Battle of Neighborhoods Report

Simon Xin

1. Introduction

The health industry and its related businesses have been growing steadily during the past few decades in North America. Given the soaring awareness of personal healthcare and pursuits of healthy lifestyles, new business interests emerge in those big metropolitan areas which cover a wide range of products and services. In this project, we investigate and compare the potential investment opportunities for healthy food business in two major financial centers: New York City and the city of Toronto.

Since the recent legalization of recreational marijuana in Canada and an increasing number of states in the U.S., cannabis and other natural herb based heath products are given more attention as diet supplement and health improvement alternatives. So here we want to forecast and compare the business potential of cannabis related heath products in those cities by studying the heath lifestyle preference statistics and based on that, further identify the most promising locations of our new business.

2. Data acquisition and pre-processing

2.1 Data sources

The datasets are obtained using Foursquare API. Foursquare Places API provides real-time location data by some of the most useful queries for venue search, venue statistics, and user information.

In this project, we use regular endpoint GET() to explore locations with specific geographical coordinates, i.e. neighborhoods in New York City (NYC) and Toronto. In particular, we search for venues of interests within the radius of 500 meters from the center of each neighborhood. The venue of interests means those venues that can be considered as indicators of a health lifestyle and likelihood of interests in cannabis and natural herb related health products. It will be discussed in details in feature section. Useful information from the query results include venue name, location, category, distance to center.

Apart from Foursquare, we further acquire information from websites such as Wikipedia and others to get demographic statistics regarding boroughs and neighborhoods in NYC and Toronto. The most relevant information used here is population.

2.2 Data pre-processing

The data pre-processing includes a number of steps to make the data obtained from sources suitable and ready for exploratory and predictive analysis. First, we want to make sure the data type of each dataset column is consistent and the missing information is properly handled with replacement. From interaction with Foursquare, we find this is well handled by the API via normalizing the JSON file to dataframe.

One thing we found of query results is that the result items are not all related directly to the search keywords. Hence, we further filter result dataset using the same keywords for category and venue name to guarantee that all items are relevant to the search criteria.

Another important point to notice in our case is the redundancy of data acquired which could interfere with data analysis and interpretations. A good example of this is the overlap of search results from adjacent neighborhoods, which produces multiple items associated with the same venue. To avoid this situation, we use venue ID as the key to remove duplicates of the same venue and only keep the item that has the smallest distance.

Moreover, to draw a fair comparison between cities with different sizes, we normalize the venue statistics with population to get the idea of the per population number of venues in both cities.

2.3 Feature determination

The proper selection of features greatly influences the accuracy of predictive modeling and reliability of the insights we draw from data analysis. Due to the nature of our business, we narrow down the search keywords to the following list:

- Gym
- Vegan
- Sports
- Nightclub
- Hookah

The first three keywords refer to indicators of healthy lifestyles, such as gym, vegan restaurant, and sports clubs and stores. The latter two words are more related to personal openness and preference for casual and entertaining events that may involve cannabis related products. We use the abovementioned list to get statistics of particular categories of venues and for each venue, we include the ‘Distance’ data to indicate the convenience of the venue location. The ‘Distance’ data also reflects the location density of interesting venues that is helpful to further pinpoint the optimum location of our potential business. As a result, we list the features used in this project:

- Gym
- Vegan
- Sports
- Nightclub
- Hookah
- Averaged distance

3. Exploratory data analysis

3.1 Number of venues by category

First we are interested to know the statistics of venues of interests in each city. Due to the quota of limited calls to Foursquare API endpoints, we use the most concentrated and popular borough to represent the two cities: Manhattan, NY and Downtown Toronto. We group the dataset by neighborhood and sum up the number of venues for each category. The top 10 neighborhoods with the highest number of venues are plotted in the bar chart to see the breakdown number of each feature venue.

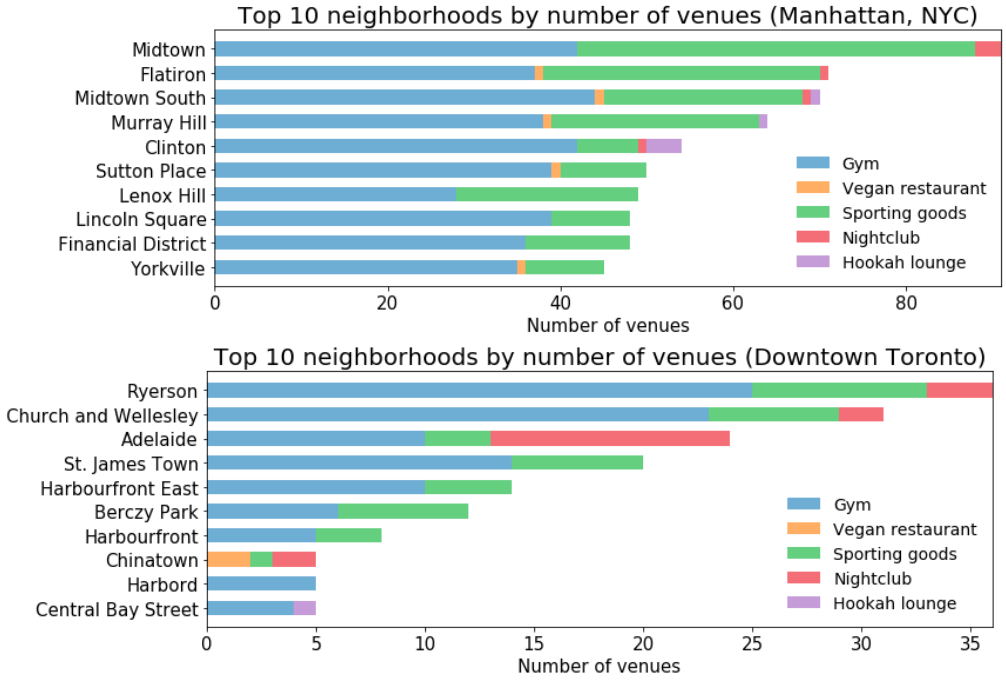


Fig. 1 Bar charts showing number of venues by neighborhoods of Manhattan and Downtown Toronto

From the plots of both cities, there are some similar features. First, for the majority of all neighborhoods in both cities, the two most popular types of venues are gym and sporting goods, which agrees with the ground truth according to my personal experience of the two cities.

However, there are still clear differences between them. One thing is that in Manhattan, the venue structure does not vary too much from neighborhood to neighborhood, gym and sports stores being the two dominant categories. And the total number of venues does not drop drastically from top to bottom, which indicates that the layout of neighborhoods is more flattened and neighborhoods do not differentiate themselves mostly by the number or type of venues they offer. Nevertheless, it is possible to identify some 'Rock star' neighborhoods, e.g. Midtown, Flatiron, Midtown south, that are more popular than others. For our business, it is interesting to see some hookah lounges in Clinton, which could be a positive factor.

While for Downtown Toronto, we see a sharp decline in the number of venues from top to bottom neighborhoods which means the district is more concentrated and there might be only a few neighborhoods ideal for business, such as Ryerson, Church and Wellesley, and Adelaide. Additionally, each of the three neighborhoods has a high presence of nightclubs, especially Adelaide, which might be a positive factor for our business during the night time business hours.

To take out the factor of population from the above venue statistics, we consider the total population in both cities:

- Manhattan: 1.66 Million (2017)
- Downtown Toronto: 0.25 Million (2016)*

* https://www.cp24.com/news/downtown-population-will-nearly-double-by-2041-amid-building-and-baby-boom-keesmaat-1.2846605

We calculate the per population number of venues:

- Manhattan: 6.67 per 10,000 people
- Downtown Toronto: 6.48 per 10,000 people

Therefore, we find in terms of the per population figures, both city centers possess similar opportunities for our business although Manhattan has many more venues in terms of the absolute number. Then, the following question is where to locate our business, which will be discussed in details in the sections below.

3.2 Venue location convenience study

In addition to the number of venues, the location convenience is also a key factor for the success of a business. Here, we plot the statistics of averaged distance to neighborhood center for top 5 neighborhoods in Manhattan and Downtown Toronto. The box plots show that Midtown South has the smallest average distance of all venues with a median distance around 320 meters, and third quartile around 400 meters. Given that the Midtown South is also among the top 3 neighborhoods in terms of the number of venues, we should locate our business within 400 meters from the centroid in order to have the best exposure to potential customers.

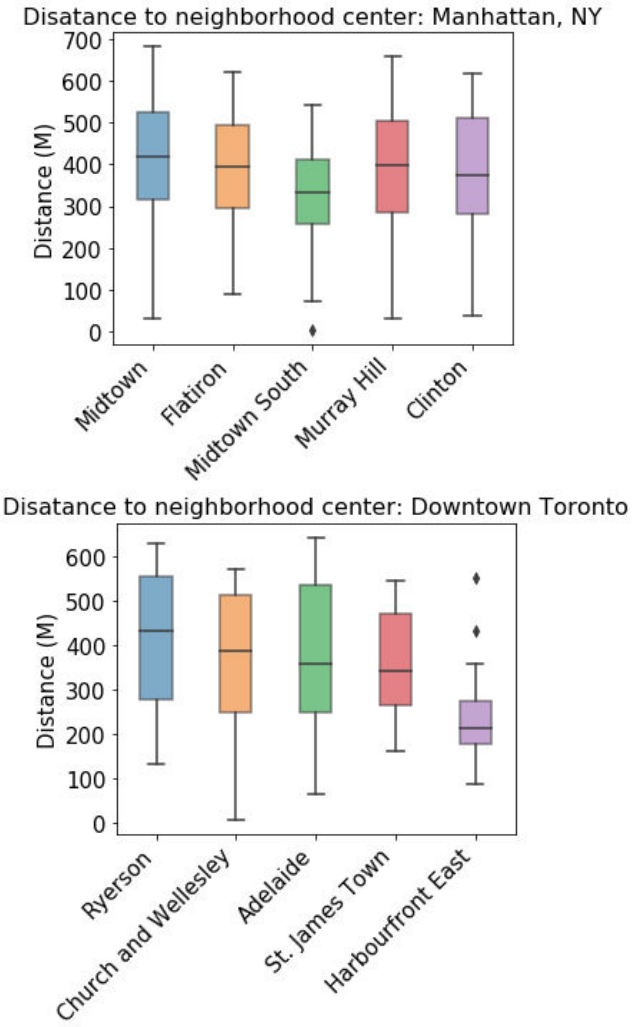


Fig. 2 Box plots of averaged distance to neighborhood center

While for Downtown Toronto, the box plot shows that the neighborhoods are much more spread out with the highest median distance ~ 420 meters and third quartile ~560 meters for Ryerson. For Church & Wellesley and Adelaide, the number does not vary too much. Based on the findings we have from previous dicussion, we might want to locate our business in Adelaide due to its high presence of nightclubs. And given the spread-out layout of the neighborhood, the ideal busines spot should be within close vicinity to those most popular nightclubs.

3.3 Visualization of venue distribution on the map

To have a more intuitive view of the venue distribution among Manhattan neighborhoods, we plot the choropleth of total number of all feature venues. The color coded map of Manhattan shows the popular neighborhoods for health businesses located near the Midtown area. Combined with life experience knowledge, the results

make sense and our best hope would be near the Midtown area where the business will be exposed to thousands of local residents and travelers from all over the world.

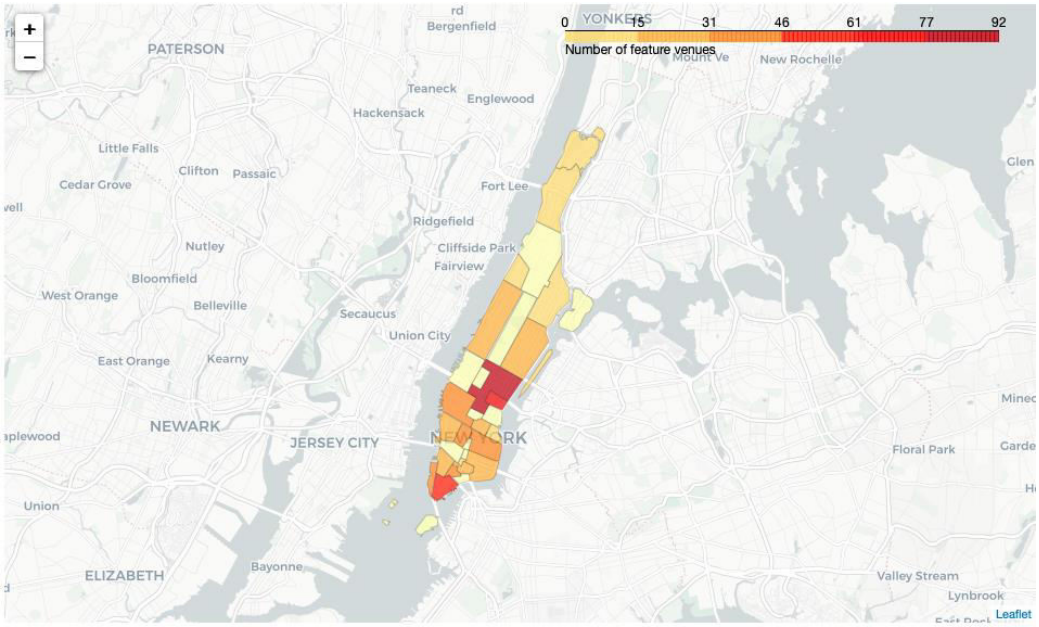


Fig. 3 Choropleth plot of neighborhoods in Manhattan

4. Predictive modeling

From previous sections of discussions, we already have critical insights into the neighborhoods in both cities and their unique potential for our new business. To further consolidate our findings, we want to build a predictive machine learning model to segment all neighborhoods into a number of clusters in terms of its unique set of features, which is repeated here below:

- Gym
- Vegan
- Sports
- Nightclub
- Hookah
- Averaged distance

4.1 K-Means model set up

In this project, we use K-means method for clustering process. First, we put all information together in the final feature dataset. Then we import the K-means function from Scikit learn. Before we set up the machine learning model, we need to first standardize feature dataset. Then, an important step in K-means modeling is to determine the number of clusters for best accuracy and justification. Here we scan the K

number and plot the Inertia vs. K number. Inertia is the sum of square distance of each data points to its centroid. The curve of Manhattan is shown below:

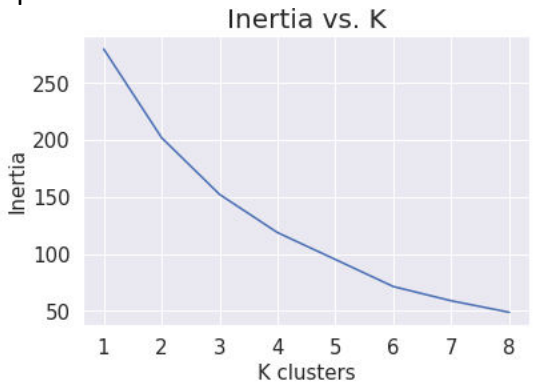


Fig. 4a Inertia vs. K of K-Means model of Manhattan

Here, we find the elbow point at K = 6 and use that for the clustering process. Similar we plot the same curve for Downtown Toronto, and show it below:

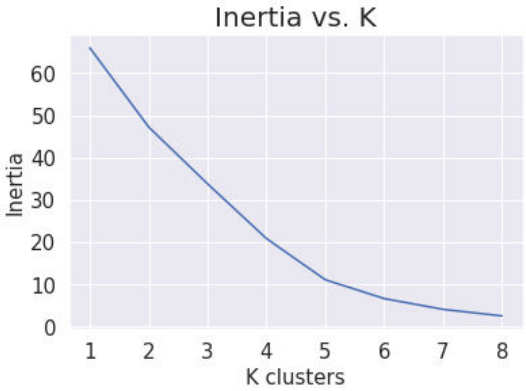


Fig. 4b Inertia vs. K of K-Means model of Downtown Toronto

In a similar way, we choose K=5.

4.2 Examine and visualization of clusters of Manhattan

First, we want to list the member neighborhoods of each cluster and check the mean statistics of all features in each cluster. Here we display the run output below.

From the cluster data, we have a few insights:

- Cluster 3 has the highest averaged numbers of gym, sports goods, and nightclubs. As a result, it represents prime locations for our new business. If we look further into the contents of cluster 3, we can find those ‘Rock star’ neighborhoods, such as Midtown, Flatiron, Midtown South, etc. Therefore, it agrees with our previous discussion.

- Cluster 4 has the highest averaged number of hookah lounges, and relatively high numbers of other feature venues. Thus, it could be a secondary consideration of business locations. Neighborhoods such as Noho and East Village are among those trending places for young population who are looking for the latest fashionable elements. As we mentioned earlier, there are quite a few hookah places in Clinton, which is considered positively in our case.

- Cluster 5 (Greenwich Village) has 5 vegan restaurants that might bring vegetarian customers to our business although there is no direction connection between vegetarian and cannabis.

To sum up, we conclude on top 3 business hubs:

Tab. 1 Top 3 business hubs in Manhattan

Tier 1 locations:	**Midtown, Chelsea, Midtown South, Flatiron**
Tier 2 locations:	Clinton, East Village, Lower East side, Noho
Tier 3 locations:	Greenwich Village

Tab. 2 Cluster details and mean statistics (Manhattan, NY)

```
The clusters in Manhattan, NYC are:
Cluster 0:
    Marble Hill
    Chinatown
    Washington Heights
    Inwood
    Hamilton Heights
    Manhattanville
    Central Harlem
    Upper East Side
    Roosevelt Island
    Tribeca
    Little Italy
    Soho
    West Village
    Manhattan Valley
    Morningside Heights
    Turtle Bay
    Tudor City
    Hudson Yards
Cluster 1:
    Yorkville
    Lenox Hill
    Upper West Side
    Lincoln Square
    Murray Hill
    Gramercy
    Battery Park City
    Financial District
    Carnegie Hill
    Civic Center
    Sutton Place
Cluster 2:
    East Harlem
    Stuyvesant Town
Cluster 3:
    Midtown
    Chelsea
    Midtown South
    Flatiron
Cluster 4:
    Clinton
    East Village
    Lower East Side
    Noho
Cluster 5:
    Greenwich Village
```

```
The averaged feature value of each cluster is:
                   Gym     Vegan      Sport  Nightclub    Hookah    Distance
Kmeans_label
0             8.000000  0.055556   2.944444   0.333333  0.222222  420.549902
1            30.090909  0.272727  10.545455   0.000000  0.272727  366.635438
2             3.500000  0.000000   2.500000   0.000000  0.000000  135.318182
3            36.250000  1.000000  28.750000   1.750000  0.250000  381.289550
4            22.750000  0.500000   6.000000   0.500000  4.250000  352.458399
5            21.000000  5.000000  15.000000   0.000000  1.000000  410.097561
```

Finally, we plot a bubble map on of neighborhoods where the color of the bubbles indicates cluster number and the size of the bubbles indicates total number of venues.

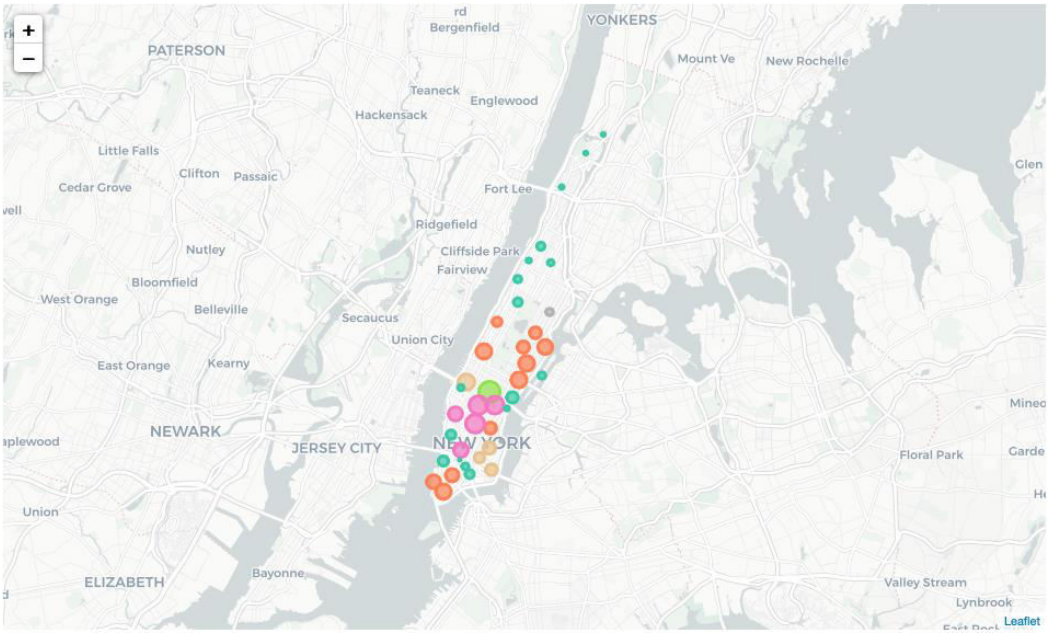


Fig. 5 Bubble map plot of neighborhoods in Manhattan

4.3 Examine and visualization of clusters of Downtown Toronto

Now let's take a look at Downtown Toronto. First we display the cluster details and mean statistics below:

Tab. 3 Cluster details and mean statistics (Downtown Toronto)

```
The clusters in Downtown Toronto are:
Cluster 0:
    Ryerson, Garden District
    St. James Town
    Berczy Park
    Church and Wellesley
Cluster 1:
    Chinatown, Grange Park, Kensington Market
Cluster 2:
    Harbourfront, Regent Park
    Harbourfront East, Toronto Islands, Union Station
    Harbord, University of Toronto
    Cabbagetown, St. James Town
Cluster 3:
    Central Bay Street
Cluster 4:
    Adelaide, King, Richmond

The averaged feature value of each cluster is:
              Gym  Vegan  Sport  Nightclub  Hookah    Distance
Kmeans_label
0            17.0    0.0   6.50       1.25     0.0  403.399014
1             0.0    2.0   1.00       2.00     0.0  442.000000
2             5.5    0.0   1.75       0.00     0.0  285.823214
3             4.0    0.0   0.00       0.00     1.0  433.500000
4            10.0    0.0   3.00      11.00     0.0  375.833333
```

From the data above, we have the following findings:

- Cluster 0 has the highest averaged numbers of gyms and sporting goods, and therefore could be a prime hub for our business. Neighborhoods such as Ryerson and Church and Wellesley provides a variety of interesting places which could be a reliable sources of customers.

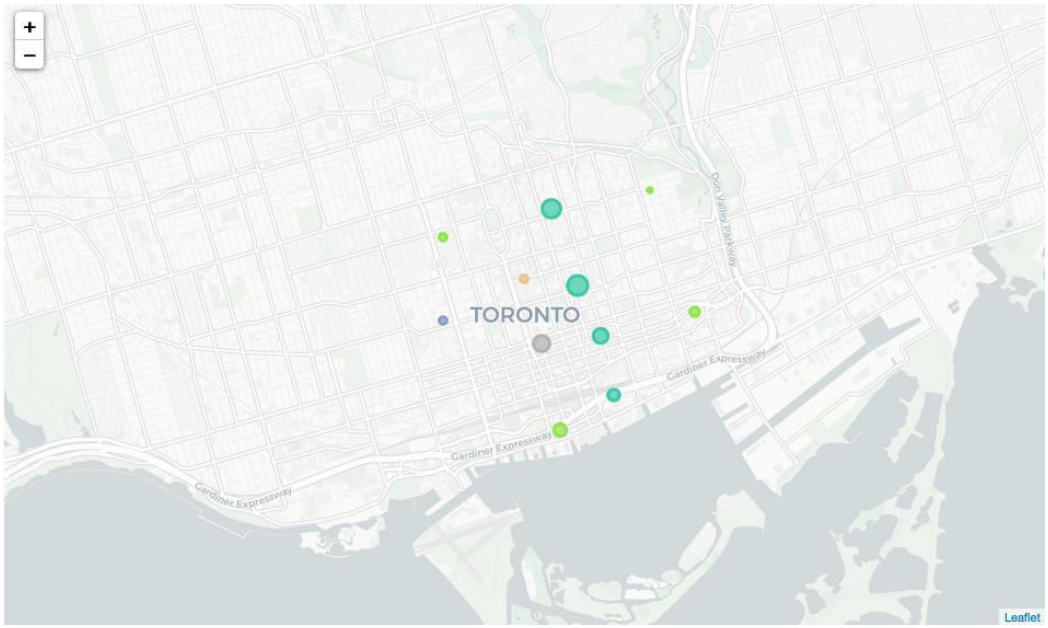
- Cluster 4 has the highest number of nightclubs. And as we discussed earlier, the nearby locations of popular nightclubs in Adelaide could be very promising spots for interested customers.

To sum up, we conclude on top 2 business hubs:

Tab. 2 Top 2 business hubs in Downtown Toronto

Tier 1 locations:	**Ryerson, St. James Town, Berczy Park, Church & Wellesley**
Tier 2 locations:	Adelaide

Finally, we plot the bubble map of all clusters of neighborhoods in Downtown Toronto.



5. Conclusions

In this report, we investigated two financial centers in North America, New York City and Toronto, in terms of potential business opportunities for cannabis and natural herb related health products. Using Foursquare API and other sources, we are able to gather relevant datasets regarding venue statistics that are indicative of healthy lifestyles and openness to causal and entertaining products. By searching for venues of different

categories in our feature set, we are able to differentiate between neighborhoods and identify prime locations of our business. Furthermore, we build predictive machine learning models to gain more accurate insights to verify our findings.