BATTLE OF NEIGHBORHOODS: RESTAURANT RECOMMENDATION

06/24/2020

Applied Data Science Capstone Project

Restaurant recommendation system is developed as a demonstration for the Applied Data Science Capstone project. It recommends restaurants in the Indian city of Bangalore based on user's likes and dislikes and their previous interest data.

Table of contents

Introduction

Background

The city of Bangalore is the capital city of the state of Karnataka, India. It is also the 3rd largest city in India with a population of over 15 million.

The city is known for its IT industry and young engineers working in the IT industry. These young employees come to Bangalore from all over the country and are used to different cuisines from their native places. Many of the IT companies are also MNCs and regularly host visitors from various parts of the world. For this reason, a diverse cuisine is available in Bangalore and reflects the demographic and economic diversity of the city. From roadside vendors to high end restaurants, one can find food from different parts of India as well as from many parts of the world.

For many new people visiting the city, the food scene in Bangalore and their own respective neighborhoods can dictate many decisions. People would prefer to find places to work and live in an area where they can find good food at reasonable prices. Therefore, a recommender system for restaurants in a neighborhood can be very useful. It can be used to answer few questions such as

- How many different cuisines are available in any given neighborhood
- Which is the closest restaurant with a good rating
- How many 'similar' restaurants of my choice are available in the neighborhood
- Are the 'similar' restaurants comparable in terms of cost and ratings

To address this issue, a restaurant recommendation system is required. The expected outcome of this recommender system is to determine

- The types of cuisines and restaurants available in any given neighborhood
- The presence of similar restaurants based on the preference to a particular cuisine/type of food
- Ranking different restaurants according to user preferences

Target Audience

The target audience for this data analysis project is anyone interested in finding a restaurant according to their preferences. People may want to try new restaurants according to their preferred food type in the same or a different neighborhoood. People who do not eat out much may want to try a restaurant based on the type of cuisine of their choice. A new person to the city might want to look for restaurants that offer food that they prefer.

Essentially, this recommender system can be used or modified to choose a restaurant based on individual preference by anyone just looking to explore the food options in the city of Bangalore, India.

Success criteria

With such diverse cuisine available, we need a system that could help residents of Bangalore explore the food scene in the city. It is humanly not possible to check with a large number of individuals manually for their preferences and recommendations. However, using the already available data from sources such as Foursquare API, computer tools can be used to create a recommendation system based on ratings from other users. This system can be used to find and explore restaurants based on personal likes and dislikes. For this project, a user will find similar neighborhoods offering vegetarian food options.

Data

Data requirements

To find a restaurant based on particular preference, the following data can be useful:

- Geographical coordinates of a venue to determine the neighborhood
- Population of the neighborhood
- Average income of neighborhood

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 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.

Data collection

1. The geographical location data can be collected from Google API to fetch the longitude and latitude of a location. Since Google API has limitations on number of API calls using a free account, this restricts the amount of data that can be gathered using Google API. Additional data can be gathered using the Beautifulsoup4 package from Wikipedia to determine the neighborhoods within Bangalore city. Once the neighborhoods are determined, it is easier to get the coordinates for a smaller dataset with less than 75 neighborhoods.

Borough	Neighborhoods	Latitude	Longitude
Central	Cantonment area	12.972442	77.580643
Central	Domlur	12.960992	77.638726
Central	Indiranagar	12.971891	77.641151
Central	Jeevanbheemanagar	12.962900	77.659500
Central	Malleswaram	13.003100	77.564300
Central	Pete area	12.962700	77.575800
Central	Rajajinagar	12.990100	77.552500
Central	Sadashivanagar	13.006800	77.581300
Central	Seshadripuram	12.993500	77.578700
Central	Shivajinagar	12.985700	77.605700

2. The population details is available from various census data. One of the sources is https://indikosh.com/dist/655489/bangalore

	Borough	Neighborhoods	Population	Normalized_population
0	Central	Cantonment area	866377	0.880810
1	Central	Domlur	743186	0.755567
2	Central	Indiranagar	474289	0.482190
3	Central	Jeevanbheemanagar	527874	0.536668
4	Central	Malleswaram	893629	0.908516

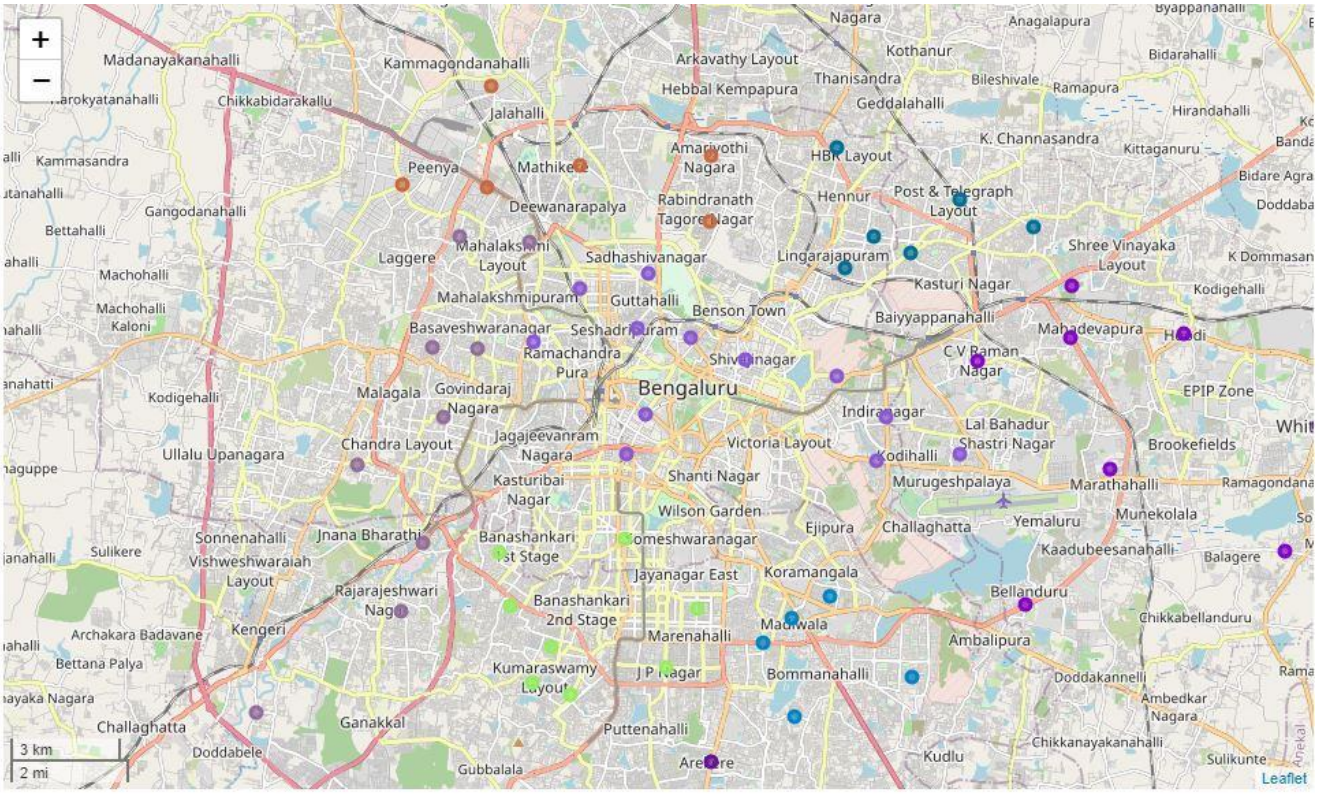
3. Neighborhood income is also available from sources such as https://en.wikipedia.org/wiki/List_of_Indian_cities_by_GDP_per_capita

	Borough	Neighborhoods	AverageIncome	Normalized_income
0	Central	Cantonment area	18944.099792	0.293051
1	Central	Domlur	56837.022198	0.879225
2	Central	Indiranagar	41991.817435	0.649581
3	Central	Jeevanbheemanagar	6667.447632	0.103140
4	Central	Malleswaram	53270.063892	0.824047

4. Foursquare API:

The Foursquare API is used to get nearest venues and to form clusters of neighborhoods. A radius of 500m is used to get the neighborhood, its coordinates, venues and location names.:

	Neighborhood	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cantonment area	Central	12.972442	77.580643	Hotel Fishland	12.975569	77.578592	Seafood Restaurant
1	Cantonment area	Central	12.972442	77.580643	Sapna Book House	12.976355	77.578461	Bookstore
2	Cantonment area	Central	12.972442	77.580643	Vasudev Adigas	12.973707	77.579257	Indian Restaurant
3	Cantonment area	Central	12.972442	77.580643	Adigas Hotel	12.973554	77.579161	Restaurant
4	Cantonment area	Central	12.972442	77.580643	Kamat Yatrinivas	12.975985	77.578125	Indian Restaurant

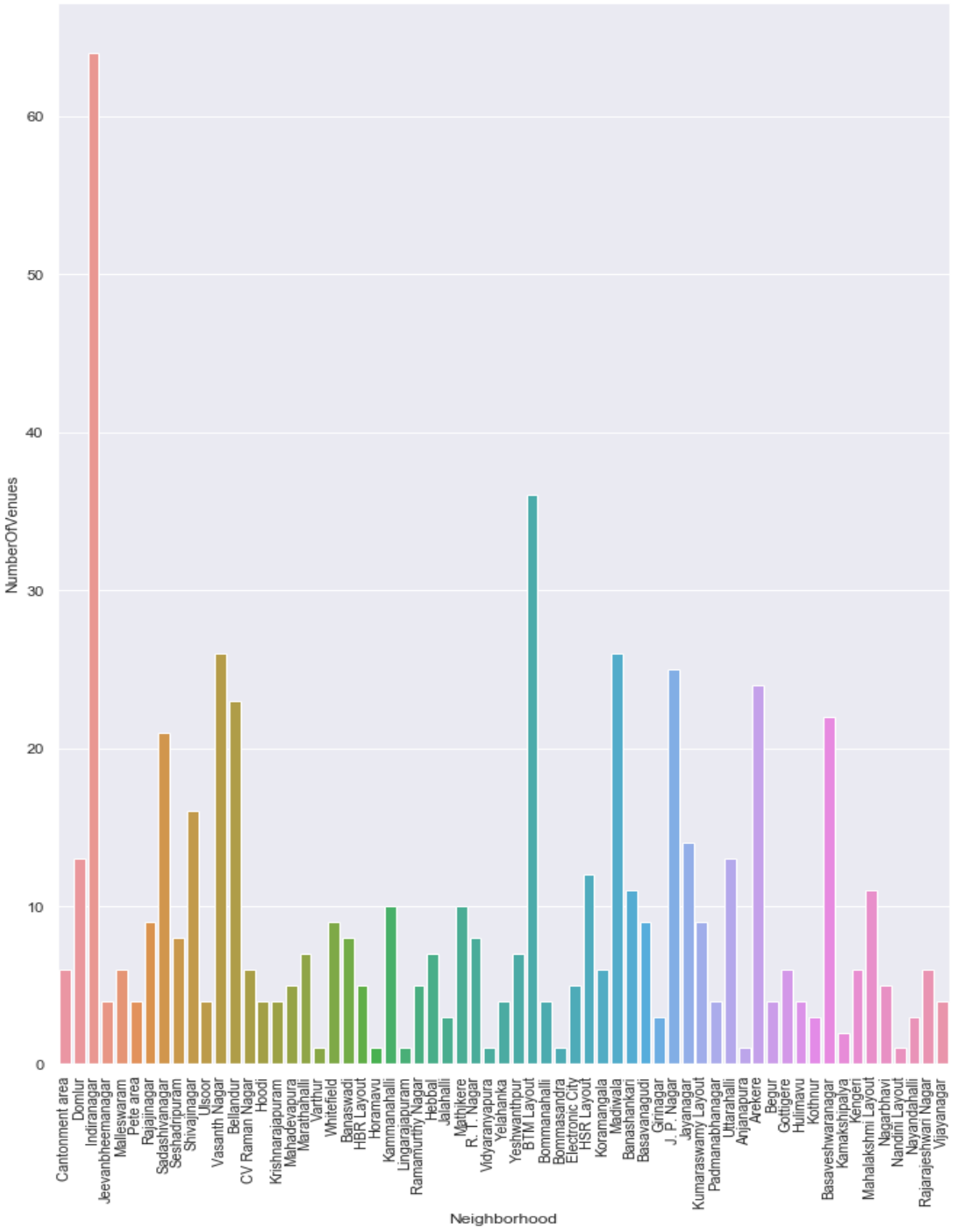


Methodology

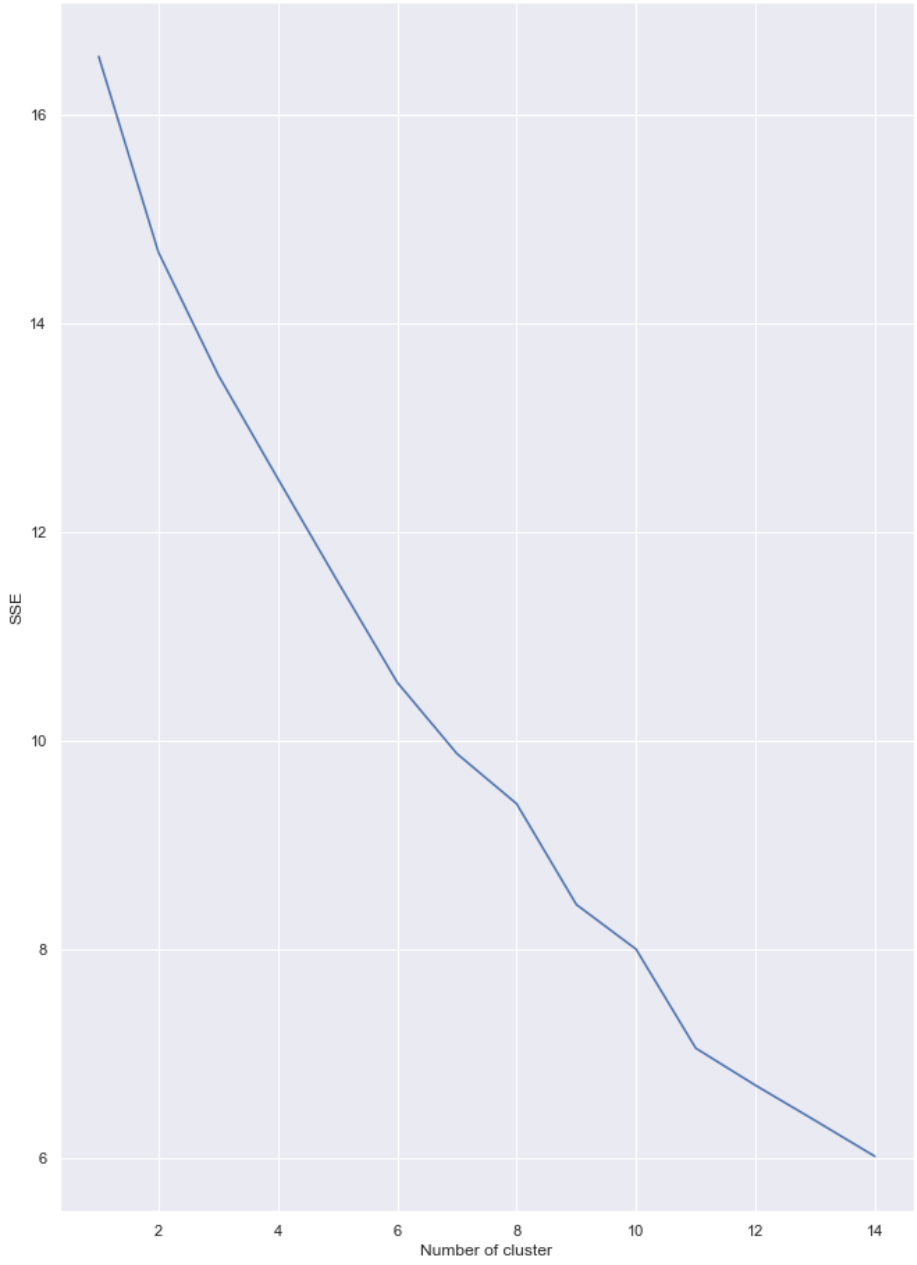
Exploratory analysis

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.

While exploring the dataset, I found out that Inderanagar has most number of venues while Varthur has the least.



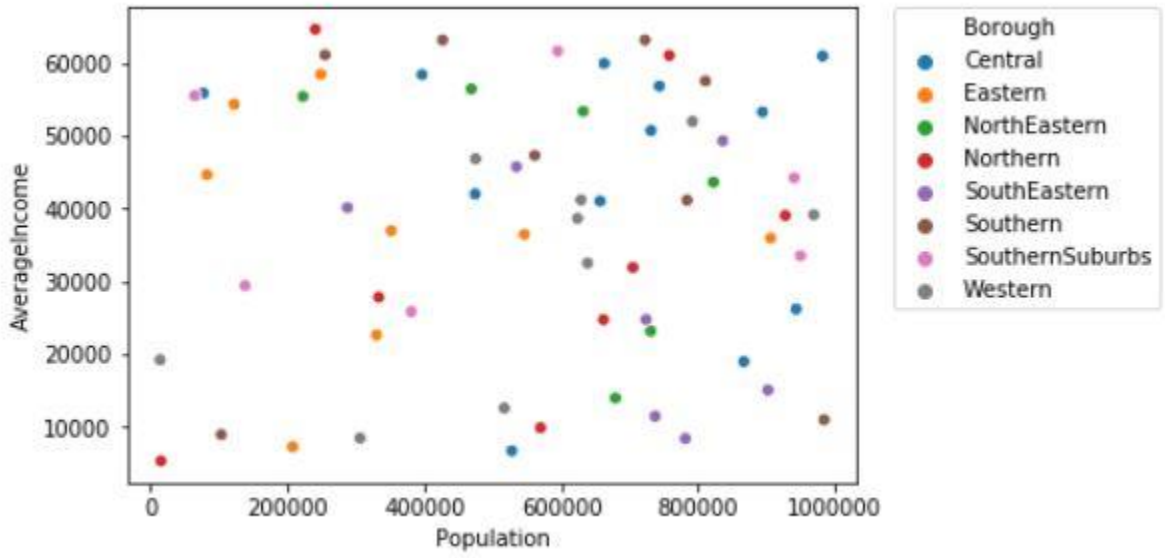
Also while producing graph for number of cluster, I produced a graph to explore all the values for n_clusters and then finding the best by exploring the elbow graph.



Inferential analysis

Most important factors while building the recommender system were population and income. They are the most important factors because they have a nonlinear relationship according to our dataset. It is 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. This can be inferred from the following graph:



Result

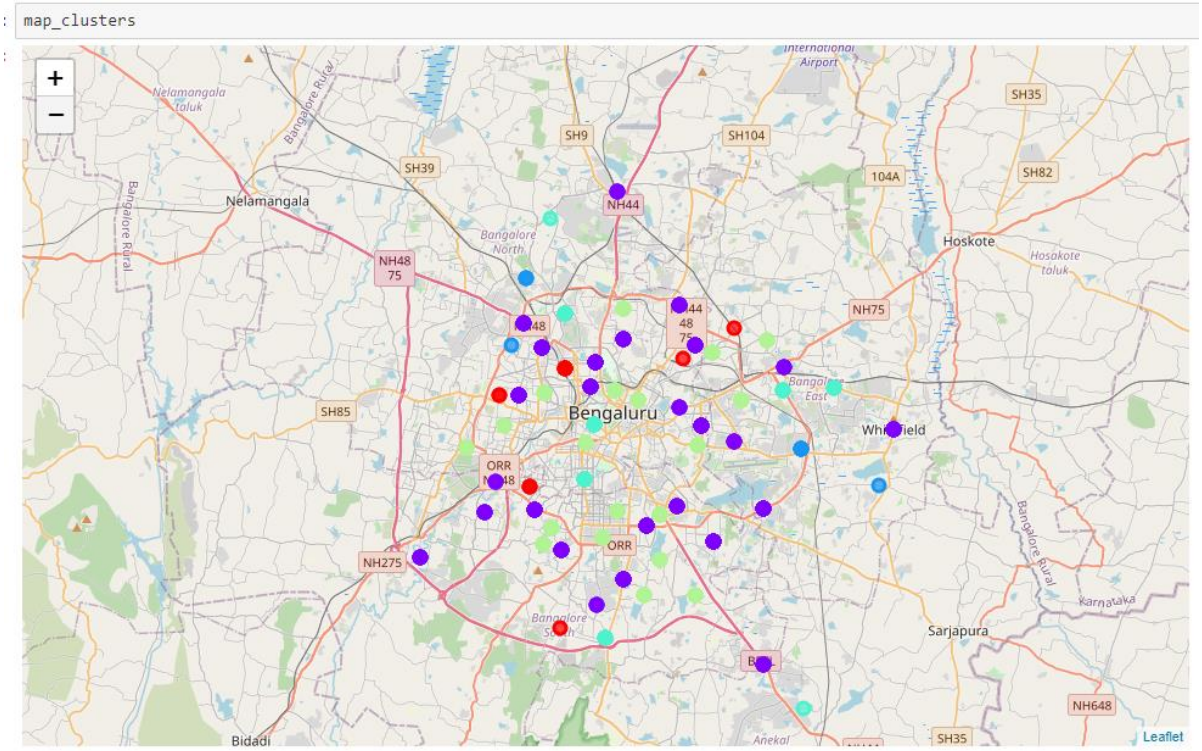
The result of the recommender system is that it produces a list of top restaurants and the most common venue item that the user can enjoy. During the runtime of the model, a simulation was done by taking 'Whitefield' as the neighborhood and then processed through our model so that it could recommend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

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Arekere	Venue Category_Indian Restaurant	Venue Category_Sporting Goods Shop	Venue Category_Pizza Place	[0.32959888840700646]
1	BTM Layout	Venue Category_Indian Restaurant	Venue Category_Ice Cream Shop	Venue Category_Chinese Restaurant	[0.6918117751640322]
2	Banashankari	Venue Category_Café	Venue Category_Indian Restaurant	Venue Category_Miscellaneous Shop	[0.8234029969357849]

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:



Another observation that we can make is that choosing number of clustering could produce very diverse results. Some may be over fitted or some may be under fitted. Hence analysis of number of clusters must be done. Ref elbow_graph in the Methodology section.

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 hunger by providing the best recommendation to fulfil all their needs.