Clustering of shopping mall locations in Delhi, India

By

Abhishek Choudhery

Coursera Capstone Project

Introduction

Shopping malls provide the population concentrated avenues of shops, multiplexes, food joints to sate their entertainment, hunger pangs, and all the other activities. The shopping mall is becoming synonymous with the one-stop-shop to all your needs. As a result, these are in great demand. With careful planning, it can reap great profits.

Business Problem

The main process to conduct this study is to create the clusters of shopping malls present in India, to create a density map to choose the best site for construction of a new shopping mall.

An ideal site would be where exists a cluster of shopping malls as it will show that the shopping mall is successful in that but also it should not be overcrowded with other malls. This condition will not be profitable for the builders.

Target Audience

The key audience for this study would be the property developers, construction companies, and investment companies who are looking to venture in the field of shopping malls.

Data

Our model would be needing the following-

1. List of constituencies in Delhi – These would markup the project boundaries to areas of Delhi
2. Coordinates of constituencies of Delhi

3. Data related to shopping malls present in Delhi, where they are present, what is their location, their density, etc.

Sources of Data

1. **List of constituencies**
   This data would be extracted from the https://ceodelhi.gov.in/Content/EntireDelhiLocalities.aspx
   We would be using BeautifulSoup to extract the list of constituencies of Delhi.
2. **Coordinates of constituencies of Delhi**
   This data, we would be extracting using Python Geocoder package which will give us the latitude and longitude coordinates of the these constituencies
3. **Venue Data**
   For the venue data, we will use the FourSquare API. This API would provide us with all the information related to our needs.

Methodology

The methodology involved is loosely based on the CRISP-DM methodology. First and foremost was our Business understanding. Our main requirement was to analyze the positioning of shopping malls in Delhi to find out the best placement for shopping malls.

The next step was for Data collection. For this, we employed the website scarping tools such as BeautifulSoup to scrape for the list of areas in Delhi. This data was present on the site https://ceodelhi.gov.in/Content/EntireDelhiLocalities.aspx.

After obtaining the list of names, we employed python Geocoder to gather the list of coordinates from the name of the location.

After that, we employed the FourSquare API to gather the data about the venues present in the locations. after gathering this data we saved it in the pandas Dataframe and visualize it over the map of Delhi using the Folium package. Using this we gather 500 venues in 5000 m radius. The data we received was in JSON format. We processed the data to remove any spurious information. We calculated the number of venues received per location. We analyzed each location by grouping it by constituency and taking the mean frequency of occurrence of each venue category.

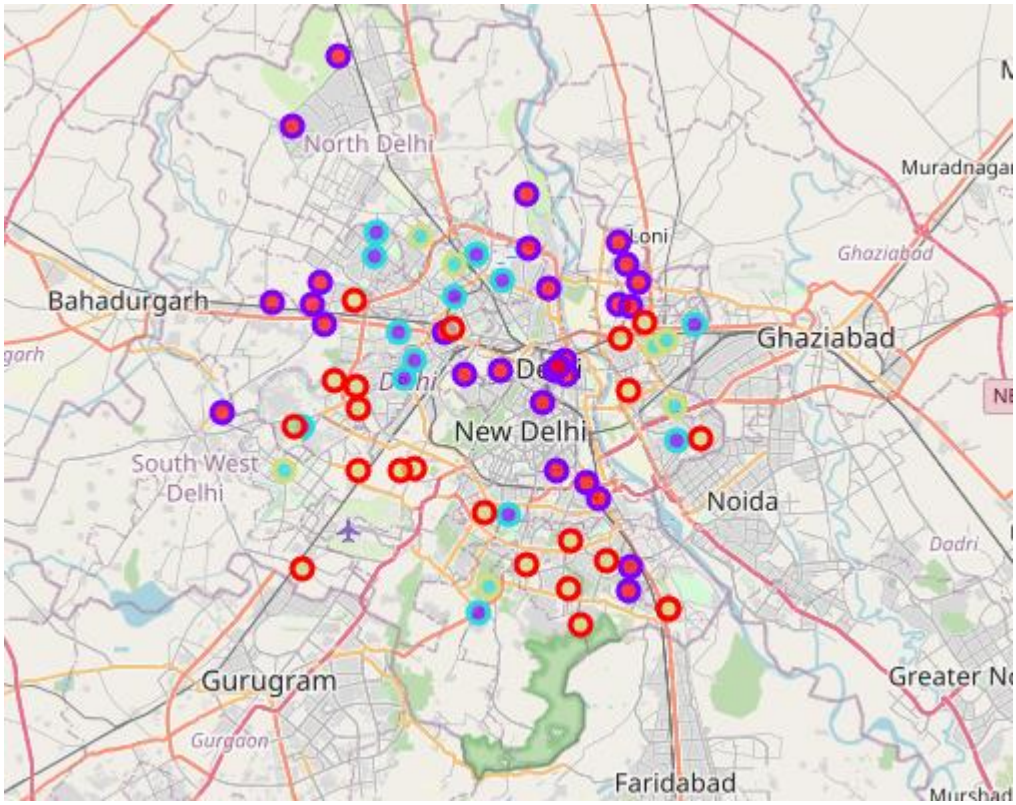
After that, we filtered the data using ‘Shopping Mall’ as the category. Our data ready, we decided to performed KMeans Clustering upon the data. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. Using the Elbow method we find the optimal cluster values to be 4. Using this model we clustered the Shopping Malls in 4 groups based on their frequency of occurrence

in the constituency. It would be helpful for us to answer the business requirement which areas are suitable to develop as a new Shopping mall.

Result

It divided the cluster into 4 groups

1. Cluster – 0 Area having lowest or no existence of shopping mall concentration
2. Cluster – 1 Area having moderate shopping mall concentration
3. Cluster – 2 Area having the highest shopping mall concentration
4. Cluster – 3 Area having lowest to moderate shopping mall concentration.



Discussion

As observed in the positioning of clusters we can conclude that the areas directed in cluster 0 have no competition, this presents a double-edged sword as these areas could be the untapped areas or areas with no demand for shopping malls. Also in areas in cluster 3 have low to moderate concentration. Both of these areas depict the opportunities to exploit. With careful extrapolation of data, the developers and investors can identify the relevant areas to build the shopping malls. Here the Generic variety of Malls can be built upon without expecting stiff competition.

Cluster 1 has a moderate concentration shopping mall, here the malls having the USP or catering to special niche or demographics can be built. The competition here will be

moderate whereas Cluster 2 has the highest mall concentration. Here, the developers should avoid it.

But this is not the only metrics that should be involved in choosing the mall location. The other host of factors should be included in creating a conclusive result. This report can act as a starting point for the research.

Conclusion

Cluster 0 and cluster 3 shows the most promising location for opening up new malls. Here we can work with the generic malls and simple services. Cluster 1 also can be utilized for the new mall. The main focus should be providing niche demographics to stand out among your competitors. Cluster 2 should be avoided.