Coursera Capstone

Opening a New Bar in Mumbai City, India

Introduction

For many extroverts, visiting Pubs and Bars is a great way to relax and enjoy themselves during weekends and holidays. They can meet new people, meet with their old friends and perform many more activities. Bars are like a one-stop meeting and hangout spot. For owners, the central location and the large crowd at the Bar provides a great distribution channel to market their services. Investors also taking advantage of this trend to build more Bars to cater to the demand. As a result, there are many bars are built in the city of Mumbai and many more are being built. Opening Bars allows Owners to earn consistent rental income.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Mumbai to open a new Bar. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Mumbai, If an investor wants to open a Bar as an Investment, where would you recommend that they open it?

Data

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

- List of neighbourhoods in Mumbai. This defines the scope of this project, which is confined to the city of Mumbai.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and to get the venue data.
- Venue data, particularly data related to Bars. We will use this data to perform clustering on the neighbourhoods.

Methodology

Firstly, we will curate a neighborhood data for the city of Mumbai. It is available on the Wikipedia Page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai).

We will do web scraping using Python requests and beautiful soup packages to extract the list of neighbourhood 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.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters.

We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a python loop. With the data, we can check how many venues were returned for each neighbourhood. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of Bar in a neighborhood. By doing so, we are also preparing the data for use in clustering.

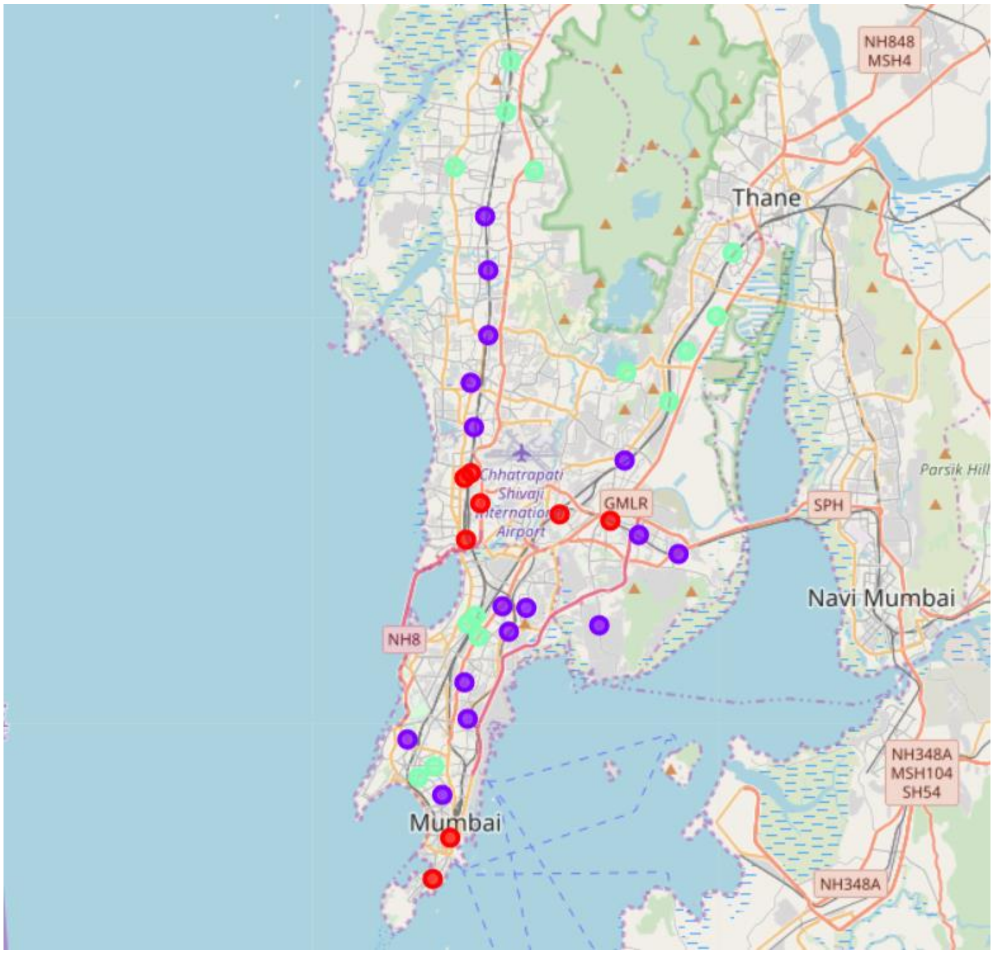
We will perform clustering by using k-means clustering. K-means clustering algorithm. It is an unsupervised machine learning algorithm and is particularly suited to solve the problem for this project, it will help us to answer the question as to which neighbourhoods are most suitable to open new Bar.

Result

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

- Cluster 0: Neighbourhoods with moderate number of bars.
- Cluster 1: Neighbourhoods with high number of bars.
- Cluster 2: Neighbourhoods with highest concentration of bars.

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in blue colour, and cluster 2 in green colour.



Discussion

As observations noted from the map in the Results section, most of the bars are concentrated in the central and southern part of the Mumbai city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number of Bars in the neighbourhoods. This represents a great opportunity and high potential areas to open new Bar as there is no competition from existing Bars. Whereas opening a Bar in central or southern part of Mumbai will suffer due to high competition.

Limitations and Suggestions

In this project we only used the frequency of bars present in a neighborhood which does not give the whole picture.

There can be other factors affecting the success of a new bar like the per capita income of the neighborhood, the age group of people who live in the neighborhood and many other.

In addition to that this project made use of a free developer API by foursquare which provides a limited call to API which works well for development. 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 and then based on the insights provided by clustering we recommended the possible solution to the business problem.

References

Suburbs in Mumbai city Wikipedia. Retrieved from https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs