Predicting Neighborhood to Open Flag-ship Restaurant for Bong Foodies Ltd

A. Introduction

A.1. Description & Discussion of the Background

Toronto is the provincial capital of **Ontario** and the most populous city in Canada, with a population of 2,731,571 as of 2016, the Toronto census metropolitan area (CMA), of which the majority is within the Greater Toronto Area (GTA), held a population of 5,928,040, making it Canada's most populous CMA. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (as of 2016) surrounding the western end of Lake Ontario. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

Being from India, I decided to use an imaginary **Indian restaurant aggregator Bong Foodies Ltd.** deciding to **open their flagship Indian restaurant in Toronto**. I was based out of Mississauga, Canada for more than 1.5 years and visited Toronto a few times. Considering the multi-cultural nature of the city along with a huge population of **Asian Origin people coming from India, Bangladesh, Pakistan** it made sense for me to choose the city as the location where the company would like to open their flag-ship restaurant.

However, considering that it has a lot of boroughs and different type of restaurants, what they wanted to understand was which would be an ideal suburb or borough where the can open it. To solve this problem, they are expecting to answer a few key questions on the neighborhoods in Toronto and recommend a best possible neighborhood. Some of the relevant questions are:

- Which are the busiest neighborhoods in terms of business, retail shops, restaurants?
- How many restaurants are there in each of these neighborhoods?
- What is the average rating of restaurants in these neighborhoods if possible?
- How many restaurants are there in these neighborhoods?
- Are there neighborhoods where there are no multi-cuisine restaurants?

A.2. Data Description

To consider the problem I have decided to use Wikipedia to get the list of suburbs for Toronto, as well as necessary demographic information of the suburbs or neighborhoods which will then be used by them as reference to extract other necessary details from Foursquare about each of the neighborhoods which includes business details, shops, retails, number of restaurants, most common restaurants etc. Further assessment and data wrangling will be done to understand which the best neighborhood is to open a multicuisine restaurant. The data sets include:

1. Neighborhood data from Wikipedia for Toronto, CA
2. Co-ordinate data sourced from Web for the neighborhoods
3. Foursquare for venue data information

Borough & Neighborhood Data from Wikipedia

Link - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Data Fields -

- Postal code - Postal Code of a Neighborhood in Toronto (Alphanumeric)
- Borough – Name of the Borough (Text)
- Neighborhood – Neighborhood areas of the Borough (Text)

Geospatial or Latitude & Longitude data for the Boroughs

Link - http://cocl.us/Geospatial_data

Data Fields -

- Postal code (Alphanumeric)
- Latitude (Alphanumeric)
- Longitude (Alphanumeric)

Data on Venues in the City

I will use **Foursquare API** to get the most common venues of given Borough of Toronto.

Link - https://foursquare.com/

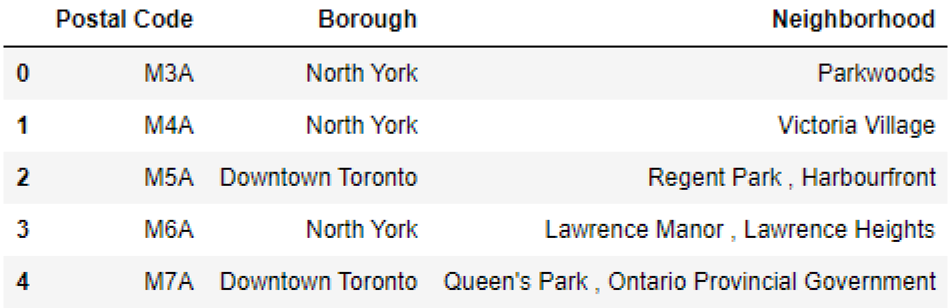
B. Data Analysis

For my code management, I have used GitHub as my repository.

B1. Data Preparation

My first data set is the Neighborhood data. There was no direct input file for this data set. So, I have used scraping method to get the data extracted from Wikipedia. Once, the data is scraped it need to be converted into a more consumable format for data analysis. In the next step, I converted the scraped data into a pandas dataframe as in Fig 1. *There are 3 main fields – Postal Code, Borough and Neighborhood*

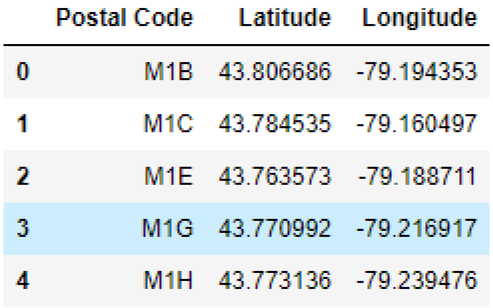
I had to do some special character & next-line characters removal to get a clean data set.



	Postal Code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park , Harbourfront
3	M6A	North York	Lawrence Manor , Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park , Ontario Provincial Government

Fig 1. Dataframe Created from Scraped data from Wikipedia

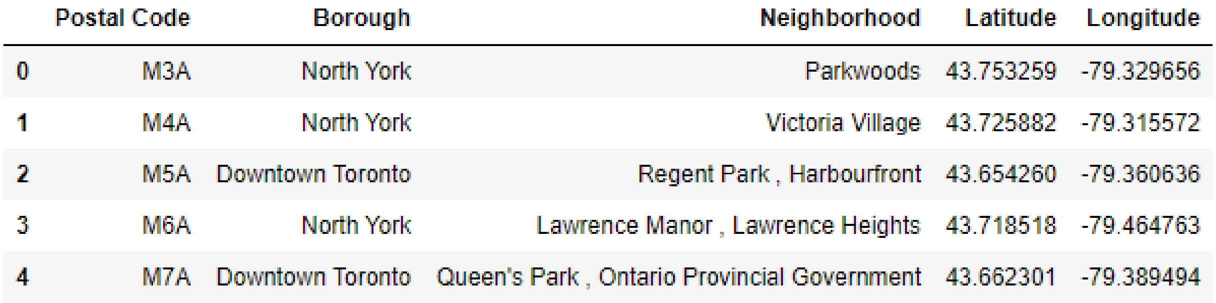
In the next step, I wanted to add the Latitude and Longitude data to the data set. I wanted to use the geocode module in pandas to get the information. However, there were some issues around it using Geocode and hence I decided to use a reference data of Latitude and Longitude mapped to Postal code as obtained during the Course.



	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

Fig 2. Dataframe Created for Postal Code, Latitude and Longitude

The next step was to merge the two data sets in Fig 1 and Fig 2 to create a new data set as in Fig 3



	Postal Code	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	Regent Park , Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor , Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park , Ontario Provincial Government	43.662301	-79.389494

Fig 3. Dataframe Created After Merge of Data Sets in Fig 1 and Fig 2

Below is a graphical representation of the Boroughs/Neighborhoods using Latitude and Longitude

```
The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

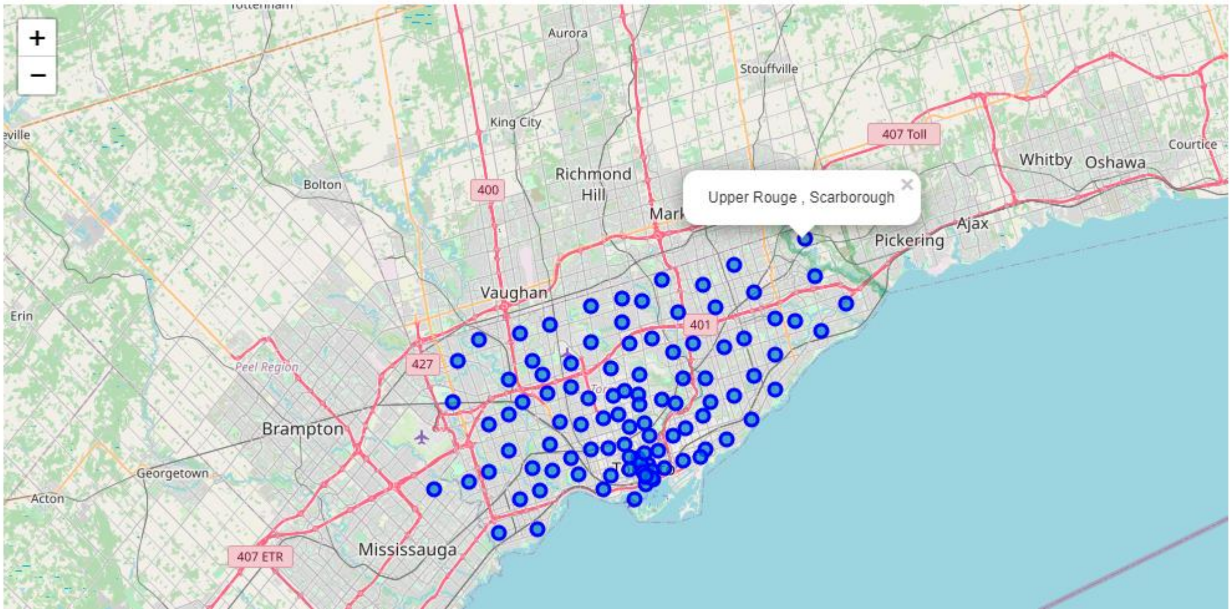
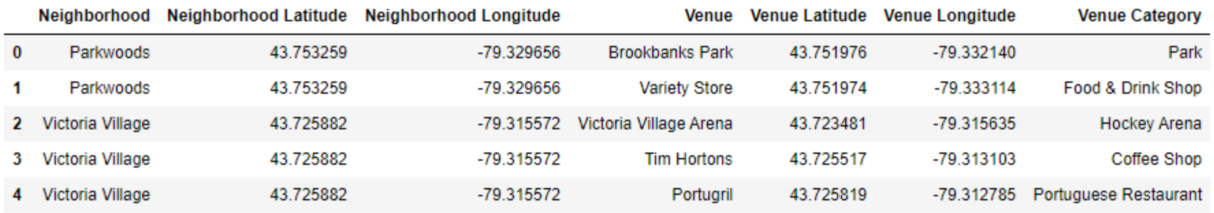


Fig 4. Boroughs Plotted on a map of Toronto, CA e.g. Upper Rouge, Scarborough

B2. Extract and Analyze Neighborhood Data for Toronto, CA

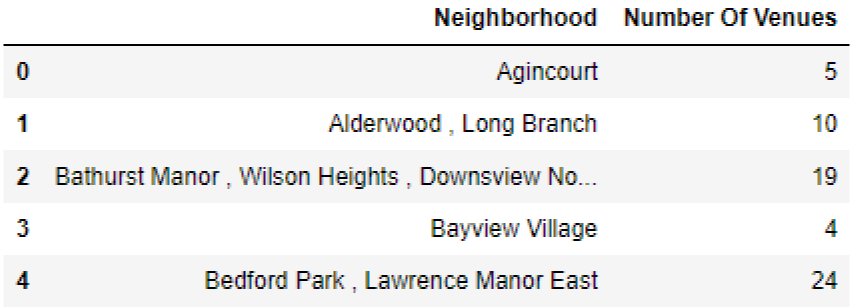
I utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as 100 venue and the radius 500 meter for each borough from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API. Fig 5 shows some of that list.



	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	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Fig 5. List of Venues Extracted using Foursquare API

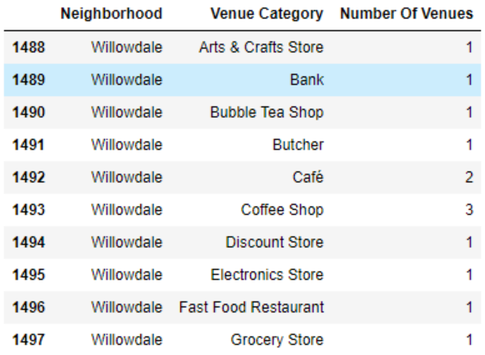
There were 2151 venues that was returned for all the Neighborhoods The next part of our analysis is with only the Neighborhood, Venue, Venue Category as we are interested in the kind of venues are there in each of these neighborhoods. I also wanted to visualize which Neighborhoods were most busy or had the greatest number of venues. Hence, I created a new dataframe grouping the Venue by Neighborhood as below Fig 6.



	Neighborhood	Number Of Venues
0	Agincourt	5
1	Alderwood , Long Branch	10
2	Bathurst Manor , Wilson Heights , Downsview No...	19
3	Bayview Village	4
4	Bedford Park , Lawrence Manor East	24

Fig 6. Number of Venues Grouped by Neighborhood

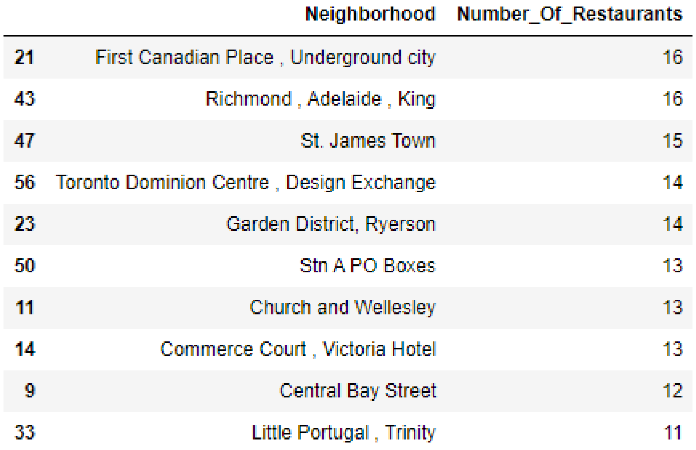
For example, when I check for **Willowdale**, we can see the category of venues and the number of each of them. As We can see, there are 3 Coffee Shops and 2 Café in that neighborhood.



	Neighborhood	Venue Category	Number Of Venues
1488	Willowdale	Arts & Crafts Store	1
1489	Willowdale	Bank	1
1490	Willowdale	Bubble Tea Shop	1
1491	Willowdale	Butcher	1
1492	Willowdale	Café	2
1493	Willowdale	Coffee Shop	3
1494	Willowdale	Discount Store	1
1495	Willowdale	Electronics Store	1
1496	Willowdale	Fast Food Restaurant	1
1497	Willowdale	Grocery Store	1

Fig 7. Number of Venues Grouped by Neighborhood & Category

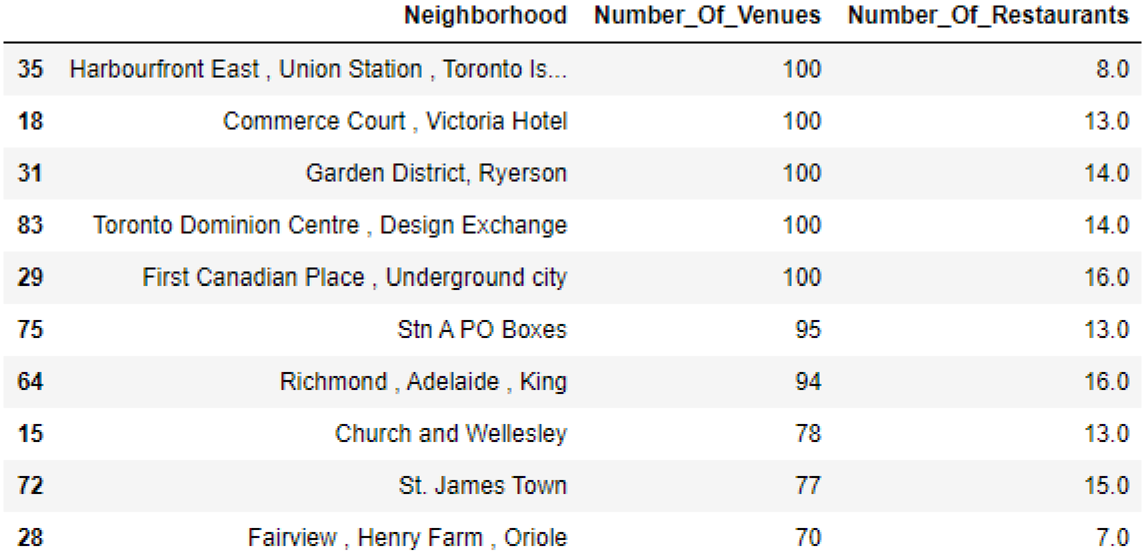
I further processed my data to find neighborhoods where there are no Indian Restaurants and find the number of other restaurants located in that neighborhood. I also wanted to see which are the Top 10 neighborhoods which had the greatest number of non-Indian restaurants.



	Neighborhood	Number_Of_Restaurants
21	First Canadian Place , Underground city	16
43	Richmond , Adelaide , King	16
47	St. James Town	15
56	Toronto Dominion Centre , Design Exchange	14
23	Garden District, Ryerson	14
50	Stn A PO Boxes	13
11	Church and Wellesley	13
14	Commerce Court , Victoria Hotel	13
9	Central Bay Street	12
33	Little Portugal , Trinity	11

Fig 7. Top 10 Neighborhoods with respect to Number of Non-Indian Restaurants

I further wanted to compare the Neighborhoods by total number of venues to the total number of non-Indian Restaurants in that Neighborhood. I sorted Neighborhoods by Total Number of Venues in Descending Order and Total Number of Non-Indian Restaurants in Ascending Order which gave me the results as in Fig 8.



	Neighborhood	Number_Of_Venues	Number_Of_Restaurants
35	Harbourfront East , Union Station , Toronto Is...	100	8.0
18	Commerce Court , Victoria Hotel	100	13.0
31	Garden District, Ryerson	100	14.0
83	Toronto Dominion Centre , Design Exchange	100	14.0
29	First Canadian Place , Underground city	100	16.0
75	Stn A PO Boxes	95	13.0
64	Richmond , Adelaide , King	94	16.0
15	Church and Wellesley	78	13.0
72	St. James Town	77	15.0
28	Fairview , Henry Farm , Oriole	70	7.0

Fig 8. Number of Venues vs Number of Non-Indian Restaurants - Neighborhoods

The Top 10 Neighborhoods identified in the above figure was chosen as the list of neighborhoods where there is a possibility of opening an Indian Restaurant considering that due to big number of venues and restaurants there would be a good influx of customers and also since there are no Indian Restaurants in these Neighborhoods makes them even more suitable. This completes the data gathering and analysis stage.

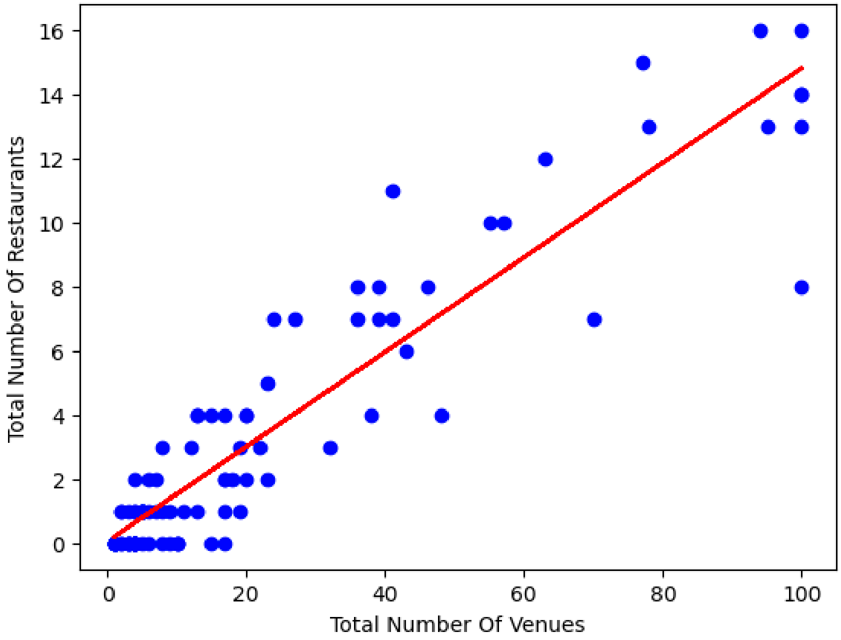
C. Methodology

The next stage was to identify which of these neighborhoods was the most ideal to open the restaurant.

I wanted to first find out if there is any relationship between the total number of venues and total number of restaurants for a neighborhood. From the looks of it, there is a very linear trend to it as displayed in Fig 9 below.

As we can see, the linear regression model fits in approximately and shows that more the number of venues, more is the number of restaurants. There is a good possibility to use this method to predict the number of venues or the number of restaurants if we have either of those numbers.

The equation is: $y= 0.1476267 + .05865559x$



I decided to next group the neighborhoods as clusters. I could see 5 distinct clusters as below with 4 of them around the same part of the city.

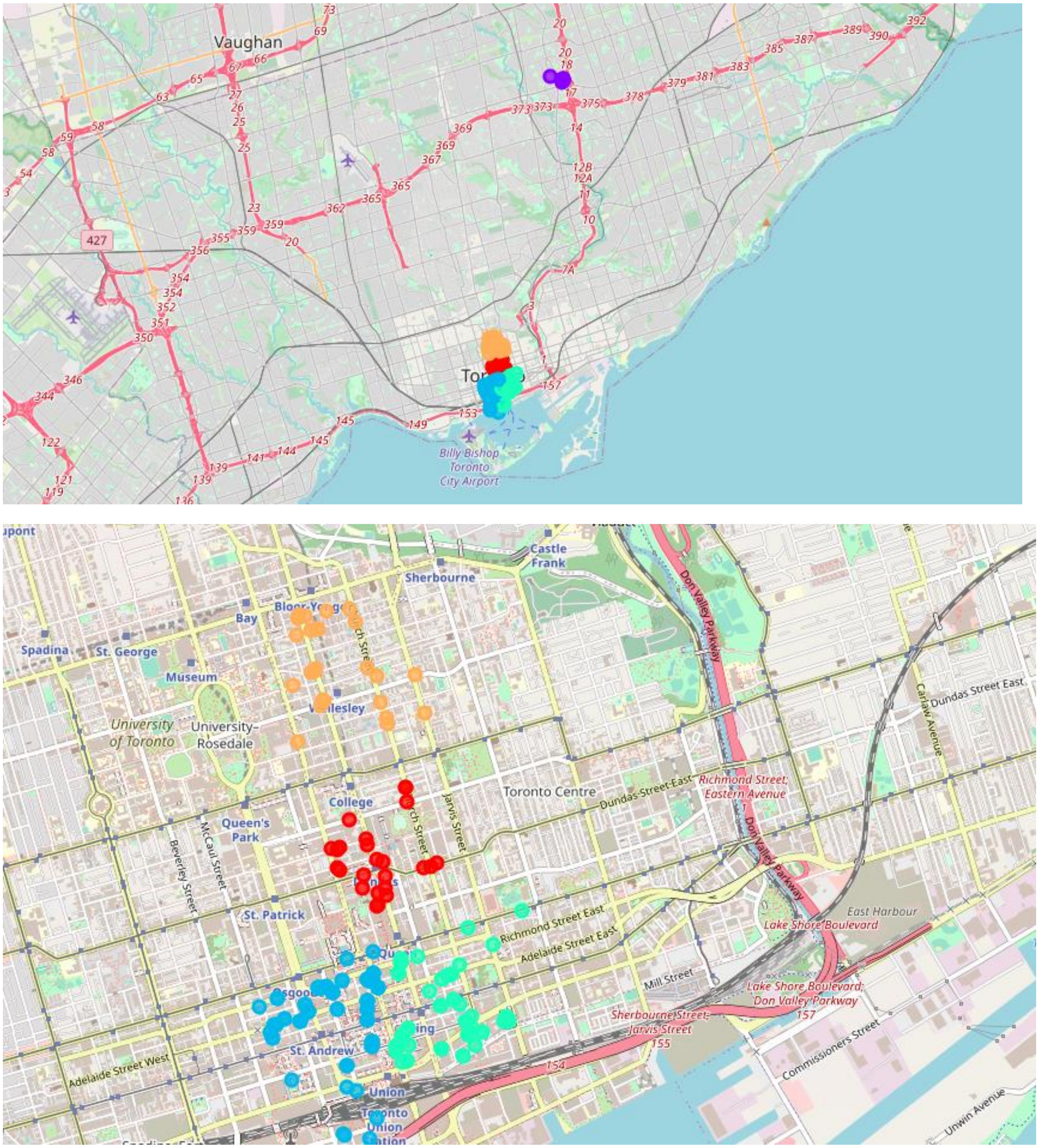


Fig 9. Clustering Neighborhoods – Version 1

C. Results

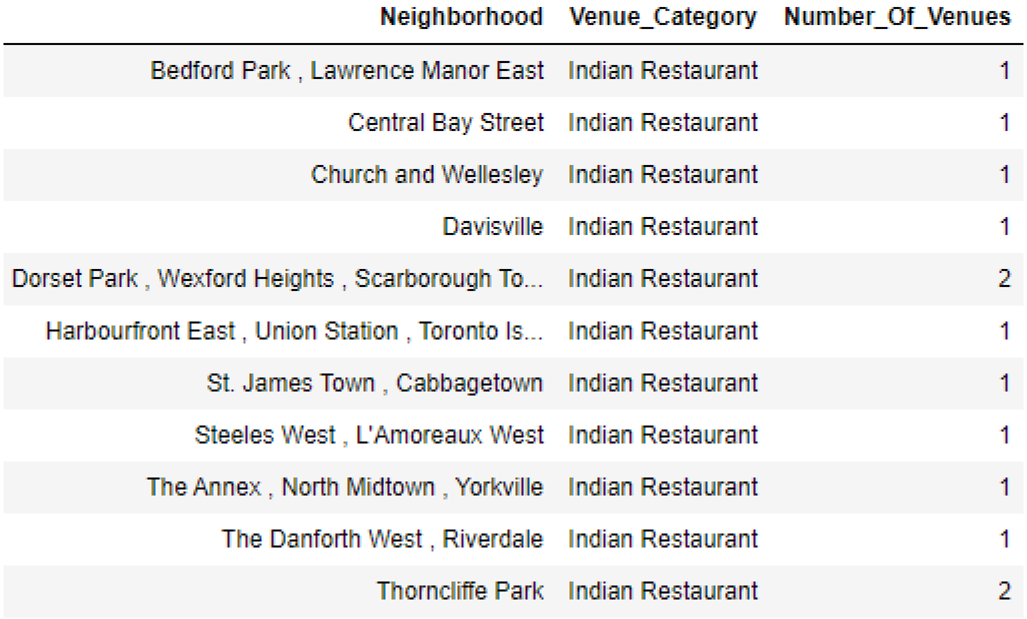
I used **Linear Regression** to get some statistical understanding of **relationship between number of venues and number of restaurants in a neighborhood**. I also did **Clustering using K-Means** for a specific clustering logic of the neighborhoods and restaurant locations.

The clustering results revealed **Church and Wellesley (Denoted by Orange)** and **St. James Town (Denoted by Red), Fairview, Henry Farm, Oriole (Denoted by Purple)** were independent of each other whereas the other 7 Neighborhoods were clustered into 2 groups which were very close to each other. Also, as noticeable from Fig 8. Number of Venues vs Number of Non-Indian Restaurants - Neighborhoods the other 7 neighborhoods account for a huge number of venues and restaurants situated close to each other which can be referenced from the clustering results.

In summary, I wanted to identify the Top 10 Neighborhoods with the greatest number of venues & restaurants and eliminate any neighborhoods which had Indian Restaurants to reduce the competition which I was able to achieve

D. Discussion

It was observed that there are 13 Indian Restaurants in the City. But at the same time, the number of other venues in these neighborhoods is quite low. Below is the list of neighborhoods with Indian Restaurants. The concern is how good the rating or footfall is for these venues. I have not done that analysis in this project, but the data set can be expanded using Foursquare API to get further insights.



Neighborhood	Venue_Category	Number_Of_Venues
Bedford Park , Lawrence Manor East	Indian Restaurant	1
Central Bay Street	Indian Restaurant	1
Church and Wellesley	Indian Restaurant	1
Davisville	Indian Restaurant	1
Dorset Park , Wexford Heights , Scarborough To...	Indian Restaurant	2
Harbourfront East , Union Station , Toronto Is...	Indian Restaurant	1
St. James Town , Cabbagetown	Indian Restaurant	1
Steeles West , L'Amoreaux West	Indian Restaurant	1
The Annex , North Midtown , Yorkville	Indian Restaurant	1
The Danforth West , Riverdale	Indian Restaurant	1
Thorncliffe Park	Indian Restaurant	2

Hence, I decided to focus on neighborhoods where there are lots of venues and restaurants but at the same time there are no Indian Restaurants. There were 2 reasons behind this:

1. More venues mean a greater number of customers
2. No Indian Restaurant means less competition in terms of the specific cuisine

I further limited my data set to Top 10 neighborhoods fitting the above 2 criteria. Further Linear Regression suggested that more the number of venues more is the number of restaurants located for that neighborhoods.

E. Conclusion

With all the data analysis and further clustering mechanism I feel the below neighborhoods could be ideal to open the flagship restaurant Church and Wellesley (Denoted by Orange) and St. James Town (Denoted by Red), Fairview, Henry Farm, Oriole (Denoted by Purple).

However, considering Oriole being quite far from Downtown, I feel Church and Wellesley, or St. James Town could be more ideal to open the new restaurants. St. James is further closer to Downtown as visible from the map and hence that could be the best location.