Predicting the Best Place for Indonesian Restaurant in Toronto

Yosia Azarya

June 24, 2020

1. Introduction

1.1 Background

Indonesian cuisine is a great food for people in Toronto, especially during winter season, because it contained various spices. In reality, there is no restaurant that served Indonesian cuisine there. So, it is a good idea to open an Indonesian restaurant in Toronto. It is quite challenging to find a place or area to open the Indonesian restaurant. This project will help the enterpreneur to find the most suitable location.

1.2 Problem

The main problem is to find the most suitable location based on the density of restautant in the area.

1.3 Interest

Enterpreneur(s) who wants to open Indonesian restaurant in Toronto, Canada.

2. Data

2.1 Data needed

Data needed for this project are shown below :

- List of neighborhood in Toronto, Canada

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

- Latitude and Longitude data for every neighborhood in Toronto, Canada

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

- Venue data related to restaurant in neighborhoods of Toronto, Canada

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Beaches	43.676357	-79.293031	Seaspray Restaurant	43.678888	-79.298167	Asian Restaurant

2.2 Data Extraction

The extraction of data needed are shown below:

- Scrapping data of Toronto neighborhoods via Wikipedia and stored into dataframe

```
[ ] url  = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
    page = requests.get(url)
```

```
[ ] df_html = pd.read_html(url, header=0, na_values = ['Not assigned'])[0]
    df_html.head()
```

- Getting location coordinates via Geospacial Data given by Coursera

```
[ ]  url_csv = 'http://cocl.us/Geospatial_data'
     df_coordinates = pd.read_csv(url_csv)
```

- Getting the venue data via API call to FourSquare API

```
[ ] def getNearbyVenues(names, latitudes, longitudes, radius=500):

        venues_list=[]
        for name, lat, lng in zip(names, latitudes, longitudes):
            print(name)

            # create the API request URL
            url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
                CLIENT_ID,
                CLIENT_SECRET,
                VERSION,
                lat,
                lng,
                radius,
                LIMIT)

            # make the GET request
            results = requests.get(url).json()["response"]['groups'][0]['items']
```

3. Methodology

3.1 Scrapping Toronto Neighborhoods Data

The Toronto Neighborhood data was scrapped from Wikipedia with pandas library in Python.

```
[ ]  url  = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
     page = requests.get(url)
     if page.status_code == 200:
         print('Page download successful')
     else:
         print('Page download error. Error code: {}'.format(page.status_code))
```

```
Page download successful
```

After that, convert the html data into dataframe with pandas.

```
[ ]  df_html = pd.read_html(url, header=0, na_values = ['Not assigned'])[0]
     df_html.head()
```

	Postal Code	Borough	Neighborhood
0	M1A	NaN	NaN
1	M2A	NaN	NaN
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

3.2 Cleaning Toronto Neighborhoods Data

The Toronto neighborhood contained "NaN" value in Borough and Neighborhood. Delete all rows contained "NaN" values of Borough and Neighborhood.

```
[ ] df_html.dropna(subset=['Borough'], inplace=True)
    df_html.head()
```

	Postal Code	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Group the dataframe with the same borough.

```
[ ] df_postcodes = df_html.groupby(['Postal Code','Borough']).Neighborhood.agg([('Neighbourhood', ', '.join)])
    df_postcodes.reset_index(inplace=True)
    df_postcodes.head(5)
```

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

Save the cleaned Toronto neighborhood data into csv file

```
[ ] df_postcodes.to_csv("torontodata.csv")
```

3.3 Combining Cleanned Data with Geospacial Data

Call the geospacial data with the link given in Coursera and store into dataframe

```
[ ] url_csv = 'http://cocl.us/Geospatial_data'
    df_coordinates = pd.read_csv(url_csv)
```

Call the Toronto Neighborhood data from csv file and store into dataframe

```
[ ] df_neighborhoods = pd.read_csv("torontodata.csv",index_col=[0])
    df_neighborhoods.head()
```

Merge both datasets with Pandas

```
[ ] df_neighborhoods_coordinates = pd.merge(df_neighborhoods, df_coordinates, on='Postal Code')
    df_neighborhoods_coordinates.head()
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	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

Filter the data only with the Borough contained "Toronto" value

```
[ ] df_toronto = dfa[dfa['Borough'].str.contains('Toronto')]
    df_toronto.reset_index(inplace=True)
    df_toronto.drop('index', axis=1, inplace=True)
    df_toronto.tail()
```

```
[ ] print(df_toronto.groupby('Borough').count()['Neighbourhood'])
```

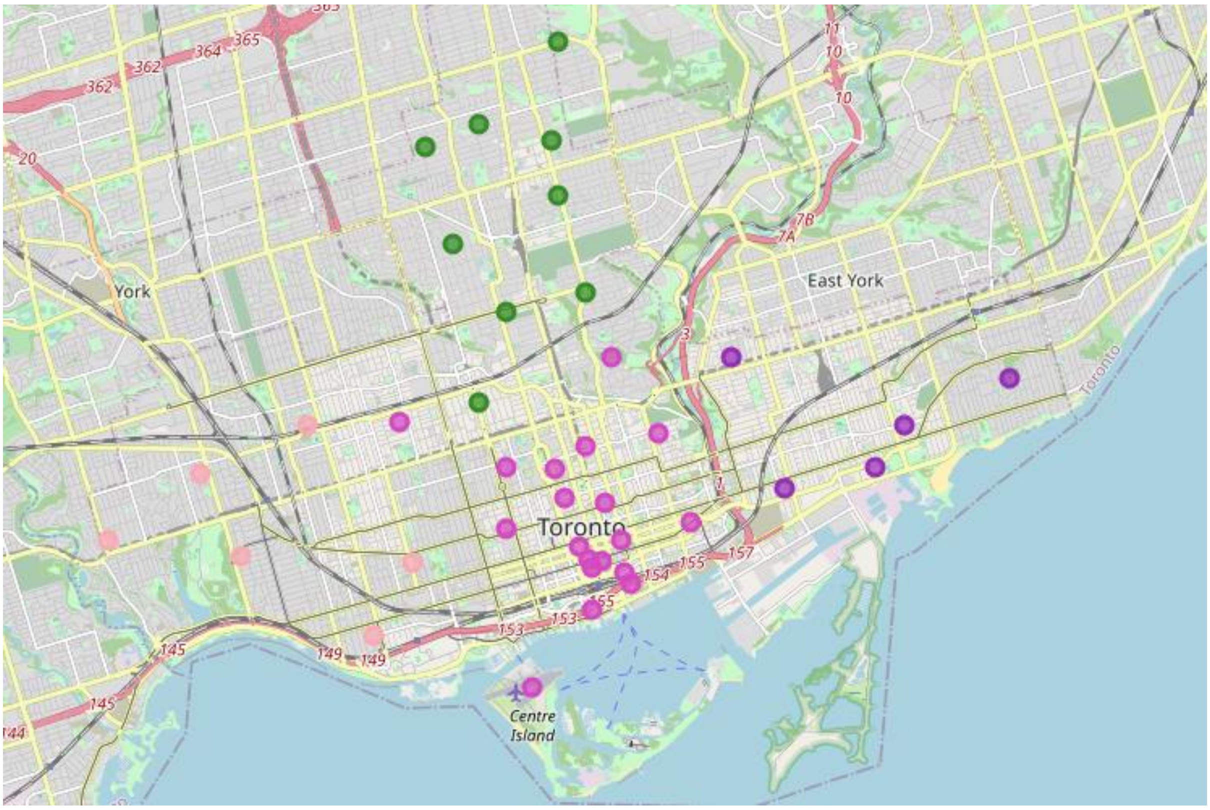
```
Borough
Central Toronto      9
Downtown Toronto    19
East Toronto         5
West Toronto         6
Name: Neighbourhood, dtype: int64
```

Show the coordinates of toronto

```
[ ] lat_toronto = df_toronto['Latitude'].mean()
    lon_toronto = df_toronto['Longitude'].mean()
    print('The geographical coordinates of Toronto are {}, {}'.format(lat_toronto, lon_toronto))
```

```
The geographical coordinates of Toronto are 43.66713498717948, -79.38987324871795
```

Show the map of Toronto with Folium



3.4 Using Foursquare API to Explore Neighborhood

Define the credential and create a function to call Foursquare API

```
[ ] CLIENT_ID = #inputclientid
    CLIENT_SECRET = #inputclientsecret
    VERSION = #YYYYMMDD
    LIMIT = 100
    radius = 500
```

Get the top 100 venues within 500m radius

```
[ ] def getNearbyVenues(names, latitudes, longitudes, radius=500):

        venues_list=[]
        for name, lat, lng in zip(names, latitudes, longitudes):
            print(name)

            # create the API request URL
            url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
                CLIENT_ID,
                CLIENT_SECRET,
                VERSION,
                lat,
                lng,
                radius,
                LIMIT)
```

```
[ ] toronto_venues = getNearbyVenues(names=df_toronto['Neighbourhood'],
                                    latitudes=df_toronto['Latitude'],
                                    longitudes=df_toronto['Longitude'])
```

Store the data from Foursquare API to dataframe

```
[ ] toronto_venues.to_csv("APIfq1.csv")
```

```
[ ] df_hasilfq = pd.read_csv("APIfq1.csv",index_col=0)
    df_hasilfq
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Beaches	43.676357	-79.293031	Seaspray Restaurant	43.678888	-79.298167	Asian Restaurant

Select only venues category contained “Restaurant” in it because we only need neighborhood with restaurant

```
[ ] array1 = df_hasilfq['Venue Category'].unique()
```

```
[ ] array1[0]
```

```
'Trail'
```

Select only venues category with "Restaurant" in it.

```
[ ] kumpulan_restaurant = []
    for i in range(len(array1)):
      # print(i)
      if "Restaurant" in array1[i]:
        kumpulan_restaurant.append(array1[i])

    print(kumpulan_restaurant)
```

```
['Asian Restaurant', 'Greek Restaurant', 'Italian Restaurant', 'Restaurant', 'Caribbean Restaurant',
```

3.5 Eliminate Neighborhoods which have no Restaurant

To get neighborhoods with restaurant, eliminate the neighborhoods which have no restaurant. In order to do that, sum all the values of restaurant column in dataframe, and eliminate rows with restaurant value of 0.

```
[ ] to_filtering = to_grouped[["Neighborhoods"]]
    to_filtering["Sum of Restaurant"] = to_asian[kumpulan_restaurant].sum(axis=1)
    to_filtering.head()
```

	Neighborhoods	Sum of Restaurant
0	Berczy Park	0.214286
1	Brockton, Parkdale Village, Exhibition Place	0.083333
2	Business reply mail Processing Centre, South C...	0.125000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
4	Central Bay Street	0.281250

Check the shape of dataframe before the filtering

```
[ ]  to_filtering.shape
```

```
(39, 2)
```

Filter the dataframe and check the final shape of dataframe

```
[ ]  resto = to_filtering["Sum of Restaurant"].values
     resto
```

```
array([0.21428571, 0.08333333, 0.125     , 0.        , 0.28125   ,
       0.11764706, 0.29333333, 0.3       , 0.27272727, 0.        ,
       0.0625    , 0.3       , 0.16666667, 0.22      , 0.12      ,
       0.32      , 0.20833333, 0.30645161, 0.        , 0.36363636,
       0.33333333, 0.2       , 0.26666667, 0.11764706, 0.09090909,
       0.23913043, 0.        , 0.        , 0.26315789, 0.26923077,
       0.23913043, 0.20833333, 0.2       , 0.25      , 0.14285714,
       0.2       , 0.39534884, 0.26      , 0.29411765])
```

```
[ ]  buangindex = []
     for i in range(len(resto)):
       if resto[i] == 0:
         buangindex.append(i)
```

```
[ ]  buangindex
```

```
[3, 9, 18, 26, 27]
```

```
[ ]  to_filtering = to_filtering.drop(to_filtering.index[buangindex])
     to_filtering.shape
```

```
(34, 2)
```

There are 5 index dropped from dataframe, it means there are 5 neighborhood with no Restaurant.

3.6 Clustering

Use the Kmeans for clustering. Kmeans used for clustering because this project aimed to find the best location based on density of neighborhood with restaurant. Kmeans clustering will use coordinates data from filtered neighborhood data.

Prepare the dataframe for fitting in Kmeans clustering

```
[ ] to_clustering = to_filtering[["Neighborhood Latitude","Neighborhood Longitude"]]
```

```
[ ] to_clustering.head()
```

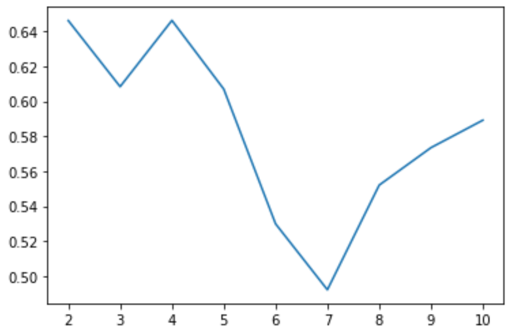
	Neighborhood Latitude	Neighborhood Longitude
0	43.644771	-79.373306
0	43.644771	-79.373306
0	43.644771	-79.373306
0	43.644771	-79.373306
0	43.644771	-79.373306

In order to find the best number of K, simulate Kmeans with K value ranged from 2-10, and check the silhouette score to decide number or K

```
[ ] sil = []
    kmax = 10
    # dissimilarity would not be defined for a single cluster, thus, minimum number of clusters should be 2
    for k in range(2, kmax+1):
      kmeansx = KMeans(n_clusters = k).fit(to_clustering)
      labels = kmeansx.labels_
      sil.append(silhouette_score(to_clustering, labels, metric = 'euclidean'))

    plt.plot (list(range(2,11)),sil)
```

```
[<matplotlib.lines.Line2D at 0x7f6c3237ca58>]
```



From the graph shown above, K=4 showed the best result based on Silhouette Score, so later K=4 will be used for Kmeans clustering.

```
[ ] # set number of clusters
    K = 4

    # run k-means clustering
    kmeans = KMeans(n_clusters=K, random_state=0).fit(to_clustering)

    # check cluster labels generated for each row in the dataframe
    kmeans.labels_[0:10]
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int32)
```

Add the cluster label to dataframe

```
[ ] # create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.
    to_merged = to_filtering.copy()

    # add clustering labels
    to_merged["Cluster Labels"] = kmeans.labels_
    to_merged.tail()
```

	Neighborhood	Sum of Restaurant	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels
38	University of Toronto, Harbord	0.294118	43.662696	-79.400049	Magic Noodle	43.662728	-79.403602	Noodle House	0
38	University of Toronto, Harbord	0.294118	43.662696	-79.400049	DT Bistro	43.662375	-79.405734	Café	0
38	University of Toronto, Harbord	0.294118	43.662696	-79.400049	Comfort Zone	43.658397	-79.400274	Nightclub	0
38	University of Toronto, Harbord	0.294118	43.662696	-79.400049	The Beer Store	43.665385	-79.403477	Beer Store	0
38	University of Toronto, Harbord	0.294118	43.662696	-79.400049	East of Brunswick	43.665609	-79.403324	Pub	0

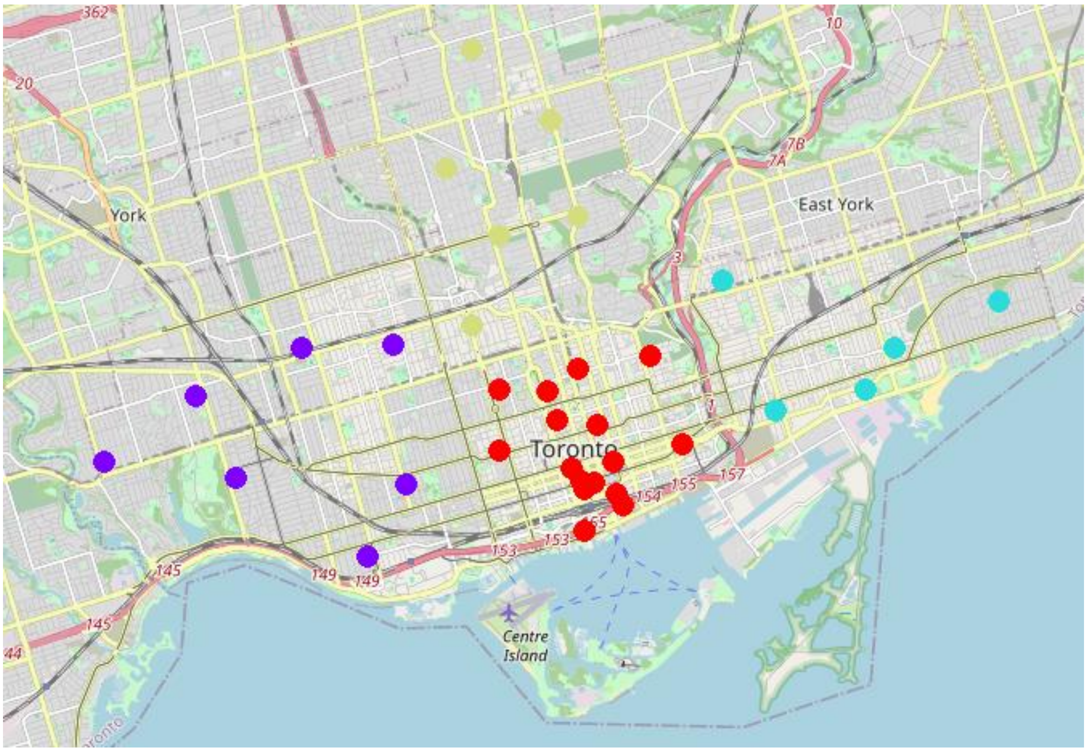
Sort the value by the cluster label

```
[ ] # sort the results by Cluster Labels
    to_merged.sort_values(["Cluster Labels"], inplace=True)
    to_merged.tail()
```

	Cluster Labels	Neighborhood	Sum of Restaurant	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
34	3	The Annex, North Midtown, Yorkville	0.142857	43.672710	-79.405678	Shoppers Drug Mart	43.674959	-79.407986	Pharmacy
34	3	The Annex, North Midtown, Yorkville	0.142857	43.672710	-79.405678	Subway	43.675650	-79.410255	Sandwich Place
34	3	The Annex, North Midtown, Yorkville	0.142857	43.672710	-79.405678	Tim Hortons	43.675800	-79.403532	Coffee Shop
34	3	The Annex, North Midtown, Yorkville	0.142857	43.672710	-79.405678	LCBO	43.675344	-79.405327	Liquor Store
8	3	Davisville	0.272727	43.704324	-79.388790	Meow Cat Cafe	43.702927	-79.388190	Café

3.7 Clustering Visualization

For the visualization, use the folium to show the clusters with different color each cluster



3.8 Clusters Examination

Cluster 0 (Red)

Cluster 0 has the most dense location with restaurant with 1181 restaurant located in this area.

```
[ ] to_merged.loc[to_merged['Cluster Labels'] == 0]
```

	Cluster Labels	Neighborhood	Sum of Restaurant	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	0	Berczy Park	0.214286	43.644771	-79.373306	LCBO	43.642944	-79.372440	Liquor Store
25	0	Richmond, Adelaide, King	0.239130	43.650571	-79.384568	Lobby Lounge at the Shangri-La Toronto	43.649155	-79.386546	Lounge
25	0	Richmond, Adelaide, King	0.239130	43.650571	-79.384568	Toronto PATH System	43.649903	-79.383053	General Travel
25	0	Richmond, Adelaide, King	0.239130	43.650571	-79.384568	Friendly Stranger - Cannabis Culture Shop	43.650387	-79.388523	Smoke Shop
25	0	Richmond, Adelaide, King	0.239130	43.650571	-79.384568	JaBistro	43.649687	-79.388090	Sushi Restaurant
...	...	...	...	...	...	...	...	...	...
11	0	First Canadian Place, Underground city	0.300000	43.648429	-79.382280	Starbucks	43.646799	-79.380690	Coffee Shop
11	0	First Canadian Place, Underground city	0.300000	43.648429	-79.382280	Noodle King	43.651706	-79.383046	Asian Restaurant
11	0	First Canadian Place, Underground city	0.300000	43.648429	-79.382280	Hudson's Bay	43.652040	-79.380391	Department Store
11	0	First Canadian Place, Underground city	0.300000	43.648429	-79.382280	Freshii	43.649273	-79.383748	Salad Place
11	0	First Canadian Place, Underground city	0.300000	43.648429	-79.382280	First Canadian Place	43.648482	-79.382443	Building

1181 rows × 9 columns

Cluster 1 (Purple)

Cluster 1 has 179 restaurent located in this area.

```
[ ] to_merged.loc[to_merged['Cluster Labels'] == 1]
```

	Cluster Labels	Neighborhood	Sum of Restaurant	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
15	1	High Park, The Junction South	0.320000	43.661608	-79.464763	Chica's Nashville Hot Chicken	43.665295	-79.464888	Fried Chicken Joint
15	1	High Park, The Junction South	0.320000	43.661608	-79.464763	Pascal's Baguette & Bagels	43.665426	-79.466176	Bakery
15	1	High Park, The Junction South	0.320000	43.661608	-79.464763	Junction Grill	43.665210	-79.468461	Diner
15	1	High Park, The Junction South	0.320000	43.661608	-79.464763	ROUX	43.665418	-79.462392	Cajun / Creole Restaurant
15	1	High Park, The Junction South	0.320000	43.661608	-79.464763	Cool Hand of a Girl	43.665410	-79.462822	Café
...	...	...	...	...	...	...	...	...	...
19	1	Little Portugal, Trinity	0.363636	43.647927	-79.419750	Founder Restaurant & Bar	43.649478	-79.425352	Restaurant
19	1	Little Portugal, Trinity	0.363636	43.647927	-79.419750	Frankie's Bar & Cafe	43.644290	-79.418481	Diner
28	1	Runnymede, Swansea	0.263158	43.651571	-79.484450	Goodfellas Wood Oven Pizza	43.648224	-79.486356	Italian Restaurant
19	1	Little Portugal, Trinity	0.363636	43.647927	-79.419750	Carmen	43.644829	-79.415872	Tapas Restaurant
19	1	Little Portugal, Trinity	0.363636	43.647927	-79.419750	apt 200	43.644026	-79.420063	Bar

179 rows × 9 columns

Cluster 2 (Blue)

Cluster 2 has 126 restaurant located in this area.

```
[ ] to_merged.loc[to_merged['Cluster Labels'] == 2]
```

	Cluster Labels	Neighborhood	Sum of Restaurant	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
36	2	The Danforth West, Riverdale	0.395349	43.679557	-79.352188	Carrot Commons	43.677485	-79.353076	Restaurant
36	2	The Danforth West, Riverdale	0.395349	43.679557	-79.352188	Dough Bakeshop	43.676643	-79.356846	Bakery
36	2	The Danforth West, Riverdale	0.395349	43.679557	-79.352188	IL FORNELLO on Danforth	43.678604	-79.346904	Italian Restaurant
36	2	The Danforth West, Riverdale	0.395349	43.679557	-79.352188	Simone's Caribbean Restaurant	43.678655	-79.346582	Caribbean Restaurant
36	2	The Danforth West, Riverdale	0.395349	43.679557	-79.352188	Athen's Pastries	43.678166	-79.348927	Greek Restaurant
...	...	...	...	...	...	...	...	...	...
35	2	The Beaches	0.200000	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
35	2	The Beaches	0.200000	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
32	2	Studio District	0.200000	43.659526	-79.340923	Jimmie Simpson Park	43.659230	-79.345063	Park
32	2	Studio District	0.200000	43.659526	-79.340923	Saulter Street Brewery	43.658412	-79.346392	Brewery
32	2	Studio District	0.200000	43.659526	-79.340923	McQueens Pub	43.661483	-79.338072	Gastropub

128 rows × 9 columns

Cluster 3 (Green-Yellow)

Cluster 3 has the least dense location with restaurant with 99 restaurant located in this area.

```
[ ] to_merged.loc[to_merged['Cluster Labels'] == 3]
```

	Cluster Labels	Neighborhood	Sum of Restaurant	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
8	3	Davisville	0.272727	43.704324	-79.388790	Pizza Pizza	43.706138	-79.389292	Pizza Place
21	3	North Toronto West, Lawrence Park	0.200000	43.715383	-79.405678	Milkcow	43.715907	-79.400125	Ice Cream Shop
21	3	North Toronto West, Lawrence Park	0.200000	43.715383	-79.405678	Drake General Store	43.714713	-79.400157	Gift Shop
21	3	North Toronto West, Lawrence Park	0.200000	43.715383	-79.405678	Olly Shoes	43.714990	-79.400119	Shoe Store
21	3	North Toronto West, Lawrence Park	0.200000	43.715383	-79.405678	St. Clements - Yonge Parkette	43.712062	-79.404255	Park
...	...	...	...	...	...	...	...	...	...
34	3	The Annex, North Midtown, Yorkville	0.142857	43.672710	-79.405678	Shoppers Drug Mart	43.674959	-79.407986	Pharmacy
34	3	The Annex, North Midtown, Yorkville	0.142857	43.672710	-79.405678	Subway	43.675650	-79.410255	Sandwich Place
34	3	The Annex, North Midtown, Yorkville	0.142857	43.672710	-79.405678	Tim Hortons	43.675800	-79.403532	Coffee Shop
34	3	The Annex, North Midtown, Yorkville	0.142857	43.672710	-79.405678	LCBO	43.675344	-79.405327	Liquor Store
8	3	Davisville	0.272727	43.704324	-79.388790	Meow Cat Cafe	43.702927	-79.388190	Café

99 rows × 9 columns

4. Conclusion and Recommendation

The conclusion is, the best location to open Indonesian Restaurant is in cluster 0 around Richmond Adelaide, Commerce Court, Berczy Park, and other location located in cluster 0 because cluster 0 has the most dense location with restaurant with 1181 restaurant located in the area. Even though Indonesian Restaurant would be new in Toronto, but it is worth to try, because Indonesian cuisine has a great taste and authentic.