Capstone Project – The Battle of Neighbourhoods

Submitted by : Ram Manohar Rai

Topic: A project to understand how to clustering of neighbourhood occurs in south east Bangalore(IT Hub).

Content:

1. **Introduction to Business Problem**: South East Bangalore is IT hub of India. It has a huge migrant population from all over the country. It is a challenge for any government agency to manage this area due its difference from other part of the city. The south east bangalore is also not a homogenous region with different area having different type of commercial setup to cater to its population. This project will identify the 5 different type of commercial areas in south east Bangalore. The four square data will supply with the list of commercial establishments along with its category in the particular area. The categories will then form the basis of classification of regions in south east Bangalore. This classification will help policy makers which includes the mayor and urban development bodies, to treat area under single classification as one, and come up with common set of policies with respect to food heath inspection infrastructure, banking infrastructure adequacy etc. This classification will also be useful for decision over establishing a new company which will require to draw a significant number of manpower. It's very important for these companies to be established in a zone where employees can get all basic necessities like restaurant, banks etc close by.
2. **Data Used for Project:** Two Sets of Data were used to complete this project. Mapsofindia.com is a website that lists location across every city in India. I used this website to extract different neighbourhoods in Bangalore. I scrapped the page: https://www.mapsofindia.com/pincode/india/karnataka/bangalore/

   This page gave the lists of various pin codes in Bangalore, which is kind of proxy of areas with reasonably equivalent population. Along with the pincode, the url gave major areas belonging to that pincode.

   For this project. I later used Nominatim to receive latitude and longitude of all the areas. Since I was focussing only on South East Bangalore, I selected only those areas whose latitude was lesser than average latitude of all listed areas, and longitudes higher than average.

   API calls from Four Square to extract the number of venues that are available in each location are made. Instead of name of the data since we are more interested in categories of data, categories of venues was specifically extracted. Number of instance of particular categories in given area gives idea about how frequent is this kind of category in given area. Top 20 categories were later identified and rest categories were discarded as these datas were too sporadic to give any character.

3. **Methodology:** Following steps were followed during the project
    a. Web Scrapping for neighbourhood location in Bangalore: To complete this project we also required the lists of neighbourhoods in Bangalore. For this we used website mapsofindia.com which had a list of neighbourhoods in Bangalore. Using BeautifulSoup library, neighbourhood table was extracted with their pincode, district and state for disambiguity. The extracted values were then converted into dataframe object named bangalore_df.
    b. Geolocator was used to get latitude and longitude of each location. Latitude and Longitude of every location was received and a dataframe was constructed with addition of lat and long to bangalore_df and named bangalore_df_clean. Since Geolocator had network issue and was unstable, try and except method was implemented and Geolocator was called again for the rows where values were not available during previous calling.
    c. Segregating areas in South East Bangalore:- Since we are interested in IT hub of South East bangalore, we created another dataframe with latitude less than average of all locations and longitude more than mean of all locations. New data frame is Bangalore_south_East
    d. Exploring the venues in region: Using four square data region was explored for different kind of venues. A dataframe with name, category of venue, location and latitude/longitude of central location was created. Region has 163 unique type of locations. Since having type of venues which are rare will not tell much about type of area, only 20 most frequently repeated venues were selected. Rest of the venues were dropped off.
    e. Mean value of time each type of venue is repeated in an area is computed. This gives idea how dominant is given type of venue in the given area. Based on these mean values all the locations are divided into 4 clusters using k-mean clustering. K mean clustering was used because it was a classification problem we wanted to bring out hidden structure in the location data. For clustering also those areas were removed that had only one venue, as data looked incomplete. Initially 5 clusters were tried but algo was identifying one cluster with only one representation. So 4 clusters were chosen to give any meaningful clustering
    f. Each kind of cluster was color coded and represented in map using folium library.

4. Result: From the representation of clusters on map we can see that cluster 1 is forming more central location. Cluster 2 is found in bit more outer area, while cluster 3 and cluster 4 are towards the outer area of south east bangalore. We look at individual cluster, we can see that cluster 1 is dominated by Indian restaurants. As cluster 1 is more towards the centre, it seems there is correlation between more central location in city and presence of more Indian restaurant. Possible reason can be that people come here for work and do more lunches in restaurant serving traditional food. Cluster is more balanced areas with diffused waits, though Indian restaurants are still on higher side. Cluster 3 and cluster 4 which represents the outer area of bangalore with increasing representation of department stores and cafes. This make sense as these areas are supposed to be residential areas.
5. Recommendation: For policy and other administrative infra purpose, more focus should be on developing good food inspection infrastructure in cluster 1. Cluster 1 is also supposed to have higher footfalls of people on a daily basis. Cluster 2 has mixed character of residential

as well as commercial area. So focus can be on providing utilities as well as food inspection infrastructure. Cluster 3 and cluster 4 looks like total residential areas so utilities infrastructure needs to be more stronger there in comparison to food inspection.

6. Conclusion: This report can be used by local mayoral authority to decide which area needs which kind of infrastructure depending upon what kind of venues are present in the area. Our assertion of cluster 1 being more commercial area can be ascertained by using time bound popularity of venues in area.