Yahuza in Toronto- Choosing the best neighbourhood

Introduction

My client, Yahuza Suya, a Nigerian-Style barbeque brand are planning to open a new spot in Toronto, to cater to the growing Nigerian population in Canada and enter new markets. Their product, being mostly chicken and beef barbeque, will appeal to a wider market since chicken and beef is eaten mostly around the world.

Problem

While they are well funded, Yahuza Suya will like to open their spot where there is a higher probability of success. They want to ensure the neighbourhood will be a safe one, with less crime likely. Another important factor is that the pricing of the product must be within the income range as they do not want to be priced out of the market.

Stakeholders

Yahuza

Other businesspersons wanting to do business in Toronto

People looking for a home and want to understand the neighbourhoods

Looking for how to fit into Toronto.

DATA

1. We will need demographics data to get a snapshot of the neighbourhoods and their income ranges. This will help us sort and get the neighbourhoods with the highest incomes and those that follow.
2. We will need the crime data which will enable us sort and filter out the safest neighbourhoods.
3. List of Metro stations to understand the metro station distribution
4. Foursquare data showing us the neighbourhoods and the venues present there. This will help us understand the characteristics of the neighbourhoods and how our spot will fit in there.

Sources

General Toronto data was obtained from

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The demographic data was obtained from
https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

The crime data was obtained from

http://data.torontopolice.on.ca/datasets?q=crime

The list of Metro stations was obtained from

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

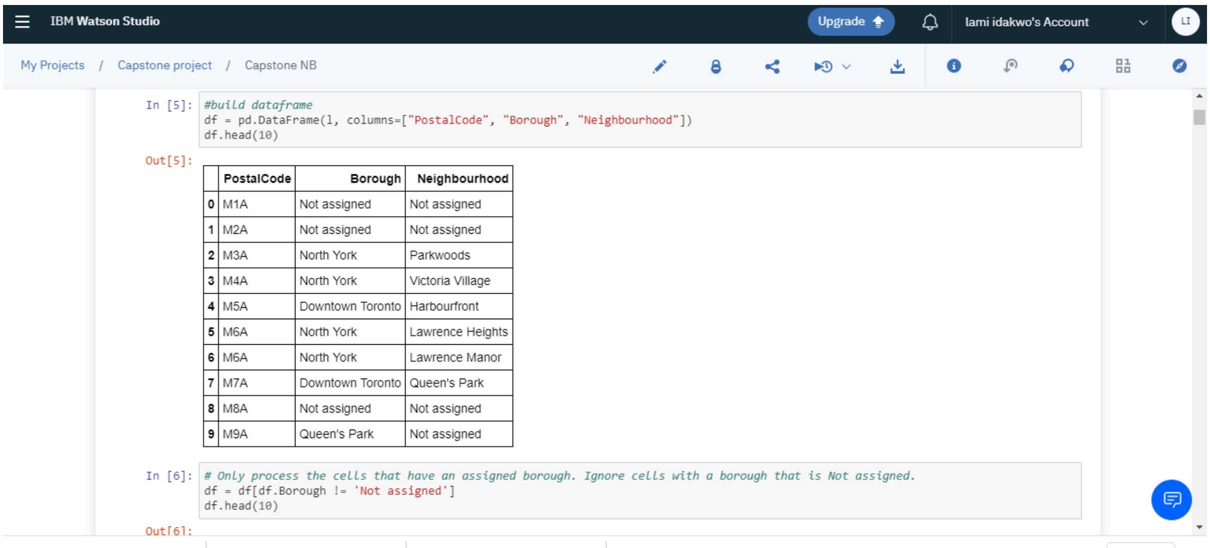
Methodology

We first import all necessary libraries. We ignore libraries like folium and sklearn until we need them so that it does not slow down the processing

Neigbourhood Data

At the basis of all the data is the list of boroughs and neighbourhoods with their postal codes. I used BeautifulSoup to scrape the data from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

I then used it to build a dataframe which looked like this:




```
#build dataframe
df = pd.DataFrame(l, columns=["PostalCode", "Borough", "Neighbourhood"])
df.head(10)
```

	PostalCode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M6A	North York	Lawrence Heights
6	M6A	North York	Lawrence Manor
7	M7A	Downtown Toronto	Queen's Park
8	M8A	Not assigned	Not assigned
9	M9A	Queen's Park	Not assigned

```
# Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.
df = df[df.Borough != 'Not assigned']
df.head(10)
```

To clean it up and ensure I don't have empty cells, I processed only the cells with assigned boroughs. I used the 'dropna()' function.

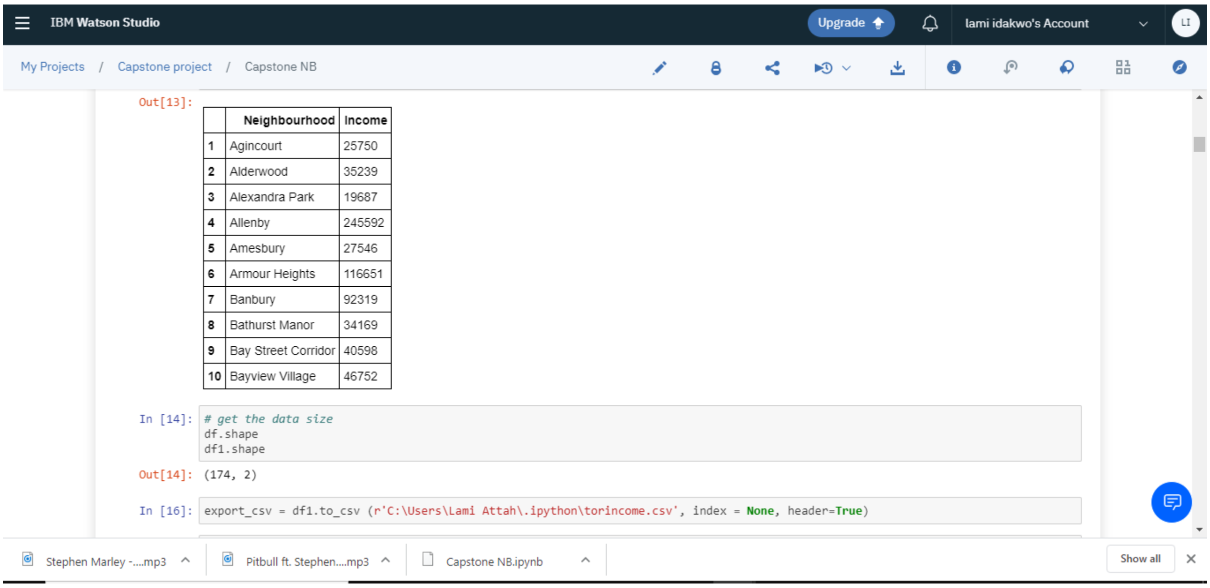
There were some boroughs with no neighbourhood. In such cases, I assigned the borough to the neighbourhoods. For neighbourhoods within the same borough and postal code, we grouped them together.

After doing this, we have a full dataframe with 103 rows and 3 columns.

Income data

To get the income data, I used BeautifulSoup again to scrape https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods and obtain the demographic data.

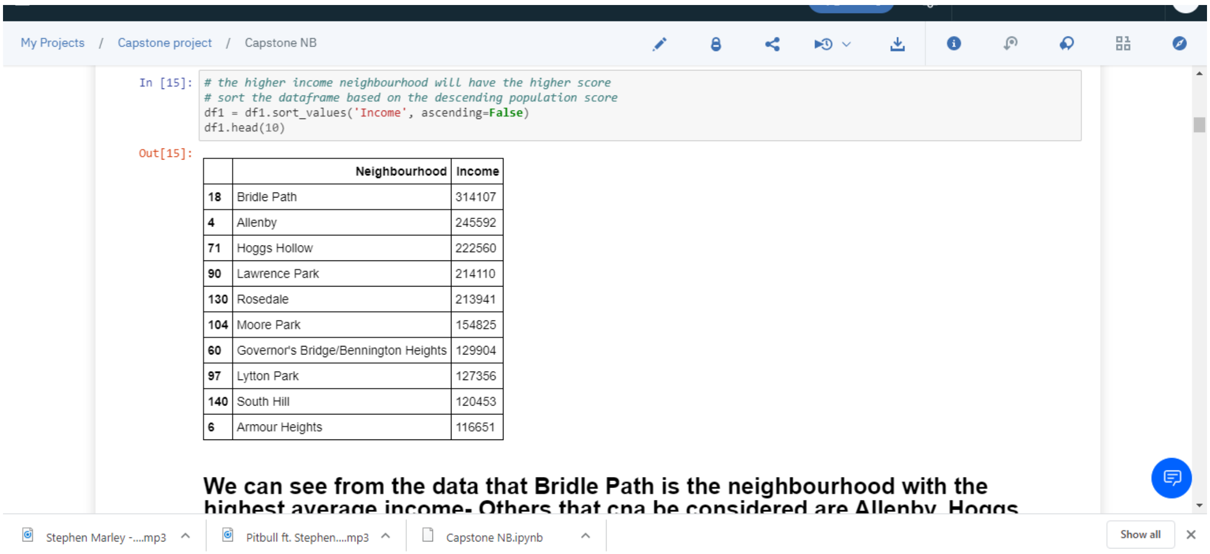
To use only the income, I dropped all other columns except the income column. This gave a dataframe that looked like this:



	Neighbourhood	Income
1	Agincourt	25750
2	Alderwood	35239
3	Alexandra Park	19687
4	Allenby	245592
5	Amesbury	27546
6	Armour Heights	116651
7	Banbury	92319
8	Bathurst Manor	34169
9	Bay Street Corridor	40598
10	Bayview Village	46752

```
In [14]: # get the data size
df.shape
df1.shape
Out[14]: (174, 2)
In [16]: export_csv = df1.to_csv (r'C:\Users\Lami Attah\.ipython\torincome.csv', index = None, header=True)
```


To get the neighbourhoods with the higher incomes, I sorted the dataframe to show the a descending order of the income. We can see the top ten (10) neighbourhoods with the highest average incomes. The dataframe looks like this:




```
In [15]: # the higher income neighbourhood will have the higher score
# sort the dataframe based on the descending population score
df1 = df1.sort_values('Income', ascending=False)
df1.head(10)
Out[15]:
```

	Neighbourhood	Income
18	Bridle Path	314107
4	Allenby	245592
71	Hoggs Hollow	222560
90	Lawrence Park	214110
130	Rosedale	213941
104	Moore Park	154825
60	Governor's Bridge/Bennington Heights	129904
97	Lytton Park	127356
140	South Hill	120453
6	Armour Heights	116651

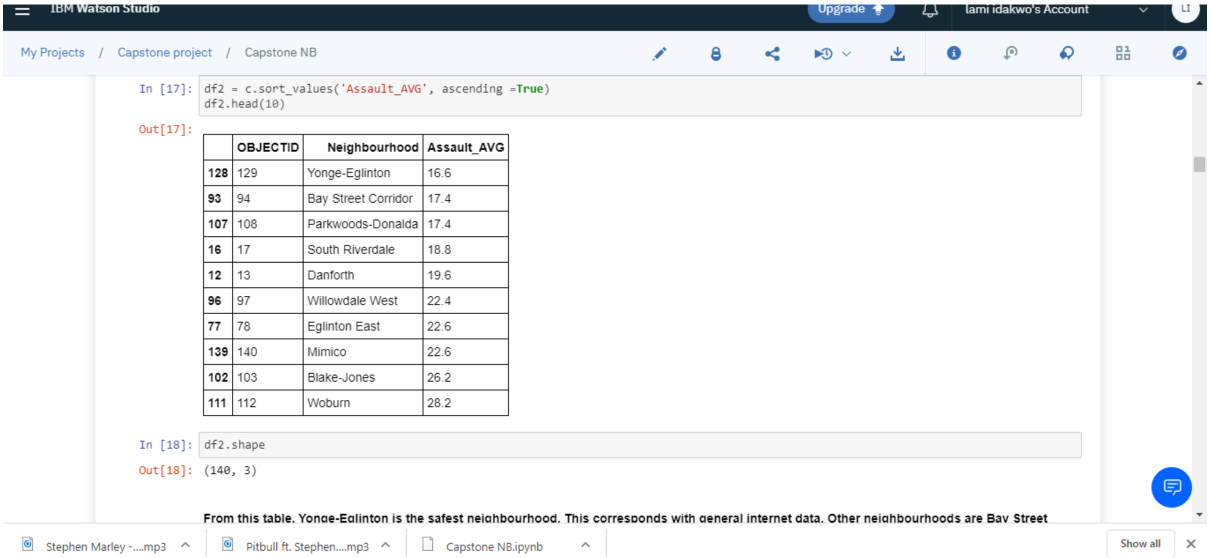
We can see from the data that Bridle Path is the neighbourhood with the highest average income- Others that cna be considered are Allenby, Hoggs

Crime data

The crime data was not importing from my computer so I decided to clean it up in Excel, then upload it to a website I own, and use read_csv to get the table.

I then sort the dataframe according to the assault average(assault is the most common form of crime in Toronto).

The dataframe looks like this :



In [17]:
```
df2 = c.sort_values('Assault_AVG', ascending =True)
df2.head(10)
```

Out[17]:

	OBJECTID	Neighbourhood	Assault_AVG
128	129	Yonge-Eglinton	16.6
93	94	Bay Street Corridor	17.4
107	108	Parkwoods-Donalda	17.4
16	17	South Riverdale	18.8
12	13	Danforth	19.6
96	97	Willowdale West	22.4
77	78	Eglinton East	22.6
139	140	Mimico	22.6
102	103	Blake-Jones	26.2
111	112	Woburn	28.2

In [18]:
```
df2.shape
```

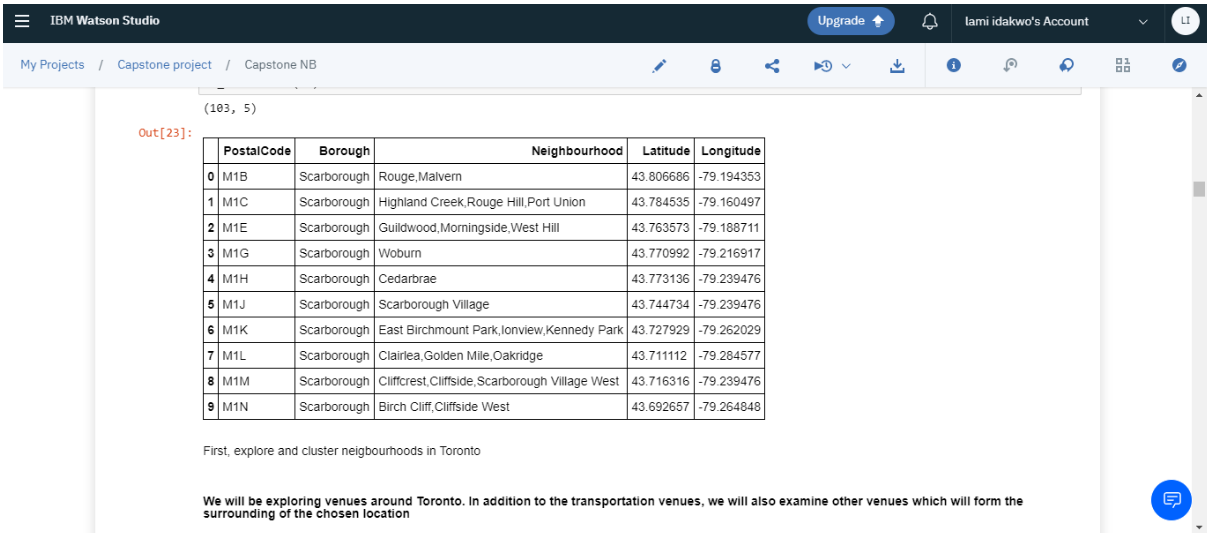
Out[18]: (140, 3)

From this table, Yonge-Eglinton is the safest neighbourhood. This corresponds with general internet data. Other neighbourhoods are Bay Street

We see the top ten safest neighbourhoods.

To complement the borough and neighbourhood dataframe and enrich our data, we obtain the coordinates and merge the two dataframes. We now have a dataframe that has the postal codes, neighbourhood, boroughs, and coordinates (latitude and longitude).

The dataframe looks like this:



(103, 5)

Out[23]:

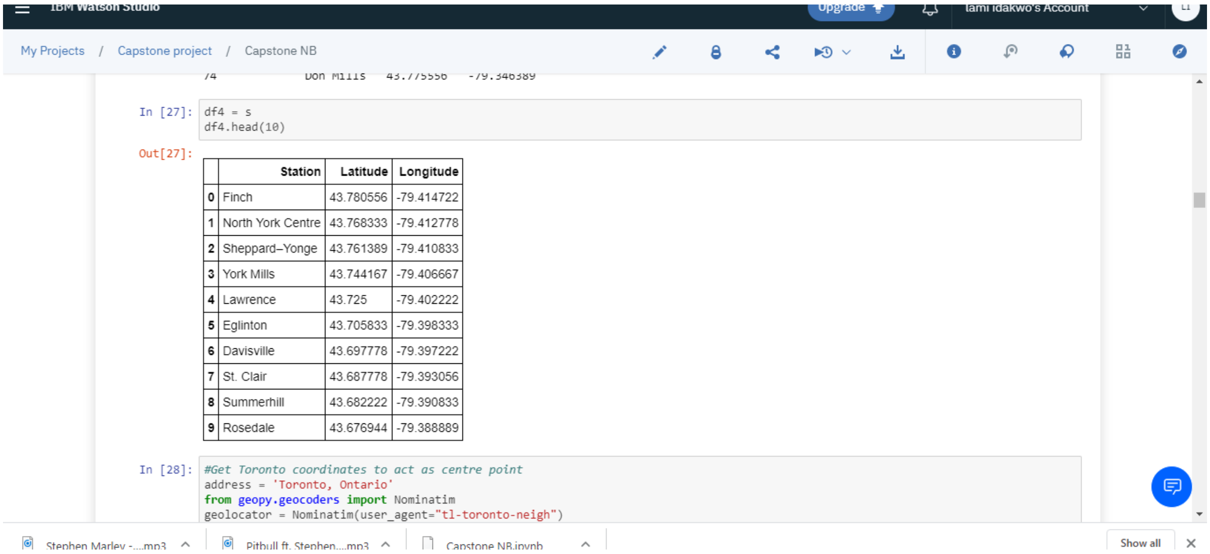
	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	East Birchmount Park,Ionview,Kennedy Park	43.727929	-79.262029
7	M1L	Scarborough	Clairlea,Golden Mile,Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffcrest,Cliffside,Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff,Cliffside West	43.692657	-79.264848

First, explore and cluster neigbourhoods in Toronto

We will be exploring venues around Toronto. In addition to the transportation venues, we will also examine other venues which will form the surrounding of the chosen location

Next, we obtained the list of Subway stations to understand their spread across Toronto. We know that the airport is a central point and that most of Toronto is not far from Pearson YYZ.

We clean the data in Excel and add the coordinate data. We have a dataframe that looks like:



```
In [27]: df4 = s
         df4.head(10)
```

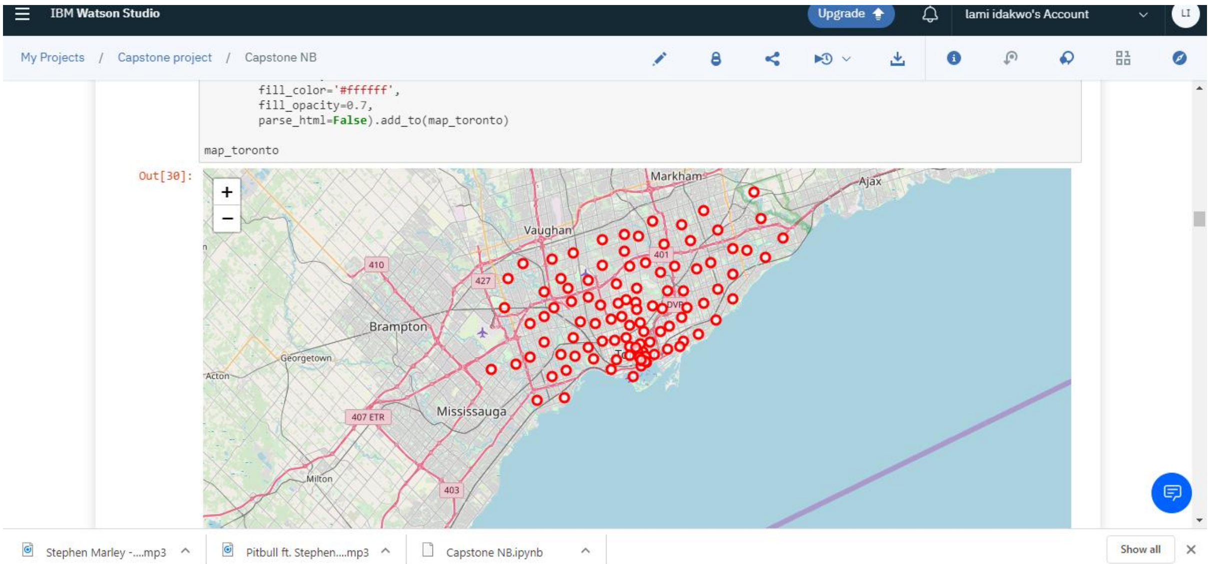
Out[27]:

	Station	Latitude	Longitude
0	Finch	43.780556	-79.414722
1	North York Centre	43.768333	-79.412778
2	Sheppard–Yonge	43.761389	-79.410833
3	York Mills	43.744167	-79.406667
4	Lawrence	43.725	-79.402222
5	Eglinton	43.705833	-79.398333
6	Davisville	43.697778	-79.397222
7	St. Clair	43.687778	-79.393056
8	Summerhill	43.682222	-79.390833
9	Rosedale	43.676944	-79.388889

```
In [28]: #Get Toronto coordinates to act as centre point
         address = 'Toronto, Ontario'
         from geopy.geocoders import Nominatim
         geolocator = Nominatim(user_agent="tl-toronto-neigh")
```

Next, we import folium and use it to visualize the Toronto map.

It looks like this:



We also visualize the Subway stations in Toronto to see the spread. We can see that the subway stations are well spread but we take note of a central point where we have most stations. The visualization is below:

```
        fill_color='#006633',
        fill_opacity=0.7,
        parse_html=False).add_to(map_metrotoronto)

map_metrotoronto
```

Out[31]:

FOURSQUARE API:

We then use Foursquare API to obtain the venues in Toronto and observe. We decided to look for general venues and not the metro stations because we can see that the subway stations are well-spread. In addition, the metro stations form a part of the foursquare results and it will be important to use the other venues to get a full understanding of the neighbourhoods.

Results:

```
(4909, 9)
```

```
In [47]: toronto_venues.groupby('Neighbourhood').count().head()
```

Out[47]:

Neighbourhood	PostalCode	Borough	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Adelaide,King,Richmond	100	100	100	100	100	100	100	100
Agincourt	45	45	45	45	45	45	45	45
Agincourt North,L'Amoreaux East,Milliken,Steeles East	29	29	29	29	29	29	29	29
Albion Gardens,Beaumond Heights,Humbergate,Jamestown,Mount Olive,Silverstone,South Steeles,Thistletown	18	18	18	18	18	18	18	18
Alderwood,Long Branch	28	28	28	28	28	28	28	28

```
In [48]: print('There are {} uniques categories.'.format(len(toronto_venues['Venue Category'].unique())))

There are 333 uniques categories.

In [49]: # one hot encoding to perform k-means clustering
```

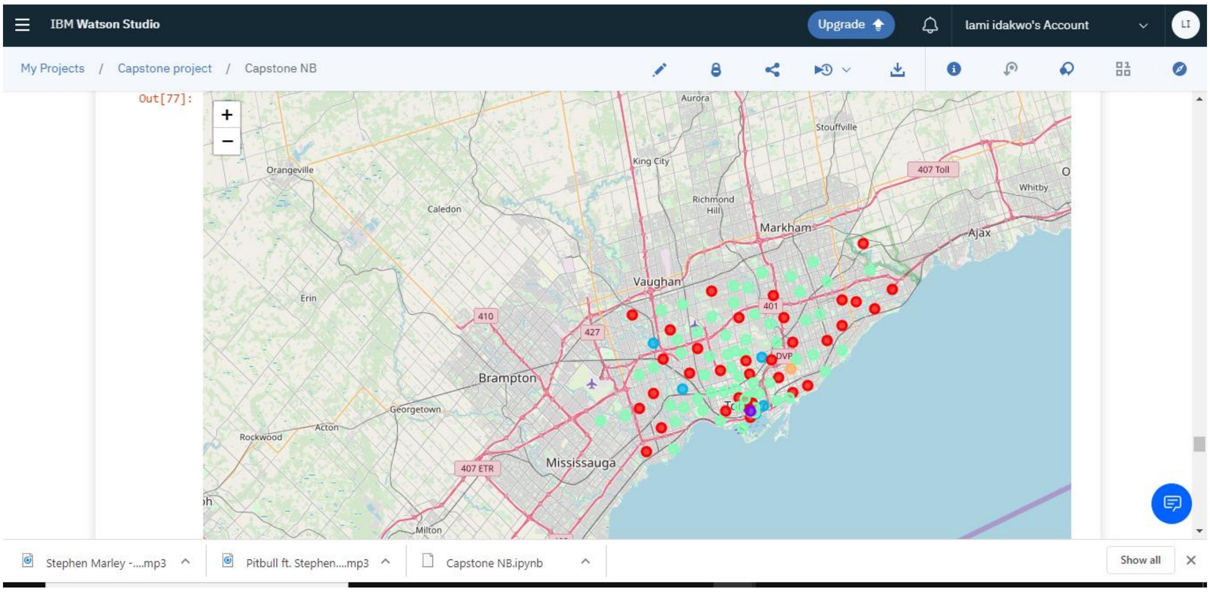
In order to perform k-means clustering, we used one-hot encoding.

We then print each neighbourhood according to the most popular venues. We then obtain the 1st ten(10) most common venues for each neighbourhood.

K-means

To perform k-means, we take our k=5. We will be observing 5 clusters. We are using k-means so that we can have a statistical visual of the characteristics of the neighbourhood. Such a view will bring certain things to light which we may have omitted in looking at income and crime data.

After perfoming k-means, we visualize the 5 clusters shown below:



We then begin to observe each clusters:

Cluster 0- We see that cluster 0 has very many spots. It also has bus stations, trains stations and metro stations among its top ten most common place. It appears to be favourable for our business line. However, it may be saturated and there may be too much competition.

Cluster 1 – Cluster one is just one neighbourhood which has foreign food restaurants and other **services. However, it does not have transportation services in its top ten common places.**

Cluster 2- While cluster two has a few varieties of venues, it seems to be quite a homely cluster, most likely family-oriented. It also does not have transportation services as common venues.

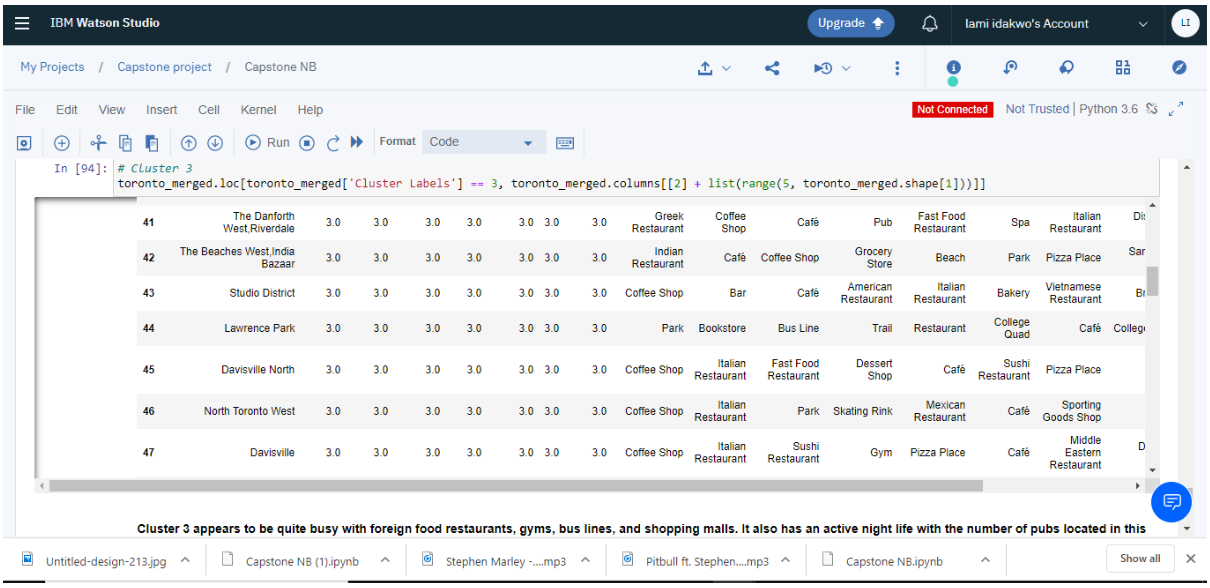
Cluster 3- Cluster 3 appears to have many spots, including transportation services, it also appears to cater to a large number of foreign nationals with the amount of foreign food restaurants around there.

Cluster 4- Cluster 4 is a bit similar to cluster 1 and appears to be focused on a certain type of market.

Discussion:

Based on the results obtained, I closed in on cluster 3, which clusters neighbourhoods with the kind of venues we will expect around our own. It also has transportation services as many of the top ten most common venues which is one of our main factors. From cluster 3, I look out for neighbourhoods which satisfy our factors- (safety and income average).

Cluster 3 is shown below:




```
In [94]: # Cluster 3
toronto_merged.loc[toronto_merged['Cluster Labels'] == 3, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

41	The Danforth West,Riverdale	3.0	3.0	3.0	3.0	3.0	3.0	3.0	Greek Restaurant	Coffee Shop	Café	Pub	Fast Food Restaurant	Spa	Italian Restaurant	Di
42	The Beaches West,India Bazaar	3.0	3.0	3.0	3.0	3.0	3.0	3.0	Indian Restaurant	Café	Coffee Shop	Grocery Store	Beach	Park	Pizza Place	Sar
43	Studio District	3.0	3.0	3.0	3.0	3.0	3.0	3.0	Coffee Shop	Bar	Café	American Restaurant	Italian Restaurant	Bakery	Vietnamese Restaurant	Br
44	Lawrence Park	3.0	3.0	3.0	3.0	3.0	3.0	3.0	Park	Bookstore	Bus Line	Trail	Restaurant	College Quad	Café	Colleg
45	Davisville North	3.0	3.0	3.0	3.0	3.0	3.0	3.0	Coffee Shop	Italian Restaurant	Fast Food Restaurant	Dessert Shop	Café	Sushi Restaurant	Pizza Place	
46	North Toronto West	3.0	3.0	3.0	3.0	3.0	3.0	3.0	Coffee Shop	Italian Restaurant	Park	Skating Rink	Mexican Restaurant	Café	Sporting Goods Shop	
47	Davisville	3.0	3.0	3.0	3.0	3.0	3.0	3.0	Coffee Shop	Italian Restaurant	Sushi Restaurant	Gym	Pizza Place	Café	Middle Eastern Restaurant	D

Cluster 3 appears to be quite busy with foreign food restaurants, gyms, bus lines, and shopping malls. It also has an active night life with the number of pubs located in this

On observation, I pick these six(6) neighbourhoods:

Lawrence Park

Rosedale

Parkwoods

Willowdale West

Riverdale West

Mimico

To narrow it down to three and have more specific results, I look closely and see that:

1.**Lawrence Park** satisfies all the three requirements and is close to a college. This will attract young adults who have an active nightlife.

2. **Willowdale West** is also a good idea as it satisfies the requirements.

3. The third option will be **Parkwoods**, which also satisfies the requirement of safety and income. Further search reveals there are many immigrants.

Conclusion

Choosing a good neighbourhood usually depends on more than one factor. In most cases, it is usually a healthy combination of important factors.

For this task, while safety and pricing were major factors, it would have been counterproductive to choose based on these two factors alone. The use of foursquare data and the venues provided gave us a more descriptive view of these neighbourhoods and we can make a more robust decision.

Such tasks show the important of Data Science as it enables fact-based decision-making, even if it means coming to the same assumption.