Analyzing the Neighborhoods in Chennai for Starting a Restaurant

Applied Data Science Capstone Project

Date of Submission: June 19, 2021 By: Karthik B

Introduction

Chennai has a wide variety of food to offer, where the Idli Sambhar is a popular dish, which is served as breakfast or dinner. Apart from regular South Indian street food, the city's streets are also filled with several North Indian street food outlets, most of them established by North Indian migrants themselves. It has many restaurants which offer Idli Sambar, Dosa, Uttapam, Paniyaram, Jigarthanda and many more. Using data science concepts I want to find which area is the best to open a new restaurant.

Data Collection

The data required for this project has been collected from multiple sources. A summary of the data required for this project is given below.

Neighborhoods Data

The data of the neighborhoods in Chennai was scraped from https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai. The data is read into a pandas data frame using the read_html() method. The main reason for doing so is that the Wikipedia page provides a comprehensive and detailed table of the data which can easily be scraped using the read_html() method of pandas.

Geographical Coordinates

The geographical coordinates for Chennai data has been obtained from the GeoPy library in python. This data is relevant for plotting the map of Chennai using the Folium library in python. The geocoder library in python has been used to obtain latitude and longitude data for various neighborhoods in Chennai. The coordinates of all neighborhoods in Chennai are used to check the accuracy of coordinates given on Wikipedia and replace them in our data frame if the absolute difference is more than 0.001. These coordinates are then further used for plotting using the Folium library in python.

Venue Data

The venue data has been extracted using the Foursquare API. This data contains venue recommendations for all neighborhoods in Chennai and is used to study the popular venues of different neighborhoods.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Chennai. 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 pandas 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 Chennai. 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 “Restaurents” data, we will filter the “Restaurents” 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 “Restaurents”. The results will allow us to identify which neighbourhoods have higher concentration of restaurents while which neighbourhoods have fewer number of restaurents. Based on the occurrence of restaurents in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new restaurents

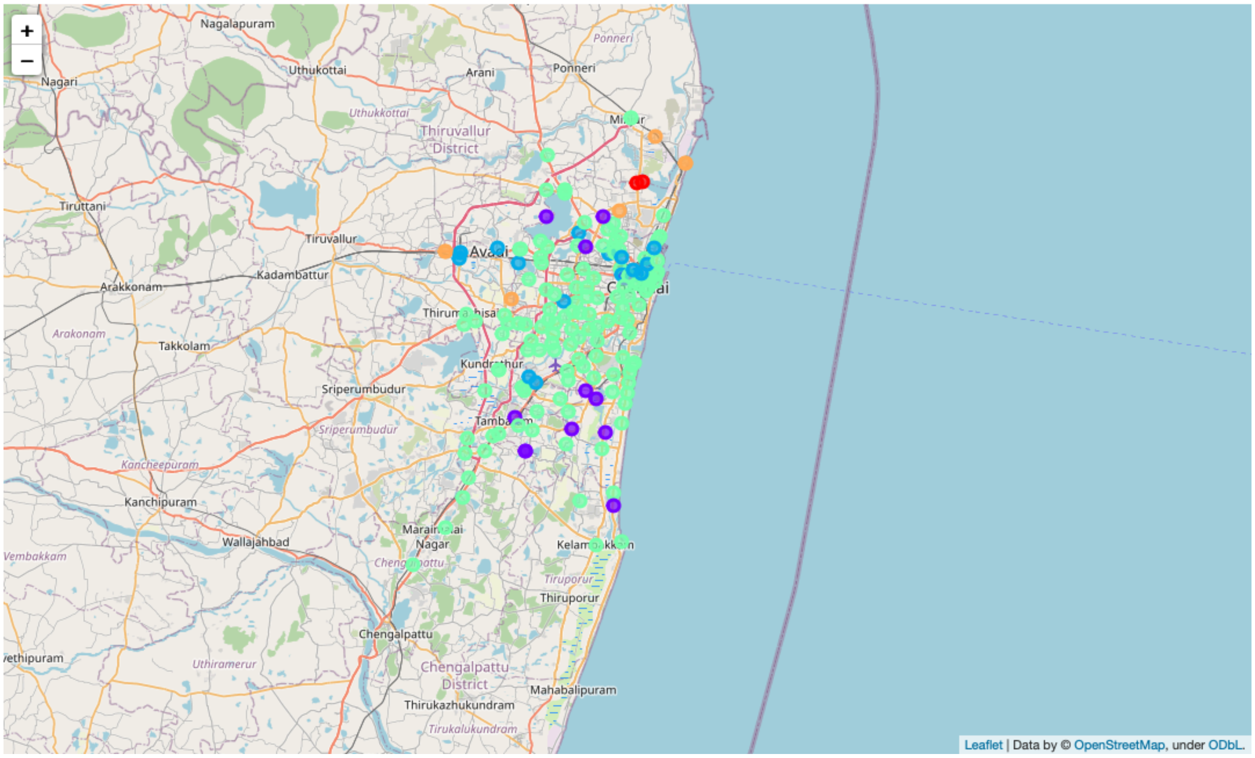
Results

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

- Cluster 0: Neighbourhoods with high concentration of restaurents
- Cluster 2: Neighbourhoods with moderate number of restaurents
- Cluster 1, 3, 4: Neighbourhoods with low number to no existence of restaurents

The results of the clustering are visualized in the map below with

- Cluster 0 in purple colour
- Cluster 1 in light blue colour
- Cluster 2 in mint green colour
- Cluster 3 in orange colour
- Cluster 4 in red color



Leaflet | Data by © OpenStreetMap, under ODbL.

Discussion

As observations noted from the map in the Results section, the highest number of restaurents are in cluster 0 and moderate number in cluster 2. On the other hand, clusters 1, 3 and 4 have very low numbers with no restaurant in the neighbourhoods.

- This represents that clusters 1, 3 and 4 have a great opportunity and are high potential areas to open new restaurents as there is very little to no competition from existing restaurents.
- Meanwhile, restaurents in clusters 0 and 2 are likely suffering from intense competition due to oversupply and high concentration of restaurents.
- From another perspective we can also say that the clusters 0 and 2 are very conducive to run a restaurant business and henceforth more restaurents are available in those areas.

- The results also show that the oversupply of restaurents mostly happened in the South and East Chennai, whereas the North and South Chennai areas have very few restaurents.
- Therefore, this project recommends restaurant owners to capitalize on these findings to open new restaurents in neighbourhoods in clusters 1, 3 and 4 with little to no competition.
- Restaurant owners with unique selling propositions to stand out from the competition can also open new restaurents in neighbourhoods in clusters 0 and 2 with moderate competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of restaurents, there are other factors such as population and income of residents that could influence the location decision of a new restaurant. However, to the best knowledge of this researcher such data are not available to the neighbourhood 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 restaurant. 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 accounts 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. property developers and investors regarding the best locations to open a new restaurant.

To answer the business question that was raised in the introduction section, the answer proposed by this project is: **The neighbourhoods in clusters 1, 3 and 4 (generally Northern and Western parts of Chennai) are the most preferred locations to open a new restaurant. We can also say that if the restaurant owner has a unique selling proposition then the restaurents can be opened in clusters 0 and 2 (generally Southern and Eastern parts of Chennai) to stand out from the competition**. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new restaurant.