Coursera Capstone Project Report

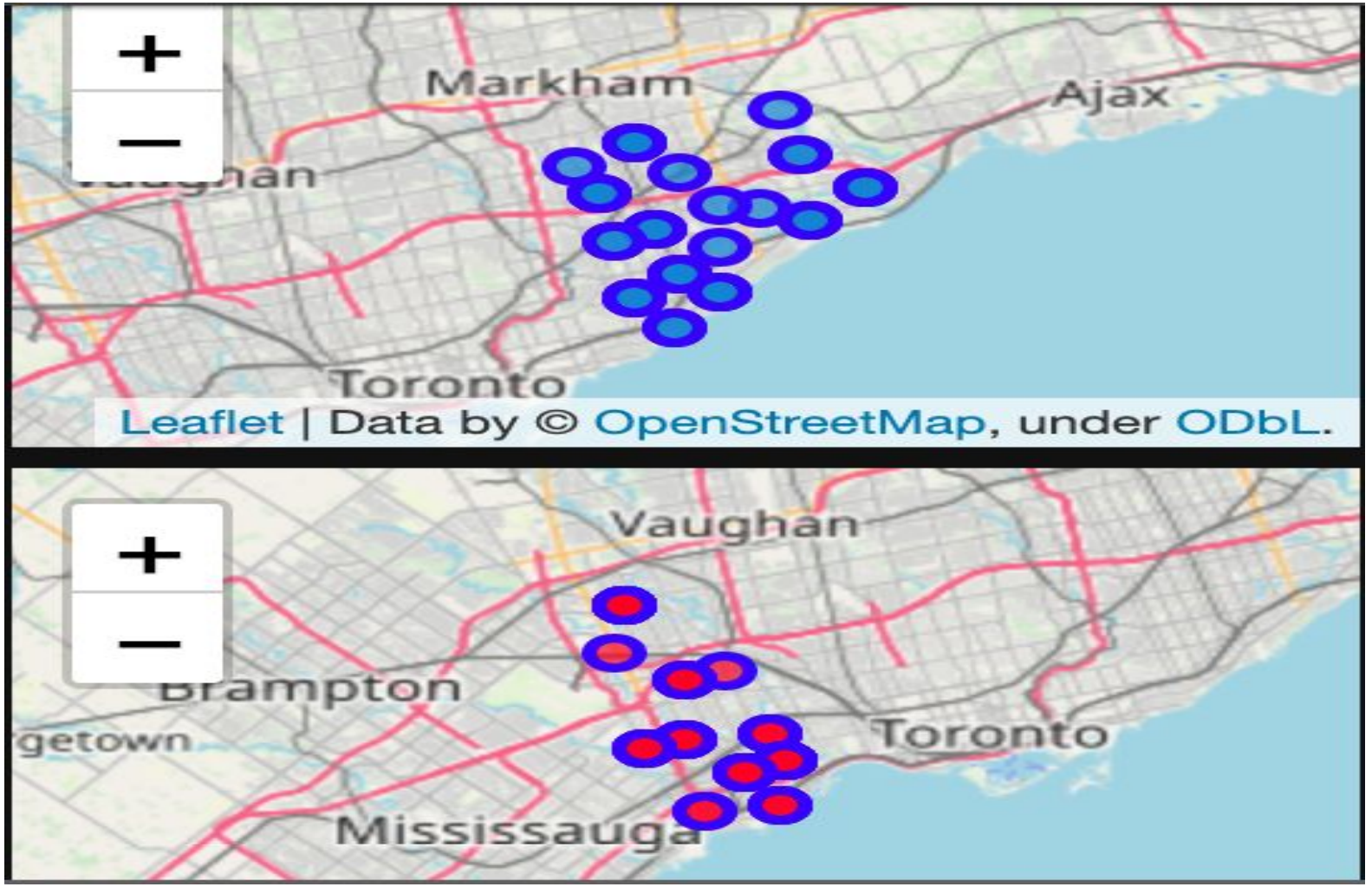
Table of contents

1. Introduction: Business Problem

In this project, we will try to find an optimal location for a bar. Specifically, this report will be targeted to stakeholders interested in opening a **bar in Scarborough or Etobicoke Canada**.

Since there are lots of bars in Canada we will try to detect **locations that are not already crowded with Bars**. We are also particularly interested in **areas with no Bars in the vicinity**. We would also prefer locations **as close to the city center as possible**, assuming that the first two conditions are met.

We will use our data science powers to generate a few most promising cluster on these criteria. The advantages of each area will then be clearly expressed so that the best possible final location can be chosen by stakeholders.


Markham
Ajax
Toronto
Leaflet | Data by © OpenStreetMap, under ODbL.
Vaughan
Brampton
Toronto
Mississauga

2. Data

-Wikipedia is the primary source of data as we can fetch locations and its neighborhood and pin codes.
-Using ‘geopy’ package in python we can get coordinates(latitude and longitude) of a location
-FourSquare API is super useful in gathering information of venues nearby the location we selected

3. **Methodology**

-Extracting the data be the primary step, which can be obtained from Wikipedia scraping the website using ‘requests’ and ‘lxml’ package.
- After extraction, we now clean the data to remove unnecessary fields, NaNs, and manipulate data set according to our needs.
- Folium is super useful for data visualization and analysis
- After gathering Neighbourhood information it can be further used with FourSquare API to get venue details nearby.
- As soon as we get the information about the existing neighborhood we calculate dis. Of the bar from other venues around. (If bar do not exist we choose city center) and append it to the existing data frame.
- Now that we have processed data we convert it to one-hot encoding which is super useful when using KNN.
- But before feeding it to the KNN algorithm. We group it according to frequency and distance and find which is the most common occurring venue in that vicinity.
- Now the grouped data is fed to the KNN algorithm, dividing it to n cluster.
- These clusters are further separated and visualized accordingly.
- As cluster size is different matrix calculation won’t work successfully, so we use the mean of the whole data frame of a single cluster and subtract it from another one to find similarity between two clusters and which cluster will be best to open a bar.

Clustering The data to groups

[833]:
```
from sklearn.cluster import KMeans
kclusters = 3

# run k-means clustering
kmeans_sc = KMeans(n_clusters=kclusters, random_state=1,algorithm='full').fit(scarborough_grouped)
kmeans_et = KMeans(n_clusters=kclusters, random_state=1,algorithm='full').fit(etobicoke_grouped)

# check cluster labels generated for each row in the dataframe
display (kmeans_sc.labels_)
kmeans_et.labels_
```

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

[833]:
```
array([1, 1, 1, 1, 1, 0, 0, 0, 0, 2], dtype=int32)
```

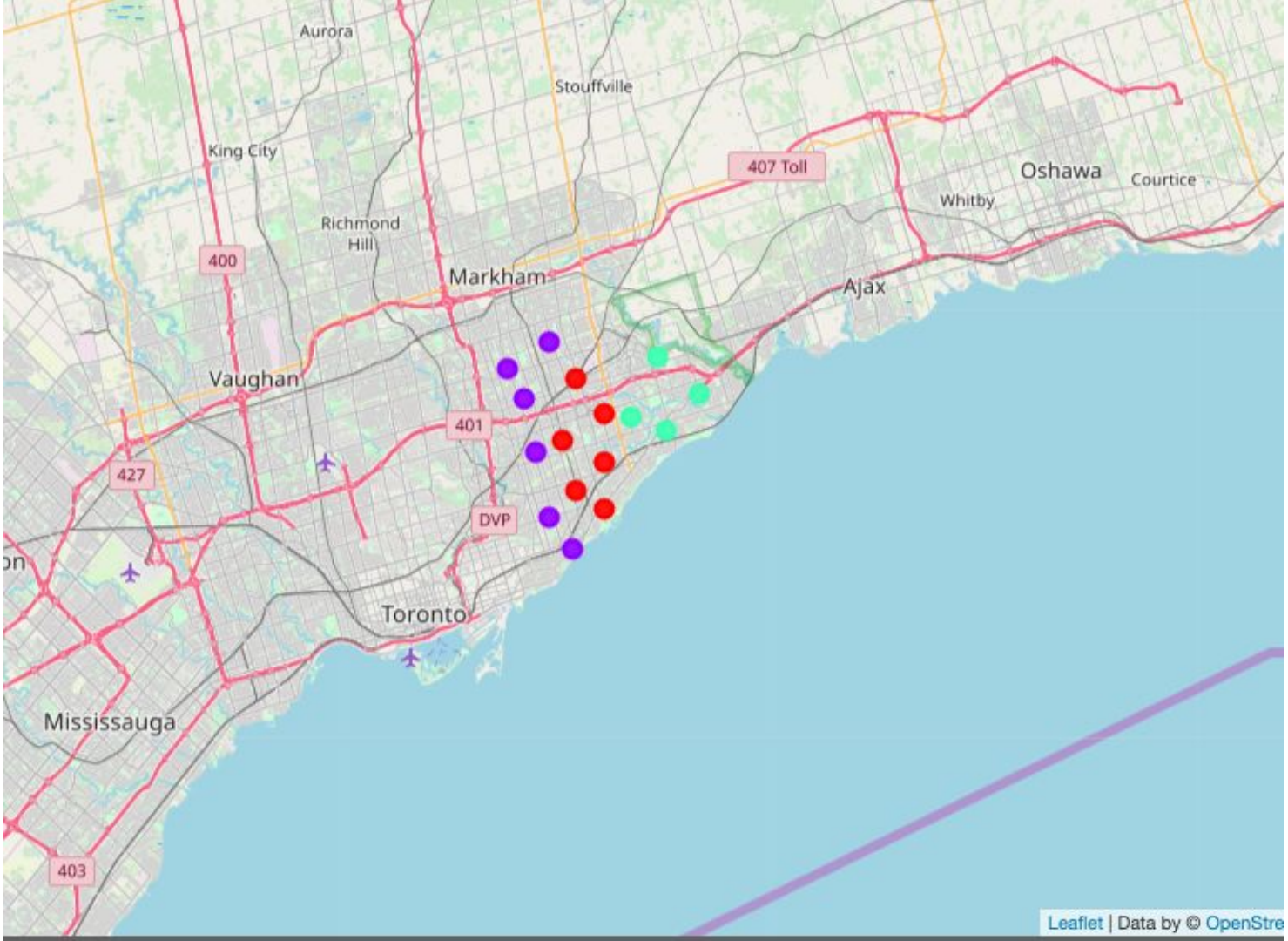
[834]:
```
# add clustering labels
neighborhood1_venues_sorted.insert(0, 'Cluster Labels', kmeans_sc.labels_)

scarborough_merged = scarborough_data
scarborough_merged = pd.merge(left=scarborough_grouped_data,right=neighborhood1_venues_sorted, left_on='Dis', right_on='Dis')
scarborough_merged=scarborough_merged.dropna()
scarborough_merged['Cluster Labels'] = scarborough_merged['Cluster Labels'].apply(np.int64)
scarborough_merged # check the last columns!
```

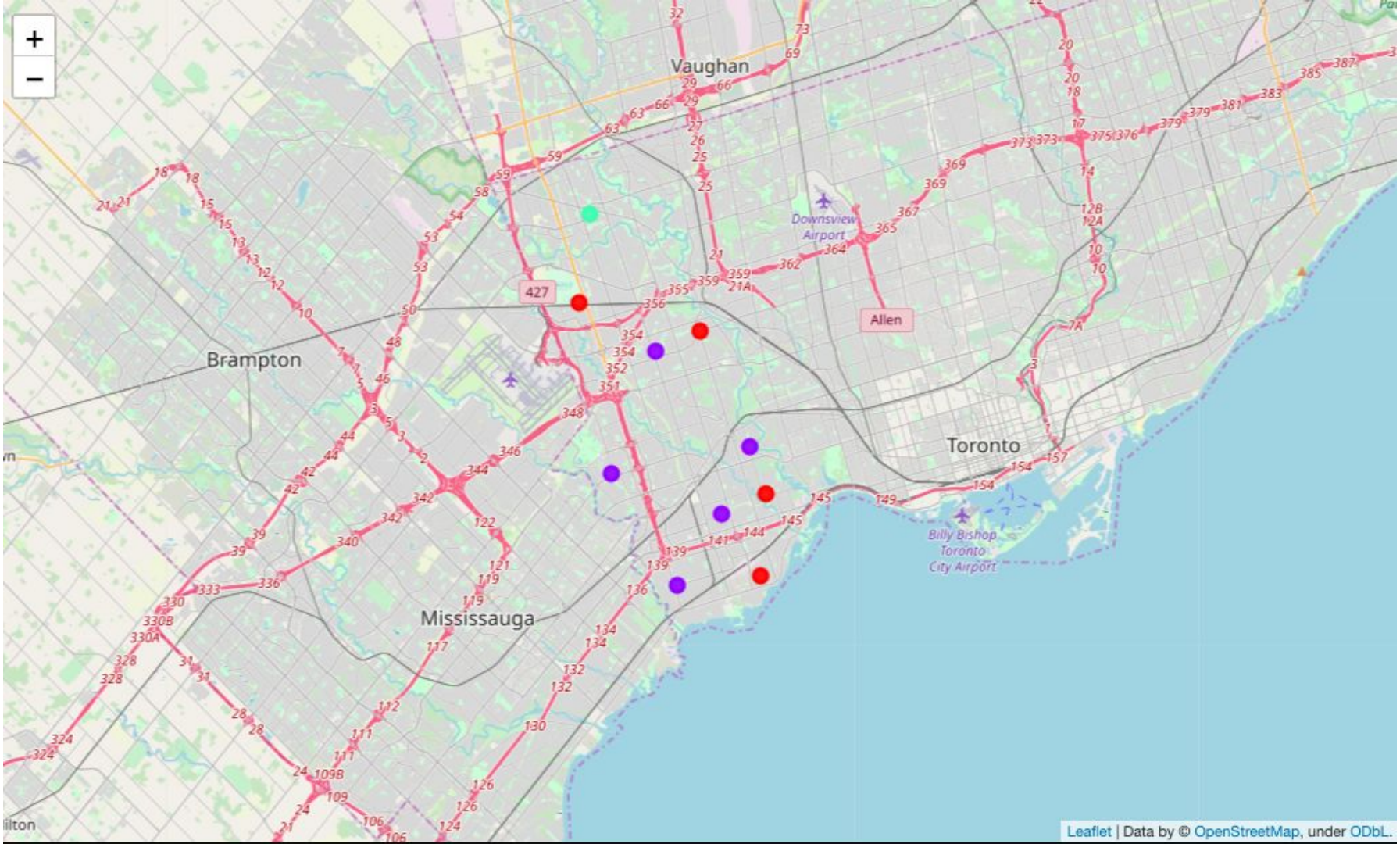
[834]:

	Dis	Neighborhood Latitude	Neighborhood Longitude	Venue Latitude	Venue Longitude	Cluster Labels	1st Most Common NHood	2nd Most Common NHood	3rd Most Common NHood	4th Most Common NHood	5th Most Common NHood	6th Most Common NHood	7th Mo Comm NHo
0	0.3045	43.784535	-79.160497	43.785644	-79.162761	2	History Museum	Bar	Vietnamese Restaurant	Caribbean Restaurant	Fried Chicken Joint	Fast Food Restaurant	Electroni Sto
1	2.9459	43.763573	-79.188711	43.766502	-79.191117	2	Mexican Restaurant	Medical Center	Electronics Store	Rental Car Location	Pizza Place	Intersection	Breakfa Sp
2	3.6758	43.806686	-79.194353	43.807448	-79.199056	2	Fast Food Restaurant	Vietnamese Restaurant	Camera Store	Fried Chicken Joint	Electronics Store	Discount Store	Departme Sto

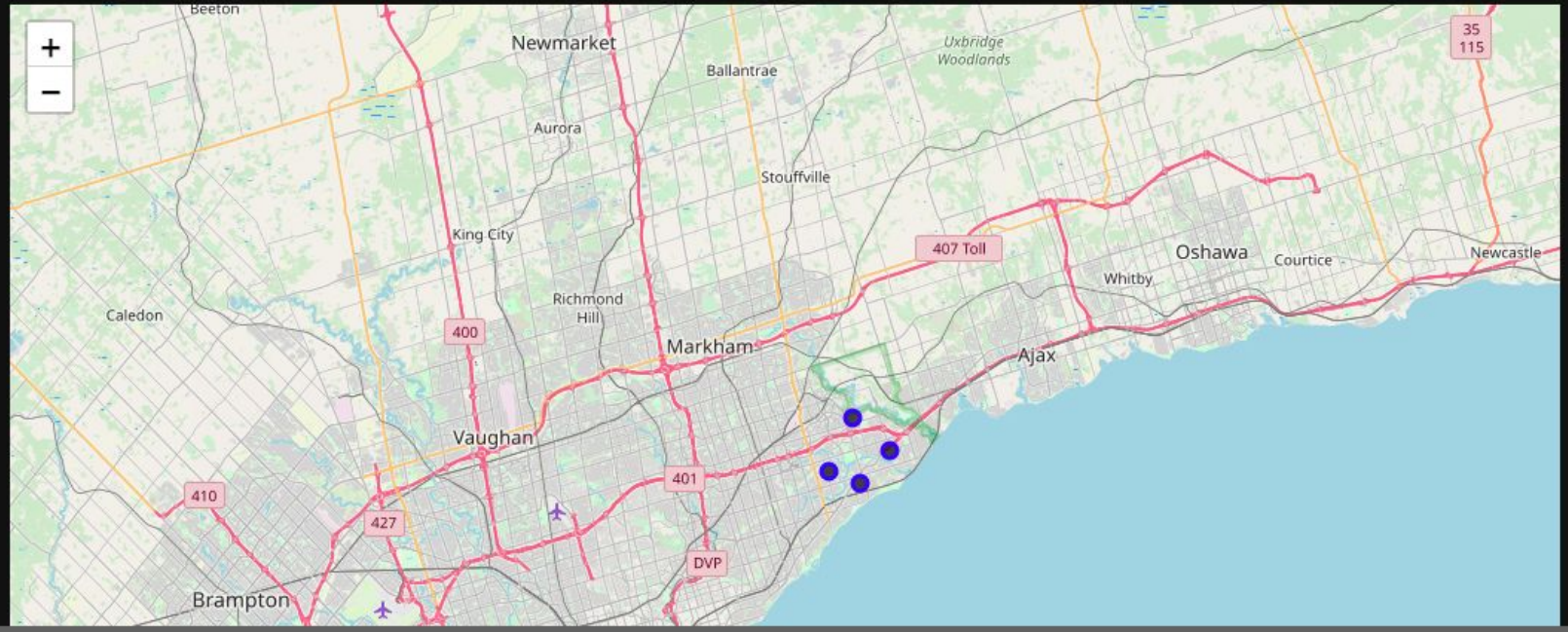
4. Results and Discussion

Our analysis shows that although there is a small number of bars in Canada (initial area of interest which was **Scarborough and Etobicoke**). These two boroughs were identified as potentially interesting (Scarborough, Etobicoke), but our attention was focused on Etobicoke which offer a combination of popularity among tourists, closeness to the city center, strong socio-economic dynamics and a number of pockets of low bars density.
After directing our attention to this more narrow area of interest we started to gather intel about the bar and its location and nearby venues in the area.
Comparison between these two areas and finding similarities between clusters let us know better which are is more suitable for opening a bar.
Those location candidates were then clustered to create zones of interest that contain the greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.
Result of all zones containing the largest number of potential new bar locations based on the number of and distance to existing venues - bar in general. This, of course, does not imply that those zones are actually optimal locations for a new bar! Purpose of this analysis was to only provide info on areas close to Etobicoke center but not crowded with existing bars - it is entirely possible that there is a very good reason for small number of bar in any of those areas, reasons which would make them unsuitable for a new bar regardless of lack of competition in the area. Recommended zones should, therefore, be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.


Aurora
Stouffville
King City
407 Toll
Oshawa
Courtice
Whitby
Richmond Hill
400
Markham
Ajax
Vaughan
401
427
DVP
Toronto
Mississauga
403
Leaflet | Data by © OpenStre


Vaughan
Brampton
Mississauga
Toronto
Allen
427
Downsview Airport
Billy Bishop Toronto City Airport
Leaflet | Data by © OpenStreetMap, under ODbL.

	Dis	Neighborhood Latitude	Neighborhood Longitude	Venue Latitude	Venue Longitude	Cluster Labels	1st Most Common NHood	2nd Most Common NHood	3rd Most Common NHood	4th Most Common NHood	5th Most Common NHood	6th Most Common NHood	7th Most Common NHood	8th Most Common NHood
0	0.3045	43.784535	-79.160497	43.785644	-79.162761	2	History Museum	Bar	Vietnamese Restaurant	Caribbean Restaurant	Fried Chicken Joint	Fast Food Restaurant	Electronics Store	Discount Store
1	2.9459	43.763573	-79.188711	43.766502	-79.191117	2	Mexican Restaurant	Medical Center	Electronics Store	Rental Car Location	Pizza Place	Intersection	Breakfast Spot	Discount Store
2	3.6758	43.806686	-79.194353	43.807448	-79.199056	2	Fast Food Restaurant	Vietnamese Restaurant	Camera Store	Fried Chicken Joint	Electronics Store	Discount Store	Department Store	Cosmetics Shop
3	4.5085	43.770992	-79.216917	43.770559	-79.219579	2	Coffee Shop	Korean Restaurant	Vietnamese Restaurant	Grocery Store	Fried Chicken Joint	Fast Food Restaurant	Electronics Store	Discount Store



5. Conclusion

The purpose of this project was to identify Canadian areas close to center with a low number of bars in order to aid stakeholders in narrowing down the search for the optimal location for a new Bars. By calculating bar density distribution from Foursquare data we have first identified general boroughs that justify further analysis (Scarborough, Etobicoke), and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby bars. Clustering of those locations was then performed in order to create major zones of interest (containing the greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decision on optimal bar location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.

Finding Similarity between clusters

[841]:
```
cluster1_onehot = pd.get_dummies(cluster1, prefix="", prefix_sep="")
cluster1_onehot=np.asarray(cluster1_onehot).flatten()
c1 = (cluster1_onehot.mean())

cluster2_onehot = pd.get_dummies(cluster2, prefix="", prefix_sep="")
cluster2_onehot=np.array(cluster2_onehot).flatten()
c2 = (cluster2_onehot.mean())
#cluster2_onehot=cluster2_onehot.reshape(1,cluster2_onehot.shape[0])

cluster3_onehot = pd.get_dummies(cluster3, prefix="", prefix_sep="")
cluster3_onehot=np.array(cluster3_onehot).flatten()
c3 = (cluster3_onehot.mean())

cluster4_onehot = pd.get_dummies(cluster4, prefix="", prefix_sep="")
cluster4_onehot=np.asarray(cluster4_onehot).flatten()
c4 = (cluster4_onehot.mean())

cluster5_onehot = pd.get_dummies(cluster5, prefix="", prefix_sep="")
cluster5_onehot=np.array(cluster5_onehot).flatten()
c5 = (cluster5_onehot.mean())

cluster6_onehot = pd.get_dummies(cluster6, prefix="", prefix_sep="")
cluster6_onehot=np.array(cluster6_onehot).flatten()
c6 = (cluster6_onehot.mean())
```

[848]:
```
import scipy.spatial.distance as distance
print ('Similarity in Bar cluster and {} Cluster is: '.format('Cluster 1')+str(1-abs(c3-c1)))
print ('Similarity in Bar cluster and {} Cluster is: '.format('Cluster 2')+str(1-abs(c3-c2)))
print ('Similarity in Bar cluster and {} Cluster is: '.format('Cluster 4')+str(1-abs(c4-c5)))
print ('Similarity in Bar cluster and {} Cluster is: '.format('Cluster 5')+str(abs(1-abs(c4-c6))))
```

```
Similarity in Bar cluster and Cluster 1 Cluster is: 0.2669983146806716
Similarity in Bar cluster and Cluster 2 Cluster is: 0.22324011150240164
Similarity in Bar cluster and Cluster 4 Cluster is: 0.7074259441413882
Similarity in Bar cluster and Cluster 5 Cluster is: 0.692553652002758
```