Data cleaning pipeline documentation:

Order to run: first run power\_outage\_data\_prep (on one’s own computer) then fasse\_data\_prep (on FASSE).

Functions contains helpers for several of the scripts in both folders.

Power outage data prep:

* This pipeline cleans the power outage data to get various measures of exposure, including daily 4+, 8+, and 12+ hour outages, number of hours without power by day and county, and some percentile-based metrics of outage that we haven’t used so far.
* WARNING: this pipeline takes a few hours (2.5) to run. The bottlenecks are b02\_expand\_outages, and the scripts that identify county-days exposed to outage in various ways. For testing, run on a subset!

Scripts and what they do:

00\_run\_all: runs the entire pipeline.

a00\_define\_continguous\_US: this produces some data files that we need to run the rest of the pipeline. This script creates files that have

1. a crosswalk of state fips codes, county fips codes, state abbreviations,

and state names for counties we wish to capture in the POUS dataset because they are in the contiguous US. This has also now been updated to include Alaska and Hawaii so contiguous is a bit wrong lol.

b) a list of counties and five-digit county fips codes that are in the contiguous US

c) a shapefile of county boundaries for those counties in the contiguous US

that we're aiming to include

These will help us filter the scripts down to the right counties every time, and also help us plot things easily for the areas we want to plot.

a01\_find\_eia\_state\_customers: This script estimates the number of electrical customers by state using the EIA data, so eventually we can generate county electrical customer estimates and coverage information, by allocating customers to county.

a02\_get\_county\_census\_cust\_est: This script uses census data to find the number of households by county and by year, and as well as the number of establishments by county and by year, which we will use to get estimates of the number of electrical customers by county, for years 2017-2020.

a03\_county\_customer\_census\_estimates: This script uses estimates from EIA of the number of electrical customers in a state, and census estimates of the number of households and establishments in counties within states, to get estimates of the number of electrical customers in each county in each state. Each county is allocated customers from EIA totals based on the percentage of households and establishments in the state that are in that county. There are a couple counties without data, in texas and hawaii. This is expected but we need to be mindful and demure of this going forward.

b01\_read\_and\_clean: This script reads in the raw power outage data and cleans the state names and county names, with the goal to assign as many accurate county FIPS codes to places referred to by text state and county names in the POUS data file. It also trims down the data to the contiguous US. It doesn't modify the data in any other way. The county fips already in POUS are unreliable so we are trying to replace them here.

b02\_expand\_to\_hourly: This script expands power outage data into a time series, from its raw form where the dataset only includes entries for changes in customers\_out (see the POUS documentation for an explanation of how the raw data is structured). It does this one county at a time, and saves each expanded county in the 'hourly county' folder in the 'data' folder. Need to clean up ordering of this script.

b03\_attach\_denoms: This script joins customer estimates from EIA to the live customer estimate data, and calculates person-coverage. Person-coverage is calculated based on two numbers. The first is the maximum number of customers recorded as ‘out’ or ‘tracked’ in the POUS dataset. The second is the estimate of the number of customers that should be in that county based on the households and number of establishments in that county, and the number of electrical customers in the state from EIA. If recorded customers in the POUS data were > 2x the EIA estimates, coverage was just reported as 1. Otherwise it’s POUS estimate/EIA estimate.

b04\_id\_outages\_continuous\_measures: This script produces a time series of days for each year and for each county, tabulating how many hours without power there were on each day for a variety of cut-points defining power outage.

b05\_identify\_daily\_binary\_exposure: This script identifies power outage events, again producing a time series of days for each year and each county, this time indicating if each day is exposed or unexposed to power outage based on a duration and a cut point. It does this for 4+, 8+, and 12+ hour outages and cut points of 0.5% of county customers without power, as well as 1%, 3%, and 5%.

b06\_identify\_outages\_percentile: This script identifies power outage events based on a percentile rather than a cut point of % of county customers without power. This is a denominator-independent metric. It classifies a county-day as exposed to outage if customers out is in the 99.9th percentile or higher for more than 1, 2, or 4 hours.

b07\_create\_exposure\_data\_for\_upload: This script combines all the exposure metrics to create an analytic dataset to upload to FASSE.