Neighbourhood Wars: School and Crime Cohabitation

(Week 1: Data)

By Nicholas Duke

28/07/2020

1. Data Acquisition

1.1 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.

1.2 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