Restaurant Recommender System

IBM Data Science

Project Report

Presented by

Sreejith S.

Table of contents

Introduction

Problem background:

Bangalore is the capital and largest city of the Indian state of Karnataka. With a population of over 15 million (as of January 2016), Bangalore is the third largest city in India and 27th largest city in the world. Bangalore is one of the most ethnically diverse cities in the country, with over 51% of the city's population being migrants from other parts of India.

The diversity of the cuisine available is reflective of the social and economic diversity of Bangalore. Roadside vendors, tea stalls, South Indian, North Indian, Arabic 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. Bangalore can also be called a foodie's paradise because of its vast variety of foods and edibles with a touch of Bangalore's uniqueness and tradition.

Problem description:

Culinary has become part of people’s lives today. Culinary is a part of life that is closely related to daily food consumption. Recently, culinary is not just about food needs, but also about a community lifestyle. People are increasingly like eating in restaurant, so the number of restaurants that are growing even more, especially in the area of Bangalore. With a growing number of restaurants, many websites or applications provide information about restaurants in Bangalore. Some information from websites or applications about restaurants in Bangalore still use filtering methods based on places or types of cuisine. However, it is still possible to display choices that are relevant to the filtering results from the user, but not according with user tastes.

The majority of people or consumers will choose restaurants with high ratings and positive reviews from other people to make choices. Therefore, ratings and reviews are very influential with consumers who have the same tastes, so consumers get the same recommendations with other consumers with the same taste.

So, the questions that must be addressed here, include:

1. What are the varieties of food available in the restaurant?

2. Which is the restaurant nearest by with a good rating?

3. Which restaurants are comparatively cheaper and also providing quality food?

The recommender system is expected to answer these questions, and it must uncover all the perspective of managing recommendations. It is sighted to show:

1. What are the diverse cuisine 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 is basically anyone in the city of Bangalore. People who have the habit of going to a particular type of cuisine will be suggested more restaurants of the similar taste. People who rarely use restaurants would prefer to have the highest rated restaurants nearby them and those could be easily suggested by our recommender system.

Data Requirements and Collection

Data requirements:

To find a solution to the questions and build a recommender model, we need data. In the restaurant recommender system, we would like to know:

1. Geographical coordinates (latitude and longitude) to find exactly where it is located.

2. Population in the neighborhood where the restaurant is located to know if it is usually crowded or not.

3. Average income of neighborhood to know how much would be the price range of food in the restaurant.

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

To access location of a restaurant, its 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.

Population of a neighborhood is very important factor in determining a restaurant's growth and number 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. More customers ensure more rating count and this helps in getting a better idea about the restaurant.

Income of a neighborhood is also an important factor. Spending of money on food and other needs is usually in proportion to the income.

Data collection:

Google maps API was used to fetch the geographical coordinates (latitude and longitude). Initially, the neighbors were scrapped using beautifulSoup4. The table headings and data became the boroughs and the neighborhoods respectively. Bangalore has 8 boroghs and 64 neighborhoods. So, each neighborhood was manually googled to find its corresponding latitude and longitude.

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

Population by neighborhood was available in

https://indikosh.com/dist/655489/bangalore

Income by neighborhood was available in

https://en.wikipedia.org/wiki/List_of_Indian_cities_by_GDP_per_capita

For some of the neighborhoods, population and income data were not available and their values are assumed. Hence, this project is just for demonstrating purposes.

	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

	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

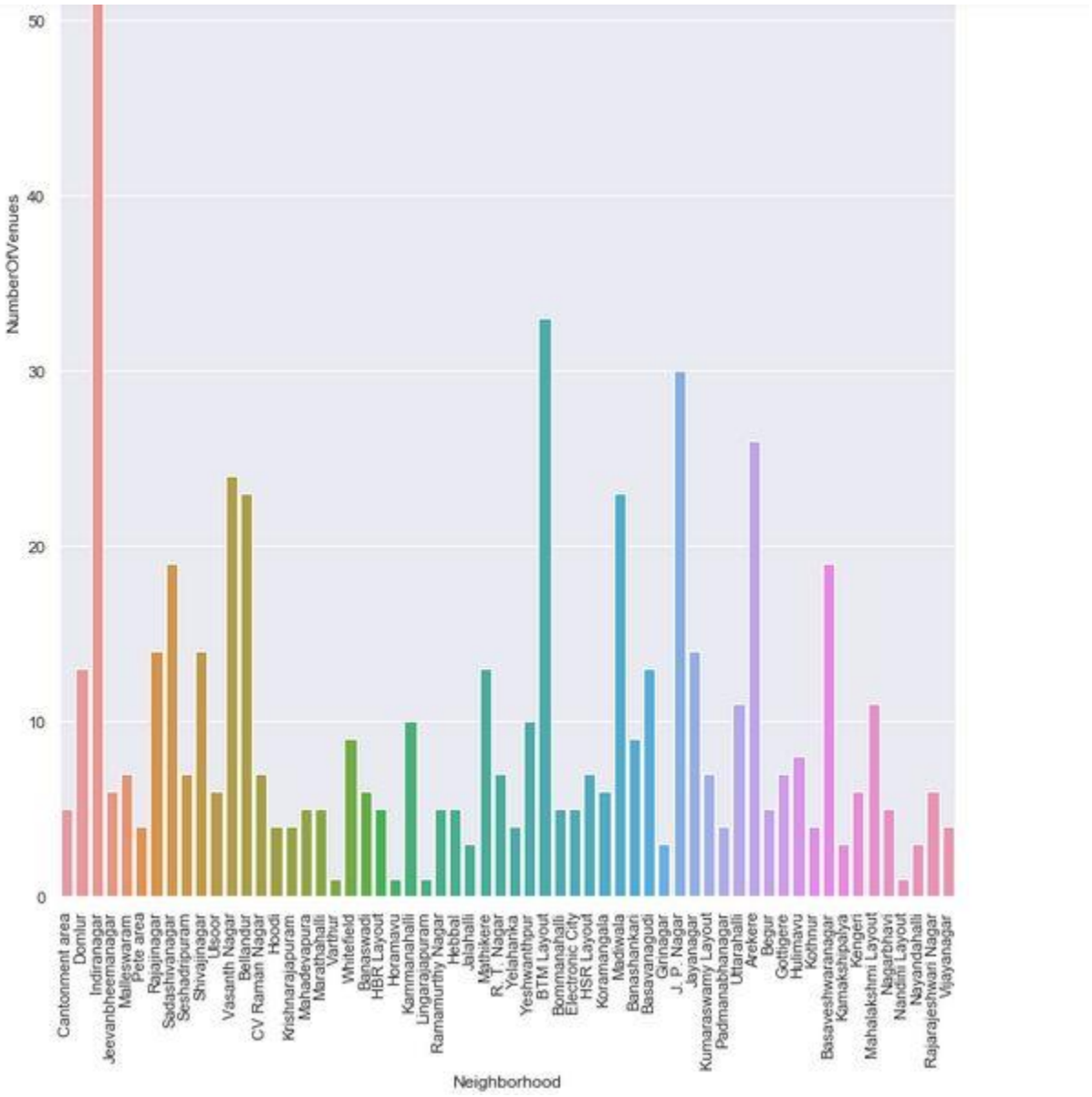
FourSquare API is focused to fetch nearest venue locations so that we can use them to form a cluster. FourSquare API leverages the power of finding nearest venues in a radius (in this project: 500mts) and also the corresponding coordinates, venue location and their names.

	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

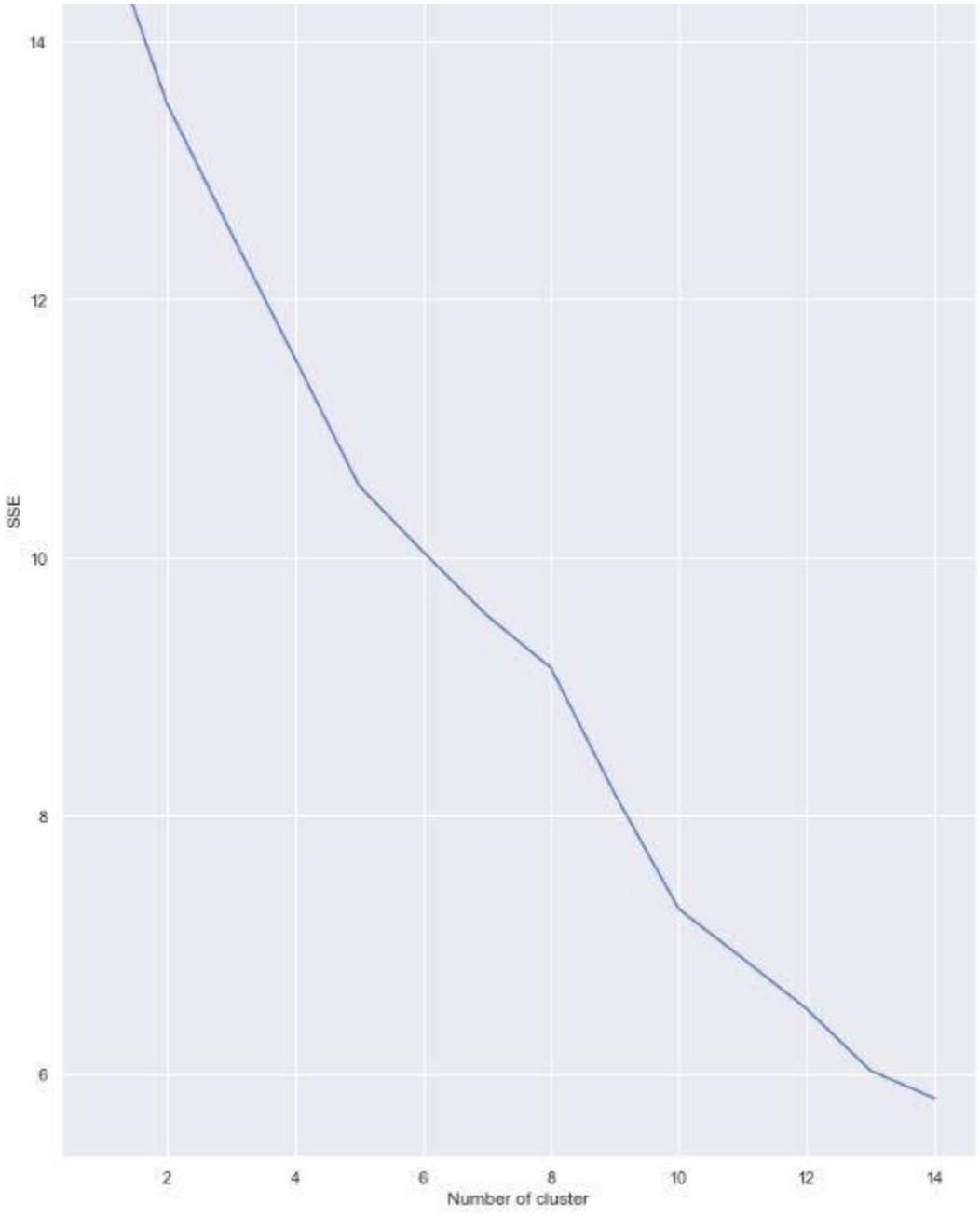
Methodology

Exploratory data 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. Here, we find that Indiranagar has most number of venues while Varthur has the least.



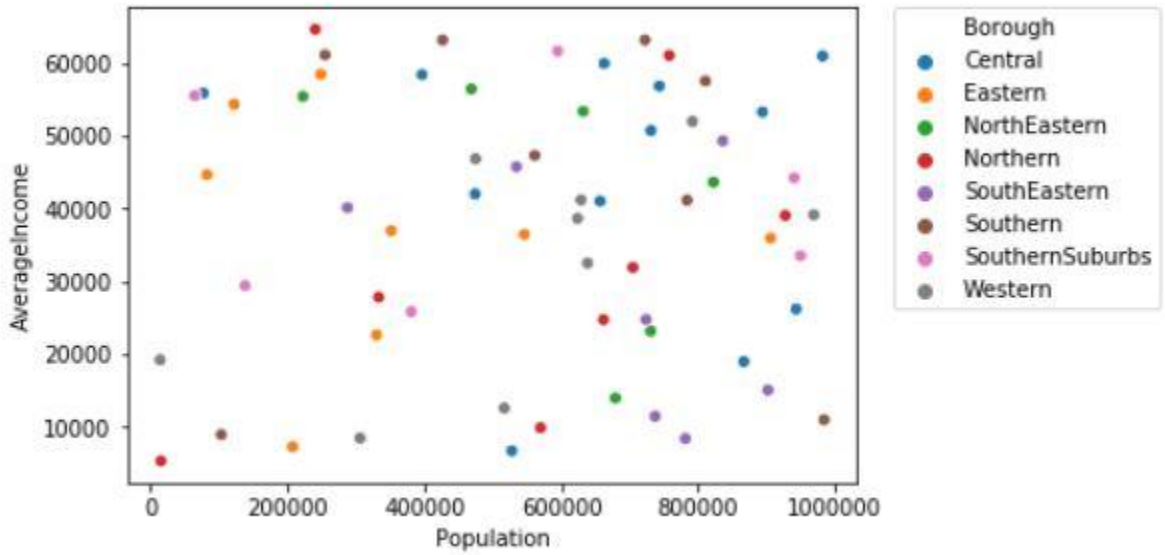
A graph was produced to explore all the values for n number of clusters and we find the optimal value for k to be 5.



Inferential analysis:

The most important factors while building the recommender system were population and income. They have a nonlinear relationship according to our dataset.

As the amount of population increases, it does not necessarily mean that average income of a neighborhood will also increase. It is true in most of the case but not generalizable. It is possible to have fewer number of people and more income and vice versa. This can be inferred from the following graph:



Result

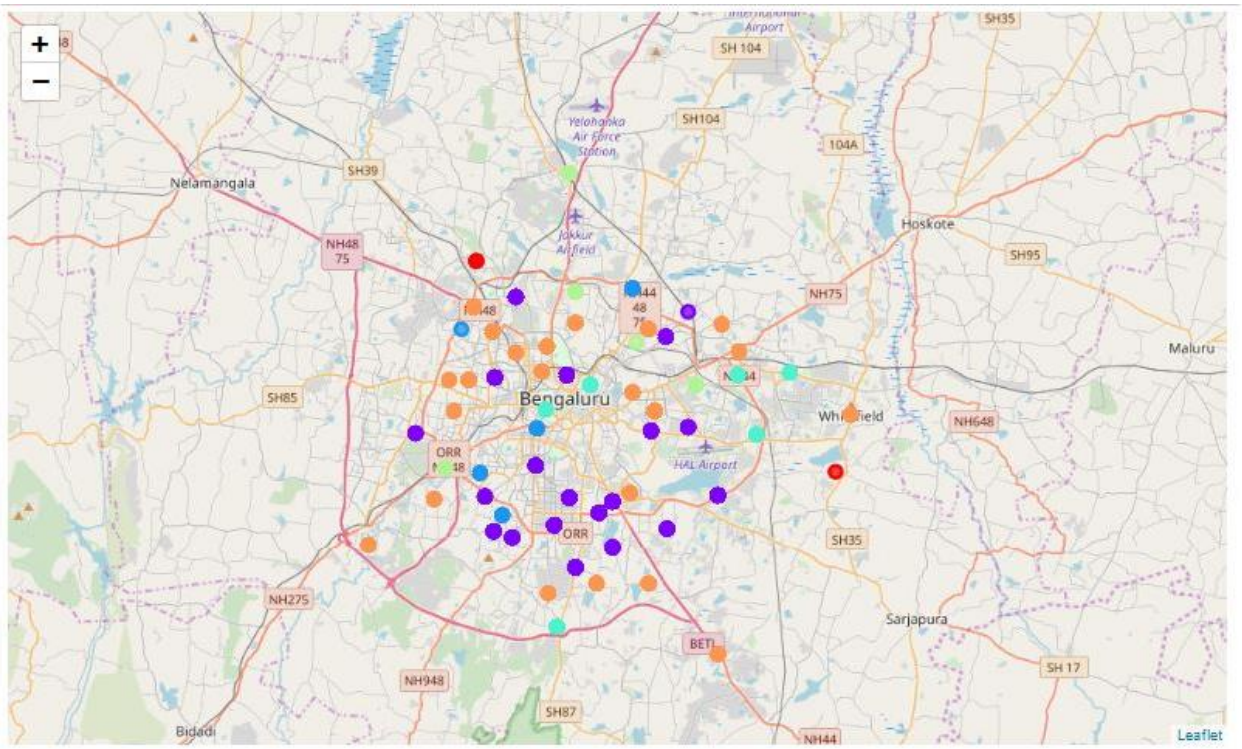
The result of the recommender system is that it produces a list of top restaurants and the most common venue item that the user can enjoy. During the runtime of the model, a simulation was done by taking 'Whitefield' as the neighborhood and then processed through our model so that it could recommend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

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Basaveshwaranagar	Venue Category_Fast Food Restaurant	Venue Category_Indian Restaurant	Venue Category_Ice Cream Shop	[0.6426377807870477]
1	Electronic City	Venue Category_Furniture / Home Store	Venue Category_Outlet Store	Venue Category_Bus Stop	[0.5423513638809381]
2	Indiranagar	Venue Category_Pub	Venue Category_Lounge	Venue Category_Indian Restaurant	[0.5684483844100865]

Discussion

Since there is 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. 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 under fitted. Hence analysis of number of clusters must be done.

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 by providing the best recommendations as per their needs.