|  |
| --- |
|  |
|  |

A machine learning model system that recommend a Business to setup in Bengaluru based on population and their respective average income, developed to demonstrate part of a capstone project to IBM through coursera.

**Introduction:**

Bangalore is the capital for Karnataka state. 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 young entrepreneurs with many schemes such as Make In India, 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 any young entrepreneurs’ choses Bangalore based locations e.g. like Kammanahali or Whitefiled as their destination to start a business, it will trigger few pre-work questions to be answered, 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, XYZ 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.

Precise expectations from this recommender system is to get answer for the questions. 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.

**Data:**

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.

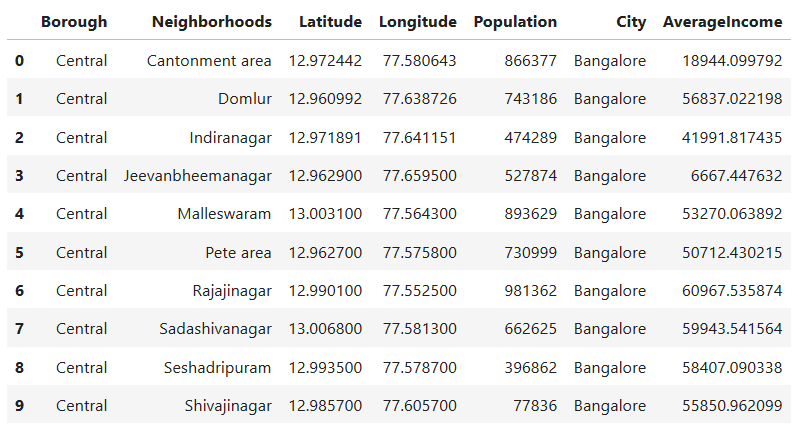
2. Population of a neighborhood is very important factor in determining a restaurant's growth and number of customers who turn up to eat. Logically, the more the population of a neighborhood, the more people will be interested to walk openly into a restaurant and less the population, a smaller number 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.

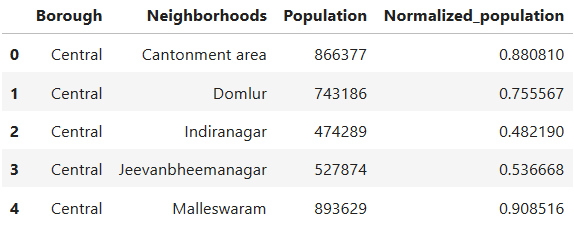
**Data collection:**

1. Collecting geographical coordinates is not difficult but after googling for more than 2 days, it was not available on open source data websites such as Wikipedia, India gov website, census report websites etc. So, I decided to use Google maps API to fetch latitude and longitude but google API has limited number of calls that I could make with my free account. So, it would take around 15 - 20 days to fetch location of all the neighborhoods in Bangalore.

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. So i manually googled each neighborhood to find its corresponding latitude and longitude. After doing so, I produced the following data frame.



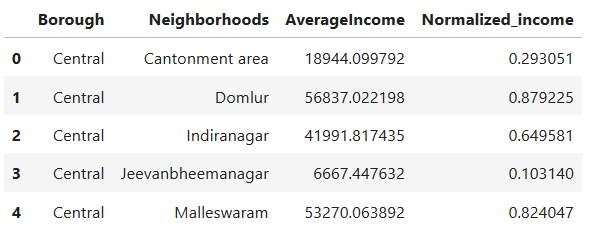
2. Population by neighborhood is again easy to find out given that it’s readily available. But in case of Bangalore, it is again not the case. i was able to find population data for few cities. [Here is the link](https://indikosh.com/dist/655489/bangalore). Rest other neighborhood population is assumed 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:



1. 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]

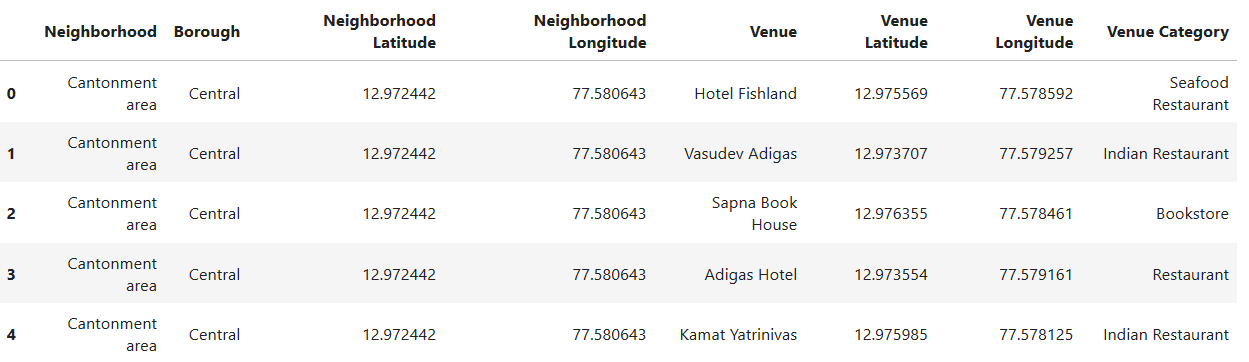
(<https://en.wikipedia.org/wiki/List_of_Indian_cities_by_GDP_per_capita>).

1. 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:



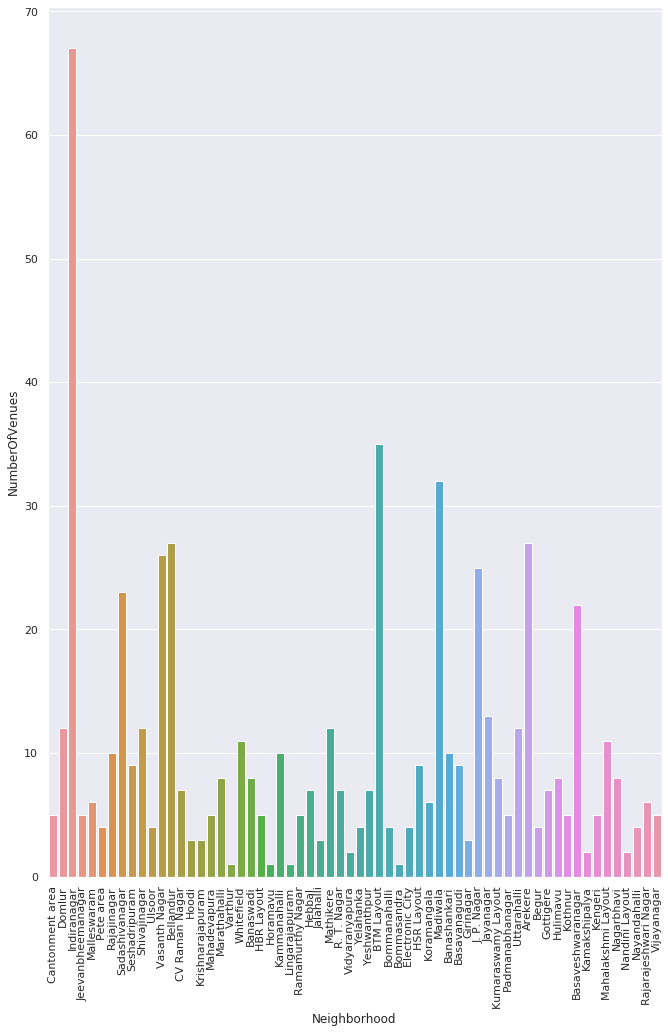
**Foursquare API:**

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:

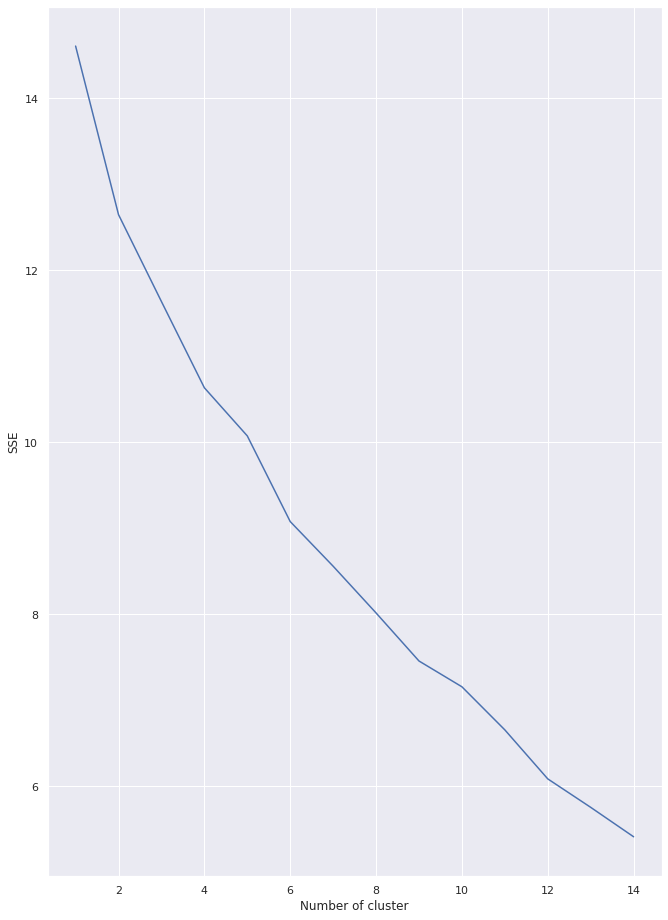


**Methodology: - Exploratory analysis:**

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 Indiranagar has most number of venues while Varthur has the least.



Also, while producing graph for number of clusters, 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 a smaller number of people may not necessarily have less average income. It is possible to have a smaller number of people and more income and vice versa. This can be inferred from the following graph:



**Result for Whitefield:**

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:



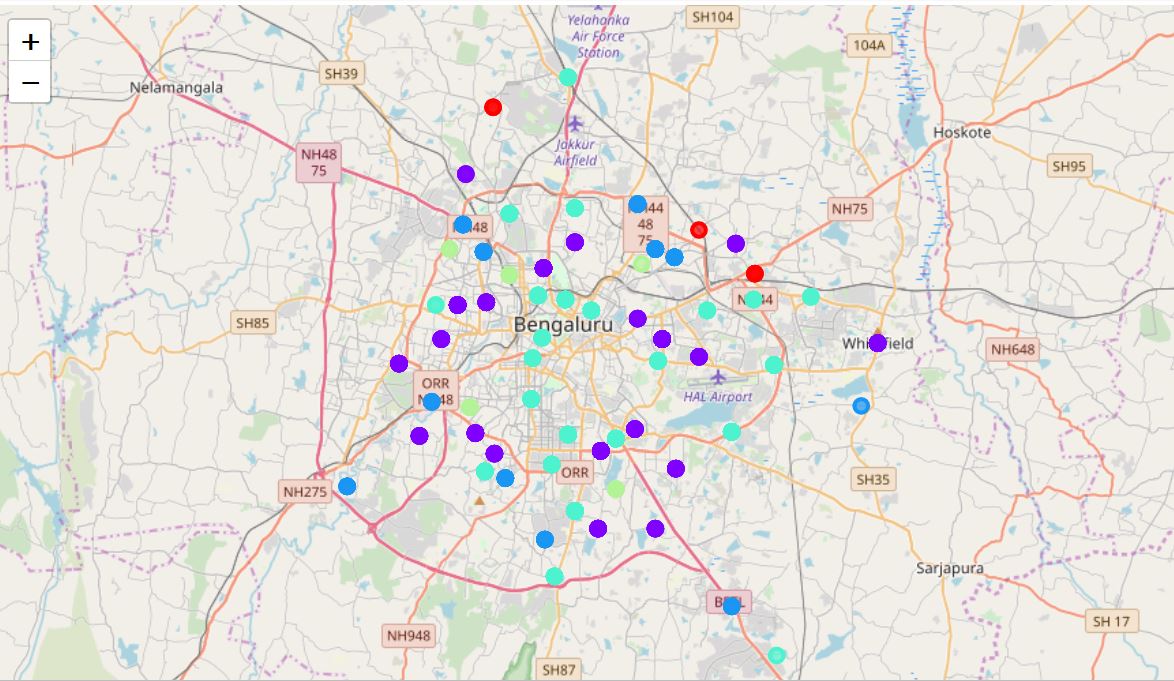
**Result for Kammanahalli:**



**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 similar neighborhood 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.