The Battle of The Neighborhoods

Aditya Gouroju

April 25 2020

Contents

1 Introduction

1.1 Background

Many people move from one city to another or maybe from one country to another due to their personal or professional causes.As Different cities in the world are filled with numerous kinds of venues that in turn define the cultures of the cities. A city not only differs from another by means of global positioning it also depends upon the interests of the people living in it .Despite the dissimilarities of being different cities it is possible to group the neighborhoods and segment the neighborhoods with the venue categories of personal choices to decide the appropriate neighborhood while moving to the city

1.2 Problem

Finding identical neighborhoods in different cities in order to help provide a perception of similar neighborhoods which may provide with a great deal of insights in order to make a decision of choosing a neighborhood that is far away, yet somewhat feels like home.

1.3 Interest

People who are moving to far away cities would be very much interested in finding out the most appropriate neighborhood they can move to which is similar to the one in which they are currently residing in

2 Data Acquisition and Cleaning

2.1 Data sources

In this project we are going to see how similar or dissimilar the neighborhoods of Toronto and New York cities are.We can get the data of Neighborhoods and its geographical data i.e, their latitudes and longitudes of Toronto by scraping from here,and of New York from here.we have to get all the data of the neighborhoods venues by Foursquare API's

2.2 Data Cleaning

The first data source in the described link is in .json format.Upon examining the data and further formatting of the .json data finally we can convert it into a dataframe that consists of 4 columns, namely: Borough, Neighborhood, Latitude and Longitude by using

Pandas Library. The second data source is a Wikipedia page that contains Postcode of the city of Toronto in a wikitable. we can scrape the page using bs4(Beautiful Soup) library and retrieve the required table as a dataframe using Pandas library .After going through a few more steps, the dataframe was obtained which consists of: PostalCode,Borough and Neighborhood.In this Dataframe the rows with Borough's which are unassigned must be dropped as they will be no use for us.After that if if there are any unassigned neighborhoods we will use them with the same names of their Borough's.

2.3 Feature Selection

Now that we have obtained the different neighborhoods and their respective geometric coordinates for the city of New York and Toronto, To dive further to the problem we are going to need the data of the venues of the respective neighborhoods.For that we are going to need the Foursquare API. Foursquare API provides with an access to an enormous database consisting of venues from all around the world including rich variety of information such as addresses, tips, photos and comments. Having signed up for a Foursquare developer, using the Client ID and Client Secret, it is possible to make API requests in order in order to retrieve venue information. By feeding a function with Neighborhood name and its geometric coordinates, using Foursquare API different venues (Restaurants, Coffee shops, etc) were extracted. After performing One-HotEncoding and grouping together the rows by neighborhoods, the NY dataset and Toronto dataset were combined into a single dataframe in order to perform clustering operation.

2.4 Dimensionality Reduction

Before going into the Analysing part we have to make sure that the model we use should not overfit the data as, overfitting of the model will not be useful in predicting the categories for new data. so we will be using **PCA - Principal Component Analysis** to reduce the number of dimensions i.e the components(columns) to consider for training the model.PCA will help us to reduce the columns and use more significant data instead with less components to get better results.For more information on PCA refer here. For applying this PCA we will normalise the data and plot a graph between the explained cummulative variance and the number of components to consider.

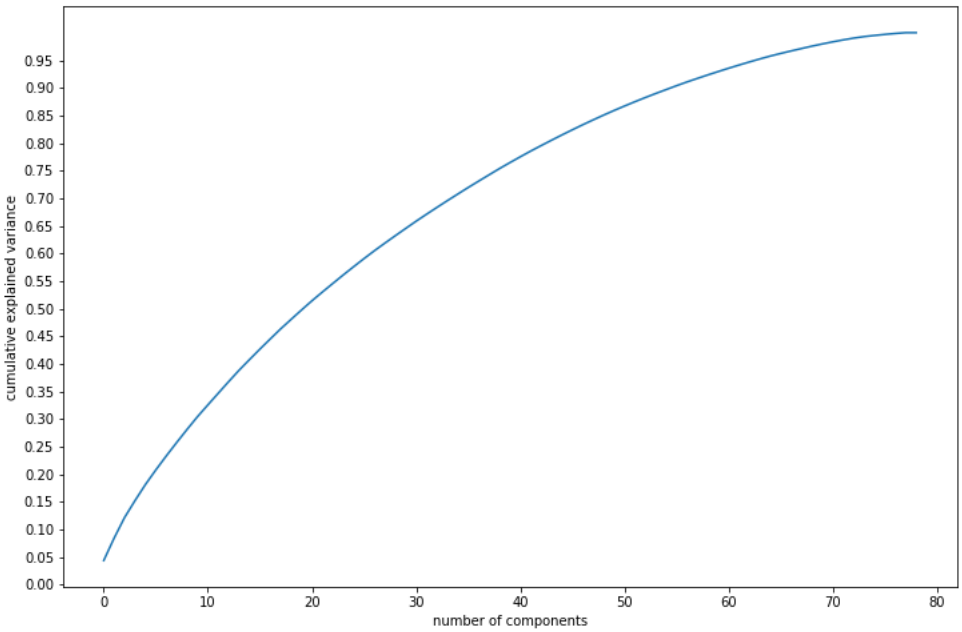

Figure 1: PCA

for optimal model without overfitting we will use 70 PCA components of the data for efficient prediction of the clusters.for that we will fit and transform the normalised data from before into a PCA dataset and use this data as training for the model we will use for more accurate clustering.

3 Methodology

3.1 Data Analysis using K-Means clustering

we will use an unsupervised machine learning model for this problem because, our main aim is to group the similar neighborhoods in both the cities which can be done by this model.

In K-Mean Clustering the program generates K(number of clusters to be formed i.e ,hyper-parameter) random centers then assignes the points to the closest center, then new centered are formed at the centroids of respective clusters and this process iterates until the

centres doesn't change any further.It has some downsides that are ,as the intial centers are random the final clusters are not global optimum. For more information on this Machine Learning method and how it works refer here
To apply **K-Means Clustering** we need to provide the model with optimal number of clusters to form.For that we use **Elbow Point Method** and **Silhouette score Method**.

3.2 Elbow Point Method

In this method to find the optimal number of clusters to form.For that we will plot a graph between the **MSE - Mean Squared Error** and the number of cluster.As the mean square error decreases as the number of clusters increases ,we will try to find the number of clusters where the decrease in MSE is drastically decreased (the point where the slope changes drastically) and use that value of **K**(number of clusters) as the optimal value for number of clusters to be formed.

3.3 Silhouette score Method

In this method we will plot a graph between Silhouette score and k (Number of clusters) where silhouette score is how perfectly a object fits in its own cluster and differs from other clusters, (ranges from -1 to +1, +1 as perfect and -1 as poor fit of k).Generally a value of K for which its silhouette score is more is choosen as optimal value of K. For more information on this method refer here

3.4 Moving Further

From the optimal value of K obtained from the above mentioned methods we will use it to define and train the model and observe the results in the next section.

4 Results

4.1 Elbow Point

The visualization of the graph that is plotted for finding the optimum value of K i.e; elbow point of the graph

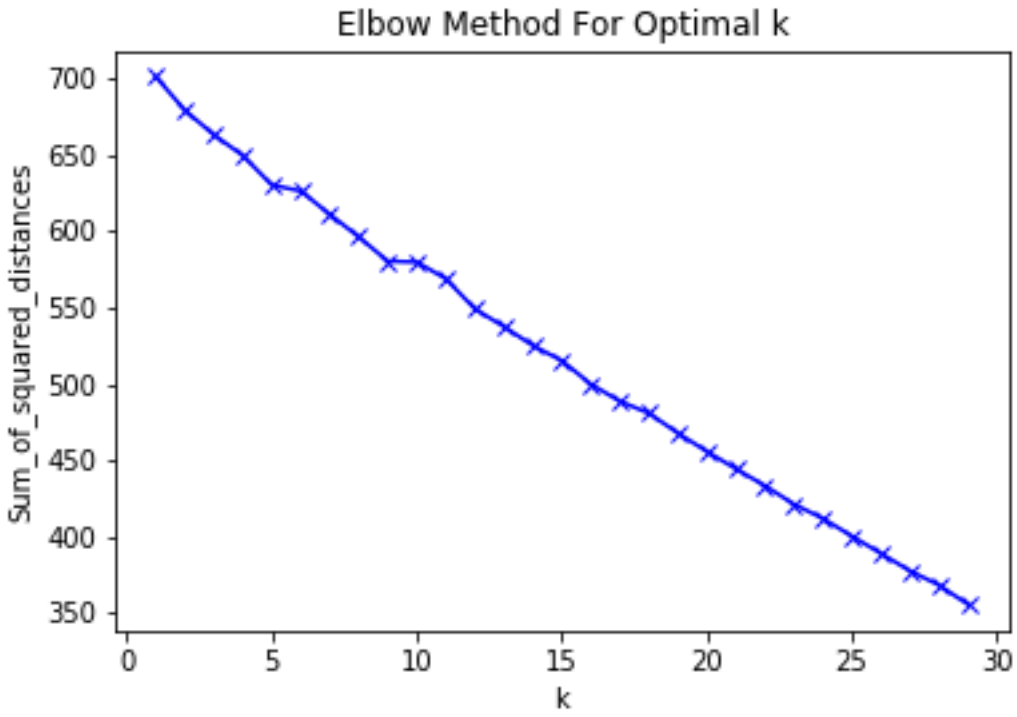

Figure 2: Determination of K by Elbow Point Method

This graph is not much helpful in determining the optimal K as the graph is almost linearly decreasing ,but at values of 5 ,10 of k there is some small decrease in the slope of the graph. Lets check the silhouette score method .

4.2 Silhouette Score

The visualization of the graph between silhouette score and values of K.

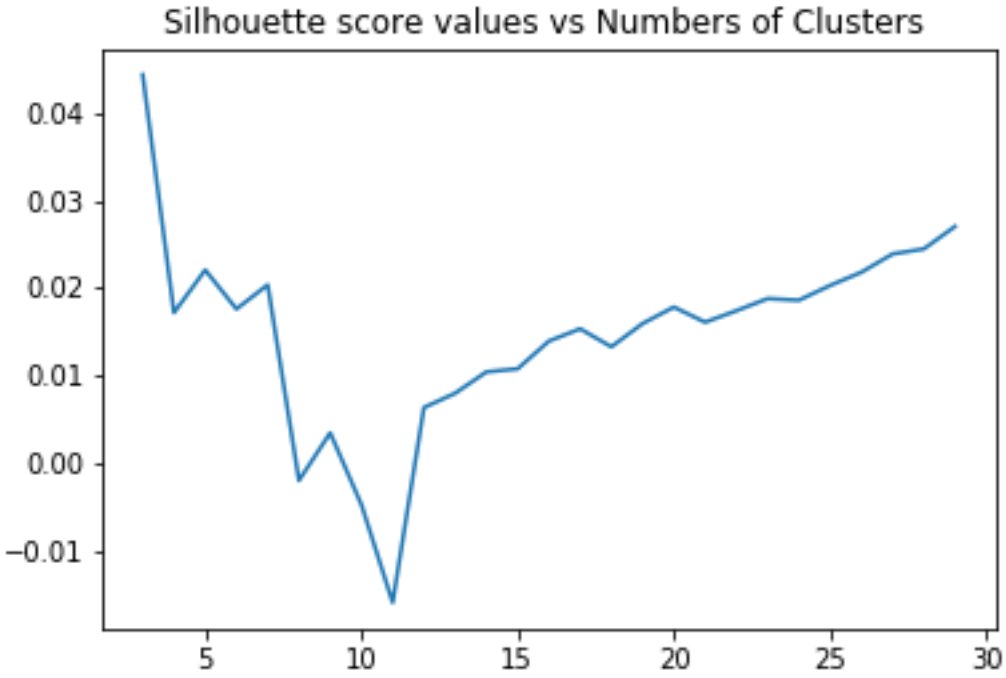

Figure 3: Determination of K by Silhouette Score Method

From this figure we can observe that the Silhouette Score is maximum for K = 3 but if we use 3 as the number of clusters to form it will not be suitable as there are many neighborhoods to consider and we cannot just segment them into only 3 different clusters, this may not be appropriate. If we consider the next highest Silhouette Score for the value of K i.e , 5 it is a good number to choose as it is the second highest value of silhouette score and segmenting the neighborhoods into 5 clusters is more favourable than just 3.

Hence we will segment the neighborhoods into 5 clusters for appropriate results.

4.3 Visualization of The Clustering in Maps

4.3.1 City of Toronto

we will now visualize the clustered neighborhoods of the city of Toronto in folium Maps

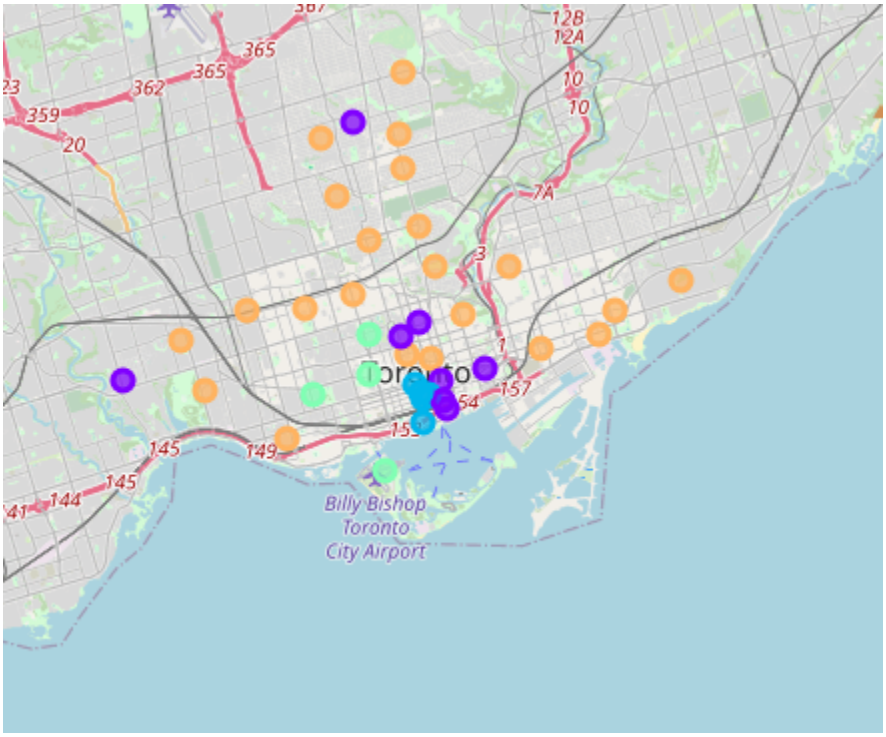

Figure 4: Neighborhoods of Toronto Segmented

The Neighborhoods with the markers of same color are of similar type of neighborhoods based on the venue categories present in the area.

4.3.2 City of New York

we will now visualize the clustered neighborhoods of the city of New York in folium Maps and compare it with the map of Toronto

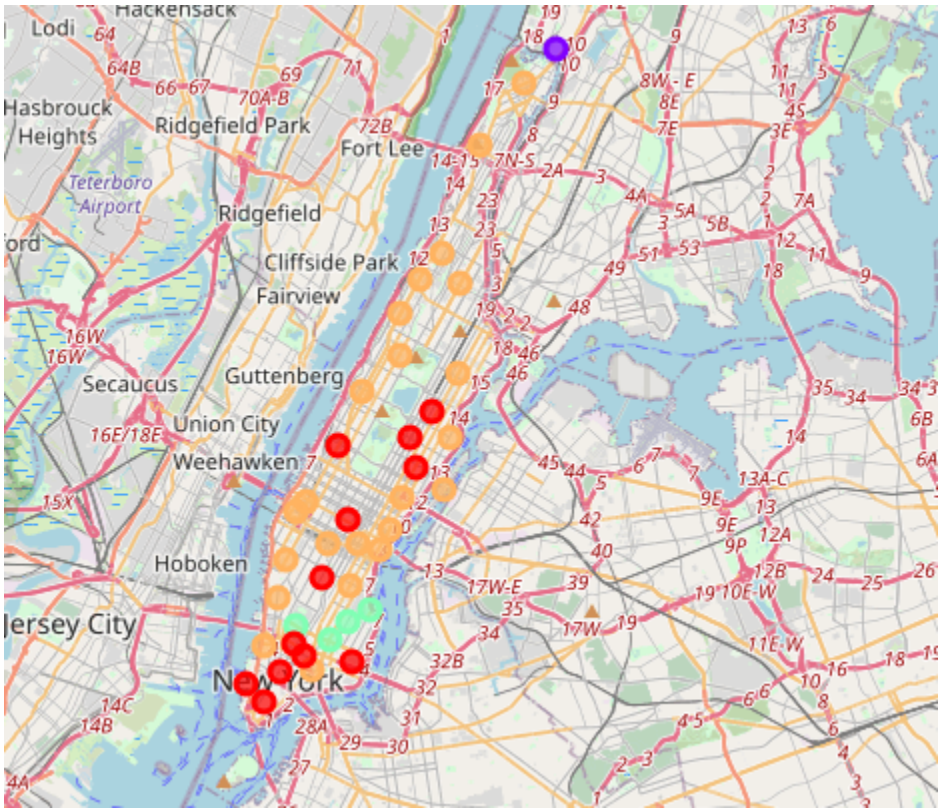

Figure 5: Neighborhoods of New York Segmented

The Neighborhoods with the markers of same color are of similar type of neighborhoods based on the venue categories present in the area. We can compare them with the Toronto Map.

5 Discussion

By comparing the maps of both the cities we can find out the similar type of neighborhoods based on the venue categories on which people moving from Toronto to New York or vice-versa can consider based on the neighborhood they are living now.

Also, For people who are more considerate on the famous or popular venue category in the neighborhood they are presently living in and they want that specific venue category must be also be famous or popular,I have also worked on it by making a dataframe of the neighborhoods with its most popular venue categories, you can find the dataframe in the project notebook I used for this project here

All the Data on the venues and the venue categories are according to the Foursquare API's .As it was not possible to request and access large amount of data on the neighborhood venues and get all the required information at a time .I limited my project only to the Data of venues in Manhattan of New York City and Main parts of Toronto City

The clustering of Neighborhoods based on their venue categories is done by K-Means clustering So, it is not possible to find the outliers in the clustering.We can even use DBSCAN clustering method .By Using DBSCAN method for clustering we can get more accurate results and identify the outliers in the clustering.For more information about DBSCAN refer here

6 Conclusion

By this project people who are moving to another place can make use of this type data which helps them to consider the most appropriate neighborhood based on the venue categories present in the neighborhood which is similar to the one they are currently residing in. As they can yet fell the new neighborhood as their home .