Coursera Capstone

IBM Applied Data Science Capstone

Opening a new Bakery in Mumbai, India

By: Ameya Thosar

July 2020

Introduction

Bakeries are an important part of our life as on occasion of Birthdays, Anniversary, a normal day etc we often go to bakery for buying Cakes, cookies, desserts. Having a Bakery in nearby location is good. In Mumbai there are many new bakeries but when you want to start your own you need to look for a location where in neighborhood there are minimum number of bakeries. Each bakery typically offers breads (bagels, buns, rolls, biscuits and loaf breads), cookies, desserts (cakes, cheesecakes and pies), muffins, pizza, snack cakes, sweet goods (doughnuts, Danish, sweet rolls, cinnamon rolls and coffee cake) and tortillas. So opening a new bakery is a good decision as people love to visit them.

Business Problem

The objective of this capstone project is to analyse and select the best location in Mumbai, India to open a new Bakery. Using Data Science Methodology and Machine Learning techniques like clustering, this project will provide a solution to the problem. In Mumbai, India if anyone wants to open a new Bakery we can suggest him where he can open it. Because when you are new in business its better that you have less competitors so that your business can grow at faster rate.

Target Audience

In this project we clearly aim at the person who want to open a new Bakery. Using this we will help him to choose the best location.

Data

To solve this problem we need following data:

- List of Neighborhood's in Mumbai.
- Latitude and Longitudes of those neighborhood's. This will be required to get plot the map and get venue data.
- Venue data particularly related to Bakery. We will use this data for clustering.

Source of data and method to extract it:

The Wikipedia page that contains a list of neighborhoods in Mumbai, (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai) with a total of 40 neighborhood's. We will use web scraping techniques to extract the data from the Wikipedia page. By this we will only get the names of the neighborhood's but with tat we also need their coordinates. For that we will use Python page, with the help of Python requests and beautiful soup Geocoder package which will give us the latitude and longitude coordinates of the neighborhood.

After that, we will use Foursquare API to get the venue data for those neighborhood's. 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 Bakery category in order to help us to solve the business problem put forward. This capstone project 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).

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_of_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 of places in Mumbai. 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 Geocoder package that will allow us to convert address into geographical coordinates i.e. 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 check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of 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. For this Foursquare developed account need to be created by using this we can make certain number of calls to the API and use it in project. 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 are there 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

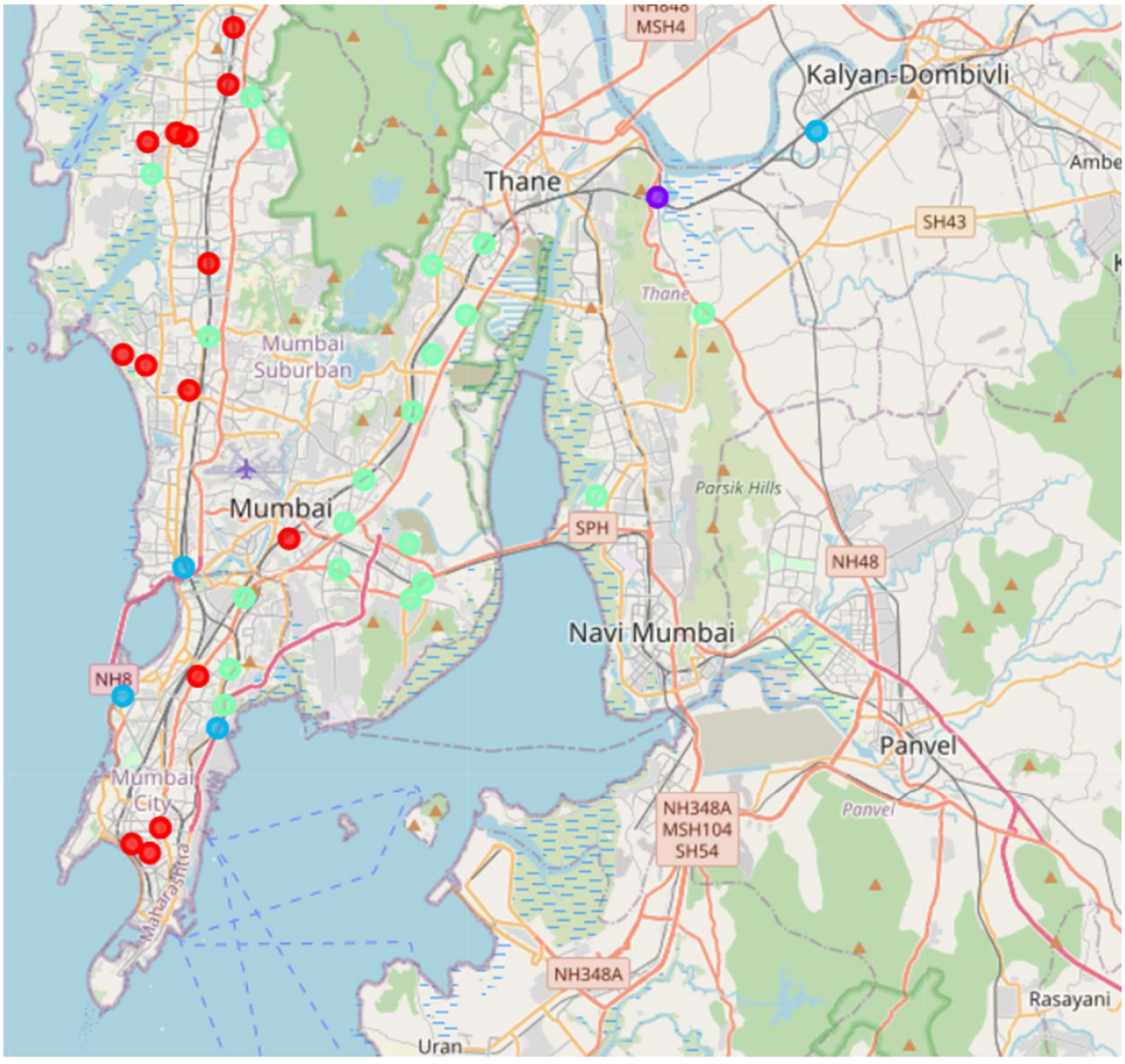
“Bakery” data, we will filter the “Bakery” as venue category for the neighbourhoods.

At last 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 5 clusters based on their frequency of occurrence for “Bakery”. The results will allow us to identify which neighbourhoods have higher concentration of bakeries while which neighbourhoods have fewer number of bakeries. Based on the occurrence of bakeries in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Bakery.

Results

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

- Cluster 0, 3: Neighbourhoods with high number of shopping malls
- Cluster 2: Neighbourhoods with moderate number to no existence of shopping malls
- Cluster 1,4 : Neighbourhoods with only one shopping malls



Discussion

As observations noted from the map shown in the Results section, higher concentration of Bakeries is seen in cluster 0 and cluster 3 and moderate concentration in cluster 2 . On the other hand, cluster 1 and cluster 4 has very low number of bakeries in the neighbourhoods. This represents a great opportunity and high potential areas to open new Bakery as there is very little to no competition from in that area. Meanwhile, Bakeries in cluster 0 and cluster 3 are likely suffering from intense competition due to high concentration of Bakeries. Therefore, this project recommends property developers to capitalize on these findings to open new bakery in neighbourhoods in cluster 1 and cluster 4 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new Bakeries in neighbourhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 0 and cluster 3 which already have high concentration of Bakeries and suffering from intense competition.

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 Bakery. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 10 and cluster 3 are the most preferred locations to open a new Bakery. 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 Bakery.