IBM Applied Data Science Capstone

Opening a Coffee shop in suburbs of Chennai , India

By: Ashok Venkatachalam

September 2020

Table of Contents

Introduction

Worldwide, experts estimate that people drink about 2.5 billion cups of coffee a day. Sales in the ready-to-drink market—which includes coffee shops—are forecast to grow by 67 percent between now and 2022. Additionally, coffee and other ready-to-drink shops show incredible resilience in volatile markets, helping to eliminate some of the uncertainty associated with small business ownership.

Business Problem

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

This project is timely as the city is currently short of coffee shop and the lifestyle of the people is changing and people are preferring to go to coffee shop for meeting and get together with friends.

Target Audience of this project

This project is particularly useful to investors looking to open or invest in new coffee shop in the suburbs of Chennai .

Data

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

- List of neighborhoods in Chennai . This defines the scope of this project which is confined to the city of Chennai , the metropolitan city in the south India.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.

- Venue data, particularly data related to coffee shops. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Chennai) contains a list of neighborhoods in Chennai , with a total of 65 neighborhoods. 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 neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. 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 Coffee Shop 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 Chennai . Fortunately, the list is available in the Wikipedia page ((https://en.wikipedia.org/wiki/Category:Suburbs_of_Chennai). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods 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 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 Chennai .

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 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 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. We group all the Restaurants into “Food Joints” category to identify where there is more “Restaurants”.

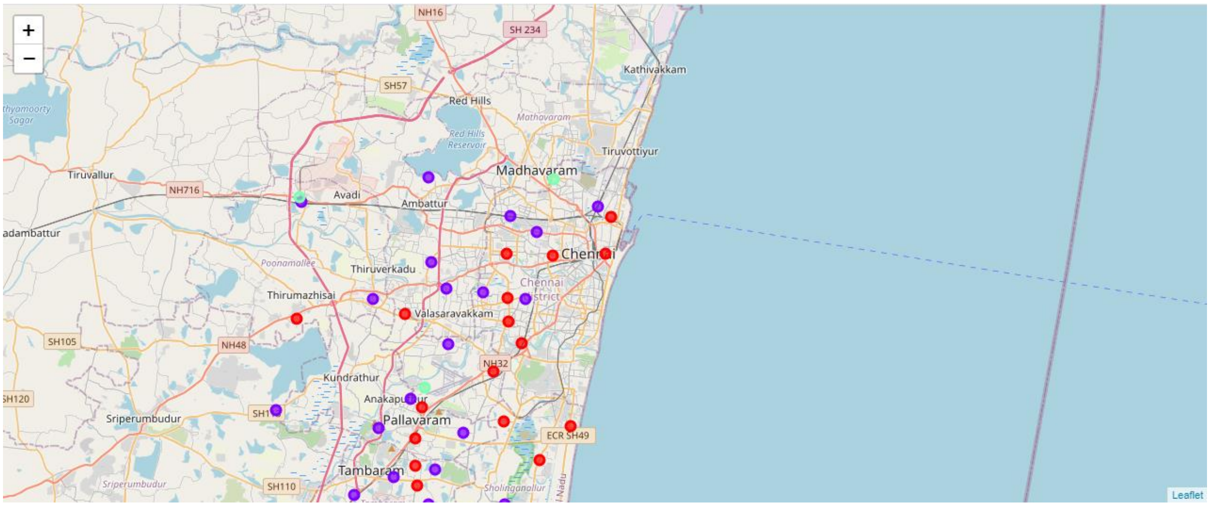
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Food Joints”. The results will allow us to identify which neighborhoods have higher concentration of Food Joints while which neighborhoods have fewer number of Food Joints. Based on the occurrence of Food Joints in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Food Joints.

Results

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

- Cluster 0: Neighborhoods with moderate number of Food Joints
- Cluster 1: Neighborhoods with low number to no existence of Food Joints
- Cluster 2: Neighborhoods with high concentration of Food Joints

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 green color.



Discussion

As observations noted from the map in the Results section, most of the Food Joints are concentrated in the central area of Chennai city, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no Food Joints in the neighborhoods. This represents a great opportunity and high potential areas to open new Food Joints as there is moderate number and hence less competition. Meanwhile, Food Joints in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Food Joints. From another perspective, the results also show that the oversupply of Food Joints mostly happened in the central area of the city, with the suburb area still have very few Food Joints. Therefore, this project recommends Investors to capitalize on these findings to open new Food Joints in neighborhoods in cluster 0. Investor with unique selling propositions to stand out from the competition can also open new Food Joints in neighborhoods in cluster 1 with sparing Food Joints. Lastly, investor are advised to avoid neighborhoods in cluster 2 which already have high concentration of Food Joints and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of Food Joints, there are other factors such as population and income of residents that could influence the location decision of a new Food Joints. However, to the best knowledge of this researcher such data are not available to the neighborhood 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 Food Joints. 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 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. investors 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 neighborhoods in cluster 0 are the most preferred locations to open a Coffee shop. This is because these area have restaurants means there is enough footfalls and starting a Coffee shop in these area would be a good alternative for the people visiting these places. Since the area is not crowded with restaurant there is a scope for new Food Joints and Coffee shop will be a good alternative here. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations in their decisions to open a new Coffee Shop.

References

Category:Suburbs in Chennai . *Wikipedia*. Retrieved from

https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur

Foursquare Developers Documentation. *Foursquare*. Retrieved from

https://developer.foursquare.com/docs

Appendix

Cluster 1

Alandur	Anna Nagar	Chromepet	Sholinganallur
Nazarethpettai	Washermanpet	Pallavaram	Thuraipakkam
Chitlapakkam	Kilpauk	Potheri village	Vadapalani
Iyyapanthangal	Madipakkam	Saidapet	Palavakkam
Guduvancheri	Senji	Selaiyur	Ashok Nagar

Cluster 2

Korukkupet	Madambakkam	Kelambakkam	Polichalur
Oragadam	Maduravoyal	Kodambakkam	Perambakkam
Panambakkam	Kamarajapuram	Sriperumbudur	Egattur
Virugambakkam	Singaperumalkoil	Perungalathur	Sembakkam
Peerkankaranai	K. K. Nagar	Ayanavaram	Keelkattalai
Tambaram	Thalambur	Thiruneermalai	Poonamallee
Mugalivakkam			

Cluster 3

Cowl Bazaar	Pattabiram	Kodungaiyur	Navalur