IBM Data Science Capstone Project

OPENING A VIETNAMESE RESTAURANT IN TORONTO, CANADA

Table of contents

1. Introduction

1.1. Background

Toronto is a diversified city which welcomes different cultures in the world, comes with it their cuisine. This makes great opportunities for restaurant business owners to blend in by bringing their great food to offer. However, this advantage could become a waste if one does not have a clear strategy, especially, of where the best location to open a restaurant should be. This is where data science comes in with great help to business owners to make the best data-driven decisions.

1.2. Topic description

On the mentioned background, I chose Vietnamese cuisine as topic for the research of the optimal neighborhood(s) to open a restaurant in Toronto, with forever love for the food, the people and the culture of Vietnam, and for the diversified beauty of Toronto.

1.3. Target audience

The main target audience of this analysis is business owners who are seeking to open a new Vietnamese restaurant in Toronto, Canada.

Besides, this report could also be useful for investors targeting to invest in Food and Beverage sector in Toronto.

1.4. How is the analysis useful for the target audience?

For business owners – the main target audience, this analysis aims to provide them a general picture on how all types of venues are distributed across all the neighborhoods in Toronto, which would be the first factor to decide on where it could be potential locations to open a new restaurants. Next, more detailed comparison on how the current restaurants are spread through the city, across the neighborhoods and by cuisine, especially Asian and Vietnamese would give the target owners a clearer view on where there direct (other Vietnamese restaurants) and indirect (other Asian restaurants) competitors are situated. With all taken into account, choosing a neighborhood where there are a high number of restaurants with high customer traffic, but not with so many Asian or Vietnamese restaurants to mitigate competition could be the optimal decision.

For investors in F&B sectors, the first part of the analysis on venue and restaurant distribution would be greatly useful as they would have the whole picture on where each types of restaurants by cuisine are located, and also which locations would be potential for their future business investments.

2. Data Collection and pre-processing

To maintain coherence and simplicity of the Capstone project as a whole, I used Toronto geographical data on borough and neighborhood from Week 3's assignment. Data on venues in the city were extracted with Foursquare API.

2.1. Data collection and summary

2.1.1. Importing and installing libraries

- Library to handle data in a vectorized manner: numpy
- Library for data analysis: pandas
- Library to handle JSON files: json
- Convert an address into latitude and longitude values: geopy
- Matplotlib and associated plotting modules
- K-means from clustering stage
- Library for map rendering: folium
- Library to display html: display_html
- Library to scrape data from Wikipedia page: BeautifulSoup

2.1.2. Geographic data on neighborhoods in Toronto

Using BeautifulSoup to scrape data from Wikipedia page [List of Postal Code of Canada](), the first few lines of the html table is displayed as below (through display_html)

```
<title>List of postal codes of Canada: M - Wikipedia</title>
```

Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	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	Not assigned

Next is to get the latitude and longitude coordinates of each neighborhood by importing the csv file containing the latitudes and longitudes of neighborhoods in Canada from [here](). The first lines of the geographical coordination displayed as below:

	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

The raw data will be further processed to take into account only boroughs and neighborhoods in Toronto.

2.1.3. Data on venues using Foursquare API

Foursquare Credentials and Version were first defined with my Client ID and Client Secret, with a limit of 100, default Foursquare API limit value.

Then, using a function to get the top 100 venues within a radius of 500 meters for all the neighborhoods in Toronto and list down the list of the extracted venues together with their category, latitude and longitude:

```
print(toronto_venues.shape)
toronto_venues.head()
```

```
(1602, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant

The raw data will be further processed to grouped by neighborhood for clustering, and also filter to analyze the distribution of restaurants by location and by cuisine across the city in the following parts.

2.2. Data pre-processing

2.2.1. Geographic data on neighborhoods in Toronto

Only the cells that have an assigned borough are to be processed, cells with 'Not Assign' borough are to be ignored.

More than one neighborhood can exist in one postal code area. These rows will be combined into one row with the neighborhoods separated with a comma.

If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

The data frame on Canada borough and neighborhood now has 3 columns and 103 rows.

	Postal Code	Borough	Neighbourhood
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
...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North
99	M4Y	Downtown Toronto	Church and Wellesley
100	M7Y	East Toronto	Business reply mail Processing Centre, South C...
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...

103 rows × 3 columns

Merging this table with the latitude and longitude table, taking into account only boroughs that contain the word Toronto to have a final clean up geographical dataframe of Toronto with Borough, and geographical coordination, with 100 rows sampled as below:

```
df_tor = df_latlon[df_latlon['Borough'].str.contains('Toronto',regex=False)]
df_tor
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
19	M4E	East Toronto	The Beaches	43.676357	-79.293031
20	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306
24	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
25	M6G	Downtown Toronto	Christie	43.669542	-79.422564
30	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568
31	M6H	West Toronto	Dufferin, Dovercourt Village	43.669005	-79.442259

2.2.2. Data on venues from Foursquare

First, the data was to be grouped by neighborhood, there are 230 unique venue categories extracted. Then, each neighborhood is to be analyzed using onehot coding and regrouped again by Neighborhood, sampled as below:

	Neighborhood	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	...
0	Berczy Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.017241	...
1	Brockton, Parkdale Village, Exhibition Place	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	...
2	Business reply mail Processing Centre, South C...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	...
3	CN Tower, King and Spadina, Railway Lands, Har...	0.066667	0.066667	0.133333	0.133333	0.133333	0.000000	0.000000	0.00	0.000000	...
4	Central Bay Street	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	...
5	Christie	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	...

10 top venues for each neighborhoods are to be displayed by frequency of each venues, sorted by descending order. This final clean-up data frame sampled as below will be used to cluster the neighborhoods by venue categories, for the objective of segmenting and selecting the most suitable cluster for opening a new restaurant.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berczy Park	Coffee Shop	Bakery	Cocktail Bar	Farmers Market	Seafood Restaurant	Restaurant	Pharmacy	Cheese Shop	Beer Bar	Café
1	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Nightclub	Coffee Shop	Yoga Studio	Performing Arts Venue	Burrito Place	Restaurant	Climbing Gym	Convenience Store
2	Business reply mail Processing Centre, South C...	Yoga Studio	Smoke Shop	Auto Workshop	Brewery	Burrito Place	Butcher	Comic Shop	Farmers Market	Fast Food Restaurant	Garden
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Airport Service	Airport Terminal	Airport	Bar	Coffee Shop	Rental Car Location	Sculpture Garden	Boat or Ferry	Boutique

3. **Exploratory Data Analysis**

3.1. Visualization of the neighborhoods in Toronto

This is to map out the geographical visualization of the neighborhoods in Toronto for further in-depth analysis, more specifically the clustering of neighborhoods by venue category using k-means clustering in the next part.

3.2. Filtering venue data to focus on restaurants, Asian restaurants and Vietnamese restaurants

From the list of all venues displayed, I now zoom in to the venues classified as 'RestFlag' to be the list of all restaurants, in other words, venues with category containing the words 'Restaurant', 'Snack Place', 'Food Court', etc.

From this set of all the restaurants, I now separated it to subsets of Asian Restaurants, which are venues that contain words as 'Asian Restaurant', 'Chinese Restaurant', 'Japanese Restaurant', 'Vietnamese Restaurant' etc., and subsets of all Vietnamese Restaurant only.

```
# numbers of restaurants by cuisine
viet_restaurants = toronto_restaurants[ toronto_restaurants['Venue Category'].isin(viet_restaurant_list) ]
asian_restaurants = toronto_restaurants[ toronto_restaurants['Venue Category'].isin(asian_restaurant_list) ]

print('Total number of restaurants:', len(toronto_restaurants['Venue'].unique()))
print('Total number of Asian restaurants:', len(asian_restaurants['Venue'].unique()))
print('Total number of Vietnamese restaurants:', len(viet_restaurants['Venue'].unique()))
```

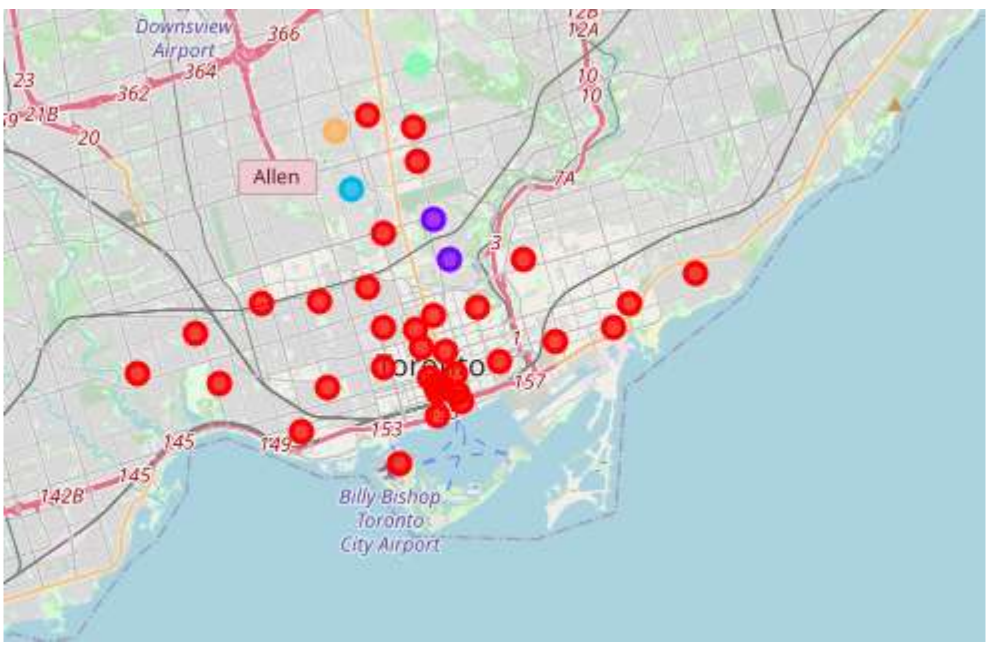
```
Total number of restaurants: 348
Total number of Asian restaurants: 83
Total number of Vietnamese restaurants: 6
```

Counting each mentioned cuisine, there are in total 348 venues classified as restaurants, in which, 83 (24%) are Asian restaurants, and 6 (1.7%) are Vietnamese restaurants. There are 35 neighborhoods that do not currently have any Vietnamese restaurants. From this observation we can later visualize the distribution restaurants by cuisine in each neighborhood across the city of Toronto.

4. Results - In-depth Data Analysis

4.1. Clustering the neighborhoods in Toronto by venues using k-means clustering

Using k-means method, the venues by neighborhood in Toronto are clustered into 5 groups based on Venue Category. The result is visualized on map as below:



At first glance, most of the venues are clustered in one single big cluster (called Cluster 1) by types of venues, the rest of the clusters contain outliers with very limited number of components.

For clearer observation, next, the clusters are listed down by neighborhood and the most common venues of each:

Cluster 1 - Red: The biggest cluster (cluster 1) sampled as following, as expected, this cluster contains restaurants, cafes and eateries. This is the type of venue that we want to focus on in this project. The cluster shows that restaurants, cafes and bar are distributed evenly and widespread across all the neighborhoods of Toronto:

```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 0, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Regent Park, Harbourfront	0	Coffee Shop	Park	Bakery	Breakfast Spot	Café	Pub	Theater	French Restaurant	Greek Restaurant	Wine Shop
4	Queen's Park, Ontario Provincial Government	0	Coffee Shop	Sushi Restaurant	College Cafeteria	Diner	Fried Chicken Joint	Sandwich Place	Burrito Place	Café	Smoothie Shop	Japanese Restaurant
9	Garden District, Ryerson	0	Clothing Store	Coffee Shop	Japanese Restaurant	Middle Eastern Restaurant	Café	Bubble Tea Shop	Cosmetics Shop	Italian Restaurant	Hotel	Bookstore
15	St. James Town	0	Coffee Shop	Café	Gastropub	American Restaurant	Cocktail Bar	Gym	Italian Restaurant	Restaurant	Farmers Market	Clothing Store
19	The Beaches	0	Asian Restaurant	Health Food Store	Trail	Pub	Yoga Studio	Dumpling Restaurant	Dog Run	Doner Restaurant	Donut Shop	Electronics Store
20	Berczy Park	0	Coffee Shop	Bakery	Cocktail Bar	Farmers Market	Seafood Restaurant	Restaurant	Pharmacy	Cheese Shop	Beer Bar	Café
24	Central Bay Street	0	Coffee Shop	Sandwich Place	Café	Italian Restaurant	Thai Restaurant	Japanese Restaurant	Burger Joint	Bubble Tea Shop	Salad Place	Portuguese Restaurant

Cluster 2: This cluster has only 2 outlier members and the most common venues are Park, Playground and Trail, listed as following:

```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 1, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
83	Moore Park, Summerhill East	1	Restaurant	Park	Trail	Dessert Shop	Event Space	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant
91	Rosedale	1	Park	Playground	Trail	Yoga Studio	Diner	Event Space	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant

Cluster 3: Another cluster comprising outliers of venues, with most common venues of Jewelry Store and Trail, listed below:

```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 2, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
68	Forest Hill North & West, Forest Hill Road Park	2	Jewelry Store	Trail	Mexican Restaurant	Sushi Restaurant	Yoga Studio	Discount Store	Event Space	Ethiopian Restaurant	Escape Room	Electronics Store

Cluster 4: Cluster with only 1 member of outliers, with most common venues of Park and Bus Line

```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 3, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
61	Lawrence Park	3	Park	Bus Line	Swim School	Yoga Studio	Discount Store	Event Space	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant

Cluster 5: Home Service, Garden.

```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 4, toronto_merged.columns[[2] + list(range(5, toronto_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
62	Roselawn	4	Health & Beauty Service	Home Service	Garden	Yoga Studio	Discount Store	Falafel Restaurant	Event Space	Ethiopian Restaurant	Escape Room	Electronics Store

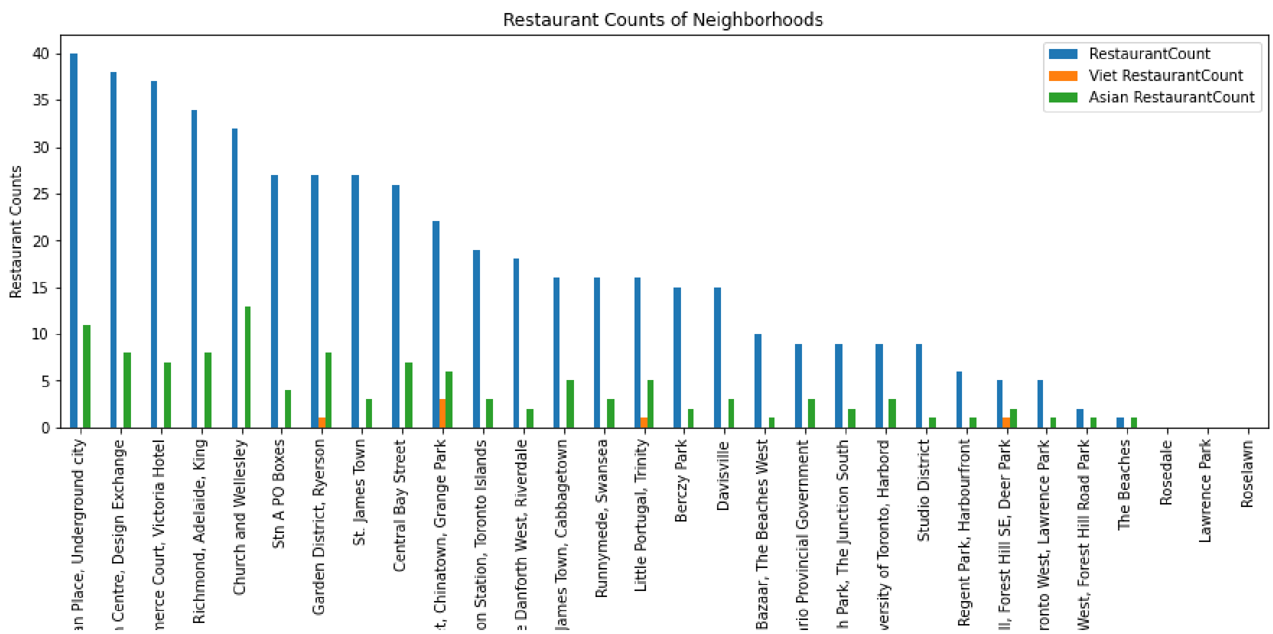
From the results of k-means clustering, it is obvious to conclude that the potential locations for a new Vietnamese restaurant would be in Cluster 1 (Red). However, this cluster includes almost all the neighborhood of the city, and this is too vague to make a decision, we need a more zoomed in result and more specific factor for the best decision. Therefore, a more detailed of how all the restaurants are distributed across all neighborhoods, and by cuisine, especially Asian and Vietnamese cuisine, presented in the next section would be crucial.

4.2. Segmentation and visualization of the restaurants distribution in each neighborhood by cuisines

As mentioned in 4.2., a more granular picture of the distribution of all the restaurants by cuisine, focusing on Asian and Vietnamese restaurants would be the most appropriate reasoning to decide on the location of a new Vietnamese restaurant in Toronto.

To get straight to the point, a visualization of bar chart was used in this case to illustrate this distribution. Using restaurant count as measure, the bar chart breakdown the number of restaurants by neighborhood through bar length, added another level of granular by color for Asian and Vietnamese restaurants.

The total number of restaurants was also sorted by descending order to facilitate analysis, because this will give us a picture of which locations being centers of restaurants, with high customer traffic.



5. Discussion

The clustering of venues by neighborhood in part 4.1. shows that restaurants and cafes are evenly spread throughout all the neighborhoods of Toronto, except The Beaches, Roselawn and the Parks, makes it not enough to just use this clustering to conclude which neighborhood is the most potential to open a new Vietnamese restaurant.

Therefore, part 4.2 is necessary to segregate and visualize the distribution of all restaurants in each neighborhood by cuisine, and to have a clearer picture on where the indirect competitors (other Asian restaurants) and the direct competitors (other Vietnamese restaurants) are situated.

6. Conclusion

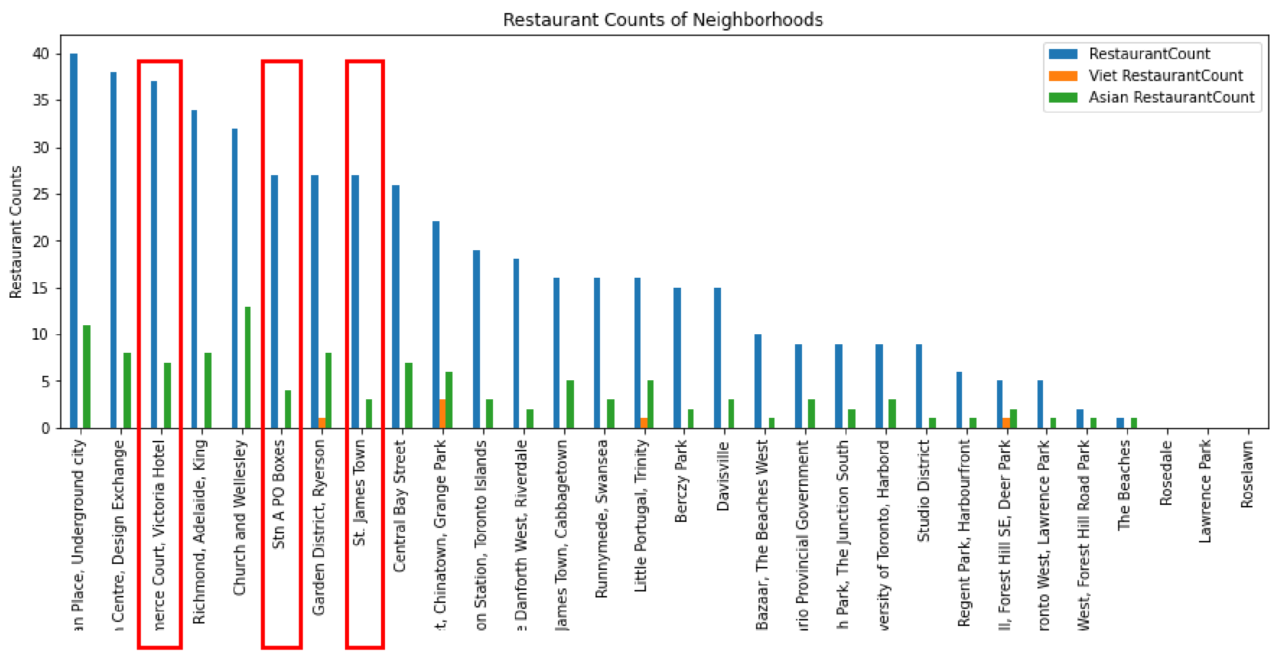
High traffic, low competition

The suitable neighborhoods for this option, based on this analysis, is neighborhoods belonging to cluster 1 of part 3., and neighborhoods that has high number of restaurants in general but not so many Asian/Vietnamese restaurants in part 4.

The neighborhoods with a high number of restaurants are potential to have high diner traffic with multiple choices. Besides, if the said neighborhoods do not have many existing Asian/Vietnamese restaurants, the possibility for a Vietnamese restaurant to be chosen would be higher.

Therefore, the most potential neighborhoods in this case could be:

- Stn A PO Boxes
- St. James Town
- Commerce Court, Victoria Hotel



Restaurant Counts of Neighborhoods
Restaurant Counts
RestaurantCount
Viet RestaurantCount
Asian RestaurantCount
0
5
10
15
20
25
30
35
40
an Place, Underground city
ı Centre, Design Exchange
nerce Court, Victoria Hotel
Richmond, Adelaide, King
Church and Wellesley
Stn A PO Boxes
Garden District, Ryerson
St. James Town
Central Bay Street
t, Chinatown, Grange Park
on Station, Toronto Islands
e Danforth West, Riverdale
James Town, Cabbagetown
Runnymede, Swansea
Little Portugal, Trinity
Berczy Park
Davisville
Bazaar, The Beaches West
ırio Provincial Government
h Park, The Junction South
versity of Toronto, Harbord
Studio District
Regent Park, Harbourfront
ll, Forest Hill SE, Deer Park
ronto West, Lawrence Park
West, Forest Hill Road Park
The Beaches
Rosedale
Lawrence Park
Roselawn