Opening an Italian Restaurant in Toronto

1. Introduction

1.1 Background

Toronto is the capital city of Ontario, Canada and has a recorded population of 2,731,571 in 2016. It is the most populous city in Canada and the fourth most populous city in North America. Toronto is home to many people from different cultures and many ethnicities. This diversity is reflected in Toronto's ethnic neighborhoods, which include Chinatown, Corso Italia, Greektown, Kensington Market, Koreatown, Little India, Little Italy, Little Jamaica, Little Portugal and Roncesvalles (Polish community).

Toronto is an international center for business and finance. It is also a cultural center with major Universities, and many arts and sports activities. This vibrant city provides many opportunities for young and enthusiastic individuals to open their own business and become a piece of this multidimensional city.

1.2 Problem

Coming from a Greek/Italian family and with a long history in the hospitality sector, opening a modern Italian restaurant in Toronto is an investment worth trying. However, even though Toronto has many opportunities to offer and it is a prosperous environment to open a business, the competition is rising day by day. Experts in the hospitality sector claim that **Location** is the number one factor in the success of a restaurant.

The main problem that this paper is going to examine is: " what would be the ideal location to open an Italian restaurant in Toronto?". Along with this question, several other subsequent questions about the demographics of the population are going to answered.

1.3 Interest

The results of this investigation will be used in building a business plan that will be used to attract investors. One example is the bank from which funds will be requested to make the initial investment. In order to receive the loan, a detailed plan and estimations need to be presented.

2. Data

2.1 Data sources

The Data that will be used in this research are extracted from the 2016 Canada Census : https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/Page.cfm?Lang=E&Geo1=CSD&Code1=3520005&Geo2=PR&Data=Count&B1=All

Several demographic features will be presented such as: Population, age, income, family status and home location (Neighborhood).

In order to determine which is the ideal location to open an Italian restaurant, the different neighborhoods of Toronto are going to be examined. The data about the neighborhoods of Toronto are from Wikipedia:

After gathering the location data and creating a dataframe, using the python **geopy** library and the python **folium,** the different neighborhoods of Toronto will be shown in a map.

Then, The location of the top 100 venues in a radius of 5000 meters in the neighborhoods of Toronto will be extracted using **Foursquare API**. More specifically, the Longitude, Latitude, Category Label, and Type of Venue will be gathered using the **GetNearbyVenues**. To analyze each neighborhood, we will use the **One hot encoding** function and then a new dataframe with the Italian restaurants in each neighborhood will be created.

Finally, the we will **cluster** the neighborhoods based on their similarities. The **best K-value** will be identified and used to determine the number of clusters. By comparing the amount of restaurants in each cluster, we will determine which cluster of neighborhoods has the less competition in Italian restaurants and also explore the demographics of that cluster.

2.2 Data Cleaning

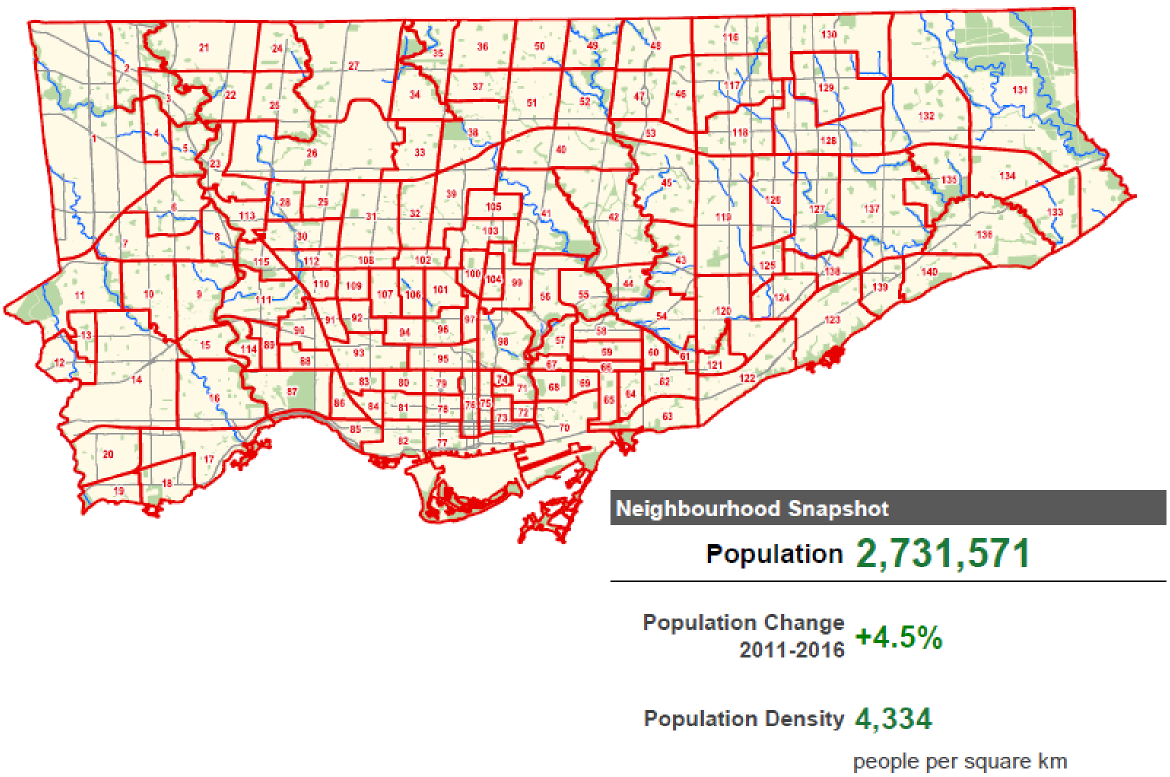
The data will be scrapped by the websites using BeautifulSoup. Unnecessary columns must be deleted and NaN values should be dropped. The postal codes and the location data must be merged in one table.

To create a new dataframe, the rows that are labeled as" Italian restaurant" need to sorted. Also, to visualize the demographics of the selected cluster, a new table needs to made based on Neighborhood ID. This table merges the cluster information from Foursquare API and the data collected for the 2016 Canada Census.

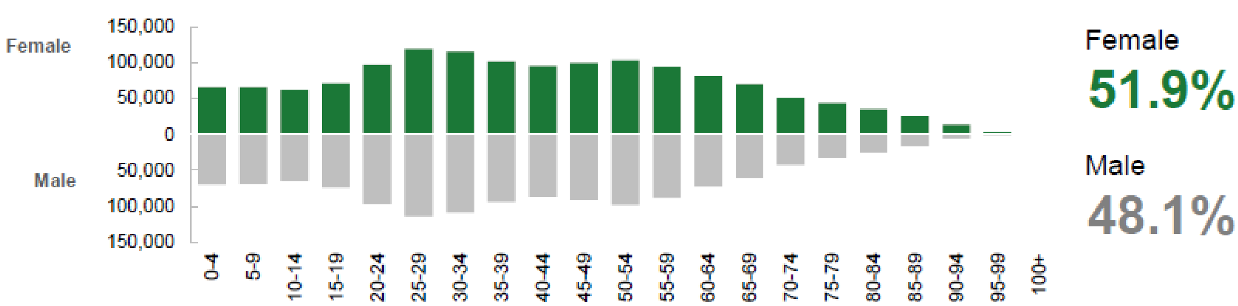
3. Exploratory Data Analysis

A) Toronto Demographics:

1. Population



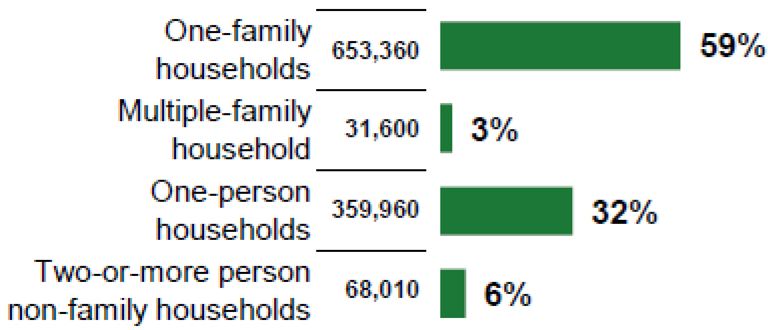
2. Age groups



- **Children 0-14 years : 398,135 (15%)**
- **Youth 15-24 years : 340,270 (12%)**
- **Working Age 25-54 years : 1,229,555 (45%)**
- **Pre-Retirement 55-64 years : 336,670 (12%)**
- **Seniors 65+ years : 426,945 (16%)**

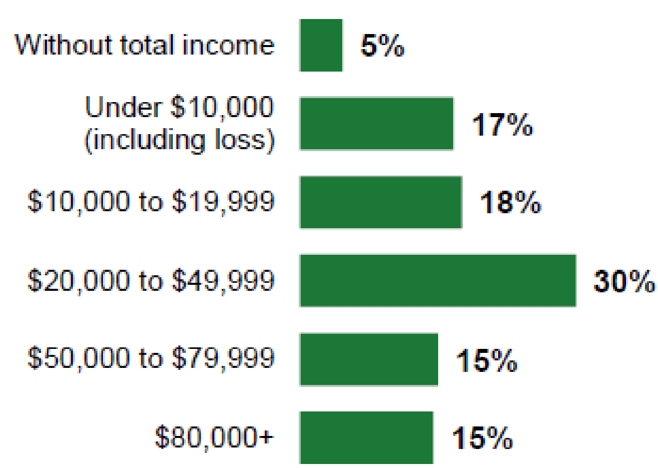
Toronto's highest age group concentration is detected in the working age. This is a good indicator for the city's cultural and financial status since more young people are living and working in this city.

3. Household types



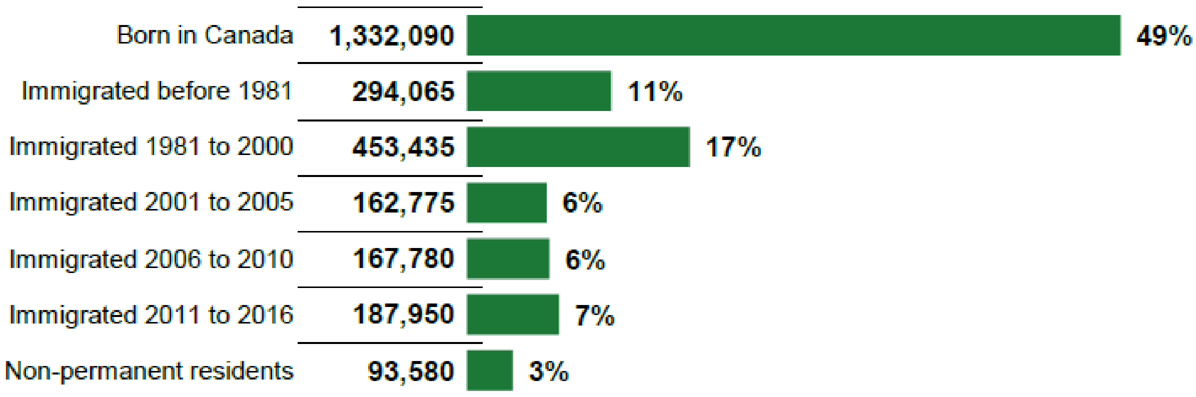
4. Total individual income

Income	Toronto
Median household income	$65,829
Median family income	$82,859
Median FY/FT work income	$55,246
Without income	4.7%
Income from gov't transfers	9.3%
Poverty (MBM)	21.9%
Low income (LIM-AT)	20.2%
Low income (LICO-AT)	17.4%

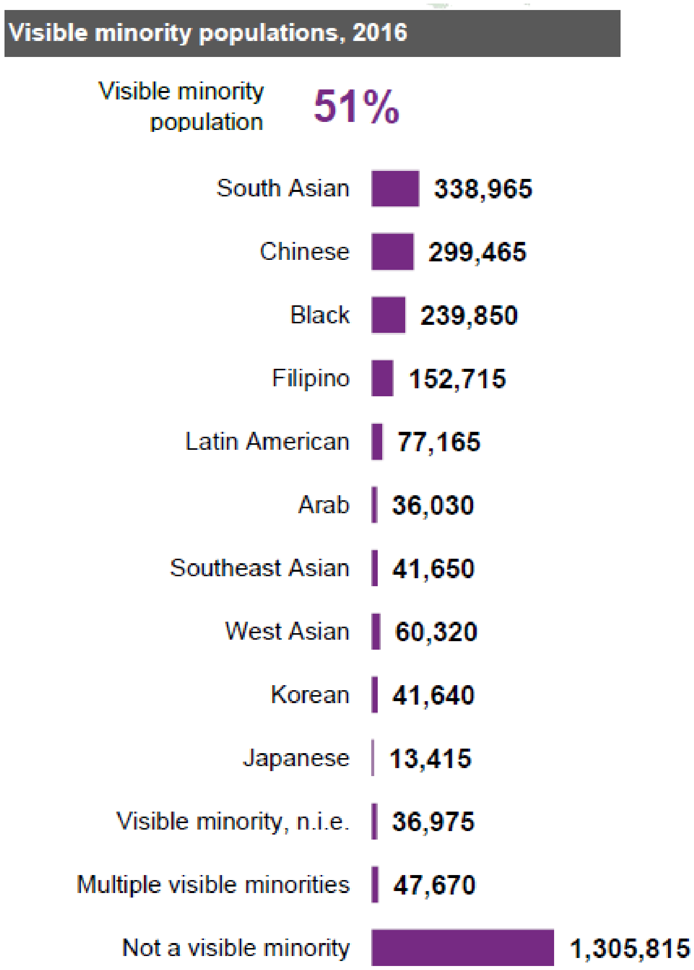


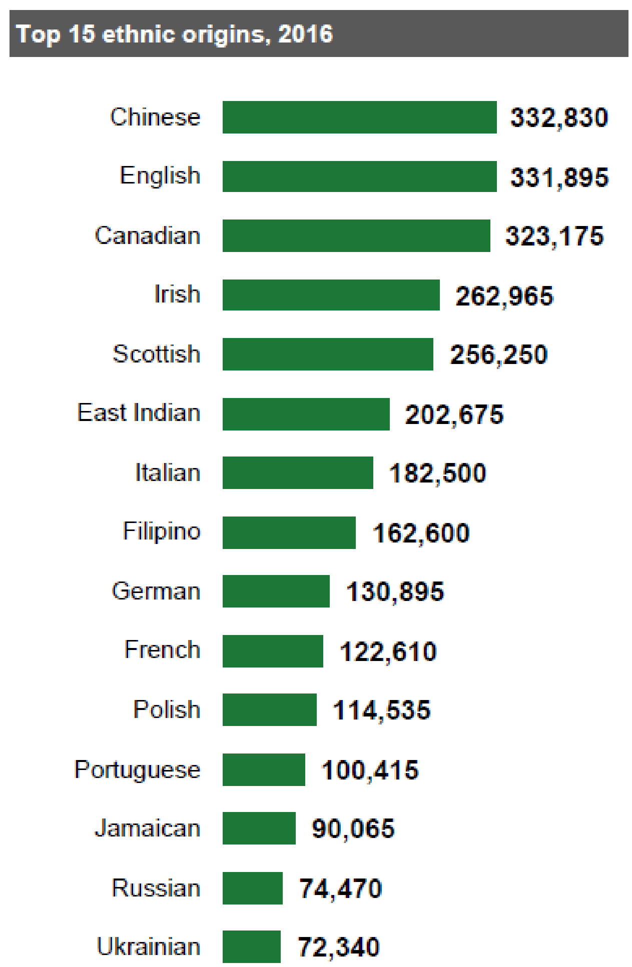
5. **Ethno-cultural diversity**

Immigration status and period of immigration:



Top 15 ethnic origins:

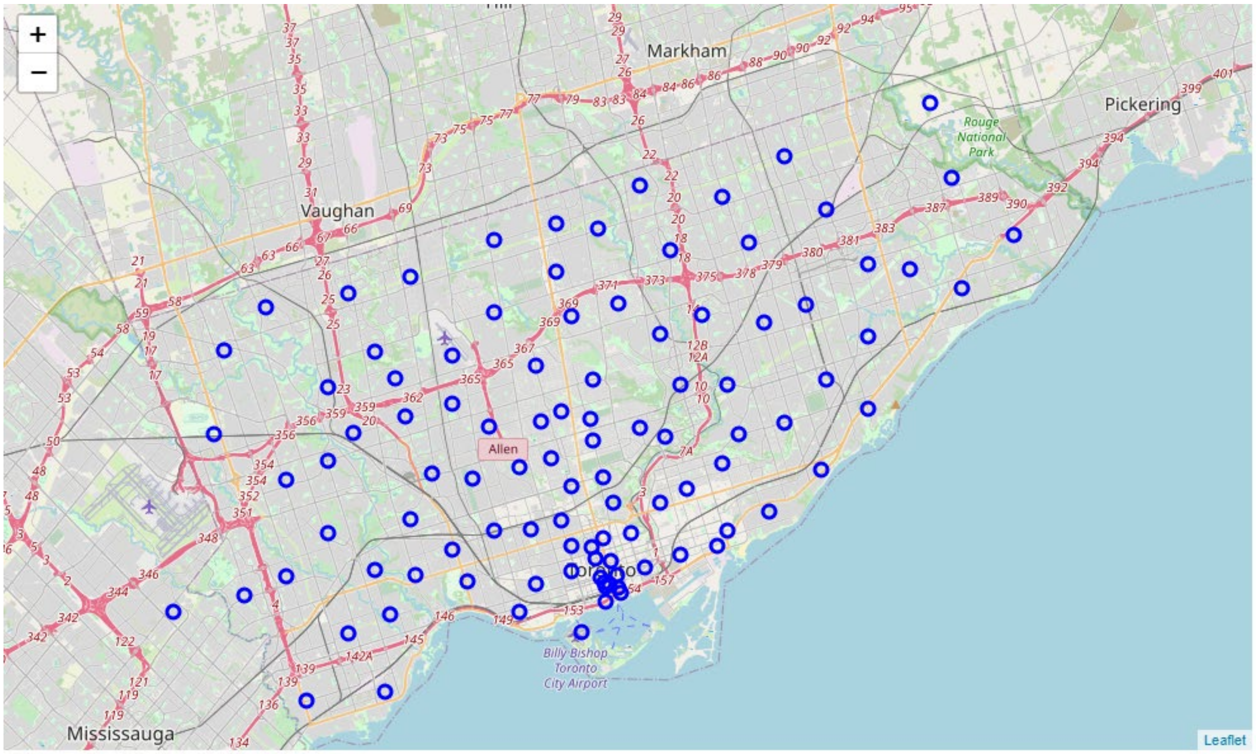




Note: "n.i.e." = not included elsewhere; "n.o.s." = not otherwise specificed

B) Toronto Neighborhoods

Below is presented the map of the Neighborhoods of Toronto using the python folium and the data for the corresponding postal codes.



Toronto is consisted by 103 Neighborhoods.

Using the Foursquare API we collected the location information of the top 100 venues in Toronto in a radius of 5000 meters.

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

```
(2153, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

2153 venues were discovered and their Venue Category is stated in the graph above.

Using the **GetNearBy** function, we identified 277 unique categories. With the **one hot encoding** function the categories were labeled and group by taking the mean of the frequency of occurrence of each category.

Group rows by neighborhood and by taking the mean of the frequency of occurrence of each category

```
toronto_grouped = toronto_onehot.groupby('Neighborhood').mean().reset_index()
toronto_grouped
```

	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquariu
0	Agincourt	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
1	Alderwood, Long Branch	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
3	Bayview Village	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
4	Bedford Park, Lawrence Manor East	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.040000	0.000000	0.00
5	Berczy Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
6	Birch Cliff,	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00

A new dataframe was created for the venues labeled as "Italian restaurants"

```
toronto_italian = toronto_grouped[["Neighborhood","Italian Restaurant"]]
```

```
toronto_italian.head(100)
```

	Neighborhood	Italian Restaurant
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.000000
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.120000
5	Berczy Park	0.017857
6	Birch Cliff, Cliffside West	0.000000
7	Brockton, Parkdale Village, Exhibition Place	0.041667
8	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
9	Caledonia-Fairbanks	0.000000
10	Cedarbrae	0.000000
11	Central Bay Street	0.042857

4. Data Modeling and results

1. Clustering

After getting the venue categories and creating a new dataframe for the Italian restaurants in order to check the competition in the different neighborhoods, the clustering of the neighborhoods follows.

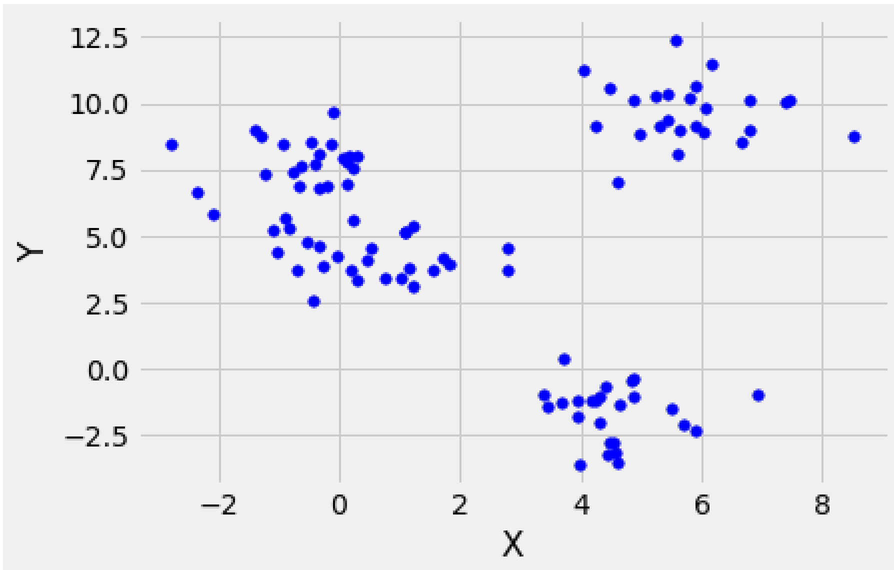
The clustering will divide the data into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.

Initially we set the **K value equal to 5**, meaning that we create 5 clusters. Each neighborhood is assigned to a cluster depending on the similarities sharing with the neighborhoods of the same cluster. Below are the first rows of the clustered neighborhoods.

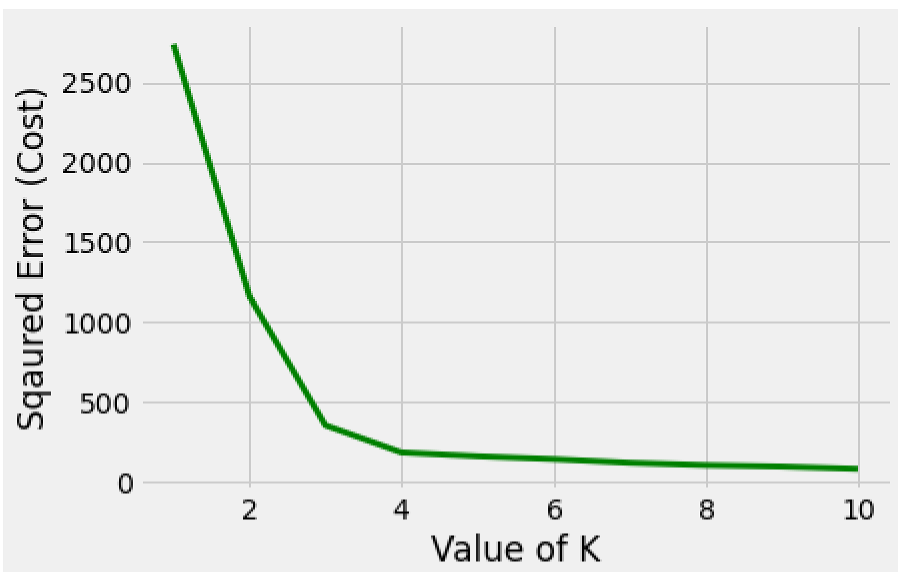
	Neighborhood	Italian Restaurant	Cluster Labels
0	Agincourt	0.000000	0
1	Alderwood, Long Branch	0.000000	0
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000	0
3	Bayview Village	0.000000	0
4	Bedford Park, Lawrence Manor East	0.083333	4
5	Berczy Park	0.017544	2
6	Birch Cliff, Cliffside West	0.000000	0
7	Brockton, Parkdale Village, Exhibition Place	0.043478	3
8	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0
9	Caledonia-Fairbanks	0.000000	0
10	Cedarbrae	0.000000	0
11	Central Bay Street	0.044118	3
12	Christie	0.066667	4
13	Church and Wellesley	0.000000	0

2. **Finding the best K-value**

We use the samples generator to create **sample points** around c center randomly chosen.



Then we calculate the **cost,** meaning the **squared error** for the clustered points.



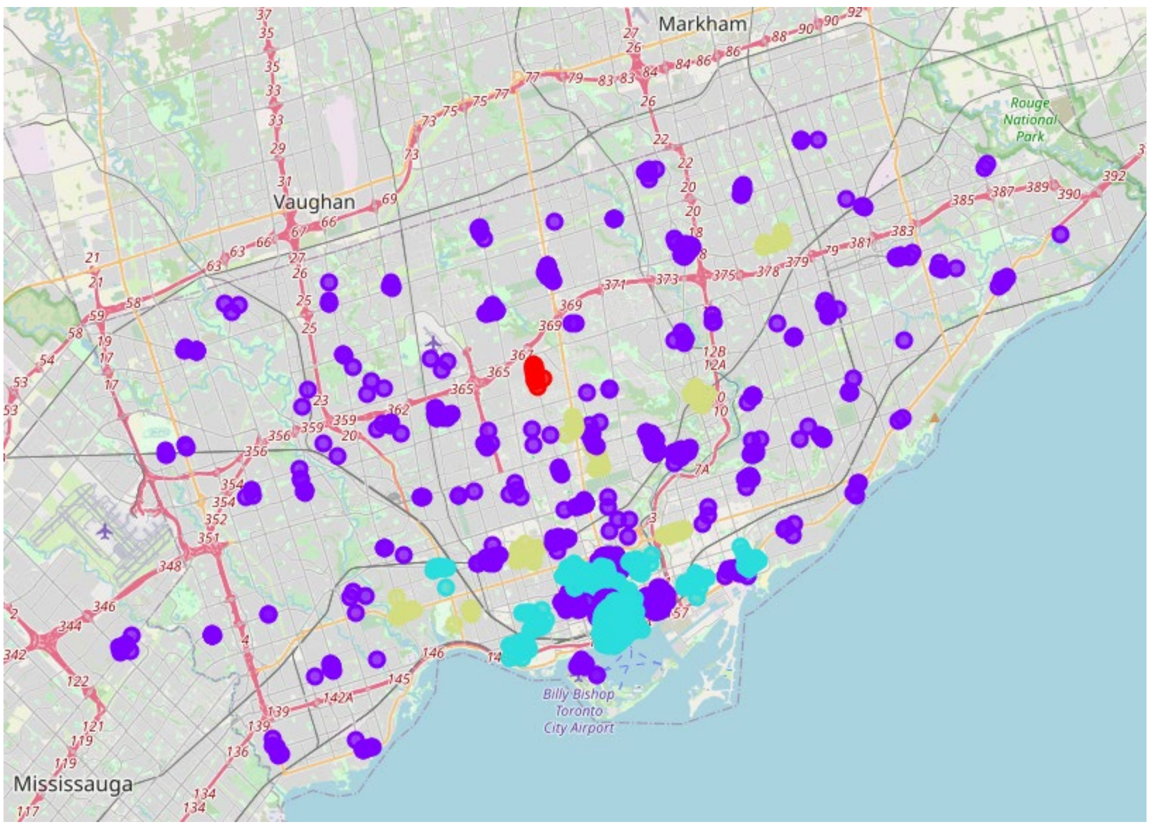
As shown in the graph above, the **best K-value is equal to 4** because from that point and after we have the least error.

3. Clustering with K=4

We run exactly the same analysis with the optimum K-value and the neighborhoods are again assigned to a cluster.

	Neighborhood	Italian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agincourt	0.0	1	43.794200	-79.262029	Panagio's Breakfast & Lunch	43.792370	-79.260203	Breakfast Spot
0	Agincourt	0.0	1	43.794200	-79.262029	El Pulgarcito	43.792648	-79.259208	Latin American Restaurant
0	Agincourt	0.0	1	43.794200	-79.262029	Twilight	43.791999	-79.258584	Lounge
0	Agincourt	0.0	1	43.794200	-79.262029	Commander Arena	43.794867	-79.267989	Skating Rink
1	Alderwood, Long Branch	0.0	1	43.602414	-79.543484	Il Paesano Pizzeria & Restaurant	43.601280	-79.545028	Pizza Place

The 4 cluster are shown in the map below.



4. Cluster examination

We examine each cluster separately in order to find the number of venues in each cluster.

cluster 0 ¶

```
toronto_italian_merged.loc[toronto_italian_merged['Cluster Labels'] == 0]
```

	Neighborhood	Italian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
4	Bedford Park, Lawrence Manor East	0.12	0	43.733283	-79.41975	Aroma Espresso Bar	43.735975	-79.420391	Café
4	Bedford Park, Lawrence Manor East	0.12	0	43.733283	-79.41975	Pheasant & Firkin	43.735173	-79.419702	Pub
4	Bedford Park, Lawrence Manor East	0.12	0	43.733283	-79.41975	Drums N Flats	43.735035	-79.420040	Comfort Food Restaurant

cluster 1

```
toronto_italian_merged.loc[toronto_italian_merged['Cluster Labels'] == 1]
```

	Neighborhood	Italian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agincourt	0.00	1	43.794200	-79.262029	Panagio's Breakfast & Lunch	43.792370	-79.260203	Breakfast Spot
0	Agincourt	0.00	1	43.794200	-79.262029	El Pulgarcito	43.792648	-79.259208	Latin American Restaurant

cluster 2

```
toronto_italian_merged.loc[toronto_italian_merged['Cluster Labels'] == 2]
```

	Neighborhood	Italian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
5	Berczy Park	0.017857	2	43.644771	-79.373306	LCBO	43.642944	-79.372440	Liquor Store
5	Berczy Park	0.017857	2	43.644771	-79.373306	The Keg Steakhouse + Bar - Esplanade	43.646712	-79.374768	Restaurant
									Vegetarian /

cluster 3

```
toronto_italian_merged.loc[toronto_italian_merged['Cluster Labels'] == 3]
```

	Neighborhood	Italian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
12	Christie	0.062500	3	43.669542	-79.422564	Fiesta Farms	43.668471	-79.420485	Grocery Store
12	Christie	0.062500	3	43.669542	-79.422564	Contra Cafe	43.669107	-79.426105	Café

Total venues in neighborhoods in clusters:

- Total venues in cluster 0 = 945
- Total venues in cluster 1 = 396
- Total venues in cluster 2 = 28
- Total venues in cluster 3 = 689

The first cluster (cluster 0) has the highest concentration of venues with 43,9 % of the total amount of restaurants. The second cluster has a medium concentration of 18,4%, while the third one has the lowest concentration of venues with only 1,3%. The fourth one has a high concentration of 32%.

Using the data from the 2016 Canada Census we import more data for the Toronto Neighborhoods such as the Total population, age, average income, gender and the neighborhood ID

	Neighbourhood	Neighbourhood Id	Combined Indicators	Total Population	Average Family Income	Pop - Males	Pop - Females	Pop 15 - 64 years
0	West Humber-Clairville	1	NaN	33312	72820	16625	16690	23285
1	Mount Olive-Silverstone-Jamestown	2	NaN	32954	57411	16070	16890	22300
2	Thistletown-Beaumond Heights	3	NaN	10360	70838	5055	5300	6760
3	Rexdale-Kipling	4	NaN	10529	69367	5130	5395	7165
4	Elms-Old Rexdale	5	NaN	9456	61196	4520	4935	6370

We drop the NaN values and merge the table with the one from the dataframe used for clustering. Below is the table with the new dataframe.

```
neighborhoods_1.tail()
```

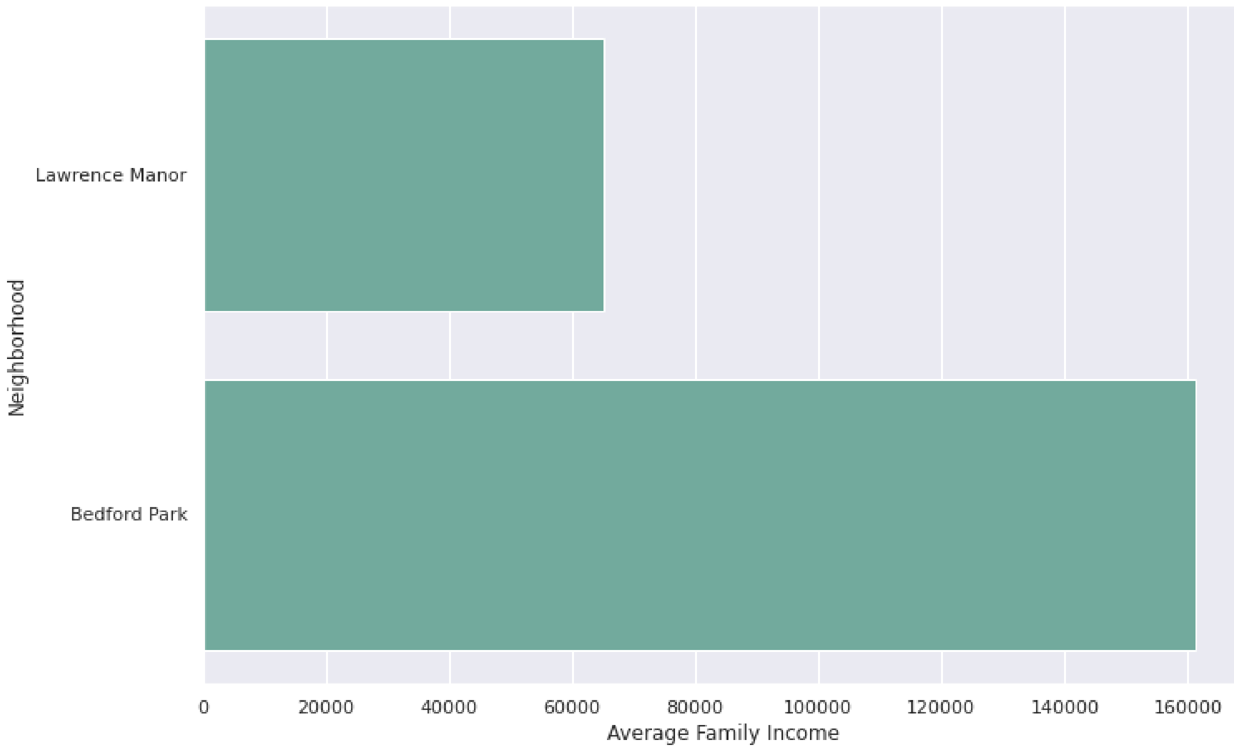
4]:

	Neighborhood	Italian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
98	York Mills West	0.0	0	43.752758	-79.400049	Tournament Park	43.751257	-79.399717	Park
98	York Mills West	0.0	0	43.752758	-79.400049	Kitchen Food Fair	43.751298	-79.401393	Convenience Store
98	York Mills West	0.0	0	43.752758	-79.400049	iRemodel Commercial Construction	43.750808	-79.402356	Construction & Landscaping
98	York Mills West	0.0	0	43.752758	-79.400049	416-Flowers, Order & Send Flowers Online	43.748405	-79.399588	Flower Shop
99	York Mills, Silver Hills	0.0	0	43.757490	-79.374714	Vyner Greenbelt	43.759642	-79.369590	Park

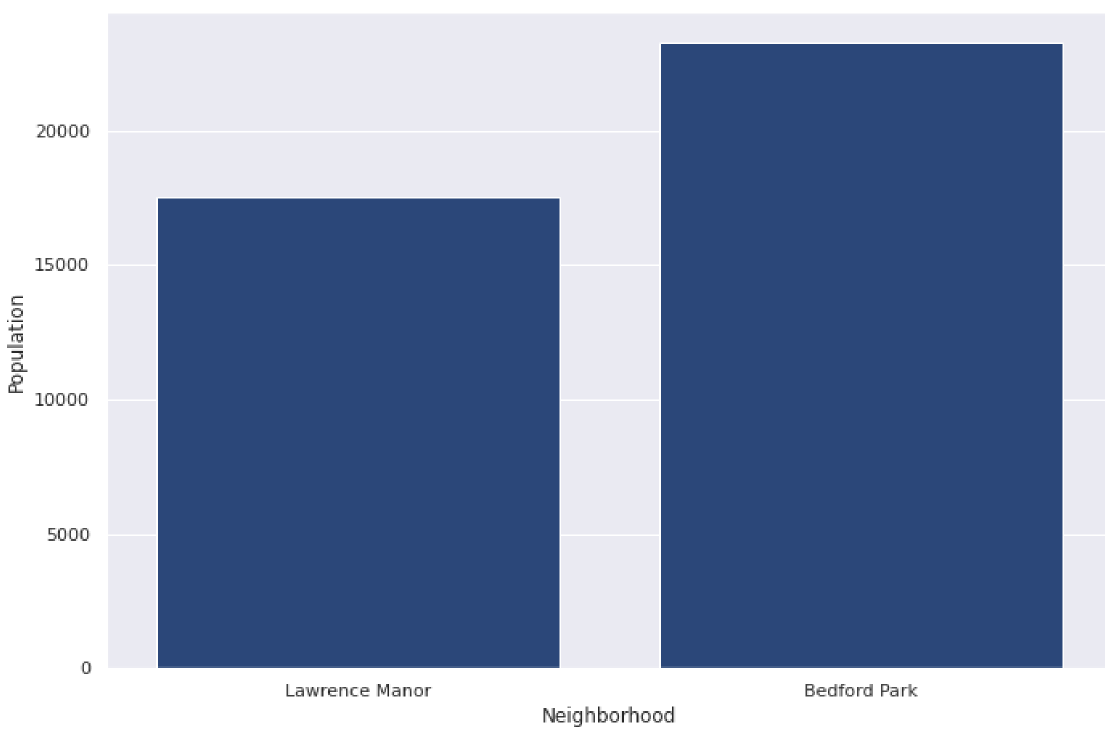
We sort the dataframe for the neighborhoods in the third cluster and we end up with 2 neighborhoods, The **Lawrence Manor** and the **Bedford Park.**

	Neighborhood	Neighbourhood Id	Population	Average Family Income	Males	Females	Age 15 - 64
0	Lawrence Manor	43	17510	65104	8110	9405	11475
1	Bedford Park	39	23236	161110	10845	12390	14700

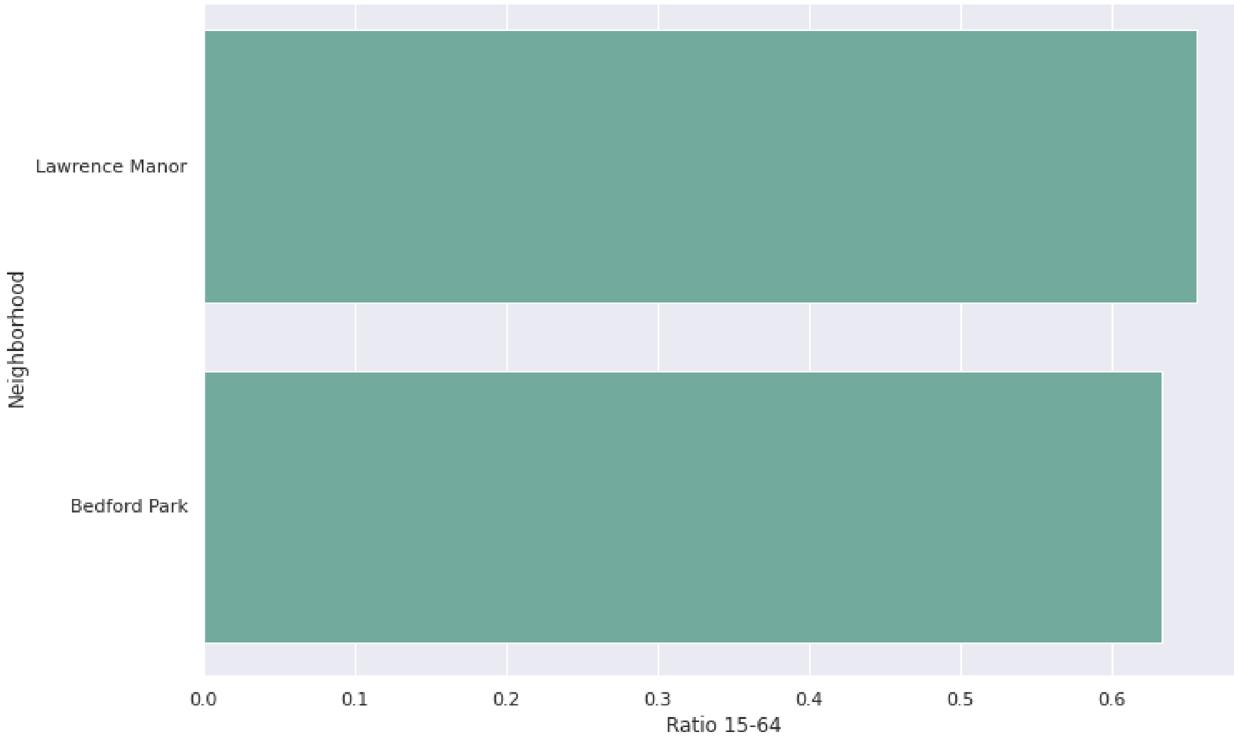
Income comparison for the 2 neighborhoods



Population comparison



Age 15-64 years as percentage of the total population



The last three graphs compare the income, total population and working age population between the 2 neighborhoods in the cluster. It is clear that **Bedford Park** 's Average Family income is more than double compared to the one of Lawrence Manor. The population of Bedford Park is also significantly higher (24.6%) that the one at Lawrence Manor. As for the 15-64 years population, the two neighborhoods share similar characteristics.

5. Conclusions

Based on the cluster analysis it is advise to open an Italian restaurant at the third cluster that is consisted by the neighborhoods of Lawrence Manor and Bedford. This cluster has only 28 restaurants in these neighborhoods, thus less competition. Both neighborhoods are located outside the big city center of Toronto.

As for the financial status of these neighborhoods, the Bedford has significantly higher income that the Lawrence Manor's and its population is also higher. Bedford has significantly higher Average Family income ($ 161,100) compared to the Toronto's Median Family income ($65,829). Therefore, Bedford is a neighborhood that you can open an Italian restaurant with less risk.

6. Future discussions

This research was conducted based on the data provided by the 2016 Canada Census. The Foursquare API was useful to identify the location of the venues and their category, however there are other several factors that influence the choice of location. For example, depending on the concept of the restaurant it will attract different kinds of people. Even though we selected the third cluster as the less competitive group of neighborhoods, it might be the case that the purpose of the restaurant is to attract tourists. Therefore, another cluster might be more appropriate for that.

Additionally, the choice of location is depending on the budget. Several locations have higher rent prices and the investment is risky. This research didn’t include the venue rent prices. For future research it would be useful to investigate how this factor influences the choice of the venue location.