Applying Data Science Methodology to Discover Business Opportunities in Toronto, Ontario, Canada

Qusay Sellat

Nov, 2019

1. Introduction

Toronto is the provincial capital of Ontario and the most populous city in Canada. It is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

However, it seems to be very hard for people outside the city to define the main characteristics of each neighborhood inside Toronto that distinguish it from the other neighborhoods.

1.1. Problem

This project analyses the data of Toronto's various neighborhoods in order to discover the best businesses that can be done in each of the respective neighborhoods by doing a cluster analysis.

1.2. Interest

Some people outside Toronto have plans to invest in the city but they don't know how and where. Therefore, this project can be helpful for those who have dreams of having a successful business into the city of Toronto.

2. Data Collecting and Pre-processing

2.1. Data Collecting

The data needed for the clustering of Toronto's neighborhoods according to their characteristics can be collected from many sources. We store collected data into Pandas dataframes. The data used in this project comes from the following sources:

- Data about Toronto's postcodes, boroughs, and neighborhoods. This data is read using Pandas library by scraping it from wikipedia url. In this report we will call this data **Neighborhoods** dataset.

	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront

- Data about the coordinates (latitude and longitude) of various Toronto's neighborhoods. This data can be collected using Geocoder Python package. However, This package is highly unreliable and I couldn't use it to download the data. Fortunately, Coursera provided the data via a reliable link. In this report we will call this data **Coordinates** dataset.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

- Data about the various types of venues located in each of the neighborhoods represented in the data collected in the above steps. For the purpose of collecting this data, we use FourSquare API. In this report we will call this data **Venues** dataset.

	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	TTC stop #8380	43.752672	-79.326351	Bus Stop
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

2.2. Data Pre-processing

In order to benefit from the collected data, we must have some processing done.

- First of all, we notice that some cells under Neighborhoods dataset are Not assigned. For convenience, any row containing Not assigned Borough was dropped. Also, any Not assigned Neighborhood is replaced by the corresponding Borough.

	Postcode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights
4	M6A	North York	Lawrence Manor

- Then we prepare Neighborhoods data set to contain latitude and longitude information by using Coordinates data set. The final **Neighborhoods Coordinates** dataset looks like this:

	Postcode	Borough	Neighborhood	latitude	longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights	43.718518	-79.464763
4	M6A	North York	Lawrence Manor	43.718518	-79.464763

- In order to do the clustering process on the neighborhoods of Toronto, we must first make a dataset that can be fit into a clustering algorithm like KMeans. One approach is to use **one-hot encoding** to represent each of the returned avenues with the corresponding neighborhoods. So first we convert to one-hot encoding:

	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Argentinian Restaurant
0	Parkwoods	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
2	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0
3	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0
4	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0

- We notice that the number of avenues is 4379 of 270 type. For each neighborhood, in order to fit the KMeans algorithm, it's important to know what avenues located in each neighborhood. For this reason we group the one-hot encoded dataset of the avenues by the neighborhood by applying the mean function to represent the importance of each avenue in describing the neighborhood. So the final dataset looks like the following, we will call it **toronto_grouped** dataset (because of sparsity nature of data, most entries are zero), and it will be the input of our clustering algorithm.

	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Argentinian Restaurant
0	Adelaide	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.030000	0.0	0.0	0.0
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
2	Agincourt North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
3	Albion Gardens	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
4	Alderwood	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
201	Woodbine Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
202	York Mills	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
203	York Mills West	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
204	York University	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
205	Yorkville	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.047619	0.0	0.0	0.0

206 rows × 271 columns

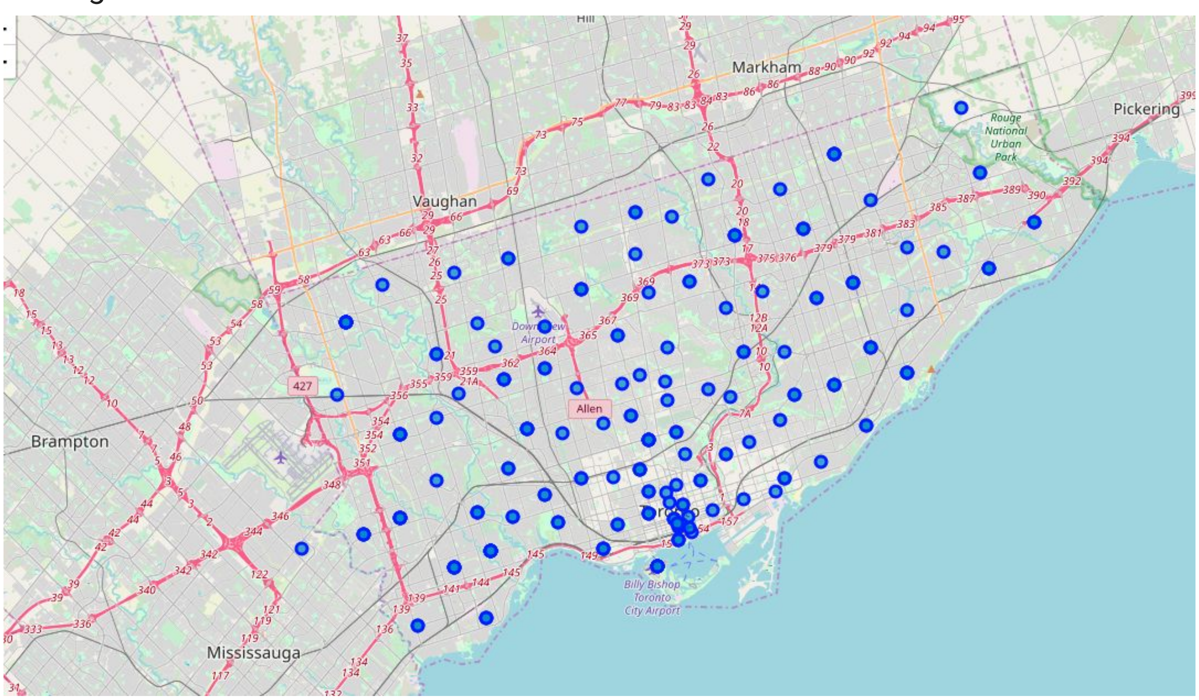
3. Exploratory Data Analysis

In this section, let's examine the statistical characteristics of the datasets we have created.

3.1. Neighborhoods of Toronto

As shown in the code, the 210 neighborhoods of Toronto belong to 11 boroughs and there are 103 distinct postcodes for them. The coordinates (latitude and longitude) of those postcodes are read into coordinates dataset and used to fill the coordinates in neighborhoods dataset.

By drawing these neighborhoods on an interactive map using folium library we got the following:



3.2. Venues of Toronto

As shown in the code there are 4379 venues of 270 types. Different types of restaurants, playgrounds, bars, shops, transportation stations, .. etc. We also notice that venues belong to 206 neighborhoods - 4 neighborhoods less than we have in neighborhoods dataset. This is due to the fact that FourSquare API didn't retrieve any venues for some neighborhoods. However, the number is very small and will not affect the final result that much and we can continue with the dataset we obtained.

The table below shows the most frequent avenues to be present in Toronto as retrieved from the avenues dataframe:

Coffee Shop	340
Café	192
Restaurant	119
Pizza Place	112
Bakery	109
Bar	102
Italian Restaurant	91
Park	90
Sandwich Place	79
Hotel	77
Fast Food Restaurant	71
Clothing Store	64
Japanese Restaurant	61
American Restaurant	59
Gym	57
Sushi Restaurant	53
Burger Joint	53
Pharmacy	50
Grocery Store	46
Pub	45

In the initial one-hot dataset, we can imagine that each of the 4379 venues is represented by a 270 vector of 0s and 1s (to indicate venue type) along with an entry filled with the name of the neighborhood the venue belongs to.

The final dataset is formed by grouping the initial one-hot dataset using the neighborhood attribute and mean method applied. This resulted in a list of each neighborhood in Toronto with its content of venues represented as a number between 0 and 1 - the higher the number, the more the neighborhood contains the respective avenue.

We also presented the most common avenues for some of the neighborhoods to get more insight about the distribution:

	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	Adelaide	Coffee Shop	Café	Thai Restaurant	Bar	Steakhouse	Sushi Restaurant	Restaurant	Burger Joint	Bakery	Cosmetics Shop
1	Agincourt	Lounge	Chinese Restaurant	Sandwich Place	Latin American Restaurant	Breakfast Spot	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop
2	Agincourt North	Playground	Park	Yoga Studio	Donut Shop	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Doner Restaurant
3	Albion Gardens	Pharmacy	Sandwich Place	Fast Food Restaurant	Beer Store	Fried Chicken Joint	Grocery Store	Pizza Place	College Stadium	Department Store	Eastern European Restaurant
4	Alderwood	Pizza Place	Pub	Gym	Coffee Shop	Pharmacy	Sandwich Place	Skating Rink	Pool	Yoga Studio	Deli / Bodega
5	Bathurst Manor	Coffee Shop	Supermarket	Pizza Place	Deli / Bodega	Bank	Sushi Restaurant	Sandwich Place	Fried Chicken Joint	Frozen Yogurt Shop	Middle Eastern Restaurant
6	Bathurst Quay	Airport Service	Airport Lounge	Airport Terminal	Harbor / Marina	Boat or Ferry	Bar	Airport	Airport Food Court	Airport Gate	Sculpture Garden
7	Bayview Village	Café	Bank	Chinese Restaurant	Japanese Restaurant	Department Store	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Doner Restaurant
8	Beaumond Heights	Pharmacy	Sandwich Place	Fast Food Restaurant	Beer Store	Fried Chicken Joint	Grocery Store	Pizza Place	College Stadium	Department Store	Eastern European Restaurant
9	Bedford Park	Coffee Shop	Italian Restaurant	Pub	Pizza Place	Indian Restaurant	Café	Sushi Restaurant	Butcher	Liquor Store	Fast Food Restaurant

We may now have a deeper sense that the various kinds of restaurants dominates the venue list. This is a sign that one of our clusters (the cluster which have the neighborhoods with a large number of restaurants like avenues) will dominate other clusters in number of avenues.

4. K-Means Clustering

In order to give people good information about the characteristics of each neighborhood, it's good to cluster the neighborhoods into groups and find the characteristics of each group and try to generalize those characteristics to each neighborhood in the cluster.
We will consider the data frame we built (toronto_grouped) as the input of the K-Means algorithm.
We can fit and display algorithm in one line of code for each thanks to Sklearn library.

```
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

In next steps, we added clustering results to the former created dataframes to result in new dataframes :

- **neighborhoods_venues_sorted** : we will alter it to also contain cluster labels.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Co V
0	0	Adelaide	Coffee Shop	Café	Thai Restaurant	Bar	Steakhouse	Resta
1	0	Agincourt	Lounge	Chinese Restaurant	Sandwich Place	Latin American Restaurant	Breakfast Spot	
2	2	Agincourt North	Playground	Park	Yoga Studio	Donut Shop	Dessert Shop	Dim Resta
3	0	Albion Gardens	Pharmacy	Sandwich Place	Fast Food Restaurant	Beer Store	Fried Chicken Joint	Gr
4	0	Alderwood	Pizza Place	Pub	Gym	Coffee Shop	Pharmacy	San

- **toronto_merged** : this will be derived from merging neighborhoods dataframe with neighborhoods_venues_sorted dataframe. It contains neighborhoods information along with cluster label and 10 most frequent venues in each neighborhood.

	Borough	Neighborhood	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th C
0	North York	Parkwoods	43.753259	-79.329656	6	Bus Stop	Park	Food & Drink Shop	Yoga
1	North York	Victoria Village	43.725882	-79.315572	0	Intersection	Coffee Shop	Portuguese Restaurant	F Rest
2	Downtown Toronto	Harbourfront	43.654260	-79.360636	0	Coffee Shop	Pub	Bakery	
3	North York	Lawrence Heights	43.718518	-79.464763	0	Accessories Store	Coffee Shop	Shoe Store	Miscella
4	North York	Lawrence Manor	43.718518	-79.464763	0	Accessories Store	Coffee Shop	Shoe Store	Miscella

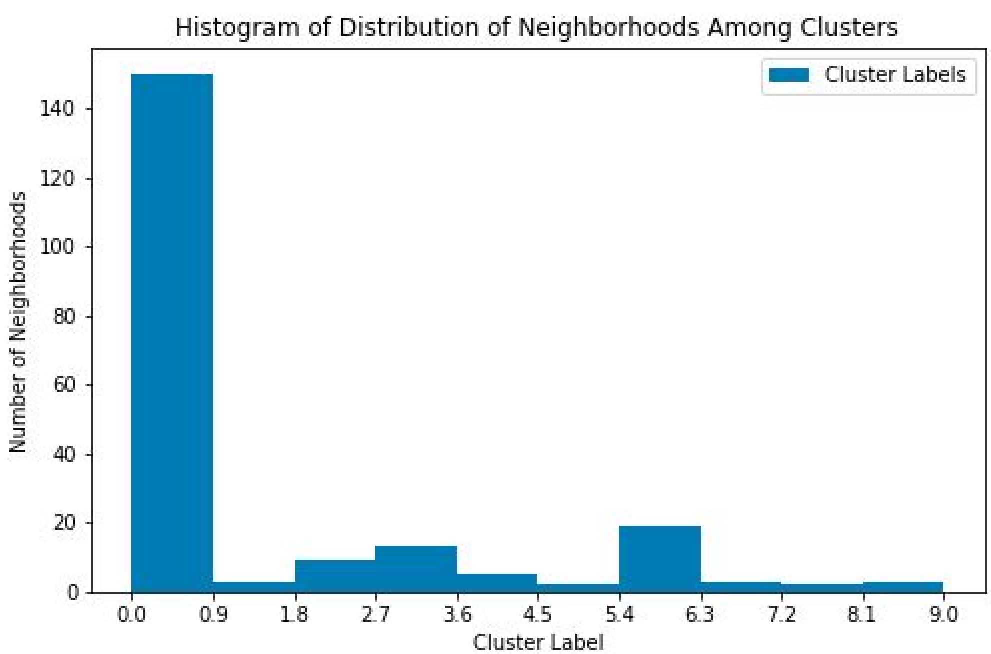
- **venue_clusters** : this will be derived from venues dataframe. For each venue it contains the cluster label for the respective neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park	6
1	Parkwoods	43.753259	-79.329656	TTC stop #8380	43.752672	-79.326351	Bus Stop	6
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop	6
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena	0
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop	0
5	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant	0
6	Victoria Village	43.725882	-79.315572	The Frig	43.727051	-79.317418	French Restaurant	0

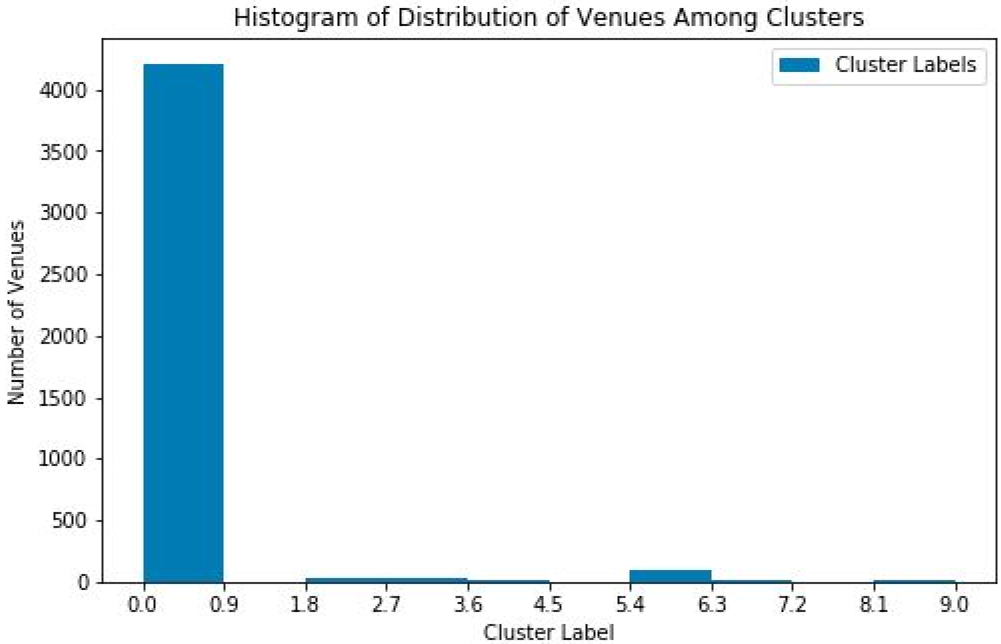
5. Result Analysis

In this section, we have to give some analysis of the clustering process we have done. First we have to get an idea of the distribution of neighborhoods and venues among clusters. We will do some plotting to notice this distribution. Then we will use folium library to get an idea of the geographical distribution of different clusters.

The number of neighborhoods in each cluster is found to be as shown in the following histogram:

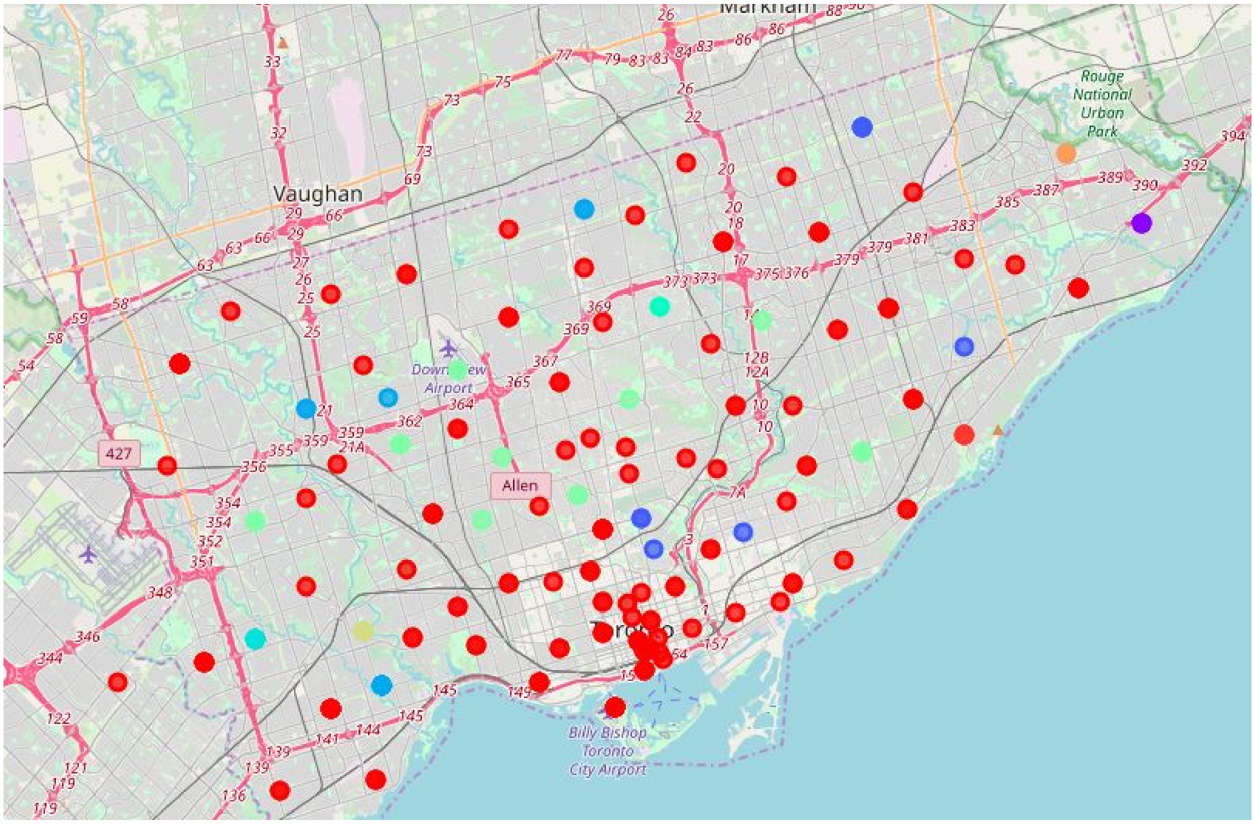


The number of venues in each cluster is found to be as shown in the following histogram:



As we can see cluster 0 has the biggest number of neighborhoods and venues. This is an indication that Toronto city is very homogeneous in nature of venues (most are some kind of restaurants). However, there are some neighborhoods that belong to other clusters. We will have a deeper look at this data later.

Using folium library, let's get an idea of the geographical distribution of different clusters:



Note: Deeper illustrations of what venues each cluster is responsible to contain are shown in the ipynb file. They are rather long to be illustrated here.

6. Conclusion and future work

Basic clustering process clearly finds that most of Toronto's neighborhoods are homogeneous and don't have that much diversity in the nature of venues. Perhaps, in the future, more data will result in more diversity, especially if it took into consideration the demographics of those neighborhoods.