**Business Location identification between New york city and Toronto - Case Study – Gopi Krishna Podicheti**

**A. Introduction**

**A.1. Description & Discussion of the Background**

* **Investors** and **new businesses** always are on the lookout for opening their business in locations that are competitive and have a thriving market. In the current world scenario there is a big focus on **Multinational market locations** to maximize on revenues/spread and risk mitigation. Similarly it is also important that if there are **similar neighborhoods** that can be identified across locations they would form a sweet spot for the new business to be setup.
* This would be really helpful in **local marketing** and designing an optimal or similar sales campaign. So I wanted to study 2 major cities in North America (New York and Toronto) and look at neighborhoods across both these cities where a new business like Coffee shops could be launched.
* Few **key factors** that would influence this decision are

1. Area of the city

2. Commonly visited places

3. Footfalls to a particular place

4. What other places are common in the area

**A.2. Assumptions and Considerations:**

1. Cost of investment and rental have not been considered for this study
2. Once the location recommendation is done based on above criteria a separate feasibility study

would be undertaken for the same

**A.3. Data Description**

 Data Source and consideration:

* There is considerable amount of data about Toronto's neighborhood available on wikipedia which I could utilize by scrapping the data and cleaning the same before getting the Geospatial data for the same along with neighborhoods.
* I also utilized data that was available from IBM on New york city neigborhoods, I had to clean this data and get it into a consistent format to match with Toronto city data.
* I used Foursquare API to get the most common venues of given Borough of Toronto and Newyork along with venues that are visited along with footfall information.

**B. Methodology**

I utilized the above mentioned data sources and stored the combined information on a csv file on my GITHUB.

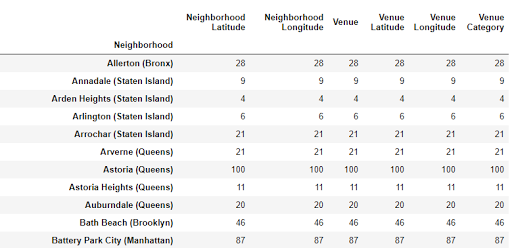
**Data cleaning:**

Data that was obtained from Wikipedia for Toronto had to be cleaned for **Borough** information like “**PO Boxes, Replymail, Postal gateway, etc**” as these were not actual boroughs. Post which we looked for data and wherever it did not have any borough information we dropped those records. We **assigned** the neighborhood information to the borough information if the **neighborhood** was **not assigned**. We tested our data with a sampling of Postal codes. We added the Latitude and Longitude information from the Geospatial coordinates sheet and further added City (‘Toronto’) as it would be helpful when we merge the data with newyork city data.

Similarly the data that was obtained from IBM data source for **New York** also needed to be re organized and there were **multiple neighborhoods** within a **borough,** so we combined the neighborhoods and the boroughs to form the new consolidated neighborhoods which was easily identifiable and in line with the data of Toronto city with respect to neighborhoods.

**Feature Selection:**

Post merging the neighborhood data for Newyork and Toronto there were **345** distinct neighborhoods and I used foursquare API's to get the Venues and Categories information for each of the neighborhoods limiting my search to **top 100** venues per neighborhood and within **500 m** range from the center of the neighborhood. This resulted in about **11737** venues with about **449** unique categories of venues.

[](https://www.blogger.com/u/1/blog/post/edit/3191831214092856324/1224895593660227733)

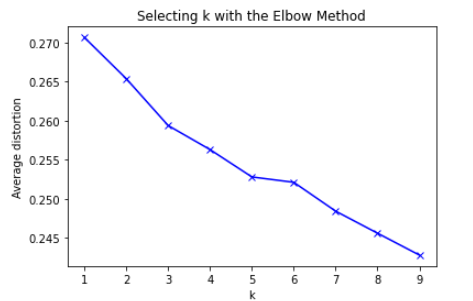
I picked the top **50** categories of Venues as going any further would not really be required for our study as it would not really make a change in our analysis.

I further got the data studied by the neighborhoods and venue categories and the mean values for the same. Following this I created a dataframe which gave me list of **top 10 most common venues for each of the neighborhood**.

[](https://www.blogger.com/u/1/blog/post/edit/3191831214092856324/1224895593660227733)

**Clustering and Segmentation:**

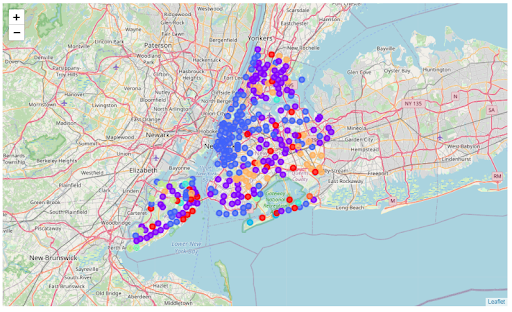
Post this I used an **unsupervised learning** algorithm to cluster the data(**K Means**). First I studied the optimal ‘k’ value - clustering segments using the **Elbow method**. Where I sampled from 1 through 10 as the cluster size and plotted the same against the average distortion.

[](https://www.blogger.com/u/1/blog/post/edit/3191831214092856324/1224895593660227733)

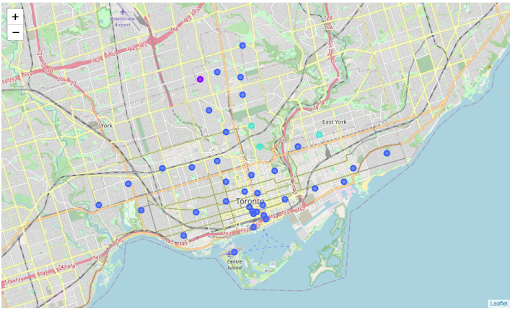
I found through this analysis that though the average distortion flattened at K value of 5 there was not significant variation between the value at 5 and 9 and hence I choose a **K value of 9**.

I then plotted the clusters using **Folium maps** to view them on both **New york city** and **Toronto**.

New York City:

[](https://www.blogger.com/u/1/blog/post/edit/3191831214092856324/1224895593660227733)

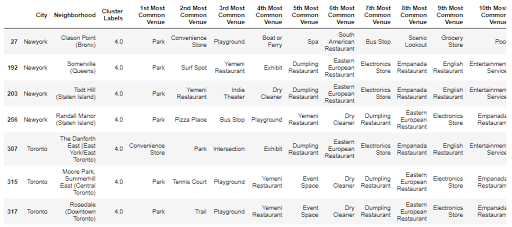
 Toronto City:

[](https://www.blogger.com/u/1/blog/post/edit/3191831214092856324/1224895593660227733)

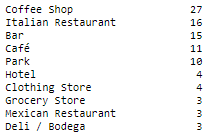
Further to this I did an analysis of the clusters with **2** criteria

1. Clusters which has a good sample size of neighborhoods from both New york and Toronto
2. Clusters where the "**1st Most common venue**" is either "**Coffee shop**" or "**Cafe**"

Interestingly **Clusters 3** and **4** were the only clusters which had neighborhoods from both Newyork and Toronto. However **Cluster 4** was primarily "**Parks**"

[](https://www.blogger.com/u/1/blog/post/edit/3191831214092856324/1224895593660227733)

So I removed Cluster 4 from my analysis. I further analyzed **Cluster 3** for "**1st Most Common Venue**" by neighborhood. I found that **Coffee Shops(27) and cafe(11)** put together was **38** followed by Italian restaurants at 16.

[](https://www.blogger.com/u/1/blog/post/edit/3191831214092856324/1224895593660227733)

In total there are **18 neighborhoods in Newyork and 20 neighborhoods in Toronto** which has either "Coffee shop" or "Cafe" as the most common venue

[](https://www.blogger.com/u/1/blog/post/edit/3191831214092856324/1224895593660227733)

[C:\Users\GUESTU~2\AppData\Local\Temp\msohtmlclip1\02\clip_image009.png](https://www.blogger.com/u/1/blog/post/edit/3191831214092856324/1224895593660227733)

**C. Discussion:**

I started my data exploration by getting the data from wiki and IBM. Further I cleaned the data and made it uniform so that I could merge the data for New York and Toronto. The number of neighborhoods was relatively high which was really helpful for my study as I was able to get rich and good amount of data with respect to Venues and categories for the neighborhoods.

Further I used K Means unsupervised learning algorithm to cluster the data. I used the Elbow method to determine the optimal value of k to determine the number of clusters and used a value of 9.

Once clustered I analyzed the data systemically to understand the clusters of interest i.e the clusters which had either “Coffee shops” or “Café” as their most common venues and also had neighborhoods in both Newyork and Toronto.

This helped to zero-in on Cluster 3 which had the exact same criteria and on further analyzing and filtering data I was able to come up with 38 different neighborhoods across Newyork and Toronto which are similar with respect to venues, footfalls and primarily “Coffee shop”/”Café” venues.

**D. Conclusion and Recommendation:**

Post segmenting and clustering for similar neighborhoods in NYC and Toronto which also have a high market for Coffee shops and Cafe footfalls.

We are sending out the following neighborhood recommendations to the investors post which they can further their cost feasibility analysis

[](https://www.blogger.com/u/1/blog/post/edit/3191831214092856324/1224895593660227733)

**E. References:**

* 1. Wikipedia - <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
  2. IBM data for New York city
  3. Foursquare API’s
  4. Geospatial data excel