Title: *What is in our water?*

Abstract: The New York City Mayor's office recognizes the deep and structural history of environmental injustice in New York: https://climate.cityofnewyork.us/ejnyc-report/history-of-environmental-injustice-and-racism-in-nyc/
To this day, there is strong evidence for water inequality throughout the United States: https://www.nature.com/articles/s41467-021-23898-z - this injustice is determined by rurality, poverty, and potentially other demographic characteristics. NYC itself is a large city with a complex plumbing and water infrastructure, and a diversity of neighborhoods with varying geographical and demographic characteristics. We ask these major questions:

1. Water quality and location: How does water quality vary from neighborhood to neighborhood? Is there a geospatial pattern, such as closeness to the Hudson or being closer to Manhattan?
2. Water quality and neighborhood: Is the variation in quality correlated to features such as median income in neighborhood, population density, racial demographics, and infrastructure?
3. Water quality and health: Is the variation in quality correlated with health outcomes such as: disease incidences, pregnancy outcomes, and hospitalizations?
4. If time permits, we may include a comparative analysis with Chicago, to see if there are different patterns between the two cities.

Data:

1. Water quality: https://data.cityofnewyork.us/Environment/Drinking-Water-Quality-Distribution-Monitoring-Dat/bkwf-xfky/about_data
    a. This data has 151,000 samples over time and across sites. The sites to location mappings are provided in a separate excel file, so this data requires some processing. Furthermore, we need to map the location (x-y coordinates) to our projection of NYC and its neighborhoods. It is possible (albeit unlikely) that some peripheral neighborhoods may not have a single site, so they will not be used in our analysis.
    b. The main measurements are chlorine, fluoride, turbidity (cloudiness), E-coli, and coliform.
2. NYC Health Atlas: https://public.tableau.com/app/profile/nyc.health/viz/NewYorkCityNeighborhoodHealthAtlas/Home
    a. This data has ~100 health measures from incidences of typhoid, chlamydia, preterm birth, number of hospital visits, access to health care, and many more for 188 neighborhoods by name. The challenge here will be to map neighborhood names to location, which may require using a different dataset (amazingly, Zillow has this data here).

3. Chicago Water Quality (if time permits): https://water.epa.state.il.us/dww/JSP/NonTcrSamples.jsp?tinwsys_is_number=716257&tinwsys_st_code=IL&history=0&begin_date=&end_date=&counter=0
    a. Getting to these data is from Chicago Government's main website here and then finding Cook county and Chicago specifically. Since the City of Chicago is smaller, there are around 262 records from the past 2 years, and they too would need to be clumped into neighborhood boundaries.
    b. The goal here would be to spot major differences in distribution of water quality between the two cities, highlighting inter-city inequality along with intra-city.

Techniques: preprocessing in Python, plotting in ggplotly, ggplot2, rgdal + raster, igraph

Visualizations:

1. Geospatial mapping: we use either geoplot or geopandas depending on final data to illustrate water quality by neighborhood. We plan to use scientific papers to identify typically safe levels for chlorine, fluoride, etc. and will plot the metrics by diverging colors (red to blue).
2. We plan to use seaborn + plotly to plot neighborhood demographics such as income, race, infrastructure, health outcomes etc. and their correlations with the aforementioned quality indexes. We plan to make both of these plots interactive, so that users can select a water quality metric (on the map) and a demographic/health-related feature simultaneously and view them on the two separate plots.
3. Using our water quality metrics, we also plan to create a network on networkx and make it interactive using plotly as well. We would color the nodes by borough and then investigate if the nodes cluster/segregate by geography.
4. Depending on our findings, we also plan to include static graphs (such as scatterplots, boxplots etc.) using seaborn that show patterns of inequality.