

Spatial Data Science I (GGR276)

Assignment 1: Understanding the “GEO” in Geostatistics

Objectives:

1. Enhance the ability and confidence to analyse and interpret non-spatial statistics through real-life examples.
2. Develop familiarity with the spatial description of point events using spatial measures of central tendency and dispersion.
3. Gain experience with the use of ArcGIS Pro and R packages, the command line interface, and scripting for the manipulation and analysis of spatial data.

DUE DATE: Friday July 12th, 2024 @ 1159 pm. | **Grace Period: 1 hour.**

Instructions: This lab is worth a total of **59 marks** and **15%** of your final course grade. You are to work alone and answer the questions on Quercus. The expectation is that you will have the assignment data and computational solutions stored in your user directory.

Introduction

A class of geographical problems exists that requires an analyst to explicitly study the location coordinates associated with specific *events* (e.g., incidence of disease, occurrence of criminal activity, forest fires, power plants). The occurrence of these events is often represented using points that identify the location of particular cases or incidents. Approaches for univariate descriptive analysis (e.g., mean, standard deviation) have been analytically extended to the bivariate case (X, Y) to facilitate spatial description of the distribution of point events. Geographical descriptive statistics of this sort are typically referred to as *geostatistics* because they can be used to summarize spatial qualities of a pattern of events located in a geographical space. In this assignment, real-life examples will be provided to enhance your confidence in analyzing and interpreting non-spatial descriptive statistics. Stepping on this foundation, you will explore the implementation, application, and interpretation of spatial measures of central tendency and dispersion.

Documentation for the lab

2 files are required:

1. Lab instructions for PART 1 & 2 are included in this PDF.
2. Documentation for PART 3 is in the file “GGR276Lab1P3_Starter_2024FT.Rmd”.

Software for this lab:

Excel, RStudio and ArcGIS Pro will be used to view the data for this exercise. You can use the lab computers or utilize the Web-based Learning Platform and access ArcGIS Pro through remote access to lab desktops.

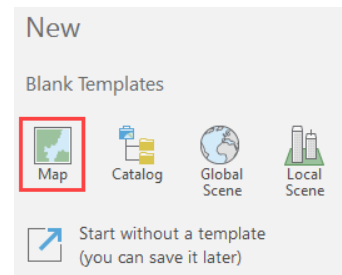
Submission guideline:

Please answer all questions in PART 1 & 2 in a word document and PART 3 in an R Markdown document. You will be asked to submit three documents to Assignment 1 on Quercus: 1) a word or pdf document, 2) an R Markdown document, and 3) an HTML file rendered from the R Markdown document.

There are many open data sources for geospatial data. In this lab, we will search for spatial data through ArcGIS Living Atlas.

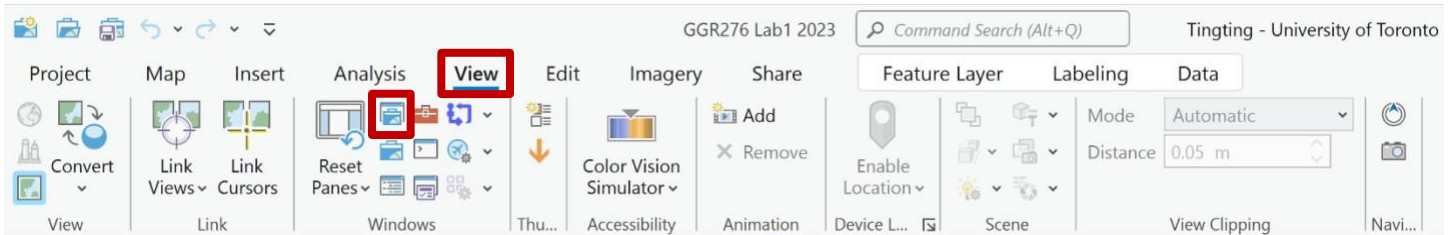
STEP 1: Start ArcGIS Pro

- Feel free to use either your laptop, lab computer or web-based platform through remote access.
- [if using lab computer] Start > Course Applications > ArcGIS Pro, and click Map in ArcGIS Pro
- Or by searching for ArcGIS Pro using Search Windows

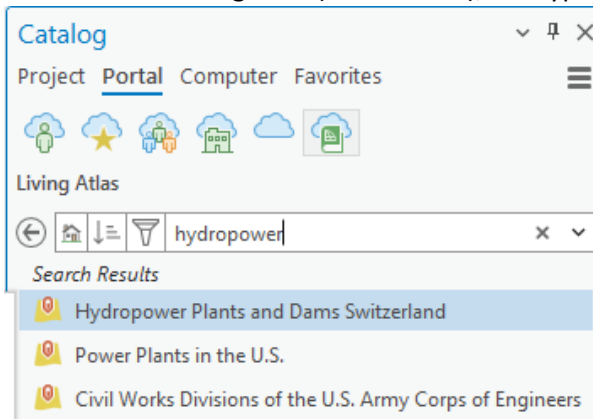


STEP 2: Search for data in ArcGIS Living Atlas

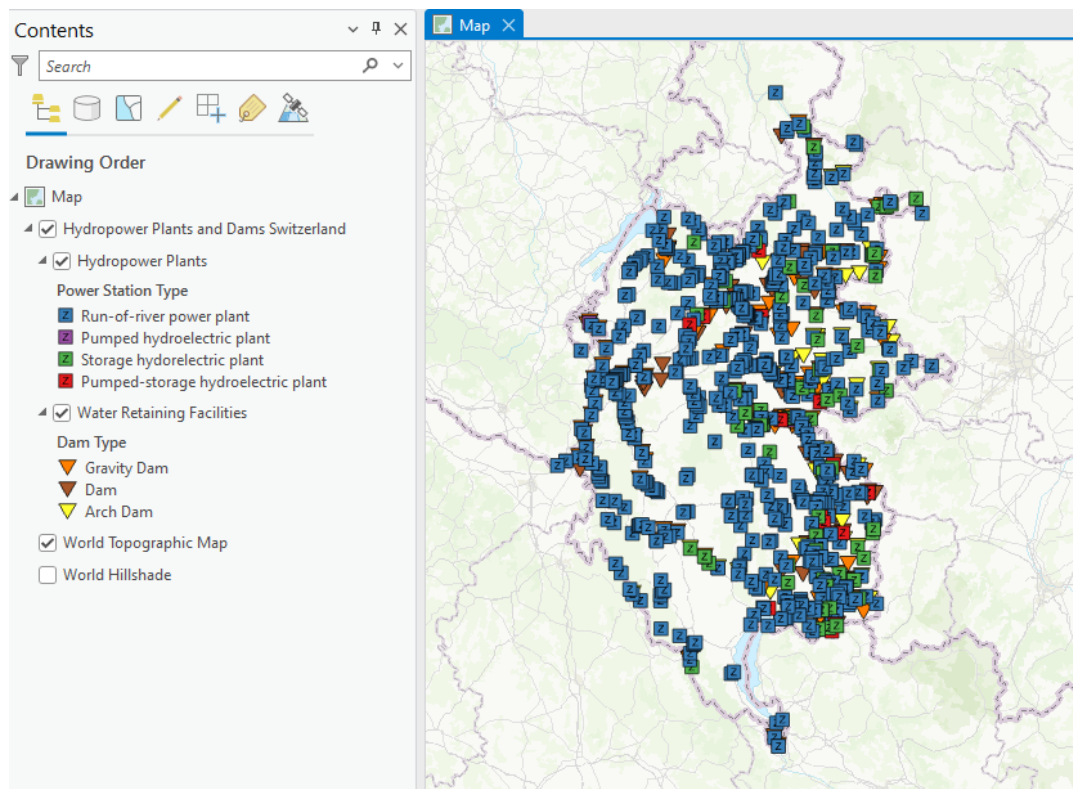
- Click on View and then Catalog in the Windows group. The Catalog will open as a pane on the right-side of the ArcGIS Pro window.



- Select Portal > Living Atlas (the last icon), and type in “hydropower” in the search bar.



- Right click on the *Hydropower Plants and Dams Switzerland* and click on Add to Current Map. You should have the power plant feature layer loaded in your Contents panel and displayed on your map.



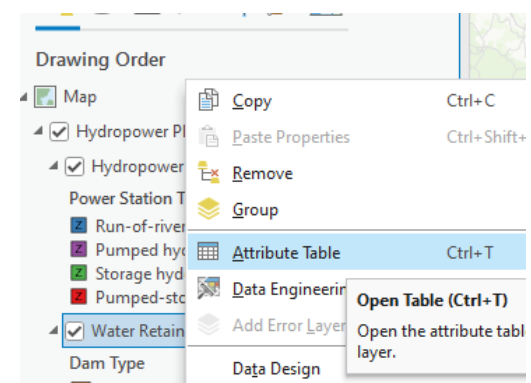
- The first part of this dataset is published by the Swiss Federal Office of Energy (SFOE), contains all hydropower facilities that are registered in the Swiss system for guarantees of origin in December 2022. This portion of the dataset includes all plants with an output capacity greater than 300 kW. There are 48.4% where power is produced by run-in river power, 47.4% storage power plants, and 4.2% pumped storage power plans. Two-thirds of the plants originate in the alpine regions of Switzerland.
- The second portion of this dataset is published by the Federal Water-retaining facilities department, and it includes all facilities that are registered in the Swiss system for guarantees of origin from November 2023. The facilities include installations designed to dam or store water or mud.
- More details for both datasets can be found here:
<https://www.arcgis.com/home/item.html?id=578f60aca19f49a68b5e59913698f096>

STEP 3: Visualize the point data

- You should be able to see the point data showing hydropower plants and water retaining facilities in Switzerland on the map.
- However, this does not tell you other information about the hydropower plants such as the power station type, address, turbine maximum, expected production, and the beginning of operation date.
 - Or the water retaining facilities such as dam type, facility aim, dam height, storage level, construction date or location.
- To access such information, you need to check the data attributes.

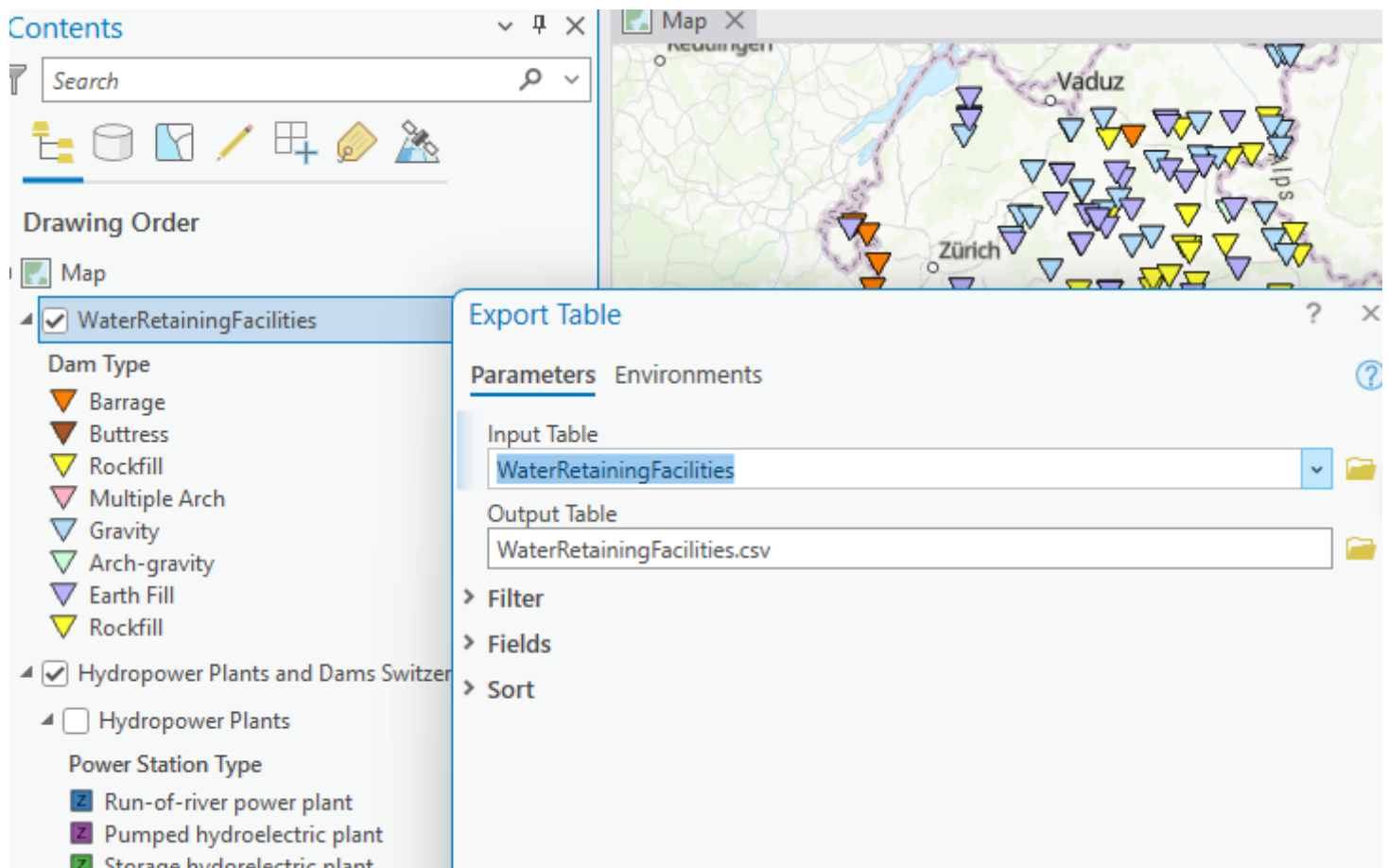
STEP 4: Check the data attributes

- In the Contents pane, right-click Water Retaining Facilities and click 'Attribute Table' in the pop-up.
- Take a look at the attribute table and scroll down to see how many rows/observations there are.
- Scroll to the right to check what fields are recorded.



STEP 5: Export the feature layer as a .csv file

- In Part 3 of the assignment, we will use RStudio to calculate central tendencies and dispersions.
- Therefore, we are going to export this attribute table to a format that R Studio can work with.
- Before we export the data, let's calculate the geometries such as x-coordinate and y-coordinate. Open a geoprocessing tool called Calculate Geometry Attributes (View > Geoprocessing > Search Bar). Select Water Retaining Facilities as the input features. Under the Geometry Attributes, add a new field easting_m with the property of Point x-coordinate, and add a new field northing_m with the property of Point y-coordinate. Keep the coordinate system the same as your input map. (If this tool fails due to not being editable, you may export the feature first, then run the tool on the exported feature.) Note: northing and eastings are in metres.
- Right click on the feature layer and select Data > Export Table. An Export Table window should pop up. For the Output Table, browse to a local directory where you want to save the csv file or to your medusa folder \\medusa\StudentWork\Your UTORid\GGR276\Lab1. Then type in a meaningful file name such as *WaterRetainingFacilities.csv*.
- Notice you may get a warning that some features were skipped in calculation. If you check the easting and northing of these features, the values are 0. Make a note of this because it is important for descriptive statistics later.
- Notice – units are listed in the ArcPro attribute table for Dam height, Crest Length, Impoundment Volume, Impoundment Level, and storage level. These units will be needed when discussing these variables. Remember descriptive statistics and measures of dispersion will have units if the variables have units.
- Data Cleaning: DamType and FacilityAim when exported to csv become codes (e.g. dt1 or fa1). In excel these can be replaced using the Find and Select option in excel. Replace the codes with their names, compare your exported dataset to the attribute table in ArcPro using Reservoir name to determine what the codes mean. E.g Aarberg – DamType is dt1 = 'Barrage' and FacilityAim is fa2 = 'Hydroelectricity'.



In the following exercise, we will inspect the following variables:

DamType: the primary dam type (barrage, rockfill dam, arch dam, gravity dam, etc.).

FacilityAim: the primary purpose of the dam (hydroelectricity, flood control, recreation, etc.).

DamHeight: height of dam from the top point to the deepest point (unit: meter, m).

ImpoundmentVolume: total volume of the dam (unit: cubic meter, m³)

ImpoundmentLevel: level used to determine the height of the dam (unit: masl, meters above sea level).

StorageLevel: The height associated with the storage volume and dammed by the barrier (unit: meter, m).

DamName: Facility Name.

easting_m: east-west position on the Earth in the projected coordinate system (unit: m)

northing_m: north-south position on the Earth in the projected coordinate system (unit: m)

Open the exported csv file with Excel, examine the data and answer **question 1-3**.

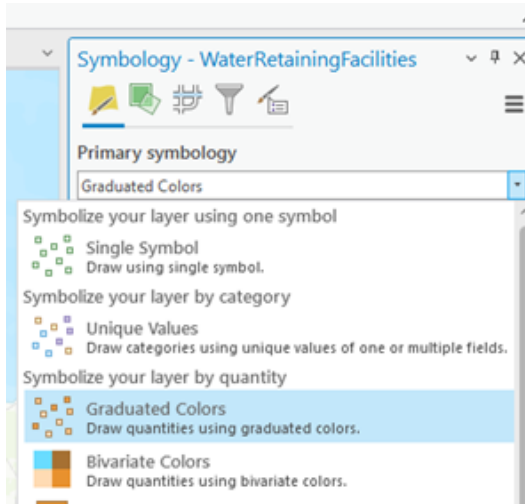
PART 1. Questions (Total Marks: 8)

- 1) Please identify the level of measurement of FacilityAim, StorageLevel, DamType, and easting_m in the .csv dataset you exported. (4 marks)
- 2) Which of the following variables are qualitative data? Multiple options are possible. (1 mark)
 - A. DamType
 - B. FacilityAim
 - C. DamHeight
 - D. Northing_m
 - E. StorageLevel
- 3) Please classify the Dam Height (m) into **5 categories** according to **quantile classification**. Recall that quantile classes contain an equal number of features. Round your final answers whole numbers. (3 marks)

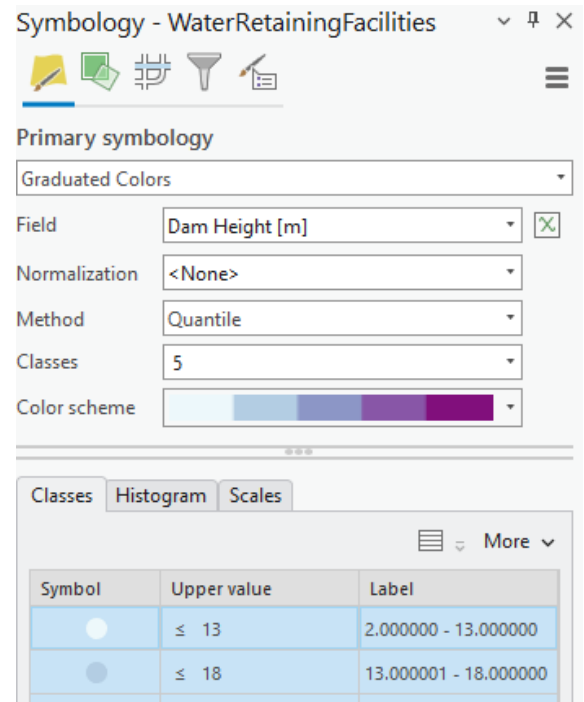
In Part 1, we classified Dam Height (m) into 5 classes according to quantile. In this part, let us experiment with classification methods and number of classes in ArcGIS Pro.

STEP 1: Classification according to ImpoundmentLevel (masl):

- Under Contents, right-click on *WaterRetainingFacilities* and click Symbology.



- Symbology will open as a pane on the right-side of the ArcGIS Pro window.
 - Primary symbology: Graduated colors**
 - Field: ImpoundmentLevel[mamsl]**
 - Normalizaton: <None>
 - Method: (choose the best method for the data)**
 - Classes: (choose the best number for the data)**
 - Color scheme: (your choice)**



PART 2 Questions (Total Marks: 12)

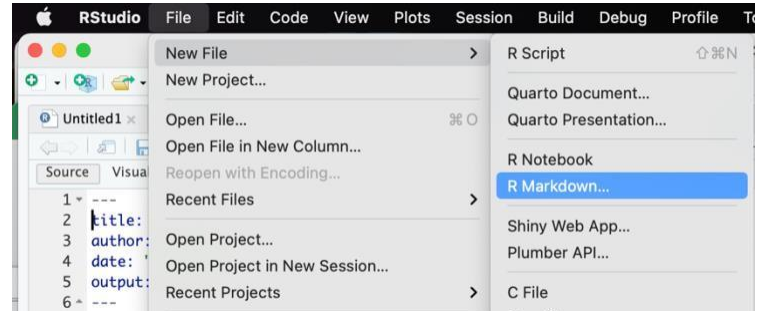
- What classification method and number of classes did you use? Justify your choice. Is the classification method used in question 3 for Dam Height (m) appropriate? Why? (6 marks)
- Examine your classified map “Impoundment Level” and describe the spatial pattern of your dataset (use location names and the hillshade to discuss features). What are some advantages of spatial data compare to non-spatial data? (6 marks)

PART 3.A: Non-spatial statistics

Next, we will start working with R (RStudio), a powerful programming language and associated user interface program. This assignment will explore the central tendency and dispersion measures using spatial data.




STEP 1: Start RStudio

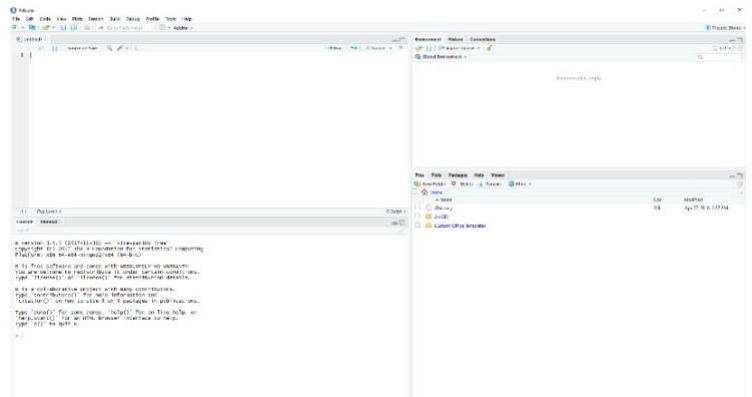
- *Start > RStudio*
- Or by searching for RStudio using Search Windows
- First open a new R Markdown Document by selecting the File > New File > R Markdown
- A new R Markdown window will appear in the top left section of RStudio
- Save your R Markdown document to your Lab 1 folder in your local directory or Studentwork folder in medusa



STEP 2: Basics in RStudio

The main RStudio interface consists of four windows.

- In the top left is the *Script* window, where you will enter most of your code to complete the assignment. You can run your scripts by highlighting the section of the code you want to run and selecting the run  **Run** button that window's top bar. Single lines of code can also be run by clicking on that line and selecting the run button. Save scripts by selecting the save  button and saving in your medusa folder.
- In the bottom left is the *Console*, where R will display any output code or errors encountered.
- In the top right is the *Environment*, where datasets (e.g., stadium_uk.csv) and variables will be displayed and can be examined at a glance.
- In the bottom right you should focus on the *Plots* window, where any output figures you create will be displayed and can be exported as pictures. Export figures by selecting the export  **Export** button and saving as an image in your local directory or medusa folder.



Your R Markdown document will also show up in the top left window. Different from the R Script, you can create HTML, PDF and MS Word documents using R Markdown. In addition, the plots will show up under the code chunks in the top left, script window, after you hit run. To begin, type the following code in the console:

```
install.packages("rmarkdown")
```

The above line of code will only install the package. You will need to type the following to load the package.

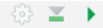
```
library(rmarkdown)
```


The codes can be embedded in the R Markdown document using R code chunks. To create an R code chunk, simply type the following in the R Markdown, and a new chunk will show up as below.

```
```${r}
```

```
```
```

```
28  ```${r}
29  ```
```



For R Markdown syntax for writing, check out page one here in this [reference](#). Here is an R Markdown [cheat sheet](#).

Once you completed all the steps below, you will be asked to render your R Markdown as an HTML file and submit both to Quercus. The HTML file should include your code chunks, visualization of the plots, and the responses to some written questions.

STEP 3: Import your *WaterRetainingFacilities.csv* data into RStudio


Next, we will import our **WaterRetainingFacilities.csv** dataset into RStudio. Note – R is case sensitive. If you get an error when loading in the file check your spelling and if you capitalized or did not capitalize any letters.

If your R Markdown is not in the same directory as your data, you will need to set the working directory (i.e., where all the data you are using is stored) first.

- It should be something like this: `\\medusa\\StudentWork\\(Your UTOR ID)\\GGR276\\Lab1` ** only if you are using medusa; if you are using your own computer then it could look like this:
`setwd("c:/users/arob1/Documents/SummerCourseInstructor")`

This working directory is just an example, you may use a different name for your folder, so use your own

name. In R the working directory is set using the `setwd()` command

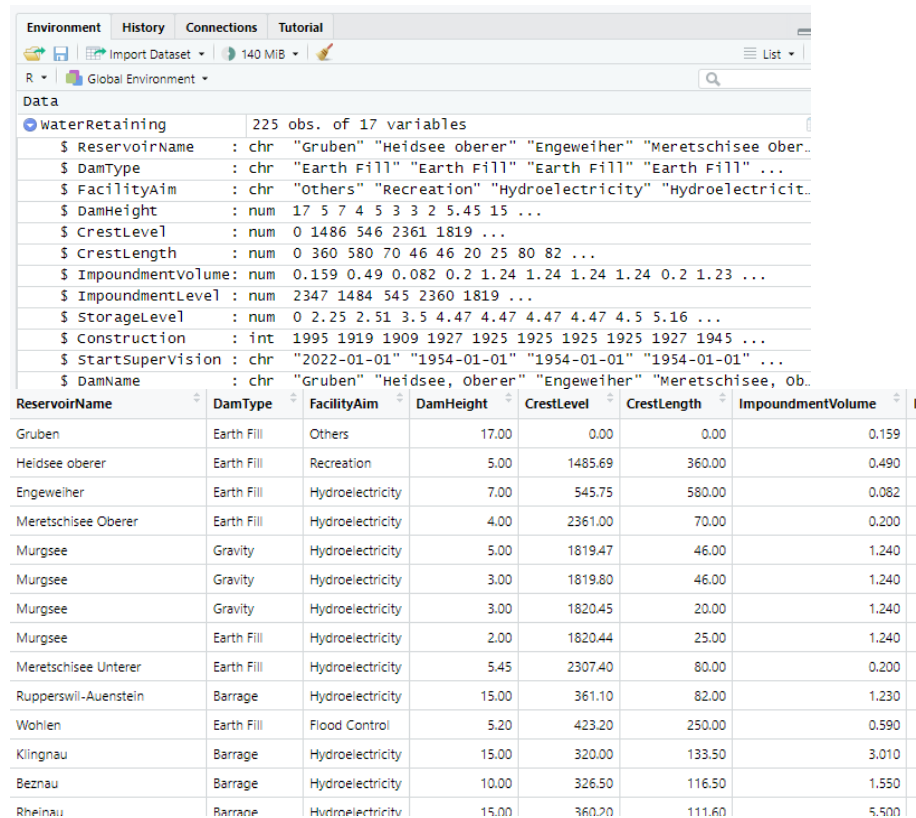
- In your *Script* window, type: `setwd("//medusa/StudentWork/(Your UTOR ID)/GGR276/Lab1")`
- Note: change the above directory to your local folder if you do not use medusa
- Run that line by either highlighting it or clicking the green triangle  to run the current chunk.
- This tells R where to look for files when you call them
- Note: remember to switch \ to / when using `setwd()`

```
> setwd("c:/users/arob1/Documents/SummerCourseInstructor")
> getwd()
[1] "c:/users/arob1/Documents/SummerCourseInstructor"
```

Next, import the csv file you have created into RStudio.

Below where you used `setwd()`, you are going to read in your csv using the `read.csv()` command. Make sure that your `read.csv()` is in the same code chunk as the `setwd()` if you used one.

- In your *Script* window type: `WaterRetaining= read.csv("WaterRetainingFacilities.csv", sep = ",", header = TRUE)`
- Think about what this line of code is doing. Remember that we have told R where to look using the `setwd()` command above.
- You should now see `WaterRetaining` in your *Environment* window showing 225 observations (i.e., Water Retaining Facilities – ie Dams) and 17 variables (e.g., `DamType`, `ImpoundmentVolume`, `easting`, `northing`). You can view this dataset in a cell-based mode by clicking it. By clicking on the little triangle in the front, you may access more information about the dataset.



| ReservoirName | DamType | FacilityAim | DamHeight | CrestLevel | CrestLength | ImpoundmentVolume |
|----------------------|------------|------------------|-----------|------------|-------------|-------------------|
| Gruben | Earth Fill | Others | 17.00 | 0.00 | 0.00 | 0.159 |
| Heidsee oberer | Earth Fill | Recreation | 5.00 | 1485.69 | 360.00 | 0.490 |
| Engeweiher | Earth Fill | Hydroelectricity | 7.00 | 545.75 | 580.00 | 0.082 |
| Meretschisee Oberer | Earth Fill | Hydroelectricity | 4.00 | 2361.00 | 70.00 | 0.200 |
| Murgsee | Gravity | Hydroelectricity | 5.00 | 1819.47 | 46.00 | 1.240 |
| Murgsee | Gravity | Hydroelectricity | 3.00 | 1819.80 | 46.00 | 1.240 |
| Murgsee | Gravity | Hydroelectricity | 3.00 | 1820.45 | 20.00 | 1.240 |
| Murgsee | Earth Fill | Hydroelectricity | 2.00 | 1820.44 | 25.00 | 1.240 |
| Meretschisee Unterer | Earth Fill | Hydroelectricity | 5.45 | 2307.40 | 80.00 | 0.200 |
| Rupperswil-Auenstein | Barrage | Hydroelectricity | 15.00 | 361.10 | 82.00 | 1.230 |
| Wohlen | Earth Fill | Flood Control | 5.20 | 423.20 | 250.00 | 0.590 |
| Klingnau | Barrage | Hydroelectricity | 15.00 | 320.00 | 133.50 | 3.010 |
| Bezau | Barrage | Hydroelectricity | 10.00 | 326.50 | 116.50 | 1.550 |
| Rheinau | Barrage | Hydroelectricity | 15.00 | 360.70 | 111.60 | 5.500 |

*The remainder of this assignment outlines an approach, using RStudio, and R's plotting capabilities, for carrying out the spatial description of the Switzerland Water Retaining Facilities dataset. All of the instructions are provided in the **GGR276Lab1P3_Starter_2024SumFT.Rmd**.

You will estimate and interpret the mean and weighted mean centre of the point distribution of Switzerland Water Retaining Facilities dataset. When the data is plotted the water retaining facility locations will be shown by the easting-northing coordinates in meters. ***You may have to make the necessary adjustments to the scripts to reflect your current work environment (e.g., location of files) and to use right case letters.**

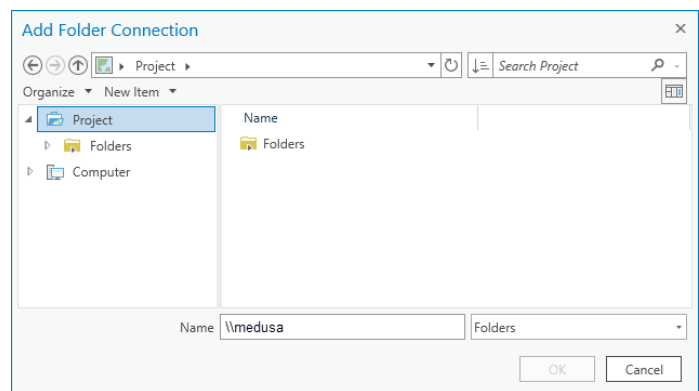
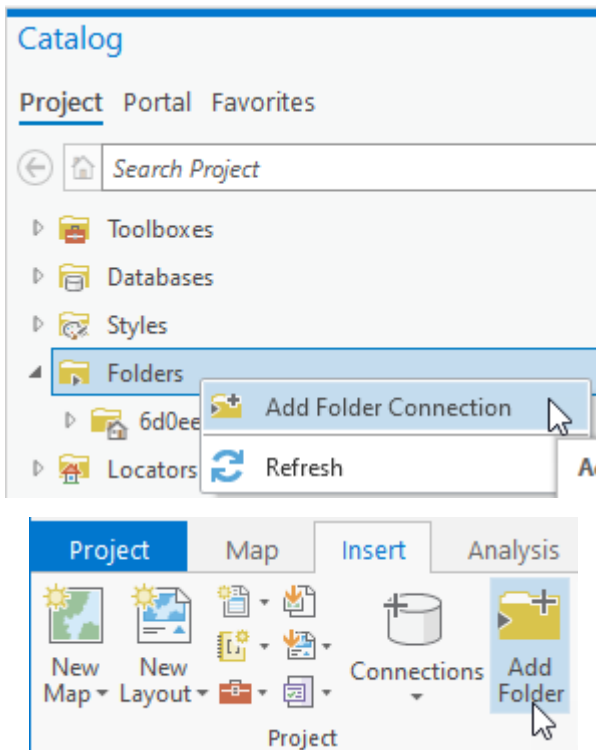
**** End of Instructions ****

Please upload Part 1&2 as a word document and Part 3 as an R Markdown document and an HTML file generated from R Markdown to Quercus. Make sure that you click submit once you complete the assignment.

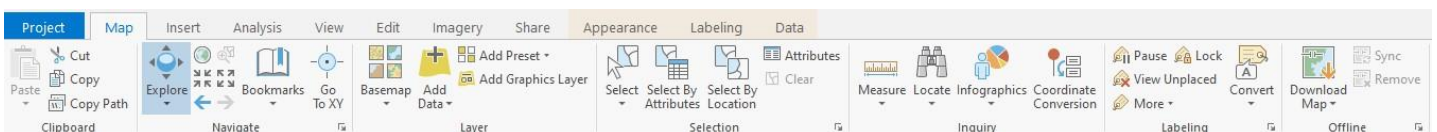
Appendix:

Appendix 1: fundamentals about ArcGIS Pro

ArcGIS Pro utilizes a sub-program called Catalog to manage the files, tools, and connections to data sources. The user has to create connections to the folder(s) where data are stored. In the case of using the web-based platform, the data for the assignment are stored on Medusa. In Catalog, click the arrow beside **Folders** to see a list of available connections. To make a new connection, right-click on the **Folders > Add Folder Connection** OR **Insert tab > Add Folder**. In the window that appears, enter `\\medusa` into the address bar and connect to the *Courses* and the *StudentWork/[YourName]* folders.

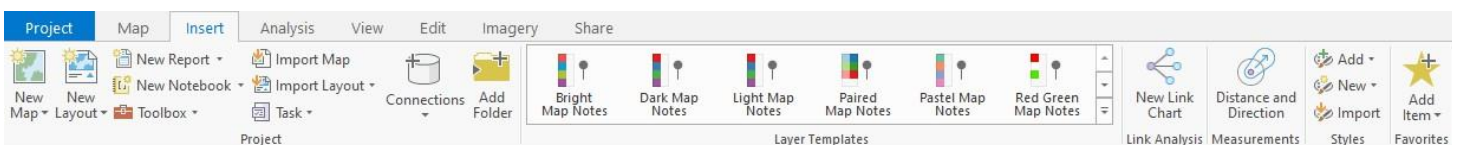


The following image highlights the commonly used tools in the main window. Please explore the program and learn how the interface is organized. Please see the list below.



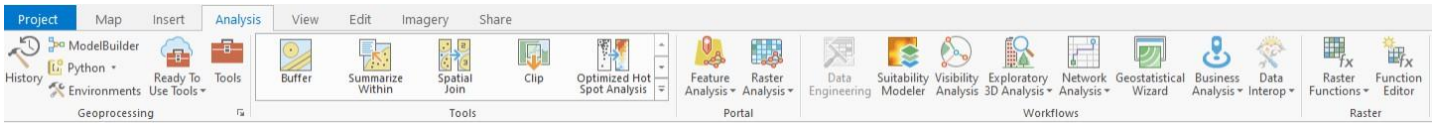
Map tab

- Navigate: tools to navigate the map
- Layer: tools to change the basemap or add data
- Selection: tools to interact with data by selecting specific areas of a map or Attribute Table



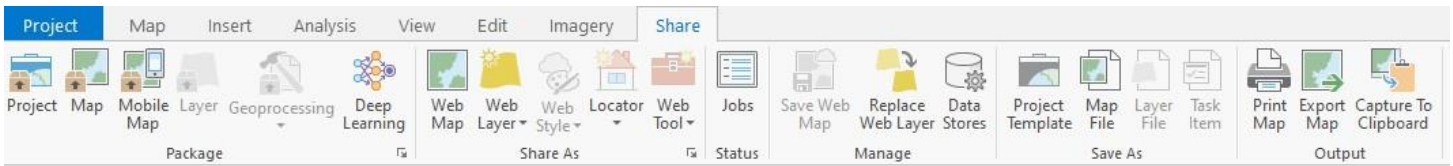
Insert tab

- Project: tools to create new maps, layouts or connections



Analysis tab

- Geoprocessing: access data analysis tools, model builder, environments and tool history
- Tools: quick access tools from the “Tools” menu

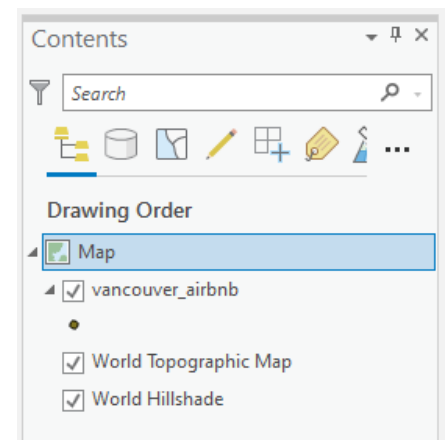


Share

- Save As: save your ArcGIS Pro document
- Output: export, print or copy your maps

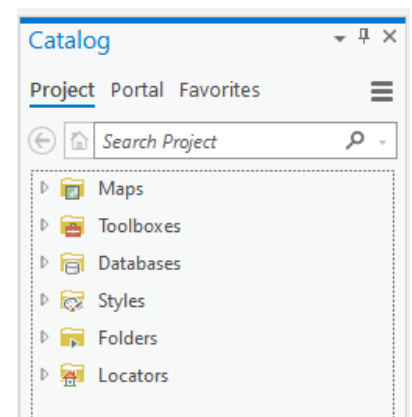
Contents

- Shows data layers
- Layers at the top are placed on top of layers underneath
- Check boxes will turn layers on and off



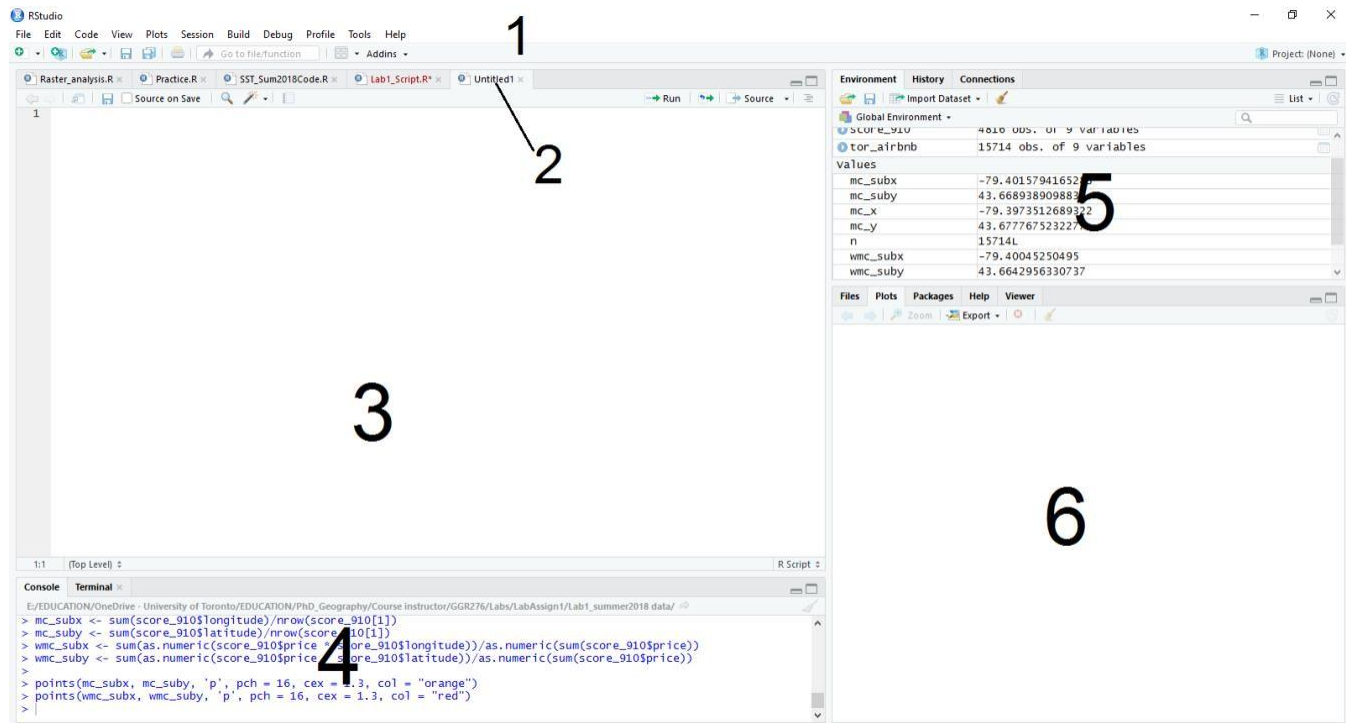
Catalog

- Access data stored in folders on your computer or cloud locations



Appendix 2: fundamentals about R programming language

R is one of the most widely used programming languages and free software environment for statistical computing and graphical visualization. The language is supported by the R Foundation for Statistical Computing. The R language is widely used in statistics and data mining by statistical software developers, data scientists, and academic researchers in many fields. R is a standalone language that allow user to write and implement codes. R can also be written and implemented in RStudio, a free and open-source integrated development environment for R programming language. The interface of RStudio is shown in the following figure.



1. This section contains tools for the user to create new or open existing project, scripts, to edit and run the R code, run debug, install existing R packages, or provide detailed documents about R programming language and RStudio.
2. The bar which shows the current opened R script files.
3. The place where we write and implement R scripts. The codes written here can be saved into a folder in our personal computer (or medusa) that can be reopened and re-run later. This is a good option if we want to run our code lines multiple times. **We recommend using it on the assignment** since you may need to run the codes multiple times.
4. Is the R console where lines of code can be written and run. However, the code lines typed here are not saved and we need to type the codes again every time we want to run it. We do not recommend using it in the assignment.
5. The window that shows the created variables and the code implementation history.
6. The window that shows the folders and files, plots, existing embedded R libraries, and detailed resources.