / lockdown recommendations and policies on pollution?
- How can we quantify the potential effects on nutrition and general health of the public, outside of those affected by the virus?
- How are refugees affected by COVID-19?

When we suggested these potential analysis questions to students we were worried that we might be giving too much direction and curbing their creativity. Fortunately, this was not the case, and students who participated in the event came up with a wide variety of questions on their own. We provide below a sample of analysis foci from the winning teams that we hope might be inspirational for educators wanting to bring COVID-19 data into their classrooms.

- Societal impacts of the COVID-19 pandemic on education in the United States: Analysis of data from surveys conducted by the US Census Bureau's Household Pulse Survey, examining the availability of devices and internet in households with children in public or private schools in the US over a period of four weeks, 23 April - 26 May 2020. (The Data Quails - University of Edinburgh)

- Relationship between dengue fever outbreak and lockdown: Investigation of whether

the dengue fever outbreak in Singapore, which coincided with Circuit Breaker (Singapore's COVID-19 lockdown measures), could be attributed to the Circuit Breaker, or alternatively if the Circuit Breaker had worsened the dengue fever outbreak. (Team lemonchocolatecheesecake - University of Edinburgh)

- Dreams in the time of COVID-19: Exploration of Google search trends as well as sentiment analysis of tweets related to people having vivid dreams during COVID-19 outbreak. (Apoorv Jha - Duke University)
- How research priorities shift as COVID-19 Progresses: Exploration of the dataset provided as part of Kaggle's COVID-19 Open Research Dataset Challenge (CORD-19) suggesting that research focus shifted from finding a cure to preventative measures for containing COVID-19. (Team N & N - Duke University)
- Purchasing behavior via Amazon and Google Trends: Analysis of purchasing behavior data based on Amazon prices and Google Trends. (Team Maskman - UCLA)
- Driving during quarantine: Investigation of traffic data to evaluate the effectiveness of the call for social distancing in Toronto measured by the decrease in the amount of people driving in residential areas of the city. (Team Shirley Eva - University of Toronto)

The projects come from the DataFest events at the University of Edinburgh, Duke University, UCLA, and University of Toronto. In the *Further Reading* section we link to event webpages where you can find the student presentations and the datasets students put together for their analyses. The variety of foci in these projects is a testament to the feasibility of engaging students with COVID-19 data without the need for epidemiological modeling expertise. We should also note that a majority of the teams worked with data provided openly by governments, suggesting that featuring COVID-19 related data in classes might also be a good way to expose students to open government datasets.

Using COVID-19 data in the classroom

At the May 2020 Electronic Conference on Teaching Statistics (eCOTS), Laura Le, Kari Lock Morgan, and Lucy McGowan spoke on the panel *Engaging Students during the COVID-19 Health Crisis* about using data related to the pandemic in the classroom[2]. One of the primary messages from the panel was that the pedagogy should be "trauma-informed" due to the potential direct impact on students. By taking this trauma-informed approach, instructors can create a classroom environment where students feel safe to discuss the subject and reduce risk of retraumatizing students impacted by the pandemic.

The panelists shared practical ways instructors can use a trauma-informed approach when discussing this data in class:

[2]Laura Le, Kari Lock Morgan, and Lucy McGowan. "Engaging Students during the COVID-19 Health Crisis." Electronic Conference on Teaching Statistics, 20 May 2020, Online. https://www.causeweb.org/cause/ecots/ecots20/panels/2.

- Start by taking an anonymous poll asking students whether or not they want to talk about data related to the pandemic in class. If the data will be used multiple times in a semester, it is good to repeat the poll to get point-in-time feedback, since students' feelings may change as the situation around the pandemic evolves.
- Indicate in the syllabus when data about the pandemic will be used, so students know when to expect the topic to be discussed in class.
- Create an alternative assignment or discussion prompt for students who do not wish to discuss the pandemic.
- If the course is designed for a more specialized audience, such as biostatistics or graduate students, the instructor can address the fact that the topic is sensitive but is also an important area of research. This is also an opportunity to talk about strategies for maintaining a healthy relationship with emotions when doing research on sensitive topics.
- As with this year's DataFest, the analysis examples can focus on societal impacts of the pandemic other than direct health outcomes.

The panelist also suggested that we should be honest about our experience with working with COVID-19 data and, where appropriate, provide a disclaimer that we are not experts in epidemiology and infectious diseases, that we have not done an exhaustive literature review, and that we can't vouch for everyone's models and predictions.

Activity: Visualizing the effects of the pandemic

Here we outline an activity for a statistics or data science course that uses data related to the pandemic. The primary goals are for students to understand how to create effective data visualizations and the ethical considerations when creating visualizations using data that are sensitive and regularly changing. This activity is largely inspired by the following data analysis exercises: Dangerous Numbers? Teaching About Data and Statistics Using the Coronavirus Outbreak, Visualizing COVID-19 , and Cumulative deaths from COVID-19. Though the data in these three activities primarily deal with direct health outcomes, the activity outlined here can be done using data about other societal impacts of the pandemic[3].

There have been numerous data visualizations used by government organizations, news media, and other public outlets to help the general public better understand the pandemic (e.g. the "flattening the curve" plots). This has resulted in a vast collection of examples that can be used to help students think about how visualizations are used to effectively (and sometimes ineffectively) communicate insights from complex data to the public. Students are able to develop their statistical literacy as they apply what they're learning in a real-world context that is current and relevant.

[3]See examples of other datasets related to the pandemic at coronavirus-teaching-resources.netlify.app/tags/alternative-data.

Part 1: Evaluate an existing visualization

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) state that one goal for introductory statistics courses is for students to be able to "demonstrate an awareness of ethical issues associated with sound statistical practice", so the activity begins by focusing on the principles of data ethics and ethical data visualizations. The textbook *Modern Data Science with R* has a chapter dedicated to ethical data science practices and professional ethics that provides a foundation for discussing the ethical considerations of working with sensitive data. More specific to data visualizations, the post "Ethical Data Viz" on the *Teach Data Science* includes recommendations for creating ethical data visualizations and examples of how visualizations can go wrong and convey misleading information. Additionally, the post *Ten Considerations Before You Create Another Chart About COVID-19* introduces a set of criteria to consider when creating visualizations specific to the COVID-19 pandemic.

After a preliminary discussion data ethics, students evaluate a visualization that conveys information about an impact of the pandemic. Using the resources above as a foundation, students can write or discuss their responses to the following:

> ***Question 1.*** *What is the topic of the visualization? What is its primary message?*
>
> ***Question 2.*** *In what ways is it effective? In what ways is it ineffective or potentially misleading?*
>
> ***Question 3.*** *(If examining an interactive visualization) What are the benefits of displaying the data using an interactive visualization? What are the limitations?*
>
> ***Question 4.*** *How can you improve the existing visualization or display the data in a new way?*

You can either have the students find a visualization they want to examine or provide one for them. If you need inspiration, the resources we list in the *Resources for teaching* section are full of data visualizations you can pick from.

Part 2: Your turn!

Applying the principles from Part 1, students can create their own visualization or replicate and improve an existing one. Along with the visualization students can write a narrative that includes a description of the data, the primary message and interesting insights from the visualization, and ideas to further improve it.

Students can then present their visualization and narrative or share them on class-wide platform, such as a discussion forum in the course's learning management system. This can involve sharing the final visualizations only or, the entire process from getting data, tidying it to prepare it for the visualization, and the steps (code) for creating the visualization. The gallery of student work provides another opportunity for discussion about the decisions they

made while completing the assignment and the challenges of working with complex real-world data.

Resources for teaching

In this column we shared how we used this year's ASA DataFest to give students an opportunity to explore societal impacts of the pandemic, some considerations when using data related to the pandemic in the classroom, and an example classroom activity. We encourage readers to visit the following resources to find data and class activities related to the pandemic and to contribute their own.

- Teaching Statistics During the COVID-19 Health Crisis: coronavirus-teaching-resources.netlify.app
- covid19-r: Collection of analyses, packages, visualizations of COVID-19 data in R. mine-cetinkaya-rundel.github.io/covid19-r
- I Eat Data Science for Breakfast: Pandemic 2020 edition. learn.themethodsection.com/workshops/ieat/pandemic2020
- Dangerous Numbers? Teaching About Data and Statistics Using the Coronavirus Outbreak. nytimes.com/2020/02/27/learning/dangerous-numbers-teaching-about-data-and-statistics-using-the-coronavirus-outbreak.html
- Visualizing COVID-19. rpruim.github.io/ds303/S20/hw/covid-19/covid-19.html
- Cumulative deaths from COVID-19. rstudio.cloud/project/1444789

Further reading

- Abuelezam, Nadia N. "Teaching Public Health Will Never Be the Same." (2020): 976-977. Available at ajph.aphapublications.org/doi/abs/10.2105/AJPH.2020.305710.
- Baumer, Benjamin S., Daniel T. Kaplan, and Nicholas J. Horton. *Modern Data Science with R*. CRC Press, 2017. Available at beanumber.github.io/mdsr2e/.
- Carver, Robert, et al. "Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016." (2016). Available at amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf.
- Hardin, Jo "Ethical Data Viz." Teach Data Science, 6 July 2020, teachdatascience.com/ethicaldataviz.
- Koerth, Maggie, et al. "Why It's So Freaking Hard To Make A Good COVID-19 Model." FiveThirtyEight, 31 March 2020, https://53eig.ht/2WSQSCs

- Makulec, Amanda. “Ten Considerations Before You Create Another Chart About COVID-19." Nightingale, 11 March 2020, link.medium.com/cJR6WVlFX7.