Clustering of Similar Neighborhoods in Different Cities

Aditya Dutta

March 2020

1. Introduction

1.1. Purpose

The purpose of this report is to explore and investigate the distribution of different venues in 2 of the most populous and popular cities in the U.S. and Canada: New York City and Toronto.

1.2. Background

Every city has its own characteristics and various venues spread all over the city. Every city has some common places where people visit often and enable other tourists to explore the city. NYC and Toronto collectively welcome close to 100 million tourists every year. Despite having many dissimilarities, it is possible to take into consideration the similar venues of different cities, segment them and group the neighborhood according to the Venue category. This can give a fair enough idea and reference to help decide when people consider moving out of a city to another and tourists who come to enjoy the stay.

The City of New York, known as New York City (NYC) is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over about 784 km^2. In 2018, NYC welcomed over 65.2 million tourists.

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 6,082,000 as of 2018 distributed over 630.2 km^2. In 2018, Toronto welcomed over 27.5 million visitors.

1.3. Method of Investigation

All the data has been extracted from secondary sources regarding the Borough and Neighborhood Venues along with location data.

2. Data Acquisition & Cleaning

2.1. Data Sources

All the data has been extracted from secondary sources and has been processed to make it ready for analysis. The analysis uses 2 sets of data consisting of NYC and Toronto Neighborhood Location data and the other 2 sets having Venue spots around these neighborhoods.

2.2. Data Cleaning

The NYC and Toronto Neighborhood locations were extracted from geolocation JSON data files. The NYC data was extracted from .json files. All the variables under ‘features’ were extracted consisting of Borough. Neighborhood, Latitude and Longitude. The Borough and Neighborhood data of Toronto was extracted by python’s web scraping package ‘Beautiful Soup’ from URL, which was followed by extracting the location data from URL and finally merging them on common key ‘Postcode’. The Toronto data had to be cleaned by clearing the Not Assigned Postcodes and replacing the Not Assigned Neighborhoods by the assigned Borough Value.

After obtaining all the required location data, we needed to figure out the venue spots near the Neighborhood locations. For this, we used the Foursquare API which grants access to an enormous database consisting of venues from all over the world including a rich variety of info such as an address, tips, photos and comments. With an already logged in credentials, the Foursquare API can be accessed by giving the 'CLIENT ID' and 'CLIENT_SECRET'.

Using Foursquare API, different venues were extracted. One Hot Encoding was performed on the datasets and grouped together. Following this, the NY and Toronto venues data was merged on common venue spots which left us with 250 venue spots per neighborhood.

After this, the data analysis has been performed.

3. Methodology

The goal of this project is to group similar neighborhoods in the city of New York and Toronto. Since the dataset is unlabeled and unsupervised,

3.1. Dimensionality Reduction

Principal Component Analysis is a dimension reduction tool that can be used to reduce a large set of variables to a smaller set by grouping the similar variables that explains the whole dataset. With the use of MinMaxScalar functionality of Scikit learn preprocessing, it scales the dataset such as all features values are in the range {0,1}. This retains the variance of the dataset.

This operation will be used before diving into Clustering of the dataset.

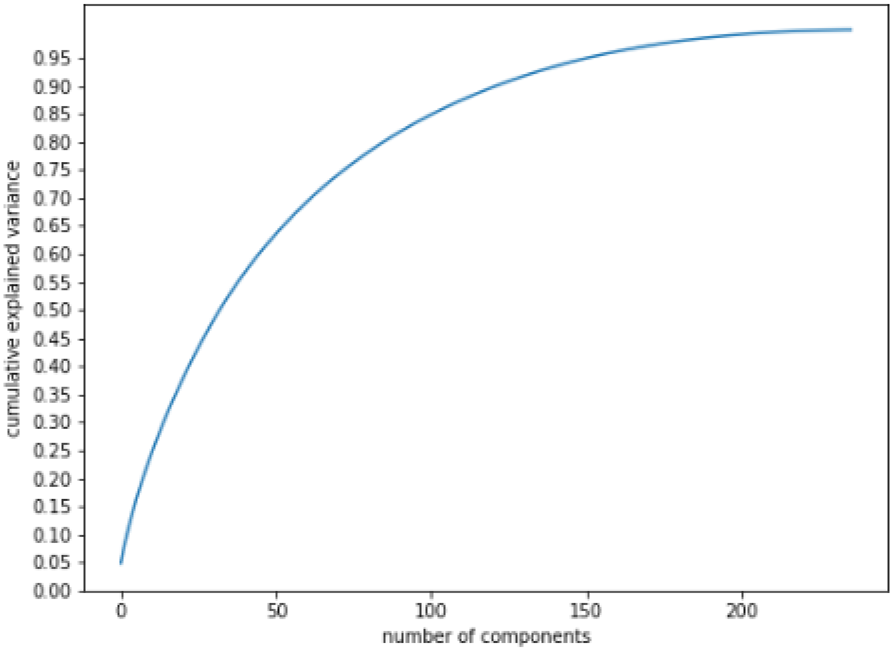


Figure 3.1.(a). PCA operation

Having performed PCA, the number of features was reduced to 175 from 250 yet retaining the maximum variance of the dataset.

3.2. Applying K-Means method of clustering

K-Means is a simple unsupervised machine learning algorithm which groups a dataset into a user-specified number (k) of clusters. Initially the k value is randomly chosen for clustering. As randomly choosing k value for making k clusters seems very vague, this needs a right way to determine the number of clusters. So, we plot a graph of SSE (Sum of Squared Errors) vs k values. If the line chart looks like an arm, then the "elbow" on the arm is the value of k to be used. The idea is that we want a small SSE, but the SSE tends to decrease towards 0 as increase k (SSE is 0 when k=number of datapoints in the dataset). So, the k value that brings a distinct elbow point of the plot gives a rough idea of the best possible number of clusters (k).

The second method to find our optimal cluster number is silhouette analysis. Silhouette analysis is a way of to measure how close each point in a cluster is to the points in its neighboring clusters. It is a neat way to find out the optimum value of k during k-means clustering. Silhouette values lies in the range [-1,1]. The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.

4. Results

4.1. Optimal Number of Clusters

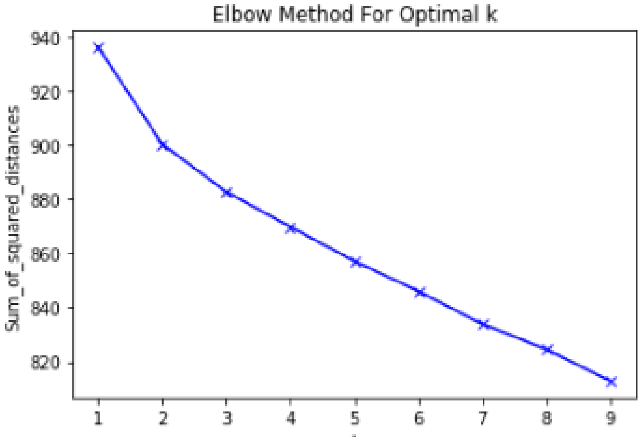


Figure 4.1.(a). Elbow method

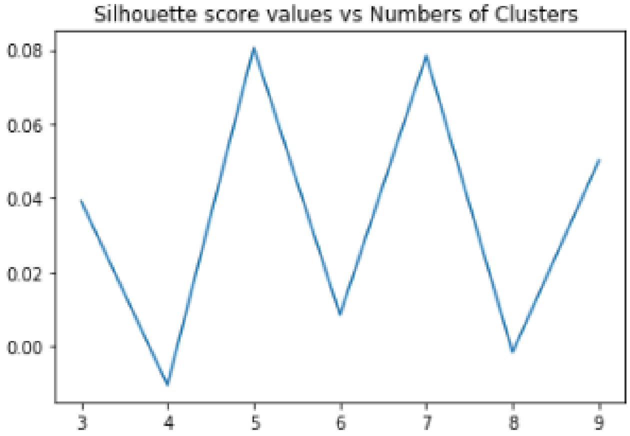
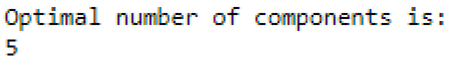
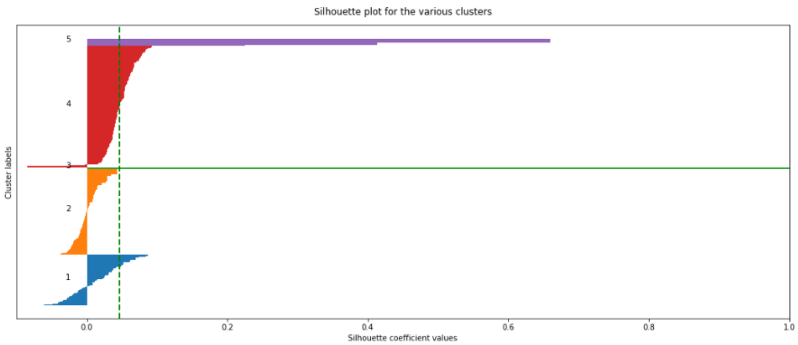


Figure 4.1.(b). Silhouette score method



```
Optimal number of components is:
5
```



As the elbow method seems couldn't provide us with a good enough value of k, Silhouette score is needed to figure out the right k clusters. As we can see that k=5 gives the highest Silhouette score, it has been chosen as the optimal k clusters.

4.2. Visualizing the Clusters on the Map

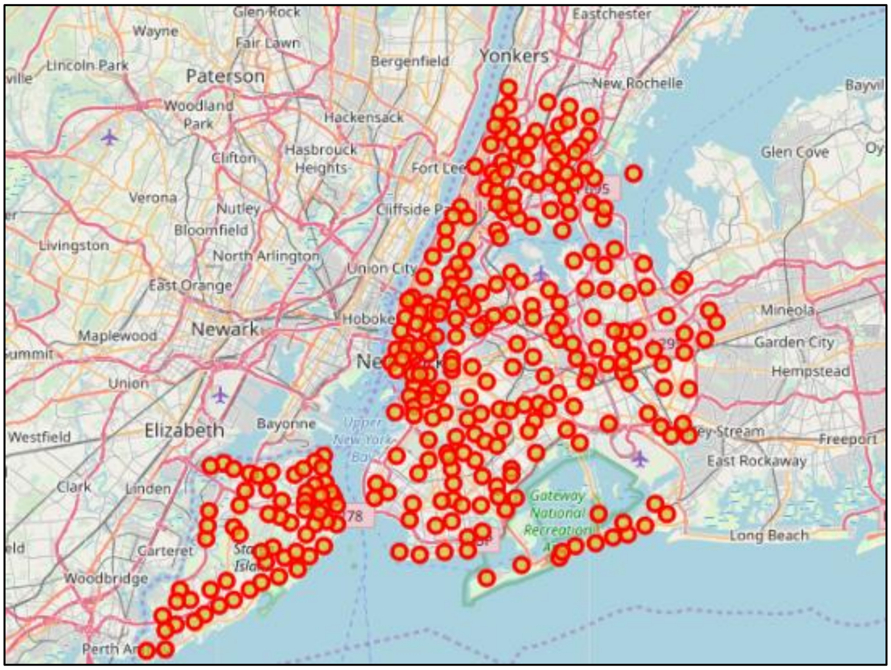


Figure 4.2.(a) New York City before clustering

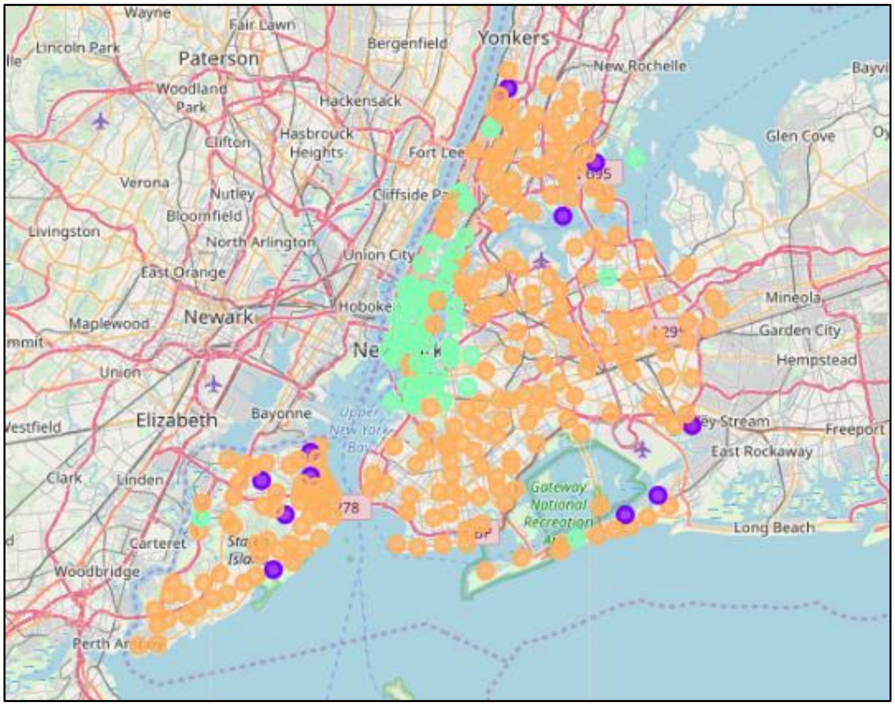


Figure 4.2.(b) New York City after clustering

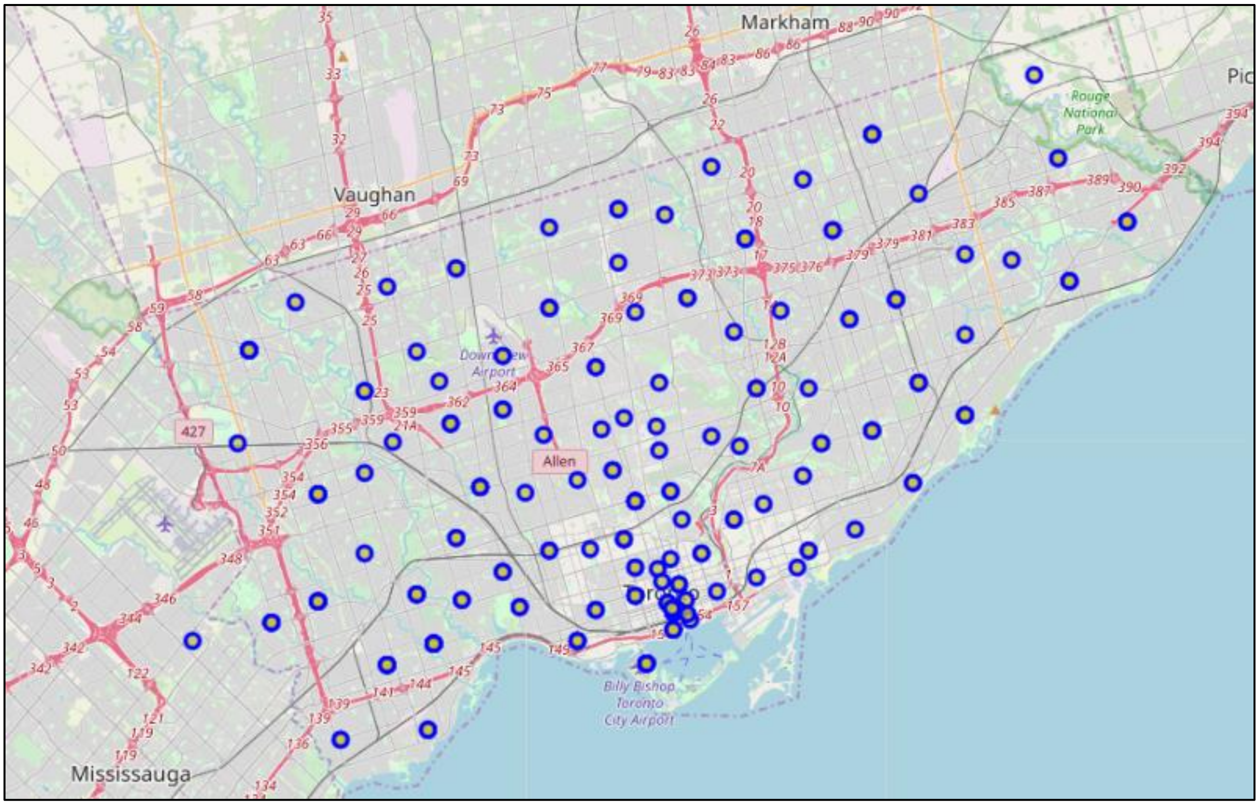


Figure 4.2.(c) Toronto before clustering

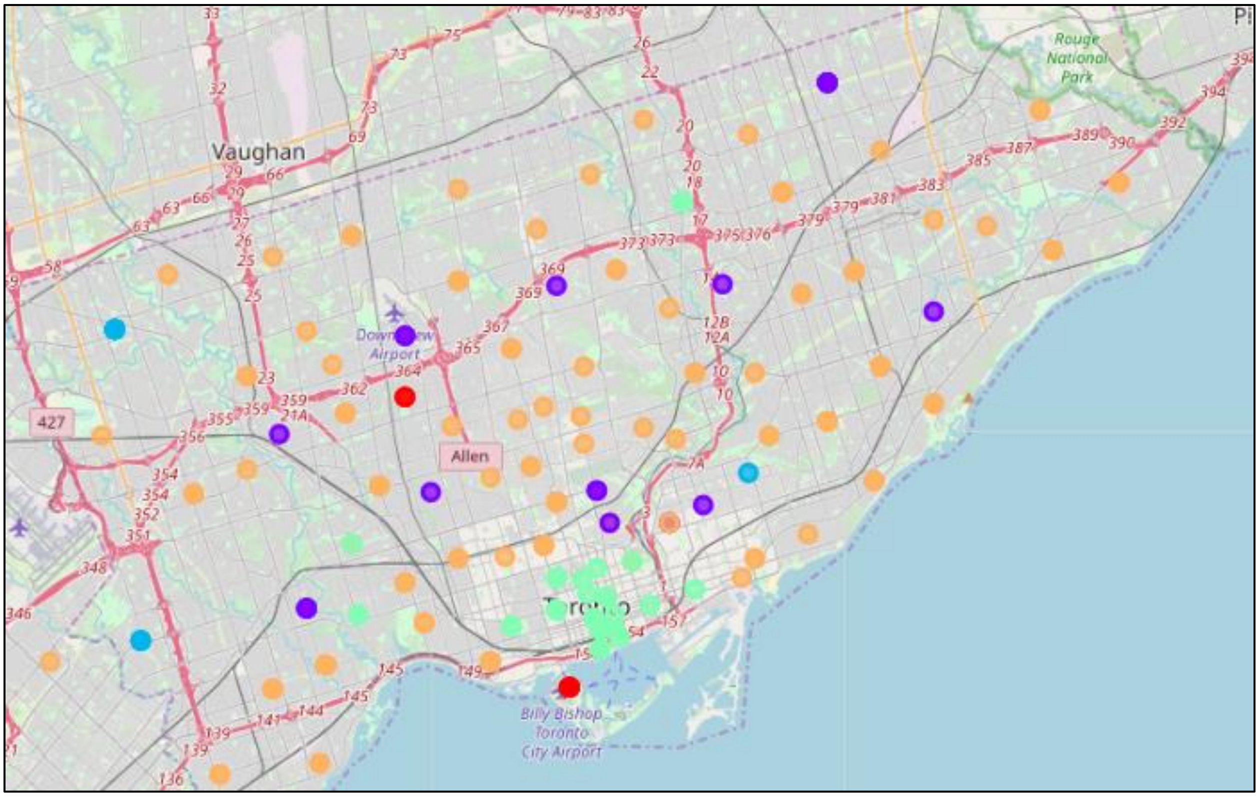


Figure 4.2.(d) Toronto after clustering

The different neighborhoods in the city of NY has been clustered into 5 distinctive clusters represented by different colors in figure 4.2.(b). In the same way the neighborhoods of Toronto city, have been clustered in to 5 distinctive clusters.

We can interpret that neighborhoods in the same clusters have similar venues nearby and a person shifting to any part of NY or Toronto can use this understand where to shift and relocate.

5. Discussions

Since this is an unsupervised clustering work, many different approaches can be adopted in order to achieve better results. The project was only done on the zip codes of New York and Toronto, each having 175 features, after performing dimensionality reduction.

For Instance, for the outliers that have been observed on the maps could be defined by using DBSCAN algorithm.

The study here has been ended by visualization of data and clustering information on the map of the City of New York and Toronto.

6. Conclusion

World is more accessible than it was 20 years ago, and people travelling have increased 100-fold. Neighborhood location clustering and analysis can help people understand new locations and avoid inconvenience and plan their stay accordingly. This can serve to be an impressive tool to better organize a city resource.