Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Cafe in Mumbai, India

By: Nishchay Nagpal

July 2020

INTRODUCTION

When I travel, wherever I travel, more than museums, monuments, temples and, even restaurants, I seek out a café.

The reasons are many. Nowhere quite represents a place like its cafés. What would Vienna, Rome and Paris be without theirs? It's not just the history or the opulence. It's the window on the world. You see ordinary people – locals as well as visitors – coming and going, taking time out between work and home, meeting friends or colleagues, joining families. If they're not right beside you in the café, you watch them on the street or from your shaded terrace.

I, myself am a huge caffeine fanatic which led me to a question of where can one open a Café shop in Mumbai, the iconic city of India.

BUSINESS PROBLEM

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

TARGET AUDIENCE

The target audience for this project would be twofold. Firstly, caffeine enthusiasts visiting Mumbai to get to know which neighborhoods have higher number of Cafes. Secondly, a company or a person looking to open a café shop in Mumbai can use the information provided here.

DATA

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

- List of neighborhoods in Mumbai: This defines the scope of this project which is confined to the city of Mumbai, the iconic city of India
- Latitude and longitude coordinates of those neighborhoods: This is required in order to plot the map and to gather the venue data from foursquare for the neighborhoods.
- Venue data, particularly data related to café shops. We will use this data to perform clustering on the neighborhoods

Sources of data and methods to extract them:

The Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai) contains a list of neighborhoods in Mumbai, with a total of 41 neighborhoods. We will be using web scraping techniques utilizing the python packages - requests and beautifulsoup, to extract the data from the Wikipedia page. Then we will be getting the geographical coordinates of the

neighborhoods using the Python Geocoder package. After that, we will be using the Foursquare API to get the venue data for all neighborhoods. Foursquare has one of the largest databases of 100+ million places and is used by over a large number of developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Cafe 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 neighborhoods 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 web scrape the URL using Python requests and beautifulsoup packages to extract the list of neighborhoods data. Now, we need to get the geographical coordinates i.e. latitude and longitude in order to be able to use Foursquare API. We will use the awesome Geocoder package that will allow us to convert neighborhood list 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 neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by

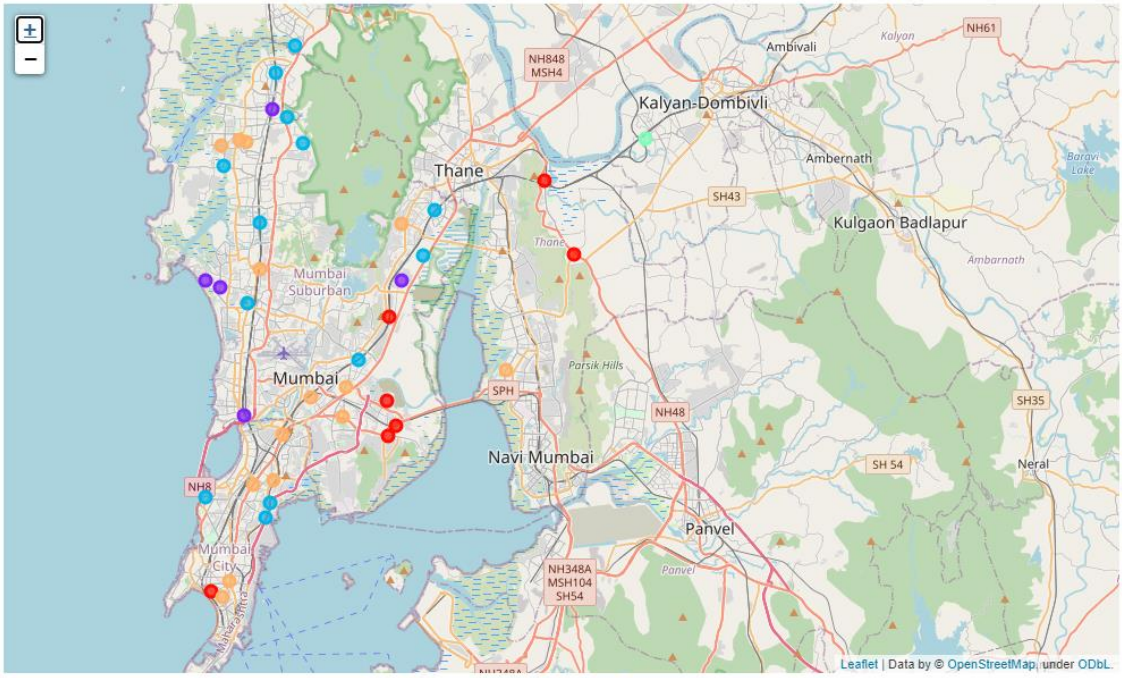
Geocoder are correctly plotted in the city of Mumbai. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key with which we can make API calls to Foursquare passing in the geographical coordinates of the neighborhoods 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 neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood 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 analyzing the "Cafe" data, we will filter the "Cafe" as venue category for the neighborhoods. 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 neighborhoods into 5 clusters based on their frequency of occurrence for "Cafe". The results will allow us to identify which neighborhoods have higher concentration of Cafe while which neighborhoods have fewer number of Cafe. Based on the occurrence of Cafe in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Café.

RESULTS

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

- Cluster 0: Neighborhoods with low number of Cafe
- Cluster 1: Neighborhoods with moderate number to no existence of Cafe
- Cluster 2: Neighborhoods with high concentration of Cafe
- Cluster 3: Neighborhoods with highest concentration of Cafe
- Cluster 4: Neighborhoods with moderate concentration of Cafe

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint skin color, cluster 3 in mint green color, cluster 4 in light blue.



LIMITATIONS OF THE PROJECT

In this project, we only consider one factor i.e. frequency of occurrence of cafes, there are other factors such as population and income of residents that could influence the location decision of a new cafes. 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 shopping mall. In addition, this project made use of the free Personal 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

Most of the Cafes are concentrated in the southern and eastern area of Mumbai, with the highest number in cluster 2 and moderate number in cluster 4. On the other hand, cluster 0 has very low number to totally no Cafes in the neighborhoods. This represents a great opportunity and high potential areas to open new Cafes as there is very little to no competition from existing shops. Meanwhile, cafes in cluster 3 are likely suffering from intense competition due to oversupply and high concentration of cafes. From another perspective, this also shows that the oversupply of cafes mostly happened in the central area of the city, with the suburb area still have very few cafes. Therefore, this project recommends developers to capitalize on these findings to open new cafes in neighborhoods in cluster 0 with little to no competition. Developers with unique selling propositions to stand out from the competition can also open new cafes in neighborhoods in

cluster 2 and cluster 4 with moderate competition. Lastly, developers are advised to avoid neighborhoods in cluster 3 which already has high concentration of cafes and is suffering from intense competition.

REFERENCES

- https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai
- https://www.telegraph.co.uk/travel/food-and-wine-holidays/best-cafes-in-the-world/
- https://developer.foursquare.com/docs