Introduction/Business Problem

Food is a feeling. You eat when you're hungry, sad, nostalgic, bored, in love, out of love and obviously when it is meal time. It soothes and it satiates. Indians take their food very seriously and the food industry is a major contributor to the economy of India.

In this project, I will be creating clusters of restaurant types in New Delhi which are not from our country (Indian food restaurants). Creating such clusters will benefit entrepreneurs and companies who wish to open their new restaurants in the region. This can also be used to classify what type of food type is popular in a given region. More number of restaurant types in an area suggest the popularity of the food in that area. In this clustering, I am removing Indian food as I wish to see the Non Indian food type restaurants which are popular in New Delhi.

Data

1. We need data about New Delhi. We know that New Delhi is divided into districts which are further divided into tehsils. Link - https://simple.wikipedia.org/wiki/List_of_districts_in_Delhi

   So we will extract and create a list of all the subdivisions in New Delhi

   Columns of the data

Sl.No.	District	Headquarter	Sub divisions (Tehsils)

2. Getting Latitude and Longitude of these tehsils

We will use the Geopy library to get the latitude and longitude. In this library, we provide address and function returns the lat and long. I will append "New Delhi" at the end of addresses as there are other places in India with the same name.

3. Using Foursquare Data to get venue related data in these tehsils.
This will help us classify the neighborhoods.

Methodology

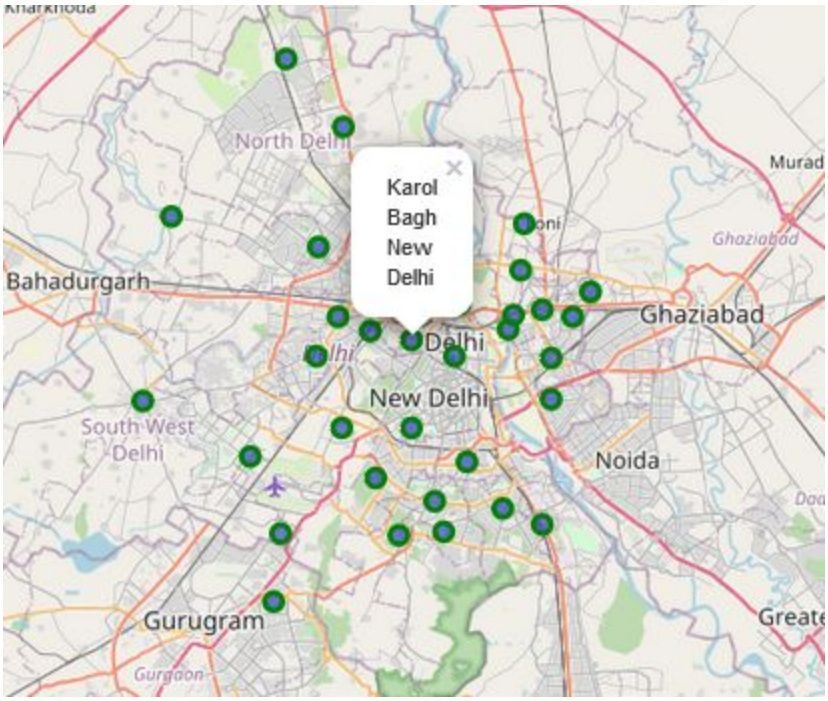
First, I had to find a way to analyze the restaurants in New Delhi for which I required a list of subdivisions in New Delhi. This is possible by extracting the list from the wikipedia page https://simple.wikipedia.org/wiki/List_of_districts_in_Delhi.

I scraped this data and stored it in a pandas dataframe using the read_html function as it can be used to extract any tabular data from the HTML.

	Sl.No.	District	Headquarter	Sub divisions (Tehsils)	Sub divisions (Tehsils).1	Sub divisions (Tehsils).2
0	1	New Delhi	Connaught Place	Chanakyapuri	Delhi Cantonment	Vasant Vihar
1	2	North West Delhi	Narela	Model Town[3]	Narela	Alipur
2	3	North Delhi	Kanjhawala	Rohini	Kanjhawala	Saraswati Vihar
3	4	West Delhi	Rajouri Garden	Patel Nagar	Punjabi Bagh	Rajouri Garden
4	5	South West Delhi	Dwarka	Dwarka	Najafgarh	Kapashera
5	6	South Delhi	Saket	Saket	Hauz Khas	Mehrauli
6	7	South East Delhi	Defence Colony	Defence Colony	Kalkaji	Sarita Vihar
7	8	Central Delhi	Daryaganj	Kotwali	Civil Lines	Karol Bagh
8	9	North East Delhi	Seelampur	Seelampur	Yamuna Vihar	Karawal Nagar
9	10	Shahdara	Shahdara	Shahdara	Seemapuri	Vivek Vihar
10	11	East Delhi	Preet Vihar	Gandhi Nagar	Preet Vihar	Mayur Vihar

There are 11 districts in New Delhi. The data is not enough. So I created a list of all the subdivisions of New Delhi.

Now, to use the data from foursquare API we need to get the latitude and longitude data of these subdivisions.I used the Geopy library to get the latitude and longitude. In this library, we provide address and function returns the latitude and longitude. After getting the data , it was checked for any errors. After gathering all these coordinates, I visualized the map of New Delhi using the Folium package to verify whether these are correct coordinates. Two columns had wrong coordinates. Since the data was small I manually corrected those coordinates.



Then, I used the Foursquare API to get top 100 venues in a 2000 radius meter. To use the API, I created an account to get the CLIENT_ID and CLIENT_SECRET key. The API returns the name, venue category and its latitude and longitude. I received data of size (1148, 7). With this data, I can check the number of unique

venues, different restaurant types and count of these restaurant types from the data.

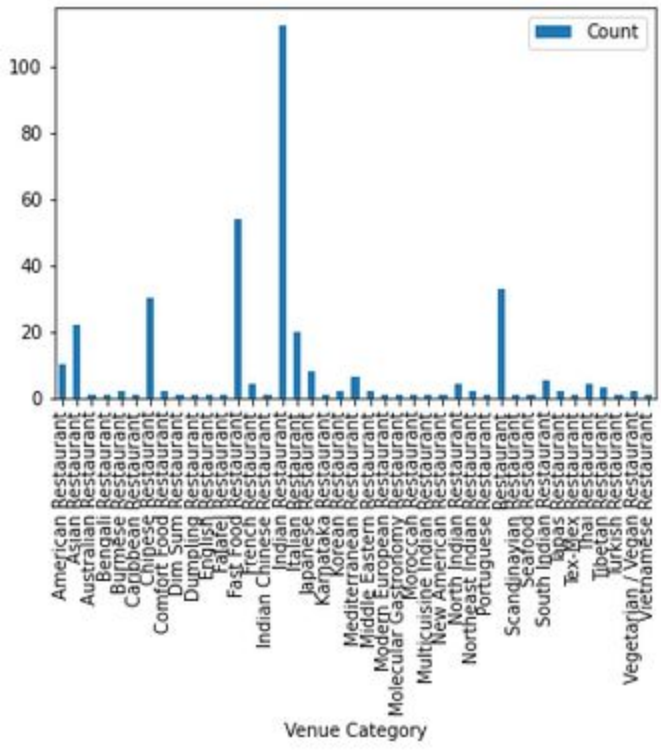
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
483	Punjabi Bagh New Delhi	28.668945	77.132461	Dunkin'	28.666258	77.126289	Donut Shop
875	Saraswati Vihar New Delhi	28.477224	77.083276	Binge Bakery	28.471200	77.102062	Bakery
1062	Sarita Vihar New Delhi	28.528574	77.288331	Pind Balluchi	28.521593	77.294188	Indian Restaurant
992	Rajouri Garden New Delhi	28.642152	77.116060	Berco's	28.638316	77.129806	Chinese Restaurant
44	Chanakyapuri New Delhi	28.594677	77.188521	My Humble House-ITC Maurya	28.597324	77.173609	Chinese Restaurant

Sample Data

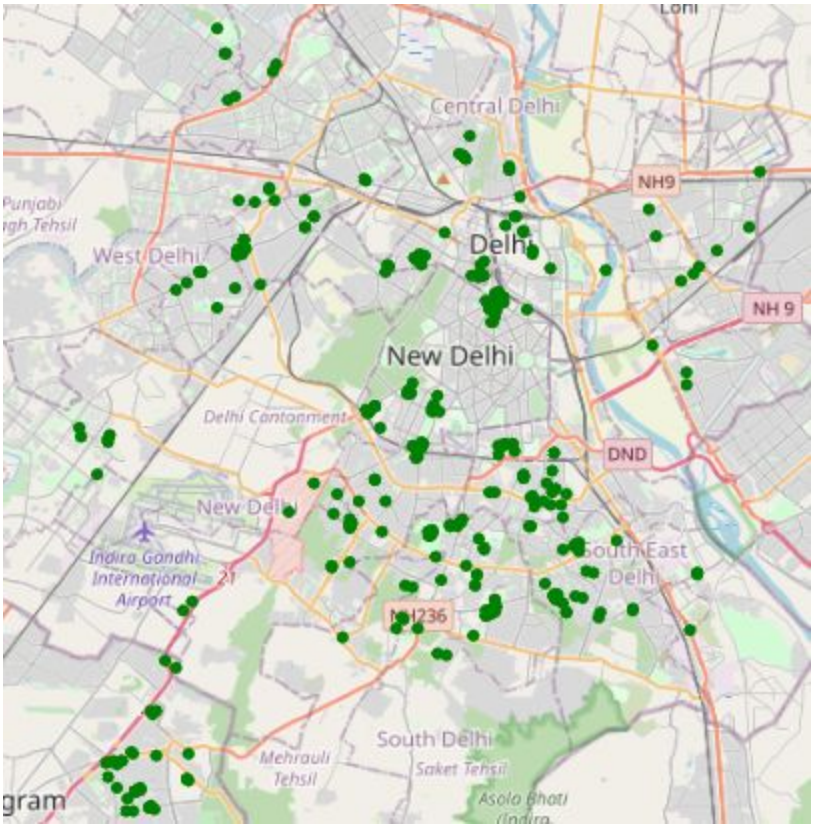
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
2	Chanakyapuri New Delhi	28.594677	77.188521	Sanadige	28.601969	77.187020	Karnataka Restaurant
3	Chanakyapuri New Delhi	28.594677	77.188521	Moti Mahal Delux	28.601677	77.187106	Indian Restaurant
4	Chanakyapuri New Delhi	28.594677	77.188521	Lázeez Affaire	28.602237	77.186044	Indian Restaurant
5	Chanakyapuri New Delhi	28.594677	77.188521	Bukhara	28.596914	77.173358	North Indian Restaurant
8	Chanakyapuri New Delhi	28.594677	77.188521	Jakoi	28.605239	77.187581	Northeast Indian Restaurant
...	...	...	...	...	...	...	...
1106	Karol Bagh New Delhi	28.652998	77.189023	Alfa Spice	28.644484	77.178748	Multicuisine Indian Restaurant
1118	Vivek Vihar New Delhi	28.669164	77.312267	Cafe Wink	28.657311	77.317098	Italian Restaurant
1135	Mayur Vihar New Delhi	28.613107	77.295722	Haldiram's	28.617935	77.279686	North Indian Restaurant
1139	Mayur Vihar New Delhi	28.613107	77.295722	Neelgiri south indian restaurant	28.608433	77.292937	Indian Restaurant
1143	Mayur Vihar New Delhi	28.613107	77.295722	Sameer's Restaurant	28.604307	77.292937	Indian Restaurant

349 rows × 7 columns

Data of Restaurants



Graph of Restaurant Types



Restaurant points on the Map

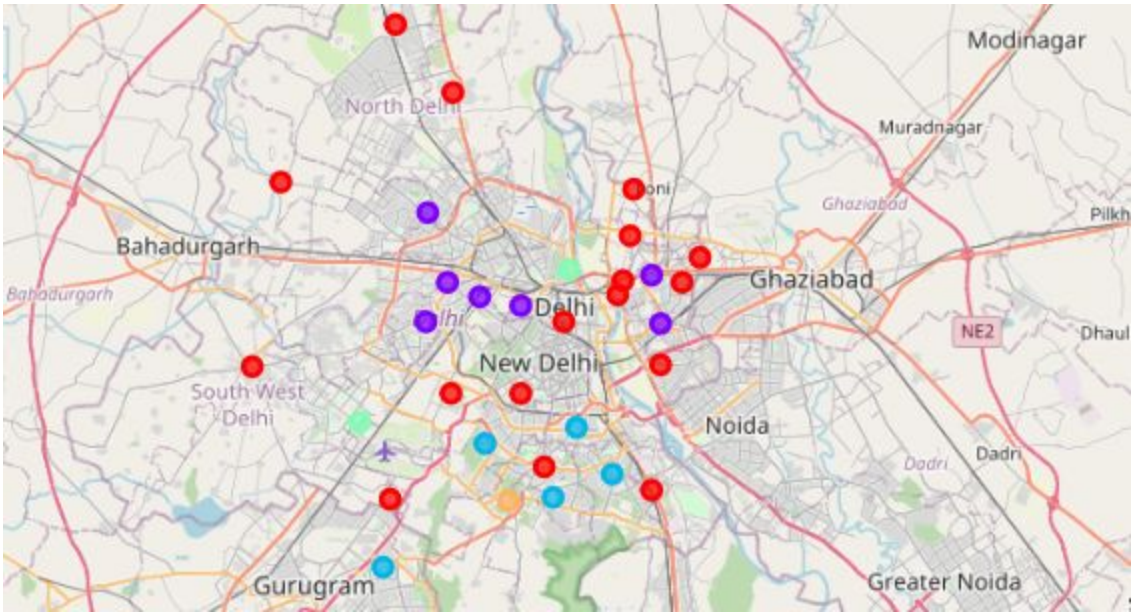
Then, I analyzed each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Since I wanted to analyse the data about the Non Indian restaurants, I will remove all the data that contains Indian Restaurant. Then, I looked at the topmost restaurant types and most common venues for a given subdivision.

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centriods, and then allocates every data point to the nearest

cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I clustered the subdivisions into 5 clusters based on frequency of occurrence of a food type restaurant.

Results



Our K means algorithm divided the area into 5 clusters as follows:

1. Vietnamese, French and Japanese restaurants (RED)
2. Fast Food Restaurants (PURPLE)
3. Chinese, Thai and American Restaurants (BLUE)
4. Asian and Fast Food Restaurants (CYAN)
5. Italian Restaurants (ORANGE)

Region names can be seen from the notebook.

Based on these clusters, decisions can be made by an entrepreneur who wishes to open a new restaurant. He/she can look to open a restaurant where the competition is low but the food type is popular. However, if the food quality is authentic and affordable, the restaurant will work.

Example : If someone wants to open a new Fast food restaurant, he can see that most of the fast food restaurants are in the cluster 1. So, the competition is high in these areas. Also, cluster 4 and 2 have less fast food restaurants. So, a new restaurant can be opened in these areas.

Limitations and Suggestions for Future Research:

1. Foursquare data is limited in India. Other apis, like zomato would give better results in India.
2. There are many other factors that can influence the decision of opening a new restaurant.
    1. population density,
    2. income of residents,
    3. rent that could influence the decision to open a new restaurant. However, to put all these data into this project is not possible to do within a short time frame for this capstone project.
3. We need more data to produce better results.

Conclusion

In this project, we have successfully identified the business problem, scraped and cleaned the data from the internet, performed exploratory analysis, used K means algorithm for clustering and provided our recommendation.