COMPARING AND SEGMENTING THE NEIGHBORHOODS OF NEW YORK CITY AND TORONTO TO SHOW HOW SIMILAR OR DISSIMILAR THEY ARE

By: Saad Zaheer

1. Introduction:

1.1. Description:

New York and Toronto are two of the most well known and crowded places in two different progressive countries, US and Canada, respectively. Both of these cities are the major financial centers of the countries they are situated in. In this study I am going to use an API called Foursquare to explore the neighborhoods in the two cities. I will be using functions such as explore that come with Foursquare API to obtain most common venue categories in each neighborhoods of both the cities. Then build on this feature to arrange the neighborhoods into clusters. For clustering, I will be using k-means. Finally, I will be using Folium library to visualize the neighborhoods of both cities and see how similar or dissimilar they are.

1.2. Audience

This problem targets people of New York City and Toronto. The analysis will help stakeholders identify similarities and dissimilarities between the people of both cities. Which can help stakeholders such a restaurant owners have a better clue if they really want to establish a branch in the other city or not. For example, if a person owns a restaurant in New York City and he wants to establish another branch in Toronto, this study will help him identify if such a venue as the one he is trying to establish will be a success or not.

2. Data Description:

A data science problem mainly depends on the availability of data. Luckily, for our problem, we have both the datasets of the neighborhoods of New York City and Toronto available online.

To be able to cluster and explore neighborhoods in those cities, we need datasets containing boroughs and neighborhoods of those boroughs. Other than that, we also need the coordinates (latitude, longitude) which we will be using with folium to visualize the neighborhoods properly on maps.

The neighborhood data for New York City is available on https://cocl.us/new_york_dataset in JSON format. We will be using our data analysis skills to analyze this JSON data and get only the neighborhoods data from it in the form of a Pandas dataframe.

The neighborhood data for Toronto is available on Wikipedia https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M which we will read into a pandas dataframe by scrapping it from the Wikipedia page. Then we will clean the data by removing rows with missing “Borough” values.

3. Analysis

Please see the analysis part in the notebook. It is easy to explain with code and figures.

4. Results and Discussion

Our analysis shows that New York is far denser than Toronto neighborhood-wise. We also saw that there are only a few categories of venues in Wakefield, New York, while in case of Regent Park / Harbourfront, Toronto, the categories are quite a few.Two venues, namely, dessert shops and ice cream shops were common to both Wakefield and Regent Park / Harbourfront. Only 8 venues were returned by Foursqaure API for wakefield, while 48 venues were returned by the API for Rogent Park / Harbourfront. The total number of venues for all of the neighborhoods in New York is 9754 while the number of venues for each neighborhood in Toronto is 1618. We also found out that there were 426 unique categories of venues in New York and 232 unique categories in case of Toronto. Then we made dataframes of the top most common venues for each neighborhood in both the cities. And finally, we clustered the neighborhoods in both cities using k-means clustering and showed the clusters on maps.

5. Conclusion

The purpose of this project was to explore neighborhoods and venues near those neighborhoods in two major cities of the world, namely, Toronto and New York. Second most important purpose of the project was to put to practice what we learned about Data Analysis and Foursquare API to real world problems and test our understanding of those skills.

Note:
Please refer to notebook for a thorough explanation of each section above.