Predicting the best location for a business startups in a particular neighborhood

Introduction

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 in 2016. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. The city is home to the Toronto Stock Exchange, the headquarters of Canada's five largest banks, and the headquarters of many large Canadian and multinational corporations. Its economy is highly diversified with strengths in technology, design, financial services, life sciences, education, arts, fashion, business services, environmental innovation, food services, and tourism.

Business Problem

A way to increase the productivity and reach of a startup is to set it up in a location which is popular and there is a better chance for business. So in order to find this place, we have to research the data of various venues and places in the region. There are various datasets and information regarding different places and venues available on the Internet. We use this data to find the best places which would enable the business startups to succeed.

Even if you have a well functioning business startup, it is important that it is set up at the right place. So, the above mentioned process will help in maximizing the efficiency of the startup. Instead of selecting with an basic assumption it would be nice to use the data relating to the nearby venues in particular neighborhood (needed by the customer) and providing him the best areas to open his/her business.

Data Section

1. Data Sources

1. Data of neighborhoods and respective postal codes are gathered by scraping the web page (Wikipedia).
2. Using the postal codes the latitude and longitude locations of those respective neighborhoods are gathered from a ".csv" file provided on the Internet.
3. Data relating to the venues in a certain neighborhood according to the postal codes are gathered using the FOURSQUARE API.

2. Data Cleaning

1.

- The data obtained from the Wikipedia page is taken as HTML text at first so it taken into a pandas data frame.
- Then the rows with 'Borough' as 'Not assigned' are dropped.
- All the rows with 'Neighborhood' feature as 'Not assigned' are replaced with it 'Borough' feature.
- Then the locations are grouped according to the postal codes.

	Postalcode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront,Regent Park
3	M6A	North York	Lawrence Heights,Lawrence Manor
4	M7A	Queen's Park	Queen's Park

2.

- The ".csv" file is converted to a data frame.
- Latitude and longitude are filtered out according to the postal codes and are stored in 'latitude', 'longitude' features of the previous data frame.

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront,Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights,Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494

3. Features

- **Postal code** - contains the postal codes of the boroughs in Toronto.
- **Borough** - stores the Borough name.
- **Neighborhood** -contains comma separated list of all the neighborhoods that come under the Postal code and Borough
- **Latitude** - contains the latitude of that region obtained using the postal code.
- **Longitude** - contains the longitude of that region obtained using the postal code.

Methodology section

Once the customer has provided you with

- The details of the business he wanted to start.
- The region he wanted to start that business.

1. The venues nearby that region are taken from API of Foursquare. After the venues are gathered the locations are plotted on the map of Toronto and are clustered as per the need which indicates the crowdedness of shops around it. For clustering the venues, the K-Means clustering is used.

	venue	latitude	longitude	venue_type
0	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Donalda Golf & Country Club	43.752816	-79.342741	Golf Course
2	Brookbanks Park	43.751976	-79.332140	Park
3	Island Foods	43.745866	-79.346035	Caribbean Restaurant
4	Graydon Hall Manor	43.763923	-79.342961	Event Space
5	Darband Restaurant	43.755194	-79.348498	Middle Eastern Restaurant
6	LA Fitness	43.747665	-79.347077	Gym / Fitness Center
7	Galleria Supermarket	43.753520	-79.349518	Supermarket
8	Tim Hortons	43.760668	-79.326368	Café

K-Means clustering

K-means clustering is a clustering algorithm that aims to partition n observations into k clusters.

There are 3 steps:

Initialization *– K initial "means" (centroids) are generated at random.*

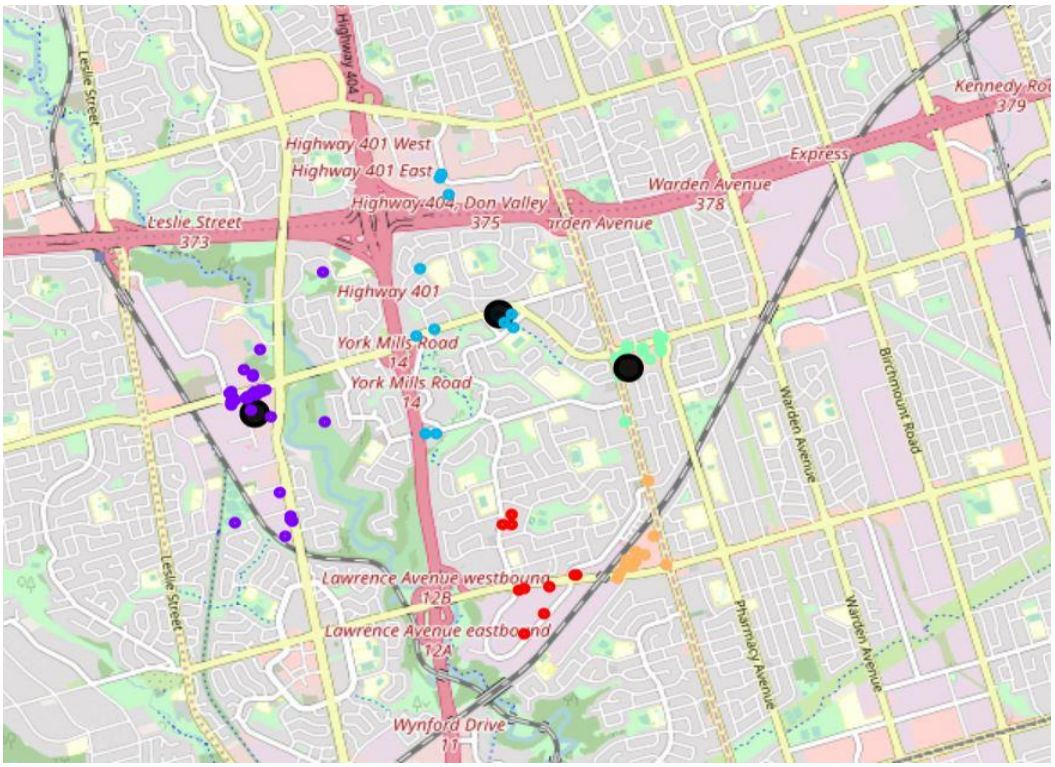
Assignment *– K clusters are created by associating each observation with the nearest centroid.*

Update *– The centroid of the clusters becomes the new mean*

Assignment and Update are repeated iteratively until convergence

The end result is that the sum of squared errors is minimized between points and their respective centroids.

Once the clustering is completed it gives the regions which are suitable for the business. As those regions will be populated due to high business region as there are many business around it.



2. Later the problem is, if there is a similar business in that area our business won't be highly profitable so we need to select the region with high population as good as possible .And also should be in area where there are no similar business. So a new data frame is created with the fields as **'cluster', 'count', 'btype_count'**

- **cluster**- the cluster id
- **count** - total number of venues in that cluster
- **btype_count**- total number of venues of the client business type in that cluster ('Shopping mall ' in this case)

	cluster	count	btype_count
0	1	37	2
1	3	15	3
2	2	14	2
3	4	12	0
4	0	10	3

In order to give maximum preference to the populated areas the data frame is sorted according to the count in decreasing order. And then the data frame is searched for a cluster which is having least no of similar business and that cluster is recommended to the customer

3. The newly created data frame is filtered for the minimum of 'btype_count'. The cluster with the minimum number of shopping malls is selected.

4. Then again the cluster with maximum number of 'count' is filtered. So the profits will increase due to the presence of famous venues nearby.

This gives the best list of clusters (location of cluster centers) in a given area where the customer can start the business with minimum risk.

Results

A list of best possible cluster centers (latitude, longitude) are give back along with the a visualization map by folium which also indicates the best locations for the business customer wanted to start in that area. With minimum risk involved, there will be less possible competition from other businesses of that area. It is taken care to provide the location with minimum competition for that business. And also making sure that the area nearby is populated or having popular venues.

Discussion Section

If someone wants to start a Coffee Shop in Parkwoods, the best way to do so, is by analyzing the number of Coffee Shops and the total number of venues present in the various clusters present in the map. We do this to ensure that the startup faces as little competition from other similar venues as possible. Here, it is better to start a Coffee Shop in the 4^{th} cluster as there are no other similar types of venues. This will help in the startup getting off to a good start.

Conclusion

People are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.