COURSERA CAPSTONE

IBM Applied Data Science Capstone

Opening New Business In Mumbai, Maharastra

BY

PRASHANTH CHITTOMPALLY

Introduction

For many entrepreneurs, it is very difficult to choose which business to start. Many entrepreneurs don't know which business will give them profits, so they need data science to help them to find out which business would be a safe venture. Choosing correct business is a tough call for entrepreneurs to make because if they choose bad one it may lead to losses. So, new business entrepreneurs need data science to help them in choosing correct business. For many shoppers, 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 Kuala Lumpur 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, India to open a new shopping mall. And also knowing whether any other business would be better than opening shopping mall in Mumbai. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer these business questions: In the city of Mumbai, if a property developer is looking to open a new shopping mall, where can he open it and whether there is any business which is in much demand than shopping mall in Mumbai?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in the city of Mumbai. Mumbai has very less shopping malls. Mumbai is the financial city of India. It is most developed city in India. So establishing a business here would be more beneficial for entrepreneurs. Establishing a business which is not there would be more profitable for companies.

Data

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

• List of neighborhoods in Mumbai. This defines the scope of this project which is confined to the city of Mumbai, city 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:Suburbs_of_Mumbai) contains a list of neighbourhoods in Mumbai, with a total of 42 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods. After that, we will 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 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 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

Firstly, we need to get the list of neighborhoods in the city of Mumbai. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai). We will do web scraping using Python requests and Beautifulsoup packages to extract the list of neighborhoods data. 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 wonderful Geocoder package that will allow us to convert address 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 coordinate's data returned by Geocoder are correctly plotted in the city of Mumbai. 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 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 neighborhood and examine how many unique categories can be curate 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. 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 occurrence for "Shopping Mall". The results will allow us to identify which neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the occurrence 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. We will then see any other category which has potential to be the better option than shopping mall and repeat the same procedure to find out best place to open that category venue.

Results

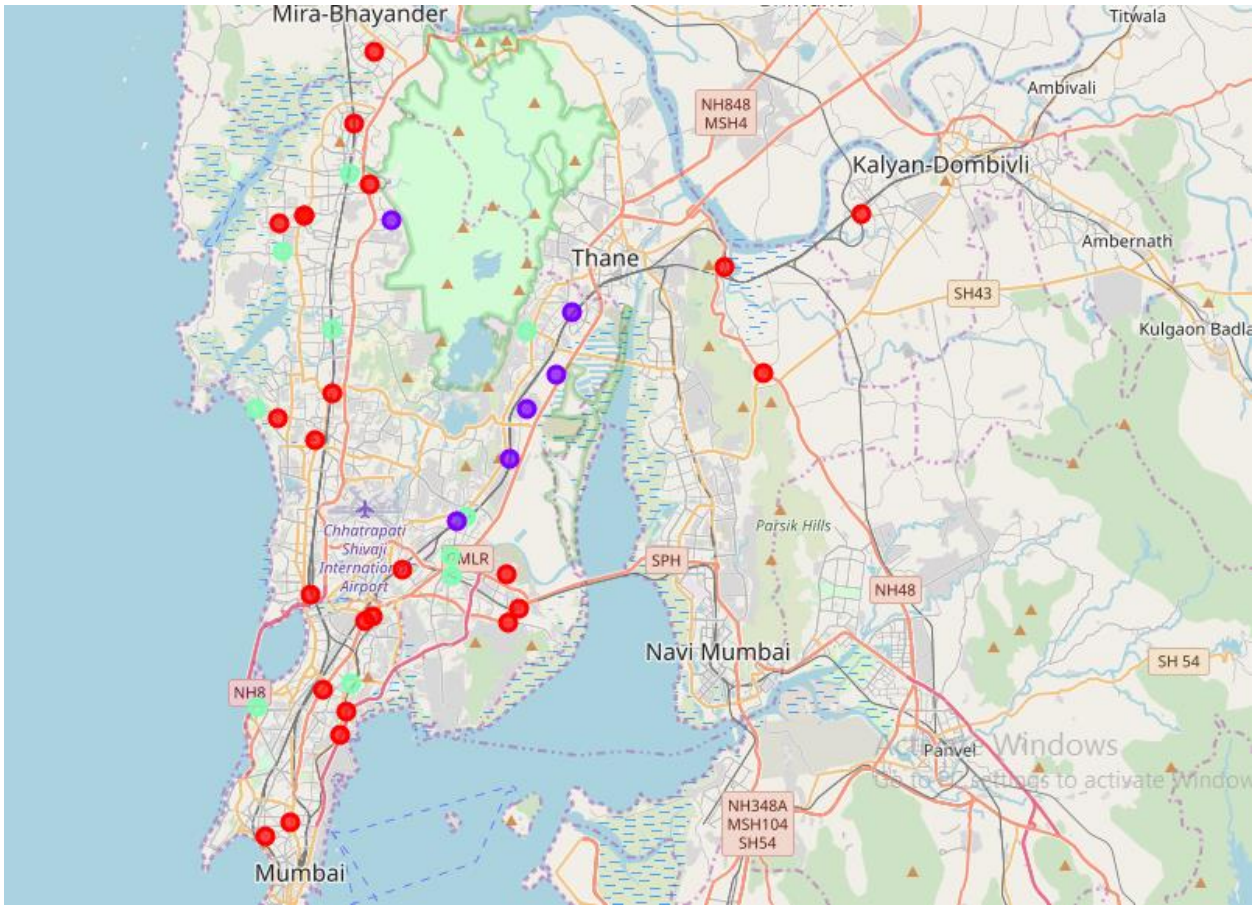
The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Shopping Mall”:

• Cluster 0: Neighborhoods with low number to no existence of shopping malls

• Cluster 1: Neighborhoods with more number of shopping malls.

• Cluster 2: Neighborhoods with moderate 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.



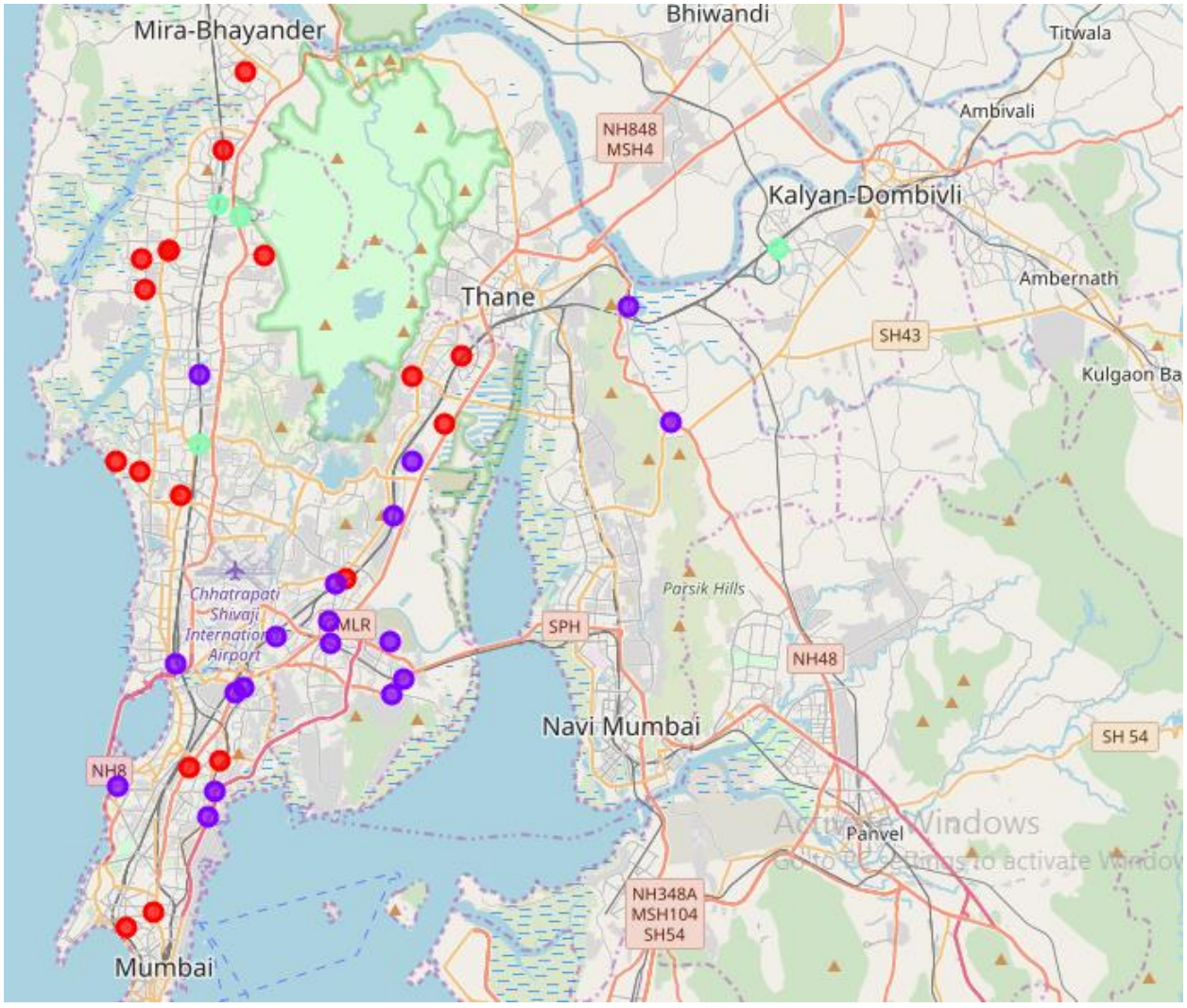
We can see Mumbai's most common venue is Indian Restuarant. We can select Indian Restuarant as other category but there are 375 Indian Restuarants in Mumbai. So, it won't be nice to start another Indian Restuarant because it would get huge competition from other Indian Restuarants. When we see Ice Cream Shop it is one of the top five most common venues. Let's see segmentation of Ice Cream Shop. The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Ice Cream Shop":

• Cluster 0: Neighborhoods with moderate number of Ice Cream Shops.

• Cluster 1: Neighborhoods with low number to no existence of Ice Cream Shops

• Cluster 2: Neighborhoods with high number of Ice Cream Shops.

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.



Discussion

We can see Mumbai's most common venue is Indian Restuarant. We can select Indian Restuarant as other category but there are 375 Indian Restuarants in Mumbai. So, it won't be nice to start another Indian Restuarant because it would get huge competition from other Indian Restuarants. When we see Ice Cream Shop it is one of the top five most common venues. Cluster 0 has moderate Ice Cream Shops, Cluster 1 has no Ice Cream Shops and Cluster 2 has most of Ice Cream Shops. So, it is better to start Ice Cream Shop in Cluster 1 areas and most of areas in Cluster 1 are good areas and they are with no Ice Cream Shops. So, it is better to start Ice Cream Shops in these areas. Ice Cream Shops are most visted places than Shopping Malls in Mumbai. So, it is better to start Ice Cream Shop than Shopping mall.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of shopping malls and Ice Cream shop, there are other factors such as population and income of residents that could influence the location decision for new shopping mall and Ice Cream Shop. However, to the best knowledge of this researcher such data are not available to the neighborhood 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 Sandbox Tier 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.

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. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new shopping mall. Ice Cream Shops are most visited places than Shopping Malls in Mumbai. So, it is better to start Ice Cream Shop than Shopping mall. The neighborhoods in cluster 1 areas are the most preferred locations to open a new Ice Cream Shop. 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 or Ice Cream Shop.

References

Category: Suburbs in Mumbai. *Wikipidea*. Retrieved from

https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai

Foursquare Developers Documentation. *Foursquare*. Retrieved from

https://developer.foursquare.com/docs

Appendix

Shopping Mall

Cluster 0

- Andheri
- Tilak Nagar (Mumbai)
- Sion, Mumbai
- Shil Phata
- Mumbra
- Mira Road
- Matharpacady, Mumbai
- Mankhurd
- Mahavir Nagar (kandivali)
- Kurla
- Kandivali
- Kalyan
- Juhu
- Jogeshwari
- Kausa
- Baiganwadi
- Eastern Suburbs (Mumbai)
- Dombivli
- Anushakti Nagar
- Grant Road
- Devipada
- Dahisar
- Bandra
- Charkop

Cluster 1

- Thakur village
- Mulund
- Bhandup
- Vikhroli
- Kanjurmarg
- Uttan
- Vashi

Cluster 2

- Wadala
- Goregaon
- Seven Bungalows
- Pestom sagar
- Borivali
- Chembur
- Western Suburbs (Mumbai)
- Ghatkopar
- Sonapur, Bhandup
- Worli

Ice Cream Shop

Cluster 0

- Andheri
- Wadala
- Uttan
- Thakur village
- Sonapur, Bhandup
- Seven Bungalows
- Mulund
- Mira Road
- Mahavir Nagar (Kandivali)
- Western Suburbs (Mumbai)
- Kandivali
- Kalyan
- Juhu
- Grant Road
- Kausa
- Ghatkopar
- Dahisar
- Bhandup
- Charkop

Cluster 1

- Mumbra
- Anushakti Nagar
- Vikhroli
- Vashi
- Baiganwadi
- Tilak Nagar (Mumbai)
- Bandra
- Sion, Mumbai
- Shil Phata
- Worli
- Goregaon
- Chembur
- Matharpacady, Mumbai

- Mankhurd
- Kurla
- Kanjurmarg
- Eastern Suburbs (Mumbai)
- Pestom sagar

Cluster 2

- Borivali
- Devipada
- Dombivli
- Jogeshwari