Project summary:

This project will consider historical and projected temperature, humidity, precipitation, population, land use and zoning to predict future grid stress at as fine spatial resolution as possible (likely by zip code / neighborhood) in a large urban area (likely new york due to availability of data). Grid stress does not necessarily scale linearly with any of these contributing factors, hence the need for training on historical relationships between these factors and corresponding grid usage.

1. Data collection (all by zipcode/neighborhood if possible, through OSF API)

   Historical (for last ~10 years or whatever is available):
   -Hourly temperature, precip, humidity levels (use a composite primary key: zipcode and date down to the hour, columns for temp, precip, humidity, zoning type)
   -As fine-grained population data as possible (Likely can't do better than every few years, but can fit a curve to this to approximate 'daily' changes) (include seasonal adjustments for tourism and 'daytime' population in commercial / industrial areas)
   -Total hourly power usage by spatial region (zipcode/neighborhood)

   Historical source data variable descriptions:

   https://www.nature.com/articles/s41597-024-04238-4/tables/1

   Future (projections for next ~10 years or whatever is available):
   -Same as Historical (except no grid usage, that is what we're trying to predict for)
   -Find future data with as fine temporal resolution as possible, temporally downscale with curve-fitting when necessary to match temporal resolution of historical data.

   Data will be collected from each coordinate dataset that corresponds to new york city, for each year, then all unioned. Currently have access to data from 550 weather stations across US, write script to only download those that correspond to a new-york burrough (coordinates are within NYC bounding polygon).

2. ETL/Database:

   Once population data is temporally downscaled to hourly, it can be joined with climate data into one table. Separate table for zoning by zipcode/neighborhood: (start with just using the 'primary' zoning type for each spatial region, this could be improved later). Table will have columns 'primary zoning type', 'zipcode/neighborhood', 'year' (no point in getting finer resolution for this). Can then be joined with the climate/population table on year and zipcode to yield a comprehensive table. Do this for both historical and future data. For historical data, can be joined with historical grid usage data. Then the future master table gets a single empty column, 'projected grid usage', that we want to predict for.

3. Analysis/ML steps (expand on this process):

   -Do some research into most appropriate models to use in this context, read into and understand mathematical justifications before proceeding.
   -Train model on the table of historical data.
   -Predict future changes.

4. Visualization step (grafana or other):

   Build a dashboard that allows trends to be visualized by desired category (could look at overall projections, projections for specific zoning types, projections by individual neighborhood/zipcode). Can sort by areas that will encounter highest stress, view likely contributors to that higher grid stress based on whether its industrial/residential, etc.

5. Documentation:

   Future improvements, current issues / oversights, theory behind particular ML model used, methods of temporal downscaling used, etc.