Best Neighborhood for Starting Coffee Shop in Kolkata

Mohammad Omar Faruque

June 09, 2020

1. Introduction

Kolkata is known as 'the city of joy'. It is the city that celebrates Indian culture and traditions throughout the year. This former capital of British rule is known for its delicious food & dessert, beautiful temples, and enlightening museums, mesmerizing parks & gardens, along with giving a glimpse of imperial India. Kolkata is one of the largest metropolitan area in the world. Around 14 million people live here. Kolkata is a pioneer in the field of drama, arts, theatre and literature with several nobel laureates contributing to the Kolkata culture etc. Talking about cuisine and restaurants Kolkata holds title being one of food heaven in India. Kolkata is also famous for coffee houses from British rule. Some best places for having coffee are: 95 Degree Cafe & Bakery, Slay Coffee, Wood House, Barista etc. College Street Coffee House is the oldest and famous coffee house in Kolkata since the British period. In this report , we will talk about Coffee Shop and opening Coffee shop at suitable place in Kolkata.

Problem:

There are many neighborhoods in Kolkata. Different types of restaurants and food shops number varys Neighborhood to Neighborhood. In some neighborhoods, there are many coffee shops compared to other types of cafe’s, restaurants and there also exists other neighborhoods where the picture is the exact opposite. So it's not easy to determine where to start a new coffee shop in the right neighborhood where the coffee house will make good business and make good profit. This problem is for businessmen or shop owners or new startup platforms who want to find a suitable neighborhood to start their coffee shop business.

Solution to Problem:

So to start a new coffee shop in a particular neighborhood or to start a specific type of cafe, we have to analyze the neighborhoods data related to its restaurants and coffee shops with machine learning and data science to determine best possible suited neighborhoods to start a coffee shop.

2. Data Collection and analyze

To solve our given problem, the 2nd step is to collect data or data acquisition. For this step, we need to collect data about different neighborhoods of Kolkata. Kolkata is a big city with over 100 plus neighborhoods, so first thing we have to do is to collect the names of neighborhoods in Kolkata city. For this:

- List of neighbourhoods in Kolkata. This defines the scope of this project which is confined to the city of Kolkata, the capital city of the country of India in South Asia.
- 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 shopping malls. 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:Neighbourhoods_in_Kolkata contains a list of neighbourhoods in Kolkata, with a total of 70 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup 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 Shopping Mall 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.

3. Methodology section & exploratory data analysis

Firstly, we need to get the list of neighbourhoods in the city of Kolkata. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Kolkata). We will do web scraping using Python requests and beautifulsoup 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 Kolkata, India.

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 “Coffee Shop” data, we will filter the “Coffee Shop” 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 “Coffee Shop”.

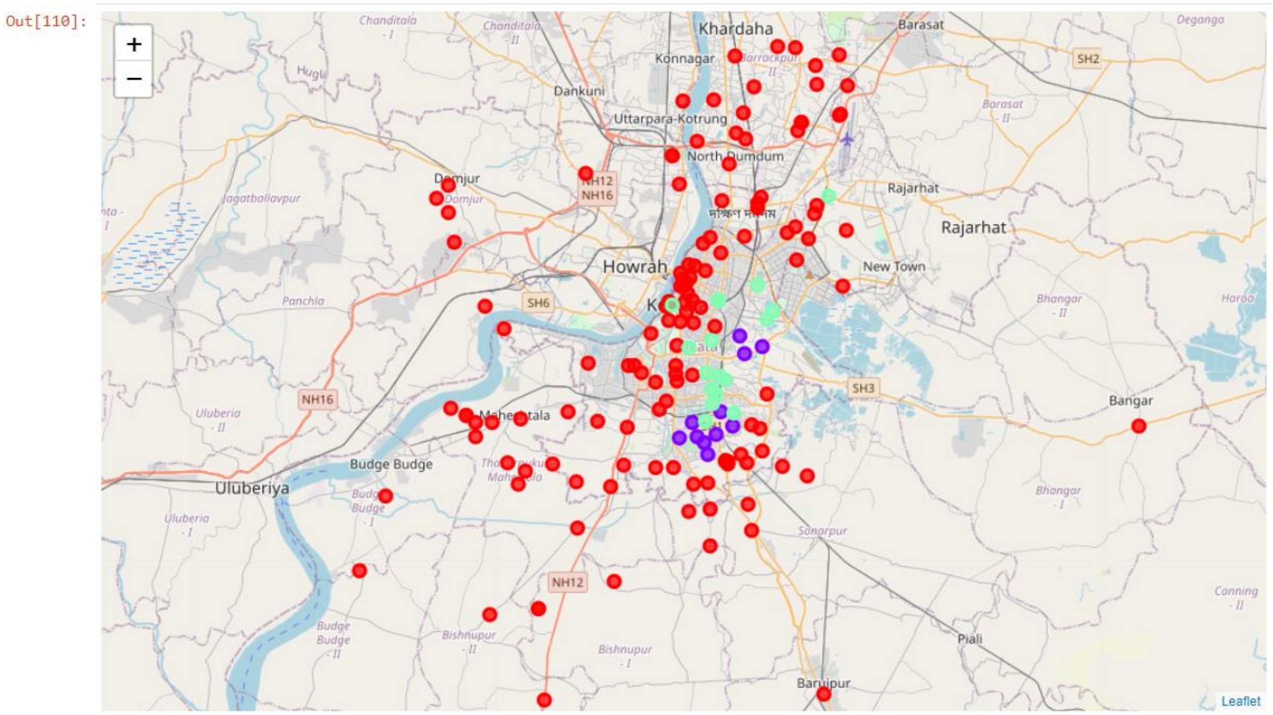
The results will allow us to identify which neighbourhoods have higher concentration of coffee shops while which neighbourhoods have fewer number of coffee shops. Based on the occurrence of coffee shops in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new coffee shop.

4. Result Section

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

- Cluster 0: Neighbourhoods with low number of coffee shop.
- Cluster 1: Neighbourhoods with moderate number to no existence of coffee shop.
- Cluster 2: Neighbourhoods with highest concentration of coffee shop.

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


Out[110]:

5. Discussion

As observations noted from the map in the Results section, most of the coffee shops or houses are concentrated in the central area of Kolkata city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to no coffee shop in the neighbourhoods. This represents a great opportunity and high potential areas to open new coffee shop as there is very little to no competition from existing coffee shops. Meanwhile, coffee shop in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of coffee shops. From another perspective, the results also show that the oversupply of coffee shops mostly happened in the central area of the city, with the suburb area still have very few coffee shops. Therefore, this project recommends entreprenuers to capitalize on these findings to open new coffee shops in neighbourhoods in cluster 0 with little to no competition. Entreprenuers with unique selling propositions to stand out from the competition can also open new shopping malls in neighbourhoods in cluster 1 with moderate competition. Lastly, entreprenuers are advised to avoid neighbourhoods in cluster 2 which already have high concentration of coffee shops and suffering from intense competition.

6. 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 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. entreprenuers, investors, interested one’s regarding the best locations to open a new coffee shop. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new coffee shop. The findings of this project will help the relevant entreprenuers or investors to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new coffee shops.