Spatial Data Science MOOC

Exercise

Explore data using data visualization techniques

Section 1 Exercise 2

02/2020

Explore data using data visualization techniques

Time to complete
80 minutes

Introduction

Data visualization helps you digest information by using symbols to visually represent quantities and categories. You can quickly make comparisons and perceive relative proportions, patterns, relationships, and trends. Data visualization is important throughout the analysis process, from exploring your data, to interpreting your results, to communicating your findings. There are various data visualization techniques available in ArcGIS. In this exercise, you will use these techniques to explore your data and look for any interesting relationships that may be useful in a predictive analysis.

Exercise scenario

Because voting is voluntary in the United States, the level of voter participation (referred to as "voter turnout") has a significant impact on the election results and resulting public policy.

Modeling voter turnout, and understanding where low turnout is prevalent, can inform outreach efforts to increase voter participation. In this exercise, you will use various visualization techniques to explore relationships and patterns of voter turnout and to identify potential variables to use in your predictive analysis.

Step 1: Open an ArcGIS Pro project

To begin, you will open the ArcGIS Pro project package that you downloaded for the first exercise in this section.

a. Start ArcGIS Pro.

b. If necessary, sign in using the provided ArcGIS account.

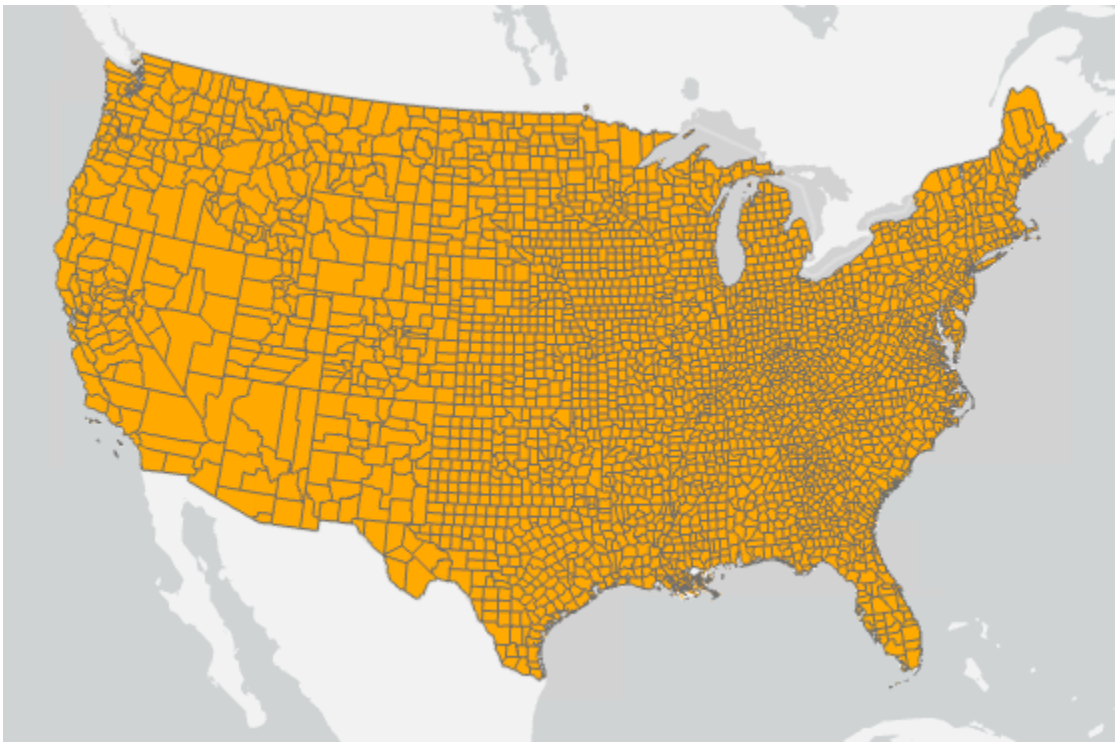
c. Under Open, click Open Another Project.

d. Browse to the Data Engineering and Visualization folder that you saved on your computer.

e. Click Data Engineering and Visualization.aprx and click OK.

f. In the Catalog pane, expand Maps.

g. Under Maps, right-click Data Visualization and choose Open.



A Data Visualization map tab opens to a gray basemap with a map layer that contains the 2016 election results for each county.

Step 2: Explore data attributes and visualize distributions

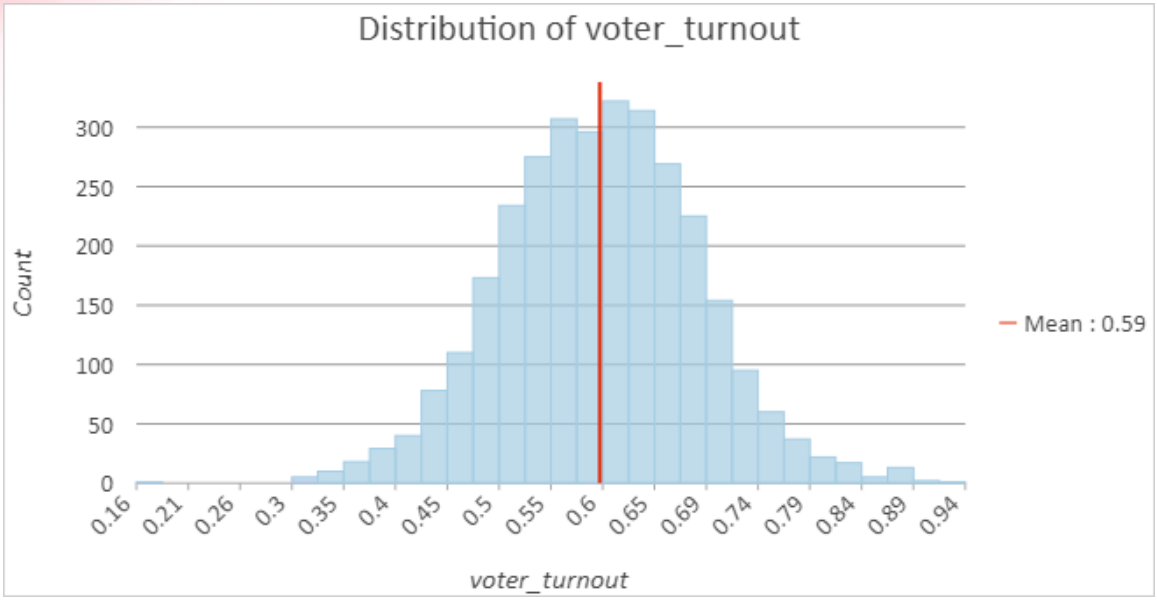
The 2016 election results layer (CountyElections2016) includes voter turnout and demographic variables geoenriched to the layer, including the variables added in the previous exercise.

a. In the Contents pane, right-click CountyElections2016 and choose Attribute Table.

The CountyElections2016 attribute table appears below the map.

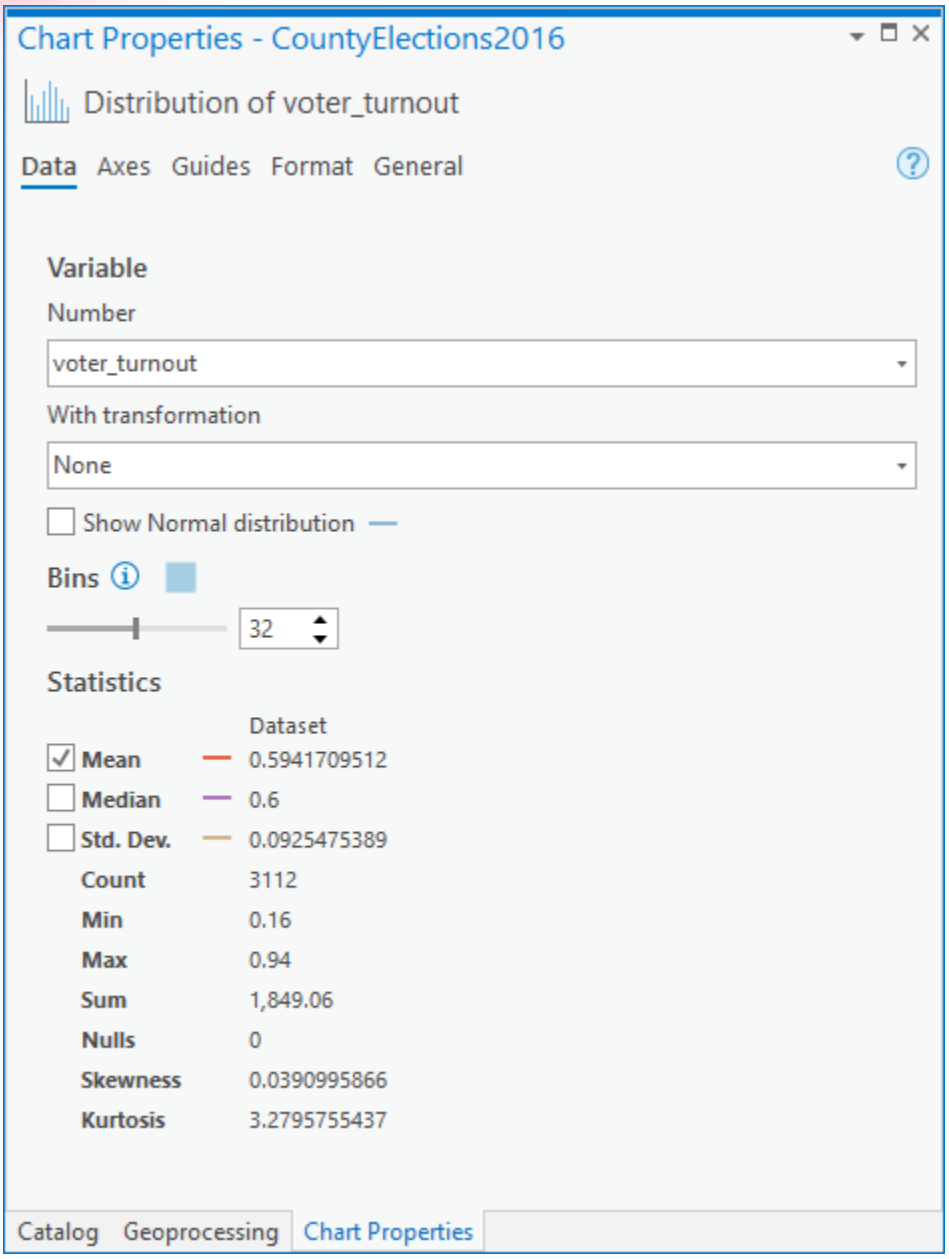
b. In the attribute table, scroll to the right to familiarize yourself with the variables included in the dataset.

c. In the attribute table, right-click the Voter_Turnout attribute column and choose Statistics.



A chart window displaying a histogram appears below the map, and the Chart Properties pane opens.

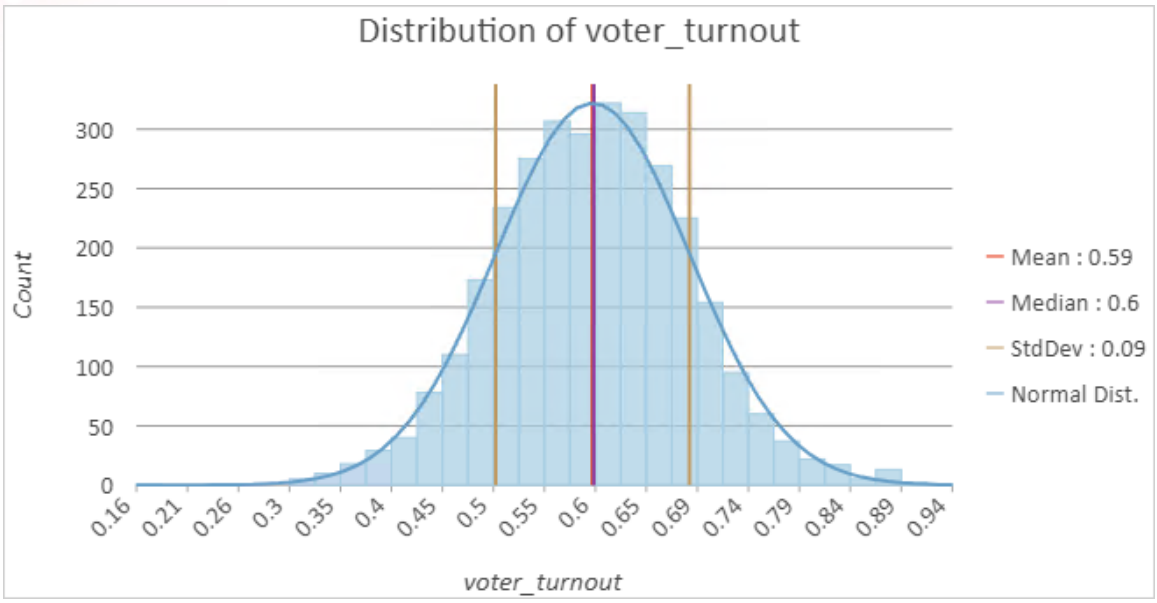
The histogram visually summarizes the distribution of voter turnout by measuring the frequency at which values appear in the dataset. The x-axis represents different ranges, or bins, of voter turnout values. The y-axis measures the number of counties with voter turnout values falling within each range. To learn more about histograms in ArcGIS Pro, see ArcGIS Pro Help: Histogram.



The histogram Chart Properties pane includes a list of statistics summarizing the Voter_Turnout variable. The mean county voter turnout value is about 0.59, with county values ranging from approximately 0.16 to approximately 0.94. By default, a line representing the mean value is overlaid on the histogram. You can add additional overlays from the Chart Properties pane.

d In the Chart Properties pane, under Variable, check the box for Show Normal Distribution.

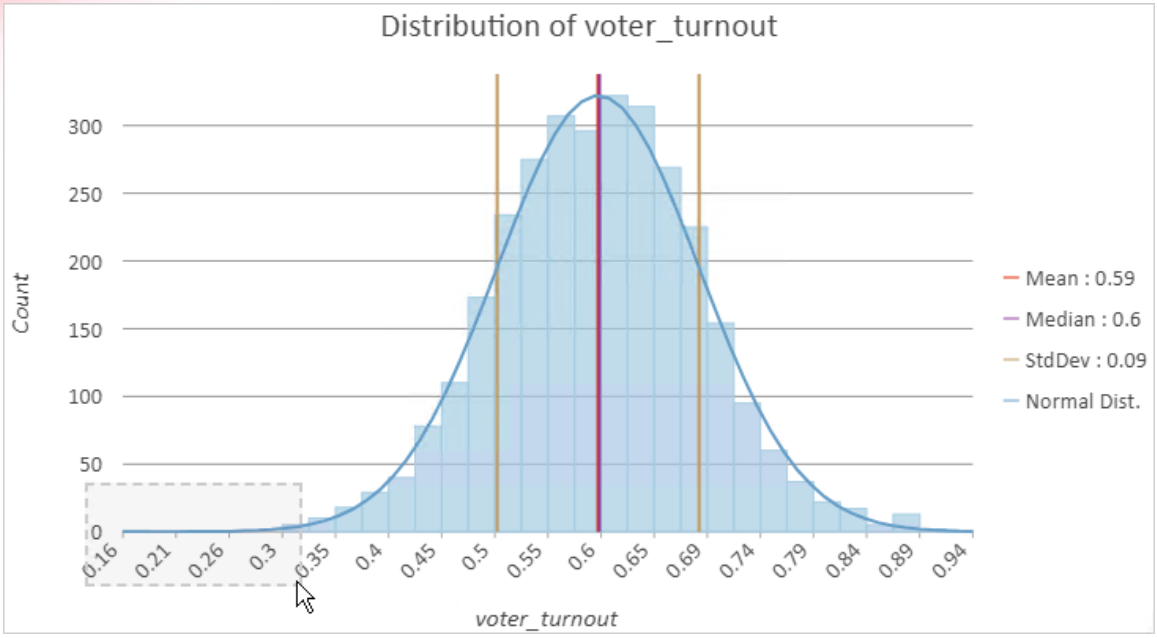
e In the Chart Properties pane, under Statistics, check the box next to Median and Std. Dev.



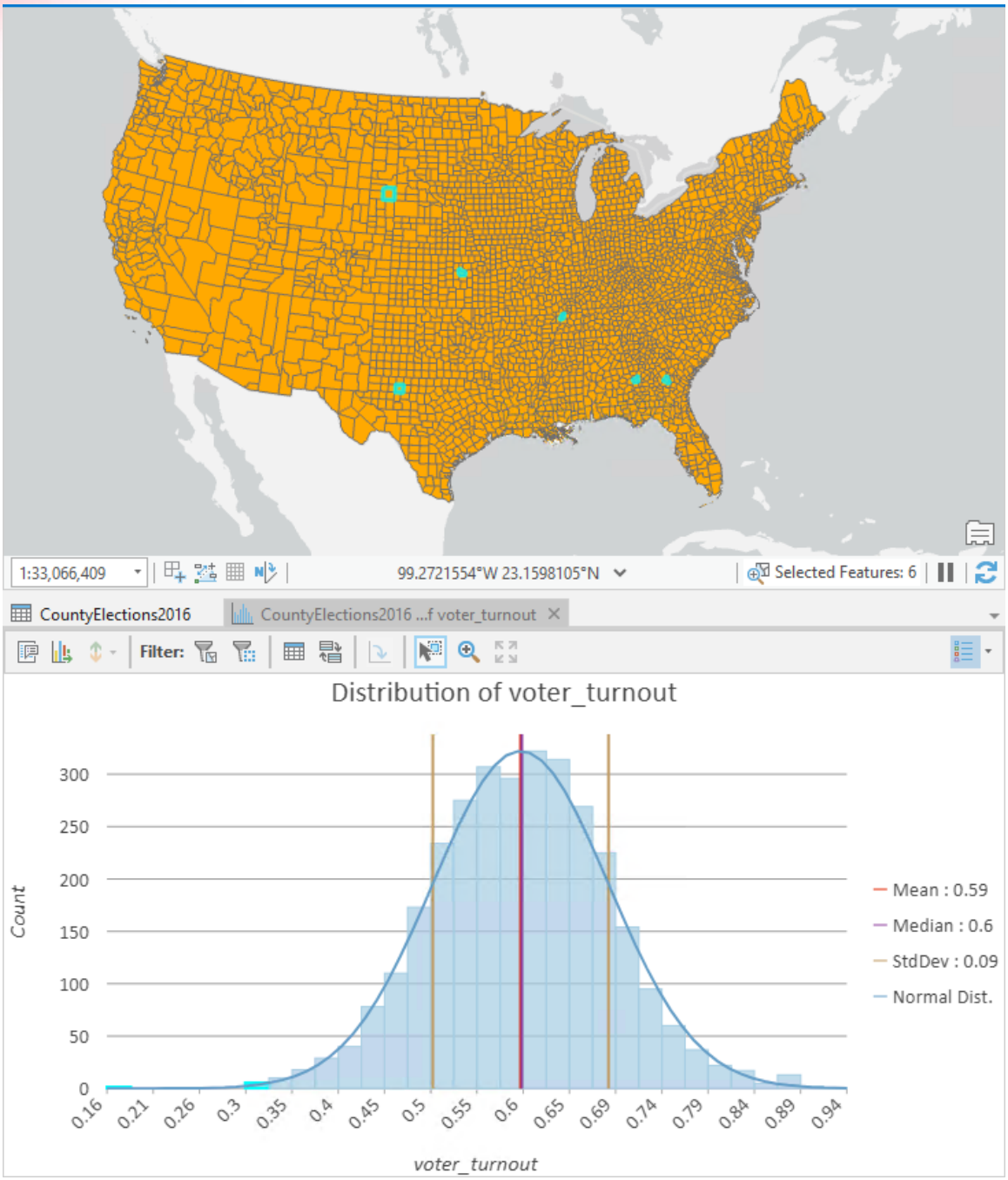
The overlays appear on the histogram. Using these overlays, you can see that the histogram has a fairly symmetrical bell shape with nearly identical mean and median values. This indicates that the Voter_Turnout variable is normally distributed. In a normal distribution, the mean, median, and mode values are equal. This means that most values fall near the average in the center of the distribution, with fewer and fewer values appearing as you move farther from the center into the left and right tails.

You can also use the histogram to interactively select features on the map based on their voter turnout values.

f In the chart window, click and drag to select the bins with the lowest voter turnout values in the left tail of the histogram.



A gray box appears to indicate which bins will be selected.

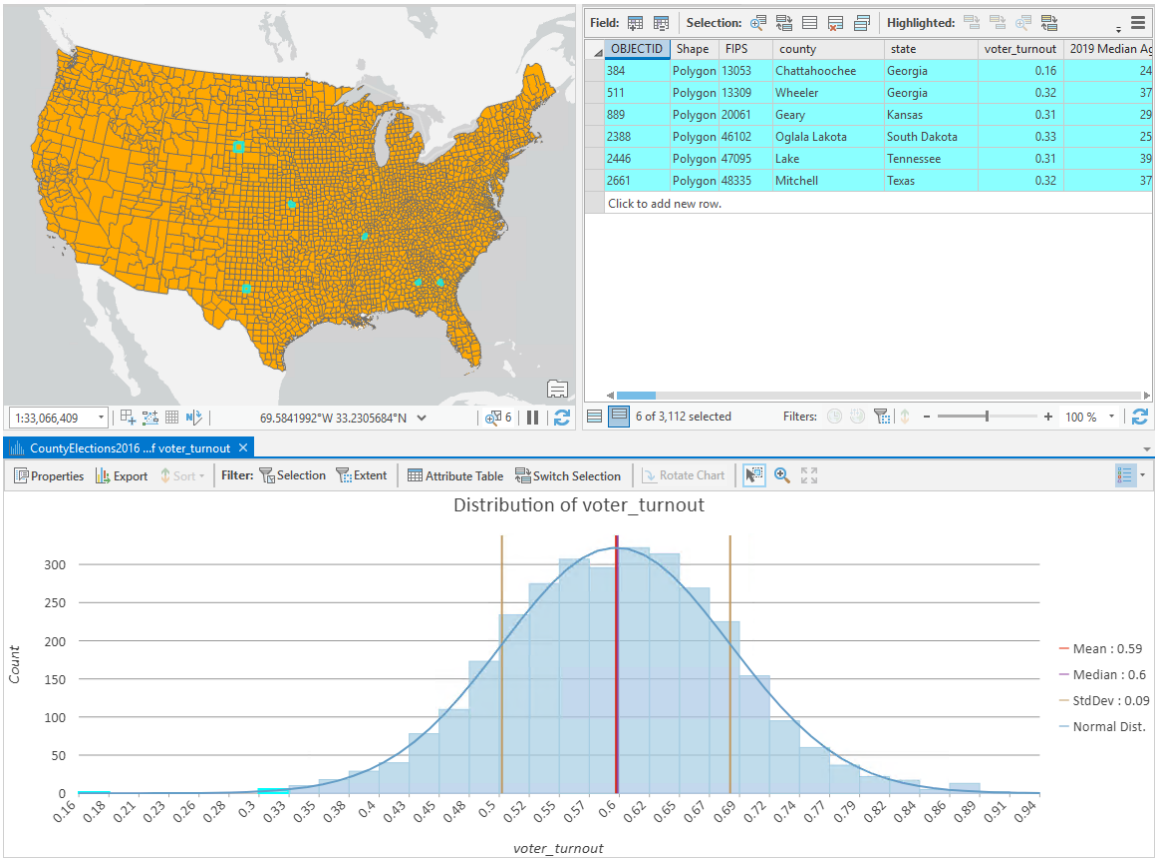


The counties with the lowest voter turnout are highlighted in the histogram and in the map. You can review these records in the attribute table.

g Click the attribute table tab and drag it to the center of the map until the docking target appears.

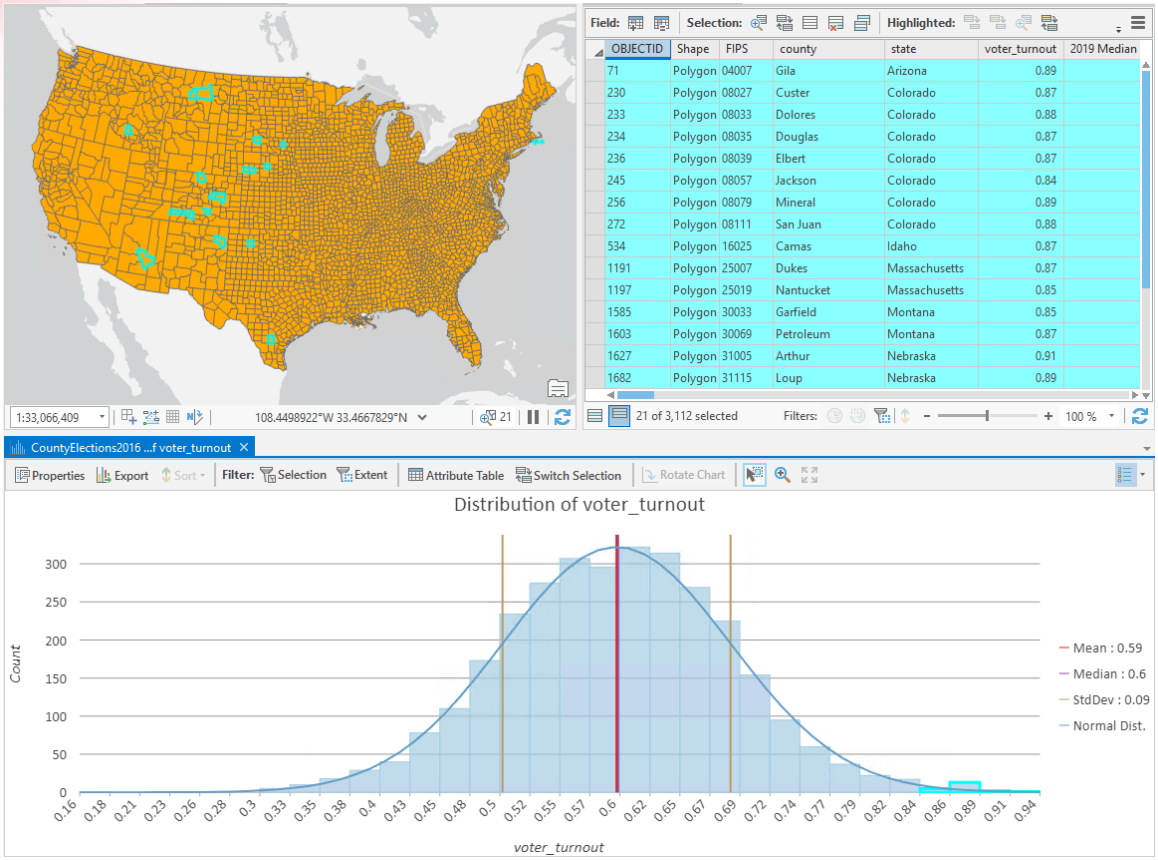
h. Dock the attribute table to the right of the map, above the histogram.

i. At the bottom of the attribute table window, click the Show Selected Records button .



The attribute table shows the records for the selected counties. You can review these records in the table to verify their voter turnout values.

j. In the histogram, click and drag to select the bins with the highest voter turnout values in the right tail of the histogram.



The counties with the highest voter turnout are selected in the histogram, map, and table.

k. In the histogram, click any blank area to clear the selection.

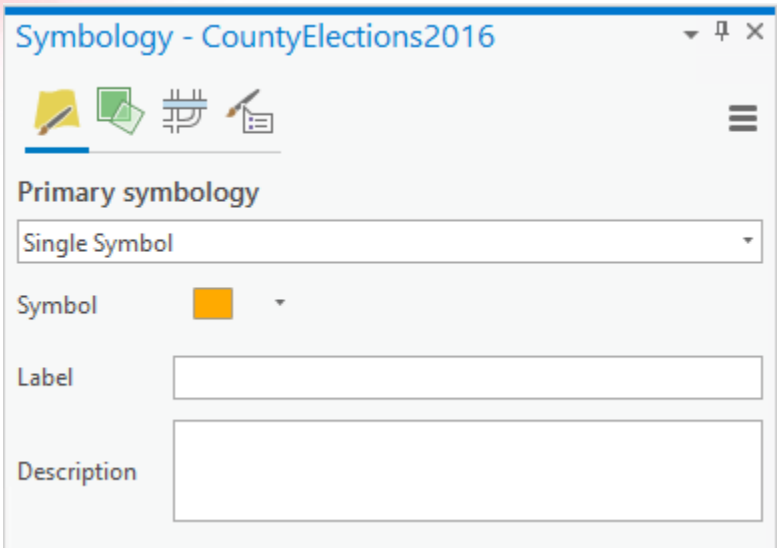
l. Close the attribute table and chart window.

You have visualized the distribution of voter turnout values and identified where the lowest and highest values fall on the map. Next, you will use layer symbology to visualize the spatial distribution of voter turnout across the country.

Step 3: Change layer symbology

Currently, every county in the 2016 election results layer is symbolized using the same color. This type of symbology is referred to as single symbol. In this step, you will change the symbology to represent the 2016 voter turnout.

a. In the Contents pane, right-click CountyElections2016 and choose Symbology.



The Symbology pane appears.

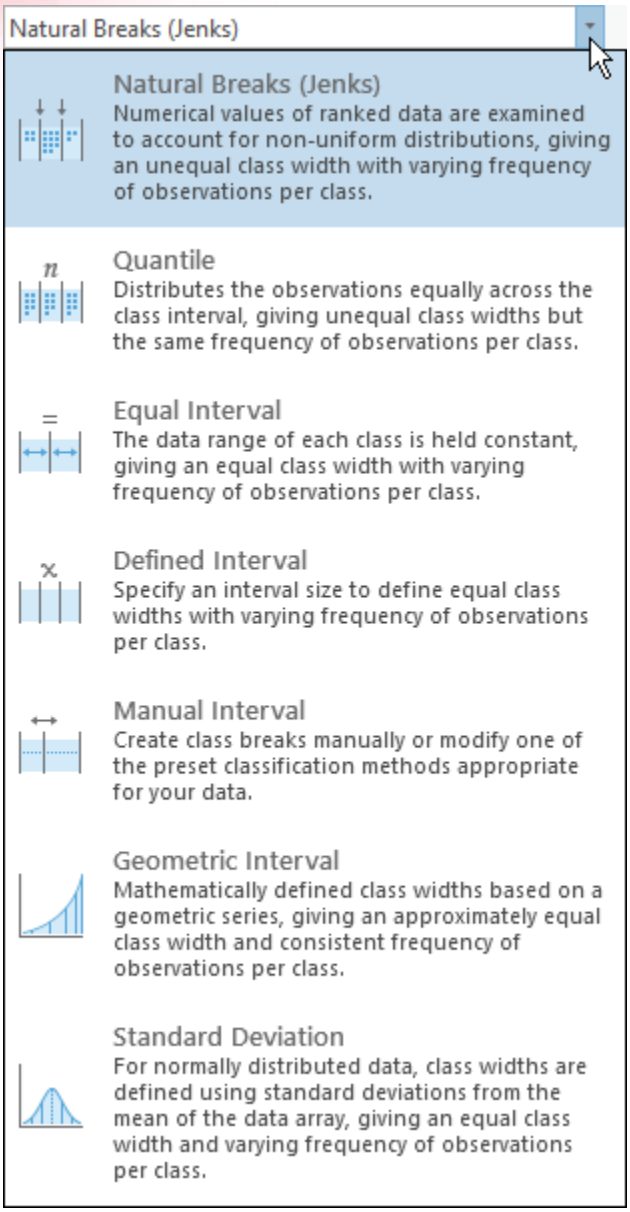
b. In the Symbology pane, under Primary Symbology, click the down arrow and choose Graduated Colors.

Graduated Colors classifies the data into different ranges based on the values of a specified attribute field. Each class is assigned a shade of color to show the relative difference between the feature values. To learn more about graduated colors, see ArcGIS Pro Help: Graduated colors.

You will specify the field and color ramp so that the symbology represents ranges of voter turnout.

c. Under Graduated Colors, next to Field, choose Voter_Turnout, if necessary.

d. Under Graduated Colors, next to Method, click the down arrow.



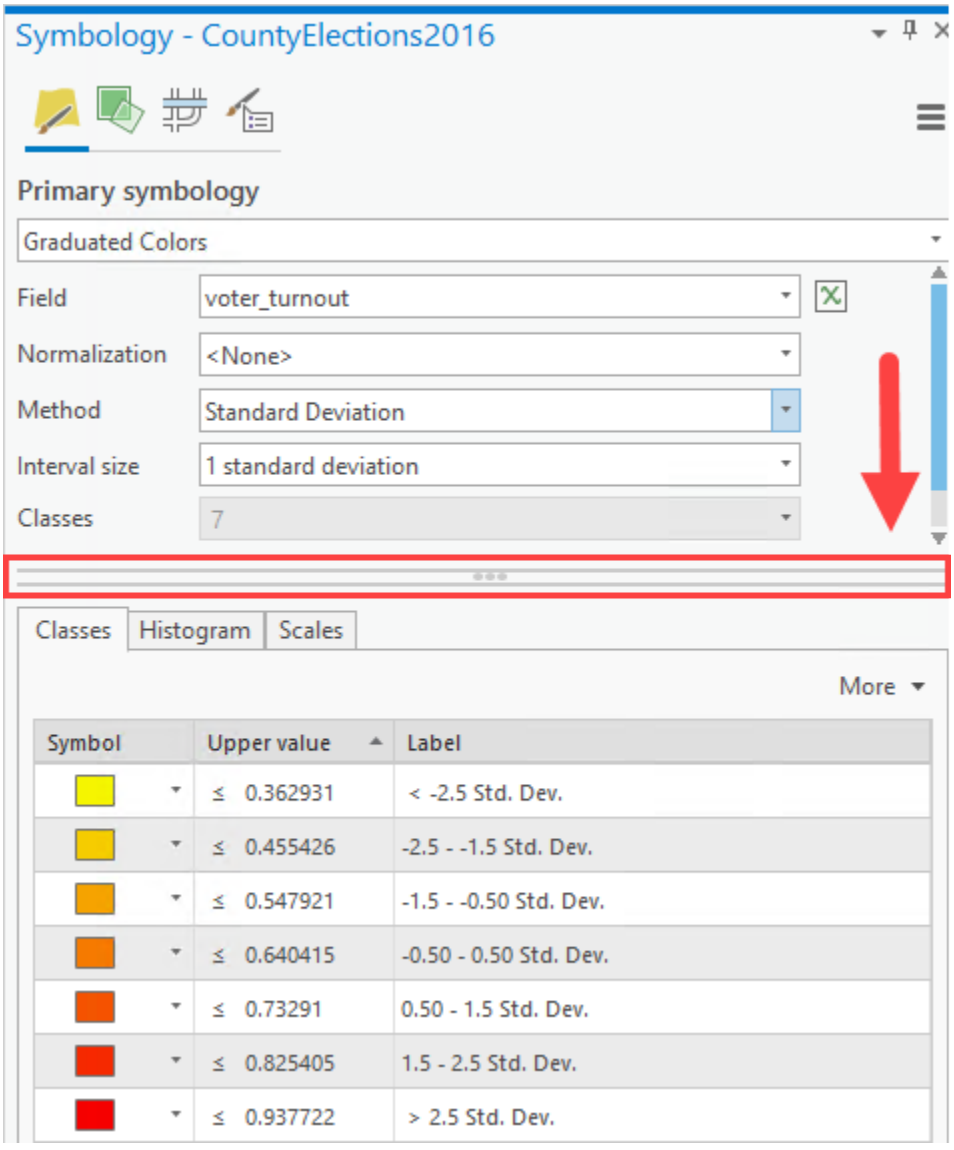
A list describing the available classification methods appears.

Natural Breaks (Jenks) is the default classification method because it is data driven. This means that the symbol ranges are calculated based on the data values, making it adaptable to different types of data distributions. Because you have determined that the voter turnout variable is normally distributed, you will use the Standard Deviation method to classify voter turnout. To learn more about data classification methods, see ArcGIS Pro Help: Data classification methods.

e From the list of classification methods, choose Standard Deviation.

f Under Classes, next to Color Scheme, click the down arrow.

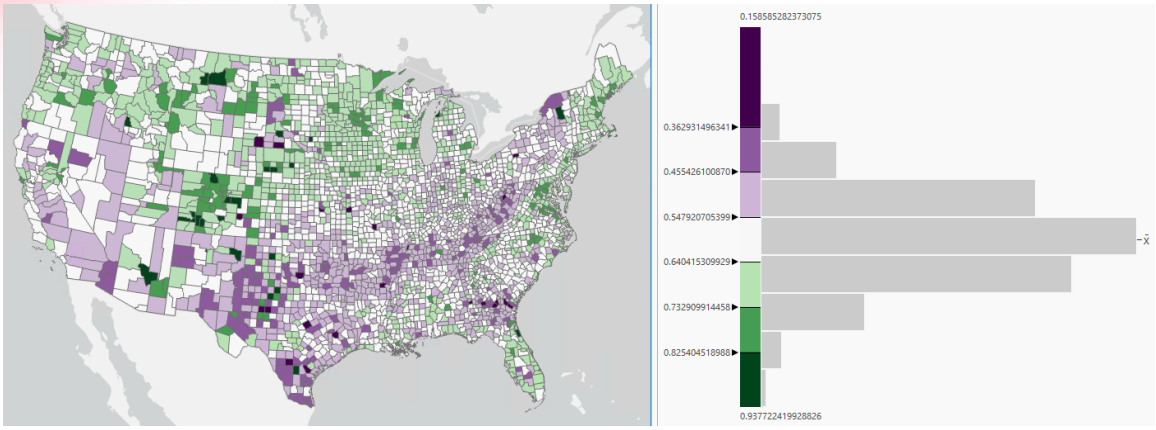
Note: If you do not see the Color Scheme parameter, you may need to scroll down further in the Symbology pane or move the pane divider down.



g In the Color Scheme window, check the box for Show Names.

h Choose the Purple-Green (Continuous) color scheme.

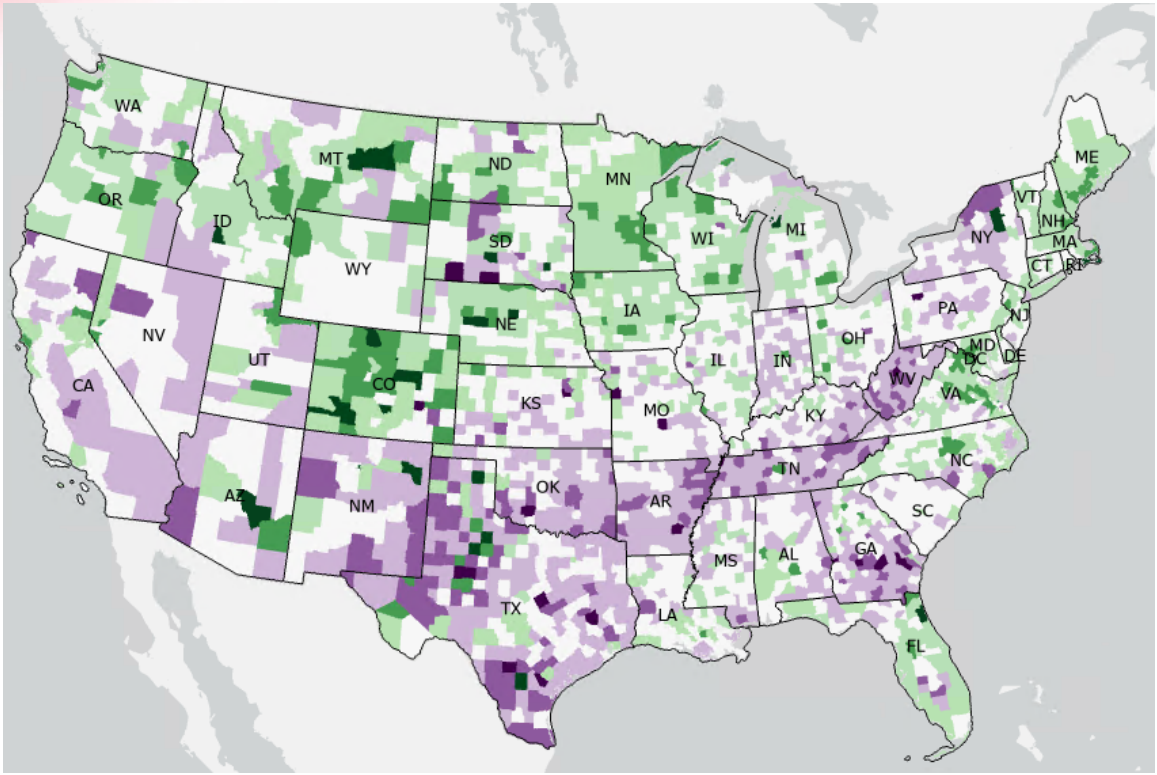
i In the Symbology pane, below the pane divider, click the Histogram tab.



By applying a Graduated Colors symbology, you created what is commonly referred to as a choropleth, or thematic, map. Choropleth maps visualize low-to-high values using light-to-dark colors. Because you are using the Standard Deviation classification method, you applied a diverging color scheme. A diverging color scheme classifies values based on how far they are from the average. On the Histogram tab, you can see how the distribution of values corresponds to the classes of color. The counties with below-average voter turnout are represented in shades of purple, and the counties with above-average voter turnout are represented in shades of green.

j. In the Symbology pane, next to Color Scheme, click the Color Scheme Options button and choose Apply To Fill And Outline.

k. In the Contents pane, turn on the US_States layer.



Combining the Graduated Colors symbology with the state layer overlay, you can begin to get a sense of how states compare to each other in terms of voter turnout. States like West Virginia (WV) and Tennessee (TN) stand out as having low voter turnout, and states like Colorado (CO) and Minnesota (MN) stand out as having high voter turnout. You will use a bar chart to summarize and compare voter turnout by state.

l. Close the Symbology pane.

m. Save the project.

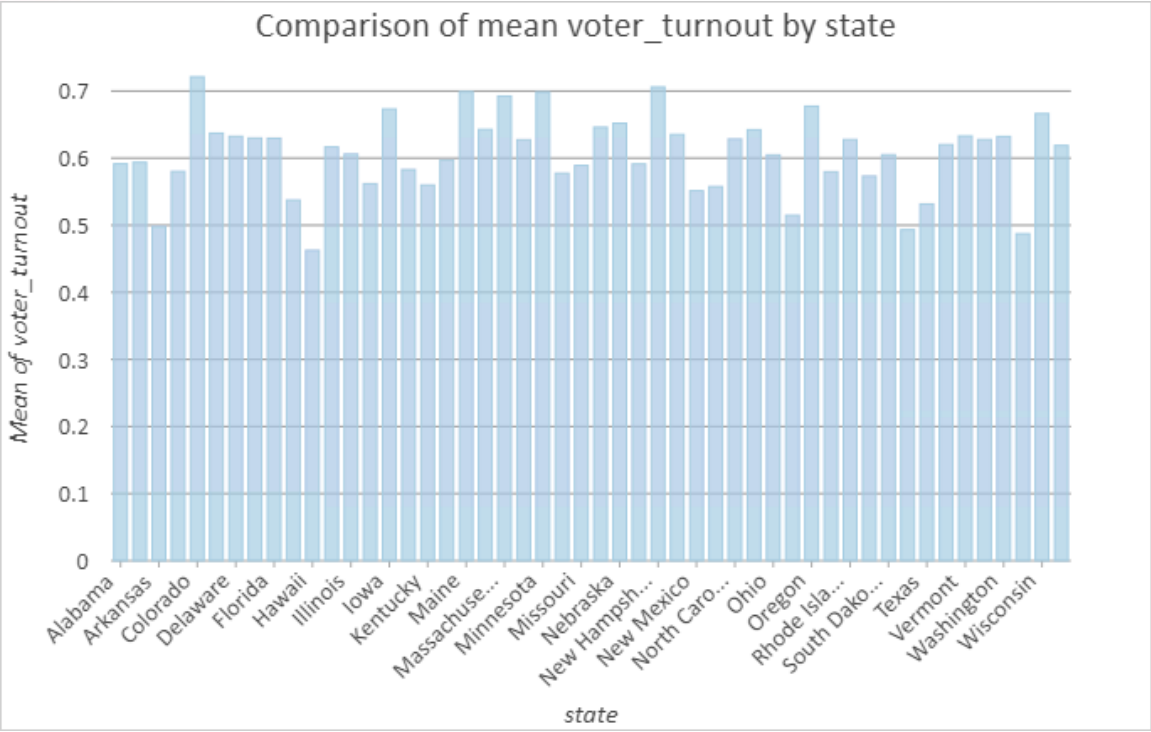
Step 4: Create a bar chart

a. In the Contents pane, right-click CountyElections2016, point to Create Chart, and choose Bar Chart.

b. In the Chart Properties pane, enter the following parameters:

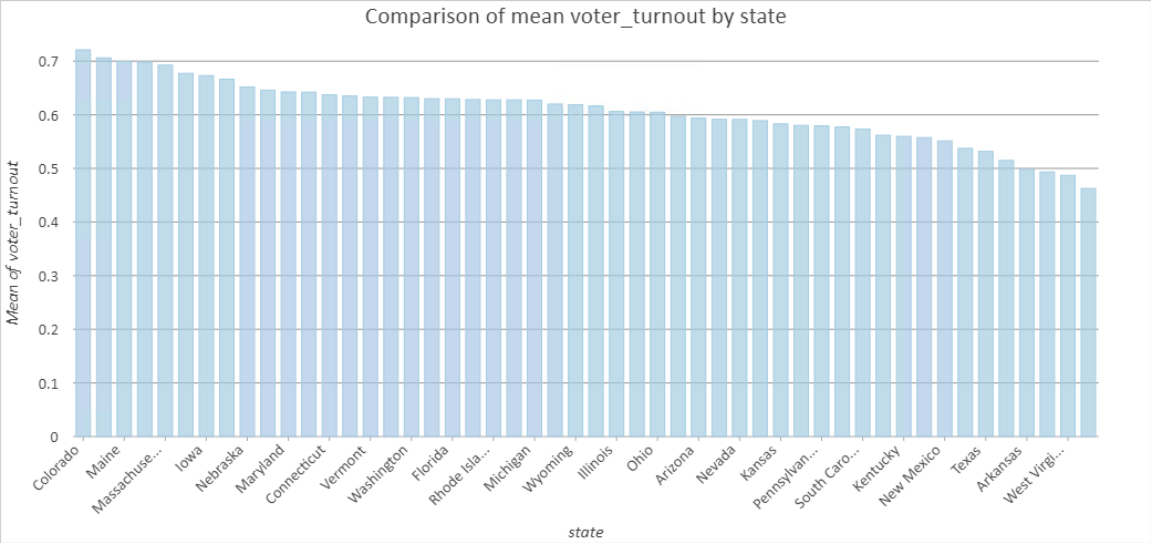
- Category Or Date: State
- Aggregation: Mean
- Numeric Field(s): Voter_Turnout

c Click Apply.



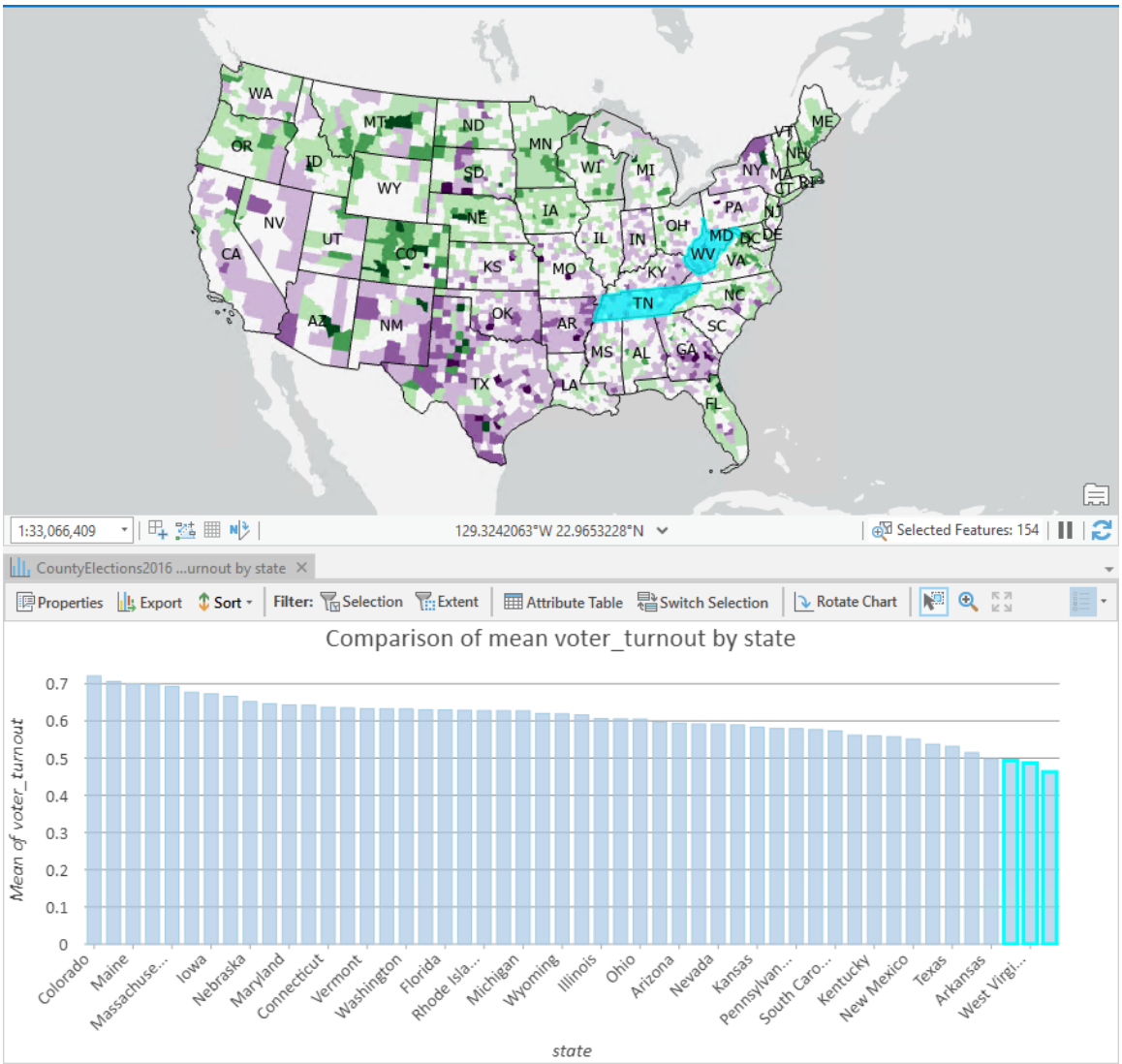
The bar chart summarizes county voter turnout values by state. Each bar represents a state, and the height of the bar corresponds to the mean voter turnout value. For more information about bar charts, see ArcGIS Pro Help: Bar chart.

d In the chart window, click Sort and choose Y-Axis Descending.



Sorting the bars by value makes it easier to visually rank the states from highest to lowest voter turnout.

e Select the three states with the lowest average voter turnout.



After Hawaii, West Virginia and Tennessee have the lowest average voter turnout, which confirms what you observed in the map. The bar chart summarizes the voter turnout for each state into a single average value. Within each state, however, there can be quite a bit of variation in voter turnout. To examine the individual county voter turnout within each state, you can use a filtered bar chart.

f Clear the selection.

Step 5: Filter a chart using a selection

a. Create a second bar chart for the CountyElections2016 layer using the following parameters:

- Category Or Date: County
- Aggregation: None
- Numeric Field(s): Voter_Turnout

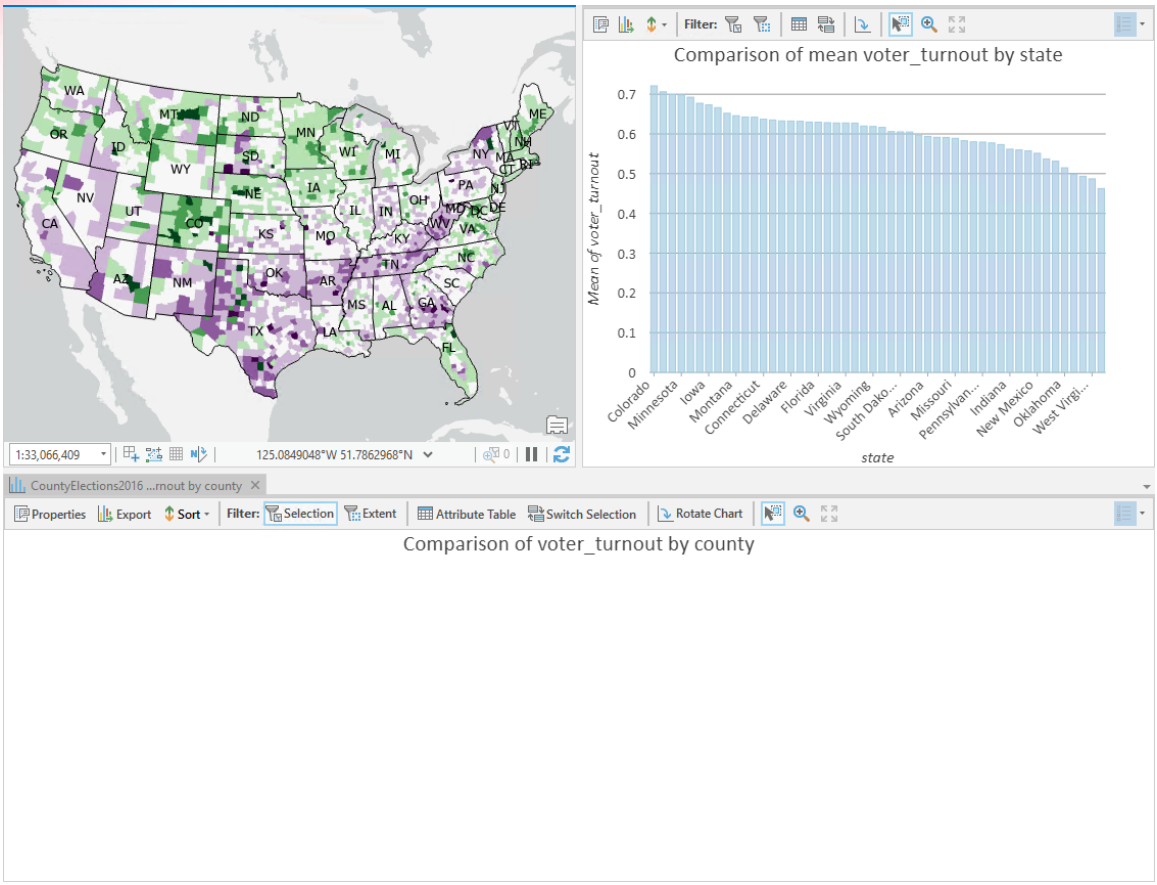
b. Click Apply.

c. In the chart window, click Sort and choose Y-Axis Descending.

d. In the chart window, next to Filter, click the Filter By Selection button  .

Filter By Selection filters the chart to only show selected features. Because no features are selected yet, the bar chart is empty.

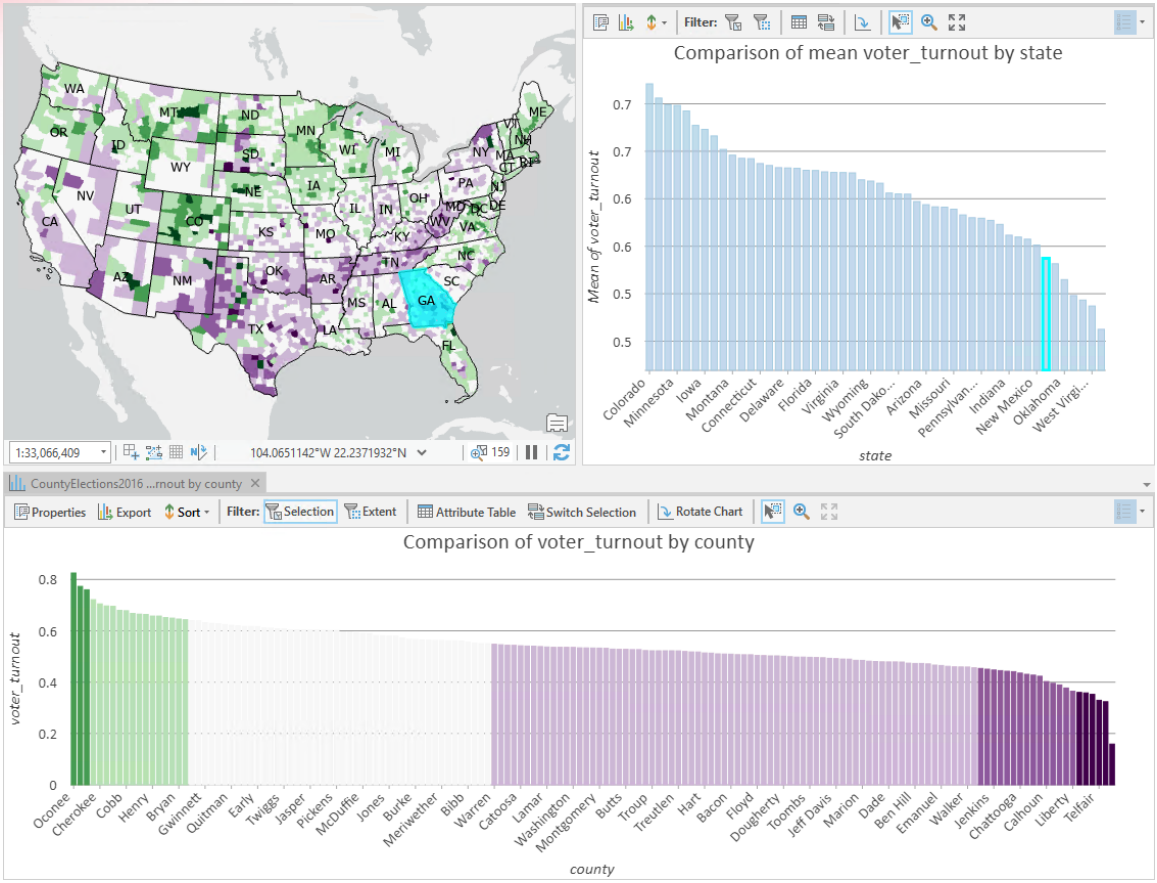
e. Dock the state bar chart window to the right of the map, above the county bar chart window.



You now have a bar chart visualizing average voter turnout by state and a bar chart visualizing individual county voter turnout values of selected features.

- In the state bar chart, select Georgia.

Note: You can use the Zoom Mode button to zoom in to the bar chart.



The county bar chart populates to show the individual county values within Georgia. Each bar in the county bar chart corresponds to a single feature on the map, so the colors of the bars match the map symbology.

You can use this interactive selection to see the range of individual county values within each state. To compare the county values to the national average voter turnout value of 0.59 (identified in the histogram), you will add a guide to your chart.

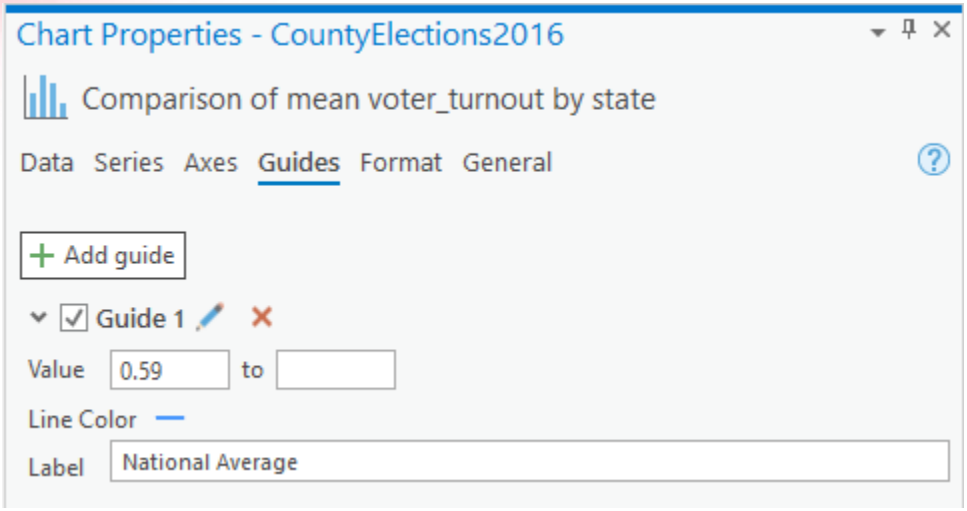
g. In the Chart Properties pane of the county bar chart, click the Guides tab.

h. Under Guides, click Add guide.

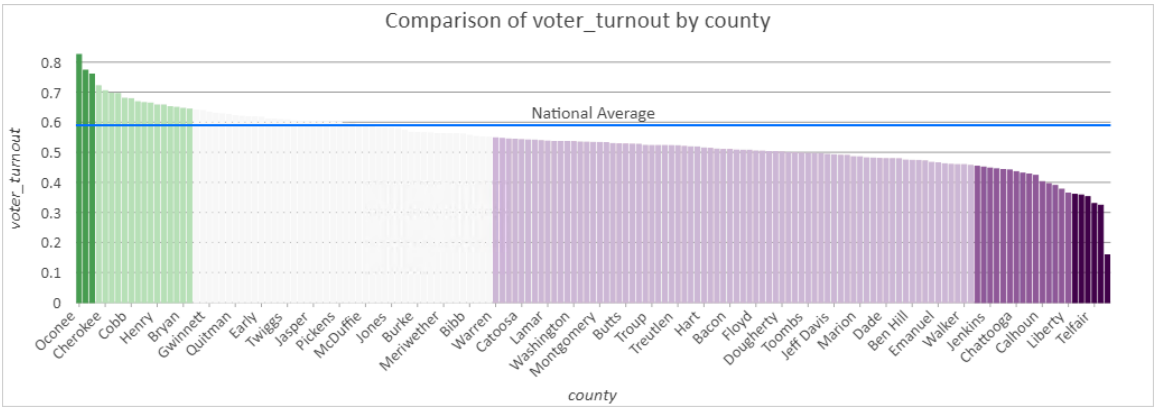
i. Next to Value, type **0.59**.

j. Next to Line Color, click the line and choose a bright blue color.

k. Next to Label, type **National Average**.



A line appears in the county bar chart marking the national average voter turnout value.



Guides allow you to reference or highlight significant values or thresholds in your charts.

l. In the state bar chart, click other states to see how their county voter turnout values vary within the state and to compare the state's average voter turnout to the national average.

m. Clear the selection and close both bar charts.

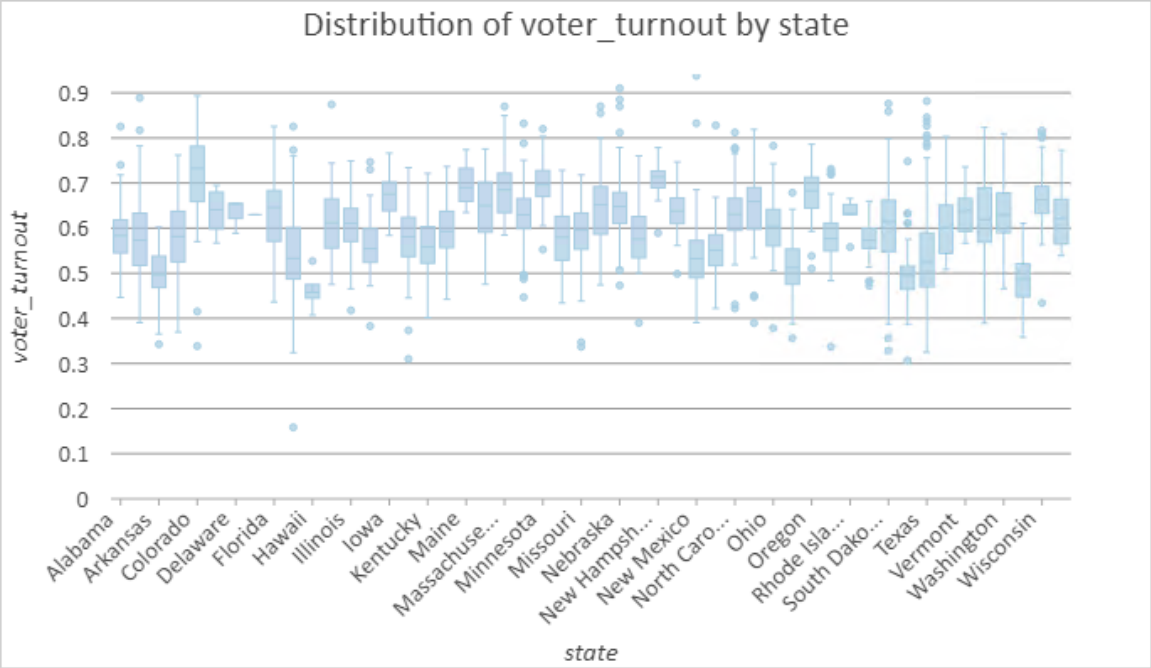
You have used bar charts and Filter By Selection to explore state voter turnout averages and to examine the individual county voter turnout values for each state. You can use interactive selection to see an overview of state averages and to investigate the range of county values within individual states. To visualize and compare the distribution of voter turnout values for every state at once, you will create a box plot.

Step 6: Create a box plot

a Create a box plot for the CountyElections2016 layer using the following parameters:

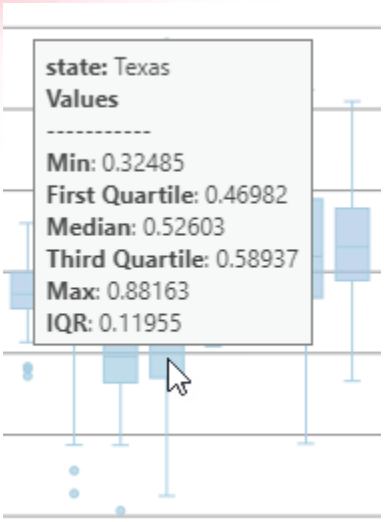
- Numeric Field(s): Voter_Turnout
- Category: State
- Show Outliers: Checked

b Click Apply.



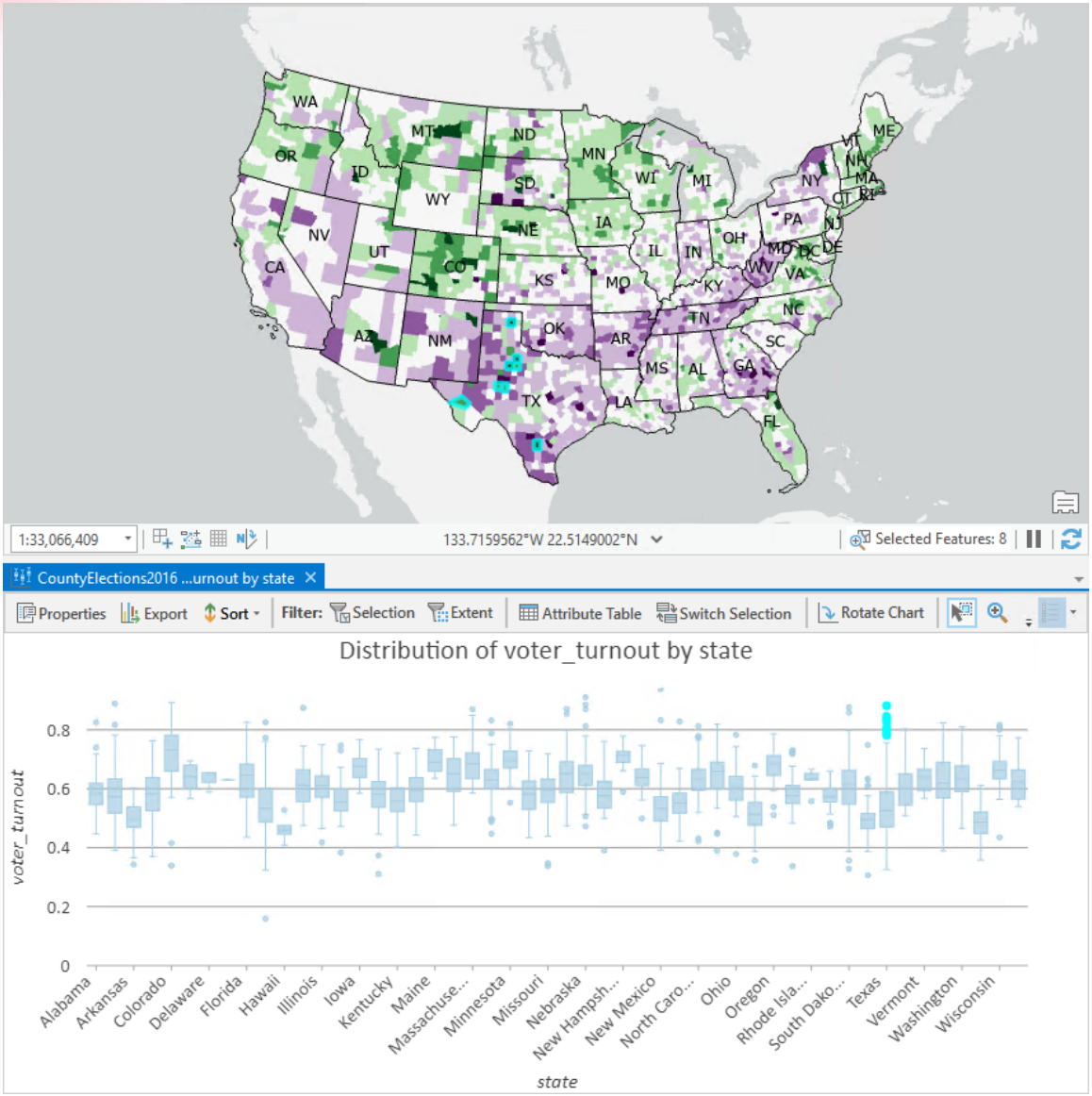
The box plot chart allows you to visualize and compare the entire distribution of county voter turnout values for each state. Box plots split numeric values into four equal quartiles and visualize five key statistics for each distribution: minimum, first quartile, median, third quartile, and maximum. The whiskers extending from the boxes span from the minimum value to the maximum value, illustrating the full range of values found in each state. The boxes span from the first quartile to the third quartile, illustrating the range of the middle half of values, or the interquartile range (IQR). The IQR indicates the size of spread, or variability, in voter turnout values in each state. For more information about box plots, see ArcGIS Pro Help: Box plot.

c In the box plot chart, pause your pointer on Texas.



The ToolTip displays the key voter turnout statistics for the state. Texas has a relatively low voter turnout average as a state. However, there is a wide range of county voter turnout values, spanning from approximately 0.32 to approximately 0.88. The counties with voter turnout values that are very different from the state average are considered outliers and are displayed as dots beyond the plot's whiskers.

d. In the box plot chart, select the Texas outliers.



The outliers are selected on the map.

Reviewing the overall distribution of voter turnout values in conjunction with individual feature locations can help you understand the data and identify areas that you may want to further investigate.

e Clear the selection and close the box plot chart.

Step 7: Explore variable relationships in a scatter plot matrix

You have used various spatial and non-spatial data visualization techniques to explore voter turnout values and distributions. You can also use data visualization tools to explore relationships in your data. Because you want to predict voter turnout, you will explore the relationship between voter turnout and other variables in your data.

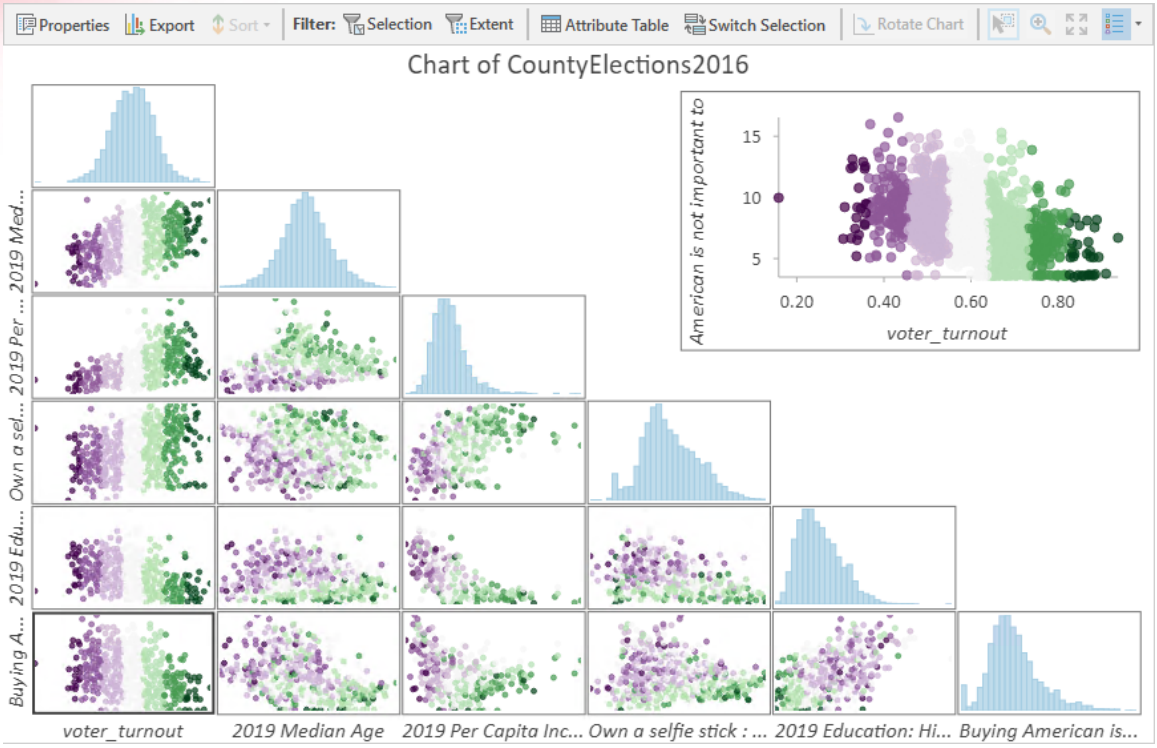
a. Create a scatter plot matrix for the CountyElections2016 layer.

b. In the Chart Properties pane, under Numeric Fields, check the box for the following fields:

- Voter_Turnout
- 2019 Median Age
- 2019 Per Capita Income
- Own A Selfie Stick : Percent
- 2019 Education: High School/No Diploma : Percent
- Buying American Is Not Important To Me : Percent

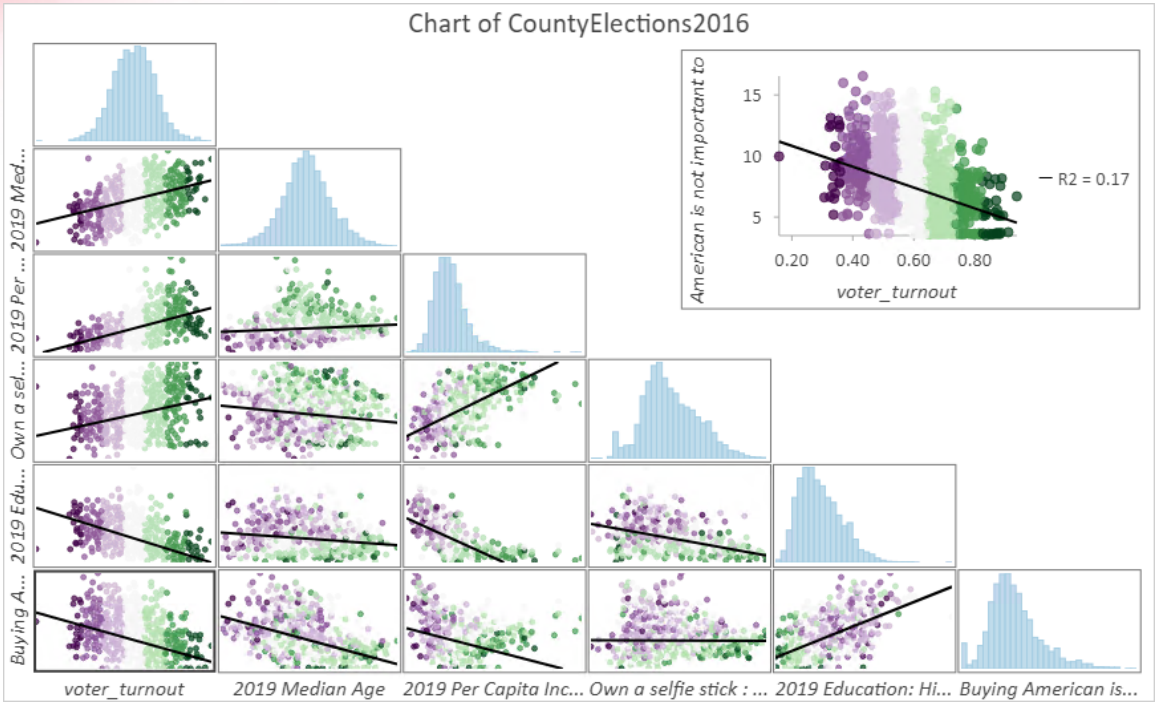
c. Under Numeric Fields, check the box for Show Histograms.

d. Click Apply.



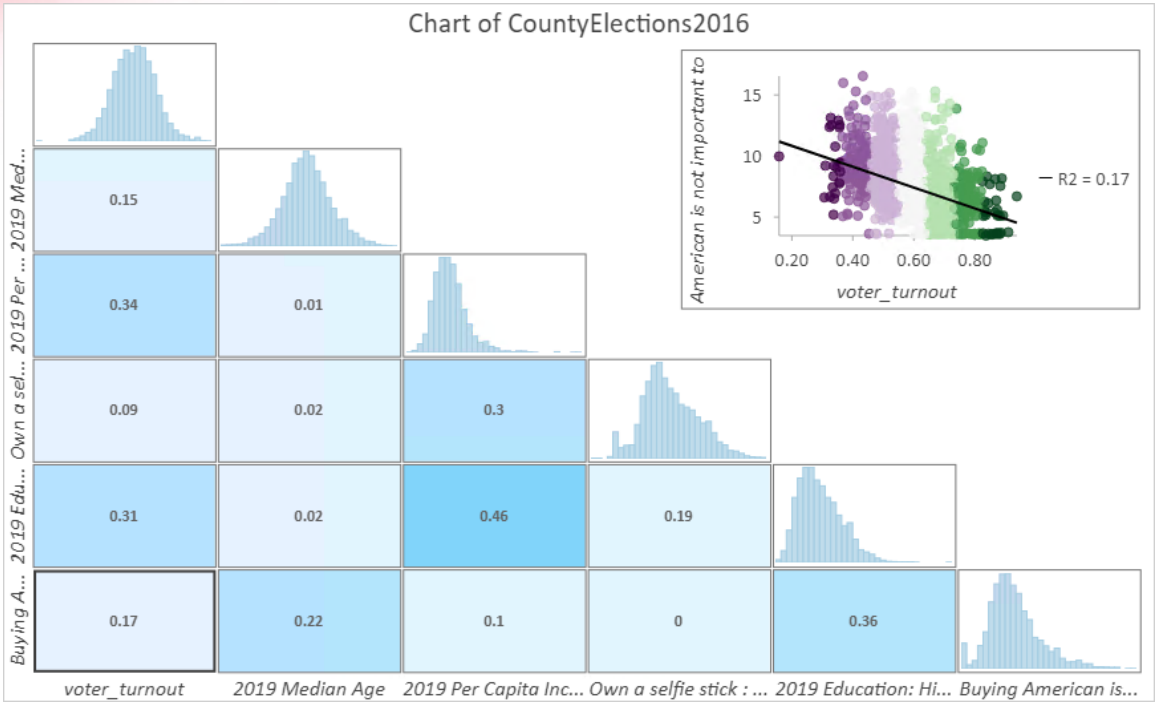
A scatter plot matrix is a grid of scatter plots, also referred to as mini-plots, used to visualize bivariate relationships between combinations of variables. Each scatter plot in the matrix visualizes the relationship between a pair of variables, allowing many relationships to be explored in one chart. A histogram visualizing the distribution of each individual variable can also be included in the matrix. For more information about scatter plot matrices, see ArcGIS Pro Help: Scatter plot matrix.

e In the Chart Properties pane, check the box for Show Linear Trend.



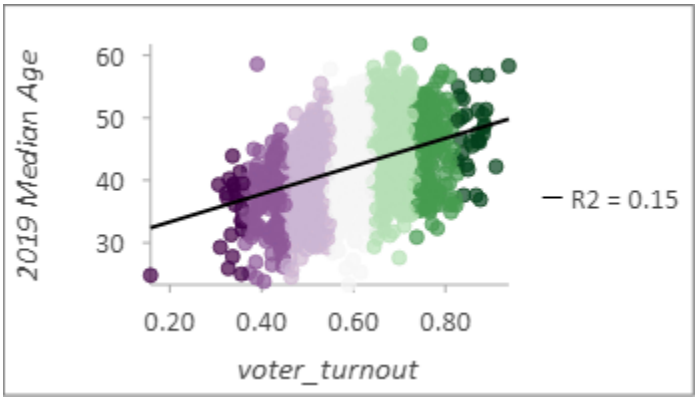
A linear trend line is added to each scatter plot in the matrix. The direction of the trend line indicates whether the variables have a positive or negative relationship, and the R-squared (R2) value indicates the strength of the relationship. For more information about scatter plots, see ArcGIS Pro Help: Scatter plot.

f In the Chart Properties pane, check the box for Show As R2.



The mini-plots in the matrix are now visualized with a color gradient corresponding to the strength of the R-squared value. You can select any mini-plot to view the relationship in more detail using the larger preview plot. While every pairwise combination of variables is plotted in the matrix, you are specifically interested how each variable relates to voter turnout. The column of mini-plots on the far left includes the relationships between voter turnout and the other variables.

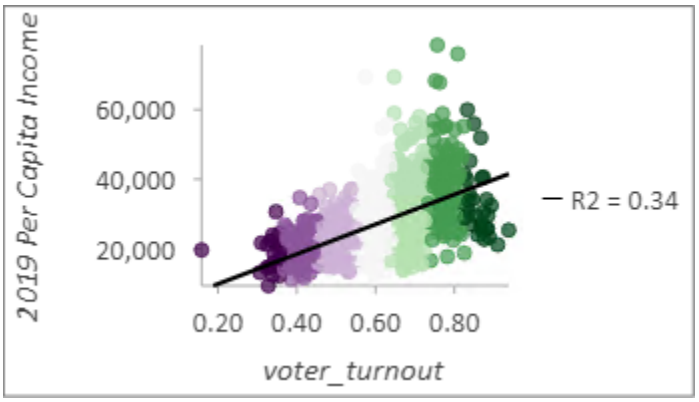
g. In the scatter plot matrix, select the mini-plot comparing Voter_Turnout and 2019 Median Age.



Median age has a positive relationship with voter turnout, where a higher median age corresponds to a higher voter turnout. However, the R-squared value for this trend is 0.15,

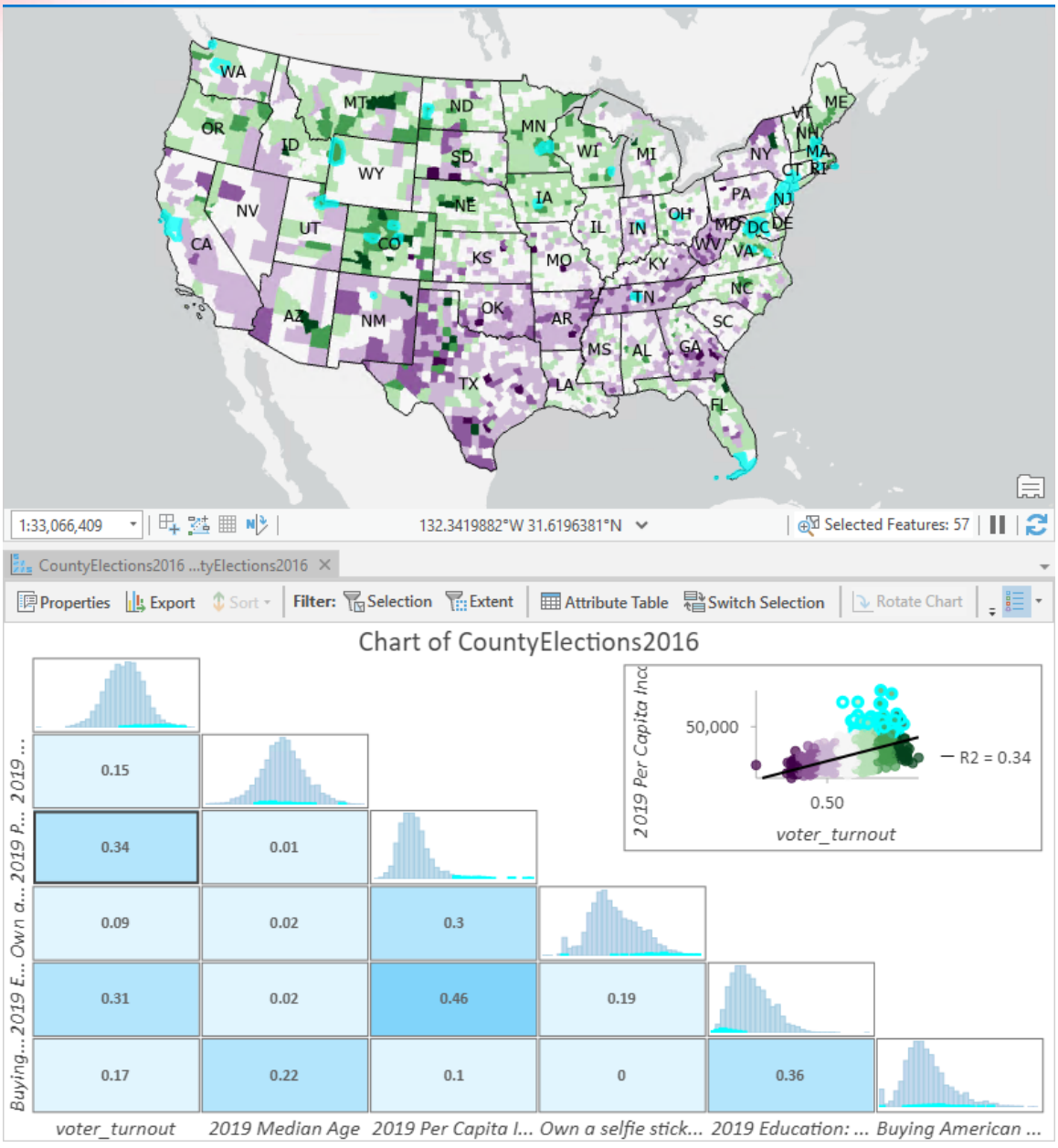
which means that median age alone can only explain about 15% of the variability in the voter turnout values.

h. Select the mini-plot comparing Voter_Turnout and 2019 Per Capita Income.



Per capita income also has a positive relationship with voter turnout, where a higher per capita income corresponds to a higher voter turnout. The R-squared value for this trend is 0.34, which means that per capita income can explain about 34% of the variability in voter turnout values. Within the preview plot, you can see where some of the points deviate from the trend. You can investigate those points using a selection.

i. In the preview plot, select points that deviate from the trend to see where they fall on the map.



The selected counties are highlighted on the map. To see if the variable relationships vary spatially, you will filter the chart by the map extent.

j. Clear the selection.

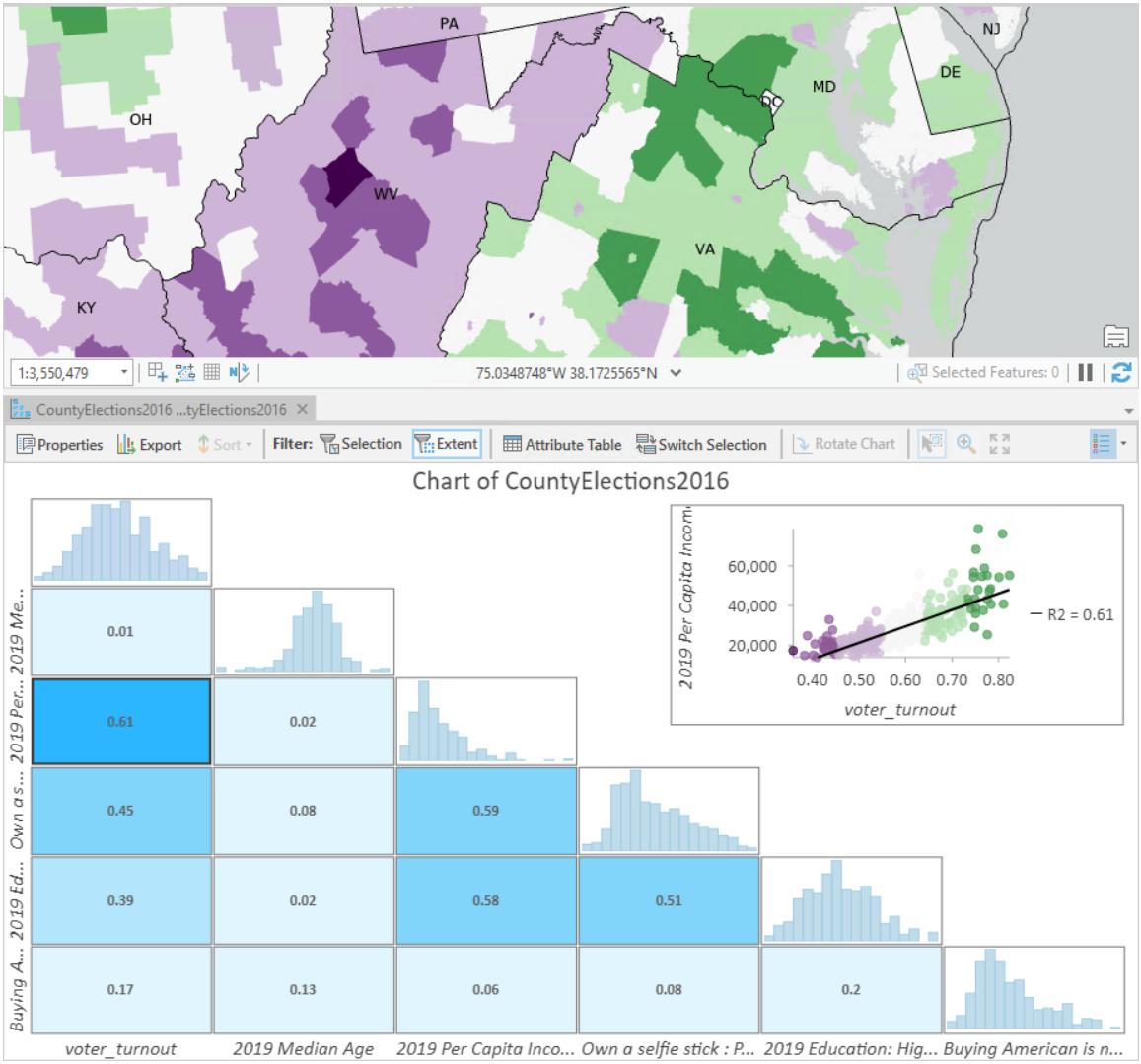
k. Save the project.

Step 8: Explore relationships at different scales

To investigate if the strength of the variable relationships vary from place to place, you can filter the scatter plot matrix to include only counties that are visible on the map.

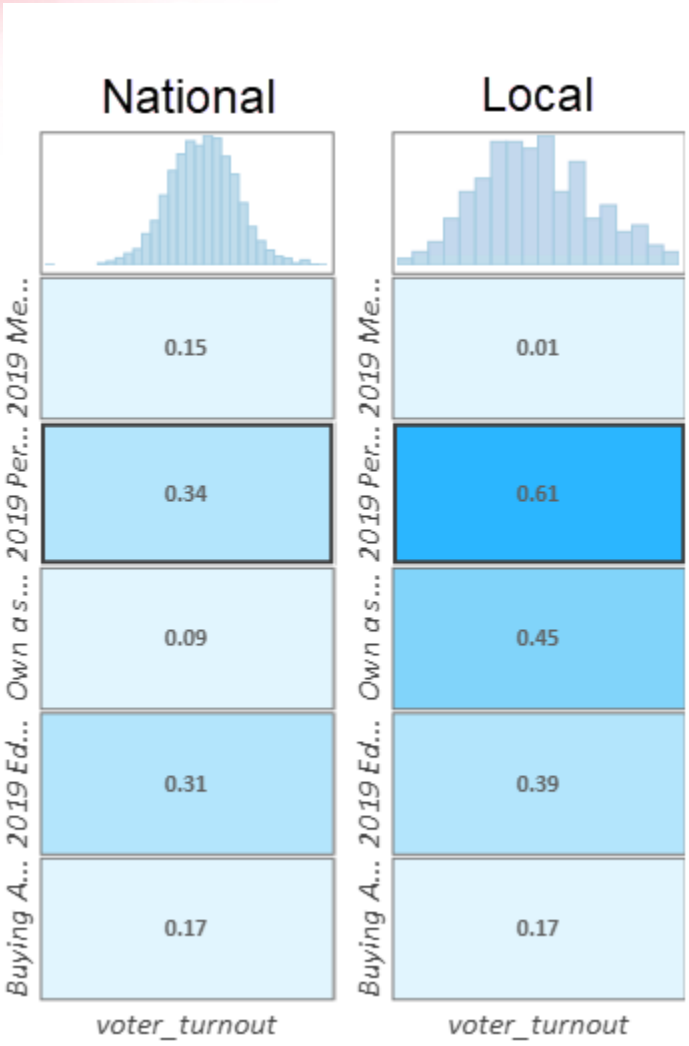
a. In the scatter plot matrix window, next to Filter, click the Filter By Extent button .

b. From the Map tab, in the Navigate group, click Bookmarks and choose WV, VA, MD.



Note: The R-squared values will vary based on the size of your map and chart windows.

The chart updates to calculate the relationships between the variables of the counties visible in the map extent.



If you compare the R-squared values at a national scale to this local scale, you can see that the relationships between voter turnout and per capita income and voter turnout and owning a selfie stick increased significantly. However, the relationship between voter turnout and median age has disappeared.

c Zoom and pan around the map to explore how variable relationships vary by scale and location.

The changes in R-squared indicate that the linear relationships between the variables vary spatially. In the next step, you will explore and quantify different types of local relationships using the Local Bivariate Relationships tool.

d Close the scatter plot matrix window.

e. In the Contents pane, uncheck CountyElections2016 to turn the layer off.

Step 9: Explore local bivariate relationships

The Local Bivariate Relationships tool quantifies the local relationship between two variables and indicates how the type of relationship varies spatially. You will use this tool to quantify the relationship between voter turnout and per capita income and determine if this relationship varies across the contiguous United States. To learn more about the Local Bivariate Relationships tool, see ArcGIS Pro Help: How Local Bivariate Relationships works.

a. Go to the United States bookmark.

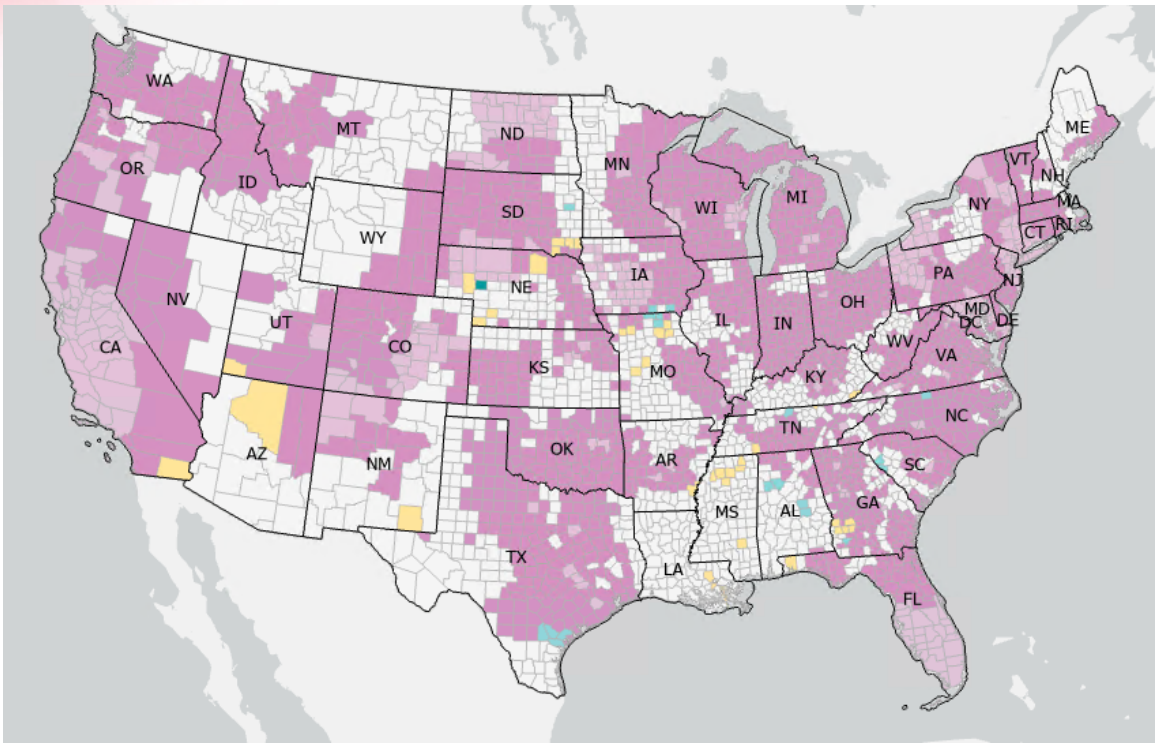
b. From the Analysis tab, in the Geoprocessing group, click Tools.

c. In the Geoprocessing pane, search for **Local Bivariate Relationships**.

d. Run the Local Bivariate Relationships (Spatial Statistics Tools) tool using the following parameters:

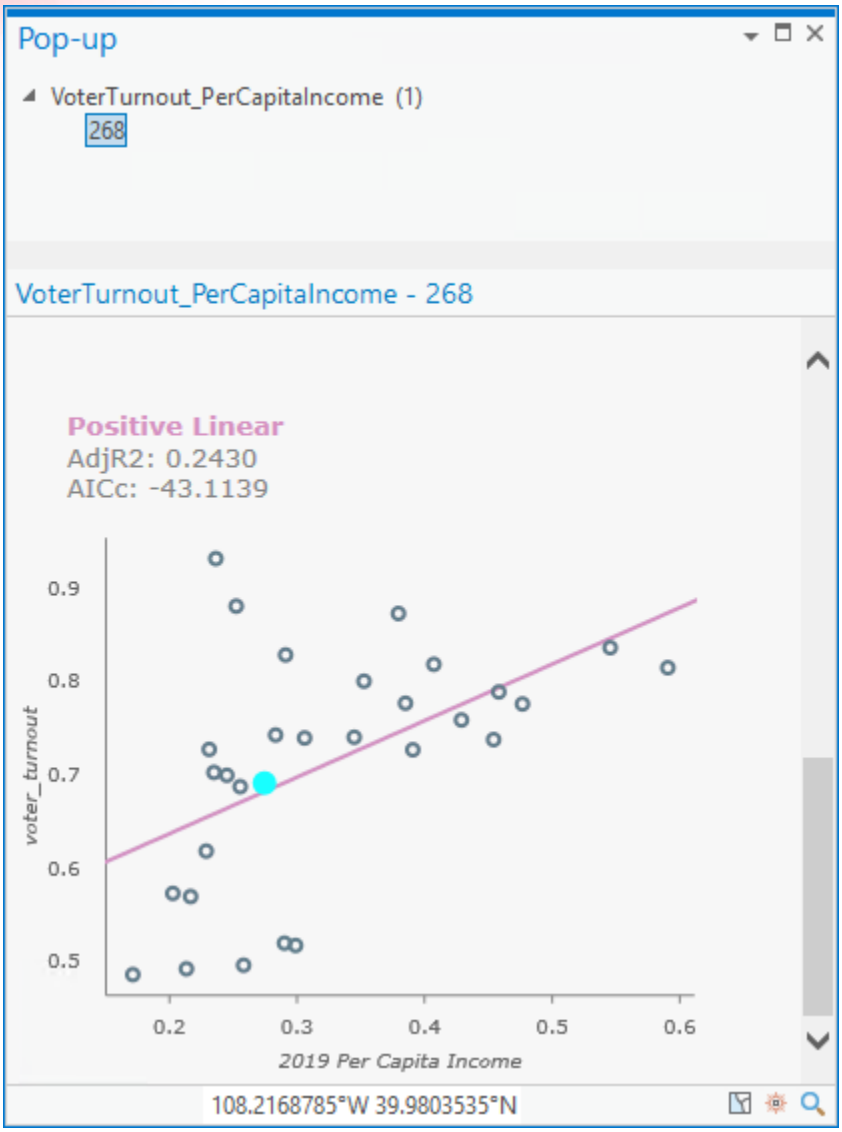
- Input Features: CountyElections2016
- Dependant Variable: Voter_Turnout
- Explanatory Variable: 2019 Per Capita Income
- Output Features: **VoterTurnout_PerCapitaIncome**

e. In the Contents pane, drag the US_States layer above the VoterTurnout_PerCapitaIncome layer.



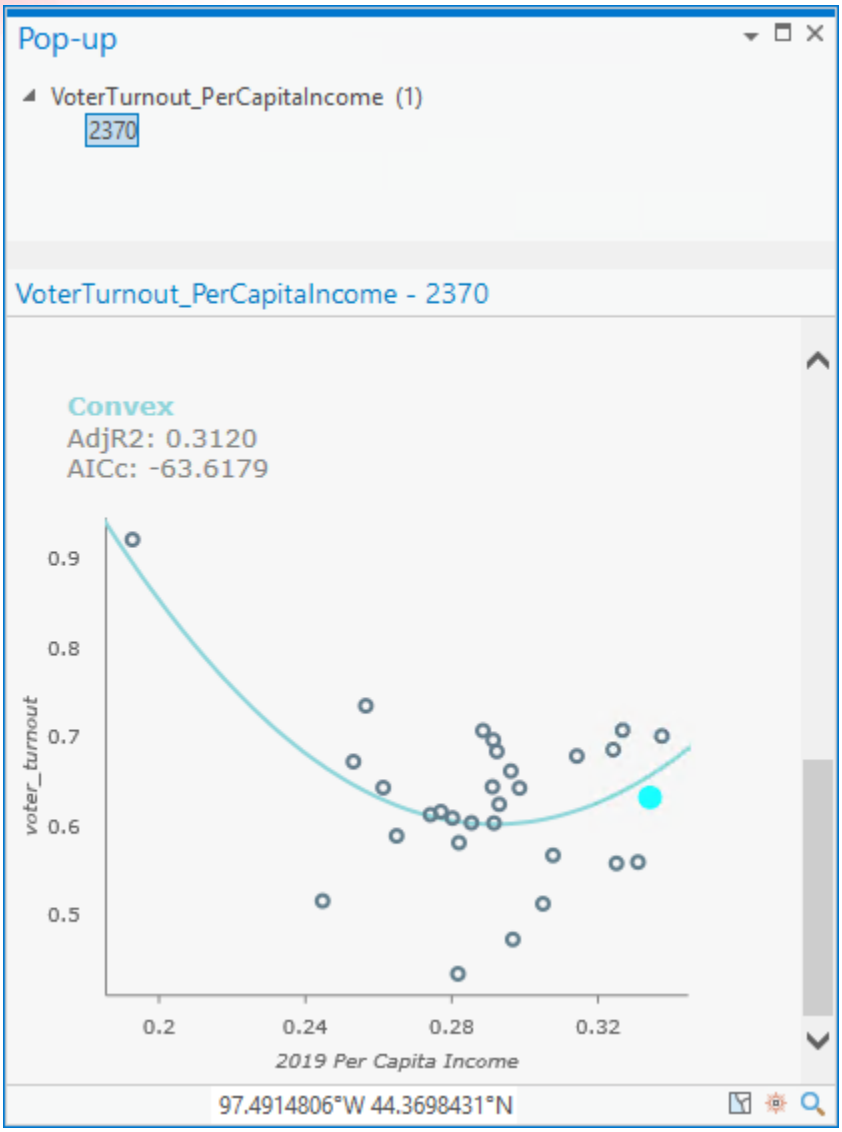
The Local Bivariate Relationships tool identifies not only linear relationships but also concave/convex and other undefined complex relationships. The colors in the map correspond to the type of relationship found in that area. Based on this output, the relationship between per capita income and voter turnout varies by location. In some areas, there is no statistically significant relationship between the two variables. In most areas, however, there is a statistically significant positive linear relationship to voter turnout, where voter turnout increases as per capita income increases.

f Click one of the counties symbolized as Positive Linear to open a pop-up window visualizing the relationship.



You can use the pop-up windows to review a scatter plot of the selected county and its neighbors, indicating the strength and type of the local variable relationship.

g Click a county symbolized as Convex to examine the shape of the relationship in the scatter plot.

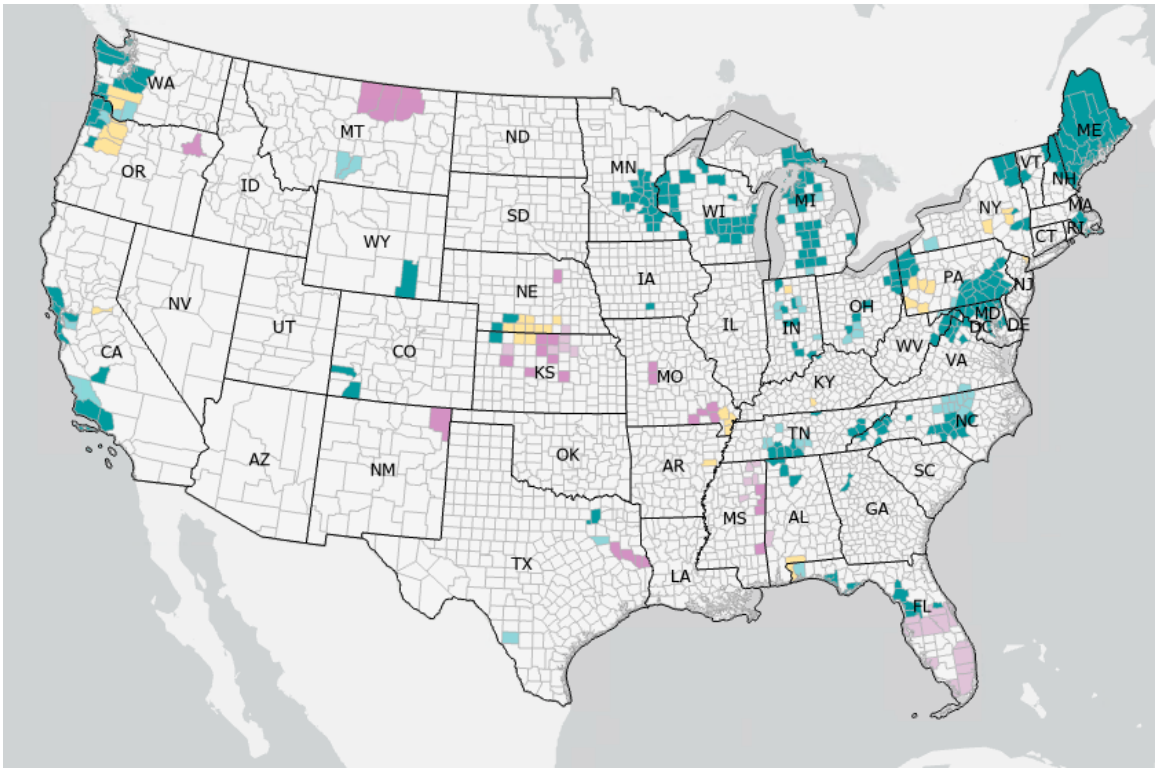


You have verified that voter turnout is related to per capita income across the majority of the country. Based on this information, per capita income will likely be useful in your prediction model. To see how the relationship between voter turnout and education vary spatially, you will run Local Bivariate Relationships again.

h Run the Local Bivariate Relationships tool using the following parameters:

- Input Features: County Elections 2016
- Dependant Variable: Voter_Turnout
- Explanatory Variable: 2019 Education: High School Diploma : Percent
- Output Features: **VoterTurnout_HSDiploma**

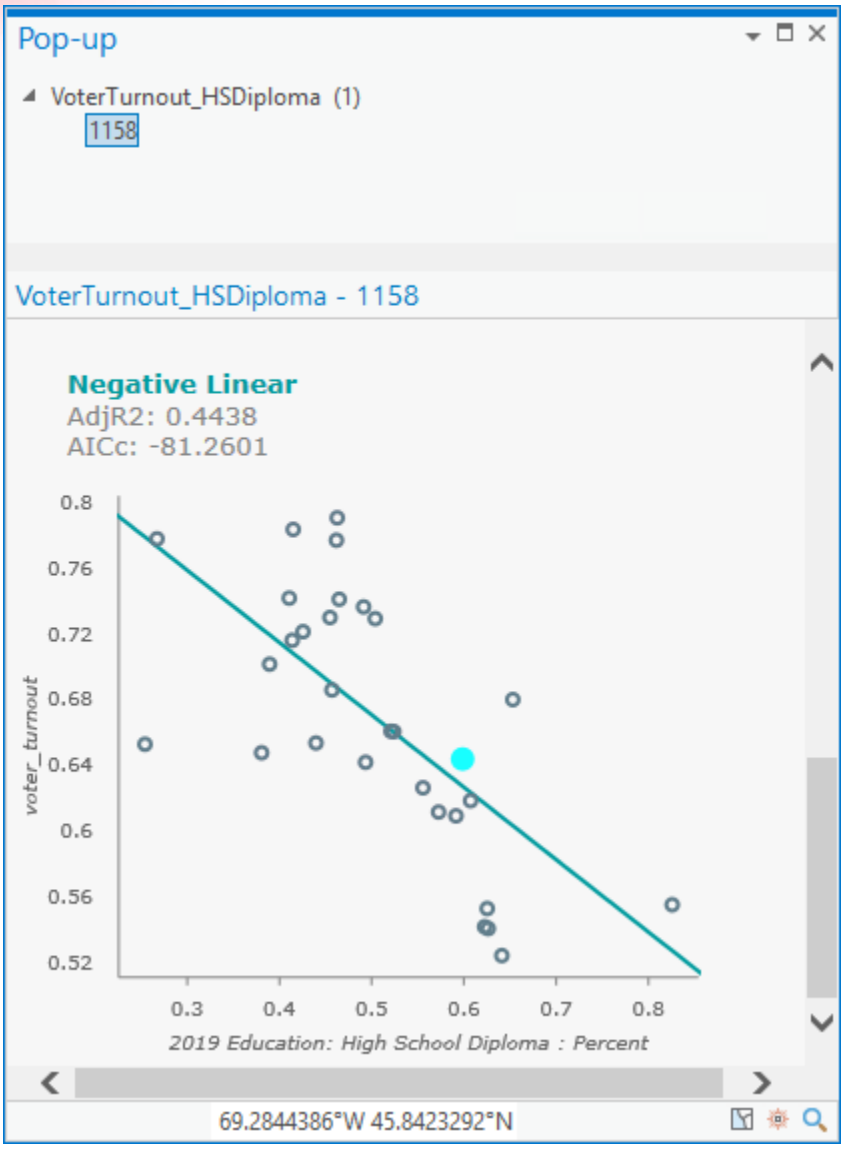
i In the Contents pane, drag the US_States layer above the VoterTurnout_HSDiploma layer.



Generally, the percent of the population with a high school diploma does not have a statistically significant relationship with voter turnout. However, there is a statistically significant negative linear relationship in most of Maine.

j In the map, zoom to Maine.

k Click one of the county polygons.



Based on this information, this variable would likely be useful if you were to create a prediction model of voter turnout in Maine. This example demonstrates how the scale of your analysis—in this case, country versus state—can impact which variables are relevant.

- If you would like to continue with this analysis, proceed to the optional stretch goal; otherwise, save the project and exit ArcGIS Pro.

Stretch goal (Optional)

If you would like to continue analyzing variable relationships, you can continue running the Local Bivariate Relationships tool on other variables in the dataset.

The following is a list of high-level tasks that you can complete to continue this analysis:

1. Run the Local Bivariate Relationships tool using other variables.
2. Share your results in the Lesson Forum, using the hashtag **#stretch** in the posting title.
3. In the Lesson Forum, identify 2-3 variables that you want to use in the prediction model.

Use the Lesson Forum to post your questions, observations, and syntax examples. Be sure to include the **#stretch** hashtag in the posting title.