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

There is a demand for a high quality authentic Japanese restaurant that will serve delicious mouth-watering meals in a city of Toronto, Canada. Investors/Restauranteurs are interested in a location that has high foot and car traffic with low competition and affluent clientele. This will help to maximize the profit on investment. The plan is to have a restaurant that will provide excellent customer service and serve high quality Japanese lunch and supper meals. The restaurant is expected to have about fifty sittings.

Targeted customers: affluent clientele (households income after tax $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 will extract insight from data to create location shortlist for a new high quality authentic Japanese restaurant in Toronto. The popularity of the venue, competition from similar kind of restaurant as well as the average household income after tax and neighbourhood population density are used as criteria to find a suitable location. Thus, the project will utilize data for locations and neighborhood demographic for Toronto. The average income per households after tax for each neighbourhood is used, 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: https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/ef0239b1-832b-4d0b-a1f3-4153e53b189e?format=csv. The data is cleaned and transformed before it's combined with other data to find a suitable location for a new restaurant. In this case, we are interested in demographics data for Toronto neighborhoods specifically population, average household income after tax, and density.

The original Toronto neighbourhood profiles with neighbourhoods as columns. However, other data have neighbourhoods as rows therefore it make sense to transpose the table in such a way neighbourhoods will be in rows.

```
df_TNDem=pd.read_csv("https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/ef0239b1-832b-4d0b-a1f3-4153e53b189e?format=csv")

df_TNDem.head()
```

5]:

	_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	Birchcliffe-Cliffside
0	1	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Number	NaN	129	128	20	95	42	34	76	52	49	39	112	127	122
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	No Designation
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	22,291
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	21,856
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%

After transpose the original data you can see there is a lot of information we don't need for this project and in addition to all the neighbourhoods we have row for city of Toronto.

```
dftn=df_TNDem.drop(['_id','Category','Topic','Data Source'], axis=1)
dftn.set_index('Characteristic',inplace=True)
df3=dftn.T

df3.head()
```

t[42]:

Characteristic	Neighbourhood Number	TSNS2020 Designation	Population, 2016	Population, 2011	Population Change 2011-2016	Total private dwellings	Private dwellings occupied by usual residents	Population density per square kilometre	Land area in square kilometres	Children (0-14 years)	Youth (15-24 years)	Working Age (25-54 years)	Pre-retirement (55-64 years)	Seniors (65+ years)	Older Seniors (85+ years)	Male: 0 to 04 years	Male: 05 to 09 years	Male: 10 to 14 years	Male: 15 to 19 years	Male: 20 to 24 years	Male: 25 to 29 years
City of Toronto	NaN	NaN	2,731,571	2,615,060	4.50%	1,179,057	1,112,929	4,334	630.2	398,135	340,270	1,229,555	336,670	426,945	66,000	69,895	69,350	64,945	74,240	97,415	113,905
Agincourt North	129	No Designation	29,113	30,279	-3.90%	9,371	9,120	3,929	7.41	3,840	3,705	11,305	4,230	6,045	925	660	695	660	840	1015	1015
Agincourt South-Malvern West	128	No Designation	23,757	21,988	8.00%	8,535	8,136	3,034	7.83	3,075	3,360	9,965	3,265	4,105	555	575	540	460	780	1000	1045
Alderwood	20	No Designation	12,054	11,904	1.30%	4,732	4,616	2,435	4.95	1,760	1,235	5,220	1,825	2,015	320	360	270	225	285	355	355

For this project we are interested in information related to income, land area, and population for each neighbourhood. The extraction of this information was done as follows:

```
df3k=df3[['Population, 2016','Population density per square kilometre','Land area in square kilometres','  Average after-tax income of households in 2015 ($)']]
df3k.reset_index(inplace=True)
df3k.rename(columns={'index':'Neighbourhood','Population, 2016':'Population','Population density per square kilometre':'Density', 'Land area in square kilometres':'Area square km','  Average afte
print(df3.shape)
df3k.head()
```

```
(141, 2383)
```

[0]:

Characteristic	Neighbourhood	Population	Density	Area square km	Average Income
0	City of Toronto	2,731,571	4,334	630.2	81,495
1	Agincourt North	29,113	3,929	7.41	427,037
2	Agincourt South-Malvern West	23,757	3,034	7.83	278,390
3	Alderwood	12,054	2,435	4.95	168,602
4	Annex	30,526	10,863	2.81	792,507

```
dfTorontoCity=df3k[0:1]
df3k=df3k[1:]
df3k.reset_index(inplace=True,drop=True)
print(df3k.shape)
df3k.head()
```

```
(140, 5)
```

[31]:

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

Toronto postal code, borough and Neighbourhood

The list of postal code, neighbourhood and borough will be scrapped from Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M using beautiful soup. The list is using combination of official and unofficial neighbourhood name. The large percentage of neighbourhood name does not match the designated neighbourhood. Thus, the list will be cleaned, saved in a file. The saved file was updated using information from Wikipedia: https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto . The update included adding missing neighbourhood and updates the one in the list to use designated neighbourhood names.

Sample Toronto postal code data scrapped from Wikipedia table using beautiful soup and transformed to data frame:

```
url = requests.get('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M').text
soup = BeautifulSoup(url,'lxml')
TNCode = soup.find('table',{'class':'wikitable sortable'})
```

```
columns=['Postcode','Borough','Neighbourhood']
df = pd.DataFrame(columns=columns)

for row in TNCode.findAll("tr"):
    trow=[]
    for cell in row.findAll("td"):
        trow.append(cell.text.strip())
        if len(trow)==3:
            df.loc[len(df)] = trow
print(df.shape)
df.head(15)
```

```
(288, 3)
```

ut[4]:

	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.

```
df=df[df['Borough']!="Not assigned"]
df.loc[(df['Neighbourhood'] == 'Not assigned'), 'Neighbourhood'] = df['Borough']
df.reset_index(drop=True,inplace=True)

print(df.shape)
df.head(15)
```

```
(211, 3)
```

	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

There are about 211 neighbourhoods listed after initial clean up. However the official list of Toronto contains 140 neighbourhoods. The name for majority of neighbourhood does not match the actual list of official neighbourhoods on neighbourhood profiles and other sources.

For example, only 2 neighbourhoods out of 5 are exactly a match on sorted data by designated neighbourhood from A-Z.

https://en.wikipedia.org/wiki/List_of_city-designated_neighbourhoods_in_Toronto

CDN number	City-designated area	Former city/borough	Neighbourhoods covered
129	Agincourt North	Scarborough	Agincourt and Brimwood
128	Agincourt South-Malvern West	Scarborough	Agincourt and Malvern
020	Alderwood	Etobicoke	Alderwood
095	Annex	Old City of Toronto	The Annex and Seaton Village
042	Banbury-Don Mills	North York	Don Mills

```
df1=df.sort_values('Neighbourhood')
df1.reset_index(drop=True, inplace=True)
df1.head()
```

17]:

	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

```
project.save_data(data=df1.to_csv(index=False),file_name='TorontoPostalCode1.csv',overwrite=True)
```

The scrapped pre-clean data is saved to csv files for further cleanup. After mapping the designated neighbourhoods to postal code the 5 top neighbourhoods sorted in ascending order.

```
dfile = project.get_file("TorontoPostalCode2.csv")

# Read the CSV data file from the object storage into a pandas DataFrame
dfile.seek(0)

dfu=pd.read_csv(dfile)
dfu1=dfu.sort_values('Neighbourhood')
dfu1.reset_index(drop=True,inplace=True)
```

```
dfu1.head()
```

29]:

	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 will make it 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.

```
dfDC1 = pd.merge(dfu1, dfRk, how='inner', on = 'Neighbourhood')
print(dfDC1.shape)
dfDC1
```

```
(132, 7)
```

t[21]:

	Postcode	Borough	Neighbourhood	Population	Density	Area square km	Average Income
0	M1V	Scarborough	Agincourt North	29113	3929	7.41	99071.293163
1	M1S	Scarborough	Agincourt South-Malvern West	23757	3034	7.83	64585.638490
2	M8W	Etobicoke	Alderwood	12054	2435	4.95	39115.154354
3	M5R	Central Toronto	Annex	30526	10863	2.81	183859.228430
4	M3B	North York	Banbury-Don Mills	27695	2775	9.98	114487.260303
5	M3H	North York	Bathurst Manor	15873	3377	4.70	58366.495521
6	M5G	Downtown Toronto	Bay Street Corridor	25797	14097	1.83	81713.511324
7	M2K	North York	Bayview Village	21396	4195	5.10	82334.335235
8	M2K	North York	Bayview Woods-Steeles	13154	3240	4.06	58703.587129
9	M5M	North York	Bedford Park-Nortown	23236	4209	5.52	167084.918989
10	M6M	York	Beechborough-Greenbrook	6577	3614	1.82	25491.827858
11	M1P	Scarborough	Bendale	29960	4011	7.47	75742.837118

Geolocation postal code longitude and latitude

The list of geographical coordinates (longitude and latitude) of each postal code: http://cocl.us/Geospatial_data for the neighborhoods in Toronto. The geolocation data are merged with data for demographic and neighbourhoods. The location latitude and longitude will be used in foursquare to find more information.

```
df_TNCoords=pd.read_csv("http://cocl.us/Geospatial_data")
df_TNCoords.rename(columns={'Postal Code':'Postcode'}, inplace=True)
print(df_TNCoords.shape)
df_TNCoords.head()
```

```
(103, 3)
```

:6]:

	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 will be used to retrieve the list of most popular top 50 venues for each of the selected popular neighborhoods location in Toronto within a radius of 5 km. The information will also help us to identify different categories of venues that exist within the selected radius in the neighbourhood.

The sample for foursquare data for venues correspond to neighbourhoods

```
: dfn.loc[0,'Neighbourhood']
ut[29]: 'Rosedale-Moore Park'

: n_latitude = dfn.loc[0, 'Latitude'] # neighborhood latitude value
  n_longitude = dfn.loc[0, 'Longitude'] # neighborhood longitude value

  n_name = dfn.loc[0, 'Neighbourhood'] # neighborhood name

  print('Latitude and longitude values of {} are {}, {}.'.format(n_name,
                                                                 n_latitude,
                                                                 n_longitude))

    Latitude and longitude values of Rosedale-Moore Park are 43.6795626, -79.37752940000001.

: # type your answer here
  import requests
  latitude=n_latitude
  longitude =n_longitude
  radius=1000
  LIMIT=100
  url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, latitude, longitude, VERSION, radius, LIMIT)
```


```
2]: {'meta': {'code': 200, 'requestId': '5d5c8b9fa6ec98002c11698c'},
     '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': 43.670562590999985, '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 search for suitable location is narrowed down to neighbourhoods which have population density above 2000, average household income after tax above $70,000, and population of greater than 4700 people. We also are looking for popular venue information from foursquare for the selected neighbourhoods. The information will help to determine which locations are ideal to open a new restaurant.

Neighbourhood Demographic and Location

Extracting neighbourhoods which fit the demographic criteria for targeted customer based on 2016 census provided by the city of Toronto. From the data you will notice that the average income per households after tax for the city of Toronto was $81,495. The estimated average income per households after tax for individual neighbourhood after tax is ~ $ 351,276. The minimum average income for the neighbourhoods is about $ 102,259.

```
) dfTNCity
35]:
```

Characteristic	Neighbourhood	Population	Density	Area square km	Average Income
0	City of Toronto	2,731,571	4,334	630.2	81495.0

```python
IncomeTN=df3k['Average Income'].mean()
PopTN=df3k['Population'].sum()
AreaTN=df3k['Area square km'].sum()

dfTorontoCity['Average Income']=dfTorontoCity['Average Income'].str.replace(',', '').astype(float)
dfTorontoCity['Area square km']=dfTorontoCity['Area square km'].str.replace(',', '').astype(float)
dfTNCity=dfTorontoCity

TorontoAvgIncome=dfTNCity['Average Income'].values[0]
TorontoDensity=PopTN/AreaTN

print("The Average Income for City of Toronto = ", TorontoAvgIncome, "\n")
print('Estimated Values of Income, Density, and Population for Toronto')
print("The estimated Density for Neighbourhoods = ", TorontoDensity, "\n")
print("The estimated Population for Neighbourhoods = ", PopTN, "\n")
print("The estimated Average Income for Neighbourhoods = ", IncomeTN)

print("Minimum Average Income = ", df3k['Average Income'].min())
df3k[['Average Income', 'Population','Density','Area square km']].describe()
```

```
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
```

34]:

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

Since the issue of average income per households from 2016 census in Toronto is complex, for simplicity the average income for each neighbourhood is divided by 4.31 as shown below. This factor will not take care of inflated average income per households in Annex and few other neighbourhoods. It would have been nice to have median Income value for each neighbourhood available on csv file for Toronto neighbourhood profiles obtained from CRA. This is because median value is good representation of the overall population household income. The average value on the other hand could be inflated or deflated by few household with higher or lower income.

```python
# the actual average income for Toronto after tax is $81,495 based on 2016 census and calculated average income is 352264.73
# The average income for individual neighbourhood was taken for smaller sample, however for simplicity we going to scale average income by pf

dfRk=df3k

#since it is average income we can divide the Average income value to match the average of City of Toronto
pf=IncomeTN/TorontoAvgIncome
dfRk['Average Income']=dfRk['Average Income']/pf

print('Income Multiplying factor = ' , pf)
dfRk.head()
```

```
Income Multiplying factor =  4.310400988667141
```

36]:

Characteristic	Neighbourhood	Population	Density	Area square km	Average Income
0	Agincourt North	29113	3929	7.41	99071.293163
1	Agincourt South-Malvern West	23757	3034	7.83	64585.638490
2	Alderwood	12054	2435	4.95	39115.154354
3	Annex	30526	10863	2.81	183859.228430
4	Banbury-Don Mills	27695	2775	9.98	114487.260303

The density value for city of Toronto is about 4334 per square kilometers. The individual sum of population for each neighbourhood divide by total land areas for these neighbourhoods will give us the same number. The mean value for population density of neighbourhoods is not going to be the same as the population density for the city. This is because 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

Thus when aggregating the data we will have to recalculate the population density by divide the sum of individual neighbourhood with the sum of land area for all neighbourhoods which share the postal code instead of using the mean value.

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

The data are aggregated so that we can have one unique postal code per location. The postal codes are used to obtain geolocation. The unique postal codes after aggregation are 98.

```
aggregations = {

    'Neighbourhood': lambda x: ','.join(x),
    'Population':'sum',
    'Area square km':'sum',
    'Density':'mean',
    'Average Income':'mean'
}
columns2=['Postcode', 'Borough', 'Neighbourhood','Population','Density','Area square km','Average Income']
dfD2=dfDC[columns2].groupby(['Postcode','Borough']).agg(aggregations).reset_index()

#recalculate Population Density for aggregated data
dfD2['Density']=(dfD2['Population']/dfD2['Area square km']).astype(int)

print(dfD2.shape)
dfD2
```

```
(98, 7)
```

[43]:

	Postcode	Borough	Neighbourhood	Population	Area square km	Density	Average Income
0	M1B	Scarborough	Malvern,Rouge	90290	45.74	1973	146431.388091
1	M1C	Scarborough	Highland Creek	12494	5.20	2402	45615.245662
2	M1C	Scarborough	Centennial Scarborough	13362	5.39	2479	49910.669695
3	M1E	Scarborough	Guildwood,Morningside,West Hill	54764	19.04	2876	59286.750198
4	M1G	Scarborough	Woburn	53485	12.31	4344	145933.058584
5	M1J	Scarborough	Scarborough Village	16724	3.10	5394	43143.781864
6	M1J	Scarborough	Eglinton East	22776	3.23	7051	79582.155095
7	M1K	Scarborough	Ionview,Kennedy Park	30764	5.53	5563	46217.973809

However, we want neighbourhoods with average income per household above $70,000. We also want areas with population above 4700 which is big enough for estimate of about 150 people per day. We are also going to consider neighbourhoods which have population density greater than 2000 per square km. This is done because we want to narrow our search on neighbourhoods that fit criteria for our targeted customer. Extraction of city of Toronto neighbourhoods which fit targeted demographic criteria:

```
#  finding neighbourhoods with income for household >70000
df_M=dfD2[dfD2['Average Income']>70000].reset_index(drop=True)

#finding Neighbourhoods with population above 4700
df_MU1=df_M[df_M['Population']>4700].reset_index(drop=True)

#finding Neighbourhoods with population density above 2000
df_MU2=df_MU1[df_MU1['Density']>2000].reset_index(drop=True)

dfM2=df_MU2.sort_values('Neighbourhood')
dfM2.reset_index(drop=True,inplace=True)
print(dfM2.shape)
dfM2
```

(42, 7)

25]:

	Postcode	Borough	Neighbourhood	Population	Area square km	Density	Average Income
0	M1V	Scarborough	Agincourt North,Milliken	55685	16.80	3314	110493.664337
1	M5R	Central Toronto	Annex,Casa Loma	41494	4.74	8754	155187.997998
2	M3B	North York	Banbury-Don Mills	27695	9.98	2775	114487.260303
3	M5G	Downtown Toronto	Bay Street Corridor	25797	1.83	14096	81713.511324
4	M2K	North York	Bayview Village,Bayview Woods-Steeles	34550	9.16	3771	70518.961182
5	M5M	North York	Bedford Park-Nortown,Lawrence Park North	37843	7.80	4851	138167.075770
6	M1N	Scarborough	Birchcliffe-Cliffside	22291	5.92	3765	85644.236109
7	M4Y	Central Toronto	Church-Yonge Corridor	31340	1.36	23044	102944.946692
8	M1L	Scarborough	Clairlea-Birchmount	26984	7.43	3631	81379.203680
9	M6E	Central Toronto	Corso Italia-Davenport	14133	1.89	7477	70527.081077
10	M2J	North York	Don Valley Village	27051	4.20	6440	96142.099329

The sample merged data that contain demographic and location data is shown below. The list corresponds to neighbourhoods which have met the demographic criteria for our ideal location and customers. It can be used for further analysis to find a suitable location for the new restaurant.

```
dfc = pd.merge(dfM2, df_TNCoords, how='inner', on = 'Postcode')
# neighbourhoods with high average income based on available data
print(dfc.shape)
dfn=dfc.sort_values('Average Income',ascending = False).reset_index(drop=True)
dfn
```

(42, 9)

27]:

	Postcode	Borough	Neighbourhood	Population	Area square km	Density	Average Income	Latitude	Longitude
0	M4W	Downtown Toronto	Rosedale-Moore Park	20923	4.65	4499	327842.352420	43.679563	-79.377529
1	M4N	Central Toronto	Lawrence Park South	15179	3.24	4684	282244.042538	43.728020	-79.388790
2	M8X	Etobicoke	Kingsway South	9271	2.58	3593	208942.045617	43.653654	-79.506944
3	M4G	East York	Leaside-Bennington	16828	4.68	3595	195253.063975	43.709060	-79.363452
4	M4V	Central Toronto	Forest Hill South	10732	2.45	4380	182035.268195	43.686412	-79.400049
5	M5R	Central Toronto	Annex,Casa Loma	41494	4.74	8754	155187.997998	43.672710	-79.405678
6	M4M	Old City of Toronto	South Riverdale	27876	8.89	3135	153733.028955	43.659526	-79.340923
7	M5E	Downtown Toronto	Waterfront Communities-The Island	65913	7.37	8943	153659.253917	43.644771	-79.373306
8	M4E	East Toronto	The Beaches	21567	3.56	6058	152930.551411	43.676357	-79.293031
9	M2L	North York	St.Andrew-Windfields	17812	7.33	2430	150633.549340	43.757490	-79.374714
10	M1G	Scarborough	Woburn	53485	12.31	4344	145933.058584	43.770992	-79.216917
11	M3A	North York	Parkwoods-Donalda	34805	7.42	4690	144639.211442	43.753259	-79.329656
12	M5M	North York	Bedford Park-Nortown,Lawrence Park North	37843	7.80	4851	138167.075770	43.733283	-79.419750
13	M9B	Etobicoke	Eringate-Centennial-West Deane,Islington-City ...	73604	29.94	2458	134993.473120	43.650943	-79.554724
14	M4V	East Toronto	Yonge-St.Clair	12528	1.17	10707	125560.708023	43.686412	-79.400049
15	M8Y	Etobicoke	Stonegate-Queensway	25051	7.83	3199	119843.838510	43.636258	-79.498509

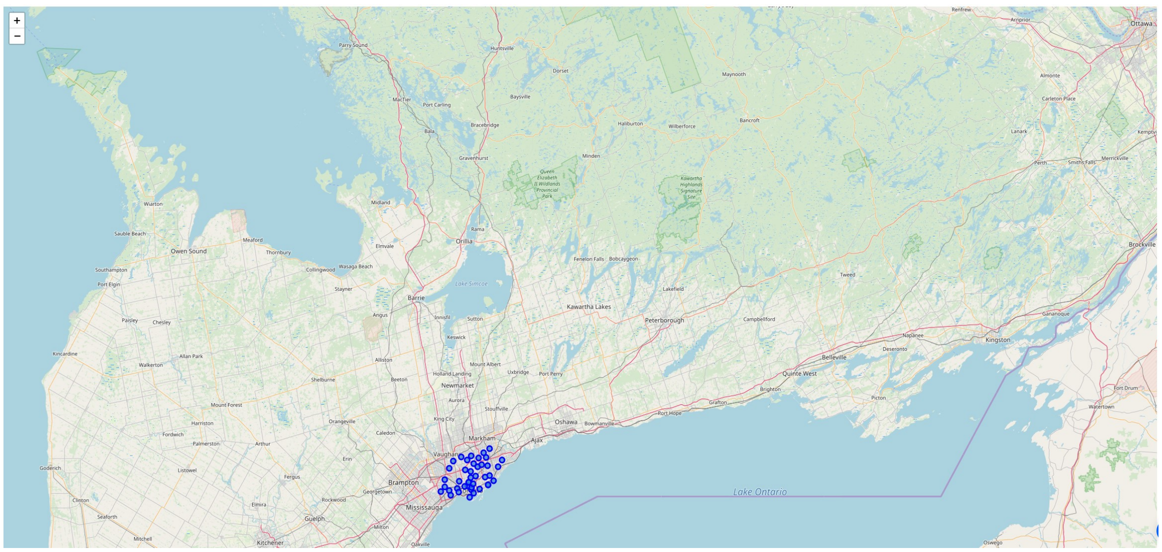
Neighbourhood Maps

The visualization of the selected neighbourhoods on the map for the city of Toronto is shown below. The sample of preprocessed data including geolocation coordinates and demographic can be seen below. The top 100 popular venues on these neighbourhoods around the radius of 1 km are extracted from foursquare.

```
map_TorontoNeighbourhoods = folium.Map(location=[latitude, longitude], zoom_start=9)

# add markers to map
for lat, lng, label in zip(dfn['Latitude'], dfn['Longitude'], dfn['Neighbourhood']):
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_TorontoNeighbourhoods)

map_TorontoNeighbourhoods
```



Neighbourhoods Venues

The venues information is obtained from foursquare. We want to understand the venues that exist on selected location. The information related to the venues will help to shortlist the neighbourhoods for suitable location to open the restaurant.

```python
venues = results['response']['groups'][0]['items']

nearby_venues = json_normalize(venues) # flatten JSON

# filter columns
filtered_columns = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng']
nearby_venues =nearby_venues.loc[:, filtered_columns]

# filter the category for each row
nearby_venues['venue.categories'] = nearby_venues.apply(get_category_type, axis=1)

# clean columns
nearby_venues.columns = [col.split(".")[-1] for col in nearby_venues.columns]

nearby_venues.head(20)
```

37]:

	name	categories	lat	lng
0	Summerhill Market	Grocery Store	43.686265	-79.375458
1	Toronto Lawn Tennis Club	Athletics & Sports	43.680667	-79.388559
2	Black Camel	BBQ Joint	43.677016	-79.389367
3	Craigleigh Gardens	Park	43.678099	-79.371586
4	Pie Squared	Pie Shop	43.672143	-79.377856
5	Tinuno	Filipino Restaurant	43.671281	-79.374920
6	Starbucks	Coffee Shop	43.671478	-79.380664
7	Manulife Financial	Office	43.672070	-79.382449
8	Booster Juice	Smoothie Shop	43.671566	-79.378581
9	Aroma Espresso Bar	Coffee Shop	43.672154	-79.377885
10	No Frills	Grocery Store	43.671616	-79.378187
11	Nijo Japanese Restaurant	Japanese Restaurant	43.671849	-79.378824

The table is the sample of venues for some of the selected locations in Toronto. There are about 257 unique categories and 1950 venues corresponding to 42 selected locations based on data from foursquare. The search was for top 100 popular venues with radius of one kilometer. However, some of the location had less than 100 venues. For example Agincourt North, Milliken had 29 venues within radius of 1 km.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt North,Milliken	29	29	29	29	29	29
Annex,Casa Loma	100	100	100	100	100	100
Banbury-Don Mills	31	31	31	31	31	31
Bay Street Corridor	100	100	100	100	100	100
Bayview Village,Bayview Woods-Steeles	15	15	15	15	15	15
Bedford Park-Nortown,Lawrence Park North	37	37	37	37	37	37
Birchcliffe-Cliffside	16	16	16	16	16	16
Church-Yonge Corridor	100	100	100	100	100	100
Clairlea-Birchmount	31	31	31	31	31	31
Corso Italia-Davenport	24	24	24	24	24	24
Don Valley Village	44	44	44	44	44	44
East End-Danforth	80	80	80	80	80	80
Edenbridge-Humber Valley	13	13	13	13	13	13
Eglinton East	11	11	11	11	11	11
Eringate-Centennial-West Deane,Islington-City Centre West,Princess-Rosethorn	16	16	16	16	16	16
Forest Hill South	77	77	77	77	77	77
Glenfield-Jane Heights	8	8	8	8	8	8
High Park North,High Park-Swansea,Junction Area	98	98	98	98	98	98
Kingsway South	46	46	46	46	46	46
L'Amoreaux,Steeles	24	24	24	24	24	24
Lansing-Westgate,Willowdale East	100	100	100	100	100	100

```python
#we want venues that contain Sushi or Japanese or Ramen
Toronto_Venues_restaurant = Toronto_venues[Toronto_venues['Venue Category'].str.contains('Sushi|Japanese|Ramen')].reset_index(drop=True)
print(Toronto_Venues_restaurant.shape)
Toronto_Venues_restaurant
```

```
(31, 7)
```

71]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Leaside-Bennington	43.709060	-79.363452	Kintako Japanese Restaurant	43.711597	-79.363962	Sushi Restaurant
1	Leaside-Bennington	43.709060	-79.363452	Maki Sushi	43.710127	-79.362466	Sushi Restaurant
2	Forest Hill South	43.686412	-79.400049	Daeco Sushi	43.687838	-79.395652	Sushi Restaurant
3	Bedford Park-Nortown,Lawrence Park North	43.733283	-79.419750	Sakura Garden	43.733398	-79.419491	Sushi Restaurant
4	Bedford Park-Nortown,Lawrence Park North	43.733283	-79.419750	Issmi Sushi	43.735269	-79.419983	Sushi Restaurant
5	Yonge-St.Clair	43.686412	-79.400049	Daeco Sushi	43.687838	-79.395652	Sushi Restaurant
6	Banbury-Don Mills	43.745906	-79.352188	Gonoe Sushi	43.745737	-79.345991	Japanese Restaurant
7	Lansing-Westgate,Willowdale East	43.770120	-79.408493	Konjiki Ramen	43.766998	-79.412222	Ramen Restaurant
8	Lansing-Westgate,Willowdale East	43.770120	-79.408493	Wako Sushi + Bar	43.770806	-79.413138	Japanese Restaurant
9	Lansing-Westgate,Willowdale East	43.770120	-79.408493	KINTON RAMEN	43.769684	-79.413049	Ramen Restaurant
10	Lansing-Westgate,Willowdale East	43.770120	-79.408493	Ajisen Ramen 味千ラーメン	43.771444	-79.413139	Ramen Restaurant
11	Lansing-Westgate,Willowdale East	43.770120	-79.408493	Aburi Room	43.769197	-79.414039	Sushi Restaurant
12	Lansing-Westgate,Willowdale East	43.770120	-79.408493	PROJECT:FISH	43.769100	-79.414305	Sushi Restaurant
13	Church-Yonge Corridor	43.665860	-79.383160	Sansotei Ramen 三草亭	43.666735	-79.385353	Ramen Restaurant
14	Church-Yonge Corridor	43.665860	-79.383160	Tokyo Grill	43.665085	-79.384707	Japanese Restaurant
15	Church-Yonge Corridor	43.665860	-79.383160	Kawa Sushi	43.663894	-79.380210	Japanese Restaurant
16	Church-Yonge Corridor	43.665860	-79.383160	Tokyo Sushi	43.665885	-79.386977	Sushi Restaurant

Below is an example which show Japanese restaurant is on the top 10 most popular venues for Bayview village, Bayview Woods-Steeles location.

```python
num_top_venues = 10
for hood in Toronto_grouped['Neighborhood']:
    print("----"+hood+"----")
    temp = Toronto_grouped[Toronto_grouped['Neighborhood'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

```
----Bayview Village,Bayview Woods-Steeles----
                venue  freq
0  Japanese Restaurant  0.13
1                 Bank  0.13
2        Grocery Store  0.13
3         Intersection  0.07
4                 Park  0.07
5                Trail  0.07
6                 Café  0.07
7   Chinese Restaurant  0.07
8        Shopping Mall  0.07
9           Skate Park  0.07


----Bedford Park-Nortown,Lawrence Park North----
                    venue  freq
0      Italian Restaurant  0.08
1             Coffee Shop  0.08
2     Fast Food Restaurant  0.05
```

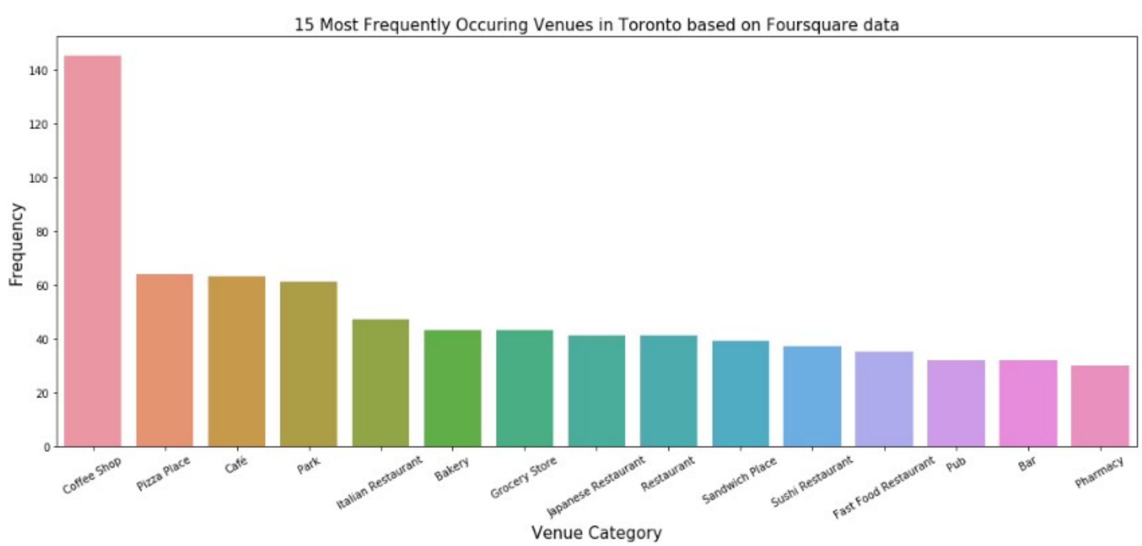
The list of the top 15 most popular venues includes Sushi and Japanese restaurant.

```
# create a dataframe of top 15 categories

Toronto_Venues_Top15 = Toronto_venues['Venue Category'].value_counts()[0:15].to_frame(name='frequency')
Toronto_Venues_Top15=Toronto_Venues_Top15.reset_index()

Toronto_Venues_Top15.rename(index=str, columns={"index": "Venue Category", "frequency": "Frequency"}, inplace=True)
Toronto_Venues_Top15
```

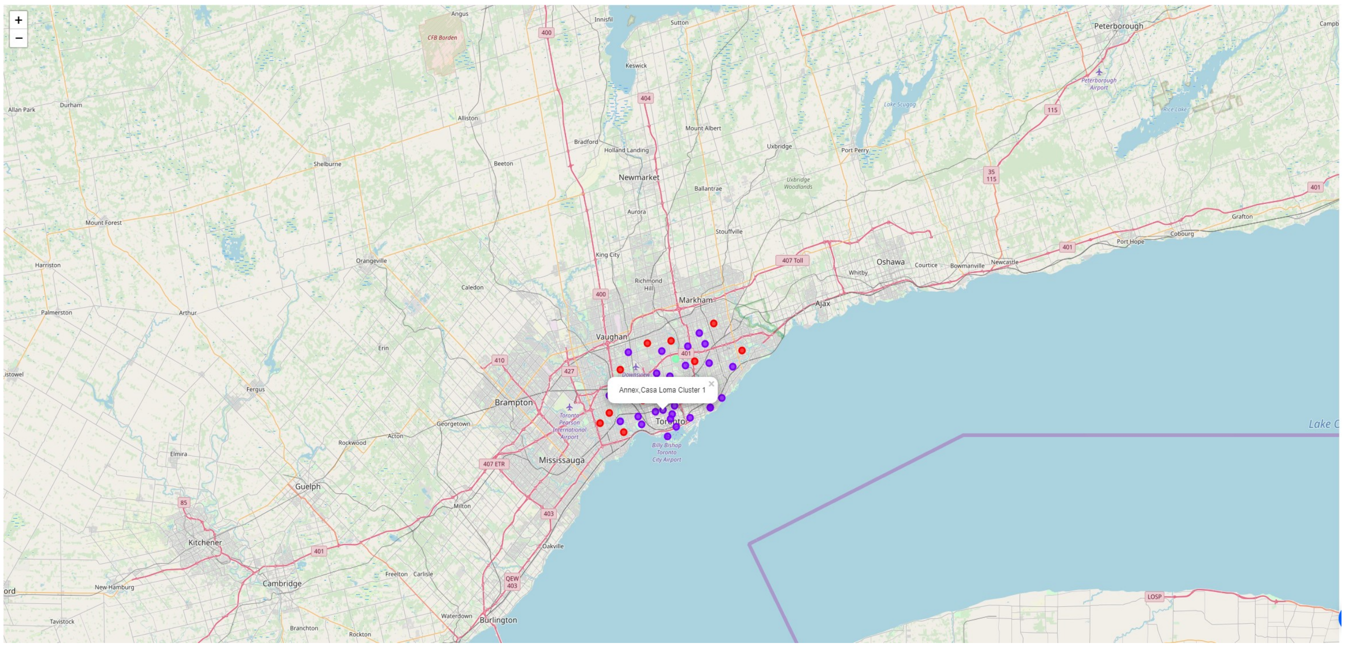
	Venue Category	Frequency
0	Coffee Shop	145
1	Pizza Place	64
2	Café	63
3	Park	61
4	Italian Restaurant	47
5	Bakery	43
6	Grocery Store	43
7	Japanese Restaurant	41
8	Restaurant	41
9	Sandwich Place	39
10	Sushi Restaurant	37
11	Fast Food Restaurant	35
12	Pub	32
13	Bar	32
14	Pharmacy	30



Neighbourhoods Clusters

The popular venues from the clusters are analyzed in order to determine suitable location for a new restaurant. The k-means clustering method is used to create 3 clusters out of the selected neighbourhoods. The high foot and car traffic of the location that is based on number of most popular business exists on a particular location will be used to create location shortlist.

Toronto map which has cluster of locations display in 3 different colors:



Cluster Labels 0

The Park is among the top ten most popular venues for all location on Cluster Labels equal to zero. There are also banks, bakeries, coffee shops, café, malls, pharmacies, stores and restaurants in most of these locations. The existence of other established popular business venue is an indication of high foot and car traffic. Hence some of the locations in this cluster are suitable for a new location of restaurant.

	Neighbourhood	Population	Average Income	Latitude	Longitude	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
0	Woburn	53485	145933.058584	43.770992	-79.216917	0	Park	Coffee Shop	Fast Food Restaurant	Pharmacy	Chinese Restaurant	Indian Restaurant	Electronics Store	Event Space	Ethiopian Restaurant	Design Studio
1	Parkwoods-Donalda	34805	144639.211442	43.753259	-79.329656	0	Park	Pharmacy	Shopping Mall	Convenience Store	Bus Stop	Skating Rink	Tennis Court	Food & Drink Shop	Café	Supermarket
2	Eringate-Centennial-West Deane,Islington-City ...	73604	134993.473120	43.650943	-79.554724	0	Park	Pizza Place	Hotel	Bank	Gym	American Restaurant	Café	Theater	Mexican Restaurant	Fish & Chips Shop
3	Stonegate-Queensway	25051	119843.838510	43.636258	-79.498509	0	Italian Restaurant	Park	Shopping Mall	Ice Cream Shop	Gym / Fitness Center	Eastern European Restaurant	Diner	Discount Store	Dive Bar	Dog Run
4	Agincourt North,Milliken	55685	110493.664337	43.815252	-79.284577	0	Chinese Restaurant	Korean Restaurant	Pizza Place	Pharmacy	Bakery	Noodle House	Park	Malay Restaurant	Dessert Shop	Caribbean Restaurant
5	Edenbridge-Humber Valley	15535	102322.730799	43.667856	-79.532242	0	Pharmacy	Playground	Shopping Mall	Bakery	Café	Baseball Field	Golf Course	Park	Convenience Store	Grocery Store
6	Glenfield-Jane Heights	30491	95121.544626	43.739015	-79.506944	0	Bank	Coffee Shop	Vietnamese Restaurant	Pizza Place	Grocery Store	Spa	Park	Shopping Mall	Dog Run	Dessert Shop
7	Corso Italia-Davenport	14133	70527.081077	43.689026	-79.453512	0	Pharmacy	Mexican Restaurant	Park	Bakery	Thai Restaurant	Market	Fast Food Restaurant	Bus Stop	Sporting Goods Shop	Beer Store
8	Bayview Village,Bayview Woods-Steeles	34550	70518.961182	43.786947	-79.385975	0	Japanese Restaurant	Grocery Store	Bank	Park	Café	Intersection	Chinese Restaurant	Trail	Shopping Mall	Skate Park
9	Westminster-Branson,Willowdale West	43210	70267.476459	43.782736	-79.442259	0	Pharmacy	Pizza Place	Bakery	Discount Store	Butcher	Eastern European Restaurant	Park	Bus Line	Convenience Store	Coffee Shop

Cluster Labels 1

The coffee shop is among the most popular venue on cluster labels 1. The other venues categories on top ten of most common venues are stores, restaurants, banks, pubs, bars, gyms, bakeries and pizza places. This cluster has thirty one locations out of the total forty two locations.

	Neighbourhood	Population	Average Income	Latitude	Longitude	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
0	Rosedale-Moore Park	20923	327842.352420	43.679563	-79.377529	1	Coffee Shop	Park	Grocery Store	Bank	Filipino Restaurant	Metro Station	Bistro	Breakfast Spot	Building	Smoothie Shop
1	Lawrence Park South	15179	282244.042538	43.728020	-79.388790	1	College Quad	Coffee Shop	Café	Park	Bus Line	Bookstore	Trail	Gym / Fitness Center	College Gym	Falafel Restaurant
2	Kingsway South	9271	208942.045617	43.653654	-79.506944	1	Coffee Shop	Park	Italian Restaurant	Burger Joint	French Restaurant	Breakfast Spot	Sushi Restaurant	Dessert Shop	Pub	Gourmet Shop
3	Leaside-Bennington	16828	195253.063975	43.709060	-79.363452	1	Coffee Shop	Sporting Goods Shop	Electronics Store	Furniture / Home Store	Brewery	Sushi Restaurant	Restaurant	Sports Bar	Burger Joint	Sandwich Place
4	Forest Hill South	10732	182035.268195	43.686412	-79.400049	1	Coffee Shop	Sushi Restaurant	Park	Italian Restaurant	Thai Restaurant	Grocery Store	Pizza Place	Liquor Store	Sandwich Place	Restaurant
5	Annex,Casa Loma	41494	155187.997998	43.672710	-79.405678	1	Café	Vegetarian / Vegan Restaurant	Coffee Shop	Italian Restaurant	Bakery	Restaurant	Museum	Gym	Pizza Place	Pub
6	South Riverdale	27876	153733.028955	43.659526	-79.340923	1	Coffee Shop	Bar	Café	Bakery	Vietnamese	American	Brewery	Italian	Diner	French

Cluster Labels 2

It only have one location. Unlike other locations this one has a farm and farmers market on its list of top ten most common venues.

	Neighbourhood	Population	Average Income	Latitude	Longitude	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
0	St.Andrew-Windfields	17812	150633.54934	43.75749	-79.374714	2	Park	Pool	Design Studio	Fast Food Restaurant	Farmers Market	Farm	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store

Results and Discussion

The results from clustering indicated that two out of the three clusters contained locations with high foot and car traffic. However, some of these locations have a lot of restaurants or don't have a lot of business to generate high foot and car traffic. The table below contain locations which have either less than 3 Japanese restaurants or more than 10 businesses. The locations which have a lot of Japanese restaurants and fewer businesses were eliminated due to high competition.

Number of Japanese restaurants and businesses within radius of 1km of the location				
Neighborhood	Japanese Cuisine restaurants	Number of Businesses	Total of All Restaurants	Population
Annex,Casa Loma	2	47	31	41,494
Bay Street Corridor	2	53	24	25,797
Bedford Park-Nortown,Lawrence Park North	2	16	13	37,843
Corso Italia-Davenport	1	11	6	14,133
Church-Yonge Corridor	10	45	35	31,340
Don Valley Village	2	29	8	27,051
East End-Danforth	4	43	15	21,381
Eglinton East	1	3	5	22,776
Forest Hill South	5	36	22	10,732
High Park North,High Park-Swansea,Junction Ar	3	50	22	60,453
Kingsway South	2	23	12	9,271
Lansing-Westgate,Willowdale East	14	48	36	66,598
Leaside-Bennington	3	36	10	16,828
Mount Pleasant East,Mount Pleasant West	8	45	35	46,433
Palmerston-Little Italy	3	50	35	13,826
Parkwoods-Donalda	0	15	3	34,805
Roncesvalles	4	43	22	14,974
Rosedale-Moore Park	1	14	4	20,923
South Riverdale	2	42	30	27,876
The Beaches	4	43	15	21,567
Waterfront Communities-The Island	4	44	19	65,913
Yonge-St.Clair	5	36	22	12,528
York University Heights	2	12	8	27,593

Keywords for Business: Bank, Mall, Grocery, Store, pub, Historic, Building, Shop, Museum, Plaza, Gallery, Business, Pub, Bar, Office, Theater, Studio, Gym, Café, Coffee, Pizza, Burger, Storage, Tea, Breakfast
Keywords for Restaurant: BBQ, Restaurant

Keywords for Japanese restaurant: Ramen, Sushi, and Japanese

The locations which are highlighted in green are the one we want to keep. The selection of these locations was based on number of existing Japanese restaurant, total number of other businesses, total number of restaurants and population. Below is the shortlist of locations for a new Japanese restaurant that need to undergo further analysis in order to pick a suitable cost effective location.

1. **Annex, Casa Loma** is among the locations in Toronto with high foot and car traffic. It has a population of over forty thousand, and a lot of businesses.
2. **Bay Street Corridor** has a lot of businesses in comparison to others and one of its most popular venues is Ramen Restaurant which has a lot of customers in the city of Toronto.

3. **Don Valley Village** it has fewer restaurants overall, about twenty nine other businesses, and Japanese restaurant is among the top ten most popular venue in this location.
4. **High Park North, High Park-Swansea, Junction Area** has a population of over sixty thousands, a lot of other businesses and fewer restaurants relative to other locations.
5. **Parkwoods-Donalda** has no Japanese restaurant and it has three restaurants and fifteen businesses based on venue data from foursquare.
6. **Rosedale-Moore Park** is high income neighbourhood with four restaurants, fifteen businesses and one Japanese restaurant
7. **Waterfront Communities-The Island** has a lot of businesses, and population of sixty five thousands

Conclusion and Recommendations

The purpose of this project was to create a location shortlist for a new restaurant based on location and demographics data. The process involved data acquisition from multiple sources, cleansing, transforming, and analyzing. The result is we were able to reduce the number of locations to seven. The selected locations have high foot and car traffic, and fewer Japanese restaurants.

Further analysis is required in order to pick one location for a new Japanese restaurant. The analysis should include cost analysis for location that factor in renting or buying cost and other operational cost including labour cost and other related fees. It should take into consideration among other things information about availability of parking space, accessibility and crime rate. Furthermore it should include demographic data related to average age and median income for reliable sources.

References

https://www.theglobeandmail.com/life/food-and-wine/restaurant-reviews/article-torontos-food-halls-point-to-a-grown-up-city-with-a-demand-for/

https://www.blogto.com/eat_drink/2018/09/toronto-ramen-joint-cant-keep-overwhelming-demand/

bentoboxmag.ca/2018/11/18/a-toronto-tradition-returns/

https://www.foodnewsfeed.com/fsr/expert-insights/8-factors-choosing-new-restaurant-location