The dataset we are using: DOHMH and DEP Waterborne Disease Risk Assessment Program Annual Report Data: Percentage of interviewed cryptosporidiosis case-patients by type of tap water exposure before disease onset, by Immune Status Link

Project Description
This project analyzes the DOHMH and DEP Waterborne Disease Risk Assessment Program Annual Report Data to investigate the relationship between tap water exposure types and cryptosporidiosis cases in immunocompromised individuals, particularly those with HIV/AIDS. We want to identify trends, predict risk factors, and provide insights for public health interventions.

Goals
1. Successfully predict the percentage of cryptosporidiosis cases for each immune status group based on tap water exposure patterns.
2. Identify the most significant tap water exposure types contributing to cryptosporidiosis cases in immunocompromised individuals.
3. Analyze trends in cryptosporidiosis cases among different immune status groups from 1995 to 2004.

Data Collection Plan
The primary dataset is already provided. After doing more research on this topic, we will collect additional data to enhance our analysis if this is needed. Some example data we might collect to enhance our prediction might includes :

1. Population demographics for the studied area: We will scrape this data from the U.S. Census Bureau website using Python libraries like BeautifulSoup or Scrapy.
2. Annual climate data for the region: This will be collected from NOAA's National Centers for Environmental Information website using similar web scraping techniques.

Modeling Plan
We think the following modeling approaches might be useful in our scenario:

1. Time Series Analysis: To analyze trends in cryptosporidiosis cases.
2. Random Forest Classifier: To identify the most important factors contributing to cryptosporidiosis cases.
3. Gradient Boosting Regressor (XGBoost): To predict the percentage of cryptosporidiosis cases for each immune status group.

Visualization Plan
We will create the following visualizations:

1. Interactive time series plots using Plotly: To show trends in cryptosporidiosis cases over time for different immune status groups.

2. Heatmaps: To visualize the correlation between tap water exposure types and cryptosporidiosis cases.
3. Stacked bar charts: To compare the distribution of tap water exposure types across immune status groups.
4. Interactive dashboard using Dash or Streamlit: To combine various visualizations for comprehensive data exploration.

Test Plan
Our test plan includes:

1. Time-based split: Train the models on data from 1995-2002 and test on data from 2003-2004.
    - We will also divide our dataset into train, validation, and test sets using a 70-10-20 ratio
2. K-fold cross-validation: Use 5-fold cross-validation on the training set to ensure the model is predicting results correctly.

Add-on based on the feedback:

Comment: make sure to detail how you'll handle any geographic alignment issues between the datasets—differences in coordinate systems or granularity can be tricky. Your use of linear regression and geospatial clustering methods makes sense, but consider mentioning advanced geospatial techniques like spatial autocorrelation for a more nuanced analysis. Validation using cross-validation and holding out city regions is a strong approach; just ensure the regions are representative of the city's diversity.

Take away:

- We need to have consistent geographic referencing within our dataset, especially the location-based tap water exposure data.
- We will have a primary dataset for the locations, and we will align any additional dataset to this primary dataset so the model is not confused
    - I think categorizing the locations using zip code might be sufficient, but there will be some manual checking to see if that is a good option. Once we decide on how to categorize the locations, we can write a script to convert all the data to use the same form of location
- Incorporate spatial autocorrelation in our project
    - Create a heat map or some sort of visualization to visualize tap water exposure and waterborne disease
    - We will use Moran's I to identify any spatial clustering of cases
- Use Chi-square test of independence to verify that the distribution of different factors in our training, validation, and test sets matches the distribution in the full dataset.
    - Account for the diversity in the dataset and see if they are being accurately represented in the dataset