MIE 223 Data Science
Lab and Assignment 10:
Spatial Data Analysis

In this lab assignment, you will learn and gain hands-on experience by working with more examples on choropleths, accessing data using OpenStreetMap, and visualizing data using Folium maps.

- Programming language: Python (Google Colab Environment)
- Due Date: Posted in Syllabus

Marking scheme and requirements: You are required to use the Jupyter notebook named `Assignment_10.ipynb` to provide appropriate answers to the questions in this assignment. All changes to the `Assignment_10.ipynb` notebook should be committed and pushed to your assignment repository on GitHub.

This assignment has *4 points* in total and the point allocation is shown below:

- Solution to Questions (1 point):
  - Q1(a): 0.1 point
  - Q1(b): 0.1 point
  - Q1(c): 0.1 point
  - Q1(d): 0.1 point
  - Q1(e): 0.1 point
  - Q2(a): 0.1 point
  - Q2(b): 0.1 point
  - Q2(c): 0.2 point
  - Q2(d): 0.1 point
- Code review (1 point)
- Quiz (2 points)

What/how to submit your work:

- All your code should be included in the provided notebook named `Assignment_10.ipynb`.
- Commit and push your work to your GitHub repository in order to submit it. Your last commit and push before the assignment deadline will be considered to be your submission. You can check your repository online to make sure that all required files have actually been committed and pushed to your repository.
- A link to create a personal repository for this assignment is posted on QUERCUS.

Notes that you should pay attention to

1. All your code should be written in the provided notebook named `Assignment_10.ipynb`.
2. All functions must *return* the specified return type.
3. Commit and push your work to your GitHub repository in order to submit it. **Your last commit and push will be considered to be your submission.** You can check your repository online to make sure that all required code has actually been committed and pushed to your repository.
4. Please do NOT rename any parameter name, function name and class name given.
5. **Your code should have output for all assignment questions unless otherwise stated**. If any of the assignment questions asks you to explain your outputs, please use markdown cells to clearly explain them.
6. As a backup plan, also upload a PDF copy of your assignment notebook to GitHub. This is useful especially when GitHub cannot render some images or outputs.
7. During the code review, TAs will go over your answer submissions to assess your understanding of the concepts relating to the assignment and the logic behind your codes. Please, ensure your codes are readable and well commented using python comments and markdown cells when necessary. Not have readable code may impact your code review score.
8. Please note the plagiarism policy in the syllabus.

1 Main Assignment

Please answer the questions below in the provided notebook named `Assignment_09.ipynb` using Google Colab. **Clearly give titles and labels to all plots unless otherwise stated.**

Q1. Health Network Fairness

In Lab 10, we delved into the issue of transportation fairness in Toronto, focusing on Forward Sortation Areas (FSAs) and TTC subway stations. However, unfairness extends beyond public transportation and permeates other sectors, including the health network. This encompasses the distribution of hospitals, clinics, emergency services, and the availability of doctors. In Questions 1 and 2, we will examine the distribution of health networks across Toronto and explore the presence of unfairness in this area.

(a) To assess the fairness of a health network, the initial step involves gathering geographical data related to the distribution of health facilities. As discussed in the lab, this can be efficiently accomplished by leveraging the OpenStreetMap API. Crafting an appropriate query is essential for retrieving information effectively. Referencing the list of OpenStreetMap Keys and their respective values in the following link: https://wiki.openstreetmap.org/wiki/Map_features, and provide three Keys along with their corresponding values that could be used in the query to retrieve geographical information of the health network.

(b) In the provided code for Q1b, the GeoDataFrame **toronto_hospital** contains a list of hospital names and their vector shapes that provide emergency services in the City of Toronto, retrieved from OpenStreetMap. Additionally, **toronto_FSA** is a GeoDataFrame containing a list of vector shapes representing each Forward Sortation Area (FSA) in the city of Toronto.

First, print out the Coordinate Reference System (CRS) of the two GeoDataFrames, and change their CRS to **EPSG:4326** if their CRS are not the same. Then, concatenate the two GeoDataFrames into a single GeoDataFrame named **gdf_all**. Display the first five rows of your concatenated GeoDataFrame.

(c) Plot the FSAs and the locations of hospitals in a single choropleth, where the hospitals are represented as red dots with a **markersize** of 20. While there are no specific style requirements, ensure that all FSAs and their boundaries are clearly visible, with appropriate sizing and a descriptive title for the choropleth.

According to this choropleth, is there any visible imbalance in the distribution of hospitals? (there is no need for any calculations; simply examine it visually)

(d) Create a new column in **gdf_all** named **centroid** such that, its value for a given row is the centroid of its shape in the geometry column if and only if its shape is a Polygon or MultiPolygon, otherwise its value should be its original shape in the geometry column. Store your result of each row in a new column named **centroid** in the **gdf_all**.

Similar to what we did in the lab, compute the distance matrix of the geometric distance between all pairs of (FSA centroid, hospital) as a **DataFrame**, where each column represents a hospital and each row represents an FSA. Ensure that you clearly label the column names with the corresponding hospital names and index each row with the FSA code. Display the first five rows of your distance matrix.

(e) Generating a heatmap based on the distance matrix. Make sure you have clear labels and titles in your plot.

Can you identify any imbalance in the distances to the hospital through the heatmap? If not, what difficulties might prevent you from gaining useful insights?

Q2.Visualizing Health Network Fairness

Note: You do not need to have a correct Q1 to attempt Q2

In the lab, we utilized the **osmnx** package to obtain the bikable street maps in Toronto. For this assignment, we will analyze the drivable street map to simulate the time required for an ambulance to reach a patient's house from the nearest hospital.

(a) Using the **osmnx** package, retrieve the drivable street map of Toronto, using 'City of Toronto, Ontario, Canada' as the place name (this is case sensitive and should be used exactly as it is), and display the retrieved graph in a plot. The background color should be white, nodes in the graph should be red with a size of 1, and edges should have a linewidth of 0.5 and be colored black. There is no requirement to include a legend or title for this plot.(Hint: The **network_type** for drivable distance is 'drive')

How many edges and nodes does this graph contain?

(b) In the given code, the GeoDataFrame **gdf_smallest_distance_by_shortest_path** contains

- **name**: the FSA code of each row
- **geometry**: the vector shape of each FSA
- **centroid**: the centroid of each vector shape
- **nearest_hospital**: the distance of the shortest driving route from the centroid to the nearest hospital.

Generate a choropleth map illustrating the shortest driving distance to the nearest hospital of each Forward Sortation Area (FSA) using this updated GeoDataFrame. The colour scheme should reflect the quantiles of the data and set the **cmap** to **OrRd**. Ensure an informative title, boundaries, and a legend in your choropleth.

Could you identify any imbalance in the distribution of distances to the nearest hospital using the choropleth map?

(c) In the provided code, the GeoDataFrame **census_data** contains:

- **name**: the FSA code of each row
- **perc_visual_minority**: the percentage of visual minorities in each FSA according to the 2016 census
- **centroid**: the centroid of each vector shape

Alongside, there's an empty folium map named **nearest_hospital_map**. Please add the following layers to the folium map:

- **Tile Layer**: Utilizing **openstreetmap** as the Tile set

- **Choropleth Layer**: Displays the choropleth of the shortest drivable distance to the nearest hospital (Similar to Q2b). The **fill_color** should be set to **OrRd**, and the **bins** should be determined based on quantiles of data. A clear legend should be included.
- **Circle Marker Layer**: Incorporates a circle marker at the centroid of each FSA. The radius of each circle corresponds to the percentage of visual minorities in that FSA, multiplied by a scalar number (e.g., 20) for better visibility. These circle markers are stored in a **FeatureGroup**.

Ensure the map includes a Layer Controller to facilitate switching between each layer, and each layer (excluding the Tile Layer) should be given a proper name.

(Note: Folium maps cannot be properly displayed on GitHub. Please take a screenshot of your final Folium map and upload it to GitHub with the name **Q2c_folium_map**. For Mac users, please use the 'command + shift + F4' to take a screenshot. For Windows users, please refer to the following article.)

(d) Can you distinguish any relationship between the percentage of visual minors and the shortest driving distance to the nearest hospital?

Q3. Raster Data (OPTIONAL: THIS QUESTION DOES NOT COUNT TOWARDS THE TOTAL GRADE)

In this question, you will analyze another instance of Urban Heat Island using near-surface temperature data for 7-8 pm on July 27, 2020, within the City of Seattle. The provided code has imported the raster data as the variable **raster_data**. Please utilize it for the following questions.

(a) Plot the raster data image in **raster_data**, ensuring a clear title and including a color bar, and set the **cmap** as **coolwarm**.

What is the shape of this image file, and what is its Coordinate Reference System (CRS)?

Based on the raster map and the concept of Urban Heat Island, provide a hypothesis regarding which areas of the map could potentially represent urban areas in Seattle.

(b) Classify the raster data using the following quantile ranges: less than 25%, 25% to 90%, 90% to 100%, and greater than 100% quantile of surface temperature. Modify the class with the highest quantile (greater than 100%) to class 0 to mitigate the influence of background color. Plot the classified raster data using the colormap **coolwarm**, ensuring a clear title and color representation.

Based on the classified raster data map, which areas of the cities in the map could potentially represent urban areas in Seattle? Does this align with your previous hypothesis?

(c) Figure 1 below is a map of Seattle (sourced from Google Maps). Does the urban area shown in the map correspond with your hypothesis regarding potential urban areas in Seattle? Provide an example of a real urban location that supports your hypothesis. (If you encounter difficulties viewing the text in the map below, you can alternatively access Google Maps and search for the map of Seattle)

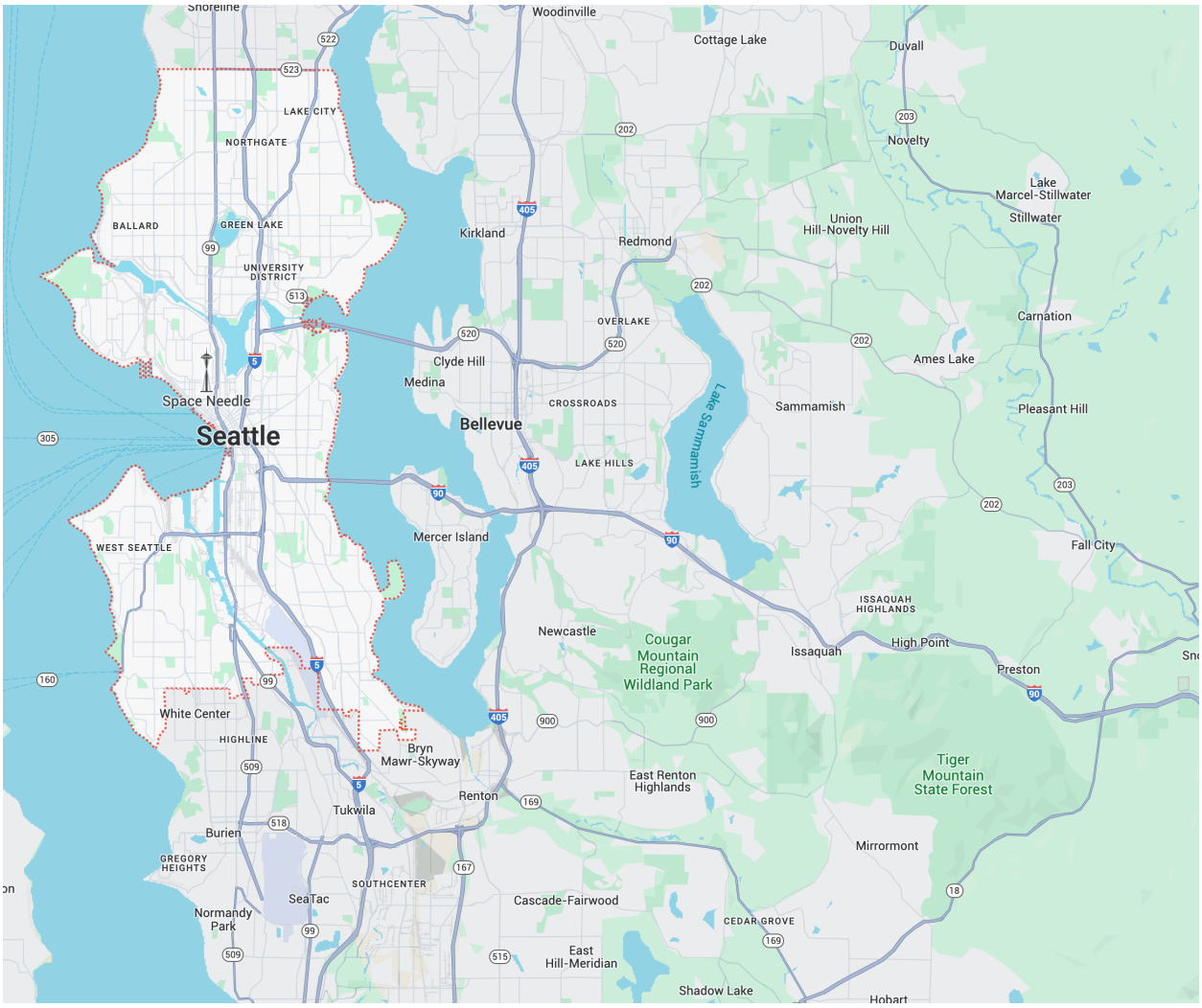

Figure 1: City of Seattle