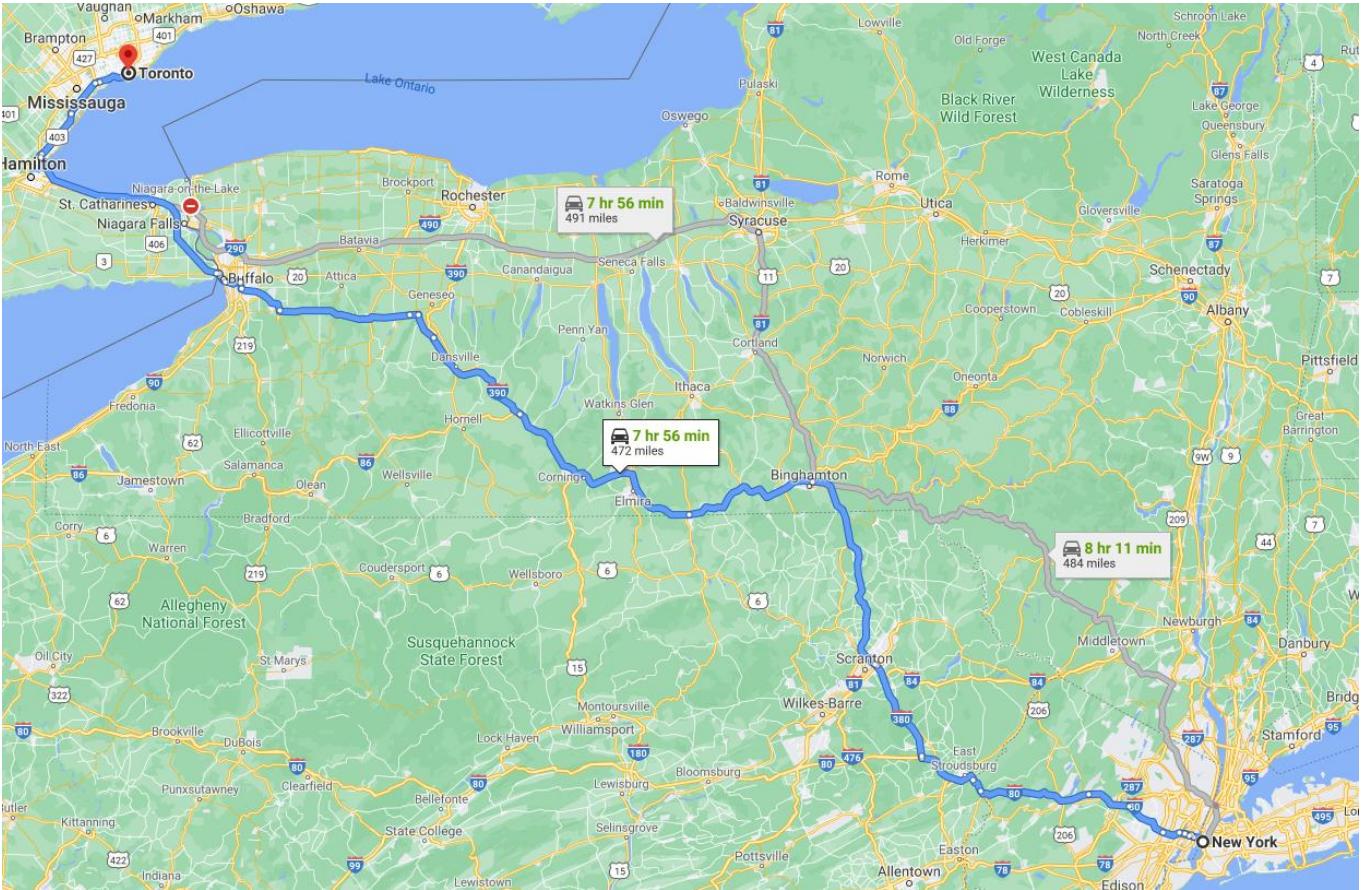


Background and business problem to solve



ABC is a successful house construction company in New York, wants to establish its business in Toronto.
ABC employs a data scientist for 2 question:

- Are the Toronto neighborhoods similar as New York?
- What kind of houses the potential clients in Toronto most want?

Data Source for analysis

New York neighborhood data: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDriverSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

Toronto post code data: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Toronto Geospatial_data: http://cocl.us/Geospatial_data

Four Square geolocation data:



The Statistic Canada Census data: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-telecharger/comp/page_dl-tc.cfm?Lang=E

This data is the 2016 census data. For each Post Code, there are 2247 lines of data to cover different aspects:

0. General (8 lines)
1. Population age distribution (26 lines)
2. Dwelling structure (28 lines)
3. Family structure (41 lines)
4. Knowledge on languages (561 lines)
5. Income (211 lines)
6. Language (263 lines)
7. Citizenship and migration status (482 line)

1619	8. Dwelling situation	Total - Private households by tenure - 25% sample data
1620		Owner
1621		Renter
1622		Band housing
1623		Total - Occupied private dwellings by condominium status - 25% sample data
1624		Condominium
1625		Not condominium
1626		Total - Occupied private dwellings by number of bedrooms - 25% sample data
1627		No bedrooms
1628		1 bedroom
1629		2 bedrooms
1630		3 bedrooms
1631		4 or more bedrooms
1632		Total - Occupied private dwellings by number of rooms - 25% sample data
1633		1 to 4 rooms
1634		5 rooms
1635		6 rooms
1636		7 rooms
1637		8 or more rooms
1638		Average number of rooms per dwelling
1639		Total - Private households by number of persons per room - 25% sample data
1640		One person or fewer per room
1641		More than 1 person per room
1642		Total - Private households by housing suitability - 25% sample data

The Toronto geojson data: Not used

Analysis Steps - I

	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

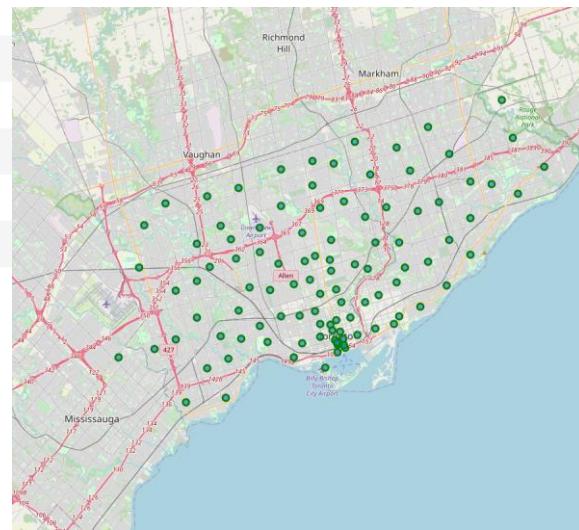
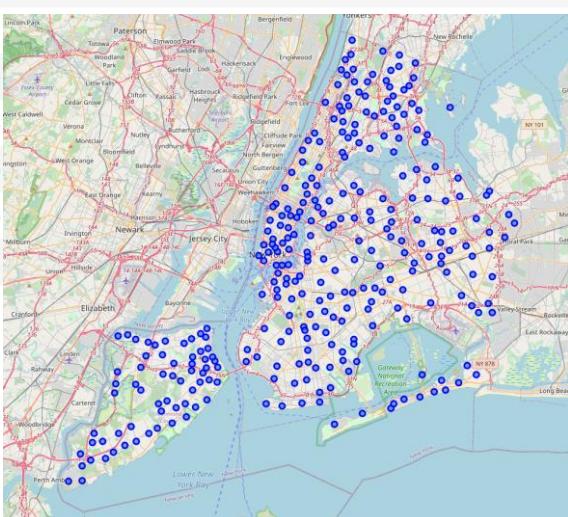
	Postal Code	Borough	Neighbourhood
M3A	North York	Parkwoods	
M4A	North York	Victoria Village	
M5A	Downtown Toronto	Regent Park, Harbourfront	
M6A	North York	Lawrence Manor, Lawrence Heights	
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	

	Borough	Neighborhood	Latitude	Longitude
0	North York	Parkwoods	43.753259	-79.329656
1	North York	Victoria Village	43.725882	-79.315572
2	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Retrieve and combined the data from different tables

Eventually link the key information in one table: Neighborhood – Latitude- Longitude

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
...
2113		43.628841	-79.520999	Islington Florist & Nursery			
2114		43.628841	-79.520999	Koala Tan Tanning Salon & Sunless Spa			
2115		43.628841	-79.520999	Once Upon A Child			
2116		43.628841	-79.520999	Kingsway Boxing Club			
2117		43.628841	-79.520999	Burrito Boyz			



Analysis Steps - II

New York

	Neighborhood	Neighborhood	Latitude	Neighborhood	Longitude	Venue	Venue	Latitude	Venue	Longitude	Venue Category
0	Wakefield		40.894705		-73.847201	Lollipops Gelato		40.894123		-73.845892	Dessert Shop
1	Wakefield		40.894705		-73.847201	Rite Aid		40.896649		-73.844846	Pharmacy
2	Wakefield		40.894705		-73.847201	Walgreens		40.896528		-73.844700	Pharmacy
3	Wakefield		40.894705		-73.847201	Carvel Ice Cream		40.890487		-73.848568	Ice Cream Shop
4	Wakefield		40.894705		-73.847201	Dunkin'		40.890459		-73.849089	Donut Shop

(10045, 7)

There are 440 uniques categories.

Toronto

	Neighborhood	Neighborhood	Latitude	Neighborhood	Longitude	Venue	Venue	Latitude	Venue	Longitude	Venue Category
0	Parkwoods		43.753259		-79.329656	Brookbanks Park		43.751976		-79.332140	Park
1	Parkwoods		43.753259		-79.329656	Variety Store		43.751974		-79.333114	Food & Drink Shop
2	Victoria Village		43.725882		-79.315572	Victoria Village Arena		43.723481		-79.315635	Hockey Arena
3	Victoria Village		43.725882		-79.315572	Portugril		43.725819		-79.312785	Portuguese Restaurant
4	Victoria Village		43.725882		-79.315572	Tim Hortons		43.725517		-79.313103	Coffee Shop

(2112, 7)

There are 265 uniques categories.

Made Four Square queries

Check the most frequent venue categories in each neighborhood

—Allerton—

	venue	freq
0	Pizza Place	0.12
1	Deli / Bodega	0.08
2	Supermarket	0.08
3	Chinese Restaurant	0.08
4	Department Store	0.04

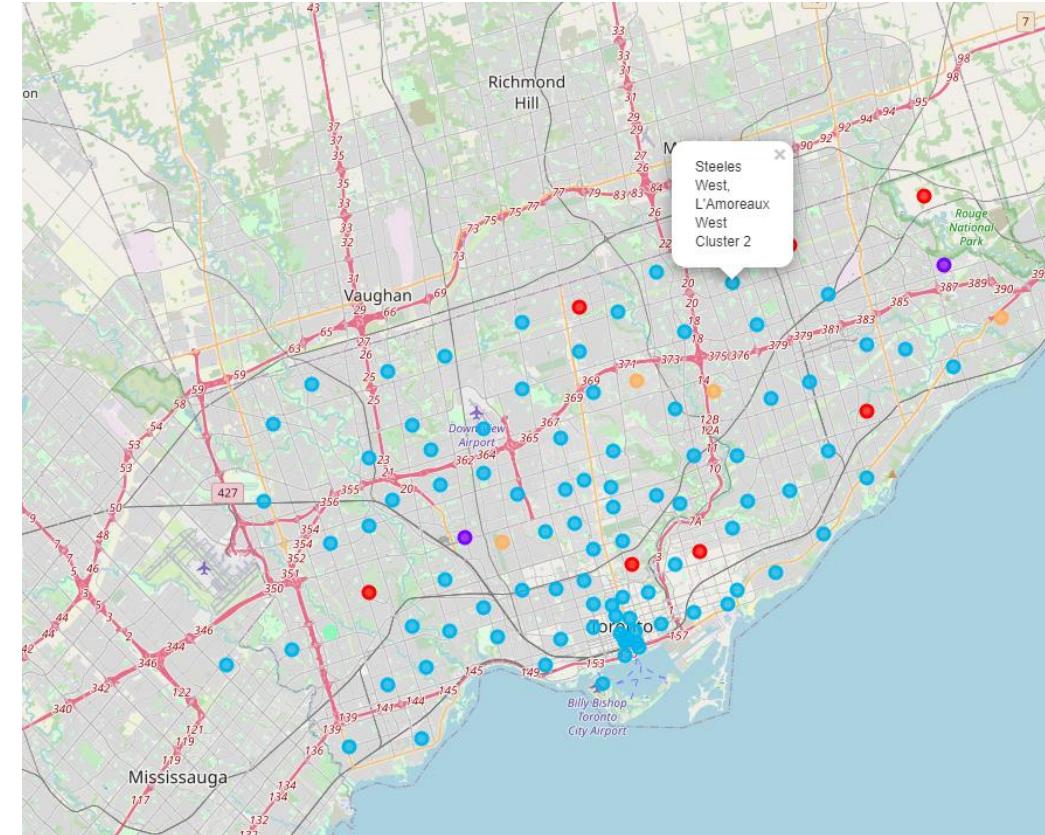
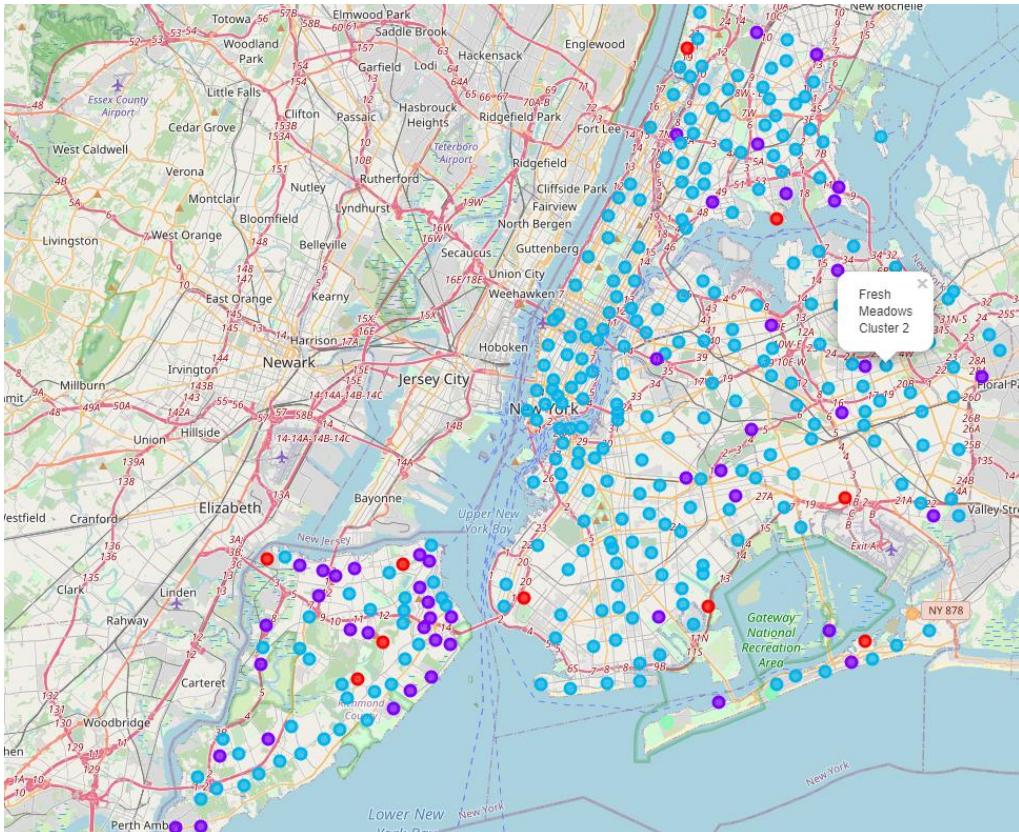
—Annadale—

	venue	freq
0	Liquor Store	0.11
1	Diner	0.11
2	Train Station	0.11
3	Park	0.11
4	Pizza Place	0.11

—Arden Heights—

	venue	freq
0	Pharmacy	0.25
1	Coffee Shop	0.25
2	Bus Stop	0.25
3	Pizza Place	0.25
4	Outlet Store	0.00

Analysis Steps - III



Did K-Means clustering for New York and Toronto separately

Analysis Steps - IV

```
newyork_merged['Cluster Labels'].value_counts()
```

```
In[1]: 2    245  
       1    48  
       0    10  
       3     2  
       4     1  
Name: Cluster Labels, dtype: int64
```

```
[44]: df_cluster = newyork_merged[(newyork_merged['Cluster Labels'] == 1)]  
df_cluster['1st Most Common Venue'].value_counts()
```

```
Out[44]: Pizza Place      37  
Italian Restaurant      23  
Deli / Bodega          22  
Coffee Shop             18  
Chinese Restaurant      13  
..  
Other Nightlife          1  
Mobile Phone Shop        1  
Dessert Shop            1  
Market                  1  
Baseball Field          1  
Name: 1st Most Common Venue, Length: 79, dtype: int64
```

```
[42]: toronto_merged['Cluster Labels'].value_counts()
```

```
Out[42]: 1    87  
       0    12  
       2     2  
       4     1  
       3     1  
Name: Cluster Labels, dtype: int64
```

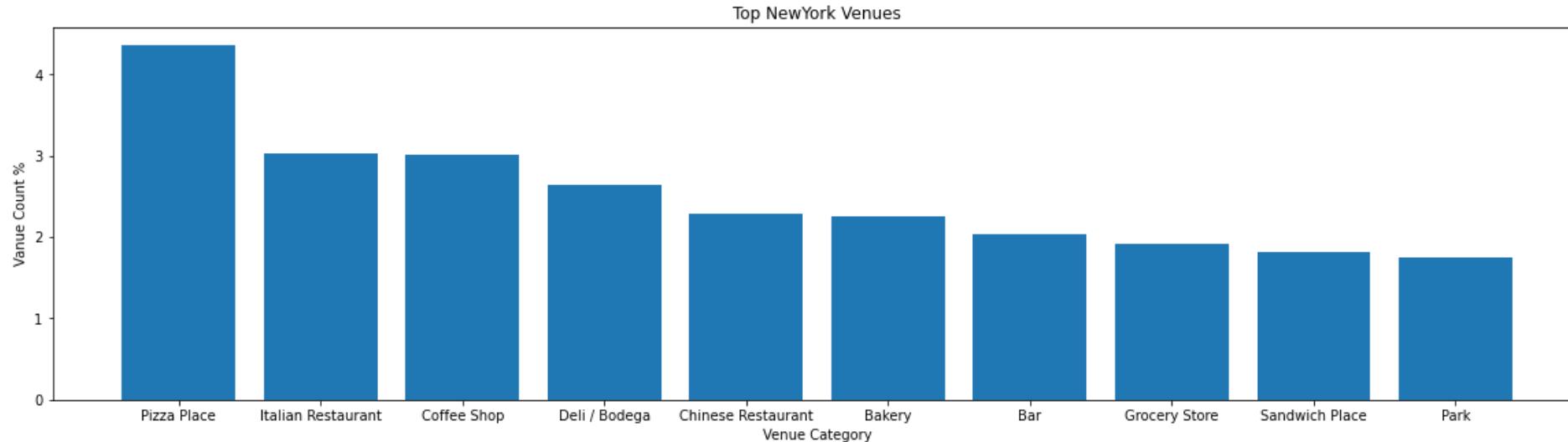
```
[45]: df_cluster = toronto_merged[(toronto_merged['Cluster Labels'] == 1)]  
df_cluster['1st Most Common Venue'].value_counts()
```

```
Out[45]: Coffee Shop      21  
Pizza Place              12  
Café                     7  
Grocery Store            6  
Pharmacy                 3  
Clothing Store           3  
Gym                      2  
  
Furniture / Home Store   2  
Bakery                   1  
Yoga Studio               1  
Vietnamese Restaurant     1  
Airport Service           1  
Name: 1st Most Common Venue, dtype: int64
```

Checked the most popular clusters in New York and Toronto

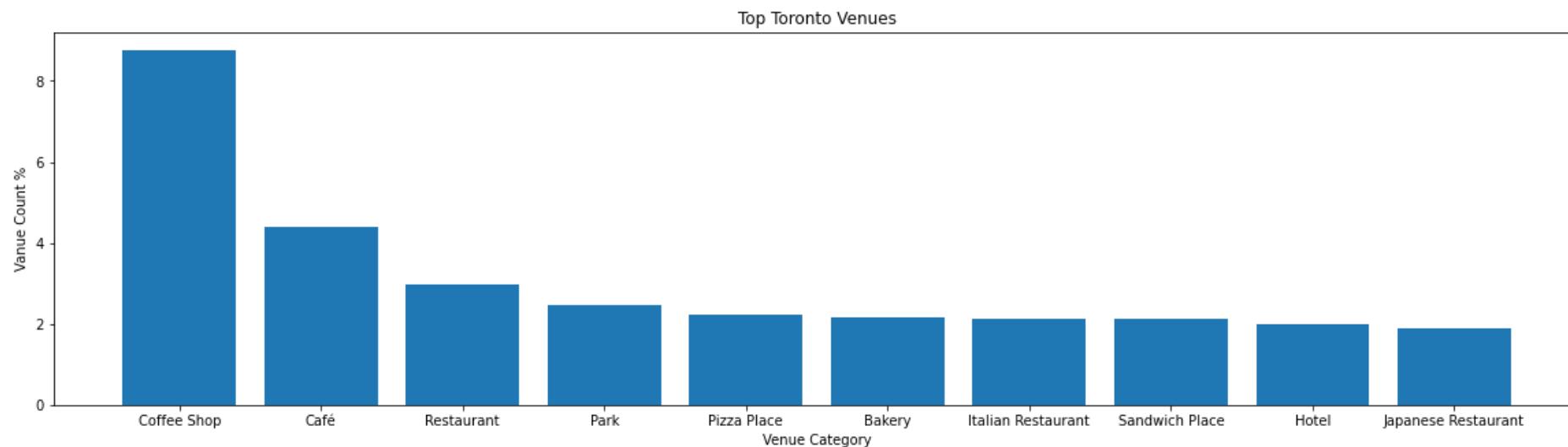
Check the most popular venue categories in top1 cluster

Analysis Steps - V



Listed the top 10 most frequent venue types in New York and Toronto

This tells what venues are more common in New York, what venues are more common in Toronto



Analysis Steps - VI

Combined the Post Code – Neighborhood – Latitude – Longitude data with Statistics Canada data

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Average Rooms	Average Value
0	M3A	North York	Parkwoods	43.753259	-79.329656	5.3	786733.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	4.7	560401.0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	3.5	573259.0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	4.6	644259.0
4	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242	5.6	1089850.0
...
91	M4X	Downtown Toronto	St. James Town, Cabbagetown	43.667967	-79.367675	3.4	873003.0
92	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944	6.4	1192475.0
93	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	3.2	501891.0
94	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509	5.0	767225.0
95	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	6.2	762796.0

Analysis Steps - VII

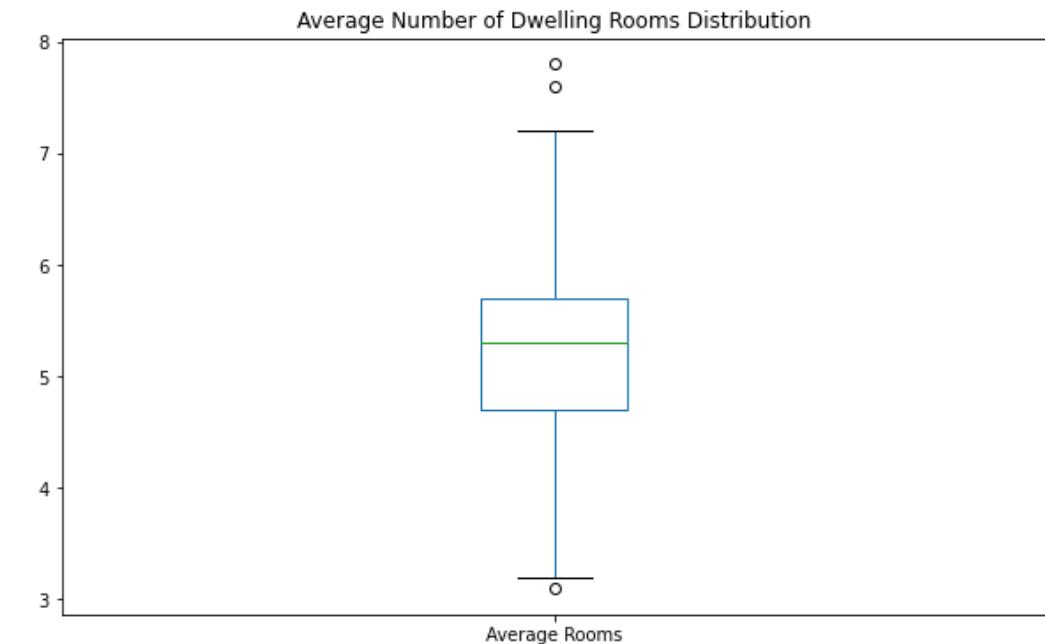
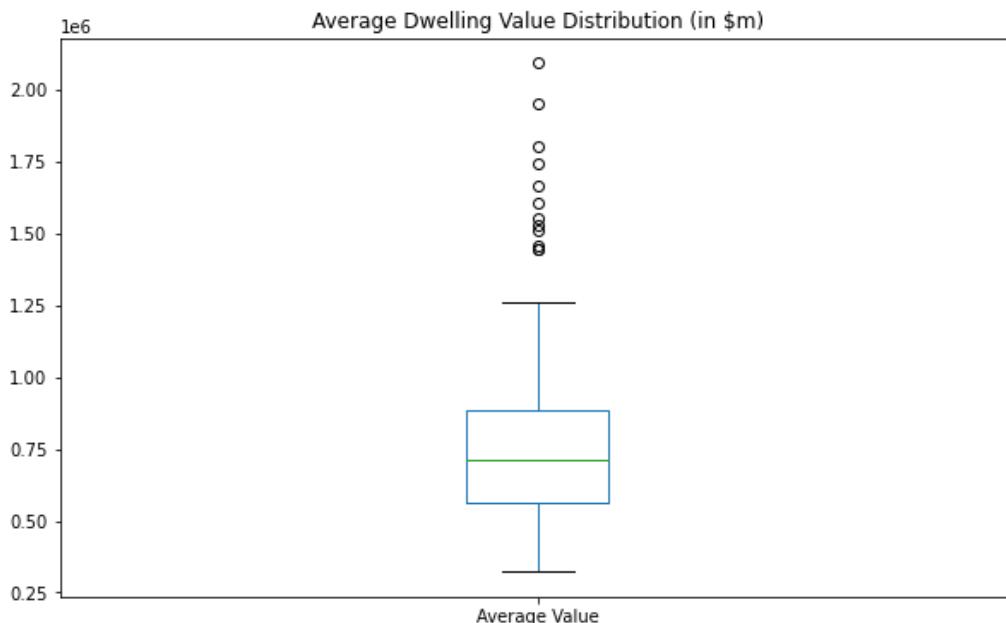
```
[32]: tr_data['Average Value'].describe()
```

```
Out[32]: count    9.600000e+01
mean     8.069142e+05
std      3.701061e+05
min     3.245700e+05
25%     5.597500e+05
50%     7.121075e+05
75%     8.859228e+05
max     2.090328e+06
Name: Average Value, dtype: float64
```

```
[33]: tr_data['Average Rooms'].describe()
```

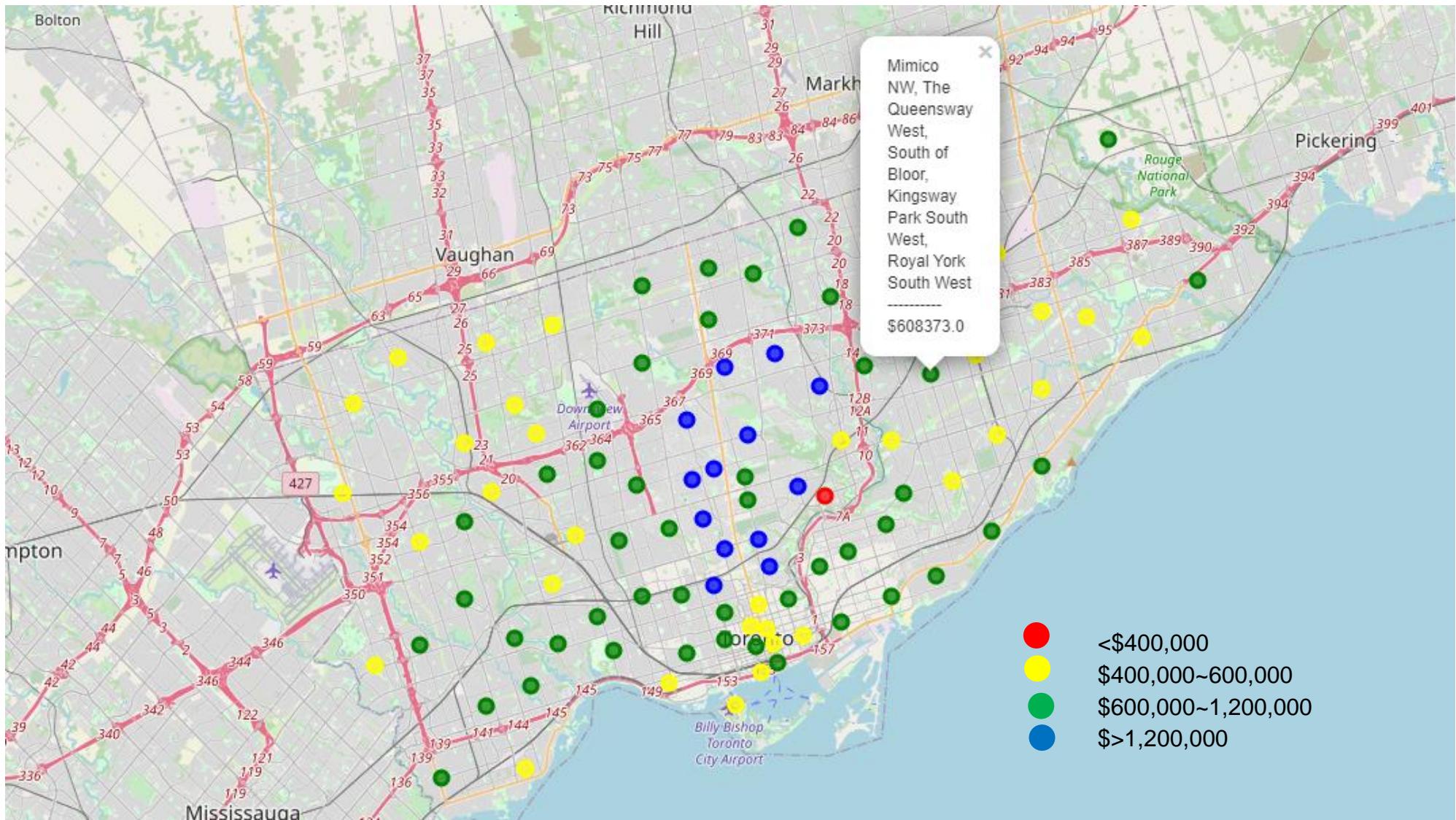
```
Out[33]: count    96.000000
mean     5.167708
std      1.001525
min     3.100000
25%     4.700000
50%     5.300000
75%     5.700000
max     7.800000
Name: Average Rooms, dtype: float64
```

Get the statistics on “Average Dwelling Value” and “Average number of Rooms”



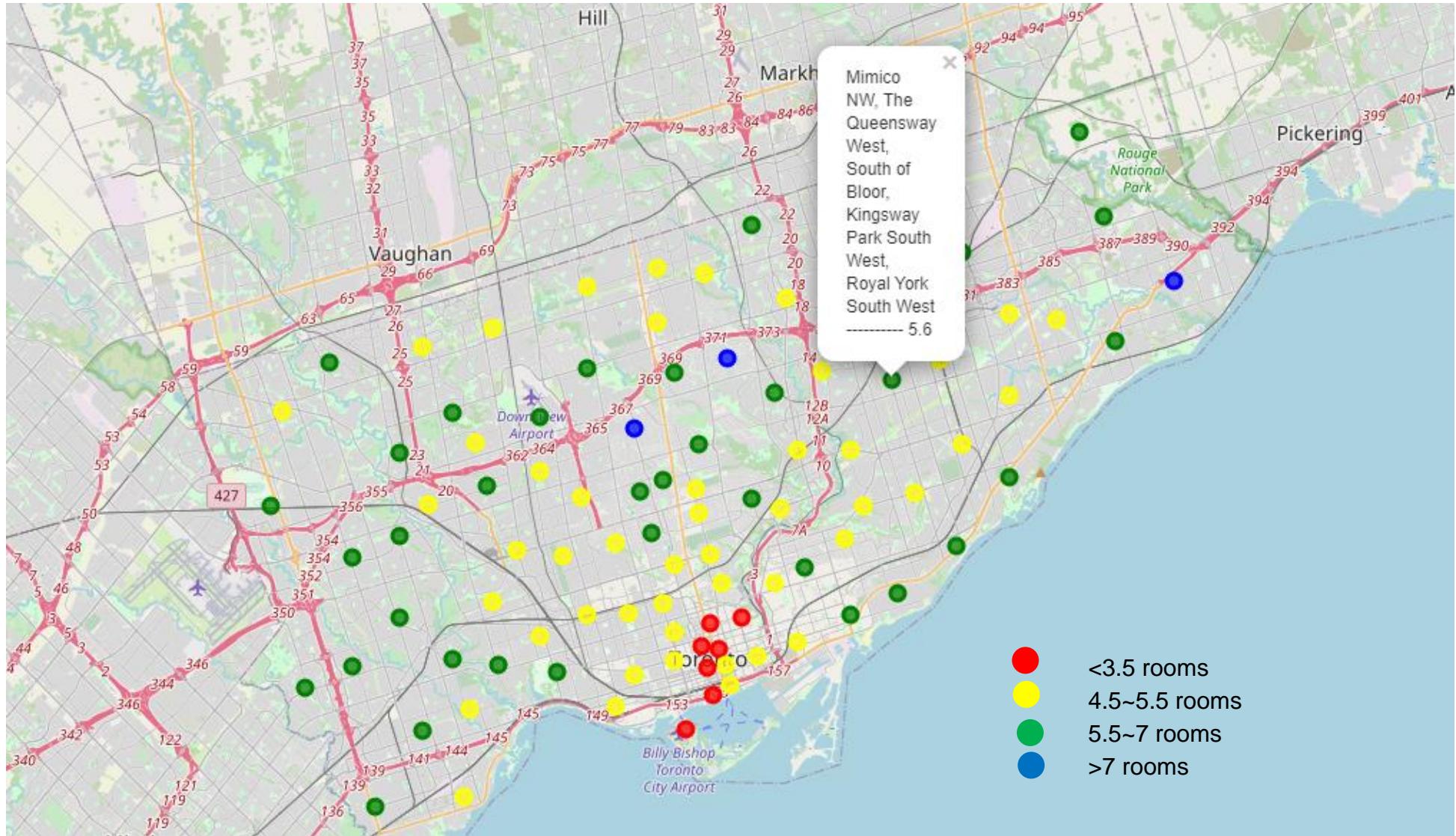
Analysis Steps - VIII

Illustrate “Average Value” on the map



Analysis Steps - IX

Illustrate “Average Rooms” on the map



Conclusion and observation

In general, we would give the information to ABC company:

- Toronto is a city with similarity to New York, but with different combinations of venues in neighborhood (e.g. more coffee shops, parks)
- We should focus on the high value areas with smaller number of rooms per dwelling
- There are many other possibilities we can do with the venue data (different queries and comments) and Statistic Canada data (2247 lines per post code).
- Most importantly, this report is to help ABC company to understand the power of data science and what potential data scientist can unlock for them to make business in Toronto.
- There are many more data sources, e.g. paid subscription to Statistics Canada for many more data combination.

Limitation for this analysis:

- Ideally, it is better to use map.choropleth to use the depth of color to reflect the "Average Dwelling Value" and "Number of Rooms". But there is a limitation to manipulate the Geojson file for Toronto. The version I found in internet, its breakdown does not match post code well and there was some problem to fetch the streambody object. This can be done later
- The Statistic Canada data is 2016, maybe too old to reflect the current situation

Conclusion:

This report and analysis give the brief to ABC company about Toronto neighborhood, what kind of houses they can build in Toronto for what price.

It is helpful to ABC company to setup their business in Toronto. And when business setup and operation goes on, we can help to discover and unlock more possibilities and potentials.


```

{'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]}}

```

	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

Postal Code	Borough	Neighbourhood
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

	Borough	Neighborhood	Latitude	Longitude
0	North York	Parkwoods	43.753259	-79.329656
1	North York	Victoria Village	43.725882	-79.315572
2	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
...
2113	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Islington Florist & Nursery	43.630156	-79.518718	Flower Shop
2114	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Koala Tan Tanning Salon & Sunless Spa	43.631370	-79.519006	Tanning Salon
2115	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Once Upon A Child	43.631075	-79.518290	Kids Store
2116	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Kingsway Boxing Club	43.627254	-79.526684	Gym
2117	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Burrito Boyz	43.626657	-79.526349	Burrito Place

1619	8. Dwelling situation	Total - Private households by tenure - 25% sample data
1620		Owner
1621		Renter
1622		Band housing
1623		Total - Occupied private dwellings by condominium status - 25% sample data
1624		Condominium
1625		Not condominium
1626		Total - Occupied private dwellings by number of bedrooms - 25% sample data
1627		No bedrooms
1628		1 bedroom
1629		2 bedrooms
1630		3 bedrooms
1631		4 or more bedrooms
1632		Total - Occupied private dwellings by number of rooms - 25% sample data
1633		1 to 4 rooms
1634		5 rooms
1635		6 rooms
1636		7 rooms
1637		8 or more rooms
1638		Average number of rooms per dwelling
1639		Total - Private households by number of persons per room - 25% sample data
1640		One person or fewer per room
1641		More than 1 person per room
1642		Total - Private households by housing suitability - 25% sample data

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allerton	Pizza Place	Bus Station	Supermarket	Chinese Restaurant
1	Annadale	Pizza Place	Pharmacy	Train Station	Liquor Store
2	Arden Heights	Pizza Place	Deli / Bodega	Pharmacy	Coffee Shop
3	Arlington	Boat or Ferry	Deli / Bodega	Intersection	American Restaurant
4	Arrochar	Bus Stop	Deli / Bodega	Pizza Place	Italian Restaurant



