CAPSTONE PROJECT-THE BATTLE OF NEIGHBOURHOOD

CH.VINAY

1.Introduction

1.1 Background :

For this Capstone project, I am creating a hypothetical scenario for a concept Burmese restaurateur who wants to explore opening an authentic Burmese restaurant in Toronto area. The idea behind this project is that there may not be enough Burmese restaurants in Toronto and it might present a great opportunity for this entrepreneur who is based in Canada. As Burmese food is very similar to other Asian cuisines, this entrepreneur is thinking of opening this restaurant in locations where Asian food is popular (aka many Asian restaurants in the neighborhood). With the purpose in mind, finding the location to open such a restaurant is one of the most important decisions for this entrepreneur and I am designing this project to help him find the most suitable location.

Business Problem

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

Target Audience

The entrepreneur who wants to find the location to open authentic Burmese restaurant

Data

To solve this problem, I will need below data :

- List of neighborhoods in Toronto, Canada.
- 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 a Burmese restaurant.

Extracting Data

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

Methodology

	Burglary	Criminal Damage	Drugs	Other Notifiable Offences	Robbery	Theft and Handling	Violence Against the Person	Total
count	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000	33.000000
mean	2069.242424	1941.545455	1179.212121	479.060606	682.666667	8913.121212	7041.848485	22306.696970
std	737.448644	625.207070	586.406416	223.298698	441.425366	4620.565054	2513.601551	8828.228749
min	2.000000	2.000000	10.000000	6.000000	4.000000	129.000000	25.000000	178.000000
25%	1531.000000	1650.000000	743.000000	378.000000	377.000000	5919.000000	5936.000000	16903.000000
50%	2071.000000	1989.000000	1063.000000	490.000000	599.000000	8925.000000	7409.000000	22730.000000
75%	2631.000000	2351.000000	1617.000000	551.000000	936.000000	10789.000000	8832.000000	27174.000000
max	3402.000000	3219.000000	2738.000000	1305.000000	1822.000000	27520.000000	10834.000000	48330.000000

Statistical summary of crimes

The count of each of the major categories of crime returns value of 33 which is the number of london boroughs.”Theft and handling” is the highest crime during the year 2016.

Methodology

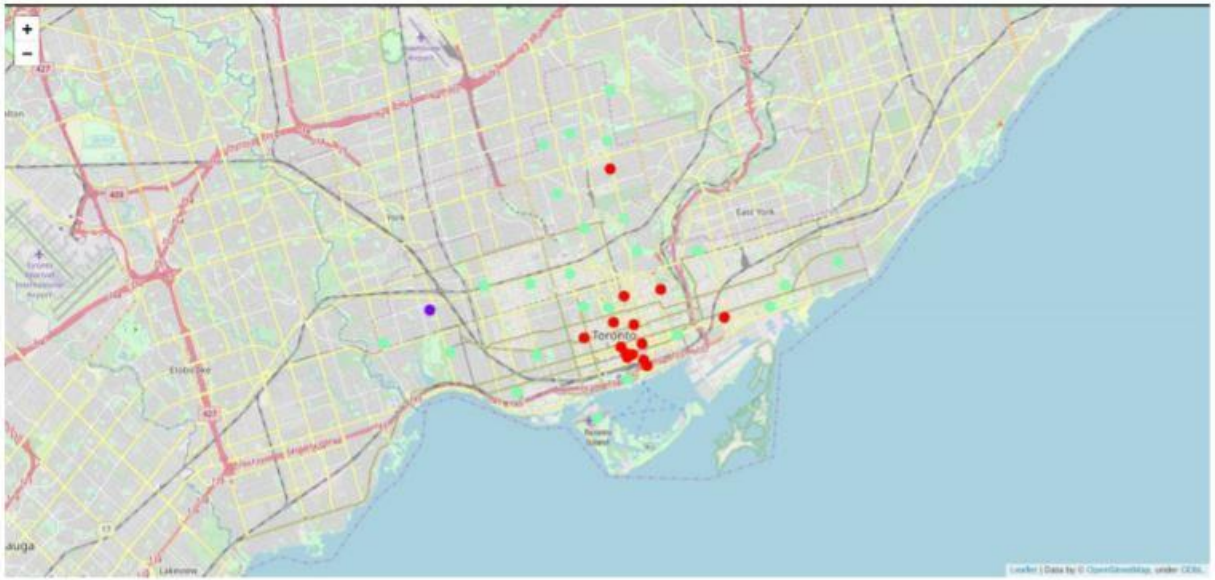
First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from wikipedia page (“ https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M ”) I did the web scraping by utilizing pandas html table scraping method as it is easier and more convenient to pull tabular data directly from a web page into dataframe. However, it is only a list of neighborhood names and postal codes. I will need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder package but it was not working so I used the csv file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering all these coordinates, I visualized the map of Toronto using Folium package to verify whether these are correct coordinates.

Methodology

Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze 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.

Here, I made a justification to specifically look for “Thai restaurants”. Previously, when I ran the model, I was looking for “Asian restaurants” but there are very few results (maybe due to Foursquare categorization) so I looked for the restaurants closest to Burmese cuisine taste.

Results


Toronto
York
East York
Etobicoke
Leaflet | Data by © OpenStreetMap, under ODbL.

Results

Clusters:

The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Thai restaurants are in each neighborhood:

- Cluster 0: Neighborhoods with little or no Thai restaurants
- Cluster 1: Neighborhoods with no Thai restaurants
- Cluster 2: Neighborhoods with high number of Thai restaurants

Recommendations

Most of Thai restaurants are in Cluster 2 which is around Adelaide, King, Richmond areas and lowest (close to zero) in Cluster 1 areas which are North Toronto West and Parkdale areas.

Also, there are good opportunities to open near Chinatown, St James town as the competition seems to be low. Looking at nearby venues, it seems Cluster 1 might be a good location as there are not a lot of Asian restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic Burmese restaurant in these locations with little to no competition. Nonetheless, if the food is authentic, affordable and good taste, I am confident that it will have great following everywhere :)

Limitations and Suggestions for Future:

In this project, I only take into consideration of one factor: the occurrence / existence of Thai restaurants in each neighborhood. There are many factors that can be taken into consideration such as population density, income of residents, 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. Future research can take into consideration of these factors. In addition, I am relying on the existence of Thai restaurants only for this project but future research can take into consideration of other variables such as existence of Asian restaurants, Asian population level in each neighborhood etc.

Conclusion:

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.

References:

List of neighborhoods in Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare Developer Documentation: https://developer.foursquare.com/docs