Optimal Location for opening Supermarket
Vaibhav Agrawal
IBM Capstone Data Science Project- 10th July' 21

Introduction:

An opening of a Supermarket is an interesting and high profit-making business avenue. However, it depends on how successfully you can run business of the Supermarket Store.

On a very broad level success of any supermarket store depends on:

a) Extremely important is - Right location of Store to make sure that customers can easily come and buy products.

b) Customer base - based on Supermarket products range, one should focus on right customer group. For example, the household's income.

c) Demand - There must be high demand which again depends on the population in the area where store is located, the customer base which one is trying to target and lastly how many such stores are already available in the vicinity.

In this project, we will attempt to solve the problem of a supermarket chain owner/ franchise owner and help them to identify which area / neighborhood in Toronto, Canada, they can open their new store. This will cater to supermarket chain owner, franchise owner for supermarket.
Thus, using the data science & machine learning techniques, this project tries to give a recommendation for an optimal location for opening of a supermarket.

Data:

Through this project, focus will be on below factors to decide optimal neighborhood for opening the store:

a) Type of Neighborhood, for example, business & offices, airports, re-creational, residential etc. - Most preferred option to target residential area as it will have maximum customer base.

b) Population & their income – For larger customer base, the neighborhood must have moderate to high population density and decent household income.

c) Current market penetration i.e., how many stores are already in the area

To work on above factor and solving the business problem, below data sets will be used:

First, we must identify the neighborhood for Toronto city. The Wikipedia page has list of neighborhoods.

We will use three-digit postal code to identify neighborhood.
Next, to use foursquare location API, we will also need latitude and longitude for each neighborhood. Using this geo-codes and Foursquare location API, we will explore each neighborhood. We will try to cluster the neighborhood based on different category of venues. This will help us further to find the residential areas.

Further, we will use census data to find out population per neighborhood and household income. This is needed to understand potential market for opening the Supermarket.
Sample screen shots:
1) Neighborhood with latitude and longitude

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure 1 – Toronto Neighborhood data

2) Census data

```
[ ] df_census.head()
```

	PostalCode	Borough	Neighbourhood Number	Population	Population density per square kilometre	Land area in square kilometres	Total - Household total income groups	Under $5,000	$5,000 to $9,999	$10,000 to $14,999	$15,000 to $19,999
0	M1B	Scarborough	263	90290	6208	45.74	26825	290	240	420	720
1	M1C	Scarborough	134	12494	2403	5.20	3700	60	25	45	60
2	M1E	Scarborough	411	54764	8570	19.04	19855	315	540	815	970
3	M1G	Scarborough	137	53485	4345	12.31	18445	435	455	685	1170
4	M1H	Scarborough	127	29960	4011	7.47	10765	615	220	255	450

Figure 2 – Toronto census data

References-

https://www.toronto.ca/city-government/data-research-maps/open-data/

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Methodology:

Following section describes methodology and step executed in the project:

a. Data Load:

1. **Load Toronto neighborhood data**
2. **Load census data**

```
import pandas as pd
df = pd.read_csv('/content/toronto.csv')
df.head()
```

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure 3 – Toronto Neighborhood dataframe

```
df_census.head()
```

	PostalCode	Borough	Neighbourhood Number	Population	Population density per square kilometre	Land area in square kilometres	Total - Household total income groups	Under $5,000	$5,000 to $9,999	$10,000 to $14,999
0	M1B	Scarborough	263	90290	6208	45.74	26825	290	240	420
1	M1C	Scarborough	134	12494	2403	5.20	3700	60	25	45
2	M1E	Scarborough	411	54764	8570	19.04	19855	315	540	815
3	M1G	Scarborough	137	53485	4345	12.31	18445	435	455	685
4	M1H	Scarborough	127	29960	4011	7.47	10765	615	220	255

Figure 4 – Toronto census dataframe

3. **Next step is merging these two data frames based on common key i.e., Postal Code.**

	PostalCode	Borough_x	Neighborhood	Latitude	Longitude	Borough_y	Neighbourhood Number	Population	Population density per square kilometre	Land area in square kilometres	Total - Household total income groups
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	Scarborough	263	90290	6208	45.74	26825
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Scarborough	134	12494	2403	5.20	3700
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	Scarborough	411	54764	8570	19.04	19855
3	M1G	Scarborough	Woburn	43.770992	-79.216917	Scarborough	137	53485	4345	12.31	18445
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	Scarborough	127	29960	4011	7.47	10765
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476	Scarborough	139	16724	5395	3.10	5920

Figure 5 – Merged Toronto Neighborhood & census data

b. Neighborhood Explorations:

1. First, we will find latitude and longitude of Toronto.

```
[ ] address = 'Toronto'
    geolocator = Nominatim()
    location = geolocator.geocode(address)
    latitude = location.latitude
    longitude = location.longitude
    print('The geograpical coordinate of Toronto are {}, {}.'.format(latitude, longitude))

    /usr/local/lib/python3.7/dist-packages/geopy/geocoders/osm.py:143: UserWarning: Using Nor
      UserWarning
    The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

Figure 6 – Toronto geo-coordinates

2. We will plot each neighborhood of Toronto on map.

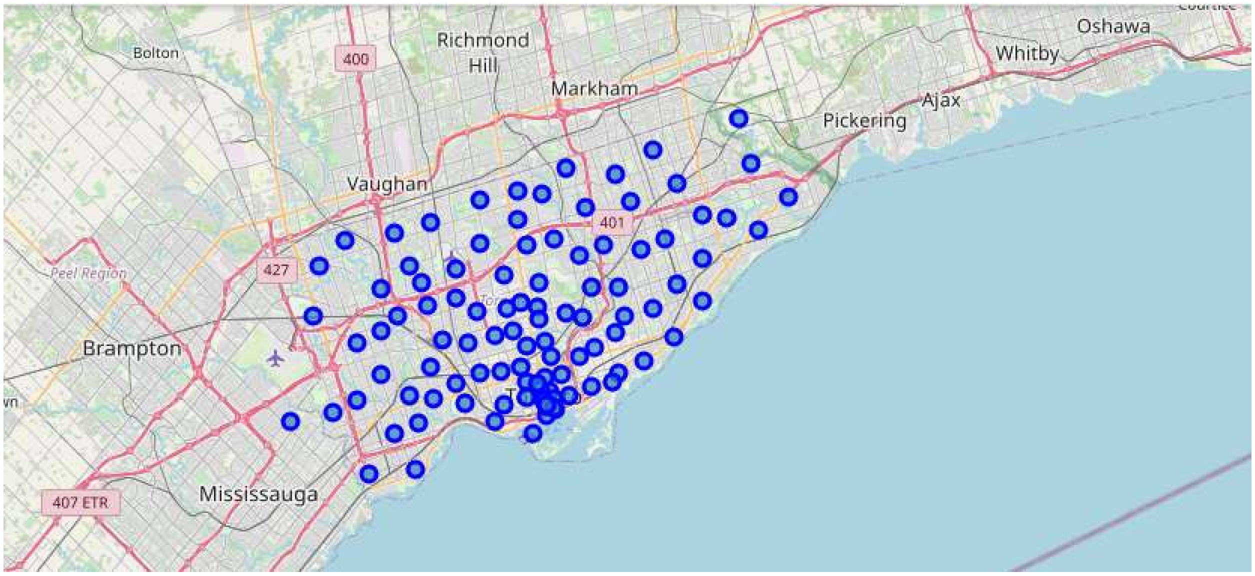

Figure 7 – Toronto Neighborhood on map

3. Next, we used the Foursquare API to explore each neighborhood and return the top 200 venues within 500 meters using longitude and latitude for each postal code.

```
print(toronto_venues.shape)
toronto_venues.head()
```

```
(2132, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Great Shine Window Cleaning	43.783145	-79.157431	Home Service
2	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store

Figure 8 – Neighborhood and venue

4. **Below gives snapshot of summary of above step.**
   It gives total number of different venue category in each of the neighborhood.

```
toronto_venues.groupby('Neighborhood').count()
```

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Adelaide, King, Richmond	98	98	98	98	98	98
Agincourt	5	5	5	5	5	5
Agincourt North, L'Amoreaux East, Milliken, Steeles East	4	4	4	4	4	4
Albion Gardens, Beaumond Heights, Humbergate, Jamestown, Mount Olive, Silverstone, South Steeles, Thistletown	9	9	9	9	9	9
Alderwood, Long Branch	8	8	8	8	8	8
Bathurst Manor, Downsview North, Wilson Heights	22	22	22	22	22	22
Bayview Village	4	4	4	4	4	4
Bedford Park, Lawrence Manor East	22	22	22	22	22	22

Figure 9 – Toronto Neighborhood vs number of venue category

5. **We have then one hot coded each venue category using get_dummies function.**

```
toronto_onehot.head()
```

Neighborhood	New American Restaurant	Nightclub	Noodle House	Office	Opera House	Optical Shop	Organic Grocery	Other Great Outdoors	Park	Performing Arts Venue	Pet Store	Pharmacy
Rouge, Malvern	0	0	0	0	0	0	0	0	0	0	0	0
Highland Creek, Rouge Hill, Port Union	0	0	0	0	0	0	0	0	0	0	0	0
Highland Creek, Rouge Hill, Port Union	0	0	0	0	0	0	0	0	0	0	0	0
Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	0	0	0
Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	0	0	0

Figure 10 – Toronto Neighborhood venues – one hot encoded

c. **Machine Learning:**

We this resulting data, now, we will try to form clusters and identify which cluster can be residential zone.

1. **First, we have used elbow method to find optimum number of clusters.**

```
1 2 3 4 5 6 7 8 9 Text(0.5, 1.0, 'Elbow Method For Optimal k')
```

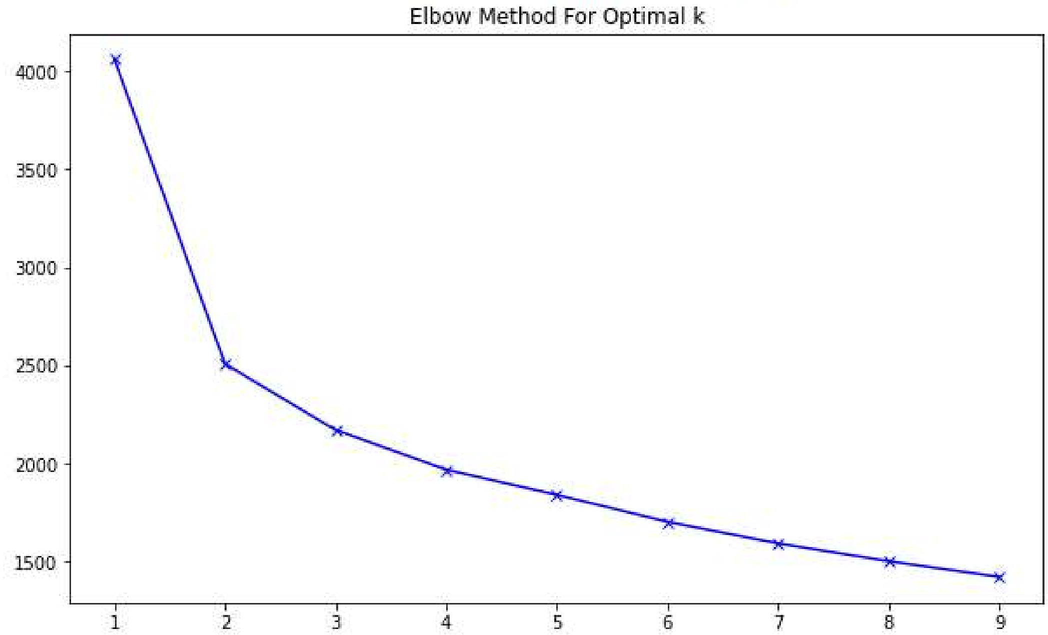

Figure 11 – Elbow method graph

Based on this graph, we will take 4 as optimum number of clusters.

2. We have used K-Means cluster to segment Toronto neighborhood.

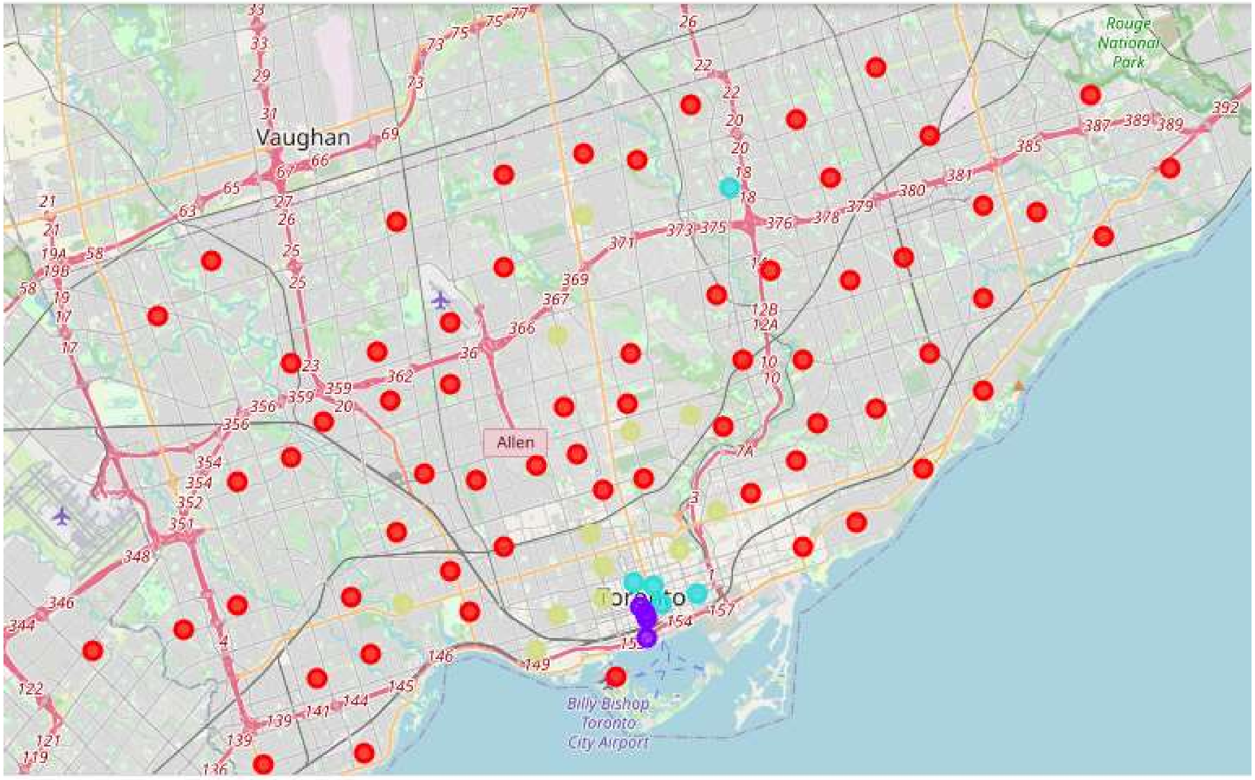

Figure 12 – Toronto Neighborhood clustered

Analysis:

Next, we will have to identify and categorize clusters:

As seen in Cluster 0, we are finding most common venues as Bank, Gym, Home Services & Park.

In cluster #1, we see most popular venues are Café and coffee shops. So, we can say it falls in re-creational zone.

Thus, we are treating cluster #0 as Residential cluster and henceforth we will only analyze more on cluster # 0.

d. Census data analysis:

1. First, let us put the data that is available.

	PostalCode	Borough	Neighbourhood Number	Population	Population density per square kilometre	Land area in square kilometres	Total - Household total income groups	Under $5,000	$5,000 to $9,999	$10,000 to $14,999	$15,000 to $19,999	$20,000 to $24,999	$25,000 to $29,999	$30,000 to $34,999
0	M1B	Scarborough	263	90290	6208	45.74	26825	290	240	420	720	730	925	955
1	M1C	Scarborough	134	12494	2403	5.20	3700	60	25	45	60	70	80	90
2	M1E	Scarborough	411	54764	8570	19.04	19855	315	540	815	970	880	890	905
3	M1G	Scarborough	137	53485	4345	12.31	18445	435	455	685	1170	825	960	910
4	M1H	Scarborough	127	29960	4011	7.47	10765	615	220	255	450	370	475	465
5	M1J	Scarborough	139	16724	5395	3.10	5920	105	180	305	330	325	345	370

Figure 12 – Toronto census with different income group

So, we have:

a. Population of neighborhood
b. Total number of households
c. Number of households falling in different income group like < 5 K, 5K – 10 K.

2. Our strategy:

2.1 To find average income, we have sort of tried to take weighted average. Hence, first we will take average of each income group and they multiply the number of households in that income group. Finally, we will add the values and then divide the same by total number of households for that neighborhood.

```
[ ] df_data["Avg_Income"] = ""
```

```
df_data["Avg_Income"]=(5000*df_data["GRP1"]+7500*df_data["GRP2"]+12500*df_data["GRP3"]+17500*df_data["GRP4"]+22500*df_data["GRP5"]+27500*df_data["GRP6"]+32500*df
 75000*df_data["GRP13"]+85000*df_data["GRP14"]+95000*df_data["GRP15"]+112500*df_data["GRP17"]+137500*df_data["GRP18"]+\
 175000*df_data["GRP19"]+200000*df_data["GRP20"])/df_data['Total - Household total income groups']
```

2.2 To find number of markets per neighborhood, we have tried to add all the venues like Supermarket, grocery store, Departmental store.

2.3 Finally, we tried to find number of people per market by dividing population of neighborhood by number of markets. We must handle neighborhood with 0 markets.

2.4 Final data frame:

```
final
```

	Borough_x	Neighborhood	Latitude	Longitude	Total_Markets	Avg_Income	Population
0	Scarborough	Rouge, Malvern	43.806686	-79.194353	0	86923.112768	90290
1	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	0	107307.432432	12494
2	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	0	76125.031478	54764
3	Scarborough	Woburn	43.770992	-79.216917	0	67249.254541	53485
4	Scarborough	Cedarbrae	43.773136	-79.239476	0	71081.049698	29960
5	Scarborough	Scarborough Village	43.744734	-79.239476	0	63167.229730	16724
6	Scarborough	East Birchmount Park, Ionview, Kennedy Park	43.727929	-79.262029	2	63864.956438	13641
7	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577	0	65490.384615	56512

Figure 14 – Toronto Neighborhood for residential zone with average income group and population

Results and discussion:

In the above table reflects how different neighborhood are placed with respect to number of markets, population, and average income. So, criterion for selection is more people per market and sufficient purchasing power. The general research shows that per capita income in Toronto is around 65 K USD. So, neighborhood hovering around 80 K - 100 K USD mark and higher population per market are our pointers towards identification of suitable location.

We can clearly see that neighborhood like "Rouge, Malvern" and "Guildwood, Morningside, West Hill" are very suitable as there are no markets and population is quite high with decent purchasing power.

It would be furthermore interesting to see, what are other shopping avenues are available in the all the neighborhoods of entire Toronto as in general we found lesser number of markets.

Conclusion:

In this project, the neighborhoods of Toronto were analyzed to find optimal location for Supermarket. We used:

Foursquare API to get location details like venues and venue category in each neighborhood.

K-means Clustering Machine learning algorithm to find different clusters.

Based on venue categories we tried to identify residential zone as it has maximum customer base for Supermarkets.

We also analyzed population data, income distribution and existing market penetration to find most suitable location.

There is always scope of improvement in any project. Some important pointers here could be more accurate classification of cluster to identify residential zones. Or in other words, there were not very suitable venue categories to pinpoints residential zones.

Although this project focuses particularly for supermarket, it could easily be extended for other cities and avenues like Restaurant, coffee shops etc.