Capstone Project

IBM DATA SCIENCE

FINDING THE OPTIMAL BUSINESS LOCATION USING GEOSPATIAL CLUSTER VISUALIZATION

Project by:-
Rahul Goswami

CONTENTS

Problem description and Business Understanding

In this project the problem that will be attempted to be solved is as follows:-

A local business in our case is a reputed chain of Chineese Restaurants who wants to extend to Toronto which experiences a decent influx of chineese people every year. Henceforth, we have been tasked with a goal to demarcate the Optimal spot(hotspot in geospatial visualization) or rephrasing the above lines: Find out the best possible location for a restaurant which will lead to an increased profitability with minimum competition and regular customer influx from around different neighborhoods.

For, this situation we will be exploring the different neighbourhoods datasets and analysing them after we build a classification model based on machine learning algorithms taking all possible parameters into consideration. We will also be using the ‘Foursquare API’ to provide us with deep Insight into every neighborhood in Toronto which will label be used to cluster venues accordingly.

Data Requirements

The dataset that will be used here is the list of all borough along with their neighbourhoods enlisted by Postal codes which can be found at:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

A web scraping tool will be utilized to scrape the tables on this page into a dataframe which can be used by our machine learning algorithms for further analysis.

Also an ethnic distributions of neighbourhoods which contain 'Chineese' people as one of the top three ethnic groups are required which is obtained from:
https://en.wikipedia.org/wiki/Demographics_of_Toronto

Further, Foursquare API will be used to find out the top venues for each neighbourhood and if there are competing local businesses in the area.This will serve as the basis for clustering the neighborhoods and hence each cluster will be analysed critically for risk evaluation for setting up businesses in the area.

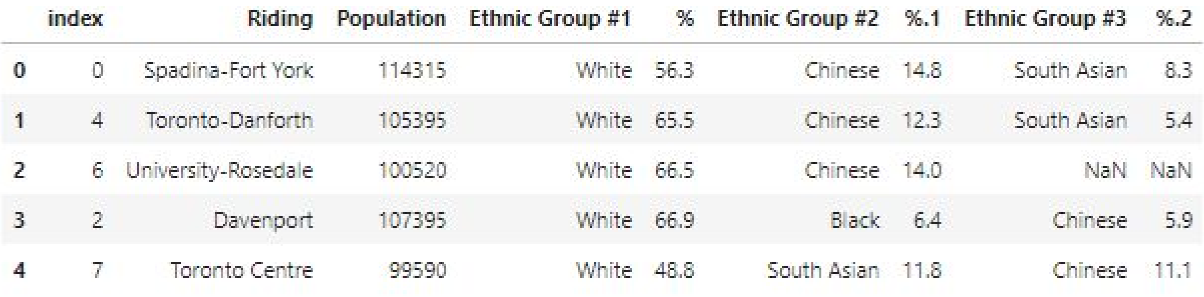
Methodology

Analytical Approach:- An effective way to approach this problem would be to categorize different neighbourhoods based on proximity to popular venues. Factors such as real estate prices, weather conditions etc can be adjusted later. We start with a simple approach as follows:

- We begin the data collection process from the demographic statistical report wiki page.
- We make an assumption here that a chineese restaurant will thrive only when there is a considerable number of chineese community people.
- Hence we shortlist the neighbourhoods that have a considerable number of chineese community people.
- From the wikipedia table we identify the borough having chineese as a major ethnic community. For each borough we consider the neighbourhood only when,'Chineese' community ranks among the top 3 ethnic communities in the neighbourhood
- From the table we infer the following:
    a. Chineese ethnic community is the topmost community with a population of 332,830
    b. The Riding with highest density is 'Scarborough-Agincourt' with a percentage of 47% which evidently becomes the sought after are for our target business.
    c. Other notable Riding include 'WillowDale' and 'Don Valley North'

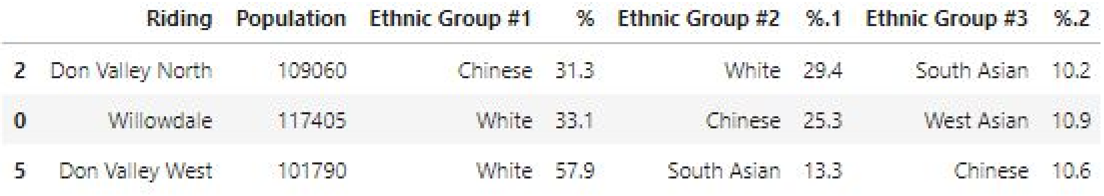
Now with the pandas framework we capture the individual neighbourhoods

With chineese community in the top3. We look up the tables of East York, North York and Scarborough which shows the most potential for chineese people. We filter out the entries based on Top3 ethnic communities.



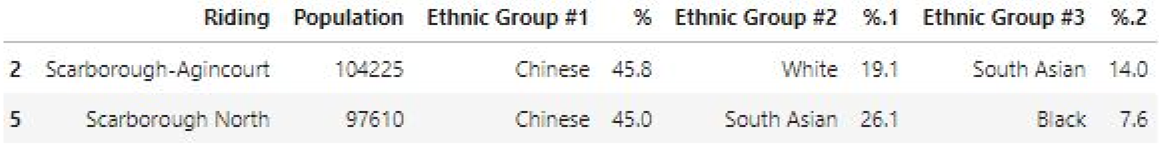
	index	Riding	Population	Ethnic Group #1	%	Ethnic Group #2	%.1	Ethnic Group #3	%.2
0	0	Spadina-Fort York	114315	White	56.3	Chinese	14.8	South Asian	8.3
1	4	Toronto-Danforth	105395	White	65.5	Chinese	12.3	South Asian	5.4
2	6	University-Rosedale	100520	White	66.5	Chinese	14.0	NaN	NaN
3	2	Davenport	107395	White	66.9	Black	6.4	Chinese	5.9
4	7	Toronto Centre	99590	White	48.8	South Asian	11.8	Chinese	11.1

Fig: Top chineese neighborhoods in Toronto



	Riding	Population	Ethnic Group #1	%	Ethnic Group #2	%.1	Ethnic Group #3	%.2
2	Don Valley North	109060	Chinese	31.3	White	29.4	South Asian	10.2
0	Willowdale	117405	White	33.1	Chinese	25.3	West Asian	10.9
5	Don Valley West	101790	White	57.9	South Asian	13.3	Chinese	10.6

Fig: Top chineese neighborhoods in North York



	Riding	Population	Ethnic Group #1	%	Ethnic Group #2	%.1	Ethnic Group #3	%.2
2	Scarborough-Agincourt	104225	Chinese	45.8	White	19.1	South Asian	14.0
5	Scarborough North	97610	Chinese	45.0	South Asian	26.1	Black	7.6

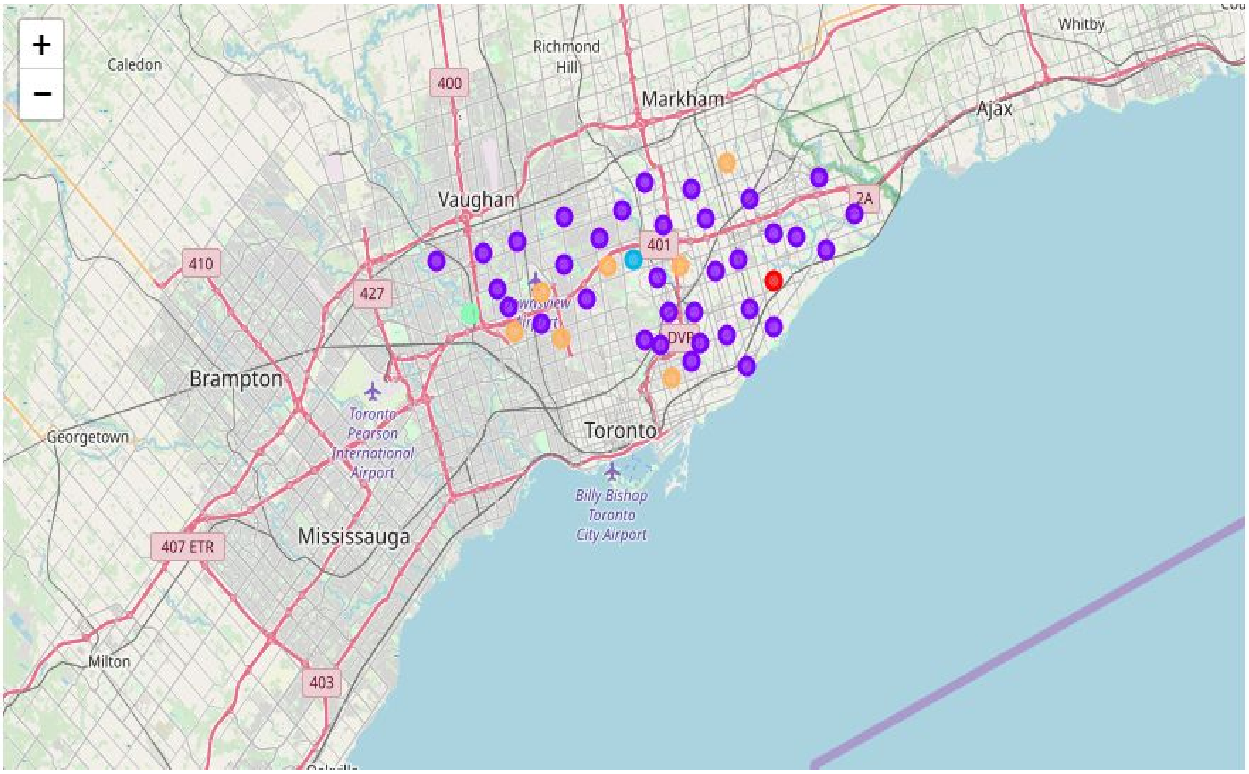
Fig: Top Chineese neighborhoods in Scarborough

As such our clustering zones are filtered, we only need to pick out the hot venues for the above enlisted boroughs our postal code dataset. We, however, consider it to be prudent to include all neighborhoods of East York, North York and Scarborough for proximity to be a parameter.

The lat-long value pairs can either be generated via geocoder library or is available at: 'https://cocl.us/Geospatial_data'. The lat-long values are already sorted via postal addresses and are appended to our dataframe as such. The resulting dataframe is stored as neighborhoods comprising of Postal Code, borough Name, Neighborhood, latitude, longitude.

Now we use the foursquare API to fetch the most popular venues for all the neighbourhoods in our dataframe and group the table according to the category of venues which are encoded via one-hot key encoding.

We use K-means clustering to generate cluster labels for each of the neighborhoods and append it to a dataframe called ‘Toronto_merged’. We take the value of k to be 5 as it denotes the elbow point for K-means and we really want to avoid redundant computations as well as keep our visualization neat. These labels are superimposed on the toronto map as markers with colors denoting the respective cluster they belong to.



RESULTS

The Map is generated using folium and some values which can't be clustered are dropped from the table. Now each clustered in analysed carefully and tallied with our search catalog.

- Cluster 1: Only 1Neighborhood in scarborough showing promise for vacancy.
- Cluster 2: Most number of neighborhoods in Scarborough, NY and EY having asian and chineese restaurants in common
- Cluster 3: Only 1 neighborhood with North York already preoccupied with european cuisine. A different flavour would be encouraged.
- Cluster4: Only 1 neighborhood with North York showing promise for vacancy .
- Cluster 5: Fitness area for most, contains parks, gyms, yoga centres playgrounds etc. Holds nutritional ingredients key. Show less promise. This includes don valley and willowdale.

Thus, as we have successfully analysed the clustered, it is now up to the client to make a decision. A number of options have to be explored further to make a decision. For instance cluster 1 could be of promise but that just includes 1 neighborhoods. Cluster 2 and 5 would be difficult to deal with and could ask some serious questions. Since the highest, number of people live in scarborough the decision could rest between 1 and 2. But people in cluster 2 are already familiar with the cuisine. Hence, a new route could be to take 3 and 4 to introduce variance. All the dots have to be joined. For example,other ethnic backgrounds apart from chineese such as south east Asian people would also find chineese food appealing and thus the relative percentages of other community also have a role to play.

Discussion

- A lot of borough were inconsistent with their data and couldn’t be clustered and thus were eventually dropped off the records. Whether or not this is normal is very debatable and open to everyone.
- Is cluster analysis essential to evaluate rivals competing for the same spot is also debatable. For instance, there could be average/mediocre level restaurants but whether our client can manage to provide better quality services is only up to them. Resource acquisition also remains an issue to be dealt with.

Conclusion

Thus, we are successfully able to categorize the different neighborhoods are rank them in the order of promise of setting up a restaurant according to community demands. Thus, the client would know have a holistic view of the situation in every area and all that remains for them is to decide how to implement this in their own way. They have been well presented with the profile of every neighbourhood with the percentages of risk involved according to area and current scenario.

Thank You