The Battle of the Neighborhoods

Brandon Truong
March 2020

1. Introduction and Background

No city is a carbon copy of another due to the variety of venues that contribute to the city's culture. Each city differs in its location and the commodities that are provided by the city and its inhabitants for everyone who comes to the city. Although they are unique, cities can be grouped together by common kinds of neighborhoods. Also, it's possible to segment the venues in a neighborhood according to the venue category, as well as grouping neighborhoods together that are similar. By grouping similar neighborhoods together, it can serve as a variable to aid in making a decision when people want to move to a different city.

The objective is to find identical neighborhoods in different cities in order to help provide people, who are making a decision to choose a neighborhood in another location, some insight.

2. Data Acquisition and Cleaning

This project works with two sets of data. The first set of data will consist of Toronto's different neighborhoods and their geometric coordinates (link is here).The second set of data will have New York's different neighborhoods and their geometric coordinates.

The data pertaining to New York is in json format. Upon examining it, the data needed was under the "features" category. We formatted the json data, resulting in a data frame consisting of 4 columns: borough, neighborhood, latitude, and longitude.

The data of Toronto is from a Wikipedia page that contains the postcode of Toronto. The data is scraped using BeautifulSoup to extract the table data. After further cleaning, the data frame is obtained,

consisting of the postcode, borough, and neighborhood. However, some of the values under the “borough” column are not assigned. To solve this problem, the value of the “borough” column of the row was copied into the “neighborhood” column.

With the obtained data, we can now come up with different venues that New York and Toronto offer. Foursquare API provides us access to a database consisting of all the venues from around the world including information such as addresses, tips, photos, and comments. By providing a function with the neighborhood name and its geographic coordinates, the datasets seemed to share features. Both the data frames were combined in order to perform a clustering operation.

3. Methodology

BeautifulSoup is used to scrape the boroughs from Wikipedia and organize a table containing information of Toronto.
Geopy is used to get the geographical location of each community.

Next, Foursquare API was used to analyze each borough and segment them. The limit was set to 100 venues, and the radius was set to 500 meters for each borough from their given longitude and latitude.

Since the goal of the project is to group together similar neighborhoods in the city of New York and Toronto, K-means will be used to cluster them together. K-means is a machine learning algorithm that groups a dataset into a user-specified number of clusters (k). Since the algorithm is naïve, users utilizing k-means clustering need some way to determine whether they are using the right number of clusters. One way is to validate the number of clusters is the elbow method, whose idea is to run k-means clustering on the dataset for a range of values. For each of those values, the sum of squared errors is calculated, and a line chart is plotted of the SSE for each value of k. If the line chart looks like an arm, then the “elbow” of the line is the value of k to use. We want a small SSE, but as the SSE decreases toward 0, the value of k increases. The goal is to

choose a small value of k that still has a low SSE, and the elbow represents where we start to have diminished returns by increasing k.

Another method to find the optimal cluster number is silhouette analysis, which is a way to measure how close each point in a cluster is to the points in its neighboring clusters. It's a nice way to find the optimum value for k during k-means clustering. The values of a silhouette analysis are in a range between [-1, 1]. A value of +1 means that the sample is far from its neighboring cluster and close to the cluster assigned. Alternatively, a value of -1 means that the point is close to its neighboring cluster and far from the cluster it's assigned to. A value of 0 means that it is at the boundary of the distance between the two clusters. Ideally, values of +1 and -1 are not preferred, so the higher the value the better the cluster configuration.

4. Results

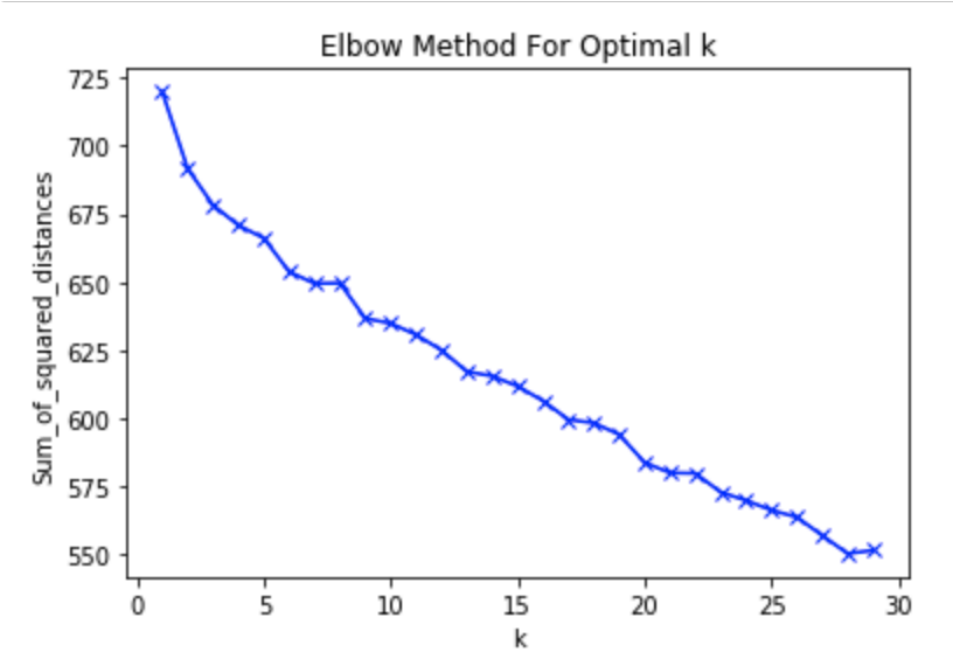

Figure 1: Elbow method to determine the k value

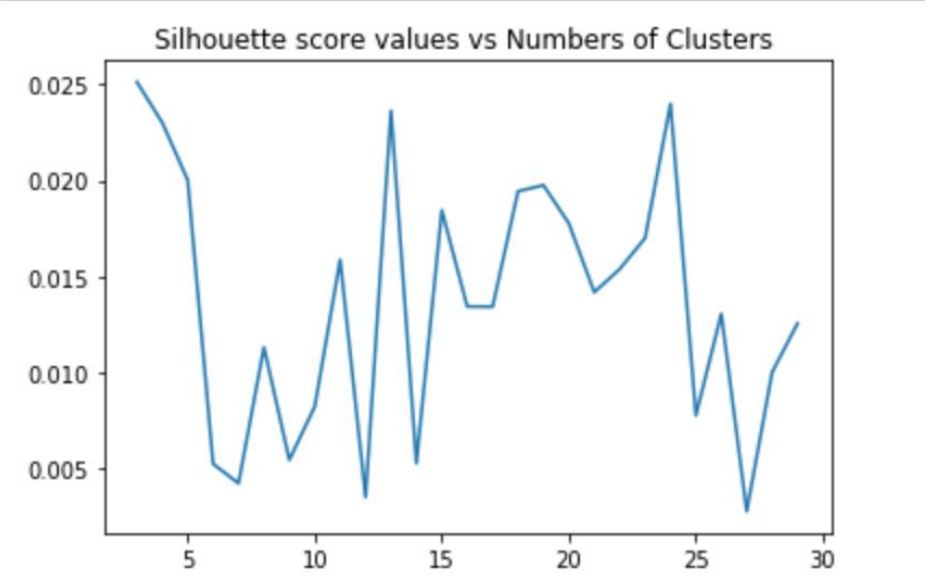

Figure 2: Silhouette Score to determine k value

The Silhouette Score confirms that a number of clusters being 5 has its peak. The Silhouette Coefficient is found using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. The elbow method shows that there is no distinct “elbow,” but we can still come to a conclusion that 15 clusters would be a reasonable choice.

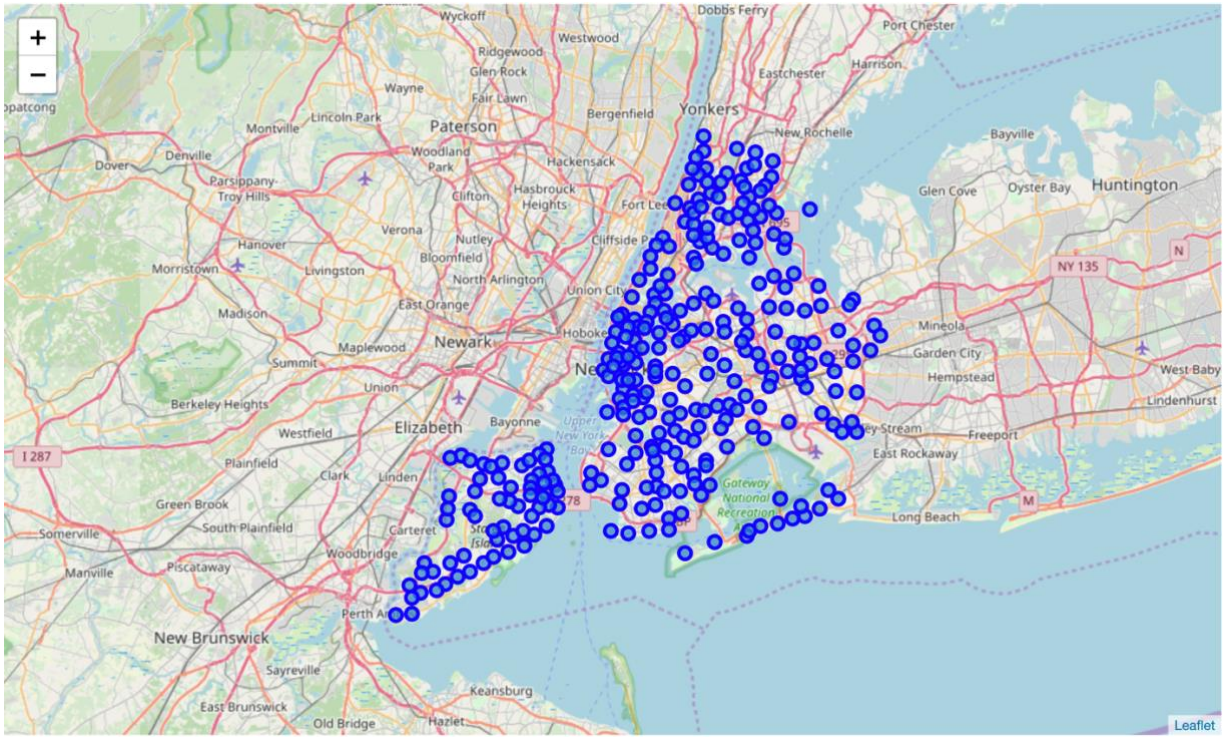

Figure 3: Venues in New York (before clustering)

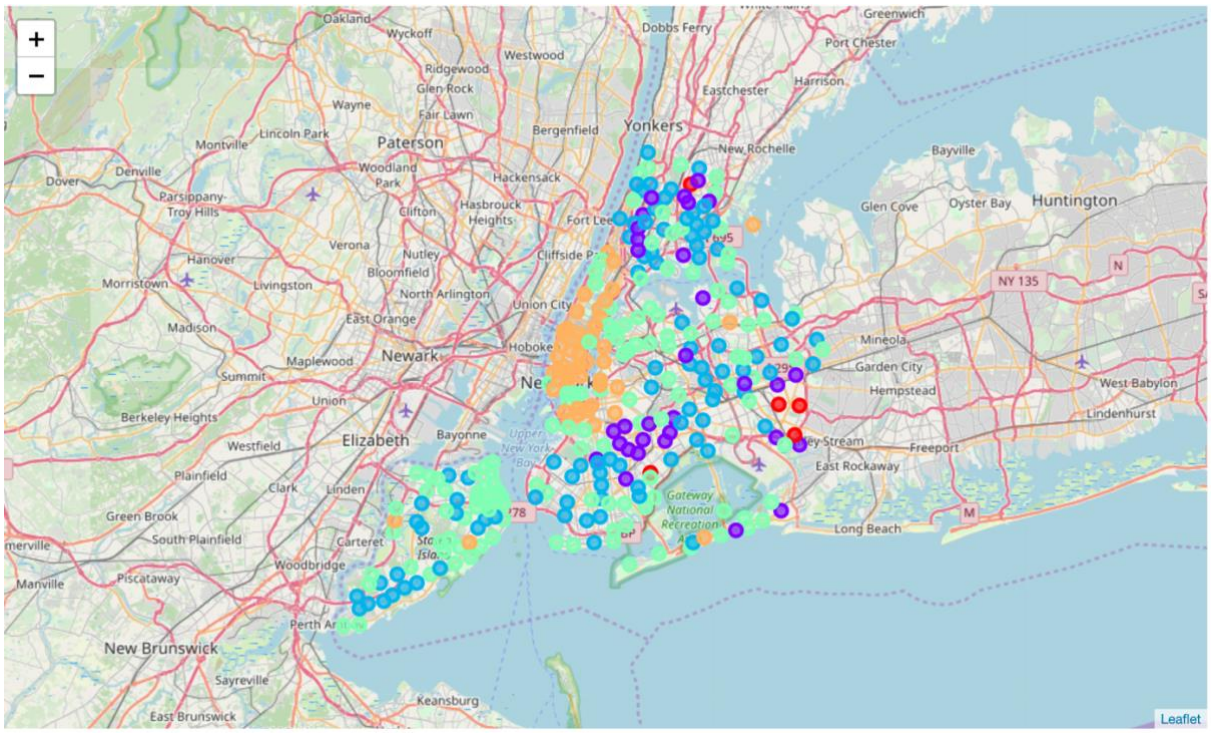

Figure 4: Venues in New York (after clustering)

Venues in New York were able to be clustered into 5 distinct clusters indicated by a different color in Figure 4. The same has been done in Toronto in Figures 5 and 6. Figure 5 shows all of the venues in Toronto, and Figure 6 shows similar neighborhoods according to the venue information obtained from Foursquare API.

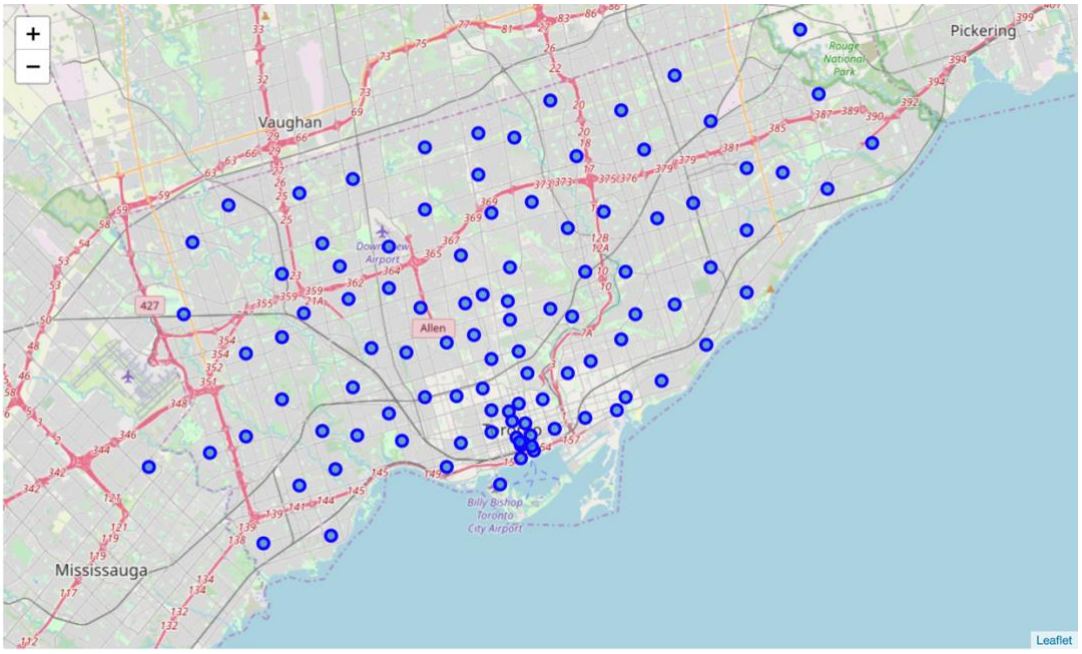

Figure 5: Venues in Toronto (before clustering)

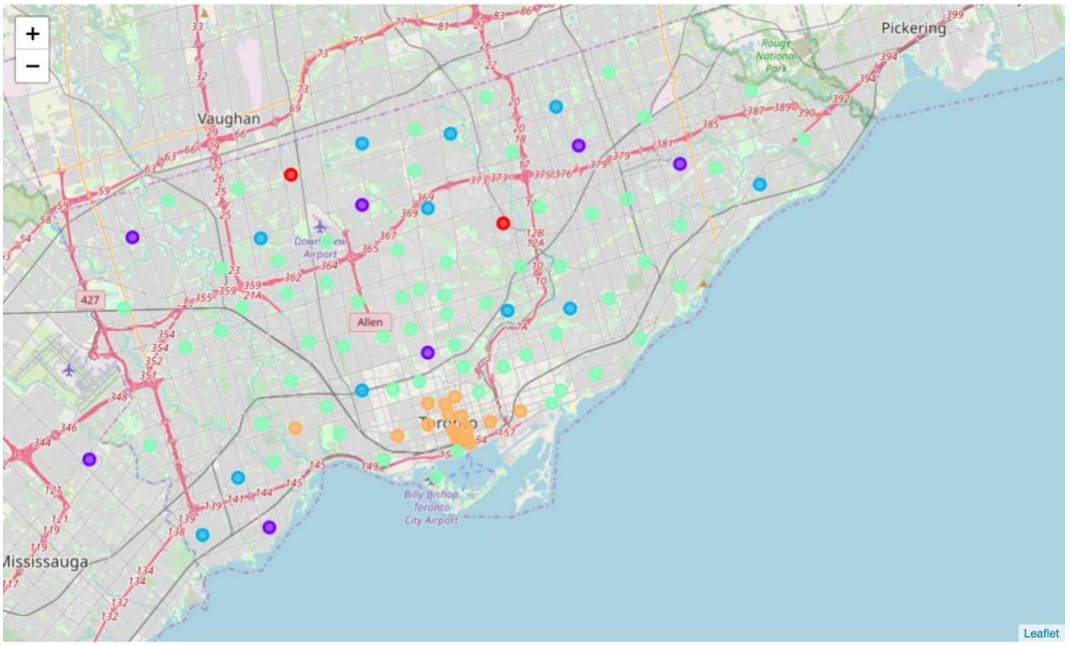

Figure 6: Venues in Toronto (after clustering)

5. Discussion

There are many different approaches to achieve these results, with some approaches better than others, since this is an unsupervised clustering work. This project was done with only the zip codes of New York and Toronto, each having 150 features. More samples introduced into the project could have produced better results. For example, outliers observed on the maps could be included in the data by using DBSCAN algorithm. In addition, further analysis of the location data on a deeper level could have produced better results. The study had stopped when data visualizations and clustering information on the maps were obtained. However, we could go even further by looking at the location data at neighborhood level.

6. Conclusion

People are always moving into new cities. In this time of age where people move to wherever their job takes them, a neighborhood recommendation system that is based on the location data is convenient. In addition, this recommendation system can be used to organize city resources, as well as be combined with crime data to indicate places with high crime rates.