Programmable toy store in Toronto

IBM Data Science Capstone Project

By Jonathan Wang

Introduction:

A client who producing and selling programmable educational robot would like to set up a store in Toronto.

To choose the location, they want to find an area which meet following requirements:

1. Number of customers aged 3-22
2. The willingness of customers to buy the toy (with their income)
3. Safety
4. Less competition

Background:

Nowadays the training of children's intelligence and logic becomes more and more important, but as a child, it is very hard to teach them to understand the difficult math and logic problems. How to make it easy for young boys and girls to become happy with logic training is a new problem of the era.

The programmable toys are designed for children and youth ages between 3 – 22 years old, it is absolutely a toy, and programmable/logic training robots as well which can make child 'Learn from Try'.

Data description

1. Neighborhood Profile of Toronto from https://open.toronto.ca/ which is recorded from Census.
   - Used to obtain population information.
   - Used to calculate children and youth aged 3 – 22 years old.
   - Used to calculate households' income.
   - Used to calculate potential consumption in the neighborhoods.
2. Neighborhood Crime Rates from http://data.torontopolice.on.ca/ which have statistics of crime data.
   - Contain different types of crime data, which can be calculated how safety the neighborhood is.

3. Neighborhood and Postal Code information from Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
   - Used to obtain the neighborhood names and postal code for target area.
4. Geography information which downloaded as Geo_Coordinates_Toronto.csv
   - Used to get the location information for particular neighborhoods.
5. Foursquare API:
   - Used to get the surroundings of target neighborhoods, to get number of toy stores around.
   - Note: in this project, Toys and Games are target information, it's categoryid = 4bf58dd8d48988d1f3941735 .

Methodology

Section 1

After obtain data source (Figure 1), data cleaning was done to the data-source and transposed to the data frame which contains population, age and income. (Figure 2).

```
[9]: TorontoProfileDFtemp.head(5)
```

[9]:

Neighborhood	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	Banbury-Don Mills	Bathurst Manor	Bay Street Corridor	Bayview Village	Bayview Woods-Steeles
Population, 2016	29,113	23,757	12,054	30,526	27,695	15,873	25,797	21,396	13,154
Children (0-14 years)	3,840	3,075	1,760	2,360	3,605	2,325	1,695	2,415	1,515
Youth (15-24 years)	3,705	3,360	1,235	3,750	2,730	1,940	6,860	2,505	1,635
Average after-tax income of households in 2015 ($)	427,037	278,390	168,602	792,507	493,486	251,583	352,218	354,894	253,036

4 rows × 140 columns

Figure 1

[12]:

Neighborhood	Population	Children	Youth	Income
Agincourt North	29,113	3,840	3,705	427,037
Agincourt South-Malvern West	23,757	3,075	3,360	278,390
Alderwood	12,054	1,760	1,235	168,602
Annex	30,526	2,360	3,750	792,507
Banbury-Don Mills	27,695	3,605	2,730	493,486

[13]:

```
TorontoProfileDFTrans['Population'] = TorontoProfileDFTrans
TorontoProfileDFTrans['Children'] = TorontoProfileDFTrans['
TorontoProfileDFTrans['Youth'] = TorontoProfileDFTrans['You
TorontoProfileDFTrans['Income'] = TorontoProfileDFTrans['In

TorontoProfileDFTrans.head()
```

[13]:

Neighborhood	Population	Children	Youth	Income
Agincourt North	29113	3840	3705	427037.0
Agincourt South-Malvern West	23757	3075	3360	278390.0
Alderwood	12054	1760	1235	168602.0
Annex	30526	2360	3750	792507.0
Banbury-Don Mills	27695	3605	2730	493486.0

Figure 2

Then Children and Youth of the population are added as Potential Consumer in Figure 3.

[16]:

Neighborhood	Population	Income	Potential Consumer
Woburn	53485	629030.0	7110
Rouge	46496	729154.0	5990
Malvern	43794	533202.0	5941
Mount Olive-Silverstone-Jamestown	32954	360648.0	5117
Willowdale East	50434	572155.0	5042

Figure 3

With the bar plot of potential consumer (Figure 4), we found that some neighborhood has much more young people than others. The top 20 neighborhoods with potential consumer were chosen to calculate their consumption capacity (Figure 5).

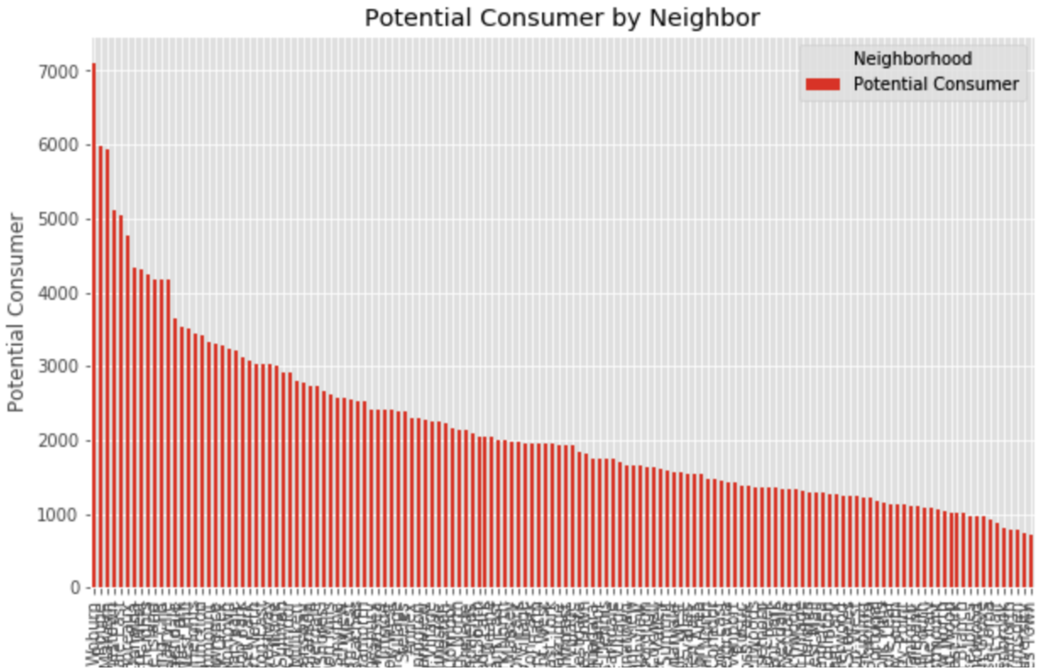

Figure 4

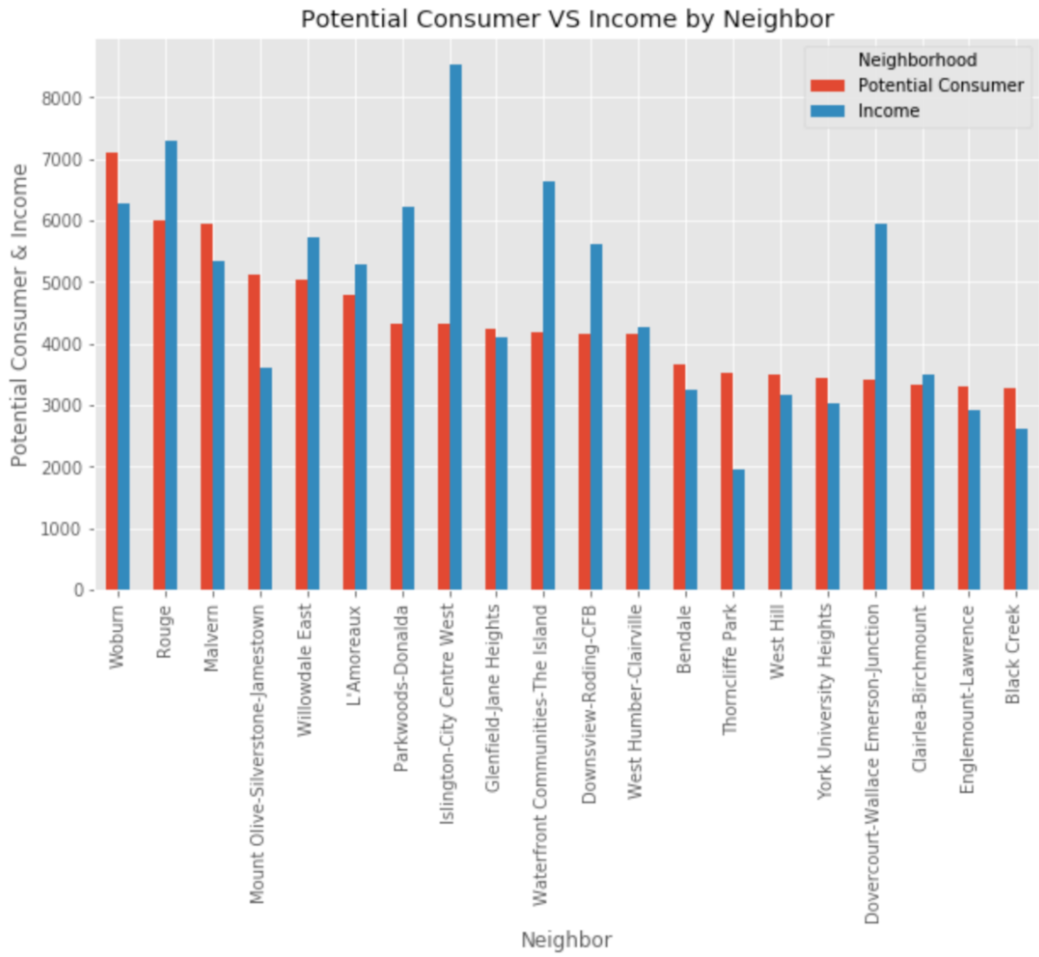

Figure 5

Section 2

With the downloaded data Toronto_Crime_Rates (Figure 6), the number of total crime by average is calculated as Figure 7.

[293]:

	OBJECTID	Neighbourhood	Hood_ID	Population	Assault_2014	Assault_2015	Assault_2016	Assault_2017	Assault_2018	Assault_2019	Assault_AVG	Assault_CHG	Assa
0	1	Yonge-St.Clair	97	12528	20	29	39	27	34	37	31.0	0.09	
1	2	York University Heights	27	27593	271	296	361	344	357	370	333.2	0.04	
2	3	Lansing-Westgate	38	16164	44	80	68	85	75	72	70.7	-0.04	
3	4	Yorkdale-Glen Park	31	14804	106	136	174	161	175	209	160.2	0.19	
4	5	Stonegate-Queensway	16	25051	88	71	76	95	87	82	83.2	-0.06	

5 rows × 60 columns

Figure 6

[125]:

	Neighborhood	Population	Total Crime
0	Yonge-St.Clair	12528	68.6
1	York University Heights	27593	665.6
2	Lansing-Westgate	16164	156.6
3	Yorkdale-Glen Park	14804	334.2
4	Stonegate-Queensway	25051	191.4

Figure 7

Then the 2 data frame above are merged into 1 data frame with Potential Consumption VS Total Crime as Figure 8 and plotted as Figure 9 and Figure 10 with top 50 neighborhoods who has highest ratio of Consumption vs Crime.

	Potential Consumption	Total Crime
Agincourt North	4479.398207	193.1
Agincourt South-Malvern West	1598.445279	275.1
Alderwood	287.963989	91.0
Annex	11732.211082	486.6
Banbury-Don Mills	5172.725332	200.8

Figure 8

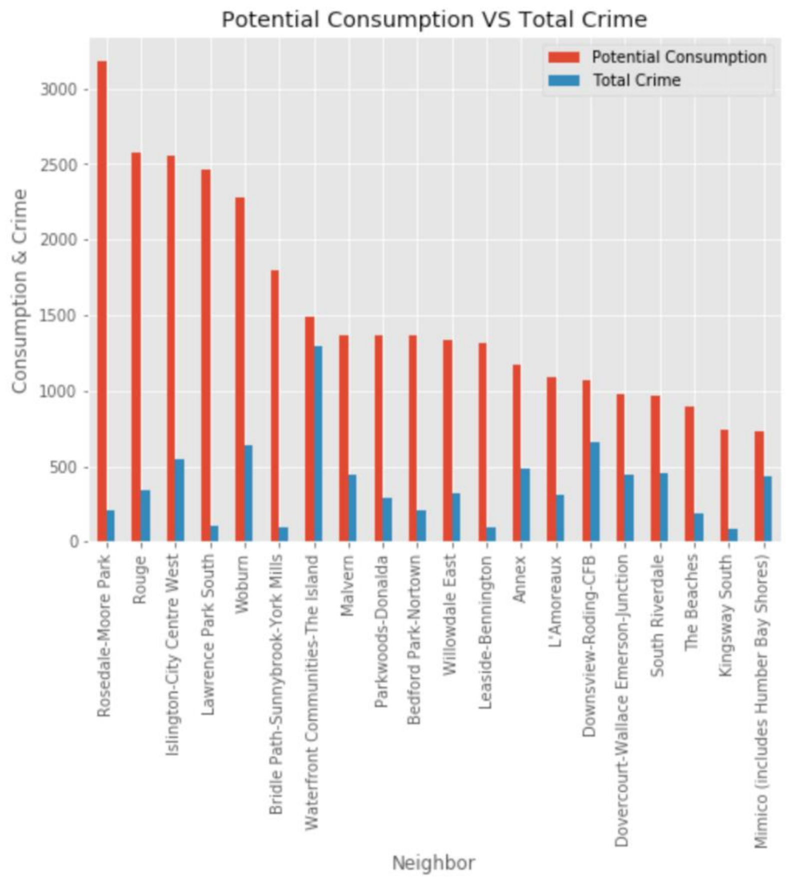

Figure 9

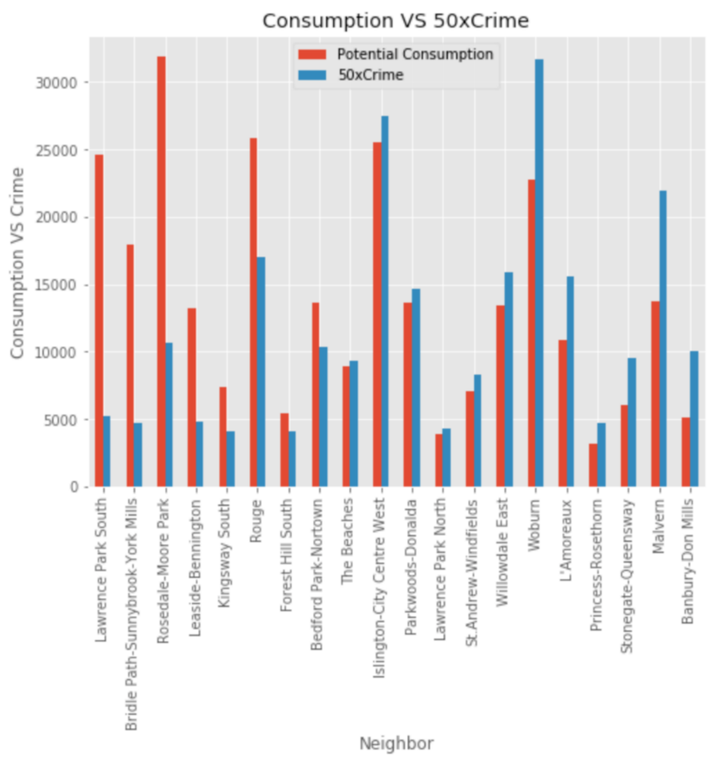

Figure 10

Then, we mean normalization to make the independent variables has reasonable effect on the output and did a Clustering with K-means algorithm to generate 4 clusters of these neighborhoods as shown in Figure 11 and Figure 12. Obviously the cluster 0 with blue dots in Figure 12 are much better to invest for they have higher consumption ability and low crime rate.

	Cluster Labels	Potential Consumption	Total Crime
Lawrence Park South	0	0.660244	-0.223354
Bridle Path-Sunnybrook-York Mills	0	0.452358	-0.241169
Rosedale-Moore Park	0	0.887812	-0.052615
Leaside-Bennington	2	0.301452	-0.235021
Kingsway South	2	0.118973	-0.258038

Figure 11

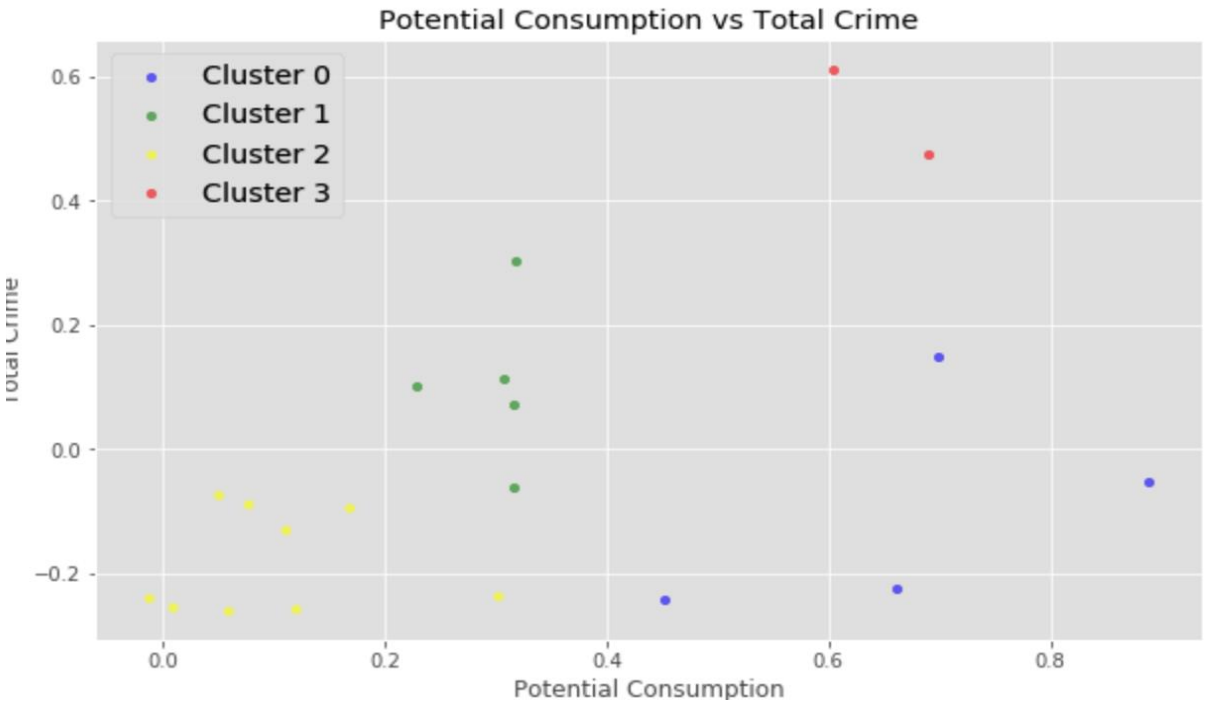

Figure 12

Section 3

A Toronto neighborhood data frame is generated by merge the neighborhood information and geography information, we select the neighborhoods in Cluster 0 we obtained in Section 2 and get below table (Figure 13) and make it visualized by Folium map.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
44	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
48	M4T	Central Toronto	Moore Park	43.689574	-79.383160
50	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529
0	M1B	Scarborough	Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill	43.784535	-79.160497
16	M1X	Scarborough	Upper Rouge	43.836125	-79.205636
20	M2L	North York	York Mills	43.757490	-79.374714
23	M2P	North York	York Mills West	43.752758	-79.400049

Figure 13

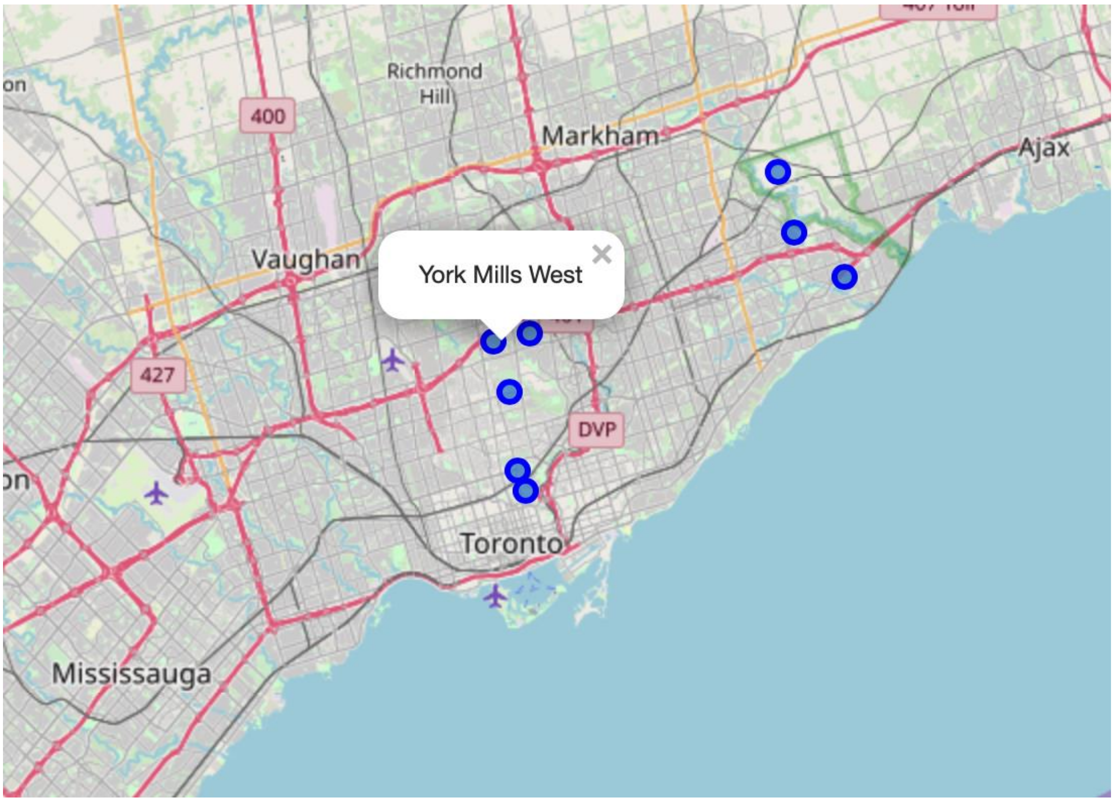

Figure 14

After Request venues from Foursquare API, we get following information as Figure 15. A data cleaning is performed and result as Figure 16, which shows the toy / game stores near our target neighborhoods.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lawrence Park	43.72802	-79.38879	Lawrence Park Ravine	43.726963	-79.394382	Park
1	Lawrence Park	43.72802	-79.38879	Sherwood Park	43.716551	-79.387776	Park
2	Lawrence Park	43.72802	-79.38879	T-buds	43.731247	-79.403640	Tea Room
3	Lawrence Park	43.72802	-79.38879	Sheridan Nurseries	43.719005	-79.400500	Flower Shop
4	Lawrence Park	43.72802	-79.38879	Bobbette & Belle	43.731339	-79.403769	Bakery

Figure 15

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
36	Lawrence Park	43.728020	-79.388790	Mastermind Toys	43.732046	-79.404141	Toy / Game Store
627	York Mills	43.757490	-79.374714	Mastermind Toys	43.769062	-79.386584	Toy / Game Store
629	York Mills	43.757490	-79.374714	The LEGO Store	43.778207	-79.343483	Toy / Game Store
733	York Mills West	43.752758	-79.400049	Mastermind Toys	43.732046	-79.404141	Toy / Game Store
754	York Mills West	43.752758	-79.400049	Mastermind Toys	43.769062	-79.386584	Toy / Game Store

Figure 16

Finally, all our final data are draw onto the map:

The blue circles are areas with high children and youth population, relatively high income and low crime rate, which we say with high potential consumption ability.

The red labels are 3 existing toy stores nearby. (there are 5 in the list but 2 of them are duplicated).

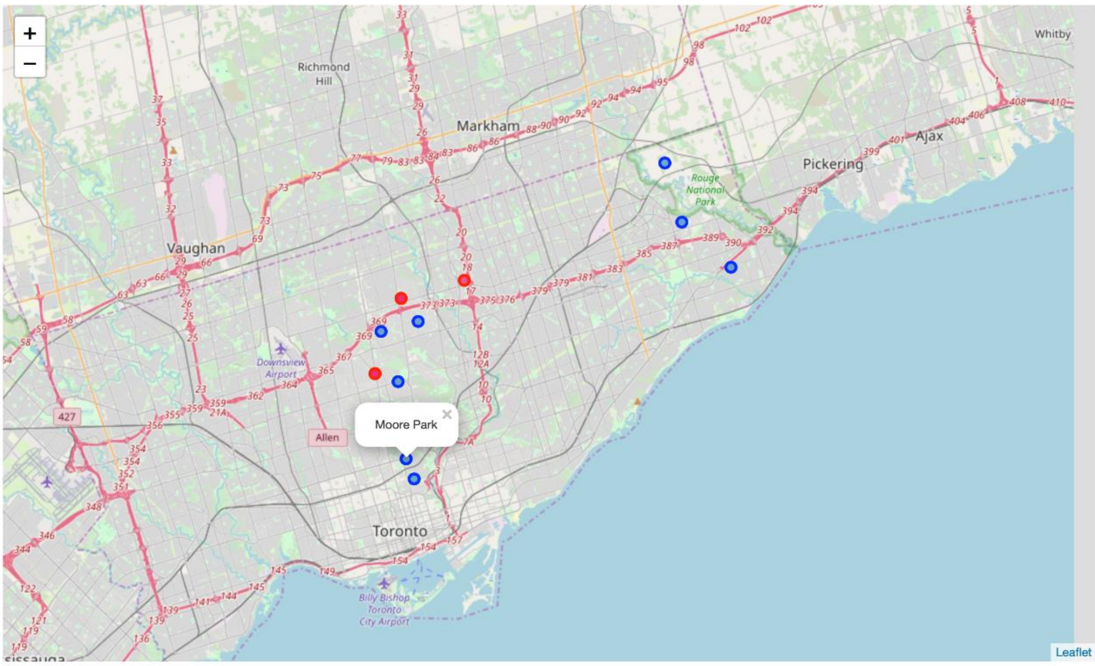

Figure 17

Results

If the client want to open their store with low competition, Moore/Rosedale/Rouge should be chosen while if they would like to fight, Lawrence and York will be good place as well.

Discussion

A lot of data is dealt with in the project, we used statistics to calculate population and consumption, we did mean normalization to make data has reasonable weight, and did K-means clustering to segment the neighborhoods. Also, we did use a lot of visualization plots to show the results more efficiently.

But there are still lot to improve:

1. If we can search more data on Foursquare, what will be the result?
2. What is the trend of children and youth in these neighborhoods? Saying if we dig deep into working age (25 – 35 years old maybe), those neighborhoods might have more babies in the recent future.

Conclusion

Although there could be more possible locations to set up the store, it's still fine to pick up location in the neighborhoods above. Not only toy stores, but also all kinds of children related shops could be established there. People could obtain a lot of information through the map services. This project gives an example of process data and chose location in a large range of area.

References

1. City of Toronto Open Data Portal
2. Toronto Police Service Public Safety Data Portal
3. Neighborhood and Postal Code information from Wikipedia
4. Geography information which downloaded as Geo_Coordinates_Toronto.csv
5. Foursquare API

Acknowledgement

I would like to express my gratitude Coursera and IBM Data Science Course for providing this course for me to learn the skills I need. Thanks

March 2020