Coursera Capstone

IBM Data Science Capstone Project

Opening a Pub in Mumbai

Yatharth Aggarwal
April, 2020

Introduction

A pub, or public house, is an establishment licensed to sell alcoholic drinks, which traditionally include beer (such as ale) and cider. It is a social drinking establishment and a prominent part of British, Irish, Breton, New Zealand, Canadian, South African and Australian cultures] In many places, especially in villages, a pub is the focal point of the community. In his 17th-century diary, Samuel Pepys described the pub as "the heart of England".

But now it has spread to almost all countries across the globe. Developing countries like India with its huge population and once a colony of Britishers has definitely moved into this trend. Westernization and night life is a common trend amongst both the upper and middle class of the society.

Mumbai is the economic capital of India. The Richest man in Antila to the Poorest man in Dharavi live in this city. Thus this city have a complete diversity among the social and economic groups, making a perfect scenario of every type of recreational facility to be built.

As a result, there are many pubs in the Mumbai and many more are being built. Opening pubs allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new pub requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the pub is one of the most important decisions that will determine whether it will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse Pub clusters in the city and select the best locations in Mumbai, India to open a new pub.
Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Mumbai, India, if a property developer is looking to open a new pub, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new pubs or similar structure in Mumbai. Commoners are also stake holder for any public space being built around them, thus they are also a target audience.

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 Mumbai city.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to pubs. We will use this data to perform clustering on the neighbourhoods.

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 40 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful soup 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 Pub 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 neighbourhoods in the city of Mumbai. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Mumbai). We will do web scraping using Python requests and beautiful soup packages to extract the list of neighbourhoods 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 neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Mumbai.

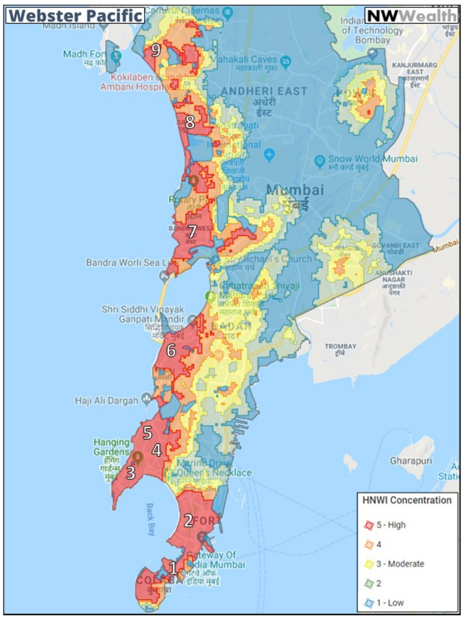

Figure I: Distribution of Rich and Poor in Mumbai

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 neighbourhoods 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 neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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 analysing the "Pub" data, we will filter the "Pub" as venue category for the neighbourhoods.

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 neighbourhoods into 3 clusters based on their frequency of occurrence for "Pub". The results will allow us to identify which neighbourhoods have higher concentration of pubs while which neighbourhoods have fewer number of pubs.

Based on the occurrence of pubs in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new pubs.

Results

The initial results from Wikipedia provided 42 suburban areas of Mumbai, namely: Andheri, Anushakti Nagar, Baiganwadi, Bandra, Bhandup, Borivali, Charkop, Chembur, Dahisar, Devipada, Dombivli, Eastern Suburbs, Ghatkopar, Goregaon, Grant Road, Jogeshwari, Juhu, Kalyan, Kandivali, Kanjurmarg, Kausa, Kurla, Mahavir Nagar (Kandivali), Mankhurd, Matharpacady, Mumbai, Mira Road, Mogra Village, Mulund, Mumbra, Pestom sagar, Seven Bungalows, Shil Phata, Sion Mumbai, Sonapur, Bhandup, Thakur village, Tilak Nagar (Mumbai), Uttan, Vashi, Vikhroli, Wadala, Western Suburbs (Mumbai) and Worli

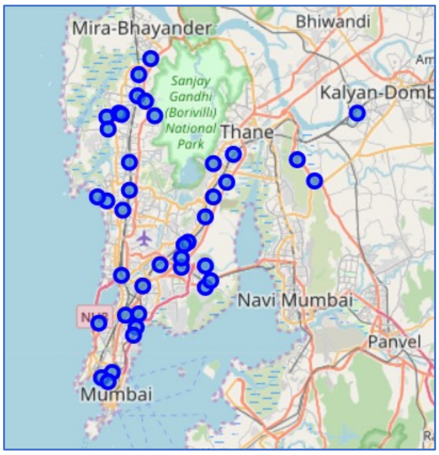

Figure II: 42 Suburbs of Mumbai

The results from the k-means clustering show that we can categorize the neighbourhoods into 5 clusters based on the frequency of occurrence for "Pub":

- Cluster 0: Neighbourhoods with high number of pubs
- Cluster 1, 2: Neighbourhoods with very low number to no existence of pubs
- Cluster 3,4: Neighbourhoods with more than 5 pubs

The results of the clustering are visualized in the map below

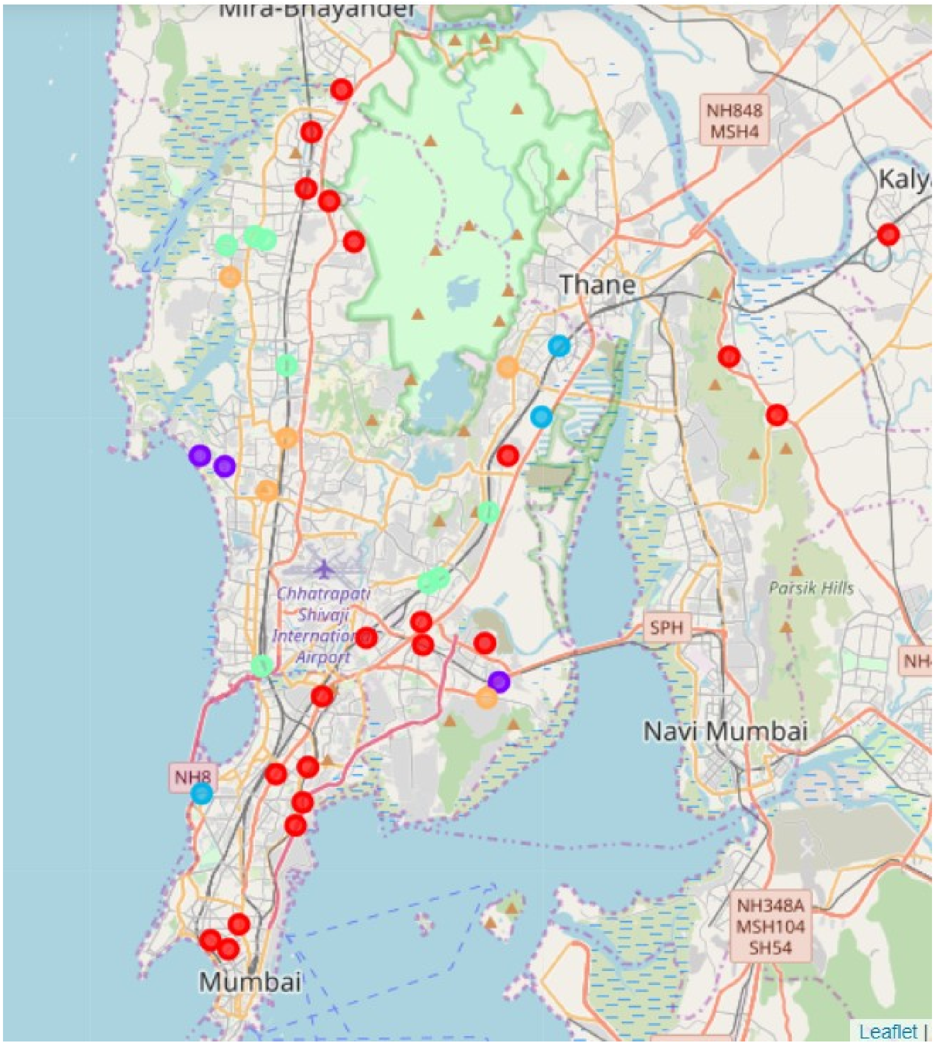

Figure III: Clustered neighbourhoods of Mumbai

Discussion

As observations noted from the map in the Results section, most of the pubs are concentrated in the central area of Mumbai city, with the highest number in cluster 0 and moderate number in cluster 3 and 4. On the other hand, clusters 1 and 2 has very low number to no pub in the neighbourhoods.

This represents a great opportunity and high potential areas to open new pubs as there is very little to no competition from existing pubs. Meanwhile, pubs in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of pubs. From another perspective, the results also show that the oversupply of pubs mostly happened in the central area of the city, with the suburb area still have very few pubs.

Therefore, this project recommends property developers to capitalize on these findings to open new pubs in neighbourhoods in cluster 3 and 4 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new pubs in neighbourhoods in cluster 1 and 2 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 0 which already have high concentration of pubs and suffering from intense competition.

Therefore, venturing in those places will not just give business dividends but also a quality life to people living in there.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of pubs, there are other factors such as population and income of residents that could influence the location decision of a new pub. However, to the best knowledge of this researcher such data are not available to the neighbourhood 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 pub. 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 5 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 pub. To answer the business question that was raised in the introduction section, the answer proposed by this project is:

The neighbourhoods in cluster 3 and 4 are the most preferred locations to open a new pub. 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 pub.