Applied Data Science Capstone Project

Recommender System for Restaurants in Chennai Area

Restaurant recommender system is a model developed to recommend/suggest restaurants to the users as they want and it also considers the users past preferences.

The data needed for this project are collected from websites like Wikipedia, latlong websites, etc

The aim of this project is to create a recommender system that direct us to the correct destination needed as per our search history and filters. We can even find the similar types of restaurants as needed by the user

Table of contents

Introduction

Problem background:

The diversity of the cuisine available is reflective of the social and economic diversity of Chennai. Roadside vendors, tea stalls, South Indian, North Indian, Muslim food, Chinese and Western fast food are all very popular in the city. Udupi restaurants, are very popular and serve predominantly vegetarian cuisine. The Chinese food and the Thai food served in most of the restaurants are can be customized to cater to the tastes of the Indian population. Chennai can also be called a foodie's paradise because of its vast variety of foods and edibles with a touch of Chennai's uniqueness and tradition.

Problem description:

If I travel and keep changing places very frequently. This is very hectic and I get to experience very different types of environment, of which I do not have much knowledge about. In such situation, food can be an important factor for decided how you rate your trips and also recommending it to the people. In such scenarios, we need to find the right place, at reasonable cost, to serve us the best possible way. So there are few questions that must be addressed, such as:

1. How many types of foods are available in the restaurant?
2. Which is the most nearest with good rating?
3. How many "similar" restaurants are available nearby?
4. Do the "similar" restaurants cost more? If so, what specialty do that have?

To address such question, XXYZ Company’s manager decides to allocate this project to me not just to find out solutions to the questions but also build a system that can help in recommending new places based on their ratings and by comparing the previous visits.

Expectations from this recommender system is to get answer for the questions, and in such a way that it uncovers all the perspective of managing recommendations. It is sighted to show:

1. What types of restaurants are present in a particular area?
2. Where are the similar restaurant present based on a preference to particular food?
3. How do different restaurants rank with respect to my preferences?

Target audience:

Target audiences for this project does not limit to a person who keeps travelling but everyone. People could simply decide to look for a similar restaurant all the time because they are addicted to a specific category of food. People who rarely use restaurants would prefer to have the most rated restaurants nearby them and all this could be easily handled by our recommender system. So target for this project is basically everyone who is exploring different places or similar places.

Success rate:

With restaurants evolving, new food categories emerge, hybrid food starts to be more popular, we need a system that could help us access vast number of food varieties. It is impossible for a person to ask each and every one about their visit to a particular place and also not everyone remembers everything. On the other hand, Computers are good at remembering things, and with Machine learning to its peak, it high time technology will by our personal guidance and help us personally based on our likes and dislikes. So people would care about this project as their personal assistance and success rate could certainly increase with time.

Data Collection:

To find a solution to the questions and build a recommender model, we need data and lots of data. Data can answer question which are unimaginable and non-answerable by humans because humans do not have the tendency to analyze such large dataset and produce analytics to find a solutions.

Let's consider the base scenario:

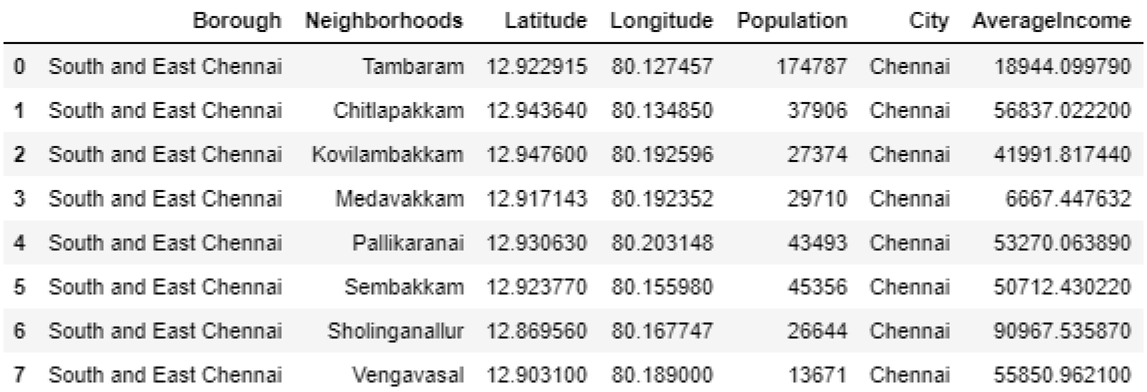
If I want to find a restaurant, then logically, I need 3 things:

1. Its geographical coordinates (latitude and longitude) to find out where exactly it is located.
2. Population of the neighborhood where the restaurant is located.
3. Average income of neighborhood to know how much is the restaurant worth.

Let’s take a closer look at each of these:

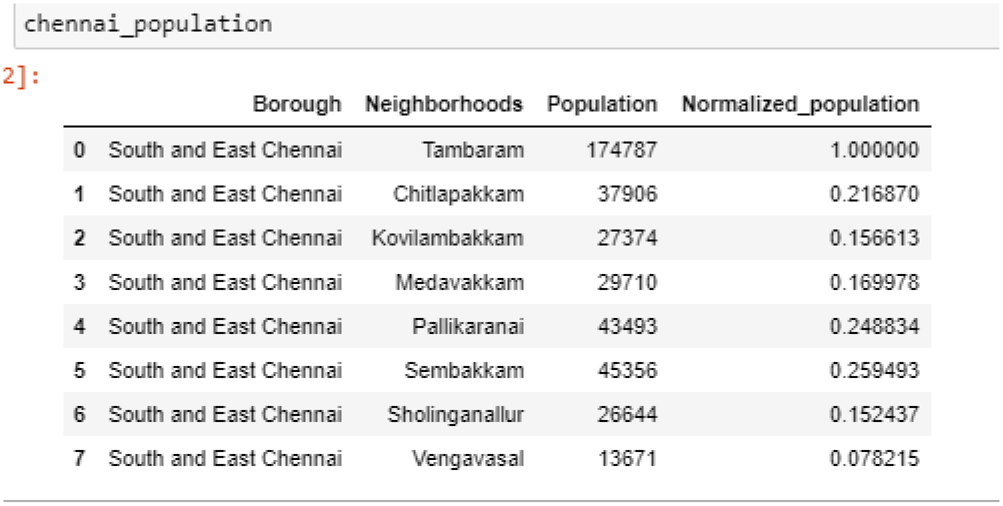
1. To access location of a restaurant, it’s Latitude and Longitude is to be known so that we can point at its coordinates and create a map displaying all the restaurants with its labels respectively.
2. Population of a neighborhood is very important factor in determining a restaurant's growth and amount of customers who turn up to eat. Logically, the more the population of a neighborhood, the more people will be interested to walk openly into a restaurant and less the population, less number of people frequently visit a restaurant. Also if more people visit, better the restaurant is rated because it is accessed by different people with different taste. Hence is is very important factor.
3. Income of a neighborhood is also very important factor as population was. Income is directly proportional to richness of a neighborhood. If people in a neighborhood earns more than an average income, then it is very much possible that they will spend more however not always true with very less probability. So a restaurant assessment is proportional to income of a neighborhood.

1. Collecting geographical coordinates is not difficult but it is not available on open source data websites such as Wikipedia, government website, census report websites etc. So I decided to use latlong.net website to fetch latitude and longitude but this has limited number of calls that I could make with my free account. So it would take around 2-3 days to fetch location of all the neighborhoods in Chennai. Initially i scrapped list of neighbor's using beautifulSoup4 from Wikipedia. The table headings becoming the boroughs and data becoming the neighborhoods. So I manually googled each neighborhood to find its corresponding latitude and longitude. After doing so, I produced the data frame.



	Borough	Neighborhoods	Latitude	Longitude	Population	City	AverageIncome
0	South and East Chennai	Tambaram	12.922915	80.127457	174787	Chennai	18944.099790
1	South and East Chennai	Chitlapakkam	12.943640	80.134850	37906	Chennai	56837.022200
2	South and East Chennai	Kovilambakkam	12.947600	80.192596	27374	Chennai	41991.817440
3	South and East Chennai	Medavakkam	12.917143	80.192352	29710	Chennai	6667.447632
4	South and East Chennai	Pallikaranai	12.930630	80.203148	43493	Chennai	53270.063890
5	South and East Chennai	Sembakkam	12.923770	80.155980	45356	Chennai	50712.430220
6	South and East Chennai	Sholinganallur	12.869560	80.167747	26644	Chennai	90967.535870
7	South and East Chennai	Vengavasal	12.903100	80.189000	13671	Chennai	55850.962100

2. Population by neighborhood is again easy to find out given that it's readily available. I have taken population data from the following link, https://www.census2011.co.in/census/metropolitan/435-chennai.html

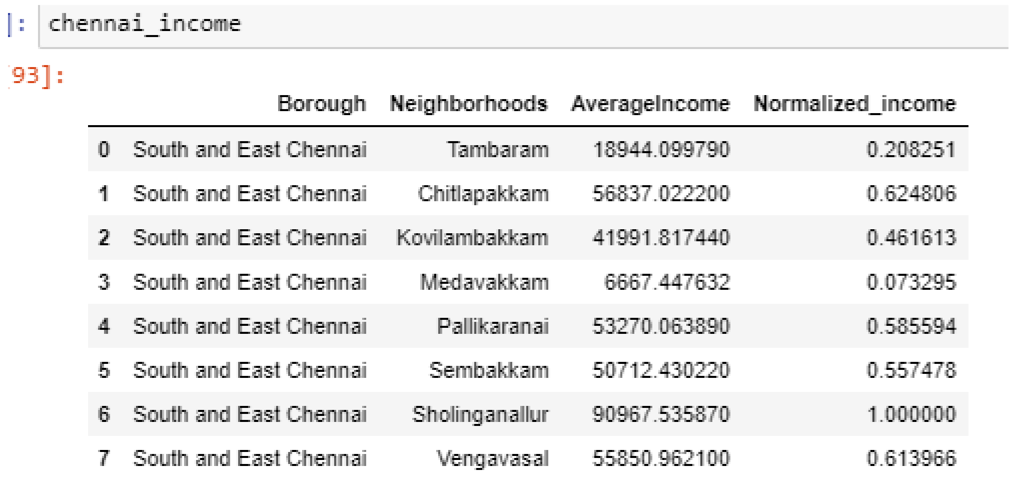


```
chennai_population
```

2]:

	Borough	Neighborhoods	Population	Normalized_population
0	South and East Chennai	Tambaram	174787	1.000000
1	South and East Chennai	Chitlapakkam	37906	0.216870
2	South and East Chennai	Kovilambakkam	27374	0.156613
3	South and East Chennai	Medavakkam	29710	0.169978
4	South and East Chennai	Pallikaranai	43493	0.248834
5	South and East Chennai	Sembakkam	45356	0.259493
6	South and East Chennai	Sholinganallur	26644	0.152437
7	South and East Chennai	Vengavasal	13671	0.078215

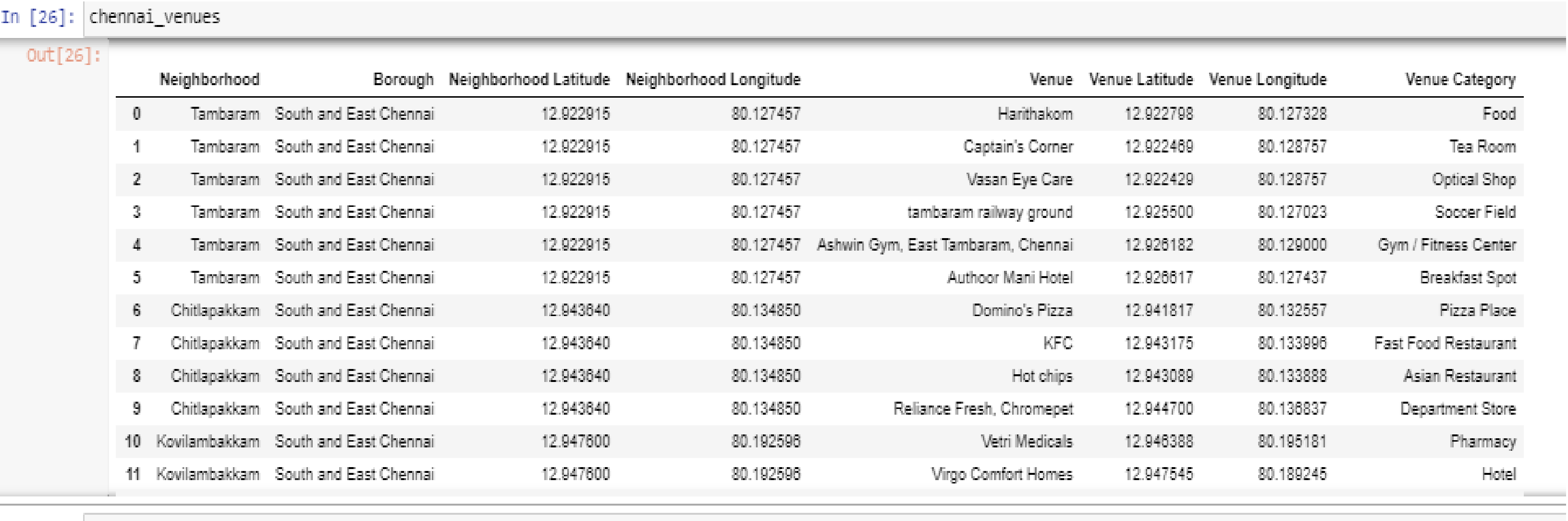
3. It was very tough to find the income of people based on the demographic and it was not available also. So the income of people are assumed.



```
|: chennai_income
93]:
```

	Borough	Neighborhoods	AverageIncome	Normalized_income
0	South and East Chennai	Tambaram	18944.099790	0.208251
1	South and East Chennai	Chitlapakkam	56837.022200	0.624806
2	South and East Chennai	Kovilambakkam	41991.817440	0.461613
3	South and East Chennai	Medavakkam	6667.447632	0.073295
4	South and East Chennai	Pallikaranai	53270.063890	0.585594
5	South and East Chennai	Sembakkam	50712.430220	0.557478
6	South and East Chennai	Sholinganallur	90967.535870	1.000000
7	South and East Chennai	Vengavasal	55850.962100	0.613966

Use of foursquare is focused to fetch nearest venue locations so that we can use them to form a cluster. Foursquare api leverages the power of finding nearest venues in a radius (in my case: 500mts) and also corresponding coordinates, venue location and names. After calling, the following data frame is created:



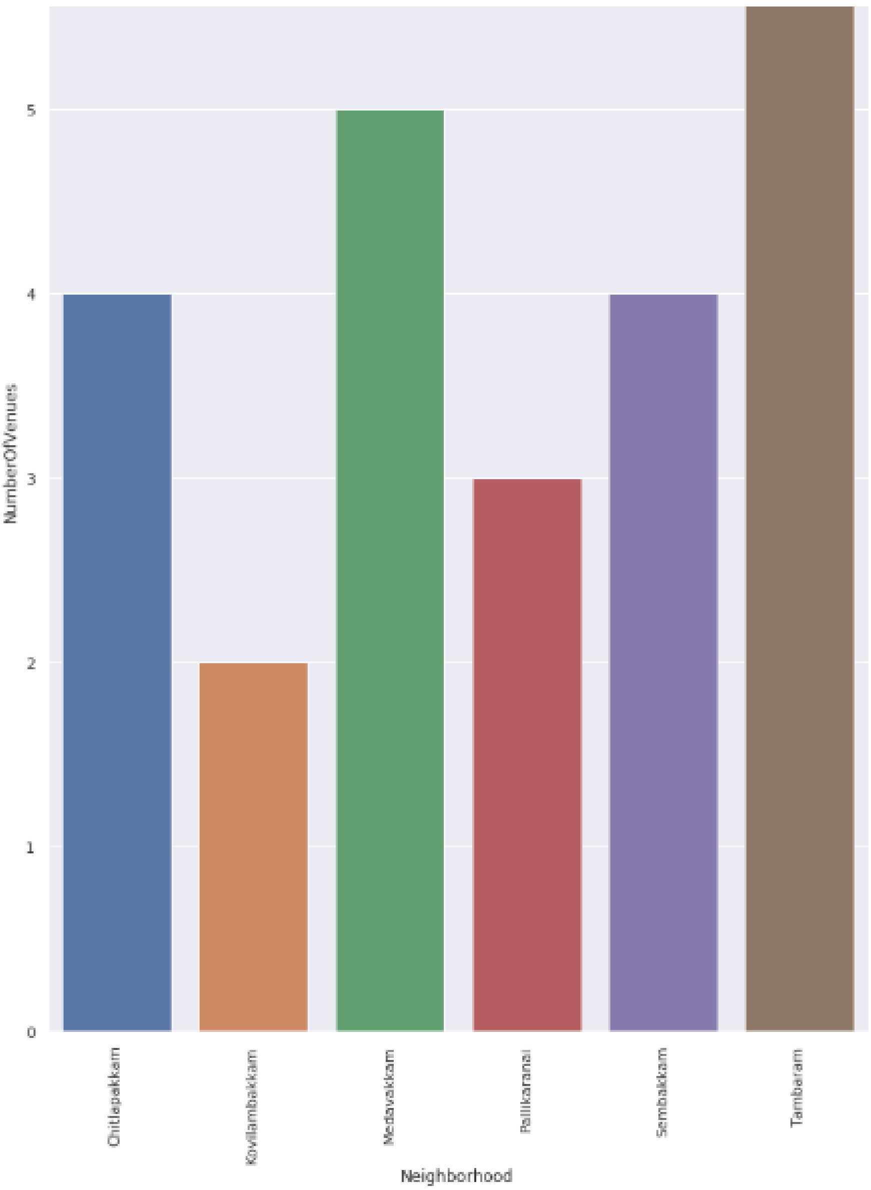
```
In [26]: chennai_venues
Out[26]:
```

	Neighborhood	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Tambaram	South and East Chennai	12.922915	80.127457	Harithakom	12.922798	80.127328	Food
1	Tambaram	South and East Chennai	12.922915	80.127457	Captain's Corner	12.922469	80.128757	Tea Room
2	Tambaram	South and East Chennai	12.922915	80.127457	Vasan Eye Care	12.922429	80.128757	Optical Shop
3	Tambaram	South and East Chennai	12.922915	80.127457	tambaram railway ground	12.925500	80.127023	Soccer Field
4	Tambaram	South and East Chennai	12.922915	80.127457	Ashwin Gym, East Tambaram, Chennai	12.926182	80.129000	Gym / Fitness Center
5	Tambaram	South and East Chennai	12.922915	80.127457	Authoor Mani Hotel	12.926617	80.127437	Breakfast Spot
6	Chitlapakkam	South and East Chennai	12.943640	80.134850	Domino's Pizza	12.941817	80.132557	Pizza Place
7	Chitlapakkam	South and East Chennai	12.943640	80.134850	KFC	12.943175	80.133996	Fast Food Restaurant
8	Chitlapakkam	South and East Chennai	12.943640	80.134850	Hot chips	12.943089	80.133888	Asian Restaurant
9	Chitlapakkam	South and East Chennai	12.943640	80.134850	Reliance Fresh, Chromepet	12.944700	80.136837	Department Store
10	Kovilambakkam	South and East Chennai	12.947600	80.192596	Vetri Medicals	12.946388	80.195181	Pharmacy
11	Kovilambakkam	South and East Chennai	12.947600	80.192596	Virgo Comfort Homes	12.947545	80.189245	Hotel

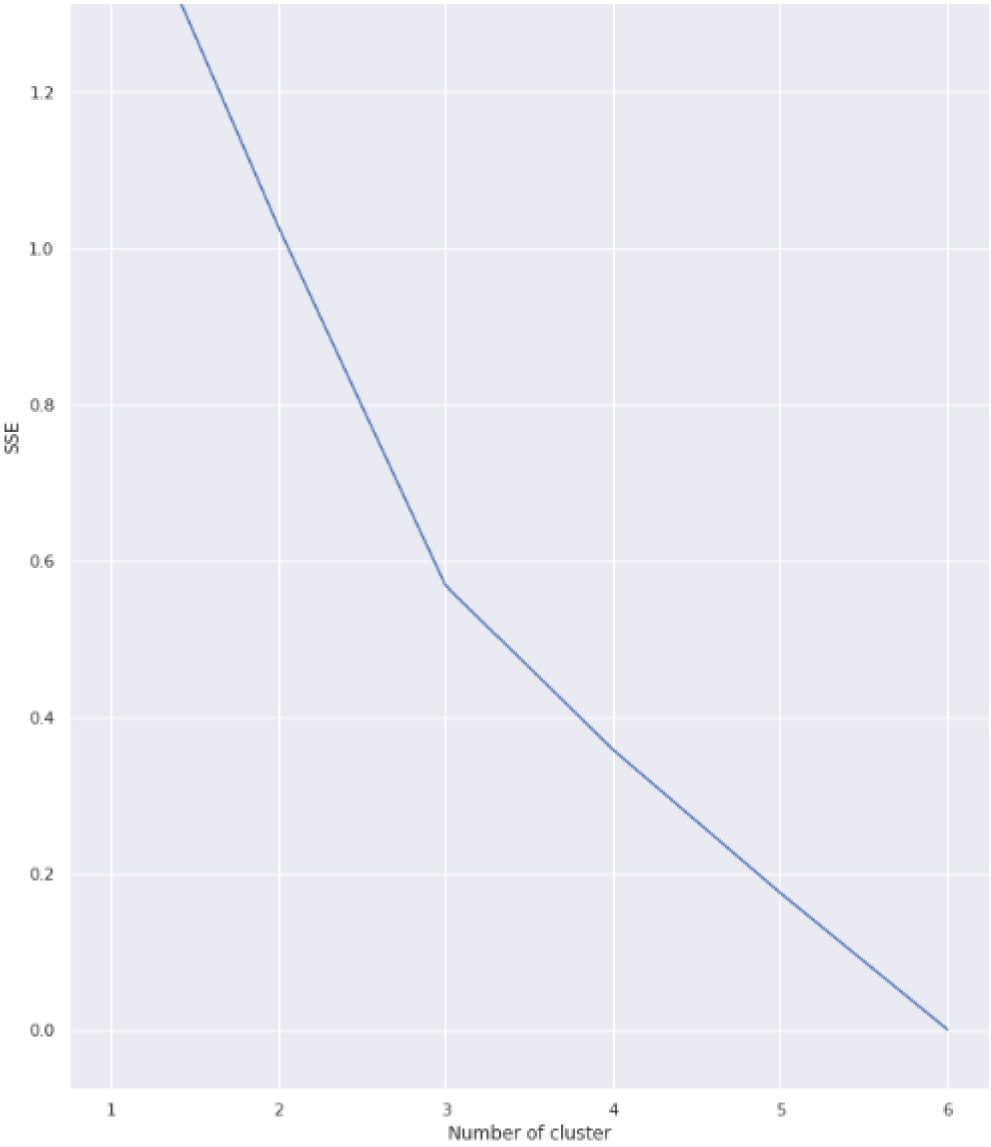
Methodology

Exploratory analysis:

Scrapping the data from different sources and then combining it to form a single-ton dataset is a difficult task. To do so, we need to explore the current state of dataset and then list up all the features needed to be fetched. Exploring the dataset is important because it gives you initial insights and may help you to get partial idea of the answers that you are looking to find out from the data.



Also while producing graph for number of cluster, I produced a graph to explore all the values for n clusters and then finding the best by exploring the elbow graph.



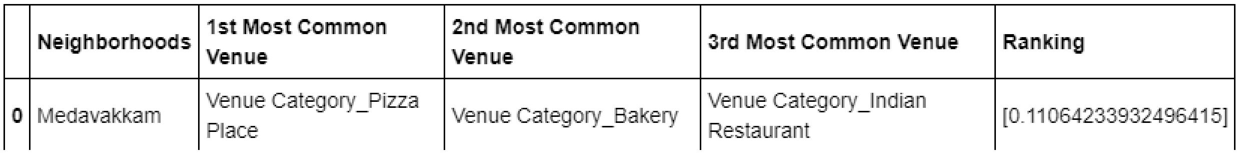
Inferential analysis:

Most important factors while building the recommender system were population and income. They are the most import factor because they have a nonlinear relationship according to our dataset. It needed to make some inferential analysis to understand this nonlinear relationship. As the amount of population increases, it does not necessarily mean that average income of a neighborhood will also increase. It is true to most of the case but also many cases differ to follow this trend. Similarly, a neighborhood with less number of people may not necessarily have less average income. It is possible to have less number of people and more income and vice versa.

Result:

The result of the recommender system is that it produces a list of top restaurants and the most common venue item that the user can enjoy. During the runtime of the model, a simulation was done by taking ‘Tambaram’ as the neighborhood and then processed through our model so that it could recommend neighborhoods with similar characters as that of ‘Tambaram’.

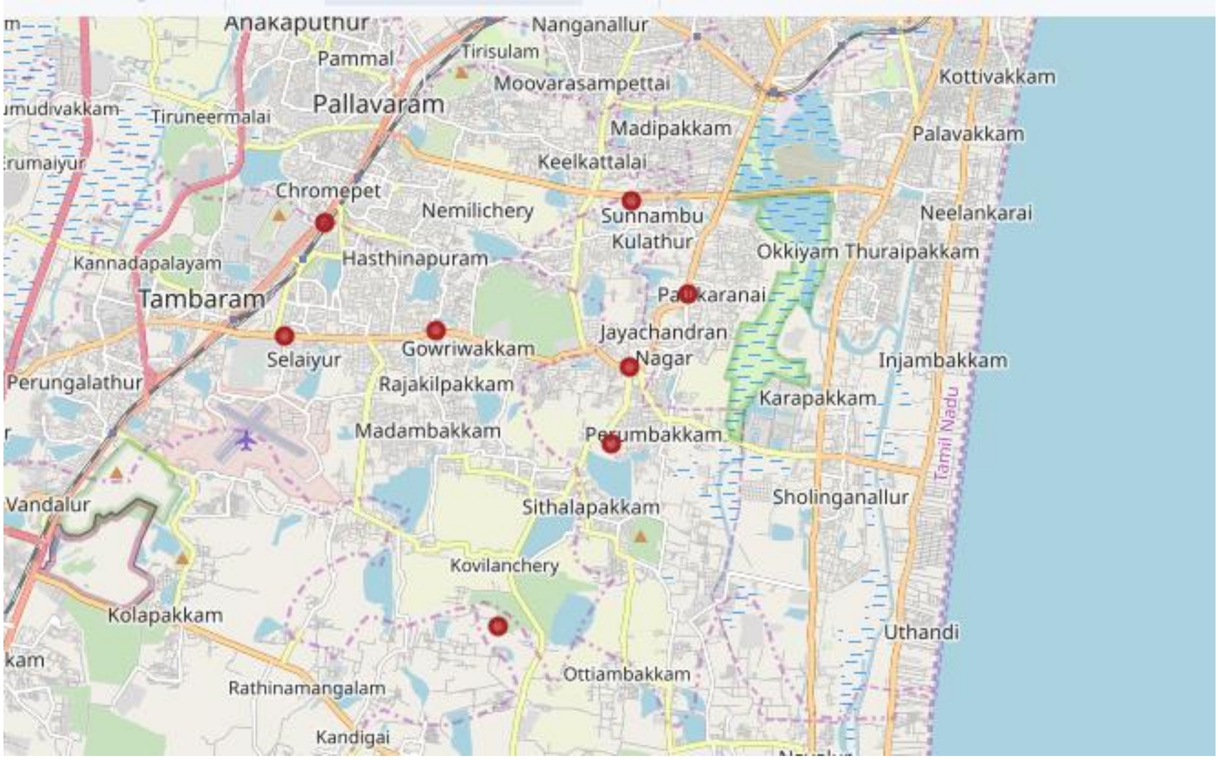
The following image shows the result:



	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Medavakkam	Venue Category_Pizza Place	Venue Category_Bakery	Venue Category_Indian Restaurant	[0.11064233932496415]

Discussion:

Since there was a nonlinear relationship between income and population, it can be concluded that we must always perform inferential approach to find relationship among different set of features. Also during clustering, similar neighborhoods must be dumped into the right cluster. The following graph shows the clusters:



Conclusion:

The recommender system is a system that considers factors such as population, income and makes use of Foursquare API to determine nearby venues. It is a powerful data driven model whose efficiency may decrease with more data but accuracy will increase. It will help users to finish their search by providing the best recommendation to fulfil their needs.