IBM Capstone Final Project

By : Shreya P Jain
October 2020

Introduction

Cinema is a major source of recreation in most countries around the world, even in India.

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

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

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.

Data Sources

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.

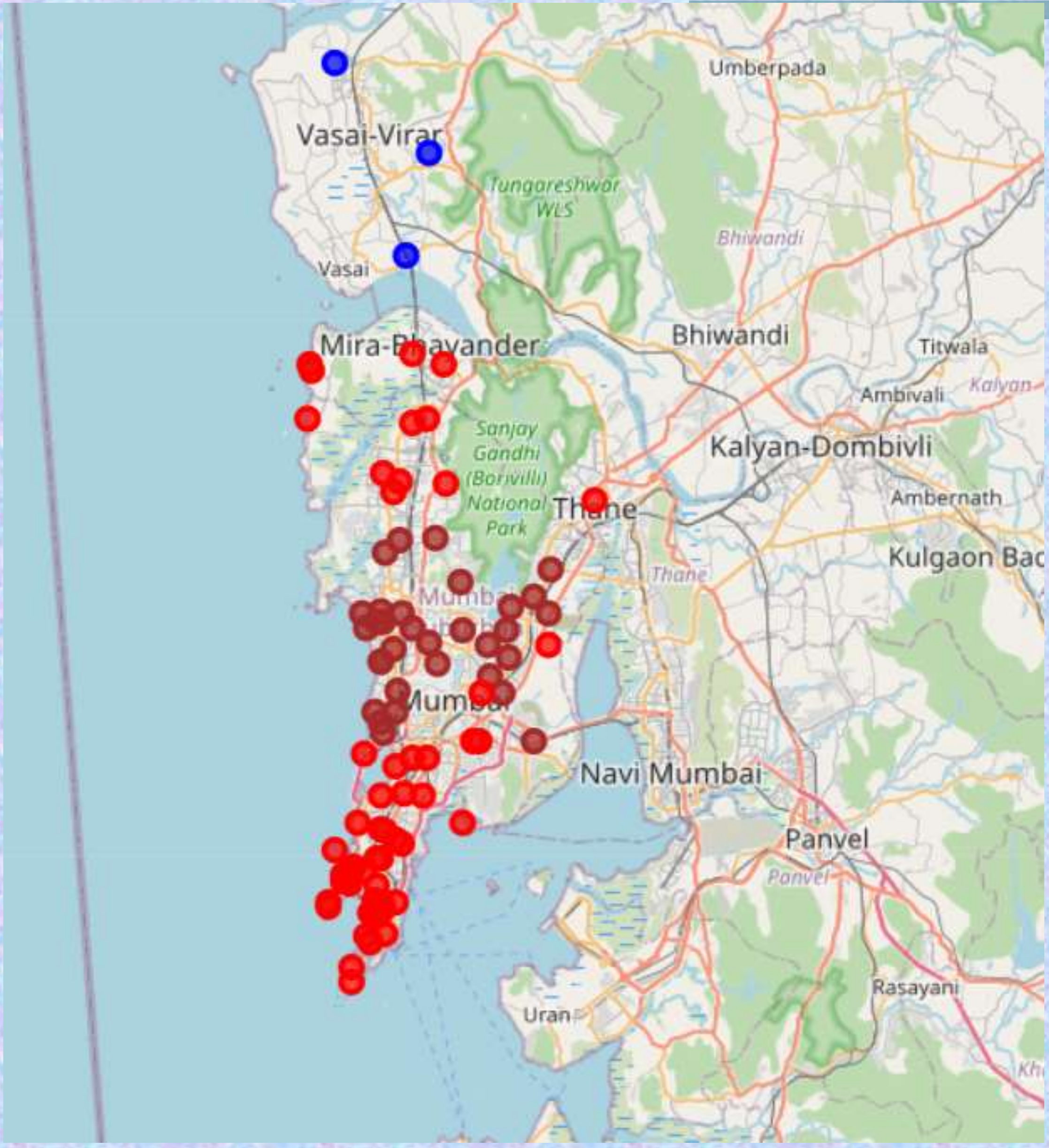
After that, we will use Foursquare API to get the venue data for those neighbourhoods.

Methodology

- ❖Web scraping Wikipedia page for list of neighbourhoods and their geographical coordinates.
- ❖Use FourSquare API to get venues data.
- ❖Group data by neighbourhood and taking mean of frequency of occurance f each venue.
- ❖Filter the data based on ‘cinema halls’, ‘multiplex’ and ‘theatre’.
- ❖Cluster the areas using k-means algorithm.
- ❖Visualize the map clusters using folium.

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.


Umberpada
Vasai-Virar
Tungareshwar WLS
Bhiwandi
Vasai
Mira-Bhayander
Bhiwandi
Titwala
Ambivali
Kalyan
Sanjay Gandhi (Borivili) National Park
Kalyan-Dombivli
Thane
Ambernath
Kulgaon Bad
Thane
Mumbai
Mumbai
Navi Mumbai
Panvel
Panvel
Rasayani
Uran

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.

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.