Strategic Location for Establishing an Asian Restaurant in Toronto, Canada

Tan Jun Wei

October 20th 2020

1. Introduction

1.1 Background

The success of establishing a new restaurant depends on several factors: demand, brand loyalty, quality of food, competition, and so on. In most cases, a restaurant's location plays an essential determinant for its success. Hence, it is advantageous and of utmost importance to determine the most strategic location for establishment in order to maximize business profits.

1.2 Business Problem

A client seeks to establish a franchised Asian restaurant, with a niche in Southeast Asian cuisine, in a Toronto neighborhood. Which neighbourhood would appear to be the optimal and most strategic location for the business operations?

The objective of this capstone project is to locate the optimal neighborhood for operation. Our foundation of reasoning would be based off of spending power, distribution of ethnic group, and competition, across each neighbourhood. We will mainly be utilizing the Foursquare API and the extensive geographical and census data from Toronto's Open Data Portal.

1.3 Interests

Fellow entrepreneurs seeking to either establish a new restaurant of a certain niche or have plans to expand their franchised restaurants would be very interested in the competitive advantages and business values this finding can potentially reap.

2. Data Acquisition and Cleaning

2.1 Data Sources

The neighbourhoods alongside their respective postal codes and boroughs were scraped from Wikipedia. Geographical coordinates for each neighbourhood were extracted from here. As for Toronto's census data---median household income, total population, and population of Southeast Asians across each neighbourhood—Toronto's Open Data Portal provides all that data here.

For returning the number of Asian restaurants in the vicinity of each neighbourhood, we will be utilizing Foursquare API, more specifically, its *explore* function. One has to register for a Foursquare developer account here to access their API credentials.

2.2 Data Cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values for certain neighbourhoods, due to lack of record keeping. Few assumptions were made to achieve the dataframe shown in *Fig 1*:

- Only the cells that have an assigned borough will be processed; boroughs that were not assigned were ignored.
- Neighbourhoods missing more than two census data value were dropped.

A column that features the percentage of distribution of Southeast Asian across each neighbourhood was calculated by dividing the population of Southeast Asians by the total population of each neighbourhood. So, the two latter columns were made redundant and dropped.

	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Southeast Asian
0	Victoria Village	43.725882	-79.315572	43743.0	0.542547
1	Rouge	43.806686	-79.194353	72784.0	0.634463
2	Malvern	43.806686	-79.194353	53425.0	0.456684
3	Highland Creek	43.784535	-79.160497	87321.0	0.680327
4	Flemingdon Park	43.725900	-79.340923	43511.0	1.048648

Fig 1: Dataframe after Data Cleaning

3. Exploratory Data Analysis

3.1 Folium Mapping

The folium library was called to help visualize, geographically, the location of each neighbourhood centered around Toronto.

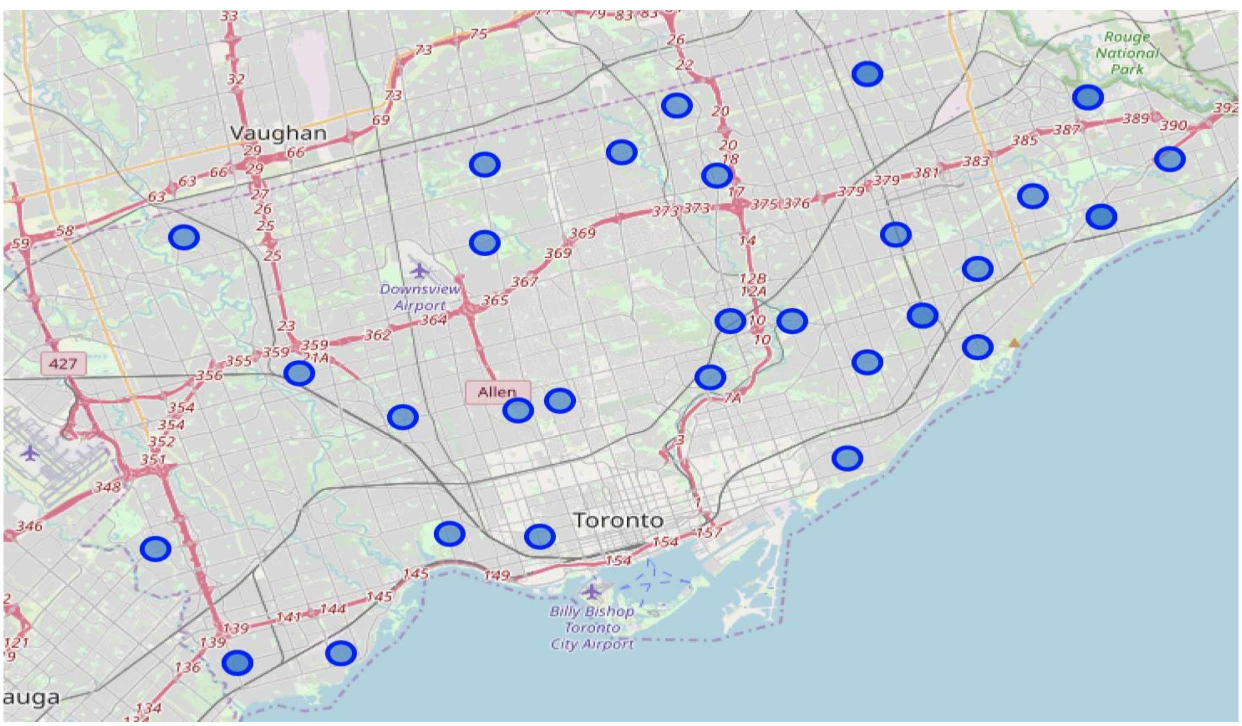

Fig 2: Neighbourhoods marked on a map of Toronto

3.2 Frequency Distribution of Asian Restaurants

Using the Foursquare API's explore function, we could return the number of Asian restaurants located in each neighbourhood. By calculating the mean respectively, it can give us a better understanding on the frequency of occurrence in each neighbourhood. The argument for the use of frequency of Asian restaurants is that I hypothesize that there would be a correlation between the number of Asian restaurants and competition. The higher the number of Asian restaurants in a neighbourhood, the stronger the competition. The assumption of our analysis is that the barrier of entry to establish a new restaurant in a competitive market is high as existing Asian restaurants may have the competitive advantage of brand loyalty.

Though, counterintuitively, the presence of Asian restaurants may even be an indicator of demand for Asian cuisine; the presence of competition may even incentivize innovation to reduce cost and increase productivity.

Hence, it would be sound to establish business operation in a neighbourhood that consists of number of restaurants around the median value.

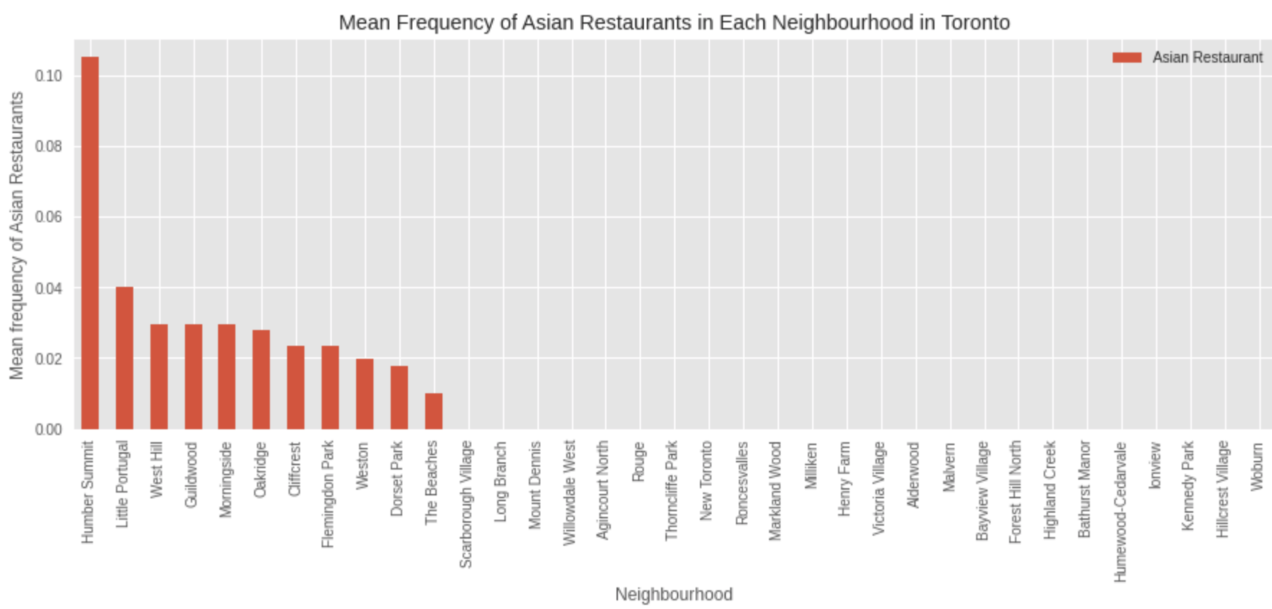

Fig 3: Mean frequency distribution of Asian restaurants in each neighbourhood

3.3 Distribution of Southeast Asians

I hypothesize that there would too exist a linear relationship between the population of a specific ethnic group and the demand of its respective cultural cuisine. Hence, it would only be sound for our clients to carry out business operations in neighbourhoods that are relatively more densely populated with Southeast Asians.

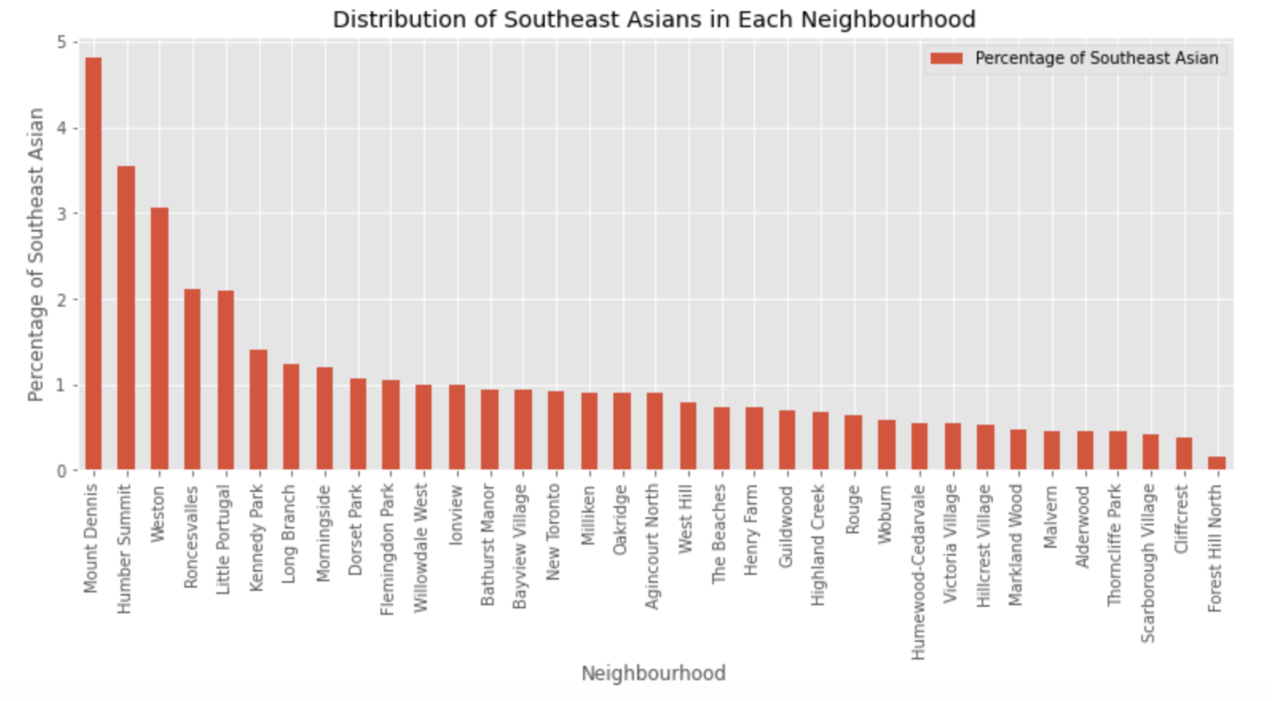

Fig 4: Distribution of Southeast Asians in each Neighbourhood by percentage

3.4 Distribution of Median Household Income

As the franchised Asian restaurant could be categorized as casual dining, the target audience is more geared towards the middle class. As can be inferred from the bar chart below, neighbourhoods distributed towards around the mean can readily afford and indulge themselves in the aforementioned Asian cuisine.

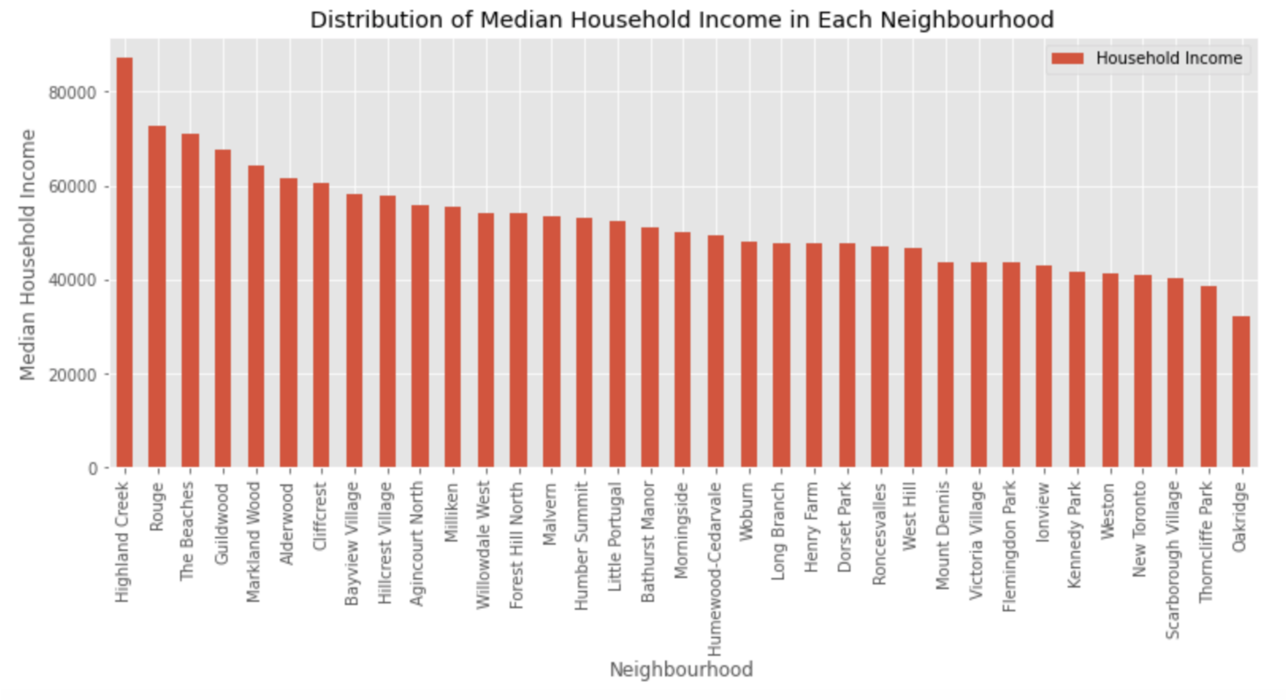

Fig 5: Distribution of Median Household Income in each Neighbourhood

4. Predictive Modeling

4.1 Data Pre-processing

To help mathematical-based algorithms--like our k-Means algorithm in this case---to interpret features with different magnitudes and distributions equally, we will have to normalize our data; as these feature columns are different in scale, we will standardize the values to a common scale. One approach of data normalization is *StandardScaler*.

	Household Income	% Southeast Asian	No. of Asian Restaurants
0	-0.763184	-0.581951	-0.501692
1	1.891815	-0.485201	-0.501692
2	0.121968	-0.672330	-0.501692
3	3.220823	-0.436926	-0.501692
4	-0.784394	-0.049234	0.647782

Fig 6: Dataframe of standardised values across all features

4.2 k-Means Clustering

Before we fit the feature values into our model, we have to pre-assign the number of clusters the algorithm should label. In order to identify the optimal number clusters to use, a range of 3 to 10 clusters were used, then the squared error calculated respectively were used as metrics of their performances.

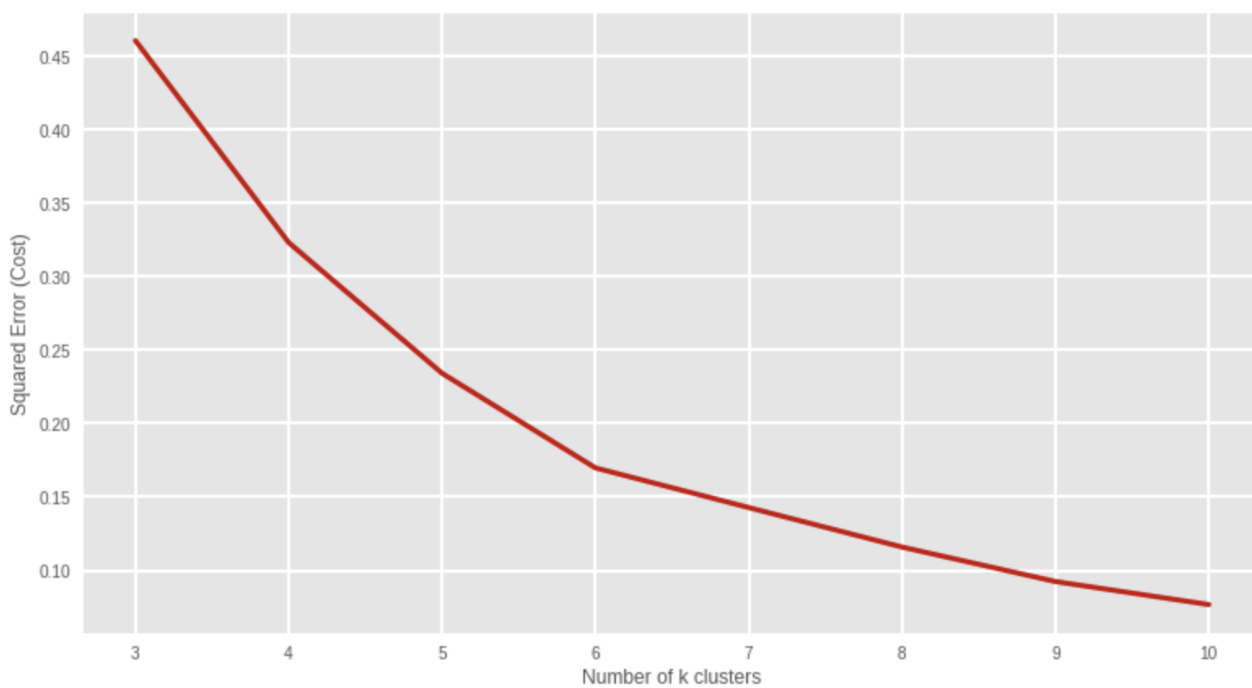

Fig 7: Relationship between the Number of k Clusters and their corresponding Squared Errors

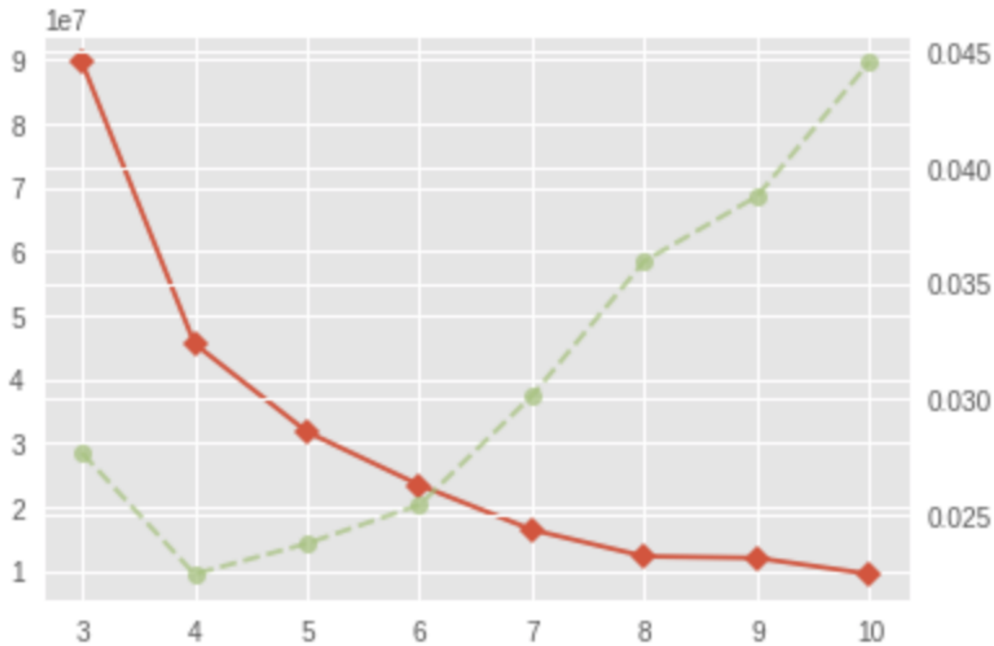

Fig 8: Using K Elbow Visualizer to identify the elbow point for the optimal k value

An analysis using K Elbow Visualizer and Squared error for each k value evident shows that k = 6 would the best value.

After identifying the number of clusters, we will fit the standardized feature values into our k-Means algorithm. The results will be clusters of neighbourhoods of similar characteristics.

4.2.1 Cluster Labels

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Southeast Asian	No. of Asian Restaurants
1	0	Rouge	43.806686	-79.194353	72784.0	0.634463	-0.501692
3	0	Highland Creek	43.784535	-79.160497	87321.0	0.680327	-0.501692
7	0	Guildwood	43.763573	-79.188711	67678.0	0.705859	0.952055
10	0	The Beaches	43.676357	-79.293031	70957.0	0.741874	-0.007418

Fig 9.1: Cluster Label 0

Cluster Label 0:

- HIGH Spending Power
- MID Percentage of Target Customers
- MID Number of Competitors

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Southeast Asian	No. of Asian Restaurants
0	1	Victoria Village	43.725882	-79.315572	43743.0	0.542547	-0.501692
5	1	Humewood-Cedarvale	43.693781	-79.428191	49252.0	0.556909	-0.501692
11	1	Woburn	43.770992	-79.216917	47908.0	0.588950	-0.501692
13	1	Bathurst Manor	43.754328	-79.442259	51076.0	0.945001	-0.501692
14	1	Thorncliffe Park	43.705369	-79.349372	38645.0	0.450066	-0.501692
15	1	Scarborough Village	43.744734	-79.239476	40181.0	0.418560	-0.501692
16	1	Henry Farm	43.778517	-79.346556	47659.0	0.731413	-0.501692
18	1	Ionview	43.727929	-79.262029	42971.0	0.989664	-0.501692
19	1	Kennedy Park	43.727929	-79.262029	41776.0	1.401624	-0.501692
29	1	Roncesvalles	43.648960	-79.456325	46883.0	2.103646	-0.501692
32	1	New Toronto	43.605647	-79.501321	40859.0	0.915991	-0.501692
34	1	Long Branch	43.602414	-79.543484	47680.0	1.239587	-0.501692

Fig 9.2: Cluster Label 1

Cluster Label 1:

- LOW Spending Power
- MID Percentage of Target Customers
- LOW Number of Competitors

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Southeast Asian	No. of Asian Restaurants
4	2	Flemingdon Park	43.725900	-79.340923	43511.0	1.048648	0.647782
8	2	Morningside	43.763573	-79.188711	50069.0	1.203094	0.952055
9	2	West Hill	43.763573	-79.188711	46803.0	0.784901	0.952055
17	2	Little Portugal	43.647927	-79.419750	52519.0	2.088823	1.475403
21	2	Oakridge	43.711112	-79.284577	32079.0	0.902853	0.871291
26	2	Dorset Park	43.757410	-79.273304	47630.0	1.059873	0.380940

Fig 9.3: Cluster Label 2

Cluster Label 2:

- LOW Spending Power
- HIGH Percentage of Target Customers
- HIGH Number of Competitors

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Southeast Asian	No. of Asian Restaurants
24	3	Mount Dennis	43.691116	-79.476013	43790.0	4.818657	-0.501692
25	3	Weston	43.706876	-79.518188	41356.0	3.056914	0.467472

Fig 9.4: Cluster Label 3

Cluster Label 3:

- LOW Spending Power
- HIGH Percentage of Target Customers
- MID Number of Competitors

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Southeast Asian	No. of Asian Restaurants
2	4	Malvern	43.806686	-79.194353	53425.0	0.456684	-0.501692
6	4	Markland Wood	43.643515	-79.577201	64297.0	0.473754	-0.501692
12	4	Hillcrest Village	43.803762	-79.363452	57682.0	0.531475	-0.501692
20	4	Bayview Village	43.786947	-79.385975	58028.0	0.934754	-0.501692
23	4	Cliffcrest	43.716316	-79.239476	60384.0	0.376530	0.647782
27	4	Forest Hill North	43.696948	-79.411307	53978.0	0.156177	-0.501692
28	4	Willowdale West	43.782736	-79.442259	54226.0	1.003779	-0.501692
30	4	Agincourt North	43.815252	-79.284577	55893.0	0.893072	-0.501692
31	4	Milliken	43.815252	-79.284577	55464.0	0.903206	-0.501692
33	4	Alderwood	43.602414	-79.543484	61402.0	0.456280	-0.501692

Fig 9.5: Cluster Label 4

Cluster Label 4:

- MID Spending Power
- MID Percentage of Target Customers
- LOW Number of Competitors

	Cluster Label	Neighbourhood	Latitude	Longitude	Household Income	Percentage of Southeast Asian	No. of Asian Restaurants
22	5	Humber Summit	43.756303	-79.565963	53272.0	3.543814	4.701191

Fig 9.6: Cluster Label 5

Cluster Label 5:

- MID Spending Power
- HIGH Percentage of Target Customers
- HIGH Number of Competitors

5. Conclusions

In this study, I have labelled the neighbourhoods corresponding to their characteristics---spending power, percentage of target customers, and number of competitors. The most promising group of neighbourhoods for opening a Asian Restaurant, with a niche in Southeast Asian cuisine, appears to be 'Cluster Label 0'.

The higher spending power of the neighbourhoods in this cluster allow them to readily afford the slightly upscaled prices of the client's Asian restaurant menu.

The average distribution of percentage of target customers---Southeast Asians---indicates a relatively reasonable demand Southeast Asian.

The number of competitors is not significant yet adequate enough to be a good indicator demand for Asian cuisine.

Our client could more specifically consider Guildwood as a location of establsihment for optimal results. However, wherever there is a shift in dynamic of business demands, we could always target different clusters of neighborhoods. Case in point, if the client has plans to expand a well-established franchised restaurant, neighbourhoods in 'Cluster Label 5' would be the optimal location---under the assumption that the aforementioned restaurant has a competitive advantage of brand loyalty against the high number of competitors in that neighbourhood.

In conclusion, the extensive analysis above would greatly increase the likelihood of the restaurant's success. Similarly, we can use this project to analyze interchangeable scenarios, such as opening a restaurant of different cuisines.

6. Future Directions

I think the model could use more improvements on capturing restaurants' individual traits. For example, two restaurants might have similar number of competitors, but one might have a smaller geographical radius while the other might a bigger radius. Another example is that restaurants which are more densely populated with Southeast Asians may not crave for their Southeast Asian cuisine any more than their ethnical counterparts. More data, especially data of different types, would help improve model performances significantly.