THE BATTLE OF CANADIAN NEIGHBORHOODS

PARAS BHATIA
September 25, 2019

1. INTRODUCTION

The Greater Toronto Area is the most populous metropolitan area in Canada. It consists of central city, Toronto, along with 25 surrounding municipalities distributed among four regional municipalities Durham, Halton, Peel and York. The Greater Toronto Area has an estimated population of 6.1 million people.
City of Toronto is the capital city of province of Ontario.
In this report we will explore the city of Toronto, we will be exploring almost 40 neighborhoods that are in city of Toronto. In terms of exploration we will try to explore the city with respect to venue category that are in the city.

1.1 Location Based Services

In today's era of digitization every location irrespective of industry of the location. We can easily find the information about a location due to data sharing and exponential rise in compute power. We have recommendation systems that recommend us the restaurants as per the user liking history, we can retrieve information such as trending location, photos of a particular location, number of likes received by a location, comments from people on a particular location and not to forget the directions to our favorite location. All of this in a couple of seconds.

1.2 OUR ATTEMPT

- With this project we try to attempt to find the trending location based on the keyword and address of a particular location.
- We rank the neighborhood in the city of Toronto based on the frequency of the occurrence of a venue category of a particular location and clustering those neighborhoods using KMeans
- We attempt to find the neighborhoods that have utmost one Pizza Place as the venue category and identifying those potential neighborhoods

1.3 INTEREST

People who are interesting in getting trending location at a particular location based on their keyword search. For example Sushi, Bay Street Toronto.
People or Business stakeholders who would like to know which neighborhood in city of Toronto is famous for what type of venue category.
People or Business stakeholders who would like to find the neighborhoods with an opportunity to open Pizza Places.

2. DATA ACQUISITION AND CLEANING

2.1 DATA SOURCES

Since, we are interested in the location based data. We will be using google API's geocode API service to get the coordinates for a location by passing its name.
We will be using Foursquare API, we will be querying Foursquare API for most of our project. Foursquare lets users search for restaurants, nightlife spots, shops, and other places of interest in their surrounding area. It is also possible to search other areas by entering the name of the remote location. It offers three types of accounts to query the places API. Personal, Start-Up, Enterprise. We will be creating a Personal account for Places API in Foursquare as that account is free of cost.
We will also collect data from city of Toronto webpage to collect the names of neighborhoods in city of Toronto region
https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/
We will web scrape the above URL page using BeautifulSoup4 to collect the names of the neighborhoods.

2.2 DATA CLEANING

Most of the data cleaning operation involves flattening JSON and parsing HTML from web scraped data to pandas dataframes.
For the JSON part we will be filtering and flattening the JSON data that is returned to us as output after querying the API.
And paring HTML returned from scraped data using BeautifulSoup4 paring engine and object.

Sample data returned by Foursquare API in the form of JSON


```
[{'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/sushi_', 'suffix': '.png'}, 'id': '4bf58dd8d48988d1d2941735', 'name': 'Sushi Restaurant', 'pluralName': 'Sushi Restaurants', 'primary': True, 'shortName': 'Sushi'}], 'hasPerk': False, 'id': '4b464cd6f964a520f11c26e3', 'location': {'address': '220 Yonge St.', 'cc': 'CA', 'city': 'Toronto', 'country': 'Canada', 'crossStreet': 'in Urban Eatery, Toronto Eaton Centre', 'distance': 398, 'formattedAddress': ['220 Yonge St. (in Urban Eatery, Toronto Eaton Centre)', 'Toronto ON M5B 2H6', 'Canada'], 'labeledLatLngs': [{'label': 'display', 'lat': 43.65480108229508, 'lng': -79.3808126449585}], 'lat': 43.65480108229508, 'lng': -79.3808126449585, 'postalCode': 'M5B 2H6', 'state': 'ON'}, 'name': 'Sushi-Q', 'referralId': 'v-1569204422'}, {'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/sushi_', 'suffix': '.png'}, 'id': '4bf58dd8d48988d1d2941735', 'name': 'Sushi Restaurant', 'pluralName': 'Sushi Restaurants', 'primary': True, 'shortName': 'Sushi'}],
```


Sample data after cleaning the JSON output from Foursquare API
The data describes the nearby sushi restaurant at Bay Street, Toronto

	Location.lat	Location.lng	Name	Address
0	43.654801	-79.380813	Sushi-Q	220 Yonge St.
1	43.661385	-79.381580	Daily Sushi	20 Carlton St.
2	43.657466	-79.380957	Tatami Sushi	335A Yonge St.
3	43.655031	-79.386724	Kathy's Sushi and Bento	187 Dundas St W
4	43.659092	-79.382789	Sushi & bbbop	384 nge St #57

Below table shows the sample cleaned data after scraping the webpage of city of Toronto to gather the names of neighborhood in Toronto area.

https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/

	Coordinates	Neighborhood
0	[43.6599082, -79.58331679999999]	Eringate-Centennial-West Deane
1	[43.6451146, -79.56877279999999]	Etobicoke West Mall
2	[43.6335688, -79.570763]	Markland Wood
3	[43.6017103, -79.5452384]	Alderwood
4	[43.593421, -79.538164]	Long Branch

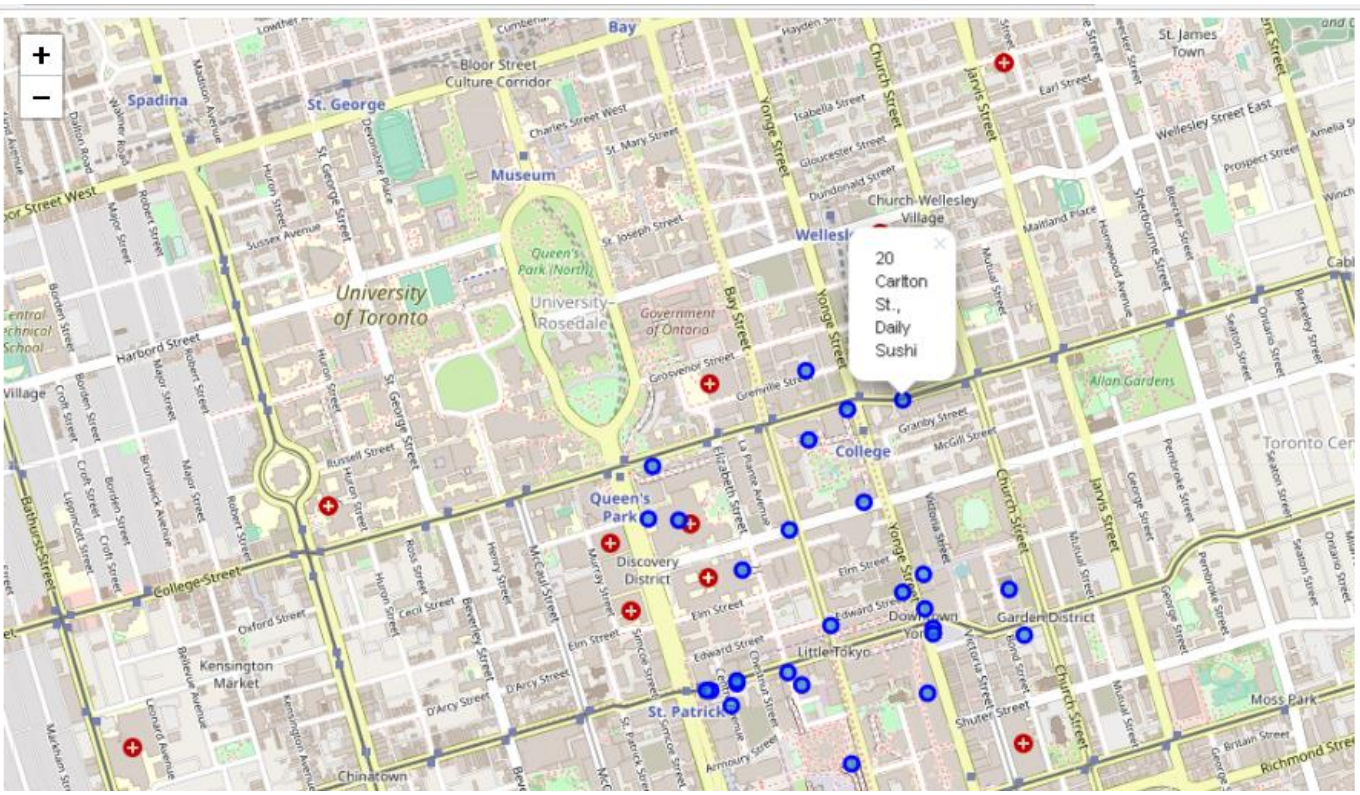
The neighborhood dataframe consists of 140 unique neighborhoods in city of Toronto region

3. METHODOLOGY

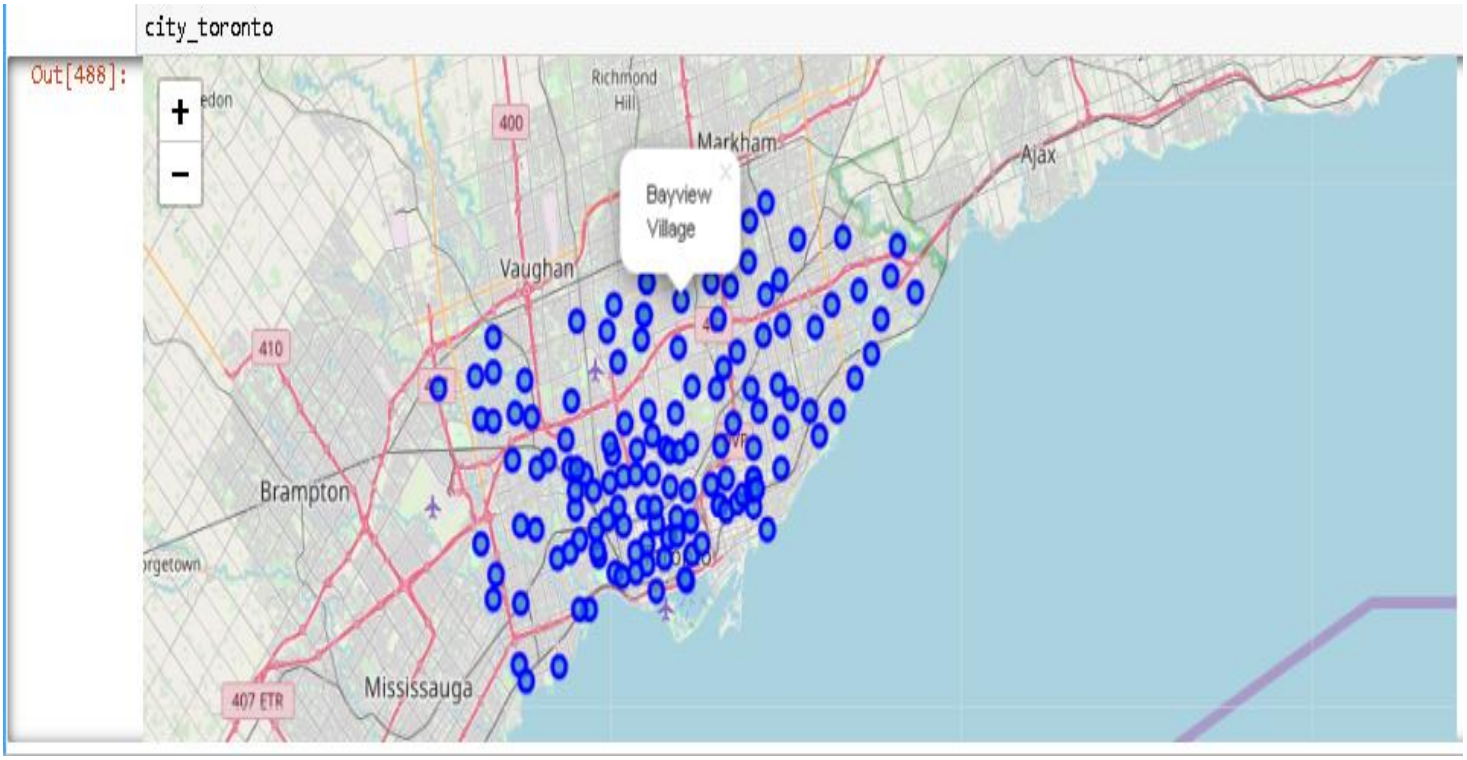
3.1 EXPLORATORY DATA ANALYSIS

First, we try to visualize the output venue category as per the location and query sent by user, and present the data returned from Foursquare API on location based map.
The map basically plots the location of all the venue category returned by the Foursquare API.

Out[478]:



The location based map shows all the 140 unique neighborhoods plotted on Toronto



3.2 ML METHODOLOGY

In this section, we plan to group neighborhood based on the frequency of a particular venue category occurring in that location.
We use the mean value of occurrence of venue categories located in a particular location and then analyzing the top 5 neighborhoods.
We then apply KMeans methodology to cluster those neighborhoods and finally identifying the potential neighborhoods with utmost one Pizza Place category.
First, we pass coordinates for each neighborhoods in Foursquare API and retrieve the venue category in that neighborhood. We use Foursquare API's explore endpoint.

Sample cleaned dataset after exploring all the neighborhoods

	Neighborhood	Neighborhood Lat	Neighborhood Lng	Venue	Venue Lat	Venue Lng	Venue Category
0	Eringate-Centennial-West Deane	43.659908	-79.583317	Centennial Park	43.656154	-79.587540	Park
1	Eringate-Centennial-West Deane	43.659908	-79.583317	Porta Via	43.663449	-79.589638	Sandwich Place

2	Eringate-Centennial-West Deane	43.659908	-79.583317	Mrakovic	43.666641	-79.578850	Eastern European Restaurant

After, performing one-hot encoding of the above dataset. We rank the venue categories in a particular neighborhood based on the mean frequency of occurrence of neighborhood

Agincourt North	
venue	freq
Indian Restaurant	0.07
Coffee Shop	0.07
Ice Cream Shop	0.07
Chinese Restaurant	0.07
Convenience Store	0.07

In the above table the frequency of occurrence of all these venue categories is 0.07 in Agincourt North neighborhood in Toronto area.

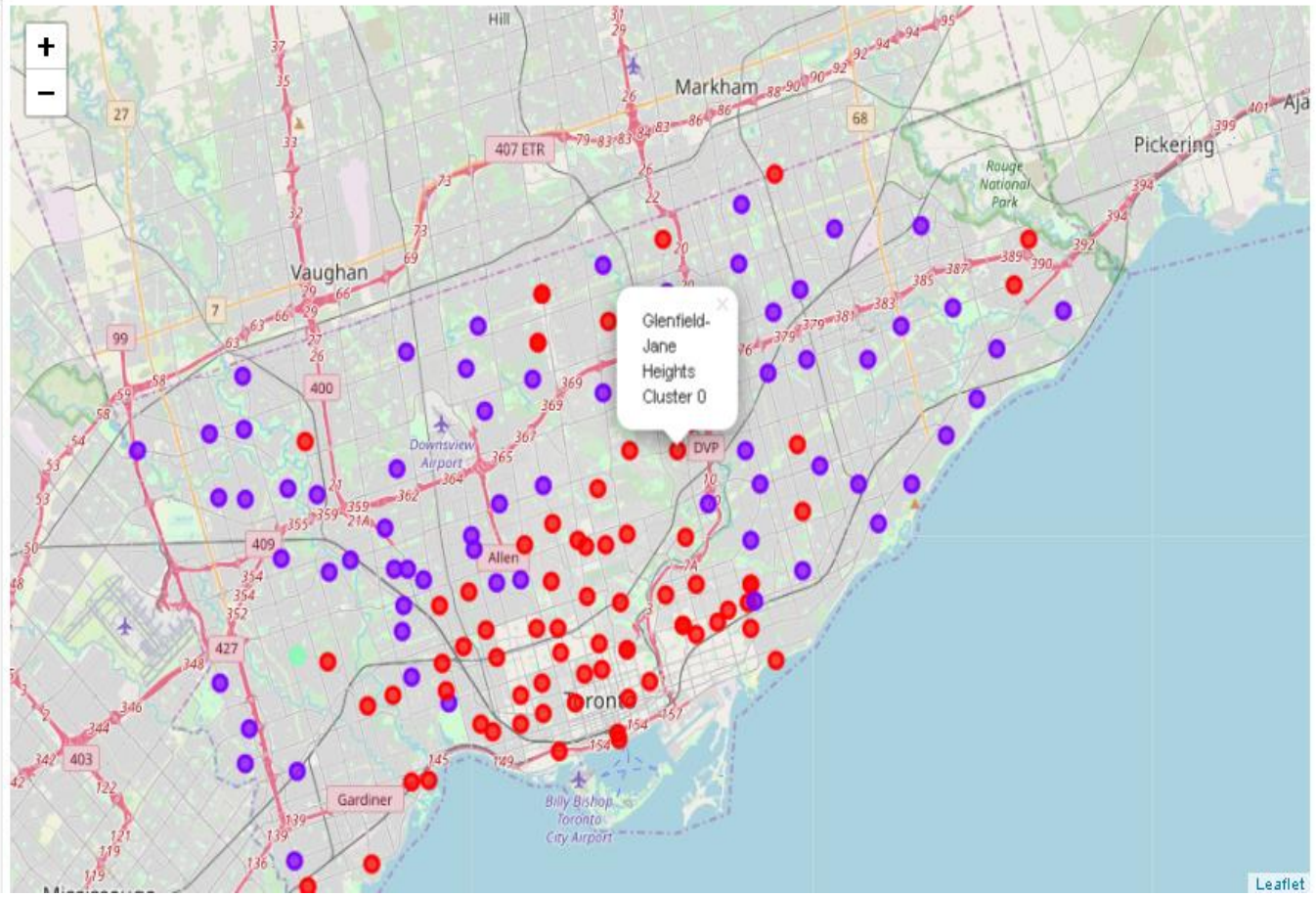
3.3 CLUSTERING

We apply KMeans algorithm on the dataset after performing one-hot encoding and after grouping all the neighborhoods and ranking all the venue categories based on ranking.

	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
0	Agincourt North	Ice Cream Shop	Chinese Restaurant	Convenience Store	Coffee Shop	Indian Restaurant	Taco Place	Bank	Bakery	Clothing Store
1	Agincourt South-Malvern West	Chinese Restaurant	Cantonese Restaurant	Korean Restaurant	Fast Food Restaurant	Coffee Shop	Pet Store	Market	Beer Store	Liquor Store
2	Alderwood	Discount Store	Pharmacy	Convenience Store	Pizza Place	Shopping Mall	Dance Studio	Coffee Shop	Park	Grocery Store
3	Annex	Café	Vegetarian / Vegan Restaurant	Beer Bar	Pizza Place	Restaurant	Bakery	Deli / Bodega	Japanese Restaurant	Coffee Shop
4	Banbury-Don Mills	Japanese Restaurant	Golf Course	Middle Eastern Restaurant	Baseball Field	Cosmetics Shop	Pharmacy	Gym / Fitness Center	Café	Caribbean Restaurant

The reason for choosing the clustering model is because we wanted to group the neighborhoods of similar categories in a city and wanted to analyze it. We choose the cluster size of 3 in our case.

Out[519]:



4. ANALYSIS

In the analysis section, we see the trend as to which neighborhoods are falling in a particular cluster and which neighborhood is famous for kind of venue category.

Neighborhoods in cluster 0

In [520]:
```
city_of_toronto_merged.loc[city_of_toronto_merged['Cluster Labels'] == 0, city_of_toronto_merged.columns
[[0] + list(range(4, city_of_toronto_merged.shape[1]))]]
```

Out[520]:

	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
4	Long Branch	Café	Bar	Coffee Shop	Pharmacy	Grocery Store	Gym	Discount Store	Sandwich Place
8	University	Coffee Shop	Vegetarian / Vegan Restaurant	Café	Restaurant	Japanese Restaurant	Bookstore	Ice Cream Shop	Italian Restaurant
9	Palmerston-Little Italy	Korean Restaurant	Café	Dessert Shop	Coffee Shop	Pizza Place	Indian Restaurant	Tapas Restaurant	Taco Place
10	Dufferin Grove	Café	Coffee Shop	Italian Restaurant	Cocktail Bar	Vegetarian / Vegan Restaurant	Bakery	Comedy Club	Restaurant

There are 73 unique neighborhoods in cluster 0.

Neighborhoods with cluster 1

```
city_of_toronto_merged.loc[city_of_toronto_merged['Cluster Labels'] == 1, city_of_toronto_merged.columns
[[0] + list(range(4, city_of_toronto_merged.shape[1]))]]
```

	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
25	Niagara	Park	Gym	Café	Historic Site	Light Rail Station	Trail	Arts & Crafts Store	Dog Run	Pizza Place
33	Edenbridge-Humber Valley	Park	Skating Rink	Baseball Field	Women's Store	Eastern European Restaurant	Fish & Chips Shop	Filipino Restaurant	Field	Fast Food Restauran
48	Lambton-Baby Point	Park	Playground	Garden	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Donut Shop

There are 66 unique neighborhoods in cluster 1.

Neighborhoods with cluster 2

```
city_of_toronto_merged.loc[city_of_toronto_merged['Cluster Labels'] == 2, city_of_toronto_merged.columns
[[0] + list(range(4, city_of_toronto_merged.shape[1]))]]
```

	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 Mc Commc Venue
6	Princess-Rosethorn	Playground	Golf Course	Garden	Yoga Studio	Farmers Market	Electronics Store	Ethiopian Restaurant	Event Service	Event Space	Falafel Restaur

As we can see there is only one unique neighborhood in cluster 2.

This analysis shows that how the neighborhoods are clustered in city of Toronto. We can use this analysis to rate the neighborhoods based on what they are famous for.

Neighborhoods in toronto region with only one Venue Category of Pizza Place ¶

In [551]:
```
one_venue_pizza
```

Out[551]:

	Neighborhood	Venue
0	Agincourt North	1
1	Agincourt South-Malvern West	1
2	Bathurst Manor	1
3	Bay Street Corridor	1
4	Bendale	1
5	Brookhaven-Amesbury	1
6	Cabbagetown-South St. James Town	1

5. RESULTS AND DISCUSSION

Using the geocode API provided by google, we can pass the type of food, category that a user wants to try out in a particular location and we can recommend the places based on the search query.
Our results and findings shows that neighborhoods in Toronto region has a great variety when it comes to different kinds of venue category.
Also we can further find which neighborhood is famous for what kind of venue category based on the frequency of occurrence of that particular venue in the neighborhood.
We were also able to find the potential neighborhoods with utmost one Pizza Place in the venue category.

6. CONCLUSION

Finally, we were able to find necessary results.
Further work that can be done to improve is that making a web application and connecting this model in the backend and serving all the necessary results in the frontend, can be pursued upon.

At last, selecting a location for a particular restaurant for example finding neighborhoods with potential to open more pizza places; as in this case, can depend on many factors such as land price, area, population demographic in that area.

The project is an attempt to take a step further and solve problem with the help of technology and any product that cannot be perfect in one attempt of building it. So, keep breaking and making things up fellas!

I thank Coursera and IBM team for bringing together all the course content, organizing it and making all the assignments and everything. It has been a great journey of learning so far. This course is fantastic for anyone who wants to learn about Data Science.