The Battle of Neighborhoods

IBM Data Science Professional Certificate - Capstone Project

Restaurant Recommender System, Delhi

Vaibhav Thakur

Introduction

Problem background:

Delhi is the capital of India. With a population of over 30 million (estimated as of 2020), Delhi is the second largest city in India and also 2nd largest city in the world.

The diversity of the cuisine available is reflective of the social and economic diversity of Delhi. Roadside vendors, tea stalls, South Indian, North Indian, Muslim food, Chinese and Western fast food are all very popular in the city. Satvik 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. Delhi can also be called a foodie's paradise because of its vast variety of foods and edibles with a touch of Delhi's uniqueness and tradition

Problem description:

Suppose someone travel and keep changing places very frequently. This is very hectic process as he had to experience very different types of environment, of which he do not have much knowledge about. In such situation, food can be an important factor for decided how you rate your trips and plus also recommending it to the people. Food can also attract people around to world to try it out if it were to be the best. In such scenarios, we need to find the right place, at reasonable cost, to serve us the best possible way. So there are few questions that must be addressed, such as:

1) How many types of foods are available in the restaurant?
2) Which is the most nearest to me with good rating?
3) How many "similar" restaurants are available nearby me?
4) Do the "similar" restaurants cost more? If so, what specialty do that have?

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 restaurants 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 my preferences?

Target Audience:

Target audiences for this project does not limit to a person who keeps travelling but everyone. People could simply decide to look for a similar restaurant all the time because they are addicted to a specific category of food. People who rarely use restaurants would prefer to have the most rated restaurants nearby them and all this could be easily handled by our recommender system. So target for this project is basically everyone who is exploring different places or similar places

Data

Data requirements:

To solve this problem, we will need the following data:

- List of Neighborhoods in Delhi
- Latitude and Longitude coordinates of all the Neighborhoods
- Venue data (powered by Foursquare), particularly related to Restaurants

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 maps.ie to fetch latitude and longitude individually of each neighborhood. I scrapped list of neighbor's using beautifulSoup4 from Wikipedia . After doing so, I produced the following data frame In general, this project would be encompassing a series of Data Science techniques, including, but not limited to, Web Scraping (using Beautiful Soup and Requests), Data Cleaning, Data Wrangling and Machine Learning (K-Means clustering algorithm)

	Neighborhood	City	Latitude	Longitude
0	Adarsh Nagar	Delhi	28.720341	77.172661
1	Ashok Vihar	Delhi	28.690420	77.176064
2	Azadpur	Delhi	28.712420	77.173111
3	Bawana	Delhi	28.797661	77.045258
4	Begum Pur	Delhi	28.732599	77.052170

2. 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	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adarsh Nagar	28.720341	77.172661	Giani's	28.717900	77.173907	Ice Cream Shop
1	Adarsh Nagar	28.720341	77.172661	My Idea Store	28.717487	77.170922	Mobile Phone Shop
2	Adarsh Nagar	28.720341	77.172661	Axis Bank ATM	28.722080	77.168740	ATM
3	Adarsh Nagar	28.720341	77.172661	Adarsh Nagar Metro Station	28.716598	77.170436	Light Rail Station
4	Adarsh Nagar	28.720341	77.172661	K.C Food Zone	28.718279	77.176971	Snack Place

Methodology

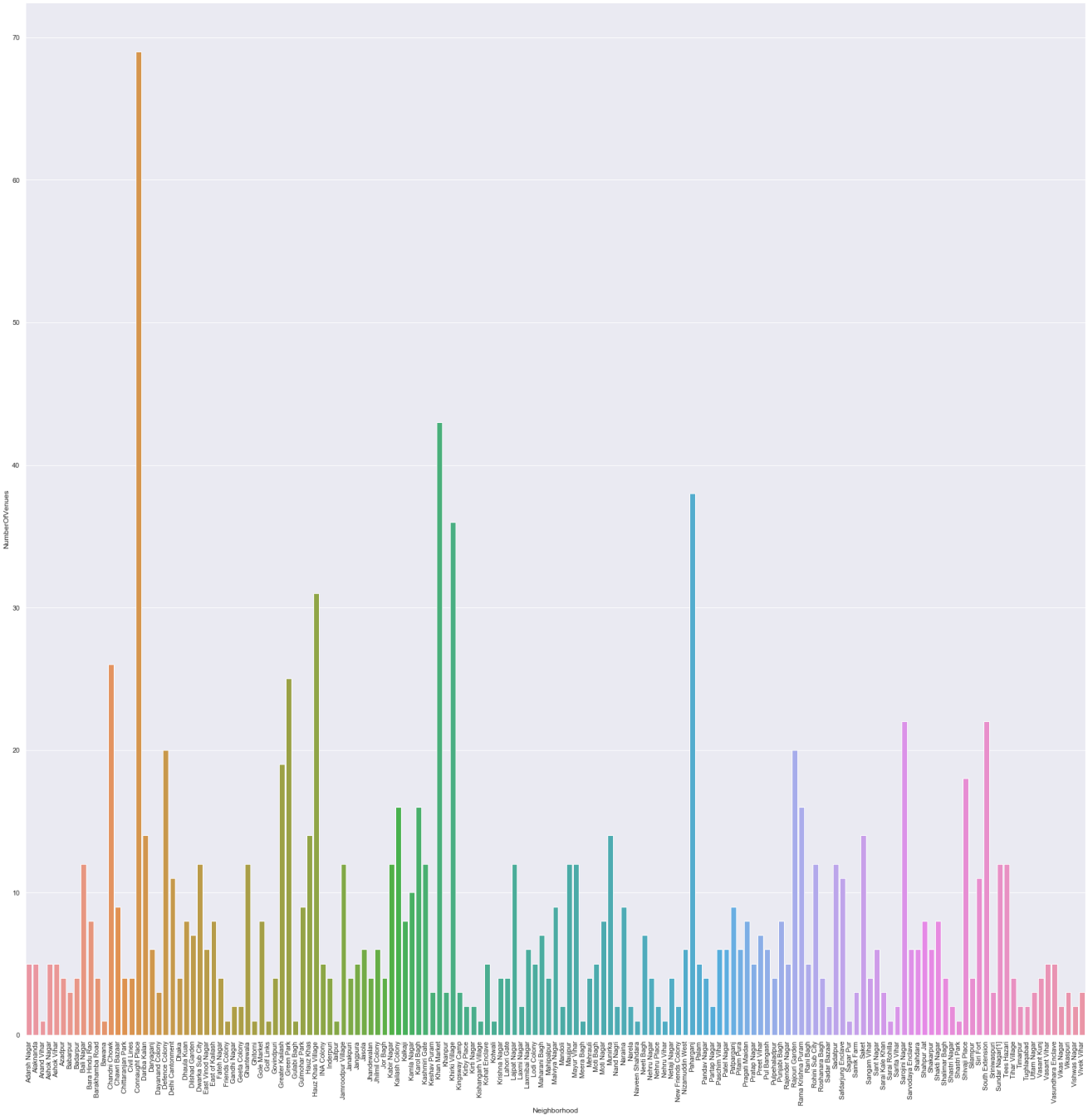
Exploratory analysis:

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 Connaught Place has most number of venues while Anand Vihar has the least.

```
Neighborhood      Connaught Place
NumberOfVenues                 69
Name: 16, dtype: object
Neighborhood      Anand Vihar
NumberOfVenues              1
Name: 2, dtype: object
```



Modelling:

Using the final dataset containing the neighborhoods in Kingston upon Thames along with the latitude and longitude, we can find all the venues within a 500 meter radius of each neighborhood by connecting to the Foursquare API. This returns a json file containing all the venues in each neighborhood which is converted to a pandas dataframe. This data frame contains all the venues along with their coordinates and category

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adarsh Nagar	28.720341	77.172661	Giani's	28.717900	77.173907	Ice Cream Shop
1	Adarsh Nagar	28.720341	77.172661	My Idea Store	28.717487	77.170922	Mobile Phone Shop
2	Adarsh Nagar	28.720341	77.172661	Axis Bank ATM	28.722080	77.168740	ATM
3	Adarsh Nagar	28.720341	77.172661	Adarsh Nagar Metro Station	28.716598	77.170436	Light Rail Station
4	Adarsh Nagar	28.720341	77.172661	K.C Food Zone	28.718279	77.176971	Snack Place

One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction). The Venues data is then grouped by the Neighborhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighborhoods.

To help people find similar neighborhoods in the safest borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster all neighborhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

Results

After running the K-means clustering we can access each cluster created to see which neighborhoods were assigned to each of the five clusters. 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 while also showing whether there is a metro station nearby.

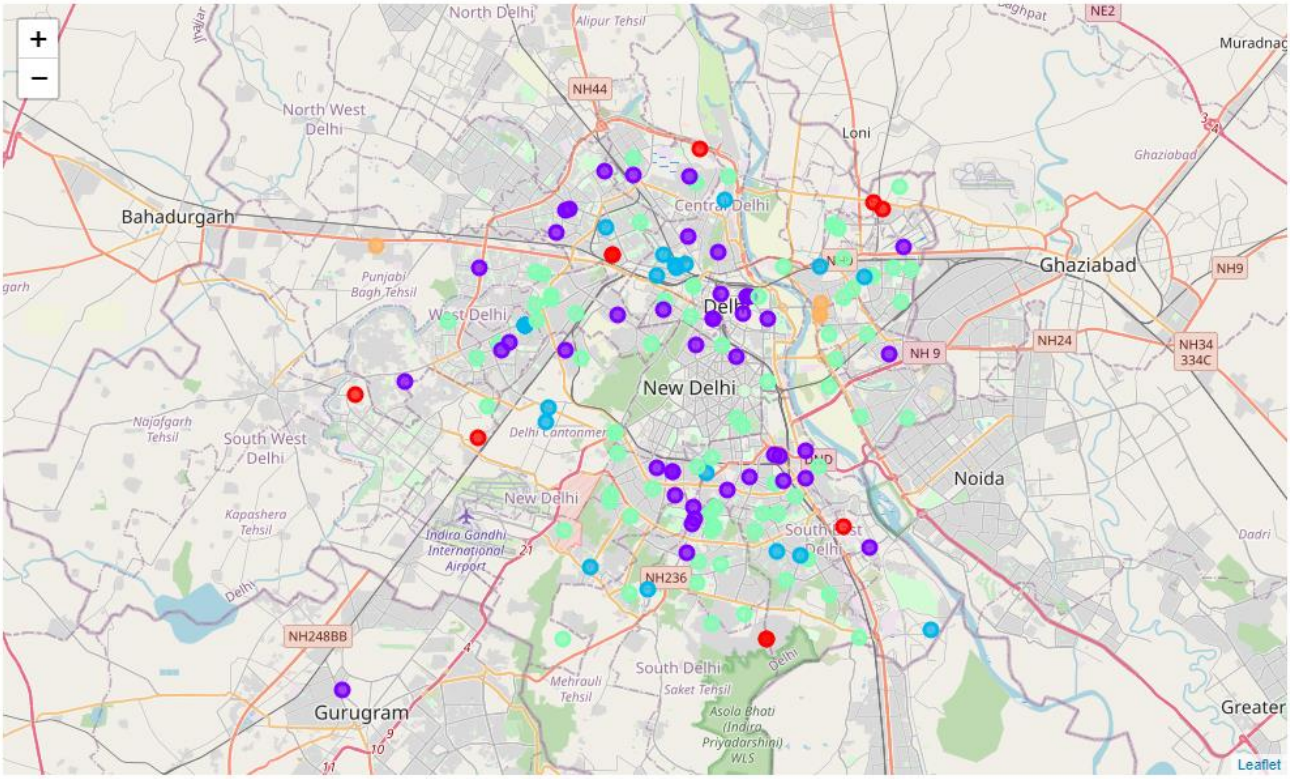
Looking into the neighborhoods in the first cluster.

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Station
3	Delhi	ATM	Dessert Shop	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Food Court	Food & Drink Shop	No
12	Delhi	ATM	Dessert Shop	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Food Court	Food & Drink Shop	No
18	Delhi	ATM	Fast Food Restaurant	Food & Drink Shop	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Food Court	No
40	Delhi	ATM	Fast Food Restaurant	Food & Drink Shop	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Food Court	No
36	Delhi	ATM	Dessert Shop	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Food Court	Food & Drink Shop	No
133	Delhi	ATM	Dessert Shop	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Food Court	Food & Drink Shop	No
50	Delhi	ATM	Smoke Shop	Fruit & Vegetable Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Food Court	Food & Drink Shop	Food	No
78	Delhi	ATM	Shipping Store	Furniture / Home Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Food Court	Food & Drink Shop	Food	No
119	Delhi	ATM	Snack Place	Smoke Shop	Indian Restaurant	Ice Cream Shop	Furniture / Home Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	No
134	Delhi	ATM	Smoke Shop	Fruit & Vegetable Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Food Court	Food & Drink Shop	Food	No
157	Delhi	ATM	Airport Lounge	IT Services	Business Service	Food & Drink Shop	Fruit & Vegetable Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Yes
183	Delhi	ATM	Park	Food	Fruit & Vegetable Store	Frozen Yogurt Shop	French Restaurant	Food Truck	Food Service	Food Court	Food & Drink Shop	No

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

Conclusion

The recommender system is a system that considers factors location and metro availability 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.