Data Science Capstone Report:

The Battle of the Neighborhoods

By:Mrigank Jain

May,2019

Introduction

Shopping malls serve as a great place for recreational activites and thus for various businesses to flourish, develop and make profits.There are restaurants, cafes, multiplexes, shopping outlets, grocery shops, arcades and much more.Shopping malls often a target the people of one or more localities by centralizing their location amongst the localities to get the right amount of inflow of customers with low competition with other vendors or shopping malls.Therefore one should consider the location to be the most influencing factor when building a new shopping mall for it be a succes or failure.

Business Problem

The objective of this capstone porject is to analyze all neighbourhoods of Delhi, India and select the best ones to open a new shopping mall.We will use data science methodology and machine learning techniqus like clustering.The problem statement that can be solved using this project is:

“In the city of Delhi,India if a property developer wants to open a new shopping mall,what locations can be recommended for it?”

Target Audience

This project is very useful for property dealer and investors looking for new profitable localities to open new shopping malls in the city of Delhi,India. There is significant overcrowding ,oversupply and unwanted competition among malls in some significant neighborhoods of Delhi while others dont have enough for basic supply.

Data

We shall need the following data:

- A dataset containing the names of all the neighborhoods in Delhi along with their latitude and longitude coordinates.
- We will use the coordinates to find venue data for each locality and and to plot them on the delhi map.
- We will use the venue data to cluster the neighbourhoods based on shopping malls.

Data sources:

- The delhi neighborhood data was downloaded from the kaggle website:https://www.kaggle.com/shaswatd673/delhi-neighborhood-data/data#
- After getting the neighbourhoods with their respective coordinates from the link above ,we'll use the Foursquare API to get the venue data for each neighborhood.Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.The API will provide many venue category but we require only the shopping mall category for each neughbourhood in order to proceed.

Methodology

First,we will download the csv file of the Delhi neighborhoods dataset from the link: https://www.kaggle.com/shaswatd673/delhi-neighborhood-data/data# .
After the data is loaded into a dataframe,we will enter our Foursquare Client ID and Client Secret that we got by registering and creating an application.
The id and secret will be used to get the top 100 venues for each neighbourhood present in our dataframe in a 2000 metres radius and store in a new dataframe.

We will analyze each neighborhood with respect to all venue categories by making one hot encoded dataframe for the same.

We will then group rows using the groupby function on the one hot encoded dataframe and take the mean of the frequency of occurence of each venue category,thus creating another dataframe.

Now we will extract the shopping mall and the neighborhood column from the recently created dataframe to perform clustering on the data by using k-means clustering.

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.

We will cluster the neighborhoods into 3 clusters based on their frequency of occurence for “Shopping Mall” venue category.

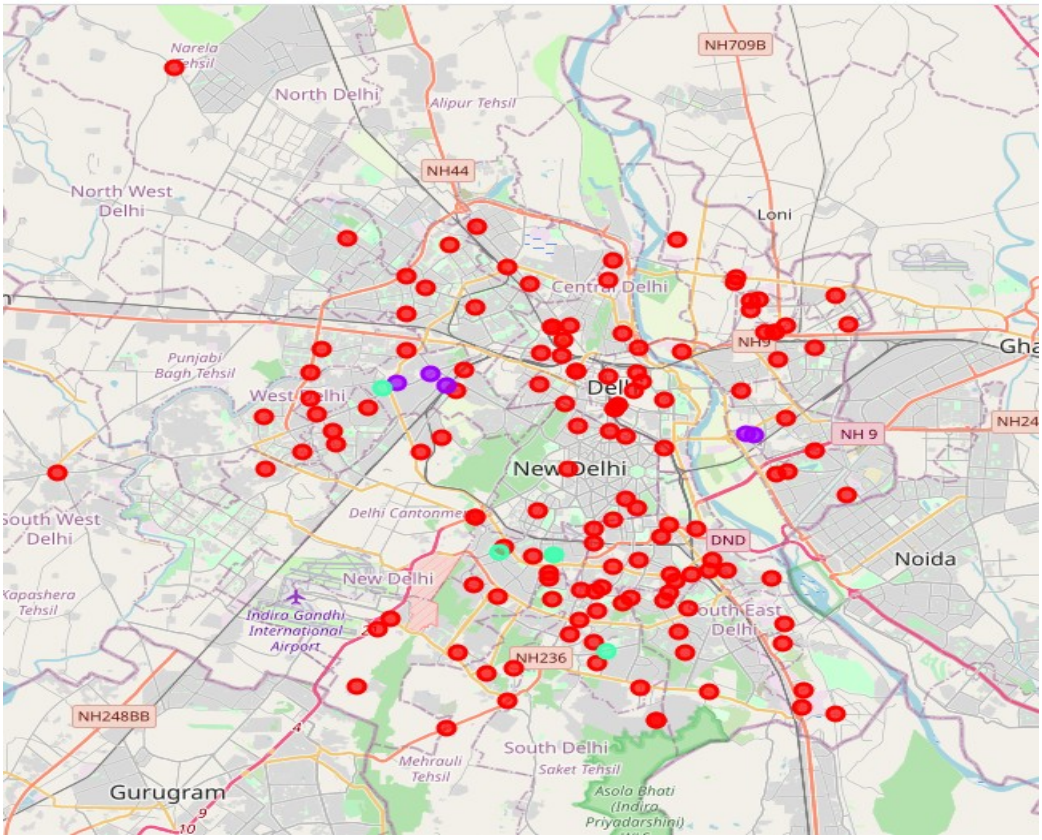
The result will allow us to identify which neighborhoods have higher concentration of shopping malls while which have lower number.

Results

The results from the k-means clustering shows that we can categorize the neighborhoods into 3 clusters based on the frequency of occurence of “Shopping Mall” venue category:

- Cluster 0 :Neighborhoods wirh high concentration of shopping malls.
- Cluster 1 & 2:Neighborhoods wirh low concentration of shopping malls.

The result are visualized in the map below with cluster 0 in red,cluster 1 in purple and cluster 2 in mint-green colour:



Conclusion

Most of the shopping malls in Delhi are concentrated in the central and the southern regions of the state(as shown by the cluster 0).While the northern,north-western and western regions have very low amount of shopping malls as shown by the clusters 1 and 2.

For a property developer who wants to find a place to build a shopping mall in Delhi,they can use this clustered data to see which area will be the most profitable due to less competition by other shopping malls.Shopping malls in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, this also shows that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb areas still having very few shopping malls.

Therefore this project recommends property developers to capitalize funds to open new shopping malls in areas of cluster 1 or 2 with little or no competition,and avoid areas of cluster 0 with high concentration of malls and oversupply of goods,thus resulting in intense competition.