Neighbourhood Wars: School and Crime Cohabitation in Toronto

By Nicholas Duke

28/07/2020

1. Introduction

1.1 Background

A group of investors intend to establish a new private middle school in Toronto Canada, and require data science insights to determine the optimal location. Opening a school for students between grade six and nine requires variables such as transport efficiency, safety and extramural activities to be taken into account. The investors are concerned that the school's success is heavily influenced by its location. The diversity of the target audience (students & their parents residing in different neighbourhoods across Toronto) requires convenient transportation and safety. In this project, school safety is determined by the number of neighbourhood crime incidents in the area in which the school is located. Examples include assault, robbery and homicide. These are student welfare metrics, where exposure to such activities before and after class, may diminish their school perception or welfare. Hence, our target audience requires a minimum crime threshold level in the school's neighbourhood. The students will be at a higher risk of crime exposure when leaving the premises to take public transport.

Many students are solely dependent on public transport such as bus lines for transport to and from school. This group of students will be more exposed to crime, thus by locating a school in a safe neighbourhood cluster, the chance of crime exposure is reduced. It is advantageous for a school to be conveniently located near to one of the public transport facilities as it will reduce commute time, transport fairs and crime risk for a large portion of students who cannot rely on their parents to take them to school. If a student has an extended commute back home, a school located near a convenience store will allow for easy access to food before the long journey.

Analysing this data will allow the stakeholders and investors to make actionable insights to locate the new private school establishment.

1.2 Problem

Data that contributes to predicting the optimal location to establish a new private middle school, includes neighbourhood crime rates in Toronto and coordinate information on the most popular venues in each of Toronto's boroughs. This project aims to narrow down the numerous possible geographic locations within Toronto in which to open a new school based on the neighbourhood's proximity to favourable venues.

1.3 Interest

The business owners, investors and other stakeholders establishing a new private school of their own will be interested to gain a holistic picture of the locational factors that will influence such venture's success. Students and parents will also be interested as it will provide piece of mind with respect to their choice of school in comparison to others schools, based on the safety and venue data.

2. Data Acquisition

2.1 Data Sources

The business problem identified the variables which will influence the new school's optimal location, namely:

1) safety levels of the nine Toronto boroughs extracted from the Toronto Crime data set, and the safest three boroughs will be shortlisted.

2) the neighbourhoods where public transport, parks and recreation are available will be shortlisted.

3) the neighbourhoods with a nearby food store will be shortlisted.

Section 1

Analysing the Toronto Police Service dataset which includes all valid assault, automobile theft, breaking and entering, homicide and robbery crimes reported to the Toronto Police Department between 2014 and 2019. That Crime data will be imported and converted into a panda's data frame. The next step will be scraping the Toronto Postal Code Wikipedia Page and extracting all the Toronto borough, postal code and neighbourhood data. Once the neighbourhood, borough and postal code information has been extracted, a new data frame will be created to merge the crime data with Toronto's neighbourhoods and boroughs. The Wikipedia neighbourhood, borough and postal code data is taken

from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:M:. The neighbourhood crime data was extracted from https://data.torontopolice.on.ca/.

Section 2

This section aims to retrieve the coordinate values matching Toronto's postal codes. This process is done by importing the latitude and longitude data from https://cocl.us/Geospatial_data. Once imported and converted into a panda's data frame, the crime data, Toronto postal code, neighbourhood, and coordinate data will be merged to create a data frame containing all the information required to determine the safest borough and neighbourhood in Toronto. Visual data analysis will be done to highlight these key locations. The Folium Library will be used to analyse the data in order to determine the safest borough as well as explore the neighbourhoods visually on a map. From this point, the three safest boroughs will be shortlisted.

Section 3

This is the Methodology section of the project. The first task will be creating a new dataset with the safest three boroughs, the neighbourhoods within the boroughs, as well as the most common venues within each neighbourhood and its coordinates. The venue and location data will be fetched using the Foursquare API.
https://foursquare.com/developers/apps/TVUZFSV3D2NTX3UH0R4UN4XPDMENTVP5YEN4CGZX2HPDNTYA/settings.

Section 4

The results and discussion section will evaluate the previous findings and critically evaluate the methods used and the results obtained. The target audience (students and parents searching for schools) problem of identifying the safest borough and neighbourhood nearest to public transport and recreational venues will be answered. Toronto student school commute data is taken from
https://smartcommute.ca/wp-content/uploads/2016/02/School_Travel_Trends_GTHA_En.pdf.

2.2 Data Manipulation

Before any data analysis could be done, the Toronto crime and neighbourhood data had to be cleaned and merged to create various coherent data frames. The following, outlines all of the Toronto crime and locational data points extracted throughout the project:

- The Toronto Crime data was taken from https://data.torontopolice.on.ca/, and uploaded into the Jupyter directory as Neighbourhood_Crime_Rates_(Boundary_File)_.csv. The dataset includes crimes committed between 2014 and 2019, however, due to the size of the dataset, only 2019 values were used. The new crime data frame columns include:
    - Toronto postal codes
    - 2019 assault incidents
    - 2019 automobile theft incidents
    - 2019 break and entering incidents
    - 2019 homicide incidents
    - 2019 robbery incidents
    - 2019 theft incidents

```
new_crime.head()
```

	Postal Code	Assault_2019	AutoTheft_2019	BreakandEnter_2019	Homicide_2019	Robbery_2019	TheftOver_2019
0	M5C	37	6	28	0	4	6
1	M3J	370	144	108	0	79	28
2	M3A	72	32	39	0	11	11
3	M4A	209	61	84	1	42	29
4	M8Y	82	34	64	0	22	4

Figure 1: Toronto crime data with postal codes and incidents

- The crime data only had the postal code for the locational requirements, therefore the neighbourhood and borough values were scraped from Wikipedia https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. The Beautiful Soup package was used to scrape the webpage. However, there were some "Not Assigned" values for both boroughs and neighbourhoods. These values were discarded. In some cases, multiple neighbourhoods were assigned to the same postal code and borough. As Seen in Figure 2 below.

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 2: Pre-processed and cleaned Wikipedia data of Toronto boroughs and neighbourhoods converted into a Pandas data frame

- In order to map the location of the Toronto crimes, coordinate values needed to be linked with the postal code, borough and neighbourhood data. The first step was to extract latitude and longitude data values of Toronto's neighbourhoods and boroughs from https://cocl.us/Geospatial_data. See Figure 3 below.

```
coordinates = pd.read_csv('https://cocl.us/Geospatial_data')
coordinates.head()
```

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Figure 3: Toronto neighbourhoods, borough and postal code latitude and longitude coordinates

- The next step was to merge Toronto's coordinate and neighbourhood data

	Postal Code	Borough	Neighbourhood	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	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Figure 4: Merged Toronto neighbourhood, borough, postal code and coordinate data

- A complete data frame was created with the tables in figure 3 and 4.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Assault_2019	AutoTheft_2019	BreakandEnter_2019	Homicide_2019	Robbery_2019	TheftOver_2019	Total
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	563	177	132	1	79	26	978
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	58	27	33	1	9	1	129
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	642	56	105	4	98	9	914
3	M1G	Scarborough	Woburn	43.770992	-79.216917	427	60	100	0	67	8	662
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	266	55	52	1	40	9	423

Figure 5: Complete data frame with crimes, and neighbourhood coordinates

- The table in figure 5 was merged with data extracted from the Foursquare API, which identifies the local venues within a 500-metre radius. See figure 6 below:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Caledonia-Fairbanks	4	4	4	4	4	4
Davisville	35	35	35	35	35	35
Davisville North	7	7	7	7	7	7
Del Ray, Mount Dennis, Keelsdale and Silverthorn	4	4	4	4	4	4

Figure 6: Venue category and neighbourhood coordinates pulled from the Foursquare API

3. Methodology

3.1 Exploratory Data Analysis

The key in channelling the focus areas of the project, is to filter out the unsafe neighbourhoods. Figure 7's bar chart of the crimes committed per borough illustrates the most unsafe and safest Toronto boroughs from the crime dataset. To proceed, the three safest boroughs were shortlisted. The chart shows that East York was the safest borough, whilst Scarborough was the most unsafe in Toronto 2019. The safest three boroughs were East York, East Toronto and York. However, I had to remove East Toronto as it had the lowest number of neighbourhoods, which would limit the potential number of nearby venue locations surrounding the school.

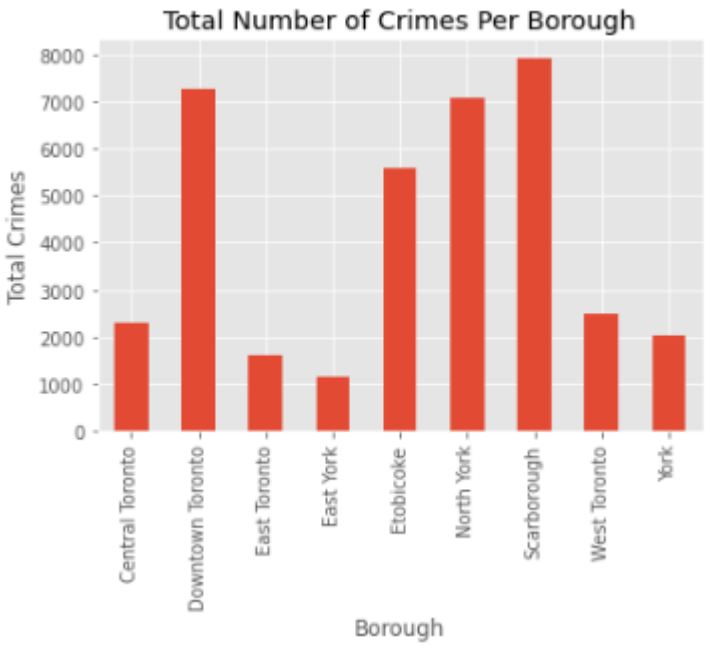
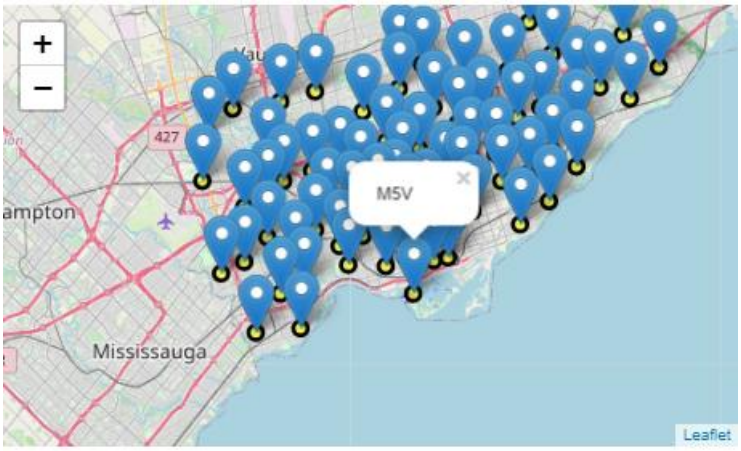


Figure 7: Number of crimes per borough in Toronto

A visual representation of the crimes committed in the neighbourhoods was created by superimposing the location of each crime using markers on a map with the help of the **folium** library. The code below looped through the crime incidents to add markers and coordinates to the map.

```
# instantiate a feature group for the incidents in the dataframe
incidents = folium.map.FeatureGroup()

# loop through the incidents
for lat, lng, in zip(complete_data2.Latitude, complete_data2.Longitude):
    incidents.add_child(
        folium.features.CircleMarker(
            [lat, lng],
            radius=5, # define how big you want the circle markers to be
            color='black',
            fill=True,
            fill_color='yellow',
            fill_opacity=0.6
        )
    )
# add pop-up text to each marker on the map
latitudes = list(complete_data2.Latitude)
longitudes = list(complete_data2.Longitude)
labels = list(complete_data2.PostalCode)

for lat, lng, label in zip(latitudes, longitudes, labels):
    folium.Marker([lat, lng], popup=label).add_to(toronto_map)



# adding incidents to map
toronto_map.add_child(incidents)
```

This Toronto crime map, clearly identifies each crime location superimposed onto the map with markers. The markers are displaying the postal codes of the crime locations as pop-up text.



The next map created illustrates the neighbourhoods of the three safest boroughs that were shortlisted. Each borough is in a different colour.

Foursquare API data was then used to find the nearest venues within a 500-metre radius of the neighbourhood coordinates. Such venues include seafood restaurants, pools, athletics and sports, gyms and cafes etc. For the purposes of this study, the most important venue categories were identified as bus lines, coffee shops, light rail stations, parks, swim schools and convenience stores. For public transport, there was a total of one bus line and one light rail station. Unfortunately, there were no bus stops or metro stations in the three shortlisted boroughs. The Foursquare API venue data was then merged with the neighbourhood data producing a data frame highlighting the nearest venues in each of the selected neighbourhoods.

3.2 Machine Learning

One Hot Encoding was necessary to convert categorical data into numerical data as a form of pre-processing to allow machine learning algorithms to work effectively. The venues in each neighbourhood were transformed into the frequency of the number of venues located in each neighbourhood. In other words, the mean frequency of the category occurrence was displayed. The top eight and ten venues in each neighbourhood were converted into a panda's data frame.

	Neighborhood	American Restaurant	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	Bank	Bar	Beer Store	Bike Shop	...	Swim School	Tennis Court	Thai Restaurant
0	Caledonia-Fairbanks	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	0.000000	0.000000
1	Davisville	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	0.028571	0.028571
2	Davisville North	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	0.000000	0.000000
3	Del Ray, Mount Dennis, Keelsdale and Silverthorn	0.0	0.0	0.0	0.0	0.0	0.0	0.25	0.0	0.0	...	0.0	0.000000	0.000000
4	East Toronto, Broadview North (Old East York)	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	0.000000	0.000000

Figure 8: Transforming categorical data into numerical data with the One Hot encoding technique.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	Caledonia-Fairbanks	Park	Pool	Women's Store	Fish & Chips Shop	Diner	Discount Store	Donut Shop	Electronics Store
1	Davisville	Pizza Place	Dessert Shop	Sandwich Place	Italian Restaurant	Coffee Shop	Sushi Restaurant	Café	Gym
2	Davisville North	Gym / Fitness Center	Food & Drink Shop	Sandwich Place	Park	Breakfast Spot	Department Store	Hotel	Garden
3	Del Ray, Mount Dennis, Keelsdale and Silverthorn	Restaurant	Sandwich Place	Discount Store	Bar	Field	Dessert Shop	Diner	Donut Shop
4	East Toronto, Broadview North (Old East York)	Convenience Store	Park	Intersection	Yoga Studio	Fish & Chips Shop	Discount Store	Donut Shop	Electronics Store

Figure 9: The top eight venues of each neighbourhood

A new merged “venues” data frame was created to display the relevant categories with the corresponding neighbourhood latitude and longitude values. This allows for a summarized view of the relevant data aiding an easier decision process regarding the optimal school location.

	Neighborhood	Bus Line	Coffee Shop	Light Rail Station	Park	Swim School	Convenience Store	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Caledonia-Fairbanks	0.0	0.000000	0.0	0.500000	0.0	0.0	5	43.689026	-79.453512	Nairn Park	43.690654	-79.456300	Park
0	Caledonia-Fairbanks	0.0	0.000000	0.0	0.500000	0.0	0.0	5	43.689026	-79.453512	Maximum Woman	43.690651	-79.456333	Women's Store
0	Caledonia-Fairbanks	0.0	0.000000	0.0	0.500000	0.0	0.0	5	43.689026	-79.453512	Fairbanks Pool	43.691959	-79.448922	Pool
0	Caledonia-Fairbanks	0.0	0.000000	0.0	0.500000	0.0	0.0	5	43.689026	-79.453512	Fairbank Memorial Park	43.692028	-79.448924	Park
1	Davisville	0.0	0.057143	0.0	0.028571	0.0	0.0	0	43.704324	-79.388790	Jules Cafe Patisserie	43.704138	-79.388413	Dessert Shop

Figure 10: Merged Venues data frame

K-Means is a method of vector quantization which aims to separate n observations into k clusters, in which each observation is situated within the cluster that is closest to the mean (cluster centroid), becoming a prototype for that specific cluster. For this project K-Means clustering enabled the neighbourhoods with a similar number of relevant venues (averaged) within each neighbourhood to be clustered together.

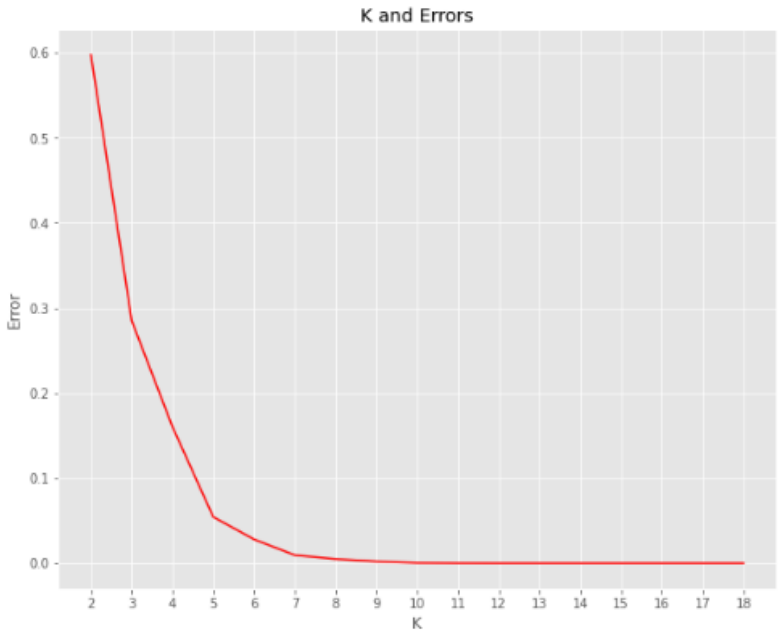


Figure 11.1: Many K cluster values and their errors

The Elbow Point technique identifies the optimal number of K clusters to be used that do not overfit or underfit the model. The best k value was chosen by comparing the accuracy of numerous k values against one another. The optimal K value was chosen at the point on the curve with the most radical kink. This point was at k = 6 clusters. The Yellowbrick package was used to fit the K-Means model to visualize the elbow point.

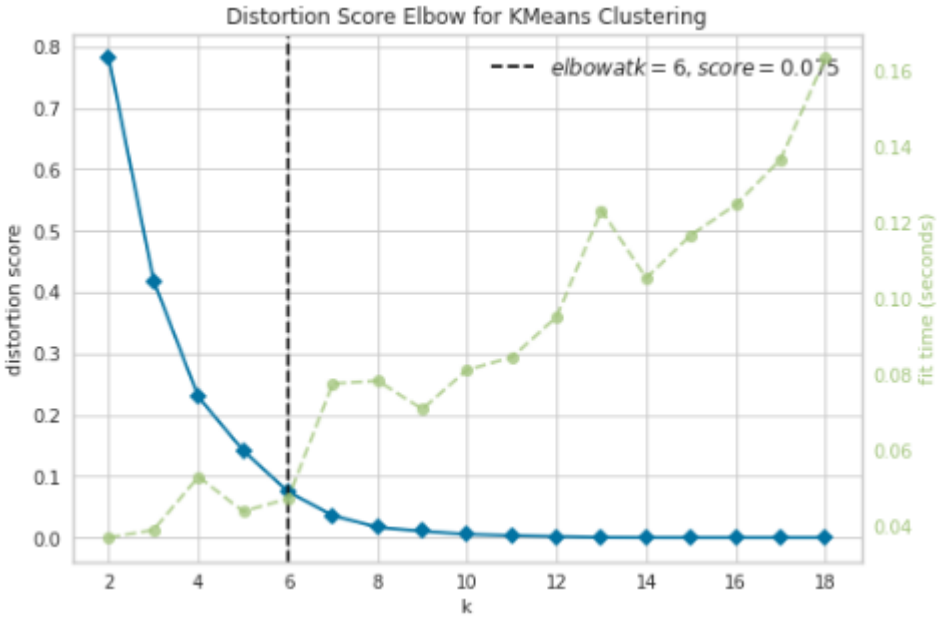


Figure 11.2: Distortion score for the optimal number of K clusters

The clusters were superimposed onto a Folium map of Toronto, and each of the six clusters was uniquely identified by a specific colour. Cluster 1 = red, cluster 2 = purple, cluster 3 = teal blue, cluster 4 = aquamarine blue, cluster 5 = light green and cluster 6 = dark orange.

It is quite common for schools to be situated in less urban and more residential neighbourhoods, which may allow parents to avoid traffic when dropping their children off at school. Residential areas offer larger plots of land for a new school to incorporate sports fields and class rooms. The bar chart below illustrates the number of neighbourhoods per cluster. As we can see the

1st cluster has significantly more neighbourhoods with 11 compared to the other five ranging between one and four.

The underlying trade-off is whether a school should be situated closer to the city centre with more neighbourhoods, or more towards the residential outskirts. The trade-off will have to take into account the convenience of school students taking the rail and being situated in the city. Alternatively, the school could be located on the outer neighbourhood ring of the city centre with access to popular bus routes.

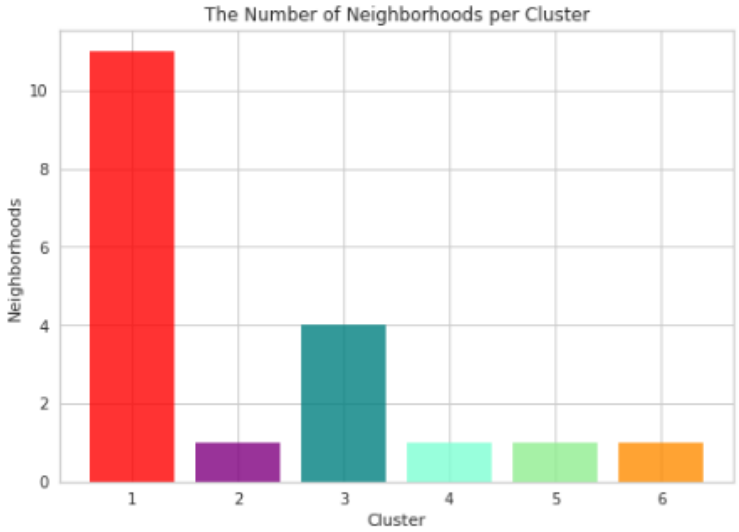


Figure 12: The number of neighbourhoods per cluster

3.3 Results: Examining Each Cluster

Analysing each cluster highlights the discriminating venue categories that distinguish each cluster. Based on these defining categories, the clusters will be assigned six unique colours. Each cluster is displayed in a panda's data frame containing borough, neighbourhood and cluster labels and venues. Figure 13 is an example of a single cluster data frame.

```
[81]: c1=merged_3.loc[merged_3['Cluster Labels'] == 0]
df_c1 = pd.merge(df_cluster, c1, on='Neighborhood')
df_c1.tail(16)
```

[81]:

	Borough	Neighborhood	Bus Line	Coffee Shop	Light Rail Station	Park	Swim School	Convenience Store	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
154	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.0	0.117647	0.058824	0.0	0.0	0.0	0	43.686412	-79.400049	LCBO	43.686991	-79.399238	Liquor Store
155	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.0	0.117647	0.058824	0.0	0.0	0.0	0	43.686412	-79.400049	The Market By Longo's	43.686711	-79.399536	Supermarket
156	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.0	0.117647	0.058824	0.0	0.0	0.0	0	43.686412	-79.400049	Daeco Sushi	43.687838	-79.395652	Sushi Restaurant
157	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.0	0.117647	0.058824	0.0	0.0	0.0	0	43.686412	-79.400049	Union Social Eatery	43.687895	-79.394916	American Restaurant
158	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.0	0.117647	0.058824	0.0	0.0	0.0	0	43.686412	-79.400049	Mary Be Kitchen	43.687708	-79.395062	Restaurant
159	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	0.0	0.117647	0.058824	0.0	0.0	0.0	0	43.686412	-79.400049	Kiva's	43.687984	-79.394715	Bagel Shop

Figure 13: Cluster one example of the merged venue values and coordinates

The ideal school location requires accepting the opportunity costs associated with the varying unique characteristics of each cluster. To answer this, the mean number of venues of each cluster will be critically evaluated.

- Coffee Shops
    - The social aspect of a coffee shop near a school is beneficial for student interaction. It may stimulate meaningful conversations and help form close friendships. Cluster one holds all the coffee venues in the data set with an average of 0.07. This is not surprising as cluster one also has the highest number of neighbourhoods with 11. Some of the neighbourhoods in cluster one include: Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park, The Annex, North Midtown, Yorkville, Runnymede, The Junction North, Roselawn and Thorncliffe Park.

```
0.07058823529411765
0.0
0.0
0.0
0.0
0.0
```

- Bus Lines
    - Bus lines are the routes over which buses regularly travel. This is arguably the most important venue impacting the lifestyles of many students. Litman (2020:87), states that buses are flexible with their routes when compared with rail transit. In addition, the bus lines do not require any special facilities. Buses are more suitable for dispersed land use when compared with rail transit. Hence, they can serve a greater rider catchment area. According to School Transportation News (2009), around 55% of Canada's K-12 students rely on school buses. The mean number of bus lines per cluster shows that the 5th cluster contains all the bus routes at 0.333. This hints at the potential trade-off that will have to be made against the light rail station as a mode of transport. The 5th cluster (light green) includes Lawrence Park in Central Toronto.

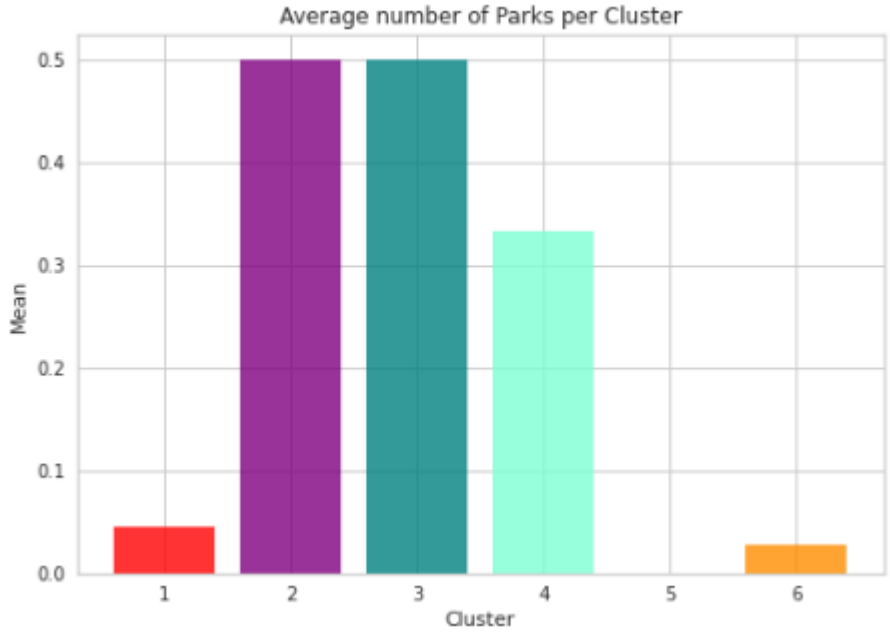
```
0.0
0.0
0.0
0.0
0.3333333333333333
0.0
```

- Light Rail Stations
    - Light rail stations are situated in the urban centres of Toronto. They relieve residents from congestion and time spent in their cars commuting to work and school. This system will save parents time, money as well as negate the risk of car accidents. Like coffee shops, all of the light rail stations are based in cluster one with 0.0058.

```
0.0058823529411764705
0.0
0.0
0.0
0.0
0.0
```

- Parks
    - Outdoor activities in time spent in nature is valuable for a student's development. With the ever-expanding increase in time spent watching screens and staying indoors, there is evidence to suggest that obesity and psychological problems are correlated to the decline in outdoor activities, suggests Louv (2005). Studies such as those done by Louv, justify the need to locate schools near a park, where the students are regularly exposed to nature. Parks were situated in all the clusters except for the 2nd. There is a relatively high concentration of parks in cluster five with 0.333. Other suitable neighbourhoods with parks include clusters three, four and six. Examples of neighbourhoods within these clusters include: Caledonia-Fairbanks, Broadview North, East Toronto, Davisville North, Woodbine Heights, Moore Park and Summerhill East.

```
0.023529411764705882
0.0
0.2
0.3333333333333333
0.3333333333333333
0.5
```



- Swim Schools
    - It would be an appealing addition, for middle school students to have access to extra swimming training. Swim schools are only situated in the 5th cluster with 0.333.

```
0.0
0.0
0.0
0.0
0.3333333333333333
0.0
```

- Convenience Stores
    - Convenience stores would be an added opportunity for students to grab a quick snack before and after class time. Convenience stores are located in clusters one, two and four. The Weston neighbourhood in the 2nd cluster being the most concentrated with 1.

```
0.005882352941176470 5
1.0
0.0
0.3333333333333333
0.0
0.0
```

3.4 Summarizing the Venues and Their Respective Clusters

- Transport Advantages
    - Light rail stations dominate cluster one, whilst bus lines dominate cluster five.

- Parks and Recreation
    - Coffee shops are situated in cluster one.
    - Parks are located in clusters one, three, four, five, and six, whilst the swim school is in cluster five.
- Convenience Stores
    - They are located in the 1st, 2nd and 4th clusters.

4. Discussion

This study aimed to identify the optimal location in which to position a private middle school in Toronto, Canada. Under the assumption that safety is one of the key elements in creating a successful middle school, the boroughs were shortlisted into the safest three. This included East York, East Toronto, and York. However, when analysing the number of neighbourhoods within each borough, it was noted that East Toronto only had two neighbourhoods. Therefore, the next best option would be to select York, as more neighbourhoods would cater for a larger catchment area of students. Furthermore, the low number of neighbourhoods in East Toronto may account for the low crime rate.

With the safest boroughs shortlisted, the next location selection step would be determined by the transport, parks and recreational venue categories. Given the availability of public transport in Toronto, a new middle school must be conveniently located near these services. With this in mind, cluster one and five are appealing, as they would reduce the cost, time and safety of many parents otherwise bound to the strenuous school commute. The transport services available in the safest three boroughs are light rail trains (cluster one) and school buses (cluster five). As expected, the rail system is located in the more densely populated urban neighbourhoods of Summerhill West and Rathnelly, whilst the bus lines cater for the more residential neighbourhoods, such as Lawrence park. Swim schools are conveniently located in cluster five with 0.33, which will aid in school extra mural sporting activities.

The majority of convenience stores are located in cluster two with a value of 1. Unfortunately, there is no convenience store locating in cluster five, and the stores in cluster one only makes up 0.05.

For the parks and recreation category, coffee shops are solely located in the 1st cluster. This ties up well with the light rail system as it is a typical beverage consumed whilst on this mode

of transport. Of the clusters available with parks, the fewest are located in cluster one with 0.02. Whereas, cluster five and six are more park friendly with 0.33 and 0.5. According to Statista (2016), fewer than 24% of children consume caffeine from coffee. Whilst this venue may be convenient for parents waiting to pick up and drop off their children from school, it does not cater for half of the target audience, who are the school children. In addition, it may only satisfy a small portion of learners commuting by train and small groups socializing after class. Thus, it can be removed as an influential venue.

Given all these factors it would be safe to say that Lawrence Park, cluster five (light green) would be the optimal location to establish a middle school. The availability of bus lines is vital in catering for a large percentage of the learner's transport needs. The psychological and recreational benefits of parks would enhance the positive impact which the school would have on the children. The convenience of the swimming school nearby would allow for extra-mural swimming activities catering for beginner to advanced swimmers between grades six and nine. Finally, zooming in on the cluster map, it is evident that there are numerous schools and colleges nearby which would allow for increased school sport rivalry, as well as social activities with the neighbouring schools.

A pitfall of the study is that it did not take into account the demographic representation of the families living in the safest three boroughs. For example, the average individual age of the three boroughs to determine the demand for grades six to nine was not studied.

5. Conclusion

In this study I analysed key geographic locations and variables which would influence the positioning of a new private middle school in Toronto. I extracted Toronto crime data and short listed the three safest boroughs. Python libraries such as matplotlib, folium and KMeans were used to import, manipulate, merge and visualize data sets. Once the crime and neighbourhood data were merged, it allowed the Foursquare API to provide coordinates of the most common venues for the three safest boroughs. The venue categories included transport, parks and recreation, and convenience stores.

Machine learning algorithms such as One Hot Encoding and K-Means clustering were used to identify the six clusters and the venue information in each. For reasons discussed earlier, cluster 5's Lawrence Park was the chosen location. Investors heading the new school project can now make credible investment decisions given this analysis on neighbourhood safety and venues in

Toronto. This study can be grown to include many more variables to enhance the credibility of the chosen location. Demographic indicators such as wage, is an example of an additional variable that can be included to gain a more thorough understanding of the target audience.