Technical Document

Data Pre-processing of HealthWebMapper2.0

Project conducted by:
San Diego State University
5500 Campanile Drive
San Diego, CA 92182-4493

Report prepared by:
Haihong Huang
Department of Geography, San Diego State University

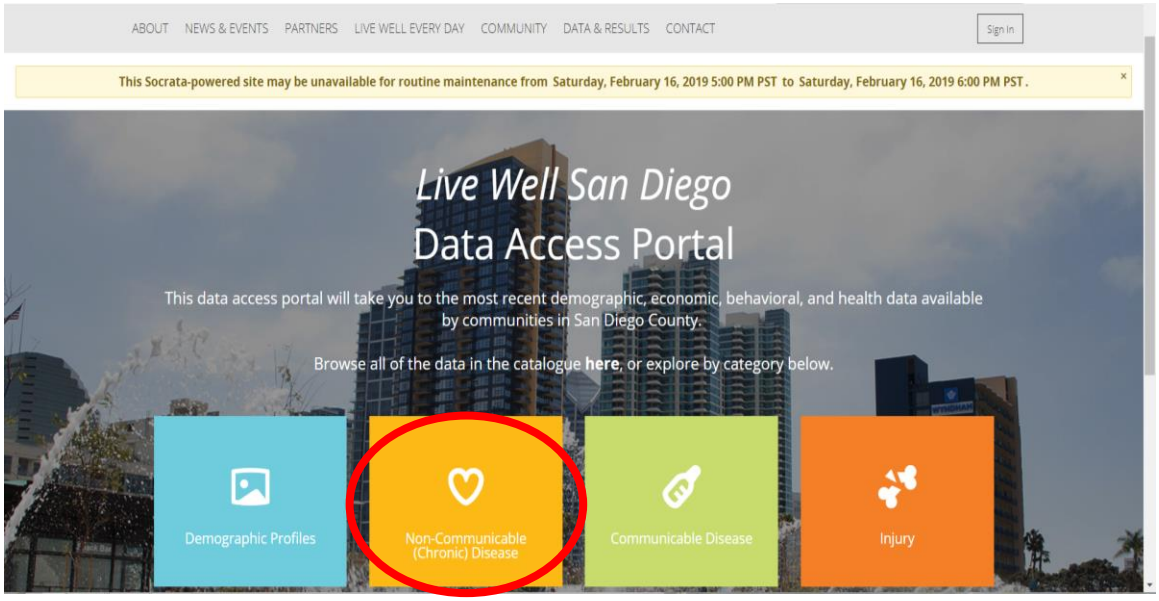
March 2019

Abstract

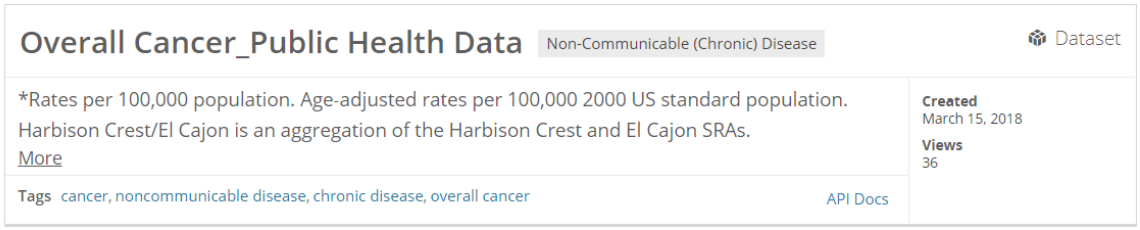
HealthWebMapper 2.0 is a web mapping application designed for visualizing cancer disparities in San Diego Sub-Regional Areas (SRAs). Its upload function allows users to import their interested cancer and socioeconomic and demographic datasets in San Diego sub-regional areas, which are available in Live Well San Diego Data Access Portal. However, the raw datasets need to be preprocessed before uploading to HealthWebMapper2.0. This technical document will help you finish data preprocessing step by step.

Part 1 Preprocessing of Cancer Datasets

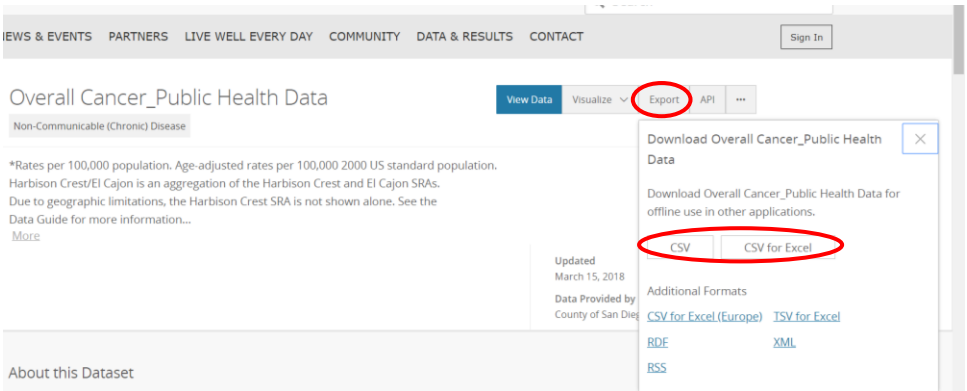
1. Visit Live Well San Diego Data Access Portal(url: https://data.livewellsd.org/) and click Non-Communicable Disease.



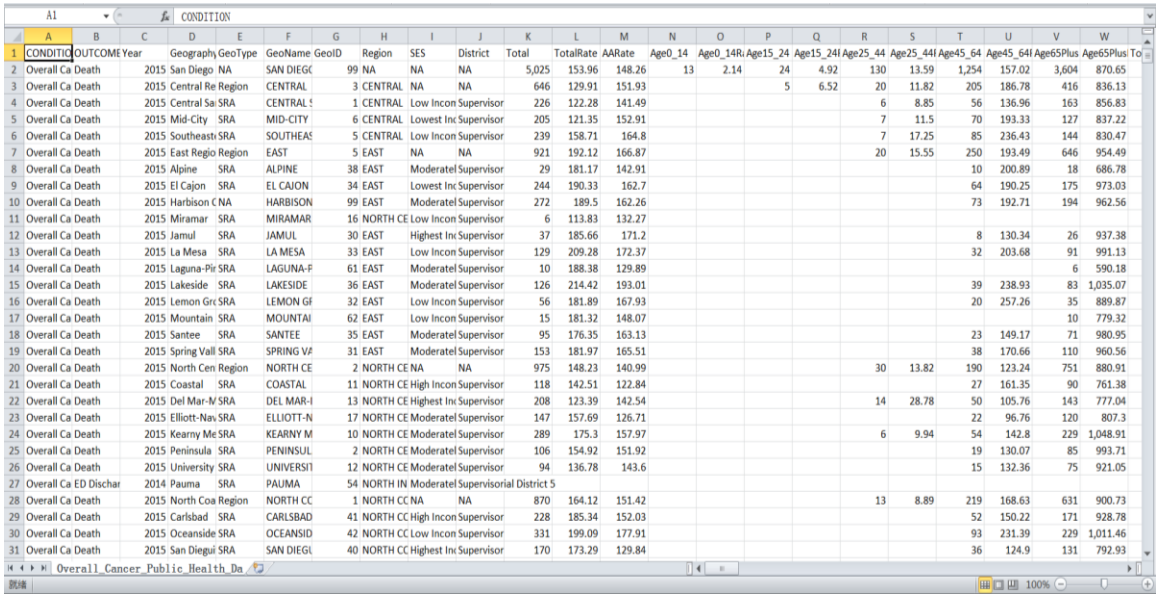
2. Download any cancer datasets you are interested in. For example, we choose “Overall Cancer public health data”, click the link.



3. Download dataset as csv by clicking “Export” button on the upper right side of the website and choose export format as CSV or CSV for Excel

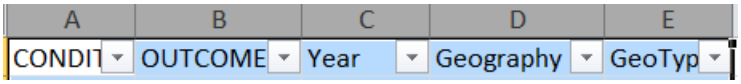


4. Open the downloaded dataset in Excel. All the cancer datasets in this data portal have the similar format. The first 10 columns are qualitative attribute such as condition, outcome, year, SES (Social Economic Status) etc.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	CONDITIO	OUTCOME	Year	Geography	GeoType	GeoName	GeoID	Region	SES	District	Total	TotalRate	AARate	Age0_14	Age0_14R	Age15_24	Age15_24	Age25_44	Age25_44	Age45_64	Age45_64	Age65Plus	Age65Plus	To
2	Overall Ca	Death	2015	San Diego	NA	SAN DIEGO	99	NA	NA	NA	5,025	153.96	148.26	13	2.14	24	4.92	130	13.59	1,254	157.02	3,604	870.65	
3	Overall Ca	Death	2015	Central Re	Region	CENTRAL	3	CENTRAL	NA	NA	646	129.91	151.93			5	6.52	20	11.82	205	186.78	416	836.13	
4	Overall Ca	Death	2015	Central Sa	SRA	CENTRAL S	1	CENTRAL	Low Incon	Supervisor	226	122.28	141.49					6	8.85	56	136.96	163	856.83	
5	Overall Ca	Death	2015	Mid-City	SRA	MID-CITY	6	CENTRAL	Lowest Inc	Supervisor	205	121.35	152.91					7	11.5	70	193.33	127	837.22	
6	Overall Ca	Death	2015	Southeaste	SRA	SOUTHEAS	5	CENTRAL	Low Incon	Supervisor	239	158.71	164.8					7	17.25	85	236.43	144	830.47	
7	Overall Ca	Death	2015	East Regio	Region	EAST	5	EAST	NA	NA	921	192.12	166.87					20	15.55	250	193.49	646	954.49	
8	Overall Ca	Death	2015	Alpine	SRA	ALPINE	38	EAST	Moderatel	Supervisor	29	181.17	142.91							10	200.89	18	686.78	
9	Overall Ca	Death	2015	El Cajon	SRA	EL CAJON	34	EAST	Lowest Inc	Supervisor	244	190.33	162.7							64	190.25	175	973.03	
10	Overall Ca	Death	2015	Harbison C	NA	HARBISON	99	EAST	Moderatel	Supervisor	272	189.5	162.26							73	192.71	194	962.56	
11	Overall Ca	Death	2015	Miramar	SRA	MIRAMAR	16	NORTH CE	Low Incon	Supervisor	6	113.83	132.27											
12	Overall Ca	Death	2015	Jamul	SRA	JAMUL	30	EAST	Highest In	Supervisor	37	185.66	171.2							8	130.34	26	937.38	
13	Overall Ca	Death	2015	La Mesa	SRA	LA MESA	33	EAST	Low Incon	Supervisor	129	209.28	172.37							32	203.68	91	991.13	
14	Overall Ca	Death	2015	Laguna-Pir	SRA	LAGUNA-F	61	EAST	Moderatel	Supervisor	10	188.38	129.89									6	590.18	
15	Overall Ca	Death	2015	Lakeside	SRA	LAKESIDE	36	EAST	Moderatel	Supervisor	126	214.42	193.01							39	238.93	83	1,035.07	
16	Overall Ca	Death	2015	Lemon Gr	SRA	LEMON GF	32	EAST	Low Incon	Supervisor	56	181.89	167.93							20	257.26	35	889.87	
17	Overall Ca	Death	2015	Mountain	SRA	MOUNTAI	62	EAST	Low Incon	Supervisor	15	181.32	148.07									10	779.32	
18	Overall Ca	Death	2015	Santee	SRA	SANTEE	35	EAST	Moderatel	Supervisor	95	176.35	163.13							23	149.17	71	980.95	
19	Overall Ca	Death	2015	Spring Vall	SRA	SPRING VA	31	EAST	Moderatel	Supervisor	153	181.97	165.51							38	170.66	110	960.56	
20	Overall Ca	Death	2015	North Cen	Region	NORTH CE	2	NORTH CE	NA	NA	975	148.23	140.99					30	13.82	190	123.24	751	880.91	
21	Overall Ca	Death	2015	Coastal	SRA	COASTAL	11	NORTH CE	High Incon	Supervisor	118	142.51	122.84							27	161.35	90	761.38	
22	Overall Ca	Death	2015	Del Mar-M	SRA	DEL MAR-I	13	NORTH CE	Highest Inc	Supervisor	208	123.39	142.54					14	28.78	50	105.76	143	777.04	
23	Overall Ca	Death	2015	Elliott-Nav	SRA	ELLIOTT-N	17	NORTH CE	Moderatel	Supervisor	147	157.69	126.71							22	96.76	120	807.3	
24	Overall Ca	Death	2015	Kearny Me	SRA	KEARNY M	10	NORTH CE	Moderatel	Supervisor	289	175.3	157.97					6	9.94	54	142.8	229	1,048.91	
25	Overall Ca	Death	2015	Peninsula	SRA	PENINSUL	2	NORTH CE	Moderatel	Supervisor	106	154.92	151.92							19	130.07	85	993.71	
26	Overall Ca	Death	2015	University	SRA	UNIVERSI	12	NORTH CE	Moderatel	Supervisor	94	136.78	143.6							15	132.36	75	921.05	
27	Overall Ca	ED Dischar	2014	Pauma	SRA	PAUMA	54	NORTH IN	Moderatel	Supervisorial District 5														
28	Overall Ca	Death	2015	North Coa	Region	NORTH CC	1	NORTH CC	NA	NA	870	164.12	151.42					13	8.89	219	168.63	631	900.73	
29	Overall Ca	Death	2015	Carlsbad	SRA	CARLSBAD	41	NORTH CC	High Incon	Supervisor	228	185.34	152.03							52	150.22	171	928.78	
30	Overall Ca	Death	2015	Oceanside	SRA	OCEANSID	42	NORTH CC	Low Incon	Supervisor	331	199.09	177.91							93	231.39	229	1,011.46	
31	Overall Ca	Death	2015	San Diegui	SRA	SAN DIEGU	40	NORTH CC	Highest Inc	Supervisor	170	173.29	129.84							36	124.9	131	792.93	

5. Go to “Data” tab in Excel, click “Filter”. A small drop-down icon will appear in headers of each column. In HealthWebMapper2.0, datasets being uploaded must have unique SRAID, which means there should have no duplicative records. Thus, filter the raw data by a unique value in columns named “CONDITION”, “OUTCOME”, “Year” and “GeoType”.



A	B	C	D	E
CONDIT	OUTCOME	Year	Geography	GeoTyp

For example, when you click drop-down icon in column “OUTCOME”, there are four types of outcomes: death, hospitalization, ED discharge, in patient treatment. Please only select one type of outcomes. Repeat the same process for “Condition” and “Year” but for “GeoType” filter only by “SRA”.

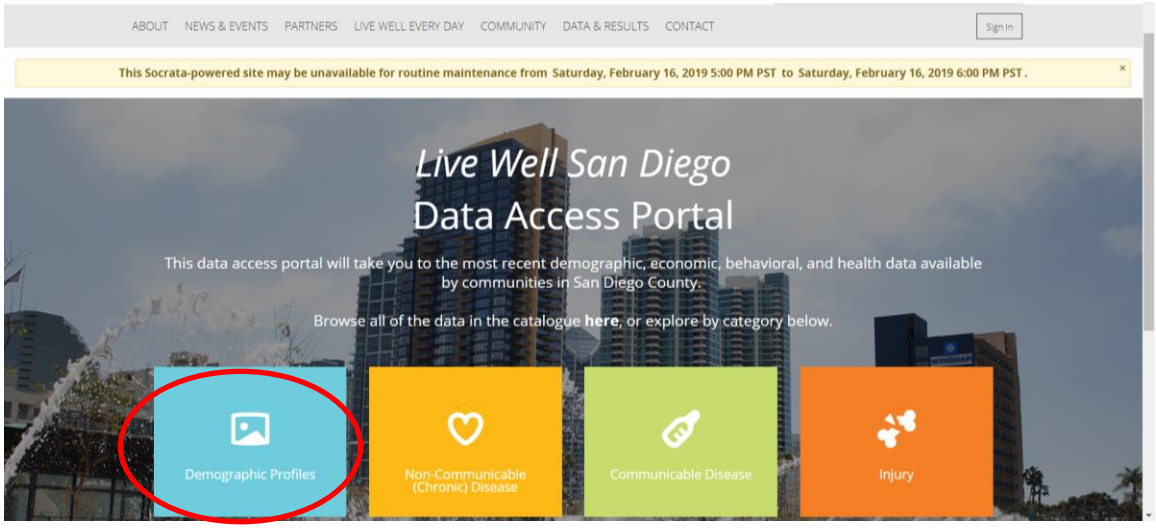
6. Change column name “GEOID” to “SRAID” (all uppercase!)
7. Now in our data table, column 1~10 are qualitative attributes such as cancer outcome, year, geographic location name and social economic status. Starting from column 11 to the end are many quantitative attributes such as Total Case, Total Rate. Due to the display and server limitation of this web application, we recommend users to streamline the datasets, keep no more than 20 columns of attributes. However, you can still keep as many attributes as you want, but the number of attributes keep may affect the display of the data tables as well as the operation speed. As you may found, there are some redundancy information and blank columns; Also, you can change the column

names (header) into more readable names in Excel if needed (e.g. AAR stand for Age Adjusted Rate).

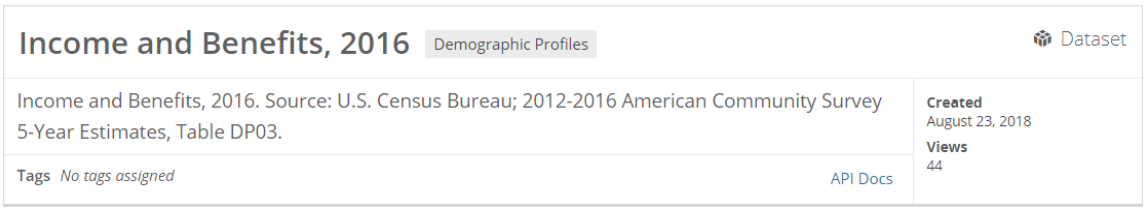
8. Copy you filtered data into a new excel file and save as csv. Your csv data table should have no more than 40 rows including header. Now, you finished the preprocessing of cancer data.

Part 2: Preprocessing of Socioeconomic and Demographic Data

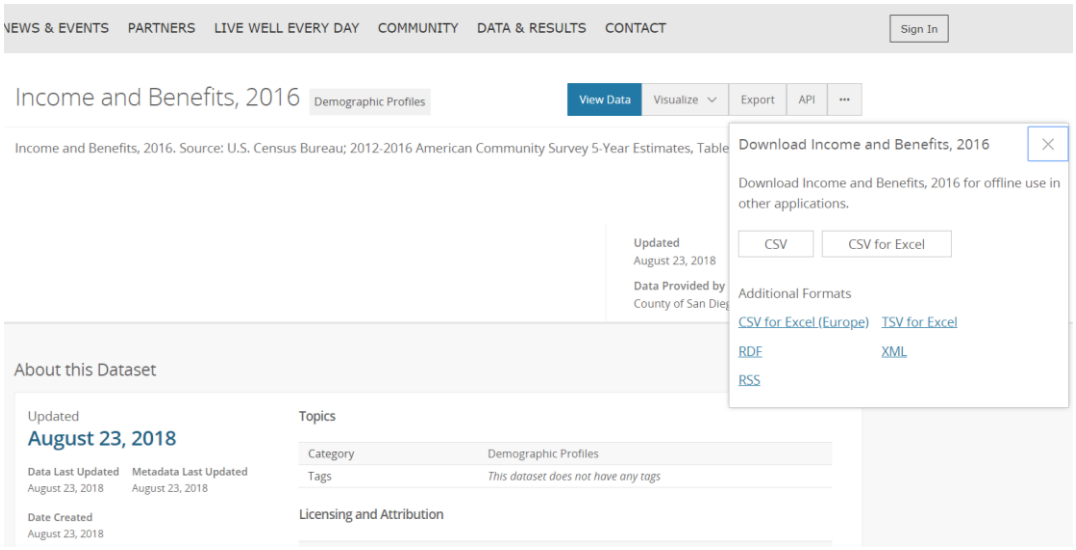
1. Similar to the preprocessing of cancer data, visit Live Well San Diego Data Access Portal (URL: https://data.livewellsd.org/) and click "Demographic Profile" to download socioeconomic and demographic data in San Diego sub-regional areas.



2. Search for the demographic factors that interest you. For example, you want to explore the relations between cancer and income and benefit, click the dataset below:



3. Export as CSV or CSV for Excel



4. Open “Income and Benefits,2016 csv” in Excel and have a look at the standard data structure for demographic profiles provided by Live Well San Diego, which contains only one qualitative attributes “Geography” and the rest are quantitative attributes.

A1 Geography

	A	B	C	D	E	F	G	H	I	J
1	Geography	Total Households	Mean Hou	Household	Household	Household	Household	Household	Households with Foc	
2	Central Region	178,522	$ 69,235.	82.6	20.2	6.4	3.7	13.4	10.7	
3	Central San Diego	77,968	$ 78,574.	83	17.4	4.7	2.3	10.4	6	
4	Mid-City	57,451	$ 62,866.	82.5	19	7.5	4.7	11.3	13.9	
5	Southeastern San Die	43,103	$ 64,892.	82.2	27	8	4.9	21.5	15	
6	East Region	162,775	$ 81,136.	78.9	28.9	7.1	4.4	20.9	11	
7	Alpine	5,755	$ 100,391	78.2	33.6	7.1	2.2	25.2	4.3	
8	El Cajon	42,064	$ 72,083.	78.4	26.6	9.6	8.7	17.4	19.4	
9	Harbison-Crest	5,405	$ 96,133.	72.3	39.9	6.8	3.1	26.8	5.1	
10	Jamul	4,790	$ 128,652	81.6	30.2	4.5	2	24.3	2.4	
11	La Mesa	24,667	$ 77,310.	76.6	29	5.7	2	20.7	8.2	
12	Laguna-Pine Valley	1,847	$ 93,036.	76.2	36.4	4.2	2.6	29.6	5	
13	Lakeside	19,870	$ 84,687.	82.7	26.4	5.8	3.4	18.7	6.8	
14	Lemon Grove	9,603	$ 69,215.	80.3	31.1	8.3	3.8	21.4	12.3	
15	Mountain Empire	2,436	$ 58,117.	58.3	43.8	9.2	1.1	25.7	5.7	
16	Santee	18,277	$ 89,929.	81.4	27.2	6	2.4	22.3	5	
17	Spring Valley	28,061	$ 80,104.	79.7	29.2	6.2	3.9	23.5	12	
18	North Central Region	239,749	$ 108,696	83.1	22.1	3.4	1.3	16.2	2.7	
19	Coastal	35,424	$ 135,661	81	23.4	2	0.7	14.7	1.7	
20	Del Mar-Mira Mesa	56,612	$ 136,314	89.8	17.7	3.5	1.4	14.8	2.5	
21	Elliott-Navajo	35,333	$ 102,405	80.6	27.7	3.2	1.1	22.7	2.6	
22	Kearny Mesa	62,020	$ 83,028.	81.2	24	4.7	1.8	16.6	4	
23	Miramar	515	$ 59,804.	95.5	1	0	0	0	0.8	
24	Peninsula	25,296	$ 99,568.	81.6	22	3.3	1.7	16.6	3	
25	University	24,549	$ 87,519.	80.4	17.9	2.1	1	11.4	1.5	
26	North Coastal Region	181,600	$ 98,015.	78.1	26.4	3.4	2.2	18	4.2	
27	Carlsbad	47,338	$ 111,871	75.4	30.2	2.6	1	20.3	1.9	
28	Oceanside	57,369	$ 76,789.	77.5	26.8	4.9	2.6	20.2	6.3	
29	Pendleton	6,948	$ 49,996.	96.1	0.2	0	2.1	3.5	4.7	
30	San Dieguito	37,249	$ 144,191	77.7	28.4	1.6	1.4	16.7	1.6	
31	Vista	32,696	$ 71,855.	79.4	23.6	4.3	4	15.5	6.8	

5. Add a column named “SRAID” (all uppercase!) in order to link this data with our preloaded SRA Polygon shapefile.

6. Some rows in those dataset are aggregated areas, please delete the rows named: *Central Region, East Region, Harbison-Crest, North Central Region, North Coastal Region, Pendleton, North Inland Region, South Region, San Diego County* in Geography column.
7. Your final data table should have 40 rows, add corresponding SRAID value to each sub-regional area (you can just copy and paste column SRAID below, they are in the same sequence as csv datasets):

Geography	SRAID
Central San Diego	1
Mid-City	6
Southeastern San Diego	5
Alpine	38
El Cajon	34
Jamul	30
La Mesa	33
Laguna-Pine Valley	61
Lakeside	36
Lemon Grove	32
Mountain Empire	62
Santee	35
Spring Valley	31
Coastal	11
Del Mar-Mira Mesa	13
Elliott-Navajo	17
Kearny Mesa	10
Miramar	16
Peninsula	2
University	12
Carlsbad	41
Oceanside	42
San Dieguito	40
Vista	52
Anza-Borrego Springs	63
Escondido	50
Fallbrook	55

North San Diego	14
Palomar-Julian	60
Pauma	54
Poway	15
Ramona	39
San Marcos	51
Valley Center	53
Chula Vista	21
Coronado	3
National City	4
South Bay	22
Sweetwater	20

The new table with column SRAID should look like below:

	A	B	C	D	E	F	G	H	I	J	K
1	SRAID	Geography	Total Households	Mean Hou	Household	Household	Household	Household	Household	Households with Foo	
2	1	Central San Diego	77,968	$ 78,574	83	17.4	4.7	2.3	10.4	6	
3	6	Mid-City	57,451	$ 62,866	82.5	19	7.5	4.7	11.3	13.9	
4	5	Southeastern San Diego	43,103	$ 64,892	82.2	27	8	4.9	21.5	15	
5	38	Alpine	5,755	$ 100,39	78.2	33.6	7.1	2.2	25.2	4.3	
6	34	El Cajon	42,064	$ 72,083	78.4	26.6	9.6	8.7	17.4	19.4	
7	30	Jamul	4,790	$ 128,65	81.6	30.2	4.5	2	24.3	2.4	
8	33	La Mesa	24,667	$ 77,310	76.6	29	5.7	2	20.7	8.2	
9	61	Laguna-Pine Valley	1,847	$ 93,036	76.2	36.4	4.2	2.6	29.6	5	
10	36	Lakeside	19,870	$ 84,687	82.7	26.4	5.8	3.4	18.7	6.8	
11	32	Lemon Grove	9,603	$ 69,215	80.3	31.1	8.3	3.8	21.4	12.3	
12	62	Mountain Empire	2,436	$ 58,117	58.3	43.8	9.2	1.1	25.7	5.7	
13	35	Santee	18,277	$ 89,929	81.4	27.2	6	2.4	22.3	5	
14	31	Spring Valley	28,061	$ 80,104	79.7	29.2	6.2	3.9	23.5	12	
15	11	Coastal	35,424	$ 135,66	81	23.4	2	0.7	14.7	1.7	
16	13	Del Mar-Mira Mesa	56,612	$ 136,31	89.8	17.7	3.5	1.4	14.8	2.5	
17	17	Elliott-Navajo	35,333	$ 102,40	80.6	27.7	3.2	1.1	22.7	2.6	
18	10	Kearny Mesa	62,020	$ 83,028	81.2	24	4.7	1.8	16.6	4	
19	16	Miramar	515	$ 59,804	95.5	1	0	0	0	0.8	
20	2	Peninsula	25,296	$ 99,568	81.6	22	3.3	1.7	16.6	3	
21	12	University	24,549	$ 87,519	80.4	17.9	2.1	1	11.4	1.5	
22	41	Carlsbad	47,338	$ 111,87	75.4	30.2	2.6	1	20.3	1.9	
23	42	Oceanside	57,369	$ 76,789	77.5	26.8	4.9	2.6	20.2	6.3	
24	40	San Dieguito	37,249	$ 144,19	77.7	28.4	1.6	1.4	16.7	1.6	
25	52	Vista	32,696	$ 71,855	79.4	23.6	4.3	4	15.5	6.8	
26	63	Anza-Borrego Springs	1,540	$ 40,464	60	49.4	15.5	4.9	31.2	14.6	
27	50	Escondido	51,902	$ 78,318	80.2	25.3	4.9	2.9	16	8.8	
28	55	Fallbrook	17,205	$ 84,685	72.3	37.1	3.6	2.3	23.4	6.2	
29	14	North San Diego	41,213	$ 116,50	83.6	23.7	4.2	1.1	18.5	2.6	
30	60	Palomar-Julian	2,174	$ 57,788	59.4	52.6	3.2	5.6	30.9	12.4	
31	54	Pauma	2,154	$ 86,663	80.5	21.8	1.2	2.2	11.9	5.8	
32	15	Poway	30,283	$ 135,75	81.5	27	3.1	1	21.2	2	
33	39	Ramona	11,645	$ 96,704	81.7	27.8	4.6	3.4	21.4	6.5	
34	51	San Marcos	31,830	$ 73,250	74.3	29.6	3.3	1.7	19	4.8	
35	53	Valley Center	7,764	$ 92,496	74.7	37.4	2.4	0.9	25.6	4.4	
36	21	Chula Vista	36,903	$ 57,674	76.7	29.7	7.5	3.9	19.9	14.6	
37	3	Coronado	8,986	$ 137,12	71.7	34.9	1.7	0.7	27.1	0.8	
38	4	National City	15,623	$ 53,862	80.4	28.9	9.1	5.2	13.8	15.7	
39	22	South Bay	38,435	$ 63,088	82.9	28.7	9	5	18.6	15.9	
40	20	Sweetwater	42,825	$ 103,49	88.2	22.1	4.9	1.8	19.9	4.1	

Now you finished the preprocessing of socioeconomic data. Repeat the same procedure for other datasets you find in Live Well San Diego Data Access Portal and visualize them in HealthWebMapper2.0