# The Battle of Neighborhoods

Clustering and Segmentation using Machine Learning Kmeans

01 **Business Problem**

Problem Definition and requirements

02 **Data Mining & Pre-processing**

Data science Methodologies

03 **Clustering Two Cities**

Modelling location data using Follium maps

04 **Conclusion**

Final Thoughts and Discussion

# Business Problem

# 01

Problem Definition and requirements

# Business Problem

- New York city is America’s hub for business and pleasure, and is one of the most visited cities across the world due to its touristic sites, landmarks and unique venues.

- Toronto on the other hand is the biggest Canadian, and also most visited city in Canada. Its full of Canada’s richest landmarks, locations, and venues.

- The goal of the business case here, is to understand the similarities and differences between 2 cities' venues (Specifically Downtown)

# Business Problem

- The Two cities' downtown neighborhoods will be compared to each other, based on the clusters they fall within. The City of NEW YORK will be compared to the city of TORONTO, to better understand the style of their venues, and how they are similar or dissimilar.

- This business case is aimed towards new business venues owners to allow them to decide on what Neighborhoods are the most suitable for their new venue investment such as; Restaurants, coffee shops or other entertainment location.

# Analytics Approach

- K means clustering will be used to segment the cities' neighborhoods and give an idea of how some Neighborhoods are similar or dissimilar to others, based on the venues' categories that exist in each of these Neighborhoods.

# Data Mining & Pre-processing

# 02

Data science
Methodologies

# Data Collection - Toronto

- The first step was to acquire a Wikipedia page including all neighborhoods of Toronto. Using Pandas method for web scraping – all data was acquired.

- Next step was to pre-process and Filter only valid Boroughs - Removing "Not Assigned" Boroughs. Then I have Included the Latitude and Longitude for each neighborhood.

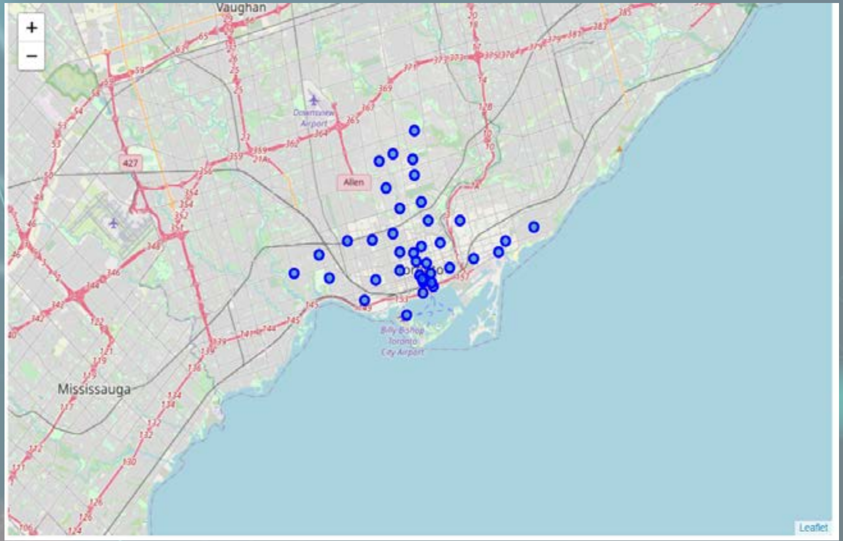
# Data Collection - Toronto

- there are 10 Unique boroughs and 103 Neighborhoods within Toronto.

```
North York          24
Downtown Toronto    19
Scarborough         17
Etobicoke           12
Central Toronto      9
West Toronto         6
York                 5
East York            5
East Toronto         5
Mississauga          1
Name: Borough, dtype: int64
```

# Data Collection - Toronto

- Follium maps are a great way to visualize location data. Using all of my data variables including Latitude, Longitude, Boroughs, and Neighborhoods, and filtering only Downtown Boroughs in Toronto results in the following data location points:



# Data Collection - Toronto

- Foursquare API database contains all the venues for different cities including ratings, trending location, and reviews etc.

- A function was created that would loop over each neighborhood, and retrieve a list of venues, with a limit of 100 venues per each neighborhood. The resulting data frame included each neighborhood and all their associated venues, including the venue name, Lat and Long coordinates, and the Venue Category,

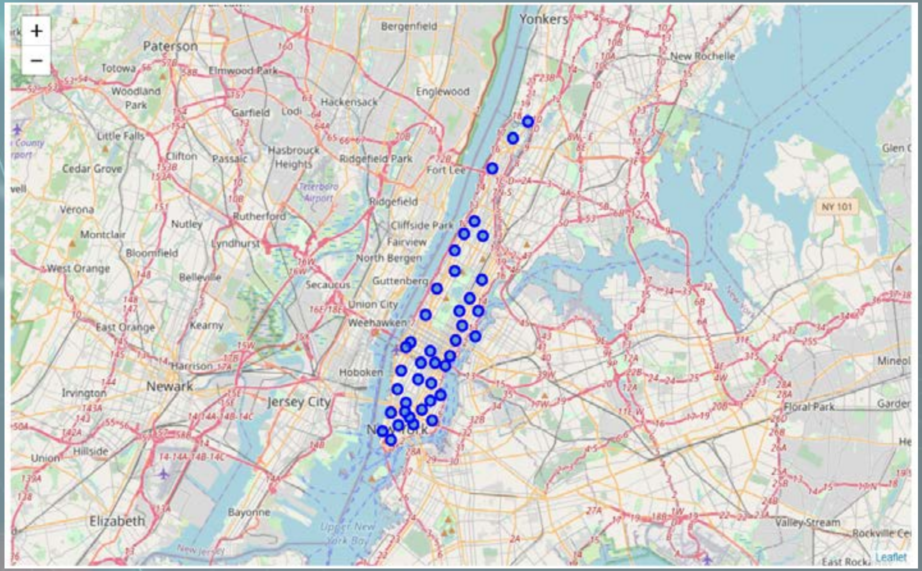
# Data Collection – New York Manhattan

- The next objective was to acquire IBM JSON file for New York city data, including all the city information. Since it’s a JSON file, the data had to be converted into a data frame.
- There are 5 boroughs and 306 neighborhoods in New York.

```
Queens           81
Brooklyn         70
Staten Island    63
Bronx            52
Manhattan        40
Name: Borough, dtype: int64
```

# Data Collection – New York Manhattan

- To decrease the size of the data frame and thus the cluster of neighborhoods, I filtered by Manhattan as the only borough to focus on, and viewed Manhattan on Follium maps.



# Data Collection – New York Manhattan

- A similar function to Toronto, was created to acquire all venues for each neighborhood in Manhattan, just like I did with Toronto. The resulting Data Frame for New York contained each neighborhood and their associated venues names, and categories, only for neighborhoods within Manhattan.

- The last stage of data collection was to merge the two data frames of the two cities vertically (New York & Toronto).

# Data Understanding & Preprocessing

- Since now we have a data frame containing two cities and their downtown boroughs, neighborhoods, venues names, and venues categories. The objective is to think of a metric that can be used as a clustering feature.

- The most suitable feature in which neighborhoods can be clustered according to, is the Venue Category. The Venue Category describes the category of each venue, which is associated with a neighborhood and a location data, that can show exactly clusters of similar venues.

# Data Understanding & Preprocessing

- Dealing with Venue Category feature was to create a one hot encoded data frame, which displays data horizontally and provides a 1 or 0 when the specific venue category matches the neighborhood in which it lies in.

| | Neighborhood | Accessories Store | Adult Boutique | Afghan Restaurant | African Restaurant | Airport | Airport Food Court | Airport Gate | Airport Lounge | Airport Service | Airport Terminal | American Restaurant | Antique Shop | Aquarium | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Battery Park City | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.0000 | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.0000 | 0.000000 | 0.000000 | 0.000000 | |
| 1 | Berczy Park | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.0000 | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.0000 | 0.000000 | 0.000000 | 0.000000 | |
| 2 | Brockton, Parkdale Village, Exhibition Place | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.0000 | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.0000 | 0.000000 | 0.000000 | 0.000000 | |
| 3 | Business reply mail Processing Centre, South C... | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.0000 | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.0000 | 0.000000 | 0.000000 | 0.000000 | |
| 4 | CN Tower, King and Spadina, Railway Lands, Har... | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.0625 | 0.0625 | 0.0625 | 0.125 | 0.1875 | 0.0625 | 0.000000 | 0.000000 | 0.000000 | |
| 5 | Carnegie Hill | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.0000 | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.0000 | 0.000000 | 0.000000 | 0.000000 | |
| 6 | Central Bay Street | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.0000 | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.0000 | 0.000000 | 0.000000 | 0.000000 | |
| 7 | Central Harlem | 0.000000 | 0.00 | 0.000000 | 0.068182 | 0.0000 | 0.0000 | 0.0000 | 0.000 | 0.0000 | 0.0000 | 0.045455 | 0.000000 | 0.000000 | |

# Data Understanding & Preprocessing

- Following the same logic and displaying the most occurrent venue category for each neighborhood, a function was created to print each neighborhood along with the top 5 most common venues

```
----Battery Park City----
          venue  freq
0          Park  0.09
1         Hotel  0.07
2           Gym  0.06
3   Coffee Shop  0.06
4 Boat or Ferry  0.04


----Berczy Park----
           venue  freq
0    Coffee Shop  0.09
1 Farmers Market  0.04
2         Bakery  0.04
3           Café  0.04
4    Cheese Shop  0.04


----Brockton, Parkdale Village, Exhibition Place----
               venue  freq
0               Café  0.14
1     Breakfast Spot  0.09
2        Coffee Shop  0.09
3  Convenience Store  0.05
4         Restaurant  0.05
```

# Data Understanding & Preprocessing

- The final pre-processioning stage included creating a data frame containing the 10 most common venues for each neighborhood, but displayed horizontally

| | 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** | Battery Park City | Park | Hotel | Gym | Coffee Shop | Boat or Ferry | Memorial Site | Playground | Shopping Mall | Sandwich Place | Gourmet Shop |
| **1** | Berczy Park | Coffee Shop | Seafood Restaurant | Bakery | Café | Farmers Market | Cocktail Bar | Cheese Shop | Beer Bar | Restaurant | French Restaurant |
| **2** | Brockton, Parkdale Village, Exhibition Place | Café | Coffee Shop | Breakfast Spot | Convenience Store | Gym | Restaurant | Italian Restaurant | Performing Arts Venue | Nightclub | Intersection |
| **3** | Business reply mail Processing Centre, South C... | Yoga Studio | Gym / Fitness Center | Comic Shop | Restaurant | Park | Skate Park | Smoke Shop | Burrito Place | Brewery | Farmers Market |
| **4** | CN Tower, King and Spadina, Railway Lands, Har... | Airport Service | Airport Lounge | Boutique | Sculpture Garden | Coffee Shop | Harbor / Marina | Rental Car Location | Boat or Ferry | Airport Terminal | Bar |

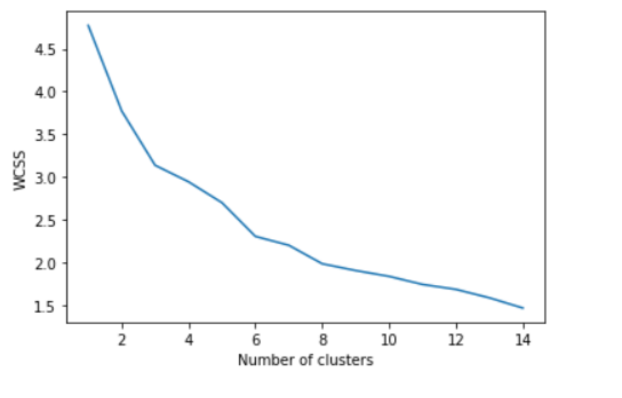
# Clustering Two Cities

# 03

Modelling location data
using Follium maps

# Data Modelling (K means Clustering)

- I have decided to utilize the use of the K means Clustering model from SKLEARN.
- First and using the Elbow Method, I wanted to know the suitable number of Clusters - using "Within Cluster Sum of Squares" WCSS
- I have decided to cluster based on 6 clusters, using an init of K means++, random state of 1, and a convergence number of 12 iterations.



# Data Modelling (K means Clustering)

- The resulted cluster labels from the trained model, can now be added to their associated neighborhoods in the last data frame we had. Now the data frame has all information about both cities' boroughs, Neighborhoods, venues, and the number of clusters their neighborhoods are associated to.

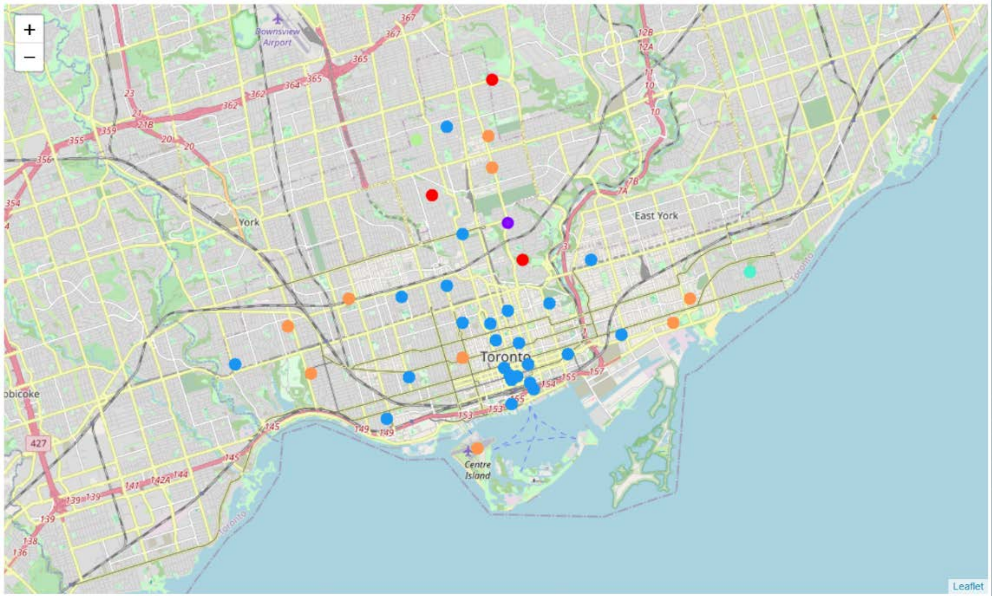
| | City | Borough | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | TORONTO | Downtown Toronto | Regent Park, Harbourfront | 43.65426 | -79.360636 | Roselle Desserts | 43.653447 | -79.362017 | Bakery | 0 | Coffee Shop | Pub | Bakery |
| 1 | TORONTO | Downtown Toronto | Regent Park, Harbourfront | 43.65426 | -79.360636 | Tandem Coffee | 43.653559 | -79.361809 | Coffee Shop | 0 | Coffee Shop | Pub | Bakery |
| 2 | TORONTO | Downtown Toronto | Regent Park, Harbourfront | 43.65426 | -79.360636 | Cooper Koo Family YMCA | 43.653249 | -79.358008 | Distribution Center | 0 | Coffee Shop | Pub | Bakery |
| 3 | TORONTO | Downtown Toronto | Regent Park, Harbourfront | 43.65426 | -79.360636 | Impact Kitchen | 43.656369 | -79.356980 | Restaurant | 0 | Coffee Shop | Pub | Bakery |
| 4 | TORONTO | Downtown Toronto | Regent Park, Harbourfront | 43.65426 | -79.360636 | Body Blitz Spa East | 43.654735 | -79.359874 | Spa | 0 | Coffee Shop | Pub | Bakery |

# Data Modelling (K means Clustering)

- Since the data frame has the two cities in 1 data frame, it is possible to see all of the clusters at once, however since Toronto is far from New York, Follium maps will have to be centered in between the two cities, and the clusters will not be properly display. So, I have decided to view the same map from Toronto Downtown point of view, and then view the same map from New York Manhattan point of view

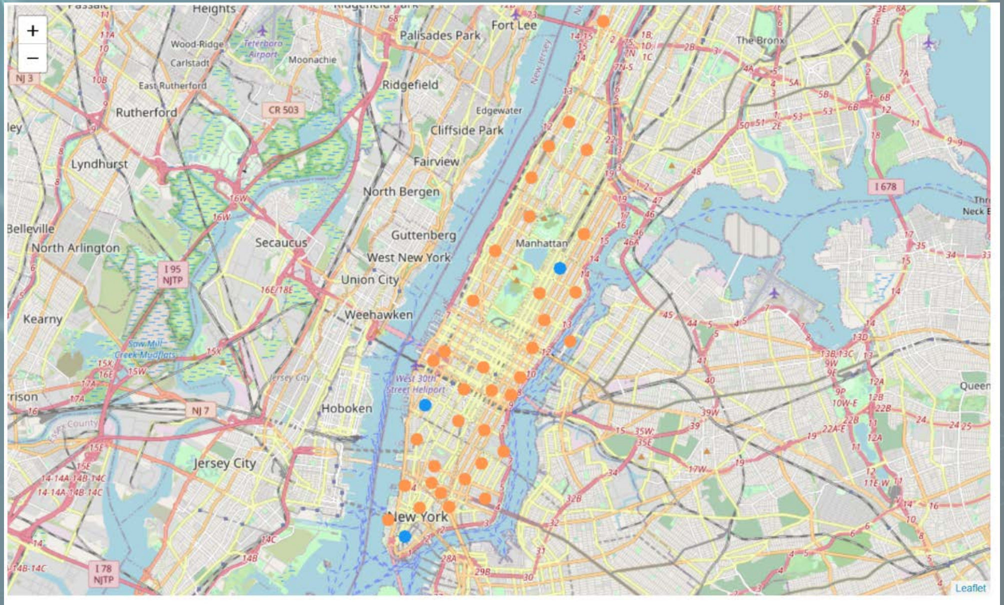
# Data Modelling (K means Clustering)

- Toronto Clusters



# Data Modelling (K means Clustering)

- Manhattan Clusters



# Conclusion

# 04

Final Thoughts and Discussion

# Conclusion

- The two cities data have been combined into 1 data frame - Containing each city Boroughs, Neighborhoods, and Venues - With the focus on downtown boroughs from each city (Downtown NY Manhattan VS Downtown Toronto)

- After Segmenting the entire data frame containing the two cities downtown regions - we can see that Downtown New York neighborhoods all fall within 2 clusters (2 and 5) whereas most of Toronto downtown neighborhoods/venues fall within the same clusters, and the rest of the neighborhoods scattered around downtown Toronto, are within the other clusters

# Conclusion

### Toronto Clusters

```
toronto_clusters = two_cities_merged_with_onehot[two_c
toronto_clusters.head(5)
```

| Cluster Labels | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **Accessories Store** | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 |
| **Adult Boutique** | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 |
| **Afghan Restaurant** | 0.0 | 0.0 | 0.000718 | 0.0 | 0.0 | 0.000000 |
| **African Restaurant** | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 |
| **Airport** | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.004545 |

### New York Clusters

```
ny_clusters = two_cities_merged_with_onehot[two_cities
ny_clusters.head(5)
```

| Cluster Labels | 2 | 5 |
|---|---|---|
| **Accessories Store** | 0.0 | 0.001026 |
| **Adult Boutique** | 0.0 | 0.000342 |
| **Afghan Restaurant** | 0.0 | 0.000000 |
| **African Restaurant** | 0.0 | 0.001026 |
| **Airport** | 0.0 | 0.000000 |

### 3) TORONTO VENUES CATEGORIES IN CLUSTER 5

```
toronto_clusters.iloc[:,5].sort_values(ascending=False).head(10)
```

```
Bar                   0.040909
Café                  0.040909
Coffee Shop           0.040909
Park                  0.040909
Pizza Place           0.031818
Dessert Shop          0.027273
Mexican Restaurant    0.027273
Bakery                0.022727
Brewery               0.022727
Italian Restaurant    0.022727
Name: 5, dtype: float64
```

### 4) NEWYORK VENUES CATEGORIES IN CLUSTER 5

```
ny_clusters.iloc[:,1].sort_values(ascending=False).head(10)
```

```
Italian Restaurant     0.041368
Coffee Shop            0.037949
Café                   0.026325
Bakery                 0.023590
American Restaurant    0.023590
Pizza Place            0.023248
Park                   0.021197
Hotel                  0.019145
Bar                    0.017778
Mexican Restaurant     0.017778
Name: 5, dtype: float64
```

# Conclusion

- Cluster 2 and 5 has all of New York downtown Neighborhood venues, and most of Toronto Downtown Neighborhood Venues
- Cluster 2 and 5 shows similarity between downtown Network and Downtown Toronto in the Type of venues categories available - With common venues such as:
    - Expensive Italian restaurants
    - Expensive Japanese Restaurants
    - Coffee shops
    - Hotels
    - Parking spots
    - Bars
    - Bakeries

# Conclusion

- For clients looking to invest into any of the above venues, it would be a great idea to invest downtown - with creative venues that offer similar categories but with different flavors - such as expensive but foreign restaurant style venues etc

# THANKS!