Lab notes for *GIS and the Digital Earth*
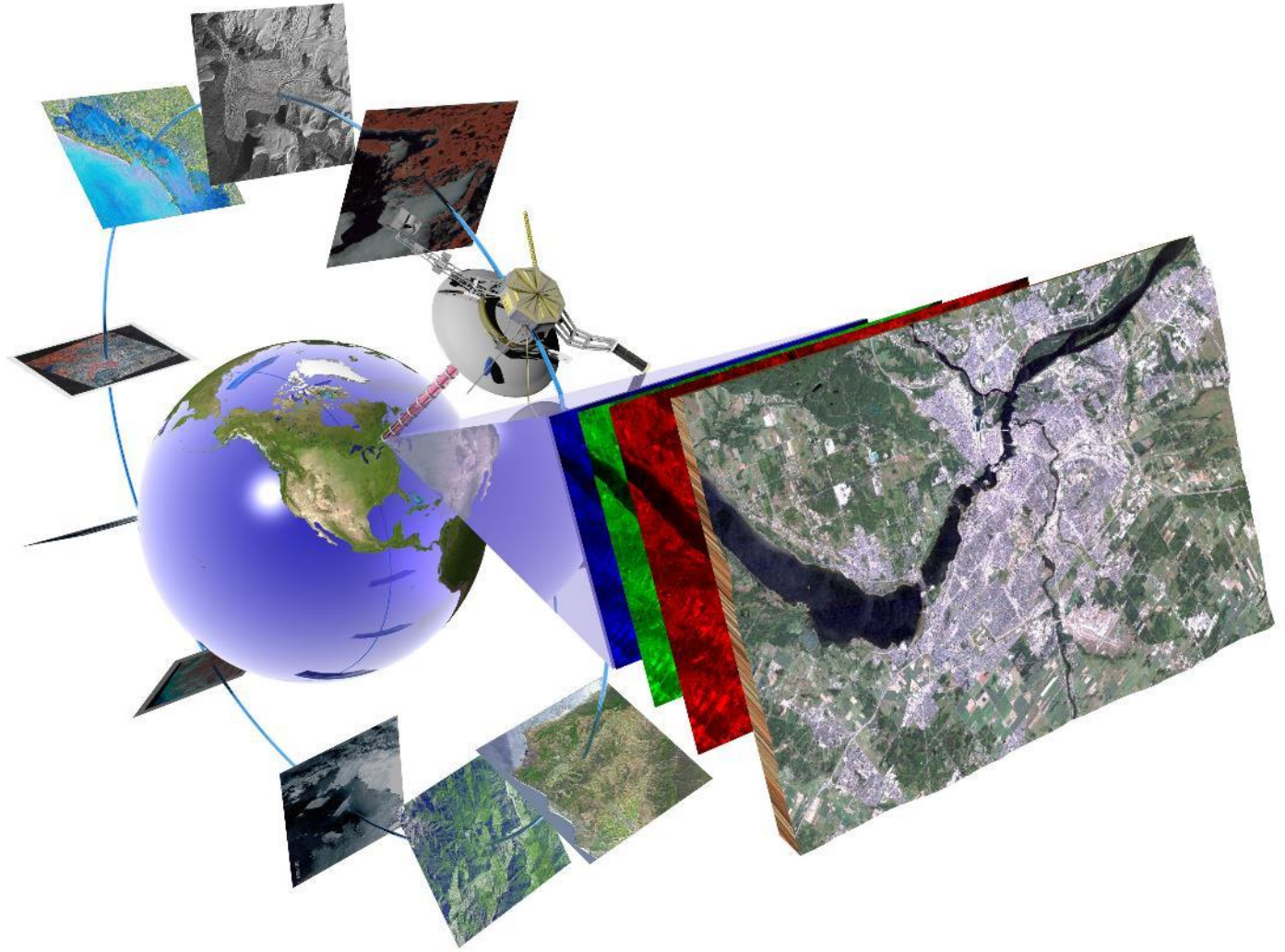
These notes were developed by Dr. Michael Sawada, Dr. Anders Knudby, Galen Richardson, and Claudia Sauro, at the University of Ottawa. They are provided free of charge to the public as an Open Educational Resource under a CC BY 4.0 license.

Student Name: _____ Student Number: _____

Contents

Student Name: _____ Student Number: _____

Introduction

This lab manual is used at the University of Ottawa for the course named GIS and the Digital Earth (course code GEG2320). It may be used freely to teach different courses, and freely modified and distributed, with proper attribution as per the [CC BY 4.0](#) license.

The document contains an introductory section with general guidelines for how to produce good figures, tables, and maps. Students are expected to refer to and abide by these guidelines when handing in their answers to each lab. Following those general guidelines are five separate lab exercises, each of which consists of i) guided steps demonstrating a data analysis performed in ArcGIS, followed by ii) a set of questions to be handed in.

How To Write a Good Report

How to make a good graph

The defaults created by Excel and other spreadsheet packages usually create poor graphs. Compare the two graphs in Figure 1 and Figure 2. In most disciplines, graphs are expected to look like Figure 2. Note that the graphs do not include a Title; the appropriate description of the figure (figure caption) is provided below the graph.
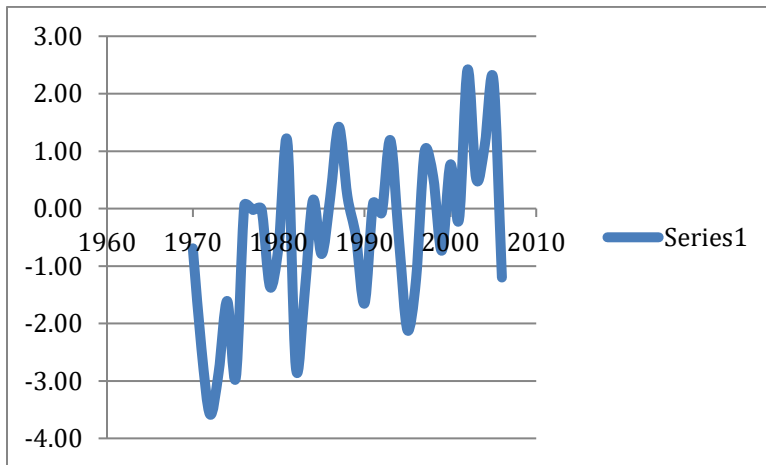
Figure 1. 1970-2006 mean annual air temperature for Whitehorse (Yukon Territory). The graph was made using the default setting in Excel and breaks a number of rules of proper graph-making.
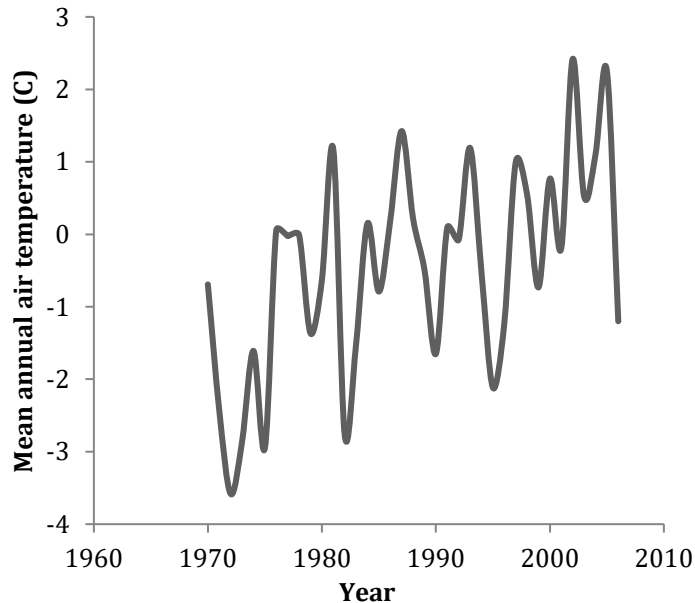
Figure 2. 1970-2006 mean annual air temperature for Whitehorse (Yukon Territory). The graph was adjusted to esthetics: removed grid lines, x-axis intersects y-axis at the minimum value, line is 1pt thick, x-axis and y-axis titles, size of the graph was reduced.

Here are some other examples of figures created in R (a statistic software; you will likely learn this software in more advanced courses) (Figure 3).
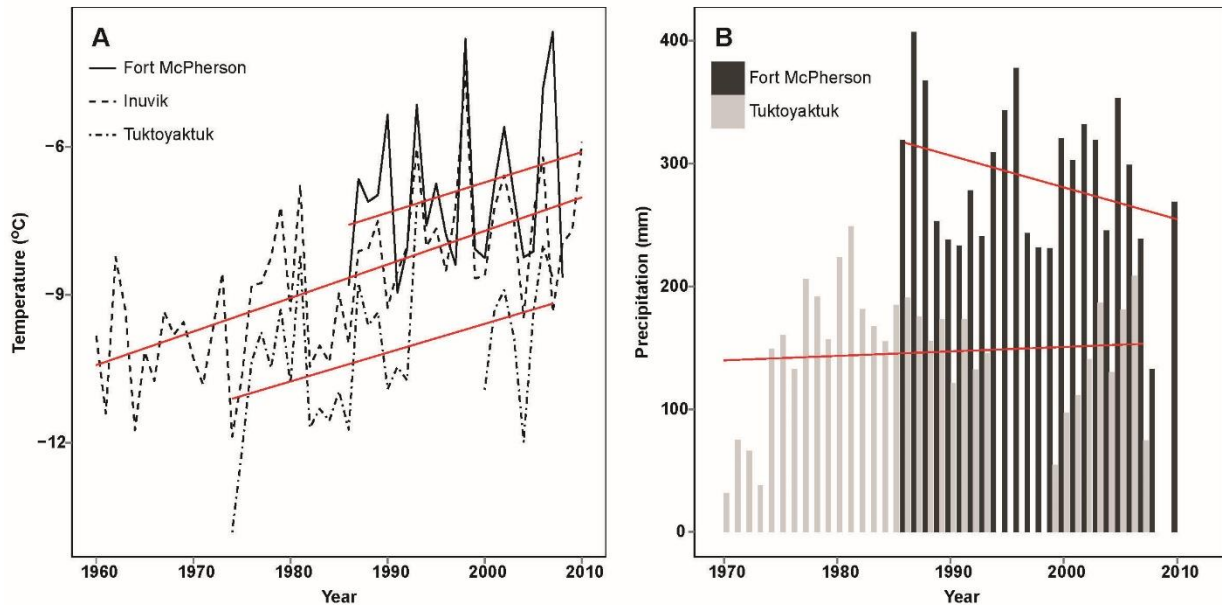
Figure 3. A) 1960-2010 mean annual air temperature for Fort McPherson, Inuvik and Tuktoyaktuk, NWT, Canada. Red lines are linear regression lines for each station. B) 1970-2010 total precipitation for Fort McPherson and Tuktoyaktuk, NWT, Canada.

Tips to make a good graph:
- You can probably use either "line" or "x-y scatter" graphs (these are Excel terms, not in general usage). "Line" graphs assume the points are evenly spaced (i.e., daily or yearly), and have fewer options; "x-y scatter" graphs are best used for unevenly spaced or missing data. If there are missing data, such as year inadvertently left out, there could be a problem with line graphs, for example. Play with the various graphs options and see which works best. In some situations, using line graph when you should use scatter will lead to problems; line or scatter plots are appropriate in different situations. Learn the circumstances under which you would you use either of these.
- Use a figure caption – below the graph or the series of graphs- rather than a title for each graph.
- If you have multiple data series, use a legend, or you can label the lines on the graph.
- Turn off the background colour and remove the grid lines. These clutter up the graph.
- On the axis scales, use appropriate decimal points and spacing. For example, it is typical to label the decade years (1900, 1910, ...) rather than some arbitrary or intermediate years. You should avoid using decimal places on the axis labels, but this depends on the particular dataset.
- Ensure that the font is large enough to read in a printed copy. A rule of thumb - the final size should be a minimum of 6-point text.
- In graphics you should use a "sans-serif" text. **Arial is a sans-serif font. Times New Roman is a serif font**. You would typically use the same font for all graphics. Avoid all

the fancy fonts that may be available. Whereas a serif font may be easier to read in a paragraph, in a figure or map it looks like it was typewritten and less professional. You should also probably stick to standard fonts; sometimes using a rare or unusual font can give problems when you switch computer or software.

- Don't use pie-charts.

How to make a good table

BOX 1 and BOX 2 show two different tables representing the same data. The table of BOX 1 is an example of a table that *does not* meet the standards expected in reports and scientific essays. The table in Box 2 shows the format you are required to use when presenting a table in your work.

BOX 1

Rank	Country of study	Population in each country	Percentage of world population
1	China	1,371,380,000	18.94289
2	India	1,275,260,000	17.6
3	United States	321,564,000	4.431
4	Indonesia	255,770,000	3.5223
5	Brazil	204,714,000	2.82

This table shows the number of people living in China, India, United States and Indonesia which are the most populated countries. The data for Brazil is based on voluntary participation of its citizen to the annual census.

BOX 2

Table 1. Top 5 most populated countries in the world and relative percentage to total world population.

Rank	Country	Population	% of world population
1	China	1,371,380,000	18.9
2	India	1,275,260,000	17.6
3	United States	321,564,000	4.4
4	Indonesia	255,770,000	3.5
5	Brazil	204,714,000	2.8

Note. Data for Brazil is based on voluntary participation to annual census.
Source: Office of International Statistics

The next two tables are examples of variations to the basic table. Table 3 shows you how to divide one column into two different columns with the use of horizontal lines. Table 4 shows how you can delineate rows by using a light shade of grey. You will revert to shading only when your table contains many columns and rows making it difficult to read across.

Table 3. Altitudinal distribution of endemic and indigenous flowers and trees of Stony Hill, Quebec, Canada.

Altitude (m)	Flowers		Trees	
	Endemic	Indigenous	Endemic	Indigenous
200	0	32	0	11
400	0	30	0	8
600	1	28	0	7
800	2	20	1	8
1000	1	21	0	3
1200	4	15	2	3

Table 4. Estimated bycatch of fishing trawlers. Data is given in tons and represents the total per month for the 2010 to 2014 period.

	Ayu	Brill	Dory	Flier	Hake	Jack	Ling	Porgy	Rudd	Sole	Walu
Jan.	12	8	45	56	85	94	82	96	98	32	36
Feb.	56	32	56	11	18	62	66	35	33	39	95
Mar.	98	11	47	89	85	8	56	6	5	12	39
Apr.	5	8	74	21	11	9	85	23	25	96	36
May	6	12	62	35	33	39	21	95	6	17	46
June	11	18	54	9	7	32	23	12	95	21	63
July	3	68	14	21	8	15	45	65	61	77	11
Aug.	98	45	56	56	11	18	86	62	35	33	39
Sept.	32	0	43	57	39	28	47	66	72	80	11
Oct.	64	0	11	18	62	35	12	33	39	95	6
Nov.	26	34	33	78	95	23	5	12	8	9	2
Dec.	12	48	26	54	85	8	2	6	5	12	39

Tips to make a good table:

- The table caption (description of the table) is always above the table, it is italicized, aligned to the left and begins with Table 1.
- The table caption is clear, concise, and informative.
- Column headings are clear, concise, and informative.
- Symbols, signs, abbreviations, or acronyms are self-explanatory or explained in a note.
- The number of decimals is consistent.
- Separating lines are absent between the columns, sufficient space is set between them.
- There are no unnecessary bold, italic or underline. These can be used to highlight individual values but then you need a note of explanation.
- Horizontal lines are kept to a minimum; they are used to separate the column headers and at the end of the table. Do not use vertical lines.
- The font is legible and consistent (always use a sans-serif font); never use smaller than 6-point text.

- Additional information, if necessary, is provided in a note located below the table and in a smaller font.
- The source of the data is stated below the table.
- The table is intelligible on its own.
- Tables are to be located closest to where they are first mentioned in the text.

How to make a good map

Figure 4 and Figure 5 show the same data presented as a thematic choropleth map. The purpose of the map is to show the population distribution in Scotland. Figure 4 is incorrect and *does not* meet the standards expected in reports and scientific essays. Figure 5 meets the minimum standards for a map.
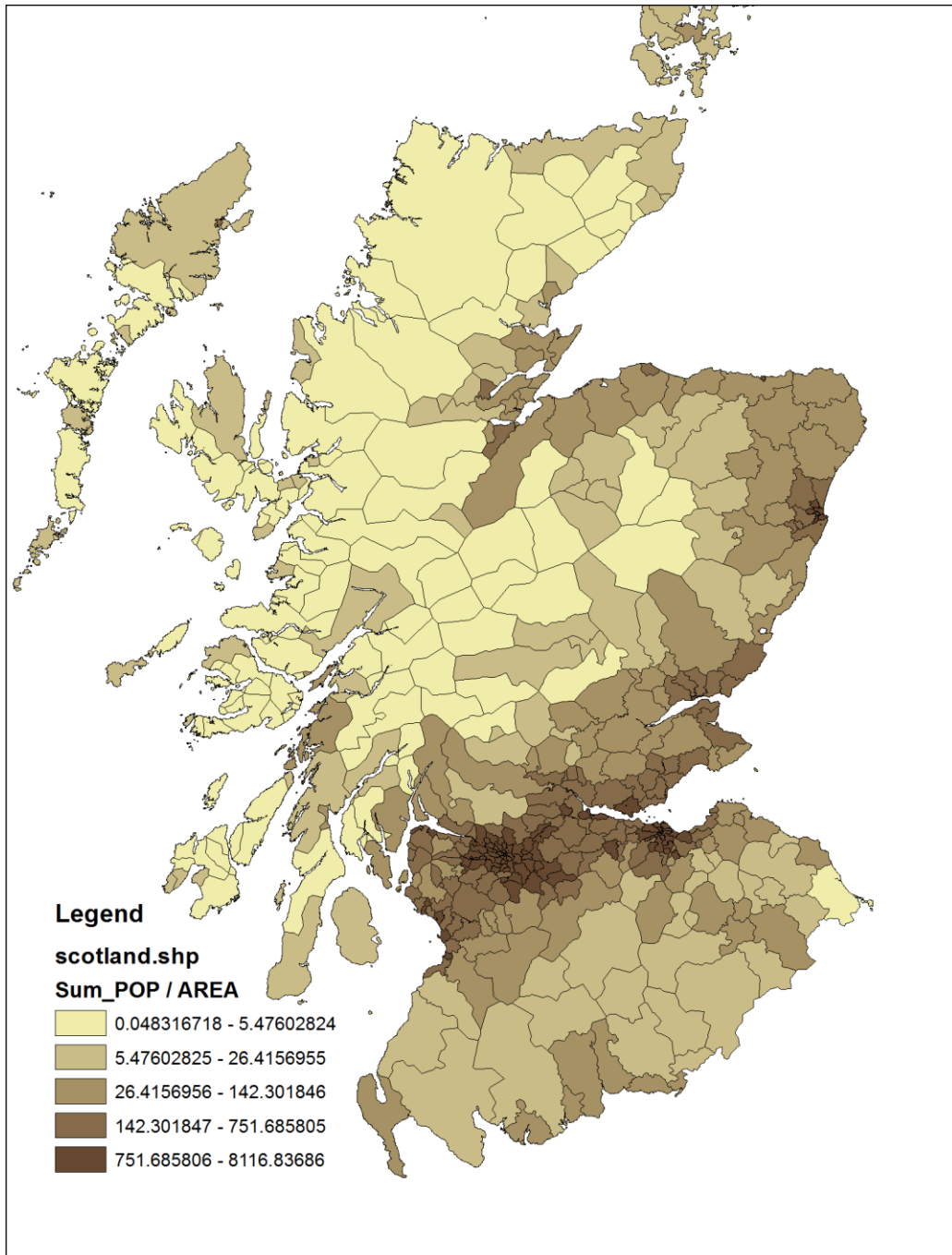
Figure 4. Population density in Scotland.

Figure 5. Scotland.
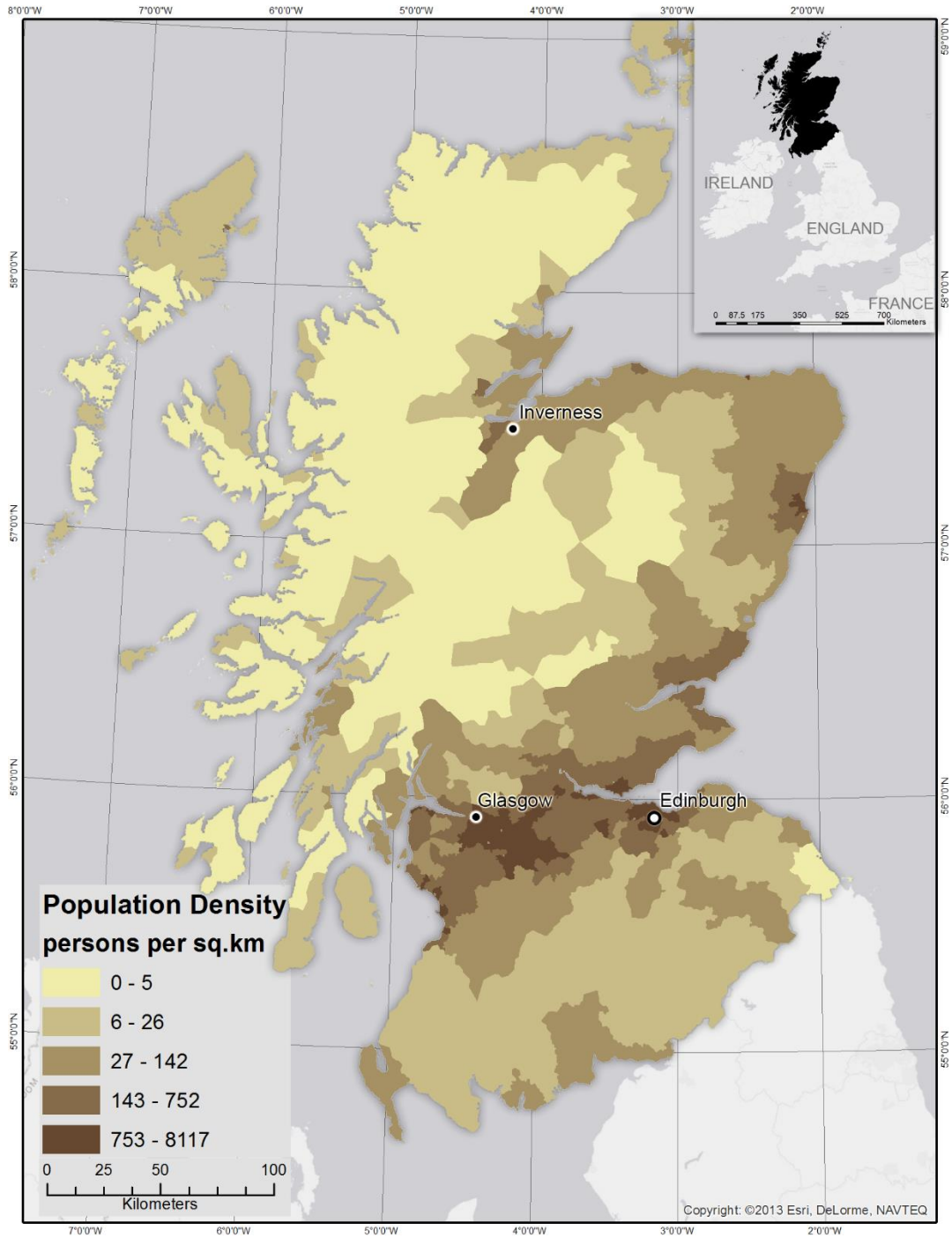
Figure 4 is an unacceptable map for numerous reasons:

- The map in Figure 4 uses polygon outlines when they add nothing to the intended purpose of the map which is to show population density – polygon outlines are a default setting in most GIS systems. Why are the polygons outlined? Is the map about administrative divisions in Scotland or is it about the distribution of population in

Scotland? If the data unit boundaries are not necessary to deliver the thematic message they should **never** be included as seen in Figure 5.

- There is no geographic context on Figure 4. If someone is not already familiar with Scotland, they will have no idea as to where geographically Scotland is found? Including in inlay map as shown in Figure 5 remedies that issue and also includes homogenously colored surrounding countries.

- In Figure 4, how can someone not familiar with Scotland explain where the population is high/low or why it may be high/low in certain regions. A few major cities included and labeled on the map would help this situation as has been done in Figure 5.

- There is no reference to North on Figure 4. If a map IS NOT showing an entire continent, then a north reference is required. Referencing the direction of north can be done using either a north arrow or, as in Figure 5, a graticule instead.

- Figure 4 includes no way to determine the coordinates of any location on the map. Including a labelled graticule as seen in Figure 5 allows for more precise positioning.

- Figure 4 includes no way for a person to ascertain the relative size of features. Figure 5 on the other hand, has a clear scale-bar.

- The legend in Figure 4 is completely incorrect. First, it is obvious that the thing in the lower left of the map is a legend or sometimes called a map key. It does not need a title that says "Legend". Moreover, what is Scotland.shp in the legend. What is Sum_POP / AREA and what are the numbers, what are the units? The labelling of the legend is incorrect. The numbers in the legend lead the reader to believe that whatever is mapped is accurate to 9 decimal places. There are few things that are accurate to 9 decimal places. Compare with the legend in Figure 5. A legend should be clear as to the variable being represented, the units of measure must be clearly labelled and present data at a reasonable level of accuracy.

- The caption for Figure 4 contains redundant information. A caption should never repeat either a map's title or content. A map legend should be sufficient so that someone reading the map knows what the map presents. In Figure 5, you can see that the map caption simply says Scotland because that is the only information needed for someone to read the map, it already has a regional context, and the legend clearly shows what the map is depicting.

There are a lot more rules in cartography but the above comparison should help you in understanding map composition at the most basic level. In general, the following rules should be applied to all maps that you produce:

- Symbolize your map according to the purpose of the map. Any excess symbols should be eliminated.

- Every map must have a title or caption, the latter being preferred, an indication of north as either an arrow or graticule, a properly composed legend, axes labels of x,y or longitude, latitude, and a scale bar.

- On thematic maps include a minimum number of contextual symbols that the reader can use to anchor map descriptions.

- Choropleth maps are only used for non-dimensional variables or density variables. They are never used for presenting count variables.
- Use proportional symbols to present count-variables on maps.

How to reference published materials in your work

In many disciplines, the usual way of referencing materials that you have consulted is the author-date method or APA format. Footnotes are not usually used (although they are common in Human Geography) and quotes are avoided. At the end of each assignment, you must reference all materials (except course or lecture notes) that you have used.

Each academic journal has its own particular format for referencing published works. In this course, you must follow the APA format: references within the text should be given in parentheses with one space separating the author(s) and the date: e.g. Slopewash processes have not been studied extensively (Clark 1988) but some work has been carried out on Banks Island (Lewkowicz 1983). Multiple authors are referred to as (Ex and Why 1986) or if more than two authors as (Ex *et al.* 1986). Please pay attention to these conventions! Nothing upsets an author more than to be incorrectly cited (e.g. if the authors are Monroe and Wicander (1993), don't cite the book as Wicander and Monroe (1993) or even worse Monroe (1993).

At the end of your work, there should be a "References" section. In this section, references are listed alphabetically by author. If two or more references are by the same author, they are listed by date from oldest to youngest. You should not cite material that you have not consulted (i.e. don't read a text that makes reference to another work and then cite the latter as if you've read it). If it is crucial that you cite the second work and you cannot obtain the original, it is permissible to cite it as Cruden (1974) in French (1976)).

The format for a journal article is as follows:
Mackay, J.R. 1971. The origin of massive icy beds in permafrost, western arctic coast. Canadian Journal of Earth Sciences, **8**: 397-422.

Format for a book:
French, H.M. 1996. The periglacial environment. 2nd Edition. Longman, London.

Format for a chapter in a book:
Van Vliet-Lanöe, B. 1985. Frost effects in soils. *In* Soils and Quaternary landscape evolution. *Edited by* J. Boardman. Wiley, Chichester, pp. 117-158.

Format for an internet site:
Give the web-site address and the date that you consulted it as follows: www.nrc.ca/cgi-bin/cisti/ October 15, 2005.

Student Name: _____ Student Number: _____

If the work you wish to cite does not fall into any of these categories, you can check on the appropriate format by consulting a recent volume of the Canadian Journal of Earth Sciences (http://www.nrcresearchpress.com/journal/cjes) or the APA style of referencing.

Student Name: _____ Student Number: _____

Laboratory #1: Cartographic Design, Symbology, and the Modifiable Areal Unit Problem

Purpose: This lab will introduce you to map layouts and guide you through how to make publishable maps and understand ways to represent data through maps. The exercises will reinforce these concepts through the creation of four map layouts that display cartographic best practices.

Learning Outcomes: By the end of this lab, students will be able to:
- Create map layouts that adhere to proper map visual hierarchy and use map elements appropriately.
- Label point and polygon features using different label placement strategies and labelling equations.
- Understand the different types of breaks and primary symbology for polygon features.
- Explain the modifiable areal unit problem.

Procedure:
- Read the example questions in Section 1 and Section 2.
- Repeat the example questions in ArcGIS Pro.
- Address the Assessment questions.

Data/Material:
- Links to the city of Surrey Open Data Catalog

Supporting Documentation:
- ArcGIS Help
- Data source: https://data.surrey.ca/
- Reading: Slocum, T.A., McMaster, R.B., Kessler, F.C., & Howard, H.H. (2022). Thematic Cartography and Geovisualization, Fourth Edition (4th ed.). CRC Press. https://doi.org/10.1201/9781003150527
- Reading: Fotheringham, A. S., & Wong, D. W. S. (1991). The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. Environment and Planning A: Economy and Space, 23(7), 1025–1044. https://doi.org/10.1068/a231025
- Statistics Canada glossary: https://www150.statcan.gc.ca/n1/pub/92-195-x/2021001/geo/ct-sr/ct-sr-eng.htm

Assignment:
- Answer the questions (#1-6)
- The total marks available for this assignment is XX marks representing XX% of your final grade.
- All lab assignments must be completed and handed in individually.

Format:
Students will hand in questions with completed answers.

Tables and figures should be produced to the standards of the Department of Geography, Environment and Geomatics, University of Ottawa.
For guidelines on proper formatting and presentation of tables and figures, please see the introductory section of this lab manual.

Section 1: Cartographic Design

Cartographic design is an iterative process where a cartographer conceptualizes and generates maps according to the needs of the target audience (Slocum, 2009). The primary steps in the cartographic process are:

1. Determining the purpose of the map,
2. Collecting data that can be used to portray the map's message,
3. Designing and creating a map layout,
4. Determining whether the map layout effectively displays the map's purpose and improves the map.

The map layouts that cartographers create are impactful in nature and can clearly explain a geographical phenomenon through maps and map elements.

One of the primary forces that guide cartographic decision making is the concept of visual hierarchy. Visual hierarchy is the principle of arranging elements to emphasize more important map features by adjusting the visual contrast, size, location, and balance of the map elements. An easier way to think of it when evaluating a map is to think "what are my eyes drawn to?". Figure 1. shows a map (left), and a version with a ranking of visual importance attributed to the map features (right).
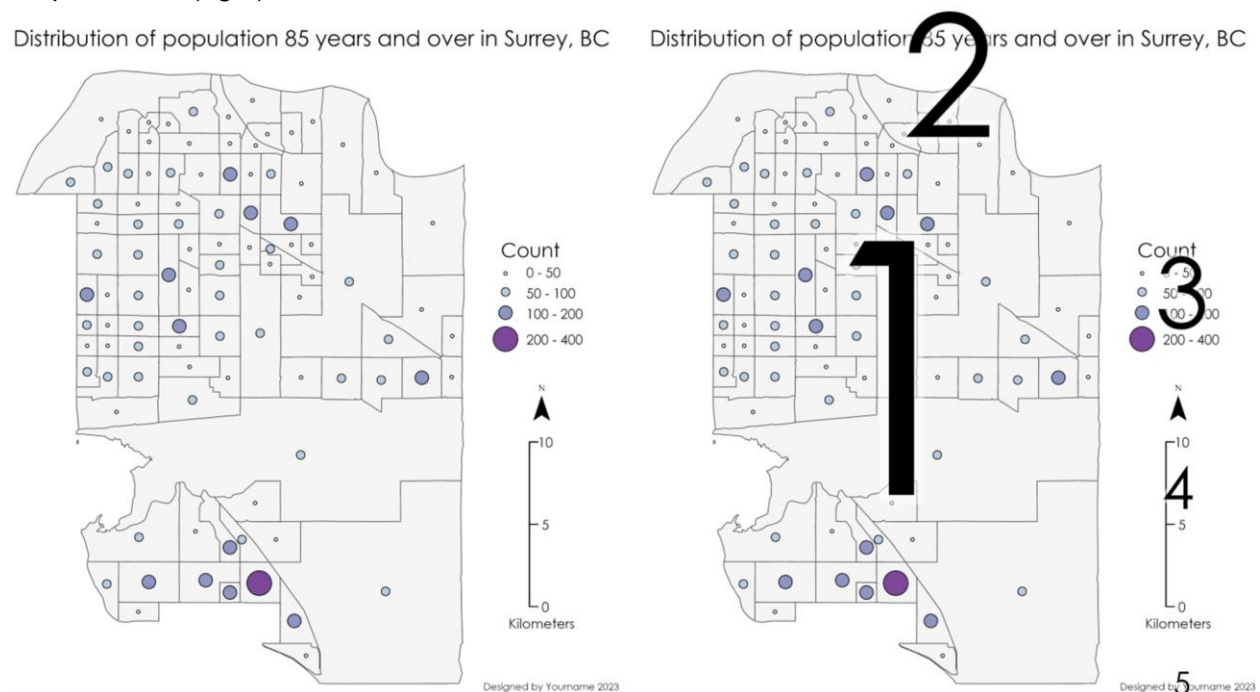

Figure 1. Two maps of Surrey, BC were created to illustrate visual hierarchy (1- highest to 5 lowest). Always remember that the geographic data is the primary object and should be the largest object taking up most of the space on a map.

The map of Surrey, BC encompasses most of the space in this layout and is the first thing that catches the viewer's eyes. The location and relatively large font size of the title make it the second most visually important thing in this layout. The size of the legend is slightly larger than

the north arrow and scale bar, while the least visually important feature is the name and date located in the bottom right corner. Proper visual hierarchy leads the reader from the map to the title and then to the other map elements. The best way to achieve this is with the effective use of size and visual contrast.

Before creating maps for the exercise and assignment, consider these tips:
1) Make a sketch before you start making the map layout.
2) Consider the relative importance of each map element when deciding the formatting.
3) Fill blank space effectively with map elements.
4) Export a draft image of your map and look at it outside of ArcGIS; this can help you notice formatting that should be addressed.
5) There are many sources for help outside of this class (ArcGIS Help, Google) that can assist you when it comes to solving cartographic problems you might encounter.

In addition, the darker an object or part of an object the more visual weight it will have. So be careful and consider that when using map surrounds.

Q1: What is the purpose of the maps created in this lab and how do I access the data?

The purpose of the maps in this lab will be to practice creating map layouts that adhere to cartographic best principals and display vector information about the city of Surrey, BC. The data used in this lab is accessible from the city of Surrey Open Data. https://data.surrey.ca/ The datasets that are required for this lab from this website include:

- Heritage sites https://data.surrey.ca/dataset/heritage-sites
- Important Trees https://data.surrey.ca/dataset/important-trees
- Bike routes https://data.surrey.ca/dataset/bike-routes
- Ecosystem sites https://data.surrey.ca/dataset/ecosystem-sites
- 2016 Surrey Census
  o Download: Census tract level and dissemination area level.

On the Surrey Open Data website, visit the web page and download each file in file geodatabase format.

Data and Resources

| CSV | **Important Trees**
Available in CSV format | | ↗ Explore ▾ |

| JSON | **Important Trees**
Available in JSON format | | ↗ Explore ▾ |

| KML | **Important Trees**
Available in KML format | | ↗ Explore ▾ |

| FGDB | **Important Trees**
Available in File Geodatabase format | | ↗ Explore ▾ |

⊕ Download
↗ Download via Map

| API | **Important Trees** 🔥
Specification at: http://resources.arcgis.com/en/help/rest/apiref/ |

Once you have downloaded all the data (or download the single data file from the lab repository), unzip the zip file and copy all the folders that end in. gdb into a single folder named "Lab Input Data". Make sure that the folder "Lab Input Data" is saved in the same folder as your assignment, so it is easy to find.

Lab Input Data

Name ^	Date modified	Type
bike_routes.gdb	2023-03-13 2:17 PM	File folder
Census_2016_Surrey_DisseminationAreas_UTM.gdb	2018-03-28 11:18 AM	File folder
Census2016_Surrey_CensusTract_UTM.gdb	2018-03-28 11:18 AM	File folder
ecosystem_sites.gdb	2023-03-13 2:17 PM	File folder
heritage_sites.gdb	2023-03-13 2:17 PM	File folder
important_trees.gdb	2023-03-13 2:15 PM	File folder

Create a New Project

Name: Carto_assignment

Location: C:\Users\student\Documents\ArcGIS\Carto_Assignment

☑ Create a new folder for this project

OK

To start working on this lab, create a new ArcGIS Pro Project. The first thing you will want to do is open the catalog pane where you can access your data. On the top menu bar, click view, then catalog pane. A new menu should pop up on the left side of your screen called catalog. Click the 'folders' icon in your catalog and you will see an existing home folder with a geodatabase inside. By default, all outputs from ArcGIS Pro tools you run will be saved to this geodatabase and thus it is called the "Default Geodatabase". Right click the 'folders' icon in the catalog menu to create a folder connection to the *Lab Input Data* folder. Once you have selected the folder containing your data (*Lab Input Data*), click ok and it will be visible in the folders section of the catalog menu. If you are encountering issues with fold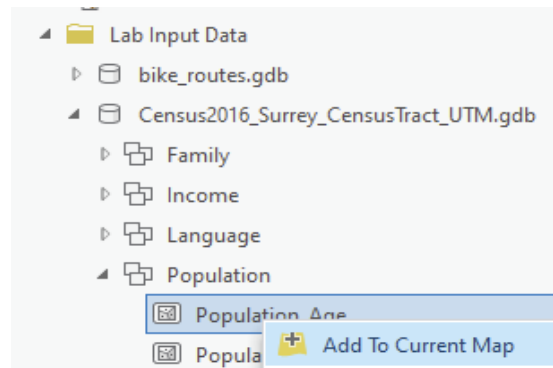er connections, please refer to this website: https://pro.arcgis.com/en/pro-app/latest/help/projects/connect-to-a-folder.htm

▲ 📁 Fol...
▷ 📁 S...
▷ 🏠 Loc...

ᚱ⁺ Add Folder Connection
🔁 Refresh
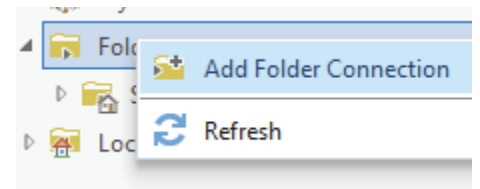
Tip: If you see your folder in Windows Explorer, you can drag and drop the folder into the catalog pane instead of using the Add Folder Connection to create a data connection.

Q2: How do I create a map and symbolize a feature class?

For Questions 2 and 3 we are going to create a map, apply basic symbology, and then create a map layout. In the Lab Input Data folder in your catalog pane, open the geodatabase called Census2016_Surrey_CensusTract_UTM.gdb, then open the population feature dataset, and right click Population_Age and click "Add to Current map". The map should zoom into the extent of the recently added feature class. In the contents pane to the right of the map, under drawing order, uncheck the base maps World Topographic Map and World hillshade. Then right click on population_age and click "Symbology". The symbology tab should pop up to the right of your map. To change the way Population_age is displayed in your map, under Primary symbology, click the dropdown menu (should be on 'single symbol' by default) to select the symbology you would like to use for this feature class. If you are encountering issues with symbolizing a feature class, please refer to this website: https://pro.arcgis.com/en/pro-app/latest/help/mapping/layer-properties/symbolize-feature-layers.htm

Q3: What are the common ways I can symbolize polygon feature class?

The first question you should ask yourself when symbolizing data in ArcGIS Pro is "What kind of data am I symbolizing?". Categorical data (names and places) should be symbolized using unique values, wherein one unique color is given to each unique category. For quantitative data (population and percentages), there are a few different ways to symbolize the data, such as Choropleth (graduated colors), graduated symbols, dot density, and proportional symbols (Figure 2).

Graduated Symbols

Population density per square kilometer

- Less than 1,000
- 1,000 - 2,500
- 2,500 - 5,000
- 5,000 - 10,000
- Over 10,000

Dot Density

Population in Surrey

100 People

Choropleth

Population density per square kilometer

- Less than 1,000
- 1,000 - 2,500
- 2,500 - 5,000
- 5,000 - 10,000
- Over 10,000

Bivariate Choropleth

Proportion of Dwellings to Population
Population Density

High

Low

Low   High

Figure 2. Comparison of different types of quantitative symbolization in Surrey, BC

The most popular method for symbolizing a polygon feature class is to use a Choropleth (graduated colors) Symbology. While Choropleth maps are easy to make and interpret, they also have several weaknesses that you should be aware of:

1. By definition, a choropleth map assumes that the data is uniformly/evenly distributed across each polygon, however, that is rarely the case. For example, in the population maps shown above, it would be unreasonable to assume the population is evenly distributed across each polygon; instead, people tend to live in residential neighborhoods, while other areas within each polygon are used mainly for

transportation, shopping, recreation etc. and thus have lower population density. Examine Figure 3:
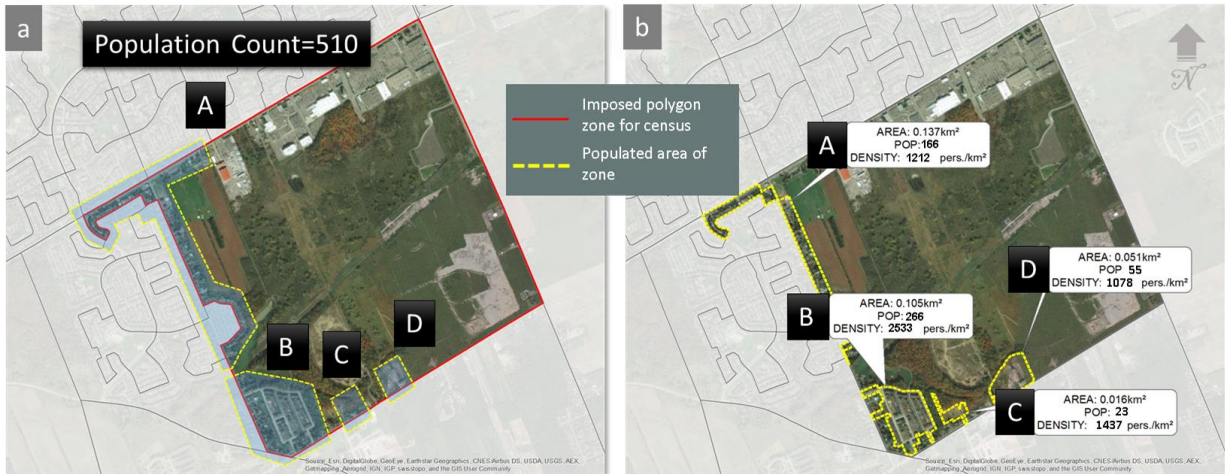
Figure 3: a) Single dissemination area zone (red outline) with a total population of 510 people. The populated part of the zone is shown in yellow and labeled A-D. Everywhere else in the zone is commercial, farmland or industrial where the population cannot reside. b) Population of 510 people disaggregated to the populated parts of the zone proportional to the area of each populated part. Notice that the assumption of homogeneity within the sub-zone populated areas is still true as the population density remains constant.

In the real world this entire polygon is allotted 510 people (Figure 3a), whereas in reality, people only inhabit the areas surrounded by the yellow dotted lines. Moreover, those little areas are not even geographically connected to each other, and each has its own unique population count and density. By now you are wondering where to stop. If we were to zoom into one of the yellow areas, we would also see more areas where people cannot live and then exclude those from population density computation to get even more accurate population densities. Should we then exclude the garage or parking spaces since people can't live there? What about backyards? Front-yards? One could go on and on, *ad infinitum*, *ad nauseam*. In the end, no map will correspond with reality because reality is infinitely complex. Here, at most we can understand that we can better approximate reality if we understand that no quantity is naturally distributed uniformly across space unless - you can think of one?

2. To produce meaningful comparisons across zones in a choropleth, the data that is going to be mapped should be normalized, such as by mapping population density instead of the total population in each zone. This can be done by plotting the data as normalized by using the normalization drop-down menu in the symbology pane. Graduated symbols, Proportional symbols and dot density maps do not require standardized data and can be used to represent count data such as population.

For your first map, select the primary symbology to be Graduated Colors, field "Average age", normalization "<None>", Method "Manual Interval", and classes "5". Set the upper

values to 35, 40, 45, 50, and 55 and update the labels. Use this map to practice cartographic elements in questions 5,6,7.

Primary symbology		
Graduated Colors		
Field	Average age	
Normalization	<None>	
Method	Manual Interval	
Classes	5	
Color scheme		

Classes Histogram Scales

Symbol	Upper value	Label
	≤ 35	30 - 35
	≤ 40	35 - 40
	≤ 45	40 - 45
	≤ 50	45 - 50
	≤ 55	50 - 55
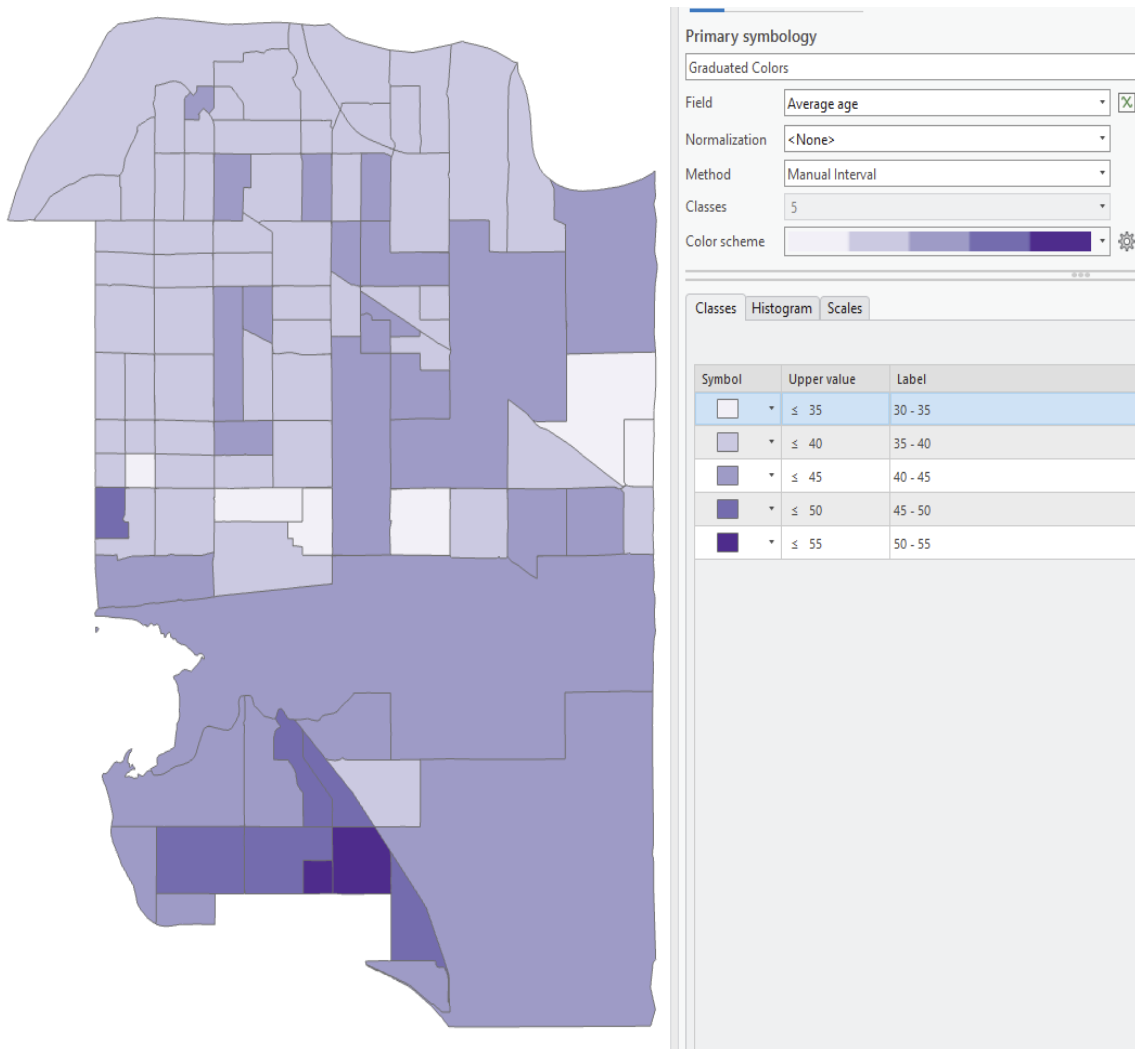
Q4: What are breaks and how do they affect the way my map looks?

Class breaks (referred to as Method in the ArcGIS symbology tab) allow you to visualize quantitative ranges of data. Class breaks methods include Natural Breaks (Jenks), Equal Interval, Manual Breaks, Standard deviation, Quantile, and Geometrical interval. The map layout below (Figure 4) shows four of the most common class break methods. More information on class breaks methods can be found here: https://pro.arcgis.com/en/pro-app/latest/help/mapping/layer-properties/data-classification-methods.htm

Natural Breaks (Jenks)

Population density per square kilometer

☐ 53 - 1,406
☐ 1,406 - 3,003
☐ 3,003 - 4,719
☐ 4,719 - 7,391
☐ 7,391 - 11,291

N

5
2.5
0
Kilometers

Equal Interval Breaks

Population density per square kilometer

☐ 53 - 2,301
☐ 2,301 - 4,548
☐ 4,548 - 6,796
☐ 6,796 - 9,043
☐ 9,043 - 11,291

N

5
2.5
0
Kilometers

Manual Breaks

Population density per square kilometer

☐ Less than 1,000
☐ 1,000 - 2,500
☐ 2,500 - 5,000
☐ 5,000 - 10,000
☐ Over 10,000

N

5
2.5
0
Kilometers

Standard Deviation Breaks

Population density per square kilometer

☐ < -1.5 Std. Dev.
☐ -1.5 - -0.50 Std. Dev.
☐ -0.50 - 0.50 Std. Dev.
☐ 0.50 - 1.5 Std. Dev.
☐ 1.5 - 2.5 Std. Dev.
☐ > 2.5 Std. Dev.

N

5
2.5
0
Kilometers

Figure 4. The most common methods used for class breaks visualized using Surrey BC population data.

Determining the appropriate class break method is a cartographic decision that can impact the message that the map is giving. Class breaks also partially determine the type of color scale that can be used and that in turn is ultimately a function of the kind of data being mapped. A guiding rule is that manual breaks should be used most of the time if you are describing your data with numeric values, such as population density per square kilometer. It is much easier for readers to understand and visualize numbers such as 1,000 people per square kilometer, rather than 1,461. If you are describing your data in categories such as low, medium, and high, Natural breaks or equal interval breaks might be more appropriate. You can also adjust what is displayed on the map next to each break by altering the label column in the symbology pane

and change the color scheme by either changing the color of each break, or by selecting a color scheme from the dropdown menu.

Classes	Histogram		
Symbol	**Upper value** ▲	**Label**	
○	≤ 50	0 - 50	
○	≤ 100	50 - 100	
●	≤ 200	100 - 200	
●	≤ 400	200 - 400	

Q5: How do I create a map layout and set the map scale?

The first step to creating a map layout is to visualize where you want to put the map and map elements. It is recommended to make a quick sketch on a piece of paper or spend a bit of time thinking about where you want to place the map and map elements (Title, legend, north arrow, scale bar). The next step is to decide on a page size and orientation that fits with your map sketch. If your map takes up an entire page, it is recommended that you select a letter page size. However, if you are making a map for digital purposes, or are generating a map that will be part of a report, it is recommended that you set up a custom page size. Squares (20cm x 20cm) are a common page size to allow for easy division of the page to make cartographic layouts. The most effective page layouts are created when cartographers look at the map, they are trying to make along with the map elements, and then choose a suitable page size. For example, since Surrey, BC has more height than width, your choice in page size should reflect this. Note that you can always change your page size after you have started working on your cartographic layout.

To create a layout on the insert tab in the Project group you just click on the New Layout button. After you have started a layout in ArcGIS Pro with a blank page, click insert, then map frame, and select the Map you would like to insert into your layout. Consider the map frame as the outline for where your map will fit on the page. Draw a map frame and then right click on it and click activate. Once you have activated the map frame, you are able to move the map around the map frame and zoom in. The map scale (with the red line underneath it in the bottom left of the image) can be manually set to zoom in or out by larger or smaller ratios respectively. Once you have placed your map appropriately, click on the "layout" button in the top menu bar and click "close activation". This will prevent further modification to your map within the map frame. Additional tips and resources for setting up a map frame and how to remove the map frame border can be found here: https://pro.arcgis.com/en/pro-app/latest/help/layouts/add-and-modify-map-frames.htm
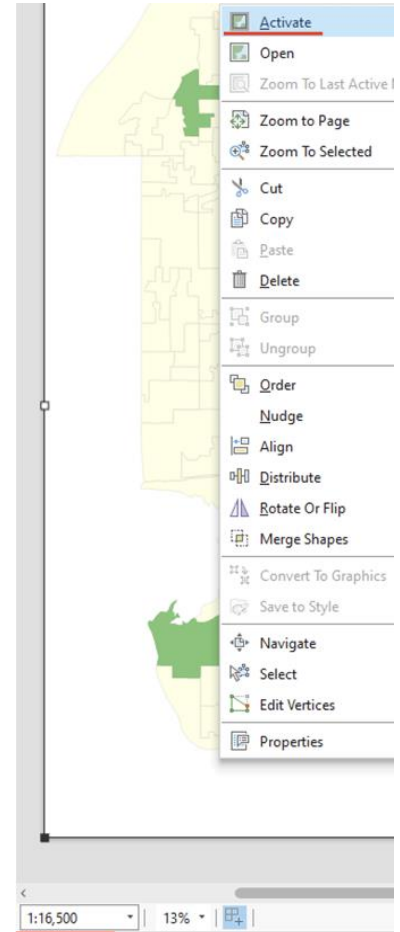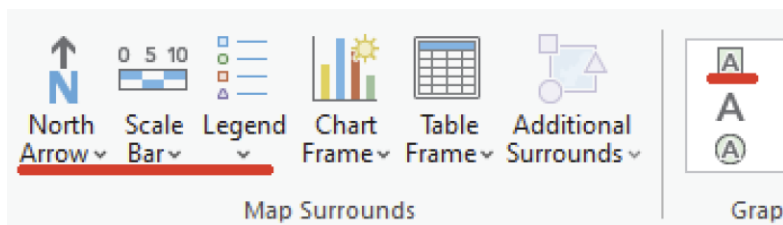
After placing your map, go to the upper menu ribbon, find the Map Surrounds group, and select an appropriate north arrow, scale bar, and legend. Also, in the Graphics and text sub-menu in the upper menu ribbon, select rectangle text and draw a box for the maps title. Rectangle text is better than drawing a straight text box, because it gives you more options if you want to add a background to your textbox. Keep in mind that you don't put a title on a map that is not a stand-alone map, that is to say, if you have a question that asks for a map only then the map should have a title, if you have a question that asks for a map and interpretation of some maps then you omit the title and instead use a numbered figure caption. This is explained in the forward with examples.

Additional resources for map layouts can be found on this page and the following sections in ArcGIS help:: https://pro.arcgis.com/en/pro-app/latest/help/layouts/layouts-in-arcgis-pro.htm

Q6: How do I design good-looking map elements?

This section will briefly go over some tips to improve your map elements and make them look more professional.

1. Legend or key

It is recommended to use the legend properties to alter the features in a legend (sometimes called a map key). Similar to the drawing order found on a map, you can turn on and off layers in the legend to decide on which symbols you want to be present. Consult Esri help for more tips on how to format legends using legend properties: https://pro.arcgis.com/en/pro-app/latest/help/layouts/work-with-a-legend.htm

One thing to note is that it is sometimes much easier to work with designing a legend when it is a graphic in ArcGIS Pro. This should only be done if you are set on the color scheme and do not plan on making any further changes to the contents of the legend (like adding another feature).

If you are planning on making multiple maps and plan on using the layout as a template, it is recommended that you update the legend in legend properties. Otherwise, when the legend is in graphic form, it is easier to fine tune the legend and make it look more professional. To convert a legend to a graphic by right-clicking on it, then click "Convert To Graphics". Then if you right click and 'ungroup' the graphic, you can easily move the individual part of the legend like symbols and text (you may need to ungroup the graphic twice depending on the complexity of your legend).
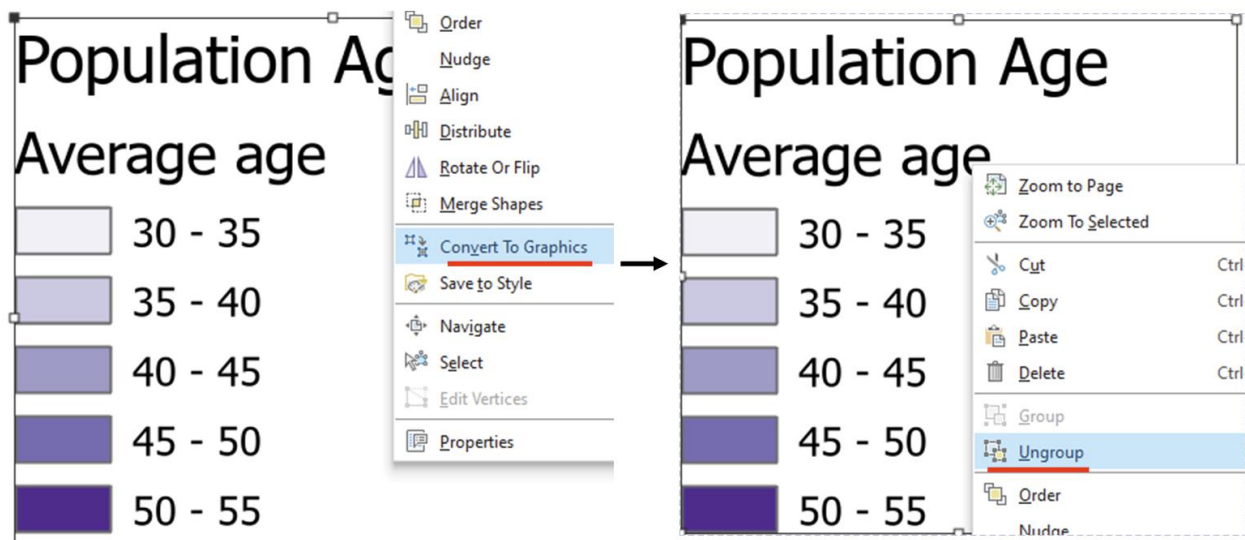
You can use guides to line up map elements and legend symbols and text to improve consistency in the spacing and layout. Once you have altered the legend elements, select all the text and symbols, and right click 'group' them so you can move the legend around your map layout as a single feature again.
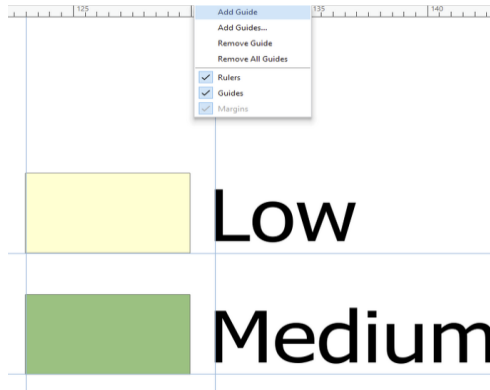
The image above displays how guidelines can be used to format a legend, with consistent spacing between the text and symbol.

2. Scale Bar

The default scale bar is usually not the most interpretable, so expect to have to alter the scale bar on most of your map layouts. Once you have placed a scale bar, click it once and the design menu should pop up in the upper menu ribbon. Open the design menu to find helpful tools for adjusting your scale bar. The fitting strategy is crucial because it determines how the scale bar is adjusted - typically "Adjust width" is the easiest to format. Then, alter the division value, divisions, and subdivisions to alter the display of the legend. Make sure you are using appropriate distance units, metric is the rule unless otherwise stated. If you want to fine tune your scale bar, you can convert it into a graphic (similar to a legend). You can do this to remove unnecessary division values, or easily flip the scale bar vertically for portrait layouts. Note: once you have a scale bar as a graphic, altering the size of the graphic distorts the scale; you should never do that.
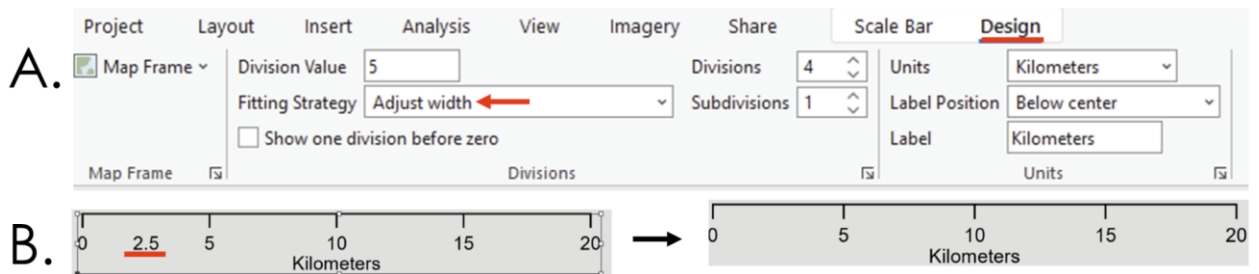
3. North Arrow

North arrows are relatively simple to implement in ArcGIS Pro. Typically you want a simple north arrow, and the size of it should be sufficient to understand what it is, but not too much larger. The North Arrow is a map element with relatively lesser importance than others and should be present on the map layout but not take up too much space. One should need to look for it rather than see it right away.

Q7: How do I export a map?

Once you have completed working on your map layout, click share in the top menu bar, then "Export Layout". You will be prompted to set a file type, and PNG is recommended due to its smaller file size. Set the DPI to at least 300 and define the file name and file path. You can use the clip to graphics extent to remove blank space in your exported map.
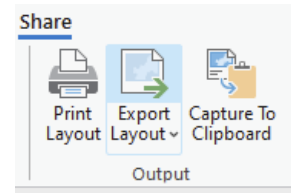
Q8: How do I label map features?

Labeling map features is important to provide context and information about the features that you are mapping. This section will go over how to label maps, label placement, and label formulas for point and polygon features.

To start, place the **heritage_sites**, **bike_routes**, and **ecosystem_sites** on your map in ArcGIS Pro. They will be located in their respective geodatabases in your **Lab Input Data** folder. In the upper menu ribbon, select you map, click on Select By Attributes, use input rows 'heritage_sites' and the clause Where SITE_NAME is equal to Port Kells Fire Hall No. 7 (see image below).
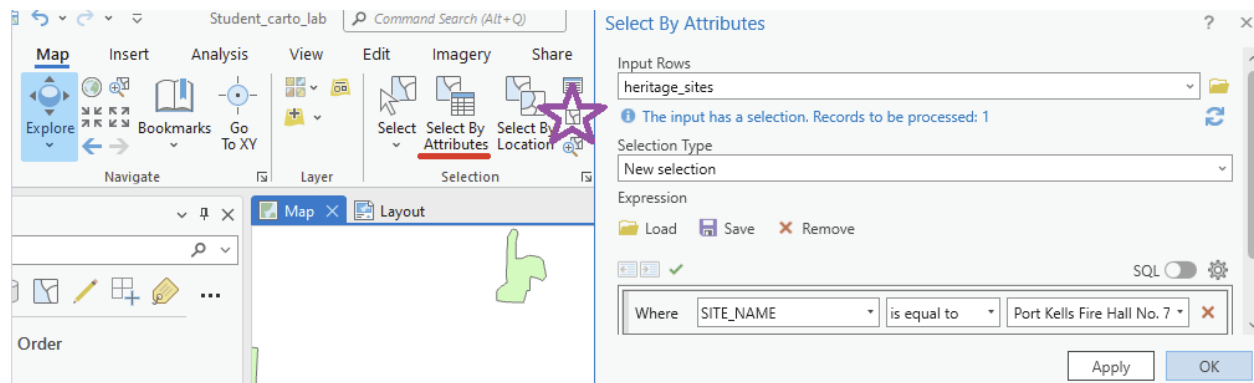
Click "Ok" and a light blue mark will pop up in the northwest corner of Surrey, BC. Zoom into that region since we will be mapping the Port Kells region of Surrey in this section of the lab. Then, in the upper menu bar, click the deselect button (purple star in the image above) to deselect the fire hall.

To enable the label for **heritage_sites**, click on **heritage_sites** once in the Content pane, and then in the upper menu bar click Labeling and then click Label. Ensure that the field is set to SITE_NAME to display the names of the heritage sites.
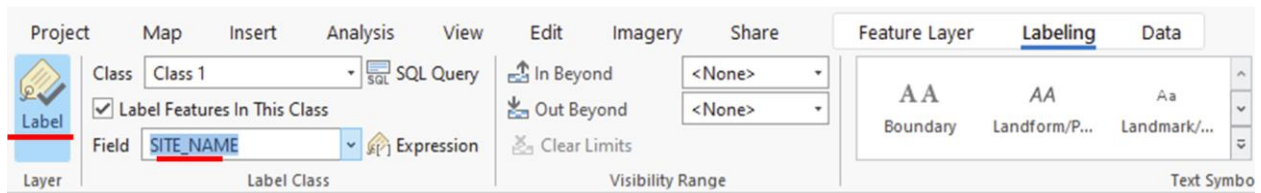
More information on labeling can be found here: https://pro.arcgis.com/en/pro-app/latest/help/mapping/text/labeling-basics.htm

Q9: How do I alter the placement of labels?

For this question, we will focus on altering the label placement of a point feature. There are two primary ways to adjust the location of point feature labels, adjusting feature weights and changing the placement preference.

The feature weight property can indicate the relative importance of certain map features from others. You can alter the feature weights by right-clicking on a map feature in your content pane and clicking "Labeling properties". In the newly opened label class pane, click position, feature weight, and you can alter the feature's importance. In the example below, the feature weight of the bike lanes was increased to prevent heritage names from overlapping with the bike lanes. More information on feature weights can be found here: https://pro.arcgis.com/en/pro-app/latest/help/mapping/text/prevent-labels-from-overlapping-certain-features.htm
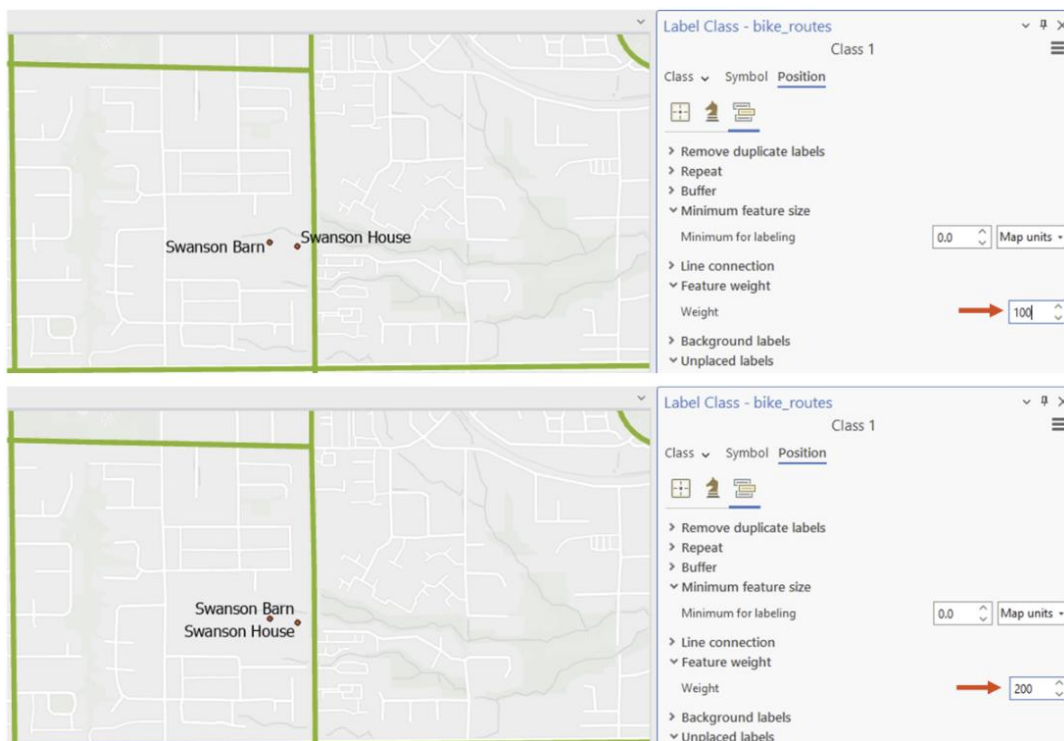
Another method for label placement, which can be used independently or with feature weights, is defining the positioning preference in the label properties. In Figure 5a, the two labels are a

bit confusing - it is not easy to see which point Swanson Barn and which point is Swanson House. By changing the label positions (swapping the 2 and 4 positions), the labels can be placed more clearly in Figure 5b.
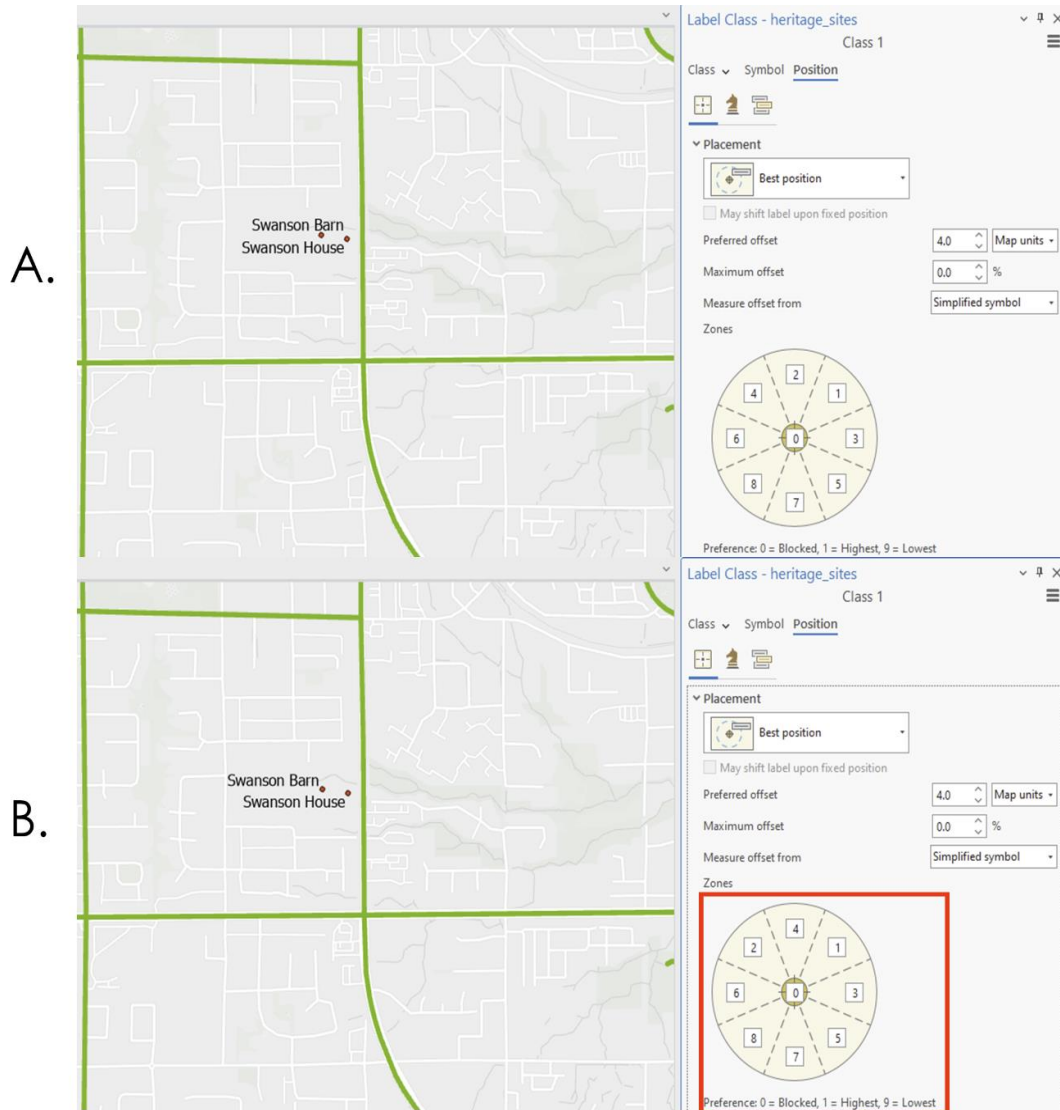
Figure 5: Labelling options for maps in ArcGIS Pro

If the features are very complex and you are finishing a map layout, sometimes it can be best to alter the label positions manually. In these instances, the convert labels to graphics tool can be used. https://pro.arcgis.com/en/pro-app/latest/tool-reference/cartography/convert-labels-to-graphics.htm

Section 2: Modifiable Areal Unit Problem

Q10: What is the Modifiable Areal Unit Problem (MAUP)?

The modifiable areal unit problem (MAUP) describes an effect whereby data reported in spatial polygons affect the mapping interpretation of patterns as well as the analytical results of data collected within those same polygons (Fotheringham et al., 1991). An extreme example of MAUP is gerrymandering in the United States, where modifying the specific boundaries of an electorate will alter the results of an election. Most instances of MAUP are more subtle in nature, but nevertheless it is important to consider the way that your spatial polygon data is defined (e.g., how the polygons are made, or the boundaries are determined, called zonation and scale (size) and how changes in polygon geometry can affect the results of any spatial analyses including how data appear on a map. The MAUP is present whenever you analyze data that is connected to polygon features or image (raster) features. When you are mapping something in a region and are using polygons, most commonly those polygons divide the entire region into a set of non-overlapping polygons and that is called a tessellation because they fill the entire space.

As an example, consider the location of some population in space, say people picnicking in a field that is flat with no trees and just short grass (a homogenous field) (Figure 6).

Figure 6. People in Krüöppen field on national festival day; Left: photos of picnickers; Right: Square quadrant overlain on portion of picnic field. Image created using DALL-E and Adobe Photoshop.

Then we overlay a quadrant (a large square) on a random portion of that picnic field (Figure 6). Next, we divide that quadrant into a set of 10 polygons and count the number of people within each of the 10 polygons. When we count points in polygons that is called a point-in-polygon operation. Next, we make another set of 10 polygons that have different sizes (scales) and

shapes (zonations) and repeat the same point-in-polygon operation. Then for each of our polygons we divide the counts by the area of the polygons to create the number of people per polygon. That is a density value, and we can show that on a choropleth map with the continuously varying color for the map scale/key. The result is illustrated in Figure 7.

A   B   C
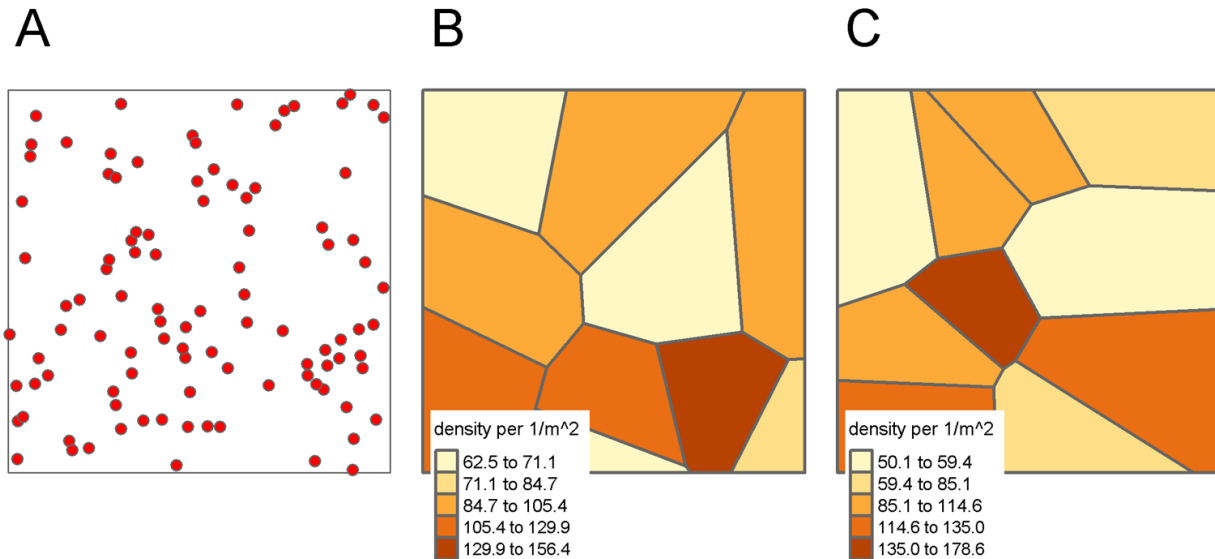
Figure 7: A) the distribution of population in some region where each red dot represents 1 thing (person, object etc.); B) a tesselation of the region with polygons showing the density of population in each region based on the points shown in A; C) A different tesselation of the region with the same number of polygons as in B but different densities and using the same population in A.

From Figure 7, it is clear that two different polygon tessellations of the same space provide very different maps of the density of picnickers. Examine just the values of minimum and maximum density in Figure 7b & c, the ranges are different. If you were asked to describe where the highest density of picnickers is you would conclude from Figure 7b that they are in the bottom right of the field, whereas if you used Figure 7c, you would conclude they are near the center of the region. Despite the picnickers remaining stationary as shown in Figure 7a, the conclusions drawn are significantly different due to the utilization of two distinct polygonal tessellations or zonations. These zonations, while containing the same quantity of polygons (10 in each map), differ in boundaries and scale. This difference in spatial partitioning has resulted in diverse values and interpretations. Such variances are a direct consequence of the Modifiable Areal Unit Problem (MAUP).

The Modifiable Areal Unit Problem (MAUP) essentially ensures that distinct results emerge when the same analysis (like the point-in-polygon analysis depicted in Figure 7) is conducted on an identical dataset, given that the dataset is represented within polygons differing solely in boundary location (zonation) and area (scale). Put simply, the geographical scale of data

analysis and the defined boundaries for the area can significantly influence the outcome of the analysis. The two components of the MAUP are thusly the:

Zoning Problem: This involves the emergence of different results when using the same underlying data when the arbitrary nature of the boundaries used to summarize that data are so called imposed spatial units (Bonham-Carter 1993) - even if the physical area/scale of those units is kept constant, the boundaries themselves could be changed and thereby the results of any analysis using the data summarized in those polygons can change as well. Thus, how you define the polygons you use for analysis directly influences the outcome of any measures.

Scale Problem: This refers to the fact that results from statistical analyses will be different based on the spatial scale of analysis. For instance, analyzing data at the level of a neighborhood (a large unit) might yield different results than analyzing it at the level of individual census tracts (smaller units). In a similar vein, a geological map might depict a substantial geological formation, such as a sedimentary basin or igneous intrusion. This larger formation may encompass smaller, nested units of diverse rock types. If you undertake an analysis, like evaluating hydrogeological parameters such as pressure per geological unit, you will find that the results vary depending on the size of the geological units you examine. In essence, the scale of the geological units used in your analysis can significantly influence the outcome. As the scale changes (from small-scale, like a country, to large-scale, like a neighborhood or a geological formation to a rock type), the data analysis will reveal different patterns or relationships.

In reality, these two aspects of the MAUP are not separable for a given polygon-based dataset. These problems highlight the challenges of spatial analysis and the importance of understanding the influence of spatial scale and zonal boundaries in analyzing geographical data. It's important to consider the MAUP in designing studies, interpreting results, and making policy decisions based on spatial data because what might be true for one zonation and scale might not be so at another. To see how this is a law of geography, see Box 1.

To understand what types of analysis might be affected by the MAUP, let's consider the correlation between two geographic variables that are distributed randomly in space. All you need to know is that correlation is a statistical measure that indicates the degree of linear relationship between two variables. Correlation ranges between +1 to -1, where +1 is perfect positive correlation and indicates that as the value of one variable increases, so does the other. Whereas a -1 is perfect negative correlation (an increase in one variable leads to a decrease in the other). A value near 0 means there is no relationship between the two variables.

In our case, we will consider two variables which are the locations of picnickers over two subsequent years (this could be any two patterns of points, e.g., assessing the locations of sampling geotechnical properties from borehole data over two years to see if the samples are random and unbiased. In any case, we hypothesize that there is no relation between the locations of picnickers in subsequent years because the park's field is homogeneous, it's only grass which is all the same height everywhere and there is no underlying difference in

permeability of the surficial geology because the soil matrix is the same everywhere (so no wet areas), the locations of picnickers in one year should have no relation to the locations of picnickers in the subsequent year. In other words, one place is just as good as another on the field and so picnickers should be randomly distributed across the field. If two geographic or spatial variables are randomly distributed then there should be no relationship between them, even when that underlying data is zoned and analyzed at the ecological level. For example, consider the locations of picnickers on two subsequent years of on the national festival day (Figure 8):
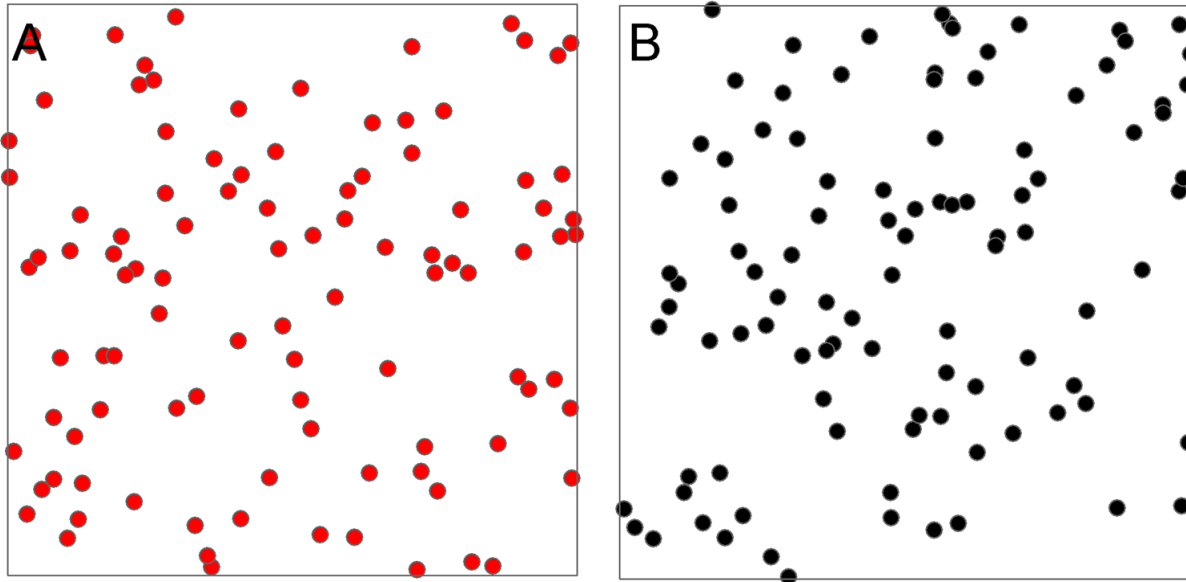
Figure 8. A) locations of picnickers in 2021; B) 2022

Again, we will choose to compare the densities of picnickers between both years by tessellating the space containing each year's locations using 10 polygons and then correlating the values of the 10 polygons with each other. There should be no correlation in the density values between the two years since the picnickers are randomly distributed each year. In fact, the correlation is near 0 which means that there is no correlation using measures specifically suited to point bivariate point pattern analysis. Therefore, in reality there is no relationship between the locations of picnickers in 2021 and 2022. However, if we did not have access to the raw data and instead the same data was presented within a set of polygons then the story changes. Depending on how the polygons are created, the number of picnickers in each polygon represented as a density value in both years could easily show a strong positive correlation or, zero correlation or even strong negative correlation (Figure 9).
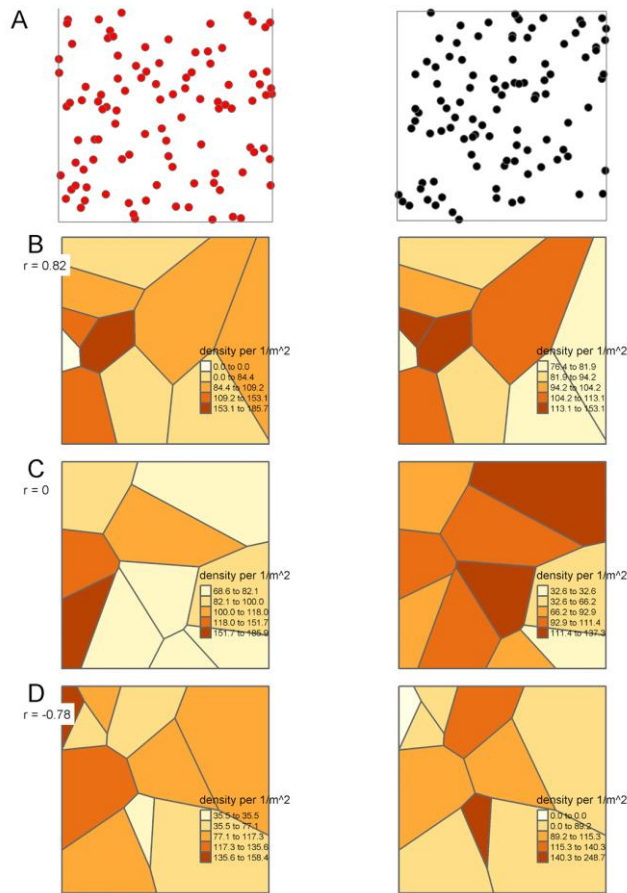
Figure 9: A) locations of picnickers in 2021 (left) and 2022 (right); B) Ten polygons that would provide a strong positive correlation; C) for the same data, a polygon set that would show no correlation; D) for the same data, a set of polygons that shows strong negative correlation. Remember the only thing that changes are the polygon configurations (zone size and zone boundaries for each zone or polygon). Note as well the density scales vary for each map in each column as explained previously. The point location never changes. The 'r = 'represents Pearson's R, the correlation coefficient and while not appropriate for geographic data, it is used here to illustrate the effect of the MAUP without using more complex spatial econometric/statistical measures.

What we see in Figure 9 is the analytical aspect of the MAUP that illustrates if we change the zonation and scale of 10 polygons within the quadrant then we will get different correlations for the same underlying data. In fact, we can randomly change the 10 polygon boundaries and sizes until we land on a relationship we like. Keep in mind that we are doing this to illustrate the MAUP and in reality, you would <u>NEVER</u> change polygons just to get values that fit with your preconceived notions. That is called **academic fraud**.

When you work with data that are given as attributes of polygon features and those polygon features are not 'natural units' like say a lake, or an island, then what you might interpret or state as an analytical outcome could just be due to the configuration of polygons. This is a major

problem in spatial data science and geographical analysis. You must be fully aware of this and convey that limitation whenever analyzing data from polygon geometries.

Q11: How do I create different zones to visualize the Modifiable Areal Unit Problem?

Census tracts (CTs) are used by Statistics Canada to divide Canada into geographic areas consisting of around 7,500 people that follow recognizable physical features such as rivers or roads. CTs are used in Surrey, BC to separate the city into zones that are used for statistical purposes. However, the definition of the CTs can influence how the census data in question are portrayed. In this question, we will make synthetic CTs and evaluate how different zone boundaries can change the density of heritage sites per CTs. This example of the MAUP will show how the division of zones in a map can change the message of a map.

The first step will be to create a new set of synthetic CTs that have different boundaries than those defined by Statistics Canada. CTs are created through the conglomeration of Dissemination Areas (DAs), which are the smallest geographic unit used in the census. We will create a new set of synthetic CTs by using ArcGIS geoprocessing tools and Dissemination Areas. The first step is to locate the dissemination area feature class, which is located within the Census_2016_Surrey_DisseminationAreas_UTM.gdb geodatabase, inside the Surrey_Without_Data feature dataset (folder), titled "Dissemination_Areas_Surrey". Add that to your map. Then, on the top menu bar click Analysis, then Tools, and the geoprocessing panel will show up on the left part of your screen. In the "find tool" search bar, look up "Build Balanced Zones" and then click on the tool. Set the input feature to "Dissemination_Areas_Surrey", Zone creation method should be set to "Defined number of zones" and define the target number of zones to be 96, since that is how many CTs there are in Surrey BC. Then click "Run". If you encounter any issues with the tool, click the upper right question mark button to access ESRI help.
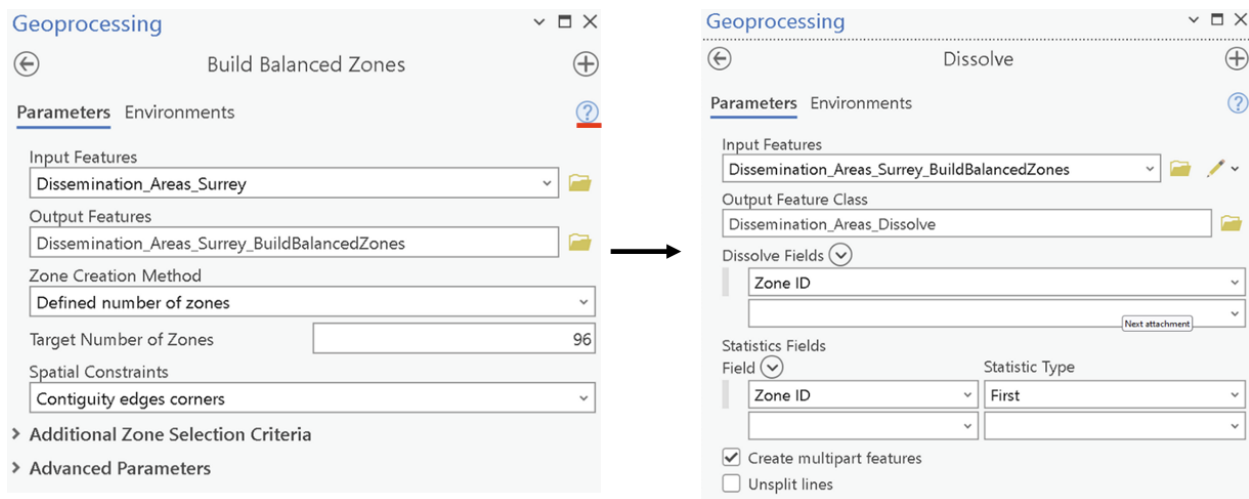
Once the balanced zones have been generated, we need to use the dissolve tool to aggregate the DAs into their synthetic CTs. Enter in the find tool search bar "Dissolve" and click on the tool. The input feature should be Dissemination_Areas_Surrey_BuildBalancedZones, and the

dissolve tool should be set to "Zone ID". Then click run to run the tool. Note that you may receive a notice that says that "The Number of Zones requested was not reached" - ignore this warning and continue to use the feature class. All outputs from geoprocessing tools default to save to the geodatabase in the home folder where your project is saved. Your output from the dissolve tool should look similar to the below map, but with different symbology for the zones.
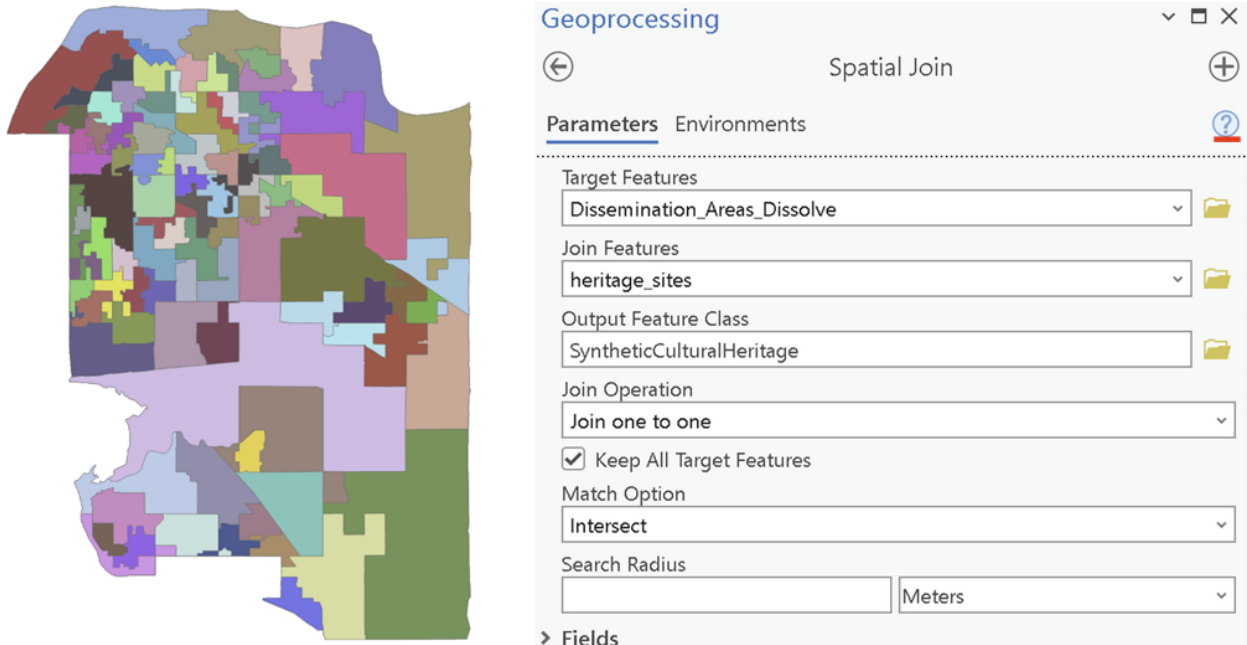
The next step is to add the heritage sites point data to the synthetic CTs using Spatial Join. A field titled 'Join_count' will be created and will allow us to know how many heritage sites are located within each synthetic CT. Enter in the Find tool search bar "Spatial Join" and click on the tool. Spatial join is a complex tool, so it is recommended to click the question mark button to better understand how the tool works before proceeding. Use the Spatial join with the following parameters:

- Target Features: Dissemination_Area_Dissolve
- Join Features: heritage_sites (note if you have not added heritage_sites to your map yet, click on the folder icon next to Join Features and access the heritage_sites feature class within heritage_sites.gdb).
- Join operation: Join one to one
- Keep all target features: Checked
- Match Option: Intersect

Then click **Run**.

Once you have completed the spatial join for the synthetic CTs, open SurreyCensusTracts in your map from the Census2016_Surrey_CensusTract_UTM.gdb. Do a similar spatial join where you join the heritage_sites to the SurreyCensusTracts feature class.
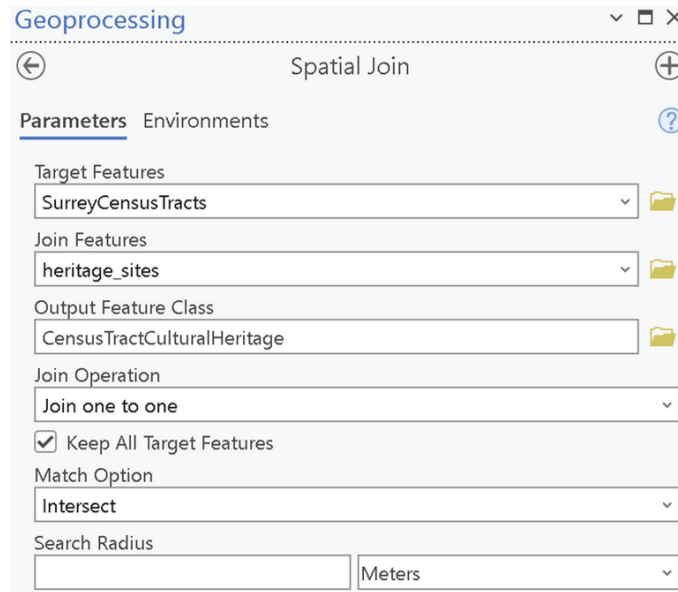
Q12: How do I symbolize two feature classes using the same symbology?

This question uses a tool that allows for the user to symbolize a feature class based on the symbolization of another feature class. It is important to have the same breaks and similar symbolization when comparing results between map elements, so this is a very important tool to know how to use.

Once you have completed both spatial joins, it is time to symbolize them using the same breaks so we can make comparisons across the maps. We will be comparing the density of cultural heritage sites per zone. Firstly, let's symbolize the SyntheticCulturalHeritagefeature class. Right click on SyntheticCulturalHeritageand, select Symbology, then symbolize the feature using:

- Primary symbology: Graduated Colors
- Field: Join_count
- Normalization: Shape_Area
- Method: Natural Breaks (Jenks)
- Classes: 3
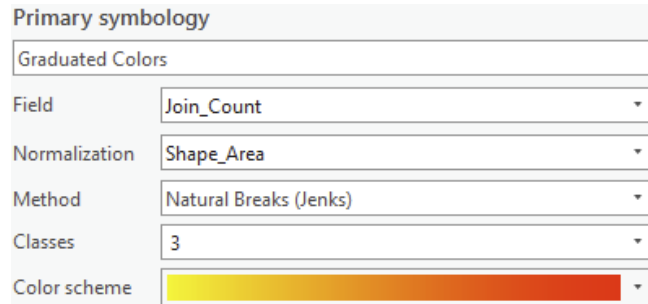
Then label the three classes, **low**, **medium**, and **high**.

Once you have finished symbolizing the SyntheticCulturalHeritage, use the geoprocessing tool "Apply Symbology From Layer" to apply similar symbology to CensusTractCulturalHeritage. Set the input layer to CensusTractCulturalHeritage, and the symbology layer to SyntheticCulturalHeritage and then run the tool.
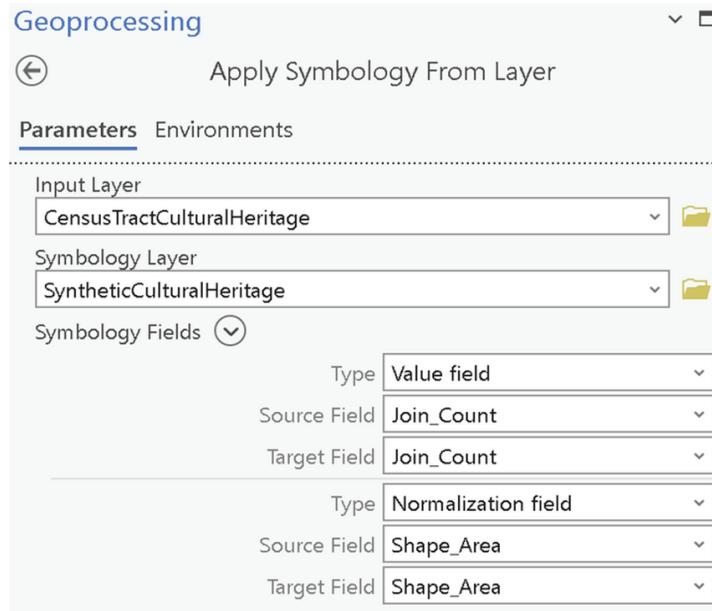
You should now have two layers with identical symbology that can be used for comparison.

Questions to hand in.

1. [5 points] Create a map layout that follows the below criteria:
   - There should be no basemap.
   - The following historical sites should be labelled: Bulman's Garage, Port Kells Post Office, Port Kells Fire Hall No. 7, Port Kells Elementary School, Church Manse, and Port Kells Park.
   - Include the Bike routes and Ecosystem sites in this map.
   - The Heritage sites should be labelled and cannot have their **site_name** touching or crossing the bike lanes.
   - The map scale must be 1:10,000.
   - Cartographic best practices must be used when creating this map layout.
   - You must use **the Arial** font, and a different scale bar and north arrow than the below map.
   - The symbology colour of the Ecosystem sites, bike routes, and heritage sites must differ from the map below.

Your map should look similar to the one below, though it does not need to be identical.

Historical Sites and Ecosystems in Port Kells
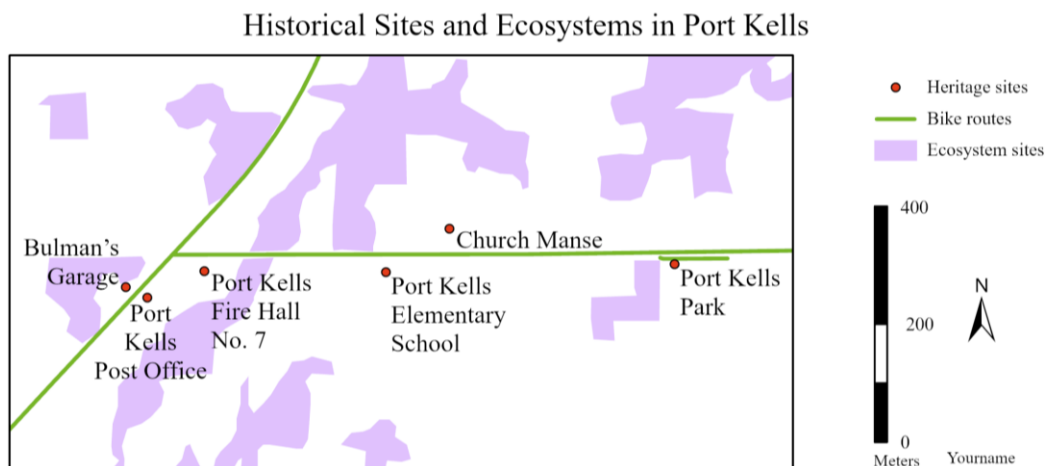
2. [3 points] In a short paragraph (max 150 words), explain how the map you submitted for Question 1 uses visual hierarchy? Why is establishing your visual hierarchy important?

3. [8 points] Similar to example questions 10 and 11, create a new synthetic census tract feature class. Then spatial join **important_trees** (located in **important_trees.gdb**) to the synthetic census tract and a census tract feature class. Using cartographic best practices, create two maps with identical symbology and map extent, with similar elements. Ensure that each map has a proper title.

4. [2 points] Explain the Modifiable Areal Unit Problem (MAUP) and how it manifests itself in the two maps you created for Question 3.

5. [1 point] If I were to create a map displaying the number of dogs found in each census tract in Surrey BC, which symbology should I use? Provide reasoning (max 100 words).

6. [1 point] If I were to create a map displaying the density of dogs found in each census tract in Surrey BC, would a Choropleth map be appropriate? Provide reasoning (max 100 words).

Student Name: _____ Student #: _____

Laboratory #2: Coordinate systems and transformations

Purpose: The purpose of this lab is to introduce geographic coordinate systems and transformations. The following exercise will allow you to both define a coordinate system and calculate the uncertainty of position within the digital geospatial layers.

Learning Outcomes: By the end of this lab, you will be able to:
- Identify the main components of the ArcGIS Desktop ArcGIS Pro user interface.
- Define a geographic coordinate system.
- Transform between geographic coordinate systems.
- Calculate location uncertainty of digital geospatial layers.

Procedure:
- Complete Group 1, Tutorials 1-5, within the ESRI Academy.
- Read and work through the examples.
- Answer the questions at the end of the lab.

Data/Materials:
- All data are provided in lab2.zip.

Assignment:
- Answer the questions (#1-9) in the spaces provided.
- The total marks available for this assignment = 27 marks for 10% total course grade.
- All lab assignments must be completed and handed in individually.

Format:
Students will hand in questions with completed answers.

Tables and figures should be produced to the standards of the Department of Geography, Environment and Geomatics, University of Ottawa.
For guidelines on proper formatting and presentation of tables and figures, please see the introductory section of this lab manual.

Q1: Why is projecting data important in GIS?

Projecting, the process of converting an ellipsoid model of the earth to a flat map, is an essential component to working with geospatial data. Projections can be used to preserve the following characteristics of features of the earth**: shape, area, direction,** and **distance**. However, there isn't a single projection which can preserve all these characteristics of features on the earth on a two-dimensional viewing device like a monitor or flat paper. So you have to select the appropriate projection for the task at hand, given the scale (city, country, continent, world) and location of what you are mapping.

The most common projection used for world maps is the Mercator projection (created by Gerardus Mercator in 1569). The Mercator projection was designed for navigation and has relatively good shape preservation. Moreover, every straight line on a Mercator map is a Rhumb line, i.e., a line of constant bearing, which a ship could use to navigate to get from one point to another. The Mercator projection, more so than any other, was one of the essential tools of global colonialism.

However, the Mercator projection is inappropriate for most maps of Canada, since it creates significant distortions of spatial objects; as one moves away from the equator, either north or south, both area and distance increase. An example of this is how the Mercator projection distorts the Great Slave Lake in the Northwest Territories. Using the Mercator coordinate system, the calculated size of the lake is approximately 120,678 km^2.
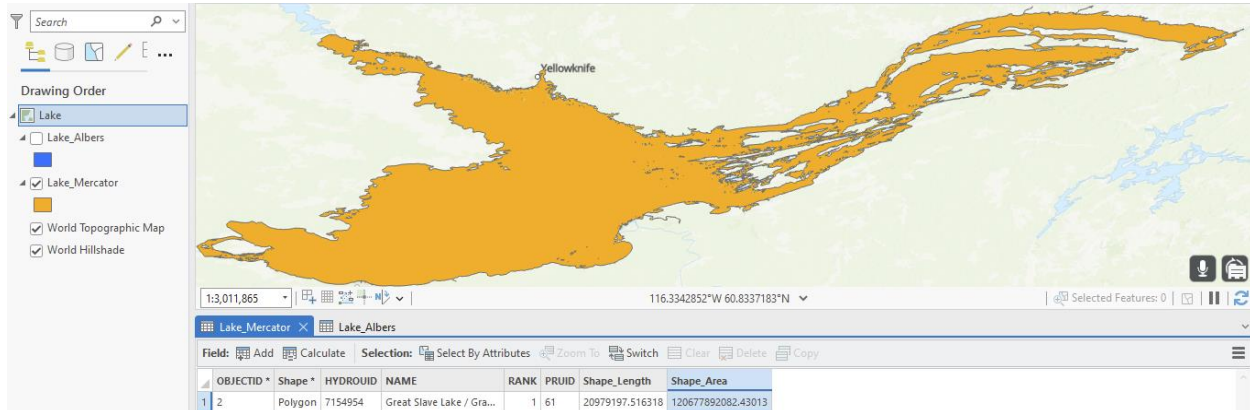
However, using a projection created specifically for Canada that is designed to preserve both shape and to some extent area as well, the Albers Equal Area, the lake size is approximately 27,065 km^2. This is substantially closer to the true size of the lake, which is estimated to be 27,200 km^2.
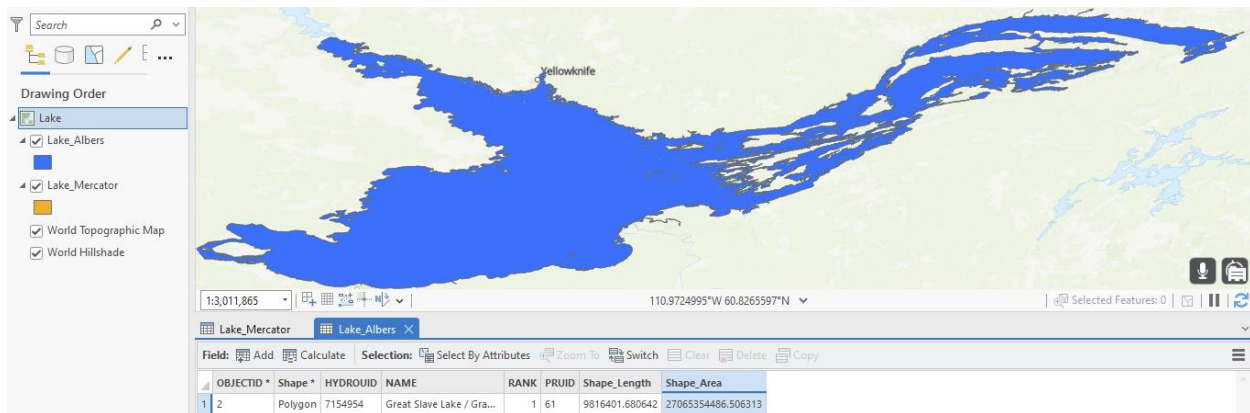
It is essential to use the correct projection when you work with geospatial data. If you are planning on conducting area-based calculations or spatial analysis with the data, it is necessary that your data is projected into a projection that preserves area. Projections such as Canada Albers Equal Area Conic would be the most appropriate for area-only project needs and, likewise, Lambert Conformal Conic is appropriate for small-scale maps of Canada.

The Project tool and Project Raster tools in ArcGIS Pro can be used to convert your data into the appropriate projection. To determine a projection of a layer, right click on it, click on properties, click source, then scroll down and open the spatial reference section. Only layers that have a projection already defined can be reprojected to another projection.
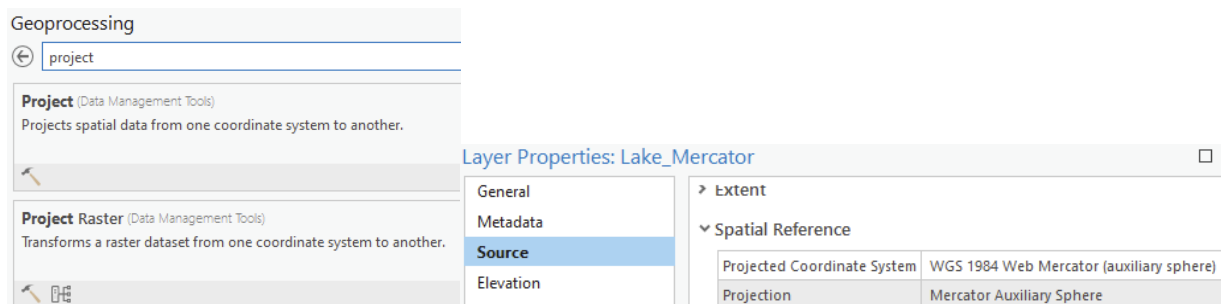
Q2: How do I convert degrees, minutes, and seconds to decimal degrees?

If you are given a coordinate such as 79°22'34" W, the first thing you need to do is notice that this coordinate references a location WEST of the prime meridian. Therefore, you will need to add a (-) sign to the result **after** you convert to decimal degrees. This is just convention, longitudes east of the prime meridian are considered positive, and those to the west are considered negative. To convert 79°22'34" W to decimal degrees, follow these equations:

$$dd = \deg + \frac{\min}{60} + \frac{\sec}{60^2}$$

$$e.g., Long = 79^o 22'34"W$$

$$dd = 79 + \frac{22}{60} + \frac{34}{60^2}$$

$$= 79 + .3666 + .00944$$

$$= \text{-}79.37611W$$

Notice at the end, a (-) sign is added. This does not mean "negative" but rather is a way to tell someone or something that this coordinate is West, and not East, of Greenwich, England.

To reverse the process, you remove the -79 portion of the decimal degree coordinate, then you are left with decimal minutes. You multiply the decimal minutes by 60 and remove the integer portion of what is left over. Then you multiply the remaining decimal portion by 60 to get the number of seconds, like this:

$$-79.37611\, dd$$

$$= 79^o \, and \, .37611 \times 60'$$

$$= 79^o \, and \, 22.5666 \, dm$$

$$= 79^o \, and \, 22' and \, .5666 * 60"$$

$$= 79^o \, and \, 22' and \, 33.996"$$

$$= 79^o 22'33.996"$$

Wait, why is this different from 79°22'34" W? Note that you would ideally do this on a calculator or computer, otherwise you would get a different result for the seconds due to numerical precision (the number of decimal places you retain vs. a computer), as happened here. Ignorance of numerical precision could lead to greater errors! Luckily, these calculations are easy to make in ArcGIS's Python window. Open the python window from the ArcGIS Pro Analysis ribbon:
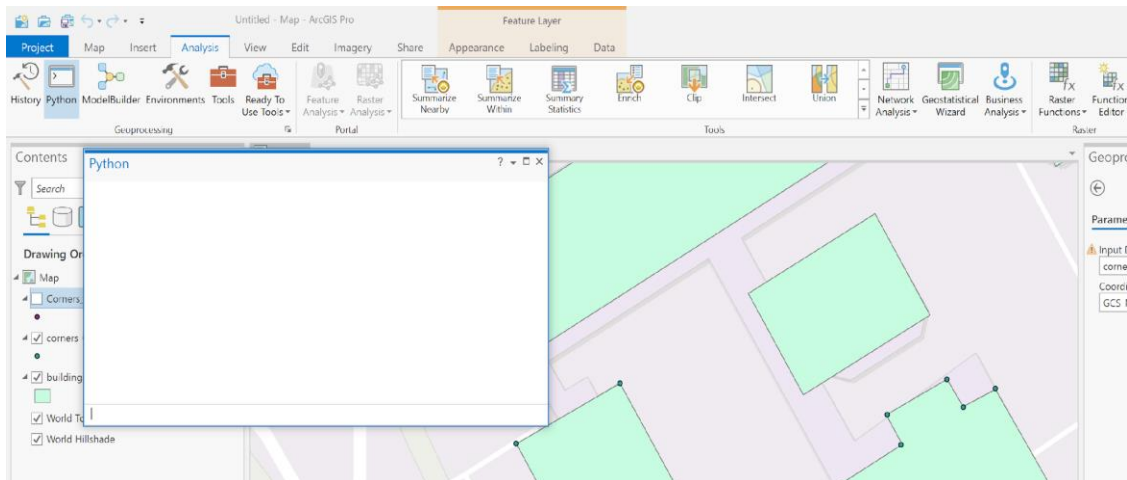
Type the following:

```
dd = 79 + 22.0/60.0 + 34.0/60.0**2
```

This statement will store the result in a variable called `dd` within the python window. To see what is stored, type the name `dd` like this, and press Enter:
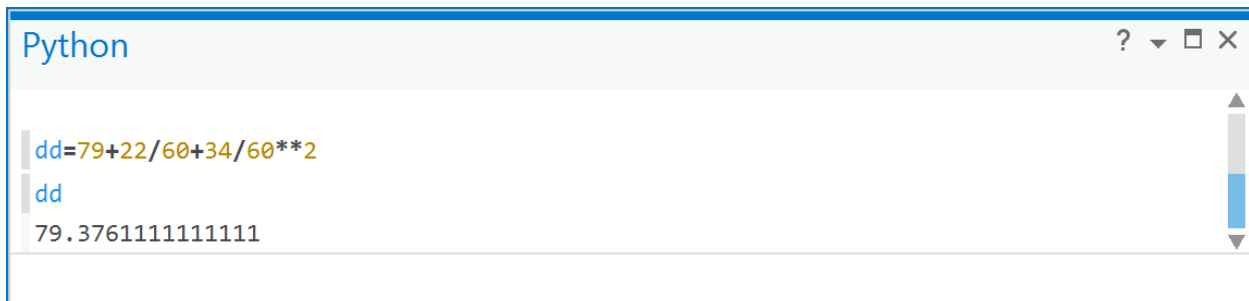
You can see there are many decimal places, in fact the 1's repeat for a long time. So, convert dd back to degrees, minutes, and seconds. First, we remove the 79, making note of it on a piece of paper, to arrive at the decimal minutes (DM),
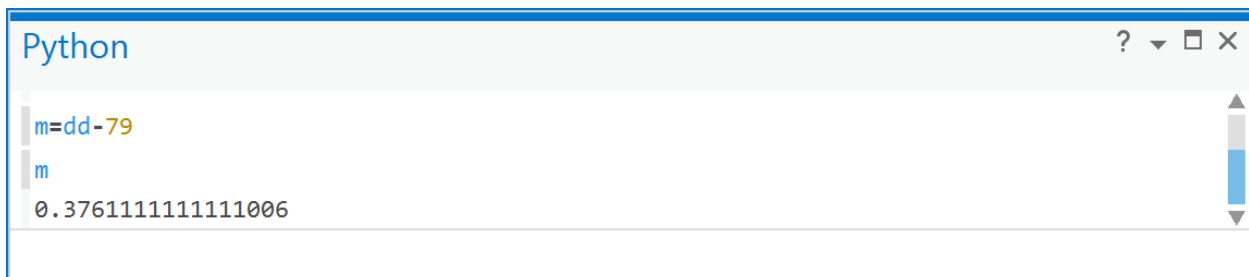
So, now in `m`, you have the decimal minutes portion, so we will multiply it by 60 and store the result in a variable called `mins`,

```
Python                                                    ? ▾ □ ✕

mins = m * 60
mins
22.566666666666038
|
```

Also make a note of 22 minutes, wherever you noted 79 degrees previously. Finally, remove the 22 minutes:

```
Python                                                    ? ▾ □ ✕

s = mins - 22
s
0.5666666666660376
|
```

and multiply the result stored in `s` by 60 to arrive at the number of seconds,

```
Python                                                    ? ▾ □ ✕

s*60
33.999999999962256
|
```

So, this can be rounded to 34". You can see readily that, even though 79°22'34" W is expressible as the sum of rational numbers, the computer cannot retain enough decimal places to ensure a perfect 34" result, and so the result can be irrational. This should be considered a source of uncertainty when converting from DMS to DD.

Q3: What is the real-world distance of 1 degree of longitude at the equator?

The distance of one degree of longitude, e.g., if you were to sail from longitude 0° to 1° W exactly along a line of latitude, based on the assumption that the Earth is a sphere, is given by:

$$d = 111\,km \times \cos(latitude)$$

Student Name: _____ Student Number: _____

Latitude is typically given in radians rather than degrees, which is $rad = deg \times \frac{\pi}{180}$. So, for example, given a latitude of 0°:

$$d = 111\,km \times \cos\left(0 \times \frac{\pi}{180}\right)$$
$$= 111\,km \times 1$$
$$= 111\,km$$

So there are 111 km in 1 degree of longitude - at the equator.

Q4: How do I calculate the real-world distance between two points using longitude and latitude?

You can calculate the distance between two points for which you know the longitude and latitude using what is called the *Haversine formula* for great-circle distance. This formula assumes that the earth is a sphere, and the following equation provides an *approximation* to the true geodesic distance (distance on an ellipsoid):

$$D_{A \to B} = r\cos^{-1}\left[\sin\phi_A \sin\phi_B + \cos\phi_A \cos\phi_B \cos(\beta)\right]$$
$$where, D_{A \to B} = \text{ distance from point A to B and,}$$
$$r = \text{ radius of earth in km or m}$$
$$\phi_A = \text{ latitude of point A}$$
$$\lambda_A = \text{ longitude of point A}$$
$$\phi_B = \text{ latitude of point B}$$
$$\lambda_B = \text{ longitude of point B}$$
$$\beta = \lambda_A - \lambda_B$$

The above equation assumes that all angular measures are in radians. So, for example, if we have the following two points:

Point 1		Point 2	
LAT	LONG	LAT	LONG
46.62006	-84.3039	46.62009	-84.3038

$$D_{A \to B} = r \cos^{-1}\left[\sin \phi_A \sin \phi_B + \cos \phi_A \cos \phi_B \cos(\beta)\right]$$

$$where, D_{A \to B} = \text{ distance from point A to B and,}$$

$$r = 6378 \text{ km}$$

$$\phi_A = 46.620061 \times \frac{\pi}{180} = 0.813673562$$

$$\lambda_A = -84.303857 \times \frac{\pi}{180} = -1.471379877$$

$$\phi_B = 46.62009 \times \frac{\pi}{180} = 0.813674138$$

$$\lambda_B = -84.30383 \times \frac{\pi}{180} = -1.471379406$$

$$\beta = -1.471379877 - -1.471379406 = -4.71239E - 07$$

$$= 6378 \cos^{-1}\left[\begin{array}{l} \sin(0.813673562)\sin(0.813674138) + \\ \cos(0.813673562)\cos(0.813674138)\cos(-4.71239E - 07) \end{array}\right]$$

$$= 0.00426 \text{ km} = 4.26 \text{ m}$$

So, the approximate distance between these two points for a spherical model of the earth, of radius 6378 km, is just over 4 m. You should always use a computer to do such calculations because the retention of many decimal places is required. Like the above example, you could do this in the python window in ArcGIS quite easily (make note of the use of parentheses!):

```
Python                                                                    □ ×
>>> Lat1 = math.radians(46.620061)
>>> Long1 = math.radians(-84.303857)
>>> Lat2 = math.radians(46.620094)
>>> Long2 = math.radians(-84.30383)
>>> 6378*math.acos(math.sin( Lat1 ) *math.sin( Lat2 )+math.cos( Lat1 )*math.cos( Lat2 )*math.cos( Long1 - Long2 ))
0.004211873988023198
>>>
```

In the above example, the units are in km because the radius of the earth was in km. You can convert to meters by multiplying the result by 1000. You could also do this in Excel using the RADIANS() and ACOS(), COS() and SIN() functions in a spreadsheet formula.

Q5: How do I calculate the location uncertainty for a digital layer?

A paper map is produced from aerial photography using photogrammetric techniques. A photogrammetrist looks at two overlapping air photos using a stereoscope and uses a range of techniques and specialized knowledge to transfer (trace) those features onto a transparent sheet of acetate. The flying height and camera lens as well as some other aspects of the air photo determine the scale and, in turn, determines the largest map scale that the cartographer can produce from such photos. You can see how a stereoscope works with airphotos here: https://youtu.be/mZkmXEFw7BE. The map that the cartographer produces is then manually or digitally digitized using either human or machine algorithms to create digital geospatial data layers used in geomatics processes within a GIS.

As a function of this process, a digital geospatial data layer that originates from a paper map has a **base scale** associated with it. The **base scale** is the map scale upon which the eventual

digital representation is based. These digital representations then become thematic layers such as roads, built-up areas, rivers, forested area etc. For example, a 1:50,000 map of the Ottawa region (NTS O31G5) looks like this:
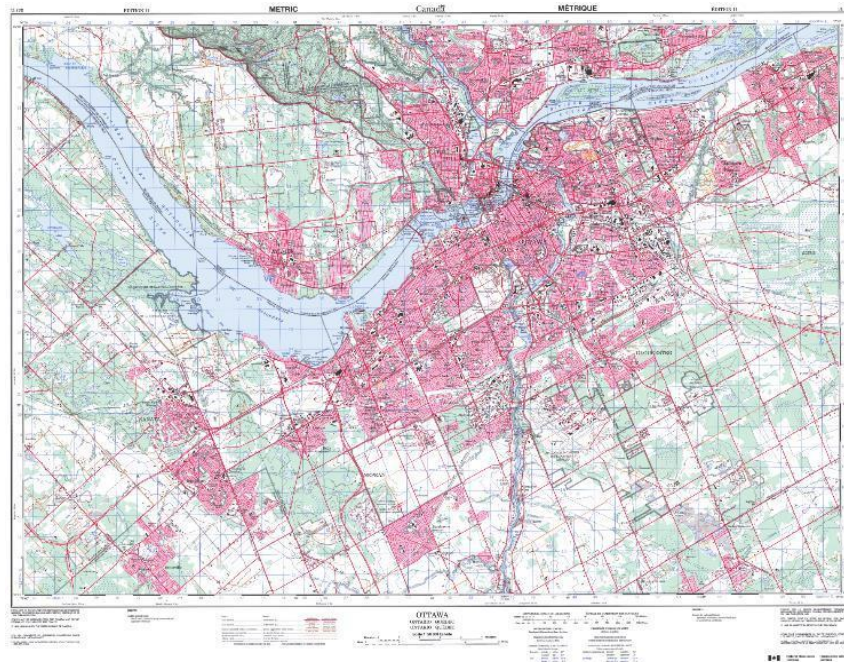
The smallest feature on a map that can be seen at a regular viewing distance on a map is 0.5 mm. So, for a map of 1:50,000 scale, that 0.5 mm line (or 0.5 mm diameter of a point) is covering a width of 0.5 mm x 50,000 = 25,000 mm, or 25 m, in the real-world.

Assuming the cartographer was unbiased in tracing objects on the air photos when making the map representation, then we can say that a position (a coordinate pair) read from a feature on that map would be at most ½ the 0.5 mm line width away from its true location. Why, because we must assume that the true location of a feature on a map is in the center of the 0.5 mm line. 0.5 x 25 m = 12.5 m, so the uncertainty of a position read from a feature on a 1:50,000 scale map would be ~12.5 m. Such positional uncertainty is important to keep in mind when you use geospatial data. If you're using a map to plan a picnic, a 12.5 m error in where you tell them to meet is not a big deal. But if you're landing a plane at night, 12.5 m off to the side could put you outside the runway, or if you're building a bridge that links to the road on the other side…

1:50,000

Coordinate read from map feature

y

x

True location

0.5 mm

Two possible locations within the line feature; both locations are 0.25 mm from true location at the center.

At a 1:50,000 scale map, the best we can do is measure a location on the line but where our location is with respect to the center of the line, we can't know.   The location we read could be as far as the either edge of the line with respect to the center.

A related concept is the minimum resolvable feature, and this is based on the map scale. The minimum resolvable feature refers to the minimum separation distance between two lines so that they are distinguishable as individual entities on a map. That value is based only on the 0.5 mm value and represents, in the case of a 1:50,000 scale map, the value of 25 m.

So, to calculate the uncertainty in position of a geospatial map layer due to the **base scale only** do the following:

$$\delta = 0.25 \ mm \times \text{scale denominator}$$
$$\text{map scale} = \frac{1}{50000}$$
$$\delta = 0.25 \ mm \times 50000$$
$$= 12500 \ mm$$

However, a digital geospatial layer that originates from a paper map involves additional uncertainty because whatever entity (human or AI) that created the digital representation had to extract features from the paper map by tracing them into a GIS system. That process is called **digitization** and itself adds more uncertainty. A machine or person cannot trace features on a paper map perfectly. A perfect trace would occur when the person/machine always traces the features at the exact centerline of the feature represented on the map, but that doesn't happen very often, even when done using machine algorithms because of simple things like image distortion when a map is photographed (digitized), paper stretching or shrinking due to humidity and/or within a scanner, etc. The figure below shows an example where the location of a digital

feature (black line) is located within, but not in the center of, the underlying feature (thick red line).
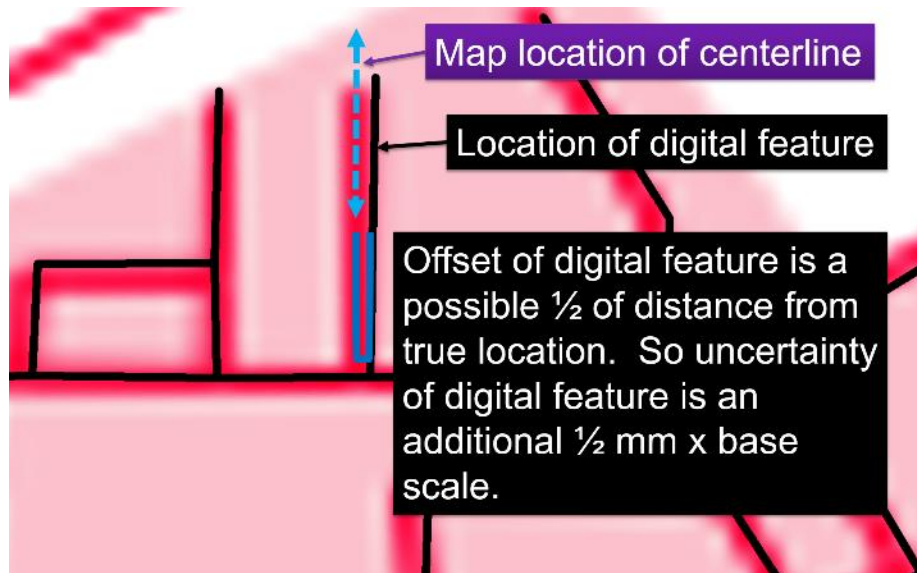
Assuming that the digital representation stays within the 0.5 mm feature width while tracing the map, then, at most, a coordinate of a location on the digital representation (black lines in Figure 7) will have an uncertainty with respect to the base-map equal to that of the base-scale of the map, specifically,

$$\delta = 0.25 \ mm \times \text{base scale denominator}$$
$$\text{base scale} = \frac{1}{50000}$$
$$\delta = 0.25 \ mm \times 50000$$
$$= 12500 \ mm$$

The total uncertainty is calculated using error propagation formulae that assume there are **no correlations** between the errors of the person making the original map and the errors committed by the entity that transformed it into a digital representation. The total error is therefore additive, and is given by:

$$total \ \delta = \sqrt{\text{base scale error}^2 + \ \text{digital error}^2}$$
$$= \sqrt{12500^2 + 12500^2}$$
$$= \sqrt{156250000 + 156250000}$$
$$= \sqrt{312500000}$$
$$= 17677.67$$
$$= 17.68 \ m$$

This formula is derived using a Taylor's series differential equation. You can read the document called *UncertaintyinGISProcesses.pdf* for derivation of different formulae.

Finally, if you decide to print or display a map at a smaller scale than the base scale, then you would have to add another term in the above equation. Also, keep in mind this is only an approximation and there are likely other sources of uncertainty in the map-making process that are not accounted for by scale alone.

Q6: What is the location uncertainty for a raster representation from a digital geospatial layer?

Every raster data layer has a spatial resolution, which refers to the size of one cell in real-world units of area. For example, a raster with 10 m spatial resolution is composed of cells that are 10 m x 10 m, or 100 m^2, in the real world. Assuming the raster cell is square, which most are, then the cell's length and width are the same. In addition, the coordinate for a raster cell is referenced to the center of the cell.
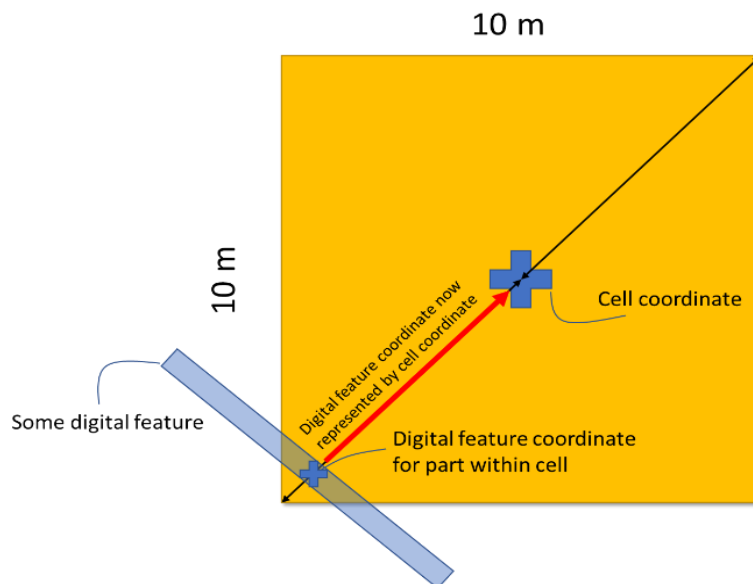
Any feature or part of a feature that has been encoded (transformed from a vector representation to a raster) to a raster could have been at most ½ the cell diagonal away from the cell center during encoding, while remaining within the cell. In such a case, the cell would be used to represent the feature or part thereof. As such, the general positional uncertainty involving the coordinate location of a feature, or part thereof, due to representation by a raster with square cells, where each cell is, for example, 100 m^2, can be given as:

$$\delta_{raster} = \frac{\sqrt{2A}}{2}$$
$$A = l \times w$$
$$l = 10m$$
$$w = 10m$$
$$\frac{\sqrt{2(100)}}{2}$$
$$= \frac{\sqrt{200}}{2}$$
$$= \frac{14.14214}{2}$$
$$= 7.071068$$

The uncertainty of position given by the above equation would be added to any other uncertainties if your raster was created from a vector digital geospatial layer. So, if you had a raster layer of roads with 25 m spatial resolution, and these originated from a vector layer of roads, that, in turn, was **digitized** from a paper map at a scale of 1:50,000, then the total positional error would be given by:

$$cell\ size = 25\ m \times 25\ m$$
$$cell\ area = 625\ m^2$$
$$raster\ error = \frac{\sqrt{2(625)}}{2} = 17.68\ m$$
$$\therefore total\ \delta = \sqrt{\text{base scale error}^2 + \text{digital error}^2 + \text{raster error}^2}$$
$$= \sqrt{12500^2 + 12500^2 + 17680^2}$$
$$= \sqrt{156250000 + 156250000 + 312500000}$$
$$= \sqrt{625000000}$$
$$= 25000$$
$$= 25\ m$$

As such we include the raster positional error within the total error. Therefore, by definition, objects or entities in raster data have greater location uncertainty than their corresponding vector representation – if indeed the raster has been created original from vector data. Note that if you were to take a raster and convert it to vector, accuracy is not increased since the raster error of position is still present.

Q7: If ArcGIS Pro does not recognize the coordinate system for a layer, how do I tell ArcGIS what the coordinate system is for the layer?

When a geospatial layer does not contain coordinate system information and it is added to ArcGIS Pro, you may see a warning, and often you just won't see your spatial data layer where you expect to find it. The figure below is a typical situation, where the data ended up near 0° latitude and 0° longitude. If you add data to your project and can't find them, this is often a good place to look for them!
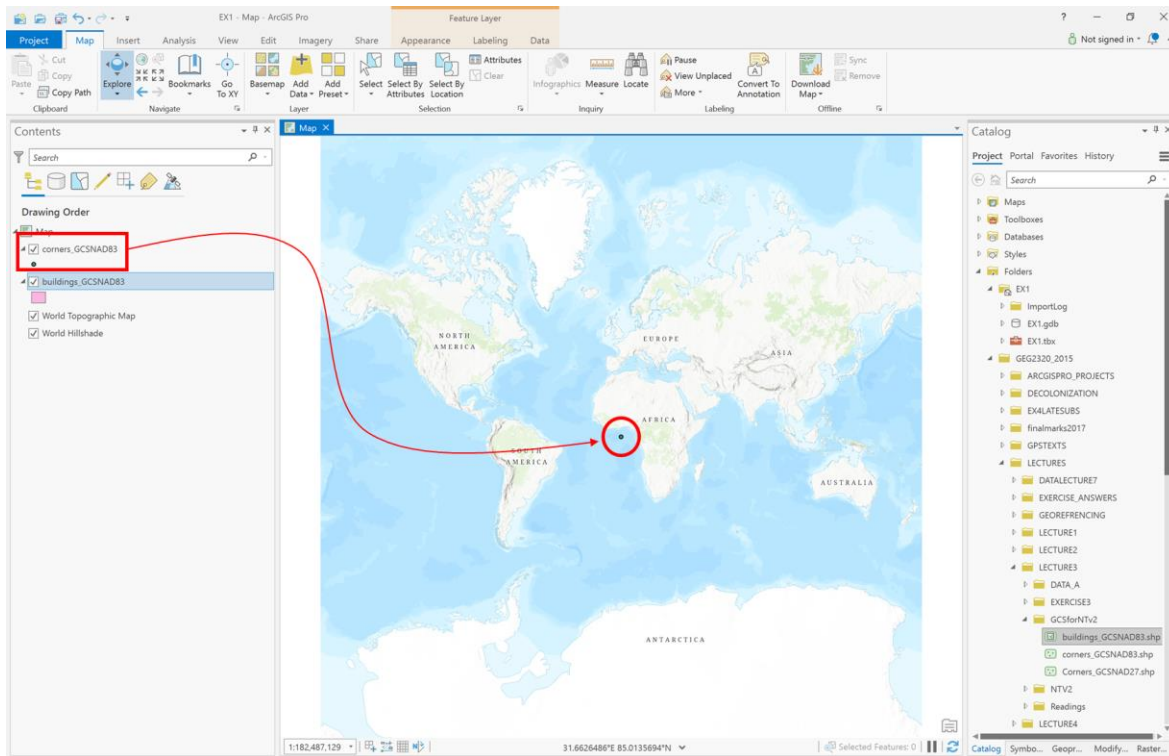
ArcGIS Pro will still display the layer, but it won't line up with other layers that have known coordinate systems.

In the GCSforNTV2 folder you have the following layers, with the following coordinate systems:

Layer	Coordinate System
corners_GCSNAD83	North American Datum of 1983
corners_GCSNAD27	North American Datum of 1927
buildings_GCSNAD83	North American Datum of 1983

1. Add the corners_GCSNAD83 layer to ArcGIS Pro in a new project.
2. Click on the "Full Extent" button in the Map tab and you will see a dot in the center of the map near 0,0 in the cartesian coordinates of the default 'WGS 1984 Web Mercator Auxiliary Sphere' coordinate system of ArcGIS Pro, just like the above figure. The small dot in the

middle of the map contains points that represent the corners of the Simard Hall Building, in Ottawa! In order to see those points in the correct location we need to define the layer's coordinate system and any geographic transformations required to bring the layer's coordinates to the coordinates of the Map so that things align correctly. To do so follow these instructions.

3. Go to Project->Options:
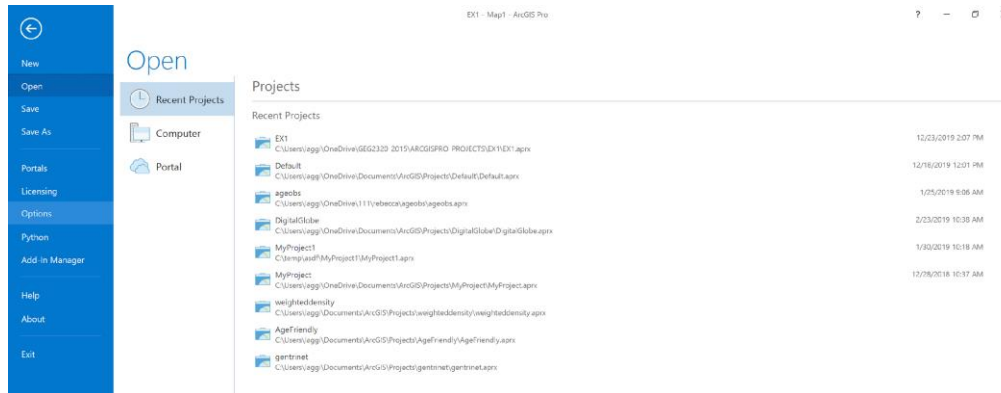
4. In the Options dialog, choose the Map and Scene item, scroll down, and put a check-mark in 'Warn if transformation between geographic coordinate systems is required to align data sources correctly'.
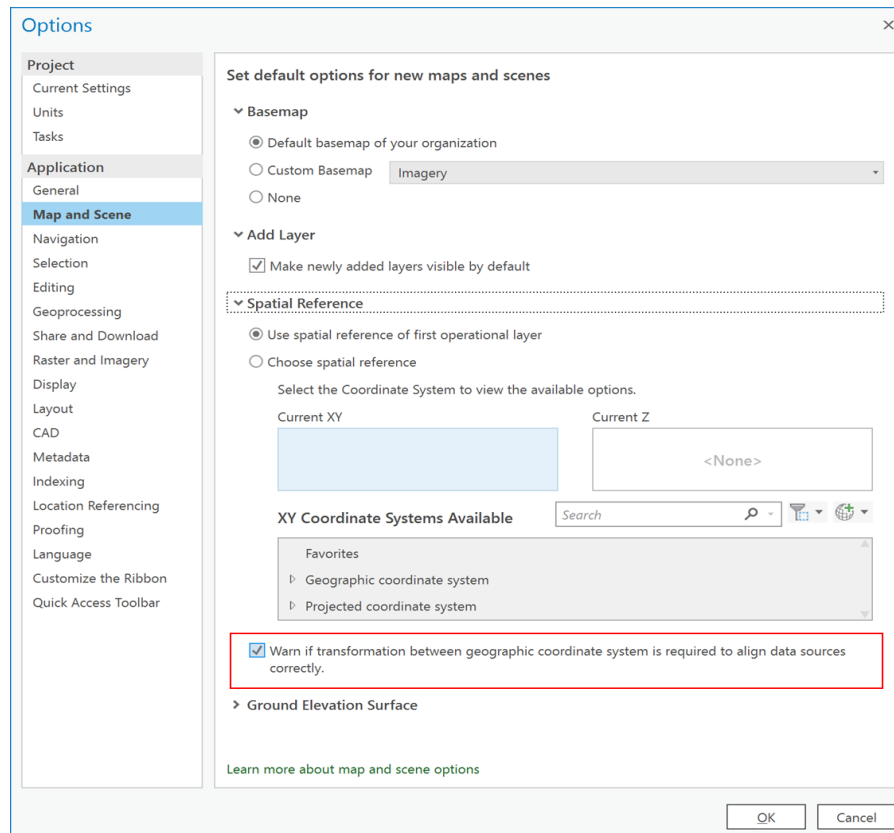
5. Go back to your Map.

6. Go to the Geoprocessing pane under Toolboxes.
7. To specify the coordinate system for the layers in ArcGIS Pro, you use the Define Projection tool under Data Management -> Projections and Transformations:
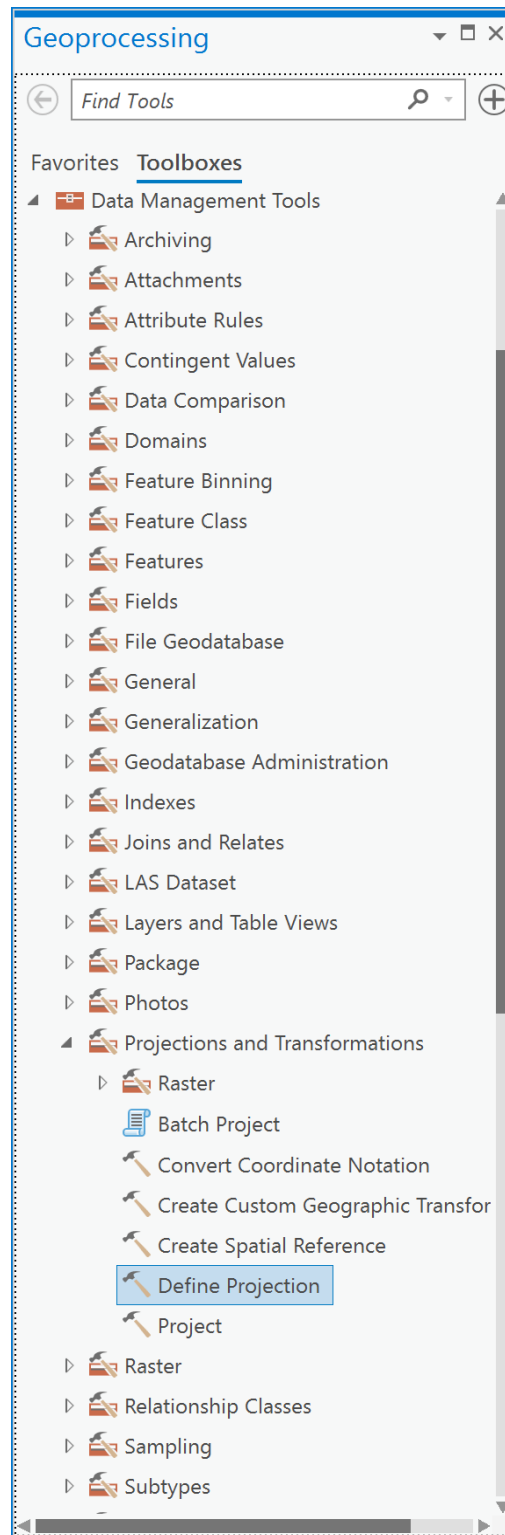
8. In the Define Projection tool, choose the layer corners_GCSNAD83 as the "Input Dataset or Feature Class":
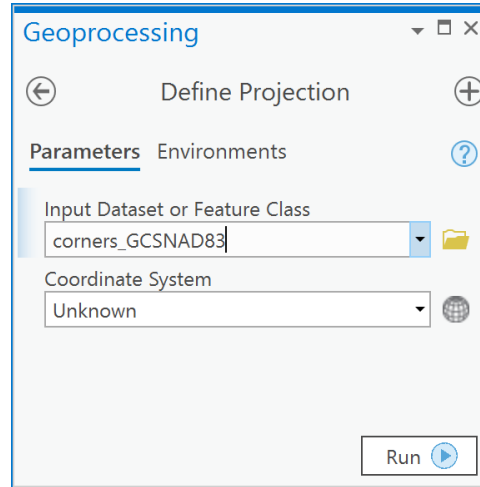
9. To the right of the "Coordinate System" input box, click on the button that looks like a globe.
10. Now, navigate to the folder Geographic Coordinate Systems -> North America -> Canada and choose "NAD 1983" and click OK.
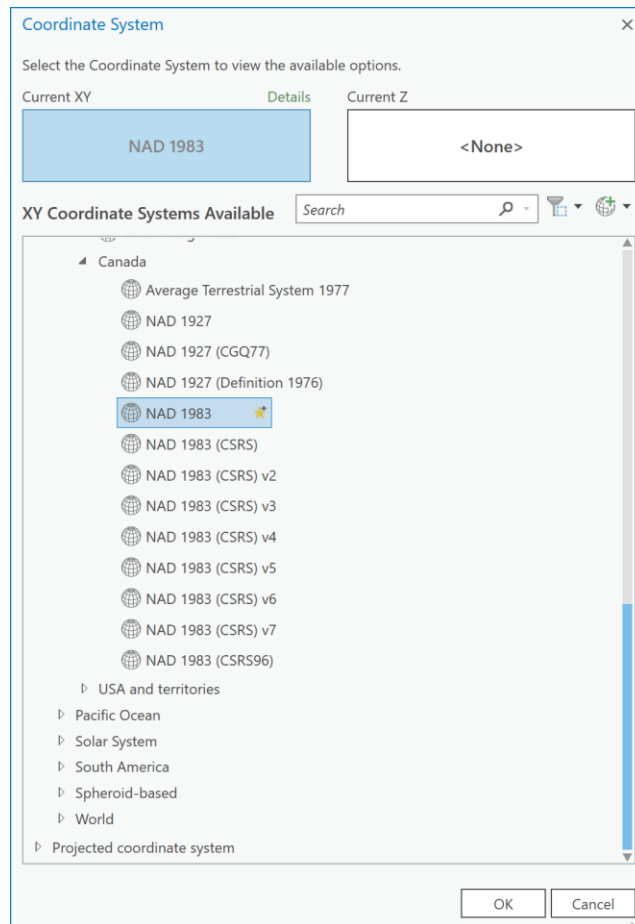
11. Your dialog will now look like this, if so, click Run:
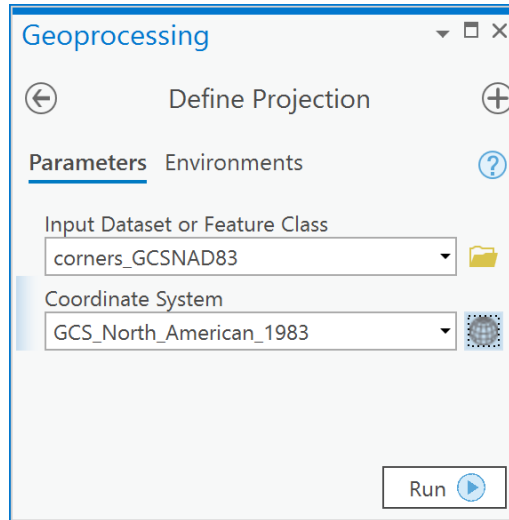
12. Once the operation is complete, you will now see the following:
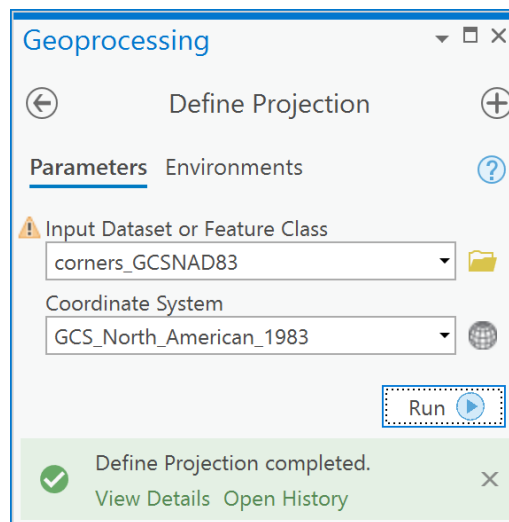
13. This means that your layer has a coordinate system defined as NAD83. Defining the projection simply tells ArcGIS what the geographic coordinate system (GCS) is for the layer. The observant student will realize that without the information in Table 1, there would be no way to know for sure what the native GCS was and, therefore, there would be no way to confidently use the Define Projection tool.

14. Now, repeat this procedure for the other two layers listed in the table, noting that corners_GCSNAD27 is in the North American Datum of 1927.

Q8: How do I convert between NAD27 and NAD83 in Canada?

If you add a layer to ArcGIS Pro that contains a geographic coordinate system that is different from the one defined for the Map, then you will see this warning:
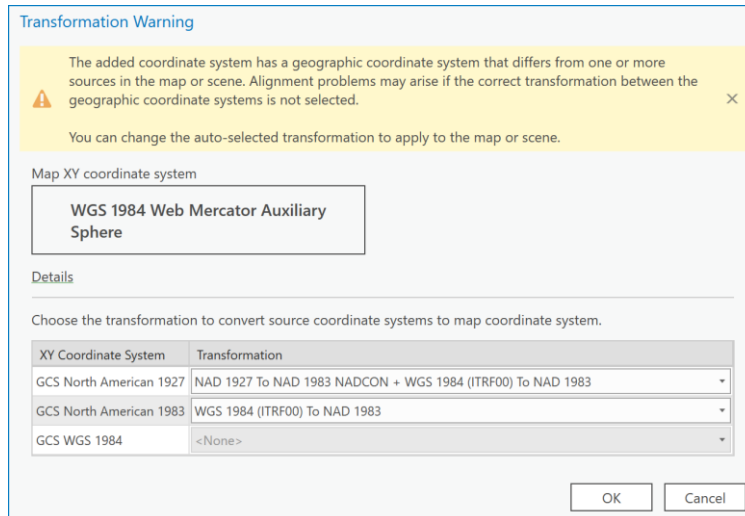
This warning means that you are trying to display two different geographic coordinate systems in the same space. ArcGIS Pro will allow you to do this if you click the 'Cancel' button. However, doing so means that the locations of common objects or entities won't line up between the layer you are adding and other layers already in your data frame. For example, in the figure below, the building corners of Simard Hall are shown in black + signs with the NAD83 coordinate system, and in red dots with the NAD27 coordinate system. The two are offset because the data shown in red dots have not been transformed correctly to the NAD83 coordinate system.
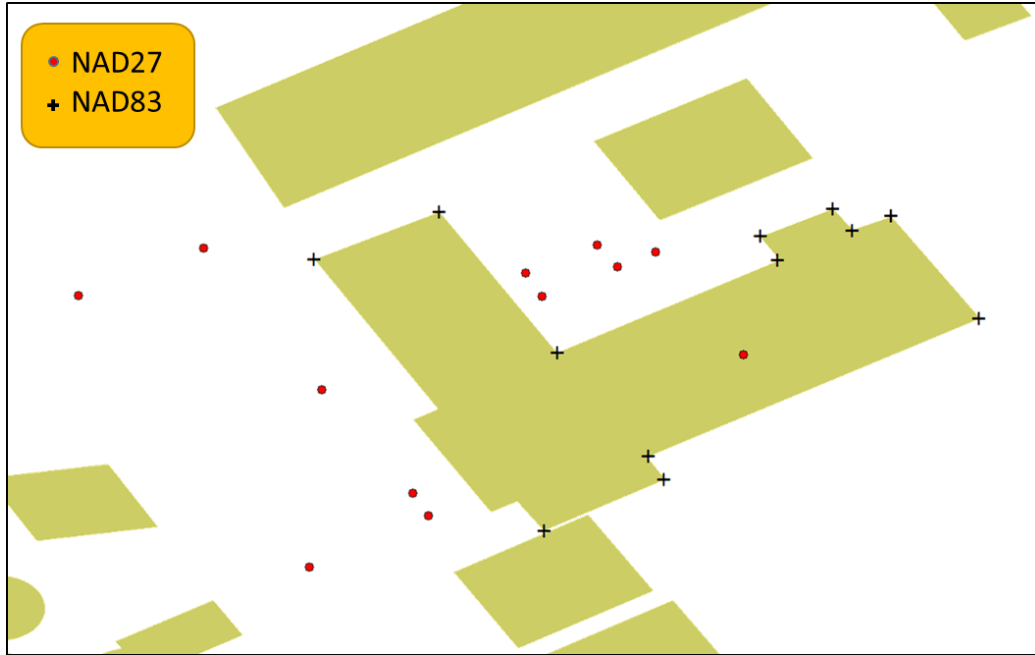
If, however, you specify the correct geographic transformation for the layers, then the data will line-up correctly, as shown below.
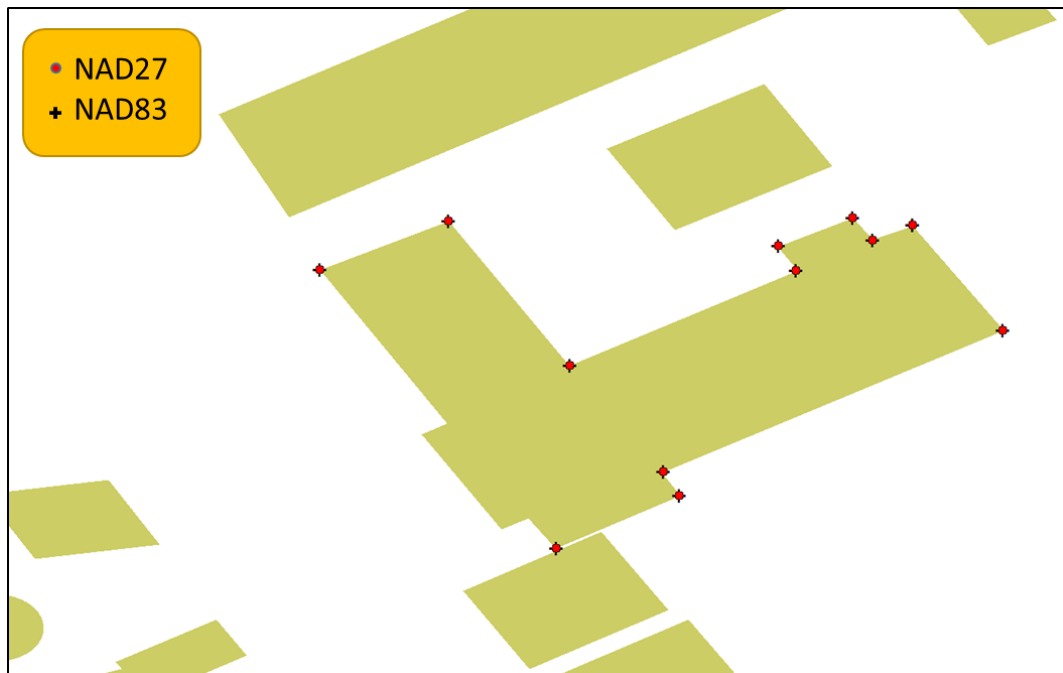
To see these effects in action,

1. Click on the Insert tab followed by the **New Map** button to create a new Map.
2. Now, from the GCSforNTv2 folder, add the layer called buildings_GCSNAD83. You will see the warning dialog, just click OK. You get this warning because the default

Longitude/Latitude system called a Geographic Coordinate System (GCS) for the basemap in ArcGIS Pro is WGS 1984 (using the WGS 84 ellipsoid for horizontal coordinates), whereas the layer you added is in the North American Datum of 1983 (using the GRS 1980 ellipsoid for horizontal coordinates). These two use different ellipsoidal models for their 3D long/lat coordinate system. Click OK.
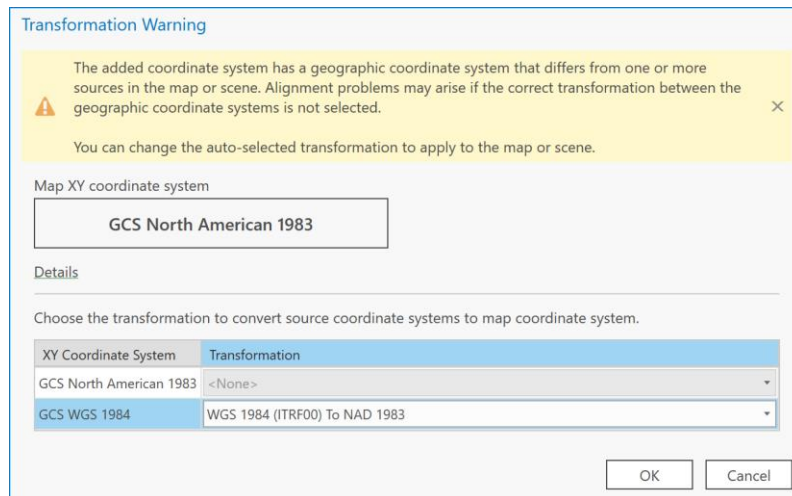
3. Now add the corners_GCSNAD83. There will be no warning dialog this time.
4. Zoom into the extent of the layer corners_GCSNAD83. You will see that the point layer called corners_GCSNAD83 lines up with the corners of the Simard Hall building:
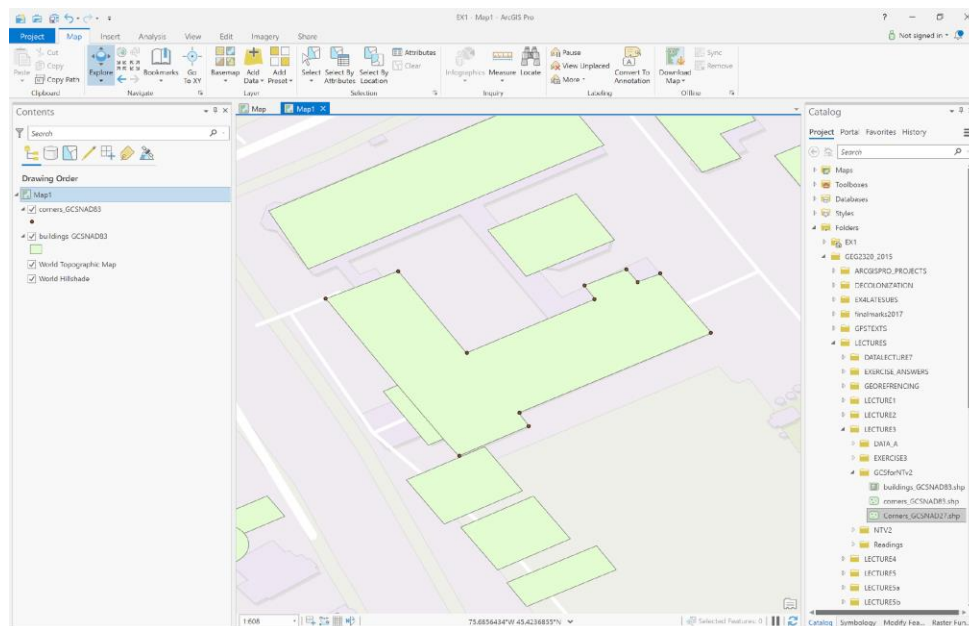
5. Then add the layer Corners_GCSNAD27. You will see a warning dialog that specifies that the GCS of the added layer, namely Corners_GCSNAD27, is different from the GCS of the map. That is because the Corners_GCSNAD27 has a geographic coordinate system of the North American Datum of 1927 (NAD 27), whereas the Map coordinate system was defined by the coordinate system of the first layer added, namely, the

Buildings_GCSNAD83 which is in the North American Datum of 1983 (NAD 83). So, this dialog is asking you how to transform between NAD 27 and NAD 83:
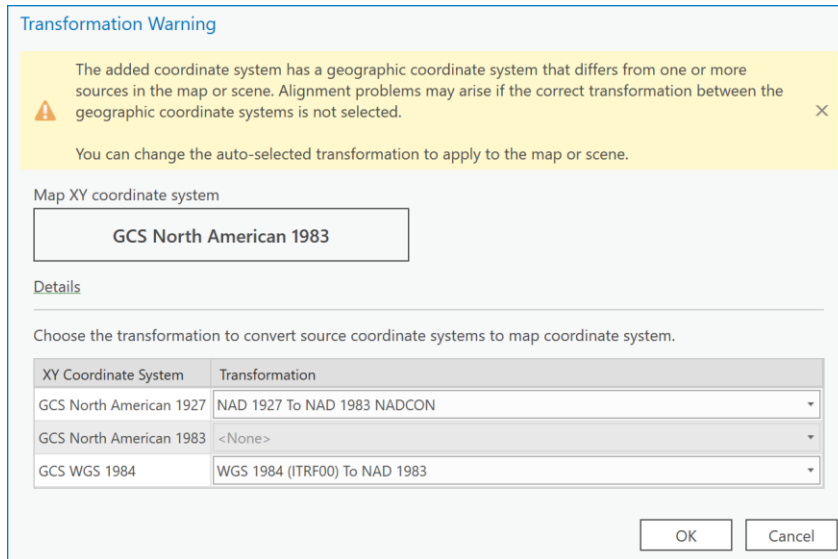
6. Click on the dropdown beside 'GCS North American 1927':
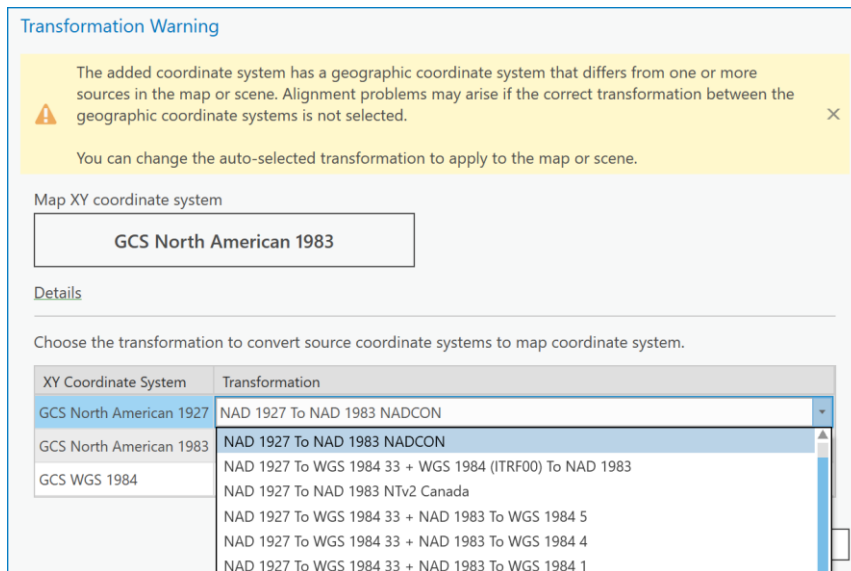
7. Choose the 'NAD 1927 to NAD 1983 NTv2 Canada' transformation.
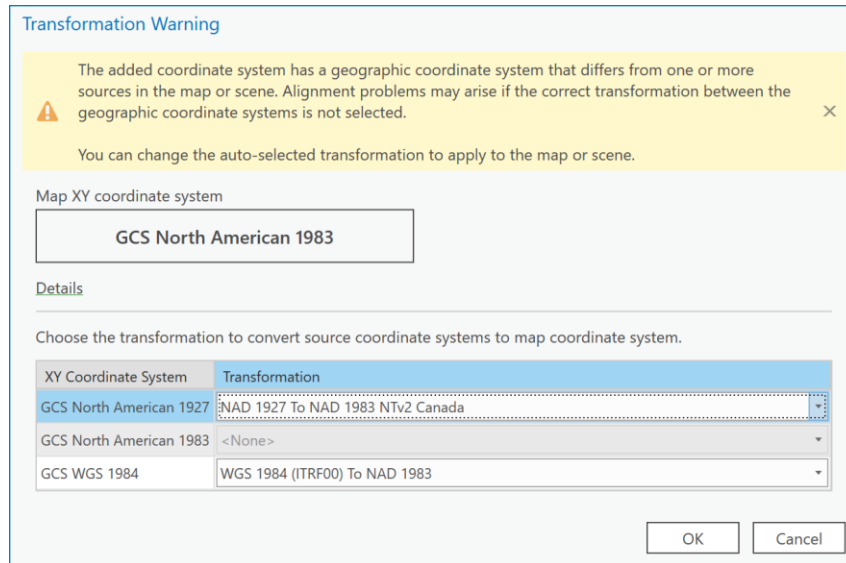
8. You will now see that all data align:
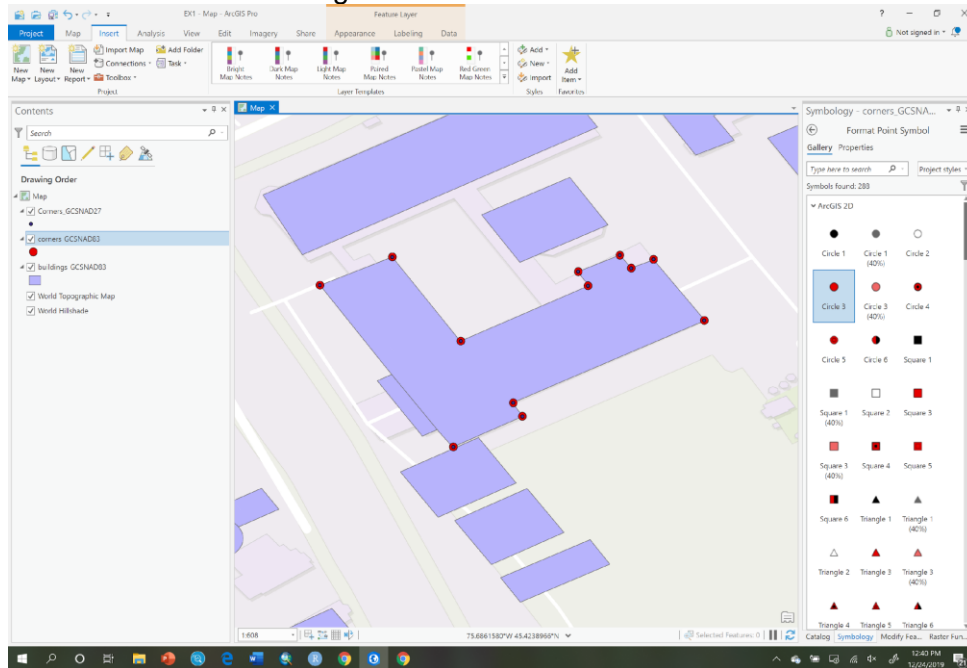
Note as well, that this process only gives a display transformation or on-the-fly transformation; the underlying data within the digital geospatial layers does not change, ArcGIS Pro has simply computed new coordinates so that all features within Corners_GCSNAD27 are displayed in coordinates of NAD83. The features stored within Corners_GCSNAD27 are still stored as NAD27, that does not change using this technique.

Q9: How do I specify the NAD27 and NAD83 conversion after I've added the data already to ArcGIS Pro?

You can specify any transformation for layers within a map in ArcGIS Pro by opening the Map properties dialog and clicking the "Transformations…" item.

1. Repeat steps 1 to 4 of example Q7.
2. Now you will see the following dialog, click "OK"
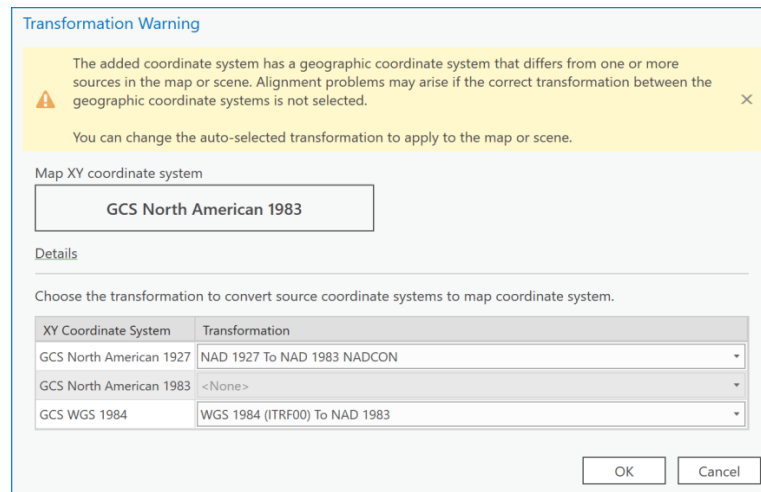
3. Your map will look like the following. When you clicked OK, you automatically chose to convert the NAD27 GCS to NAD83 by using the NADCON transformation. That transformation is only good for the lower 48 US states. NADCON is not defined for Canada. While NADCON is better than nothing, you can see that the 'Corners_GCSNAD27' layer does not align with the other two layers. So we need to specify the correct transformation for Canada, which is the NTV2 transform. This is easy to correct in the next step.
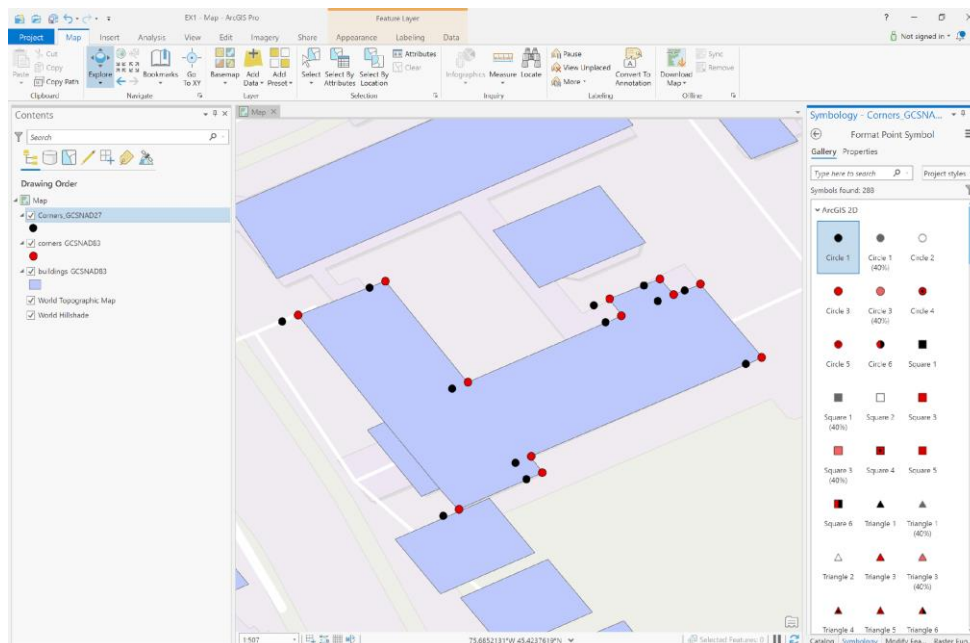
4. Right click on the Map and choose Properties
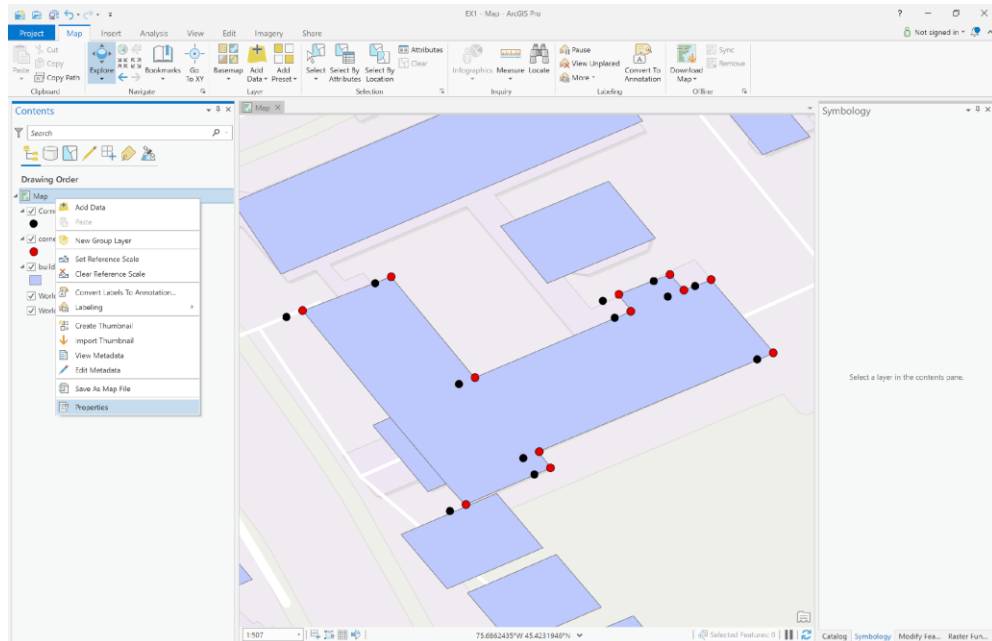
5. In the Map Properties dialog, choose the *Transformation* item:
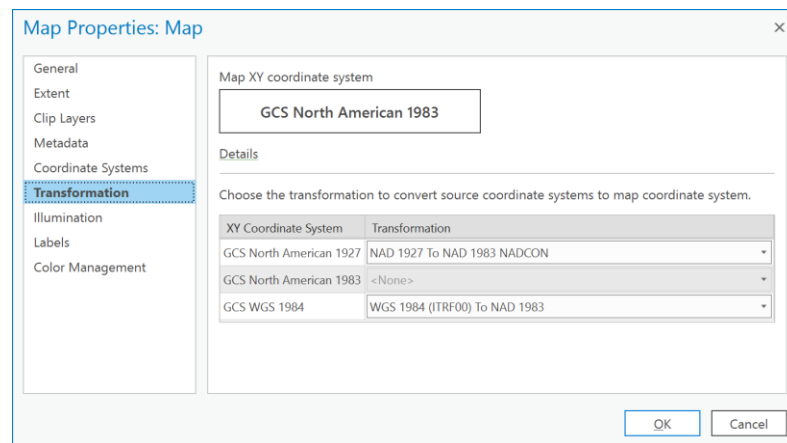
6. Click on the drop-down menu beside 'GCS North American 1927' and choose the 'NAD 1927 to NAD 1983 NTv2 Canada' transformation and then click OK:
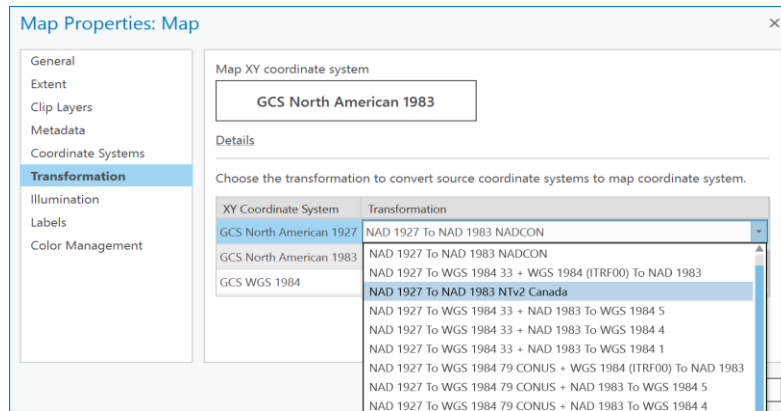
7. Now everything aligns:
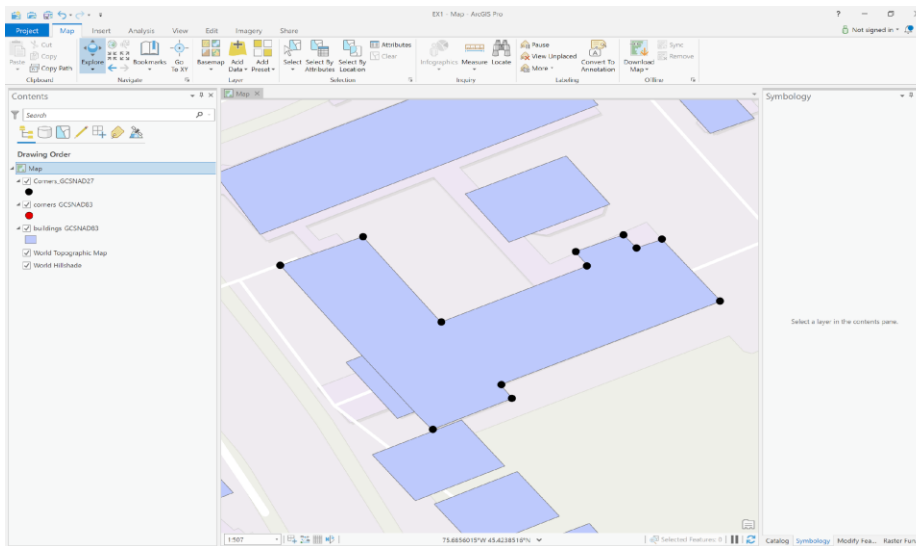
Q10: How do I get coordinates of both NAD27 and NAD83 for each point in a set of points?

To acquire coordinates for a set of points in a predefined coordinate system, there are three methods:

1. Use the identify tool and click on a point and examine the coordinates.
2. Use the 'Add XY Coordinates' tool in Data Management Tools -> Features (the easiest).
3. Use the Add Field Tool in the Geoprocessing Fields toolbox and then calculate geometry in those fields in the native coordinate system of the layer(s).

The second and third approaches allow you to acquire coordinates for more than one point at a time with high precision, so you will likely use either of those most of the time. To build additional GIS skills, this example will use the third method, which takes a bit longer but uses more tools that are useful in other contexts.

1. Create a new Map.

2. Add the Corners_GCSNAD27 layer to ArcGIS Pro. Click OK to any dialog that comes up. We are not concerned with alignment here but rather with adding coordinates in the native coordinate system, so it doesn't matter what transformation we choose.
3. Go to Data Management Tools -> Fields and right-click on the Add Field tool and choose Batch
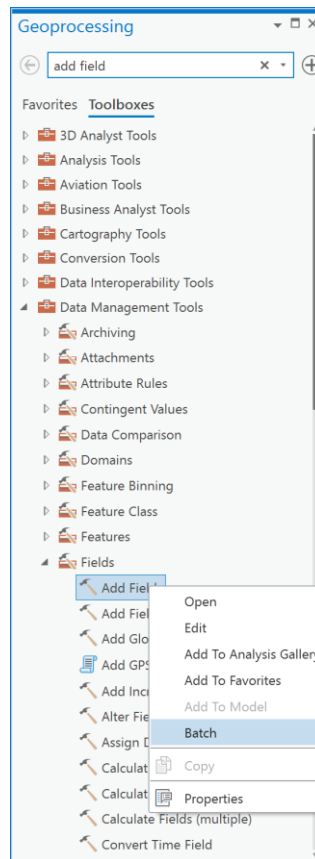
4. You will now see the batch processing dialog. In the 'Choose a batch parameter' dropdown choose 'Field Name' and click Next:
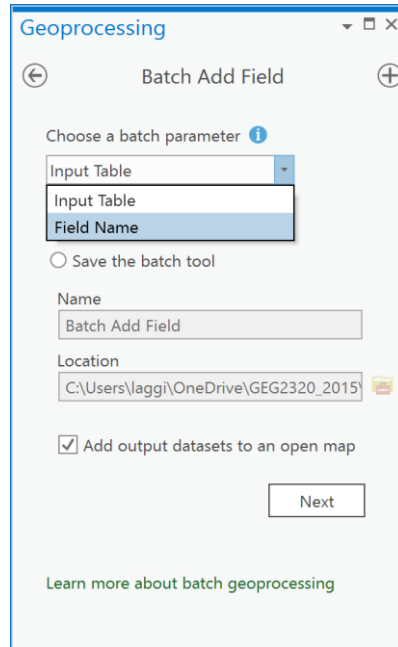
5. Now you will see Parameters dialog. In this dialog choose 'Corners_GCSNAD27' from the dropdown under 'Input table'. Under 'Batch Field Name' type in LON27 and LAT27 as you see below. Finally, under 'Field type' choose 'Double' and click Run:
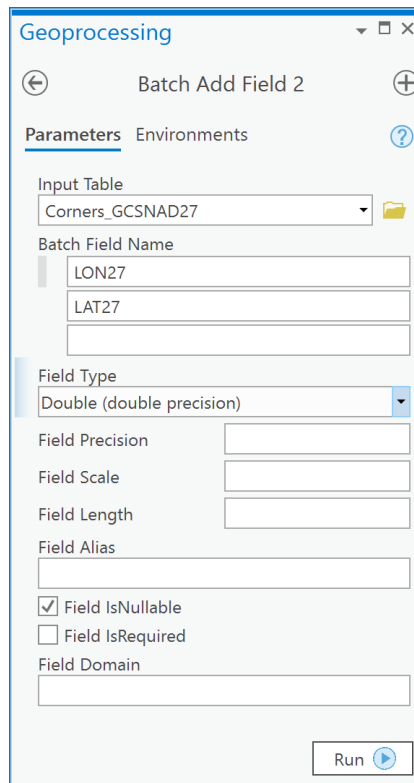
6. Now, right click on the 'Corners_GCSNAD27' layer and choose Attribute Table. You will see that there are now two new attributes, one called LON27 and one called LAT27.
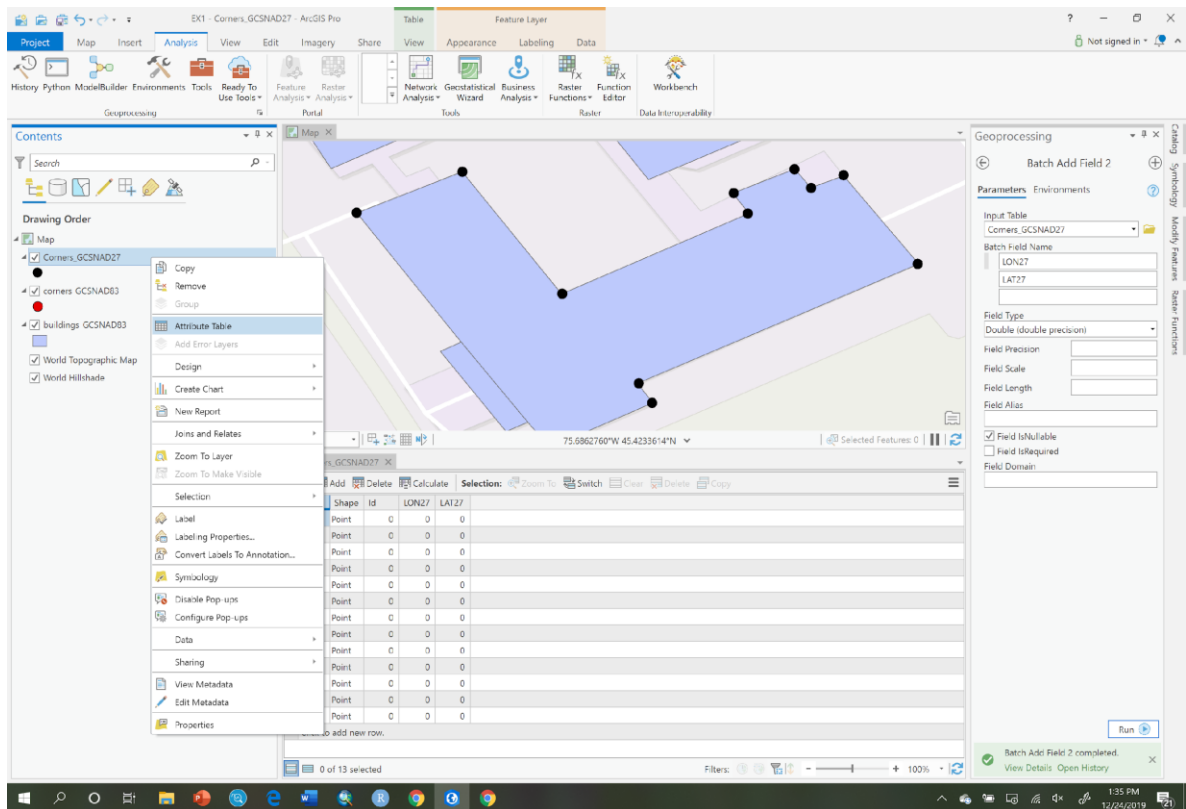
7. Right click on the LON27 attribute and choose 'Calculate geometry':
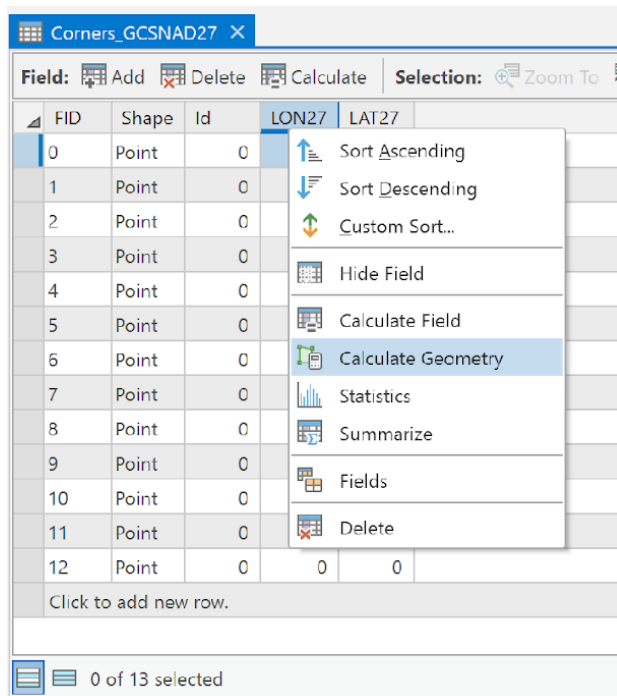
8. Under Geometry Property-Target Field, choose LON27 and then LAT27. Beside LON27 in the Property dropdown, choose Point x-coordinate and beside LAT27 choose Point y-coordinate. Under the Coordinate System, choose 'Corners_GCSNAD27' and the coordinate system will then specify GCS_North_American_1927 as the coordinate system to calculate the GCS values within. Your dialog should look like the example below, if so, click Run:
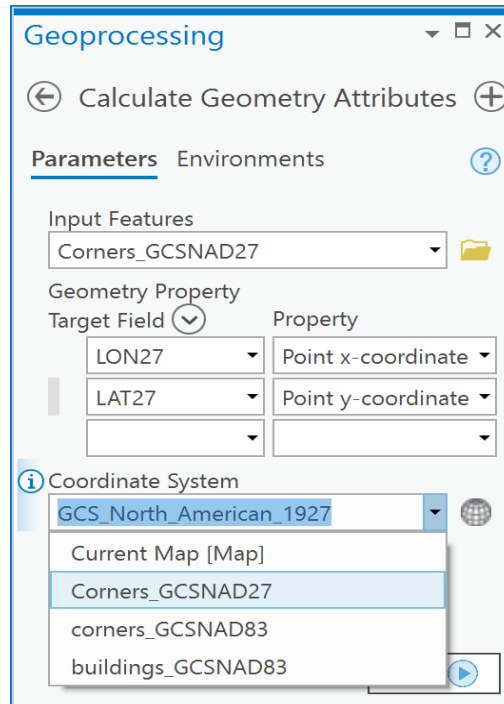
9. You should now see the following in your Attributes pane:

10. You have now calculated the coordinates of each building corner in NAD 27 Longitude and Latitude.
11. Now change the display map coordinate system to NAD 83 by right clicking on the Map in the Table of Contents and choosing Properties. In the Properties dialog choose the Coordinate System item. Then choose Geographic Coordinate System -> North America -> Canada -> NAD 1983:
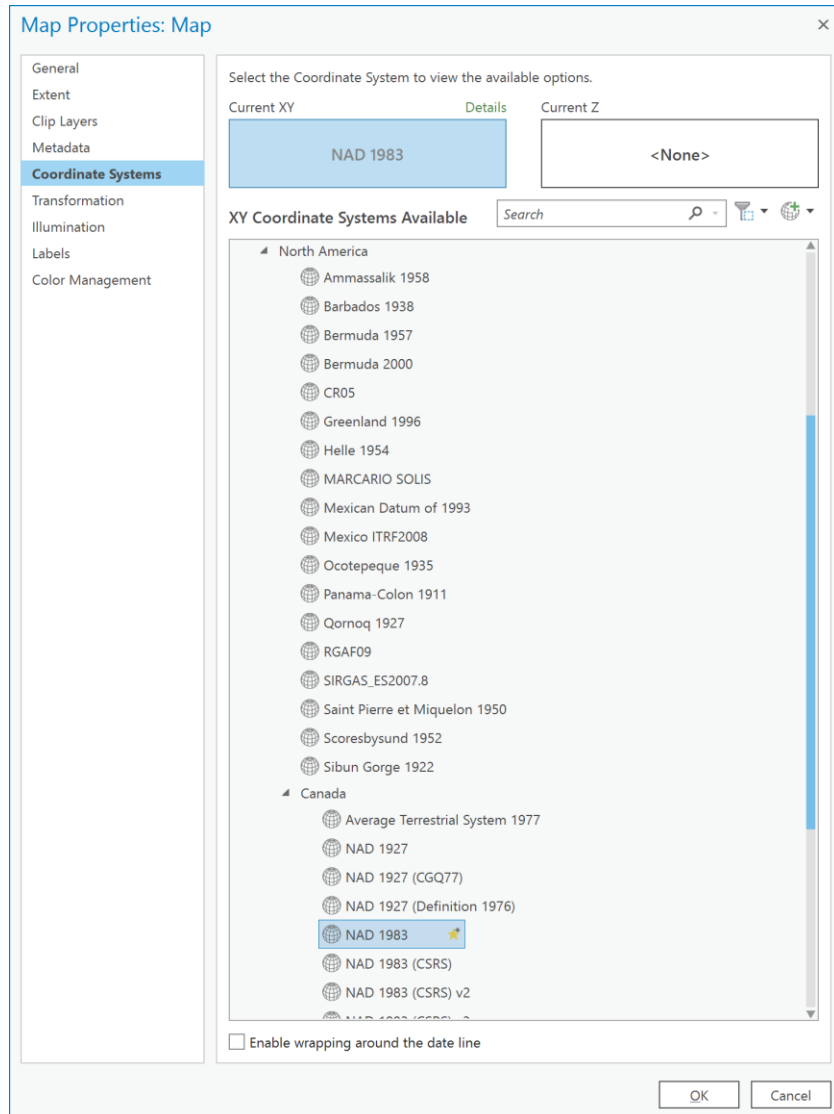
12. Now switch to the Transformation item and choose 'NAD 1927 to NAD 1983 NTv2 Canada' and click OK.
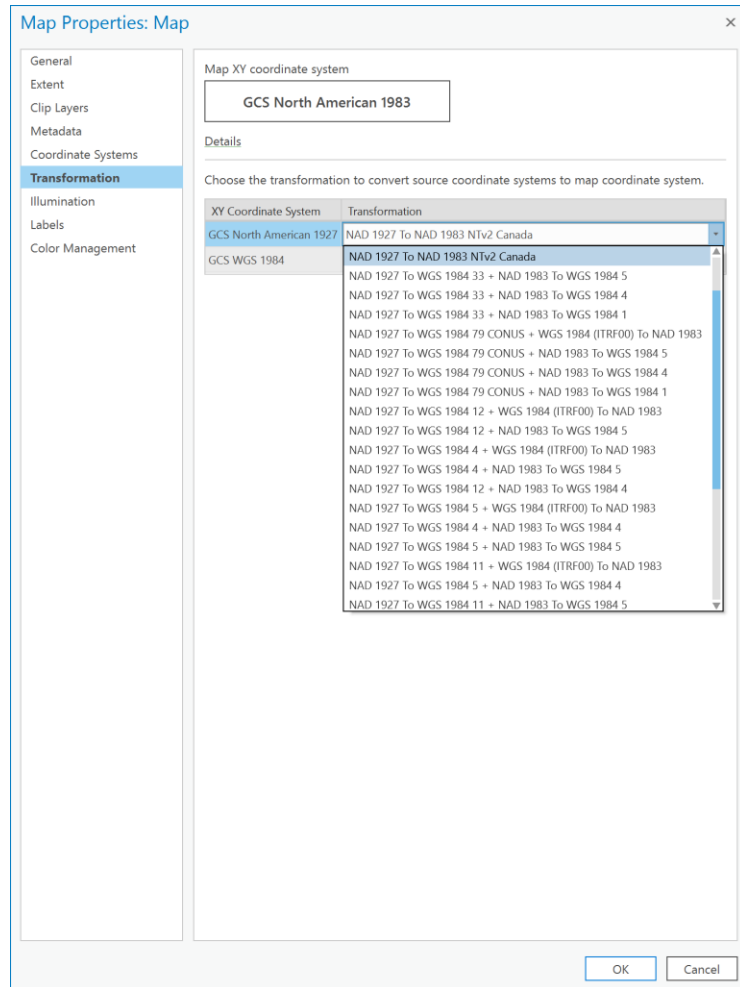
13. Next, repeat steps 3 and 4. You will see the Parameters dialog. In this dialog choose 'Corners_GCSNAD27' from the dropdown under 'Input table'. Under 'Batch Field Name' type in LON83 and LAT83 as you see below. Finally, under 'Field type' choose 'Double' and click Run:
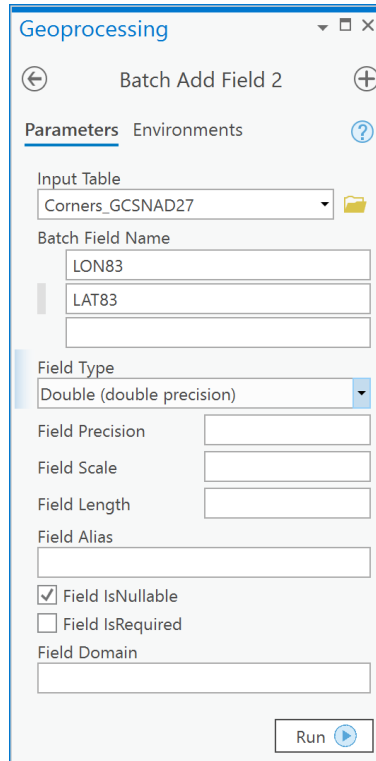
14. Next, in the Attributes pane, right click on the LON83 field and choose 'Calculate Geometry':
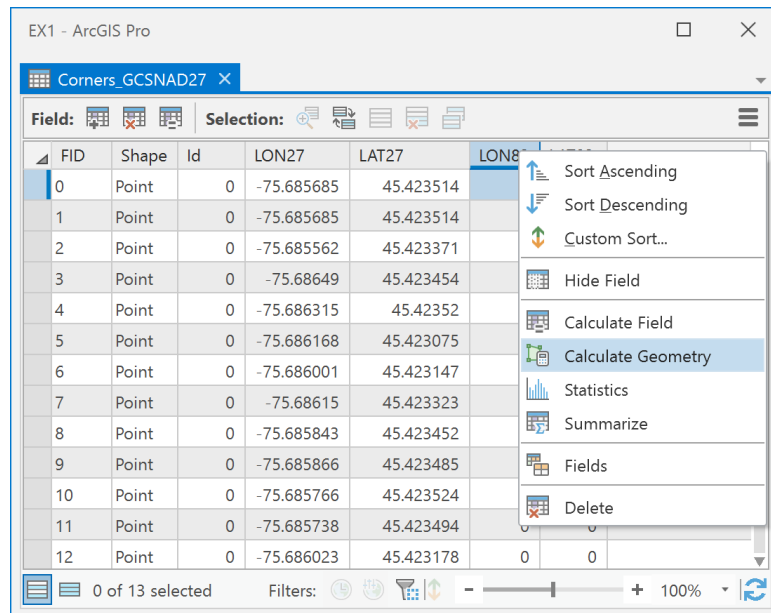
15. In the Calculate Geometry Attributes dialog, fill in the following parameters to calculate the coordinates in NAD83. Click Run once your dialog looks like the one below:
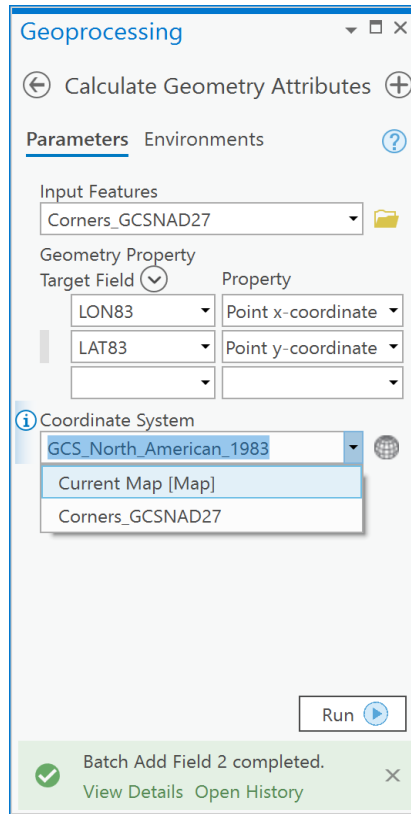
16. Now you will see that the NAD27 coordinates and NAD83 coordinates differ in the 3-5th decimal places:

Q11: How do I make a permanent change into NAD83 (or any coordinate system) for a digital geospatial layer?

As noted earlier, a display transformation is not a permanent change of the coordinates stored within a file. To permanently change coordinates there are two methods:

1. Export the current layer from the TOC using the 'Feature Class to Feature Class' Geoprocessing tool (also under Conversion Tools -> To Geodatabase) and specify the geographic transformation within the Environment settings of the Feature Class to Feature Class tool.
2. Use the Toolbox tool called 'Project' under Data Management -> Projections and Transformations. This method is often preferred because it is made specifically to change coordinate systems of vector spatial data.
3. In the Geoprocessing pane, open the Data Management Tools -> Projections and Transformations -> Project tool:
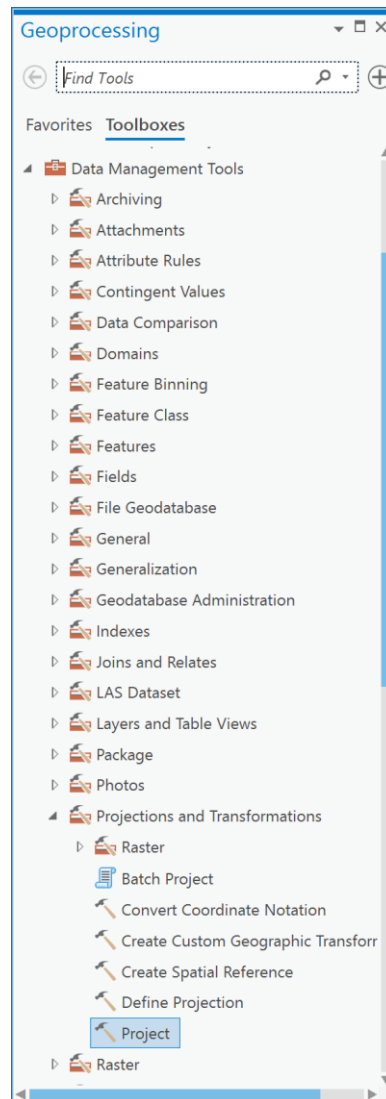
4. In the Project tool dialog, under Input Dataset of Feature Class, choose the 'Corners_GCSNAD27' layer. Under Output Dataset or Feature Class, name your output 'NewCornersNAD83'. Under output Coordinate System, choose GCS_North_American_1983. Finally, under Geographic Transformation, choose 'NAD 1927 to NAD 1983 NTv2 Canada' and click Run:
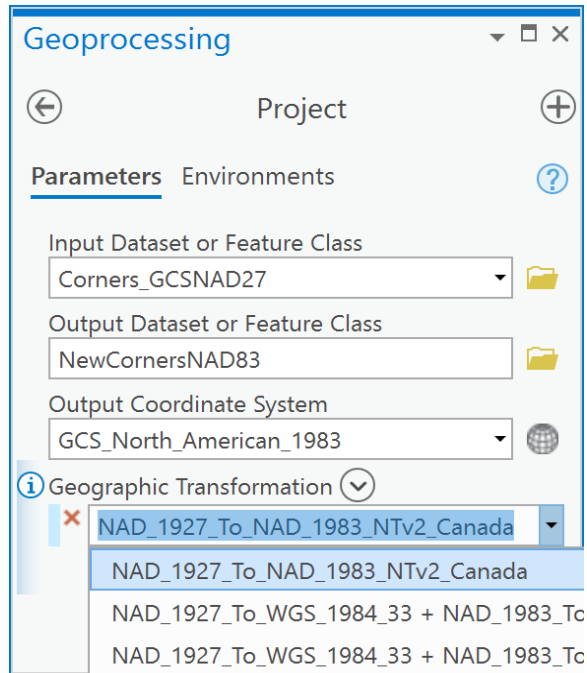
5. Now, we will see if everything worked as expected. Create a new Map. Add the 'Buildings_GCSNAD83', 'Corners_GCSNAD83' and the 'NewCornersNAD83' to the map. They should all align without specifying a transformation:
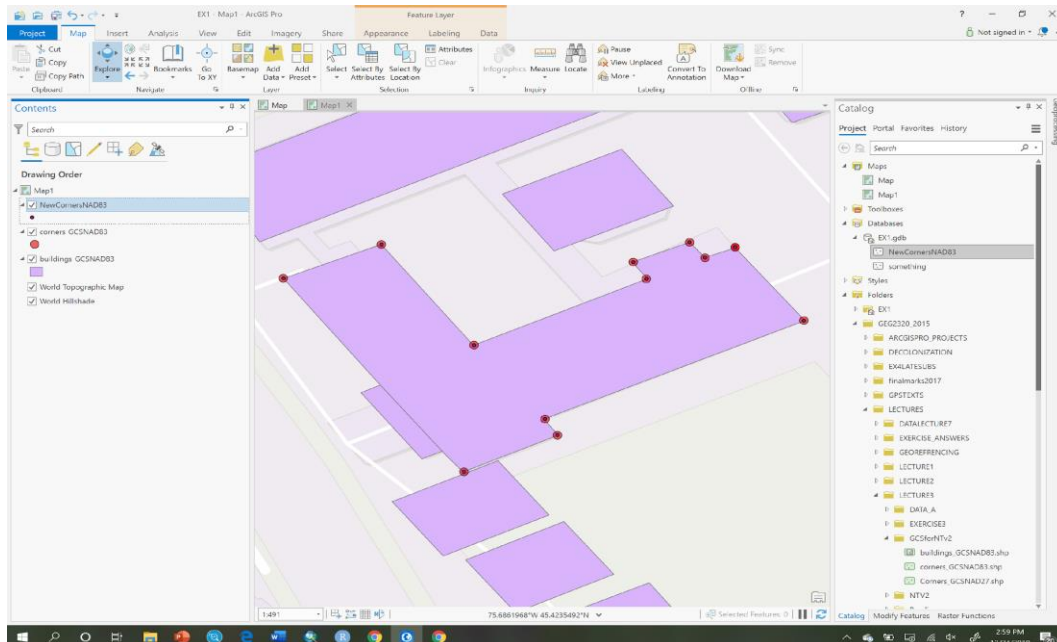
Questions to hand in.

1. [2 point] Provide the real-world distance of 1 degree of longitude at the following latitudes assuming a sphere as the model for the earth (see Q3 and Q4):

Latitude	Distance (km)
5	
40	
56	
65	
90	

What do you notice about the real-world distance of 1 degree of longitude, as latitude decreases towards the equator? Why?

2. [2 points] Calculate the decimal degree equivalents of the following coordinates to 5 decimal places (see Q2).

Latitude	Longitude	Latitude dd	Longitude dd
45°6'37" N	75°40'12" W		
45°12'12" N	75°54'46" W		
45°15'15" N	75°26'2" W		
45°19'27" N	76°2'52" W		
45°17'34" N	75°54'22" W		
45°18'30" N	75°35'6" W		
45°14'44" N	75°53'60" W		
45°20'31" N	76°2'51' W		
45°20'33" N	75°43'25" W		
45°21'2" N	75°45'53" W		
45°21'11" N	75°21'10" W		
45°25'42" N	75°47'23" W		
45°22'14" N	75°51'25" W		
45°25'25" N	75°57'20" W		
45°23'0" N	76°51'39" W		
45°55'55" N	75°54'28" W		
45°25'54" N	75°39'41" W		
45°18'1" N	75°29'26" W		
45°21'19" N	75°32'58" W		
45°29'1" N	76°46'4" W		

3. [2 points] Calculate the *base scale* location uncertainty for the following map scales (see Q5):

Scale	uncertainty (m)
1 :500	
1 :20000	
1 :50000	
1 :300000	
1 :4000000	

4. [2 point] Why does uncertainty increase with decreasing map scale?

5. [2 points] Calculate the location uncertainty for the following vector *digital* geospatial layers that were digitized from maps at the base scales listed (see Q5),

Base scale	uncertainty (m)
1 :5000	
1 :20000	
1 :50000	
1 :300000	
1 :4000000	

6. [3 points] Calculate the uncertainty in location for the following raster layers (*all of which have square cells and were encoded from digital geospatial vector layers*), given the information provided in the table (see Q6):

Base scale	Raster resolution in m (one side of a cell)	Uncertainty (m)
1 :1000	0.50	
1 :25000	4	
1 :63360	30	
1 :250000	60	
1 :5000000	2000	

7. [5 points] You are provided with a layer called mine_entrances.shp in lab2.zip. This data represents the locations of mine entrances in the North American Datum of 1927 (NAD27) (you must define that coordinate system using the define projection tool so that ArcGIS recognizes that it as NAD27). Create a table that shows the longitude and latitude of each point (mine entrance) in both NAD27 and NAD83. Include a fifth column in which you calculate the real-world distance (in meters) between the NAD27 and NAD83 coordinates. Examine the table you have created and explain what you observe

and why. **Hint**: You may want to sort your table using columns of interest in order to uncover a pattern.

	NAD27		NAD83		
MineID	*Longitude*	*Latitude*	*Longitude*	*Latitude*	*Distance*

8. [5 points] Read the following blog posts by Sarah Sibbett:
- https://resources.esri.ca/getting-technical/your-guide-to-canadian-datum-transformations-and-arcgis-online-part-1
- https://resources.esri.ca/getting-technical/your-guide-to-canadian-datum-transformations-and-arcgis-online-part-2
- https://resources.esri.ca/getting-technical/your-guide-to-canadian-datum-transformations-and-arcgis-online-part-3

   o What coordinate systems are most Canadian data within?
   o What is the default coordinate system for ArcGIS Pro?
   o What transformation is required to convert coordinates from NAD83 CSRS to WGS84 in Canada?
   o What transformation is required to convert coordinates from NAD83 Original to NAD83 CSRS in Ontario?
   o In Ottawa (see figure on the next page), what are the differences in meters between:
      ▪ NAD27 and NAD83 (original).
      ▪ NAD83 (original) and NAD83 CSRS.
      ▪ NAD83 CSRS and WGS84.

Notes: You can find Ontario NTv2 transformation files here: https://www.arcgis.com/home/item.html?id=7d26e16ba42040d7a4ae66a4e02d16a0. Such files would be necessary when working with municipalities that have not adopted NAD83 or WGS84 coordinate systems. To download province-specific files to transform between NAD83 and NAD83 CSRS see https://webapp.geod.nrcan.gc.ca/geod/data-donnees/transformations.php - you would need to create an account and sign in to download them. You might want to know about the 2022 change in the US for data there and can read it here: https://geospatial.blogs.com/geospatial/2018/08/new-2022-state-plane-coordinate-system-will-change-locations-by-up-to-15-meters.html
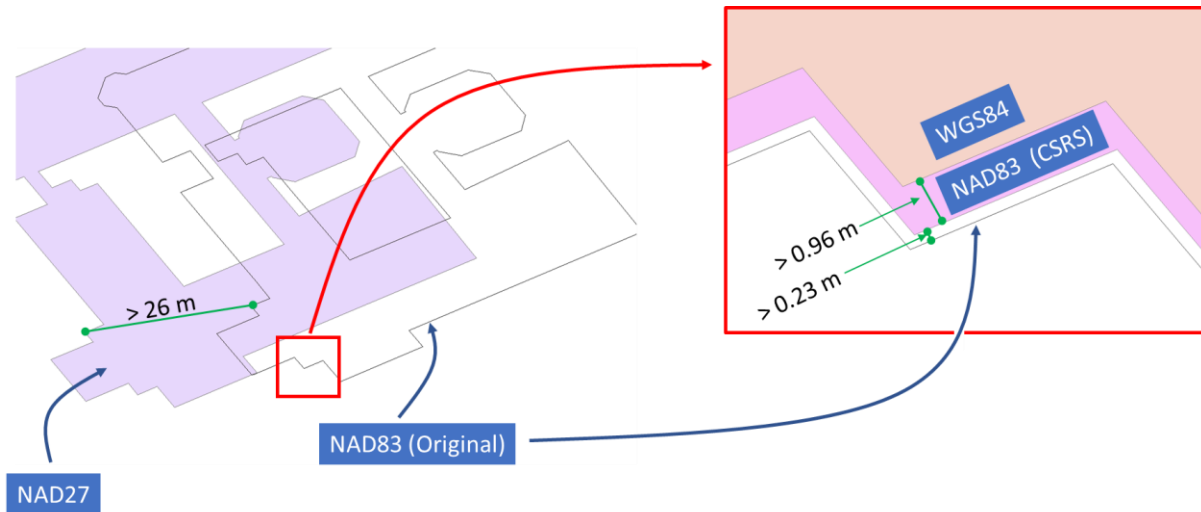
9. [4 points] Create a Map (e.g., a layout) in the coordinate system of NAD 1983 Modified Transverse Mercator Zone 9. None of the layers you will use have a pre-defined coordinate system/projection. Your map should contain the following layers that you can find in lab2.zip:

Layer	Coordinate System Metadata
buildings_1_NAD83.shp	North American Datum of 1983
buildings_2_NAD27.shp	North American Datum of 1927
buildings_3_NAD27_UTM18N.shp	Universal Transverse Mercator Zone 18 N based on NAD27 geographic coordinates
buildings_4_NAD83_CANADA_ALBERS.shp	Canada Albers Equal Area Conic based on NAD83 geographic coordinates

Student Name: _____ Student Number: _____ _____

Laboratory #3: Working with geographic databases

Purpose: The purpose of this lab is to provide you with experience in creating a geographic database.

Learning Outcomes: By the end of this lab, you will be able to:
- Create a geodatabase.
- Import data to the geodatabase.
- Create a relationship class within the geodatabase.
- Answer spatial questions using aspatial ad-hoc queries.
- Demonstrate calculating fields.
- Demonstrate map creation.

Procedure:
- Read and work through the examples.
- Answer the questions at the end of the lab.

Data/Materials:
- All data are provided in lab3.zip. You can use these data if you do not have access to Scholars Geoportal.

Assignment:
- Answer the questions (#1-6) in the spaces provided.
- The total marks available for this assignment = 27 marks for 12.5% total course grade.
- All lab assignments must be completed and handed in individually.

Format:
Students will hand in questions with completed answers.

Tables and figures should be produced to the standards of the Department of Geography, Environment and Geomatics, University of Ottawa.
For guidelines on proper formatting and presentation of tables and figures, please see the introductory section of this lab manual.

Student Name: _____ Student #: _____

Q1: How do I download data from the Scholars GeoPortal?

The Scholars GeoPortal is a hardware and software infrastructure that gives Ontario-based students and researchers access to large-scale geospatial datasets. These include land-based vector data (water, cultural features, etc.), census data, and some raster imagery.

You can only download certain data from Scholars Geoportal when you are authenticated through a university in Ontario, either by accessing the data on campus or by using a VPN. If you can, you should follow the instructions below to download the data for this exercise. Learning how to download data from Scholars GeoPortal is very useful.

Note: if you cannot access Scholars GeoPortal, you can find all the same data sets in the file called lab3.zip, you just need to sort them out based on the filenames.

To download the census tracts for Ottawa-Gatineau that you need for this exercise:

1. Go to http://geo2.scholarsportal.info/

2. In the Search tab, search box type "Census tracts 2011" and click the Search button:
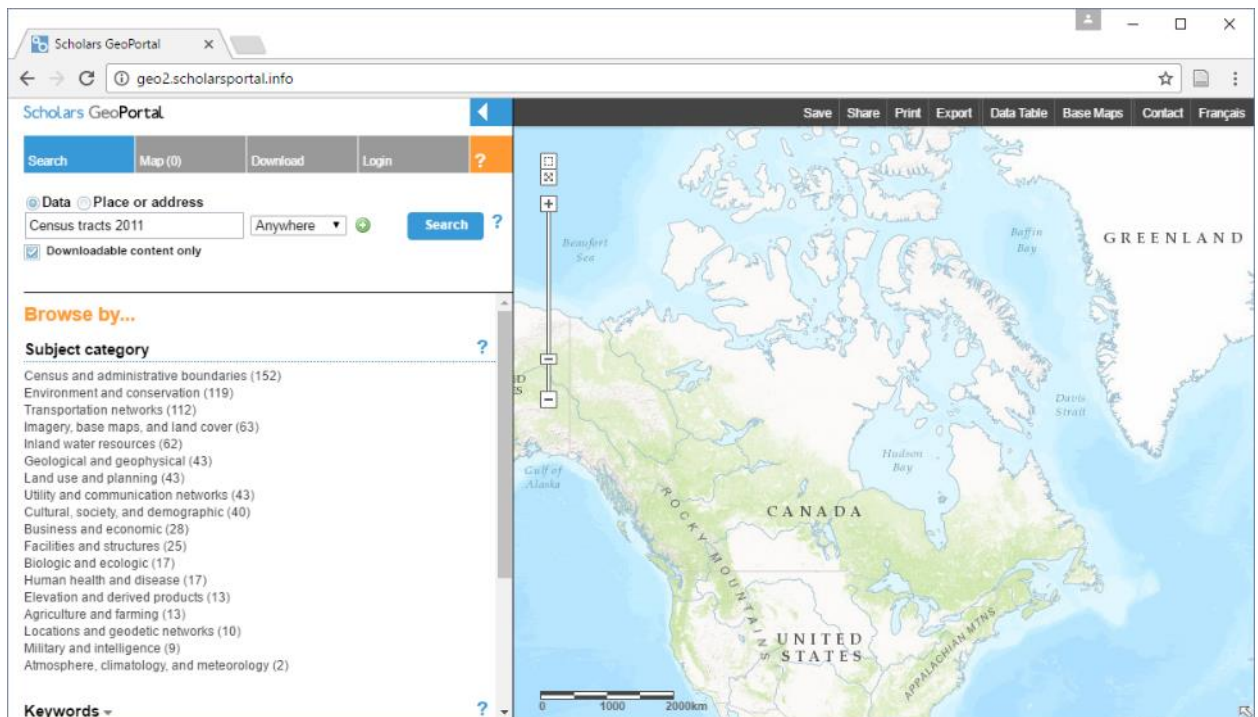
3. When the search is complete, you will see the following list of items (note that the list may not be in the same order as you see below). Find the "Cartographic Boundary Files (CBF), 2011 Census" item:
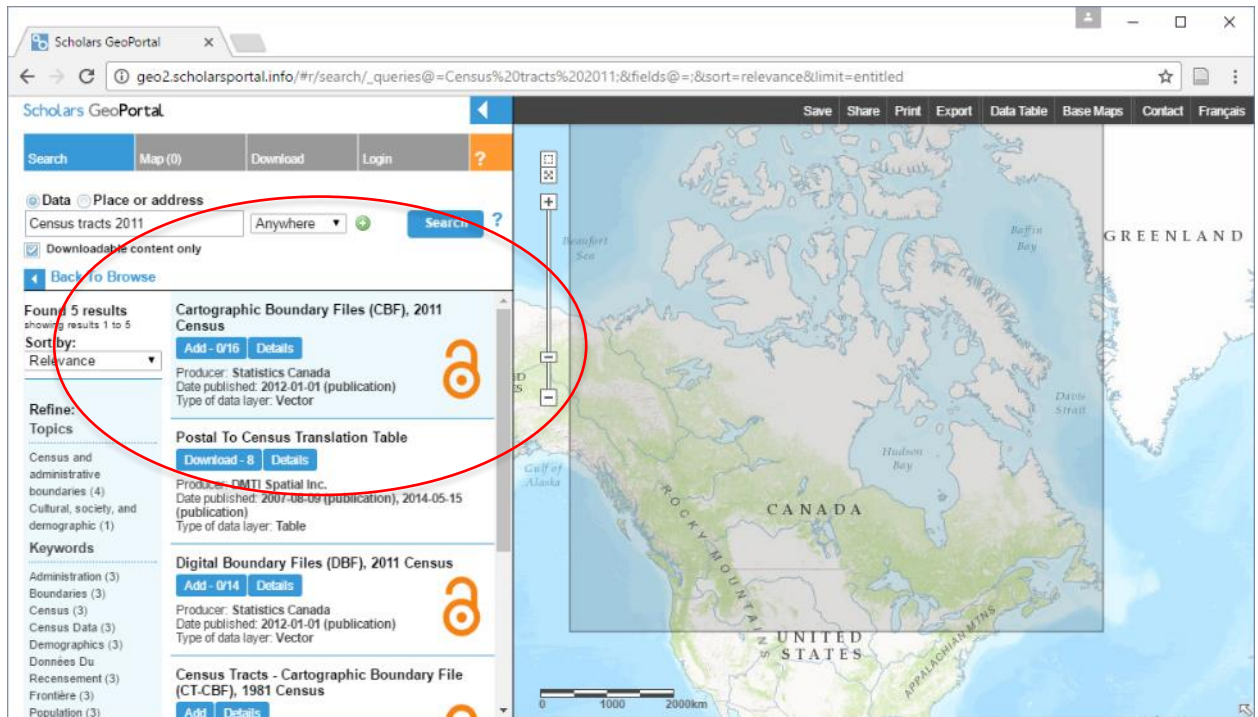
4. For the "Cartographic Boundary Files (CBF), 2011 Census" click on the Add – 0/20 button and you will see another list. In this list, look for the item called "Census Tracts *Vector*" and click the Add button and your map will look something like this:
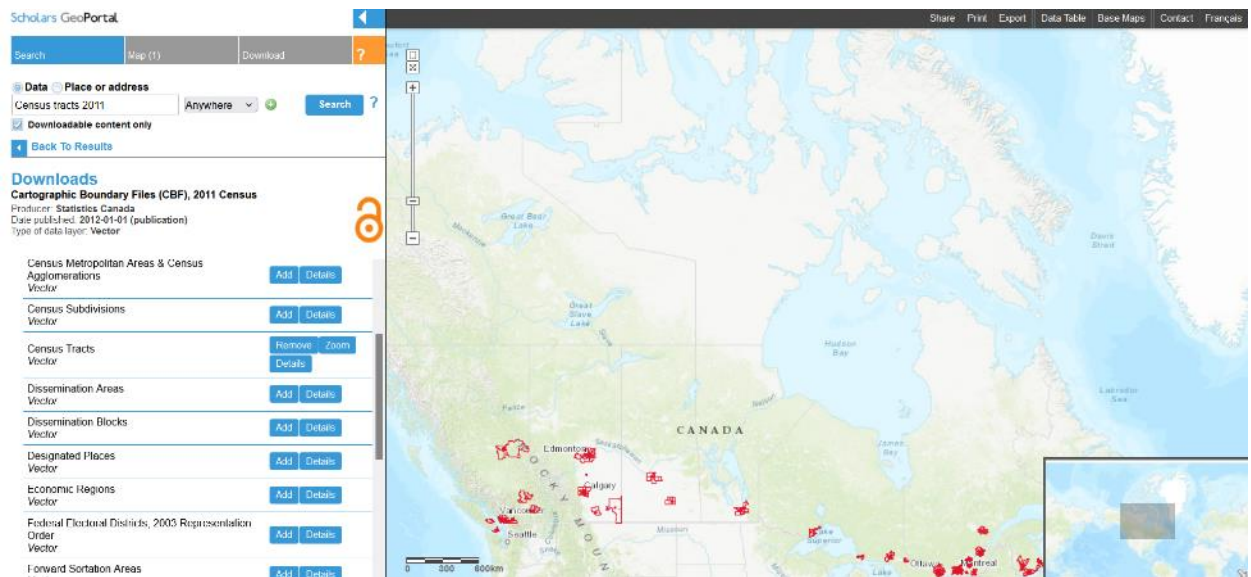
5. You can zoom into the Ottawa region using the zoom tools:
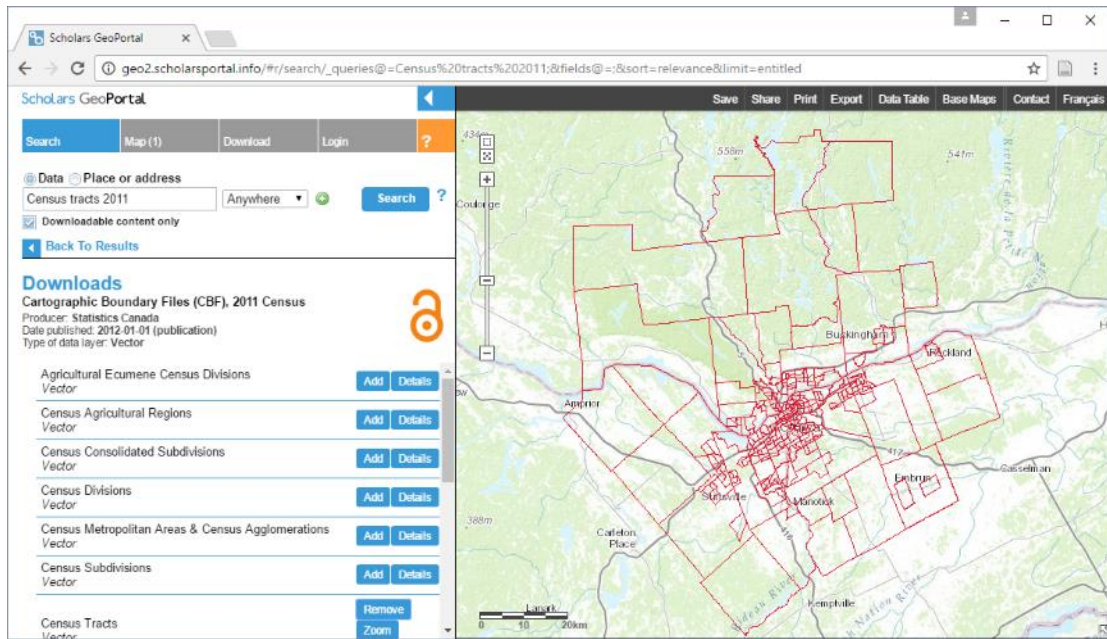
6. Now, click on the Download tab at the top of the page.

7. Do the following:
   - Select "Download by area of interest".
   - Select "Draw an area".
   - Click on the rectangle tool.
   - Click and drag a box around the Ottawa-Gatineau region so that all red lines are completely contained within the box you draw.
   - In the "Choose an output file format" (vector data only)" section, choose File Geodatabase – (.gdb).
   - In the "Choose an output coordinate system (vector data only)" section, choose UTM – Zone 18N.
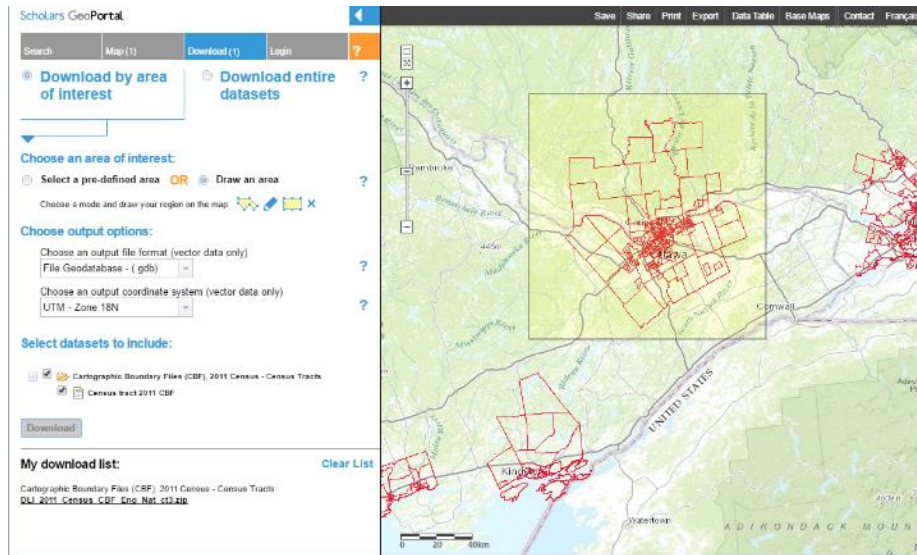
8. Click on the Download button. When the download is prepared, you will see a .zip file for download in the 'My download list': section.
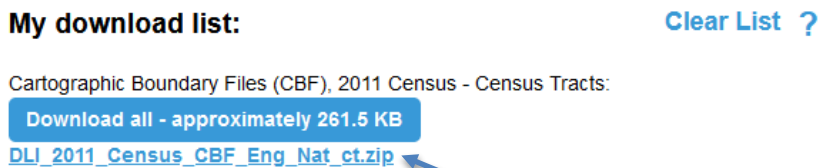
9. Download the file by clicking the **link with the bolded and underlined blue text**, and unzip the file to a location that you will remember.

10. Open ArcGIS Pro and add the folder you unzipped in the previous step.

11. In the **data.gdb** geodatabase within that folder, you will find a single feature class that will be called something to the effect of "DLI_2011_Census…. In the Catalog pane, rename that feature class to "Census_tracts" (right click on the Feature Class and choose "Rename")
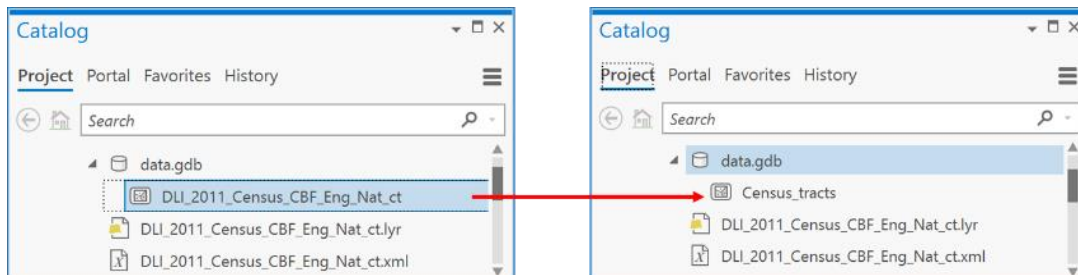
12. Now, add "Census_tracts" to a new map:

Student Name: _____ Student #: _____
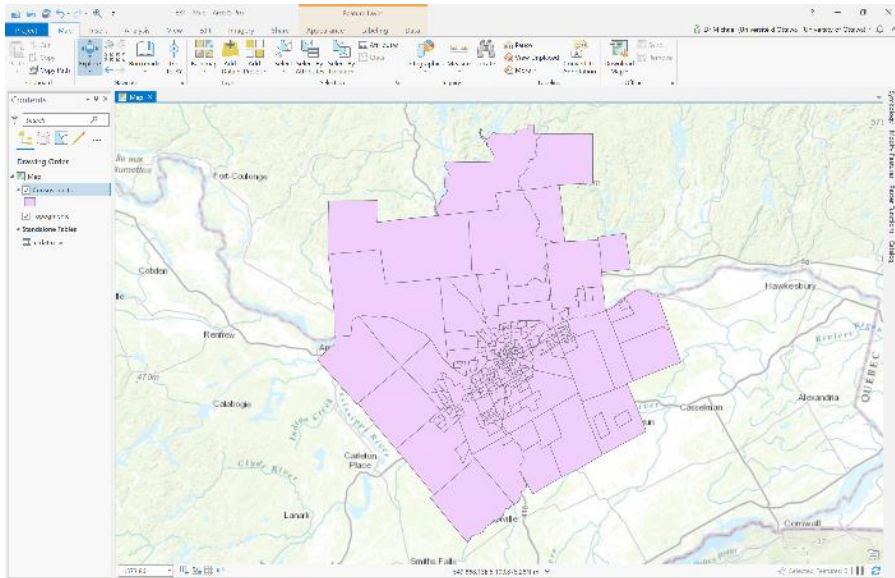
13. Open the attribute table. In the attribute table, you can see that the columns or fields only identify each of the census tracts. In particular, the **CTUID** field is important because it contains a unique identifier for each census tract.

Q2: Where do I get data to go with the Census Tracts?

Data for each census tract can be accessed through the Canadian Census Analyzer web-tool, which is available from campus or when you are authenticated with the uOttawa VPN. The census analyzer is not a data source, but rather an interface or web program. The data source is always the institution, organization or person that created the data, in this case Statistics Canada.

Note: As for the data from Scholars GeoPortal, if you cannot access data from the Canadian Census Analyzer, you can find the data you need in the lab3.zip file.

Student Name: _____ Student #: _____

1. Go to http://datacentre.chass.utoronto.ca/census/

2. Next, under Census Profiles Tables, choose 'Census Tract'
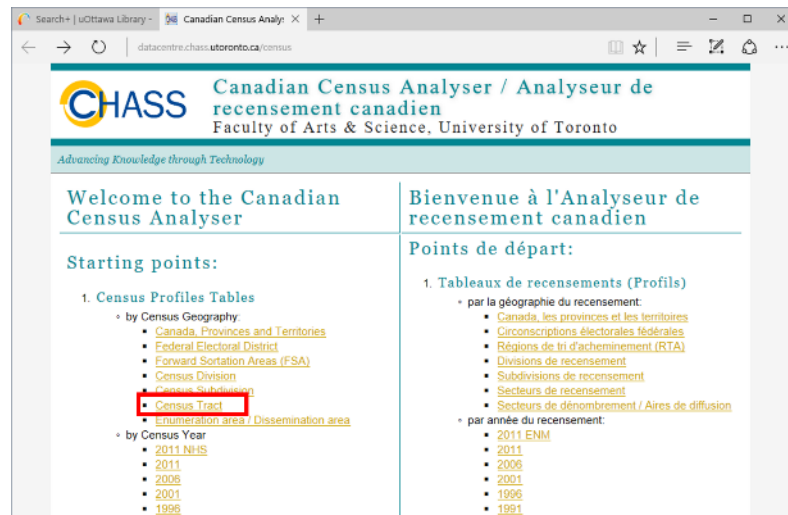
3. Now you will see a list of dates, and we are interested in the 2011 Census data. There are two types of 2011 census data:
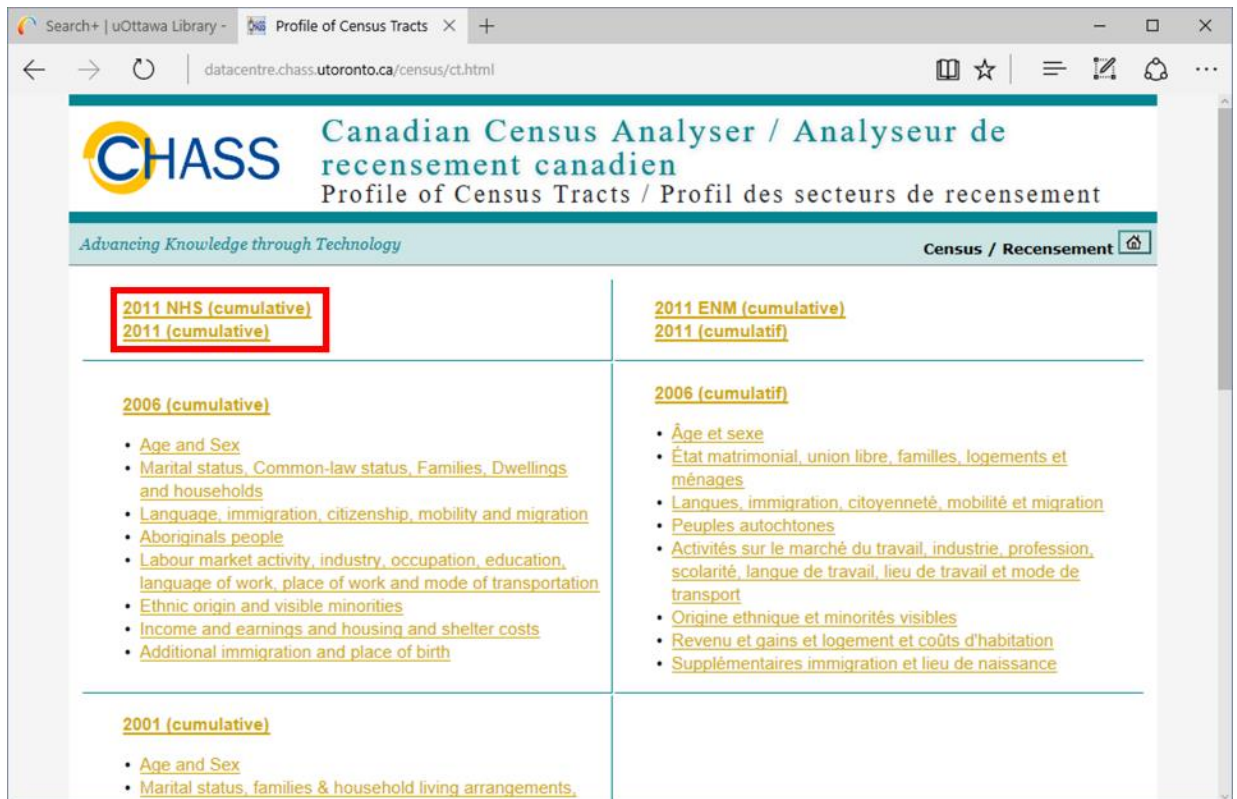
a. 2011 NHS (cumulative): the long-form census data that contains some 1000+ variables that are reported by approximately 20% of the population and contain details about work, transportation etc.
b. 2011 (cumulative): the full census that everyone fills out, which contains data about language, age, sex, marital status and population and age breakdowns.

Data is available at the census tract level for both types of census data.

4. Click on '2011 (cumulative)', then, under "Step 1: Specify Census Geography for retrieval", put a checkmark in 'O' and below that, a check mark in 'Ottawa-Gatineau (Ont.)'. You should see the following:
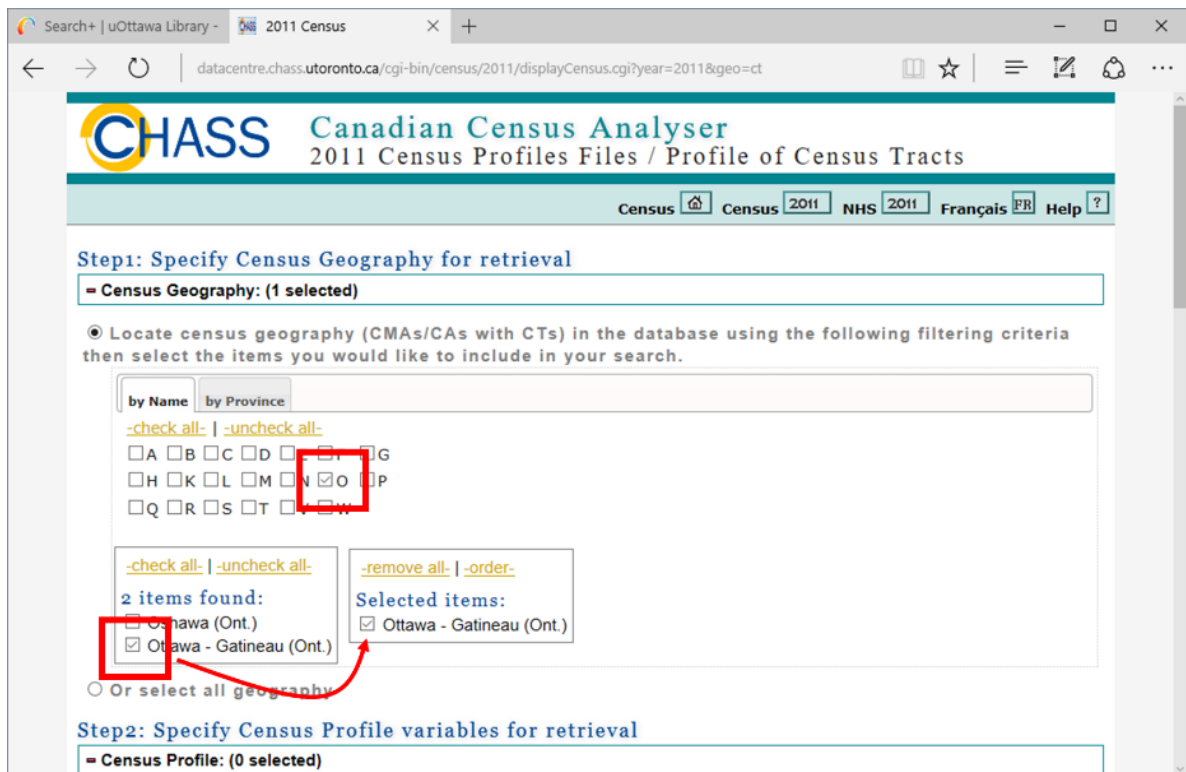
5. Now, under "Step2: Specify Census Profile variables for retrieval", place a check mark in each of Population, 2011 (v1) and Population, 2006 (v2) to specify that you want to download those two census variables:

Student Name: _____ Student #: _____
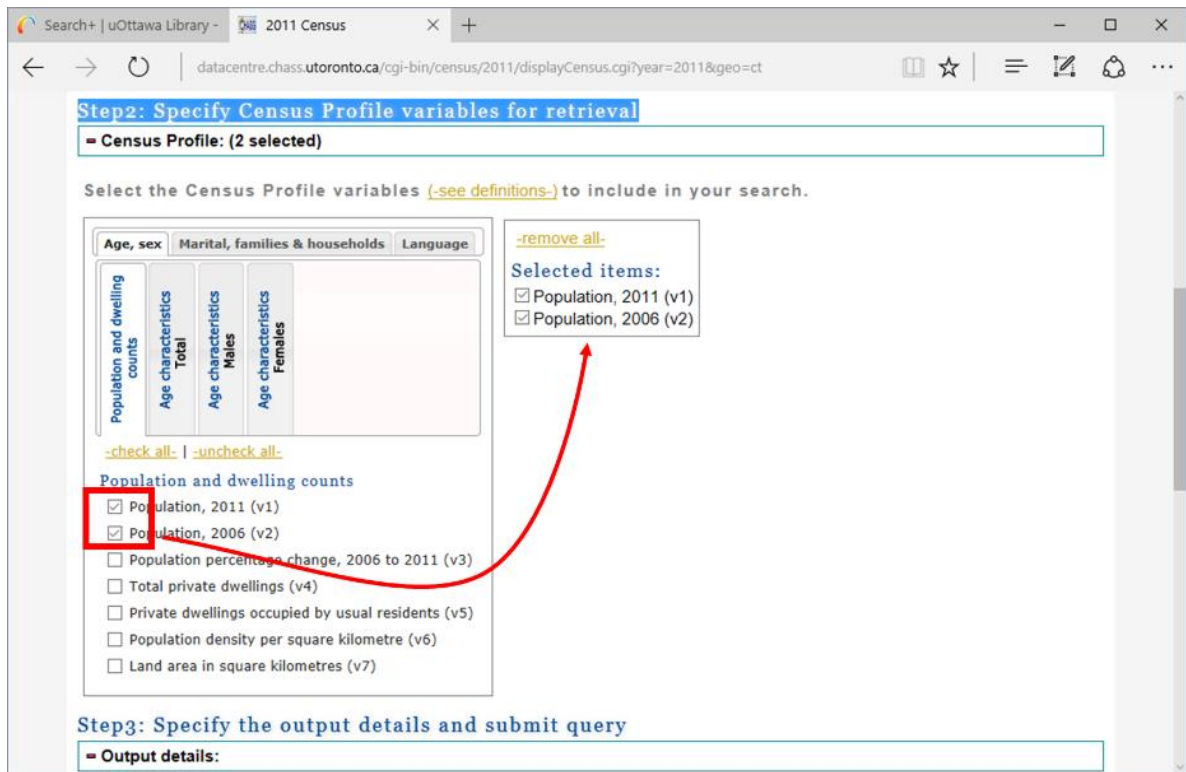
6. Next, you need to know what you are downloading, i.e., what do these variables mean, and how are they defined. So click on the (-see definitions-) link:
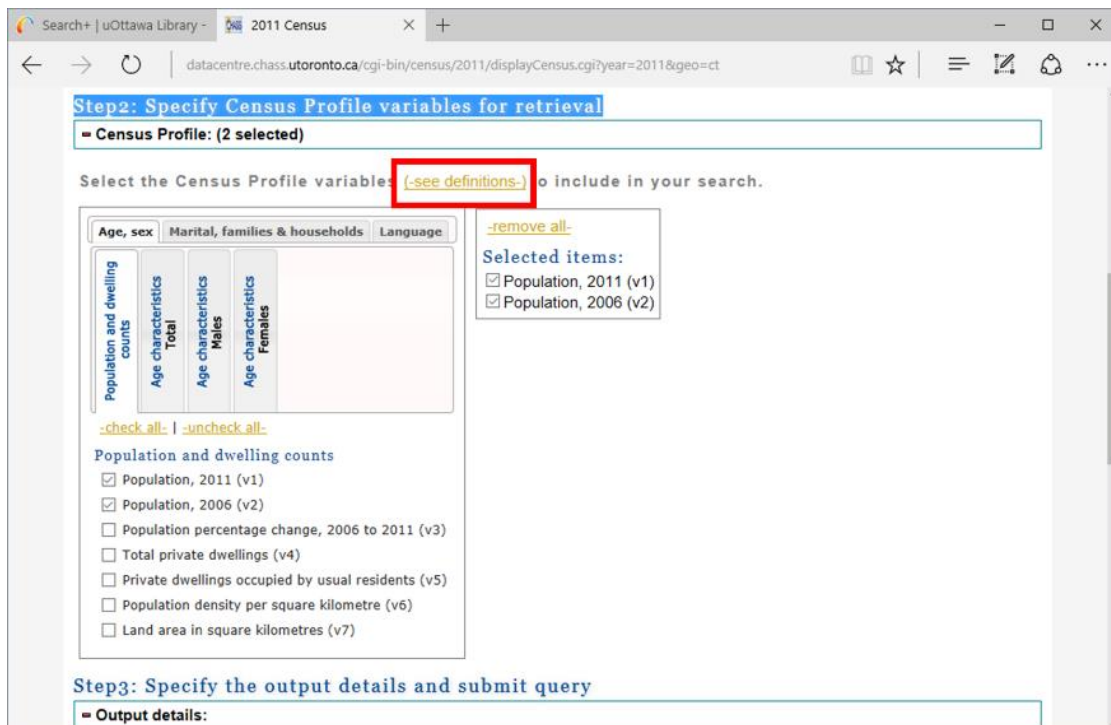
7. Read the information about the 2011 and 2006 population:
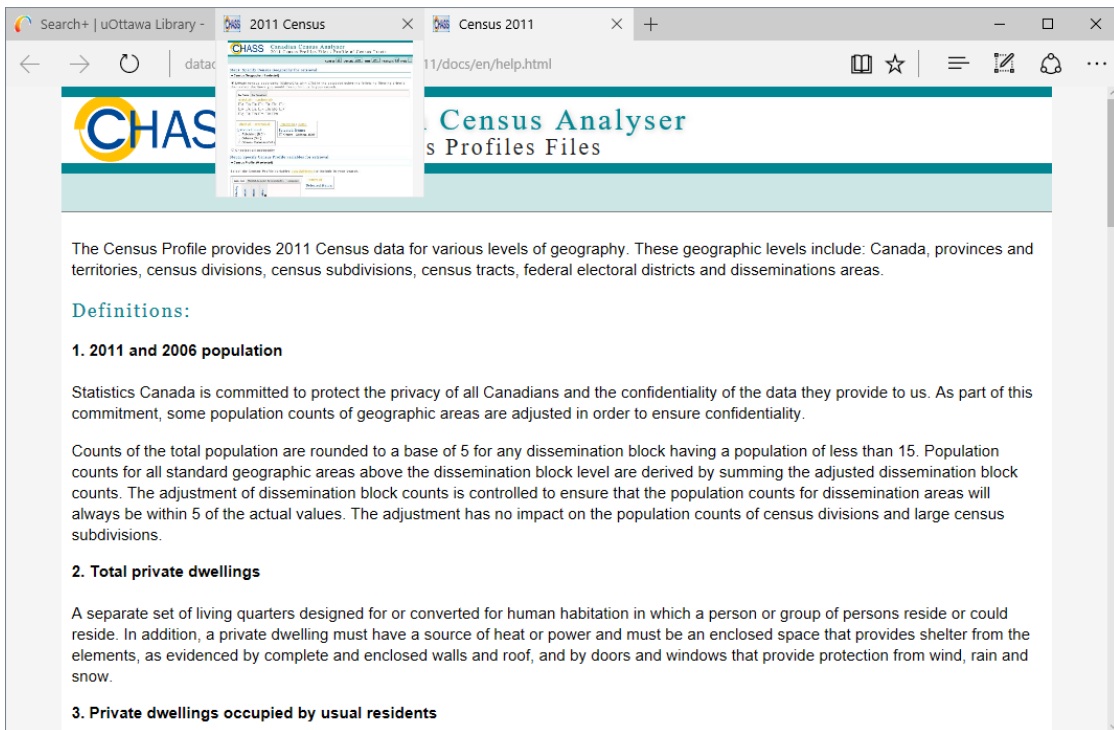
8. Return to the 2011 Census tab in your browser. Under "Step3: Specify the output details and submit query", under "Select output format", choose "Comma-Separated Values (CSV) for spreadsheet" and click the Submit Query button:
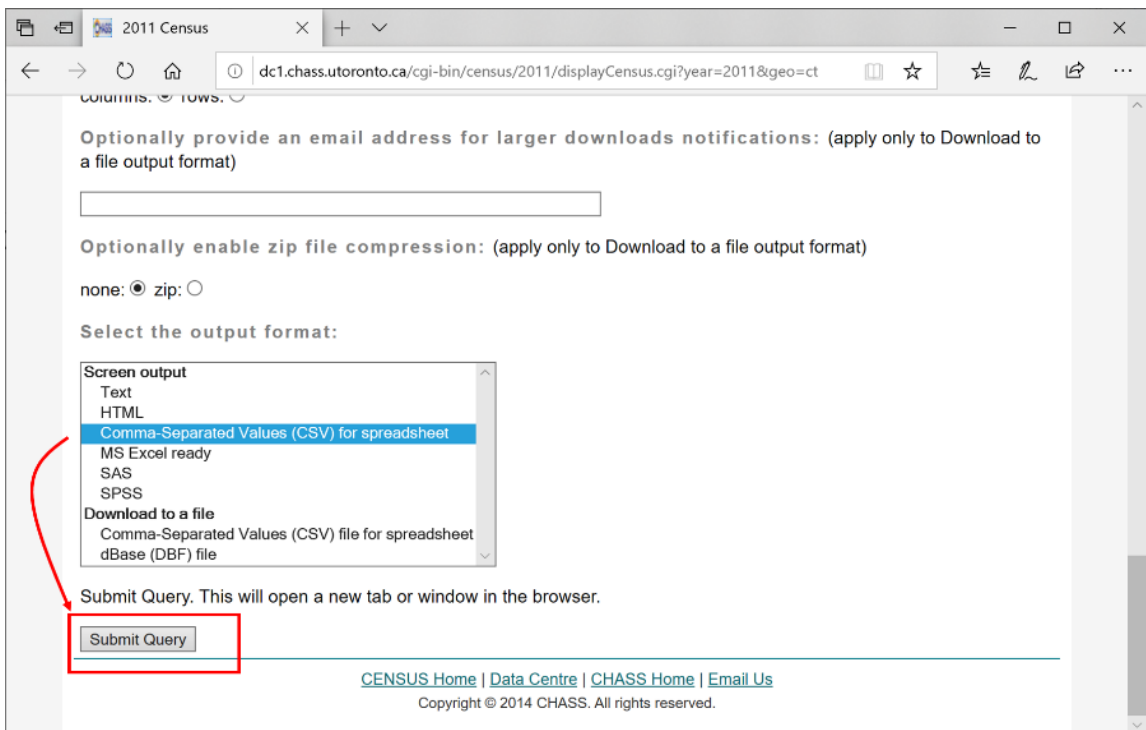
9. You will now see a webpage with the data that looks like this:
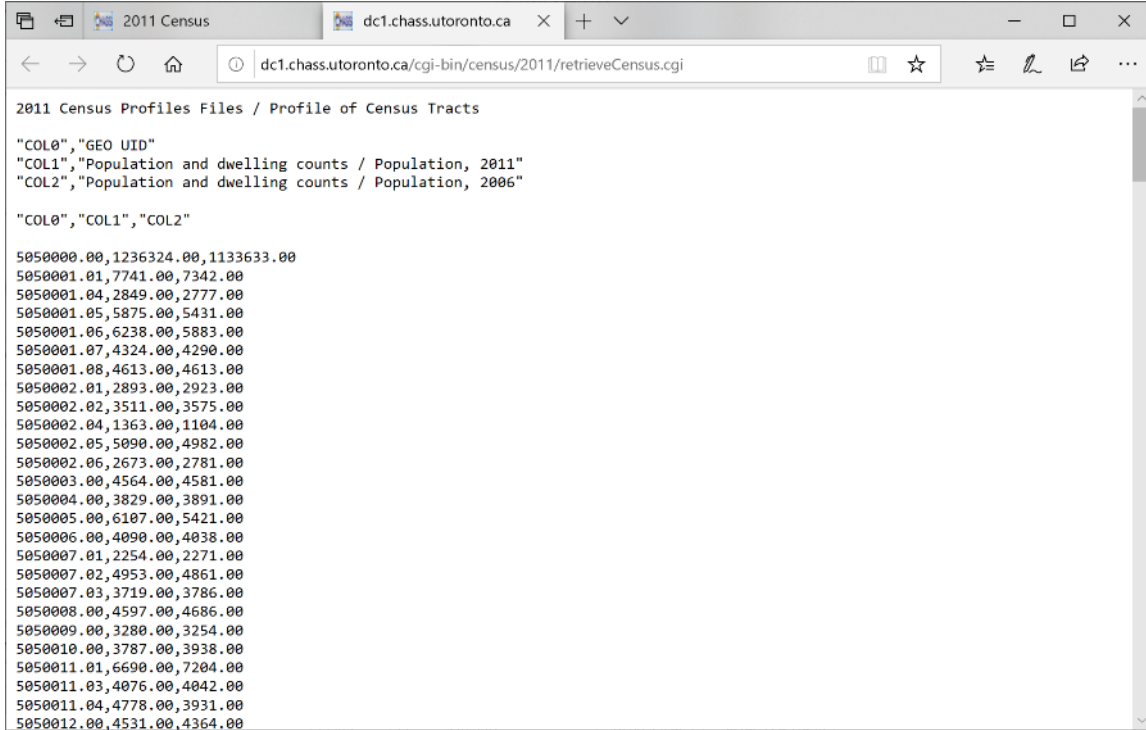
10. Select all the text on the webpage by using the key sequence Ctrl+A on the keyboard, followed by Ctrl+C to copy the data to the clipboard:
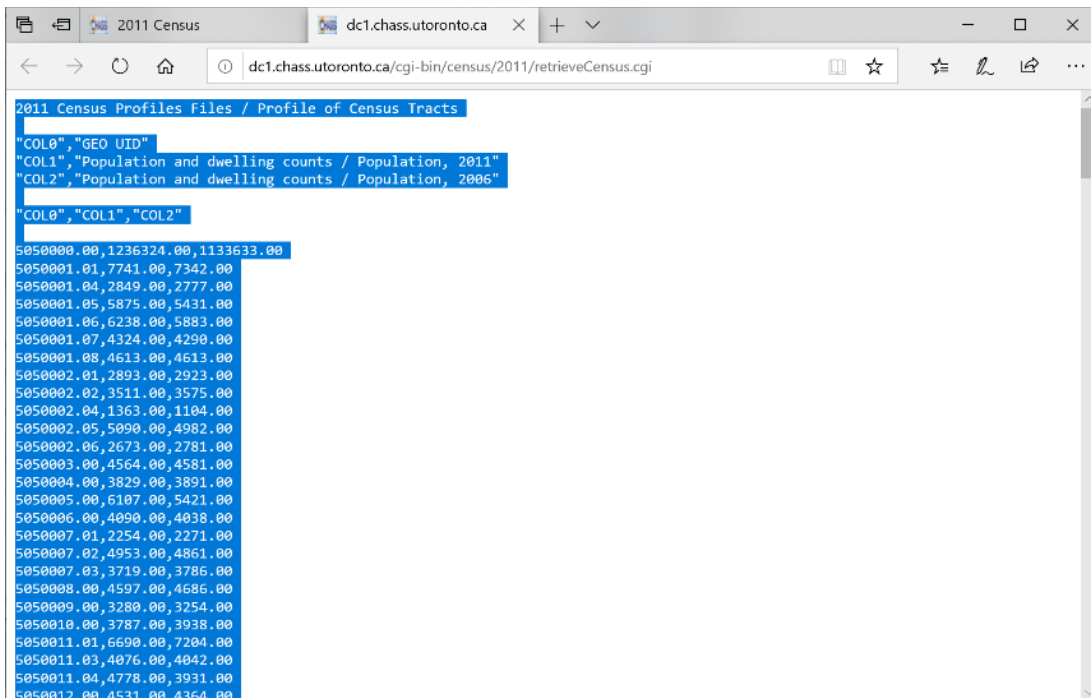
11. Open Notepad on your desktop and then use the key sequence Ctrl+V to paste the data into notepad:

12. Now you need to edit the data so that it can be opened as a ".csv" file in ArcGIS Pro. To do so, make your data look like the following below. It is important that there are no blank lines between the header (the first row that has the column labels) and the data. Then save the file with the name "ctdata.csv". Use the Windows Explorer to check that you did not accidentally use another file extension for the file (it can easily end up as ctdata.csv.txt, which will not work!)

13. Now return to ArcGIS Pro. Add the folder where you saved the '.csv' file, and add the .csv file to the project contents:
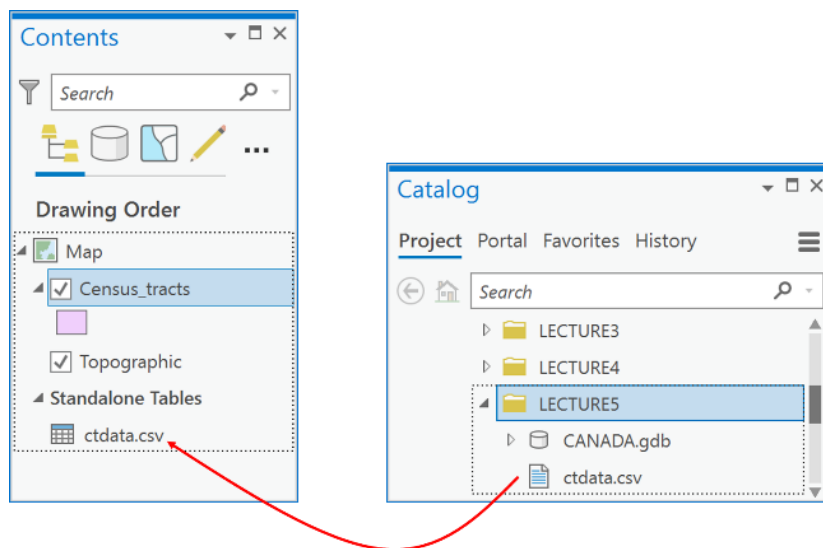
14. Open and examine this stand-alone attribute table and examine the table of the Census_tracts feature class that you downloaded, renamed and added to ArcGIS Pro in Q1 above. You will see that both have a CTUID field but that the field values are in different formats (see the figure below). Specifically, the ctdata.csv file has interpreted the CTUID as a numeric field, whereas the Census_tracts interprets the CTUID field as text. Fields can only be joined if they a) are both either text or integer data types, and b)

if they contain common values that are identical for the same features. In this case, the format for a CTUID must be a text value of the form "#######.##" without exception, like '5050151.08'. The confusion arises because these text fields use numbers, but the numbers have no numeric meaning, they are just unique labels for each census tract. Therefore, we will need to do some data manipulation on the ctdata.csv file before we can get access to the population fields for mapping or further analysis.
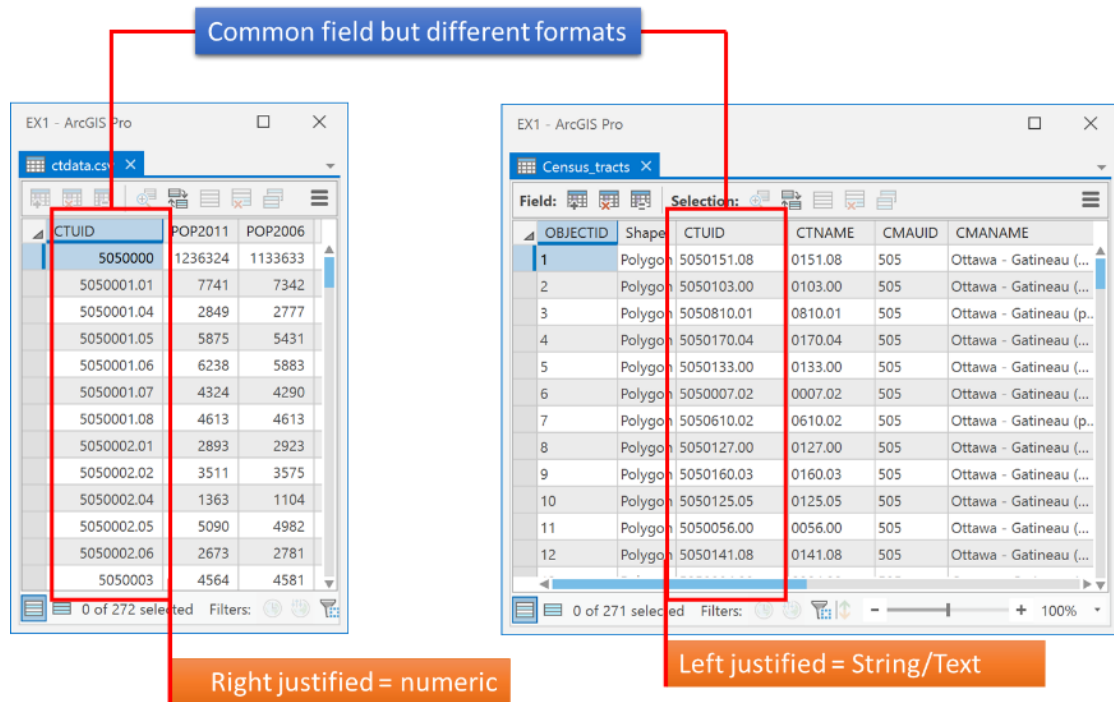
15. Open the Table to Table tool, and select your ctdata.csv file as "Input Rows", and the data.gdb geodatabase from Q1 as the "Output Location". Call the Output Name 'population' as the new table. Then click Run.
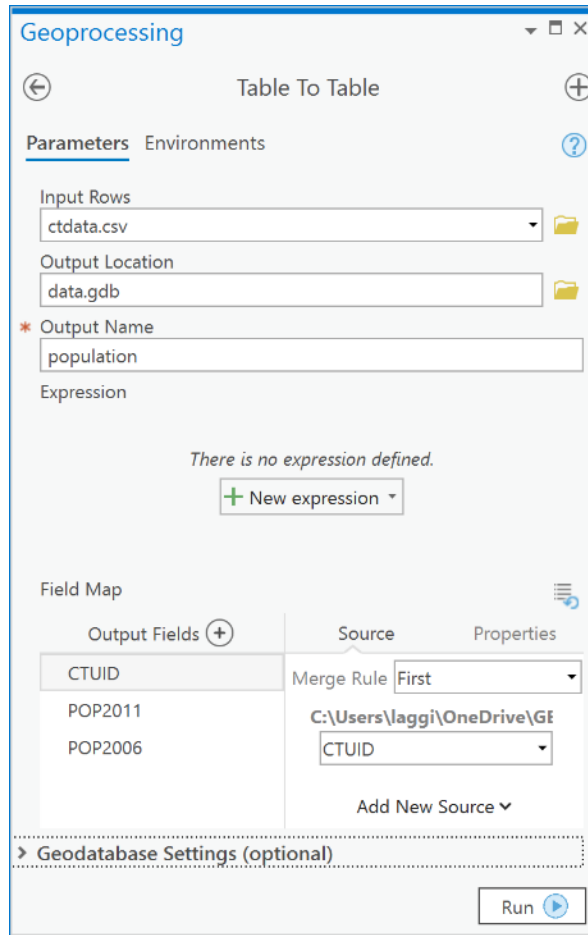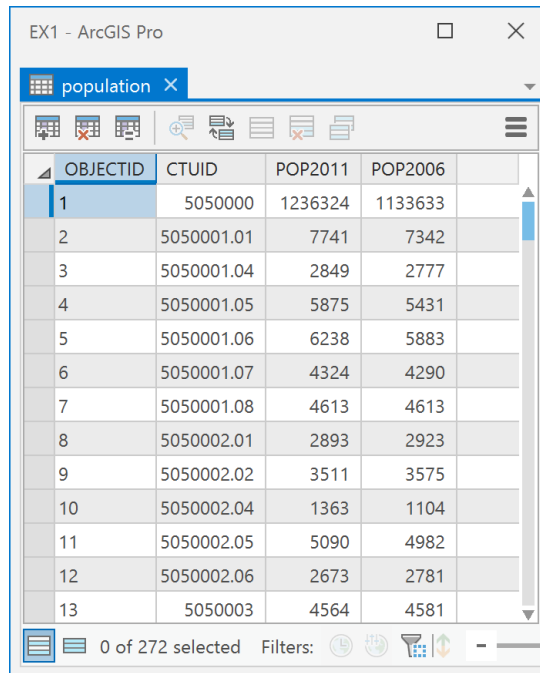
16. You should now have a population table in your data.gdb.
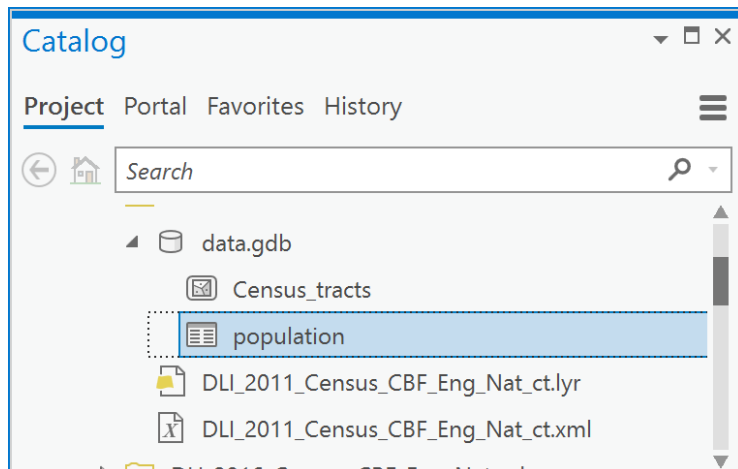
17. Open the population table and you will see that there is a new attribute added called OBJECTID. That is automatically added to any table within a geodatabase. You will notice that the first row has very large populations; that is because it is the total population for each year. You can delete that row or leave it, it will not affect the rest of
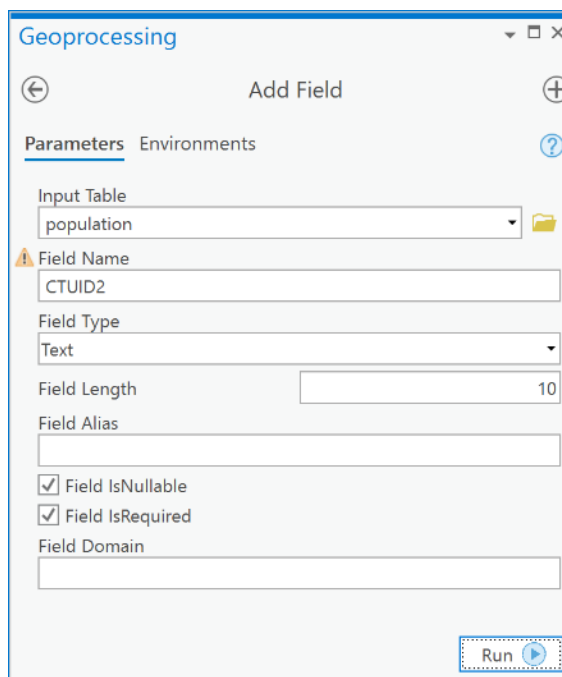
this exercise. More importantly, you will also notice that the CTUID is still in numeric format and many of the values are missing the ".##" (such as ".00") parts of the CTUIDs:

18. Add a new field (Data Management Tools -> Fields -> Add Field) to the 'population' table called CTUID2. Make it of type "Text" and a field length of 10 (digits). Run it.

19. Now, in the population table you should see a CTUID2 field. Right click on the field and choose Calculate field:
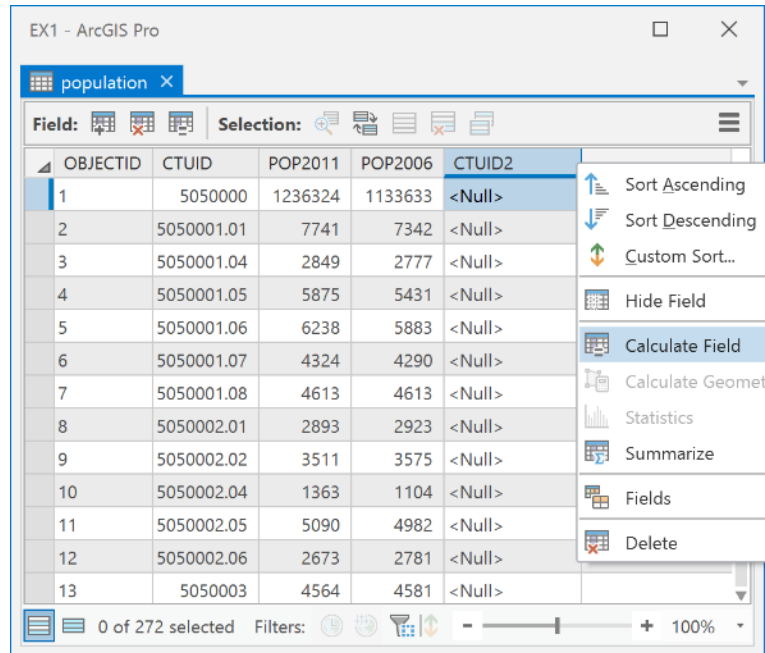
20. In the Calculate Field tool, type, under CTUID =, this statement:

"{:.2f}".format(!CTUID!)

21. Then click Run
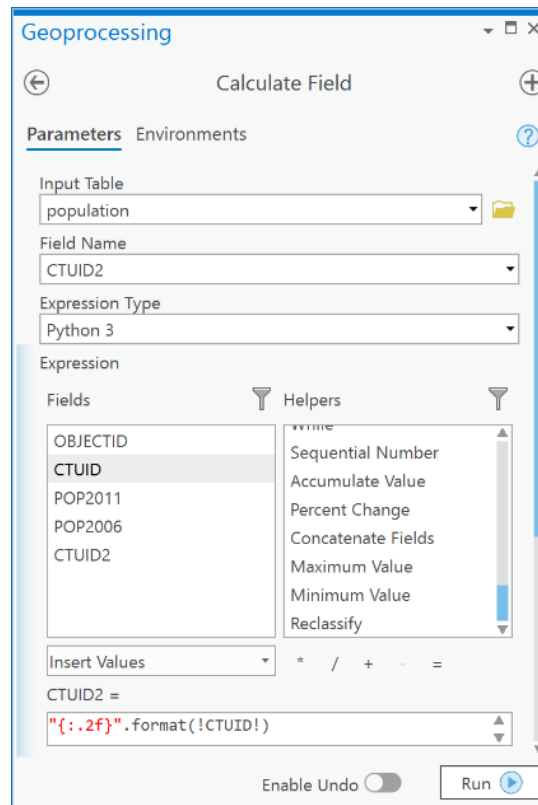
22. You will now see your population table has the correctly formatted CTUIDs in the CTUID2 field and the values are left justified – a visual indicator that this is a text field.

Q3: How can I relate the geospatial data layer Census_tracts to the attribute data table with the population information?

One way to access the fields in the population table from your census tracts layer is to tell ArcGIS how to match the CTUID field values in the census tract layer to the CTUID values in the population table. You can do this by creating a 1-to-1 *relationship class* within the geodatabase and then use that as a basis for querying and selecting census data in the population table to the Census_tracts attribute table.

1. In ArcCatalog, right-click on the **data.gdb** geodatabase and choose New -> Relationship Class:
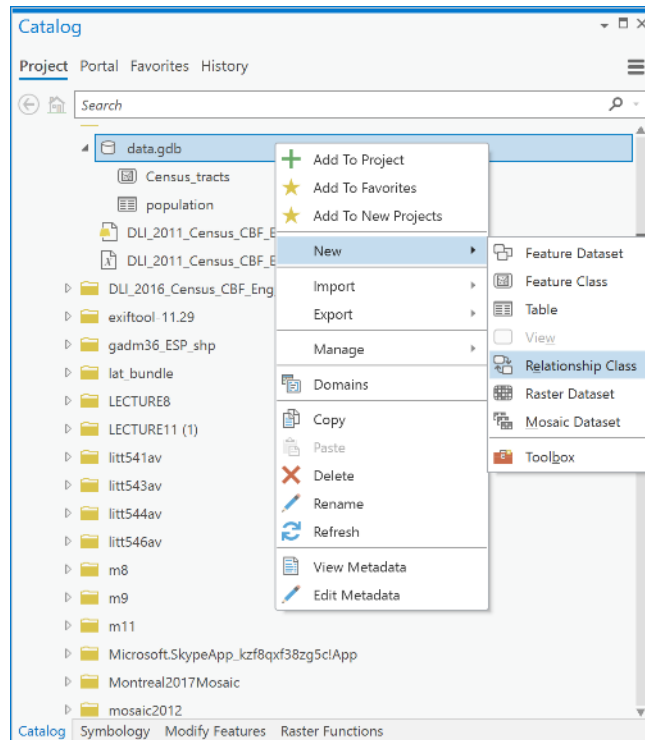
2. Set the following in the Create Relationship Class dialog box:
   - ***Origin Table***: Choose *Census_tracts*.
   - ***Destination Table***: Choose *population*. You have now specified that the Census_tracts table contains the source records for this relationship and that those records are related to the population table.
   - ***Output Relationship Class***: This is the name in the geodatabase that this relationship will be visible as, call it *pop*.
   - ***Relationship Type***: Simple.
   - ***Forward path label***: Type in 'population', which specifies the label when navigating from the origin class (Census_tracts) to the destination class (population). So we call this population to remind us that the relationship goes from Census_tracts to population.
   - ***Backward path label***: Type 'Census tracts'.
   - ***Message Direction***: Choose 'None (no messages propagated)', since this relationship class won't be used for editing features. If it was used in editing, then deleting a feature in one feature class would send a message to the other and tell it that there was a deletion. Such operations are critical and are part of what is called referential integrity: search for 'referential integrity' in the ArcGIS Pro help file to understand this type of relationship.
   - ***Cardinality***: This specifies the type of relation, in our case, each census tract can have at most one record in the population attribute table, so we choose a 'One to one (1:1)' relation.
   - ***Relationship class is attributed***: We do not need any new attributes to be stored for this relation so leave the "Relationship class is attributed" check box blank.
   - ***Origin Primary Key***: Specify the common field that contains a unique identifier for each spatial unit within the Census_tracts layer (origin), in this case, it is the CTUID field.

- ***Origin Foreign Key***: Specify the common field in the population table, so select CTUID2 in the dropdown.

3. Click Run.
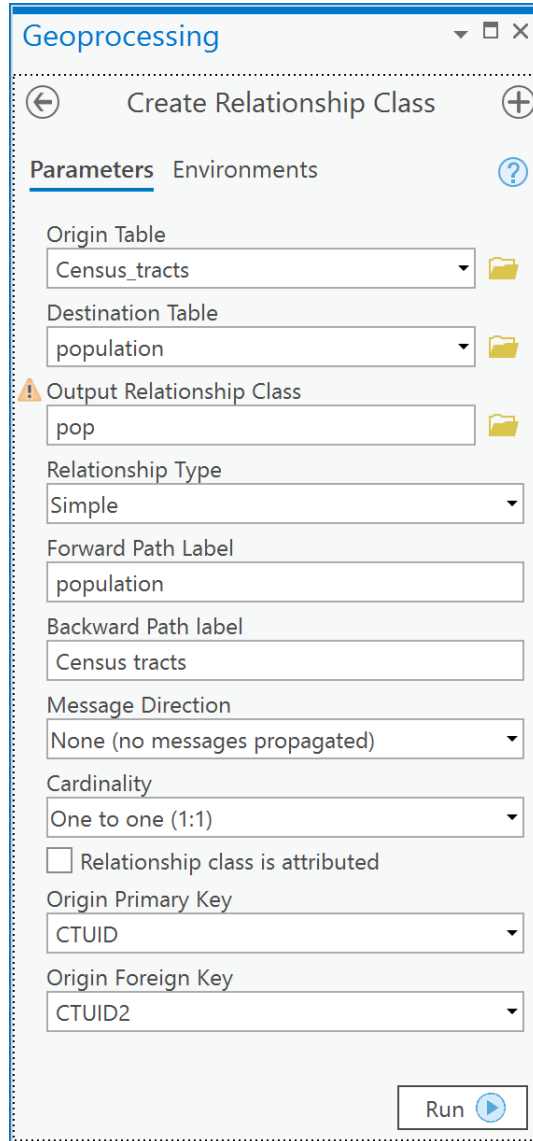
4. Now examine the data.gdb geodatabase and you should see a relationship class item added called 'pop':
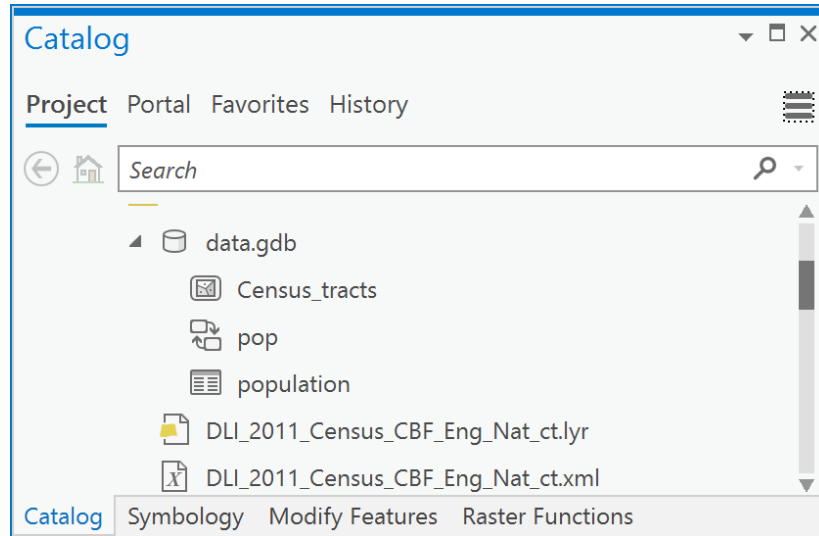
Q4: What can I do now that I have established a relationship class between my census population table and my census tracts feature class?

The relationship class is like a table join, but you do not see the records from the related table (population) within the Census_tracts table. There are limitations to what can be done with relationship classes. For example, you can query using related tables in SQL, but you cannot use fields in tables participating in a relationship class for field calculations in the origin table. Let's see what we can do:

1. Arrange the Census_tracts and population tables side by side. In the menu, choose "Select related records" in both tables:
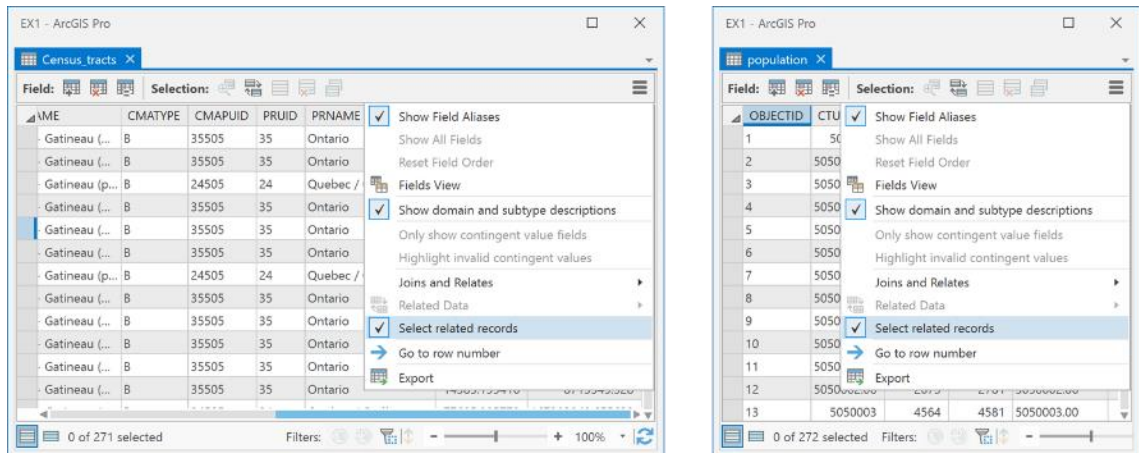
2. Next, select a few records in the Census_tracts table and examine the population table. You will see that corresponding records are selected in the population table!

3. To have effect in the other direction, select some records in the population table and then from the menu choose Related Data -> Census_tracts – Census tracts. You will now see that those features will be selected in the Census_tracts table and on the Map:

4. The relationship class adds this link between the two attribute tables that have the common values in the CTUID and CTUID2 fields. So you can query the Census_tracts table using SQL statements such as this:

    CTUID IN (SELECT CTUID2 FROM population WHERE POP2011 > **7000**)
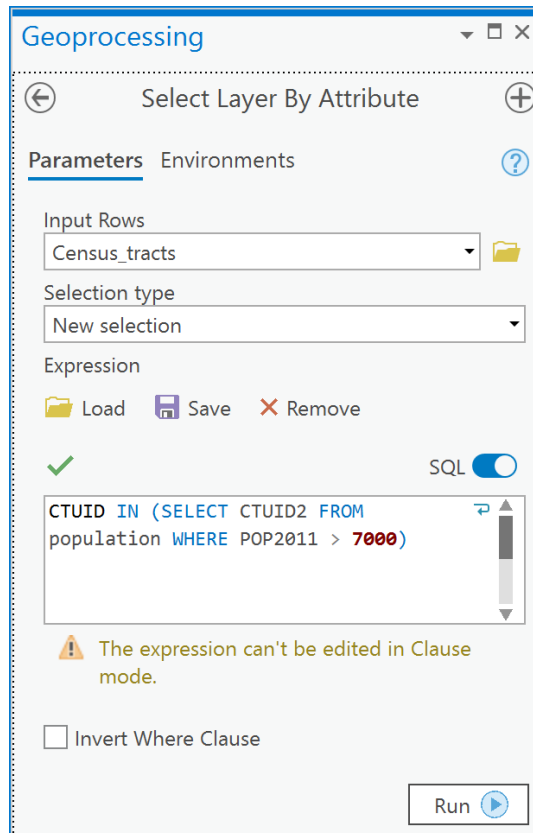
You will see all census tracts selected whose population is greater than 7000 people.

Note: One caveat of using census data that is reported within polygon areas like census tracts is that there is an implicit assumption that the number of things, like people, are uniformly distributed across the area of each tract. In reality, that is not true because people can't live everywhere within a census tract, e.g., people don't live in forests, or on streets etc., because people live in houses. Dasymetric mapping deals with increasing accuracy of maps by only showing data where people can live rather than everywhere. We deal with this in a more advanced course.

5. You could also use more complex SQL statements that evaluate arithmetic expressions on the related table's fields. Perhaps you want to select all census tracts that have had a decrease in the population from 2006 to 2011:

CTUID IN (SELECT CTUID2 FROM population WHERE (POP2011 - POP2006) < 0)
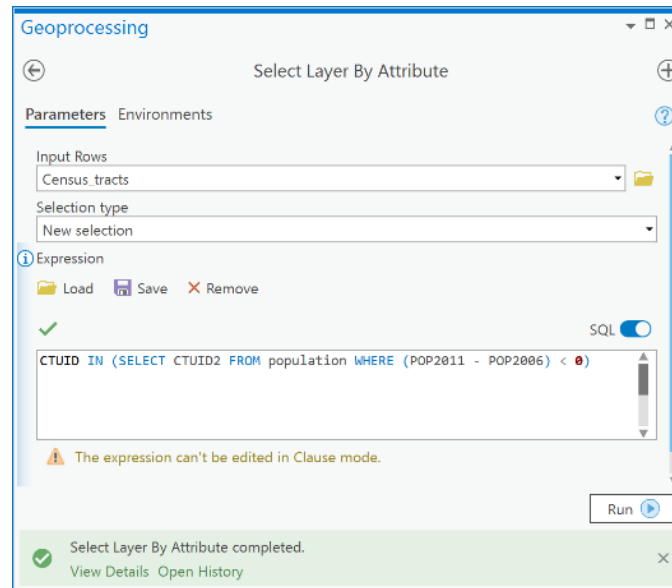
Q5: How can I map the population within Census tracts using my population table?

Relationship classes are limited when it comes to symbolizing fields in related tables. So we can't map the population within the Census_tracts layer using relationship classes. Therefore, we need to rely on table joins to append the population table's attributes to the Census_tracts layer. You can quickly do an **ad-hoc join** of the population table to the census tract layer so that you can have direct access to the POP2011 and POP2006 fields, so that they are visible in the Census_tracts layer as if they were part of the census tract layer's attribute table - but be aware that this is only a virtual join when we use the Add Join tool.

1. Right-click on the census tract layer in the Contents and choose Joins and Relates -> Add Join:
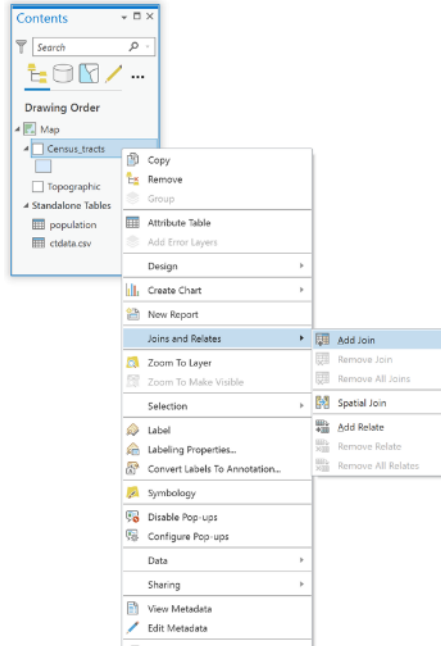
2. In the Joins & Relates dialog box, under 'Layer Name or Table View', specify Census_tracts. Under 'Input Join Field' choose CTUID. Under 'Join Table' choose population and finally under 'Output Join Field' choose CTUID2. Click OK.
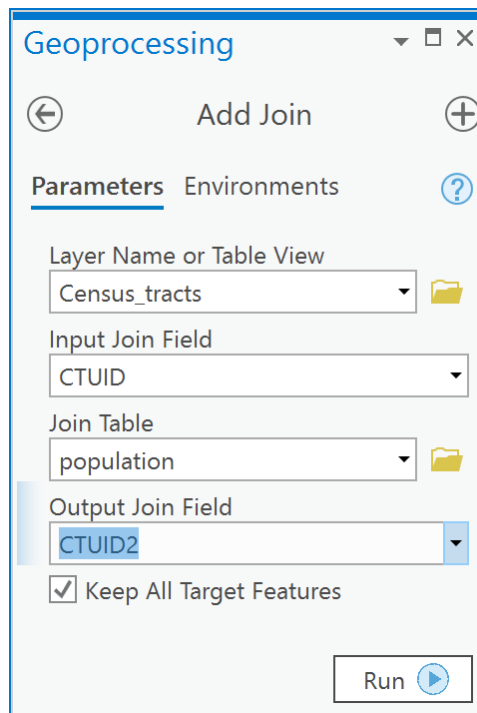
3. Open the Census_tracts attribute table, scroll to the right, and you should see the POP2011 and POP2006 fields appended to the end of the attribute table's list of fields:

Joined records from the *population* table

6. You can now Query, Map or Calculate Fields using these appended records within the Census_tracts attribute table. For example, using the **Select by Attributes** tool on the Map tab to open the "Select Layer by Attribute" tool and, for example, show which census tracts have a population decrease between 2011 and 2006 (note this is just a different way of doing what we did in Q5). You could query for:

population.POP2011 < population.POP2006
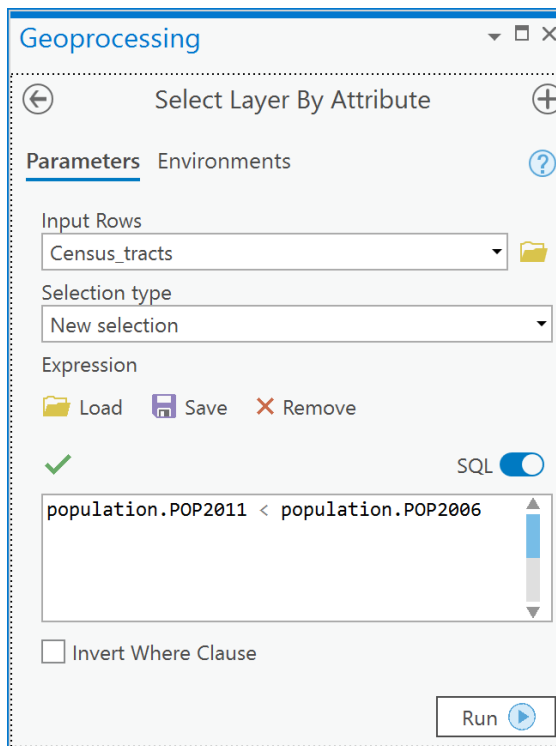
4. You will now see the selection made within the census tracts' attribute table and the map.
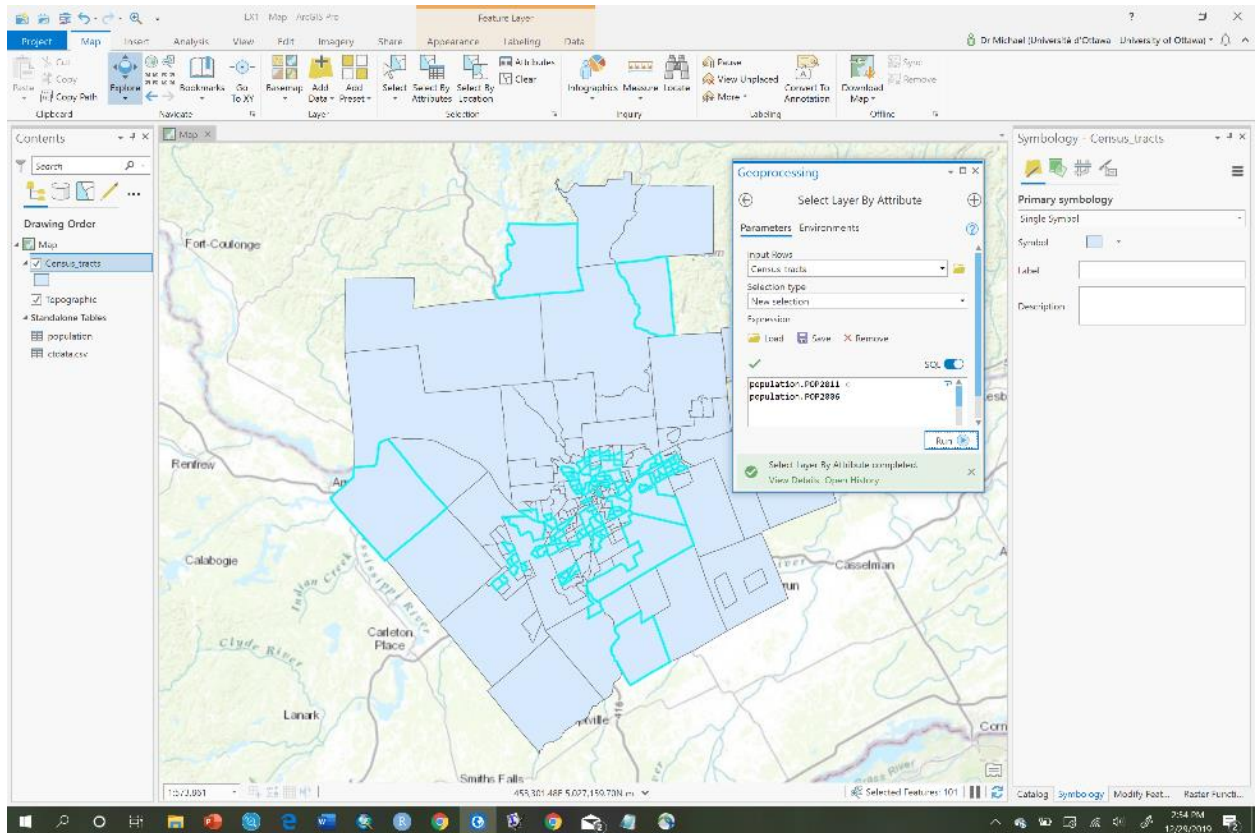
Q6: I want to calculate population density; how do I do that?

To calculate values using existing fields you will add a new field to your attribute table. For example, if I want to calculate population density in 2011 in each census tract, I would add a field of type Double (because density is a ratio and ratios will likely be real numbers) to the attribute table and then proceed to calculate the population density as number of persons per square kilometer,

1. Go to Data Management Tools -> Field-> Add Field,
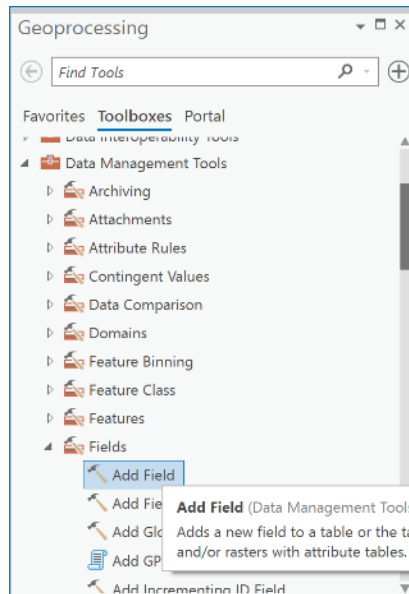
2. In the Add Field dialog box, choose your census tracts layer in the 'Input Table' dropdown, then under 'Field Name' type 'POPULATION DENSITY 2011' and choose 'Field Type' as Float or Double, and click Run:
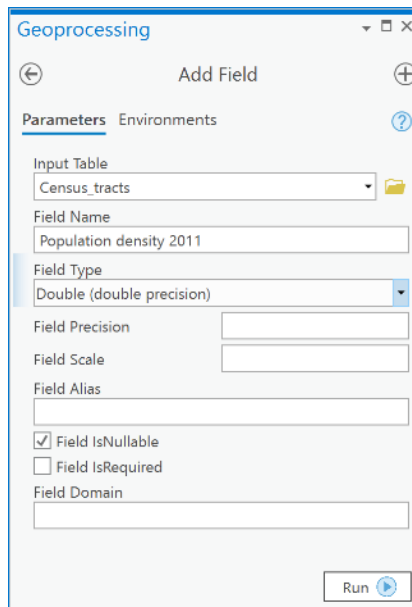
3. Examine your census tracts attribute table, you should see a new field called POPULATION_DENSITY_2011 with all <Null> values. Notice that ArcGIS replaced the spaces in the name you typed with underscore characters "_". It did this because spaces are not allowed in database field names.

4. Next, make sure nothing is selected in your Census_tracts attribute table by clicking on the Clear button in the Selection group of the Map tab of ArcGIS Pro.

5. Next, open the Calculate Field tool in ArcToolbox under Data Management Tools -> Fields:
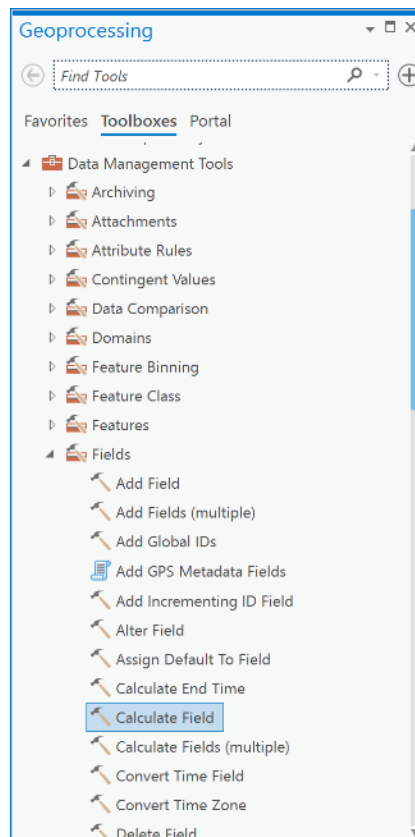
6. In the Calculate Field dialog box, (i) choose your census tracts layer under 'Input Table', (ii) under 'Field Name', choose the Population_density_2011 field, (iii) enter the expression: **!population.POP2011! / !Census_tracts.Shape_Area!**, and (iv) click Run.
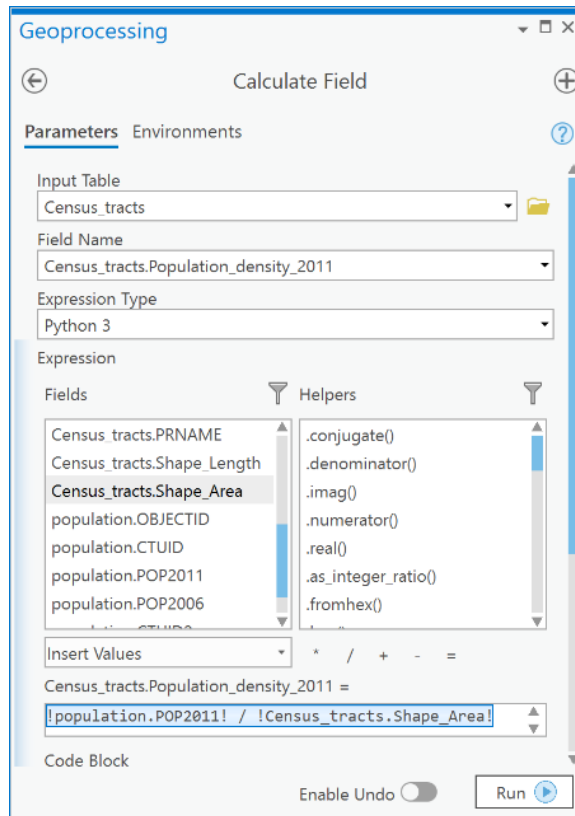
7. Open your census tracts attribute table, remove all joins, and examine the population density:

8. The population density values are in units of persons per square meter which is not a normal way to represent population density; usually you would have this in persons per square kilometer. To change this we will multiply this field by 1000^2 , or 1,000,000, because there are 1000^2 m^2 in one 1 km^2. In the Calculate Field dialog box, (i) choose your census tracts layer under 'Input Table', (ii) under 'Field Name', choose the

Census_tracts.Population_density_2011 field, (iii) enter the
expression **!Census_tracts.Population_density_2011! * 1000**2**, and (iv) click Run.
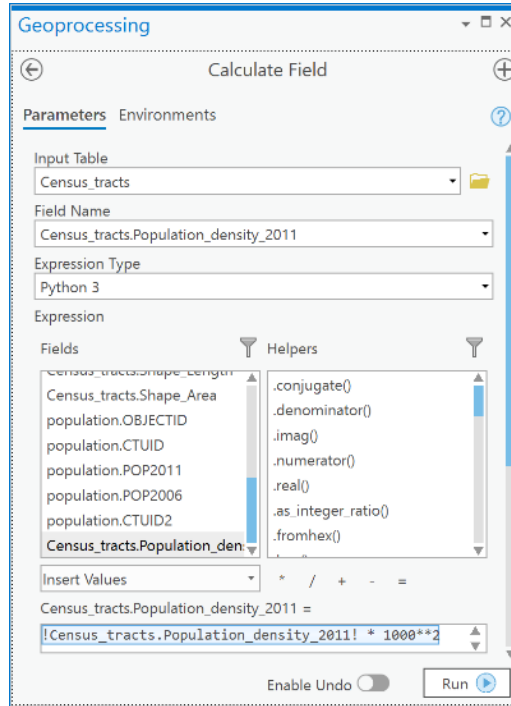

9. Now examine the census tracts attribute table and you will see that the population
density values are now more reasonable (and interpretable) as persons per square
kilometer:

10. Now, symbolize the census tracts layer by population density using a Graduated colors
map. Right-click on the Census_tracts layer in the Contents and choose Symbology.
Select Graduated colors as Primary Symbology, then set the 'Field' as

Population_Density_2011, and under 'Method' select a Quantile classification. The result should something like this:
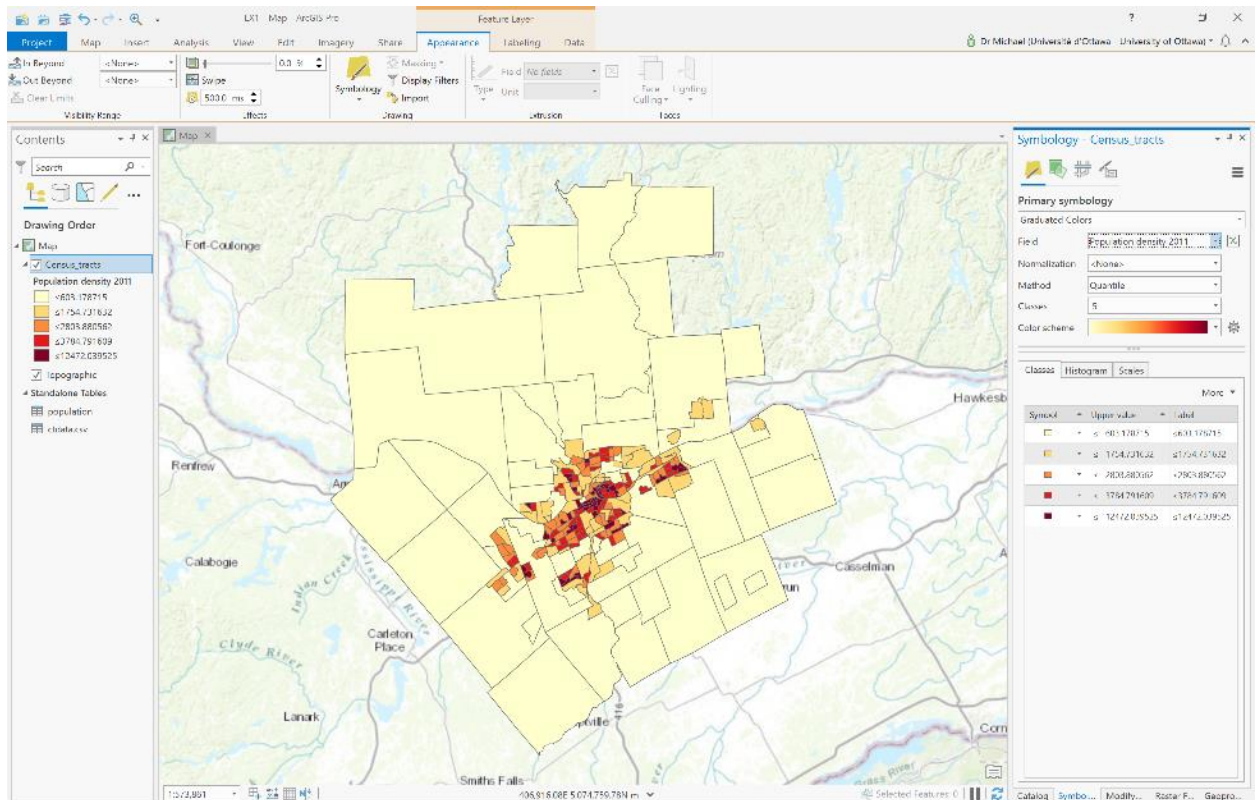

Questions to hand in.

Note: Some of these questions require an explanation, so read carefully. Some questions also require a map as part of the answer and this must be presented correctly, so ensure that you review the ESRI Training course "Creating a Map Layout". Marks will be deducted for incorrect presentation (also remember scale-bars should be metric).

Note: Some questions require you to download data from Scholars GeoPortal or the Canadian Census Analyser. If you do not have access to these resources, you can find the relevant data files in lab3.zip.

1. [2 points] How can a relationship class enforce referential integrity in the geodatabase?

2. [2 points] In Q6, you calculated population density in the Census_tracts layer. In what coordinate system was the area for each Census tract calculated? How do you know this? Is your choice of CS a good choice for calculating area? Why or why not?

3. [1 points] Do you think that the population density map is spatially accurate as to where people are clustered? Why or why not?

4. [9 points] Download the 2016 census tracts for Ottawa-Gatineau. Next download from the 2016 tables in the Canadian Census Analyser (see Q2) the following variables for the Ottawa-Gatineau area (note that in questions 5 and 6 below you also need to download more variables, you may prefer to do all of them at once, it is up to you):
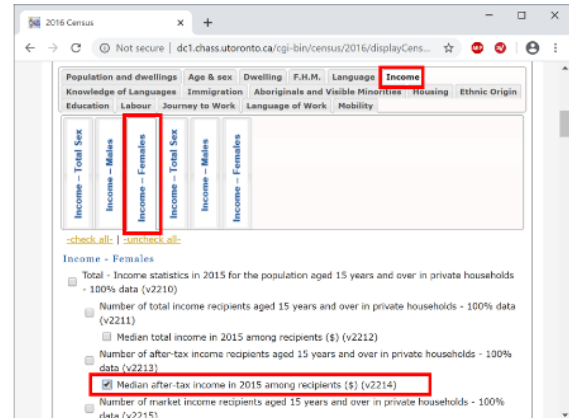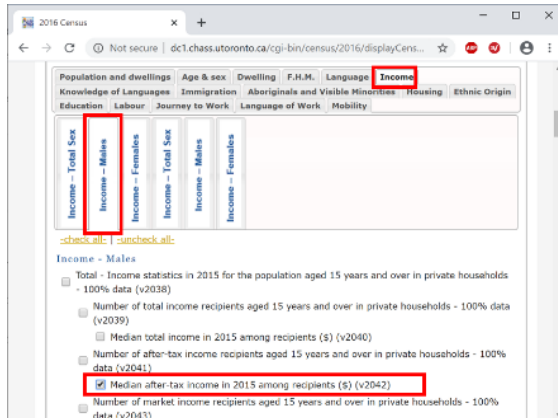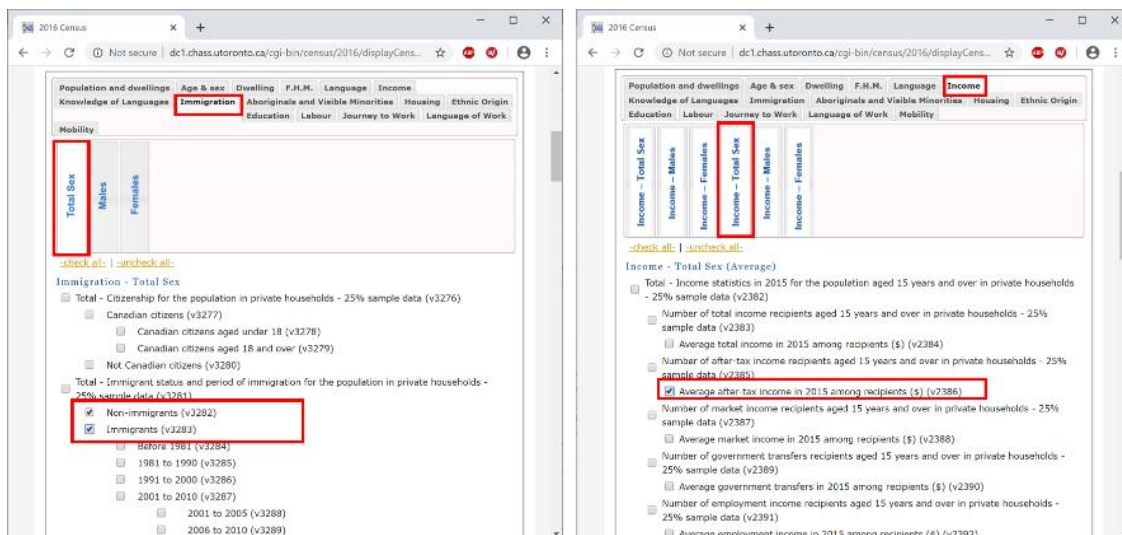

Join your data table to your census tract layer for 2016 (see Q5 for that procedure). Answer the following questions [1 point each], only the last one (i) requires a map to be created:

   a. In *how many* census tracts do males make less than the median after-tax income of 35,000$? (you need a query to answer this so see Q5 for query examples)
   b. In *how many* census tracts do females make less than the median after-tax income of 35,000$?
   c. In *how many* census tracts do males have a greater median after-tax income than females?
   d. In *how many* census tracts do females have a greater median after-tax income than males?
   e. In *how many* census tracts do females AND males have a median after-tax income less than 35000$?
   f. In *how many* census tracts do males have a median after-tax income between 35000$ and 45000$?
   g. In *how many* census tracts do females have a median after-tax income between 35000$ and 45000$?
   h. In *how many* census tracts do females AND males have a median after-tax income between 35000$ and 45000$?
   i. Describe the spatial pattern of the male/female median after-tax income ratio (either create a new field and calculate the ratio or use the female median income as a normalization field in the symbology for graduated colors)? For this question present a map as a figure.

5. [6 points] Download from the 2016 tables in the Canadian Census Analyser (see Q2) the following variables:
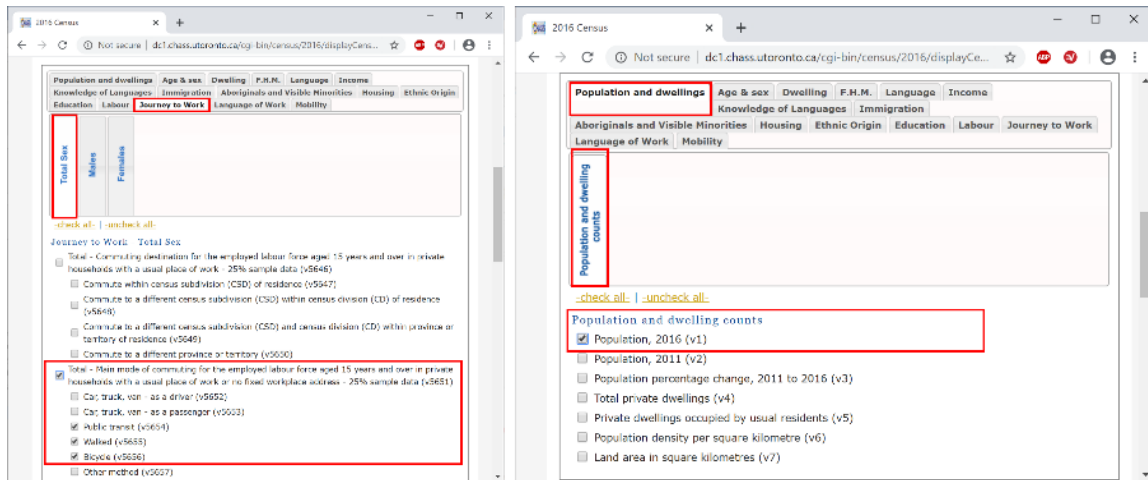
Join your data table to your census tract layer for 2016 (see Q5 for that procedure).
Answer the following questions [1 point each]:

    a. In how many census tracts do more people bike than walk as their mode of transportation to work?

    b. In how many census tracts do more people walk than bike as their mode of transportation to work?

    c. What is the average after-tax income of census tracts where more people walk than bike to work?

    d. What is the average after-tax income of census tracts where more people bike than walk to work?

    e. Create a map showing the ratio of biking to walking and explain the pattern you observe? (You need a map and explanation).

    f. Create a map showing the proportion of the total population that bikes to work and explain the pattern that you observe.

6. [7 points] Download from the 2016 tables in the Canadian Census Analyser (see Q2) the following variables:

Join your data table to your census tract layer for 2016 (see Q5 for that procedure).
Answer the following questions [1 point each]:
  a. In how many census tracts are there more immigrants than non-immigrants?
  b. In how many census tracts is the ratio of immigrant to non-immigrant greater than 0.25?
  c. In how many census tracts is the ratio of immigrant to non-immigrant greater than 0.5?
  d. In how many census tracts is the ratio of immigrant to non-immigrant greater than 0.4?
  e. What is the average income of census tracts that have an immigrant/non-immigrant population ratio of greater than 0.4?
  f. What is the average income of census tracts that have an immigrant / non-immigrant population ratio less than 0.4?
  g. What is the pattern of immigrant / non-immigrant population ratio?

Laboratory #4: Spatial Analysis in Vector

Purpose: The purpose of this lab is to provide you with an introduction to basic spatial analysis concepts and capabilities in geographic information systems science.

Learning Outcomes: By the end of this lab, students will be able to:
- Execute a spatial query.
- Differentiate between a spatial query and an attribute query.
- Undertake a date-based query.
- Assess centrographic measures of a point pattern.
- Create and explain a kernel density (KDE) estimate of a point pattern.
- Use point-in-polygon operations to calculate density.
- Operationalize union overlay.

Procedure:
- Read and work through the examples.
- Answer the questions at the end of the lab.

Data/Materials:
- Provided with laboratory outline:

Assignment:
- Answer the questions (#1-5).
- The total marks available for this assignment = 15 marks for 12.5% total course grade.
- All lab assignments must be completed and handed in individually.

Format:
Students will hand in questions with completed answers.

Tables and figures should be produced to the standards of the Department of Geography, Environment and Geomatics, University of Ottawa.
For guidelines on proper formatting and presentation of tables and figures, please see the introductory section of this lab manual.

Student Name: _____ Student Number: _____

This section presents examples of things you will need to know to address the questions you will be handing in for marks. Your task is to review and repeat the examples in this section and then determine which of these are relevant to helping you address the questions that are at the end of the exercise.

Q1: How do I undertake a spatial query?

In Lecture 5 and Exercise 2, you learned about structured query language (SQL) and how to use queries to answer questions about attribute data using "Select Layer By Attribute" in ArcGIS. A spatial query employs the "Select Layer By Location" dialog and usually two or more spatial layers within the same spatial domain. For example, suppose we have two point layers: one layer called **Crime**[1], shows the locations of crimes and the second layer called **Schools** shows the location of schools (Figure 6):
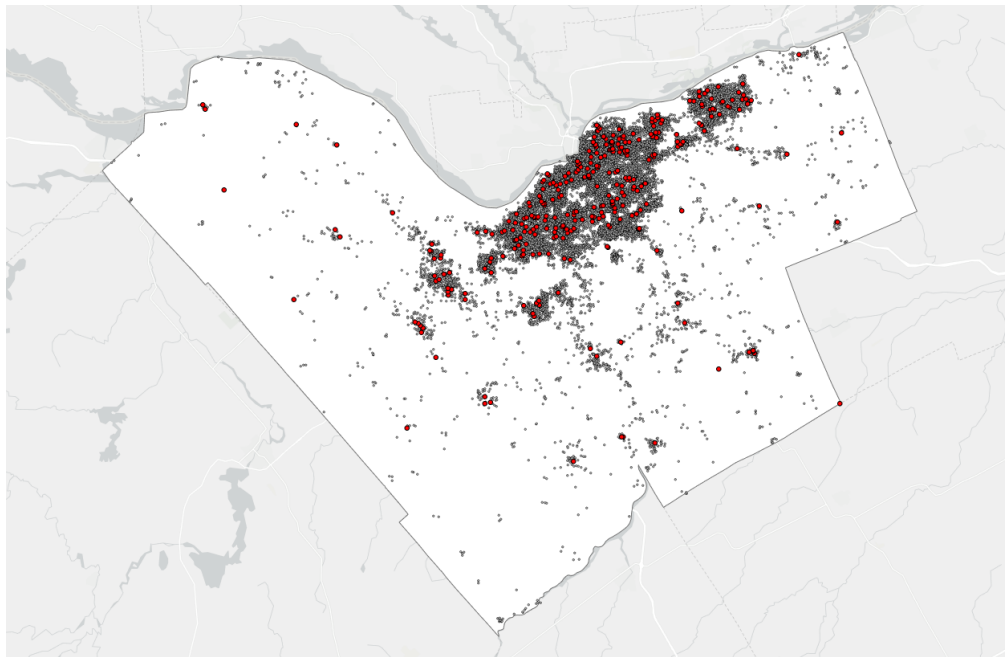

*Figure 6: Two layers for use in spatial and attribute query denoted by different colored points.*

We might want to select all crimes within 250 m of schools in Ottawa. The most efficient way to complete such a task is to use the "Select Layer By Location" tool to select all Crime points within 250 m of Schools.

1. Add Crime and Schools to ArcGIS Pro (from the Ottawa feature dataset in the OTTAWADATA geodatabase in the Lecture7 folder).
2. In the Map tab, find the 'Selection' group and Click on "Select Layer By Location" to open the Select Layer by Location tool:

---

[1] The Crime layer is a synthetic layer representing calls-for-service (when someone calls 911) and does not represent actual crimes committed at the locations specified. At a small scale, the dataset is representative of the real patterns of crimes in Ottawa over the period of 1998-2001. It was created by using probability surfaces derived from actual calls for service over that time period specified. It is not locally accurate.
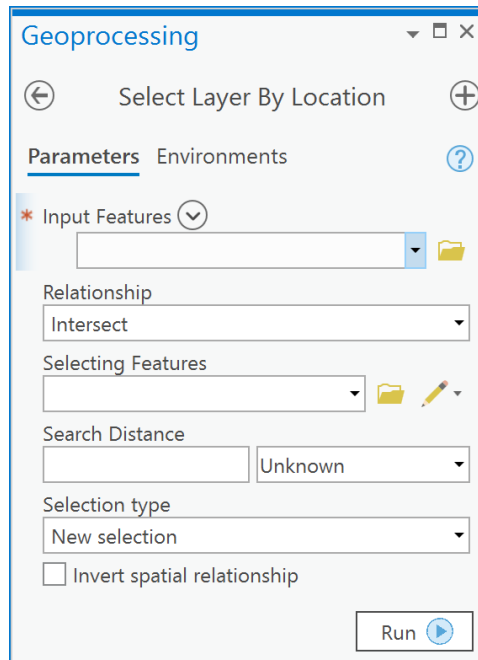
3. In the dialog, the Input Features are the target of the selection and here because we want to select all crimes within 250 m of a school, the feature class Crime should be selected under 'Input Features'. Under 'Relationship' choose Within a distance. Under 'Selecting Features' choose Schools. Type in 250 within the 'Search Distance' input box. Click Run.
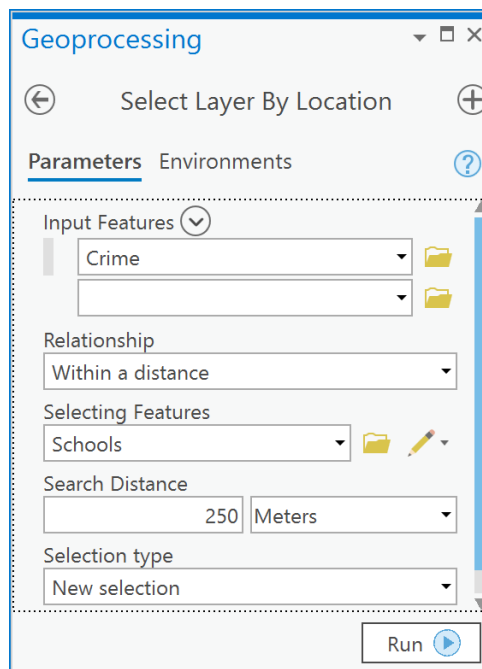
4. You should see a result that looks like the following:
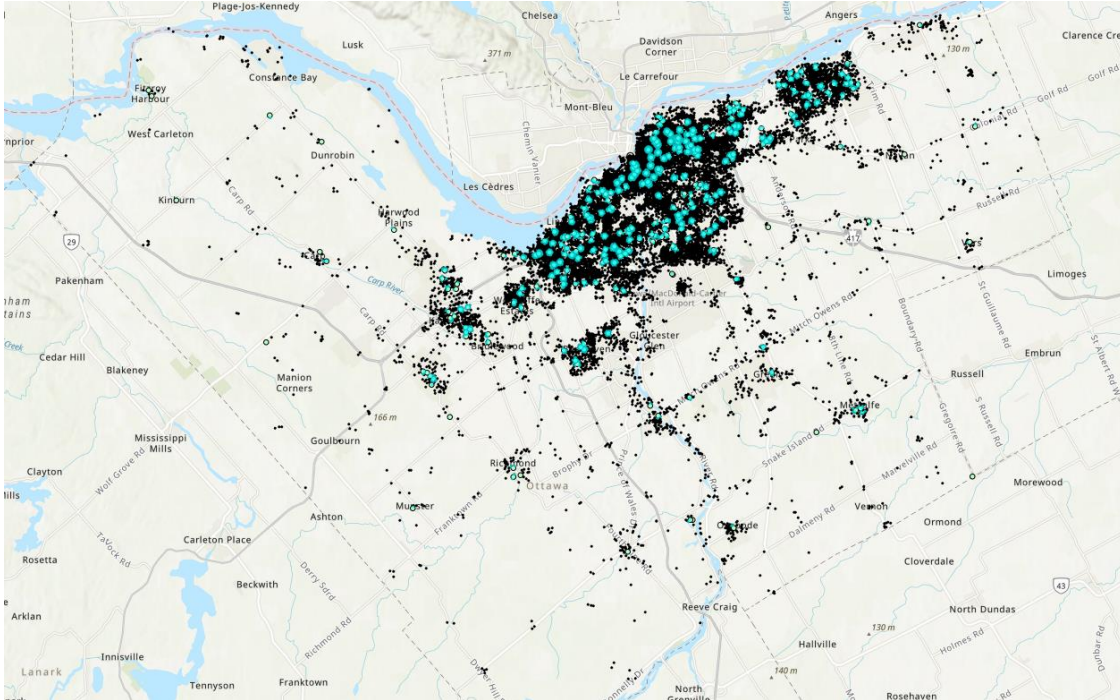
You have now undertaken a spatial query. You have used a spatial relationship - "Within a distance" - between two feature classes within the same spatial domain to select features within one of the participating feature classes. The "Within a distance" operation "creates buffers using the buffer distance around the source features and returns all the features intersecting the buffer zones." (ESRI Help).

Q2: How are SQL and spatial queries (select layer by location) related?

Using Select Layer By Location with Select Layer By Attribute together is a powerful combination. For example, you could address a question such as: *How many assaults have occurred within 250 m of a school in Ottawa*? Answering this question requires a spatial query and an SQL query together:

1. Repeat Example Question Q1 to select all crime within 250 m of each School.
2. Open the attribute table for the Crime layer.
3. Notice that there are 6251 Crime points (out of 25454) selected. Also, notice that there are other attributes such as "SUMMARY" which provides information on the type of crime for each point:
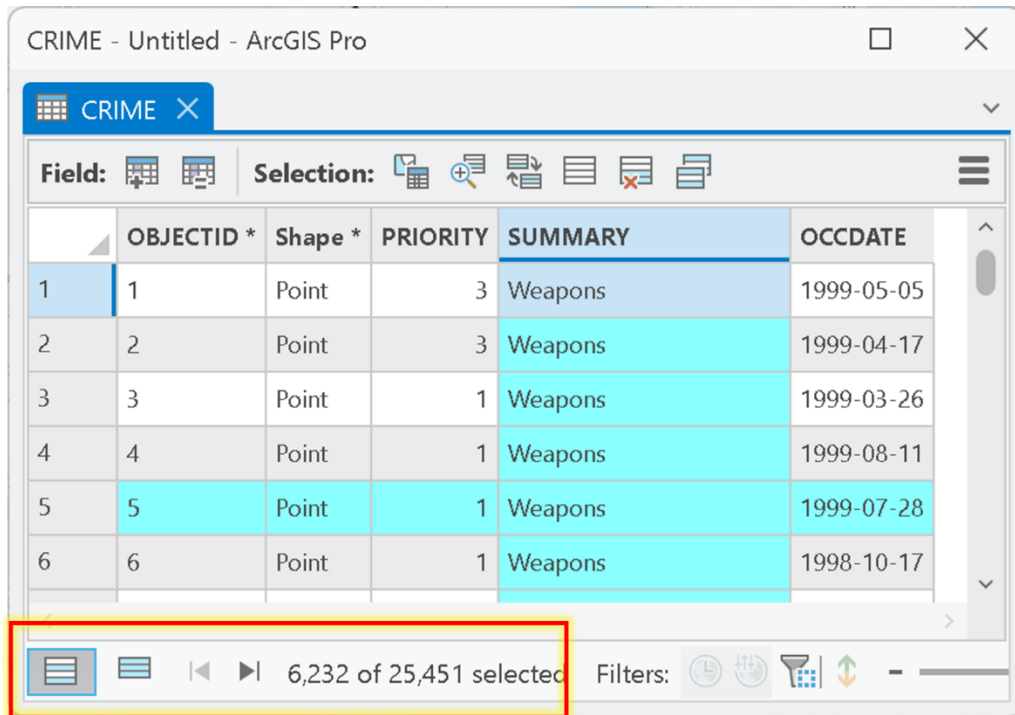
4. So, *from* the 6251 selected features that resulted from the spatial query in Example Question Q1, you now want to select only those that are 'Assault'. To make this selection, keep the Crime table open and go to the Map tab and find the "Selection" group and choose Select Layer By Attribute.
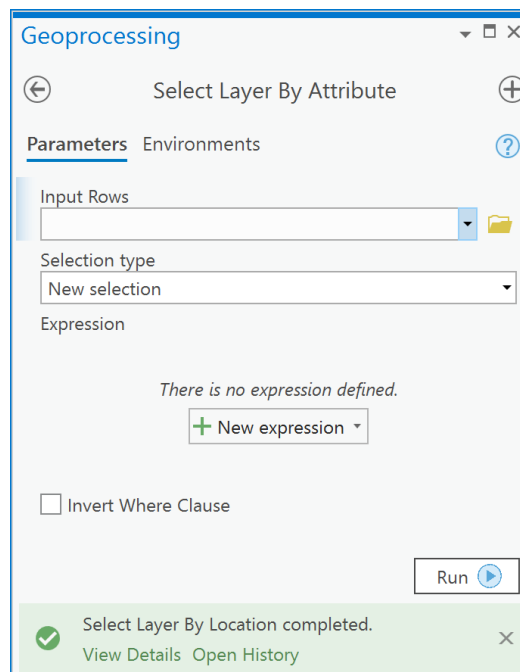
5. Examine the Select Layer By Attribute dialog box below and make yours look the same, the key here is to tell ArcGIS that you want to '**Select subset from the current**

selection' so that only those points that meet the criteria of having a SUMMARY type equal to 'Assault':
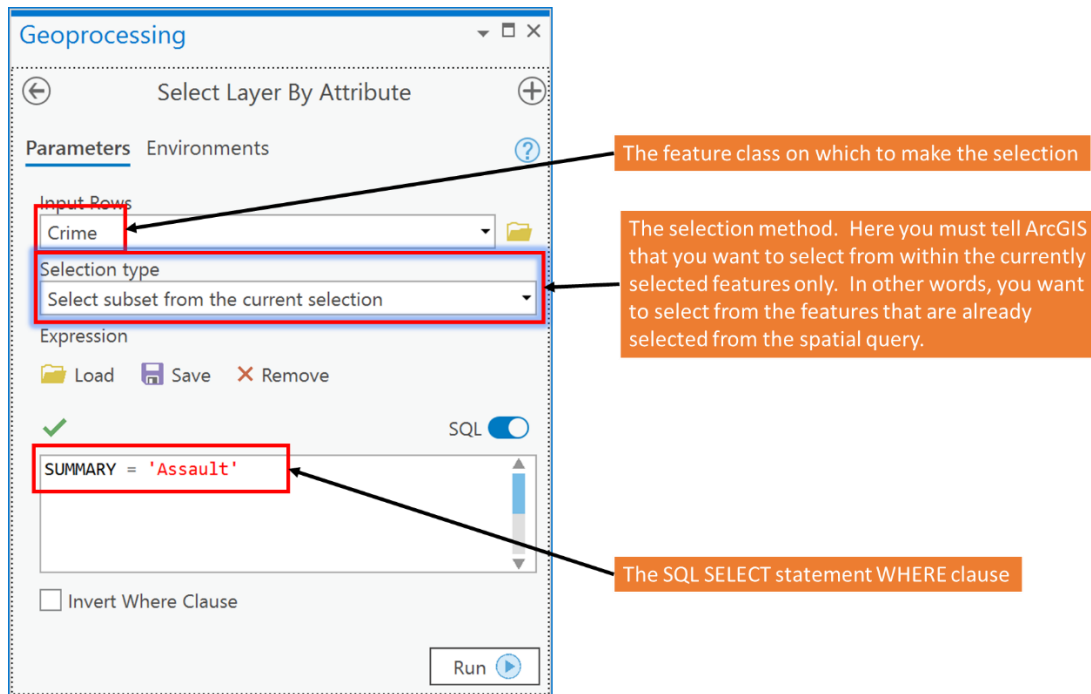
6. Click Run.

7. Examine the attribute table for Crime, you can see now that there are 909 assaults within 250 m of schools in Ottawa. This represents about 14% (909/6232 * 100) of the total crimes within 250 m of schools.

## Q3: How do I create an SQL query with a date field?

Time and Date are implicit attributes of spatial data. Why? Because all spatial data is collected at a given date and time. Some spatial data, like the Crime layer, has information on the date for at which an observation occured. In the case of the Crime layer, there is an OCCDATE field that tells us the date at which a given point observation occurred. The date format is given as:

month/day/year

So, for example, a point with an OCCDATE of 7/1/1998 means that the event in question occurred July 1st, 1998. To query a Date field in a file geodatabase or shapefile you need to precede the date value with the keyword **date** (without quotes), for example, if I wanted all crimes that occurred on July 1st, 1998, I would enter the following in the Select Layer By Attribute dialog:

OCCDATE = date '07/01/1998'

This would select all records (and features) that occurred in July 1st, 1998. For example, make the above query on the Crime layer using the Select Layer By Attribute dialog:
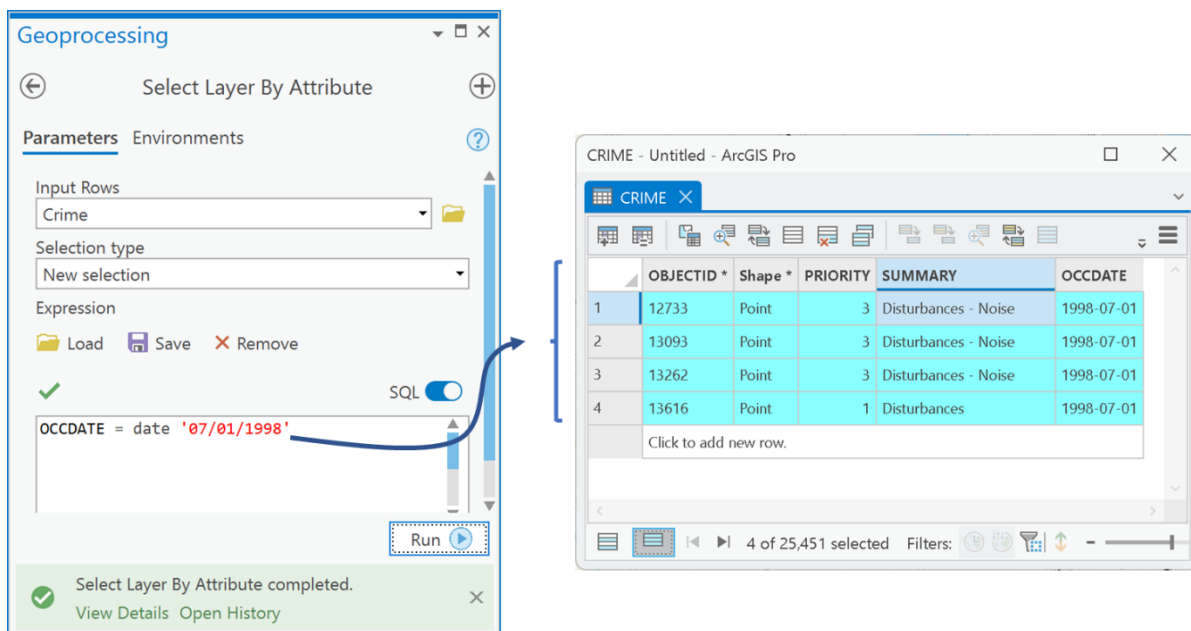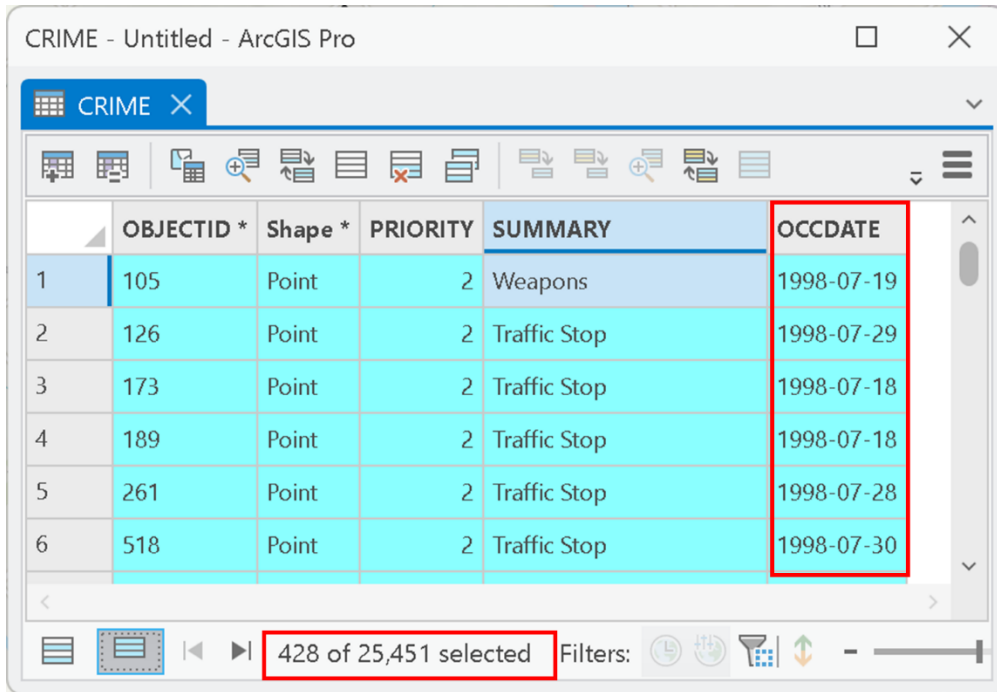

You can also query for all dates within a range of dates using relational operators like <,>,<=,>=. For example, if you want all crimes within the month of July 1998 you could use the following SQL statement:

OCCDATE >= date '07/01/1998' AND OCCDATE < date '08/01/1998'

This would select all dates in July since the "<" symbol is used to ensure that no date, after or including August 1st is selected.

You can also use date conditions together with conditions based on other fields. Suppose that you wanted to get all assaults in the month of January 1999, you could use the following statement:

OCCDATE >= date '01/01/1999' AND OCCDATE < date '02/01/1999' AND SUMMARY = 'Assault'

Q4: How do I undertake centrographic measures for a point set?

In Lecture 6 you were introduced to centrographic measures of point sets. These measures include the *mean center*, *standard distance (SD)* and *standard deviation ellipse*. These centrographic measures are often used for comparing distributions of point phenomena at different times within one area or spatial domain (Figure 7). Figure 7a illustrates a south-east shift in the mean center of assault crimes in Ottawa between January and June 1999. Moreover, the standard distance circles for these two months indicate overall dispersion of assaults remained the same as indicated by the size of the SD circles (Figure 7a). However, the standard deviation ellipses for the two months indicate that, for January 1999, that the directional dispersion was slightly N-S direction and the orientation shifted more NNE-SSW in to July 1999 (Figure 7b).
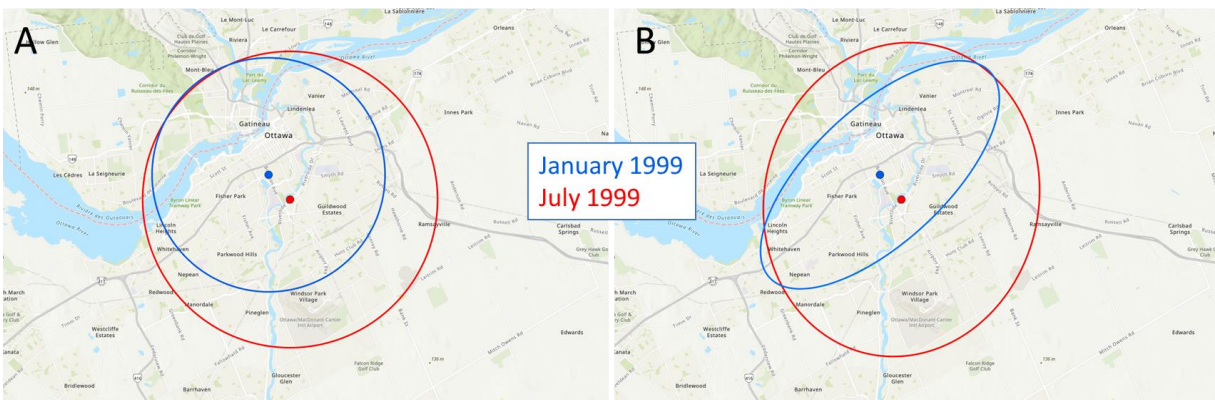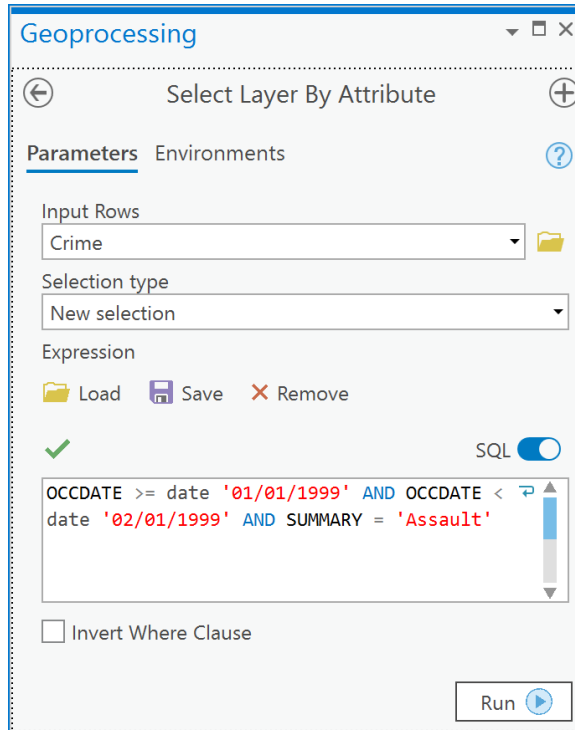

Figure 7: A) Standard distance circles and B) Standard deviation ellipses for Assaults in Ottawa 1999.
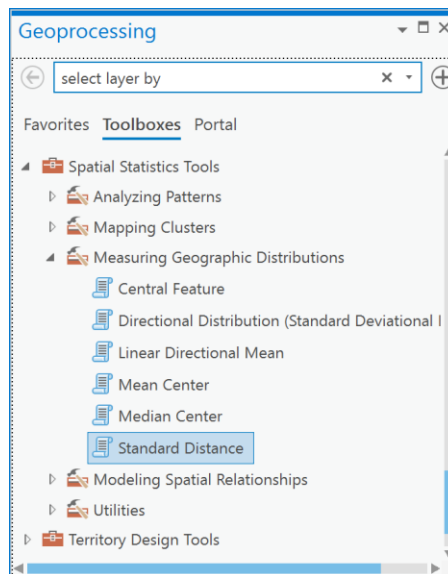
Suppose you wanted to compare the dispersion of assaults in Ottawa between January 1999 and February 1999. To do this analysis you would:
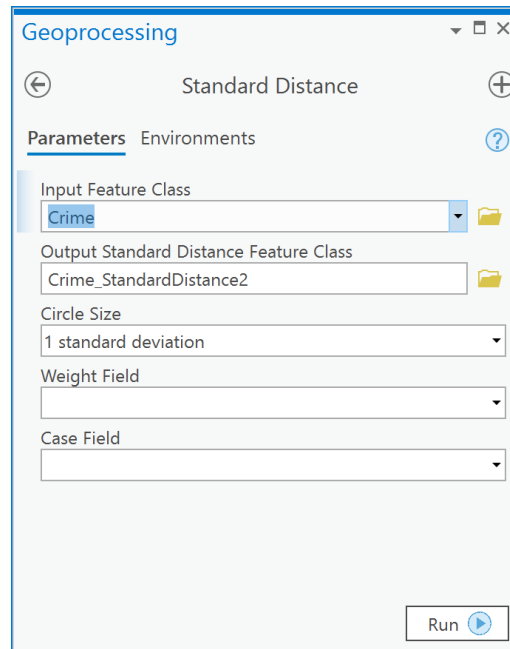
1. Use Select Layer By Attribute to query for assaults in January 1999:

2. Go to Spatial Statistics tools -> Measuring Geographic Distributions -> Standard Distance,

3. In the 'Input Feature Class' drop-down choose Crime and Click Run.

4. The result is a polygon shapefile that shows a circle with 1 standard distance that represents the dispersion of assaults in January. Because we used a query to select a subset of features from the entire crime dataset, **only the selected records were used** in calculating the standard distance. This is the default behavior of almost all tools, because when there is a selection within a layer in ArcGIS and that layer is used in a Tool, the tool only operations on the selected features.
5. Next, repeat 1 to 3 for February 1999.
6. The result (Figure 8) will show 2 standard distance circles on your map, one for January 1999 assaults and 1 for February 1999 assaults:
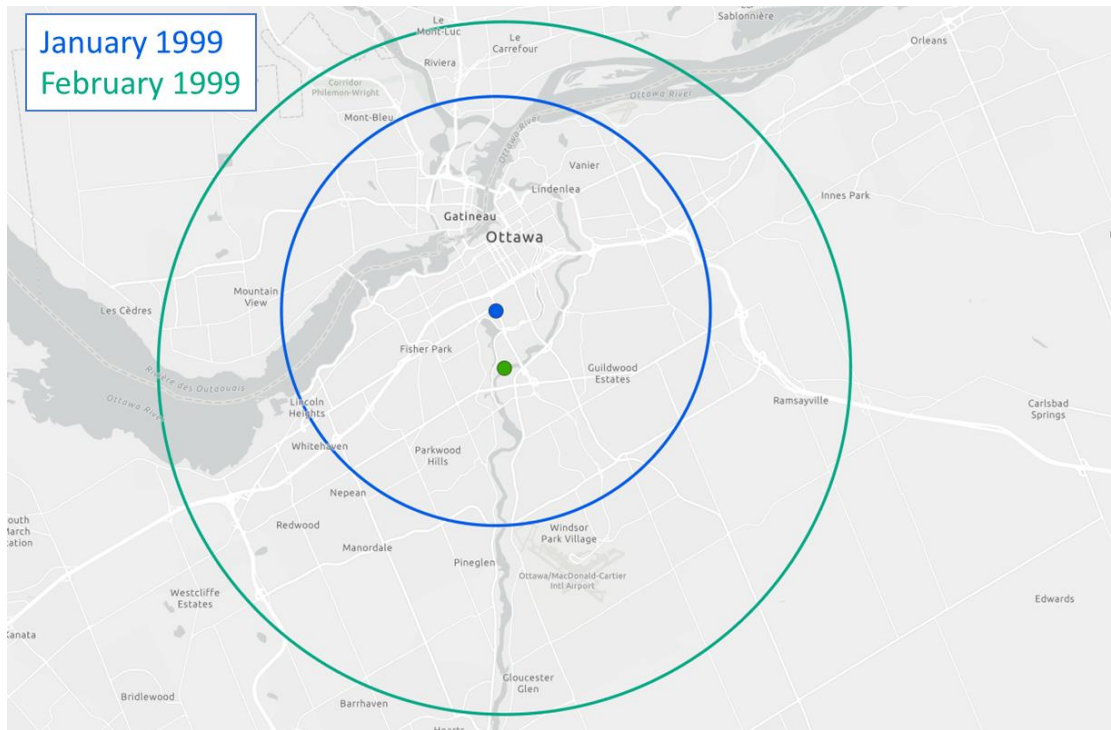
Figure 8: Standard distance circles comparing dispersion of assaults in January and February 1999 within Ottawa

Assaults are more *dispersed* in February when compared to January 1999 as shown by the larger standard distance circle in February (Figure 8). Since the centers of each circle are not coincident there seems to be a shift of the *mean center* of activity towards the east in February.

We could have used the Python pane instead to undertake the above analysis, for example to acquire all the centrographic measures – mean center, standard distance and standard deviation ellipse - for January assault crimes,

```python
# Create string variables to hold
# the SQL where clause as well
# as the names of the input and
# output layers
input_lyr = 'crime'
out_mean_center = 'january_mean_center'
out_sd = 'january_sd'
out_sde = 'january_sde'
query_string = "OCCDATE >= date '01/01/1999' AND OCCDATE < date '02/01/1999' AND SUMMARY = 'Assault'"

# Select crimes within the date range
SelectLayerByAttribute(input_lyr, where_clause=query_string)

# Create a layer with a point for the mean center
MeanCenter(input_lyr, out_mean_center)
```

```
# Create a layer with a circle for standard distance
StandardDistance(input_lyr, out_sd)

# Create a layer with a standard deviation ellipse
DirectionalDistribution(input_lyr, out_sde)
```

The advantage of the above code is that when we want to re-run the python statements for February, all we must do is change the string variables at the top, mainly replacing 'january' with 'february' and modifying the month's in the SQL where clause:

```
input_lyr = 'crime'
out_mean_center = 'february_mean_center'
out_sd = 'february_sd'
out_sde = 'february_sde'
query_string = "OCCDATE >= date '02/01/1999' AND OCCDATE < date '03/01/1999' AND SUMMARY = 'Assault'"

SelectLayerByAttribute(input_lyr, where_clause=query_string)

MeanCenter(input_lyr, out_mean_center)

StandardDistance(input_lyr, out_sd)

DirectionalDistribution(input_lyr, out_sde)
```

Q5: How can I visualize the spatial variability in point intensity within a region?

In Lecture 6 you were introduced to the kernel density estimate (KDE). A KDE shows the variation in intensity, called lambda, λ, across a spatial domain. A KDE is represented as a kernel density surface that is created from a set of points. The KDE is created by placing a 3D mathematical function, called a kernel, whose maximum value is centered at the location of each point in a pattern. Then, the z-values (height values) of all the 3D mathematical functions (the kernels) are added together (Figure 9) at each location in space to create the KDE surface. This process results in a raster layer that shows high values where there are lots of points and lower values where there are fewer points (Figure 9).
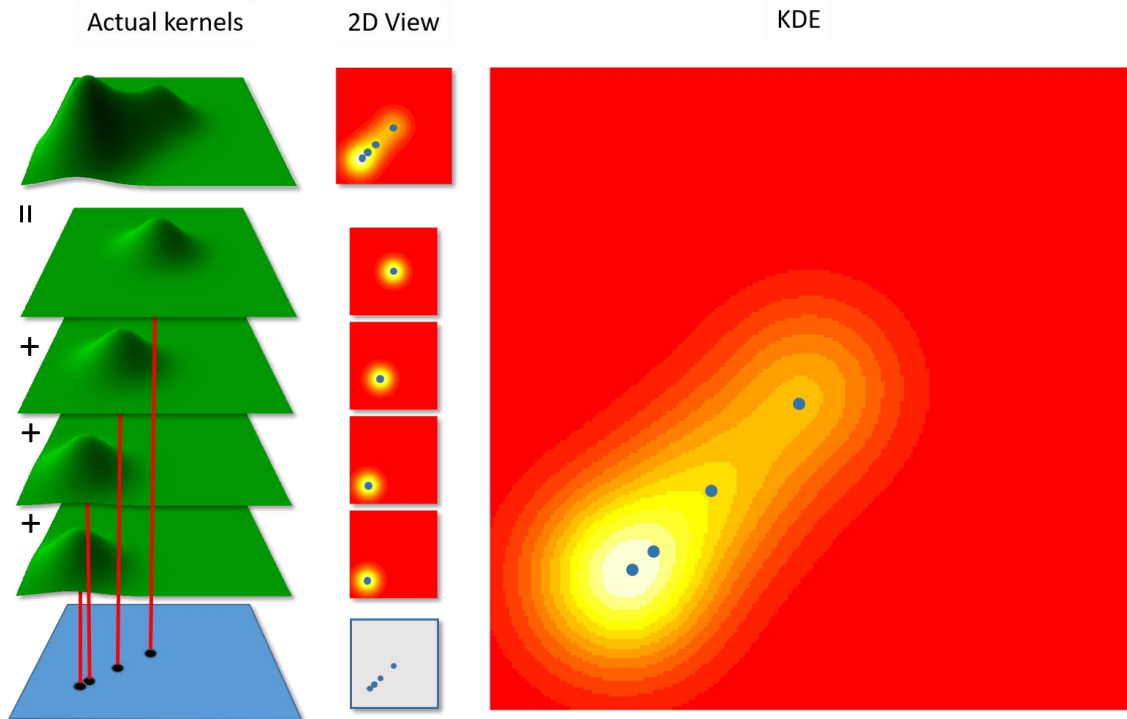
Figure 9: The creation of a KDE. Gaussian 3D kernels (left) placed on top of each point in a point pattern are summed at each location within the study region.

For example, you could create a KDE for assaults in January 1999 and compare this to the KDE for assaults in February 1999:

1. You will need to ensure in ArcGIS that you load the Spatial Analyst extension. To do so, click on the Project tab and then look under 'Esri Extensions' for Spatial Analyst – you may need to scroll down. If there is a checkbox there check it and return to the map tab, if there is no checkbox and the word 'no' appears, then click on the button that says, 'Configure your licensing options'. You will see a list and again find the one that says Spatial Analyst. Thereupon, you should see a checkbox. Check it and click OK and return to your map tab. If there is no checkbox then choose the line that contains Spatial Analyst and click the Authorize button. Once finish, click OK and return to your Map Tab. Yours will be somewhat similar or identical to what you see below:
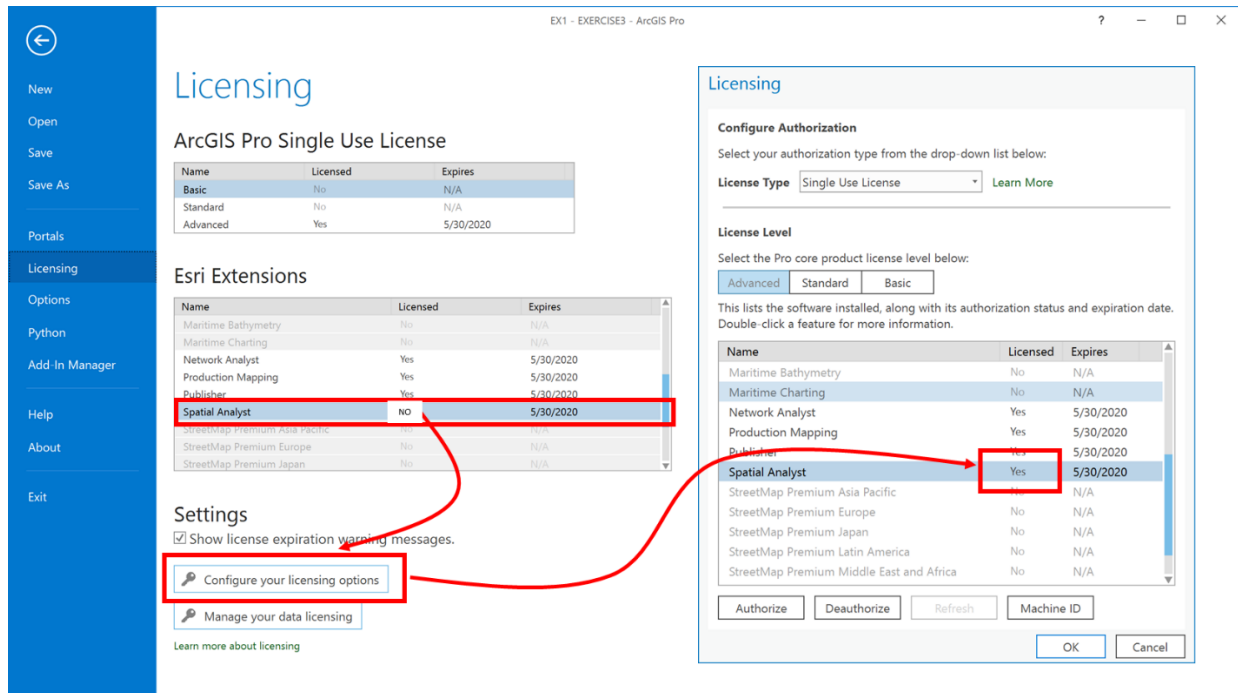
2. Now, add the layer called Boundary from the OTTAWADATA feature dataset to your map.

3. You need to set a number of key Environment parameters in order to work with the Spatial Analyst tools in the geoprocessing pane. Specifically, you must set the Processing Extent, Cell Size and Analysis Mask within the ArcGIS Pro's Environment settings. Go to the Analysis tab and in the Geoprocessing group click on the Environments button and set your parameters to the same as you see in 1,2,3 below:
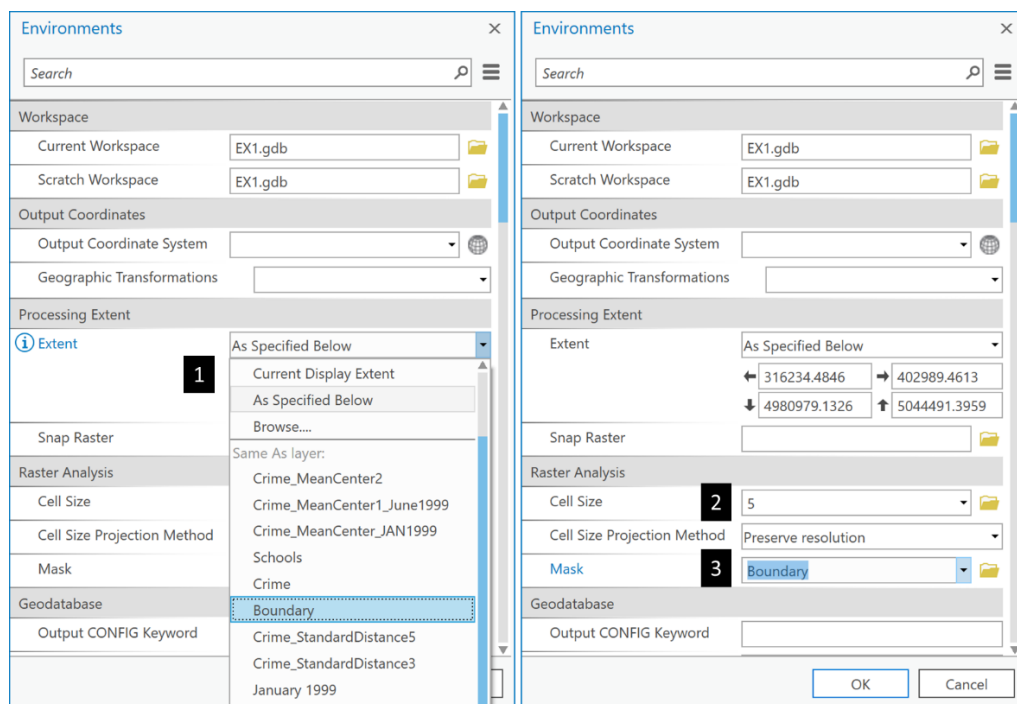
4. Now, select all Crime points that are assaults in January 1999 as you did in EX Q4 above:
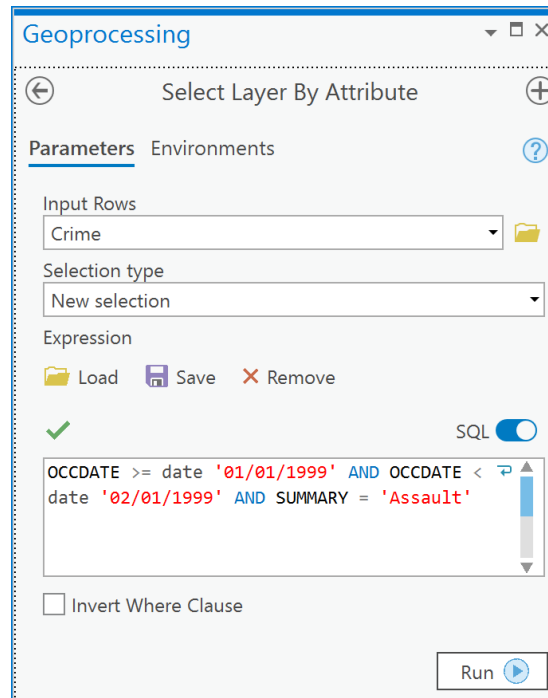
5. Then go to Spatial Analyst Tools -> Density -> Kernel Density. Under 'Input point or polyline featuers' choose Crime. Under 'Output raster' call it 'january'. Here the 'Search radius (optional)' is the bandwidth setting. ArcGIS calculates a reasonable default for most cases using Silverman's Rule (see help file). Ensure all other settings are as you see below and click Run:
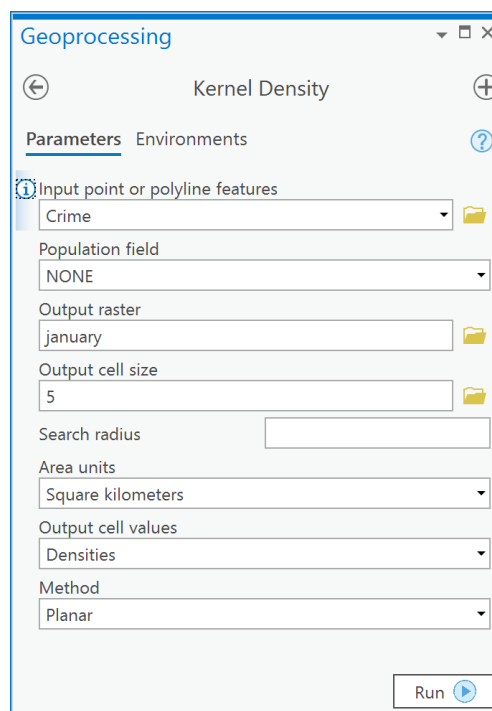
6. Set critical environment settings for the tool-level by selecting the Environments tab and fill in the following:
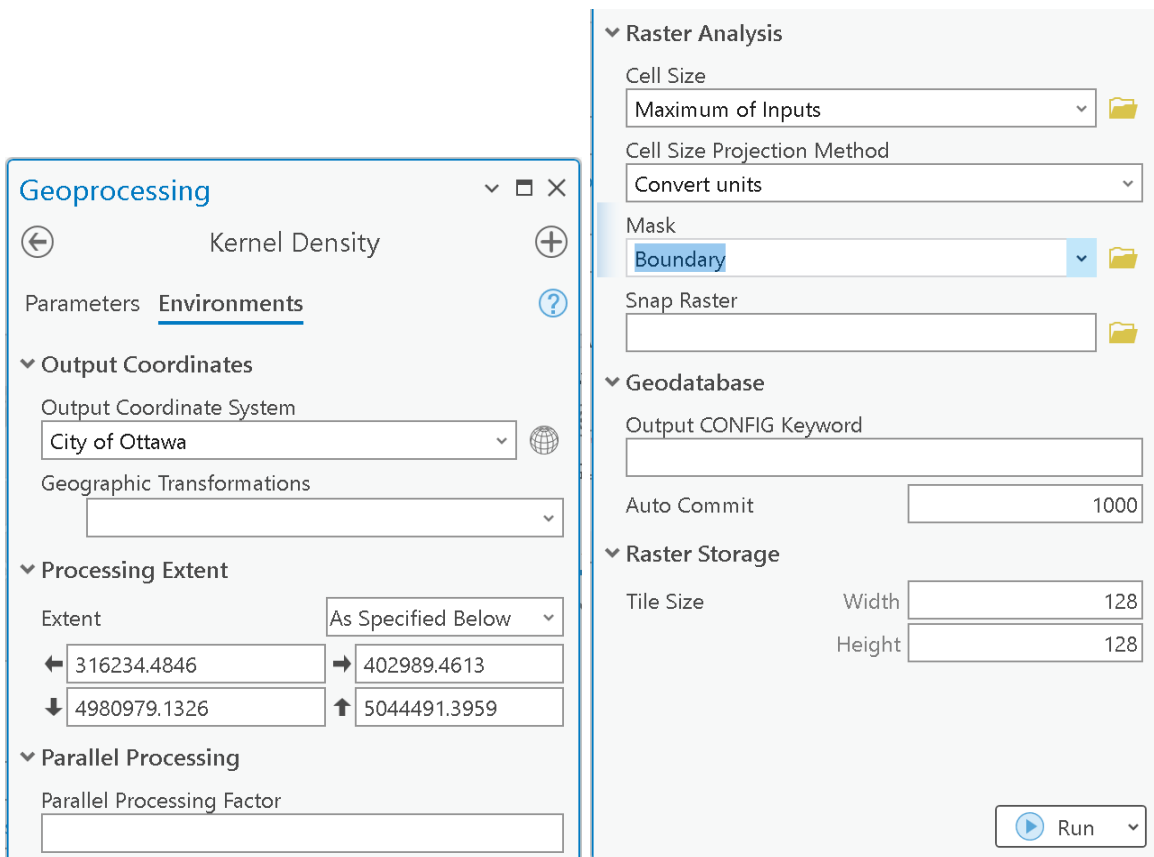
7. Repeat 3-6 for February 1999 calling the output KDE 'february'.
8. To compare two or more KDE estimates, the **z-scales are the same on both maps must be the same, o**therwise a color value such as the darkest blue on one map would not correspond to the same magnitude of density on the other – likewise with any color on the z-scale. The easiest way to make two maps with the same z-scale is to use the Apply Symbology from Layer geoprocessing tool to import the symbology of an existing layer you have already created a classification for. In the case of the assaults, I import from the layer that has the maximum density, which in our case is January, so I import January's symbology to February's:
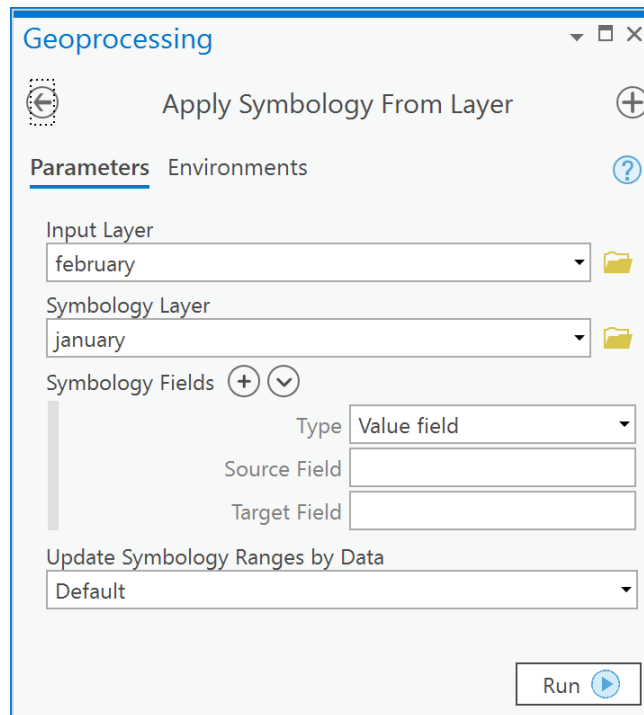
9. Your results (Figure 10) allow you to compare the two months intensity surfaces.
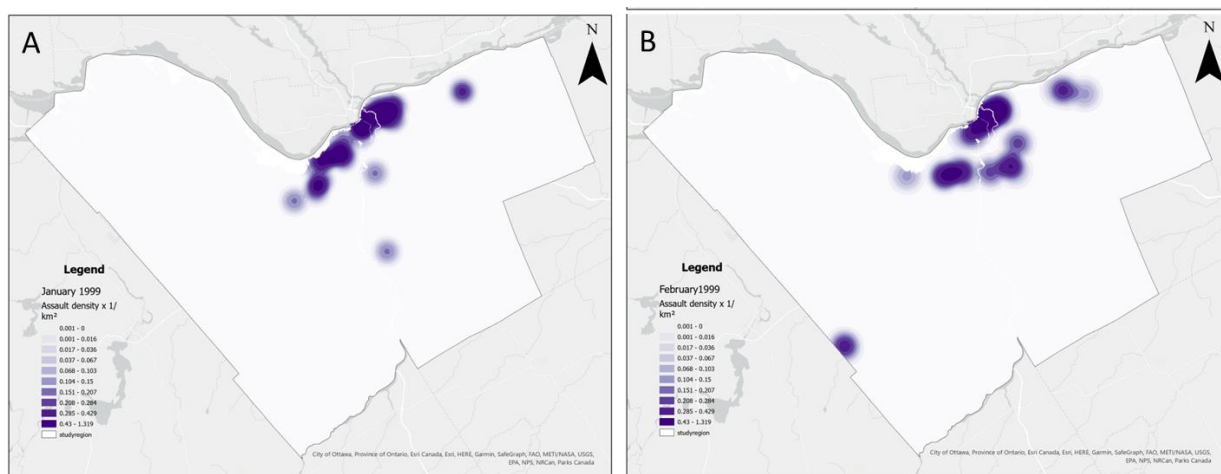

Figure 10: Kernel density estimates for assaults in A) January 1999; B) February 1999

You can see that the intensity is higher between the Rideau River and Rideau canal in January 1999 when compared to February 1999 (Figure 10). In addition, there is more spread to the intensity values in February 1999 (Figure 10). These observations are consistent with the standard distance observations in EX Q4, the higher intensity in the same location in January 1999 leads to a smaller standard distance, whereas the more diffuse intensity in February 1999 leads to a larger standard distance.

Another, more efficient way to compare to maps of the same variable is to compute the difference between them. Think about it. When comparing two quantitative maps you are
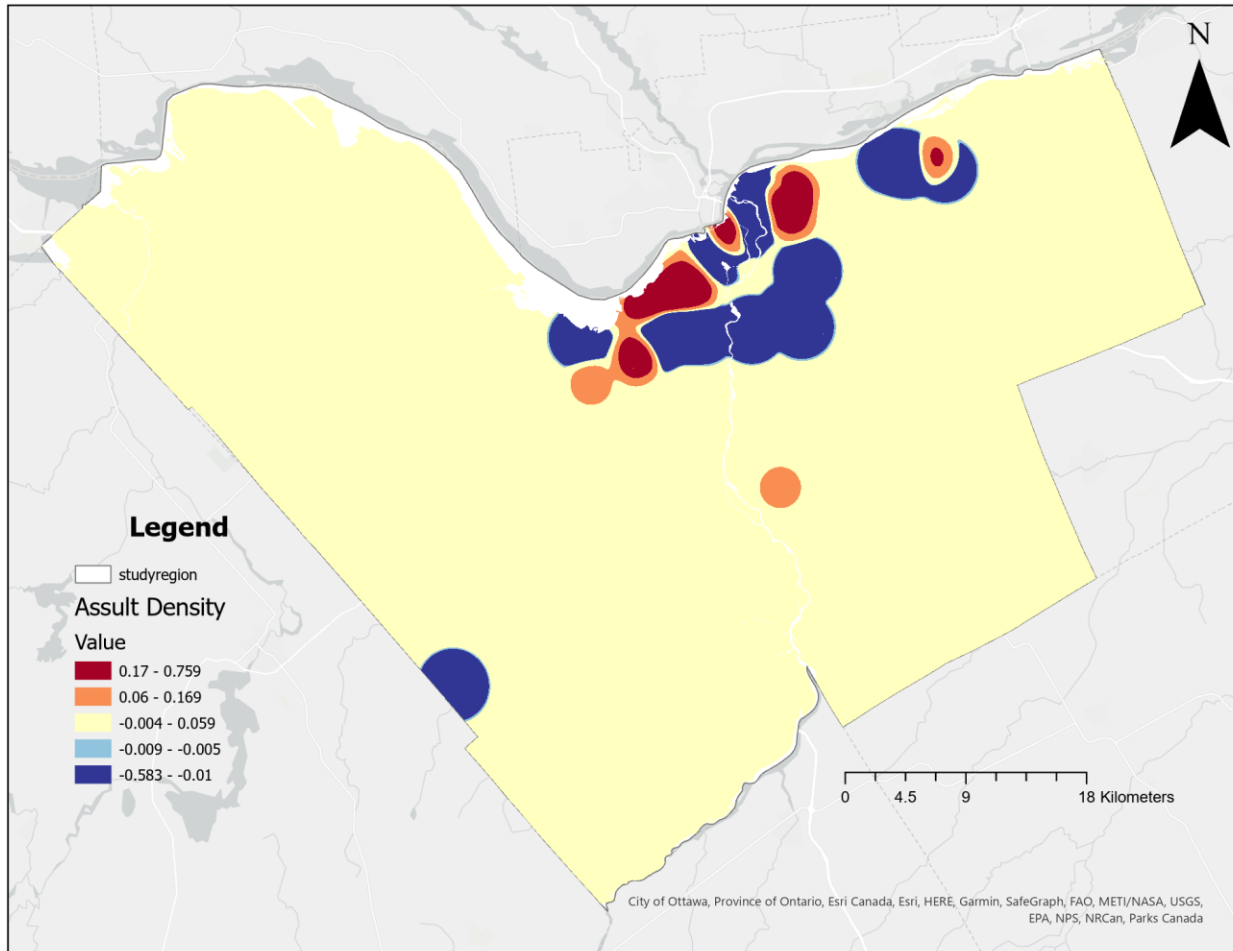
concerned with where they are different and where they are the same. A difference map can help to quickly identify incongruities between two time periods. We can easily compute the difference between the assault densities between January and February using the Python tool in ArcGIS Pro.

1. Go to the Analysis tab and click on the Python button.
2. Enter the following lines except for the ones preceeded by a # which are comments for you to read in order to understand what the python statements mean.

```python
# Import the spatial analyst toolbox
# so we can call tools by their name
from arcpy.sa import *

# Create python raster variables from
# the january and february rasters
jan = Raster('january')
feb = Raster('february')

# Compute the difference between January and
# February and make a new layer called diff
diff = jan - feb
```

3. The result will be a map of differences. Where the map has positive values, January had a higher intensity of assaults than February. Where the values are less than zero, February had a higher intensity of assaults,

Legend

studyregion

Assult Density
Value

- 0.17 - 0.759
- 0.06 - 0.169
- -0.004 - 0.059
- -0.009 - -0.005
- -0.583 - -0.01

City of Ottawa, Province of Ontario, Esri Canada, Esri, HERE, Garmin, SafeGraph, FAO, METI/NASA, USGS, EPA, NPS, NRCan, Parks Canada

Q6: How do I calculate assaults per unit area within a region?

In Lecture 6 you were introduced to some point-polygon relations. One common relation that is used is to examine the density of points within polygons. To do so requires that you count each point within each polygon and then divide that count by the area of the polygon to arrive at a measure of points per unit area. This type of operation is easily done in GIS because point-in-polygon is a common operation. To count the number of points in any number of polygons within the same spatial domain, you only need to undertake a spatial join using the Spatial Join tool in ArcToolbox in order to undertake spatial joins when you only want to utilize the selected records from a query within the spatial join operation.

Suppose that you want to calculate, within Ottawa neighborhoods, the density of assaults from January-March 1999 (inclusive) and compare this with the dispersion of assaults in June-August 1999 (inclusive):

1. Create a query on the Crime layer for assaults between January 1999 and March 1999,

OCCDATE >= date '01/01/1999' AND OCCDATE < date '04/01/1999' AND SUMMARY = 'Assault'
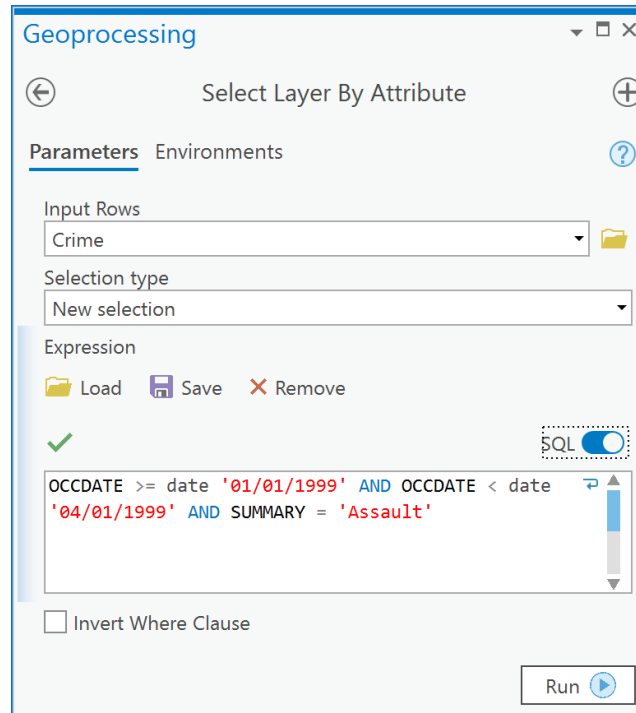
2. Go to Analysis Tools -> Overlay -> Spatial Join. The 'Target Features' will be the polygon layer **Boundary** which contains the Neighborhoods, the 'Join Features' is the **Crime** layer containing the selected records from the query in Step 1 above. The 'Output Feature Class' should be called **Neighbourhood_Assaults**. Put a check mark in the 'Keep All Target Features". Under 'Field Map of Join Features', delete all fields except 'names' which will retain only the neighborhood name in the output. This tool will automatically create a 'Join_Count' field in the output feature class that contains the number of points found in each neighbourhood polygon in **Boundary**. Before clicking Run, ensure your Spatial Join dialog options are the same as those below:
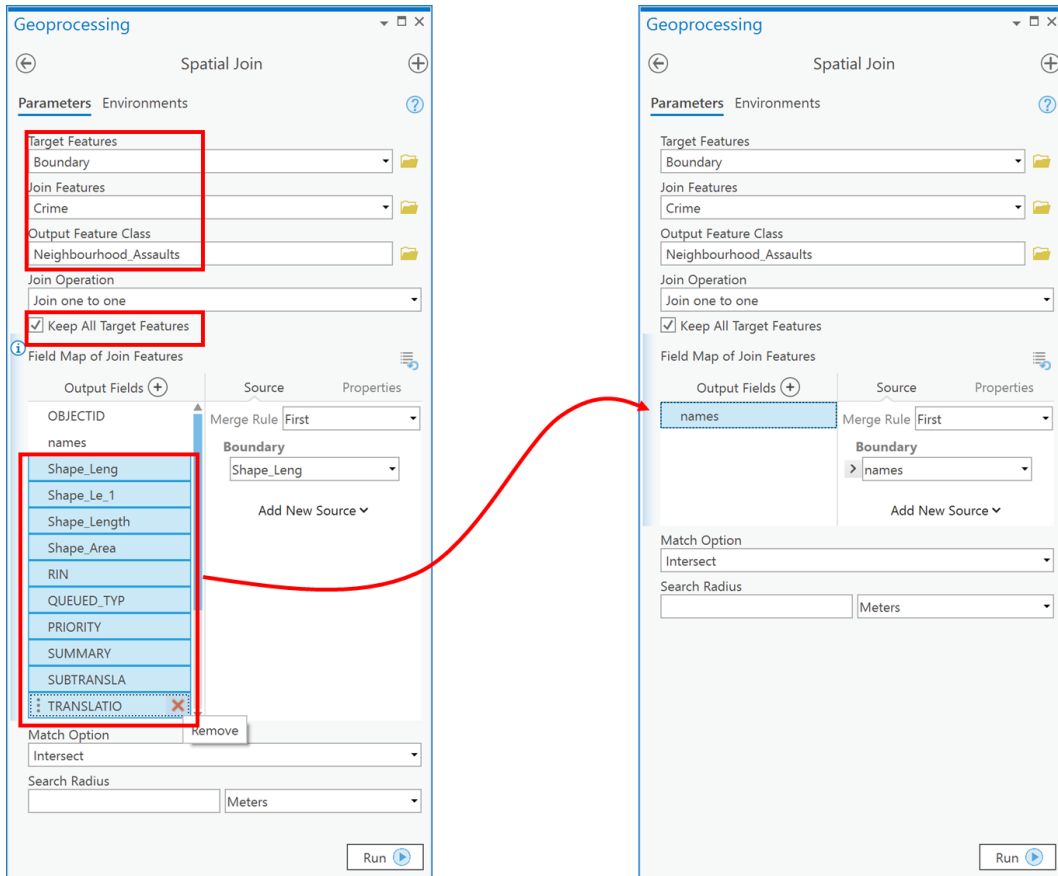
3. You will now have a new feature class in your table of contents. Open the attribute table for the new feature class and you will see the following table with a new field called 'Join_Count' that contains the number of assaults within each neighborhood. The 'Shape_Area' field contains the area of each neighborhood in the map units (m^2) of the current coordinate system (which is the same coordinate system by default as the target layer):
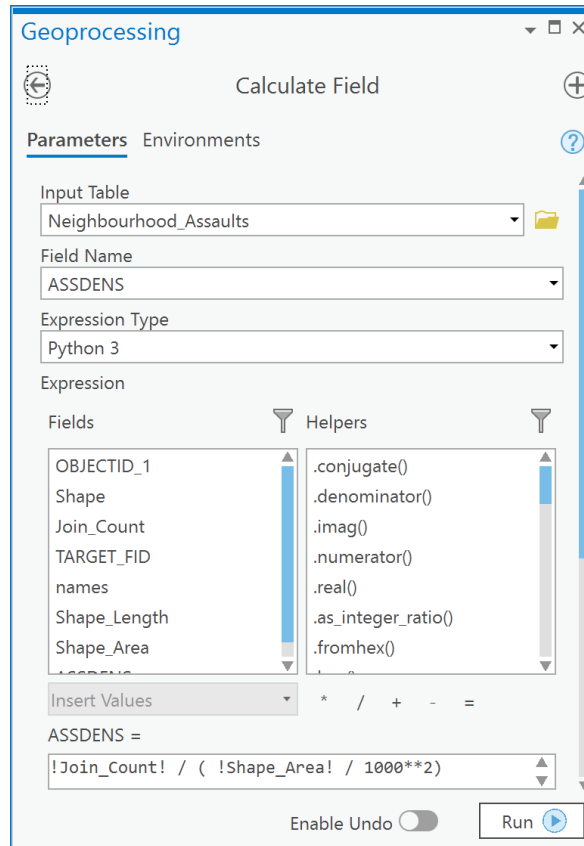
4. Next add a new field to the table called ASSDENS for assault density and make it of type double (Review EXERCISE 2 if you do not remember how to do this).
5. Now, calculate, in the ASSDENS field, the assault density in persons per square kilometer using an expression such as:

$$!Join\_Count! / (\ !Shape\_Area! / 1000**2)$$

6. Next, create a choropleth map of the Assault density using the ASSDENS field. Click on the Neighbourhood_Assaults layer in the Contents pane. Then in the Feature Layer ribbon, choose the Appearance tab. In the Drawing group click on the Symbology button and choose Graduated Colors. In the Symbology pane, under 'Field' choose ASSDENS by choosing Graduated colors in the Symbology tab, e.g.:
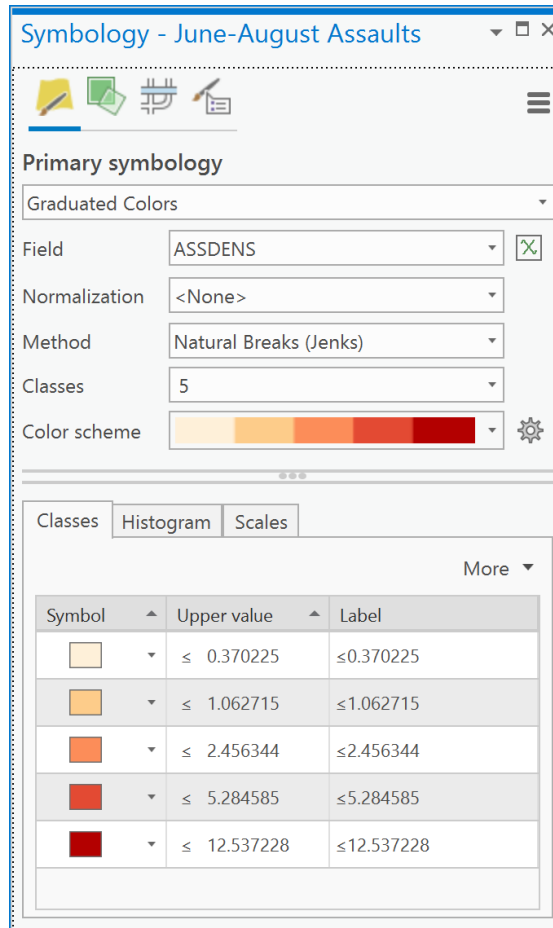
7. Repeat steps 1-6 for June-August 1999 (inclusive).
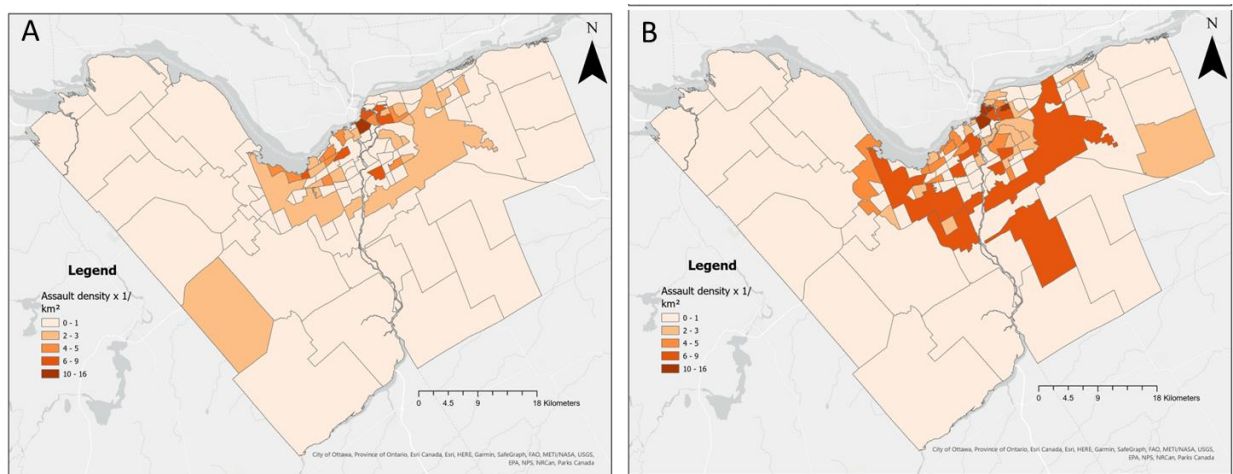8. The results (Figure 11) allow you to compare the density by neighborhoods between the two timer periods.

Figure 11: Assaults in Ottawa Neighborhoods in A) January-March 1999; B) June-August 1999

Notice again, that when comparing two maps of density, the z-scale of assaults per square kilometer is the same for both maps (Figure 11), otherwise the same color on each map would

represent a different value and that would lead to incorrect comparisons. Overall the patterns in the two seasons are quite similar, however, the density of assaults is higher in more neighborhoods in the summer in central Ottawa when compared to winter. The pattern in summer also exhibits more spatial similarity in the central neighborhoods.

This example took a number of steps and these could be completed in the Python pane as well. The advantage again being that to complete the same analysis for another time window, you would only need to change the values in the statements that are highlighted in yellow:

```python
# Import the management & analysis
# toolbox tool functions.
from arcpy.management import *
from arcpy.analysis import *

input_lyr = 'crime'
query_string = "OCCDATE >= date '01/01/1999' AND OCCDATE < date '04/01/1999' AND SUMMARY = 'Assault'"
nbrhoods = 'Boundary'
spatial_join_output = 'Neighbourhood_Assaults'
density_field_name = 'ASSDENS'
field_to_keep = 'names'

# First, select the assaults using the SQL where
# clause held in query_string
SelectLayerByAttribute(input_lyr, where_clause=query_string)

# Create a field mapping object. This is complex but
# the idea is that this ensures that only the
# neighbourhoood names are transferred in the spatial join.
fm = arcpy.FieldMap()
fms = arcpy.FieldMappings()
fm.addInputField(nbrhoods, field_to_keep)
of = fm.outputField
of.name = field_to_keep
fm.outputField = of
fm.mergeRule = 'first'
fms.addFieldMap(fm)

# Spatially join the selected crime points
# to the neighbourhoods and transfer
# only the names field to the output
# using the fieldmapping variable fms
SpatialJoin(nbrhoods,
 input_lyr,
 spatial_join_output,
```

```
    'JOIN_ONE_TO_ONE',
    'KEEP_ALL',
    field_mapping=fms
    );

# Add a field called "ASSDENS" to the spatially
# joined layer
AddField(spatial_join_output,
 density_field_name,
 'DOUBLE'
 );

# Now calculate the field as the number of
# assaults per square kilometer
CalculateField(spatial_join_output,
 density_field_name,
 '!Join_Count! / ( !Shape_Area! / 1000**2)'
 );
```

Q7: How do I create a union overlay and why?

NOTE: *Create a new ArcGIS Pro map before starting this example. Also, if you have kept ArcGIS Pro open from the previous example(s), ensure you reset the Geoprocessing Environments-> Processing Extent back to option 'Default'.* ***If you do not do so*** *then the spatial extent of Ottawa will apply for all analytical outputs and this will lead to you not seeing the results at the scale of North America.*

In Lecture 6 you were introduced to the intersection and union overlay. A union overlay is used when you want to determine the geometric union of any number of polygons. The output of a union operation will "contain polygons representing the geometric union of all the inputs as well as all the fields from all the input feature classes" (ESRI). Thus, a union combines both polygons and their attributes (Figure 12). In Figure 12, there are two layers, one layer has an orange polygon with an attribute ORANGE and a green polygon with attribute GREEN. The second layer is a yellow triangle with the attribute YELLOW. When the yellow polygon is unioned with the orange and green polygon layer, wherever the yellow polygon intersected the orange polygon, the attributes of ORANGE and YELLOW are combined, and a new polygon is created with the two attributes ORANGE and YELLOW in two separate fields. By examining the unioned attribute table we can see which of the new polygons generated intersected because they have values in both new polygon's attributes. The exact same process is true for the yellow and green polygon intersection. Where there were no intersections taking place, the new polygons generated simply have either an attribute that is ORANGE or GREEN.
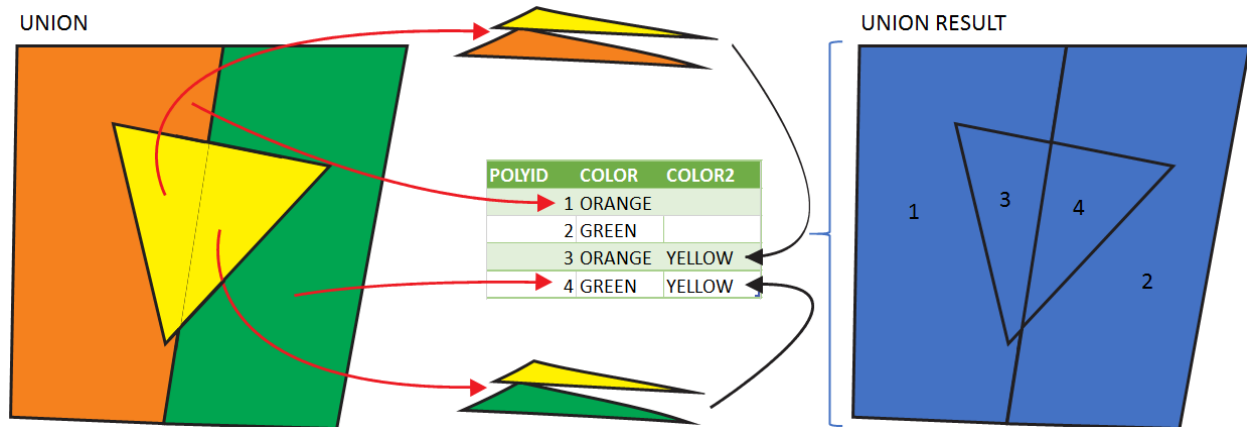
Figure 12: Geometric union example

If you were studying tree species richness (the number of different types of species in a given area) in North America, you could union overlay a set of polygons that represent range boundaries for different species.

Consider three range boundaries (Figure 13), for three species of Walnut (*Jugulans* spp.) in North America (from https://www.fs.fed.us/nrs/atlas/littlefia/species_table.html), namely Black walnut, White walnut and Little walnut. These are added to ArcGIS Pro from the TREES feature dataset in the OTTAWADATA.gdb file geodatabase in the Lecture7/Data folder. Each of the walnut layers contains a set of polygons. In the attribute table for each layer, there is a field called CODE that specifies the value of '1' for each of the polygons in each layer (Figure 13). Placing these three layers into a union overlay processes produces a new polygon layer (Figure 13). The new unioned polygon layer has a set of new polygons and a new attribute table. The new attribute table contains the combined set of attributes for all the input layer attribute tables. Specifically, you can now see the combination of CODE fields (Figure 13). The value of CODE was '1' in each of the input layers for each input polygon and thus, for a given output polygon in the unioned layer, a value of '1' in any one of the three unioned CODE fields (CODE, CODE1, or CODE2) signifies that, for the unioned polygon in question, one of the three input layers contained a tree species at that given location (If you had undertaken an intersection overlay of the three input layers, the only polygons returned would be ones in which all three input layers contained a polygon). A new field can be added to the unioned table called "SUMCODE" and, when calculated as the sum of the individual unioned CODE fields, this represents the number of species of walnut that overlap in each unioned polygon.
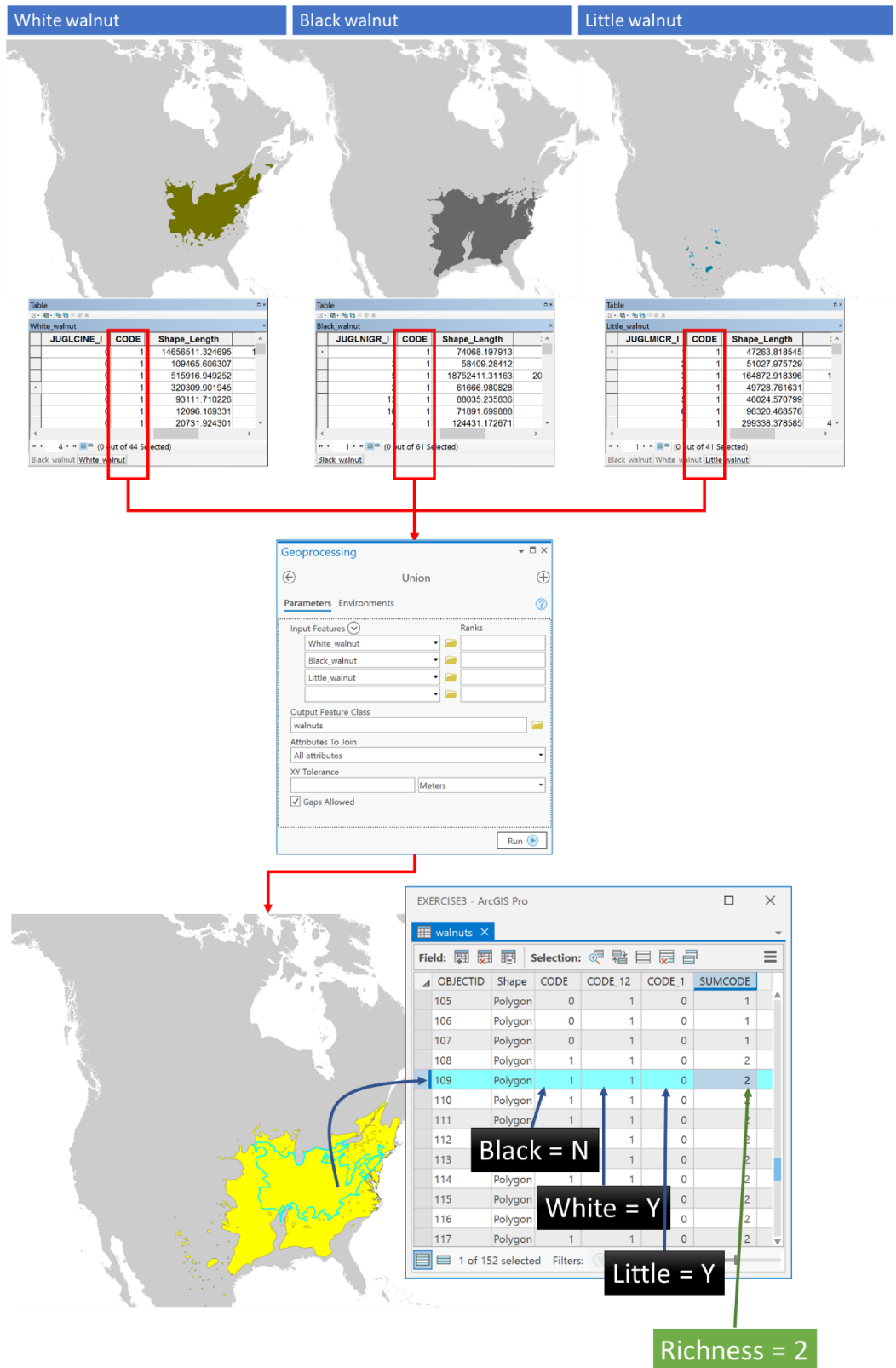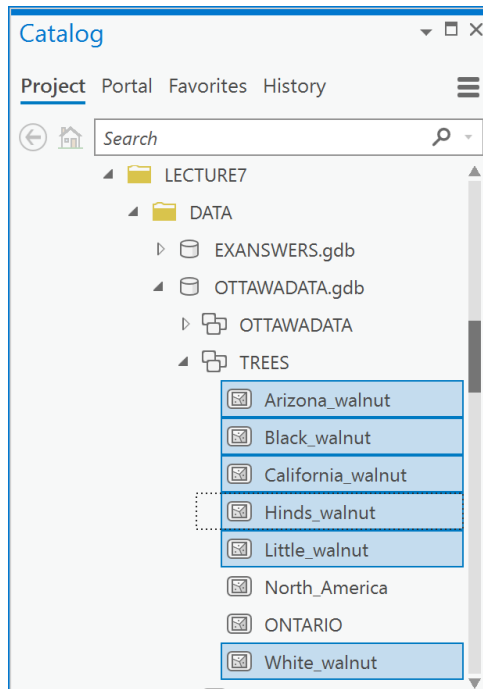
Figure 13: Process example of union overlay combining geometry and attributes in the output for three species of walnut (Jugulans spp.)
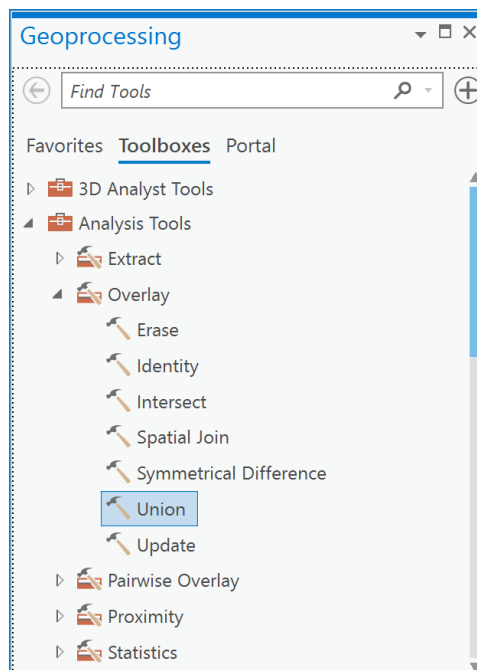
To try this process for yourself:

1. Add the following layers to ArcGIS Pro:

2. Choose Analysis Tools -> Union in the Geoprocessing pane.

3. In the Union dialog box, add the six input layers and call the output 'walnutunion',
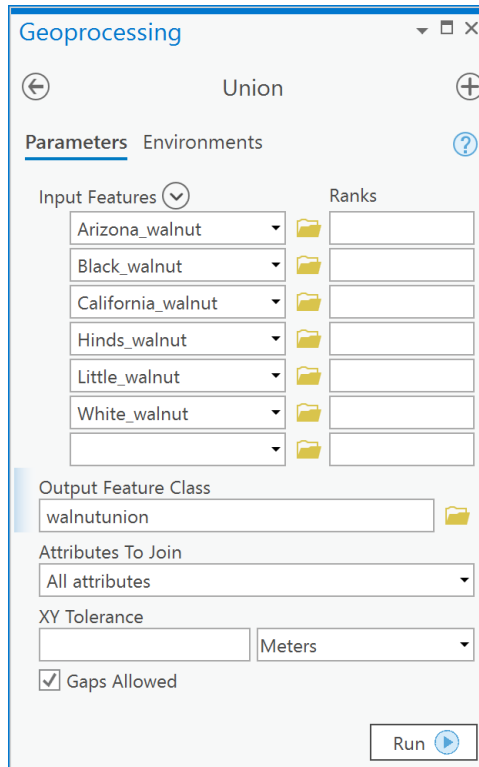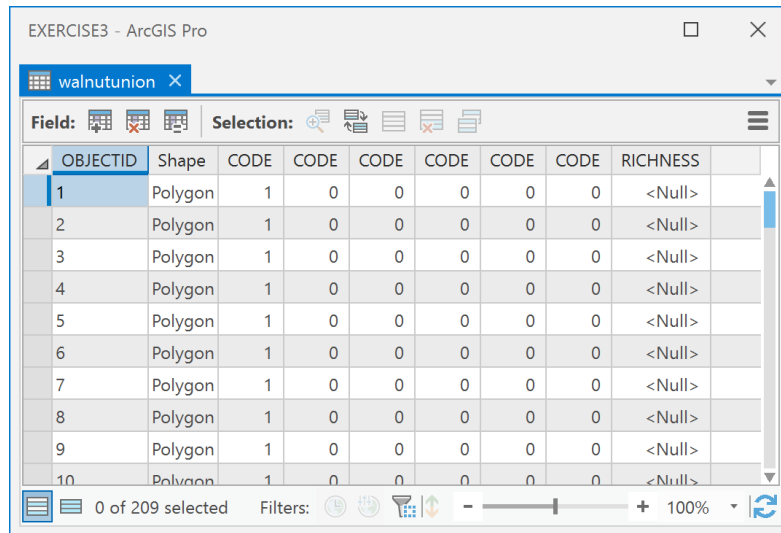
4. Open the attribute table for the resulting layer (Note that you will see many more fields than this because I have, in the fields view, turned off all fields but the CODE fields)

5. Add a new field called "RICHNESS" of type "Long Integer".

OBJECTID	Shape	CODE	CODE	CODE	CODE	CODE	CODE	RICHNESS
1	Polygon	1	0	0	0	0	0	<Null>
2	Polygon	1	0	0	0	0	0	<Null>
3	Polygon	1	0	0	0	0	0	<Null>
4	Polygon	1	0	0	0	0	0	<Null>
5	Polygon	1	0	0	0	0	0	<Null>
6	Polygon	1	0	0	0	0	0	<Null>
7	Polygon	1	0	0	0	0	0	<Null>
8	Polygon	1	0	0	0	0	0	<Null>
9	Polygon	1	0	0	0	0	0	<Null>
10	Polygon	1	0	0	0	0	0	<Null>

0 of 209 selected   Filters: – ———|—— + 100%

6. Calculate the new field as the sum of each of the CODE fields (CODE, CODE_1, CODE_12). Use a python statement such as:

!CODE! + !CODE_1! + !CODE_12! + !CODE_12_13! + !CODE_12_13_14! + !CODE_12_13_14_15!
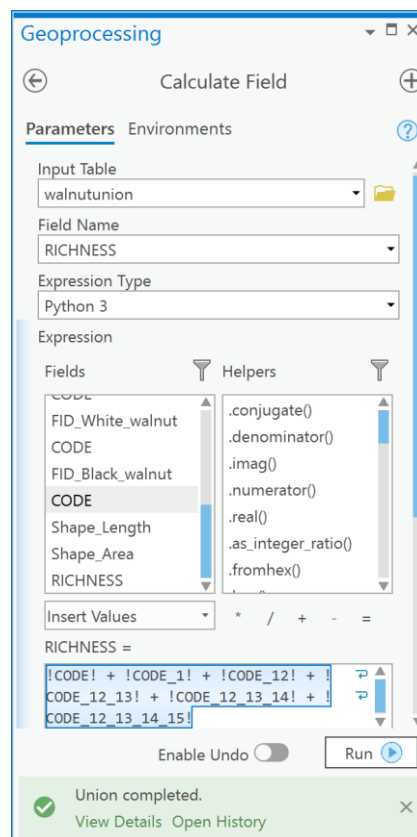
In the field calculator. Remember, an attribute table cannot have two fields of the same name in the same table, thus, when the union process occurs, the original fields called "CODE" in the input layers to the union process are automatically renamed by ArcGIS by appending a "_1" or "_12" etc…

7. Create a choropleth map using a quantitative symbology,
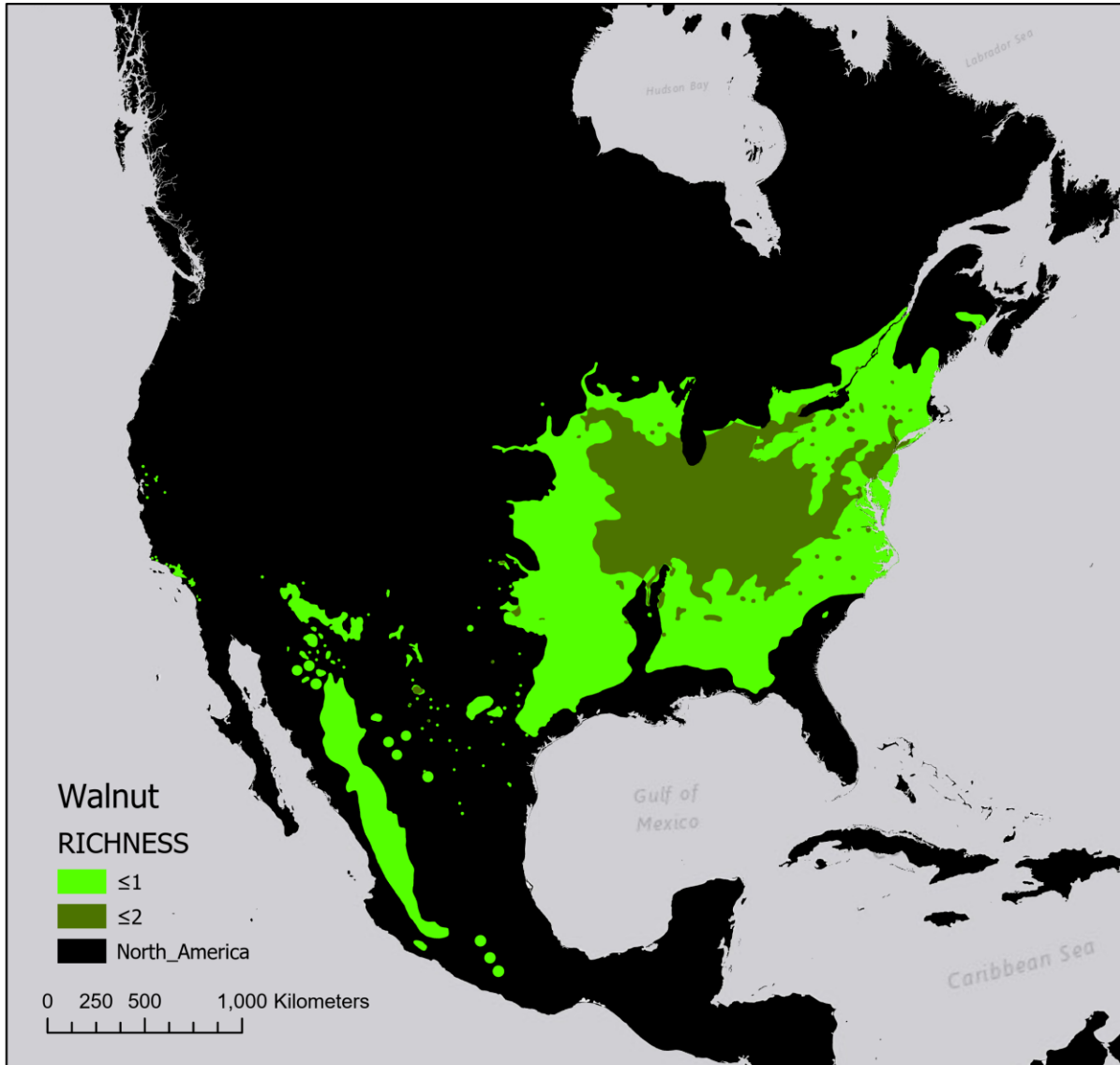
Like our other examples, we could also formulate a set of Python statements to complete this task. This one is a little more complex and you may not fully understand all the statements and that is OK. The advantage of such a set of python statements is that you would only need to change the highlighted statements to create a species richness column for any number of trees.

```python
# Import the management & analysis
# toolbox tool functions.
from arcpy.management import *
from arcpy.analysis import *

# Define output feature class name
# and species richness field name
out_fc="walnuts"
richnessfld="RICHNESS"

# Create a list of layers to union
input_features=['White_walnut',
 'Black_walnut',
 'Little_walnut',
 'California_walnut',
 'Arizona_walnut',
 'Hinds_walnut'];

# Union all the tree layers keeping all
# attributes because we need the CODE
# field from each
ulayer=Union(input_features,out_fc,"ALL")

# Create a list of field names because
# we don't know how many layers will
# be unioned, nor what all the CODE
# fields will be named during the
# union process. We only know that
# they will start with the string
# 'CODE'.
fldlist=["!{}!+".format(fld.name) for fld in
arcpy.ListFields(ulayer,"CODE*")]

# Create the field calculator expression
# which will be like '!CODE!+ !CODE_1!+...'
calcexpr=' '.join([s for s in fldlist])
calcexpr=calcexpr[0:-1]

# Add a field called "RICHNESS" to the
# unioned layer
AddField(ulayer,
 richnessfld,
 'LONG'
 );
```

```
# Now calculate the field RICHNESS
# using the expression string
# stored in calcexpr
CalculateField(ulayer,
 richnessfld,
 calcexpr
 );
```

Questions to hand in.

Each of the questions requires an explanation - so read carefully. Each question requires a map as part of the answer, and this must be a correct presentation map (excluding the graticule is ok but not necessary). Ensure you review Section 2.3 of the style manual that is found at the beginning of these lab notes.

1. [2 points] Using standard distance circles and mean centers, compare the distribution of accidents ('Accident' in the SUMMARY field of the Crime feature dataset) between winter and summer in Ottawa in 1999. Define winter as January to March and summer as June to August. Present a map of the results and explain what these results tell you about the distribution of accidents between seasons.

2. [2 points] For the same winter and summer periods in (1), create KDE estimates for accidents in each season. Compare the maps for each season.

3. [5 points] As part of your job at the City of Ottawa, you are asked to determine which areas of the city should be targeted for selective tree injections of TreeAzin, a systemic insecticide, that kills Emerald Ash Borer larvae. The injection programme is selective and targets "ash tree protection in neighbourhoods with a high percentage of forest cover from ash species" (http://ottawa.ca/en/residents/water-and-environment/plants-and-animals/invasive-species). Go to the City of Ottawa Open Data website and search for the "Tree Inventory". Download and unzip the shapefile. Import that the shapefile into your OTTAWADATA.gdb geodatabase (in the Lecture7 folder). Calculate the density of Ash trees (**trees / km^2**) in Ottawa Neighborhoods. Use the neighborhoods feature class in the OTTAWADATA.gdb/OTTAWADATA feature dataset. In calculating density, ensure you only include *Fraxinus* spp. when creating your query, use a query statement such as:

   "SPECIES" LIKE 'Fraxinus%'

   to select only Ash species. Present a map of the density of ash trees in Ottawa neighborhoods. Provide an explanation of the pattern of Ash tree density in Ottawa. Based on the observed pattern, what recommendation would you provide the city manager regarding the injection programme?

4. [2 marks] Present a KDE estimate for the same Ash trees in (3) and compare the results with (3). What are the similarities and differences you notice? Why might there be similarities and differences based on the methods used to calculate density vs. intensity?

Would your recommendations change or stay the same? Why? (Hint, it may be informative to overlay the neighbourhood boundaries with no color fill on top of the KDE surface).

5. [4 marks] Download the shapefiles for all the *Fraxinus* spp. found at https://github.com/wpetry/USTreeAtlas. There are 16 species in total. You might want to download the entire repository as a zip file and then unzip it so that you can easily access the folders for each species - rather than downloading each individual species. To do so just click on the green button and choose Download Zip:
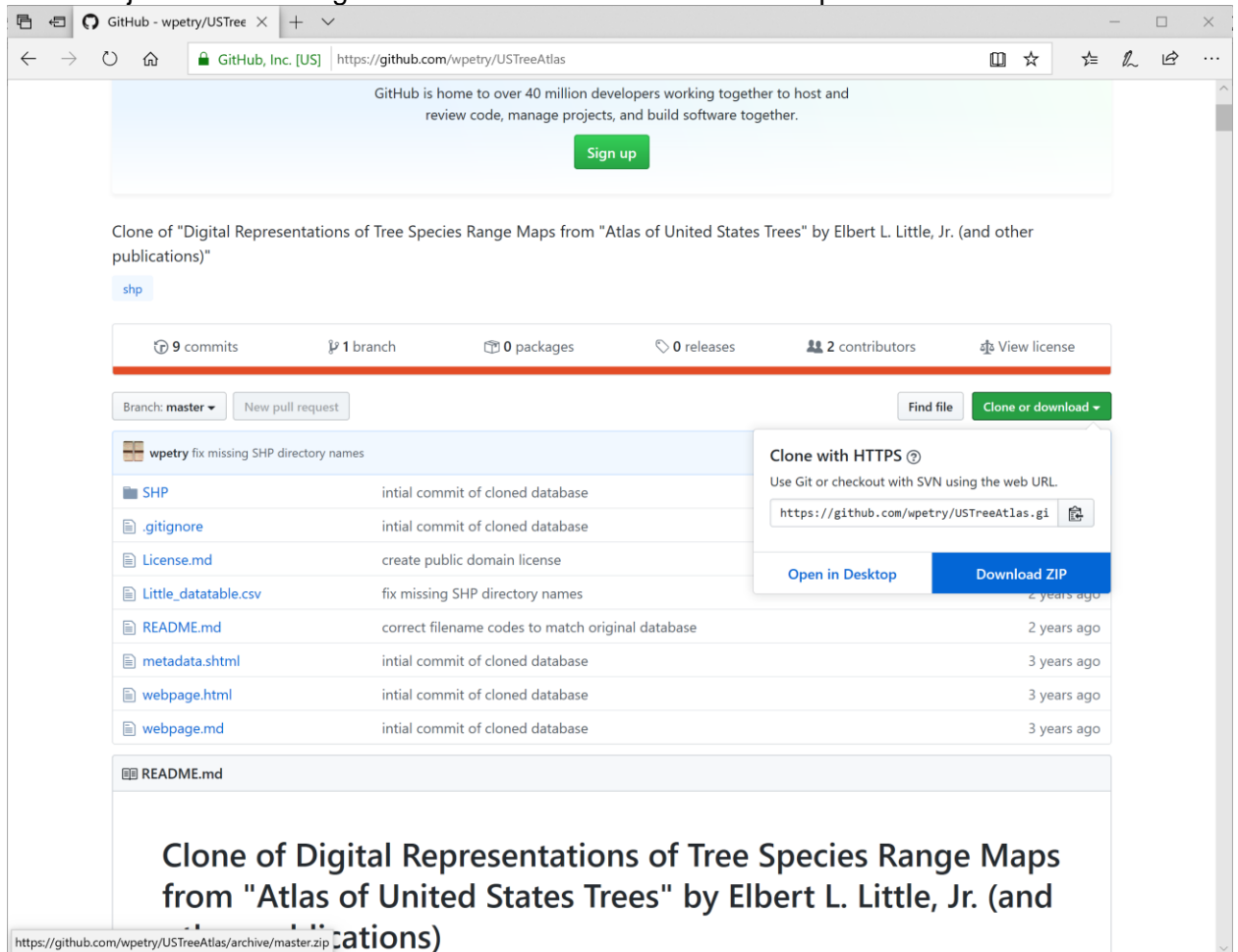
Using the union overlay tool, create a map that shows the species richness of *Fraxinus* in North America (any folder that starts with "Frax…" in the github repository. Note that you will need to use the Define Projection tool for each of the files you download. The information on the coordinate system is on the GitHub page (the metadata). Present a map of *Fraxinus* species richness in the same coordinate system as the TREES feature dataset in the OTTAWADATA.gdb. Describe the pattern of *Fraxinus* spp. richness in North America.

Student Name: _____ Student Number: _____

Laboratory #5: Remote Sensing and Raster GIS

Purpose: The purpose of this lab is to provide you with an introduction to a basic remote sensing classification workflow and to assess change detection in forested regions using raster GIS.

Learning Outcomes: By the end of this lab, you will be able to:
- Undertake spectral enhancement using band combinations and ratios.
- Train an image classifier.
- Explain a signature file.
- Undertake a supervised classification using high and medium resolution remotely sensed imagery.
- Clean and generalize a thematic raster layer.
- Assess a classified result for accuracy.

Procedure:
- Read and work through the examples.
- Answer the questions at the end of the lab.

Data/Materials:
- All data are provided in lab5.zip.

Assignment:
- Answer the questions (#1-8) in the spaces provided.
- The total marks available for this assignment = 17 marks for 15% total course grade.
- All lab assignments must be completed and handed in individually.

Format:
Students will hand in questions with completed answers.

Tables and figures should be produced to the standards of the Department of Geography, Environment and Geomatics, University of Ottawa.
For guidelines on proper formatting and presentation of tables and figures, please see the introductory section of this lab manual.

Student Name: _____ Student Number: _____

Forest loss is an ever-increasing problem in the world today (see
https://earthenginepartners.appspot.com/science-2013-global-forest). In the example questions
you will learn how to classify forest and non-forest in two satellite images from 2001 and 2016
respectively, and then determine how much forest has been lost or gained in the meantime in a
small region in the south-east of Ottawa. To achieve an estimate of forest loss, in this
laboratory, you will be undertaking post-classification change detection using a 2001 Landsat 7
ETM+ image at 30 m resolution, and a RapidEye image at 5 m resolution from 2016. A post-
classification change detection allows you to avoid many pre-processing steps necessary for
other change detection methods, and it also allows you to directly use the digital numbers within
the images without, for example, undertaking absolute radiometric corrections or corrections to
ground reflectance for the different images used.

This section presents examples of things you will need to know to address the questions you
will be handing in for marks. Your task is to review and repeat the examples in this section, and
the derived layers you produce in this section are used directly in the questions at the end of the
exercise. Other questions are relevant to help you address questions at the end of the exercise.

Q1: How do can I view different band combinations of a remotely sensed image?

Various band combinations can help you identify different features in remotely sensed images.

1. Add the image called "RE2016_07_20" from the IMAGERY.GDB in the lab5.zip file to
   ArcGIS.
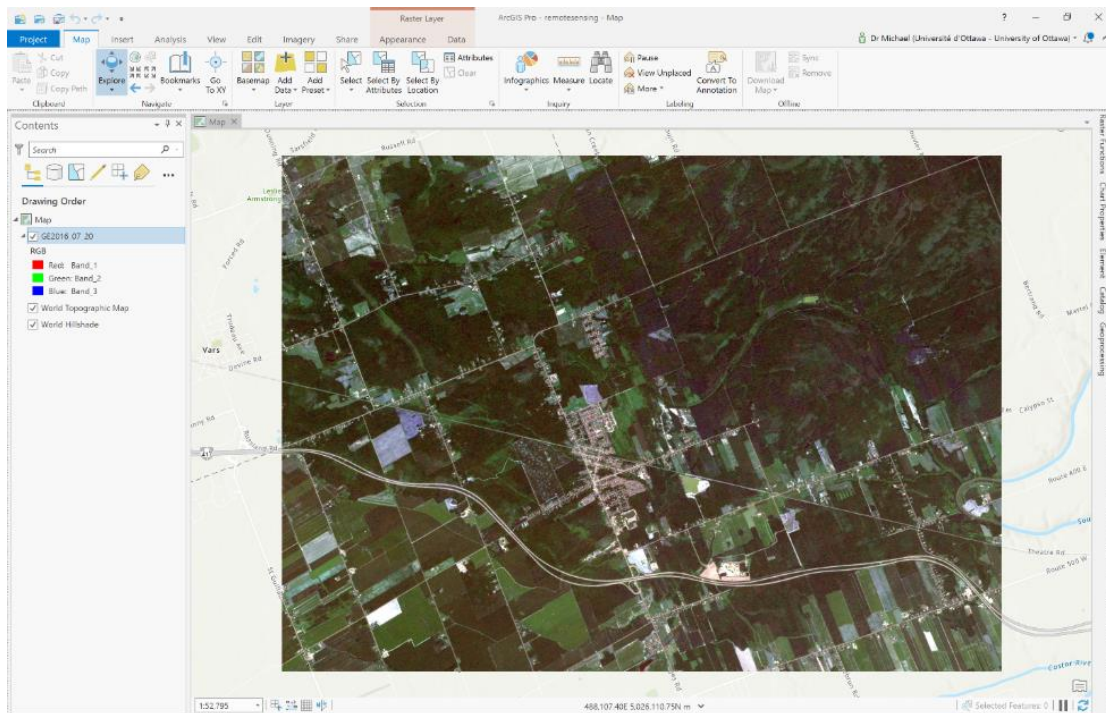
2. This is a section of a RapidEye multispectral image acquired on July 20th, 2016. This satellite constellation is used in agricultural and phenological monitoring and research because of the daily revisit time (off-nadir, 5.5 day at nadir) due to a 5-satellite constellation. This makes the sensor useful for monitoring crop stress or disease outbreaks in large regions. The RapidEye sensor is multispectral, with the following bands:

<div align="center">

440 – 510 nm (Blue)
520 – 590 nm (Green)
630 – 685 nm (Red)
690 – 730 nm (Red Edge)
760 – 850 nm (Near IR)

</div>

This image is not atmospherically corrected. In most cases, using this image for assessing crop stress or phenology would require atmospheric correction, as you can learn about in another course, but we will not do that here. Furthermore, the image you are using here is orthorectified, but only using a constant elevation, which is not super-accurate.

3. To view other band combinations of the image, right-click the "RE2016_07_20" layer in the contents pane and then select Symbology:
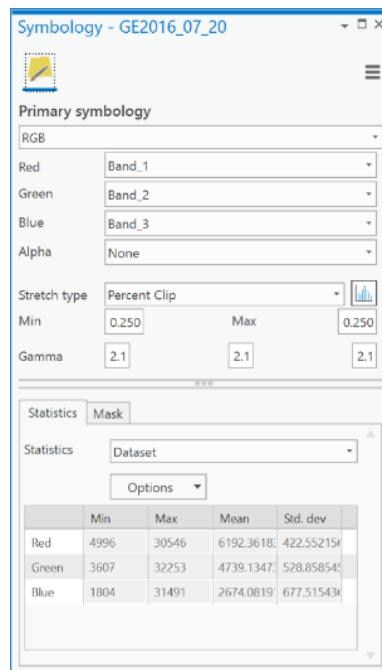
4. By default, ArcGIS will display the visible color band combination 3, 2, 1 for Red, Green and Blue. You can change this to False Color Infrared, for example, by choosing 4,1,2 (Band_4 as the Red band, Band_3 as the Green band, and Band_2 as the Blue band) in the Band column of the Symbology tab:
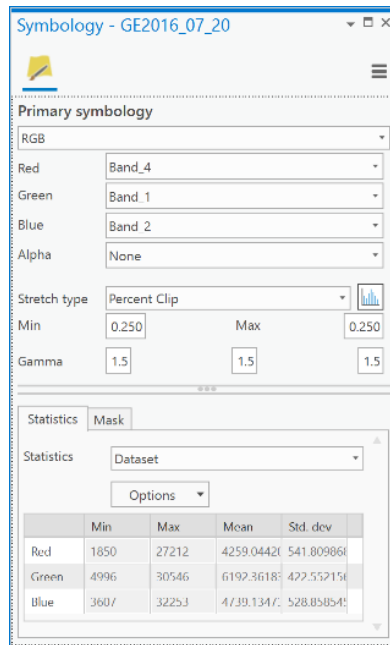
5. The resulting map is a "False color composite" and will show red where vegetation is present.
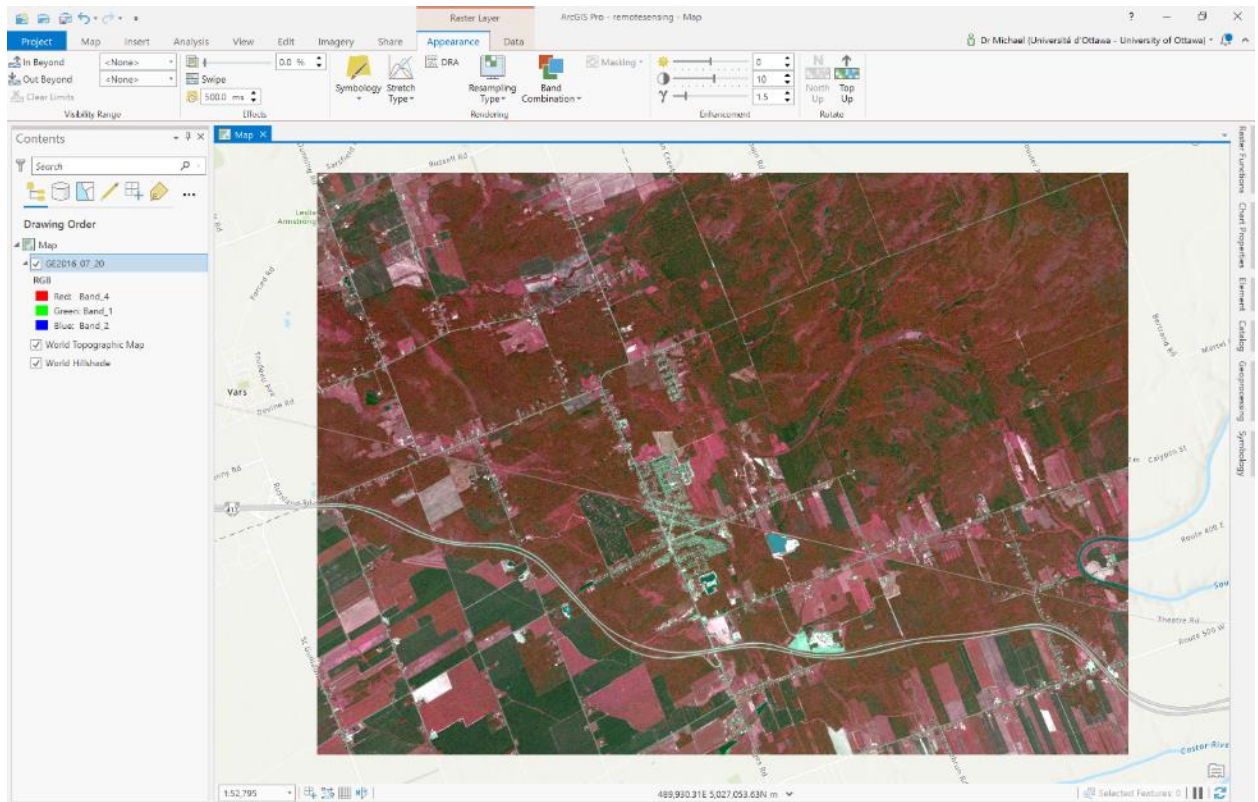
6. You can try other band combinations to create different false color composites, some of which may visually highlight different elements of the imaged landscape.

Q2: How do I classify a remotely sensed image?

In this lab, we will conduct what's called "supervised classification", in which the image analyst (you) provides the computer with examples of pixels containing different landscape features to be treated as information classes. An information class is a nominal category that you want to extract from a remotely sensed image. For example, if you want to extract forested areas from an image then you could create an information class called **Forest** and one called **NonForest**. You could then find examples of pixels that belong to the user-defined information class called **Forest** and another set of pixels for the class called **NonForest**. You could have many other information classes depending on the type of information you want to extract from an image. For example, information classes could include 'water', 'cropland', 'residential', 'commercial', 'dense forest', 'open forest', etc. Depending on the type of remotely sensed image and its spatial and spectral resolution there are many kinds of information that are possible to extract.
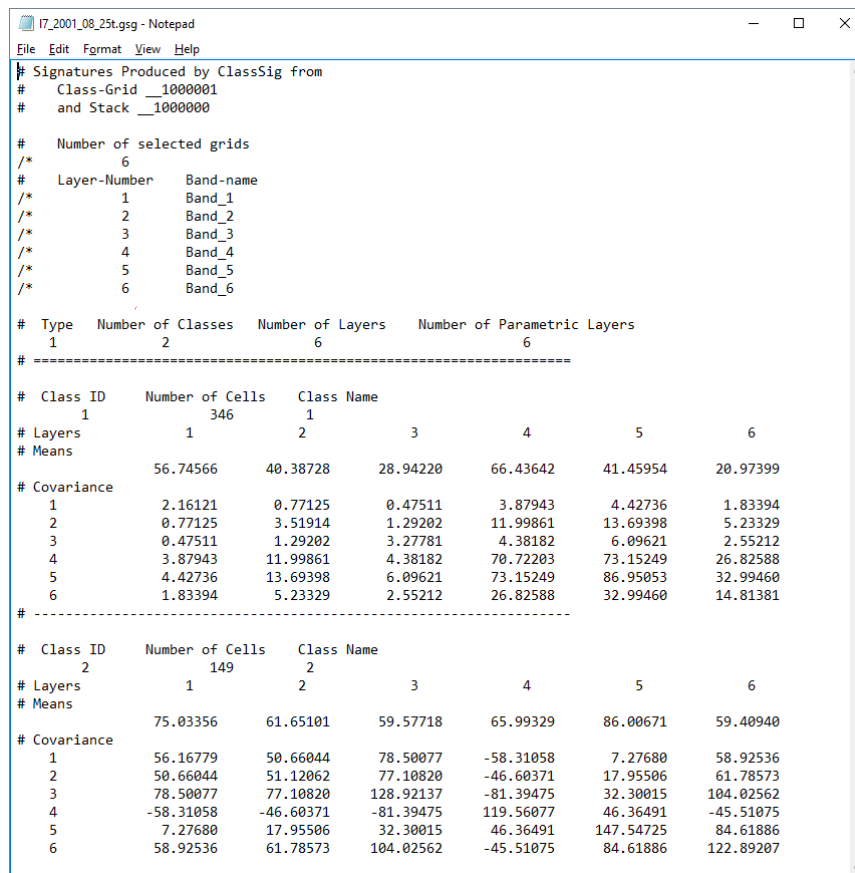
To conduct a supervised classification, you need two things:

1. Training data: The training data consists of pixels within the image that belong to different information classes that you want to extract from the image to create a thematic raster layer. For example, pixels that belong to **Forest** and pixels that belong to **NonForest** are two examples of information classes. Examine the figure below. The green polygon is a training polygon that encloses a set of pixels (across all bands) that the analyst has identified as belonging to the information class called **Forest** in the image. The absolute frequency (the count of pixels) of 'Forest' pixel values with different digital numbers (DN), within the green polygon, are shown as green histograms for each band. Likewise, the blue polygon contains **NonForest** pixels and is a training site containing non-forest pixels across all bands in the image. The absolute frequency of DNs within the **NonForest** information class for each band are shown as blue histograms. For the classification to be accurate, the training pixels for each information class should have separation at least one of the histograms, i.e., the histograms for each training class should have minimal overlap. This is discussed in detail later. In this case, separability is seen in Bands B1, B2, B3, but not B4.
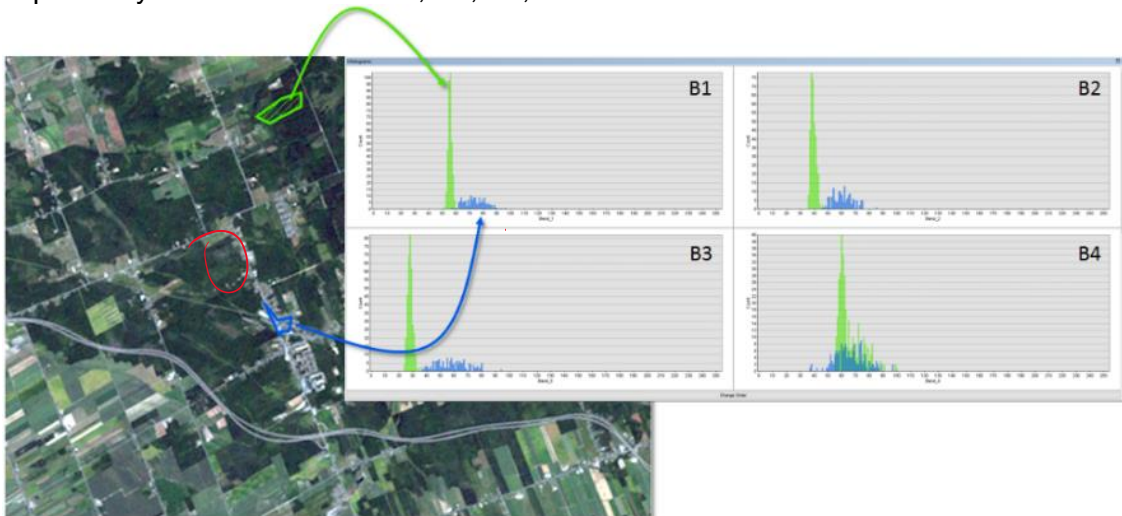
2. <u>A signature file</u>: A signature file has a ".gsg" file extension and contains a statistical summary of the different pixel values within each of the information classes specified in the training dataset:

The signature file above specifies the mean values of the histograms and their variability (covariance) within each image band. The signature file is a statistical summary of the histograms for each band, for each information class. For example, the file contains the mean and covariance of DNs for pixels belonging to the class **Forest**, and likewise for **NonForest**, for each band. This information allows a machine classifier to place each pixel in the image into its respective class.

Using the signature file created from the training histograms, a supervised classification is one that assigns all pixels in the image to the information classes specified in the signature file. The classifier will look at your signature file and then assign each pixel in your image to one of the information categories that you defined in the training dataset. The outcome of the classifier is a thematic raster layer that could then be used in map algebra or other raster modelling processes:
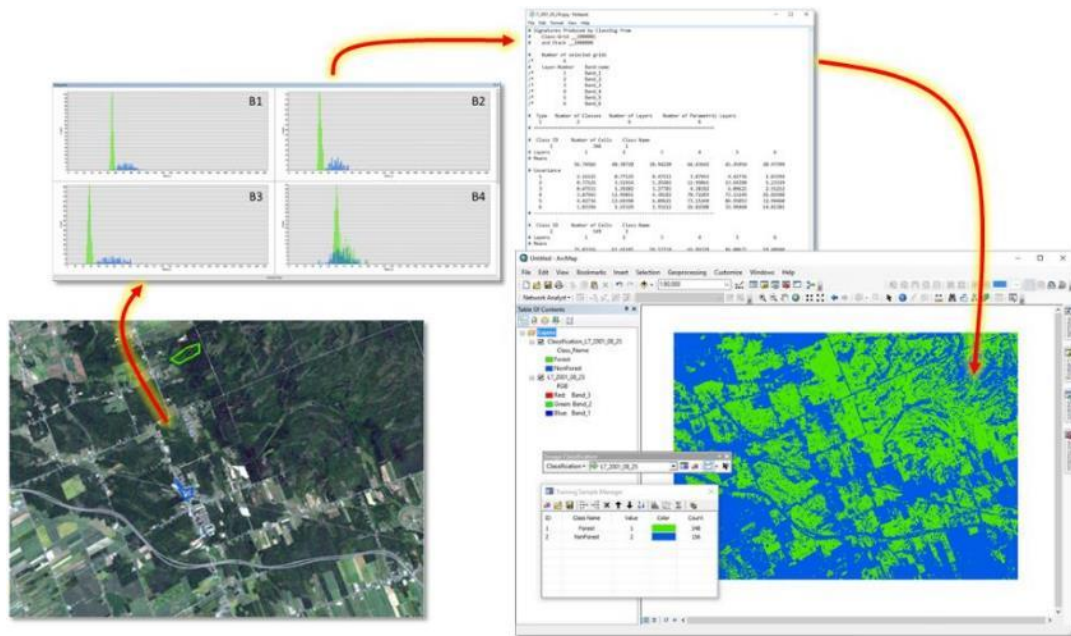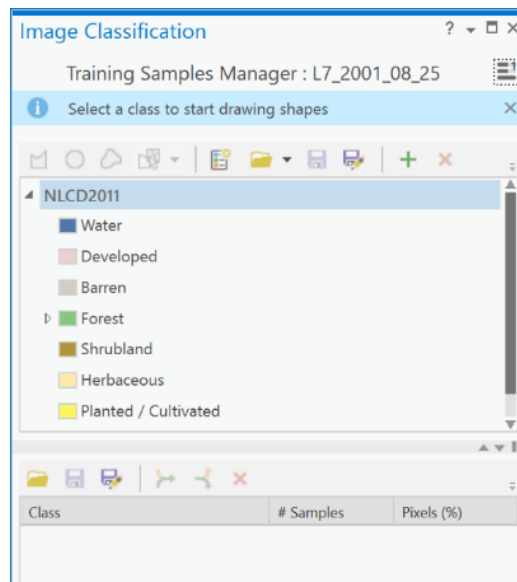
Classifiers are AI/machine learning algorithms that learn from your training data (in the signature file) how to classify any pixel in the image, including those pixels that are not part of your training set. There are many types of classifiers that have various benefits and drawbacks; these go by names such as Maximum Likelihood, Neural Networks, Support Vector Machines, K-means, Random Forest, and Iso-Cluster. Regardless of which classifier you work with, the quality of the training data is paramount and has the greatest effect on the accuracy of the resulting map layer.

Ok, let's get started. We'll first work with a Landsat 7 image from 2001, August 25th called "L7_2001_08_25" in the IMAGERY.GDB:

3. Add the "L7_2001_08_25" to ArcGIS and ensure that you choose it in the Contents pane:
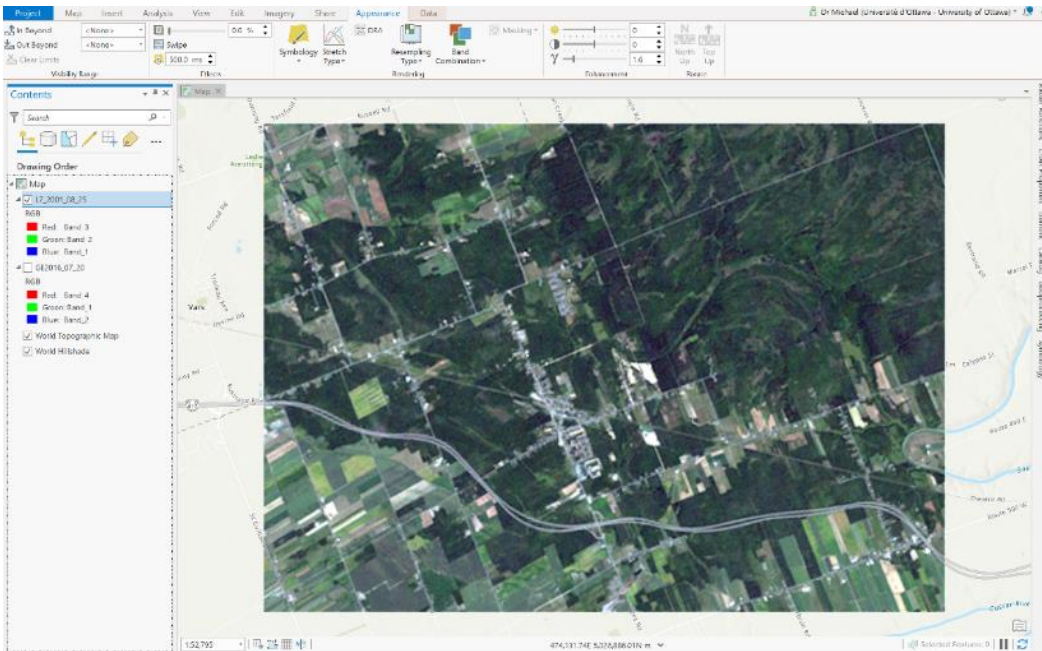
4. Click on the Imagery tab and then click on the button 'Classification Tools' and choose the Training Samples Manager. There may be a default set of information classes, don't worry, we will deal with that in the next steps.


5. Below is the description of the *Training Samples Manager*.

| Tool	Function
	Create a training sample by drawing a polygon around pixels or objects in the raster.
	Create a training sample by drawing a circle around pixels or objects in the raster.
	Create a training sample by drawing a freehand shape around pixels or objects in the raster.
	Create a training sample by selecting a segment from a segmented layer. This option is only available if there is a segmented layer in the **Contents** pane. Activate the **Segment Picker** by highlighting the segmented layer in the **Contents** pane, then select the layer from the **Segment Picker** drop-down list.
	Create a new classification schema. Right-click the **New Schema** title and click **Add New Class** to begin creating class categories.
	Select a classification schema option. • Browse to an existing schema. • Generate a new schema from an existing training sample feature class. • Generate a new schema from an existing classified raster. • Use the default 2011 National Land Cover Database schema.
	Save changes to the schema.
	Save a new copy of the schema.
	Add a class category to the schema. Select the name of the schema first to create a new parent class at the highest level. Select the name of an existing class to create a subclass.
	Remove the selected class or subclass category from the schema.

6. Using the Training Samples Manager, you will open a pre-existing training set of polygons within a feature class called "L7_2001_08_25_Training" within the IMAGERY.GDB geodatabase. First, click on the 'Classification schema' button and choose Browse to an existing schema:
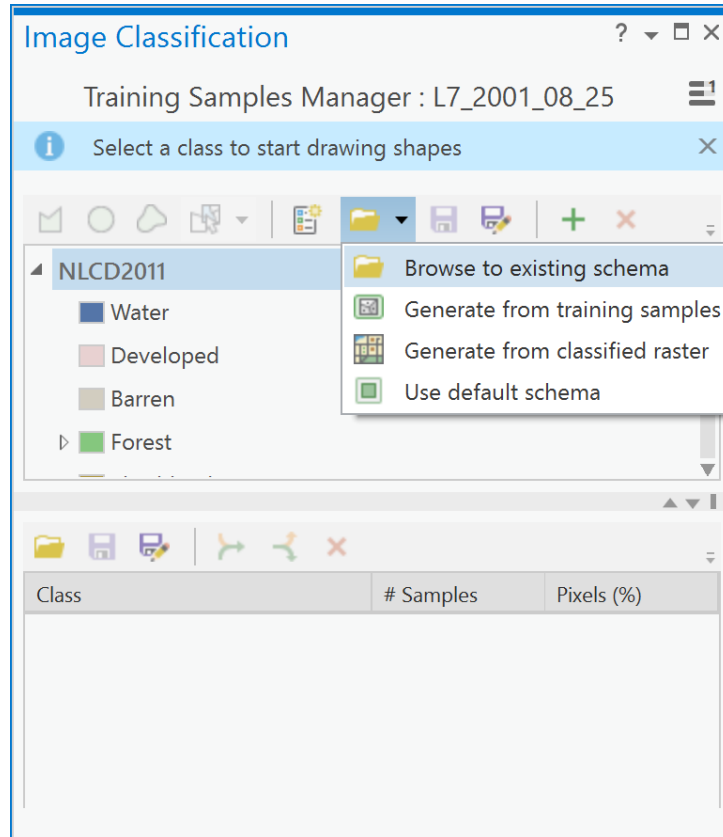
7. Navigate to where you unzipped the lab5.zip file, choose the "Ottawa.ecs" file, and click OK:
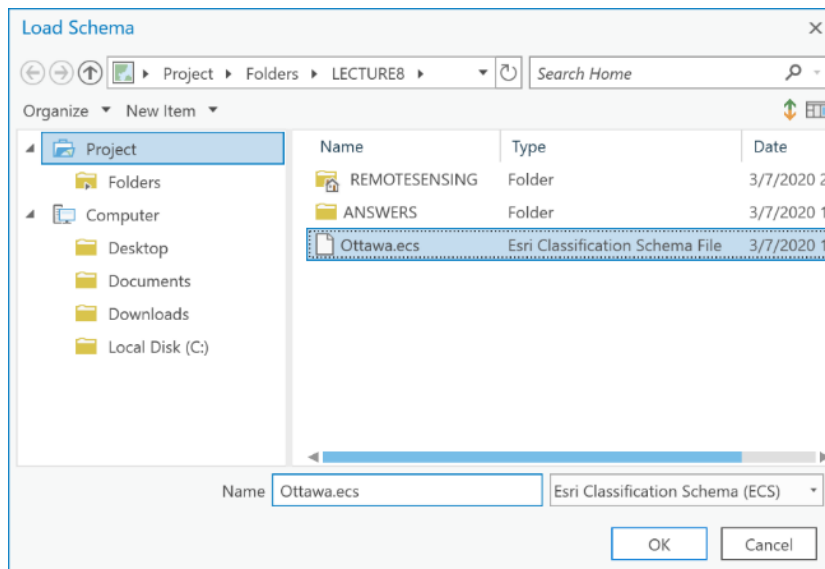
8. Now, click on the 'Load training samples' button in the lower half-pane and navigate to the and choose the "L7_2001_08_25_Training" feature class from IMAGERY.GBD. Your Training Samples Manager pane should now look like this:
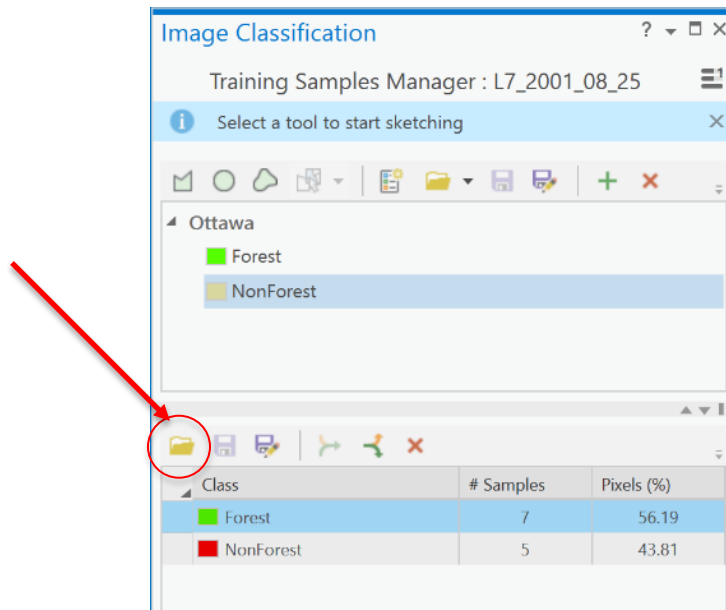

You can see that the **Forest** class corresponds to the color green and the training sites are shown as green polygons in the map. Likewise, for the **NonForest** information class which has polygons shown as red. What is important here is that the two information classes, **Forest** and **NonForest** have approximately the same percentage of pixels in each class. Right now, the **Forest** class has 56% and the **NonForest** has 44%. That's not too bad, but we will still add a new polygon to the **NonForest** class to increase its percentage.

9. To add another example of **NonForest** to the training samples, zoom into the bottom left quarter of the map:
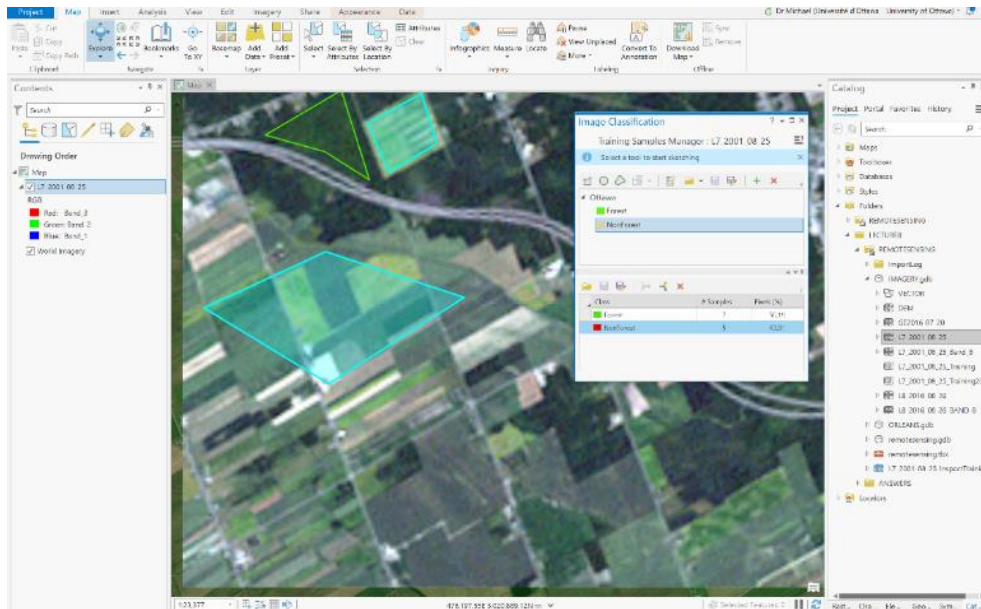
10. Next, in the Training Samples Manager, in the upper half-pane, under the schema called Ottawa, click on the **NonForest** information class. You will see that the polygon tool in the toolbar becomes active (this is the left-most tool; *newer versions of ArcGIS may have a rectangle button first, making the Polygon button second. Make sure to select the proper tool*).
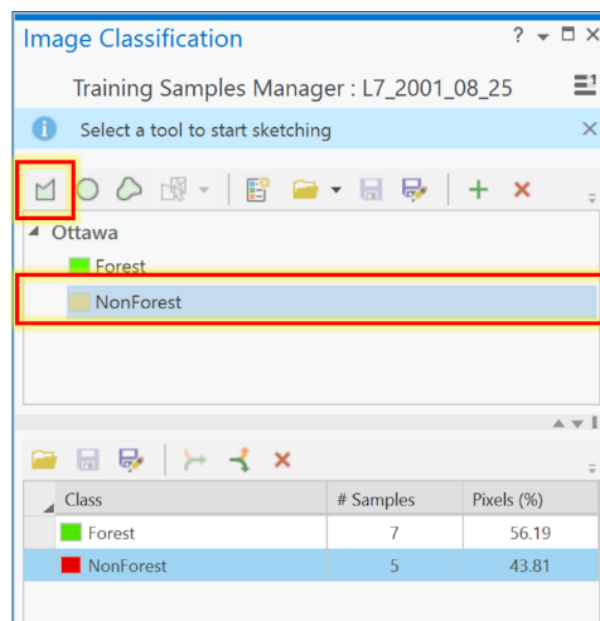

11. Now, click on the polygon tool and draw a polygon in any non-forested area on the satellite image. **Do not choose any other drawing tool than the Polygon tool to draw your polygon**:
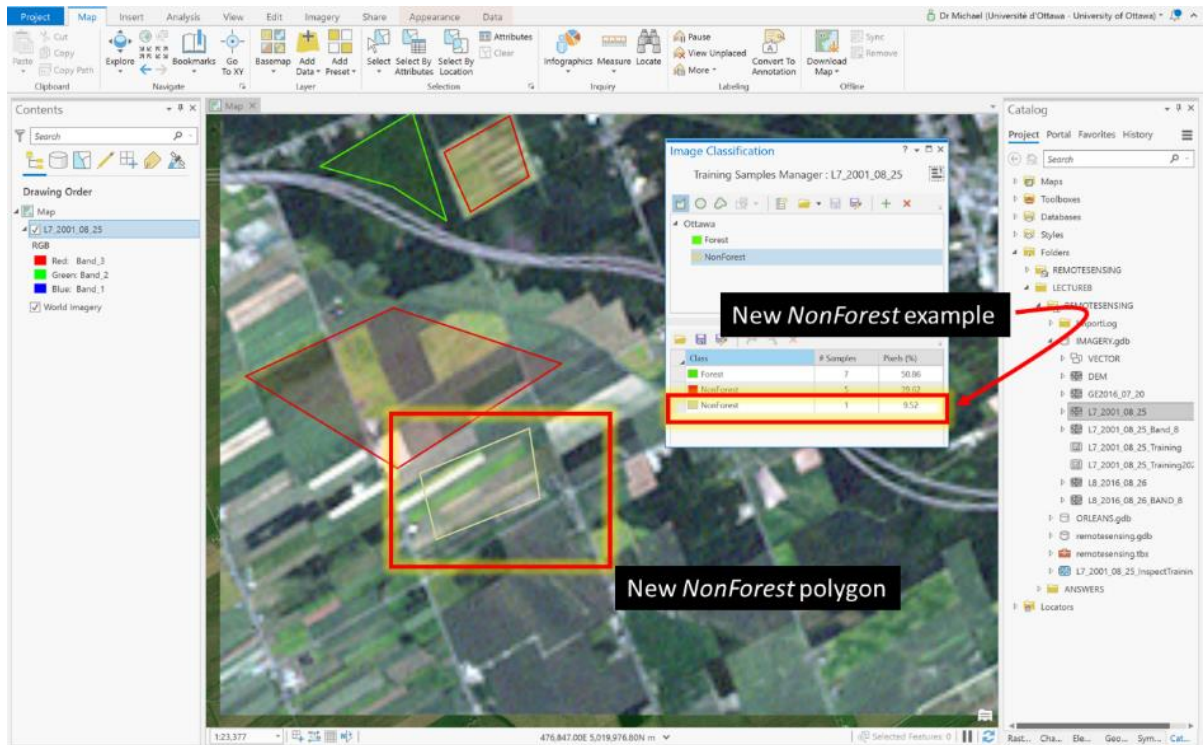
12. Examine the **Training Samples Manager** pane, the lower half-pane, and you will now see the original **NonForest** example polygons in red and the new **NonForest** example that you just made. Hold down your Ctrl key and click on both **NonForest** examples and click on the 'Collapse' button to merge the two **NonForest** examples into a single set:
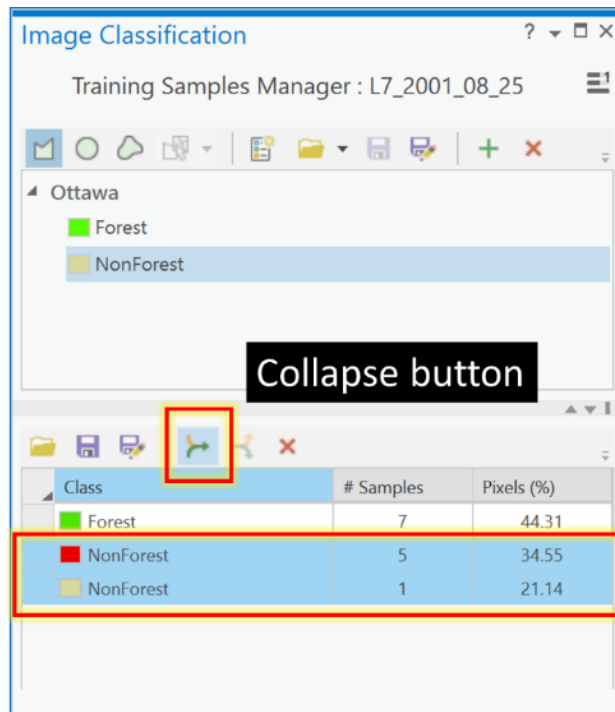
13. Now your Training Samples Manager pane will look like this:
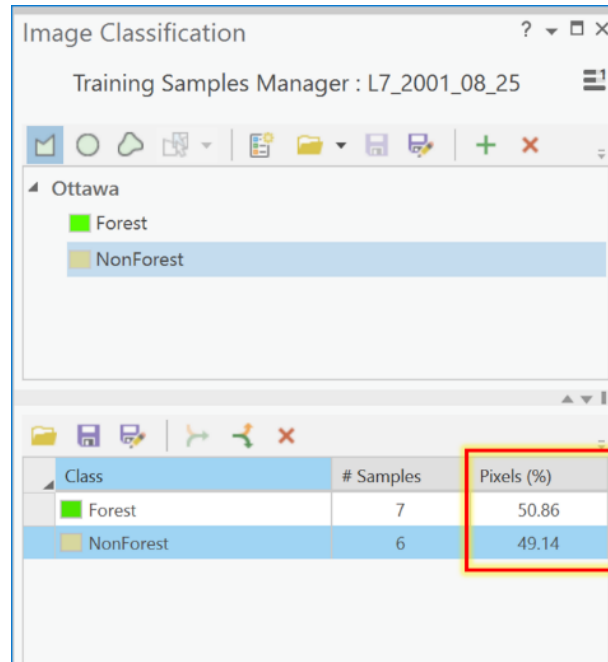
14. Check that the two information classes now have approximately 50% pixels each. If the polygon you added was too large then the NonForest class will now have much more than 50% of the pixels, in which case you did not succeed in balancing the size of the training samples from the two classes. If that happened you can either add a polygon to the Forest class (and so on, until you get it right), or you can do the following:

   - Select the NonForest class in the bottom half of the Training Samples Manager, and click the "Expand" button next to the "Collapse" button.
   - You will now see several rows labeled NonForest, each corresponding to an individual polygon.
   - Delete the last one – the one you just created.
   - Add a new (smaller) one.
   - Select all the NonForest polygons, and collapse them.
   - Check that you ended up more or less at a 50/50 distribution between the two classes.

15. Examine the **Training Samples Manager** pane carefully, it provides a lot of important information about the training samples for each information class. One thing not shown is that **Forest** (the first-listed class) is given a pixel value of 1 and **NonForest** (the next-listed class) is given a pixel value of 2 (and so on, if you had more classes). These values will become the cell values for any classification of the image. The 'Pixels %' column tells you how many pixels are in each information training class. In the above case, 50.86% of the training pixels are in the **Forest** information class and 49.14% are in the **NonForest** training set. So here we have achieved an almost equal proportion and that is acceptable and better than what we started with.

16. Next, we need to save our new training polygons to a feature class. To do so, click on the 'Save' button in the lower-pane of the **Training Samples Manager** and navigate to the IMAGERY.GDB and save the new training file with the name "L7_2001_08_25_Training2023". We don't need to save the 'Ottawa' schema in the top half-pane because it hasn't changed. Had we added a new information class then we would also have had to save the new schema.

17. This next step is fundamental because we will explore how 'pure' our information classes are based on the pixels that we chose for each class. We want to find out which bands have the most *separability* between **Forest** and **NonForest**. *Spectral separability* refers to the spectral distance between the pixel values that compose each information class in a training samples dataset. Ultimately, we want to know which bands of our remotely sensed image would be best for classifying our image into **Forest** and **NonForest**. The bands that are best will clearly distinguish between our two information classes and there will be little to no overlap between the pixel values for **Forest** and **NonForest** in those bands.

To understand spectral separability, look at the figure below to see histograms for four bands and two information classes differentiated as red and green colors. Each histogram shows the frequency of digital numbers for each class, red class or green class, on the y-axis and the range of digital numbers associated with the red or green class on the x-axis. For example, in Band_1 we see that the green class ranges from about 52-60 and the red class from around 55-150. So, while the two histograms are clearly different, there is also some overlap around values of 55-60, shown in dark red. However, there is no overlap in the peaks of the red and green histograms which is a good thing. In general, a 'good' band for classification purposes is one that would have no overlap at all between the red and green histograms. If there is no overlap between information classes in a band then the red and green classes are distinct and completely separated. As such, such a band would be good for classifying the image into red and green information classes. Here, none of the image bands show complete separability. The band with the smallest overlap is Band_2. Why? Because there is little overlap between the red and green histograms. Conversely, the worst band is Band_4, where there is considerable overlap and so Band_4 would not help in classifying the corresponding image into red and green bands.
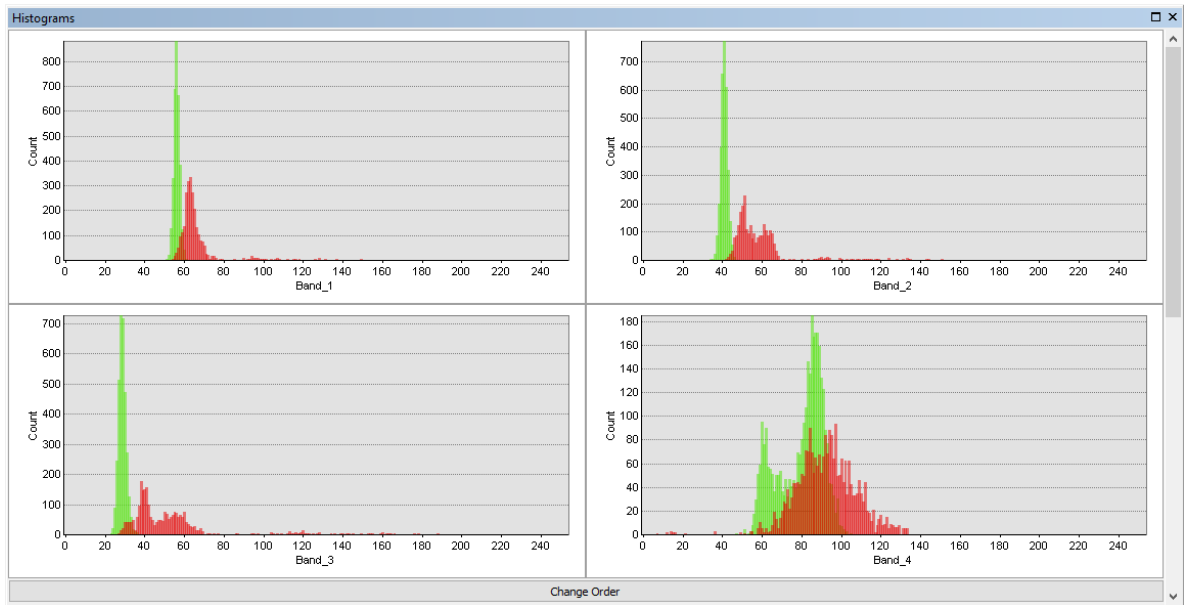
We don't need to generate histograms for each band to examine separability. The most common way to examine spectral separability among the different bands is to generate a plot of the statistics of each histogram for each band in a spectral profile.

18. To create a spectral plot, first, ensure you have added your "L7_2001_08_25_Training2020" layer to the contents pane and symbolize it using unique values and the ClassName field, and make **Forest** as green and **NonForest** red.
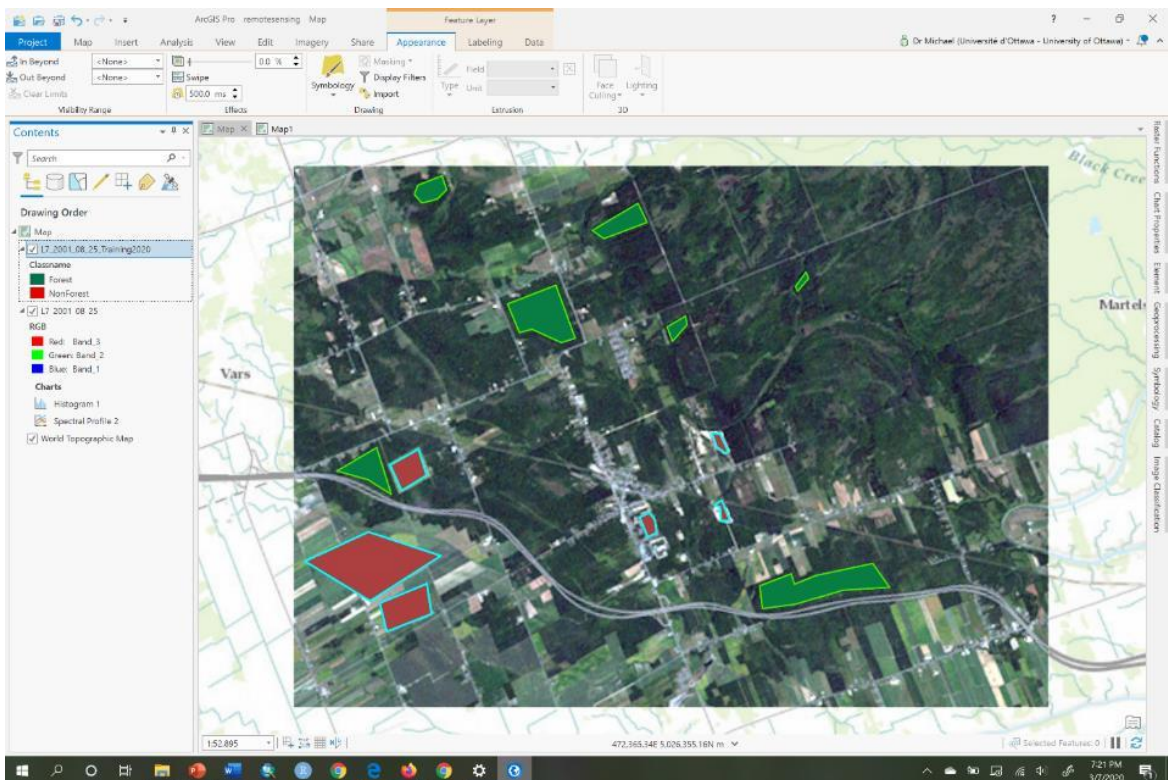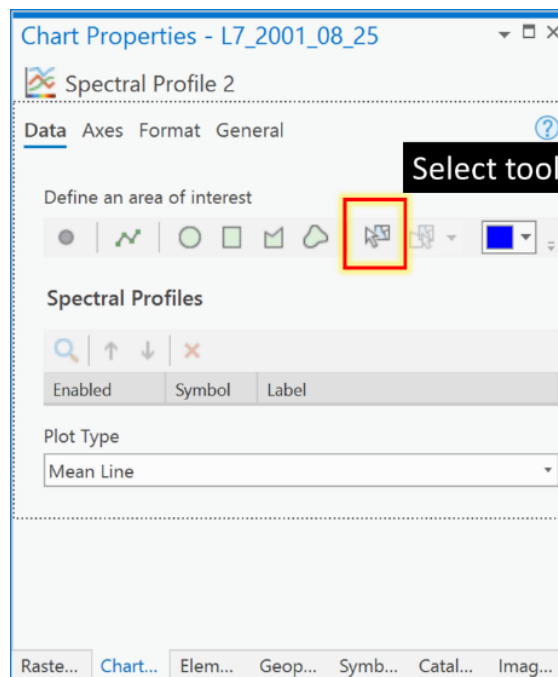
19. Next, choose the "L7_2001_08_25" layer in the contents pane. Then, click on the Data tab. Now click on the Create Chart and choose 'Spectral Profile':
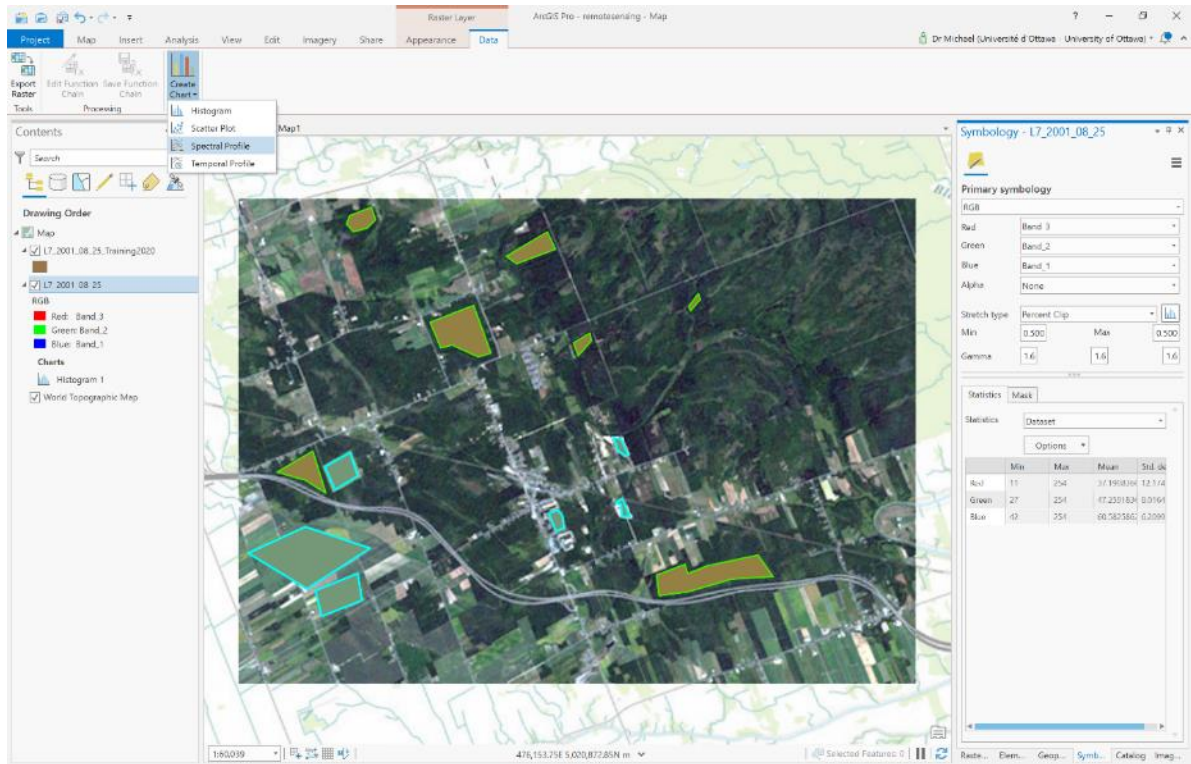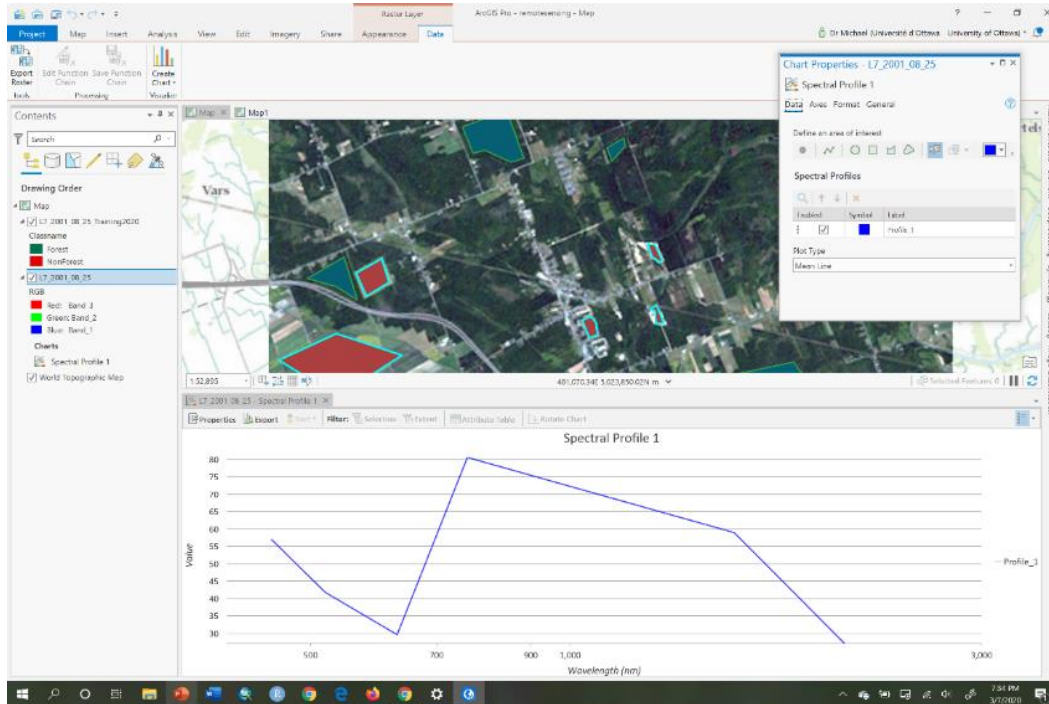

20. Next, in the Chart Properties pane, click on the Feature selector button:

21. Now, click on any of the **Forest** polygons in the map pane. You will see a chart created with a single line called "Profile_1" that represents the average or mean of the DNs across all bands for that information class called **Forest**.
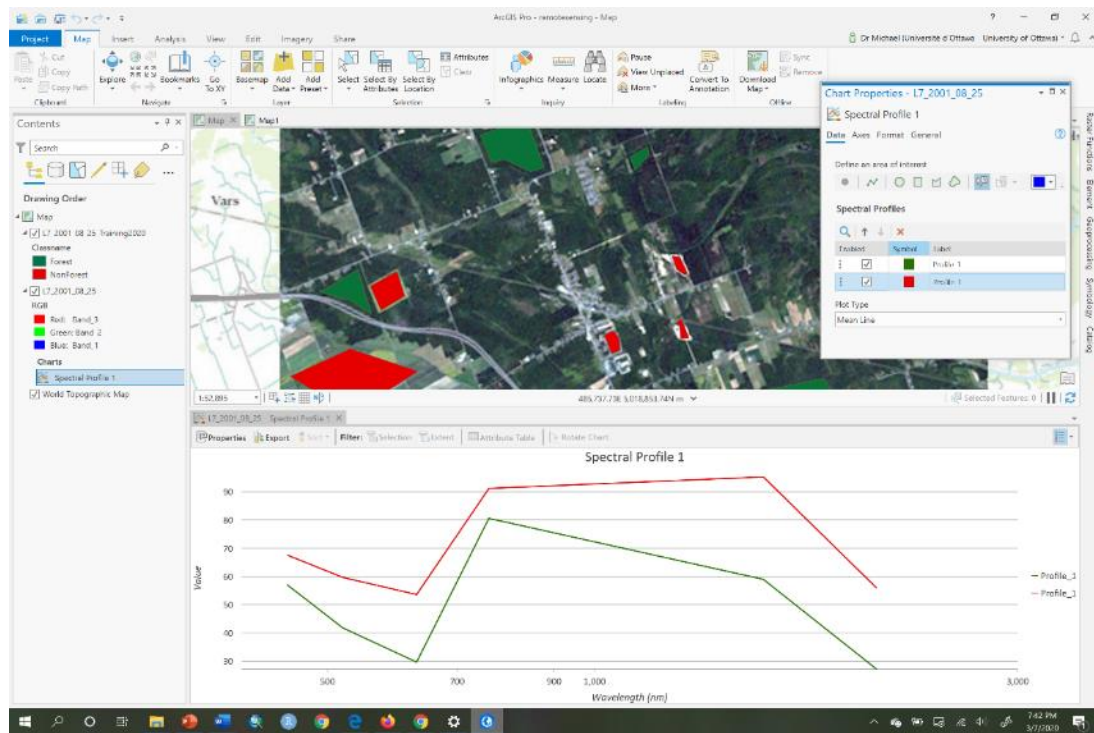

22. Now, in the Chart Properties pane, change the label for this spectral mean line from "Profile_1" to **Forest** (if it works!) and change the color to green and you will see your line become green:
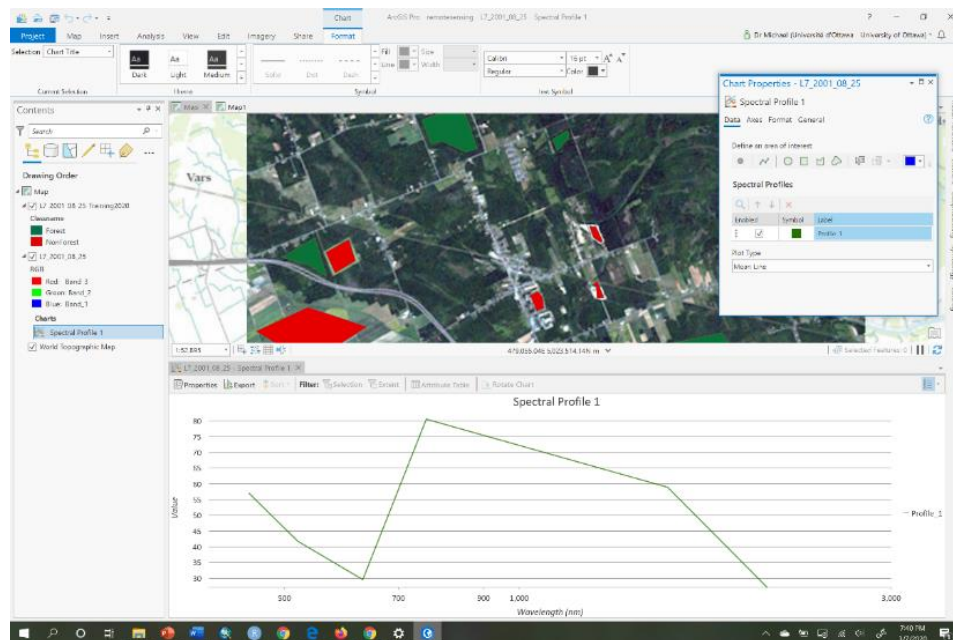
23. Now repeat 21-22 but for the **NonForest** training polygons and make the new line red to correspond to **NonForest** and rename the label if you can in your version of ArcGIS Pro:


24. Now examine the Spectral Profile 1 graph:

This graph shows wavelength on the x-axis rather than band labels because the metadata of the image contains information on the wavelength of each band. You can infer the bands by seeing where the lines have their 'kinks'. For example, Band_1 is at the far left, Band_2 is just after 500 nm and Band_3 is just before 700 nm, etc. What we see is that in each band there is a difference between the mean values for each information class. Perhaps the least separability might be in Band 4, just after 700 nm, since the lines are closest there. We can check with a more detailed statistical plot called a box-and-whisker diagram.

25. To create the box-and-whisker plot of the spectral separability for each band we only need to change the plot type in the Chart Properties pane. In the pane change the Plot Type from 'mean' to 'boxes':
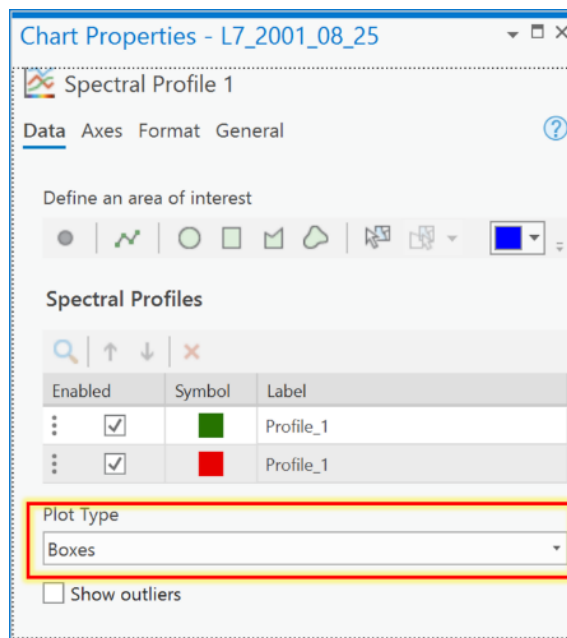

26. Now examine the box and whisker plot.

In this plot, there is actually a box-and-whisker plot for each band, each one showing the **Forest** (green) and **NonForest** (red) DNs. The top and bottom of the solid-colored boxes are the upper and lower quartiles of the DN histogram for the information class. The whiskers are the two lines outside the box that extend to the minimum and maximum DNs for the information class. So basically, these are a visual summary of the histograms for each information class, in which we can now examine which bands have the best separability. There is much more information here than in the Spectral Profile graph. What really care about seeing here is that the boxes don't overlap, and they don't for all bands except for Band_4, which is the one that was also closest together in the Spectral Profile plot. So, we can consider excluding Band_4 from any classification of **Forest** and **NonForest**, but we should probably keep all the others.

27. Create a signature file from the training sites that we can use in classification. To create the file, go to the Analysis tab, click on Tools, and search for the Create Signatures tool from the Spatial Analyst toolbox. Undock the Create Signatures tool so that you can see that and your Catalog pane at the same time. Then, in the Catalog pane, navigate to IMAGERY.GDB and expand the "L7_2001_08_25" image so that you can see all the individual bands (Band_1, Band_2,…,Band_6). Drag all bands **except Band_4** (since we know it has bad separability) into the 'Input raster bands', under 'Input raster or feature sample data' choose the "L7_2001_08_25_Training2020" feature class, under 'Sample field' choose Classname, and save the file to the folder where you have your data, and call it "l7_2001_08_25.gsg". Click Run.
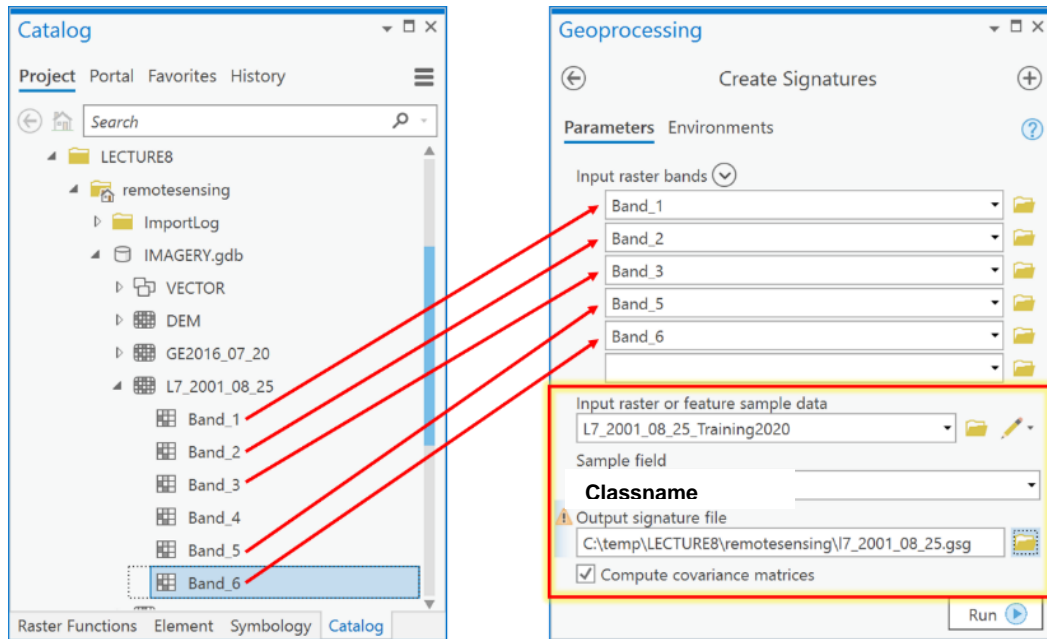
28. Now, use the signature file to classify the image into the two information classes of **Forest** and **NonForest**. To do so, search for the Spatial Analyst tool "Maximum Likelihood Classification". Undock the Maximum Likelihood Classification tool so that you can see it with your Catalog pane at the same time. Then, in the Catalog pane, navigate to IMAGERY.GDB and expand the "L7_2001_08_25" image so that you can see all the individual bands (Band_1, Band_2, …, Band_6). Drag all bands **except** Band_4 into the 'Input raster bands', under 'Input signature file', choose the signature file created in the previous step called "l7_2001_08_25.gsg", under 'Output classified raster' call it "MLCLASS_L7_21", and save it in IMAGERY.GDB. Leave all other values as defaults and click Run.
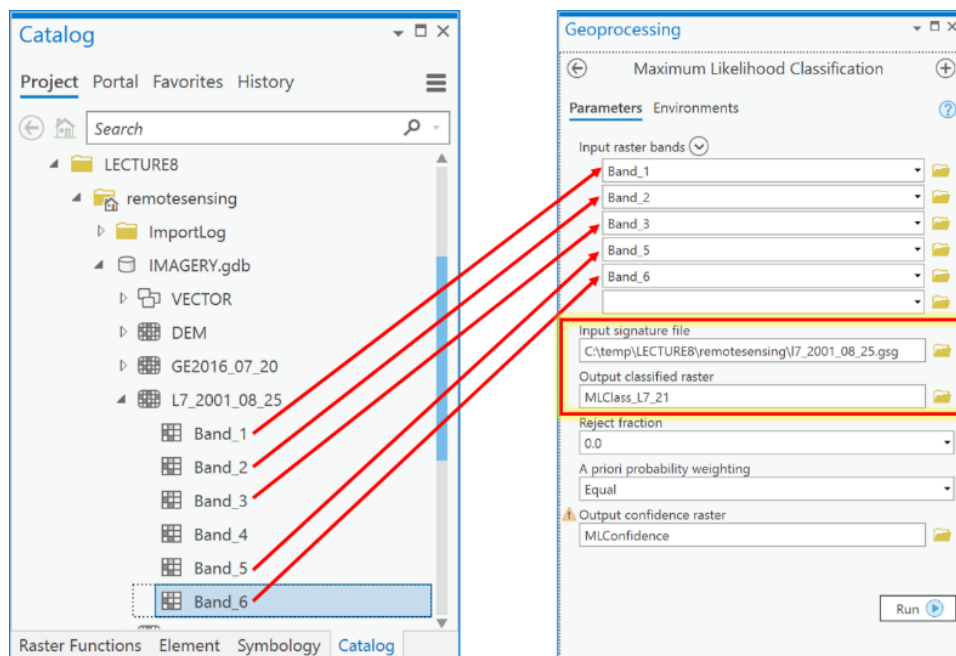
29. Examine the resulting classified raster (note that you may want to change value 1 to green and value 2 to red). We now have converted a remotely sensed image to a thematic raster layer.
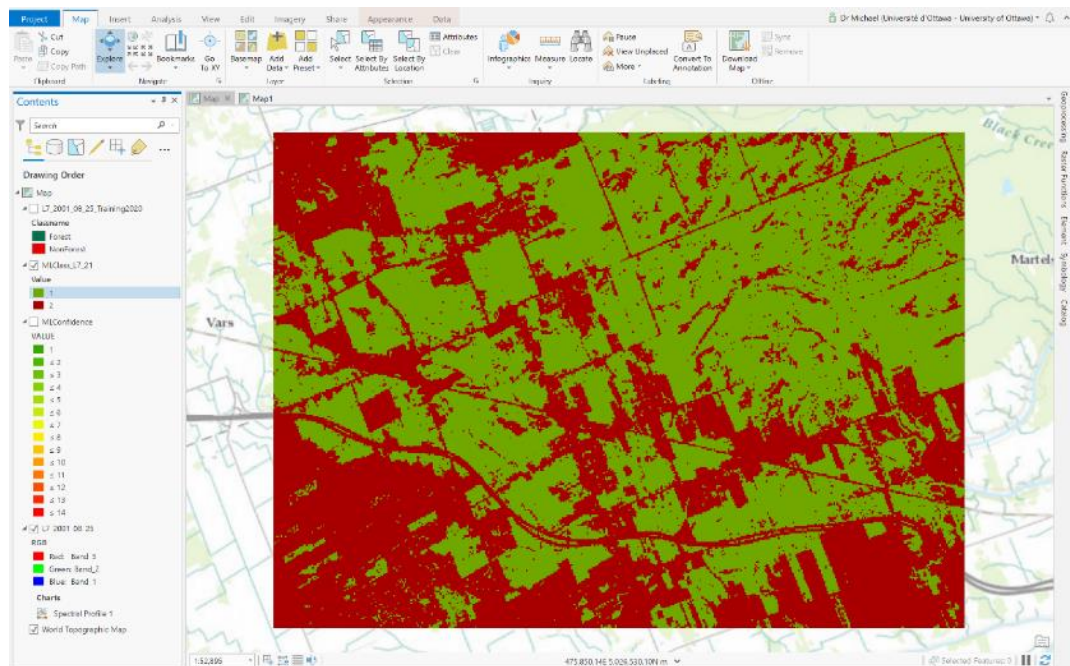
**Q3: How do I create a training dataset corresponding to my information classes?**

Training data consists of pixels that belong to different information classes. For example, all pixels that belong to the class called **Forest**, and another set of pixels for the class called **NonForest**.

1. If it's not there already, add the "RE2016_07_20" image to ArcGIS from IMAGERY.GDB.
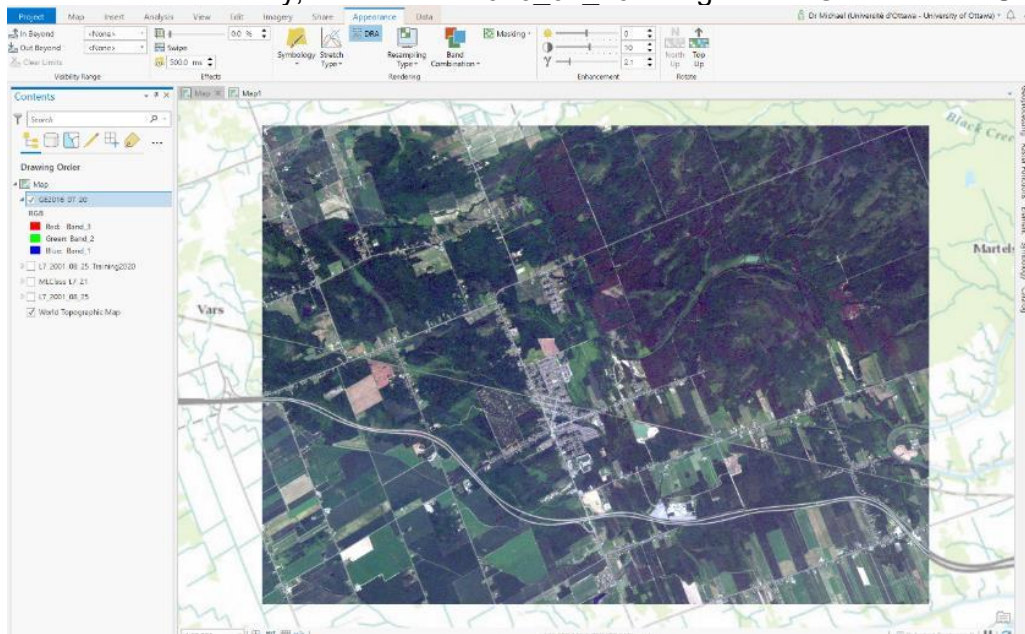
2. Click on the Imagery tab and then click on the button 'Classification Tools' and choose the Training Samples Manager.

3. Using the Training Samples Manager, open the Ottawa.ecs schema, as you also did earlier.

4. Now we will create a new set of training polygons for the **Forest** and **NonForest** information classes, to be used with the RapidEye image. Note that we cannot simply reuse the polygons we used for the Landsat image, because areas within those polygons may have changed from **Forest** to **NonForest**, or vice versa, in the period between 2001 (when the Landsat image was acquired) and 2016 (when the RapidEye image was acquired). To do this, in the Training Samples Manager, in the upper half-pane, under the schema called Ottawa, click on the **Forest** information class. You will see that the polygon tool in the toolbar becomes active.
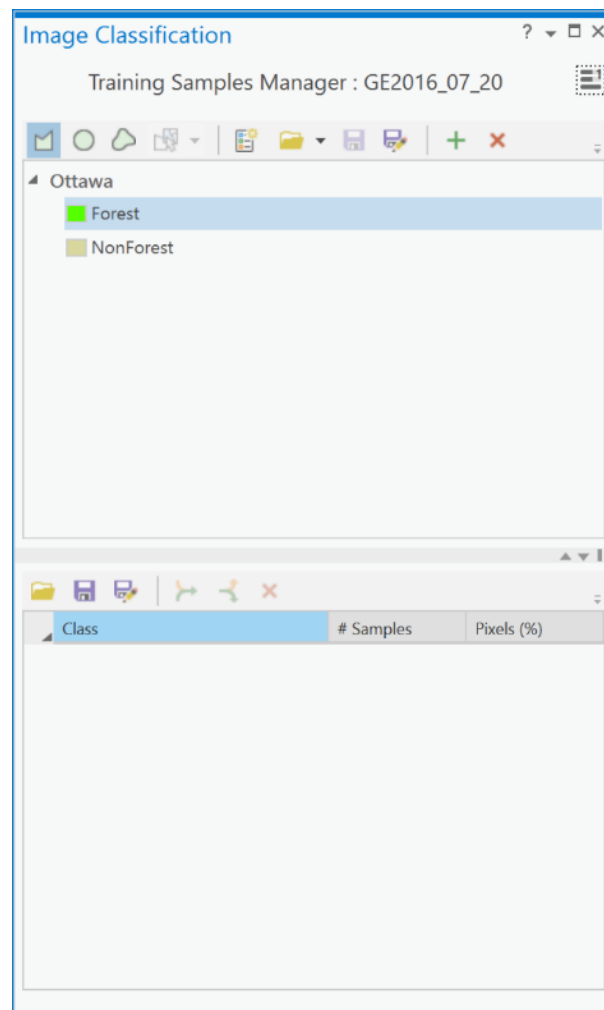

5. Now, click on the polygon tool and draw a polygon in any forested area on the satellite image.
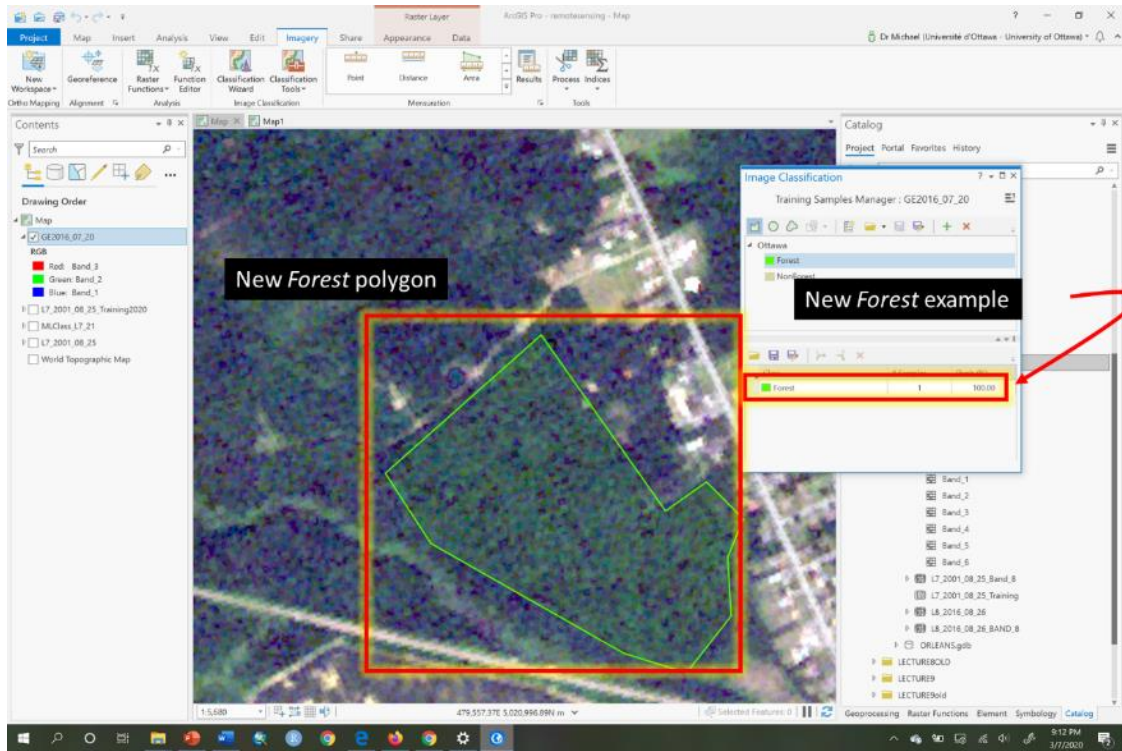
6. Now repeat the previous step nine more times in different forested areas. You will end up with something that looks like this:
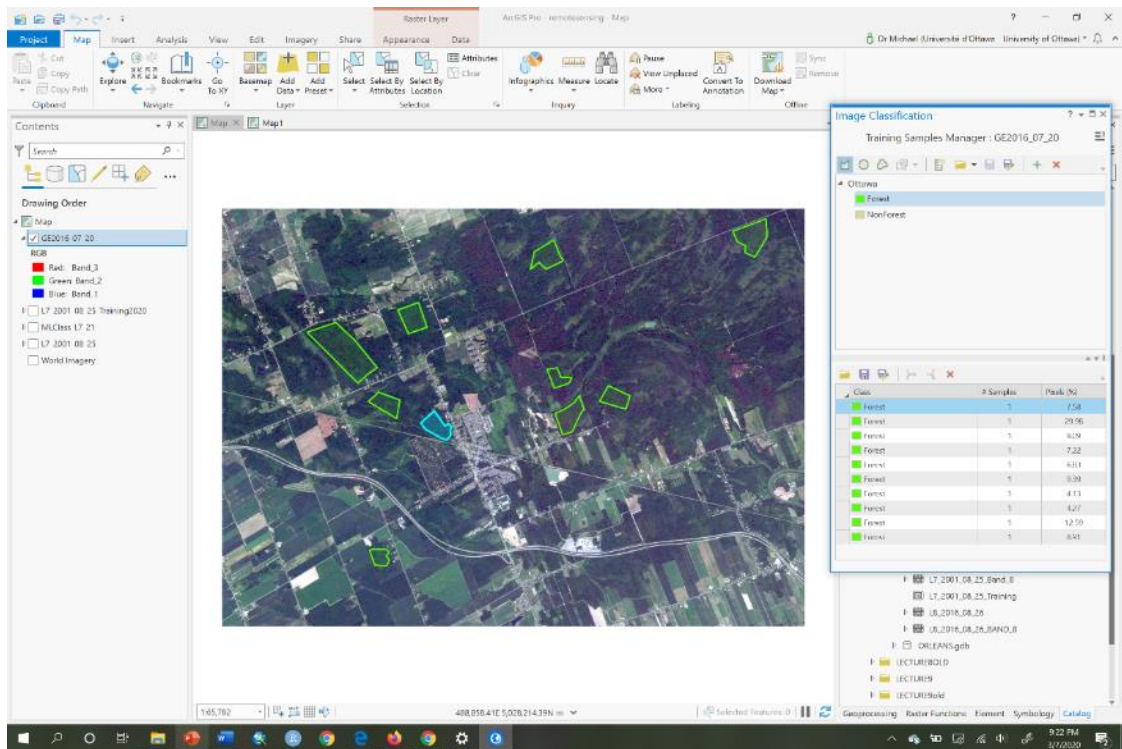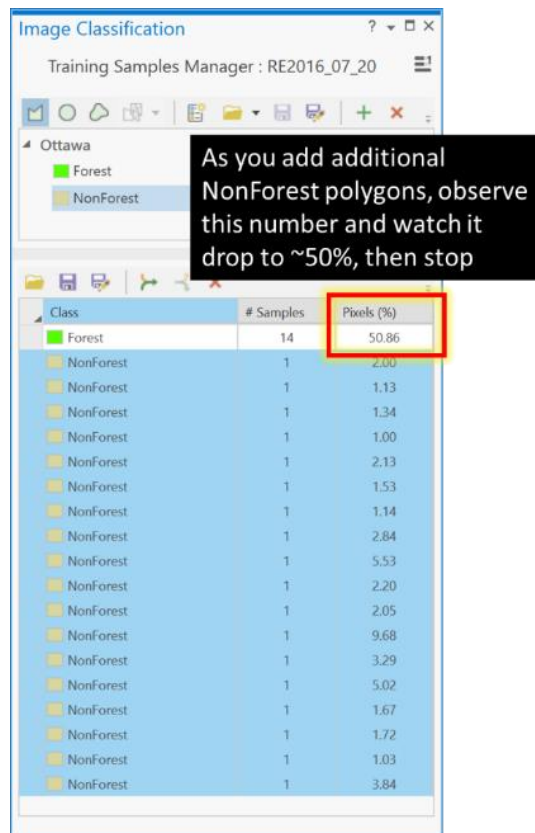
7. Examine the Training Samples Manager pane, the lower half-pane, and you will now see the 10+ **Forest** polygons listed. Hold down your Shift key and click on the first and last **Forest** polygon example to select all the **Forest** examples. Then click the 'Collapse' button to merge the **Forest** examples into a single group:
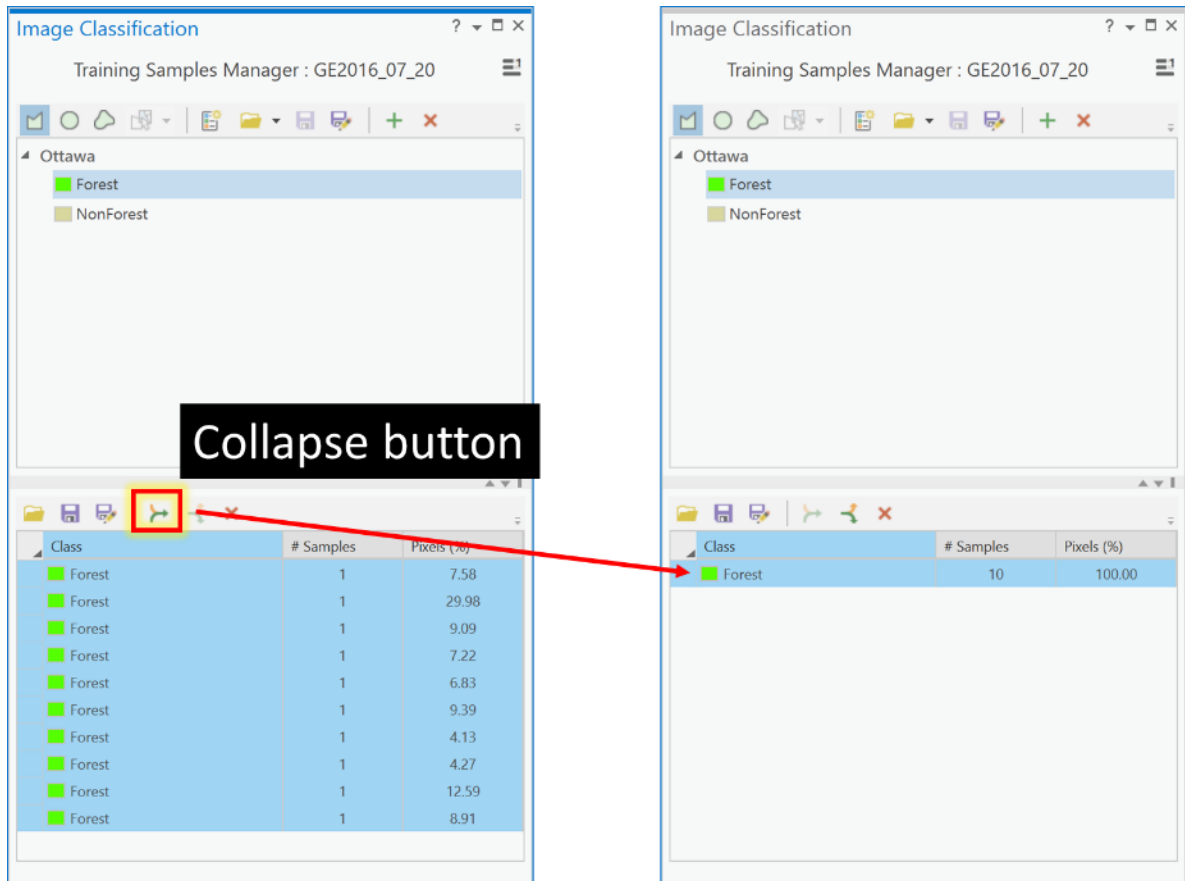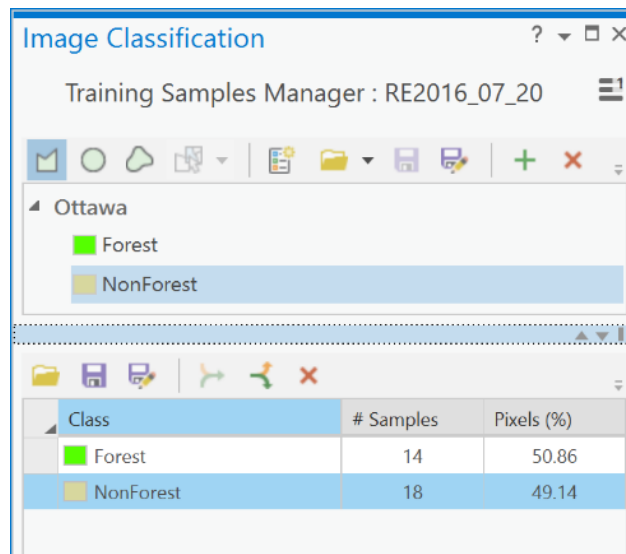

8. Next, in the Training Samples Manager pane, click on the **NonForest** and repeat steps 5-7 for the **NonForest** information class. In the end you should have a Training Samples Manager pane that looks like the one below, with near equal % of pixels in each information class. To ensure you create equal classes, the number of polygons is not what you monitor as you add additional examples of **NonForest**, instead you observe the Pixel % of the **Forest** class and try to get it to about 50%. Then merge all the **NonForest** polygons at that point:

9. Now your Training Samples Manager pane will look something like this:

10. Next, we need to save our new training polygons to a feature class. To do so, click on the 'Save' button in the lower-pane of the Training Samples Manager and navigate to IMAGERY.GDB and save the new training file with the name "RE2016_07_20_training".

Q4: How do I create a signature file from my training polygons?

To create a signature file you need a training polygon dataset and the multiband image that you want to use to extract the pixel information from each of the training information classes. Then you need to assess separability again, before creating the signature file, because you are using a new image.

1. Add the "RE2016_07_20_training" feature class from Q3 above to the Contents pane pane and symbolize it using unique values and the Classname field, and make **Forest** as green and **NonForest** red.
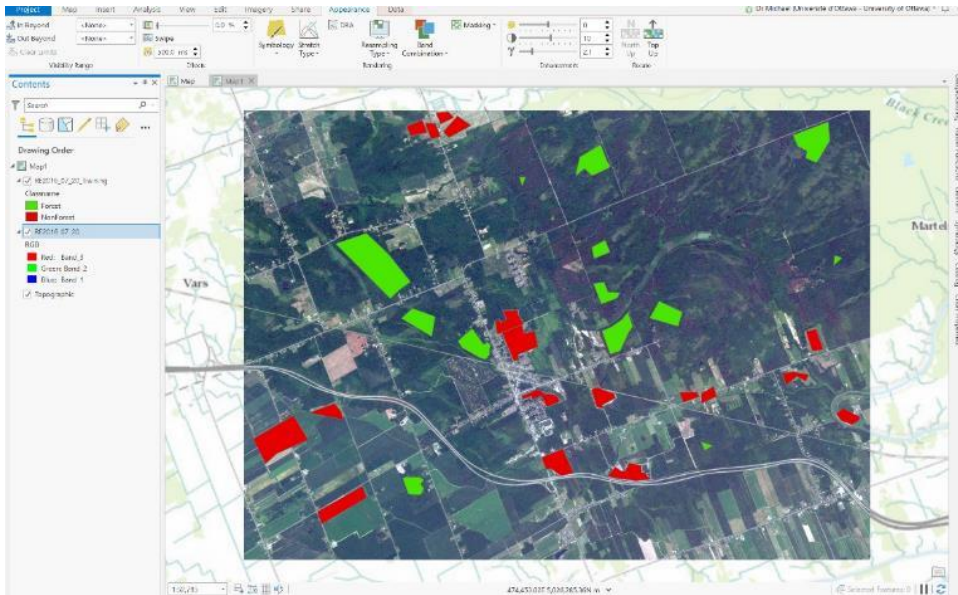

2. Next, choose the 'RE2016_07_20' layer in the contents pane. Then, click on the Data tab. Now click on the Create Chart and choose 'Spectral Profile':
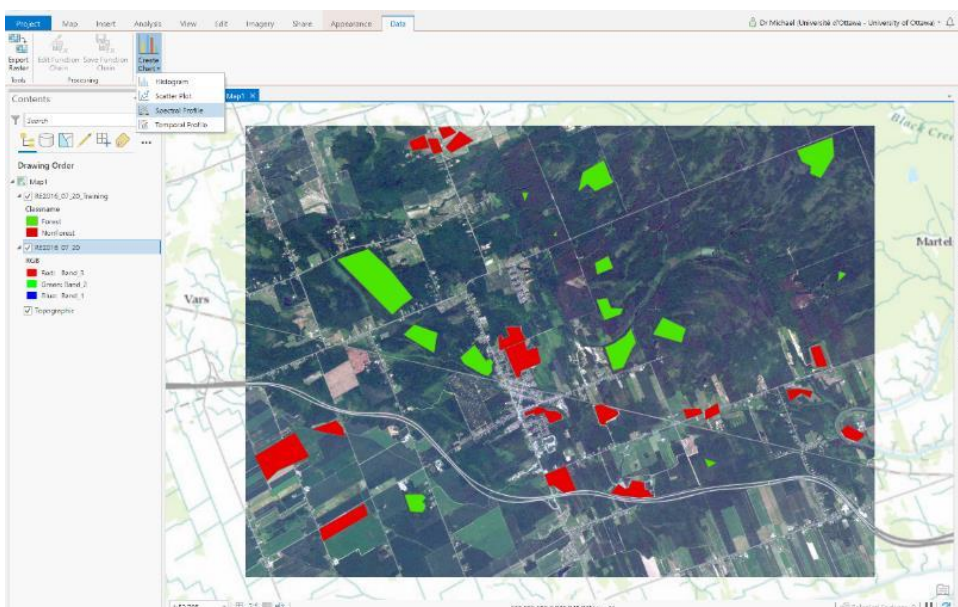
3. Next, in the Chart Properties pane, click on the Feature selector button:
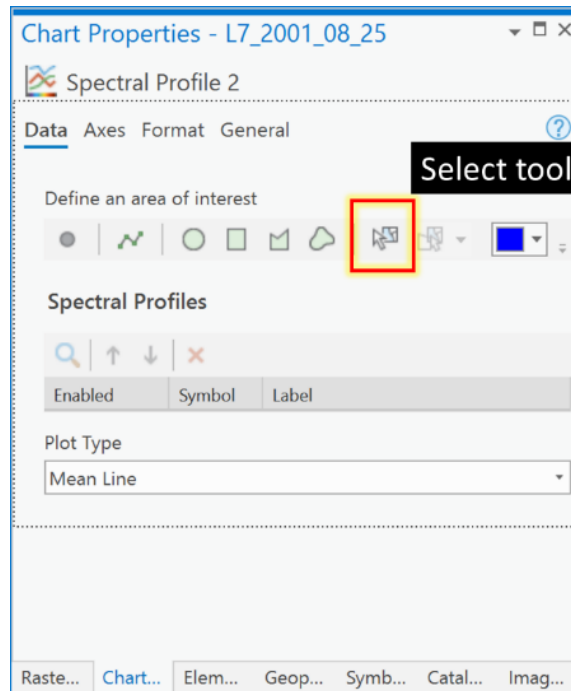
4. Now, click on any of the **Forest** polygons in the map pane. You will see a chart created with a single line called "Profile_1" that represents the average or mean of the DNs across all bands for that information class called **Forest**. In the Chart Properties pane, change the label for this spectral mean line from "Profile_1" to **Forest** (if it works!) and change the color to green and you will see your line become green:
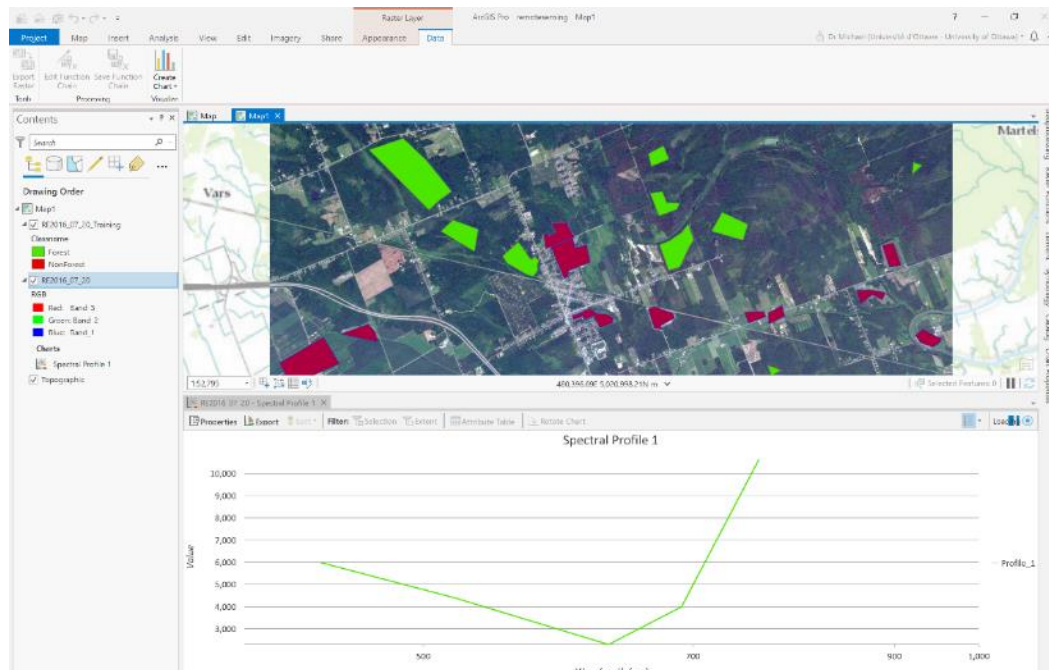
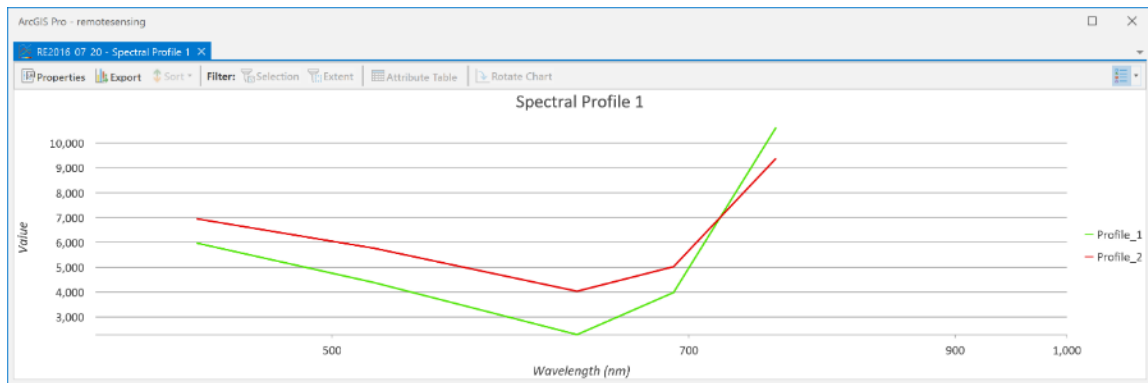5. Now repeat steps 3-4 for the **NonForest** training polygons, and make the new line red to correspond to **NonForest.** Rename the label if you can in your version of ArcGIS Pro:
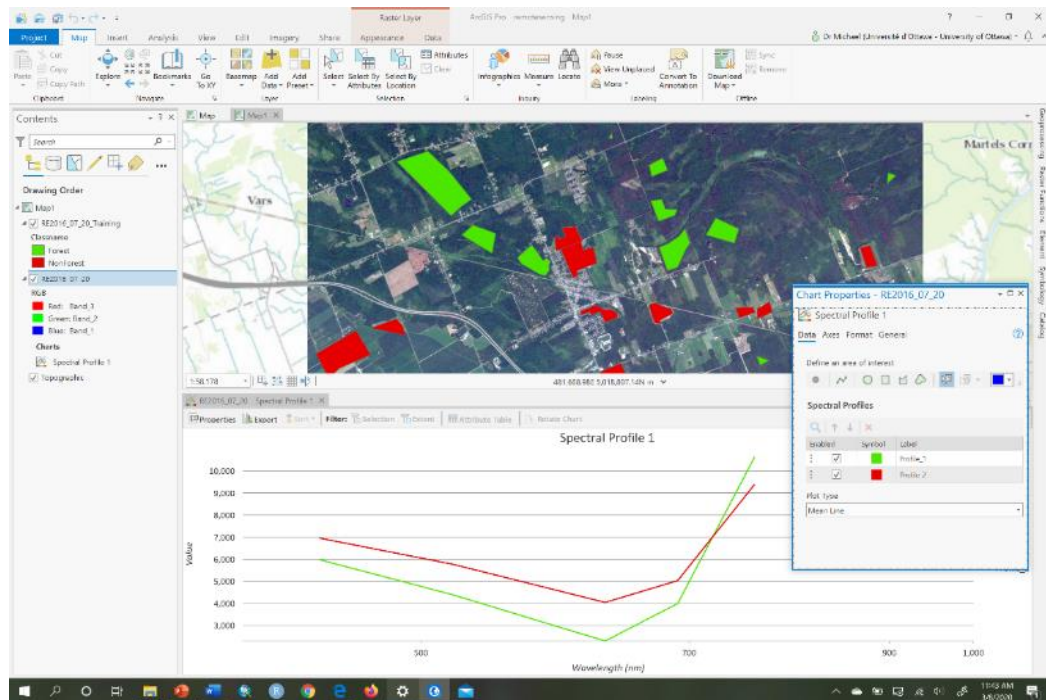
6. Now examine the Spectral Profile graph:

Just as in Q2, this graph shows wavelength on the x-axis rather than band labels, because the metadata of the image contains information on the wavelength of each band. The mean DNs show consistent differences, but it's hard to judge separability from the means alone so we need to examine the box-and-whisker plots to better choose those bands whose boxes do not overlap and so have good separability.

7. To create the box and whisker plot of the spectral separability for each band we only need to change the plot type in the Chart Properties pane. In the pane change the Plot Type from 'mean' to 'boxes':
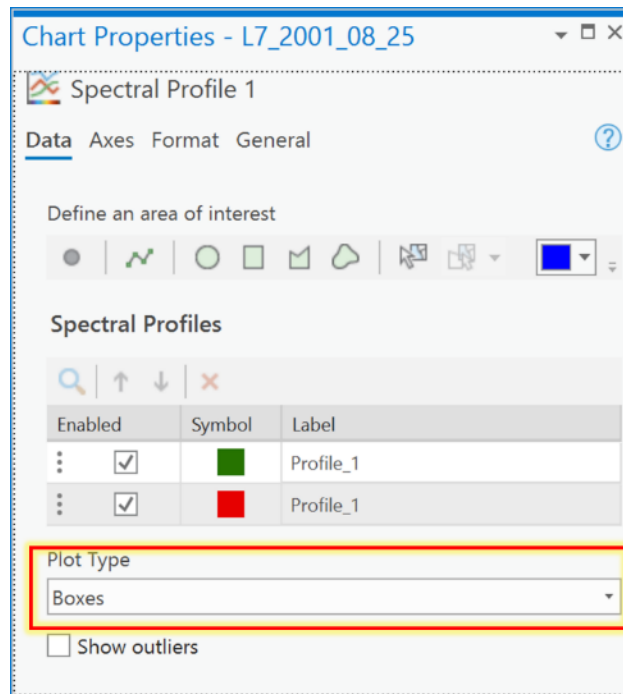
8. Now examine the box and whisker plot.
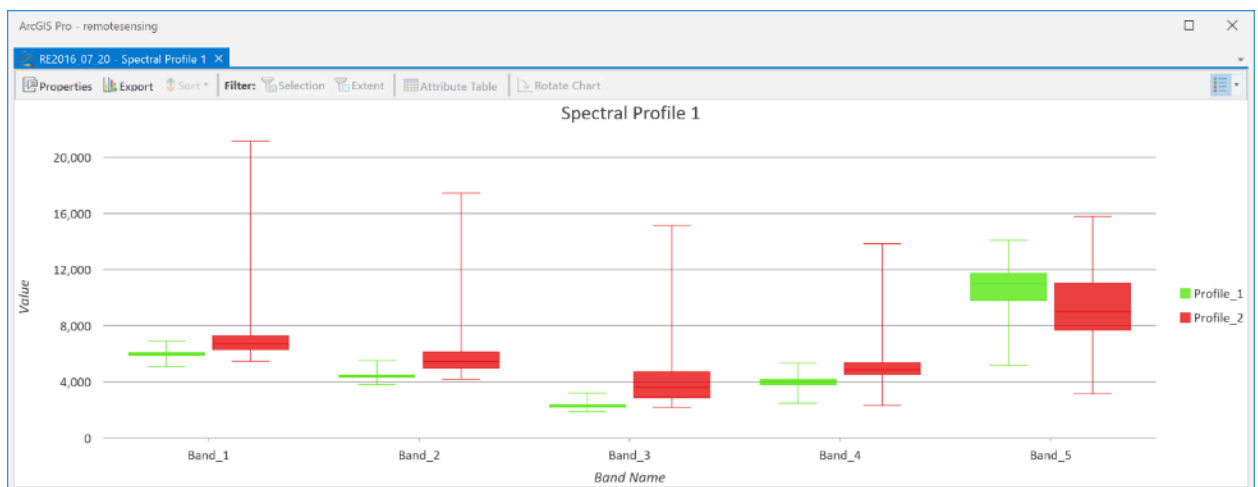
Most bands show decent separability, except for Band_5 where the boxes overlap. The spectral bandwidth of Band_5 on this RapidEye sensor is 760-850 nm, which is in the near-infrared part of the spectrum. This is the same part of the spectrum in which the "L7_2001_08_25" image showed poor separability for **Forest** and **NonForest** (in Band 4, which covers 750-900 nm for the Landsat sensor). So, for our present image, Band_5 will be excluded from any classification of **Forest** and **NonForest**.

9. Create a signature file from the training sites that we can use in classification. To create the file, go to Spatial Analyst -> Multivariate and choose the Create Signatures tool. Undock the Create Signatures tool so that you can see that and your Catalog pane at the same time. Then, in the Catalog pane, navigate to IMAGERY.GDB and expand the "RE2016_07_20" image so that you can see all the individual bands (Band_1, Band_2,…,Band_5). Drag all bands **except** Band_5 (since we know it has bad separability) into the 'Input raster bands', under 'Input raster or feature sample data' choose the "RE2016_07_20_training" feature class, under 'Sample field' choose Classname. Save the file to the folder where your data are and call it "RE2016_07_20.gsg". Click Run.
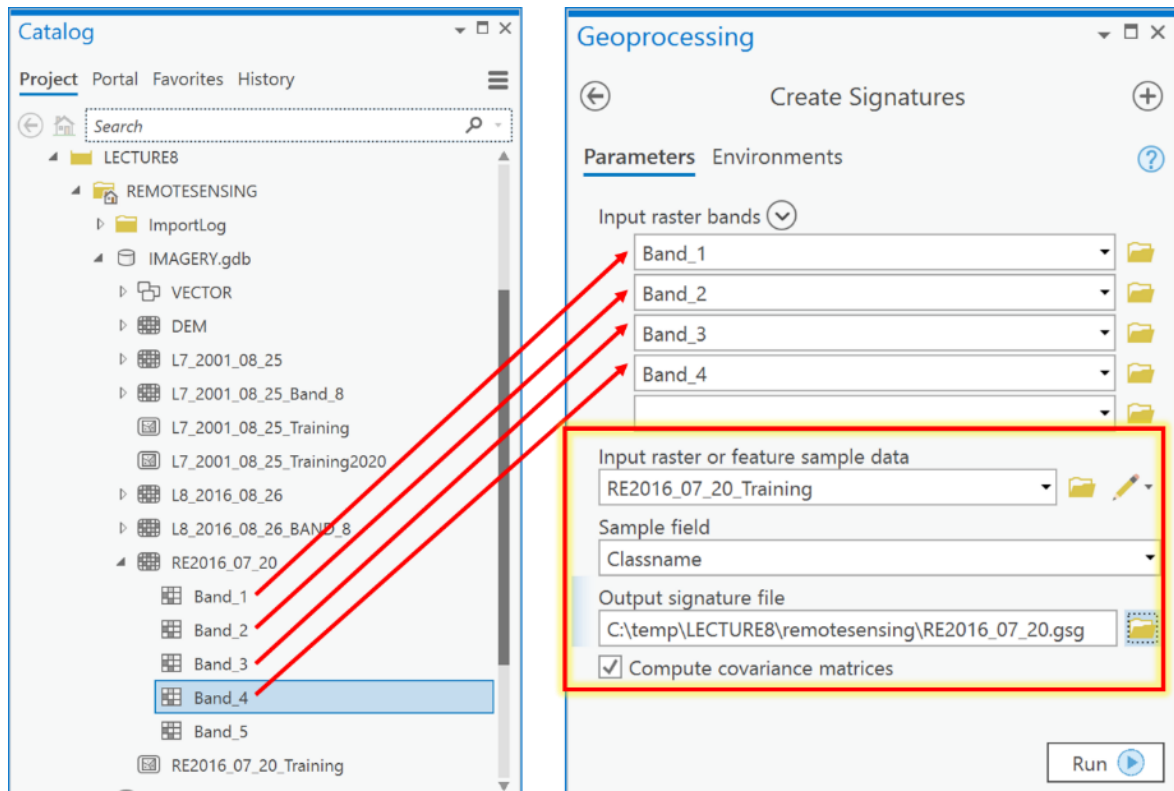
10. You now have a signature file that you can use in image classification.

Q5: How do I classify my image using supervised image classification?

To classify an image, you need to have completed the above example questions because you require the signature file "RE2016_07_20.gsg" for the 'RE2016_07_20' RapidEye image in order to undertake supervised classification.

Now that you have a signature file from the previous example question, you can go ahead and undertake a supervised classification:

1. Use the signature file to classify the RapidEye image into the two information classes of **Forest** and **NonForest**. To do so, go to Spatial Analyst -> Multivariate and choose Maximum Likelihood Classification. Undock the Maximum Likelihood Classification tool so that you can see it with your Catalog pane at the same time.

2. In the Catalog pane, navigate to IMAGERY.GDB and expand the RE2016_07_20 image so that you can see all the individual bands (Band_1, Band_2,…,Band_5). Drag all bands **except** Band_5 into the 'Input raster bands', under 'Input signature file', choose the signature file created in Q4 called "RE2016_07_20.gsg", under 'Output classified raster' call it "MLCLASS_RE_21", and save it in IMAGERY.GDB. Leave all other values as defaults and click Run:
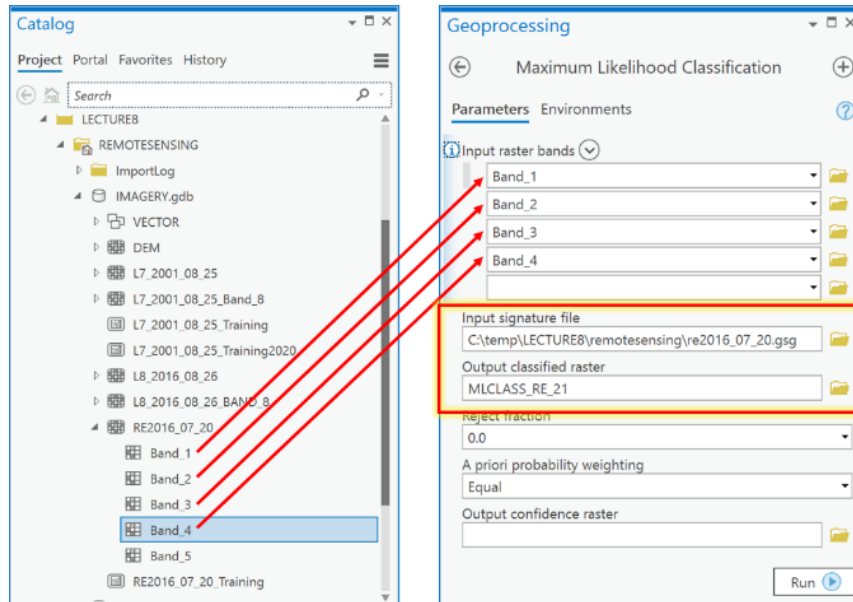

3. Examine the resulting classified raster (note that you may want to change value 1 to green and value 2 to red. We now have converted the image into a thematic raster layer.
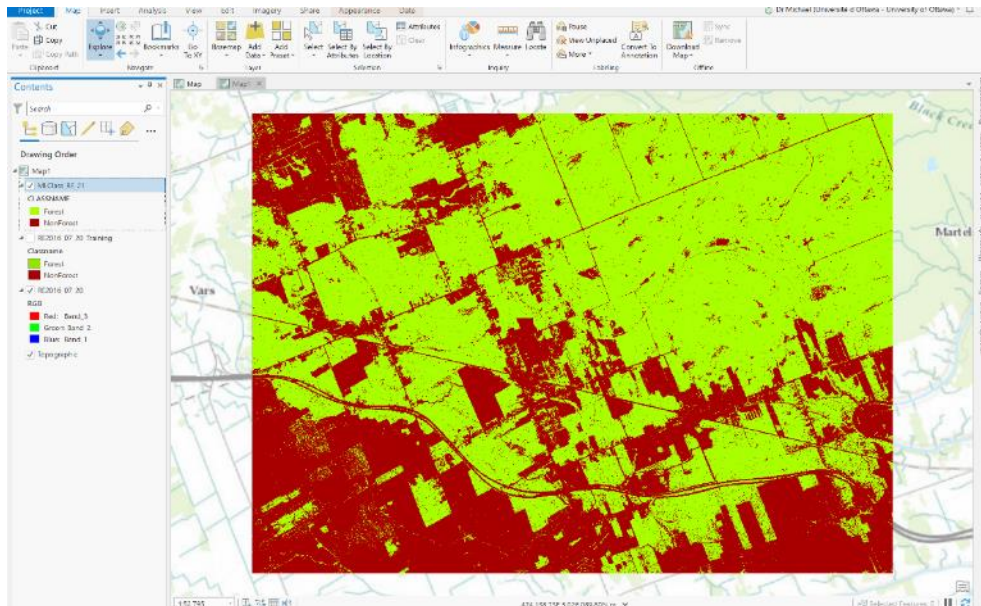
## Q6: How can I clean up my classified raster layer?

In the classified result, you will notice many artifacts of the classification process; these are particularly evident as **Forest** pixels surrounded by **NonForest** pixels or vice-versa. As a general rule, the higher the image resolution (the smaller the pixels), the more misclassified pixels there are likely to be. Given that the Landsat 7 layer was at a 30 m resolution, and the RapidEye was at 5 m resolution, it is not surprising that there are more misclassified pixels in the map made from the RapidEye image. For example, looking at the Landsat and RapidEye classification results side-by-side (as shown below), we can see that the **Forest** from Landsat 7 is more solid green with fewer misclassified pixels than the RapidEye in the same location.
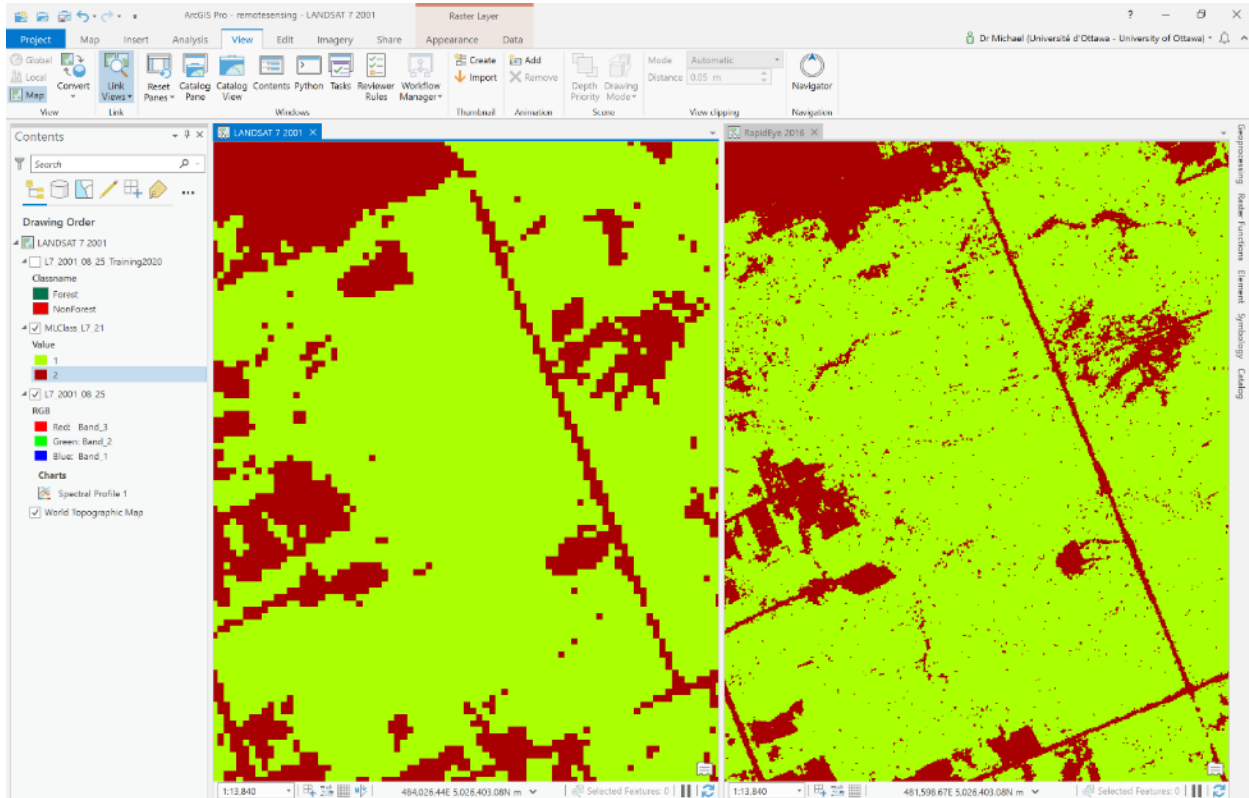

Single **NonForest** pixels surrounded by **Forest** are unlikely to represent forest loss at that pixel in the majority of cases. It partly comes down to the definition of what a "Forest" actually is?!? While this is not a formal definition, many would consider a forest an areas that is dominated by trees – but that doesn't mean there has to be trees everywhere! Forests can have clearings without trees, but those clearings are still part of the forest. With the RapidEye imagery, such learings, or areas that only contain a small tree, or an unhealthy tree, may have been misclassified as **NonForest** (see examples in the figure below).

We can filter out some of the individual erroneous pixels using a majority filter:

1. Open the Majority Filter tool under Spatial Analyst Tools -> Generalization:
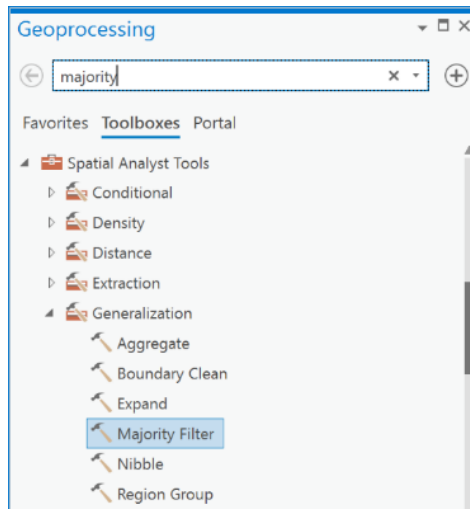
2. In the Input raster drop-down, select your "MLCLASS_RE_21" classified raster layer. Call the output "MLCLASS_RE_21_filtered" and save it to IMAGERY.gdb. Choose the Number of neighbors to use (optional) as eight neighbors, and Run it.
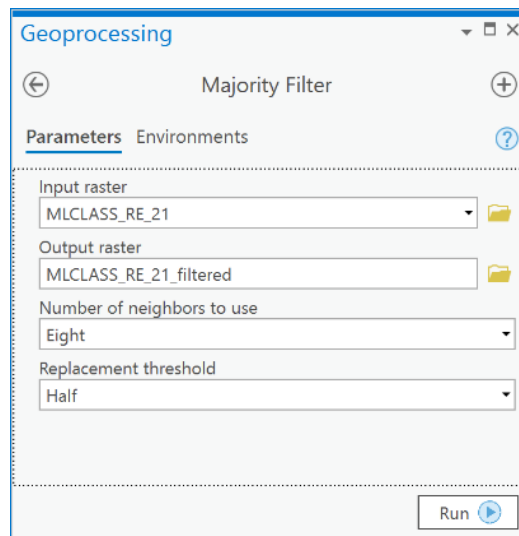
3. The result is a cleaner classified layer with less misclassified/abhorrent pixels:

Sometimes you may run the majority filter (or similar tools) a few times to get the result at the level of generalization that is required for a particular purpose.

Q7: How do I generalize my "MLCLASS_RE_21_filtered" classified raster so that the cell size is the same as the layer "MLCLASS_L7_21" from Q2?

The Landsat 7 scene that was classified in Q2 is in the layer called "MLCLASS_L7_21" and has a spatial resolution of 30 meters (i.e., each cell is 30 x 30 meters). Your "MLCLASS_RE_21_filtered" layer has the same resolution as the RapidEye image which is 5 m (see comparison below).
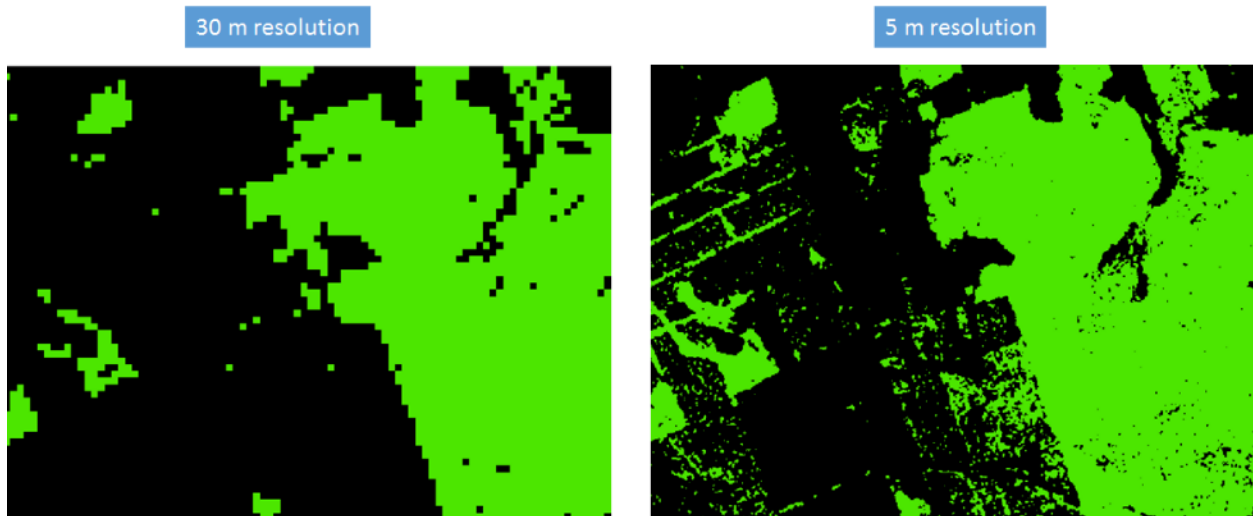

Because there are 36 cells in "MLCLASS_RE_21_filtered" layer for every cell in "MLCLASS_L7_21", you will be increasing the cell size for "MLCLASS_RE_21_filtered" to match the cell size of "MLCLASS_L7_21".
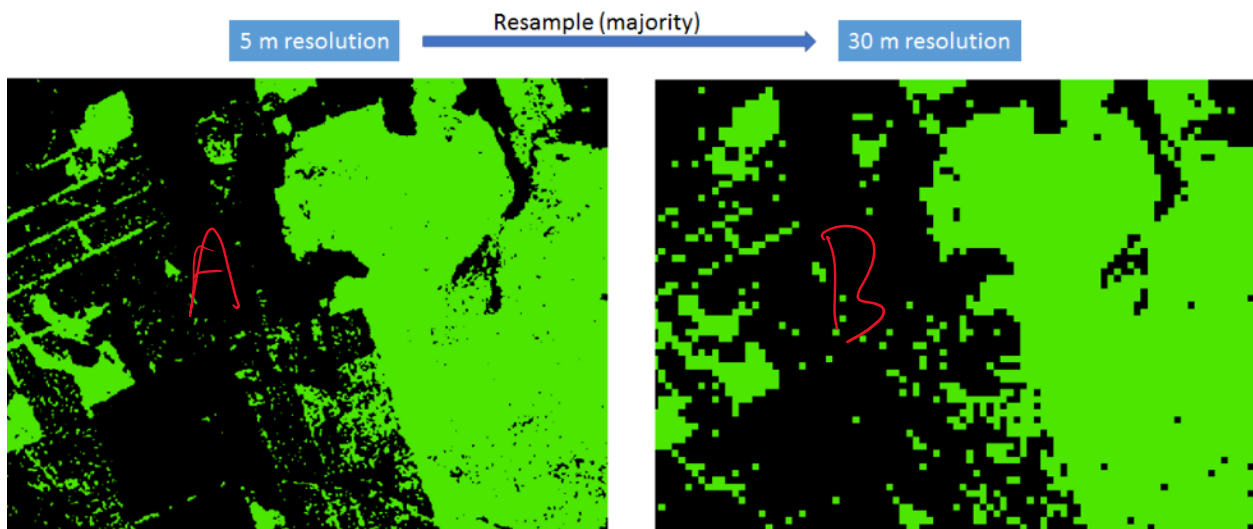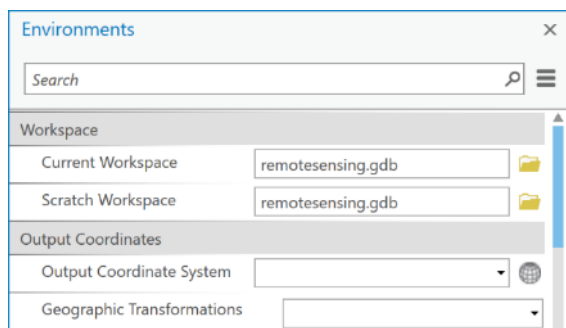
So, you must tell ArcGIS how to summarize the "MLCLASS_RE_21_filtered" cells when increasing the cell size of that layer. In other words, within each new 30 x 30 m cell, ArcGIS must look at the 36 smaller cells that it is made of and decide how to calculate the value (**Forest** or **NonForest**) within the new cell size. Because we're dealing with nominal data (**Forest** and **NonForest**), ArcGIS will by default simply take the 5 x 5 m cell value closest to the cell center of the new 30 x 30 m cell. However, with nominal data, a more appropriate aggregation function would be for the new 30 x 30 m cell to take on a value of the majority of values within it. For example, if there are 24 **Forest** cells and 12 **NonForest** cells and within the new 30 m cell, then the new cell at 30 m should be a **Forest** cell, as in the example below.
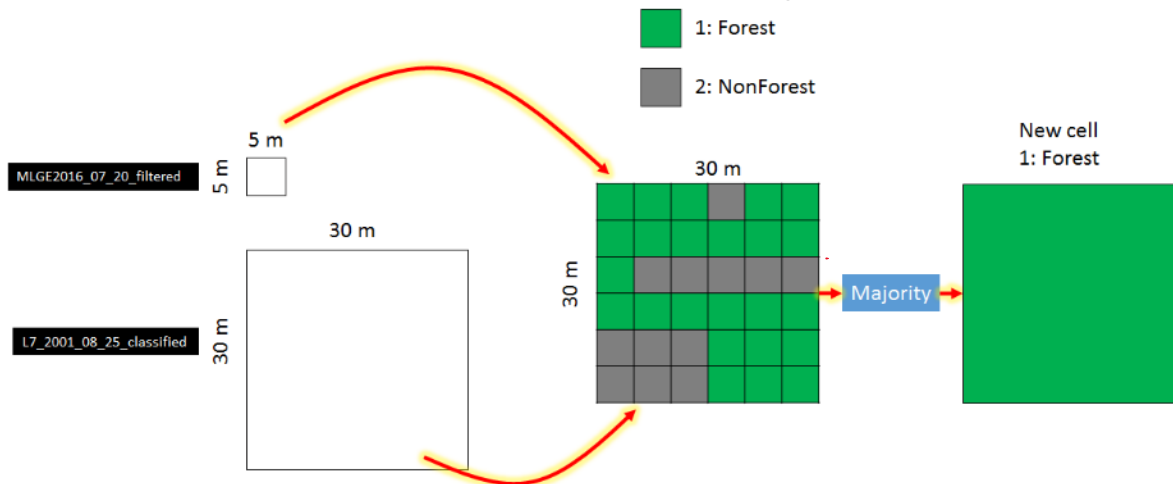
When you generalize your raster layer to the same cell size as the Landsat 7 classification result, you will also want to ensure that the cells in the generalized layer line up correctly with the cells in the Landsat 7 layer. This is done by setting the "Snap Raster" in the global environment settings as well as the Processing Extent:

1. Click on the Analysis tab and then click on the Environments button.
2. In the Environment Settings dialog, under Processing Extent, under Extent choose "MLClass_L7_21" and under Snap Raster, also choose "MLClass_L7_21". Then under Raster Analysis, Cell Size choose "MLClass_L7_21" and under Mask also choose "MLClass_L7_21" and then click OK:

By completing Step 2, you have ensured that any processing output in raster analysis will conform to the cell size and extent of the "MLCLASS_L7_21" layer. Keep in mind that these settings will be maintained until you close and reopen ArcGIS or change them again. So, if you decide to do other analyses, with other layers, that have a different extent than "MLCLASS_L7_21", results will only be produced within the specified extent. Thus, you may need to clear the extent property or set it back to its default value when you are done with the raster analyses.

3. To actually convert the RapidEye product to 30-m resolution, you use the Resample tool under Data Management -> Raster -> Raster Processing. Open that tool:

Student Name: _____ Student Number: _____
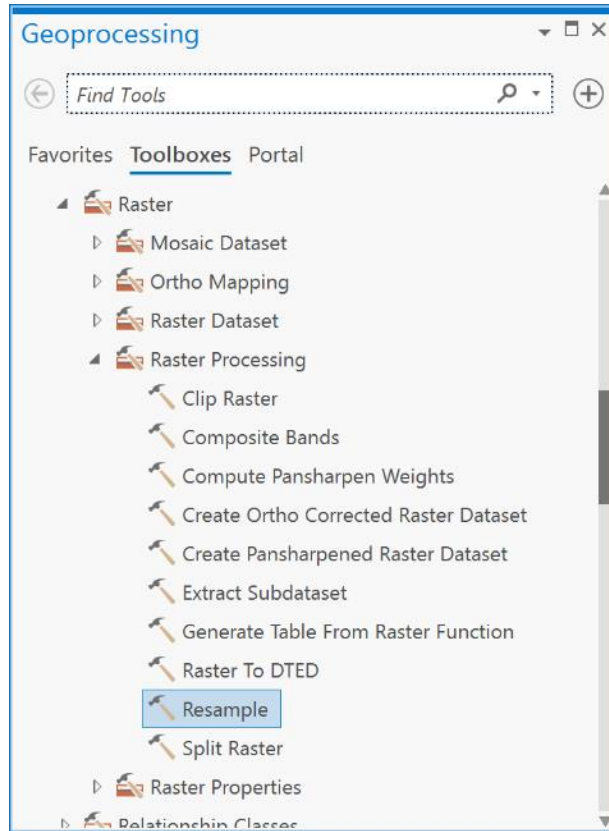
4. In the Resample dialog, your input raster should be specified as "MLCLASS_RE_21_filtered", the Output Raster should be called "MLCLASS_RE_21_classified" and the Output Cell Size (optional) should be specified "MLClass_L7_21". Most importantly, the Resampling Technique (optional) should be set to MAJORITY. Then click OK:
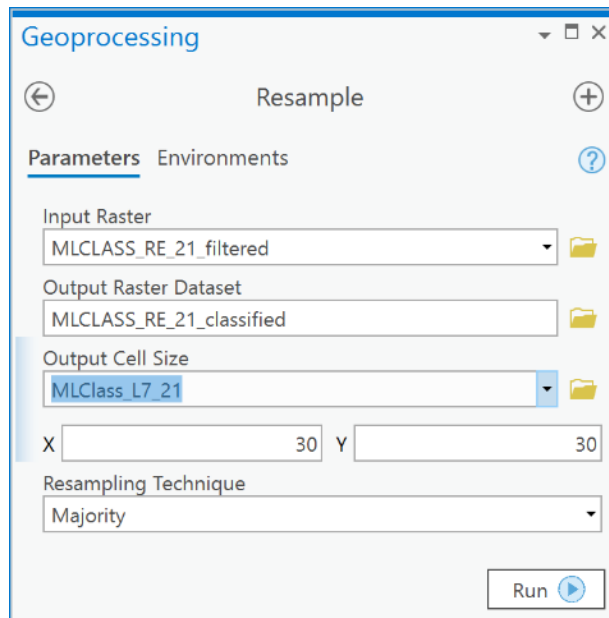
5. Your aggregated result will look like this:
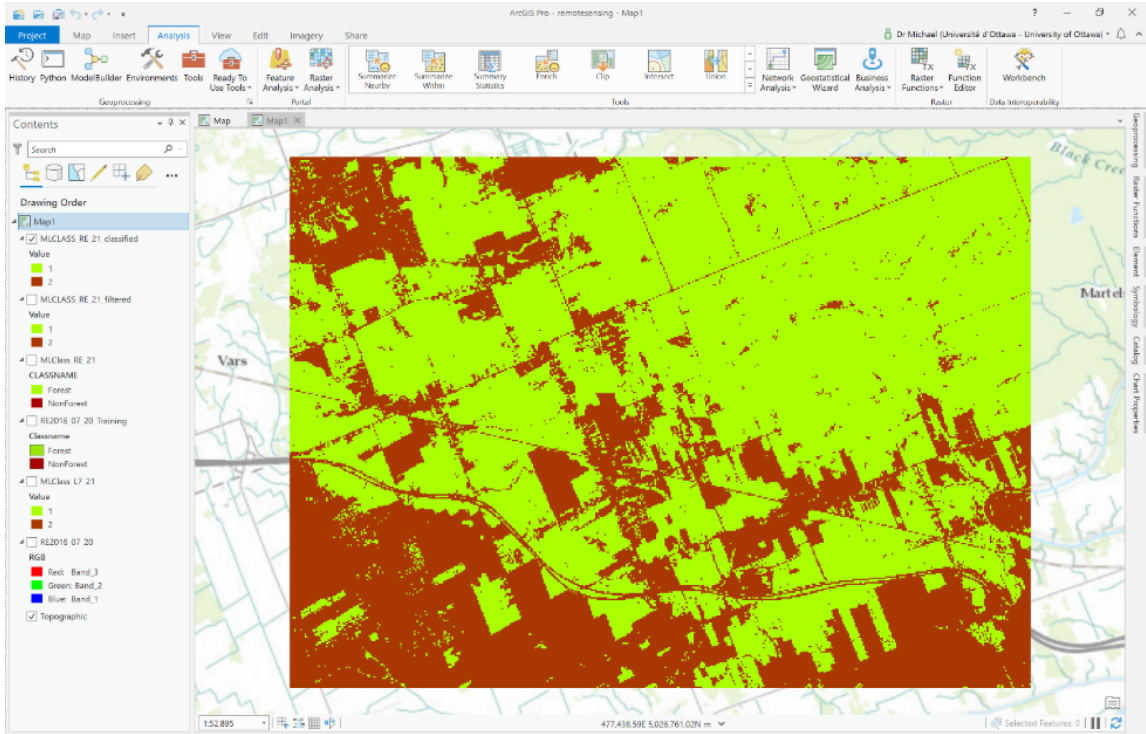
Q8: How do I test the accuracy of a classified result?

The accuracy of a classification is tested by comparing classified pixels to what are called 'ground-truth' data. The ground-truth (GT) data are locations that have not been used in the classification but are known to belong to one of the information classes that you have defined, either belonging to **Forest** or **NonForest**. In other words, the GT data is a "test dataset" and is independent of the data used to create the signature file that trained the classifier. GT data can be acquired by either going into the field with a GPS unit and observing multiple locations for the presence of any particular information class. However, unless absolutely necessary, other methods of collecting independent GT data are typically used because field-based GT data is often difficult, slow, and expensive to acquire. Thus, you can usually collect examples of GT data by directly observing features within the pre-classified image and labelling those as one of the information classes, in our case as either **Forest** or **NonForest**. To assess the accuracy of a classification you would collect such data at several randomly selected points and compute a confusion matrix that would allow you do quantify how often the classification is correct, and how often it isn't.

Note: The term 'ground-truth' is frowned upon by some, because it implies that these data are inherently "true". This obviously isn't correct, because they are made by humans, and humans make mistakes. Nonetheless we use the term here because it is common, and you will encounter it in the future if you work more with satellite imagery. Alternative terms used for the same thing is "validation data" or "test data", but those terms themselves can mean different things to different people, so they are not ideal either. We'll stick with 'ground-truth' for this exercise!
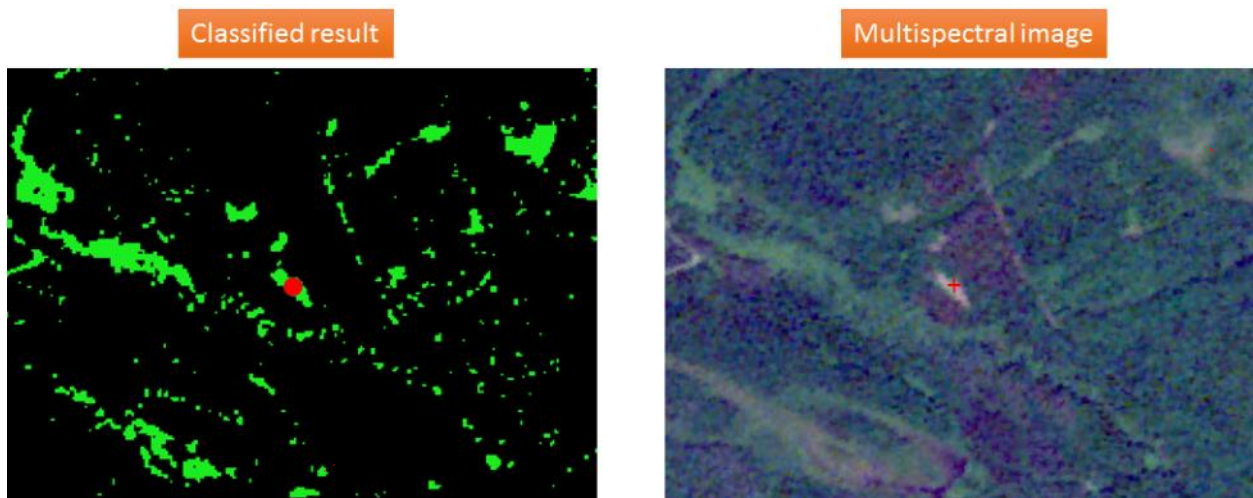
Example of a misclassification. Pixel classified as Forest (left) is clearly not a forest (right) as can be seen in a ground truth point (red cross, right) identified manually from a multispectral image.

You will calculate the accuracy of the "MLCLASS_L7_21" result, since everyone will have the same result from Q2 above.

1. Go to Spatial Analyst -> Segmentation and Classification and choose the tool Create Accuracy Assessment Points.
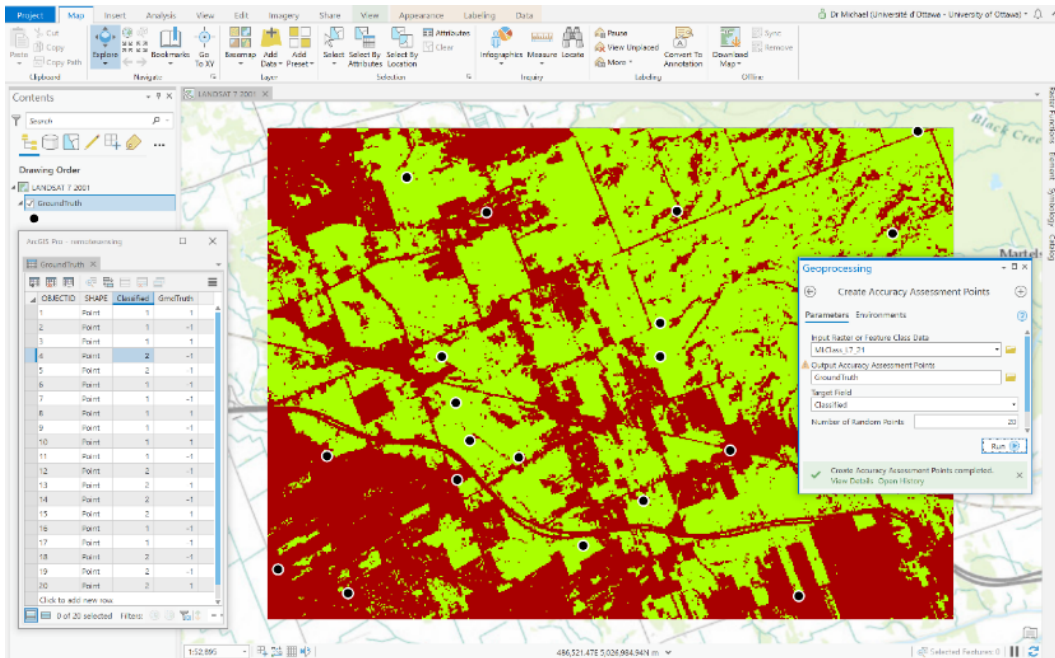
2. In the tool, under Input raster or feature class data, choose the "MLCLASS_L7_21" layer and under Output Accuracy Assessment Points, call the output "GroundTruth" and ensure that you save it in the IMAGERY.gdb geodatabase. Under Target Field choose Classified. Under Number of Random Points choose 20, and finally under Sampling Strategy choose Random.

   This will create 20 points randomly (where the x and y coordinates for each point are chosen, within the extent of the image, from the uniform probability density function independently and with equal probability) within each of the classes **Forest** and **NonForest**. For each random point, the pixel value from the "MLCLASS_L7_21" layer will be extracted at the point location within the "MLCLASS_L7_21" layer. These values will be stored in the "GroundTruth" table in a column called "Classified" containing the classified pixel values for each of the points.
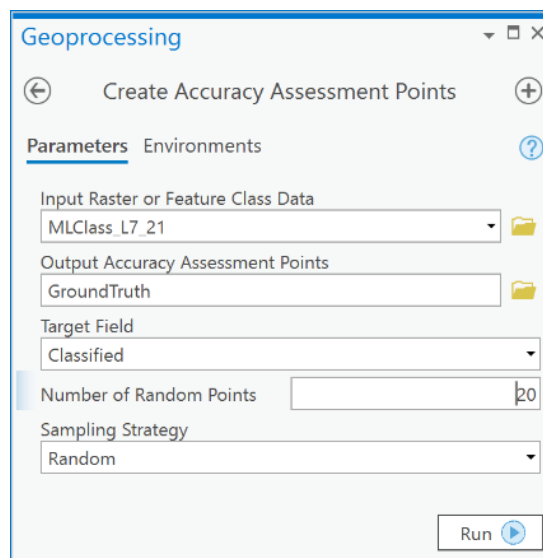
3. The result will be 20 randomly located points within the image:

4. Now open the table called "GroundTruth". It should look something like this:

5. Each random point has assigned to it the information class value for either **Forest** = 1 or **NonForest** = 2 depending on which class contained the point. This data will be in the "Classified" column of the "GroundTruth" attribute table (see the above table).

6. Now we need to fill in the GrndTruth column of the table which has values of -1 right now. The -1 values tell us that the 'true' class for that point has not been observed and entered into the table. So we need to fill these 'true' class values, remembering that the value 1 means **Forest** and 2 means **NonForest**. To do so, ensure the only layers visible on the map are "L7_2001_08_25" and "GroundTruth". Now select the first row in the "GroundTruth" table, to identify the first point. For the currently selected row, right-click and choose Zoom To. Then zoom a little closer until you can clearly see what information class the currently selected point is in. Now, enter either a 1 or 2 in the GrndTruth column, depending on your interpretation of the image at that point. In the below example, you can see that **Forest** is under the selected point, so you would enter a value of 1 in the GrndTruth column for that row by double-clicking the cell where you want the 1.
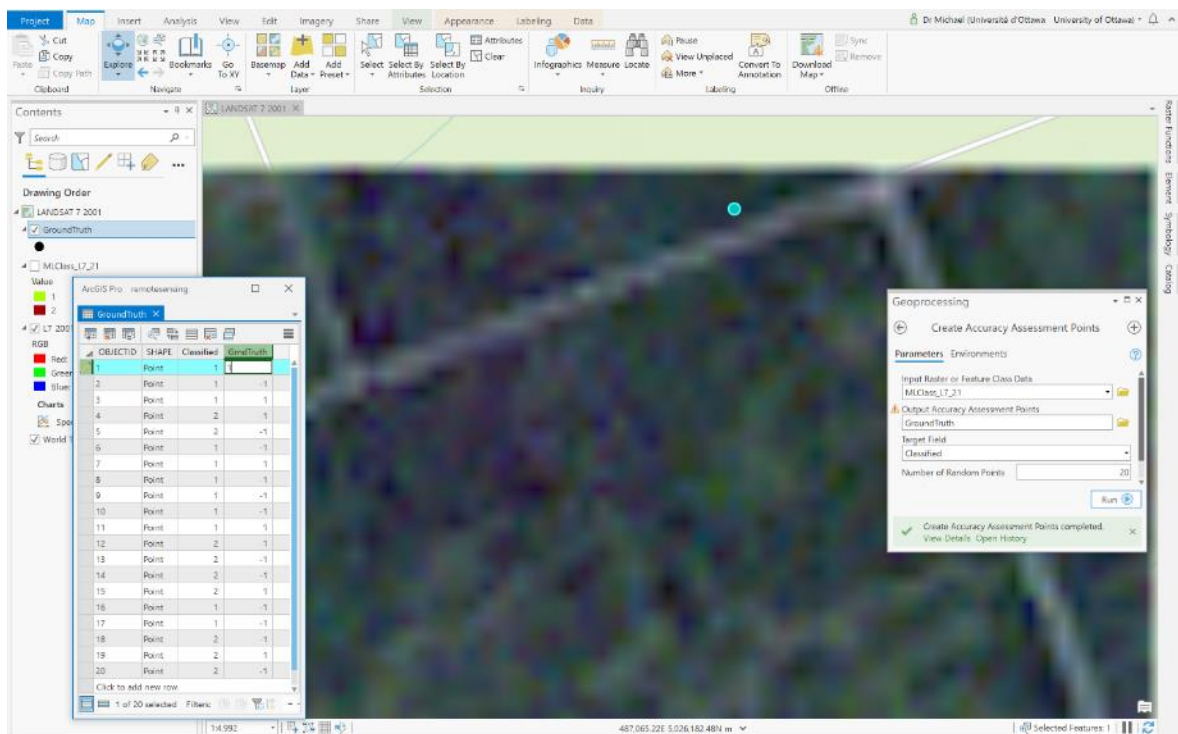
7. Now select the second row, Zoom To, and examine whether or not there is **Forest** or **NonForest** at that location and enter either a value of 1 or 2, respectively in the GrndTruth column of the "GroundTruth" table. In this example, again, it is **Forest** so you would enter a value of 1:
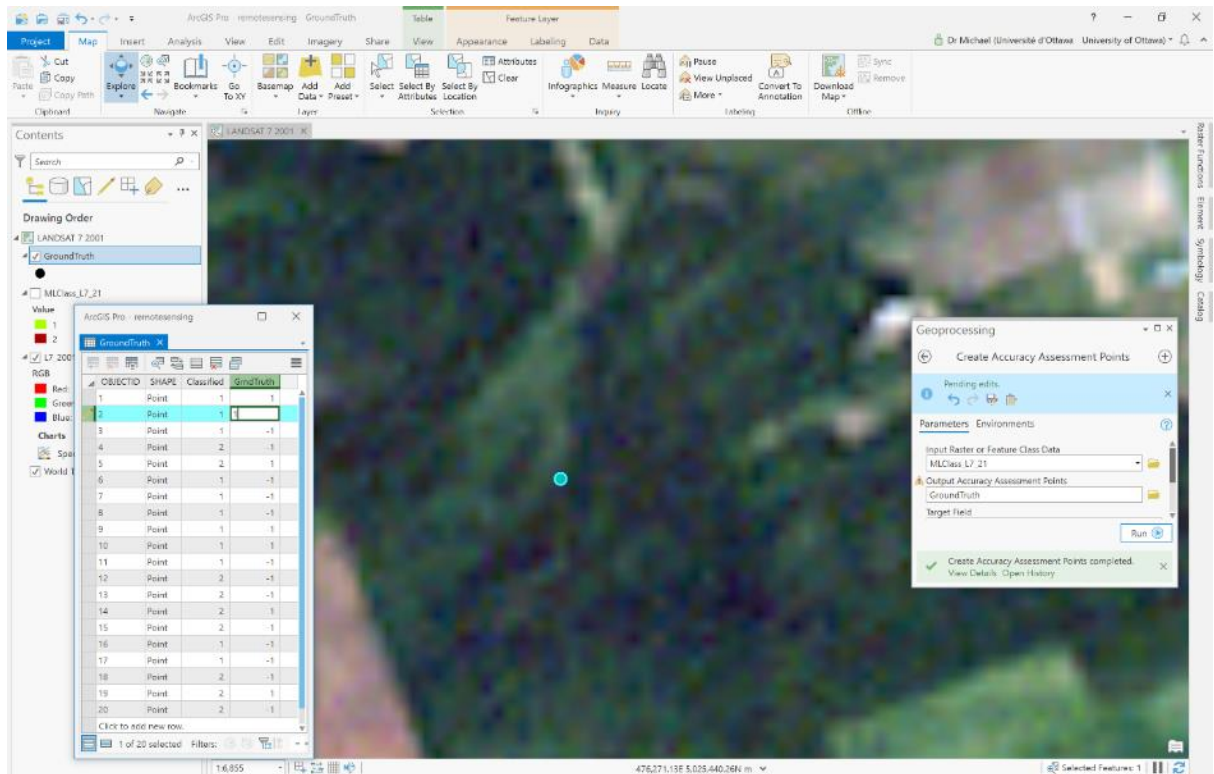
8. Repeat these steps for all the points in the "GroundTruth" table, to create a table that is complete and ready for accuracy assessment, e.g.,:

9. Now, go to the Edit tab and click on Save to save the edits to the "GroundTruth" table. Then click on the Clear button in the Selection group on the same tab:
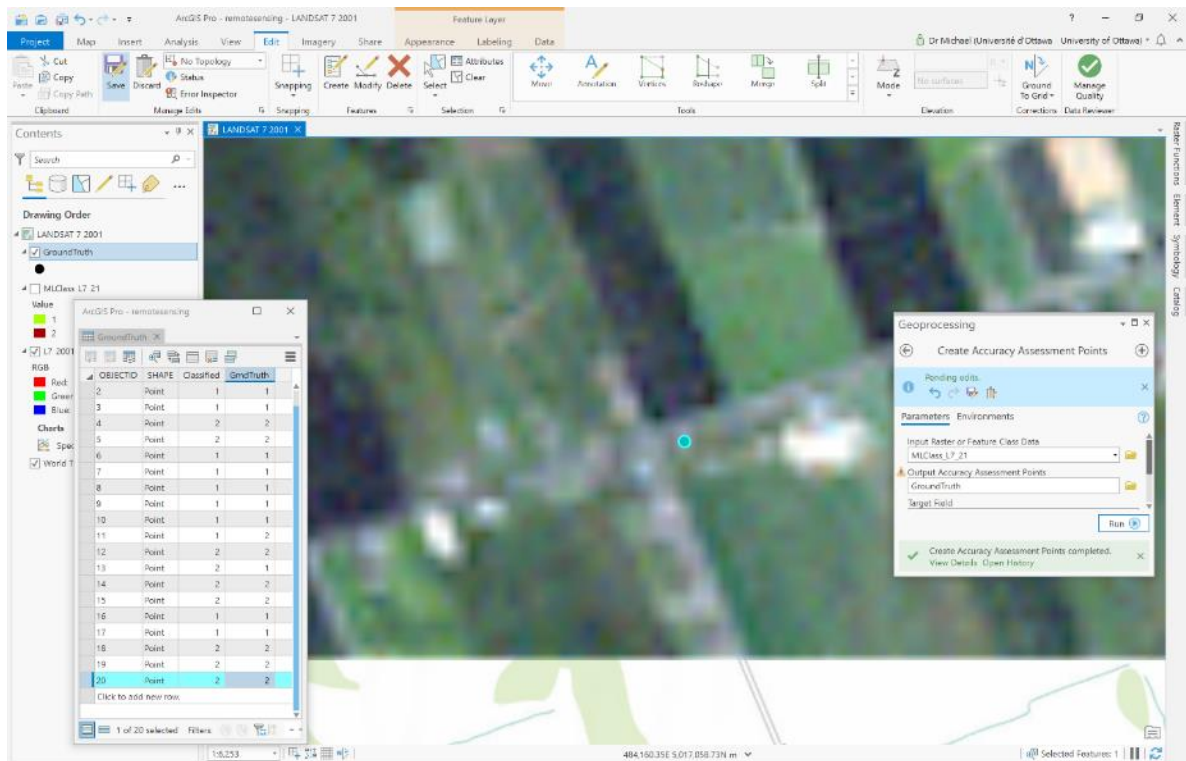
10. Your "GroundTruth" table now contains a column called Classified that contains the output of the classifer at each point and a column called GrndTruth that contains the observed information class at each point. Now you want to tabulate how many times **Forest** (value of 1) corresponds to **NonForest** (value 2), how many times **Forest** (value of 1) corresponds to **Forest** (value of 1), how many times **NonForest** (value 2) corresponds to **NonForest** (value 2) and finally how many times **NonForest** (value 2) corresponds to **Forest** (value of 1). Because we're only dealing with two classes in this example you could easily do this manually, but ArcGIS has a tool that does it for us, called the Compute Confusion Matrix which is under Spatial Analyst -> Segementation and Classification in ArcToolbox. Open that tool and under Input Accuracy Assessment Points choose your "GroundTruth" table. Call the output ConfusionMatrix and save it to your IMAGERY.gdb:
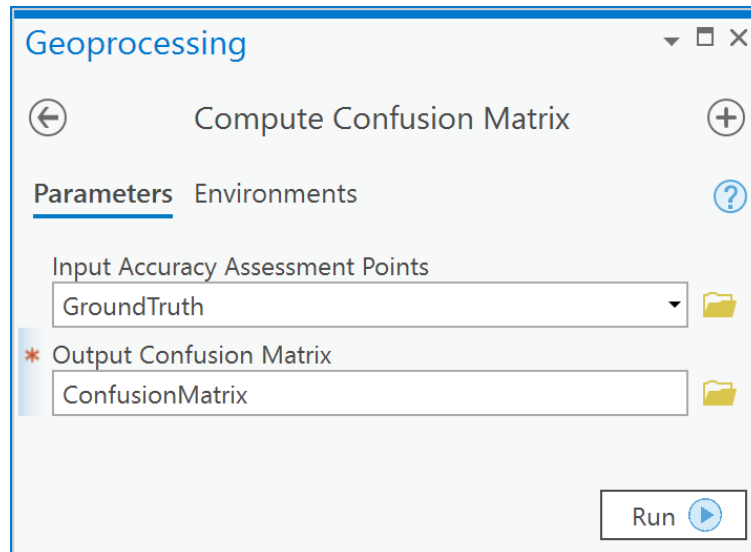
11. Open the ConfusionMatrix table:

In the table, C_1 means class value 1 (**Forest)**, and C_2 means class value 2 (**NonForest)**. The numbers in the cells where a column and row of the same name intersect indicate the number of times the two corresponding classes were correctly classified. For example, in the table illustrated above, there were 10 times where a C_1 corresponded to a C_1 (where C_1 column intersects with C_1 row) and 1 time a C_1 corresponded with a C_2 classified result. Thus, 10 **Forest** points were correctly classified and 1 was incorrectly classified. Similarly, there were 8 times where a C_2 was correctly classified as a C_2, and one time where it was classified as a C_1. Thus, 8 out of 9 times **NonForest** was correctly classified. The Kappa value is a measure, ranging from -1 to 1, of how much better than random the classifier performed. Here it is 0.797, which is considered a modestly good result. You can read the help file for the Compute Confusion Matrix tool for more information.

12. Now repeat all the steps for this question with your "MLCLASS_RE_21_classified" layer using the RE2016_07_20 image as the ground truth layer.

Q9: How can I create a raster with only Forest from each of my classified results, and determine where forest has been lost?

Since you now have two rasters, both containing information classes called **Forest** and **NonForest**, you can use a simple map algebra expression to create a Boolean raster layer where the value of 1 means **Forest** and 0 means **NonForest**. You can do this in several ways, here we'll try to use the Python programming language that's integrated as part of ArcGIS:

1. Open the Python pane:
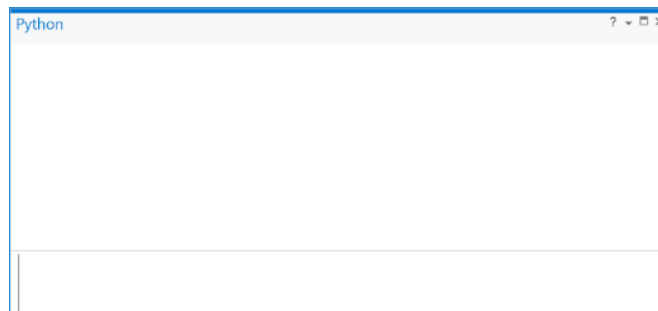
2. Type the following three lines and hit the Enter key (note that you don't need to enter the lines in green that are preceded by the # sign, since these are only comments to tell you what each line does):

```python
# access Spatial Analyst functions
from arcpy.sa import *

# Create a raster variable from the filtered classification results
x = Raster("MLCLASS_RE_21_classified")

# Create a boolean layer using the relational operator for logical equals
forest_2016 = x == 1
```

The first line tells ArcGIS to allow you to access Spatial Analyst functions within the python pane. The second line uses the Raster() function to create a raster object that can be used in a Python map algebra expression. The third line is a relational map algebra expression using logical equals (==). That expression looks at each cell in the raster layer called x, and if the cell has a value of 1 (**Forest**), then the new raster cell's output will also have a value of 1. If the cell's value is anything other than 1 then the output raster will have a value of zero for the cell in question.

3. The result will be a raster layer called "forest_2016" in the table of contents but with pixel values of 1 for cells that are forest and 0 otherwise.

4. Now modify these lines, to create a binary raster for the classified Landsat image:

```
# access Spatial Analyst functions
from arcpy.sa import *

# Create a raster variable from the filtered classification results
y = Raster("MLClass_L7_21")

# Create a boolean layer using the relational operator for logical equals
forest_2001 = y == 1
```

5. Now, to examine the amount of forest that has been lost vs. gained we can subtract the forest_2001 from the forest_2016:

```
# Subtract 2016 from 2001 to see where forest has been
# gained and/or lost.
forest_change = forest_2016 – forest_2001
```

6. Your resulting map will look something like this (it won't be exactly the same because your training data for the RapidEye sensor was different):
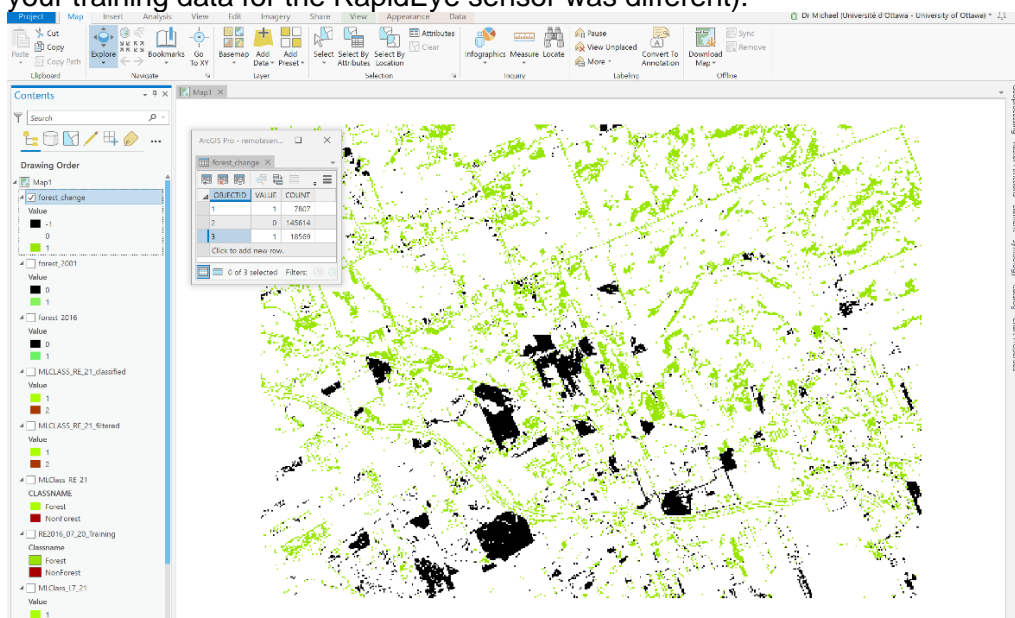
Here the attribute table tells us in the Count field how many total pixels are for each value.

Questions to hand in.

1. [1 point] What is the positional uncertainty due to cell resolution within the "MLCLASS_L7_21" layer?

2. [3 points] In Q2 you created a thematic raster layer called "MLCLASS_L7_21", and in Q3-Q7 you created a layer called "MLCLASS_RE_21_classified". In Q8 you then created confusion matrices for both layers. Present a map for each result and confusion matrix. Explain which result is more accurate, how you know, and what you think the reason is.

3. [3 points] In the raster you created at the end of Q9, you had cell values of -1, 0, and 1. Describe what these values represent in the landscape, e.g., what does a -1 cell value mean, what does a 0 cell value mean and what does a +1 cell value mean? Present the map with your explanation. What is your overall conclusion about forest loss in the region?

4. [3 points] In the differenced layer you created at the end of Q9, you may be interested in the large contiguous *regions* of cells with values that indicate forest loss (conversion from forest to non-forest). Where are these large contiguous regions? Use the Region Group tool under Spatial Analyst -> Generalization with your result from Q9 as the input layer. Read the help file for the Region Group tool by choosing the Tool Help button. Explain what the region group tool does. Now, open the attribute table of the region group result raster, right-click on the "Count" column and sort in descending order. Examine the "Link" field. How do the values in the "Link" field relate to the cell values in the result from Q9? Provide a table of area in map units for the 10 largest contiguous groups of cells that have changed from **Forest** to **NonForest** from 2001 to 2016. Also present a map showing these regions.

5. [2 points] What is the average region size (patch size) in square meters of regions that have changed from **Forest** to **NonForest** from 2001 to 2016? You will need to undertake an attribute query on the attribute table from the region grouped layer from (4) and right click on the "Count" column and choose Statistics to get the mean, in order to answer this question.

6. [2 points] Select only those regions that changed from **Forest** to **NonForest** between 2001 to 2016 and are greater than or equal to 3000 square meters in size. What is the average slope for each region? You will need to use Spatial Analyst -> Surface -> Slope with the DEM data found in IMAGERY.gdb, as well as an attribute query on the region grouped layer from (4) and Spatial Analyst -> Zonal Statistics as Table, to address this question. Present a table showing the average slope for each selected region.

7. [1 point] Examine the layer called L8_2016_08_26, which is a Landsat 8 OLI image. Use your identify tool to click on a few different locations and note the DN each time. Compare those values with the DN values for similar features in the "L7_2001_08_25" layer. Why are the DNs in the Landsat 8 image greater than those in the Landsat 7 image for similar features?

8. [2 points] Create and present an NDVI map using the L8_2016_08_26 image. Consult your lecture notes, or the Internet, to determine the appropriate bands to use for calculating NDVI with Landsat 8 data, as they are not the same as for Landsat 7. What does this map tell you? Why?