# BATTLE OF THE NEIGHBORHOODS

Ajaykumar Mudaliar

## Introduction/Business Problem

We are looking to open a restaurant in Toronto, the goal is to maximize business profits. The Chef is a keen Indian Chef but is also capable of cooking other cuisines.

Using Data Analysis, we will try to find the answers to the following questions:

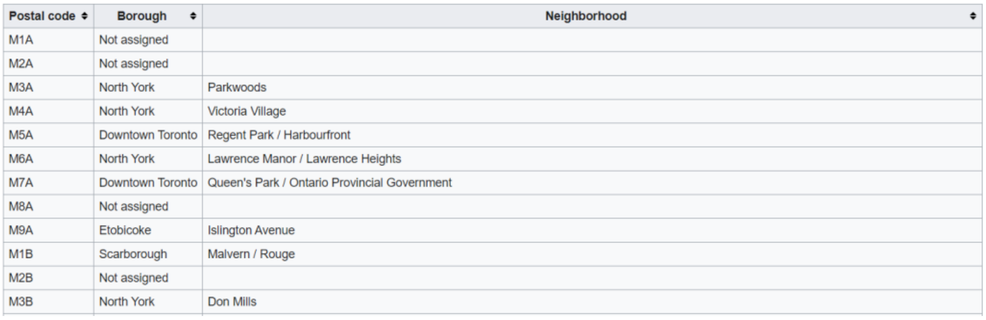
- When neighborhoods are the best to open an Indian restaurant?
- Is opening an Indian the best option in Toronto?
- If not, Indian which other cuisines are famous and bound to earn a better profit?
- Are Restaurants famous in a concentrated neighborhood or are they distributed?

## Approach

- ➢ We will first retrieve the neighborhoods of Toronto and get the most famous venues in each neighborhood.
- ➢ Then we filter the venues to just restaurants
- ➢ We group neighborhoods by unsupervised clustering to check if there is any underlying pattern in the distribution of the restaurants.
- ➢ We will analyse which are the most famous cuisines in Toronto and where they are located.
- ➢ We will analyse the distribution of Indian restaurants across Toronto.
- ➢ We will analyse the number of famous Indian restaurants all around Toronto.
- ➢ Using this information, we will reach a conclusion on the questions described above.

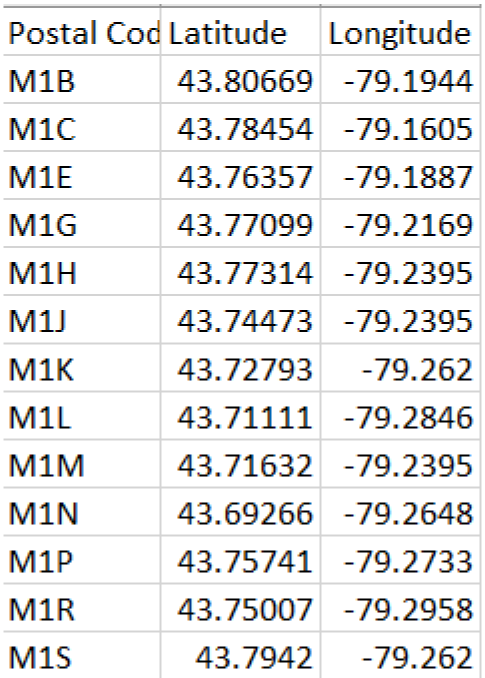
## Data:

- We will Wikipedia to retrieve the postcodes of neighborhoods in Toronto, link = [https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)



| Postal code ⬍ | Borough ⬍ | Neighborhood ⬍ |
|---|---|---|
| M1A | Not assigned | |
| M2A | Not assigned | |
| M3A | North York | Parkwoods |
| M4A | North York | Victoria Village |
| M5A | Downtown Toronto | Regent Park / Harbourfront |
| M6A | North York | Lawrence Manor / Lawrence Heights |
| M7A | Downtown Toronto | Queen's Park / Ontario Provincial Government |
| M8A | Not assigned | |
| M9A | Etobicoke | Islington Avenue |
| M1B | Scarborough | Malvern / Rouge |
| M2B | Not assigned | |
| M3B | North York | Don Mills |

- We use a python modules and a .csv file containing latitude and longitudes of the neighborhoods of Toronto to update each neighborhood with their longitude and latitudes.



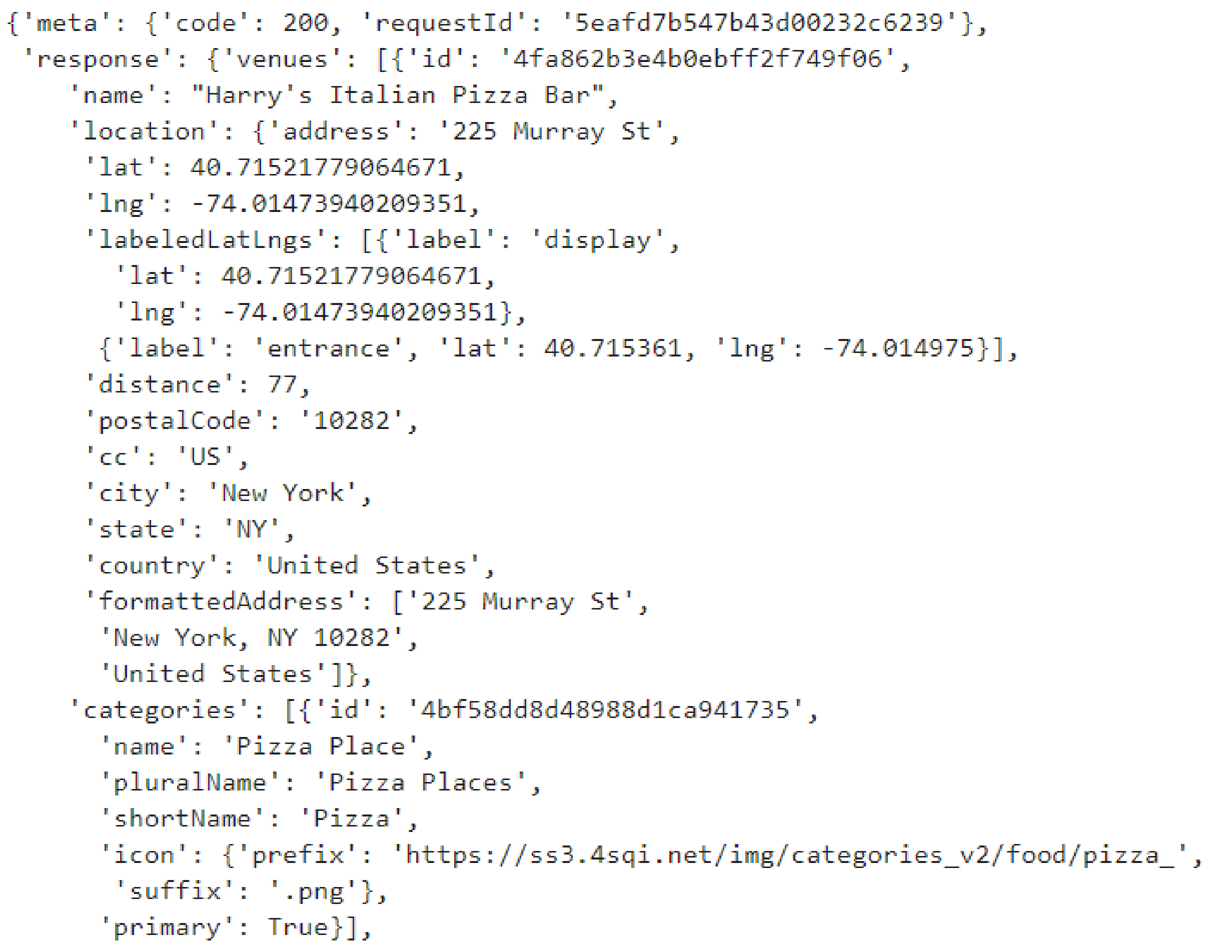
| Postal Cod | Latitude | Longitude |
|---|---|---|
| M1B | 43.80669 | -79.1944 |
| M1C | 43.78454 | -79.1605 |
| M1E | 43.76357 | -79.1887 |
| M1G | 43.77099 | -79.2169 |
| M1H | 43.77314 | -79.2395 |
| M1J | 43.74473 | -79.2395 |
| M1K | 43.72793 | -79.262 |
| M1L | 43.71111 | -79.2846 |
| M1M | 43.71632 | -79.2395 |
| M1N | 43.69266 | -79.2648 |
| M1P | 43.75741 | -79.2733 |
| M1R | 43.75007 | -79.2958 |
| M1S | 43.7942 | -79.262 |

- We will use Foursquare to retrieve reviews and popular venues in Toronto.

A Query request to get the venues will look like this:

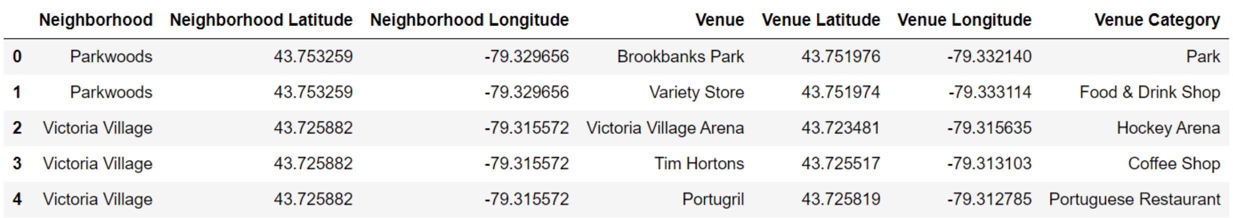
```
https://api.foursquare.com/v2/venues/ search ?
client_id= CLIENT_ID &client_secret= CLIENT_SECRET &ll= LATITUDE , LONGITUDE &v= VERSION &query= QUERY &radius= RADIUS &limit= LIMIT
```

The data is retrieved in form of a json file, a snapshot of which is shown below:



```
{'meta': {'code': 200, 'requestId': '5eafd7b547b43d00232c6239'},
 'response': {'venues': [{'id': '4fa862b3e4b0ebff2f749f06',
    'name': "Harry's Italian Pizza Bar",
    'location': {'address': '225 Murray St',
     'lat': 40.71521779064671,
     'lng': -74.01473940209351,
     'labeledLatLngs': [{'label': 'display',
       'lat': 40.71521779064671,
       'lng': -74.01473940209351},
      {'label': 'entrance', 'lat': 40.715361, 'lng': -74.014975}],
     'distance': 77,
     'postalCode': '10282',
     'cc': 'US',
     'city': 'New York',
     'state': 'NY',
     'country': 'United States',
     'formattedAddress': ['225 Murray St',
      'New York, NY 10282',
      'United States']},
    'categories': [{'id': '4bf58dd8d48988d1ca941735',
      'name': 'Pizza Place',
      'pluralName': 'Pizza Places',
      'shortName': 'Pizza',
      'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/pizza_',
       'suffix': '.png'},
      'primary': True}],
```

This information will be processed to be of the form:



| | 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 |

## Data acquisition and cleaning

The sources for the data have been discussed in the above section.

All the data retrieved from the sources will have to be processed to a desirable format so EDA can be done.

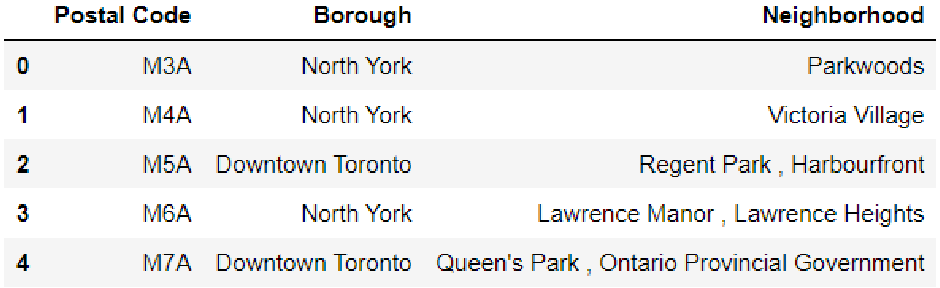
## **The methods used for data cleaning is discussed below:**

## 1: Data from Wikipedia:

This data is retrieved by scraping the Wikipedia website using beautiful soup from the bs4 module in python.

This is done in the following steps:

1. Download the HTML using the request module by passing the required web address
2. Create a Beautiful Soup object
3. Parse the html using the html.parser of beautiful soup.
4. Locate the table in the html text
5. Loop through each cell and retrieve the Postcode, Borough and Neighborhood information.
6. Ignore postcodes with missing information
7. Zip all the information and store it in a dataframe.
8. The result looks like the following:



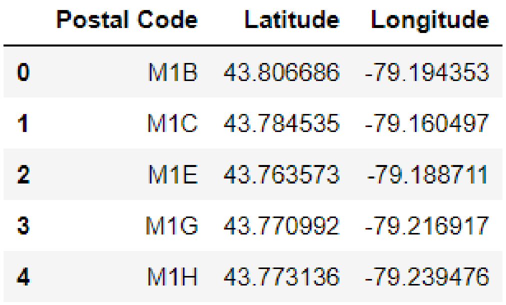
| | Postal Code | Borough | Neighborhood |
|---|---|---|---|
| **0** | M3A | North York | Parkwoods |
| **1** | M4A | North York | Victoria Village |
| **2** | M5A | Downtown Toronto | Regent Park , Harbourfront |
| **3** | M6A | North York | Lawrence Manor , Lawrence Heights |
| **4** | M7A | Downtown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

## 2: Data about postcodes, latitude and longitude:

We need the latitude and longitude information to locate the neighborhoods on a folium map, but this information is unavailable on Wikipedia.

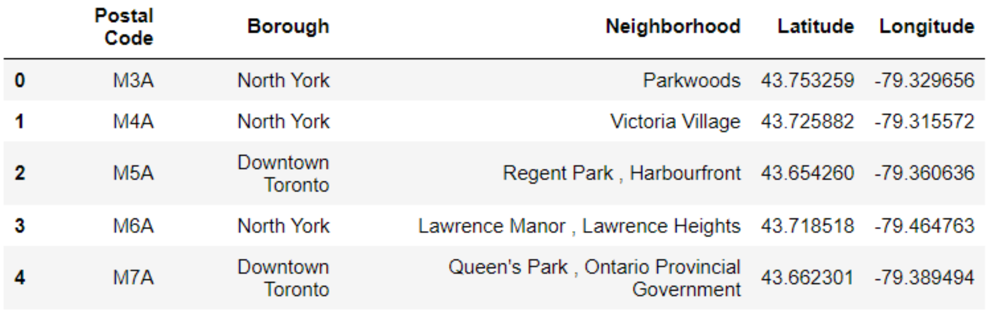
So first we try to retrieve this information using Geopy a python module, but information for many postcodes is missing, so we use a geospatial data csv, load it into a dataframe and then merge it with the original dataframe on the Postal Codes, the result looks like the following:

Csv file loaded into dataframe:



| | 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 |

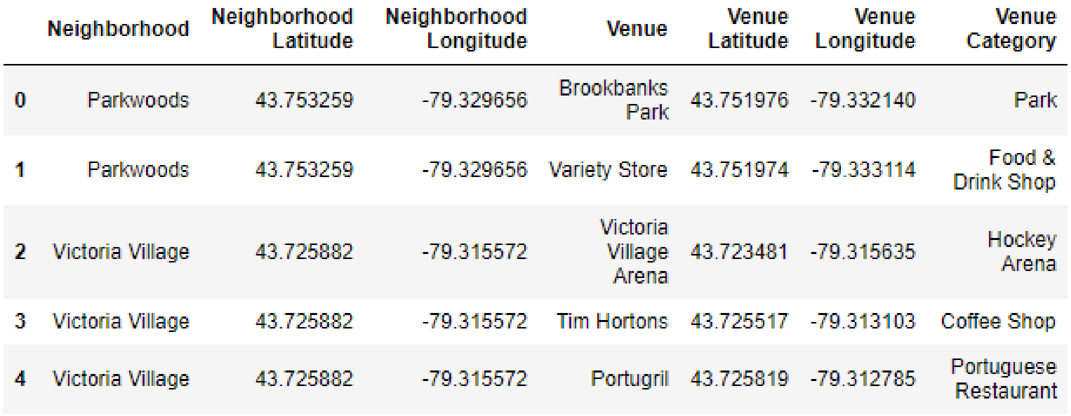
Complete dataframe:



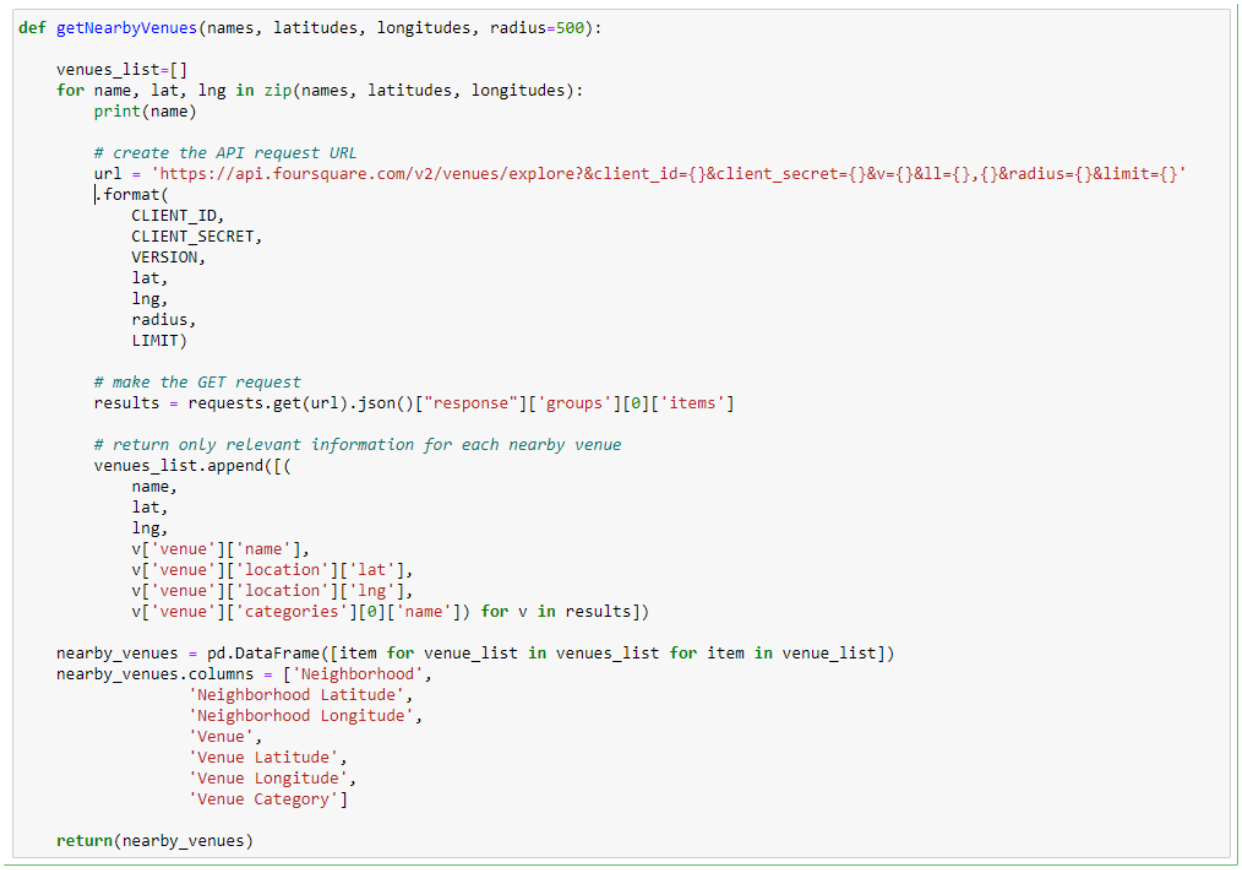
| | Postal Code | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|
| **0** | M3A | North York | Parkwoods | 43.753259 | -79.329656 |
| **1** | M4A | North York | Victoria Village | 43.725882 | -79.315572 |
| **2** | M5A | Downtown Toronto | Regent Park , Harbourfront | 43.654260 | -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 |

## 3: Data from Foursquare:

Data from foursquare after the GET request is retrieved in form of a json file and using the code shown in the next page we parse this information in a dataframe shown below:

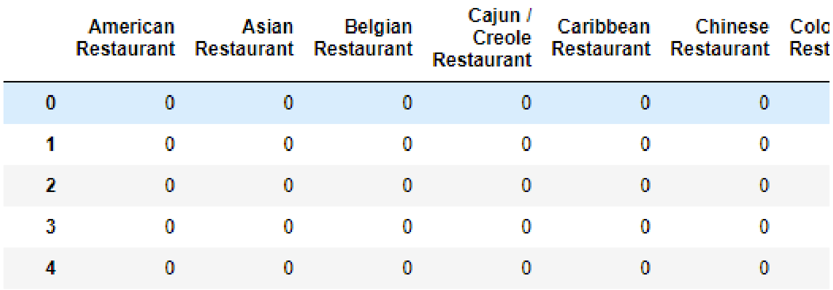


| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| **0** | Parkwoods | 43.753259 | -79.329656 | Brookbanks Park | 43.751976 | -79.332140 | Park |
| **1** | Parkwoods | 43.753259 | -79.329656 | 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 |



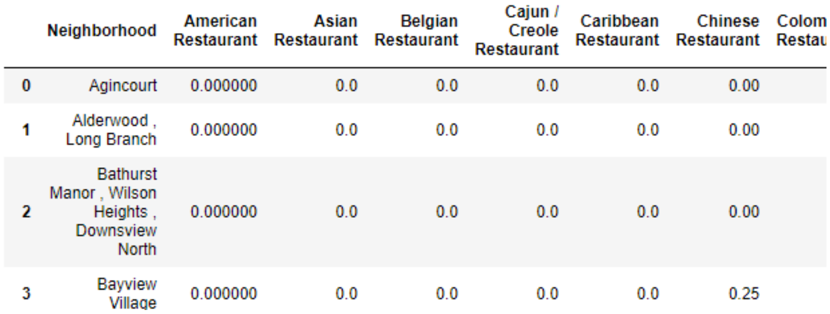
```
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'
        .format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

Then we filter this One hot encode this dataframe on the Venue Category column and then filter to just restaurants



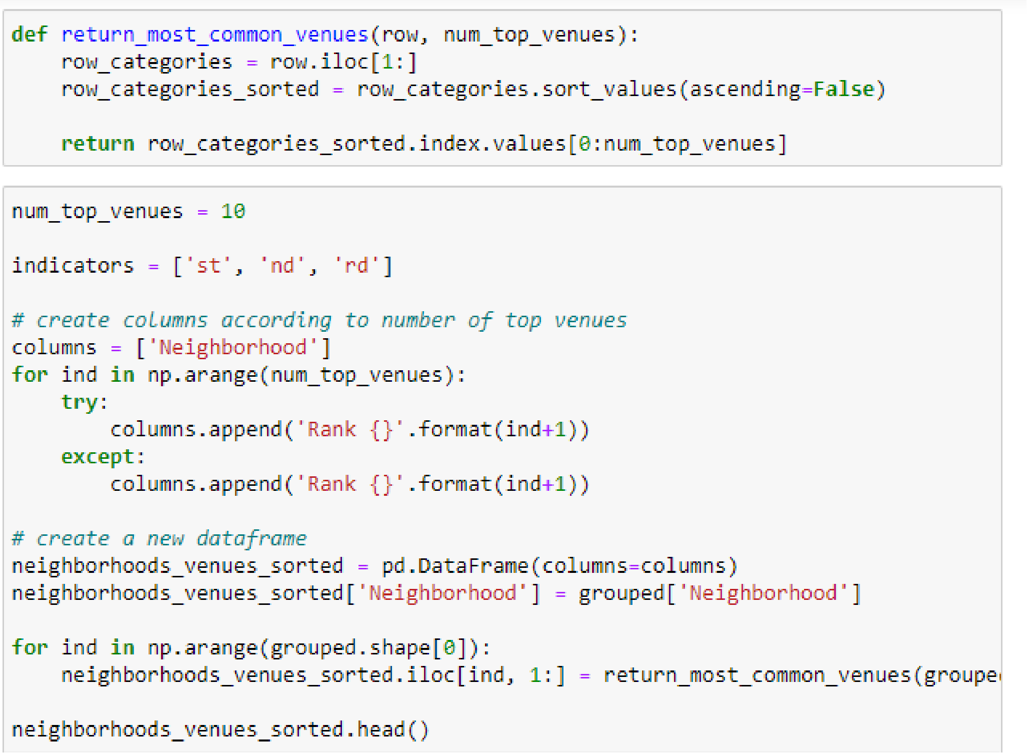
| | American Restaurant | Asian Restaurant | Belgian Restaurant | Cajun / Creole Restaurant | Caribbean Restaurant | Chinese Restaurant | Colc Rest |
|---|---|---|---|---|---|---|---|
| **0** | 0 | 0 | 0 | 0 | 0 | 0 | |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | |
| **2** | 0 | 0 | 0 | 0 | 0 | 0 | |
| **3** | 0 | 0 | 0 | 0 | 0 | 0 | |
| **4** | 0 | 0 | 0 | 0 | 0 | 0 | |

We then group each category by neighborhood:



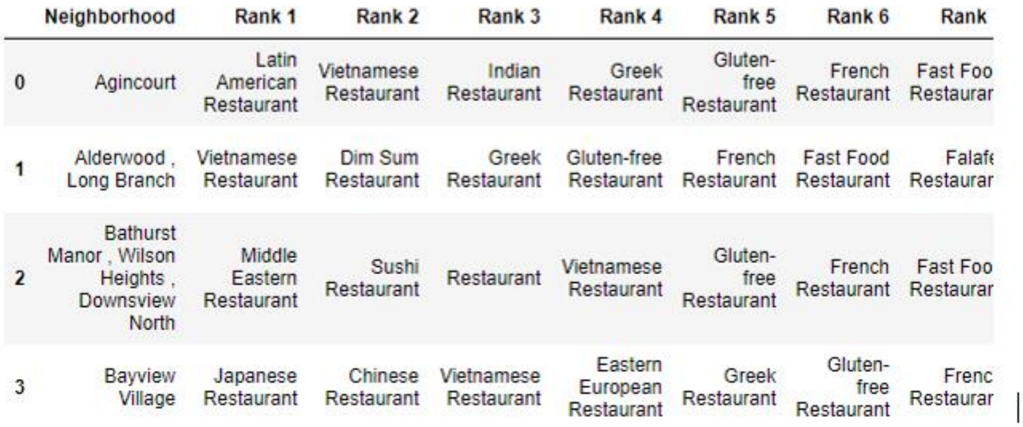
| | Neighborhood | American Restaurant | Asian Restaurant | Belgian Restaurant | Cajun / Creole Restaurant | Caribbean Restaurant | Chinese Restaurant | Colom Restau |
|---|---|---|---|---|---|---|---|---|
| **0** | Agincourt | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | |
| **1** | Alderwood , Long Branch | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | |
| **2** | Bathurst Manor , Wilson Heights , Downsview North | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | |
| **3** | Bayview Village | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.25 | |

Using another function shown below we find out the most popular venues in each neighborhood and put that information into a dataframe as well.



```
def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending=False)

    return row_categories_sorted.index.values[0:num_top_venues]
```

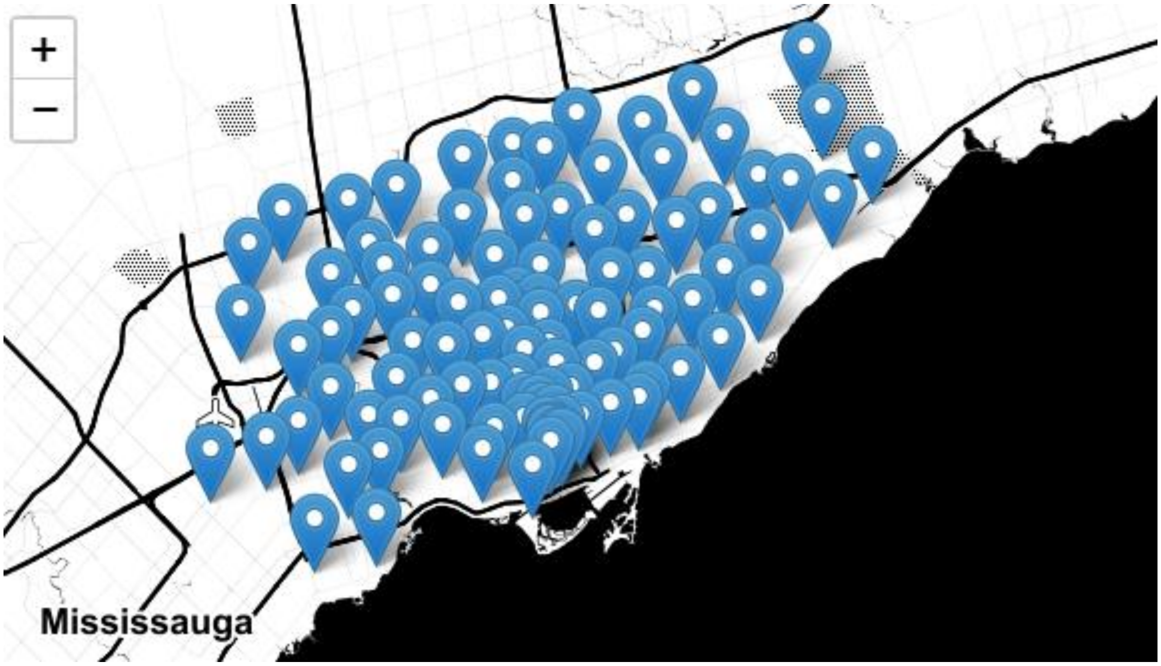
```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('Rank {}'.format(ind+1))
    except:
        columns.append('Rank {}'.format(ind+1))

# create a new dataframe
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = grouped['Neighborhood']

for ind in np.arange(grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(groupe

neighborhoods_venues_sorted.head()
```



| | Neighborhood | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 | Rank 6 | Rank |
|---|---|---|---|---|---|---|---|---|
| 0 | Agincourt | Latin American Restaurant | Vietnamese Restaurant | Indian Restaurant | Greek Restaurant | Gluten-free Restaurant | French Restaurant | Fast Foo Restaurar |
| 1 | Alderwood , Long Branch | Vietnamese Restaurant | Dim Sum Restaurant | Greek Restaurant | Gluten-free Restaurant | French Restaurant | Fast Food Restaurant | Falafe Restaurar |
| 2 | Bathurst Manor , Wilson Heights , Downsview North | Middle Eastern Restaurant | Sushi Restaurant | Restaurant | Vietnamese Restaurant | Gluten-free Restaurant | French Restaurant | Fast Foo Restaurar |
| 3 | Bayview Village | Japanese Restaurant | Chinese Restaurant | Vietnamese Restaurant | Eastern European Restaurant | Greek Restaurant | Gluten-free Restaurant | Frenc Restaurar |

## Insights found from data:

First looking at the distribution of the neighborhoods in Toronto for a rough understand of the map:

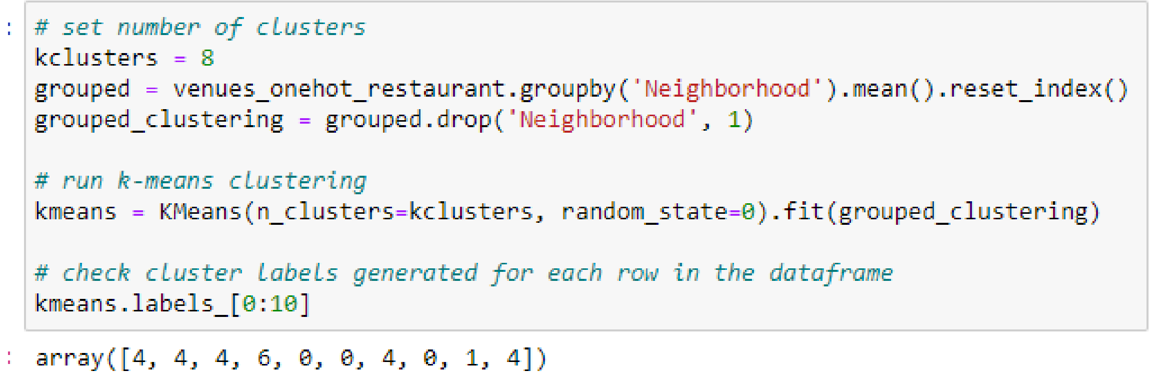


In all these neighborhoods there are 38 different types of restaurant and many unlabeled restaurants which will be treated as multi-cuisine restaurants.

Let’s apply some Machine Learning algorithms to discover insights.

Let’s try to group the neighborhoods according to their preference in restaurant type. Using K Means unsupervised model with 7 clusters we Label each neighborhood.

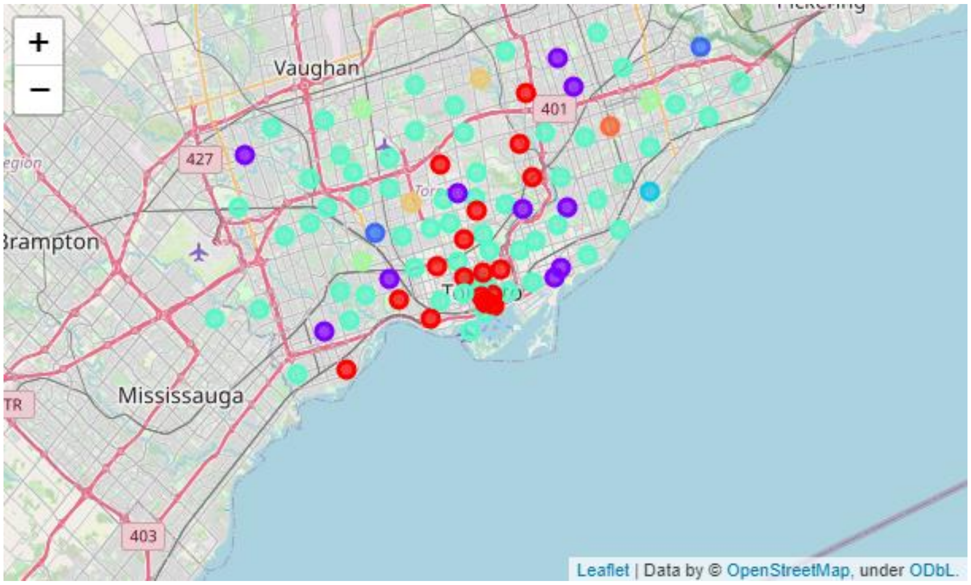
This is done by using training a Sklearn.cluster.KMeans model on the onehot_encoded dataset, this is shown in the code below:



```
# set number of clusters
kclusters = 8
grouped = venues_onehot_restaurant.groupby('Neighborhood').mean().reset_index()
grouped_clustering = grouped.drop('Neighborhood', 1)

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

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

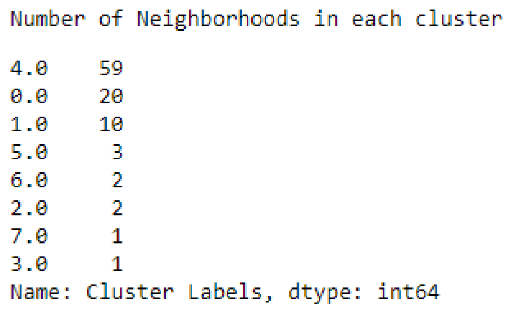
```
array([4, 4, 4, 6, 0, 0, 4, 0, 1, 4])
```

Afters clustering the neighborhoods according to their restaurant preferences lets look at the map:

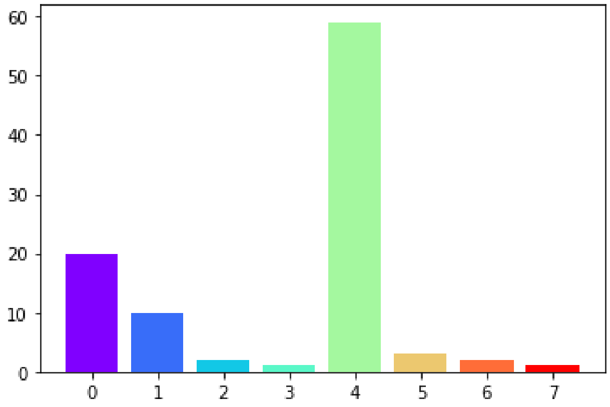


Looks like there is an uneven distribution of preferences, some clusters are much larger than the others.

***Let’s look at their sizes:***



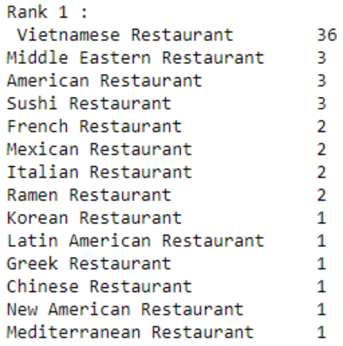
```
Number of Neighborhoods in each cluster

4.0    59
0.0    20
1.0    10
5.0     3
6.0     2
2.0     2
7.0     1
3.0     1
Name: Cluster Labels, dtype: int64
```



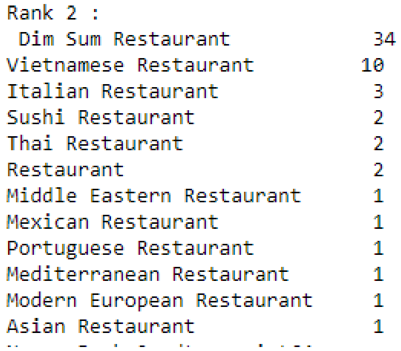
The light blue clusters on the map can be identified as Cluster 4 and the red ones as cluster 0 and the purple ones as cluster 0.

***Let's look what is similar in each cluster.***

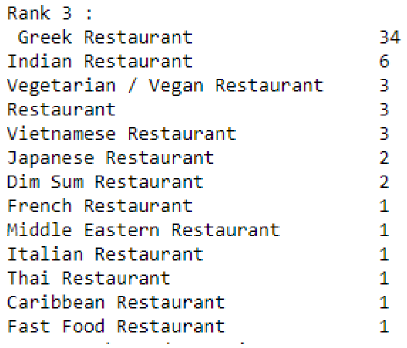
For cluster 4: (most popular cluster for Restaurants)



```
Rank 1 :
 Vietnamese Restaurant          36
Middle Eastern Restaurant       3
American Restaurant             3
Sushi Restaurant                3
French Restaurant               2
Mexican Restaurant              2
Italian Restaurant              2
Ramen Restaurant                2
Korean Restaurant               1
Latin American Restaurant       1
Greek Restaurant                1
Chinese Restaurant              1
New American Restaurant         1
Mediterranean Restaurant        1
```



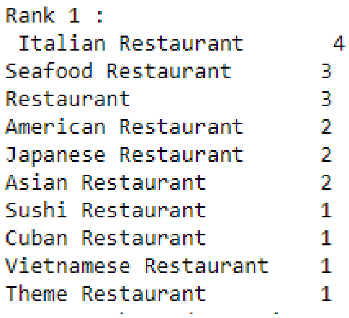
```
Rank 2 :
 Dim Sum Restaurant             34
Vietnamese Restaurant          10
Italian Restaurant              3
Sushi Restaurant                2
Thai Restaurant                 2
Restaurant                      2
Middle Eastern Restaurant       1
Mexican Restaurant              1
Portuguese Restaurant           1
Mediterranean Restaurant        1
Modern European Restaurant      1
Asian Restaurant                1
```



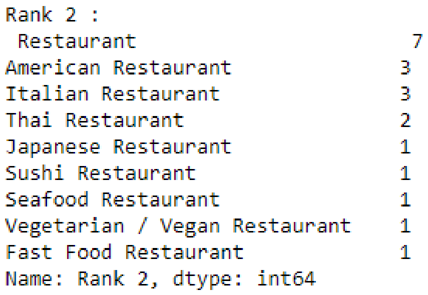
```
Rank 3 :
 Greek Restaurant                34
Indian Restaurant                6
Vegetarian / Vegan Restaurant    3
Restaurant                       3
Vietnamese Restaurant            3
Japanese Restaurant              2
Dim Sum Restaurant               2
French Restaurant                1
Middle Eastern Restaurant        1
Italian Restaurant               1
Thai Restaurant                  1
Caribbean Restaurant             1
Fast Food Restaurant             1
```

It looks like Vietnamese, Dim sum and Greek food is very popular in these neighborhoods.

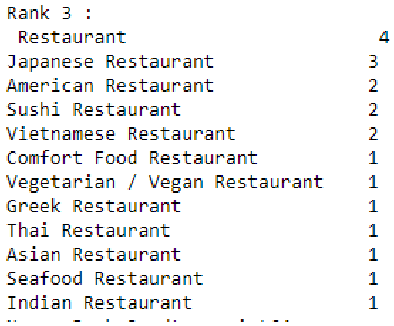
For cluster 0 (second most popular cluster for Restaurants)



```
Rank 1 :
 Italian Restaurant         4
Seafood Restaurant         3
Restaurant                 3
American Restaurant        2
Japanese Restaurant        2
Asian Restaurant           2
Sushi Restaurant           1
Cuban Restaurant           1
Vietnamese Restaurant      1
Theme Restaurant           1
```



```
Rank 2 :
 Restaurant                        7
American Restaurant               3
Italian Restaurant                3
Thai Restaurant                   2
Japanese Restaurant               1
Sushi Restaurant                  1
Seafood Restaurant                1
Vegetarian / Vegan Restaurant     1
Fast Food Restaurant              1
Name: Rank 2, dtype: int64
```

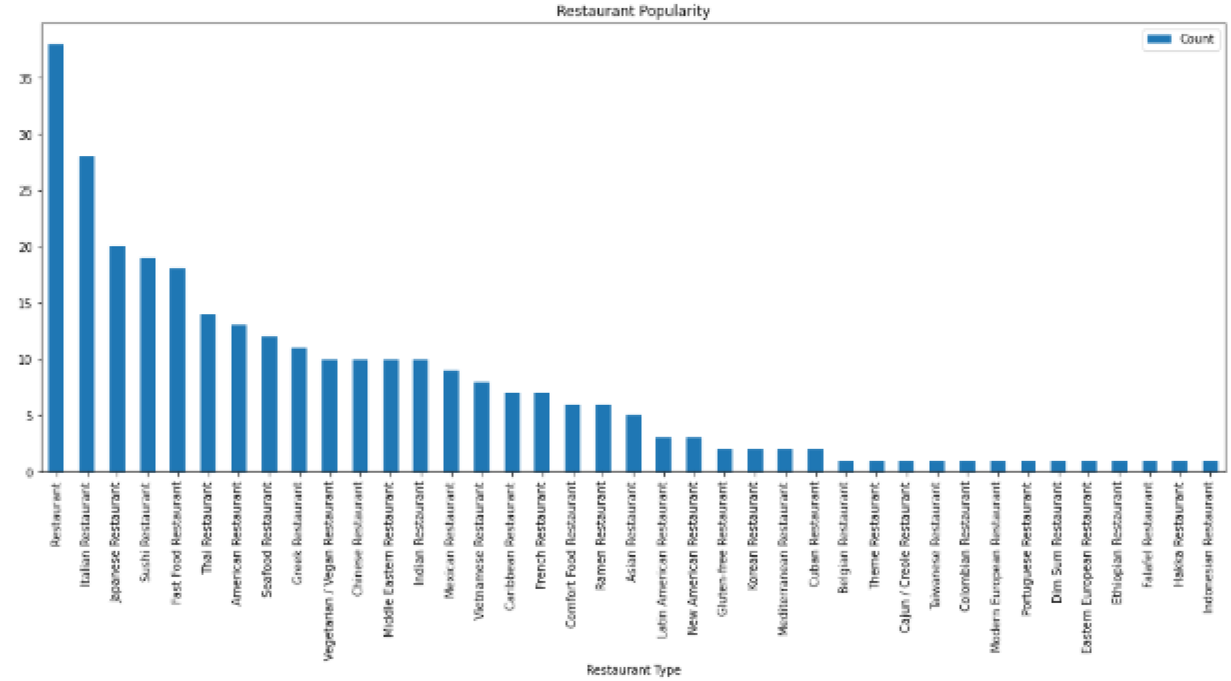


```
Rank 3 :
 Restaurant                       4
Japanese Restaurant              3
American Restaurant              2
Sushi Restaurant                 2
Vietnamese Restaurant            2
Comfort Food Restaurant          1
Vegetarian / Vegan Restaurant    1
Greek Restaurant                 1
Thai Restaurant                  1
Asian Restaurant                 1
Seafood Restaurant               1
Indian Restaurant                1
```

In these neighborhoods Italian, American and unlabelled restaurants seem to be famous.

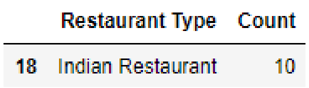
*From this we can say that Indian food does not seem to be famous in most of the area having very low number of well-known famous restaurant. Lets future analyse this fact ->*

Let’s look at the distribution of famous restaurants according to their cuisines:



A lot of Unlabeled restaurants, Japanese, Italian, Sushi and fast food chains seem to exist with only 10 popular Indian Restaurants

Now let’s look at the rank (according to number of popular restaurants) of Indian restaurants:



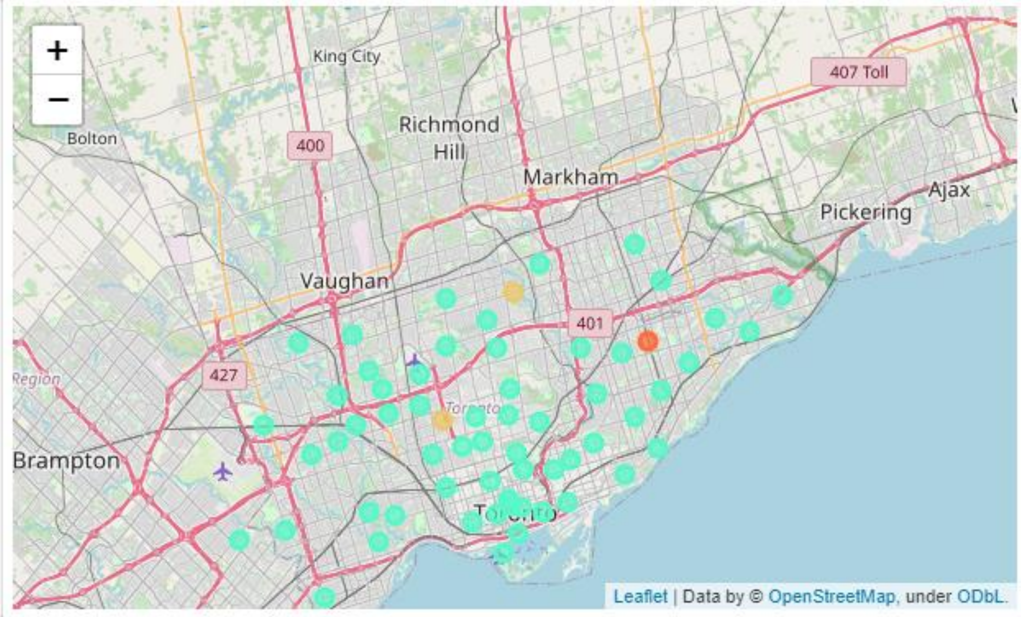
| | Restaurant Type | Count |
|---|---|---|
| **18** | Indian Restaurant | 10 |

Rank 18/37, Indian restaurant are not so popular in Toronto.

*If say, we had to open an Indian restaurant beside this fact, let us try to find the best location. Let’s explore the location of the existing famous Indian Restaurants. ->*

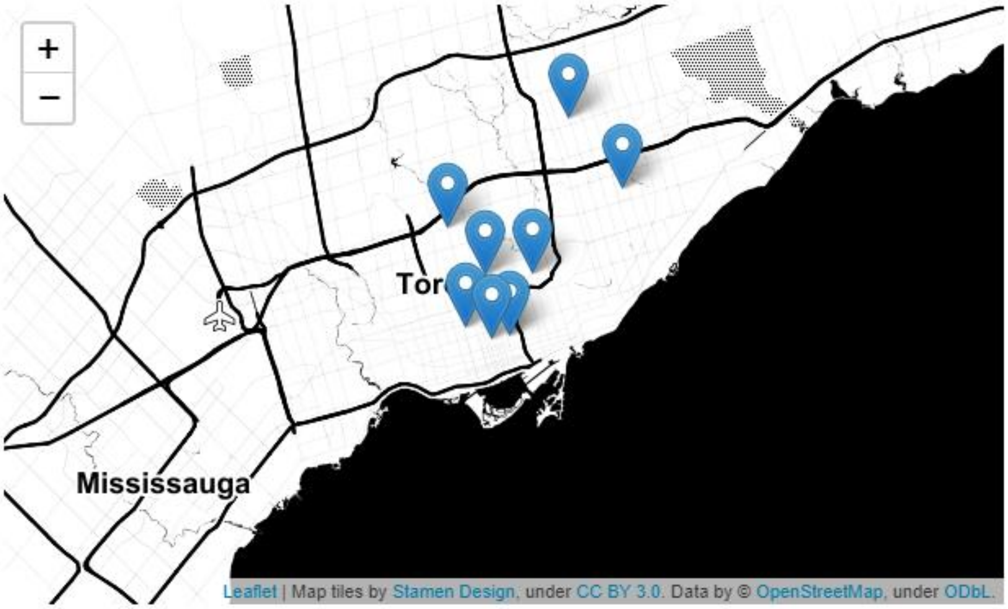
Upon more analysis it is found that these restaurants lie in neighborhoods in clusters 4, 6, and 7

Let’s see where these clusters of neighborhoods lie on the map.



These neighborhoods are well distributed, which means there is no particular area which Is famous for Indian food.

Lets see the specific neighborhoods with the famous Indian restaurants:


+
−
Tor
Mississauga
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under ODbL.

They all lie close to downtown toronto.

## Conclusion:

The distribution of the cluster containing famous Indian restaurants is un-even, i.e. they are located across Toronto. The specific restaurants lie close to downtown Toronto.

However, they fall at the 18th rank for famous cuisines in Toronto and there are only 10 famous Indian restaurants in Toronto.

Based on this, we can say that the location will not have a big impact on the profits of the new restaurant apart from the fact that a commercial area like downtown Toronto will attract more crowd and is more likely to become famous.

So, we can conclude that if the chef is confident in his ability to impress people with his food, business in crowded areas like near a shopping center will flourish as there are not many well renowned Indian restaurants, however this choice will be risky.

A safer option is to open an Italian or Vietnamese restaurant if the chef can prepare this cuisine as these restaurants are very popular all-around Toronto.

## Future direction:

After deciding which option, the individual will make, more research on what people prefer in each cuisine will be conducted to refine the menu that will be served in the restaurant. Using the feedback and menus of famous restaurants from Foursquare we can retrieve what are the most liked dishes in the famous restaurant and align our menu to the same.