Location Analytics for Finding the Best Location for a New Indian Restaurant in Toronto

Capstone Project - The Battle of Neighborhoods

IBM Data Science Professional Certificate

by

Usman Ahmed

musman.ahmed@hotmail.com

September 27th, 2018

TABLE OF CONTENTS

ABSTRACT

Toronto is a host to large number of immigrants who settle there and explore various job and business opportunities. Among various business opportunities, many of them consider investing in new traditional restaurant. Those who decide to do so, have one important question in mind: what is the best location to start a new exotic restaurant? In this project, we apply data science techniques, to answer this question for an Indian/Pakistani restaurant. We collect data from various public sources, clean it, and apply various statistical and machine learning techniques and come up with the best candidate for hosting such restaurant.

1. INTRODUCTION

Canada is a popular immigrants' destination and over the years people from all around the world have chosen it as their country of residence. One of the most important decisions they have to make while they are planning to move to Canada is to choose the city where they would like to live. The statistics available on various web-sites and forums suggest that the most preferable destination of these new immigrants is Greater Toronto Area. The rationale behind this preference is basically the comparative good weather and available opportunities for starting a new business.

The people immigrating to Canada come from different social, geographical and economic backgrounds. By nationality, Indians make a large part of these new immigrants. Apart from getting a job, these Indian nationals always look forward to start their own business. Among various business options for Indians, restaurants serving traditional Indian/Pakistani food is the most one. As there are many Indians and Pakistani immigrants living in Greater Toronto Area (GTA) who would like to savor the food of their countries of origin, this option seems economically feasible.
The next stage after having an economic feasible option for the business is to decide about the location of the restaurant. From location perspective, there are various factors that contribute to the success of a restaurant. These include the presence of potential customer base, presence of competitors, crime rate in the area, foot traffic, accessibility by public transport or availability of the parking slots, visibility of the restaurant, conditions of the lease agreement and safety conditions of the building.

The field of data science provides the tools and ability to analyze such factors and facilitate the perspective restaurant owners to find out the best location for opening the new restaurant. In this project, we perform spatial analysis through the visualization of the mentioned factors, correlation analysis and cluster analysis of various factors. We analyze the clusters and come up with the best neighborhood in the GTA for opening new restaurant that targets middle aged, and middle class Indian/Pakistani customers.

The rest of the report is structured as follows:

- The section 2 describes the methodology including the data collection, preparation and analysis.
- The section 3 presents the results of the analyses.
- In section 4, discussion on the observation of the results, recommendation and limitations is presented.
- The section 5 concludes the report while outlining some important further works.

2. METHODOLOGY

In this section, we describe our methodology including the phases of data collection, data preparation, spatial visualization, correlation analysis, model building and cluster analysis.

A. Data Collection and Preparation

In order to identify the most suitable neighborhood for hosting a new restaurant, we identified the following factors potentially affecting the decision to choose the area for opening a new restaurant.

- Presence of customer base
    - Total population of the area
    - Population change in recent years
    - Population density
    - Population of target customers according to age
    - Population of target customers according to the ethnicity (as the restaurant would offer traditional Indian/Pakistani food)
    - Population change of immigrants in recent years
- Presence of competitors
- Crime rates in area
- Accessibility
    - Access to public transport
    - Availability of parking slots in the area
    - Traffic flow in the area
- Safety
    - Traffic accidents
    - Pedestrian accidents

The next task was to collect the relevant data and prepare it for the analysis. In the following, we describe our data collection and preparation methodology.

A.1. Data Collection

The list of data sets and sources used in this project is presented in the following.

i. First of all, we need to have the **list of neighborhoods** in the GTA and their spatial boundaries. This data is available through Toronto City's Open Data Catalogue in form of shapefile.

ii. The data regarding indicators for identifying the presence of customers base is also available in the form of **neighborhood profiles** through Toronto City's Open Data Catalogue. This data set is available in CSV and contains various social and economic indicators of each neighborhood.

iii. The **presence of competitors**, i.e. Indian/Pakistani restaurants, in the area was collected using **Foursquare API**.

iv. The Toronto Police Open Data Catalogue offers the data regarding **crime rates** by neighborhood so this data was downloaded from their website. The data is available in shapefile.

v. **Access to public transport and transport safety** data was downloaded from Toronto City's Open Data Catalogue in the CSV format.

vi. Location data of **parking slots** in the city of Toronto is also available on Toronto City's Open Data Catalogue in JSON format.

A.2. Data Preparation

The collected data sets are very rich in data and record various important indicators. However, all of them are not important for our project. In the following we briefly describe each of these data sets, processing carried out for transformation and the extracted features.

i. The **neighborhoods** data contains the boundaries of each neighborhood with its official id and name. The data set is available in shapefile. We use all the features available in our project.

ii. The **neighborhood profiles** data set use census data to provide a portrait of the demographic, social and economic characteristics of the people and households in each City of Toronto. The data set records hundreds of indicators in CSV format. For simplicity, we cleaned the file to remove unnecessary features before using the data set in our project. We extracted the following indicators from this data set for use in our study.

   - Population
   - Rate of Population change
   - Population density
   - Population from 14 years – 60 years
   - Population of Indian/Pakistani origin people
   - Rate of immigrants population change
   - Average income of households

iii. The **Foursquare API** was used to query for the presence of Indian/Pakistani restaurants in each neighborhood. We recorded the number of restaurants returned for each query as the number of competitors available in the area.

iv. The data set available through the Toronto City's Police Open Data platform records counts of major crimes by neighborhood. The counts are available for Assault, Auto Theft, Break and Enter, Robbery, Theft Over and Homicide. The data also includes four year averages and crime rates per 100,000 people by neighborhood based on 2016 Census Population. We extracted the overall **crime rate** for each neighborhood for the use in our project.

v. The data set used for the **access to public transport and safety** records following indicators for each neighborhood: number of TTC (public transport for Toronto) stops, number of crowded routes for TTC, number of pedestrian collisions, number of traffic collisions, length or road in kilometers and roads volume. In our project, we use all indicators except road length and volume.

vi. The data sets for **parking slots** records location of parking space in the GTA with other details such as parking slots, timings, fees, rates, payment options, etc. For our project, we just use the location and parking slots. For each parking space we query the location to find the neighborhood it is situated. The sum of parking slots grouped by neighborhoods determine the parking slots available in each neighborhood.

B. Data Visualization and Analysis

In the following, we discuss the analysis performed to understand the data and to find out the most suitable candidate for hosting a new Indian/Pakistani restaurant.

B.1. Spatial Visualization

Once the data was ready for the analysis, we decided to visualize the spatial distribution of each feature. For the purpose, we used choropleth maps which helped us understand how various features are distributed across neighborhoods in the GTA. To be able to create and visualize the choropleth maps, the neighborhoods shapefile was converted to JSON format.

B.2. Correlation Analysis

After understanding the spatial distribution of features, it is important to investigate if there are any clear determinants for deciding the location of a restaurant. For this purpose, we decided to calculate the correlation between restaurants count and other features. For simplicity, correlation matrix was used to calculate and visualize the results of correlation analysis.

B.3. Clustering and Segmentation

The next step was to perform clustering of neighborhoods based on the selected features.

Before, we could proceed to clustering, we needed to normalize the features as the magnitudes of the features were very different and without normalization, the resulting clusters would be biased in favor of some features. Therefore, we used standard scalar normalization.

The clustering was performed on the normalized data set and for this purpose, we used **kmeans** algorithm to create 5 clusters. The returned clusters were visualized using the choropleth map to understand the distribution of clusters geographically. The clusters were also examined based on their centroids to segment the neighborhoods based on the characteristics of the features.

The segment with the best combination of the following characteristics represents the most suitable candidate for hosting the new restaurant.

3. RESULTS

In this section, we present the results of our analysis.

A. Spatial Visualization of Features

In the following, the results of the visualization representing the spatial distribution of various feature is presented. The features are visualized using choropleth maps.

A.1. Total Population

Figure 3-1 displays the distribution of population in City of Toronto's neighbourhoods. The population would contribute favourably towards opening a new restaurant. We see that the most populous neighbourhood are located in north-eastern and southern part of the city.

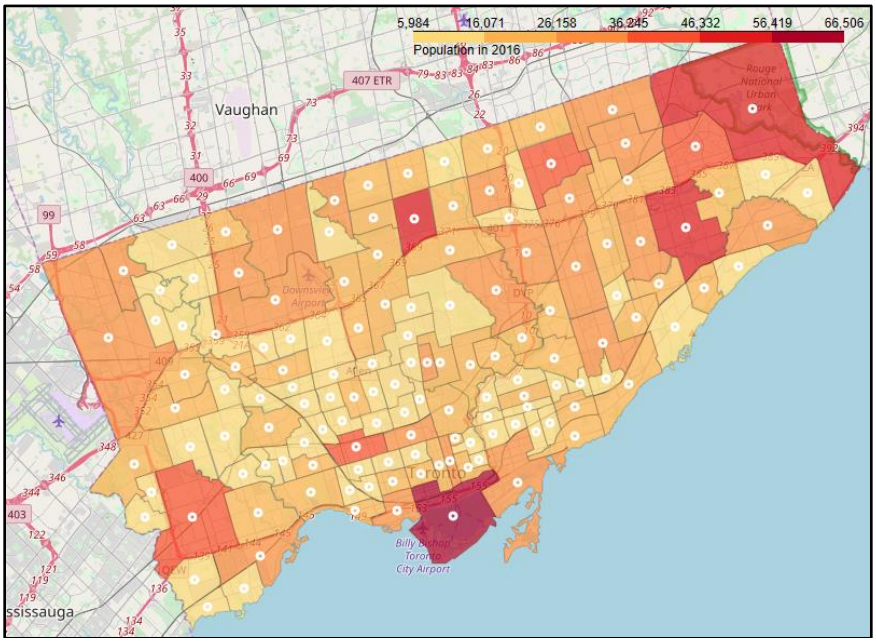

Figure 3-1: Total population of Toronto City by neighborhood in 2016

A.2. Population Change

Figure 3-2 shows the spatial distribution of rate of population change in last five years. We can observe that the highest rate of population increase is towards the southern part of the city. The increase in population in recent years mean that there would be more potential customers in recent future.

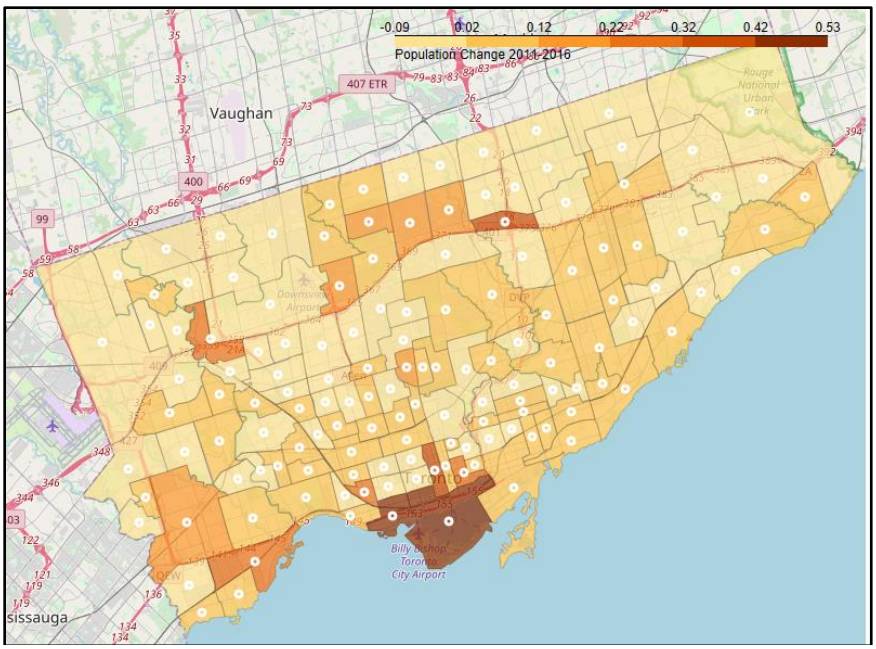

Figure 3-2: Change in Toronto City population from 2011 to 2016

A.3. Population Density

Population density in the city of Toronto is also concentrated towards the southern part as evident from the Figure 3-3.

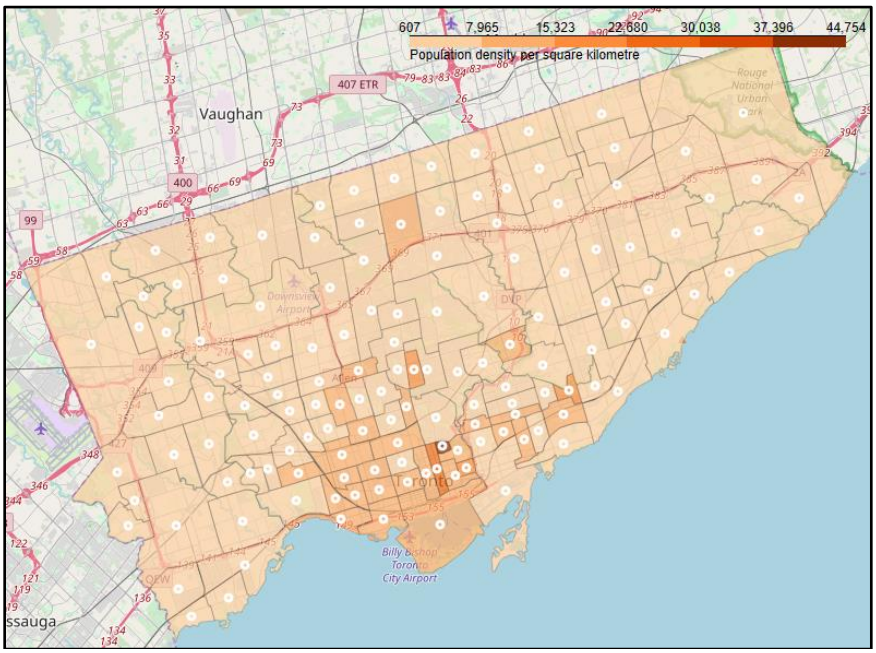

Figure 3-3: Population density of Toronto City

A.4. Average Income

Average income is also expected to play a positive role towards the success of restaurant as the people who are good economically visit the restaurants more often. In Figure 3-4, we see that the people living in central neighbourhoods are better economically. The southern and southern part also host reasonable population of middle classed households.

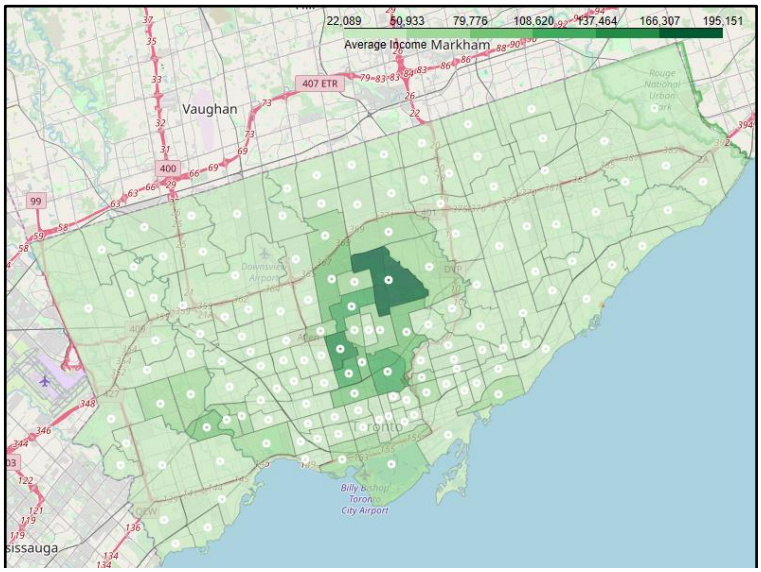

Figure 3-4: Average income of households by neighborhoods

A.5. Presence of Customer Base According to Age

The population of age-based potential customers' base is in high numbers towards the extremes of the city. The southern coastal part ranks high in this criterion too as shown in Figure 3-5.

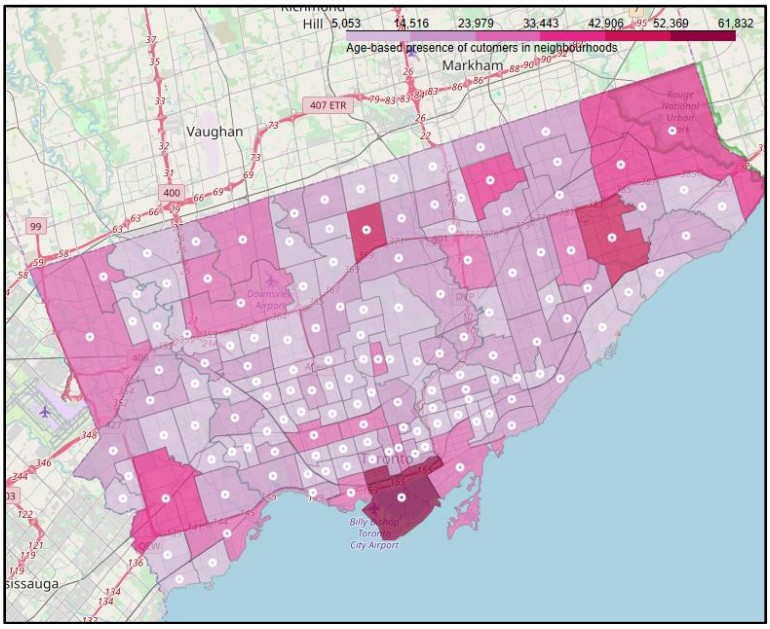

Figure 3-5: Population of 14 to 60 years old people in Toronto City

A.6. Presence of Customer Base According to Ethnicity

Figure 3-6 presents the spatial distribution of Indian/Pakistani origin's immigrants in the City of Toronto. We see the highest number of such population eastern and western extremes.

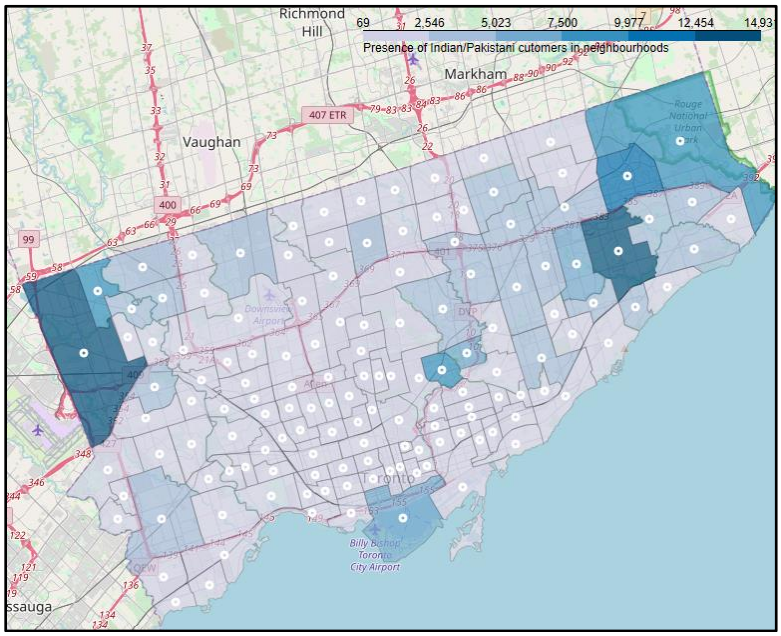

Figure 3-6: Population of Indian/Pakistani origin people in Toronto

A.7. Immigrants Population Change

The increase in immigrants' population in a neighbourhood would contribute positively towards the success of a restaurant. Figure 3-7 depicts that the southern and central neighbourhoods rank better in this regard.

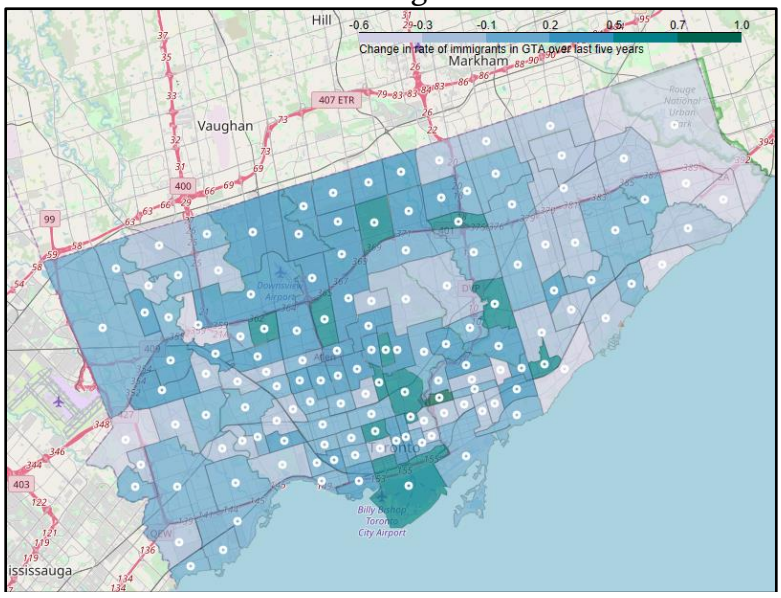

Figure 3-7: Change in immigrants population in last 5 years

A.8. Crimes Rate

Crimes are one of the main deterrents for people to visit any area, therefore higher the crime rate in area, less the probability of success of any business. Figure 3-8 presents the spatial distribution of crime rates in Toronto City. We see that the crime rate is highest in some of the central neighbourhoods followed by north-western region. The southern part ranks in the middle in this regard.

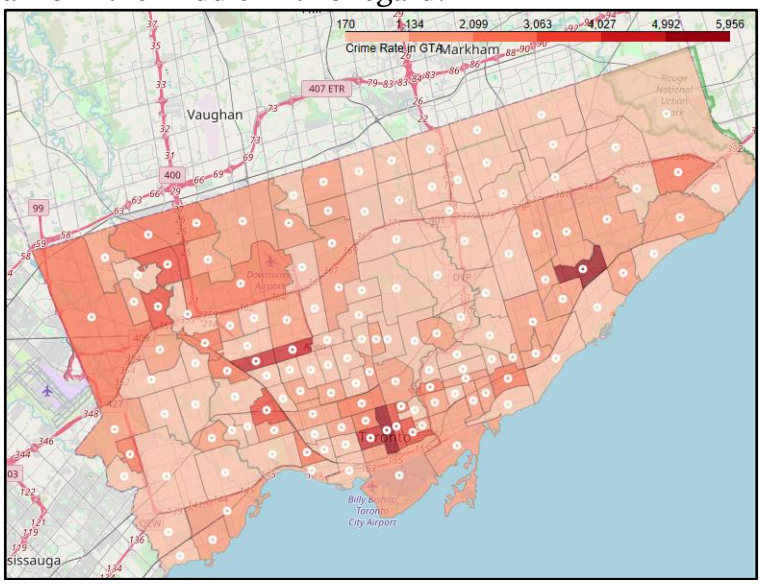

Figure 3-8: Crime rates by neighborhood in Toronto

A.9. Access to Public Transport

We measure the access to public transport by the number of TTC stops in the neighbourhood and display the distribution in Figure 3-9. The eastern and western extremes rank higher according to this criterion. The southern part ranks in the middle.

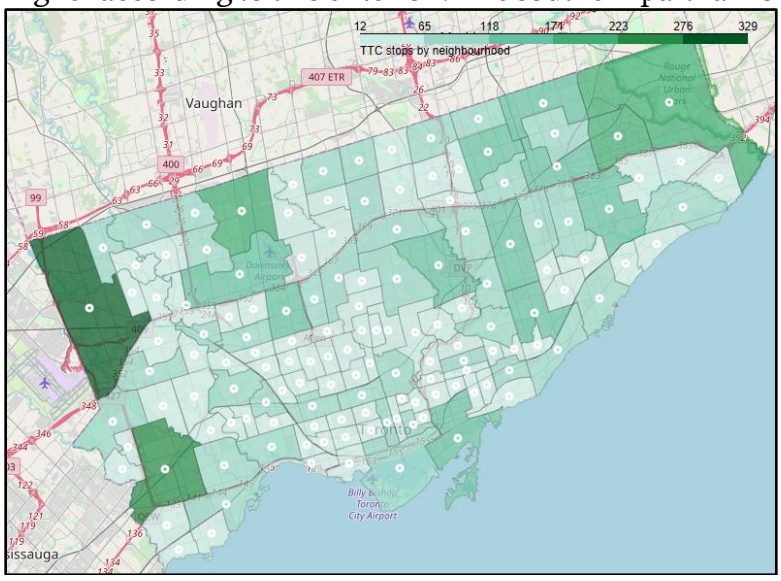

Figure 3-9: Number of TTC stops in each neighborhood

A.10. Crowded TTC Routes

Despite the availability of higher number of TTC stops, we observe, from Figure 3-10, that the public transport is quite crowded in eastern and western extremes. This would have a negative impact on the success of restaurant as the visitor would avoid those routes.

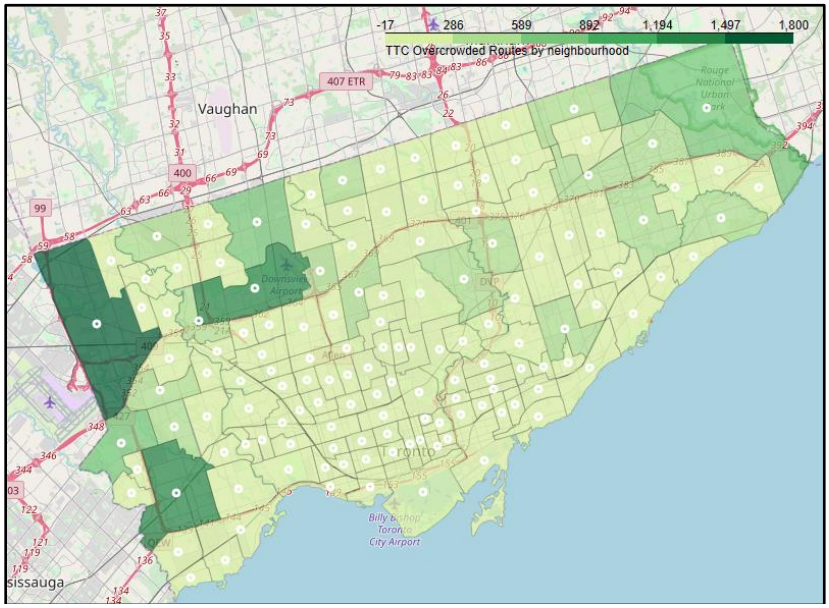

Figure 3-10: Crowded public transport rates by neighborhood

A.11. Pedestrian Collisions

Figure 3-11 presents the geographic distribution of pedestrian collisions in the City of Toronto. Pedestrian collisions may prove an important deterrent for the visitors as they might not feel safe in that area.

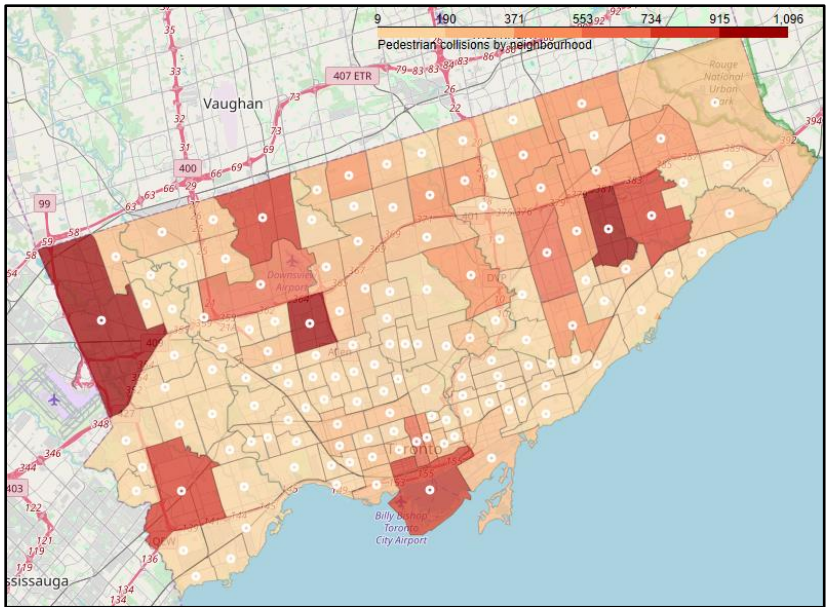

Figure 3-11: Number of pedestrian collisions in City of Toronto

A.12. Traffic Collisions

The geographic distribution of traffic collisions can be visualized in Figure 3-12. Like pedestrian collisions, the traffic collisions are also a risk for public safety and would play a negative role in people's decision to visit the neighbourhood with higher traffic collisions.

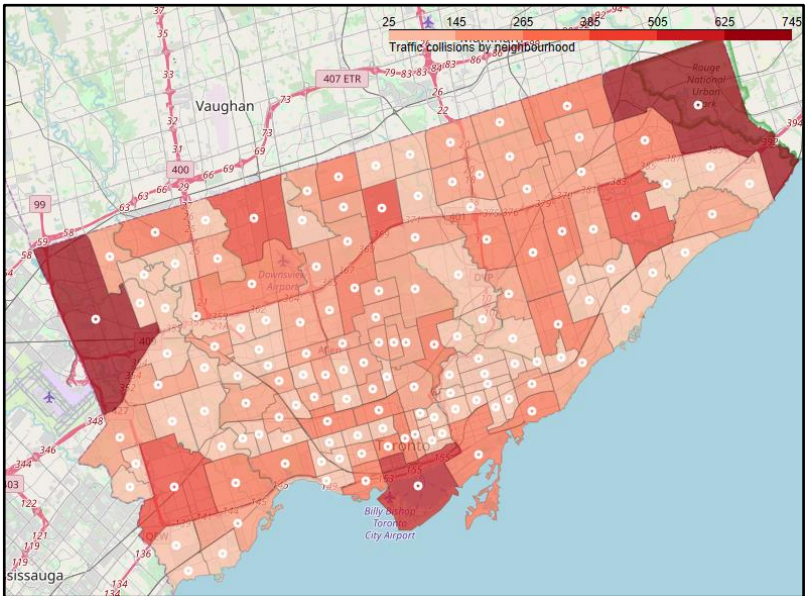

Figure 3-12: Number of traffic collisions in Toronto City

A.13. Parking Spaces

Availability of parking near the restaurant is one of the important factors in customers' choice to visit a particular place. According to Figure 3-13, southern and south-western neighbourhoods rank better in this regard.

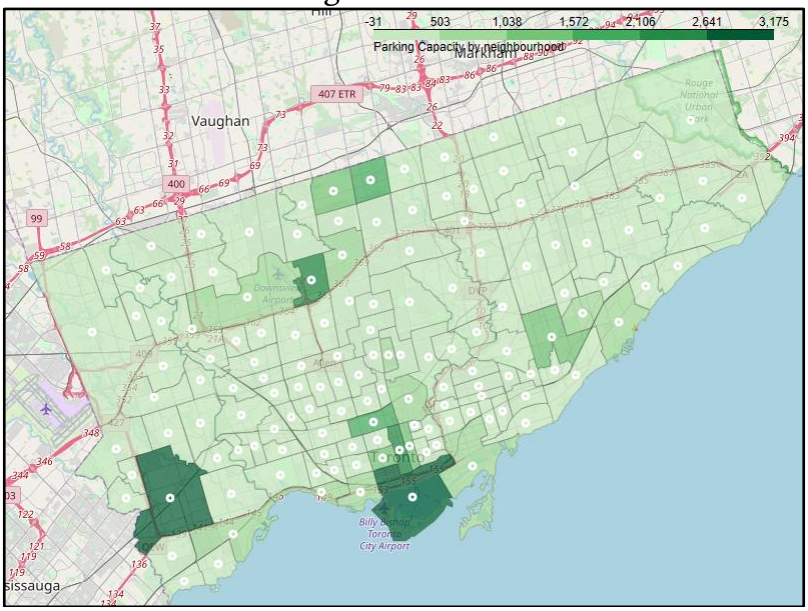

Figure 3-13: Availability of parking spaces in each neighborhood

A.14. Presence of Competitors

Apart from the factors discussed above, an important factor in success of a restaurant would be the presence of competitors near-by. Figure 3-14 displays the distribution of Indian/Pakistani restaurants in each neighbourhood.

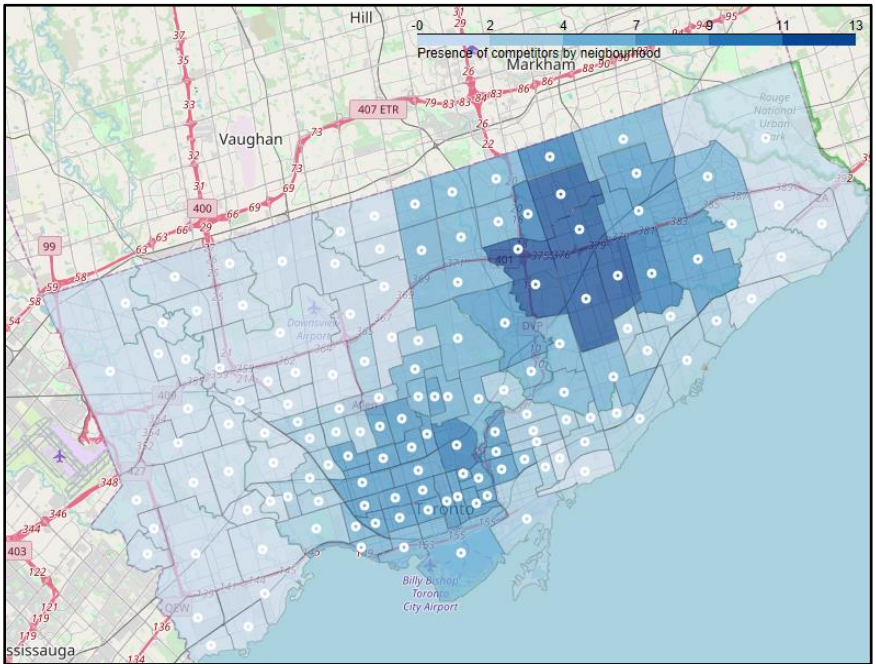

Figure 3-14: Presence of other Indian/Pakistani restaurants by neighborhood

B. Correlation Analysis

After visualizing the spatial distribution of features, we analysed the correlation among various features. For this purpose, we used a correlation matrix plot as presented in Figure 3-15. The diagonal representing the correlation among the features themselves is exactly 1 as expected. Apart from this, we see high correlation among various features based on population. Like, population by age and total population are highly correlated. Similarly, populous neighbourhoods have higher number of TTC stops or number of collisions represented by high positive correlation.

However, we are unable to find any clear determinants for the restaurants count. This implies that the decision to find the location of a restaurant cannot be determined by any single feature. Therefore, we need to analyse all the features in combination and group similar neighbourhoods together. For this purpose, we performed cluster analysis and the results are available in the following section.

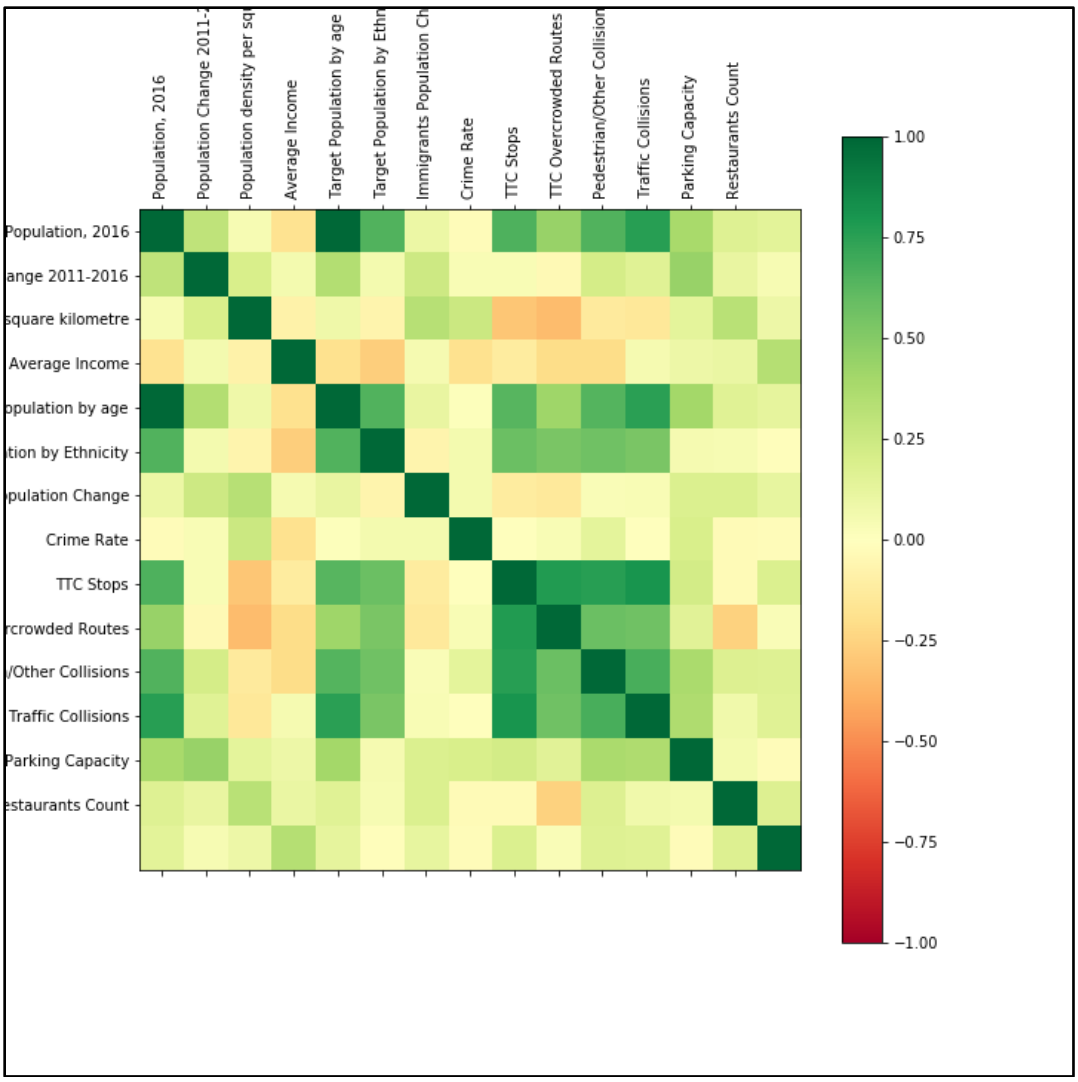

Figure 3-15: Correlation matrix

C. Clustering and Segmentation

As described in section 2, clustering was performed using **kmeans** algorithm to group the neighbourhoods in five clusters. The members of each of these clusters can be visualized in Figure 3-16. The distribution of neighbourhood in each cluster is presented in Table 3-1.

Table 3-1: Memership count for each cluster

Cluster Label	Number of neighborhoods
1	9
2	30
3	30
4	9
5	62

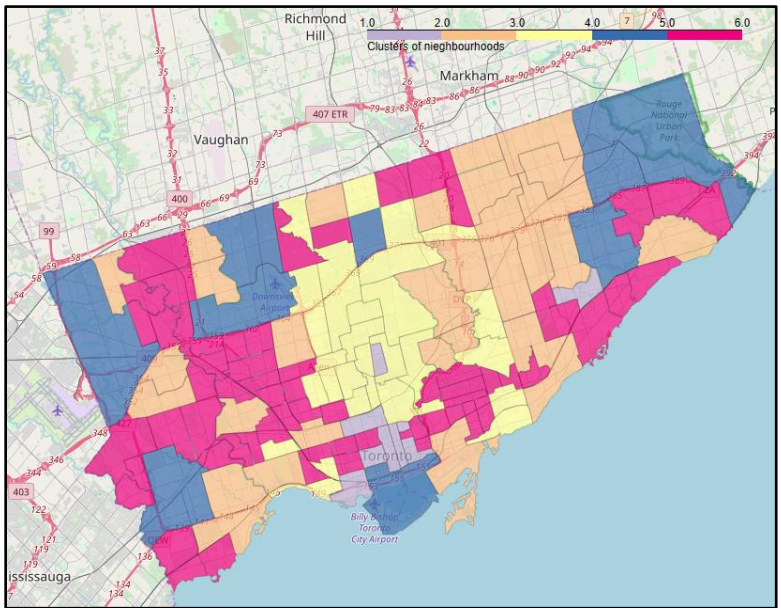

Figure 3-16: Neighborhood membership in resulting clusters

To understand the characteristics of clusters, we calculated the centroids of the clusters as presented in Table 3-2. The analysis of the table demonstrates that the cluster 4 is the most suitable cluster to host the new restaurant. The cluster has 9 restaurants so we decided to re-cluster the cluster 4. The visualization of the resulting clusters is presented in Figure 3-17. The analysis of these resultant clusters indicate that the **122 Waterfront Communities-The Island** is the most suitable candidate for hosting the new restaurant.

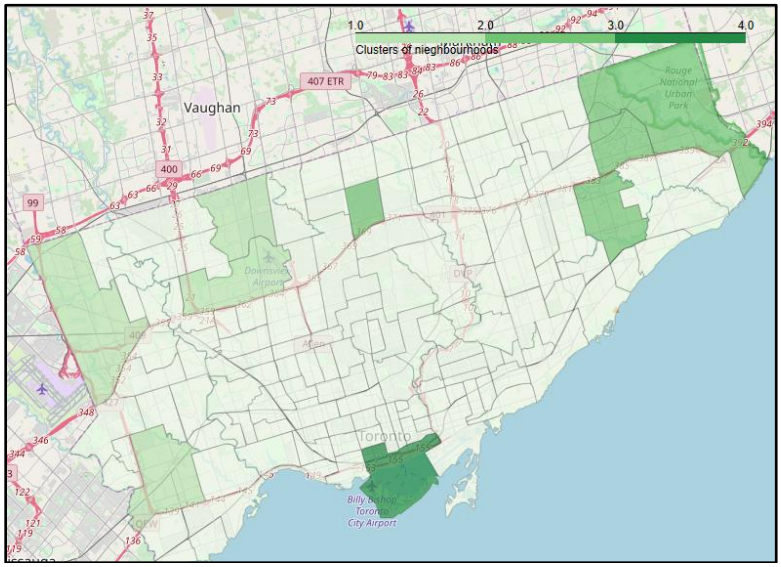

Figure 3-17: Neighborhood membership of re-clustering

Table 3-2: Centroids of the returned clusters

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
1	25371.44	0.14	17562.22	44379.22	22342.78	1852.78	0.26	3318.34	59.44	79.78	302.11	223.89	970.56	6.89
2	26843.43	0.03	4766.57	32857.53	22312.33	2819.83	0.03	1128.94	96.87	240.37	356.17	227.37	170.23	5.00
3	17001.90	0.06	6162.43	66497.83	14243.83	761.33	0.32	966.09	51.73	103.27	140.20	180.70	310.37	5.03
4	44449.33	0.08	4203.89	33961.67	38571.67	7398.33	0.13	1381.25	180.56	783.44	633.33	484.89	813.44	2.78
5	12706.84	0.01	5690.23	38432.10	10643.71	932.02	0.00	1471.82	42.08	137.84	106.19	106.02	90.26	3.06

A1 – Cluster Labels
A2 – Population
A3 - Population Change 2011-2016
A4 – Population Density
A5 – Average Income
A6 – Target Population by Age
A7 – Target Population by Ethnicity
A8 – Immigrant Population Change
A9 – Crime Rate
A10 – TTC Stops
A11 – TTC Overcrowded Rates
A12 - Pedestrian/Other Collisions
A13 - Traffic Collisions
A14 - Parking Capacity
A15 - Restaurants Count

4. DISCUSSION

In the previous section, we discussed the results of our analysis and based on the presented results, we concluded that **122 Waterfront Communities-The Island** is the most suitable candidate for hosting the new restaurant.

We performed a careful analysis based on the available data. However, despite having used the data sets from credible sources, we still had some limitations. We discuss these limitations in the following.

1. The population data in the neighborhood profiles dates back to 2016, so it does not reflect the current population estimates of the neighborhoods. Same applies to other features used from the same data set, like population density, population change, average income, presence of target population by age and ethnicity.
2. The parking spaces in each neighborhood were collected from Toronto Parking Authority's data sets. The data set contains parking spaces information operated the authority only.
3. The restaurants count for each neighborhood is found using Foursquare API's regular calls. For this purpose, centroids of the neighborhoods' boundaries were used as central location and the restaurants were retrieved for 5km radius. This may not be the right way as some queries could overlap and some of the restaurants could be missed.
4. The decision to choose restaurant's location is also based on some factors which were not included due to the limited scope of the project. Such as, foot traffic, presences of offices/schools/universities in the area, taxes and lease costs, etc.

Despite some limitations, this work presents an approach to find out the best location to host a new restaurant. The results of the work could be improved and made more realistic by working on the limitations mentioned above.

5. CONCLUSION

In this project, we worked for identifying the best neighborhoods for a new Indian/Pakistani restaurant. The highlights of the work are as follows:

1. We collected data from various sources, such as Foursquare API, Toronto City's Open Data Catalogue, Toronto Police Open Data Catalogue, etc.
2. The collected data was available in various formats including CSV, shapefile, JSON and xls.
3. The data preparation required lot of wrangling and cleaning before being available for the analysis.
4. For better understanding, we visualized the distribution of variables of the dataset using choropleth maps
5. Correlation analysis was performed to investigate the determinants of restaurant count
6. Data was normalized before modeling the data
7. Cluster analysis was done to segment the neighborhoods for the suitability of hosting a new Indian/Pakistani Restaurant.
8. The best neighborhoods was determined based on the cluster analysis, i.e. **122 Waterfront Communities-The Island**