IBM Applied Data Science Capstone Project

Kalaivanan.S

13-February-2020

Setting up a Coffee shop in Chennai, India

Objective:

The objective of this project is to analyze a suitable location to open up a coffee shop in Chennai, India.

Introduction:

Chennai is a metropolitan, capital city located in the southern state of Tamilnadu, India. Being a cosmopolitan city, people from all over India stay and work in this city, the major employers being Information Technology industry and Automobile sector. In addition to the people living there, foreign tourists flock this place all through the year.

This has led to a spur in the food and hospitality industry. People throng to the various cuisines located in the suburbs of Chennai on a daily basis. The most common place visited by people of all cultures (both domestic and foreign) would be a nice coffee shop.

Business Problem:

This project analyzes the neighborhoods present in Chennai, and the distribution of coffee shops in these neighborhoods. The final objective would be to identify a suitable place to open up a new coffee shop in the suburbs of Chennai. This project clusters the neighborhoods according to the number of coffee shops present in each neighborhood. Web scraping and machine learning techniques are utilized to arrive at a conclusion.

Data required:

The below mentioned Data would be required to solve the objective

- The list of neighborhoods in the city of Chennai.
- The geographical coordinated of the neighborhoods
- The list of venues, particularly coffee shops located in these neighborhoods

Data Sources:

The neighborhoods data of Chennai is scraped from the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Chennai). The list contains a total of 61 neighborhoods located in Chennai. BeautifulSoup package is used to scarp the data from Wikipedia page and store the neighborhoods in the dataframe.

Geocoder from geopy package is used to extract the geographical coordinates of the stored neighborhoods. The latitudes and longitudes are append to the neighborhoods dataframe.

Foursquare API is used to get the details of trending venues for the corresponding neighborhoods in the dataframe. Among the trending venues, the coffee shop venues are segregated and clustering algorithms are applied to find out a suitable place for opening up a new coffee shop. The methodology and inference are detailed in the next part of the documentation.

Methodology:

The packages required for execution of the project are imported in python. To analyze the neighborhoods in Chennai, the list of neighborhoods must be imported to a dataframe. This data is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Chennai). This page has information about 61 neighborhoods located in the city of Chennai, India. Web scraping is done via python using BeautifulSoup package and stored in a pandas dataframe. The dataframe consists only the names of neighborhoods. The next step would be to extract the geographical coordinates of the stored neighborhoods.

The Geocoder package is utilized for this purpose. The extracted coordinates are appended to the neighborhood dataframe to include the corresponding location information of the neighborhoods. Folium package is used to visualize the neighborhoods in a map.

The next step would be to get the details of commercial establishments located in the neighborhood. FourSquare API is used to extract this information. The top 100 venues located within a radius of 2 kilometer is extracted using foursquare API. The names of the business establishment, and the category to which it belong are extracted and appended to the neighborhoods dataframe. A total of 147 unique categories are found to be present in the resulting dataframe. Onehot encoding is used to group the neighborhoods according to the category.

The objective of the project is to find a suitable place to set up a coffee shop. The frequency of the venues are analyzed and the mean is stored. A new dataframe named as coffe shops consisting the mean of coffee shops in each neighborhood is created.

k-means clustering is applied to group the neighborhoods by the presence of coffee shops. A cluster size of 5 is used for this project. The resulting dataframe consists of the name of each neighborhood, the occurrence of coffee shops, and the respective cluster to which it belongs. The latitude and longitude information is appended to this pandas dataframe. This information is utilized to visualize the clusters on the map of Chennai, India. Figure 1 is the map depicting the information about the location of coffee shops in the neighborhoods of Chennai, clustered according to the frequency of coffee shops in each location.

Results:

The clustering algorithm has clustered the neighborhoods into a total of five clusters based on the frequency of occurrence of coffee shops in each location.

Cluster 0 has a total of 42 neighborhoods represented in red color

Cluster 1 has a total of 5 neighborhoods represented in purple color

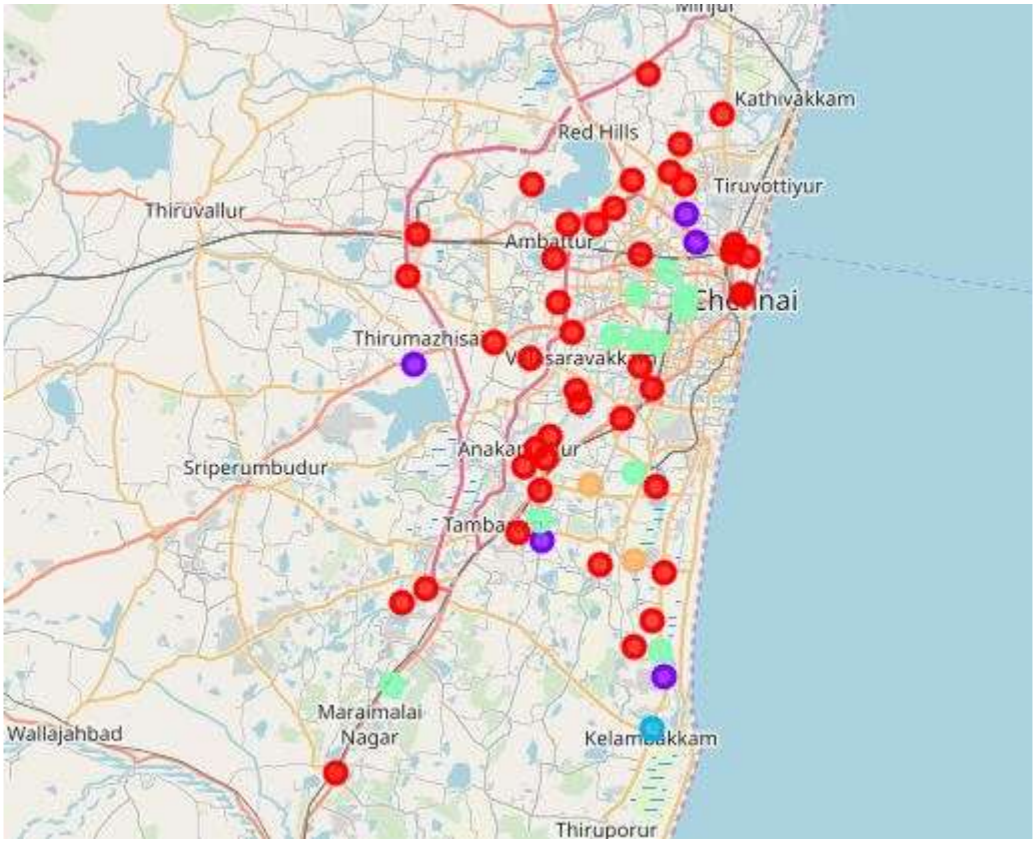

Figure 1. Neighborhoods clustered according to the frequency of occurrence of coffee shop in each location

Cluster 2 has a total of 1 neighborhoods represented in blue color

Cluster 3 has a total of 11 neighborhoods represented in pale green

Cluster 4 has a total of 2 neighborhoods represented in pale brown color

Discussion:

The neighborhoods with highest frequency of coffee shops are present in cluster 4. One is located in the IT corridor and the other in a highly populated locality.

The second highest frequency of coffee shops in a neighborhood are present in cluster 2. This place also tends to be the entrance to the city of Chennai located near the International Airport.

Next is cluster 1 with medium occurrences of coffee shops. It can be observed that the neighborhoods in this cluster are present in four different directions around the city of Chennai. One of the neighborhood is situated in the IT corridor, the other at locations centered on huge working populations and residential locality.

Cluster 3 has few coffee shops located in the neighborhoods. Few neighborhoods in this cluster are home to the lower middle class working population. One location situated in a hospital zone and one in the outskirts of the city which is currently being habited by people working in IT sector.

Cluster 0 has zero to very few coffee shops in the neighborhoods. The reasons for less or no coffee shops in this location can be attributed to one of the following:

- The neighborhoods in this cluster are located in the outskirts of Chennai
- They are inhabited by poor people who work as laborers.
- The neighborhoods are not safe, prone for theft activities
- They are not residential areas

From the analysis, it can be inferred that neighborhoods in cluster 0 would not be a viable place to set up a new coffee shop. The neighborhoods in cluster 4 are well populated, but has also a high concentration of coffee shops already serving customers. From a business stand point, this would mean a lot of competition, and ground work to develop a new customer base.

The neighborhood in cluster 2 is located on the outskirts of the city, but scores on two major points as listed below

- This location is a niche place for high valued residential properties
- Its proximity to the International airport tends to attract business travelers and foreign tourists to flock for a nice place to relax and have a cup of coffee.

But the cost of setting up a coffee shop would also be high in this neighborhood. This is suitable if one is capable of investing a lot of money.

Cluster 1 also has neighborhoods located at the outskirts of the city with a decent distribution of coffee shops. The cost of setting up a new shop would be considerably less, and would have a decent amount of travelling customers.

The neighborhoods in cluster 3 are located either in the middle of the city or in the IT corridor. This would be a good place to set up a coffee shop if one looks out to develop a base for regular customers to flock in.

Conclusion:

In this project, I have identified a business problem of finding a suitable location to setup a new coffee shop in the metropolitan city of Chennai, India. I have utilized web scraping for identifying the neighborhoods, and Foursquare API to identify the trending venues within a specified radius. An unsupervised machine learning known as k-means is used to cluster the neighborhoods based on the mean of the frequency of coffee shops in each neighborhood. The pros and cons of each neighborhood is detailed. Three clusters have been identified to be suitable for setting up a coffee shop. The question of which would be best depends on factor such as

- The initial investment planned
- The target customers (tourists, rich people or working class people)

If the investment is high and one targets tourist people and the rich class, the neighborhoods in cluster 2 would be a great place.

If the objective to invest a considerable amount and focus on highway travelers the neighborhoods in the cluster 1 would be the place to set up a new coffee shop.

If one wants to set up a coffee shop in the heart of the city and have a fan following of regular customers, then the ideal place would be the neighborhoods in cluster 3.