Clustering & Comparing Neighbourhoods In New York & Toronto

Induru Bimsara Wijesinghe
IBM

1.Introduction

1.1 Background

In this project we will study,analyze,cluster and compare the neighbourhoods of two important cities in the world. Newyork and Toronto are largest as well as financial and tourist capitals of the countries United state and Canada respectively.There are roughly about 8.39million residents in Newyork and 2.93 million residents in Toronto. Newyork and Toronto are both huge,diverse and cosmopolitan cities.
New York City (NYC) is one of the most populous cities in the United States of America. Also, NYC is the most linguistically diverse city in the world: as many as 800 languages are spoken in it. Moreover, NYC plays an essential role in the economics of the USA: if New York City were a sovereign state, it would have the 12th highest GDP in the world. New York City consists of five boroughs: Brooklyn, Queens, Manhattan, The Bronx, and Staten Island.
The second city of interest in this project is Toronto. As with NYC in the USA, Toronto is the most populous city in Canada. It's recognized as one of the most multicultural and cosmopolitan cities in the world. Toronto also is a very diverse city: over 160 languages are spoken in it. On the economic side, Toronto is an international centre for business and finance and it is considered the financial capital of Canada.

Since both countries are near, people move to two cities for various reasons.they move in order to get a job,to start a business ,to shop,to travel and various other reasons.

1.2.Problem statement

Suppose a person wants to move from Newyork to Toronto for a job.This person does not know anything about Toronto and he would like to move into a place where he lives now.
Is it possible to create a system that can help our users showing similarities between two cities?

Suppose a businessman wants to move from Newyork to Toronto to start a business.He wants to know what type of what types of businesses are more likely to thrive in both cities, what are the neighborhoods that are suitable for each type of business, and what

types of businesses are not very desirable in each city.If he knows this he can get better and more effective decisions regarding where to open their businesses.
Is it possible to create a system that can help our businessman showing similarities and differences in businesses between two cities?

1.3 Approach

Foursquare is a website where people comment and rank food sites, coffee sites, malls and parks. For instance, let's think that a Foursquare user had to move from New York city, USA to the city of Toronto, Canada. Foursquare location data along with a clustering algorithm can suggest a neighborhood in order to help this user to live in Toronto in a similar place. The neighborhood that will be suggested, will not be a random suggestion, but instead will be a place for his pleasure. Thus, previous data from New York and Toronto will be used to predict a good living neighborhood for him.

1.4 Target Audience

People who seek a new job in Toronto/New york
Businessmen start new business in these cities
Residents who move between these two cities.

2. Data

In order to analyse the cities on a meaningful level, they need to be divided into different areas, e.g. neighborhoods, boroughs.A list of neighborhoods in New York and Toronto is downloaded and their respective location in longitude and latitude coordinates is obtained. The sources are the following:

Newyork
https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json

Toronto
https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&direction=next&oldid=942655364

Foursquare API will be used for this project.Moreover, their specific coordinates are merged. Only Manhattan neighborhoods and boroughs that contain the string "Toronto" are taken into account. A Foursquare API GET request is sent in order to adquire the surrounds venues that are within a radius of 500m. The data is formated using one hot encoding with the categories of each venue. Then, the venues are grouped by neighborhoods computing the mean of each feature.
The similarities will be determined based on the frequency of the categories found in the neighborhoods. These similarities found are a strong indicator for a user and can help him to decide whether to move in a particular neighborhood near the center of Toronto or not.

3. Methodology

3.1 Feature Extraction

For feature extraction One Hot Encoding is used in terms of categories. Therefore, each feature is a category that belongs to a venue. Each feature becomes binary, this means that 1 means this category is found in the venue and 0 means the opposite. Then, all the venues are grouped by the neighborhoods, computing at the same time the mean. This will give us a venue for each row and each column will contain the frequency of occurrence of that particular category.

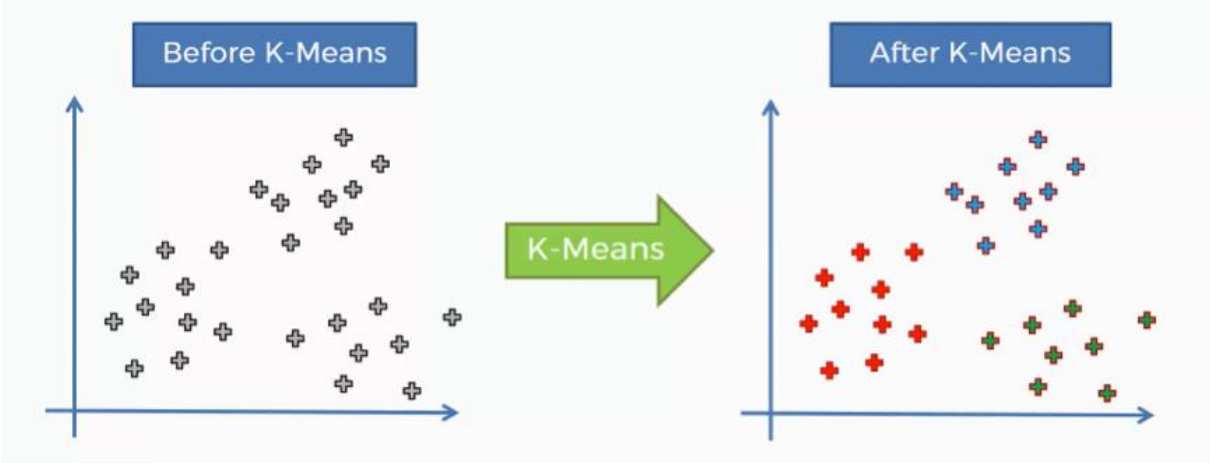
3.2 Unsupervised Learning

For the purpose of doing unsupervised learning to found similarities between neighborhoods, a clustering algorithm is implemented. In this case K-Means is used due to its simplicity and its similiraty approach to found patterns.

- K-Means:

K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the muldimensional features.

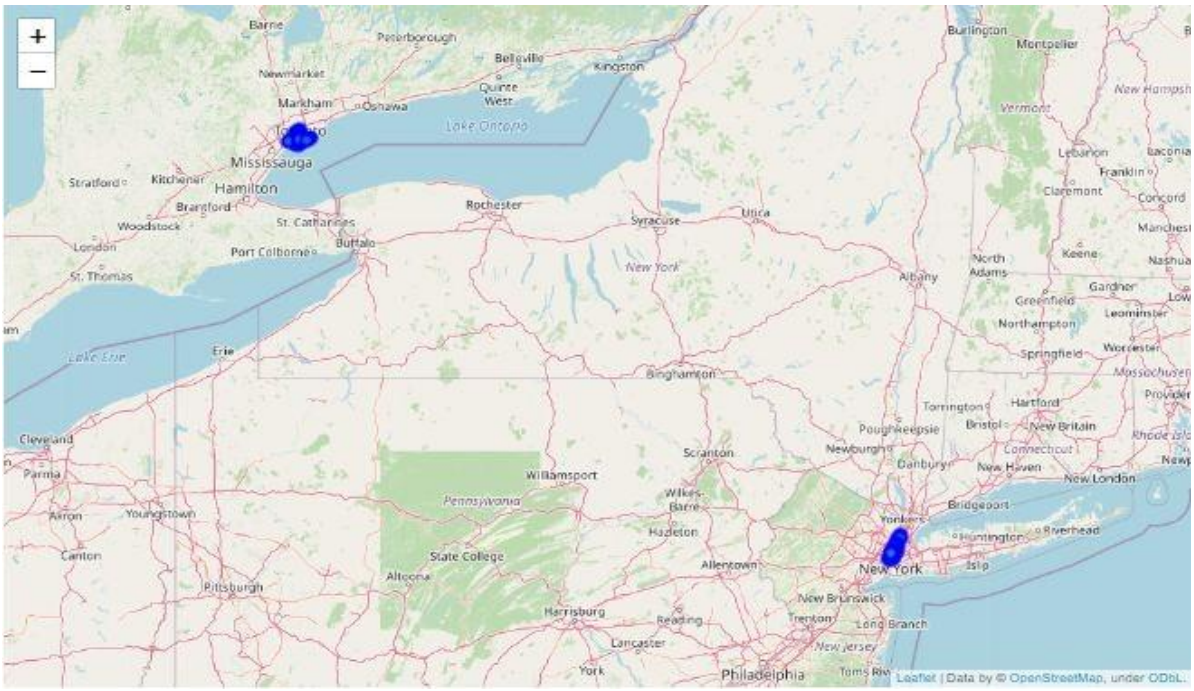
In the following figure there is a graphical example of how a K-Means algorithm works. As it is possible to see, dispersion is minimized by representing all clustered data into one group or cluster.



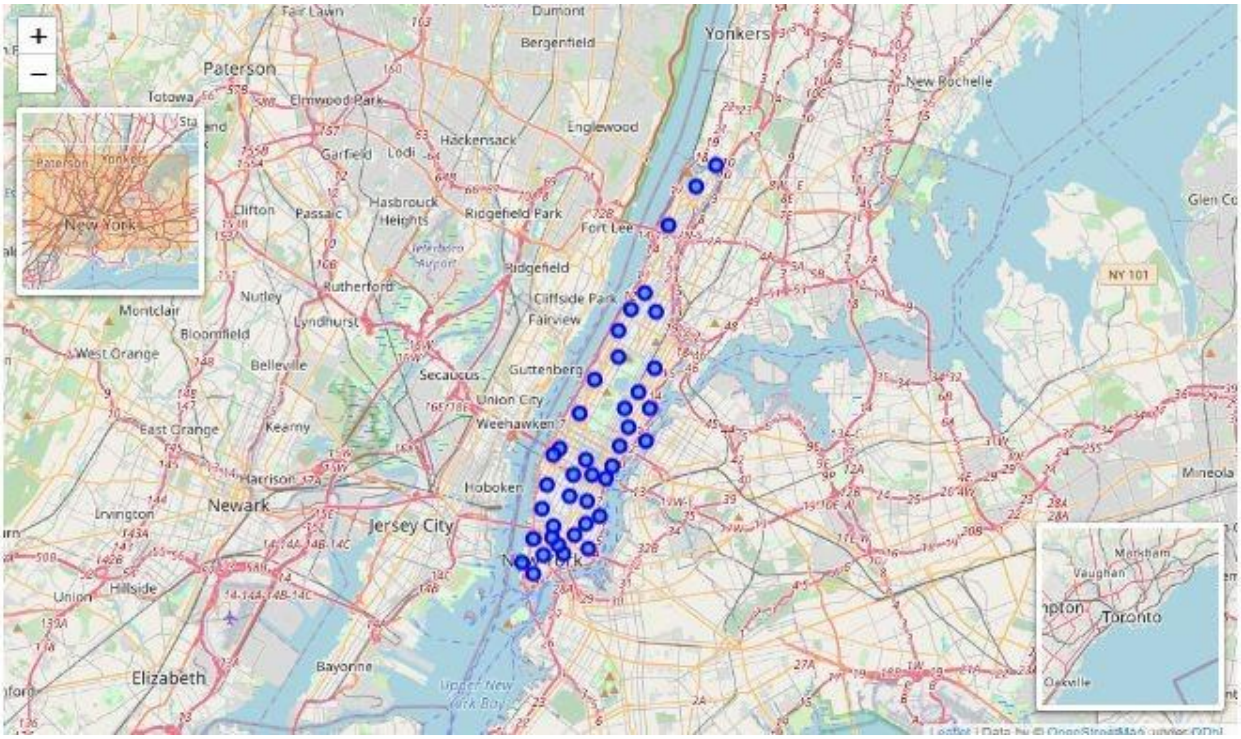
It is necessary for this algorithm to have a prior idea about the number of clusters since it is considered an input of this algorithm. For this reason, the elbow method is implemented. A chart that compares error vs number of cluster is done and the elbow is selected. Then, further analysis of each cluster is done.

4. Results

Initially, data is plotted in a geographical map to get a notion of the world location. First image shows cluster of neighborhoods in two cities and other two images are shown the neighborhoods in Newyork and Toronto.



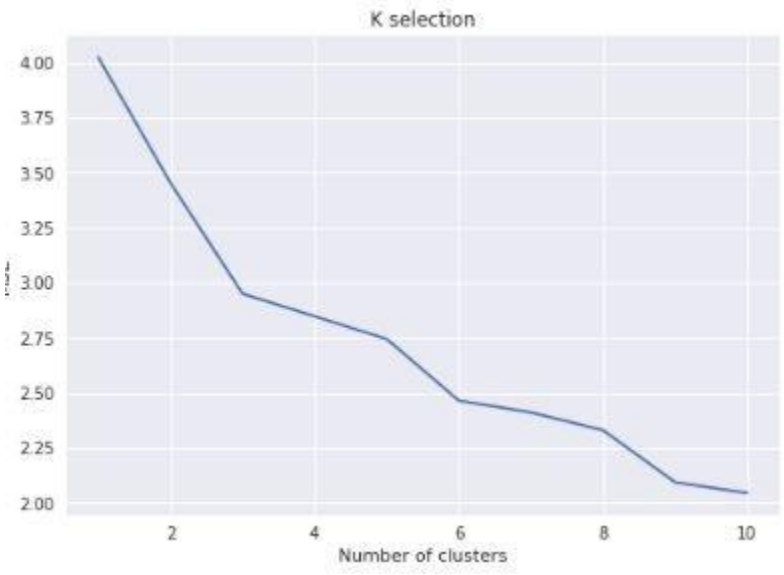
Clustered image of neighborhoods in newyork and Toronto



Neighborhoods in Newyork

Neighborhoods in Toronto

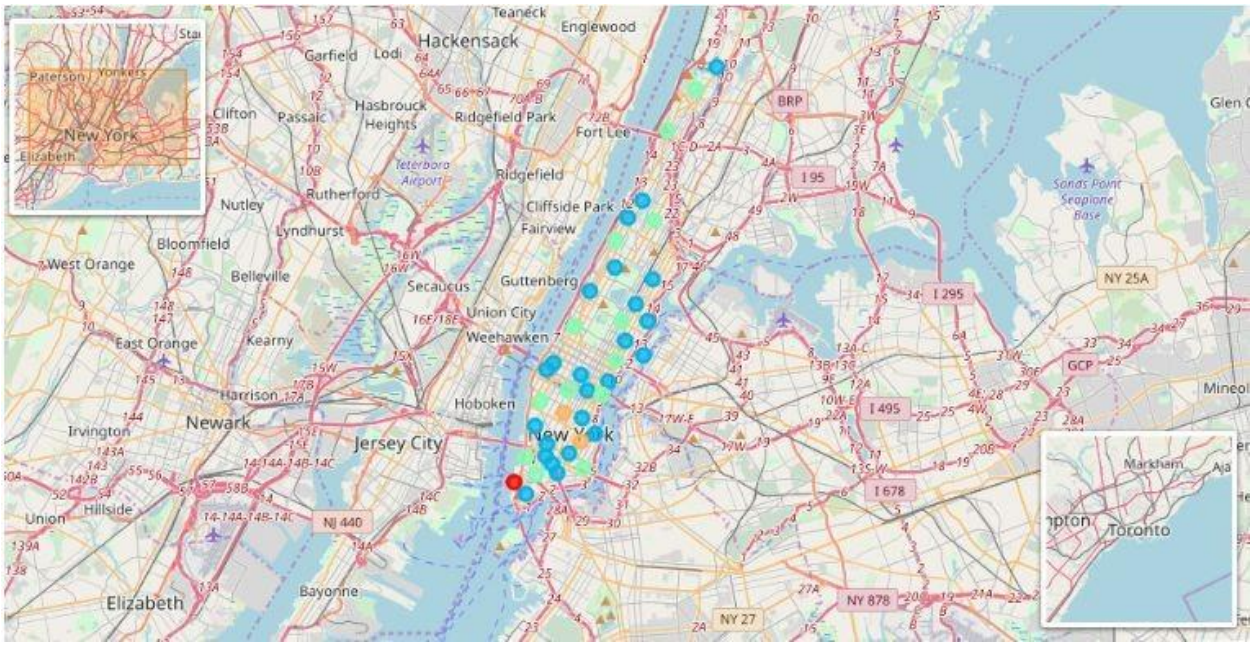
After That,the cluster algorithm is implemented. For this purpose, it is necessary to have a prior idea about the number of clusters. Therefore, the mean squared error (MSE) is plotted vs the number of clusters. The number of clusters start with a value of 1 increasing until a value of 10. This chart is shown in the image below.



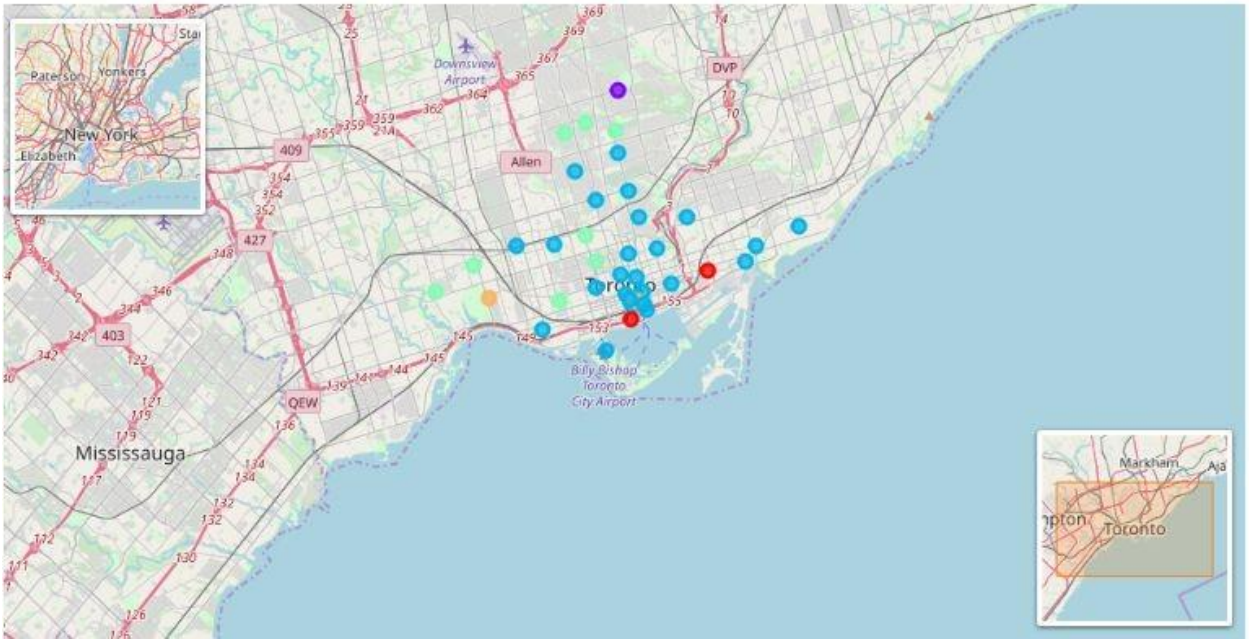
K-selection for number of clusters

As it is expected, the MSE decreases over the number of clusters. The elbow method here is implemented in order to select the appropriate number of groups. In this case, it is possible to see that the elbow is found more or less around 5. The MSE found below this number shows little changes rather than big ones. Finally, once the number of clusters is fixed, the clustering algorithm is repeated through samples and each neighborhood is labeled according to the clusters found.

For visualization purposes, the geographical data is again plotted but with different colors. Each color represents the cluster for which that neighborhood belongs. This image is shown below.



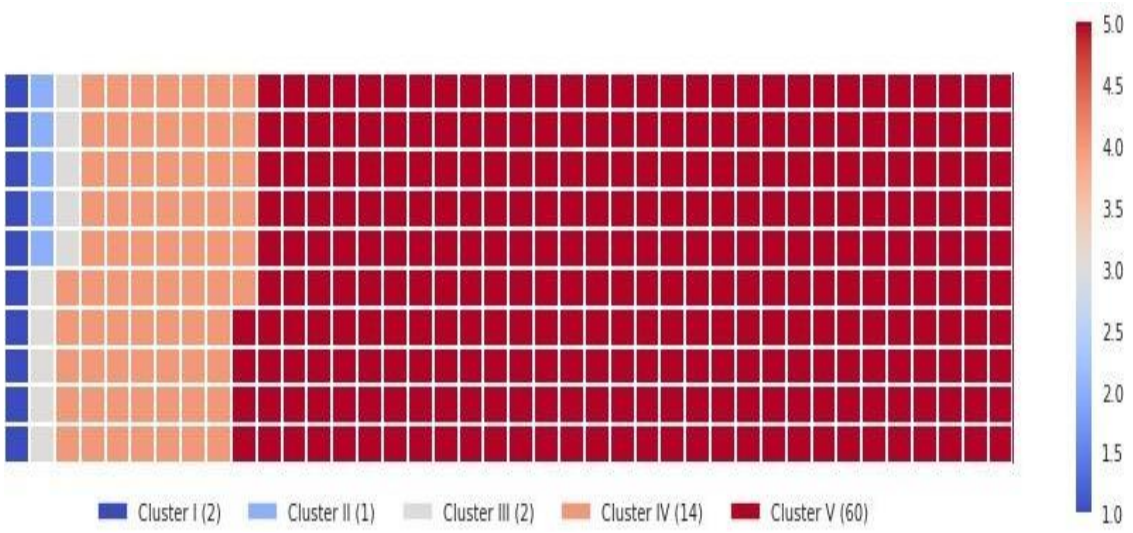
Neighborhoods in Newyork



Neighborhoods in Toronto

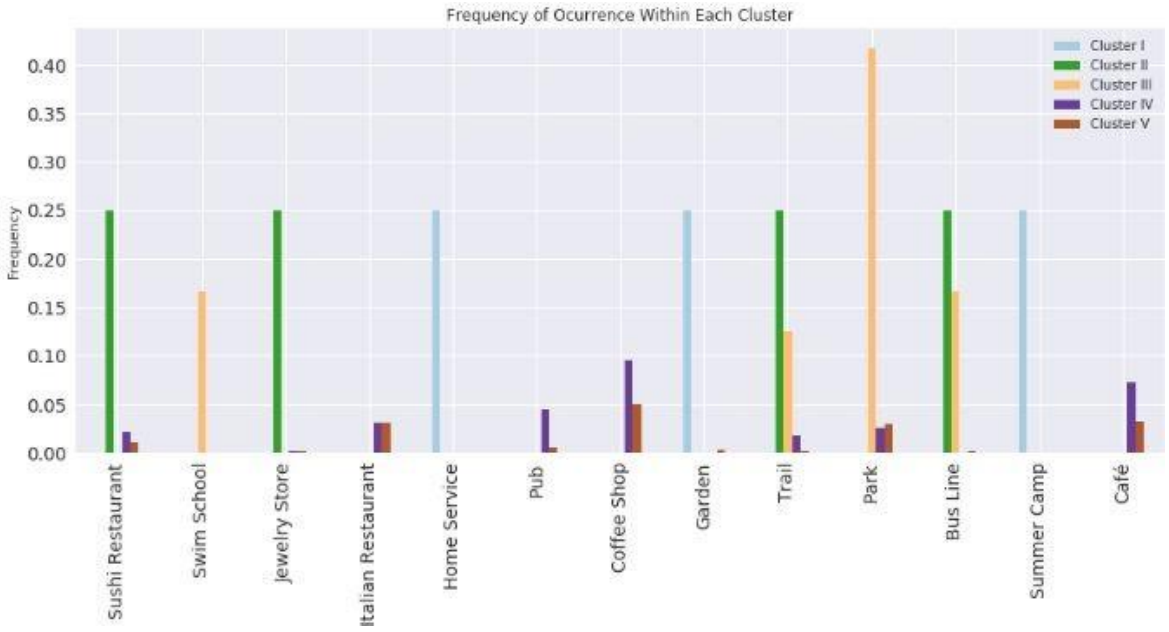
In above images, it is evident that cluster algorithm is not segmenting the neighborhoods for location areas. This means that it is not true that geolocation of neighborhoods is correlated with the categories of the venues around each neighborhood. Yet, it is possible to see which neighborhoods within Manhattan, New York are more similar to the neighborhoods within Toronto. Those neighborhoods that are similiar among them belong to the same cluster. Hence, they have the same color in the image above.

In the image below it is found the proportion of the neighborhoods assigned to each cluster. For this reason a waffle chart is implemented. There are two major clusters and three minor clusters.Cluster V obtained 60 neighborhoods which is the highest proportion.Cluster II obtained lowest frequency



For practical purposes, a word cloud is shown in the image below. In this way, a person trying to locate a similar neighborhood in Toronto can locate it looking for the neighborhoods with the same color. Here we can see that Chinatown, WashingtonHeights, HamiltonHeights and so on belong are similar among them. On the other hand, StJamesTown, MarbleHill North Midtown and Manhattanville so on are different from those we mentioned earlier but again they are similar among them.

for research purposes, bar charts are employed to found insights within the clusters. The bar chart that is shown below shows the features with higher frequency in the centroids found by the algorithm. In this way we can learn what the algorithm is finding.



The image above shows a particular category "Home Service" and “Summer Camp” with a high frequency of 1. This category is related to the I cluster.
It is possible to see that I cluster focuses on neighborhoods that have around Home Service, Garden and Summer Camp. On the other hand II cluster focuses on neighborhoods that have around ,Sushi Restaurant,Jewellery Store,trail and Bus Line Third (III) cluster focuses on neighborhoods that have around Park, and Swim School . The (IV) cluster focuses on neighborhoods that have around coffee shops,pubs and cafes .The final cluster focuses on neighborhoods that have around Italian Restaurant.

5. Discussion

It is worth to note that this work is useful only for those who live in Manhattan, New York or in the neighborhoods near the center of Toronto. The reason is because there is a limited amount of data we can request using de Foursquare API. Consequently, it will has a greater cost than the Lite version.

6. Conclusion

Section where you conclude the report.

In this work a segmentation between two different countries is done. This segementation involves the neighborhoods in Manhattan, New York and the neighborhoods near to the center of Toronto. The data is downloaded and the venues around the neighborhoods is adquired using the Foursquare API. One Hot Encoding is used for converting the categories of the venues into a feature matrix. Then, all venues are grouped by neighborhoods and at the same time the mean is calculated. Hence, the resulting features used are the frequency of occurrence from each category in a neighborhood.

The K-Means clustering algorithm is used for finding similirities between all the neighborhoods listed in the feature matrix. The elbow method is used for selecting the appropriate number of clusters. Hence, the K selected is 5. Results show that there are 2 major groups and 2 minor groups. In adition, there is one group that contains only one neighborhood that is isolated from others. The description of the clusters is the following:

Cluster

- I: Neighborhoods that have around garden,Home services,Summer camp .
- II: Neighborhoods that have around sushi Restaurant,Jewellery store,Trail and Busline
- III: Neighborhoods that have around park,swim school.
- IV: Neighborhood that have around Cofe shops,pubs and cafes.
- V: Neighborhoods that have around Italian Restaurant.

Finally, any user who wants to move from manhattan to toronto and viceversa can use this system to get a notion or idea about what is the best suitable place for him.