Applied Data Science Capstone Project

Restaurant Recommendation System

1. Introduction

Problem background:

Bangalore (now Bengaluru) is the capital 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.
The diversity of the cuisine available is reflective of the social and economic diversity of Bangalore. Roadside vendors, tea stalls, South Indian, North Indian, Chinese and Western fast food are very popular throughout the city. Udupi chain of restaurants are very popular and serve Vegetarian cuisine. The Chinese and Thai food served in most of the restaurants are 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:

Since we all are travelers and keep changing places very frequently it becomes very hectic and we continuously get exposed to different types of environment about which we do not have much knowledge. In such situations, food can be considered as an important factor based on which we rate our trips and also recommend it to other people. Food has the power to attract people around to world! In such scenarios, we need to find the right place with the appropriate quality and at reasonable cost. So there are few questions that must be addressed, such as:

1. How many types of cuisines are available in the restaurant?
2. Restaurants near me with a good rating?
3. How many ‘similar’ restaurants are available in the vicinity?
4. Does the ‘similar’ restaurants cost more? If so, what is the differentiating factor?

To address the above questions, ABC Company’s official 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 rankings.

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

1. What types of restaurant 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 our preferences?

Target Audience:

Target audience for this project is not limited to a person who is a traveler but everyone including the residents of the city. People decide to look for a similar kind of restaurant most of the time because they have developed a liking for a particular cuisine. People who seldom use restaurants would prefer to know about the highest rated restaurants near them. All this could be easily handled by our recommender system. So in short the target for this project is everyone who is an explorer.

Success rate:

With restaurants continuously evolving, new food categories emerge, hybrid (mix of cuisines) food start to become more popular, we need a system that could help us access vast number of food varieties. It is impossible for any person to ask each and every one about their visit to a particular place and also not everyone remembers everything. On the other hand, we can exploit the power of computers as they are good at remembering things, and with Machine learning to its peak, its high time we use technology for our personal guidance to help us based on our likes and dislikes.

2. Data :

Data requirements:

To find a solution to the questions and build a recommender model, we need data and lots of data. Data can answer questions that are unimaginable and unanswerable by humans because we do not have the capacity to analyze such large dataset and produce analytics to find a solution.

Let's consider a scenario:
Suppose I want to find a restaurant, then logically, I need 3 things:

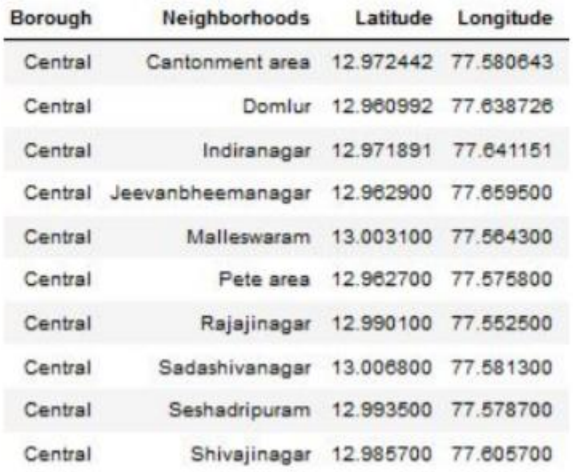
1. Its geographical co-ordinates (latitude and longitude) to find out where exactly it is located.
2. Population of the neighborhood where the restaurant is located.
3. Average income of neighborhood to find out the economical value.

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 pin-point its coordinates and create a map displaying all the restaurants with its respective labels.
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 into a restaurant and vice-versa. Also if more people visit, better is the rating of the restaurant as it is accessed by different people with different tastes. Hence, it becomes a very important factor.
3. Income of a neighborhood is also a very important factor. Income is directly proportional to spending capacity of a neighborhood. If people in a neighborhood earn more than an average income, then it is highly likely that they will spend more however it is not always true. So, it can be assumed that a restaurant assessment is proportional to income of a neighborhood.

Data collection:

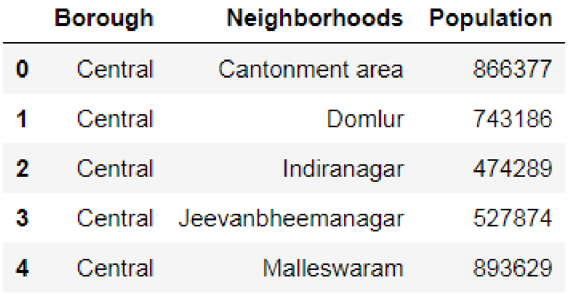
1. Collection of geographical coordinates is not very difficult. Initially I scrapped list of neighborhoods using beautifulSoup4 package from the following link : [wikipedia](https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore). Bangalore has 8 boroughs and 64 neighborhoods. So the co-ordinates for each neighborhood was manually searched on Google. After doing the above the following dataframe was compiled:



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

Table 1: Co-ordinates of Borough & Neighborhoods

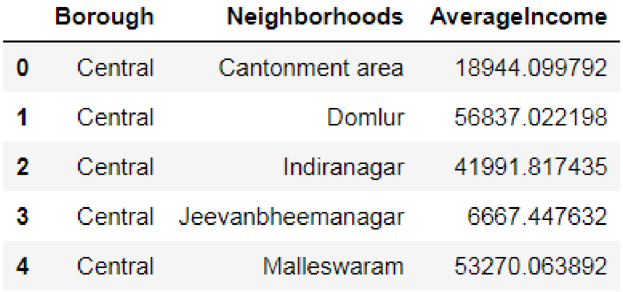
2. Population of neighborhood was found out using the following link: (https://indikosh.com/dist/655489/bangalore). Since this is a demonstrating project the data may be inaccurate, the main aim of this project is to get a working model. The data frame for Bangalore neighborhood population looks like:



	Borough	Neighborhoods	Population
0	Central	Cantonment area	866377
1	Central	Domlur	743186
2	Central	Indiranagar	474289
3	Central	Jeevanbheemanagar	527874
4	Central	Malleswaram	893629

Table 2: Neighborhood Population

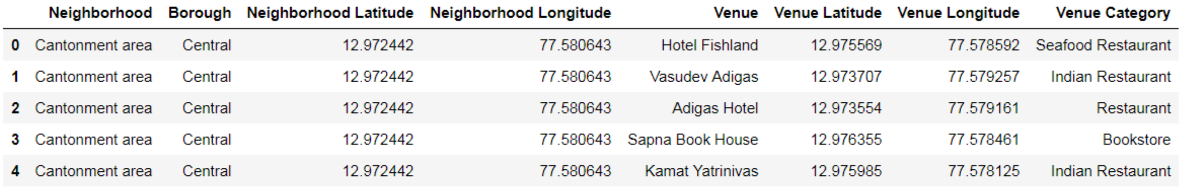
3. Income of the neighborhood was found out using the link: (https://en.wikipedia.org/wiki/List_of_Indian_cities_by_GDP_per_capita). Neighborhood Income is assumed and may be inaccurate. The data frame for Bangalore neighborhood population looks like:



	Borough	Neighborhoods	AverageIncome
0	Central	Cantonment area	18944.099792
1	Central	Domlur	56837.022198
2	Central	Indiranagar	41991.817435
3	Central	Jeevanbheemanagar	6667.447632
4	Central	Malleswaram	53270.063892

Table 3: Neighborhood Income

4. Foursquare API was used 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 case: 500m) and also corresponding coordinates, venue location and names. After calling, the following data frame is created:



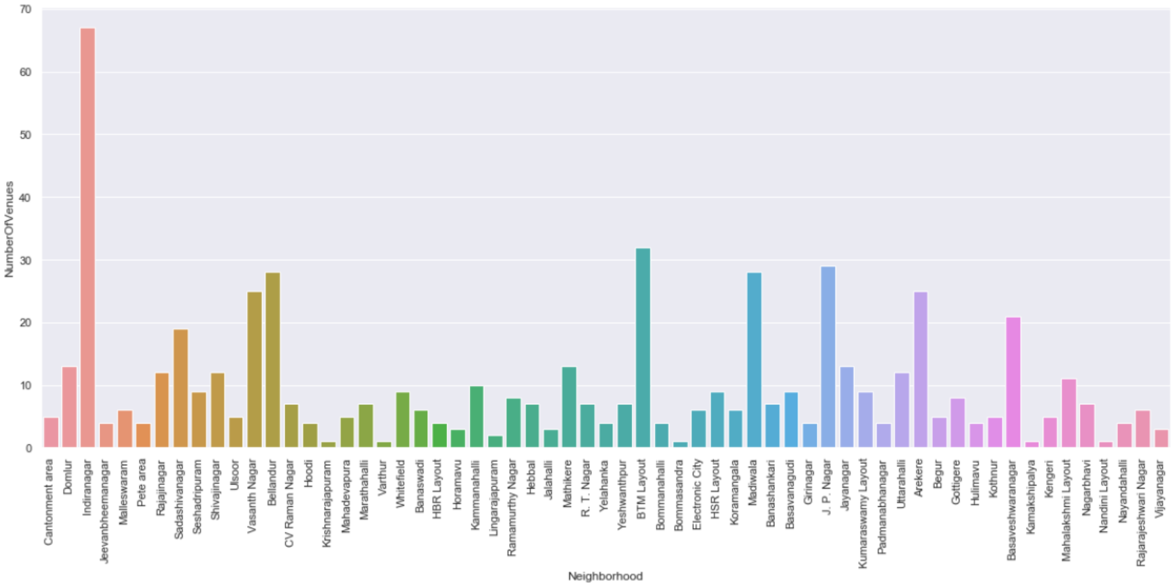
	Neighborhood	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cantonment area	Central	12.972442	77.580643	Hotel Fishland	12.975569	77.578592	Seafood Restaurant
1	Cantonment area	Central	12.972442	77.580643	Vasudev Adigas	12.973707	77.579257	Indian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

Table 4: Foursquare API data

3.Methodology:

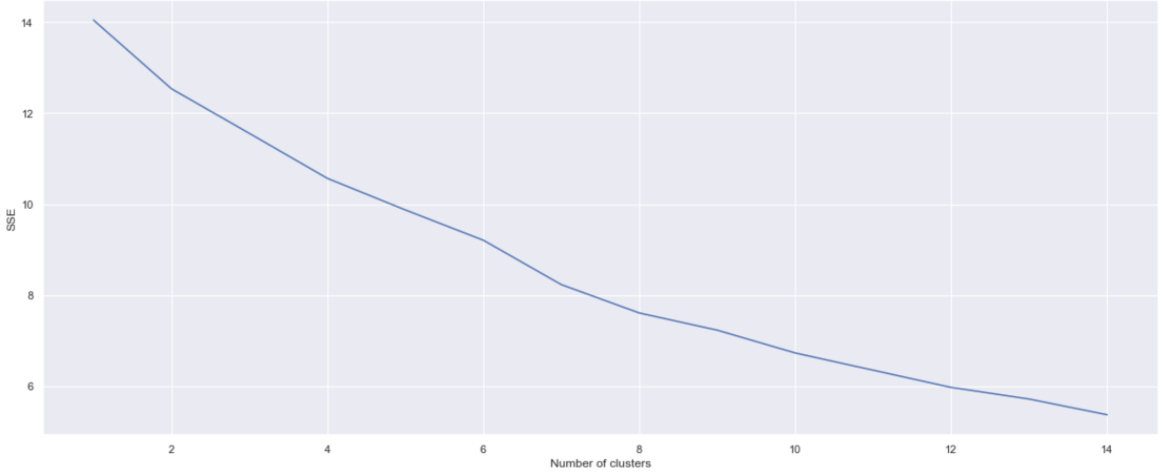
Exploratory analysis:

The data was scrapped from different sources and then combined. 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 us initial insights and may help us get partial idea of the answers that we are looking to find out from the data. While exploring the dataset, I found out that Indiranagar has most number of venues while Varthur has the least. The analysis can be interpreted from the following graph:



Graph 1: Exploratory analysis of Neighborhoods

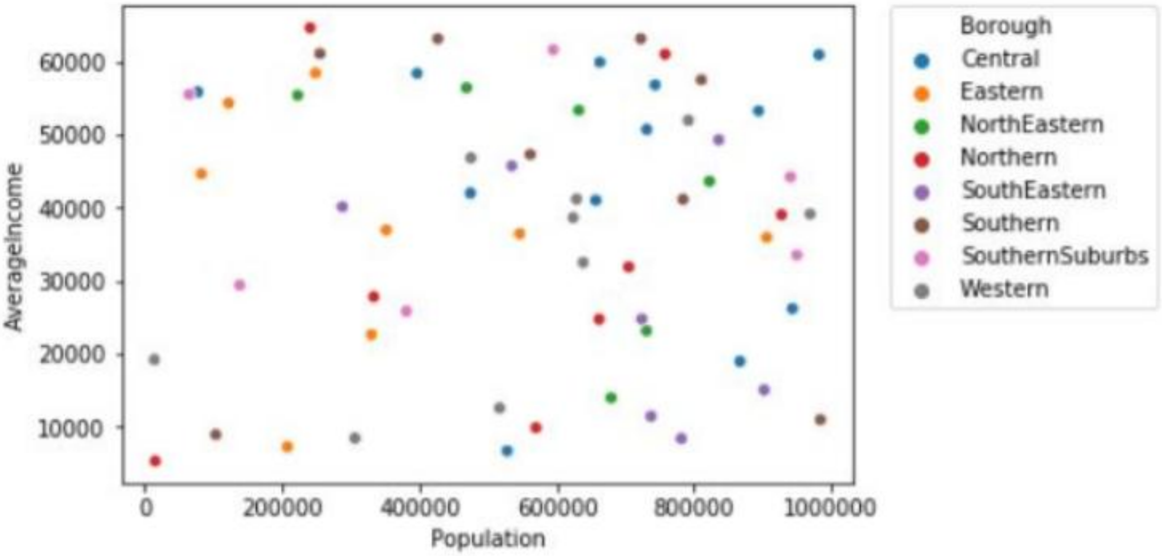
Also while producing graph for number of cluster, a graph was produced to explore all the values for ‘n_clusters’ and to find out the optimal value of ‘k’ (in this case k=4) by exploring the elbow graph as shown below:



Graph 2: Elbow graph for optimal ‘k’

Inferential analysis:

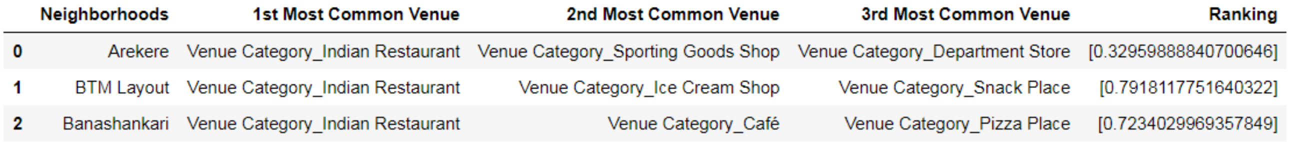
Most important factors while building the recommender system were population and income. They were the most import factor because they had a nonlinear relationship according to our dataset. So it was required to make some inferential analysis to understand the nonlinear relationship. As the amount of population increases, it does not necessarily mean that average income of a neighborhood will also increase. It is true for most of the cases but many cases do not 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:



Graph 3: Comparison of Average Income and Population in different boroughs

4. 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 ‘HSR Layout’ as the neighborhood and then further processed through our model so that it could recommend neighborhoods with similar characters as that of ‘HSR Layout’. The following image shows the result:

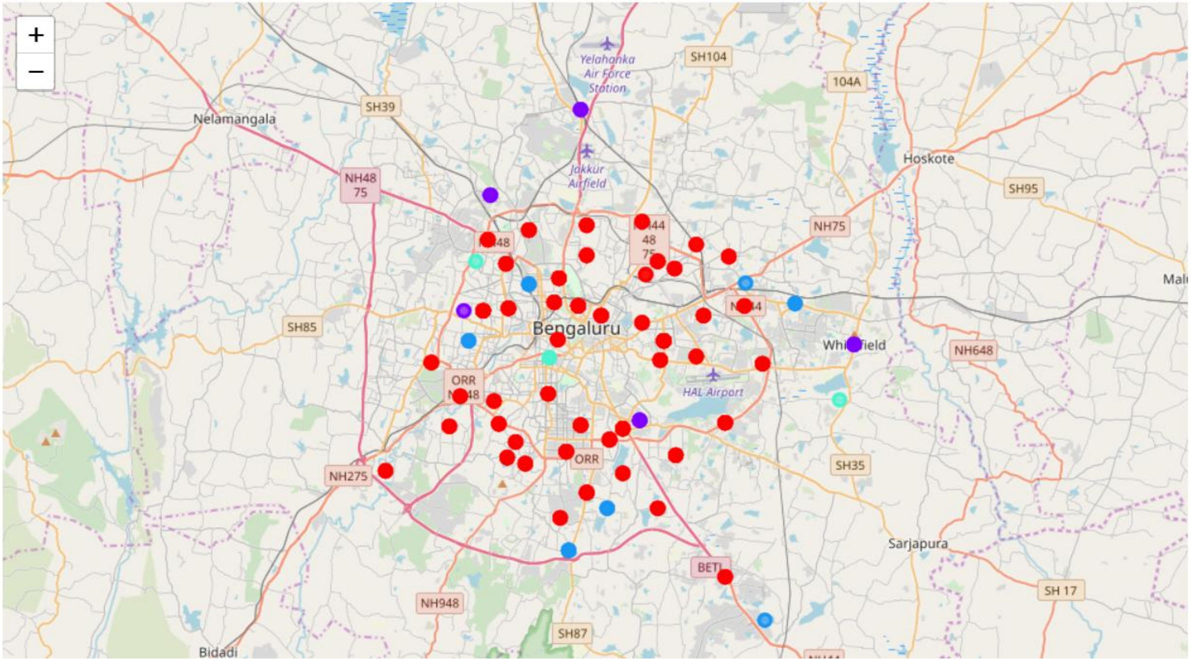


	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Arekere	Venue Category_Indian Restaurant	Venue Category_Sporting Goods Shop	Venue Category_Department Store	[0.32959888840700646]
1	BTM Layout	Venue Category_Indian Restaurant	Venue Category_Ice Cream Shop	Venue Category_Snack Place	[0.7918117751640322]
2	Banashankari	Venue Category_Indian Restaurant	Venue Category_Café	Venue Category_Pizza Place	[0.7234029969357849]

Table 5: Result

5. 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 map shows the various clusters:



Map 1: Map showing various clusters

Another observation that we can make is that varying the number of clusters (‘k’) 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 to find out the value of optimal ‘k’. Refer to Graph 2 in Methodology Section.

6. Conclusion :

The recommender system we modeled considers various 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 recommendation.