Lesson 2: Introduction to Geospatial Analysis

## INTRODUCTION

## This lesson is an introduction to geospatial analysis. You will learn about descriptive statistics including central tendency of the data, distribution of the data, and the shape of the data. In this lesson you will gain an understanding of how to classify data using various statistical methods. You will also learn about creating graphic statistical models and the elements associated with them.

## LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Explore data relationships using geospatial data.

2. Create simple data sets using a table operation method.

3. Classify quantitative data using a variety of statistical methods.

4. Create a scatter plot of data.

5. Analyze scatter plot data to produce a presentation of results.

## LEARNING SEQUENCE

|  |  |
| --- | --- |
|  | |
| Required Reading | Read the following:  Introduction to Geospatial Analysis   * Data Classification * Classification Methods * Table Operations * Creating Graphic Statistical Models |
| Assignments | Complete the following:   * Lab: Introduction to Geospatial Analysis * Quiz: Introduction to Geospatial Analysis |

## INSTRUCTION

**Introduction to Geospatial Analysis**

## Descriptive Statistics

Descriptive statistics are quantitative descriptions of data that provide some basic descriptions of the data set.

Cartographers use descriptive statistics to explore the character of the data. Descriptive statistics can describe things such as the central tendency of the data, the dispersion of the data, and the shape of the data.

|  |
| --- |
| **Key Facts**  **Descriptive Statistics**   * Are quantitative descriptions of data * Allow a cartographer to explore character of data * Describe central tendency, dispersion, and shape |

## Central Tendency

Central tendency of data describes the distribution of the data. There are six measures that we commonly used to describe the central tendency of a data set: maximum, minimum, range, mean, median, and mode.

* The **maximum** measure of central tendency simply reports the maximum value in a data set.
* The **minimum** measure of central tendency returns the minimum value in a data set.
* The **range** measures how high and low a series of numbers span. In order to determine the range you would identify the highest and lowest numbers in the series and then subtract the lowest number in from the highest. The range measure returns the maximum value minus the minimum value of a data set.
* The **mean** measure of central tendency is the average value of the data set. The average value is defined by summing all values in the data set and dividing the sum by the number of values in the data set.
* The **median** measure of central tendency reports the middle point of the data. For instance, if we had three observations in our data set, we would sort the observations in ascending order and then look at the value at the midway point which would be the second observation of this case. If there is an even number of observations in a data set, then the median will be the average of the two most central observations.
* The **mode** measure of central tendency reports the most common value found in the data set. If multiple values are tied as the most common value, the data set has multiple mode values.

**Dispersion**

Dispersion measures the variability of the data set. There are two measures of dispersion: variance, and standard deviation.

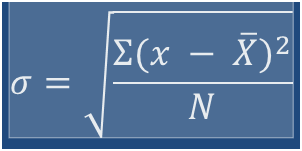
**Variance**

The variance measure of dispersion takes the sum of the squares of the deviations divided by the number of observations.

The units of variance are identical to the original units of measure. So for instance, if the original units of the observations were in feet, then they variance will report how much the observations vary on average in feet.

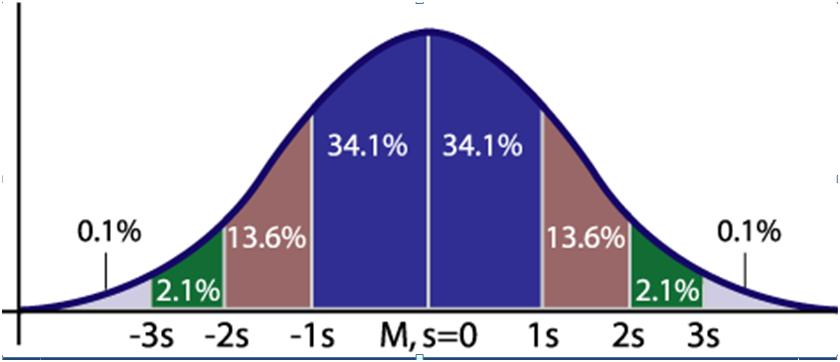
**Standard Deviation**

The standard deviation is similar to the variance except that it takes the square root of the sum of the squares of the variance divided by the number of observations.



The standard deviation measures dispersion in a standard way so that two different data sets can be compared. Standard deviation does not report the dispersion in the original units of the observations.

Review the normal distribution of data provided which illustrates that most of the examples of the data set are close to the average value while relatively few examples tend to be positioned at one extreme or the other. The x-axis is the value in question and the y-axis is the number of observations for each value on the x-axis.



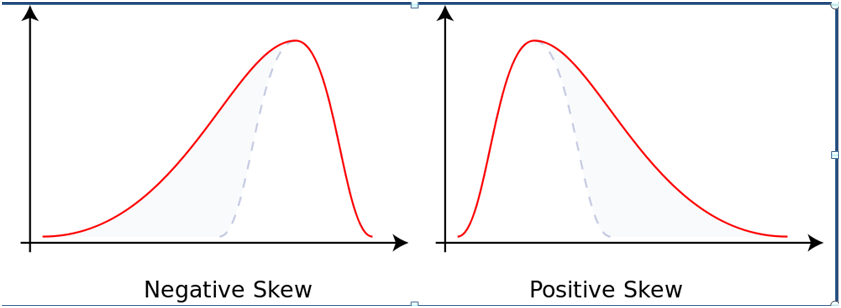
Normal Distribution of Data

The standard deviation tells us how tightly all the various observations are clustered around the mean of the data set. When the examples are pretty tightly bunched together, the bell-shaped curve is steep, and the standard deviation is small. When the examples are spread apart, the bell curve is relatively flat, which tells you that you have a reasonably large standard deviation.

One standard deviation away from the mean in either direction on the horizontal axis accounts for about 68% of the observations in the data set. Two standard deviations away from the mean of the four areas closest to the center accounts for about 95% of the observations of the data set. Three standard deviations account for about 99% of all the observations of the curve.

## Skewness

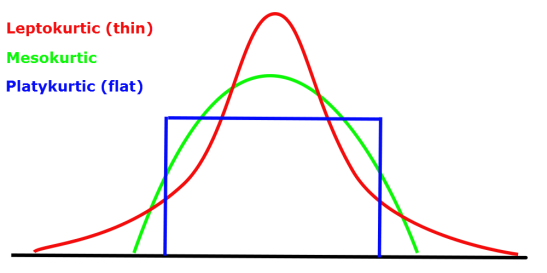
The skewness measure tells us whether the peak of a distribution is to one side of the mean or the other. If a data set has a negative skew, then the peak of the distribution is above the average, if a data set has a positive skew, the peak of the data set is below the average.



## Kurtosis

The final measure of central tendency is kurtosis. Kurtosis describes the flatness or peakedness of a distribution. For normal distribution, the kurtosis is equal to the value of 3.0. A normally distributed data set having the kurtosis value of 3.0 is a mesokurtic distribution. A value above 3.0 is a leptokurtic distribution. A value less than 3.0 is a platykurtic distribution.

Below is an illustration of the three types of kurtosis.

* The green line is the normal distribution with a value of 3.0 and is mesokurtic.
* The red line with a strong peak is the leptokurtic distribution and has a value greater than 3.0.
* The blue line which has a flat top is the platykurtic distribution which has a value less than 3.0.

**Data Classification**

## What is Data Classification?

Data classification categorizes objects based on a set of conditions into separate bins or classes. Classification may add to or modify attribute data for each geographic object. For example, a classification could add a new code such as large or small.

Classification could also recode an attribute, such as changing urban to dense. One attribute can yield many different maps depending on which classification method is chosen. Different classification methods will have a direct effect on how the map and data are perceived by the map user, therefore much care must be taken when choosing a data classification method.

## Classifying Data

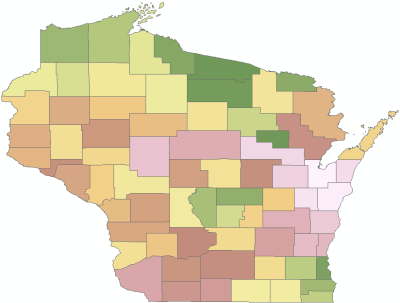
The primary reason to classify data is to simplify the data for visual display. There are three goals for data classification with regards to cartography.

**Goal 1:** Simplify the visualization so that spatial patterns of distribution can

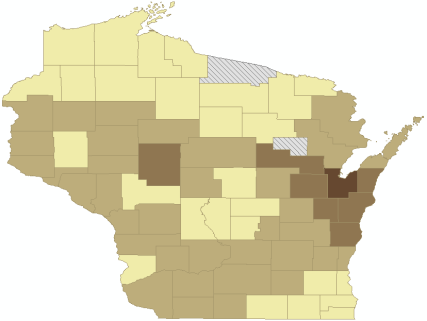
become more easily viewable by the map reader.

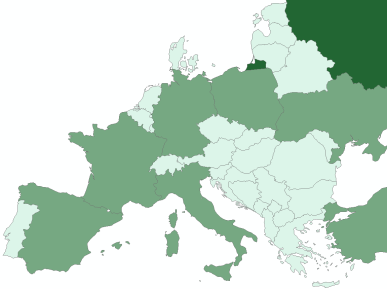
**Goal 2:** Group similar observations together.

**Goal 3:** Show the difference between the groups.



For instance, this map assigns each value a unique color, which means, that each county has its own unique color. It is very difficult to look for patterns when each county has its own unique color.

This map uses the same data, but uses a data classification method to categorize the counties to make it easier to visualize the data. The lighter values that represent less of an item and the darker values represent more of an item; spatial patterns of distribution begin to emerge.



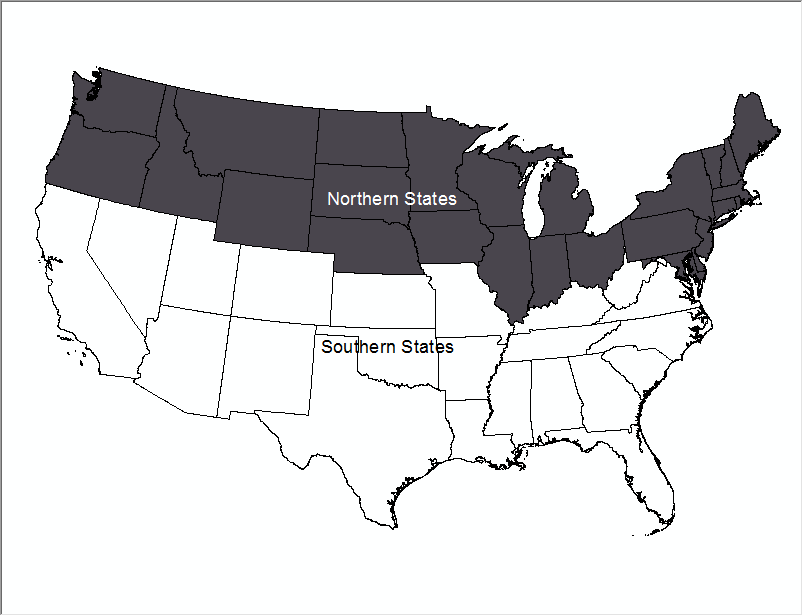
Here is an example of a classification showing countries with high, medium, and low values of population.

**Classification Methods**

## Binary Classification Method

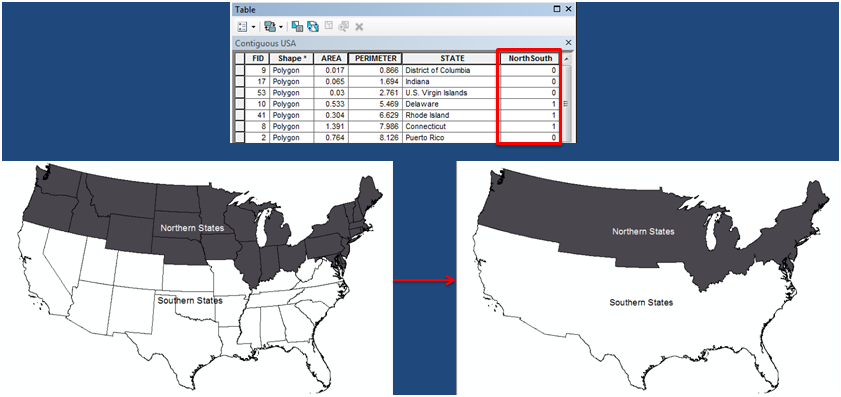
Binary classification places objects into two classes. The two classes can be the value of 0 and 1, true and false, or any other dichotomy that you can think of. Typically, a binary classification is used to store the results of complex operations when the operation returns either a yes or no answer.

In this example, the states were classified into two binary classes: one class for the northern states and one class for the southern states.



## Dissolve Classification Method

The dissolve method combines similar features within a data layer based on a shared attribute. For example, continuing the northern and southern states example, we can dissolve all the states based on whether they are a northern state or a southern state. The dissolve operation creates new geometry and in this case, two polygons. In regards to this pair of new polygons, one represents the combined extent of all states classified as northern, while the other represents the combined extent of all states classified as southern. Take note that the new geometry does not transfer any of the attributes of the dissolved.

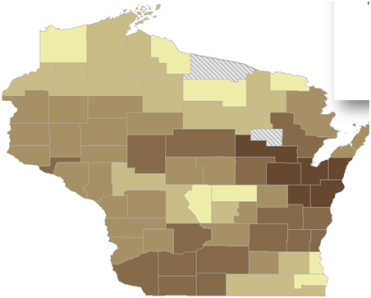


**Automatic Classification**

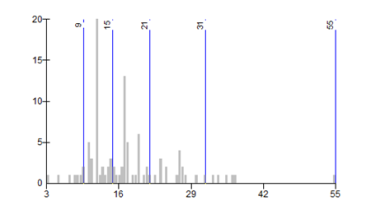
An automatic classification method is where you set up rules for the computer to follow and then the computer executes the data classification. For example, let’s pretend we are building a vending machine. In order for our machine to determine whether the user has inserted enough money, we need the machine to determine which type of coin was inserted. Therefore, we set up rules for the vending machine to sort the coins based on their diameter and width. With those rules in place, the vending machine will be able to sort the coins into different bins, thereby determining the amount of money the user inserted. This is an example of automatic classification. In a GIS, it is a similar operation. We tell the computer that we want data within a certain range of values to go into different bins, thereby sorting them and preparing them for visualization on a map.

## Jenks Natural Breaks

The Jenks Natural Breaks classification method aims to maximize homogeneity in classes. It uses breaks in a histogram as class breaks and assumes that group data are alike.



Here is an example of the Wisconsin map classified using the natural breaks method. The map displays the result of the classification visually and shows the five classes varying from a light brown to a dark brown in a sequential color. This map is interesting to look at, and displays interesting patterns in the data.

This histogram illustrates the values where the x-axis is the observation value, and the y-axis is the number of observations for each value. The vertical blue lines show the extent of each class and are considered to be the class breaks. Looking at the histogram and the blue lines, you can see that natural breaks tend to place the class breaks where there are natural valleys or vacancies in the data set.

**Advantages and Disadvantages of the Jenks Natural Breaks Classification Method**

Advantages

* Considers distribution of data
* Minimizes in-class variance
* Maximizes between class variance
* Produces classification with high accuracy

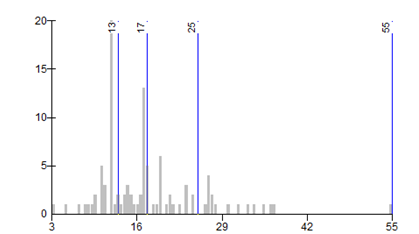
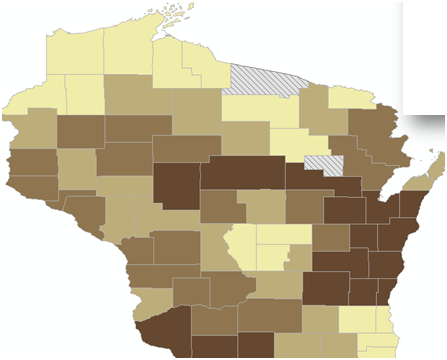
Disadvantages

* Complicated
* Difficult to understand procedure for grouping

## Nested Means

The nested means automatic classification method creates classes about the arithmetic means of the data set. Additional means can be calculated about the first mean to create additional classes.

The way the nested means works is at the first two classes are created 1 above the mean and 1 below the mean. The third and fourth classes are created above and below the means from the first two classes. In this case, it creates a reasonably interesting map, for each class has about the same number of observations.



**Advantages and Disadvantages of using the Nested Means Classification Method**

Advantages

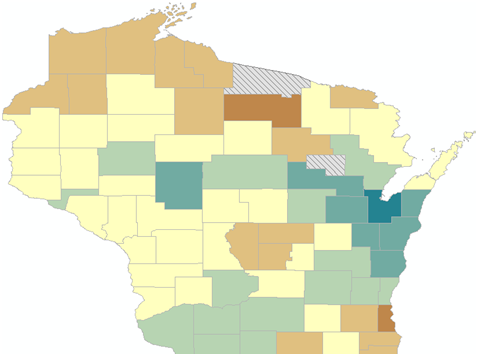
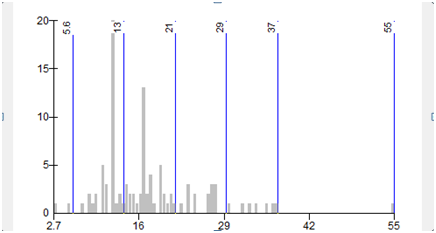
* Easily Computed
* Mathematically intuitive

Disadvantages

* Limited to classes: 2, 4, 8, ...
* Does not consider distribution of data

## Mean and Standard Deviation

The mean and standard deviation automatic classification method creates classes about the arithmetic mean and standard deviations above and below the mean. For this method, since we are looking at divergent behavior, you should use a diverging color. Provided is an example of a diverging color map created using colors tending toward brown as being identified as observations below the mean, and observations towards blue being above the mean. The data for the state has a positive skew, it is not normally distributed; therefore, this method is not ideal for this data set, but does still tend to display interesting geographic patterns.

**Advantages and Disadvantages of using the Mean and Standard Deviation Classification Method**

Advantages

* Good for data with normal distribution
* Considers distribution of data
* Produces constant class intervals

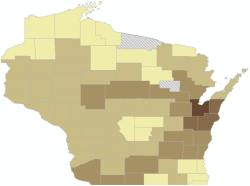
Disadvantages

* Most data are not normally distributed
* Requires understanding of statistics

Not easily understood by map reader

## Equal Interval

The equal interval automatic classification method creates classes with equal ranges. The class range is calculated by taking the maximum value of the data set, subtracting the minimum value from it, and then dividing that by the number of observations in the data set.

Looking at the map we can see interesting patterns are now hidden since the class ranges are including data that is different from each other..

If we look at the histogram of the data, we can see that the class breaks are evenly distributed which is the goal of the equal interval classification method.

**Advantages and Disadvantages of the Equal Interval Classification Method** Advantages

* Easily understood
* Simple to compute
* No gaps in legend

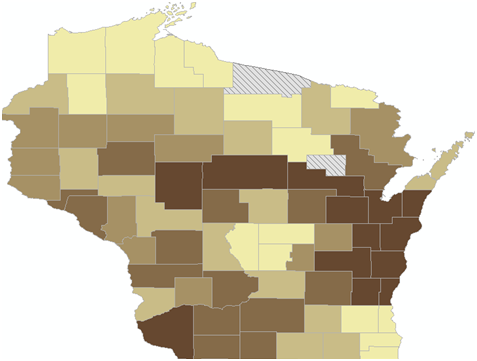
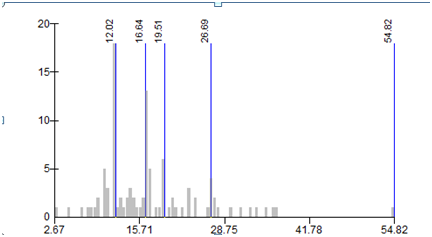
Disadvantages

* Does not consider distribution of data
* May produce class with zero observations

## Equal Frequency

The equal frequency automatic classification method is also known as the quantile classification method. The goal of the equal frequency method is to distribute observations equally among classes, that is, each class will have the same number of observations. If the number of observations does not divide equally into the number of classes, then you should place the extra observations across the lower classes, thereby overloading the lower classes.

As the equal frequency automatic classification method aims to have the same number of observations in each class, there should be about the same amount of each color on the map, which permits a balanced looking map. Again, a major negative, is that a class may contain data values that are dissimilar, which goes against the purpose of classifying data.

**Advantages and Disadvantages of the Equal Frequency Classification Method**

Advantages

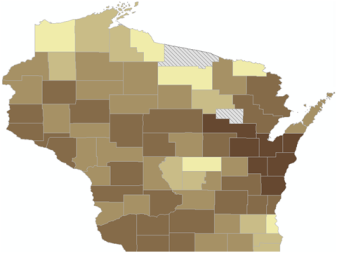
* Easily calculated
* Applicable to ordinal data
* No classes with zero observations

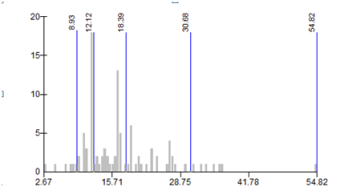
Disadvantages

* Does not consider distribution of data
* Gaps in legend
* Observation value may be closer to value in a different class

## Arithmetic and Geometric Intervals

The arithmetic and geometric intervals automatic classification methods create class boundaries that change systematically with a mathematical progression. This classification method is useful when a range of observations are significant and the observations follow some sort of mathematical progression that can be followed with the classes.

On the map, this classification method creates a reasonably interesting map. Additionally, it does not over fit the data but in this case does seem to over fit the data creating patterns where they do not exist.



If we look at the histogram, you can see that the class ranges slowly increase along the x-axis as this classification method is fitting the data in increasingly larger classes.

**Advantages and Disadvantages of the Arithmetic and Geometric Intervals Classification Method**

Advantages

* Good for data with large ranges
* Break points determined by rate of change in the data set

Disadvantages

* Not appropriate for data with small ranges or linear trends

**Table Operations**

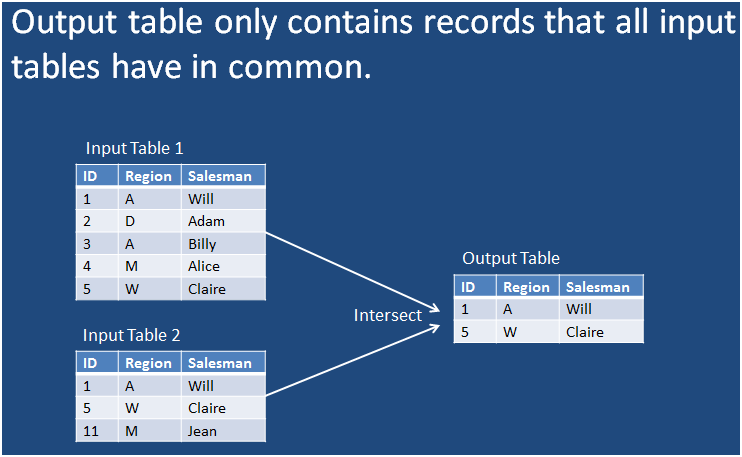
## Common Table Operations

Combining tables of attributes is a powerful analysis technique. There are three common types of table operations: intersect, union, and join. Each one of these operations combines tables in different ways, and for different purposes.

It is important to note that in order for multiple tables to be inputs for intersection or union operations, the tables must have the same fields and field types, which are more commonly referred to as the schema. Let’s take a look at each one of these three table operations in more detail.

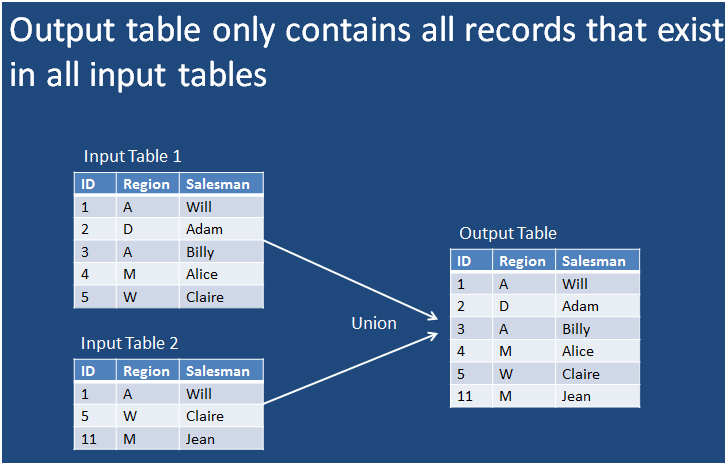
### Intersect

The intersect table operation produces an output table that only contains records that all of the input tables had in common. That means that every attribute for the entire row was identical. As an example, we have two input tables that have the same schema, as all input tables must have the same schema for the intersect operation. The output table only includes the two records that were found in both input tables. In this case, the row where the ID is 1, the region is A, and the salesman is Will, and in the second row the ID is 5, the region is W, and the salesman is Claire.



### Union

The union table operation produces an output table that contains all records that existed in all input tables. For example, we have the same two tables as before, and notice at the output table combines all the rows from both input tables. Again, note that both input tables have the same schema, as this is required for union.

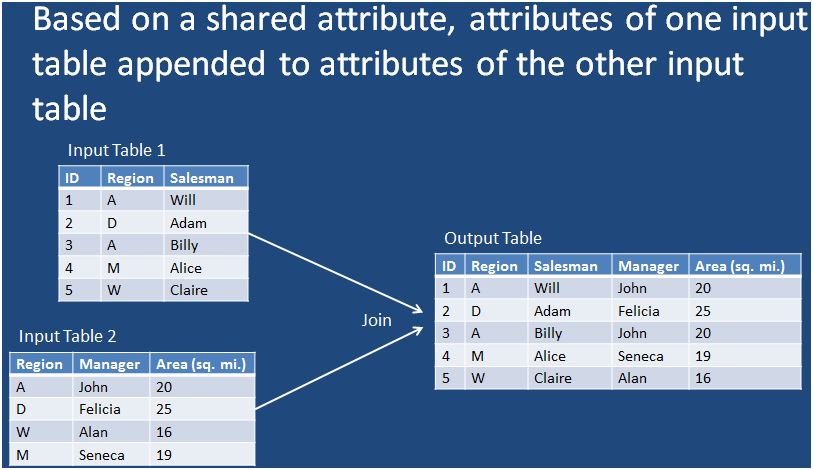


### Join

The join operation joins together two tables based on a shared attribute. Based on the shared attribute, attributes of one input table are appended to the attributes of the other input table. For example, we have two input tables that share the common attribute of region. Based on the region attribute, the manager and area attributes of Input Table 2 are appended to the rows of Input Table 1 in order to create the Output Table.

The region attribute, matches its value between the two tables, which determines which attributes of which row from Input Table 2is copied over. For example, if we look at the Output Table, the row with ID 1 had a region A, which means that the manager is John, and the area is 20 sq. miles. In row where the ID was 2, since the region was the value of D, then the row with the same region of value D in the second input table is copied to the output table.

Also note, that for the record where the ID was 3, the region is A again, so once again, the manager John, and the area 20 are copied to the output table for that row.



**Creating Graphic Statistical Models**

## Graphic Statistical Models

A graphic statistical model is another form of cartography that provides visualization for analyzing data in a form other than a map. Graphic statistical models are excellent at allowing the user to identify trends in the data.

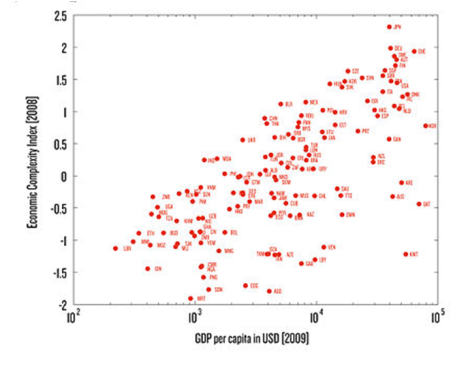
Additionally, graphic statistical models provide an insight into the distribution of the data over the entire data set. There are many different graphic statistical models available for use, but here we will only cover the four most common, which are scatterplots, bar graphs, vertical line graphs, and histograms.

|  |
| --- |
| **Key Facts**  Graphic Statistical Models   * Provide a visualization for analyzing data * Allow the user to identify trends * Provide an insight into the distribution of the data * Scatter plots, bar graphs, vertical line graphs, and histograms all provide this type of visual representation |

### Scatter Plots

A scatter plot is a good method of viewing relationships between variables. A scatter plot allows the reader to identify relationships between two variables where each variable is on its own axis. With both variables on separate axis, relationships are visible by a pattern in the plot which is commonly known as a trend and is traced using trendlines.

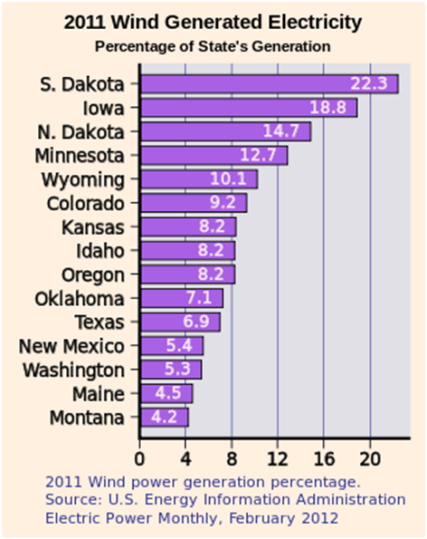
If there is a trend inside the data set, its scatter plot will visually describe it for you. It may also describe no relationship as the data seems randomly scattered.



This is an example of a scatter plot showing the GDP per capita of the United States in the year 2009 and the economic complexity index of the United States in year 2008. As we can see in the scatter plot, as the GDP increases, so does the economic complexity index. This is a good example of positive correlation between the two variables. Also, since there seems to be a trend in this data set, you could draw a trend line starting at the bottom left through the center of the data points to represent the general trend of the data.

### Bar Graph

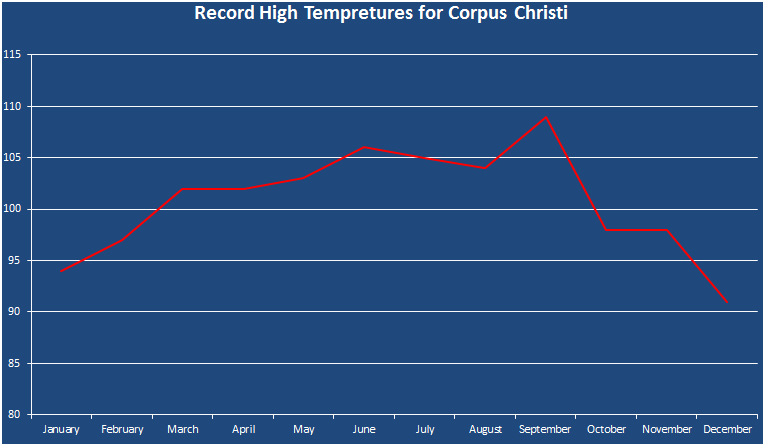
A bar graph shows rectangular bars that are in proportion to the values they represent. This allows for comparison between categories of items in a nice visual form. On a bar graph, one axis shows the categories of the items while the other shows a discrete value for each one of those items.

Here is an example of a bar graph showing the wind generated electricity as a percentage of a state’s energy generation for the year 2011. On the vertical axis we see the name of each item, in this case state names. On the horizontal axis, we see numbers that represent percentages.

For each state, a bar is drawn that is proportional in size to the percentage value of a state’s wind generated. With all of the states’ bars drawn, the bar graph presents the user with a nice visual medium for comparing the percentages between states. It should be noted that while this bar graph is ordered in descending order from top to bottom, this is not required.

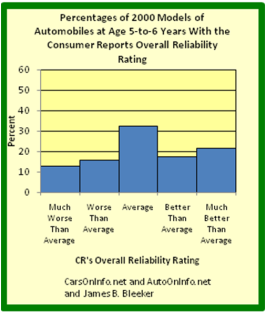
### Vertical Line Graph

A vertical line graph displays information through plotting a series of data points that are then connected with lines. Vertical line graphs are excellent for visualizing trends within a data set and are very simple to create and interpret.

In this vertical line graph, on the vertical axis are temperature values between 80 and 115. On the horizontal axis, the 12 months of the year are represented. For each month, a dot is placed above that month at the location where the temperature was recorded. This is done for all 12 months. Next, a line is drawn between all 12 dots to show continuity.

### Histogram

A histogram is used to plot the density of data. On a histogram, for each range, or value of ranges, a bar is extended for each observation that falls on the value or range of values. A histogram allows the user to see the frequency of observations in relation to specified ranges within a dataset.



This histogram shows the percentages of 2000 Models of Automobiles at Age 5-to-6 Years with the Consumer Reports overall reliability rate. On the vertical axis, we see the percentage value. On the horizontal axis, we see five qualitative measures of reliability. The height of the bars varies depending on how many values fall within each measure of reliability. As can be seen here, the majority of values fell within the average reliability rating.

## SUMMARY

## This lesson introduced you to geospatial analysis. You learned about descriptive statistics including central tendency examples. Standard deviation was explained as were the three types of Kurtosis. The lesson explained how to classify data and the methods used in this process. There were many examples of equal intervals, equal frequency, and arithmetic and geometric intervals. In this lesson you also learned about the importance of table operations and how to create graphic statistical models.

## ASSIGNMENTS

1. Lab: Introduction to Geospatial Analysis

2. Quiz: Introduction to Geospatial Analysis