


Exercise 2: Explore data using data visualization techniques

 How can I print an exercise to PDF format?

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. Various data visualization techniques are 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.

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.

Note: The exercises in this course include View Result links. Click these links to confirm that your results match what is expected.

Estimated completion time in minutes: 80 minutes

Expand all steps ▼

Collapse all steps ▲

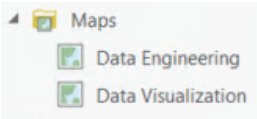
- Step 1: Open an ArcGIS Pro project

To begin, you will open the ArcGIS Pro project package that you downloaded previously.

- a Start ArcGIS Pro.
- b If necessary, sign in with the provided course ArcGIS account.
- c Near Recent Projects, click Open Another Project.

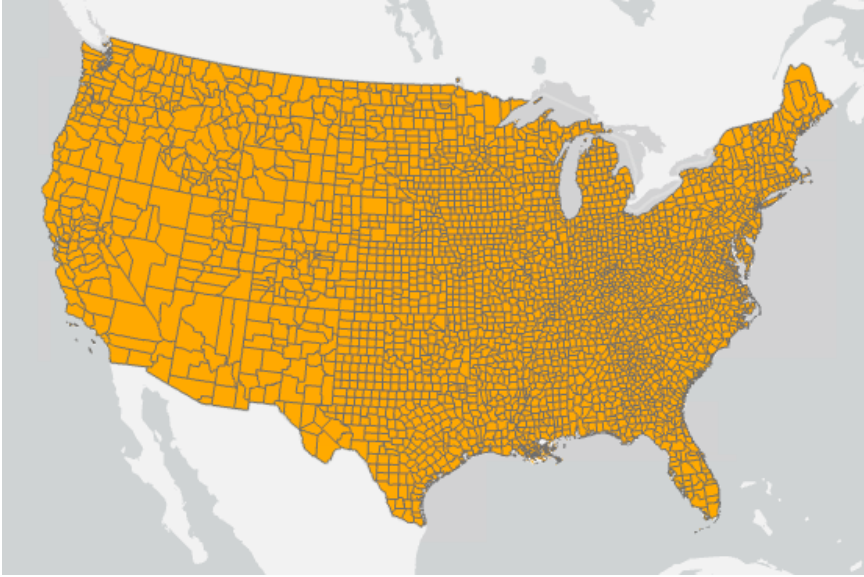
Note: If you have configured ArcGIS Pro to start without a project template or with a default project, you will not see the Start page. On the Project tab, click Open, and then click Open Another Project.

- d Browse to the DataEngineering_and_Visualization folder that you saved on your computer.
- e Click DataEngineering_and_Visualization.aprx to select it, and then click OK.
- f In the Catalog pane, expand Maps.



Step 1f***: Open an ArcGIS Pro project.

- g Right-click Data Visualization and choose Open.



*Step 1g***: Open an ArcGIS Pro project.*

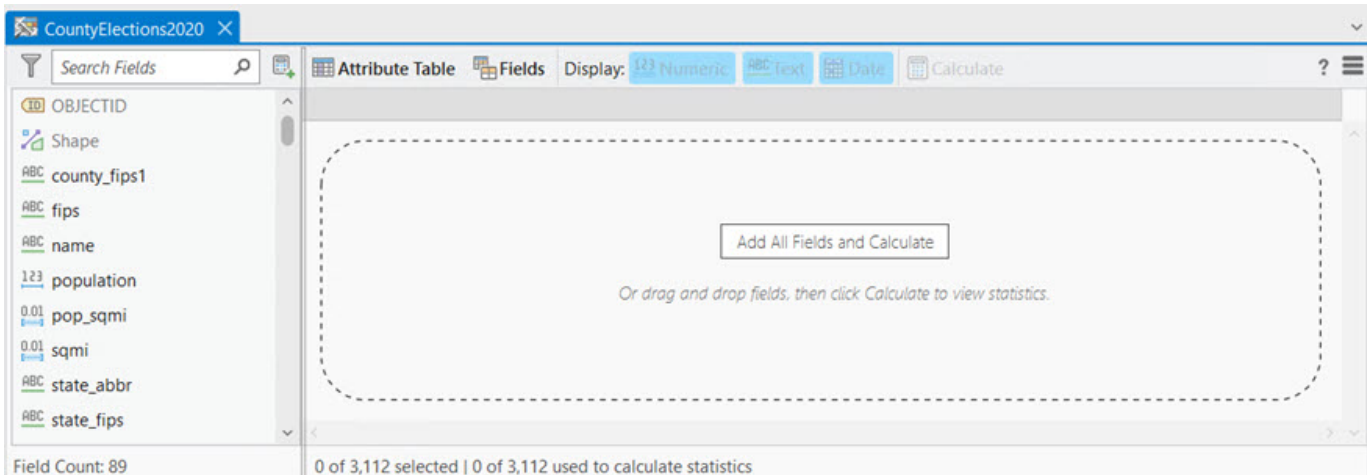
The color in your map might be different.

A Data Visualization map tab opens to a gray basemap with a map layer that contains the 2020 election results for counties that have been enriched with demographic variables. The CountyElections2020 layer has also been projected to the USA Contiguous Equidistant Conic projected coordinate system. The feature class that you created using the Data Engineering Notebook is projected in the WGS84 coordinate system, which is a standard coordinate system for web mapping applications. However, this projection does not preserve areas, distances, or angles. Because this layer will be used for a distance-based analysis, it is best practice to use an equidistant projection to preserve true distance measurements on your map.

- Step 2: Explore fields in the Data Engineering view

The 2020 election results layer (CountyElections2020) includes voter turnout and demographic variables geoenriched to the layer. This layer was created using the same workflow that you explored in the first exercise of this section. First, you will explore some of the fields in the 2020 election results layer using the Data Engineering view to ensure that the data is ready to use in your predictive analysis.

- In the Contents pane, right-click CountyElections2020 and choose Data Engineering.



*Step 2a***: Explore fields in the Data Engineering view.*

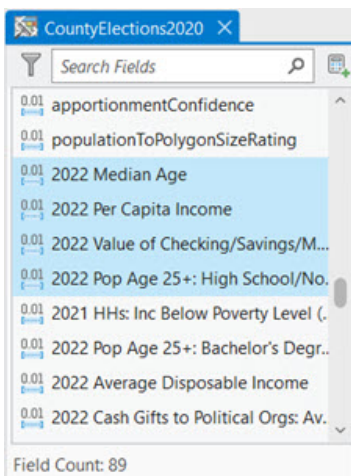
The Data Engineering view opens. You will explore the fields in the layer and select four fields to view their statistics to gain a better understanding of the values and distribution of the fields.

- In the fields panel, scroll down to familiarize yourself with the fields, or variables, that are included in the dataset.

Note: You can adjust the width and height of any of the docked windows in ArcGIS Pro to view the information.

- Locate and click the 2022 Median Age variable.

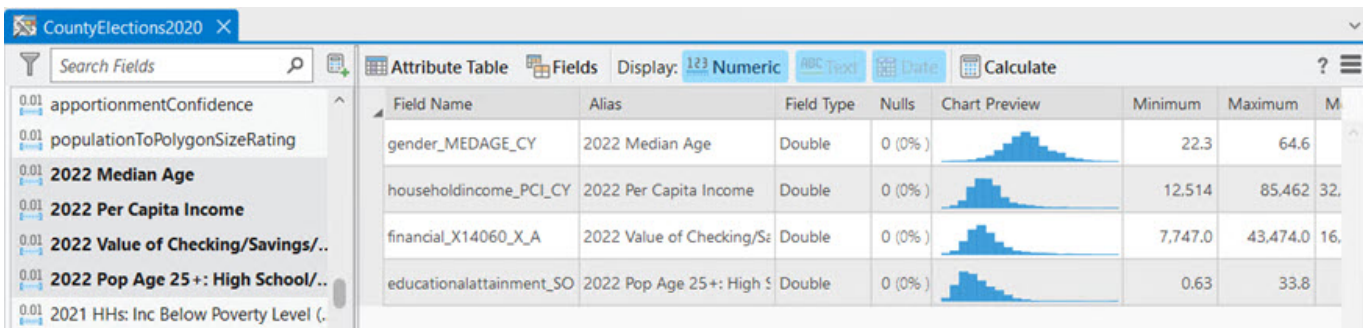
- d On your keyboard, press and hold the Shift key and click the 2022 Pop Age 25+: High School/No Diploma field.



*Step 2d***: Explore fields in the Data Engineering view.*

The four fields that you want to explore are selected in the Data Engineering view.

- e Right-click the selected fields and choose Add To Statistics And Calculate.



*Step 2e***: Explore fields in the Data Engineering view.*

The fields that you selected are added to the statistics panel, and the statistics are calculated. Each row corresponds to a field, and each column shows a different metric or statistic for each field.

- f In the statistics panel, scroll to the right and explore the statistics available for each field.

This quick review of your data shows that these fields are not skewed and can be used in the predictive analysis.

- g Save the project.

Next, you will use charts to visualize and explore the distribution of a field.

- Step 3: Visualize the data distribution

In this step, you will explore the data attributes of the 2020 election results layer and visualize the data distribution.

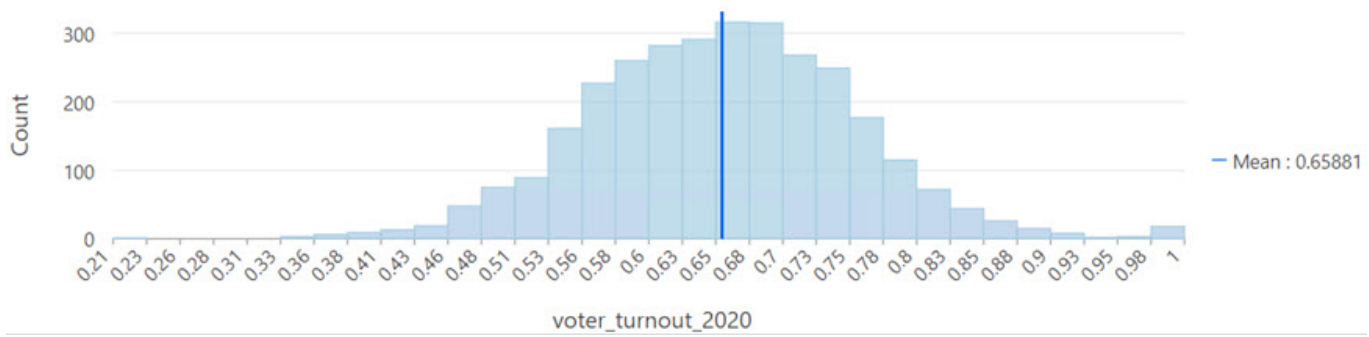
- a In the Data Engineering view, click Attribute Table.

The CountyElections2020 attribute table opens in a new tab. Earlier, you explored these same fields.

- b To the left of the attribute table tab, on the Data Engineering view tab, click the Close button to close the Data Engineering view.

- c In the CountyElections2020 attribute table, scroll to the right and, right-click the Voter_Turnout_2020 column and choose Statistics.

Distribution of voter_turnout_2020



Step 3c***: Visualize the data distribution.

Note: Your figures may be slightly different in this step.

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

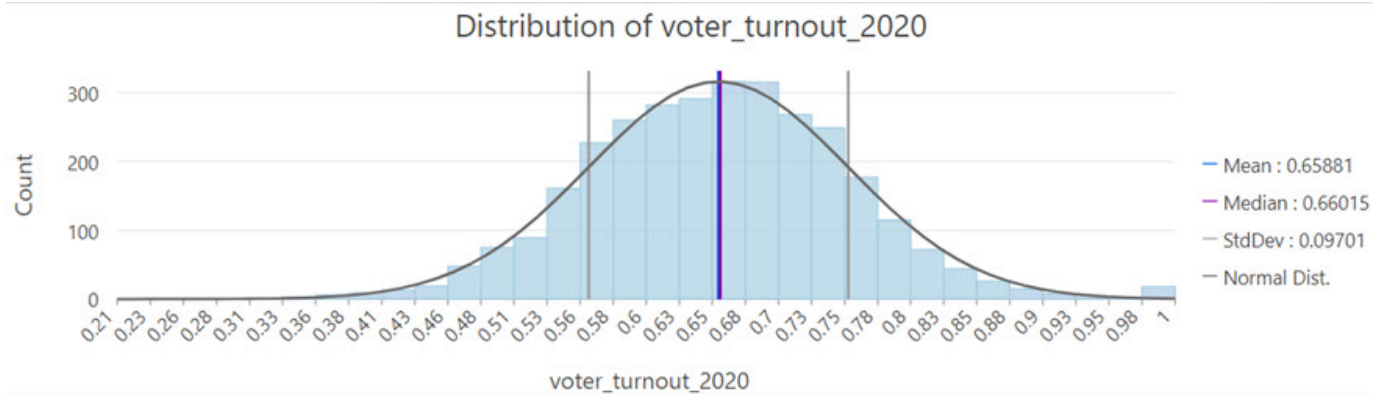
The histogram visually summarizes the distribution of voter turnout by measuring the frequency at which values appear in the dataset. The x-axis represents different ranges, or bins, of voter turnout values. The y-axis measures the number of counties with voter turnout values falling within each range.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

d Examine the Chart Properties pane.

The Chart Properties pane includes a list of statistics summarizing the Voter_Turnout_2020 variable. The mean county voter turnout value is about 0.66, with county values ranging from approximately 0.21 to approximately 1. By default, a line representing the mean value is overlaid on the histogram. You can add additional overlays from the Chart Properties pane.

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

f Under Statistics, check the Median and Std. Dev. boxes.



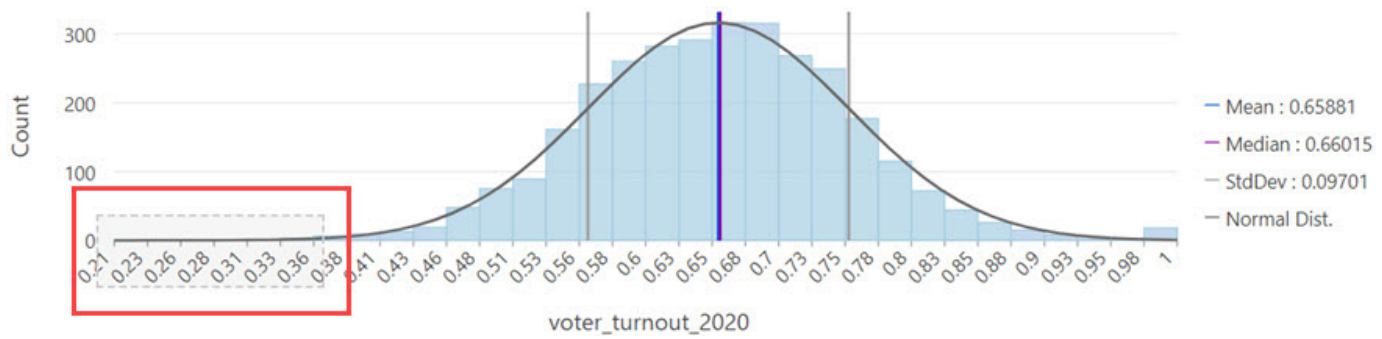
Step 3f***: Visualize the data distribution.

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 outcome indicates that the Voter_Turnout_2020 variable is normally distributed. In a normal distribution, the mean, median, and mode values are equal. This outcome 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.

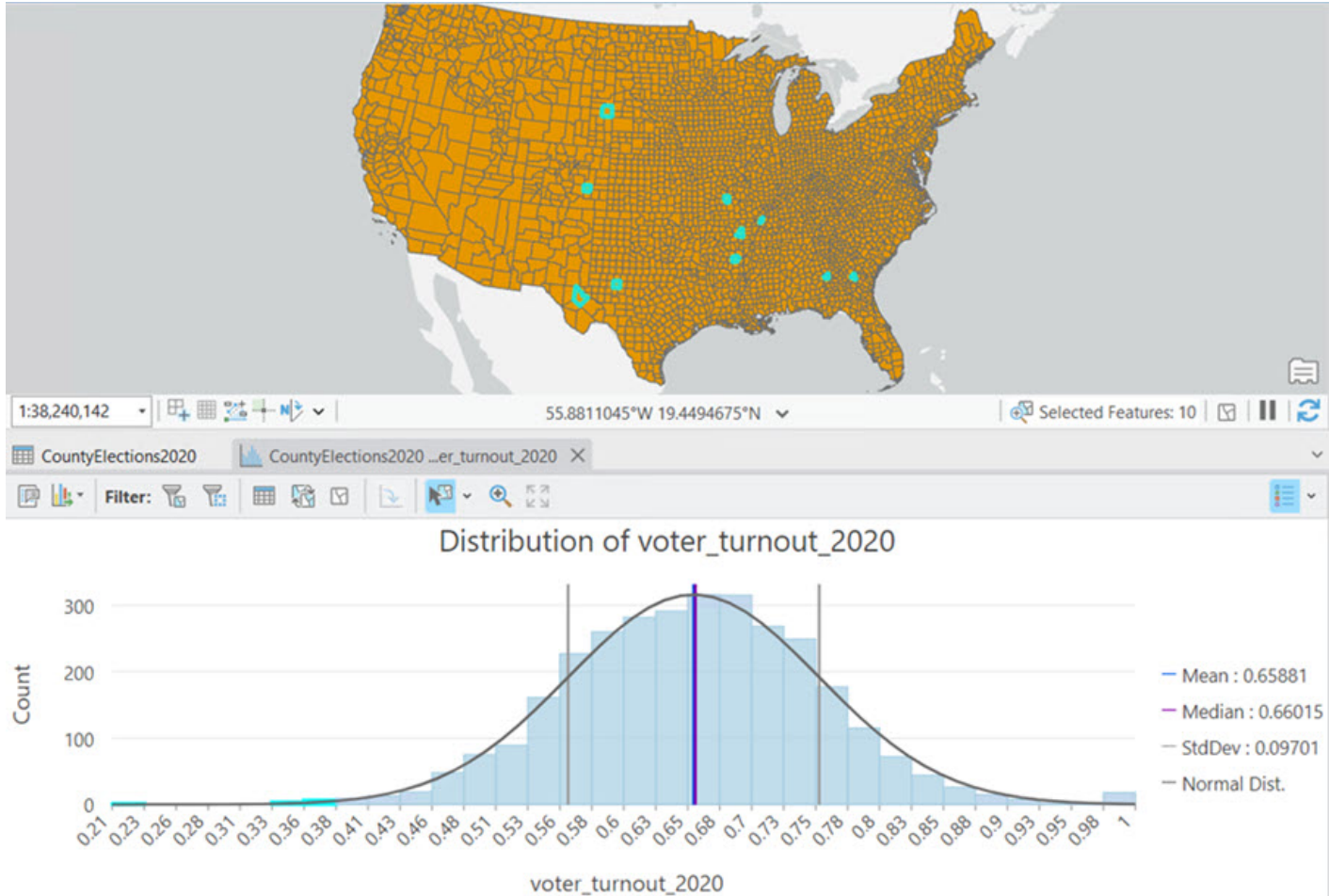
You can also use the histogram to interactively select features on the map based on their voter turnout values.

g In the chart view, drag a box around the bins with the lowest voter turnout values in the left tail of the histogram, as shown in the following graphic.

Distribution of voter_turnout_2020



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

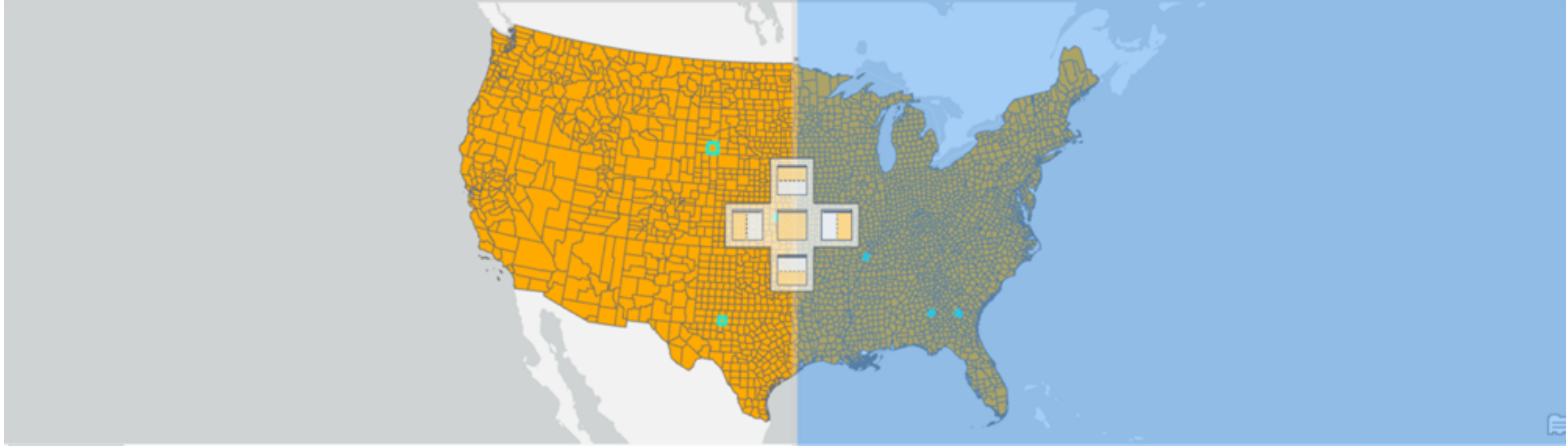


Step 3g***: Visualize the data distribution.


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.

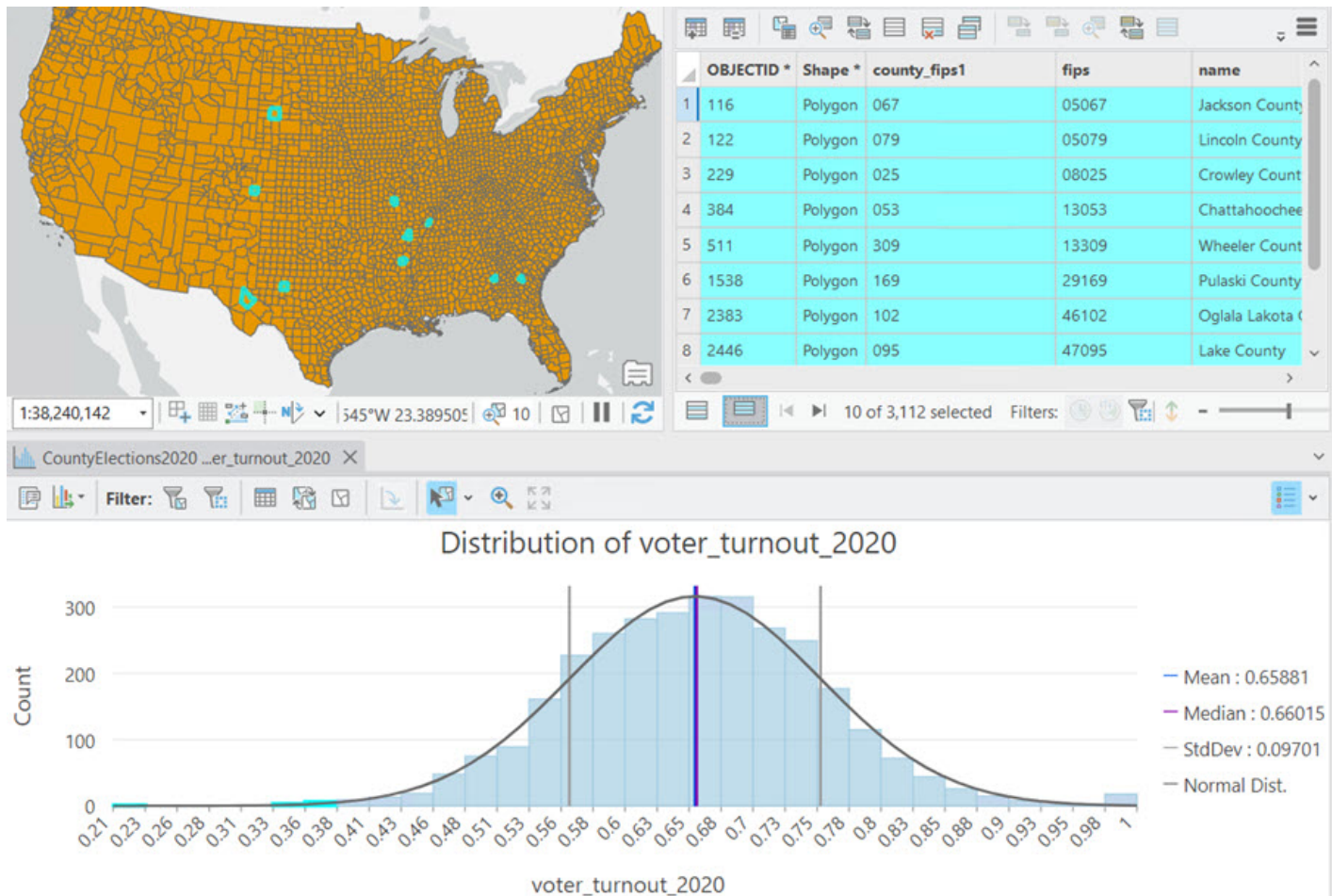
h Click and hold the CountyElections2020 attribute table tab.

i Drag the attribute table tab to the center of the map until the docking target appears, as shown in the following graphic.



The attribute table is represented by a blue shadow, and docking targets appear in the center of the map view and at the edges of the application window. Each target represents an area where the window can be positioned. The blue shadow displays where the attribute table will be docked when you release the click.

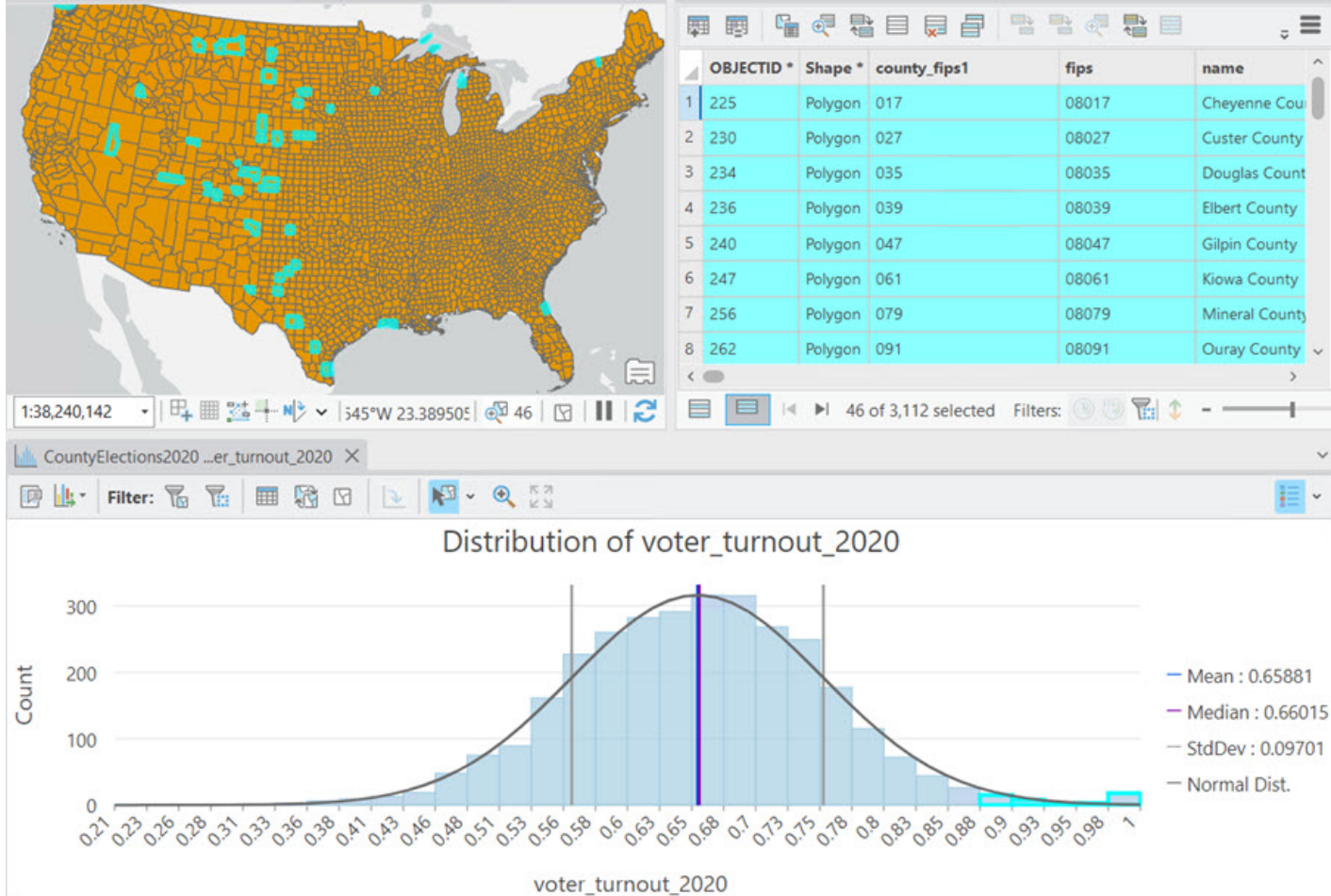
- j Pause over the right-hand docking target, and then release the click to dock the attribute table to the right of the map and above the histogram.
- k At the bottom of the attribute table window, click the Show Selected Records button .




Step 3k***: Visualize the data distribution.

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

- l In the right tail of the histogram, drag a box around the bins with the five highest voter turnout values, as shown in the following graphic.



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

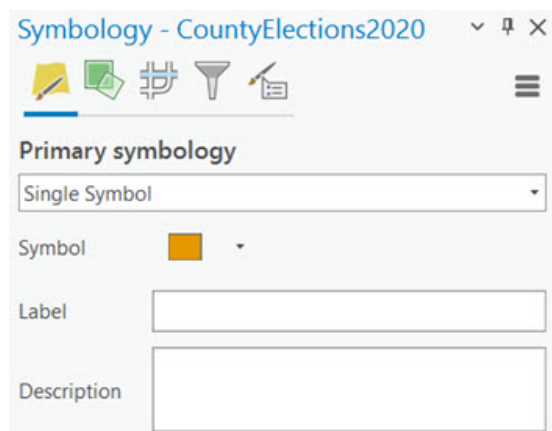
- m In the histogram, click any blank area to clear the selection.
- n Close the attribute table and the histogram chart view by clicking the Close button  on each tab.

You 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 4: Change layer symbology

Currently, every county in the 2020 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 2020 voter turnout.

- a In the Contents pane, right-click CountyElections2020 and choose Symbology.



Step 4a***: Change layer symbology.

The Symbology pane appears.

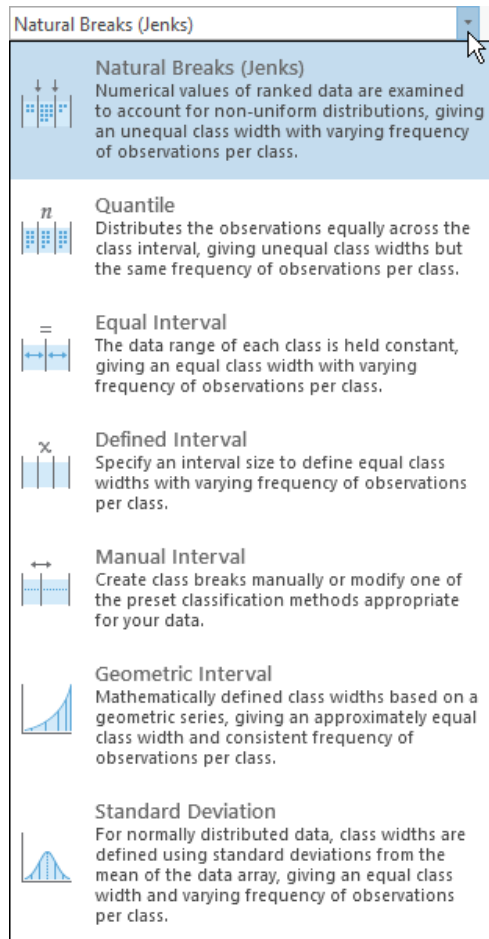
b In the Symbology pane, under Primary Symbolology, click the down arrow and choose Graduated Colors.

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 color symbology, go to 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, for Field, choose Voter_Turnout_2020.

d For Method, click the down arrow.



Step 4d***: Change layer symbology.

A list describing the available classification methods appears.

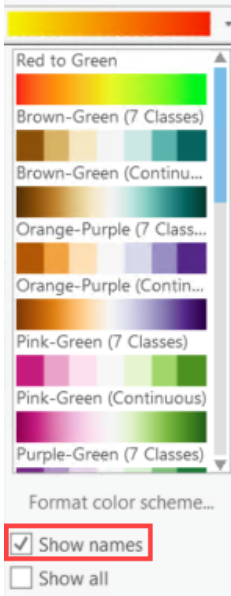
Natural Breaks (Jenks) is the default classification method because it is data driven. That means the symbol ranges are calculated based on the data values, making it adaptable to different types of data distributions. Because you have determined that the voter turnout variable is normally distributed,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

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

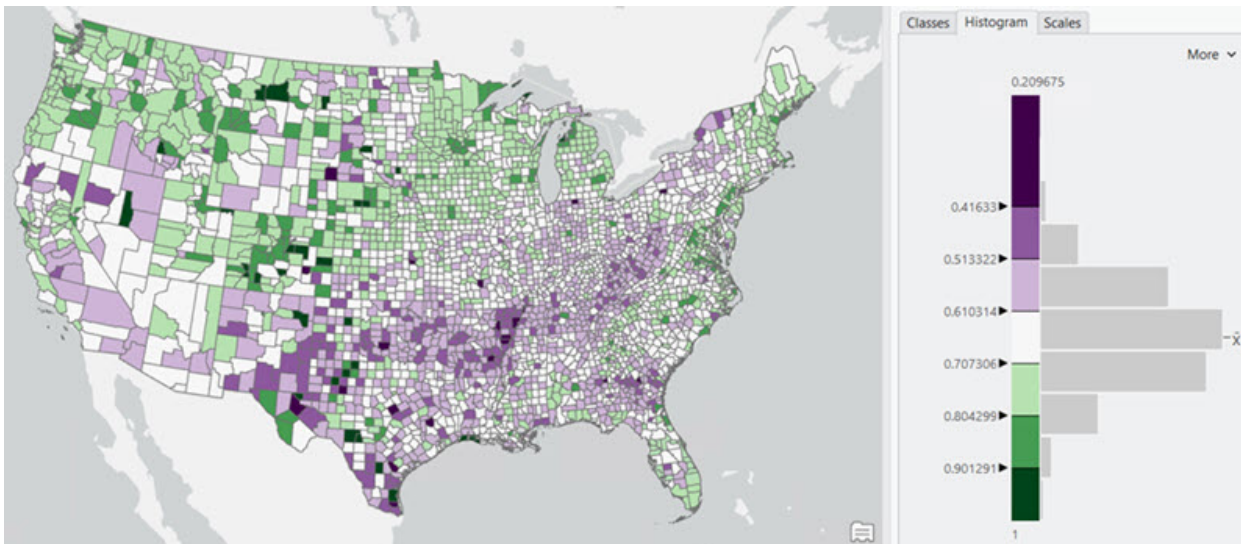
f For Color Scheme, click the down arrow.

Note: If you do not see the Color Scheme parameter, below Classes, drag the pane divider (the bar with the three horizontal dots) down until you see the Color Scheme parameter.

g In the Color Scheme window, check the Show Names box, as shown in the following graphic.




- h Choose the Purple-Green (Continuous) color scheme.
- i In the Symbology pane, below the pane divider, click the Histogram tab.

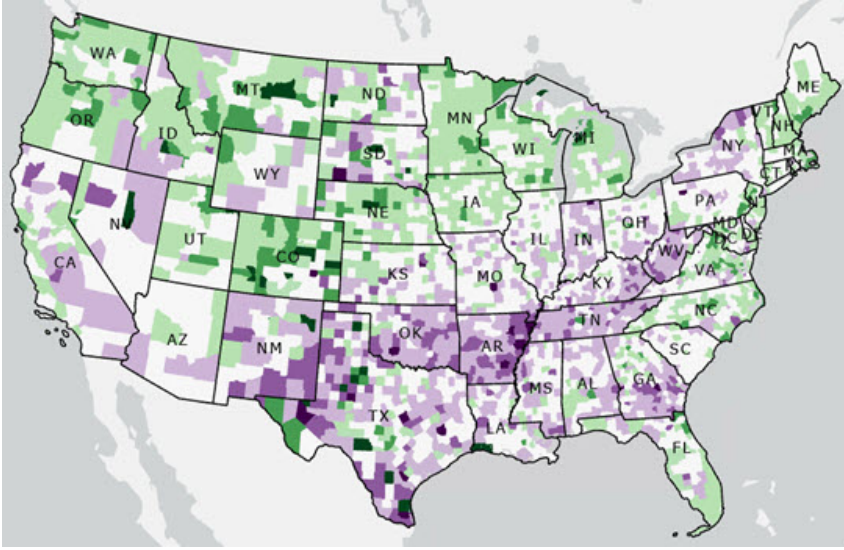


*Step 4i***: Change layer symbology.*

Note: You may need to widen the Symbology pane to better see the histogram.

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.



Step 4k***: Change layer symbology.

After 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), Tennessee (TN), and Arkansas (AR) stand out as having low voter turnout, and states like Colorado (CO), Minnesota (MN), and Wisconsin (WI) stand out as having high voter turnout. You will use a bar chart to summarize and compare voter turnout by state.

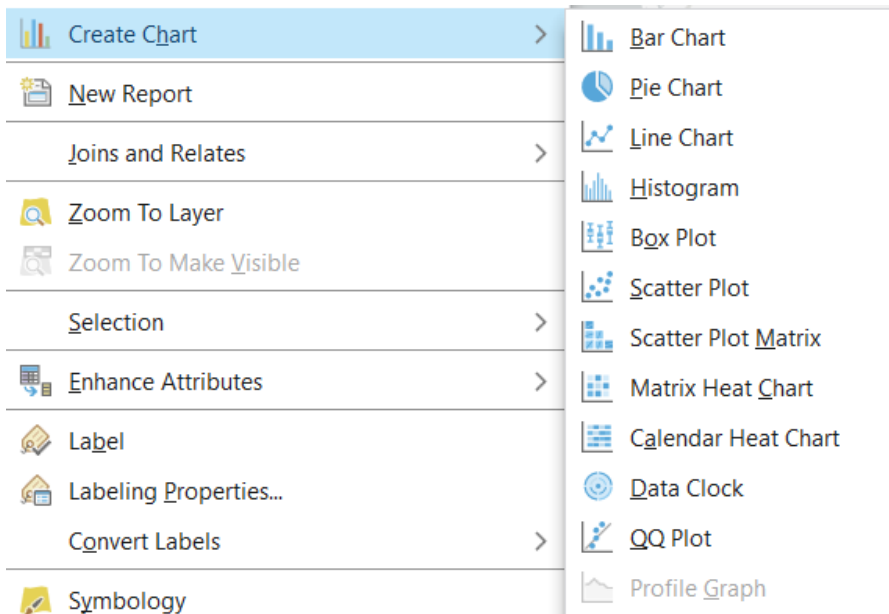
l Close the Symbology pane.

m Save the project.

- Step 5: Create a bar chart

ArcGIS Pro provides numerous ways to visualize your data in different charts. You will now visualize the data as a bar chart.

a In the Contents pane, right-click CountyElections2020, point to Create Chart, and view the available options.



Step 5a***: Create a bar chart.

b Choose Bar Chart.

c In the Chart Properties pane, specify the following parameters:

- For Category Or Date, choose State_Name.
- For Aggregation, choose Mean.

- For Numeric Fields, click Select and check the box for Voter_Turnout_2020.

Chart Properties

Mean of by state_name

Data Series Axes Guides Format General ... ?

Variables

Category or Date

state_name

Aggregation

Mean

* Numeric field(s)

+ Select

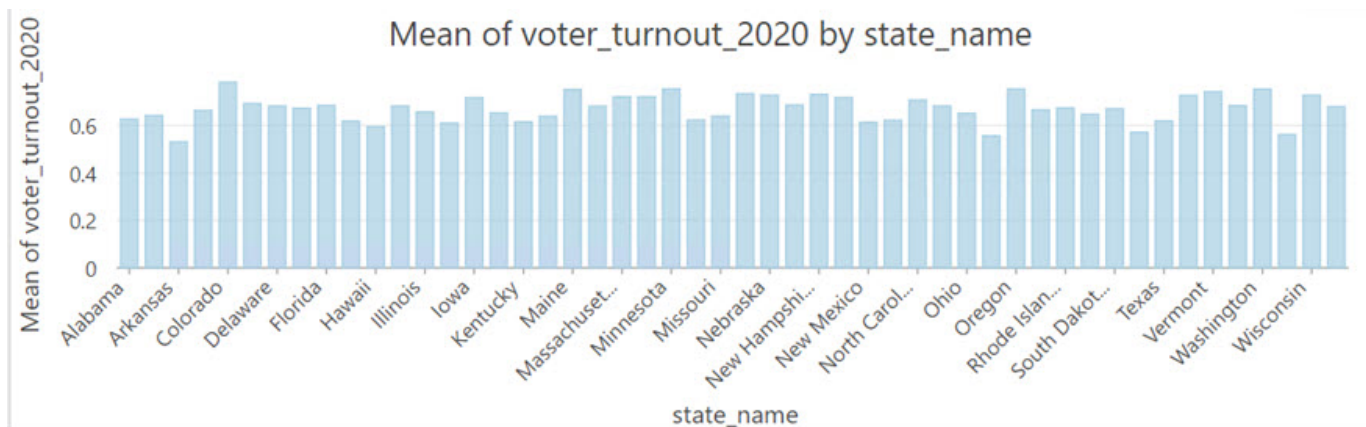
Search

- ☐ voter_turnout_2016
- ☒ voter_turnout_2020
- ☐ voter_turnout_dem_2008
- ☐ voter_turnout_dem_2012
- ☐ voter_turnout_dem_2016
- ☐ voter_turnout_dem_2020
- ☐ voter_turnout_gop_2008
- ☐ voter_turnout_gop_2012
- ☐ voter_turnout_gop_2016
- ☐ voter_turnout_gop_2020
- ☐ pctdiff_dem_vs_gop_2008

Apply Cancel

Step 5c***: Create a bar chart.

- d Click Apply.

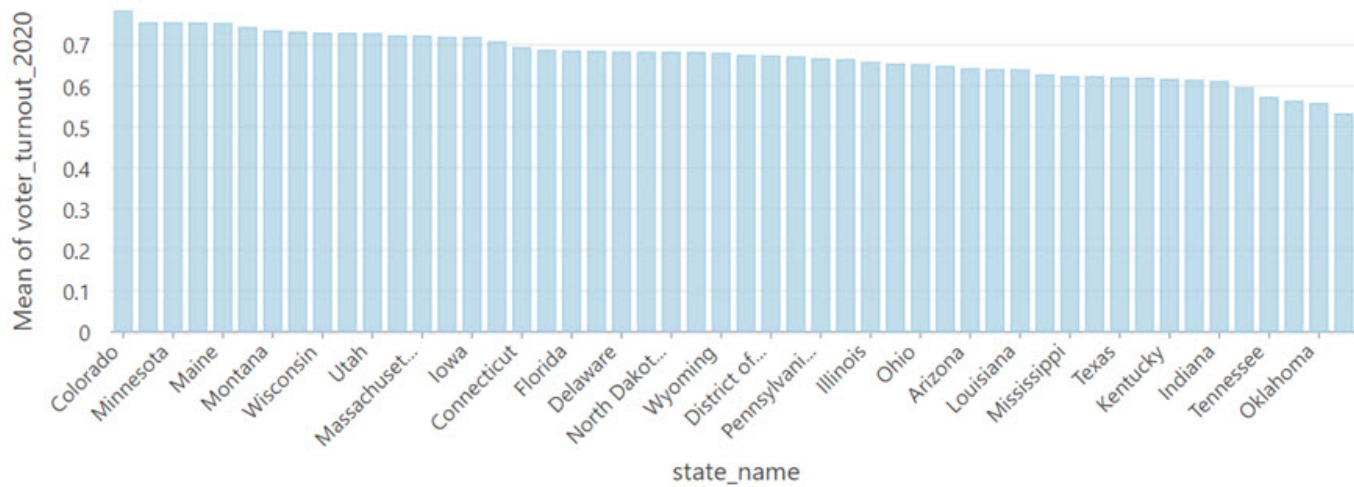


Step 5d***: Create a bar chart.

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, go to ArcGIS Pro Help: Bar chart.

- e At the bottom of the Chart Properties pane, under Sort, click the down arrow and choose Y-Axis Descending.

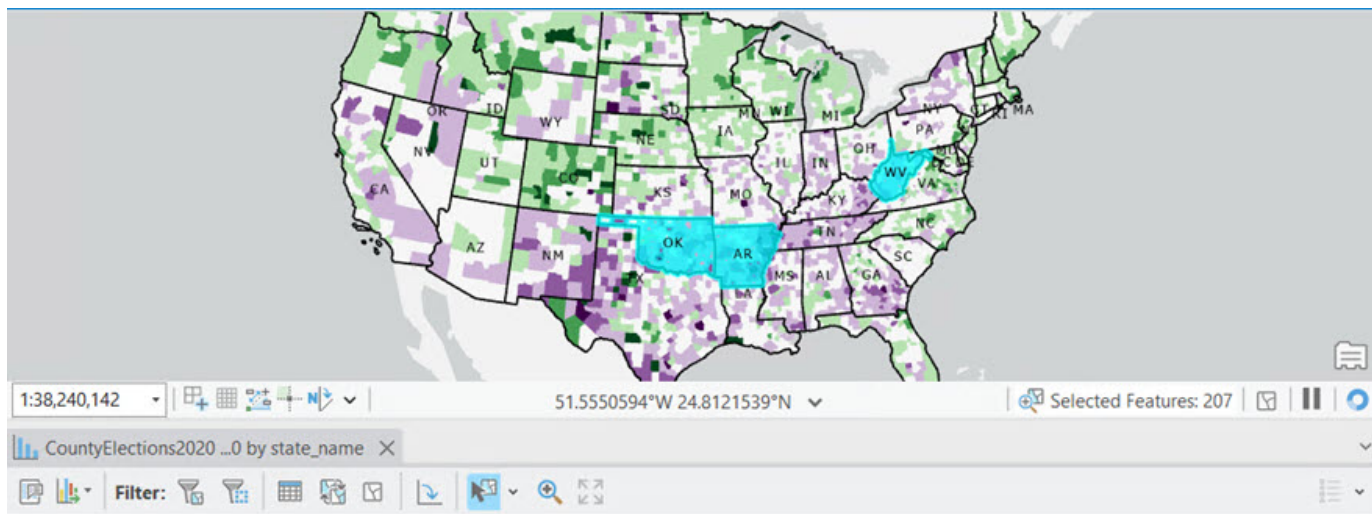
Mean of voter_turnout_2020 by state_name



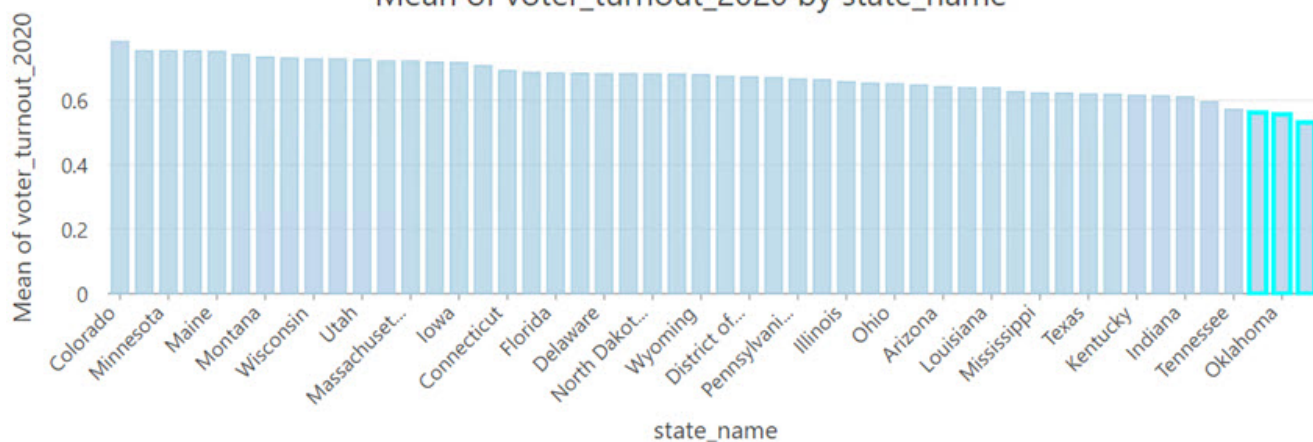
Step 5e***: Create a bar chart.

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

f In the chart view, select the three states whose counties have the lowest voter turnout on average.



Mean of voter_turnout_2020 by state_name



Step 5f***: Create a bar chart.

West Virginia, Arkansas, and Oklahoma have the lowest county average voter turnout for 2020, 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.

g In the chart view, clear the selection.

- Step 6: Filter a chart using a selection

In this step, you will use a filtered bar chart to examine the individual county voter turnout within each state.

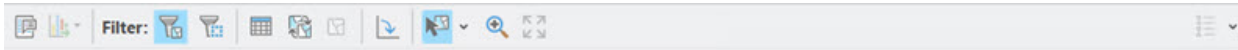
a Create a second bar chart for the CountyElections2020 layer using the following parameters:

- Category Or Date: Name
- Aggregation: <None>
- Numeric Fields: Voter_Turnout_2020

b Click Apply.

c At the bottom of the Chart Properties pane, under Sort, click the down arrow and choose Y-Axis Descending.

d In the chart view, on the toolbar next to Filter, click the Filter By Selection button .

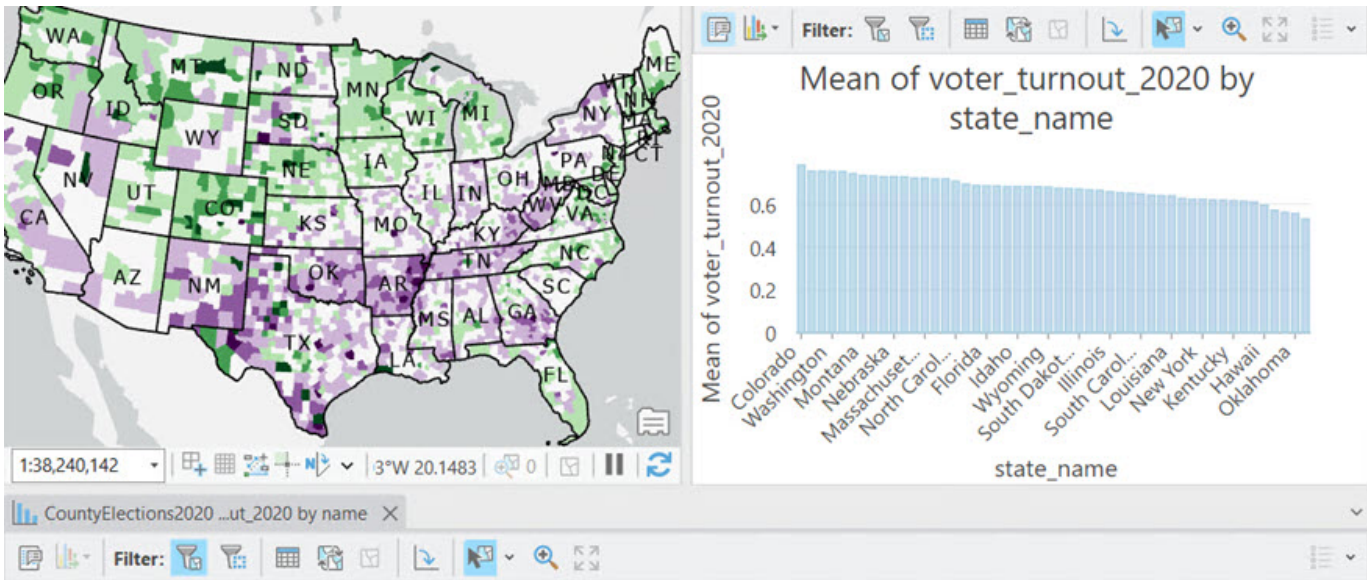


No data is available to display due to empty data field(s) and/or combination of filters.

*Step 6d***: Filter a chart using a selection.*

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

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






No data is available to display due to empty data field(s) and/or combination of filters.

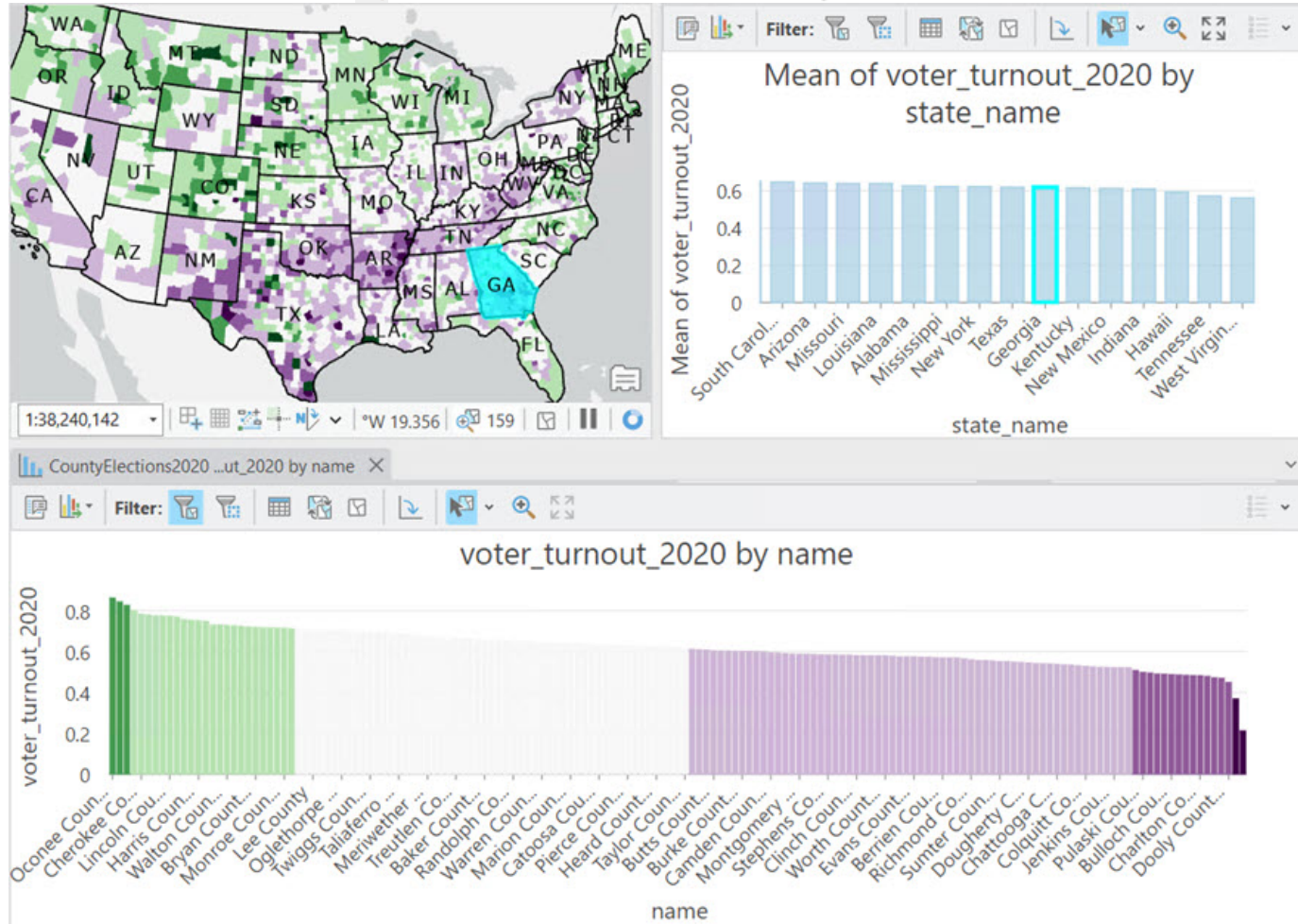
*Step 6e***: Filter a chart using a selection.*

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

f In the state bar chart, select Georgia.

Georgia is located between Texas and Kentucky in the chart. Point to the bar to see which state it represents. You can use the Zoom Mode button  or your mouse to zoom in to the bar chart.

If you use the Zoom Mode button , click the Select Interaction Mode button  so that you can select a state.



Step 6f***: Filter a chart using a selection.

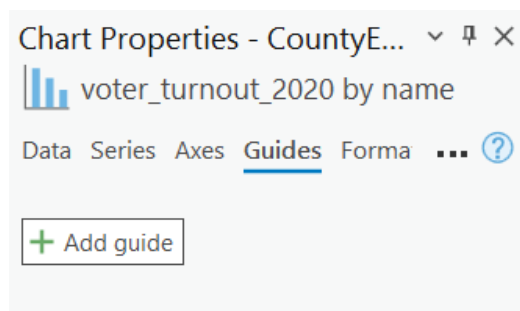
The county bar chart populates to show the individual county values within Georgia. Each bar in the county bar chart corresponds to a single feature on the map, so the colors of the bars match the map symbology.

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 Click the county (name) bar chart view tab to activate that chart.

The Chart Properties pane updates to show county information.

h At the top of the Chart Properties pane, click the Guides tab.



Step 6h***: Filter a chart using a selection.

i Under Guides, click Add Guide.

- j For Value, type **0.66** in the first field.
- k For Line Color, click the line and choose a bright blue color.
- l For Label, type **National Average**.

Chart Properties

voter_turnout_2020 by name

Data Series Axes **Guides** Format General ... ?

+ Add guide

▼ ☒ Guide 1 ✎ ✕

Value to

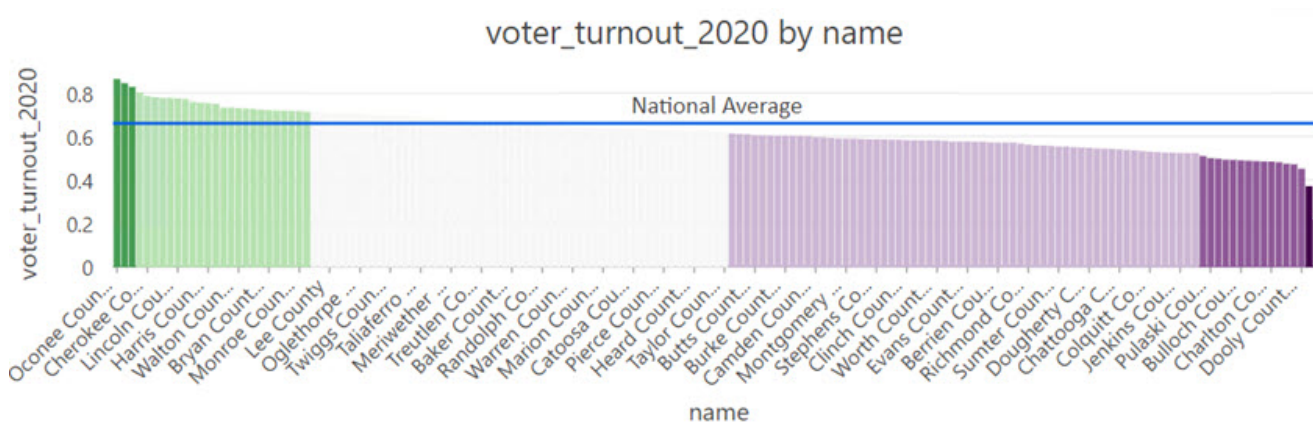
Line Color

Line type ▼

Label

Step 6l***: Filter a chart using a selection.

In the county bar chart, a line appears, marking the national average voter turnout value.



Step 6l***: Filter a chart using a selection.

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

- m 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.
- n Clear the selection and close both chart views.

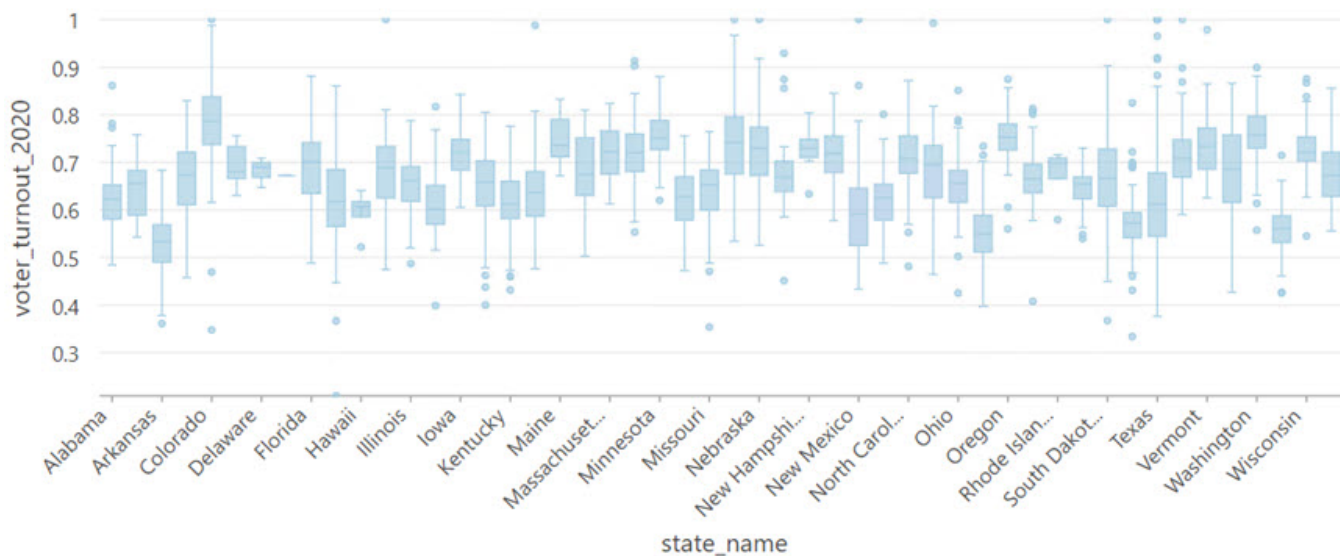
You used bar charts and the Filter By Selection method to explore state voter turnout averages and to examine the individual county voter turnout values for each state.

- Step 7: Create a box plot

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 one time, you will create a box plot.

- a Create a box plot for the CountyElections2020 layer using the following parameters:
 - For Numeric Fields, click Select, check the box for Voter_Turnout_2020, and then click Apply.
 - For Category, choose State_Name.
 - For Show Outliers, check the box.

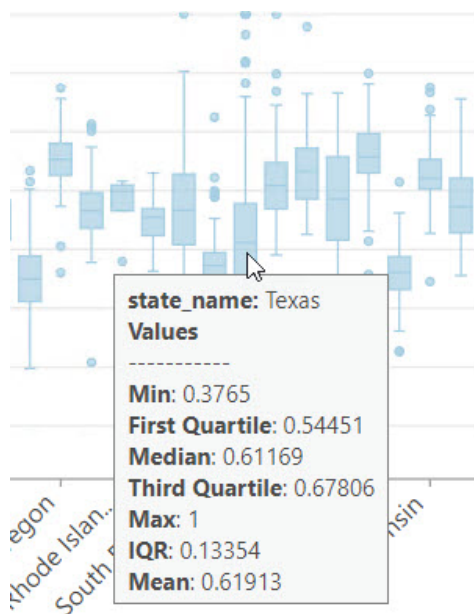
Distribution of voter_turnout_2020 by state_name



Step 7a***: Create a box plot.

Box plots are automatically sorted alphabetically by their categories (x-axis ascending). You can change the sorting order in the Chart Properties pane. 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 they 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, go to ArcGIS Pro Help: Box plot.

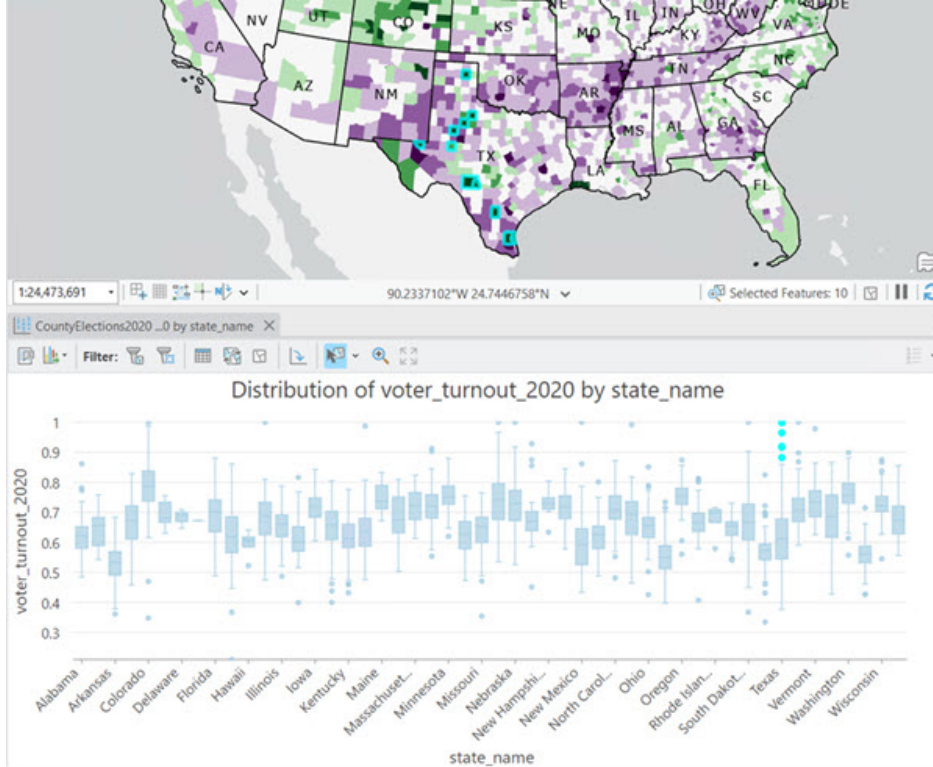
b In the box plot chart, point to Texas.



Step 7b***: Create a box plot.

The ToolTip displays the key voter turnout statistics for the state. Texas overall has a relatively low voter turnout average. However, there is a wide range of county voter turnout values, spanning from approximately 0.38 to approximately 1. 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.

c In the box plot chart, select the Texas outliers, as shown in the following graphic.



The outliers are selected on the map.

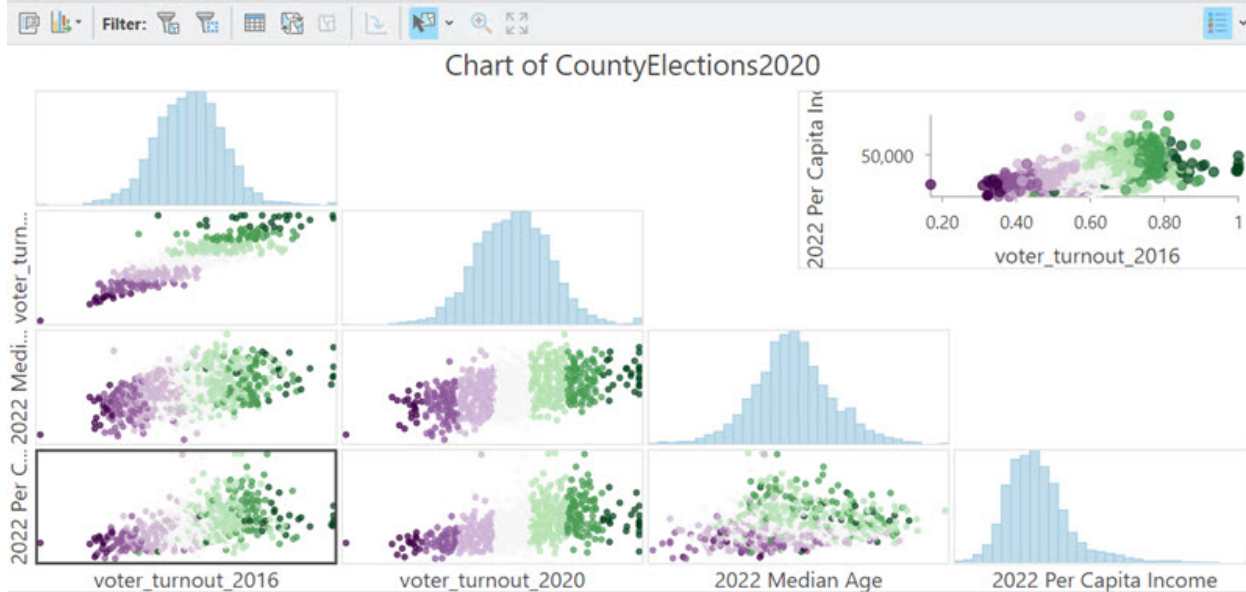
- d Clear the selection and close the box plot chart.

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.

- Step 8: Explore variable relationships in a scatterplot matrix

You have used various spatial and nonspatial 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.

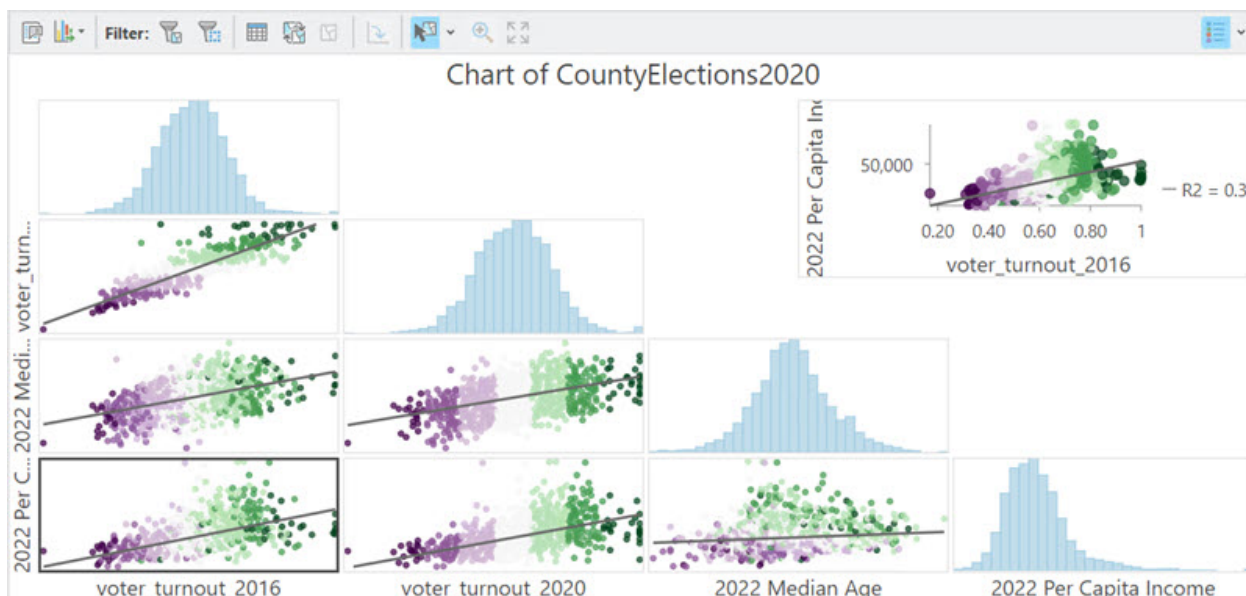
- a Create a scatterplot matrix for the CountyElections2020 layer.
- b In the Chart Properties pane, under Numeric Fields, click Select and check the boxes for the following fields:
 - Voter_Turnout_2016
 - Voter_Turnout_2020
 - 2022 Median Age
 - 2022 Per Capita Income
- c Click Apply.
- d Under Matrix Layout, next to Diagonal, click the down arrow and choose Histograms.



Step 8d***: Explore variable relationships in a scatterplot matrix.

A scatterplot matrix is a grid of scatterplots, also referred to as mini-plots, used to visualize bivariate relationships between combinations of variables. Each scatterplot 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 scatterplot matrices, see ArcGIS Pro Help: Scatter plot matrix.

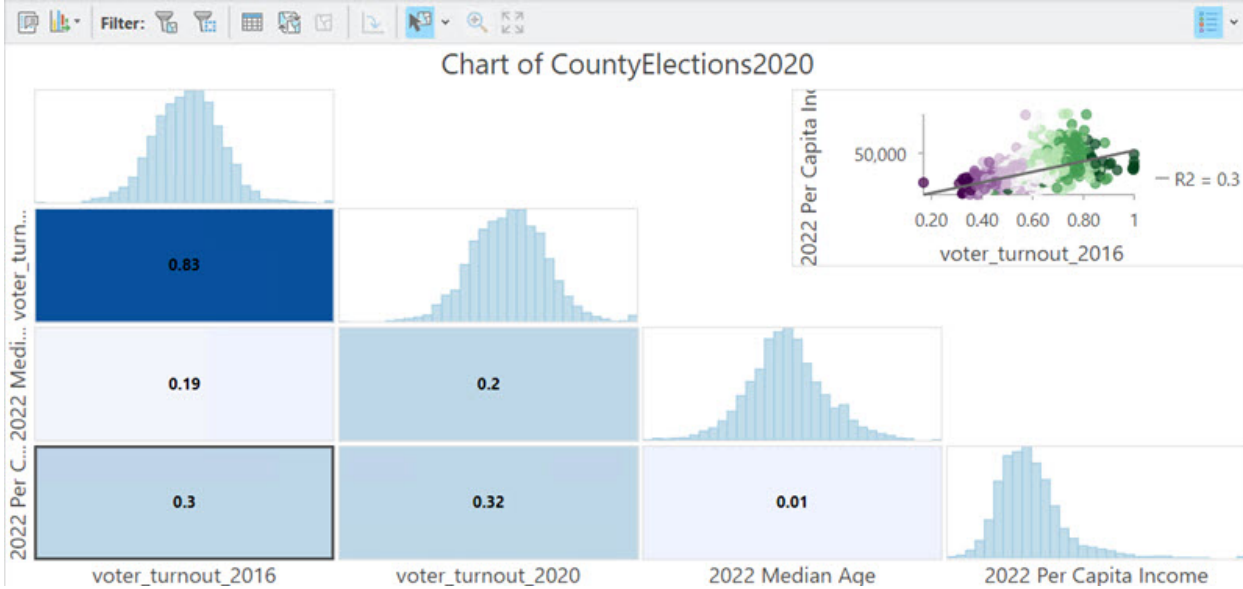
- e In the Chart Properties pane, check the box for Show Linear Trend.



Step 8e***: Explore variable relationships in a scatterplot matrix.

A linear trend line is added to each scatterplot 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 scatterplots, go to ArcGIS Pro Help: Scatter plot.

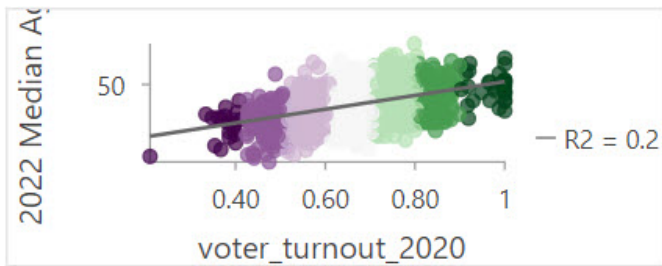
- f Under Matrix Layout, next to Lower Left, click the down arrow and choose R-Squared.



Step 8f***: Explore variable relationships in a scatterplot matrix.

The mini-plots in the matrix are now visualized with a color gradient that corresponds 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 in 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.

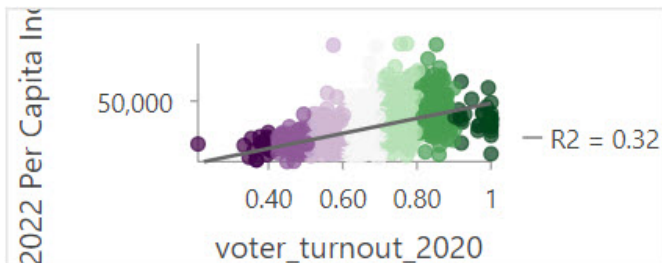
- g In the scatterplot matrix, select the mini-plot that compares Voter_Turnout_2020 and 2022 Median Age.



Step 8g***: Explore variable relationships in a scatterplot matrix.

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 0.2, which means that median age alone can explain only about 20 percent of the variability in the voter turnout values.

- h Select the mini-plot that compares Voter_Turnout_2020 and 2022 Per Capita Income.



Step 8h***: Explore variable relationships in a scatterplot matrix.

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 0.32, which means that per capita income can explain about 32 percent 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, as shown in the following graphic.

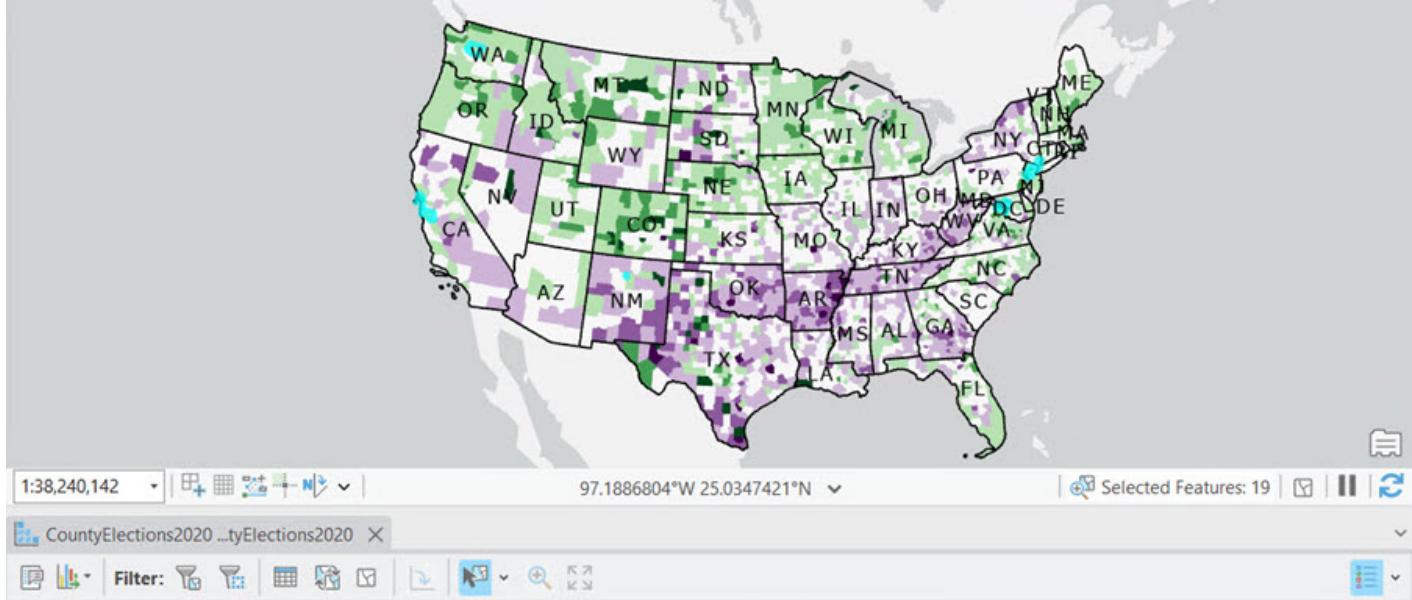
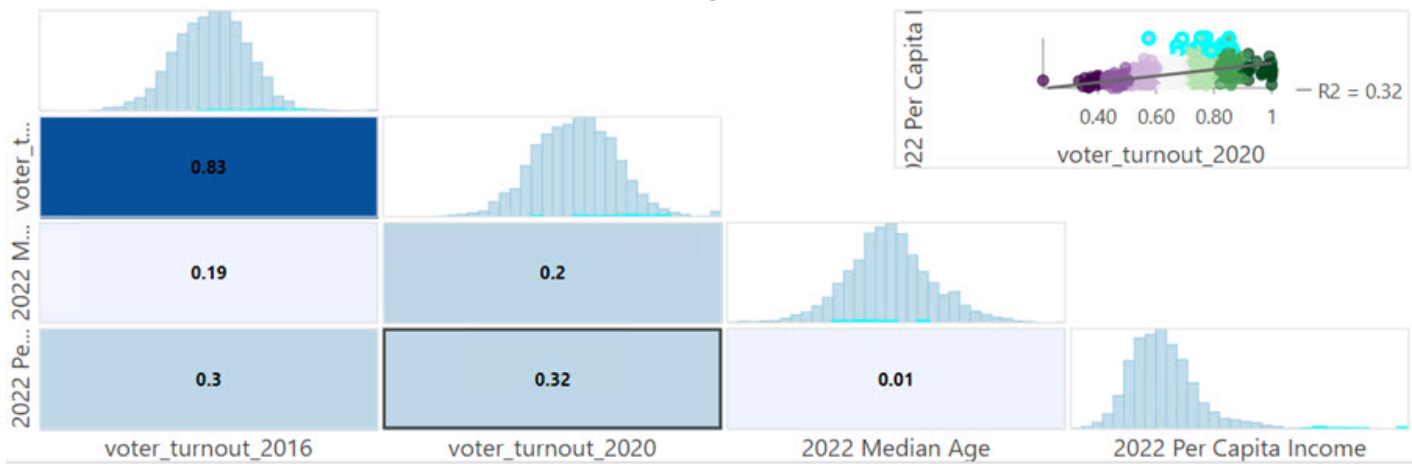


Chart of CountyElections2020




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

j Clear the selection, and then save the project.

- Step 9: Explore relationships at different scales

To investigate whether the strength of the variable relationships varies from place to place, you can filter the scatterplot matrix to include only counties that are visible on the map.

a In the scatterplot matrix view, on the toolbar next to Filter, click the Filter By Extent button .

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

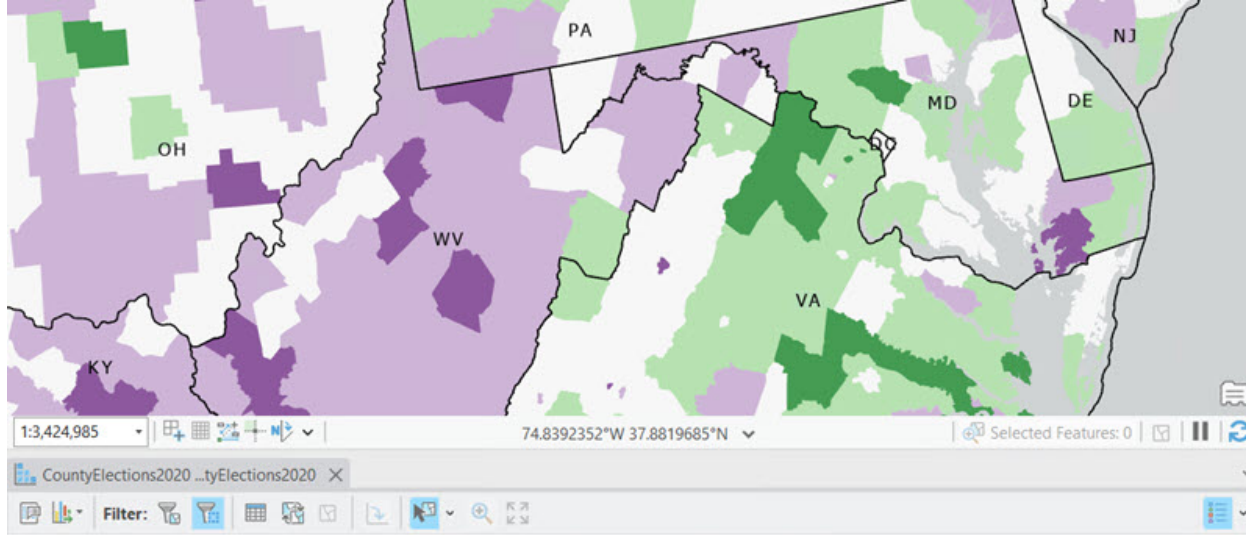
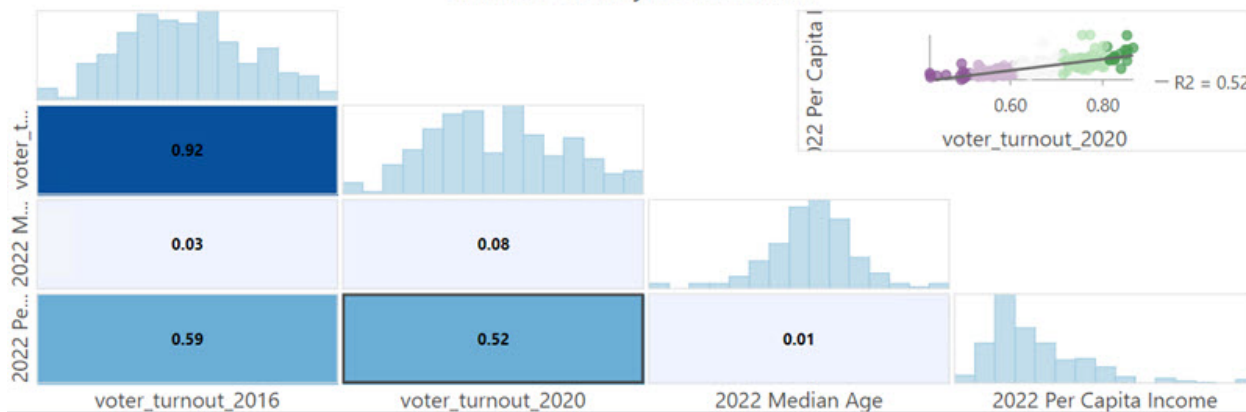


Chart of CountyElections2020



Step 9b***: Explore relationships at different scales.

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

The chart updates to calculate the relationships between the variables of the counties that are 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 have increased. However, the relationship between voter turnout and median age has decreased.

- c Zoom and pan around the map to explore how variable relationships vary by scale and location.
- d When you are done, close the scatterplot matrix view and the Chart Properties pane.
- e In the Contents pane, uncheck CountyElections2020 to turn the layer off.

The changes in R-squared values indicate that the linear relationships between the variables vary spatially. In the stretch goal, you can explore and quantify different types of local relationships by using the Local Bivariate Relationships tool.

- Step 10: Stretch goal (Optional)

ArcGIS Pro includes a comprehensive suite of geoprocessing tools that can perform spatial analysis or manage GIS data in an automated way. 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 then determine whether this relationship varies across the contiguous United States. To learn more about the Local Bivariate Relationships tool, go to ArcGIS Pro Help: How Local Bivariate Relationships works. 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 that is found in that area.

- a Use the following high-level tasks to continue this analysis:
 1. Run the Local Bivariate Relationships tool using voter turnout and per capita income variables.
 2. Turn on Enable Local Scatterplot Pop-ups in the tool.

3. Click a county to open a pop-up window that visualizes the relationship.

4. Run the tool again to explore different variables.

- b Use the Lesson Forum to post your questions and observations. Be sure to include the **#stretch** hashtag in the posting title.
- c Also, in the Lesson Forum, identify two to three variables that you want to use in a prediction model.
- d When you are finished, close the map and Geoprocessing pane, and then save the project and exit ArcGIS Pro.

Based on the information that you find using this tool, you can see how the scale of your analysis (for example, county versus state) can impact which variables are relevant for a prediction analysis.