TUTORIAL 7.1 | FINDING AND CLEANING CENSUS DATA

Goals

- Search, download, and clean census data for QGIS from the Census FactFinder website.
- Join the relevant shapefile.
- Visualize additional variables.

Introduction

In previous tutorials, you used census data that had already been downloaded and cleaned for use with the TIGER shapefiles. In this tutorial, you'll learn how to use the census platform directly, how to "harmonize" the data with the TIGER shapefiles, and how to simplify the number of data columns.

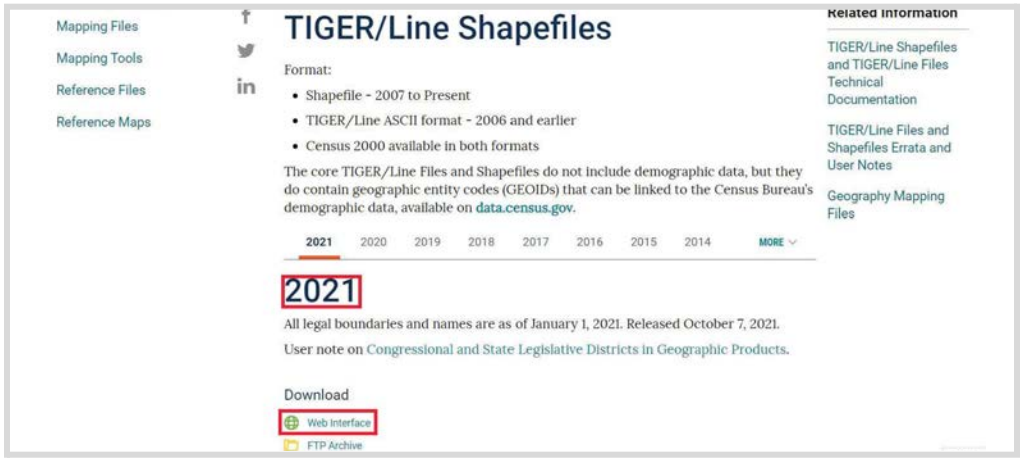
It's helpful to come into the census website, called "FactFinder", with an idea of what variable you want, since you can search directly by variable name (eg. B25003, the ACS table for the census race question we used in an earlier tutorial). However, the FactFinder website also allows you to search and filter by topic and by geography. I like to start filtering by geography, which allows you to choose the geographic resolution such as "census tract", "census block", "county", "state", and so on, since I usually know what kind of question I'm trying to answer, and what resolution I'll need to answer it. Once I have the geography defined, if I don't know the exact table I'm looking for, I'll filter the results by "topic".

When you've found a table you want to download, it's often helpful to preview it first in the census website's map feature. You can choose which variable in the table to preview and get a preliminary sense of what geographical patterns you might see.

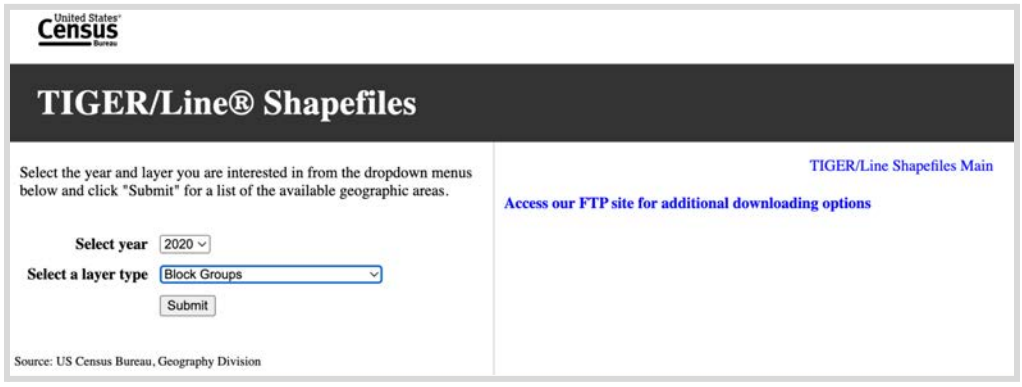
In this tutorial, you'll be looking at the same geographic level (block groups), and a new census topic (health insurance coverage).

Step 1: Download the Block Group level TIGER shapefiles for 2010 and 2020.

1a First, go to https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html
Select **"Web Interface"** at the bottom.



1b On the next page, select **"2020" and "Block Group"**. On the following page, select "**Virginia**" and **download**. Because you downloaded the entire state and you only want your city, you'll need to filter that file later in QGIS.



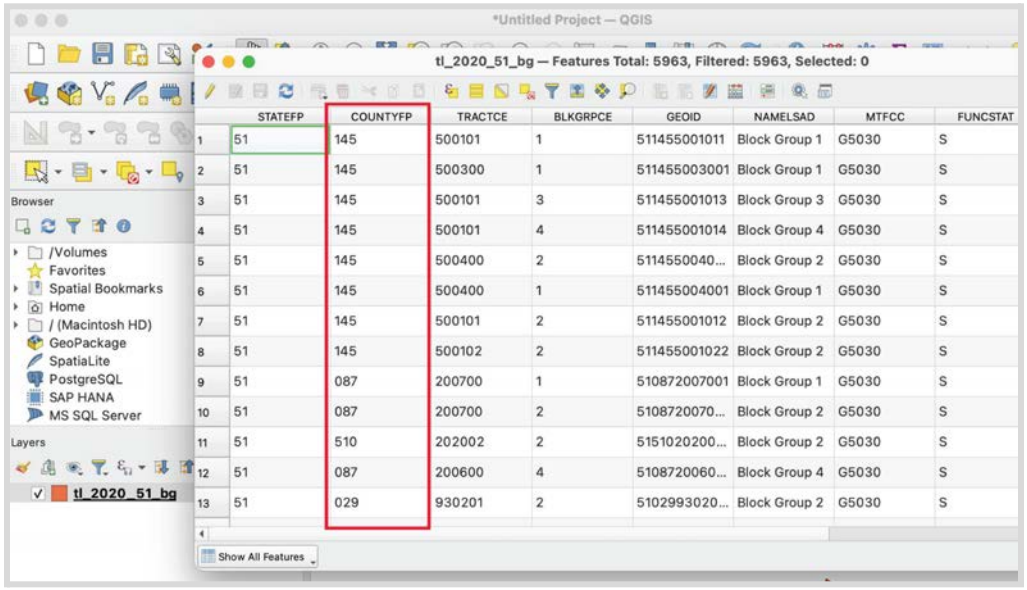
1c Next, go back a page, change the year to **"2010" and "Block Group".** On the following page, you'll have the option to select a smaller geography: select "Virginia", and then the city that you're working on. I selected "Roanoke City". Again, **download**.

Step 2: Open and adjust TIGER files in QGIS.

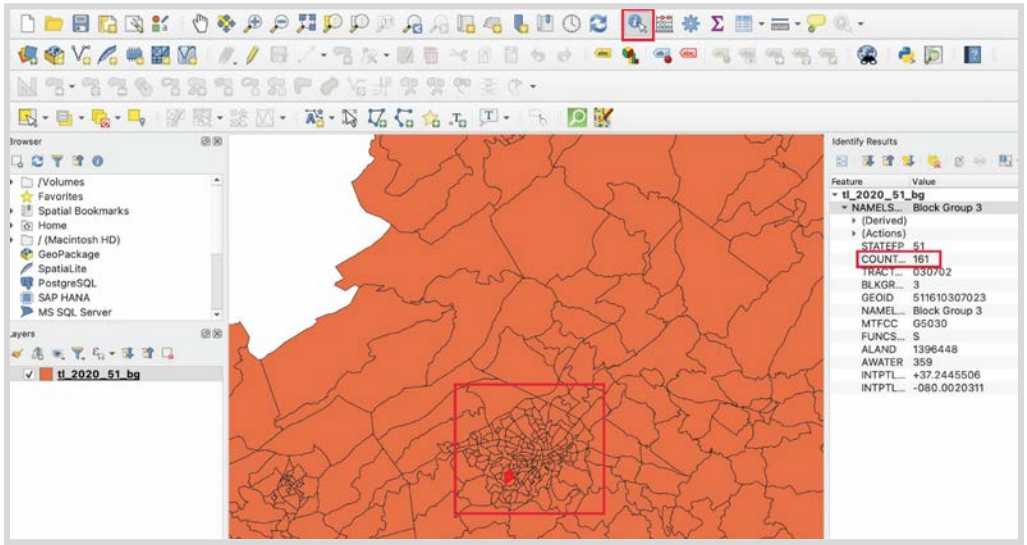
2a First, create a new QGIS file and set the **CRS to UTM 17N**.

2b Import the 2010 and 2020 Block Group Shapefiles. The 2020 Shapefile needs to be downsized to just your city. There are a few ways to do this. You can select the appropriate block groups manually. You can select them by overlap of the 2010 layer, or of the city's administrative boundary. However, we're going to select them using the Attribute Table.

2c First, open the **Attribute Table** of your 2020 Shapefile and take a look at the columns. Notice that the "State FP" column is the same for every block group, "51", since they're all in Virginia, but the "**County FP**" varies by county. By identifying the County FP of your city you can select all the Block Groups in your city. Using the **Select tool** I can click on one block group in the city to identify the County FP.

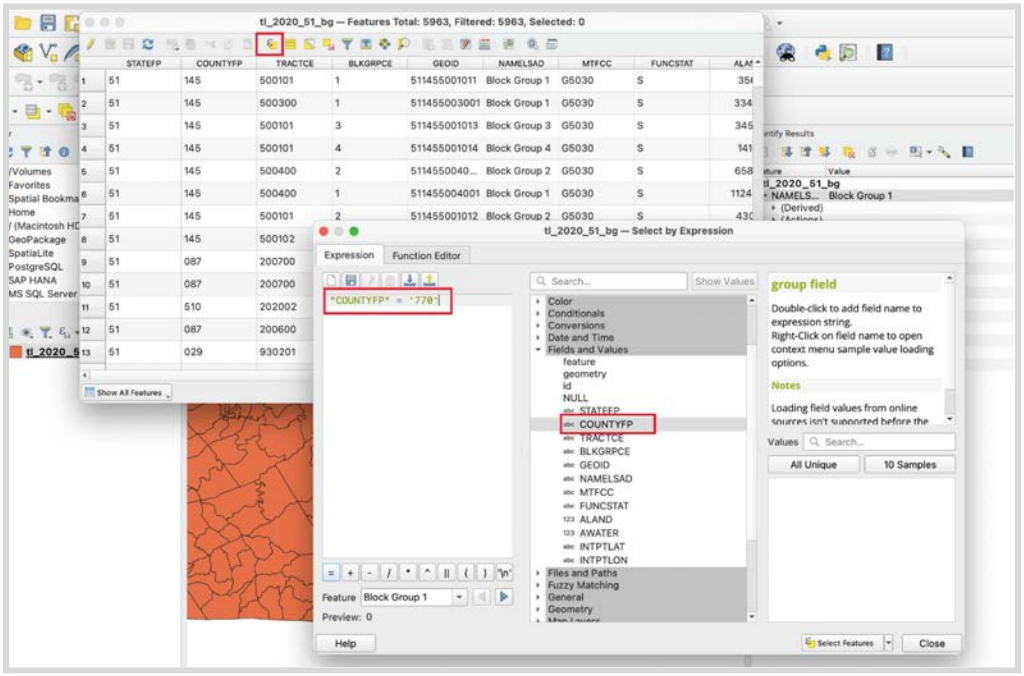


	STATEFP	COUNTYFP	TRACTCE	BLKGRPCE	GEOID	NAMELSAD	MTFCC	FUNCSTAT
1	51	145	500101	1	511455001011	Block Group 1	G5030	S
2	51	145	500300	1	511455003001	Block Group 1	G5030	S
3	51	145	500101	3	511455001013	Block Group 3	G5030	S
4	51	145	500101	4	511455001014	Block Group 4	G5030	S
5	51	145	500400	2	5114550040...	Block Group 2	G5030	S
6	51	145	500400	1	511455004001	Block Group 1	G5030	S
7	51	145	500101	2	511455001012	Block Group 2	G5030	S
8	51	145	500102	2	511455001022	Block Group 2	G5030	S
9	51	087	200700	1	510872007001	Block Group 1	G5030	S
10	51	087	200700	2	5108720070...	Block Group 2	G5030	S
11	51	510	202002	2	5151020200...	Block Group 2	G5030	S
12	51	087	200600	4	5108720060...	Block Group 4	G5030	S
13	51	029	930201	2	5102993020...	Block Group 2	G5030	S

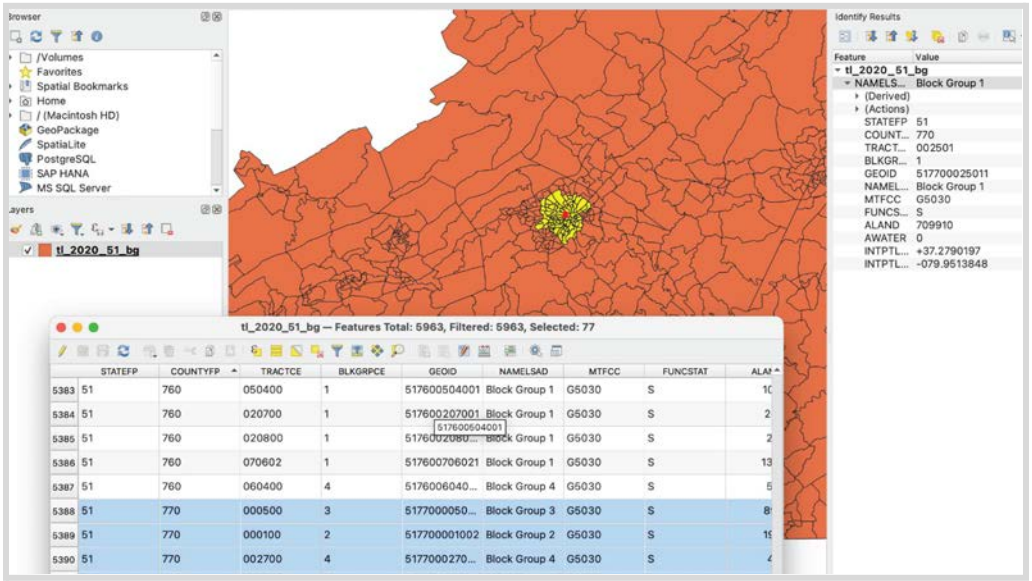


County FP
Roanoke: 770
Richmond: 760
Lynchburg: 680

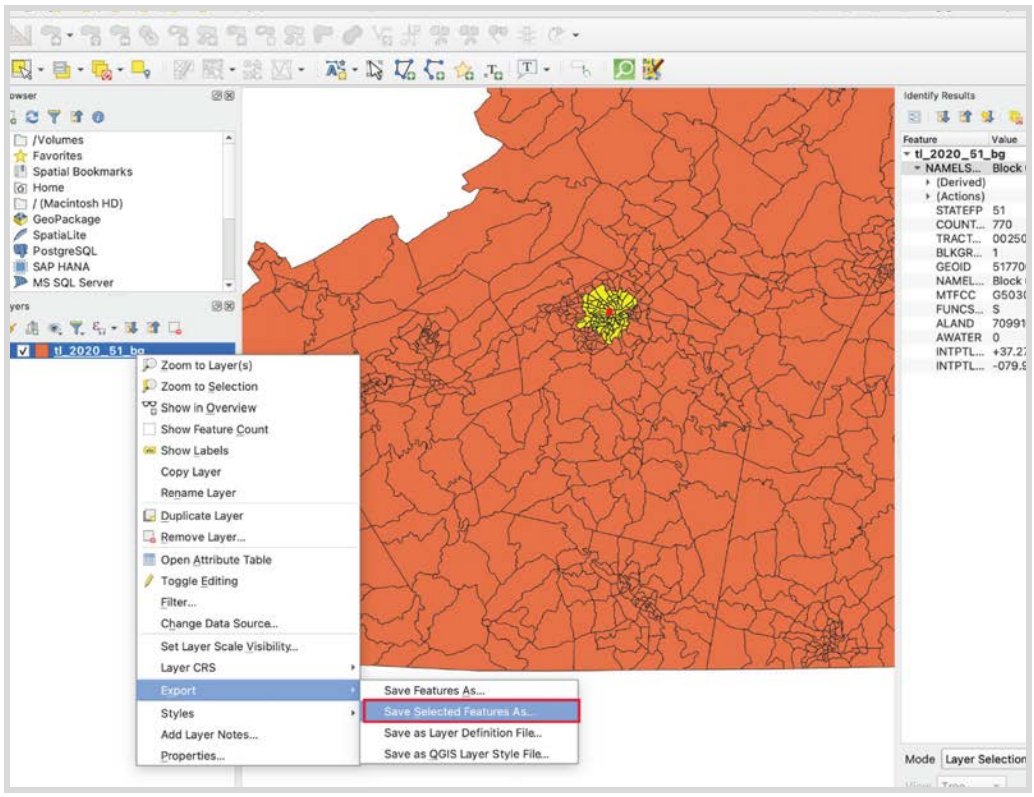
2d In the **Attribute Table**, open the **Expression Editor** (click “Select by Expression”, the yellow square on the sigma symbol along the top toolbar). Like with the symbology expression editor, you can select variables in the attribute table from ”Fields and Values”. Select “COUNTYFP” and then equals, and your FP in single quotation marks. So, in my case, “COUNTYFP” = ‘770’.



Click “**Select Features**” and then check the map to see that the block groups in your city are selected.

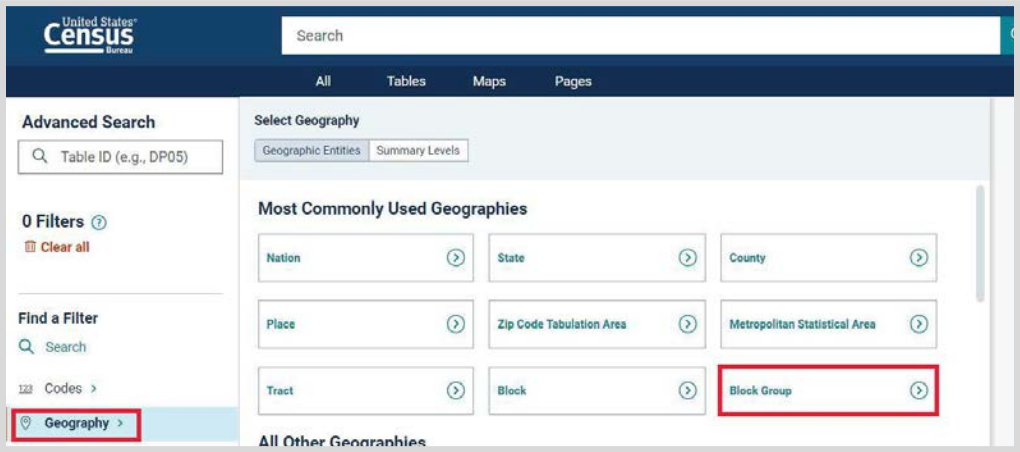


2e Right click on the Shapefile layer and select **Export > Save Selected Features As…** (make sure to choose "selected" features). Save the selected block groups as a new layer in your data folder. Check "**Add to map**" and then Save.

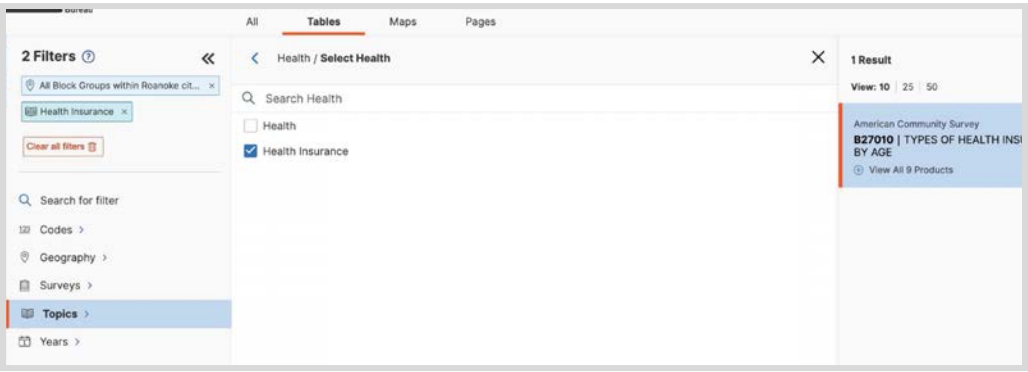


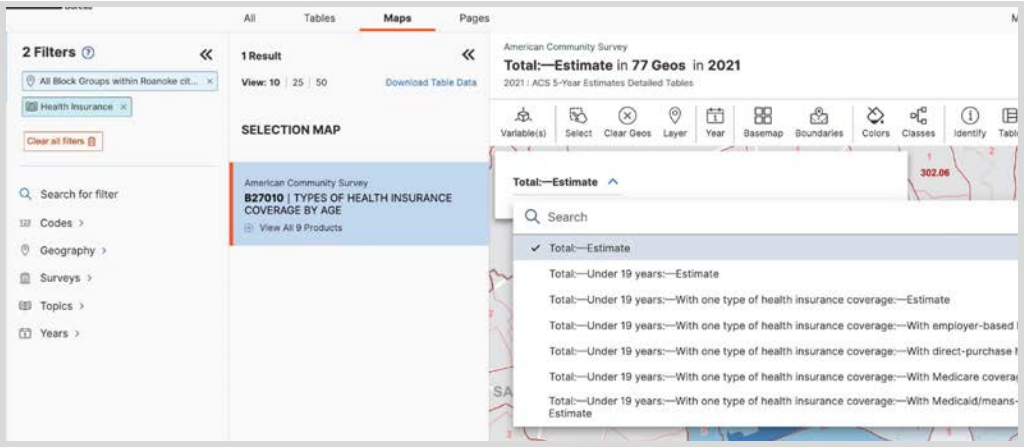
Step 3: download tables from the Census FactFinder

3a Go to **Census FactFinder**: https://data.census.gov/cedsci/table. You can filter the available datasets with "**Advanced Search**". First, filter by **"Geography"** on the left**: "Block Groups" > "Virginia" > Your City** (eg "Roanoke City") **>** then select **"All Block Groups** within that city". You'll see around 1200 tables available at this geographic level.



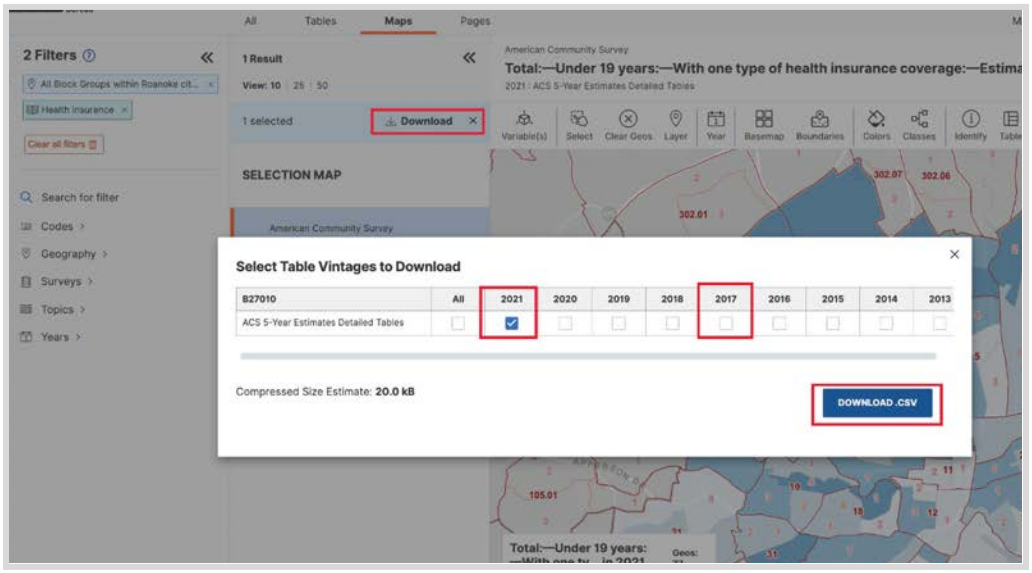
3b Next, filter by **"Topics" ("Health" > "Health Insurance")**. Look for the table B27010 in the search results. Before we download it, check to see what the data looks like with the census "Maps" tab at the top toolbar. This is also a useful way to see what data each table has available with the dropdown menu "**Variables**" in the map.





From this, we can see that the data are broken down by age. Since we want to see the total insured population, we'll need to combine some of these columns and remove others. We'll do this in a spreadsheet editor after downloading the tables.

3c Download the table by **checking the box** beside the table and clicking "**download**" at the top. Select the 5-year estimates "2021" (including 2017-2021) and "2017" (including 2013-2017), then click "download".

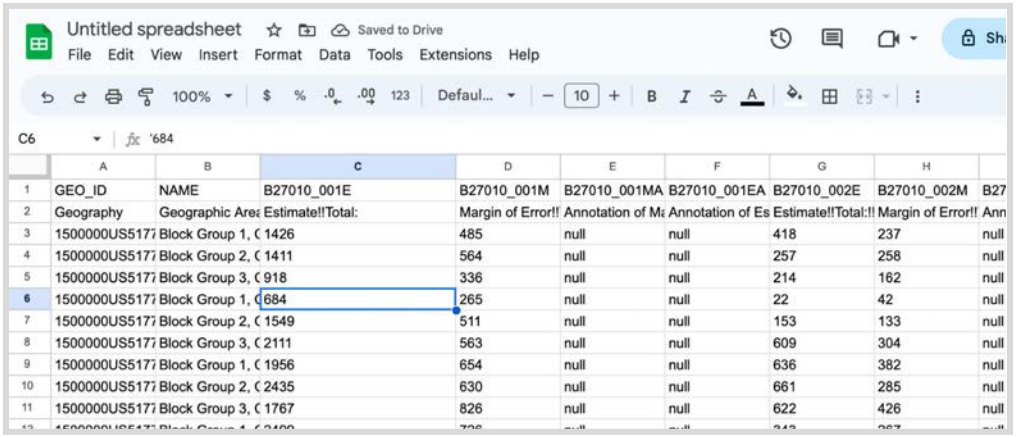


Step 4: Clean up Census data spreadsheets.

I prefer to use Google Sheets for this step, but any spreadsheet editor will work similarly.

Before we import the data, we're going to clean it up. We want to **1) simplify the data, 2) rename the columns and remove the extra header row, and 3) convert the GEOID column to match the shapefile.** Note that the downloaded census data contains three files: the table notes, the metadata, and the data. You only need the data file, but the table notes contain an explanation of the data itself, which can be helpful to read through.

4a Unzip the census file. Then, go to Google Sheets (type **"sheet.new**" into your browser) and upload the 2021 **data** csv for table B27010 that you downloaded (File>Import).



	A	B	C	D	E	F	G	H	
1	GEO_ID	NAME	B27010_001E	B27010_001M	B27010_001MA	B27010_001EA	B27010_002E	B27010_002M	B27
2	Geography	Geographic Area	Estimate!!Total:	Margin of Error!!	Annotation of Ma	Annotation of Es	Estimate!!Total:!!	Margin of Error!!	Ann
3	1500000US5177	Block Group 1, C	1426	485	null	null	418	237	null
4	1500000US5177	Block Group 2, C	1411	564	null	null	257	258	null
5	1500000US5177	Block Group 3, C	918	336	null	null	214	162	null
6	1500000US5177	Block Group 1, C	684	265	null	null	22	42	null
7	1500000US5177	Block Group 2, C	1549	511	null	null	153	133	null
8	1500000US5177	Block Group 3, C	2111	563	null	null	609	304	null
9	1500000US5177	Block Group 1, C	1956	654	null	null	636	382	null
10	1500000US5177	Block Group 2, C	2435	630	null	null	661	285	null
11	1500000US5177	Block Group 3, C	1767	826	null	null	622	426	null

4b Firstly, we'll remove the columns we won't use. We need the GEO_ID column to match the spreadsheet data with the shapefile, and we want to see the total insured population. Rather than add up all the possible insurance options included, we'll simply add all the "uninsured" columns together.

Keep the "GEO_ID" and "Total"columns as well as all columns with "no health insurance" (eg under 19 no insurance, 20-31 no insurance, and so on). Remove all the other columns, including "NAME", "ANNOTATION", and all "Margin of Error" columns (ctrl+click to select multiple, then **right click, "Delete Column"**). Note: to see the full column name, pull the column's right edge from the upper, lettered gray header section.

Note: this is a long and tedious process. I like to look for the Estimate columns that contain only the information I need first (eg with "no health insurance" in them, which you can also find with the search function), and then re-name the top row. Then I go back and delete all the columns without a re-named top row.

You should end up with 4 "no health insurance" rows, one total row, and one GEOID row.

	A	B	C	D	E	F	G	H
1	GEO_ID	NAME	TOTAL	UNINSURED1	UNINSURED2	UNINSURED3	UNINSURED4	
2	Geography	Geographic Area	Estimate!!Total:	Estimate!!Total:!!	Estimate!!Total:!!	Estimate!!Total:!!	Estimate!!Total:!!65 years	
3	1500000US5177	Block Group 1, C	1426	0	81	57	0	
4	1500000US5177	Block Group 2, C	1411	0	84	0	0	
5	1500000US5177	Block Group 3, C	918	18	47	61	0	
6	1500000US5177	Block Group 1, C	684	0	76	54	0	
7	1500000US5177	Block Group 2, C	1549	0	68	0	0	
8	1500000US5177	Block Group 3, C	2111	0	75	112	0	
9	1500000US5177	Block Group 1, C	1956	0	81	125	0	
10	1500000US5177	Block Group 2, C	2435	0	115	76	0	
11	1500000US5177	Block Group 3, C	1767	257	32	47	0	
12	1500000US5177	Block Group 1, C	2499	32	40	406	48	
13	1500000US5177	Block Group 2, C	660	0	9	59	0	

4c Next, rename the columns. Put the new names in the **top row**, replacing the esoteric census abbreviations (P001002, etc.) with shortened versions of the descriptions in row 2. Note: **avoid using special characters or spaces in your column names.** For instance, I'll change "Estimate!!Total:" to "TOTAL".

Once you've got your new names in row 1, **delete row 2** (**select the row, right click, "Delete Row"**). QGIS will automatically read the first row as "titles", and everything starting in the second row as data, so you don't want to confuse it by including two header rows.

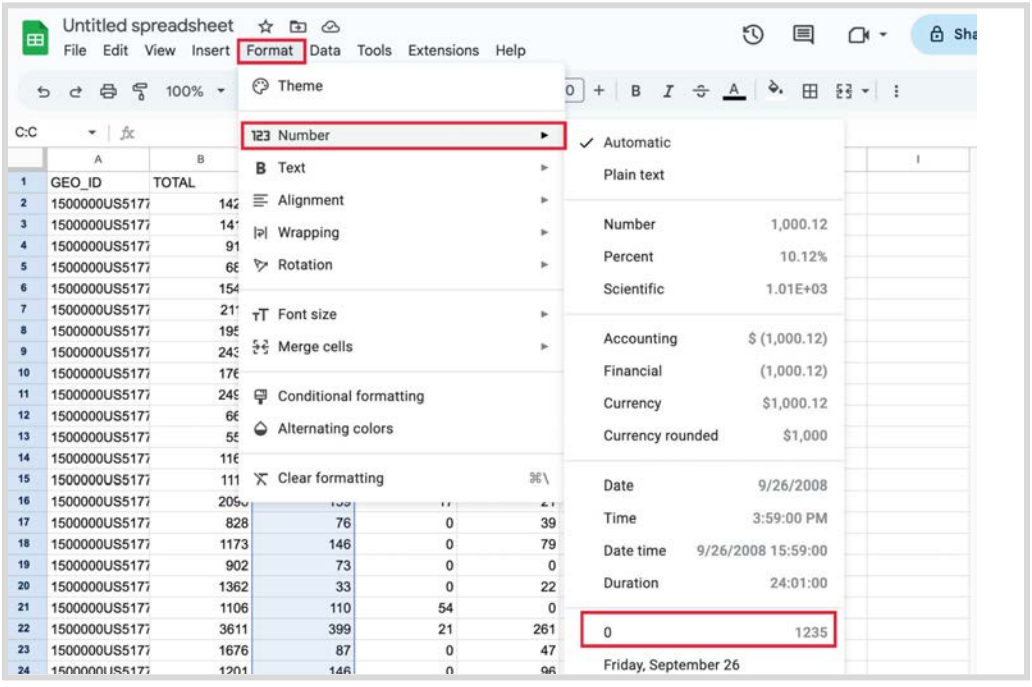
	A	B	C	D	E	F	G
1	GEO_ID	TOTAL	UNINSURED1	UNINSURED2	UNINSURED3	UNINSURED4	
2	1500000US5177	1426	0	81	57	0	
3	1500000US5177	1411	0	84	0	0	
4	1500000US5177	918	18	47	61	0	
5	1500000US5177	684	0	76	54	0	

4d Now we can combine the "uninsured" columns. Add a column after "TOTAL" called "UNINSURED". In the first cell, type =SUM( ) and select the four cells of the uninsured columns, eg. **=SUM(D2:G2)**

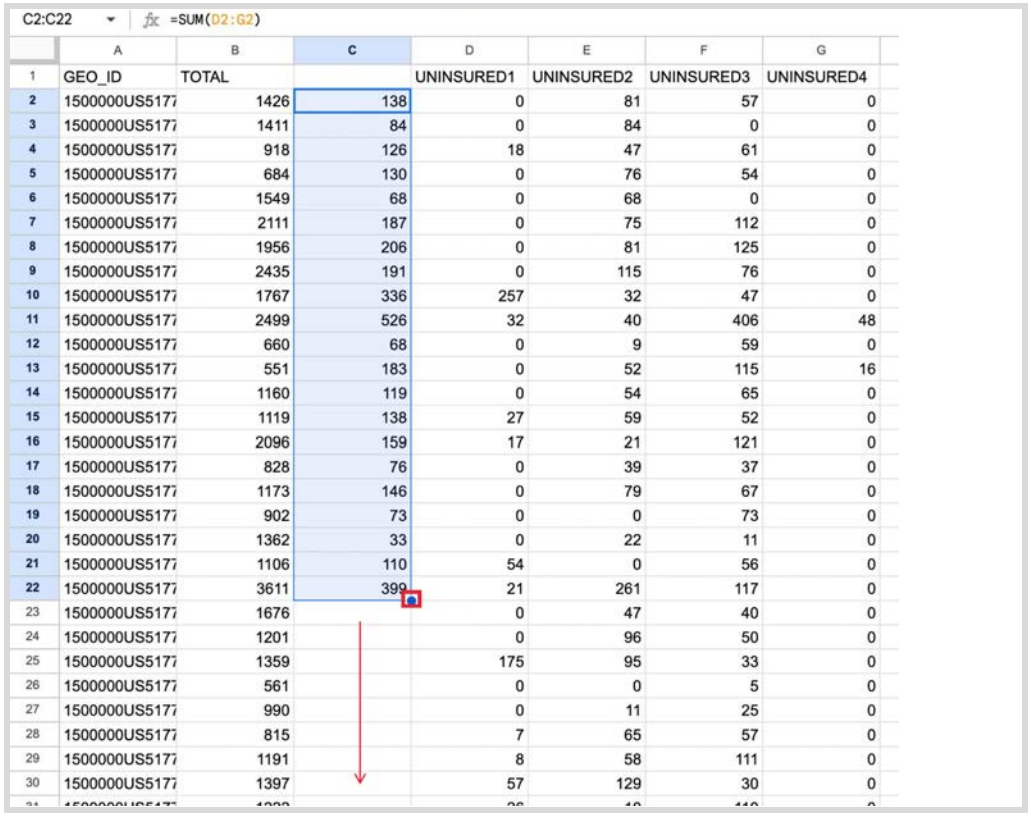
C2 =SUM(D2:G2)

	A	B	C	D	E	F	G
1	GEO_ID	TOTAL	0 ×	UNINSURED1	UNINSURED2	UNINSURED3	UNINSURED4
2	1500000US5177	1426	=SUM(D2:G2)	0	81	57	0
3	1500000US5177	1411		0	84	0	0
4	1500000US5177	918		18	47	61	0
5	1500000US5177	684		0	76	54	0
6	1500000US5177	1549		0	68	0	0
7	1500000US5177	2111		0	75	112	0

Note: if the formula doesn't work or returns 0, it might be because the spreadsheet interprets the data as text instead of a number. You can select the column, then choose **Format > Number > "0"** to format it as a number.

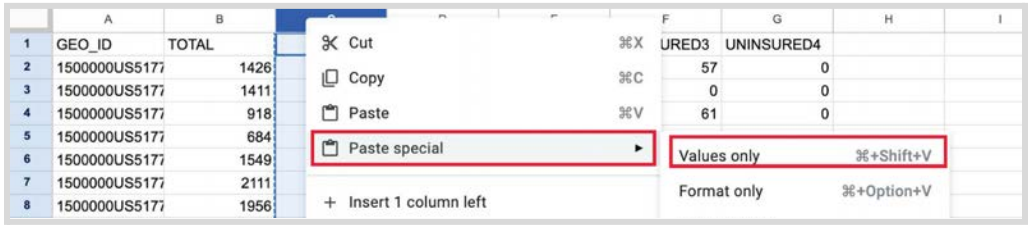


Click and drag the square in the bottom right of the cell all the way down the column to fill in the other cells. You should see the formula automatically update per row, eg D3:G3, D4:G4, and so on.



	A	B	C	D	E	F	G
1	GEO_ID	TOTAL		UNINSURED1	UNINSURED2	UNINSURED3	UNINSURED4
2	1500000US5177	1426	138	0	81	57	0
3	1500000US5177	1411	84	0	84	0	0
4	1500000US5177	918	126	18	47	61	0
5	1500000US5177	684	130	0	76	54	0
6	1500000US5177	1549	68	0	68	0	0
7	1500000US5177	2111	187	0	75	112	0
8	1500000US5177	1956	206	0	81	125	0
9	1500000US5177	2435	191	0	115	76	0
10	1500000US5177	1767	336	257	32	47	0
11	1500000US5177	2499	526	32	40	406	48
12	1500000US5177	660	68	0	9	59	0
13	1500000US5177	551	183	0	52	115	16
14	1500000US5177	1160	119	0	54	65	0
15	1500000US5177	1119	138	27	59	52	0
16	1500000US5177	2096	159	17	21	121	0
17	1500000US5177	828	76	0	39	37	0
18	1500000US5177	1173	146	0	79	67	0
19	1500000US5177	902	73	0	0	73	0
20	1500000US5177	1362	33	0	22	11	0
21	1500000US5177	1106	110	54	0	56	0
22	1500000US5177	3611	399	21	261	117	0
23	1500000US5177	1676		0	47	40	0
24	1500000US5177	1201		0	96	50	0
25	1500000US5177	1359		175	95	33	0
26	1500000US5177	561		0	0	5	0
27	1500000US5177	990		0	11	25	0
28	1500000US5177	815		7	65	57	0
29	1500000US5177	1191		8	58	111	0
30	1500000US5177	1397		57	129	30	0

One last step: The formula depends on the other columns, and we want to simplify this spreadsheet to just a few columns. To delete the “no insurance” columns without losing the new “uninsured total” column, you first need to copy / paste the “uninsured total” column. Select it, copy, and then right click **“Paste Special” > “Values Only**”.

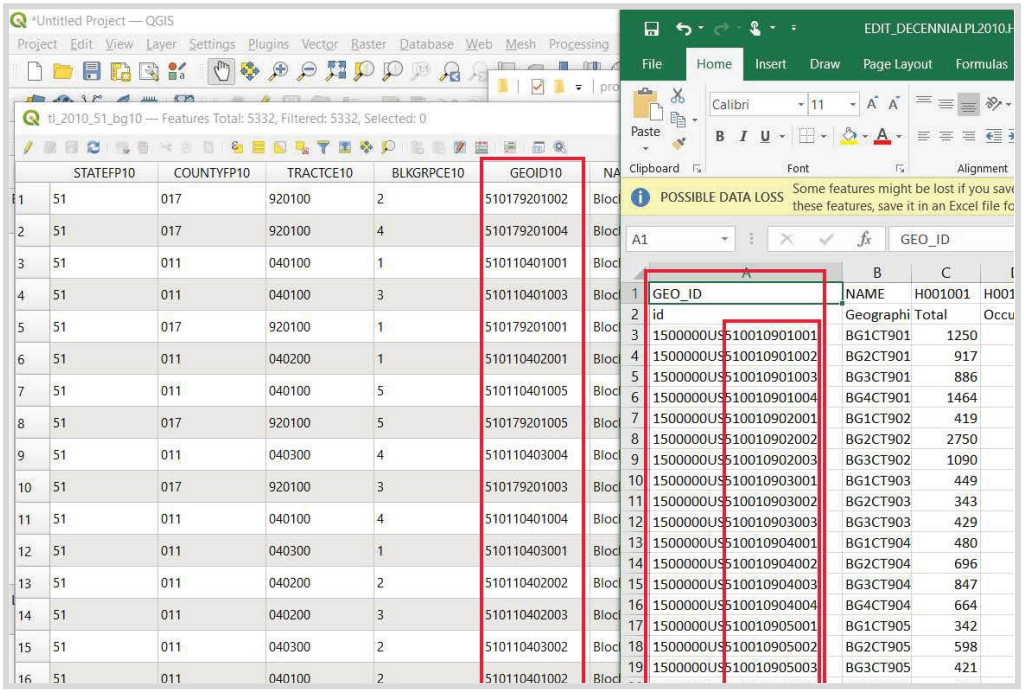


	A	B	F	G	H
1	GEO_ID	TOTAL	URED3	UNINSURED4	
2	1500000US5177	1426	57	0	
3	1500000US5177	1411	0	0	
4	1500000US5177	918	61	0	
5	1500000US5177	684			
6	1500000US5177	1549			
7	1500000US5177	2111			
8	1500000US5177	1956			

Now you can delete the four individual “no insurance” columns and keep only the **GEOID, TOTAL, and UNINSURED** total columns.

	A	B	C	D
1	GEO_ID	TOTAL	UNINSURED	
2	1500000US5177	1426	138	
3	1500000US5177	1411	84	
4	1500000US5177	918	126	
5	1500000US5177	684	130	

4e Nearly there! Lastly, we need to **convert the GEOID column**. Looking at the attribute table of the shapefiles in QGIS, we can see that the GEOID numbers are only the last 12 digits of the GEOIDs in the census tables.

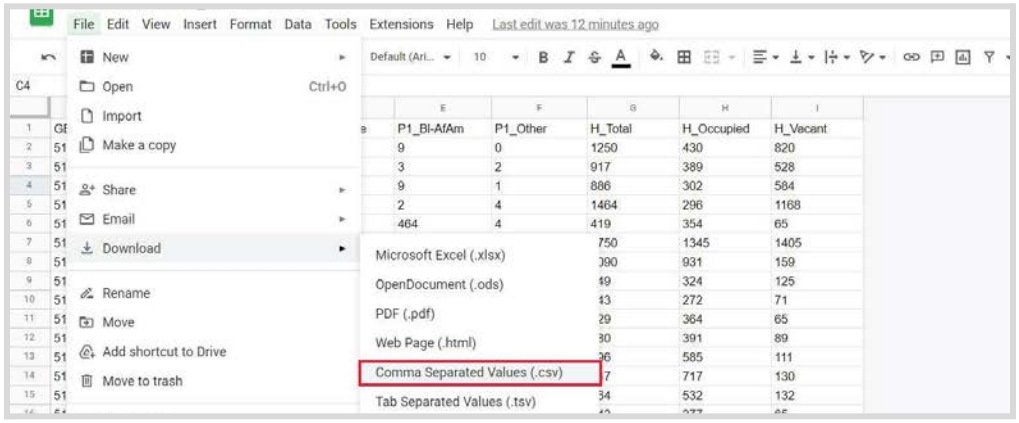


	STATEFP10	COUNTYFP10	TRACTCE10	BLKGRPCE10	GEOID10	NA
1	51	017	920100	2	510179201002	Bloc
2	51	017	920100	4	510179201004	Bloc
3	51	011	040100	1	510110401001	Bloc
4	51	011	040100	3	510110401003	Bloc
5	51	017	920100	1	510179201001	Bloc
6	51	011	040200	1	510110402001	Bloc
7	51	011	040100	5	510110401005	Bloc
8	51	017	920100	5	510179201005	Bloc
9	51	011	040300	4	510110403004	Bloc
10	51	017	920100	3	510179201003	Bloc
11	51	011	040100	4	510110401004	Bloc
12	51	011	040300	1	510110403001	Bloc
13	51	011	040200	2	510110402002	Bloc
14	51	011	040200	3	510110402003	Bloc
15	51	011	040300	2	510110403002	Bloc
16	51	011	040100	2	510110401002	Bloc

	A	B	C
1	GEO_ID	NAME	H001001
2	id	Geographi	Total
3	1500000US510010901001	BG1CT901	1250
4	1500000US510010901002	BG2CT901	917
5	1500000US510010901003	BG3CT901	886
6	1500000US510010901004	BG4CT901	1464
7	1500000US510010902001	BG1CT902	419
8	1500000US510010902002	BG2CT902	2750
9	1500000US510010902003	BG3CT902	1090
10	1500000US510010903001	BG1CT903	449
11	1500000US510010903002	BG2CT903	343
12	1500000US510010903003	BG3CT903	429
13	1500000US510010904001	BG1CT904	480
14	1500000US510010904002	BG2CT904	696
15	1500000US510010904003	BG3CT904	847
16	1500000US510010904004	BG4CT904	664
17	1500000US510010905001	BG1CT905	342
18	1500000US510010905002	BG2CT905	598
19	1500000US510010905003	BG3CT905	421

To do this, use the “Search and Replace” function (usually ctrl+F or find it in the toolbar). Search for the standard leading numbers you want to remove (in this case “1500000US”) – it’s best to copy / paste this from one of the cells – and then leave the “replace” box empty to delete these characters without removing the rest. Click “Find and replace all” and watch as your GEOIDs correct.

4g Lastly, **download** the 2021 files as **.csv** files.



	P1_BI-AfAm	P1_Other	H_Total	H_Occupied	H_Vacant
	9	0	1250	430	820
	3	2	917	389	528
	9	1	886	302	584
	2	4	1464	296	1168
	464	4	419	354	65
			750	1345	1405
			090	931	159
			49	324	125
			13	272	71
			29	364	65
			80	391	89
			96	585	111
			7	717	130
			34	532	132

4f Open a new tab or spreadsheet and **Repeat** this step for the 2017 csv.

Step 5: Import the TIGER files to QGIS and join the tables.

Referring to the earlier tutorials about importing and joining census data, import the .csv files you edited to your QGIS file and join the **2017 csv to the 2010 Shapefile and the 2021 csv to the 2020 Shapefile** using the GEOID columns

As always, check that the data joined correctly in each **shapefile**'s attribute table.

Step 5: Style data and create Print Layout.

Style the new census data to show the **percentage of insured population**. Like last time, you'll need to use the Expression Calculator to show the percentage. Keep in mind that you created a column for total "uninsured" not total "insured", so you'll need to subtract the uninsured from the total to get the number insured, and then divide that by the total to get the percentage: (Total - Uninsured) / Total. You may also need to use the "to_int" formula around the entire expression: to_int ((Total - Uninsured) / Total).

Make sure to use "pretty breaks" and set your classes to "10" to get 10 even percentage point intervals. Also be sure to remove the outlines from your block groups to make data patterns more visible.

In a new Print Layout, Compare the 2017 and 2021 data. Include **at least one previous data layer** for comparison, for example the median income or racial data. The previous layers used Census Tracts and not Block Groups, but you can still identify general patterns in the city's neighborhoods.

As always, include a title, legend, scale, north arrow, map labels, your name, and sources.