

## Join, normalize, clip, analyze

In this tutorial we are proceeding to drill in to reveal and analyze Census data for Richmond, California. This is not easy, because Census-Tract data does not line up with Richmond. Therefore we will need to do the following:

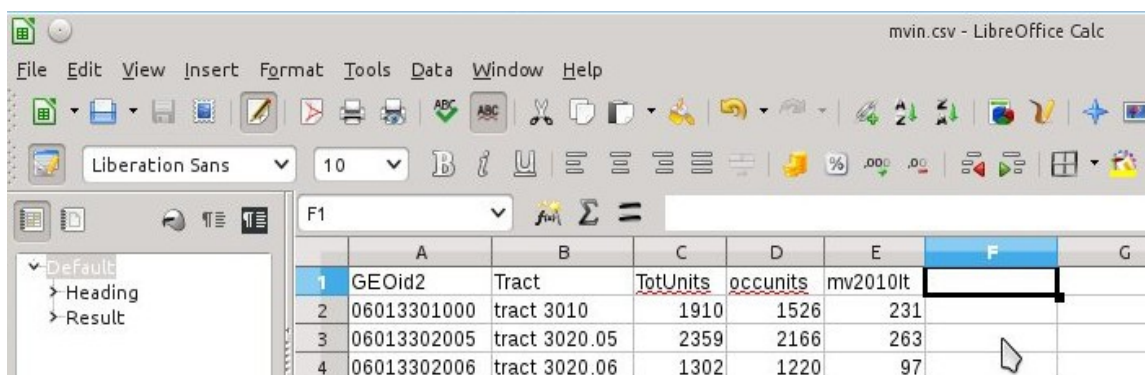
1. Join in the data we are interested in, but for all of Contra Costa;
2. **Normalize the data** we will analyze for Richmond by calculating **area densities**;
3. Use a boundary file of the city of Richmond to clip away all the rest of Contra Costa County;
4. Recalculate the areas of the partial Census-Tracts for Richmond only;
5. And then recalculate the raw figures (populations, housing, etc) for the partial Census-Tracts of Richmond itself.

I will continue to use QGIS to show how this is done. The *process* is identical in ArcMAP; all that differs is the *method*. In ArcMAP, the geoprocessing functions can be found in the Toolbox panel. In QGIS, they are organized under the drop-down menus from the top, like most forms of software.

*Note: I have run out of time to annotate this tutorial. I am posting the PDF with screen-shots only. Hopefully most of the example will be clear. PC*

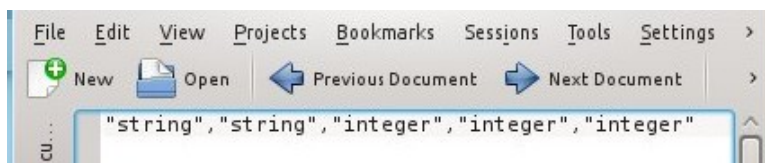
### 1. Chunk down the FactFinder CSV even further.

Before joining my Census data to my shapefile, I shaved down the data to only 5 fields and saved the file as “mvin.csv”. That way, when I join it, the Attribute-names will be reasonably short.



	A	B	C	D	E	F	G
1	GEOid2	Tract	TotUnits	occunits	mv2010lt		
2	06013301000	tract 3010	1910	1526	231		
3	06013302005	tract 3020.05	2359	2166	263		
4	06013302006	tract 3020.06	1302	1220	97		

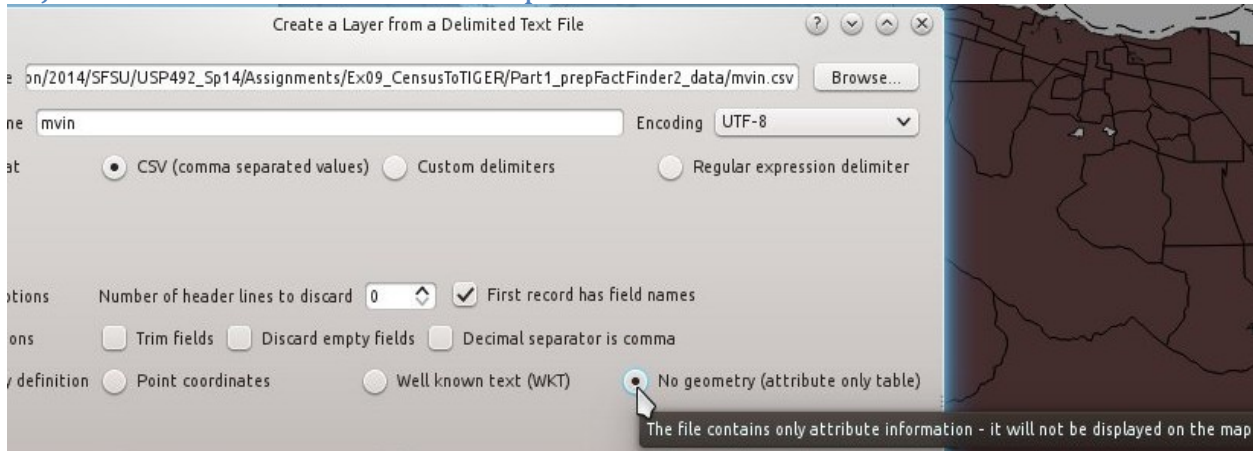
To bring the CSV into QGIS, I then take that extra step of creating a CSVT text file that describes the data:



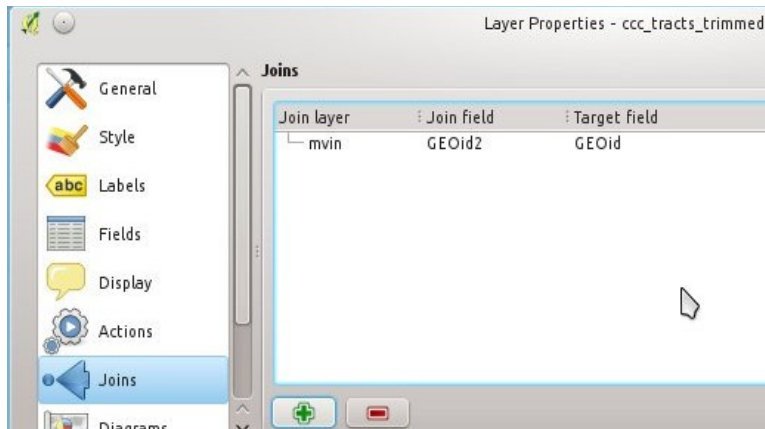
Two more things to note about CSVT files. FIRST: they can specify only three types of data: text **string**, whole number **integer**, and **real** numbers with decimal places. The file I am joining only has “string” and “integer” data, but if I had a field with fractional/decimal data I would specify “real”.

SECOND: if you are using a Mac, the most basic text editor is called TextEdit, and ironically, it cannot save as raw text (.txt) files by default! Here is the fix: 1) start TextEdit. 2) Open the Preferences dialog. 3) Under New Document Format, click the Plain Text option. 4) under Saving, make sure to check the box ‘Append “.txt” extension to plain text files.’

## 2. Join the CSV file to the Tracts shapefile.



Now, add the CSV file to QGIS via **Layer > Add delimited text layer...** which brings up the **Create a Layer from a Delimited Text File** dialog (above). Note that if you have coordinate-data in a CSV file, you could turn it into a spatial layer right away upon import. However we will specify **No geometry (attribute only table)**.



Then double-click the Tracts shapefile layer (or right-click and select "Properties"), and in the left-hand pane, select **Joins**. In the right-hand panel, click the green plus-sign at the bottom. Choose the "mvin" file to join it.



After all our grief with GEOid fields, here is the moment where they must line up in order for the join to succeed.

Note: there is another option, which I noticed as I was building this tutorial. Shapefiles downloaded directly from FactFinder many not include the 11-digit GEO.id2 field, but the *do* include the original, 20-digit GEO\_ID field. Tabular data downloaded from AFF in CSV format also includes this GEO\_ID field. The reason I disregarded that older, longer geocode was that I built this tutorial off of older tutorials. In the future I will just work with the full GEO\_ID field.

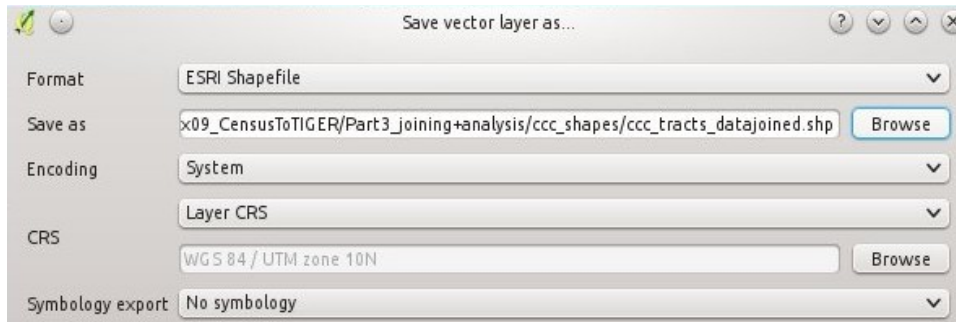
Below is the attribute table for this example. Both QGIS and ArcMAP include the CSV filename **and** the field name in the header, so if your file name was long, your header name may be difficult to read unless you make the columns very wide.

Attribute table - ccc\_tracts\_trimmed :: Features total: 207, filtered: 207, selected: 0

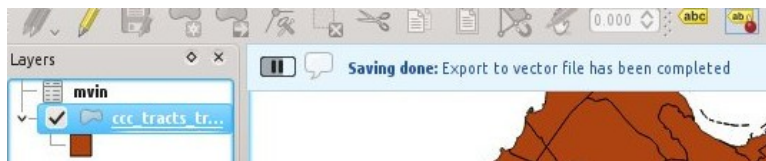
	NAME	GEOid	area_sqm	area_acre	mvin_Tract	mvin_TotUnits	mvin_occunits	mvin_mv2010lt
0	3020.06	06013302006	6143394	1518.066	tract 3020.06	1302	1220	97
1	3031.02	06013303102	6525422	1612.467	tract 3031.02	2365	2186	190
2	3131.02	06013313102	2004941	495.432	tract 3131.02	1415	1257	135
3	3132.06	06013313206	1718676	424.694	tract 3132.06	1639	1545	249

### 3. Save As a new file with data permanently joined.

Under some circumstances you may want to keep the data-joins temporary, especially if you are working with huge and constantly-changing sets of data. In our case, a permanent is useful so that it is easier to move our data around (from home machine to laptop to HSS lab), and to avoid any problems with the **Field Calculator**, which we are about to use a lot.

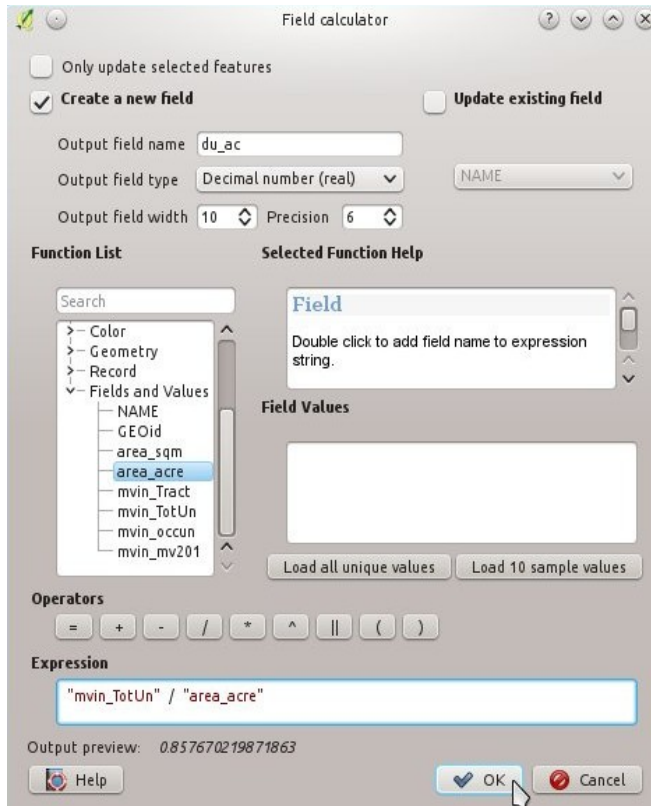


I saved mine as a shaefile called “ccc\_tracts\_datajoined”.



#### 4. Normalize the data: calculate densities for each spatial attribute.

Remember: all we have is county-wide data from FactFinder. Later we will clip away everything but Richmond data. In many places, 5/7ths or 1/9th of a tract may lie within Richmond. How will we know how many houses lie within these fractions of whole tracts? For this exercise, we will assume that the entities are evenly-distributed across the tract (In more detailed analysis you could show how distributions are uneven, and process the data differently). If we assume that the items (houses, people, recent move-ins) are evenly distributed, then we can calculate simple *densities* based on the area of each tract.



RELEASE THE FIELD CALCULATOR!!!

Okay, I am getting a bit loopy while compiling this tutorial. But seriously: make a nice short **Field name**. In this case, I am using the well-known planning term for dwelling units per acre: “du\_ac”.

Set the **Precision** to at least three decimal places. In my case, I am going for higher precision to begin with, and then I will step-down the precision as I pull in coarser data.

The **Expression** is pretty straightforward. However, a typo can cause trouble with the finicky syntax. So I just double-click the items in the **Function list** that I want to add to the **Expression**.

In this case the **Output preview** is misleading. The fact that anything shows up means that the **Expression** is valid, but it will not calculate to twenty decimal places.

I use the same method with the **Field calculator** to normalize the remaining data:

- “mvin\_occun” > “ocdu\_ac”; Precision 6; Expression = “mvin\_occun” / “area\_acre”
- “mvin\_mv201” > “mvin10\_ac”; Precision 6; Expression = “mvin\_mv201” / “area\_acre”

Once I calculated these density fields, I deleted the raw-number fields:



## 5. Now get the city boundary file of Richmond.

As in the previous tutorial, I am going to FactFinder to get this shapefile. Under **Geographies**, go to the second tab, **Name**, and search for “Richmond, California”. Under **Geography Filter Options**, select “City or Town”. Then in the **Geography Results** pane, select “Richmond city, California”. This now appears in the left panel under **Your Selections**. Then click the **Map** tab in the **Select Geographies** pane.

The screenshot shows the FactFinder interface. On the left, the 'Your Selections' pane shows 'Richmond city, California' selected under 'Geographies'. The 'Select Geographies' pane has the 'Name' tab selected, showing search results for 'Richmond city, California'. The 'Geography Filter Options' pane shows 'City or Town (1)' selected. The 'Geography Results' pane shows a list of results, with 'Richmond city, California' selected.

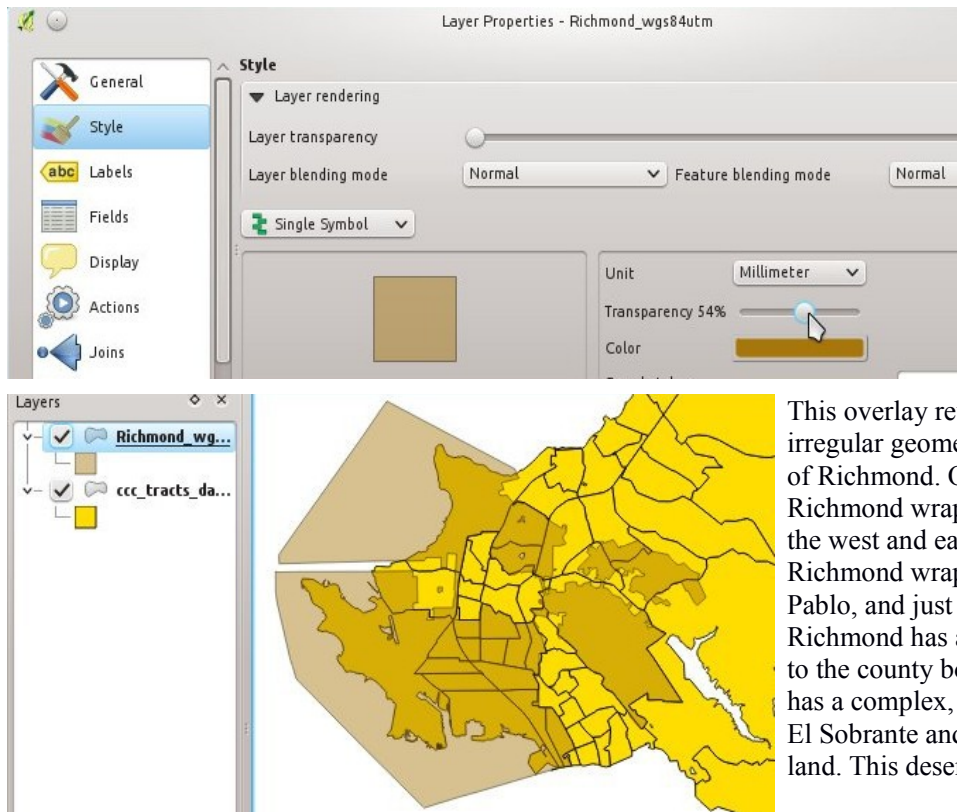
Geography Name	Geography Type
Richmond city, California	Place within State
All Blocks within Richmond city, California	Block
All Census Tracts (or parts) fully within/partially within Richmond city, California	Census Tract (or within Place)

After a few moments a preview of your shapefile will appear. Click the **Download** icon (note the yellowed cursor's position in the screenshot below).

The screenshot shows the FactFinder interface with the 'Map' tab selected in the 'Select Geographies' pane. A map of Richmond, California, is displayed, showing the city boundary in yellow. A yellow cursor is hovering over the 'download' icon in the top right corner of the map. The 'Legend' pane shows 'Your Selections' as 'undefined'.

As in the previous tutorial, you will need to load this shapefile, then **Save As** and specify the WGS 84/UTM Zone 10 North projection. Then reload it into the same workspace as your “ccc\_tracts\_joined” shapefile.

The next thing I did was open the **Layer Properties** of this shapefile, go to the **Style** pane, and set the **Transparency** of the layer to about 50% (screenshot below).



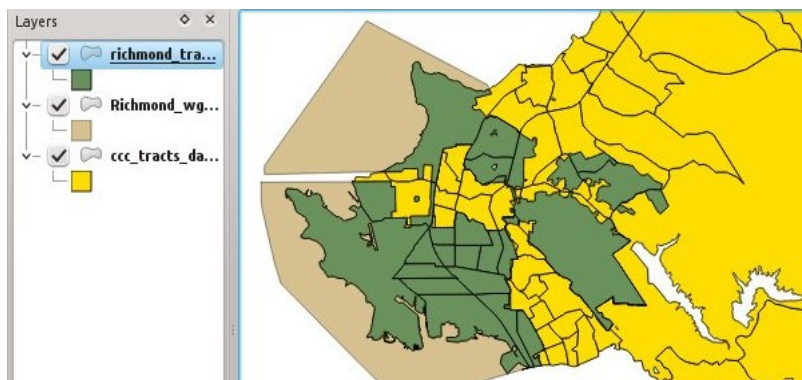
This overlay reveals the disturbingly irregular geometry of the city boundaries of Richmond. On the Southeast, Richmond wraps around El Cerrito on the west and east. In the middle, Richmond wraps completely around San Pablo, and just west of that, North Richmond has a sliver that extends west to the county border. In the northeast it has a complex, interdigitated border with El Sobrante and unincorporated county land. This deserves a research paper.

## 6. Clip the County shapefile using the city boundary file.

In any case, I can now clip my “ccc\_tracts\_datajoined” shapefile using Richmond’s boundary.

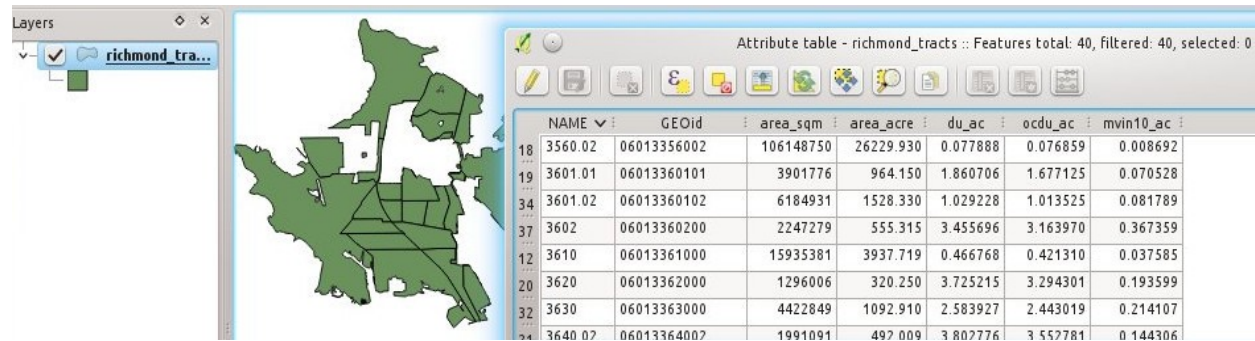
- **Vector > Geoprocessing Tools > Clip**
- **Input vector layer:** the Tracts file with the data.
- **Clip layer:** the city boundary file.
- **Output shapefile:** {path}/richmond\_tracts.shp

Remember, it will only work if both shapefiles are in the same datum / projection. If successful, QGIS will offer to add it to the current workspace. As you can see, what remains is only the land-area of the tracts within the city of Richmond (in green).



## 7. Recalculate the data for the clipped tracts.

In the screenshot below you can see that the clipped shapefile retains the Attribute data of the county (yay!).



	NAME	GEOid	area_sqm	area_acre	du_ac	ocdu_ac	mvin10_ac
18	3560.02	06013356002	106148750	26229.930	0.077888	0.076859	0.008692
19	3601.01	06013360101	3901776	964.150	1.860706	1.677125	0.070528
34	3601.02	06013360102	6184931	1528.330	1.029228	1.013525	0.081789
37	3602	06013360200	2247279	555.315	3.455696	3.163970	0.367359
12	3610	06013361000	15935381	3937.719	0.466768	0.421310	0.037585
20	3620	06013362000	1296006	320.250	3.725215	3.294301	0.193599
32	3630	06013363000	4422849	1092.910	2.583927	2.443019	0.214107
24	3640.02	06013364002	1991091	497.009	3.807776	3.552781	0.144306



Field calculator

☐ Only update selected features

☒ Create a new field ☐ Update existing field

Output field name:

Output field type:

Output field width:  Precision:

Function List

- Operators
- Conditionals
- Math
- Conversions
- Date and Time
- String
- Color
- Geometry
  - \$area
  - \$length
  - \$perimeter

Selected Function Help

**\$area function**

Returns the area size of the current feature.

**Syntax**

\$area

**Arguments**

None

**Example**

\$area = 42

Operators

= + - / \* ^ || ( )

Expression

Output preview: 13.2328108549118

Using the **Field Calculator**, the first step is to calculate the area of the clipped Census Tracts.

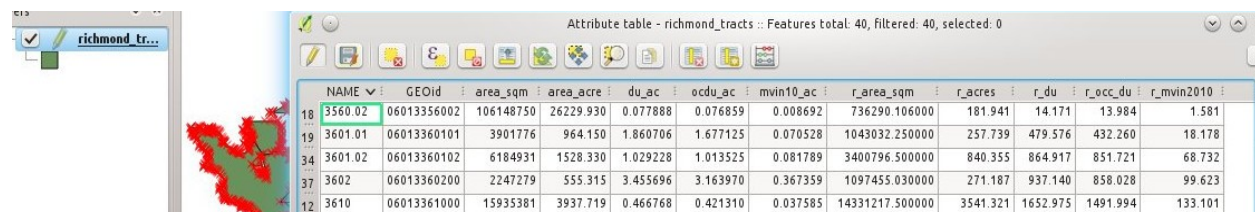
**Name:** "r\_area\_sqm"  
**Type:** "Decimal number (real)"  
**Precision:** 6  
**Expression:** \$area

Then I use the **Field Calculator** to calculate the acreage of each Tract-remainder:

**Name:** "r\_acres"  
**Type:** "Decimal number (real)"  
**Precision:** 3 [because my conversion number only has 3 decimal places]  
**Expression:** "r\_area\_sqm" / 4046.856

Finally, I use the density-fields to recalculate estimated raw numbers for Dwelling Units, Occupied Dwelling Units, and Move-ins postdating 2010. The resulting Attribute Table is shown below.

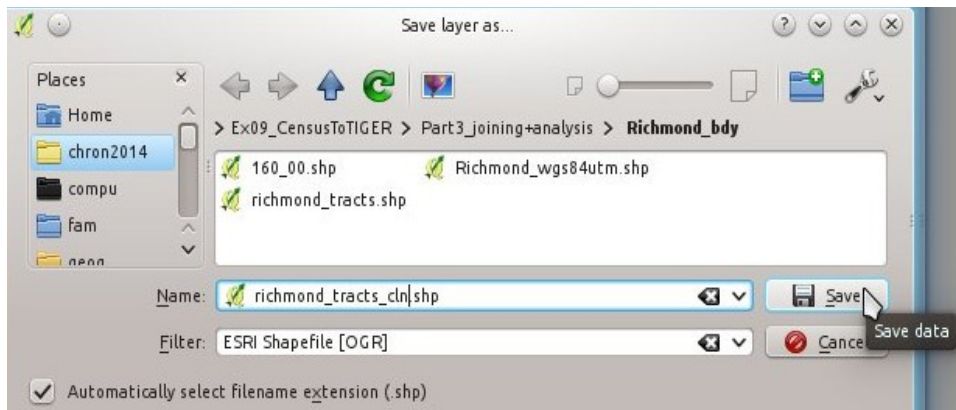
- "du\_ac" > "r\_du"; Type: real; Precision: 3; Expression = "du\_ac" \* "r\_acres"
- "ocdu\_ac" > "r\_ocdu"; Type: real; Precision: 3; Expression = "ocdu\_ac" \* "r\_acres"
- "mvin10\_ac" > "r\_mvin2010"; Precision 3; Expression = "mvin10\_ac" \* "r\_acres"



	NAME	GEOid	area_sqm	area_acre	du_ac	ocdu_ac	mvin10_ac	r_area_sqm	r_acres	r_du	r_ocdu	r_mvin2010
18	3560.02	06013356002	106148750	26229.930	0.077888	0.076859	0.008692	736290.106000	181.941	14.171	13.984	1.581
19	3601.01	06013360101	3901776	964.150	1.860706	1.677125	0.070528	1043032.250000	257.739	479.576	432.260	18.178
34	3601.02	06013360102	6184931	1528.330	1.029228	1.013525	0.081789	3400796.500000	840.355	864.917	851.721	68.732
37	3602	06013360200	2247279	555.315	3.455696	3.163970	0.367359	1097455.030000	271.187	937.140	858.028	99.623
12	3610	06013361000	15935381	3937.719	0.466768	0.421310	0.037585	14331217.500000	3541.321	1652.975	1491.994	133.101

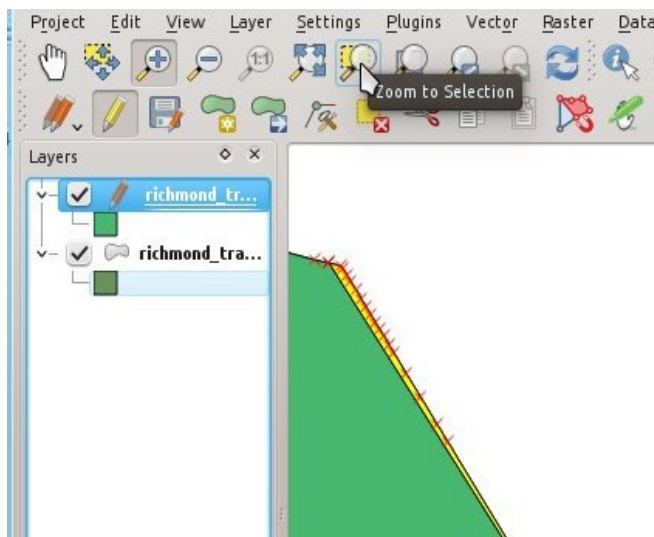
## 8. Clean up the clipped shapefile.

For this step, I am going to be a little conservative and “Save As” this shapefile as a separate copy that I can go to clean up. I will keep both shapefiles loaded so that I can compare them and make sure I have not cleaned off too much.



Then, using the Attribute Table (below), I seek out any tiny scrap polygons that have no area and effectively no data. Record #4 has an area of 0.000003 square meters; that is a sliver-artifact that should just be deleted.

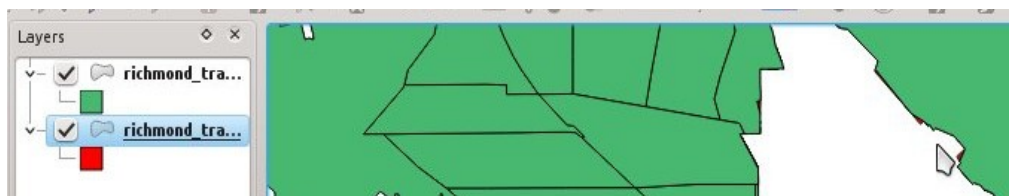
	NAME	_sqm	area_acre	du_ac	ocdu_ac	mvin10_ac	r_area_sqm
2	3870	802032	198.187	5.868195	5.474627	0.565123	178.071593
3	3660.01	06013366001	1011113	5.287130	4.794839	0.400237	1159.219280
4	3680.02	06013368002	668285	5.764910	5.262297	0.775114	0.000003
5	3740	06013374000	994035	7.946847	7.165190	0.846795	994035.398000
6	3820	06013382000	1942415	5.881483	5.185622	1.050042	1940511.110000



When I find a record that might be a candidate for deletion, I can also visually inspect it with the **Zoom to selection** tool. Here I can see that the sliver is the artifact of a slight mismatch between the Census Tract and city-boundary polygons.

There were only about 50 entities in the shapefile, so I could scroll up and down the calculated **area** Attribute to remove the “shavings” from the clipping process. Since the Tract file and the City file were both in the same datum/projection, and both shapefiles came from the Census Bureau, I am going to assume that these little variations are not caused by disagreements between my data-sources.

The last part of the cleaning-process is to change the initial clip-layer to bright red, and do a visual inspection of the overlying cleaned file. In the screenshot below I find two little areas that poke out (near the cursor), and they are about the size of a shopping-cart. Once I am satisfied that I have a properly clean shapefile I unload the previous-step clip file and move on with analysis.

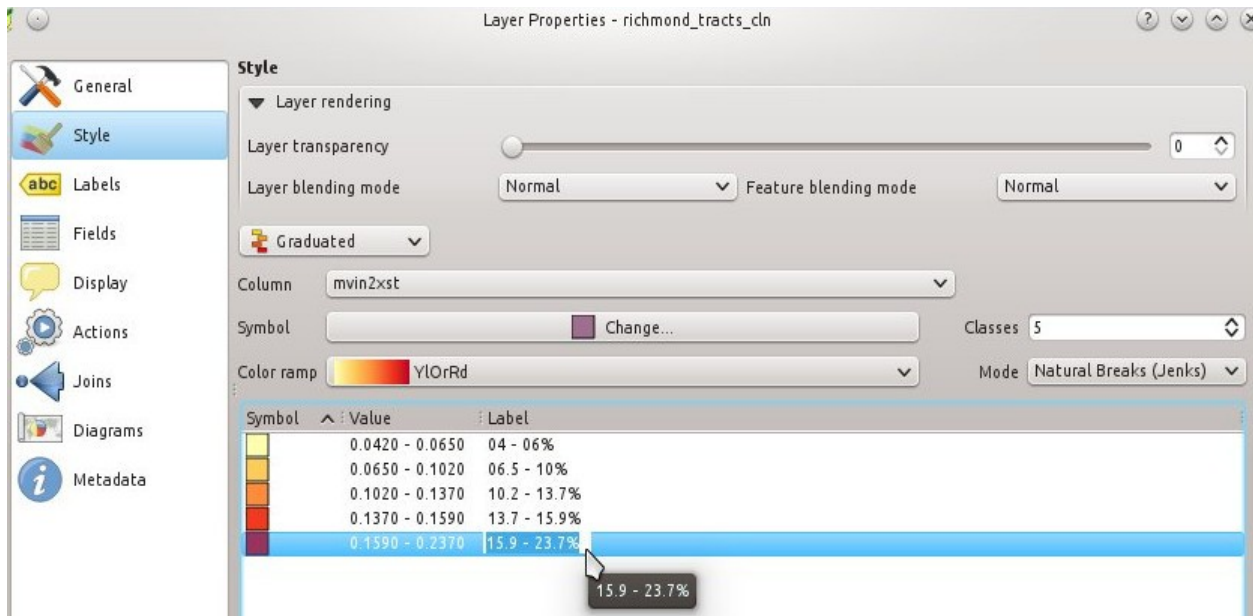


## 9. Create a choropleth map

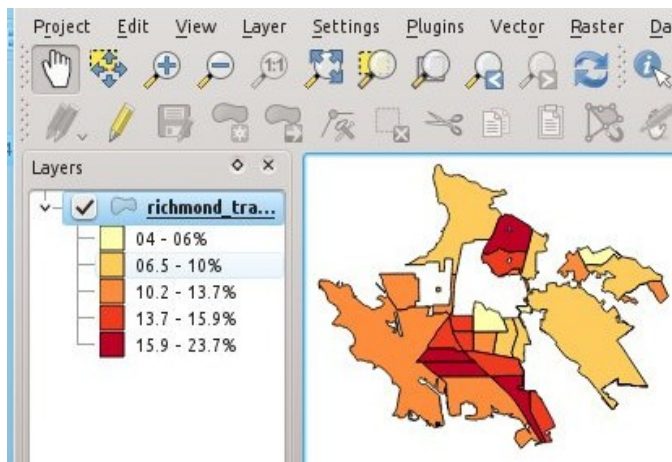
Here I am going to create a choropleth map of the relative density of recent move-ins. For this analysis I use the Field Calculator to generate another attribute: “mvin2xst”. The equation is:

$$\text{“r\_mvin2010”} / \text{“r\_occ\_du”}$$

Then I go **Layer properties > Style** (the **Symbology** tab in ArcMAP) and select **Graduated Symbols**:



I manually retype the data into more human-readable form, because the **Label** field is the one that will be used for final displays of this map.



Finally, I have created a choropleth map of Richmond, showing the proportion of households that moved in after 2010.

If I want to add any more Census data to this shapefile, I will need to step back to the Contra Costa shapefile, join tract-level data to that, calculate densities with that whole file, and then use the Richmond City boundary file to clip.

Or, if I know that I will be doing that repeatedly, I can set up a “clip-ratio” field where I calculate the proportion of each county tract that lies in Richmond, and then calculate fractions directly.