Brandon Lasher
CS 522
11/2/2022

Zip codes and Clustering via DBScan

Project Description

Use a zip code database to predict the location of major cities base on the density of the zipcodes within a given area.

Approach

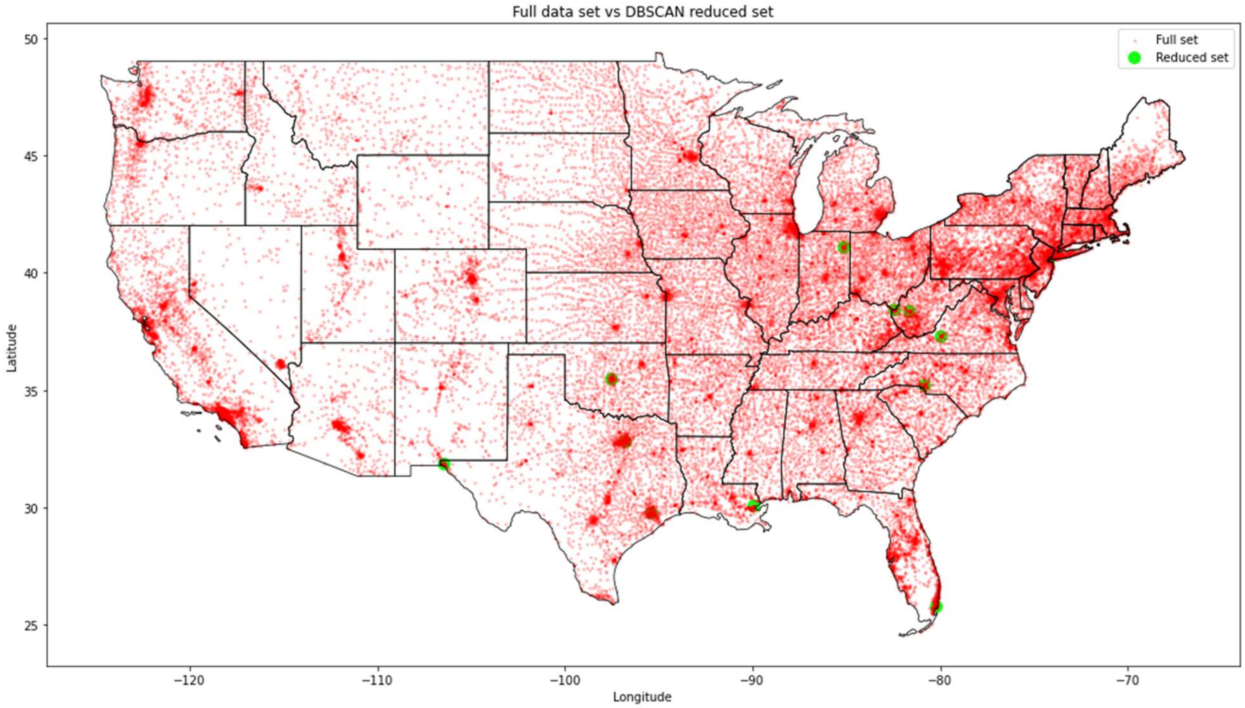
For this project I used the SK-learn's DBScan method for clustering and a zip code database found at https://www.unitedstateszipcodes.org/zip-code-database/. I also used a .shp file of the lower 48 states and the geopandas library to add a background and allow for a better visualization.

In order to do a DBScan, we first need to setup the parameters for the distance matrix generation. Because this uses Latitude and Longitude, I decided to use great circle distance to account for the curvature of the earth between the points. We must also define the requirements to be a cluster. There are two factors which come in to play. The maximum distance to the next point and the number of points found in that locality. These both can play a key role is removing noise from the dataset.

I tried a several different variations of these parameters and the results are shown below.

Test case 0

Max distance between 2 points: .5 km
Min points in a cluster: 25
Found: 12 clusters

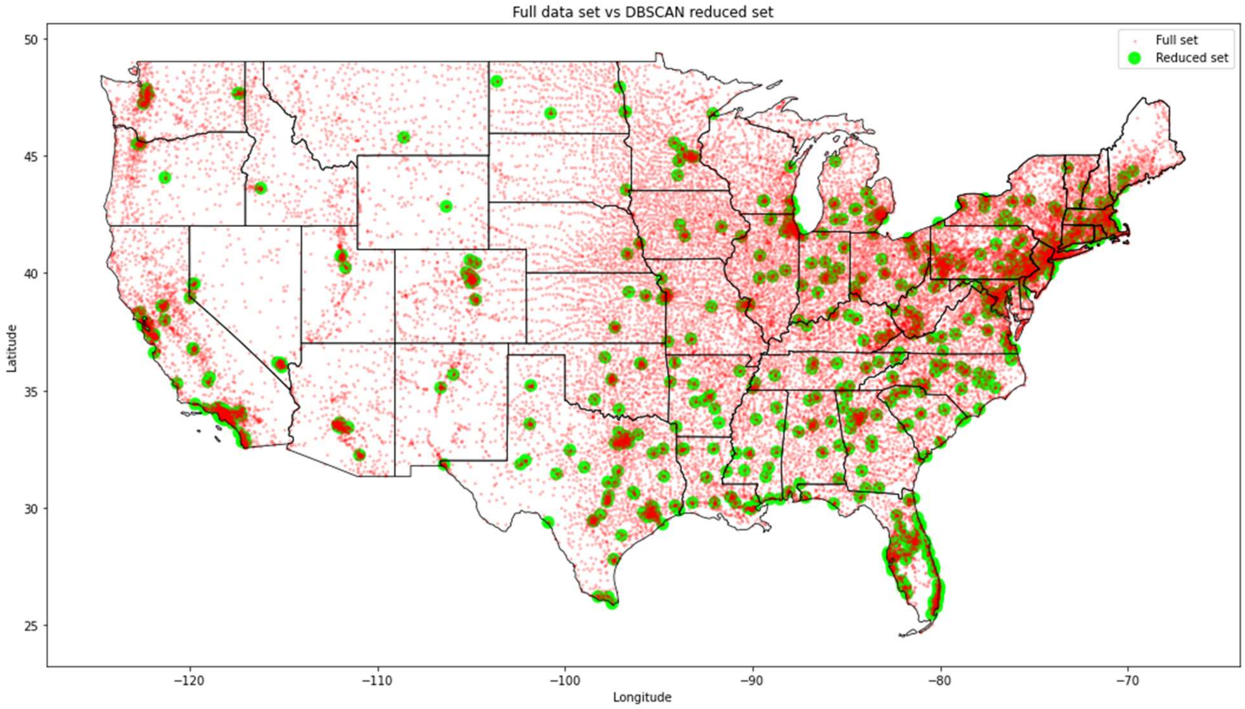


Latitude	Longitude	City
40.71	-73.99	New York City, NY
37.27	-79.95	Roanoke, VA
38.35	-81.63	Charleston, WV
38.41	-82.43	Huntington, WV
35.19	-80.83	Charlotte, NC
25.77	-80.20	Miami, FL
41.07	-85.13	Fort Wayne, IN
30.06	-89.93	New Orleans, LA
35.46	-97.51	Oklahoma City, OK
32.79	-96.76	Dallas, TX
29.76	-95.38	Houston, TX
31.84	-106.43	El Paso, TX

The parameters for this test case are chosen such that a city would need have lots, 25, of zip code points near one another, .5 km in order to create a cluster. The main issue with this approach is that there may be large cities that due to geographical reasons do not have zip code points that close to eachother.

Test case 1

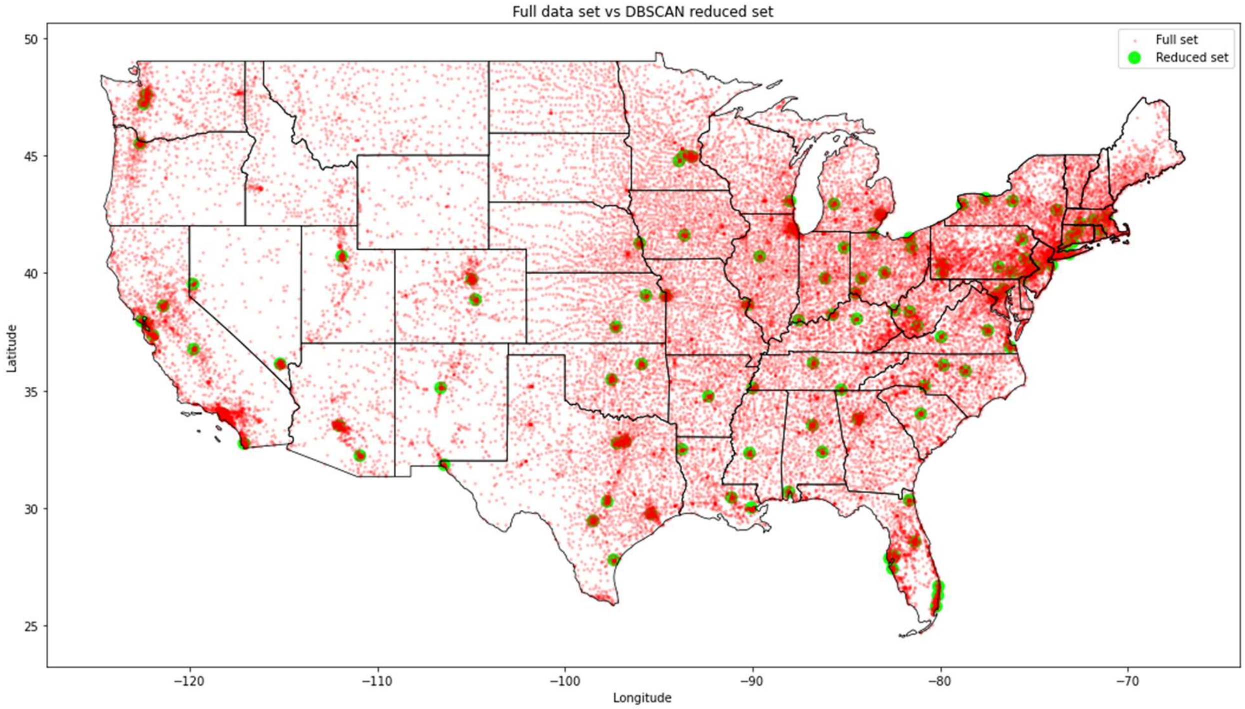
Max distance between 2 points: 1.5 km
Min points in a cluster: 3
Found: 635 clusters



This approach too few zip codes in a cluster, 3, causes many places to be classified as a city. With this it would be hard to isolate the largest of the cities.

Test case 2

Max distance between 2 points: 10 km
Min points in a cluster: 20
Found: 110 clusters



This parameter set increases the max distance between the zip code points and keep a fairly large amount of cluster points. Though it still picked out 110 cities, I found this approach to the parameters to appear to be the most effective.

Conclusion

As demonstrated by the previous figures this type of clustered is extremely sensitive to the parameters which define a cluster. As they become more strenuous, (e.g smaller radius and more points required) we can find the larger of the likely cities based on the zip code density.

In testcase 0 some very large cities in the US were excluded from the list and it is likely that these tight parameters cannot account for geographic barriers in the distance between zipcode points. It is because of factors such as this, I would suggest that this method be used for a general purpose as to find many large cities and not the “largest”.

Additional information

See the project code at https://github.com/blasher565/dm_cs522/blob/main/zipcodes/zipcodes.ipynb