Exercise 16: Data Summarization

We have found that data summarization is one of the most common tasks we perform inside of our GIS software. Data summarization is often required before you can answer a question, perform statistical analysis, create a model, or write up a report. Here, you will explore data summarization using ArcGIS Pro.

Topics covered in this exercise include:

1. How to count the number of points occurring in a polygon
2. How to summarize the length of lines within a polygon
3. How to find the area of each class in a categorical raster grid within polygons
4. How to calculate summary statistics for a continuous raster grid within polygons
5. How to calculate summary statistics for a continuous raster grid relative to line features
6. How to extract raster values at point locations

Step 1. Open a Map Project

First, we need to download and open the **Exercise_16.aprx** file.

☐ Download the **Exercise_16** 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_16** folder where it was saved on your local machine.

☐ Select **Exercise_16.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_16.aprx** file within the uncompressed folder directly to launch ArcGIS Pro.

This project contains a single map that contains five data layers.

Discharge_Points: hypothetical pollution discharge points

streams: line features presenting streams

**watersheds**: polygon watershed boundaries

nlcd_2011: categorical land cover raster grid from the 2011 National Land Cover Database (NLCD) as a 30 x 30 meter grid

elevation: digital elevation model (DEM) for the study area extent in meters as a 30 x 30 meter grid
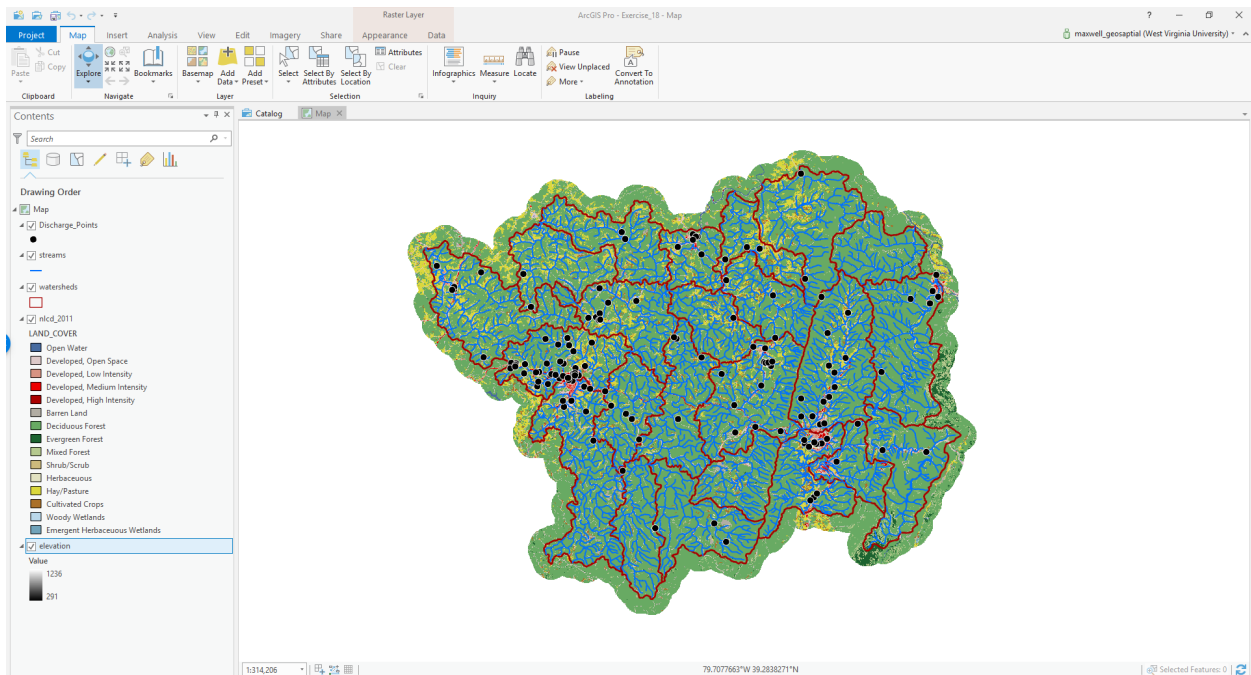
Step 2. Set Environment Settings

☐ Navigate to the Analysis Tab then select Environments in the Geopocessing Area.

☐ Change the Current Workspace and the Scratch Workspace (if needed) to the **Exercise_16.gdb** geodatabase in the **Exercise_16** folder in your downloaded copy of the data. This will cause all of the outputs to automatically be saved to this location.

☐ Click Okay to accept these settings.
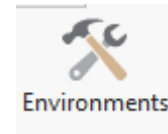
Step 3. Count Points in Polygons

You will start by counting the number of discharge points in each watershed. This can be accomplished using the **Summarize Within Tool**.

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.
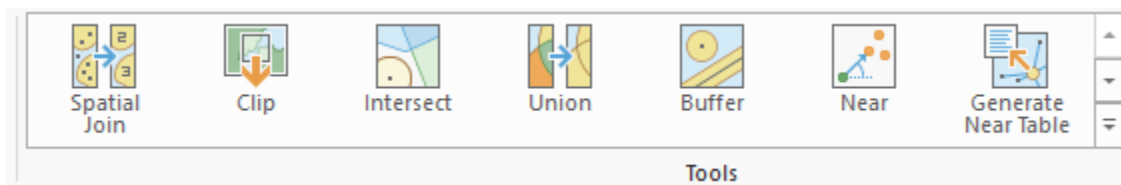
Once you open the Geoprocessing Pane, you can access your favorite tools or search for Tools.
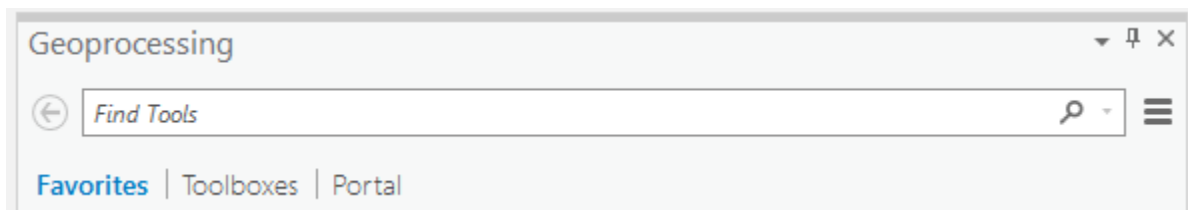
We have decided to demonstrate ArcToolbox here so that you get a sense of where the tools are located in the Toolbox directory. You can also create your own tools and toolboxes, which will be investigated in later labs.

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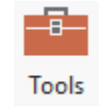
Note: We will not provide these directions for accessing other tools. We will
just tell you where to find them within ArcToolbox.

☐ Navigate to Analysis Tools followed by the Statistics subtoolbox.
☐ Open the **Summarize Within Tool**.
☐ Set the Input Polygons to the **watersheds** layer.
☐ Set the Input Summary Features to the **Discharge_Points** layer.
☐ Name the Output Feature Class **watershed_cnt_discharge**. Note
  that this will automatically save to the **Execise_16** directory, as
  specified in the environments.
☐ You do not need to change any of the other settings.
☐ Click Run to execute the tool. The output should automatically be
  added to your map.
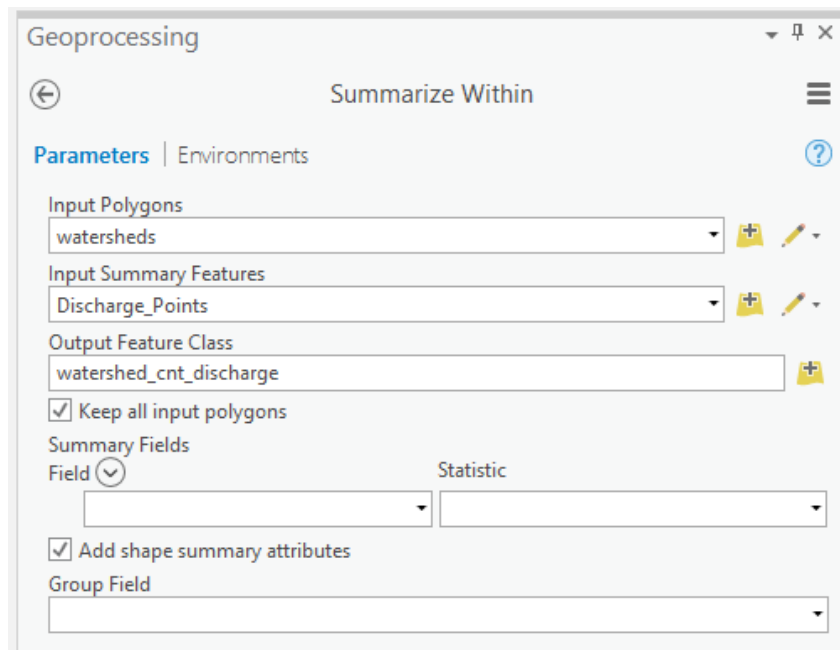
Use the attribute table for the output to answer the following questions. Note
that a field named "Count of Points" should have been added to the table as
the last attribute column. This is the number of discharge points counted in
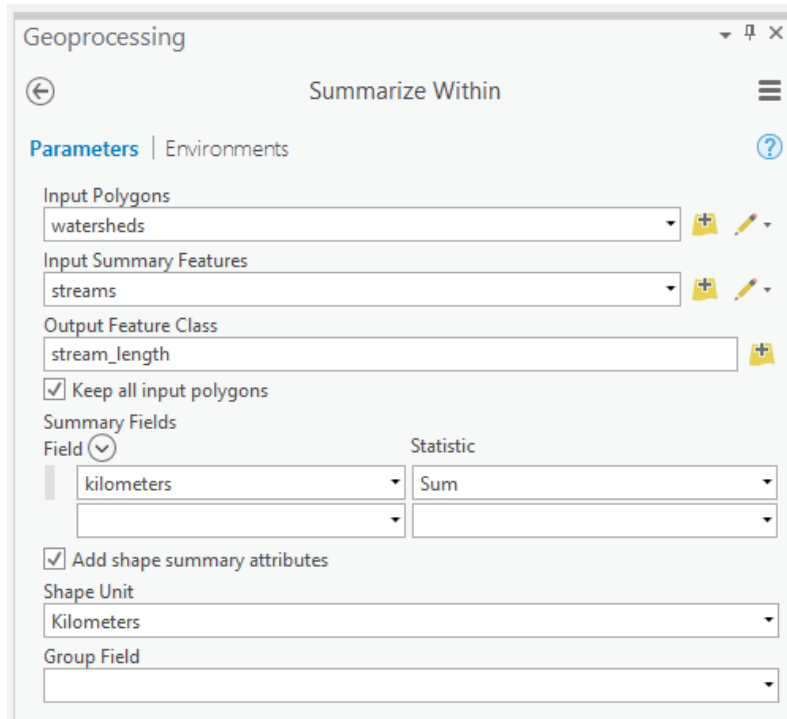that watershed.

Question 1. What is the name of the watershed with the largest number of discharge points (the name is provided in the "HU_12_NAME" field)?
(2 Points)

Question 2. Did all the watersheds have at least one mapped discharge point? (2 Points)

Step 4. Sum Length of Lines in Polygons

You will now sum the length of streams in each watershed. This can also be done using the **Summarize Within Tool**. Note that the **streams** layer contains an attribute column named "kilometers." This is the length in kilometers. It is important to make sure geometric fields are correct before using them in calculations, as geometry fields can be incorrect as a result of geoprocessing operations. It has been confirmed that this field is correct.

☐ Navigate to Analysis Tools followed by the Statistics subtoolbox.
☐ Open the **Summarize Within Tool**.
☐ Set the Input Polygons to the **watersheds** layer.
☐ Set the Input Summary Features to the **streams** layer.
☐ Name the Output Feature Class **stream_length**.
☐ Under Summary Fields, add the "kilometers" field and set the Statistic to "Sum."
☐ You do not need to change any additional settings.
☐ Click Run to execute the tool. The output should automatically be added to your map.

Use the attribute table for the output to answer the following questions. Note that a field named "Sum kilometers" should have been added to the table as the last attribute column. This is the length of streams by watershed.

Question 3. What watershed has the longest length of streams in it (the name is provided in the "HU_12_NAME" field)? (2 Points)

Question 4. What is the average length of streams per watershed in kilometers? (2 Points)

Hint: You can obtain the average by using the **Summary Statistics Tool**.

Step 5. Calculate Area of Categorical Raster in Polygons

You will now calculate the area of each class in a categorical raster by polygon. This can be accomplished using the **Tabulate Area Tool**. However, before you start, let's reclassify the land cover grid into fewer classes using the table below.

Old Value	Old Class	New Value	New Class
11	Open Water	10	Open Water
21	Developed, Open Space	20	Developed
22	Developed, Low Intensity	20	Developed
23	Developed, Medium Intensity	20	Developed
24	Developed, High Intensity	20	Developed
31	Barren Land	30	Barren Land
41	Deciduous Forest	40	Forest
42	Coniferous Forest	40	Forest
43	Mixed Forest	40	Forest
52	Shrub/Scrub	70	Other Vegetation
71	Grassland/Herbaceous	70	Other Vegetation
81	Pasture/Hay	70	Other Vegetation
82	Cultivated Crop	70	Other Vegetation
90	Woody Wetland	90	Wetland
95	Emergent Herbaceous Wetland	90	Wetland

☐ Navigate to Spatial Analyst Tools followed by the Reclass subtoolbox.
☐ Open the **Reclassify Tool**.
☐ Set the Input Raster to the **nlcd_2011** layer.
☐ Make sure the Reclass Field is set to "LAND_COVER."
☐ Use the values specified above to reclassify the raster grid.
☐ Name the output **lc_general**.
☐ Click Run to execute the tool. The output should automatically be added to your map.
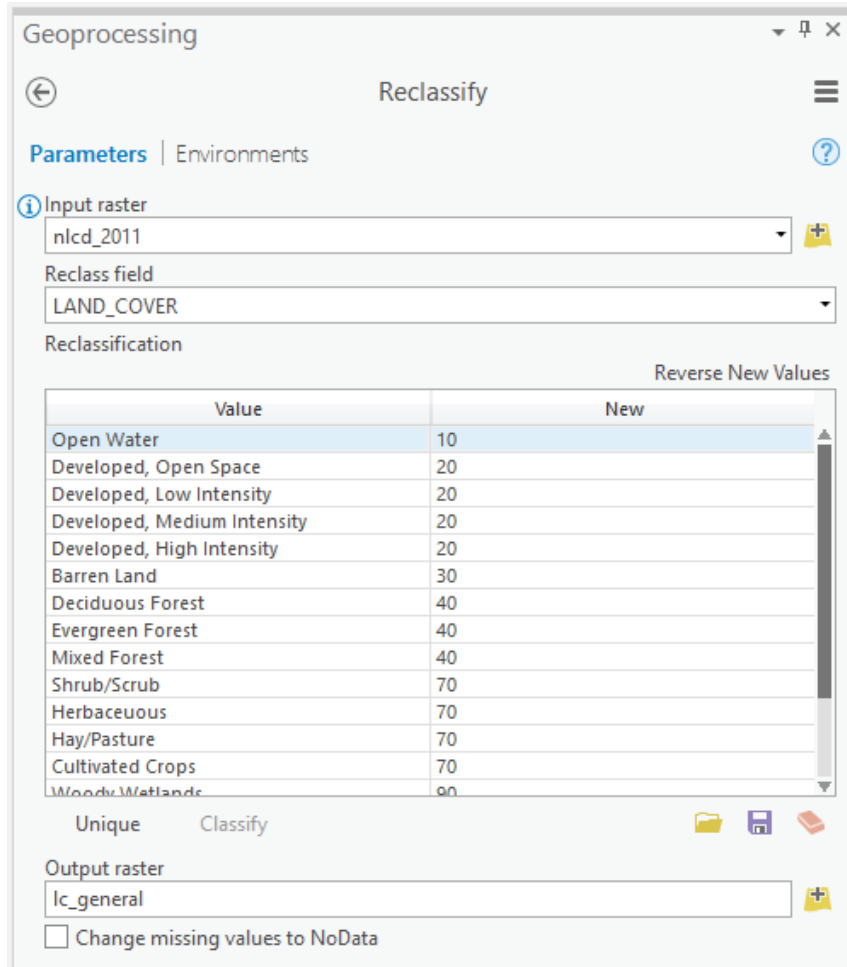
You are now ready to summarize the categorical land cover raster data relative to the watershed boundaries using the **Tabulate Area Tool**.

- ☐ Navigate to Spatial Analyst Tools followed by the Zonal subtoolbox.
- ☐ Open the **Tabulate Area Tool**.
- ☐ Set the Input Raster or Feature Zone Data to the **watersheds** layer.
- ☐ Set the Zone Field to the "HU_12_NAME" field.
- ☐ Set the Input Raster or Feature Class Data to the **lc_general** layer.
- ☐ Make sure the Class Field is set to "VALUE."
- ☐ Name the Output Table **lc_table**.
- ☐ Make sure the Processing Cell Size is set to 30 (this should happen by default).
- ☐ Click Run to execute the tool. The output should automatically be added to your map.
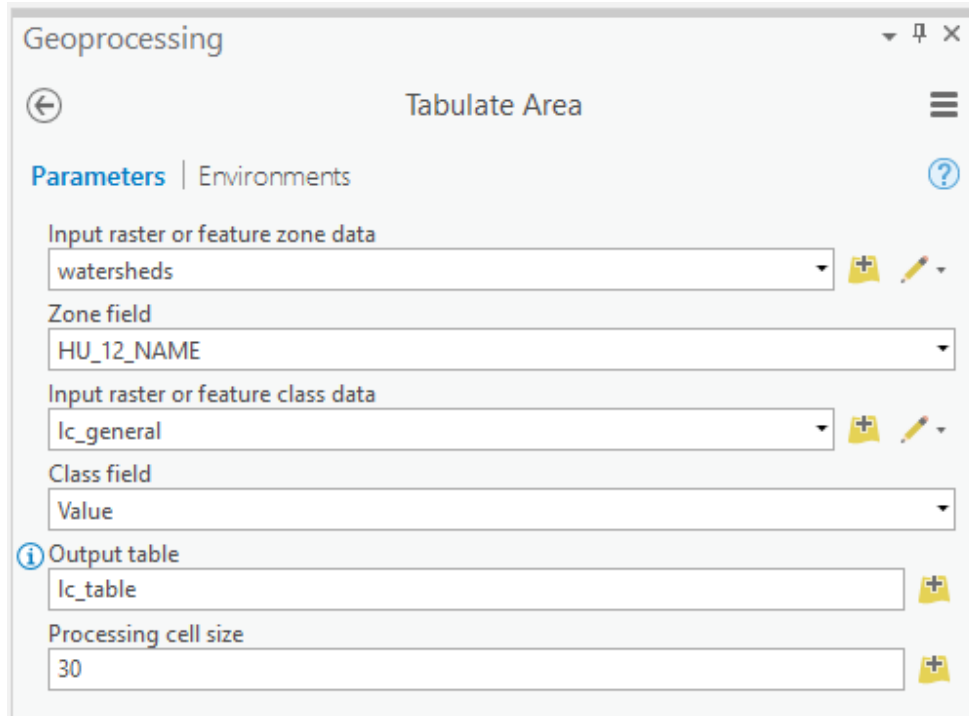
The output table provides the area of each land cover type in square meters. The areas are reported in square meters since the UTM projection uses meters as the length units. Let's now calculate the percent forest by watershed using table calculations.

☐ Open the **lc_table** attribute table.
☐ Add a new field. Name it "per_forest" and set the Data Type to Double.
☐ Save the field.
☐ In the field, calculate the percent forest as follows:

((!VALUE_40!) / (!VALUE_10! + !VALUE_20! + !VALUE_30! + !VALUE_40! + !VALUE_70! + !VALUE_90!)) * 100

Use your results to answer the following questions.

Question 5. Which watershed had the largest land area as forest? (2 Points)

Question 6. What watershed had the largest percentage of its land area as forest? (2 Points)

Question 7. Which watershed had the largest land area as developed? (2 Points)

It is possible to join the table to the watershed boundaries using a table join. Note that table joins are temporary. So, if you want to permanently include the new attribute columns, you will need to export the data layer.

☐ Right-click on the **watersheds** layer in the Contents Pane. Select Joins and Relates followed by Add Join. This will open the **Add Join Tool**.
☐ Make sure the **watersheds** layer is set as the Layer Name or Table View.
☐ Make sure the Input Join Field is set to the "HU_12_NAME."
☐ Set the Join Table to the **lc_table** layer.
☐ Make sure the Output Join Field is set to "HU_12_NAME."
☐ Click Run to execute the tool.
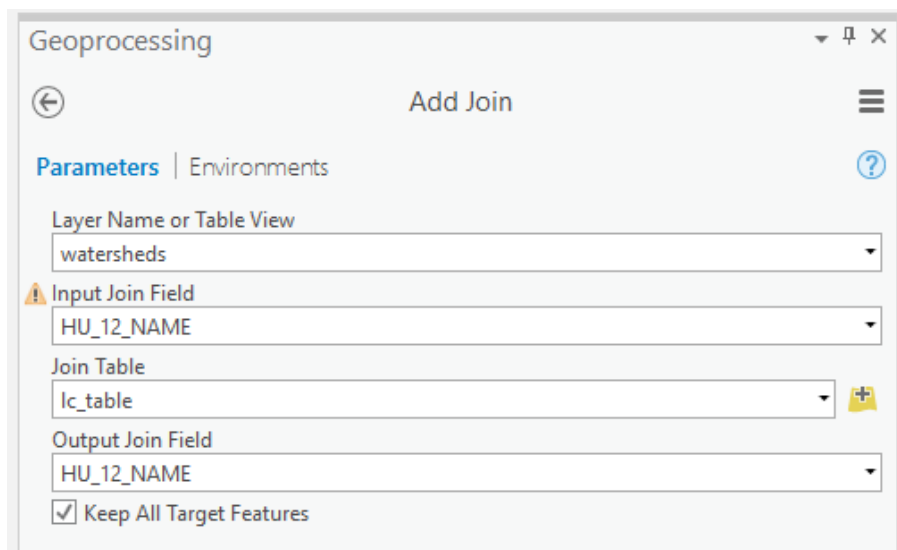
If you open the attribute table for the watersheds layer you can see that the new attributes were added. Again, this is a temporary join, so you will have to export the layer to permanently attach these fields.

You do not need the join for the remainder of the exercise, so you will now remove it.

☐ Right-click on the watershed layer in the Contents Pane. Select Joins and Relates followed by Remove All Joins.
☐ Click Yes on the Remove Joins message.

Step 6. Calculate Summary Statistics for a Continuous Raster in Polygons

The **Tabulate Area Tool** is used to summarize categorical raster data. If you want to summarize continuous raster data as a table, use the **Zonal Statistics as Table Tool**. Instead of generating areas for each class, this tool calculates summary statistics for the continuous data within each polygon. Here, you will calculate summary statistics for elevation data within each watershed.

☐ Navigate to Spatial Analyst Tools followed by the Zonal subtoolbox.
☐ Open the **Zonal Statistic as Table Tool**.
☐ Set the Input Raster or Feature Zone Data to the **watersheds** layer.
☐ Set the Zone Field to the "HU_12_NAME" field.
☐ Set the Input Value Raster to the **elevation** layer.
☐ Name the Output Table **ele_stats**.
☐ Calculate "All" statistics.
☐ Click Run to execute the tool. The output should automatically be added to your map.
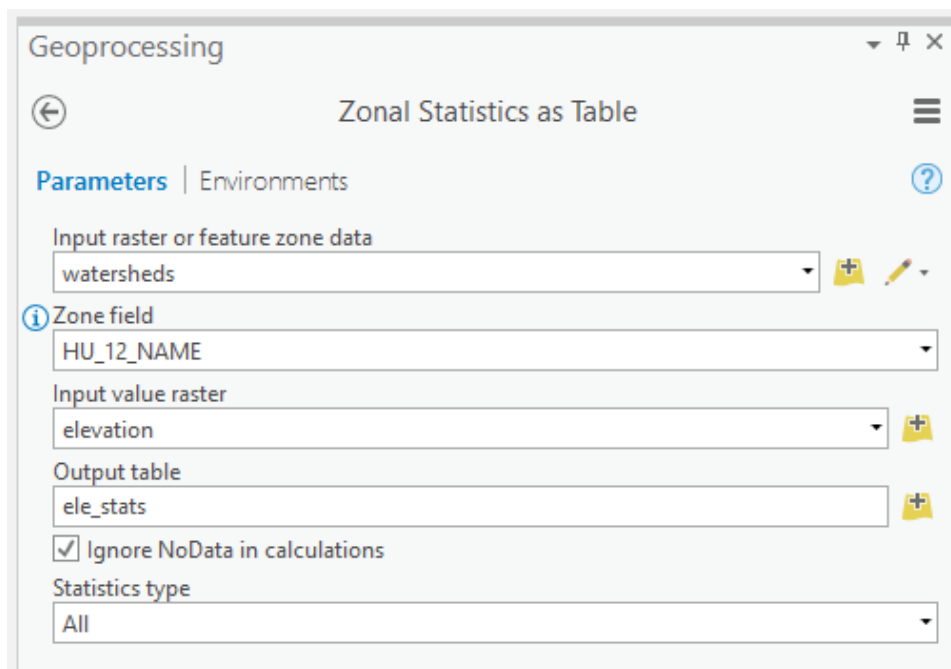

Note that this result could also be joined to the watersheds layers using a table join, as described above. Use the resulting table to answer the following questions.

Question 8. Which watershed has the largest mean elevation? (2 Points)

Question 9. Which watershed had the largest range of elevations? (2 Points)

Question 10. Which watershed has the highest recorded maximum elevation? (2 Points)

Note: It is also possible to calculate individual summary statistics as raster grid surfaces. This can be accomplished with the **Zonal Statistics Tool**. For example, the grid below shows the average elevation for each watershed. Note that all cells in the same watershed are assigned the same value, the mean elevation in that watershed.
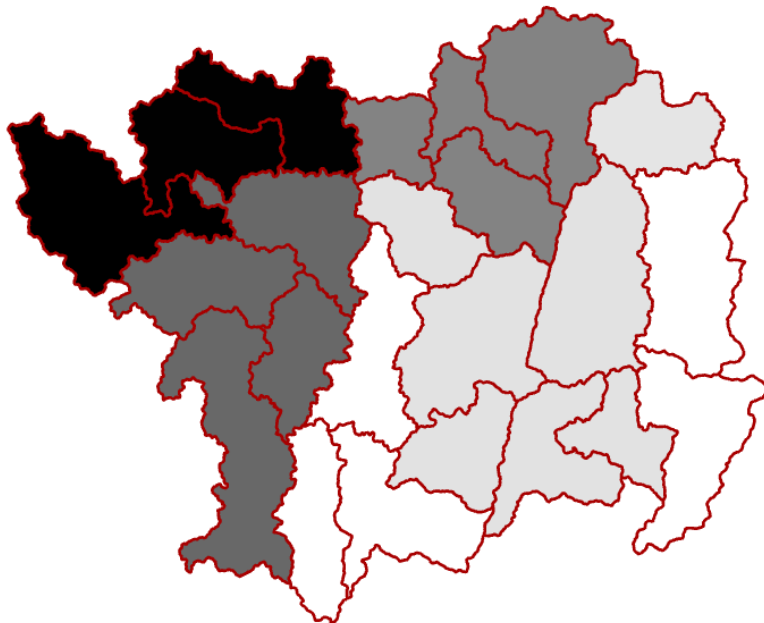

Step 7. Calculate Summary Statistics for a Continuous Raster for Lines

You can also calculate summary statistics for a continuous raster grid relative to line features as opposed to polygon features. This can also be accomplished using the **Zonal Statistic as Table Tool**. You will now calculate summary statistics for the elevation data relative to the mapped streams.

- ☐ Navigate to Spatial Analyst Tools followed by the Zonal subtoolbox.
- ☐ Open the **Zonal Statistic as Table Tool**.
- ☐ Set the Input Raster or Feature Zone Data to the **streams** layer.

- Set the Zone Field to the "FINALNAME2" field. This field provides a unique name for each stream.
- Set the Input Value Raster to the **elevation** layer.
- Name the Output Table **stream_stats**.
- Calculate "All" statistics.
- Click Run to execute the tool. The output should automatically be added to your map.
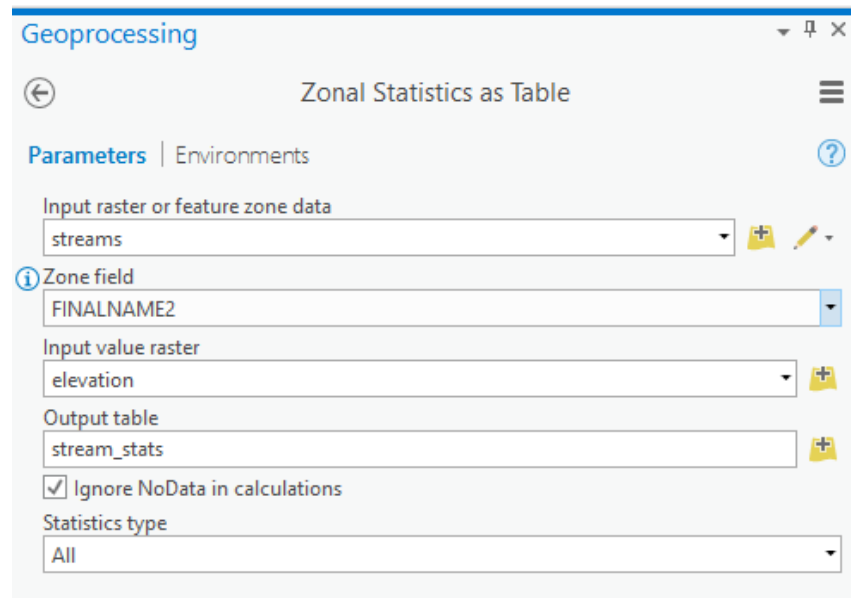
Use the result to answer the following questions.

Question 11. Which stream has the highest recorded elevation? (2 Points)

**Question 12.** What is the mean elevation of Red Creek? (2 Points)

Question 13. Which stream experiences the largest drop in elevation along its course? (2 Points)

Step 8. Extract Raster Values at Points

Lastly, you will extract raster values at points. The Extraction subtoolbox in Spatial Analyst Tools provides two tools to accomplish this task: **Extract Values to Points** and **Extract Multi Values to Points**. Here, we will demonstrate the **Extract Multi Values to Points Tool**. This tool allows you to extract multiple values from different raster grids at once. However, it can also be used for a single grid. We suggest using this tool as opposed to the **Extract Values to Points Tool** because it allows you to name the resulting attribute field.

☐ Navigate to Spatial Analyst Tools followed by the Extraction subtoolbox.
☐ Open the **Extract Multi Values to Points Tool**.
☐ Set the Input Point Features to the **Discharge_Points** layer.
☐ Add the **lc_general** and the **elevation** data layers to the Input Raster List. Under Output Field Name, name the field associated with the land cover "lc" and the field associated with the elevation "elev."
☐ Do not select the "Bilinear interpolation of values at point locations" option.
☐ Click Run to execute the tool.

Note: Unlike the other tools, this tool will add the columns to the original point layer. It will not make a copy of the data. If you don't want to change the original layer, we would recommend doing this work on a copy of the data.
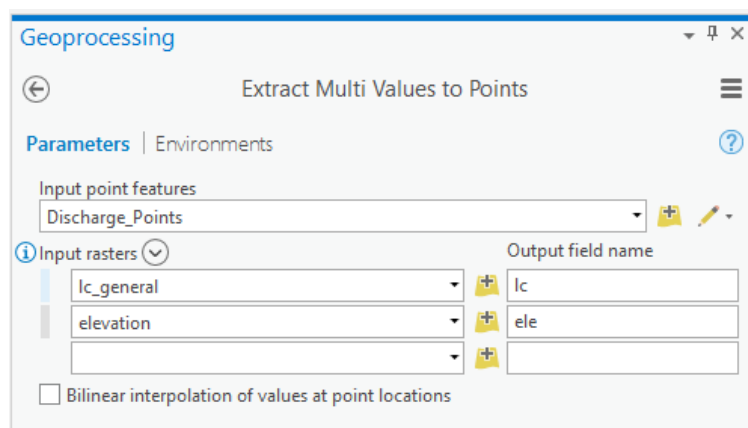
Open the attribute table for the **Discharge_Points** layer to answer the following questions.

Question 14. What was the maximum recorded elevation for any of the discharge points? (2 Points)

Question 15. What land cover type occurred most often at the points? (2 Points)

Hint: You will have to use the **Summary Statistics Tool**.

END OF EXERCISE