The Battle of Neighborhoods of two cities

Coursera Capstone

IBM Data Science Professional Certificate

By: Rajeev Kavety

The Battle of Neighborhoods of two cities

1. Introduction

For this project, I have chosen to compare the precedence of food courts among the most popular venue categories in both Toronto, CA and Manhattan, NY. As people like to visit most places that have multiple shopping choices; with niche visitors preferring to visit the location that has food courts as well, and also for the start-up entrepreneurs who are looking to start a business where there is a lucrative opportunity in places that have high traffic of shoppers.

2. Problem

The problem with the top 10 most common venues is that, it misses the niche market, such as, food courts. This information can be useful for the entrepreneurs who want to setup a business in the top most venues, where possibility of running a business can have good return-of-investment (ROI), or is profitable. Hence, the challenges of identifying these locations, and overlaying the top most common venues having food courts on the maps of Toronto and Manhattan, can address both preferences of the public.

3. Data

For the Toronto neighborhoods, using scrapping techniques demonstrated by beautiful soup4, and using the python results package, the neighborhood data was extracted from the Wikipedia link. For Toronto map coordinates, the information in a CVS file was downloaded from the link http://cocl.us/Geospatial_data , and was merged with the neighborhoods DataFrame.

For the Manhattan neighborhoods dataset, the link is in the json format. https://geo.nyu.edu/catalog/nyu_2451_34572.

Foursquare API

For the purpose of getting the venues information, Foursquare API can be used. Foursquare API is a location data provider that uses RESTful API calls to retrieve data about most common venue categories in each neighborhood. A Foursquare API GET request is sent in order to extract the surrounding venues that are within a radius of 500m. The 'results' explored using Foursquare venue category is in hierarchy. The venues retrieved for all the neighborhoods are categorized into top 10 most common venues having various businesses.

4. Methodology

4.1 Exploratory Data Analysis

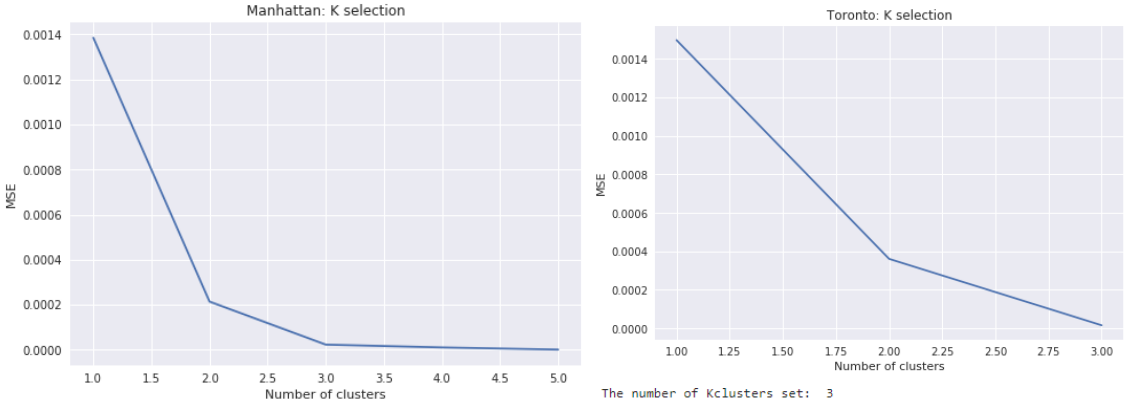
The data downloaded are the neighborhoods located in New York and Toronto. For this project only Manhattan neighborhoods and boroughs are grouped into a DataFrame. For Toronto city, Neighborhoods of North Toronto, Central Toronto, West Toronto, East Toronto, Downtown Toronto, North York, East York, York, and Etobicoke are taken into account. The data is formatted using one hot encoding with the categories of each venue. Then, the venues are grouped for neighborhoods by computing the mean of each feature based on the frequency of occurrence of the categories for the neighborhoods.

Further, to focus on the scope of this project to food courts along with the venues, the food court data was combined to the top 10 venues list.

4.2 Unsupervised Machine Learning

The Battle of Neighborhoods of two cities

Using the k-means clustering algorithm- Elbow method, we can find the number of clusters that can explain the information pertaining to the problem being solved.



The number of the clusters that explains the data for Manhattan food courts are 5, and for Toronto are 3.

5. Results

Using Foursquare API GET request, the data is in the hierarchical format, as shown here.

```
neighborhoods_data[0]
```

```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```

The Battle of Neighborhoods of two cities

After, applying data wrangling and data cleaning techniques, the data is put in a readable format by transforming into a DataFrame, as shown here.

Load and explore the data

Next, let's load the data.

```
searchfor=('Toronto', 'York','Etobicoke')
toronto_data = df_new[df_new['Borough'].astype(str).str.contains('|'.join(searchfor))]

toronto_data= toronto_data.reset_index(drop=True)

toronto_data.head()
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M2H	North York	Hillcrest Village	43.803762	-79.363452
1	M2J	North York	Fairview,Henry Farm,Oriole	43.778517	-79.346556
2	M2K	North York	Bayview Village	43.786947	-79.385975
3	M2L	North York	Silver Hills,York Mills	43.757490	-79.374714
4	M2M	North York	Newtonbrook,Willowdale	43.789053	-79.408493

The data is summarized on the frequency of occurrences of unique venues for each neighborhood. After this step, we need to make sure the category of food courts is available in the venues.

Let's find out how many unique categories can be curated from all the returned venues

```
print('There are {} uniques categories.'.format(len(manhattan_venues['Venue Category'].unique())))
```

```
There are 321 uniques categories.
```

Let's find out if Food court as a unique venue category are available.

```
"Food Court" in manhattan_venues['Venue Category'].unique()
```

```
True
```

As the food courts venue is not among the top 10 venues for most of the neighborhoods, it has to be merged to the DataFrame of top 10 venues.

The Battle of Neighborhoods of two cities

```
toronto_foodcourt=toronto_grouped[['Neighbourhood', 'Food Court']]
toronto_foodcourt.shape
```

```
(83, 2)
```

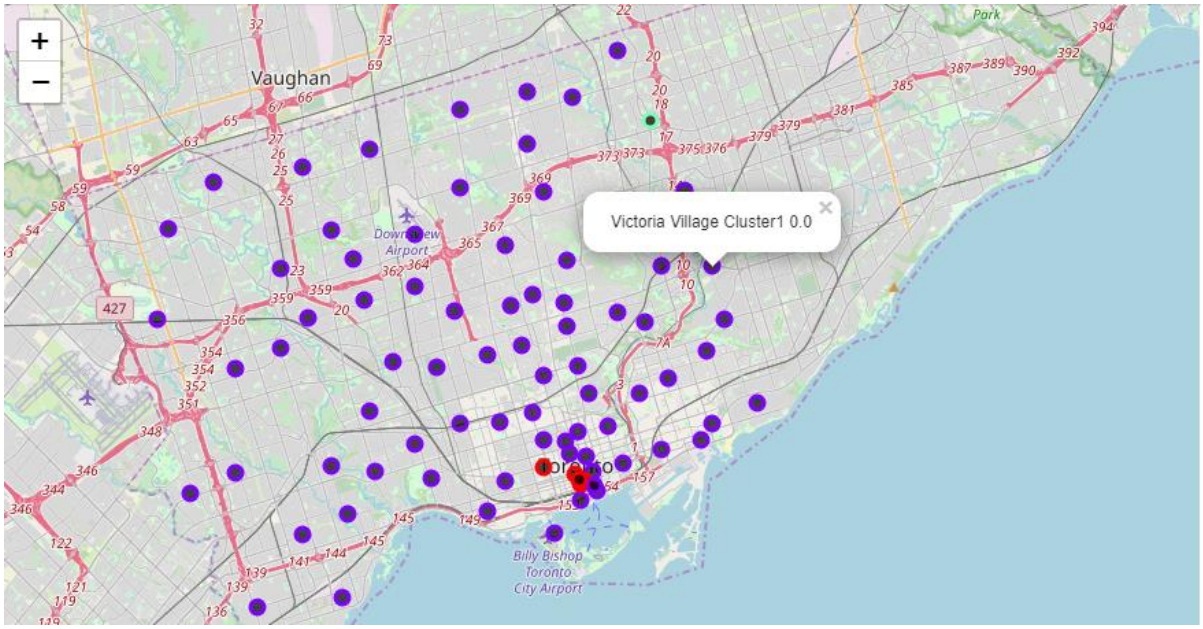
```
toronto_grouped["Food Court"].value_counts()
```

```
0.000000    77
0.010000     4
0.014493     1
0.031250     1
Name: Food Court, dtype: int64
```

After applying the K-means algorithm, the number of clusters that can explains the precedence of food courts in the top 10 venues for Toronto city are three clusters, and for Manhattan are four clusters.

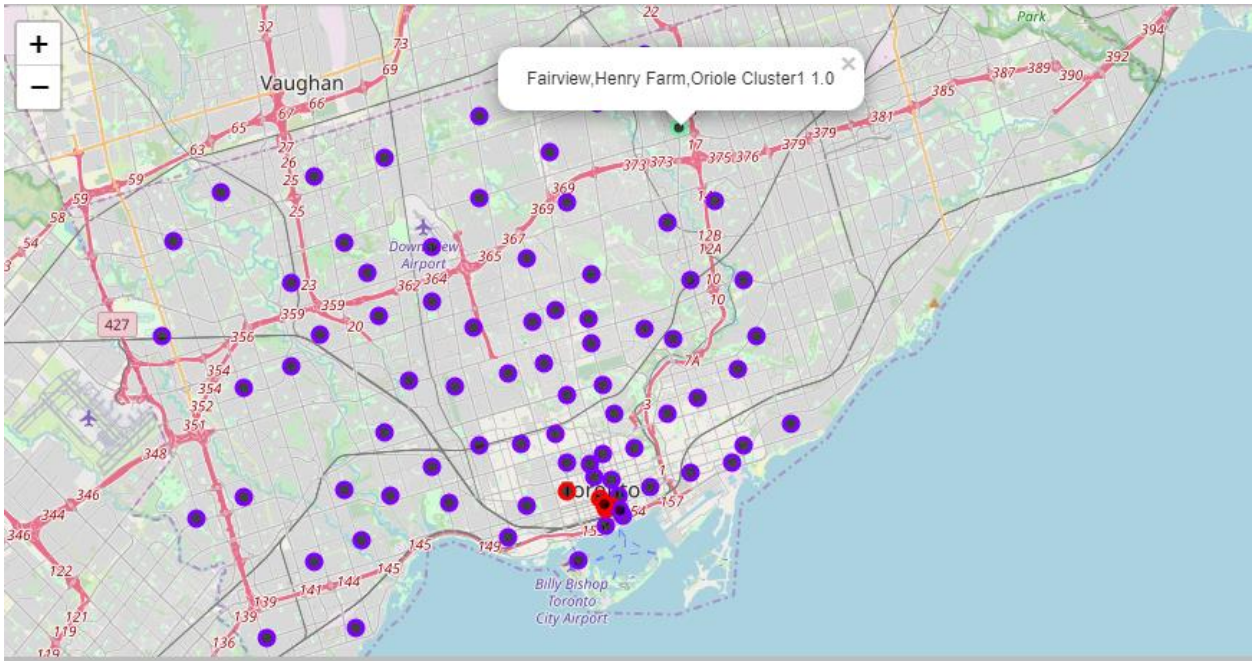
Finally, to visualize the clusters, Folium library is used to generate the maps showing the clusters for the top 10 most common venues, along with focus on food courts for the neighborhoods of both Toronto and Manhattan cities.

Cluster 0- Example of Top venues without the food court-Toronto

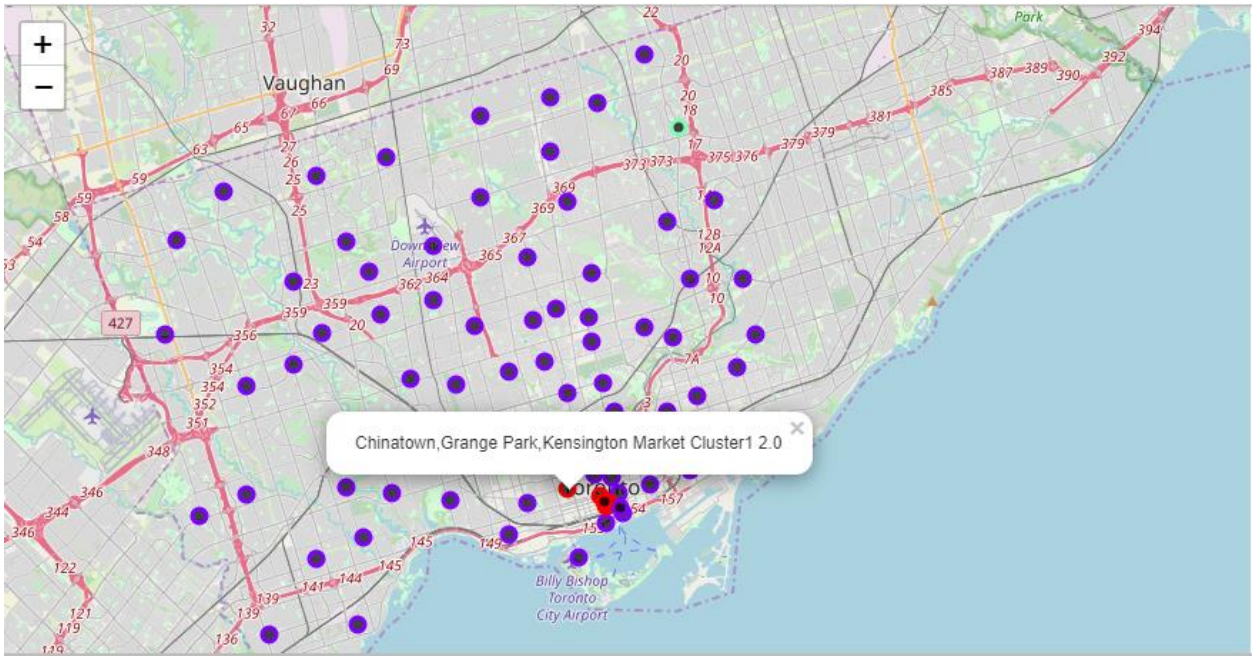


The Battle of Neighborhoods of two cities

Cluster 1- Example Top venues with the food court- Toronto

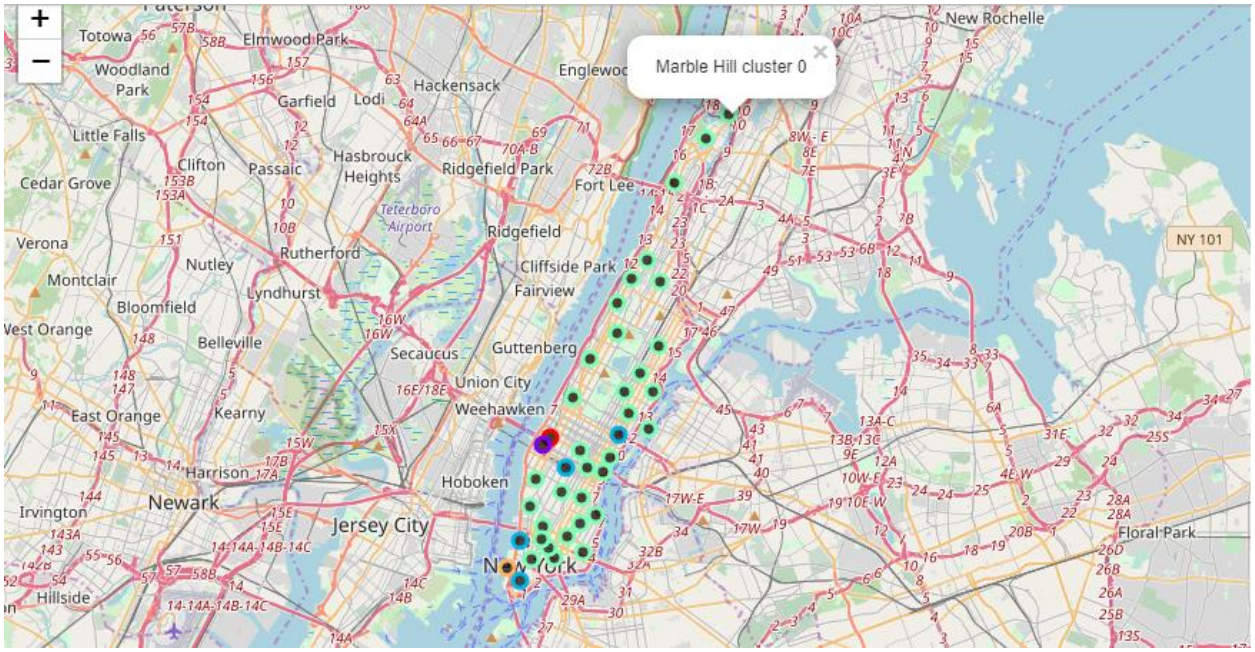


Cluster 2- Example of Top venues with the food court- Toronto

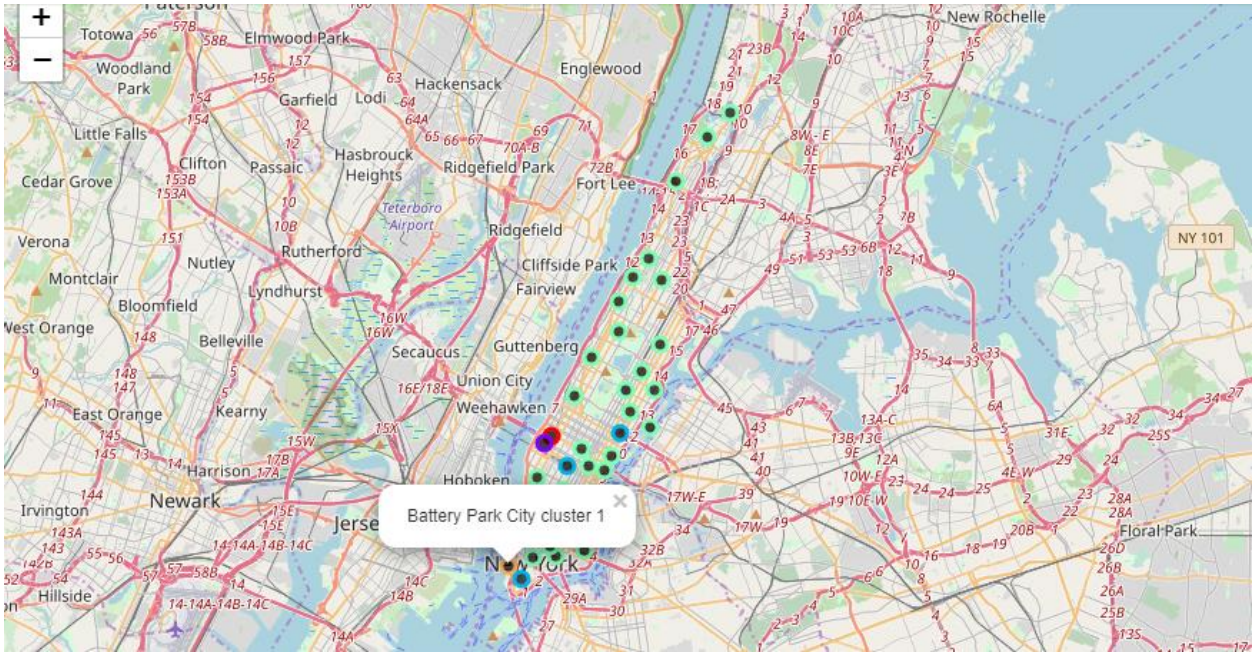


The Battle of Neighborhoods of two cities

Cluster 0- Example of Top venues without the food court-Manhattan

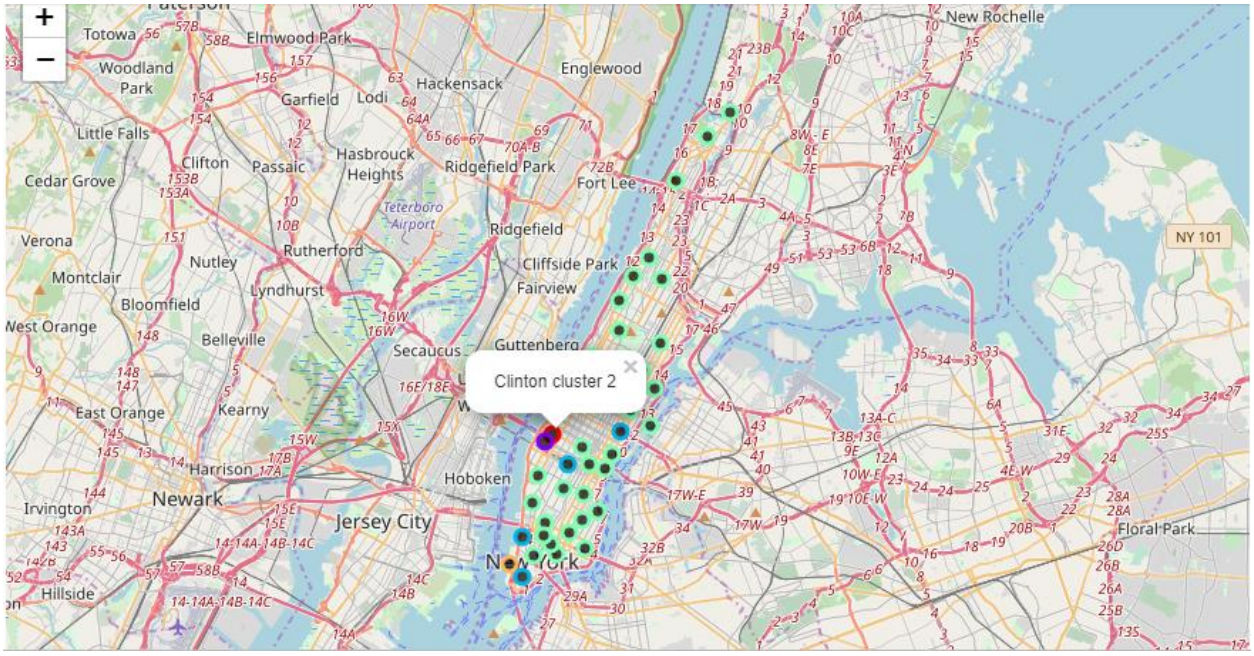


Cluster 1- Example of Top venues with the food court-Manhattan

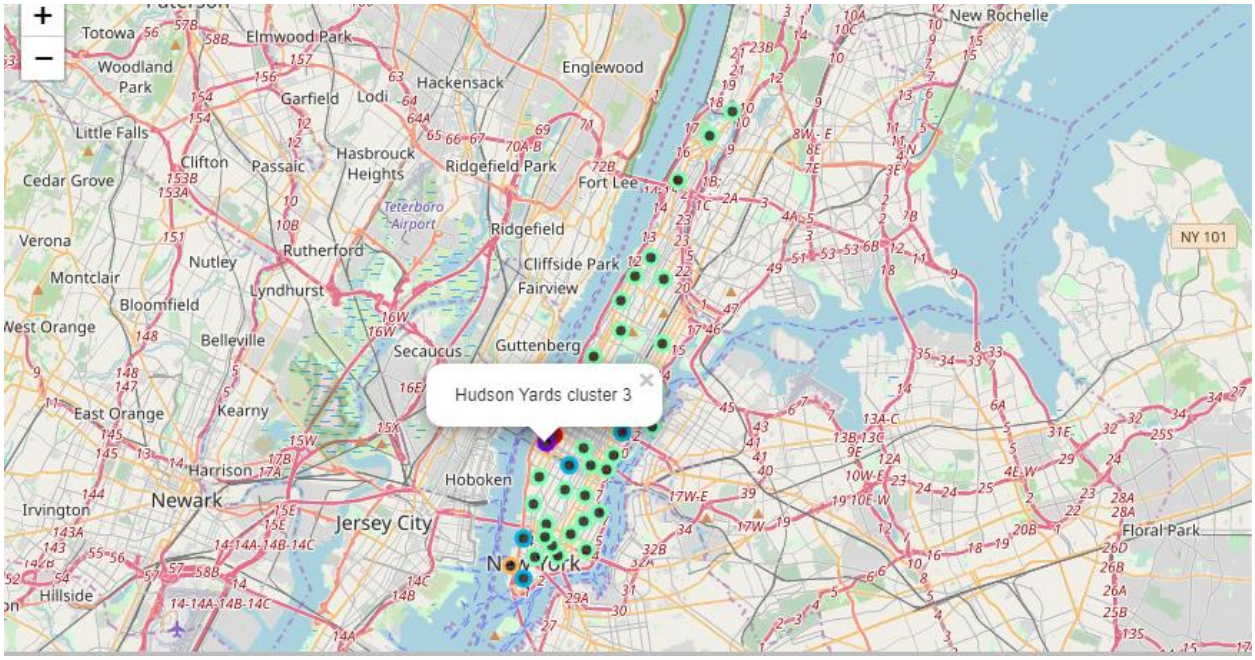


The Battle of Neighborhoods of two cities

Cluster 2- Example of Top venues with the food court-Manhattan

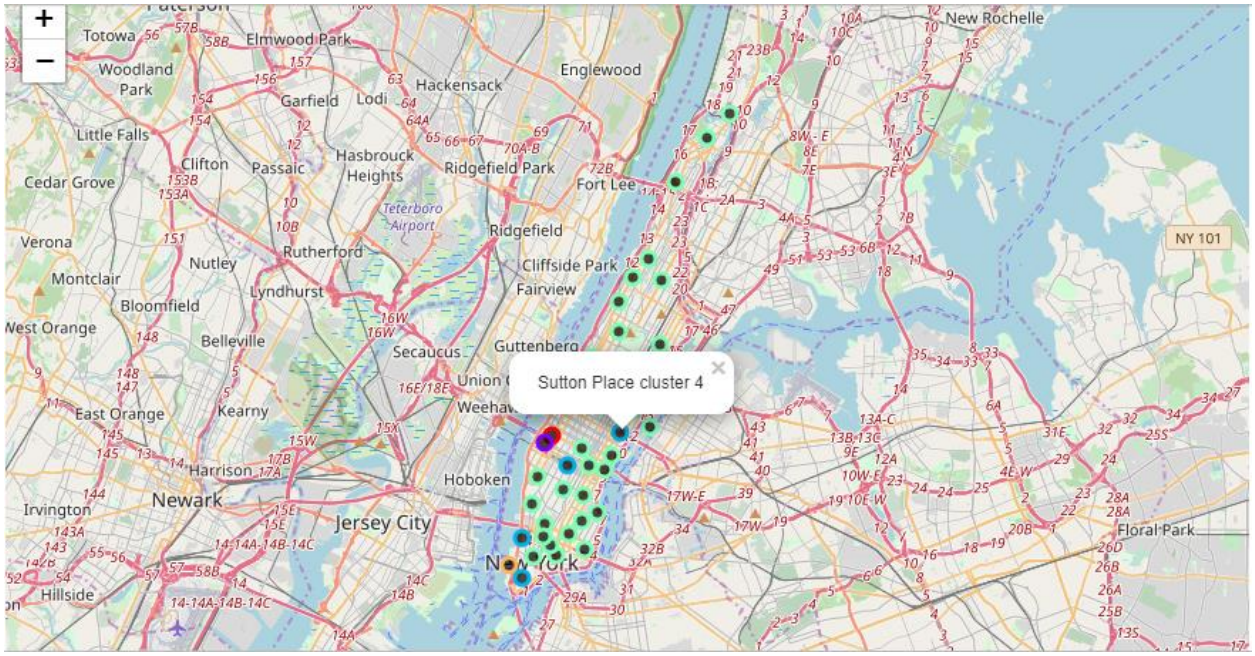


Cluster 3- Example of Top venues with the food court-Manhattan



The Battle of Neighborhoods of two cities

Cluster 4- Example of Top venues with the food court-Manhattan



6. Discussions

From the clusters for Toronto city, it can be deduced that the cluster1 shows most food courts are concentrated at locations that are easily accessible by efficient public transport points, such as, subways linked to bus routes. In comparison to Toronto, Manhattan cities Clusters0-4, show that most of the food courts are located near touristic places also accessed by sight- seeing buses or the trains.

7. Conclusions

The purpose of analyzing the precedence of food courts in some of the top 10 venues, gives an insight for the niche market of food courts. As people like to visit most places that have multiple shopping choices, visitors prefer to visit the location that has food courts as well, and also for gives information to the start-up entrepreneurs for

starting a business. As the speculation of observing the clusters at specific locations in both the cities doesn't include population density, it would be helpful to include population density to provide more information for the entrepreneurs.