IBM Capstone Final Project

Opening a new Cinema Hall in Mumbai, India

By: Shreya P. Jain

October 2020

Introduction

Cinema is a major source of recreation in most countries around the world, even in India. Multiplex segment is growing while single screen segment is declining. As of March 2005, there were approximately 13000 cinemas in India out of which 73 were multiplexes with total 276 screens. Multiplexes constitute only 0.6% of about 12000 cinema halls, but account for 28% to 34% of box office collection for the top 50 films.

People are turning towards multiplexes due to various reasons, some of which are safety, better ambience, eateries, etc.

For many movie lovers, visiting multiplexes is a great way to relax and enjoy themselves during weekends and holidays. Multiplexes re like a one-stop destination for all types of movie-watchers. The location of multiplex is one of the most important decisions that will determine whether the multiplex 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, India to open a new multiplex mall. Using data science methodology and machine learning techniques like clustering. This project aims to

provide solutions to answer the business question: Where should a property developer or investor open a multiplex in the city of Mumbai, India?

Data

To solve the problem, we will need the following data

- List of Neighbourhoods in Mumbai. This defines the scope of the project, which is confined to the city of Mumbai.
- The latitude and longitudes of those neighbourhoods, required to plot the map, as well as get the venues data.
- Venues data, particularly data related to multiplexes, cinema halls, theater, etc.

The Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai#Mumbai_neighbourhood_coordintes) contains the list of neighbourhoods and their latitude and longitude. We will use pandas for web scraping.

After that, we will use Foursquare API to get the venue data for those neighbourhoods.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Mumbai, India, and their geographical locations. Luckily the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai#Mumbai_neighbourhood_coordintes). We will do the web scraping using Python requests and pandas package. We then convert the list obtained to a pandas DataFrame.

We use folium package to visualize the neighbourhoods. This allows us to check that the geographical coordinates we have are approximate.

Next, we will use FourSquare API to get the top 200 venues that are within the radius of 5000 metres. We need to register a FourSquare Developer Account in order to get the credentials required to make an API call. We then make a loop to call for data for all the coordinates. The data obtained is in json format, which we clean and convert to a pandas DataFrame.

With the data, we check how many venues were returned for each neighbourhood, and how many unique categories can be curated from all the returned venues. Then we analyse each neighbourhood by grouping the rows by neighbourhood, and taking mean of the frequency of occurrence of each venue category. By doing so, we are preparing the data for clustering.

Since we are analysing for multiplex, we use all the related coloumns only, which include – multiplex, movie theater, indie movie theater, etc.

Lastly, we perform clustering on the data using k-means clustering.

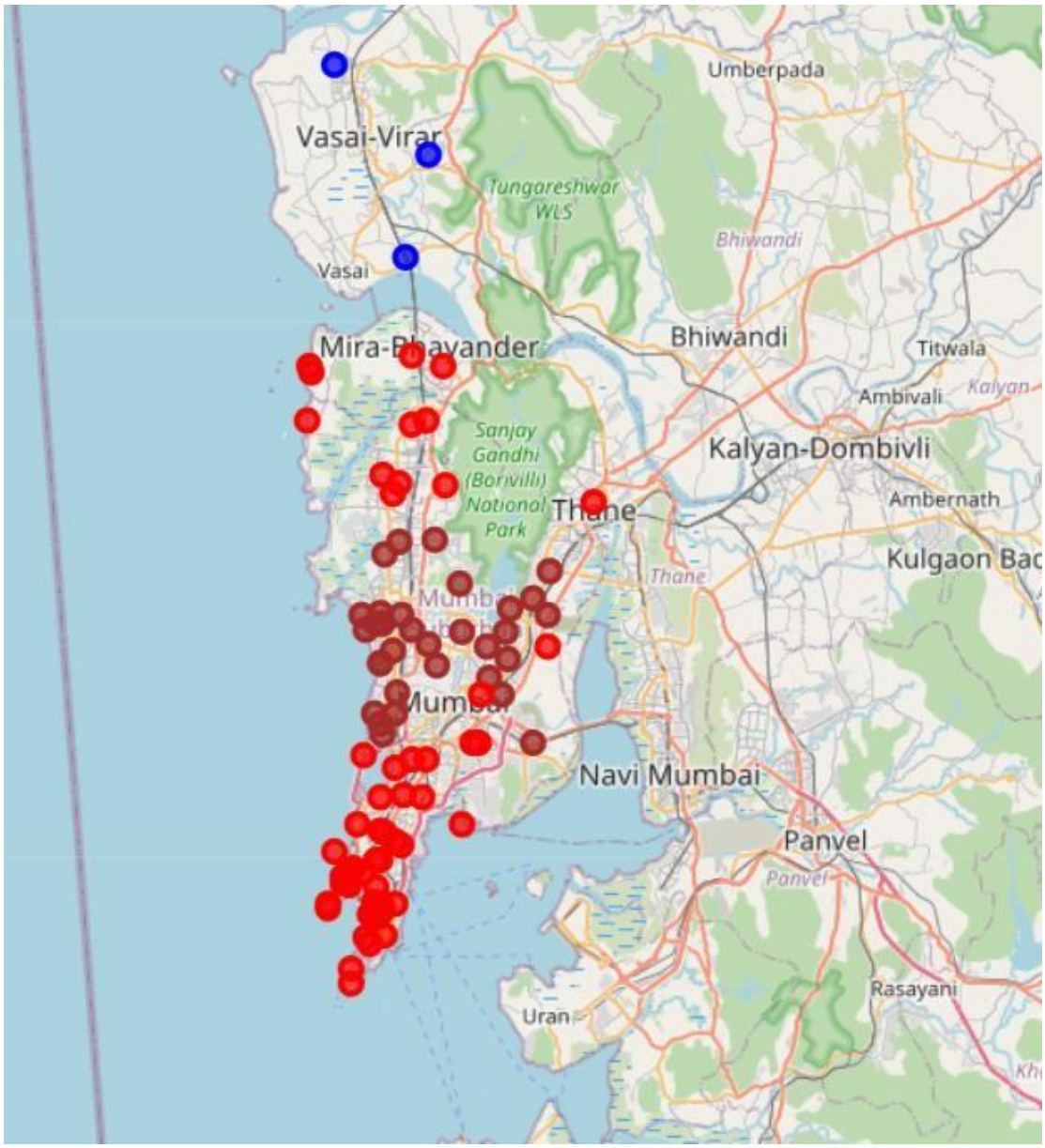
K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms. We cluster the neighbourhoods into 3 clusters, based on the frequency of occurrence of cinemas/multiplexes. The result will allow us to identify the locations suitable for opening new multiplexes.

Results

The results from the k-means clustering show that we can categorise the neighbourhoods into 3 clusters based on the frequency of occurrence of multiplexes and theatres :

- Cluster 0 : Areas with moderate frequency of multiplexes and theaters
- Cluster 1: Areas with low number to no existence of multiplexes and theaters
- Cluster 2: Areas with high concentration of multiplexes and theaters.

The resulting clusters are visualized in the map below, with cluster 0 in brown, cluster 1 in red, cluster 2 in brown.



Discussions

As noted from the map, cluster 2 has the highest number of multiplexes and cinemas, while cluster 1 has very low number of multiplexes. This represents a great opportunity and high potential areas to open new multiplexes as there is very little competition from the exisiting ones. Meanwhile, multiplexes in cluster 2 are likely suffering from intense competition.

Therefore, this project recommends property developers and investors to capitalize on these findings to open new multiplexes in cluster 1 areas. Investors with unique selling propositions to stand out from the competition can also open new multiplexes in cluster 0. Lastly, it is advised to avoid areas in cluster 2 which already have high concentration of multiplexes and cinema halls.

Limitations and Suggestions for Future Research

In this project, we only consider one factor, i.e the frequency of multiplexes, and theaters. There are however other factors such as population, income of residents, etc. that could influence the location of new multiplex. However, to the best knowledge of the researcher, such data is not available to the neighbourhood level required by this project.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required extracting and preparing the required data, analysing the data by using k-means clustering algorithm, and lastly, providing recommendations to the relevant stakeholders. To answer the business question: Cluster 1 is the most preferred locations to open new multiplexes.