Capstone Project (The Battle of Neighborhoods)

IBM Data Science

Rupali Panda

City of Mumbai (Photo credit: https://en.wikipedia.org/wiki/Bangalore)

Introduction

Shopping malls have become a very important part of our daily lives in metro cities. With increase in working hours and busy lifestyles, many households and professionals find it difficult to shop for merchandise and daily needs in places which are geographically apart. In addition to the basic household needs, shopping malls provide other amenities such as entertainment, health-related services and restaurants among many others, under one roof. These centres have always attracted people, especially the younger generation, who can shop for the latest fashion trends, do shopping (including grocery and home essentials), dine in restaurants, go to movies and let kids play in well-maintained facilities meant for them. Shopping malls provide a wider variety of good and services, which was not available in local stores.

Bangalore is one of the biggest cities of India. This city with a population of 10 million, is also known as the silicon valley of India due to a large number of software professionals working

for many national and international companies. Busy schedules and modern life choices leave little time for these professionals, who have set the trend of "quality shopping under one roof" and spend considerable amount of money at these malls. Therefore, opening a shopping mall is not only a lucrative business for the property developers but proves very useful for the consumers. However, opening a shopping mall requires a lot of strategic planning, especially the success of such a venture depends heavily on the choice of a suitable location.

Business Problem:

The property developers and investors need to review and then strategically target a suitable location for the shopping mall to maximize the return on their investments. The objective of this project is to analyze and select the best locations in the neighbourhoods of Bangalore to open a new shopping mall. For this analysis, Data science methodology and machine learning techniques such as clustering were used.

Target Audience:

This project will particularly be very useful for the property developers who are looking for investment in a new shopping mall in the Bangalore metropolitan area. It will also be very helpful for the investors/businesses who can open the stores inside the mall.

Data:

The following data are required for this project:

- List of all the neighbourhoods in Bangalore, which is needed to determine the location of existing shopping malls.
- Latitude and Longitude of the neighbourhoods - required to plot and map venue data.
- Venue data, which will be used to perform clustering of the neighbourhoods based on the existing malls in the city.

Data Sources and Data Science Methodology

- The data were collected by using web scarping techniques from the Wikipedia page: https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore.
- Then the Python Geocoder package was used to get the geographical co-ordinates of the neighbourhoods of Bangalore city.
- Foursquare API was used to get venue data for the neighbourhoods.
- Finally to analyze each neighbourhood and clustering the k-means clustering is used.

Methodology

As the first step in data gathering, list of all the neighbourhoods were obtained from the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore) using web scrapping methods of Python and data extraction using the beautifulsoup package. Geographical coordinates (latitudes and longitudes) of these neighbourhoods were obtained from the Geocoder package to make them usable on the Foursquare platform. The locational information, including the latitude and longitude data, were populated in a Pandas dataframe and plotted in a map using the Folium package of Python. This ensures a visual inspection of proper representation of the geographical locational data obtained from Geocoder.

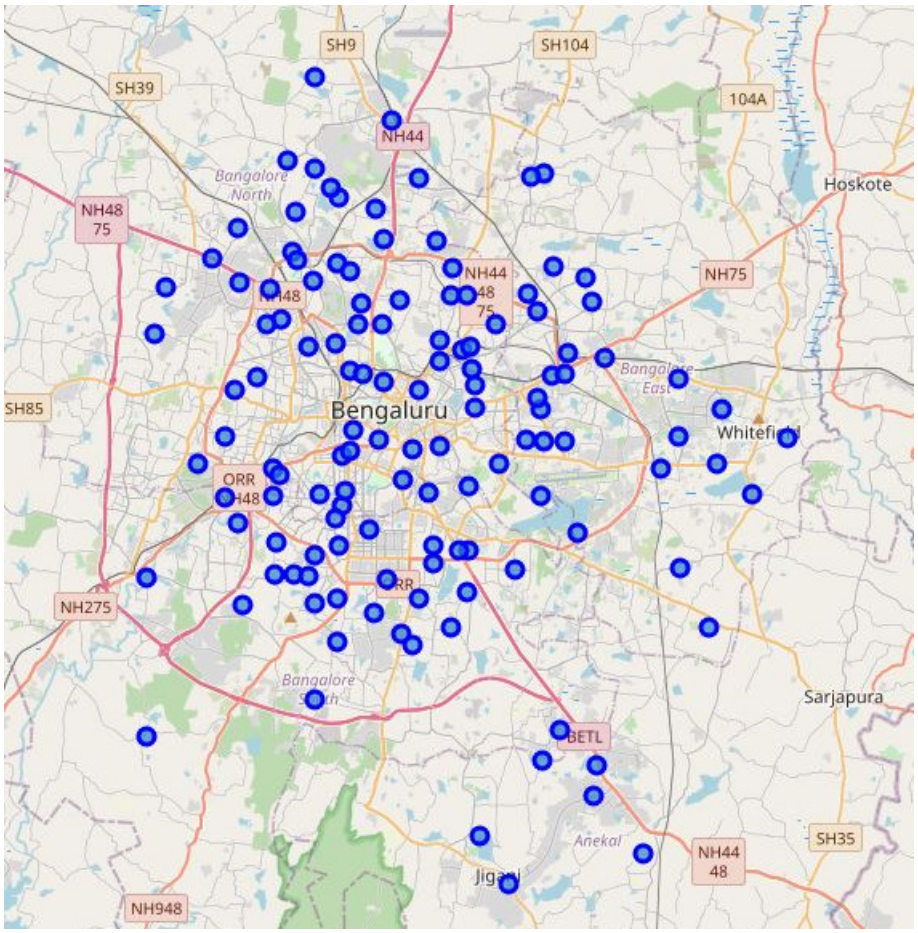

Fig. 1: Neighbourhoods of metropolitan Bangalore city.

The above data, in tabular form, were used in the Foursquare API to obtain top 100 venues, which fall within a radius of 1000 m of Bangalore. For this purpose, we created an API Developer account to obtain a Foursquare ID and key. Subsequently, API calls were made to Foursquare and the locational information (latitude and longitude) was passed in Python. Venue names and categories along with the latitudes and longitudes, in JSON format, were extracted from Foursquare API. These data

helped in understanding the unique categories existing in each neighbourhood. The neighbourhoods were then grouped by the frequency of each venue category. Then, the data was filtered for the relevant "Shopping Mall" category for this analysis.

In the final step, k-means clustering, the most popular unsupervised clustering algorithm in machine learning, was used to identify 5 centroids for 5 clusters. Each data point is assigned to its nearest cluster, while keeping the centroid a minimum. The clustering was based on the concentration of shopping malls in the neighbourhoods. We attempt to find the most suitable neighbourhood locations based on this criterion.

Results

Based on the frequency of occurrence of "Shopping Mall" neighbourhoods in the Bangalore metropolitan area are categorized into 5 clusters by k-means clustering method (Fig. 2):

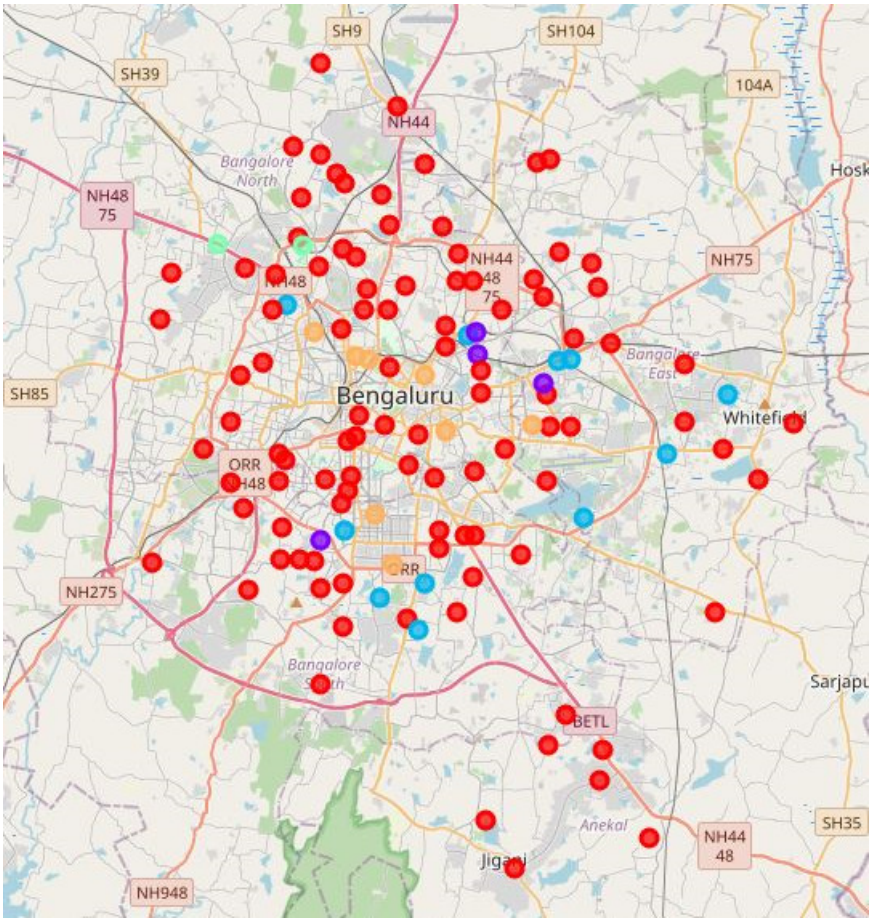

Fig. 2: Clustering of Shopping Mall categories in Bangalore

Cluster 1 (Red): Neighbourhoods with lowest concentration of shopping malls

Cluster 2 (Purple): Neighbourhoods with moderate number of shopping malls

Cluster 3 (Blue): Neighbourhoods with high concentration of shopping malls
Cluster 4 (Green): Neighbourhoods with 2nd highest concentration of shopping malls
Cluster 5 (Orange): Neighbourhoods with highest number of shopping malls

Discussion

Most of the shopping malls are in the central area of Bangalore, with the highest numbers in cluster 5, then in cluster 4 and moderate numbers in cluster 2 and 3. However, cluster 1 has very low number of shopping mall in the neighborhoods. Hence it is advisable for the property developers to capitalize on these findings and open new shopping malls in Cluser 1 neighbourhood. As there are not many shopping malls in Cluster 1, the new shopping mall will be very successful as there will be very little to nil competition. However, it seems that many shopping malls already exist in Cluster 5 and cluster 4. Therefore, it is advisable to avoid those neighbourhoods for a new shopping mall project.

This project considers only one aspect (i.e., concentration of shopping malls) of the neighbourhoods in Bangalore to come up with the best possible venues for a new shopping mall. However, other factors such as demographic data and real estate value affect the decision as well. Since these data are not readily available, reliability of this model may have been compromised. Also, an advanced user account of the Foursquare API (instead of the free Sandbox Tier account used in this report) would yield better results. Different clustering algorithms could also be used to compare the results.

Conclusion

In this project, we have attempted to find the best location for opening a Shopping Mall in Bangalore metropolitan area. We have identified the business problem; specified the data required and how to obtain them from the appropriate source; extracted and prepared the data for analysis; conducted clustering analysis using k-means clustering algorithm of machine learning technique and creating 5 clusters based on the concentration of shopping malls in the neighbourhoods. The outcome of the project analysis culminated in recommendations that Cluster 1 is the most preferable location for a new shopping mall. This finding would help all relevant stakeholders to make informed decisions.

References

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

https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore