TORONTO CLUSTERING – TO OPEN A CAFÉ IN SCARBOROUGH

- ## Business Problem

  Scarborough is a former municipality and an administrative division of Toronto, Ontario, Canada. Situated atop the Scarborough Bluffs, it occupies the eastern part of the city. It contains the Scarborough Bluffs, huge cliffs overlooking Lake Ontario, lined with parks, beaches, and hiking trails. The area is also known for its **diverse spread of restaurants**, including regional Southeast Asian, Chinese, and **Indian cuisine**.

  Our client, from India, wants to open a dessert café in Scarborough named "Sweet Tooth Cafe". He is excited to open the café as new immigrants are always welcomed in Scarborough. His idea is to make and sell famous & favorite homemade dishes from all the corners of the world along with desserts too. It will represent idea of cultural diversity.

  His idea is to open this café in a location which is accessible by young people, working class and old people. It should be located in an area which should not be far away from the city. It should be in the middle of the town. A place whose permits will be easily available and rent shouldn't be high enough. The neighborhood of the café should have other cafes or restaurants but not more than 5. It can contain other shops so that customers will be available. He wants to ensure that the neighborhood of the café should not get disturbed because of it.

  This are the requirements of the client to open café in Scarborough.

- ## Source & Usage of Data

  The majority of the data will be covered from Foursquare API. We will need geospatial data of Scarborough, Toronto for longitudes and latitudes of registered regions.

  From Foursquare API, we will call venues filtered with specialities like dessert, restaurants, cafes and so on. That data shall be from Scarborough, Toronto. We will also require data of different types of shops in the neighborhood of the finalized area. Merged dataset will be created from both datasets and then analysis will be done.

  For finalizing the location, we tend to ensure either government permission is there to run a café or not. And moreover, its neighborhood is fine with café or not. We need to make sure the eatery shops should not be more than 5 in the nearby neighborhood as per client's requirements. The data should say whether its far away from the town or not. The location of the café should be in the middle of the town accessible by people from all the corners.

  Data for permission of government will be got from Canadian Government website. Or it can be asked from official as well. Data stating which dishes to be sold according to its visitors and residents will also be needed. That can be taken with help of Foursquare API by calling tips of users on venues. Or a survey can be conducted for the same which will be helpful in many ways.

- ## Exploratory Data Analysis & Model Development

  First, we imported location data of Toronto. Cleaning was required in that dataset. It was done using dropping a column, grouping by Postal Code column with neighborhoods in a single cell.

That was done with Apply method. Two datasets were extracted from a single one with different conditions and then merged on a single column Postal Code.
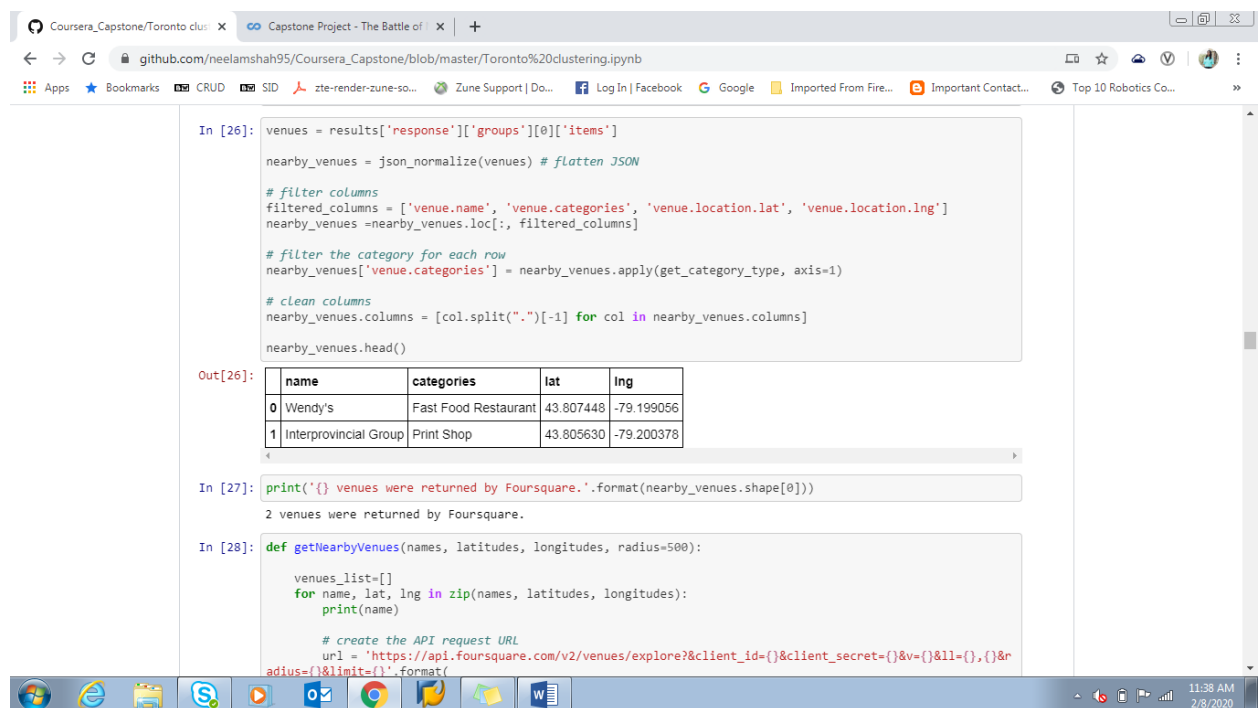
Now, geospatial data of Toronto was downloaded and cleaning was done as required.

Both the final datasets were merged on Postal Code column and new dataframe is created. That dataframe consists of postal codes of different regions with their neighborhoods and values of latitude & longitude of respective areas.

Scaling of data is also done using Preprocessing Scalar library. Data type of Cluster Label column is changed to integer type from float type.

K-means Clustering method is used to cluster venues identified in Scarborough area. Dataset of shops in Scarborough is made by making a call for venues with categories in Foursquare API.

Both final data frames are merged and what shops under which categories in a particular area is made.

```
Clarks Corners, Sullivan, Tam O'Shanter
Agincourt North, L'Amoreaux East, Milliken, Steeles East
L'Amoreaux West
Upper Rouge
```

In [30]:
```
print(scar_venues.shape)
scar_venues.head()
```

(84, 7)

Out[30]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge, Malvern	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop
2	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Chris Effects Painting	43.784343	-79.163742	Construction & Landscaping
3	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
4	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Affordable Toronto Movers	43.787919	-79.162977	Moving Target

In [31]:
```
scar_venues.groupby('Neighborhood').count()
```

Out[31]:

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Agincourt	5	5	5	5	5	5

Before clustering the data, **One hot Encoding** technique is used to convert categorical variables into a form that could be provided to ML algorithms to do a better job in prediction. 0 indicates false values or non-existent and 1 indicates true values or existing ones. Example from current project as shown below:

Out[35]:

	Neighborhood	American Restaurant	Athletics & Sports	Bakery	Bank	Bar	Breakfast Spot	Burger Joint	Bus Line	Bus Station	Café	Car Res
0	Agincourt	0.0	0.000	0.000000	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.00	0.0
1	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.0
2	Birch Cliff, Cliffside West	0.0	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.25	0.0
3	Cedarbrae	0.0	0.125	0.125000	0.125000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.1
4	Clairlea, Golden Mile, Oakridge	0.0	0.000	0.222222	0.000000	0.000000	0.000000	0.000000	0.222222	0.111111	0.00	0.0
5	Clarks Corners, Sullivan, Tam O'Shanter	0.0	0.000	0.000000	0.076923	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.0
6	Cliffcrest, Cliffside, Scarborough Village West	0.5	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.0
7	Dorset Park, Scarborough Town Centre, Wexford ...	0.0	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.0

```
for ind in np.arange(scar_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(scar_grouped.iloc[ind, :], num_t
op_venues)

neighborhoods_venues_sorted.head()
```

Out[39]:

	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 Most Common Venue	8th Most Common Venue	9t...
0	Agincourt	Clothing Store	Skating Rink	Breakfast Spot	Latin American Restaurant	Lounge	Coffee Shop	General Entertainment	Gas Station	Fr Ch Jo
1	Agincourt North, L'Amoreaux East, Milliken, St...	Playground	Park	Vietnamese Restaurant	Clothing Store	Grocery Store	General Entertainment	Gas Station	Fried Chicken Joint	Fa Re
2	Birch Cliff, Cliffside West	General Entertainment	Skating Rink	Café	College Stadium	Vietnamese Restaurant	Coffee Shop	Grocery Store	Gas Station	Fr Ch Jo
3	Cedarbrae	Hakka Restaurant	Athletics & Sports	Bakery	Bank	Gas Station	Fried Chicken Joint	Caribbean Restaurant	Thai Restaurant	Cc & La
4	Clairlea, Golden Mile, Oakridge	Bakery	Bus Line	Soccer Field	Ice Cream Shop	Intersection	Park	Bus Station	Vietnamese Restaurant	El St

In [43]: # set number of clusters

Clustering is used on above shown data with definition of k_clusters = 5. Embedding clusters of places on map, it becomes easy to identify which area has more than 5 restaurants or cafes. Folium map is used to represent map.

- **Results**

  The final location decided, based on the dataframe scar_merged, is **L'Amoreaux West**. The photo of the dataframe is attached below. Same is embedded on folium map in the form of clusters. Final map and dataframe is shown to client. 3 areas are selected and government permissions will be available for which area is checked. Moreover, data from the survey can also be used to decide the area. Survey is conducted in selected 3 areas and opinions of people is taken. This survey is conducted on social networking sites so that data is easily available for compilation and analysis.

  After making clarity on all such points, final area selected becomes L'Amoreaux West.

- **Discussions**

  Previously, this project was to be conducted in Vadodara, Gujarat, India but due to lack of time and data availability, I decided to conduct it in Toronto itself.

  Observations made during this project is Folium map is not always visible in the code. The area was decided on its first visibility.

- **Conclusions**

  Hence, we can conclude on an area as of now to open a café in Scarborough. We used K-means Clustering and one hot encoding technique to reach to the result. This project does not end here unless all obstructions are cleared. Obstructions include government permissions, actually the café going onfloor and resident's reviews too.