

# EXERCISE 3

## Spatial Statistics: Mapping Clusters



### Introduction

This exercise will walk students through the process of summarizing and evaluating geographic distributions; accessing broad geographic patterns and trends over time and in identifying where spatial clustering occurs and where spatial outliers are located.

### Objectives

- Explore the tools and functionality within the Mapping Clusters, Measuring Geographic Distributions, and Analyzing Patterns toolsets
- Evaluate and statistically validate if features, or the values associated with features, form clustered, dispersed, or random spatial patterns
- Perform cluster analysis to identify the locations of statistically significant hot spots, cold spots, and spatial outliers.

### Prerequisites

- ArcGIS Pro installed on your machine



### USDA Non-Discrimination Statement

In accordance with Federal civil rights law and U.S. Department of Agriculture (USDA) civil rights regulations and policies, the USDA, its Agencies, offices, and employees, and institutions participating in or administering USDA programs are prohibited from discriminating based on race, color, national origin, religion, sex, gender identity (including gender expression), sexual orientation, disability, age, marital status, family/parental status, income derived from a public assistance program, political beliefs, or reprisal or retaliation for prior civil rights activity, in any program or activity conducted or funded by USDA (not all bases apply to all programs). Remedies and complaint filing deadlines vary by program or incident.

Persons with disabilities who require alternative means of communication for program information (e.g., Braille, large print, audiotape, American Sign Language, etc.) should contact the responsible Agency or USDA's TARGET Center at (202) 720-2600 (voice and TTY) or contact USDA through the Federal Relay Service at (800) 877-8339. Additionally, program information may be made available in languages other than English.

To file a program discrimination complaint, complete the USDA Program Discrimination Complaint Form, AD-3027, found online at [How to File a Program Discrimination Complaint](#) and at any USDA office or write a letter addressed to USDA and provide in the letter all of the information requested in the form. To request a copy of the complaint form, call (866) 632-9992. Submit your completed form or letter to USDA by: (1) mail: U.S. Department of Agriculture, Office of the Assistant Secretary for Civil Rights, 1400 Independence Avenue, SW, Washington, D.C. 20250-9410; (2) fax: (202) 690-7442; or (3) email: [program.intake@usda.gov](mailto:program.intake@usda.gov).

USDA is an equal opportunity provider, employer, and lender.

---

## Table of Contents

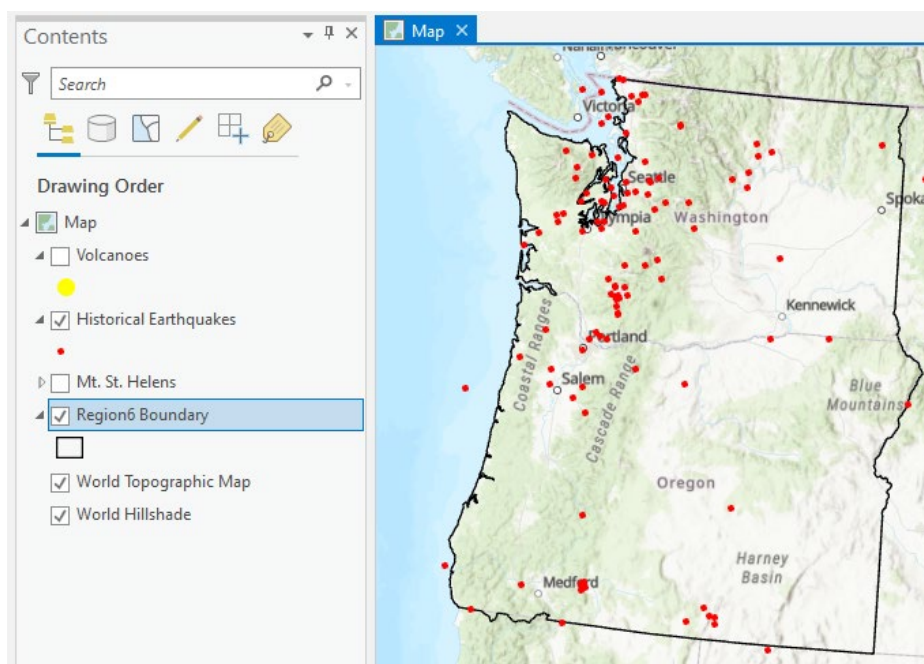
Part 1: Explore the Project Data .....	4
A. Open the Exercise 3 project file.....	4
B. Calculate a total yearly count and averages for Historical Earthquakes. ....	4
Part 2: Explore the Spatial Distribution of the Data .....	5
A. Use the Standard Distance tool to help identify trends in earthquakes. ....	5
B. Use the Standard Distance tool to help identify trends in earthquakes based on the year they occurred. ....	9
C. Explore the attribute data of earthquakes occurring in the 1980 Standard Distance ellipse. ....	11
Part 3: Identify if the Spatial Clustering has Statistical Significance .....	14
A. Use the Cluster and Outlier Analysis (Anselin Local Moran's I) tool to identify statistically significant spatial clusters of earthquakes within the same Year. ....	14
B. Select the Historical Earthquakes with statistically significant clustering and review their attributes. ....	16
Part 4: Utilize the Optimized Hot Spot Tool to Identify High (hot) & Low (cold) Value Clustering of Earthquake Magnitudes. ....	20
A. Make sure your data is projected correctly.....	20
B. Use the Optimized Hot Spot tool to identify statistically significant spatial clustering or sparsity of point locations. ....	20
C. Use the Optimized Hot Spot Analysis tool to identify statistically significant clustering of historical earthquake magnitudes high values (hot spot) and low values (cold spot). ....	22
D. Visualize your results. Use your results to create an interpolated surface and prepare for presentation.....	23

# Part 1: Explore the Project Data

## A. Open the Exercise 3 project file

1. Open Windows Explorer and navigate to your course folder  
....\ArcGISProSpatialStatistics\Data\Exercise\_3
2. Open the **Exercise\_3** ArcGIS Project File.
3. Take a look at the **Historical\_Earthquakes** layer features and attributes.

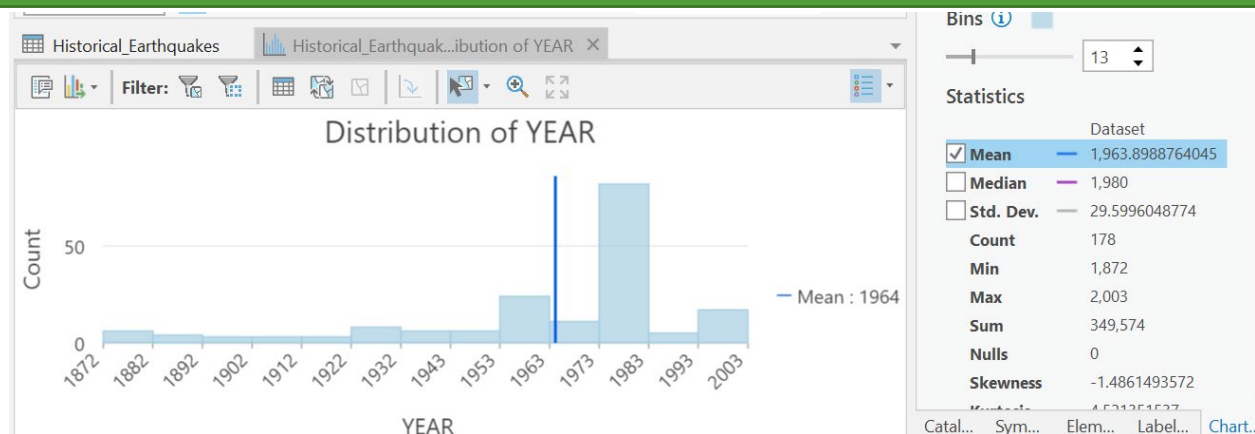
*The Historical Earthquake dataset is an ESRI dataset that presents the locations of significant, historic earthquakes that caused deaths, property damage, and geological effects, or were otherwise experienced by residents in the area.*



4. Is there any noticeable clustering or dispersion amongst the features? Is there any noticeable clustering or dispersion amongst the features' attributes? These are questions we can answer using Spatial Statistics Tools.

## B. Calculate a total yearly count and averages for Historical Earthquakes.

1. Open the **Attribute Table** for **Historical\_Earthquakes**.
  - i. *Right-click* the **Year** field heading then choose **Statistics**.
  - ii. To have the years in the chart displayed as whole numbers, select **Axes** in the Chart Properties then set the X-axis Number format to **Custom** and the Format string to **0**.



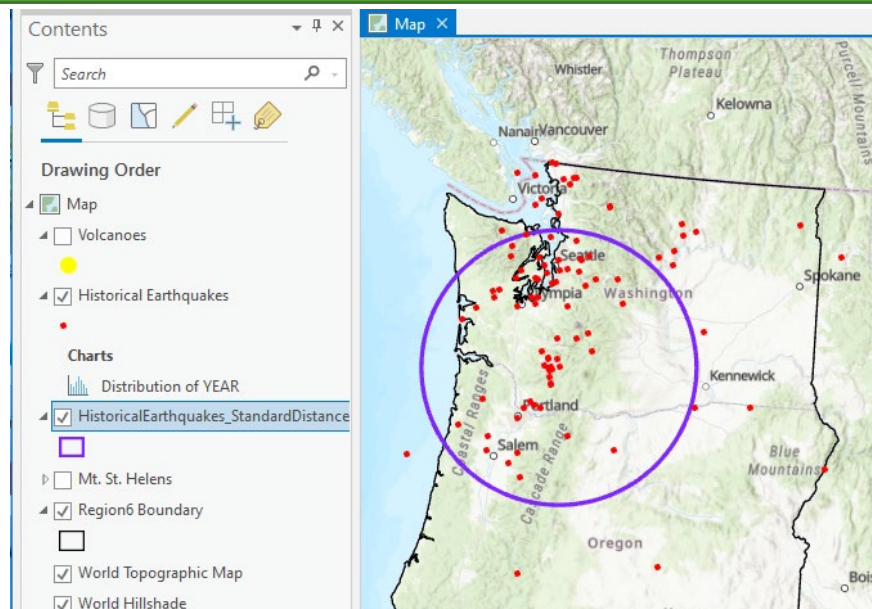
- iii. QUESTION – What year did the most earthquakes occur?
  - iv. QUESTION – What is the total amount of earthquakes in this area for all years?
  - v. QUESTION – Is there a trend such as an increase or decrease of earthquakes over time?
2. Close the statistics chart.

## Part 2: Explore the Spatial Distribution of the Data

### A. Use the Standard Distance tool to help identify trends in earthquakes.

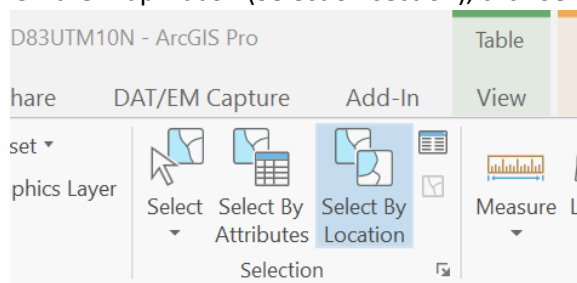
The Standard Distance tool will show us the distribution of the data. We used this tool earlier to determine the distribution of BigFoot sightings.

1. In the **Geoprocessing** pane, go to the **Toolboxes** and expand the **Spatial Statistics** toolbox.
2. Expand the **Measuring Geographic Distributions** toolset then open the **Standard Distance** tool.
3. In the tool window enter the following parameters:
  - i. Input Feature Class: **Historical\_Earthquakes**.
  - ii. Accept the other defaults and click **Run**.
4. The output will look like the following screenshot.



*In the standard distance ellipse we can see that a majority of earthquakes occurred within a specific region which indicates possible clustering. Let's look at statistics for those earthquakes. But first we must select the earthquakes that fall within the standard distance ellipse.*

5. On the Map ribbon (Selection section), click **Select by Location**.



4. In the Select by Location window set the following parameters:

- i. Input Features: **Historical\_Earthquakes**.
- ii. Relationship: **Completely within**.
- iii. Selecting Features: **Historical\_Earthquakes\_StandardDistance**.

?

×

Select By Location

Input Features

Historical Earthquakes

Relationship

Completely within

Selecting Features

HistoricalEarthquakes\_StandardDistance

Search Distance

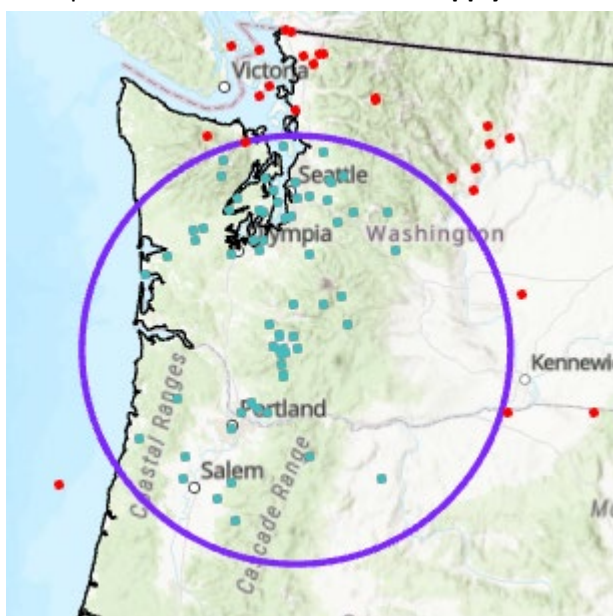
Meters

Selection type

New selection

☐ Invert spatial relationship

iv. Accept the other defaults and click **Apply**.



6. Open the attribute table for **Historical\_Earthquakes**. Notice that *128 out of the 178 Earthquakes are selected*.



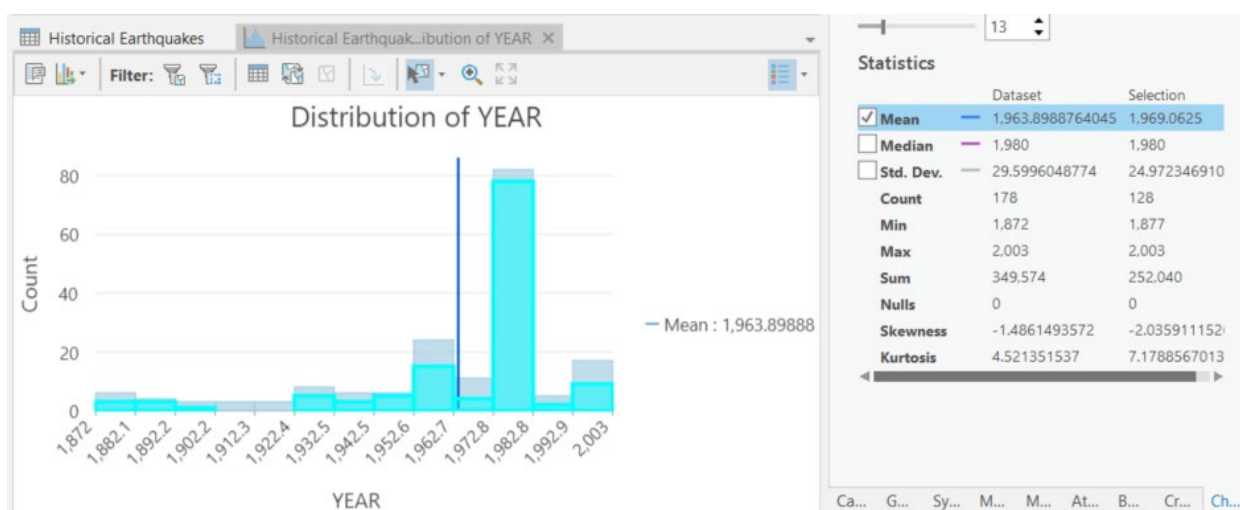
Historical Earthquakes X

Field: Selection:

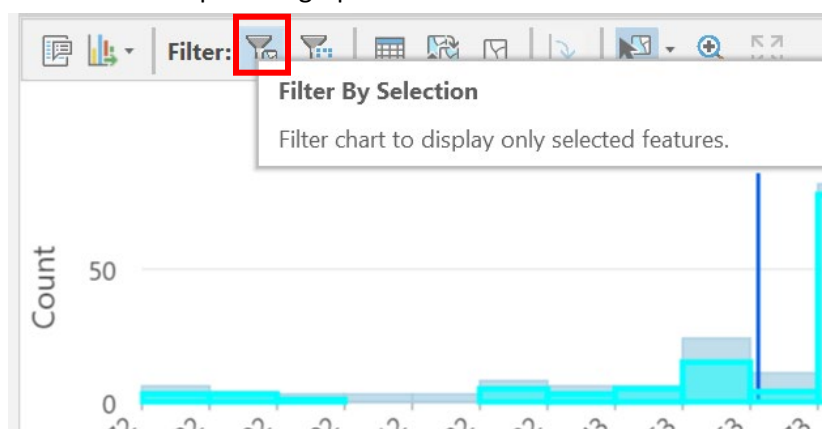
	DEPTH	MAG	MMI	LOCATION	YEAR	MONTH	DAY
6	2	4.5	6	In Washington	1980	4	3
7	5	4.5	5	In Washington	1980	4	9
8	2	4.7	5	In Washington	1980	4	15
9	0	4.7	5	In Washington	1980	4	15

128 of 178 selected Filters:

7. In the attribute table, *right-click* **Year** field heading and select **Statistics**.



i. You can make the chart show only the selected features by clicking the **Filter by Selection** button at the top of the graph.

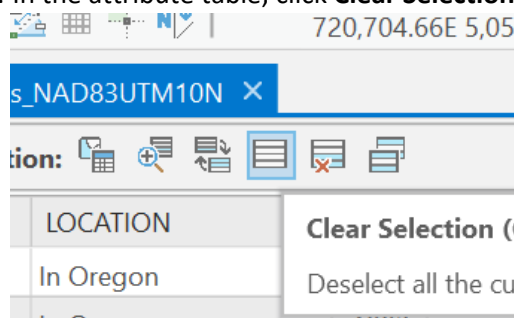


**INTERPRETATION:** Of the earthquakes that fell within our Standard Distance ellipse more than half occurred during 1980. Maybe we should look a little closer at the earthquakes that occurred during that year.

8. Close the *Statistics* chart



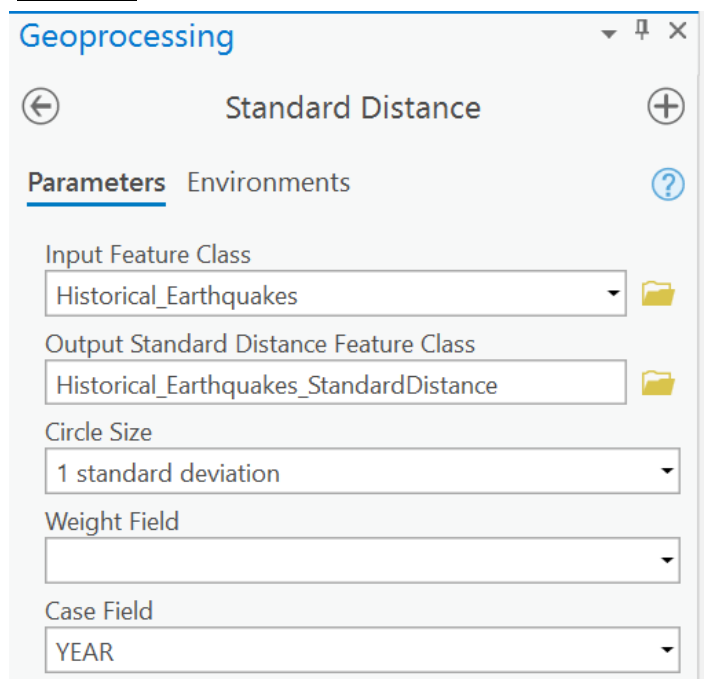
9. In the attribute table, click **Clear Selection**.



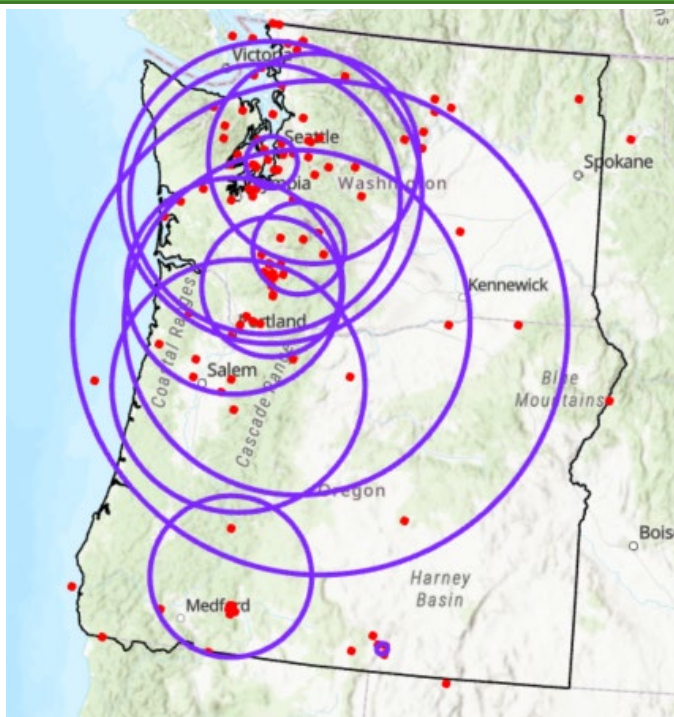
## B. Use the Standard Distance tool to help identify trends in earthquakes based on the year they occurred.

Let's run the standard distance tool again only this time we will examine the annual spatial distribution of the earthquakes by adding YEAR as the optional Case Field. This will create separate distance ellipses for each unique year.

1. From the Measuring Geographic Distributions toolset, open the **Standard Distance** tool.
2. Enter the following parameters:
  - i. Input Feature Class: **Historical\_Earthquakes**.
  - ii. Output Standard Distance Feature Class: **Historical\_Earthquakes\_StandardDistanceYear**.
  - iii. Case Field: **Year**.



- iv. Accept the other defaults and click **Run**. The tool will add a layer to the map view displaying a separate Standard Distance Ellipse for each unique year in the dataset. You can ignore the warnings about too few features.



*Remember, we are only interested in the standard distance ellipse for earthquakes that occurred during the year 1980. We will set a definition query in the layer's properties to only display the ellipse for that year.*

3. In the Contents pane, double-click the **HistoricalEarthquakes\_StandardDistanceYear** layer.
4. In the Layer Properties select **Definition Query** then click **New Definition Query**.
5. Create this query: **Where Year is equal to 1980.**

### Layer Properties: Historical\_Earthquakes\_StandardDistanceYear

General

Metadata

Source

Elevation

Selection

Display

Cache

**Definition Query**

Time

Range

Definition Queries + New definition query

☒ Query 1

Where

YEAR

is equal to

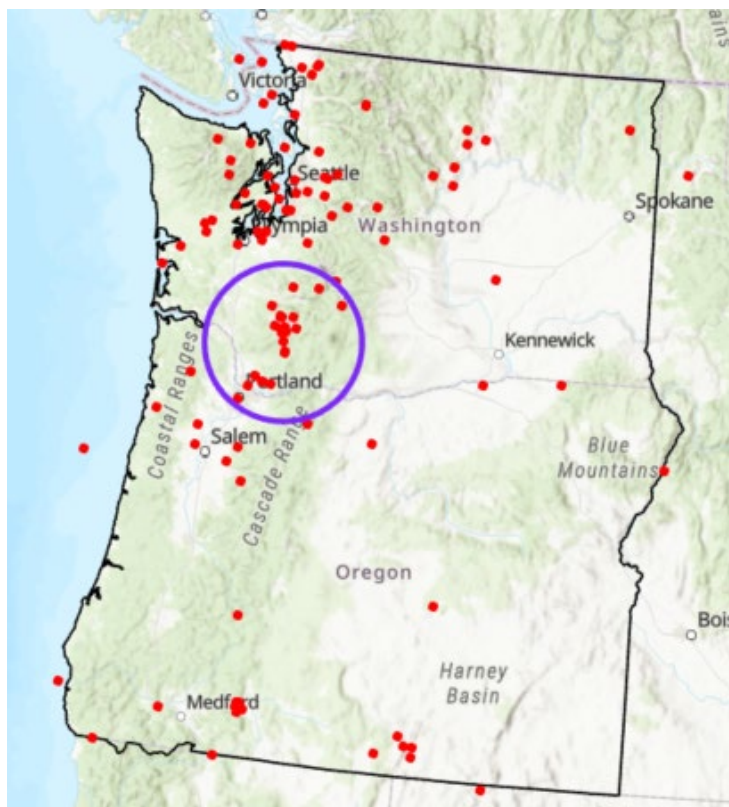
1980

+ Add Clause

SQL ☐

Apply

Cancel



*Evaluating the attributes for earthquakes occurring within the 1980 Standard Distance ellipse may help us identify a possible cause of the increase of earthquakes during that year.*

### C. Explore the attribute data of earthquakes occurring in the 1980 Standard Distance ellipse.

1. Open the **Select by Location** tool and set the following parameters:
  - i. Input Features: **Historical\_Earthquakes**.
  - ii. Relationship: **Completely Within**.
  - iii. Selecting Features: **Historical\_Earthquakes\_StandardDistanceYear**.

**Select By Location**

Input Features: Historical Earthquakes

Relationship: Completely within

Selecting Features: HistoricalEarthquakes\_StandardDistanceYear

Search Distance: Meters

Selection type: New selection

☐ Invert spatial relationship

Apply OK

- iv. Click **Apply** and close the Select by Location window.
2. Now let's look at the attributes of the selected earthquakes. Open the attribute table of **Historical\_Earthquakes**.
3. Click the **Show Selected Records** button near the bottom of the window.

Historical Earthquakes

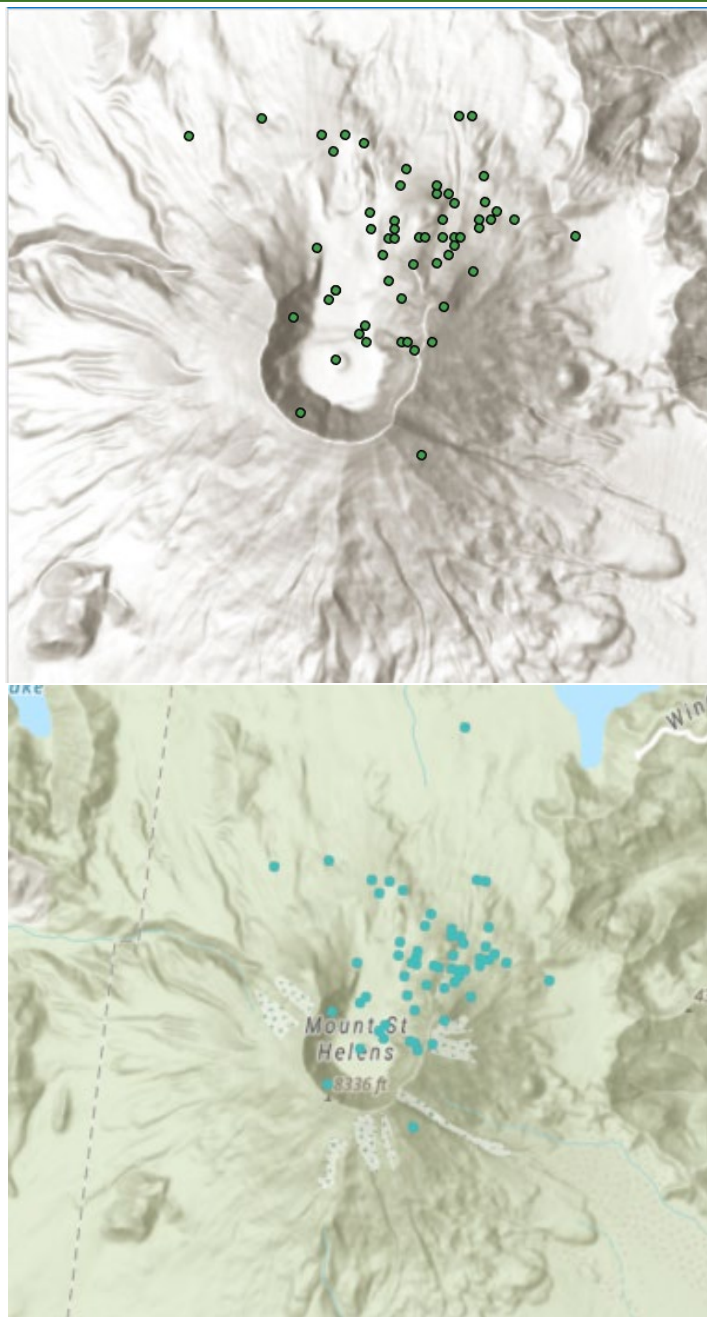
Field: Selection: Highlighted:

OBJECTID_1 *	Shape *	OBJECTID	DEPTH	MAG	MMI	LOCATION	YEAR	MONTH
20	Point	1966	7	4.8	5	In Washington	1961	9
21	Point	1967	7	5.1	6	In Clark County, near t...	1961	9
22	Point	1968	10	4.5	5	In Washington	1980	4
23	Point	1969	-9999	5	7	North of Portland, Ore...	1877	10
24	Point	1970	-9999	5	5	WASHINGTON-OREGO...	1892	2
25	Point	1971	-9999	5	6	Near Portland, Oregon	1941	12

81 of 178 selected Filters: 100%

Show selected records

4. Close the attribute table.
5. In the Contents pane, right-click the **Mt\_St\_Helens** layer then click **Zoom to Layer**. Notice that most of the selected features are right around Mt. Saint Helens.



*As you can see, in May of 1980 Mount Saint Helens erupted. This could be a major factor in the increase of earthquakes that year. We can verify this assumption with further analysis.*

**NOTE:** 62 of the earthquakes that occurred in 1980 are within two miles of the volcano's center.

6. Clear Selected features.
7. Zoom to the extent of the **Region6\_Boundary** layer.

## Part 3: Identify if the Spatial Clustering has Statistical Significance

We have already identified clustering amongst some of the earthquakes, but is there clustering of earthquakes based on the year they occurred or their magnitudes? And is the clustering statistically significant? These questions can be answered with the Cluster and Outlier Analysis tool.

### A. Use the Cluster and Outlier Analysis (Anselin Local Moran's I) tool to identify statistically significant spatial clusters of earthquakes within the same Year.

*Given a set of weighted features, the Cluster and Outlier Analysis tool identifies statistically significant hot spots, cold spots, and spatial outliers using the Anselin Local Moran's statistic.*

1. Go to the Spatial Statistics toolbox and expand the **Mapping Clusters** toolset
2. Open the **Cluster and Outlier Analysis (Anselin Local Moran's I)** tool and use the following parameters:
  - i. Input Feature Class: **Historical\_Earthquakes**
  - ii. Input Field: **YEAR**.
  - iii. Output Feature Class: **HistoricalEarthquakes\_ClustersOutliers**
  - iv. Conceptualization of Spatial Relationships: **Fixed Distance Band**.
  - v. Distance Band or Threshold Distance: **8045**. (Note: our data is in meters; 1 mile = 1609 meters, we want a distance band of 5 miles).



**Geoprocessing**

Cluster and Outlier Analysis (Anselin...)

**Parameters** Environments

Input Feature Class  
Historical Earthquakes

Input Field  
YEAR

Output Feature Class  
HistoricalEarthquakes\_ClustersOutliers

Conceptualization of Spatial Relationships  
Fixed distance band

Distance Method  
Euclidean

Standardization  
Row

Distance Band or Threshold Distance  
8045

☐ Apply False Discovery Rate (FDR) Correction

Number of Permutations  
499

- vi. Accept the other defaults and click **Run**. You will receive a warning that some of the features in the dataset have no neighboring features within the given specified distance threshold. This is ok.
- vii. Turn off **Historical Earthquakes** and **HistoricalEarthquakes\_StandardDistanceYear** layers in the Contents pane.
- viii. Examine the output in the map.

*INTERPRETATION: The output feature class has the following attributes for each feature in the input feature class: local Moran's I index, z-score, p-value, and cluster/outlier type (COType).*

*The output field, COType, distinguishes between a statistically significant cluster of high values (HH), a cluster of low values (LL), an outlier in which a high value is surrounded primarily by low values (HL), or an outlier in which a low value is surrounded primarily by high values (LH). Statistical significance is set at the 95 percent confidence level for this tool.*

*Our results show statistically significant clustering in two locations of our study area.*





## B. Select the Historical Earthquakes with statistically significant clustering and review their attributes.

Again, evaluating the attributes for earthquakes occurring in locations with statistical significance may help us identify trends in the data.

1. On the Map tab (Selection group) click the **Select by Attribute** button and set the following parameters:
  - i. Input Rows: **Historical\_Earthquakes\_ClustersOutliers**
  - ii. Selection Type: **New selection.**
  - iii. Query: **Where COType Fixed 8045 RS is equal to HH.**

**Select By Attributes**

Input Rows  
Historical\_Earthquakes\_ClustersOutliers

Selection type  
New selection

Expression  
Load Save Remove

Where COType Fixed 8045 RS is equal to HH

+ Add Clause

SQL ☐

- iv. Click **Apply** then close the Select by Attributes window.
2. Open the attribute table for **Historical\_Earthquakes\_ClustersOutliers**.
  - i. Click the **Show selected records** button. Notice there are 68 features selected that have statistically significant clustering.

OBJECTID *	Shape *	SOURCE_ID	YEAR	LMIndex Fixed	LMiZScore Fixed	LMiPVal
79	Point	79	1980	0.115423	4.18651	
80	Point	80	1980	0.115423	3.999417	
81	Point	81	1980	0.115423	4.440543	
82	Point	82	1980	0.115423	4.31384	
83	Point	83	1980	0.115423	4.199121	
84	Point	84	1980	0.115423	4.081374	

68 of 178 selected Filters: 100%

- ii. Examine the **Year** field.

*There are three different years having statistical significance; 6 in 1993, 1 in 1994, and 61 in 1980. We have already determined the probable cause of earthquakes for 1980, so we will eliminate those and look closer at the others.*

3. Open the Select by Attributes tool and set the following parameters:
  - i. Input Rows: **Historical\_Earthquakes\_ClustersOutliers**.
  - ii. Selection Type: **Remove from the current selection**.
  - iii. Query: **Where Year is equal to 1980**.

**Select By Attributes**

Input Rows  
Historical\_Earthquakes\_NAD83UTM10N\_ClustersOutliers

Selection type  
Remove from the current selection

Expression  
Load Save Remove

SQL ☐

Where YEAR is equal to 1980

Apply OK

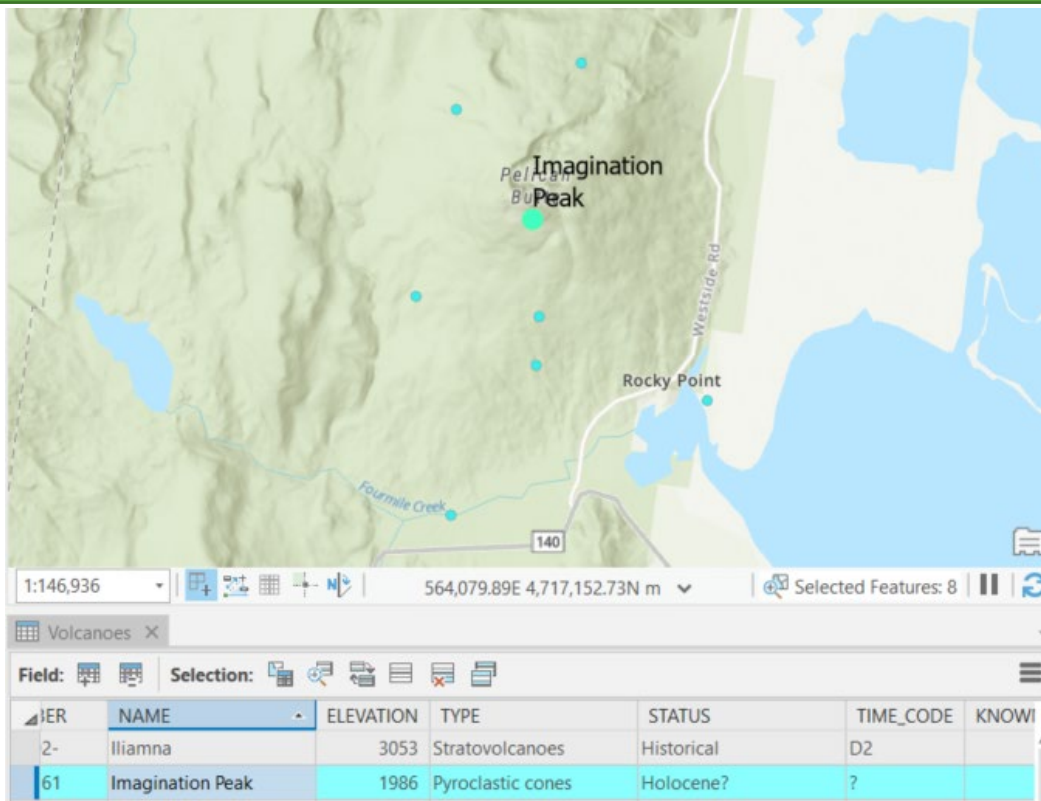
iv. Click **Apply** then close the Select by Attributes window.

*The earthquakes occurring between the years 1993-1994 are all clustered above the southern boundary of our study area. The increase of earthquakes during 1980 appears to be caused by volcanic activity; can we make the same assumption with clustering of earthquakes during 1993-1994?*

OBJECTID *	Shape *	SOURCE_ID	YEAR	LMiIndex Fixed	LMiZScore Fixed	LMiPVal
2	Point	2	1993	0.749543	2.235	
3	Point	3	1993	0.749543	2.287254	
4	Point	4	1993	0.752432	1.817865	
5	Point	5	1993	0.743764	1.703334	
6	Point	6	1993	0.743764	1.761037	
7	Point	7	1993	0.749543	2.230047	
8	Point	8	1994	0.778438	1.830878	

7 of 178 selected Filters: 100%

4. In the Contents pane, turn on the **Volcanoes** layer and open its attribute table.
  - i. Right-click the Name field header then click **Sort Ascending**.
  - ii. Locate the **Imagination Peak** record in the table then double-click the left end of that row to make the map zoom to that feature.



The table below defines the values of the *TIME\_CODE* field in the *Volcanoes* attribute table. Because the date of an eruption from Imagination Peak is “uncertain”, we cannot confirm volcanic activity as an explanation for the increase in earthquakes during these years, however we do know there is less than a 5% chance that the clustering of the earthquakes in this area occurred by random chance.

D1	Last known eruption 1964 or later
D2	Last known eruption 1900-1963
D3	Last known eruption 1800-1899
D4	Last known eruption 1700-1799
D5	Last known eruption 1500-1699
D6	Last known eruption A.D. 1-1499
D7	Last known eruption B.C. (Holocene)
U	Undated, but probable Holocene eruption
Q	Quaternary eruption(s) with the only known activity being
?	Uncertain Holocene eruption

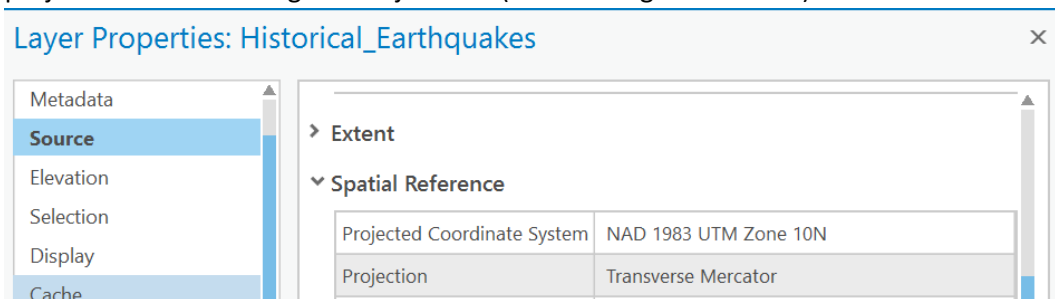
- After examining the results uncheck all the layers in the Contents pane except for **Region6\_Boundary**.
- Then right-click on **Region6\_Boundary** and **Zoom to Layer**.

## Part 4: Utilize the Optimized Hot Spot Tool to Identify High (hot) & Low (cold) Value Clustering of Earthquake Magnitudes.

*The Optimized Hot Spot tool identifies statistically significant spatial clusters of high values (hot spots) and low values (cold spots). It automatically aggregates incident data, identifies an appropriate scale of analysis, and corrects for both multiple testing and spatial dependence. This tool interrogates your data in order to determine settings that will produce optimal hot spot analysis results. If you want full control over these settings, use the Hot Spot Analysis tool instead.*

### A. Make sure your data is projected correctly

1. In the Contents pane, double-click **Historical\_Earthquakes** to open the Layer Properties window.
2. Click on the **Source** tab then expand the **Spatial Reference** section.
3. Ensure that the layer is set to **NAD\_1983\_UTM\_Zone\_10N**. For this example we have the correct projected coordinate system. If the projection wasn't correct you would have to re-project the dataset using the Project tool (Data Management Tools).



4. Close the Layer Properties window.

### B. Use the Optimized Hot Spot tool to identify statistically significant spatial clustering or sparsity of point locations.

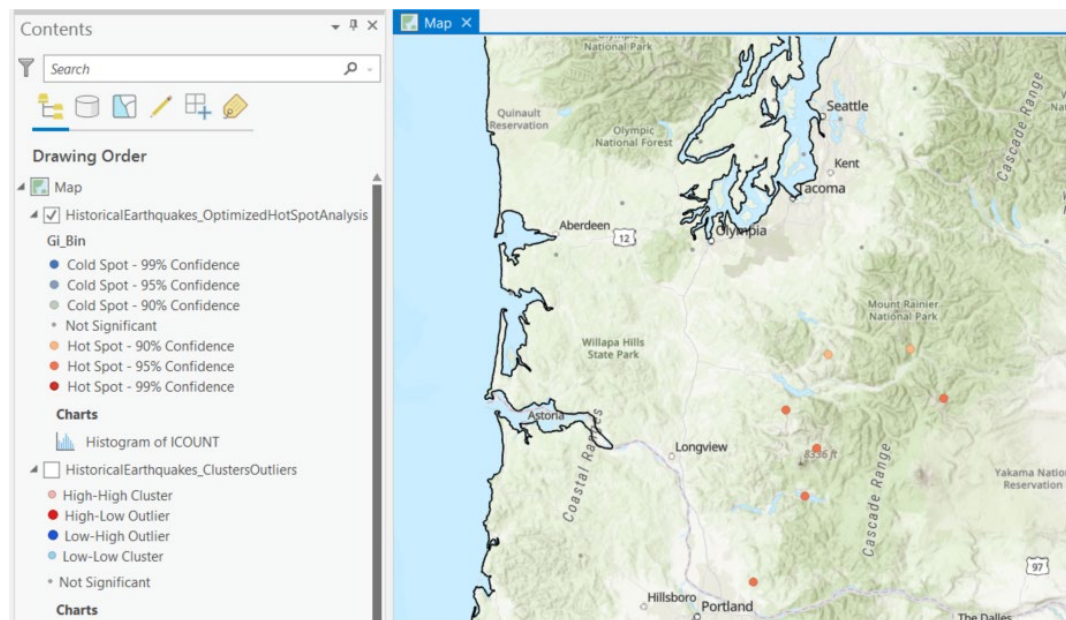
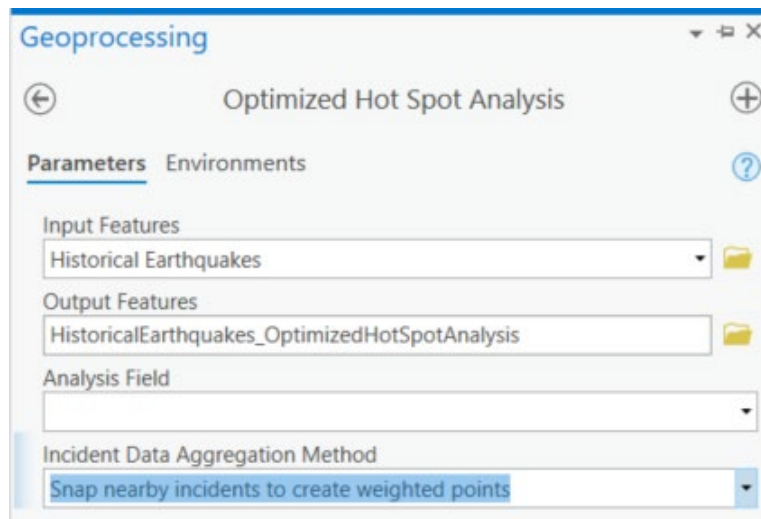
With point data you will sometimes be interested in analyzing data values associated with each point feature and will consequently provide an Analysis Field. In other cases you will only be interested in evaluating the spatial pattern (clustering) of the point locations or point incidents. The decision to provide an Analysis Field or not will depend on the question you are asking.

When the Input Features you provide represent incident data (when you don't provide an Analysis Field), the tool will aggregate the incidents and the incident counts will serve as the values to be analyzed.

1. In the Mapping Clusters toolset open the **Optimized Hot Spot Analysis** tool and set the following parameters:
  - i. Input Features: **Historical\_Earthquakes**.

*Leave the Analysis field blank. Analyzing point features when there is no Analysis Field allows you to identify where point clustering is unusually (statistically significant) intense or sparse. This type of analysis answers questions like: Where are there many points? Where are there very few points? When you do not provide an Analysis Field the tool will aggregate your points in order to obtain point counts to use as an analysis field.*

ii. Incident Data Aggregation Method: **Snap nearby incidents to create weighted points**



*Looking at the results we can see some statistically significant clustering of historical earthquakes (hot spots) centered on Mt. Saint Helens. We also see many “not significant” values. No cold spots were identified.*

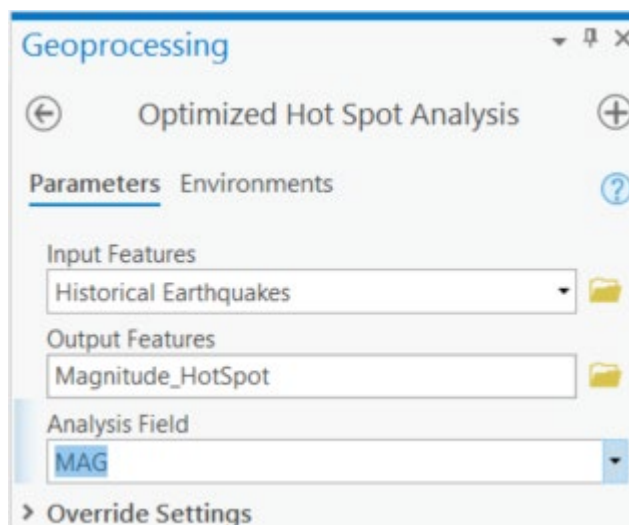


2. When you are done examining the data, uncheck HistoricalEarthquakes\_OptimizedHotSpotAnalysis and zoom back out to the Region6\_Boundary extent.

C. Use the Optimized Hot Spot Analysis tool to identify statistically significant clustering of historical earthquake magnitudes high values (hot spot) and low values (cold spot).

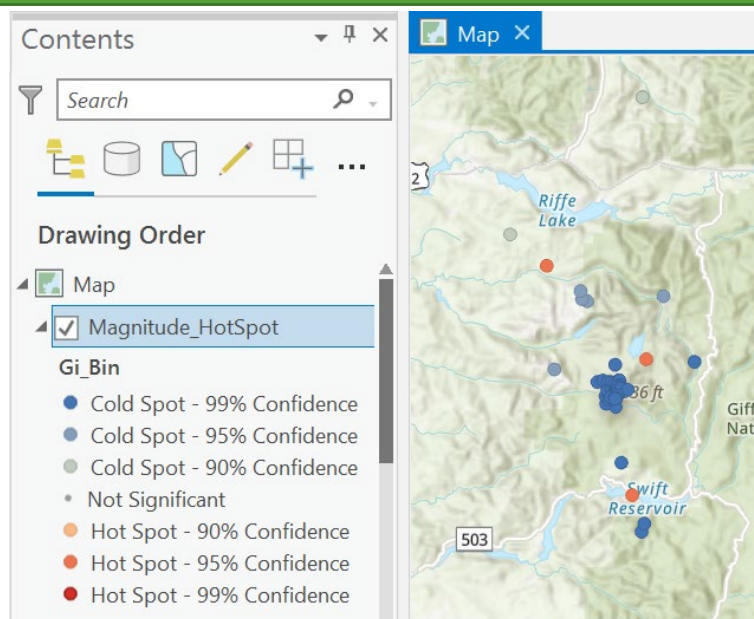
Analyzing point features **with** an Analysis Field allows you to answer questions like: Where do high and low values cluster? The points are not aggregated when using the Analysis Field like they were in the previous step when your input features represented incident data.

1. Open the **Optimized Hot Spot Analysis Tool** and enter the following parameters:
  - i. Input Features: **Historical\_Earthquakes**
  - ii. Name the Output Feature ...\**Magnitude\_HotSpot**.
  - iii. Analysis Field: **MAG** (for Magnitude).



- iv. Click **Run**.
2. Examine the output in the map.





3. Open the Magnitude\_HotSpot attribute table and examine the data.

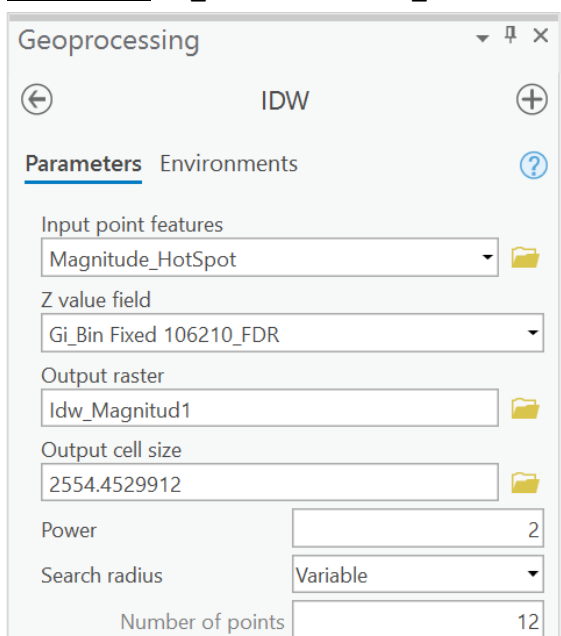
Magnitude_HotSpot						
Field:	Selection:					
OBJECTID *	Shape *	SOURCE_ID	MAG	GiZScore Fixed 106210	GiPValue Fixed 106210	
1	Point	1	4.6	1.760759	0.078279	
2	Point	2	4.4	0.758721	0.448019	
3	Point	3	6	0.758721	0.448019	
4	Point	4	4.5	0.746907	0.45512	
5	Point	5	6	0.758721	0.448019	

**INTERPRETATION:** The result of the Hot Spot Analysis tool is a new feature class where every feature in your dataset is symbolized based on whether it is part of a statistically significant hot spot, a statistically significant cold spot, or is not statistically significant. As expected there is clustering around Mt. Saint Helens again but this time as cold spots. Meaning that there is a clustering of lower magnitude earthquakes in this area compared to the rest of the data. Surrounding the Puget Sound we have a number of hot spots meaning that area has a clustering of earthquakes with a higher magnitude relative to the rest of the Pacific Northwest.

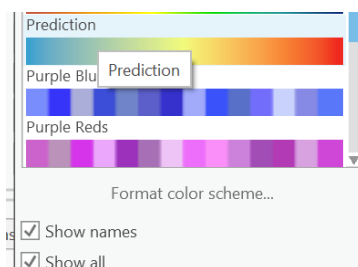
D. Visualize your results. Use your results to create an interpolated surface and prepare for presentation.

*While you as the analyst understand that each location has a particular z-score and a particular p-value that determine the symbology and can guide decision-making, you may be required to share your analysis with a broader audience who may be expecting the results to look like the heat maps people have become accustomed to in the media. Thus, the next step is to create a continuous surface to visualize the results. Remember that this surface is for visualization purposes only, and the true statistical analysis happens feature by feature. Showing both the surface and the true results of the hot spot analysis at the same time is a great way to present both the statistical results and the more approachable visualization. The method that you will use in this tutorial, IDW (Spatial Analyst), interpolates a raster surface from points using an inverse distance weighted technique.*

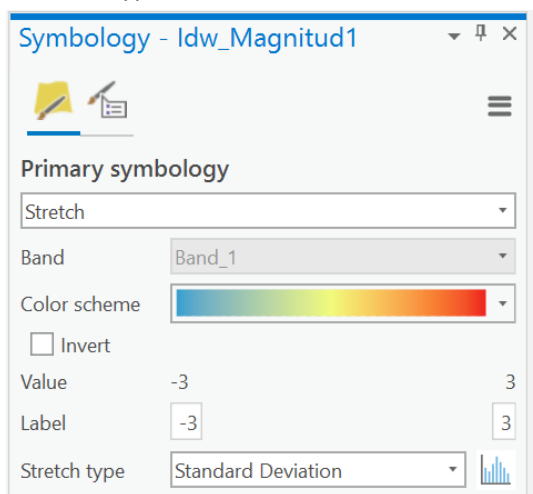
1. Go to the **Spatial Analyst** toolbox (you need the Spatial Analyst license for this), then expand the **Interpolation** toolset.
2. Open the **IDW** tool and set the following parameters.
  - i. Input point features: **Magnitude\_HotSpot**.
  - ii. Z value field: **Gi\_Bin Fixed 106210\_FDR**.



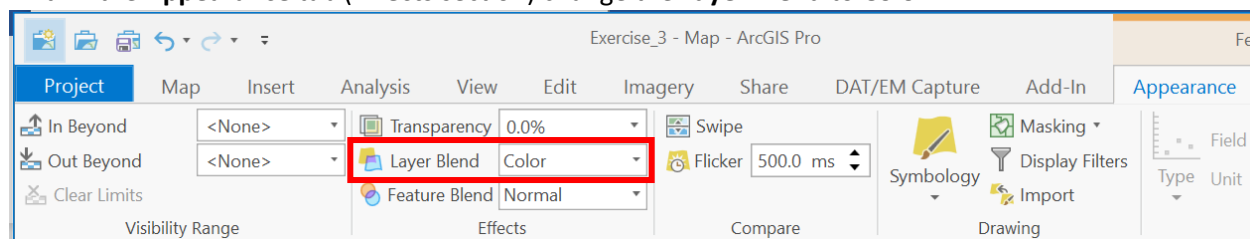
- iii. Accept the other defaults and click **Run**.
3. Right-click the new **Idw\_Magnitud1** layer and click **Symbology** and set the following options:
  - i. Primary Symbology: **Stretch**.
  - ii. Color scheme: **Prediction** (To match our Magnitude\_HotSpot layer color scheme - high values are red and low are blue). Make sure the **'Show names'** and **'Show all'** boxes are checked.



iii. **Stretch Type: Standard Deviation**

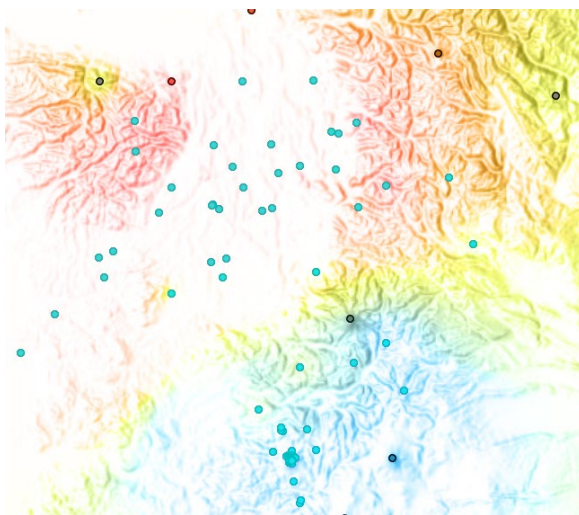


4. Within the **Appearance** tab (**Effects** section) change the **Layer Blend** to **Color**.



5. Turn off all other layers besides World Hillshade and Idw\_Magnitud1. The display should resemble the screen grab below.

*Note: The best results from IDW are obtained when sampling is sufficiently dense with regard to the local variation you are attempting to simulate.*



*One application of spatial statistics and the Hot Spot Analysis tool is to help to resolve resource allocation problems. Where would you allocate additional earthquake emergency response resources?*

**Congratulations! You've completed Exercise 3.**