Coursera Capstone Project

IBM Applied Data Science

By

Shivani Mahaddalkar

August 2020

Introduction

For the Applied Data Science Capstone Project, this attempts to determine the potential neighborhoods for a new restaurant for Indo-Chinese cuisine in the city of Mumbai in India. The two most popular cuisines in the city are the Indian and Chinese cuisines, the idea to have a fusion restaurant where both cuisines are available in one restaurant could be successful. Mumbai is India's financial capital and the seventh most populous city in the world. For an investor looking to invest in a new Indo-Chinese restaurant, Mumbai can prove to be very profitable. With the growing population of the city and its increasing standard of living, such a restaurant can be a good investment for entrepreneurs looking to profit from the economic situation of the city. As the demand for fine dining increases, the patrons for the restaurants increases and therefore, this project aims to identify neighborhoods with such profitable opportunities.

Business Problem

In a city like Mumbai, a good fine dining experience is the optimal one, where the patron could enjoy both of the popular cuisines at one place in a single meal. Restaurants prove to get high footfall even during weekdays and profits from investing in such ventures are guarantees. In the project, similar neighborhoods are clustered together to find viable options for a new Indo-Chinese restaurant.

Data

The data required for the analysis is obtained in three steps.

1. List of neighborhoods in Mumbai. They are categorized by their pin codes in a csv file obtained from a Mumbai Tour Guide page.
2. The geographical coordinates of the neighborhoods.
3. Venues present in these neighborhoods within 500 m radius of their geographical coordinates.

Data Sources

The list of neighborhoods is obtained from a website in a list from a Mumbai Tour

Guide page (https://mumbai7.com/postal-codes-in-mumbai/). The code reads the file using a command and converts it into a pandas data-frame with the neighborhood names and the postal codes of the same. The pandas data-frame has 157 neighborhoods in Mumbai.

The Python Geocoder package gives the latitudes and the longitudes of each of those neighborhoods. It uses the postal codes and geographical coordinates are stored in another data-frame in the order of the postal codes table. First few entries of the data-frame are given below.

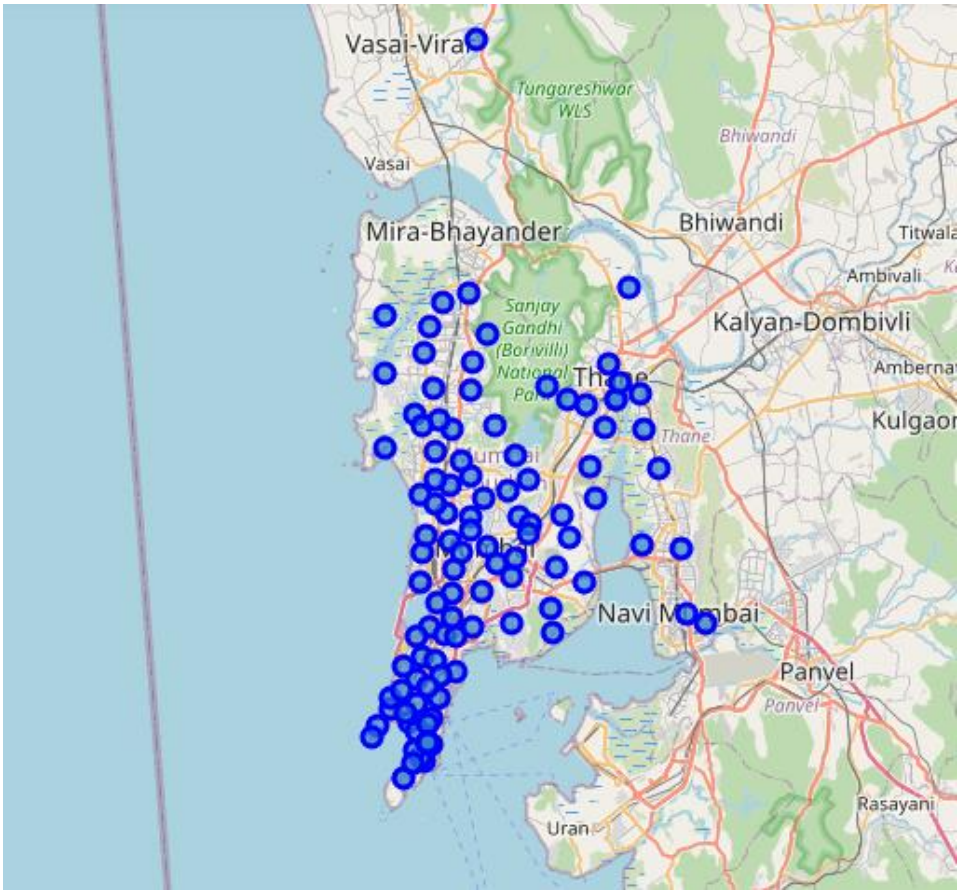
	latitude	longitude
0	18.964005	72.807983
1	19.161085	72.884394
2	19.119298	72.851100
3	19.122935	72.840610
4	19.020313	72.868280

Then using the Foursquare API, upto 100 venues within 500 m of radius of the neighborhood coordinates are extracted. Venue names, categories and coordinates are fetched for further understanding. A few venue categories are Brewery, Bookstore, Indian Restaurant, Chinese Restaurant etc.

Methodology

This section explains the process used to extract, analyze and the data is obtained from a csv file from a Mumbai Tour website in a pandas data-frame. The list is downloaded from 'Mumbai Guide' website (https://mumbai7.com/postal-codes-in-mumbai/). The data extracted has the name of the city (Mumbai), the name of the neighborhood and the pin code of the neighborhood.

Further, to get the geographical coordinates for the neighborhoods present in the list, the Geocoder package is used. The latitudinal and longitudinal coordinates are found out using the pin code of the neighborhood from the data-frame above. These coordinates are then populated in the neighborhoods' data-frame. Then using the folium package the neighborhoods are visualized on the map of Mumbai. The picture below is the map of Mumbai with the neighborhoods in the data-frame highlighted.



Then, the Foursquare API is used to bring upto 100 venues in each neighborhood within 500 m radius of the latitudinal and longitudinal coordinates of the neighborhood. To use the Foursquare API, a developer account needs to be created and using the Client ID and Client Secret Key the API can be accessed. Making calls to the API, we get the name of the Venues, the coordinates of the Venues and the category of the Venues. The frequencies of each venue category for both Indian and Chinese restaurants for every neighborhood are calculated. Then both the frequencies are added. The thought behind it is, for an Indo-Chinese restaurant both Indian and Chinese restaurants have to be considered as competition. Hence, both the frequencies are added to form a new column. Adding the frequencies of occurrence of Indian Restaurants and Chinese Restaurants to the neighborhoods' name prepares the data-frame for k-means clustering.

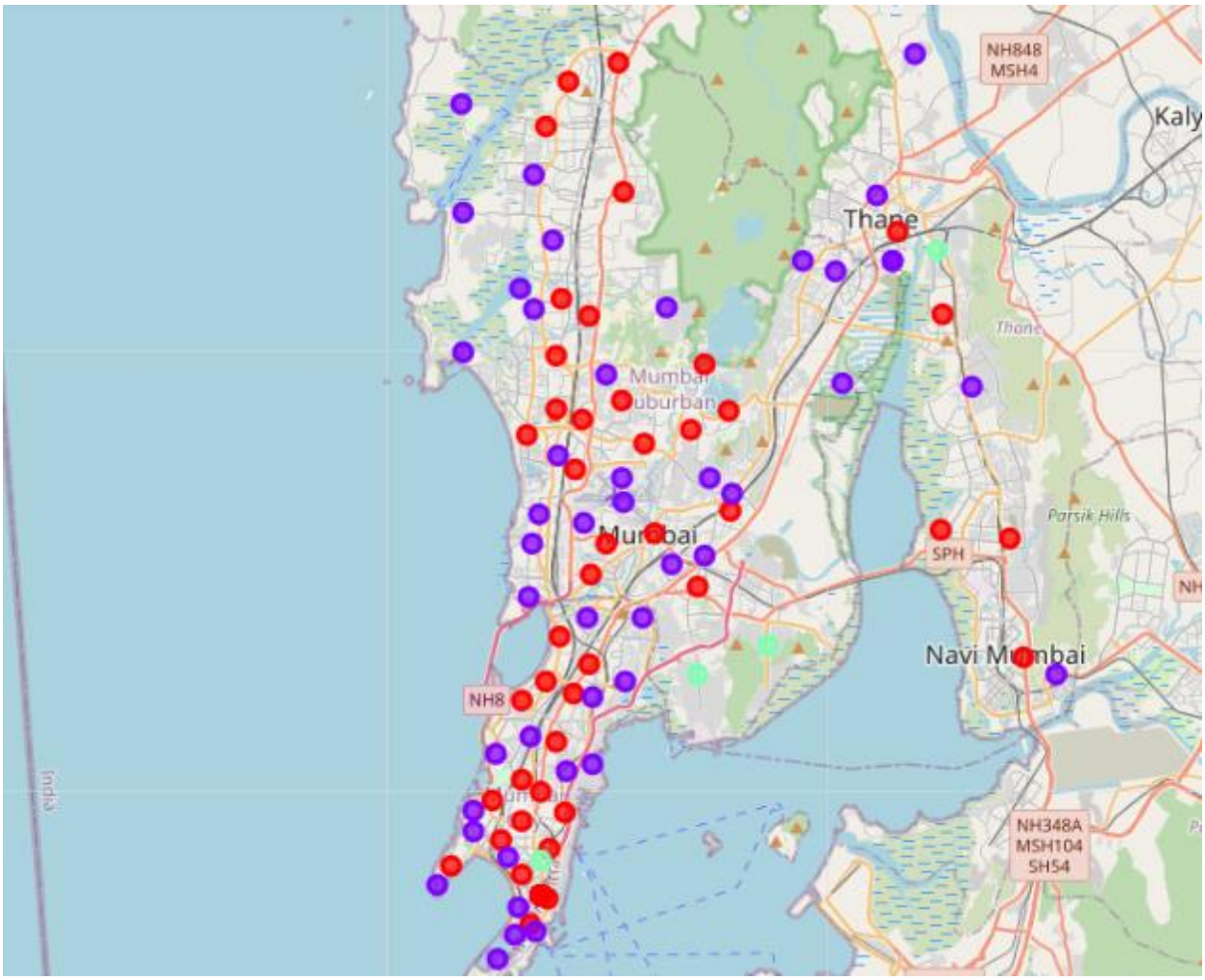
K-means clustering is an unsupervised machine learning algorithm where similar neighborhoods according to the frequencies are grouped together. For k-means clustering, k number of centroids are identified and all data points are assigned one of the k clusters. The assignment happens such that the distance between a data point and all k centroids is calculated and the nearest centroid to the data point is the cluster to which the data point becomes a part of. Here, neighborhoods are assigned to one out of three clusters. The clusters can be later classified as high, medium and low presence of Indian and Chinese restaurants. The neighborhoods present in each of the neighborhoods are a strong indicator to which neighborhoods are viable for an Indo-Chinese restaurant business venture. We visualize the three clusters on the Mumbai map as well. The image below shows it with three different colors.

Results

The clusters are classified into high, medium and low concentration of Indian and Chinese restaurants. The clusters are as follows:

- Cluster 0: Medium frequency of Indian and Chinese restaurants
- Cluster 1: Low frequency of Indian and Chinese restaurants
- Cluster 2: High frequency of Indian and Chinese restaurants

In the map, cluster 0 is indicated in red, cluster 1 is indicated in purple and cluster 2 is indicated in mint green.



Neighborhoods in cluster 0:

Mahim	Kasa	Suryanagar
Talasari	Kandivli (East)	Mantralaya
Malabar Hill	Thane (H Q)	Rajawadi
Wagle Industrial Estate	Kalbadevi	Prabhadevi
Kurla	Juhu	Poonam Ngr Jogeshwari (E)
Krishi Utpanna Bazar	Jogeshwari (West)	Parel
Kosbad Hill	Jawhar	Papdi
Tarapur	SEEPZ	Palghar H O
Tarapur App	Jacob Circle	Nirmal
Tarapur J/A	JNPT Town Ship	Nerul Mode
Kelwa Mahim	Mandpeshwar	Naupada
Kelwa	J B Nagar	Nalasopara (East)
N I T I E	N A D Karanja	Mumbra
Mumbai G P O	Mumbai Central	Satpati
Mokhada	Mira Road	Mira
Mazgaon	Matunga	Sopara
I I T Mumbai	Tulsiwadi	Grant Road
Borivli (West)	Bordi	Turbhe
Bhayander (East)	Bhayandar	Bhavani Shankar Road
Belapur	Bassien Road	Bassien
Vidyanagari	Vikramgad	Vile Parle (East)
Bandra (East)	Ballard Estate	Balcum
Arnala	Virar	Andheri (West)
Andheri (East)	Airoli Mode	Agashi
Veer Jijamata Bhosle Udyan	Chakala MIDC	Boisar
Goregaon (West)	Goregaon (East)	Umbarpada
Chinch Bunder	Gholvad	Uran
Uttan	Ganeshpuri	Vajreshwari
Vangam	Chembur	Dapcheri
Saki Naka	Chinchani	Dahanu Road
Vashi	Dahanu	Dadar
Council Hall	Vasai Road East	Dahisar

Neighborhoods in cluster 1:

Santacruz P&T Colony	Sandoz Baug	Tank Road
Wadala	Vile Parle (West)	Tilak Nagar
Santacruz (East)	Santacruz (West)	Vesava (Versova)
Thane (East)	Sewri	Sion
Vasai East I/E	Aarey Milk Colony	Rajbhavan
Girgaon	Ghatkopar (West)	Ghansoli
Dharavi	Delisle Road	Cumballa Hill
Sahar	Colaba	Bhandup (East)
Barve Nagar	Bangur Nagar	Bandra (West)
August Kranti Marg	Antop Hill	Borivli HO
Jakegram	Hutatma Chowk	Worli
Kandivli (West)	Khar	Kharodi
Nehru Nagar	Konkan Bhawan	Kopri Colony
Malad (West)	Jogeshwari (East)	Marine Lines
Motilal Nagar	Mulund (West)	Nariman Point

Neighborhoods in cluster 2:

Anu Shakti Nagar	Mandvi
Kalwa	F C I Mumbai
Manor	

Discussion

In the cluster 0, moderate number of Indian and Chinese restaurants are present in these neighborhoods. A lot of neighborhoods are present in this cluster. These are mostly in the southern parts of Mumbai. The moderate number of already established ventures can prove to be an obstacle. Hence, these neighborhoods are not recommended for a new Indo-Chinese restaurant. In cluster 1, low number of Indian and Chinese restaurants are present. These neighborhoods are present in huge numbers in the western and the suburban side of Mumbai, the less affluent parts of Mumbai. These will have lesser competition for fine dining restaurants. Cluster 2 with high frequency of Indian and Chinese restaurants are not a lot in number. These are present in the higher income neighborhoods. Although, these neighborhoods seem to be lucrative based on the income group of the population, the competition in these neighborhoods is high and the running establishments in these neighborhoods make them not lucrative neighborhoods for a new Indo-Chinese restaurant venture. Hence, the neighborhoods in cluster 1 should be focused on by someone looking to invest in an Indo-Chinese restaurant.

The limitation for this is the only feature taken into consideration is the frequency of Indian and Chinese neighborhoods. Other important factors to be considered can be population, mean income, reasons for the lack of restaurants, which could prove to be important for decisions in establishing a restaurant and the type of restaurant.

Conclusion

With the growing rate of the middle-class population and the increase in demand for food outlets and restaurants, investment in an Indo-Chinese restaurant. With a venture like this, with a lot of decisions to be made in starting such a business, a well-researched location for a restaurant can be a good starting point in planning to invest in an Indo-Chinese restaurant. In this project, using an unsupervised machine learning algorithm and an API which gives details about venues in a neighborhood, we were able to narrow down potential neighborhoods in the city. The investors could add in more constraints to narrow down the pool further from the present one present in cluster 1.