****The following is a novel unsupervised machine learning project*

1. Introduction

Canada is one of the most eclectic countries in the world. As such entrepreneurs have diverse scope in terms of the kind of food & beverage ventures they can undertake within neighborhoods in Toronto. As of 2016 Toronto has a population in excess of 5 million. The population, in particular the population density per square kms, and the presence of other food & beverage companies in a given neighborhood in Toronto will determine the feasibility of opening a new restaurant in Toronto.

If any given Toronto neighborhood has a high density of restaurants per km and these restaurants are already well established and liked by the community then opening another restaurant in that community would be ill-advised. Further, if a given neighborhood has a below average population/density this would lessen Person X's potential customer base, and by extension lessen his potential revenue which in turn will affect profitability.

For this project we assume Person X is considering opening a restaurant in the Harbard and Harbor/University of Toronto neighborhoods. In this report we discuss the intuitive feasibility of undertaking such a venture by drawing on unsupervised machine learning techniques viz., clustering.

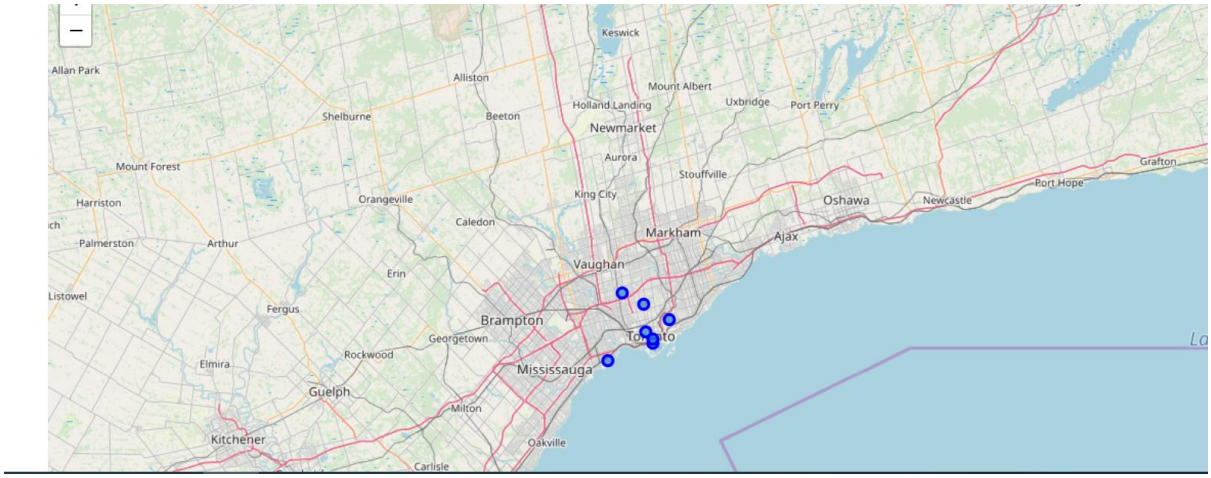
2. Data

To pull postal code data of all the boroughs and neighborhoods in Canada we web scrape the Canada wikipedia page. We remove all end of line new line escape characters in the last column by slicing in python. The final table is put into pandas dataframe;

Out[27]:

	Postcode	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M5A	Downtown Toronto	Regent Park
6	M6A	North York	Lawrence Heights
7	M6A	North York	Lawrence Manor
8	M7A	Queen's Park	Not assigned
10	M9A	Etobicoke	Islington Avenue
11	M1B	Scarborough	Rouge
12	M1B	Scarborough	Malvern
14	M3B	North York	Don Mills North

We then use the geocoder python package to get the longitude and latitude of every postal code and append the values to the dataframe. After that we filter the dataframe to get all neighborhoods in Toronto and display this area on a map using folium;



We then used the Fouresquare API to get the most common venue categories in all the Toronto neighborhoods. The foursquare location data is always returned in json format, for our purposes the keys we are interested in are the “response” and “venue” keys. Within those keys we are interested in the “name”, “categories” and “location” keys, so we need to do a bit of data engineering and cleaning to filter the result json. The categories key will have all the data about the category of venues in a given location/neighborhood.

3. Methodology

After we filter the json for all the venues in Toronto neighborhoods we use onehot encoding to count the frequency of each category of venue in all the Toronto neighborhoods.

Out[54]:

	Yoga Studio	Airport	American Restaurant	Aquarium	Art Gallery	Asian Restaurant	Bakery	Bar	Baseball Stadium	Basketball Stadium	...	Steakhouse	Supermarket	Sushi Restaurant	Tea Room	Thai Restaurant	Theater	Train Station	Vegetarian / Vegan Restaurant
0	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

5 rows × 98 columns

We then print the top 5 venue categories by frequency percentage;

```
num_top_venues = 5


for hood in toronto_grouped['Neighborhood']:
    print("----"+hood+"----")
    temp = toronto_grouped[toronto_grouped['Neighborhood'] ==
hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq',
ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

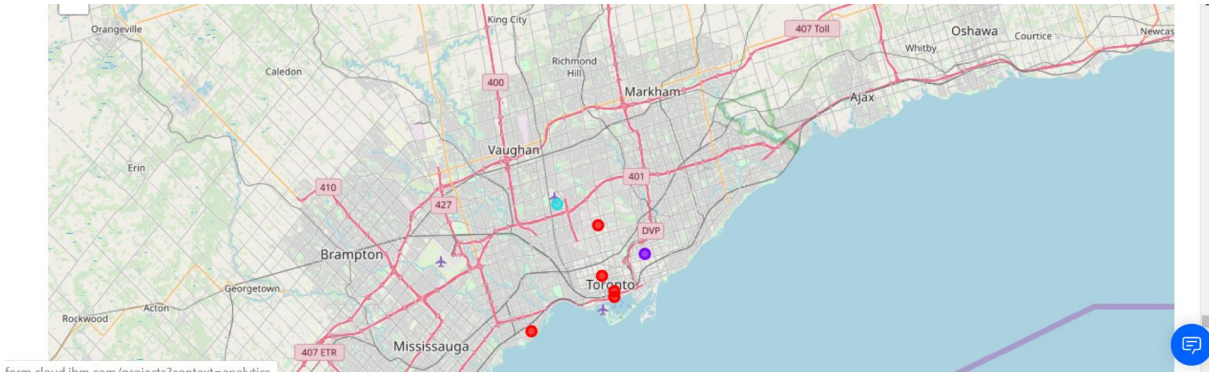
And the output;

```
----CFB Toronto, Downsview East----
                      venue  freq
0                   Airport  0.25
1                      Park  0.25
2                Playground  0.25
3  Construction & Landscaping  0.25
4                      Lake  0.00


----Design Exchange, Toronto Dominion Centre----
                venue  freq
0         Coffee Shop  0.15
1                Café  0.08
2               Hotel  0.06
3          Restaurant  0.04
4  American Restaurant  0.04
```

```
 ----East Toronto----
               venue  freq
0               Park  0.50
1  Convenience Store  0.25
2        Coffee Shop  0.25
3        Yoga Studio  0.00
4             Office  0.00


 ----Harbord, University of Toronto----
                venue  freq
0               Café  0.14
1         Restaurant  0.06
2             Bakery  0.06
3           Bookstore  0.06
4  Italian Restaurant  0.06


 ----Harbourfront East, Toronto Islands, Union Station----
         venue  freq
0  Coffee Shop  0.13
1     Aquarium  0.05
2        Hotel  0.05
3         Café  0.04
4      Brewery  0.03
```

```
----Humber Bay Shores, Mimico South, New Toronto----
                venue  freq
0          Restaurant  0.08
1  Fried Chicken Joint  0.08
2  American Restaurant  0.08
3         Liquor Store  0.08
4               Bakery  0.08


----North Toronto West----
                venue  freq
0         Coffee Shop  0.10
1       Clothing Store  0.10
2  Sporting Goods Shop  0.10
3          Yoga Studio  0.05
4         Burger Joint  0.05
```

We then run k-means to cluster the Toronto neighborhoods into 4 clusters. We display the clusters superimposed on a map using folium;



After examining the clusters we notice that the location Person X is considering as the location for his new restaurant falls into cluster 0 (the red dots) and we call this the coffee shop cluster;

4. Examine Clusters

Now, we examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories.

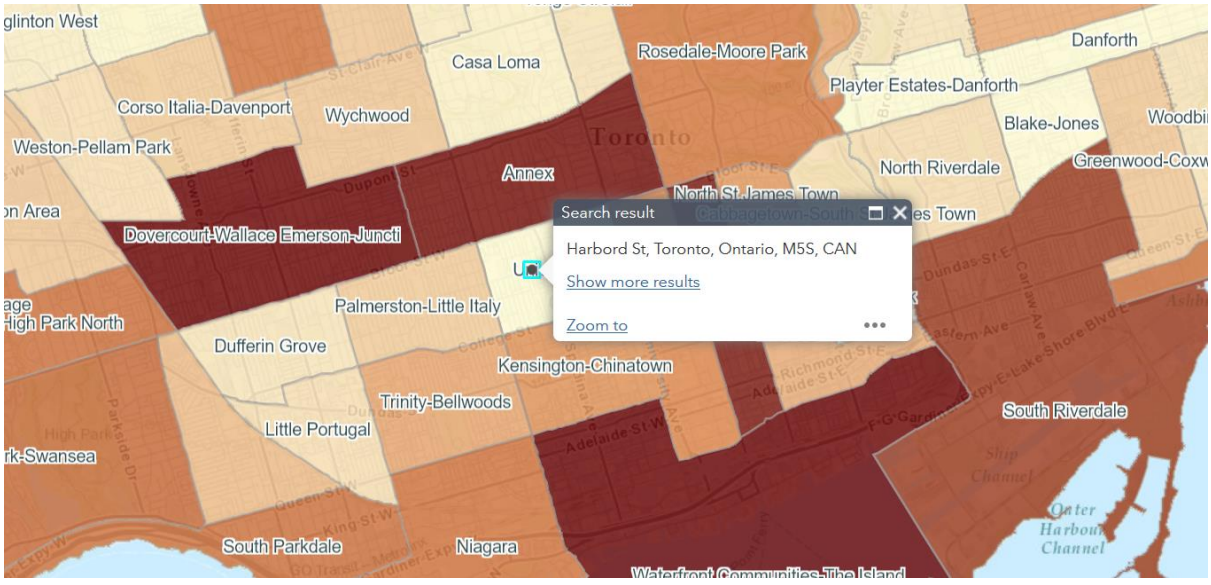
```
[71]: toronto_merged.loc[toronto_merged['Cluster Labels'] == 0, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1]))]]
```

Out[71]:

	Borough	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
2	Central Toronto	0	Coffee Shop	Clothing Store	Sporting Goods Shop	Yoga Studio	Mexican Restaurant	Rental Car Location	Restaurant	Chinese Restaurant	Café	Salon / Barbershop
3	Downtown Toronto	0	Coffee Shop	Aquarium	Hotel	Café	Italian Restaurant	Brewery	Scenic Lookout	Fried Chicken Joint	Restaurant	Sports Bar
4	Downtown Toronto	0	Coffee Shop	Café	Hotel	Restaurant	American Restaurant	Bar	Italian Restaurant	Gastropub	Deli / Bodega	Seafood Restaurant
5	Downtown Toronto	0	Café	Bakery	Bar	Bookstore	Italian Restaurant	Restaurant	Japanese Restaurant	Coffee Shop	Poutine Place	Sandwich Place
6	Etobicoke	0	Pharmacy	Restaurant	Sandwich Place	Café	Pizza Place	Fast Food Restaurant	Liquor Store	Fried Chicken Joint	Bakery	Gym

4. Results

From our analysis we can see that Harbard and University of Toronto neighborhoods fall into the coffee shop cluster. In this cluster the 1st most common venue are coffee shops, then cafe's. The cluster does not have a high density of restaurants. Prima facie this might seem like an optimal location to place a restaurant however if we observe a population density map of the University of Toronto neighborhood we can see that this cluster has very low population density and most of the directly adjacent neighborhoods have low population density as well;



Further, the majority of person X's customer base in this area would presumably consist of University students. These people would have much rather buy restaurant food from cheaper alternatives on campus as they have limited budgets.

5. Discussion

Most of the neighborhoods in Toronto already have established restaurants. Coffee shops and restaurants feature as the highest occurring venue in all the boroughs. This coupled with the

fact that the entrepreneur's chosen neighborhood is low population dense means opening a restaurant in Toronto is not a good idea.

However we notice that clothing stores are not high occurring venues in all but one of the Toronto boroughs. So maybe the entrepreneur should consider switching his line of business. The clothing store market in Toronto is less competitive and non-monopolistic so the entrepreneur will not face any barriers to entry.

6. Conclusion

This analysis will equip the entrepreneur with the relevant information to make an intuitive decision on what type of shop he should open in any given Toronto neighborhood. The good thing about Toronto is that its very diverse, so if the entrepreneur can find a niche and figure out innovative ways to promote his product, he will not struggle to build a customer base quickly.