Atlan | Data Science Intern Challenge

Aim: Identify commercial centers using Points of Interest (POI) data

General Terminolgy: POI Data provides location of different places with their tags like school, hospital,etc. (co-ordinates + tag). In POI Data, we have use of following terminologies which are very important to understand -

- Node
  It is single point on Map e.g. Single Restaurant, Tree, etc
- Way
  Line or Area (Space) e.g. Road, Lake, Building, etc.
- Tag
  Data about Node or Way e.g. Restaurant's name.
- Features
  Things that appears on map respresented using points, lines and areas.

Approach:

(A)

1. Get Points of Interest from open data sources like open street maps (OSM).

   I have used OSmnx (https://www.researchgate.net/publication/309738462_OSMnx_New_Methods_for_Acquiring_Constructing_Analyzing_and_Visualizing_Complex_Street_Networks) which is a python module that can be used to retrieve, construct, analyze , visualise street networks from OpenStreetMap and collect POI Data. One the most useful features that OSMnx provides is an easy-to-use way of retrieving OpenStreetMap data (using OverPass API).

2. Understanding how spatial location data works
   1. There are different Spatial Data types such as :
      i. Vector Data - It is best describes as the graphical representation of real world which is composed of points, lines, ploygons. It can be stored in shapefile fomat (.shp).
      ii. Raster Data - Imagery data that is presented in a grid of pixels/aerial or satellite taken data.
   2. Understanding necessary GIS concepts like projections, spatial clustering, etc. :

      Geographic Information System is a system designed to capture, store, manipulate, analyze, manage, and present spatial or geographic data.

While Creating Visuals there are a variety of geographic elements to consider. One of them is Projection.
- Projection basically converts the 3D Data into 2D differing on shape, size, area and accuracy.

Data Source(contd Approach):

POI (OSM) data- https://www.openstreetmap.org/#map=11/28.6518/77.2219

- I had a lot of different data sources from where we could have collected data and ==then have performed join and aggregations== on them to get the desired set of Data. eg https://data.gov.in/catalog/status-geographical-information-system-gis-pradhan-mantri-jan-dhan-yojana-pmjdy?filters%5Bfield_catalog_reference%5D=416081&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc ; https://www.earthdatascience.org/courses/earth-analytics-python/spatial-data-vector-shapefiles/intro-vector-data-python/ ; https://freegisdata.rtwilson.com/ , etc.

- But due to limited knowledge as of now, I could only use OSmnx to retrieve POI Data from OSM and use it for clustering. I ==propose== a following ==Block Diagram== which we can use (http://gallery.dataiku.com/projects/DKU_GEO_CLUS_NYC/flow/) .

- Now, the data could be collected in two ways-
    - First way is to collect the data (amenity type or tags like Water Capacity, Electricity usage, Firestation, etc) corresponding to different latitudes and longitudes of City. The intuitive fact that high electricity uses, with Water Capacity and markets nearby can be attributed to the fact that it might be a supermarket.
    - Second Way is to collect the data using OSmnx for different amenity type and apply Clustering Technique to group them to consider them as like Commercial centre, etc. Markets, Cafe, etc having higher densities in particular area can be thought of as being a Mall, SuperMarket, etc.
- I used the second Approach and used KNN, DBSCAN, Clustering algorithms.
- Please find the associated labels in the .csv file after running the GeoVERSION.ipynb.