Estimating the Best Location for a Restaurant Business in Toronto

Tenzin Jangchup

February 17, 2021

1. Introduction

The Toronto downtown core has a ton of restaurants of all sizes and cuisine types because of high foot traffic and these locations also have a relatively high cost of operation which comes as no surprise. Individuals who are interested in starting a cost-efficient restaurant business with potentially low cost of operation should ideally take all Toronto neighborhoods into consideration in order to pick the optimal location. Therefore, for this project, the stakeholder/audience is anyone who is looking to find the best location to start a restaurant business. This study will explore neighborhood population, crime rate, the number of existing businesses in the area to determine restaurant market saturation as factors to inform those interested in opening a new restaurant business to choose the best location in Toronto.

In order to have a successful restaurant business in a given area, there are a number of extrinsic factors that we need to take into consideration. For purpose of this project, the top 3 requirement in the order of priority are assumed as following:

1. Less number of existing restaurant businesses
2. Neighborhood population
3. Low crime rate.

For criteria #1, the Foursquare API was used to find the top 10 most common nearby venues. Toronto crime data was also leveraged to further narrow down our search for the neighborhood which fits criteria #2 and #3 best.

The target audience of this study is anyone who is looking to start a restaurant business in the Toronto area and would like to pick a location with a relatively low crime rate so that their business has a better chance of thriving.

2. Data

Toronto Neighbourhoods Boundary File [1] from Kaggle [2] was used as the source for crime data. It included 2014-2019 Crime Data by Neighbourhood and yearly counts are available for Assault, Auto Theft, Break and Enter, Robbery, Theft Over and Homicide. This data also included the neighbourhood population based on the 2016 Census. The population data was used for two things:

1. To normalize the data from yearly crime count to crime per 1000 people
2. To use it as an indicator for restaurant business demand.

From this crime data, features related to different types of crime such as Assault, Auto Theft, Break and Enter, Robbery, Theft Over and Homicide committed in the Toronto area ranging from 2014-2019 were selected.

In addition to the crime data, geopy.geocoders [3] library was used to get the coordinates for each of the neighbourhoods in the crime data. The coordinates were used to explore the Foursquare API to enrich the dataset by finding nearby venue categories for each neighbourhood. Based on Foursquare venue data, the summary of existing restaurants and businesses in each neighbourhood were obtained.

3. Methodology

The crime data from Kaggle contained 140 Toronto neighbourhoods and 60 features. These features included Assault, Auto Theft, Break and Enter, Robbery, Theft Over and Homicide counts for years 2014-2019. In addition, it also included crime rate and change in crime rate since 2014. However, only the raw crime occurrence count, population, and the neighbourhoods were retained for analysis.

After the latitude and longitude for the neighbourhoods were downloaded via geocoders library, preliminary data screening showed that there were 31 neighbourhoods for which the coordinates were unobtainable. Therefore, those neighbourhoods were dropped from the dataset. In the final dataset, there were 109 neighbourhoods and their crime records along with their coordinates.

In order to enrich the dataset with local business data, the Foursquare API was used to explore nearby venues for each neighbourhood and the various business categories that currently exist in the area were appended to the dataset. This data enrichment allowed us to see what sort of businesses are common in the area and found potential market gaps in each neighbourhood. The neighbourhood search radius was limited to 500 and the venue limit to 150. From the Foursquare data, only the Venue Category feature was kept for k-means clustering purposes.

After the data cleaning and feature selection process, there were two dataframes to work with:

1. toronto_venues: Contained Neighbourhood list, their coordinates, and their respective venue categories in a 500m radius.
2. df: Contained Neighbourhood list, Population, counts of crime (Assault, Auto Theft, Break and Enter, Robbery, Theft Over and Homicide) between the years 2014-2019.

Our exploratory data analysis can be broken down to two phases:

1. Analysis on venue category dataframe where the categories were converted into numerical data in order to fit the dataframe to K-means cluster.

2. Analysis on crime dataframe where the result from the cluster was used to analyze crime rate and trend between years 2014-2019.

3.1 Analysis on Venue Category

We utilized one-hot encoding on the venues and grouped them by their neighborhoods so that we can convert the data to numerical value which can then be used for K-means clustering to cluster the dataset based on our selected features. After this exercise, our dataset had 2498 rows with 268 features and contained Neighbourhood names along with their various venue categories. However, we saw that our one-hot encoding process had created a new record for each venue category. So in order to represent this data more comprehensively, we grouped the dataset by Neighbourhood and used the mean value for the venue category.

3.2 Analysis on Population and Crime Trend

We merged Cluster 1 data with Toronto crime data and filtered our the neighbourhoods from other clusters. By looking at the neighbourhood list in cluster 1, our objective was to:

1. Pick the top 5 most populated neighbourhoods
2. Visualize the crime rate to population ratio
3. Visualize the trend of crime rate

We first took the neighbourhoods from cluster 1 and determined the most 5 populated neighbourhoods within that cluster. Fig 3.1 below shows the population distribution among the neighbourhoods in cluster 1. We can see that Rouge has a substantially greater population than the rest of the neighbourhoods. This will be taken into account when crime occurrence count is analyzed later.

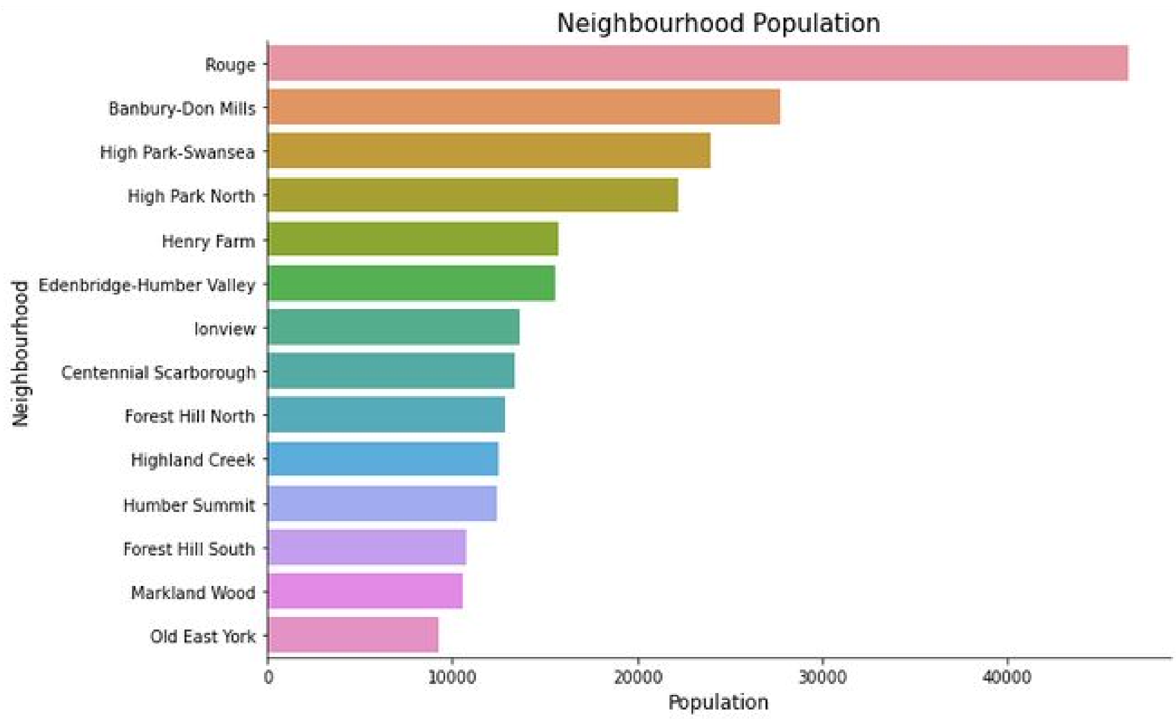

Fig. 3.1 - Population distribution in cluster 1 neighbourhoods

Rouge, Banbury- Don Mills, High Park Swansea, High Park North, and Henry Farm were identified as the top 5 most populated neighbourhoods within cluster 1. These 5 neighbourhoods were then analyzed to visualize the crime trend as shown in Fig. 3.2 below.

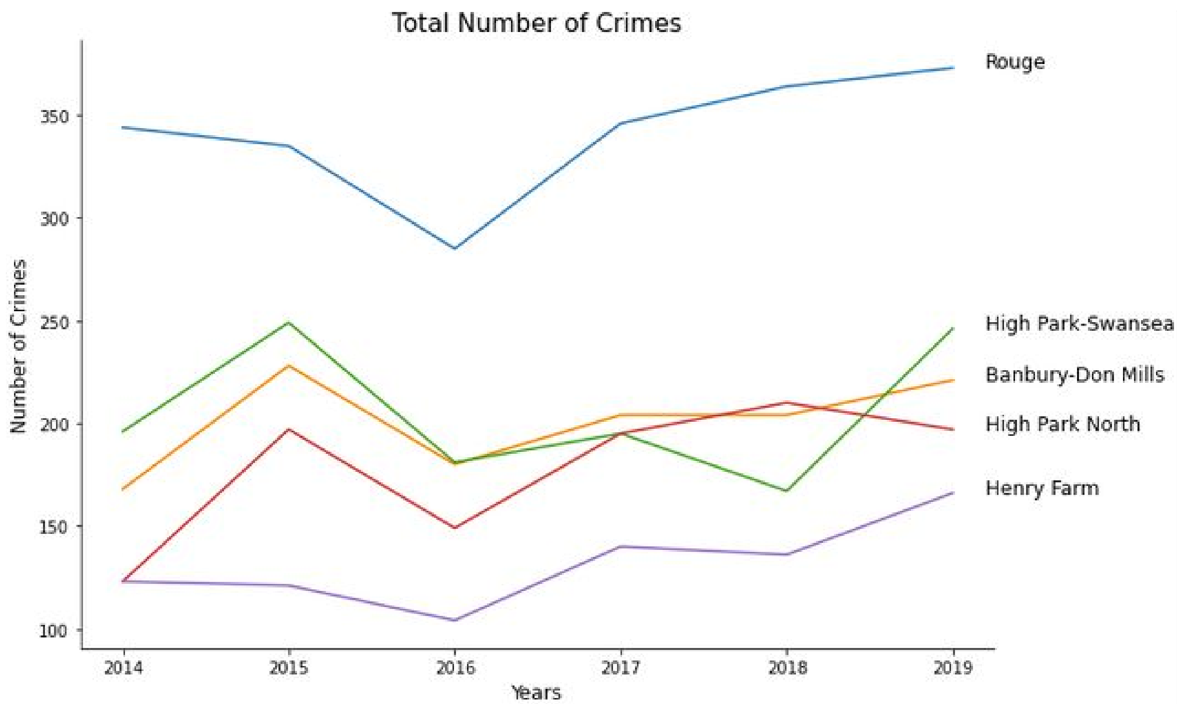

Fig. 3.2 - Crime trend in cluster 1 neighbourhoods

4. Results

4.1 Finding the Best K-Value for K-Means Clustering

We used K-means clustering to segregate each row based on their data points. This allowed us to visualize the market gap in terms of what businesses are currently available in the neighbourhood. The elbow graph shown below shows that K value of 3 is where the change in distance score diminishes. The elbow graph method indicates that the K value of 3 is the ideal choice. However, this method is only a good starting point to find our best K. So we will experiment with the K value of 3 and 4 to see which one presents use with a more distinct dataset.

Using inertia value from the sklearn library, we can easily find the sum of distances of samples to their closest cluster center. We can see the inertia value plot against its K-value to find the best K in Fig 3.1 below.

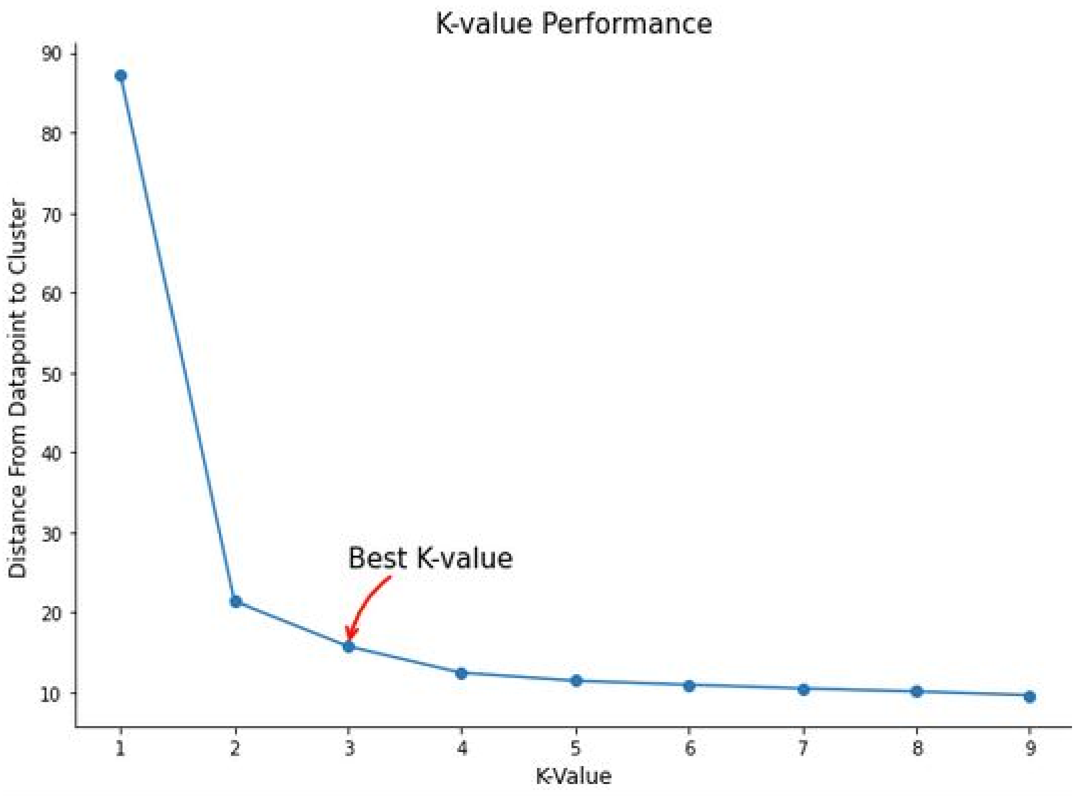

Fig. 4.1 - Line plot of inertia value for different K-Values

4.2 Cluster Visualization

Since we have 268 features, it is not possible to visualize each feature on the cluster. So we will use the PCA module to estimate the principal component from our features and reduce our feature down to 2 and visualize how the clusters are distributed. We combined our 3 dataframes created above and also selected the features that will be helpful to visualize our data as follows. These 3 dataframes can be joined by using Neighborhood as index.

- df (Selected features: Neighbourhood, Population, Latitude, Longitude)

- neighborhoods_venue_sorted (Selected features: top 10 most common venue categories)
- df_group (Selected features: Neighborhood, Cluster Labels)

We then used the combined dataframe for mapping our cluster to visualize the geographical location and their cluster category. The image below (Fig. 3.2) shows that the data points have clearly been clustered well, into 4 distinct clusters despite a small number of outliers.

Fig. 4.2 - Cluster visualization using principal component analysis

4.3 Cluster Examination

We examined the four clusters closely to determine the distinctions between them.
As we've outlined in the beginning our criteria for opening a restaurant business, we want a neighbourhood that fits the following 3 requirements:

1. Less number of existing restaurant businesses
2. Neighborhood population
3. Low crime rate

From this observation, neighbourhoods in Cluster 1 provides strong support for our requirements and seems to be a good starting place to further explore. The map (Fig. 4.1) plots the clusters on a geographical map to show each location in the data set. The colour indicates which cluster the neighbourhood belongs to. The red markers represent cluster 0, greens show cluster 1, blue indicate cluster 2, and black represents cluster 3.

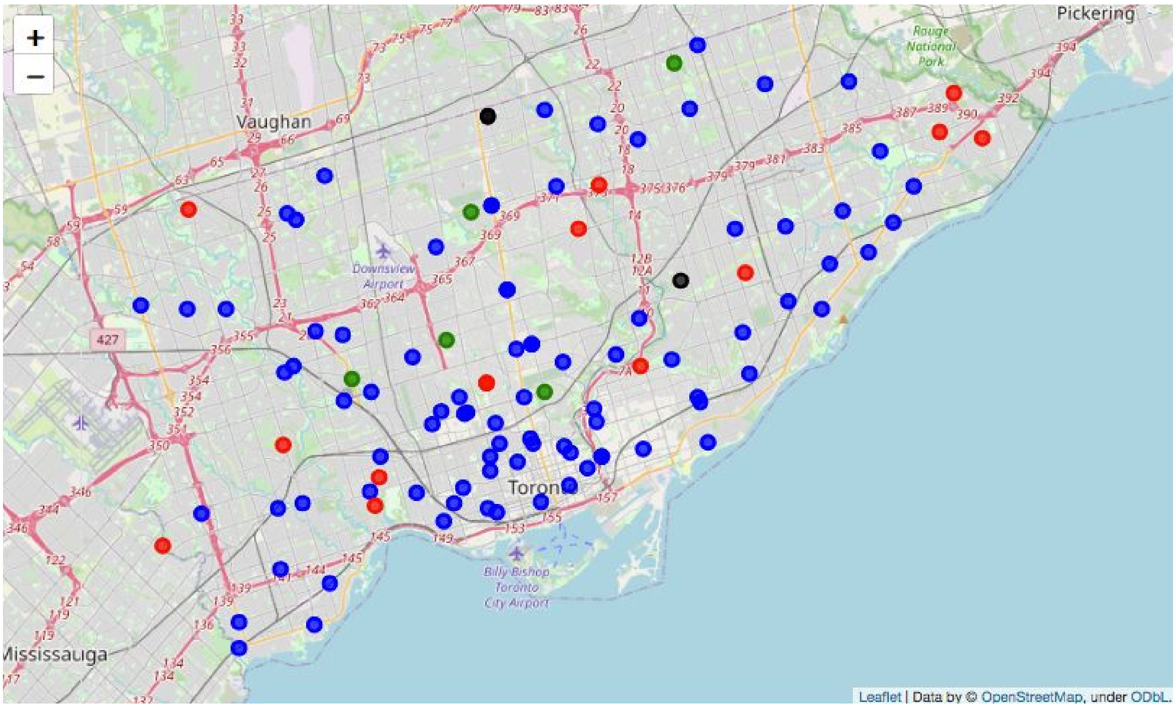

Fig. 4.3 - Geographical plot of neighbourhoods color coded per cluster

Cluster 0, indicated as red markers in the map below, includes a variety of restaurants, particularly eastern and middle eastern cuisine, as the top 10 most common venues. Since these neighbourhoods are oversaturated with restaurant businesses, opening a new restaurant venue in these neighbourhoods would be risky, particularly for those interested in starting an eastern or middle eastern restaurant.

Neighbourhood	Population	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
Victoria Village	17510	Thai Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Event Space	Dumpling Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Elementary School	Yoga Studio
Newtonbrook West	23831	Korean Restaurant	Greek Restaurant	Thai Restaurant	Middle Eastern Restaurant	Cosmetics Shop	Dog Run	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market
Newtonbrook East	16097	Korean Restaurant	Greek Restaurant	Thai Restaurant	Middle Eastern Restaurant	Cosmetics Shop	Dog Run	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market

Fig. 4.4 - Cluster 0 Datasets

Cluster 1, highlighted as green markers in the geographical map below, have parks and recreational activity businesses as the dominant venues in that neighbourhood. Restaurant businesses rank from 6-10th most common venues in these areas therefore, cluster 1 meets requirement 1 (fewer existing restaurant businesses). A market gap exists providing the best opportunity to establish a new restaurant venue in these neighbourhoods.

Neighbourhood	Population	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
Humber Summit	12416	Construction & Landscaping	Gift Shop	Bakery	Park	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Factory	Event Space
Centennial Scarborough	13362	Park	Yoga Studio	Elementary School	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Event Space
Ionview	13641	Construction & Landscaping	Deli / Bodega	Metro Station	Park	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Factory
Forest Hill North	12806	Playground	Bank	Park	Mediterranean Restaurant	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant
Old East York	9233	Park	Intersection	Plaza	Pastry Shop	Liquor Store	Pub	Elementary School	Donut Shop	Dumpling Restaurant
Forest Hill South	10732	Playground	Bank	Park	Mediterranean Restaurant	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant
Henry Farm	15723	Intersection	Park	Restaurant	Tennis Court	Farmers Market	Falafel Restaurant	Factory	Event Space	Fast Food Restaurant
Edenbridge-Humber Valley	15535	Park	Fast Food Restaurant	Bus Stop	Yoga Studio	Dumpling Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Elementary School
High Park North	22162	Park	Convenience Store	Mattress Store	Baseball Field	Gym / Fitness Center	Tennis Court	Event Space	Eastern European Restaurant	Egyptian Restaurant

Fig. 4.5 - Cluster 1 Datasets

Cluster 2, shown as blue markers, includes mostly fast food restaurants and coffee shops. These businesses rank from the 2nd-5th most common venues in these neighbourhoods. Opening a restaurant in these areas would also be risky. Since there are fewer dine in restaurants in these areas, these neighborhoods may get less foot traffic compared to other neighbourhoods.

Neighbourhood	Population	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
Lansing-Westgate	16164	Playground	IT Services	Health & Beauty Service	Event Space	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Elementary School
Rosedale-Moore Park	20923	Playground	Park	Gym	Tennis Court	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Factory	Event Space	Elementary School
Englemount-Lawrence	22372	Metro Station	Coffee Shop	Bakery	Shopping Mall	Playground	Tennis Court	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant
Black Creek	21737	Playground	Construction & Landscaping	Café	Coffee Shop	Event Space	Dumpling Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Elementary School
Steeles	24623	Playground	Dive Bar	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Factory	Event Space	Elementary School

Fig. 4.6 - Cluster 2 Datasets

Cluster 3, shown as black markers in the map below, has a mix of different business types with pizza places and restaurants being the most common venues in these areas. Opening a new restaurant in these areas may be risky.

Neighbourhood	Population	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
Yonge-St.Clair	12528	Coffee Shop	Grocery Store	Italian Restaurant	Sushi Restaurant	Café	Thai Restaurant	Bagel Shop	Pizza Place	Bank	Sandwich Place
York University Heights	27593	Pizza Place	Discount Store	Fast Food Restaurant	Grocery Store	Gas Station	Coffee Shop	Caribbean Restaurant	Sandwich Place	Falafel Restaurant	Beer Store
Yorkdale-Glen Park	14804	Bank	Argentinian Restaurant	Gas Station	Clothing Store	Fast Food Restaurant	Italian Restaurant	Coffee Shop	Paintball Field	Print Shop	Bike Shop
Stonegate-Queensway	25051	Burrito Place	Bank	Yoga Studio	Asian Restaurant	Shopping Mall	Mattress Store	Middle Eastern Restaurant	Liquor Store	Sushi Restaurant	Eastern European Restaurant
The Beaches	21567	Beach	Pizza Place	Breakfast Spot	Japanese Restaurant	Bar	Restaurant	Nail Salon	Park	Pub	Skating Rink
...	...	...	...	...	...	...	...	...	...	...	...
Little Portugal	15559	Bar	Café	Coffee Shop	Cocktail Bar	Restaurant	Bakery	Korean Restaurant	Health & Beauty Service	French Restaurant	Japanese Restaurant
Milliken	26572	Chinese Restaurant	Japanese Restaurant	Bakery	Pet Store	Noodle House	Asian Restaurant	Intersection	Juice Bar	Bookstore	Furniture / Home Store
Pleasant View	15818	Breakfast Spot	Burger Joint	Bank	Bakery	Fast Food Restaurant	Thrift / Vintage Store	Japanese Restaurant	Italian Restaurant	Shopping Mall	Pizza Place
Wychwood	14349	Ice Cream Shop	Restaurant	Italian Restaurant	Coffee Shop	Sushi Restaurant	Bakery	Pizza Place	Indian Restaurant	Café	Juice Bar
Leaside-Bennington	16828	Bank	Bakery	Grocery Store	Indian Restaurant	Breakfast Spot	Kids Store	Restaurant	Bar	Bagel Shop	Boutique

Fig. 4.7 - Cluster 3 Datasets

4.4 Crime Rate Examination

In order to provide a more realistic view of the crime trend, we accounted for the population differences among the neighbourhoods. So we calculated the crime per 1000 population and observed the result shown below in Fig 4.6.

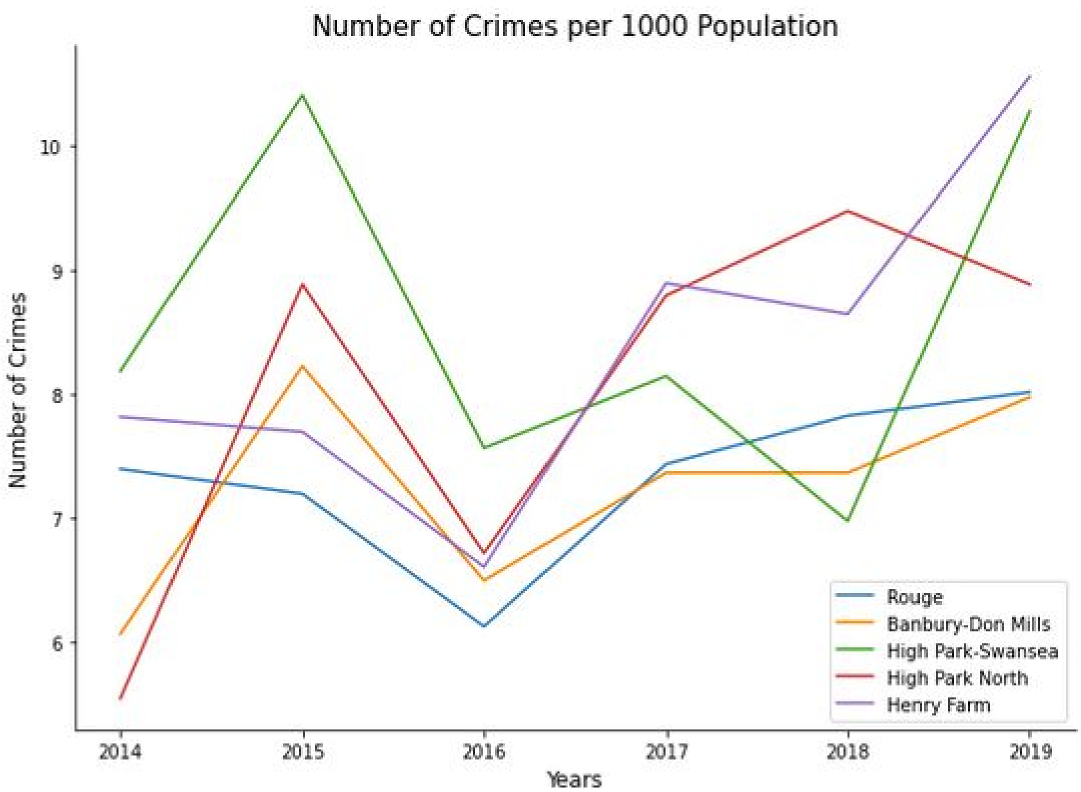

Fig. 4.8 - Crime per 1000 population in cluster 1 neighbourhoods

This graph indicates that Henry Farm neighbourhood crime trend is on the rise while Rouge and Banbury has a fairly stable crime trend. We see some minor fluctuation in the crime reports in the High Park-Swansea neighbourhood.

5. Discussion

From our results, we observed that the majority of neighbourhoods follow a similar trend in terms of popularity of the business categories. The neighbourhoods in cluster 0 had mostly Pizza place, Fast food restaurant and coffee shops as top 5 while in cluster 1, we see restaurants of different cuisines as the most common business. And in cluster 2, parks and recreational businesses are the top 5 common categories followed by restaurants of various cuisines.

From this observation, neighbourhoods in Cluster 1 provide strong support for our requirements. To narrow down further, based on our findings, the ideal neighbourhood to start a restaurant business where there is a good balance of neighbourhood population and crime rate could be either Rouge or Banbury since they both have the highest population in the cluster and also lowest crime rate. In addition, these 2 neighbourhoods also don't have many existing restaurant businesses which eliminates a lot of competition.

It is also interesting to dive further into cluster 2. There are a large number of elementary schools in the areas saturated with fast food businesses and coffee shops. It would be interesting to see whether or not dine in family owned restaurants would succeed in these areas. In addition, it could be worth looking into other businesses that would succeed alongside middle eastern and eastern restaurants in cluster 0. For instance, perhaps opening an ethnic supplies store in this area would be beneficial, or perhaps it would be a great neighbourhood to host a cultural street festival. It might also be interesting to see affluence through house-hold income as a factor whether there's a reason cluster 3 has a large number of fast-food restaurants in those areas.

6. Conclusion

Given the diverse cultural background of the Toronto population, we could see the contrasting features of the community reflect in the various clusters. This study explored the most common businesses in Toronto neighbourhoods and also its crime data in order to resolve a few options of neighbourhood that could be presented to a potential client who might want to start a restaurant business in Toronto. As a result, with the evidence and visual support provided above, the Rouge or Banbury neighbourhood stands out as an ideal choice to start a restaurant business.

7. References

[1]: https://www.kaggle.com/alincijov/toronto-crime-rate-per-neighbourhood
[2]: https://www.kaggle.com/
[3]: https://pypi.org/project/geocoder/