IBM Applied Data Science Capstone Report

EVALUATION OF LOCATIONS FOR A NEW HIGH QUALITY JAPANESE RESTAURANT IN TORONTO BASED ON DEMOGRAPHICS, AND LOCATIONS

Hendrika Kuffar

Table of Contents

Introduction

The goal of the project is to create location shortlist for a new high quality Japanese restaurant in the City of Toronto, Ontario, Canada.

Business Problem

Toronto is a home to some of the best restaurant around the world. However, several sources [1][2][3] have indicated there is a demand for high quality Japanese restaurant that serves delicious meals. The plan is to have a restaurant that will provide excellent customer service and serve lunch and supper meals.

The restaurant industry in Toronto is very competitive, and the strategy is to find locations with high foot and car traffic, affluent clientele, and low competition. Therefore, the criteria for suitable locations are:

1. Median family income above $70,000
2. Population density above 2000 per square kilometer
3. 20 or more business on a list of hundred most common venues

4. High population per restaurant

Targeted customers: affluent clientele (family income $70000+) in most densely populated neighborhoods as well as corporate

Audience & Stakeholders: restaurateurs/investors, restaurant critics, other businesses and people living in the neighborhood and their surrounding neighborhoods

Data

This project extracts insight from data to create location shortlist for a new high quality Japanese restaurant in Toronto. The project will utilize data for Toronto locations and neighborhood demographics. The average income per households after tax for each neighbourhood is used on preliminary analysis, because the median income values for individual neighbourhoods are not available on Toronto neighbourhood profiles csv files for 2016 census.

Toronto demographics

The demographic data from [Toronto neighbourhood profiles csv files for 2016 census](). The data is cleaned and transformed before it's combined with other data to find a suitable location for a new restaurant.

	_id	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	Banbury-Don Mills	Bathurst Manor	Bay Street Corridor	Bayview Village	Bayview Woods-Steeles	Bedford Park-Nortown	Beechborough-Greenbrook	Bendale	Birchcliff-Cliffsic
0	1	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Number	NaN	129	128	20	95	42	34	76	52	49	39	112	127	12
1	2	Neighbourhood Information	Neighbourhood Information	City of Toronto	TSNS2020 Designation	NaN	No Designation	No Designation	No Designation	No Designation	No Designation	No Designation	No Designation	No Designation	No Designation	No Designation	NIA	No Designation	N Designatic
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	25,797	21,396	13,154	23,236	6,577	29,960	22,29
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	19,348	17,671	13,530	23,185	6,488	27,876	21,85
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%	33.30%	21.10%	-2.80%	0.20%	1.40%	7.50%	2.00

Other data set which are used for this project have neighbourhoods as rows. Thus, the data is transposed. The information related to average family income after tax, land area, and population for each neighbourhood are sliced from the data.

Neighbourhood	Population	Density	Area square km	Average Income
Agincourt North	29,113	3,929	7.41	427,037
Agincourt South-Malvern West	23,757	3,034	7.83	278,390
Alderwood	12,054	2,435	4.95	168,602
Annex	30,526	10,863	2.81	792,507
Banbury-Don Mills	27,695	2,775	9.98	493,486

Toronto postal code, borough and Neighbourhood

The sample of Toronto postal code data scraped from Wikipedia table using beautiful soup and transformed to data frame is shown below. The data consist of combination of official and unofficial neighbourhood name.

	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M5A	Downtown Toronto	Regent Park
6	M6A	North York	Lawrence Heights
7	M6A	North York	Lawrence Manor
8	M7A	Queen's Park	Not assigned
9	M8A	Not assigned	Not assigned
10	M9A	Etobicoke	Islington Avenue
11	M1B	Scarborough	Rouge
12	M1B	Scarborough	Malvern
13	M2B	Not assigned	Not assigned
14	M3B	North York	Don Mills North

The initial cleaning of Toronto postal code data was done to remove 'Not Assigned' Borough and renaming 'Not Assigned' Neighbourhood.

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

The city of Toronto has 140 designated neighborhoods which are on census data. However, the remaining number of neighborhoods after cleaning up the Wikipedia data is 211. The problem is most neighborhood names from this dataset do not match the Toronto designated neighborhood names. For example, only two neighborhoods from postal code dataset out of five are exactly a match on sorted data.

	Postcode	Borough	Neighbourhood
0	M5H	Downtown Toronto	Adelaide
1	M1S	Scarborough	Agincourt
2	M1V	Scarborough	Agincourt North
3	M9V	Etobicoke	Albion Gardens
4	M8W	Etobicoke	Alderwood

The cleaned postal code data was saved to object storage on the IBM cloud as a CSV file. The file was updated using information from the Wikipedia list of designated Toronto neighborhoods. Part of the update was to add the missing designated neighborhoods. It was also to replace old names with designated names. For example, "The Annex" was replaced with Annex.

	Postcode	Borough	Neighbourhood
0	M1V	Scarborough	Agincourt North
1	M1S	Scarborough	Agincourt South-Malvern West
2	M8W	Etobicoke	Alderwood
3	M5R	Central Toronto	Annex
4	M3B	North York	Banbury-Don Mills

The change made easy to merge the demographic data with the postal code ones based on 'Neighbourhoods' without losing a lot of important information prior to analysis. Below is the sample of data after merging postal code and demographic data for the city of Toronto.

Postcode	Borough	Neighbourhood	Population	Density	Area square km	Average Income
M1V	Scarborough	Agincourt North	29113	3929	7.41	99071.293163
M1S	Scarborough	Agincourt South-Malvern West	23757	3034	7.83	64585.638490
M8W	Etobicoke	Alderwood	12054	2435	4.95	39115.154354
M5R	Central Toronto	Annex	30526	10863	2.81	183859.228430
M3B	North York	Banbury-Don Mills	27695	2775	9.98	114487.260303

Geolocation postal code longitude and latitude

The list of postal code geographical locations in Toronto is used. The location latitude and longitude values are needed in order to obtain information about the venue from foursquare.

	Postcode	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

Foursquare location venues information

The Foursquare API was used to retrieve the list of most popular top 100 venues for each of the selected popular neighborhoods location in Toronto within a radius of 1 km. The information is important in understanding different categories of venues that exist within the selected radius of the neighbourhood. Below is the sample for foursquare data for venues for Rosedale-Moore Park neighbourhood in Toronto.

```
{'meta': {'code': 200, 'requestId': '5d5dc7ecbf7dde002ce26574'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Rosedale',
  'headerFullLocation': 'Rosedale, Toronto',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 27,
  'suggestedBounds': {'ne': {'lat': 43.68856260900001,
    'lng': -79.36510816548741},
   'sw': {'lat': 43.67056259099985, 'lng': -79.38995063451262}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '4adcb343f964a520e32e21e3',
       'name': 'Summerhill Market',
       'location': {'address': '446 Summerhill Ave',
        'crossStreet': 'btwn. MacLennan Ave. and Glen Rd.',
```


Methodology

The data are acquired, cleaned, scaled, aggregated, merged, filtered, analyzed, and visualized. The demographics and locations data for Toronto neighborhoods are from multiple sources.

1. The sample of Toronto postal code data scraped from Wikipedia table using beautiful soup
2. The demographic data from Toronto neighbourhood profiles CSV files for 2016 census.
3. The list of geographical locations for Toronto
4. Top 100 venues for locations in Toronto within a 1-kilometer radius of from Foursquare API

For this project, the location data for venues are used to create new features. The features and demographics are then used to create locations clusters with a k-means algorithm.

Neighbourhood Demographic and Location

The dataset is explored to check if it requires adjustment. The quick review indicated average family income for the city of Toronto in the dataset is $81,495. However, the estimated average family income from neighborhoods is $ 351,276, and a minimum family income is $ 102,259. The online review from neighborhoods documents for the city of Toronto also indicated the average family income values listed on the CSV file are too high. For Example, the average family income listed by the city of Toronto for Annex is $203,150, and the 2015 average household income after tax listed on the CSV file is $792,507.

```
The Average Income for City of Toronto =  81495.0

Estimated Values of Income, Density, and Population for Toronto
The estimated Density for Neighbourhoods =  4334.175869510028

The estimated Population for Neighbourhoods =  2731571

The estimated Average Income for Neighbourhoods =  351276.1285714286
Minimum Average Income =  102259.0
```

Characteristic	Population	Density	Area square km	Average Income
count	140.000000	140.000000	140.000000	1.400000e+02
mean	19511.221429	6261.135714	4.501714	3.512761e+05
std	10033.589222	4840.359075	4.544665	2.309379e+05
min	6577.000000	1040.000000	0.420000	1.022590e+05
25%	12019.500000	3595.250000	1.852500	1.953375e+05
50%	16749.500000	5071.500000	3.275000	2.915495e+05
75%	23854.500000	7621.250000	5.382500	4.305408e+05
max	65913.000000	44321.000000	36.890000	1.413132e+06

The issue with average family income from the 2016 census in Toronto is complex. Complex in the sense that average family income can be inflated or deflated by a few households with higher or lower income. There is also another issue of bias that could arise if data collected is not representative of the population majority.

It would have been nice to have family median income value for each neighborhood available on the CSV file for Toronto neighborhood profiles. The median value is a good representation of the majority family income. For simplicity, the average income for each neighborhood is divided by 4.31 (351,276/81,495). This factor will not take care of inflated or deflated average family income in some of the neighborhoods.

The density value for the city of Toronto is about 4334, with a calculated mean of 6261 per square kilometer. The mean value for the population density of neighborhoods is not the same as the population density for the city. Mathematically:

$$\frac{\sum x}{\sum y} \neq \frac{1}{n}\sum \frac{x}{y}$$

Where:

x=Population of each neighbourhood
y=Land area per square km of each neighbourhood

The formula below is used to calculate population density for the city of Toronto. The estimation of population density from neighborhood values resulted in 4334 per square kilometer. Similarly, the new aggregated population density value for the unique postal code will be calculated using the same formula.

$$\frac{\sum x}{\sum y}$$

The data are aggregated so that we can have one unique postal code per location. The postal codes are used to obtain geolocation.

Postcode	Neighbourhood	Population	Area square km	Density	Average Income
M1B	Malvern,Rouge	90290	45.74	1973	146431.388091
M1C	Centennial Scarborough,Highland Creek	25856	10.59	2441	47762.957679
M1E	Guildwood,Morningside,West Hill	54764	19.04	2876	59286.750198
M1G	Woburn	53485	12.31	4344	145933.058584
M1J	Eglinton East,Scarborough Village	39500	6.33	6240	61362.968479

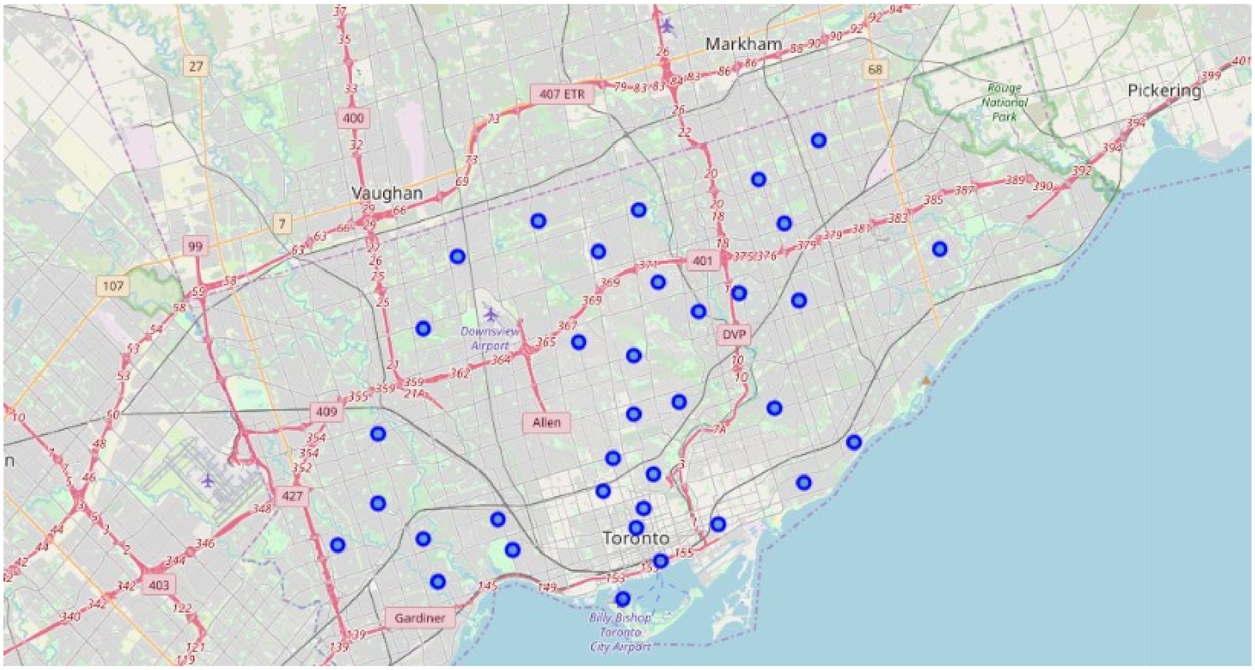
The dataset is then filtered to ensure the locations meet the demographic criteria below:

1. Average Family Income > 70000
2. Population Density > 2000

Postcode	Neighbourhood	Population	Area square km	Density	Average Income
M1G	Woburn	53485	12.31	4344	145933.058584
M1N	Birchcliffe-Cliffside	22291	5.92	3765	85644.236109
M1R	Wexford/Maryvale	27917	10.25	2723	78067.678827
M1T	Tam O'Shanter-Sullivan	27446	5.41	5073	110348.666226
M1V	Agincourt North,Milliken	55685	16.80	3314	110493.664337

Neighbourhood Maps

The visualization of the city of Toronto selected locations on the map.



Neighbourhoods Venues

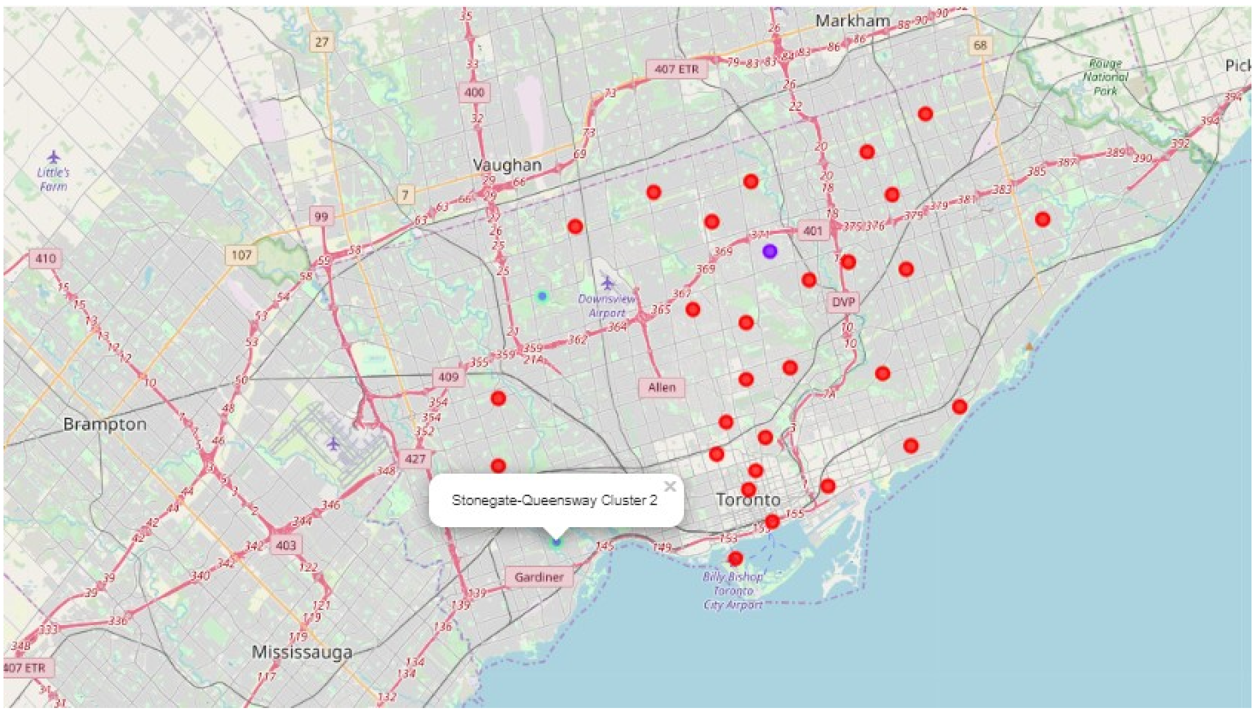
The venues information is crucial in determining the locations which are suitable to open a new restaurant. Foursquare API is used to get the top 100 venues around 1 km radius. The table below shows some of the locations have less than a hundred venues. For example, Agincourt, Milliken location has 31 venues.

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Agincourt North,Milliken	31	31	31	31	31	31
Annex,Casa Loma	100	100	100	100	100	100
Banbury-Don Mills	32	32	32	32	32	32
Bay Street Corridor	100	100	100	100	100	100
Bayview Village,Bayview Woods-Steeles	14	14	14	14	14	14
Bedford Park-Nortown,Lawrence Park North	41	41	41	41	41	41
Birchcliffe-Cliffside	15	15	15	15	15	15
Church-Yonge Corridor	100	100	100	100	100	100
East End-Danforth,The Beaches	80	80	80	80	80	80
Edenbridge-Humber Valley	13	13	13	13	13	13
Eringate-Centennial-West Deane,Islington-City Centre West,Princess-Rosethorn	15	15	15	15	15	15
Forest Hill South,Yonge-St.Clair	78	78	78	78	78	78
Glenfield-Jane Heights	8	8	8	8	8	8
High Park North,High Park-Swansea,Junction Area	100	100	100	100	100	100
Kingsway South	44	44	44	44	44	44
L'Amoreaux,Steeles	24	24	24	24	24	24
Lansing-Westgate,Willowdale East	100	100	100	100	100	100

Neighbourhoods Clusters

The k-means unsupervised machine learning algorithm is used to cluster the locations in two different forms. First to create **3 clusters based on similarity of venues** out for the selected locations. Then to **create 9 clusters of locations which have similar demographic characteristics and aggregated venues features**. The second set of clusters involved the creation and analysis of the new features.

Toronto map which has 3 clusters based similarity of venues correspond to each location are display below. Where cluster 0 is display in red, cluster 1 in purple and cluster 2 in cyan.

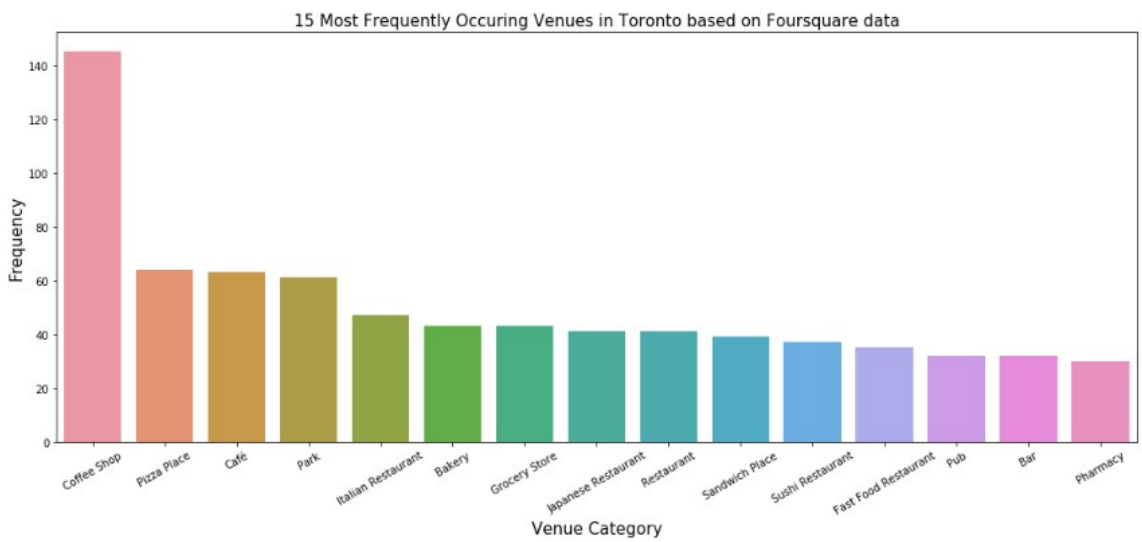


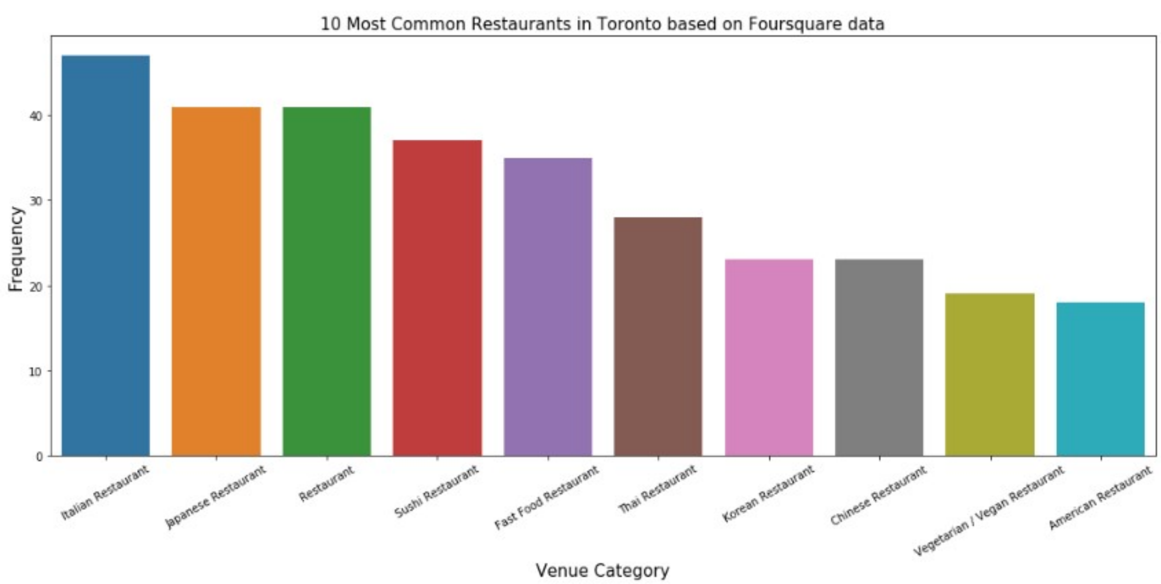
Cluster Labels 0: The **coffee shop/Cafe** is among the most popular venues for this cluster. The other venues categories on most common venues are stores, restaurants, banks, pubs, bars, gyms, bakeries and pizza places.

Cluster Labels 1: It only have one location St.Andrews-Windfield which has 4 venue around the radius of 1 km. The most frequent visited places are the **park** followed by the **pool**.

Cluster Labels 2: The **Park** is among the most popular venues for all locations corresponding to this cluster. There are also banks, bakeries, coffee shops, café, malls, pharmacies, stores and other restaurants in most of these locations.

The top 15 most popular venues and the top 10 most popular restaurants graph are plotted using venues data from foursquare.





The clustering above is based on popularity of the venues and not demographic data. The k-means is used to cluster the locations based on demographic characteristics and aggregated venues features similarity. The number of total venue categories is over 240. From the venue categories, new features are created by aggregating the venues data to represent[1]:

- Number of Businesses
- Number of Restaurants
- Number of Japanese Restaurants
- Population per Restaurant

The demographics data merged with new features

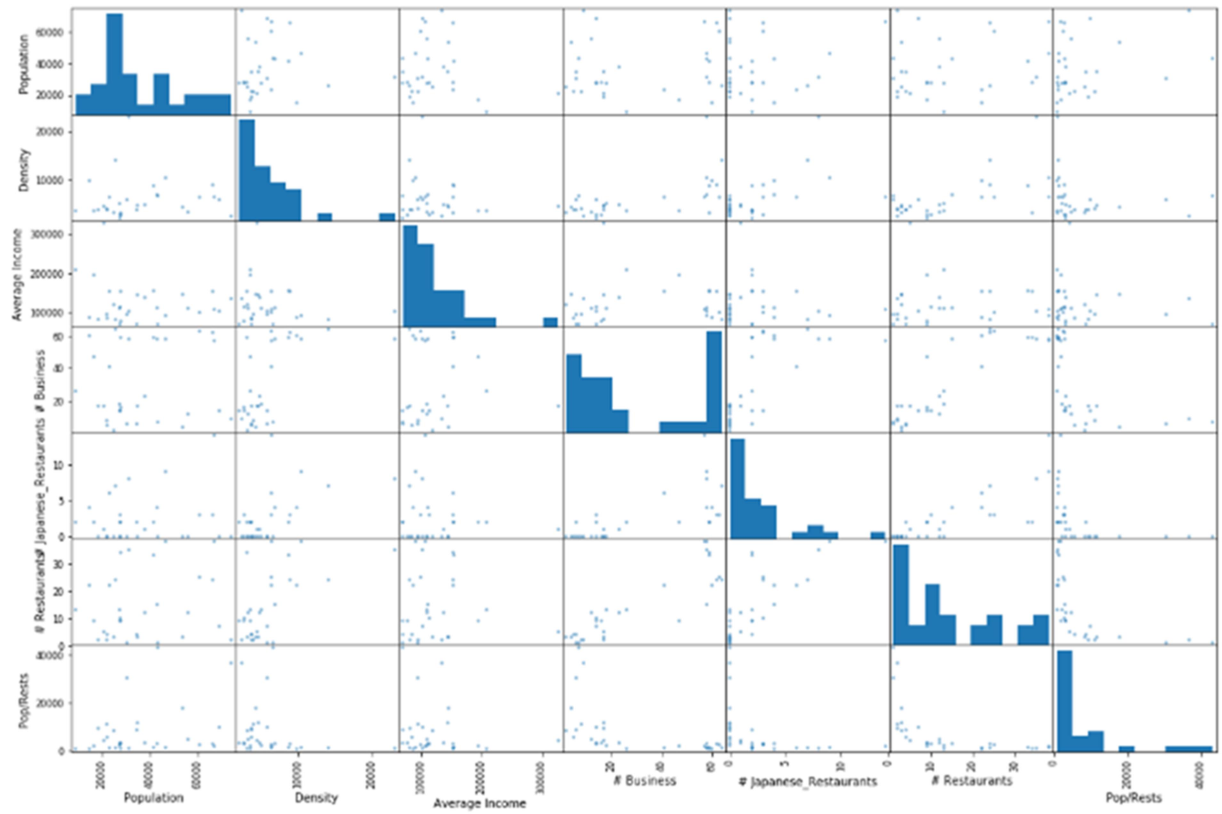
	Population	Area square km	Density	Average Income	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
Population	1.000000	0.638612	0.045264	-0.149964	0.106025	0.154360	0.180663	0.296867
Area square km	0.638612	1.000000	-0.468073	-0.091335	-0.378187	-0.265576	-0.282620	0.436741
Density	0.045264	-0.468073	1.000000	-0.144308	0.602487	0.574797	0.597725	-0.195620
Average Income	-0.149964	-0.091335	-0.144308	1.000000	0.064677	-0.091322	-0.009626	-0.172777
# Business	0.106025	-0.378187	0.602487	0.064677	1.000000	0.696377	0.878492	-0.534484
# Japanese_Restaurants	0.154360	-0.265576	0.574797	-0.091322	0.696377	1.000000	0.791813	-0.409118
# Restaurants	0.180663	-0.282620	0.597725	-0.009626	0.878492	0.791813	1.000000	-0.560650
Pop/Rests	0.296867	0.436741	-0.195620	-0.172777	-0.534484	-0.409118	-0.560650	1.000000

The Pearson correlation is calculated, and the values of correlation are used to determine the features that are not correlated. The correlation matrix below shows that there is **a very strong correlation** between the number of businesses and restaurants (~0.9), and the number of restaurants and Japanese restaurants (~0.8). There is also **a strong correlation** between the number of businesses and Japanese restaurants (~0.7) and that of population density with number of restaurants, number of businesses and Japanese restaurant (~0.6). Moreover, there is **a moderate correlation** between land area and population per restaurant.

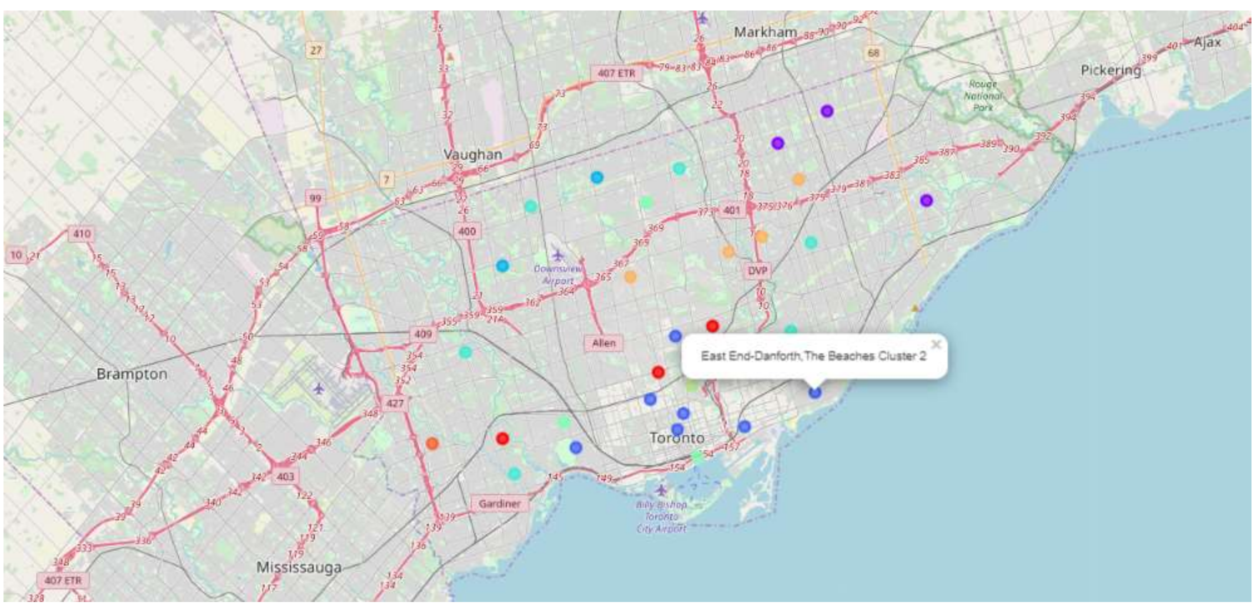
The correlated features are eliminated to improve the result of clustering by k-means. K-Mean clustering is unsupervised machine learning technique. The correlation coefficient threshold of 0.4 is used for this project. Thus the features used for clustering are population, average family income, population per restaurant and the number of businesses. The values of these features are standardized first before the locations data are fitted into nine clusters using k-means.

---

[1]**Keywords for Business Categories:** Bakery, Bank, Bar, Brewery, Building, Business, Breakfast, Taco, Burrito, Butcher, store, Boutique, Coffee, Café, Creperie, Club, Chiropractor, Concert, Deli, Entertainment, Athletics, Service, Pharmacy, Hotel, Hostel, Health, Storage, Soup, Snack, Sandwich, Salon, Tea, Grocery, Store, pub, Historic, Gym, Studio, Rock, Shop, Museum, Plaza, Gallery, Pub, Office, Theater, Mall, Market, Wings, Gas, Rental, Pizza
Keywords for Restaurant Categories: Restaurant, BBQ, Bistro, Steakhouse, Noodle, Diner, Poke
Keywords for Japanese restaurant Categories: Ramen, Sushi, and Japanese



The color-coded markers correspond to different clusters superimposed on the map for Toronto based on their location.



Cluster Labels 0: locations with high income based on 2016 Toronto neighbourhood profile

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
0	Leaside-Bennington	16828.0	4.68	3595.0	47.0	2.0	9.0	1869.777778
0	Forest Hill South,Yonge-St.Clair	23260.0	3.62	6425.0	43.0	6.0	23.0	1011.304348
0	Kingsway South	9271.0	2.58	3593.0	24.0	2.0	12.0	772.583333

Cluster Labels 1: locations with high population per restaurants and fewer businesses

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
1	Woburn	53485.0	12.31	4344.0	3.0	0.0	3.0	17828.333333
1	Agincourt North,Milliken	55685.0	16.80	3314.0	12.0	0.0	12.0	4640.416667
1	L'Amoreaux,Steeles	68616.0	11.69	5869.0	15.0	0.0	7.0	9802.285714

Cluster Labels 2: locations with a lot of businesses

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
2	East End-Danforth,The Beaches	42948.0	6.22	6904.0	58.0	4.0	15.0	2863.200000
2	South Riverdale	27876.0	8.89	3135.0	59.0	2.0	34.0	819.882353
2	Mount Pleasant East,Mount Pleasant West	46433.0	4.45	10434.0	59.0	8.0	37.0	1254.945946
2	Church-Yonge Corridor	31340.0	1.36	23044.0	59.0	8.0	33.0	949.696970
2	Bay Street Corridor	25797.0	1.83	14096.0	64.0	6.0	25.0	1031.880000
2	Annex,Casa Loma	41494.0	4.74	8754.0	59.0	3.0	33.0	1257.393939
2	Roncesvalles	14974.0	1.52	9851.0	60.0	4.0	22.0	680.636364

Cluster Labels 3: locations few business and restaurants, and high population per restaurants.

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
3	Westminster-Branson,Willowdale West	43210.0	6.49	6657.0	9.0	0.0	1.0	43210.0
3	Glenfield-Jane Heights	30491.0	5.20	5863.0	4.0	0.0	1.0	30491.0

Cluster Labels 4: locations with few businesses and restaurants

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
4	Birchcliffe-Cliffside	22291.0	5.92	3765.0	5.0	0.0	4.0	5572.750000
4	Wexford/Maryvale	27917.0	10.25	2723.0	15.0	0.0	9.0	3101.888889
4	Bayview Village,Bayview Woods-Steeles	34550.0	9.16	3771.0	5.0	2.0	4.0	8637.500000
4	York University Heights	27593.0	13.23	2085.0	15.0	1.0	7.0	3941.857143
4	O'Connor-Parkview	18675.0	4.94	3780.0	15.0	0.0	2.0	9337.500000
4	Stonegate-Queensway	25051.0	7.83	3199.0	2.0	0.0	3.0	8350.333333
4	Willowridge-Martingrove-Richview	22156.0	5.53	4006.0	9.0	0.0	2.0	11078.000000

Cluster Labels 5: locations with high population and businesses

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
5	Lansing-Westgate,Willowdale East	66598.0	10.32	6453.0	56.0	14.0	39.0	1707.641026
5	Waterfront Communities-The Island	65913.0	7.37	8943.0	62.0	3.0	24.0	2746.375000
5	High Park North,High Park-Swansea,Junction Area	60453.0	9.42	6417.0	68.0	3.0	22.0	2747.863636

Cluster Labels 6: location with high income and population per restaurants

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
6	Rosedale-Moore Park	20923.0	4.65	4499.0	16.0	1.0	4.0	5230.75

Cluster Labels 7: locations with number of businesses between 15 -21

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
7	Tam O'Shanter-Sullivan	27446.0	5.41	5073.0	15.0	0.0	13.0	2111.230769
7	Parkwoods-Donalda	34805.0	7.42	4690.0	17.0	0.0	3.0	11601.666667
7	Banbury-Don Mills	27695.0	9.98	2775.0	18.0	3.0	9.0	3077.222222
7	Bedford Park-Nortown,Lawrence Park North	37843.0	7.80	4851.0	21.0	2.0	13.0	2911.000000

Cluster Labels 8: locations with high population and fewer businesses and restaurants

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
8	Eringate-Centennial-West Deane,Islington-City ...	73604.0	29.94	2458.0	12.0	0.0	2.0	36802.0

Results and Discussion

The Cluster Labels 2 & 5 contained locations with high foot and car traffic. Some of these locations have high competition and may not be ideal for new restaurants. For example, the table below contains locations with high population density, number of businesses and restaurants.

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
2	Mount Pleasant East,Mount Pleasant West	46433.0	4.45	10434.0	59.0	8.0	37.0	1254.945946
2	Annex,Casa Loma	41494.0	4.74	8754.0	59.0	3.0	33.0	1257.393939
2	Church-Yonge Corridor	31340.0	1.36	23044.0	59.0	8.0	33.0	949.696970
2	South Riverdale	27876.0	8.89	3135.0	59.0	2.0	34.0	819.882353
5	Lansing-Westgate,Willowdale East	66598.0	10.32	6453.0	56.0	14.0	39.0	1707.641026

The population per restaurant, the total number of other businesses, family income, population density, and population are used to select a suitable location. The location shortlists for a new Japanese restaurant that needs to undergo further analysis:

1. **Bedford Park-Nortown and Lawrence Park North** are among the neighborhoods with high family medium income and population per restaurant.
2. **High Park North, High Park-Swansea, and Junction Area** have a combined population of over sixty thousand, high number of established businesses and fewer restaurants relative to other locations.
3. The venue data from foursquare show **East End-Danforth and The Beaches** have high number of businesses and population per restaurant.
4. **Waterfront Communities-The Island** has a lot of businesses and a population of over sixty-five thousand

Cluster Labels	Neighbourhood	Population	Area square km	Density	# Business	# Japanese_Restaurants	# Restaurants	Pop/Rests
2	East End-Danforth,The Beaches	42948.0	6.22	6904.0	58.0	4.0	15.0	2863.200000
5	Waterfront Communities-The Island	65913.0	7.37	8943.0	62.0	3.0	24.0	2746.375000
5	High Park North,High Park-Swansea,Junction Area	60453.0	9.42	6417.0	68.0	3.0	22.0	2747.863636
7	Bedford Park-Nortown,Lawrence Park North	37843.0	7.80	4851.0	21.0	2.0	13.0	2911.000000

Neighborhood	Median Income	Density	Population	Businesses	Restaurants	Japanese Restaurants	Population/ Restaurants
Bedford Park-Nortown, Lawrence Park North	$184,168	4,851	37,843	21	15	1	2911
High Park North,High Park-Swansea,Junction Area	$107,012	6,417	60,453	68	22	3	2748
East End-Danforth,The Beaches	$127,704	6,904	42,948	58	15	4	2863
Waterfront Communities-The Island	$108,199	8,943	65,913	62	24	3	2746

Note: The family median income values were extracted from neighborhood documents which are available through Toronto wellbeing map. The locations which had more than one neighbourhood used mean value of individual neighbourhood median family income.

Conclusion and Recommendations

The purpose of this project is to create a location shortlist for a new restaurant based on location and demographics data. The process involved data acquisition from multiple sources, cleanse, transformation, analyzation, and visualization. The result is a suitable location shortlist for a new restaurant. The selected locations have family median income above $70,000, are densely populated, and have high foot and car traffic.

Additional analysis is required to select the location for a new Japanese restaurant. The analysis should factor in renting or buying cost, labor cost, and other related costs. It should also take into consideration among other things information related to the availability of parking space, accessibility of the venue, and crime rate.

References

[1] Jason Chow, 2018. Toronto's food halls point to a grown-up city with a demand for quality, https://www.theglobeandmail.com/life/food-and-wine/restaurant-reviews/article-torontos-food-halls-point-to-a-grown-up-city-with-a-demand-for/ Last Accessed August 10, 2019

[2] Amy Carlberg, 2018. Toronto ramen joint can't keep overwhelming demand. https://www.blogto.com/eat_drink/2018/09/toronto-ramen-joint-cant-keep-overwhelming-demand/, Last Accessed August 15, 2019

[3] Walter Muschenheim, 2018. A Toronto Tradition Returns, http://bentoboxmag.ca/2018/11/18/a-toronto-tradition-returns/, Last Accessed August 10, 2019

[4] Tom Larkin, 201 8. Factors for Choosing a New Restaurant Location, https://www.foodnewsfeed.com/fsr/expert-insights/8-factors-choosing-new-restaurant-location, Last Accessed August 17, 2019

[5] http://map.toronto.ca/wellbeing, by City of Toronto

[6] https://thepostalcodelookup.com/