Spatial Clustering Tutorial using GeoDa:

A Training Module for the CDC/ATSDR Guidelines for Examining Unusual Patterns of Cancer and Environmental Concerns

Table of Contents

Background

GeoDa is an open-source software package that was first introduced by Dr. Luc Anselin in 2003. GeoDa provides a user-friendly graphical interface for analyzing spatial patterns in point and polygon data.

GeoDa allows users to interactively explore relationships between maps and statistical plots and identify patterns of spatial clustering and hotspots. It supplements traditional GIS and statistical software by focusing on understanding geographic phenomena, through exploratory spatial data analysis, geo-visualization, and spatial modeling. Some key functionalities of GeoDa include spatial autocorrelation statistics (global and local), basic linear regression, and spatial regression models (e.g., lag and error models).

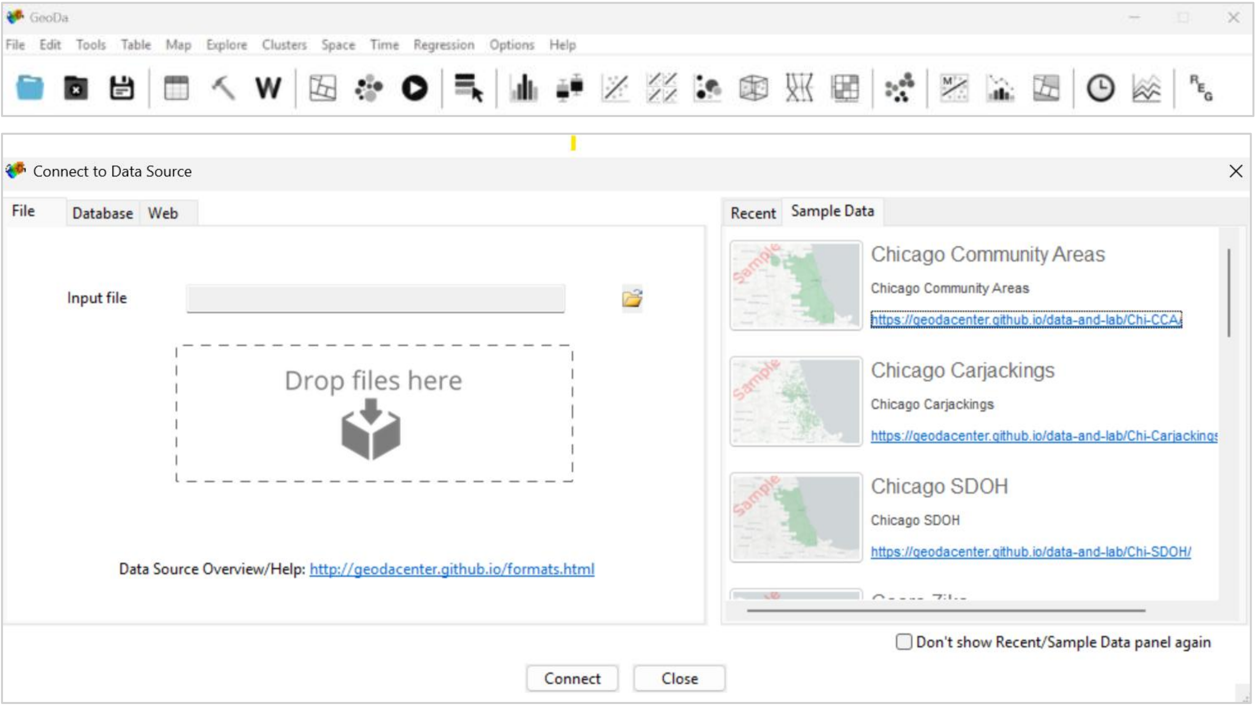
This tutorial will focus on using GeoDa to explore prostate cancer cases among 18 year or older in Pennsylvania's 3,218 census tracts for 2010 to 2019 from CDC's National Environmental Public Health Tracking Network Data Explorer Tool. We also used population data from the U.S. Census Bureau to get information about the age distribution of the male population in each tract. These data are all provided in the **PAcancerUpdated.xlsx** file in the tutorial folder. This analysis is only intended as a demonstration of how to use GeoDa for cancer cluster investigations and the sample results and findings presented as part of this tutorial should not be interpreted as real-world conclusions.

Launch GeoDa

1. Launch the GeoDa software by double-clicking on the **GeoDa shortcut** on your Desktop.

   ***Note**: If you do not see the GeoDa app icon, search for the application on your computer or navigate to the GeoDa folder where the software was downloaded and open the GeoDa.exe file.*

2. GeoDa will open in 2 windows: the **toolbar** and the **Connect to Data Source** window.


GeoDa
File Edit Tools Table Map Explore Clusters Space Time Regression Options Help
Connect to Data Source
File
Database
Web
Input file
Drop files here
Recent
Sample Data
Chicago Community Areas
Chicago Community Areas
https://geodacenter.github.io/data-and-lab/Chi-CCA/
Chicago Carjackings
Chicago Carjackings
https://geodacenter.github.io/data-and-lab/Chi-Carjackings
Chicago SDOH
Chicago SDOH
https://geodacenter.github.io/data-and-lab/Chi-SDOH/
□ Don't show Recent/Sample Data panel again
Connect
Close

Data Source Overview/Help: http://geodacenter.github.io/formats.html

Import Data

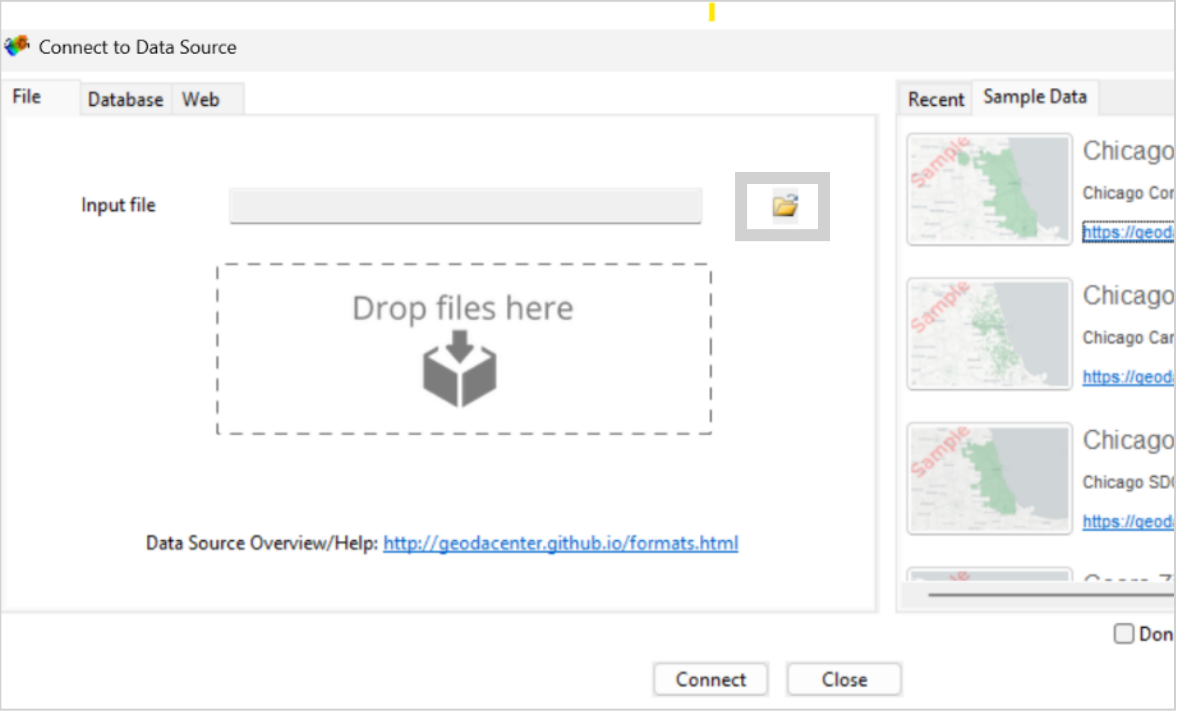
The first pane of the **Connect to Data Source** window includes 3 main tabs that you can use to import or connect to your data:

- **File tab –** lets you import data from your computer
- **Database tab –** helps you connect to a database where your data is stored
- **Web tab –** allows you to connect to data that you have a URL for (e.g., GeoJson URL, WFS URL)

The second pane on the right, is known as the **Recent/Sample Data Panel**. This panel has 2 tabs:

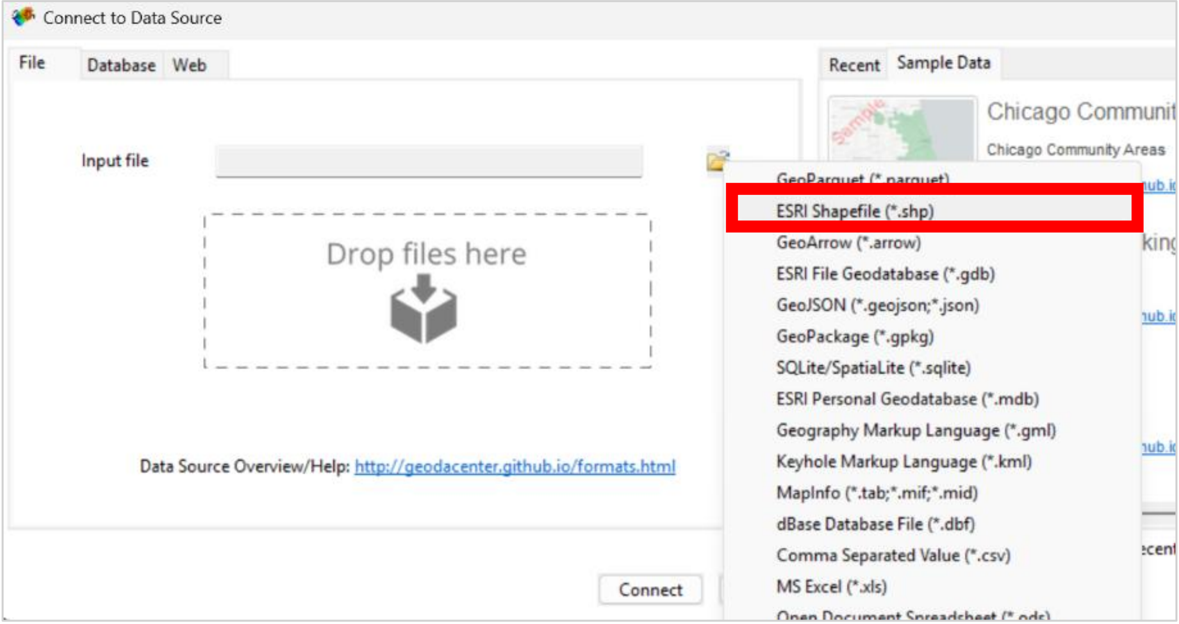
- **Recent Data tab** – allows you to quickly find and add data from a list of files that you recently opened in GeoDa
- **Sample Data tab** – includes sample datasets that you can use to test out different features in GeoDa

1. Open the prostate cancer data in GeoDa by first selecting the **folder** icon in the **File** tab of the **Connect to Data Source** window.

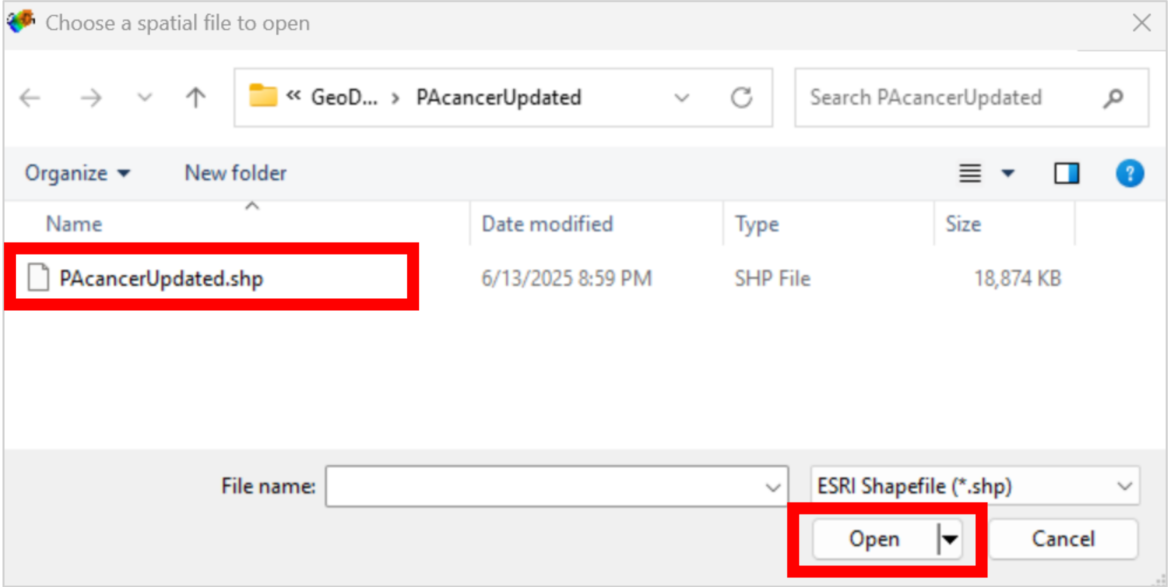


2. There are several types of files we can import to GeoDa. We will be using a shapefile for this analysis, so select "**ESRI Shapefile (*.shp)**" from the drop-down menu.

***Note**: Make sure that your tutorial data folder is unzipped before trying to import your data, so you are able to find and import your data.*


Connect to Data Source
File
Database
Web
Input file
Drop files here
Data Source Overview/Help: http://geodacenter.github.io/formats.html
Recent
Sample Data
Chicago Communit
Chicago Community Areas
GeoParquet (*.parquet)
ESRI Shapefile (*.shp)
GeoArrow (*.arrow)
ESRI File Geodatabase (*.gdb)
GeoJSON (*.geojson;*.json)
GeoPackage (*.gpkg)
SQLite/SpatiaLite (*.sqlite)
ESRI Personal Geodatabase (*.mdb)
Geography Markup Language (*.gml)
Keyhole Markup Language (*.kml)
MapInfo (*.tab;*.mif;*.mid)
dBase Database File (*.dbf)
Comma Separated Value (*.csv)
MS Excel (*.xls)
Open Document Spreadsheet (*.ods)
Connect

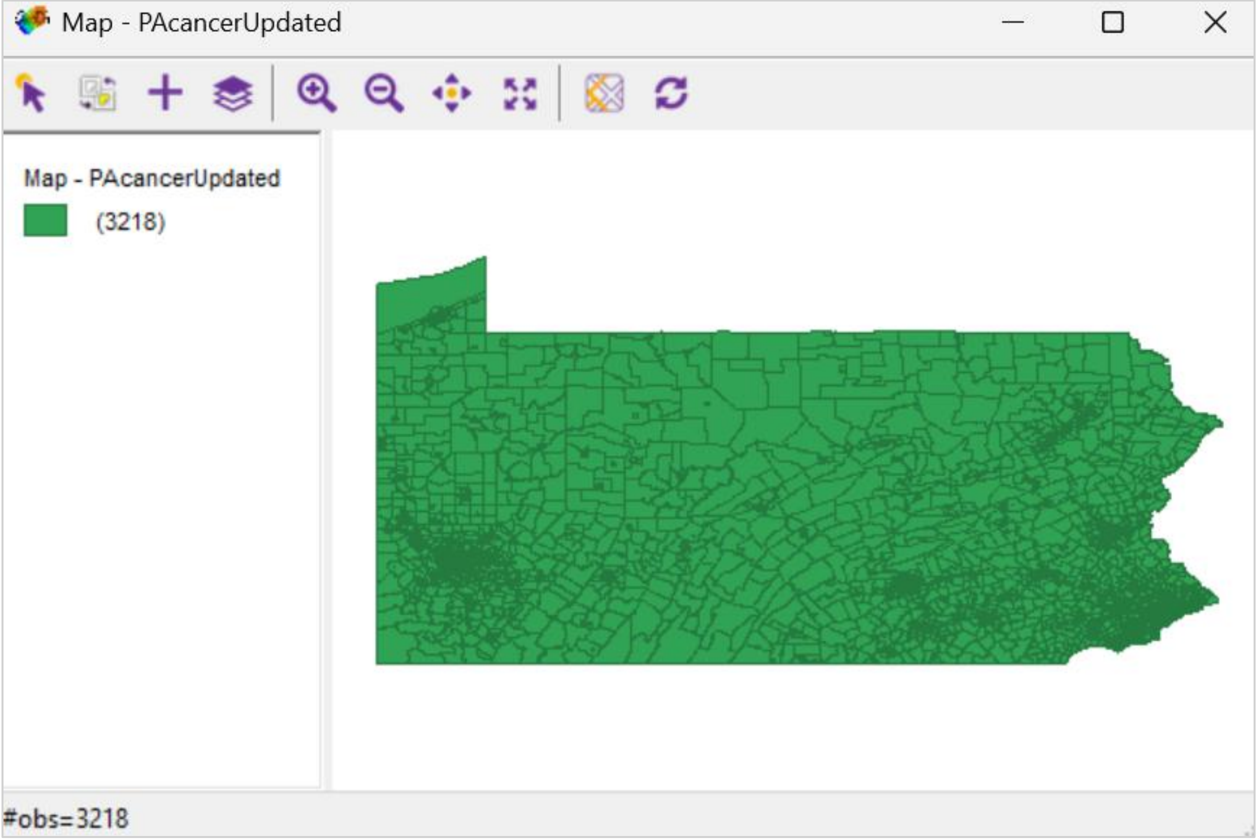
3. Next, navigate to where your tutorial data is stored and select the shapefile called "**PAcancerUpdated.shp**" from the folder. Use the "**Open**" button (in the bottom right) to finish importing it into GeoDa.



***Note:** You can also drag and drop your data into GeoDa, by dragging the file from your computer into the "**Drop files here**" box in the **Connect to Data Source** window. Once you have imported your data into GeoDa, you can quickly access the file again from the **Recent Data** tab in the **Recent/Sample Data Panel**.*

Explore & Visualize the Data

After we have imported our shapefile into GeoDa, we will be able to see the map of the **3,218 PA census tracts** below.

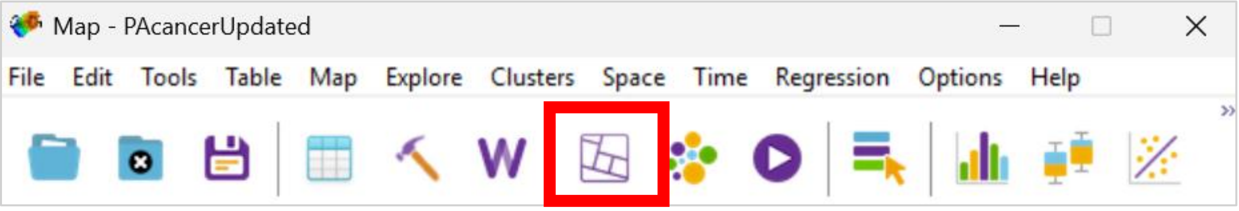


GeoDa includes several mapping tools that we can use to help explore and visualize our data before we begin our analysis. For this tutorial, we'll look at how to **create a choropleth map** of the data, **customize our map**, and **explore the data** further in a data table.

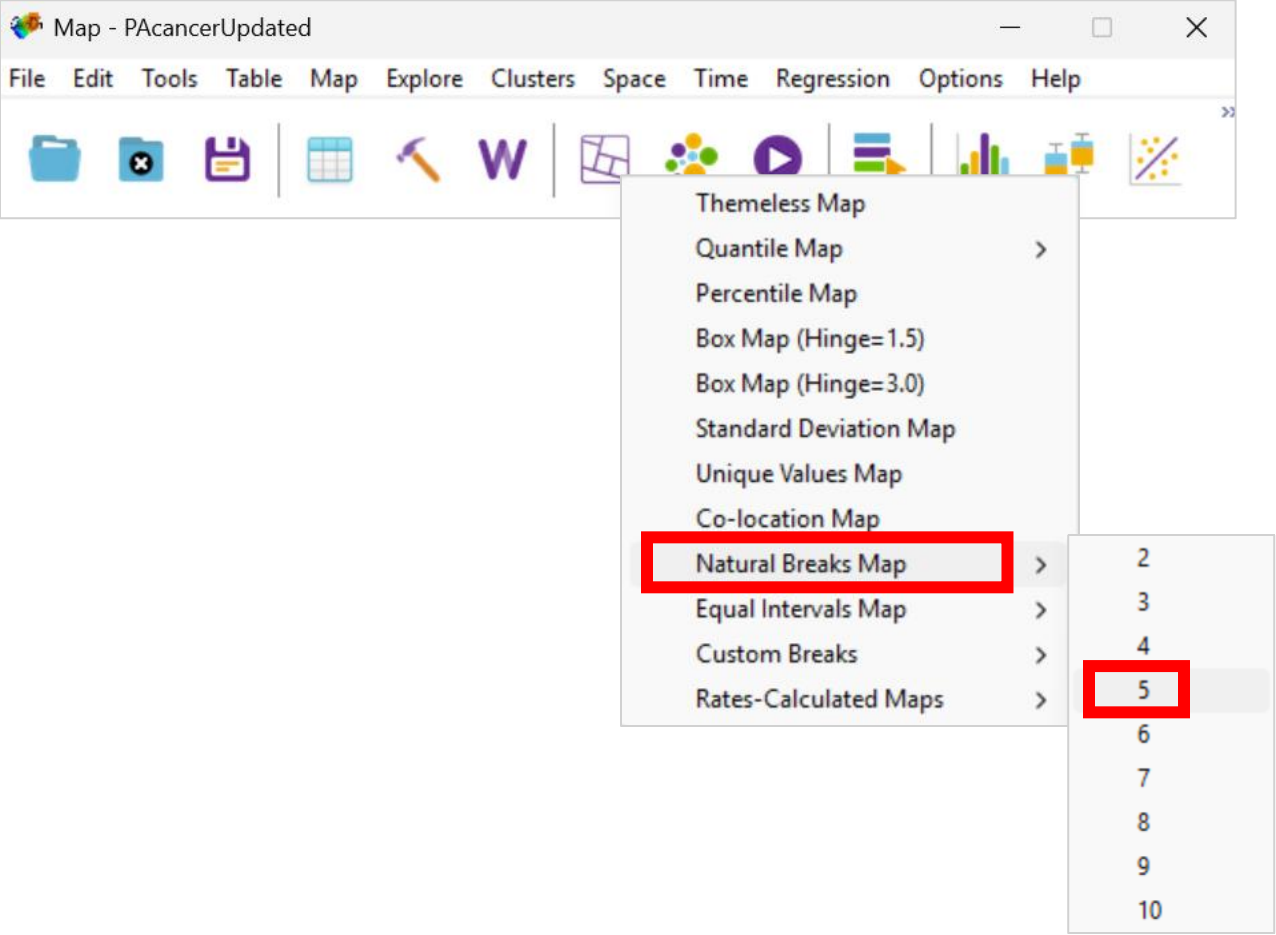
There are many types of choropleth maps that we can create in GeoDa, including natural breaks, quartile, and equal interval maps. For this tutorial, we will be creating a **natural breaks** map, which groups data based on natural cut points (i.e., **breaks**) in the data that help **maximize differences between groups**, while **minimizing differences within each group**. Natural breaks are often used for choropleth maps as they aim to represent the data's true distribution – making for a more accurate and easily understandable visualization, especially for highlighting clusters within the data.

Create a Choropleth Map

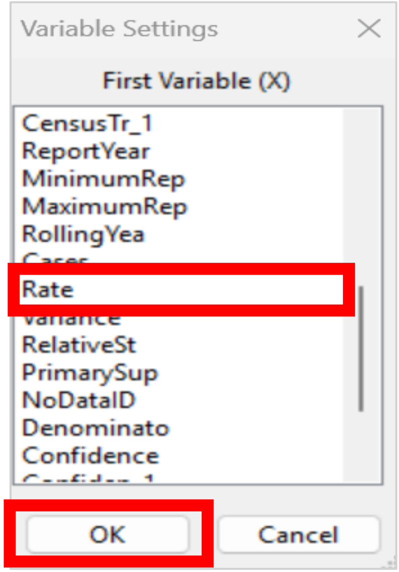
1. We'll start by creating a choropleth map of the data. First, select the **Map** icon from the **GeoDa Toolbar**.



2. Then, select "**Natural Breaks Map**" from the menu and specify that GeoDa should use "**5**" categories (i.e., breaks).



3. The **Variable Settings** window (to the right) will appear, to let us select the variable that we would like to visualize in our map. For this tutorial, let's view prostate cancer rates, by scrolling down to select the [**Rate**] variable from the menu. Next, select the "**OK**" button (in the bottom left) to create the map.

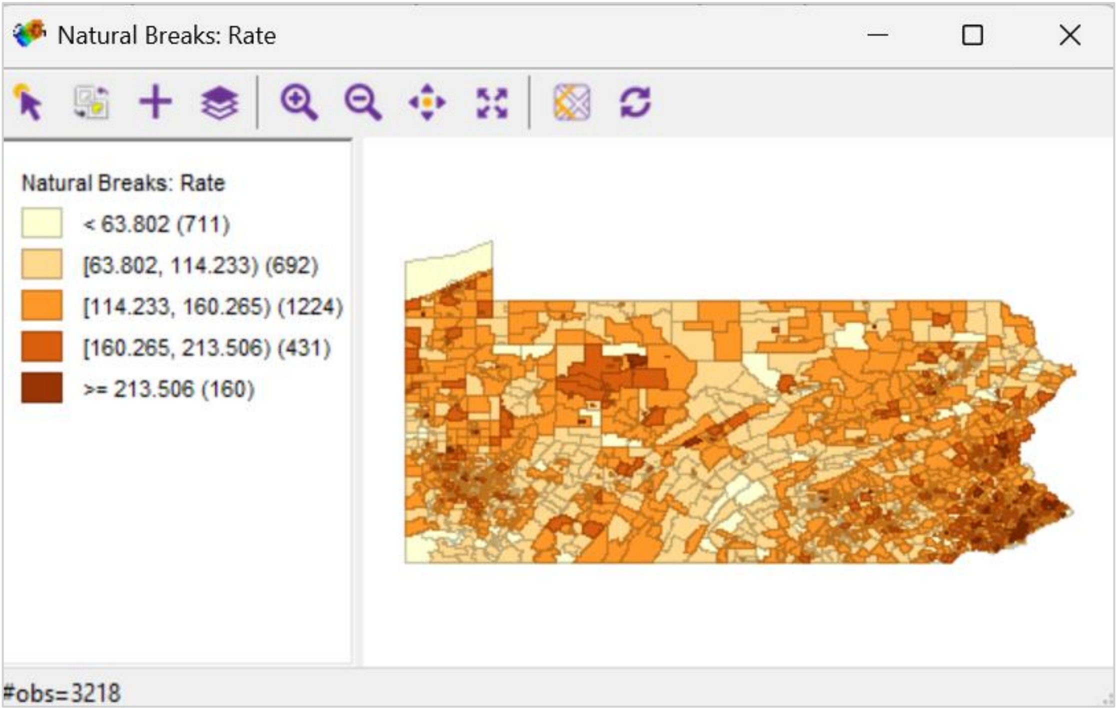


Variable Settings

First Variable (X)

CensusTr_1
ReportYear
MinimumRep
MaximumRep
RollingYea
Cases
Rate
Variance
RelativeSt
PrimarySup
NoDataID
Denominato
Confidence
Confiden_1

GeoDa will open a new window with our new choropleth map.



Natural Breaks: Rate

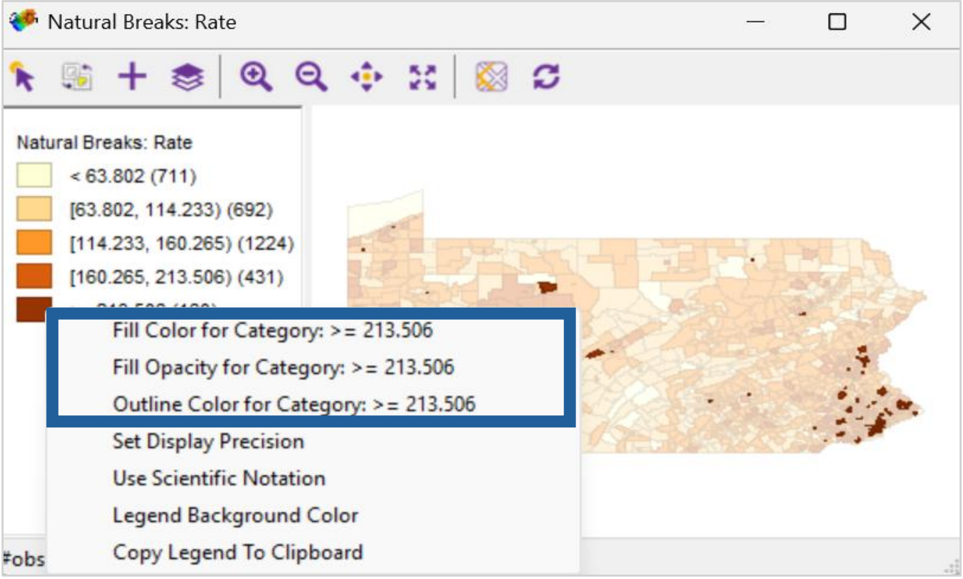
Natural Breaks: Rate
- < 63.802 (711)
- [63.802, 114.233) (692)
- [114.233, 160.265) (1224)
- [160.265, 213.506) (431)
- >= 213.506 (160)

#obs=3218

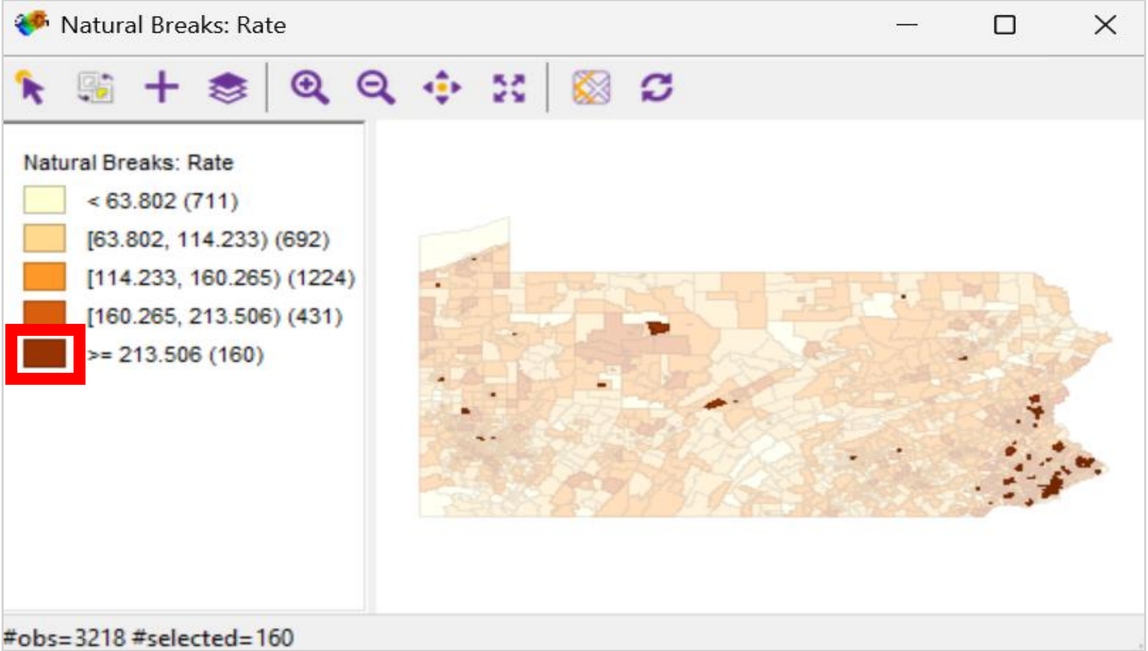
Customize the Map

There are a few ways that we can customize our choropleth map, including using different **map colors** or **highlighting specific categories** (e.g., highest rates) on the map.

1. To choose new map colors, right click a color in the legend. A menu will appear to allow you to choose a new **fill color, fill opacity, or outline color**. For this tutorial, we will keep the default colors.

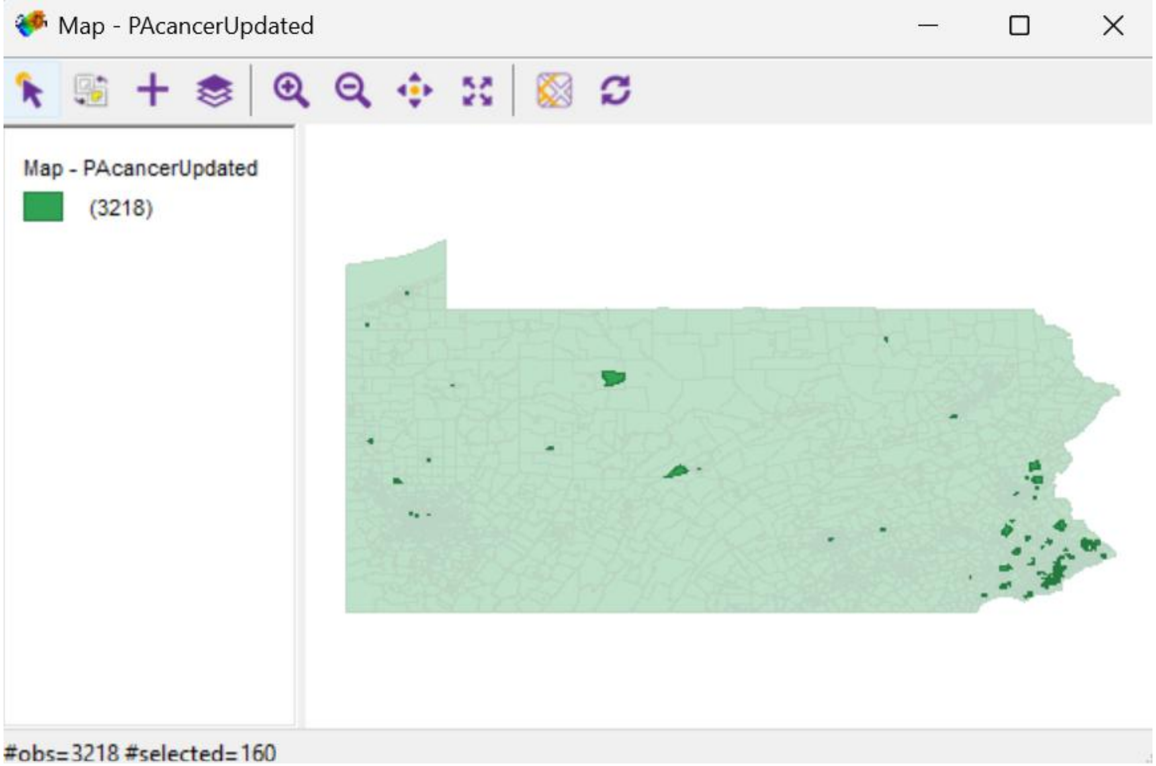


2. Next, let’s learn how to highlight a specific category on the map. Click on the **legend color box** for the **highest rates** to see the **160 tracts** with the highest prostate cancer rates highlighted on the map.



#obs=3218 #selected=160

Highlighting tracts on this map, will also highlight them in all of the other windows we have open, such as the map that opened when we imported our data, which now highlights those tracts as well.

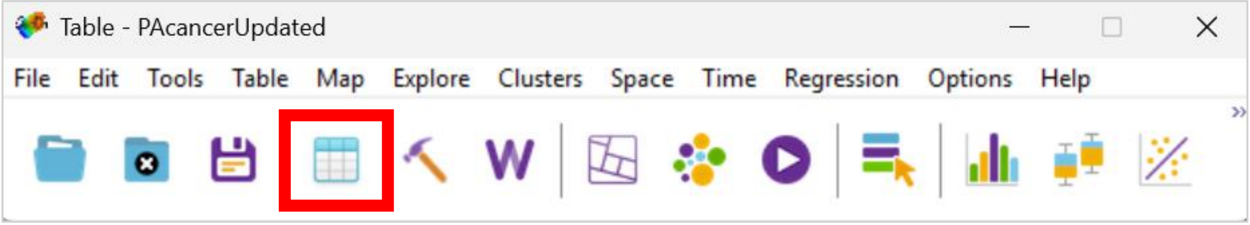


#obs=3218 #selected=160

Explore the Data Table

We can further explore the data by opening it up in a **Data Table**.

1. Leaving the tracts with the highest cancer rates selected in our choropleth map, select the **Table** icon to open the data table.



2. The resulting table has **49 variables** for each of the **3,218 census tracts**, which can be seen using the scroll bars along the bottom and on the right side of the table. We can also see the total number of records (i.e., census tracts) and the **160 tracts** with the highest rates of prostate cancer are still selected in the bottom left of the table.

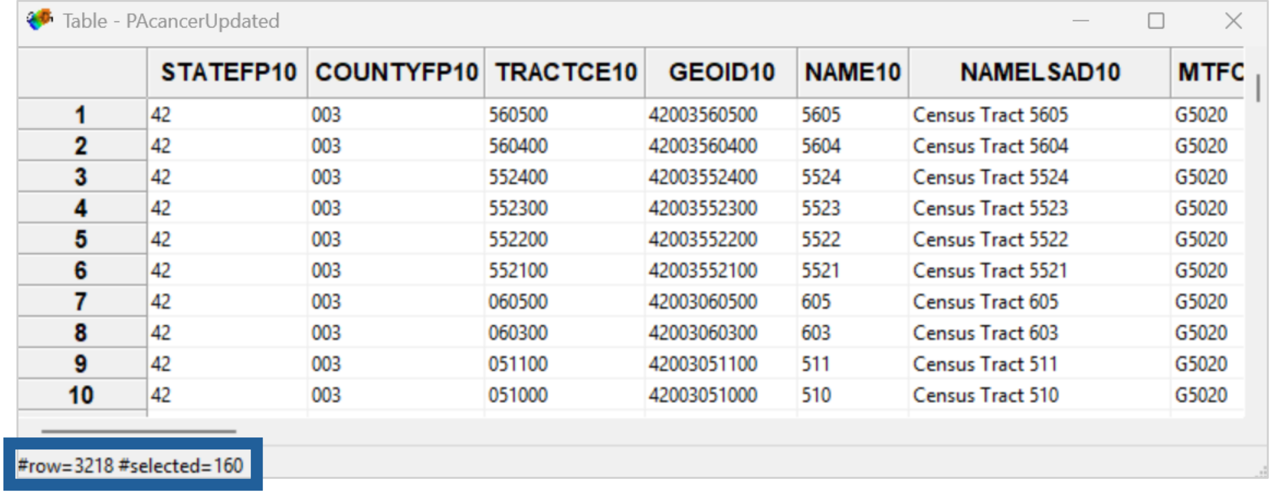


Table - PAcancerUpdated

	STATEFP10	COUNTYFP10	TRACTCE10	GEOID10	NAME10	NAMELSAD10	MTFC
1	42	003	560500	42003560500	5605	Census Tract 5605	G5020
2	42	003	560400	42003560400	5604	Census Tract 5604	G5020
3	42	003	552400	42003552400	5524	Census Tract 5524	G5020
4	42	003	552300	42003552300	5523	Census Tract 5523	G5020
5	42	003	552200	42003552200	5522	Census Tract 5522	G5020
6	42	003	552100	42003552100	5521	Census Tract 5521	G5020
7	42	003	060500	42003060500	605	Census Tract 605	G5020
8	42	003	060300	42003060300	603	Census Tract 603	G5020
9	42	003	051100	42003051100	511	Census Tract 511	G5020
10	42	003	051000	42003051000	510	Census Tract 510	G5020

#row=3218 #selected=160

3. To make it easier to explore the data for the highlighted tracts, right click one of the columns at the top of the table, such as the [**STATEFP10**] column, and select the option to “**Move Selected to Top**”.

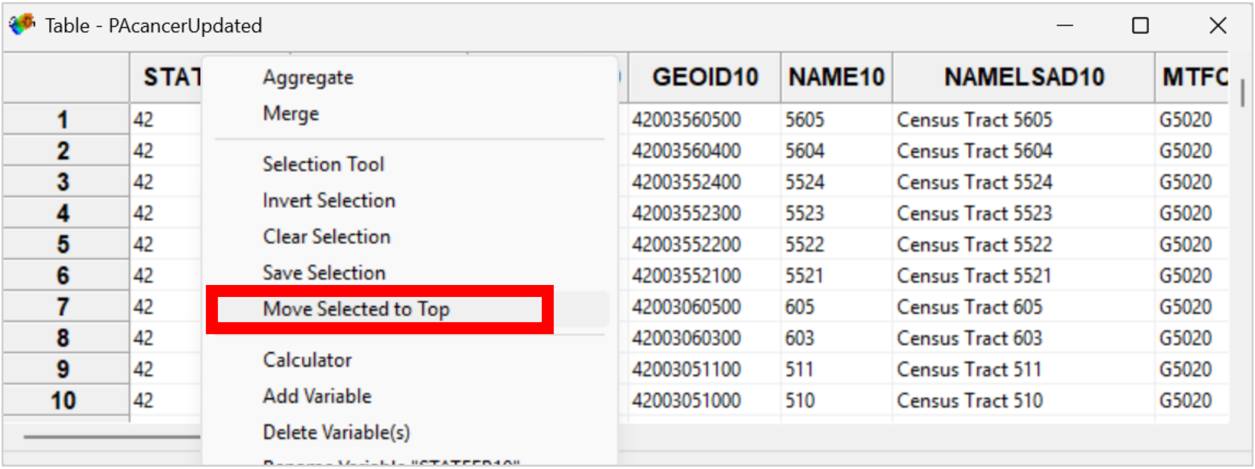


Table - PAcancerUpdated

	STAT	GEOID10	NAME10	NAMELSAD10	MTFC
1	42	42003560500	5605	Census Tract 5605	G5020
2	42	42003560400	5604	Census Tract 5604	G5020
3	42	42003552400	5524	Census Tract 5524	G5020
4	42	42003552300	5523	Census Tract 5523	G5020
5	42	42003552200	5522	Census Tract 5522	G5020
6	42	42003552100	5521	Census Tract 5521	G5020
7	42	42003060500	605	Census Tract 605	G5020
8	42	42003060300	603	Census Tract 603	G5020
9	42	42003051100	511	Census Tract 511	G5020
10	42	42003051000	510	Census Tract 510	G5020

The **160 tracts** with the **highest prostate cancer rates** are now highlighted in yellow at the top of the table.

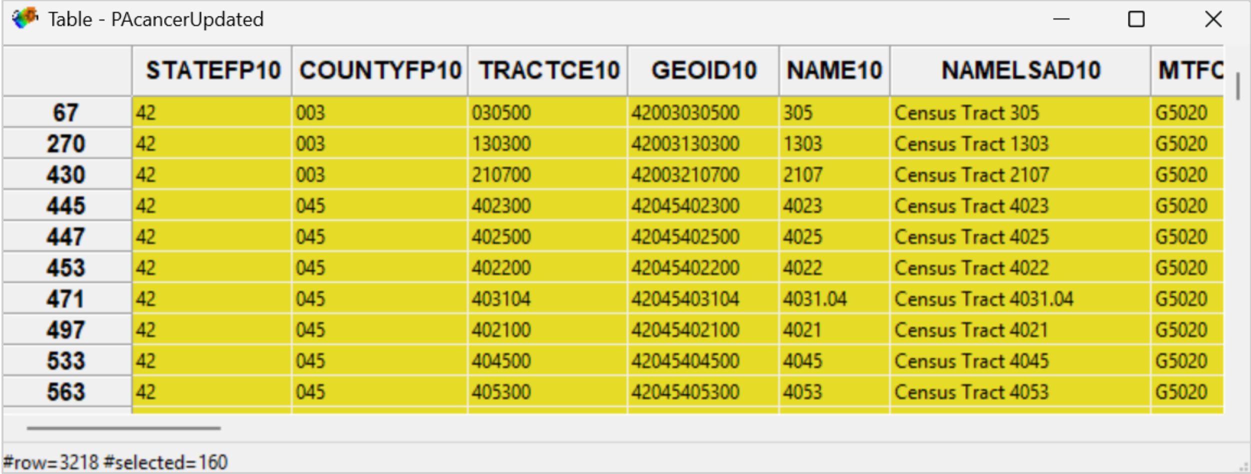


Table - PAcancerUpdated

	STATEFP10	COUNTYFP10	TRACTCE10	GEOID10	NAME10	NAMELSAD10	MTFC
67	42	003	030500	42003030500	305	Census Tract 305	G5020
270	42	003	130300	42003130300	1303	Census Tract 1303	G5020
430	42	003	210700	42003210700	2107	Census Tract 2107	G5020
445	42	045	402300	42045402300	4023	Census Tract 4023	G5020
447	42	045	402500	42045402500	4025	Census Tract 4025	G5020
453	42	045	402200	42045402200	4022	Census Tract 4022	G5020
471	42	045	403104	42045403104	4031.04	Census Tract 4031.04	G5020
497	42	045	402100	42045402100	4021	Census Tract 4021	G5020
533	42	045	404500	42045404500	4045	Census Tract 4045	G5020
563	42	045	405300	42045405300	4053	Census Tract 4053	G5020

#row=3218 #selected=160

Spatial Autocorrelation Analysis

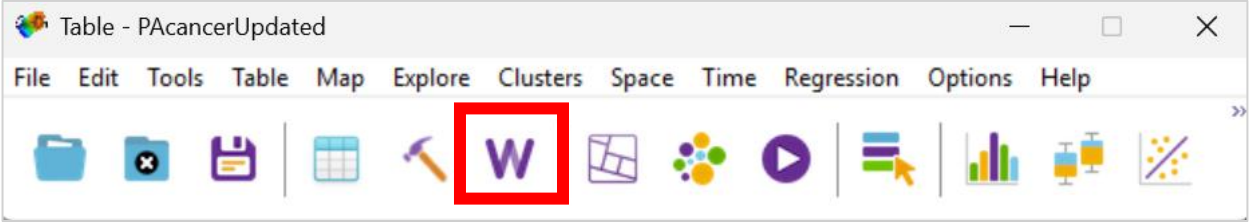
For our spatial autocorrelation analysis today, we will be creating a **spatial weights** file and then using it to run our **Global Moran's I** and **Local Moran's I** analyses.

Add Spatial Weights File

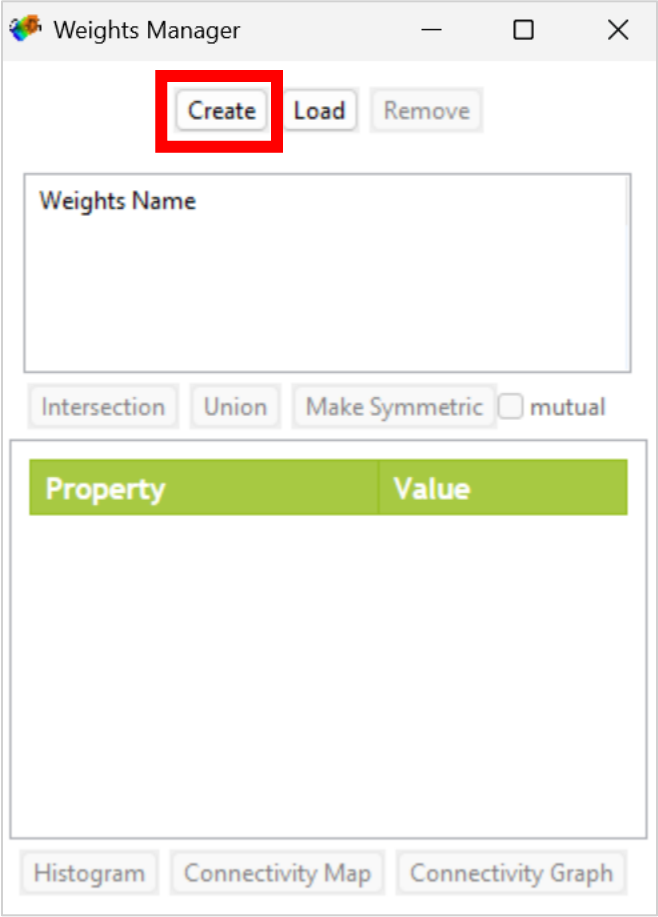
Before we can run our spatial autocorrelation analysis, we need to create or import a **spatial weights** file, to tell GeoDa how to compare neighboring tracts. Spatial weights are different than what you might be used to using (e.g., weights in a survey analysis), in that spatial weights are used to characterize the relationships between geographies (e.g., census tracts), rather than to adjust the data like survey weights.

For this analysis, we will be creating our own spatial weights file that uses **Queen Contiguity** as our weight measurement, which will look for neighboring tracts that share a common edge (e.g., sharing a census tract boundary) or vertex (e.g., tract boundaries that meet at a single point or corner). You may also be interested in Rook Contiguity when conducting your own analysis, which only identifies tracts as neighbors if they share a boundary or border.

1. First, select the **Weights Manager** icon from the toolbar to create or load a spatial weights file.



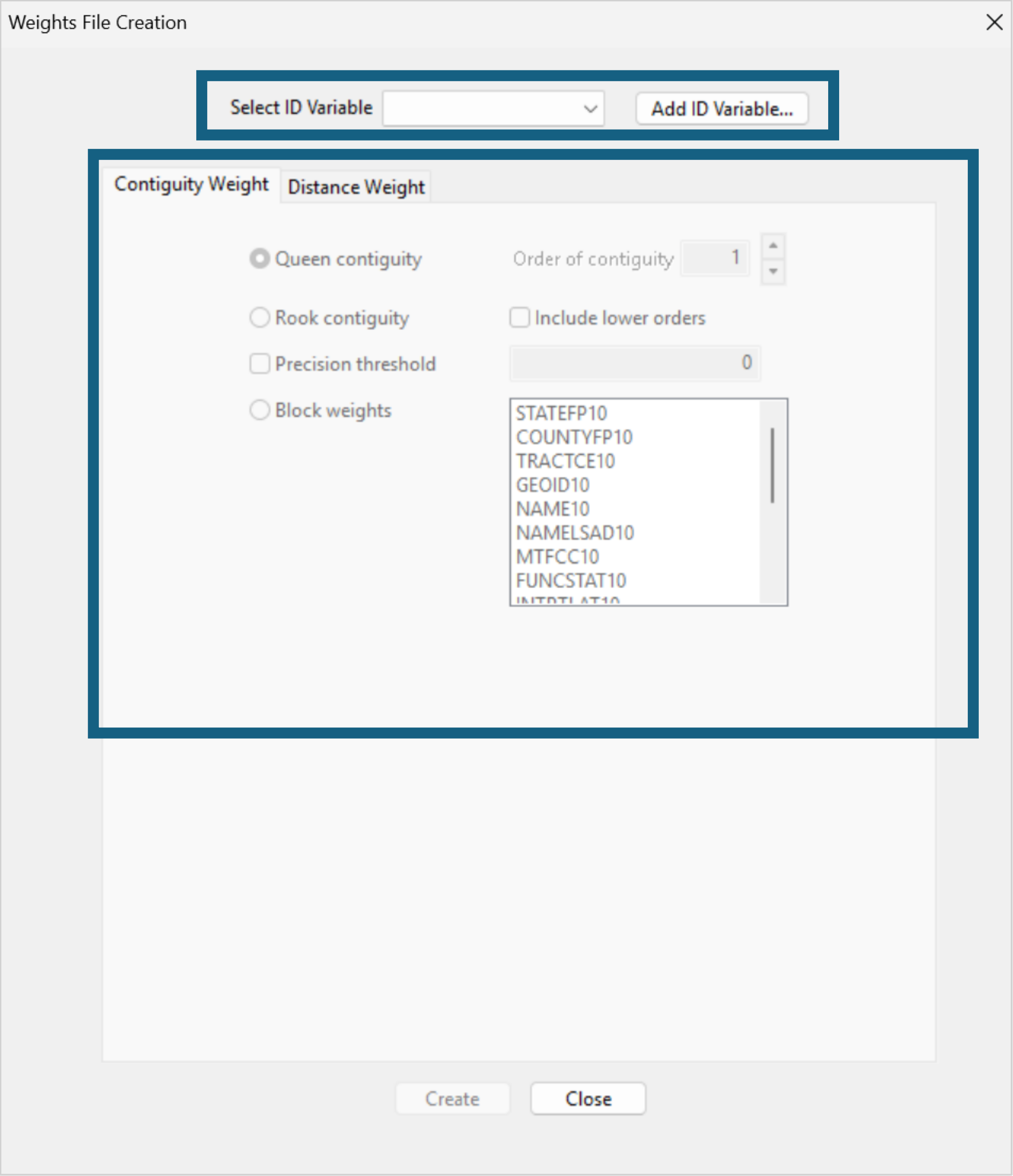
2. Next, select "**Create**" from the top left of the **Weights Manager** menu.



Note: *If you already have a weights file, you can select the "Load" button to import your weights file into GeoDa instead.*

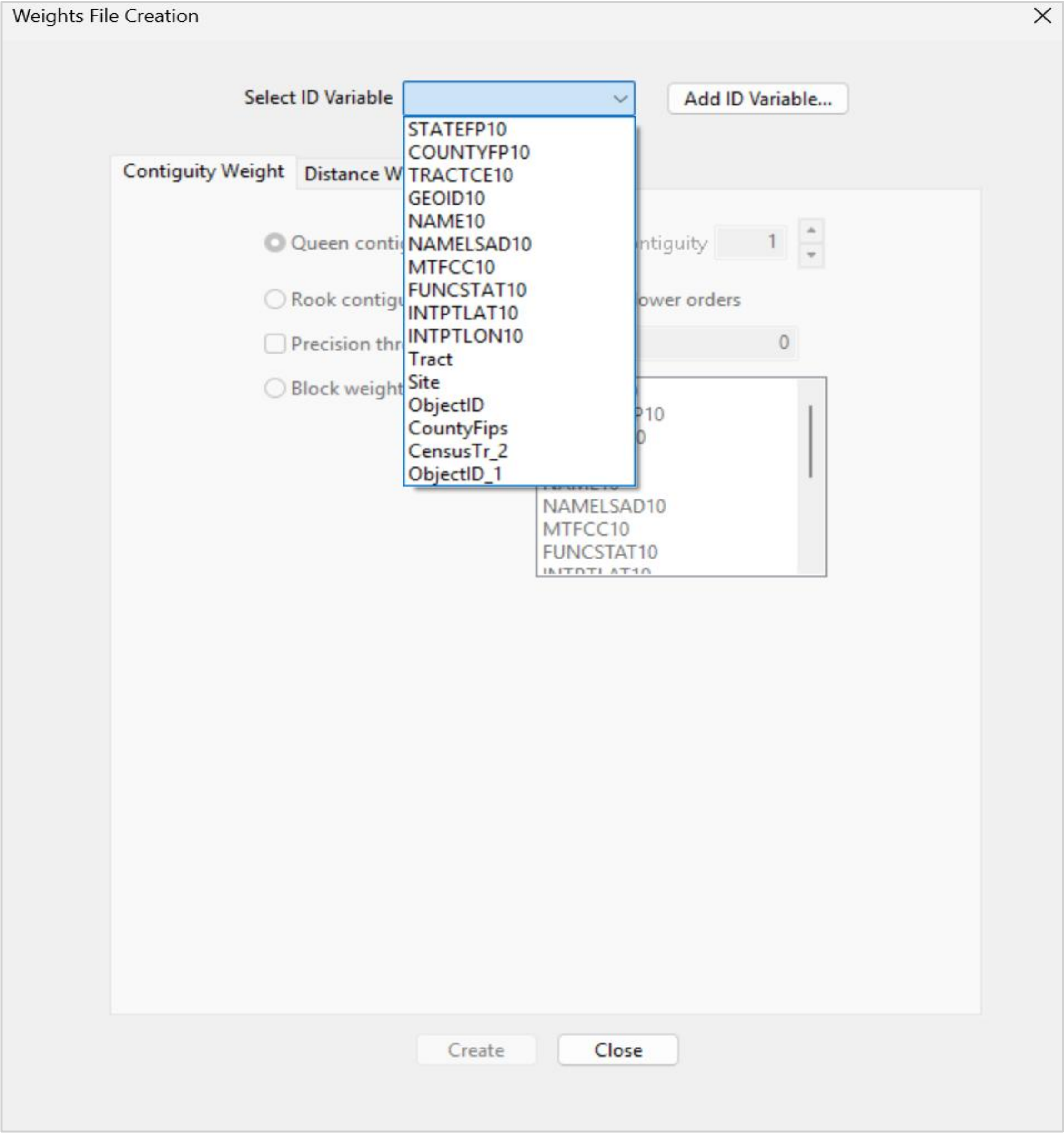
The following window will appear to allow us to select our **ID Variable (i.e., the variable that we use to distinguish unique records)**, as well as our spatial **weights measurement type** in the "**Contiguity Weight**" and "**Distance Weight**" tabs.

***Note**: All options in the "Contiguity Weight" and "Distance Weight" tabs will be grayed out and unable to be selected (like the image below) until after we have selected our ID variable. The "Create" button at the bottom of the screen will also be grayed out until we select our ID Variable.*



2. The **ID Variable** for this analysis is the **census tract identifier**, represented by the [**GEOID10**] variable in our data. Select [**GEOID10**] from the drop-down menu at the top of the screen.

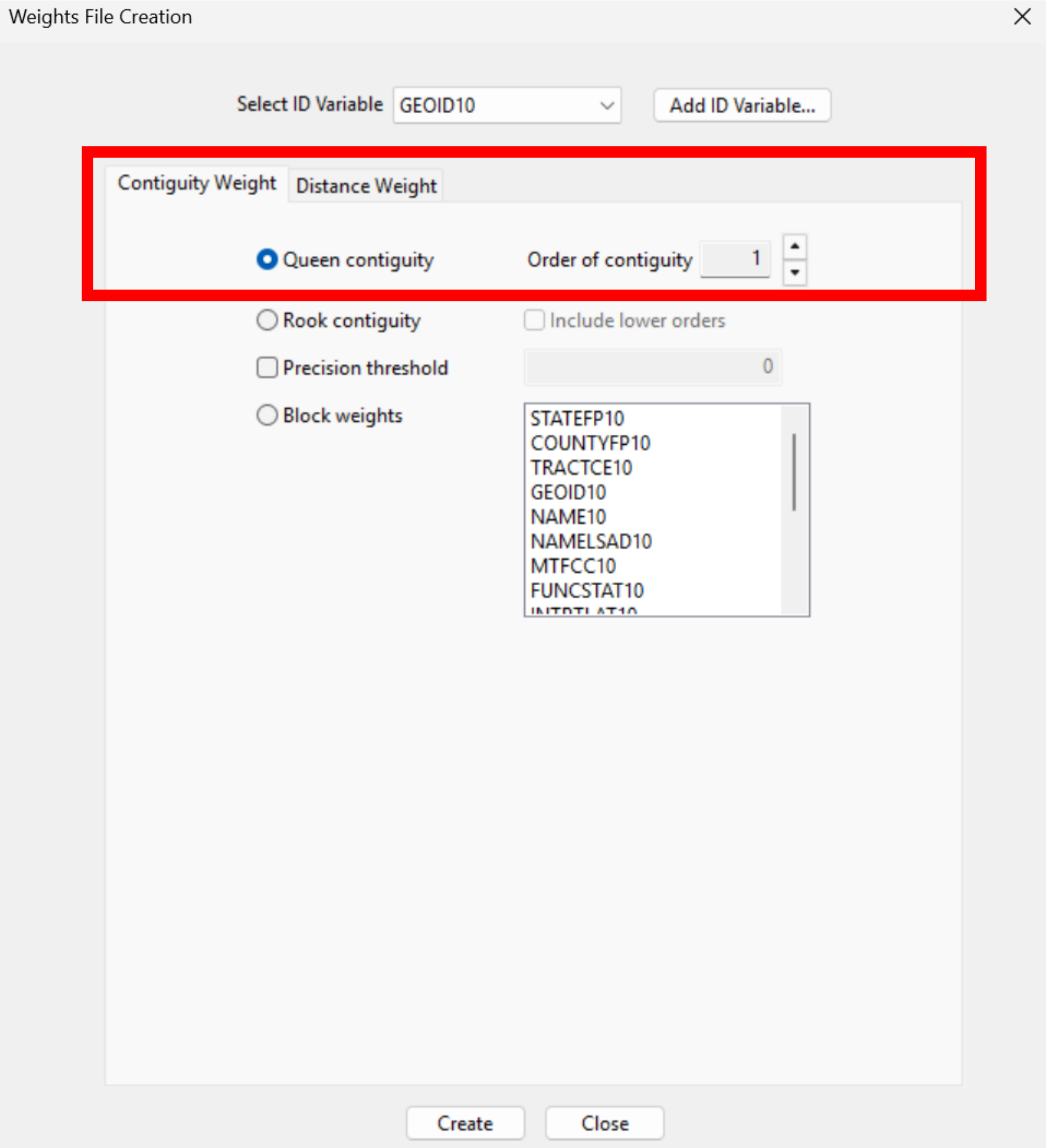
   Note: *The ID Variable can also be set using the "Add ID Variable…" button to the right of the drop-down menu.*



3. Once we have selected our **ID Variable**, we will then be able to select our **spatial weights measurement type**. For this analysis, we are using **Queen Contiguity**. Navigate to the "**Contiguity Weight**" tab and select the "**Queen Contiguity**" option. Next, for **Order of Contiguity**, select **"1"**.

   "Contiguity" defines how neighboring areas are related, and the terms "Rook contiguity" and "Queen contiguity" are inspired by the moves of chess pieces. For a central square in a grid, Rook contiguity would identify the four squares directly adjacent to its top, bottom, left, and right sides. Queen contiguity would identify those four squares plus the four squares diagonally adjacent to its corners. Rook contiguity might be preferred when you want a more restrictive definition where direct adjacency along a border is paramount. Queen contiguity is often the default because it captures a broader sense of spatial interaction, including connections through corners. It's generally more robust for irregularly shaped polygons where a shared corner might still represent a meaningful connection.

   First-Order Contiguity (Order 1) is the most common and often the default. With Queen contiguity, it considers only immediate neighbors – those polygons that share at least one border or vertex with the central polygon. This captures very localized spatial relationships. Higher-Order Contiguity (Order > 1) expands the definition of neighbor. Order 2 includes polygons that are neighbors of neighbors (i.e., polygons that share a border/vertex with a first-order neighbor) and Order 3 includes polygons that are neighbors of those neighbors, and so on. By using higher orders, you can detect spatial patterns that operate over a wider geographic extent, but with increased complexity and computational cost, more difficult interpretation, and diminishing returns.
   For more information on the different spatial weighting options that are available in GeoDa, see the GeoDa spatial weighting tutorial and guidance here: https://geodacenter.github.io/workbook/4a_contig_weights/lab4a.html.



Weights File Creation

Select ID Variable: GEOID10

Contiguity Weight

● Queen contiguity — Order of contiguity: 1

○ Rook contiguity — □ Include lower orders

□ Precision threshold — 0

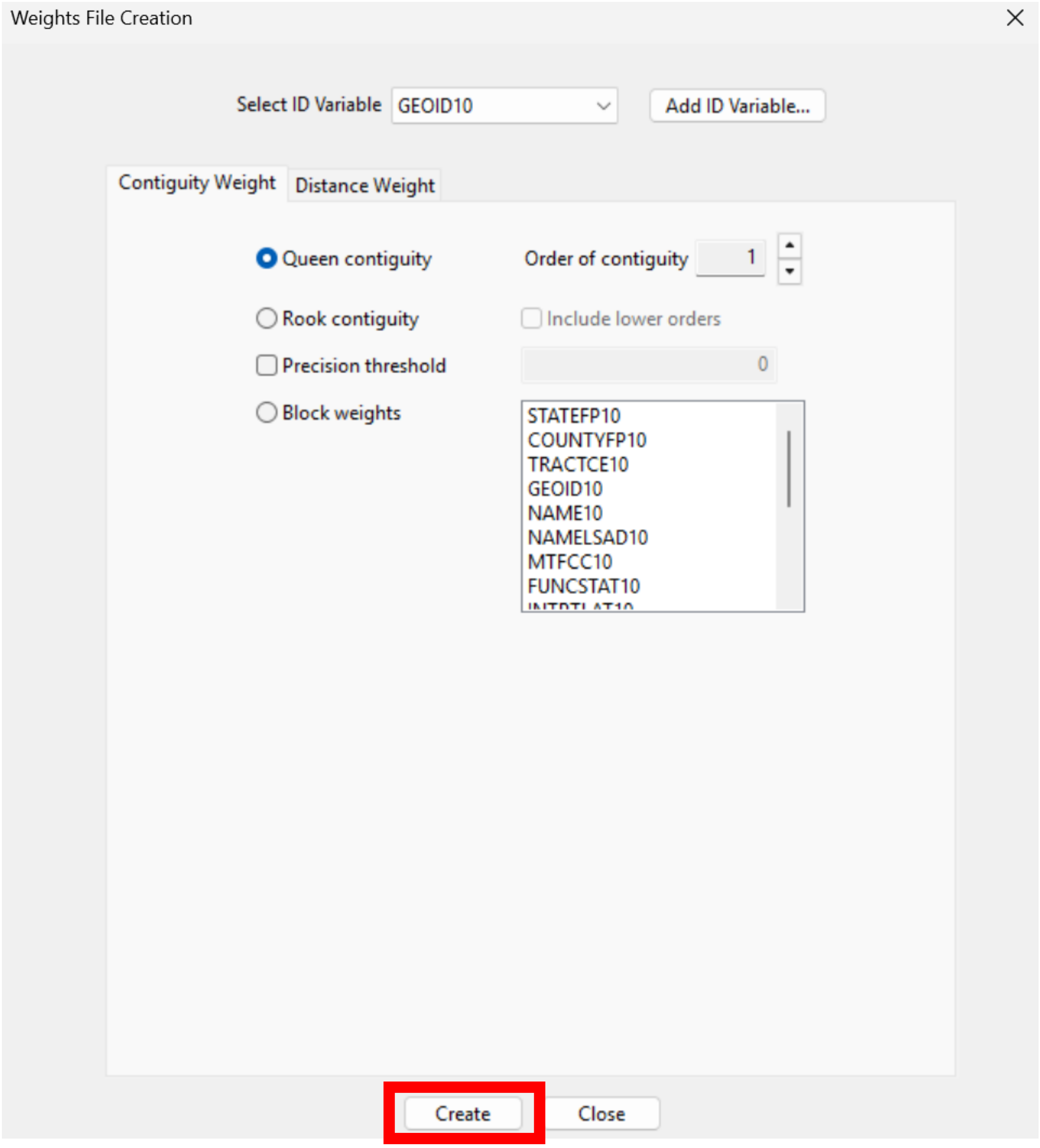
○ Block weights

STATEFP10
COUNTYFP10
TRACTCE10
GEOID10
NAME10
NAMELSAD10
MTFCC10
FUNCSTAT10

***Note**: Some of these options may already be selected by default, but you will be unable to make any changes until selecting your ID Variable in the previous step.*

4. Select the "**Create**" button at the bottom of the screen to create the weights file.

   Note: *The "Create" button will be unable to be selected until after you have set your ID Variable at the top of this window.*



Weights File Creation

Select ID Variable: GEOID10

Contiguity Weight

● Queen contiguity   Order of contiguity: 1

○ Rook contiguity   □ Include lower orders

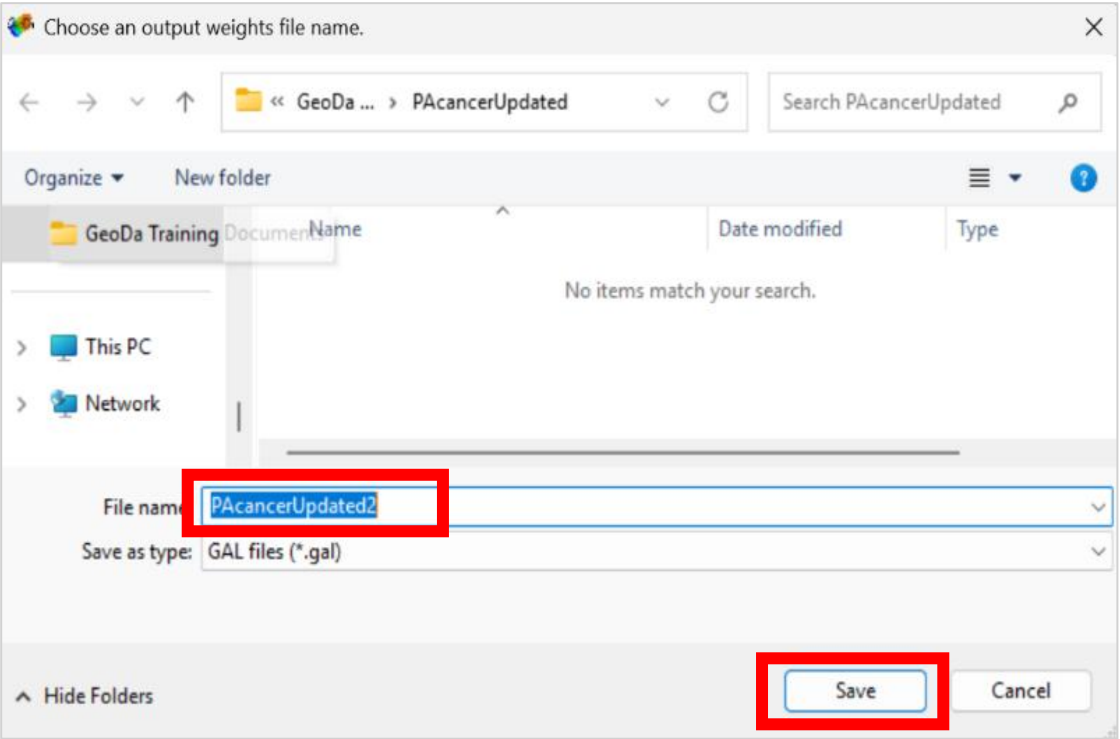
□ Precision threshold   0

○ Block weights

STATEFP10
COUNTYFP10
TRACTCE10
GEOID10
NAME10
NAMELSAD10
MTFCC10
FUNCSTAT10
INTPTLAT10

5. Next, the following window will appear to let you decide what to name your results and where you would like to save them. For this tutorial, name the new weights file "**PAcancerUpdated2.gal**" and select the "**Save**" button (in the bottom right) to save the results.

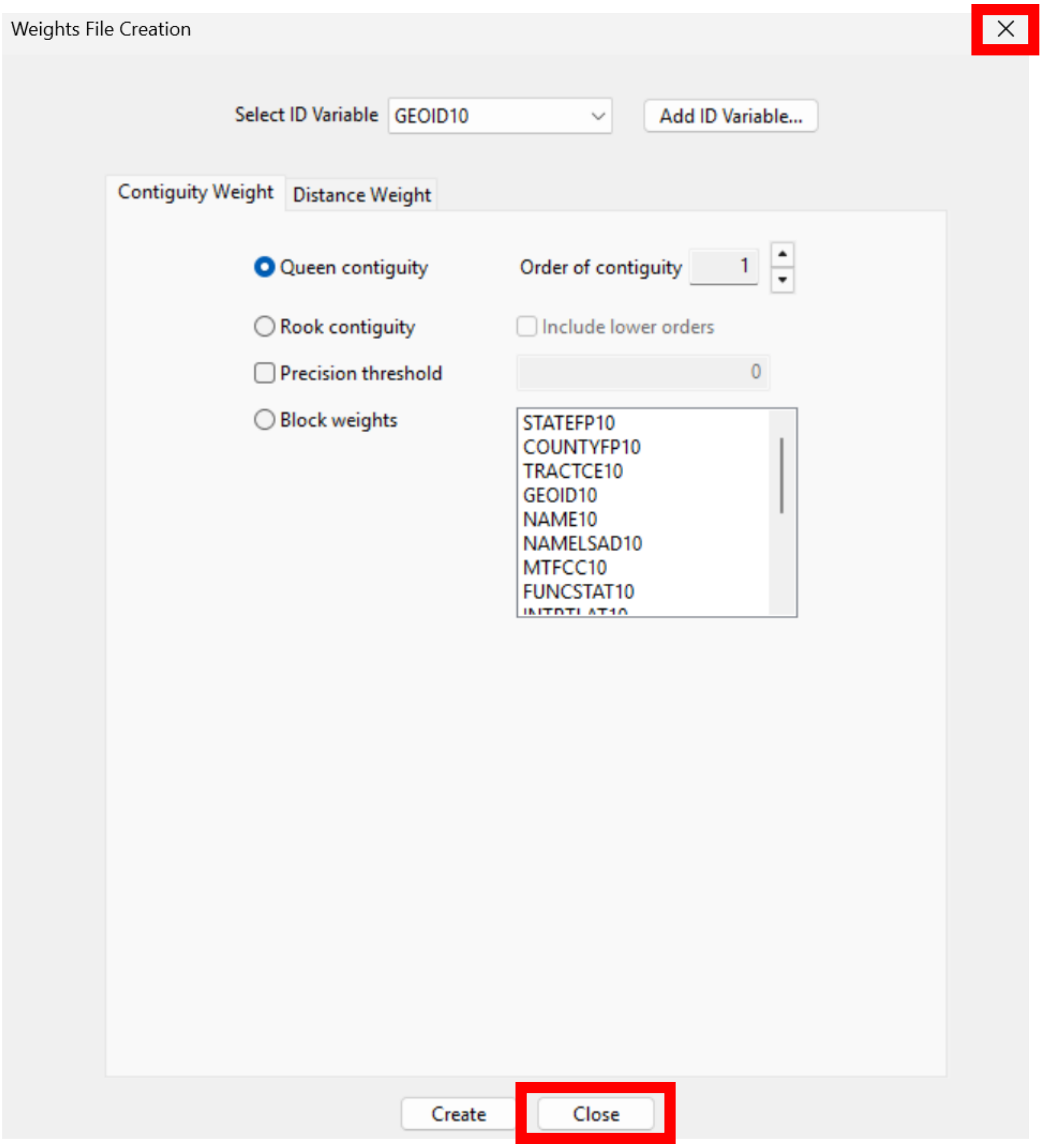
   Note: *After you have created your weights file the first time, you can import it into GeoDa the next time you run your analysis, by using the "Load" button in the Weights Manager.*



File name: PAcancerUpdated2

Save as type: GAL files (*.gal)

6. Once you have saved your file, close the Weights File Creation window, using the “**x**” in the upper right or the “**Close**” button at the bottom of the window.



Weights File Creation

Select ID Variable: GEOID10

● Queen contiguity — Order of contiguity: 1

○ Rook contiguity — □ Include lower orders

□ Precision threshold — 0

○ Block weights

STATEFP10
COUNTYFP10
TRACTCE10
GEOID10
NAME10
NAMELSAD10
MTFCC10
FUNCSTAT10

Explore the Weights File

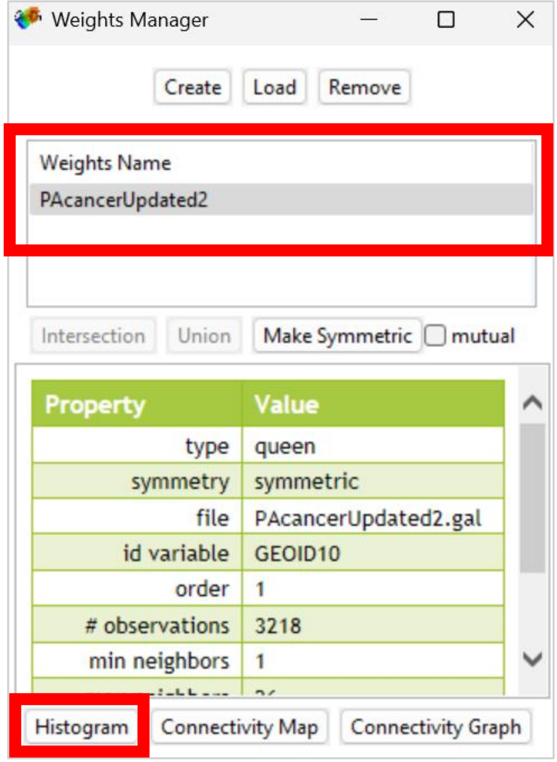
Now that we have created our weights file, we can learn more about the data in our weights file, using the **Histogram**, **Connectivity Map**, and **Connectivity Graph** tools in the **Weights Manager**.

Histogram

Let's start by using the **Histogram** tool to explore our weights file to visualize how many tracts have a certain amount of neighbors (i.e., tracts they share a border with).

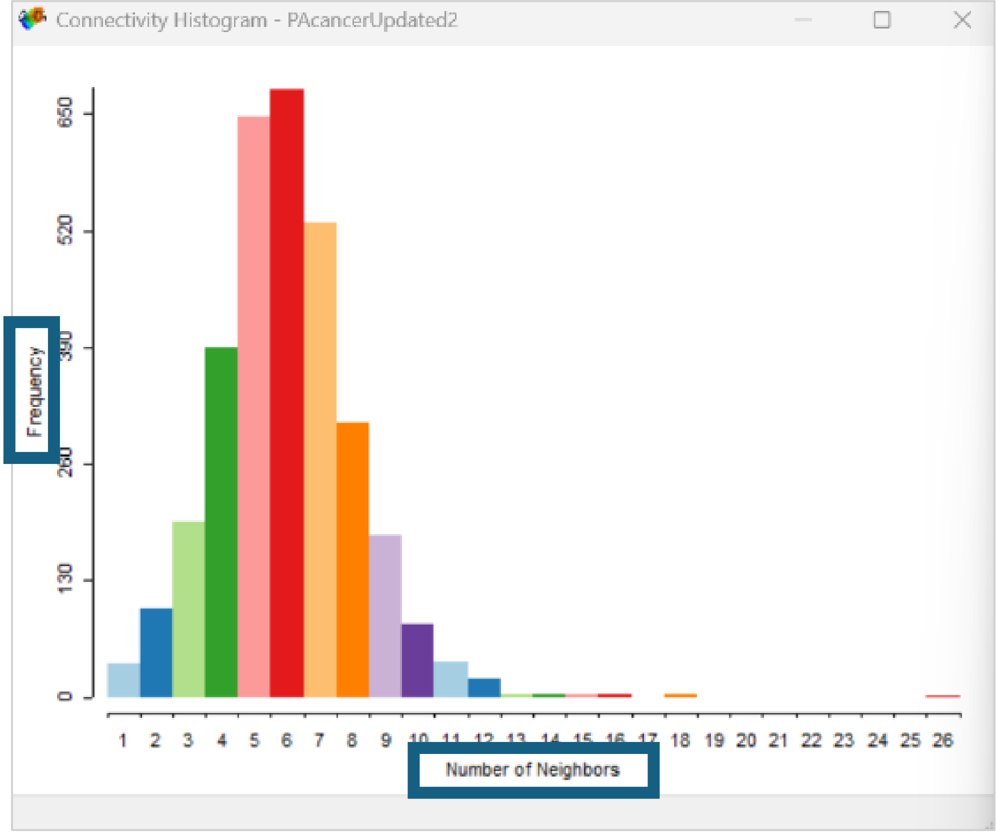
1. First, select the "**PAcancerUpdated2**" weights file we created from the menu in the top of the **Weights Manager**. Then, select the "**Histogram**" button in the bottom left.

   Note: *If you are unable to select the "Histogram" button, please be sure that you have created or loaded your weights file (i.e., the "PAcancerUpdated2" file) and that it is selected and highlighted in the top of the Weights Manager.*



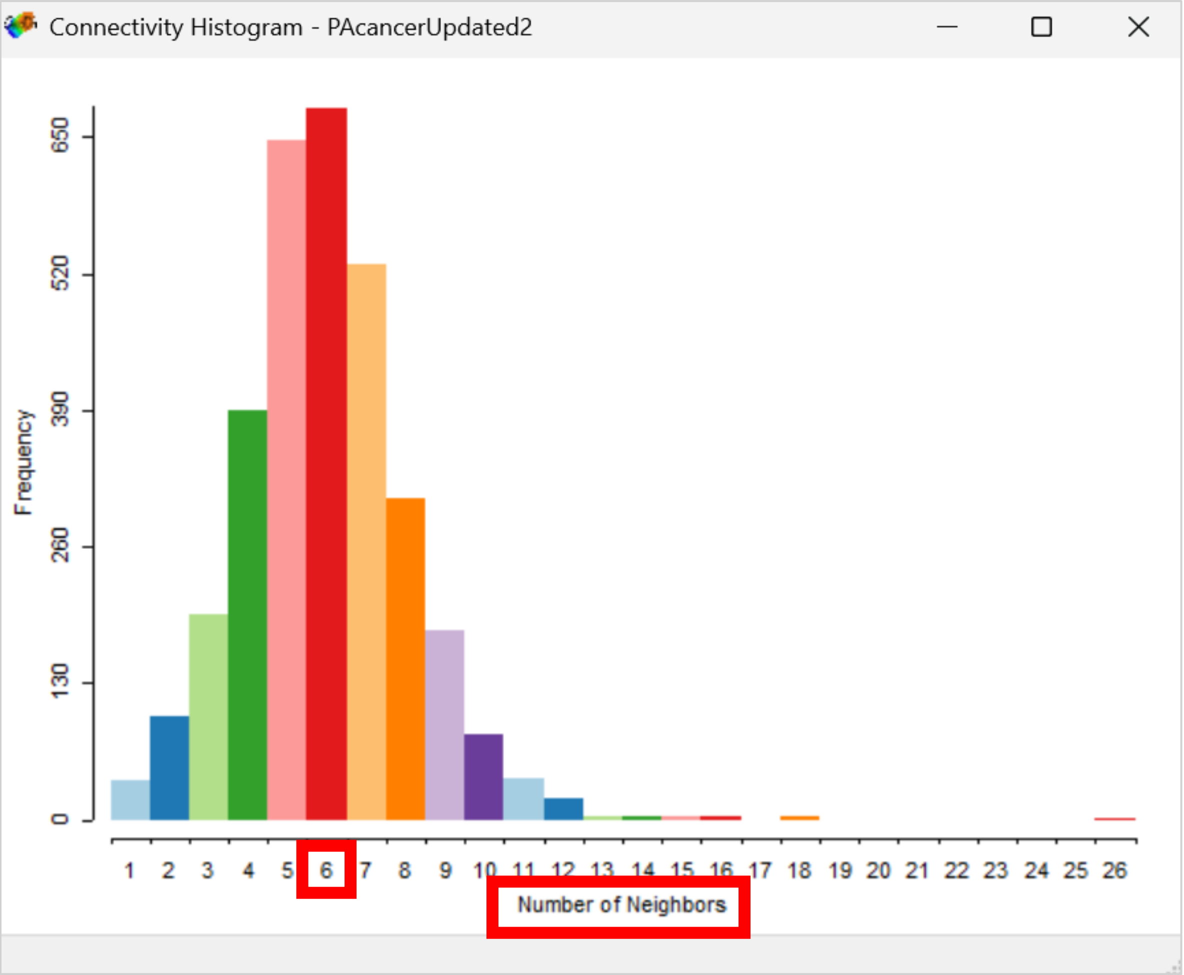
Property	Value
type	queen
symmetry	symmetric
file	PAcancerUpdated2.gal
id variable	GEOID10
order	1
# observations	3218
min neighbors	1

GeoDa will automatically create and open the connectivity **histogram** for us to explore, which shows the **frequency of tracts (y-axis)** with a **certain number of neighbors (x-axis)**.

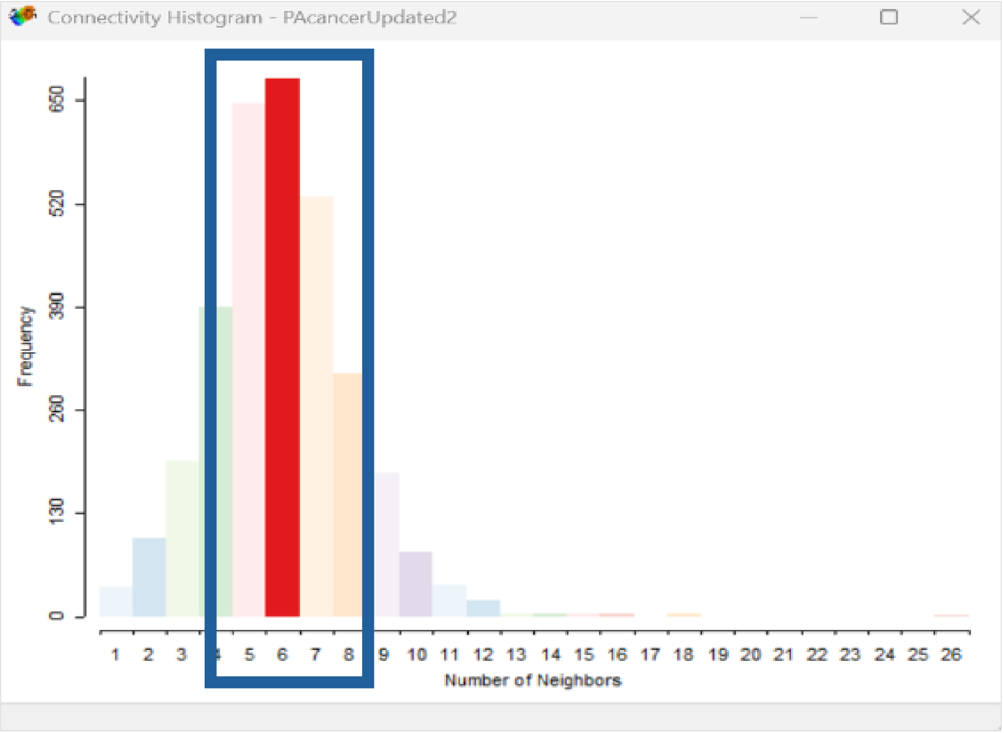


Note: *If you still have the tracts with the highest rates selected, those tracts will also be highlighted by default in the histogram when it opens. You can click any white space within the histogram chart to clear the filter or select any bar on the histogram to apply a new filter instead.*

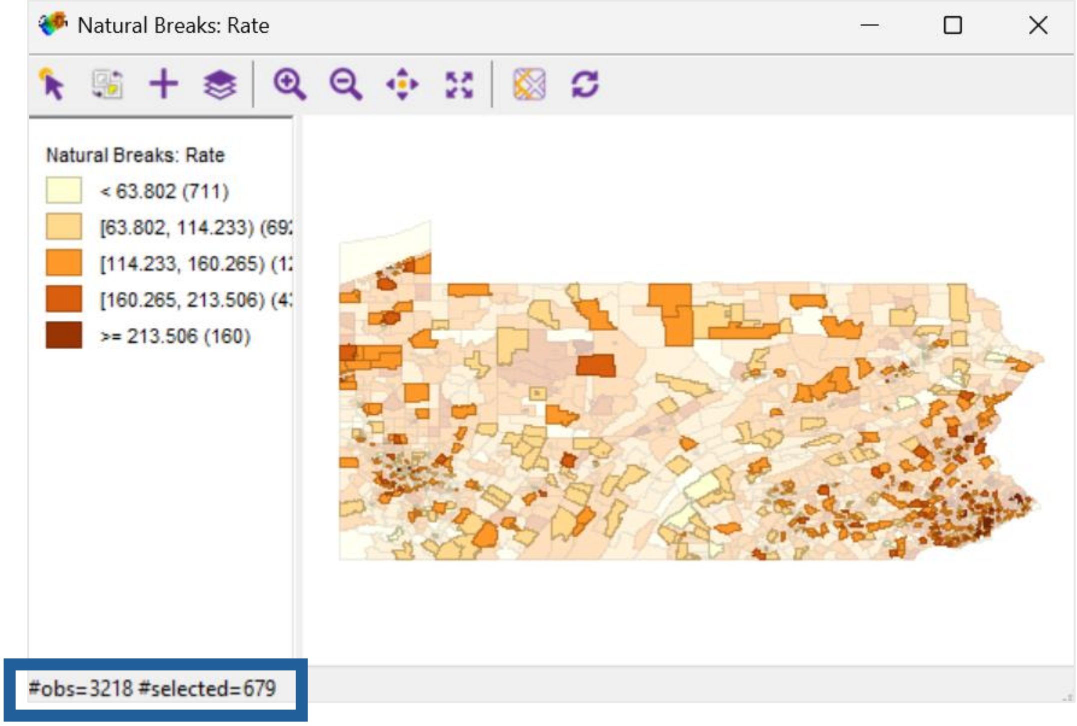
2. The bars of the histogram are **interactive** and selecting one will highlight all tracts with that number of neighbors in the other maps and tables that we have open. To give this a try, select the bar that represents tracts with **6 neighbors** from the **x-axis**.



The bar will be highlighted in the histogram and will act as a filter for all the other windows we have open.



For example, the natural breaks map is now highlighting the **679 census tracts** that have **6 neighbors**.



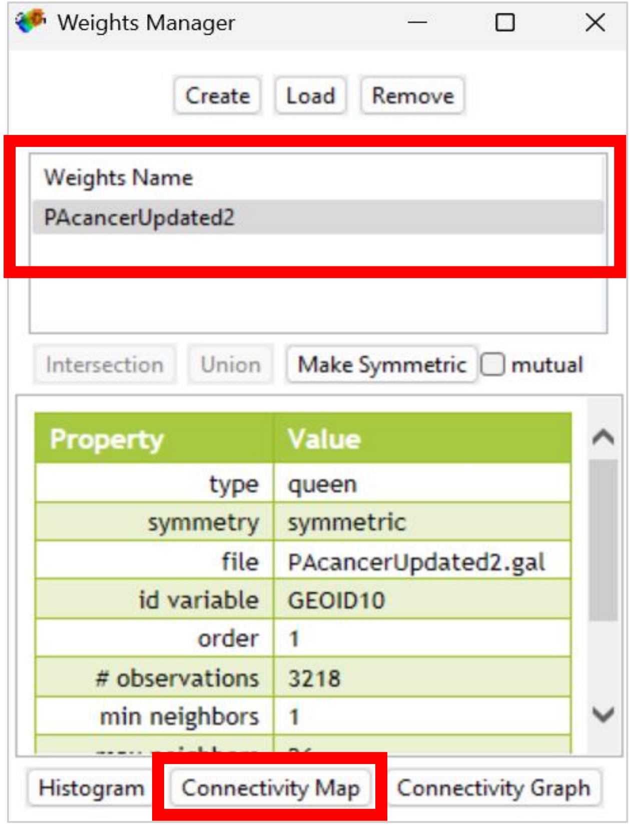
```
#obs=3218 #selected=679
```

Connectivity Map

Next, let's create and explore the **Connectivity Map**.

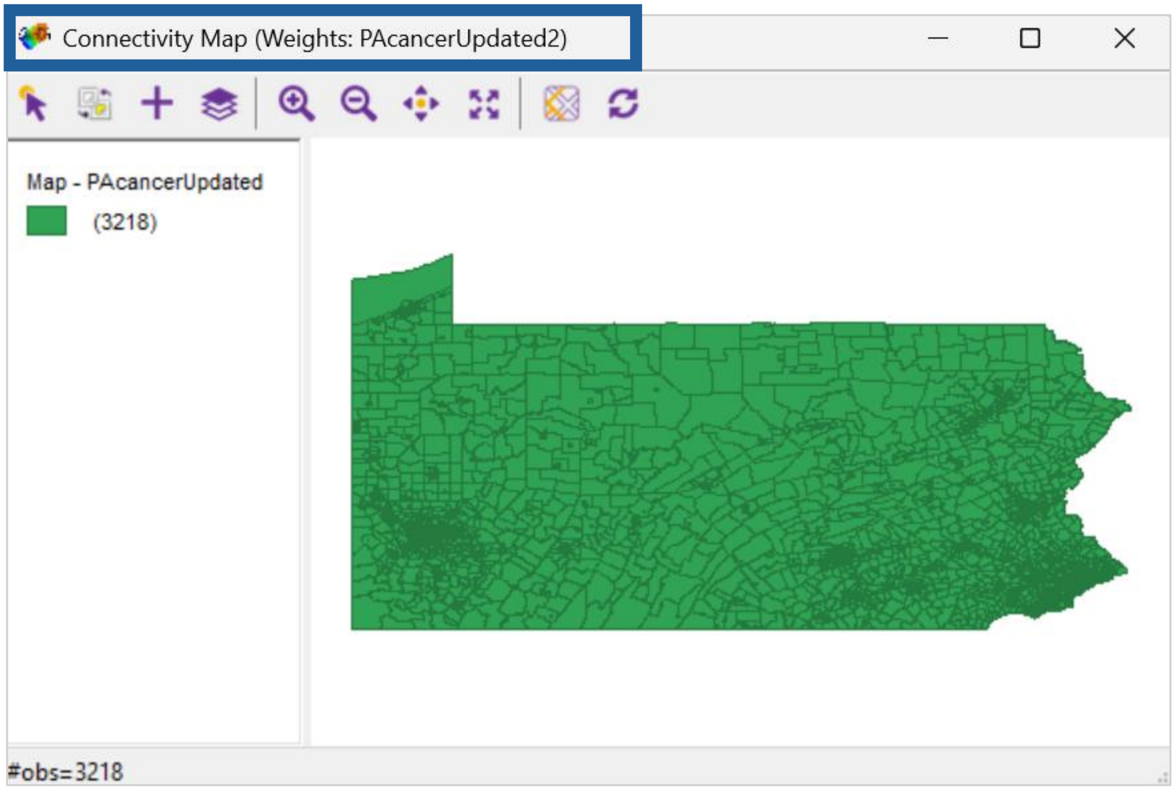
1. Select the "**Connectivity Map**" button in the bottom center of the **Weights Manager** window.

   Note: *If you are unable to select the "Connectivity Map" button, please be sure that you still have the "PAcancerUpdated2" file selected and highlighted in the top of the Weights Manager.*



Property	Value
type	queen
symmetry	symmetric
file	PAcancerUpdated2.gal
id variable	GEOID10
order	1
# observations	3218
min neighbors	1

GeoDa will automatically create and open the **Connectivity Map**, which will look a lot like the map from when we imported the data, until you hover above a community or tract on the map. You can tell this is the **Connectivity Map**, by the title in the upper left of the window.



Connectivity Map (Weights: PAcancerUpdated2)

Map - PAcancerUpdated

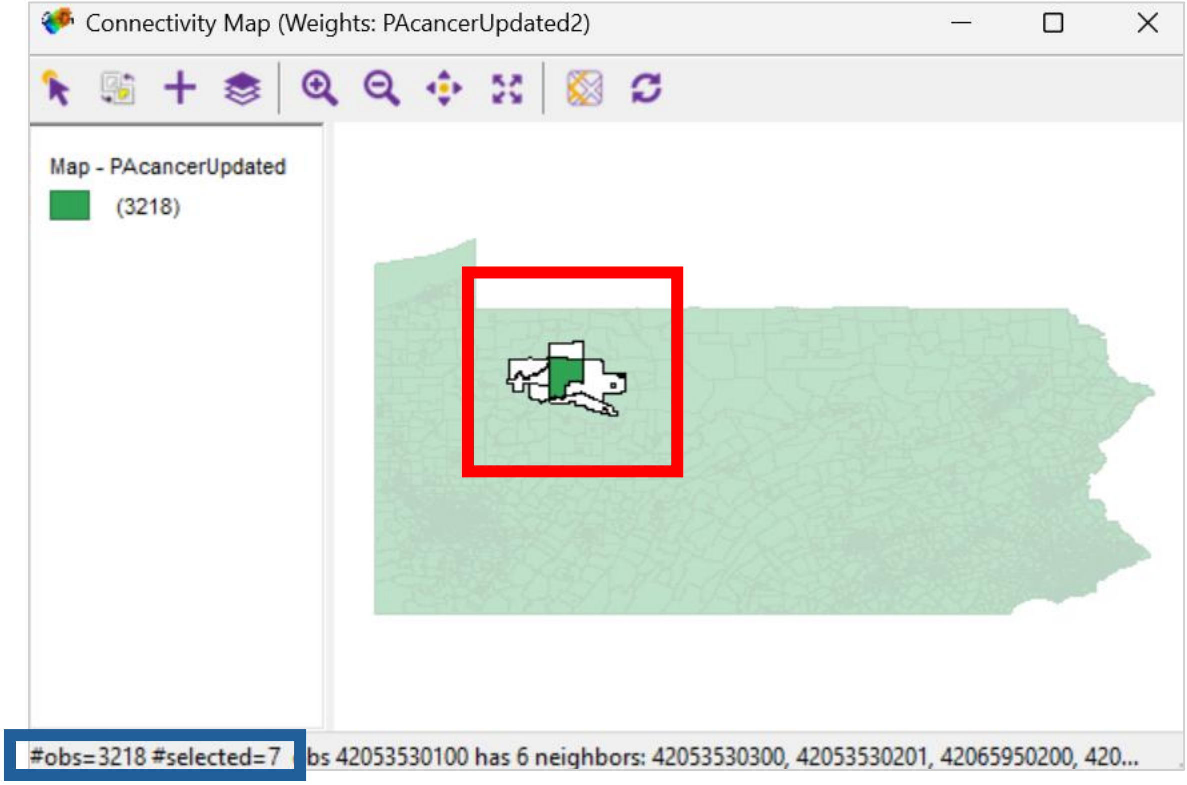
🟩 (3218)

#obs=3218

Note: *If you still have tracts selected from another window when you create this map, those tracts will also be highlighted by default when it opens. Hovering over the connectivity map will clear the filter.*

2. To see which tracts share a border with the tract of interest, **hover above** a tract to highlight that tract in dark green and all its neighbors that it is connected with in white.

   The number of tracts that are currently selected will appear in the bottom left of the **Connectivity Map** window.



Connectivity Map (Weights: PAcancerUpdated2)

Map - PAcancerUpdated

🟩 (3218)

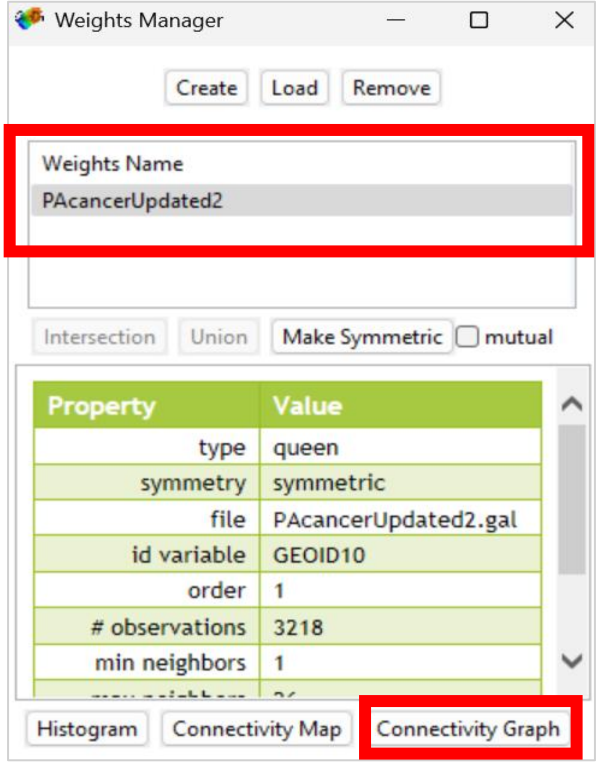
#obs=3218 #selected=7 bs 42053530100 has 6 neighbors: 42053530300, 42053530201, 42065950200, 420...

Connectivity Graph

Lastly, let's use the **Connectivity Graph** tool to help us visualize how tracts are connected to one another.

1. Select the "**Connectivity Graph**" button in the bottom right of the **Weights Manager** window.

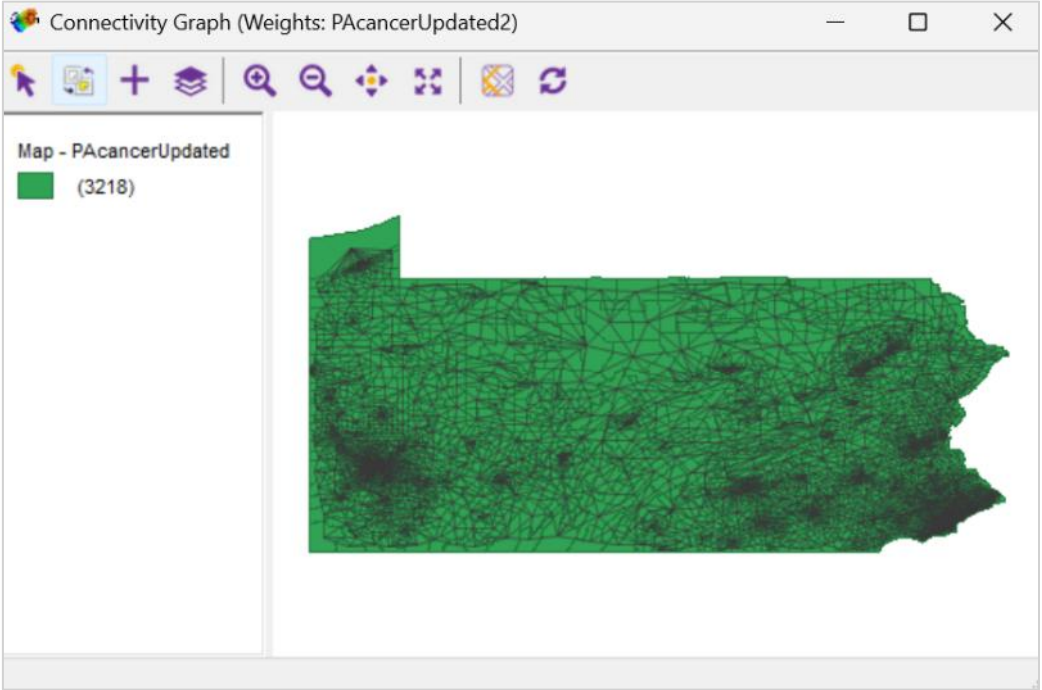
   Note: *If you are unable to select the "Connectivity Graph" button, please be sure that you still have the "PAcancerUpdated2" file selected and highlighted in the top of the Weights Manager.*



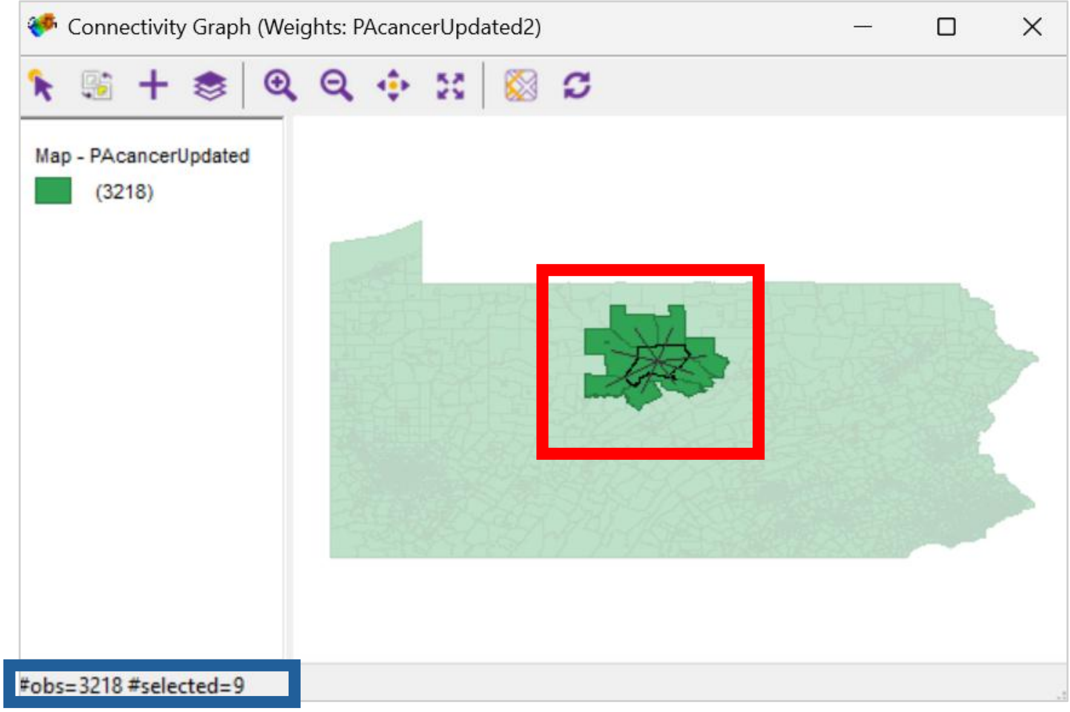
Weights Name
PAcancerUpdated2

Property	Value
type	queen
symmetry	symmetric
file	PAcancerUpdated2.gal
id variable	GEOID10
order	1
# observations	3218
min neighbors	1

GeoDa will automatically create and open the **Connectivity Graph**, which looks like the **Connectivity Map**, but with webs of lines that connect the center (i.e., centroid) of each tract with the centroids of its neighbors.



2. Select a **tract** on the map to **highlight** it and see which tracts are its neighbors. The number of selected tracts will also appear in the bottom left of the **Connectivity Graph** window.



#obs=3218 #selected=9

Global Moran’s I

For this tutorial, we will use **Global Moran’s I** (i.e., **Univariate Moran’s I**) to measure how similar neighboring tracts are to each other in terms of prostate cancer rates.

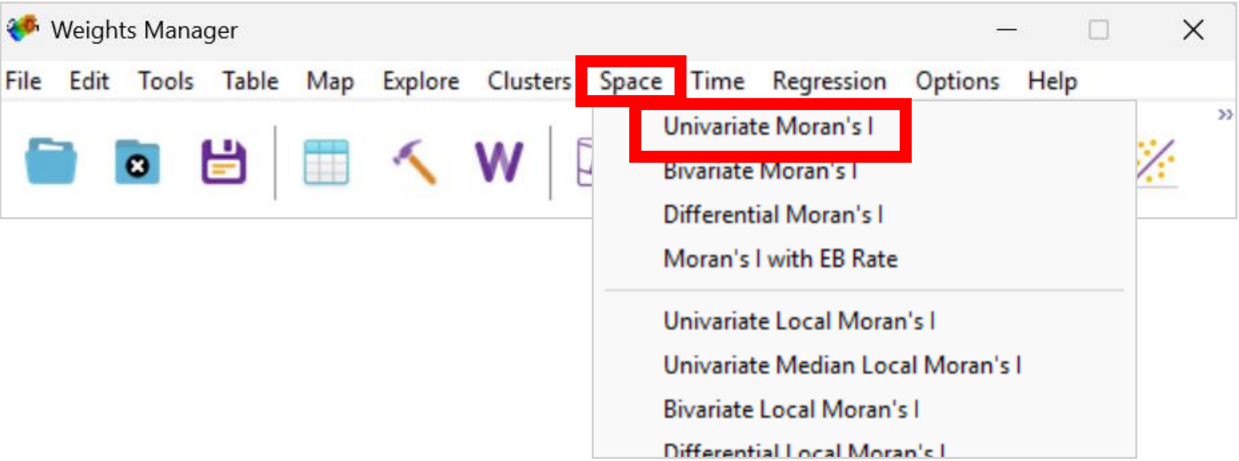
The results of our analysis will be the **Moran’s I scatterplot.** The **scatterplot points** help us visualize if neighboring tracts have similar cancer rates, by showing us how similar the cancer rate of each tract is to the cancer rates of its neighbors . The results also show us how close the cancer rates of a tract and its neighbors are to the overall mean.

The scatterplot results also include the **regression line** (i.e., the line of best fit through the scatterplot points), which helps us check for **spatial clustering** – in other words, whether or not tracts with high cancer rates are grouped together or spread out randomly.

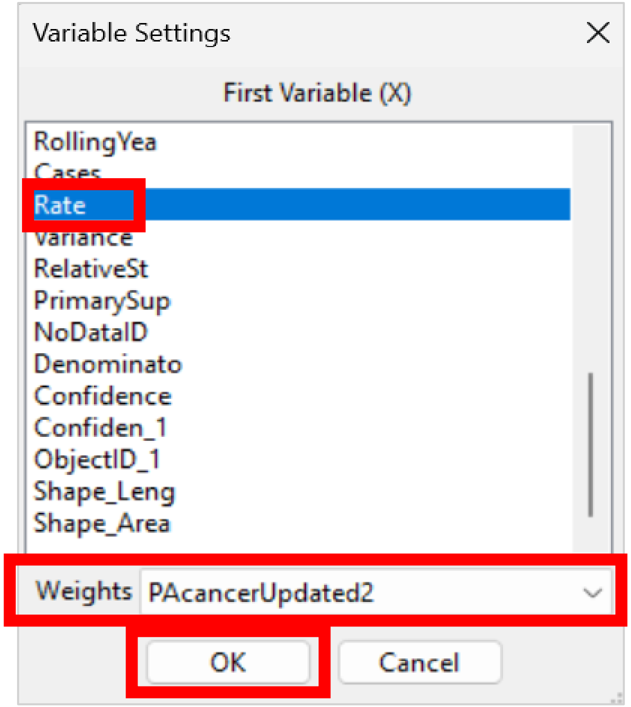
The **slope of the regression line** is our **Moran’s I statistic**, which will always be a value between **-1 and 1**. A **positive slope** means tracts with similar values are located near each other (i.e., **spatial clustered**), while a **negative slope** means tracts with high and low values are mixed together (i.e., **spatial dispersed**).

To learn more about the Moran’s I scatterplot results and how to interpret the results, see: https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#moran-scatter-plot

1. To run Global Moran’s I, select the “**Space**” option from the toolbar and “**Univariate Moran’s I**” from the first section of the **Space** menu.



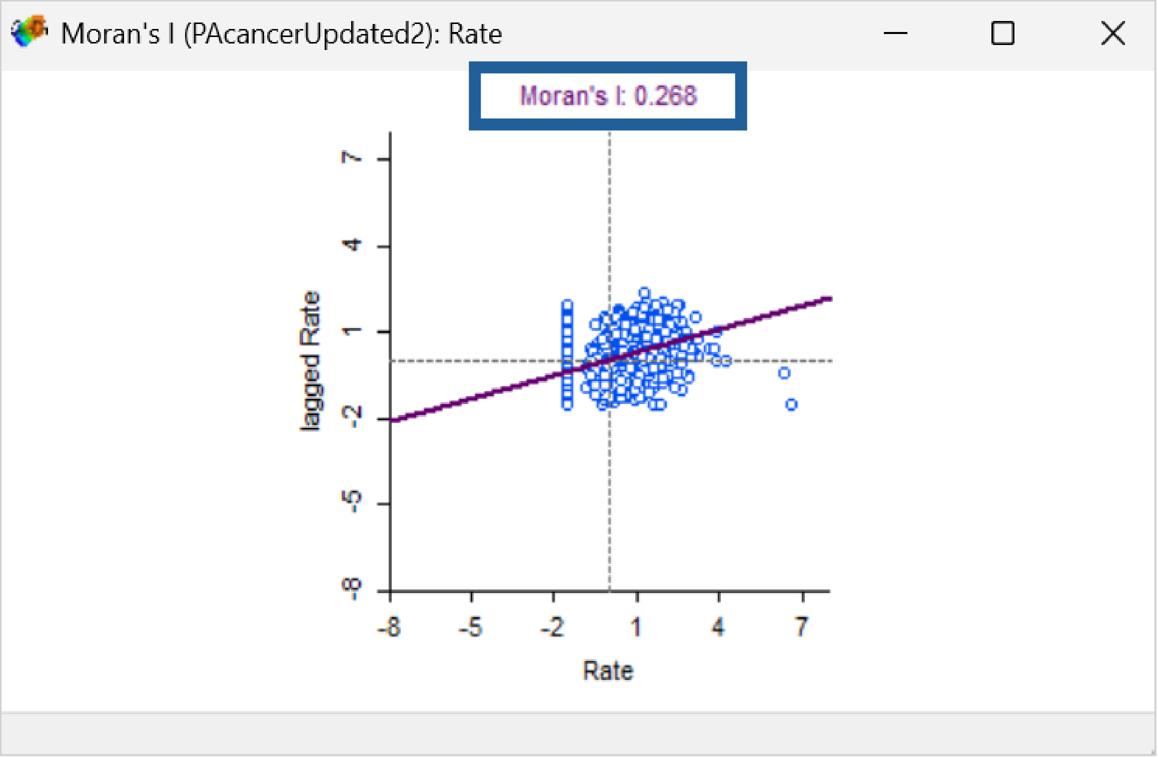
2. Next, we need to select the value that we want to compare between tracts. For this analysis, scroll down and select the [**Rate**] variable. Ensure that the **spatial weights file** (**PAcancerUpdated2)** we created is selected as the **Weights** file in the drop-down menu at the bottom. Then select "**OK**" to run the analysis.



Weights PAcancerUpdated2

The **Moran's I scatterplot** (below) will appear, with the **prostate cancer rates** on the x-axis and their **spatially lagged (weighted sum of the neighboring location values) counterparts** on the y-axis. Our plot shows a **positive slope** to the scatterplot points and includes the slope (i.e., **Moran's I statistic**) at the top of our scatterplot, as **Moran's I = 0.268**. The Moran's I index typically ranges from -1 to +1, like a Pearson correlation coefficient. A positive Moran's I indicates positive spatial autocorrelation or clustering. Similar values (either high values or low values) tend to be located near each other. A negative Moran's I indicates negative spatial autocorrelation or dispersion. Dissimilar values tend to be located near each other. This is often described as a "checkerboard" pattern. Moran's I near zero suggests no significant spatial autocorrelation, meaning the spatial distribution of the variable is essentially random. The values are distributed independently of their location. The Moran's I value is essentially a Pearson correlation coefficient (r) and can be interpreted as such. A value of 0.268 in our example would be indicative of weak positive autocorrelation across the study area.

Note: *If you had any tracts still selected from exploring the weights file, those tracts will be highlighted in the scatterplot when it initially opens. Click any white space in the scatterplot results to see all results or highlight different tracts in the plot.*



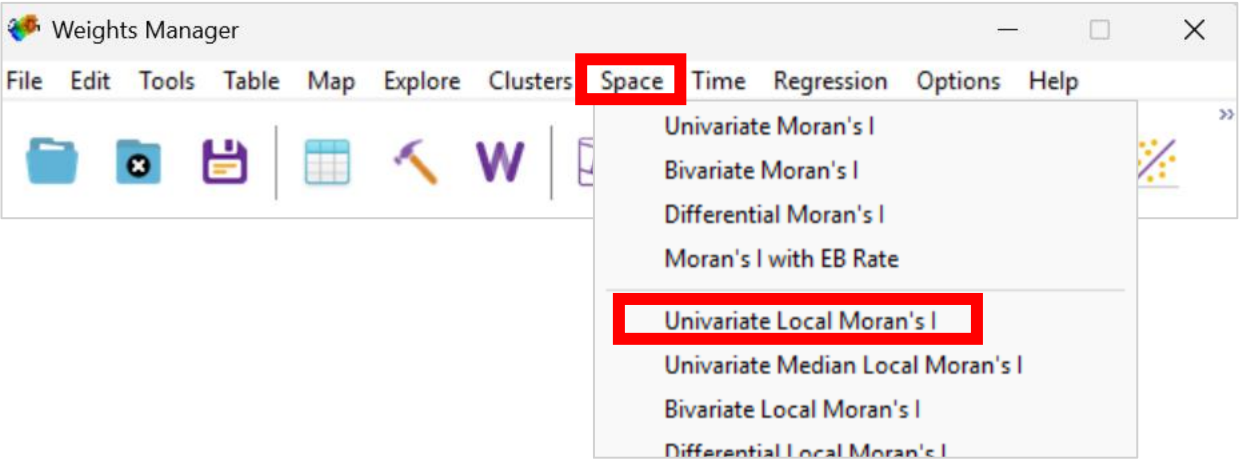
Local Moran’s I

Next, let’s run **Local Moran’s I** (i.e., **Univariate Local Moran’s I**), to help us determine where clusters (i.e., tracts with similar cancer rates) are located.

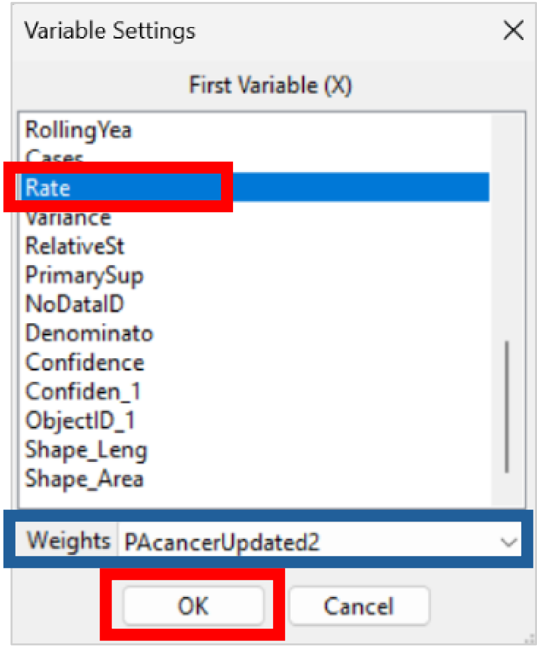
Moran's I is a global statistic that measures the overall spatial autocorrelation of a variable across an entire study area. While it tells you if spatial autocorrelation exists globally, it doesn't tell you where those clusters are located within the study area. **Local Moran's I** is a common type of **LISA (Local Indicators of Spatial Association.)** It is a local statistic that decomposes the global Moran's I into the contribution of each individual observation. It identifies specific locations where significant spatial clustering hot spots or cold spots occur.

The results of the Local Moran’s I includes a **Significance Map**, **Cluster Map**, and **Moran’s I Scatterplot**.

1. To run Local Moran’s I, select the “**Space**” option from the toolbar again, but this time select the “**Univariate Local Moran’s I**” from the top of the second section of the **Space** menu to run the **Local Moran’s I** analysis.

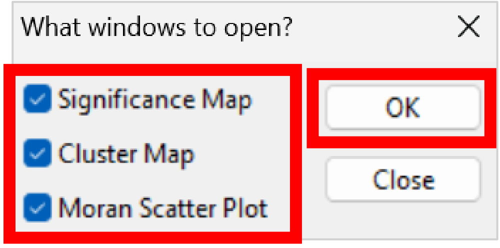


2. Next, scroll down and select the [**Rate**] variable. Ensure that our weights file, **PAcancerUpdated2,** is selected as the Weights file in the drop-down menu at the bottom. Then select "**OK**" to run the analysis.



Weights PAcancerUpdated2

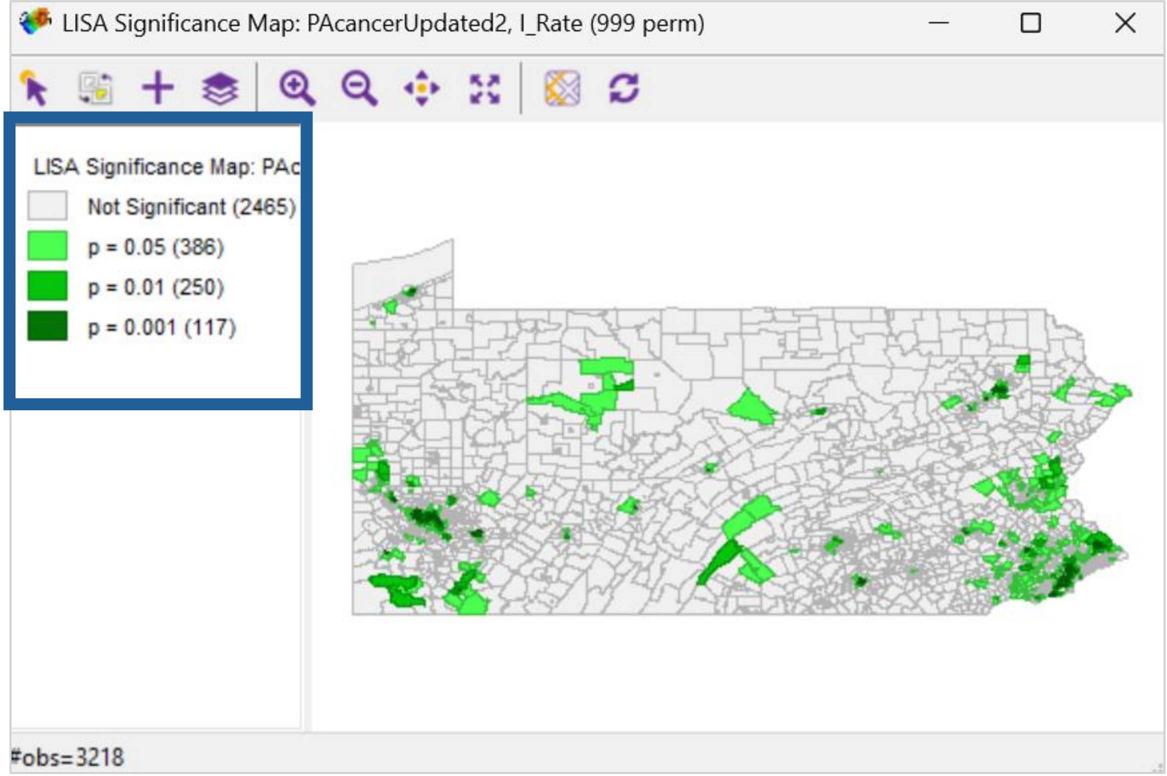
3. Another window will appear to allow you to select what windows you would like to open as part of this analysis. Check the boxes for all three options, the: **Significance Map**, **Cluster Map**, **Moran Scatter Plot**. Then select the "**OK**" button.



☑ Significance Map
☑ Cluster Map
☑ Moran Scatter Plot

Significance Map

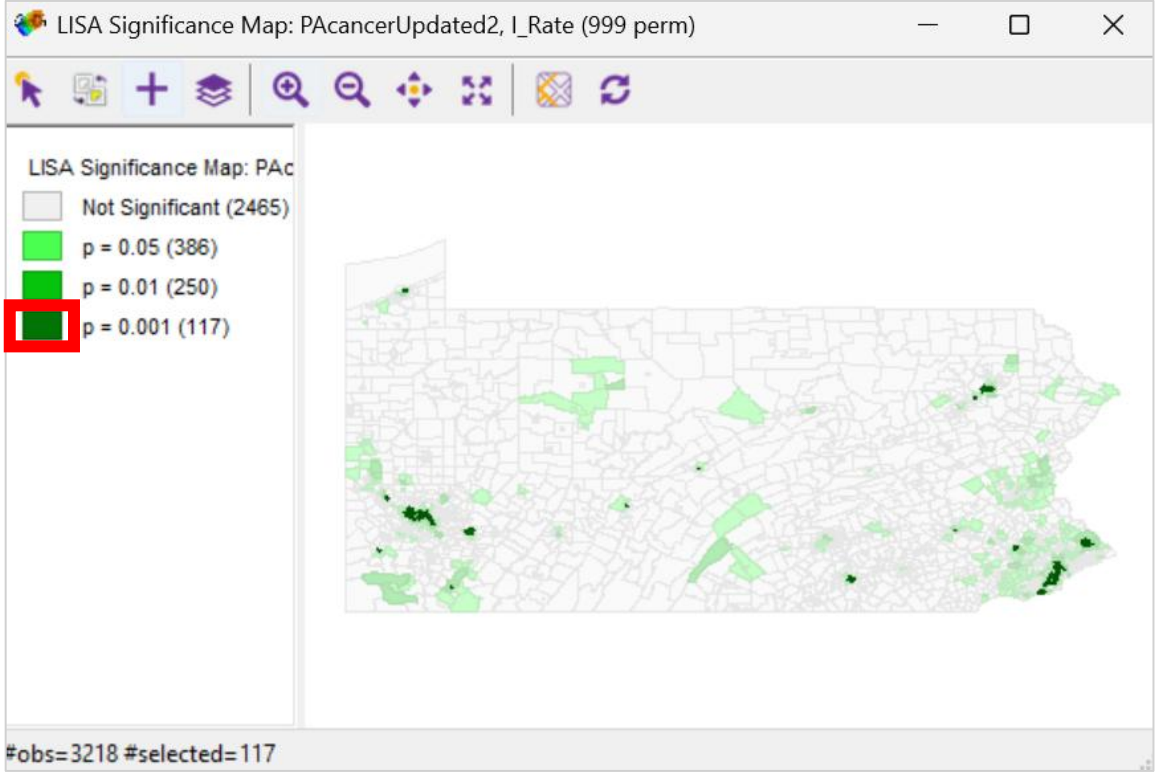
The **Significance Map** shows where clustering is statistically significant, with the darker greens signifying more **statistically significant clustering**. The legend also includes information for how many tracts are in each category.


LISA Significance Map: PAcancerUpdated2, I_Rate (999 perm)
LISA Significance Map: PAc
Not Significant (2465)
p = 0.05 (386)
p = 0.01 (250)
p = 0.001 (117)
#obs=3218

In this example, we see:

- **2,465** census tracts have **no statistically significant** clustering (**~77%** of all PA census tracts)
- **386** census tracts have a **p-value between 0.05 - 0.01**
- **250** census tracts have a **p-value between 0.01 - 0.001**
- **117** census tracts have the **most statistically significant clustering** (**p-value < 0.0001**)

1. Click on a legend color to highlight a specific significance category. For example, click the color for the **darkest green category** (**p = 0.001**), to highlight the **117 census tracts** where we see the **most statistically significant clustering**.



LISA Significance Map: PAcancerUpdated2, I_Rate (999 perm)

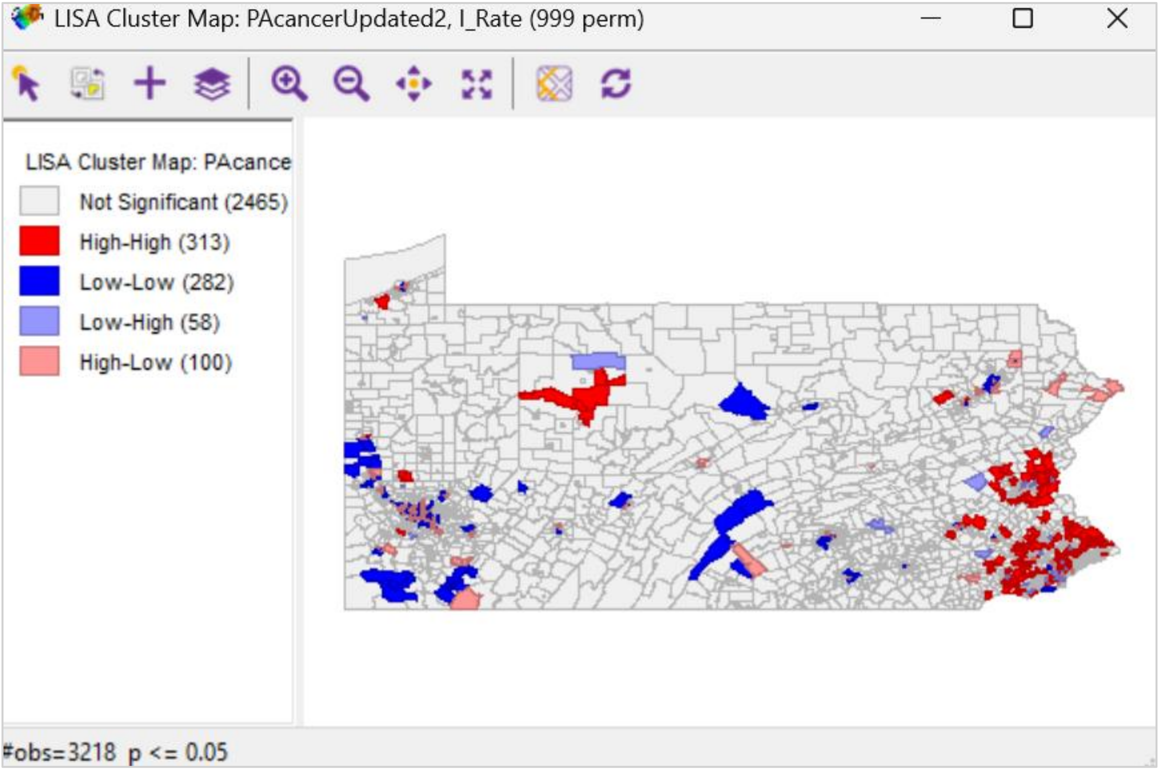
LISA Significance Map: PAc

- Not Significant (2465)
- p = 0.05 (386)
- p = 0.01 (250)
- p = 0.001 (117)

#obs=3218 #selected=117

Cluster Map

The **cluster map** allows us to see **what kind** of clustering pattern each tract has: **High-High**, **Low-Low**, **Low-High**, and **High-Low**.



LISA Cluster Map: PAcancerUpdated2, I_Rate (999 perm)

LISA Cluster Map: PAcance

⬜ Not Significant (2465)
🟥 High-High (313)
🟦 Low-Low (282)
🟪 Low-High (58)
🟥 High-Low (100)

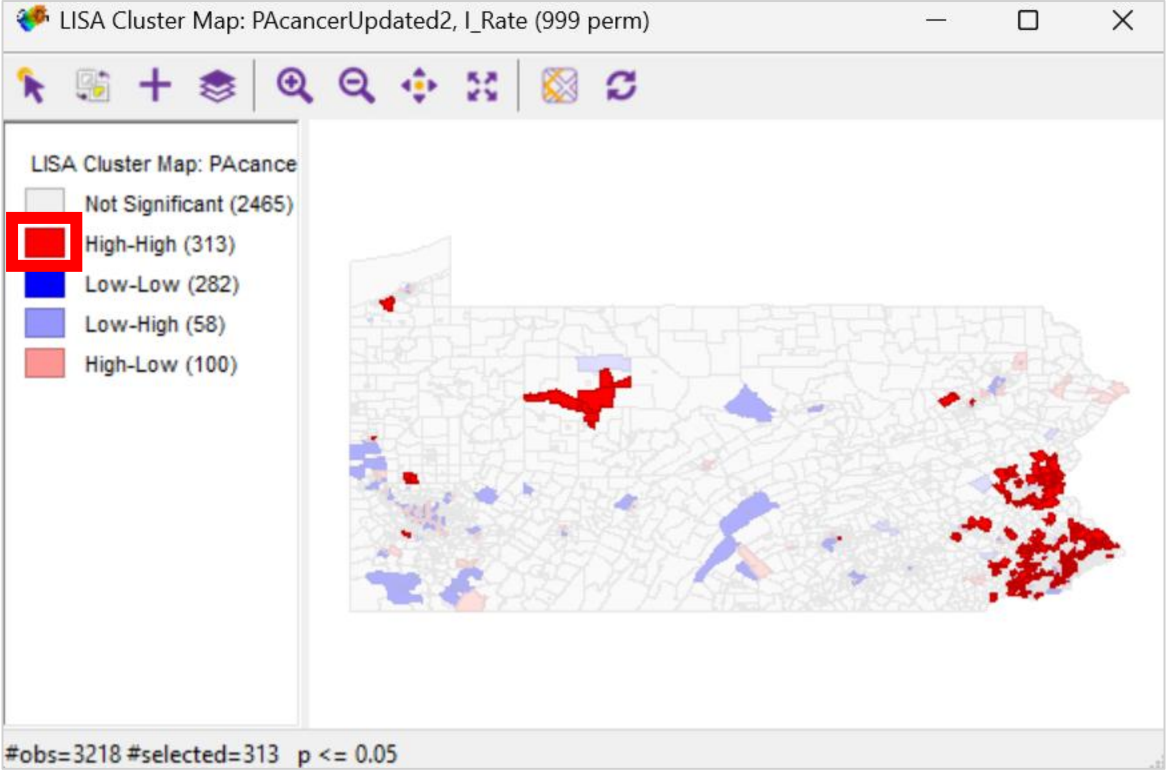
#obs=3218 p <= 0.05

Just like with the **Significance Map**, we can see how many tracts have each clustering type in the legend:

Cluster Type	Meaning	# of Tracts
High-High	Tracts with **high prostate cancer rates** that are surrounded by other tracts with **high prostate cancer rates**	**313**
Low-Low	Tracts with **low prostate cancer rates** that are surrounded by other tracts with **low prostate cancer rates**	**282**
Low-High	Tracts with **low prostate cancer rates** that are surrounded by tracts with **high prostate cancer rates** (**outliers**)	**58**
High-Low	Tracts with **high prostate cancer rates** that are surrounded by tracts with **low prostate cancer rates (outliers)**	**100**

1. We can also use the legend to highlight one particular category, like we did with the **Significance Map**. For example, click the legend color for the **High-High** category, to highlight the **313** tracts with high prostate cancer rates that are surrounded by other tracts that also have high cancer cluster rates.

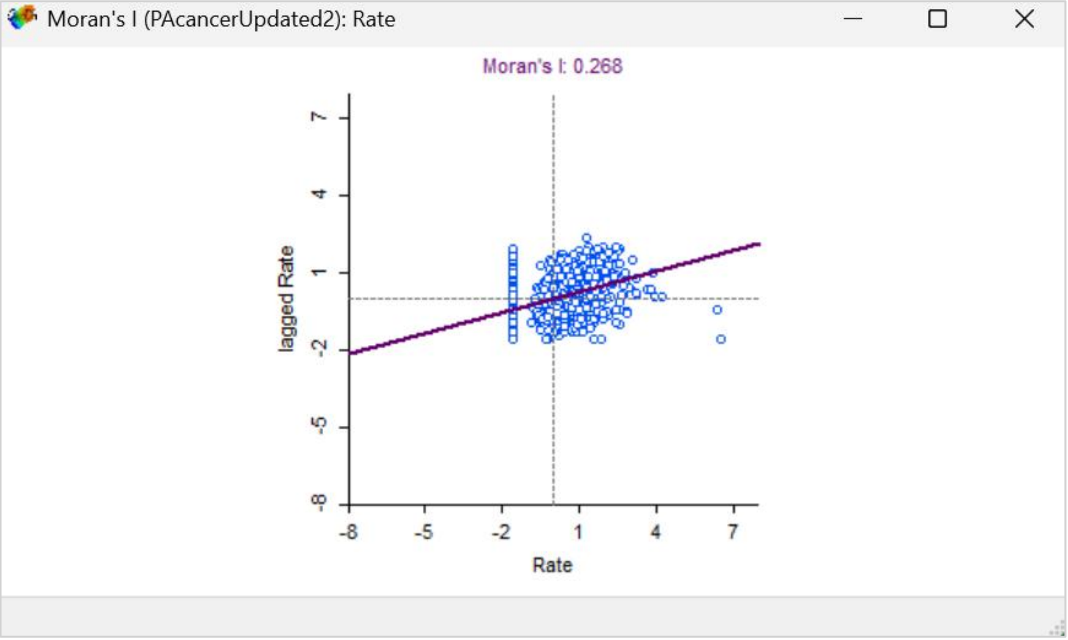
   Note: *Note how these tracts are different from the 117 tracts with the most statistically significant clustering that we saw highlighted in the Significance Map.*


LISA Cluster Map: PAcancerUpdated2, I_Rate (999 perm)
LISA Cluster Map: PAcance
Not Significant (2465)
High-High (313)
Low-Low (282)
Low-High (58)
High-Low (100)
#obs=3218 #selected=313 p <= 0.05

   Clear your selection, by selecting any white space in the map window before moving forward. This will allow us to save and join our **Local Moran's I** results to our **Data Table** in the next section.

Moran’s I Scatterplot

The **Moran’s I Scatterplot** will appear and look similar to what we saw with our Global Moran’s I scatterplot.

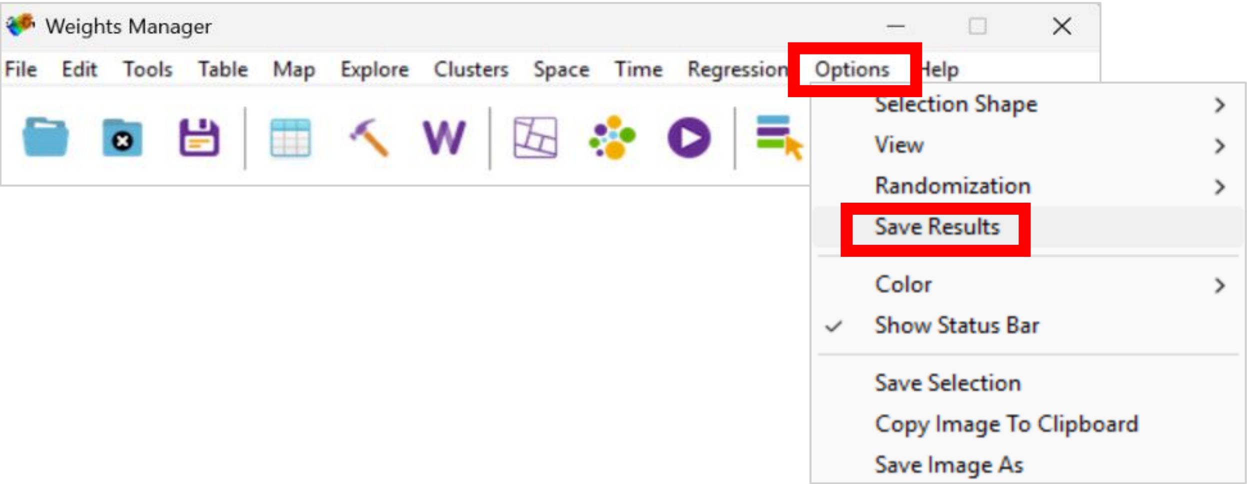


Results

Join and Save the Results

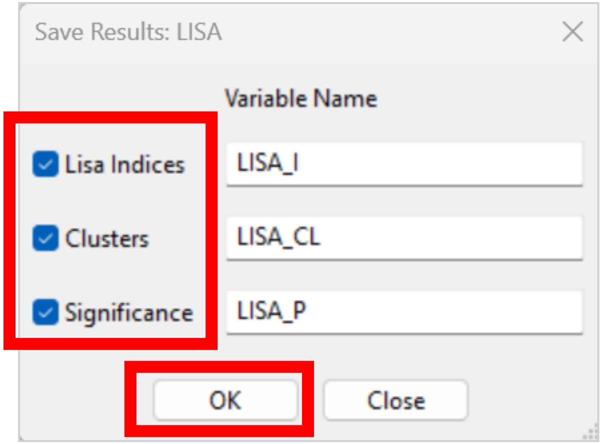
Now, let's join and save our **Local Moran's I results** to our **Data Table**, to be able to view and explore all of our data in one place.

1. First, make sure that you are on the **Cluster Map** results window from the last section and that you clear your current highlighted selection by clicking any white space on the map. Then, click the "**Options**" menu from the toolbar and select "**Save Results**" from the drop-down menu.



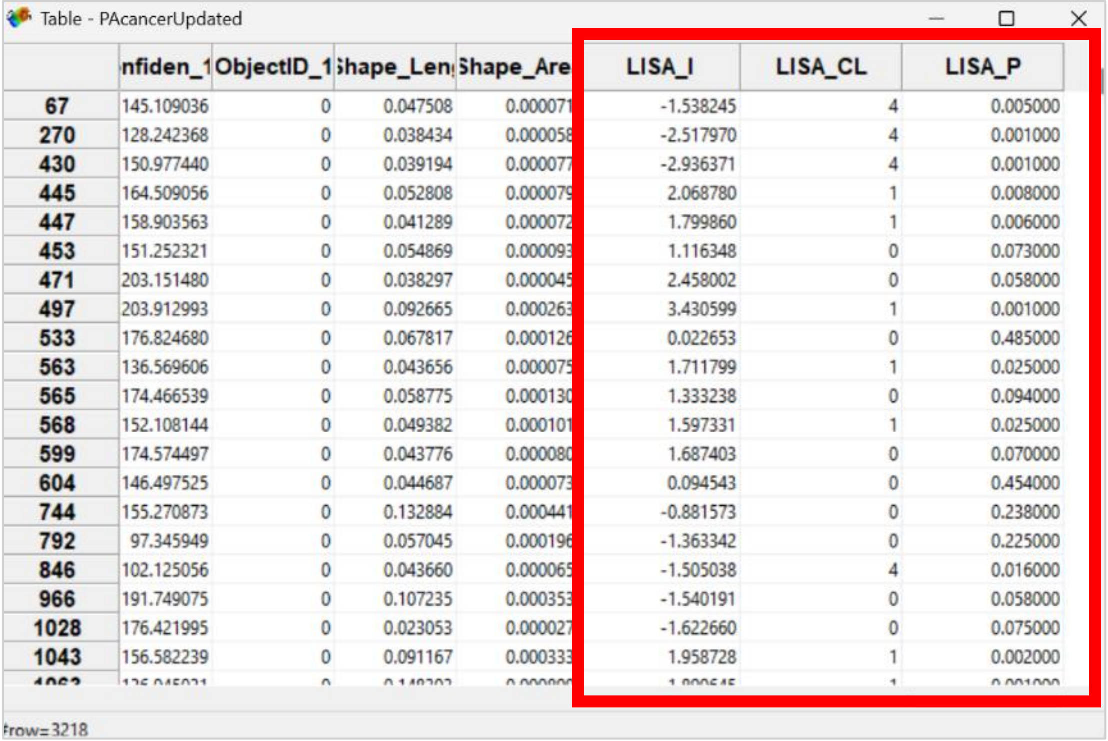
2. A window will then appear to let you select what 3 variables you would like to save to the **Data Table**. Check all options: **LISA Indices ([LISA_I]), Clusters ([LISA_CL])**, and **Significance ([LISA_P])**. Leave the variable names the same and select the "**OK**" button in the bottom left to save them to the **Data Table**.

Note: *If you are not seeing these variable options, return to the previous step and make sure that your Cluster Map window is selected and that no tracts are highlighted before saving your results.*


Save Results: LISA

	Variable Name
☑ Lisa Indices	LISA_I
☑ Clusters	LISA_CL
☑ Significance	LISA_P

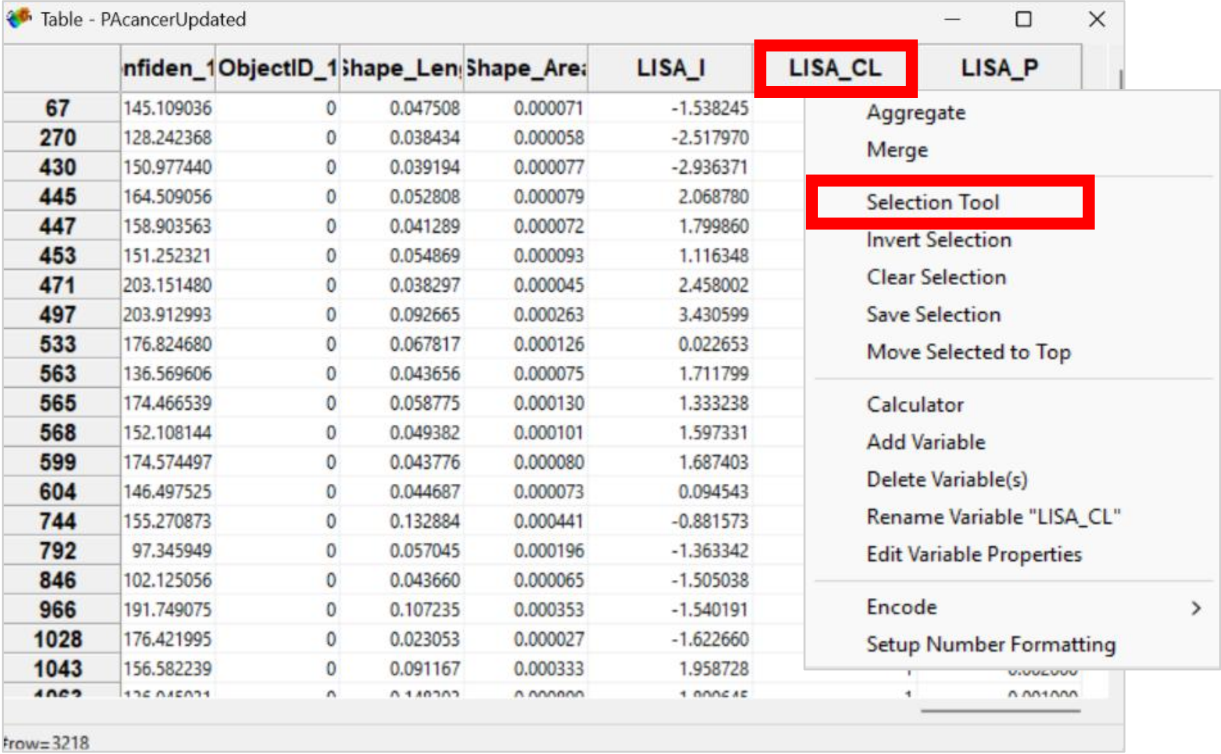
3. Next, let's check that the results were joined to the **Data Table**. Go back to the window with your table in it (or reopen it with the table icon in the toolbar) and use the scroll bar at the bottom to scroll all the way to the right. The last 3 columns should now match what we set for our results in the last step: [**LISA_I**], [**LISA_CL]**, and [**LISA_P**].



	nfiden_1	ObjectID_1	;hape_Len	Shape_Are	LISA_I	LISA_CL	LISA_P
67	145.109036	0	0.047508	0.000071	-1.538245	4	0.005000
270	128.242368	0	0.038434	0.000058	-2.517970	4	0.001000
430	150.977440	0	0.039194	0.000077	-2.936371	4	0.001000
445	164.509056	0	0.052808	0.000079	2.068780	1	0.008000
447	158.903563	0	0.041289	0.000072	1.799860	1	0.006000
453	151.252321	0	0.054869	0.000093	1.116348	0	0.073000
471	203.151480	0	0.038297	0.000045	2.458002	0	0.058000
497	203.912993	0	0.092665	0.000263	3.430599	1	0.001000
533	176.824680	0	0.067817	0.000126	0.022653	0	0.485000
563	136.569606	0	0.043656	0.000075	1.711799	1	0.025000
565	174.466539	0	0.058775	0.000130	1.333238	0	0.094000
568	152.108144	0	0.049382	0.000101	1.597331	1	0.025000
599	174.574497	0	0.043776	0.000080	1.687403	0	0.070000
604	146.497525	0	0.044687	0.000073	0.094543	0	0.454000
744	155.270873	0	0.132884	0.000441	-0.881573	0	0.238000
792	97.345949	0	0.057045	0.000196	-1.363342	0	0.225000
846	102.125056	0	0.043660	0.000065	-1.505038	4	0.016000
966	191.749075	0	0.107235	0.000353	-1.540191	0	0.058000
1028	176.421995	0	0.023053	0.000027	-1.622660	0	0.075000
1043	156.582239	0	0.091167	0.000333	1.958728	1	0.002000
1063	[illegible]	0	[illegible]	[illegible]	[illegible]	1	[illegible]

#row=3218

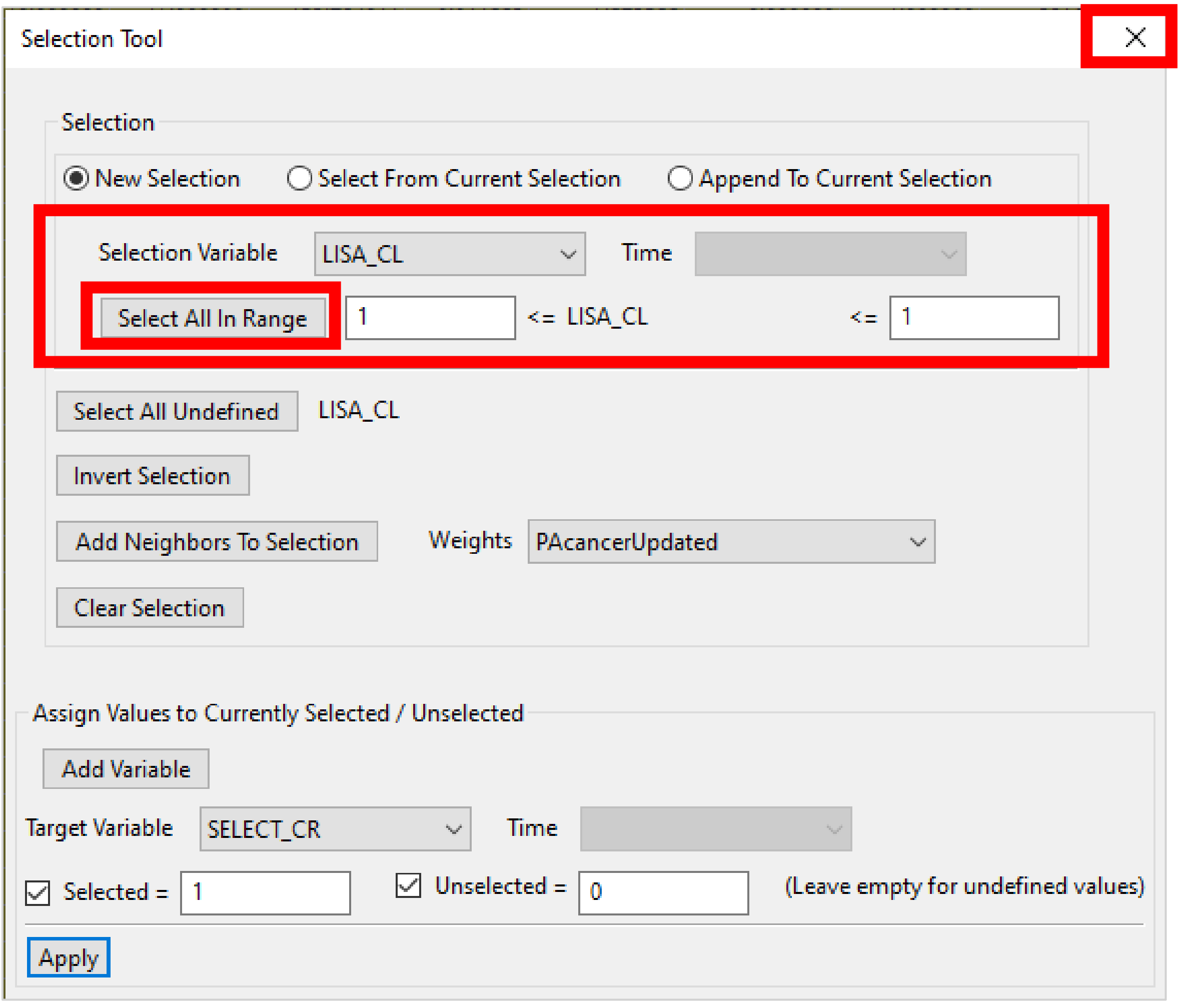
4. Next, right click the **[LISA_CL]** variable column name and choose the “**Selection Tool**” from the drop-down menu.



	nfiden_1	ObjectID_1	Shape_Len	Shape_Are	LISA_I	LISA_CL	LISA_P
67	145.109036	0	0.047508	0.000071	-1.538245		
270	128.242368	0	0.038434	0.000058	-2.517970		
430	150.977440	0	0.039194	0.000077	-2.936371		
445	164.509056	0	0.052808	0.000079	2.068780		
447	158.903563	0	0.041289	0.000072	1.799860		
453	151.252321	0	0.054869	0.000093	1.116348		
471	203.151480	0	0.038297	0.000045	2.458002		
497	203.912993	0	0.092665	0.000263	3.430599		
533	176.824680	0	0.067817	0.000126	0.022653		
563	136.569606	0	0.043656	0.000075	1.711799		
565	174.466539	0	0.058775	0.000130	1.333238		
568	152.108144	0	0.049382	0.000101	1.597331		
599	174.574497	0	0.043776	0.000080	1.687403		
604	146.497525	0	0.044687	0.000073	0.094543		
744	155.270873	0	0.132884	0.000441	-0.881573		
792	97.345949	0	0.057045	0.000196	-1.363342		
846	102.125056	0	0.043660	0.000065	-1.505038		
966	191.749075	0	0.107235	0.000353	-1.540191		
1028	176.421995	0	0.023053	0.000027	-1.622660		
1043	156.582239	0	0.091167	0.000333	1.958728		

Note: *You can right click any variable name to access and use the Selection Tool. For example, right clicking the [GEOID10] variable would also allow us to access the selection tool and filter the [LISA_CL] variable here.*

5. Select all records where **[LISA_CL]= 1**, indicative of **High-High** tracts, by first selecting the [**LISA_CL**] variable from the drop-down menu next to **Selection Variable**. Next, set the range of values to select to "**1 <= LISA_CL <= 1**" and then select the "**Select All in Range**" button to the left of the range to apply the selection to your **Data Table**. When you are done, use the "**x**" button in the top right corner of the **Selection Tool** to close it.



Selection Tool

Selection

● New Selection ○ Select From Current Selection ○ Append To Current Selection

Selection Variable: LISA_CL Time:

Select All In Range 1 <= LISA_CL <= 1

Select All Undefined LISA_CL

Invert Selection

Add Neighbors To Selection Weights: PAcancerUpdated

Clear Selection

Assign Values to Currently Selected / Unselected

Add Variable

Target Variable: SELECT_CR Time:

☑ Selected = 1 ☑ Unselected = 0 (Leave empty for undefined values)

Apply

The data table will now highlight all tracts where **[LISA_CL] = 1**.

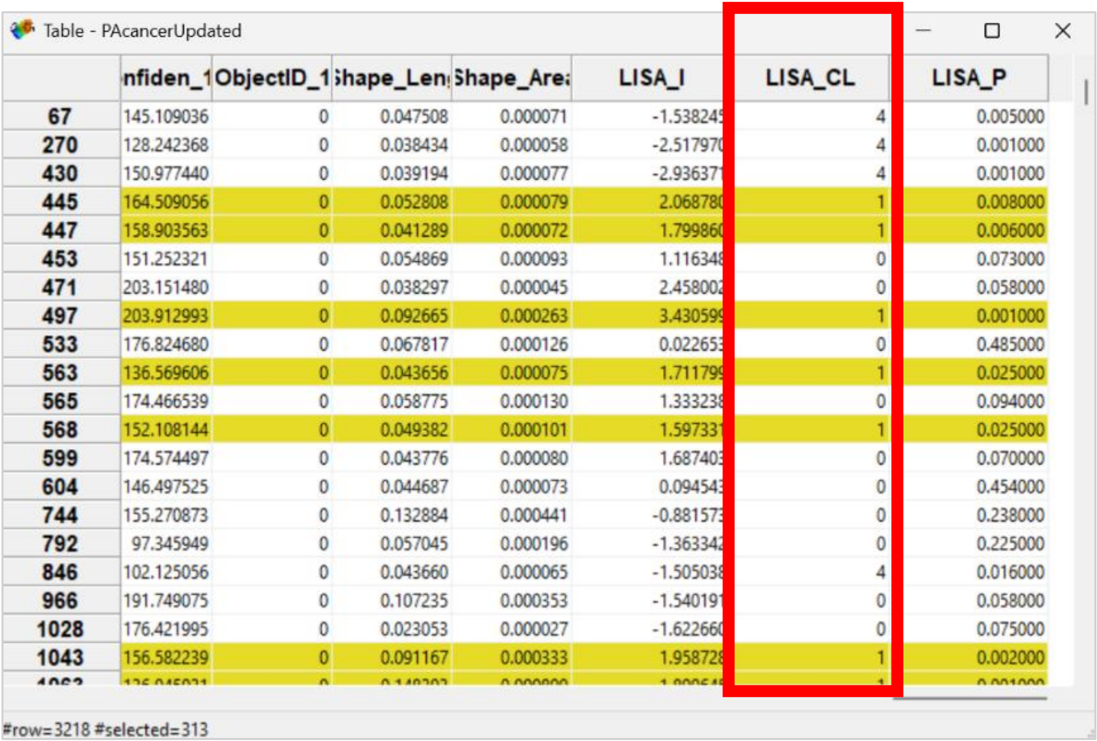


Table - PAcancerUpdated

	nfiden_1	ObjectID_1	Shape_Len	Shape_Area	LISA_I	LISA_CL	LISA_P
67	145.109036	0	0.047508	0.000071	-1.538245	4	0.005000
270	128.242368	0	0.038434	0.000058	-2.517970	4	0.001000
430	150.977440	0	0.039194	0.000077	-2.936371	4	0.001000
445	164.509056	0	0.052808	0.000079	2.068780	1	0.008000
447	158.903563	0	0.041289	0.000072	1.799860	1	0.006000
453	151.252321	0	0.054869	0.000093	1.116348	0	0.073000
471	203.151480	0	0.038297	0.000045	2.458002	0	0.058000
497	203.912993	0	0.092665	0.000263	3.430599	1	0.001000
533	176.824680	0	0.067817	0.000126	0.022653	0	0.485000
563	136.569606	0	0.043656	0.000075	1.711799	1	0.025000
565	174.466539	0	0.058775	0.000130	1.333238	0	0.094000
568	152.108144	0	0.049382	0.000101	1.597331	1	0.025000
599	174.574497	0	0.043776	0.000080	1.687403	0	0.070000
604	146.497525	0	0.044687	0.000073	0.094543	0	0.454000
744	155.270873	0	0.132884	0.000441	-0.881573	0	0.238000
792	97.345949	0	0.057045	0.000196	-1.363342	0	0.225000
846	102.125056	0	0.043660	0.000065	-1.505038	4	0.016000
966	191.749075	0	0.107235	0.000353	-1.540191	0	0.058000
1028	176.421995	0	0.023053	0.000027	-1.622660	0	0.075000
1043	156.582239	0	0.091167	0.000333	1.958728	1	0.002000

#row=3218 #selected=313

6. Group the selected results at the top of the screen as we did before, by right clicking any of the column headers and selecting the "**Move Selected to Top**" option.

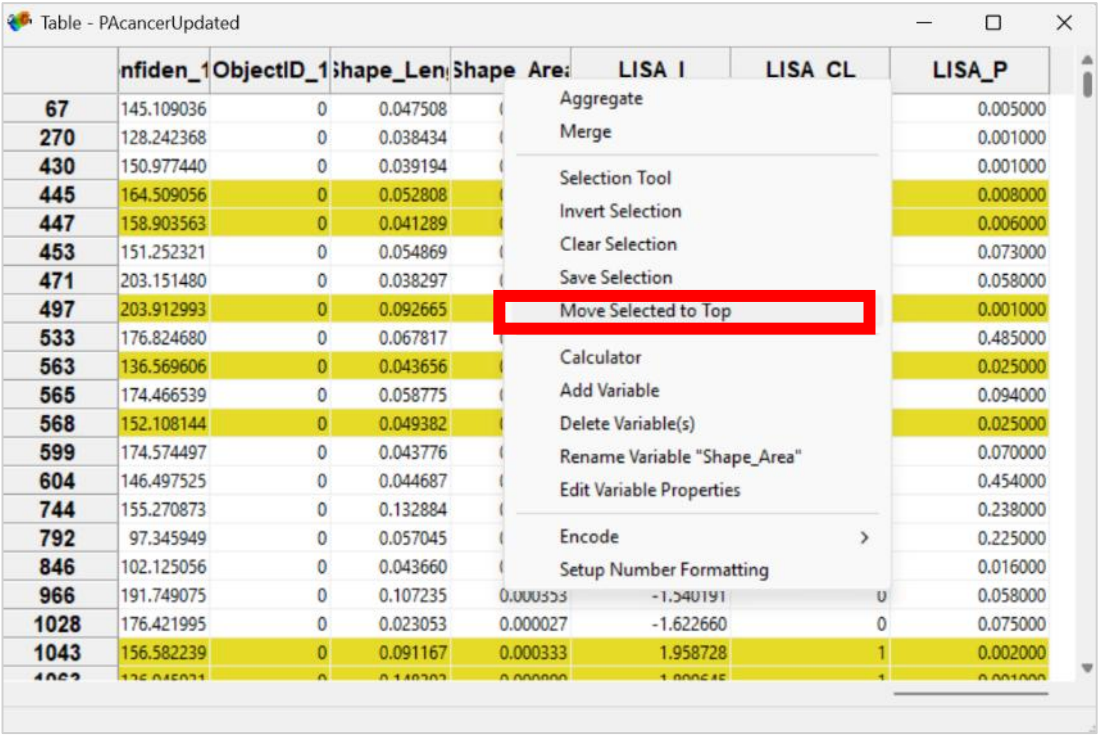


Table - PAcancerUpdated

	nfiden_1	ObjectID_1	Shape_Len	Shape_Area	LISA_I	LISA_CL	LISA_P
67	145.109036	0	0.047508				0.005000
270	128.242368	0	0.038434				0.001000
430	150.977440	0	0.039194				0.001000
445	164.509056	0	0.052808				0.008000
447	158.903563	0	0.041289				0.006000
453	151.252321	0	0.054869				0.073000
471	203.151480	0	0.038297				0.058000
497	203.912993	0	0.092665				0.001000
533	176.824680	0	0.067817				0.485000
563	136.569606	0	0.043656				0.025000
565	174.466539	0	0.058775				0.094000
568	152.108144	0	0.049382				0.025000
599	174.574497	0	0.043776				0.070000
604	146.497525	0	0.044687				0.454000
744	155.270873	0	0.132884				0.238000
792	97.345949	0	0.057045				0.225000
846	102.125056	0	0.043660				0.016000
966	191.749075	0	0.107235	0.000353	-1.540191	0	0.058000
1028	176.421995	0	0.023053	0.000027	-1.622660	0	0.075000
1043	156.582239	0	0.091167	0.000333	1.958728	1	0.002000

The **Data Table** will now show all of the highlighted tracts where **[LISA_CL] = 1 (high-high tracts)** at the top of the table.

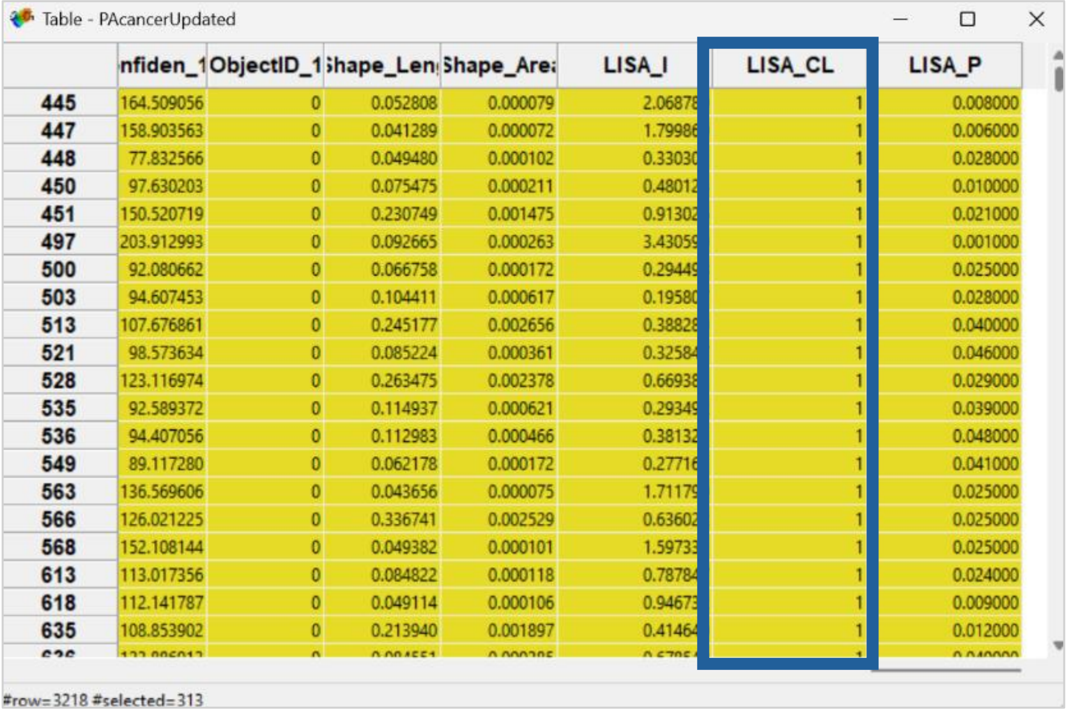
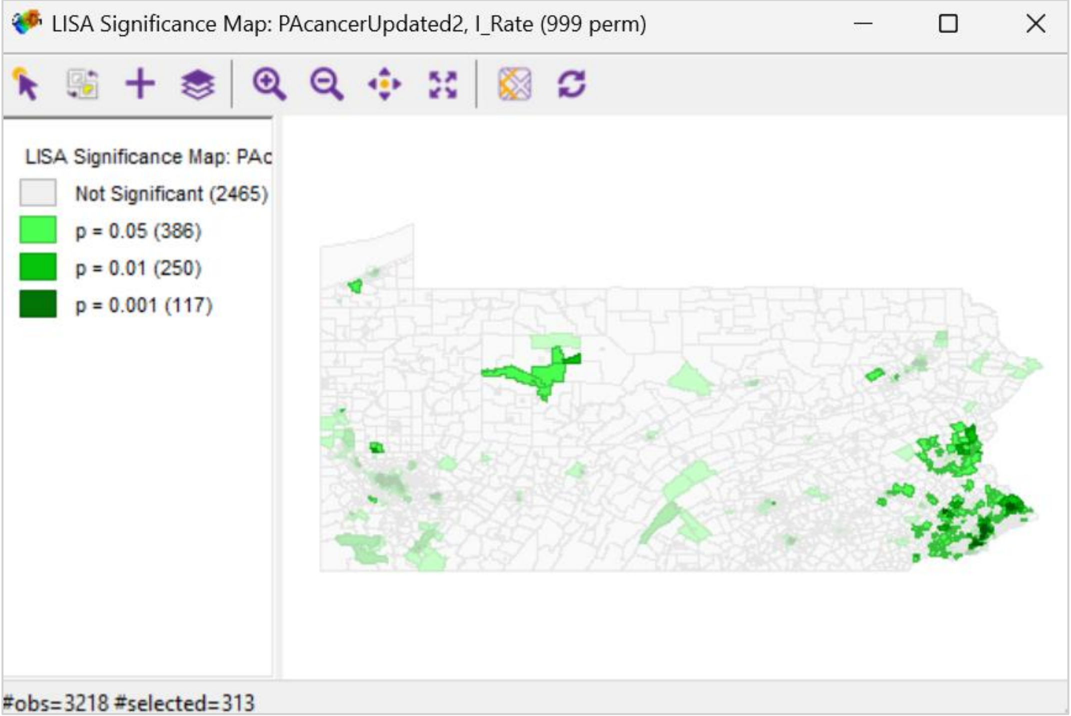


Table - PAcancerUpdated

	nfiden_1	ObjectID_1	;hape_Len	;hape_Are:	LISA_I	LISA_CL	LISA_P
445	164.509056	0	0.052808	0.000079	2.06878	1	0.008000
447	158.903563	0	0.041289	0.000072	1.79986	1	0.006000
448	77.832566	0	0.049480	0.000102	0.33030	1	0.028000
450	97.630203	0	0.075475	0.000211	0.48012	1	0.010000
451	150.520719	0	0.230749	0.001475	0.91302	1	0.021000
497	203.912993	0	0.092665	0.000263	3.43059	1	0.001000
500	92.080662	0	0.066758	0.000172	0.29449	1	0.025000
503	94.607453	0	0.104411	0.000617	0.19580	1	0.028000
513	107.676861	0	0.245177	0.002656	0.38828	1	0.040000
521	98.573634	0	0.085224	0.000361	0.32584	1	0.046000
528	123.116974	0	0.263475	0.002378	0.66938	1	0.029000
535	92.589372	0	0.114937	0.000621	0.29349	1	0.039000
536	94.407056	0	0.112983	0.000466	0.38132	1	0.048000
549	89.117280	0	0.062178	0.000172	0.27716	1	0.041000
563	136.569606	0	0.043656	0.000075	1.71179	1	0.025000
566	126.021225	0	0.336741	0.002529	0.63602	1	0.025000
568	152.108144	0	0.049382	0.000101	1.59733	1	0.025000
613	113.017356	0	0.084822	0.000118	0.78784	1	0.024000
618	112.141787	0	0.049114	0.000106	0.94673	1	0.009000
635	108.853902	0	0.213940	0.001897	0.41464	1	0.012000

#row=3218 #selected=313

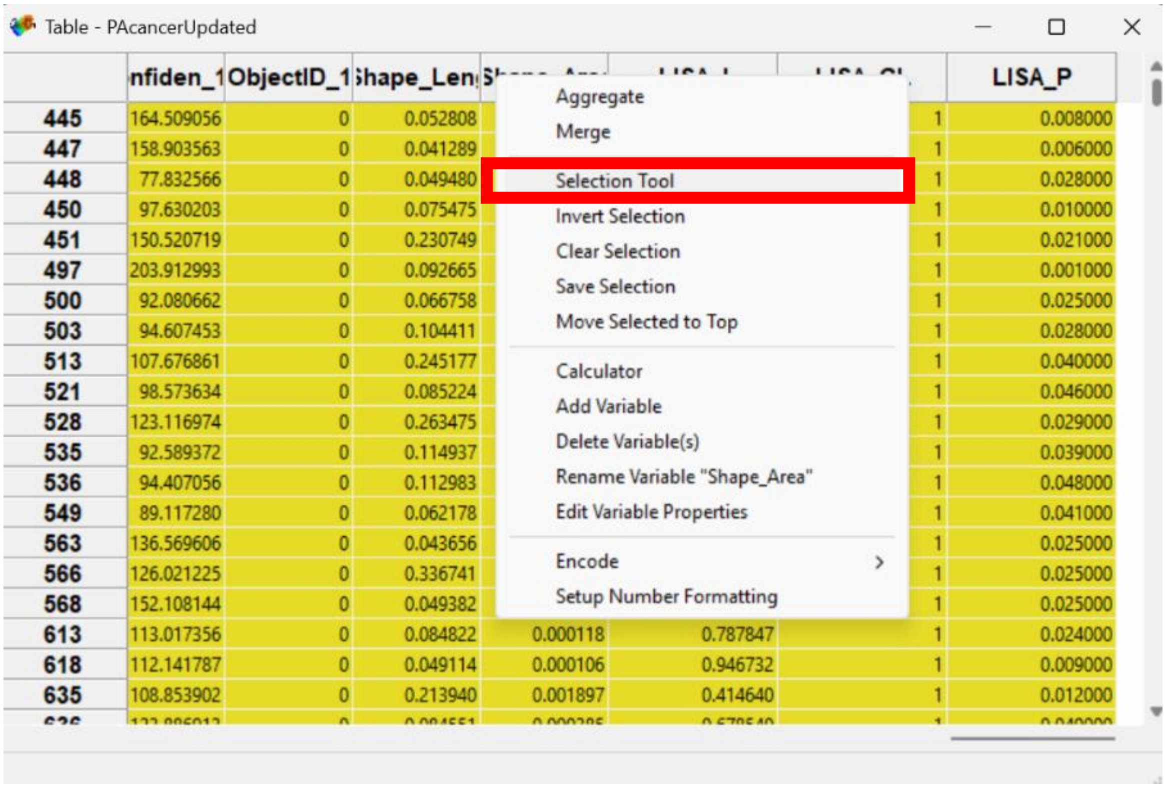
Our maps will also be filtered to these tracts as well, for example here are the **313 high-high tracts** highlighted on the **Significance Map**:



LISA Significance Map: PAcancerUpdated2, I_Rate (999 perm)

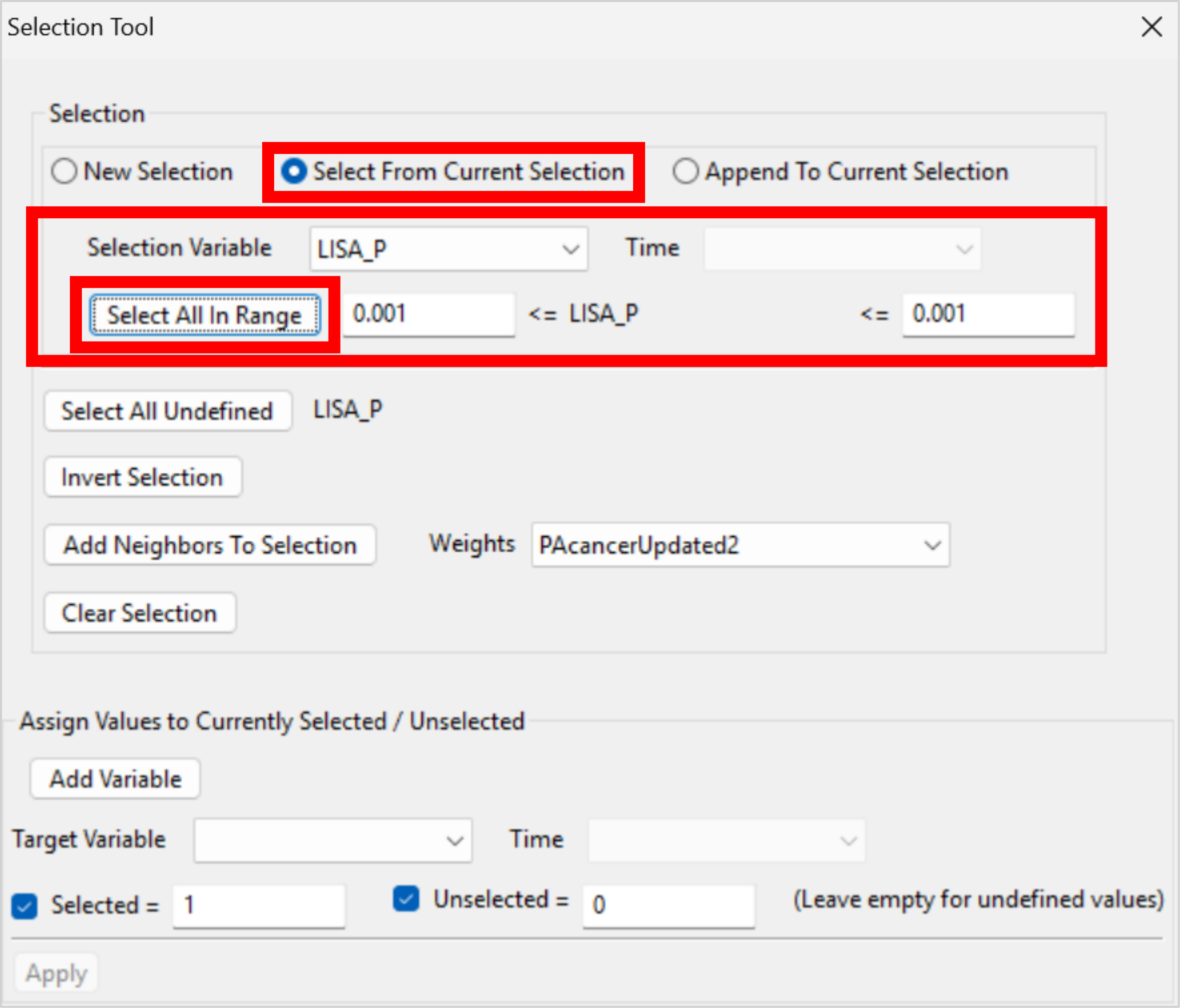
#obs=3218 #selected=313

7. Let’s further filter the **Data Table** to highlight only the **313 high-high tracts** that have the **most statistically significant clustering (i.e., p=0.001)**, by going back to the **Data Table**, right clicking any of the column headers in the table, and selecting the **Selection Tool** option from the menu to open it.



	nfiden_1	ObjectID_1	Shape_Len	Shape_Area	LISA_I	LISA_CL	LISA_P
445	164.509056	0	0.052808			1	0.008000
447	158.903563	0	0.041289			1	0.006000
448	77.832566	0	0.049480			1	0.028000
450	97.630203	0	0.075475			1	0.010000
451	150.520719	0	0.230749			1	0.021000
497	203.912993	0	0.092665			1	0.001000
500	92.080662	0	0.066758			1	0.025000
503	94.607453	0	0.104411			1	0.028000
513	107.676861	0	0.245177			1	0.040000
521	98.573634	0	0.085224			1	0.046000
528	123.116974	0	0.263475			1	0.029000
535	92.589372	0	0.114937			1	0.039000
536	94.407056	0	0.112983			1	0.048000
549	89.117280	0	0.062178			1	0.041000
563	136.569606	0	0.043656			1	0.025000
566	126.021225	0	0.336741			1	0.025000
568	152.108144	0	0.049382			1	0.025000
613	113.017356	0	0.084822	0.000118	0.787847	1	0.024000
618	112.141787	0	0.049114	0.000106	0.946732	1	0.009000
635	108.853902	0	0.213940	0.001897	0.414640	1	0.012000

8. As we want to build on the previous selection of **[LISA_CL] = 1** to include the most statistically significant tracts from [**LISA_P**], click the "**Select from Current Selection**" button at the top of the **Selection Tool**. Then, select the [**LISA_P**] variable as our **Selection Variable**, set the selection range to **"0.001 <= LISA_P >= 0.001"**, and click the "**Select All in Range**" button to the left of the **selection range** to apply the new highlight.



Selection Tool

Selection

○ New Selection ● Select From Current Selection ○ Append To Current Selection

Selection Variable LISA_P Time

Select All In Range 0.001 <= LISA_P <= 0.001

Select All Undefined LISA_P

Invert Selection

Add Neighbors To Selection Weights PAcancerUpdated2

Clear Selection

Assign Values to Currently Selected / Unselected

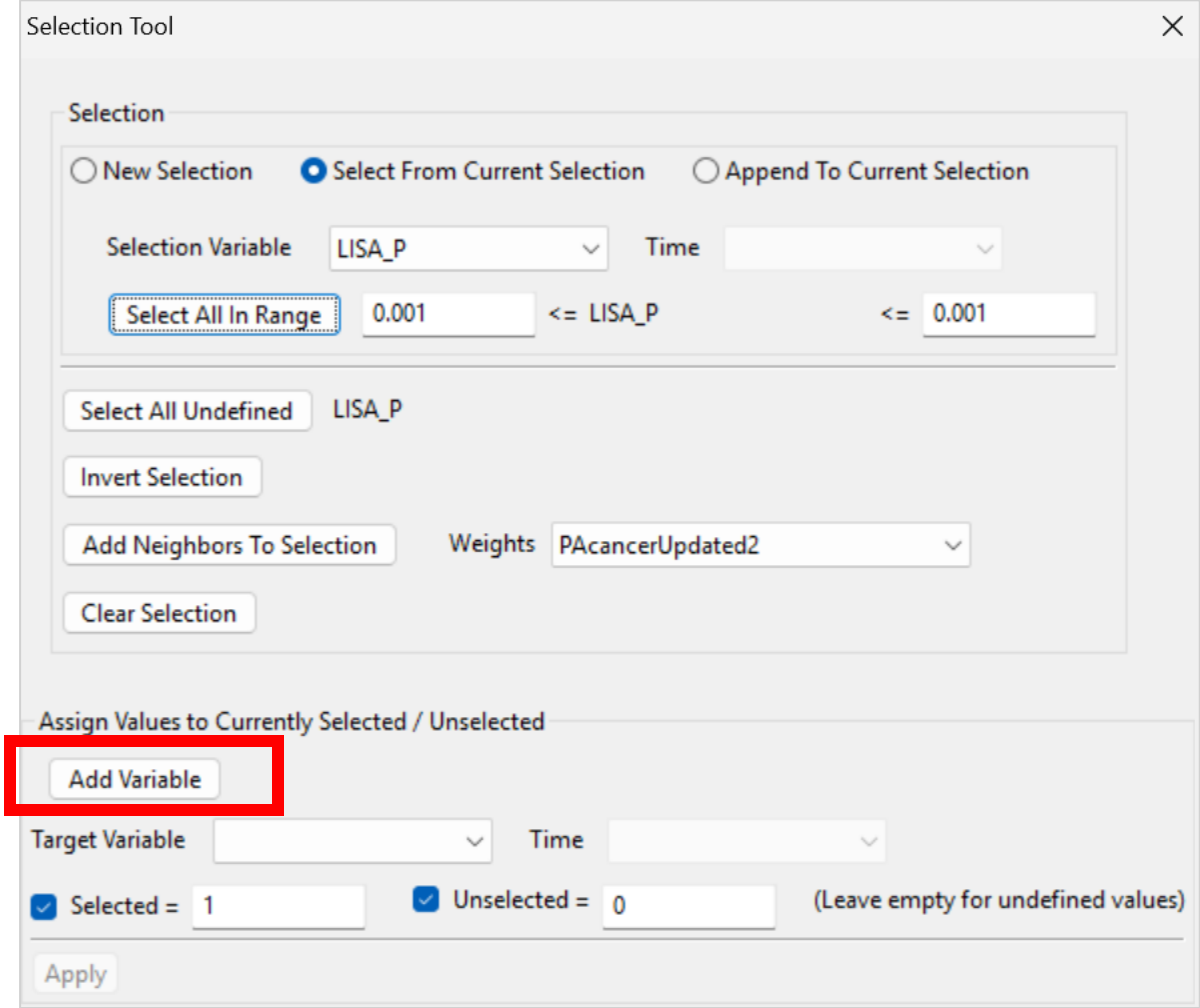
Add Variable

Target Variable Time

☑ Selected = 1 ☑ Unselected = 0 (Leave empty for undefined values)

Apply

9. Before we close the **Selection Tool**, let’s tell GeoDa to create a new variable, called “**SELECT_P**” in our **Data Table** from our current selection, by selecting the “**Add Variable**” button in the “**Assign Values to Currently Selected / Unselected**” section.



Selection Tool

Selection

○ New Selection ● Select From Current Selection ○ Append To Current Selection

Selection Variable: LISA_P

Time:

0.001 <= LISA_P <= 0.001

Select All Undefined LISA_P

Weights: PAcancerUpdated2

Assign Values to Currently Selected / Unselected

Target Variable:

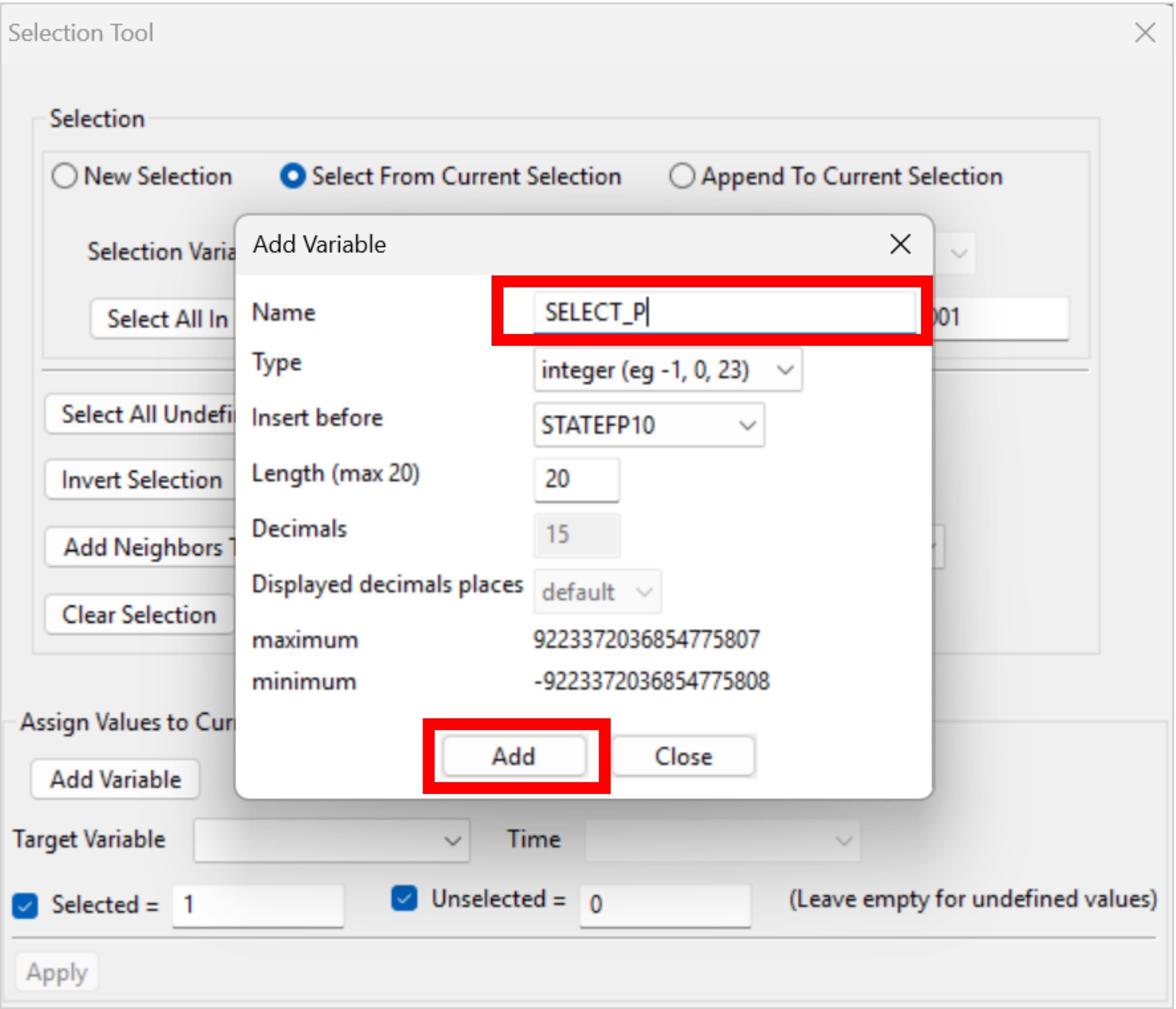
Time:

☑ Selected = 1

☑ Unselected = 0

(Leave empty for undefined values)

10. Name the new variable “**SELECT_P**” in the window that appears and click the “**Add**” button to add it to the **Data Table**.



Selection Tool

Selection

○ New Selection ● Select From Current Selection ○ Append To Current Selection

Add Variable

Name SELECT_P

Type integer (eg -1, 0, 23)

Insert before STATEFP10

Length (max 20) 20

Decimals 15

Displayed decimals places default

maximum 9223372036854775807

minimum -9223372036854775808

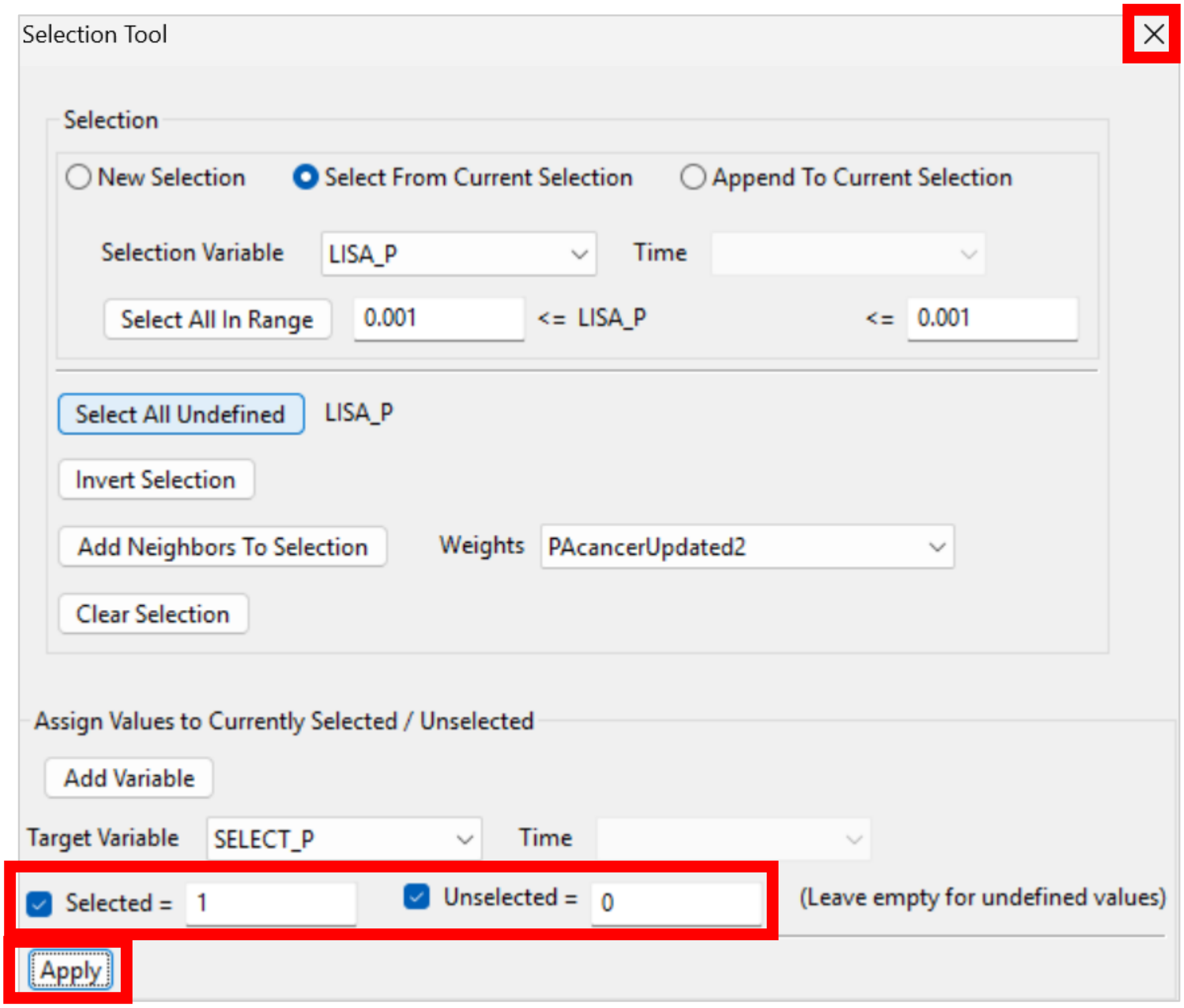
Assign Values to Cur

Target Variable

Time

☑ Selected = 1 ☑ Unselected = 0 (Leave empty for undefined values)

11. Make sure that the "**Selected**" and "**Unselected**" boxes are checked to create flags for selected records (i.e., tracts) in the new column, where **Selected tracts = 1** and **Unselected tracts = 0**. Then, click the "**Apply**" button to apply these changes to the table and maps. Close the **Selection Tool** window using the "**x**" in the top right corner.



Selection Tool

Selection

○ New Selection ● Select From Current Selection ○ Append To Current Selection

Selection Variable LISA_P Time

Select All In Range 0.001 <= LISA_P <= 0.001

Select All Undefined LISA_P

Invert Selection

Add Neighbors To Selection Weights PAcancerUpdated2

Clear Selection

Assign Values to Currently Selected / Unselected

Add Variable

Target Variable SELECT_P Time

☑ Selected = 1 ☑ Unselected = 0 (Leave empty for undefined values)

Apply

The **44 high-high tracts** with the **highest significance** are now highlighted in our table. The new [**SELECT_P**] variable has also been added as the first column to our table:

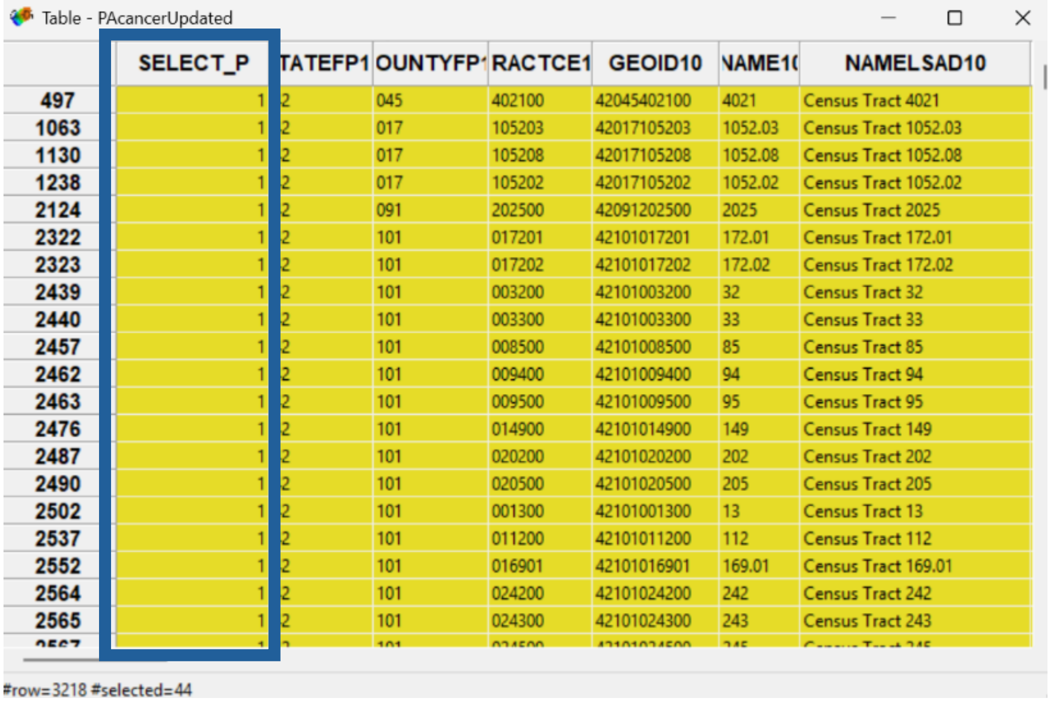
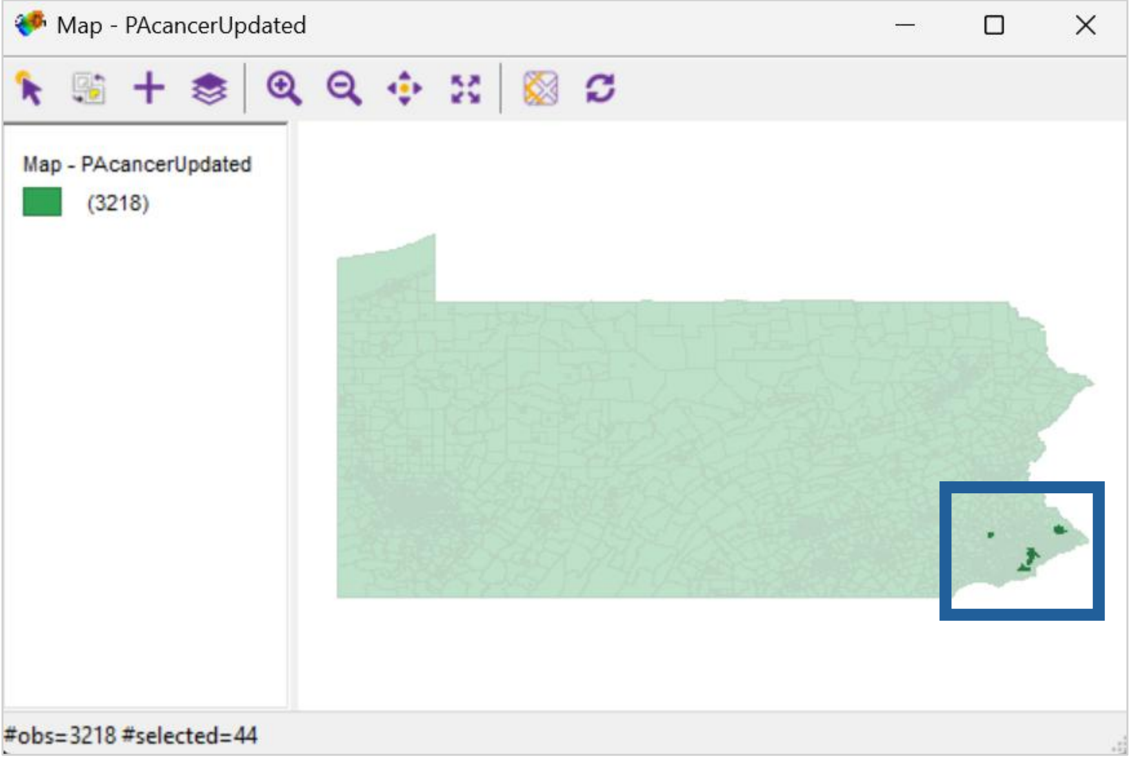


Table - PAcancerUpdated

	SELECT_P	TATEFP1	OUNTYFP1	RACTCE1	GEOID10	NAME1(	NAMELSAD10
497	1	2	045	402100	42045402100	4021	Census Tract 4021
1063	1	2	017	105203	42017105203	1052.03	Census Tract 1052.03
1130	1	2	017	105208	42017105208	1052.08	Census Tract 1052.08
1238	1	2	017	105202	42017105202	1052.02	Census Tract 1052.02
2124	1	2	091	202500	42091202500	2025	Census Tract 2025
2322	1	2	101	017201	42101017201	172.01	Census Tract 172.01
2323	1	2	101	017202	42101017202	172.02	Census Tract 172.02
2439	1	2	101	003200	42101003200	32	Census Tract 32
2440	1	2	101	003300	42101003300	33	Census Tract 33
2457	1	2	101	008500	42101008500	85	Census Tract 85
2462	1	2	101	009400	42101009400	94	Census Tract 94
2463	1	2	101	009500	42101009500	95	Census Tract 95
2476	1	2	101	014900	42101014900	149	Census Tract 149
2487	1	2	101	020200	42101020200	202	Census Tract 202
2490	1	2	101	020500	42101020500	205	Census Tract 205
2502	1	2	101	001300	42101001300	13	Census Tract 13
2537	1	2	101	011200	42101011200	112	Census Tract 112
2552	1	2	101	016901	42101016901	169.01	Census Tract 169.01
2564	1	2	101	024200	42101024200	242	Census Tract 242
2565	1	2	101	024300	42101024300	243	Census Tract 243
2567	1	2	101	024500	42101024500	245	Census Tract 245

#row=3218 #selected=44

All of our maps will also be filtered to the **44 high-high tracts with the highest significance**, which in this case appears to be centered in the Philadelphia area:



#obs=3218 #selected=44

Results Summary

Before we dive into the results, remember:

This analysis is only intended as a demonstration of how to use GeoDa for cancer cluster investigations and the sample results and findings presented as part of this tutorial should not be interpreted as real-world conclusions.

In this GeoDa tutorial, we highlighted a few areas with spatial clustering based on the rates of prostate cancer within the census tracts. We started by using choropleth mapping to visualize the prostate cancer rates within each of the census tracts in Pennsylvania. We then viewed the highest clustering of rates in the southeastern part of the state with clustering of other census tracts spread throughout the central and western portions of the state.

The **Global Moran's I statistic** showed us that some **weak positive spatial autocorrelation** was present. The local cluster significance map then highlighted areas where clustering of rates (i.e. similarity of tract level rates with neighbor rates) was statistically significant. These areas were primarily in the southeastern and southwestern parts of the state, with some additional clustering throughout the central part of the state.

Next, we used a cluster map to assess what kind of clustering pattern was present in each census tract. In this example, the cluster maps identified **High-High** clustering, in **313** of the **3,218** census tracts in Pennsylvania,, showing that census tracts with higher prostate cancer rates were near other census tracts with higher prostate cancer rates (based on the p-value selected for the analysis).

Note: *If your analysis did find clustering present, it does not necessarily mean there is a single cause or environmental cause for the pattern; it is information to consider for further investigation. Similarly, if your analysis did not find clustering, it does not necessarily mean that analyses are complete. Investigators can use findings, including examination of potential clusters with borderline (near threshold value) significance, from these analyses to inform on future work and investigations.*

References

GeoDa. Accessed: https://geodacenter.github.io/.

GeoDa Documentation. Accessed: https://geodacenter.github.io/documentation.html.

GeoDa, Moran's Scatterplot: Accessed: https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#moran-scatter-plot.

Pennsylvania, Age-Adjusted Rate of Prostate Cancer (Males Only) per 100,000 Population, Census Tract, 2010-2019 - https://ephtracking.cdc.gov/DataExplorer/?query=ab3f12d9-49b6-4e74-b6fe-e2d80ff0cb5d.

CDC, Centers for Disease Control and Prevention. Guidelines for Examining Unusual Patterns of Cancer and Environmental Concerns. December 2022. Accessed: https://www.cdc.gov/cancer-environment/php/guidelines/index.html.