

Census Data Cleaning Guide

Authors

Scot Hall
Annie Cerminara
Ademola Adeyemo

Terms of Use

This script is licensed under a [Creative Commons — Attribution 4.0 International — CC BY 4.0](https://creativecommons.org/licenses/by/4.0/). It is attributed to Scot Hall, Annie Cerminara and Ademola Adeyemo.

Usage Guide

These instructions support a workflow for simplifying census data. The workflow includes obtaining census shapefile, data and metadata from the U.S. Census and using MS Excel to create wanted fields.

Obtain Census TIGER Shapefile:

1. Go to <http://www.census.gov/>
2. Click on the **Geography** link at the top of the page
3. Click on the **Maps & Data** box



4. Click on **TIGER Products**
5. Click on **TIGER/Line Shapefiles** for current year

TIGER Products

TIGER = Topologically Integrated Geographic Encoding and Referencing
TIGER products are spatial extracts from the Census Bureau's MAF/TIGER
Census Bureau offers several file types and an online mapping application.

- [TIGER/Line Shapefiles - New 2016 Shapefiles](#)
- [TIGER/Line Geodatabases](#)
- [TIGER/Line with Selected Demographic and Economic Data](#)
- [Cartographic Boundary Shapefiles](#)
- [KML - Cartographic Boundary Files](#)
- [TIGERweb](#)

6. Click on the **Download** dropdown and then **Web interface**



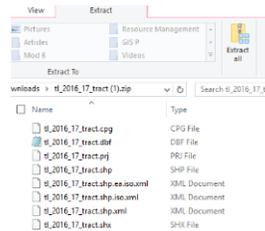
7. Select **Year** and **Layer Type** then **Submit**



8. Follow next cues until you reach the desired download

9. The result is a Zip file

a. Extract the file to the desired folder



10. In this tool, we will be using the .shp file but files included in the download will be required for it to operate correctly.

Obtain Census Attribute Data File:

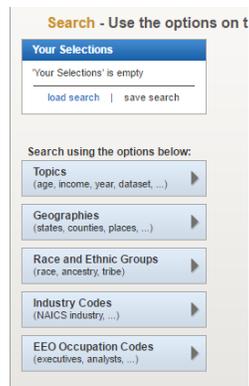
1. Go to <http://www.census.gov/>
2. Click on the **Data** link at the top of the page
3. Click on the **Data Tools & Apps** box
4. Click on **American Fact Finder**



- Click on **Advanced Search** and then **SHOW ME ALL**



- Work through the options on the left to find desired data



Make sure the geographies match the attribute data

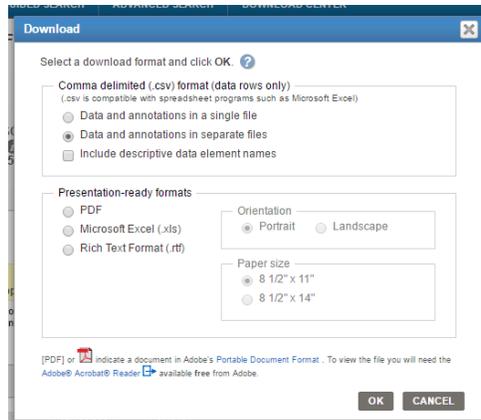
- Click on row to select desired data

The screenshot shows the search results page for 'School Enrollment'. The table has columns for 'ID', 'Table, File or Document Title', 'Dataset', and 'About'. The second row is selected, indicated by a checkmark in the first column.

ID	Table, File or Document Title	Dataset	About
<input type="checkbox"/>	S1401 ★ SCHOOL ENROLLMENT	2014 ACS 5-year estimates	?
<input checked="" type="checkbox"/>	B14001 ★ SCHOOL ENROLLMENT BY LEVEL OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER	2014 ACS 5-year estimates	?
<input type="checkbox"/>	B14007 ★ SCHOOL ENROLLMENT BY DETAILED LEVEL OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER	2014 ACS 5-year estimates	?
<input type="checkbox"/>	B14003 ★ SEX BY SCHOOL ENROLLMENT BY TYPE OF SCHOOL BY AGE FOR THE POPULATION 3 YEARS AND OVER	2014 ACS 5-year estimates	?

- Click **Download**

- In Download dialog box:
 - Select Data and annotations in separate files
 - Uncheck Include descriptive data element names
 - Do not select any options under Presentation-ready formats
 - Select OK



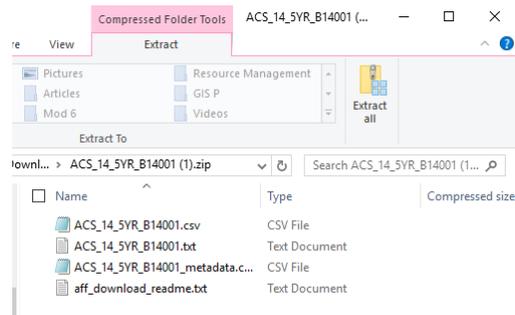
10. When prompted, select **Download** from the pop up window



11. The result is a Zip file

a. Extract the file to the desired folder

12. This file contains a .csv file(CensusData.csv) and associated text file, metadata.csv file and a readme.txt file.



Create a .csv file for User Designated Fields

Procedure:

1. **Open** the Census metadata file you downloaded in the previous step. This should be clearly marked as the metadata file and will look similar to this (Please note that your file will not contain the same data as the example):

	A	B	C	D	E	F	G	H	I
1	GEO.id	id							
2	GEO.id2	id2							
3	GEO.display-label	Geography							
4	HD01_VD01	Estimate; Total:							
5	HD02_VD01	Margin of Error; Total:							
6	HD01_VD02	Estimate; Enrolled in school:							
7	HD02_VD02	Margin of Error; Enrolled in school:							
8	HD01_VD03	Estimate; Enrolled in school: - Enrolled in nursery school, preschool							
9	HD02_VD03	Margin of Error; Enrolled in school: - Enrolled in nursery school, preschool							
10	HD01_VD04	Estimate; Enrolled in school: - Enrolled in kindergarten							

It will contain a list of id's and a description of what those codes correspond to.

2. **Create** a new document and save it in *.csv format with a new file name ([Desired_Census Data.csv](#)). Note the file name as it will need to be inserted when running the script.
3. **Copy** the code along with the description from the metadata file, and paste it into your new file that you created. Repeat this for each field you would like to show up in your User Designated Fields file. There is no need to copy anything other than census codes and their descriptions.

	A	B	C	D	E	F	G
1	HD01_VD02	Estimate; Enrolled in school:					
2	HD01_VD04	Estimate; Enrolled in school: - Enrolled in kindergarten					
3	HD01_VD07	Estimate; Enrolled in school: - Enrolled in grade 9 to grade 12					
4	HD01_VD09	Estimate; Enrolled in school: - Graduate or professional school					
5	HD02_VD10	Margin of Error; Not enrolled in school					
6							

(Actual results will vary depending on the data you copy)

4. Once all the fields you wish to keep are transferred to the new file, **Rename** the descriptions with the field names you wish to have in the final document.

	A	B	C	D
1	HD01_VD02	Enrolled		
2	HD01_VD04	Kindergarden		
3	HD01_VD07	9th - 12th		
4	HD01_VD09	Graduate		
5	HD02_VD10	Not Enrolled		

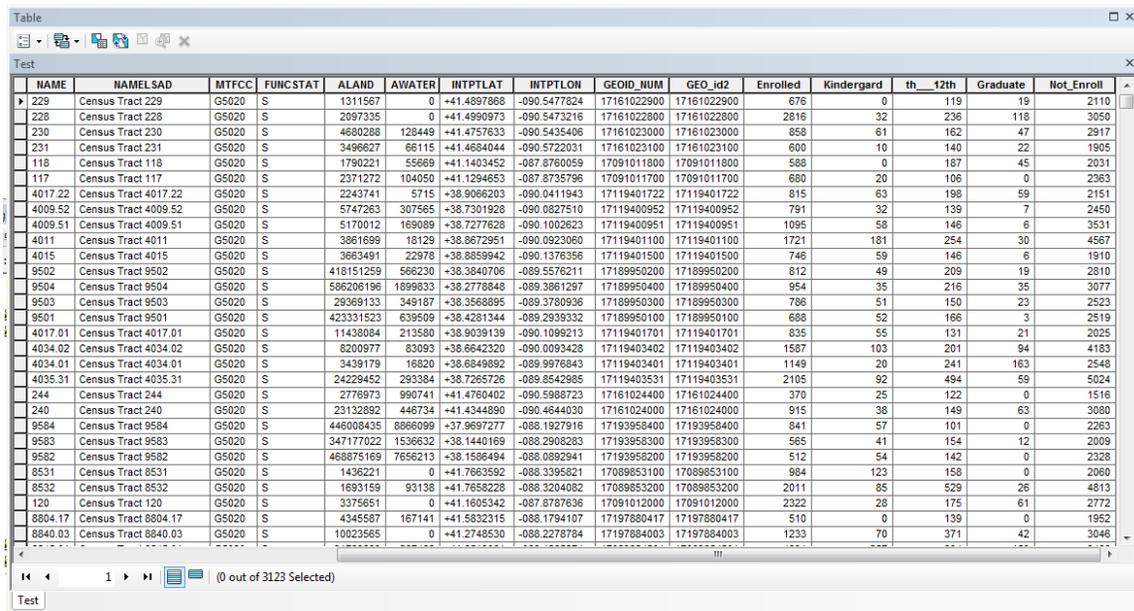
5. **Save** your new file (make sure it is *.csv format).

Run the Tool

Run the tool using the necessary parameters

Results

A copy of your shapefile will be added to your map with the designated fields you created joined with your feature class data. Please note that the field names you created might be renamed based on length and what characters are allowed. Below is an example of the joined attribute table that is created:



The screenshot shows a software window titled "Table" with a toolbar and a data table. The table has 14 columns: NAME, NAME_LASD, MTFCC, FUNCSTAT, ALAND, AWATER, INTPTLAT, INTPTLON, GEOID_NUM, GEO_id2, Enrolled, Kindergard, th_12th, Graduate, and Not_Enroll. The data rows represent various census tracts, such as Census Tract 229, 228, 230, etc., with corresponding values for each field. The status bar at the bottom indicates "(0 out of 3123 Selected)".

NAME	NAME_LASD	MTFCC	FUNCSTAT	ALAND	AWATER	INTPTLAT	INTPTLON	GEOID_NUM	GEO_id2	Enrolled	Kindergard	th_12th	Graduate	Not_Enroll
229	Census Tract 229	G5020	S	1311567	0	+41.4897868	-090.5477824	17161022900	17161022900	676	0	119	19	2110
228	Census Tract 228	G5020	S	2097335	0	+41.4990973	-090.5473216	17161022800	17161022800	2816	32	236	118	3050
230	Census Tract 230	G5020	S	4680288	128449	+41.4757633	-090.5435406	17161023000	17161023000	858	61	162	47	2917
231	Census Tract 231	G5020	S	3496627	66115	+41.4684044	-090.5722031	17161023100	17161023100	600	10	140	22	1905
118	Census Tract 118	G5020	S	1790221	55669	+41.1403452	-087.8760059	17091011800	17091011800	588	0	187	45	2031
117	Census Tract 117	G5020	S	2371272	104050	+41.1294653	-087.8735796	17091011700	17091011700	680	20	106	0	2363
4017.22	Census Tract 4017.22	G5020	S	2243741	5715	+38.9066203	-090.0411943	17119401722	17119401722	815	63	198	59	2151
4009.52	Census Tract 4009.52	G5020	S	5747263	307565	+38.7301928	-090.0827510	17119400952	17119400952	791	32	139	7	2450
4009.51	Census Tract 4009.51	G5020	S	5170012	169089	+38.7277628	-090.1002623	17119400951	17119400951	1095	58	146	6	3531
4011	Census Tract 4011	G5020	S	3861699	18129	+38.8672951	-090.0923060	17119401100	17119401100	1721	181	254	30	4567
4015	Census Tract 4015	G5020	S	3663491	22978	+38.8859942	-090.1376356	17119401500	17119401500	746	59	146	6	1910
9502	Census Tract 9502	G5020	S	418151259	566230	+38.3840706	-089.5576211	17189950200	17189950200	812	49	209	19	2810
9504	Census Tract 9504	G5020	S	586206196	1899833	+38.2778848	-089.3861297	17189950400	17189950400	954	35	216	35	3077
9503	Census Tract 9503	G5020	S	29369133	349187	+38.3568895	-089.3780936	17189950300	17189950300	786	51	150	23	2523
9501	Census Tract 9501	G5020	S	423331523	639509	+38.4281344	-089.2939332	17189950100	17189950100	688	52	166	3	2519
4017.01	Census Tract 4017.01	G5020	S	11438084	213580	+38.9039139	-090.1099213	17119401701	17119401701	835	55	131	21	2025
4034.02	Census Tract 4034.02	G5020	S	8200977	83093	+38.6642320	-090.0093428	17119403402	17119403402	1587	103	201	94	4183
4034.01	Census Tract 4034.01	G5020	S	3439179	16820	+38.6849892	-089.9976843	17119403401	17119403401	1149	20	241	163	2548
4035.31	Census Tract 4035.31	G5020	S	24229452	293384	+38.7265726	-089.8542985	17119403531	17119403531	2105	92	494	59	5024
244	Census Tract 244	G5020	S	2776973	990741	+41.4760402	-090.5988723	17161024400	17161024400	370	25	122	0	1516
240	Census Tract 240	G5020	S	23132892	446734	+41.4344890	-090.4644030	17161024000	17161024000	915	38	149	63	3080
9584	Census Tract 9584	G5020	S	446008435	8860999	+37.9697277	-088.1927916	17193958400	17193958400	841	57	101	0	2263
9583	Census Tract 9583	G5020	S	347177022	1536632	+38.1440169	-088.2908283	17193958300	17193958300	565	41	154	12	2009
9582	Census Tract 9582	G5020	S	468875169	7656213	+38.1586494	-088.0892941	17193958200	17193958200	512	54	142	0	2328
8531	Census Tract 8531	G5020	S	1436221	0	+41.7663592	-088.3395821	17089853100	17089853100	984	123	158	0	2060
8532	Census Tract 8532	G5020	S	1693159	93138	+41.7658228	-088.3204082	17089853200	17089853200	2011	85	529	26	4813
120	Census Tract 120	G5020	S	3375651	0	+41.1605342	-087.8787636	17091012000	17091012000	2322	28	175	61	2772
8804.17	Census Tract 8804.17	G5020	S	4345587	167141	+41.5832315	-088.1794107	17197880417	17197880417	510	0	139	0	1952
8840.03	Census Tract 8840.03	G5020	S	10023565	0	+41.2748530	-088.2278784	17197884003	17197884003	1233	70	371	42	3046