The Battle of Neighborhoods

Analysis of Opening Japanese Restaurant in Toronto

ABSTRACT

This is a report from my project - Analysis of Opening Japanese Restaurant in Toronto and this is part of my Data Science Professional Certification from IBM at Coursera.

AUTHOR

Paulina Kossowska

Table of Contents

1. Introduction 2
   1.1. Background 2
   1.2. Problem Description 2
   1.3. Target Audience 2
2. Data Preparation 3
   2.1. Data Sources 3
   2.2. Data Cleaning 4
      2.2.1. WebScrapping using BeautifulSoup 4
      2.2.2. GeoSpatial DataSet 5
      2.2.3. Toronto GeoJSON file 5
      2.2.4. Number of Japanese Restaurants in Toronto: using Foursquare API 5
      2.2.5. Coordinates: City of Toronto 6
      2.2.6. East and Southeast Asian origins Population in Toronto 6
      2.2.7. Business Improvement Area 7
3. Methodology 8
4. Exploratory Data Analysis 9
   4.1. Explore the neighbourhoods in Toronto 9
   4.2. Explore East and Southeast Asian population in Toronto 9
   4.3. Explore Japanese Restaurants in Toronto 11
   4.4. Create a choropleth map of Toronto with Japanese Restaurants 12
   4.5. Create a heatmap for Japanese Restaurants 13
   4.6. Create map of Business Improvement Area of Toronto with Japanese Restaurants 14
5. Density based Clustering DBSCAN 15
   5.1. Data preparation for DBSCAN 15
   5.2. DBSCAN Modelling 16
   5.3. Distinguish outliers 17
   5.4. DBSCAN Visualization 18
6. Discussion 18
7. Conclusion 21
8. Limitations 22
9. Reference 22

1. Introduction

1.1. Background

Toronto - the city of possibilities. Toronto is Canada's largest city and a world leader in such areas as business, finance, technology, entertainment and culture. Its large population of immigrants from all over the globe has also made Toronto one of the most multicultural cities in the world. This veracity can be seen in the name of Toronto's neighbourhoods like: Little Japan, east Chinatown, Greektown, Corso Italia and many more. What also distinguishes Toronto is the fact that it is the most populous city in Canada and the fourth most populous city in North America. According to the information's set at Wikipedia page the population of the city is over 6 billion. All these factors make Toronto an attractive place to live, work or start a business. What is definitely worth mentioning is the fact that thanks to the multiculturalism of this city, we are able to taste the best flavours of cuisines from around the world.

1.2. Problem Description

Imagine a situation where there is a descendant of immigrants from Japan (let's call him Paul) and he is currently living in the most density city in Canada - Toronto. Paul has a lot of doubts where he should open his business. Through this analysis I will help Paul decide to choose the best neighborhood in Toronto for his Japanese restaurant.

1.3. Target Audience

The aimed audience for this project could be an entrepreneur or business owner who want to open a Japanese restaurant in Toronto or who is interesting to improve the business proficiency of existing business. I believe that the result of this analysis will be beneficial for mentioned groups.

2. Data Preparation

2.1. Data Sources

In this project I will use Foursquare location data and machine learning clustering algorithm to find the best place, the best neighborhood in Toronto to open a Japanese restaurant. I will find the most suitable location for an entrepreneur to open a new Japanese restaurant in Toronto, Canada.

For analysis in this project I will use the following data:

1. List of postal codes from Toronto
    a. Data source: WebScrapping methods from wikipedia page
    b. Data description: Dataset contains information about different borough and their neighbourhood's in Toronto
2. Geographical coordinates of each neighbourhood's in Toronto
    a. Data source: csv file named **Geospatial_Coordinates.csv** provided by Coursera
    b. Data description: GeoSpatial Dataset will be use to get information about coordinates of neighbourhoods of Toronto
3. Japanese restaurants in each neighbourhood's in Toronto
    a. Data source: Foursquare API
    b. Data description: By using Foursquare API I will get all venues in different neighbourhood's in Toronto and filter it to only Japanese restaurant
4. Toronto GeoJSON file
    a. Data source: City of Toronto Open Data - click here and here
    b. Data description: I will use this GeoJSON file to create choropleth map for Toronto
5. Immigration and ethnocultural diversity statistics
    a. Data source: City of Toronto Open Data - click here
    b. Data description: I will use this information to create a population density choropleth map

6. Business Improvement Areas: is an association of commercial property owners and tenants within a defined area who work in partnership with the City to create thriving, competitive, and safe business areas that attract shoppers, diners, tourists, and new businesses
    a. Data source: City of Toronto Open Data - click here
    b. Data description: I will use GeoJSON file of BIA to create a population density choropleth map

2.2. Data Cleaning

In this section I will use different methods to obtain all data mentioned in section above.

2.2.1. WebScrapping using BeautifulSoup

Beautiful Soup is a Python library for pulling data out of HTML and XML files. After scratching Wikipedia page, I did some cleaning techniques to get clear table with Toronto's Postal Codes. First, I removed all borough where value was not assigned. In the situation were neighborhood value was not assigned, I took the value from borough column and finally if postal code appears more than once and had more than one neighborhood, then I combined them into one row with the neighbourhood's separated with a comma.

The final results look as follows:

5) Finally I will check the efects of processing above steps

```
df = results.reset_index()
df.head(10)
```

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights, Lawrence Manor
4	M7A	Downtown Toronto	Queen's Park
5	M9A	Etobicoke	Islington Avenue
6	M1B	Scarborough	Rouge, Malvern
7	M3B	North York	Don Mills North
8	M4B	East York	Woodbine Gardens, Parkview Hill
9	M5B	Downtown Toronto	Ryerson, Garden District

```
df.shape
```

```
(103, 3)
```

2.2.2. GeoSpatial DataSet

I will use GeoSpatial DataSet provided by Coursera and merged it with previous dataframe. It will give me information about latitude and longitude for each postal code.

```
toronto_df = pd.merge(df, geo, on='PostalCode')
toronto_df.head()
```

	PostalCode	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	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494

```
toronto_df.shape
```

```
(103, 5)
```

2.2.3. Toronto GeoJSON file

In this section I loaded information from Open Data City of Toronto into two variables called **geo_toronto** and **ward_geo**. Later, I will use them to create choropleth maps for Toronto.

2.2.4. Number of Japanese Restaurants in Toronto: using Foursquare API

After defined Foursquare credentials and version I was able to scratched information about all Japanese restaurants in different neighbourhood's in Toronto. I created dataframe **toronto_japan** and loaded it into csv file. Anyway, some cleaning techniques must occur because it is turned out that primary table contains information about 20 different venues. I was only interested in Japanese Restaurant so I dropped any irrelevant data. In final I got the following table:

```
toronto_japan.shape
```

```
(1498, 7)
```

```
toronto_japan.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Matsuda Japanese Cuisine & Teppanyaki	43.745494	-79.345821	Japanese Restaurant
1	Parkwoods	43.753259	-79.329656	Gonoe Sushi	43.745737	-79.345991	Japanese Restaurant
2	Parkwoods	43.753259	-79.329656	Sushi Ichiban	43.758912	-79.310811	Japanese Restaurant
3	Parkwoods	43.753259	-79.329656	Teriyaki Experience	43.754468	-79.351424	Japanese Restaurant
4	Parkwoods	43.753259	-79.329656	Katsura Japanese Restaurant 桂	43.756259	-79.349571	Japanese Restaurant

2.2.5. Coordinates: City of Toronto

I used here GeoPy package to locate the coordinates of Toronto as follow:

2.2.5. Coordinates: City of Toronto ¶

```
address = 'Toronto, ON'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto are {}, {}.'.format(latitude, longitude))
```

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

2.2.6. East and Southeast Asian origins Population in Toronto

Data about different population in neighbourhood's in Toronto was obtained from Open Data City of Toronto. City of Toronto Neighbourhood Profiles use Census data to provide a portrait of the demographic, social and economic characteristics of the people and households in each City of Toronto neighbourhood. Data are stored in csv file called **neighbourhood-profiles-2016-csv**.

	_id	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	Banbury-Don Mills	Bathurst Manor	Bay S Cor
0	1	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Number	NaN	129	128	20	95	42	34	
1	2	Neighbourhood Information	Neighbourhood Information	City of Toronto	TSNS2020 Designation	NaN	No Designation	No Designation	No Designation	No Designation	No Designation	No Designation	Desig
2	3	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2016	2,731,571	29,113	23,757	12,054	30,526	27,695	15,873	2
3	4	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2011	2,615,060	30,279	21,988	11,904	29,177	26,918	15,434	1
4	5	Population	Population and dwellings	Census Profile 98-316-X2016001	Population Change 2011-2016	4.50%	-3.90%	8.00%	1.30%	4.60%	2.90%	2.80%	3

```
eastasia.shape
```

```
(2383, 146)
```

The primary table needed some cleaning techniques to obtain information only about East and Southeast Asian origins I am interested to analysis. In the result I got a clean table containing information about number of populations in different neighbourhood's in Toronto. What have to be noticed here is the fact that according to the Census 2016 the Toronto was split into 140 neighbourhoods.

	Neighborhood	Neighbourhood Number	Population
6	Agincourt North	129.0	18575.0
7	Agincourt South-Malvern West	128.0	13075.0
8	Alderwood	20.0	1295.0
9	Annex	95.0	3945.0
10	Banbury-Don Mills	42.0	6920.0

```
eastasia.shape
```

```
(140, 3)
```

2.2.7. Business Improvement Area

In this section I loaded information from Open Data City of Toronto into **geo_toronto** variable Later, I will use them to create choropleth maps for Toronto.

3. Methodology

As I am mentioned at the beginning of this project Toronto is the most populous city in Canada and the fourth most populous city in North America. According to the information posted at webpage in 2018 the population in city centre of Toronto was about 3 million people and in entire Toronto region more than 6 million people. Since the second half of the 20th century the city has grown phenomenally, from a rather sedate provincial town – "Toronto the Good" - to a lively, thriving, cosmopolitan metropolitan area.

My first step in discovering potential areas for a new Japanese restaurant in Toronto will be an exploratory data analysis. Through this I will visualise interesting perceptions like:

- total number of neighbourhood's in different Toronto boroughs
- explore more about East and Southeast Asia population in Toronto and I will try to find the most and lowest dense neighbourhood's in Toronto by these origins
- total number of Japanese restaurants in different neighbourhood's and boroughs and also plot all of this information inside choropleth map using heatmap plugins to try find patterns in Japanese restaurants location in Toronto city
- using Business Improvement Area data, I will show on choropleth map how they look like in Toronto comparing to localisation of Japanese restaurants

My second step in this analysis will be performing machine learning algorithm. For the purpose of this project I choose DBSCAN: a density-based clustering algorithm which is appropriate to use when examining spatial data.

This analysis will help me identifying different cluster groups and defined it as potential candidates' locations for a new Japanese restaurant in Toronto.

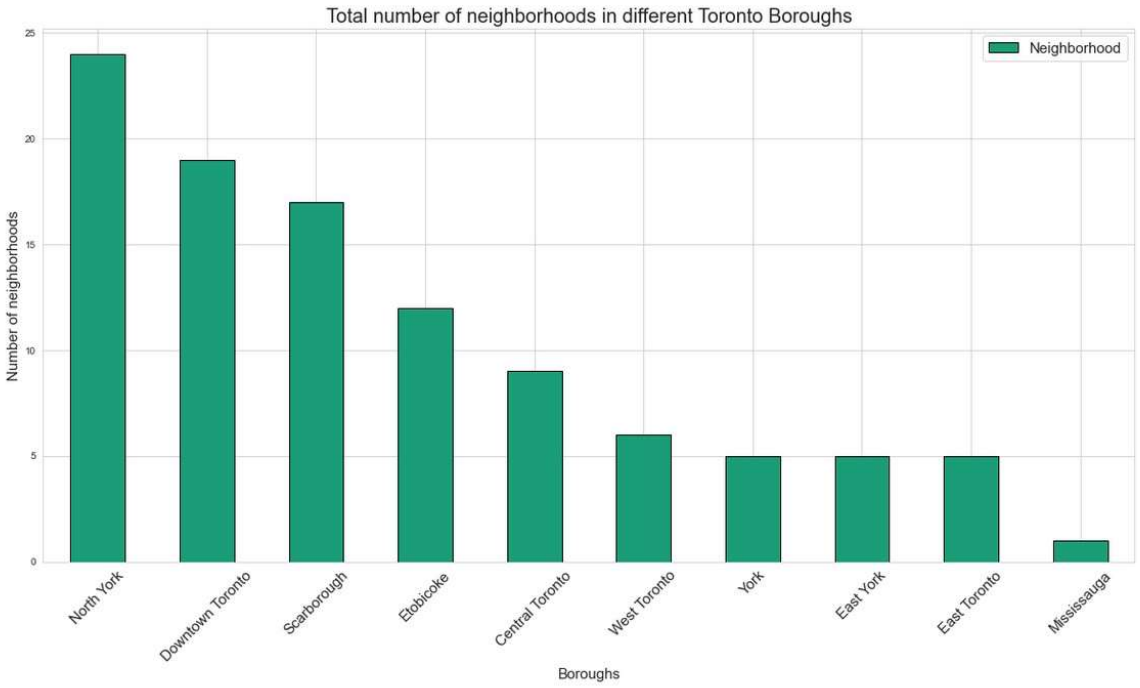
4. Exploratory Data Analysis

Having all of this data into different dataframes allows me to perform some analysis to get a look of what kind of data we have and work on. I divided this into different sections containing different type of visualisation.

4.1. Explore the neighbourhoods in Toronto

Before I started my work, I wanted to know a little more about neighbourhoods in Toronto. I found out that there are 10 Boroughs in Toronto: North York, Downtown Toronto, Etobicoke, Scarborough, East York, York, East Toronto, West Toronto, Central Toronto, Mississauga.

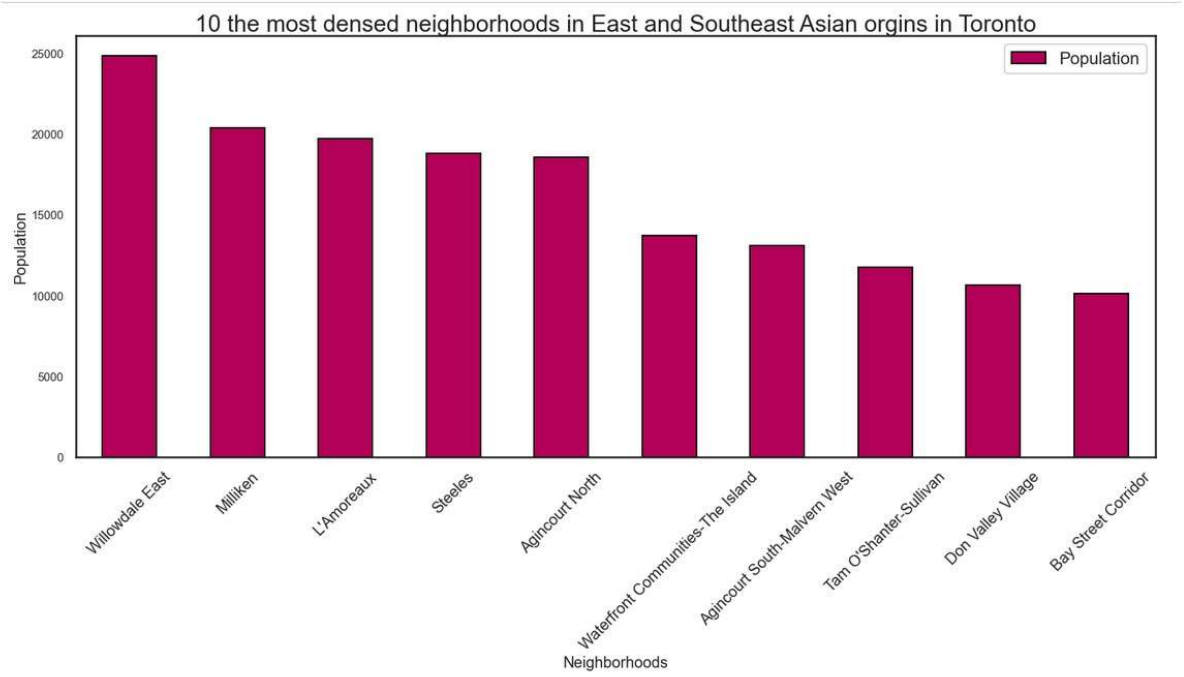
I also wanted to know which one had the highest number of neighbourhoods inside them. It come clear to me that these were: North York, Downtown Toronto and Scarborough.

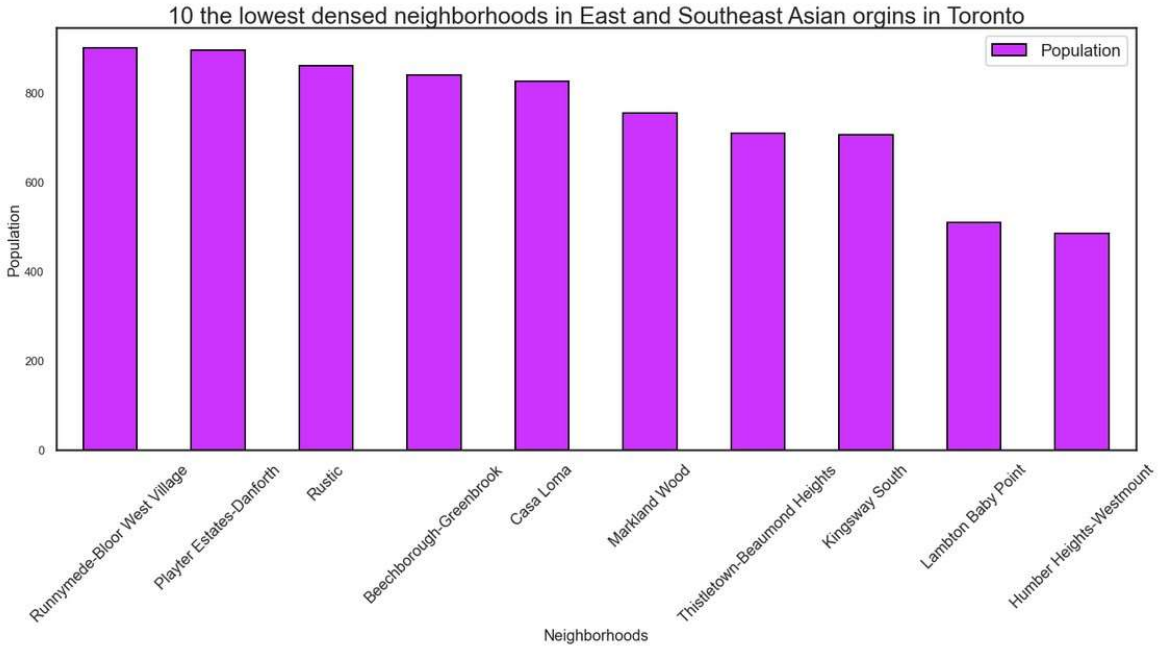


4.2. Explore East and Southeast Asian population in Toronto

In this section I concentrated to explore more from **eastasia** dataframe. I was interested which neighbourhoods were the most and the lowest populated by

East and Southeast Asian origins. I built two bar charts showing mentioned tendency.





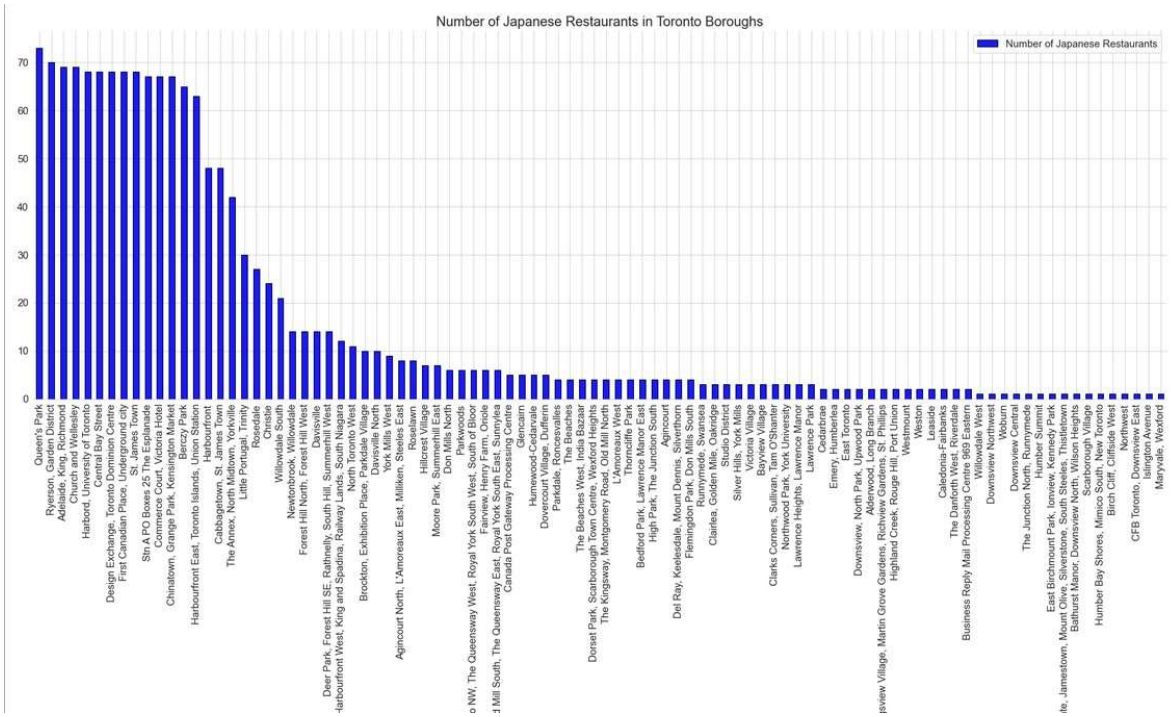
Further analysis shows that the most densely populated areas are: North York, Scarborough, Don Valley Village, Markham, Toronto Islands and Downtown Toronto. In the other hand Parkdale - High Park, Toronto

Danforth, York South-Weston, Etobicoke are the poorly densely populated areas in East and Southeast Asian origins.

4.3. Explore Japanese Restaurants in Toronto

This part of project is very interesting. After cleaning **toronto_japan** dataframe I was able to took a deeper look inside. I was curious about two things: how many Japanese restaurants were in each neighborhood in Toronto and of course how those numbers look for different boroughs.

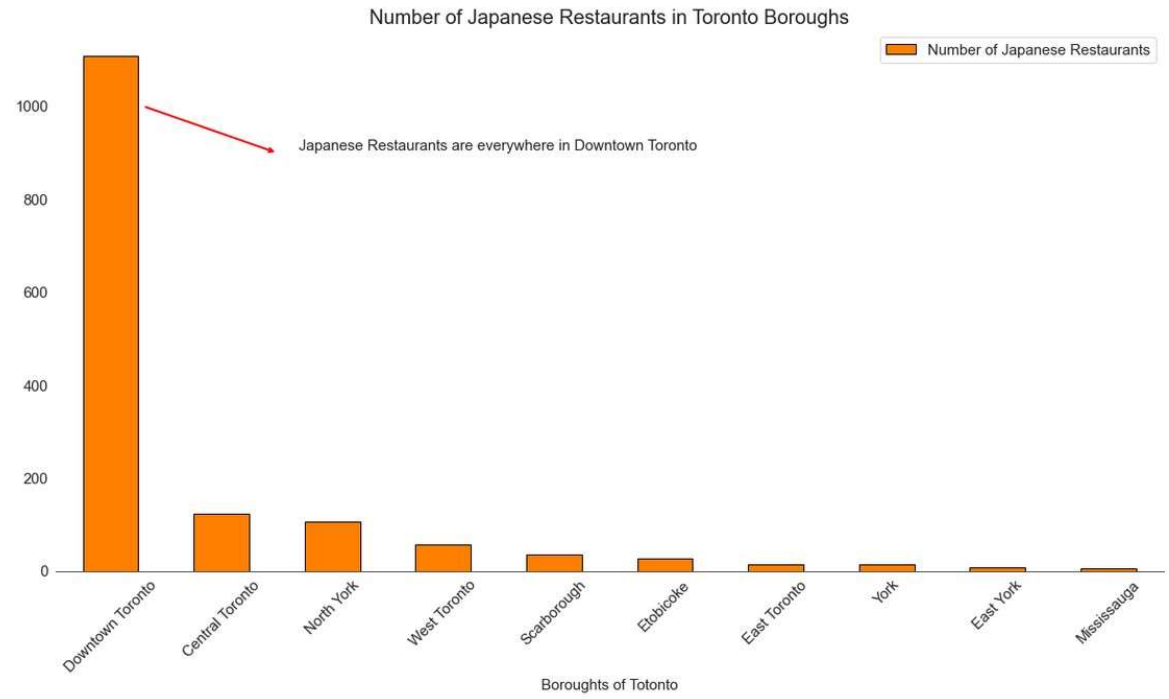
I got the following results:



From the first graph above we know that there is 1 neighbourhood where number of Japanese restaurants is above 70, 13 where number of Japanese restaurants is above 60 and 16 where the number is equal to 1. However further analysis is difficult because the total number of neighbourhoods in Toronto is 103, so better idea is to get look at total number of Japanese restaurants in different boroughs in Toronto.

In the other hand the second graph below shows us total number of Japanese restaurants in different Toronto's borough. It is clearer and more visible to

analysis. We can observe that the highest number of Japanese restaurants are in Downtown Toronto with total number equal to 1075.

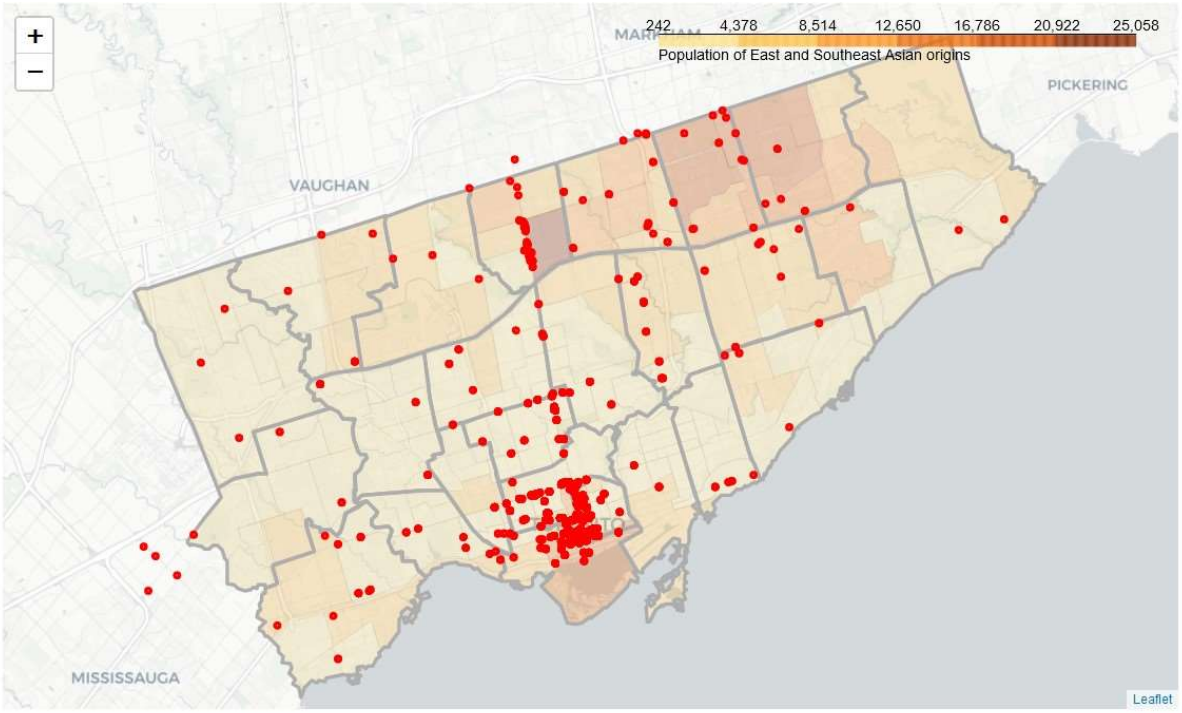


On the second, third and fourth place with highest number of Japanese restaurants are: Central Toronto, North York and West Toronto.

What is absolutely worth to notice is a fact that the density of Japanese restaurants is the highest in central Toronto boroughs with relatively small number of Japanese restaurants on the outskirts of Toronto.

4.4. Create a choropleth map of Toronto with Japanese Restaurants

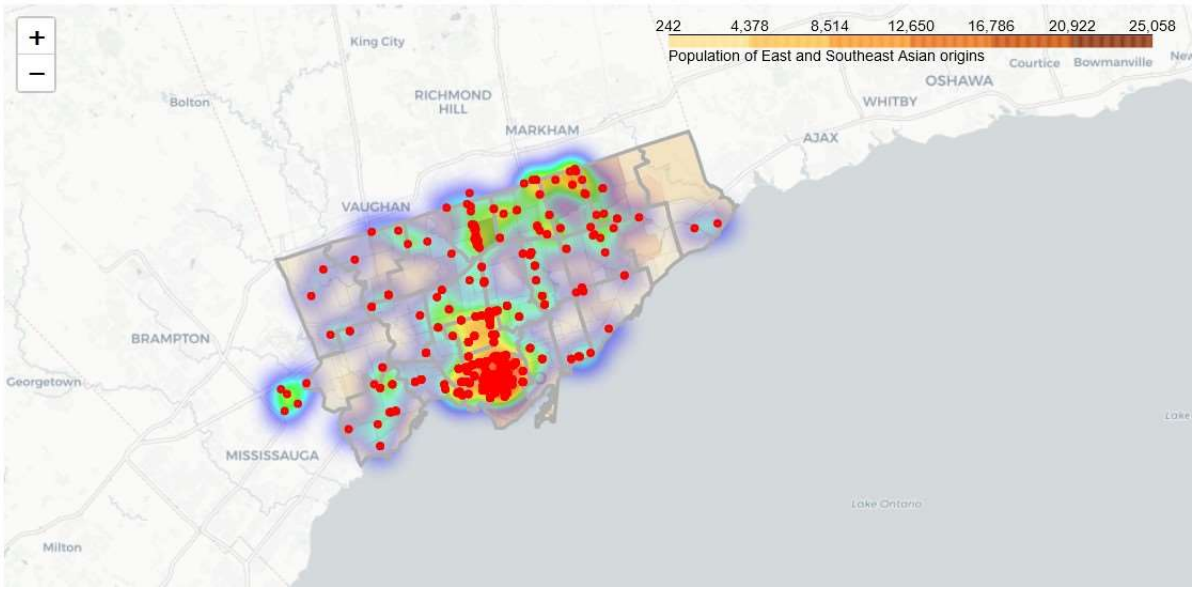
Folium makes it easy to visualize data that's been manipulated in Python. I used this library many times in this project to visualise some interesting maps of Toronto. One of them was built by using GeoJSON data of Toronto and eastasia dataframe containing population information. The darker the colour on the map, the more densely populated the area is. Additionally, I added information about Toronto ward and split whole Toronto city into wards created based on last Census in 2016. The red dots represent Japanese restaurants in city of Toronto.



According to the map above we can conclude that the higher number of Japanese restaurants is positively correlated with East and Southeast Asian population in Toronto. We can also see high density of Japanese restaurants in city centre of Toronto.

4.5. Create a heatmap for Japanese Restaurants

Folium plugins HeatMap will give better insights in density of Japanese restaurants in Toronto as shown below.

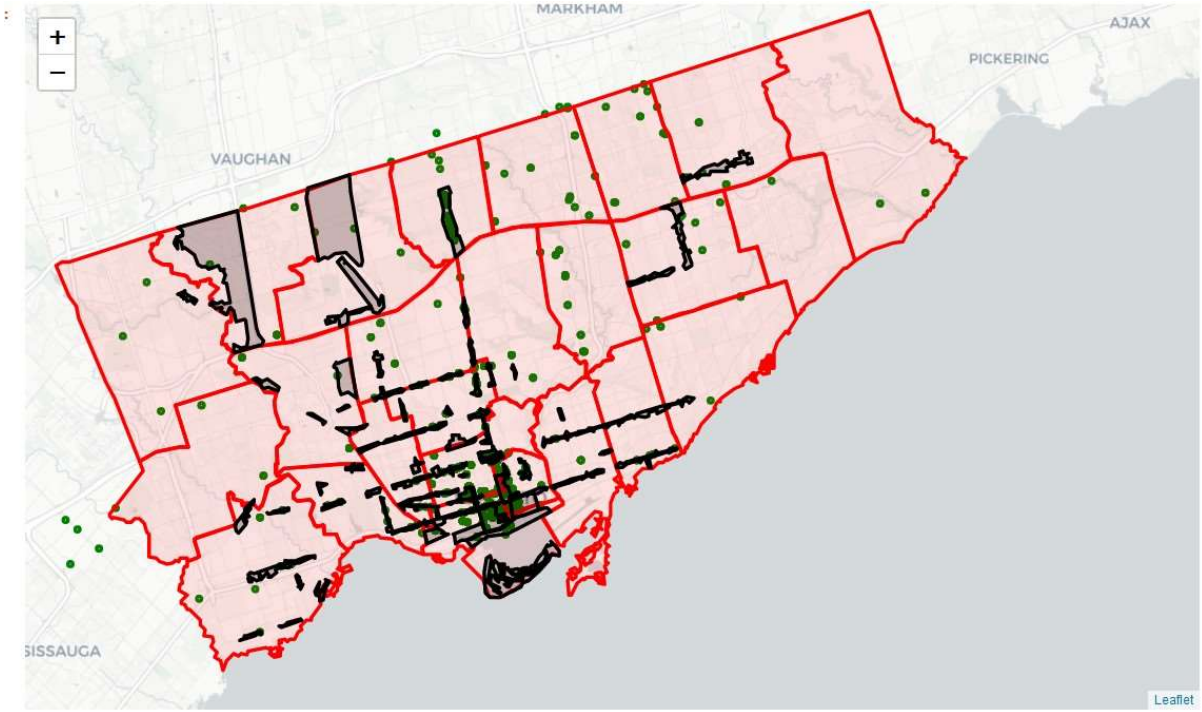


4.6. Create map of Business Improvement Area of Toronto with Japanese Restaurants

A Business Improvement Area (BIA) is an association of commercial property owners and tenants within a defined area who work in partnership with the City to create thriving, competitive, and safe business areas that attract shoppers, diners, tourists, and new businesses.

Data needed to create this business area was obtained from Open Data City of Toronto and like wards information was also created based on last Census in 2016.

This time a green dot represents Japanese restaurants. The map of Toronto was split into wards. The black areas on this map represent mentioned Business Improvement Areas.



We can see that Japanese restaurants are highly located in city centre of Toronto and this is also located in BIA areas. What we can also clearly see and this map is another proof of it, is a fact poorly located Japanese restaurants on the outskirts of Toronto.

5. Density based Clustering DBSCAN

DBSCAN - a density-based clustering algorithm which is appropriate to use when examining spatial data. Most of the traditional clustering techniques such as K-Means clustering can be used to group data in an unsupervised way, however, when applied to tasks with arbitrary shaped clusters or clusters within clusters, traditional techniques might not be able to achieve good results, and elements in the same cluster might not share enough similarity or the performance may be poor. Density-based clustering algorithms locate regions of high density that are separated from one another by regions of low density

Advantages of using DBSCAN algorithm:

- unlike K-means, DBSCAN does not require the user to specify the number of clusters to be generated
- DBSCAN can find any shape of clusters. The cluster doesn't have to be circular.
- DBSCAN can identify outliers

5.1. Data preparation for DBSCAN

To perform this machine learning algorithm, I will use data from **toronto_japan** dataframe. First, I will clean coordinates data for each venue in this dataframe.

```
xs, ys = np.asarray(toronto_japan['Venue Latitude']), np.asarray(toronto_japan['Venue Longitude'])
toronto_japan['xm'] = xs.tolist()
toronto_japan['ym'] = ys.tolist()
```

```
cluster_data = toronto_japan[['xm', 'ym']]
cluster_data = np.nan_to_num(cluster_data)
cluster_data
```

```
array([[ 43.74549367, -79.34582141],
       [ 43.74573682, -79.3459912 ],
       [ 43.75891159, -79.31081063],
       ...,
       [ 43.6247282 , -79.50990398],
       [ 43.625863  , -79.504237  ],
       [ 43.62598209, -79.50349796]])
```

5.2. DBSCAN Modelling

As is the case in most machine learning algorithm, the model's behaviour is dictated by several parameters. In this example, I will touch the following three:

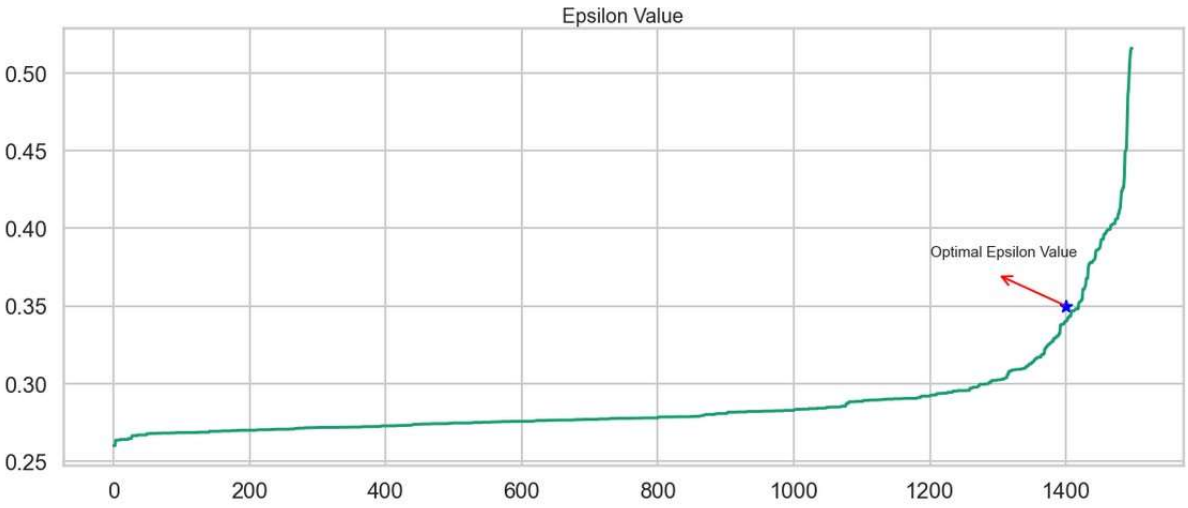
- **eps:** two points are considered neighbours if the distance between the two points is below the threshold epsilon
- **min_samples:** the minimum number of neighbours a given point should have in order to be classified as core point
- **metric:** the metric to use when calculating distance between instances in a feature array (i.e. **default: Euclidean distance**)

One limitation of DBSCAN is that it is sensitive to the choice of eps, in particular if clusters have different densities. If eps are too small, sparser clusters will be defined as noise. If eps are too large, denser clusters may be merged together.

I can calculate the distance from each point to its closest neighbour using the **NearestNeighbors**. The point itself is included in n_neighbors. The kneighbors method returns two arrays, one which contains the distance to the closest n_neighbors points and the other which contains the index for each of those points.

After sorting I was able to plot the results.

The optimal value for epsilon will be found at the point of maximum curvature. In this case **eps = 0.35.**



5.3. Distinguish outliers

I used min_samples as 6 and proceed the following code to identify outliers and gave them number -1.

```
core_samples_mask = np.zeros_like(db.labels_, dtype=bool)
core_samples_mask[db.core_sample_indices_] = True
labels = db.labels_
toronto_japan["Cluster_DB"]=labels

realClusterNum=len(set(labels)) - (1 if -1 in labels else 0)
clusterNum = len(set(labels))
```

After performing DBSCAN I finally got a table containing information about different cluster number for each venue located in Toronto.

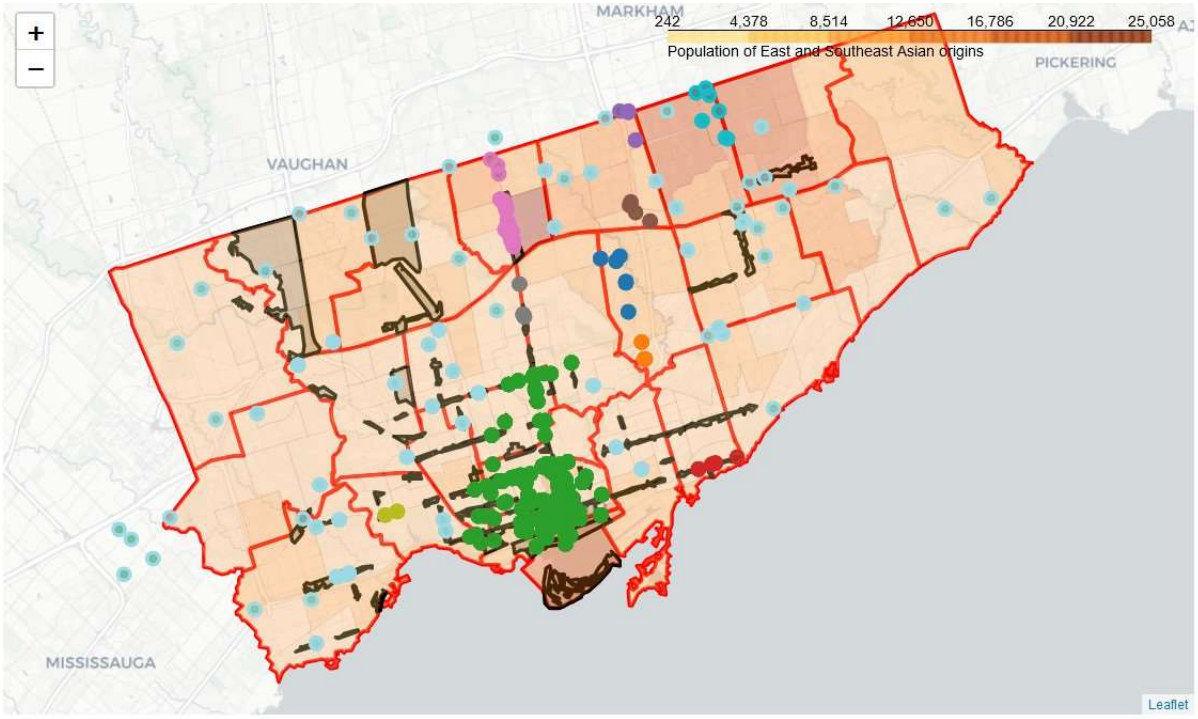
```
In [141]: toronto_japan[['Neighborhood', 'Venue Name', 'Cluster_DB']]
```

Out[141]:

	Neighborhood	Venue Name	Cluster_DB
0	Parkwoods	Matsuda Japanese Cuisine & Teppanyaki	0
1	Parkwoods	Gonoe Sushi	0
2	Parkwoods	Sushi Ichiban	-1
3	Parkwoods	Teriyaki Experience	0
4	Parkwoods	Katsura Japanese Restaurant 桂	0
5	Parkwoods	Gyu-Kaku	5
7	Victoria Village	Tatami Sushi	-1
8	Victoria Village	Kaiseki Yu-Zen Hashimoto	1
9	Victoria Village	Teriyaki Experience	-1
11	Harbourfront	Kinka Izakaya Original	2
12	Harbourfront	NAMI	2

5.4. DBSCAN Visualization

Finally, I was able to put all of this information's about new clusters into choropleth map of Toronto.



6. Discussion

After performing DBSCAN algorithm I finally have potential candidates for localisation a new Japanese restaurant. Now I have to consider few things:

- localisation of Business Improvement Area
- high or low population density in East and Southeast Asian origins in Toronto
- density of Japanese restaurants in each cluster

Outlier identified with number -1 will not be taking into consideration during following analysis.

Cluster 0: Fair

- Out of BIA
- Moderate population density of East and Southeast Asian origins
- Moderate density of Japanese Restaurants

Cluster 1: Not Recommended

- Out of BIA
- Low population density of East and Southeast Asian origins
- Low density of Japanese Restaurants

Cluster 2: Potential

- In of BIA
- High population density of East and Southeast Asian origins
- High density of Japanese Restaurants

Cluster 3: Fair

- In of BIA
- Low population density of East and Southeast Asian origins
- Moderate density of Japanese Restaurants

Cluster 4: Potential

- Out of BIA
- High population density of East and Southeast Asian origins
- Low density of Japanese Restaurants

Cluster 5: Potential

- Out of BIA
- Moderate population density of East and Southeast Asian origins
- Low density of Japanese Restaurants

Cluster 6: Not Recommended

- In of BIA
- High population density of East and Southeast Asian origins
- High density of Japanese Restaurants

Cluster 7: Potential

- In of BIA
- Low population density of East and Southeast Asian origins
- Low density of Japanese Restaurants

Cluster 8: Potential

- In of BIA
- Low population density of East and Southeast Asian origins
- Low density of Japanese Restaurants

Cluster 9: Not Recommended

- Out of BIA
- High population density of East and Southeast Asian origins
- Moderate density of Japanese Restaurants

Cluster 10: Fair

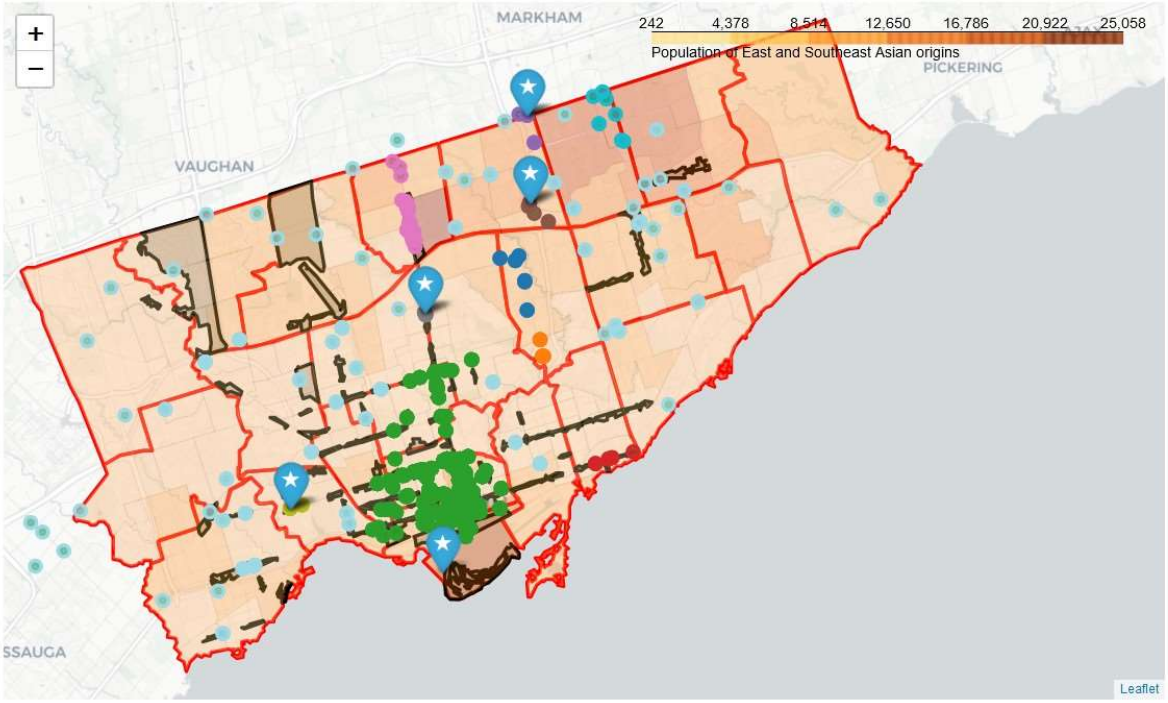
- In of BIA
- Low population density of East and Southeast Asian origins
- Moderate density of Japanese Restaurants

7. Conclusion

I have spotted 5 points with high potential to open new business. Now I give it star label with colour blue.

The Boroughs that should be considered when opening new Japanese restaurants are:

- Scarborough
- North York
- Downtown Toronto outskirts
- Toronto Islands



A very important note is that all of above information conclude my analysis. I found 4 potential areas to open Japanese restaurant. However please notice that there exist thousands of different factors that should be considered when opening a new business like: rent prices, average revenue of restaurants in different areas, potential client's accessibility, sales volume, and many many more.

8. Limitations

One of important limitations worth to mention is a fact that data used in this project was obtained from Open Data city of Toronto and based on Census results from 2016 year.

Immigration and ethnocultural diversity statistics like neighbourhood profiles data, Business Improvement Areas data and GeoJSON ward data are based on results of 2016 Census in Canada.

Current 2021 year are the year for conducting new Census in Canada, so if the data will be available could be used to improve density population map or business improvement area map. New Census data could bring new insight about neighbourhoods' profiles in city of Toronto.

9. Reference

1. https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050
2. https://open.toronto.ca/dataset/business-improvement-areas/
3. https://open.toronto.ca/dataset/neighbourhood-profiles/
4. https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/ward-profiles/44-ward-model/
5. https://open.toronto.ca/dataset/neighbourhoods/
6. https://developer.foursquare.com/
7. https://towardsdatascience.com/machine-learning-clustering-dbscan-determine-the-optimal-value-for-epsilon-eps-python-example-3100091cfbc
8. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
9. https://www.britannica.com/place/Toronto
10. https://www.google.com/maps/d/viewer?mid=1GVzAjujA652hcRVy7TdHpWwDM3Y&vpsrc=0&ctz=300&ie=UTF8&t=h&oe=UTF8&msa=0&ll=43.724315723883095%2C-79.32511256096308&z=11