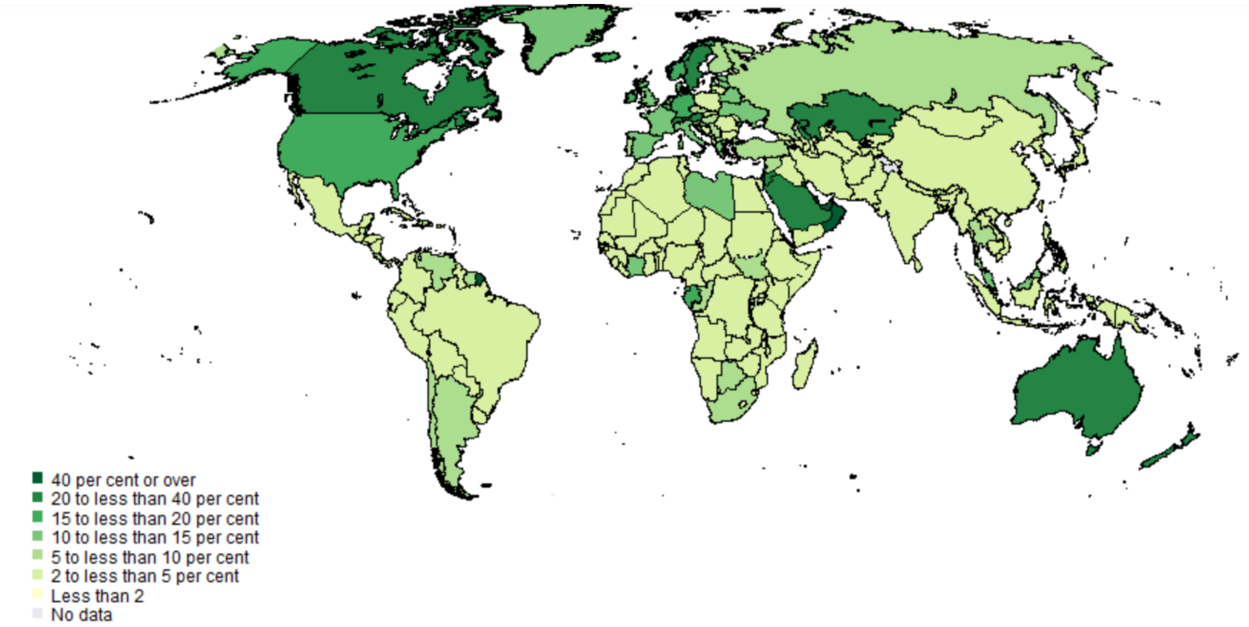
Battle of Neighborhoods

North York, Toronto - Average Housing Prices & School Ratings in the Neighbourhood

1. Introduction & the business problem

According to US News, Canada is a top major destination for immigrants across the world. As per the latest statistics by the United Nations, between 20% to 40% of the Canadian Population in 2019 were immigrants.



40 per cent or over
20 to less than 40 per cent
15 to less than 20 per cent
10 to less than 15 per cent
5 to less than 10 per cent
2 to less than 5 per cent
Less than 2
No data

Every immigrant looks for a better quality of life in the new country than her/his country of origin. A comfortable living with basic facilities is the needs of each immigrant families. Happy immigrants result in better and prosper.

Project Description:

As mentioned in the background, we are trying to solve a real-life macro problem of thousands of immigrants flocking to Canada. Immigrants coming to Canada needs a livelihood and hence stays near the economic center of the country. Toronto becomes the natural choice for immigrants as per Wikipedia the city is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. However, an immigrant moving to a large city like Toronto needs to decide the neighbourhood to stay out of the various choices.

The major determinable factors to settle in would be **housing prices** and **ratings of schools**. The end result provides users an overview of the places before they are moving to Toronto as their new country or city as their place for work or to begin a new life.

In summary, this Project would assist an immigrant or users have more informed decision on choosing the best neighborhood out of many neighborhoods to move into Toronto city based on the housing prices and school ratings.

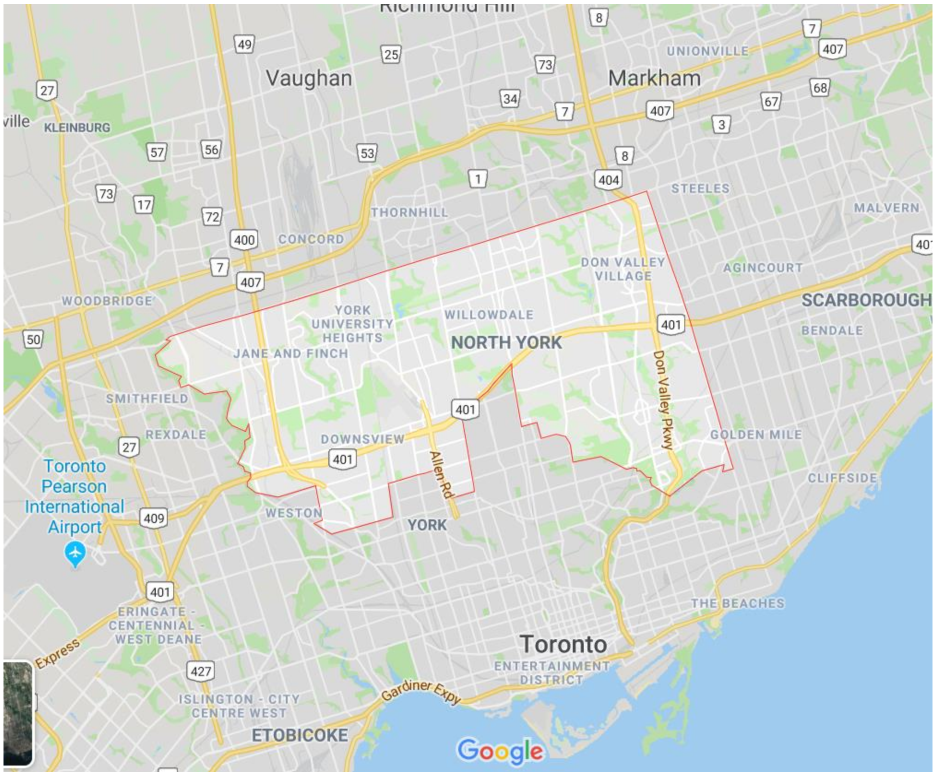
Selection criteria

For the purposes of this project, the definition of a good neighborhood is one that has an appreciable commercial presence within a given community as well as:

1. Compare mean housing prices
2. Compare school ratings

The Location:

North York is a popular destination for new immigrants in Canada to reside. As a result, it is one of the most diverse and multicultural areas in the Greater Toronto Area, being home to various religious groups and places of worship. Although immigration has become a hot topic over the past few years with more governments seeking more restrictions on immigrants and refugees, the general trend of immigration into Canada has been one of on the rise. Again referring to Wikipedia over the late 20th century and early 21st century, North York City Centre have emerged as secondary business districts outside Downtown Toronto. High-rise development in these areas has given the former municipalities distinguishable skylines of their own with high-density transit corridors serving them.
Thus, this projects aim to create an analysis of features for **North York neighborhood**. The features include like mean house price, school ratings, population rate, crime rates, recreational facilities, etc.



Foursquare API:

This project would use Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

Work Flow:

Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

Clustering Approach:

To compare the similarities of two cities, we decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods in a big cities like New York and

Toronto. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm

Libraries: List of Python libraries in alphabetical order with a brief overview

Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.

Geocoder: To locate the coordinates of a given addresses

Geopy: To retrieve Location Data

JSON: Library to handle JSON files

Matplotlib: Python Plotting Module

Pandas: For creating and manipulating data frames

Requests: Library to handle http requests

Scikit Learn: For importing k-means clustering

XML: To parse and modify XML document

2. Data section

We will discuss the data used in this problem in three buckets. Firstly, the geo location data, then the local search API data namely Foursquare and lastly the web research data that was fed in the Python analytical model.

Longitude and Latitude Data:

We will need geo-locational information about that specific borough and the neighborhoods in that borough. It is "North York" in Toronto. This project will require knowledge of the different neighborhoods in Toronto, school ratings and mean house prices. As such the neighborhood data required will be:

1. Neighborhood location in terms of latitude and longitude
2. School Ratings
3. Average Housing Prices

Dataset comprising latitude and longitude, zip codes is already available through the previous notebook. The location of North York would be filtered using the same:

https://github.com/SaunakBhattacharyya/Coursera_Capstone/blob/master/Task%202:%20geographical%20coordinates.ipynb

We get the list of neighborhoods, boroughs and postal codes from the Wikipedia website. https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto

Then for further deep dive and study of the demographics of each of the neighbourhoods, we use Wikipedia information again.

https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

In order to establish the targeted neighborhood(s), we will explore the demographics of the neighborhoods in the city of Toronto by segmenting the data and conducting descriptive analysis using Panda. Additional data will be gleaned by web scraping and API will be used to generate data.

Foursquare API Data:

We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue
5. Name of the venue e.g. the name of a store or restaurant
6. Venue Latitude

7. Venue Longitude
8. Venue Category

Secondary Research Data:

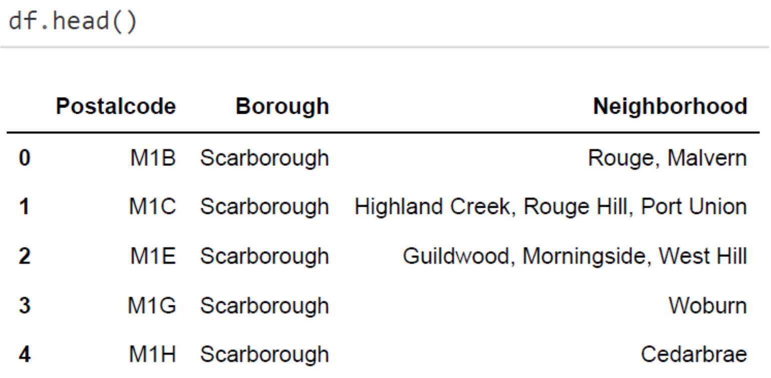
North York as a borough occurs in multiple post codes (M2, M3, M6, etc.) it brings in the chart neighbourhoods from other borough also. So, we create a simple excel file with the 102 neighbourhoods names listed above and two columns containing the average housing price and school ratings. Using ranking method, we arrive at the probable best neighbourhoods in North York, Toronto.

- Now we do some extensive secondary research over websites to find the average housing price for these over 100 neighbourhoods. As there is no single website that provides the average housing price in one place or one table. We collect information from 3 websites namely:
    - Numbeo https://www.numbeo.com/cost-ofliving/in/Toronto
    - Point2homes https://www.point2homes.com/CA/Real-Estate-Listings/ON/Toronto.html
    - Squarespace Report - https://static1.squarespace.com/static/546bbd2ae4b077803c592197/t/5c5c92ae15fcc0cc392edc40/15495707760 and collate the information on an excel sheet for further data processing and ranking
- We face the same problem with the school ratings of the neighbourhoods. There is no single website or a report that provides this information in one table or place. So, we use various filters and arrive at the school ratings for the neighbourhood using Compare School Rankings website
    - http://ontario.compareschoolrankings.org/elementary/SchoolsByRankLocationName.aspx?schooltype=elementary

Note we use only elementary school rankings here for simplicity in decision making.

3. Methodology & Results section

Data processing - First we convert the extracted the content of Postal Code HTML table as data frame, for further usage.

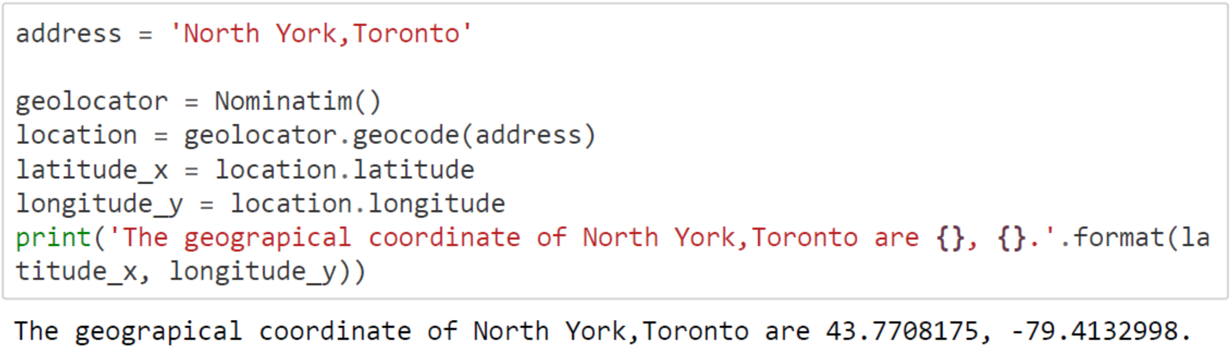


```
df.head()
```

	Postalcode	Borough	Neighborhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Data Cleaning - The web page data shows missing data. So, we need to drop the "None" rows in Data Frame. Thus, we will drop any row which contains 'Not assigned' value. All "Not assigned" will be replace to 'NaN' using numpy library.

Data Import - Let us find our objective of interest the North York neighbourhood and print its geo coordinates.

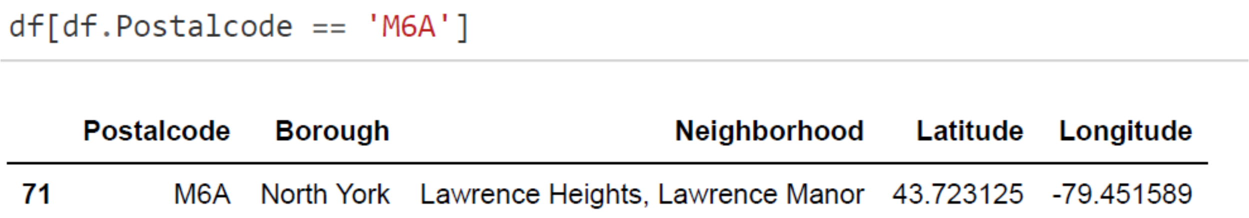


```
address = 'North York,Toronto'

geolocator = Nominatim()
location = geolocator.geocode(address)
latitude_x = location.latitude
longitude_y = location.longitude
print('The geograpical coordinate of North York,Toronto are {}, {}.'.format(la
titude_x, longitude_y))
```

```
The geograpical coordinate of North York,Toronto are 43.7708175, -79.4132998.
```

Post Code 'M2', 'M3', 'M4', 'M6', 'M9' are the codes for North York from the Wikipedia list imported above. Note we are getting an approximate geo location/coordinates for the entire Borough of North York and not a single Neighbourhood. Let us get the information for Lawrence Heights neighbourhood out of North York.

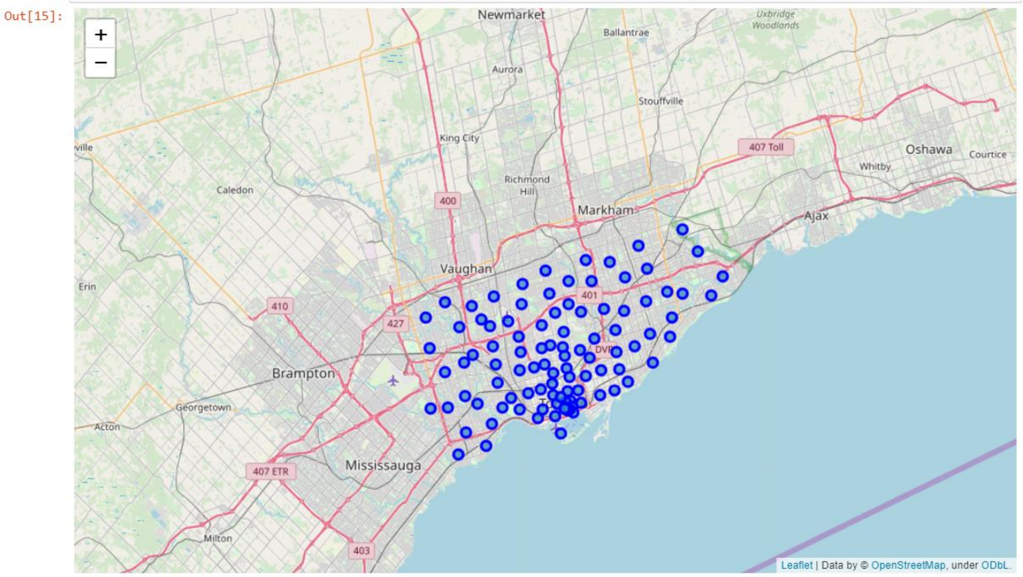


```
df[df.Postalcode == 'M6A']
```

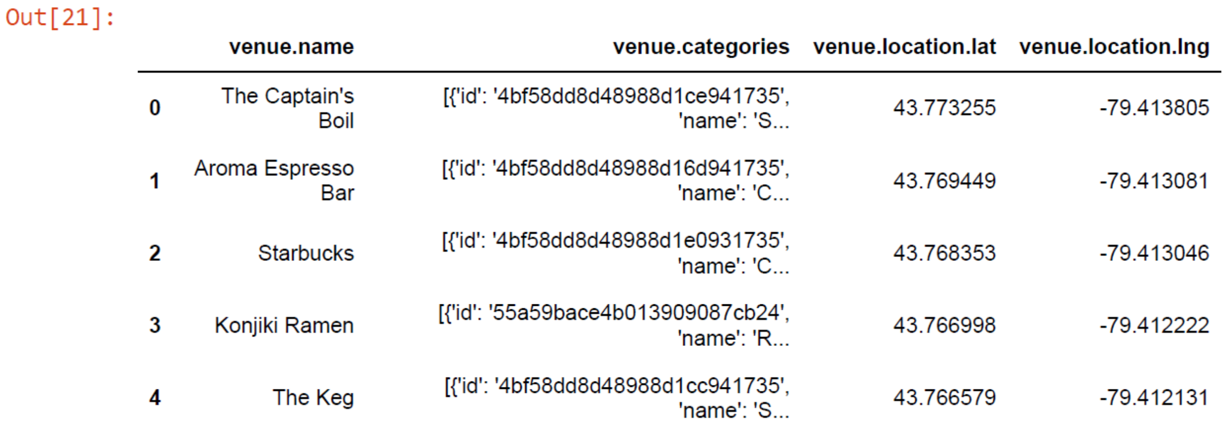
	Postalcode	Borough	Neighborhood	Latitude	Longitude
71	M6A	North York	Lawrence Heights, Lawrence Manor	43.723125	-79.451589

Visual Analytics:

Next, we create Map of the North York neighbourhood in Toronto using Foursquare API.

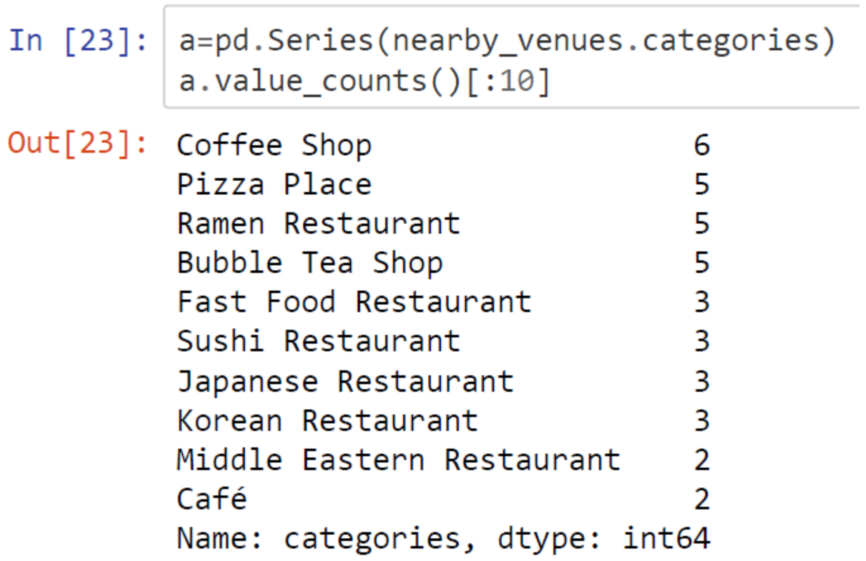

Out[15]:

For an immigrant it is very important to understand the nearby attraction and venues to be aware of her/his probable surroundings. So, we study the nearby attractions in North York, Toronto.


Out[21]:

	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	The Captain's Boil	[{'id': '4bf58dd8d48988d1ce941735', 'name': 'S...	43.773255	-79.413805
1	Aroma Espresso Bar	[{'id': '4bf58dd8d48988d16d941735', 'name': 'C...	43.769449	-79.413081
2	Starbucks	[{'id': '4bf58dd8d48988d1e0931735', 'name': 'C...	43.768353	-79.413046
3	Konjiki Ramen	[{'id': '55a59bace4b013909087cb24', 'name': 'R...	43.766998	-79.412222
4	The Keg	[{'id': '4bf58dd8d48988d1cc941735', 'name': 'S...	43.766579	-79.412131

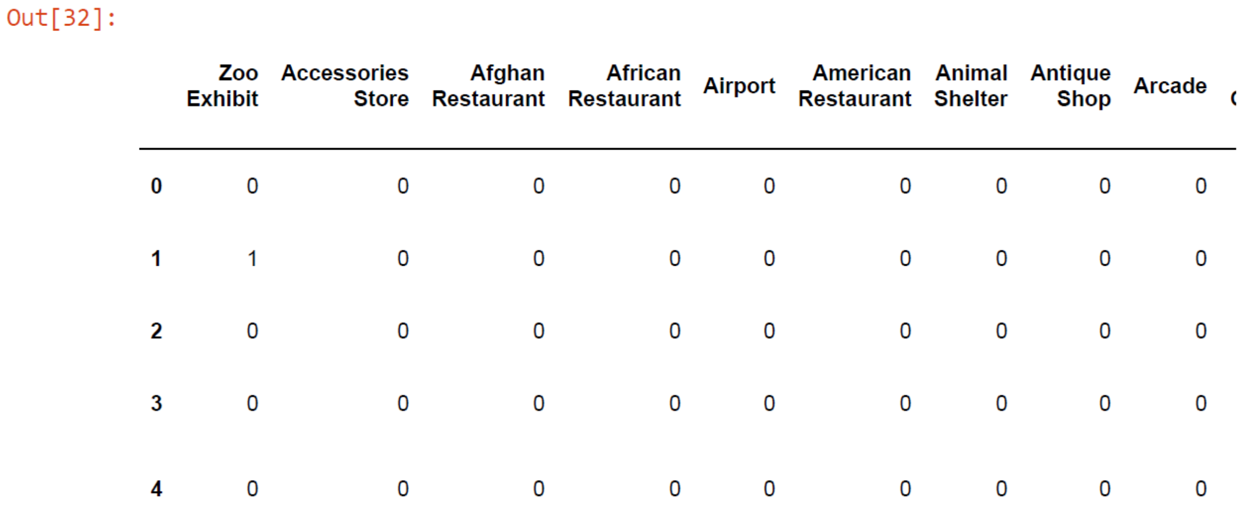
Though our selection of neighbourhood is dependent on housing price and school ratings, it might be worthwhile for an immigrant to get to know the popular nearby attractions. Hence, listing the top nearby attraction by categories in North York.



```
In [23]: a=pd.Series(nearby_venues.categories)
         a.value_counts()[:10]

Out[23]: Coffee Shop                  6
         Pizza Place                  5
         Ramen Restaurant             5
         Bubble Tea Shop              5
         Fast Food Restaurant         3
         Sushi Restaurant             3
         Japanese Restaurant          3
         Korean Restaurant            3
         Middle Eastern Restaurant    2
         Café                         2
         Name: categories, dtype: int64
```

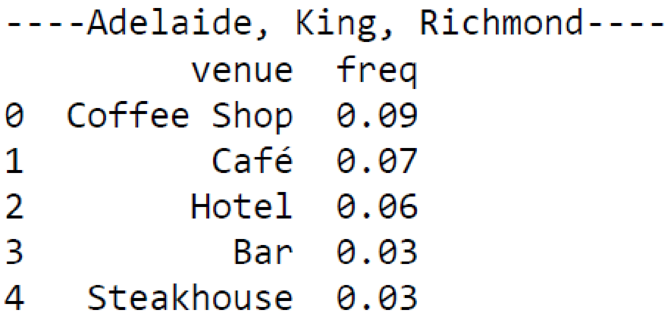
Next, we do One Hot Encoding of categorical variables like the features to Binary values. One hot encoding is a process by which categorical variables are converted into a form that could be provided to Machine Learning algorithms to do a better job in prediction. We also add neighborhood column back to data frame.



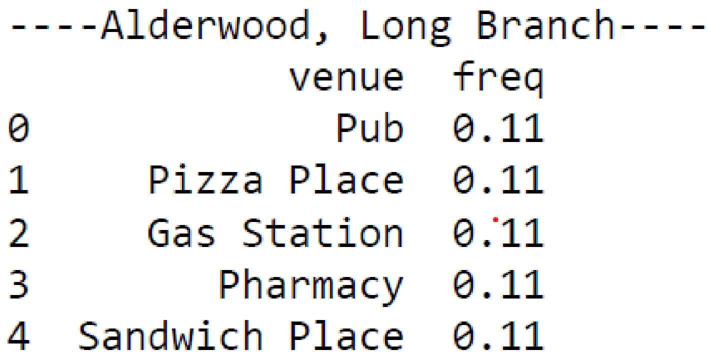
Out[32]:

	Zoo Exhibit	Accessories Store	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Animal Shelter	Antique Shop	Arcade
0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0

Then for each of the neighbourhood let us list the top 5 venue frequency - keeping in mind all options open for the immigrant while decision making. Some examples below:

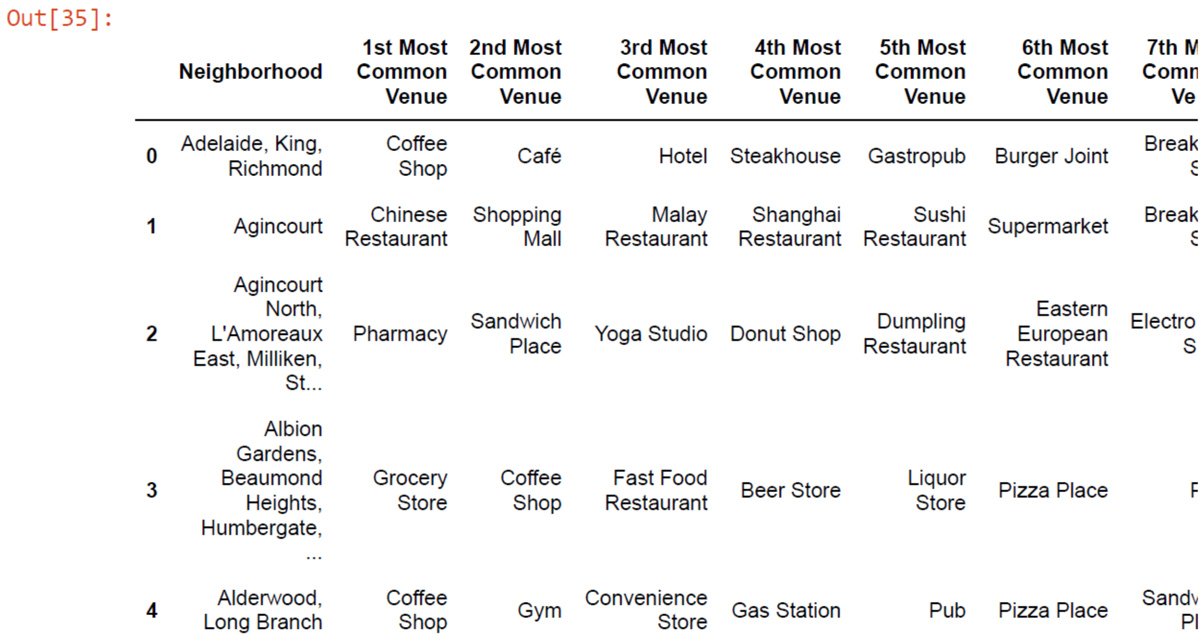


```
----Adelaide, King, Richmond----
         venue  freq
0  Coffee Shop  0.09
1         Café  0.07
2        Hotel  0.06
3          Bar  0.03
4   Steakhouse  0.03
```



```
----Alderwood, Long Branch----
            venue  freq
0             Pub  0.11
1     Pizza Place  0.11
2     Gas Station  0.11
3        Pharmacy  0.11
4  Sandwich Place  0.11
```

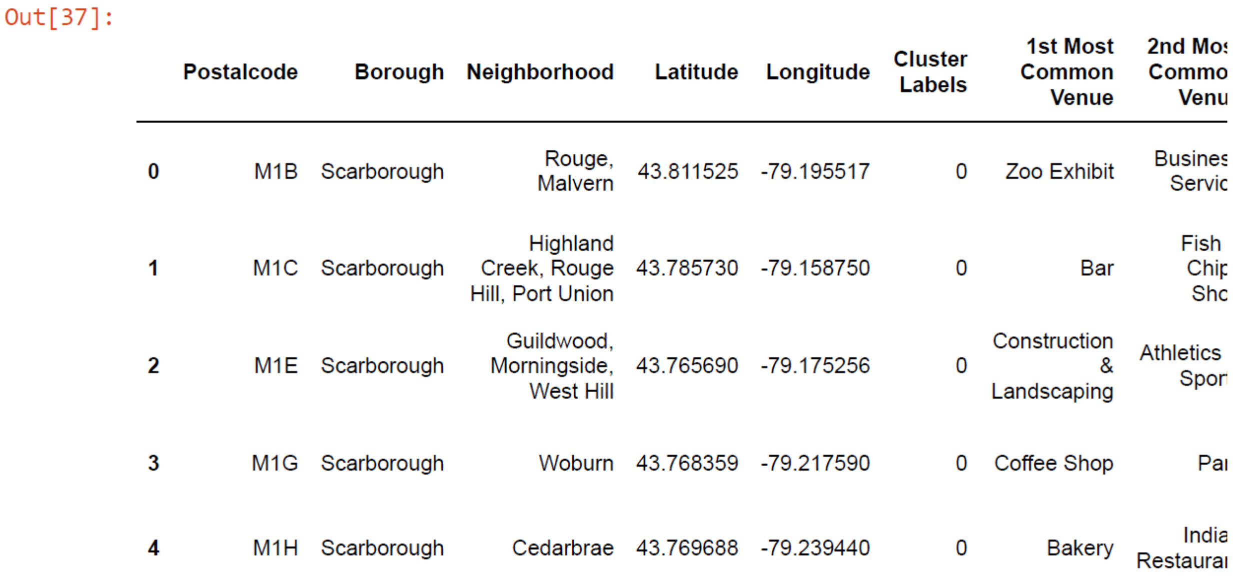
Next, we create a new data frame to get details of the most common venues near Neighborhood.



Out[35]:

	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 M Comn Ve
0	Adelaide, King, Richmond	Coffee Shop	Café	Hotel	Steakhouse	Gastropub	Burger Joint	Break S
1	Agincourt	Chinese Restaurant	Shopping Mall	Malay Restaurant	Shanghai Restaurant	Sushi Restaurant	Supermarket	Break S
2	Agincourt North, L'Amoreaux East, Milliken, St...	Pharmacy	Sandwich Place	Yoga Studio	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Electro S
3	Albion Gardens, Beaumond Heights, Humbergate, ...	Grocery Store	Coffee Shop	Fast Food Restaurant	Beer Store	Liquor Store	Pizza Place	F
4	Alderwood, Long Branch	Coffee Shop	Gym	Convenience Store	Gas Station	Pub	Pizza Place	Sandv Pl

Clustering of the neighbourhood by k-means is done. At first, we need to set number of clusters. For using k-means technique, let us assign the neighborhood into 3 clusters. Then, k-means clustering is run. We check cluster labels generated for each row in the data frame.

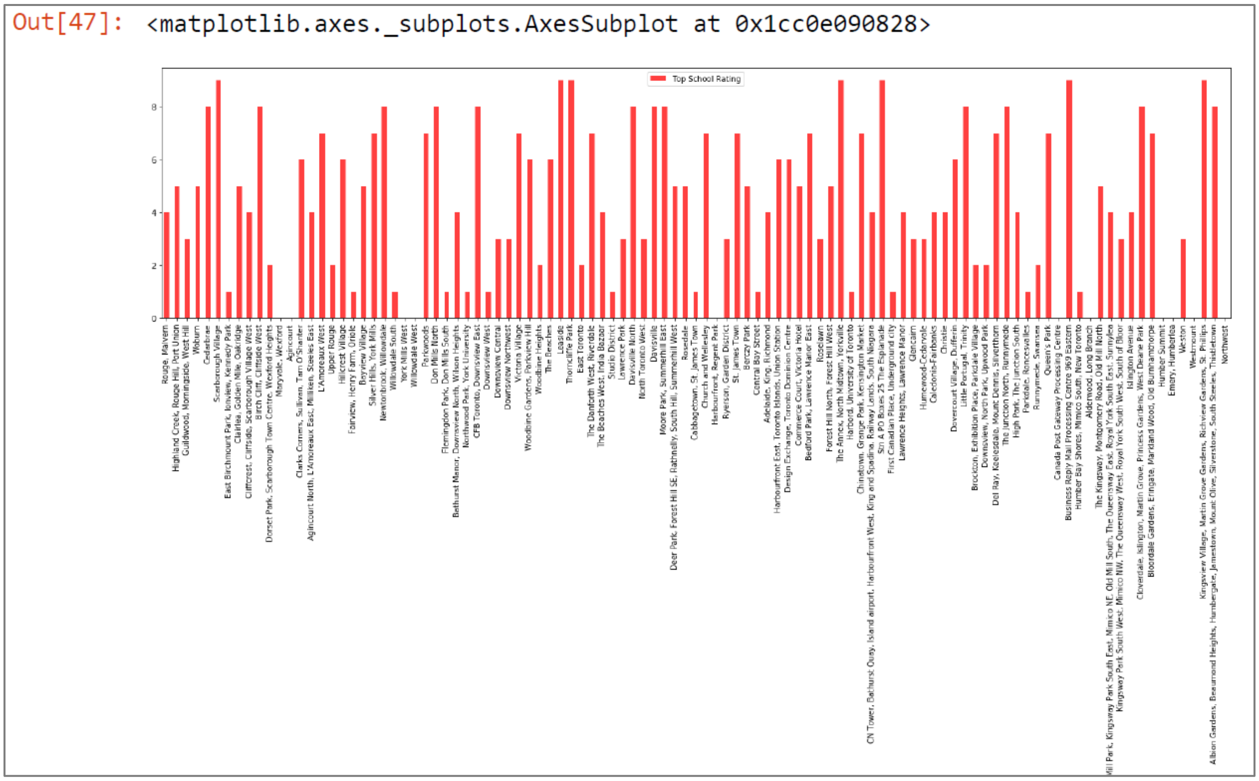


Out[37]:

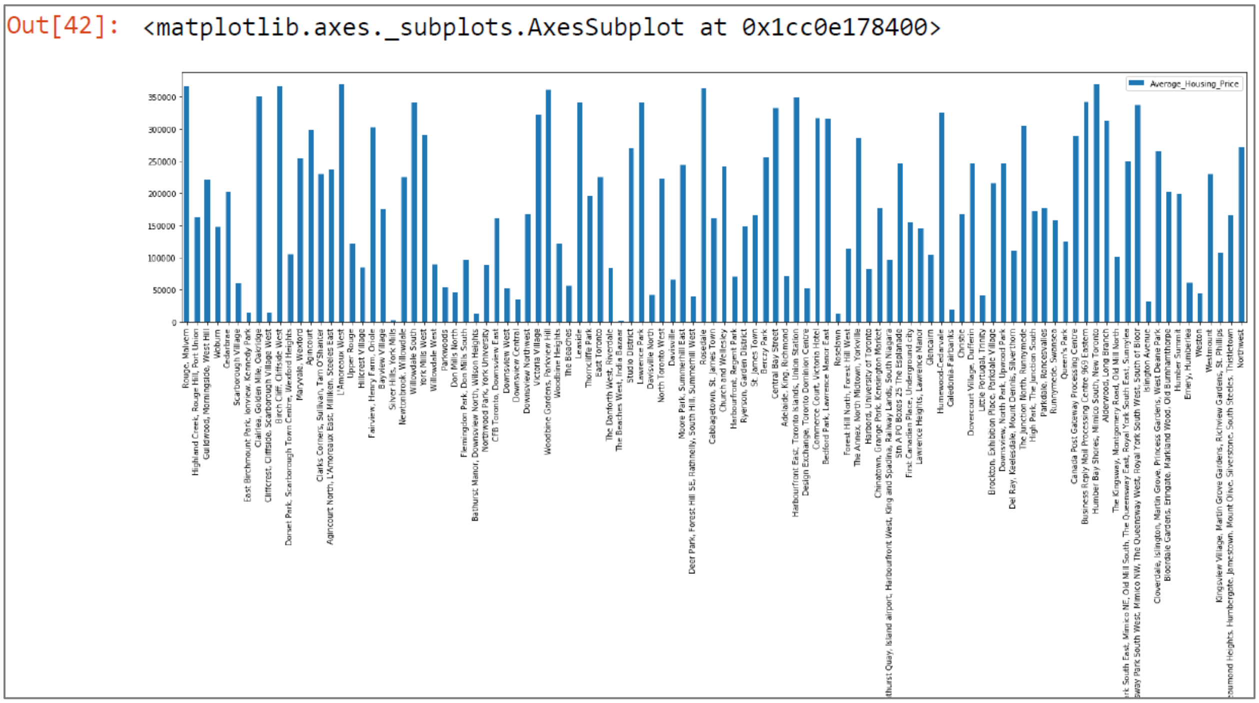
	Postalcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Mos Commo Venu
0	M1B	Scarborough	Rouge, Malvern	43.811525	-79.195517	0	Zoo Exhibit	Busines Servic
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.785730	-79.158750	0	Bar	Fish Chip Shc
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.765690	-79.175256	0	Construction & Landscaping	Athletics Spor
3	M1G	Scarborough	Woburn	43.768359	-79.217590	0	Coffee Shop	Pai
4	M1H	Scarborough	Cedarbrae	43.769688	-79.239440	0	Bakery	India Restaurai

Visual Analytics: Bar Chart of Average Housing Price for each of the neighbourhood were charted along with Bar Chart of Average School Ratings for each of the neighbourhood.

Secondary web research was done to find out the average housing prices and school ratings in the neighbourhoods of Toronto. As mentioned in the ‘Data section – as this information was not to be found in a singular website in a tabular format – various websites were referred. [see data section]

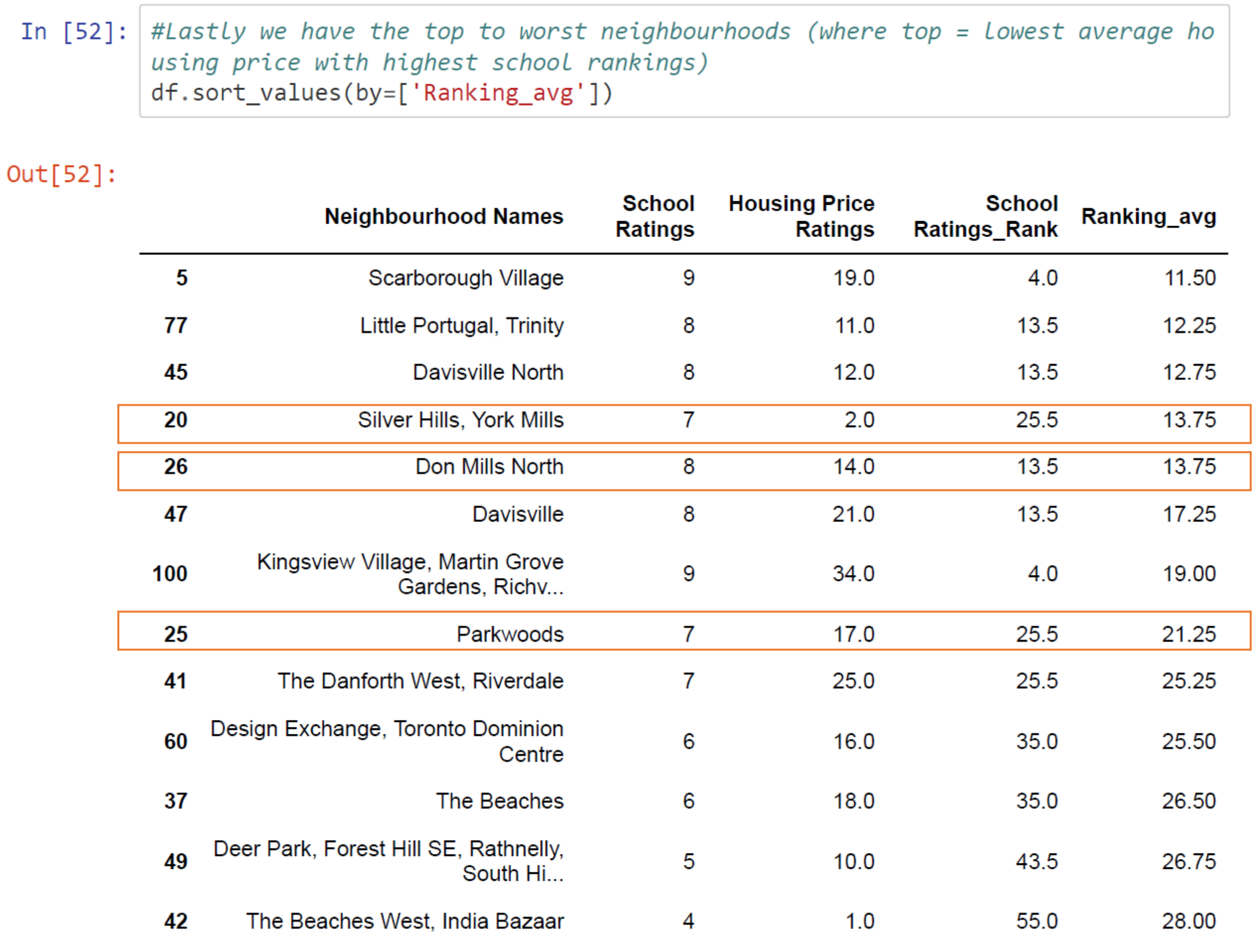


```
Out[47]: <matplotlib.axes._subplots.AxesSubplot at 0x1cc0e090828>
```



```
Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x1cc0e178400>
```

The above charts are good at providing a one shot visual when we look into the Jupiter Notebook or Python Kernel. However, to take a decision for an immigrant for staying in a neighborhood further analysis with both factors needs to be treated. Both the charts are a good visual of the two factors i.e. average housing price and school ratings of the neighborhoods. However, as North York as a borough occurs in multiple post codes (M2, M3, M6, etc.) it brings in the chart neighborhoods from other borough also. So, we create a simple excel file with the 102 neighborhoods names listed above and two columns containing the average housing price and school ratings. Using ranking method, we arrive at the probable best neighborhoods in North York, Toronto.



In [52]:

```
#Lastly we have the top to worst neighbourhoods (where top = lowest average ho
using price with highest school rankings)
df.sort_values(by=['Ranking_avg'])
```

Out[52]:

	Neighbourhood Names	School Ratings	Housing Price Ratings	School Ratings_Rank	Ranking_avg
5	Scarborough Village	9	19.0	4.0	11.50
77	Little Portugal, Trinity	8	11.0	13.5	12.25
45	Davisville North	8	12.0	13.5	12.75
20	Silver Hills, York Mills	7	2.0	25.5	13.75
26	Don Mills North	8	14.0	13.5	13.75
47	Davisville	8	21.0	13.5	17.25
100	Kingsview Village, Martin Grove Gardens, Richv...	9	34.0	4.0	19.00
25	Parkwoods	7	17.0	25.5	21.25
41	The Danforth West, Riverdale	7	25.0	25.5	25.25
60	Design Exchange, Toronto Dominion Centre	6	16.0	35.0	25.50
37	The Beaches	6	18.0	35.0	26.50
49	Deer Park, Forest Hill SE, Rathnelly, South Hi...	5	10.0	43.5	26.75
42	The Beaches West, India Bazaar	4	1.0	55.0	28.00

So, the top 3 neighbourhoods in North York are Silver Hills, Don Mills North and Parkwoods.

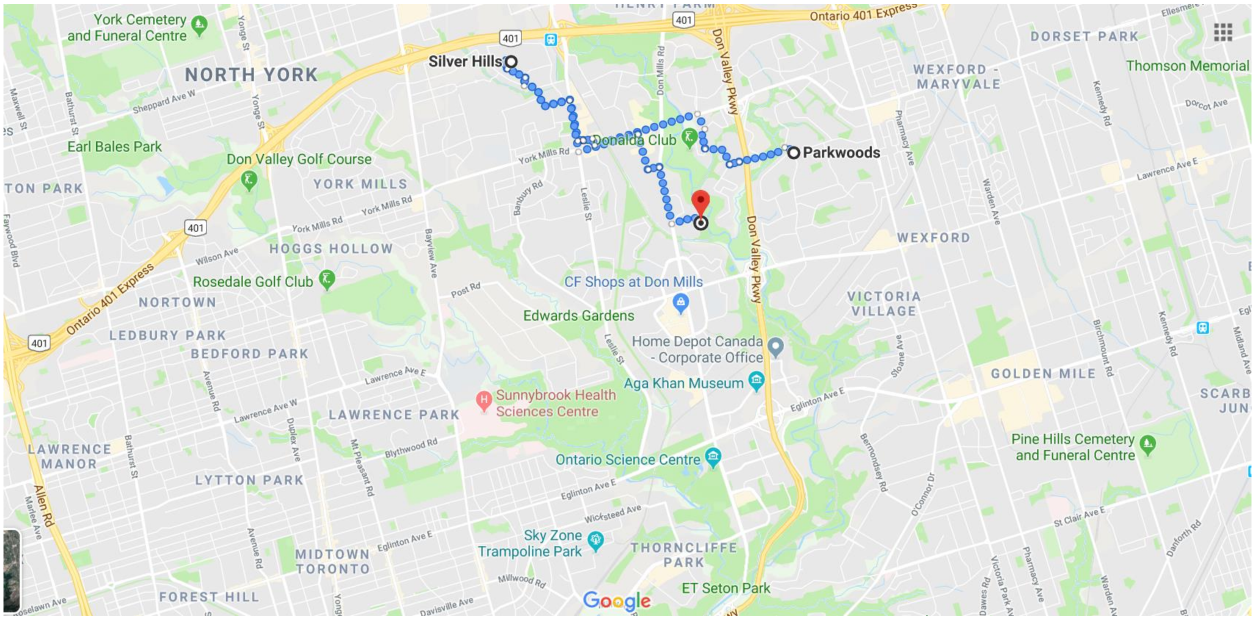
4. Discussion section

As a firm believer of fact-based decision making, this Capstone project was an eye opener for me. Through this project, I could address a macro problem yet touch individual lives. Knowing Canada to be an excellent host nation welcoming thousands of immigrant families across the globe this project is relevant and meaningful.

As you might have noticed the approach is practical and data intensive covering all aspects of settlement. It is a great assistance to a stranger and non-inhabitant migrating to a large city like Toronto. Toronto is a city with a high population and population density. It is very easy to get lost in this crowded city and take a wrong decision of landing in a not so economical neighbourhood. This in turn impacts the immigrant family economically and psychologically.

These where the K-means algorithm as part of this clustering study comes of value. When I tested the Elbow method using Python, I set the optimum k value to 3. However, only 102 neighbourhood coordinates were used. If we wish to be more detailed and have an accurate guidance, the data set can be expanded and the details of the neighborhood or street can also be drilled. Moreover, data analysis is also performed through this information by adding the coordinates of neighbourhood, average school ratings and home sales price averages as static data on GitHub. In the future, such data can also be processes and analyzed dynamically by number of platforms or packages.

Lastly, we finish the study by visualizing the data and clustering information on the Toronto map. The 3 selected neighbourhoods are depicted which have come out as the top places to stay for an immigrant based on school education quality and home prices in North York. Also, other attractions and nearby venues of these three neighbourhoods are also listed in the above section of ‘Results’.



Lastly, due to the flexibility of Python – as a data scientist I wish to stress that the entire coding/programming can be customized. Not only the cities can be changed but the neighbourhoods can be modified. So, the notebook in this project has a universal usage for the future. This is the essence learnt from this ‘Data Science’ course that any solution should be scalable and have wide adoption.

5. Conclusion section

To summarize, the scope of this of the analysis has great potential. The immigration issue will be more dynamic in the future and the information afforded us may be dated due to relying on personal user information via Foursquare. As it might have personal bias, that was not looked upon and was not treated. The accuracy of data depends purely depends on the data provided by FourSquare. Also, it must be noted that ‘average price’ as a factor is also changing periodically. Price is a determinant of multiple function. Overall though, the model created can easily be replicated again and used in the coming years. With the data analyzed and scoring system established the recommendations are good enough for the next couple of years.

Overall, I am highly impressed with the course, content and lab works presented during the Coursera IBM Certification Course. I believe this Capstone project was a great opportunity to practice and apply the Data Science tools and methodologies learned. With the concepts learnt over the last 8 courses in this professional certification I was able to create a real-life relevant project that showcased my data science skills and future potential.

Most importantly, I am confident now that I have gained necessary capabilities to become a professional Data Scientist and I am excited to explore practical cases in the future. I feel rewarded with the efforts, time and money spent. I highly appreciate and recommend this course with all the topics covered. I thank the faculty and the Coursera team for this.