MAY 2020

Mumbai

Analysing Neighbourhoods in the Suburbs and the City

Submitted by: **Abhiraj Rao**

IBM Data Science Professional Certificate Capstone Project

Introduction

Mumbai (also known as Bombay, the official name until 1995) is the capital city of the Indian state of Maharashtra. It is the financial, commercial, and entertainment capital of India. Mumbai is the most populous city in India, and the seventh most populous city in the world with a population of over 20 million.

The city of Mumbai consists of two distinct regions: Mumbai City district and Mumbai Suburban district, which form two separate revenue districts of Maharashtra. The Mumbai Suburban Railway, popularly referred to as Locals forms the backbone of the city's transport system. It is operated by the Central Railway and Western Railway zones of the Indian Railways.

Mumbai's culture is a blend of traditional festivals, food, music, and theatres. The city offers a cosmopolitan and diverse lifestyle with a variety of food, entertainment, and night life, available in a form and abundance comparable to that in other world capitals. Mumbai's history as a major trading centre has led to a diverse range of cultures, religions, and cuisines coexisting in the city.

Problems to be Addressed

Mumbai suffers from the same major urbanisation problems seen in many fast growing cities in developing countries: widespread poverty and unemployment, poor public health and poor civic and educational standards for a large section of the population. With available land at a premium, Mumbai residents often reside in cramped, relatively expensive housing, usually far from workplaces, and therefore requiring long commutes on crowded mass transit, or clogged roadways. Many of them live in close proximity to bus or train stations although suburban residents spend significant time travelling southward to the main commercial district.

The number of households in Mumbai is forecast to rise from 4.2 million in 2008 to 6.6 million in 2020. As the city grows and develops, it becomes increasingly important to examine and understand it quantitatively. This is important not only for business owner, entrepreneurs, and developers, but also the general population as well as people looking to move into this large and diverse city.

Exploring the neighborhoods of Mumbai in an attempt to classify them as well as possible can be a huge boon to the general public, and help us in identifying the underlying patterns, as well as potential business opportunities.

DATA

The following data was used in the project:

- **List of neighbourhoods in Mumbai from Wikipedia:**

  https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai

- **Venues geo-data using the Foursquare developer API. This was used to get a detailed list of venues present around the neighbourhoods:**

  https://developer.foursquare.com/

Methodology

The following action steps were taken in the project completion process:

1 Scraping the Wiki data

After getting the source of the Wiki data using *requests*, we use the *BeautifulSoup* library to 'scrape' the data from the webpage. This is stored in a *pandas* DataFrame, and is the base of our main data.

Given the slightly unreliable nature of data from Wikipedia, some of the entries manually need to be replace or deleted from the DataFrame.

2 Getting the geodata

Using the *geopy* library in Python, mainly the *Nominatim* module, we pass the list of neighbourhoods to get the latitudes and longitudes for each entry, and append it to the main dataset. Inaccuracies in the *Nominatim* geodata is corrected, and the neighbourhoods without any geodata even after data cleaning and manipulation are 'dropped' from the dataset.

3 Getting the venues data

Each neighbourhood, along with its geolocation data is passed to the *Foursquare API* using our unique developer credentials and a specific API link for the **GET** request. This inclusive list of all nearby venues in Mumbai will be stored in a new DataFrame, on which we execute data manipulation and analysis.

4 Data Cleaning

This is a crucial step in making sure our data has no inaccuracies or redundancies that will cause problems in further operations or Machin Learning implementations.

Firstly, we drop the duplicate venue entries. Due to there being inconsistencies in the Foursquare venue data, mainly the 'Category' values, we will check the unique categories present and assign them to our own custom Category list, one that is a bit more generalise and consistent, and can make our features more robust. This is added to a new column and the old values are discarded, along with incorrect/unnecessary venues.

5 Data Pre-processing

We use one-hot encoding to turn the categorical variable 'Category' into columns of its values as binary representations. 1 if the venue belongs to that category, otherwise 0. We group this DataFrame by the neighbourhoods, averaging the encoded categorical variable values. This in turn, gives us a dataset with the frequencies of venue categories in each neighbourhood.

We use the above data to form a new data set which has the neighbourhoods, and their ten most common venue categories. This is our final dataset which will be used for classification.

5 K-Means Clustering

Our classification algorithm of choice will be the K-Means Clustering Algorithm, implemented using the *scikit-learn* library. Using both the Elbow Method and Silhouette Score Method to find the optimal value of K.

K-means clustering is run on our data and we get 5 clusters. We represent these clusters on a map as well as print each one out to explore the results, and name and describe each cluster.

Results

Based on further exploration into the clustering results, we observe these characteristics of our five clusters.

Cluster '0':
This is the most populated cluster out of all. Most of the neighbourhoods have more than half of the 10 most common venues as eateries, generally small-scale Indian restaurants, for cheap and quick meals, as well as fast food chains and street food stalls. The most characteristic thing about this cluster is that these spots all lie very close to the Western and Central Railway Lines of the city, more so than other neighbourhoods.

Cluster '1':
Neighbourhoods with Fitness Centres like Club Houses, Gyms, Yoga Studios, etc, as the most common venues, while having other amenitues also available. Train and Bus stations in the vicinity for good connectivity. There is a characteristic lack of eateries in the list of common venues among these neighbourhoods.

Cluster '2':
This cluster has neighbourhoods having bars, pubs, nightclubs, and cafes as the most common venues. This includes the famous Parsi/Irani cafes in the Western Suburbs. This cluster has a larger spread of multi-cuisine restaurants available as compared to the rest. Non-eateries are mainly sports and fitness centres, clothing stores, and arts and entertainment centers. Notably, it is more concentrated in the Southern Suburbs and Mumbai City area.

Cluster '3':
This cluster has neighbourhoods nearest to Train Stations. Interestingly, as a consequence of this, the next most common venue category is street food stalls, supplying the needs of the busy, fast-moving crowd of the city.

Cluster '4':
Neighbourhoods having Parks in the top 3 most common venues are present in this cluster. These neighbourhoods are present mainly in the northern part of Mumbai, i.e. the Suburban region.

Discussion and Conclusion

This project forms a sufficient baseline to the problem of categorising neighbourhoods in a city as diverse as Mumbai. As developed as the city may be, the developing status of the country and its populace means there is not as much data, especially crowd-sourced like that of Foursquare, available to perform much more in-depth classifications. It would be interesting to see how more detailed and accurate data, and more feature sets would result in.

The clustering results show how the different neighbourhoods' characteristics are so interdependent with the transport infrastructure of the city. With the lower and middle class populace being the majority of the users of public transit, the neighbourhoods have shaped themselves to follow the trends of day to day flow of the workforce.

Mumbai City and the Southern part of the Suburbs boasts a large concentration of commercial and industrial buildings, while the Northern part of the suburbs is full of residences. This infrastructural diversity can be seen to impact the nature of the neighbourhoods.

Neighbourhoods with 'more developed' venues like expensive eateries, nightclubs, arts and fitness centers, are clustered in the City area, being more condensed. While in the Suburban region, more affordable food options, and more open spaces like parks and sports grounds are available, with less congestion.

The Railway lines can be seen, aspecially visually on ur map, as the 'backbone' of the city, with the largest cluster characteristically having neighbourhoods along its path, as if being dependent on it.

The airport appears to case a big 'divide' between the South and the North, and there have not yet been any neighbourhood clusters that seem to co-depend with it, as one might expect from a developed city.

These preliminary clusters can serve as a tool for both individuals as well as companies, who wish to set up base in the city in accordance with their needs.