Employment Inventory Tagging Overview

Purpose of “tagging”

- Make it easier to identify and remove wrong Data Axle records by:
  - Flagging records that are likely duplicates
  - Checking if record had matching record in 2016
  - If record is considered “verified” by data axle
  - Has match in a supplementary data source (e.g. schools)

The MOST important tag

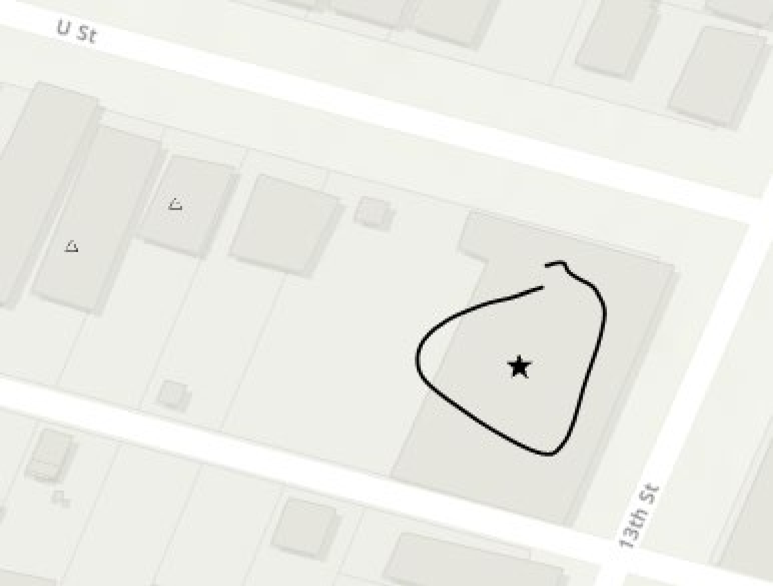
- FULTYP
- Discovered late in process
- The FIRST tag that anyone should look at when cleaning data
- Indicates if a given record has been verified, not verified, is suspect, etc.
- If ever in doubt about any set, always check FULTYP

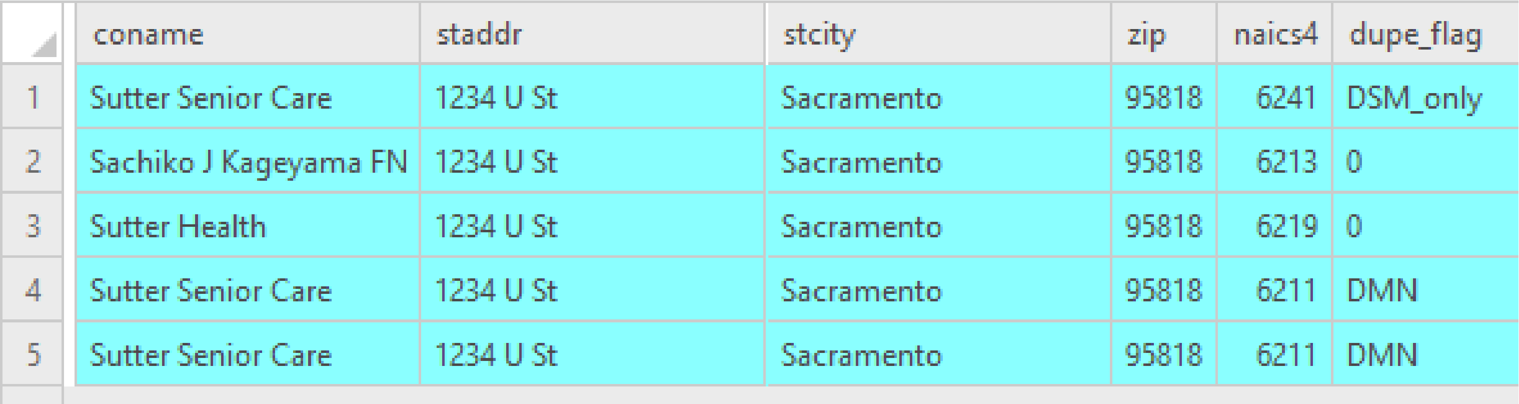
Duplicate identification process

- Link to flow chart

Duplicate flag, example 1

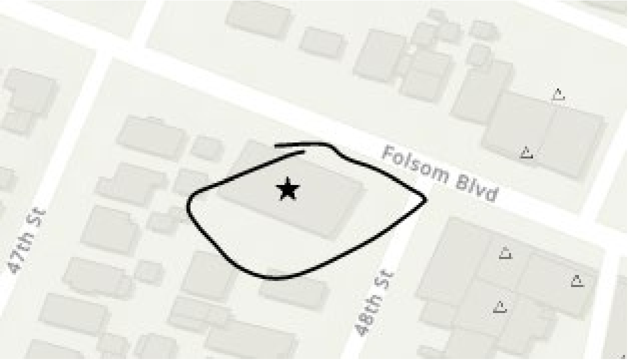
- Purpose: flag instances in which the same business is listed twice

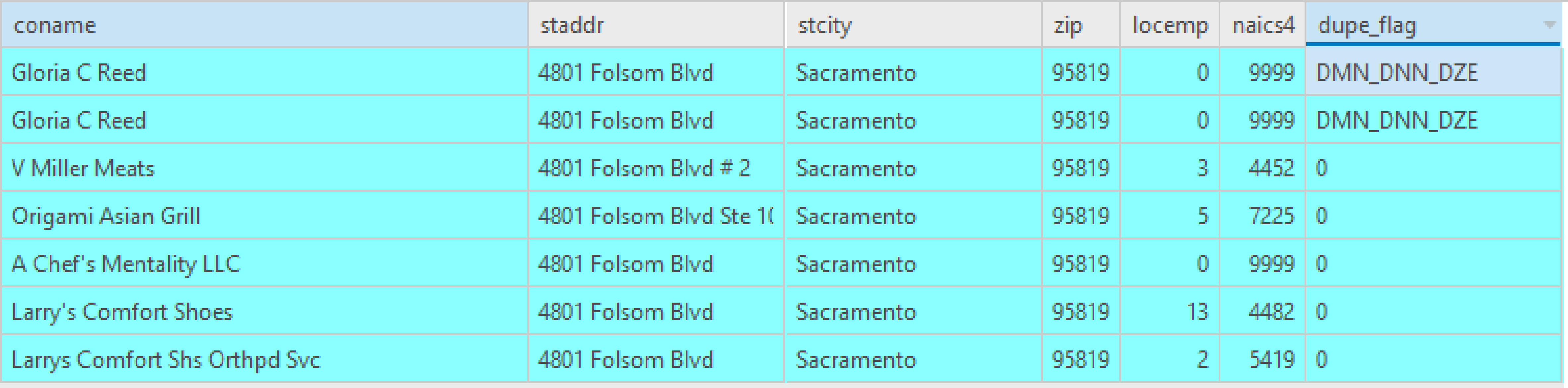




	coname	staddr	stcity	zip	naics4	dupe_flag
1	Sutter Senior Care	1234 U St	Sacramento	95818	6241	DSM_only
2	Sachiko J Kageyama FN	1234 U St	Sacramento	95818	6213	0
3	Sutter Health	1234 U St	Sacramento	95818	6219	0
4	Sutter Senior Care	1234 U St	Sacramento	95818	6211	DMN
5	Sutter Senior Care	1234 U St	Sacramento	95818	6211	DMN

Duplicate flag, example 2

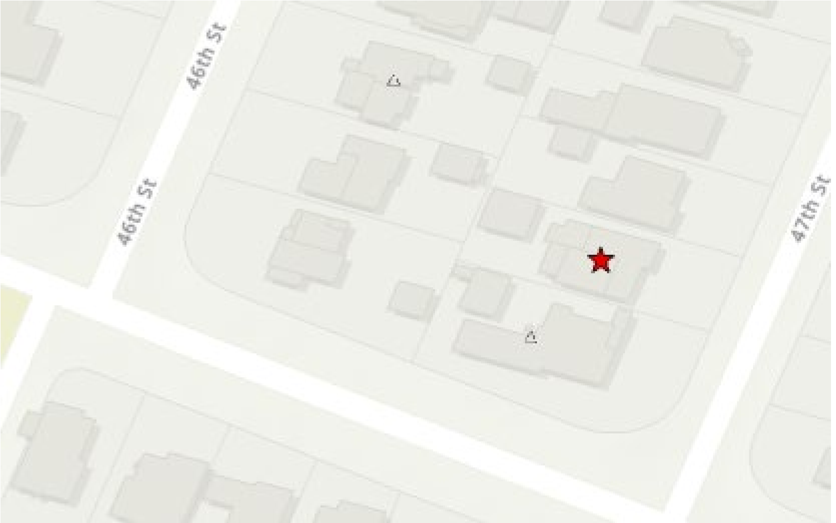


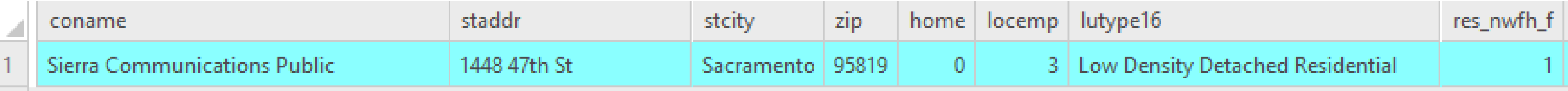


coname	staddr	stcity	zip	locemp	naics4	dupe_flag
Gloria C Reed	4801 Folsom Blvd	Sacramento	95819	0	9999	DMN_DNN_DZE
Gloria C Reed	4801 Folsom Blvd	Sacramento	95819	0	9999	DMN_DNN_DZE
V Miller Meats	4801 Folsom Blvd # 2	Sacramento	95819	3	4452	0
Origami Asian Grill	4801 Folsom Blvd Ste 1(	Sacramento	95819	5	7225	0
A Chef's Mentality LLC	4801 Folsom Blvd	Sacramento	95819	0	9999	0
Larry's Comfort Shoes	4801 Folsom Blvd	Sacramento	95819	13	4482	0
Larrys Comfort Shs Orthpd Svc	4801 Folsom Blvd	Sacramento	95819	2	5419	0

Add land use flag

- Combined with "Home" flag, identify records that are not home-based businesses but are located in residential areas



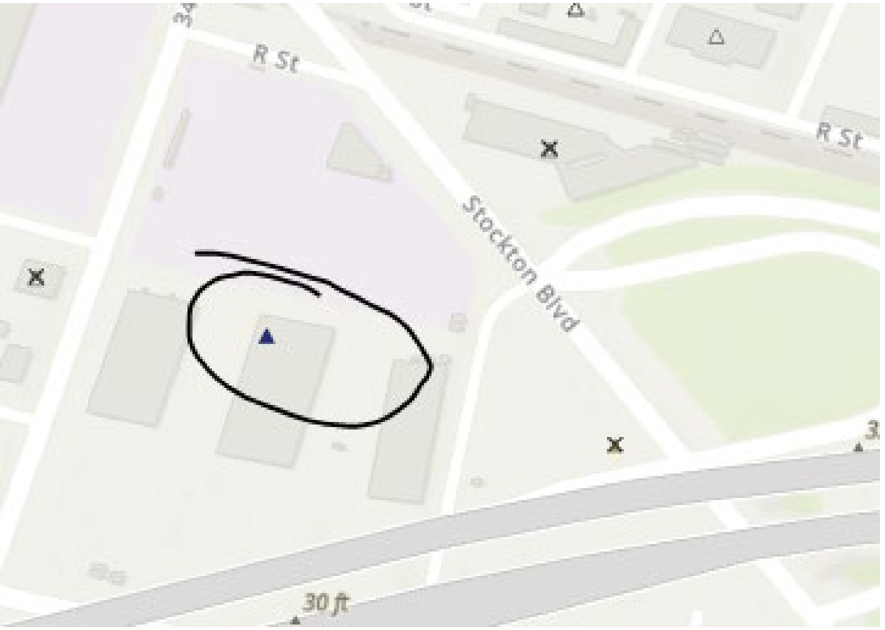


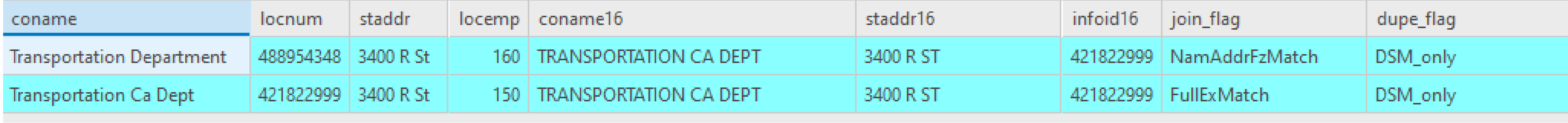
	coname	staddr	stcity	zip	home	locemp	lutype16	res_nwfh_f
1	Sierra Communications Public	1448 47th St	Sacramento	95819	0	3	Low Density Detached Residential	1

Matching to 2016 data

- VERY imperfect process
  - Businesses come and go, many businesses in 2016 no longer exist, change names, or new businesses come along
- But can help choose which duplicate is the “real” one
  - E.g. if two records are duplicates, pick the one that has a matching record in 2016

Example: Matching to 2016 data



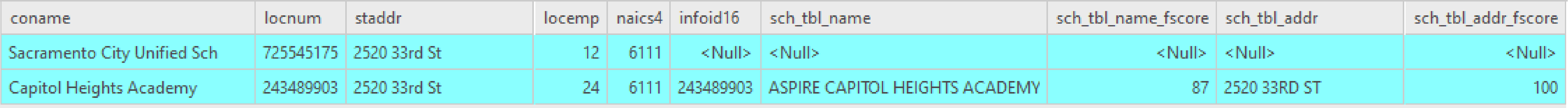


coname	locnum	staddr	locemp	coname16	staddr16	infoid16	join_flag	dupe_flag
Transportation Department	488954348	3400 R St	160	TRANSPORTATION CA DEPT	3400 R ST	421822999	NamAddrFzMatch	DSM_only
Transportation Ca Dept	421822999	3400 R St	150	TRANSPORTATION CA DEPT	3400 R ST	421822999	FullExMatch	DSM_only

School tag

- Checks if a Data Axle school record also appears in a school inventory sheet.

Y St
X St Y St Al
33rd St
e Al



coname	locnum	staddr	locemp	naics4	infoid16	sch_tbl_name	sch_tbl_name_fscore	sch_tbl_addr	sch_tbl_addr_fscore
Sacramento City Unified Sch	725545175	2520 33rd St	12	6111	<Null>	<Null>	<Null>	<Null>	<Null>
Capitol Heights Academy	243489903	2520 33rd St	24	6111	243489903	ASPIRE CAPITOL HEIGHTS ACADEMY	87	2520 33RD ST	100

How can this help clean???

Possible workflow

1. Figure out what the protocols will be for filtering by FULTYP
2. Assume that all records with ‘0’ for dupe_flag value are not duplicates, at least to start
3. Be very wary of any records with “res_nwfh_f” = 1
4. 2016 data can help figure out which duplicate to keep (but again, FIRST check FULTYP value for duplicates)

More information

- ReadMe that I’m maintaining is on GitHub (Click here to access, no login required)
- Script does NOT remove any records. In only adds columns that help people who are working to remove records.

Questions?