# Related Documents

* Note – these docs will ultimately have copies in this repository, once they’re finalized
* [Teams folder for EI 2020](https://teams.microsoft.com/_#/files/Baseyear%20Update?groupId=cef7cf03-9cde-4ad6-9e24-0c524a2346fa&threadId=19%3A545e292551414831967a9b043fc0d41f%40thread.skype&ctx=channel&context=2020%2520employment%2520inventory%2520work&rootfolder=%252Fsites%252FDataDevelopmentMonitoringandSupport%252FShared%2520Documents%252FBaseyear%2520Update%252F2020%2520employment%2520inventory%2520work)
* Raw data from Pete – “P:\Employment Inventory\Employment 2020\Data Axle Raw - DO NOT MODIFY”
* Emp Inventory GIS work - [I:\Projects\Darren\EmpInventory](file:///I:\Projects\Darren\EmpInventory)
* [Shengyi’s python scripting and process notes on Sharepoint](https://sacog.sharepoint.com/:w:/s/DataDevelopmentMonitoringandSupport/EbZGyvTm-eZKpUJwqdZ6VQIBmhPm5z-2YPc__zOkUZlQYA?e=vSI5UO)

# Dealing with Home-Based Businesses

* If HOME = 1, it means it’s a home-based business—what does it mean? Are we including these?

# Exploring the 2020 Data Axle Raw Data Set

## Guidelines for quality

* Consider ignoring records with fewer than 5 employees, i.e., not removing or flagging for QA

## Types of Duplication

* Multiple rows with ALL of the following qualities:
  + Same or very similar company name,
    - Use fuzzywuzzy library
  + Same address or very similar address
  + Same lat-long
    - \*\*\*if the lat-long is either parcel or entry point level of granularity
      * Could there be duplication where same biz has point for both a parcel point and an entrance point? Or where the same biz has a point at multiple geolevels? I.e., same name, same address, but multiple geo levels?>>>no this doesn’t appear to be an issue.
  + More than 0 employees
    - Can’t we just eliminate rows with no employees? Tina has concerns about it.
  + Valid NAICS code (not 9999904)
* Locational duplication
  + Could same address have multiple lat-longs?
  + Or single lat-long has multiple addresses associated with it

### Possible causes of duplication

* Old businesses wrongly left in (toss)
* Individual employees within the business are listed as separate businesses (keep)
* Other error/noise in the data set (depends)

### Potential resolutions to duplication

* Getting total emp at each unique lat-long?
  + Appears to be, at least in part, the approach that Shengyi took in 2016
    - Make table with distinct businesses at each lat/long
  + Problems with this approach
    - Many lat-longs are for ZIP code (though definitely not the majority)
* Cannot use SITE variable
  + SITE value is not present at majority of location, even when they are a distinct parcel
* Businesses that no longer exist
* What about getting total emp by NAICS code, lat, and long for all locnums where emp > 0 and geo\_level was parcel (P) or entry point (0)?
  + Issues:
    - Still have many potential obsolete businesses that are no longer at location
    - Omits employees that were tagged to a ZIP code centroid rather than parcel.
    - Same company may have duplicate record at same location.

## Hard to locate sites

* No valid address (doesn’t start with a number), no site-level lat-long,

# Making a list of duplicate entries

* Might help to have a list of all duplicate entries with fields:
  + Company name
  + Address
  + ZIP
  + NAICS
  + Emp count
  + Date added/modified
* To build the list:
  + Go through each row (again, maybe omit all rows with zero employees right off—but still need to know from Tina why this can’t be done)
  + For each row in master list
    - Select all rows in dataframe with following relationship with the row:
      * Have high fuzzywuzzy score with coname
      * High fuzzywuzzy score with address
      * Address numbers match
      * ZIPs match
    - If more than 2 rows selected (including the row being iterated on), take the selected rows (confirm that they have the comparison row included) and append them to a new dataframe, along with a dupe\_id value
* Result will be a list of all locnums that have at least one duplicate value
* Can add dupe\_id values to the master table with a join based on locnum
* Resulting master table will have a dupe\_id for all locnums that are potential duplicates.
  + If a locnum does NOT have a dupe\_id, it could mean:
    - The record still shares a location with other locnum, but is a distinct and valid company (e.g. 76 and Spinners occupying same land)
    - The record is still invalid/duplicate because it is for a business that formerly occupied the site (e.g. Ford’s burgers used to be where Dali’s is now).
    - The record is unique and distinct (Yay!)
* Problems:
  + Will get lots of confusing, multiple dupe\_ids for same business, e.g.
    - If row1 = “Kaiser Permanente” and row2 = “Kaiser Permanente Health”, then row1 will put 2 the 2 records into the dupe table, but so will row2, so you’ll have 4 rows put into the dupe table, but should only have put 2 in.
      * Could perhaps “extract” the rows so that they don’t get passed again? Risk (using same example): if there’s a row3 that has “J P Kaiser Law”, then that could get tossed in with the “Kaiser” dupe ID.

## Maybe working just with lat, long, and coname?

## Search for dupes at the location level

* Lat-longs MUST be at the parcel/address level—for the ~7,000 records not at a parcel level, will need to try and geocode their addresses, then run this process again.
* Add NAICS4 field, which is the 4-letter NAICS code—use this instead of full NAICS value.
* Add a “dupe\_flag” field to master df
  + Logic/rationale: something can only be a duplicate if it has the same company name and is at the same location.
* Possible duplicate flags to make:
  + See the check\_flags.csv
  + For each record, ALL non-null check flag values assume all the following conditions are met:
    - High match or perfect match with at least one other record on company name
    - Same lat/longs as at least one other record
* Add “latlong ID” field of ints, each int represents a unique lat-long pair (about 146,000 in the data set)
* For each ID in the latlong IDs:
  + Select all records in master that have that latlong UID
  + If only one record, set dupe\_flag = ‘0’ (in theory it’s a unique and valid business) and go to the next record
  + If > 1 record at the location, then for each record at the location:
    - If the record’s dupe\_flag value is blank/null,
      * compare the record’s company name against all other company names in the selection, doing the following at each check:
        + If fuzzy ratio above X (maybe 90), it means it’s a possible dupe of the check value, so:

If zero employees: set dupe\_flag = “DZE”

If >0 employees:

If NAICS = 9999, dupe\_flag = “DNN”

If NAICS != 9999:

If NAICS does not match NAICS code of any other records with same company name at same lat-long, dupe\_flag = “DUN”

If NAICS matches (in addition to lat-long and company name), dupe\_flag = “DMN” >> these are the lowest-hanging fruit, i.e. almost certainly duplicates (matching name, location, and NAICS code).

* + - * + If fuzzy ratio below X, assume it’s a unique and valid business
    - If the record’s dupe\_flag is not blank or null, then skip it because it’s already been marked as a dupe of a previously checked value.
* \*\*Possible pitfalls of this approach:
  + First and foremost, requires doing another run-through to de-dupe records that only have ZIP-code granularity. For these cases, may want to use Shengyi’s address-based system.
  + Does not tag whether a business is obsolete or no longer actually there
  + Some entities (e.g. company “Ucd School of Medicine”) have what appear to be many duplicates, but by looking at the employment totals, it seems more reasonable to keep all the duplicates than to ignore them.
    - Also for weird ones like the UCD med school one, these might be replaced by some of the supplemental data sources that Tina mentioned.
  + For some cases, employees of a bigger company may be listed as their own business (e.g. for UCD med center, individual doctors are listed as separate records, but they may also be counted under total employment for one or more of the records corresponding to UCDMC)
    - How many case are there like this? How much are we likely to overcount if we just say “okay it’s fine if these are kept as separate businesses even if they get double counted”?
  + In rare cases (how many?), one unique lat-long pair could have more than one address affiliated with it. Worst case example is if company X has 2 locations at a latlong pair, but those are two legitimately different locations because they have different addresses.

### Notes on fuzzywuzzy thresholding

* Fuzz.ratio goes from 0 (no similarity) to 100 (strings match exactly)
* Fuzz partial ratio?

## Tests and Questions for any method

* What does total employment look like for 2016 vs. 2020? How does this difference vary at smaller geographic levels?