# 1. Introduction: Business Problem

In this project, our main objective is to look for an optimal location to open a Chinese restaurant. Main target audiences are aspiring chefs or business owners who are looking to open a Chinese restaurant in the city of Toronto.

One of the key considerations when deciding to open a restaurant is to firstly identify a location which is not overly crowded with existing restaurants as it means that competition will be greater. Specifically, the ideal location should have minimal or close to zero existing Chinese restaurants to minimize competition. Besides competition, the ideal location should also be in an area which is bustling with a significant population of Chinese race to cater to their palettes.

We will make use of data analytics to create a model that aims to recommend the ideal location which addresses the 3 key criteria identified above: (1) number of existing restaurants (2) number of existing Chinese restaurants (3) population with high number of Chinese. The final recommendations will be tabled together to allow targeted stakeholders to weigh out the pros and cons before deciding on the final location.

# 2. Data

- Neighborhood data for the city of Toronto. For this, there is a Wikipedia page that contains all the information that we need to explore and cluster the neighborhoods in Toronto https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Data of number of existing restaurants and type of restaurants in each neighborhood to be obtained from Foursquare API
- Coordinates of each neighborhood in Toronto: http://cocl.us/Geospatial_data
- Population demographics of each neighborhood. We obtained this from the open data portal of Toronto Statistics Department: https://open.toronto.ca/dataset/wellbeing-toronto-demographics/

## 2.1. Data Cleaning

To begin, we downloaded and installed all dependencies and packages which were required for this project. These include:

- Numpy
- Pandas
- Json
- Requests
- Matplotlib
- Kmeans needed for our K-means clustering
- Folium packages

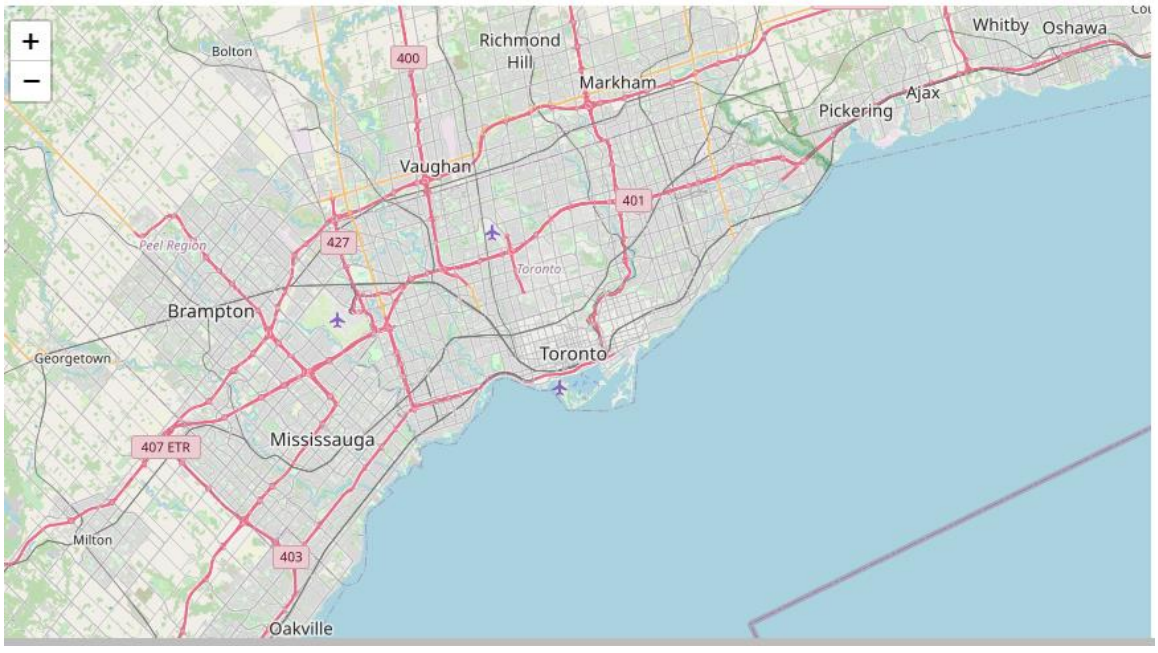
From the raw demographics data extracted as an excel format, this was downloaded into python and read into a pandas Dataframe. To ensure that we have only the relevant data which we required, we only extracted data involving Neighborhoods and the various Demographics (i.e Chinese, Latin American, Arabs etc.). To make the data more comprehensive, we downloaded OpenCage geocoder to obtain the coordinates of the various neighborhoods. These coordinates are necessary for plotting onto the Toronto map for visualization purposes in the later parts of the projects. Prior to this, we had to generate a key to access the OpenCage API. We extracted the coordinates and named this dataframe df_demo.

| | Neighborhood | Chinese | South Asian | Black | Filipino | Latin American | Southeast Asian | Arab | West Asian | Korean | Japanese | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Milliken | 16790 | 3780 | 1365 | 1145 | 125 | 290 | 170 | 135 | 20 | 60 | 43.823174 | -79.301763 |
| 2 | Steeles | 16705 | 1895 | 660 | 755 | 50 | 95 | 360 | 115 | 140 | 55 | 43.816178 | -79.314538 |
| 3 | Agincourt North | 16565 | 5160 | 1530 | 1355 | 230 | 230 | 370 | 75 | 155 | 135 | 43.808038 | -79.266439 |
| 4 | L'Amoreaux | 16455 | 8285 | 3875 | 1905 | 385 | 635 | 985 | 930 | 205 | 245 | 43.799003 | -79.305967 |
| 5 | Willowdale East | 14860 | 1700 | 845 | 520 | 440 | 190 | 485 | 3395 | 4265 | 285 | 43.761510 | -79.410923 |

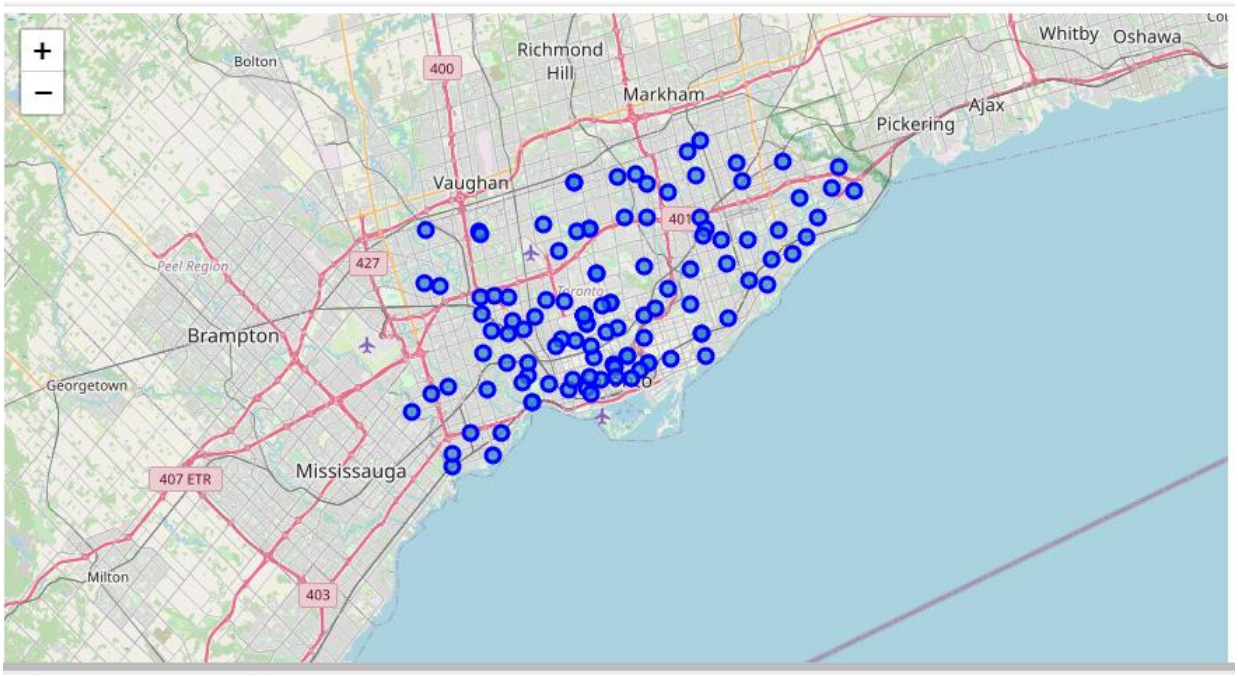
df_demo has a total of 140 rows and 13 columns.

## 2.2. Explore and Cluster Neighborhoods in Toronto

We first made use of Geopy library to obtain the latitude and longitude of Toronto. With that, the geographical coordinates of Toronto are 43.6534817, -79.3839347. Using the coordinates, we created a map of Toronto using Folium as seen below.



The next step was to add markers to the map of Toronto which superimposes the neighborhood on top. We named this map_toronto, as seen below.




In order to explore these neighborhoods further and segment them, we made use of Foursquare API to extract the information. Every time information is extracted, it requires the user's ID and secret passwords. With the credentials on hand, we did a trial by first exploring the first neighborhood in df_demo. The first neighborhood was Milliken with latitudes and longitudes of 43.8231743 and -79.3017626 respectively. Before extracting the details, a url had to be created and run through requests.get(url). A sample of the json results extracted is shown below.


```
{'meta': {'code': 200, 'requestId': '5ed323729da7ee001b6f27ce'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Scarborough',
  'headerFullLocation': 'Scarborough',
  'headerLocationGranularity': 'city',
  'totalResults': 40,
  'suggestedBounds': {'ne': {'lat': 43.8276743045, 'lng': -79.29553706206312},
   'sw': {'lat': 43.818674295499996, 'lng': -79.30798813793689}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '56945d1c498e11466e96405f',
       'name': 'Planet Fitness North Scarborough',
       'location': {'lat': 43.824095167666584,
        'lng': -79.30141064389495,
        'labeledLatLngs': [{'label': 'display',
          'lat': 43.824095167666584,
          'lng': -79.30141064389495}],
        'distance': 106,
```

To make the data more readable, we had to extract the required information (Venue Name, Venue Category, Latitude and Longitude) and structure it into a pandas dataframe. We named this dataframe nearby_venues.

| | name | categories | lat | lng |
|---|---|---|---|---|
| 0 | Planet Fitness North Scarborough | Gym | 43.824095 | -79.301411 |
| 1 | Deer Garden Signatures 鹿園魚湯米線 | Noodle House | 43.821898 | -79.298857 |
| 2 | Nichiban Sushi | Sushi Restaurant | 43.823172 | -79.306064 |
| 3 | Aka-Oni Izakaya | Japanese Restaurant | 43.822372 | -79.298905 |
| 4 | Sun's Kitchen 拉麵王 | Chinese Restaurant | 43.825282 | -79.306231 |
| 5 | Allan's Pastry Shop | Bakery | 43.820953 | -79.304564 |
| 6 | Uncle Tetsu's Japanese Cheesecake | Bakery | 43.825150 | -79.305954 |
| 7 | New Northern Dumplings 新北方餃子館 | Dumpling Restaurant | 43.821886 | -79.298751 |
| 8 | Kim Po Vietnamese Cuisine - 金寶越南美食 | Vietnamese Restaurant | 43.823292 | -79.305257 |
| 9 | Fish Ball Place 真之味小食屋 | Snack Place | 43.825290 | -79.306202 |
| 10 | Tim Hortons | Coffee Shop | 43.825423 | -79.297787 |
| 11 | Pacific Heritage Town 太古民族村 | Miscellaneous Shop | 43.825796 | -79.306242 |
| 12 | RBC Royal Bank | Bank | 43.825195 | -79.299575 |
| 13 | Brothers Bakery 兄弟餅店 | Bakery | 43.825443 | -79.306984 |
| 14 | Tak Fu Seafood Restaurant 德福點心皇 | Chinese Restaurant | 43.822633 | -79.298958 |
| 15 | Fat Ninja Bite | Japanese Restaurant | 43.823043 | -79.306446 |

After the trial exploration, we expanded our analysis to all neighborhoods in Toronto. We named this toronto_venues.

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Milliken | 43.823174 | -79.301763 | Planet Fitness North Scarborough | 43.824095 | -79.301411 | Gym |
| 1 | Milliken | 43.823174 | -79.301763 | Deer Garden Signatures 鹿園魚湯米線 | 43.821898 | -79.298857 | Noodle House |
| 2 | Milliken | 43.823174 | -79.301763 | Nichiban Sushi | 43.823172 | -79.306064 | Sushi Restaurant |
| 3 | Milliken | 43.823174 | -79.301763 | Aka-Oni Izakaya | 43.822372 | -79.298905 | Japanese Restaurant |
| 4 | Milliken | 43.823174 | -79.301763 | Sun's Kitchen 拉麵王 | 43.825282 | -79.306231 | Chinese Restaurant |

There were a total of 261 unique venue categories. We obtained this information by sorting Toronto_venues according to Neighborhoods using the groupby function and returning the total count for venue category per neighborhood. This was done using the count function.

## 2.3. Analyze Each Neighborhood in Toronto

The purpose of this section is to make use of onehot function to have a broad overview of what are the types of venues available in each neighborhood using a binary display (i.e 0 for not present; 1 for present). We labelled this new dataframe Toronto_grouped.

| | Neighborhood | Longitude | ATM | Accessories Store | Afghan Restaurant | American Restaurant | Animal Shelter | Antique Shop | Arcade | Art Gallery | Art Museum | Arts & Crafts Store | Asian Restaurant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Agincourt North | -2140.253999 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | Agincourt South-Malvern West | -317.040209 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | Alderwood | -795.453404 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | Annex | -3493.906322 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | Banbury-Don Mills | -396.783017 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

From here, we could see that the top 5 neighborhoods with most number of Chinese Restaurants are:

1. North Riverdale

2. South Riverdale

3. Milliken

4. L'Amoreaux

5. Tam O'Shanter-Sullivan

Per our introduction of our business problem, one of the criteria for choosing an optimal location was to avoid places which are currently packed with existing Chinese Restaurants and the above 5 will definitely not be part of the final choices.

We wanted a more in-depth breakdown of the top venues for each neighborhood. This will give us a clearer picture of which neighborhoods are currently crowded with existing Chinese Restaurants from those which are not. We printed out the top 5 venues for each neighborhood, sample as shown below.

```
----Agincourt North----
                venue  freq
0  Chinese Restaurant   2.0
1             Bakery   2.0
2               Bank   2.0
3     Discount Store   1.0
4           Pharmacy   1.0


----Agincourt South-Malvern West----
                      venue  freq
0                       ATM   1.0
1  Latin American Restaurant   1.0
2                    Lounge   1.0
3            Breakfast Spot   1.0
4               Pizza Place   0.0


----Alderwood----
          venue  freq
0   Pizza Place   2.0
1          Pool   1.0
2           Pub   1.0
3           Gym   1.0
4  Skating Rink   1.0
```

The above results were subsequently converted into a dataframe, labelled as Toronto_venues_sorted.

| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|
| 0 | Agincourt North | Bakery | Bank | Chinese Restaurant | Fast Food Restaurant | Frozen Yogurt Shop |
| 1 | Agincourt South-Malvern West | Breakfast Spot | Lounge | Latin American Restaurant | ATM | Fish & Chips Shop |
| 2 | Alderwood | Pizza Place | Coffee Shop | Dance Studio | Pub | Skating Rink |
| 3 | Annex | Pizza Place | Coffee Shop | Thai Restaurant | Indian Restaurant | Ice Cream Shop |
| 4 | Banbury-Don Mills | Botanical Garden | Park | Trail | Coffee Shop | Intersection |

To make our data more comprehensive for our model to run, we include attributes like number of Chinese and coordinates of each neighborhood into the above dataframe. The purpose of this was t have at one glance the top 5 most common venues in each neighborhood, the number of Chinese population in each neighborhood as well as the coordinates for plotting later. We labelled this new dataframe df_toronto_merged.

| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | No of Chinese | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Agincourt North | Bakery | Bank | Chinese Restaurant | Fast Food Restaurant | Frozen Yogurt Shop | 16565 | 43.808038 | -79.266439 |
| 1 | Agincourt South-Malvern West | Breakfast Spot | Lounge | Latin American Restaurant | ATM | Fish & Chips Shop | 9810 | 43.795223 | -79.260241 |
| 2 | Alderwood | Pizza Place | Coffee Shop | Dance Studio | Pub | Skating Rink | 70 | 43.601717 | -79.545232 |
| 3 | Annex | Pizza Place | Coffee Shop | Thai Restaurant | Indian Restaurant | Ice Cream Shop | 1695 | 43.670338 | -79.407117 |
| 4 | Banbury-Don Mills | Botanical Garden | Park | Trail | Coffee Shop | Intersection | 3535 | 43.734804 | -79.357243 |

Before fitting the dataframe into our model, we had to ensure that the data type is correctly classified. Initial checks showed that No of Chinese is categorized as object, which is incorrect. This had to be changed to an integer. After the amendments, data types have been confirmed to be correct.

```
Neighborhood              object
1st Most Common Venue     object
2nd Most Common Venue     object
3rd Most Common Venue     object
4th Most Common Venue     object
5th Most Common Venue     object
No of Chinese              int64
Latitude                 float64
Longitude                float64
dtype: object
```

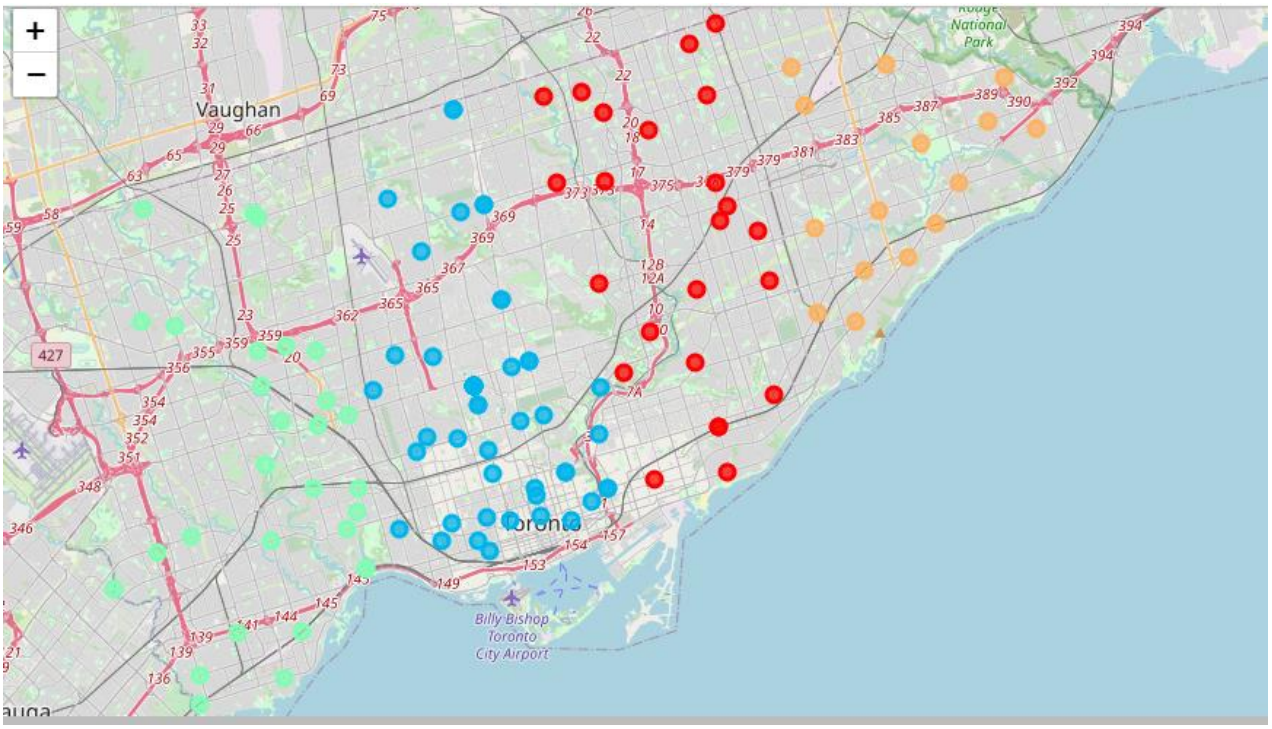
# 3. Model

In this project, we will be using k-means clustering to cluster the neighborhoods into 5 clusters. The neighborhoods were plotted based on their latitude and longitude and clustered accordingly. The cluster labels were included into the dataframe.

| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | No of Chinese | Latitude | Longitude | Cluster Labels |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Agincourt North | Bakery | Bank | Chinese Restaurant | Fast Food Restaurant | Frozen Yogurt Shop | 16565 | 43.808038 | -79.266439 | 4 |
| 1 | Agincourt South-Malvern West | Breakfast Spot | Lounge | Latin American Restaurant | ATM | Fish & Chips Shop | 9810 | 43.795223 | -79.260241 | 4 |
| 2 | Alderwood | Pizza Place | Coffee Shop | Dance Studio | Pub | Skating Rink | 70 | 43.601717 | -79.545232 | 3 |
| 3 | Annex | Pizza Place | Coffee Shop | Thai Restaurant | Indian Restaurant | Ice Cream Shop | 1695 | 43.670338 | -79.407117 | 2 |
| 4 | Banbury-Don Mills | Botanical Garden | Park | Trail | Coffee Shop | Intersection | 3535 | 43.734804 | -79.357243 | 0 |

## 3.1. Visualising the Model

We plotted the clusters onto the map of Toronto.




The dataframe was sorted in accordance to the cluster labels to determine the number of neighborhoods in each cluster.

| Cluster Labels | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | No of Chinese | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 71 | 71 | 71 | 71 | 71 | 71 | 71 | 71 | 71 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| 3 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 |
| 4 | 32 | 32 | 32 | 32 | 32 | 32 | 32 | 32 | 32 |

In the following sections, we will be analyzing each of the clusters to determine its suitability as a location for setting up a new Chinese Restaurant.

## 3.2. Analyzing Cluster 1

For each cluster, we extracted the top 5 most common venues across all the neighborhoods within the cluster along with the statistics for number of Chinese in that cluster.

| 1st Most Common Venue in Cluster 1 | 2nd Most Common Venue in Cluster 1 | 3rd Most Common Venue in Cluster 1 | 4th Most Common Venue in Cluster 1 | 5th Most Common Venue in Cluster 1 |
|---|---|---|---|---|
| Italian Restaurant | Chinese Restaurant | Mobile Phone Shop | Pharmacy | Cafe |

- Max number of Chinese: 14860
- Min number of Chinese: 75
- Mean number of Chinese: 1447

The neighborhoods in Cluster 1 are 'Banbury-Don Mills', 'Bayview Village', 'Bayview Woods-Steeles', 'Da nforth', 'Danforth-East York', 'Don Valley Village', 'Dorset Park', 'East End-Danforth', 'Flemingdon Park', ' Henry Farm', 'Hillcrest Village', 'Ionview', "L'Amoreaux", 'Milliken', "O'Connor-Parkview", 'Oakridge', 'Ple asant View', 'Steeles', "Tam O'Shanter-Sullivan", 'Taylor-Massey', 'The Beaches', 'Thorncliffe Park', 'Victor ia Village', 'Wexford/Maryvale'.

## 3.3. Analyzing Cluster 2

As there is only 1 neighborhood under Cluster 2, we will not by performing the statistical analysis of number of chinese and the mode of each most common venue.

| 1st Most Common Venue in Cluster 2 | 2nd Most Common Venue in Cluster 2 | 3rd Most Common Venue in Cluster 2 | 4th Most Common Venue in Cluster 2 | 5th Most Common Venue in Cluster 2 |
|---|---|---|---|---|
| Grocery Store | Shopping Mall | Japanese Restaurant | Furniture/Home Store | Print Shop |

- Number of Chinese: 3065

The neighborhood in Cluster 2 is 'St.Andrew-Windfields'.

## 3.4. Analyzing Cluster 3

| 1st Most Common Venue in Cluster 3 | 2nd Most Common Venue in Cluster 3 | 3rd Most Common Venue in Cluster 3 | 4th Most Common Venue in Cluster 3 | 5th Most Common Venue in Cluster 3 |
|---|---|---|---|---|
| Fast Food Restaurant | Bank | Discount Store | Beer Store | Department Store |

- Max number of Chinese: 4620
- Min number of Chinese: 320
- Mean number of Chinese: 1465

The neighborhoods in Cluster 3 are 'Annex', 'Bathurst Manor', 'Bay Street Corridor', 'Bedford Park-Nortown', 'Beechborough-Greenbrook', 'Birchcliffe-Cliffside', 'Blake-Jones', 'Briar Hill-Belgravia', 'Bridle Path-Sunnybrook-York Mills', 'Broadview North', 'Brookhaven-Amesbury', 'Cabbagetown-South St.James Town', 'Caledonia-Fairbank', 'Casa Loma', 'Church-Yonge Corridor', 'Clairlea-Birchmount', 'Clanton Park', 'Corso Italia-Davenport', 'Dovercourt-Wallace Emerson-Juncti', 'Downsview-Roding-CFB', 'Dufferin Grove', 'Englemount-Lawrence', 'Eringate-Centennial-West Deane', 'Forest Hill North', 'Forest Hill South', 'Greenwood-Coxwell', 'Humbermede', 'Humewood-Cedarvale', 'Kensington-Chinatown','Kingsview Village-The Westway', 'Lansing-Westgate', 'Lawrence Park North', 'Lawrence Park South', 'Leaside-Bennington', 'Little Portugal', 'Moss Park', 'Mount Olive-Silverstone-Jamestown', 'Mount Pleasant East', 'Mount Pleasant West', 'Newtonbrook East', 'Newtonbrook West', 'Niagara', 'North Riverdale', 'North St.James Town', 'Oakwood Village', 'Old East York', 'Palmerston-Little Italy', 'Parkwoods-Donalda', 'Playter Estates-Danforth', 'Princess-Rosethorn', 'Regent Park', 'Rockcliffe-Smythe', 'Roncesvalles', 'Rosedale-Moore Park', 'South Riverdale', 'Thistletown-Beaumond Heights', 'Trinity-Bellwoods', 'University', 'Waterfront Communities-The Island', 'West Humber-Clairville', 'Westminster-Branson', 'Weston-Pellam Park', 'Willowdale East', 'Willowdale West', 'Willowridge-Martingrove-Richview', 'Woodbine Corridor', 'Woodbine-Lumsden', 'Wychwood', 'Yonge-Eglinton', 'Yonge-St.Clair', 'Yorkdale-Glen Park'.

## 3.5. Analyzing Cluster 4

| 1st Most Common Venue in Cluster 4 | 2nd Most Common Venue in Cluster 4 | 3rd Most Common Venue in Cluster 4 | 4th Most Common Venue in Cluster 4 | 5th Most Common Venue in Cluster 4 |
|---|---|---|---|---|
| Fast Food Restaurant | Coffee Shop | Bus Line | Park | Chinese |

- Max number of Chinese: 16790
- Min number of Chinese: 595
- Mean number of Chinese: 5324

The neighborhoods in Cluster 4 are 'Alderwood', 'Black Creek', 'Edenbridge-Humber Valley', 'Elms-Old Rexdale', 'Etobicoke West Mall', 'Glenfield-Jane Heights', 'High Park North', 'High Park-Swansea', 'Humber Heights-Westmount', 'Humber Summit', 'Islington-City Centre West', 'Junction Area', 'Keelesdale-Eglinton West', 'Kingsway South', 'Lambton Baby Point', 'Long Branch', 'Maple Leaf', 'Markland Wood', 'Mimico', 'Mount Dennis', 'New Toronto', 'Pelmo Park-Humberlea', 'Rexdale-Kipling', 'Runnymede-Bloor West Village', 'Rustic', 'South Parkdale', 'Stonegate-Queensway', 'Weston', 'York University Heights'.

## 3.6. Analyzing Cluster 5

| 1st Most Common Venue in Cluster 5 | 2nd Most Common Venue in Cluster 5 | 3rd Most Common Venue in Cluster 5 | 4th Most Common Venue in Cluster 5 | 5th Most Common Venue in Cluster 5 |
|---|---|---|---|---|
| Coffee Shop | Coffee Shop | Discount Store | Cafe | Doner Restaurant |

- Max number of Chinese: 9810
- Min number of Chinese: 50
- Mean number of Chinese: 850

The neighborhoods in Cluster 5 are ‘Agincourt North', 'Agincourt South-Malvern West', 'Bendale', 'Centennial Scarborough', 'Cliffcrest', 'Eglinton East', 'Guildwood', 'Highland Creek', 'Kennedy Park', 'Malvern', 'Morningside', 'Rouge', 'Scarborough Village', 'West Hill', 'Woburn'.

# 4. Analysis and Conclusion

For cluster 1, one of the glaring features is that Chinese Restaurants is the top 2 most common venues, coupled with the acceptable number of Chinese (Mean of 1447), does not make it an ideal cluster to set up another Chinese Restaurant. Competition is expected to be tough and the restaurants have to compete with an average number of Chinese populace.

Even though Cluster 5 does not have any Chinese Restaurants as its top 5 most common venues, the mean number of Chinese populace is only 850. This also makes it not an ideal location as the pool of potential customers will be low.

Cluster 3 does not have any Chinese Restaurants as its top 5 most common venues but its mean number of Chinese populace of 1465 does not make it an ideal location as well.

Eventually, we narrowed our choices to either Cluster 2 or 4. Cluster 2 has a respectable number of Chinese at 3065 with no Chinese Restaurants at its top venues. Cluster 4, albeit having Chinese Restaurants as its 5th most common venue, this is more than compensated by high average number of Chinese of 5324. Hence, the aspiring Chinese chef or business owner can select neighborhoods from either Cluster 2 or 4.