A Crash Course in Deterministic Matching

Presented by Lauren Chenarides

Evidence-Based Policymaking for Applied Economists

Day 2: October 8, 2021, *12-5 pm Eastern*

Activity Overview

Objective: ***Leave with the confidence, and practical tools, to link two datasets based on deterministic matching strategies***

Recall: What is deterministic matching?

Deterministic matching strategies

Use case: *What is the effect of food pantry presence on food retailer donation volumes?*

Simple shortcuts

- Standardizing variables
- Geocoding addresses

Learning objectives

- **Define** deterministic matching, and provide examples.
- **Describe** the steps involved in matching two datasets that share a common identifier.
- **Refine** two datasets that require preprocessing.
- **Apply** the tools to your own research problem.

Required materials

- Two linkable datasets that share a common identifier
- Access to Stata software
  - I’m using version 17, but these can be used in earlier versions
  - Similar packages and code exist in R and SAS
- Admin rights to your computer
  - You’ll need to install some packages
- An API key

Two forms of data linkages

Deterministic

- Looks for an exact match between two pieces of data
  - Two datasets share a common identifier (e.g., SSN, firm, ISBN, SKU, UPC)
  - The researcher can link the two datasets directly by using the shared identifier

Probabilistic

- No common identifier
- Datasets share *similar* identifying features that are not unique (e.g., age, race, gender, # employees, industry classification)
- The researcher can calculate a probability that the same person (or firm) from one dataset is represented in another dataset

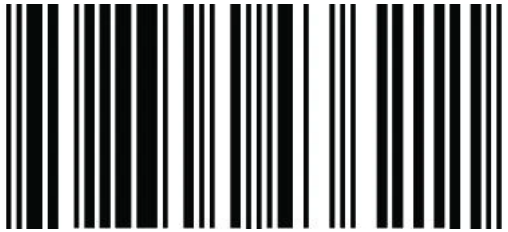
Examples of deterministic matching

Deterministic

- Looks for an exact match between two pieces of data
  - Two datasets share a common identifier (e.g., SSN, firm, ISBN, SKU, UPC)
  - The researcher can link the two datasets directly by using the shared identifier

- Matching transaction level data with product characteristic data via a Universal Product Code (UPC)
- Matching geographical information with local area average income via County FIPS code
- Matching nearest water supply point sources to disease outbreaks

A perfect world



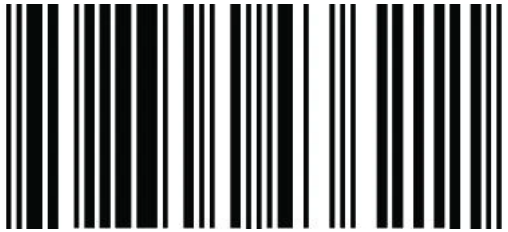

```
merge m:1 UPC using
"product_dictionary.dta"
```

Result	Number of obs	
Not matched	0	
Matched	62,647	(_merge==3)

	UPC	PrivateLabel	Manufacturer	ProductDescription
56683	855564003079	Branded	Way Better Snacks	Simply Sprouted Way Better Snacks Simply Sunny Mu…
56684	855564003093	Branded	Way Better Snacks	Simply Sprouted Way Better Snacks Simply Sweet Po…
56685	5013351042129	Branded	Two Birds Cereals	Two Birds Blueberry & Acai Super Seeds Breakfast …
56686	3421331514016	Branded	Marie de Livinhac	La Vérité Bio Purée aux Légumes Oubliés Bio (Orga…
56687	9332045000419	Branded	Bellamy's Organic	Bellamy's Organic Tao Say Huu Co (Apple Snacks) a…
56688	857245003514	Branded	OMG! Food Company	OMG! Organic Turmeric Powder is said to be one of…
56689	898999010175	Branded	All Market Europe	Vita Coco Lait de Coco Light (Light Coconut Milk)…
56690	5060406460391	Branded	Life Health Foods	Gluten Free Nutri-Brex Coconut & Crispy Rice Whol…
56691	4710887945986	Branded	Taiwan Smile Food	Pop-Smile Let's Party! Merry Christmas Popcorn Pa…
56692	751320721686	Branded	Alimentos Da Vila	Da Vila Pão de Cacau (Cocoa Bread) is free from l…
56693	5030343832681	Branded	Belazu Ingredient	Belazu Ingredient Co. Single Estate Verdemanda Ex…
56694	5060594060069	Branded	Poptails & Dreams	BioPop Sorbets Passion Ananas Fruit de la Passion…
56695	8719189106176	Branded	Boon-Foodconcepts	Boon Burger Chili (Chilli Burger) is free from so…
56696	7897534838390	Branded	Duprata Alimentos	Natufruit Linha Saúde e Bem Estar Barra de Banana…
56697	69593120816	Branded	La Maison Orphée	Maison Orphée Vegan Caesar Vinaigrette is made wi…
56698	609722910828	Branded	Kitz Living Foods	Kitz Living Foods Fresh Herb Dulse & Spirulina Cr…
56699	4011800181017	Branded	Schwartauer Werke	Schwartau Samt Erdbeer-Rote Johannisbeere Fruchta…
56700	4260133141063	Branded	Beltane Naturkost	Beltane Bio Fix Paprika-Hähnchen in Ungarischer R…
56701	5281117007686	Branded	Al Wadi Al-Akhdar	Alwadi Hummus Tahina Kikärtsdip (Hummus Tahini Ch…

Match score = 100%

Reality…



??

```
merge m:1 UPC using
"product_dictionary.dta"
```

Result	Number of obs	
Not matched	18,225	
from master	18,225	(_merge==1)
from using	0	(_merge==2)
Matched	62,647	(_merge==3)

Match score = 77%

	UPC	PrivateLabel	Manufacturer	ProductDescription
53790	8717853492457	Branded	Hari's Treasure/La …	Hari Crunchy Crunchy Müsli Himbeere & Kokosnuss (…
53791	8717853492624	Branded	Hari Tea / La Alter…	Hari Crunchy Cereal Arándano Cardamomo Coco (Exot…
53792	8717853492655	Branded	Hari Tea / La Alter…	Hari Crunchy Cereal Frambuesa Coco (Exotic Wild R…
53793	8717948490214	Branded	De Zuivelmaatschapp…	Creamy Moments Yoghurt Kokos Naturel (Natural Coc…
53794	8717953103901	Branded	Yogi & Yousef	Yogi & Yousef's 100% Natuurlijke Dadels (100% Nat…
53795	8717953103925	Branded	Yogi & Yousef	Yogi & Yousef's 100% Natural Dates are described …
53796	8717953127495	Branded	Clean Foods	Clean Foods RawPasta Spaghetti (Spaghetti) is now…
53797	8717953155078	Branded	Chocodelic	Choco Delic Rauwe Chocolade Reep met Hazelnoten (…
53798	8717953190406	Branded	ProViand	ProViand Gemüse Filet Streifen Rindfleisch-Art (B…
53799	8717953190413	Branded	ProViand	ProViand Gemüse Filet Streifen Hähnchen Art (Chic…
53800	8717953190420	Branded	ProViand	Proviand Bacon Art Gemüse Filet Würfel (Bacon Sty…
53801	8717953204134	Branded	Rose&Vanilla	Rose&Vanilla Appeltaart (Apple Pie) is gluten- an…
53802	8717953204141	Branded	Rose&Vanilla	Rose&Vanilla Walnoten Brownie (Walnut Brownie) is…
53803	8717953204172	Branded	Rose&Vanilla	Rose&Vanilla Wortel Taart (Carrot Cake) is now av…
53804	8717953228437	Branded	Orangefit	Orangefit Choco FitBar + Proteïne (Choco Fit Bar …
53805	8717953254603	Branded	Seamore	Seamore I Sea Pasta Vild Økologisk Havtang (Wild …
53806	8717953254641	Branded	Seamore	Seamore I Sea Bacon Alghe Selvagge Biologique (Wi…
53807	8717953254696	Branded	Seamore	Seamore I Sea Wraps Tortilha de Algas Marinhas (S…
53808	8717953282200	Branded	Choco & Things	Choco and Things Raw Organic Chocolate Nibs King …

Goal: Improve match scores

It is rare that the “linkability” between two datasets is 100%.

- What happens when two datasets share the same identifier (e.g., store name), but the identifiers are not identical (e.g., “Dollar Tree” vs. “DLR TREE”)?
- What happens when two datasets have identical identifiers (e.g., latitude and longitude), but they produce false matches?

Deterministic matching (DM) requires establishing a process so you can **increase the match scores** between two different datasets.

How linkable are your datasets?

DM is ideal if your source data collect:

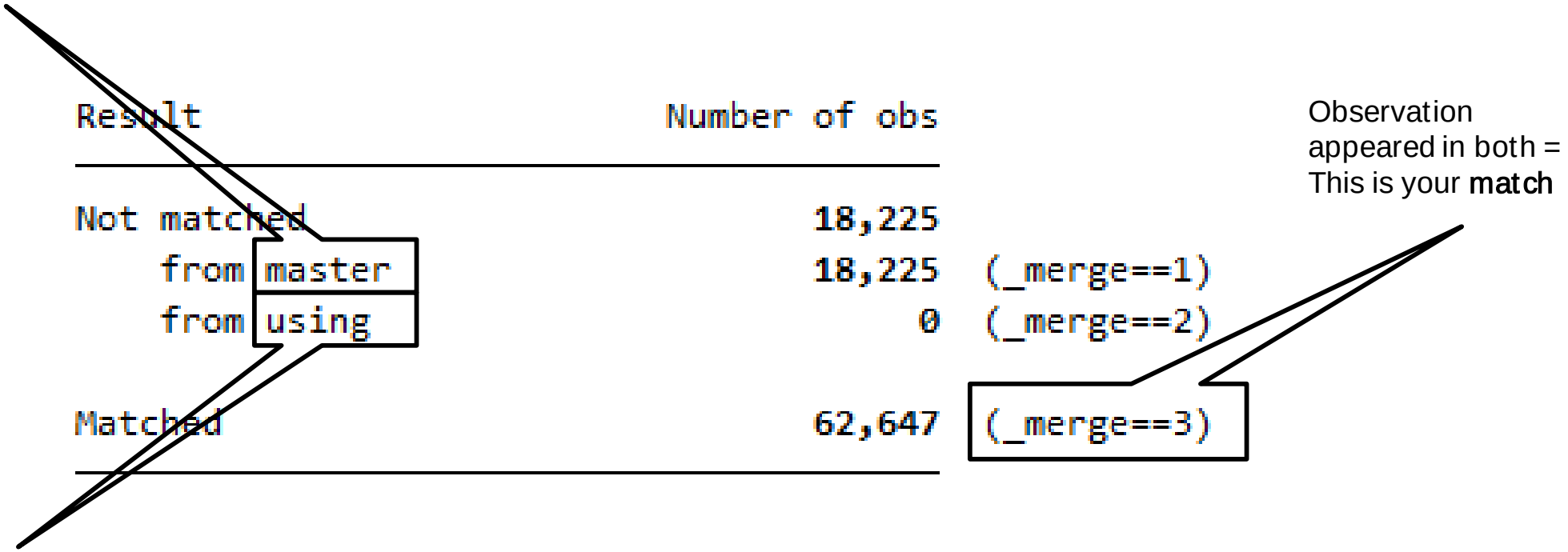
- The **same unique identifiers** like UPC or Census block group
- In the **same format** like a text string or numeric value

In most cases, any two data purveyors do not collect:

- unique identifiers that can be mapped across files, or
- the same piece of information in the same format

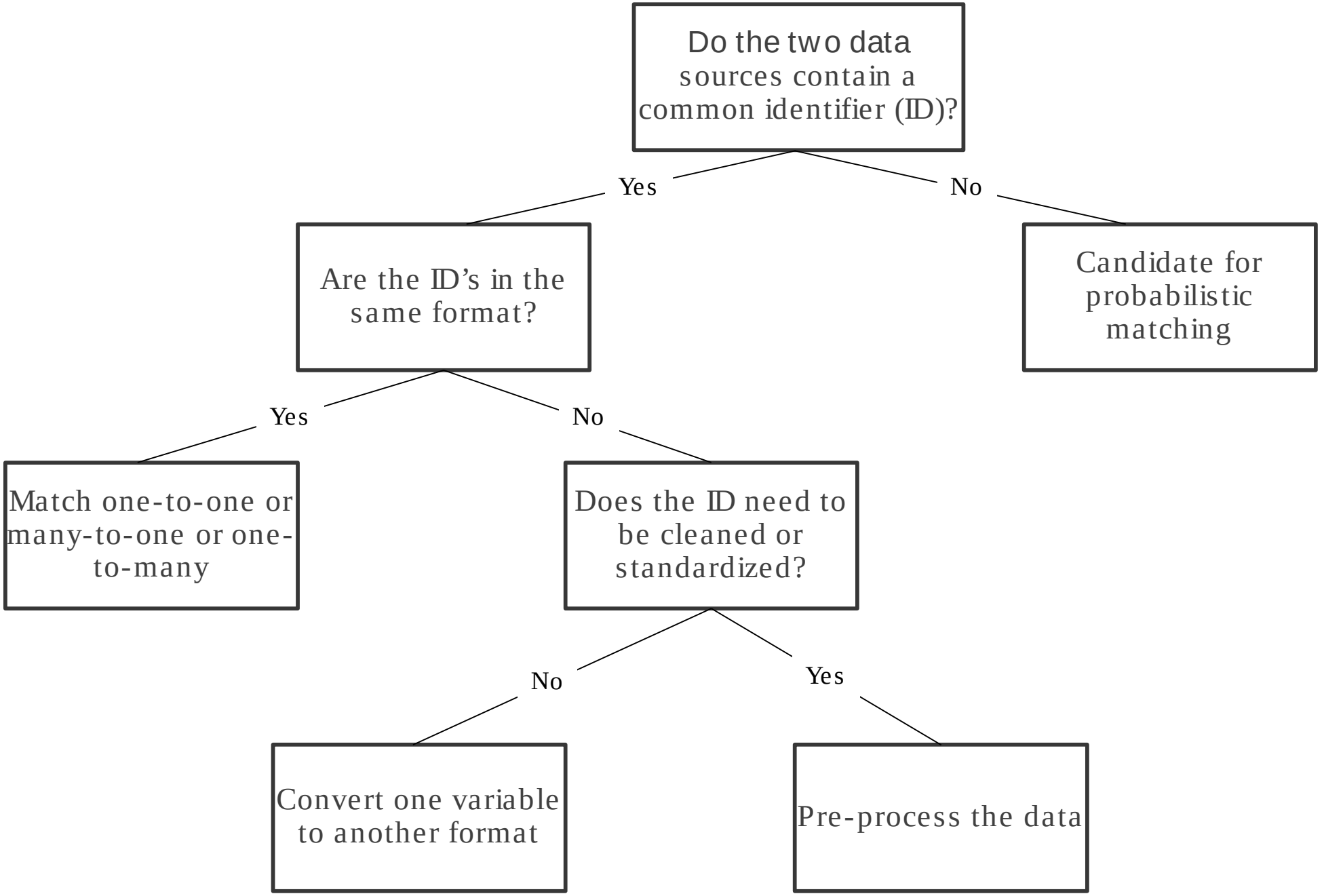
```
merge m:1 varname using filename
```

Master file = the file in memory



Using file = the file being merged *into* the master file

Match score = $$\frac{\_merge == 3}{\_merge == 1 + \_merge == 2 + \_merge == 3}$$

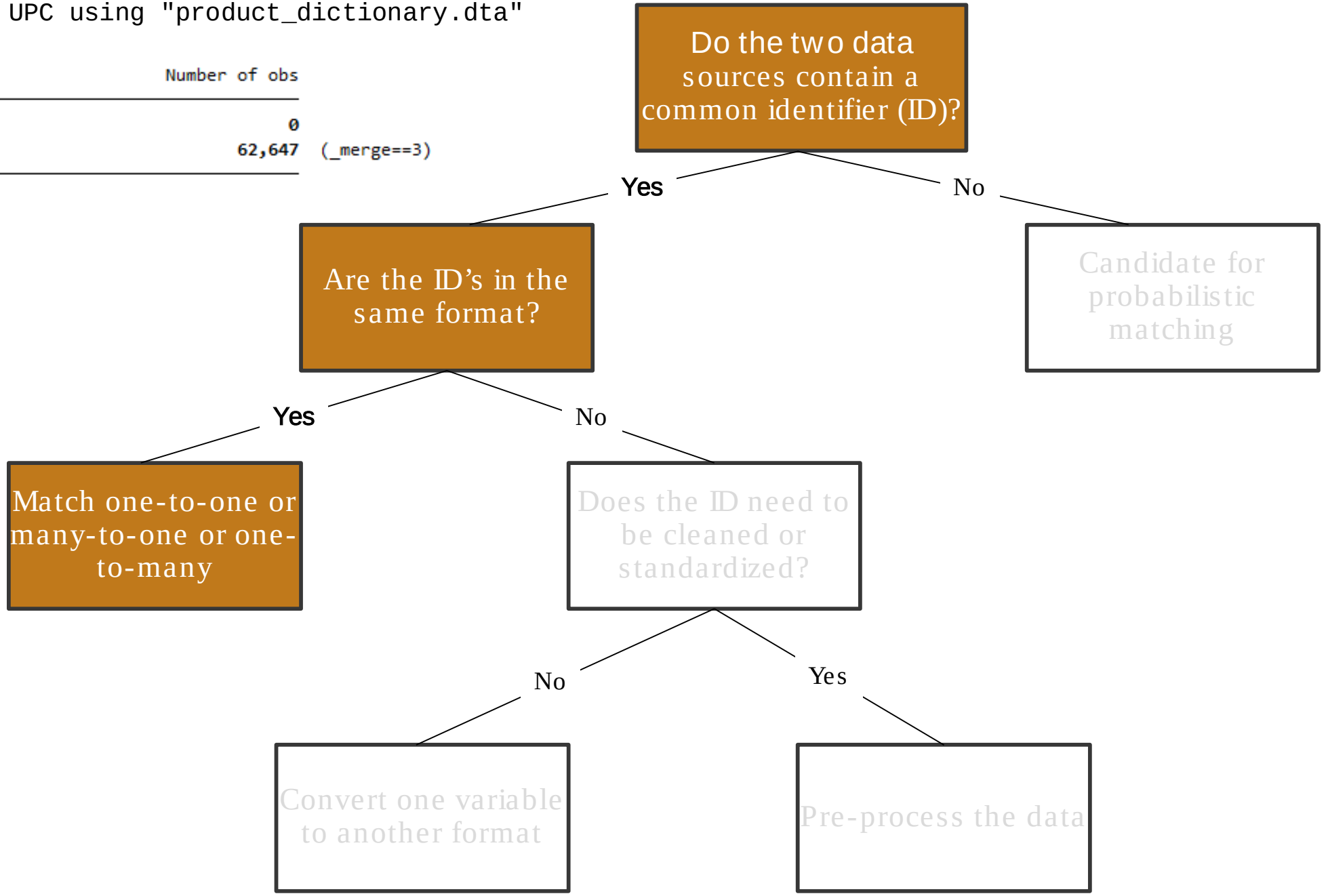

Do the two data sources contain a common identifier (ID)?
Yes
No
Are the ID's in the same format?
Candidate for probabilistic matching
Yes
No
Match one-to-one or many-to-one or one-to-many
Does the ID need to be cleaned or standardized?
No
Yes
Convert one variable to another format
Pre-process the data

Example #1

Using store level transactions and a UPC product dictionary, graphically depict monthly sales by food category.

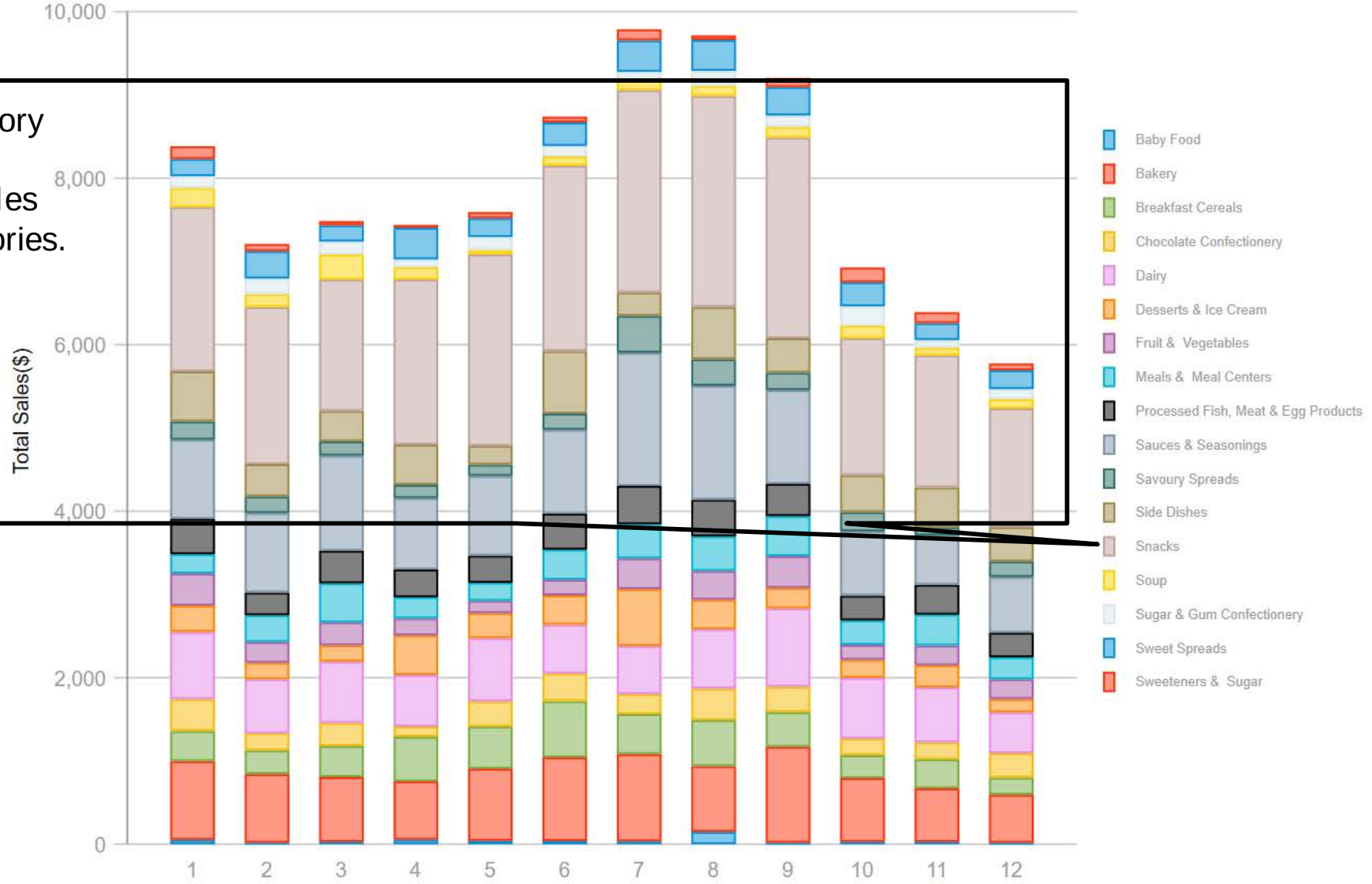
```
merge m:1 UPC using "product_dictionary.dta"
```

Result	Number of obs	
Not matched	0	
Matched	62,647	(_merge==3)



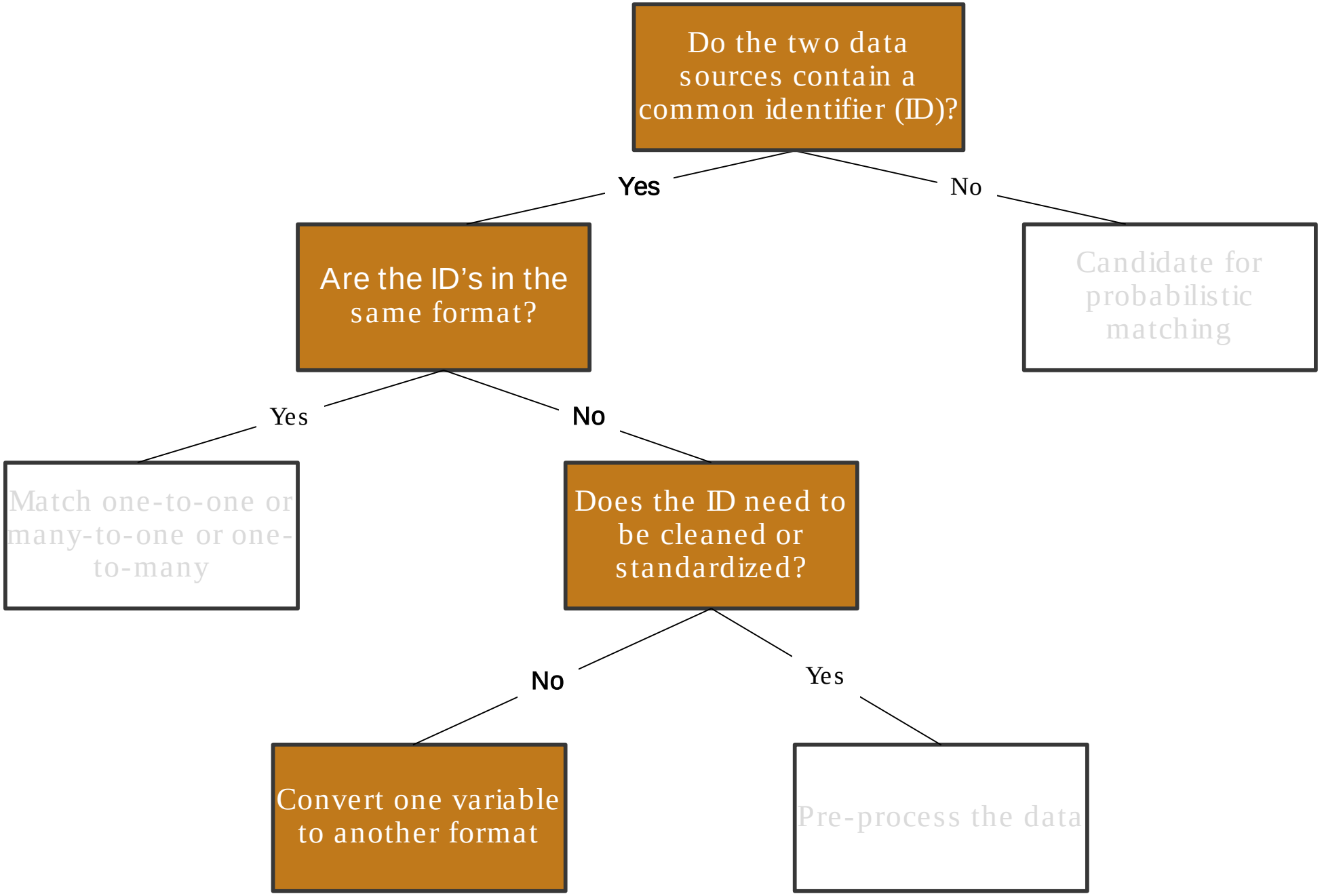
Category-Level Food Sales by Month

The **snack** category has the greatest share of food sales across all categories.



Example #2

Determine the per capita income by Census block group and create quantiles to highlight regional differences.


Do the two data sources contain a common identifier (ID)?
Yes
No
Are the ID's in the same format?
Candidate for probabilistic matching
Yes
No
Match one-to-one or many-to-one or one-to-many
Does the ID need to be cleaned or standardized?
No
Yes
Convert one variable to another format
Pre-process the data

TIGER/Line Files and Shapefiles contain geographic codes that can be linked to the Census Bureau's demographic data.

Name	Label	Type	Format
OBJECTID	OBJECTID	long	%10.0f
STATE_FIPS	STATE_FIPS	str2	%9s
CNTY_FIPS	CNTY_FIPS	str3	%9s
STCOFIPS	STCOFIPS	str5	%9s
TRACT	TRACT	str6	%9s
BLKGRP	BLKGRP	str1	%9s
FIPS	FIPS	str12	%12s
POPULATION	POPULATION	long	%10.0f
POP_SQMI	POP_SQMI	float	%19.11f
POP2010	POP2010	long	%10.0f
POP10_SQMI	POP10_SQMI	float	%19.11f
WHITE	WHITE	int	%10.0f
BLACK	BLACK	int	%10.0f
AMERI_ES	AMERI_ES	int	%10.0f
ASIAN	ASIAN	int	%10.0f
HAWN_PI	HAWN_PI	int	%10.0f
HISPANIC	HISPANIC	int	%10.0f
OTHER	OTHER	int	%10.0f
MULT_RACE	MULT_RACE	int	%10.0f
MALES	MALES	int	%10.0f
FEMALES	FEMALES	int	%10.0f
AGE_UNDER5	AGE_UNDER5	int	%10.0f

	OBJECTID	STATE_FIPS	CNTY_FIPS	STCOFIPS	TRACT	BLKGRP	FIPS	POPULATION	POP_SQMI	POP2010	POP10_SQMI	WHITE
1	1386	06	083	06083	003102	1	060830031021	842	4.30000019073	825	4.19999980927	680
2	1387	06	083	06083	980100	1	060839801001	11	0.10000000149	11	0.10000000149	6
3	1388	06	111	06111	980000	1	061119800001	59	2.59999990463	56	2.50000000000	44
4	1389	06	083	06083	001906	5	060830019065	1143	116.90000152588	1084	110.80000305176	788
5	1390	06	083	06083	001903	3	060830019033	847	1486.00000000000	822	1442.09997558594	731
6	1391	06	083	06083	001903	4	060830019034	2402	906.40002441406	2027	764.90002441406	1549
7	1392	06	083	06083	002932	1	060830029321	609	7.30000019073	577	7.00000000000	396
8	1393	06	083	06083	001903	5	060830019035	1124	5915.79980468750	1082	5694.70019531250	807
9	1394	06	083	06083	001903	2	060830019032	849	1845.69995117188	797	1732.59997558594	754
10	1395	06	083	06083	002930	2	060830029302	1427	1603.40002441406	996	1119.09997558594	741
11	1396	06	083	06083	002909	2	060830029092	1663	6396.20019531250	1694	6515.39990234375	1181
12	1397	06	083	06083	002930	1	060830029301	1748	5462.50000000000	1677	5240.60009765625	1027
13	1398	06	083	06083	002932	2	060830029322	2010	302.29998779297	1922	289.00000000000	1570
14	1399	06	083	06083	002930	5	060830029305	1209	10075.00000000000	1202	10016.70019531250	741
15	1400	06	083	06083	001906	4	060830019064	927	9.89999961853	910	9.69999980927	741
16	1401	06	083	06083	002930	6	060830029306	1570	22428.59960937500	1414	20200.00000000000	880
17	1402	06	083	06083	002909	3	060830029093	2269	6302.79980468750	2063	5730.60009765625	1459
18	1403	06	083	06083	002915	1	060830029151	573	1005.29998779297	580	1017.50000000000	348
19	1404	06	083	06083	002930	3	060830029303	1314	4692.89990234375	1285	4589.29980468750	863
20	1405	06	083	06083	002909	1	060830029091	2305	6229.70019531250	2206	5962.20019531250	1531
21	1406	06	083	06083	002928	1	060830029281	2917	6341.29980468750	2748	5973.89990234375	2131
22	1407	06	083	06083	002930	4	060830029304	775	25833.30078125000	754	25133.30078125000	395
23	1408	06	083	06083	001208	4	060830012084	777	7770.00000000000	727	7270.00000000000	596
24	1409	06	083	06083	001208	3	060830012083	766	4505.89990234375	727	4276.50000000000	618
25	1410	06	083	06083	001304	8	060830013048	707	7070.00000000000	678	6780.00000000000	596
26	1411	06	083	06083	001208	2	060830012082	802	4010.00000000000	869	4345.00000000000	738
27	1412	06	083	06083	002906	2	060830029062	1462	4300.00000000000	1423	4185.29980468750	1133
28	1413	06	083	06083	002907	3	060830029073	1218	110.00000000000	1154	104.19999694824	846
29	1414	06	083	06083	002922	4	060830029224	1972	2240.89990234375	1404	1595.50000000000	1031
30	1415	06	083	06083	002922	1	060830029221	5030	6448.70019531250	4673	5991.00000000000	2727
31	1416	06	083	06083	003004	3	060830030043	1488	7085.70019531250	1429	6804.79980468750	1118
32	1417	06	083	06083	002913	1	060830029131	1811	3853.19995117188	1667	3546.80004882813	1441
33	1418	06	083	06083	003001	3	060830030013	980	9800.00000000000	935	9350.00000000000	549
34	1419	06	083	06083	002922	5	060830029225	1671	11935.70019531250	1455	10392.90039062500	884
35	1420	06	083	06083	002906	3	060830029063	1693	4979.39990234375	1679	4938.20019531250	1425
36	1421	06	083	06083	002924	1	060830029241	1685	56166.69921875000	1530	51000.00000000000	1197

The American Community Survey 5-year estimates contain demographic data that can be linked to the TIGER/Line Files via geographic codes.

Name	Label	Type	Format
year		float	%9.0g
BLOCKGROUP	FIPS	double	%15.0g
GEOID	Geographic Identifier	str40	%19s
STATE	State (FIPS)	str2	%9s
COUNTY	County	str3	%9s
TRACT	Census Tract	str6	%9s
BLKGRP	Block Group	str1	%9s

	year	BLOCKGROUP	GEOID	STATE	COUNTY	TRACT	BLKGRP	T001_00~2006	T002_001_2~6	T002_002_2~6	T002_003_2~6	T003_001_2~
1	2006	10010201001	15000US010010201001	01	001	020100	1	547	547	334.0027	1.6377114	1.651733
2	2006	10010201002	15000US010010201002	01	001	020100	2	1262	1262	586.996	2.1499293	2.149925
3	2006	10010202001	15000US010010202001	01	001	020200	1	1313	1313	1652.092	.79475001	.79475
4	2006	10010202002	15000US010010202002	01	001	020200	2	707	707	1428.207	.49502623	.497283
5	2006	10010203001	15000US010010203001	01	001	020300	1	2825	2825	1892.327	1.4928706	1.49636
6	2006	10010203002	15000US010010203002	01	001	020300	2	718	718	1254.158	.5724957	.572495
7	2006	10010204001	15000US010010204001	01	001	020400	1	1310	1310	1242.787	1.0540825	1.05447
8	2006	10010204002	15000US010010204002	01	001	020400	2	2022	2022	2938.048	.68821207	.694093
9	2006	10010204003	15000US010010204003	01	001	020400	3	972	972	3352.941	.28989478	.289894
10	2006	10010204004	15000US010010204004	01	001	020400	4	536	536	1240.204	.43218694	.432186
11	2006	10010205001	15000US010010205001	01	001	020500	1	1692	1692	1015.795	1.66569	1.6656
12	2006	10010205002	15000US010010205002	01	001	020500	2	6034	6034	3005.093	2.0079247	2.026617
13	2006	10010205003	15000US010010205003	01	001	020500	3	2212	2212	3042.341	.72707171	.727071
14	2006	10010206001	15000US010010206001	01	001	020600	1	2441	2441	1514.457	1.611799	1.631157
15	2006	10010206002	15000US010010206002	01	001	020600	2	961	961	647.1896	1.4848818	1.488705
16	2006	10010207001	15000US010010207001	01	001	020700	1	1483	1483	196.0323	7.5650798	7.864417
17	2006	10010207002	15000US010010207002	01	001	020700	2	1181	1181	1086.677	1.0868	1.089223
18	2006	10010208011	15000US010010208011	01	001	020801	1	666	666	21.30675	31.257708	34.30716
19	2006	10010208012	15000US010010208012	01	001	020801	2	2194	2194	131.1582	16.727886	16.81126
20	2006	10010208021	15000US010010208021	01	001	020802	1	3096	3096	78.86757	39.255678	39.37286
21	2006	10010208022	15000US010010208022	01	001	020802	2	3691	3691	362.4075	10.184668	10.1867
22	2006	10010208023	15000US010010208023	01	001	020802	3	1641	1641	155.5481	10.549794	10.67283
23	2006	10010208024	15000US010010208024	01	001	020802	4	1988	1988	145.2965	13.682371	13.70197
24	2006	10010209001	15000US010010209001	01	001	020900	1	1422	1422	36.0455	39.450138	39.49559
25	2006	10010209002	15000US010010209002	01	001	020900	2	1637	1637	49.6876	32.945848	32.99478
26	2006	10010209003	15000US010010209003	01	001	020900	3	1364	1364	65.03829	20.972259	21.04819
27	2006	10010209004	15000US010010209004	01	001	020900	4	1058	1058	53.79899	19.665796	19.69478
28	2006	10010210001	15000US010010210001	01	001	021000	1	625	625	7.840792	79.711331	80.17864
29	2006	10010210002	15000US010010210002	01	001	021000	2	2259	2259	32.43163	69.654222	69.79982
30	2006	10010211001	15000US010010211001	01	001	021100	1	2192	2192	35.13802	62.382572	62.8681
31	2006	10010211002	15000US010010211002	01	001	021100	2	356	356	8.190034	43.467463	43.61122
32	2006	10010211003	15000US010010211003	01	001	021100	3	750	750	9.52116	78.771917	83.51822
33	2006	10030101001	15000US010030101001	01	003	010100	1	644	644	5.123158	125.70372	135.8159
34	2006	10030101002	15000US010030101002	01	003	010100	2	1230	1230	14.05305	87.525459	91.58623
35	2006	10030101003	15000US010030101003	01	003	010100	3	1682	1682	10.39025	161.88252	162.2051
36	2006	10030102001	15000US010030102001	01	003	010200	1	1454	1454	32.43922	44.822286	44.90093
37	2006	10030102002	15000US010030102002	01	003	010200	2	1158	1158	28.38717	40.793079	40.85273
38	2006	10030103001	15000US010030103001	01	003	010300	1	2617	2617	110.5087	23.681384	23.69812
39	2006	10030103002	15000US010030103002	01	003	010300	2	3216	3216	73.49739	43.756657	46.5195
40	2006	10030103003	15000US010030103003	01	003	010300	3	1610	1610	24.66473	65.275385	71.4446
41	2006	10030104001	15000US010030104001	01	003	010400	1	727	727	5.427578	133.94557	134.7075

Approach 1: Convert string to numeric

```
use "Data\USA_Block_Group_Boundaries_Lower_48.dta", clear

destring FIPS, gen(BLOCKGROUP)
format BLOCKGROUP %15.0g

merge 1:1 BLOCKGROUP using "Data\acs_2006_2010_cbg.dta"
```

Turn FIPS into a numeric value and call it 'BLOCKGROUP'

Merge with master file on 'BLOCKGROUP'

```
Result                           Number of obs
-----------------------------------------
Not matched                         4,471
    from master                        14  (_merge==1)
    from using                      4,457  (_merge==2)

Matched                           215,877  (_merge==3)
-----------------------------------------
```

Match score = 98%

Approach 2: Convert numeric to string

```
use "Data\acs_2006_2010_cbg.dta", clear
tostring BLOCKGROUP, gen(FIPS) format(%012.0f)
save "Data\acs_2006_2010_cbg-2.dta", replace

use "Data\USA_Block_Group_Boundaries_Lower_48.dta", clear
merge 1:1 FIPS using "Data\acs_2006_2010_cbg-2.dta"
```

Turn BLOCKGROUP into a string and call it 'FIPS'

Merge with master file on 'FIPS'

Result	Number of obs	
Not matched	4,471	
from master	14	(_merge==1)
from using	4,457	(_merge==2)
Matched	215,877	(_merge==3)

Match score = 98%

Approach 2: Convert numeric to string

```
use "Data\acs_2006_2010_cbg.dta", clear
gen str12 FIPS = string(BLOCKGROUP, "%012.0f")
save "Data\acs_2006_2010_cbg-3.dta", replace

use "Data\USA_Block_Group_Boundaries_Lower_48.dta", clear
merge 1:1 FIPS using "Data\acs_2006_2010_cbg-3.dta"
```

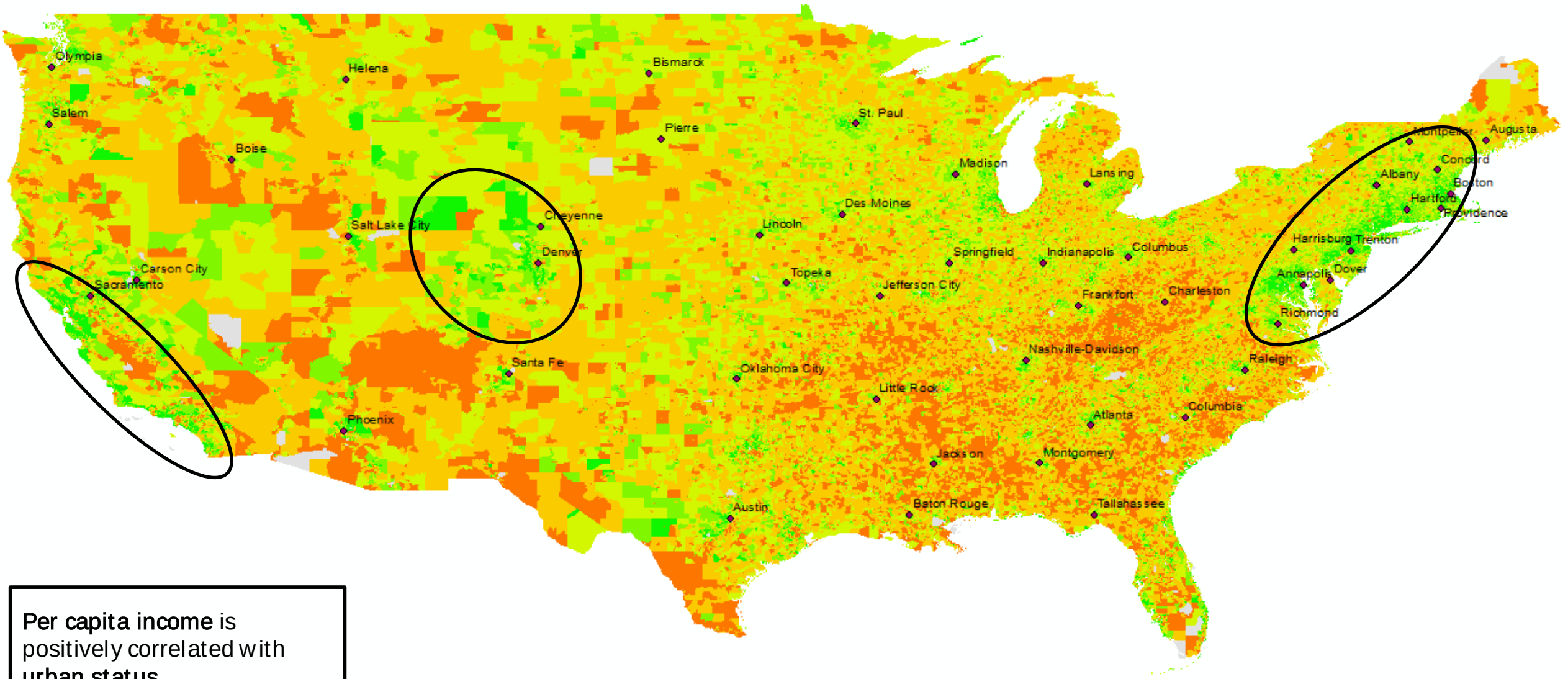
Turn BLOCKGROUP into a string and call it 'FIPS'

Merge with master file on 'FIPS'

Result	Number of obs	
Not matched	4,471	
from master	14	(_merge==1)
from using	4,457	(_merge==2)
Matched	215,877	(_merge==3)

Match score = 98%

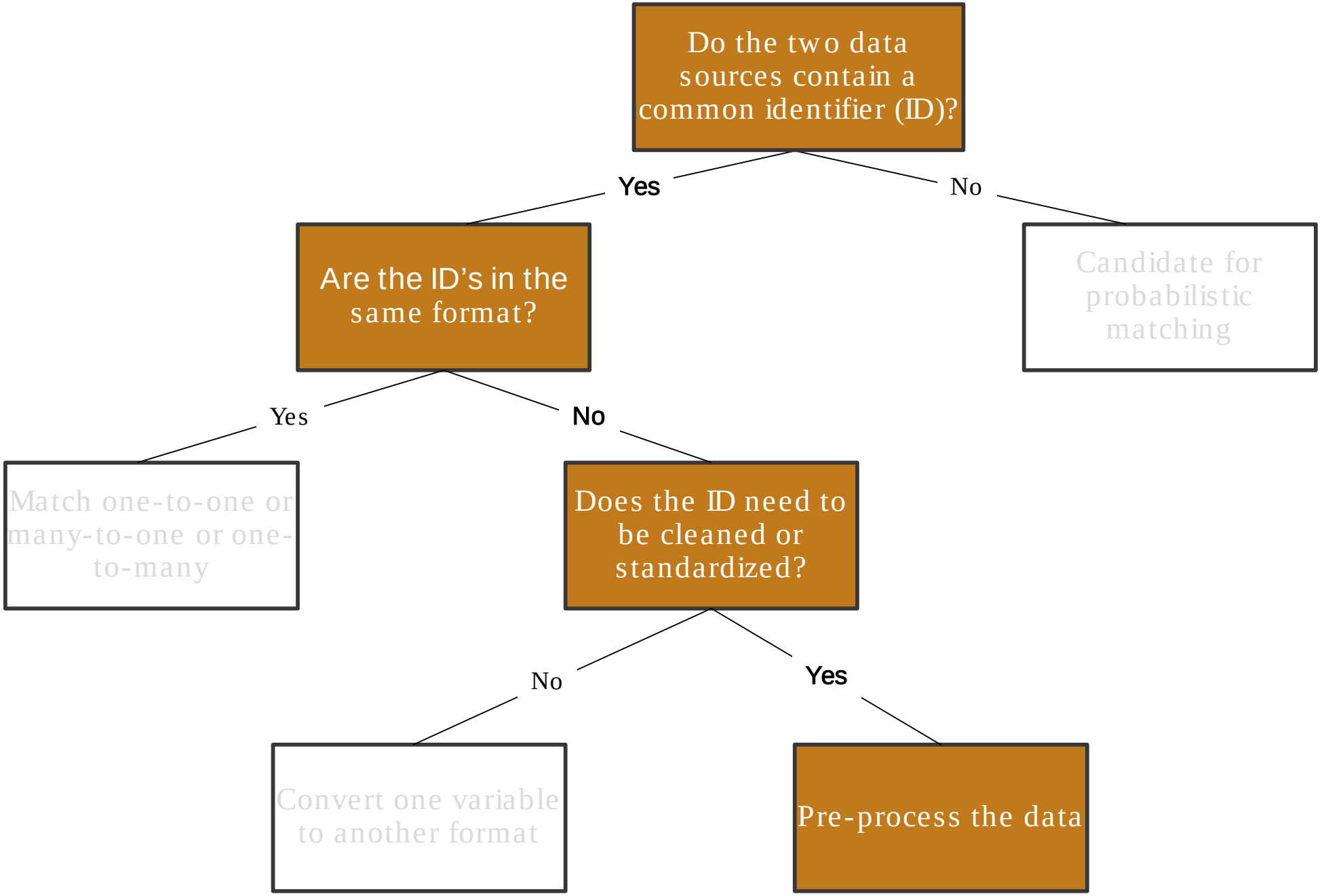
Per capita income by Census block group



Per capita income is positively correlated with **urban status**.

Example #3

Map donation flows between food retail-donors and food pantries.


Do the two data sources contain a common identifier (ID)?
Yes
No
Are the ID's in the same format?
Candidate for probabilistic matching
Yes
No
Match one-to-one or many-to-one or one-to-many
Does the ID need to be cleaned or standardized?
No
Yes
Convert one variable to another format
Pre-process the data

Research question

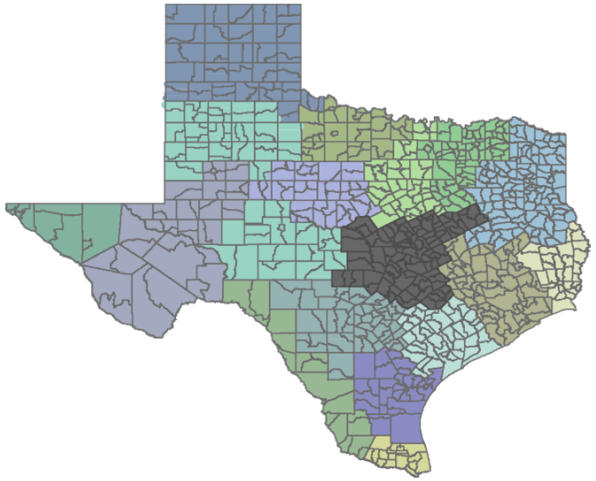
What is the effect of **food pantry presence** on **food retailer donation volumes**?

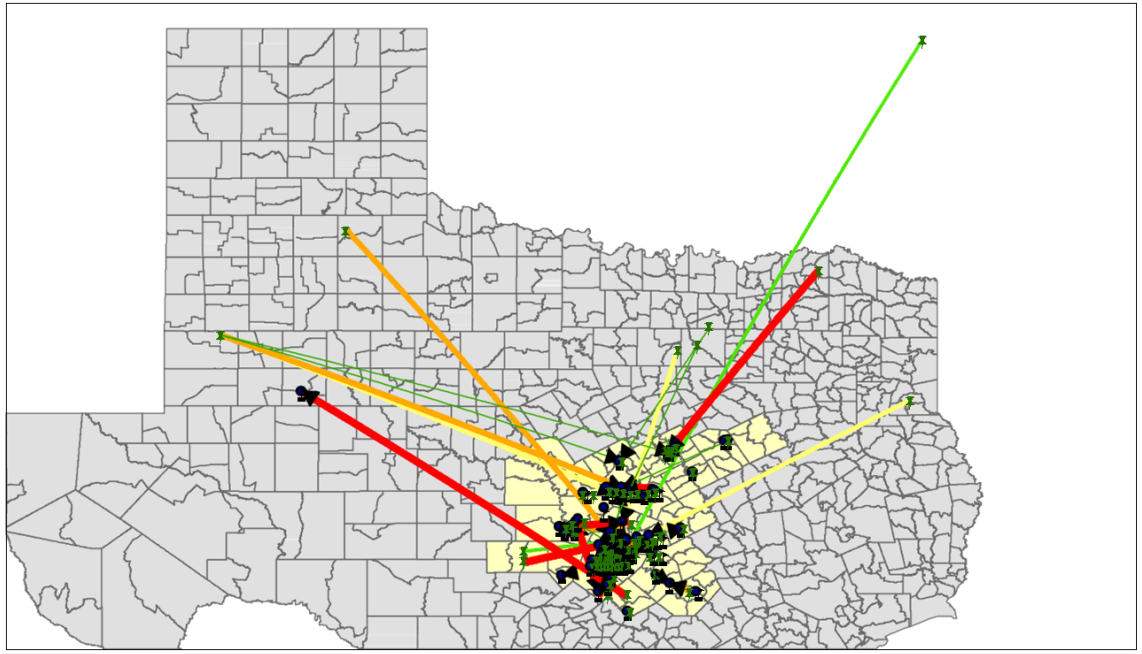
Data requirements:

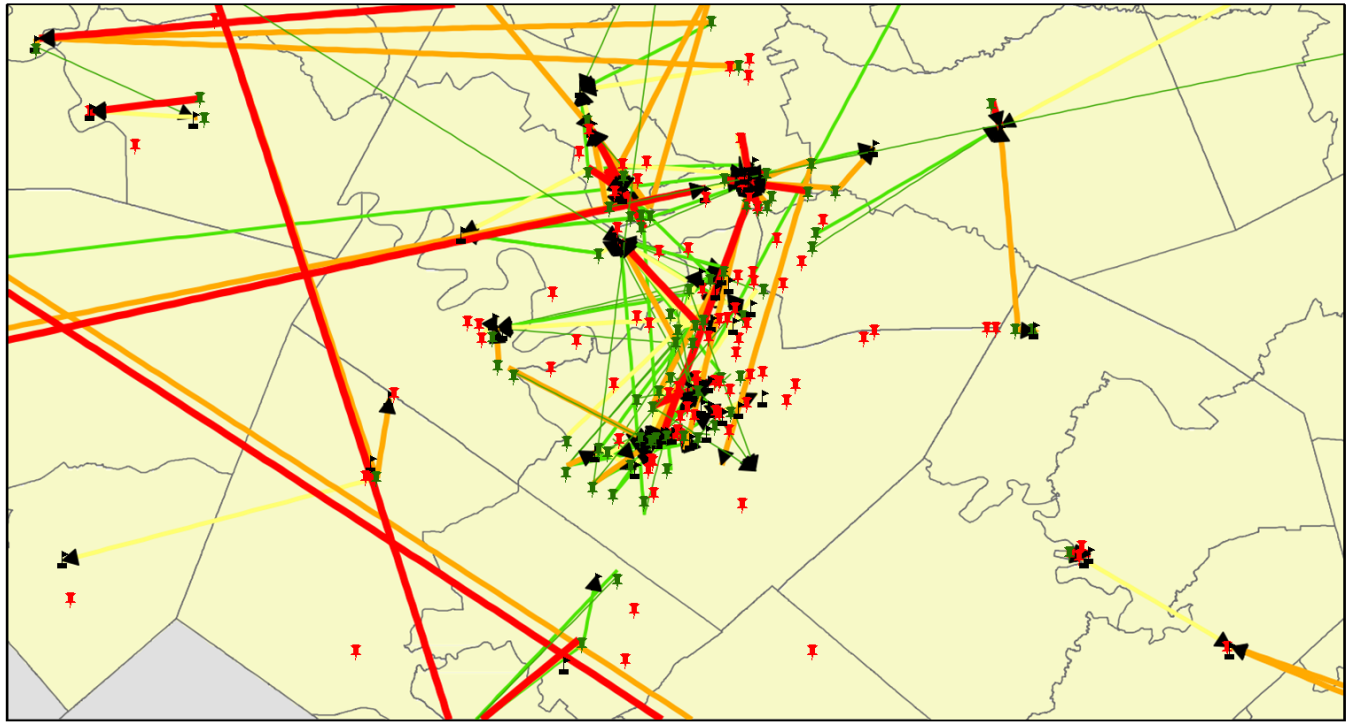
- Pounds of food donated by each retailer to each food pantry
- Distance between food pantry and food retailer
  - Food pantry location
  - Food retailer location

These data come from either (1) the food bank's internal record-keeping system or (2) a receipting platform managed by a third-party vendor in the form of Excel or CSV files.

Pre-processing requirements: **Standardize** common identifiers to link datasets and **Geocode** addresses to generate distances.





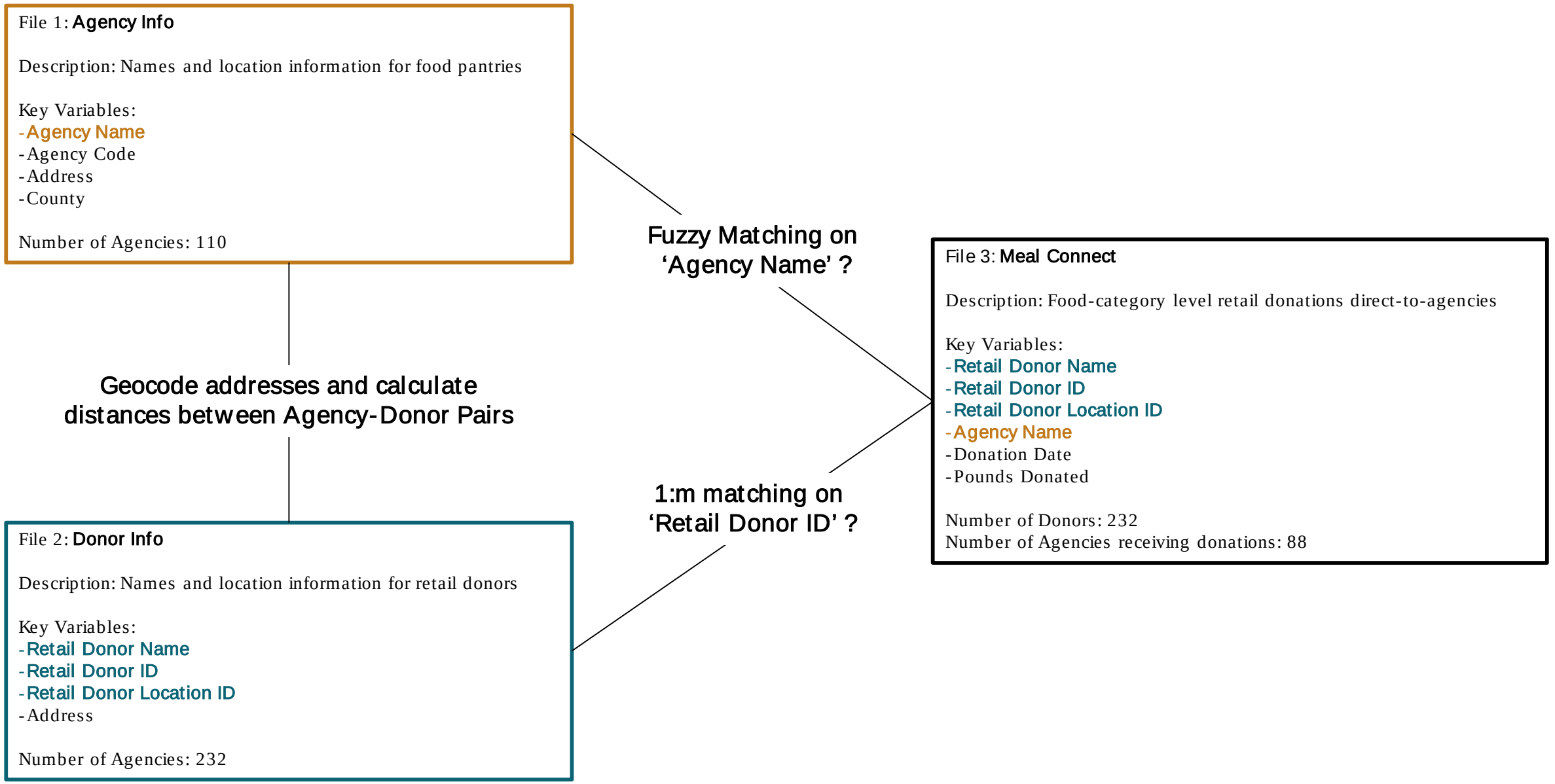


Legend

Donor-Agency Flows

Pounds Donated

- Quintile 1
- Quintile 2
- Quintile 3
- Quintile 4
- Quintile 5
- Donor Retailer
- Non-Donor Retailer
- Food Pantry


File 1: Agency Info
Description: Names and location information for food pantries
Key Variables:
-Agency Name
-Agency Code
-Address
-County
Number of Agencies: 110
Fuzzy Matching on 'Agency Name' ?
File 3: Meal Connect
Description: Food-category level retail donations direct-to-agencies
Key Variables:
-Retail Donor Name
-Retail Donor ID
-Retail Donor Location ID
-Agency Name
-Donation Date
-Pounds Donated
Number of Donors: 232
Number of Agencies receiving donations: 88
Geocode addresses and calculate distances between Agency-Donor Pairs
1:m matching on 'Retail Donor ID' ?
File 2: Donor Info
Description: Names and location information for retail donors
Key Variables:
-Retail Donor Name
-Retail Donor ID
-Retail Donor Location ID
-Address
Number of Agencies: 232

File 1: **Agency Info**

	agencyno	agencyname	agencyaddress	county
1.	OP020	Caritas of Austin	611 Neches, Austin, TX 78701	TRAVIS
2.	OP202	Project Transitions	5606 Roosevelt Ave. #109, Austin, TX 78756	TRAVIS
3.	OS016	Austin Baptist Chapel	908 E. Cesar Chavez, Austin, TX 78702	TRAVIS
4.	OS048	Casa Marianella	821 Gunter, Austin, TX 78702	TRAVIS
5.	OS090	Ebenezer CDC	1014 E. 10th Street, Austin, TX 78702	TRAVIS
6.	OS207	River City Youth	5209 S. Pleasant Valley Rd, Austin, TX 78744	TRAVIS
7.	OS219	Salvation Army Shelter Austin	501 E. 8th St., Austin, TX 78701	TRAVIS
8.	OS220	Salvation Army Rehab	4216 S. Congress Ave, Austin, TX 78745	TRAVIS
9.	OS275	Trinity CDC	5801 Westminster Dr., Austin, TX 78723	TRAVIS
10.	OS373	Front Steps (ARCH)	500 E. 7th St., Austin, TX 78701	TRAVIS

Task #1 - Standardizing variables

Fuzzy Matching on 'Agency Name'

File 3: **Meal Connect**

	year	donorid	donorl~d	donorname	donors~e	agencyname	pounds
1.	2015	776	HEB #34	H-E-B Food Stores	TX	Round Rock Area Serving Center	971
2.	2015	776	HEB #34	H-E-B Food Stores	TX	Storehouse Pflugerville	323
3.	2015	776	HEB #373	H-E-B Food Stores	TX	Round Rock Area Serving Center	1527
4.	2015	776	HEB #495	H-E-B Food Stores	TX	Round Rock Area Serving Center	853
5.	2015	776	HEB #611	H-E-B Food Stores	TX	Church of the Springs	74
6.	2015	776	HEB #668	H-E-B Food Stores	TX	Operation Phantom Support	263
7.	2015	776	HEB #673	H-E-B Food Stores	TX	Round Rock Area Serving Center	835
8.	2015	3198	B148	Sam's Club	TX	Caritas of Waco	145
9.	2015	3198	B153	Sam's Club	TX	Operation Phantom Support	2237
10.	2015	3201	B189	Walmart Stores, Inc.	TX	Storehouse Pflugerville	4700

'Donor ID' ⊃ 'Donor Location ID'

1:m matching on 'Donor Location ID'

File 2: **Donor Info**

	donorid	donorl~d	storen~r	donorname	donorstreetaddress1	donors~2	donorc~y	donors~e	donorp~e
1.	776	HEB #467	467	H-E-B Food Stores	701 Milam Street		Mexia	TX	76667
2.	6228	B151	T1542	Target	5621 N I H 35		AUSTIN	TX	78723
3.	11496	B159	W45912	Whole Foods Market	11920 Domain Drive		AUSTIN	TX	78758
4.	6228	B423	T1817	Target	12901 N I H 35 Ste 3-300		AUSTIN	TX	78753
5.	6228	B476	T1061	Target	5300 S Mo Pac Expy		AUSTIN	TX	78749
6.	6228	B475	T0096	Target	2300 W Ben White Blvd		AUSTIN	TX	78704
7.	6228	B463	T0095	Target	8601 Research Blvd		AUSTIN	TX	78758
8.	11972	B127	2483	Randall's	715 S. EXPOSITION		AUSTIN	TX	78703
9.	11972	B126	2482	Randall's	8040 MESA DRIVE		AUSTIN	TX	78731
10.	3201	B208	4554	Walmart Stores, Inc.	2525 W ANDERSON LN		AUSTIN	TX	78757

Goal: Create a crosswalk between Agency Info and Meal Connect

Overview of steps to perform fuzzy matching

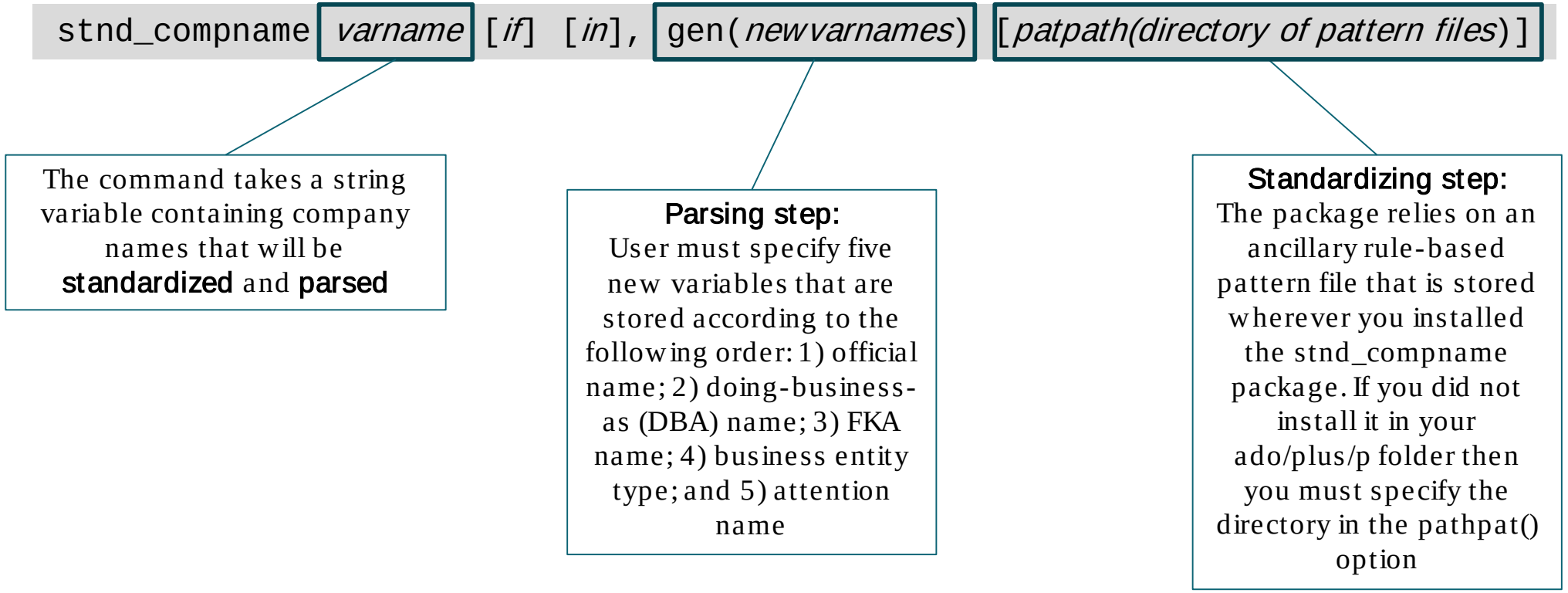
Step 1	**Standardize** variables	`stnd_compname`
Step 2	**Link** files	`reclink2`
Step 3	**Review** matches	`clrevmatch`

Install the following package:

```
ssc install dm0082
```

Step 1 - Run the **stnd_compname** command for both Agency Info and Meal Connect files.

The stnd_compname and stnd_address commands parse and standardize company names and addresses to improve the match score when linking.



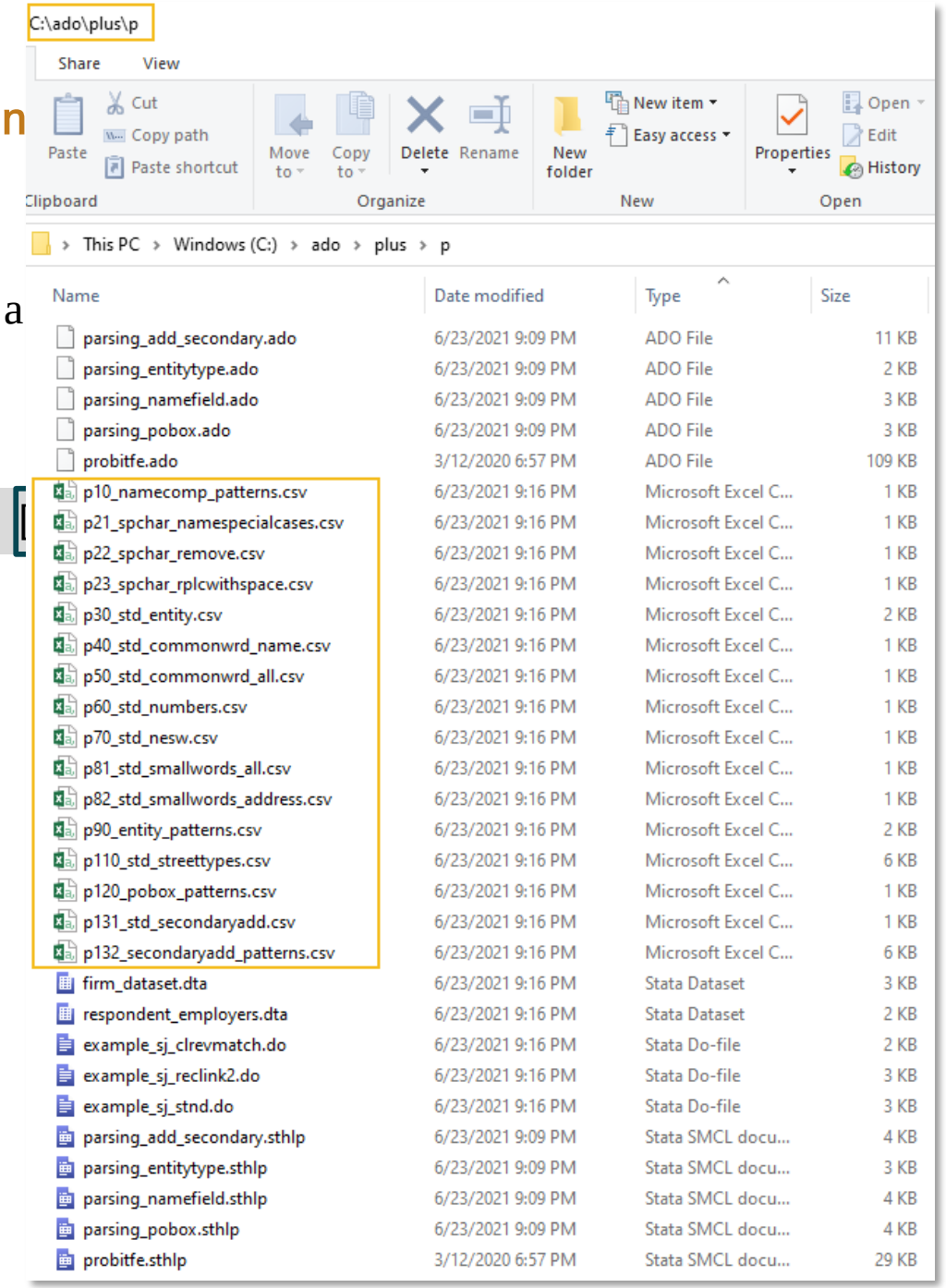
Step 1 - Run the `stnd_compname` command for both Agency In

The `stnd_compname` and `stnd_address` commands parse a and addresses to improve the match score when linking.

```
stnd_compname varname [if] [in], gen(newvarnames) [
```

The command takes a string variable containing company names that will be **standardized** and **parsed**

Parsing step:
User must specify five new variables that are stored according to the following order: 1) official name; 2) doing-business-as (DBA) name; 3) FKA name; 4) business entity type; and 5) attention name



Step 1 - Run the **stnd_compname** command for both Agency Info and Meal Connect files.

```
** Standardize agencies from Agency Info
use "1_agency_lookup.dta", clear
stnd_compname agencyname, gen(ag_name ag_dbaname ag_fkaname ag_entityname attn_name)
sort agencyname
gen ag_id_1=_n

save "1_agency_lookup_standardized.dta", replace

** Standardize agencies from Meal Connect
use "3_pounds_donated.dta", clear
stnd_compname agencyname, gen(ag_name ag_dbaname ag_fkaname ag_entityname attn_name)
sort agencyname
gen ag_id_2=_n

save "3_pounds_donated_standardized.dta", replace
```

Variable to be parsed

Variables must be labeled the same

Must specify an identifier

Step 2 – Use the `reclink2` command to assign a match score between records.

The `reclink2` command performs probabilistic record linkage between two datasets that have no joint identifier necessary for standard merging.

```
reclink2 varlist using filename, idmaster(varname) idusing(varname)
  gen(newvarname) [wmatch(match weight list) wnomatch(nonmatch weight list)
  orblock(varlist) required(varlist) exactstr(varlist) exclude(filename)
  merge(newvarname) uvarlist(varlist) uprefix(text) minscore(#) minbigram(#)
  manytoone npairs(#)]
```

The command takes the names of the variables common in both the master and using files that it will search over in order to find the most likely match with the using data

This is the using file, where the master dataset is the dataset in memory

These are the ID's that you gave in the `stnd_compname` command step

Step 2 – Use the `reclink2` command to assign a match score between records.

The `reclink2` command performs probabilistic record linkage between two datasets that have no joint identifier necessary for standard merging.

```
reclink2 varlist using filename, idmaster(varname) idusing(varname)
  gen(newvarname) [wmatch(match weight list) wnomatch(nonmatch weight list)
  orblock(varlist) required(varlist) exactstr(varlist) exclude(filename)
  merge(newvarname) uvarlist(varlist) uprefix(text) minscore(#) minbigram(#)
  manytoone npairs(#)]
```

Generate a variable that is the record linkage score (scaled 0-1)

The score is determined by how much "weight" you assign to any of the common variables

Step 2 – Use the `reclink2` command to assign a match score between records.

The `reclink2` command performs probabilistic record linkage between two datasets that have no joint identifier necessary for standard merging.

```
reclink2 varlist using filename, idmaster(varname) idusing(varname)
  gen(newvarname) [wmatch(match weight list) wnomatch(nonmatch weight list)
  orblock(varlist) required(varlist) exactstr(varlist) exclude(filename)
  merge(newvarname) uvarlist(varlist) uprefix(text) minscore(#) minbigram(#)
  manytoone npairs(#)]
```

Specifies that `reclink2` will allow records from the using dataset to be matched to multiple records from the master dataset (a many-to-one linking procedure)

Specifies that the program retain the top # potential matches from the using dataset that correspond to a given record in the master dataset

Step 2 – Use the **reclink2** command to assign a match score between records.

```
** Begin fuzzy matching (start with the Meal Connect file)

use "3_pounds_donated_standardized.dta", clear

reclink2 ag_name ag_dbaname ag_fkaname ag_entityname attn_name using
1_agency_lookup_standardized.dta, idmaster(ag_id_2) idusing(ag_id_1) gen(rlsc) wmatch (10
6 3 2 2) manytoone npairs(2)

gsort -_merge -rlsc

save "MealConnect_Agency_Crosswalk_1_forreview.dta", replace
```

This is the master file

This is the match score, weighted by wmatch()

This is the using file

	agencyname	ag_name	Uag_name	ag_id_2	rlsc	ag_id_1	_merge
1.	Travis Heights Food Pantry	TRAVIS HTS FOOD PANTRY	TRAVIS HTS FOOD PANTRY	543	1.0000	105	3
2.	Travis Heights Food Pantry	TRAVIS HTS FOOD PANTRY	TRAVIS HTS FOOD PANTRY	537	1.0000	105	3
3.	Travis Heights Food Pantry	TRAVIS HTS FOOD PANTRY	TRAVIS HTS FOOD PANTRY	541	1.0000	105	3
4.	Travis Heights Food Pantry	TRAVIS HTS FOOD PANTRY	TRAVIS HTS FOOD PANTRY	542	1.0000	105	3
5.	Travis Heights Food Pantry	TRAVIS HTS FOOD PANTRY	TRAVIS HTS FOOD PANTRY	511	1.0000	105	3
6.	Travis Heights Food Pantry	TRAVIS HTS FOOD PANTRY	TRAVIS HTS FOOD PANTRY	561	1.0000	105	3
7.	Micah 6	MICAH 6	MICAH 6	292	1.0000	68	3
8.	Travis Heights Food Pantry	TRAVIS HTS FOOD PANTRY	TRAVIS HTS FOOD PANTRY	534	1.0000	105	3
9.	Heaven's Harvest	HEAVENS HARVEST	HEAVENS HARVEST	161	1.0000	54	3
10.	Hope Food Pantry Austin	HOPE FOOD PANTRY AUSTIN	HOPE FOOD PANTRY AUSTIN	236	1.0000	55	3

Step 3 – Use the `clrevmatch` command to spot check possible matches.

The `clrevmatch` command allows users to review matches, interactively, without having to exit out of Stata.

```
clrevmatch using filename, idmaster(varname) idusing(varname) varM(varlist)
  varU(varlist) clrev result(newvarname) clrev note(newvarname)
  [reclinkscore(varname) rlscoremin(#) rlscoremax(#) rlscoredisp(on|off)
  fast clrev label(label) nobssave(#) replace newfilename(newfilename)
  saveold]
```

File to be reviewed, which is what you save after running `reclink2`

This dataset must contain the record identifiers from the master and the using datasets

Specify the set of variables in the master and using datasets, respectively, that will be displayed during the review

Step 3 – Use the `clrevmatch` command to spot check possible matches.

```
** Begin interactive clerical review
clrevmatch using MealConnect_Agency_Crosswalk_1_forreview.dta, idm(ag_id_2)
idu(ag_id_1) varM(ag_name) varU(Uag_name) clrev_result(crev) clrev_note(crnote)
replace

save "MealConnect_Agency_Crosswalk_1.dta", replace
```

File to be reviewed, which is what you save after running `reclink2`

ag_name and Uag_name will be displayed during the review

Let's see how it works in action!

File 1: **Agency Info**

Description: Names and location information for food pantries

Key Variables:
- Agency Name
- Agency Code
- Address
- County

Number of Agencies: 110

Geocode addresses and calculate distances between Agency-Donor Pairs

File 2: **Donor Info**

Description: Names and location information for retail donors

Key Variables:
- Retail Donor Name
- Retail Donor ID
- Retail Donor Location ID
- Address

Number of Agencies: 232

File 3: **Meal Connect**

Description: Food-category level retail donations direct-to-agencies

Key Variables:
- Retail Donor Name
- Retail Donor ID
- Retail Donor Location ID
- Agency Name
- Donation Date
- Pounds Donated

Number of Donors: 232
Number of Agencies receiving donations: 88

Overview of steps to generate distances

Step 1	**Standardize** address	`stnd_address`
Step 2	**Geocode** address	`geocodehere`
Step 3	**Map** distances	`georoute`

Install the following packages:

```
ssc install dm0082
ssc install insheetjson
ssc install libjson
ssc install georoute
net describe geocodehere,
from(https://raw.githubusercontent.com/simonheb/geocodehere/master/)
```

An application programming interface (API) key is an access code used to authenticate a user who wants to call a certain program hosted by a third-party developer.

It is required to obtain HERE maps API credentials, and is cost-free, provided that you stay below 100,000 monthly transactions.

Donor Addresses

	donorname	donorstate	donorstreetaddress1	donorstree~2	donorcity	donorposta~e
1	Sprouts Farmers Markets	TX	4006 South Lamar Blvd.		Austin	78704
2	Sprouts Farmers Markets	TX	4006 South Lamar Blvd.		Austin	78704
3	Sprouts Farmers Markets	TX	6920 Manchaca Road		Austin	78745
4	Sprouts Farmers Markets	TX	6920 Manchaca Road		Austin	78745
5	Sprouts Farmers Markets	TX	6920 Manchaca Road		Austin	78745
6	Sprouts Farmers Markets	TX	6920 Manchaca Road		Austin	78745
7	Amazon	TX	2209 Rutland #B		Austin	78758
8	Randall's	TX	9911 BRODIE LANE		Austin	78748
9	Randall's	TX	9911 BRODIE LANE		Austin	78748
10	Randall's	TX	9911 BRODIE LANE		Austin	78748
11	Randall's	TX	2727 EXPOSITION BLVD.		AUSTIN	78731
12	Randall's	TX	2727 EXPOSITION BLVD.		AUSTIN	78731
13	Randall's	TX	5311 BALCONES DRIVE		Austin	78731
14	Randall's	TX	5311 BALCONES DRIVE		Austin	78731
15	Randall's	TX	6600 S. Mopac		Austin	78749
16	Randall's	TX	1500 W 35th Street		Austin	78703
17	Randall's	TX			n	78703
18	Randall's	TX			n	78703
19	Randall's	TX			n	78703
20	Randall's	TX			n	78703
21	Randall's	TX			N	78731
22	Randall's	TX			N	78731
23	Randall's	TX			N	78731
24	Randall's	TX			N	78731
25	Randall's	TX			N	78703
26	Randall's	TX			N	78703
27	Randall's	TX	715 S. EXPOSITION		AUSTIN	78703
28	Randall's	TX	715 S. EXPOSITION		AUSTIN	78703
29	Randall's	TX	10900-D Research Blvd		Austin	78759
30	Randall's	TX	10900-D Research Blvd		Austin	78759
31	Randall's	TX	10900-D Research Blvd		Austin	78759
32	Randall's	TX	10900-D Research Blvd		Austin	78759
33	Sam's Club	TX	9900 S. IH 35 Bldg J-34		Austin	78748
34	Sam's Club	TX	10901 Lakeline Mall Dr.		Austin	78717
35	Sam's Club	TX	130 Sundance Pkwy Ste 300		Round Rock	78681
36	Sam's Club	TX	130 Sundance Pkwy Ste 300		Round Rock	78681
37	Target	TX	5621 N I H 35		AUSTIN	78723
38	Target	TX	5621 N I H 35		AUSTIN	78723
39	Whole Foods Market	TX	4301 West William Cannon		Austin	78749
40	Whole Foods Market	TX	4301 West William Cannon		Austin	78749
41	Whole Foods Market	TX	4301 West William Cannon		Austin	78749
42	Whole Foods Market	TX	4301 West William Cannon		Austin	78749

Agency Addresses

	agencyname	agencyno	agencyaddress	county
1	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
2	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
3	Casa Marianella	OS048	821 Gunter, Austin, TX 78702	TRAVIS
4	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
5	South Austin Neighborhood Ctr.	PA235	2508 Durwood, Austin, TX 78704	TRAVIS
6	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
7	Heaven's Harvest	PA512	1734 Rutland Dr, Austin, TX 78758	TRAVIS
8	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
9	Bannockburn Baptist Church	PA022	7100 Brodie Lane, Austin, TX 78745	TRAVIS
10	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
11	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
12	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
13	Heaven's Harvest	PA512	1734 Rutland Dr, Austin, TX 78758	TRAVIS
14	Heaven's Harvest	PA512	1734 Rutland Dr, Austin, TX 78758	TRAVIS
15	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
16	Micah 6	PA379	2203 San Antonio St., Austin, TX 78705	TRAVIS
17			TX 78745	TRAVIS
18			TX 78705	TRAVIS
19			TX 78745	TRAVIS
20			TX 78751	TRAVIS
21			TX 78758	TRAVIS
22			TX 78758	TRAVIS
23			TX 78758	TRAVIS
24			TX 78758	TRAVIS
25	Sal		TX 78701	TRAVIS
26			TX 78705	TRAVIS
27	Salvation Army Shelter Austin	OS219	501 E. 8th St., Austin, TX 78701	TRAVIS
28	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
29	Heaven's Harvest	PA512	1734 Rutland Dr, Austin, TX 78758	TRAVIS
30	Salvation Army Shelter Austin	OS219	501 E. 8th St., Austin, TX 78701	TRAVIS
31	Heaven's Harvest	PA512	1734 Rutland Dr, Austin, TX 78758	TRAVIS
32	Salvation Army Shelter Austin	OS219	501 E. 8th St., Austin, TX 78701	TRAVIS
33	Heaven's Harvest	PA512	1734 Rutland Dr, Austin, TX 78758	TRAVIS
34	Austin Baptist Chapel	OS016	908 E. Cesar Chavez, Austin, TX 78702	TRAVIS
35	Austin Baptist Chapel	OS016	908 E. Cesar Chavez, Austin, TX 78702	TRAVIS
36	Austin Baptist Chapel	OS016	908 E. Cesar Chavez, Austin, TX 78702	TRAVIS
37	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
38	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
39	El Buen Samaritano	PA091	7000 Woodhue, Austin, TX 78745	TRAVIS
40	El Buen Samaritano	PA091	7000 Woodhue, Austin, TX 78745	TRAVIS
41	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
42	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS

	donorname	donorstate	donorstreetaddress1	donorstree~2	donorcity	donorposta~e
1	Sprouts Farmers Markets	TX	4006 South Lamar Blvd.		Austin	78704
2	Sprouts Farmers Markets	TX	4006 South Lamar Blvd.		Austin	78704
3	Sprouts Farmers Markets	TX	6920 Manchaca Road		Austin	78745
4	Sprouts Farmers Markets	TX	6920 Manchaca Road		Austin	78745
5	Sprouts Farmers Markets	TX	6920 Manchaca Road		Austin	78745
6	Sprouts Farmers Markets	TX	6920 Manchaca Road		Austin	78745
7	Amazon	TX	2209 Rutland #B		Austin	78758
8	Randall's	TX	9911 BRODIE LANE		Austin	78748
9	Randall's	TX	9911 BRODIE LANE		Austin	78748
10	Randall's	TX	9911 BRODIE LANE		Austin	78748
11	Randall's	TX	2727 EXPOSITION BLVD.		AUSTIN	78731
12	Randall's	TX	2727 EXPOSITION BLVD.		AUSTIN	78731
13	Randall's	TX	5311 BALCONES DRIVE		Austin	78731
14	Randall's	TX	5311 BALCONES DRIVE		Austin	78731
15	Randall's	TX	6600 S. Mopac		Austin	78749

Donor address is (mostly) already parsed into street address, city, state and zip code.

```
** Standardize Donor address
use "2_donor_lookup.dta", clear

stnd_address donorstreetaddress1, gen(add1 pobox unit bldg floor)


** Geocode street address
geocodehere, apikey(7L5cZaHzptNXjEWeDTOycdcdTOF1HO7xBwCH6O1NoJ4)
postalcode(zip) street(add1) searchtext(donorstreetaddress1) replace
```

	donorl~d	geocod~at	geocodeh~n	geocodehe~et	geocod~r	ge~lcode
1.	B732	-34.0768	150.57057	[]	[]	2570
2.	B778	-25.41846	-57.57124	San Antonio	[]	2650
3.	B782	6.812324	-75.239734	Guadalupe	2	051820
4.	B181	9.90642	-84.13842	San Antonio	[]	10202
5.	B186	26.17018	-97.78355	East	[]	78552
6.	HEB #591	26.17018	-97.78355	East	[]	78552
7.	B793	29.41935	-98.5233	Guadalupe St	2021	78207
8.	HEB #104	29.55014	-98.37935	Live Oak	[]	78233
9.	B419	29.84608	-97.97031	Barnes Dr	700	78666
10.	B146	29.85121	-97.95347	Leah Ave	1350	78666

```
. tab geocodehere_match_code, miss
```

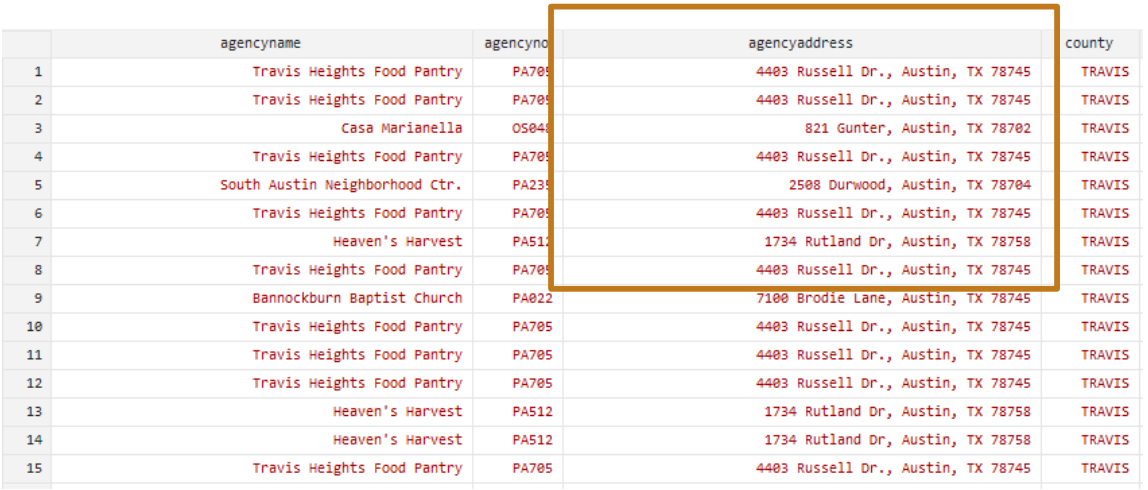
geocodehere_match_code	Freq.	Percent	Cum.
	43	18.53	18.53
ambiguous	6	2.59	21.12
ambiguousUpHierarchy	7	3.02	24.14
exact	156	67.24	91.38
upHierarchy	20	8.62	100.00
Total	232	100.00	

Geocoded addresses = 81%

	donorl~d	donorstreetaddress1	geoco~at	geocod~n	geoco~et	geocod~r	ge~lcode
222.	B776	10th & Congress	.	.			
223.	B757	183A & New Hope	.	.			
224.	B479	3550 S General Bruce DrBldg A-100	.	.			
225.	B761	Slaughter & Brodie	.	.			
226.	B756	Hwy 620 & Hwy 2222	.	.			

Need to do more pre-processing to fix these addresses

Agency address is not parsed, and stored as a string in a single variable.



	agencyname	agencyno	agencyaddress	county
1	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
2	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
3	Casa Marianella	OS045	821 Gunter, Austin, TX 78702	TRAVIS
4	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
5	South Austin Neighborhood Ctr.	PA235	2508 Durwood, Austin, TX 78704	TRAVIS
6	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
7	Heaven's Harvest	PA511	1734 Rutland Dr, Austin, TX 78758	TRAVIS
8	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
9	Bannockburn Baptist Church	PA022	7100 Brodie Lane, Austin, TX 78745	TRAVIS
10	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
11	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
12	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS
13	Heaven's Harvest	PA512	1734 Rutland Dr, Austin, TX 78758	TRAVIS
14	Heaven's Harvest	PA512	1734 Rutland Dr, Austin, TX 78758	TRAVIS
15	Travis Heights Food Pantry	PA705	4403 Russell Dr., Austin, TX 78745	TRAVIS

```
** Standardize Agency address
use "1_agency_lookup.dta", clear

split agencyaddress, p(",")
split agencyaddress3, p(" ")

stnd_address agencyaddress1, gen(add1 pobox unit bldg floor)

** Geocode street address
geocodehere, apikey(7L5cZaHzptNXjEWeDTOycdcdTOF1HO7xBwCH6O1NoJ4)
postalcode(agencyaddress32) street(agencyaddress1)
searchtext(searchtext) replace
```

	agencyname	geoco~at	geocode~n	geocodehere_street	geocod~r	ge~lcode
1.	A New Entry Inc.	30.28248	-97.67705	Webberville Rd	1808	78721
2.	A New Entry McCabe Center	30.27974	-97.7203	Martin Luther King Blvd E	1915	78702
3.	AIDS Services of Austin	30.32785	-97.69127	Cameron Rd	[]	78752
4.	ATCIC Project Recovery	30.27559	-97.73574	E 15th St	403	78701
5.	Abiding Love Lutheran Church	30.21778	-97.84582	Brush Country Rd	7210	78749
6.	Austin Baptist Chapel	30.26097	-97.73478	Cesar Chavez St E	908	78702
7.	Austin Restoration Ministries	30.36668	-97.68325	N Interstate 35	10206	78753
8.	Austin Shelter: Women/Children	30.28268	-97.66951	Tannehill Ln	4523	78721
9.	Austin Spanish Seventh Day Adv	30.36172	-97.69339	W Rundberg Ln	100	78753
10.	Austin Voices Burnet	30.36381	-97.72598	Hathaway Dr	8401	78757

```
. tab geocodehere_match_code, miss
```

geocodehere_match_code	Freq.	Percent	Cum.
	3	2.73	2.73
exact	107	97.27	100.00
Total	110	100.00	

Geocoded addresses = **97%**

	agencyname	agencyaddress	geoco~at	geocod~n	geoco~et
108.	Travis Co Comm Crt Pflugervill	15822 Foothill Farm Loop, BldD, Pflugerville, TX 78660	.	.	
109.	Foundation Comm-Sierra Ridge	201 W. St. Elmo, Austin, TX 78704	.	.	
110.	Trav Co Com Ctr at Palm Square	100 N. IH 35 #1000, Austin, TX 78701	.	.	

Need to do more pre-processing to fix these addresses

```
georoute, herekey(pe62h3tQj58pWMkAwYIUYNRKOfCpNJn0MUw0WgXQkLZ)
startxy(d_geocodehere_lat d_geocodehere_lon) endxy(ag_geocodehere_lat
ag_geocodehere_lon) di(dist) ti(time) co(p1 p2)
```

georoute results

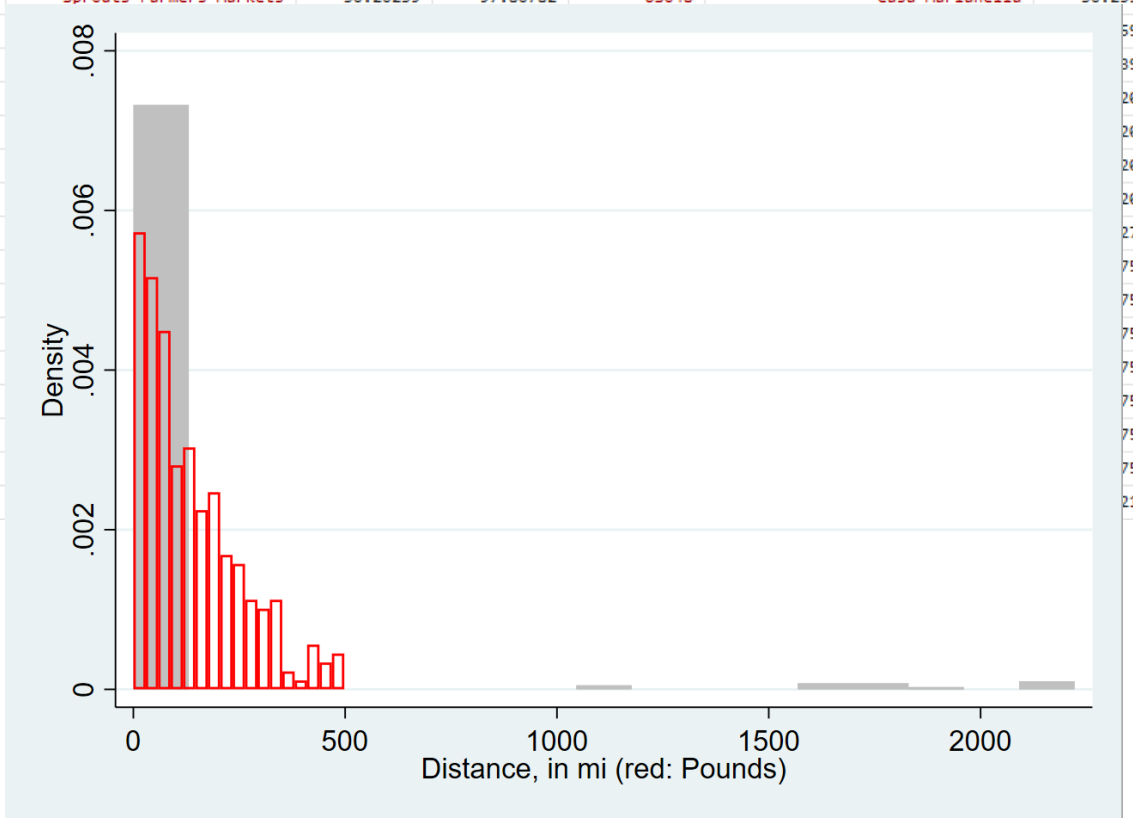
	donorlocat~d	donorname	d_geocodeh~t	d_geocodeh~n	agencynumber	agencyname	ag_geocode~t	ag_geocode~n	pounds	dist	time	georoute_diagnostic
1	B157	Whole Foods Market	30.27065	-97.75356	OS219	Salvation Army Shelter Austin	30.26835	-97.73742	100	1.317307	6.55	OK
2	B185	Walmart Stores, Inc.	30.41358	-97.67568	PA512	Heaven's Harvest	30.37546	-97.71018	27	5.113885	15.81667	OK
3	B533	Target	30.39352	-97.7459	PA705	Travis Heights Food Pantry	30.22794	-97.78591	324	13.69254	17.33333	OK
4	HEB #465	H-E-B Food Stores	30.26048	-97.71157	PA691	Foundation Comm-Arbor Terrace	30.23142	-97.74139	200	5.440726	13.63333	OK
5	HEB #714	H-E-B Food Stores	30.34422	-97.96642	PA156	Lake Travis Crisis Ministries	30.36585	-97.95116	164	1.818753	4.166667	OK
6	B028	Sprouts Farmers Markets	30.20239	-97.80782	OS048	Casa Marianella	30.25988	-97.701	142	11.87192	20.21667	OK
7	B028	Sprouts Farmers Markets	30.20239	-97.80782	OS048	Casa Marianella	30.25988	-97.701	131	11.87192	20.21667	OK
8	B028	Sprouts Farmers Markets	30.20239	-97.80782	PA235	South Austin Neighborhood Ctr	30.23957	-97.76028	35	4.727392	14.1	OK
9	B028	Sprouts Farmers Markets	30.20239	-97.80782	PA368	South Austin Church of Nazaren	30.2049	-97.80619	38	.2379852	1.583333	OK
10	B028	Sprouts Farmers Markets	30.20239	-97.80782	PA368	South Austin Church of Nazaren	30.2049	-97.80619	124	.2379852	1.583333	OK
11	B028	Sprouts Farmers Markets	30.20239	-97.80782	PA368	South Austin Church of Nazaren	30.2049	-97.80619	41	.2379852	1.583333	OK
12	B028	Sprouts Farmers Markets	30.20239	-97.80782	PA368	South Austin Church of Nazaren	30.2049	-97.80619	240	.2379852	1.583333	OK
13	B028	Sprouts Farmers Markets	30.20239	-97.80782	PA705	Travis Heights Food Pantry	30.22794	-97.78591	16	2.524631	8.716666	OK
14	B103	Amazon	49.917558	-119.38694	PA512	Heaven's Harvest	30.37546	-97.71018	72	2221.913	2090.65	OK
15	B103	Amazon	49.917558	-119.38694	PA512	Heaven's Harvest	30.37546	-97.71018	1095	2221.913	2090.65	OK
16	B103	Amazon	49.917558	-119.38694	PA512	Heaven's Harvest	30.37546	-97.71018	171	2221.913	2090.65	OK
17	B103	Amazon	49.917558	-119.38694	PA512	Heaven's Harvest	30.37546	-97.71018	28	2221.913	2090.65	OK
18	B103	Amazon	49.917558	-119.38694	PA512	Heaven's Harvest	30.37546	-97.71018	1140	2221.913	2090.65	OK
19	B103	Amazon	49.917558	-119.38694	PA512	Heaven's Harvest	30.37546	-97.71018	500	2221.913	2090.65	OK
20	B103	Amazon	49.917558	-119.38694	PA512	Heaven's Harvest	30.37546	-97.71018	48	2221.913	2090.65	OK
21	B117	Randall's	30.18452	-97.84858	PA022	Bannockburn Baptist Church	30.2123	-97.83224	61	2.272976	6.7	OK

Export to ArcMap or use mapping packages in Stata

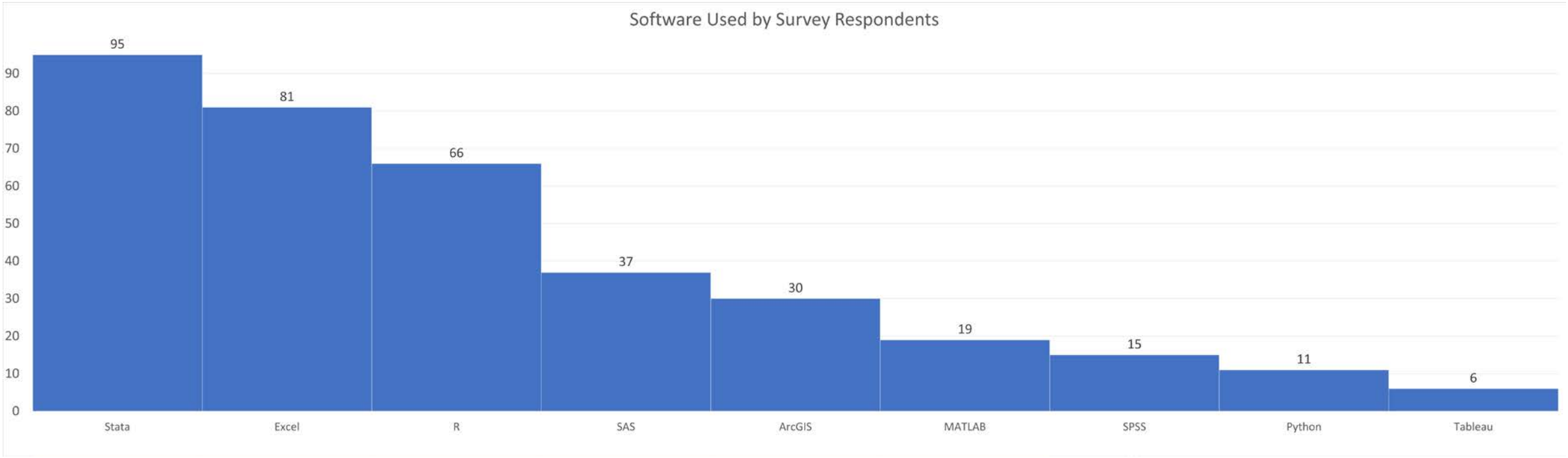
```
georoute, herekey(pe62h3tQj58pWMkAwYIUYNRKOfCpNJn0MUw0WgXQkLZ)
startxy(d_geocodehere_lat d_geocodehere_lon) endxy(ag_geocodehere_lat
ag_geocodehere_lon) di(dist) ti(time) co(p1 p2)
```

georoute results

	donorlocat~d	donorname	d_geocodeh~t	d_geocodeh~n	agencynumber	agencyname	ag_geocode~t	ag_geocode~n	pounds	dist	time	georoute_diagnostic
1	B157	Whole Foods Market	30.27065	-97.75356	OS219	Salvation Army Shelter Austin	30.26835	-97.73742	100	1.317307	6.55	OK
2	B185	Walmart Stores, Inc.	30.41358	-97.67568	PA512	Heaven's Harvest	30.37546	-97.71018	27	5.113885	15.81667	OK
3	B533	Target	30.39352	-97.7459	PA705	Travis Heights Food Pantry	30.22794	-97.78591	324	13.69254	17.33333	OK
4	HEB #465	H-E-B Food Stores	30.26048	-97.71157	PA691	Foundation Comm-Arbor Terrace	30.23142	-97.74139	200	5.440726	13.63333	OK
5	HEB #714	H-E-B Food Stores	30.34422	-97.96642	PA156	Lake Travis Crisis Ministries	30.36585	-97.95116	164	1.818753	4.166667	OK
6	B028	Sprouts Farmers Markets	30.20239	-97.80782	OS048	Casa Marianella	30.25988	-97.701	142	11.87192	20.21667	OK
7	B028						5988	-97.701	131	11.87192	20.21667	OK
8	B028						3957	-97.76028	35	4.727392	14.1	OK
9	B028						2049	-97.80619	38	.2379852	1.583333	OK
10	B028						2049	-97.80619	124	.2379852	1.583333	OK
11	B028						2049	-97.80619	41	.2379852	1.583333	OK
12	B028						2049	-97.80619	240	.2379852	1.583333	OK
13	B028						2794	-97.78591	16	2.524631	8.716666	OK
14	B103						7546	-97.71018	72	2221.913	2090.65	OK
15	B103						7546	-97.71018	1095	2221.913	2090.65	OK
16	B103						7546	-97.71018	171	2221.913	2090.65	OK
17	B103						7546	-97.71018	28	2221.913	2090.65	OK
18	B103						7546	-97.71018	1140	2221.913	2090.65	OK
19	B103						7546	-97.71018	500	2221.913	2090.65	OK
20	B103						7546	-97.71018	48	2221.913	2090.65	OK
21	B117						2123	-97.83224	61	2.272976	6.7	OK



Export to ArcMap or use mapping packages in Stata

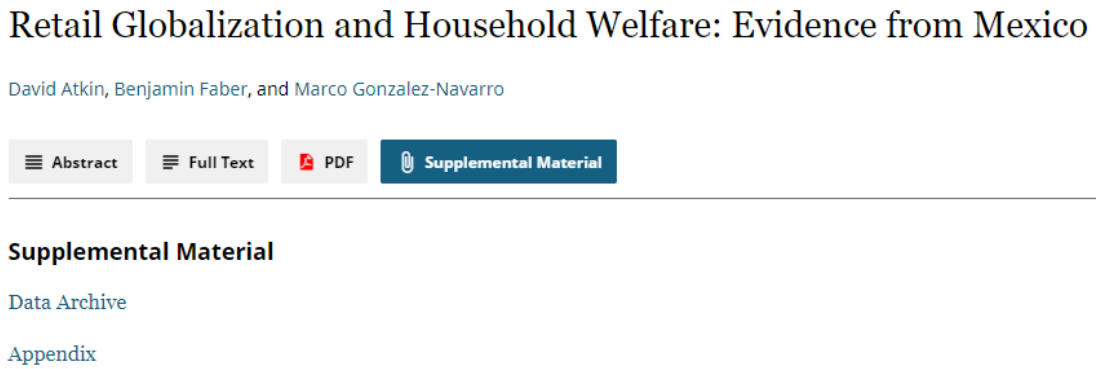


Stata	R	SAS
stnd_compname		TRIM, TRIMN, or STRIP
stnd_address	postmastr	
reclink2	diyar or links	COMPGED or SPEDIS
clrevmatch		
geocodehere	geocode	GEOCODE
georoute	drive_time	ZIPCITYDISTANCE or GEODIST

Takeaways

- 80% of the time, merge is the first-best (and cleanest) option
- Computers make errors too
- Match score from the merge command vs. fuzzy matching
- Match validity from the reclink2 command – weights are arbitrary
- Thresholds – what is good enough?
- Coordinates can be wrong – Re-geocoding with an API is good practice and (in most cases) free to do

Closing Remarks


Retail Globalization and Household Welfare: Evidence from Mexico

David Atkin, Benjamin Faber, and Marco Gonzalez-Navarro

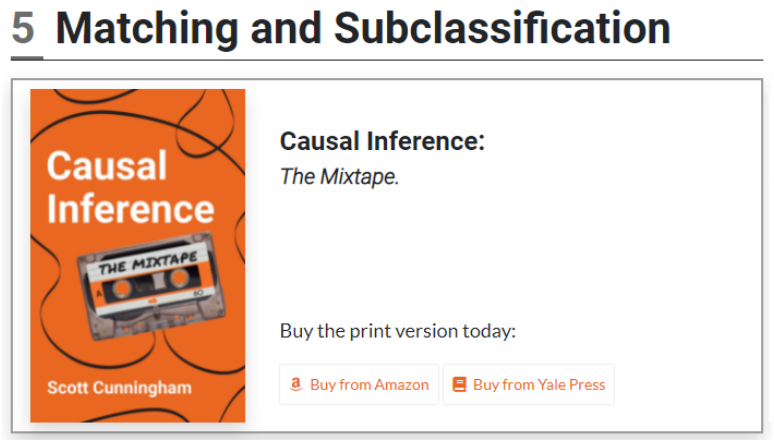
Abstract | Full Text | PDF | Supplemental Material

Supplemental Material

Data Archive

Appendix

1. After reading a paper, find the replication code.
2. Ask for help.
3. Sharing is caring.
4. Don't throw the baby out with the bathwater.
5. Jack of all trades or master of none?


5 Matching and Subclassification

Causal Inference

THE MIXTAPE

Scott Cunningham

Causal Inference:
The Mixtape.

Buy the print version today:

Buy from Amazon | Buy from Yale Press

On this page

5 Matching and Subclassification

5.1 Subclassification

5.2 Exact Matching

5.3 Approximate Matching

5.4 Conclusion

Appendix and Additional Slides

Important terms

Crosswalk is a file you generate that maps a variable in one dataset to a variable in a second dataset.

One-to-one is matching where one observation in one dataset has exactly one match in a second dataset.

Many-to-one is matching where many observations in one dataset have exactly one match in a second dataset.

One-to-many is matching where one observation in one dataset has many possible matches in a second dataset.

Fuzzy matching is when two datasets collect the same piece of information, but the information is recorded differently.