NY vs. Toranto:

Hannah L. Weinberger

- ## Introduction:

In this project, we will study, analyze, cluster and compare the neighborhoods of two different cities in the world.
Doing this project, we will understand that cities in the world have many different venues that define the culture of the cities. yet it is somewhat possible to group together the similar kind of neighborhoods in different cities. You can segment the different venues in the neighborhood according to venue category, and then group neighborhoods together that incorporate similar kind of neighborhoods. Having the similar neighborhoods grouped together it can help make decision when people consider moving out of a city to another.

What we will gain is helping any person who needs to move from his neighborhood to a different city. And he is having a hard time leaving home. This project will help him find a neighborhood far from home with as much similarity to home.

- ## The data:

We will use 2 datasets. The first consists of New York's different neighborhoods and their respective geometric coordinates. https://cocl.us/new_york_dataset
The second dataset consists of Toronto's different borough and their respective postcodes. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
As we import the data frames, we will clean them and make the data frame comfortable to read.
Next we will use Foursquare for each data frame to find the venues in the neighborhoods. We will make a new df for each city with the venues clustered into groups.
Last step will to merge both data frames and cluster the same venues.

- ## Methodology:

The goal of this project is to group the similar neighborhoods.

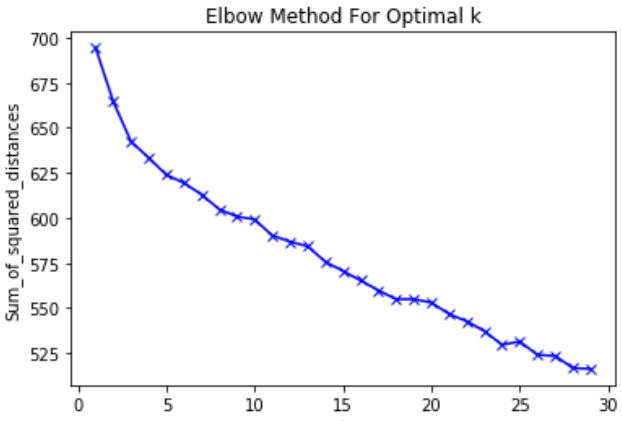
1. we will determine the optimal cluster number. Using K-means a machine learning algorithm that groups a dataset into a user specified number (k) of clusters. The algorithm clusters the data into k clusters, even if k is not the right number of clusters to use. Therefore, when using k-means users need some way to determine whether they are using the right # of clusters. One method to find out the right number of k is the elbow method. The idea if the elbow method is to run the K-means clustering on the dataset for a range of k's (e.g. k from 1 to 20), and for each K calculate the sum of squared errors. Then plot a line chart of the SSE for each value of k. if the line chart looks like an arm, then the "elbow" on the arm is the value of k to be used. We want a small SSE, but not an SSE of 0 (the SSE is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster). The goal is to choose a small value of k that still has a low SSE. We can also use silhouette method to find the optimal cluster number. Silhouette analysis measures how close each point in a cluster is to the point in its neighboring cluster. Silhouette values lies

in the range of [-1,1]. A value of 1 indicates that the sample is far away from its neighboring cluster and very close to the cluster it is assigned. A value of -1 indicates that the sample is closer to the neighboring cluster then from the cluster it is assigned. A value of 0 means it's at a boundary of the distance between the two clusters. We would like a value as high as possible.

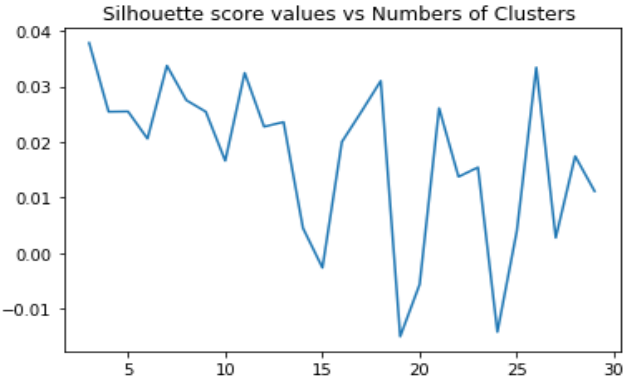
2. Because K-measures incorporate a heuristic approach, it does not guarantee convergence in the global optima in each iteration depending on how the initial location of the clusters is determined, it may be converge on a different local optima. In order to overcome this issue, many reps various random starts were made to find the best convergence set.

- **Results:**

1. Optimal # of clusters:



Elbow method.

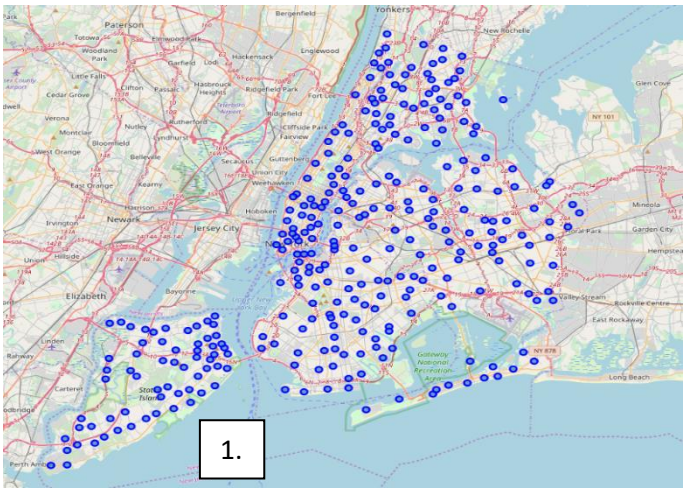


Silhouette method.

```
Optimal number of components is:
3
```

The Number of clustering being 3 experiences a decrease before it and a gradual regular decrease after it in the elbow method. The Silhouette score confirms number of clusters being 3 has its peak in the screen shot added. The silhouette coefficient is calculated using the average distance between clusters and the average distance to the closest cluster for each sample.

2. <u>Visualizing the clusters on the map:</u>
   We created folium maps to help get a visual idea of what the different clusters look like on the map of New York and Toronto.



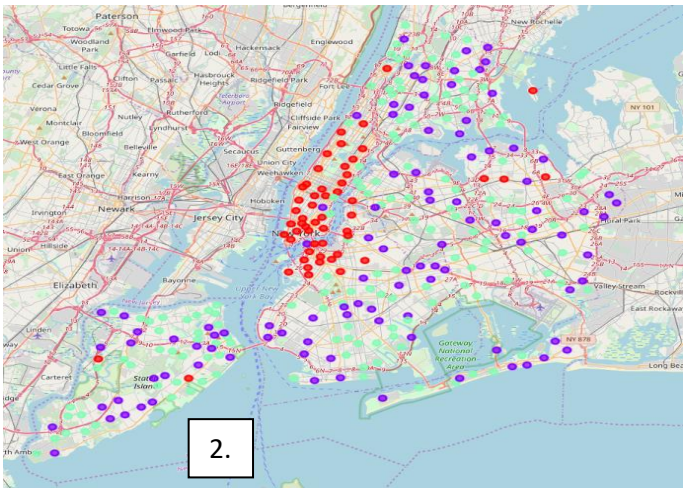
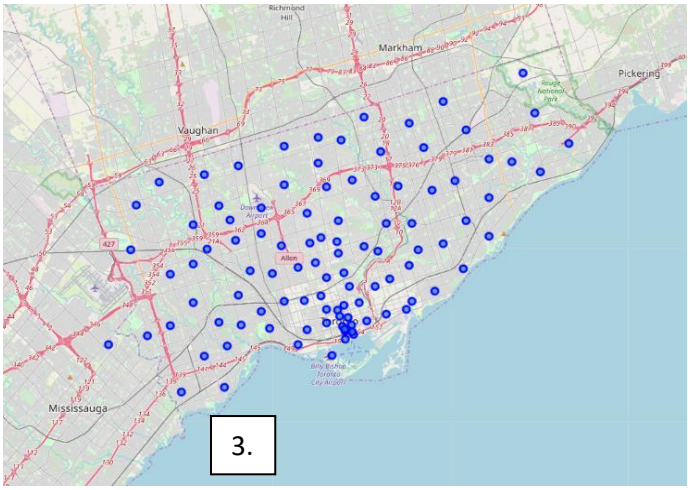


Image #1 is Venues pinned on NY before clustering. Image #2 is venues pinned on Ny after clustering.



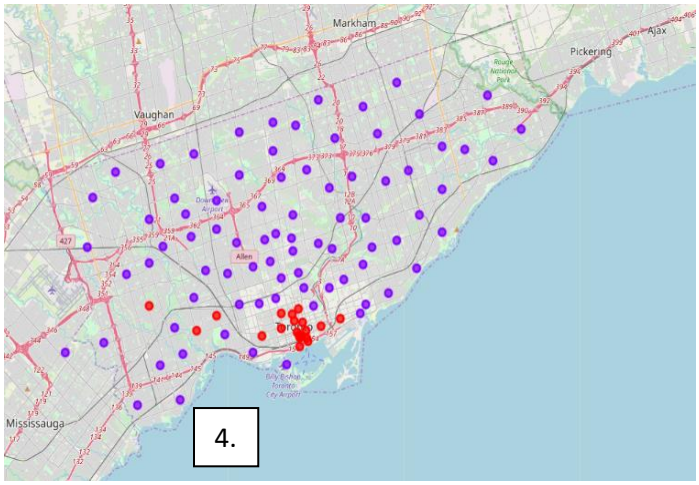


Image #3 is Venues pinned on Toronto before clustering. Image #4 is venues pinned on Toronto after clustering.

The different venues in New York City were grouped into 3 unique clusters represented by different colors in image #2. Now, to meet the initial goal of finding similar neighborhoods in a different city we will create the same map for the other city. The points of similar color in image #2 and again, in image #4 represent a similar neighborhood in terms of venue information we obtained from the Foursquare API.

- **Discussion**:

Because this is a cluster work without supervision, many different approaches can be adopted in order get better results. The project was only done in New York and Toronto, even after performing a dimensional reduction they have many features. Having more samples may result in better clustering. For example, the map viewers can be configured by using DBSCAN algorithm.
Dealing with location data at a deeper level, for example at the neighborhood level may be better clustering of similar data points that could eventually lead to a better cluster. The study Here ends by displaying the data and clustering information on the map of New York and Toronto

- **Conclusion:**

People often move to new cities. And in this growing world full with technology, recommending a neighborhood based on location data is something that should be considered basic now-a-days. The application of neighborhood segmentation is beyond this application too. It can serve as an impressive tool for better organization of city resources. Furthermore, it can be used as a security measurement tool if it is combined with crime data.