Helping Tourism in Chennai, India

Tushar Raju

October 15,2019

1. Introduction

1.1 Background

Chennai also known as Madras is the capital of the Indian state of Tamil Nadu. Chennai is among the most-visited Indian cities by foreign tourists. Tourism-guide publisher Lonely Planet named Chennai as one of the top ten cities in the world to visit in 2015. Chennai was ranked the 43rd-most visited city in the world for the year 2015. The Quality of Living Survey rated Chennai as the safest city in India. Chennai attracts 45 percent of health tourists visiting India, and 30 to 40 percent of domestic health tourists. National Geographic mentioned Chennai as the only South Asian city to feature in its 2015 "Top 10 food cities" list. Chennai was also named the ninth-best cosmopolitan city in the world by Lonely Planet. In October 2017, Chennai was added to the UNESCO Creative Cities Network (UCCN) list for its rich musical tradition.

1.2 Problem

In recent days, the tourism of this beautiful city has decreased. Tourism not only adds to the country's economy but is a core part of income. How do we increase tourism and bring back the popularity of the city? In-order to achieve this, we need the tourists to have a good experience and moreover a personalized one. We can achieve this with the help of data science.

1.3 Interest

With all this tourist attraction taken into account, a system that can find a place suitable for the tourist to visit during their stay will be helpful. Our goal is to identify places based on their rating & pricing and make it visible to the tourists in-order for them to choose a place to visit based on their budget and based on the venue's rating. This will eventually increase the tourism level of Chennai.

2. Data Acquisition & Cleaning

2.1 Data Sources

Our sources are the FoursquareAPI & ZomatoAPI. With the help of the co-ordinates of Chennai i.e., 13.0827° N, 80.2707° E we establish the centre point and search for a radius of 5000m i.e., 5km for all venues using the FoursquareAPI. The data we collected included the venue name, category, latitude & longitude. Using the ZomatoAPI, by providing the data we obtained from FoursquareAPI we were able to get the average price for two, rating, price range & address for each corresponding venue.

2.2 Data Cleaning

Data collected from both the FoursquareAPI & the ZomatoAPI were stored in two different data-sets. We had to combine these two data-sets and remove the repeated & unwanted data. For this, we calculated the latitude & longitude difference for each venue in the two datasets & removed the venues which had a latitude difference & longitude difference of greater than 0.0004. On combination of these two data-sets, we were left with a single data-set but we had a few repeated columns such as venue name, latitude & longitude. In-order to remove this, we dropped one of the two repeated columns for each type and finally we were left with a single data-set with no repeated columns. Some of the data had their rating column value to be 0, which clearly meant that the venue was not rated. Hence, we removed the venues with a rating of 0, so that we do not advertise wrong information. Now the data is clean and ready for the next step.

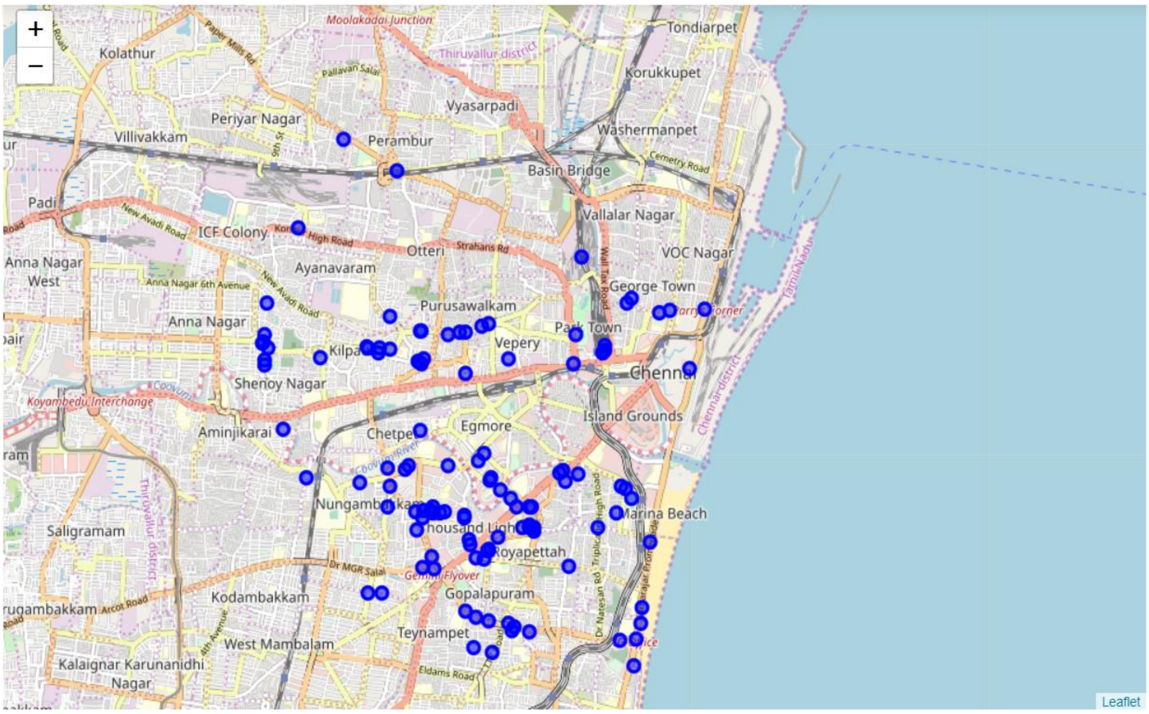
2.3 Feature Selection

After data cleaning, we were left with 75 venues & 8 features. On examining the data, it was found that there was some redundancy in the features. For example, the address column represents the location, but two other features – latitude & longitude also represent the location of the venue. So, to remove the redundancy, we chose to drop the address column because latitude & longitude were given by numerical values which are essential for the clustering process. We will replace the categorical values which are strings with numerical values where 0 represents leisure places such as the park, pool, … 1 represents restaurants, 2 represents quick eats like the donut shop, bakery, …Other features like venue name & review which are strings will be dropped for the next step.

3. Exploratory Data Analysis

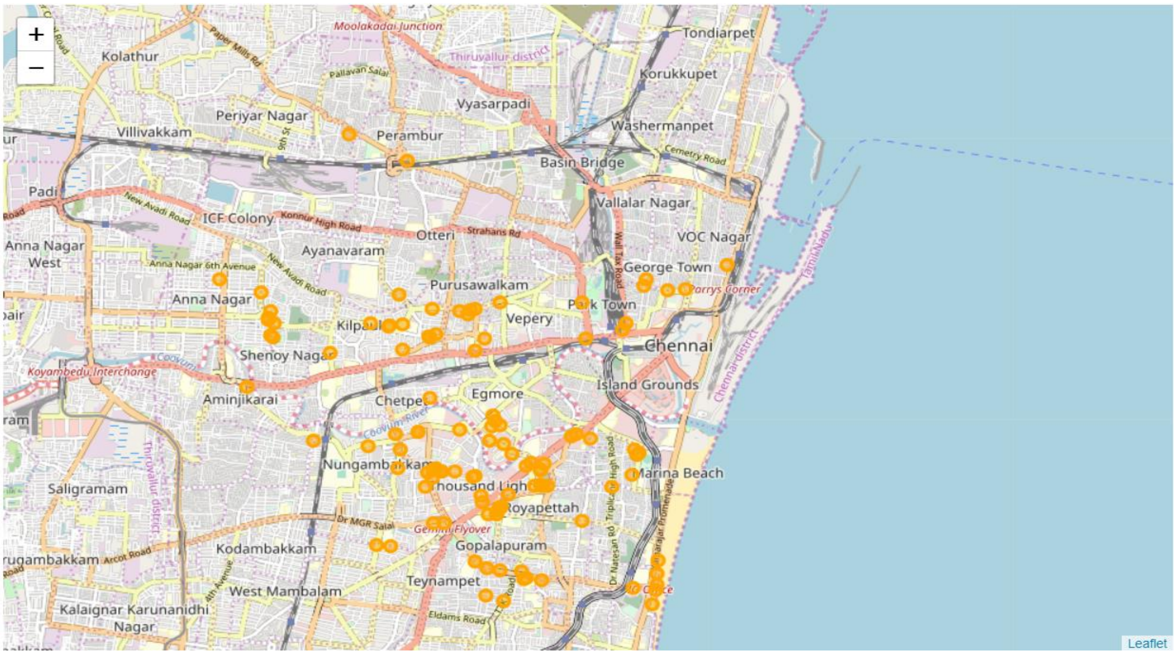
3.1 Initial Maps

FoursquareAPI Venue Map



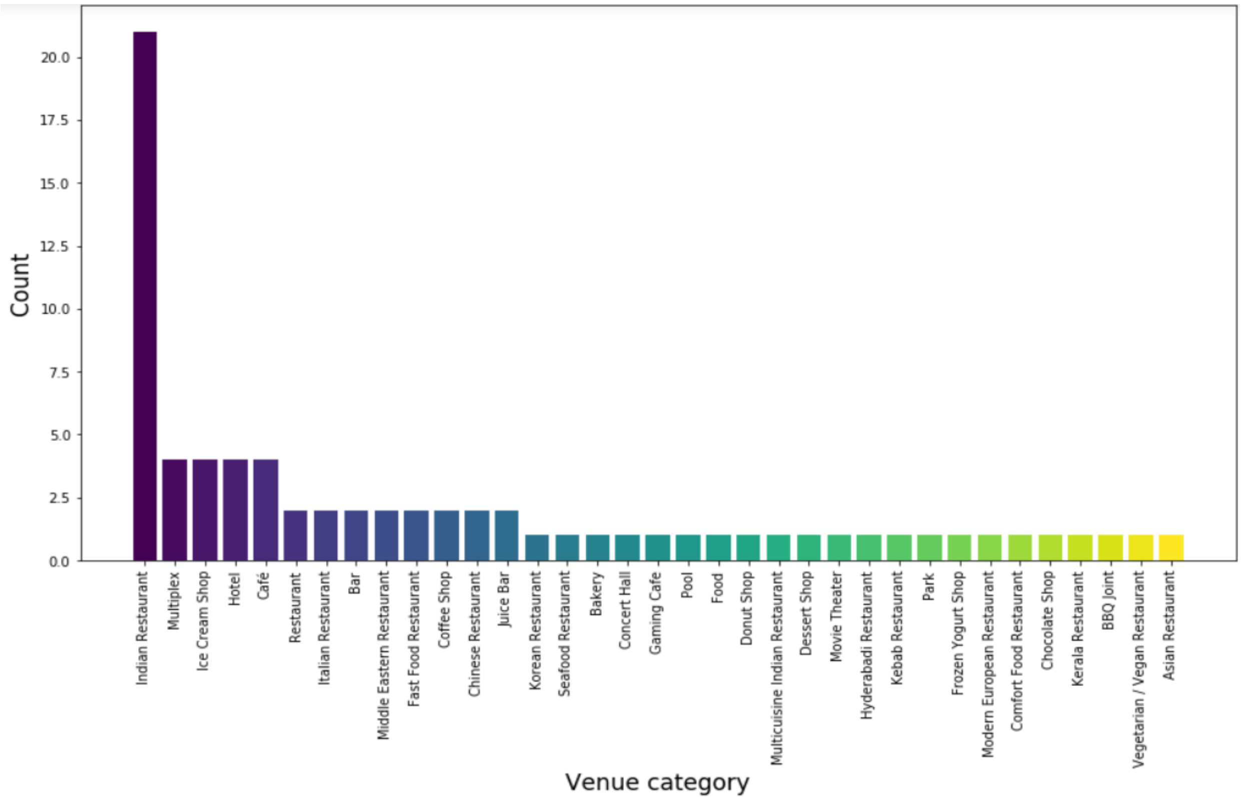
With the help of the FoursquareAPI data-set we retrieved, we plot the venues with the help of their latitude & longitude co-ordinates using the folium library. From the map we can see that the Nungambakkam & Thousand Lights area have the most number of venues in and around them.

ZomatoAPI Venue Map



With the help of the ZomatoAPI data-set we retrieved, we plot the venues with the help of their latitude & longitude co-ordinates using the folium library. From the map we can see that this data-set also has the most number of venues located in and around Nungambakkam & Thousand Lights area but there are slight deviations in the latitude & longitude positions of the venues when compared with the FoursquareAPI venues.

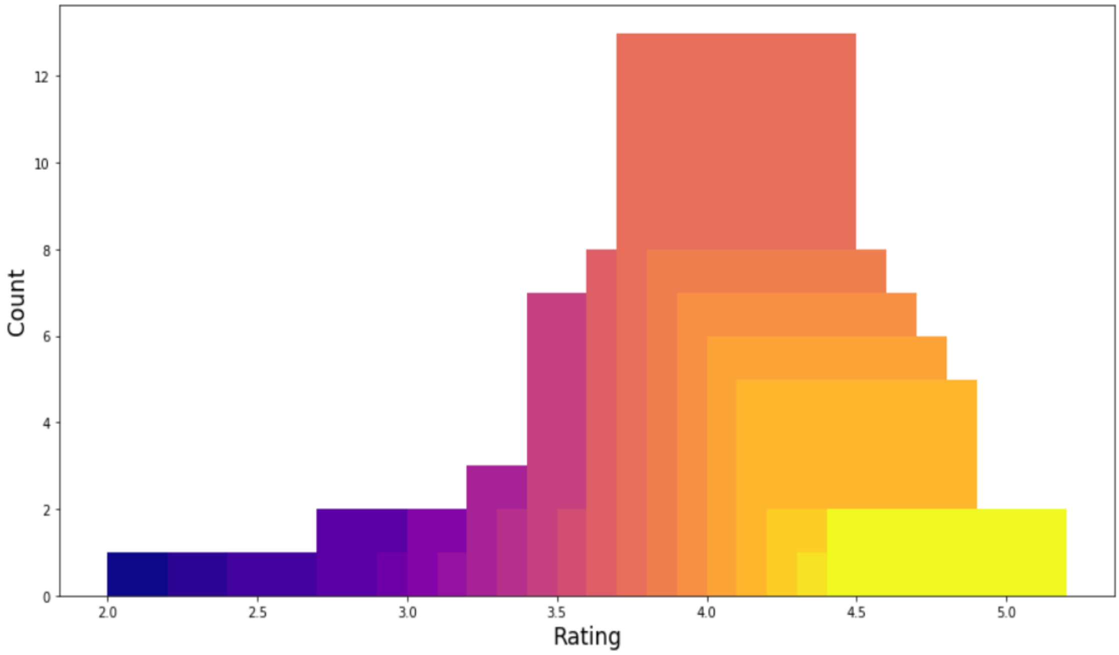
3.2 Venue Category Count -> Bar Plot



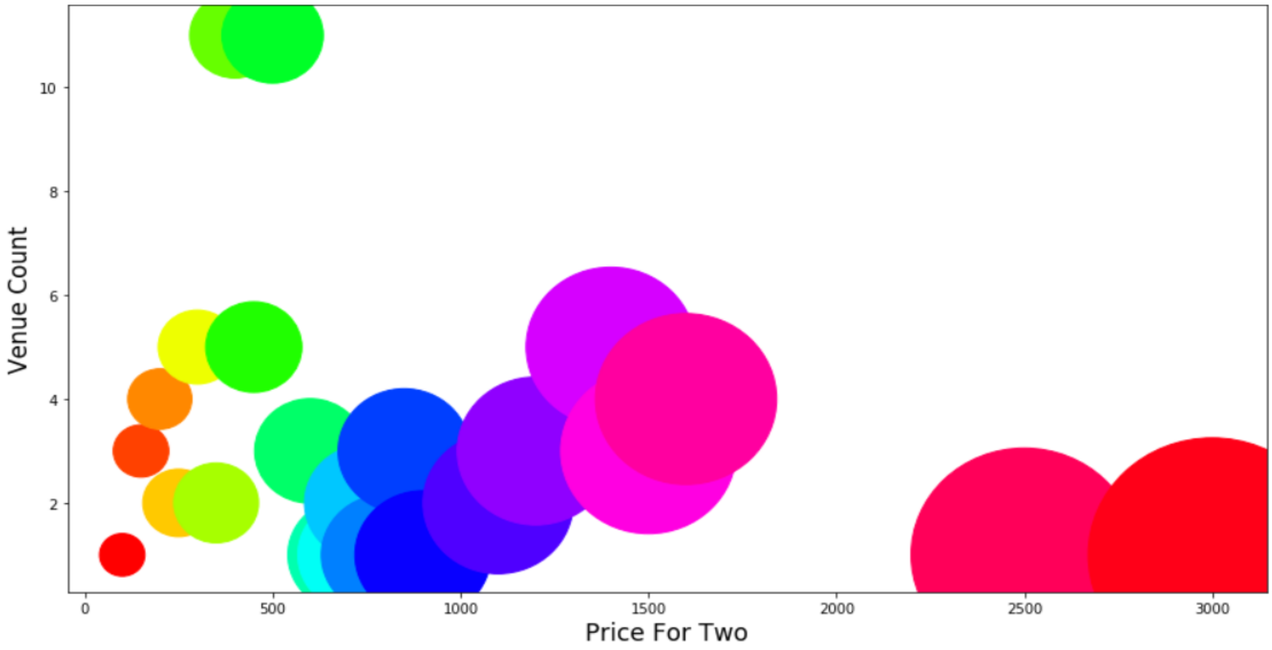
Now to visualize the combined data of both the data-sets, we use a bar plot to plot the venue's category against the count to see what kind of venues are present within a radius of 5km of Chennai and to see the variety of venues also. From the graph we can see that Indian Restaurants are the most common venue in the city. It is great for tourists, since they can have a taste of the south-Indian cuisine and culture.

3.3 Venue Rating Count -> Bar Plot

Now let us visualize the rating of each venue along with their count to see the average rating of the venues within in Chennai with the help of a bar plot. From the graph we can see that, most of the venues have a rating between 3.5 & 4.5. This is good, it ensures the tourists that the venues in Chennai are well taken care of and their customers are satisfied with their experience at the venue.



3.4 Price for two Count -> Scatter Plot



Now let us visualize the average price for two against the number of venues with the help of a scatter plot. From the graph we can see that the most number of venues exist between 400 Rupees and 1500 Rupees. This shows that the venues in the city of Chennai are mostly affordable places. This is another plus point for tourists on a budget trip.

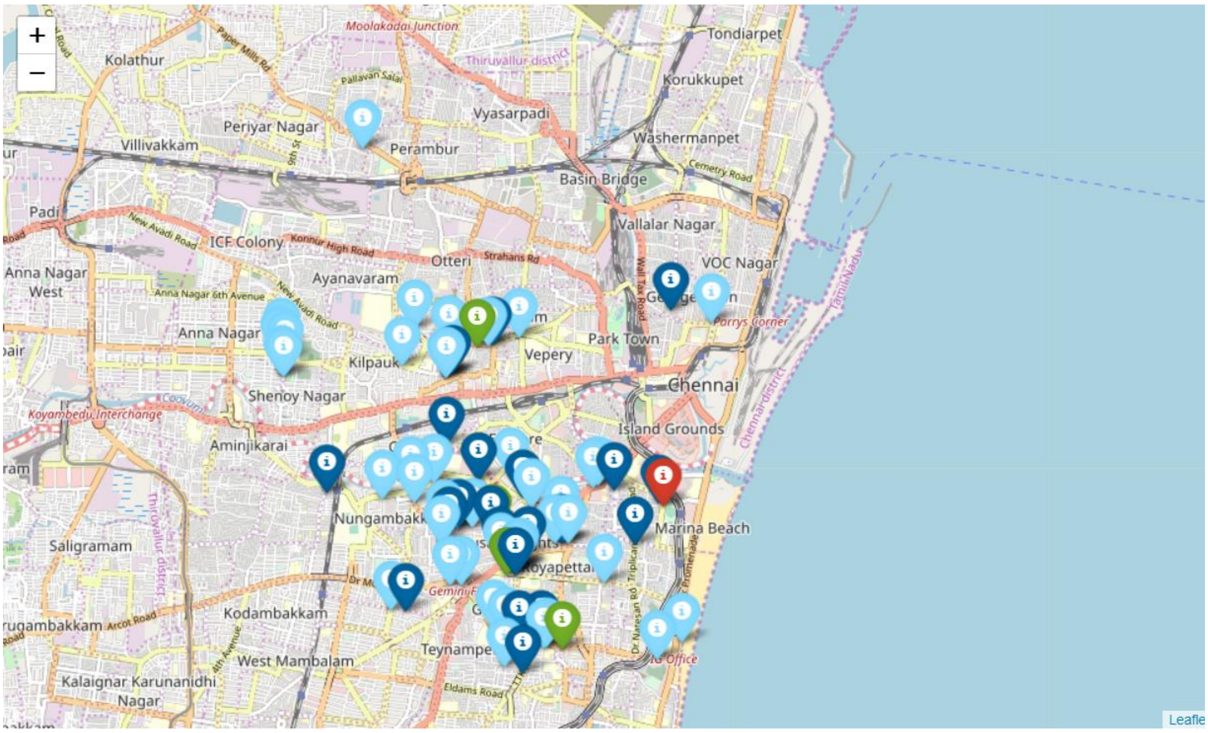
3.5 Map View

We now plot the venues on the map with the help of the folium library and assign markers to them based on their rating.

Now we assign colours according to their rating bins. The corresponding colours are:

- 0-2.5 (Bad) rating is red
- 2.5-3.2 (Average) rating is green
- 3.2-4.2 (Good) rating is light-blue
- 4.2-5.0 (Excellent) rating is dark-blue

Rating Map

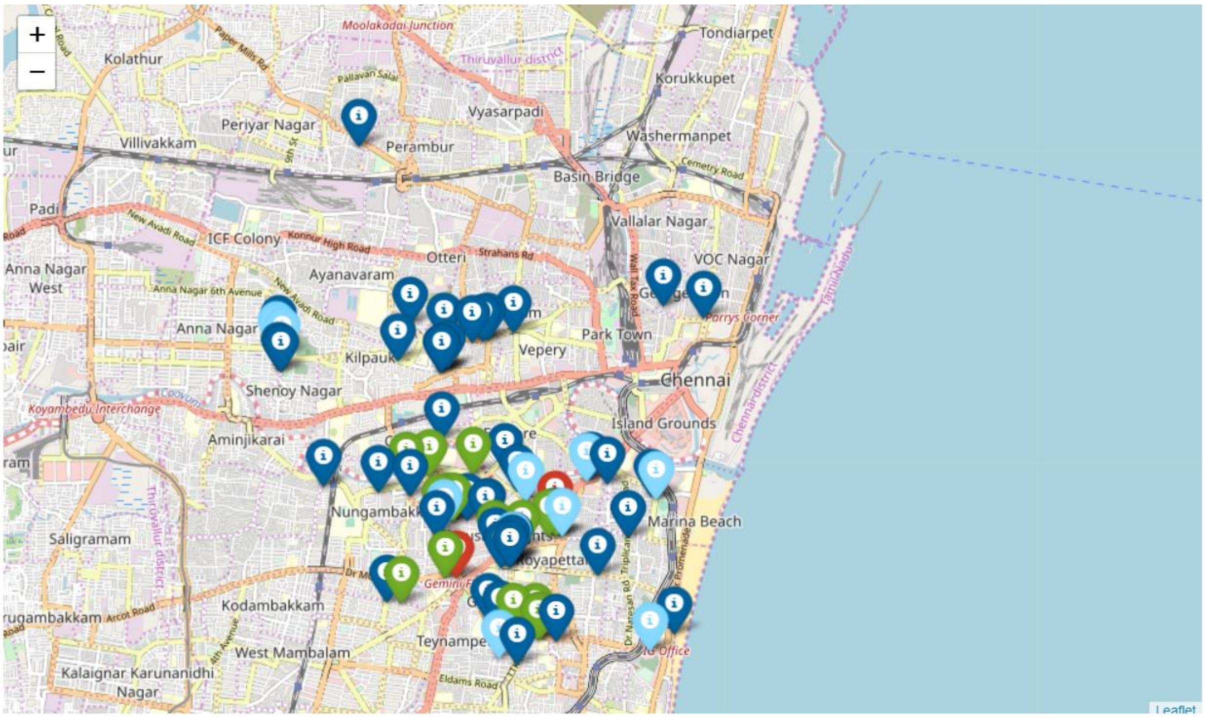


We now plot the venues on the map with the help of the folium library and assign markers to them based on their price.

Colour code:

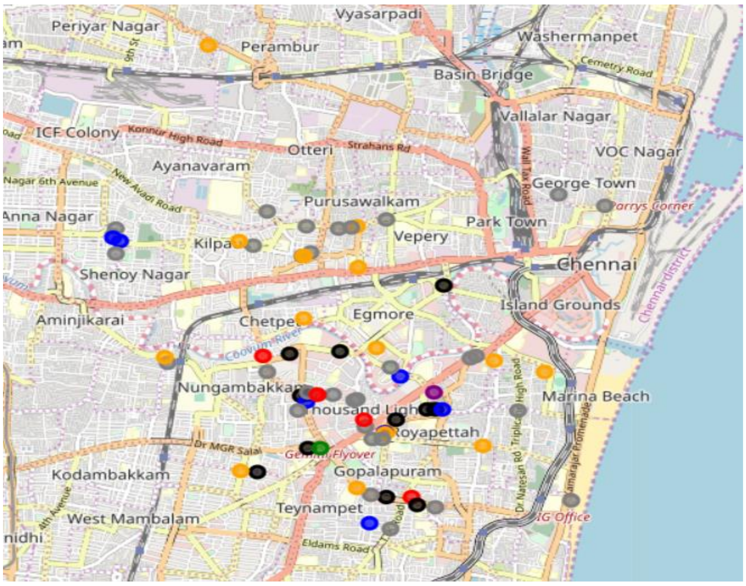
- Cheap is dark-blue
- Average is light-blue
- Affordable is green
- Expensive is red

Pricing Map



4. Clustering

Let us begin the clustering phase. We will be using a method known as Affinity Propagation Clustering. Affinity Propagation is a newer clustering algorithm that uses a graph-based approach to let points 'vote' on their preferred 'exemplar'. The end result is a set of clusters from which we derive clusters by essentially doing what K-Means does and assigning each point to the cluster of its nearest exemplar. Affinity Propagation has some advantages over K-Means. First of all, the graph-based exemplar voting means that the user doesn't need to specify the number of clusters. Second, due to how the algorithm works under the hood with the graph representation it allows for non-metric dissimilarities. On clustering, we are left with 6 clusters. We then plot these clusters onto the map with the help of the folium library.



5.Results & Discussion

Our analysis provides an insight on the venues within a 5km radius of Chennai. We had found that most venues were located in and around Nungambakkam & Thousand Lights. The most common venue category was Indian restaurants, tourists can enjoy the Indian cuisine & Chennai's culture. The city of Chennai mostly consists of venues rated between 3.5 & 4.5, this is evidence that the venues are well taken care of and satisfy their customers. Most venues had their prices between 500 Rupees & 1500 Rupees, this is also evidence that the city of Chennai is an affordable place where you can have a great time at decent prices.

6.Conclusion

The whole purpose of this analysis is to provide information to the tourists visiting Chennai to make decisions based on their preferences and also promote tourism in Chennai. We initially acquired data from FoursquareAPI & ZomatoAPI. we cleaned the data by combining the two datasets & removing venues with similar latitude & longitude co-ordinates. We dropped venues which were not given ratings yet. We the visualized the data in the form of bar plots, scatter plots & maps to come the conclusions mentioned in the results section. We then grouped the data into 3 clusters with the help of Affinity Propagation Clustering. In the end this information will only act as a guide or insight to the tourist. The tourist is the one who will make the final decision on where they want to visit based on the rating, pricing, location etc., this clustered data will assist them in making this decision.