Lab Report

Title: Lab01-ETL Data Pipelines
Notice: Dr. Bryan Runck
Author: Megan Marsolek
Date: 10/3/2021

Project Repository: https://github.com/mmarsole/GIS5571
Google Drive Link: NA
Time Spent: 9 hours ?

Abstract

For this lab, I extracted data from three separate websites via python code (a.k.a. web scraped data). Each website required different extraction methods, for example, extracting shapefiles from MN Geospatial Commons involved using CKAN and requests package, while data from Google Places required an API key (needed to make an account to attain the key). I found NDAWN the easiest to understand and extract data from amongst all three websites.

After extracting data, I then spatially joined the two shapefiles from MN Geospatial Commons using python code. I corroborated the results by spatially joining the same data within ArcPro GUI. I found it easier to spatially join the data via open-sourced packages instead of Esri's ArcPy, since I am still unfamiliar with its functions and syntax.

Problem Statement

Acquiring data from three different websites (NDAWN, MN Geospatial Commons, and Google Places) by coding with Python ETLS to extract the data directly from the web. You'll notice based on Table 1 that in all I extracted four separate datasets, two shapefiles from MN Geospatial commons, and html based data that was scraped directly from the website. From NDAWN I extracted data within a table and converted it to a csv, while from Google I extracted 'place details' that was then stored as a dictionary.

Table 1. Data Scraped via ETL Pipeline

#	Requirement	Defined As	(Spatial) Data	Attribute Data	Dataset	Preparation
1	**MNDNR Native Plant Communities**	Shapefiles	Polygon geometry	NPC_descri (gives the plant description for each observation)	Mn GeoSpatial Commons	Import 'requests' package
2	**State Parks, Recreation Areas, and Waysides**	Shapefiles	Polygon geometry	UNIT_NAME (name of park), UNIT_TYPE (classificatio	Mn GeoSpatial Commons	Import 'requests' package

				n for type of public area)		
3	**ND Current Weather observations**	A table from a html website (converted to csv)	None (each station name did indicate what city within ND each measurement was from)	Station, Data of Acquisition, air temp, Wind direction, Current and Peak Wind Speed, Relative Humidity	From: NDAWN	Import 'requests' and 'BeautifulSoup' package
4	**Bockley Art Gallery google reviews and ratings ('place details')**	Information pulled form google maps (loads as a dictionary using 'requests' package)	Written address	Ratings, reviews, address, phone number, etc.	From: Google Maps	Created a Google places account to access an API key

Input Data

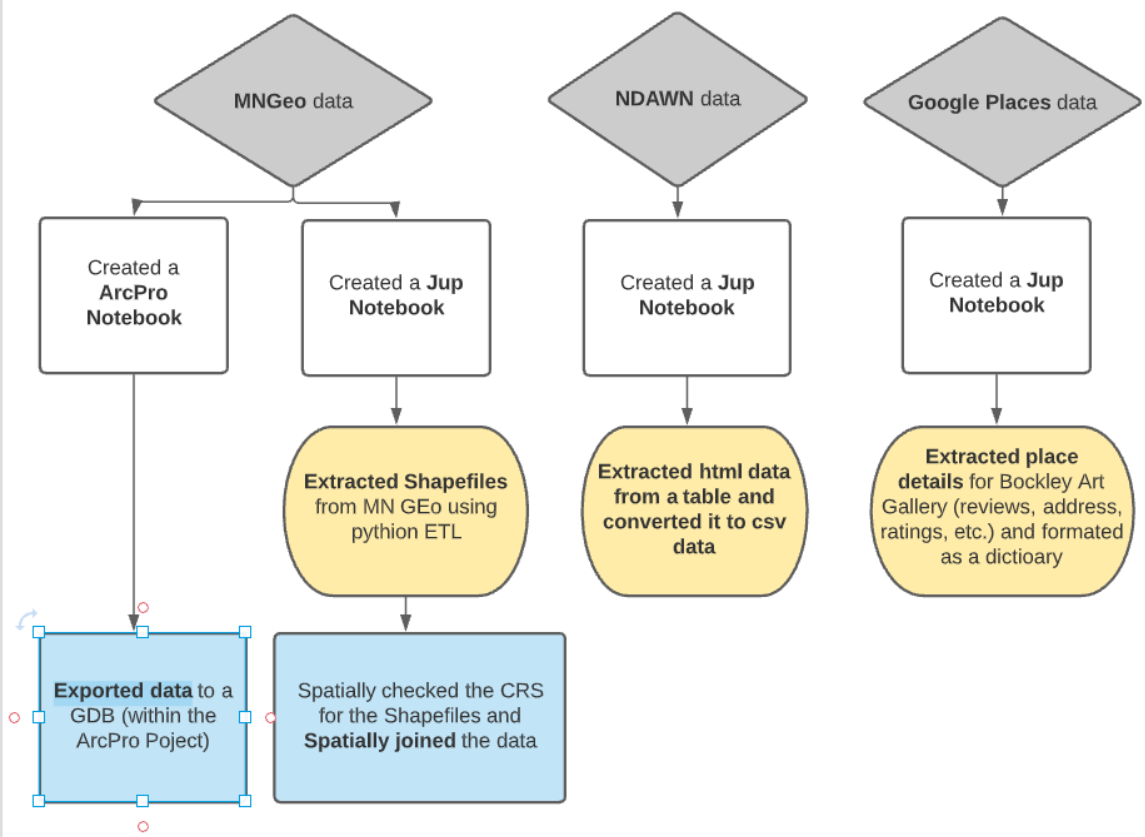
For this assignment we also had to create a spatial join and then export the data to a GDB (GeoDataBase). Using the Shapefiles from MN Geospatial Commons, I spatially joined MN State Parks to MN Plant Communities. The intention was to produce a subsetted version of the Plant Communities Data that only retained Plant observations that occurred within MN State Parks and indicated what park they belonged to with a new attribute column (UNIT-NAME). Please see Table 2 for further details about the data used and acquired.

Table 2. Data used in Spatial Join

#	Title	Purpose in Analysis	Link to Source
1	MN Parks	An additional attribute added to the Plant Communities data that will tell us within what park each plant observation occurred	Mn GeoSpatial Commons
2	MN Plant Communities	Subsetted to include only plant observations within MN State Parks	Mn GeoSpatial Commons

Methods

Figure 1: Python Extraction



Results

Image 1: State Wide view of Spatially Joined data and a Zoomed in view to the right

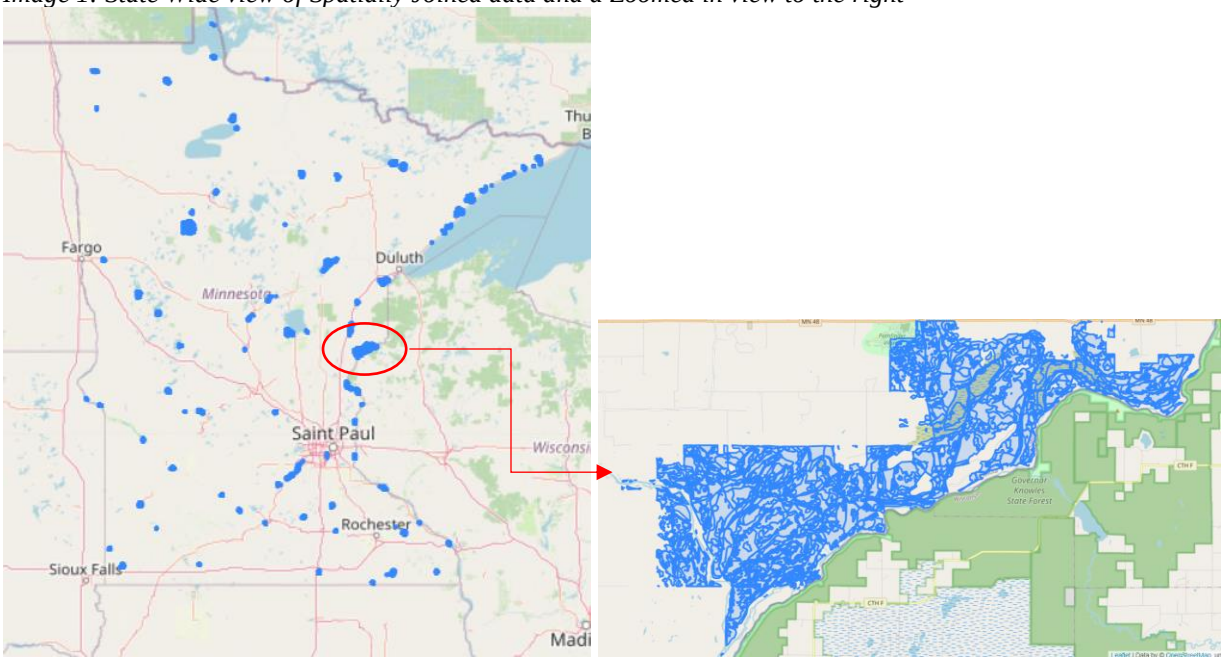


Image 2: Initial inspection of csv converted NDAWN data

	Stations	Date of acquisistion	Air temp	Wind Direction	Current Wind Speed	Peak Wind Gust	Relative Humidity
0	Ada 1N	02 Oct 11:05 CDT	59°	NNW	9 mph	4 mph	88 %
1	Adams 5N	02 Oct 11:05 CDT	62°	NNE	6 mph	4 mph	62 %
2	Alamo 2S	02 Oct 11:05 CDT	60°	NNW	11 mph	8 mph	50 %
3	Alexander 7SW	02 Oct 11:05 CDT	59°	NNW	10 mph	8 mph	49 %
4	Alvarado 4N	02 Oct 11:05 CDT	59°	NNW	10 mph	8 mph	86 %
...	...	...	...	...	...	...	...
163	Williston 5SW	02 Oct 11:05 CDT	62°	ESE	4 mph	2 mph	47 %
164	Wishek 5W	02 Oct 11:05 CDT	62°	NW	9 mph	7 mph	69 %
165	Wolford 4E	02 Oct 11:05 CDT	62°	WSW	5 mph	3 mph	73 %
166	Wolverton 2E	02 Oct 11:05 CDT	61°	N	12 mph	8 mph	87 %
167	Zeeland 7NE	02 Oct 11:05 CDT	63°	NNW	10 mph	7 mph	72 %

Image 3: Inspection of the Places details for Bockley Art Gallery (formatted as a dictionary)


```
{
   "html_attributions" : [],
   "result" : {
      "formatted_phone_number" : "(612) 377-4669",
      "name" : "Bockley Gallery",
      "rating" : 4.2,
      "reviews" : [
         {
            "author_name" : "Brian Moe",
            "author_url" : "https://www.google.com/maps/contrib/115278262360979468986/reviews",
            "language" : "en",
            "profile_photo_url" : "https://lh3.googleusercontent.com/a-/AOh14GhbShWMJofLFe61tzxTw86DGadLQzPfHjB8_y4DBCU=s128-c0x00000000-cc-rp-mo-ba3",
            "rating" : 5,
            "relative_time_description" : "8 months ago",
            "text" : "High-quality diverse curation. Todd, the owner, is great and is always there to talk about the artwork. My gallery of choice in Minnesota.",
            "time" : 1611327435
         },
         {
            "author_name" : "Karin Erickson",
            "author_url" : "https://www.google.com/maps/contrib/109393339186196940960/reviews",
            "language" : "en",
            "profile_photo_url" : "https://lh3.googleusercontent.com/a/AATXAJzzTqx-fI_ZkS3TNEXkxWza3Dn3Mp7Y3bWx_Hmo=s128-c0x00000000-cc-rp-mo",
            "rating" : 1,
            "relative_time_description" : "2 years ago",
            "text" : "I had a painting I left there on consignment for 18 months...a George Morrison. It did not sell. I called several times and left a message that I wanted to r
etrieve it..  I figured that if it was not going to sell we might as well just enjoy it. Todd was never available nor did he call back. Finally his assistant, Emily called and I t
old her what I wanted and she said she had it and I said I was coming to get it.\nWhen I got there Todd called and said I could not take it. I showed all my ownership and consignm
ent papers to Emily and I did take it but not before Todd arrived and screamed that I \"could not do this to him\".\nI left, but felt threatened by Todd and quite fearful. I would
urge people to find a gallery with a more sane owner for their consignments.",
            "time" : 1558038859
```


In all, I managed to extract and save a csv version of some of the NDAWN weather data, store the reviews for Bockley Art Gallery, extract shapefiles via python ‘requests’, spatially join the shapefiles, and save the data to a GDB.

Results Verification

I performed the spatial join within Jupyter Notebooks using ‘.sjoin()’ function, and visually inspected the output (using folium maps), I further performed the Spatial Join within ArcPro GUI and found visual results upon inspection.

For verification on the extraction of data I compared the output for the NDAWN data to what was present on the page, and via this cursory glance I found no obvious errors (page updated every five minutes, which maked it hard to do more than a cursory glance). The same visual comparison was performed on the data I extracted about Bockley Art Gallery (I looked at the Google Maps to compare reviews, and other extracted data).

Discussion and Conclusion

Overall, I found the most challenging part of this assignment trying to use Arcpy functions to spatially join data. I am not familiar with this package's format, syntax, or its function names, and have trouble using it when I am more familiar with other open-sourced packages that can perform the same functions (all except export data to a GBD, my research found that since this is an Esri proprietary product I cannot save my shapefiles using open source means).

Otherwise, the next biggest challenge was learning to extract data from any web source via coded ETLs. This is the first time I have ever built or coded such commands, and I am grateful for open-sourced tutorials that guided me through extracting from open domain websites (Goggle was trickier since you have to use an API key). I still found it very intimidating, I don't know JavaScript, inspecting a webpage takes time to locate my desired data, and CKAN still feels awkward, given I don't know its other functions or attributes.

I think with time I might get better at ETL Pipeline coding, but recognize I am severely hampered by my lack of understanding and experience in JavaScript, API keys, and general extraction packages ('requests', 'BeautifulSoup', etc.)

References

Clever Programmer'. (2019, April 4). *20 - web scraping with python using beautiful soup & requests (Python tutorial for beginners 2019)*. Www.youtube.com. https://www.youtube.com/watch?v=E5cSNSeBhjw

Self-score

Fill out this rubric for yourself and include it in your lab report. The same rubric will be used to generate a grade in proportion to the points assigned in the syllabus to the assignment.

Category	Description	Points Possible	Score
Structural Elements	All elements of a lab report are included **(2 points each)**: Title, Notice: Dr. Bryan Runck, Author, Project Repository, Date, Abstract, Problem Statement, Input Data w/ tables, Methods w/ Data, Flow Diagrams, Results, Results Verification, Discussion and Conclusion, References in common format, Self-score	28	**28**
Clarity of Content	Each element above is executed at a professional level so that someone can understand the goal, data, methods, results, and their validity and implications in a 5 minute reading at a cursory-level, and in a 30 minute meeting at a deep level **(12 points)**. There is a clear connection from data to results to discussion and conclusion **(12 points)**.	24	**22**
Reproducibility	Results are completely reproducible by someone with basic GIS training. There is no ambiguity in data flow or rationale for data operations. Every step is documented and justified.	28	**24**
Verification	Results are correct in that they have been verified in comparison to some standard. The standard is clearly stated **(10 points)**, the method of comparison is clearly stated **(5 points)**, and the result of verification is clearly stated **(5 points)**.	20	**20**
		100	**94**