Table of Contents

Table of Figures

The Battle of Neighbourhoods

Manasa Krishnan

July 21, 2021

1. Introduction

1.1 Background

Chennai, the city of multiple cultures, is the economic capital of the state of Tamil Nadu located in India. Known for its welcoming nature to citizens of other countries, it is no surprise that restaurants and food outlets play a major role. Spread across a huge land area, with beautiful beaches, one cannot resist visiting this place. With this project, one would know what to look for exactly, ranging from restaurants to food trucks, and enjoy as a tourist, or better *a Chennaite!*. Business People and young entrepreneurs will find this project especially helpful as it will lead to specific locations where their start-ups might become a hit.

1.2 Problem

The main aim of the project is to identify several venue categories using latitude and longitude values of the areas of Chennai and merge it with the Foursquare API so that business-minded people and even the Chennaites can make use of the clustered data to make smart decisions about starting a business in a given location depending on its popularity.

1.3 Target Audience

The target audience are the business people and young entrepreneurs who want to start their businesses in and around Chennai. The results obtained here would be useful for their endeavours. Moreover, local citizens and foodaholics can make use of these visualisations to try out new restaurants.

2. Data Acquisition and Cleaning

2.1 Data Sources

Chennai is a widely spread city with multiple areas consisting of hotels and food outlets. The link to dataset used here is: **https://chennaiiq.com/chennai/latitude_longitude_areas.asp** . The .csv file consists of columns named Neighborhood, Latitude and Longitude. The data for venues across the areas of Chennai will be obtained later with the Foursquare API credentials, Client ID and Client Secret.

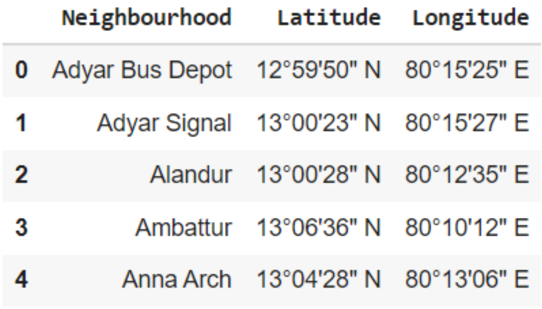
2.2 Data Pre-processing

After sucessfully importing all the necessary libraries, we go ahead with extracting necessary information from the html page using **soap** object and **html parser**. We collect only the location, latitude and longitude information, convert it into a dataframe, rename the columns and finally get the resulting dataframe with column names **Neighbourhood, Latitude and Longitude**.

The raw dataset obtained from the above-mentioned link consists of the latitude and longitude values in degree, minutes and seconds. We will change these to numerical values using the formula:

$$x = degree + (minute/60) + (second/(60 * 60)),$$

where x is the resultant float value obtained after the calculation. After this, we will successfully obtain a dataframe with float values for latitudes and longitudes.



	Neighbourhood	Latitude	Longitude
0	Adyar Bus Depot	12°59'50" N	80°15'25" E
1	Adyar Signal	13°00'23" N	80°15'27" E
2	Alandur	13°00'28" N	80°12'35" E
3	Ambattur	13°06'36" N	80°10'12" E
4	Anna Arch	13°04'28" N	80°13'06" E

Fig 2.2.1 Sample Dataframe

3. Methodology

3.1 Visualising Chennai

Folium Library is one of the most widely used library for visualisation of maps. In this step, we will find the latitude and longitude values of Chennai and visualise the areas given in the dataset. With this, we will know how wide spread Chennai is and what areas have the greatest number of restaurants, shopping malls and others with the help of marker cluster algorithm**.**

3.2 Exploring Food Outlets

Foursquare is a social networking service available for common smartphones, including the iPhone, BlackBerry and Android-powered phones. The app's purpose is to help you discover and share information about businesses and attractions around you.

Here, we use our Foursquare credentials to extract popular venues in and around the areas of Chennai. We use our Client ID and Client Secret to access our Foursquare account and find the venue names. We convert all these details into a dataframe understandable by all the users. The exact co-ordinates of the location of the venues are also extracted.

3.3 Finding Unique Categories

We use the groupby method to find the unique number of venue categories from the dataset that we derive after using Foursquare API to get the venue category, latitude and longitude of the areas of Chennai. The final dataframe obtained consists of both neighbourhood latitudes and longitudes and also venue latitudes and longitudes. We get a total of 141 unique venue categories.

3.4 Plotting Charts to Check for Outliers and Bias

Matplolib library is a powerful tool for plotting graphs and charts. With its easy-to-use interface, we will plot a bar chart for the venue categories at hand. We have 141 unique venue categories. From this, we plot bar charts to check for bias in the dataset. If there is a bias, our methodology might be skewed. We will also remove venue categories that are less than 10 in numbers.

3.5 Executing One-hot Encoding

One-hot encoding produces a sparse matrix, where most of the entires of the m*n matrix are **zeroes**. A column which has a value of 1 indicates that the row is associated to that column of value. This is applied to Venue Category as cluster analysis will be performed based on those categories.

3.6 Clustering Neigbourhoods using K-Means Technique

K-means clustering algorithm is one of the most powerful techniques of unsupervised machine learning algorithms. With this algorithm, we can finally visualise Chennai city's map with superimposed markers. First, we identify the appropriate number of clusters needed for our problem at hand. For that, we use silhouette score from sklearn.metrics package.

We will get a value of 5 which has the highest silhouette score from the graph we will plot. So, the number of clusters will be taken as 5. We will produce a dataframe with each neigbourhood’s first to tenth most common venues. These venues range from restaurants, multiplex, electronic stores, shopping malls, bus stations, grocery stores etc.

4. Result

The final step is to analyse the clusters obtained. We take each cluster and produce results with the distinct venue categories and the total count of such categories. After examining 5 clusters which were our optimised number of clusters as per the silhouette score, we plot a map with these clusters marked by the Folium library. For example, cluster label 1 consists food outlets like bakery, cake shops, restaurants etc and cluster label 2 consists of shopping malls and multiplexes.

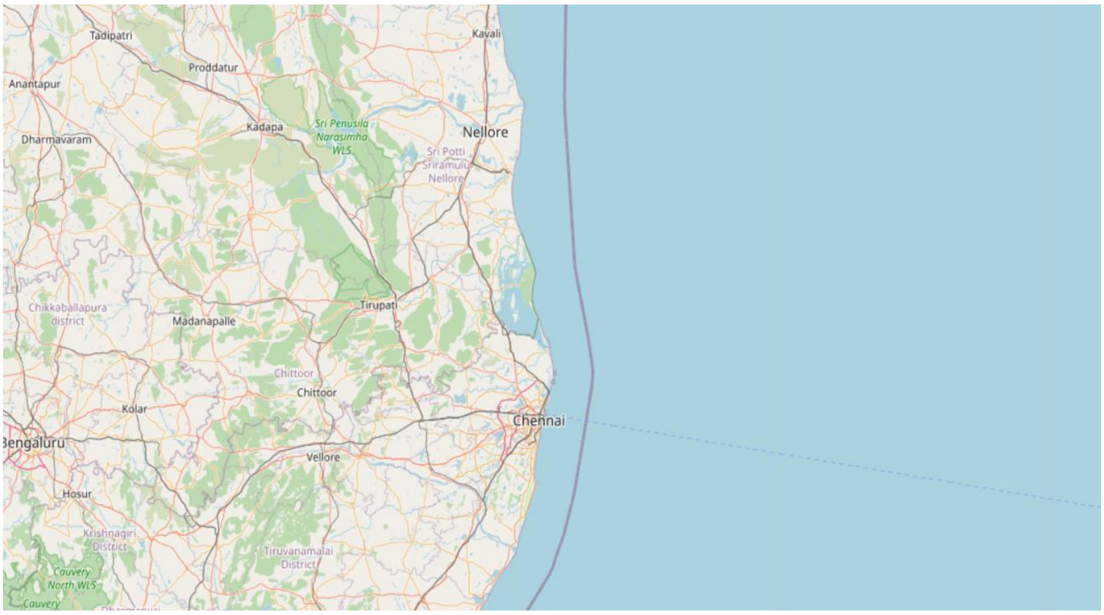


Fig 4.1 Chennai city’s location with its respective latitude and longitude visualised using Folium library

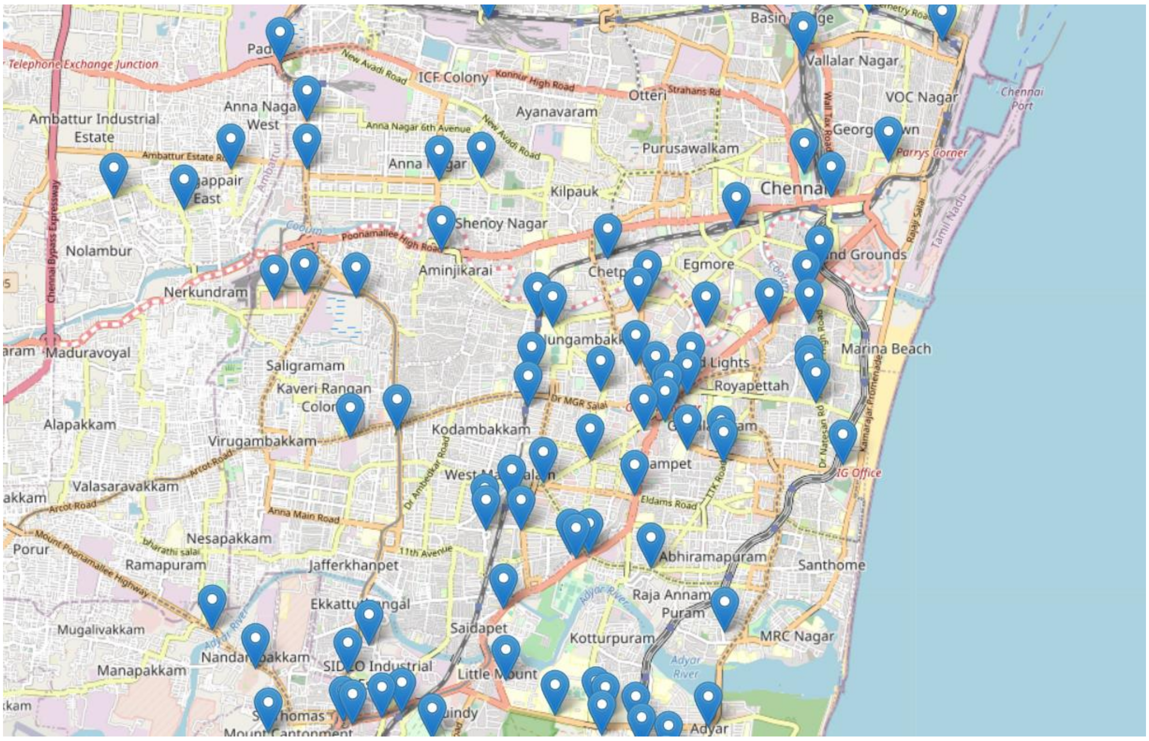
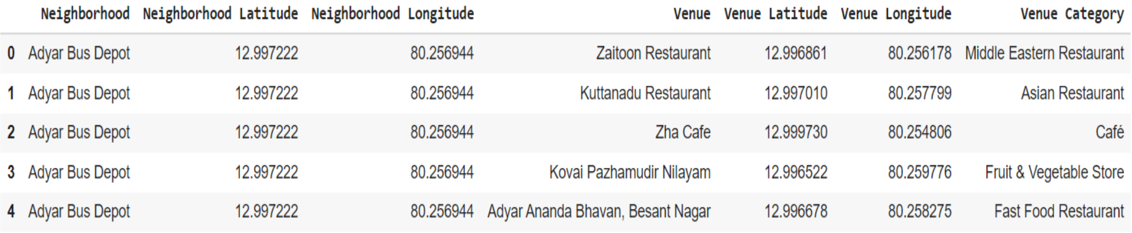
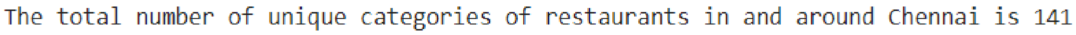


Fig 4.2 The areas of Chennai visualised using geopy and folium library plotted using marker cluster algorithm



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adyar Bus Depot	12.997222	80.256944	Zaitoon Restaurant	12.996861	80.256178	Middle Eastern Restaurant
1	Adyar Bus Depot	12.997222	80.256944	Kuttanadu Restaurant	12.997010	80.257799	Asian Restaurant
2	Adyar Bus Depot	12.997222	80.256944	Zha Cafe	12.999730	80.254806	Café
3	Adyar Bus Depot	12.997222	80.256944	Kovai Pazhamudir Nilayam	12.996522	80.259776	Fruit & Vegetable Store
4	Adyar Bus Depot	12.997222	80.256944	Adyar Ananda Bhavan, Besant Nagar	12.996678	80.258275	Fast Food Restaurant

Fig 4.3 The venue details obtained for areas of Chennai using Foursquare API



```
The total number of unique categories of restaurants in and around Chennai is 141
```

Fig 4.4 The total number of unique venue categories for the given dataset is found to be 141 which will later be clustered.

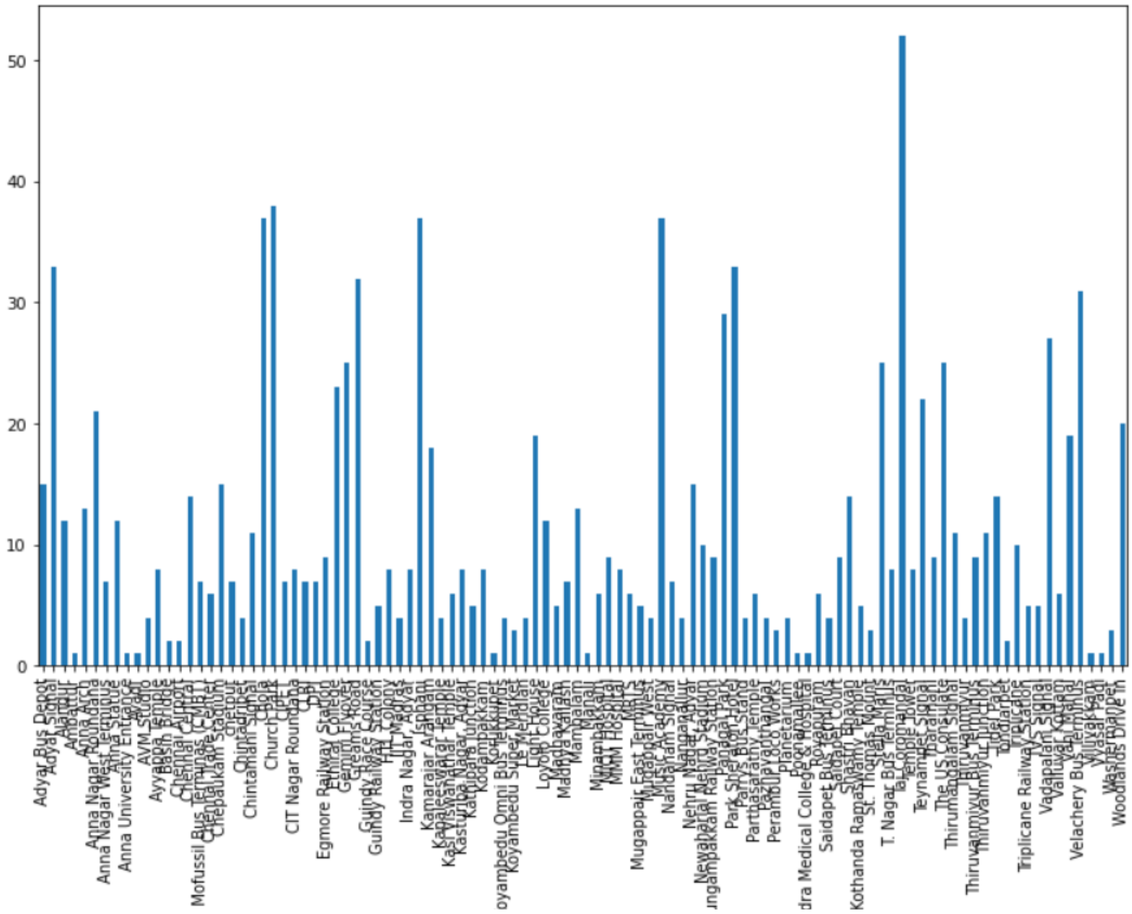


Fig 4.5 A bar chart showing the number of venues present in Chennai

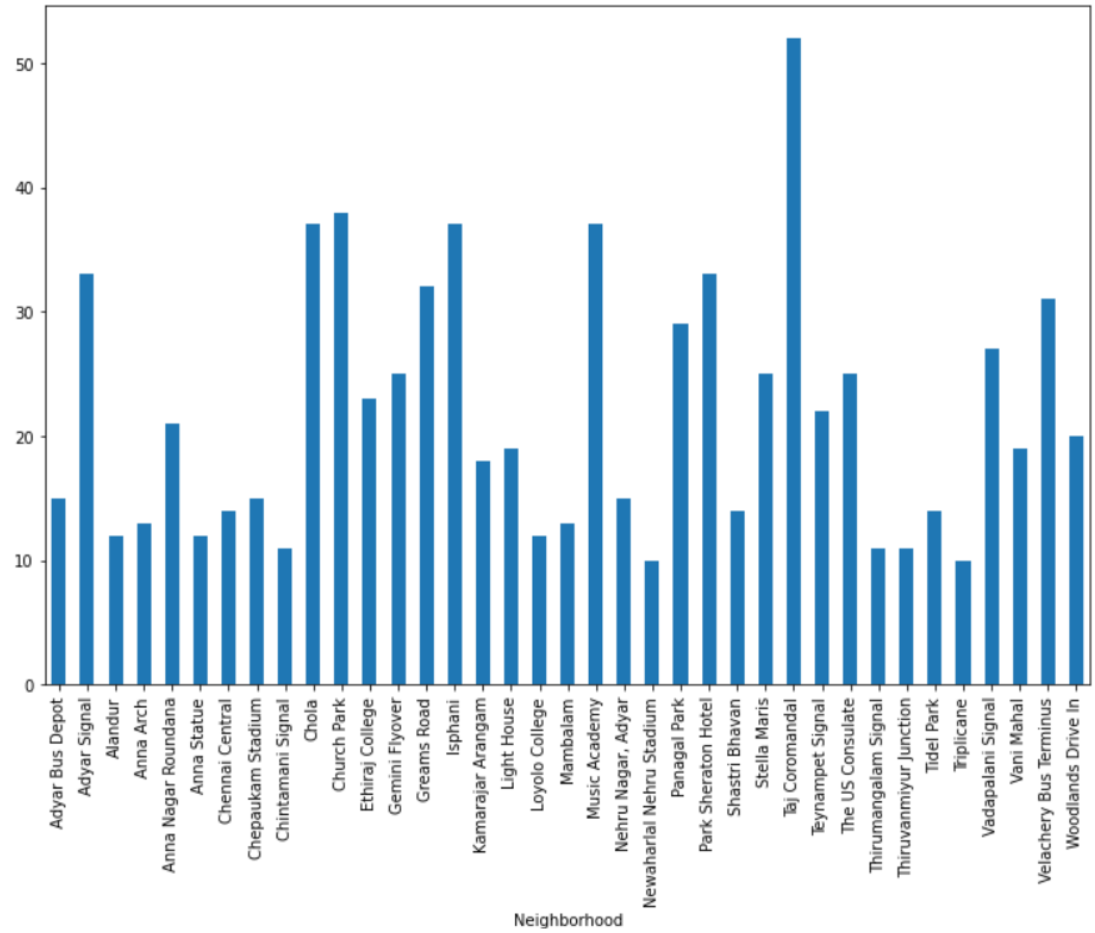
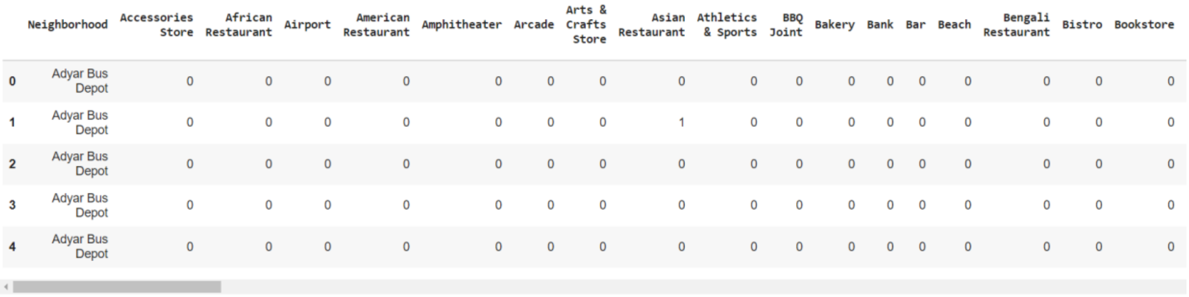
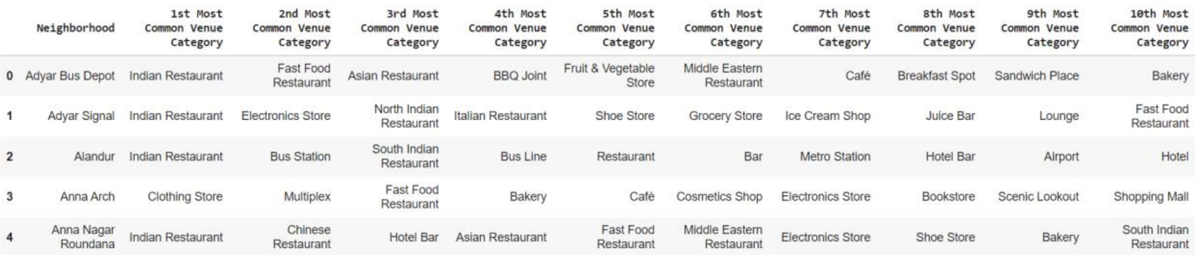


Fig 4.6 A bar chart after removing bias showing the venues which are equal to or more than 10 in number



	Neighborhood	Accessories Store	African Restaurant	Airport	American Restaurant	Amphitheater	Arcade	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bank	Bar	Beach	Bengali Restaurant	Bistro	Bookstore
0	Adyar Bus Depot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Adyar Bus Depot	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	Adyar Bus Depot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Adyar Bus Depot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Adyar Bus Depot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig 4.7 On-hot encoding for the given venue categories



	Neighborhood	1st Most Common Venue Category	2nd Most Common Venue Category	3rd Most Common Venue Category	4th Most Common Venue Category	5th Most Common Venue Category	6th Most Common Venue Category	7th Most Common Venue Category	8th Most Common Venue Category	9th Most Common Venue Category	10th Most Common Venue Category
0	Adyar Bus Depot	Indian Restaurant	Fast Food Restaurant	Asian Restaurant	BBQ Joint	Fruit & Vegetable Store	Middle Eastern Restaurant	Café	Breakfast Spot	Sandwich Place	Bakery
1	Adyar Signal	Indian Restaurant	Electronics Store	North Indian Restaurant	Italian Restaurant	Shoe Store	Grocery Store	Ice Cream Shop	Juice Bar	Lounge	Fast Food Restaurant
2	Alandur	Indian Restaurant	Bus Station	South Indian Restaurant	Bus Line	Restaurant	Bar	Metro Station	Hotel Bar	Airport	Hotel
3	Anna Arch	Clothing Store	Multiplex	Fast Food Restaurant	Bakery	Café	Cosmetics Shop	Electronics Store	Bookstore	Scenic Lookout	Shopping Mall
4	Anna Nagar Roundana	Indian Restaurant	Chinese Restaurant	Hotel Bar	Asian Restaurant	Fast Food Restaurant	Middle Eastern Restaurant	Electronics Store	Shoe Store	Bakery	South Indian Restaurant

Fig. 4.8 A dataframe consisting of the top 10 venue categories for each area of Chennai

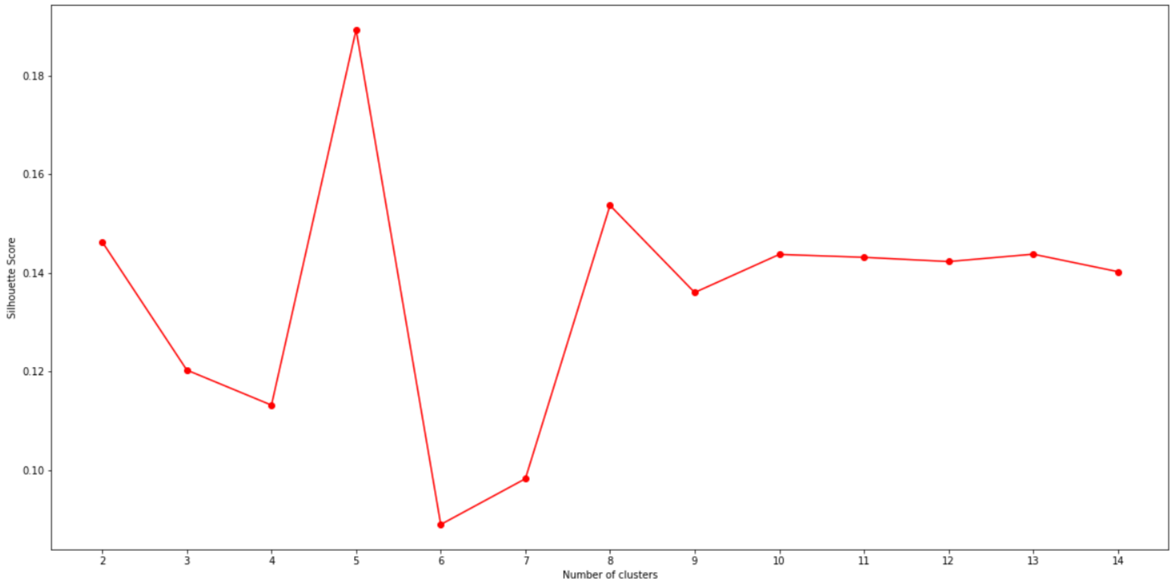
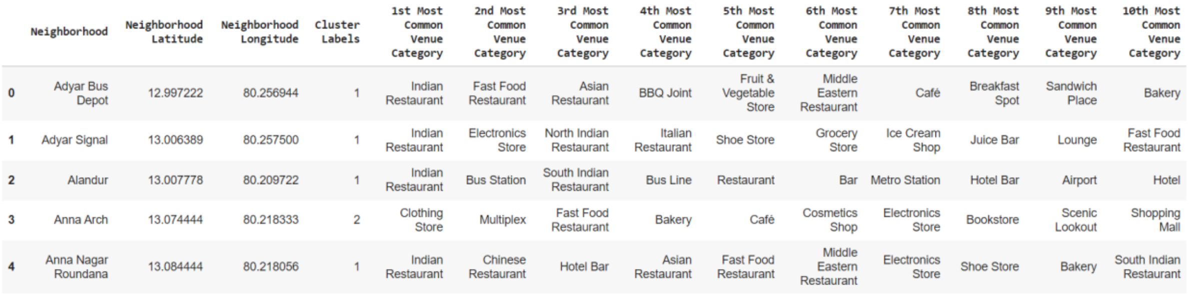


Fig 4.9 The optimised number of clusters is found to be 5 using silhouette score



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Cluster Labels	1st Most Common Venue Category	2nd Most Common Venue Category	3rd Most Common Venue Category	4th Most Common Venue Category	5th Most Common Venue Category	6th Most Common Venue Category	7th Most Common Venue Category	8th Most Common Venue Category	9th Most Common Venue Category	10th Most Common Venue Category
0	Adyar Bus Depot	12.997222	80.256944	1	Indian Restaurant	Fast Food Restaurant	Asian Restaurant	BBQ Joint	Fruit & Vegetable Store	Middle Eastern Restaurant	Café	Breakfast Spot	Sandwich Place	Bakery
1	Adyar Signal	13.006389	80.257500	1	Indian Restaurant	Electronics Store	North Indian Restaurant	Italian Restaurant	Shoe Store	Grocery Store	Ice Cream Shop	Juice Bar	Lounge	Fast Food Restaurant
2	Alandur	13.007778	80.209722	1	Indian Restaurant	Bus Station	South Indian Restaurant	Bus Line	Restaurant	Bar	Metro Station	Hotel Bar	Airport	Hotel
3	Anna Arch	13.074444	80.218333	2	Clothing Store	Multiplex	Fast Food Restaurant	Bakery	Café	Cosmetics Shop	Electronics Store	Bookstore	Scenic Lookout	Shopping Mall
4	Anna Nagar Roundana	13.084444	80.218056	1	Indian Restaurant	Chinese Restaurant	Hotel Bar	Asian Restaurant	Fast Food Restaurant	Middle Eastern Restaurant	Electronics Store	Shoe Store	Bakery	South Indian Restaurant

Fig 4.10 Cluster numbers are assigned to each area depending on their venue categories and prospective of a new venture

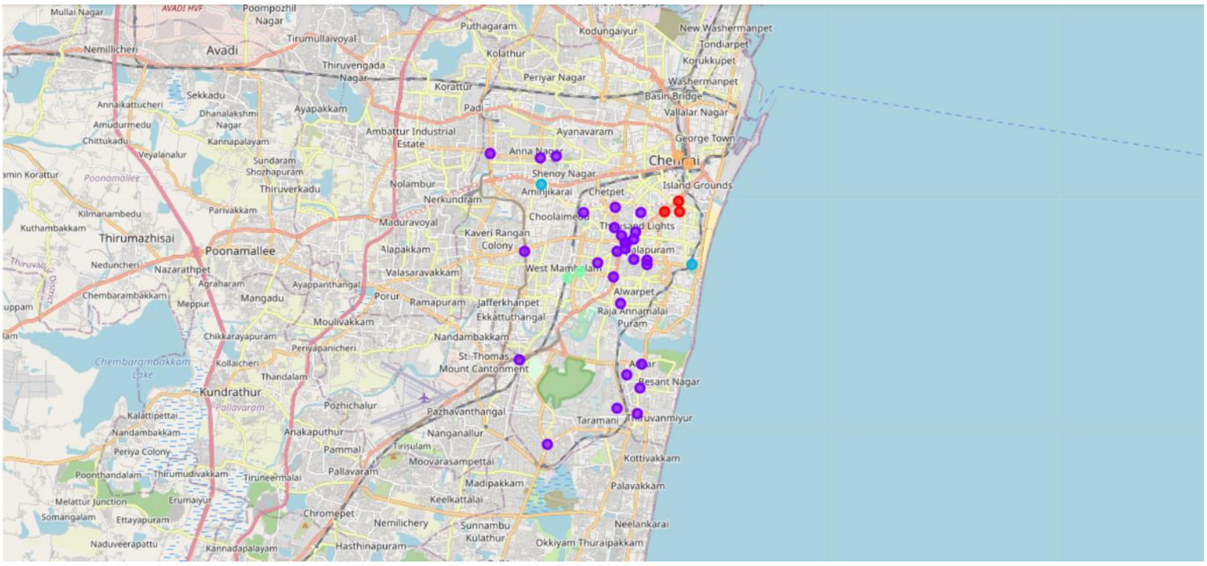


Fig 4.11 Clustering is done on leaflet map using Folium and geopy. Circle markers are used to mark the areas within each cluster. There are a total of 5 clusters

5. Conclusion

This project is successful in deriving clusters of venue categories onto the Leaflet maps which can be used by stakeholders to decide and analyse what kind of business becomes a hit in certain locations. We made use of a dataset that consists of latitude and longitude values of areas of Chennai and merged this with the Foursquare API to get the venue categories of the restaurants located in the city. This will inevidently be useful for both common people and stakeholders as it is visualised in the form of a dataframe (i.e) a table which is user-friendly.

6. Future Works

As days go on, several startups lay their foot on Chennai and so the dataset must be updated regularly. With this dataset, not only more kind of maps can be coded but also with the help of tools such as "R" and "Matlab", one can also make interactive visualisations of the information at hand so it is pleasing to our eyes. A dashboard can be coded with inputs such as area, latitude and longitude and outputs as venue categories.