The battle of neighborhoods in the city of Toronto

by Tomas Stankevičius

2021

Content

Introduction / Business Problem

The company wants to start a restaurant business and is trying to decide which area in Toronto could be a right spot to do that. An area should have the biggest potentiality in developing restaurant business. Location should help company to reach as wide customers auditory as possible to make business profitable. As company does not possess a knowledge of Toronto districts and neighborhoods, it decided to hire a business consultant company which provides various consultant services for companies which want to establish new business in Toronto area. A consultant company utilizing Data Science methods is going to cluster Toronto city area and provide recommendations where is the best place to start a restaurant business.

As there are a lot of restaurants in Toronto already, we will take into consideration every district's population and its density too. We are very interested in districts located in Toronto Downtown as tourists can generate additional revenues too.

Explanation about Data

Consultant company is going to use Foursquare location data services to build a data base of various venues in every neighborhood of Toronto city. Then this data base will enriched with additional data provided by Toronto Open Data portal.

At first we are going to gather geospatial information from Toronto Open Data portal about every neighborhood in Toronto City. From page - https://open.toronto.ca/dataset/neighbourhoods/

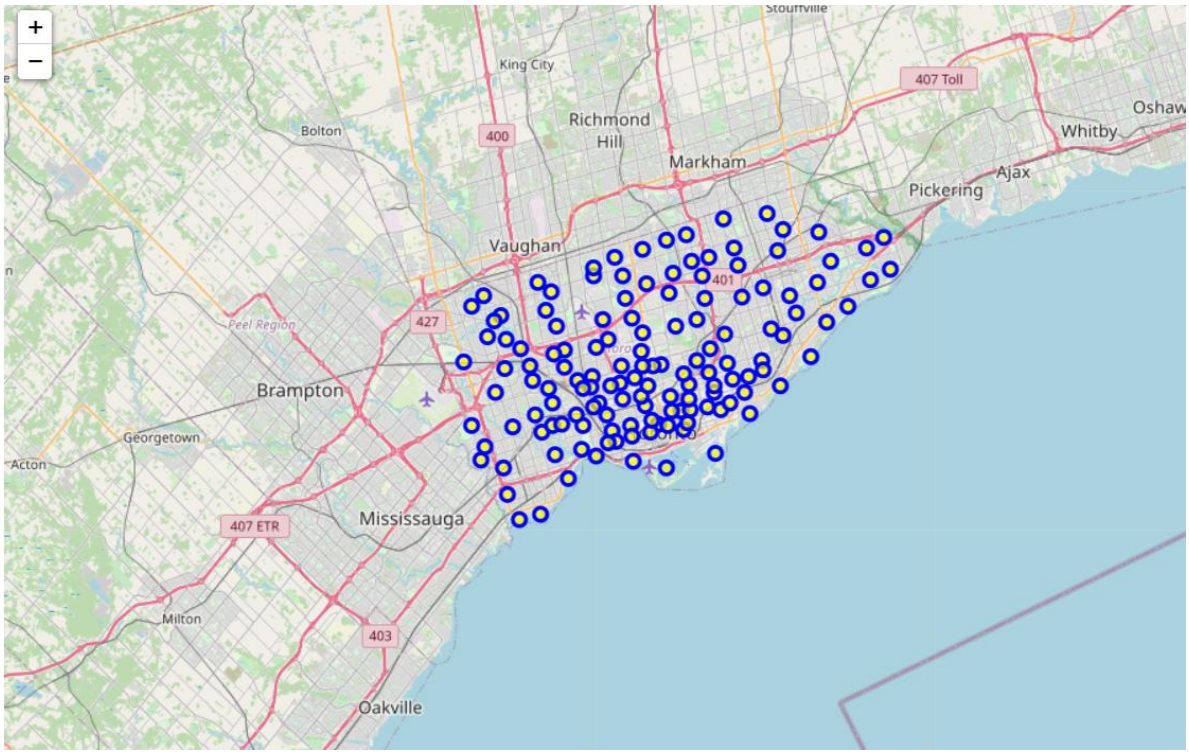
Next, we will collect population and density information about every neighborhood in Toronto City from the page - https://open.toronto.ca/dataset/neighbourhood-profiles/

Lastly, a **Foursquare API** explore endpoint provides a list of recommended venues near the current location. We are going to use this endpoint to build our venues data base. Folium package provides a nice maps building functionality, so, we will be using it too to visualize results.

Our 5 rows of processed and combined data from mentioned Toronto Open Data portal in pandas data frame looks like this:

	Neighbourhood	Latitude	Longitude	Population, 2016	Population density per square kilometre
0	Agincourt North	43.806043	-79.259428	29113	3929
1	Agincourt South-Malvern West	43.789804	-79.265165	23757	3034
2	Alderwood	43.606948	-79.544768	12054	2435
3	Annex	43.673365	-79.402468	30526	10863
4	Banbury-Don Mills	43.738617	-79.348636	27695	2775

Using Folium package we can map those neighborhoods on Toronto map to get better understanding of their locations:



A Foursquare API explore endpoint provides a list of recommended venues near the current location. We are going to use this endpoint to build our venues data base. A fraction example of response code looks like this:


```
{
  "meta": {
    "code": 200,
    "requestId": "5ac51ef86a607143de8eg5cb"
  },
  "response": {
    "warning": {
      "text": "There aren't a lot of results near you. Try something more general, reset your filters, or
    },
    "suggestedRadius": 600,
    "headerLocation": "Lower East Side",
    "headerFullLocation": "Lower East Side, New York",
    "headerLocationGranularity": "neighborhood",
    "totalResults": 230,
    "suggestedBounds": {
      "ne": {
        "lat": 40.724216906965616,
        "lng": -73.9896507407283
      },
      "sw": {
        "lat": 40.72151724718017,
        "lng": -73.98693222860872
      }
    },
    "groups": [
      {
        "type": "Recommended Places",
        "name": "recommended",
```

Lastly, a Foursquare API explore endpoint provides a list of recommended venues near the current location. We are going to use this endpoint to build our venues data base. Folium package provides a nice maps building functionality, so, we will be using it too to visualize results.

Methodology

As mentioned, this project goal is to find area in Toronto city with high population and enough density for a newly established restaurant to be able to attract sufficient amount of clients from the very beginning.

First, we have collected and processed required data such as Toronto neighborhoods geospatial data, as well as population and density in those neighborhoods. Using Foursquare API we were able to build a categorized database of venues located in every Toronto district.

Next step in our analysis will be to analyze gathered venue data and process and scale it, in order to build a Machine Learning model based on unsupervised K-Means algorithm. K-Means lets us to cluster districts according to their similar features. Districts inside clusters are very similar, but clusters from each other are very different.

We will build several graphs and maps for better understanding of K-Means model results as well as to help us draw a final conclusions

Let's sort Venue Category in our data frame. As we can see top 10 most popular venue categories are: Coffee Shop, Cafe, Park, Restaurant, Pizza Place, Sandwich Place, Italian Restaurant, Bakery, Bar, Grocery Store.

So, all categories are related to food, except park. That's quite promising result for our further analysis.

```
toronto_venues_filt.groupby(['Venue Category'])['Venue Category'].count().sort_values(ascending=False)[:15]
```

```
Venue Category
Coffee Shop            193
Park                   147
Café                   108
Pizza Place             94
Italian Restaurant      75
Bakery                  71
Sandwich Place          69
Restaurant              68
Grocery Store           63
Bank                    55
Pharmacy                53
Fast Food Restaurant    48
Bar                     47
Sushi Restaurant        40
Indian Restaurant       36
Name: Venue Category, dtype: int64
```

Further we perform one hot encoding on Toronto Venues data frame and calculate an average frequency of every category in every neighborhood. Please, check a data frame an example of first 5 rows bellow.

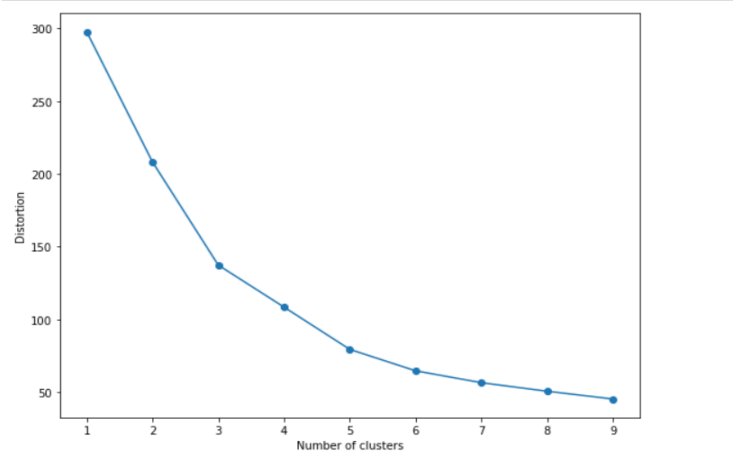
	Neighbourhood	Accessories Store	Afghan Restaurant	African Restaurant	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	I
0	Agincourt North	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	Agincourt South-Malvern West	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	Alderwood	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	Annex	0.0	0.0	0.0	0.018868	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	Banbury-Don Mills	0.0	0.0	0.0	0.023256	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Next we create a pandas data frame of 10 most popular venues in each neighborhood. Here is example of first 5 rows.

	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
0	Agincourt North	Indian Restaurant	Beer Store	Fast Food Restaurant	Salon / Barbershop	Coffee Shop	Shopping Mall	Ice Cream Shop	Antique Shop	Fish Market	Farm
1	Agincourt South-Malvern West	Chinese Restaurant	Park	BBQ Joint	Lounge	Clothing Store	Sandwich Place	Latin American Restaurant	Malay Restaurant	Shopping Mall	Noodle House
2	Alderwood	Pizza Place	Gas Station	Pharmacy	Skating Rink	Park	Hotel	Moroccan Restaurant	Convenience Store	Intersection	Construction & Landscaping
3	Annex	Café	Pub	Italian Restaurant	Coffee Shop	Gym	Sandwich Place	Mexican Restaurant	Electronics Store	Social Club	Church
4	Banbury-Don Mills	Restaurant	Coffee Shop	Bank	Café	Pizza Place	Clothing Store	Gourmet Shop	Sushi Restaurant	Optical Shop	Movie Theater

After that we need to prepare a data for K-Mean clustering algorithm. We need to scale Population and Population density for model to be accurate. We use StandarScaler module from sklearn library.

At first we set a random number for max cluster numbers as we don't know the best value so far. Let's set total cluster number to 3. We fit a model with prepared data. After that let’s find a best value of k for our model.



As we see from the graph a distortion / inertia falls quite rapidly till we reach max number of clusters equal to 5. So let’s choose k value as 5 as it is the best fit for our model.

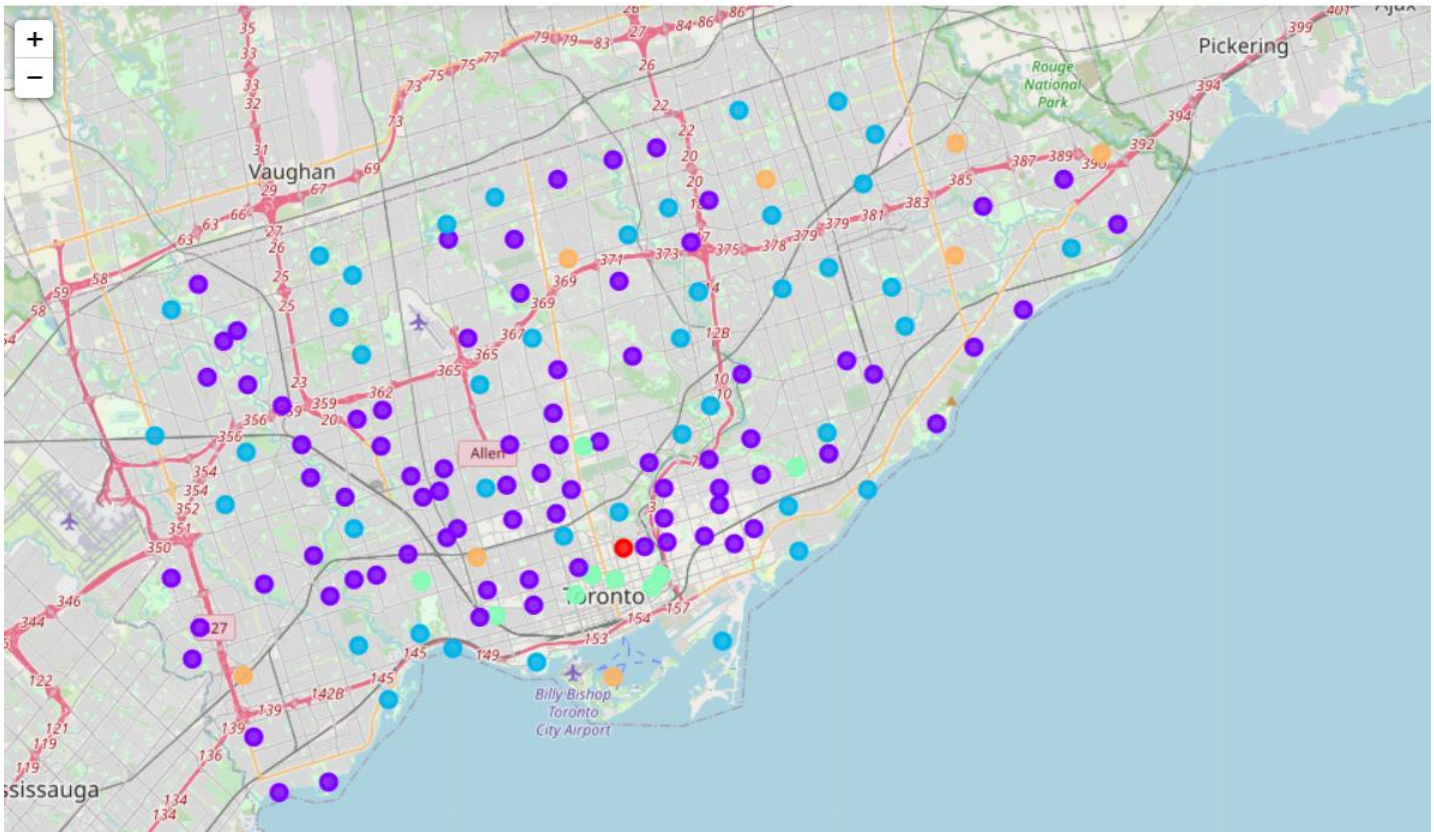
Clustering Analysis

Now let's prepare a final data frame for visualization and further analysis. We merge tow pandas data frames – Toronto population data with Toronto venues data, which includes clustering labels.

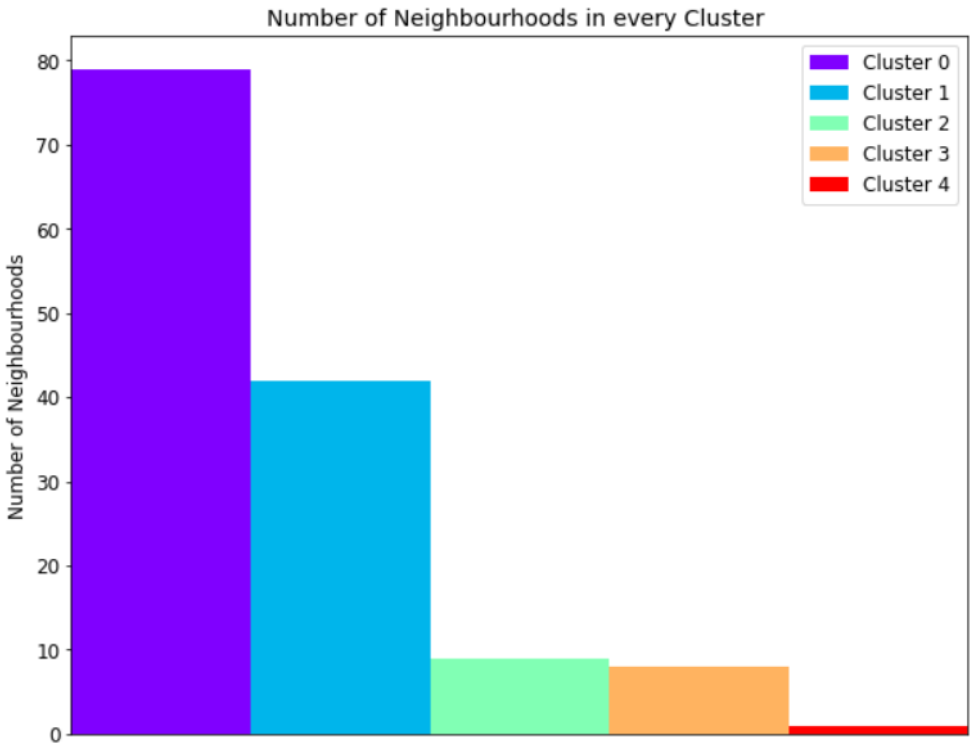
```
tor_final = tor_final.merge(neighborhoods_venues_sorted, left_on=['Neighbourhood'], right_on=['Neighbourhood'])
tor_final.head()
```

	Neighbourhood	Latitude	Longitude	Population, 2016	Population density per square kilometre	Cluster Labels	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	8tl Co
0	Agincourt North	43.806043	-79.259428	29113	3929	1	Indian Restaurant	Beer Store	Fast Food Restaurant	Salon / Barbershop	Coffee Shop	Shopping Mall	Ice Cream Shop	A
1	Agincourt South-Malvern West	43.789804	-79.265165	23757	3034	1	Chinese Restaurant	Park	BBQ Joint	Lounge	Clothing Store	Sandwich Place	Latin American Restaurant	Rest
2	Alderwood	43.606948	-79.544768	12054	2435	0	Pizza Place	Gas Station	Pharmacy	Skating Rink	Park	Hotel	Moroccan Restaurant	Conve
3	Annex	43.673365	-79.402468	30526	10863	1	Café	Pub	Italian Restaurant	Coffee Shop	Gym	Sandwich Place	Mexican Restaurant	Elec
4	Banbury-Don Mills	43.738617	-79.348636	27695	2775	1	Restaurant	Coffee Shop	Bank	Café	Pizza Place	Clothing Store	Gourmet Shop	Rest

Let's visualize results using Folium package. Clusters are marked on map. Different color represents a neighborhoods dependency to different clusters.

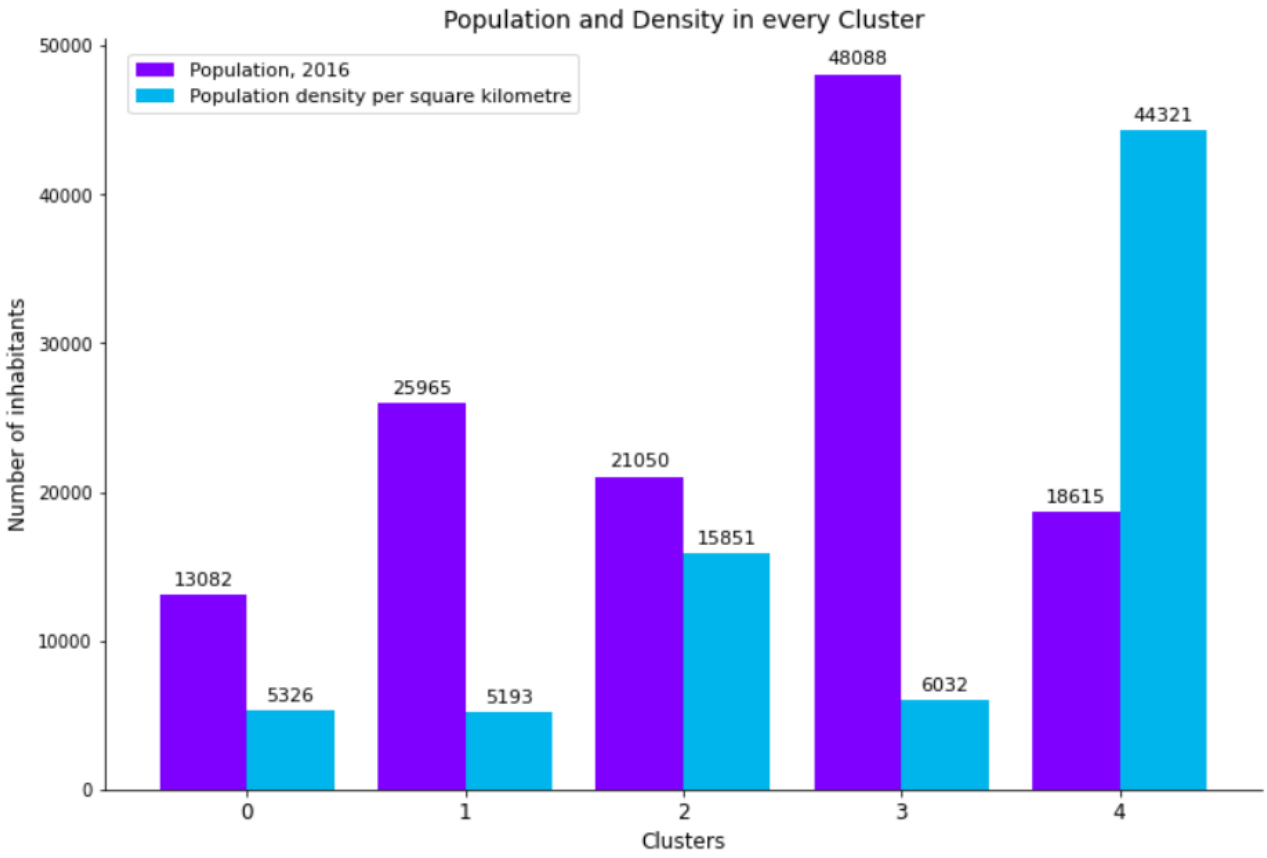


We check how many districts there are in every clusters. Results are bellow.



As we can see the most popular cluster is number 0, with almost 80 neighborhoods in that cluster.

Let's check average population and density in every cluster.



We see interesting population and density distribution. Although Cluster 3 has highest population, its density is one of the lowest. Cluster 4 has highest population density, but inhabitants number is only at the 4th place. It is because cluster 4 has only one district.

Let's examine every cluster separately.

```
In [136]: clus_0 = tor_final.loc[tor_final['Cluster Labels'] == 0, tor_final.columns[[0] +[3] +[4]+ list(range(6, tor_final.shape[1]))]]
clus_0
```

Out[136]:

	Neighbourhood	Population, 2016	Population density per square kilometre	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
2	Alderwood	12054	2435	Pizza Place	Gas Station	Pharmacy	Skating Rink	Park	Hotel	Moroccan Restaurant	Convenience Store
5	Bathurst Manor	15873	3377	Lawyer	Trail	Park	Skating Rink	Food	Farmers Market	Food Court	Event Space
8	Bayview Woods-Steeles	13154	3240	Park	Dog Run	Zoo	Event Space	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
10	Beechborough-Greenbrook	6577	3614	Furniture / Home Store	Discount Store	Convenience Store	Video Store	Fast Food Restaurant	Grocery Store	Sandwich Place	Italian Restaurant
14	Blake-Jones	7727	8134	Café	Pizza Place	Fast Food Restaurant	Burger Joint	Park	Department Store	Brewery	Mexican Restaurant
15	Briar Hill-Belgravia	14257	7791	Furniture / Home Store	Coffee Shop	Sandwich Place	Pharmacy	Japanese Restaurant	Falafel Restaurant	Bike Shop	Park

Cluster 0 is most popular cluster. It contains districts with medium size and low density populations. Because of it and proximity to Toronto Downtown, this cluster is not the best candidate. Also if we look at the top venues, restaurant or similar food service places are not amongst the most popular venues.

```
In [137]: clus_1 = tor_final.loc[tor_final['Cluster Labels'] == 1, tor_final.columns[[0] +[3] +[4]+ list(range(6, tor_final.shape[1]))]]
clus_1
```

Out[137]:

	Neighbourhood	Population, 2016	Population density per square kilometre	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 M Comr Ve
0	Agincourt North	29113	3929	Indian Restaurant	Beer Store	Fast Food Restaurant	Salon / Barbershop	Coffee Shop	Shopping Mall	Ice Cream Shop	Antique Shop	Ma
1	Agincourt South-Malvern West	23757	3034	Chinese Restaurant	Park	BBQ Joint	Lounge	Clothing Store	Sandwich Place	Latin American Restaurant	Malay Restaurant	Shop
3	Annex	30526	10863	Café	Pub	Italian Restaurant	Coffee Shop	Gym	Sandwich Place	Mexican Restaurant	Electronics Store	Social
4	Banbury-Don Mills	27695	2775	Restaurant	Coffee Shop	Bank	Café	Pizza Place	Clothing Store	Gourmet Shop	Sushi Restaurant	Op S
7	Bayview Village	21396	4195	Metro Station	Furniture / Home Store	Moving Target	Park	Deli / Bodega	Sandwich Place	Fish Market	Breakfast Spot	Fast F Restau
9	Bedford Park-	23236	4209	Italian	Indian	Café	Greek	Pet Store	Pharmacy	Sushi	Restaurant	Con

Neighborhoods in cluster 1 has high population numbers but low density. From the locations on the Toronto map we see it mainly a residential areas with houses and cottages. Although venues related to restaurants are still popular, but because of density and districts distance to the Toronto Downtown, this cluster is not so promising.

```
In [125]: clus_2 = tor_final.loc[tor_final['Cluster Labels'] == 2, tor_final.columns[[0] +[3] +[4]+ list(range(6, tor_final.shape[1]))]]
          clus_2
```

Out[125]:

	Neighbourhood	Population, 2016	Population density per square kilometre	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
6	Bay Street Corridor	25797	14097	Coffee Shop	Café	Sushi Restaurant	Sandwich Place	Bubble Tea Shop	Middle Eastern Restaurant	Chinese Restaurant	Tea Room	Japanese Restaurant	Ice Crear Sho
23	Church-Yonge Corridor	31340	23044	Coffee Shop	Clothing Store	Japanese Restaurant	Theater	Cosmetics Shop	Café	Diner	Ramen Restaurant	Fast Food Restaurant	Burge Joir
49	High Park North	22162	11726	Park	Café	Convenience Store	Pub	Mexican Restaurant	Baseball Field	Mattress Store	Grocery Store	Sandwich Place	Gym Fitnes Cente
62	Kensington-Chinatown	17945	11806	Café	Vegetarian / Vegan Restaurant	Coffee Shop	Mexican Restaurant	Bar	Bakery	Grocery Store	Caribbean Restaurant	Beer Bar	Arts Craft Stor
71	Little Portugal	15559	12859	Bar	Restaurant	Coffee Shop	Café	Thrift / Vintage Store	Grocery Store	Cocktail Bar	Pizza Place	Park	Sandwic Plac
79	Moss Park	20506	14753	Coffee Shop	Park	Restaurant	Bakery	Breakfast Spot	Thai Restaurant	Theater	Pub	Italian Restaurant	Gyr
83	Mount Pleasant West	29658	21969	Coffee Shop	Italian Restaurant	Sushi Restaurant	Restaurant	Pub	Movie Theater	Middle Eastern Restaurant	Gastropub	Bookstore	Caribbea Restaurar
100	Regent Park	10803	16880	Coffee Shop	Pizza Place	Sushi Restaurant	Curling Ice	Brewery	Breakfast Spot	Falafel Restaurant	Thai Restaurant	Bar	Baker
115	The Beaches	15683	15528	Beach	Pub	Bakery	Café	Breakfast Spot	Pizza Place	Park	Japanese Restaurant	Bar	Nail Salo

Cluster 2 contains 9 districts. As we see from the table every neighborhood has high population and high enough density. Top 3 most common venues are consist mainly of restaurants or entertainment places such as parks and beaches. Cluster 2 looks very promising as we have a nice balance of population, density and various venues situated in clustered districts.

```
In [126]: clus_3 = tor_final.loc[tor_final['Cluster Labels'] == 3, tor_final.columns[[0] +[3] +[4]+ list(range(6, tor_final.shape[1]))]]
          clus_3
```

Out[126]:

	Neighbourhood	Population, 2016	Population density per square kilometre	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	1
32	Dovercourt-Wallace Emerson-Junction	36625	9819	Park	Café	Brewery	Portuguese Restaurant	Coffee Shop	Bar	Bakery	Pharmacy	Liquor Store	Sup
58	Islington-City Centre West	43965	2712	Fast Food Restaurant	Supermarket	Concert Hall	BBQ Joint	Camera Store	Mobile Phone Shop	Furniture / Home Store	Supplement Shop	Chinese Restaurant	Ga
65	L'Amoreaux	43993	6144	Mexican Restaurant	Caribbean Restaurant	Hotpot Restaurant	Greek Restaurant	Chinese Restaurant	Bank	Vietnamese Restaurant	Bakery	Restaurant	Pi
73	Malvern	43794	4948	Fast Food Restaurant	Pharmacy	Pizza Place	Grocery Store	Restaurant	Gym / Fitness Center	Sandwich Place	Park	Supermarket	B
105	Rouge	46496	1260	Trail	Coffee Shop	Mobile Phone Shop	Fish & Chips Shop	Fried Chicken Joint	Mexican Restaurant	Campground	Breakfast Spot	Shopping Mall	
122	West Hill	65913	8943	Gym / Fitness Center	Gymnastics Gym	Grocery Store	Park	Athletics & Sports	Zoo	Exhibit	Falafel Restaurant	Farm	
129	Willowdale West	50434	10087	Park	Road	Financial or Legal Service	Exhibit	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	R
132	Woodbine Corridor	53485	4345	Grocery Store	Hungarian Restaurant	Café	Gas Station	Baseball Field	Zoo	Financial or Legal Service	Farm	Farmers Market	R

Cluster 3 has high enough population, but low density. From the cluster map we see, that majority of cluster 4 districts are located quite far away from Toronto Downtown.

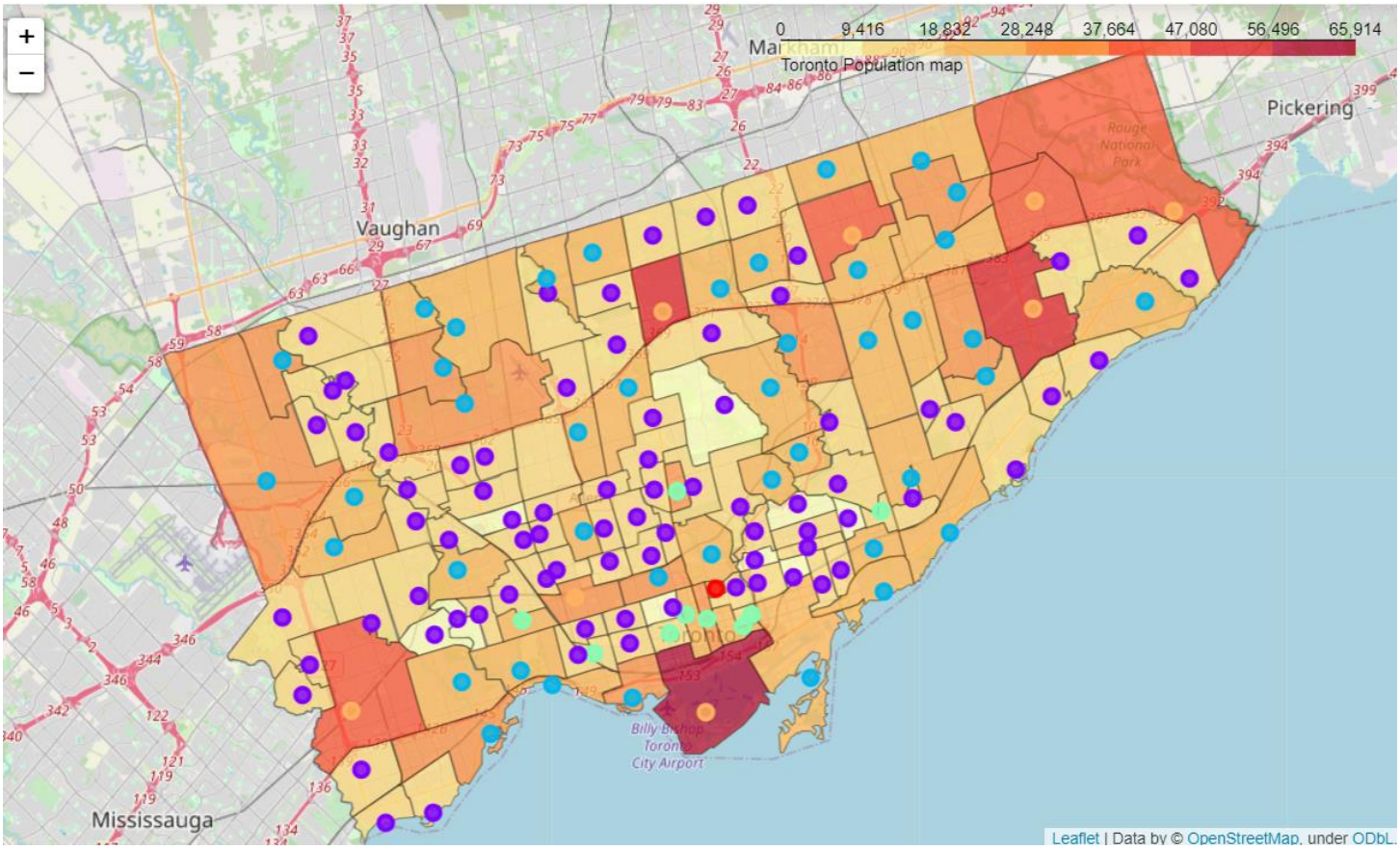
In [127]:
```
clus_4 = tor_final.loc[tor_final['Cluster Labels'] == 4, tor_final.columns[[0] +[3] +[4]+ list(range(6, tor_final.shape[1]))]]
clus_4
```

Out[127]:

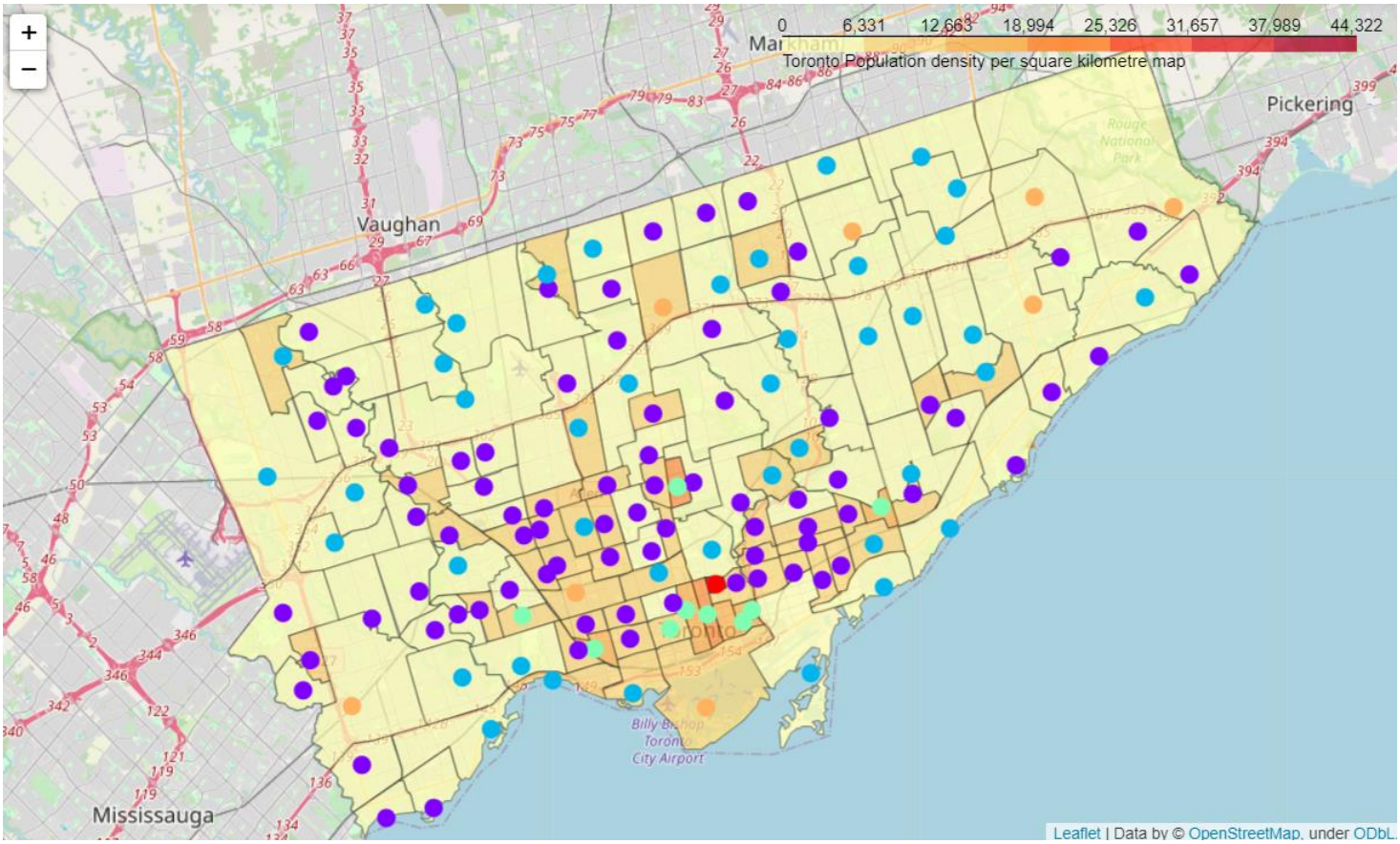
	Neighbourhood	Population, 2016	Population density per square kilometre	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
89	North St.James Town	18615	44321	Coffee Shop	Restaurant	Grocery Store	Breakfast Spot	Hotel	Café	Sandwich Place	Park	Pharmacy	Filipino Restaurant

We can call Cluster 4 an outlier cluster as it contains only one neighborhood in it. This cluster has high enough population and high density and restaurant venues by popularity are among most popular types of venues. It's quite promising candidate for our analysis.

Now let's see how every cluster looks on a population map. Folium package provides a nice functionality to build these kind of maps for us to better understand our problem and make correct conclusions.



Also let's see how every cluster looks on a Toronto population density map.



Results and Discussion

Our clustering analysis show that taking into consideration such factors as population density, total neighborhood population number, a distance to Toronto Downtown as well as top10 venues category popularity distribution in every district, majority of Cluster 2 districts falls under those criterions. As Cluster 4 contains only one district it can also be included into potential areas to establish a new restaurant.

We see that some districts from Cluster 0 also falls under our criterions. A further more narrow and concentrated analysis is required which is out of scope of this project.

Conclusion

The goal of this project was to identify and locate Toronto districts which are close to a Downtown and has high population numbers as well as high density. Using Foursquare API we took into consideration a most popular venues types in Toronto districts to narrow our results.

Our Machine Learning clustering model build on combined data provided 5 different clusters of which Cluster 2 and 4 looks most promising.

For final quality decision a further additional analysis of those clusters including more variables are required. A type and number of tourists attractions and their annual visiting number in each neighborhood, attractiveness of each neighborhood based on e.g. residents income level, noise levels, social and economic factors, crime situation and etc. are among those additional factors a company should take into consideration.