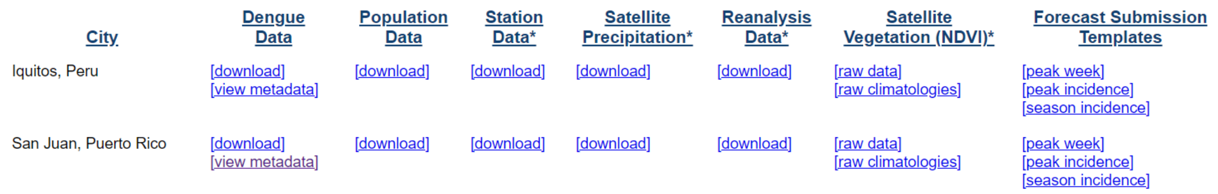
15.1 Capstone 2 Project Ideas and Proposal

Topic 1: DengAI - Predicting Disease Spread

Goal is to predict the total cases label for each (city, year, weekofyear) in the test set. There are two cities, **San Juan** and **Iquitos**, with test data for each city spanning 5 and 3 years respectively. You will make one submission that contains predictions for both cities. The data for each city have been concatenated along with a city column indicating the source: sj for San Juan and iq for Iquitos. The test set is a pure future hold-out, meaning the test data are sequential and non-overlapping with any of the training data. Throughout, missing values have been filled as NaNs.

Data Source: https://dengueforecasting.noaa.gov/



City	Dengue Data	Population Data	Station Data*	Satellite Precipitation*	Reanalysis Data*	Satellite Vegetation (NDVI)*	Forecast Submission Templates
Iquitos, Peru	[download] [view metadata]	[download]	[download]	[download]	[download]	[raw data] [raw climatologies]	[peak week] [peak incidence] [season incidence]
San Juan, Puerto Rico	[download] [view metadata]	[download]	[download]	[download]	[download]	[raw data] [raw climatologies]	[peak week] [peak incidence] [season incidence]

San Juan Dengue Data contains 988 Rows

Iquitos Dengue Data contains 468 Rows

Training Data - 24 Columns and 416 Rows

https://www.kaggle.com/datasets/qianyigang129/dengai-dataset?select=DengAI_Predicting_Disease_Spread_-_Training_Data_Features.csv

Topic 2: Heart Disease Prediction

Data Source: Acknowledgement: This data comes from the University of California Irvine's Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Heart+Disease

https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction?select=Heart_Disease_Prediction.csv

The leading cause of death in the developed world is heart disease. Therefore there needs to be work done to help prevent the risks of having a heart attack or stroke.

The above dataset will be used to predict which patients are most likely to suffer from a heart disease in the near future using the features given
Dataset contains 270 records with 14 columns.

Topic 3: Keeping it fresh: Predict Restaurant Inspections

https://www.drivendata.org/competitions/5/keeping-it-fresh-predict-restaurant-inspections/page/17/

Data Source:
https://www.yelp.com/dataset
https://drivendata.s3.amazonaws.com/data/5/public/restaurant_ids_to_yelp_ids.csv
https://drivendata.s3.amazonaws.com/data/5/public/AllViolations.csv

The City of Boston is home to thousands of restaurants and just a handful of health inspectors. Can the inspectors use the Yelp reviews that citizens generate to get a better view of active risks to public health?

The goal for this competition is to use data from social media to narrow the search for health code violations in Boston. Access the historical hygiene violation records from the City of Boston — a leader in open government data — and Yelp's consumer reviews. The challenge: Figure out the words, phrases, ratings, and patterns that predict violations, to help public health inspectors do their job better.