Introduction / Business Problem

One restaurant chain managed to grow its business significantly over the last 5 years. They started with small restaurant in Venice, in 2 years opened 5 more restaurants in Italy and now they have restaurants in many other European countries - France, Spain, Germany and Austria.

Their main specialty is Italian cuisine, but recently they added few American dishes in menu because they found out that it attracts even more customers. And since healthy trends are becoming more popular every day, they also created special menu for those who care a lot about their health. This strategy turned out to be successful - lots of customers, high revenue and ambitious plans for the future development.

Now they are considering entering new continent - North America. The management believes they should start with big, international city since there are higher chances that some people there might be familiar with the chain.

Currently two options are considered - New York and Toronto.

Strategic planning team asked me to conduct research and provide a reasonable proposal of the best option. Along with the city choice, I should provide recommendations of optimal neighborhoods where to open the restaurant, considering the following:

- Competition - estimation of how many similar places already exist nearby
- Presence of Fitness Centers and parks nearby - because restaurant wants to promote its special healthy menu

Data Section

In order to conduct necessary research, the following data sources will be used:

New York Dataset

- Dataset of New York neighborhoods - contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. Source: https://geo.nyu.edu/catalog/nyu_2451_34572

Toronto Dataset

- List of postal codes of Toronto - Wikipedia page for scrapping data that is in the table of postal codes. Source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- CSV file with Toronto data - file that has the geographical coordinates of each postal code. Source: http://cocl.us/Geospatial_data

Comparing the cities and their neighborhoods

- Foursquare location data - to explore venues, cluster neighborhoods based on their similarity and discover restaurants, fitness centers and parks in neighborhoods to make the best choice of where to open the restaurant.

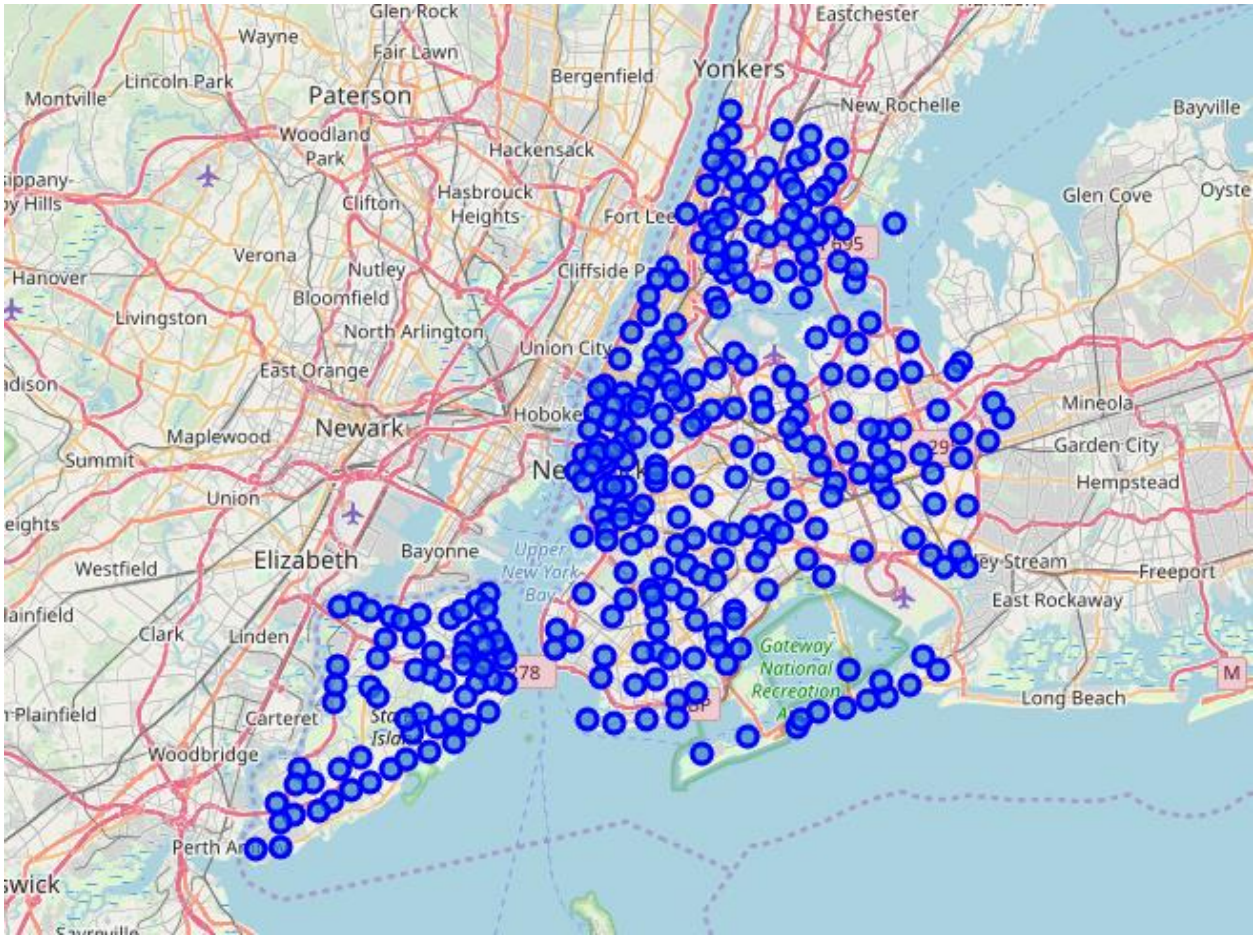
Methodology

Research was done in Jupiter Notebook, using Python and its libraries. There are 2 parts – first deals with New York data, and second is for Toronto data. I will go through each part in detail to explain every step done.

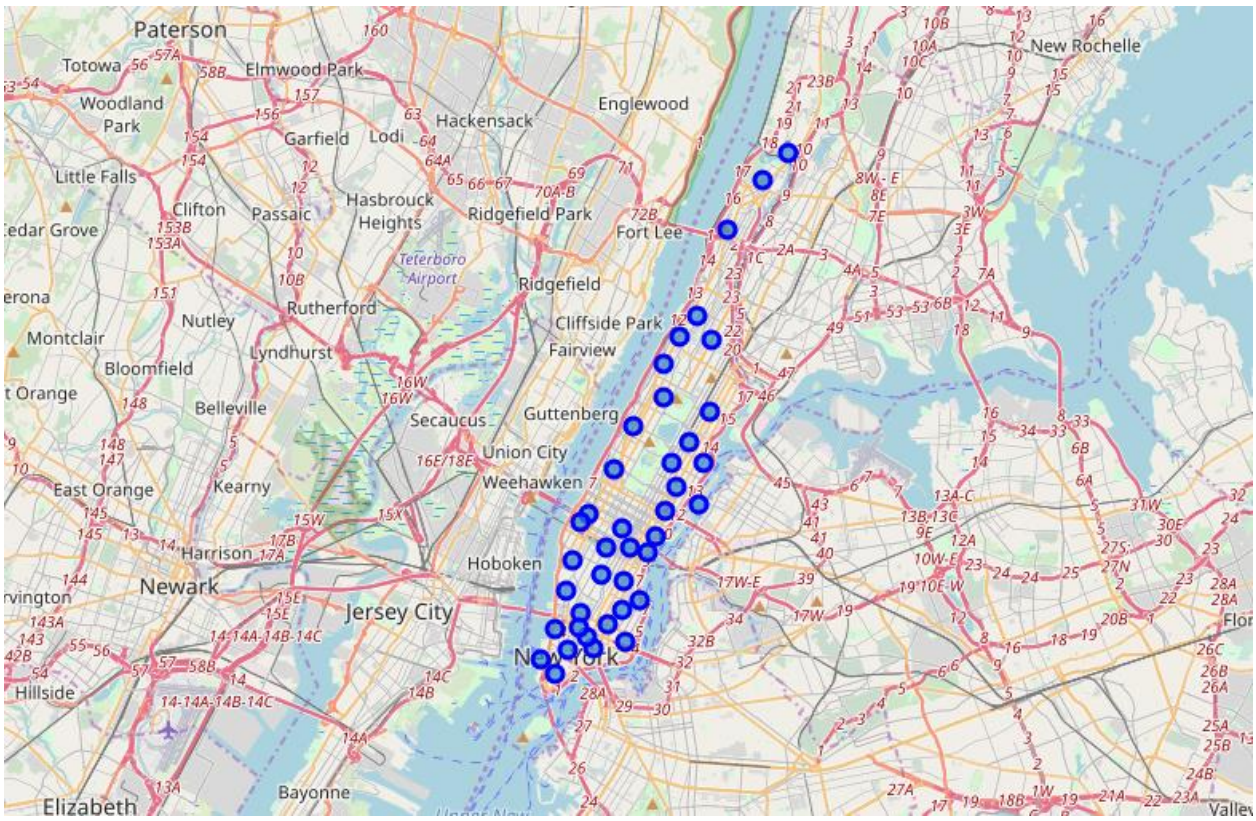
Part 1. New York dataset

I've started with importing necessary libraries and loading provided JSON file with geographical coordinates of all NY neighborhoods. Then this data was loaded to pandas dataframe, my first discovery was that there were 5 boroughs and 306 neighborhoods in total.

With help of geopy library I found out coordinates of NY and created a map of the city with neighborhoods superimposed on top:



I decided to work with the central part – Manhattan and sliced original dataframe to include only this borough. Later I used geopy library again to discover coordinates of the whole borough and visualized its neighborhoods on the map:

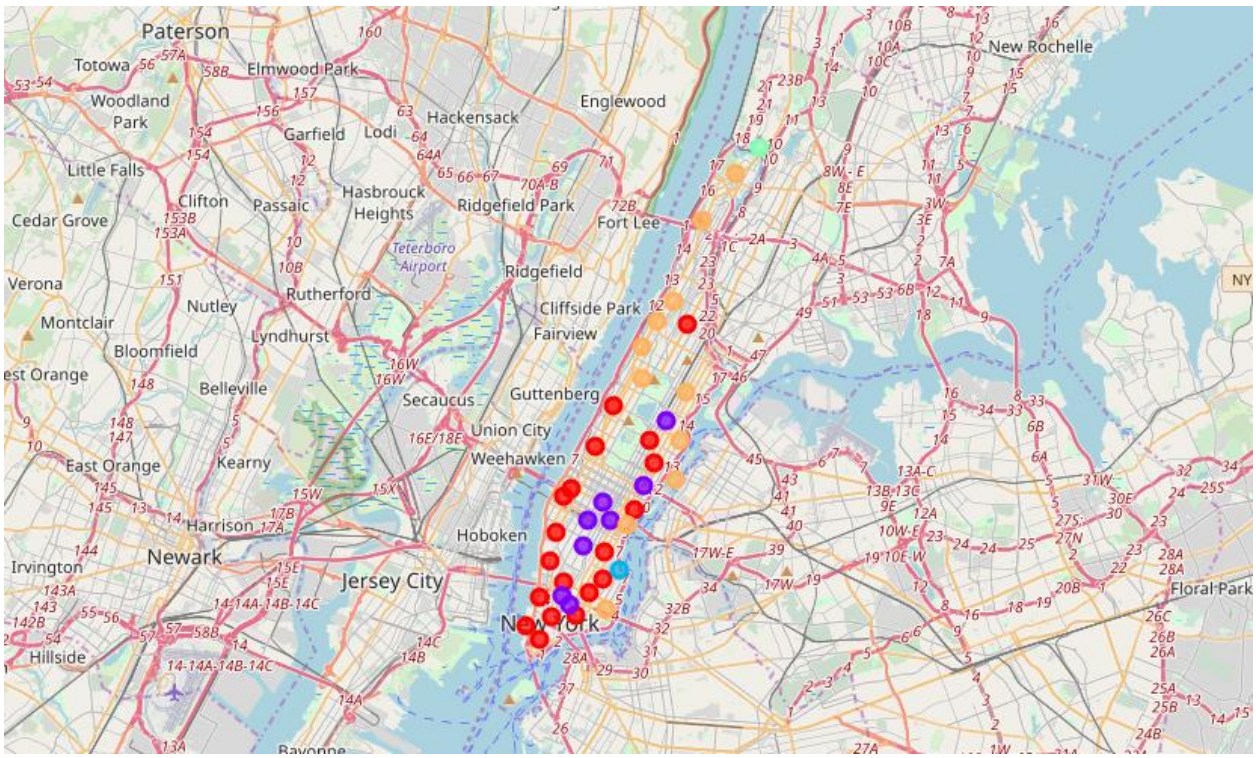


Then I've used the function that was created by the lecturers of this course to get a list of nearby venues and used foursquare API to get names of venues in each neighborhood. In total, there were 341 unique categories.

In order to apply ML algorithm, I needed to turn categorical names into numbers so I've done one-hot encoding and created new dataframe. Then I grouped rows by neighborhood and took the mean of the frequency of occurrence of each category.

Out of received 100 venues for each neighborhood I wanted to analyze top-10 most common venues so I've used a function to rearrange my dataset again.

I chose k-means algorithm for clustering because it's easy and intuitive way to identify different groups of neighborhoods. Number of cluster was set to 5, after fitting into a model Cluster labels were added to the final NY dataset. The following map shows the result of clustering:



There were 5 clusters in my data, so I examined each of them. In order to understand what is typical about each cluster I counted how many times each venue type appeared among top-10 list. Here are my results for 1st cluster:

```
Italian Restaurant      13
Coffee Shop             12
American Restaurant      9
Hotel                    9
Cocktail Bar             8
Café                     7
Gym / Fitness Center     7
Pizza Place              6
Bakery                   6
Park                     6
dtype: int32
```

Italian and American restaurants are popular there, as well as Gyms and Parks. This cluster seems like a good potential fit for further analysis.

Let's check 2nd cluster:

```
Bakery                      5
Coffee Shop                 5
Gym / Fitness Center        5
Café                        4
Yoga Studio                 3
Mediterranean Restaurant    3
Hotel                       3
Italian Restaurant          3
Japanese Restaurant         3
Gym                         3
dtype: int32
```

There are neither a single American restaurant in the top nor park. This cluster doesn't seem like a good option.

Cluster 3 and cluster 4 consist only of 1 neighborhood and unfortunately none of them had American or Italian restaurant as most common venue, so we won't consider them.

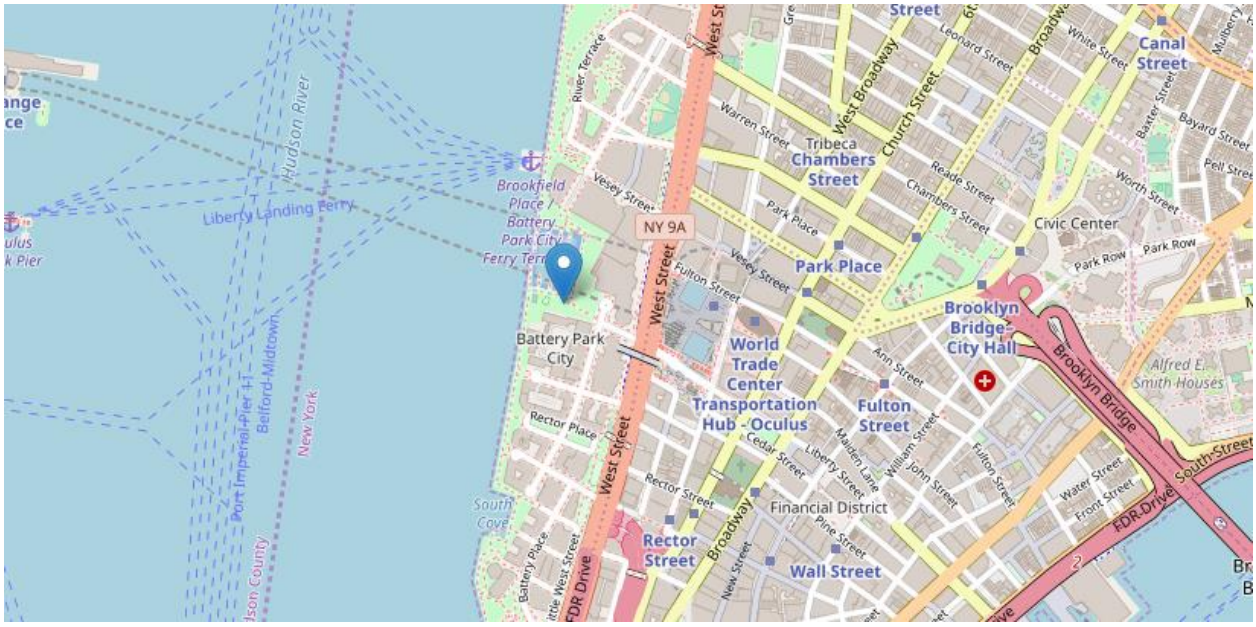
The overview of cluster 5 below:

```
Coffee Shop            9
Deli / Bodega          9
Pizza Place            8
Mexican Restaurant     8
Park                   6
Café                   5
Sandwich Place         4
Bakery                 4
Gym                    3
Japanese Restaurant    2
dtype: int32
```

This cluster has parks and gyms as most common venues but not any of our cuisines that we are looking for.

So I've decided to have a closer look on Cluster 1 since it has all the features that we were looking for. I've filtered the Cluster 1 dataset in order to exclude neighborhoods with high competition – i.e. those that already have Italian and American restaurants, and I wanted to exclude neighborhoods where bars were among top-10 most common venues – since it doesn't match with healthy trends.

After applying the following filter, only one neighborhood was left - Battery Park City. So my option for NY city is this neighborhood:



Part 2. Toronto dataset

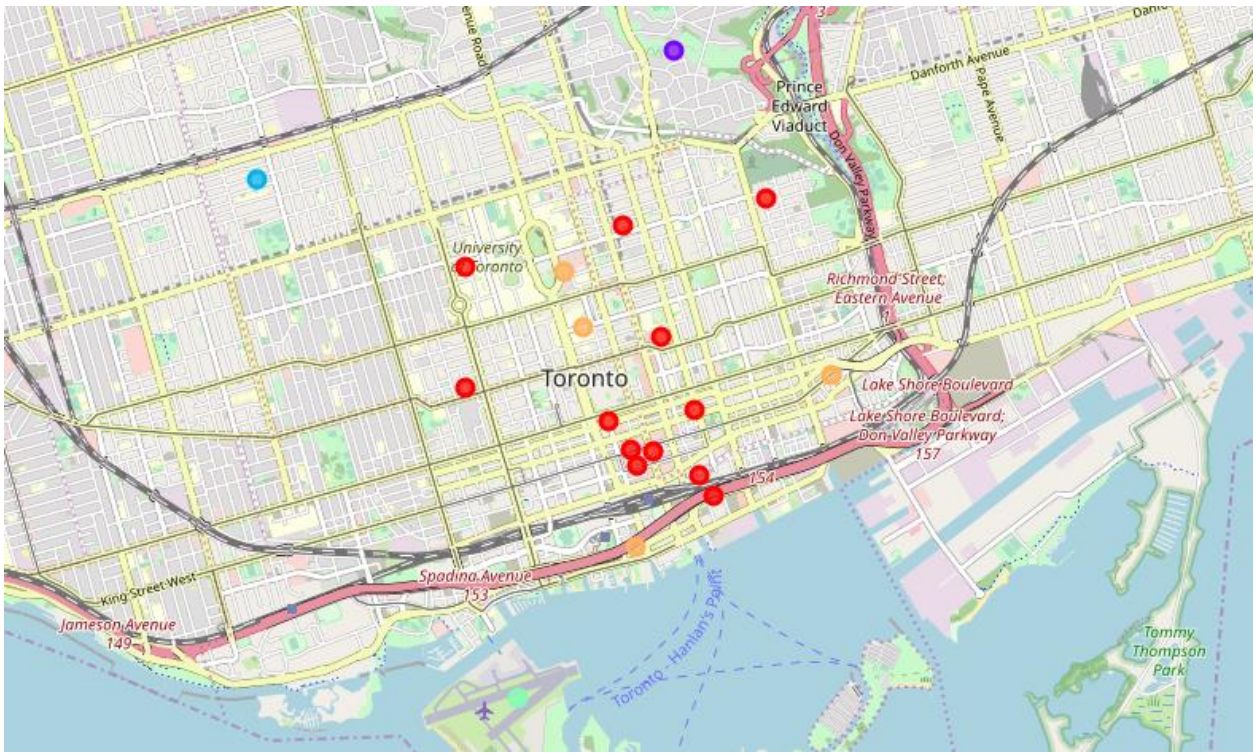
In order to compile Toronto dataframe I had to use 2 sources – Wikipedia page and provided csv file with geographical coordinates of postal codes in Toronto.

I started with scrapping Wikipedia page with BeautifulSoup library. Then I cleaned dataset from ‘Not assigned’ values and grouped neighborhoods with the same postal codes in one line. I’ve merged this table with second file, adding coordinates for each post code.

To keep the same logic as with New York dataset, I’ve decided to work only with Downtown Toronto borough, so I’ve limited my dataframe to that borough. Here on the map you can see all neighborhoods in Downtown Toronto:

Then I’ve applied the same function that I used with New York data to return venues from Foursquare API. I found out that there were 207 unique categories.

I’ve done the same transformation to Toronto dataset that I’ve done to New York dataset to prepare it for fitting the model. K-means algorithm was set to 5 clusters and applied to the data.



Cluster 1, that represents red markers had the following characteristics:

```
Café                   12
Restaurant             11
Coffee Shop            11
Hotel                   7
Japanese Restaurant     6
Bakery                  5
Seafood Restaurant      5
Italian Restaurant      5
Gastropub               4
Bar                     4
dtype: int32
```

We see that Italian restaurant here is in top-10 most common venues, but this cluster doesn't have gyms or parks in the list.

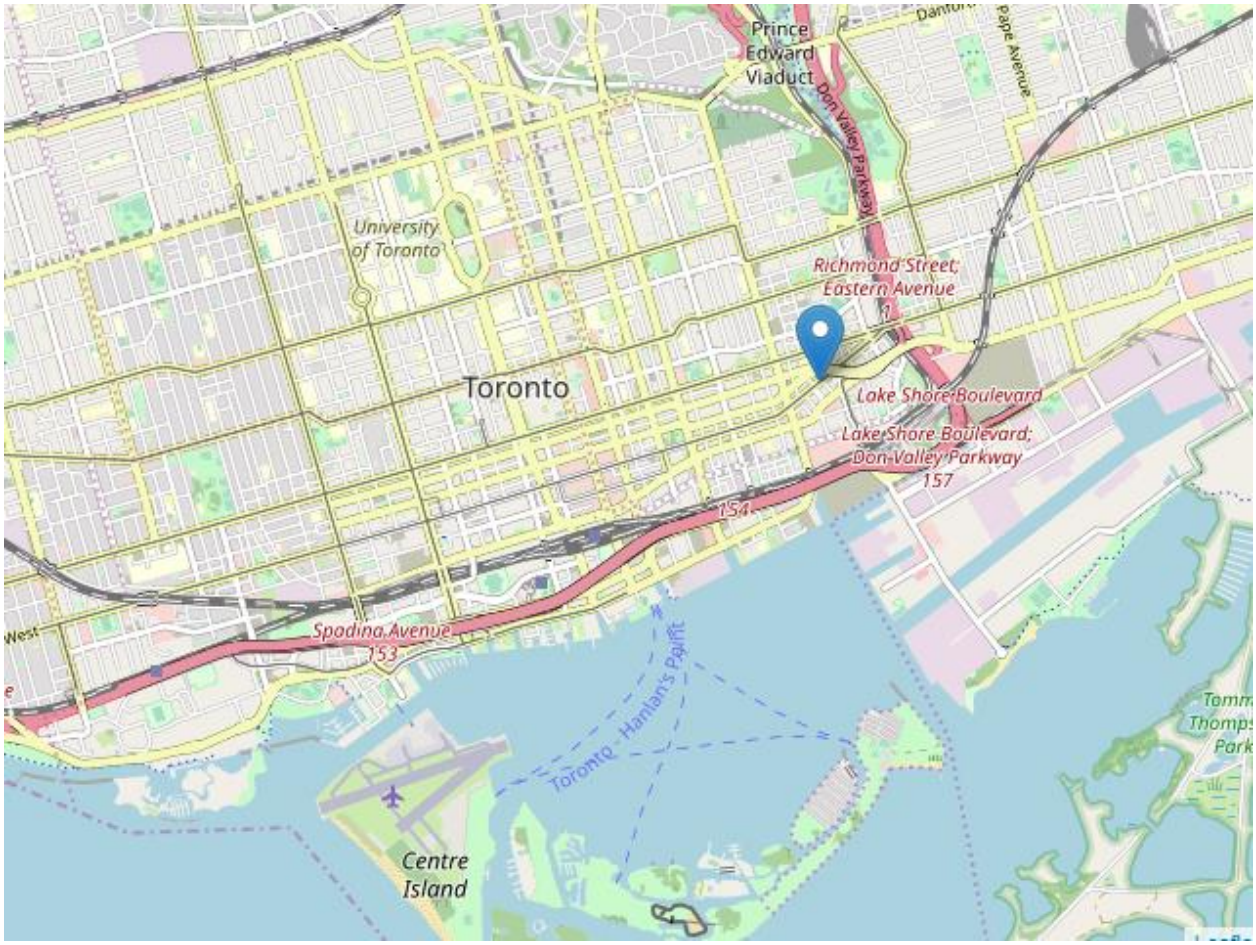
Each of Clusters 2-4 was represented by 1 neighborhood, and one of them was located around airport area. Their location and description don't seem to be a good option.

Finally, Cluster 5 (orange marker) had the following characteristics:

```
Coffee Shop            4
Restaurant             3
Café                   3
Italian Restaurant     3
Park                   2
Mexican Restaurant     2
Japanese Restaurant    2
Yoga Studio            1
Diner                  1
Bakery                 1
dtype: int32
```

Italian restaurant, followed by the park are among the most common venues in the area. This option is the best.

After applying the same filters – excluding the neighborhoods that already have Italian restaurant, and neighborhoods with bars, I've ended up with the following option – **Harbourfront** neighborhood:

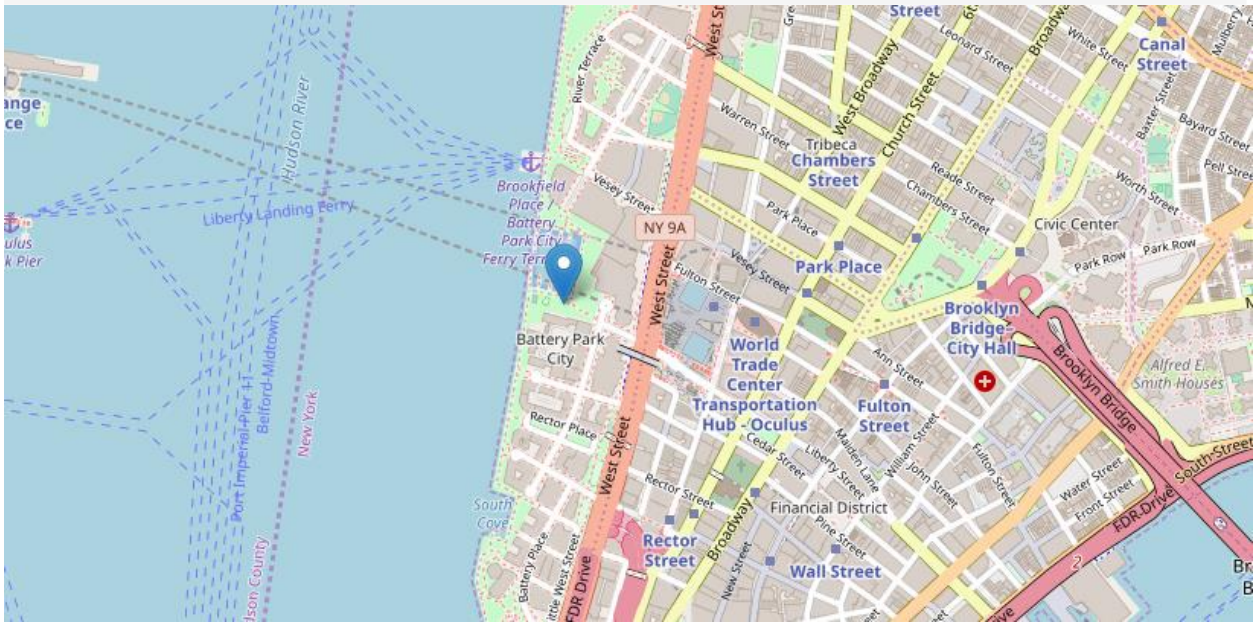


Results

After doing all the necessary research I've come up with these 2 options:

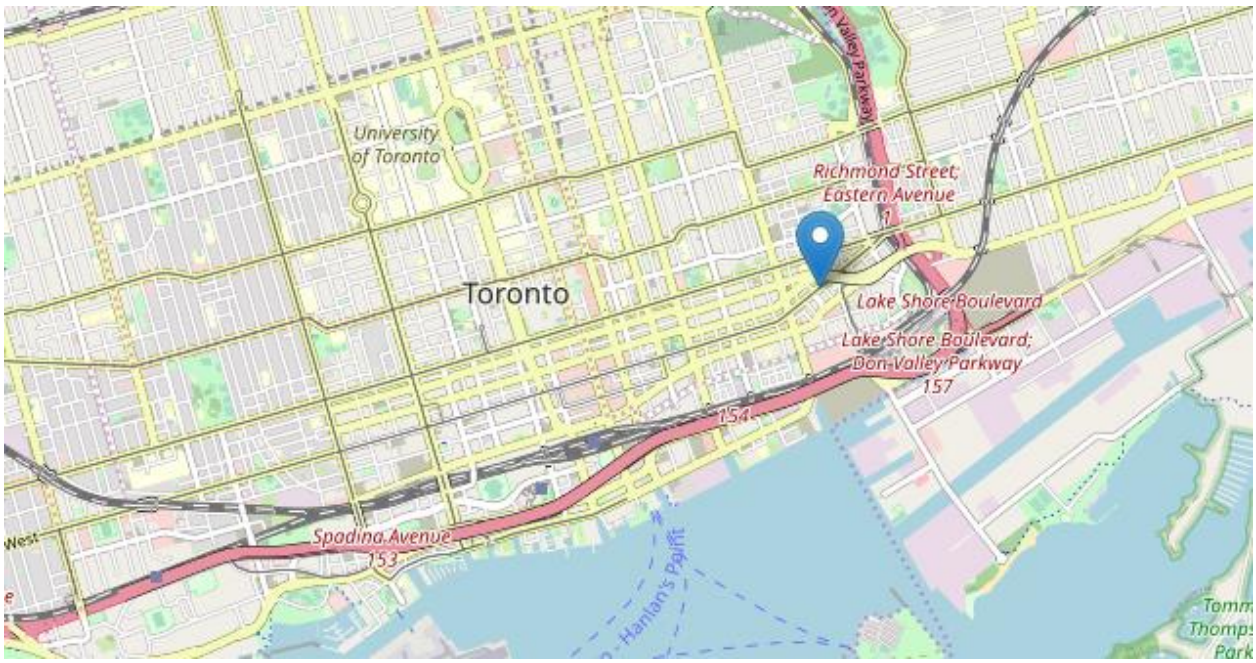
for New York

Neighbourhood	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
Battery Park City	Park	Coffee Shop	Hotel	Wine Shop	Gym	Shopping Mall	Women's Store	Memorial Site	Food Court	Men's Store



and Toronto

Neighbourhood	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
Harbourfront	Coffee Shop	Park	Café	Pub	Bakery	Mexican Restaurant	Restaurant	Theater	Breakfast Spot	Performing Arts Venue



As you can see, neighborhood in New York is located along the waterfront, making it more attractive in terms of relaxing atmosphere. People who worked all day surrounded by the skyscrapers may feel better when they go out for a dinner somewhere where they can see a horizon line, plus west side makes this place fabulous during the sunsets.

From my point of view, out of these 2 options New York neighborhood **Battery Park City** is the best optimal place to open the restaurant. As for Toronto neighborhood **Harbourfront** – it may be considered as the second step in expanding on the North American market.

Discussion

Overall this analysis was very interesting – I've discovered different clusters of New York and Toronto and shared my opinion on best neighborhood to open a restaurant.

This research could go further and analyze rental and other costs, coming up with exact address where this opening in the chosen neighborhood could be most reasonable. I will leave it for the next time because it would make this analysis long and frustrating for the reader.

Conclusion

In this report I've done research on choosing the best city and neighborhood among New York and Toronto for European restaurant chain to open their first restaurant in North America.

I've considered all concerns the management had: the area should have high people traffic, not so many Italian or American restaurants, it should be somewhere near to the park and has some amount of fitness centers around.

Central districts of 2 cities were analyzed, their data was fitted into a model, clustered based on their similarity and one neighborhood in each city was selected as the optimal option. In Results section I've explain how I chose the best out of them – this decision didn't need math or anything – so I didn't include it into methodology section.

My final choice is **Battery Park City** neighborhood of Manhattan, New York. This place has all the characteristics requested by the strategic planning team and has relatively low competition. Thus the goal of this analysis was fulfilled, leaving it for the next time to find out appropriate building for the restaurant.

Thank you for your time and attention.