Use of k-Means Clustering to build a new Jollibee Franchise in Toronto

Jul 25, 2020 • data, analysis • learning, personal development, goal

Note: This is my final project for the IBM Data Science Professional Certificate in Coursera. The business problem is purely fictional.

1. Introduction

1.1 Background

Jollibee is the largest fast food restaurant chain in the Philippines, while very well known around Southeast Asia and Middle East, there is only one Jollibee franchise in Toronto, Canada.

1.2 Business Problem

Jollibee Foods Corporation (JFC) looks to expand its footprint in Toronto, Canada. As of July 2020, the only Jollibee franchise is in Scarborough and the company is looking for a similar neighborhood where they can open another franchise.

2. Data Acquisition and Cleaning

2.1 Data Sources

1. Borough, neighborhood, and postal code - Wikipedia
2. Geographical coordinates provided by the IBM Data Science Professional Course - cocl.us
3. Neighorhood Profiles - Toronto Open Data Portal
4. Foursquare Venue API - Explore Venues

2.2 Data Preparation

1. Borough, neighborhood, and postal code
   Data was obtained by scraping the Wikipedia page using the `urllib.requests` library provided by Python and the Beautiful Soup library. Only data rows with proper values were collected, none of the `Not assigned` rows were collected.
2. Geographical coordinates provided by the IBM Data Science Professional Course
   The dataset provided is in CSV format and the `pandas.read_csv` function was used to read the data into a `DataFrame`.
3. Neighborhood Profiles
   Only specific rows were used from the CSV dataset, the total population of 2016, the total Filipino population, and the average income of each neighborhood were used. The dataset was using the neighborhood names as columns which is why each data selected was transposed.
4. Foursquare Venue API
   Using the latitude and longitude with a radius of 1000, all venues were queried from the API.

After scraping and cleaning each data source, all of data were combined into a single data frame. The head of the final data frame is shown below.

	Neighborhood	Latitude	Longitude	Average Income	Percentage of Filipino	Fast Food Restaurant
0	Victoria Village	43.725882	-79.315572	35786.0	7.367219	0.0
1	Regent Park	43.654260	-79.360636	34597.0	2.823290	0.0
2	Malvern	43.806686	-79.194353	29573.0	11.634014	1.0
3	Rouge	43.806686	-79.194353	39556.0	9.387904	1.0
4	Highland Creek	43.784535	-79.160497	40972.0	7.043381	0.0

Fig 1: Head of Final DataFrame

3. Methodology

3.1 Visualizing the Neigborhood Data

The `folium` library was used to visualize the location of each neighborhood in the final data frame.

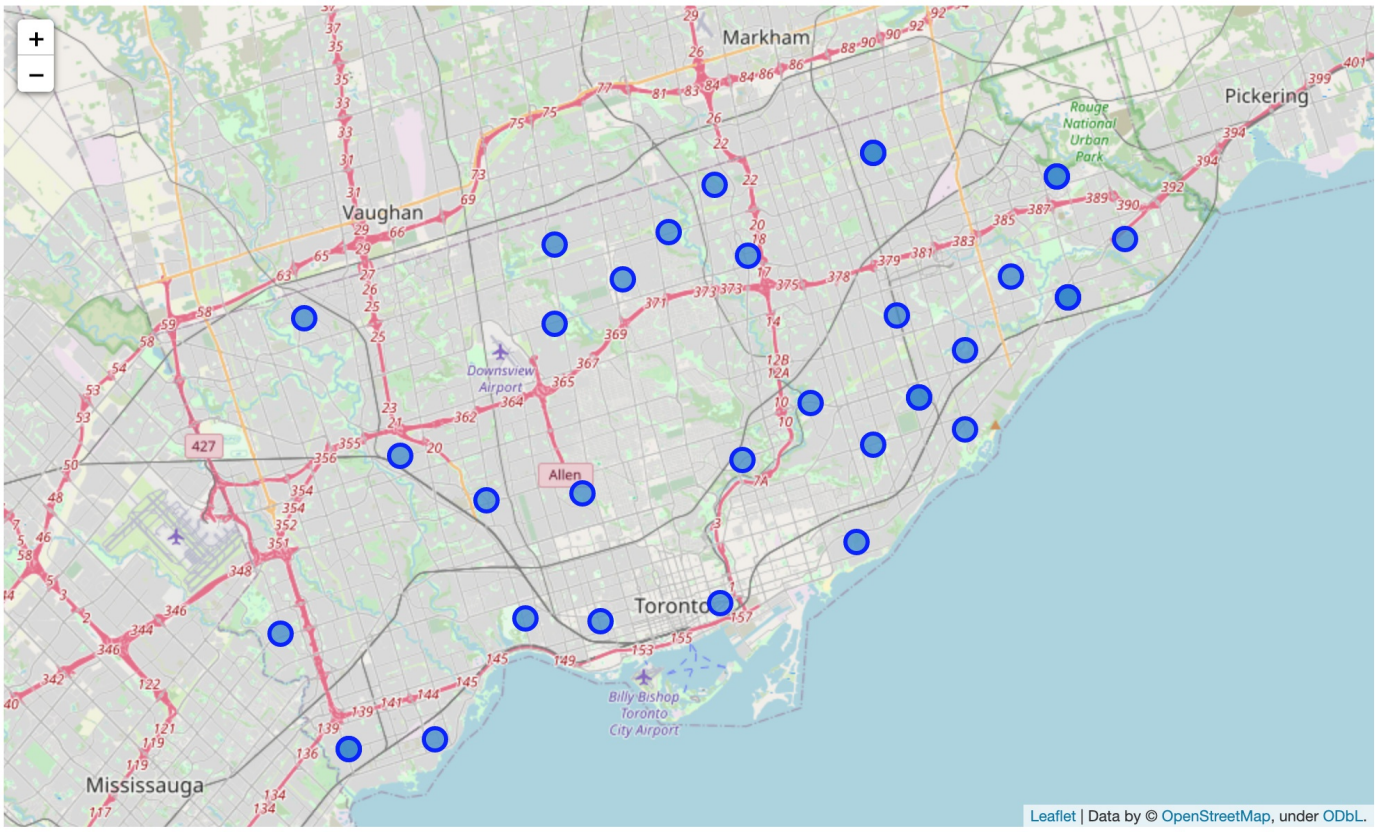

Fig 2: Neighborhoods of Toronto

3.2 Exploring the Data

3.2.1 Distribution of Fast Food Restaurants

As mentioned in the **Data Acquisition and Cleaning** section, the Foursquare API was used to explore the venues within 1000 miles of Toronto, Canada. The frequency distribution of fast food restaurants for each neighborhood was calculated, to determine where's the best place to build another Jollibee franchise. Typically, fast food restaurants tend to group together as their target market is similar.

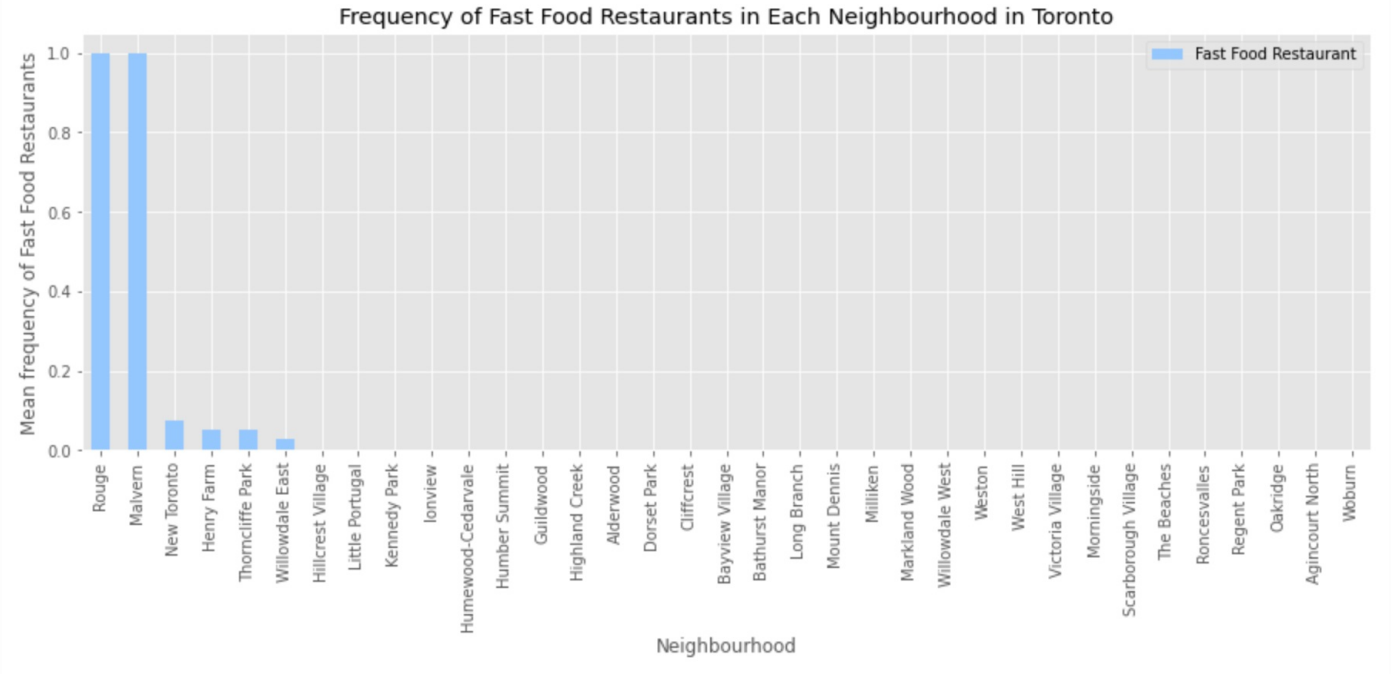

Fig 3: Frequency Distribution of Fast Food Venues in Toronto

3.2.2 Distribution of Filipino Population

Jollibee is a Filipino staple, in fact, it is the only fast food restaurant that beat McDonanld's and this is the reason why I used this data. Jollibee can leverage nostalgia to help attract customers and opening in a neighborhood with a good amount of Filipino will help with that.

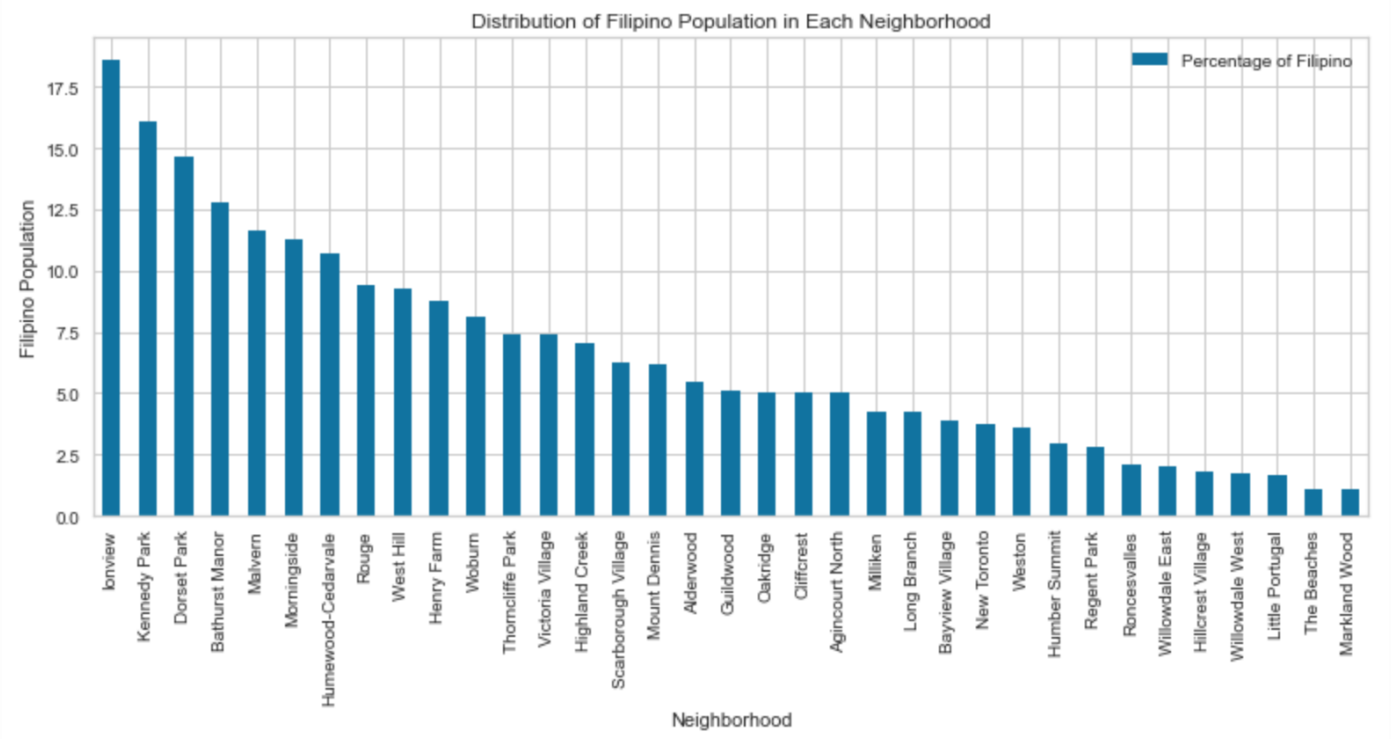

Fig 4: Frequency Distribution of Filipino Population in Toronto

3.2.3 Distribution of Average Household Income

Lastly, the average houshold income is used to determine which neighborhoods will be able to afford to buy from Jollibee. Jollibee value meal price ranges from around 5 to around 8 dollars, we can assume that a neighborhood with an average household income around the median should be a good spot.

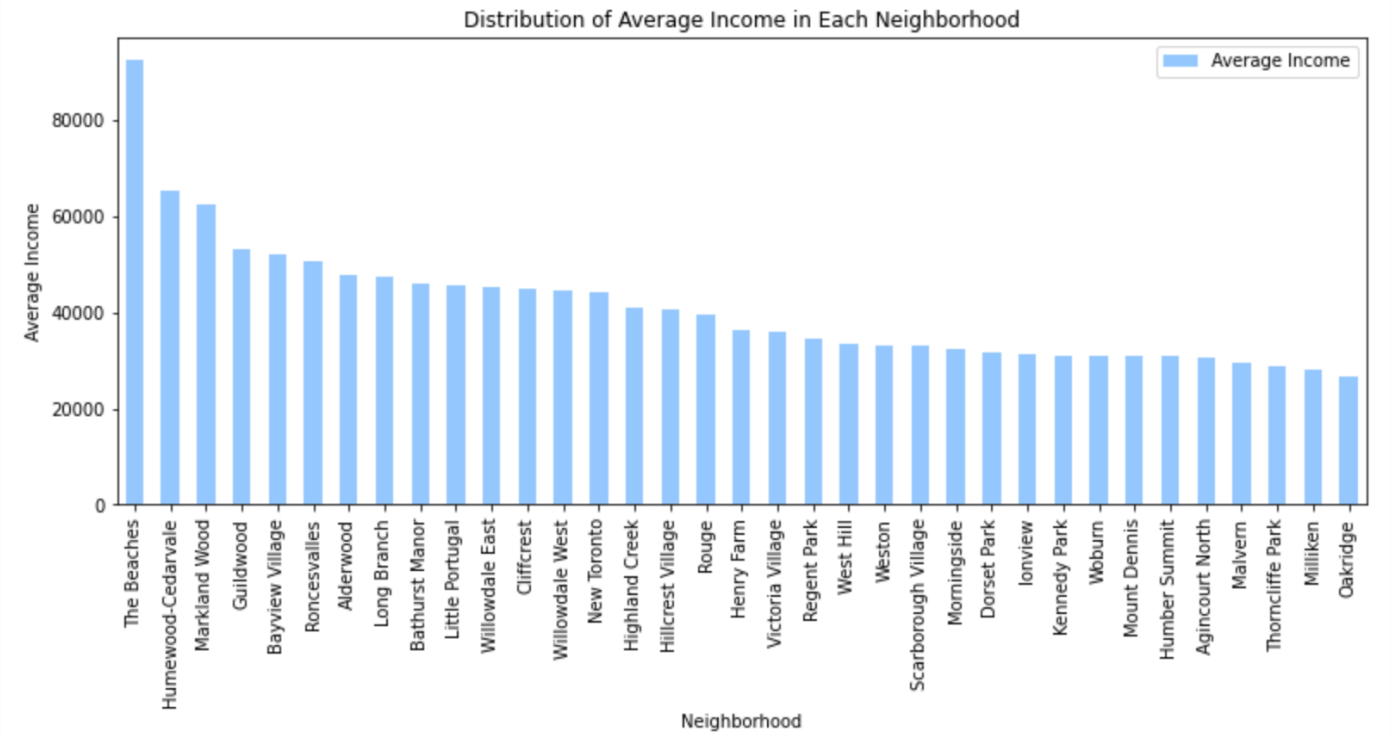

Fig 5: Frequency Distribution of Average Income in Toronto

3.3 Clustering the Data

For the model, we will be using the distribution of fast food restaurants, distribution of Filipino population. Let's start with normalizng the dataset.

3.3.1 Normalizing the data

Using the `StandardScaler` of the `sklearn.preprocessing` library, we can normalize the data, which results in:

	Average Income	% Filipino	No. of Fast Food Restaurants
0	-0.390734	0.193622	-0.292960
1	-0.481835	-0.844644	-0.292960
2	-0.866772	1.168565	4.023428
3	-0.101878	0.655340	4.023428
4	0.006616	0.119627	-0.292960

Fig 6: Normalized Data

3.3.2 Finding the optimzal `k` for clustering

To determine the optimal `k`, the *Elbow Method* was used with the help of the `yellowbrick` library. To summarize, the optimal `k` is 5.

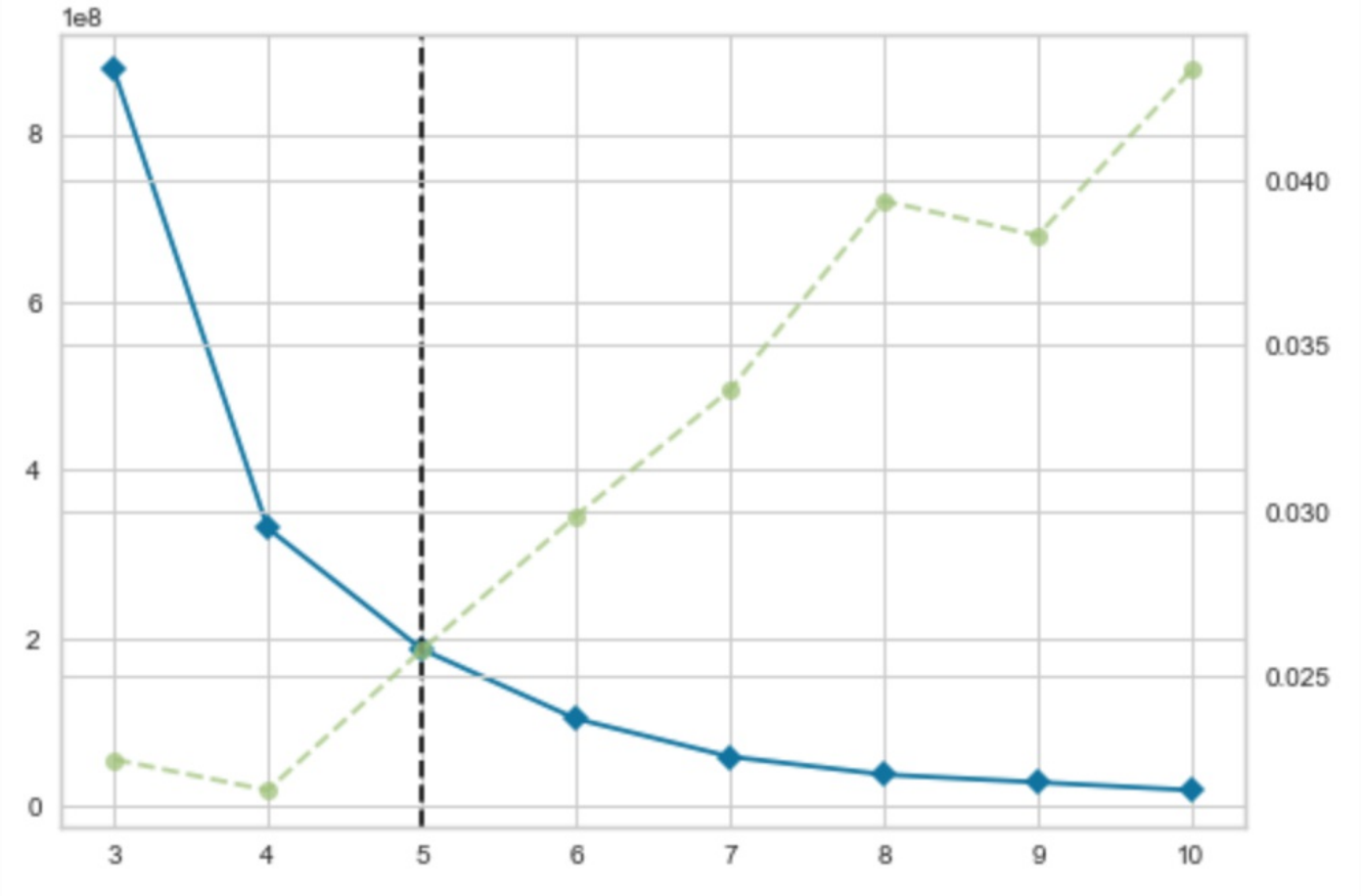

Fig 7: Optimal `k` is 5 using the Elbow Method

Results

Visualizing Clustered Data

Using the `folium` library, we can visualize the clustered data.

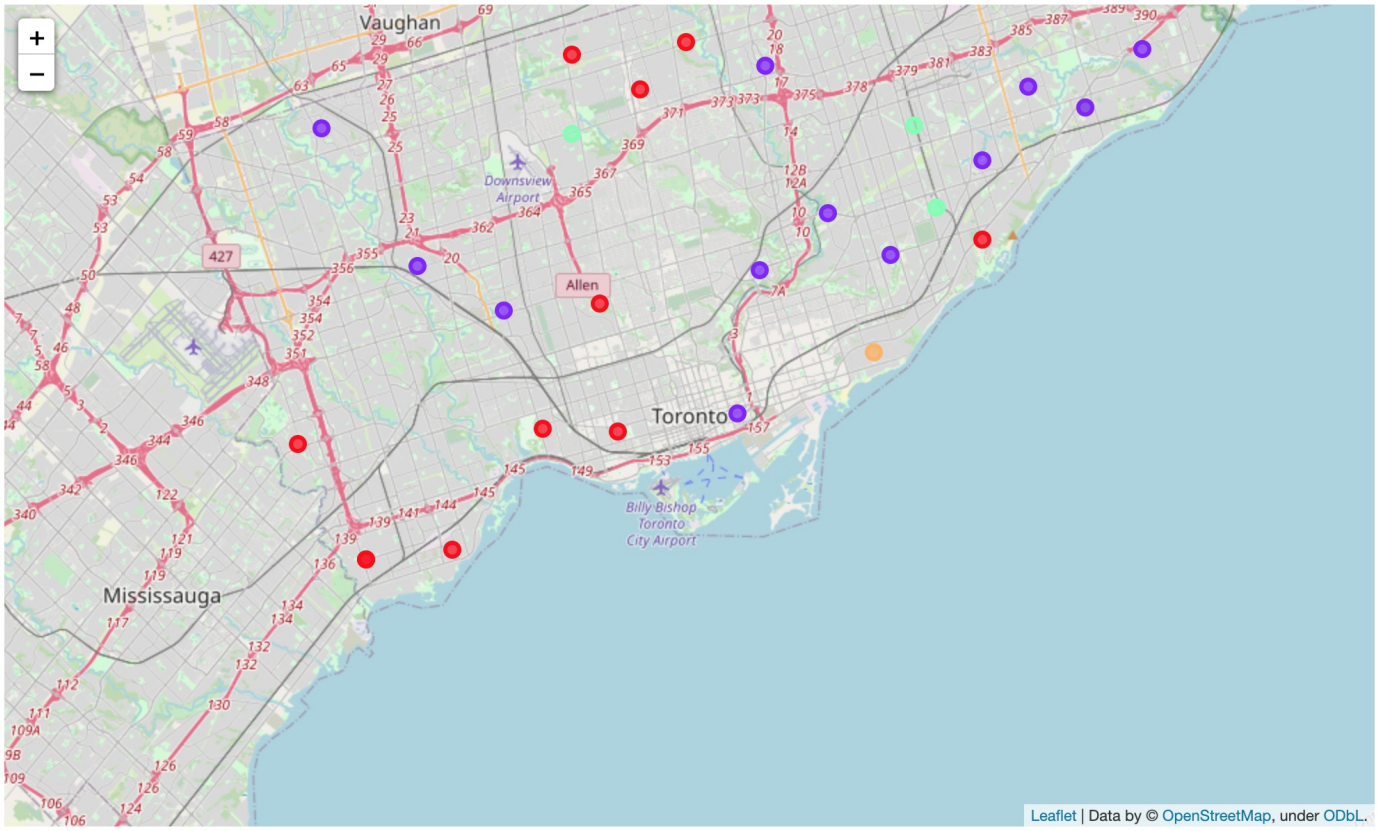

Fig 8: Clustered Folium Visualization

Examining the clusters

Cluster 0

- MID Average Income
- LOW-MID Filipino Population
- LOW Fast Food Venues

	Neighborhood	Latitude	Longitude	Average Income_x	Percentage of Filipino	No. of Fast Food Restaurants	Cluster Label
5	Humewood-Cedarvale	43.693781	-79.428191	65274.0	10.720501	-0.301026	0
6	Markland Wood	43.643515	-79.577201	62378.0	1.089634	-0.301026	0
7	Guildwood	43.763573	-79.188711	53203.0	5.092266	-0.301026	0
12	Hillcrest Village	43.803762	-79.363452	40442.0	1.830637	-0.301026	0
17	Little Portugal	43.647927	-79.419750	45737.0	1.671059	-0.301026	0
20	Bayview Village	43.786947	-79.385975	52035.0	3.879230	-0.301026	0
23	Cliffcrest	43.716316	-79.239476	44718.0	5.051773	-0.301026	0
25	Willowdale East	43.770120	-79.408493	45326.0	2.002617	-0.047535	0
28	Willowdale West	43.782736	-79.442259	44576.0	1.741852	-0.301026	0
29	Roncesvalles	43.648960	-79.456325	50580.0	2.103646	-0.301026	0
32	New Toronto	43.605647	-79.501321	44101.0	3.707581	0.417199	0
33	Alderwood	43.602414	-79.543484	47709.0	5.475361	-0.301026	0
34	Long Branch	43.602414	-79.543484	47384.0	4.214597	-0.301026	0

Fig 9: Cluster 0

Cluster 1

- LOW-MID Average Income
- MID Filipino Population
- LOW Fast Food Venues

	Neighborhood	Latitude	Longitude	Average Income_x	Percentage of Filipino	No. of Fast Food Restaurants	Cluster Label
0	Victoria Village	43.725882	-79.315572	35786.0	7.367219	-0.301026	1
1	Regent Park	43.654260	-79.360636	34597.0	2.823290	-0.301026	1
4	Highland Creek	43.784535	-79.160497	40972.0	7.043381	-0.301026	1
9	West Hill	43.763573	-79.188711	33323.0	9.291034	-0.301026	1
11	Woburn	43.770992	-79.216917	30878.0	8.123773	-0.301026	1
14	Thorncliffe Park	43.705369	-79.349372	28875.0	7.414251	0.129909	1
15	Scarborough Village	43.744734	-79.239476	32913.0	6.218608	-0.301026	1
16	Henry Farm	43.778517	-79.346556	36359.0	8.776951	0.213524	1
21	Oakridge	43.711112	-79.284577	26793.0	5.055977	-0.301026	1
22	Humber Summit	43.756303	-79.565963	30731.0	2.939755	-0.301026	1
24	Mount Dennis	43.691116	-79.476013	30827.0	6.179651	-0.301026	1
26	Weston	43.706876	-79.518188	32997.0	3.612717	-0.301026	1
30	Milliken	43.815252	-79.284577	28085.0	4.271414	-0.301026	1
31	Agincourt North	43.815252	-79.284577	30414.0	5.032116	-0.301026	1

Fig 10: Cluster 1

Cluster 2

- LOW-MID Average Income
- MID-HIGH Filipino Population
- HIGH Fast Food Venues

	Neighborhood	Latitude	Longitude	Average Income_x	Percentage of Filipino	No. of Fast Food Restaurants	Cluster Label
2	Malvern	43.806686	-79.194353	29573.0	11.634014	4.008325	2
3	Rouge	43.806686	-79.194353	39556.0	9.387904	4.008325	2

Fig 11: Cluster 2

Cluster 3

- HIGH Average Income
- LOW Filipino Population
- LOW Fast Food Venues

	Neighborhood	Latitude	Longitude	Average Income_x	Percentage of Filipino	No. of Fast Food Restaurants	Cluster Label
10	The Beaches	43.676357	-79.293031	92580.0	1.112811	-0.301026	3

Fig 12: Cluster 3

Cluster 4

- MID Average Income
- HIGH Filipino Population
- LOW Fast Food Venues

	Neighborhood	Latitude	Longitude	Average Income_x	Percentage of Filipino	No. of Fast Food Restaurants	Cluster Label
8	Morningside	43.763573	-79.188711	32291.0	11.257519	-0.301026	4
13	Bathurst Manor	43.754328	-79.442259	45936.0	12.789013	-0.301026	4
18	Kennedy Park	43.727929	-79.262029	30974.0	16.060270	-0.301026	4
19	Ionview	43.727929	-79.262029	31383.0	18.583682	-0.301026	4
27	Dorset Park	43.757410	-79.273304	31692.0	14.638243	-0.301026	4

Fig 13: Cluster 4

Conclusion

In this study, I analyzed the neighborhoods of Toronto, Canada to determine where is the best place to build another Jollibee franchise. Using the average income, Filipino population, and fast food venues data, the most promising group is Cluster 2.

Typically, fast food restaurants tend to open near each other and with cluster 2 having the most number of fast food restaurants, a low-to-mid average income and mid-to-high Filipino population, `Cluster 2` fits the bill.

Specifically, the client can consider Rouge, Toronto to open another Jollibee franchise.

Future Directions

The model could use more improvments in capturing if the population of a neighborhood has more families with kids of a certain or age. Another thing we can consider is how much distance between each franchise.