IBM DATA SCIENCE CAPSTONE PROJECT

OPENING AN ASIAN RESTAURANT IN CHICAGO

by Prakhar Rajput

Feb 2020

Introduction

For this Capstone project, My client wants to open an authentic asian restaurant in Chicago area. The idea behind this project is that there may not be enough Asian restaurants in Chicago and it might present a great opportunity for this entrepreneur who is based in India.This entrepreneur is thinking of opening this restaurant in locations where Asian food is popular where many Asian restaurants are in the neighborhood. With this in mind, finding the location to open such a restaurant is one of the most important decisions for my client and I am creating this project to help him find the best place for opening his restaurant.

Business Problem

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new asian restaurant in Chicago. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: In Chicago, if an entrepreneur wants to open a asian restaurant, where should they consider opening it?

Data

To solve this problem, I will need below data:

- List of neighborhoods in Chicago,IL,USA.
- Latitude and Longitude of these neighborhoods.
- Venue data related to Asian restaurants. This will help us find the neighborhoods that are most suitable to open an asian restaurant.

Extracting the data

- Scrapping of Chicago neighborhoods from Wikipedia
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

Methodology

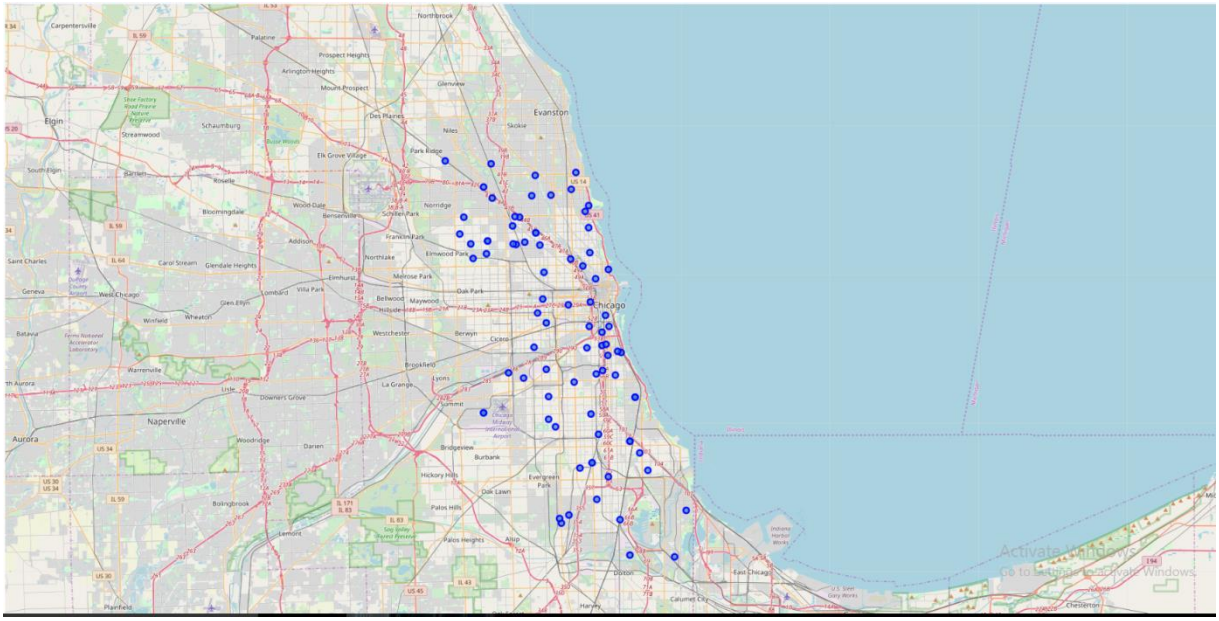
Getting the list of neighborhoods from Chicago, IL, USA. Using the list of neighborhoods from Wikipedia page("'https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago").
Web scraping was done by using the pandas. The table was extracted from the web page and stored in a dataframe. Below is the picture of the head of the extracted dataframe.

	Neighborhood	Community area
0	Albany Park	Albany Park
1	Altgeld Gardens	Riverdale
2	Andersonville	Edgewater
3	Archer Heights	Archer Heights
4	Armour Square	Armour Square
5	Ashburn	Ashburn

After scraping was done, the next task was to find the co-ordinates of the neighborhoods in the dataframe. For this task, the geocoder lib is used to find the latitude and longitude of the neighborhoods. (Note: geocoder does return coordinates of only 17 or 18 rows in one call, to get coordinates of all the rows one must call 17 rows multiple times.)
The dataframe below contains rows with latitude and longitude.

	Neighborhood	Community area	LAT	LONG
0	Albany Park	Albany Park	41.9719	-87.7162
1	Altgeld Gardens	Riverdale	41.6549	-87.6004
2	Andersonville	Edgewater	41.9771	-87.6693
3	Archer Heights	Archer Heights	41.8114	-87.7262
4	Armour Square	Armour Square	41.84	-87.6331
5	Ashburn	Ashburn	39.0046	-77.4908

Using the folium library I visualized the neighborhoods on the map. Below is the picture of the Chicago map with neighborhoods marked.

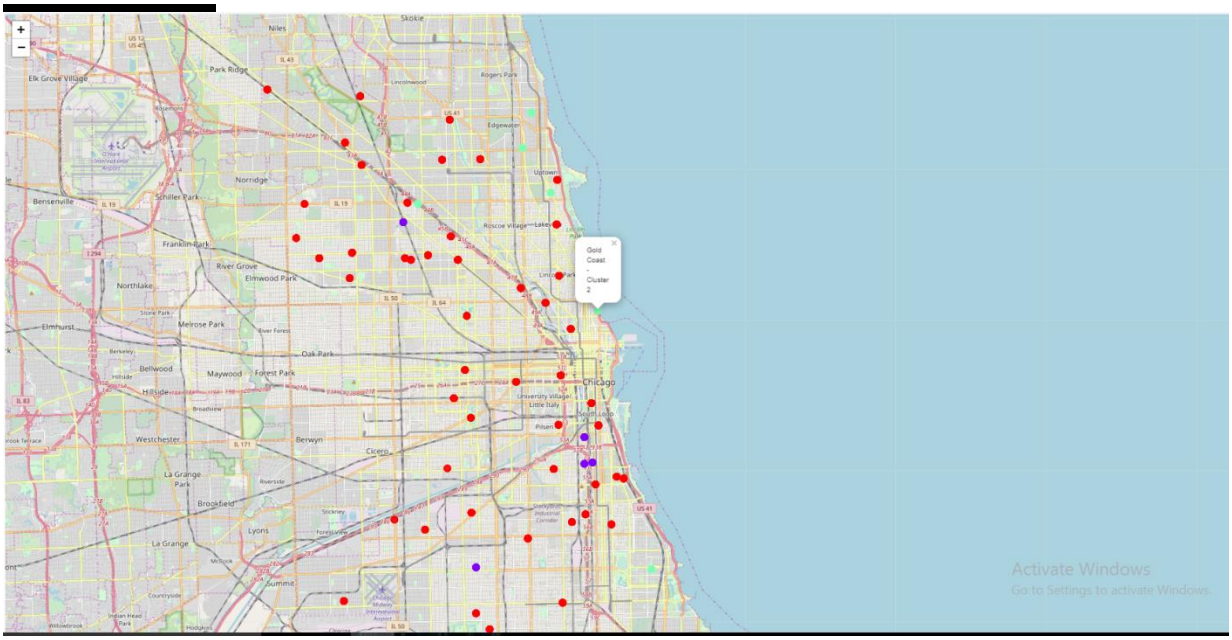


Now using the Foursquare API, I extracted the list of top 100 venues in a radius of 500 meters Using my API key I extracted names, categories, latitude, and longitude of the venues. After that grouping the rows by neighborhoods and taking the mean on the frequency of occurrence of each venue category, this is done for doing clustering.

Then the clustering is performed by using k-means since it is easy and simple to use.k-means is an unsupervised algorithm and can distinguish data points into clusters

The neighborhood has been divided into three clusters on the basis of frequency of "Asian restaurant" .Based on the result the recommendations can be given.

Result



In the above map, the neighborhoods have been divided into three clusters :
Cluster 0: Less Frequency of Asian restaurants, close to zero (Recolor)
Cluster 1:No Asian restaurants(Purple color)
Cluster 2:High Frequency of Asian Restaurants(Light Green)

Discussion

Cluster 2 represents most no of Asian restaurants in places like Andersonville and East Hyde Park and cluster 1 have lowest(or none) Asian restaurants. Our client could open his restaurant in places like Gage Park or Fernwood with less or no competition. If the services provided are of good quality then the restaurant could be successful.
In the above scenario, only the frequency of occurrence Asian restaurants was taken, while the demographics of neighborhoods could also be taken as a factor if the data is available.

Conclusion

Chicago is a big city and opening an Asian restaurant in places with less competition is more likely to grow, also keeping in mind the quality of services been offered. Using k-means we were able to cluster the neighborhoods and get the solution to our problem.

References

List of neighborhoods of Chicago:
https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago
Foursquare Documentation:
https://developer.foursquare.com/docs