Analyzing and Predicting optimal locations for popular restaurants in Chennai city

Mohammed Ismail P

April 22, 2021

1. Introduction

1.1 Background

Chennai, also known as Madras, is the capital of the Indian state of Tamil Nadu. Located on the Coromandel Coast of the Bay of Bengal, it is one of the largest cultural, economic, and educational centers of south India. According to the 2011 Indian census, it is the sixth-most populous city and fourth-most populous urban agglomeration in India. The Chennai Metropolitan Area is the 36th-largest urban area by population in the world. The traditional and de facto gateway of South India, Chennai is among the most visited Indian cities by foreign tourists. It was ranked the 36th-most visited city in the world for the year 2019. The Quality of Living Survey rated Chennai as the safest city in India. Chennai has the fifth-largest urban economy in India.

The diversity of the cuisine available is reflective of the social and economic diversity of Chennai. Indian, Chinese, Asian, Italian, Middle Eastern, Thai are some of the most popular in the city. We will analyze each of these restaurants to predict a suitable location for the respective restaurant.

1.2 Problem Definition

In this capstone project, we will try to visualize all major parts and areas of Chennai city and try to predict optimal locations for each of the restaurants. Since there are lots of restaurants in Chennai, we try to select those locations that are not already crowded with restaurants within the region. Due to low restaurant density, the chance of setting up a new restaurant might be a great success.

1.3 Interest

In particular, the concluded project report will target stakeholders who want to set up one of the popular restaurants in Chennai, Tamil Nadu, India. We will use various data science and analysis techniques to reach our goal of selecting optimal locations. The advantages of each area will then be clearly expressed so that the best possible final location can be chosen by stakeholders.

2. Data acquisition

The data and their respective data sources for our projects are discussed below,

2.1 Chennai city's major areas and neighborhoods

Data Source: Wikipedia (https://en.wikipedia.org/wiki/Areas_of_Chennai)

Data description: This data contains all the information about the major areas and neighborhoods in Chennai city. We will initialize the crawler to scrape and extract data and information about the areas and locality from the Wikipedia web page. The data we get is only the names of all the neighborhoods in Chennai city. A total of 163 neighborhood names are extracted. Sample data is shown in Figure 1,

	Neighborhood
0	Adambakkam
1	Adyar
2	Alandur
3	Alapakkam
4	Alwarpet

Figure 1: Names of Chennai city's major areas

2.2 Geographical coordinate of the areas

Data Source: Python's geocoder library (https://geocoder.readthedocs.io/index.html)

Data description: Geocoder is a simple and consistent geocoding library written in Python. Dealing with multiple different geocoding providers such as Google, Bing, OSM, and many more. Specifically, we will be using ArcGIS, as it is reliable and free. This data contains all the geographical locations of each of the neighborhoods, that are merged with previously scraped neighborhood data. The coordination data is used to plot the neighborhood on the Chennai city map. Sample data is shown in Figure 2,

	Neighborhood	Latitude	Longitude
0	Adambakkam	12.99192	80.20603
1	Adyar	13.00304	80.25187
2	Alandur	13.00013	80.20049
3	Alapakkam	13.04613	80.16501
4	Alwarpet	13.03471	80.25416

Figure 2: Chennai city's Neighborhood with their coordinates

2.3 Venues in each locality of Chennai city

Data Source: Foursquare API (https://foursquare.com/)

Data description: Foursquare API allows us to fetch information, in form of a JSON file, about all the venues in the neighborhood of Chennai city, using Foursquare credentials. The coordinates data is used by the Foursquare API to fetch the venue within the specified radius (we used the radius of 1500 meters). We extract only the relevant information like Venue name, Venue coordinate, and Venue Category. Chennai city has a total of 211 unique venue categories and 1864 venues, which includes various restaurants, Shopping mall, Multiplex, etc., Sample data is shown in Figure 3,

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adambakkam	12.99192	80.20603	Luxe Cinemas	12.991041	80.216962	Multiplex
1	Adambakkam	12.99192	80.20603	Phoenix Market City	12.991710	80.217297	Shopping Mall
2	Adambakkam	12.99192	80.20603	Mainland China	12.991028	80.217084	Chinese Restaurant
3	Adambakkam	12.99192	80.20603	IMAX®	12.990639	80.216310	Multiplex
4	Adambakkam	12.99192	80.20603	Sukkkubai Beef Biryani Shop	12.998769	80.201381	Indian Restaurant

Figure 3: Neighborhoods with surrounding venue details

3. Methodology

As discussed above, we will initialize a crawler to scrape the data about the areas of Chennai city, from the Wikipedia webpage. It contains only the names of the neighborhoods in Chennai city. We then utilize python's geocoder library, with ArcGIS as a geocode provider, to get the coordinates of the neighborhood. The neighborhoods are visualized using python's folium package. This allows us to perform a sanity check to make sure that the geographical coordinate's data returned by the geocoder are plotted correctly. Map of Chennai city along with the neighborhood is shown in Figure 4,

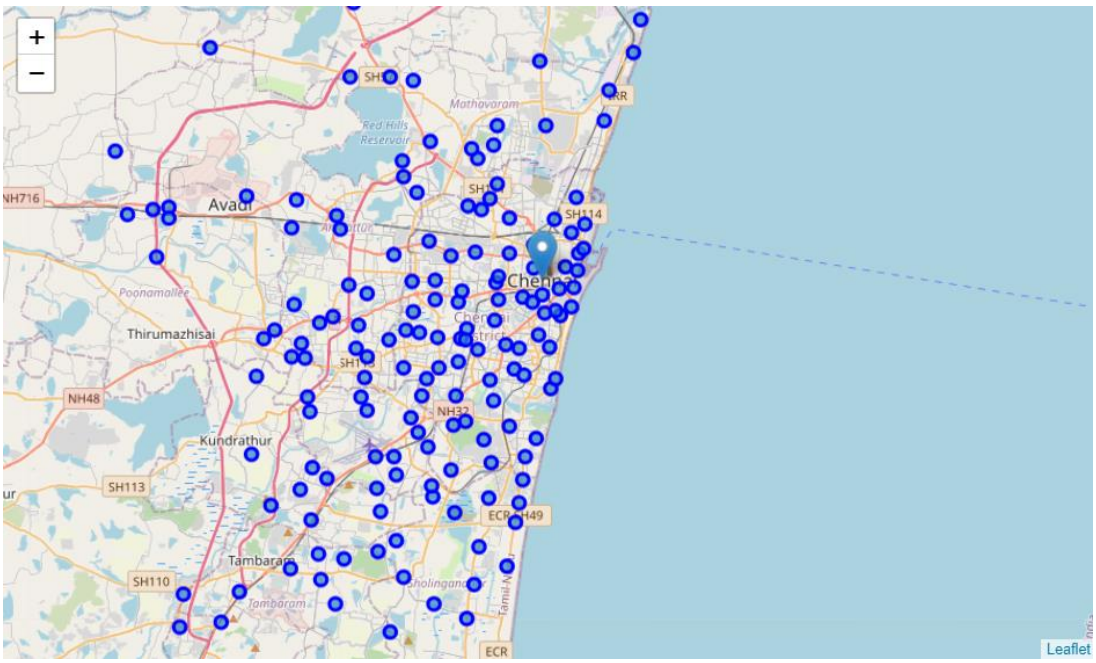

Figure 4: Chennai city map with neighborhoods

Then, the coordinates data is used by the Foursquare API to fetch information about all the venues in each of the neighborhoods of Chennai city, targeting only the relevant information like Venue name, Venue coordinate, and Venue Category. In the end, we get a table as shown in Figure 3.

We then use one-hot encoding which allows the representation of categorical data to be more expressive. Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers. In our case, venue categories grouped by the neighborhood and are expressed as the mean of the frequency of occurrence of each category. Sample data is shown in Figure 5,

	Neighborhood	ATM	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Terminal	American Restaurant	Andhra Restaurant	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store
0	Adambakkam	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.00
1	Adyar	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.014286	0.0	0.00
2	Alandur	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.00
3	Alapakkam	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.00
4	Alwarpet	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.010000	0.0	0.01

Figure 5: Venue density distribution in each neighborhood

We then apply the clustering algorithm, KMeans, on the above data to cluster the neighborhood based on general venue density and analyze & compare the sets in each cluster to conclude the most promising and optimal locations for each restaurant type. We then filter out only the target restaurant of interest in each neighborhood, to analyze within the clusters.

4. Analysis

The analysis phase starts immediately after applying a clustering algorithm to cluster the neighborhood, based on general venue density. The results will allow us to identify similar neighborhoods with the concentration of each restaurant type. Based on the occurrence of these restaurants in different neighborhoods, we will try to answer the question as to which neighborhood is most suitable to open respective restaurants. In particular, we will target those neighborhoods that have less than 20% of the entire restaurant density and no target restaurant.

We identified 4 as optimal cluster value to group similar neighborhoods according to the Silhouette score, which computes the mean Silhouette Coefficient of all samples. The best value is 1 and the worst value is -1. In other words, the best cluster value has a greater Silhouette score that is close to 1. After applying the KMeans clustering algorithm, all the neighborhoods get separated and form different clusters. The resulted sample output is as shown in Figure 6,

	Neighborhood	Latitude	Longitude	Cluster Labels	ATM	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Terminal	American Restaurant	Andhra Restaurant	Antique Shop
0	Adambakkam	12.99192	80.20603	3	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0
1	Adyar	13.00304	80.25187	0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0
2	Alandur	13.00013	80.20049	0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0
3	Alapakkam	13.04613	80.16501	3	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0
4	Alwarpet	13.03471	80.25416	3	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0

Figure 6: Venue density with Cluster label in each neighborhood

The above result in Figure 6, is used to create map using folium library with different markers for each cluster. The cluster with 4 different colours along with the cluster legends is shown in Figure 7,

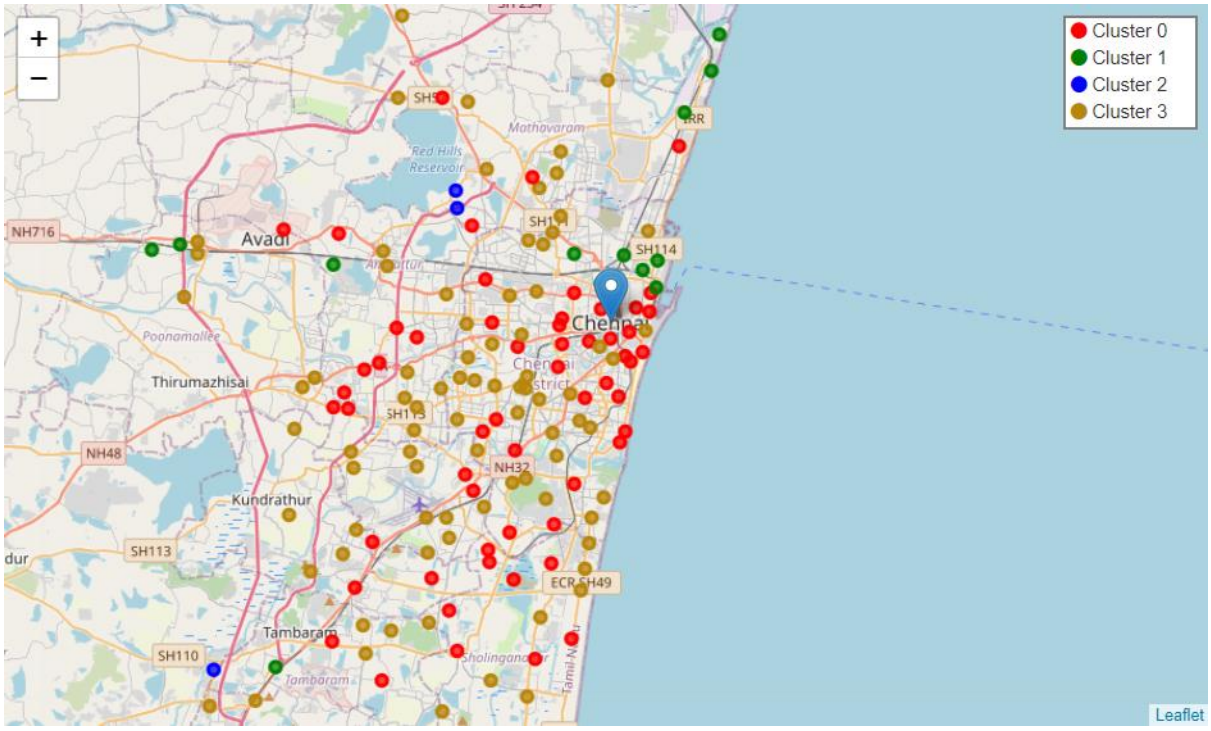

Figure 7: Chennai city map with clustered neighborhoods

Each of the clusters are analyzed separately to find the opportunity of opening a new restaurant with respect to each type, within each of the cluster. Each of the clusters are thoroughly examined to uncover some useful insights about the restaurant and general venue density.

5. Result

After thoroughly analyzing the clusters and their respective neighborhoods, the findings are concluded as follows,

5.1 Cluster 0

Cluster 0 has a total of 57 neighborhoods. It has full of restaurants and other food shops. Opening a new restaurant in an already established market is not a good idea. Indian restaurant, Fast Food restaurant, Café, Pizza place, Hotel, Food, Train station, Food & Drink shop, Bakery, and Chinese restaurant are some of the most common venues in Cluster 0. Only 8% of the areas have the opportunity to open new restaurants.

Table depicting the proportion of each restaurant type with respect to other venues within cluster 0,

Restaurant	Percentage
Indian restaurant	19.4015
Fast Food restaurant	4.6979
Vegetarian restaurant	2.5274
Chinese restaurant	1.2936
Asian restaurant	0.6925
Italian restaurant	0.7958

Pie chart that depicts the proportion of each restaurant type in cluster 0 is shown in the Figure 8,

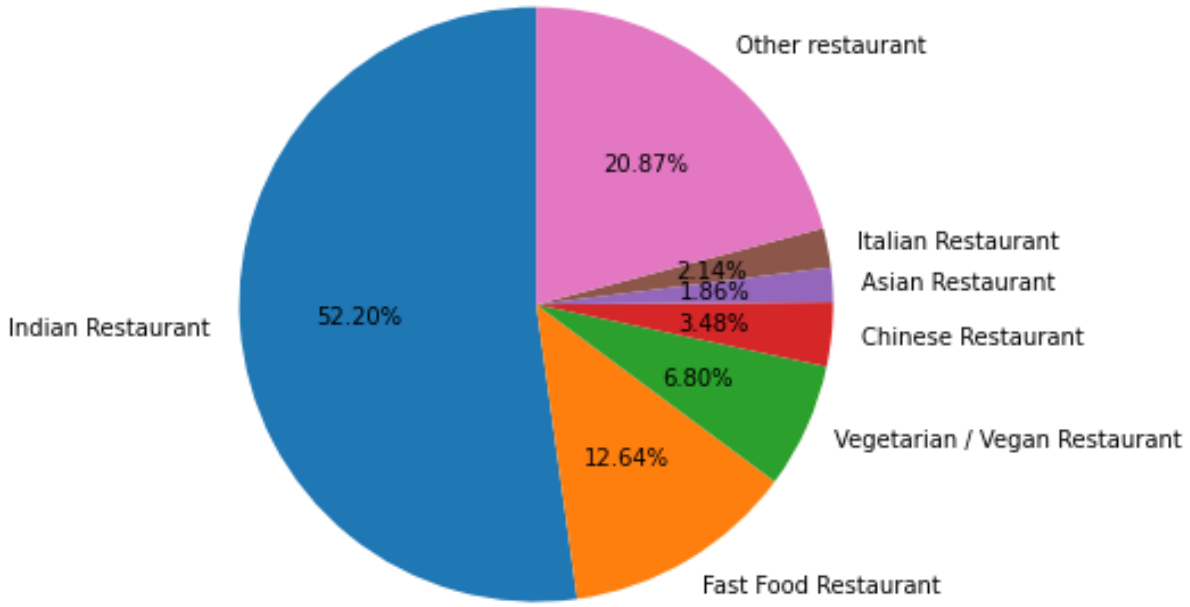

Figure 8: Proportion of restaurants in Cluster 0

Based on the cluster 0 analysis, the following areas are concluded as the optimal locations for opening each of the restaurant type,

Any type of restaurant (preferably Indian)

- Madhavaram
- Madhavaram Milk Colony
- Vallalar Nagar

Fast Food, Chinese, Asian and Italian restaurant

- Kazhipattur (also Vegetarian)
- Mangadu (also Indian)

5.2 Cluster 1

Cluster 1 has a total of 14 neighborhoods. It has high market density with fewer medium popular restaurant types. This cluster can be considered as one of the optimal places to open medium popular restaurant type. Railway station, Food, Flower shop, Flea market, Fish market, Food & Drink shop, Field, Yoga studio, Fast Food Restaurant, and Indian restaurant are some of the most common venues in Cluster 1. About 74% of the area has promising opportunity for opening a new restaurant. Point to be noted that this cluster has no Asian and Italian restaurant.

Table depicting the proportion of each restaurant type with respect to other venues within cluster 1,

Restaurant	Percentage
Indian restaurant	7.3696
Fast Food restaurant	3.6990
Vegetarian restaurant	2.4490
Chinese restaurant	1.8141
Asian restaurant	0.0000
Italian restaurant	0.0000

Pie chart that depicts the proportion of each restaurant type in cluster 1 is shown in the Figure 9,

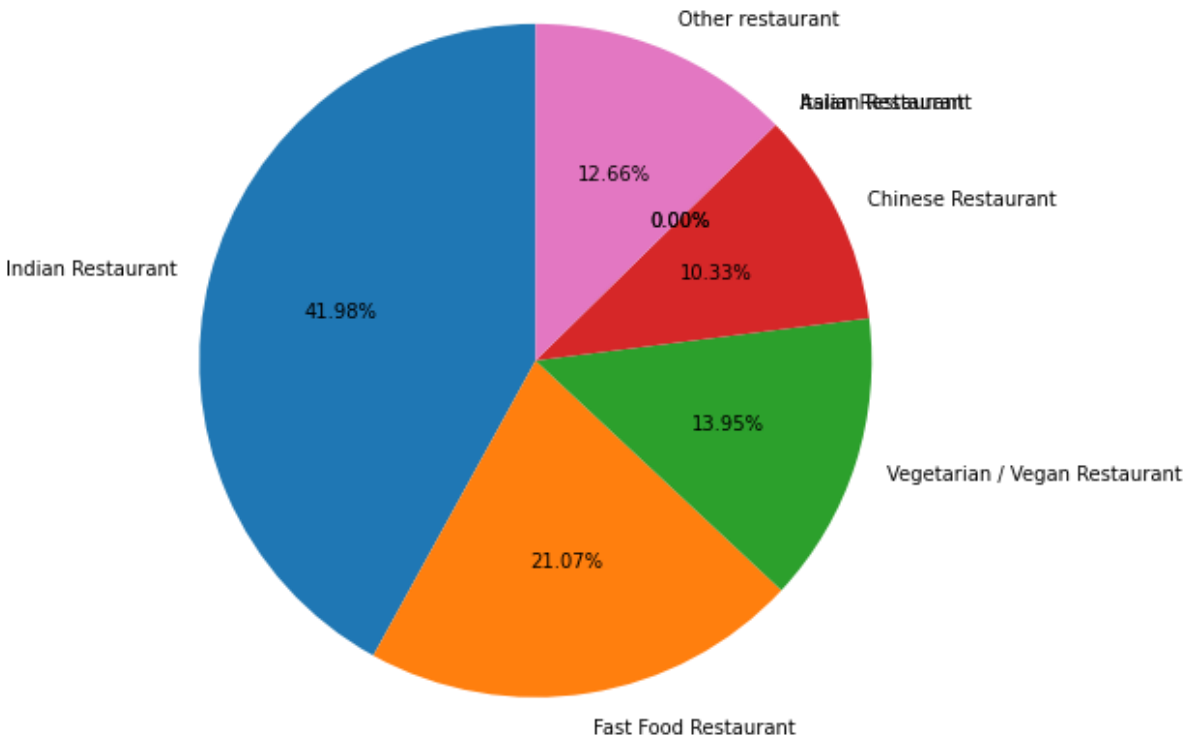

Figure 9: Proportion of restaurants in Cluster 1

Based on the cluster 1 analysis, the following location are concluded as the optimal locations for opening each of the restaurant type,

Any type of restaurant

- Eranavur, Ennore, Mathur, Nemilichery, Peerkankaranai, Pulianthope (preferably Indian)
- Korukkupet (except Vegetarian)
- Royapuram, Washermanpet (except Fast Food)
- Thiruninravur (except Indian)

5.3 Cluster 2

Cluster 2 has only 3 neighborhoods. It has negligible restaurant density. Opening a new restaurant in this cluster can be profitable. Pharmacy, Electronics store, Food & Drink shop, Food, Flower shop, Flea market, Fish market, Field, Yoga studio and Fast Food restaurant are some of the most common venues in Cluster 2. 2 out of 3 areas do not have a single restaurant.

Table depicting the proportion of each restaurant type with respect to other venues within cluster 2,

Restaurant	Percentage
Indian restaurant	2.6667
Fast Food restaurant	0.0000
Vegetarian restaurant	1.3333
Chinese restaurant	1.3333
Asian restaurant	2.6667
Italian restaurant	2.6667

Pie chart that depicts the proportion of each restaurant type in cluster 2 is shown in the Figure 10,

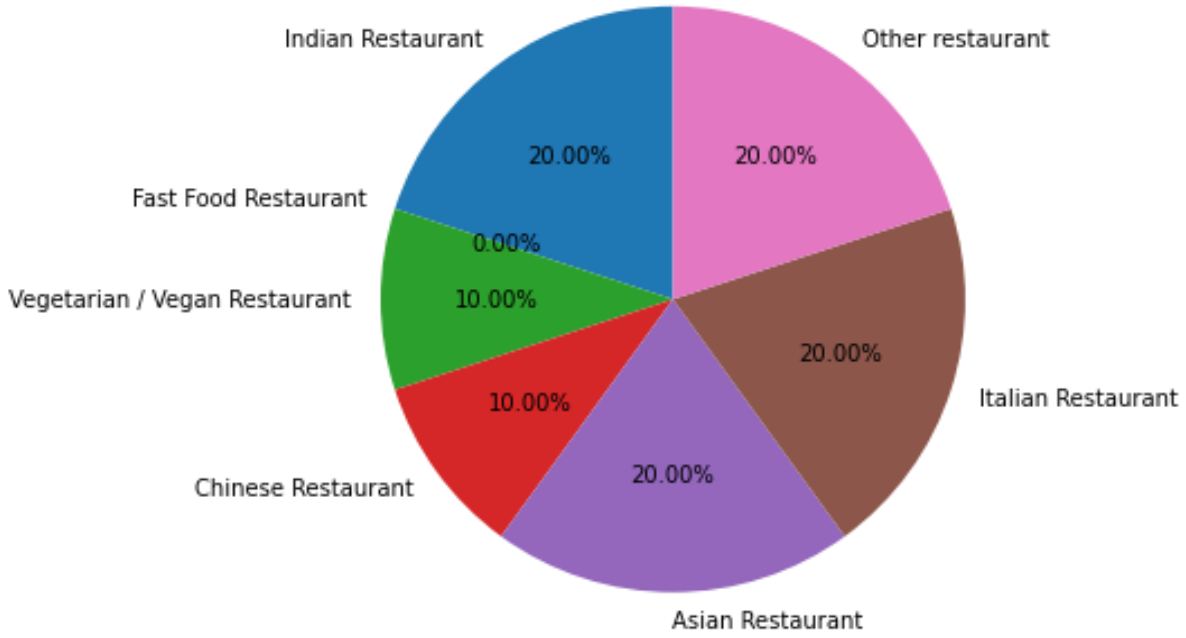

Figure 10: Proportion of restaurant in Cluster 2

Based on the cluster 2 analysis, the following areas are concluded as the optimal locations for opening each of the restaurant type,

Any type of restaurant (preferably Indian)

- Puthagaram
- Surapet

5.4 Cluster 3

Cluster 3 has a total of 83 neighborhoods. It is very similar to Cluster 0 with high restaurant density. Indian restaurant, Pizza place, Café, Flower shop, Fast Food restaurant, Food, Bakery, Coffee shop, Hotel and Clothing store are some of the most common venues in Cluster 3. About 36% of the areas have the opportunity to open new restaurant.

Table depicting the proportion of each restaurant type with respect to other venues within cluster 3,

Restaurant	Percentage
Indian restaurant	9.7639
Fast Food restaurant	3.9858
Vegetarian restaurant	2.4506
Chinese restaurant	1.5784
Asian restaurant	1.5971
Italian restaurant	0.7607

Pie chart that depicts the proportion of each restaurant type in cluster 3 is shown in the Figure 11,

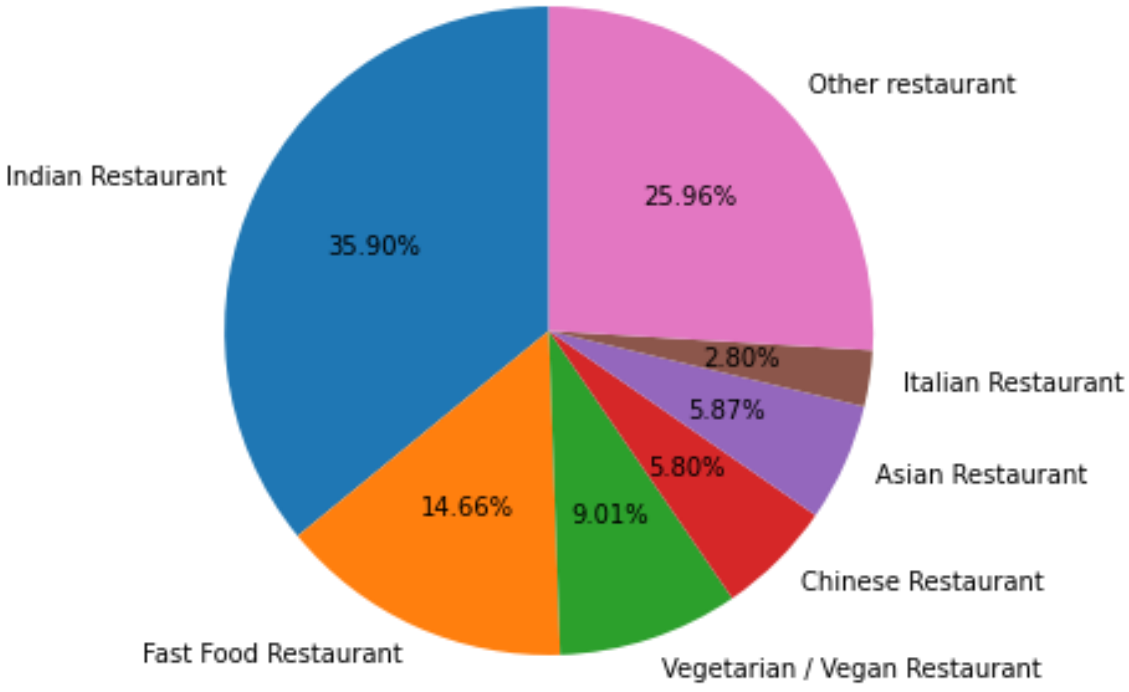

Figure 11: Proportion of restaurant in Cluster 3

Based on the cluster 3 analysis, the following areas are concluded as the optimal locations for opening each of the restaurant type,

Any type of restaurant

- Ayanavaram, K.K. Nagar, Kodungaiyur, Korattur, Kundrathur, Manali New Town, Minjur, Moolakadai, Mudichur, Pammal, Pattabiram, Pattaravakkam, Perumbakkam, Perungalathur, Poonamallee, Puzhal, Sholavaram, Vandalur (preferably Indian)
- Chitlapakkam, Kottur, Madambakkam, Mannady, Nanganallur, Selaiyur, Sholinganallur, Thiruneermalai (except Indian)
- Perambur (except Chinese)
- Peravallur (except Asian)

6. Discussion

There are some clusters that do not have some specific restaurant type. For example, Cluster 1 does not have a single Asian and Italian restaurant, and Cluster 2 does not have a single Fast Food restaurant. Thus, opening these types of restaurants have the high possibility of success as well as failure, due to various factors like populations, income of residents, financial reports, etc., Thus further analysis and survey of these clusters and areas of interest are needed to be done before taking further actions.

We can also apply the same approach for the larger dataset. This approach can also be applied to identify the promising locations for other venues like shopping malls, movie theatres, salons, and other related businesses. In this project, we considered only one factor, that is, venue density distribution. There are some other factors like population and income of residents, real estate availability, surrounding localities, financial report, etc., that can be really helpful to further narrow down the choice of selecting the optimal locations with a more accurate analysis.

7. Conclusion

To summarize, the purpose of this project was to identify the optimal locations for opening some of the popular restaurant type namely, Indian, Fast Food, Vegetarian, Chinese, Asian and Italian restaurant. Factors like venue density distribution is used to achieve our task.

First, we scraped the neighborhood information of Chennai city from the Wikipedia website and used python’s geocoder library to get the coordinate information of each of the neighborhoods. Then, we utilized Foursquare API to fetch and identify all the venues within each neighborhood provided the radius. We then clustered the neighborhood based on the venue density. Each cluster provided useful information and insights about the venues and restaurants that are briefly discussed in analysis and result section, that are helpful for the stakeholders.

Final decision on selecting the optimal location for the restaurant will be made by stakeholders based on specific characteristics of neighborhoods and locations, like competition, future growth, Health regulations, Safety / Crime rate, and taking additional factors into consideration that are discussed above.