Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Shopping Mall in Mumbai, Maharashtra

By: Arohi Narang

Introduction

Malls are a great source of entertainment especially during weekends and holidays. People can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more recreational activities. It is a place enjoyed by all age groups and no one can ever get bored at a mall. Malls are the paradise 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. Builders also try to buy land close to shopping malls. As a result, there are many shopping malls in the city of Mumbai and many more are being built. Opening shopping malls allows 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 analyse and select the best locations in the city of Mumbai, Maharashtra to open a new shopping mall. Using data science and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: Which are the best locations in Mumbai where a property developer should look to open a new shopping mall?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in the city of Mumbai

.

Data

To solve the problem, we will need the following data:

- List of neighbourhoods in Mumbai. This defines the scope of this project which is confined to the city of Mumbai, the financial capital city of the country of India.

- Latitude and longitude coordinates of those neighbourhoods. 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 neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page(https://en.wikipedia.org/wiki/Category:Suburbs_in_Mumbai) contains a list of 42 neighbourhoods in Mumbai. We use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful soup packages. Then we get the geographical coordinates of the neighbourhoods using Python Geocoder package which will gives us the latitude and longitude coordinates of the neighbourhoods.

After that, we use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.
Foursquare API will provides 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 put forward. This is a 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). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

1. Data Extraction: We get the list of neighbourhoods in the city of Mumbai from the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Mumbai). We will do web scraping using Python requests and beautiful soup packages to extract the list of neighbourhoods data.

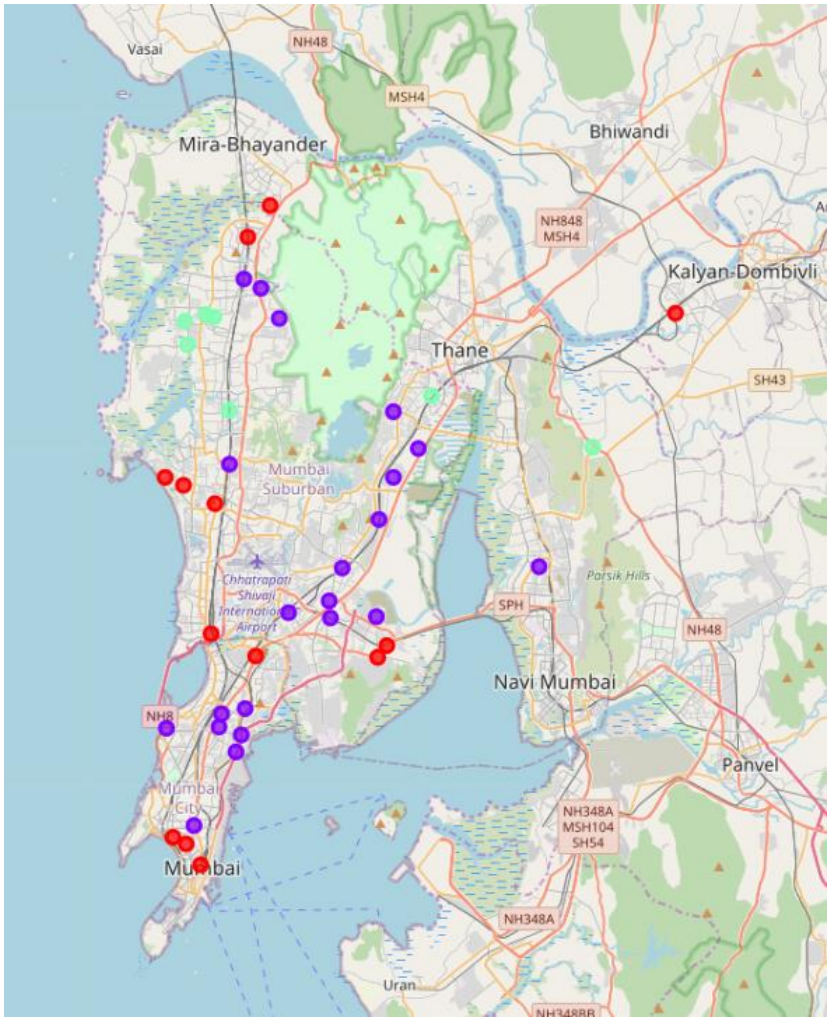
2. Obtain Geographical Coordinates: We get the geographical coordinates in the form of latitude and longitude by using the wonderful Geocoder package. It converts the address into geographical coordinates in the form of latitude and longitude.
3. Visualization on Map: After gathering the data, we will populate the data into a pandas Data Frame and then visualize the neighbourhoods 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 Mumbai.
4. Use Of Foursquare API: We use this tool 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 neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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 analysing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighbourhoods.
5. K means Clustering: Lastly we 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 neighbourhoods into 3 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

Results

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

- Cluster 0: Neighbourhoods with least number of shopping malls
- Cluster 1: Neighbourhoods with moderate number of shopping malls
- Cluster 2: Neighbourhoods with highest concentration of shopping malls

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



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. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. 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 account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results. We have estimated the radius to get the information on the different venue categories. Each suburb in Mumbai has a different size. In the above project the radius of 5000 and if we take a higher radius it will be too big for the other suburbs. Thus Andheri bandra fall in cluster 0 but in reality there are malls in that suburb. They belong to the western suburb which is in cluster 2 and maybe if we increase the radius we will get a more accurate result.

Conclusion

Most of the shopping malls are concentrated in the western suburbs of Mumbai city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to totally no shopping mall in the neighbourhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 2 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 western suburbs of the city. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighbourhoods in cluster 0 with

little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighbourhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

References

Category:Suburbs in Kuala Lumpur. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Category:Suburbs_in_Mumbai

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