Comparison between neighborhoods of New York City and the city of Toronto

1-Introduction

1.1-Background

New York city is most populous city in United Sates with 86 lakh population and same way Toronto is capital city of capital of the province of Ontario, is a major Canadian city with 27 lakh population.

Living cost and many other parameters are high in New York in comparison of Toronto while both are same level city in their respective countries.

As we know that prosperity of neighborhoods of any country creates life standard of that country. In this project I would like to do analysis on neighborhoods of each country so that we can compare how neighborhoods of one country is different from neighborhoods of another country.

1.2 Problem

if any investor has investing strategy for one country and now he wants to invest in different country then he would have to change his strategy according to new country.

In this project I will dig out some specific pattern in neighborhoods of both and will investigate about specific properties of Neighborhoods of both countries E.g. (Diversity of neighborhoods of both Countries, Capital strength of neighborhoods, ….) which will help them to make strategy accordingly

2. Data acquisition and cleaning

2.1- Data sources

Data has been taken form below resources:

1- New York data will be scraped from below link
https://cocl.us/new_york_dataset

2-Toronto Data will be scraped form below link

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

3-geocoder package for collecting Postal code, latitudes, longitudes of Neighborhoods.

4-We will use **Foursquare API** for collecting venues data of New York and Toronto within 500 meter radius.

2.2- Data cleaning and Preprocessing

From above URL we get raw data where many rows have Borough and Neighborhoods as 'Not Assigned'.

We will drop those rows where Borough is not assigned.

We will replace value of Neighborhoods with name of Borough if Neighborhoods is not assigned.

We will put all neighborhoods in same cell if postal code is same.

2.3- Feature selection

I will select below features form venues and city data.

Borough, Neighborhood, Neighborhood Latitude, Neighborhood Longitude , Venue,Venue Latitude, Venue Longitude, Venue Category

3-Methodology

I will use below libraries in this project.

1-Pandas

2-Numpy

3-ScikitLearn

4-Matplotlib

5- Beautiful Soup

3.1 Model evolution

I will use sum of squared error to figure out number of cluster required for clustering of New York and Toronto data

4-Data Analysis

In this project I will focus on venue data in both countries from there I will analyze distribution of venues in both countries, With the help of distribution of venues I will summarize my results regarding how neighborhoods of both countries differ.

Below are limitations which is applicable in my analysis.

1- I will consider only those venues which are within 500-meter radius from Neighborhood.

2- Maximum Number of venues per neighborhood which I am considering for analysis is 100

After preprocessing and transformation of data we got below data.

total 20551 venues in New York data and 4842 in Toronto (Above condition is applicable radius= 500 Meter and Maximum number of venues in each Neighborhoods=100)

Total neighborhoods in New York=306

Total neighborhoods in Toronto=103

Analysis 1-For Small Neighborhoods of both countries:

From total number of venues in both countries it is clear that many neighborhoods of both countries have less than 100 venues with in 500 meter radius since we had set limit 100 in our code to get maximum 100 venues per Neighborhood.

Ideally value should be below:

Total neighborhoods in New York=30600

Total neighborhoods in Toronto=10300

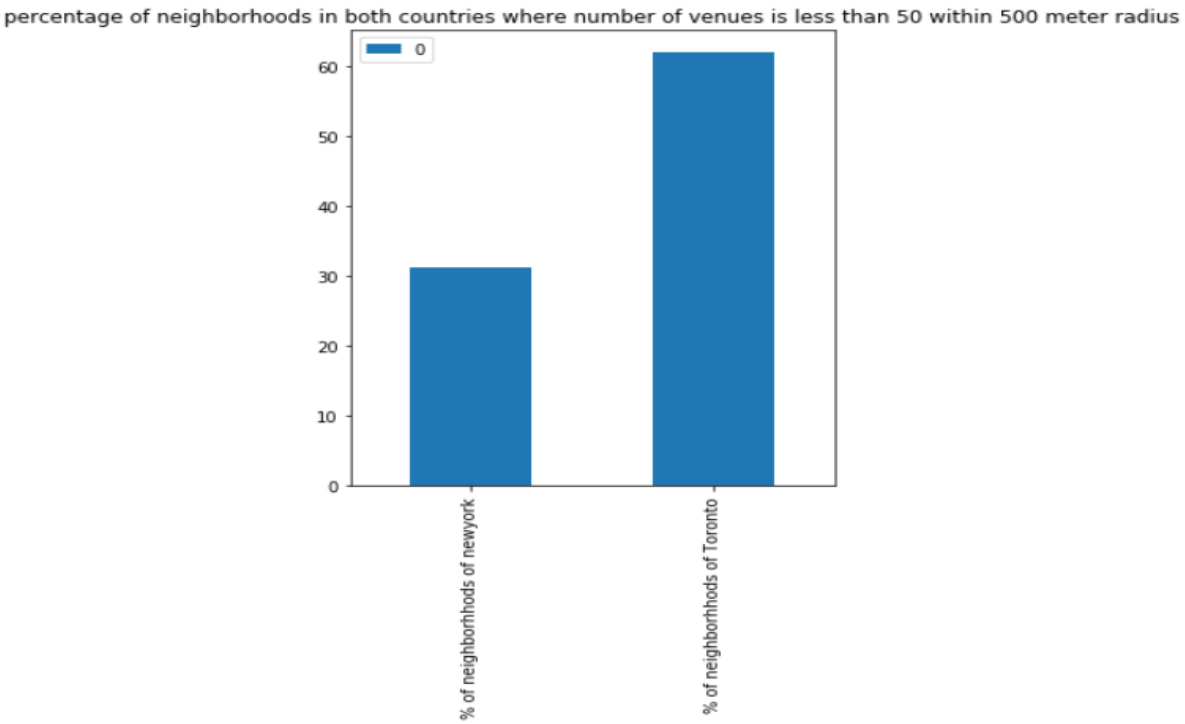
After doing analysis for small neighborhoods of both countries where venues frequency is less than 50 below is result

Result 1:

63% Neighborhoods of Toronto has less than 50 venues within 500-meter radius while percentage in New York is 30%

From this result we can conclude below:

If we want to divide Neighborhoods of Toronto with respect to density of venues, then 63% neighborhoods of Toronto will come under small category but in New York only 30 % would come in small category.



Analysis2- For Large Neighborhoods of both countries:

In this analysis we will focus only top 20 neighborhoods of both countries where venue density is 100

We will check which country's neighborhood wins the race regarding variety of venues.

Here I am referring variety with number of unique categories in any neighborhood

we can say that this result will explain which country has more diversity in nature of population

I am assuming here if any neighborhood consists more type of venues then that neighborhood is more dynamic and there is more chances to get profit in opening new type of venue.

After analysis we got below result.

Result2: From above graph it is clear that neighborhoods of new York wins the race regarding diversity of categories of venues.

Analysis3: Clustering of neighborhoods of both countries

In this analysis I checked how many clusters are required for clustering of both countries data, it will help us to understand that how neighborhoods of each countries are similar with in country.

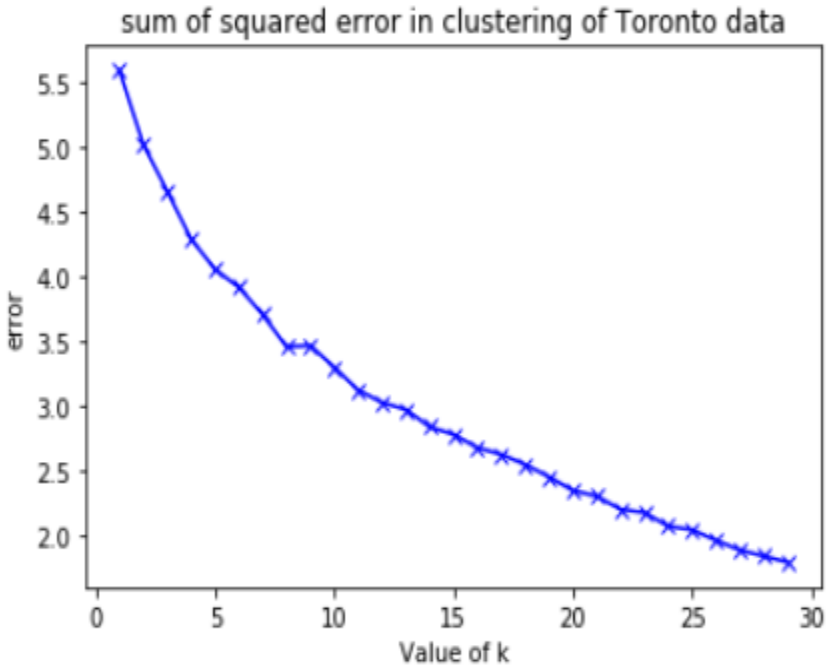
In this analysis I focused only categories of venues so after one hot encoding of venue categories I grouped data on neighborhoods and took mean of each venue category (Please refer to code for more explanation)

We applied K-means clustering algorithm on venue data of both countries.

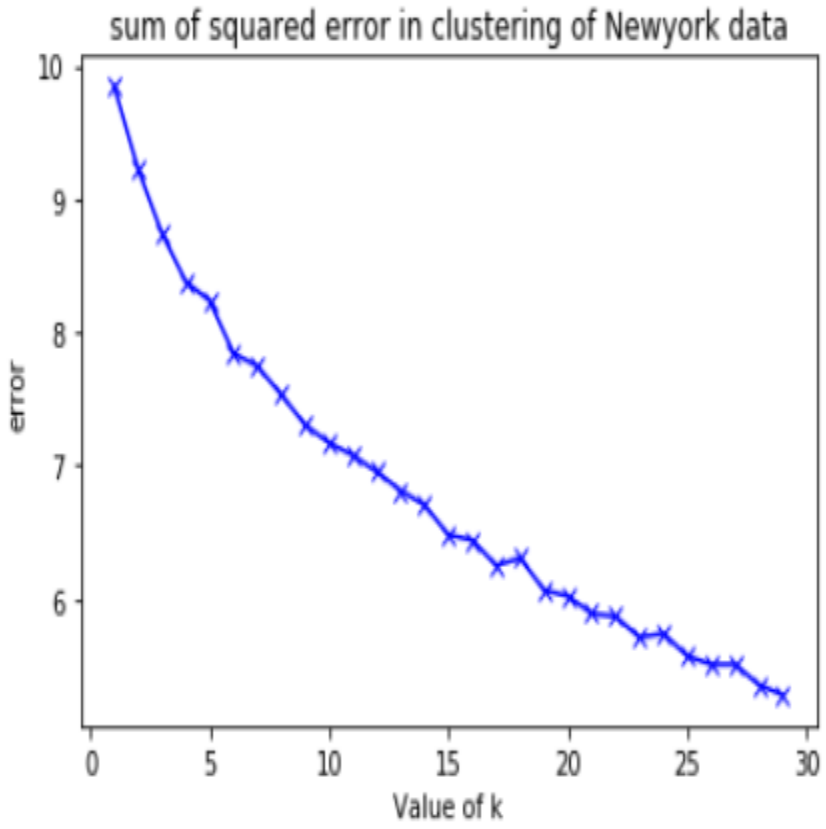
Finally, we plotted graph for sum of squared error for clustering of each countries.

Results are below.

Graph for Toronto:



Graph for New York:



Result 3: From above two graphs it is clear that if we want to divide neighborhoods of each countries in group then we need more groups for New York data.

Results:

1- If we want to divide Neighborhoods of Toronto with respect to density of venues, then 63% neighborhoods of Toronto will come under small category but in New York only 30 % would come in small category.

2- Neighborhoods of New York wins the race regarding diversity of categories of venues. We can conclude that Neighborhoods of New York has more unique type of venues than Toronto.

3- if we want to divide neighborhoods of each countries in groups then we need more groups for New York data. We can conclude that neighborhoods of New York have more difference with each other while Toronto Neighborhoods have less difference between each other.

Conclusion:

In this analysis I divided each country in below two parts.

1-Small neighborhoods

2-Large neighborhoods

Division has been done on density of venues and after that we analyzed it on frequency of unique categories of venue.

After analysis we found that Toronto is collection of small neighborhoods while New York is collection of big neighborhoods. There are more results which have been explained above.

From above analysis we can check how neighborhoods of each countries are different with each other.

By this result investor can change their strategy accordingly.

This is a limited analysis we can do a lot of interesting analysis further which I would like to do in next project.