IBM Watson IoT

IBM

Capstone Project

Advanced Data Science

Powered By

coursera

IBM Digital Credential

A machine learning model that recommends Ideal Business to be setup in Bengaluru based on volume of population, and their average income and various other category of business or shops that are already in place, also by identifying more common venues by comparing with candidate /model neighborhood. This is to demonstrate part of capstone project to IBM through Coursera.

Introduction:

Problem background:
Bangalore is not just capital for the state of Karnataka, but also it is the IT Capital of India with diverse population from all over India, such that it is one of the Cosmopolitan City in India. With a population of over 15 million (as of January 2016), Bangalore is the third largest city in India.

As the Govt. of India and Govt. of Karnataka encourages it's citizens with many schemes such as Make In India or to start a small-scale Enterprises, the young entrepreneurs may stuck with basic questions of what business to start, and in which location to start to keep their avenues optimal and boost their ability to sustain their business and to take further level in a reasonable timeframe.

Problem description:
If one choses Bangalore as their destination to start a business, it will trigger few basic questions to be answered to further start their pre-work upon given set of answers and recommendations. Such as:

1. Which business should I pick, and why?
2. Which location best suites?
3. Is there enough population to maximize the opportunities?
4. From the recommended locations, what percentage of population falls into the spending range? Such as average income of the location population., etc.

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 rankings compared to the previously visited by me.

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 venues are present / business in the given area?
2. Find similar nature of business present in other neighboring locations based on a preference?
3. How do different business rank with respect to the preferences?

Target Audience:
Target audience for this project are those who interested in starting a new business of their interest, specific to Bangalore location, and looking for a recommender system that help understanding the current dynamics of the business that they are about to start.

Success rate:
Success rate would be unpredictable, when there is no proper evaluation of various parameters such as location, size of population, accessibility to the place where business get started, average income of the public in that specific area, or those who commute around the place.

This recommender system is expected to fill such gaps, and scientifically explore, evaluate and recommend optimal place, and nature of business that suites to the place based on the above mentioned parameters, and rank them how such business gets performed in similar neighborhoods.

1. 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 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 solution.

Let's consider the base scenario:

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

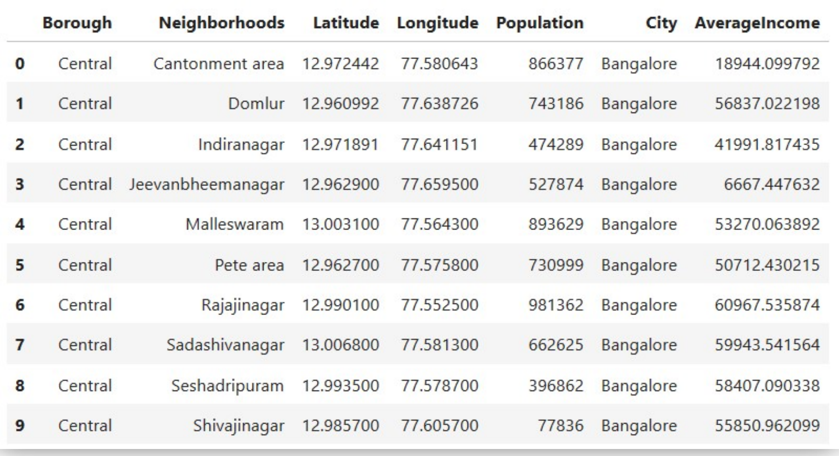
1. Its geographical coordinates (latitude and longitude) to find out where exactly it is located.
2. Population of the neighborhood where the restaurant is located.
3. Average income of neighborhood to know how much the restaurant would worth.

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 the 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.

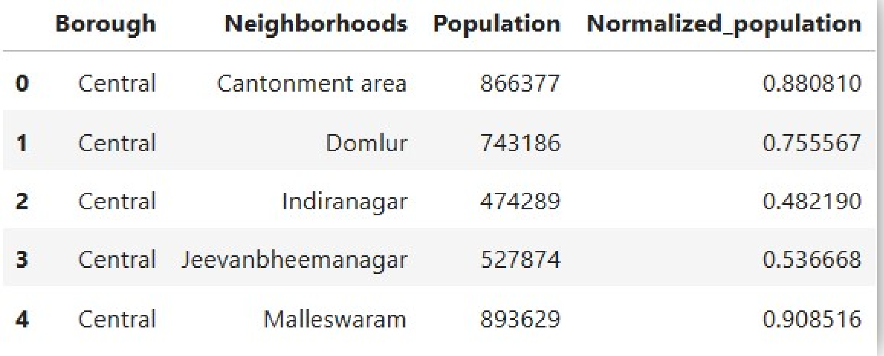
Data collection:

1. Collecting geographical coordinates is not difficult job, as I could fetch them directly from foursquare API by giving the location / neighborhood details to fetch latitude and longitude.
2. Initially I scrapped list of neighbors’ using beautifulSoup4 from [wikipedia] (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore). The table headings becoming the boroughs and data becoming the neighborhoods. Bangalore has 8 boroughs and 64 neighborhoods.



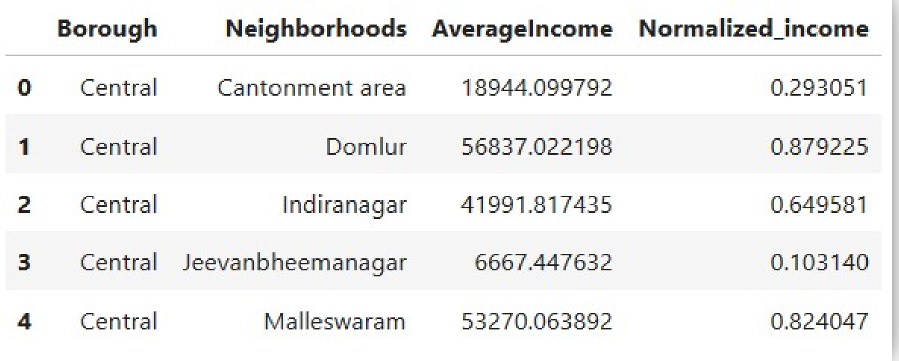
	Borough	Neighborhoods	Latitude	Longitude	Population	City	AverageIncome
0	Central	Cantonment area	12.972442	77.580643	866377	Bangalore	18944.099792
1	Central	Domlur	12.960992	77.638726	743186	Bangalore	56837.022198
2	Central	Indiranagar	12.971891	77.641151	474289	Bangalore	41991.817435
3	Central	Jeevanbheemanagar	12.962900	77.659500	527874	Bangalore	6667.447632
4	Central	Malleswaram	13.003100	77.564300	893629	Bangalore	53270.063892
5	Central	Pete area	12.962700	77.575800	730999	Bangalore	50712.430215
6	Central	Rajajinagar	12.990100	77.552500	981362	Bangalore	60967.535874
7	Central	Sadashivanagar	13.006800	77.581300	662625	Bangalore	59943.541564
8	Central	Seshadripuram	12.993500	77.578700	396862	Bangalore	58407.090338
9	Central	Shivajinagar	12.985700	77.605700	77836	Bangalore	55850.962099

3. Population by neighborhood is again easy to find out given that it's readily available. But in case of Bangalore, it is again not the case. I was able to find population data for few cities. [Here is the link] (https://indikosh.com/dist/655489/bangalore). Rest other neighborhood population is assumed and may be inaccurate but since this is a demonstrating project, the main idea to get the working model. The data frame for Bangalore neighborhood population looks like:



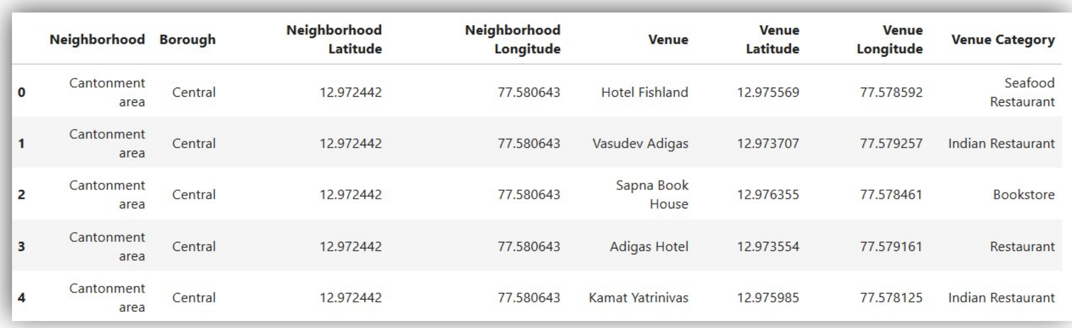
	Borough	Neighborhoods	Population	Normalized_population
0	Central	Cantonment area	866377	0.880810
1	Central	Domlur	743186	0.755567
2	Central	Indiranagar	474289	0.482190
3	Central	Jeevanbheemanagar	527874	0.536668
4	Central	Malleswaram	893629	0.908516

4. Income by neighborhood is again easy to find out given that it's readily available. But in case of Bangalore, it is again not the case. I was able to find Income data for main city. [Here is the link] (https://en.wikipedia.org/wiki/List_of_Indian_cities_by_GDP_per_capita). Neighborhood Income is assumed and may be inaccurate but since this is a demonstrating project, the main idea to get the working model. The data frame for Bangalore neighborhood population looks like:



	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

5. Foursquare API: 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:

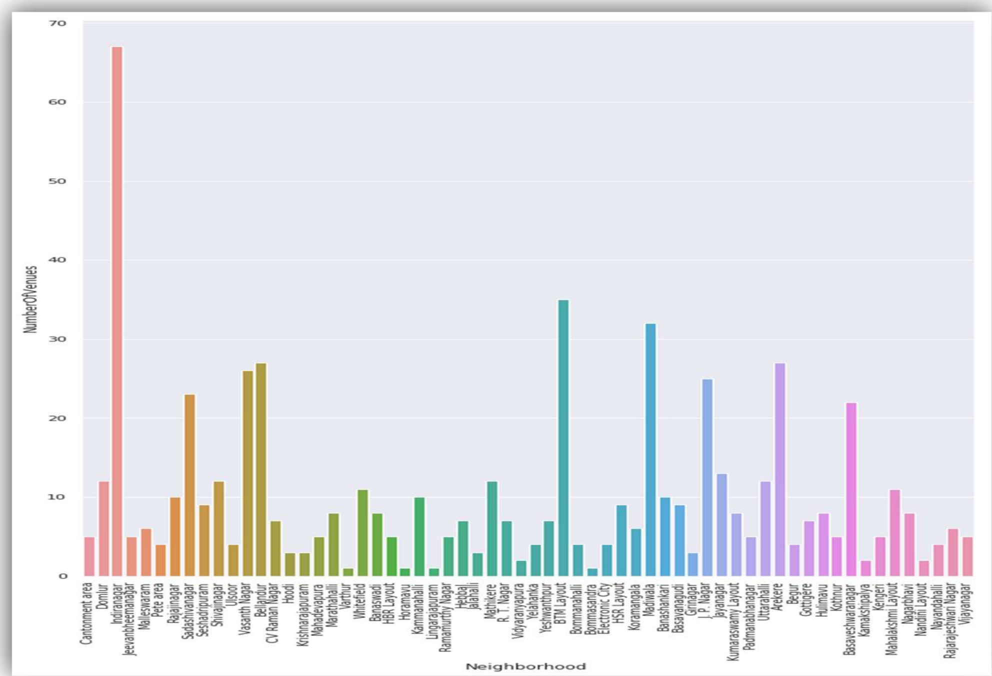


	Neighborhood	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cantonment area	Central	12.972442	77.580643	Hotel Fishland	12.975569	77.578592	Seafood Restaurant
1	Cantonment area	Central	12.972442	77.580643	Vasudev Adigas	12.973707	77.579257	Indian Restaurant
2	Cantonment area	Central	12.972442	77.580643	Sapna Book House	12.976355	77.578461	Bookstore
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

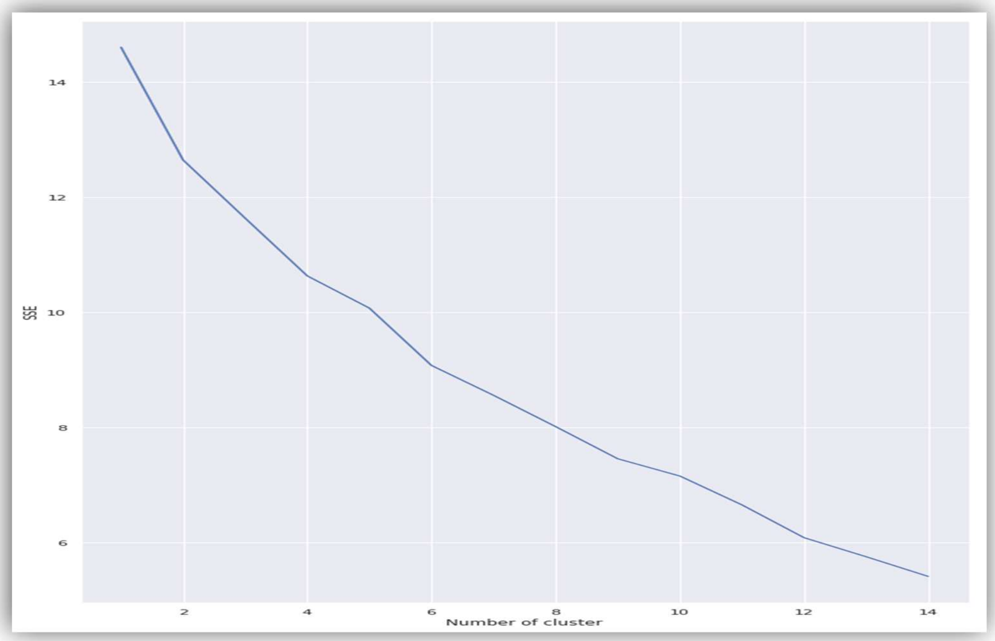
2. Methodology :

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 that Indira Nagar has most number of venues while Varthur has the least.



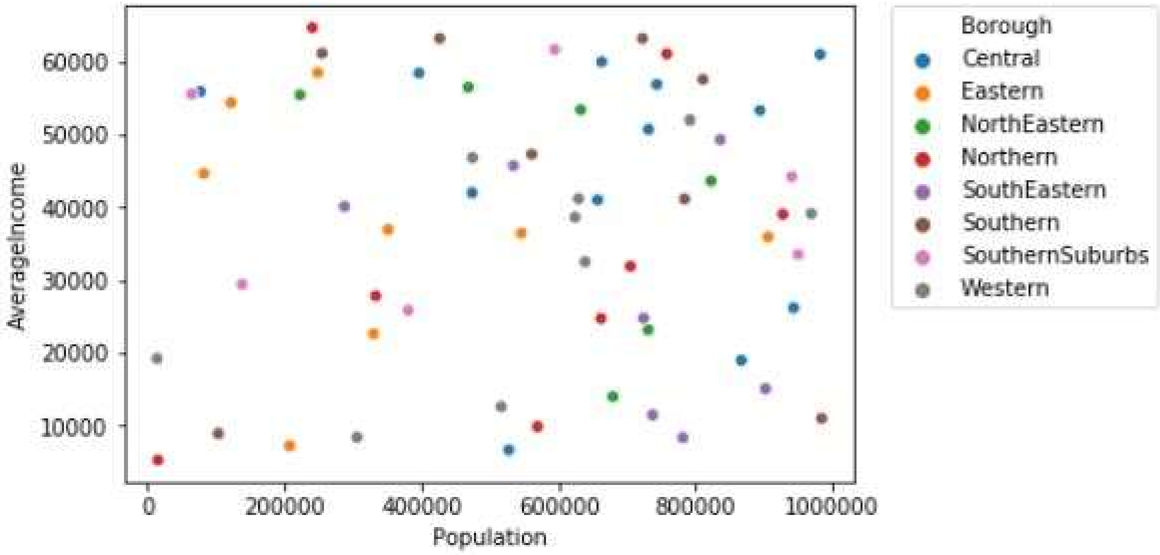
Also while producing graph for number of cluster, I produced a graph to explore all the values for n_clusters and then finding the best by exploring the elbow graph shown below.



Inferential analysis:

Most important parameters for building a recommender system of this nature are population and income. They are the most import factor as they have a nonlinear relationship as notice from the dataset.

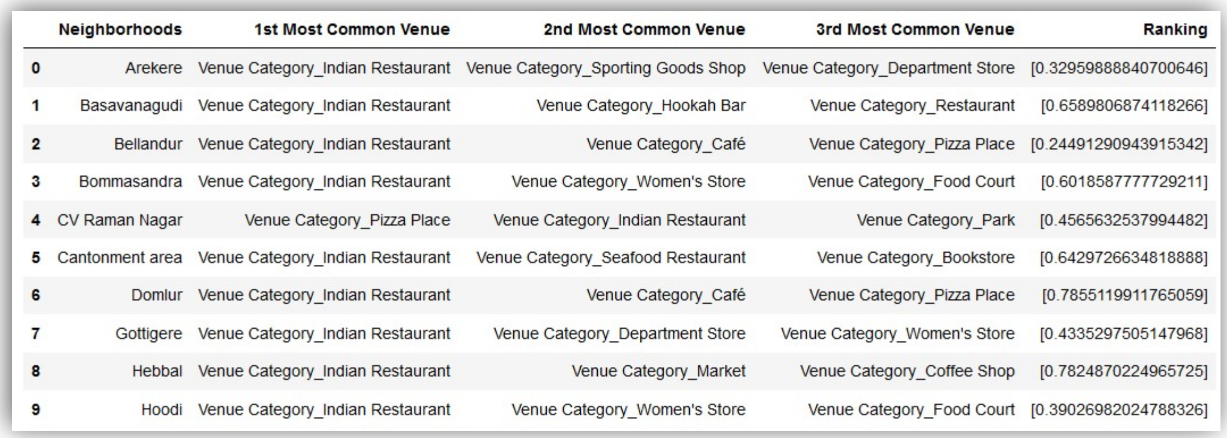
Ought to make some inferential analysis to understand this nonlinear relationship. As the amount of population increases, it doesn't necessarily mean the average income of a neighborhood also increases. It might be true in most cases, at the same time many cases differ to follow such trend. Similarly, a neighborhood with less number of people 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:



3. Result :

The result of the recommender system produces a list of top business opportunities by location, and their ranking with the most common venue items that the user can look at. During the runtime of the model, a simulation was done by taking 'Hebbal' as the neighborhood and then processed through our model so that it could recommend neighborhoods with similar characteristics as that of 'Hebbal'.

The following image shows the result:

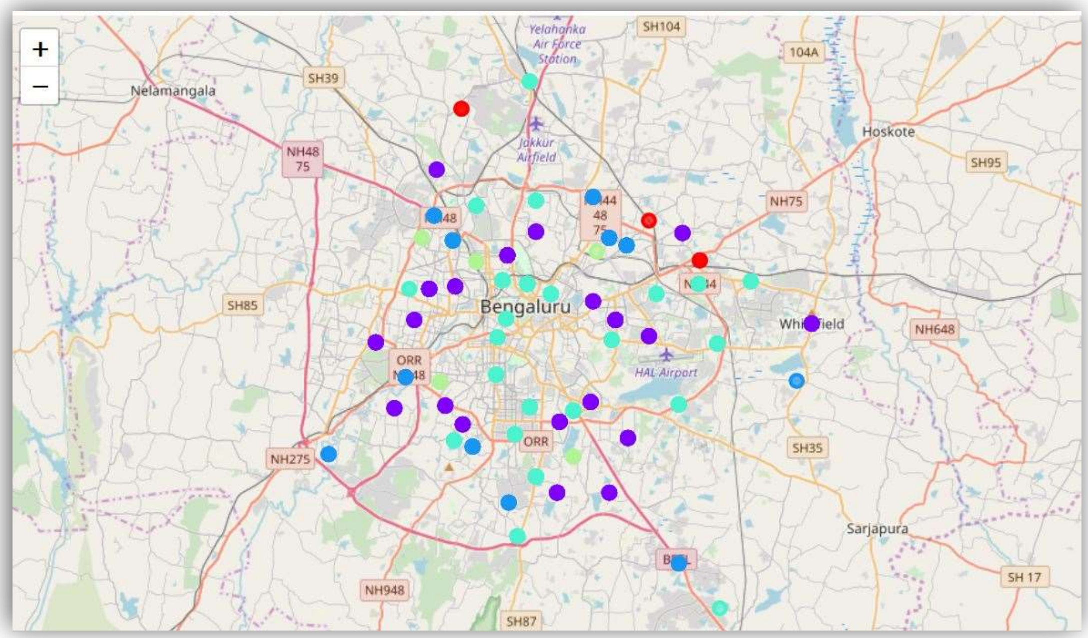


	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Arekere	Venue Category_Indian Restaurant	Venue Category_Sporting Goods Shop	Venue Category_Department Store	[0.32959888840700646]
1	Basavanagudi	Venue Category_Indian Restaurant	Venue Category_Hookah Bar	Venue Category_Restaurant	[0.6589806874118266]
2	Bellandur	Venue Category_Indian Restaurant	Venue Category_Café	Venue Category_Pizza Place	[0.24491290943915342]
3	Bommasandra	Venue Category_Indian Restaurant	Venue Category_Women's Store	Venue Category_Food Court	[0.6018587777729211]
4	CV Raman Nagar	Venue Category_Pizza Place	Venue Category_Indian Restaurant	Venue Category_Park	[0.4565632537994482]
5	Cantonment area	Venue Category_Indian Restaurant	Venue Category_Seafood Restaurant	Venue Category_Bookstore	[0.6429726634818888]
6	Domlur	Venue Category_Indian Restaurant	Venue Category_Café	Venue Category_Pizza Place	[0.7855119911765059]
7	Gottigere	Venue Category_Indian Restaurant	Venue Category_Department Store	Venue Category_Women's Store	[0.4335297505147968]
8	Hebbal	Venue Category_Indian Restaurant	Venue Category_Market	Venue Category_Coffee Shop	[0.7824870224965725]
9	Hoodi	Venue Category_Indian Restaurant	Venue Category_Women's Store	Venue Category_Food Court	[0.39026982024788326]

4. 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 of similar neighborhoods:

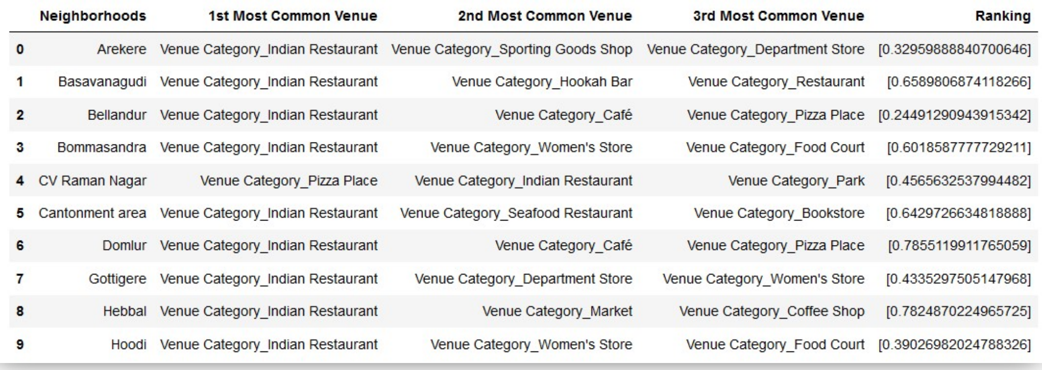


Another observation is that choosing number of clusters could produce very diverse results. Some may be overfit, and some may underfit. Hence analysis about the number of clusters must be done.

Ref elbow_graph in the Methodology section.

5. 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.



	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	Basavanagudi	Venue Category_Indian Restaurant	Venue Category_Hookah Bar	Venue Category_Restaurant	[0.6589806874118266]
2	Bellandur	Venue Category_Indian Restaurant	Venue Category_Café	Venue Category_Pizza Place	[0.24491290943915342]
3	Bommasandra	Venue Category_Indian Restaurant	Venue Category_Women's Store	Venue Category_Food Court	[0.6018587777729211]
4	CV Raman Nagar	Venue Category_Pizza Place	Venue Category_Indian Restaurant	Venue Category_Park	[0.4565632537994482]
5	Cantonment area	Venue Category_Indian Restaurant	Venue Category_Seafood Restaurant	Venue Category_Bookstore	[0.6429726634818888]
6	Domlur	Venue Category_Indian Restaurant	Venue Category_Café	Venue Category_Pizza Place	[0.7855119911765059]
7	Gottigere	Venue Category_Indian Restaurant	Venue Category_Department Store	Venue Category_Women's Store	[0.4335297505147968]
8	Hebbal	Venue Category_Indian Restaurant	Venue Category_Market	Venue Category_Coffee Shop	[0.7824870224965725]
9	Hoodi	Venue Category_Indian Restaurant	Venue Category_Women's Store	Venue Category_Food Court	[0.39026982024788326]