Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Pizza Shop in Chennai, India

By: Karthick Arulraj

May 2020

INTRODUCTION

For many Food lovers, visiting Pizza shops is a great way to relax and enjoy the taste of pizza. Pizza is one of the famous Food all over the world especially among kids. The reason for pizza spread is that you can add anything to it and eat it anytime of the day. It was originally dough with topping of any ingredients as meat or vegetables that baked in the oven. People can have different variety of pizza in whatever price range they need. For Entrepreneur, the central location and the large crowd place provides a great opportunity to Run a pizza shop. As a result, there are many Pizza shops in the city of Chennai and many more are being built. Opening Pizza shop allows Entrepreneur to earn consistent money. Of course, as with any business decision, opening a new Pizza shop requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the Pizza shop is one of the most important decisions that will determine whether the Pizza shop will be a success or a failure.

Business Problem

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

Target Audience

The entrepreneur who wants to find the location to open a New Pizza shop.

Data

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

• List of Neighbourhoods in Chennai, India. This defines the scope of this project which is confined to the city of Chennai, India.

• 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 Pizza shops. 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/List_of_neighbourhoods_of_Chennai) contains a list of Neighbourhoods in Chennai. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and Beautiful-soup 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 Pizza 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 neighbourhoods in the city of Chennai, India. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai). 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. 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 Data-Frame 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 Chennai.

Next, I use Foursquare API to pull the list of top 100 venues within 2000 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I'm able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Here, I made a justification to specifically look for "Pizza place"

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of Centeriods , 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 it is highly suited for this project as well. I have clustered the neighborhoods in Chennai into 3 clusters based on their frequency of occurrence for "Pizza place". Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the Pizza shop.

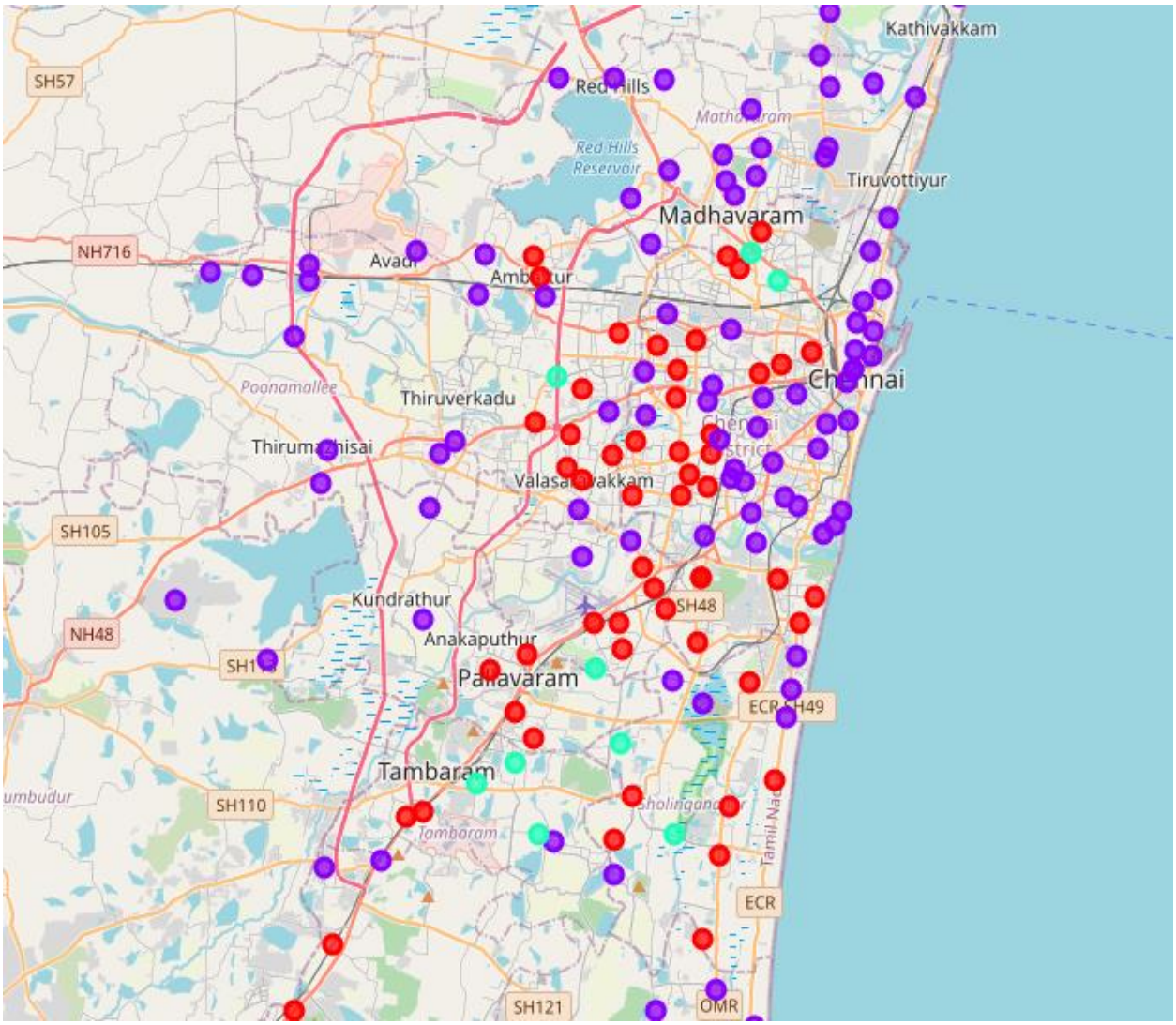
Results

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

- Cluster 0 - Neighbourhoods with moderate number of Pizza shops
- Cluster 1 - Neighbourhoods with low number to no existence of Pizza shops
- Cluster 2 - Neighbourhoods with high count of Pizza shops

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.

`



Discussion

Most of Pizza shops are in Cluster 2 and very low count (close to zero) in Cluster 1. Also, There are good opportunities to open near Park Town (City centre place), Panagal park (Top shopping place in Chennai) and Siruseri (IT Zone of Chennai) in Cluster 1. Looking at nearby venues, it seems Cluster 1 might be a good location as there are very less Pizza shops in these areas. Therefore, this project recommends the entrepreneur to open a Pizza shop in these locations with little to no competition.

Limitations and Suggestions for Future Research

In this project, I only take into consideration of one factor: the occurrence / existence of Pizza shops in each neighborhood. There are many factors that can be taken into consideration such as population density, income of residents, rent that could influence the decision to open a new Pizza shop. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration of these factors. 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 Entrepreneur regarding the best locations to open a new Pizza shop. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new pizza shop. The findings of this project will help the relevant Entrepreneur to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Pizza shop.

References

List of neighborhoods in Chennai, India:

https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai

Foursquare Developer Documentation: https://developer.foursquare.com/docs