Your last assignment in this class is to be a partner project in which you compile and then utilize a database to investigate a topic or question of interest. You are entirely free to choose a topic and database centered around whatever types of data you'd prefer, but here are some general guidelines about each "piece":

**Database construction:** Your created database should pull from *at least two* ***different*** sources of data. Reasonably, that means it should be comprised of at least two different database tables, which should have some shared relationship between them. Note though that you are free to create as many tables as you want, potentially even from a single source of data, to help structure your database. Keep track of where exactly you got your data sources, as you should be prepared to cite them.

In addition to bringing the data into your database, you should also ensure that other good database practices are followed:

- that tables have primary keys defined
- that relationships between tables are clearly defined with foreign keys
- that indexes have been created where appropriate
- and that any other obvious constraints have been set up.

If you are searching for good data sources, some excellent places to start might include (but are certainly not limited to):

- https://www.kaggle.com/datasets
- https://data.gov/
- https://data.worldbank.org/
- https://data.fivethirtyeight.com/
- https://www.census.gov/data.html
- https://www.who.int/data/gho/
- https://data.unicef.org/

If you are willing to do a bit of programming in either Python or R, you can likely access other large repositories of information utilizing either publically provided APIs or some basic webscraping.

**Database Usage:** You should then investigate a question or topic for which your database of information may provide an answer or some insight. Again, you have total flexibility here as to what might be interesting to investigate. You should be prepared to use reasonable SQL queries to extract the information from your database, but you do not need to feel constrained to using *only* SQL for your analysis. If you are looking to do some more complicated statistical analysis of the data or visualize it in some way, bringing it into something like R or Python would make sense at some point.

# Deliverables

You will be responsible for submitting a paper describing and summarizing your database and analysis by the end of the day on April 22, 2022, which will be done by uploading the paper and any other supporting materials to your pair's GitHub repository. A decent rule-of-thumb for length should be around 6-8 pages, including visuals. You should plan to cover within your paper such topics as:

- What question/topic were you investigating?
- Where did you get your data from?
- How was the database constructed? (Entity Relationship Diagrams can be very useful here)
- How did you do your analysis to investigate your question?
- What conclusions did you draw?
- What hurdles did you encounter along the way? Or what would you change about your approach for a future project?

# Scoring

Projects will be scored on the following rough rubric, so make sure you provide evidence of the following:

- (40%) Construction of database
  - (20%) Good data sources with relations to one another
  - (10%) Primary and foreign keys defined
  - (10%) Database tidiness: good table names, constraints and indexes defined
- (40%) Analysis of database
  - (10%) Question/topic clearly stated and capable of being investigated by data
  - (20%) Analysis clearly explained and supports the fundamental question/topic
  - (10%) Conclusions easy to understand and with appropriate visuals
- (20%) Quality of paper/writing: text is well constructed, proofread, and organized, visuals are clearly labeled and explained in the text, data and references are cited.