# Introduction

### Background

Canada is known as the most immigrant welcoming and multicultural western country. In 2019, Canada admitted roughly 340,000 new permanent residents, of which 45% chose to settle around Toronto which is the biggest metropolis of Canada and continuously appraised as one of the topmost livable cities in the world.

### Problem

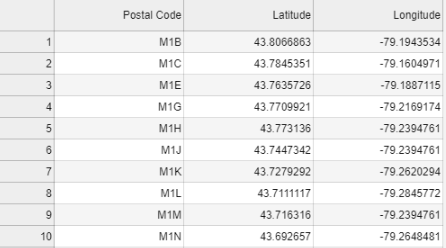
In 2020 November, Canada announced a 3-year open work permit and immigration pathway for Hong Kong residents. The policy has drawn strong attention of Hong Kong youths who are desperate for an immediate move. This report aims to study the living environment of Toronto neighbourhoods in 4 areas – house price, community safety, ethnic diversity and neighbourhood venues. Target audience of this report is Hong Kong residents who decided to make a hasty trip to Toronto but have not determined a suitable place to settle down. Particularly I will focus the study on the most populous urban centre - Toronto City. Toronto City is 630 square kilometres in size and of estimated 2.9M population. I will dissect the 140 neighbourhoods in Toronto City.

# Data Source

Below are the data sources for this study, with a fraction of the datasets being shown.

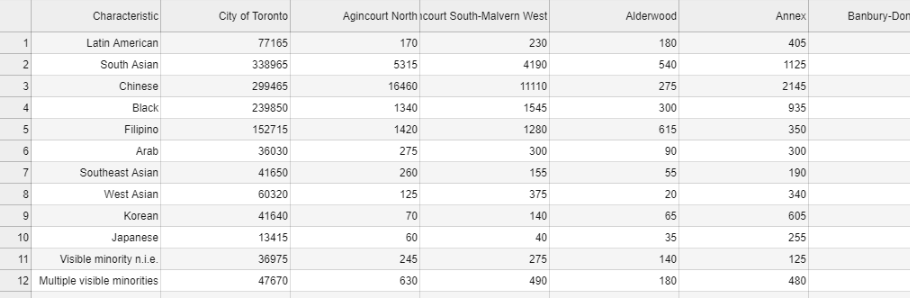
1. [Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) provides a table of postal codes and neighbourhoods.



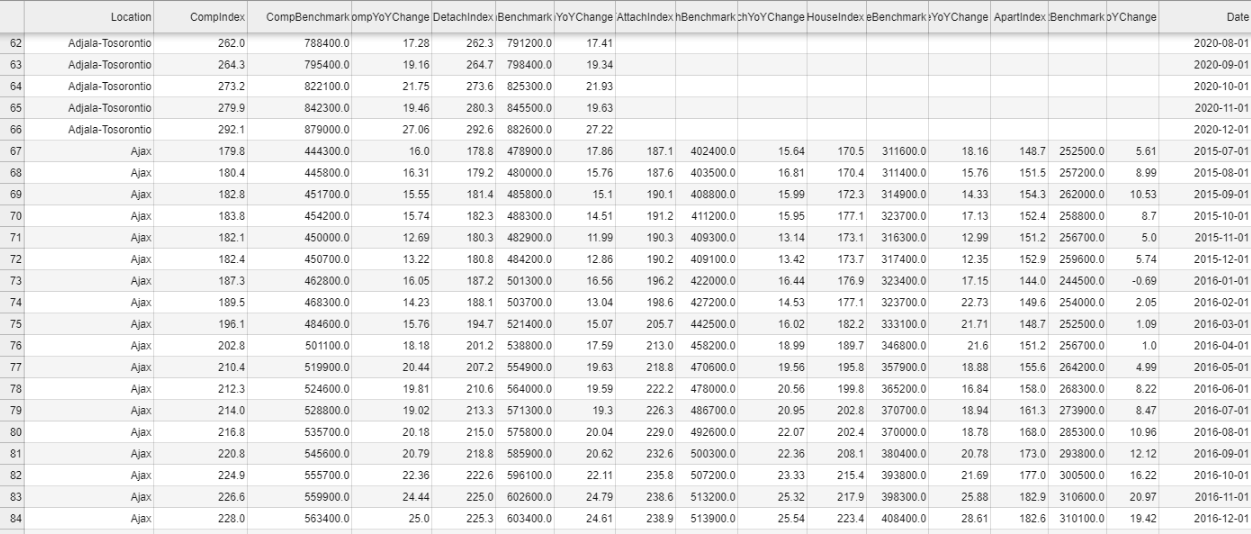
1. [Geocode Python](https://geocoder.readthedocs.io/index.html) provides latitude and longitude coordinates of the postal codes. 
2. [Toronto Open Data Portal](https://open.toronto.ca/) provides Toronto City's Neighbourhood geojson file for creating neighbourhoods map.



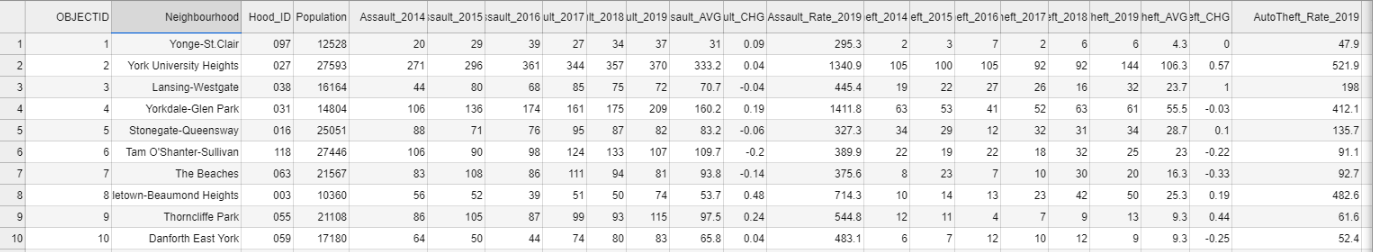
It also provides Neighbourhood Profiles basing on tabulations of 2016 Census of Population data, which portraits demographic and social characteristics of people in each neighbourhood. Please note that Census of Population is held every 5 years in Canada, I look to a refreshed dataset be coming soon.



1. [Kaggle](https://www.kaggle.com/) provides the latest datasets of Toronto House Price information originated from the [Toronto Regional Real Estate Board (TRREB)](https://trreb.ca/). Data is broken down into 4 house types – Single Family Detached, Single Family Attached, Townhouse, Apartment, of individual neighbourhood in the months of 2015-2019. Apart from house price benchmark of each neighbourhood, TRREB also uses various housing attributes to derive House Price Index [HPI](https://trreb.ca/index.php/market-news/mls-home-price-index) which is a measure of weighted house price change.



Kaggle also provides 2014 -2019 Crime Rate, data is broken down into individual neighbourhood’s figures.

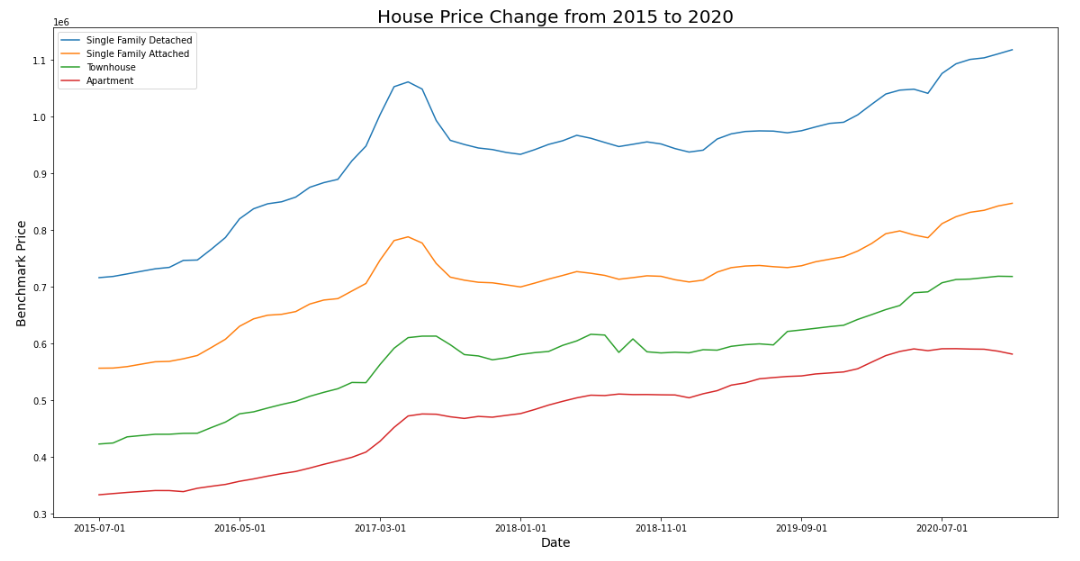


1. Foursquare API for exploring common venues in the 140 Toronto City neighbourhoods.

# Methodology

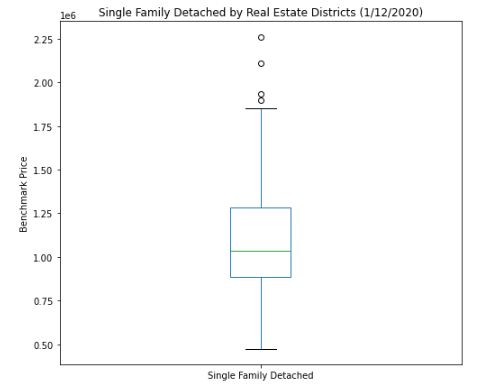
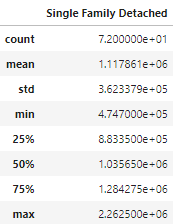
### Plot Price Trend by Line Chart

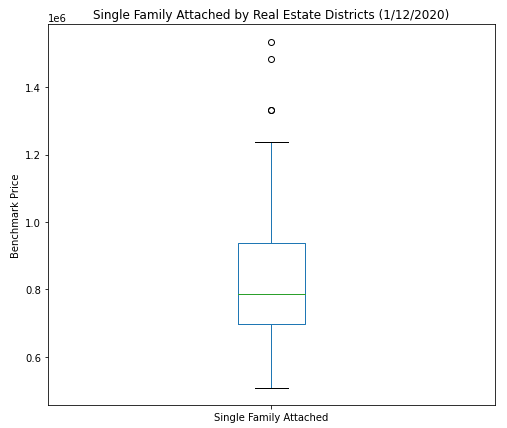
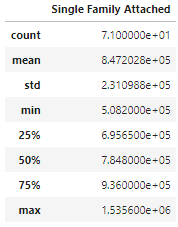
I plot the composite benchmark price of the 4 house types in 2015 to 2019 to compare house types and visualize their price trend. This assists house buyers on making buy or not buy decision.

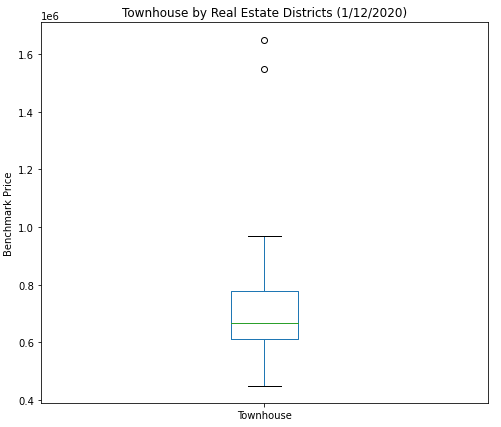
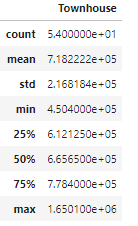


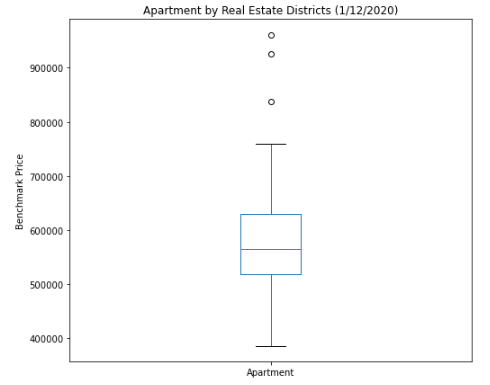
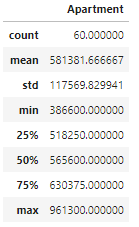
### Box Plot the Price Range

I use box plot to depict the price range of individual house type and see the price distribution. This helps house buyers to understand the affordable house types.

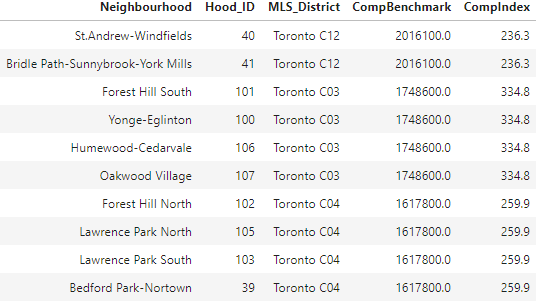
### Rank and Compare Price of Neighbourhoods

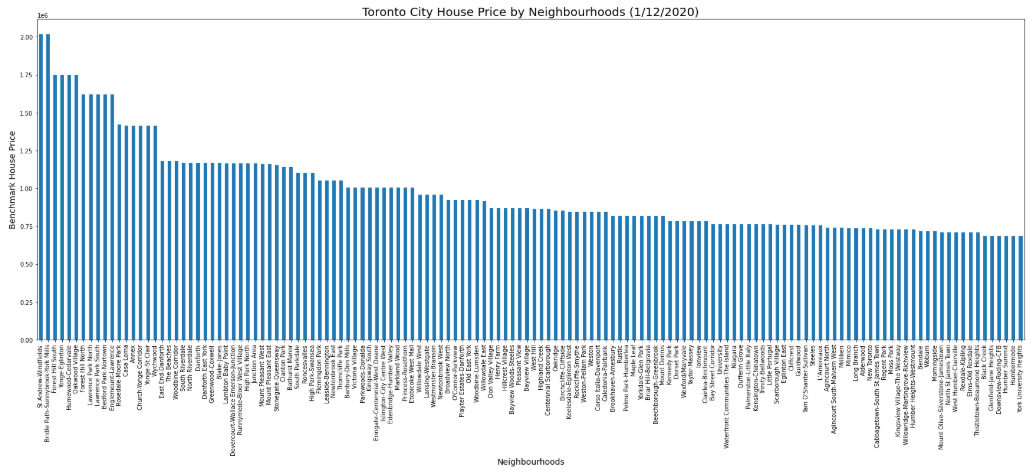
Toronto City is divided into 140 neighbourhoods. TRREB however divides house price data into 35 real estate districts of Toronto City. Although the borders of neighbourhoods and real estate districts may not perfectly align, I manually map the 140 neighbourhoods approximately into 35 real estate districts. Then by merging with house price TRREB house price data, I obtain a table of neighbourhood house price index (HPI) and house price benchmark. Among all neighbourhoods, 2 of them do not have house price data thus they are dropped from the table.



Merging Neighbourhoods with TRREB dataset’s composite house benchmark price and HPI I obtain 138 neighbourhoods’ house price and index. The below bar chart shows the neighbourhoods’ benchmark price, from high to low order, enabling house buyers to position their potential destinations among 140 neighbourhoods given a limited budget.

Top 10 neighbourhoods benchmark price in CAD

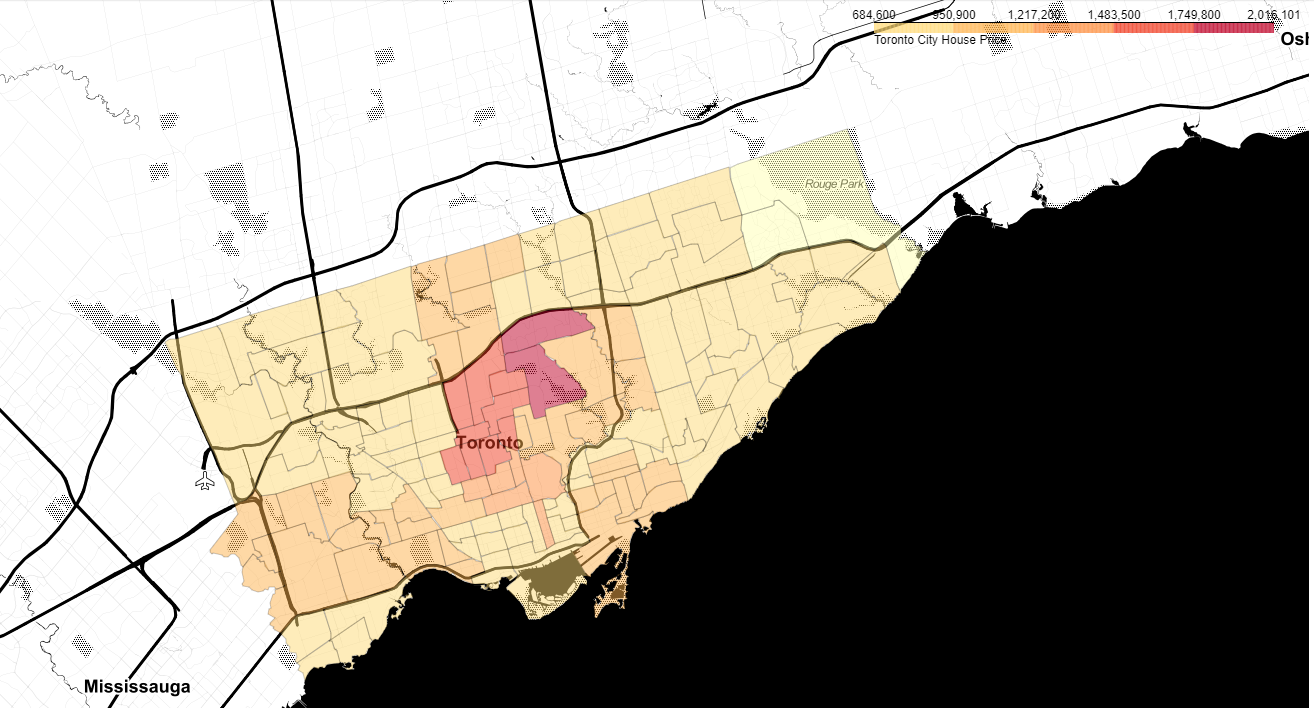




### Visualize House Price Level on Choropleth Map

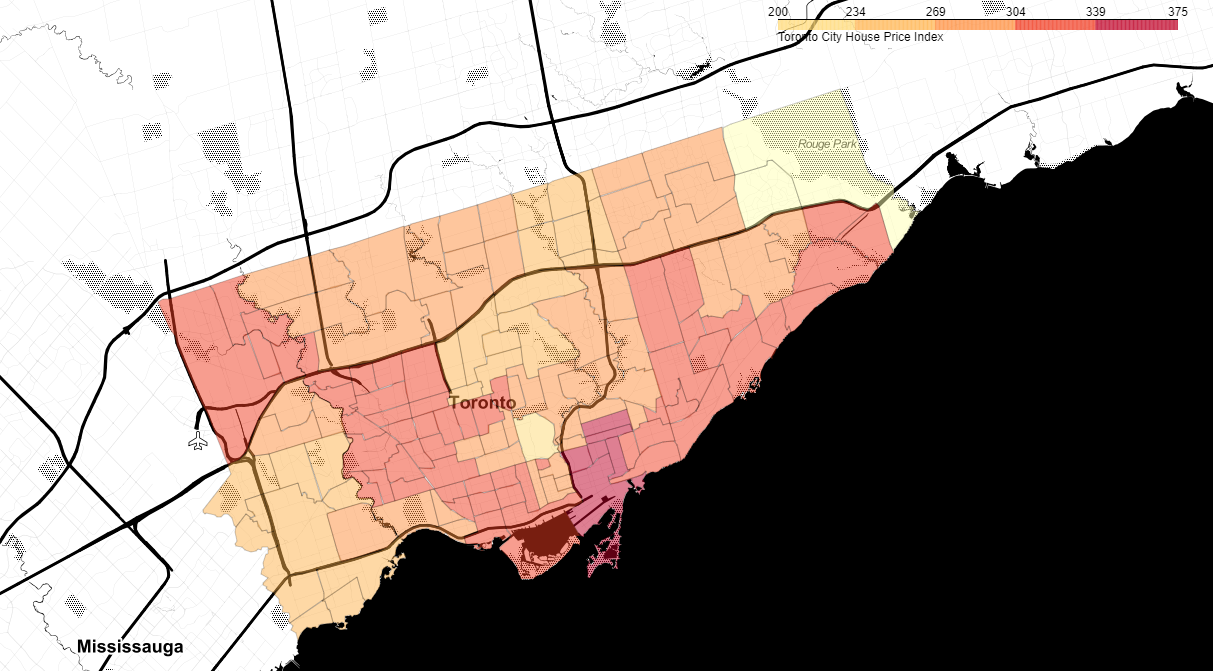
I overlay the 1/12/2020 house benchmark price on Choropleth Map. Benchmark price of Toronto City ranges from CAD 684,600 to 2,016,100. Higher-priced houses are located at the central of Toronto City.

Please note although there are 138 neighbourhoods’ price data, Toronto Open Data Portal’s neighbourhood geojson file only contains coordinates of 100 neighbourhoods. Neighbourhoods on the outskirts of Toronto City are thus not depicted on the Choropleth Map.



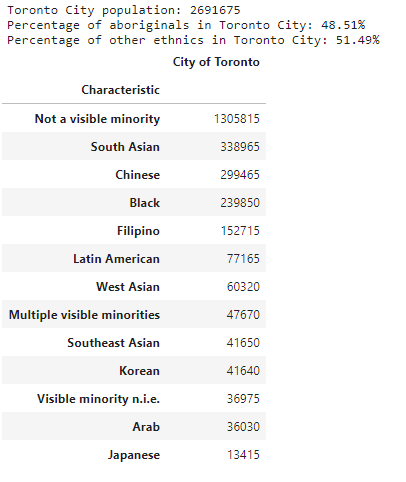
### Visualize House Price Index (HPI) on Choropleth Map

I also create a Choropleth Map to display the neighbourhoods’ HPI as of 1/12/2020. Settlers who concern about price rate change can reference this map for information.

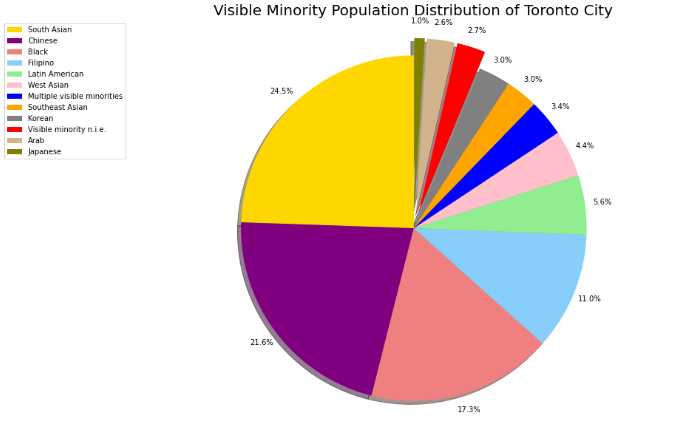


### Illustrate Population Distribution of Visible Minority by Pie Chart

According to 2016 census profile, half of Toronto City's population are identified as visible minorities. A visible minority is defined by the Government of Canada as “persons, other than aboriginal peoples, who are non-Caucasian in race or no-white in color.”. Under this definition, Toronto City population is observed to be contributed by 13 ethnics – Latin American, South Asian, Chinese, Black, Filipino, Arab, Southeast Asian, West Asian, Korean, Japanese, Visible minority n.i.e. (not in elsewhere), Multiple visible minorities, Not a visible minority.



The bar chart below illustrates the proportions of each minority group that totally contributes to 51.49% of Toronto City total residences. South Asian, Chinese and Black are identified as the top 3 minority groups.



### Highlight Other Ethnics-Dominated Neighbourhoods on Choropleth Map

From the above analysis I learnt that the average visible minority population in Toronto City is 51.49%. Next I will check which neighbourhoods specifically have more condensed minority groups than the overage average, and identify the individual group that dominates the residency.

Toronto Open Data Portal provides breakdown of 140 neighbourhoods’ population count of 13 ethnic groups. I transformed the figures into “each ethnic’s proportion in the 140 neighbourhoods”.

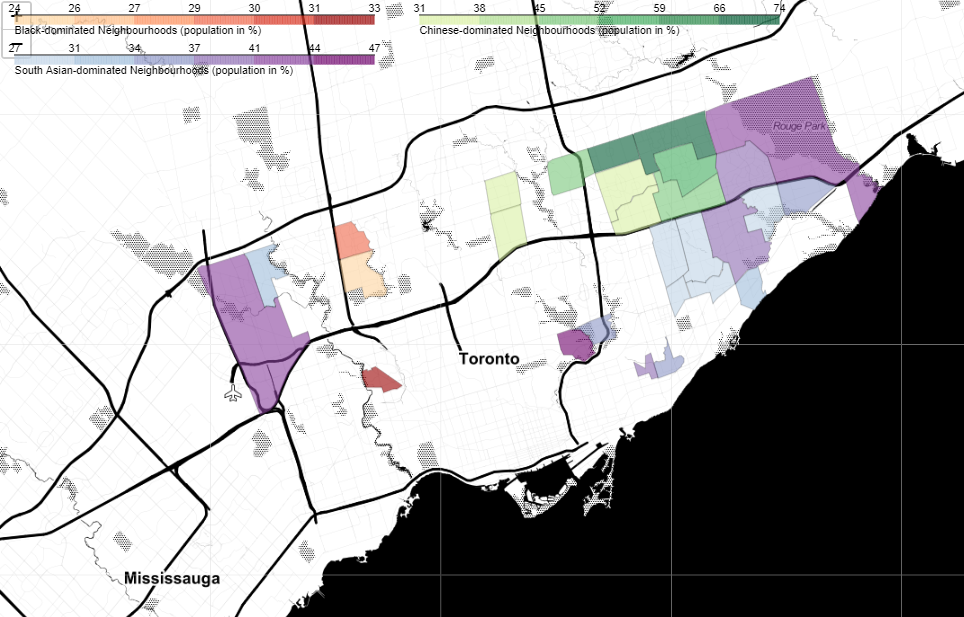
Snapshot of ethnics proportion in 140 neighbourhoods



Below are the findings, telling the identified neighbourhoods, the biggest minority group in these neighbourhoods and this group’s population percentage in the neighbourhoods. Settlers who have preference on either living in close proximity to their own ethnics or bending into the major Canadian aboriginals’ community can reference this information.



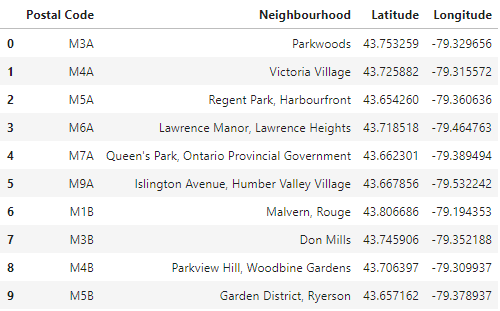
Accordingly, I created 3 individual geojson files to represent the Black, Chinese & South Asian-dominated neighbourhoods. The geojson files are then overlaid on folium as a Choropleth Map to provide a one-view picture of these neighbourhoods’ locations and the ethnics’ population proportions.



### Explore Neighbourhoods’ Nearby Venues by Foursquare API

Toronto City is divided into 140 neighbourhoods which are assigned to 103 postal codes. Below shows a fraction of the merged data of Wikipedia’s postal codes to neighbourhoods mapping and Geocode Python’s postal codes coordinates.

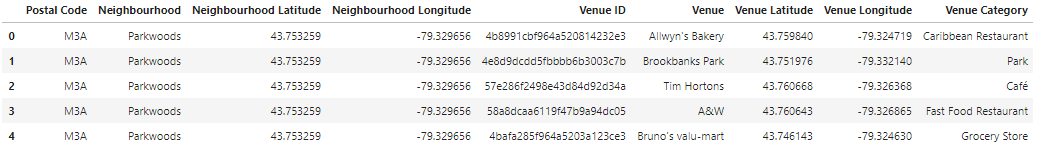
postal codes and coordinates



Note that among all neighbourhoods, only Downsview has multiple postal codes.

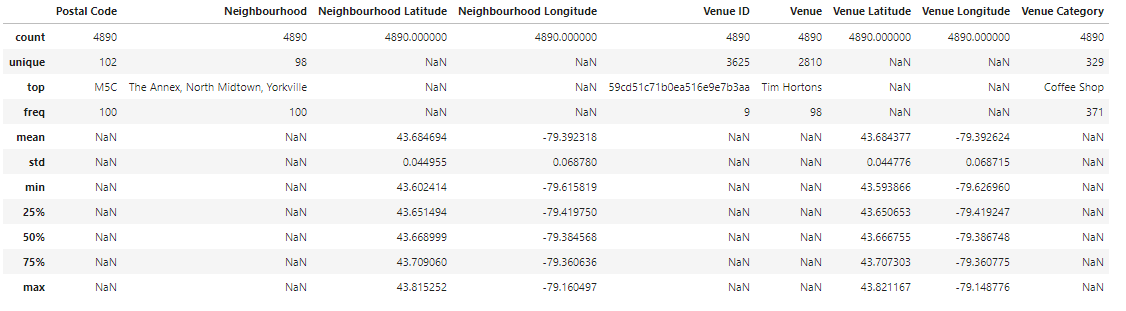


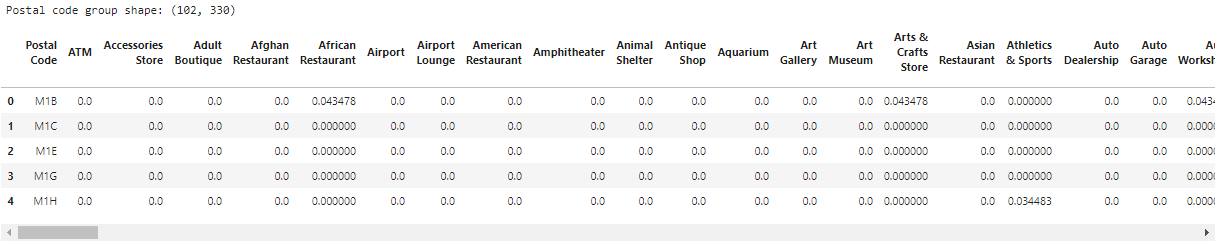
I query each postal code’s popular venues within 1km proximity. Below is a fraction of the result sets.



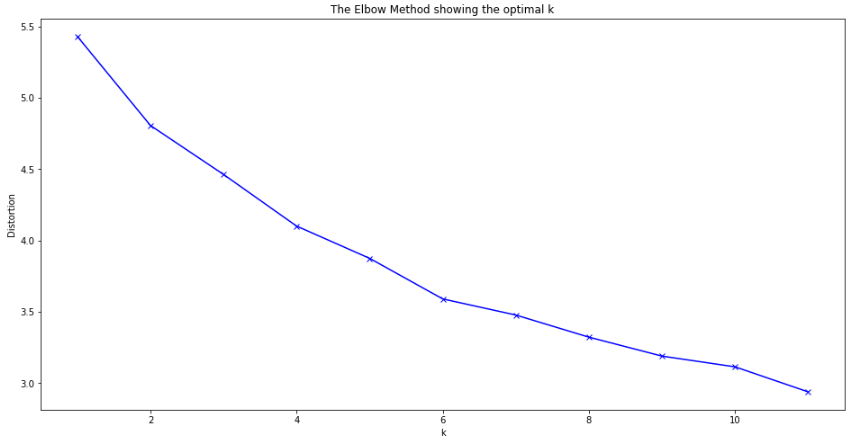
Describe the result set to learn that:

* 4890 venues are discovered
* There are 3625 unique venues and 339 unique categories which means 1265 venues are overlapped in multiple neighbourhoods’ circles.
* Venues are found in 102 postal codes and 98 neighbourhoods. 1 out of total 103 postal codes has no nearby venue discovered.



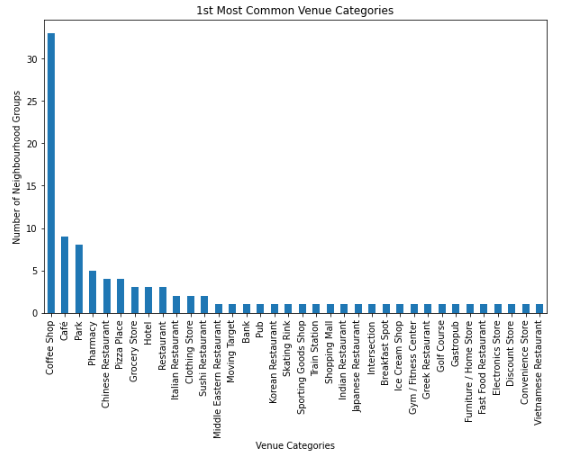
I group the result set by postal code and venue category then transform the data set into a list of unique postal codes and % of each venue category being discovered in the postal code area. 

I use K-Means methodology to cluster the neighbourhoods by commonality of their nearby venue categories. I try to find the optimal k of K-Means by elbow method but the elbow is not acute. I decide to use k=6 for clustering.

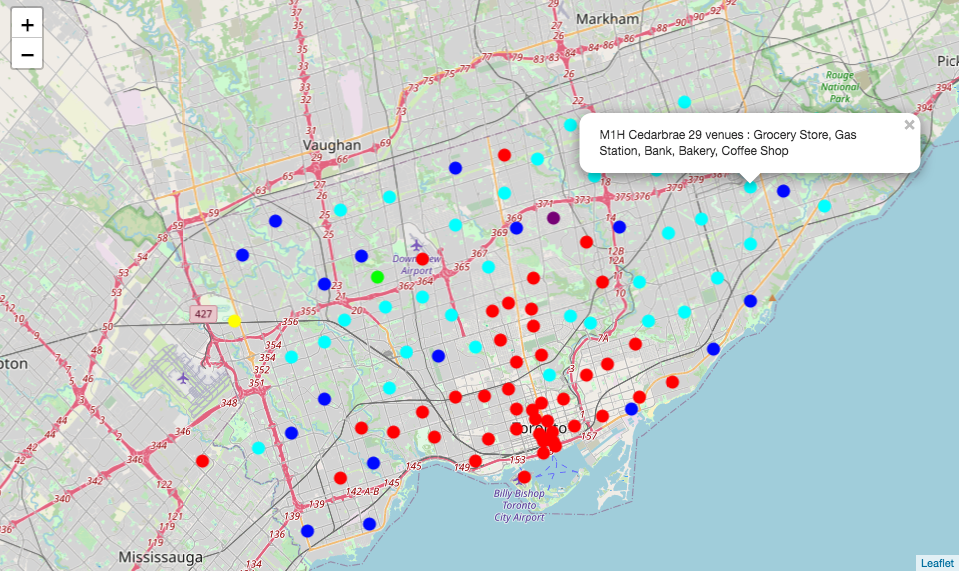


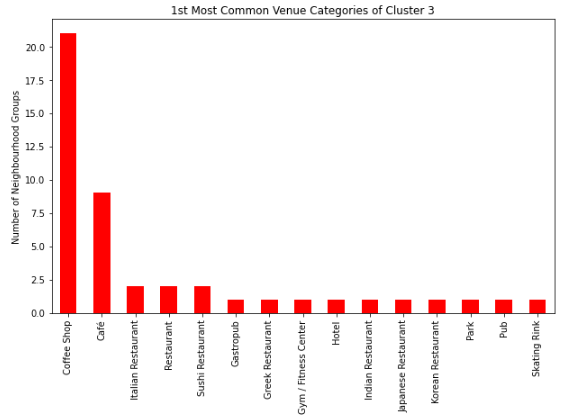
I sort out the 5 most common venue categories and venues count of each neighbourhood. Joining back postal code coordinates and the corresponding neighbourhoods.

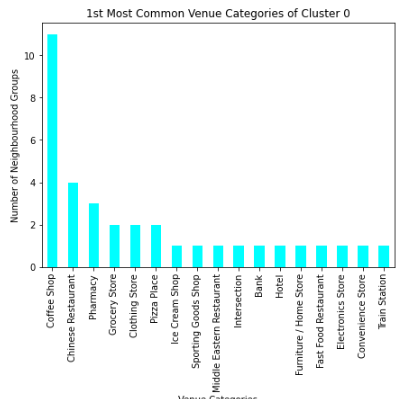


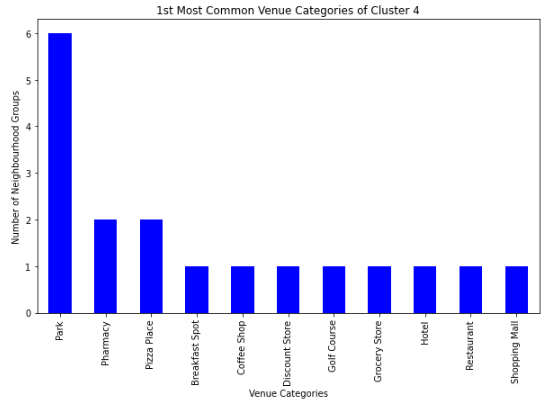
Below bar chart shows the 1st most common venues in all neighbourhoods. 

I visualize the neighbourhoods on folium map using different colors to represent the cluster group that they belong to. Markers are labelled to describe neighbourhood name, number of venues and the 1st-5th most common venue categories.









Foursquare API only discovers 3 to 4 venues in LIME, PURPLE, YELLOW clusters. They can be regarded as outliers. K-means uses % of each venue category in a neighbourhood to calculate the clustering. With only a few venues dominating the % of venue categories in one neighbourhood will tend to form a highly distinctive cluster by the neighbourhood itself. This may explain why the elbow chart is flattened and unable to form an acute elbow as k increases.

LIME Cluster



PURPLE Cluster



YELLOW Cluster



### Analyze Crimes Rates by Crime Type

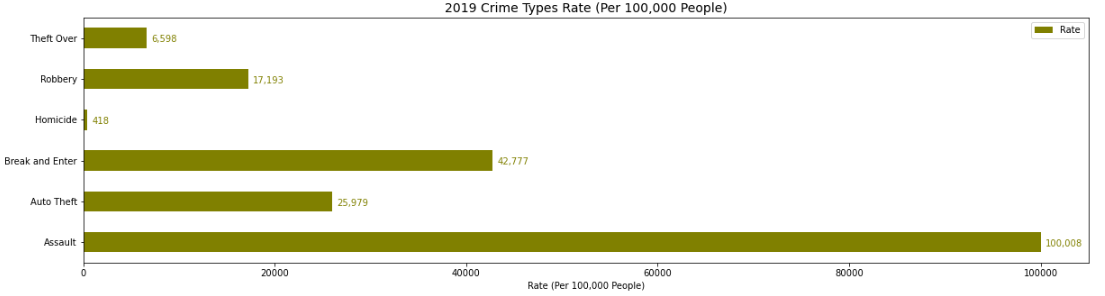
Kaggle provides 2014-2019 crime rates, i.e. number of crimes per 100,000 people, of each individual neighbourhood. I will study the 2019 data as it is the latest. Counts are available for Assault, Auto Theft, Break and Enter, Homicide, Robbery and Theft Over.

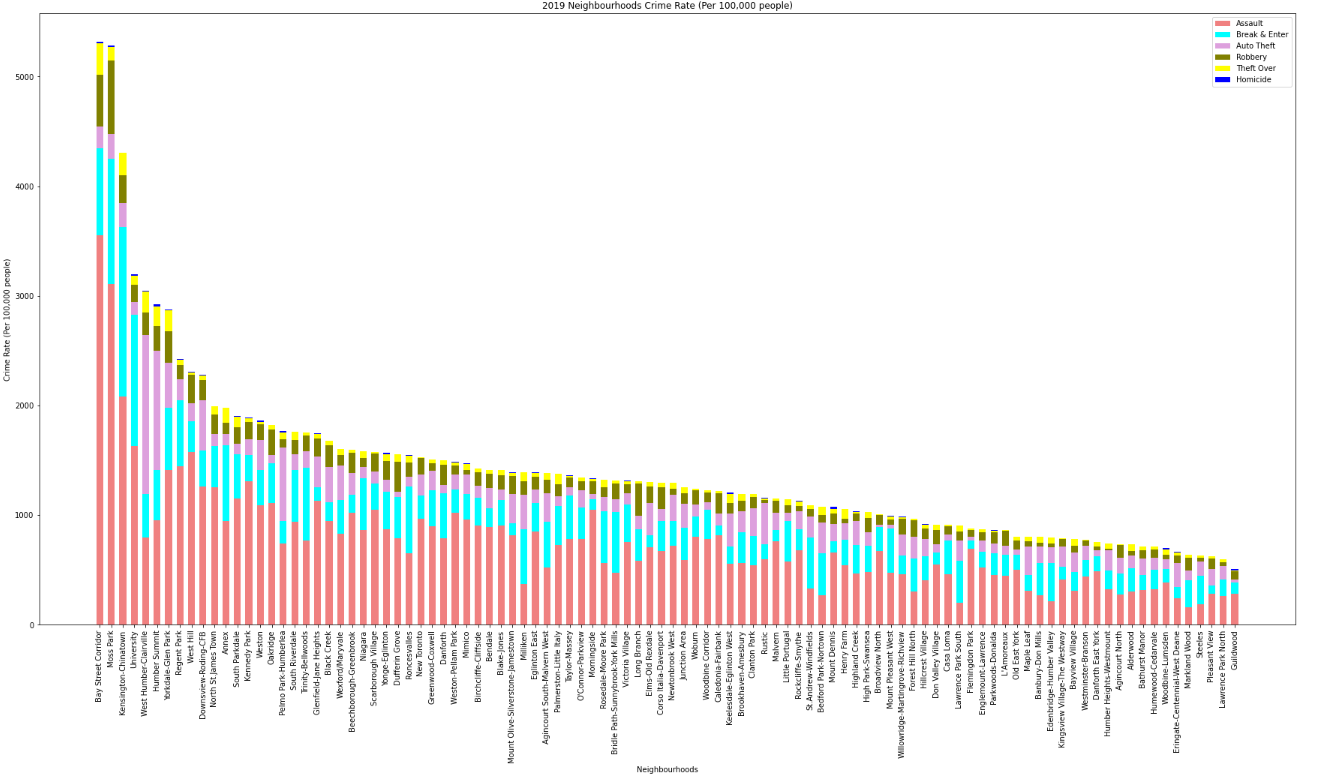
Fraction of 140 neighbourhoods’ 2019 crime rates



Describe the dataset to find that Assault has the highest overall crime rate among the 6 crime types in Toronto City. A comparison of crime types can be clearly presented with the below horizontal bar chart.





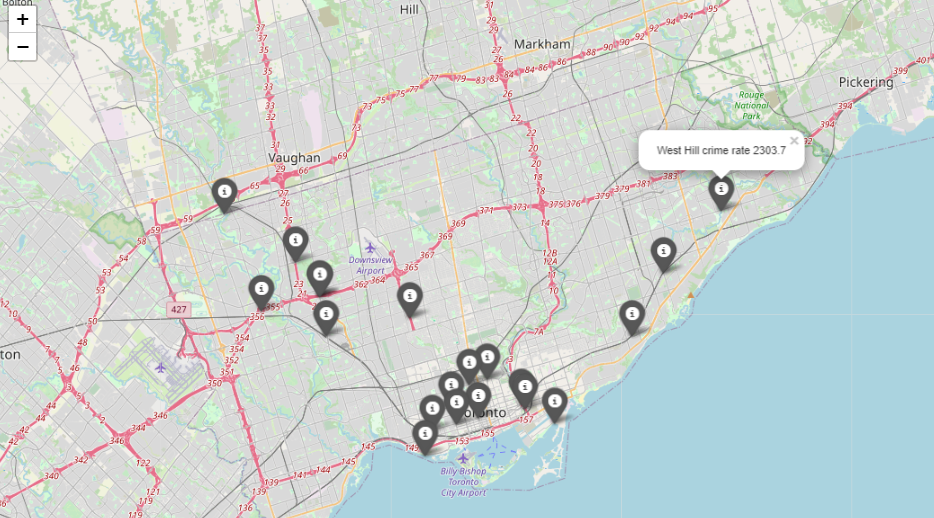
I rank the neighbourhoods by their total crime rates and plot them all in stacked bar charts, in high to low order. People who are sensitive to community safety level can easily spot the neighbourhoods which they would like to avoid. By inspecting the stacked bars, the graph can tell 3 neighbourhoods have exceptionally high Auto Theft rate than other neighbourhoods, they are West Humber-Clairville, Humber Summit and Pelmo Park-Humberlea.

### Pin the Least Safe Neighbourhoods on Folium Map

Considering that settlers will be more attentive to of high-crime-rate neighbourhoods, I further pin 20 highest-crime-rate neighbourhoods on folium map for anyone to overview the locations of these neighbourhoods on map. The neighbourhoods are pinned on an interactive folium map by markers as well as tips labelling neighbourhood names and crime rate.

20 highest crimes rate neighbourhoods

Bay Street Corridor, Moss Park, Kensington-Chinatown, University, West Humber-Clairville, Humber Summit, Yorkdale-Glen Park, Regent Park, West Hill, Downsview-Roding-CFB, North St.James Town, Annex, South Parkdale, Kennedy Park, Weston, Oakridge, Pelmo Park-Humberlea, South Riverdale, Trinity-Bellwoods, Glenfield-Jane Heights



# Results

### House Price

Toronto house price shows an upwards trend in past 5 years. Special attention should be paid to apartment which begins to go downwards since 2020. Townhouse’s rise is flattening as well.

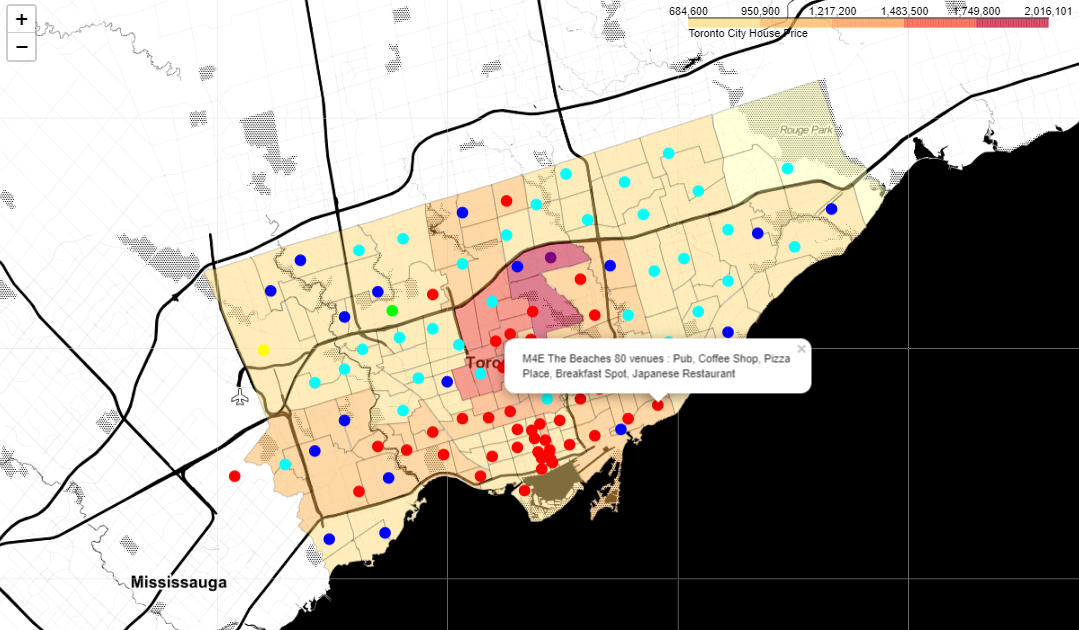
Benchmark price by different house types are summarized in below table. Potential house buyers can reference this to determine their affordable house types.

|  |  |  |
| --- | --- | --- |
| **House Type**  **(as of 1-Dec-2020)** | **Mean Price (CAD)** | **Standard Deviation from Mean (CAD)** |
| Single Family Detached | 1,117,861 | 362,338 |
| Single Family Attached | 847,203 | 231,099 |
| Townhouse | 718,222 | 216,818 |
| Apartment | 581,382 | 117,570 |

When measuring by the composite house price, the highest-priced houses are located in 17 neighbourhoods at central of Toronto City. Potential house buyers can reference below table and map to determine their affordable destinations. There color dots will be explained in later section.

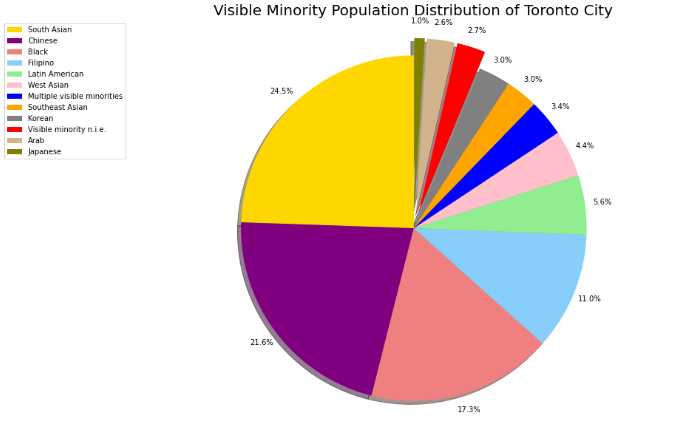
|  |  |  |
| --- | --- | --- |
| **Neighbourhood** | **Composite Benchmark Price (CAD)** | **Price Bin** |
| St.Andrew-Windfields | $2,016,100 |  |
| Bridle Path-Sunnybrook-York Mills | $2,016,100 |  |
| Forest Hill South | $1,748,600 |  |
| Yonge-Eglinton | $1,748,600 |  |
| Humewood-Cedarvale | $1,748,600 |  |
| Oakwood Village | $1,748,600 |  |
| Forest Hill North | $1,617,800 |  |
| Lawrence Park North | $1,617,800 |  |
| Lawrence Park South | $1,617,800 |  |
| Bedford Park-Nortown | $1,617,800 |  |
| Englemount-Lawrence | $1,617,800 |  |
| Rosedale-Moore Park | $1,423,000 |  |
| Casa Loma | $1,413,700 |  |
| Annex | $1,413,700 |  |
| Church-Yonge Corridor | $1,413,700 |  |
| Yonge-St.Clair | $1,413,700 |  |
| Wychwood | $1,413,700 |  |

Price Level Map



### Ethnics Population & Distribution

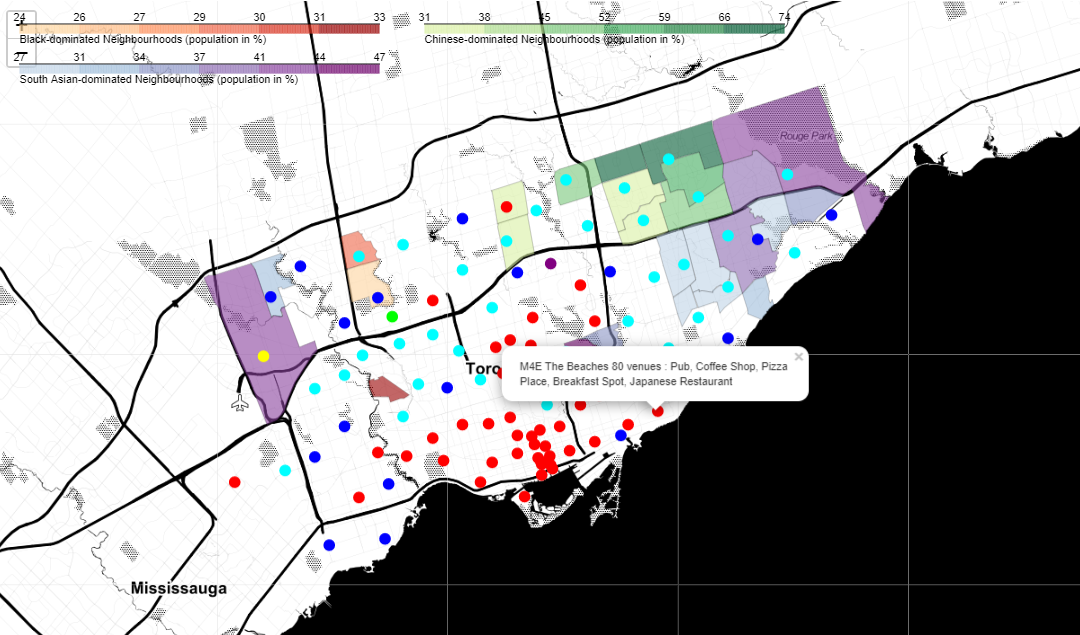
According to 2016 census profile, 51.49% of Toronto City's population are identified as visible minorities. A breakdown of the minority by 12 ethnic groups is shown in below pie chart. **South Asian, Chinese and Black** are the 3 biggest ethnic groups which largely contribute to the 51.49% minority residences, that is equivalent to about **33% of 2.9M Toronto City Total Population**.



In 27 out of the total 140 neighbourhoods, residency % from a single ethnic outreaches that of the aboriginals. Potential settlers who have preference on either living in close proximity to their own ethnics or bending into the major Canadian aboriginals’ community can reference below table and map (the colored dots will be explained in next section).



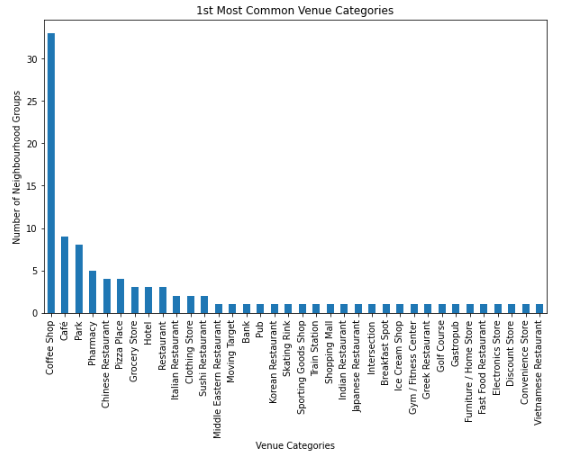
Minority Dominated Neighbourhoods MAP



### Neighbourhood Features

K-means methodology is applied to cluster neighbourhoods by the venue categories explored by Foursquare API. 3625 unique venues of 339 unique categories are found in Toronto City neighbourhoods. 6 clusters are overlaid on the above maps, representing by dots in different colors. The pins are labeled to describe the neighbourhood name, number of venues found in 1 km proximity and the 1st-5th most common venue categories.

Toronto City neighbourhoods’ 1st most common venues are summarized in below bar charts. Coffee Shop completely stands out, following it are Café, Park, & Pharmacy.



**Red Cluster**

It can be seen that coffee shop and cafe are highly centralized in the RED neighbourhoods which are largely located along Yonge Street. RED cluster also major in wide variety of international cuisines.

**Cyan Cluster**

In CYAN neighbourhoods, coffee shop is still the 1st most common venue although the number is just half of that of RED cluster. Dining places appear to have less variety in CYAN cluster, chinese restaurant stands out among all dining options. CYAN cluster is well-occupied by variety of merchandizes including pharmacy, grocery store, clothing store, furniture store, electronics store, convenience store. It also has park, hotel, transportations and intersection. CYAN cluster is more multi-functional.

**Blue Cluster**

BLUE cluster is specially featured with an abundance of parks. BLUE neighbourhoods also serve larger leisure venues such as discount store, shopping mall and golf course

The LIME PURPLE YELLOW clusters comprised of only 1 neighbourhood each, and are regarded as outliers since only 3 to 4 venues are discovered by Foursquare API.

**Lime Cluster** 2 Vietnamese restaurant, 1 Baseball Field, 1 Restaurant

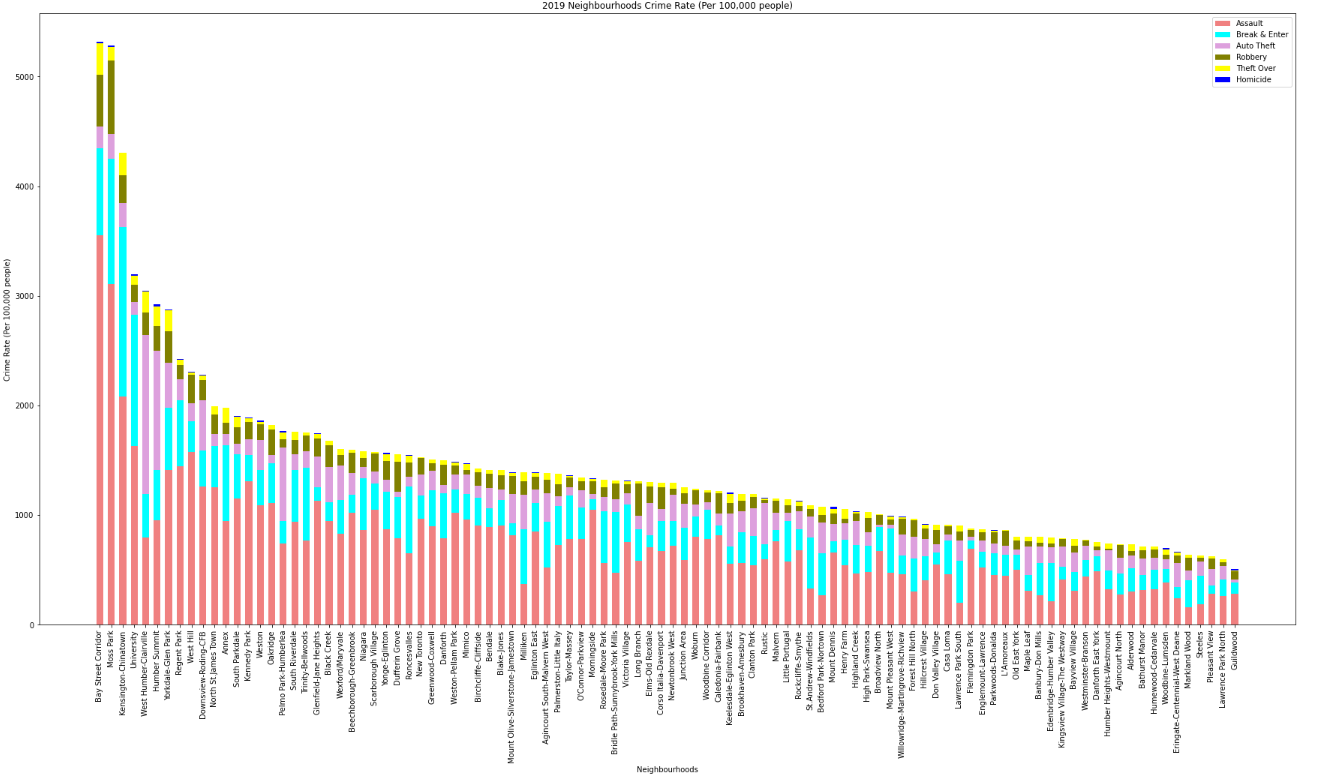
**Purple Cluster** 1 Pool, 3 Parks

**Yellow Cluster** 1 Coffee Shop, 1 Lounge, 1 Moving Target

### Community Safety

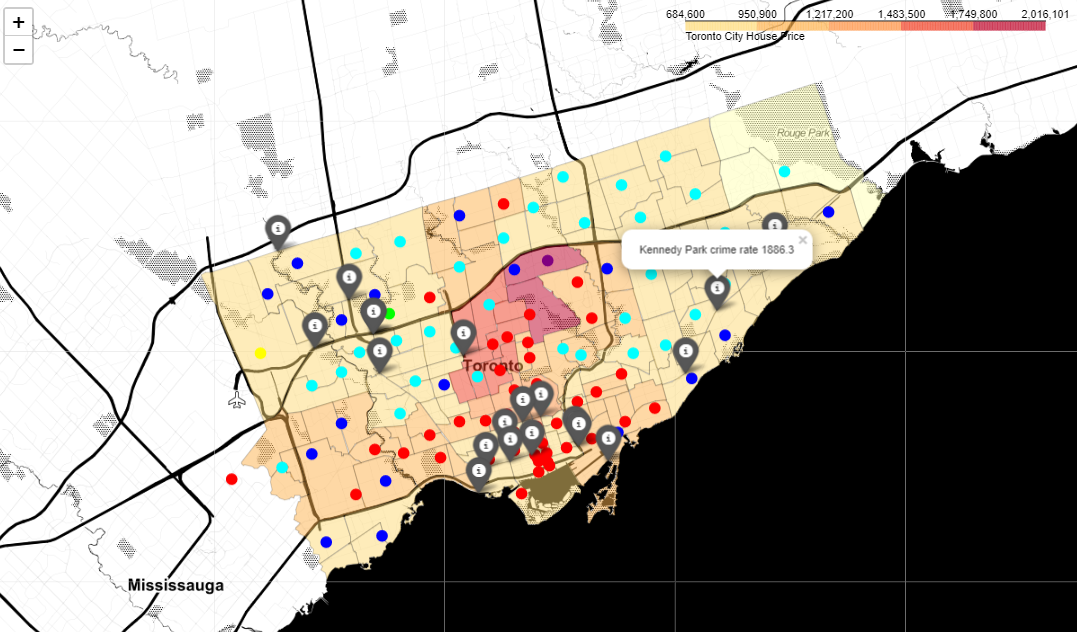
This section is for people who are sensitive to community safety to shortlist the least safe neighbourhoods which they may wan to avoid. 2019 Crime Rate (per 100,000 people) per neighbourhood is laid out in below stacked bar chart which provides 3 insights:

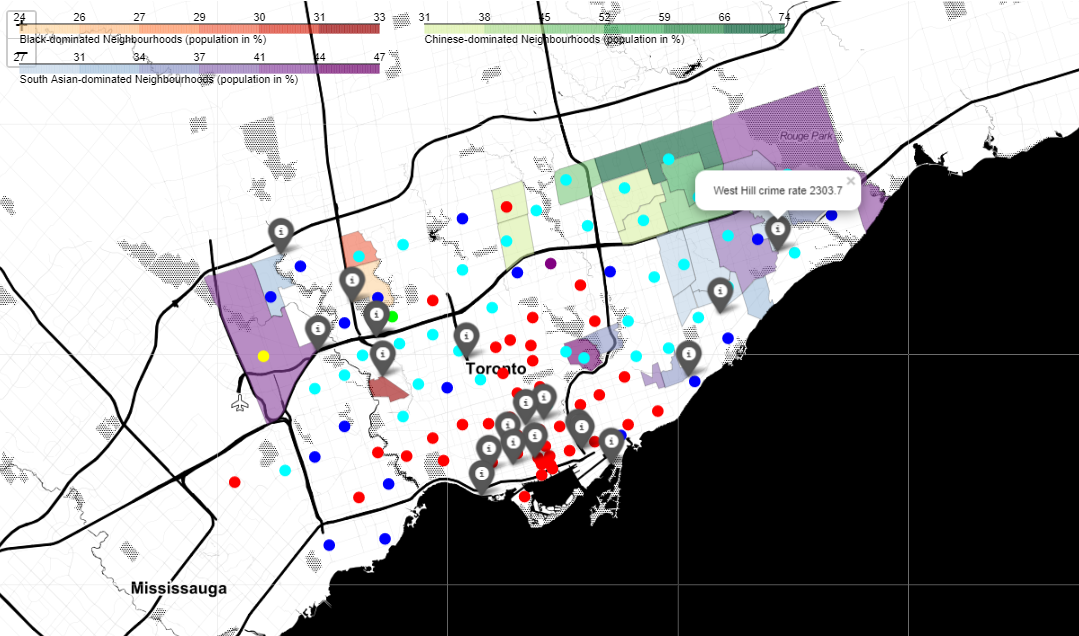
* Assault has the highest overall crime rate among the 6 crime types. Following after are Break and Enter, Auto Theft, Robbery. Crime rates of Theft Over and Homicide are comparatively low.
* 3 neighbourhoods have exceptionally high Auto Theft rate than other neighbourhoods, they are West Humber-Clairville, Humber Summit and Pelmo Park-Humberlea.
* The 20th highest-crime-rate neighbourhoods are: Bay Street Corridor, Moss Park, Kensington-Chinatown, University, West Humber-Clairville, Humber Summit, Yorkdale-Glen Park, Regent Park, West Hill, Downsview-Roding-CFB, North St.James Town, Annex, South Parkdale, Kennedy Park, Weston, Oakridge, Pelmo Park-Humberlea, South Riverdale, Trinity-Bellwoods, Glenfield-Jane Heights.



Maps are further enhanced to add markers for the 20 highest-crime-rate neighbourhoods. Markers are labeled to describe the neighbourhood name and the respective 2019 total crime rate. Readers can overview neighbourhood clusters and least safe neighbourhood on either the Price Level Map or the Minority-dominated Map.

Price Level Map



Minority Dominated Neighbourhoods MAP

# Discussion

Looking at the enhanced Price Level Map above, I would suggest the CYAN cluster neighbourhoods are more favorable for living. The neighbourhoods are found to be featured with multi-functional venues around, fair house price and less crimes. While whether ones prefer to stay close to specific ethnics’ zone is very personal consideration.

However, there are a few missing pieces in this study due to the limitation of data sources.

As mentioned earlier, TRREB divides Toronto City’s houses into 35 real estate districts. The house price statistics is detailed down to 35 districts but the entire area actually covers 140 neighbourhoods. The research can be more precise if each neighbourhood’s house price is available as input for my study.

Besides, Toronto Open Data Portal’s neighbourhood geojson file only contains 100 neighbourhoods’s coordinates. 40 neighbourhoods on the outskirts of Toronto City are missing from the Choropleth Map sketch.

The statistics of visible minorities is originated from 2016 Census of Population. The recent actual figures could have changed. Canada’s Census of Population is usually held every 5 year, I look to a refreshed dataset be coming soon.

In the neighbourhood clustering study, totally 4890 venues are discovered within 1km proximity and among these venues, only 3625 venues are unique. That means 1265 venues are overlapped in more than 1 neighbourhoods’ circles. This problem is unavoidable due to difference size and irregular boundary shape of the neighbourhoods. The trial of using 500m proximity proves that it does not significantly reduce the overlaps, rather it reduces the number of venues by half which does not provide sufficient venue category features and accurate averaged category counts for the clustering algorithm. After all, I determine 1km as an appropriate walkable distance. Although the overlap is not perfect condition, it indeed does not violate the principle in this neighbourhood clustering model.

I try to find the optimal k for K-Means clustering by elbow method but the elbow is not acute. I ended up take k=6. Among the 6 clusters, Foursquare API only discovers 3 to 4 venues in LIME, PURPLE, YELLOW clusters. K-means uses % of each venue category in a neighbourhood to calculate the clustering. With only a few venues dominating the % of venue categories in one neighbourhood will tend to form a highly distinctive cluster by the neighbourhood itself. This may explain why the elbow chart is flattened and unable to form an acute elbow as k increases.

I am wondering if there is another clustering algorithm or a bigger venue dataset that yields better results.

# Conclusion

In spite of some limitations, this study achieves overview of the latest house price by type and the price trend of Toronto City; exploration of livable neighbourhood zones in terms of community environment and safety; and distribution of major ethnics. As discussed above, it is proposed to further deepen the research on house price and neighbourhood features.

When ones ponder immigration and the destination to settle in, the decision is influenced by a basket of factors and concerns. This study only focus on 4 aspects. Statistical researches can possibly be extended to include e.g. general living costs, job opportunities, educations etc.