Capstone Project – Thai Restaurant in Toronto

krishna

March 7, 2020

1. Introduction

Toronto is the most populous city in Canada, with a population over than 2.9 million people in 2018. In additional, the city is the center of international business which is included finance, arts, culture, etc. recognized as one of the most multicultural and cosmopolitan cities in the world. Therefore, one of the best things about Toronto's multiculturalism is all the great food from diverse cultures. Among one of the best cuisines in Toronto is Thai food. Mostly, people are known for their Pad Thai, curry, and Thai ice tea, but Thai cuisine goes far beyond that.

This report will show the location of recommended Thai restaurant in Toronto for suggesting the opportunities either business development or sight-seeing at there. By the way, the raw data was limited to describe the Thai restaurants' detail such as recommended dishes, I hope this idea would be useful for the reader who would like to open a new Thai restaurant in Toronto.

2. Data description

The information of Toronto would be based on the following data :

- The Toronto Postal Code, Borough, and Neighborhood was extracted from Wikipedia information (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- The latitude and longitude of Postal Code was extracted from IBM support (http://cocl.us/Geospatial_data)
- The number of restaurants within the certain radius of each borough was come from Foursquare for Developer (Foursquare_API)

```
import pandas as pd
data = pd.read_html('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M',skipro
df = data[0]

df.columns=['PostalCode','Borough','Neighborhood']
df = df[df.Borough != 'Not assigned']

df.head()
```

	PostalCode	Borough	Neighborhood
1	M3A	North York	Parkwoods
2	M4A	North York	Victoria Village
3	M5A	Downtown Toronto	Harbourfront
4	M6A	North York	Lawrence Heights
5	M6A	North York	Lawrence Manor

Fig-1 Toronto's Postal Code in Wikipedia

```
import pandas as pd
df2 = pd.read_csv('http://cocl.us/Geospatial_data')
df2.head()
```

	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

Fig-2 Importing Geospacial Data

After that they were combined the information into one table.

```
df3 = pd.merge(df, df2, left_on = 'PostalCode', right_on = 'Postal Code', how = 'left')

df3 = df3.drop(columns = ['Postal Code'])
df3.head()
```

	PostalCode	Borough	Neighborhood	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

Fig-3 Toronto's Borough and Geospacial Data

3. Methodology and Analysis

After cleaning and preparing the data, we could utilize this data by searching and identifying the venues in the area. This report was used Foursquare Developer as the based platform to get the specific information on each venue.

```
toronto_venues = getNearbyVenues(names=df4['Neighborhood'],
                                 latitudes=df4['Latitude'],
                                 longitudes=df4['Longitude']
                                )
```

```
The Beaches
The Danforth West,Riverdale
The Beaches West,India Bazaar
Studio District
Lawrence Park
Davisville North
North Toronto West
Davisville
Moore Park,Summerhill East
Deer Park,Forest Hill SE,Rathnelly,South Hill,Summerhill West
Rosedale
Cabbagetown,St. James Town
Church and Wellesley
Harbourfront
Ryerson,Garden District
St. James Town
Berczy Park
Central Bay Street
Adelaide,King,Richmond
Harbourfront East,Toronto Islands,Union Station
Design Exchange,Toronto Dominion Centre
Commerce Court,Victoria Hotel
Roselawn
Forest Hill North,Forest Hill West
The Annex,North Midtown,Yorkville
Harbord,University of Toronto
Chinatown,Grange Park,Kensington Market
CN Tower,Bathurst Quay,Island airport,Harbourfront West,King and Spadina,Railway Lands,South Niagara
Stn A PO Boxes 25 The Esplanade
First Canadian Place,Underground city
Christie
Dovercourt Village,Dufferin
Little Portugal,Trinity
Brockton,Exhibition Place,Parkdale Village
High Park,The Junction South
Parkdale,Roncesvalles
Runnymede,Swansea
Queen's Park
Business Reply Mail Processing Centre 969 Eastern
```

Fig-4 Extracting data from Foursquare

To be focusing on restaurant in Toronto, we have to drop other venue from the dataframe.

```
toronto_res = toronto_venues[toronto_venues['Venue Category'].str.contains("Restaurant")]
print(toronto_res.shape)
toronto_res.head()
```

```
(411, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
4	The Danforth West,Riverdale	43.679557	-79.352188	Pantheon	43.677621	-79.351434	Greek Restaurant
6	The Danforth West,Riverdale	43.679557	-79.352188	Mezes	43.677962	-79.350196	Greek Restaurant
7	The Danforth West,Riverdale	43.679557	-79.352188	Cafe Fiorentina	43.677743	-79.350115	Italian Restaurant
11	The Danforth West,Riverdale	43.679557	-79.352188	Messini Authentic Gyros	43.677827	-79.350569	Greek Restaurant
15	The Danforth West,Riverdale	43.679557	-79.352188	7 Numbers	43.677062	-79.353934	Italian Restaurant

Fig-5 Selecting restaurant in dataframe

Currently, we could identify the category of venue which related to Neighborhood area. it will be better to make it ease understand with “One-Hot Encoder” method.

```
# one hot encoding
toronto_res_onehot = pd.get_dummies(toronto_res[['Venue Category']], prefix="", prefix_sep="")
toronto_res_onehot['Neighborhood'] = toronto_res['Neighborhood']

fixed_res_columns = [toronto_res_onehot.columns[-1]] + list(toronto_res_onehot.columns[:-1])
toronto_res_onehot = toronto_res_onehot[fixed_res_columns]

#Due to the unidentified restaurant will mislead the analysis of Toronto's taste, I would rather to drop this term
toronto_res_onehot = toronto_res_onehot.drop(['Restaurant'], axis=1)
toronto_res_onehot.head()
```

	Neighborhood	Afghan Restaurant	American Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Colombian Restaurant	Comfort Food Restaurant	Cuban Restaurant	Di Rest
4	The Danforth West,Riverdale	0	0	0	0	0	0	0	0	0	0	0	
6	The Danforth West,Riverdale	0	0	0	0	0	0	0	0	0	0	0	
7	The Danforth West,Riverdale	0	0	0	0	0	0	0	0	0	0	0	
11	The Danforth West,Riverdale	0	0	0	0	0	0	0	0	0	0	0	
15	The Danforth West,Riverdale	0	0	0	0	0	0	0	0	0	0	0	

Fig-6 Turning data into One-Hot Encoder

```
toronto_res_grouped = toronto_res_onehot.groupby('Neighborhood').mean().reset_index()

print(toronto_res_grouped.shape)
toronto_res_grouped.head()
```

```
(33, 45)
```

	Neighborhood	Afghan Restaurant	American Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Colombian Restaurant	Comfort Food Restaurant	Cuban Restaurant
0	Adelaide,King,Richmond	0.0	0.064516	0.064516	0.0	0.032258	0.0	0.000000	0.000000	0.032258	0.0	0.0
1	Berczy Park	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.1	0.0
2	Brockton,Exhibition Place,Parkdale Village	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0
3	Business Reply Mail Processing Centre 969 Eastern	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0
4	Cabbagetown,St. James Town	0.0	0.083333	0.000000	0.0	0.000000	0.0	0.083333	0.166667	0.000000	0.0	0.0

Fig-7 Grouping One-Hot Encoder data

After we got the one-hot encoder data, it could be grouped together to make it clear understand for popular location. the data was aligned and give a prioritization as a recommended restaurant in their area.

```
def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending=False)

    return row_categories_sorted.index.values[0:num_top_venues]

print('Done')
```

```
Done
```

```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

neighborhoods_venues_res_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_res_sorted['Neighborhood'] = toronto_res_grouped['Neighborhood']

for ind in np.arange(toronto_res_grouped.shape[0]):
    neighborhoods_venues_res_sorted.iloc[ind, 1:] = return_most_common_venues(toronto_res_grouped.iloc[ind, :], num

print(neighborhoods_venues_res_sorted.shape)
neighborhoods_venues_res_sorted.head()
```

```
(33, 11)
```

	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	9th Most Common Venue	10th Most Common Venue
0	Adelaide,King,Richmond	Thai Restaurant	Sushi Restaurant	American Restaurant	Asian Restaurant	Seafood Restaurant	Vegetarian / Vegan Restaurant	Gluten-free Restaurant	Latin American Restaurant	Fast Food Restaurant	Greek Restaurant
1	Berczy Park	Seafood Restaurant	Comfort Food Restaurant	Vegetarian / Vegan Restaurant	Thai Restaurant	Eastern European Restaurant	Japanese Restaurant	French Restaurant	Dim Sum Restaurant	Fast Food Restaurant	Falafel Restaurant
2	Brockton,Exhibition Place,Parkdale Village	Italian Restaurant	Japanese Restaurant	Vietnamese Restaurant	Dim Sum Restaurant	Filipino Restaurant	Fast Food Restaurant	Falafel Restaurant	Ethiopian Restaurant	Eastern European Restaurant	Dumpling Restaurant
3	Business Reply Mail Processing Centre 969 Eastern	Fast Food Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	French Restaurant	Filipino Restaurant	Falafel Restaurant	Ethiopian Restaurant	Eastern European Restaurant	Dumpling Restaurant	Doner Restaurant
4	Cabbagetown,St. James Town	Chinese Restaurant	Italian Restaurant	American Restaurant	Thai Restaurant	Taiwanese Restaurant	Indian Restaurant	Caribbean Restaurant	Japanese Restaurant	Vietnamese Restaurant	Doner Restaurant

Fig-8 Top Ten Restaurant in Toronto area

This stage got the information of recommended venue on each area. By the way, we would like to analysis the Thai restaurant on Toronto to identify the suitable location for developing a restaurant business that we should be creating a model to classify the best location is. That is our next step on k-mean method.

```
# set number of clusters
kclusters_res = 5
toronto_grouped_res_clustering = toronto_res_grouped.drop('Neighborhood', 1)
kmeans_res = KMeans(n_clusters = kclusters_res, random_state=0).fit(toronto_grouped_res_clustering)
kmeans_res.labels_[0:100]
```

```
array([1, 1, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 0, 1, 0, 1, 4, 1, 1,
       4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2], dtype=int32)
```

```
# add clustering labels
neighborhoods_venues_res_sorted.insert(0, 'Cluster Labels', kmeans_res.labels_)
print('Done')
```

```
Done
```

```
toronto_res_merged = df4
toronto_res_merged = toronto_res_merged.join(neighborhoods_venues_res_sorted.set_index('Neighborhood'), on='Neighbo
toronto_res_merged = toronto_res_merged.dropna()
toronto_res_merged['Cluster Labels'] = toronto_res_merged['Cluster Labels'].astype(int)
print(toronto_res_merged.shape)
toronto_res_merged.head()
```

```
(33, 16)
```

	PostalCode	Borough	Neighborhood	Latitude	Longitude	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	8t Co
41	M4K	East Toronto	The Danforth West,Riverdale	43.679557	-79.352188	2	Greek Restaurant	Italian Restaurant	American Restaurant	Caribbean Restaurant	Vietnamese Restaurant	Doner Restaurant	Filipino Restaurant	Fas Res
42	M4L	East Toronto	The Beaches West,India Bazaar	43.668999	-79.315572	1	Fast Food Restaurant	Sushi Restaurant	Italian Restaurant	Vietnamese Restaurant	Cuban Restaurant	Filipino Restaurant	Falafel Restaurant	Et Res
43	M4M	East Toronto	Studio District	43.659526	-79.340923	1	American Restaurant	Italian Restaurant	Middle Eastern Restaurant	Thai Restaurant	Seafood Restaurant	Latin American Restaurant	Comfort Food Restaurant	Res
46	M4R	Central Toronto	North Toronto West	43.715383	-79.405678	4	Fast Food Restaurant	Chinese Restaurant	Mexican Restaurant	Vietnamese Restaurant	Dim Sum Restaurant	Filipino Restaurant	Falafel Restaurant	Et Res
47	M4S	Central Toronto	Davisville	43.704324	-79.388790	1	Thai Restaurant	Sushi Restaurant	Italian Restaurant	Seafood Restaurant	Greek Restaurant	Indian Restaurant	Japanese Restaurant	Vietr Res

Fig-9 K-Mean labels on each area

This analysis is assumed the appropriated clustering value at 5. Then, generating and giving a label into dataframe to be used in the next step on visualization with Folium.

```
# create map
latitude = 43.657952
longitude = -79.387383
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=11)
kclusters_res = 5

# set color scheme for the clusters
x = np.arange(kclusters_res)
ys = [i + x + (i*x)**2 for i in range(kclusters_res)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(toronto_res_merged['Latitude'], toronto_res_merged['Longitude'], toronto_res_merg
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```

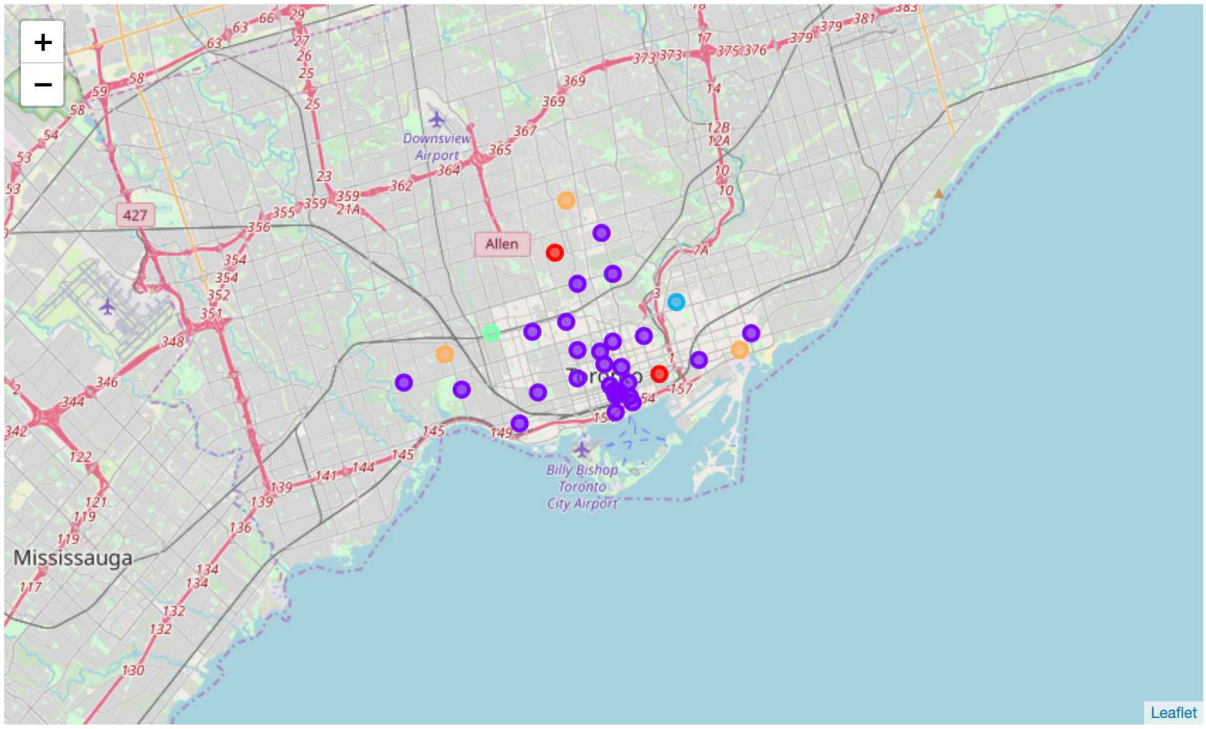

Fig-10 Restaurant in Toronto via Folium's Visualization

Let's repeat these again with Thai restaurant only due to we have to compare the distribution of Thai Restaurant on each clustering to ensure that our next business will be successful and do not open in the red ocean market.

```
#Exploring Thai Restaurant in top 5 areas
thai_res_top = toronto_res_merged.drop(columns = ['6th Most Common Venue','7th Most Common Venue',
                        '8th Most Common Venue','9th Most Common Venue',
                        '10th Most Common Venue'])

thai_1 = thai_res_top[thai_res_top['1st Most Common Venue']=='Thai Restaurant']
thai_2 = thai_res_top[thai_res_top['2nd Most Common Venue']=='Thai Restaurant']
thai_3 = thai_res_top[thai_res_top['3rd Most Common Venue']=='Thai Restaurant']
thai_4 = thai_res_top[thai_res_top['4th Most Common Venue']=='Thai Restaurant']
thai_5 = thai_res_top[thai_res_top['5th Most Common Venue']=='Thai Restaurant']

th_i = [thai_2,thai_3,thai_4,thai_5]
for i in th_i:
    thai_1 = thai_1.append(i)
thai_1
```

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
47	M4S	Central Toronto	Davisville	43.704324	-79.388790	1	Thai Restaurant	Sushi Restaurant	Italian Restaurant	Seafood Restaurant	Greek Restaurant
58	M5H	Downtown Toronto	Adelaide,King,Richmond	43.650571	-79.384568	1	Thai Restaurant	Sushi Restaurant	American Restaurant	Asian Restaurant	Seafood Restaurant
82	M6P	West Toronto	High Park,The Junction South	43.661608	-79.464763	4	Mexican Restaurant	Thai Restaurant	Fast Food Restaurant	Italian Restaurant	Cajun / Creole Restaurant
57	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	1	Italian Restaurant	Japanese Restaurant	Thai Restaurant	Vegetarian / Vegan Restaurant	Korean Restaurant
43	M4M	East Toronto	Studio District	43.659526	-79.340923	1	American Restaurant	Italian Restaurant	Middle Eastern Restaurant	Thai Restaurant	Seafood Restaurant
51	M4X	Downtown Toronto	Cabbagetown,St. James Town	43.667967	-79.367675	1	Chinese Restaurant	Italian Restaurant	American Restaurant	Thai Restaurant	Taiwanese Restaurant
56	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	1	Seafood Restaurant	Comfort Food Restaurant	Vegetarian / Vegan Restaurant	Thai Restaurant	Eastern European Restaurant
61	M5L	Downtown Toronto	Commerce Court,Victoria Hotel	43.648198	-79.379817	1	American Restaurant	Seafood Restaurant	Japanese Restaurant	Thai Restaurant	Vegetarian / Vegan Restaurant
70	M5X	Downtown Toronto	First Canadian Place,Underground city	43.648429	-79.382280	1	American Restaurant	Asian Restaurant	Seafood Restaurant	Japanese Restaurant	Thai Restaurant

Fig-11 Filtered Thai restaurant in dataframe

```
# create map
latitude = 43.657952
longitude = -79.387383
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=12)
kclusters_res = 5

# set color scheme for the clusters
x = np.arange(kclusters_res)
ys = [i + x + (i*x)**2 for i in range(kclusters_res)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(toronto_res_merged['Latitude'], toronto_res_merged['Longitude'], toronto_res_merg
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)


thai_list = thai_1[['Latitude','Longitude']].to_numpy()

for lat_th, lon_th in thai_list:
    folium.Marker(
    location = [lat_th,lon_th],
    popup = folium.Popup(max_width=450).add_child(
    folium.Vega(vis1, width=450, height=250))
    ).add_to(map_clusters)

map_clusters
```

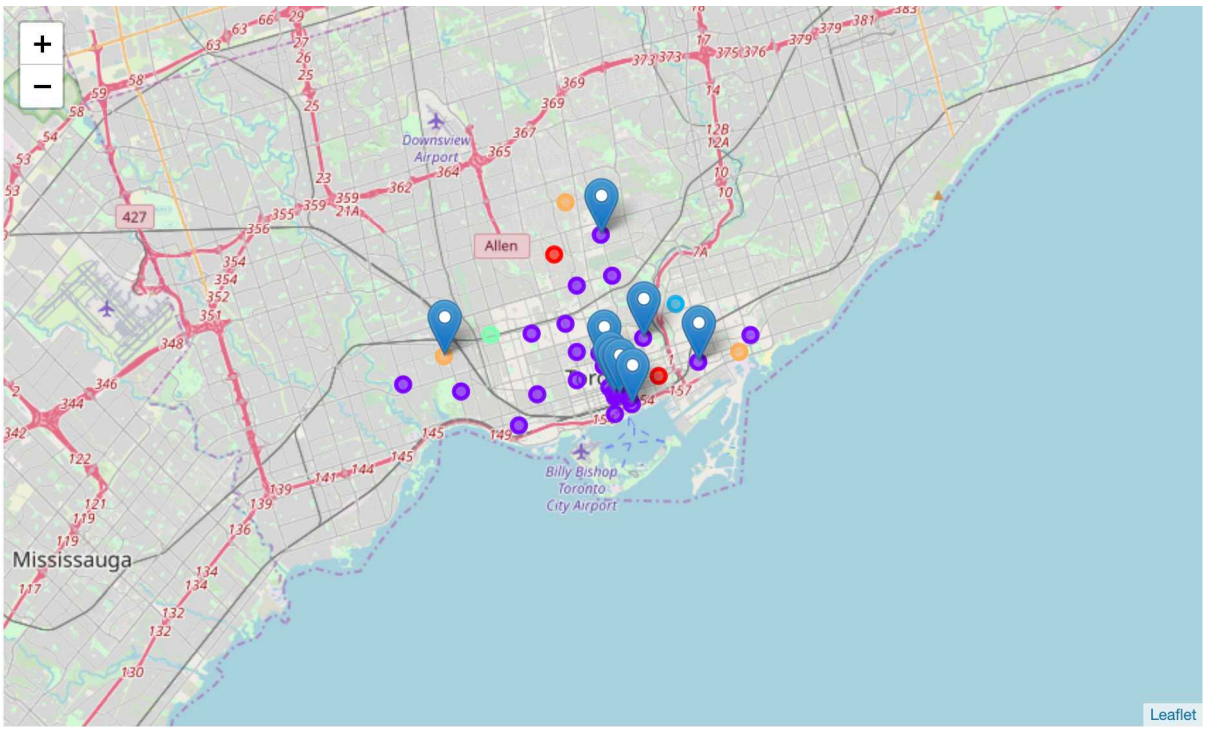

Fig-12 Re-Plot Thai restaurant in Folium

4. Results and discussion

During the analysis, five clusters were defined. The result could be implied that one cluster from this analysis was centroid in the middle of Toronto while the others would be the boundary. In additional, Thai restaurant is located mostly on the single cluster. If you woud like to run the Thai restaurant business, you should consider the other cluster to develop with the lean financial and optimized the marketing cost.

Perhaps, you can develop the focused market at a premium customer in the red ocean which your proposition in the market is high-end only. This strategy would help you deduct educating market about Thai food because there are a lot of Thai restaurant and being top five of the famous venue in those areas.

What could be done better?

The data from Foursquare was limited. Reality, you have to make a market survey to check a demand on High-End User or General User on your business. By the way, this information is useful to understand a current market situation to decide or support the business direction.

5. Conclusion

To conclude, the basic data analysis was performed to identify the most popular restaurant on each boroughs in the city of Toronto. During this analysis, it was required a several important statistical features to explore and visualize the data. Furthermore, clustering helped to highlight the group of optimal areas for forecasting the next step of business. Finally, the data is not limited to do only once, it needs to update the data for real time monitoring on the market likes data-driven way.