KOLLAM TO VELLORE RELOCATION USING K-MEANS CLUSTERING

by Shaun Oommen Alexander

INTRODUCTION/ BUSINESS PROBLEM

This project is based on a problem that my friend faced in college. I currently study at Vellore Institute of Technology, Vellore as a 4th year Engineering student. My friend and I both moved from Kerala, India to Tamil Nadu, India for our higher studies. While moving to a new location it can be really tough to find out the right location on the basis of accessibility, distance, services etc. With this project I am looking to help my friend who is based in Kollam, Kerala to find the right location to shift to in Vellore, Tamil Nadu.

The reason I consider this to be a tough problem is because Vellore is quite underdeveloped and it can be hard to find a place to live, while Kollam is quite developed with lots of different venues and locations of different types. So for someone who is moving from an area like Kollam to VIT Vellore, this project will try to find similar locations in Vellore to locations in Kollam using clustering. Here the k means clustering algorithm is used to achieve the task. Folium library can be used to visualize the clusters in both cities. Here the k means clustering algorithm is used to achieve the task. Folium library can be used to visualize the clusters in both cities The major Target Audience for this project will be those looking to relocate from one city to another. In this project we are exploring the possibility of moving from Kollam to Vellore but this idea can be utilized in other similar cases as well.

DATA DESCRIPTION

For this project I will first leverage data from the web to find out the neighbourhoods in Kollam and Vellore using BeautifulSoup for web scraping.

https://en.wikipedia.org/wiki/List_of_areas_of_Vellore

https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Kollam

I also used the FourSquare API to obtain all the venues within 2000 metres of the location. Using this venue data we will be able to find locations in vellore with a similar profile. In this case I will not be filtering down the venues based on the number of venues returned since it is important that people looking for places with high accessibility and low accessibility both find their area of interest. The following data are obtained from the Foursquare API:

1. Venue
2. Venue Latitude
3. Venue Longitude
4. Venue Category Data

Initially we obtained a total of 81 different categories but there were a lot of categories that were similar and could be combined like Fast Food Restaurant and Fried Chicken Restaurant. This took some hands on work but I reduced the categories down to 70 by combining those that were really similar.

METHODOLOGY

After obtaining the list of neighbourhoods from the links and obtaining the venues near the neighbourhoods using the FourSquare API. After this I one hot encoded the data. Then used K-Means algorithm to cluster the data with number of clusters set to 7. The clusters obtained can be analyzed to find which locations are similar between Kollam and Vellore. After obtaining the results we can sort the them based on the distance to VIT university and select the best options.

ANALYSIS

In this section we will be analyzing the results of K Means clustering with the help of folium maps to visualize the various locations and using colours to denote the cluster labels.

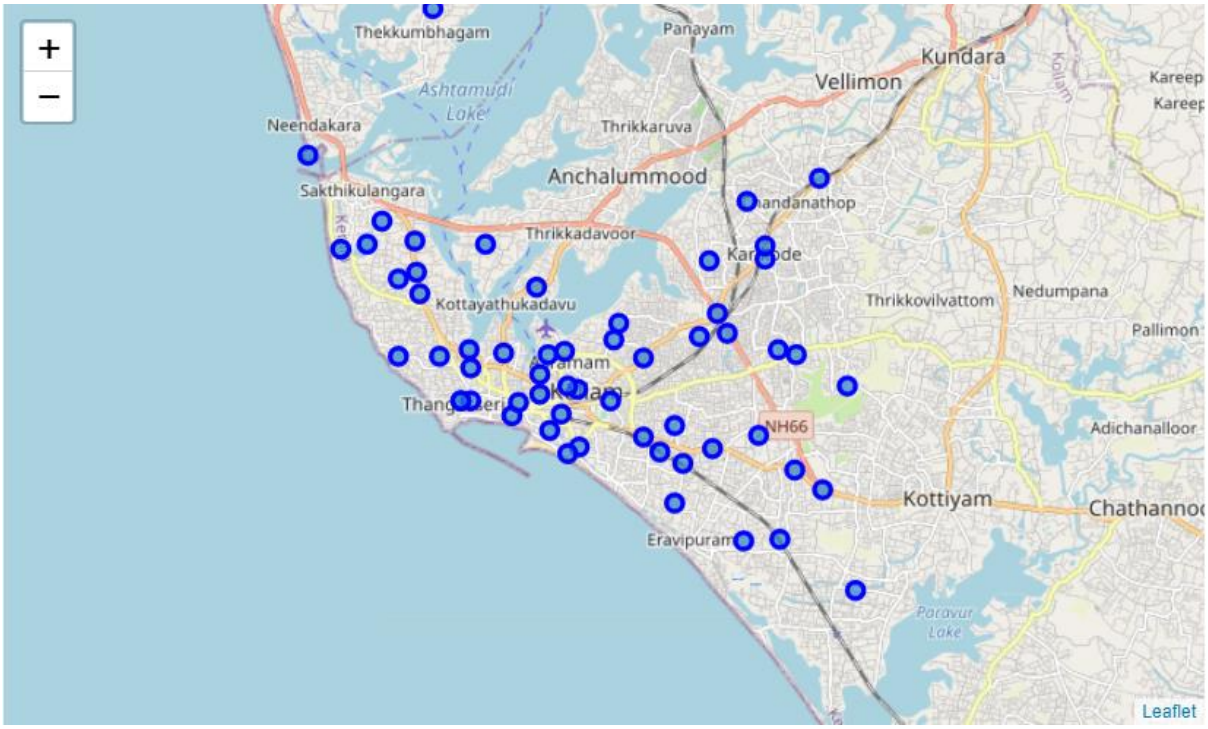

Figure : Neighbourhoods at Kollam before clustering

After using K Means clustering with k = 7 the neighbourhoods in Kollam are clustered as follows :

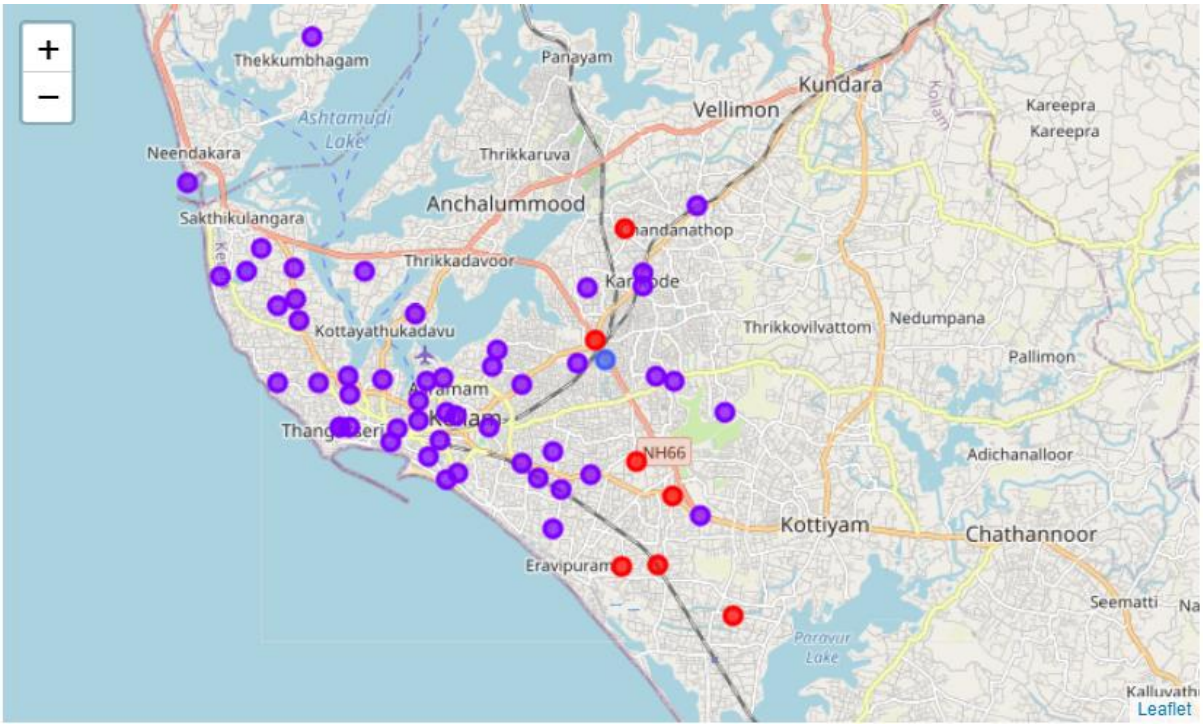

Figure : Clustered Neighbourhoods in Kollam after K Means

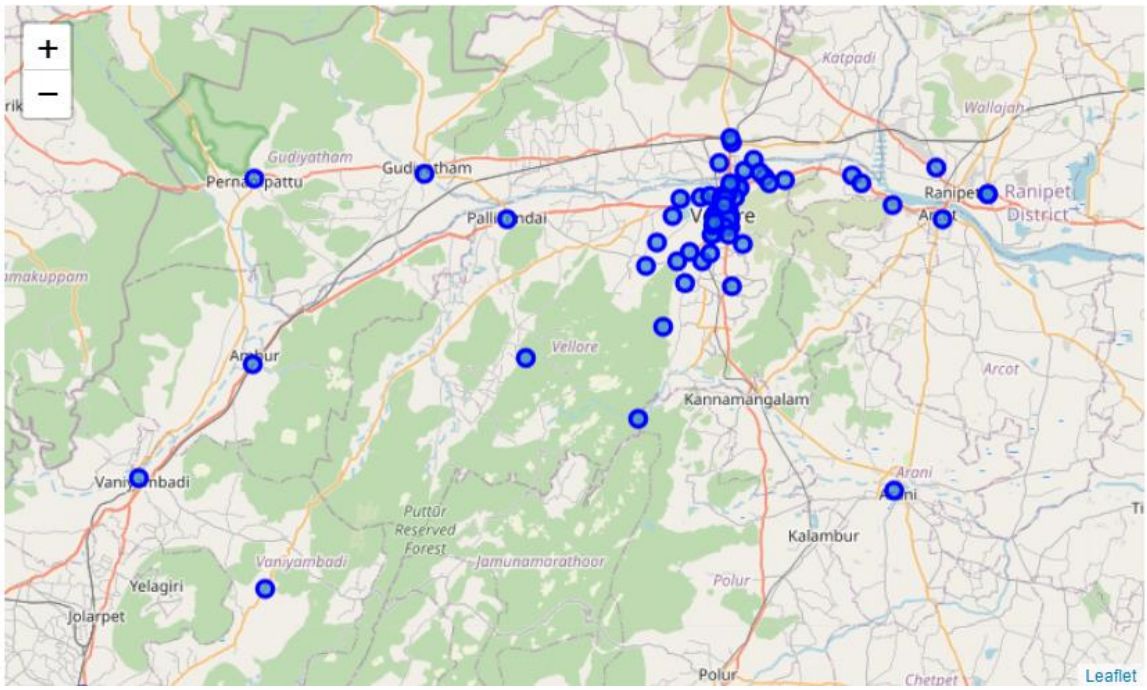

Figure : Neighbourhoods in Vellore

After using K Means clustering with k = 7 the neighbourhoods in Vellore are clustered as follows :

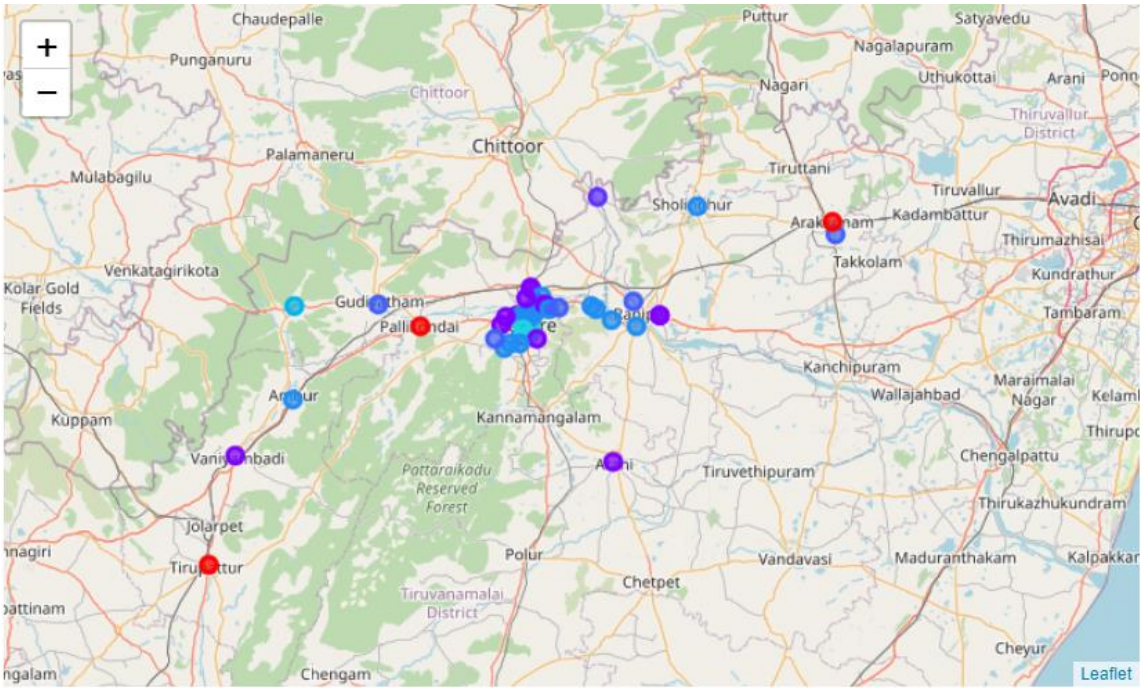

Figure : Clusters in Vellore as obtained K Means algorithm

As we can see there are a lot of different types of neighbourhoods in Vellore as compared to Kollam. To explain this, we need some background on Vellore. The reason being that Vellore is home to two major institutes Vellore Institute of Technology and Christian Medical College or CMC. The area around these institutes have much more venues than other remote locations. This results in neighbourhoods belonging to a more number of clusters than when compared to Kollam where all the neighbourhoods have similar nearby venues.

RESULTS AND DISCUSSION

In this Section we will be visualizing the various clusters obtained from our methodology and the neighbourhoods most likely to relocate to from Kilikollur, Kollam which is the neighbourhood my friend is currently living in.

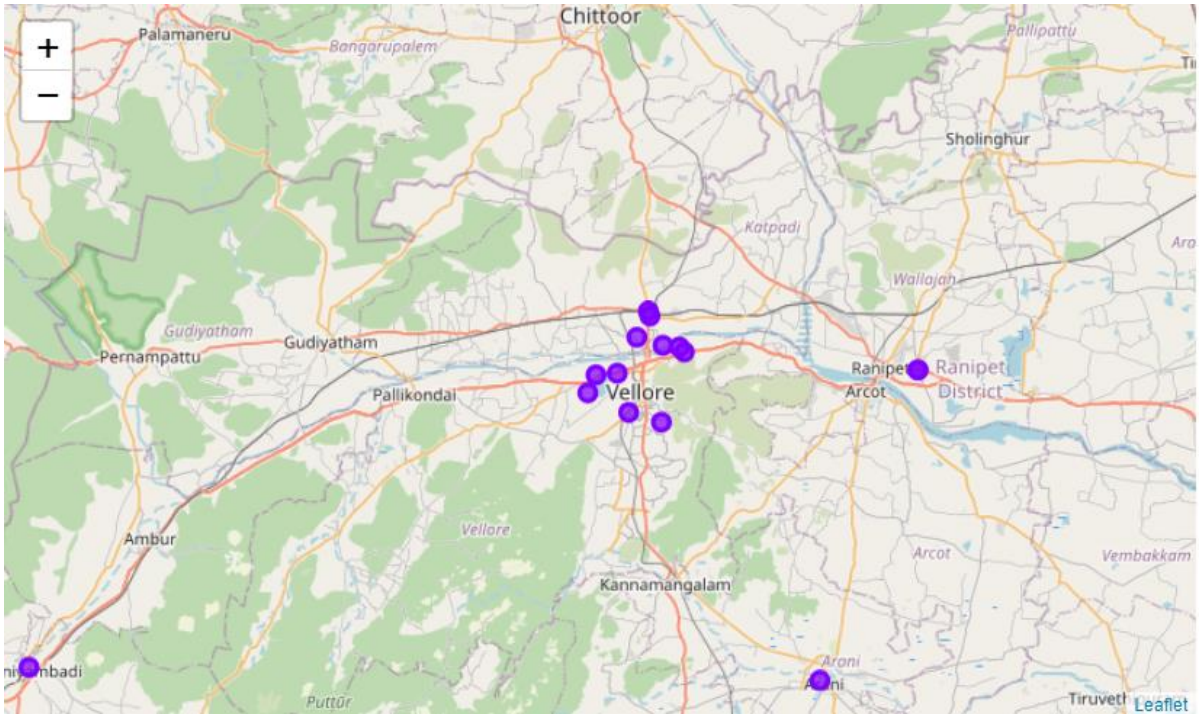

Figure : Neighbourhoods in Vellore similar to Kilikollur, Kollam

As we can see the clusters are distributed across the map at varying distances. Coming back to our original problem. Our aim is to find the best location for my friend to say outside college and hence we need to find places closer to his University. For this we have used the geolocator package to get the latitude and longitude of VIT university and for each neighbourhood in the cluster found the distance and added it to the dataframe. I then sorted this dataframe based on the distance to VIT University and from this we obtained 6 best outcomes from 14 which were closest to the university

	Neighbourhood	City	Latitude	Longitude	Cluster Labels	Distance to VIT
5	Dharapadavedu	Vellore	12.9695	79.1386	1.0	2.269835
17	Katpadi	Vellore	12.9734	79.1369	1.0	2.448409
1	Gandhi Nagar	Vellore	12.9460	79.1492	1.0	3.051000
36	Vallalar	Vellore	12.9443	79.1633	1.0	3.060659
2	Sathuvachari	Vellore	12.9399	79.1681	1.0	3.642485
39	Kazhinjur	Vellore	12.9527	79.1280	1.0	4.003187

Figure : Top 6 Neighbourhoods to relocate to from Kilikollur, Kollam

CONCLUSION

In this project I have successfully found relocation options for my friend who is considering moving from Kilikollur, Kollam to Vellore and filtered the options with respect to closeness to his University as well as similarity to his current location of residence. This can also be considered to be a proof of concept to finding similar location in an unknown region using K Means clustering and location data.