**Overall workflow**

1. Update operations data with subsequent [naatbatt](https://www.nrel.gov/transportation/li-ion-battery-supply-chain-database.html) data releases (~ every 6 months) and additional sources (as requested by Chris). Add timeline estimates for facilities that are planned or under construction.
2. Update CEJST data with [Climate and Economic Justice Screening Tool](https://screeningtool.geoplatform.gov/en/#3/33.47/-97.5) data releases (have not yet had to do this)
3. Update state data with [Oxfam](https://www.oxfamamerica.org/explore/countries/united-states/poverty-in-the-us/best-states-to-work-2022/) data releases (have not yet had to do this)
4. Monitor and complete any [Running updates](https://docs.google.com/document/d/17Y-7XkBvAOX2afRpBrRegEWvTZpcafqfsMaCOctQeDI/edit?usp=drive_link) (data cleaning/update tasks that Chris adds to this doc as needed; he will also email to make sure you see them)
5. Add updated data to [map](https://ucsc.maps.arcgis.com/home/item.html?id=557b51c627a94344902c25c9baaa5627)
   1. Operations data >> *Operations* feature layer
   2. State data >> *Labor Policy Indicators* group layer
   3. CEJST data >> *Climate & Economic Justice Screen Tool* group layer
6. Find, clean, and add new layers for map (as requested by Chris)

**Operations data**

Files:

* [lithiumSupplyChainMapData](https://docs.google.com/spreadsheets/d/1m5g9kwD8Js02a1Dx_063WJxkUQBjQcL8cOuPzlcwZ-s/edit?usp=drive_link)(final operations data)
* operations >> [lithiumSupplyChainMapData\_update](https://docs.google.com/spreadsheets/d/1auDwhc6o1NOX9W4V87LBkSOnO6XDbLb_nc_r_Oqhtvc/edit?usp=drive_link) (final output of [lithiumSupplyChainMapData\_wrangling.R](https://drive.google.com/file/d/1KWIyPyU5cIg15CUfULx_r_mFUbgZzpYS/view?usp=drive_link))
* [lithiumSupplyChainMapData\_wrangling.R](https://drive.google.com/file/d/1KWIyPyU5cIg15CUfULx_r_mFUbgZzpYS/view?usp=drive_link) (R script for initial cleaning and subsequent updates)
* ​​[operations\_naatbatt\_June2023.xlsx](https://docs.google.com/spreadsheets/d/177RA4FMdUMP-vcJNhh_wJy03n3m6oMqK/edit?usp=drive_link&ouid=111513418839683461668&rtpof=true&sd=true) (June 2023 [naatbatt](https://www.nrel.gov/transportation/li-ion-battery-supply-chain-database.html) data)
* [operations\_chargedMap\_June2023.xlsx](https://docs.google.com/spreadsheets/d/1GX3GQMhtkNK4THqyc04_ByoCpGTn8LNy/edit?usp=drive_link&ouid=111513418839683461668&rtpof=true&sd=true) (June 2023 [Charged](https://www.charged-the-book.com/na-ev-supply-chain-map) data)
* [operations\_naatbatt\_Dec2022.xlsx](https://docs.google.com/spreadsheets/d/19VTCNLq32t4d56yGK0nAJn0qxDn2ldgp/edit?usp=drive_link&ouid=111513418839683461668&rtpof=true&sd=true) (Dec 2022 naatbatt data)
* [operations\_EV](https://docs.google.com/spreadsheets/d/1wtgvgRcimpzmFBgLzR5tEbKqD_Sb_o2zpf3scEeVZRE/edit?usp=drive_link) (Aug 2022 EV data compiled by Chris’s team)
* [operations\_naatbatt\_Dec2021.xlsx](https://docs.google.com/spreadsheets/d/1c9gfglUA2wgBZiF47e7ehivCZty-ahhV/edit?usp=drive_link&ouid=111513418839683461668&rtpof=true&sd=true) (Dec 2021 naatbatt data)

Prior work:

* Processing raw naatbatt data. Challenges:
  + did not use consistent field names across supply chain segments
  + created separate entries for separate products coming from a single facility, even in the same supply chain segment (we want separate entries for facilities operating at multiple supply chain segments, but there should only be one entry per facility per supply chain segment)
  + did not standardize website URL data (i.e., many different URL formats),
  + was inconsistent in breakdown of workforce estimates (i.e., if a given facility had multiple entries, it sometimes split the overall workforce across the multiple entries, and sometimes used the total workforce for every entry)
  + did not include a timeline field for planned or under-construction facilities
  + had a number of errors (e.g., flipped lat/long coordinates, typos)
* Updating data with new naatbatt data releases. Challenges:
  + have not used consistent entry ID’s across data releases (\*they seemingly reassign them in alphabetical order with every release)
  + have changed some field names across data releases
  + have not always update the quality check date field after updating a given entry (e.g., company or facility name, status, workforce, supply chain segment)
  + can use QC date field to add new/updated entries, but *replacing* old entries with updated entries is tricky given inconsistent ID’s, company/facility names, supply chain segments, and our own manual updates/tweaks and addition of timeline data.
* Updating data with additional entries from other sources (e.g., Charged map, DOE grant announcements). Challenges:
  + Manually finding data to complete fields for added entries (e.g., lat/long, workforce, timeline)
  + Inconsistent company/facility names, supply chain segments, and workforce estimates between naatbatt and other sources (i.e., check for duplicates)
  + Future naatbatt data releases including the added entries (i.e., check for duplicates)

Suggestions:

* Further automate and streamline naatbatt updates. Given the complications listed above (especially the changing entry ID’s), some options I see:
  + Continue to use/update existing dataset, with some manual work needed for every data release:
    - Rewrite initial cleaning and update scripts to be more flexible/robust given all we know now. As needed, tweak the update script for every new data release to account for changing field names, etc. Add new entries based on QC Date field, and manually comb through updated dataset to remove duplicates. Update QC Date and Team Member fields to track manual changes from our team.
    - Pros: Data is cleanest and most accurate
    - Cons: Most time-intensive
  + Create a new dataset from the most recent naatbatt data release, do not manually tweak/fix any fields or entries, and build on it with every new data release. Use a combination of QC Date and Company/Facility Name to automate data updates.
    - Pros: Less time-intensive.
    - Cons: More errors, duplications, inconsistencies, etc. Must still track down timeline data.
  + Rebuild and automate workflow from the ground-up, with a new operations dataset created with every data release (vs. new entries identified and added onto existing dataset):
    - Rewrite script to clean every new naatbatt data release and combine it with our alternative entry data (e.g., from Charged)
    - Pros: Most time-efficient
    - Cons: Most errors, duplications, inconsistencies, etc. Must still track down timeline data.

**CEJST data**

Files:

* [cejst\_final](https://docs.google.com/spreadsheets/d/1bSV13a_90VdYHx3rRIQl1iqZkB-J33s7vrjMyIg_kg4/edit?usp=drive_link)(final CEJST data)
* [cejst\_wrangling.R](https://drive.google.com/file/d/1MJRdk0fxvbK8H6QYdrBSCodJHpmTKM3D/view?usp=drive_link) (R script for initial cleaning)
* ACS data had to be brought in separately, as CEJST census data race variables were not granular enough (i.e., we wanted race by latinx identification). used ACS 2019 as that is what most recent CEJST used.
* Also brought in data directly from CDC dataset to calculate asthma/heart/diabetes rates, as the numbers in the CEJST appeared to be errors (way too high)

**State data**

Files:

* [state\_final](https://docs.google.com/spreadsheets/d/1pSdYl70jLZOvBi0xn3bjUHNQ0L2bc5zwrKZPGCj9oUg/edit?usp=drive_link)(final state data)
* [state\_wrangling.R](https://drive.google.com/file/d/15VDJURGPQ-7eRrm1DYso3zgi9z_xn5-x/view?usp=drive_link) (R script for initial cleaning)
* Combined 2022 Oxfam data with DOL minimum wage data and BEA GDP data