GIS for Designers
VT | A+D | SU22

TUTORIAL 3 | IMPORTING CENSUS DATA

Zoom: Wednesday 06.01, 6:30-7:30pm
https://virginiatech.zoom.us/j/2981092726

Goals

- Download and parse census data for your city's state.
- Add census data to the state shapefile.
- Examine opportunities and limitations of census data.
- Produce choropleth map of chosen census feature.

Introduction

In this tutorial, you'll be looking at census data for the state where your city from Tutorial 1-2 is located.

Any GIS data has two components: the data itself, usually in spreadsheet form, and the geographical shape that the data links to. The shapefiles are defined as points connected by lines (eg, the outline of a census district or building); each polygon within the shapefile has an associated ID number. The data spreadsheet will have a corresponding ID column that links data (aka "attributes") with shape.

To use the Census data, we need to download both the **shapefile** (also called TIGER), and the **spreadsheet data.** The Census has several levels of data detail, from the state to county, down to census tracts, census block groups, and as small as census blocks (generally corresponding to a city block). For privacy reasons, they don't publish household-level data and they randomly rearrange a percent of all data. Not all types of data are available at all levels, or for all years, so it's good to check the Census Finder data before you download the shapefile.

Note 1: your census data and shapefile year must match. The boundaries of geographical areas (census tracts, blocks, etc) CHANGE every 10-year census (so, your 2020 census data won't accurately match your 2010 shapefile). We'll be using both the 2010 and 2020 shapefiles in this tutorial to compare changes across the decades.

Note 2: the US government does a major, Decennial census every 10 years. The American Community Survey (ASC) supplements the major census by collecting annual data through a smaller, randomized poll that is then aggregated and used to predict the un-polled population changes. We'll use Decennial data in this tutorial, and talk more about the ACS in Tutorial 4.

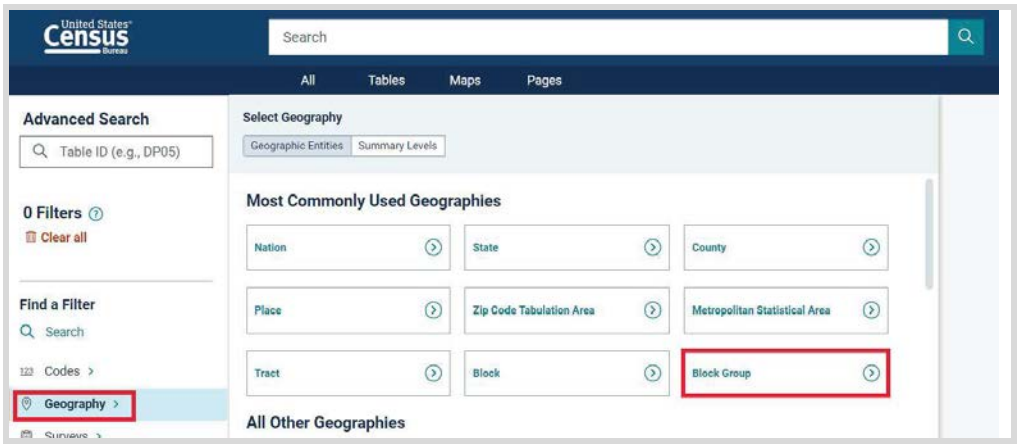
Step 1: Download US Census shapefiles and data.

1b First, we'll download the **TIGER Block Group Shapefile for 2020 and 2010 in the state of your city from Tutorials 1 and 2**: https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

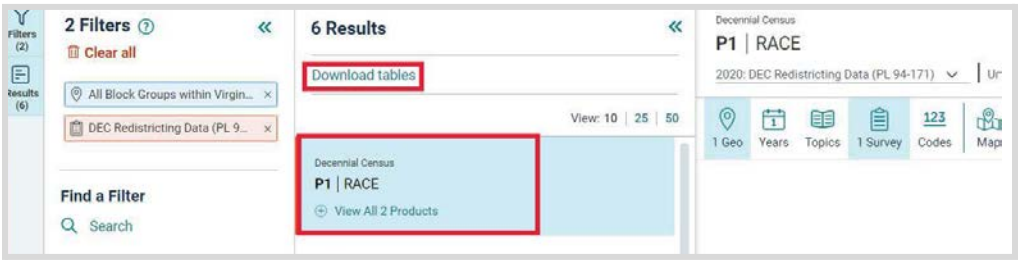
> **Note:** Since I used Blacksburg in the last tutorials, I'll be downloading Virginia data. If you used Pittsburg, for instance, follow this tutorial for the state Pennsylvania. If you used Charlotte, download North Carolina. Etc.
>
> Since we want to compare census data between 2010 and 2020, we need to download both the 2020 and 2010 shapefiles. Select **"Web interface"** at the bottom. On the next page, select **"2020" and "Block Group"**. Finally, select **your state**, and download. Note: download the entire state.

Then, go back a page, change the year to **"2010", "Block Group"**, then **your state**. Download.

1b Second, go to **Census Finder**: https://data.census.gov/cedsci/table. You can filter the available datasets with "Advanced Search", first by **"Geography" ("Block Groups" > your state > "All")**, and then by **"Survey" ("Decennial" > "Redistricting")**. We're looking for the data available from the most recent ten-year (decennial) survey. Then, click "Search".



We're going to be looking at **"P1 | Race" data table**. Click on the P1 result, and then **"Download tables"** at the top of the list. **Check "All"** to download 2020 and 2010, and **"Download"**. You'll see yet another popup, which, once loaded, you can again "Download".

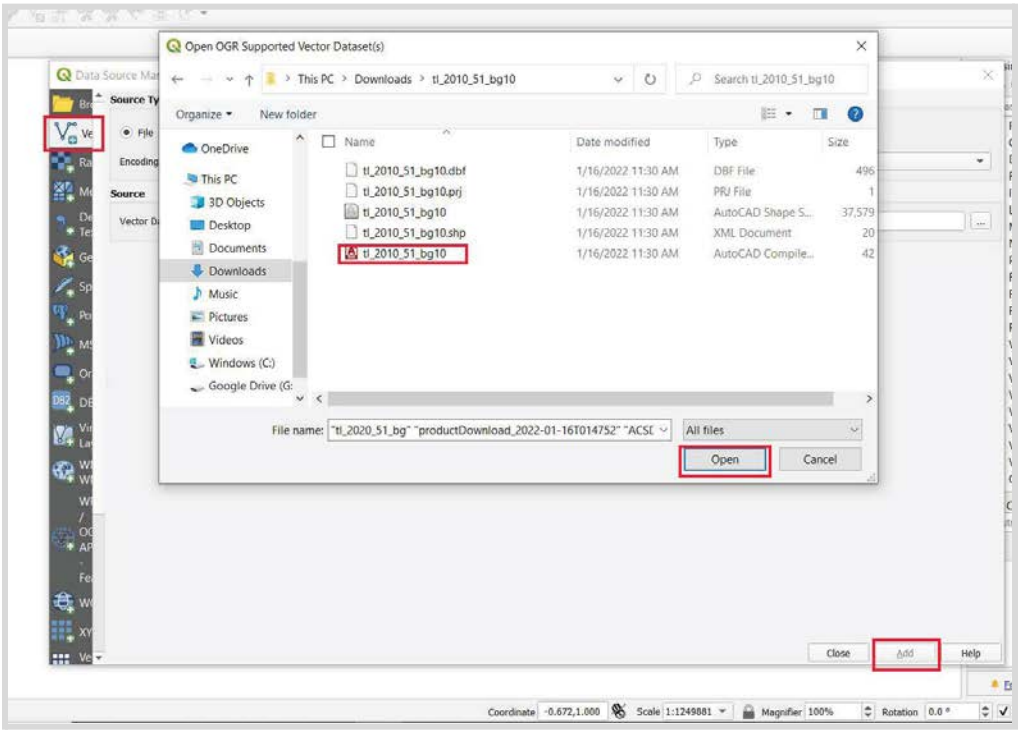


Step 2: Import the TIGER Shapefiles to QGIS.

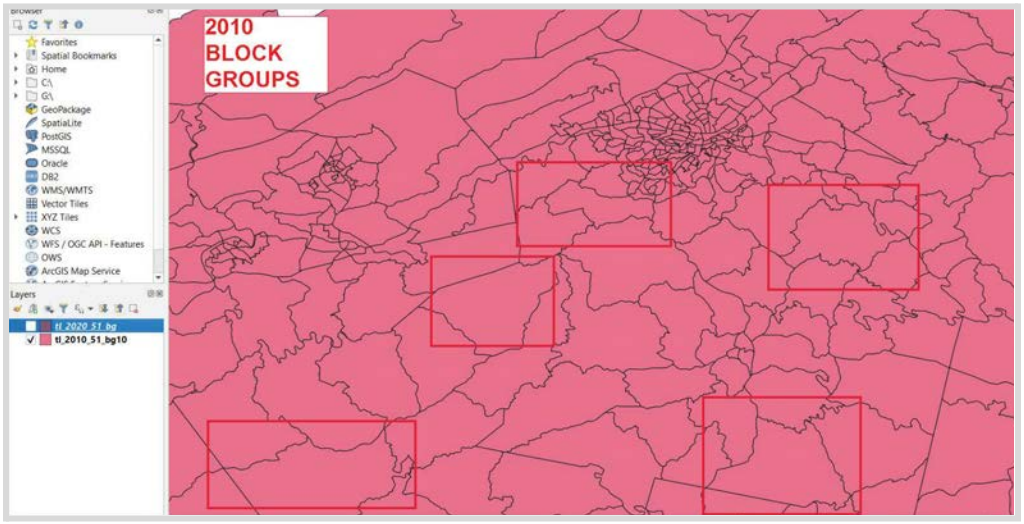
2a Open a new QGIS file.

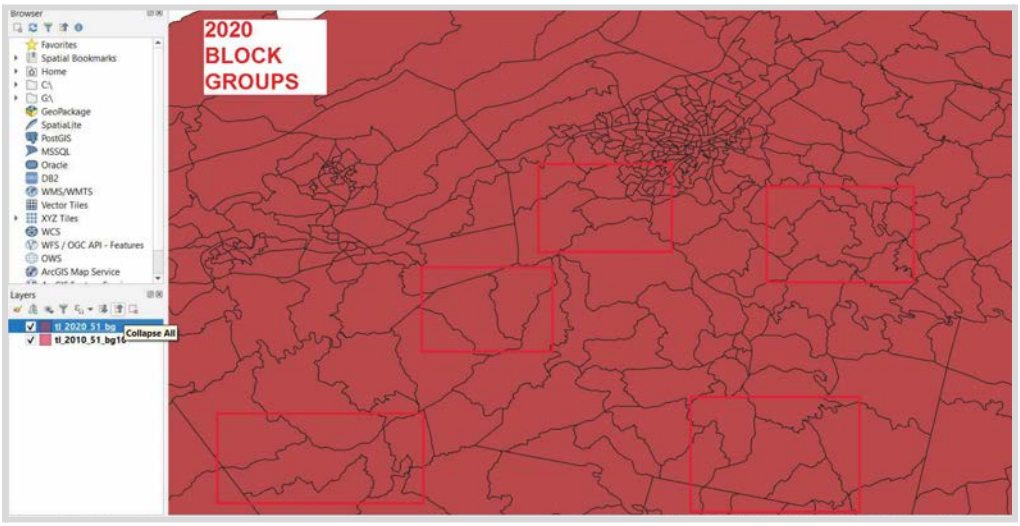
2b Set your CRS as in Tutorial 1 step 2b. **Note:** since the TIGER shapefiles use NAD83 projection, they should show up well in your NAD83/UTM projection.

2c **Layer > Add Layer > Add Vector Laye**r; select the 2020 shapefile; **add**; then select the 2010 shapefile; **add**.



Notice the differences in the Block Group outlines between 2010 and 2020.

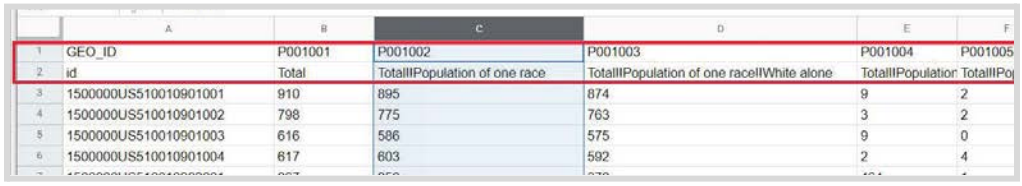




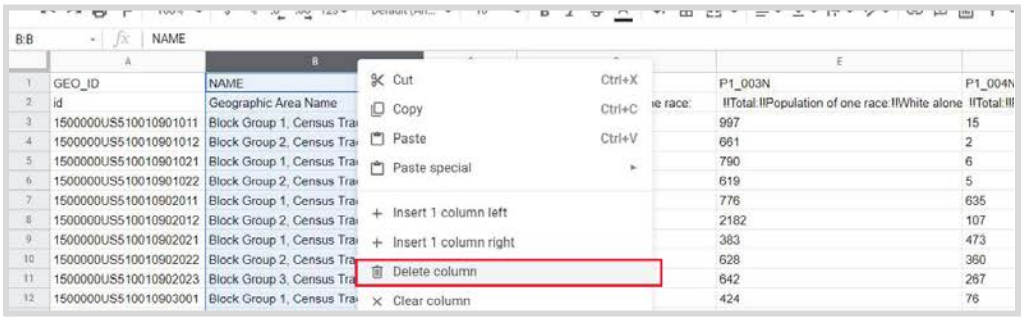
Step 3: Clean up Census data spreadsheets.

3a Before we import the data, we're going to clean it up. Go to Google Sheets (type **"sheet.new**" into your browser) and upload the Census csv 2020 P1 file (File>Import).. **Make sure to uncheck "Convert Numbers"**. 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.

1) 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 compare White and NonWhite populations. Keep the "GEO_ID", "Total", and "Total Population of One Race White" columns, and remove the others (**right click, "Delete Column"**). We do not need the "NAME" column. Note: to see the full column name, pull the column's right edge from the upper, lettered gray header section.

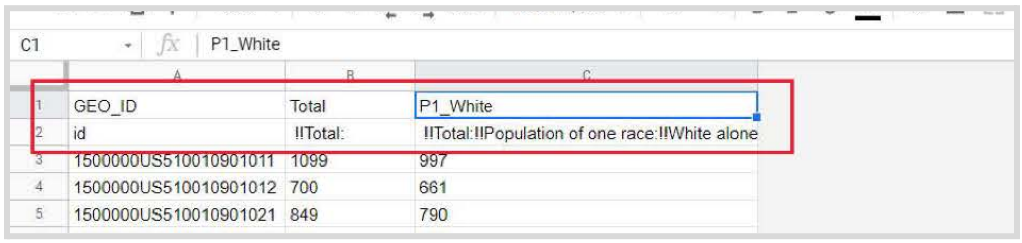


	A	B	C	D	E	F
1	GEO_ID	P001001	P001002	P001003	P001004	P001005
2	id	Total	Total!!Population of one race	Total!!Population of one race!!White alone	Total!!Population	Total!!Po
3	1500000US510010901001	910	895	874	9	2
4	1500000US510010901002	798	775	763	3	2
5	1500000US510010901003	616	586	575	9	0
6	1500000US510010901004	617	603	592	2	4



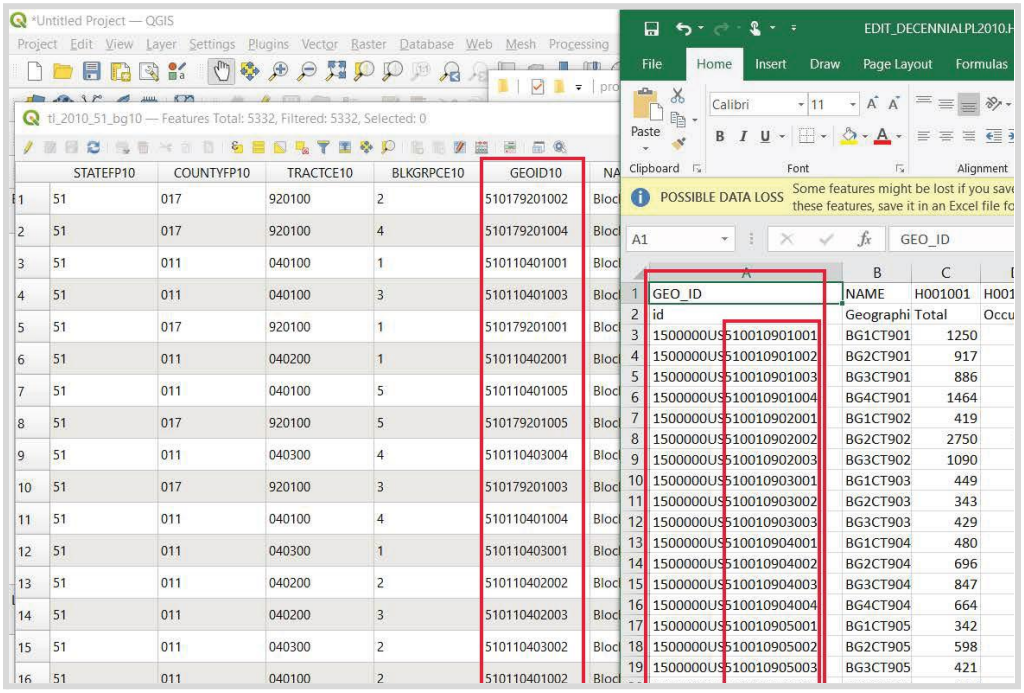
	A	B	E	F
1	GEO_ID	NAME	P1_003N	P1_004N
2	id	Geographic Area Name	!!Total:!!Population of one race:!!White alone	!!Total:!!
3	1500000US510010901011	Block Group 1, Census Tra	997	15
4	1500000US510010901012	Block Group 2, Census Tra	661	2
5	1500000US510010901021	Block Group 1, Census Tra	790	6
6	1500000US510010901022	Block Group 2, Census Tra	619	5
7	1500000US510010902011	Block Group 1, Census Tra	776	635
8	1500000US510010902012	Block Group 2, Census Tra	2182	107
9	1500000US510010902021	Block Group 1, Census Tra	383	473
10	1500000US510010902022	Block Group 2, Census Tra	628	360
11	1500000US510010902023	Block Group 3, Census Tra	642	267
12	1500000US510010903001	Block Group 1, Census Tra	424	76

2) Next, we want to 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 "Total!! Population of one race!!White alone" to "P1_White". Once you've got your new names in row 1, delete row 2 (**select the row, right click, "Delete Row"**).



	A	B	C
1	GEO_ID	Total	P1_White
2	id	!!Total:	!!Total:!!Population of one race:!!White alone
3	1500000US510010901011	1099	997
4	1500000US510010901012	700	661
5	1500000US510010901021	849	790

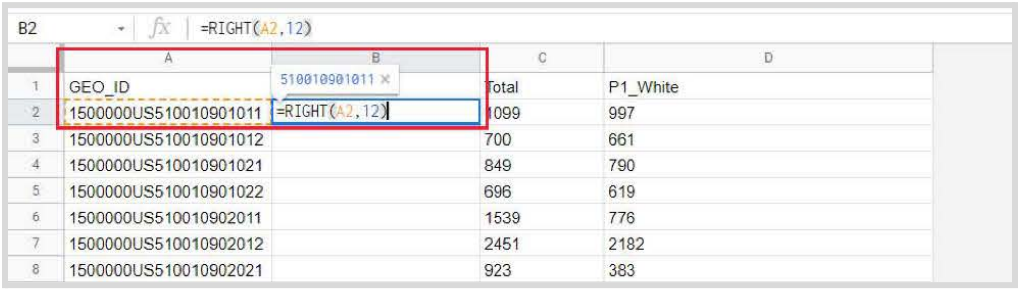
3) Nearly there! Lastly, we need to **convert the GEOID column**. Looking at the attribute table of the shapefiles in QGIS, we can see that the GEOID numbers are only the last 12 digits of the GEOIDs in the census tables.



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

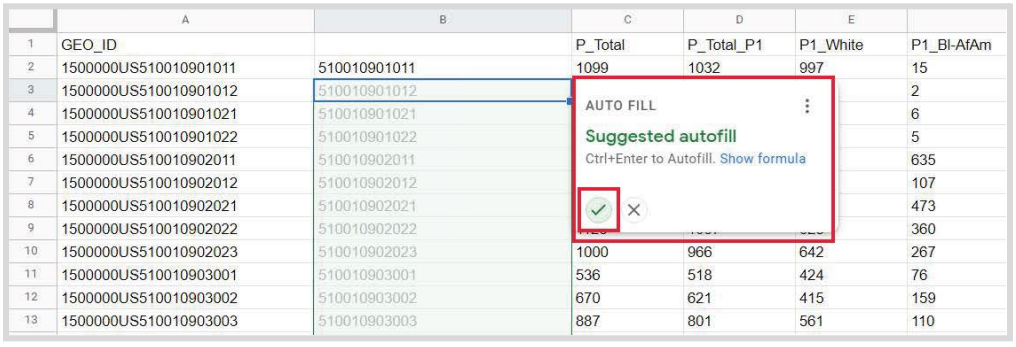
	GEO_ID	NAME	H001001	H001
2	id	Geographi	Total	Occu
3	1500000US510010901001	BG1CT901	1250	
4	1500000US510010901002	BG2CT901	917	
5	1500000US510010901003	BG3CT901	886	
6	1500000US510010901004	BG4CT901	1464	
7	1500000US510010902001	BG1CT902	419	
8	1500000US510010902002	BG2CT902	2750	
9	1500000US510010902003	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 convert these, we can use the formula in Google Sheets " =RIGHT(cell, # char)". Insert a column to the right of the GEOID column (**select GEOID column, right click, "Insert 1 column right**"). In the new cell B2 insert the formula "**=RIGHT(A2,12)**" (without the quotation marks).



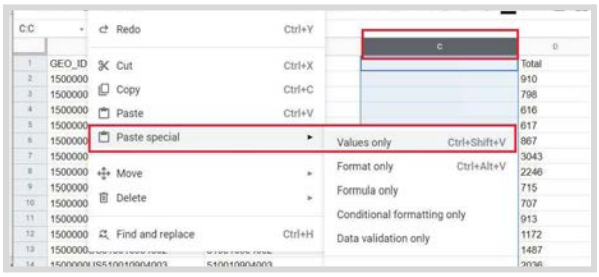
	A	B	C	D
1	GEO_ID		Total	P1_White
2	1500000US510010901011	=RIGHT(A2,12)	1099	997
3	1500000US510010901012		700	661
4	1500000US510010901021		849	790
5	1500000US510010901022		696	619
6	1500000US510010902011		1539	776
7	1500000US510010902012		2451	2182
8	1500000US510010902021		923	383

When Google Sheets helpfully suggests to autofill the column, click the green check mark. (if Google Sheets doesn't suggest autofill, drag the formula down with the small gray square on the lower right of the cell):

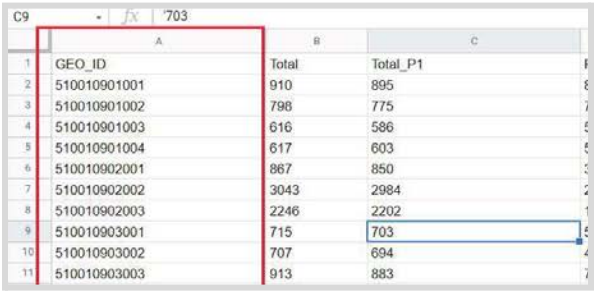


	A	B	C	D	E	
1	GEO_ID		P_Total	P_Total_P1	P1_White	P1_BI-AfAm
2	1500000US510010901011	510010901011	1099	1032	997	15
3	1500000US510010901012	510010901012				2
4	1500000US510010901021	510010901021				6
5	1500000US510010901022	510010901022				5
6	1500000US510010902011	510010902011				635
7	1500000US510010902012	510010902012				107
8	1500000US510010902021	510010902021				473
9	1500000US510010902022	510010902022				360
10	1500000US510010902023	510010902023	1000	966	642	267
11	1500000US510010903001	510010903001	536	518	424	76
12	1500000US510010903002	510010903002	670	621	415	159
13	1500000US510010903003	510010903003	887	801	561	110

Because the new, simplified GEOID column is **formula-dependent** on the original column, you can't delete the original without losing the new one. As a final step, add a new column beside the first two. Select the **shortened GEOID column and "copy"**. Then, in the third (empty) column, **Edit>Paste Special>Values Only**.

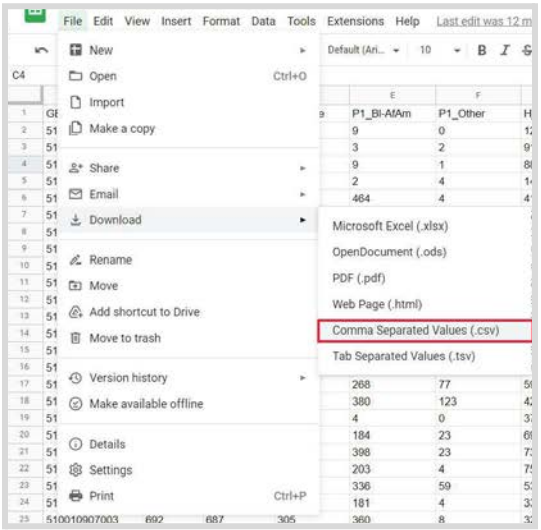


You can now delete the first two columns (the original GEO_ID and the formula-dependent shortened GEO_ID). Rename your new, 12-character GEO_ID column as "GEO_ID".



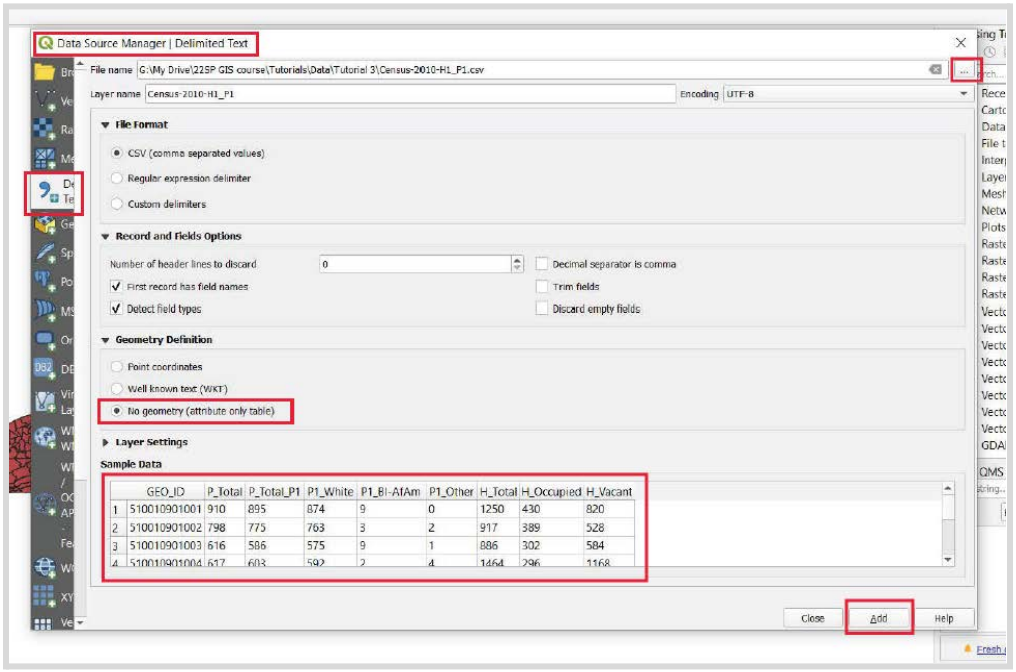
3b **Repeat** Step 3a for the 2010 census data. **Note:** the GEOIDs will be slightly different for 2010, so you need to repeat all three steps (several GEOIDs were added in 2020 for new census block groups).

3c Lastly, **download** both 2010 P1 and 2020 P1 files as **.csv** files**.**

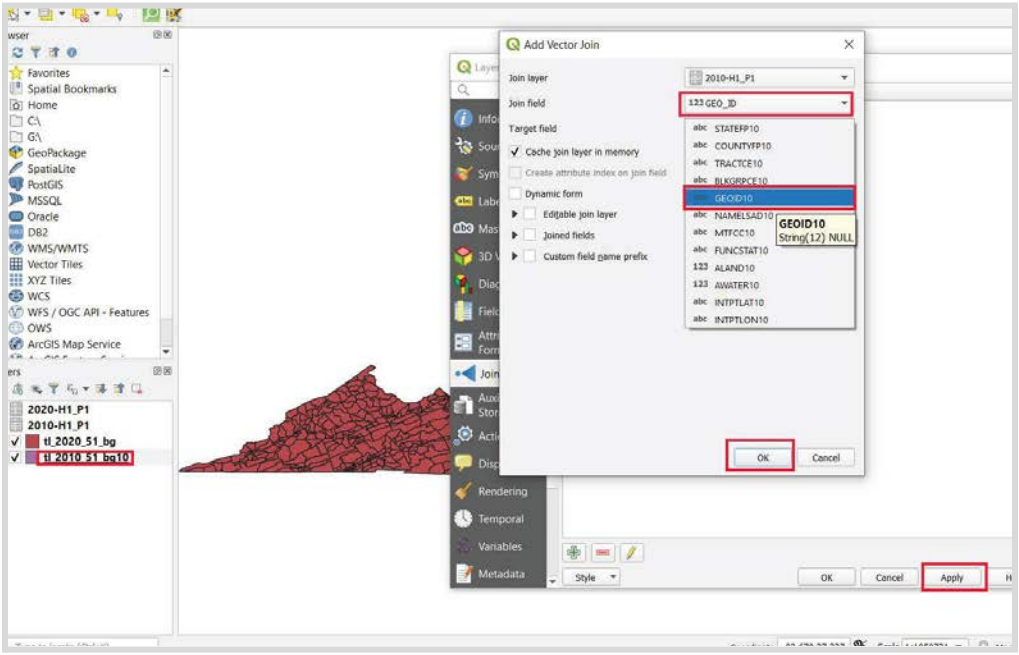


Step 4: Associate Census data with TIGER shapefiles.

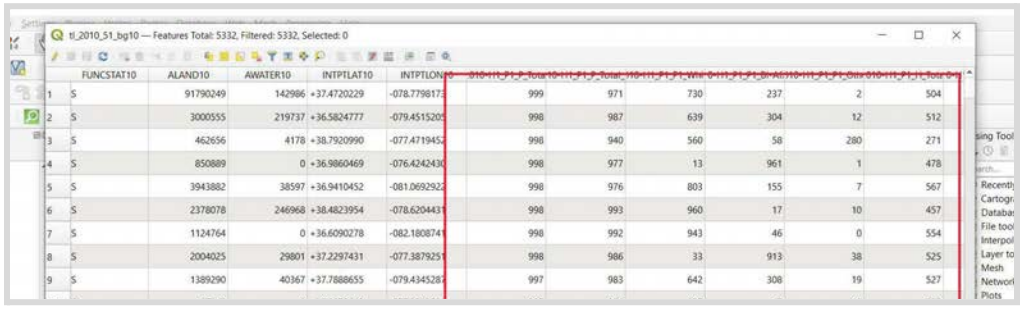
4a In QGIS go to **Layer > Add Layer > Add Delimited Text Layer** and select your **2010 and 2020** csv files. Make sure that under “Geometry Definition” that **“No Geometry”** is selected. Check that the Sample Data display looks ok and click **“Add”**.

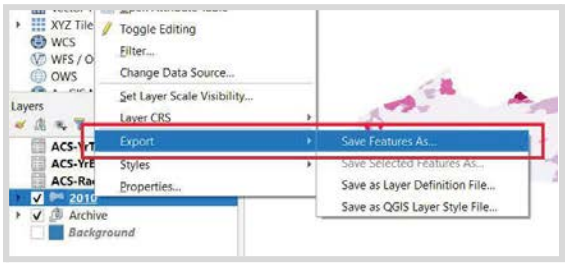


4b Doubleclick the **2010 shapefile** layer for Layer Properties, select the **“Joins”** tab (blue arrow). Merge your **P1 2010 data** layer with the category **“GEOID” for both Join Field and Target Field.**



4c Open the Layer Attribute Table, and make sure that you don't see "NULL" for the fields (this happens when the merge fails). Right click on your shapefile layer and select **Export > Save Features As…**

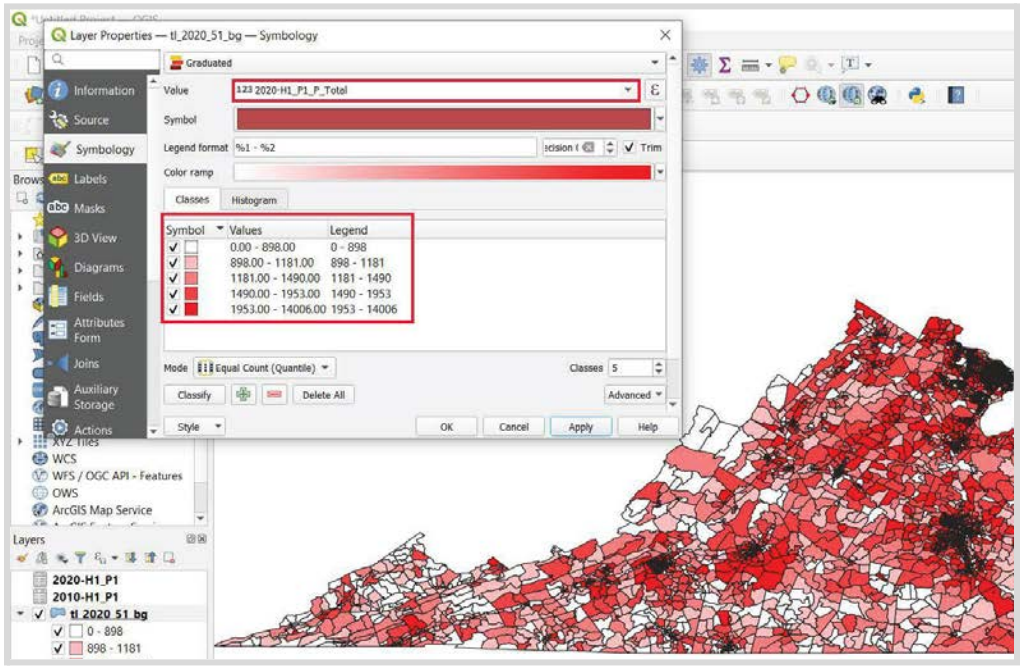




4d **Repeat** Steps 4b-4c for the 2020 shapefile.

Step 5: Color map according to total population.

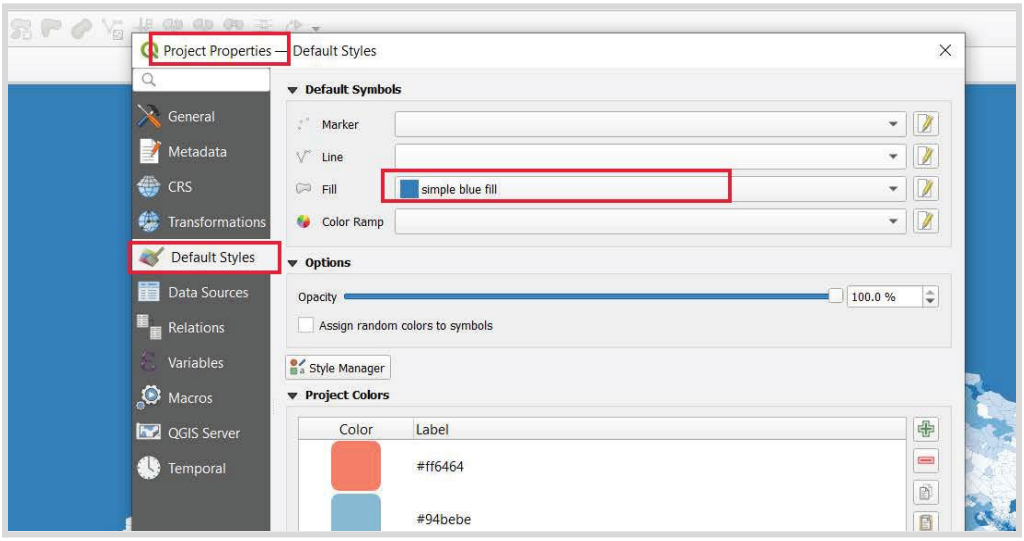
5a First, let's look at the total population data. Color the 2020 shapefile by **"Graduated"**, selecting your total population field. Color the 2010 layer the same way, making sure that your values are the same range (you might need to manually edit them to match).



Now, flip between the two maps. We can see immediately that the population in larger, rural census block groups has decreased. Presumably, people are moving to the cities.

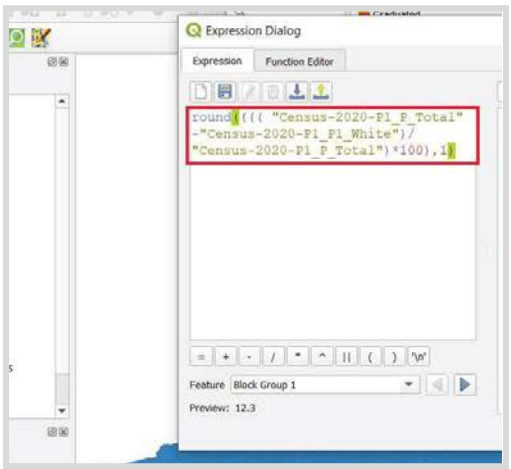
Step 6: Color map according to percentage of non-white population.

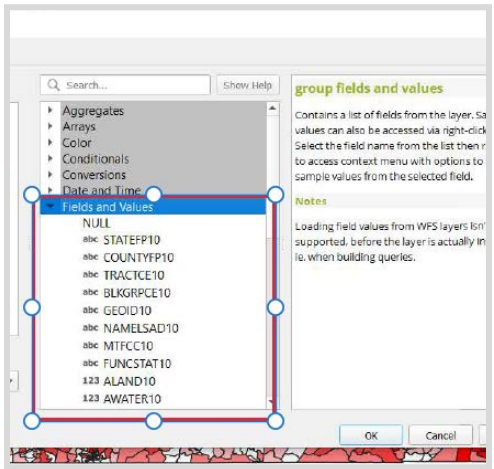
6a To ensure that your polygon layers will automatically NOT include outlines, and save a lot of trouble testing gradients, go to **Project > Properties > Default Styles**, and change the **"Fill"** to a symbol with the border pen type set to **"No Pen"**.



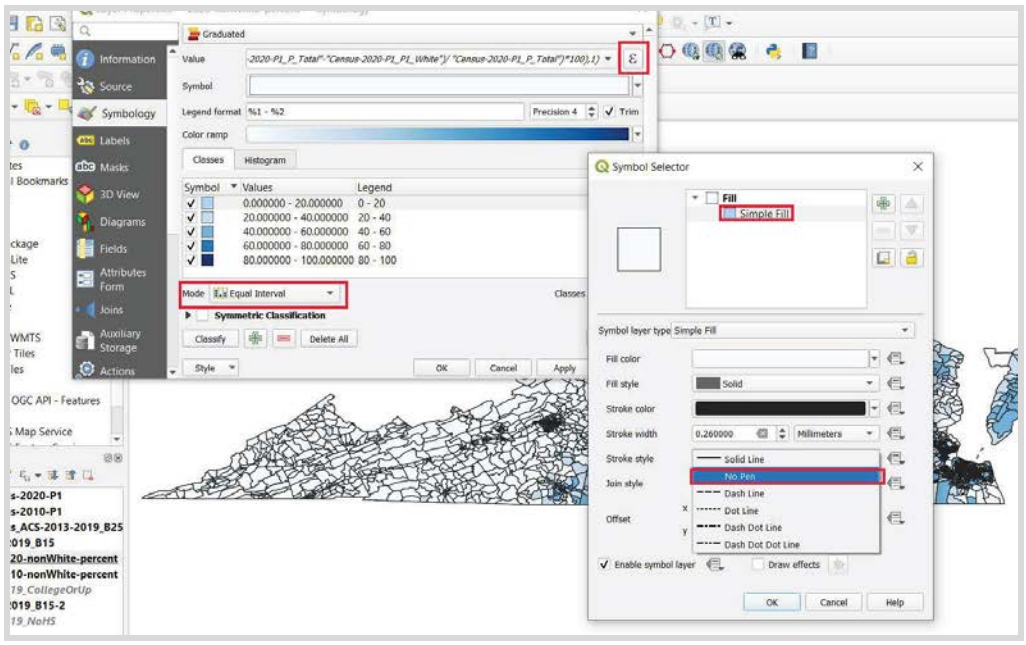
6b To get the percentage of non-white population, we need to use an equation in the Symbology > Graduated "Value" field. In the "Value" field, click **sigma (Σ symbol)**. Here, you can see the possible components of expressions such as AND, OR, NOT operators, IF, THEN conditionals, math operations, conversions like the "to_int" we used in Tutorial 2, and **"Fields and Values"** which list the attributes of your data. You can double click to add these to the expression – very useful in case your fields have complicated names like mine do.

> We have attributes for "Total population" and "Total white population". To get the "Total non-white population", subtract these two. Then, to get the percentage, divide that number by the Total population and multiply by 100. As a last step, round the percentage to an even number with ROUND(). Your final equation should be something like: **ROUND ((("Total"-"TotalWhite") / "Total")*100),1)**. Here, the "1" at the end refers to the number of places to ROUND to.





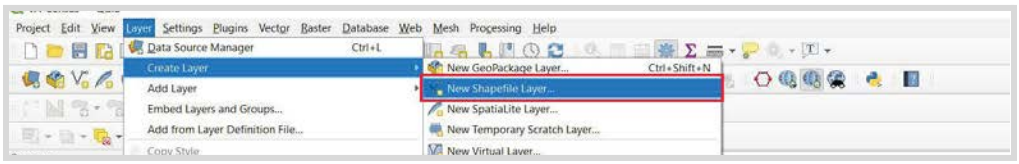
6c Now, make sure that your percentages are nice, even numbers: set the **“Mode” to “Equal Interval”.** If your symbols have an outline, be sure to remove them in order to see the smaller polygons. Click on the “Symbol” square, select “Simple Fill”, and then dropdown “Stroke style” to “No Pen”.

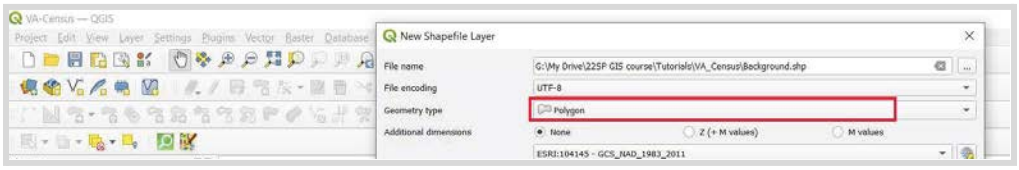


6c **Repeat** this step for both 2010 and 2020 layers.

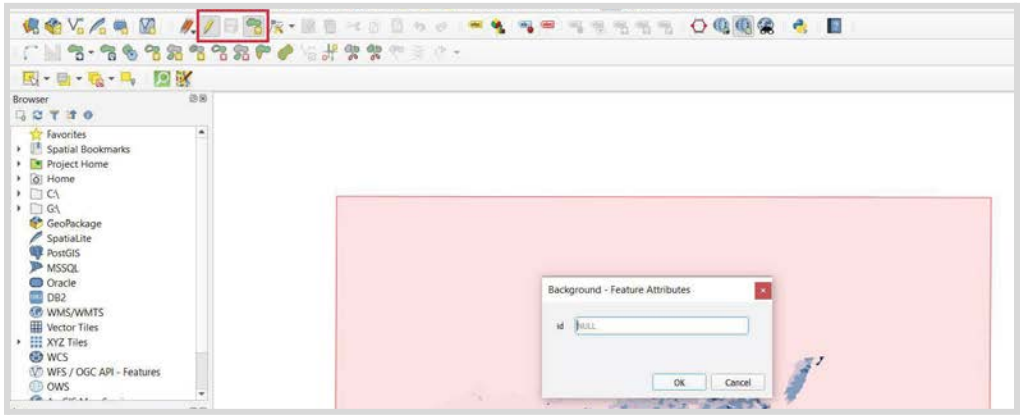
Step 7: Add a background polygon for contrast.

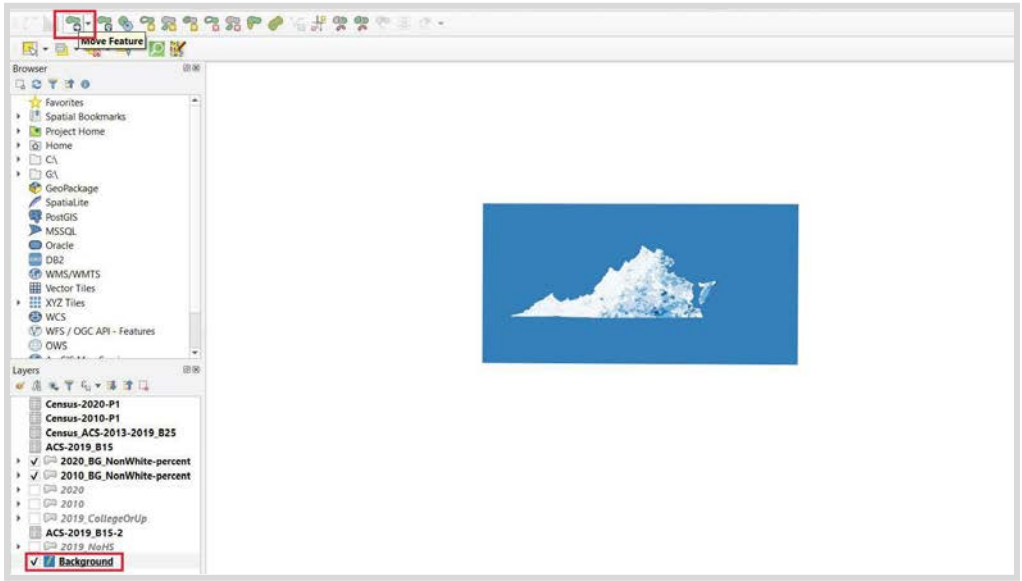
7a Add a new Shapefile layer, and set the Geometry type to “Polygon”. Name your layer something like “Background”





7b Toggle on **“Edit”,** then select the **“Add Polygon Feature”** tool. Click to draw a rectangle behind your map.
NOTE: to finish the polygon, right click. Make sure to toggle “Edit” off when you’re done. If your rectangle isn’t in the right place, you can use the “Move Feature” tool.

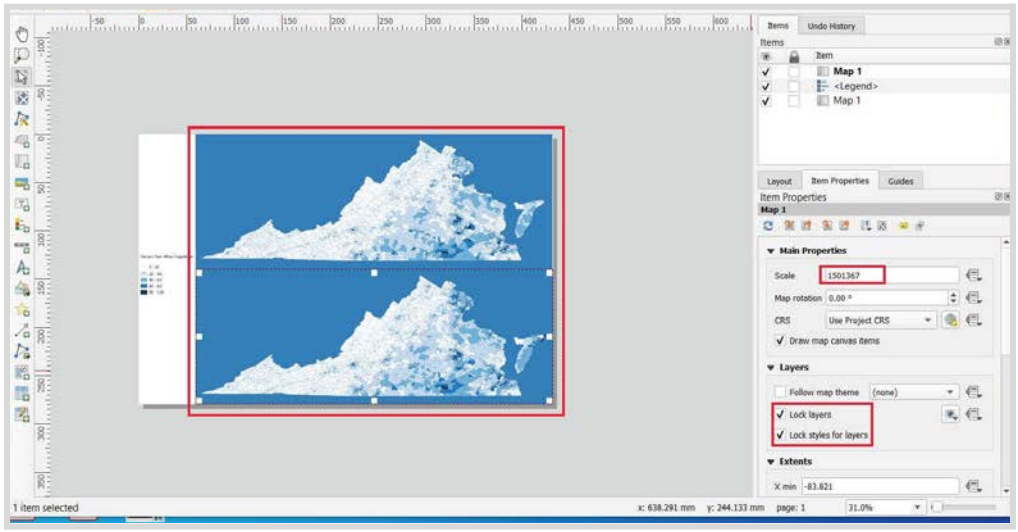




Step 8: Export map as pdf with legend. (or continue to Optional section)

8a Set your page to 11”x17”.

8b Draw your first map. In your map view’s properties, check **“Lock Layers” and “Lock styles for layer”.** Then, in your main map, turn on your other census population layer. In the Layout, draw the second map. Again, “Lock Layers” and styles. Make sure your scale is the same for both maps.



8c Lastly, add the Legend and Title(s).

- **Bonus (STRONGLY RECOMMENDED) -**

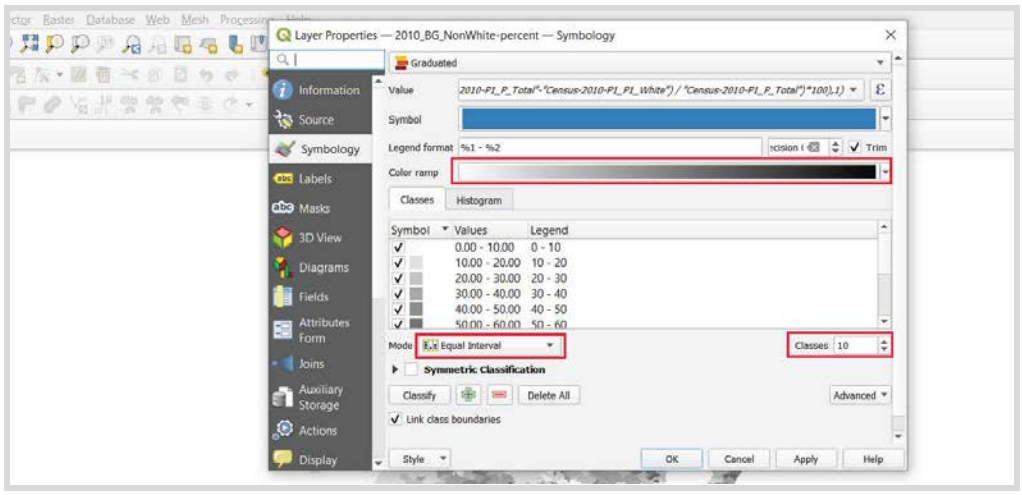
Step 9: Raster analysis.

In the previous steps, you visually analyzed the difference between 2010 and 2020. However, it's difficult to see the difference without flipping between the layers. You could create a GIF to make the change more evident, but you can also perform raster analysis on the two images. QGIS has a tool called the Raster Calculator that allows you to perform math on the pixels in images. In other words, you can export the two years of census data and use simple subtraction to visualize the change between them as a new image.

This is especially useful when comparing the changes between census decades, since their shapefile outlines have changed. Because of this, you can't get an entirely accurate picture just by comparing their spreadsheet tables. The Raster Calculator will compare them pixel by pixel, rather than using attribute tables to compare them block ID by block ID.

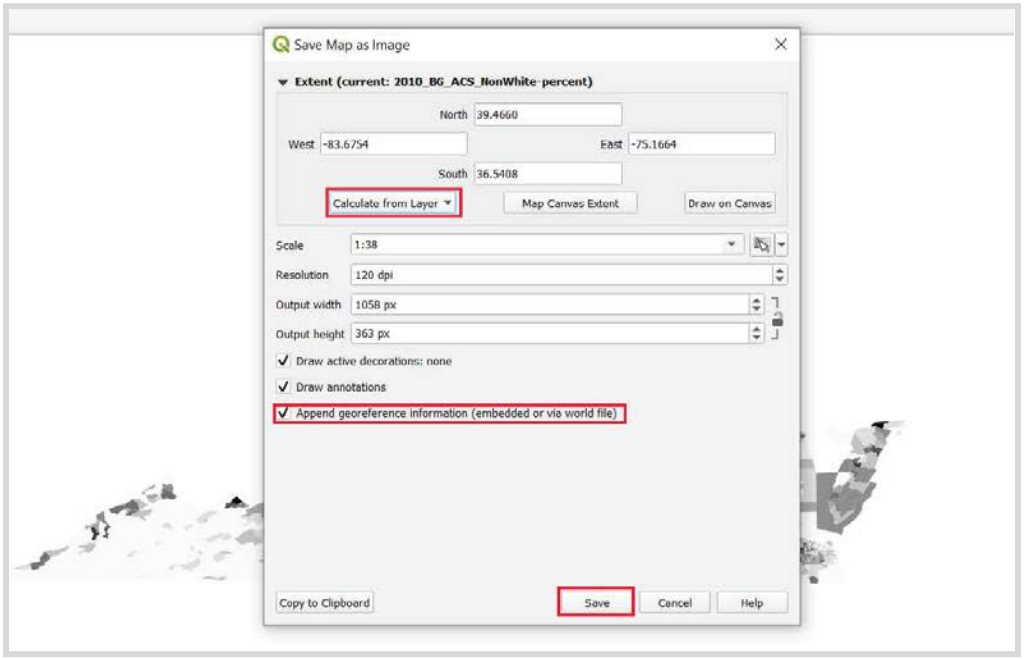
NOTE: to accurately compare the two datasets, first make sure that they're styled identically. In other words, both maps should have the same color scheme and the same percentage ranges (use "Equal Interval" in the graduated styling mode and set the same number of "Classes").

9a First, style your 2010 and 2020 maps with a **black and white gradient** rather than a color gradient. This will make the Raster Calculation easier. Double check that your percentage ranges match between the two layers.

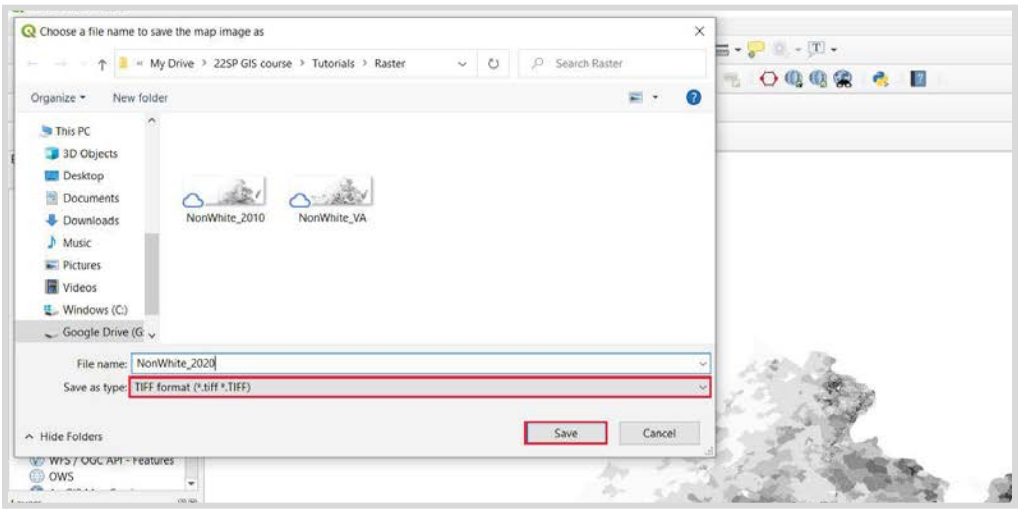


9b Next, export both the 2010 and 2020 Nonwhite percentage maps as images. Turn off all your other layers, and only leave on the 2010 map. Go to **Project > Import/Export > Export Map to Image…**

In the dialogue box, set the **Extent with "Calculate by Layer",** and select your 2010 layer. Leave the other fields as their defaults; make sure that "Append georeference information" is checked.



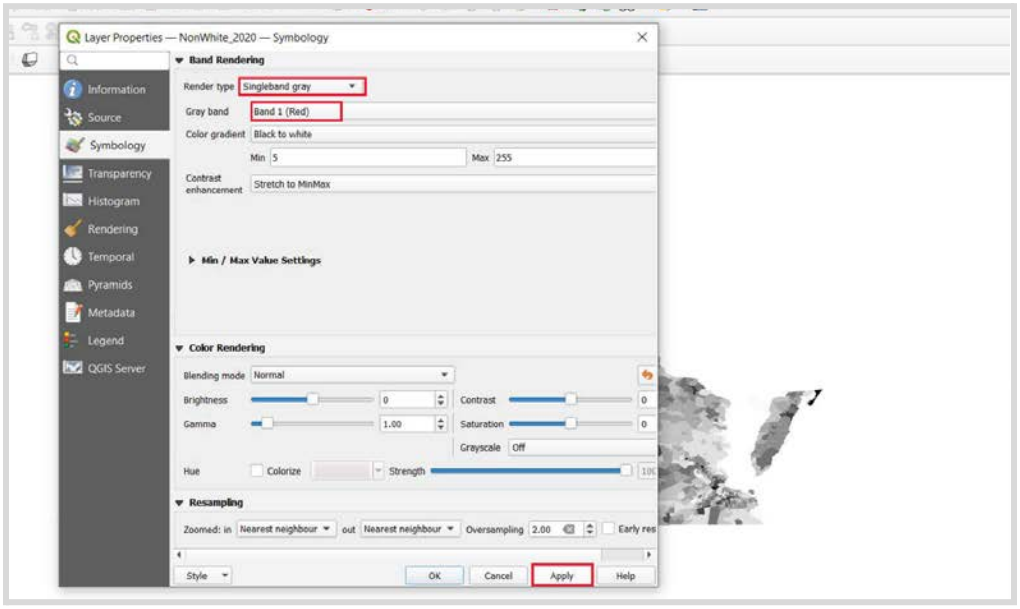
Click Save, select **"TIFF",** and save your new Raster image somewhere safe. Repeat this step for the 2020 layer.



9c Add back in your new Raster layers with Layer > Add Layer > Add Raster Layer…

9d Double click on one of your census raster layers.You'll notice this looks a little different from the Vector layer property box. Notice the property "Band Rendering" (the first property under "Symbology"). This will automatically set to "Multiband Color". However, when you open the Raster Calculator in the next step you'll see that it asks you to choose a band to compare between the two raster layers. Since we don't want to compare the data across three color bands, we'll simplify it to one.

Select “**Singleband Grey**”. You can leave the “Grey Band” to “Band 1 (Red)”. Make a note of this, since we’ll want to compare the Band 1 of both layers in the next step. Select “Apply”

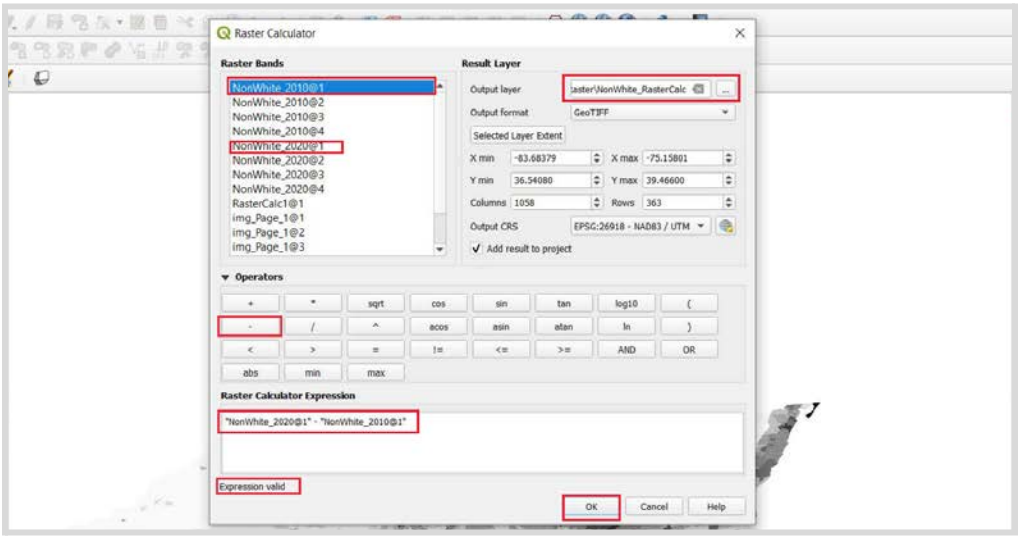


Repeat this for both layers.

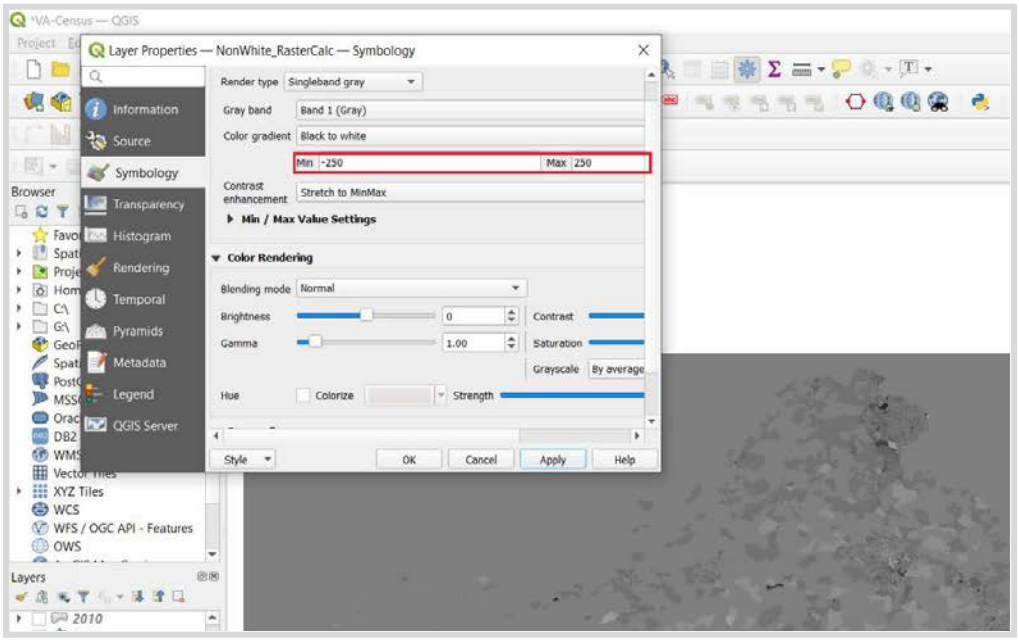
9e Now, open **Raster > Raster Calculator**. Since we can see that the Nonwhite percentage has grown between 2010 and 2020, we want to subtract 2010 from 2020. As with the Graduated Symbology expression editor, you can double click to add layers to your Raster Calculator expression. First, Double click on the 2020 Band 1 layer listed in the “Raster Bands” list (upper left). Then click the minus (-) symbol in “Operators”.

Note: Notice that at this point, before completing the expression, you have a text message in the bottom left that says **“Expression invalid”**. This text, which appears in all QGIS expression editors, is a useful indicator to keep an eye on.

Now, double click on your 2010 Band 1 layer in the “Raster Bands” list to complete the expression. In the upper right, set the “Output Layer” to a safe place with a logical name. Click “OK”.



9e You should now see the new Raster Calculated layer added to your map. It will automatically set its color scale from -250 to 250. The grays that are LIGHTER than the background gray are negative values (eg areas where NonWhite percentage DECREASED from 2010 to 2020); the grays that are DARKER than the background are positive values (where NonWhite percentage INCREASED).



If you only want to see the decrease, double click on the layer and set your Band Rendering MIN to 0. Leave your MAX at 250. Likewise, if you only want to see the increase, set the MIN to -250 and the MAX to 0.