The Battle of Neighborhoods : Toronto vs New York

IBM Applied Data-Science Capstone Project

Maruf Ansari

INTRODUCTION

The aim of this project is to explore the neighborhoods of two major economic capitals, Toronto (CA) and New York City (USA), and group them by common nearby venues. In this project i have compared the neighborhoods of both the cities with respect to places to eat, better connectivity to several useful regions and how they are distributed around both cities. The places I have considered are airports, metros, coffee-shops, restaurants, schools, colleges, general stores, hospitals etc.

The target audience will be anyone visiting, relocating or expanding their businesses between the financial capitals of the two countries to search for better neighborhoods suited for their needs and therefore might offer their preferred range of amenities.

DATA

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

```
[15] #Shape of our table
     print('Shape of New York dataframe is {} with {} unique boroughs and {} neighborhoods.'.format(
             newyork_df_new.shape, len(newyork_df_new['Borough'].unique()), newyork_df_new.shape[0] ))

     Shape of New York dataframe is (306, 4) with 5 unique boroughs and 306 neighborhoods.
```

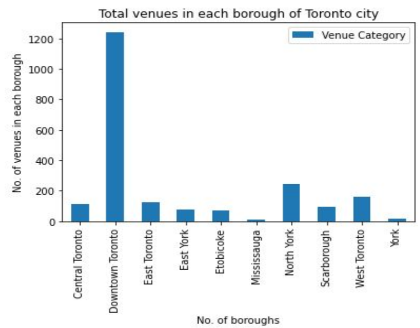
	Borough	Neighborhood	Latitude	Longitude
0	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	Scarborough	Woburn	43.770992	-79.216917
4	Scarborough	Cedarbrae	43.773136	-79.239476

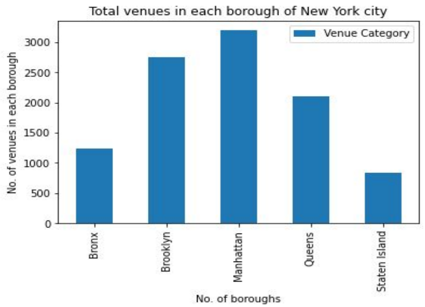
```
[8] # Shape of our table
    print('Shape of Toronto dataframe is {} with {} unique boroughs and {} neighborhoods.'.format(
            toronto_df_new.shape, len(toronto_df_new['Borough'].unique()), toronto_df_new.shape[0] ))

    Shape of Toronto dataframe is (103, 4) with 10 unique boroughs and 103 neighborhoods.
```

- The data set for Toronto is available at wikipedia. It includes postal-code, boroughs and neighborhood names. Along with it i have to add location coordinates of each neighborhood, which i found on a coursera given link.
- The data set for New York was given by coursera. It includes boroughs, neighborhood names and location of each neighborhood.
- I used the Foursquare API to analyze the nearby places to these neighborhoods and see the proximity of important places from the corresponding neighborhoods

METHODOLOGY





Analyze each neighborhood

- Firstly I obtained our data and cleaned it and prepared it for further processes and took initial visualization of the neighborhood dataset for both the cities with a folium map in order to understand the layout of the neighborhoods and their location in the respective cities.
- For analyzing the neighborhoods of Toronto and New York, I was able to generate a list of popular/trending places near each neighborhood based on the corresponding map coordinates in the data sets using the explore section of the Foursquare API.
- After it, I have generated tables showing the number of venues in each of the boroughs and then visualize them using a bar chart of matplotlib library to get a better idea about it.
- Finally i have generated tables for the top 10 most common venues around each of the neighborhoods of both the cities by using one hot encoding and then display them after grouping them on neighborhoods.

METHODOLOGY

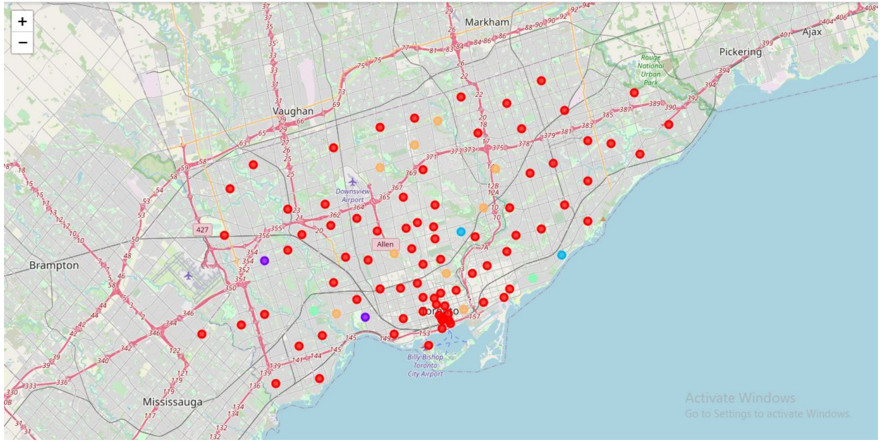
Cluster neighborhoods

- I imported the scikit-learn library in order to use the k-means algorithm to cluster venues and see similarities for different neighborhoods of each of the cities. Using this algorithm i have successfully labeled each of the neighborhood venues in suitable clusters and generated a well formatted data frame containing all the required information.
- Along with it, I have visualized the total no. of 1st most common venues in each of the clusters of both the cities using a bar graph of matplotlib library. After this I have also plotted the count of all unique venues in each cluster of 1st most common venues and visualized it using seabron library.
- Finally the visualization for both cities' neighborhood clusters is done through the folium library.

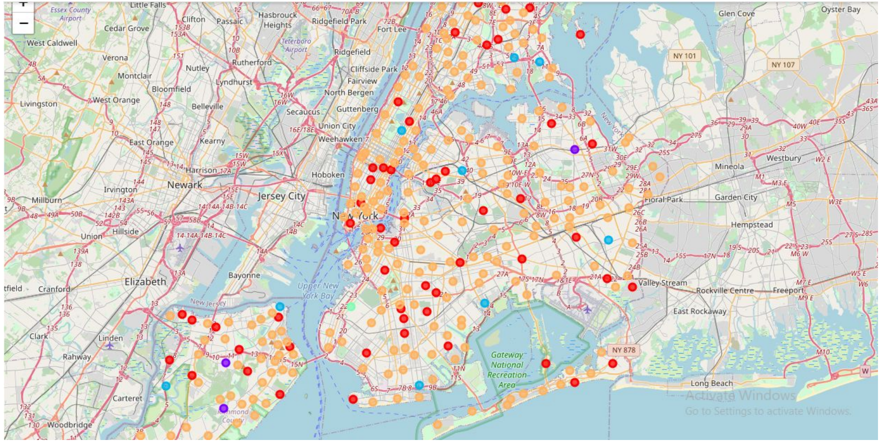
RESULT

The results obtained after applying the k-means clustering on both the dataset show that we can categorise the neighborhoods into 5 clusters based on the frequency of occurrence of venues.
Here is the map of both cities containing clusters of top 10 most common venues obtained after visualising the clustered dataset.

Toronto city cluster map :



New York city cluster map :



RESULT

Cluster Labels	Toronto Cluster 0	Toronto Cluster 1	Toronto Cluster 2	Toronto Cluster 3	Toronto Cluster 4	|	New York Cluster 0	New York Cluster 1	New York Cluster 2	New York Cluster 3	New York Cluster 4
1st Most Common Venue	Pizza Place	Baseball Field	Garden	Playground	Park	---- ----	Pizza Place	Park	Bus Station	Italian Restaurant	Tennis Court
2nd Most Common Venue	Sandwich Place	Yoga Studio	Home Service	Park	Convenience Store	---- ----	Deli / Bodega	South American Restaurant	Bus Stop	Yoga Studio	Playground
3rd Most Common Venue	Dessert Shop	Donut Shop	Dim Sum Restaurant	Yoga Studio	Yoga Studio	---- ----	Supermarket	Grocery Store	Grocery Store	Fish & Chips Shop	Yoga Studio
4th Most Common Venue	Coffee Shop	Diner	Diner	Doner Restaurant	Donut Shop	---- ----	Check Cashing Service	Bus Stop	Yoga Studio	Event Space	Ethiopian Restaurant
5th Most Common Venue	Café	Discount Store	Discount Store	Dim Sum Restaurant	Diner	---- ----	Mexican Restaurant	Boat or Ferry	Fish & Chips Shop	Exhibit	Event Space

One of my aims was also to show the top 5 venues information for each cluster of both the cities in a single table. This will help us to easily compare and analyze the clusters of both the cities.
The resultant clustered map and the above table will allow us to identify which neighborhoods have higher concentration of venues and places to visit while which have lower concentration.
Based on this, it will help us to answer the question as to which neighborhoods are most suitable based on our requirements.

DISCUSSION

The results show that there is a similarity for both cities' first few common places. The 1st cluster and 5th cluster of Toronto city and New York city respectively, shows that they have a very high concentration of venues around their neighborhoods and that's why they can be compared directly. But the results may also show that there are many clusters that do not have direct comparison.

Limitations and future research

Here it may be necessary to better clean the venue data returned by Foursquare API. As we can see, some of the top venues listed include 'Bus Stop' or 'Miscellaneous Shop' or 'Discount Store'. This may or may not be a significant venue. However based on the results it is clear we need more holistic data to improve the accuracy and usefulness of our neighborhood clustering. Along with this, it can be considered to include data on population, cost of living, demographics, schools and transportation. In addition, these data can also be accessed dynamically from specific platforms or packages.

CONCLUSION

- In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing K-means algorithm for clustering and lastly providing recommendations for stakeholders.
- This project has given us some insight into the amenities in the selected neighborhoods of both the cities most common places. As a result, people are turning to big cities as they can achieve better outcomes through their access to the platforms where such information is provided.

To the future,

Maruf Ansari

Thank you

Contact

Maruf Ansari

Github.com/MarufAnsari/
Linkedin.com/in/maruf5
marufansari0305@gmail.com