Spatial Data Science & Engineering

Assignment 5

Maximum points Possible – 10

Tasks:

You are given a Python file Assignment5.py containing an empty function *find_hotspot* with the following parameters:

- taxi_zones: it is a GeoPandas DataFrame containing two attributes – *LocationID* and *geometry*. Each row in *taxi_zones* geodataframe represents a small part of the New York city, having a shape of polygon or multi-polygon.
- taxi_pickups: it is a Pandas DataFame containing two attributes - *pickup_longitude* and *pickup_latitude*. Each row represents a point location where a taxi pickup happened.
- output_path: a string path to write the final output.

Your task is to complete the *find_hotspot* function. The job of *find_hotspot* function is to find the hotspot taxi zones in terms of the number of taxi pickups happened inside each taxi zone. Whether a zone is hotspot or not is determined based on a statistical value called G_i^*. The value of $G_i^*$is calculated as follows:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}}$$

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2}$$

Here, X_j is the attribute value of taxi zone j. Attribute value of a taxi zone is the number of taxi pickups inside that zone. Don't consider pickups happened on the exterior boundary of a zone. n is the number of taxi zones, $W_{i,j}$ is the spatial weight between zones i and j. Spatial weight between zones i and j is 1 if zone i intersects zone j and i != j, otherwise 0. If the denominator in the equation of G_i^* becomes 0, replace the denominator with 0.0000000001.

Calculate the value of G_i^* for each taxi zone i and write the LocationID and G_i^* score of top 50 zones having the highest G_i^* score to the given *output_path*. If two taxi zones have similar G_i^* score, sort them in the ascending order of LocationID. For output format check the file *output/g_scores.txt*.

You can test your method by running tester.py. Don't import anything from tester.py to Assignment5.py, it will raise error during testing in our side. tester.py file is only for your testing purpose only.

Submission Instructions:

- Submit only Assignment5.py file.

Warning:

- Use only geopandas, pandas, and numpy libraries for your computation. Don't import any extra module in the Assignment5.py file that require installation of that module. You can import those modules which come default with Python and do not require separate installation.