Capstone Project - The Battle of the Neighborhoods

1.Introduction:

In this project, we'll be using foursquare API to get the data regarding two cities, New York City an The city of Toronto. Then we'll find their neighborhoods and cluster them according to their venue categories. As a result, we can find out which portion of each city is similar to the other and which are not.

If anyone wants to examine a portion of a particular city, say, someone wants to check out a restaurant in a particular city then what is the best possibility to find that kind of restaurant to the other city and in which portion. Neighborhoods that are in the same cluster will be similar to their categories.

2. Data acquisition and preprocessing:

2.1. Data Sources:

Firstly we'll download the Newyork data with the borough, neighborhood, latitude, and longitude of each city from the link,

"https://cocl.us/new_york_dataset"

Then acquire the Toronto data from the wikipedia page. Link is,

"https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"

2.2. Data preprocessing:

After getting the Toronto data we'll convert it into csv file which will make our use of data more efficient. Then we'll merge those two datasets into one and by using "foursquare API" we'll collect the venues for each neighborhood. Each venue consists of four items. Unique id, name, categories, and location. Our main focus will be on venue categories. We'll extract different venue categories along with the venues and by using them we'll cluster the full dataset.

3.Methodology:

- We'll collect the latitude and longitude for the Newyork city and The city of Toronto.

- Then we will use the **foursquare API** and explore the neighborhoods of these cities within 500 radius.

- From API call we'll collect the neighborhoods name along with their locations and and venue categories.

- Find out the unique categories and create a onehot dataframe with the columns of the names of venue categories.

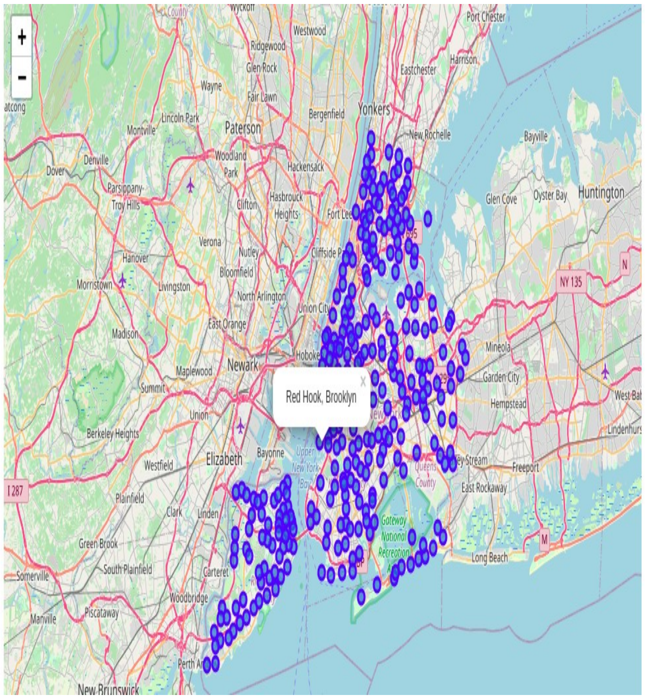
- Then keep the top venues according to venues and specify the num of cluster(cluster = 10).

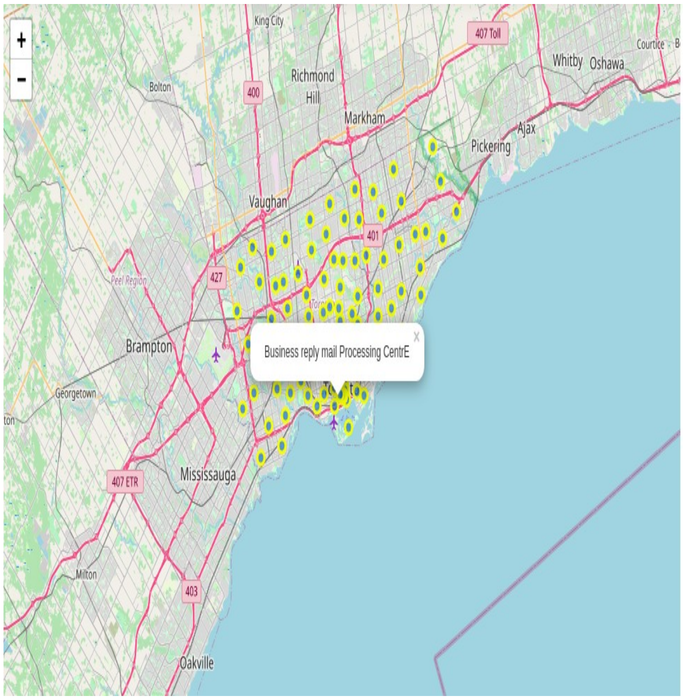
- Apply **K-means** clustering algorithm to find out the similar neighborhoods

- Finding the number of venues in a cluster and information regarding their name and show them in the folium map

4. Analysis:

Visualizing all the venues regarding the places the Newyork and the city of Toronto:





\

Fig1: New York

Fig2: Toronto

Here we can say that number of neighborhoods regarding to the Newyork city is much higher than the city of Toronto. For The Newyork city it is 306 and for The city of Toronto the number is 103.

After clustering:

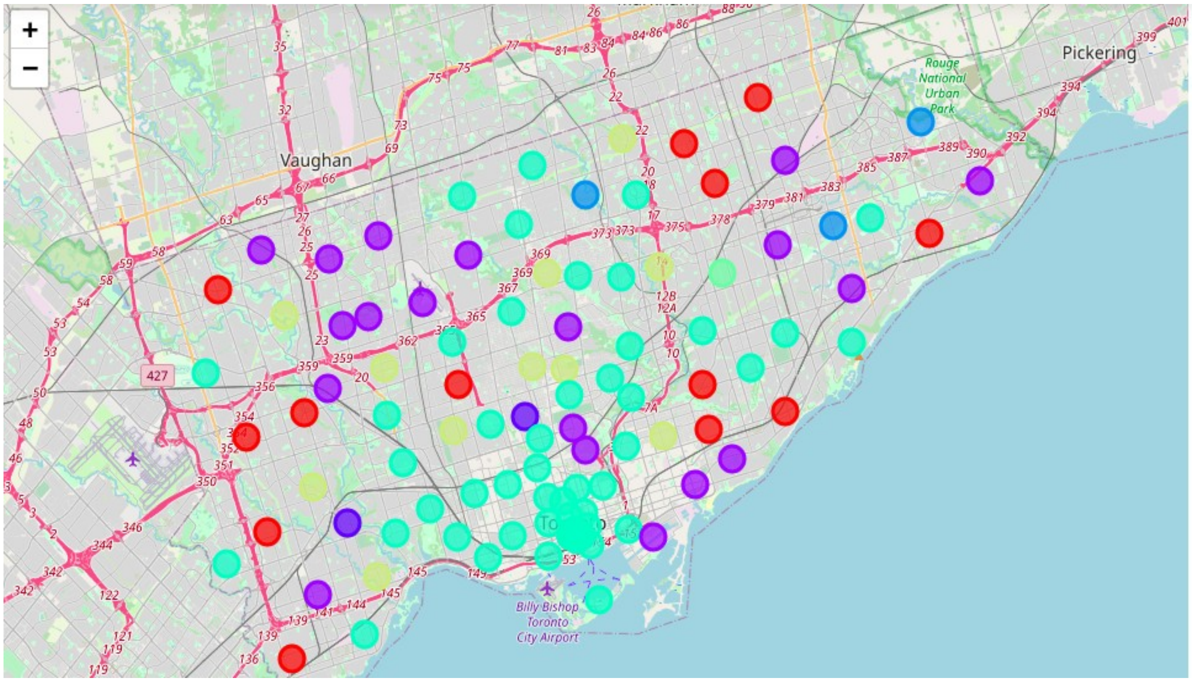


Fig3: New York after clustering

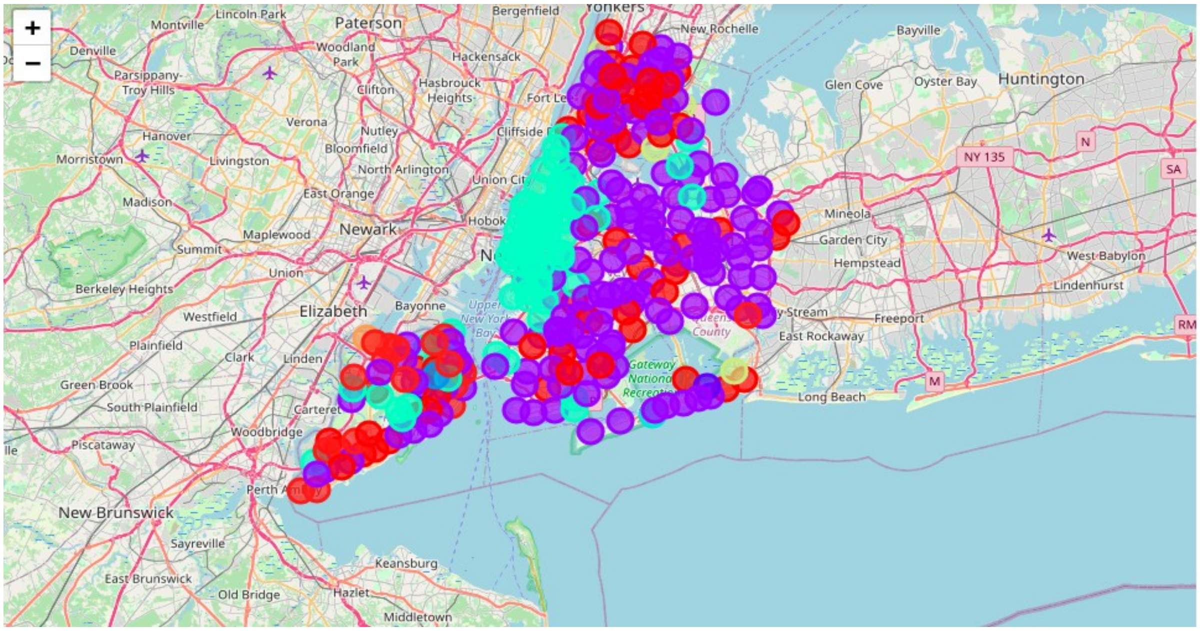


Fig4: Toronto after clustering

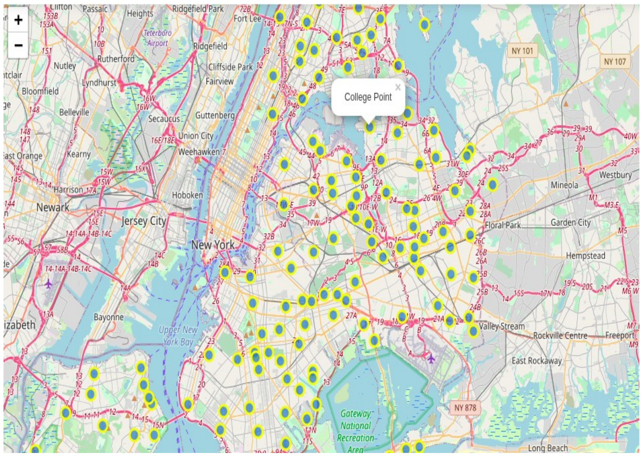


fig5: Cluster 1

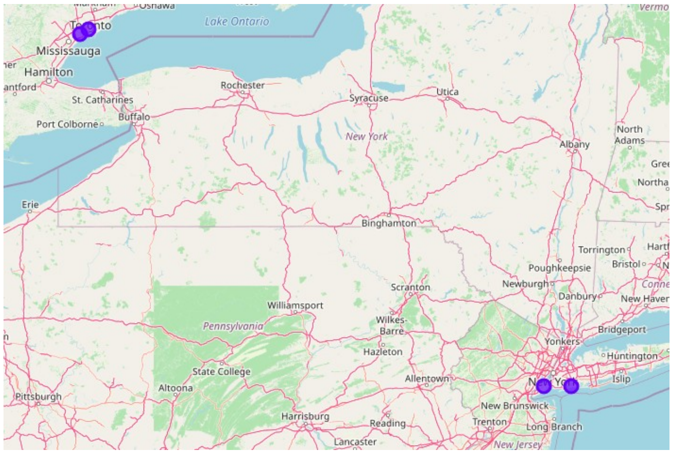


fig6: Cluster 2

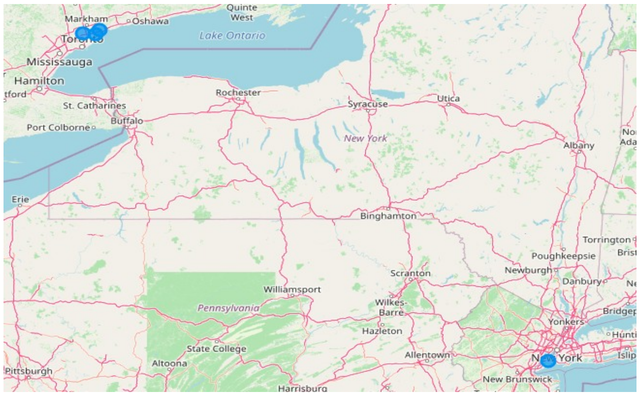


fig7: Cluster 3

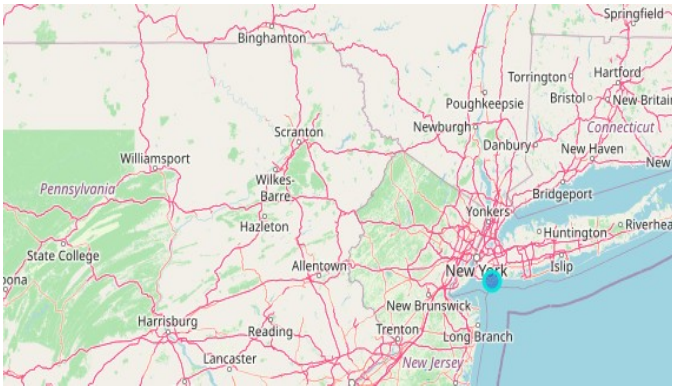


fig8: Cluster 4

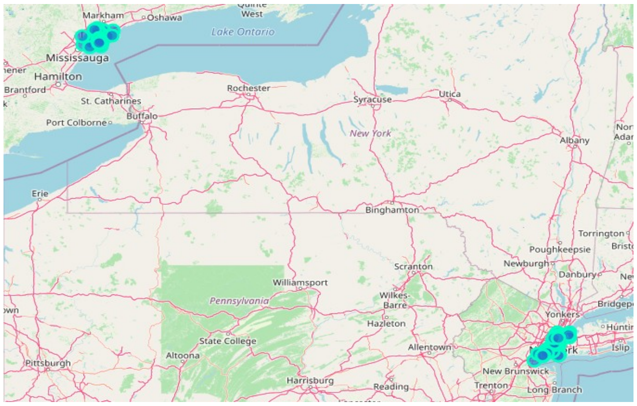


Fig9: Cluster 5

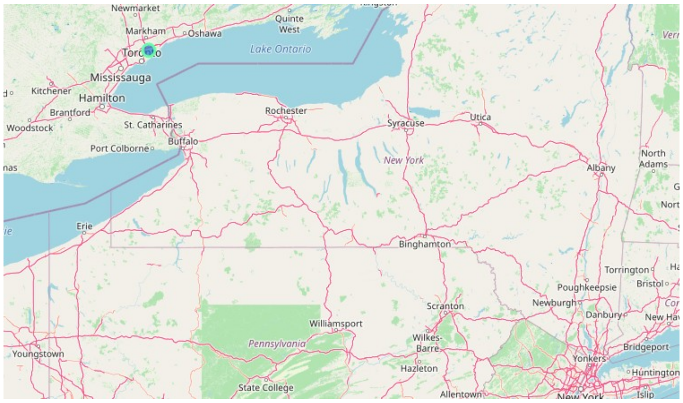


Fig10: Cluster 6

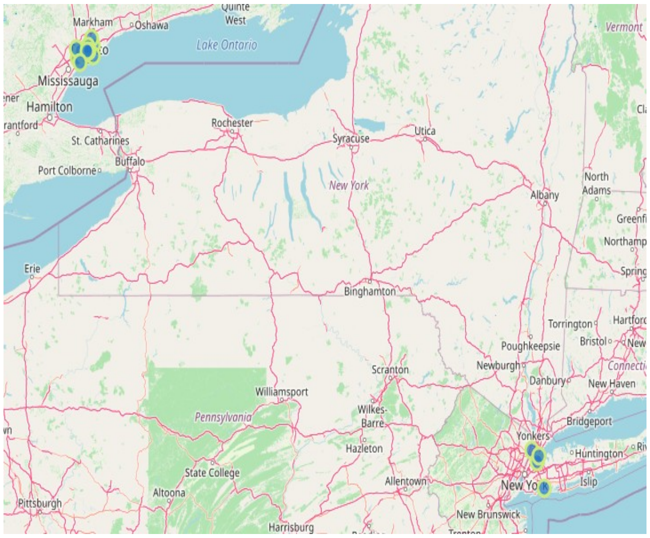


Fig11: Cluster 7

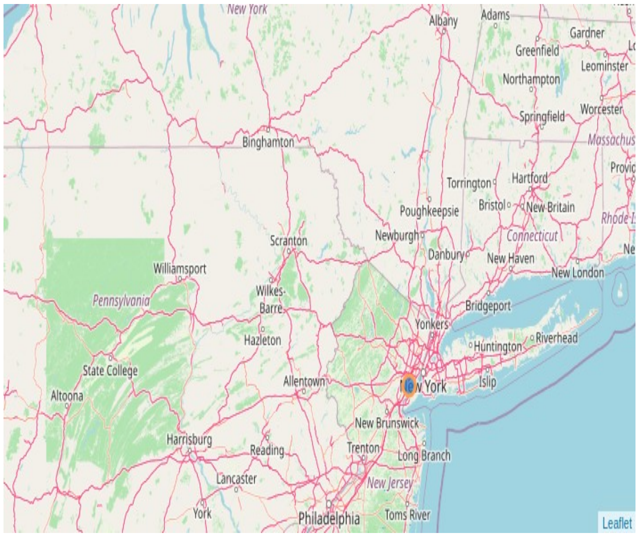


Fig12: Cluster 8

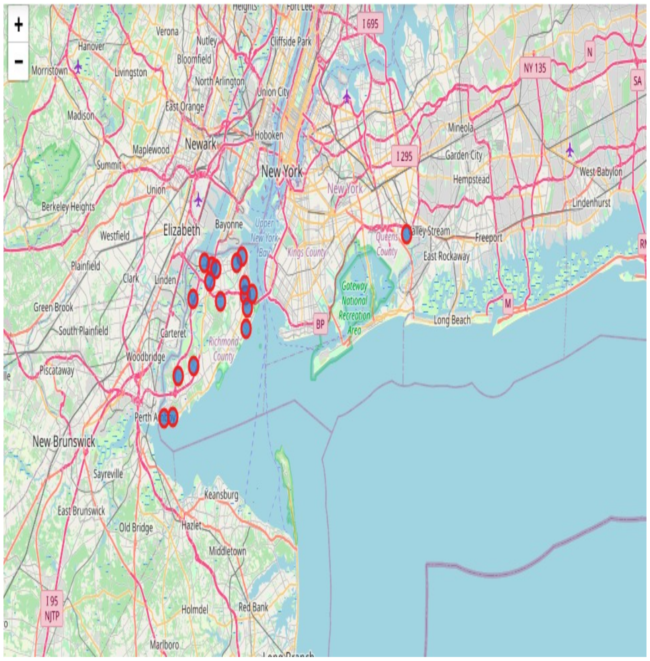


Fig13: Cluster 9

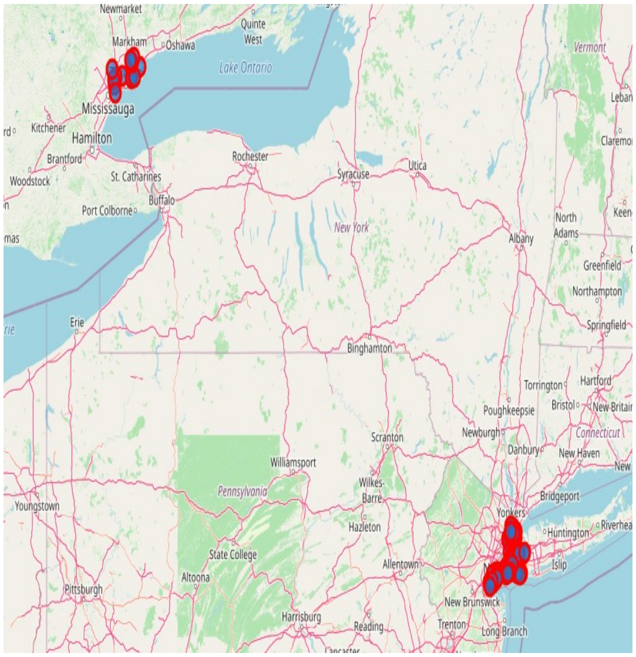


Fig14: Cluster 10

5.Result & Discussion:

Total number of 10 clusters are there, and each cluster consists some values. In cluster number 1,2,3,5,7,9 & 10 consist of values which are on both cities which in case we can say are similar places. but cluster number 4,6 & 8 consist of only one values. So, they are the unique values with unique characteristics which are not similar to the others.

In cluster no 1 there are 164 values. That means 164 values are similar according to their categories.

In cluster no 2 there are 4 values. That means 4 values are similar according to their categories.

In cluster no 3 there are 4 values. That means 4 values are similar according to their categories.

In cluster no 4 there are 1value. That means there is just one unique value in that cluster.

In cluster no 5 there are 126values. That means 126 values are similar according to their categories.

In cluster no 6 there are 1value. That means there is just one unique value in that cluster.

In cluster no 7 there are 15 values. That means 15 values are similar according to their categories.

In cluster no 8 there are 164 values. That means there is just one unique value in that cluster.

In cluster no 9 there are 18 values. That means 18 values are similar according to their categories.

In cluster no 10 there are 72 values. That means 72 values are similar according to their categories.

6.Conclusion:

In this project we wanted to cluster those places in both the Newyork city and the city of Toronto which are more similar to each other. So that, if we are in a part of a city we can assume that in what part of the other city can be the similar like present city. We have used total 10 clusters to apply the algorithm and got the result. Apart from that we could use less number of clusters to get more compact values as some of our clusters have only one item in it. By using this foursquare API we can also predict the location of a venue of certain category by using other machine learning algorithms which I want to implement in the future.