Exploring the neighborhoods of Toronto with location data – a tourist's perspective

Lifu Deng

Introduction

Toronto is at the heart of the most vivid and populous metropolitan area in Canada, an international center of business, finance, arts, and culture. It is recognized as one of the most multicultural and cosmopolitan cities in the world and is a great destination of tourists [1][2][3][4]. This means that while people from around the world visit Toronto for different purposes, many of them are happy to spend some spare time touring around the city. For example, I can imagine myself attending a psychology conference someday (post-pandemic, of course) at University of Toronto, and I will have a few hours free to spend after the conference ends, during which I can wander around in the neighborhood near the hotel before heading to the airport.

Such kind of brief city trips will perhaps have at least one of the following characteristics: 1) they are often without specific destination; 2) they are like mini explorations, where the variety of places visited can significantly add to the fun of such experience; and 3) the tourists may walk a lot instead of driving or taking public transportation, because they are not really familiar with the city. While mobile apps such as Google Map and Yelp may offer great recommendations of places to go, the information are often non-quantitative and lack big picture of the city. Therefore, **the overarching purpose of this project is to provide some quantitative information from public data sources to make such brief trips in Toronto more fun and enjoyable**.

More specifically, the goal of this project can be broken down to several questions.

- What are the most common types of places/venues in Toronto?
- Which parts of the city contain a high density and a good variety of venues to explore?

- Can different neighborhoods be categorized into several different groups, with each demonstrating a unique combination of venues?

To answer these questions, let's take a glance at the data we have in the next section.

Data description

This section provides the overview of the data used in the project, while the detailed approaches of data acquisition can be found in the ***Methods*** section.

The neighborhoods of Toronto are distributed across multiple urban and suburban boroughs and are labelled with unique 3-digit postal codes. A complete list of neighborhoods can be found in Wikipedia [5] or Canada Post website. In this project, the data of neighborhood is loaded from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M using Pandas 1.0.4 package in Python 3.6. Because in our scenario the tourists are likely staying in the downtown area, I excluded the neighborhoods in the outskirts of the city by keeping only the neighborhoods in boroughs with key word 'Toronto' in their names. This results in 39 neighborhoods with unique postal codes located within four boroughs, namely, East Toronto, West Toronto, Central Toronto, and Downtown Toronto. The postal codes are associated with geospatial coordinates, which are loaded from the follow link http://cocl.us/Geospatial_data.

Given the geospatial coordinates, the places (venues) of the neighborhoods can be found using **Foursquare** API. Here, search requests are defined as the inquiries into venues near the location within a 500-meter radius and are sent to Foursquare and a limit of 150 venue records. Foursquare then returns information in JSON format, which includes all known nearby venues, specified with their names, categories, and geospatial coordinates. 1,618 different venues in 234 categories are obtained as a result.

Methods

The following analyses are performed Python 3.6 run on Jupyter Notebook 6.0.3. The code can be found in my github (https://github.com/FugaDeng/Coursera_Capstone).

The neighborhoods of Toronto are loaded using Pandas.read_html and Pandas.read_csv methods. Postal codes associated with 'Not assigned' boroughs or neighborhoods are excluded. The result is shown as follow:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Total_venues	Venue_types
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	46	30
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	33	26
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	100	62
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	81	52
4	M4E	East Toronto	The Beaches	43.676357	-79.293031	4	4

Requests to Foursquare are sent in the format of URLs, which contains information such as the client ID and credential, geospatial coordinates, the radius of search, and the limit of maximum records that Foursquare can return. The result is shown as follow:

	PostalCode	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M5A	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	M5A	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	M5A	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	M5A	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	M5A	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant

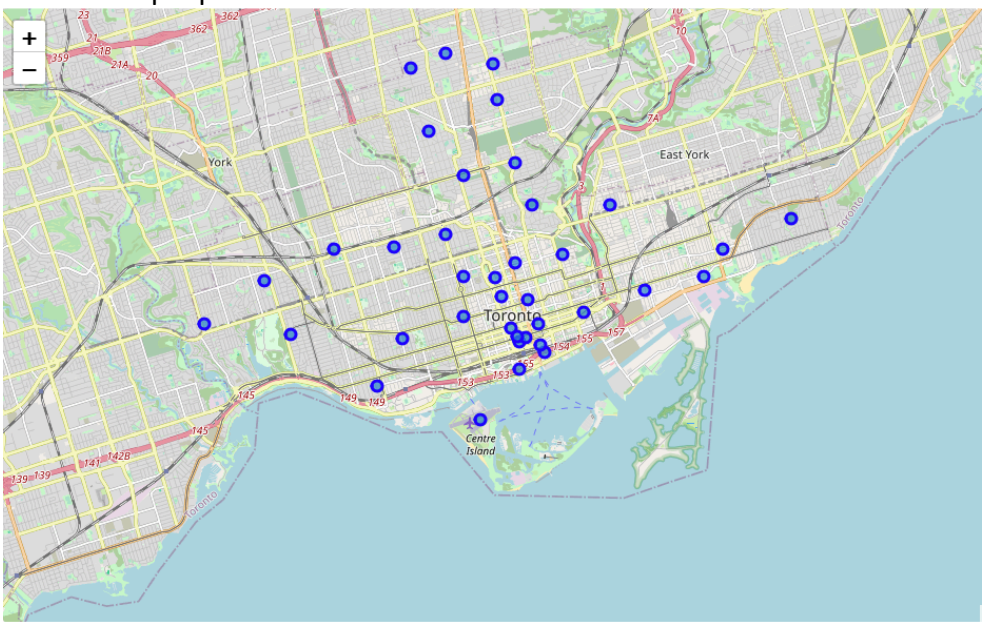
To visualize the data, several Python packages are used, including:

- Matplotlib 3.2.1 and seaborn 0.10.0, for plotting bar charts, scatter plots, histograms, and regression lines
- Folium 0.5.0, for visualizing neighborhoods and their associated values onto the city map
- Wordcloud 1.4.1, for turning the frequently-appearing venue types into word clouds.

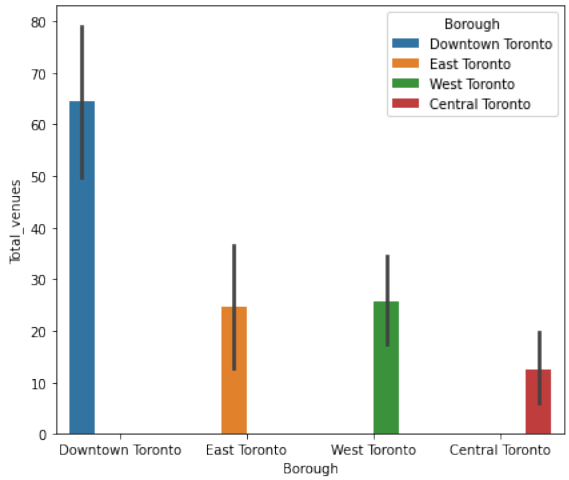
In order to find the similarities between neighborhoods and sort them into several ***similar groups*** in a data-driven manner, I will use the K-means clustering algorithm implemented in scikit-learn 0.22.1. The feature vector of each neighborhood is defined as the numbers of venue in each category.

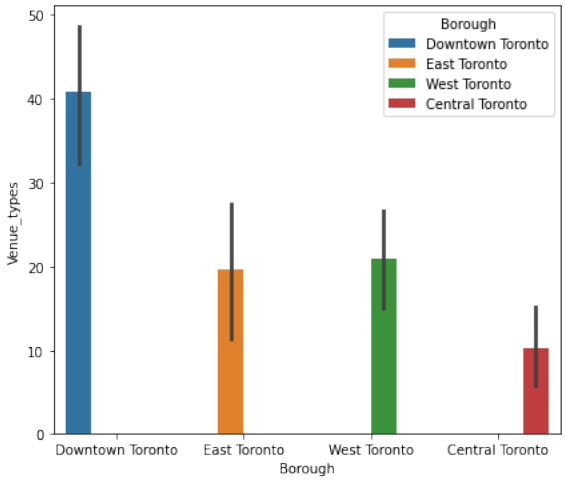
Results

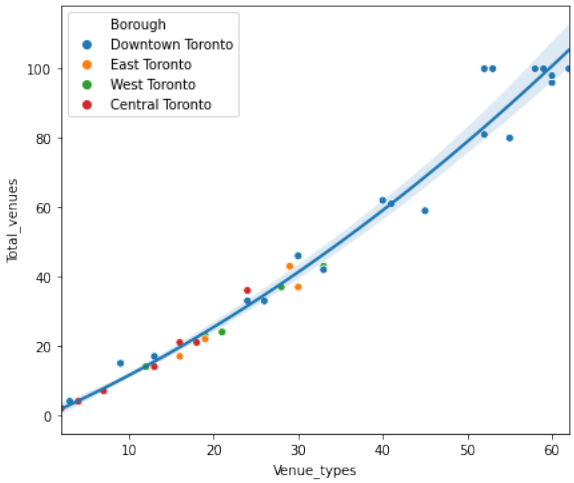
First off, let's take a glance at the geographical distribution of the neighborhood in the figure below, rendered using Folium. Each blue dot represents the center of one neighborhood associated with a unique postal code.

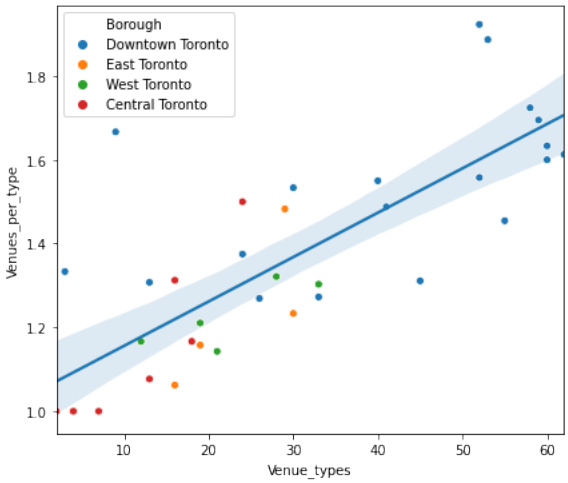


Turning to the number of venues, we can calculate the **number of venues** for each neighborhood and the **number of venue types**. The two measures are averaged according to the boroughs and plotted in the top-left and top-right panel below. Note that the numbers of venues of different neighborhoods are only the venues falling into the 500-meter search circle rather than all venues in that neighborhood. Therefore, this value in fact reflects the **spatial density** of venues, indicating whether the neighborhood is located in a business district or buzzy downtown area. Indeed, Downtown Toronto is the borough with highest venue density AND venue variety – an exciting area to explore! In the bottom left panel, each datapoint represents one neighborhood. We can see that venue density and venue variety are positively related and can be predicted almost perfectly using a second-order quadratic function. To go a step further, such non-linear relationship is explained by the fact shown in the bottom right panel: as the venue variety (i.e., Venue_type) increases in one neighborhood, it is more likely to see more than one similar venue near it. This observation may have important implication regarding customer choice and business competition.

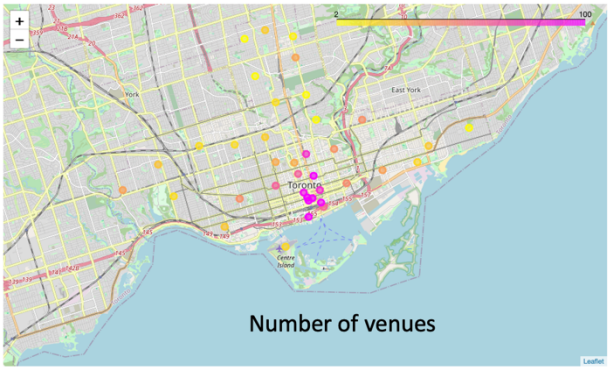


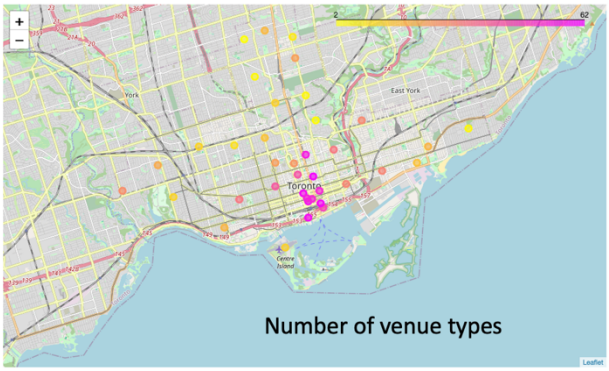




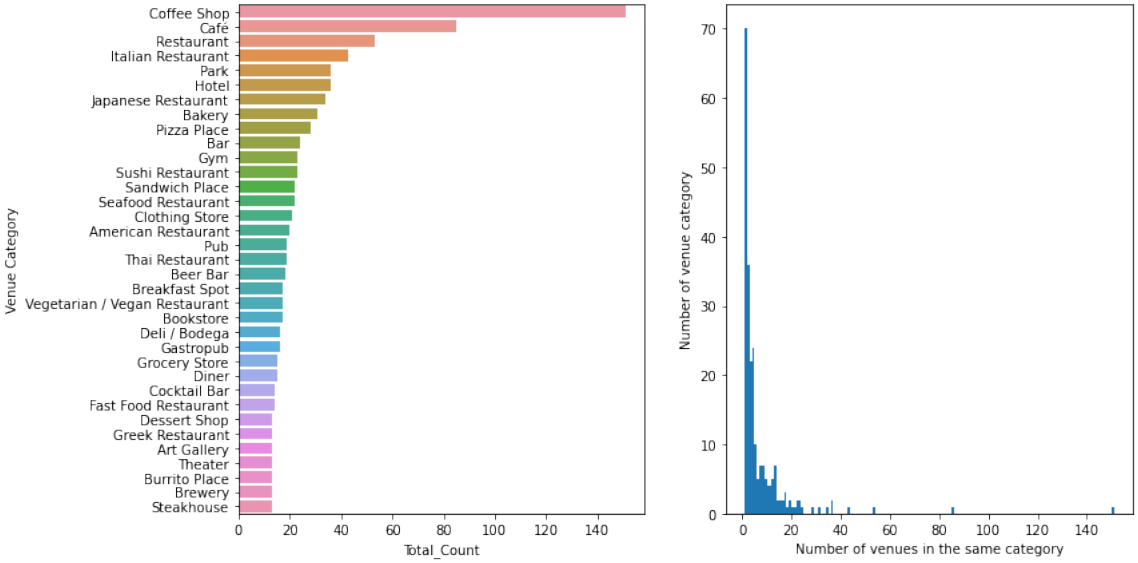


The spatial density (i.e. number of venues) and the variety (i.e. number of venue types) across different neighborhoods are visualized in the two maps below. As we can see, the two measures are spatially correlated. **As an interim summary, Downtown Toronto is a great tourist attraction due to its high density and variety of places one can visit.**





Next, we can take a closer look at the types of venues in Toronto. In the bottom left panel, the counts of the top 35 most common venues across all neighborhoods are plotted. Not surprisingly, coffee shop comes to the top of the list. **Among the popular venues are also many types restaurants offering different kinds of cuisine, which indicates the multicultural characteristic of the city.** On the right, a histogram shows that many venue types only have a very limited number of venues.

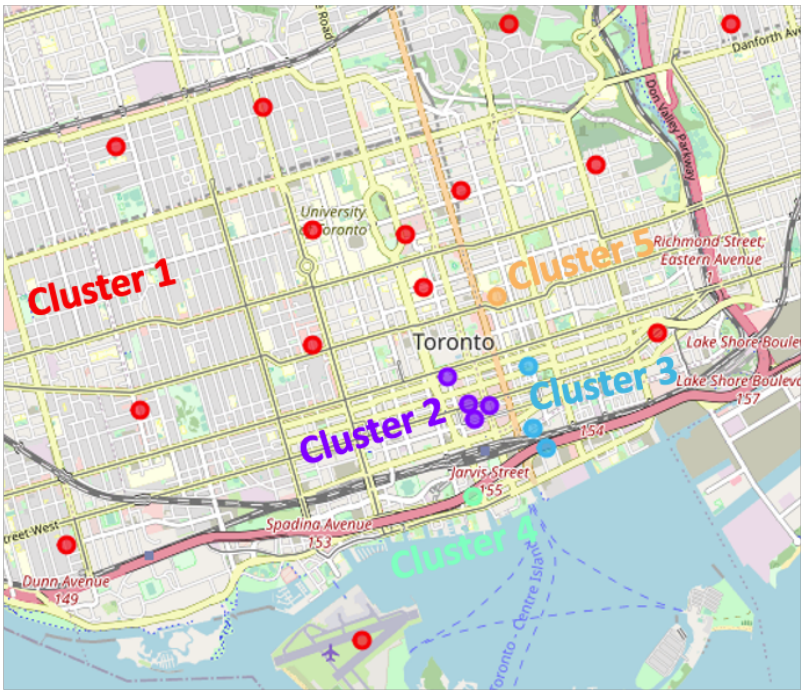


We can count the number of venues in each unique category for each unique neighborhood, as is shown below. In fact, each column can serve as a feature vector of the corresponding neighborhood. Intuitively, if two neighborhoods have similar venue composition (i.e., feature vectors), they can be considered as the 'same' group of neighborhoods. In order to assign neighborhoods into different groups, we can feed the feature vectors into the k-means clustering algorithm. Here, I remove the 'Coffee Shop' entries of the feature vectors, because coffee shop is ubiquitous and does not tell us about the uniqueness of the neighborhoods.

Venue Category	Total_Count	M5A	M7A	M5B	M5C	M4E	M5E	M5G	M6G	...	M4T	M5T	M4V	M5V	M4W	M5W	M4X	M5X	M4Y	M7Y
Coffee Shop	151	7	7	10	6	0	5	11	1	...	0	4	2	1	0	11	3	10	6	0
Café	85	3	1	4	4	0	1	3	3	...	0	5	0	0	0	3	3	7	2	0
Restaurant	53	1	1	1	2	0	2	1	1	...	1	0	1	0	0	3	2	4	3	1
Italian Restaurant	43	0	1	2	2	0	1	3	1	...	0	0	0	0	0	4	2	1	1	0
Park	36	3	1	1	2	0	1	1	2	...	1	2	0	0	2	2	1	0	1	1
Hotel	36	1	0	3	2	0	1	0	0	...	0	0	0	0	0	2	0	5	2	0
Japanese Restaurant	34	0	1	3	2	0	1	1	0	...	0	1	0	0	0	3	1	4	5	0
Bakery	31	3	0	1	1	0	3	0	0	...	0	1	0	0	0	3	2	1	0	0
Pizza Place	28	0	0	2	0	0	0	0	0	...	0	1	1	0	0	0	2	2	2	1
Bar	24	0	1	0	0	0	0	1	0	...	0	2	0	0	0	0	0	2	0	0

For the k-means algorithm, I set k=5 and ran the algorithm 10 times, as k-means is non-deterministic. However, in our case, the algorithm **produced consistent clustering results**, as is shown below. The most significant observation is that neighborhoods outside Downtown Toronto are so similar that they have been group into a single cluster (shown in red).

Downtown neighborhood, in comparison, are group into 4 different clusters, indicating the variety of neighborhood types. **This also sounds like a good news for tourists, as they only need to walk within a short distance in order to experience the many different facets of the Toronto city.**



Now, let's visualize the most popular venues in each neighborhood cluster. The following word clouds highlight the most common venue types in each neighborhood cluster, where larger text indicates higher frequency. In order to make the word clouds more representative, I excluded words that are not so informative ('Coffee', 'Shop', 'Restaurant', 'Café'). Hopefully this may help a tourist get a better idea where he/she wants to go for a half-day tour in the city of Toronto!


Cluster 1
Store
Bar
Park
Italian
Pizza
Place
Bakery
Sushi
Japanese
Pub
Sandwich Place
Cluster 2
American
Hotel
Place
Gym
Bar
Seafood
Japanese
Salad Place
Steakhouse
Concert Hall
Bodega
Deli
Cluster 3
Store
Cocktail
Bar
Seafood
Cheese
Italian
Bakery
Food
Japanese
Farmers Market
Creperie
Pharmacy
Cluster 4
Bar
Hotel
Aquarium
Place
Stadium
Lookout
Scenic
Fried
Brewery
Venue
Chicken
Italian
Cluster 5
Clothing Store
Store
Place
Cosmetics
Theater
Japanese
Bubble
Middle
Eastern
Lingerie
Italian
Tea

Discussion

Downtown Toronto is a great tourist attraction due to its high density and variety of places one can visit. The city seems quite pedestrian-friendly, as a variety of venues can be found within a small range of area. The types of venues reveal the multicultural character of the city. There are some reliable differences between neighborhoods, which make them fall into different categories. The exploration of neighborhoods can be guided by the frequently appearing venues. However, if you are not picky about the places you want to visit, there are always a lot of coffee shops for you ☺

The project may benefit from refinements in the following aspect. Firstly, the sampling of venue may not be adequate for the neighborhoods located in the suburban areas. Secondly, the search area may overlap for adjacent downtown neighborhoods, suggesting the use of an adaptive search radius in Foursquare API. Lastly, it is worth exploring other types of data Foursquare can offer, such as customer ratings.

Conclusion

As the availability of geolocation data has been increasing over the recent years, analyzing such data can bring new understanding to problems we frequently encounter. Helping people to decide a part of the city to explore is one example of its application. Even when the available data are relatively simple and limited (as in this case), the observation can still be exciting, insightful, and persuasive, when combined with proper presentation of data visualization.

References

- 1. https://en.wikipedia.org/wiki/Toronto
- 2. https://books.google.com/books?id=_p7CDgAAQBAJ&pg=PP147#v=onepage&q&f=false
- 3. https://books.google.com/books?id=uifwpL0qZ_EC&pg=PA3#v=onepage&q&f=false
- 4. https://books.google.com/books?id=WhtwAgAAQBAJ&pg=PA163#v=onepage&q&f=false
- 5. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M