Capstone Project: Segmenting and Clustering Neighborhoods in Toronto City

Pashkov Daniel

22.10.2020

1 Introduction: Description the Problem.

Today in modern city there are hundreds of venues and at each of its decade's restaurants, shops, houses etc. For city government and businessman, who decide open new spa or park, or for citizens, who want to buy new home, very important know what is difference between city areas, which often unequal to districts borders.

So, needs some clusterisation, that dividing city on clusters based on most using places in that area. That is the main purpose of this project.

2 Data cleaning and feature selection

2.1 Data sourse

For this project neighborhoods data was scraped from wikipedia page using BeautifulSoup library (for more information link here).This data includes postal code for each borough.

For exploring neighborhoods in this project was use Foursquare API. Foursquare is a technology company that built a massive dataset of location data. They actually crowd-sourced their data and had people use their app to build their dataset and add venues and complete any missing information they had in their dataset. Currently its location data is the most comprehensive out there, and quite accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, and is currently being used by over 100,000 developers, and this number is only growing.

2.2 Data cleaning

Original data set have 217 neighborhoods. (Figure 1)

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park
3	M5A	Downtown Toronto	Harbourfront
4	M6A	North York	Lawrence Manor
...	...	...	...
212	M8Z	Etobicoke	Mimico NW
213	M8Z	Etobicoke	The Queensway West
214	M8Z	Etobicoke	South of Bloor
215	M8Z	Etobicoke	Kingsway Park South West
216	M8Z	Etobicoke	Royal York South West

217 rows × 3 columns

Figure 1. Original Dataset.

Further, get the coordinates for every neighborhood using Nominatim search engine and join this data with original dataset.

However, for some neighborhoods engine not found coordinates, so this row was dropped. After that, clean data includes 198 neighborhoods.(Figure 2, Figure 3)

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.761124	-79.324059
1	M4A	North York	Victoria Village	43.732658	-79.311189
2	M5A	Downtown Toronto	Regent Park	43.660706	-79.360457
3	M5A	Downtown Toronto	Harbourfront	43.640080	-79.380150
4	M6A	North York	Lawrence Manor	43.722079	-79.437507
...	...	...	...	...	...
193	M8Y	Etobicoke	Royal York South East	43.648183	-79.511296
194	M8Z	Etobicoke	Mimico NW	43.616677	-79.496805
195	M8Z	Etobicoke	The Queensway West	43.623618	-79.514764
196	M8Z	Etobicoke	South of Bloor	43.670489	-79.386465
197	M8Z	Etobicoke	Royal York South West	43.648183	-79.511296

198 rows × 5 columns

Figure 2. Data with coordinates after cleaning.

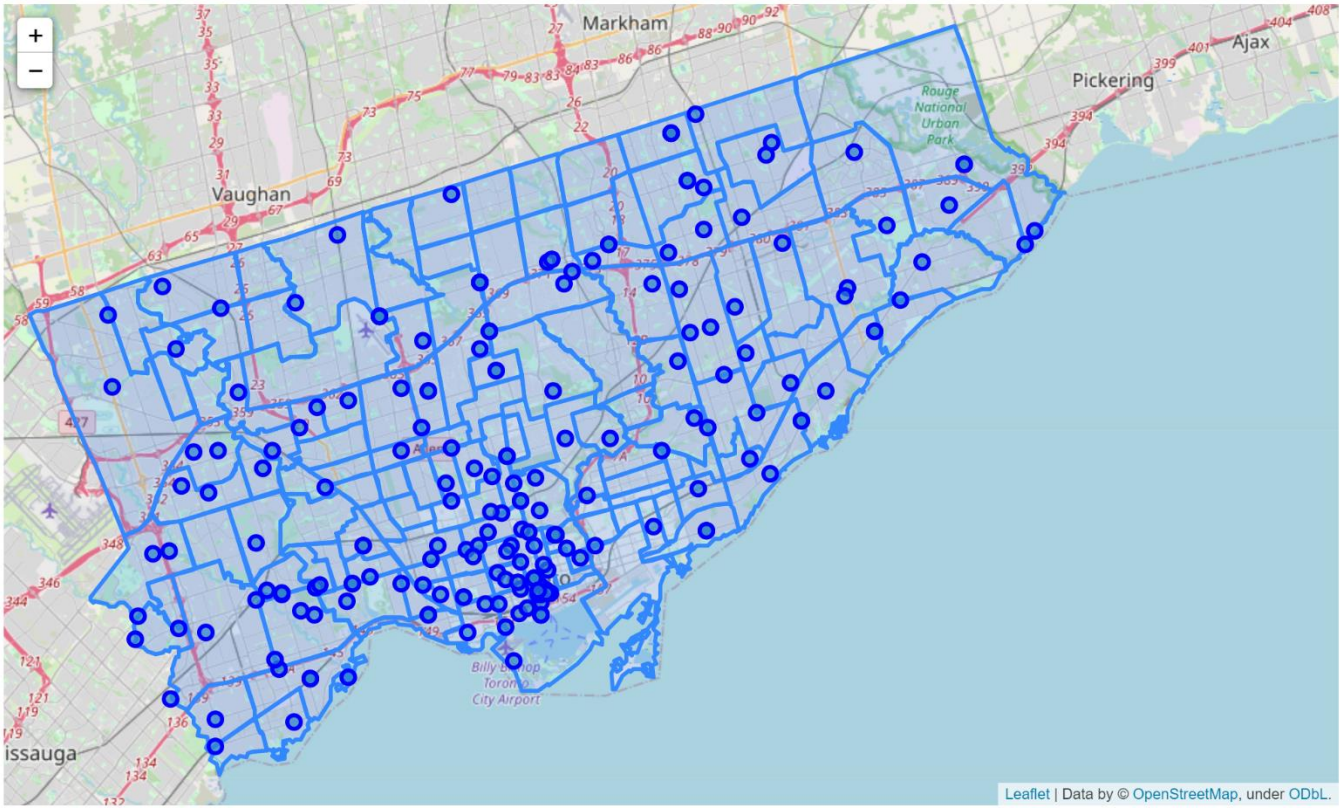

Figure 3. Neighborhoods in Toronto map. Straight line denotes districts borders.

According to map above (Figure 3), we see that some points located in a distance between others. So, for analyze neighborhoods needs calculate maximum distance between closest points, and this distance is about 3 kilometers.

3 Exploratory Analysis

3.1 Explore Neighborhoods in Toronto

For using Foursquare API was created account in website for developer. This account give credentials for connecting to the servers.

To get most common venues for each neighborhoods use explore request for venues with radius, which was calculated in previous chapter (half of 3 kilometers).

Example for explore request:

url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION, lat, lng, radius, LIMIT)

In example there are LIMIT parameters – is default REST API's limit of value in response.

After, using response for this request (in JSON format), creates dataframe with next columns/parameters (Figure 4):

1. 'Neighborhood',
2. 'Neighborhood Latitude',
3. 'Neighborhood Longitude',
4. 'Venue',
5. 'Venue Latitude',
6. 'Venue Longitude',
7. 'Venue Category'

Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Parkwoods	43.761124	-79.324059	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
Parkwoods	43.761124	-79.324059	Tim Hortons	43.760668	-79.326368	Café
Parkwoods	43.761124	-79.324059	A&W	43.760643	-79.326865	Fast Food Restaurant
Parkwoods	43.761124	-79.324059	LCBO	43.757774	-79.314257	Liquor Store
Parkwoods	43.761124	-79.324059	Dollarama	43.758135	-79.310672	Discount Store

Figure 4. Data from API response.

3.2 Analyze Each Neighborhood

After getting venues, use onehot encoding venues category for each neighborhoods. In result – dataset there for each neighborhoods there are columns with 0/1 label for venues category. (Figure 5)

	Neighbourhood	Accessories Store	Afghan Restaurant	Airport	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	Arcade	...	Volleyball Court	Warehouse Store	Whisky Bar	Wine Bar
0	Parkwoods	0	0	0	0	0	0	0	0	0	...	0	0	0	0
1	Parkwoods	0	0	0	0	0	0	0	0	0	...	0	0	0	0
2	Parkwoods	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	Parkwoods	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4	Parkwoods	0	0	0	0	0	0	0	0	0	...	0	0	0	0

Figure 5. Data after onehot encoding.

Furter, first, group rows by neighborhood and by taking the mean of the frequency of occurrence of each category and, second, create the new dataframe with top 10 venues for each neighborhood. (Figure 6)

	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	Adelaide	Coffee Shop	Restaurant	Café	Hotel	Beer Bar	Japanese Restaurant	Gastropub	Thai Restaurant	Pizza Place	Plaza
1	Agincourt North	Indian Restaurant	Coffee Shop	Pizza Place	Vietnamese Restaurant	Sandwich Place	Chinese Restaurant	Bank	Park	Pharmacy	Bubble Tea Shop
2	Bathurst Quay	Park	Coffee Shop	Gym	Café	Restaurant	Yoga Studio	Italian Restaurant	Bakery	Seafood Restaurant	Dog Run
3	Bloordale Gardens	Coffee Shop	Sandwich Place	Beer Store	Bank	Gym	Pizza Place	Liquor Store	Clothing Store	Pharmacy	Convenience Store
4	Broadview North (Old East York)	Greek Restaurant	Park	Café	Italian Restaurant	Pub	Bakery	Coffee Shop	Pizza Place	Burger Joint	Flower Shop

Figure 6. Dataset with feature columns.

4 Clustering Neighborhoods

4.1 K-means Clustering Algorithm

In this project for clustering using k-means algorithm. K-means can group data only unsupervised based on the similarity to each other. There are various types of clustering algorithms such as partitioning, hierarchical, or density based clustering.

K-means is a type of partitioning clustering. That is, it divides the data into k non-overlapping subsets or clusters without any cluster internal structure or labels. This means, it's an unsupervised algorithm. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar.

For selecting the k-means parameters use the GridSearchCV library. After training, best models includes 7 clusters. In map below (Figure 7), each color of markers is different cluster.

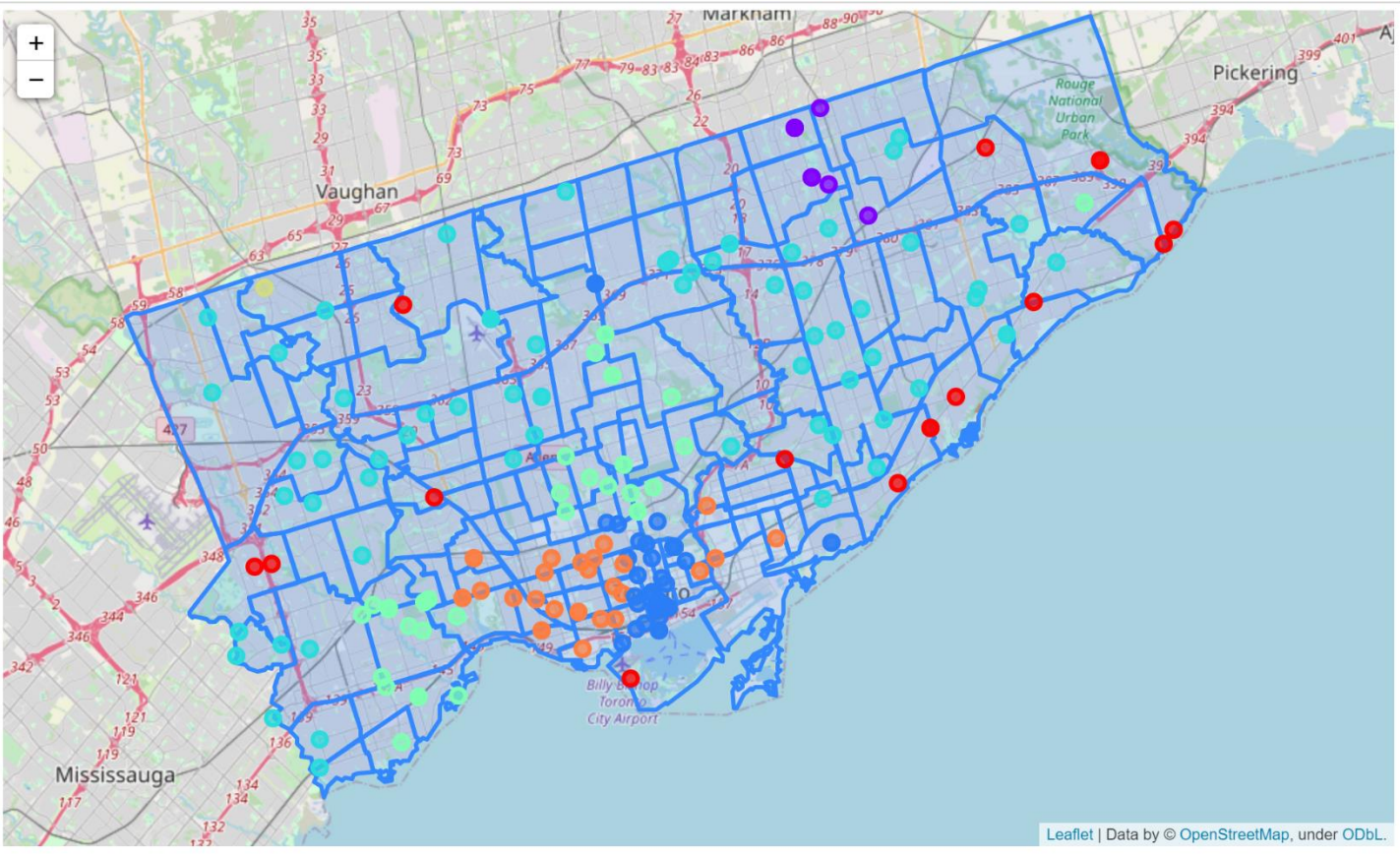

Figure 7. Toronto map with markered clusters.

4.2 Examine Clusters

In this section, examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, possible assign a name to each cluster.

Cluster 1

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
90	Scarborough	0	Harbor / Marina	Pizza Place	Fast Food Restaurant	Chinese Restaurant	Asian Restaurant	Grocery Store	Gift Shop	Liquor Store	Park	Hardware Store
102	Scarborough	0	Park	Coffee Shop	Pub	Harbor / Marina	Breakfast Spot	Grocery Store	Seafood Restaurant	Hardware Store	Sushi Restaurant	Sandwich Place
98	York	0	Coffee Shop	Park	Asian Restaurant	Convenience Store	Gas Station	Tennis Court	Golf Course	Beer Store	Supermarket	Thrift / Vintage Store
21	Scarborough	0	Park	Sandwich Place	Convenience Store	Pet Store	Fast Food Restaurant	Pizza Place	Coffee Shop	Beer Store	Train Station	Zoo
10	Scarborough	0	Park	Trail	Gas Station	Pizza Place	Pharmacy	Bank	Gym / Fitness Center	Campground	Garden	Discount Store
60	Downtown Toronto	0	Harbor / Marina	Park	Lighthouse	Pizza Place	Dog Run	BBQ Joint	Boat or Ferry	Café	Nudist Beach	Track
101	Scarborough	0	Park	Pizza Place	Chinese Restaurant	Café	Golf Course	Gym	Diner	Filipino Restaurant	Bus Stop	Hotel

According map and cluster description this cluster represents areas, which located on coast or near harbor and rivers. So, marked this cluster as 'Coasts Areas'.

Cluster 2

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
152	Scarborough	1	Chinese Restaurant	Coffee Shop	Fast Food Restaurant	Vietnamese Restaurant	Pharmacy	Bank	Grocery Store	Athletics & Sports	Pizza Place	Intersection
172	Scarborough	1	Chinese Restaurant	Coffee Shop	Fast Food Restaurant	Vietnamese Restaurant	Pharmacy	Bank	Grocery Store	Athletics & Sports	Pizza Place	Intersection
151	Scarborough	1	Chinese Restaurant	Bubble Tea Shop	Bakery	Japanese Restaurant	Tea Room	Vietnamese Restaurant	Korean Restaurant	Park	Gas Station	Bank
135	Scarborough	1	Chinese Restaurant	Shopping Mall	Cantonese Restaurant	Bank	Fast Food Restaurant	Sandwich Place	Bookstore	Filipino Restaurant	Pizza Place	Coffee Shop
141	Scarborough	1	Chinese Restaurant	Coffee Shop	Bubble Tea Shop	Vietnamese Restaurant	Dessert Shop	BBQ Joint	Bakery	Pharmacy	Park	Gas Station
149	Scarborough	1	Chinese Restaurant	Park	Bubble Tea Shop	Bakery	Pizza Place	Noodle House	Japanese Restaurant	Gym	Bank	Korean Restaurant
165	Etobicoke	1	Chinese Restaurant	Bubble Tea Shop	Bakery	Japanese Restaurant	Tea Room	Vietnamese Restaurant	Korean Restaurant	Park	Gas Station	Bank
171	Scarborough	1	Chinese Restaurant	Bubble Tea Shop	Bakery	Japanese Restaurant	Tea Room	Vietnamese Restaurant	Korean Restaurant	Park	Gas Station	Bank

According the 1st common venues, obvious, it's 'Chinatown'.

Cluster 3

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
46	Downtown Toronto	2	Coffee Shop	Restaurant	Café	Hotel	Beer Bar	Japanese Restaurant	Gastropub	Thai Restaurant	Pizza Place	Plaza
161	Downtown Toronto	2	Park	Coffee Shop	Gym	Café	Restaurant	Yoga Studio	Italian Restaurant	Bakery	Seafood Restaurant	Dog Run
180	Downtown Toronto	2	Coffee Shop	Park	Café	Diner	Bakery	Restaurant	Italian Restaurant	Gastropub	Japanese Restaurant	Gay Bar
71	Downtown Toronto	2	Café	Restaurant	Coffee Shop	Hotel	Japanese Restaurant	Beer Bar	Park	Plaza	Pizza Place	Movie Theater
148	Downtown Toronto	2	Coffee Shop	Café	Restaurant	Beer Bar	Mexican Restaurant	Sandwich Place	Hotel	Taco Place	Arts & Crafts Store	Art Gallery
3	Downtown Toronto	2	Coffee Shop	Café	Hotel	Park	Japanese Restaurant	Restaurant	Theater	Scenic Lookout	Brewery	Farmers Market
160	Downtown Toronto	2	Coffee Shop	Café	Hotel	Park	Japanese Restaurant	Restaurant	Theater	Scenic Lookout	Brewery	Farmers Market

In this cluster the most common the coffee shops, cafes, pubs and points are located in center of Toronto, so this is 'City Center'.

Cluster 4

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
150	Scarborough	3	Indian Restaurant	Coffee Shop	Pizza Place	Vietnamese Restaurant	Sandwich Place	Chinese Restaurant	Bank	Park	Pharmacy	Bubble Tea Shop
28	Etobicoke	3	Coffee Shop	Sandwich Place	Beer Store	Bank	Gym	Pizza Place	Liquor Store	Clothing Store	Pharmacy	Convenience Store
76	Scarborough	3	Bakery	Pizza Place	Coffee Shop	Bank	Bus Line	Fast Food Restaurant	Intersection	Restaurant	Park	Sandwich Place
19	Etobicoke	3	Coffee Shop	Grocery Store	Sandwich Place	Bank	Vietnamese Restaurant	Fast Food Restaurant	Department Store	Pharmacy	Beer Store	Sporting Goods Shop
43	North York	3	Sandwich Place	Gym / Fitness Center	Flea Market	Athletics & Sports	Skating Rink	Coffee Shop	Playground	Escape Room	Theater	Climbing Gym
...	...	...	...	...	...	...	...	...	...	...	...	...
1	North York	3	Coffee Shop	Fast Food Restaurant	Middle Eastern Restaurant	Grocery Store	Pizza Place	Gym	Chinese Restaurant	Intersection	Shoe Store	Rental Car Location
120	Etobicoke	3	Pizza Place	Coffee Shop	Train Station	Plaza	Park	Sandwich Place	Ice Cream Shop	Flea Market	Fried Chicken Joint	Café

According number and location of this cluster, it's 'Low Cost Residential Areas'.

Cluster 5

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
157	Central Toronto	4	Park	Italian Restaurant	Sushi Restaurant	Café	Coffee Shop	Mexican Restaurant	Gym	Burger Joint	Spa	Gastropub
117	Central Toronto	4	Italian Restaurant	Sushi Restaurant	Coffee Shop	Café	Bank	Park	Liquor Store	Grocery Store	Ice Cream Shop	Pizza Place
156	Central Toronto	4	Coffee Shop	Pizza Place	Bank	Italian Restaurant	Liquor Store	Japanese Restaurant	Sushi Restaurant	Restaurant	Sandwich Place	Bagel Shop
22	Scarborough	4	Burger Joint	Coffee Shop	Athletics & Sports	Italian Restaurant	Breakfast Spot	Gym / Fitness Center	Neighborhood	Gym	Fish & Chips Shop	Zoo
190	Etobicoke	4	Coffee Shop	Italian Restaurant	Sushi Restaurant	Bank	Park	Pub	Breakfast Spot	Bar	Bakery	Eastern European Restaurant
167	Etobicoke	4	Park	Italian Restaurant	Bank	Indian Restaurant	Sushi Restaurant	Breakfast Spot	Café	Pizza Place	Grocery Store	Bar
18	Etobicoke	4	Coffee Shop	Italian Restaurant	Sushi Restaurant	Bank	Pub	Grocery Store	Bakery	Restaurant	Fast Food Restaurant	Breakfast Spot

This cluster include residential areas with parks, verious restaurants, spas. So it's 'Middle Cost Residential Areas'.

Cluster 6

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
88	North York	5	Electronics Store	Park	Sports Bar	Mexican Restaurant	Coffee Shop	Bank	Financial or Legal Service	Event Space	Falafel Restaurant	Farm

This cluster are riddlest. One of the common venues is mexican restaurant and there are bar, store, so called it 'Latino-Americans Area'.

Cluster 7

	Borough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
57	East York	6	Greek Restaurant	Park	Café	Italian Restaurant	Pub	Bakery	Coffee Shop	Pizza Place	Burger Joint	Flower Shop
147	Downtown Toronto	6	Coffee Shop	Café	Bar	Caribbean Restaurant	Beer Bar	Mexican Restaurant	Vegetarian / Vegan Restaurant	Restaurant	Art Gallery	Italian Restaurant
49	West Toronto	6	Café	Coffee Shop	Korean Restaurant	Park	Grocery Store	Bar	Italian Restaurant	Bakery	Cocktail Bar	Middle Eastern Restaurant
74	West Toronto	6	Coffee Shop	Café	Restaurant	Park	Bakery	Bar	Gift Shop	Italian Restaurant	Theater	Dog Run
138	Downtown Toronto	6	Café	Cocktail Bar	Bar	Grocery Store	Dessert Shop	Coffee Shop	Italian Restaurant	Beer Bar	Pub	Vegetarian / Vegan Restaurant
159	Downtown Toronto	6	Coffee Shop	Café	Yoga Studio	Bakery	Italian Restaurant	Spa	Dessert Shop	Beer Bar	Caribbean Restaurant	Sandwich Place
73	West Toronto	6	Coffee Shop	Restaurant	Park	Bakery	Café	Indian Restaurant	Gift Shop	Bar	Italian Restaurant	Tibetan Restaurant

The last cluster are closest to the city center, but it is residental area. So called cluster at 'Center Residential Areas'.

5 Conclusion

Main purpose of this project is **segmenting and clustering neighborhoods in Toronto City**, based on a **real data**, which parsing on internet, using **BeautifulSoup** library.

Using **Nominatim search engine**, were receiving coordinates for each neighborhoods and after that, using **Forsquare API**, for each neighborhoods were find **top 10 most common venues** within a radius. Further, top venues transfer to ***k*-means algorithm** as a features, and using **GridSearchCV** library, was find **7 clusters** with the best accuracy. In 'Examine Clusters' sections each clusters was described and labeled.

In result there are clusters:

- **'Coasts Areas'**
- **'Chinatown'**
- **'Low Cost Residential Areas'**
- **'Middle Cost Residential Areas'**
- **'City Center'**
- **'Latino-Americans Area'**
- **'Center Residential Areas'**