

**CRP 4080: Introduction to Geographic Information Systems
Fall 2024**

Instructor: Wenzheng (wl563)

Lab TAs: Gauri Nagpal (gn247); Shubham Singh (ss3736); Anika Sinthy (ats243)

Location: Sibley 305, Barclay Gibbs Jones Computer Lab

Points Possible: 90

In today's Lab, you will learn how to:

1. Select features by their attributes (or querying data)
2. Export selected data into a new shapefile
3. Use styling to represent univariate and multivariate spatial relationships
4. Use exploratory data analysis to identify outliers
5. Normalize variables

Getting Started

1. Start by creating a folder (also known as a “directory”) in your personal drive.
2. Add the Lab 2 data folder from the course directory.
3. Open ArcGIS Pro,
 - a. start a new project in your folder
 - b. Add the data directory
 - c. Add COUNTIES (US counties) and STATES (all 50 states)

Selecting a Feature: NY State

Features need to be selected for several reasons. Selecting features is often required to manipulate GIS data. You may export selected features to create a subset of your data. You can also use feature selection to visually match the attribute table with the map view, since selected features can be seen on the map view with a different styling than the rest of the layers. If you are doing an operation that includes editing, extracting, or otherwise referring to one feature within a shapefile, you will likely need to Select that feature. Before we move on to the tutorial, let's take a second to learn how to Select a feature.

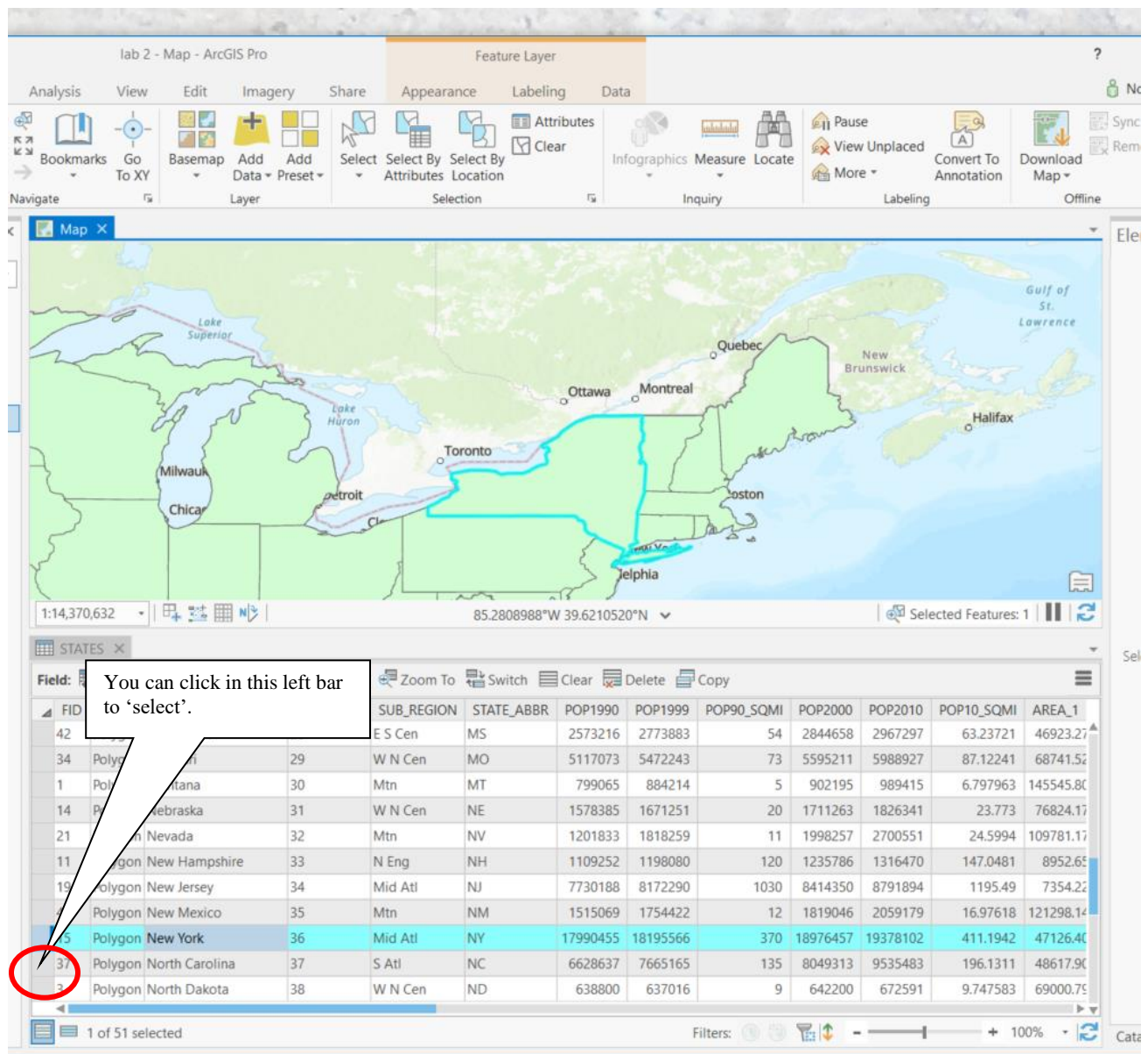


Figure 1. Selecting a feature from the Attribute Table.

- 1) Right click on the “STATES” layer and select **Open Attribute Table**.
- 2) Right click on column header for the “STATE_NAME” column, choose **Sort Ascending**.
- 3) Click in the left margin (“left bar”) of the row that contains New York (FID 15). See Figure 1. The left bar is the section of the Attribute Table window with no column header left of FID in Figure 1. Clicking in this area selects the feature. If you look at the map pane in your window, notice that the border of New York State is highlighted with the same color as it was highlighted in the attribute table (fluorescent blue).
- 4) To zoom to the selected feature: Right click on the left bar of the selected state (New York) in the attribute table.

- 5) Click **Zoom To**. This zooms the display to the extent of the feature, in this case, New York. Note that you may also do the same thing by **simply double clicking the “left bar”** of the row that contains New York.

Note that if Counties is listed above States in your Layers pane (which is likely, since they are in alphabetical order in your directory) you will see the outlines of US counties within the highlighted area of New York State (Figure 2). This is somewhat misleading because you have not selected the counties.

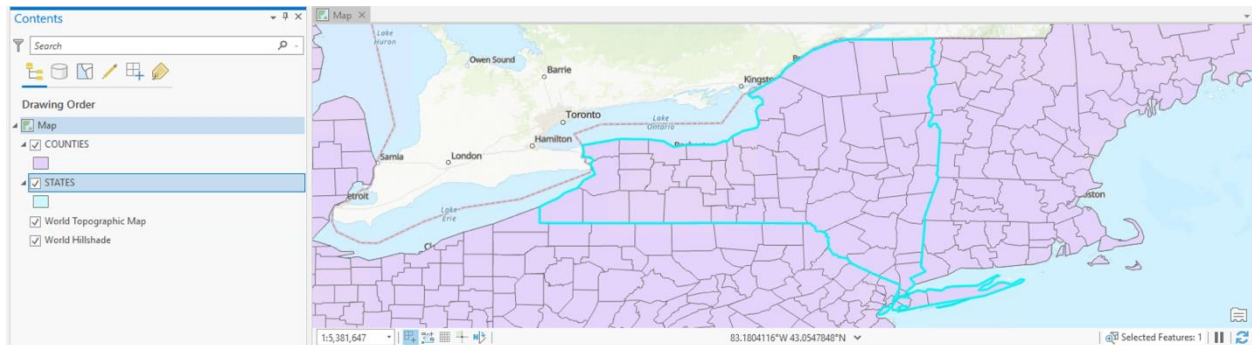




Figure 2. NY State is selected in the STATES layer, but it is drawn below the COUNTIES layer. Note that if we turn off the STATES layer, the outline of New York State disappears.

Toggling layers on and off is helpful for troubleshooting.

- 1) In your Layers pane, toggle on and off the Counties and States layers to see what happens. Note that when the STATES layer is turned off, the light blue selection outline disappears, even if New York state is still selected.
- 2) If you are still not sure what we mean, try selecting all or a few of the counties in New York and see what that looks like. Close the Attribute Table by clicking on the X in the upper right corner of the frame .

To unselect features, click **Clear Selected Features**.  in the Selection group.

You can also select and deselect features using the Select Features tooltip , and by clicking on whitespace in the Map pane. To select a feature this way:

- 1) Click on the tooltip to activate it,
- 2) then click on the feature you wish to Select. You can also draw a rectangle to select multiple features at one

Play around with this by selecting and deselecting 2 states and 2 counties. When you are done, clear all selections before we continue with the exercise.

Part 1: Creating Layers of New York States and Counties

We will now create a separate layer of only New York State and New York counties. Creating new layers limited to your study area minimizes the RAM, or memory required for processing. Extracting the study area (as this procedure is often called) is a simple way to prepare data for analysis in ArcGIS¹.

- 1) Select New York State, as we did above. With New York State selected, right-click on the States layer in the table of content.
- 2) Right-click on the States layer in the table of contents (or Layers pane), choose **Data** and click **Export features** (see Figure 2).

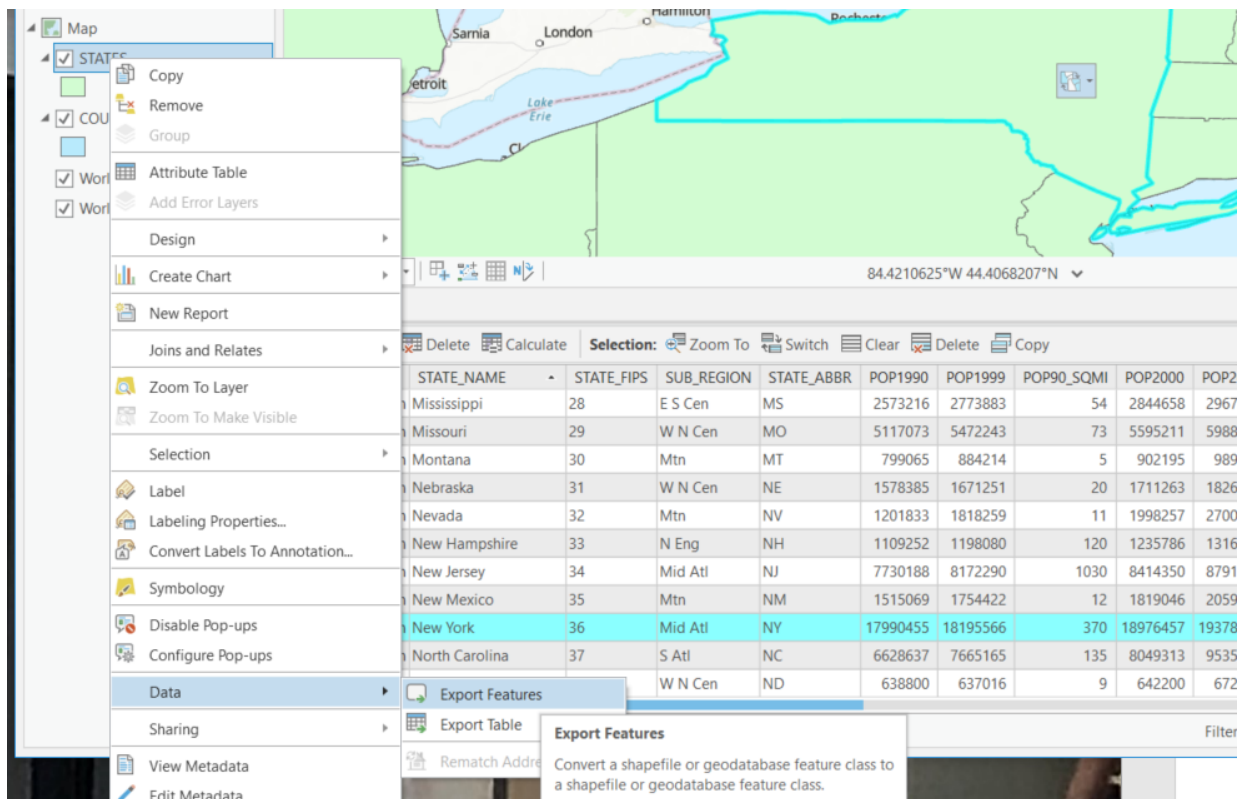


Figure 3. Export Data to create new layer.

- 3) A dialog box to 'Export Features' should appear (Figure 4).
 - Choose an output location where is easy for you to find the output data, e.g., the Lab2 output folder you just created.
 - Give the output the name 'nystate' (see **Error! Reference source not found.**). This tool creates a new layer only of the selected features.

¹ More advanced users often skip this step and use other tools or databases to limit operations, which also makes file management more streamlined. We may demonstrate these in labs later in the semester.

- click OK. You will see that a new layer, nystate, is added to the table of contents and a layer of only NY State is added to the view. The file is saved in the location you designated in the previous step, and you will later add this layer to a different view.

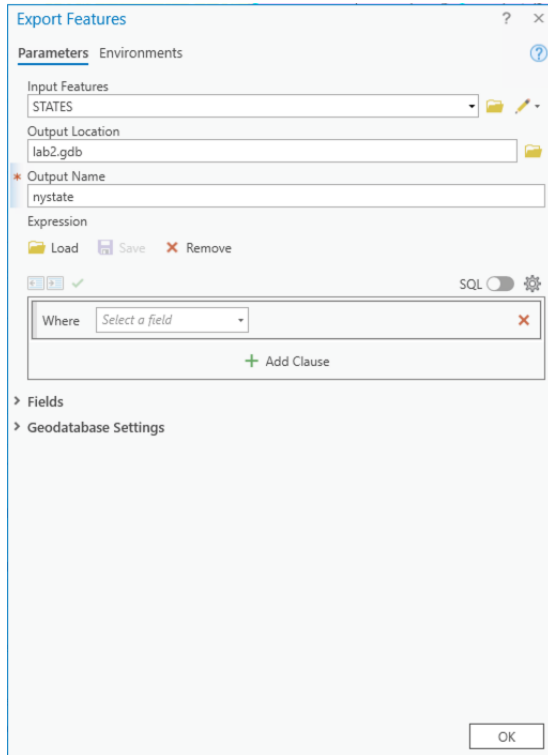


Figure 4. Create a new layer named “nystate” in the project geodatabase

- 4) Click **Clear Selected Features**. This unselects any features you have selected. This is important, because it prevents you from doing operations on features you don’t intend to alter. To clear selected features, you must have the main window activated. You may not be able to clear selected features if your Attribute Table window is open and activated.
- 5) Change the symbology of nystate so that it has no fill and a slightly thicker border than the counties borders.

Note that you can skip the selection step and export files directly to a new layer using the “where” options in the export dialog. Still, selections are useful for visually seeing the results of a “where” clause and verifying selections before committing them to a file.

Selecting by Location and Creating a Layer of New York State Counties

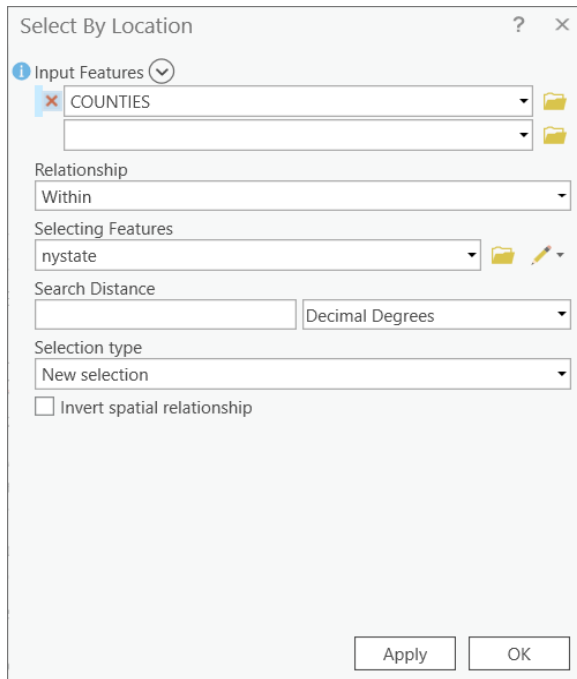



Figure 5. Select by Location Dialog Box.

1) Click **Select By Location** in the Selection group. This brings up the Select by Location dialog (Figure 5).

- We are going to select all the counties that are contained in New York.
- Select **COUNTIES** as the input feature, '**Within**' as the relationship. And '**nystate**' as the selecting feature.
- Click Apply and OK
- Right Click on the Counties layer, choose **Open Attribute Table**. In the lower left portion of the table frame, click the on the '**Show selected records**' icon . This filters the table selected features. You can see that all counties of New York state are selected. Leaving the counties selected, close the Attribute Table.

- 2) Right Click on the **Counties** layer in the table of contents, choose **Data** and click **Export Features** as we did above. Name the new layer "**nycounties**".
- 3) As a point of comparison, you can explore different select by location relationships (for example, try selecting counties that are 'completely within' New York State. You will note that this produces a different selection pattern). You can also use using search distance measurements.
- 4) Clear Selected Features.
- 5) Change the symbology of **nycounties** so to have no fill and a thinner border than the **nystate** border. Turn off the original STATES and COUNTIES layer – we will no longer need these for this analysis.

Selecting features by attribute/Querying the data

Now we will run a query to find the number of counties with a population greater than 300,000. To begin, you should have a map document open with the "**nycounties**" and "**nystate**" layers in the table of contents. To do this,

- 1) Select the **Select by Attributes** option from the **Selection** group on the toolbar. A dialog box will appear. The dialog box uses a guided form to create SQL (structured query language) syntax, which is what most databases use.

- 2) In the dialog box, select nycounties for the input rows field and 'New selection' for the selection type.

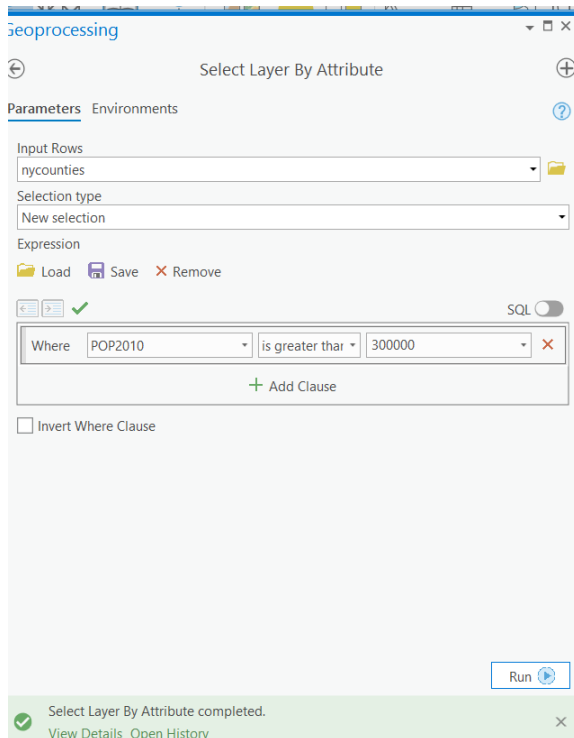



Figure 6. Select by Attributes Dialog Box.

- 3) Click "New Expression". Notice the 'Where' clause. A "WHERE clause" is part of a [SQL SELECT](#) statement that specifies which rows to select.
- 4) Create the following expression: Where POP2010 is greater than 300000 (See Figure 6).
- 5) Click **Apply** and close the dialog box. Both the attribute table and the layer for nycounties should now indicate 14 counties selected.
- 6) Open the Attribute Table and click the 'Show selected records'  as we did before. This helps you identify the names of selected counties.
- 7) Export the selected counties and create a new shapefile of NY counties with a population greater than 300,000.
- 8) Under the options menu in the attribute table, clear the selection and close the attribute table.
- 9) Save the ArcGIS project to your Lab2 folder.

Part 2: Univariate Statistical Analysis and Deliverables

Now we will create a series of maps that use data classification schemes to depict population density (per square mile, POP10_SQMI). You must submit a separate map layout for each classification method! However, you can create and reuse a single layout for each of them.

LAB 2 DELIVERABLES

Reminders before working on your assignment:

- Clean up maps by renaming labels.
- Choose shading patterns or color patterns that make intuitive sense (i.e., shading should get darker as population increases).
- In order to save your map as a jpg to be inserted into your Word document, use *share/export*, and save in your folder.
- Depict title, scale, north arrow, and metadata (e.g., data sources, **classification scheme**, authors) on your layout.

For each map requested below, create a new Map Pane and Map Layout BEFORE you change the styling. You can copy and paste from the Catalog View Pane to facilitate this.

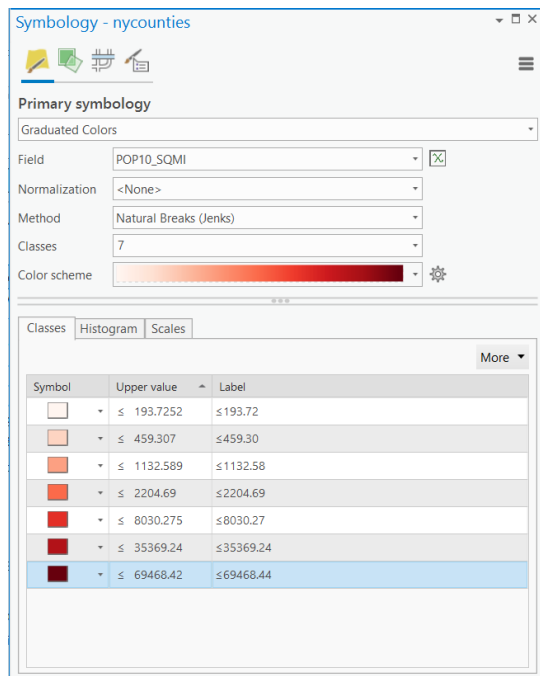
Keeping yourself organized and working methodically, rather than quickly, will be faster in the end.

Map #1: Basic classification

Lay out a map displaying NYS counties with a 2010 population more than 300,000.

Map #2: Natural Breaks

1. In the table of contents, right click on the layer name “nycounties” and select Symbology tab. Under primary symbology select **Graduated colors** (See **Error! Reference source not found.**).
2. Select “POP10_SQMI” as the field. (“POP10_SQMI” represents population density of 2010, measured in population of 2010 per square mile.)
3. Based on the population density classification field, use the default Classification of **Natural Breaks (Jenks)**. You can experiment with the number of classes, but it should be between 5 to 7 (This will be the rule for all thematic maps. No more than 7 classes!). Your window should be like **Error! Reference source not found.**




Error! Reference source not found.

Experiment with the number of classes and settle on which one you find appropriate. Note that we did not use “normalization.” We will explore this later.

Next, select an appropriate color scheme (Generally monochromatic works well for the given data range).

4. Under Advanced Symbol Options

, you can adjust the number of significant digits under ‘Label’.

5. Close the dialog box when you are satisfied with your creation.

Figure 7. Natural Breaks in Layer Properties Dialog Box.

Map 3: Equal Interval Map

1. Follow the same procedure as for Map 2, but this time use the equal interval classification method to create a new layout with all of the appropriate elements.
2. Click on the 'Histogram' tab and notice how the data classification intervals differ. This time, most of the counties fall into the lowest category.

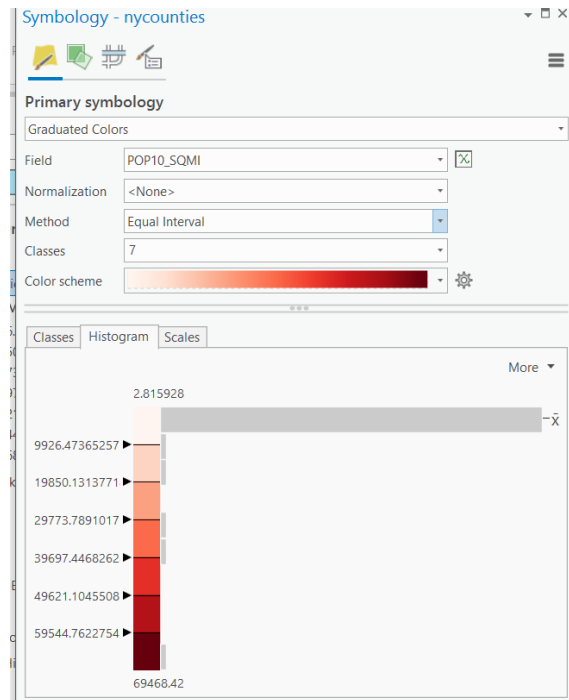


Figure 8. Equal Interval in Layer Properties Dialog Box.

3. Experiment with the number of classes and settle on which one you find appropriate.

- If you are unhappy with the color schemes available in ArcGIS, you can make your own and save it. The [Colorbrewer website](http://colorbrewer.org) is helpful for selecting colors for graduated and categorized maps.
- Also, note that you can change the position of the bars by clicking and dragging them (if you wanted to create a custom or user defined classification). Notice that as soon as you drag one of the triangles, the 'Method' switches to 'Manual Interval'
- Categorization is for display purposes only – it does not alter the underlying continuous data set. If we had 62 classes (one for each county), it would be very hard for the eye to distinguish between classes. **The human eye can only decipher at most 6-7 categories.**

Map 4: Quantile

Follow the same procedure but use the *quantile* classification method and create a new layout with all the appropriate elements. Experiment with the number of classes and settle on one you find appropriate. As you toggle between the different classification methods, please take a moment to notice the effect on the summary statistics, histogram, bars and break values.

Map 5: Standard Deviation

Follow the same procedure but use the *standard deviation* classification method and create a new layout with all the appropriate elements. Experiment with the number of classes and settle on one you find appropriate.

Reflection Question 1: Choose two maps and discuss the differences in patterns based on classification schemes. Which do you feel provides a better representation of population density? Provide a brief discussion of your reasoning.

Normalizing Data

When presenting data for enumerated units, such as political or administrative boundaries, researchers often normalize data. One way to accomplish this is by dividing each value by the column total (e.g. the proportion of population in each county). Another way is to divide it by another variable (e.g. the ratio of white to non-white population in each county). The population density variable we already used represents one method of normalizing data based on the area. Now let's explore other ways to normalize.


Map #6: Normalized Black population by State (Except Hawaii and Alaska)

- 1) Toggle off the 'nystate' and 'nycounties' layers you created and toggle on the 'STATES' and 'COUNTIES' layers from the original lab data.
- 2) Change the layer symbology to style with graduated colors (Figure 8):
 - Right click the layer 'STATES.'
 - In the Symbology tab, select Graduated colors.
 - Under the 'Field:' menu, select 'BLACK'.

This displays the total African American population for 2010 by state. Note the pattern – states with high populations tend to have a lot of African American people.

- 3) Now under the normalization menu, select 'POP2010' – notice the change in the pattern. Note that there are ratios rather than numbers in the "Label" columns. Experiment with the number of classes and the classification method.

We can change the ratio to a percentage.

- 4) Click on the 'Advanced Symbol options' .
- 5) Under Category, select percentage.
- 6) Check the box 'Numbers represent a fraction. Adjust it to show as a percentage' and reduce the decimal places to 2 (see box below).

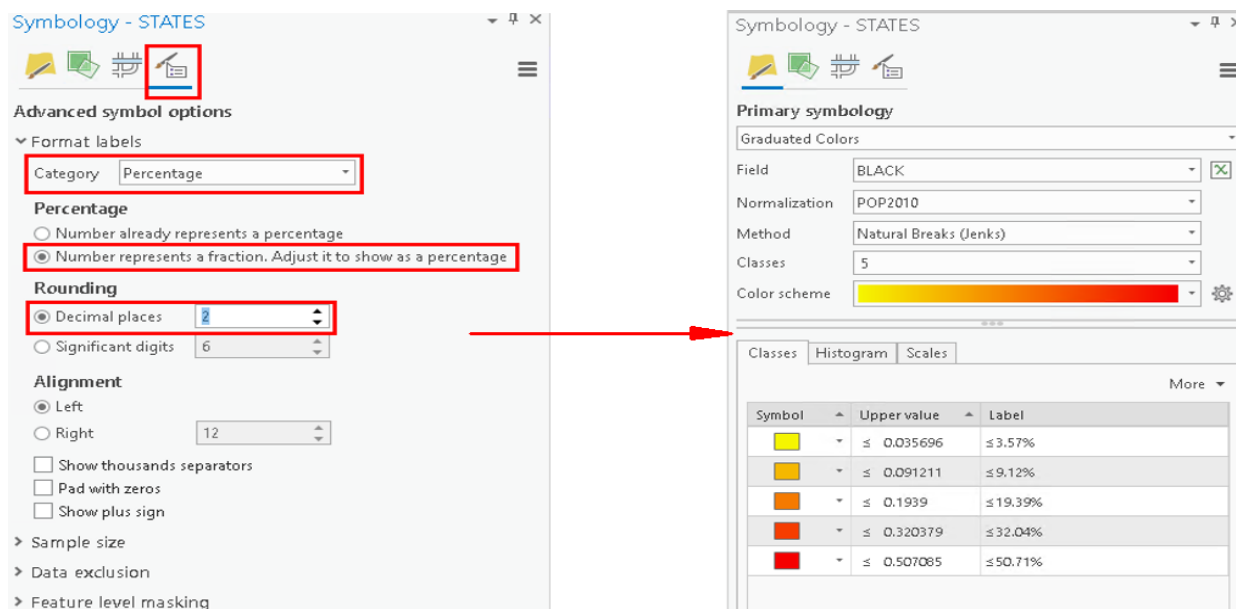


Figure 8. Normalizing Data and Representing it using Percentage.

This will give you a more straightforward classification that's easier to interpret.

Map #7 Multivariate state map depicting ratio of Black and Hispanic population to total

Let's use your normalized layer of African Americans to see if there is a relationship with Hispanic population.

We will explore 2 options to create a multivariate map:

1) Multiple symbols:

- Start with the normalized black population map.
- Copy and paste your states layer within the table of contents. Check that you have two states layers in the Contents Pane.
- Drag the new STATES layer to the top of the draw order, if necessary.
- Create a graduated symbol map of the population Hispanic, normalized by the 2010 population on the top STATES layer.
- Adjust the colors and the size as you see fit. Your map should look like Figure 9.

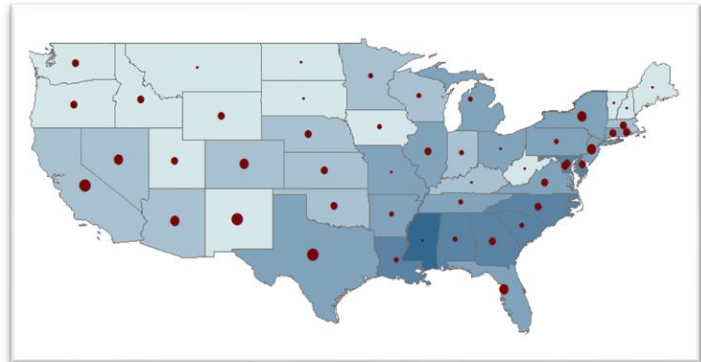



Figure 9. Multivariate map of normalized Black and Hispanic Populations.

2) Bivariate Colors:

- In the Contents pane, right-click the STATES layer and click Symbolology. The Symbolology pane appears.
- For the Primary Symbolology parameter, choose Bivariate Colors.
- For Field 1, choose population Hispanic, normalize accordingly
- For Field 2, choose Black Population, normalize accordingly
- Experiment with an appropriate color scheme.

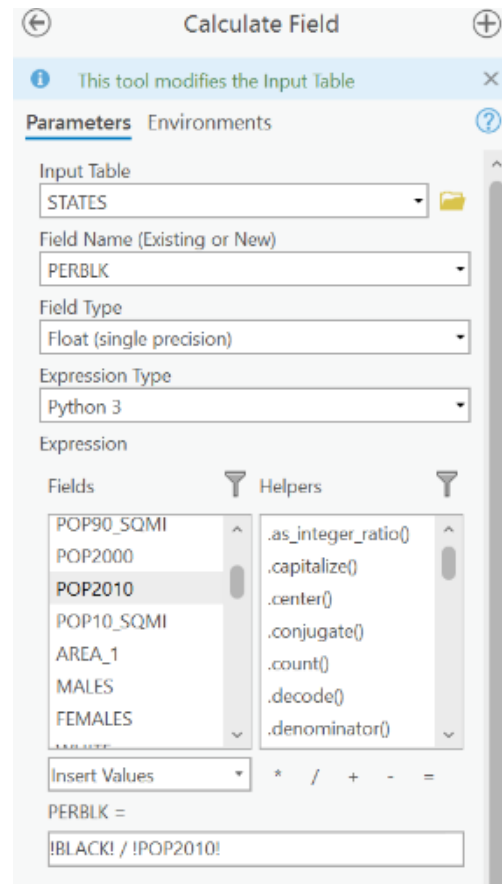
Reflection Question 2a. Do you see a relationship between Black and Hispanic state populations? Briefly explain how this relationship changes across space. Which version of these multivariate maps communicates this relationship better?

Creating a scatterplot: We can visually compare these 2 variables, but let's create a scatterplot to better quantify the relationship. In order to create a scatterplot, we first need to create new variables that contain normalized values (we cannot simply normalize the data visually as we did with the previous maps). In this case, we will create a new field with the percent of the population that is Black (PERBLK), and another field that depicts the percent of the population that is Hispanic (PERHISP)

- Click Tools  in the Analysis tab. The Geoprocessing pane appears.

- In the Geoprocessing pane, search for the *Calculate Field* (Data Management Tools) and open it. Choose STATES as the input table.
- For Field Name, type PERBLK. For Field Type, choose Float. Under Expression, in the Fields list, double-click BLACK. In the Expression text box, add a divide sign. Then add POP2010. The formula is $PERBLK = !BLACK! / !POP2010!$.
- Click RUN. If you scroll over in the attribute table, you should see the new field PERBLK (percent of the population that is black). Now create a similar field for Hispanic (PERHISP).
- Now to create the scatterplot, In the Contents pane, right click STATES. In the Create Chart, choose Scatter Plot. In the Chart Properties pane, for the X-axis Number parameter, choose PERBLK. For Y-axis Number, choose PERHISP.

Reflection Question 2b. What is the nature of the relationship? What is the R²? Does this confirm your visual interpretation?



Save your project.

Map 8: Excluding outliers from a classification

As mentioned in the lecture, sometimes the distribution of a variable of interest skews the classification. This can cause a map to conceal variation between many features. To address this, sometimes we need to change the classification scheme or omit some variables from it.

- 1) Create a proportional symbol map of US states using the field POP10_SQMI using any classification scheme (Figure 10). If the displayed symbol is too large, check the "Maximum size" option to adjust the maximum size of the symbol.
- What happened? Washington DC has a very high population density and is dominating the classification (Figure 10). We still cannot see the variation in many features, because the classification is being dominated by one outlier. In some cases, it is not as obvious, and we may need to undertake exploratory spatial data analysis to investigate our data.

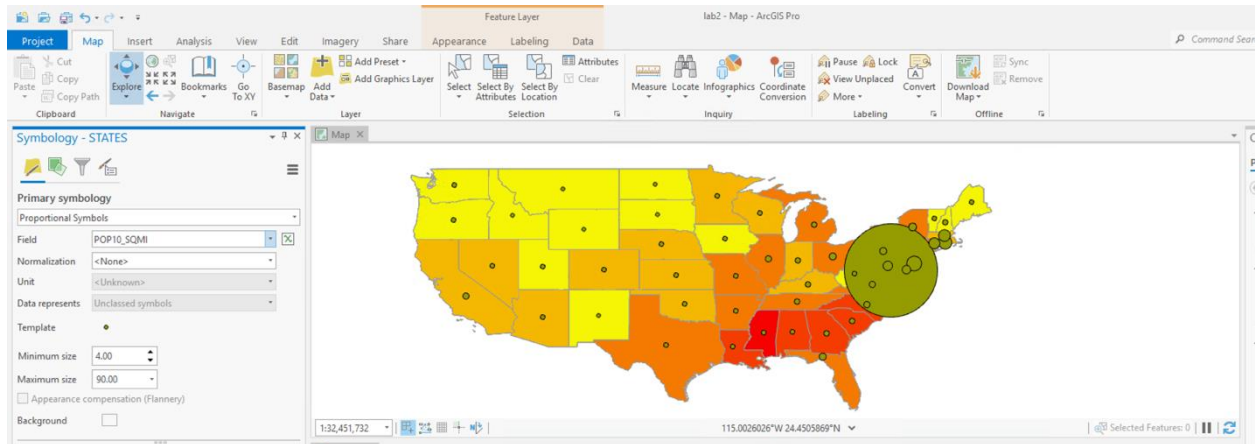


Figure 10. An easily-identifiable outlier can be seen in Washington DC. The Feature Layer tab is highlighted above the toolbar.

- 2) One such strategy is to examine a QQ plot (quantile, quantile):
 - In the Create Chart, under 'Create Chart', select QQplot.
 - Choose POP10_SQMI where it says "Compare the distribution".

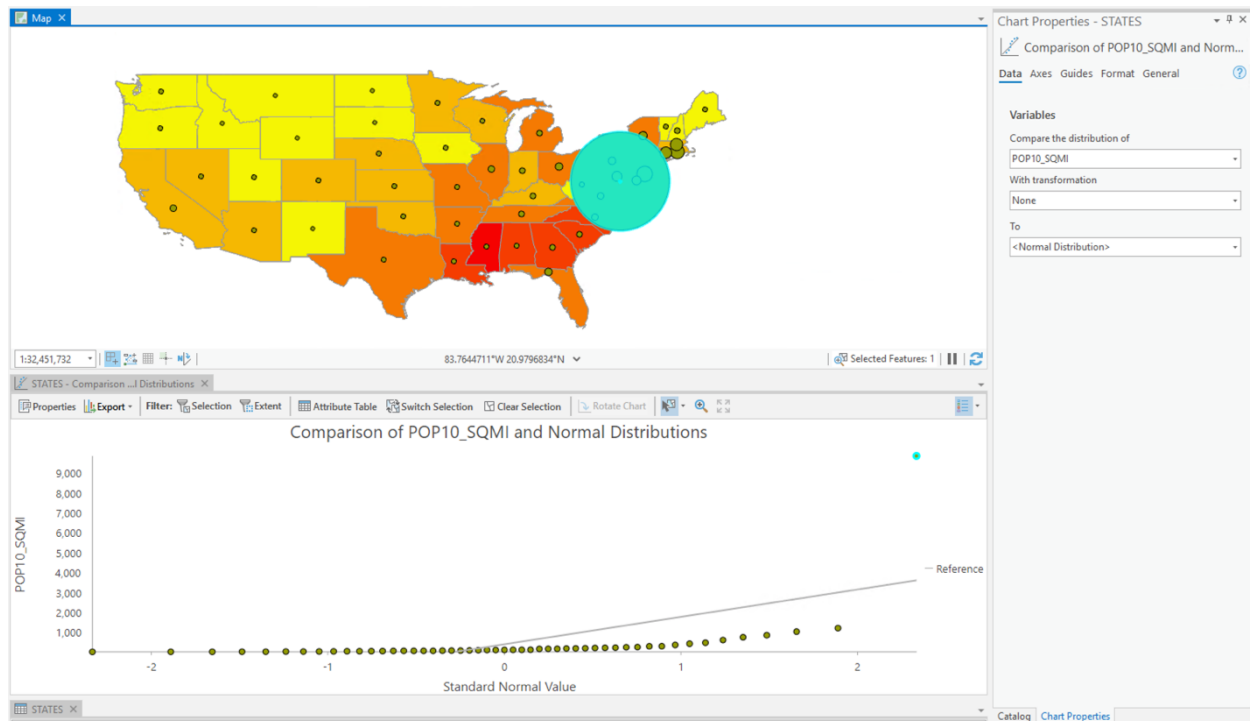


Figure 11. Normal QQPlot Chart with Washington, DC selected in the chart and map.

You will see the distribution of population density by state. Highlight the outlier in the upper right of the chart. You will see that it corresponds to Washington, DC (Figure 11). Note

how much the inclusion of DC is throwing off the distribution—the next highest value (New Jersey) is below the line that represents a ‘normal’ distribution.

- 3) Another strategy is to use a histogram.
 - Add another chart using the Chart Type, Histogram
 - Select POP10_SQMI as the Number. A histogram should appear at the bottom with a chart properties box that contains some additional summary statistics.
 - Examining the Statistics section of the Chart Properties Pane (Figure 12)

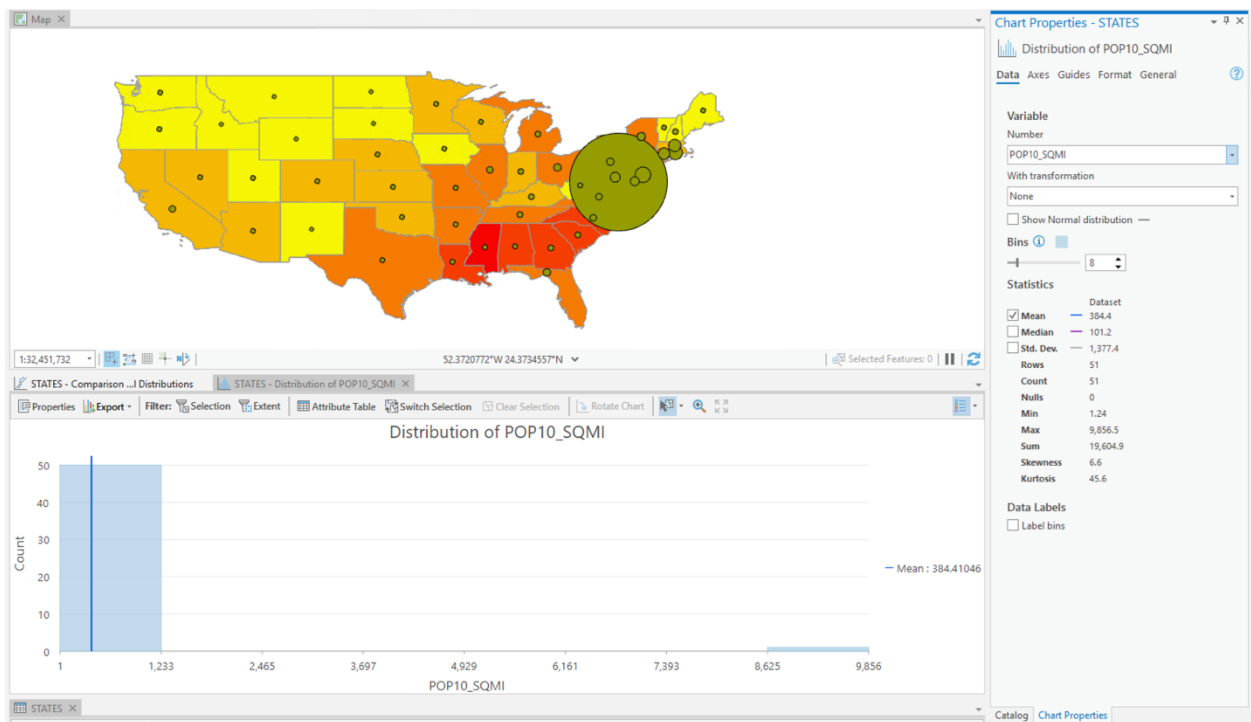




Figure 12. Histogram chart type; see the Statistics section of the Chart Properties Pane on the right side.

You will see several statistics in the Chart Properties Pane, described below:

- *Skewness* is a measure of the symmetry of a distribution. For symmetric distributions, the coefficient of skewness is zero. If a distribution has a long right tail of large values (i.e., Washington DC), it is positively skewed, and if it has a long-left tail of small values, it is negatively skewed.
- In addition, note that the mean (384) is larger than the median (101), also indicating a positively skewed distribution (the opposite would be true for a negatively skewed distribution).
- *Kurtosis* characterizes the relative “peakedness” or flatness of a distribution compared to the normal distribution. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution. Normal distributions produce a kurtosis statistic of about zero. A positive *Kurtosis* value indicates the possibility of a *leptokurtic* distribution (that is, too tall), a negative value indicates the possibility of a *platykurtic* distribution (that is, too flat). The existence of flat or peaked distributions as

indicated by the kurtosis statistic is important too, as it indicates violations of the assumption of normality that underlies many of the other statistics.

- 1) Let's exclude Washington, DC using a SQL query to create a better classification.
 - Find and click on Advanced Symbol options in the symbology pane,  and 
 - Click on Data Exclusion
 - Add an expression with a "where clause" that excludes the District of Columbia (not Washington).
 - Apply the changes

Examine the resulting effect. Does this solve the problem?

Map #8: Create a map layout of US states population density (using graduated color), sans DC.

LAB 2 DELIVERABLES

- **Map 1:** Map of NYS counties with a 2010 population greater than 300,000. (10 points)
- **Maps 2 - 5.** Each layout should be on its own page. (5 points each)
- **Reflection Question 1:** Choose two maps and discuss the differences in patterns based on classification schemes. Which do you feel provides a better representation of population density? Provide a brief discussion of your reasoning. (5 points)
- **Map 6:** Normalized Black population by State. Include a sentence or 2 about interesting spatial relationships you may notice. (Note: you only have to depict the continental US for your map!) (10 points)
- **Map 7:** Multivariate map depicting normalized Black and Hispanic population by state. Either approach (multiple symbols or Bivariate Colors) is fine. Include a line about which you thought more effectively conveyed the data (10 points)
- **Reflection Question 2a:** Do you see a relationship between Black and Hispanic state populations? Briefly explain how this relationship changes across space. Which version of these multivariate maps communicates this relationship better? **2b.** Having completed the scatterplot, what is the nature of the relationship? What is the R^2 ? Does this confirm your visual interpretation? (5 points)
- **Map 8:** Population Density 2010 map using graduated colors by state (sans Washington DC) (10 points)

Additional Homework Maps

MAP 9: Create a multiple attributes map of New York State counties displaying the % of housing units which are renter occupied (normalized to the total number of housing units) and the % population that has never been married, using "NMARRY00" (normalized to 1999 population). Justify your choice of classification scheme. Discuss any patterns you may notice. (10 points)

MAP 10: Create a map layout of the continental US by county depicting population density. Use whatever classification scheme and number of classes you wish and make any exclusions you think are appropriate. Justify your decision using the Normal QQ plot as well as the kurtosis and skewness statistics. Hint: as before, you will need to write a query using SQL, but keep in mind the principles of Boolean logic if you chose to exclude more than 1 county! (10 points)