The Battle of Neighborhoods: Part I & II

IBM Data Science Professional Certificate: Capstone Project

Authored by Hritik Panchasara on March 31, 2020

1. Introduction and Business Problem

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 as of 2016. There are a variety of attractions and venues, and the city of Toronto has a lot of them. The distribution of said venues can be mapped out across neighborhoods in different boroughs. Instead of localizing this project to one stakeholder, such as a restaurant, we can address a multitude of prospective issues. If someone is looking to open a restaurant, where could they open it? Similarly, if a contractor is trying to start their own business, where could they setup their office?

We can harness Data Science techniques and data sets provided by the city of Toronto to provide prospective stakeholders with enough information in order for them to make better decisions with choosing the vest venue category for the best localities. Common venue types in neighborhoods will be analyzed, and this will help in finding the optimal venue category that is best suited in the best neighborhoods of Toronto.

2. Data Sources and Formatting

For analyzing neighborhood data in Toronto, we need to explore and cluster neighborhood data sets from Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). The python library pandas will be used to read the table into a pandas data frame. Geographical information such as latitude and longitude information can be accessed from the csv file uploaded online (https://cocl.us/Geospatial_data) using pandas data frame. The FourSquare API will be harnessed to get venue details in the city and help categorize them before clustering the neighborhoods.

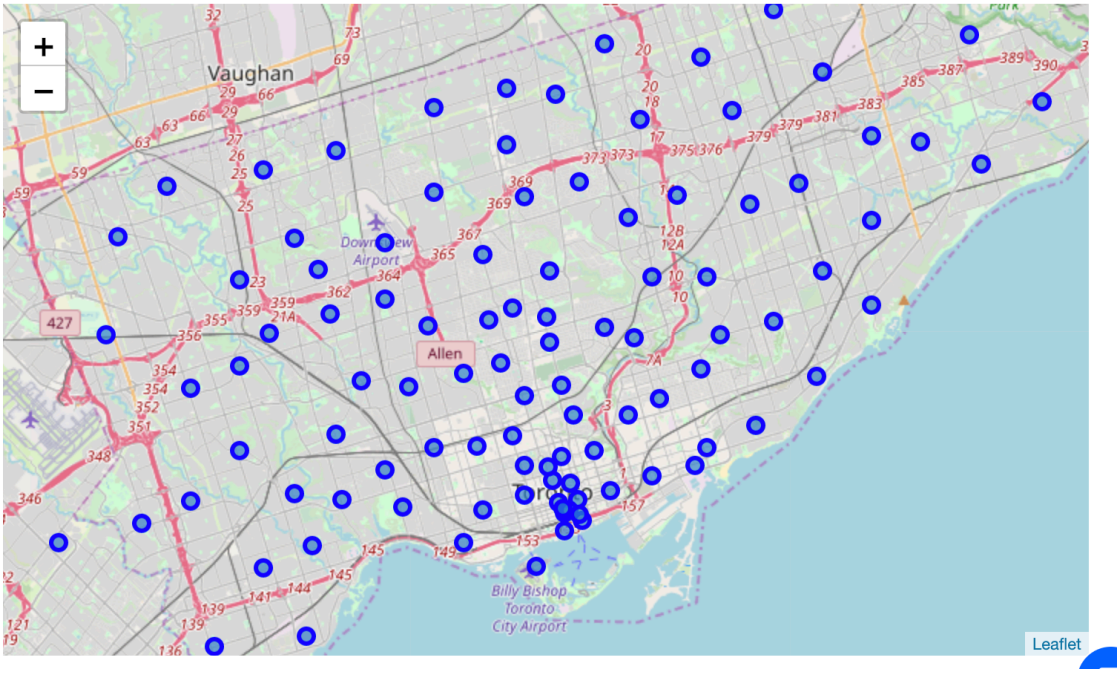
Data will be captured and formatted into a single table. Some information related to neighborhoods and boroughs are missing or multiple copies exist that will be eradicated or combined into single tuples to improve data standards. These tuples will be separated using commas and the 'groupby' function.

3. Methodology

Detecting the most common type of venue for each neighborhood in different boroughs in the main approach here. First, data needs to be collected from the Wikipedia page and geographical coordinates needs to be assigned to each location. Next, once the venue details are collected, the folium library will be used to analyze promising areas

according to ratings and frequency against most common areas. The final step will include creating clusters using the k-means clustering algorithm.

The dataframe generated analyzed 10 boroughs and 103 neighborhoods. The geographical coordinates of Toronto are 43.6534817, -79.3839347. The folium generated map of Toronto is displayed below:

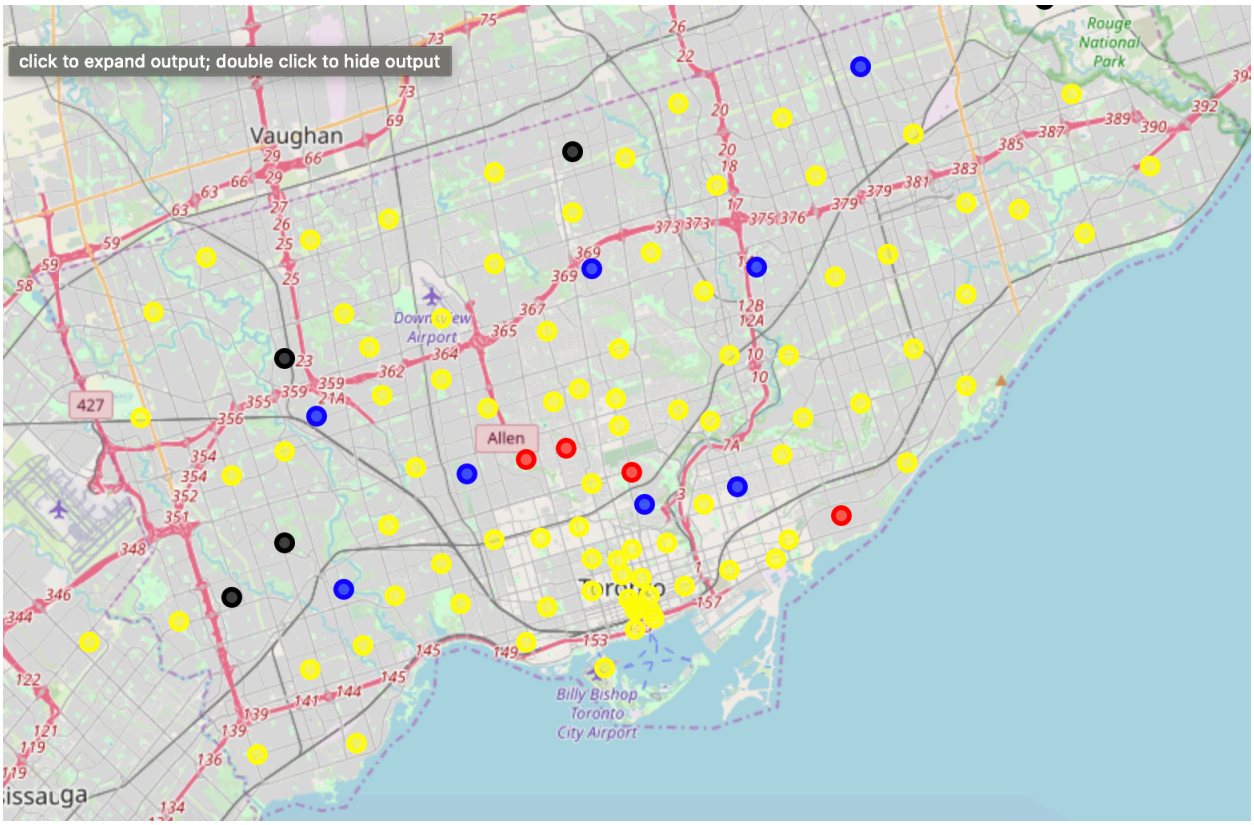


Leveraging the FourSquare API provides us with neighborhood data and venue information. For neighborhood Rogue Hill, Port Union, and Highland Creek, we can differentiate the different venues with their categories and geolocations as shown below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern / Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill / Port Union / Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Rouge Hill / Port Union / Highland Creek	43.784535	-79.160497	Scarborough Historical Society	43.788755	-79.162438	History Museum
3	Guildwood / Morningside / West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store
4	Guildwood / Morningside / West Hill	43.763573	-79.188711	Marina Spa	43.766000	-79.191000	Spa

Next, we create another dataframe to display frequently occurring venue types for each neighborhood. Some neighborhoods have the same common frequenting venues in its top 10 shortlist, whereas for some that isn't the case.

The K-means clustering algorithm model aims at portioning *n* observations into *k* clusters having the nearest mean values. The number of clusters taken are 3 for this case scenario. The clusters are then visualized using the folium library. Cluster 1 is represented by the red color, cluster 2 is represented by the blue color, and cluster 3 is represented by the yellow color. Cluster 1 is smaller than cluster 2 and cluster 3. Cluster 2 is larger than cluster 1, and cluster 3 is the largest.



We can derive from cluster 1 that the most common venue type or category is

	Borough	Cluster Labels BON	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
14	Scarborough	0.0	1.0	Park	Playground	Doner Restaurant	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Donut Shop
23	North York	0.0	1.0	Park	Bank	Convenience Store	Drugstore	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant
25	North York	0.0	1.0	Park	Food & Drink Shop	Convenience Store	Women's Store	Donut Shop	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant
40	East York	0.0	1.0	Park	Coffee Shop	Convenience Store	Empanada Restaurant	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Dessert Shop	Drugstore
50	Downtown Toronto	0.0	1.0	Park	Playground	Trail	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Ethiopian Restaurant	Dumpling Restaurant	Drugstore	Department Store
74	York	0.0	1.0	Park	Women's Store	Market	Doner Restaurant	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Drugstore
90	Etobicoke	0.0	1.0	Smoke Shop	River	Park	Women's Store	Dog Run	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Donut Shop
98	York	0.0	1.0	Convenience Store	Park	Drugstore	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Women's Store

We can derive from cluster 2 that the most common venue type or category is

	Borough	Cluster Labels BON	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
14	Scarborough	0.0	1.0	Park	Playground	Doner Restaurant	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Donut Shop
23	North York	0.0	1.0	Park	Bank	Convenience Store	Drugstore	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant
25	North York	0.0	1.0	Park	Food & Drink Shop	Convenience Store	Women's Store	Donut Shop	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant
40	East York	0.0	1.0	Park	Coffee Shop	Convenience Store	Empanada Restaurant	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Dessert Shop	Drugstore
50	Downtown Toronto	0.0	1.0	Park	Playground	Trail	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Ethiopian Restaurant	Dumpling Restaurant	Drugstore	Department Store
74	York	0.0	1.0	Park	Women's Store	Market	Doner Restaurant	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Drugstore
90	Etobicoke	0.0	1.0	Smoke Shop	River	Park	Women's Store	Dog Run	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Donut Shop
98	York	0.0	1.0	Convenience Store	Park	Drugstore	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Women's Store

Finally, we can derive from cluster 3 that the most common venue type or category is

	Borough	Cluster Labels BON	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
0	Scarborough	1.0	2.0	Fast Food Restaurant	Dessert Shop	Event Space	Ethiopian Restaurant	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Donut Shop
1	Scarborough	1.0	2.0	History Museum	Bar	Drugstore	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Women's Store	Dim Sum Restaurant
2	Scarborough	1.0	2.0	Electronics Store	Rental Car Location	Mexican Restaurant	Bank	Intersection	Moving Target	Medical Center	Spa	Breakfast Spot	Donut Shop
3	Scarborough	1.0	2.0	Coffee Shop	Korean Restaurant	Women's Store	Drugstore	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Eastern European Restaurant
4	Scarborough	1.0	2.0	Fried Chicken Joint	Gas Station	Hakka Restaurant	Bank	Athletics & Sports	Caribbean Restaurant	Thai Restaurant	Bakery	Dim Sum Restaurant	Diner
5	Scarborough	1.0	2.0	Grocery Store	Playground	Donut Shop	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Drugstore

4. Results

We found each neighborhood in each borough to show consistencies within top frequenting venue categories or the most common types of venues. The table below lists boroughs with their most occurring venue categories.

Borough	Most Common Venue Category
East Toronto	Cafes and Restaurants
Central Toronto	Gym and Park
York	Trails and Parks
Scarborough	Fast Food Restaurants
North York	Grocery Stores, Recreational Area/Parks
Etobicoke	Pizza Place
Downtown Toronto	Coffee Shop
East York	Restaurants

Cluster 1 shows that the most common venue categories are scenic spaces such as trails. Cluster 2 shows that parks and other recreational, open spaces are frequently occurring venue categories in that area. Cluster 3 is the largest, showing common venue categories to be fast food restaurants and cafes. This data was collected using the Wikipedia dataset combined with geographical data and then further segmented and clustered using FourSquare API and the K-mean clustering algorithm.

5. Conclusion

The purpose of this project was to identify most common venue category in neighborhoods of Toronto, in order to aid stakeholders in narrowing down the search for the optimal venue category in the neighborhoods of Toronto. A final decision can be made by stakeholders based on a given neighborhood's specific characteristics. There is additional scope to this project as more factors can be considered such as crime occurrences in certain boroughs, levels of noise/proximity to major roads, demographic information to target the right audience, real estate availability, socioeconomic groups residing in different areas and so on.