Exploring the Similarity and Dissimilarity Between New York City (USA) and Toronto (Canada)

1. Introduction

Both Toronto and New York are the most popular city and financial capital of their own country. Most of the foreigner, including me, might to live or travel or know the similarity and disimilarity of these two diverse city. Therefore, in this project, we will segment and compare this two cities based on:

1. The most popular restaurant of the two cities in order to know whether they have the same preferable food or not
2. Their Average Income
3. Population Desity

If we can figure out the above problem, it will be helpful not only especially for those who would like to migrate and work at one of these cities but also for the tourists who would like to explore the most popular restaurant among the city. Moreover, it will also helpful for the person who is confusing about whether he/she should open their new restaurant in New York or Toronto because this project result can help them to understand what is the favorite/popular foods for the city dwellers.

2. Data Description

The required raw data for segmentation and Comparing are as follows:

1. Neighborhoods of New York data : https://geo.nyu.edu/catalog/nyu_2451_34572
2. Neighborhood of Toronto data: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
3. Demographics data of Toronto: https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

4. Demographics data of New York: https://en.wikipedia.org/wiki/Demographics_of_New_York_City
5. Geospatial Coordinates of Toronto: http://cocl.us/Geospatial_data

3. Methodology

3.1. Python libraries

Before doing anything, we need to import all the required libraries (such as Pandas, Matplotlib, Numpy, Seaborn, Sklearn, Folium ,Beautifulsoup, Wikipedia) for data collecting, data analysis, modeling and visualization. The required libraries are as follows:

- Pandas: Library for data analysis
- Numpy: Library for handle data in a vectorized manner
- Matplotlib: Library for plotting modules
- Seaborn: Library for plotting modules
- Scikit Learn: Library for building algorithm model
- Folium: Library for visualizing geospatial data
- Beautifulsoup: Library for scraping the HTML data from webpage
- Wikipedia: Library for scraping the data from webpage

3.2. Data Preparation

There are two parts for data preparation process. The first part is for boroughs and neighborhood list of New York and Toronto. The second part is for Demographics data of New York and Toronto.

3.2.1. Boroughs and Neighborhood List of New York and Toronto

Firstly, the boroughs and neighborhood list of New York was downloaded and converted it into pandas data frame by using pandas package in python. The boroughs and neighborhood list of Toronto was also imported from Wikipedia by using Beautifulsoup library and converted it into pandas data frame. Since, the imported list of Toronto did not include Location data, another data set comprised of location data of neighborhood and boroughs was imported and converted it from csv.format to pandas data frame. After the dataset was cleaned, the two tables were merged to get the final Toronto dataset.

	Postcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure: Example of Dataset

Secondly, based on the location data in the dataset, we can use Foursquare API for importing all venues into the both Toronto and New York dataset.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898276	-73.850381	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station

Figure: Dataset of New York including Venues info

3.2.2. Demographics data of New York and Toronto

Population density and average income data for all boroughs of New York and Toronto were imported from Wikipedia webpage by using Wikipedia library and converted them to a new data frame.

3.3. Data Analysis

3.3.1. Analyzing the dataset of Boroughs and Neighborhoods list of New York and Toronto

When all the required data is ready, we will start analyzing the data based on venue category for exploring the most popular restaurant in Toronto and New York separately. Firstly, the related venue category of each neighborhood was grouped by ‘Neighborhood’. After that, the restaurant from the venue category of neighborhood was filtered out. Then, according to the frequency of occurrence of the restaurant, the data frame was created and the top 10 restaurant for each neighborhood was displayed.

	Neighborhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th M Comn Resta
0	Adelaide,King,Richmond	American Restaurant	Thai Restaurant	Asian Restaurant	Sushi Restaurant	Restaurant	Vegetarian / Vegan Restaurant	Indian Restaurant	Colon Resta
1	Agincourt	Vietnamese Restaurant	Dumpling Restaurant	Hakka Restaurant	Greek Restaurant	Gluten-free Restaurant	German Restaurant	French Restaurant	Filipin Resta
2	Agincourt North,L'Amoreaux East,Milliken,Steel...	Vietnamese Restaurant	Dumpling Restaurant	Hakka Restaurant	Greek Restaurant	Gluten-free Restaurant	German Restaurant	French Restaurant	Filipin Resta
3	Albion Gardens,Beaumond Heights,Humbergate,Jam...	Fast Food Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Hakka Restaurant	Greek Restaurant	Gluten-free Restaurant	German Restaurant	Frenc Resta
4	Alderwood,Long Branch	Vietnamese Restaurant	Dumpling Restaurant	Hakka Restaurant	Greek Restaurant	Gluten-free Restaurant	German Restaurant	French Restaurant	Filipin Resta

Figure: Example of top 10 restaurants for each neighborhood

3.3.2. Analyzing the dataset of Demographics data of New York and Toronto

When the data is imported into pandas data frame, the dataset is needed to be cleaned and modified in order to just filter out the required dataset. After that the dataset is ready to visualize and compare for the average income and population density.

	Borough	Population	Average Income
0	The Bronx	1471160	19570
1	Brooklyn	2648771	23900
2	Manhattan	1664727	378250
3	Queens	2358582	31310
4	Staten Island	479458	23460

Figure: Example of New York Demographics datasets

3.4. Modeling

As we would like to classify the similarity and dissimilarity of New York and Toronto, the clustering algorithm will be used to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields. Although there is more than one clustering algorithm, K-means clustering algorithm will be used to generate the cluster of this two cities based on common restaurants because it is easily to understand and implement in code. Finally, these clustering will be visualized by using folium library.

```
#Cluster Neighborhoods of toronto
# set number of clusters
kclusters = 5

toronto_grouped_clustering = toronto_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([0, 0, 0, 2, 0, 0, 0, 0, 0, 0], dtype=int32)
```

```
#Cluster Neighborhoods
# set number of clusters
kclusters = 5

ny_grouped_clustering = ny_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(ny_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([1, 0, 0, 0, 3, 0, 0, 1, 0, 1], dtype=int32)
```

4. Result and Discussion Section

4.1. The most popular restaurant of the two cities

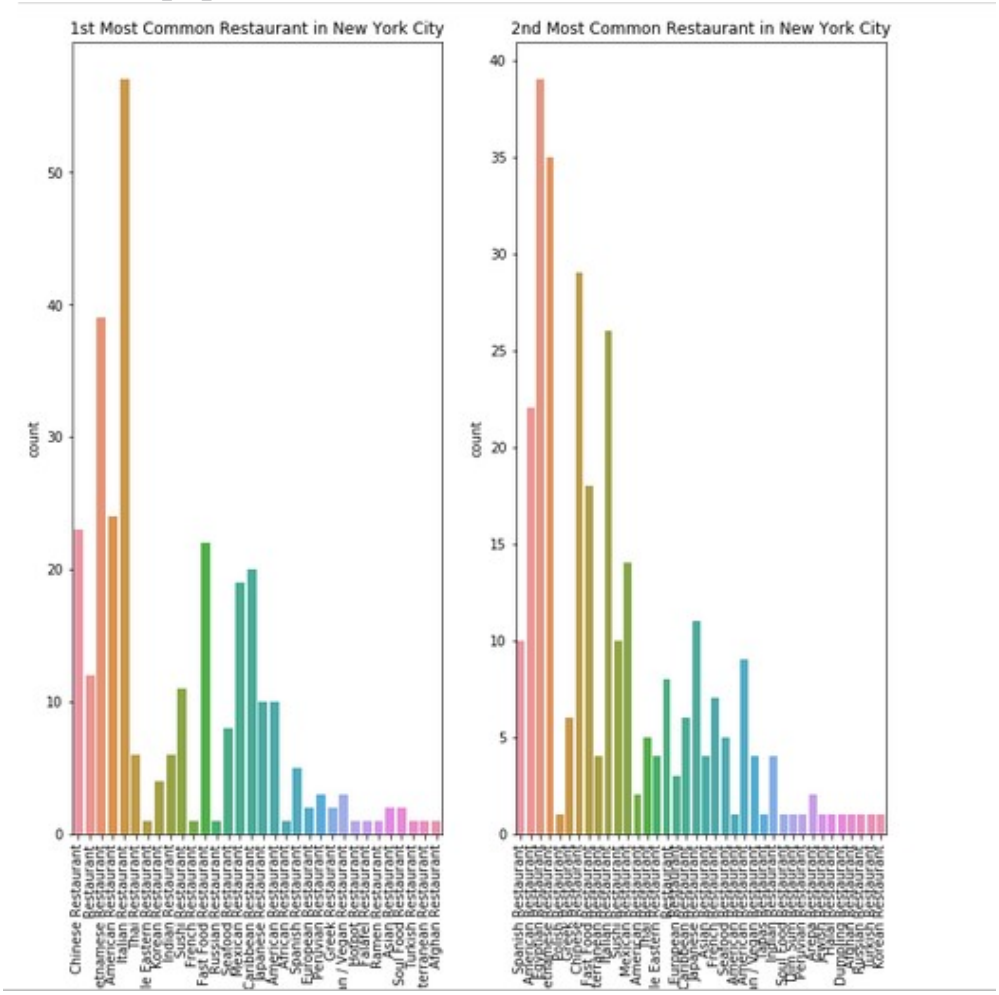


Figure: The first and second popular restaurant in New York

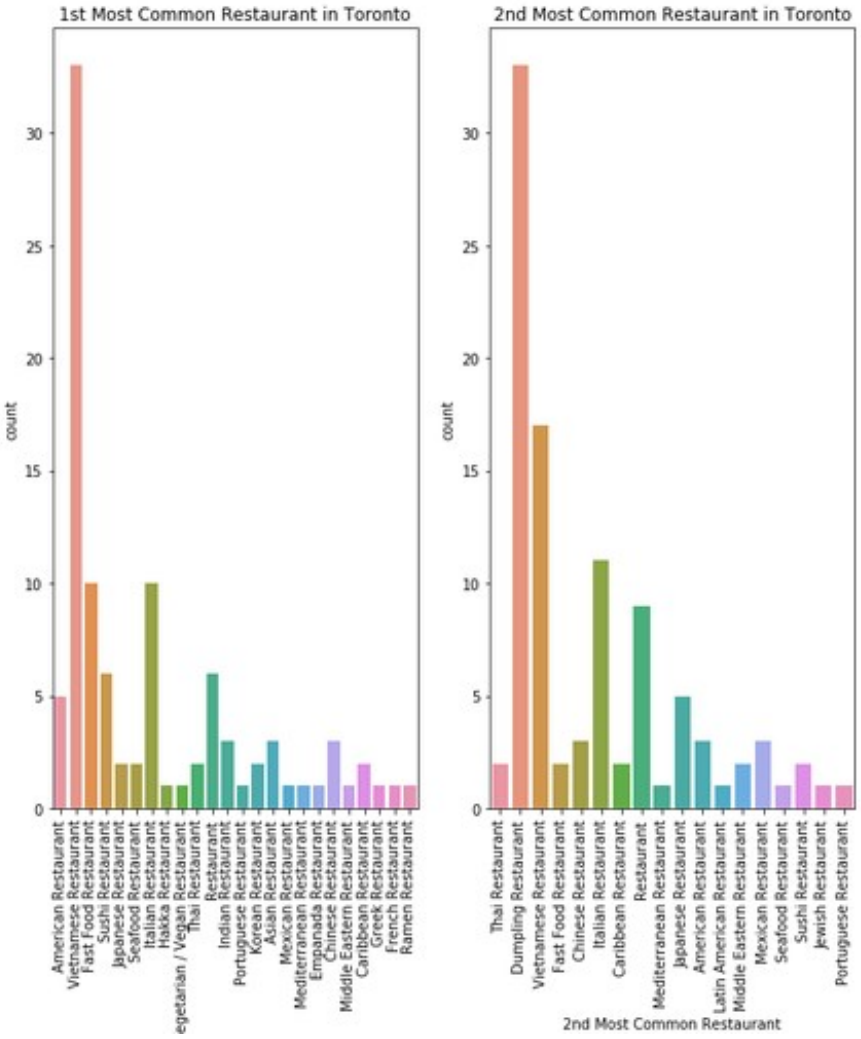


Figure: The first and second popular restaurant in Toronto

In New York City, the Italian Restaurant dominates in the 1st most common restaurant. On the other hand, Vietnamese restaurant is the most popular restaurant in Toronto. According to this result, it is clear that these two cities have different food taste.

4.2. Clustering of Neighborhoods based on Most Common Restaurant

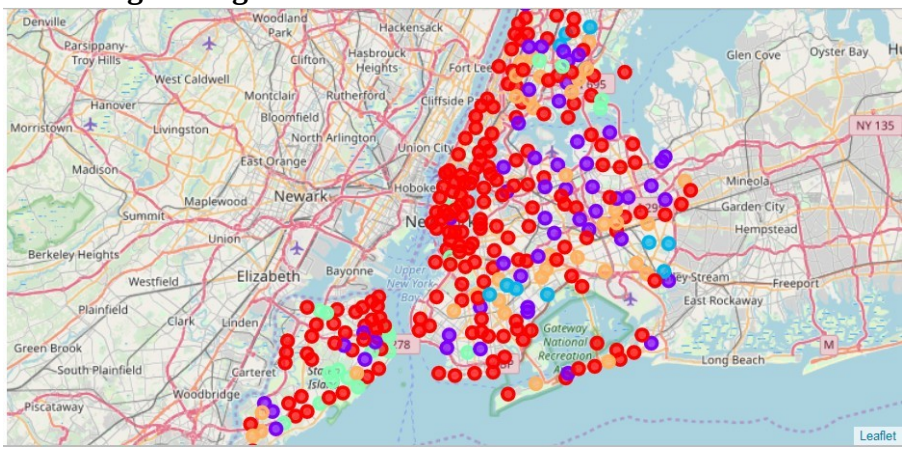


Figure: Cluster of New York

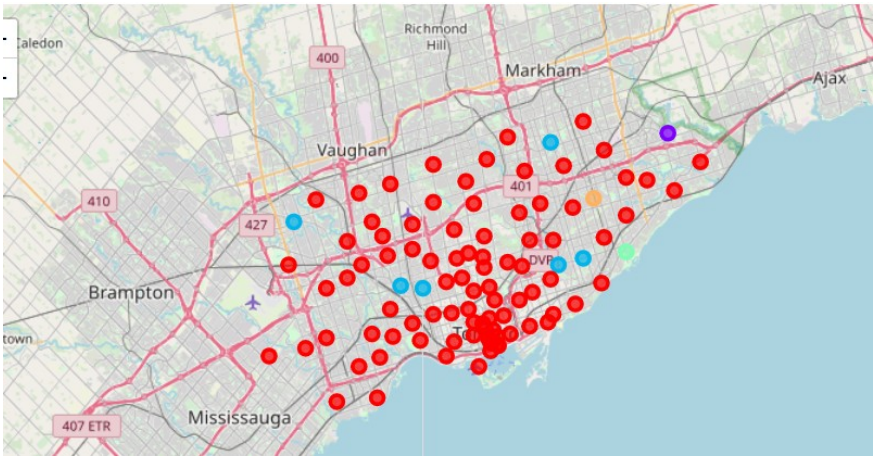
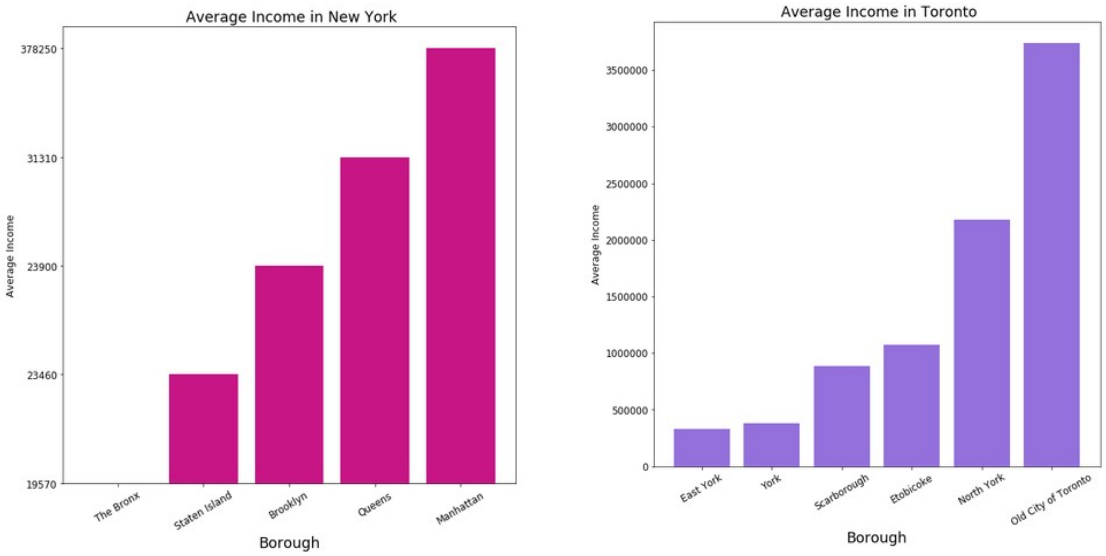


Figure: Cluster of Toronto

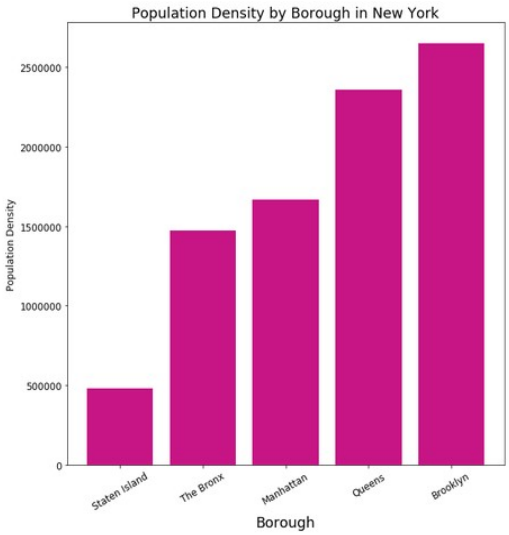
In the above cluster, it can be seen that New York has one big cluster (62% of the neighborhoods, one smallest cluster (3% of the neighborhoods) and three mid size clusters. For Toronto, there are one big cluster (91% of the neighborhoods) and four clusters (sharing the remaining 9% of the neighborhoods). So, it is clear that most of the Toronto Neighborhoods have the same prefer in foods New York has more variable in taste.

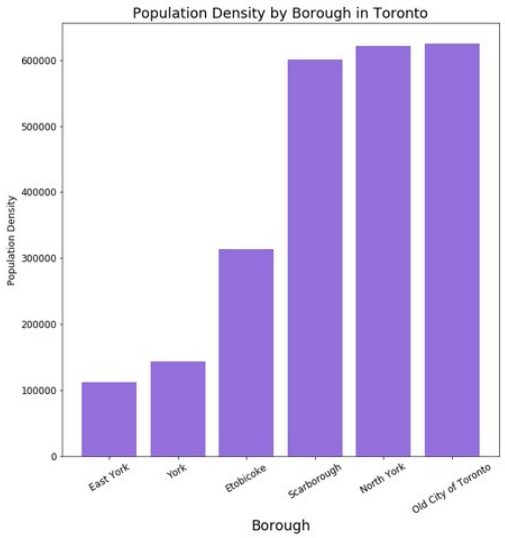
4.3. Average Income



In the given chart, Manhattan has by far the highest income when comparing to the other four boroughs in New York. At the same time, Old city of Toronto has highest average income but not very far from other four in Toronto. So, the Toronto has the uniform distribution of income than New York.

4.4. Population Density





In New York, Brooklyn has the most population whereas in Toronto Scarborough, North York and Old City of Toronto have almost the same highest population. However, New York is more crowded than Toronto when comparing with the overall population.

5. Conclusion

In this capstone project, I analyzed the relationship between New York and Toronto based on restaurants, income and population.