**Battle of Finding the Similar Neighbourhood**

**Shivam Agarwal**

March 20, 2019

1. **Introduction**

A friend of mine is moving from New York City (Brooklyn) to Toronto (Canada).

He is searching for a place in Toronto which is similar in neighbourhood as in Brooklyn (New York City).

* 1. **Business Problem**

As he is moving to Toronto he asked me to help him in finding a place which has same surrounding like malls, bar, restaurants, gym etc. nearby to his new place.

1. **Data acquisition and cleaning**
   1. **Data Source**

New York City neighbourhood has a total of 5 boroughs and 306 neighbourhoods in order to segment the neighbourhood and explore them we need a dataset. The data set is available free available at

<https://geo.nyu.edu/catalog/nyu_2451_34572> (json file).

For the Toronto neighbourhood data, a Wikipedia page exists that has all the information we need to explore and cluster the neighbourhoods in Toronto.  You will be required to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a *pandas* data frame so that it is in a structured format like the New York dataset.

* 1. **Data Cleaning**

For New York City data we read a json file and notice that relevant data is in the “features” key, which is basically a list of neighbourhoods.

Next we transformed a data into data frame by looping through the data and fill the data frame one by one. And finally we ensure that the data frame contains 5 boroughs and 306 neighbourhoods.

For Toronto City we scrap a Wikipedia page: -https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M

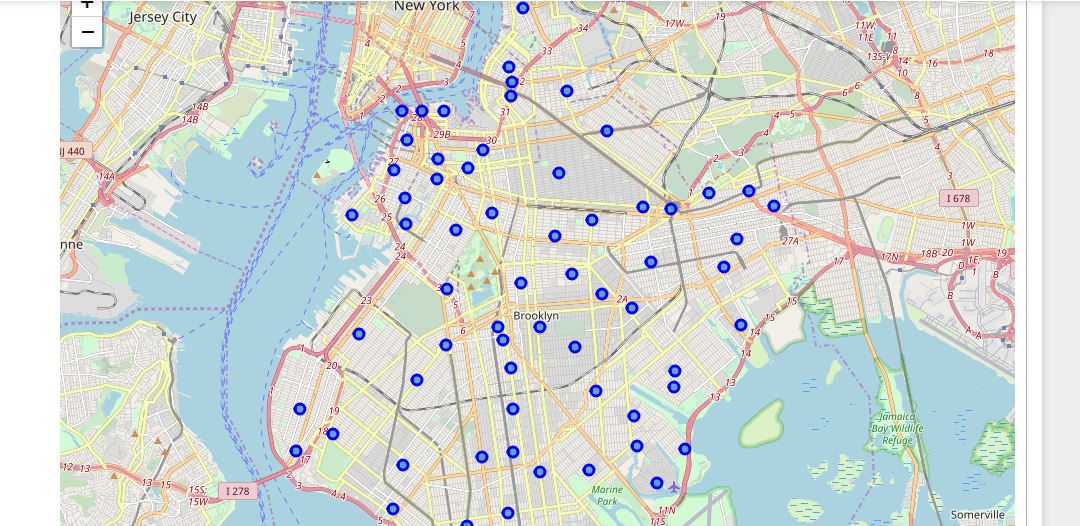
From where we get a list of postal codes of Toronto borough and along with their neighbourhoods.

First of all to clean a data we remove all the observation which has missing borough. Then we merge multiple neighbourhoods in single columns according to their postal codes. Finally we merge the latitude and longitude of boroughs of Toronto with the Postal codes data frame to obtain a desired data frame with 103 neighbourhoods.

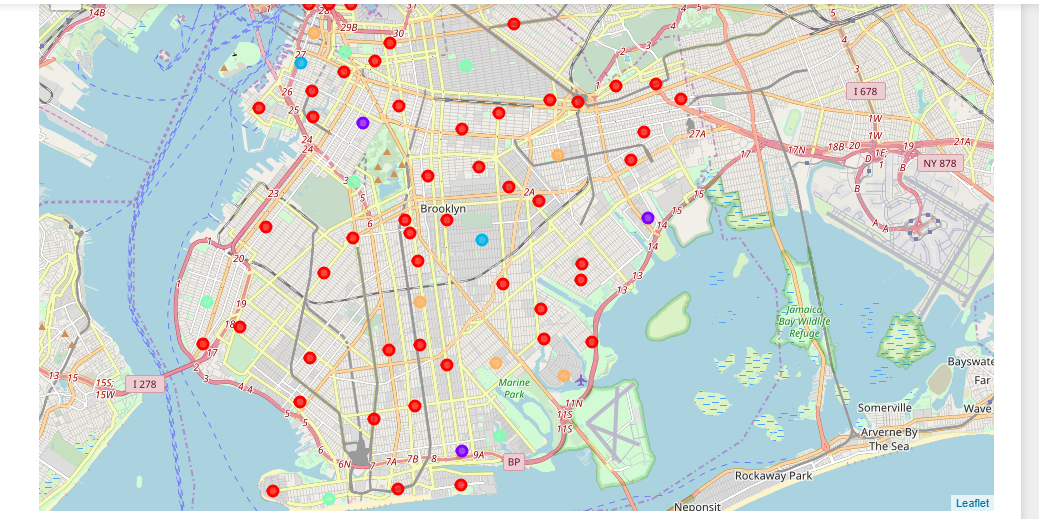
* 1. **Features Selection**

After cleaning the data for New York City we have 306 neighbourhoods and 5 boroughs. As my friend lives in Brooklyn we consider only Brooklyn having 70 neighbourhoods. We didn’t find any correlation between any independent variables so we finally selected all the variables for our model development.

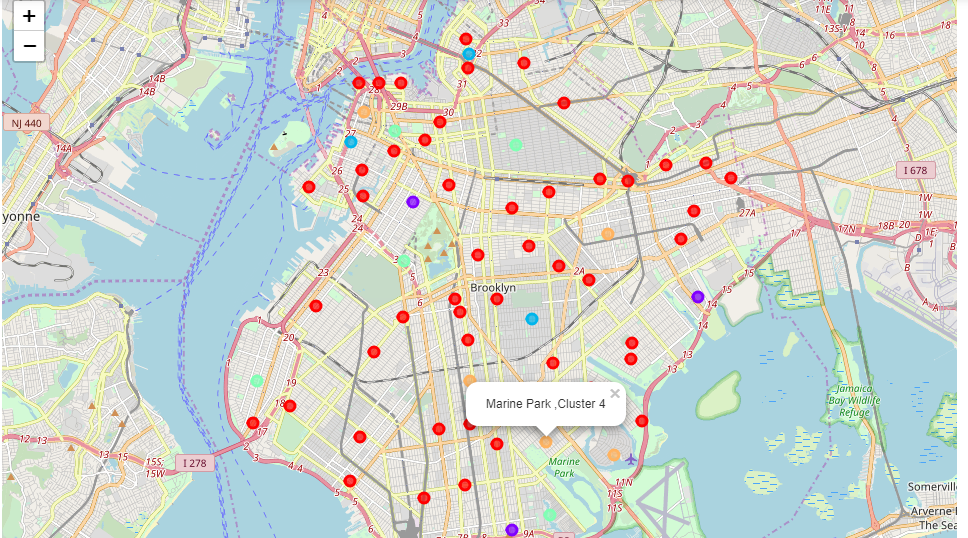
1. **Methodology**
2. In order to get the accurate Brooklyn neighbourhood segmentation we create a map marking all the neighbourhoods of Brooklyn.



1. Access Foursquare location data of each neighbourhoods, using one hot technique to get top 20 venues of Brooklyn’s neighbourhood.
2. Apply K-means Cluster to segment all 70 neighbourhoods of Brooklyn into 5 clusters.



1. As he lives in Marine park which comes in Cluster 4.



1. Now we have to find the similar neighbourhood in Toronto.
2. We repeat the above process:-

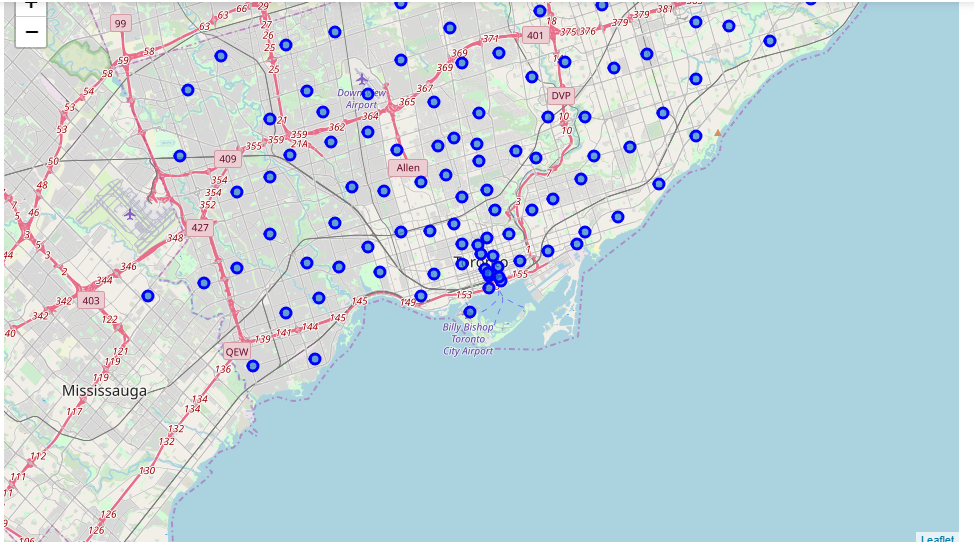


Fig: - Showing all the 103 neighbourhoods in Toronto.

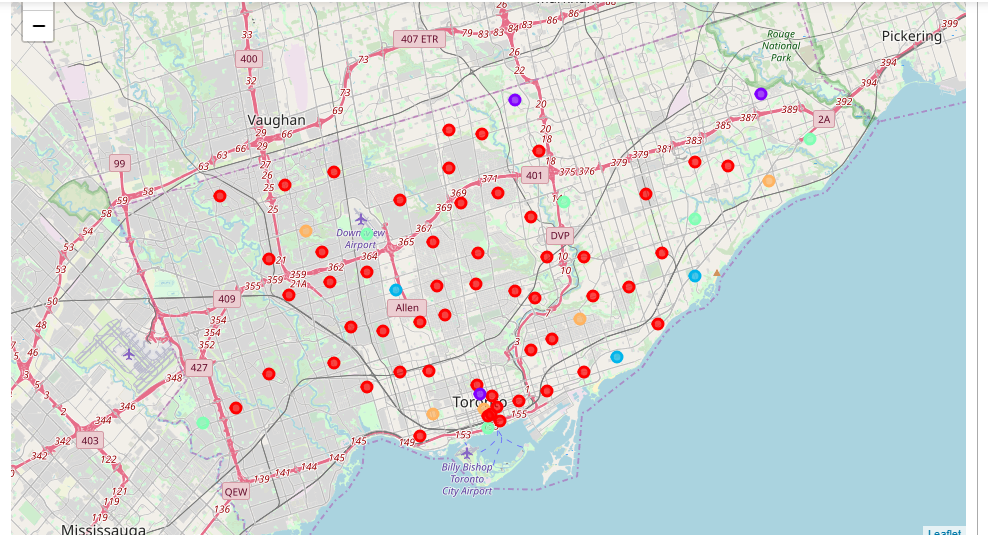


Fig: - Showing division of 103 neighbourhoods of Toronto into 5 Clusters.

1. **Results**

In New York City Marine Park falls under cluster 4. So in Toronto 4 Boroughs namely North York, Etobicoke, East York , Central Toronto falls in Cluster 4 showing similar neighbourhood as Marine Park in Brooklyn.

1. **Discussion**

The neighbourhoods of cluster 4 in Toronto are very much similar to Marine Park in New York City. So finally we have four choices out of 103 neighbourhoods of Toronto. Thus making our search limited to these four neighbourhoods only in place of searching in 103 neighbourhoods.

1. **Conclusion**

This result has limitations. The venues used in the project are from the top 20 venues of each neighbourhood, which might neglect neighbourhoods with more venues.