Predicting the Neighborhood to get settled in India

-Kapil Bansal

1. Introduction

1.1. Background

India is a country in South Asia. It is the second-most populous and seventh-largest country by area. Delhi, Mumbai, Chennai and Kolkata are four most important metro cities in India. Metro city are the urban cities that are highly populated. People move towards metro cities in search of better job and opportunities and a better life. These cities are popular and a place of interest for many due to their superior infrastructure like road, metro, safety, good quality education etc.

1.2. Problem

It is often difficult to decide to choose one of these cities for settlement. The deciding factor would be the superior and unique facilities these cities provide when compared to each other. This project aims to predict the best place to get settled in these metro cities.

1.3. Interest

People might be interested in knowing the analyze of different neighborhoods and the facilities and opportunities these neighborhoods can provide before settling or investing their money in any of these metro cities.

2. Data acquisition and Cleaning

2.1. Data Sources

For any data science project or analysis, data is most important. For this project, data can be found at government portal here. This dataset contains Indian postal codes along with their state name and coordinates. We have to download the CSV files and then load the data. This dataset, however, lacks data for latitudes and longitudes. We will use Google Geocoding APIs for filling data.
We will also use Foursquare APIs to get the venues in each neighborhood.

2.2. Data Preprocessing

There are several problems with the dataset. The dataset is huge and contains the data of all the states. However, we need data of the four metropolitan cities only. Also, there is a lot of missing data too.

2.2.1. Data Cleaning

We will select only those rows which have name of those four cities in their taluk (administrative district). Also, there are same pin codes for different entries so we will keep the first entry only.

2.2.2. Filling missing data

We don't have coordinates values in our data. Therefore, we will use Google Geocoding APIs to get latitudes and longitudes value using pin codes. For better understanding we will add one more column as neighborhood in our dataset (as the office name is not that insightful). However, some errors might get crept in.

We will manually remove those rows which contain coordinates outside of India or the rows for which we are unable to fetch coordinates values.

2.2.3. Feature Selection

There are a lot of features in the dataset. However, we need only neighborhood data and its coordinates for analyzing. Therefore, we will use Neighborhood, Taluk, Pin code, latitude and longitude. By using Foursquare APIs, we will extract different venues in each of the cities.

3. Methodology

3.1. Analyzing the data

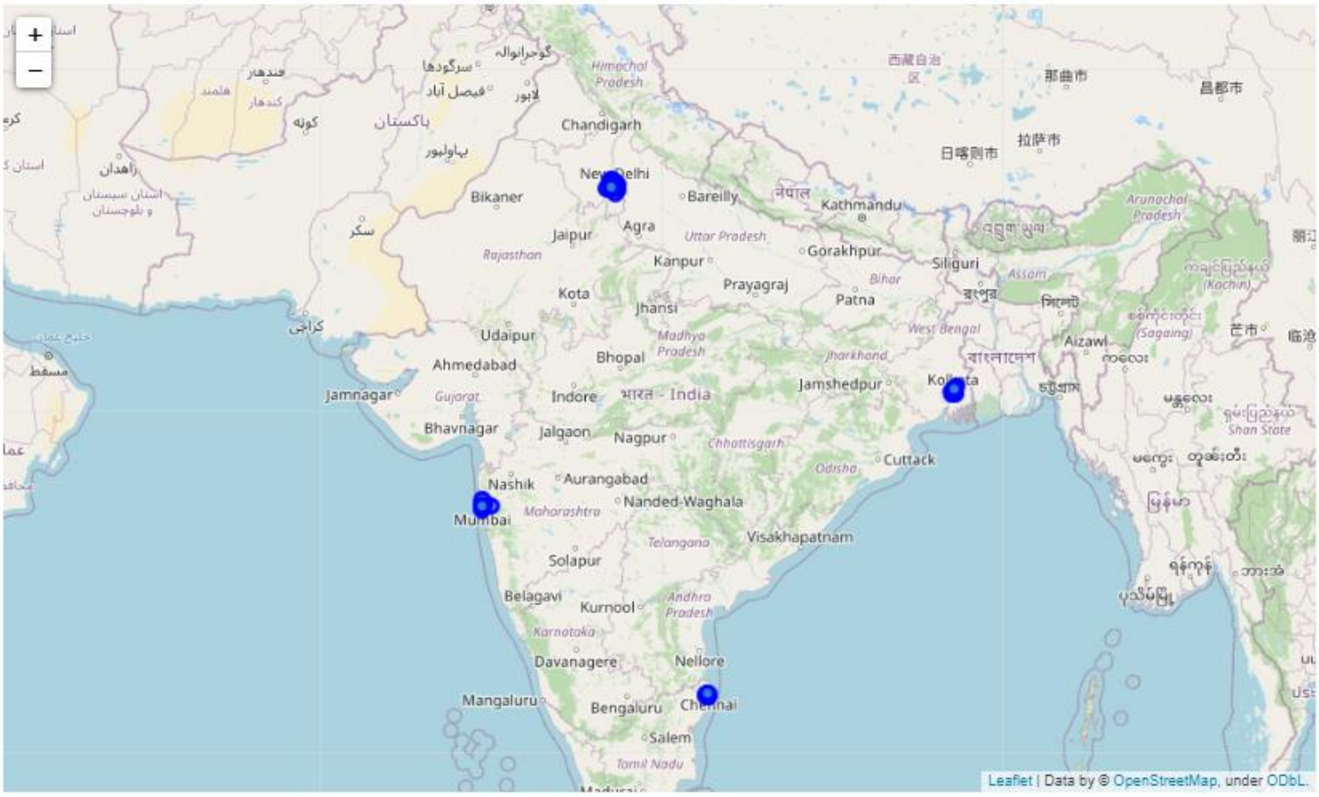
We will use one-hot encoding for the venues category and group the data by their neighborhood. For analyzing the data, we will extract top 10 most venues of the data.

3.2. Modelling the data

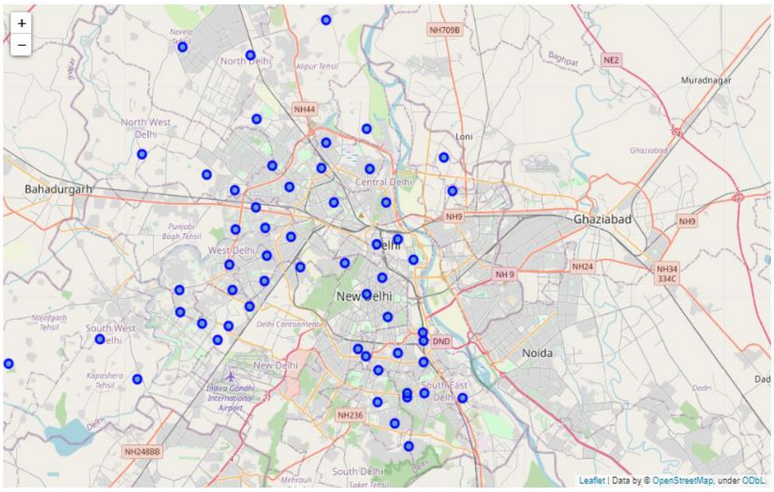
We will use k-means clustering algorithm to cluster the venues in each city in five clusters.

3.3. Visualizing the data

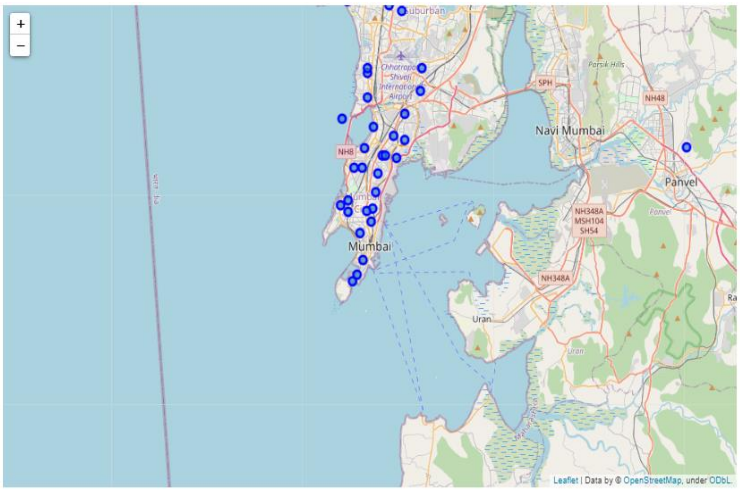
We will use folium library to visualize the map.



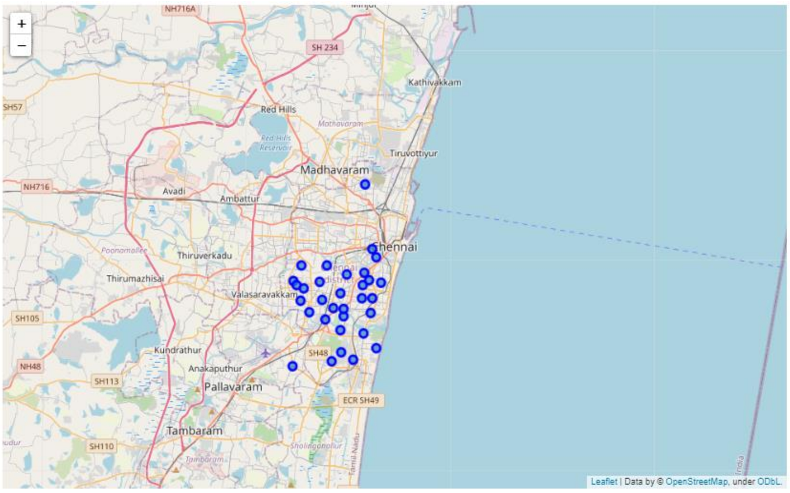
Map of India showing the four metro cities.



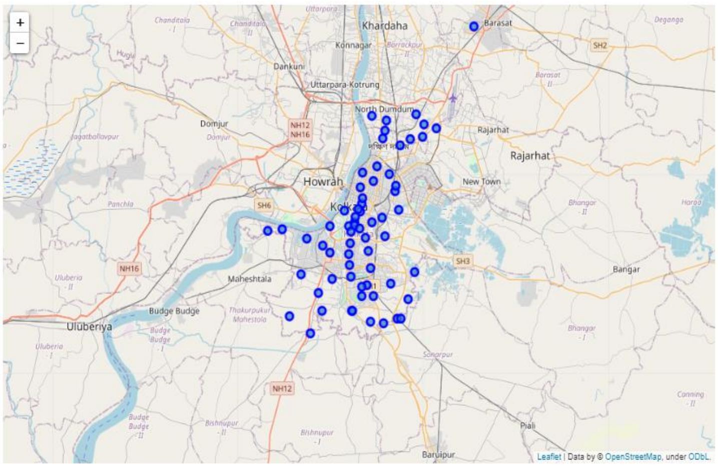
Delhi



Mumbai

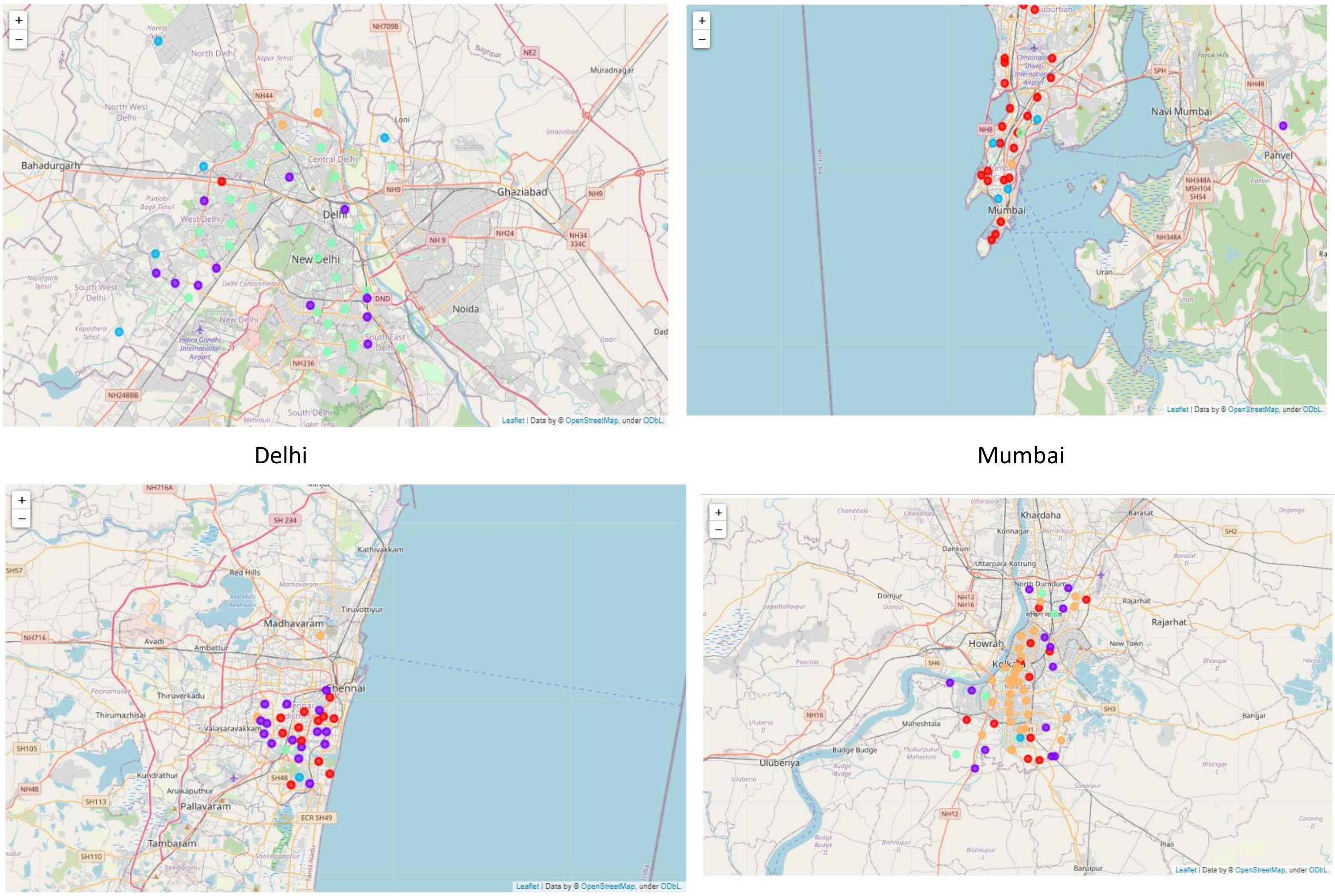


Chennai



Kolkata

Maps of four metro cities showing different neighborhoods plotted on them.



Delhi

Mumbai

Chennai

Kolkata

Maps of four metro cities showing different neighborhood clusters.

4. Results and Discussion

In this project, we attempted to load dataset of India's metropolitan cities and try to analyze neighborhood in these metro cities based on most popular venues they have. We used k-means clustering algorithm to cluster the neighborhood.

The main aim of this project was to help people to relocate or settle in these metro cities. Given the cluster information of all metro cities.

We can say that **Delhi** people are good for art, craft, playgrounds and malls. Also, Delhi contains Indian Restaurants and Fast food places. So, it might not be good international visitors.

In **Mumbai** Cafe, Pub, Gym and Spa are quite famous here. These people are more health conscious. Also due to close proximity to sea shore, fish markets are also famous there.

Chennai and its neighborhoods are a great place for foodie. There are various types of restaurants and hotels in Chennai.

Kolkata and its neighborhoods contain Shopping malls, Multiplex, ATMs etc. It contains Park, Gyms, Cafe and Hotel too.

We can also analyze each neighborhood cluster wise to find the best neighborhood for targeted persons.

5. Conclusion

This project helps us to get better understandings of neighborhoods with respect to most common venues. For example, people who are foodies and like to taste different foods should search for neighborhood in Chennai etc.

The future of this project can be to include data related to crimes, pricing and more to get better insights of the neighborhoods and suggest better neighborhood.