On Cleaning Data for SCRIPT

Proposed cleaning process:

1. Receive new files
2. Clean new files
3. Break up new cleaned files into the same smaller groups (eg. by zipcode or driver ID) as previous data
4. Add/append the new data pieces into those files in S3, avoiding reopening the massive original files or having to reclean all the old data each time

Cleaning process implemented for recent Residential Data:

\*In folders “SCRIPT/Cleaning/Residential/\*

* Unite.ipynb
  + Opens different time periods, ie. original data dump + 5 more months + 1 more month, and combines them into one file
  + Does this for both the intervals and sessions data
* Clean\_Intervals.ipynb and Clean\_Sessions.ipynb
  + Cleans the data in each file
  + Cleans based on power values, adds new columns in sessions related to date\_time, and also cleans for things like charger type, time zone, driver country, and currency
* Join\_Session\_Interval.ipynb
  + Makes sure the sessions are all paired, i.e. no sessions appear with intervals in the intervals file which don’t have a match in the sessions file, etc
  + Saved outputs e.g. in bucket 'script.chargepoint.data' with path 'Residential\_Data/clean/sessions\_paired.csv' are the full final cleaned data
* Reduce\_Into\_Smaller\_Groups.ipynb
  + Split up the large files e.g. by zipcode to make them easier to work with in the analysis, all saved to S3
  + Also includes some late cleaning of the zipcode column, it would be good to have that earlier with the rest of the cleaning

Cleaning process for commercial data:

* Oskar’s scripts are in SCRIPT/Cleaning
* The Residential scripts above are based on his and use his functions, but all the file names and locations are different. I also tried to document which of his cleaning steps I took out in the comments in the corresponding residential files
* I would recommend using the scripts in SCRIPT/Cleaning/Residential as a starting point, then pulling pieces from SCRIPT/Cleaning to fill-in a couple of other cleaning steps specific to the commercial set