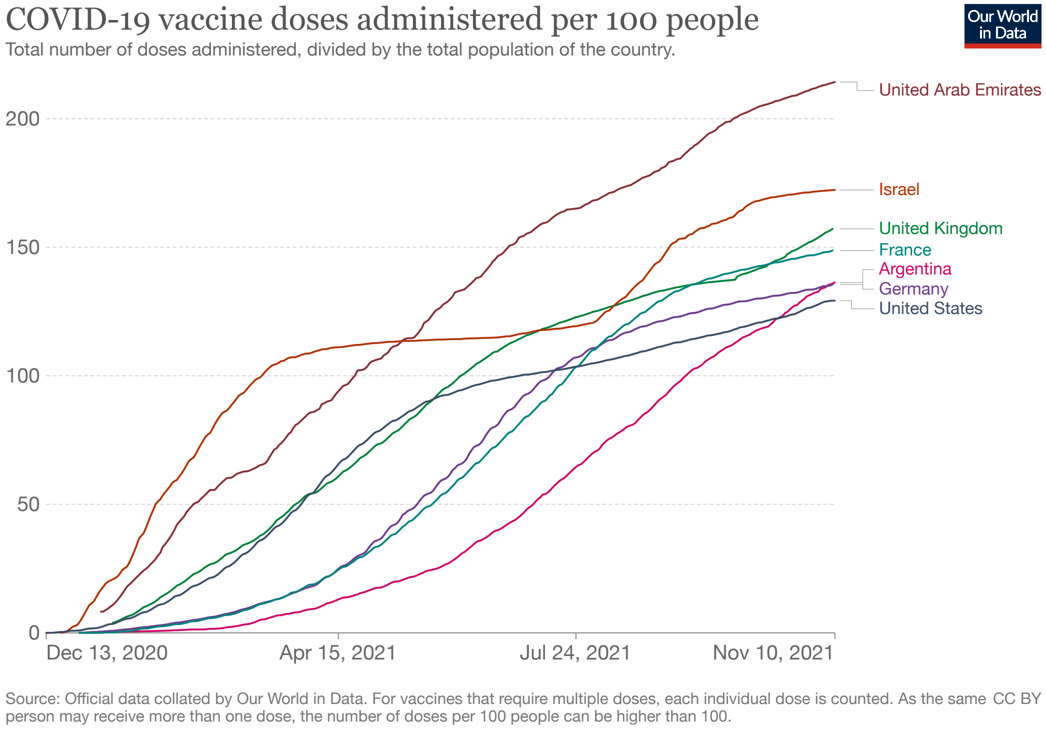
STATS 604 F21 Final Project

Introduction

The COVID-19 pandemic has indelibly altered societies and economies around the world, and caused the death of as many as 20 million people worldwide. In response to this grave and historic challenge, researchers and pharmaceutical companies mounted an unprecedented effort to develop safe and effective COVID-19 vaccines, accelerating a process that normally takes decades to just under one year. The announcement of successful clinical trials by Pfizer/BioNTech in November 2020 was celebrated as the beginning of the end of the pandemic.

Although enthusiasm for vaccination was initially high, many areas have since seen vaccination rates plateau (Figure 1). As a result, the pandemic has not abated in the way that hoped for back in spring 2021.


COVID-19 vaccine doses administered per 100 people

Total number of doses administered, divided by the total population of the country.

Source: Official data collated by Our World in Data. For vaccines that require multiple doses, each individual dose is counted. As the same person may receive more than one dose, the number of doses per 100 people can be higher than 100. CC BY

Figure 1: COVID vaccination rates over time.

Assignment

You will use statistics and data science to study the phenomenon of vaccination and vaccine hesitancy. Examples of the types of questions you could potentially address include:

- What predicts vaccine hesitancy?
- Which groups stand to lose the most/least from vaccine hesitancy?
- How much additional loss of life has there been due to vaccine hesitancy?
- What is the relationship between vaccine hesitancy and political views?
- How much sooner would we reach herd immunity if everyone were vaccinated?
- What sort of interventions are effective at reducing vaccine hesitancy?
- How/why have attitudes towards vaccines changed over time?
- Are "vaccination passports" effective?
- How is ICU availability related to vaccine hesitancy? Can one be used to predict the other?

These questions are just examples, and are intentionally open-ended and qualitative. Part of your assignment is to formulate an appropriate question that can be answered quantitatively using available data, and then do so using the techniques we have learned in this class.

Project proposal

To maximize the chance of a successful project, each group will submit a project proposal which the instructor will use to offer feedback. The proposal should include (preliminary) descriptions of:

- Your proposed research topic;
- The data sources you will use;
- The methods/models you will use; and
- The division of labor for each group member.

These are due on **Friday, November 19**, in lieu of the weekly problem set.

Submission requirements

There are two components to the final project: a written report and a presentation.

Written report (20% of overall grade) You will submit a written report in a similar format as labs 1-4. Your report should clearly document each step of the data science lifecycle, including question formulation formulation, data collection/cleaning/filtering, exploratory data analysis, predictive/inferential analysis, model checking and validation, and communication of the results. Careful attention should be paid throughout to the issues of predictability and stability that we have emphasized throughout the course. There is no minimum or maximum page requirement—your report should be sufficiently long to clearly explain what you did and why, and not longer. Judicious use of visualizations to support your claims is strongly encouraged.

Presentation (5% of overall grade) Each report must be accompanied by a twenty-minute presentation summarizing your data, methods, analysis, and results. The presentations will be delivered to the class on the evening of **Tuesday, December 14** at a time and location to be determined.

Data sources

You may use any publicly accessible data source for your project. There are a huge number of sources for COVID-19 data, and part of your assignment is to locate ones that are appropriate for answering your chosen question, and/or develop new sources. A few starting points are:

- Vaccines/vaccine hesitancy:
  - CDC
  - IHME
  - OWID
  - WHO
  - KFF
- COVID-19 data hubs:
  - https://covid19datahub.io
  - US census
  - UN
  - HHS
  - AWS
  - JHU
  - NYT
  - Schools

- GISAID (huge database of COVID genetic data)

Timeline

- Wednesday, November 10: project assigned.
- Friday, November 19: project proposal due.
- December 10, 11:59pm: final report due.
- December 14, late afternoon/evening (details TBD): group presentations.

Logistical details

- The project is due one month from today, at 11:59pm on December 10, 2021. No extensions or exceptions to this deadline will be granted, for any reason. Submit early and repeatedly.
- Your submission should be in the same format as the lab reports, and contain both a PDF and a .zip file containing all the code necessary to reproduce your analysis.
- You may collaborate in groups of up to four students. Please indicate your group membership by adding yourself to one of the groups on this Canvas page.