Capstone Project Report - The Battle of Neighborhoods

Samira Amirli

May 7, 2020

1. Introduction

1.1. Background

Various cities in every part of the world contain various kinds of venues and their categories in itself which inform about the specific local cultures of those cities to the mankind. Let me introduce the most important cities which attract the people's attention from all over the world, such as New York City which is the main capital, international and most populated city of the United States, and Toronto which is the main financial capital, and also most populated city of Canada.

1.2. Business Problem

Let me introduce my tourism agent that is led by me and my friend which is located in Munich, Germany. We are working here as the consultants regarding the travelling issues to various points of the world, trying to offer the best opportunities to our customers and making them become satisfied with the results. As it is obvious that most people in Munich are working at very prestigious work places and their salaries are very high pointing to the high level of living standards of those people. For that reason, after getting retired, older people get very high retirement salaries as a result of their previous high work performance.

Here in my research, the main target audiences are older people above age of 50-60 years. Every month various number of those older people visit our travel

agency and indicate that they want to travel to the USA or Canada. They want from us to provide them with the best opportunities, give them advices about the details of the travel to abovementioned countries and want from us to find the best locations in those countries according to their preferences and tastes. Therefore, in order to provide our customers with necessary information regarding travelling tips and also provide with the best opportunities, starting from hotels to entertainment places, we conduct a small survey and prepare special questions for them, which help us to learn about their preferences. Thus, the following survey questions for the customers are prepared by us:

1. How old are you and what is the purpose of your travelling?

2. Are you travelling with your family/friends or alone?

3. Are you interested in entertainment, such as going to pubs, clubs, cinema or others?

4. Are you interested in doing various sport types and do you want to go for some winter sports?

5. Would you like mainly to visit the old places, such as not investigated remote places and historical museums?

In addition, for my research, I will choose the main boroughs from abovementioned two countries, such that I will choose Brooklyn from New York, and York from Toronto, which are one of most important and famous places of their corresponding cities. After that, according to customers' answers to the survey questions, I will recommend them which one of the places to visit.

1.3. Objective

The purpose of this project is to categorically segment the neighborhoods of New York City and Toronto (Brooklyn, York respectively) into major clusters and

examine those clusters to find the appropriate travel places considering the preferences and tastes of our customers who are in the category of above 50- 60 years.

2. Data Description

2.1. Data Sources

In this project, I will be working with two data sets. The first dataset of New York City consists of 5 boroughs and neighborhoods in each borough, and also geometric coordinates, such as latitude and longitude coordinates of each neighborhood. The link to this dataset can be found easily on the web and is the following: https://geo.nyu.edu/catalog/nyu_2451_34572

The second dataset of Toronto city consists of different boroughs, neighborhood in each borough and their respective postal codes. The link to this dataset is taken from Wikipedia page and is the following: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Also the data for the geographical coordinates for each neighborhood in Toronto are collected, thus, the following csv file contains those data: https://cocl.us/Geospatial_data

The Foursquare API will be utilized to obtain the geographical location data, such as for Brooklyn in New York, and for York in Toronto. These datas will be used to explore the venues in the neighbourhoods of Brooklyn and York,respectively.The venues will provide the categories needed for my dataset analysis : https://developer.foursquare.com/

3. Methodology

The first dataset of New York City is in .json format which contains 5 boroughs and its neighborhoods. After .json file is downloaded, we need to analyze it in order to get detailed information about the structure of the data. All necessary data is found to be in 'features' key, which lists the neighborhoods. Then, the python dictionary should be transformed into pandas dataframe by looping through the data and filling in the dataframe one row at a time. Finally, the dataframe is created consisting of four columns: Borough, Neighborhood, Latitude and Longitude variables.

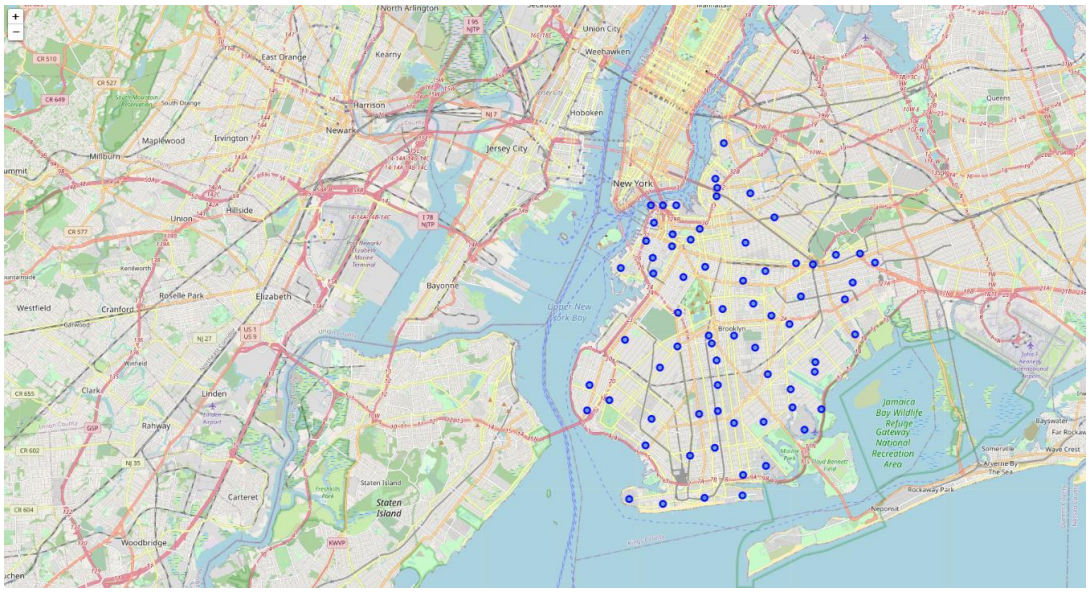
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

If we analyze, the dataset consists of 5 boroughs and 306 neigborhoods.

As part of my analysis, I will segment and cluster only the neighborhood in Brooklyn. Thus, the original dataframe is sliced and new dataframe of Brooklyn data is created.

	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

For the purpose of obtaining the latitude and longitude values of Brooklyn, ‘geopy’ library is utilized, which return Latitude as 40.6501038, and Longitude as -73.9495823. Following this step, a map of Brooklyn is created with neighborhoods superimposed on top. To illustrate the map, python ‘folium’ library is imported.



Next, all the same abovementioned procedures are applied to the second dataset of Toronto city. This dataset is a Wikipedia page which contains all information about the variables in a wikitable. Therefore, I load and use BeautifulSoup package to transform the data in a wikitable into the pandas dataframe. In the next step, there is a need for the cleaning of the data because a few several values in the column “Borough” were not assigned. Thus, the values with “Not assigned” in “Borough” column will be dropped. Next, also “Neighborhood” column contains some “Not assigned values”, so, they are assigned the values from the respective row values in “Borough” column. The coordinate data is fetched for all neighborhoods in Toronto city using the provided csv file and merging it with the pandas dataframe. The final dataframe of Toronto city is the following:

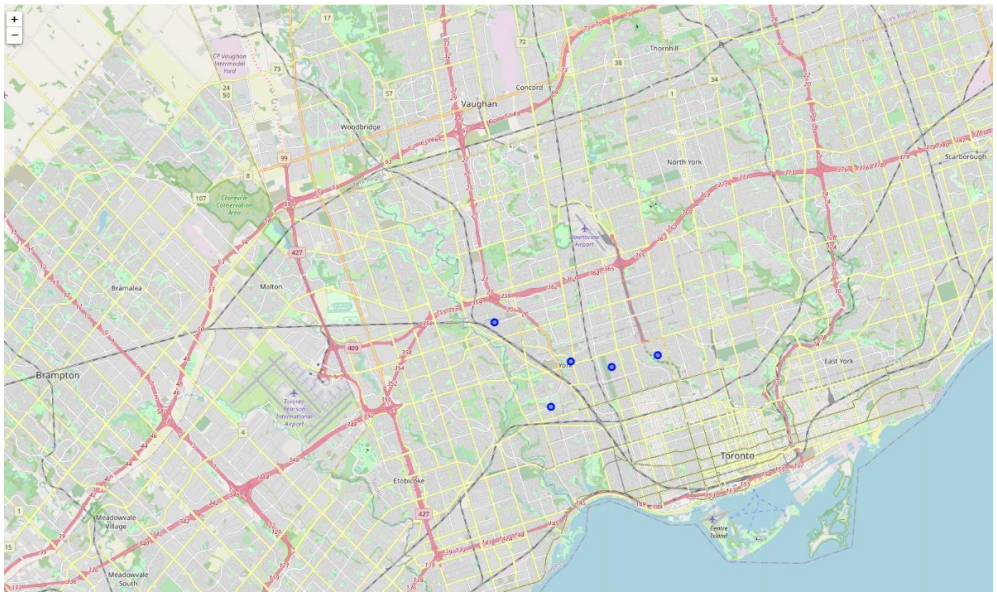
As part of my analysis, I will segment and cluster only the neighborhood in York. Thus, the original dataframe is sliced and new dataframe of York data is created.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern , Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill , Port Union , Highland Creek	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

As part of my analysis, I will segment and cluster only the neighborhood in York. Thus, the original dataframe is sliced and new dataframe of York data is created.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M6C	York	Humewood-Cedarvale	43.693781	-79.428191
1	M6E	York	Caledonia-Fairbanks	43.689026	-79.453512
2	M6M	York	Del Ray , Mount Dennis , Keelsdale and Silvert...	43.691116	-79.476013
3	M6N	York	Runnymede , The Junction North	43.673185	-79.487262
4	M9N	York	Weston	43.706876	-79.518188

For the purpose of obtaining the latitude and longitude values of York, ‘geo py’ library is utilized, which return Latitude as 43.67910515, and Longitude as -79.4 9118414007154. Following this step, a map of Brooklyn is created with neighborhoods superimposed on top. To illustrate the map, python ‘folium’ library is imported.



3.1. Exploratory Data Analysis

Moreover, I will utilize the Foursquare API to explore the neighborhoods and segment them, and it is necessary to define the Foursquare credentials and version. Thereafter, 'getNearbyVenues' function is created, which functions loop through all the neighborhoods of Brooklyn and sets an API request URL by defining radius = 500 and LIMIT = 50. This function returns maximum 50 nearby venues. Then, following the same steps I define a new function which returns a list of all venues for all neighborhoods in Brooklyn. The result indicates that there is a total of 2091 venues in all neighborhoods in Brooklyn.

```
In [18]: print(brooklyn_venues.shape)
         brooklyn_venues.head()
```

```
(2091, 7)
```

Out[18]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
1	Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
2	Bay Ridge	40.625801	-74.030621	Leo's Casa Calamari	40.624200	-74.030931	Pizza Place
3	Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
4	Bay Ridge	40.625801	-74.030621	The Bookmark Shoppe	40.624577	-74.030562	Bookstore

Applying the same procedures, a list of all venues for all neighborhoods in York are obtained. As a result, it is observed that there is a total of 17 venues in all neighborhoods, which is a very small number compared with Brooklyn.

In [46]:
```
print(york_venues.shape)
york_venues.head()
```

```
(17, 7)
```

Out[46]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Humewood-Cedarvale	43.693781	-79.428191	Cedarvale Park	43.692535	-79.428705	Field
1	Humewood-Cedarvale	43.693781	-79.428191	Cedarvale Ravine	43.690188	-79.426106	Trail
2	Humewood-Cedarvale	43.693781	-79.428191	Phil White Arena	43.691303	-79.431761	Hockey Arena
3	Humewood-Cedarvale	43.693781	-79.428191	Prince's Parkette	43.697385	-79.424704	Park
4	Caledonia-Fairbanks	43.689026	-79.453512	Nairn Park	43.690654	-79.456300	Park

3.2. Feature Engineering

Aiming to analyze each neighborhood in a deeper level, 'onehot encoding' function will be used. This function is applied for the purpose of converting the categorical variables, which is the "Venue Category" in my dataframe for both Brooklyn and York respectively, into dummy variables in order to let the machine learning algorithms exhibit the best performance in the prediction.

```
print(brooklyn_onehot.shape)
brooklyn_onehot.head()
```

```
(2091, 263)
```

Out[47]:

	Yoga Studio	Accessories Store	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Baseball Field
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

```
print(york_onehot.shape)
york_onehot.head()
```

```
(17, 14)
```

Out[50]:

	Neighborhood	Bus Line	Caribbean Restaurant	Convenience Store	Fast Food Restaurant	Field	Grocery Store	Hockey Arena	Park	Pool	Restaurant	Sandwich Place	Trail	Women's Store
0	Humewood-Cedarvale	0	0	0	0	1	0	0	0	0	0	0	0	0
1	Humewood-Cedarvale	0	0	0	0	0	0	0	0	0	0	0	1	0
2	Humewood-Cedarvale	0	0	0	0	0	0	1	0	0	0	0	0	0
3	Humewood-Cedarvale	0	0	0	0	0	0	0	1	0	0	0	0	0
4	Caledonia-Fairbanks	0	0	0	0	0	0	0	1	0	0	0	0	0

As the LIMIT is defined as 50 in the previous section, the Foursquare API will return many venues. However, to keep the analysis more understandable and easier, it is possible to create a new dataframe which illustrates each neighborhood along with the top 10 most common venues in both Brooklyn and York respectively.

```
for ind in np.arange(brooklyn_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(brooklyn_grouped.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

Out[52]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bath Beach	Chinese Restaurant	Pharmacy	Gas Station	Bubble Tea Shop	Italian Restaurant	Fast Food Restaurant	Sushi Restaurant	Check Cashing Service	Kebab Restaurant	Sports Bar
1	Bay Ridge	Spa	Italian Restaurant	Pizza Place	Chinese Restaurant	Greek Restaurant	Grocery Store	Ice Cream Shop	Hookah Bar	American Restaurant	Lounge
2	Bedford Stuyvesant	Pizza Place	Coffee Shop	Café	Bar	Fried Chicken Joint	Bagel Shop	Cocktail Bar	Gift Shop	Gourmet Shop	Boutique
3	Bensonhurst	Sushi Restaurant	Bakery	Ice Cream Shop	Chinese Restaurant	Italian Restaurant	Donut Shop	Factory	Road	Noodle House	Liquor Store
4	Bergen Beach	Harbor / Marina	Baseball Field	Athletics & Sports	Playground	Donut Shop	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field

```
for ind in np.arange(york_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(york_grouped.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

Out[56]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Caledonia-Fairbanks	Women's Store	Pool	Park	Trail	Sandwich Place	Restaurant	Hockey Arena	Grocery Store	Field	Fast Food Restaurant
1	Del Ray, Mount Dennis, Keelsdale and Silverthorn	Sandwich Place	Restaurant	Fast Food Restaurant	Convenience Store	Women's Store	Trail	Pool	Park	Hockey Arena	Grocery Store
2	Humewood-Cedarvale	Trail	Park	Hockey Arena	Field	Women's Store	Sandwich Place	Restaurant	Pool	Grocery Store	Fast Food Restaurant
3	Runnymede, The Junction North	Grocery Store	Convenience Store	Caribbean Restaurant	Bus Line	Women's Store	Trail	Sandwich Place	Restaurant	Pool	Park
4	Weston	Park	Convenience Store	Women's Store	Trail	Sandwich Place	Restaurant	Pool	Hockey Arena	Grocery Store	Field

3.3. Machine Learning Algorithm

K-means is an unsupervised machine learning algorithm which groups certain data points into certain user-specified number of k clusters based on their similar characteristics or features. I will apply this algorithm to assign each neighborhood into certain cluster specifying each cluster label and variable cluster size. There are several ways to apply this algorithm and identify the number of clusters. To keep it simple, I am going to apply random initialization, such that I set the number of

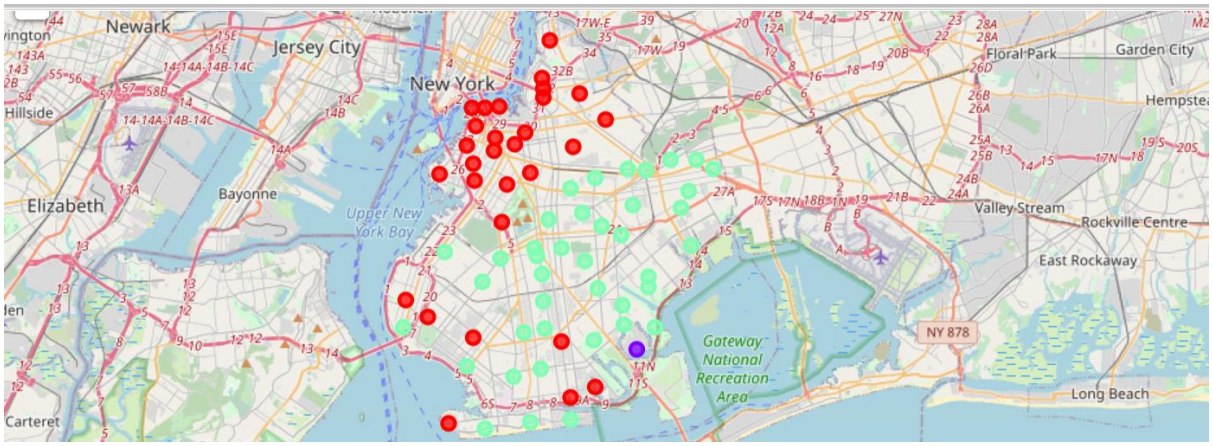
clusters to 3 and run the k-means to cluster each neighborhood into 3 clusters, in both Brooklyn and York respectively. My result will segment each neighborhood in both Brooklyn and York based upon the most common venues in its near surrounding.

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Brooklyn	Bay Ridge	40.625801	-74.030621	0	Spa	Italian Restaurant	Pizza Place	Chinese Restaurant	Greek Restaurant	Grocery Store	Ice Cream Shop	Hookah Bar	American Restaurant	Lounge
1	Brooklyn	Bensonhurst	40.611009	-73.995180	0	Sushi Restaurant	Bakery	Ice Cream Shop	Chinese Restaurant	Italian Restaurant	Donut Shop	Factory	Road	Noodle House	Liquor Store
2	Brooklyn	Sunset Park	40.645103	-74.010316	2	Bank	Latin American Restaurant	Pizza Place	Bakery	Mexican Restaurant	Gym	Fried Chicken Joint	Pharmacy	Deli / Bodega	Mobile Phone Shop
3	Brooklyn	Greenpoint	40.730201	-73.954241	0	Coffee Shop	Bar	Cocktail Bar	Yoga Studio	Spa	Café	Bakery	Record Shop	Pizza Place	French Restaurant
4	Brooklyn	Gravesend	40.595260	-73.973471	2	Pizza Place	Bakery	Bus Station	Lounge	Diner	Chinese Restaurant	Breakfast Spot	Furniture / Home Store	Bar	Electronics Store

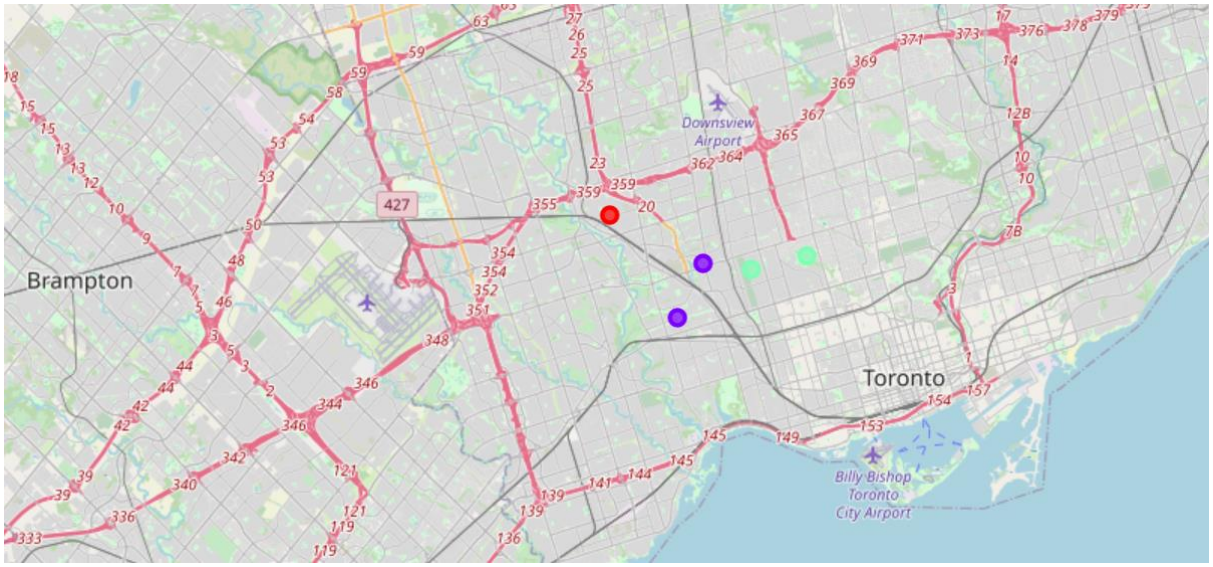
Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
M6C	York	Humewood-Cedarvale	43.693781	-79.428191	1.0	Trail	Tennis Court	Hockey Arena	Field	Women's Store	Turkish Restaurant	Skating Rink	Sandwich Place	Restaurant	Pool
M6E	York	Caledonia-Fairbanks	43.689026	-79.453512	0.0	Park	Women's Store	Pool	Turkish Restaurant	Trail	Tennis Court	Skating Rink	Sandwich Place	Restaurant	Pizza Place
M6M	York	Del Ray , Mount Dennis , Keelsdale and Silvert...	43.691116	-79.476013	1.0	Turkish Restaurant	Skating Rink	Sandwich Place	Restaurant	Bar	Women's Store	Trail	Tennis Court	Pool	Pizza Place
M6N	York	Runnymede , The Junction North	43.673185	-79.487262	2.0	Pizza Place	Caribbean Restaurant	Bus Line	Breakfast Spot	Women's Store	Turkish Restaurant	Trail	Tennis Court	Skating Rink	Sandwich Place
M9N	York	Weston	43.706876	-79.518188	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Finally, the neighborhoods in Brooklyn and York are visualized based on different clusters based on their characteristics using the pthon's 'folium' library.

Map of Brooklyn with grouped clusters:



Map of York with grouped clusters:



Compared to Brooklyn, here we observe very few neighborhoods or even a single neighborhood in specific clusters in terms of venue information. The reason is maybe that there are very few, only total of 17 venues and less diverse venue categories in York in comparison with Brooklyn.

4.Results

In addition, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, you can also assign a name to each cluster consisting of various neighborhoods.

Cluster 1 in Brooklyn:

	Neighborhood	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bay Ridge	-74.030621	0	Spa	Italian Restaurant	Pizza Place	Chinese Restaurant	Greek Restaurant	Grocery Store	Ice Cream Shop	Hookah Bar	American Restaurant	Lounge
1	Bensonhurst	-73.995180	0	Sushi Restaurant	Bakery	Ice Cream Shop	Chinese Restaurant	Italian Restaurant	Donut Shop	Factory	Road	Noodle House	Liquor Store
3	Greenpoint	-73.954241	0	Coffee Shop	Bar	Cocktail Bar	Yoga Studio	Spa	Café	Bakery	Record Shop	Pizza Place	French Restaurant
6	Sheepshead Bay	-73.943186	0	Dessert Shop	Turkish Restaurant	Sandwich Place	Harbor / Marina	Yoga Studio	Grocery Store	Hotel	Creperie	Restaurant	Outlet Store
12	Windsor Terrace	-73.980073	0	Deli / Bodega	Café	Diner	Park	Plaza	Bakery	Coffee Shop	Sushi Restaurant	Chinese Restaurant	Beer Store
13	Prospect Heights	-73.964859	0	Bar	Mexican Restaurant	Cocktail Bar	Beer Bar	Coffee Shop	Wine Bar	Café	Thai Restaurant	Bakery	Ice Cream Shop
15	Williamsburg	-73.958115	0	Coffee Shop	Bar	Bagel Shop	Yoga Studio	Taco Place	Breakfast Spot	Burger Joint	Café	Clothing Store	Playground

Cluster 2 in Brooklyn:

	Neighborhood	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
30	Mill Island	-73.908186	1	Pool	Locksmith	Women's Store	Filipino Restaurant	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	Fish & Chips Shop
100	Mill Island	-73.908186	1	Pool	Locksmith	Women's Store	Filipino Restaurant	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	Fish & Chips Shop

Cluster 3 in Brooklyn:

	Neighborhood	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Sunset Park	-74.010316	2	Bank	Latin American Restaurant	Pizza Place	Bakery	Mexican Restaurant	Gym	Fried Chicken Joint	Pharmacy	Deli / Bodega	Mobile Phone Shop
4	Gravesend	-73.973471	2	Pizza Place	Bakery	Bus Station	Lounge	Diner	Chinese Restaurant	Breakfast Spot	Furniture / Home Store	Bar	Electronics Store
5	Brighton Beach	-73.965094	2	Restaurant	Russian Restaurant	Beach	Eastern European Restaurant	Gourmet Shop	Sushi Restaurant	Bank	Mobile Phone Shop	Taco Place	Korean Restaurant
7	Manhattan Terrace	-73.957438	2	Pizza Place	Donut Shop	Ice Cream Shop	Steakhouse	Coffee Shop	Grocery Store	Convenience Store	Restaurant	Bank	Cosmetics Shop
8	Flatbush	-73.958401	2	Mexican Restaurant	Coffee Shop	Juice Bar	Bank	Caribbean Restaurant	Pizza Place	Lounge	Chinese Restaurant	Bagel Shop	Sandwich Place
9	Crown Heights	-73.943291	2	Pizza Place	Café	Museum	Bakery	Pharmacy	Coffee Shop	Salon / Barbershop	Candy Store	Burger Joint	Playground

If we analyze each cluster thoroughly, it becomes apparent that the first two clusters (Cluster 1 and 3) contain lots of various venue categories relating mostly to entertainment venues, such as mostly restaurants from various cultures, also cafes,

bars, pubs and so on. However, in Cluster 2, we observe only one neighborhood and entirely no entertainment venues.

Now let's analyze the clusters in York.

Cluster 1:

	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	York	43.706876	-79.518188	0	Park	Convenience Store	Women's Store	Trail	Sandwich Place	Restaurant	Pool	Hockey Arena	Grocery Store	Field

Cluster 2:

	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	York	43.691116	-79.476013	1	Sandwich Place	Restaurant	Fast Food Restaurant	Convenience Store	Women's Store	Trail	Pool	Park	Hockey Arena	Grocery Store
3	York	43.673185	-79.487262	1	Grocery Store	Convenience Store	Caribbean Restaurant	Bus Line	Women's Store	Trail	Sandwich Place	Restaurant	Pool	Park

Cluster 3:

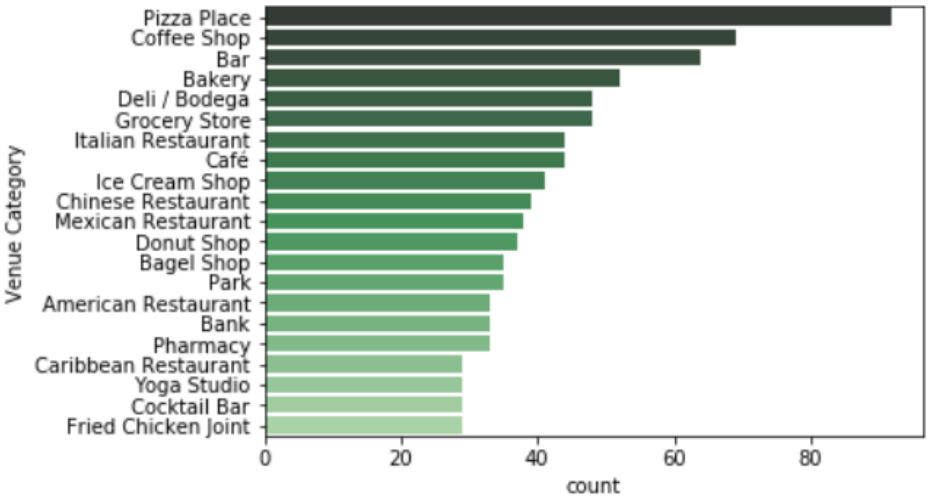
	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	York	43.693781	-79.428191	2	Trail	Park	Hockey Arena	Field	Women's Store	Sandwich Place	Restaurant	Pool	Grocery Store	Fast Food Restaurant
1	York	43.689026	-79.453512	2	Women's Store	Pool	Park	Trail	Sandwich Place	Restaurant	Hockey Arena	Grocery Store	Field	Fast Food Restaurant

If we analyze the all three clusters chosen in York area, it becomes clear that it is not possible to find enough venues for entertainment in the neighborhoods in these clusters compared to venues in Brooklyn. These clusters are mainly better for various kinds of sport activities purposes or visiting parks.

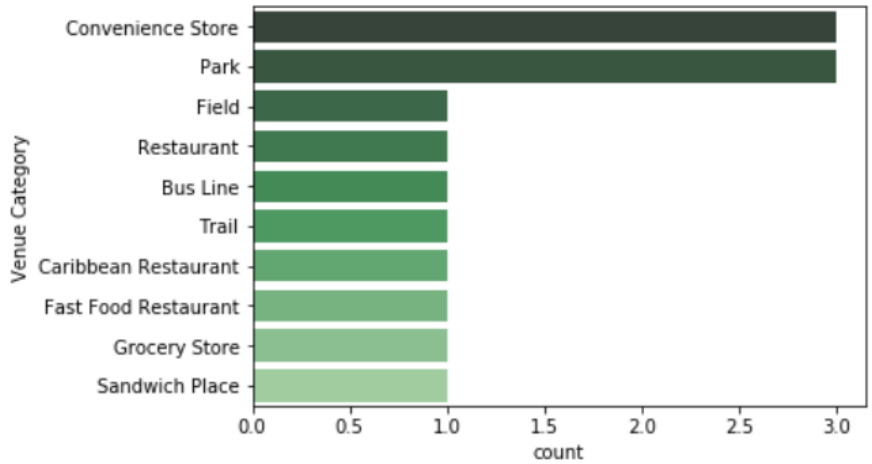
4.2. Data Visualization

In order to analyze the frequency of the top most common venues based on their venue categories, I am going to create a bar plot, which will help me to differentiate the venue categories and their densities between Brooklyn and York.

Bar plot for Brooklyn:



Bar plot for York:



5. Discussion

From the analysis of each cluster consisting of various neighborhoods and also by examining the bar graphs depicted based on most common venues and their categories in both Brooklyn and York, we can definitely state that Brooklyn is much more diverse than York in terms of entertainment places, such as restaurants, cafes, bars, museums and so on. On the other side, York is mainly diverse for various sport activities, also diverse for its parks, calm rest places and small stores.

If we consider that the main target audience in this research were the ones with ages above 50-60 years, and based on their answers to my survey questions, let's take into account that they are not so much interested in heavy sport activities or just going to vegetable fields or parks, instead, they prefer mostly calm entertainment places, such as tasting the foods from various cultures in various restaurants, visiting the historical places as museums, or going to cinema or pub places. For that reason, as a consultant in a tourism agency, I would recommend them travelling to Brooklyn instead of York, taking into account their answers to previously mentioned survey questions.

However, these recommendations and results can change if I add more specific questions to my survey. For example, we can ask them whether they have special health problems or not, and so on. Moreover, the results can change if we add other kinds of venue categories, such pharmacy places, hospitals, and other categories.

Furthermore, while segmenting the neighborhoods into various cluster numbers, it could be better to search for the optimal number of clusters instead of random initialization. For instance, as a further research, various machine learning algorithm methods as Elbow Method, Silhouette Method, or even Density Based Clustering Method can be applied for the purpose of finding the optimal k for better prediction, which can modify the results in this research.

6. Conclusion

All in all, I analyzed the venues and their categories in Brooklyn and Toronto by segmenting the neighborhoods in those places into different clusters. Moreover, data visualization tools, such as bar graphing helped me a lot in finding the better prediction. For the overall analysis, I applied the machine learning algorithm K-means to a multi-dimensional dataset for 2 various regions and tried to find the main differences between them in order to provide my target customers with the best opportunities based on their preferences, where the survey questions came for my assistance.

To recapitulate, these datasets can be applied to many various problems for the further research along with new datasets and API platforms.

References

1. New York City data: https://geo.nyu.edu/catalog/nyu_2451_34572

2. Wikipedia page for Toronto data: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

3. Geographic location data for Toronto : https://cocl.us/Geospatial_data