Datasets

We've identified the following sources of data that we recommend using for your project. You are free to use other datasets if you prefer, but please take the teaching team approval first.

1. **Poverty Statistics**
    - Download link: Poverty Data
    - Source: World Bank Data

2. **Consumer Complaints**
    - Download link: Consumer Complaints (zipped csv file)
    - Source: Consumer Complaint Database

3. **USA's Consumer Price Index**
    - Download link: historicalcpi.xls
    - Source: United States Department of Agriculture Economic Research Service

4. **Indicators on Women and Men**
    - Download links:
        - Legal Age for Marriage
        - Marriages
        - Maternity leave benefits
        - Part-time employment
        - Teaching staff
        - Women legislators and managers
    - Source: United Nations Statistics Division (UNSD)

5. **Startups: Funding and Acquisitions**
    - Download link: Data on Startup Companies, Investments, and Acquisitions (zipped folder with many csv files included)
    - Source: Crunchbase

6. **Crime and Socioeconomic Indicators**
    - Download links:
        - Crimes - 2001 to present (Chicago) (Press "Export")
        - Census Data - Selected socioeconomic indicators in Chicago, 2008 - 2012
        - Crime in the United States (USA)
    - Source: City of Chicago, census.gov (via Big Data for Social Good Challenge)

7. **New York City**
    - Download links:
        - New York City Open Data
        - New York City Restaurant Inspection Results
    - Source: data.ny.gov

8. **Walmart**
    - Download links:
        - stores.csv, features.csv, train.csv (download all button)

9. **World Health**
    - Download link: Indicators - we recommend any of the datasets in the *Health* section
    - Source: World Bank Data

10. World Cup

- Download links: Players.csv, Teams.csv

11. Coronavirus Datasets

- Johns Hopkins dataset
- Kaggle datasets (includes country datasets and links to useful sites like WHO and CDC)
- EU CDC - publishes downloadable data daily
- Downloadable versions of ALL data from EU CDC (collected when it is published daily on the EU CDC website)
- State and county level data for coronavirus in the U.S. (compiled by The New York Times)
- https://www.worldometers.info/coronavirus/
- https://www.worldometers.info/
- https://covid19.who.int/table
- https://covid19.who.int/region/emro/country/eg
- https://ourworldindata.org/covid-vaccinations
- https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html
- https://www.bloomberg.com/graphics/covid-vaccine-tracker-global-distribution/
- https://www.usnews.com/news/best-countries/articles/covid-19-vaccination-rates-by-country

12. Cybersecurity

- https://www.unb.ca/cic/datasets/index.html
- https://www.stratosphereips.org/datasets-ctu13

Project Proposal

Due Date: Monday, 7/3, 2021, 11:59 PM

The main purpose of the proposal is for us to give feedback on whether the scope of the project is in the range of what we're expecting, and in cases where you plan to use a different dataset than one from the list above, whether it looks suitable and promising. On average we expect proposals to be about half-a-page long, though we know the lengths will vary. Please create a document containing the following **two** parts.

1. **Dataset**
   - State what data you plan to use -- either which one of the datasets we've suggested, or another dataset of your choosing.
   - Describe the data. As part of this, please include the total size of the dataset (e.g. number of rows) and a small sample of the data.
   - Include a link to the source of the data, and discuss any difficulties you anticipate getting the data ready for analysis.
2. **Goals**
   - Formulate a specific set of questions you want to answer, points you want to make, or issues you wish to explore through the data. Be as concrete as possible.