IBM Applied Data Science Capstone

Finding Best Location to Open New Shopping Mall in New Delhi, India

By: Aditya Uniyal

Introduction

For many people, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city of New Delhi and many more are being built. Opening shopping malls allow property developers to earn consistent rental income.

Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best location in the city of New Delhi, India to open a new shopping mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question:

Considering existing competitors in the market, what could be the best location in the city of New Delhi, India to open a new shopping mall?

Data

To solve the problem, we need the following data:

- List of neighborhoods in New Delhi. This defines the scope of this project which is confined to the city of New Delhi, that serves as a capital of India.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi) contains a list of neighborhoods in New Delhi, with a total of 141 neighborhoods. Web scraping techniques such as requests module and beautiful-soup packages can be used to extract the data from the Wikipedia page. Then I will be using the Python Geocoder package which will return the latitude and longitude coordinates of the neighborhoods.

After that, I will be using Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.
Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem.
This project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling to machine learning (K-means clustering) and map visualization (Folium).

Methodology

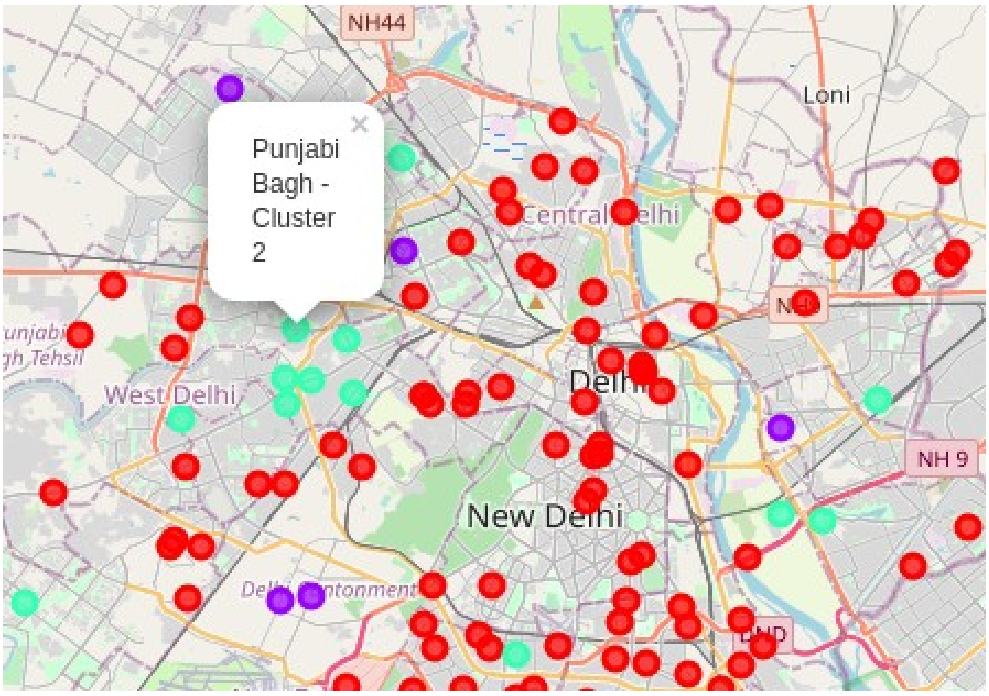
Firstly, we need to get the list of neighborhoods in the city of New Delhi. The list is available on (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi). Using Python requests module and beautiful-soup package we can extract the list of neighborhoods. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder package that will allow us to convert addresses into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of New Delhi.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With this data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighborhoods.

Lastly, we will perform clustering on the data by using k-means clustering.
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 occurrence for Shopping Mall. The result will allow us to identify the concentration of shopping malls in neighborhoods. Based on the occurrences of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Shopping Mall”.



The results of the clustering are visualized in the map above with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color

Cluster 0: Neighborhoods with highest number of shopping malls

Cluster 1: Neighborhoods with lowest number of shopping malls

Cluster 2: Neighborhoods with moderate concentration of shopping malls

Discussion

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of New Delhi, with the highest number in cluster 0 and moderate number in cluster 2.
On the other hand, cluster 1 has very low number of shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still having very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 0 which already have high concentration of shopping malls.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of shopping malls, there are other factors such as population and income of residents that could influence the location decision of a new shopping mall. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall.
In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could also make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into “3” clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall. The answer to the business question that was raised in the introduction section is: The neighborhoods in cluster 1 are the most suitable locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

References

Category Neighborhoods in Delhi: Wikipedia
https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi

Foursquare Developers Documentation. *Foursquare*
https://developer.foursquare.com/docs