

LAB 5 – CALCULATING & QUERYING ATTRIBUTE DATA: PART 1

What you'll Learn: This is the first of a two-part lab where we learn how to query, calculate, and rank attribute data. You will learn how to apply an attribute join.

Data: Kentucky counties and block groups as a geodatabase and tabular data as a text file.

What You'll Submit: 1) Answers to all questions in a .doc or .pdf, and 2) a first-draft map (in .pdf or .png) of Kentucky Distressed Block Groups that shows ranked areas of economic distress based on the latest unemployment rate, educational attainment rate, and per capita market income (PCI). This data is derived from the 2009-2013 American Community Survey 5-year estimates.

Lab naming conventions: Tools that you click will be bolded, e.g., **QGIS Menu > File > New** to create a new QGIS project file. Text that you'll type will have quotes around it, such as "MyNewProject.qgs" and names of existing datasets and directories will be *italicized*, e.g., *DataToUse.zip*. Key terms will be underlined. **Important tips and key instructions will be in bold red font.**

Note: This is part one of a two-part lab that will continue next week. **Keep the data!**

STEP 1: BACKGROUND

Typically, I put background info in the top bar, but it's important enough this week to constitute it's own step – so read carefully!

Most spatial data in a GIS consist of at least two types of data: those (spatial) data depicting the *location and shape of features*, and text or numerical (descriptive) data *describing the attributes of features*. These text and numerical data are contained in tables, like Excel spreadsheets or .csv files. We have many tools to query, select, and calculate new fields in these tables.

This data can inform significant social and economic policy. For example, Census data defines how many representatives each state has in congress. In this lab we will be using Census data to wrestle with the issue of socioeconomic distress in Kentucky.

Social and economic (socioeconomic) distress is federally defined for special areas in the U.S. using census and other federal agency data. This distress designation has considerable impact, because it focuses more resources and attention on those areas. Appalachia has long been one of those areas.

The agency tasked with managing millions of federal funding for the region is the **Appalachian Regional Commission (ARC)**. The ARC allocates this funding to economically distressed counties in the region, using a measure of distress based on three economic indicators:

1. Three-year average **unemployment rate** that is greater than 1.5 times the U.S. average (2008-2010) of 8.2%
2. A **per capita market income** (PCI) that is less than two-thirds of the U.S. average of \$27,344
3. AND a **poverty rate** that is greater than 1.5 times the U.S. average of 15.1%;
4. OR they have greater than 2 times the U.S. poverty rate and qualify on the unemployment or income (PCI) indicator.

To reiterate: when appropriate criteria from the above list are met, an area is considered socioeconomically distressed. One problem this federal program has faced is using the county as the resolution of analysis. Eastern Kentucky has the majority of distressed counties in the region, yet it is one of the least populated areas of Appalachia. This analysis might be seen as a Map of Rural Distress, since wealthier areas in urban counties could skew Per Capita Income for the whole county, for example, and make the county appear less distressed.

We are going to find distress in Kentucky at the smallest unit for which data is available, the block group, for the whole state. We will use similar measures that ARC uses to calculate and rank distress. We'll use slightly different measures, because our data is different.

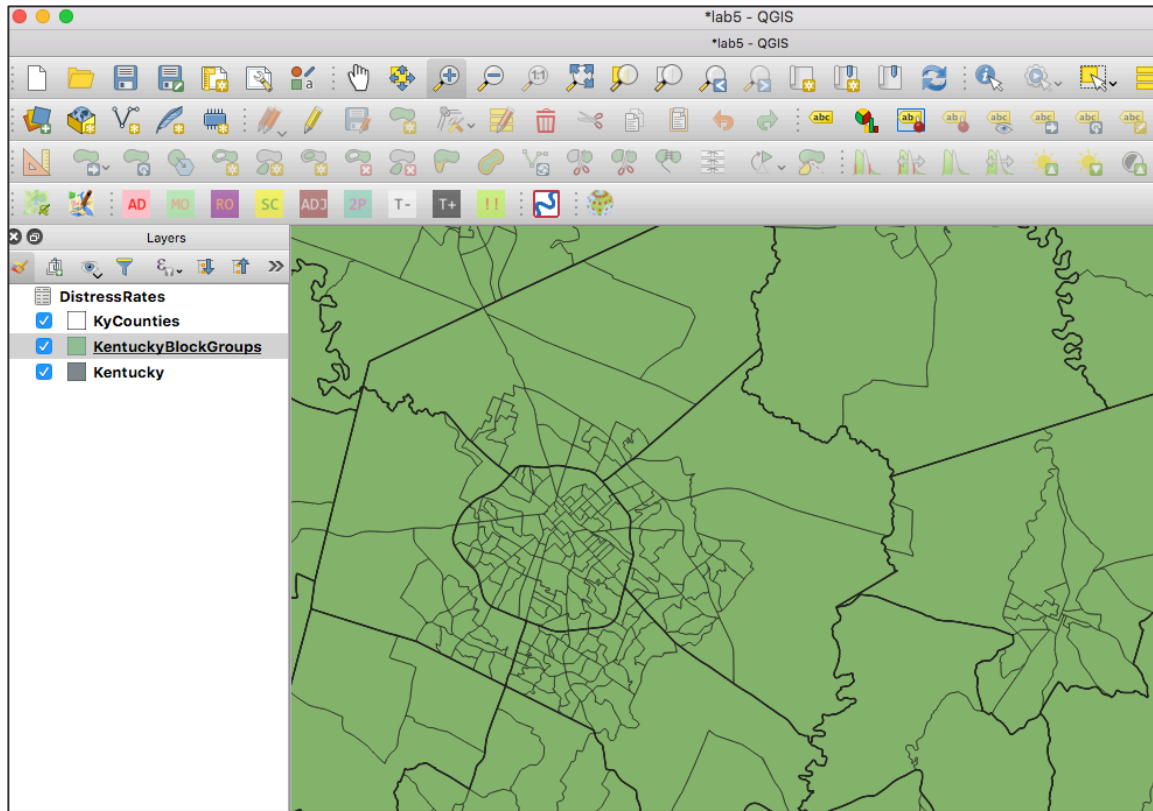
STEP 2: SET UP YOUR WORKSPACE

As usual, set up an appropriate workspace (see previous labs for reference if you have questions). Download the data from Canvas and unzip it.

Create a new map document in QGIS and add the following files to it:

- Shapefile titled "Kentucky"
- Shapefile titled "KyCounties"
- Shapefile titled "KentuckyBlockGroups"
- Comma separated values (.csv) file titled "DistressRates"

Organize the layers to match the screenshot below, making the fill hollow (e.g., set the fill opacity to 0) in the *KyCounties* layer so you can see the block groups:



You may also want to thicken the line. Considering points on map design from last week, what colors might you want to set for the lines in the block group layer versus the counties layer to make them pop out or fade away?

Be sure your data is projecting in the **Single Zone Kentucky State Plane Coordinate System (FIPS 1600)**.

Save your project entitled the following: *Ky_Distressed_Areas_[Your Last Name].qgs* in your Lab5 folder.

STEP 3: JOINING EXISTING TABLES

The table join is a key function in GIS work. Table joins allows us to connect spatial data (e.g., block group polygons) to descriptive data (e.g., distress indicators in a .csv) via a common field (e.g., column). You will see this common field referred to as a primary key, a joint item, or sometimes a GEOID (note that this is literally short for “geography identification” and not to be confused with the geometric geoid!).

Let’s begin by exploring your data. Select the layer *KentuckyBlocksGroups* and **right+click > Attribute Table**. Examine the table. Each record in this table corresponds to a block group, the smallest unit of geography we can use for demographic mapping. Note the fields, especially one called [GEOID_Data] and [GEOID].

Now **Right+click** on *DistressRates.csv* to see its attribute table. Notice that this table also has a field named [GEOID]. The file *DistressRates.csv* contains the following attributes:

1. [PercentUnemployed], which shows the ratio of people in the civilian labor force 16 years and older not employed.
2. [PCI] is the average income per person.
3. [PercentNoHS] shows the ratio of people 25 years and older that have had some high school education, but didn't graduate (didn't receive a diploma).
4. [TotPop] is the total count of folks that live in the block group.

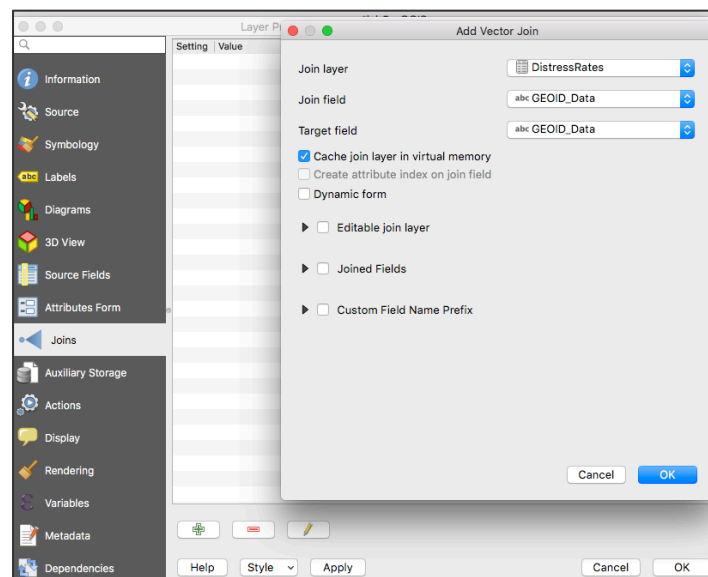
Compare these two attribute tables. Note that they have the same number of records – both datasets refer to Kentucky's block groups, but contain different data. We want to merge these files and attach all of the rich Census information in the descriptive data with the spatial data that currently contains little in the way of descriptive attributes.

Now we'll begin the process of actually joining these tables!

First, **right+click** on the *KentuckyBlocksGroups* layer in the TOC and select **Properties > Joins**.

This interface will allow you to select attribute table from another layer in QGIS, and attach it to the current layer.

Click the green plus and imitate the screenshot to the right. The options refer to the following:



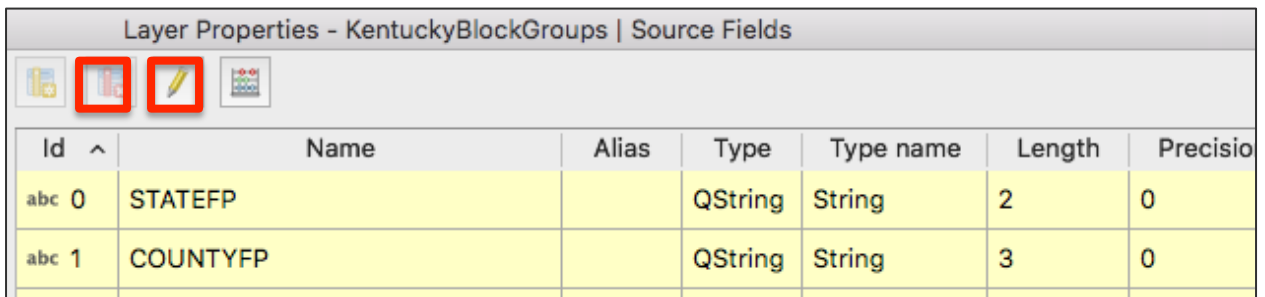
1. **Join layer** – The table you wish to join to the current table. It's the .csv file, *DistressRates.csv*.
2. **Join field** – The field in this layer used to match records. It's [GEOID_Data].
3. **Target field** – The field in the joining table that you will use to match records. It's [GEOID_Data].

Once you're done, press OK, and you should see a new entry in the window. What do you see when you close the **Properties** window and open the attribute table for *KentuckyBlockGroups*?

You've just joined the two tables, matching the records in one table to the records in

another table. **Right now, however, the join is only temporary, and exists only in QGIS.** In other words, we don't yet have a shapefile that contains all of this information in the same place. Let's do a little clean up and make our join permanent.

- a. In *KentuckyBlockGroups*, navigate **Properties > Source Fields**. At the top of the screen, you should see a toolbar that allows you to initiate **Editing Mode** (the pencil) and **Add** or **Remove Fields** (the buttons to the left and right).



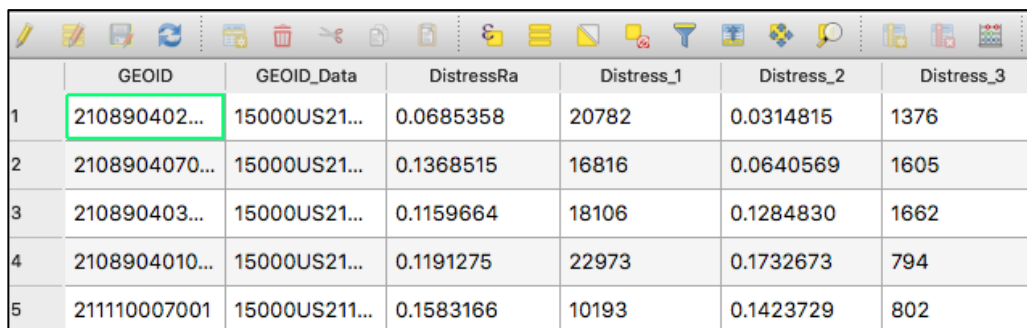
Id ^	Name	Alias	Type	Type name	Length	Precisio
abc 0	STATEFP		QString	String	2	0
abc 1	COUNTYFP		QString	String	3	0

First, **click** the pencil to activate **Editing Mode**. Then, by clicking on fields and then selecting the red **Remove Field** button, remove all but the following fields: GEOID, GEOID_Data, TotPop, PercentNoHS, PCI, and PercentUnemployed. **Click** the pencil again and save your changes.

- b. As I said, this is a temporary join; the original files/data have not been modified. QGIS keeps track of joins in views, and how to display the various joined files. The data are not copied to a new, combined file.

Export layer to a new feature class via **right+click** the **layer > Export > Save Features As...** Follow the instructions **to save this as an ESRI shapefile, in a new folder called "created-data"**. Make sure you keep the correct coordinate system and you it in. Call the new feature class, "KyDistressBG" and click OK.

When you complete all this, your attribute table should look like the below:

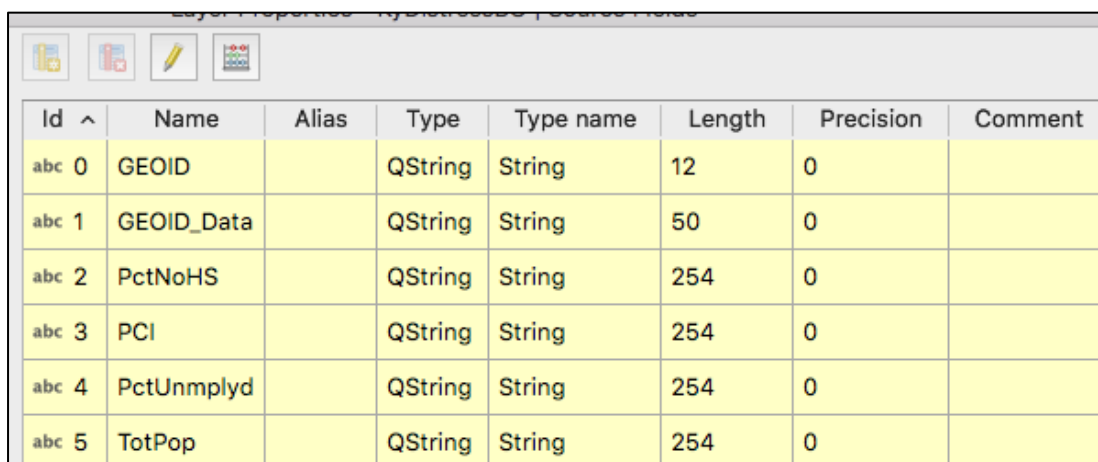


	GEOID	GEOID_Data	DistressRa	Distress_1	Distress_2	Distress_3
1	210890402...	15000US21...	0.0685358	20782	0.0314815	1376
2	2108904070...	15000US21...	0.1368515	16816	0.0640569	1605
3	210890403...	15000US21...	0.1159664	18106	0.1284830	1662
4	2108904010...	15000US21...	0.1191275	22973	0.1732673	794
5	211110007001	15000US211...	0.1583166	10193	0.1423729	802

Much cleaner -- but what happened to our field names??

QGIS has a string limit of 10 characters for field names when it exports to new

shapefiles, which means the field names from our join got cut off. Let's rename them by returning to the **Properties > Source Fields** window in the new *KyDistressBG* layer and using the **Editor Mode** to update field names as such:



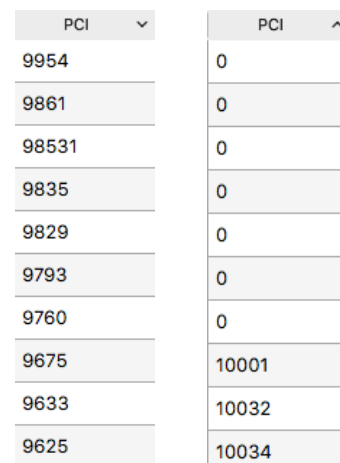
Id ^	Name	Alias	Type	Type name	Length	Precision	Comment
abc 0	GEOID		QString	String	12	0	
abc 1	GEOID_Data		QString	String	50	0	
abc 2	PctNoHS		QString	String	254	0	
abc 3	PCI		QString	String	254	0	
abc 4	PctUnmplyd		QString	String	254	0	
abc 5	TotPop		QString	String	254	0	

Finally, you'll note that all of these fields are type "string," which means we can't actually run numerical calculations on them. In fact, if you go to the attribute table and sort the fields, you'll notice a peculiar organization:

On the left, we have PCI sorted high to low, while on the right, we have PCI sorted low to high. They aren't sorting numerically, but rather *alphabetically* – which is to say, "1000000" is a *lower value* than "2" – because the data types are "string."

So if we want to perform any analysis on these fields, we'll need to update their data types to "real".

You can fix this fairly easily for all 4 fields by opening the **Attribute Table**, enabling the **Editing Mode** (if it isn't already enabled), and opening the **Field Calculator**.



PCI ▾	PCI ^
9954	0
9861	0
98531	0
9835	0
9829	0
9793	0
9760	0
9675	10001
9633	10032
9625	10034

First, create a **New Field** titled "PCI_" with the output field type "Whole Integer," output field length "7," and enter the following text in the expression window:

```
to_real("PCI")
```

Click OK, and a new field should appear in your attribute table. Replicate this process for the other three fields, mimicking the inputs as seen in screenshots below (PCI included). **Be sure you are working in editing mode; make careful note of the field length and precision; and be sure to copy the expressions exactly!**

☒ Create a new field

☐ Create virtual field

Output field name:

Output field type:

Output field length: Precision:

Expression:

Function Editor:

☒ Create a new field

☐ Create virtual field

Output field name:

Output field type:

Output field length: Precision:

Expression:

Function Editor:

☒ Create a new field

☐ Create virtual field

Output field name:

Output field type:

Output field length: Precision:

Expression:

Function Editor:

☒ Create a new field

☐ Create virtual field

Output field name:

Output field type:

Output field length: Precision:

Expression:

Function Editor:

If completed properly, you should see an attribute table that resembles the following:

KyDistressBG :: Features Total: 3285, Filtered: 3285, Selected: 0

KyDistressBG :: Features Total: 3285, Filtered: 3285, Selected: 0

abc: GEOID

	GEOID	GEOID_Data	PctNoHS	PCI	PctUnmplyd	TotPop	PctNoHS_	PCI_	PctUnmply_	TotPop_
1	211110116043	15000US211...	0.0000000	52821	1.0000000	34	0.0000000	52821	1.0000000	34
2	2111706700...	15000US211...	0.1134565	10282	0.6857143	424	0.1134565	10282	0.6857143	424
3	211110043021	15000US211...	0.2842105	6440	0.6265060	358	0.2842105	6440	0.6265060	358
4	2113395040...	15000US211...	0.0381232	14215	0.5970149	478	0.0381232	14215	0.5970149	478
5	210039203...	15000US21...	0.2776280	9954	0.5415162	566	0.2776280	9954	0.5415162	566
6	211110059002	15000US211...	0.2365662	7993	0.5354331	1914	0.2365662	7993	0.5354331	1914
7	2111100590...	15000US211...	0.0569106	12520	0.5098814	453	0.0569106	12520	0.5098814	453

Show All Features

Go ahead and delete the original fields, keeping all fields with underscores, and close editing mode before moving on.

STEP 4: BUILDING A QUERY TO FIND DISTRESSED AREAS

Now that we've got our descriptive data joined to spatial data – and properly formatted, at last – what do we do with it in order to find distressed areas?

Recall that criteria for socioeconomic distress that I mentioned earlier in Step 1:

1. Three-year average **unemployment rate** that is greater than 1.5 times the U.S. average (2008-2010) of 8.2%
2. A **per capita market income** (PCI) that is less than two-thirds of the U.S. average of \$27,344
3. **AND** a **poverty rate** that is greater than 1.5 times the U.S. average of 15.1%;
4. **OR** they have greater than 2 times the U.S. poverty rate and qualify on the unemployment or income (PCI) indicator.

We are going to categorize Kentucky Census Block Groups based on the definition of distress similar to ARC method. We will then rank them in a custom scheme:

Not Distressed (Rank 0)

1. An **Unemployment Rate** that is **below** the U.S. average of 9.0%
2. An **Income** (PCI) that is **above** of the U.S. average of \$28,155
3. **AND** a **No Diploma Rate** (no high school diploma) that is **below** the U.S. average of 8.0%;

Transitional (Rank 1)

1. Not in Ranks 0, 2 or 3.

Distressed (Rank 2)

1. An **Unemployment Rate** that is greater than 1.5 times the U.S. average
2. An **Income** that is less than 80% of the U.S. average
3. A **No Diploma Rate** greater than the U.S. average
4. **OR** they have an **Income** less than 60% the US average rate **AND** a **No Diploma Rate** greater than 1.5 times the US average

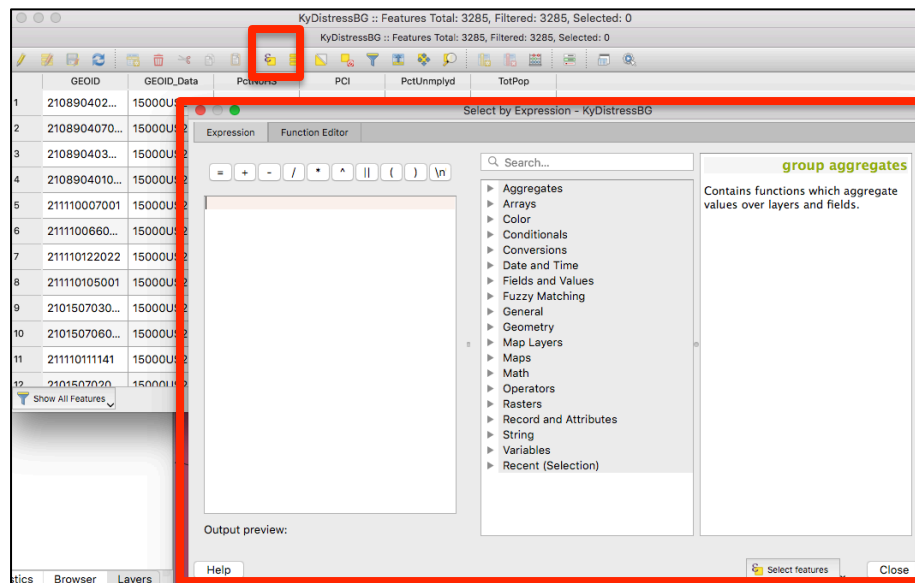
Severely Distressed (Rank 3)

1. An **Unemployment Rate** that is greater than 2 times the U.S. average
2. An **Income** that is less 60% of the U.S. average
3. A **No Diploma Rate** greater than 1.5 times the U.S. average
4. **OR** they have an **Income** less than 40% the US average rate **AND** a **No Diploma Rate** greater than 2 times the US average

“HOLY COW!!!” you might be saying; or, “My head is spinning! How can I keep this straight??” I am here to assure you that it's actually going to be simpler than you think.

QGIS has a powerful functionality for filtering and selecting certain records with a **Select**

by **Expression** tool. To open the **Select by Expression** tool, navigate to *KyDistressBG* > **Attribute Table** and click the button that is highlighted in the small red box below:



You should then see the window indicated by the much bigger red box appear. This is your **Select by Expression** tool.

Let's start with the first rank and try to build the query to select those block groups that meet our requirements.

STEP 4.1: RANK 0 QUERY

Ok, so you know the three fields we're working with, but what is the query you'll use to find the first rank? It is suggested that you write out the query to help understand how to formulate the syntax. Consider the following simple equation:

Rank 0 = (PCI > \$28,155) AND (UNEMPLOYMENT < 9.0%) AND (EDUCATION < 8.0%)

This query fits the Rank 0 requirement, **BUT** we cannot use this syntax in QGIS. We have to follow rules (bummer ☹), but what's nice in QGIS is that query builder in the **Select by Expression** menu will help you make the query in the proper format. If you ever need help with syntax, reference the windows in the middle (double click items for auto-formatting) and right (a useful help tab).

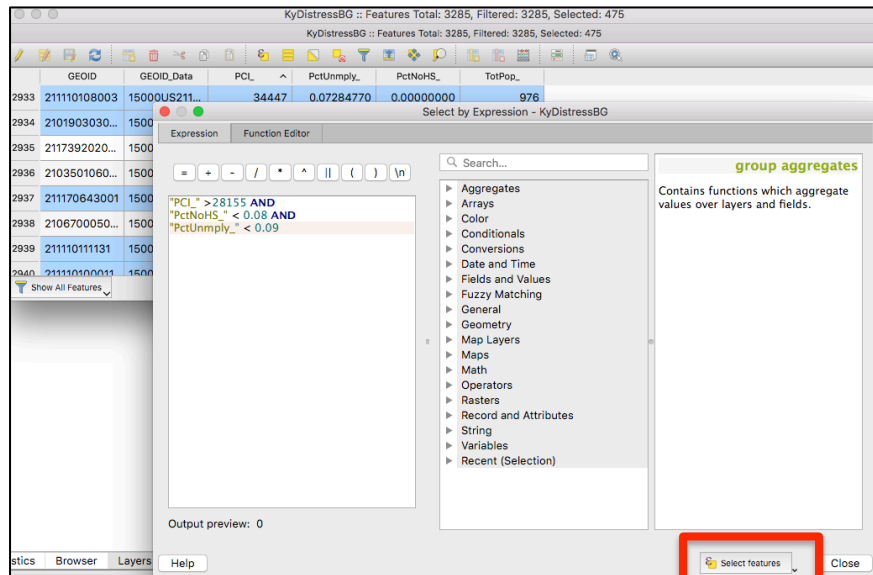
Make sure you *KyDistressBlockGroups* attribute table is open. Open the **Select by Expression** in attribute table. We'll build the query simply selecting fields, operators and typing in values. In the **Select by Expression** menu, populate your query by clicking on buttons and field names.

Your final query should like the following (see the screenshot):

```
"PCI_" > 28155 AND  
"PctNoHS_" < 0.08 AND  
"PctUnmply_" < 0.09
```

Once you **click** the **Select features** button, indicated by the red box, your attribute table should automatically select 475 features. **All of these features meet the criteria of a Rank 0 distress, but as of now, our only evidence is a selection query.**

We need to encode this data in a new field if we want it to stick.

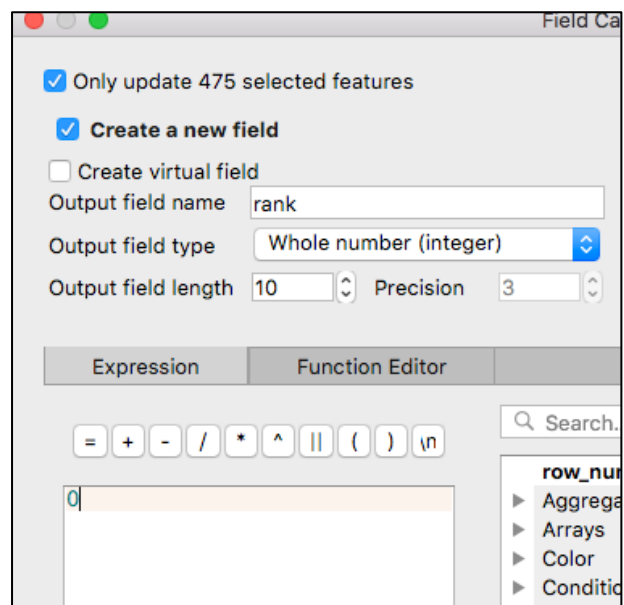


STEP 4.2: CALCULATE FIELD VALUES FOR RANK 0

In this section we are creating a new field and adding a unique value that indicates Rank 0. This is helpful for quickly selecting areas of a particular rank and for making a map that displays the geographic extent of distress.

Toggle **Editor Mode** in the attribute table back on and open the **Field Calculator**. Select **Create a new field**, name it “rank,” and select field type as “whole integer” and length of “1”. **Be sure that “Only update 475 selected features” is checked!**

Once you click OK, your attribute table should now show a field that includes values either “0” or “NULL” – the NULL values contain no data, while the 0 values contain a numeric value of 0. **The 0 represents a socioeconomic distress rank of 0 within the scale we established above.** The new field should resemble the screenshot below:



KyDistressBG :: Features Total: 3285, Filtered: 3285, Selected: 475

KyDistressBG :: Features Total: 3285, Filtered: 3285, Selected: 475

abc GEOID = [] Update All Update Selected

	GEOID	GEOID_Data	PCI_	PctUnmply_	PctNoHS_	TotPop_	rank
2517	2119993000...	15000US211...	28160	0.09727030	0.06920190	1391	NULL
2518	2109196020...	15000US21...	28167	0.01675980	0.07185630	1608	0
2519	2104720090...	15000US21...	28168	0.09866390	0.03906900	1838	NULL
2520	2108302070...	15000US21...	28169	0.05353320	0.11722140	992	NULL
2521	2101507031...	15000US21...	28187	0.05550880	0.02506270	1878	0
2522	2122701190...	15000US212...	28190	0.08333330	0.02725970	903	0
2523	2105900180...	15000US21...	28200	0.08700570	0.08191400	1635	NULL
2524	211450313012	15000US211...	28220	0.10815050	0.07363180	1406	NULL
2525	2105900110...	15000US21...	28245	0.15040000	0.06802720	1185	NULL
2526	211110111065	15000US211...	28250	0.07034220	0.09772420	999	NULL
2527	210590006...	15000US21...	28255	0.07060060	0.00000000	2246	0
2528	2101903060...	15000US21...	28261	0.10578840	0.10633210	1044	NULL
2529	211110124101	15000US211...	28268	0.03719910	0.03060050	2459	0
2530	211110100051	15000US211...	28283	0.06654680	0.05797100	891	0
2531	211170642001	15000US211...	28316	0.13181240	0.05413690	1418	NULL

Next, click the **Invert Selection** button – depicted inside the red box above – and repeat the field calculation. This time, of course, instead of creating a new field, you should be sure to check the **Update existing field** box.

After we have block groups calculated as either 0 or 1, we now can find the distressed block groups (e.g., all block groups ranked 1). The following two queries will find the more economically distressed areas that are currently ranked 1. And for those block groups, we'll replace that value with either a 2 or a 3.

PART 4.3: QUERING AND CALCULATING RANKS 2 AND 3

Here's where it gets a little tricky, since we are doing a more complex query that uses AND and OR, plus mathematical operators. Here's what want to query in simple syntax:

Rank 2 = (Condition 1) OR (Condition 2) where

Condition 1 = (PCI < 80% of \$28,155) AND
(UNEMPLOYMENT > 1.5 times the rate of 9.0%) AND
(NO DIPLOMA > 8.0%, the average rate in the U.S.)

OR

Condition 2 = (PCI < 60% of \$28,155) AND
(NO DIPLOMA > 1.5 times the rate of 8.0%)

Rank 3 = (Condition 1) OR (Condition 2) where

Condition 1 = (PCI < 60% of \$28,155) AND
(UNEMPLOYMENT > 2 times the rate of 9.0%) AND
(NO DIPLOMA > 1.5 times the rate of 8.0%)

OR

Condition 2 = (PCI < 40% of \$28,155) AND
(NO DIPLOMA > 2 times the rate of 8.0%)

Using the **Select by Expression** tool, replicate the previous steps to calculate Ranks 2 and 3. Remember, you need to cut and paste your query and put it in the Data Sheet to answer questions 1 and 2. **And don't forget to do this within the Editing Mode, saving at key checkpoints (e.g., after a successful calculation), so that if you mess something up you don't have to start all over!**

QUESTION 1: Copy and paste your query for Rank 2.

QUESTION 2: Copy and paste your query for Rank 3.

QUESTION 3: Why did we need a query for Ranks 0, 2, and 3, but not for Rank 1?

QUESTION 4: In 2-4 sentences, what cursory observations can you make about socioeconomic distress in Kentucky? And what unanswered questions do you still have from looking at the data?

QUESTION 5: What is a table join, why is it important, and how is it done?

TIPS:

- Use your math operators – e.g., selecting block groups with twice the Unemployment Rate can be expressed as: **(UnemploymentRate > 2 * 0.09)**
- Use parentheses to separate conditions, i.e.,
**((condition 1a) AND (condition 1b)) OR
((condition 2a) AND (condition 2b))**
- Copy and paste the conditions as they are written out above directly into the expression box; you'd obviously have to delete it before running your selection query, but that might help streamline the process of turning simple syntax into the syntax that will be recognizable in QGIS

Once you are done, clear any selected features and save your edits.

STEP 5: FIRST DRAFT AT A CHOROPLETH MAP

This has been a challenging lab that probably forced a lot of you out of your comfort zones. That said, you should feel proud having completed it, especially if this is not the kind of thing you were familiar with at the beginning of the semester!

To finish Lab 5, mock up a first draft choropleth map of socioeconomic distress in Kentucky, laid out and exported from the print composer with the following elements:

1. A choropleth map of Kentucky of socioeconomic distress – in *KyDistressBG*:
 - a. Open the layer styling panel
 - b. Select “categorized”
 - c. Use the “rank” column
 - d. Use the “reds” color ramp
2. Include visible counties with dark gray lines and hollow fill
3. *KyDistressBG* stroke lines should be transparent
4. A title
5. A legend
6. Data source, date, and your name
7. Any other cartographic accouterment is welcome but not required

Note that Lexington, Louisville, and Covington are hard to see. But (thankfully) that’s a problem we can save until next week to fix!

STEP 6: REVIEW & SUBMIT

You should submit completed materials for Lab 5 via Canvas by Monday, 2/24 at 11:59pm. The completed materials include:

- A Word document containing answers to the questions posed at the end of Step 4 in these lab instructions
- One map, formatted in print composer w/ all requisite elements, in .pdf or .png

Be in touch if you have any questions!