Capstone Project – IBM Data Science

The Battle of the Neighbourhoods (Week 2)

Best Location for a New Fast-Food Restaurant in Toronto

By: Mary A.

1. Introduction

The project is aimed at leveraging the Foursquare location data to solve the problem of identifying the best neighborhood in which to open a new fast-food restaurant in Toronto. Details of the project as well as processing of the data used will be described and a neighborhood will be recommended to the client.

1.1 Background Information

One of the many decisions that needs to be made when opening a new business is about the location in which the business should be opened. There are usually a lot of factors that may be considered for a location depending on the specific business. One good pointer as to whether a specific area will be potentially good for a business is to consider if there is a similar business in the same or similar areas to the one being considered. In this project, we will explore how information about different locations can be used to decide where to open a new business. Specifically, we are interested in helping a client find the best location for a new fast food (FF) restaurant in Toronto.

1.2 Problem Statement

Our client, Maria is planning to open a new fast food (FF) restaurant in Toronto but needs guidance on the best neighborhood in which she can open the restaurant. Mary is not yet decided on the specific fast food (e.g., KFC, McDonalds, Burger King) but she is certain that it will be a fast-food restaurant. Based on initial conversations with Maria, she believes that neighborhoods with an existing FF restaurant is a viable option. Neighborhoods that also look similar (in terms of surrounding venues) to another neighborhood with a FF restaurant are also likely to be a good location. However, opening a FF restaurant in a location that already has many FF restaurants or within close proximity to another FF restaurant will create more competition that Maria would prefer to avoid for her new business. Since there are so many neighborhoods, Maria would like to know which neighbourhood is the best location in which to open a new FF restaurant in Toronto.

1.3 Target Audience

This project is targeted at helping a client - Maria to know where to open a new FF restaurant in Toronto based on how many restaurants are currently in each neighborhood and the similarity of each neighborhood to ones with existing fast-food restaurants. A similar approach to this analysis can be employed to find the best location for other businesses as well.

2. Data

This project uses data from different sources to address the problem stated. Information about neighborhoods in Toronto will be extracted from a web page. Detailed information about the venues in each neighborhood will then be gathered from Foursquare for further processing.

2.1 Description of Data

The following data are used in the project.

1. Borough and neighborhoods information: This information is gathered by web scraping a Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) that contains a list of the postal codes, borough, and neighborhoods of Toronto. It is then stored in a data frame.
2. Longitude and Latitude information: In order to explore a neighborhood using Foursquare, we need the latitude and longitude of the locations. There are two ways in which this information could be gathered. The first approach is to make use of the geocoder package, which provides the latitude and longitude of a given location/postal code. The alternative is to use a CSV file in which this information had already been stored at this URL. The project uses this alternative method and reads the latitude and longitude of each neighborhood form the CSV file.
3. Venues in each neighborhood: To analyze the neighborhoods in Toronto, we need more information about the venues around each neighborhood. Foursquare location services provide a way to gather this information and will be used to get data on venues in each neighborhood of Toronto.

Although it is possible to limit our analysis to only neighborhoods with 'Toronto' in its name, we decided against this in order to have more neighborhood venues to be explored for the best location for a FF restaurant.

2.2 How the Data will be used

The data will be used to arrive at a recommendation of a neighbourhood where a new FF restaurant is best located.

We will start out with a data frame containing all the neighbourhoods in Toronto. Latitudes and Longitudes of the neighborhoods will then be added to the data frame. This information will be used to query Foursquare location services to obtain venues in each neighborhood. Each venue usually belongs to a venue category in the obtained data. For the purpose of the project, we are particularly interest in the Fast-Food Restaurant category since it describes exactly the business that our client is interested in opening. However, to better understand the neighborhoods and how similar they are to each other, the complete set of data for all neighborhoods and venue categories will be used to cluster the neighborhoods into groups of similar neighborhoods.

Using the data frame that contains the complete list of venues in each neighborhood, we do some data cleaning to prepare the data for one-hot ending of the venue categories. We will then use the resulting data of one-hot encoding to do some exploratory analysis to see some of the most common venues in

each neighborhood. To identify neighborhoods that are similar to one another, the k-means model will be used to cluster the neighborhood into a number of clusters. The method employed to identify the best number for k is the Elbow method.

Throughout the project report, we will use plots and Folium maps for visual exploratory analysis and understanding of the neighborhoods and other data.

Once we have the clusters defined, we will explore each cluster more closely and find the average number of FF restaurants in each cluster. Our goal is to find the top 2 clusters with the highest average number of restaurants within the cluster. We decided to use the top 2 clusters to have more neighborhoods to further explore. Although having highest average number of FF restaurants in a cluster would imply that there is demand for fast food in that cluster, we also want to limit the amount of competition that the business would face when opened in a neighborhood within the cluster. To find the best neighborhood for the business, we will calculate a distance matrix to show the distance between neighborhoods with no existing FF restaurant and those with at least one FF restaurant within the best clusters earlier identified. Distance to neighborhoods with multiple FF restaurant will be multiplied by the number of restaurants in the neighborhood for a better analysis of the competition that may be faced. Finally, we find the maximum average distance across all neighborhoods that currently have no FF restaurant, but which are within the best clusters identified. The maximum average distance implies that it has the least amount of competition with existing FF restaurants that are in the best clusters earlier identified but it is still within a neighborhood where FF restaurants are in high demand.

The analysis will lead to the identification of the best 2 neighborhoods in which our client may open a new fast- food restaurant in Toronto. The neighborhoods are shown on the map together with other neighborhoods and their clustering. Neighborhoods with existing fast-food restaurants will also be identified on the final map for a visual identification of the best neighborhoods found for the new business.

2.3 Data Gathering

The first step in the data gathering phase was to use web scraping to obtain the list of neighborhoods in Toronto together with their corresponding postal code and borough. This information was store in a data frame shown in Figure 1.

The coordinate of the locations is obtained from the CSV file mentioned above and merged to the information previously obtained so that each neighborhood information also lists the coordinate of the neighborhood. The resulting data after this stage is shown in Figure 2 with a total of 103 neighborhoods data. Finally, we use the latitude and longitude information for each neighborhood to find venues close to the neighborhood from Foursquare API [1]. This information is stored in a new data frame and has the structure shown in Figure 3.

The collected data venue had 2136 venues across the various neighborhoods. This complete set of data regarding the venues will be analyzed to make recommendation regarding where to open a fast-food restaurant.

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills North
8	M4B	East York	Parkview Hill, Woodbine Gardens

Figure 1: Neighborhoods in Toronto

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills North	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937

Figure 2: Data frame with Coordinate information

	PostalCode	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M3A	North York	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	M3A	North York	Parkwoods	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant
2	M3A	North York	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	M4A	North York	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	M4A	North York	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

Figure 3: Toronto Venue Data

3. Methodology

In this section, we describe the major part of the project. Exploratory data analysis was done to better understand the different neighborhoods, the venues in them and other details about the venue. Machine learning techniques is applied in our analysis to arrive at a solution to the stated problem.

3.1 Exploratory Data Analysis

In our data exploration, we will look at the categories to better understand each category. But we can already see from the second row of the sample in Figure 3 that there is a Fast Food Restaurant category, which is exactly the category for the business that our client wants to open. We explore further to see how many neighborhoods have each of the categories, the total number of distinct categories of venues and the number of fast-food restaurants in each neighborhood. Figure 4 shows the grouping of the venue category to see how may of each is in a neighborhood.

Venue Category	PostalCode	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Accessories Store	2	2	2	2	2	2	2	2
Adult Boutique	1	1	1	1	1	1	1	1
Airport	2	2	2	2	2	2	2	2
Airport Food Court	1	1	1	1	1	1	1	1
Airport Gate	1	1	1	1	1	1	1	1
Airport Lounge	2	2	2	2	2	2	2	2
Airport Service	3	3	3	3	3	3	3	3
Airport Terminal	2	2	2	2	2	2	2	2
American Restaurant	19	19	19	19	19	19	19	19
Antique Shop	3	3	3	3	3	3	3	3

Figure 4: Count of Neighborhoods with each Venue Category

Reviewing the data at this point shows that there is a venue category called Neighborhood. Since we already have a column called 'Neighborhood' in our dataset, we will rename this category to 'Neighborhood Area' to avoid confusion. We also note that there are 274 unique categories and the number of fast-food restaurants in each neighborhood as shown in Figure 5.

The next step of our exploratory analysis was to view the existing fast-food restaurants on a map. For this, we used Folium map and add a label to show the name of the fast-food restaurant on the map. As shown in Figure 6, this provided a good way of knowing how existing fast-food restaurant are located in Toronto.

	Neighborhood	No_of_restaurants
0	Bedford Park, Lawrence Manor East	1
1	Church and Wellesley	2
2	Clarks Corners, Tam O'Shanter, Sullivan	2
3	Commerce Court, Victoria Hotel	1
4	Enclave of M4L	1
5	Enclave of M5E	1
6	Fairview, Henry Farm, Oriole	4
7	Garden District, Ryerson	2
8	High Park, The Junction South	1
9	Hillcrest Village	1
10	India Bazaar, The Beaches West	1
11	Malvern, Rouge	1
12	Mimico NW, The Queensway West, South of Bloor,...	1
13	New Toronto, Mimico South, Humber Bay Shores	1

Figure 5: Number of Restaurant in a Neighborhood

Figure 6: Existing Fast-Food Restaurants

3.2 Analyzing the Neighborhoods

In our analysis of the neighborhood, we use one-hot encoding to identify how many of each venue category are in each neighbourhood. We also use the encoding to further analyze each neighborhood to see which venues are the 10 most common in the neighborhoods. The summary information obtained from our data after using one-hot encoding is shown in Figure 7.

	Neighborhood	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store
0	Agincourt	4	4	4	4	4	4	4	4	4	4	4	4	4	4
1	Alderwood, Long Branch	6	6	6	6	6	6	6	6	6	6	6	6	6	6
2	Bathurst Manor, Wilson Heights, Downsview North	23	23	23	23	23	23	23	23	23	23	23	23	23	23
3	Bayview Village	4	4	4	4	4	4	4	4	4	4	4	4	4	4
4	Bedford Park, Lawrence Manor East	25	25	25	25	25	25	25	25	25	25	25	25	25	25
5	Berczy Park	58	58	58	58	58	58	58	58	58	58	58	58	58	58

Figure 7: Output of One-hot Encoding Summary

Knowing the most common venues in east neighborhood gives us an idea of what the neighborhood might look like. To do this, we define a function to sort venues in descending order, create the new data frame and display the top 10 venues for each neighborhood. This result is shown in Figure 8.

	Neighborhood	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	Agincourt	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	College Stadium
1	Alderwood, Long Branch	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	College Stadium
2	Bathurst Manor, Wilson Heights, Downsview North	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	College Stadium
3	Bayview Village	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	College Stadium
4	Bedford Park, Lawrence Manor East	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	College Stadium

Figure 8: Common Venues in Neighborhoods

3.3 Clustering the Neighborhoods

To find neighborhoods that are similar in terms of venues in the neighborhood we will create a cluster of neighborhoods using k-means algorithm. Using this algorithm requires that a value if provided for k. Since we do not know how well to cluster the neighborhoods or how many clusters is most suitable. We employ the Elbow method to find the best value of k and then cluster the neighborhood into this number of clusters.

Finding the best value of *k*

Since our target business is a fast-food restaurant, we use the data on fast food restaurant and the elbow method to determine the best way to cluster the neighbourhoods. Figure 9 shows the result of the KElbow visualizer and identifies the best value of k as 5.

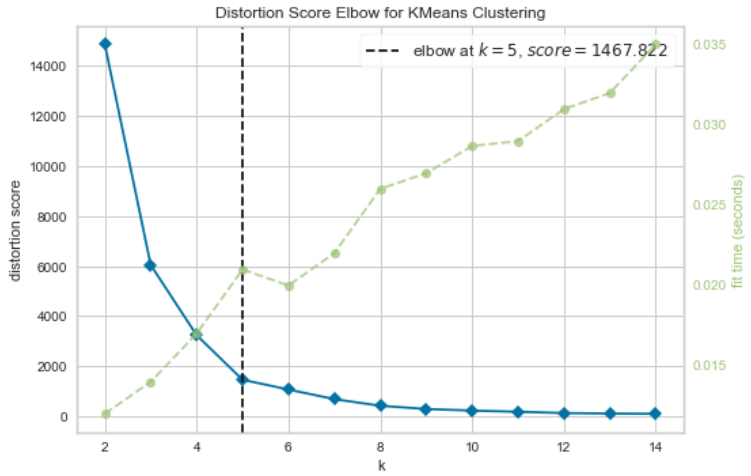

Figure 9: KElbow Visualizer Output

The k-means algorithm is used to cluster the neighborhood into 5 clusters. Each cluster show that the neighborhoods are quite similar. The data frame containing the list of neighborhoods and venues is then updated to include the cluster which each neighborhood falls into. To visualize the clusters, we show each neighborhood on Folium map and identify similar neighborhood (in the same cluster) with the same color as shown in Figure 10. Lastly, we examined each of the clusters to observe their similarities. For example, the details of cluster 1 is shown in Figure 11.

Our next goal is to examine each cluster more closely and calculate the average number of restaurants in each cluster. We are interested in the cluster with the highest average number of restaurants since this indicates that fast food (FF) restaurant is common in the cluster. The analysis resulted in Cluster 1 with an average of 1 FF restaurants as the best cluster containing a neighborhood in which a fast-food restaurant may be opened. The second-best cluster identified was Cluster 2 with an average of 0.85 FF restaurants in the cluster.

We can visualize the average number of restaurants in each cluster as shown in Figure 12 to see the best clusters to open a new fast-food restaurant based on those with the higher average number of FF restaurants, since the higher averages shows that fast food restaurant is a common venue in the cluster, most likely owing to demand for fast food. However, to decide the best neighborhood in these clusters, we need to explore the clusters even more closely.

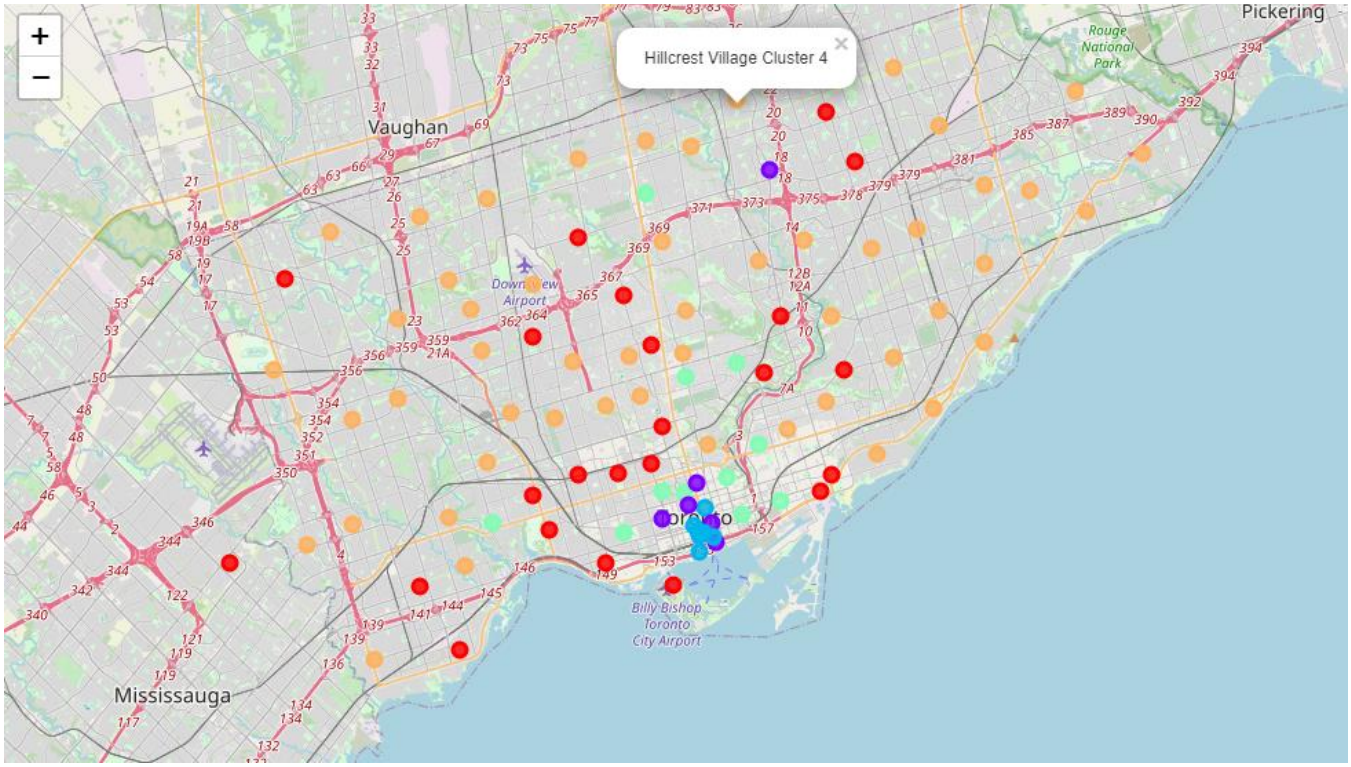

Figure 10: Clustering the Neighborhoods

	PostalCode	Borough	Neighborhood	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
14	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	1	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore
19	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	1	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore
23	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	1	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore
32	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556	1	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore
80	M5T	Downtown Toronto	Kensington Market, Chinatown, Grange Park	43.653206	-79.400049	1	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore
94	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	1	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore

Figure 11: Examining Cluster 1

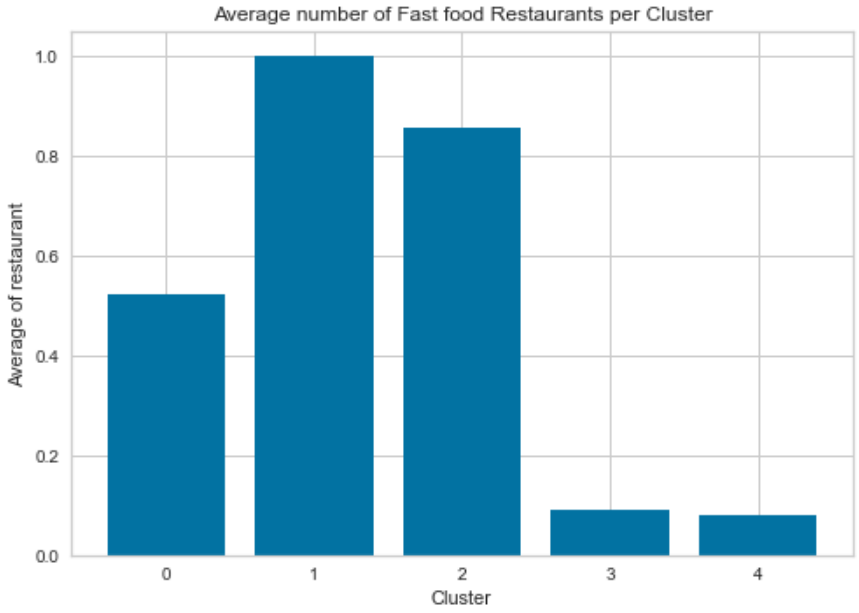

Figure 12: Average Number of FF-Restaurants Per Cluster

3.4 Finding the Best Neighborhood for a new Fast-food Restaurant

In our analysis until this point, we have found the best two clusters that have the highest number of fast-food restaurants on average. It will be best to open a new fast-food restaurant in a location where of fast-food restaurants are located but we also want to avoid opening a fast-food restaurant in a neighborhood that already has one. The aim of this section is to find which neighborhood in the two clusters found would be the best neighborhood for the restaurant. Simply put, we want to find a

neighborhood in these clusters that currently has no fast-food and which has the average farthest distance from other neighborhoods with a fast-food restaurant. To achieve this, we find the list of neighborhoods with a fast-food restaurant in the best clusters, and the list of those with no fast-food restaurants. We then find the average distance between the set. For neighborhoods with multiple fast-food restaurants, this is considered by multiplying the distance by the number of restaurants.

The task of finding the average distance from a neighborhood with no FF restaurant within the identified cluster was completed by using a distance matrix, as shown in Figure 13 to capture the distances between neighborhoods with no ff restaurants and those with a ff restaurant within the identified clusters.

Neighborhood	Fairview, Henry Farm, Oriole	Church and Wellesley	Garden District, Ryerson	Richmond, Adelaide, King	Toronto Dominion Centre, Design Exchange	Commerce Court, Victoria Hotel	Enclave of M5E
St. James Town	0.521045	0.032639	0.013343	0.009196	0.007521	0.005496	0.005091
Berczy Park	0.545583	0.046555	0.027221	0.012667	0.008613	0.007358	0.002267
Central Bay Street	0.509161	0.017929	0.016965	0.007900	0.012240	0.012344	0.017024
Kensington Market, Chinatown, Grange Park	0.545008	0.042208	0.042959	0.015704	0.019432	0.020843	0.026097
Harbourfront East, Union Station, Toronto Islands	0.568515	0.050167	0.033174	0.010154	0.006364	0.007632	0.008904
First Canadian Place, Underground city	0.539618	0.034906	0.018701	0.003134	0.001437	0.002474	0.007697

Figure 13: Distance Matrix for Neighborhoods in the recommended clusters

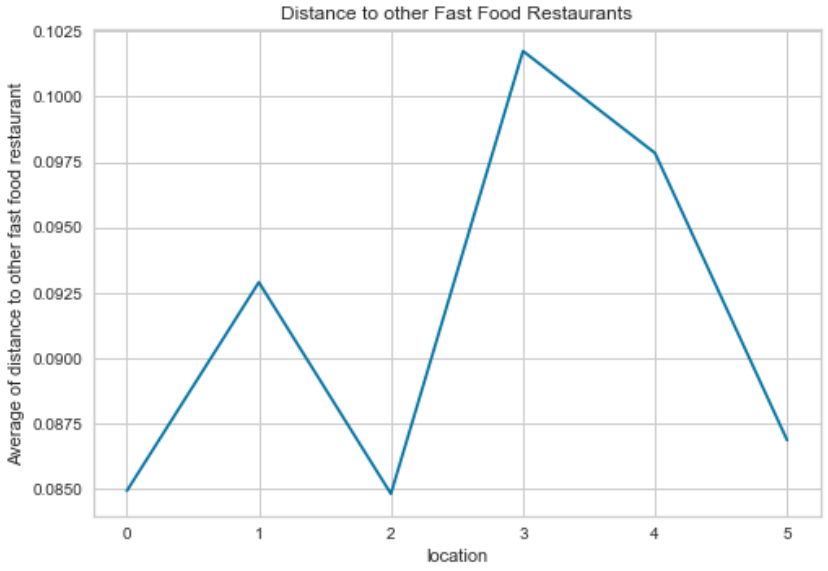

Figure 14: Average Distance to other Fast-food Restaurants

We can visualize the average distance of the neighborhoods with no FF restaurants to others with an existing fast-food restaurant in the best two clusters as shown in Figure 14. The result show that the best two neighbourhoods to open a new fast-food restaurant are those with postal codes M5T and M5J as shown in Figure 15.

	PostalCode	Borough	Neighborhood	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
3	M5T	Downtown Toronto	Kensington Market, Chinatown, Grange Park	43.653206	-79.400049	1	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore
4	M5J	Downtown Toronto	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	2	Yoga Studio	Dumpling Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore

Figure 15: Best Neighborhoods for a new fast-food business

4. Results and Discussion

As shown above, the neighborhood - Kensington Market, Chinatown, Grange Park in cluster 5 is the best neighborhood to open a new fast-food restaurant. The neighborhood with the second highest distance may also be considered if there are other factors that are being considered.

To visualize these locations in contrast to existing fast-food restaurants, the map of all the clustered neighborhoods and the neighborhood with at least one fast-food restaurant is shown as in Figure 16. The red marked with a home icon shows the neighborhood in which the new fast-food restaurant is best located whereas the blue markers shows neighbours with at least a fast-food restaurant already existing.

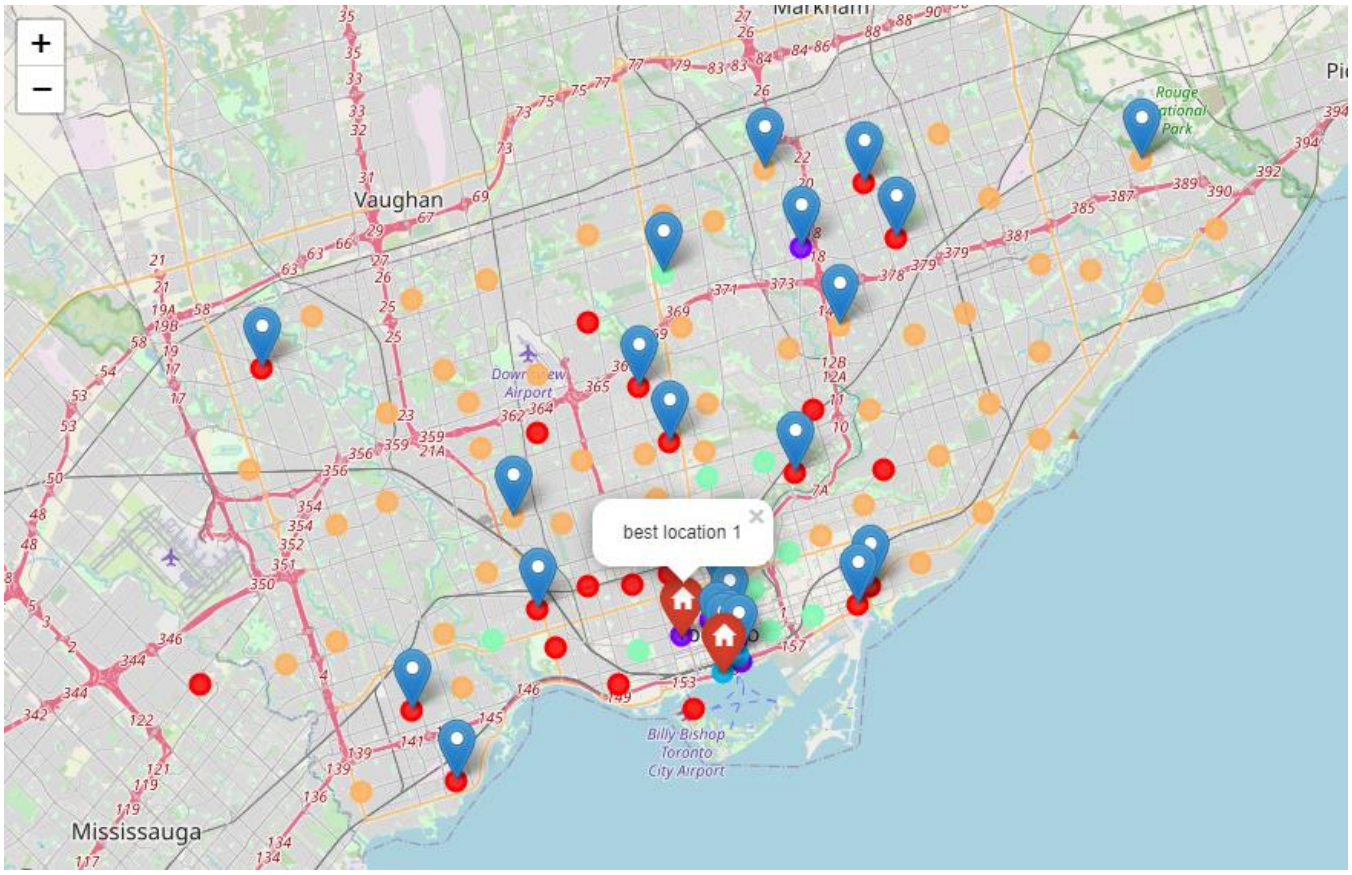

Figure 16: Best 2 locations for new fast-food Restaurant

As seen above and from the analysis, the best neighborhoods to open a fast-food restaurant based on the location data available is in the **Kensington Market, Chinatown, Grange Park** neighborhood or the **Harbourfront East, Union Station, Toronto Islands** neighborhood. Looking at the map, these two locations seem indeed like the best options as they are close to many various other locations but still away from other neighbourhoods that already have one or more fast food restaurants.

5. Conclusion

The goal of this project was to find a neighborhood that was most suitable to open a new fast food (FF) restaurant. We used the information about the neighborhoods in Toronto as well as data from Foursquare about these neighborhoods to find the neighborhoods that are most similar by generating clusters of neighborhoods. Neighborhoods with a high number of existing restaurants typically implies that there is demand for fast food in such neighborhood. This was useful in determining which cluster a new FF restaurant would be best located. As with any business, our client is also interested in limiting the amount of competition that her new business would have to face while starting out. Our approach for finding the best neighborhood for a new FF restaurant involved evaluating the distance to existing fast foods in the identified neighbourhood before arriving at a conclusion on the best two neighborhoods in which the fast-food restaurant may be opened.

It is important to note that this decision is based on the location information only. As mentioned earlier, opening a business involves a lot of factors and there may be other factors that are important to the client that may need to be further explored for each of these neighborhoods. The recommendation to the client would be to consider these top two neighborhoods identified.

References

[1] Foursquare (https://developer.foursquare.com/docs/places-api/endpoints/)

[2] Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)