IBM Data Science Professional Certificate

Capstone Project

Similarities and differences between neighbourhoods of New York and Toronto

Author: Tomasz Tomaszewski

May, 2021

PART 1

Introduction

This work is a capstone project submission for "IBM Data Science Professional Certificate" online course on Coursera. Additionally it can showcase my knowledge and skills in data science/data analysis of real-world datasets and different scenarios.

In this report I will try to compare the cities of Toronto (Canada) and New York (USA). I will find some public datasets, explore them, anylize them and visualise them. All this work will be later presented in the form of final report presentation document.

Comparison of neighbourhoods of those cities will help to get insights on what kind of venues and points of interests are common for both cities and what kinds are very different between the cities.

New York, with an estimated 2019 population of over 8 millions is the most populous city in the United States and also the most densely populated major city in the United States. New York City serves as the cultural and financial capital of USA and possibly the world.

The second city of interest is Toronto, the capital of Canadian province of Ontario. The city it is the most populous city in Canada and the fourth most populous city in North America, with population of over 2.5 millions. Like New York, Toronto is also an international centre of business and culture, and is widely renowned as one of the most multicultural and cosmopolitan cities in the world.

Business problem

The report should give readers better understanding of similarities and differences between the two cities. This in turn will help to find suitable location either if one is interested in setting up a business in the city, or considers moving to the city or wants to visit city as a tourist.

Both final deliverable and the notebook can facilitate making decisions for many various stakeholders, including for example:
1) Students considering taking studies in any of the city in question.
2) Somebody who got an job offer in either NY or Toronto and would like to get to know the city before making decision to move.

3) Business management - exploration of the city could help in making decision in which district would be best to open a business (like bank, coffee shop, restaurant etc).
4) Citizens of various countries could find the area of the city that suits them best before they decide to move.
5) Turists will find it easier to make decision what to visit in the city or which city to choose for a visit.

PART 2

Description of data.

During earlier labs and courses modules of IBM Data Science Professional Certifficate there were many datasets presented and explored. Specifically for preparation of this report following sources of data will be used.

Datasets listing names of the neighborhoods of New Yorkand Toronto and their latitude and longitude coordinates.

For New York the source was provided by course teachers and the data set is extracted in the form of json file from following url:

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

	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
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

Data for Toronto were extracted via web scrapping technique with help of BeautifulSoup. Data source is the following link:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Thereafter Pandas dataframe has been created (see picture below):

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills North
8	M4B	East York	Parkview Hill, Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson

The source of data, hence also dataframe contains only PostalCode, Borough and Neighbourhood of Toronto. This problem will be solved by finding latitude and longitude data.

In order to make the data source complete, geospatial coordinates of Toronto were extracted from a csv file (file and location provided in IBM Data Science course):

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv

See the snipets of the code for geospatial coordinates extraction:

In [28]:
```
#Using read CSV method - getting Toronto coordinates
geo_coordinates="https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-
df_toronto_coord = pd.read_csv(geo_coordinates)
df_toronto_coord.head()
```

Out[28]:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Finally, both Toronto data frames were merged together which resulted in final Toronto data frame similar to the data for New York (see below).

:
```
toronto_merged_all = pd.merge(df, df_toronto_coord2, on="PostalCode")
toronto_merged_all.head(10)
```

:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills North	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937

Venues data were downloaded with help of Foursquare API, which is popular source of venue data and location data and utilisation of this tool was introduced during the course.

Different numbers of venues were found in different neighborhoods for respective city. Data were also saved in the form of pandas data frame.

Foursqure date retrieved for New York:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
1	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop
4	Marble Hill	40.876551	-73.91066	Astral Fitness & Wellness Center	40.876705	-73.906372	Gym

Foursqure date retrieved for Toronto:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant
4	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa

All retrieved data will be later wrangled, processed and analysed in later parts of this project and report in coming sections.

PART 3

Working with data.

In order to get some insights about neighborhoods for Toronto and Manhattan, there had been some work with retrieved data performed. All datasets were processed, cleaned and analyzed. It resulted in gathering some information about the most common venues in each neighborhood, which in turn helped in group ing neighbourhoods and venues into clusters.

As already mentioned in Part 2 Foursquare API was used to get the nearby venues for both cities. Thereafter one-hot encoding and k-means clustering algorithm was utilised to analyze the neighbourhoods.

Eploratory data analysis:

Collected data has been later analysed, plotted and visualised. Several techniques and libraries were used. Please refer to Jupyter notebook which entire Python code, which as a part of this capstone project.

Below comes example for Manhattan with snippets of the code regarding data analysis of venues and neighbourhoods:

Manhattan data:

```
In [50]: mht_bar1=manhattan_venues["Venue Category"].value_counts()
         mht_bar1.head(10) #data for barcharts visualisation performed later in the notebook

Out[50]: Coffee Shop            144
         Italian Restaurant     133
         Café                    77
         Bakery                  77
         Pizza Place             77
         American Restaurant     77
         Park                    73
         Hotel                   69
         Bar                     66
         Mexican Restaurant      58
         Name: Venue Category, dtype: int64
```

In [135]:
```
mht_bar2=mht_g1.drop('Neighborhood', axis=1).sum().sort_values(ascending=False)
mht_bar2.head(10)
```

Out[135]:
```
Coffee Shop            39
Italian Restaurant     33
Pizza Place            33
Bakery                 32
Café                   32
American Restaurant    30
Bar                    28
Park                   28
Gym                    27
Mexican Restaurant     27
dtype: int64
```

In [62]:
```
trt_bar1=toronto_venues["Venue Category"].value_counts()
trt_bar1.head(10) #data for barcharts visualisation performed later in the notebook
```

Out[62]:
```
Coffee Shop            150
Café                    84
Restaurant              50
Italian Restaurant      41
Hotel                   39
Park                    37
Japanese Restaurant     32
Bakery                  31
Pizza Place             27
Gym                     24
Name: Venue Category, dtype: int64
```

Toronto data:

In [79]:
```
trt_g1["Coffee Shop"].sum() # in how many neighborhoods coffee shops exists
```

Out[79]:
```
26
```

In [80]:
```
trt_bar2=trt_g1.drop('Neighborhood', axis=1).sum().sort_values(ascending=False)
```

In [81]:
```
trt_bar2.head(10)
```

Out[81]:
```
Park                   27
Coffee Shop            26
Café                   26
Restaurant             25
Italian Restaurant     22
Bakery                 19
Pizza Place            18
Bar                    15
Japanese Restaurant    15
Pub                    14
dtype: int64
```

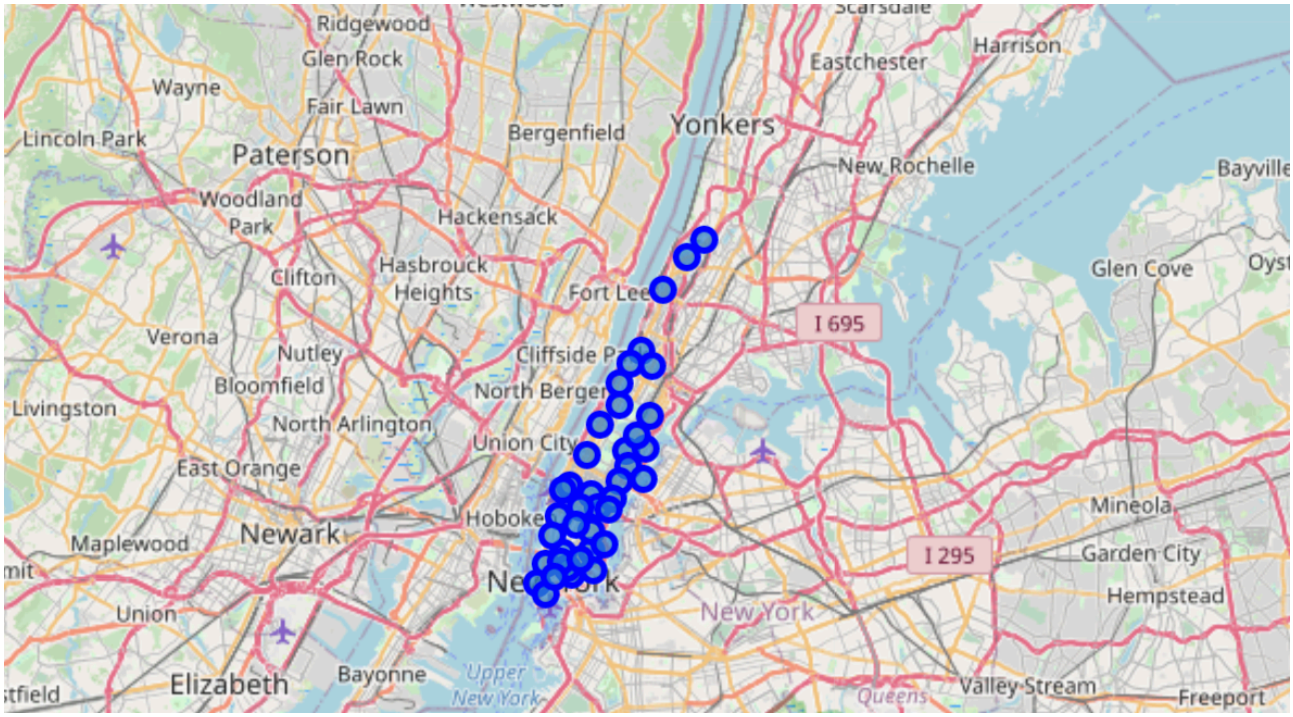
PART 4

Results.

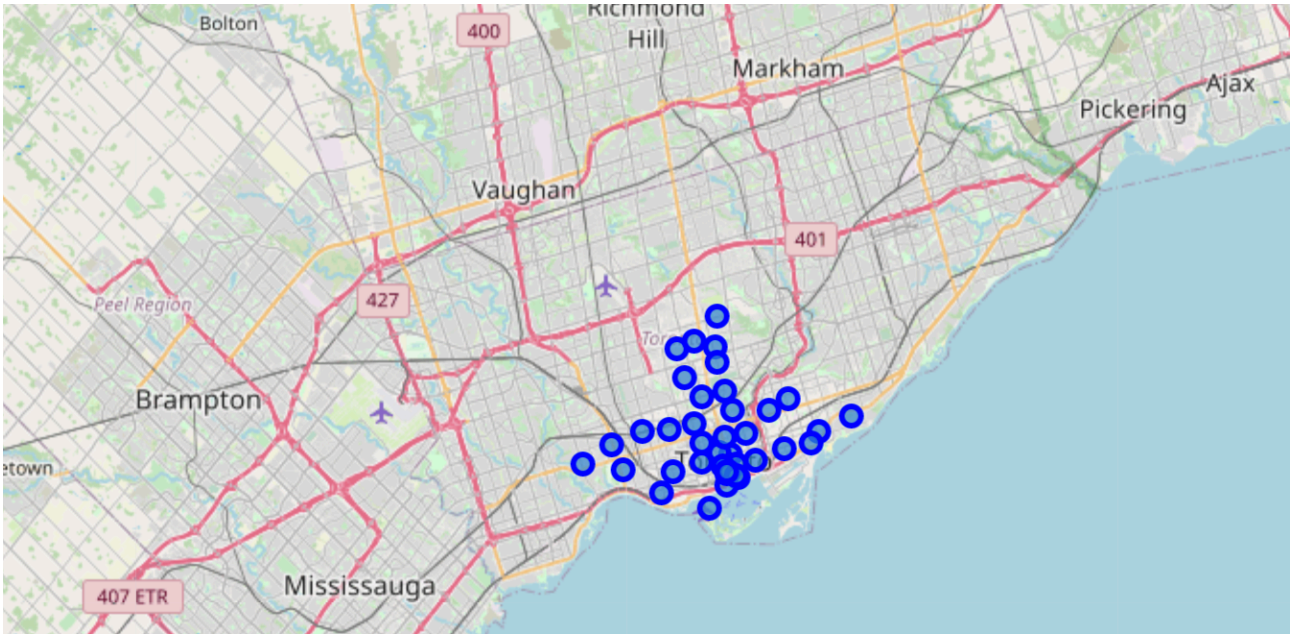
The next step was presentation of the results based on findings mentioned above. One of the best method helping stakeholders making their choice could probably be visualising data on maps and plotting some graphs do get better overview and insights about both cities and neighborhoods.

Folium Map tool is of a great help to get a quick glimpse on Toronto and New York, and with help of that tool the following maps of neighbourhoods were visualised.

Neighbourhoods of Manhattan:



Neighbourhoods of Toronto:



Clustering of neighbourhoods.

In order to find similarities and differences data were distributed into five clusters with help of K-means technique, discussed during the IBM Data Science Professional certificate course. Before that happened data had to be transformed into binary labels with one hot encoding. This is because the algorithm works on digital data and numbers only and our date contains strings of information as labels for each neighborhood. “One hot encoding” basically transforms string labels and creates new columns per each label and using 1 or 0. This enables to determine if row of table has that feature or not.

```
In [70]: mht_g1 = manhattan_onehot.groupby('Neighborhood').max().reset_index()
         mht_g1.head(5)
```

Out[70]:

	Neighborhood	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auditorium	Australian Restaurant
0	Battery Park City	0	0	0	1	0	0	0	0	0	0	1	1	0
1	Carnegie Hill	0	0	0	1	0	1	0	1	0	0	0	0	0
2	Central Harlem	0	0	1	1	0	0	1	0	0	0	0	0	0
3	Chelsea	0	0	0	1	0	0	1	0	0	1	0	0	0
4	Chinatown	0	0	0	1	0	0	0	0	0	1	0	0	0

Thereafter the data was organised in five clusters for each city by K-means algorithm method.

Example for Manhattan:

Building clusters of neighbourhoods

In [89]:
```
#The choice was made to select 5 clusters for the purpose of this task
```

In [90]:
```
#Clusters NY
# set number of clusters
kclusters = 5

manhattan_grouped_clustering = manhattan_grouped.drop('Neighborhood', 1)

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

# check cluster labels generated for each row in the dataframe
kmeans_manhattan.labels_[0:100]
```

Out[90]:
```
array([1, 0, 0, 0, 0, 1, 1, 3, 0, 1, 0, 0, 0, 3, 1, 3, 0, 0, 0, 3, 3, 0,
       4, 1, 1, 3, 1, 0, 3, 0, 2, 0, 0, 3, 0, 1, 0, 3, 0, 0], dtype=int32)
```

Out[143]:

	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
0	Manhattan	Marble Hill	40.876551	-73.910660	1	Gym	Sandwich Place	Coffee Shop	Yoga Studio	Deli / Bodega	Supplement Shop	Steakhouse	Seafood Restaurant
1	Manhattan	Chinatown	40.715618	-73.994279	0	Chinese Restaurant	Bakery	Cocktail Bar	American Restaurant	Salon / Barbershop	Spa	Optical Shop	Dessert Shop
2	Manhattan	Washington Heights	40.851903	-73.936900	0	Café	Bakery	Grocery Store	Mobile Phone Shop	New American Restaurant	Supplement Shop	Latin American Restaurant	Gym
3	Manhattan	Inwood	40.867684	-73.921210	0	Mexican Restaurant	Restaurant	Lounge	Café	Caribbean Restaurant	Bakery	Wine Bar	Pizza Place
4	Manhattan	Hamilton Heights	40.823604	-73.949688	0	Pizza Place	Café	Coffee Shop	Deli / Bodega	Park	Mexican Restaurant	Yoga Studio	Cocktail Bar
5	Manhattan	Manhattanville	40.816934	-73.957385	1	Coffee Shop	Italian Restaurant	Deli / Bodega	Mexican Restaurant	Bar	Supermarket	Burger Joint	Spanish Restaurant
6	Manhattan	Central Harlem	40.815976	-73.943211	1	Seafood Restaurant	Public Art	African Restaurant	American Restaurant	French Restaurant	Bar	Gym / Fitness Center	Chinese Restaurant

Example of cluster number 2 for both cities:

In [108]:
```
#Cluster 2
```

In [146]:
```
mht_clust_2=manhattan_merged2.loc[manhattan_merged2['Cluster Labels'] == 1, manhattan_merged2.columns[[1] + list(range(5, man
hattan_merged2.shape[1]))]]
mht_clust_2.head(8)
```

Out[146]:

	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	Marble Hill	Gym	Sandwich Place	Coffee Shop	Yoga Studio	Deli / Bodega	Supplement Shop	Steakhouse	Seafood Restaurant	Pizza Place	Department Store
5	Manhattanville	Coffee Shop	Italian Restaurant	Deli / Bodega	Mexican Restaurant	Bar	Supermarket	Burger Joint	Spanish Restaurant	Bus Station	Café
6	Central Harlem	Seafood Restaurant	Public Art	African Restaurant	American Restaurant	French Restaurant	Bar	Gym / Fitness Center	Chinese Restaurant	Music Venue	Spa
9	Yorkville	Italian Restaurant	Gym	Coffee Shop	Bar	Deli / Bodega	Wine Shop	Japanese Restaurant	Mexican Restaurant	Sushi Restaurant	Bagel Shop
14	Clinton	Italian Restaurant	Theater	Gym / Fitness Center	American Restaurant	Sandwich Place	Gym	Coffee Shop	Spa	Wine Shop	Hotel
23	Soho	Clothing Store	Italian Restaurant	Boutique	Mediterranean Restaurant	Bakery	Coffee Shop	Sporting Goods Shop	Women's Store	Art Gallery	Shoe Store
24	West Village	Italian Restaurant	New American Restaurant	Park	American Restaurant	Cocktail Bar	Ice Cream Shop	Wine Bar	Cosmetics Shop	Coffee Shop	Gay Bar
26	Morningside Heights	Coffee Shop	Bookstore	Park	American Restaurant	Burger Joint	Café	Deli / Bodega	Farmers Market	Supermarket	Salad Place

In [110]:
```
trt_clust_2=toronto_merged2.loc[toronto_merged2['Cluster Labels'] == 1, toronto_merged2.columns[[1] + list(range(5, toronto_m
erged2.shape[1]))]]
trt_clust_2
```

Out[110]:

	Borough	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
21	Central Toronto	1	Trail	Jewelry Store	Sushi Restaurant	Bus Line	Wine Shop	Diner	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant
26	Central Toronto	1	Dessert Shop	Sandwich Place	Gym	Sushi Restaurant	Italian Restaurant	Café	Pizza Place	Coffee Shop	Dance Studio	Brewery

Example of cluster number 5 for both cities:

In [117]:
```
#Cluster 5
```

In [118]:
```
mht_clust_5=manhattan_merged2.loc[manhattan_merged2['Cluster Labels'] == 4, manhattan_merged2.columns[[1] + list(range(5, man
hattan_merged2.shape[1]))]]
mht_clust_5.head()
```

Out[118]:

	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
22	Little Italy	Bakery	Italian Restaurant	Café	Ice Cream Shop	Hotel	Salon / Barbershop	Thai Restaurant	Chinese Restaurant	Mediterranean Restaurant	Sandwich Place

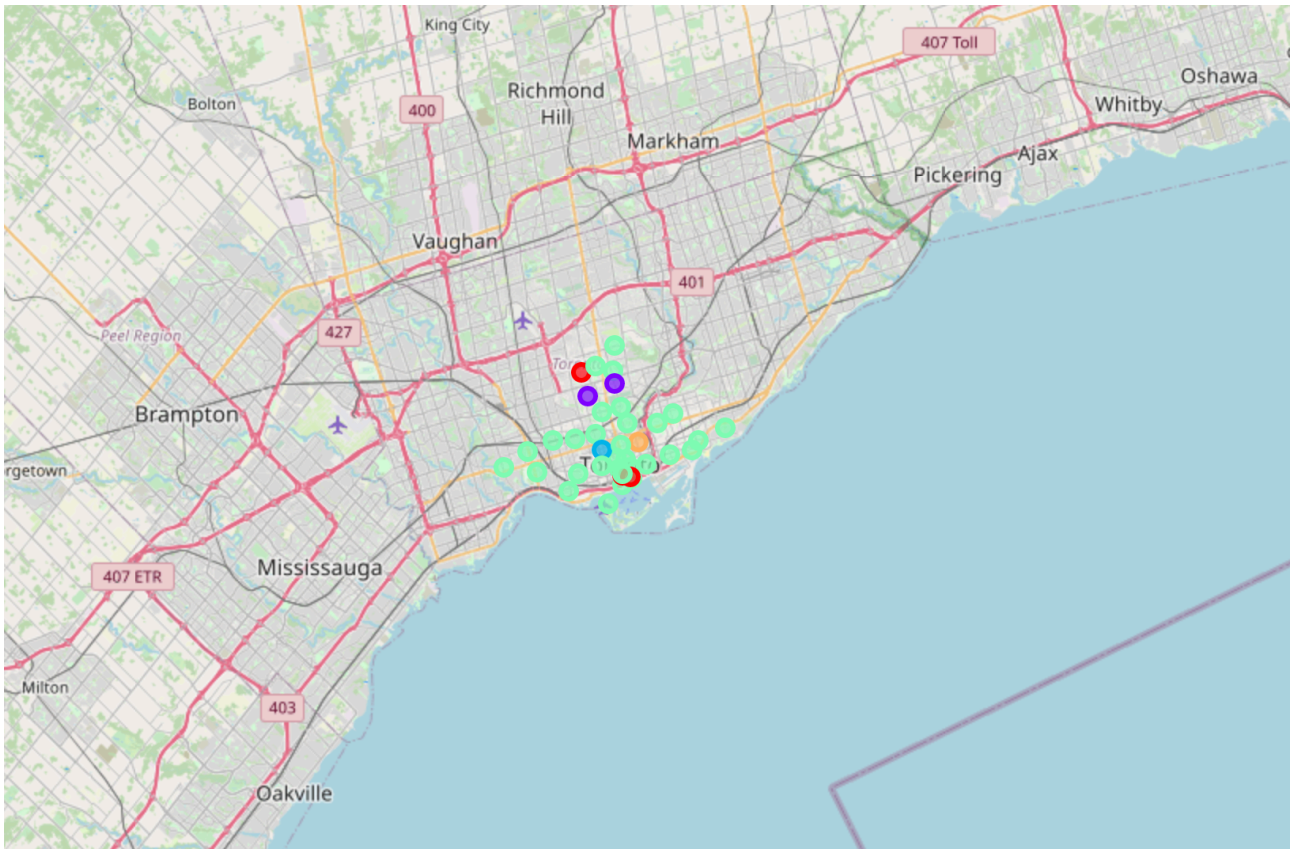
In [151]:
```
trt_clust_5=toronto_merged2.loc[toronto_merged2['Cluster Labels'] == 4, toronto_merged2.columns[[1] + list(range(6, toronto_m
erged2.shape[1]))]]
trt_clust_5
```

Out[151]:

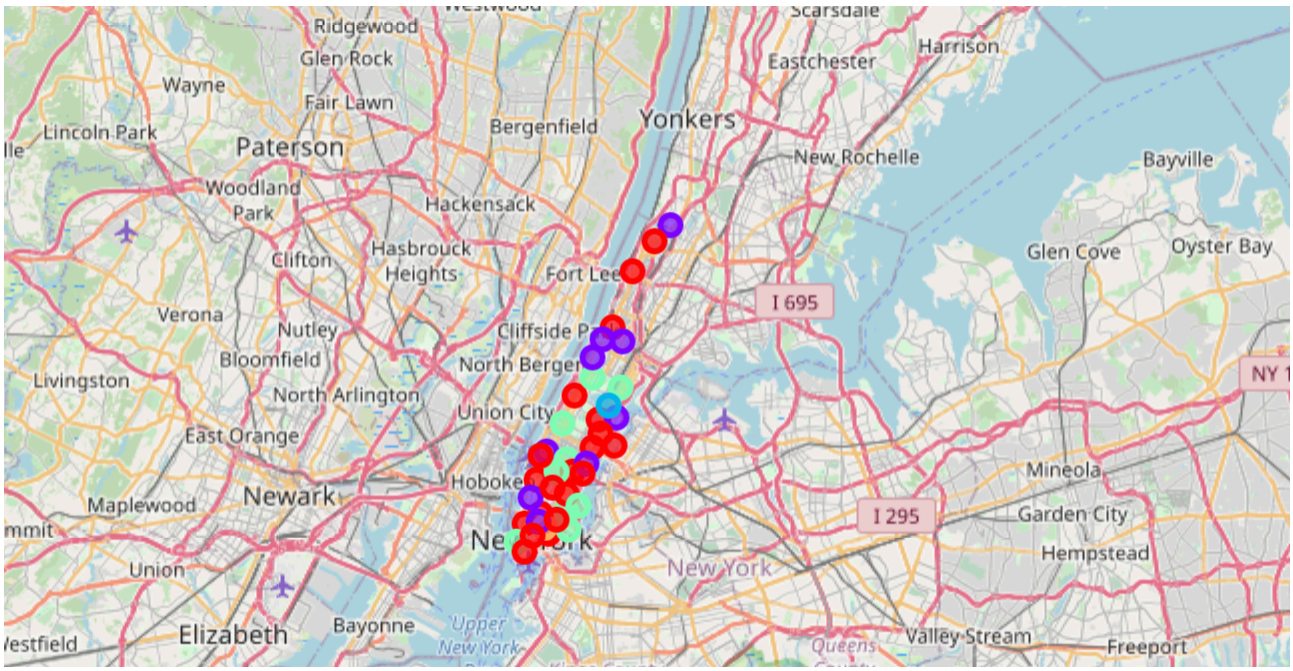
	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
35	Downtown Toronto	Coffee Shop	Café	Chinese Restaurant	Italian Restaurant	Pub	Pizza Place	Pharmacy	Bakery	Restaurant	American Restaurant

For better overview the clusters are visualised on Folium maps as below:

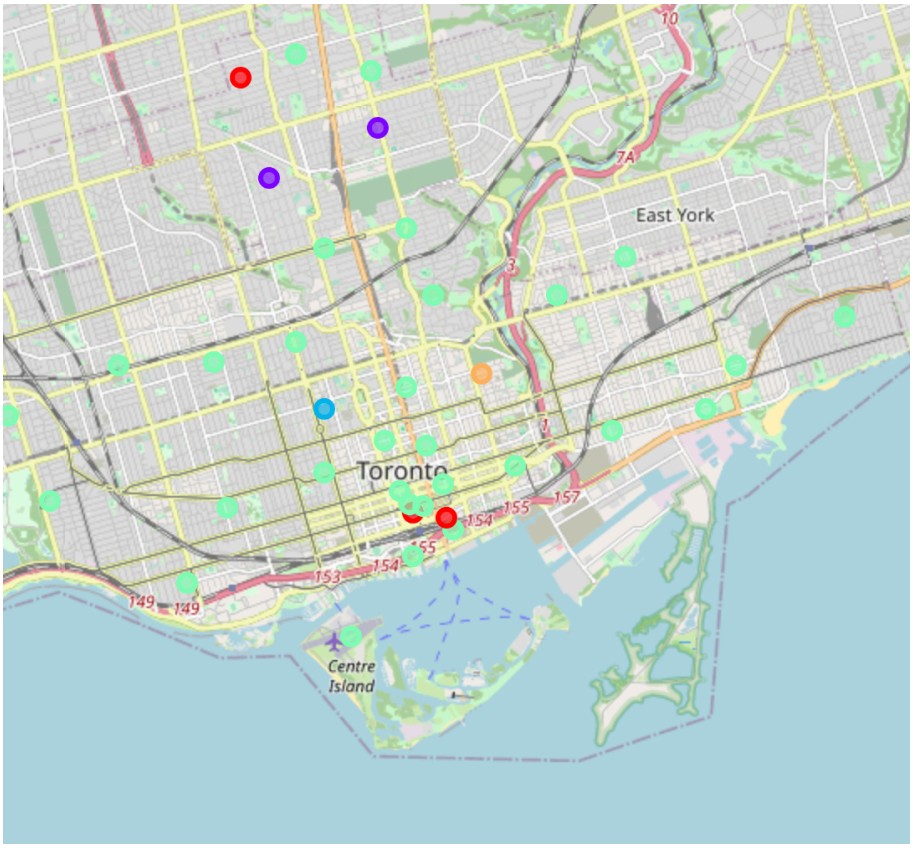
Toronto (default zoom):



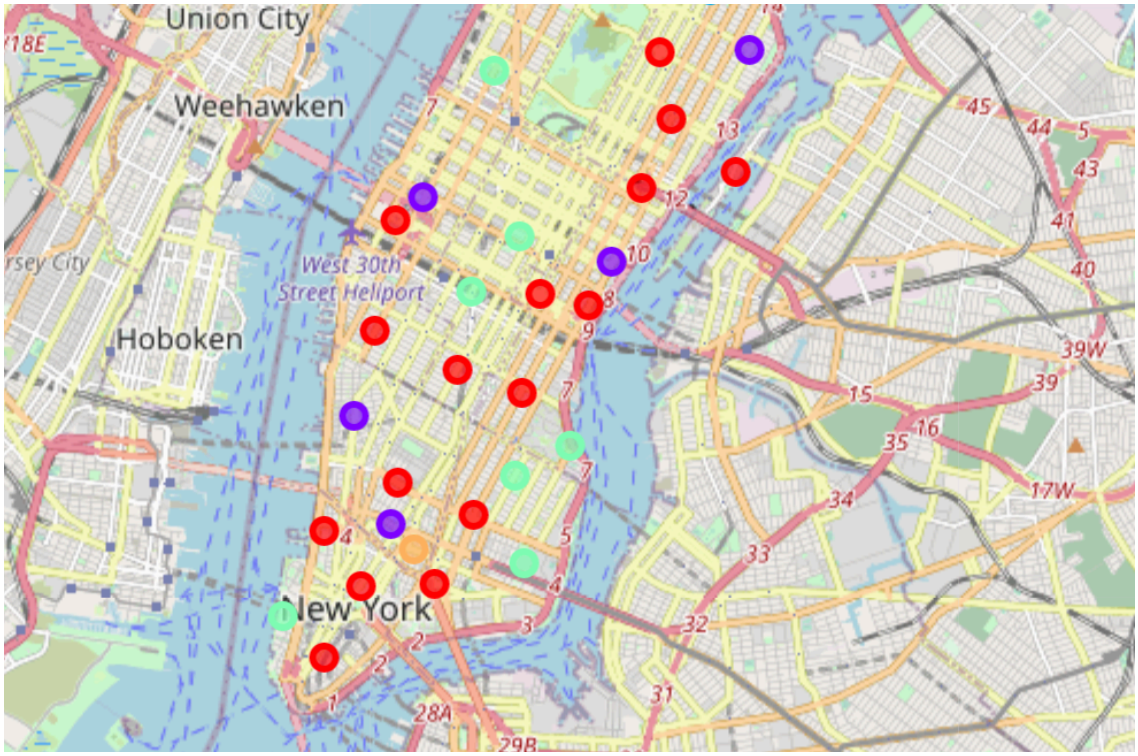
New York (default zoom):



Toronto (increased zooming):



New York (increased zooming)

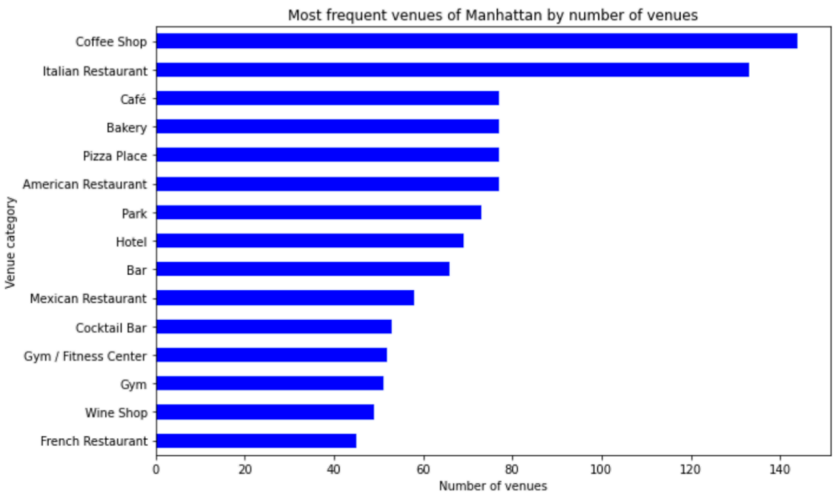


Finding most common venues.

To get better insights most common venues in both cities and also the frequency of venues occurrence in particular neighbourhoods were found. Dataframes for both metropolis have been built and later visualised with help of Matplotlib Pyplot in the form of horisontal bar charts.

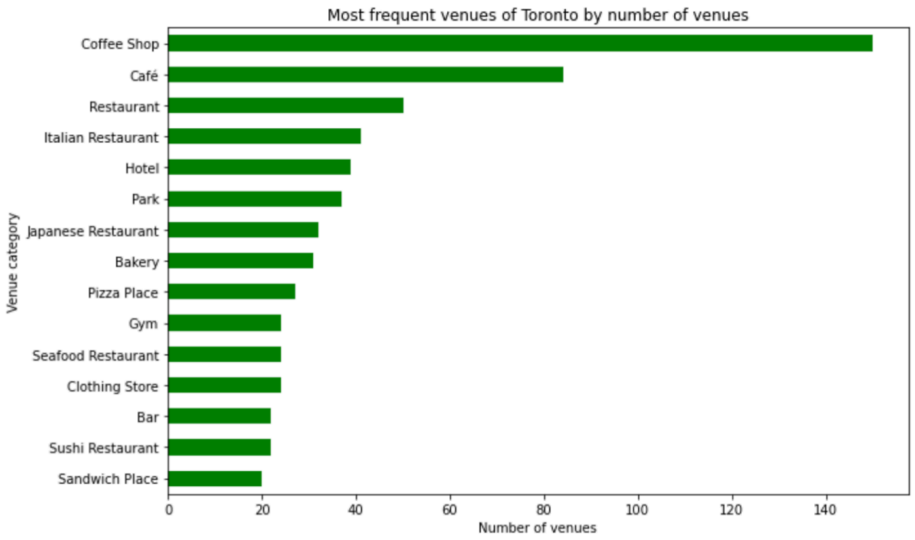
- **Top 15 venues for New York.**

	Venue Category
Coffee Shop	144
Italian Restaurant	133
Café	77
Bakery	77
Pizza Place	77
American Restaurant	77
Park	73
Hotel	69
Bar	66
Mexican Restaurant	58
Cocktail Bar	53
Gym / Fitness Center	52
Gym	51
Wine Shop	49
French Restaurant	45



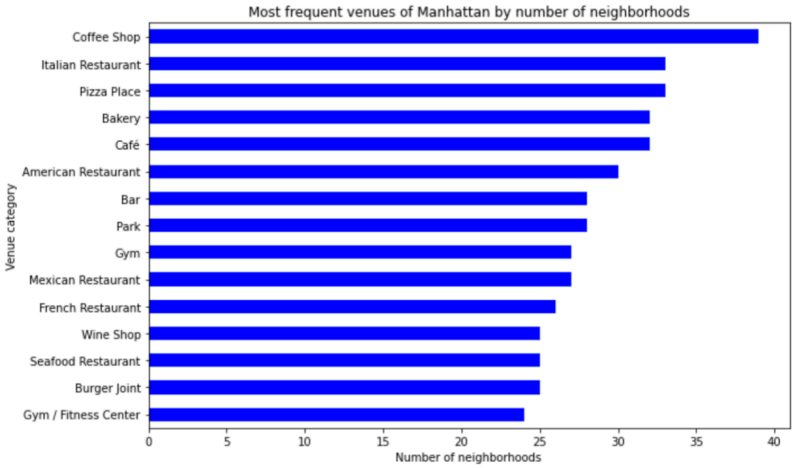
- **Top 15 venues for Toronto.**

	Venue Category
Coffee Shop	150
Café	84
Restaurant	50
Italian Restaurant	41
Hotel	39
Park	37
Japanese Restaurant	32
Bakery	31
Pizza Place	27
Gym	24
Seafood Restaurant	24
Clothing Store	24
Bar	22
Sushi Restaurant	22
Sandwich Place	20



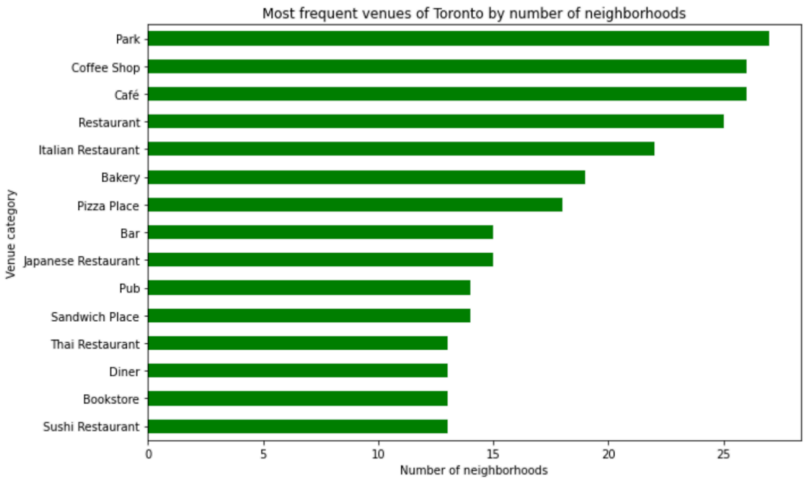
- **New York venues in number of neighbourhoods.**

	Number of neighbourhoods
Coffee Shop	39
Italian Restaurant	33
Pizza Place	33
Bakery	32
Café	32
American Restaurant	30
Bar	28
Park	28
Gym	27
Mexican Restaurant	27
French Restaurant	26
Wine Shop	25
Seafood Restaurant	25
Burger Joint	25
Gym / Fitness Center	24



- **Toronto venues by number of neighbourhoods.**

	Number of neighbourhoods
Park	27
Coffee Shop	26
Café	26
Restaurant	25
Italian Restaurant	22
Bakery	19
Pizza Place	18
Bar	15
Japanese Restaurant	15
Pub	14
Sandwich Place	14
Thai Restaurant	13
Diner	13
Bookstore	13
Sushi Restaurant	13



PART 5

Findings.

After analysis, visualising and presentation of the dataset we can clearly see that there can be identified clusters of neighbourhoods within the cities based on similarities between neighbourhoods.

We clearly see that for both Toronto and Manhattan most common venues are coffee shops, cafe's, restaurants and parks.

Distributions of various venues in the cities can help to make decision which neighbourhood is most suitable for a person or stakeholder based on his/hers needs.

Analysing of extracted date will contribute to easier decision making in for example:

- choosing most suitable location to move and/or live,
- finding based place to work, based on neighbourhood facilities,
- exploring the city as tourist, based on desired venues.

Conclusion.

Extracted data can serve as beginning of decision making proces for stakeholders. Having the data in the form of tables, graph and finaly visualising them on the map will be contributing factor for drawing conclusions. Now we have some overview of both Toronto and New York. Looking at the data and maps one can easily identify the area of interest for stakeholder, based on personal needs.
Main drawbacks of this analysis is only one source of data for clustering (from the Foursquare API). Possibly more data about both cities could lead to different outcome. It would be also helpful to add another criteria to analysis (more categories, rating of venues etc).

The results shows that one can find similarity for both cities but differences also exist. For example both New York and Toronto have plenty of coffee shops, restaurants and cafee s for certain locations of the cities. The analysis might help in finding best location to live, to stay, or to visit.