Garry

About

Decision Making for Opening a New Restaurant in Toronto

Garry Just now · 9 min read

Sometimes we will have a question. Can we put a point on the map and know that we can start a business there? This will certainly help make the right business decisions. As a matter of fact, data can be utilized in order to aid the decision making process. Specifically, it is possible aggregate information from different neighborhoods and find out the most suitable location to start a business.

The Problem

According to Wikipedia [1], there are 140 officially recognized neighborhoods. Supposing we would like to open a restaurant, it can be a difficult task to decide on the location to open the restaurant. Another question to be ask is the type of restaurant to be opened. Searching for the answer for this question is even more difficult during the COVID-19 pandemic. To this end, the business question we are asking here are:

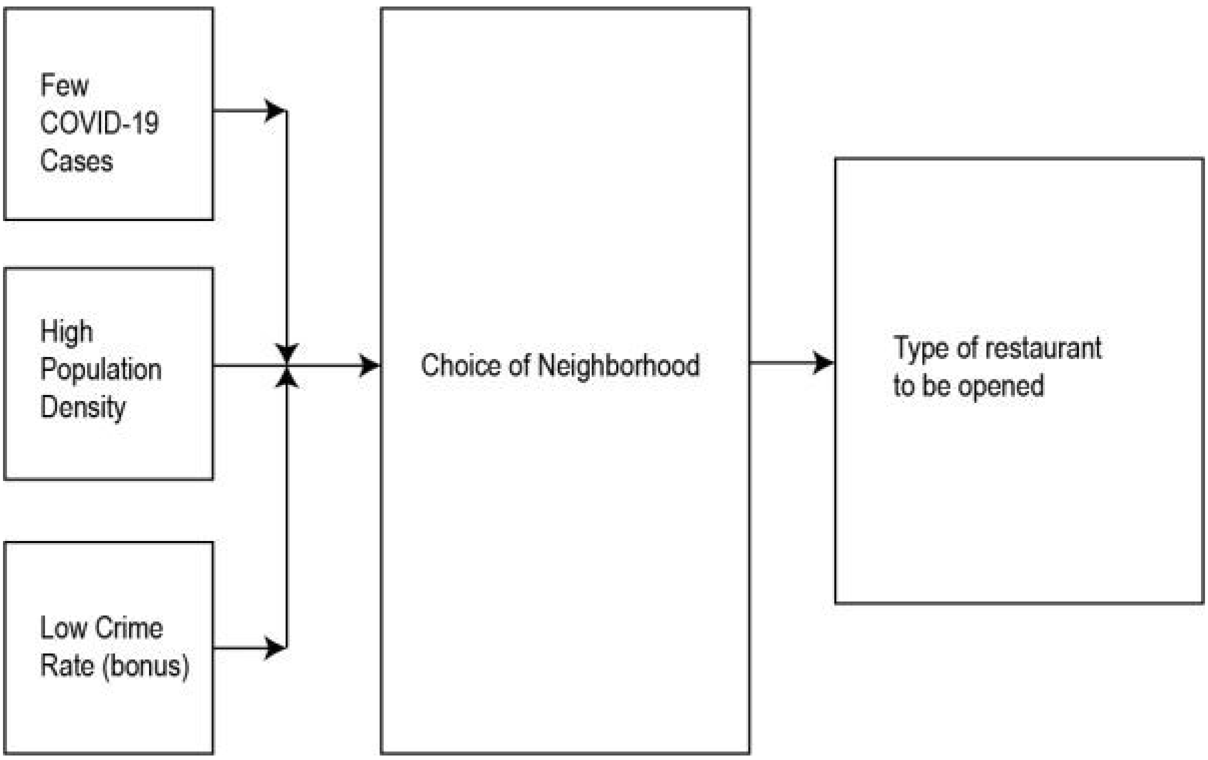
> *During the COVID-19 Pandemic,*
>
> *What is the the most suitable neighborhood to start a new restaurant in Toronto?*
>
> *What should be the type of restaurant to be opened?*

So, we will attempt to answer this two questions by using data and information of the neighborhoods.

Methodology

Before actually collecting data to answer the above questions, it is essential to have a framework to make the decision. In fact, the first aspect to consider is the spread of COVID-19. If there are a lot of individuals being infected in a neighborhood, there can be a regional lockdown in the neighborhood. Besides, the fear of being infected can also hinder people to eat out. As a result, it will be a sensible decision to start a restaurant in a neighborhood with fewer COVID-19 cases. The next aspect to consider is the population in the neighborhood. The best bet here will be search for the neighborhood with higher density. This means that there will be more people in the neighborhood on average and increase the chance of getting new customers. The neighborhood itself should also be considered. Are there are different types of venues or facilities in the neighborhood? What are the different type of restaurants which are already in the neighborhood? This will help to answer the question of what type of restaurant to be opened. It will also be a bonus if the crime rate in the neighborhood is low. This will imply that the neighborhood is a safe one. People may be more willing to hang out in this case.

The decision framework can be summarized by using the following figure.



Decision Framework for the problem of starting a new restaurant

B. Data Requirements

will be obtained from this Open Data Portal:

1. Toronto Neighborhood Profile (https://open.toronto.ca/dataset/neighbourhood-profiles/)
2. Toronto Neighborhood Crime Rate (https://open.toronto.ca/dataset/neighbourhood-crime-rates/)
3. COVID-19 Cases in Toronto (https://open.toronto.ca/dataset/covid-19-cases-in-toronto/)

Besides these three datasets, it is also needed to aggregate the information of different facilities, especially restaurants, of the neighborhoods. For this type of data, the Foursquare API (https://developer.foursquare.com/) will be used.

C. Data Extraction and Cleaning

There are a few steps in the data analysis stage. First of all, relevant information from the datasets will be extract. There is a lot of information which is not needed from the Neighborhood Profile dataset. Only data related to population density will be extracted and utilized.

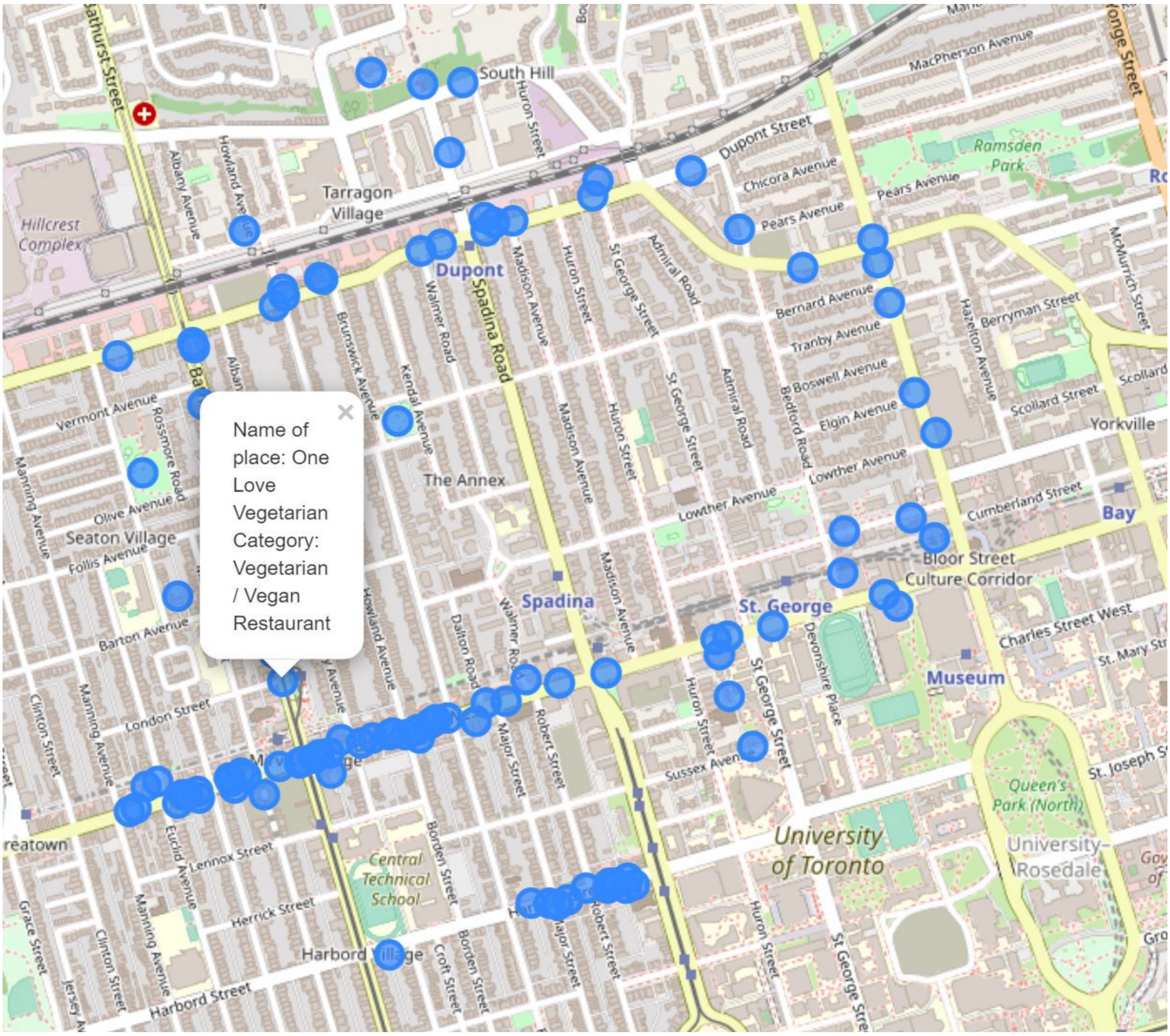
	A	B	C	D	E	F	G	H	I	J
1	_id	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt Nort	Agincou	Alderw	Annex
2	1	Neighbourhood	Neighbourhood Information	City of Toronto	Neighbourhood Number		129	128	20	95
3	2	Neighbourhood	Neighbourhood Information	City of Toronto	TSNS2020 Designation		No Designation	No Desigr	No Desigr	No Desigr
4	3	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2016	2,731,571	29,113	23,757	12,054	30,526
5	4	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2011	2,615,060	30,279	21,988	11,904	29,177
6	5	Population	Population and dwellings	Census Profile 98-316-X2016001	Population Change 2011-:	4.50%	-3.90%	8.00%	1.30%	4.60%
7	6	Population	Population and dwellings	Census Profile 98-316-X2016001	Total private dwellings	1,179,057	9,371	8,535	4,732	18,109
8	7	Population	Population and dwellings	Census Profile 98-316-X2016001	Private dwellings occupie	1,112,929	9,120	8,136	4,616	15,934
9	8	Population	Population and dwellings	Census Profile 98-316-X2016001	Population density per sq	4,334	3,929	3,034	2,435	10,863
10	9	Population	Population and dwellings	Census Profile 98-316-X2016001	Land area in square kilom	630.2	7.41	7.83	4.95	2.81
11	10	Population	Age characteristics	Census Profile 98-316-X2016001	Children (0-14 years)	398,135	3,840	3,075	1,760	2,360
12	11	Population	Age characteristics	Census Profile 98-316-X2016001	Youth (15-24 years)	340,270	3,705	3,360	1,235	3,750
13	12	Population	Age characteristics	Census Profile 98-316-X2016001	Working Age (25-54 year	1,229,555	11,305	9,965	5,220	15,040
14	13	Population	Age characteristics	Census Profile 98-316-X2016001	Pre-retirement (55-64 yea	336,670	4,230	3,265	1,825	3,480
15	14	Population	Age characteristics	Census Profile 98-316-X2016001	Seniors (65+ years)	426,945	6,045	4,105	2,015	5,910
16	15	Population	Age characteristics	Census Profile 98-316-X2016001	Older Seniors (85+ years)	66,000	925	555	320	1,040
17	16	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 0 to 04 years	69,895	660	575	360	445
18	17	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 05 to 09 years	69,350	695	540	270	365
19	18	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 10 to 14 years	64,945	660	460	225	325
20	19	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 15 to 19 years	74,240	840	780	285	465
21	20	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 20 to 24 years	97,415	1015	1000	355	1215
22	21	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 25 to 29 years	113,905	1015	1045	355	2080
23	22	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 30 to 34 years	108,895	835	820	410	1610
24	23	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 35 to 39 years	94,070	680	625	455	1055
25	24	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 40 to 44 years	86,535	760	610	420	835
26	25	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 45 to 49 years	90,860	890	760	440	850
27	26	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 50 to 54 years	98,735	1160	970	515	920
28	27	Population	Age characteristics	Census Profile 98-316-X2016001	Male: 55 to 59 years	88,145	1060	850	540	855

neighbourhood-profiles-2016-csv

The above figure is the dataset Toronto Neighborhood Profile. It can be seen that there is a lot of data and information. However, we will only extract the population in different neighborhood. The cells in the rectangle are some of the information to be extracted from this dataset.

Similarly, it is also need to extract data required from the crime dataset. Only the number of cases of different types of crimes will be extracted from the dataset. For the COVID-19 dataset, the original CSV file comprises information of each case. As we are only interested in the number of cases in each neighborhood, we will group the neighborhoods and count the total number of cases in each neighborhood.

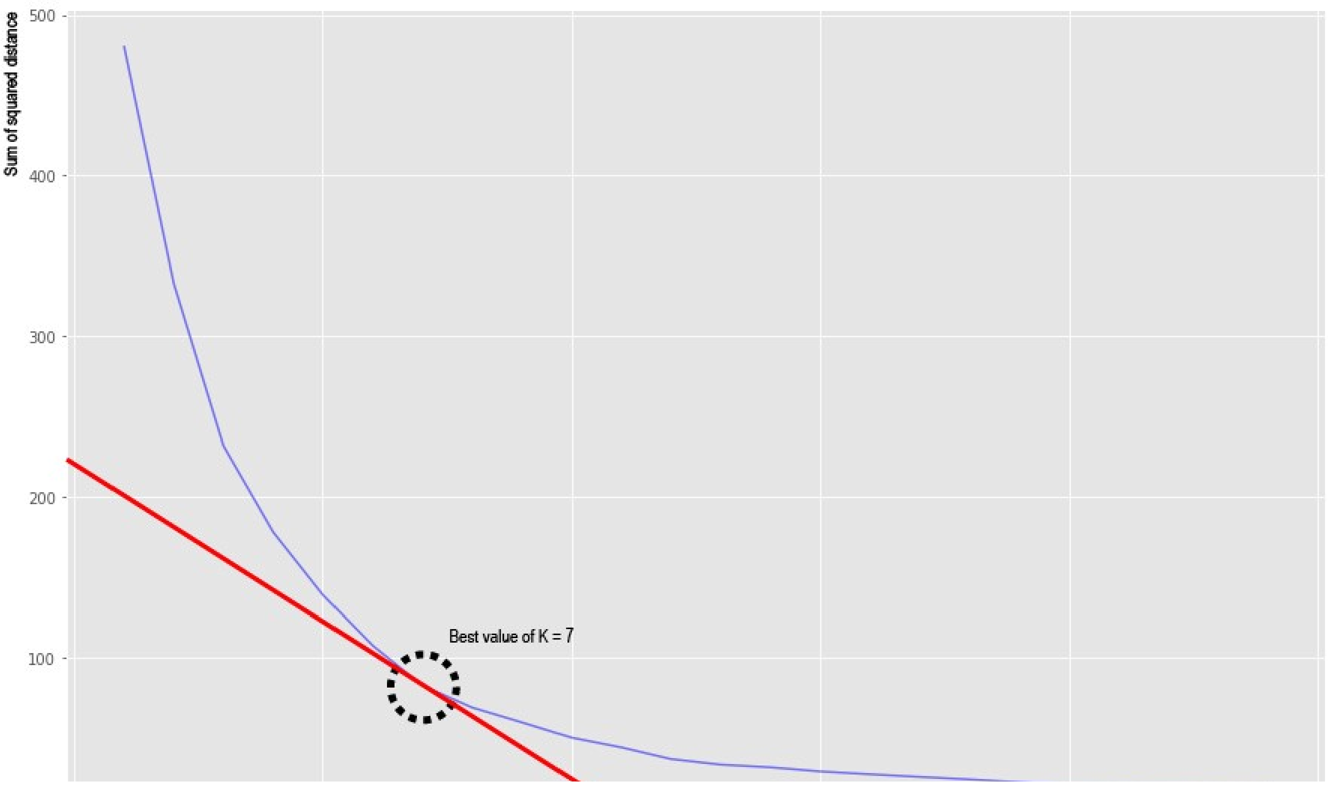
Meanwhile, the information of each neighborhood will be extracted by using Foursquare API. In particular, information of various venues, including but not limited to, restaurants, bookstores and parks, can be extracted.



The above is a screenshot of the information of venues extracted from Foursquare for the neighborhood Annex. Each blue point represents a venue. Here, the name of the venue and the category of the venue are extracted and shown on the map.

D. Data Analysis

To analyze the data, the data of venues in each neighborhood will be extracted using Foursquare API. The number of venues in various categories as a ratio to the total number of venues in each neighborhood will then be calculated. Information about the population density, crime rate and number of COVID-19 cases of each neighborhood will be appended to data extracted from Foursquare. The combined dataset will be analyzed by using the techniques of clustering analysis [2]. In particular, K-means Clustering Algorithm will be adopted [3]. When performing K-means Clustering, it is needed to determine, the number of clusters K. The elbow method will be adopted to determine the best value of K. The sum of squared distance from the centroid to the points in each clusters will be calculated and plotted against the value of K. Here, it can be intuitively imagined that this sum of squared distance will be smaller when the value of K increases. The sum will attain zero when the value of K is equal to the total number of points. Usually, the best value of K is where the absolute value of the slope at that point is the greatest, as shown in the following figure.



Search for the best value of K

For each cluster, the mean population and number of COVID-19 cases can be obtained. Because, the various types of restaurants in the cluster will also be found.

Results

The first step of the project was to gather all the data and transformed it to the format required for analysis. Data analysis was down by using Python with Jupyter Notebook. For the data from the Toronto Open Portal, CSV files were downloaded. The following figure shows the code to download the CSV file of Toronto Neighborhood Profiles.

```
[ ]: import urllib.request

     #get the data from toronto neighborhood profile
     url = 'https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/ef0239b1-832b-4d0b-a1f3-4153e53b189e?format=csv'
     filename = 'toronto_neighborhood_profile_raw.csv'
     urllib.request.urlretrieve(url, filename)
```

Code to download CSV file of Toronto Neighborhood Profiles

As discussed, there a information which was not needed in the CSV file, it will be essential to extract the required information from the CSV file. In the following figures, the code used to clean and extracted data from the Neighborhood Profiles CSV file is presented. The data in the Pandas dataframe is also shown in the figure.

```
import pandas as pd

#read the downloaded csv
toronto_neighborhood_df = pd.read_csv('toronto_neighborhood_profile_raw.csv')
toronto_neighborhood_df = toronto_neighborhood_df.iloc[:,4:]

check_rows = toronto_neighborhood_df["Characteristic"].str.contains(
    'Population, 2016|Total private dwellings|Population density per square kilometre',
    case=False, regex=True)

toronto_neighborhood_df = toronto_neighborhood_df[check_rows]

toronto_neighborhood_df.set_index('Characteristic', inplace=True)
toronto_neighborhood_df.index.name = None
toronto_neighborhood_df = toronto_neighborhood_df.iloc[:,1:].T

for col in toronto_neighborhood_df.columns:
    toronto_neighborhood_df[col] = toronto_neighborhood_df.apply(lambda x: x[col].replace(",",""), axis=1)
    toronto_neighborhood_df[col] = toronto_neighborhood_df[col].astype("float64")

toronto_neighborhood_df.reset_index(inplace=True)
toronto_neighborhood_df.rename(columns={'index': 'Neighborhood'}, inplace=True)

print(toronto_neighborhood_df.shape)
toronto_neighborhood_df.head()
```

```
(140, 4)
```

Neighborhood	Population, 2016	Total private dwellings	Population density per square kilometre

2	Alderwood	12054.0	4732.0	2435.0
3	Annex	30526.0	18109.0	10863.0
4	Banbury-Don Mills	27695.0	12473.0	2775.0

Code to extract required information from Neighborhood Profiles CSV file. The table shows the information stored in the Pandas dataframe after data extraction.

The above process was also performed for the crime rate and COVID-19 cases CSV files. Eventually, the following data was extracted from these two files.

(140, 2)

[4]:

	Neighborhood	Crime Cases
0	South Parkdale	416
1	South Riverdale	490
2	St.Andrew-Windfields	194
3	Taylor-Massey	214
4	Humber Summit	363

(140, 2)

[5]:

	Neighborhood	COVID Cases
0	Agincourt North	884
1	Agincourt South-Malvern West	685
2	Alderwood	274
3	Annex	654
4	Banbury-Don Mills	564

Data extracted from the Crime and COVID-19 cases CSV files

Histograms of Population density, Number of Dwellings, Crimes and COVID-19 Cases of Neighborhood and scatter plots between population and (1) Crime Rate, (2) COVID-19 Rate lower sides of population density, number of dwelling, crime and COVID-19 cases. At this point, it will be essential to combine this data with the information of neighborhoods from Foursquare API in order to gain more insight.

The information of neighborhoods were extracted by using the Foursquare API. In this step, the number of venues in each neighborhood was found, as shown in the following figure.

[49]:

	Neighborhood	Number of Venues
53	Humber Summit	3
50	Highland Creek	4
13	Black Creek	5
90	Oakridge	7
21	Centennial Scarborough	7
20	Centennial	7
75	Martingrove	8
24	Clanton Park	9
101	Rexdale-Kipling	9
59	Kennedy Park	10

Number of venues in some of the Neighborhoods

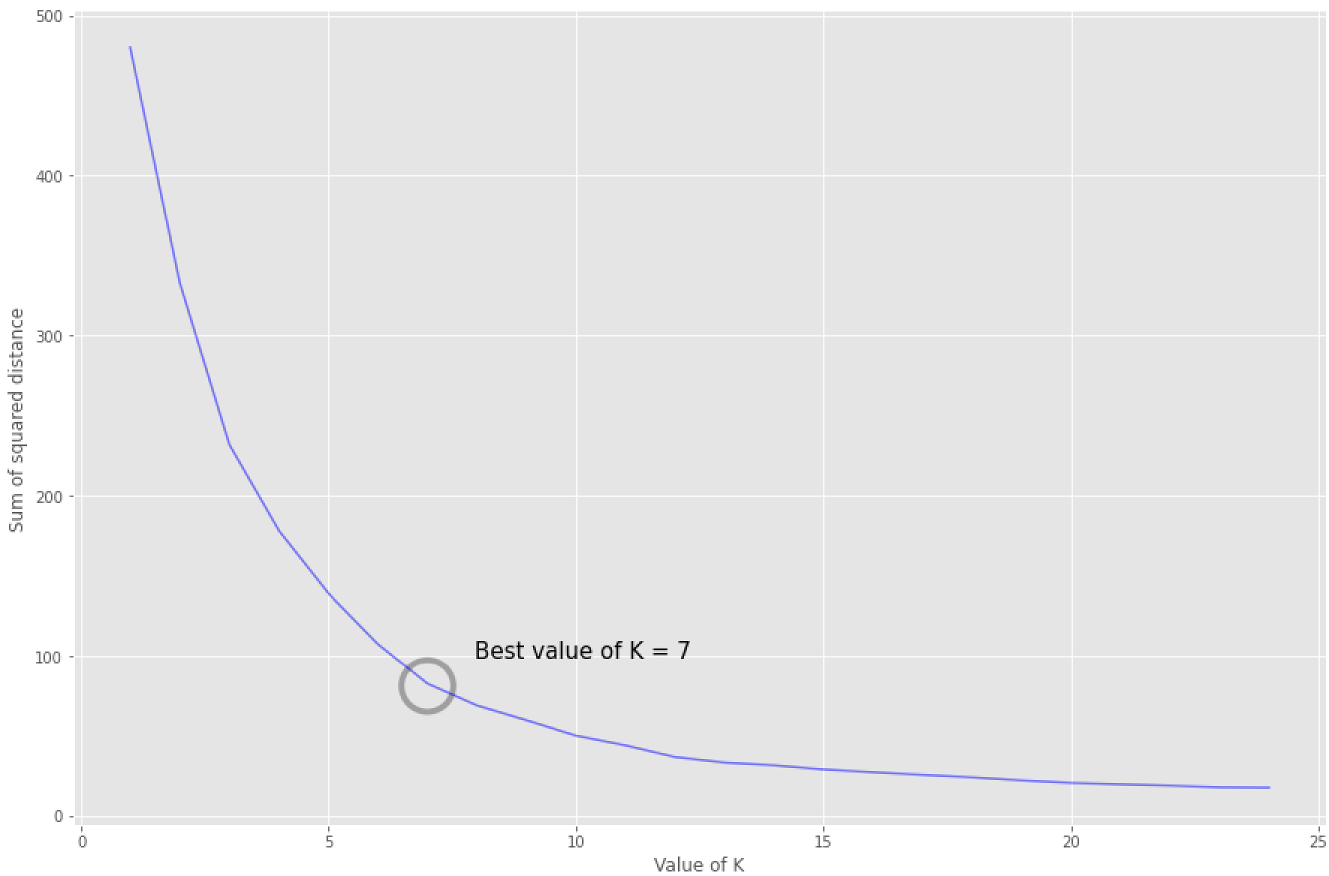
It can be clearly seen that the are neighborhood with only few venues. As the task was to find a place to open a new restaurant. Neighborhoods with too few venues might not considered suitable as this might be a sign that people might not eat out in these neighborhoods. As a result, it has decided that neighborhoods with venue number smaller or equal to 15 would be remove from the analysis.

venue categories in each neighborhood can be found:

	Neighborhood	Population	Total private dwellings	Population density per square kilometre	Crime Cases	COVID Cases	latitude	longitude	Crime Rate	COVID Rate	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 Mo Commo Venu
0	Agincourt North	29113.0	9371.0	3929.0	214.0	884.0	43.808038	-79.266439	0.054467	0.224994	Coffee Shop	Bakery	Indian Restaurant	Bank	Chinese Restaurant	Park	Salon / Barbershop	Sandwic Plac
1	Agincourt South-Malvern West	23757.0	8535.0	3034.0	329.0	685.0	43.781969	-79.257689	0.108438	0.225775	Clothing Store	Restaurant	Gym / Fitness Center	Coffee Shop	Bakery	Bank	Sandwich Place	Gy
2	Alderwood	12054.0	4732.0	2435.0	88.0	274.0	43.601717	-79.545232	0.036140	0.112526	Pharmacy	Discount Store	Pizza Place	Convenience Store	Donut Shop	Shopping Mall	Park	Skatin Ri
3	Amesbury	17757.0	6667.0	5045.0	212.0	1025.0	43.706162	-79.483492	0.042022	0.203171	Portuguese Restaurant	Fast Food Restaurant	Coffee Shop	Pizza Place	Supermarket	Breakfast Spot	Big Box Store	Pa
4	Annex	30526.0	18109.0	10863.0	604.0	654.0	43.670338	-79.407117	0.055602	0.060204	Café	Italian Restaurant	Korean Restaurant	Bakery	Restaurant	Coffee Shop	Vegetarian / Vegan Restaurant	Groce Sto

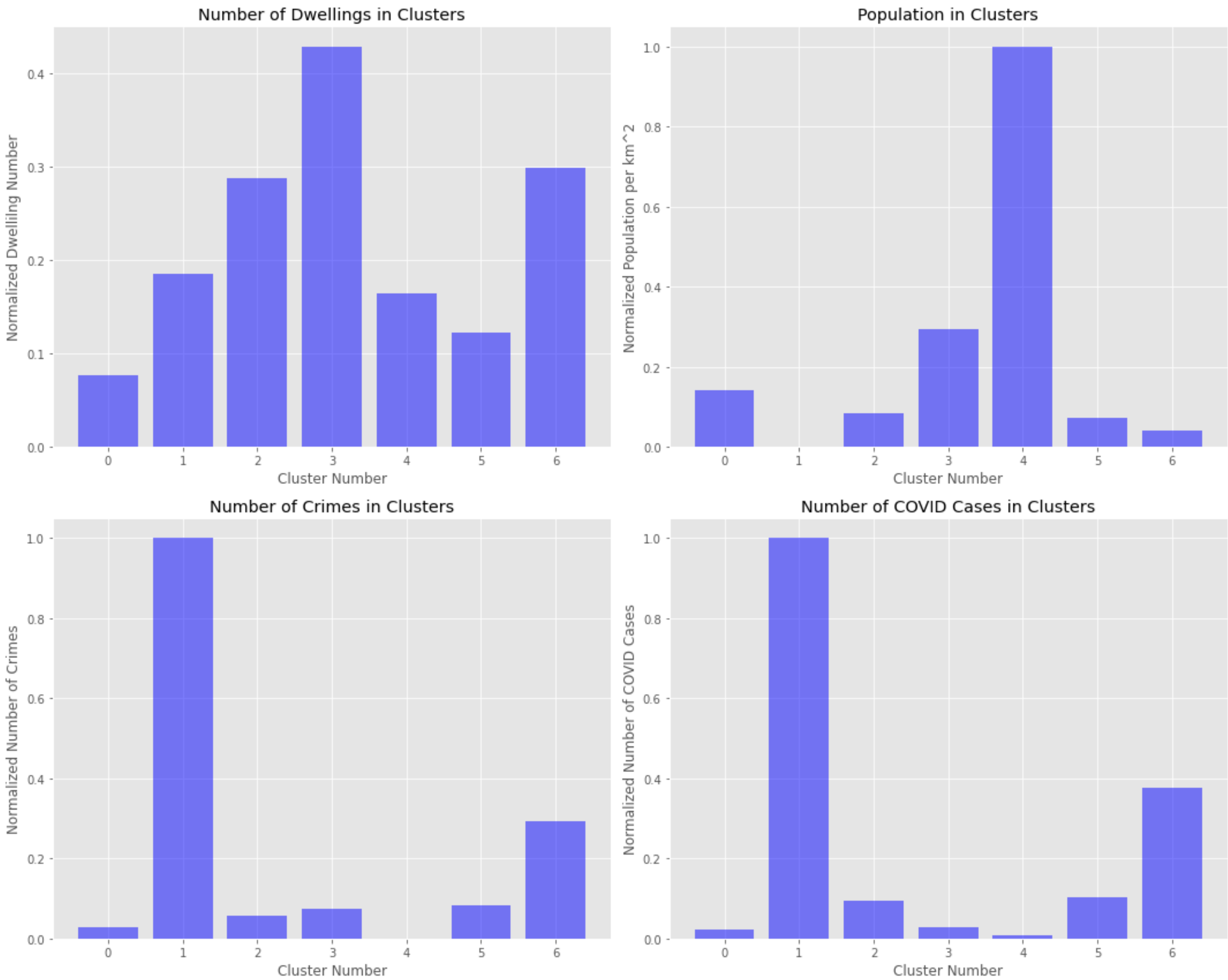
A part of screenshot of the top 10 venue categories in each included neighborhood. Information about COVID-19, crime cases and population are also included in this data frame.

The data was analyzed by using K-means clustering algorithm. By using the elbow method, it has been determined that there should be 7 clusters.



Elbow to determine the best value of K as 7

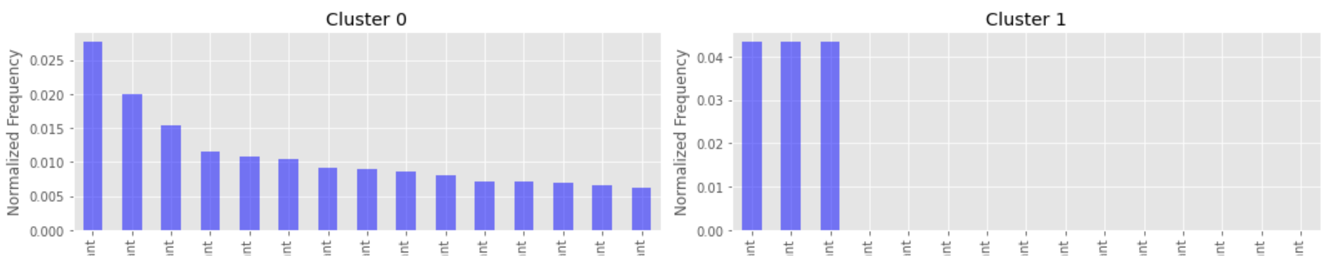
With this cluster, the K-means clustering analysis was run again with K value 7. The resulted cluster label (0 to 6) of each neighborhood was found. The information of

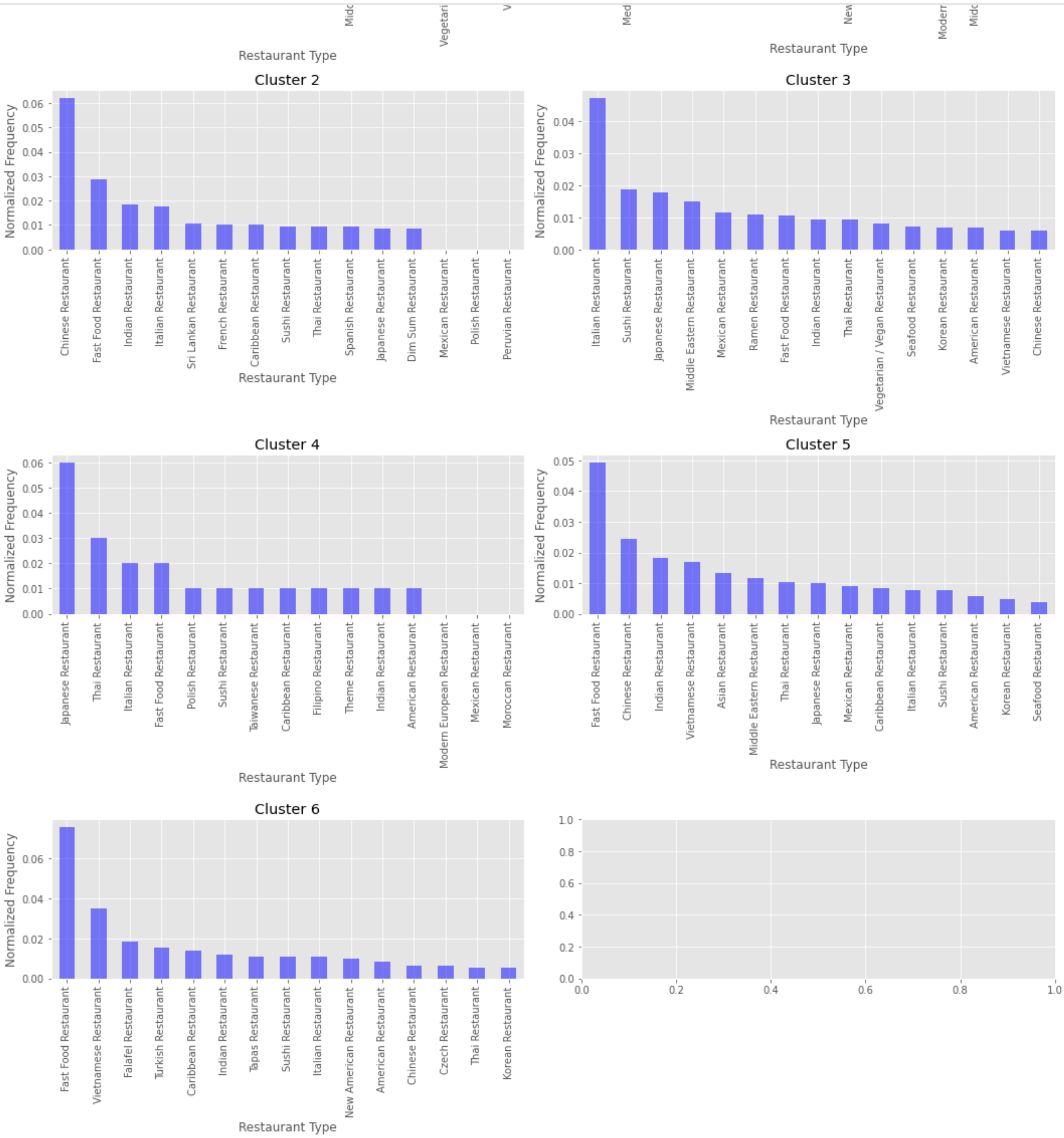


Bar Chart Showing Information about number of dwelling, populations, crime and COVID-19 cases in the clusters. Numbers shown on the y-axes are normalized numbers.

From the above graphs, it can be seen that Cluster 4 was of the higher population, lowest crime and COVID-19 cases. It can be said that Cluster 4 was determined to be the best cluster for opening a new restaurant.

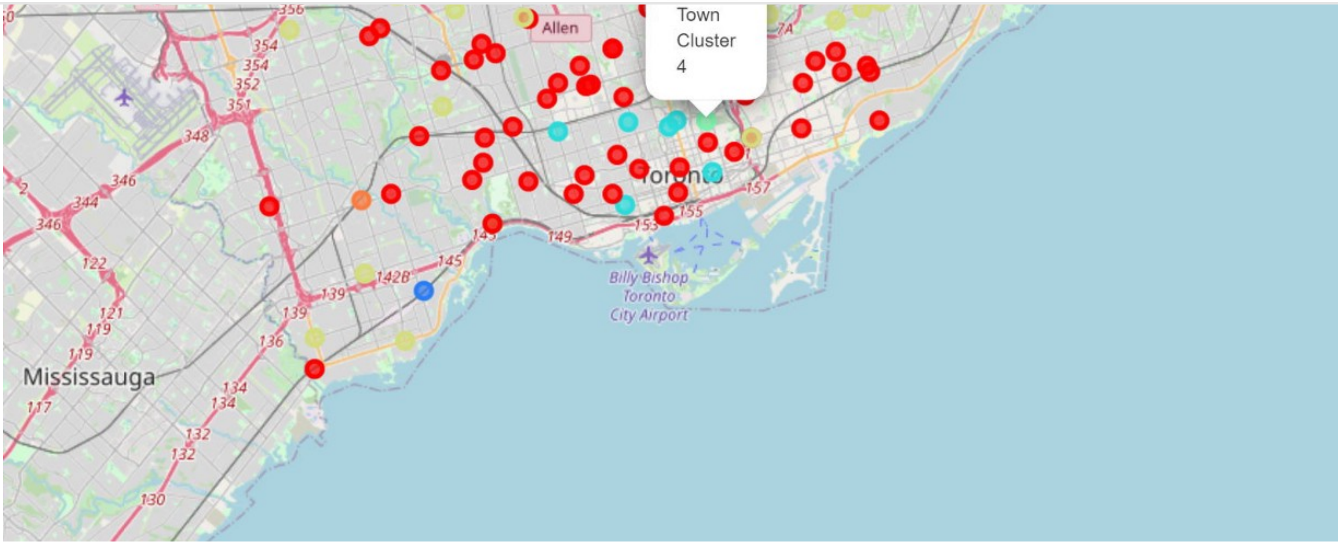
Meanwhile, it was also of importance to under the types of restaurants already in the cluster. We mainly pay attention to Cluster 4. The number of Japanese restaurants in this cluster was of the highest, followed by Thai and Italian restaurants.





Normalized number of types of restaurants in each cluster

So, the next question will be which neighborhoods were in Cluster 4. It is shown in the following figure.



Distribution of neighborhoods in different clusters

Conclusion and Discussion

In this project, a decision framework to choose a neighborhood to open a new restaurant in Toronto. According to our analysis, it has been determined that Cluster 4 was the best neighborhood. In fact, only North St. James Town was in this cluster. It probably implied that this neighborhood is special and distinct compared to other neighborhoods in Toronto. The population in this neighborhood was relative high. The cases of crimes and COVID-19 were low. During the pandemic, it is essential go find a place with low COVID-19 cases. This is why this was included in the analysis.

Regarding the type of restaurant to be open, it should be up to the business owner to make the final decision. If a business owner believes that opening a restaurant of popular type is a better choice, then Japanese restaurant will be the way to go. On the contrary, Mexican restaurant might be considered the best choice for a business owner who believes the opposite.

In summary, the decision framework developed in this project can be applied to other places, so that business owners can use this framework to aid their decision to open new restaurants.

References:

[1] List of neighbourhoods in Toronto, Wikipedia

[2] Cluster Analysis — an overview, ScienceDirect

[3] K-means clustering tutorial, Teknomo Kardi, Medicine 2006