Exercise 28: Introduction to Spatial Statistics

This exercise will explore a variety of spatial statistical analysis methods made available in ArcGIS Pro. These tools allow you to measure spatial central tendency, spatial variability, and spatial autocorrelation. This lab is simply an introduction, as there are a variety of statistical analysis techniques that can be applied to spatial data. Here, we will make use of data for the city of Pittsburgh and also county-level data for the western United States. The data for Pittsburgh can be obtained here: http://pittsburghpa.gov/dcp/gis/data-maps.html. The county-level data for the western United States is from the 2010 United States Census.

Topics covered in this exercise include:

1. Measure spatial central tendency and dispersion
2. Analyze point patterns
3. Analyze global spatial autocorrelation
4. Find local hot spots and cold spots

Step 1. Open a Map Project

First, we need to download and open the **Exercise_28.aprx** file.

- ☐ Download the **Exercise_28** data from https://www.wvview.org/. All lab materials are available on the course webpage and linked to the exercise. You will need to extract the compressed files and save it to the location of your choosing.
- ☐ Open ArcGIS Pro. This can be done by navigating to All Apps followed by the ArcGIS Folder. Within the ArcGIS Folder, select ArcGIS Pro. Note that you can also use a Task Bar or Desktop shortcut if they are available on your machine.
- ☐ After ArcGIS Pro launches, select "Open another project."

Open another project

- ☐ Navigate to the directory that houses the material for this course. The project files are in the **Exercise_28** folder where it was saved on your local machine.
- ☐ Select **Exercise_28.aprx**. Click OK to open the project.
- ☐ If necessary, navigate to the WV map.

Note: If you'd prefer, you can also just click on the **Exercise_28.aprx** file within the uncompressed folder directly to launch ArcGIS Pro.

This project contains two maps. The Pittsburgh map contains the following layers.

universities: point locations of universities

trees: point location of trees

schools: point locations of schools

PAT_stops: point locations of bus stops along the Port Authority of Allegheny County transit routes

hospitals: point locations of hospitals

City_Boundary: polygon municipal boundary of the city of Pittsburgh

The West map only contains one layer: **west_counties**. This is a polygon layer that provides all of the county boundaries in the western United States.
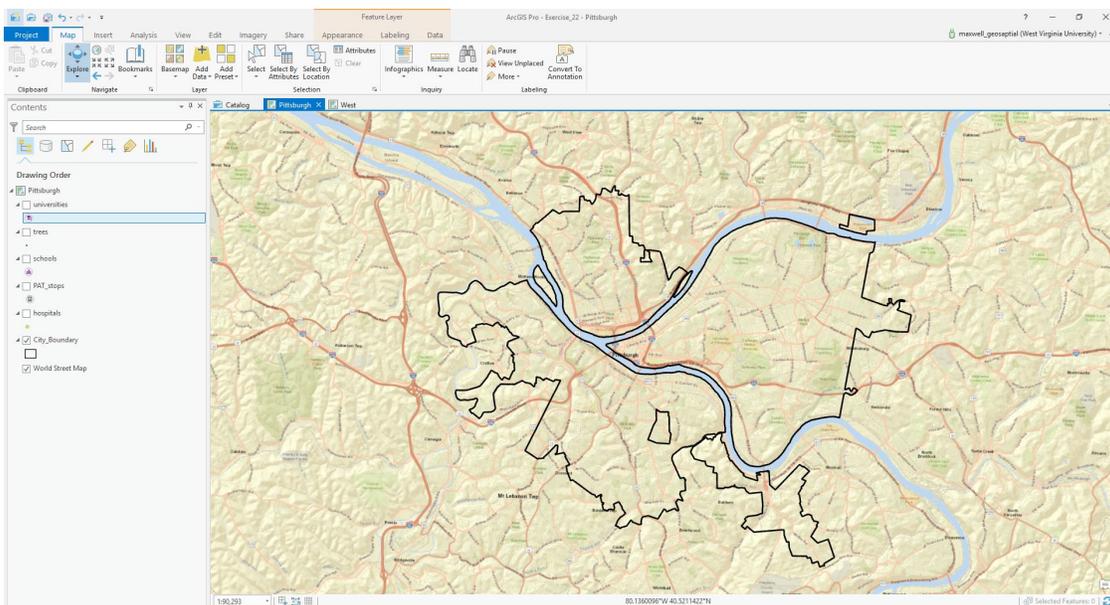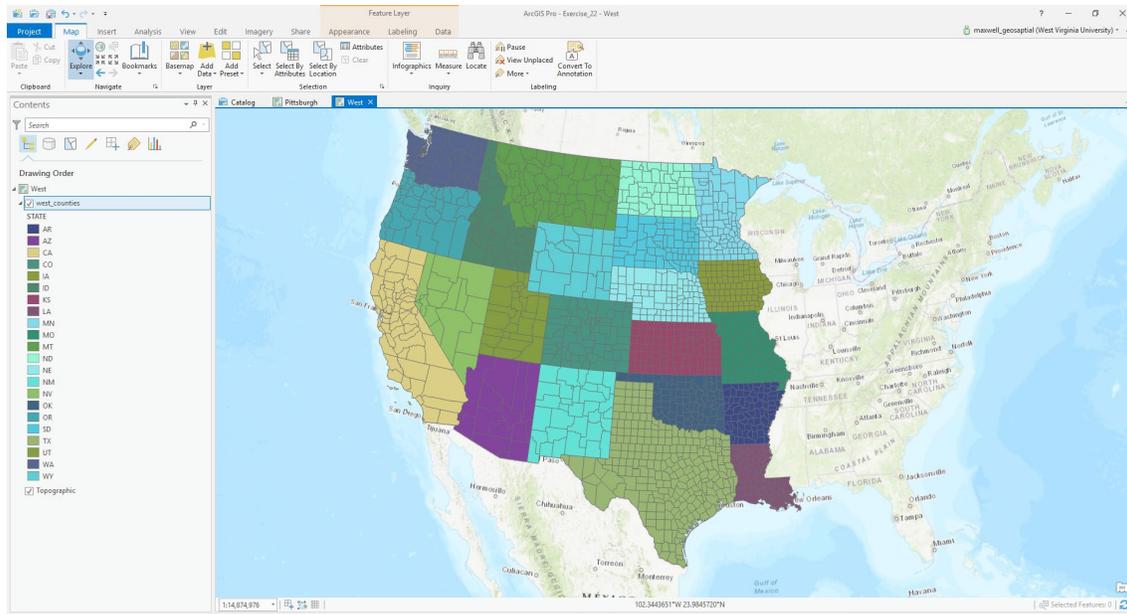
You will start the analysis with the Pittsburgh data.

☐ Navigate to the Pittsburgh map.

Step 2. Set Environment Settings

Before you start performing analyses, you will set some environment settings.

☐ Click on the Environments button in the Geoprocessing area of the Analysis Tab.

☐ Under Workspace, set the Current Workspace and Scratch Workspace to the **Exercise_28.gdb** geodatabase in the **Exercise_28** folder.

☐ You do not need to change any other settings.

☐ Click OK to save the Environment changes.
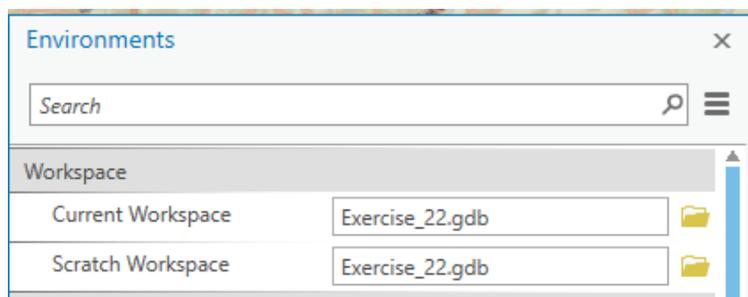
Step 3. Measuring Spatial Central Tendency

You will begin your analysis by finding the most centrally located university in Pittsburgh. This can be accomplished using the **Central Feature Tool**.

3

Using this tool, the most central feature is defined as the feature with the smallest accumulated distance to all other features in the dataset.

Note: In this lab, we will access tools from ArcToolbox. However, there are many ways to access tools in ArcGIS Pro. For example, some of the more common tools are provided in the Tools list in the Analysis Tab.

Once you open the Geoprocessing Pane, you can access favorite tools or search for Tools.

We have decided to demonstrate ArcToolbox here so that you get a sense of where the tools are located in the Toolbox directory.

- ☐ In the Analysis Tab, select Tools from the Geoprocessing Area. This should open the Geoprocessing Pane.
- ☐ In the Geoprocessing Pane, navigate to the Toolboxes.
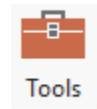
Note: We will not provide these directions for accessing other tools. We will just tell you where to find them within ArcToolbox.

- ☐ Navigate to Spatial Statistic Tools followed by the Measuring Geographic Distributions subtoolbox.
- ☐ Click on the **Central Feature Tool**.
- ☐ Set the Input Feature Class to the **universities** layer.
- ☐ Name the Output Feature Class **central_university**. Note that the output will automatically save to the location specified in the environment settings: **Exercise_28.gdb**.
- ☐ You do not need to change any of the other settings.
- ☐ Click Run to execute the tool.
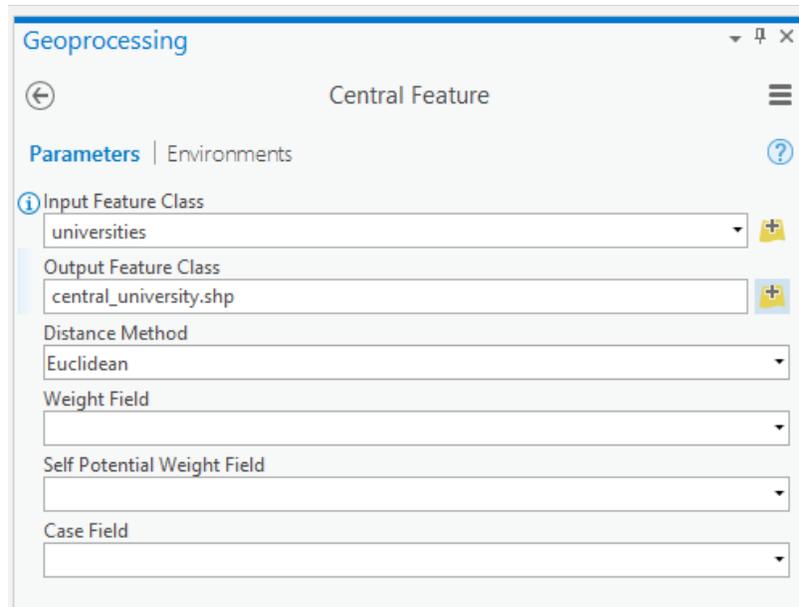
This analysis will produce a new vector layer containing one point: the university that was calculated to be most central. Use the attribute table of your output to answer the following question. Note that the name of the university is in the "NAME" field.

Question 1. What is the most central university in Pittsburgh? (2 Points)

☐ Conduct a similar analysis for the hospitals. Find the most central hospital. Note that the hospital name is in the "name" field.

Question 2. What is the most central hospital in Pittsburgh? (2 Points)

Instead of finding the most central feature, you can also find the spatial mean or median of the data. This is defined as the geographic center of the data. This is calculated from the average X and Y coordinates of the input features. If the input features are polygons, the average X and Y values of the centroids are used. This geographic center can be defined as a mean or as a median. You will first calculate the mean using the **Mean Center Tool**.

☐ Navigate to Spatial Statistic Tools followed by the Measuring Geographic Distributions subtoolbox.
☐ Click on the **Mean Center Tool**.
☐ Set the Input Feature Class to the **schools** layer.
☐ Name the Output Feature Class **mean_center_schools**.
☐ You do not need to change any of the other settings.
☐ Click Run to execute the tool.
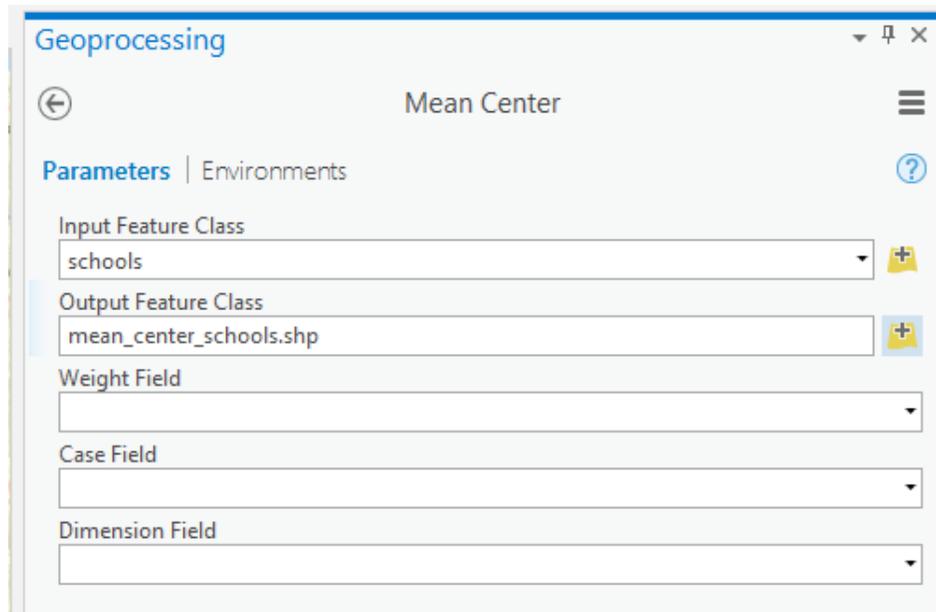
Now, you will calculate the median center using the **Median Center Tool**.

- ☐ Navigate to Spatial Statistic Tools followed by the Measuring Geographic Distributions subtoolbox.
- ☐ Click on the **Median Center Tool**.
- ☐ Set the Input Feature Class to the **schools** layer.
- ☐ Name the Output Feature Class **median_center_schools**.
- ☐ You do not need to change any of the other settings.
- ☐ Click Run to execute the tool.
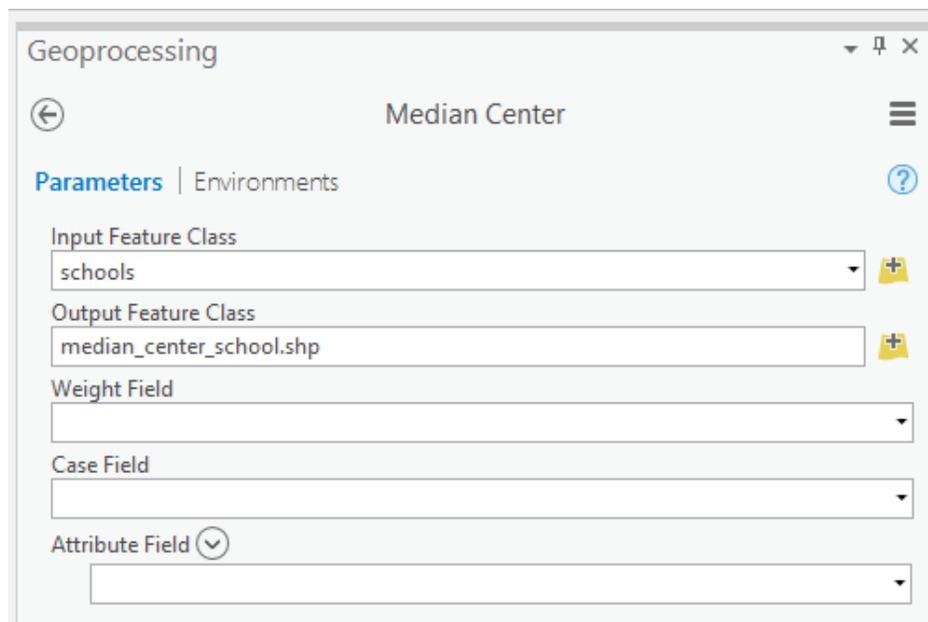
You will notice that the mean and median center are not at the same location. Also, they are not at the location of one of the schools, in contrast to the central feature.

☐ Use the measuring tool to measure the distance between the mean and median centers in meters.
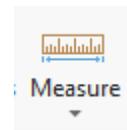
Question 3. How far apart are the mean and median centers for the hospitals in meters? (2 Points)
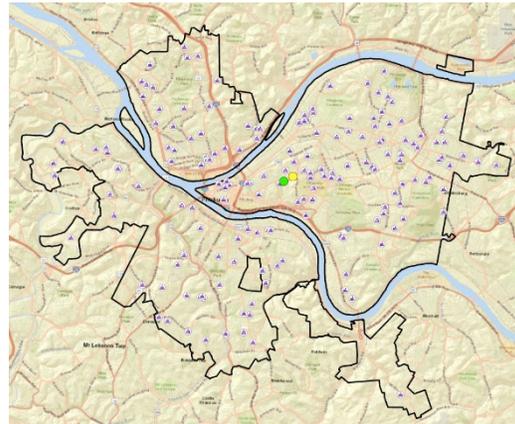
It is also possible to apply a weight to the calculations. As an example, we will find the mean center for the schools using the number of students as a weight. So, schools with more students will have a higher weight in the calculation of the mean. These numbers are stored in the "students" field. Note that these are synthetic data and are not the real school population sizes. The results will not represent a true measure. Note that these coordinates are relative to the map projection: NAD 1983 State Plane Pennsylvania South Feet.

☐ Navigate to Spatial Statistic Tools followed by the Measuring Geographic Distributions subtoolbox.
☐ Click on the **Mean Center Tool**.
☐ Set the Input Feature Class to the **schools** layer.
☐ Name the Output Feature Class **mean_center_schools_weighted**.
☐ Set the Weight Field to the "students" field.
☐ You do not need to change any of the other settings.
☐ Click Run to execute the tool.
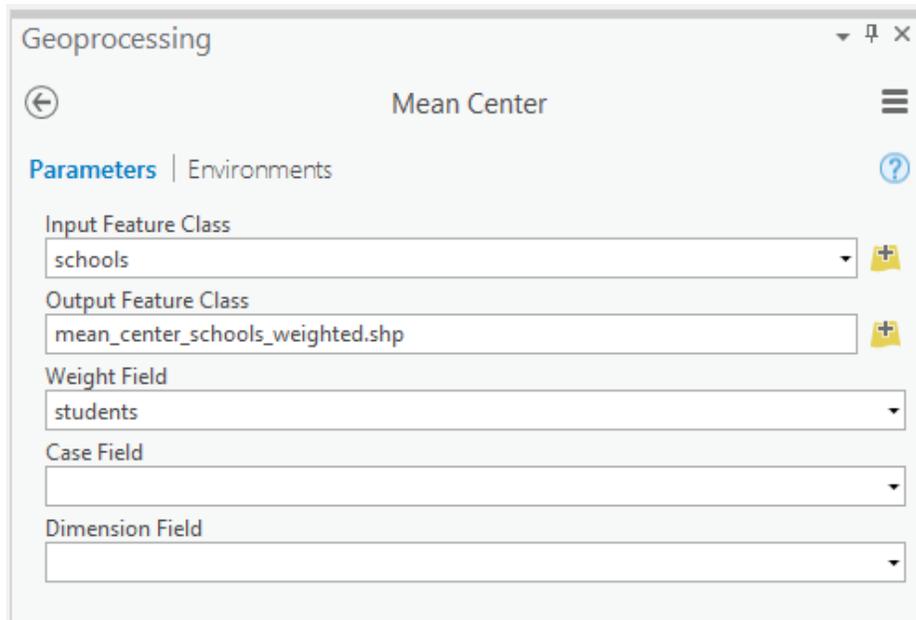
Question 4. How far apart are the mean and weighted mean center locations in meters? (2 Points)

☐ Now, calculate the unweighted mean center for the hospitals layer. You can answer the questions below using the attribute table of the results.

Question 5. What is the X coordinate of this location? (2 Points)

Question 6. What is the Y coordinate of this location? (2 Points)

Step 4. Measures of Spatial Variance and Directional Trends

The measures investigated above provide estimates of spatial central tendency. You will now investigate measures of spatial variability or dispersion. First, you will calculate a standard distance for the **hospitals** layer using the **Standard Distance Tool**. The standard distance is a measure of spatial standard deviation. The larger the circle, the more spatially dispersed the data. The center of the circle will be the mean center.

☐ Navigate to Spatial Statistic Tools followed by the Measuring Geographic Distributions subtoolbox.
☐ Click on the **Standard Distance Tool**.
☐ Set the Input Feature Class to the **hospitals** layer.
☐ Name the Output Standard Distance Feature Class **hospitals_sd**.
☐ Set the Circle Size to "1 standard deviation."
☐ You do not need to change any of the other settings.
☐ Click Run to execute the tool.
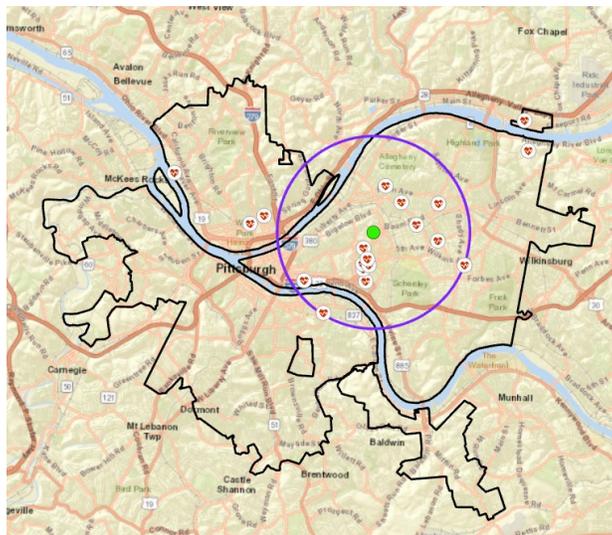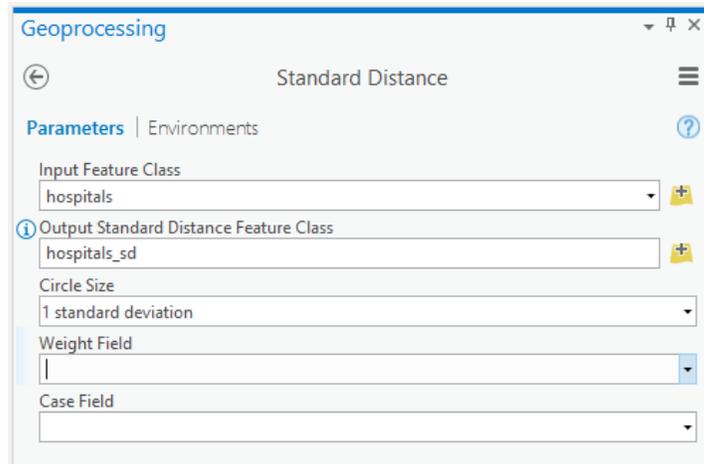
The spatial standard deviation (in the map units, in this case feet) is noted in the attribute table of the output.

Question 7. What is the spatial standard deviation value for the hospitals? (2 Points)

☐ Complete a similar analysis to find the spatial standard deviation for the universities.

Question 8. What is the spatial standard deviation value for the universities? (2 Points)

You will now experiment with the **Directional Distribution (Standard Deviational Ellipse) Tool**. This tool is used to describe the spatial central tendency, spatial dispersion, and directional trends. The created ellipse will be centered on the mean center. The larger the ellipse, the more dispersed the data. The long access of the ellipse will be aligned with the direction in

which the data are most dispersed. You will now calculate this for the **universities** layer.

- ☐ Navigate to Spatial Statistic Tools followed by the Measuring Geographic Distributions subtoolbox.
- ☐ Click on the **Directional Distribution (Standard Deviational Ellipse) Tool**.
- ☐ Set the Input Feature Class to the **universities** layer.
- ☐ Name the Output Ellipse Feature Class **universities_dd**.
- ☐ Set the Ellipse Size to "1 standard deviation."
- ☐ You do not need to change any of the other settings.
- ☐ Click Run to execute the tool.
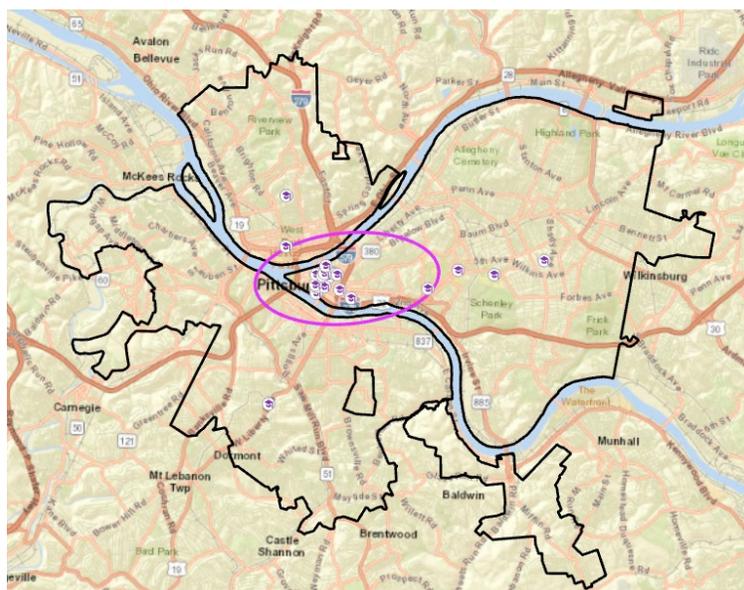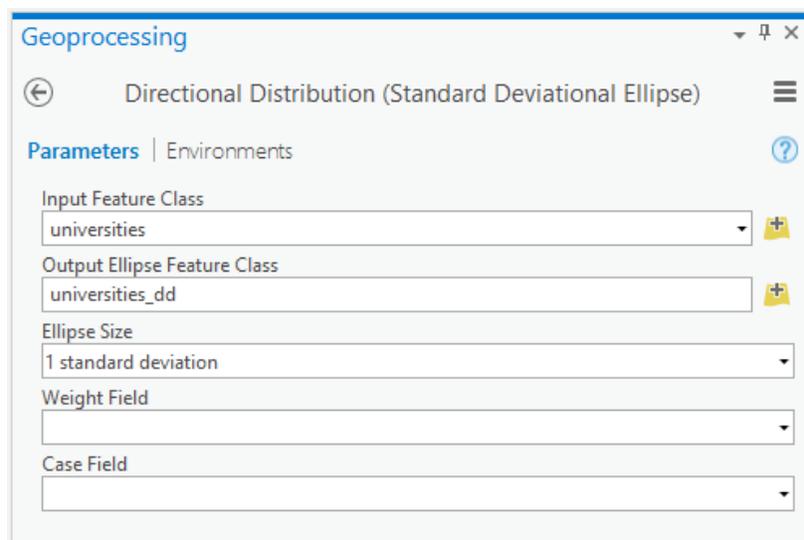
Use the result to answer the following question.

Question 9. Describe the directional trends in the university data. (4 Points)

☐ Repeat this analysis for the schools layer with the "students" field applied as a weight.

Question 10. Describe the directional trends in the schools data. Also, compare the universities and schools results. (4 Points)

Step 5. Analyzing Point Patterns

You will now analyze spatial patterns using the **Average Nearest Neighbor Tool**. This tool provides a statistical test of dispersion, randomness, or clustering based on the average distance from each feature to its nearest neighbors. It does not look at the attribute data associated with the points. It only looks at the point distribution. It is a statistical test that produces a z-score and a *p*-value. You will run this tool on the **trees** layer to determine if the spatial pattern of the trees are dispersed, random, or clustered.

☐ Navigate to Spatial Statistic Tools followed by the Analyzing Patterns subtoolbox.
☐ Click on the **Average Nearest Neighbor Tool**.
☐ Set the Input Feature Class to the **trees** layer.
☐ Leave the Distance Method set to Euclidean.
☐ Make sure Generate Report is selected.
☐ You do not need to change any of the other settings.
☐ Click Run to execute the tool.
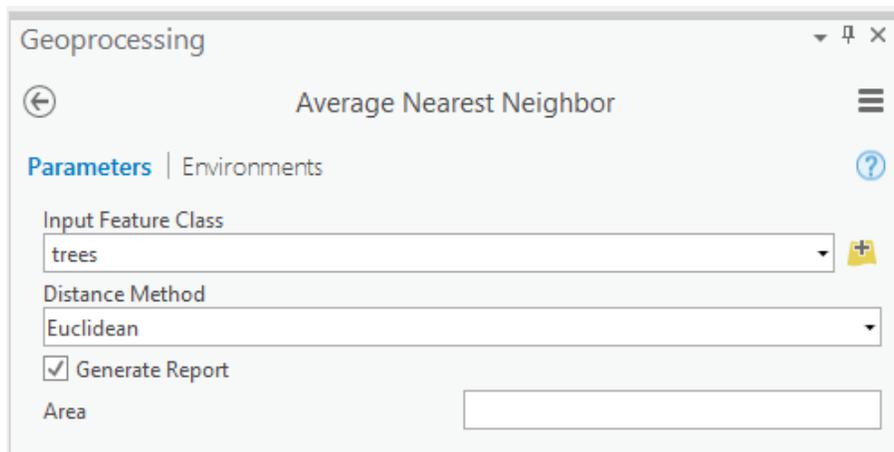
This tool does not produce a new layer. It generates a report.

☐ In the result pop-up click on the Report File. This will open in a web browser as an HTML report.
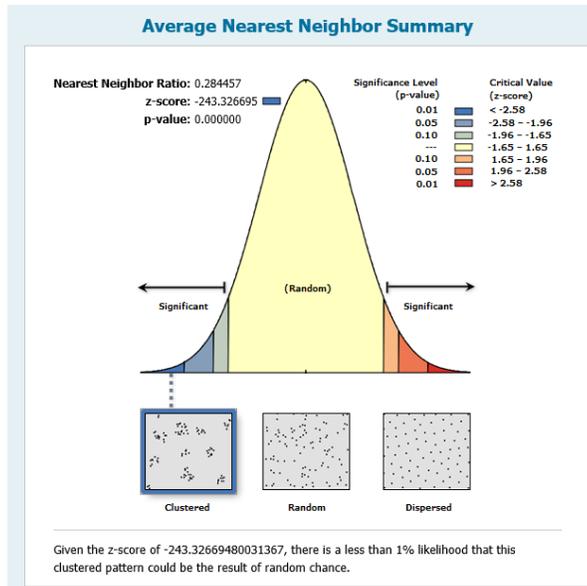
Average Nearest Neighbor Summary

Observed Mean Distance:	41.6558 US_Feet
Expected Mean Distance:	146.4400 US_Feet
Nearest Neighbor Ratio:	0.284457
z-score:	-243.326695
p-value:	0.000000

The *p*-value for this test was 0, suggesting statistical significance. Note that this result suggests that the trees are clustered as opposed to spatially random or dispersed.

Question 11. What is the null hypothesis for this test? (4 Points)

☐ Complete a similar analysis using the **PAT_stops** layer. Use Euclidean for the Distance Method.

Question 12. What is the nearest neighbor ratio? (2 Points)

Question 13. What is the z-value? (2 Points)

Question 14. What is the *p*-value? (2 Points)

Question 15. According to this analysis, are the stops dispersed, random, or clustered? (2 Points)

Step 6. Analyzing Global Spatial Autocorrelation

The average nearest neighbor statistic looks at spatial patterns, not the attributes associated with the features. Exploring the correlation of feature attributes across space is known as spatial autocorrelation. The question here is whether or not the attributes or values associated with features are spatially clustered (near features tend to have more similar values), spatially dispersed (near features tend to have more dissimilar values or distant features tend to have more similar values), or spatially random (there is no pattern in the distribution of the values). The Global Moran's I Test can be used to assess global spatial autocorrelation.

☐ You are done using the Pittsburgh map. Switch to the West map.

You will explore whether median income aggregated at the county level in the western United States shows spatial randomness, dispersion, or clustering. You will do this using the **Spatial Autocorrelation (Global Moran's I) Tool**.

☐ Navigate to Spatial Statistic Tools followed by the Analyzing Patterns subtoolbox.
☐ Click on the **Spatial Autocorrelation (Global Moran's I) Tool**.
☐ Set the Input Feature Class to the **west_counties** layer.
☐ Set the Input Field to "med_income."
☐ Make sure to select Generate Report.
☐ Set the Conceptualization of Spatial Relationships to "Contiguity edges only."
☐ Leave the Standardization as None.
☐ Click Run to execute the tool.
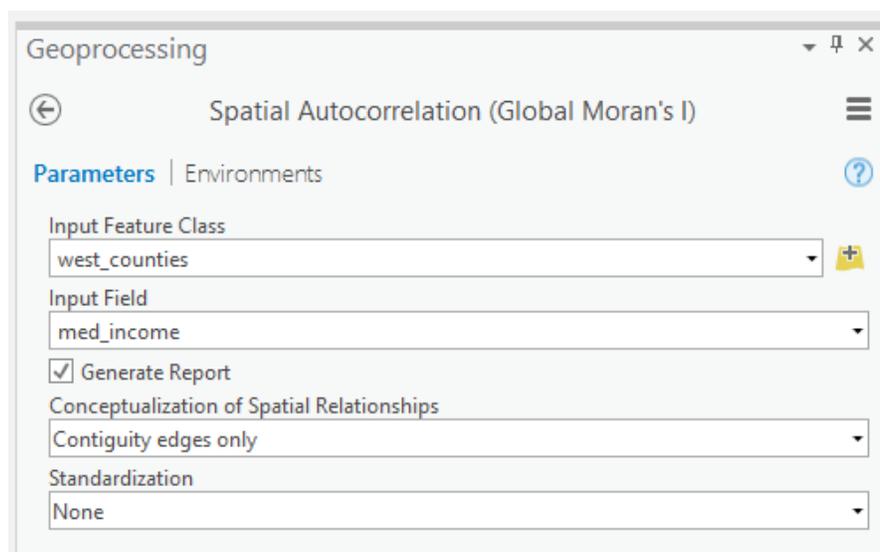
Similar to the Average Nearest Neighbor Tool, a new output is not generated. Instead, an HTML report is produced.

- ☐ Open the HTML report.

Spatial Autocorrelation Report
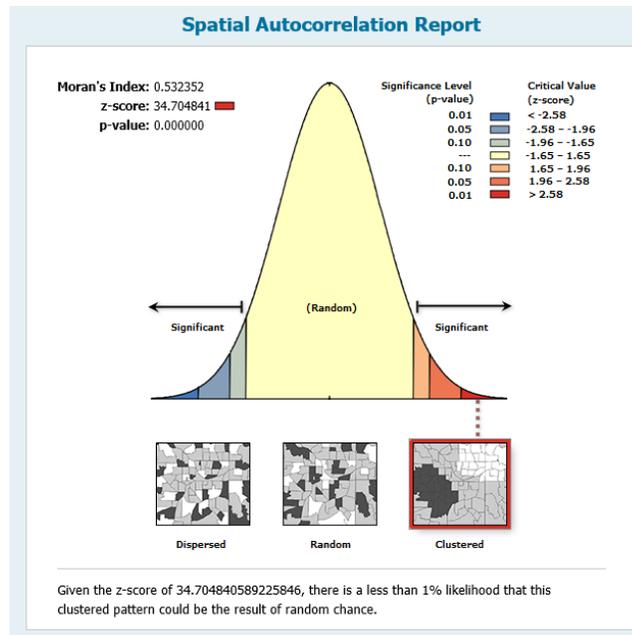
Global Moran's I Summary	
Moran's Index:	0.532352
Expected Index:	-0.000666
Variance:	0.000236
z-score:	34.704841
p-value:	0.000000

Note that this test suggests that median income shows a clustered pattern when evaluated globally. A positive Moran's I index suggests clustering whereas a negative result suggests dispersion. Values near zero indicate randomness. So, this test suggests that near counties tend to have more similar income than more distant counties.

Question 16. What is the null hypothesis for this test? (4 Points)

This test is impacted by how spatial relationships are defined. Use the tool help to answer the following questions.

Question 17. Describe the contiguity edges only method for defining spatial relationships. (4 Points)

Question 18. Describe the inverse distance method for defining spatial relationships. (4 Points)

14

☐ Complete a similar analysis to investigate global spatial autocorrelation at the county level for the percentage of the population that is older than 55. This information is provided in the "per_gt55" field. Make sure to use the "Contiguity edges only" option for the conceptualization of spatial relationships. Also remember to generate the report. Use the results to answer the following questions.

Question 19. What was the reported Moran's I Index? (2 Points)

Question 20. Does this suggest dispersion, clustering, or randomness? (2 Points)

Step 7. Analyzing Local Spatial Autocorrelation

Lastly, you will analyze local spatial autocorrelation, as opposed to assessing autocorrelation globally as investigated above. Here, you will specifically investigate hot spot analysis. The goal of this tool is to identify local clusters of high values (hot spots) and local clusters of low values (cold spots) using the Getis-Ord GI* statistic. This analysis can be completed using the **Hot Spot Analysis (Getis-Ord Gi*) Tool**. You will start by performing this analysis for median income.

☐ Navigate to Spatial Statistic Tools followed by the Mapping Clusters subtoolbox.
☐ Click on the **Hot Spot Analysis (Getis-Ord Gi*) Tool**.
☐ Set the Input Feature Class to the **west_counties** layer.
☐ Set the Input Field to "med_income."
☐ Name the Output Feature Class **med_income_hot_spots**.
☐ Set the Conceptualization of Spatial Relationships to "Contiguity edges only."
☐ You do not need to change any additional settings.
☐ Click Run to execute the tool.
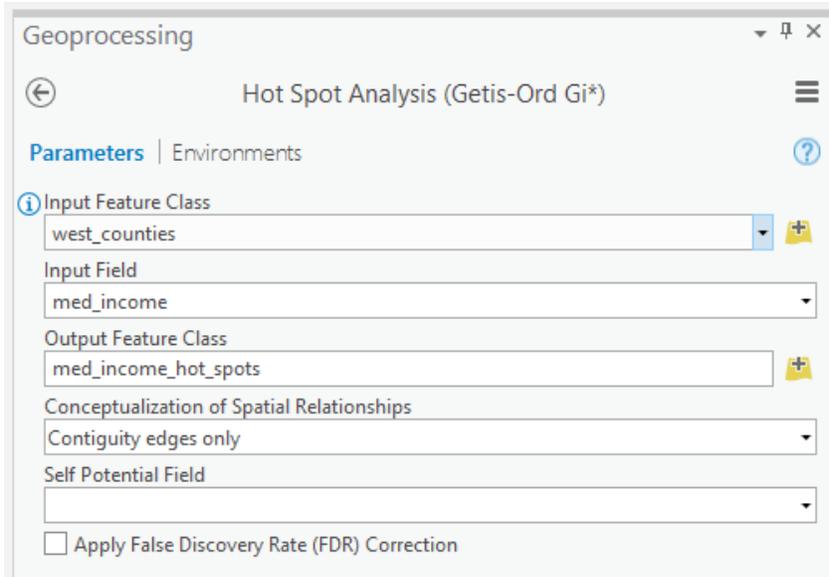
In contrast to the global statistics, a new spatial output will be generated and added to your map. The output is shown below.
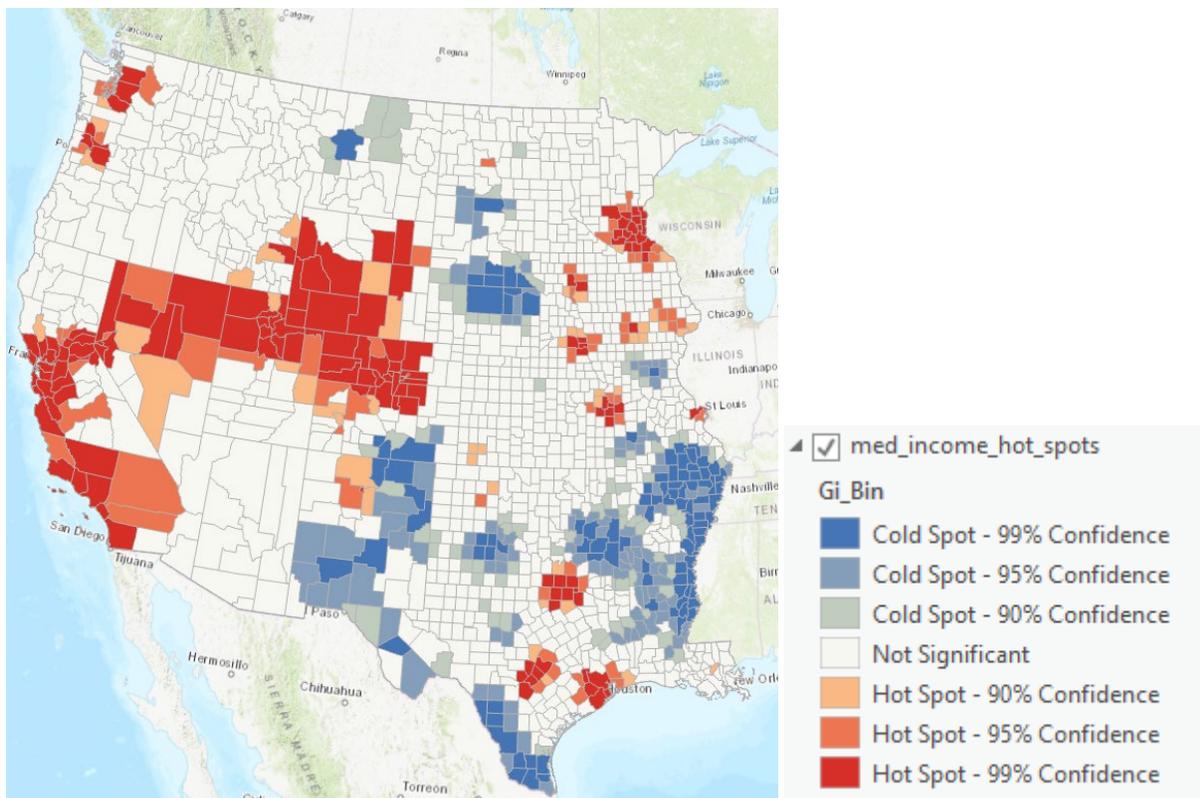
Areas in red indicate hot spots, or clusters of counties with a high median income. Blue areas indicate cold spots, or clusters of counties with low median incomes. White areas suggest randomness or patterns that are not statistically significant. Darker shades or red or blue indicate increasing

statistical significance at the 90%, 95%, and 99% confidence intervals. Note that no global measure is produced since this is not a global statistical test. If you open the attribute table, you will see that different *p*-values were generated for each county. Again, this is a local test.

OBJECTID	Shape	SOURCE_ID	med_income	Shape_Length	Shape_Area	GiZScore Contiguity	GiPValue Contiguity	NNeighbors Contiguity	Gi_Bin Contiguity
1	Polygon	1	31728	1015424.824093	29055887659.976841	-1.578685	0.114408	9	0
2	Polygon	2	44000	510184.566395	16105034948.7941	-0.135433	0.892269	6	0
3	Polygon	3	47933	1238796.603981	48331699096.773514	-0.527593	0.597782	7	0
4	Polygon	4	38405	820821.824126	12422439946.458048	0.413442	0.679283	7	0
5	Polygon	5	40902	570246.381247	12047560238.565525	-0.119138	0.905166	8	0
6	Polygon	6	53654	393007.364413	4759059881.163303	-1.323317	0.18573	7	0
7	Polygon	7	32973	588811.661134	11689012929.469244	0.369999	0.711383	8	0
8	Polygon	8	56511	811121.040806	23891620302.433167	0.050847	0.959448	7	0
9	Polygon	9	38641	1070524.221019	34865483707.600502	0.726661	0.467434	9	0
10	Polygon	10	39416	966438.438959	25795010890.214333	-0.98398	0.325125	6	0
11	Polygon	11	46653	716322.297603	23801036762.361328	0.418789	0.67537	7	0

To finish the lab, you will conduct a hot spot analysis to look for local hot spots and cold spots of counties based on the percentage of the population that is older than 55. You will then generate a map of your output and describe the results.

☐   Use the **Hot Spot Analysis (Getis-Ord Gi*) Tool** to investigate local hot spots and cold spots relative to the percentage of the population in each county that is older than 55. Reminder that this information is in the "per_gt55" field. Make sure to use the "Contiguity edges only" method as the conceptualization of spatial relationships.

Deliverable 1 (10 Points)

Make a map of the output. Submit it as a PDF. It will be graded relative to the following rubric.

1.   Descriptive title provided (2 Points)
2.   Scale bar and north arrow provided (2 Points)
3.   Legend is provided using the default legend from the analysis (2 Points)
4.   A credit is provided for the 2010 US Census (1 Point)
5.   Your name is provided (1 Point)
6.   The map is overall neatly organized and uses space well (2 Points)

Deliverable 2 (10 Points)

Write a paragraph that provides a brief description of the spatial patterns observed.

As a final note, this is simply an introduction to spatial statistics, and there are many methods that were not discussed here. This is a field of study in its own right. If you are interested in this, we would recommend taking a statistics course and/or a spatial statistics course. Other than just exploring global and local patterns, spatial statistical methods, such as geographically weighted regression (GWR), can be used to make prediction over a map space.

END OF EXERCISE