K-Means Clustering: Asian Cuisine Restaurant Startup Location Recommendation In Toronto

1st Zhehao Xu
Applied Data Science Course Student
z352xu@uwaterloo.ca

I. Introduction

It was sometimes confusing and frustrating when you were trying to start up a new restaurant business in Toronto but had not a clue where should your new restaurant locate at that could maximize your possibility to success and make a profit from it. In this capstone project, a closer analysis of the best locations for starting up a new Asian cuisine restaurant was conducted using data from multiple sources such as Foursquare, Kaggle. Insights would be provided to those who would like to start an Asian restaurant business in Toronto as the final delivery of this project.

II. Data

The data this project needed was mainly focusing on factors that could influence the success of a restaurant including:

- Competition
- Area Average Income
- Target Customers
- Safety

Each above feature would be explained here why they were chosen to evaluate a place if it was an ideal candidate. The 'Safety' here was considering if the location had a low crime rate or not as an unsafe area may have a greater chance of getting robbed and this section of data was acquired from the robbery dataset on Toronto public safety data portal website. 'Area Average Income' was the average income in each neighborhood as it indicated how much customers could afford as a large portion of target customers would live nearby in the same area. 'Competition' was a measure of how many similar Asian cuisine restaurants were in that neighborhood that may compete with. 'Target Customer' was the group of customers that the Asian restaurant aimed to attract. It was important that you open the restaurant in a place with a large number of potential customers. In this case, Asians would be a good targeting group as people tended to have meals that they were familiar with.

III. Methodology

In order to reach the goal of providing valuable recommendations for new Asian cuisine restaurants, A formal comprehensive dataset must be constructed to drive insights which required multiple data sources as each one of them contributed single or multiple features that may help in evaluating a location following the above disciplines discussed in the data section.

A. Data Collecting & Cleaning

1) Neighborhood: The initial step of constructing dataset was to confirm the primary key for it which in this case would be names of each neighborhood in Toronto. By searching on the Internet, there were in total 140 neighborhoods in the city of Toronto. Original plan was to use geopy to find the latitudes and longitudes for all those neighborhoods. However, after discovering some neighborhood could not be located this way, an alternative solution was to import the neighbourhoods dataset from city of Toronto open data website which contained all 140 neighborhoods along with their latitudes and longitudes.Since the project did not interest in neighborhoods' geometry data, only name, latitude, longitude for each neighborhood was remained.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude
0	Wychwood	43.676919	-79.425515
1	Yonge-Eglinton	43.704689	-79.403590
2	Yonge-St.Clair	43.687859	-79.397871
3	York University Heights	43.765736	-79.488883
4	Yorkdale-Glen Park	43.714672	-79.457108
...	...	...	...
135	Kennedy Park	43.725556	-79.260382
136	Kensington-Chinatown	43.653554	-79.397240
137	Kingsview Village-The Westway	43.698993	-79.547863
138	Kingsway South	43.653520	-79.510577
139	L'Amoreaux	43.795716	-79.314084

140 rows × 3 columns

Fig. 1: Neighborhood Data from City of Toronto Open Data

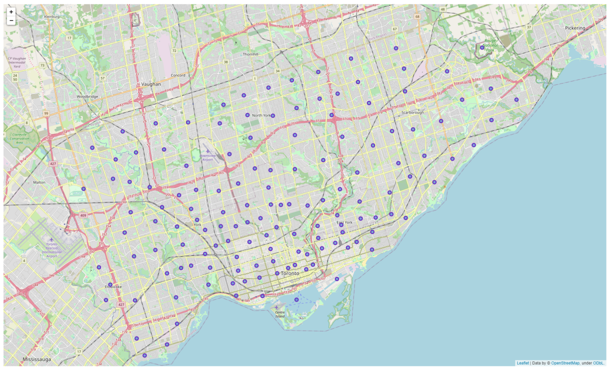

Fig. 2: Neighborhoods on Map

2) Competition: The competition feature evaluated how intense that similar Asian cuisine restaurants were competing with each other in the same neighborhood area. This feature had a float data type ranging from 0 to 1. The higher the value was, the more intense the competition was. It was calculated based on the number of Asian restaurants divided by the maximum number of Asian restaurants in all neighborhoods. The data was initially collected from Four Square by querying a list of Asian restaurants for each neighborhood according to their latitudes and longitudes. The search was limited to 100 rows per query.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Yonge-Eglinton	43.704689	-79.40359	Sake Bar Kushi	43.704923	-79.406954	Japanese Restaurant
1	Yonge-Eglinton	43.704689	-79.40359	Mandarin Buffet	43.705819	-79.398051	Chinese Restaurant
2	Yonge-Eglinton	43.704689	-79.40359	Teriyaki Experience	43.706748	-79.398564	Japanese Restaurant
3	Yonge-Eglinton	43.704689	-79.40359	Lemongrass	43.706764	-79.398521	Asian Restaurant
4	Yonge-Eglinton	43.704689	-79.40359	Kanda Izakaya	43.703606	-79.397896	Japanese Restaurant

Fig. 3: List of Venues from Four Square

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Restaurants
0	Wychwood	43.676919	-79.425515	0
1	Yonge-Eglinton	43.704689	-79.403590	9
2	Yonge-St.Clair	43.687859	-79.397871	13
3	York University Heights	43.765736	-79.488883	4
4	Yorkdale-Glen Park	43.714672	-79.457108	4
...	...	...	...	...
135	Kennedy Park	43.725556	-79.260382	0
136	Kensington-Chinatown	43.653554	-79.397240	100
137	Kingsview Village-The Westway	43.698993	-79.547863	1
138	Kingsway South	43.653520	-79.510577	0
139	L'Amoreaux	43.795716	-79.314084	7

140 rows × 4 columns

Fig. 4: Restaurants in Each Neighborhood

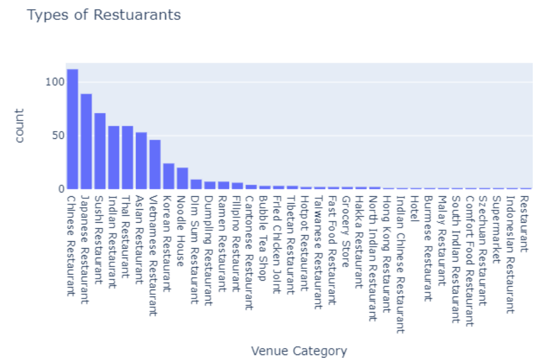

Fig. 5: Common Types of Asian Cuisine Restaurants

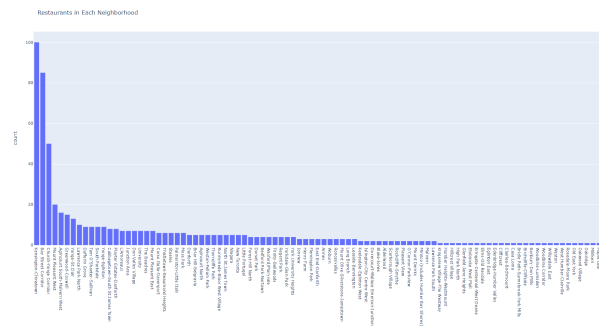

Fig. 6: Count of Asian Cuisine Restaurants by Neighborhoods

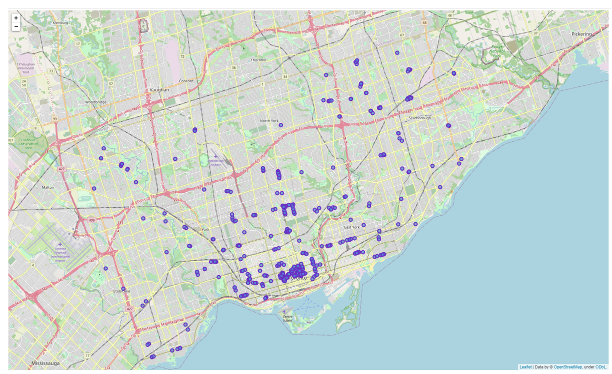

Fig. 7: Asian Cuisine Restaurants on Map

3) Area Average Income: Another feature that the target audience of this project may concern about was that the price setting for the food. Should it be luxury and expensive or cheap and lower quality? It really depended on the income of the neighborhood the restaurant was located at because customers that attended regularly were those who lived nearby as many people did not willing to drive for a long distance for a meal. It was essential to know the economic condition of the neighborhood before making any critical decision. This section of data was obtained in the Toronto 2016 census profile where there was a set of individual average income grouped by neighborhoods.

	Neighborhood	Average Income($)
0	Agincourt North	30414
1	Agincourt South-Malvern West	31825
2	Alderwood	47709
3	Annex	112766
4	Banbury-Don Mills	67757
...	...	...
135	Wychwood	54460
136	Yonge-Eglinton	89330
137	Yonge-St.Clair	114174
138	York University Heights	29958
139	Yorkdale-Glen Park	38527

140 rows × 2 columns

Fig. 8: Area Average Income for Each Neighborhood

4) Target Customers: If the customers weren't been taken into consideration when trying to find the ideal location candidate, then the outcome would miss a huge part of the puzzle. In this case, the new Asian restaurant would much likely to attract customers whose origin was also from Asia whether it was south-eastern, southern, western Asia. Therefore, the population of Asians was extracted from the Toronto 2016 census profile and combined it with the total population in each neighborhood to find the density of Asian people for each area. As discussed before, the higher the density was, the more target customers were.

	Neighborhood	Total Asian Population	Population	Asian Population Percentage
0	Agincourt North	22290	29113	0.765637
1	Agincourt South-Malvern West	16010	23757	0.673907
2	Alderwood	990	12054	0.082130
3	Annex	4660	30526	0.152657
4	Banbury-Don Mills	9090	27695	0.328218
...	...	...	...	...
135	Wychwood	1475	14349	0.102795
136	Yonge-Eglinton	2080	11817	0.176018
137	Yonge-St.Clair	1625	12528	0.129709
138	York University Heights	8080	27593	0.292828
139	Yorkdale-Glen Park	2395	14804	0.161781

140 rows × 4 columns

Fig. 9: Population in Each Neighborhood

5) Safety: Safety was an important feature to look at as if the person could safely open the restaurant and not worried about getting robbed. Even more, they could extend the restaurant opening hours when it was located in a low crime rate area to midnight or late at night. In this way, a dataset of robbery records was extracted from the "Toronto Police Safety Service" open data portal which recorded all robbery events caught from 2014 to 2019. Considering that was a large amount of data, this project would only examined all cases that happened in 2019.

	Neighborhood	Crime Cases
0	Church-Yonge Corridor	143
1	Moss Park	137
2	Bay Street Corridor	123
3	Waterfront Communities-The Island	95
4	York University Heights	79
...	...	...
135	Rustic	3
136	Woodbine-Lumsden	3
137	Humber Heights-Westmount	2
138	Leaside-Bennington	2
139	Centennial Scarborough	1

140 rows × 2 columns

Fig. 12: Crime Cases in Each Neighborhood

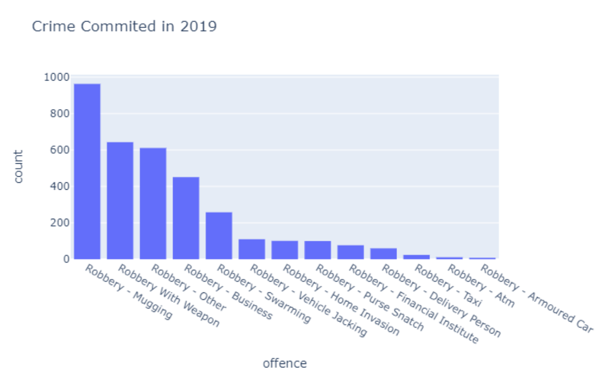

Fig. 13: Common Types of Robbery in 2019

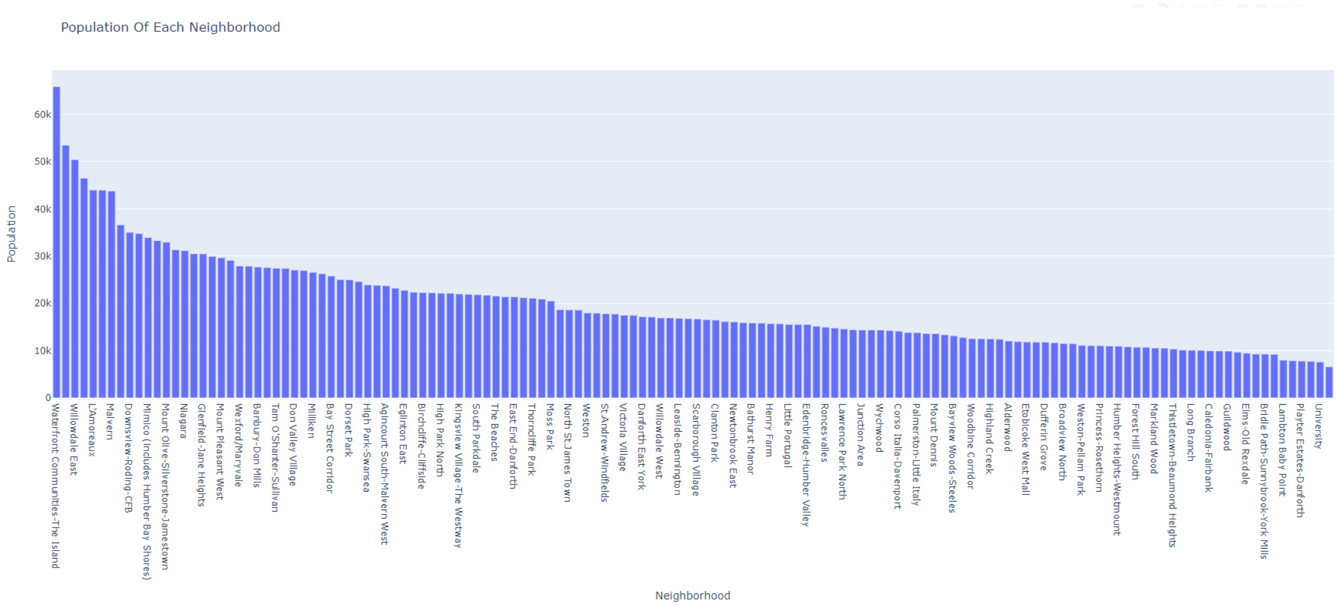

Fig. 10: Total Population in Each Neighborhood

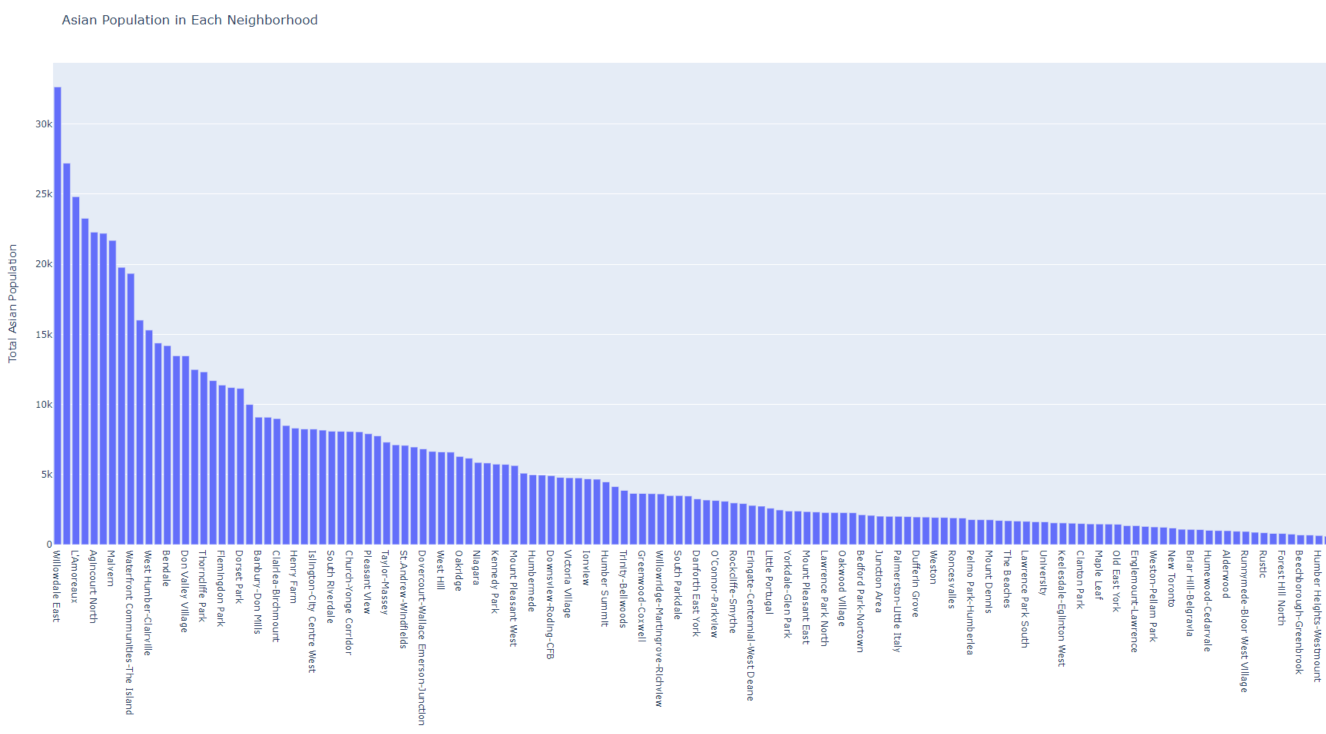

Fig. 11: Asian Population in Each Neighborhood

B. Dataset Combining & Finalization

After gathering all the required features, a final dataset could be obtained by combing all the above sub-dataset into one. Merging all datasets on the 'Neighborhood' column. Features excluding average income were feature scaled ranging from 0 to 1 for a better understanding. The finalized dataset could be used for further analysis in order to drive insights into which neighborhoods were good candidates and which were not.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Competition	Area Average Income	Target Customers	Safety
0	Wychwood	43.676919	-79.425515	0.00	54460	0.102795	0.937063
1	Yonge-Eglinton	43.704689	-79.403590	0.09	89330	0.176018	0.860140
2	Yonge-St.Clair	43.687859	-79.397871	0.13	114174	0.129709	0.972028
3	York University Heights	43.765736	-79.488883	0.04	29958	0.292828	0.447552
4	Yorkdale-Glen Park	43.714672	-79.457108	0.04	38527	0.161781	0.706294
...	...	...	...	...	...	...	...
135	Kennedy Park	43.725556	-79.260382	0.00	30974	0.335222	0.811189
136	Kensington-Chinatown	43.653554	-79.397240	1.00	37422	0.455001	0.678322
137	Kingsview Village-The Westway	43.698993	-79.547863	0.01	36674	0.231364	0.895105
138	Kingsway South	43.653520	-79.510577	0.00	144642	0.080897	0.916084
139	L'Amoreaux	43.795716	-79.314084	0.07	31826	0.564067	0.580420

140 rows × 7 columns

Fig. 14: Final Dataset Overview

C. K-Means Clustering

1) Feature Normalization: Since the task for this project was to find ideal neighborhoods to start an Asian cuisine restaurant, unsupervised machine learning techniques could be applied here as neighborhoods could be clustered into different groups while members in the same group shared similar patterns that may not be very obvious at the first glance. Those four features were firstly normalized to have a mean of 0 and a standard deviation of 1. The formula below was the Z Score normalization while each sample was subtracted by its mean μ and then divided by its standard deviation σ.

$$x_{new} = \frac{x_{old} - \mu}{\sigma} \tag{1}$$

The reason for that was to minimize the distance calculation while K-Means tried to obtain max distance between groups and the least distance between samples within the same group. A bigger value range here may result in a hard time converging and find the optimal cluster numbers.

2) K Parameter Tuning: The optimal number of clusters K was found by iterating through a set of possible K values and observing the rate of decreasing in the sum of distance. This was also called the elbow point finding as the graph would like an exponentially decreasing curve and ideal K value was the point the decreasing rate changed the most. Observing from the below figure, 4 clusters seemed could obtain a relatively low distance summation and not been overfitting.

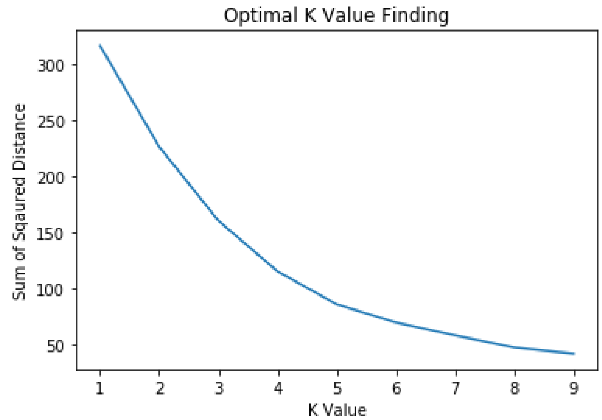

Fig. 15: Optimal K Value Finding

With the optimal number of clusters found, the K means model could be implemented to fit the dataset. The model would assign each neighborhood into one of the four clusters.

IV. Results

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Competition	Area Average Income	Target Customers	Safety	Clusters
0	Wychwood	43.676919	-79.425515	0.00	54460	0.102795	0.937063	1
1	Yonge-Eglinton	43.704689	-79.403590	0.09	89330	0.176018	0.860140	1
2	Yonge-St.Clair	43.687859	-79.397871	0.13	114174	0.129709	0.972028	3
3	York University Heights	43.765736	-79.488883	0.04	29958	0.292828	0.447552	0
4	Yorkdale-Glen Park	43.714672	-79.457108	0.04	38527	0.161781	0.706294	1

Fig. 16: Dataset with Assigned Clusters

A. Cluster 0

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Competition	Area Average Income	Target Customers	Safety	Clusters
61	Waterfront Communities-The Island	43.633880	-79.377202	0.00	70600	0.293417	0.335664	0
81	Bayview Village	43.776361	-79.377117	0.00	52035	0.546597	0.902098	0
82	Bayview Woods-Steeles	43.796802	-79.382118	0.00	47246	0.529117	0.965035	0
24	Newtonbrook East	43.791536	-79.405937	0.00	45212	0.620923	0.916084	0
69	Willowdale West	43.771210	-79.427558	0.00	44576	0.501299	0.804196	0
125	Highland Creek	43.790775	-79.177472	0.00	40972	0.465824	0.937063	0
44	Rouge	43.821201	-79.186343	0.00	39556	0.500581	0.776224	0
62	West Hill	43.767490	-79.176676	0.00	33323	0.241129	0.503497	0
85	Bendale	43.760366	-79.257400	0.00	33256	0.473632	0.720280	0
17	Morningside	43.782399	-79.207041	0.00	32291	0.380693	0.860140	0
135	Kennedy Park	43.725556	-79.260382	0.00	30974	0.335222	0.811189	0
128	Humber Summit	43.758920	-79.556175	0.00	30731	0.359617	0.804196	0
54	Taylor-Massey	43.694998	-79.295901	0.00	30430	0.465472	0.909091	0
129	Humbermede	43.743430	-79.542367	0.00	29528	0.320360	0.811189	0
87	Black Creek	43.764890	-79.521979	0.00	25989	0.283388	0.706294	0
68	Willowdale East	43.770602	-79.401484	0.01	45326	0.647282	0.720280	0
126	Hillcrest Village	43.802988	-79.354804	0.01	40442	0.661391	0.888112	0
98	Clairlea-Birchmount	43.713592	-79.281382	0.01	36232	0.332975	0.741259	0
63	West Humber-Clairville	43.716180	-79.596356	0.01	31771	0.459294	0.510490	0
111	Eglinton East	43.740922	-79.245598	0.01	30033	0.361784	0.811189	0
15	Milliken	43.820691	-79.275009	0.01	28085	0.835466	0.769231	0
119	Glenfield-Jane Heights	43.745636	-79.513465	0.01	27984	0.254173	0.650350	0
30	Oakridge	43.697408	-79.279708	0.01	26793	0.453954	0.776224	0
37	Pleasant View	43.786982	-79.334948	0.02	36346	0.499747	0.895105	0
47	Scarborough Village	43.738652	-79.216813	0.02	32913	0.394044	0.804196	0
12	Malvern	43.803658	-79.222517	0.02	29573	0.495387	0.678322	0
122	Henry Farm	43.771144	-79.341241	0.03	36359	0.528525	0.951049	0
71	Woburn	43.766740	-79.228586	0.03	30878	0.508647	0.531469	0
116	Flemingdon Park	43.715930	-79.332646	0.03	28654	0.518853	0.909091	0
20	Mount Olive-Silverstone-Jamestown	43.746868	-79.587259	0.03	26548	0.408600	0.615385	0
39	Regent Park	43.659992	-79.360509	0.04	34597	0.443395	0.902098	0
105	Dorset Park	43.759274	-79.278908	0.04	31692	0.445547	0.755245	0
3	York University Heights	43.765736	-79.488883	0.04	29958	0.292828	0.447552	0
28	North St.James Town	43.669623	-79.375247	0.05	32648	0.307279	0.769231	0
74	Agincourt North	43.805441	-79.266712	0.05	30414	0.765637	0.755245	0
57	Thorncliffe Park	43.707749	-79.349984	0.05	28875	0.583665	0.909091	0
18	Moss Park	43.656518	-79.367297	0.06	58915	0.177997	0.041958	0
51	Steeles	43.812959	-79.321207	0.06	31786	0.803111	0.937063	0
104	Don Valley Village	43.783294	-79.353644	0.07	37379	0.497579	0.762238	0
139	L'Amoreaux	43.795716	-79.314084	0.07	31826	0.564067	0.580420	0
53	Tam O'Shanter-Sullivan	43.780130	-79.302919	0.09	34200	0.523756	0.790210	0
75	Agincourt South-Malvern West	43.788658	-79.265612	0.16	31825	0.673907	0.797203	0

Fig. 17: Cluster with Label 0

B. Cluster 1

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Competition	Area Average Income	Target Customers	Safety	Clusters
50	St.Andrew-Windfields	43.756246	-79.379037	0.00	100516	0.397204	0.916084	1
38	Princess-Rosethorn	43.666051	-79.544559	0.00	99055	0.137997	0.895105	1
5	Lambton Baby Point	43.657420	-79.496045	0.00	76629	0.074515	0.958042	1
27	North Riverdale	43.671995	-79.351260	0.00	73253	0.162387	0.853147	1
6	Lansing-Westgate	43.754271	-79.424748	0.00	72371	0.294172	0.923077	1
124	High Park-Swansea	43.645065	-79.467872	0.00	71204	0.095507	0.790210	1
130	Humewood-Cedarvale	43.691370	-79.427683	0.00	65274	0.071006	0.930070	1
52	Stonegate-Queensway	43.635518	-79.501128	0.00	64140	0.092611	0.846154	1
14	Markland Wood	43.633542	-79.573432	0.00	62378	0.082907	0.916084	1
0	Wychwood	43.676919	-79.425515	0.00	54460	0.102795	0.937063	1
96	Centennial Scarborough	43.782376	-79.150843	0.00	54045	0.259317	0.993007	1
49	South Riverdale	43.649292	-79.335651	0.00	53803	0.290214	0.755245	1
121	Guildwood	43.748829	-79.195055	0.00	53203	0.136130	0.951049	1
103	Danforth East York	43.689468	-79.331403	0.00	51846	0.189464	0.958042	1
99	Clanton Park	43.741978	-79.446303	0.00	51181	0.091064	0.881119	1
113	Englemount-Lawrence	43.720345	-79.437409	0.00	48134	0.060343	0.888112	1
79	Bathurst Manor	43.764813	-79.456055	0.00	45936	0.123795	0.916084	1
91	Broadview North	43.688825	-79.355630	0.00	44557	0.127402	0.930070	1
70	Willowridge-Martingrove-Richview	43.683645	-79.554221	0.00	44177	0.163161	0.776224	1
34	Parkwoods-Donalda	43.755033	-79.330180	0.00	42516	0.261026	0.748252	1
35	Pelmo Park-Humberlea	43.717515	-79.528282	0.00	39003	0.166014	0.944056	1
60	Victoria Village	43.728489	-79.314874	0.00	35786	0.272130	0.902098	1
94	Caledonia-Fairbank	43.688569	-79.455212	0.00	35112	0.067805	0.874126	1
25	Newtonbrook West	43.785830	-79.431422	0.00	34904	0.337585	0.909091	1
40	Rexdale-Kipling	43.723725	-79.566228	0.00	34418	0.191376	0.839161	1
107	Downsview-Roding-CFB	43.733292	-79.490497	0.00	34168	0.140220	0.552448	1
84	Beechborough-Greenbrook	43.693216	-79.479473	0.00	33829	0.103391	0.916084	1
64	Westminster-Branson	43.778813	-79.452418	0.00	32724	0.133021	0.902098	1
92	Brookhaven-Amesbury	43.701326	-79.485589	0.00	32483	0.164724	0.881119	1
46	Rustic	43.711609	-79.498091	0.00	31800	0.085002	0.979021	1
110	Edenbridge-Humber Valley	43.670886	-79.522458	0.01	101551	0.079820	0.965035	1
78	Banbury-Don Mills	43.737657	-79.349718	0.01	67757	0.328218	0.930070	1
123	High Park North	43.657565	-79.466302	0.01	57465	0.123635	0.909091	1
72	Woodbine Corridor	43.676773	-79.315407	0.01	55199	0.150706	0.923077	1
86	Birchcliffe-Cliffside	43.694682	-79.265093	0.01	54360	0.142658	0.811189	1
32	Old East York	43.696781	-79.335488	0.01	53315	0.157045	0.944056	1
73	Woodbine-Lumsden	43.694107	-79.311164	0.01	47710	0.164654	0.979021	1
114	Eringate-Centennial-West Deane	43.658017	-79.580445	0.01	47002	0.150097	0.916084	1
127	Humber Heights-Westmount	43.692233	-79.522416	0.01	45615	0.058915	0.986014	1
100	Cliffcrest	43.721121	-79.235530	0.01	44718	0.259492	0.825175	1
31	Oakwood Village	43.688566	-79.439785	0.01	39803	0.107496	0.706294	1
115	Etobicoke West Mall	43.645063	-79.568939	0.01	38255	0.191594	0.930070	1
13	Maple Leaf	43.715574	-79.480758	0.01	37108	0.145386	0.965035	1
137	Kingsview Village-The Westway	43.698993	-79.547863	0.01	36674	0.231364	0.895105	1
65	Weston	43.702716	-79.515723	0.01	32997	0.107548	0.818182	1
112	Elms-Old Rexdale	43.721519	-79.548983	0.01	32012	0.181895	0.902098	1
16	Mimico (includes Humber Bay Shores)	43.615924	-79.500137	0.02	54438	0.145890	0.888112	1
132	Islington-City Centre West	43.633463	-79.543317	0.02	52787	0.187308	0.755245	1
88	Blake-Jones	43.676173	-79.337394	0.02	48511	0.217419	0.930070	1
76	Alderwood	43.604937	-79.541611	0.02	47709	0.082130	0.965035	1
29	O'Connor-Parkview	43.706800	-79.312228	0.02	43907	0.168407	0.895105	1
106	Dovercourt-Wallace Emerson-Junction	43.665677	-79.438541	0.02	39740	0.186212	0.671329	1
41	Rockcliffe-Smythe	43.674790	-79.494420	0.02	34059	0.133507	0.923077	1
134	Keelesdale-Eglinton West	43.685727	-79.471437	0.02	33316	0.139718	0.923077	1
19	Mount Dennis	43.688144	-79.499989	0.02	30827	0.130214	0.909091	1
109	East End-Danforth	43.684174	-79.299359	0.03	54324	0.169777	0.811189	1
42	Roncesvalles	43.646123	-79.442992	0.03	50580	0.127221	0.867133	1
11	Long Branch	43.592362	-79.533345	0.03	47384	0.078838	0.790210	1
131	Ionview	43.735364	-79.272470	0.03	31383	0.343450	0.923077	1
117	Forest Hill North	43.704218	-79.428143	0.04	85099	0.061690	0.867133	1
58	Trinity-Bellwoods	43.650176	-79.415342	0.04	50694	0.233450	0.832168	1
4	Yorkdale-Glen Park	43.714672	-79.457108	0.04	38527	0.161781	0.706294	1
67	Wexford/Maryvale	43.748572	-79.298637	0.04	35047	0.254863	0.811189	1
45	Runnymede-Bloor West Village	43.659269	-79.485708	0.05	71888	0.092850	0.881119	1
26	Niagara	43.636681	-79.412420	0.05	70623	0.187781	0.832168	1
102	Danforth	43.684025	-79.329819	0.05	55225	0.183633	0.874126	1
10	Little Portugal	43.647536	-79.430323	0.05	45737	0.166463	0.930070	1
23	New Toronto	43.600589	-79.510358	0.05	44101	0.102940	0.881119	1
89	Briar Hill-Belgravia	43.699024	-79.452851	0.05	34768	0.075752	0.811189	1
66	Weston-Pellam Park	43.673962	-79.460244	0.05	33528	0.113534	0.937063	1
33	Palmerston-Little Italy	43.659157	-79.418409	0.06	58071	0.145740	0.895105	1
101	Corso Italia-Davenport	43.677661	-79.447469	0.06	41717	0.075709	0.804196	1
56	Thistletown-Beaumond Heights	43.737988	-79.563491	0.06	32815	0.298263	0.951049	1
55	The Beaches	43.671050	-79.299601	0.07	92580	0.078824	0.888112	1
21	Mount Pleasant East	43.704852	-79.384924	0.07	85340	0.140089	0.909091	1
59	University	43.662506	-79.401180	0.07	50757	0.212962	0.916084	1
133	Junction Area	43.667890	-79.471440	0.07	49709	0.140958	0.902098	1
36	Playter Estates-Danforth	43.679700	-79.354887	0.08	70831	0.128139	0.930070	1
93	Cabbagetown-South St.James Town	43.667648	-79.366107	0.08	63012	0.132831	0.881119	1
1	Yonge-Eglinton	43.704689	-79.403590	0.09	89330	0.176018	0.860140	1
108	Dufferin Grove	43.655428	-79.437341	0.09	40588	0.167162	0.776224	1
48	South Parkdale	43.636700	-79.439338	0.09	35207	0.159962	0.769231	1
120	Greenwood-Coxwell	43.672612	-79.324318	0.15	46206	0.252827	0.930070	1
22	Mount Pleasant West	43.704435	-79.393360	0.20	57039	0.189999	0.895105	1

Fig. 18: Cluster with Label 1

C. Cluster 2

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Competition	Area Average Income	Target Customers	Safety	Clusters
97	Church-Yonge Corridor	43.659649	-79.379017	0.50	53583	0.257339	0.000000	2
80	Bay Street Corridor	43.657511	-79.385721	0.85	56526	0.483971	0.139860	2
136	Kensington-Chinatown	43.653554	-79.397240	1.00	37422	0.455001	0.678322	2

Fig. 19: Cluster with Label 2

D. Cluster 3

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Competition	Area Average Income	Target Customers	Safety	Clusters
118	Forest Hill South	43.694526	-79.414318	0.00	204521	0.088520	0.888112	3
138	Kingsway South	43.653520	-79.510577	0.00	144642	0.080897	0.916084	3
90	Bridle Path-Sunnybrook-York Mills	43.731013	-79.378904	0.01	308010	0.258472	0.909091	3
43	Rosedale-Moore Park	43.682820	-79.379669	0.01	207903	0.118052	0.867133	3
95	Casa Loma	43.681852	-79.408007	0.01	165047	0.099380	0.944056	3
8	Lawrence Park South	43.717212	-79.406039	0.02	169203	0.109032	0.909091	3
9	Leaside-Bennington	43.703797	-79.366072	0.03	125564	0.118552	0.986014	3
77	Annex	43.671585	-79.404001	0.03	112766	0.152657	0.783217	3
83	Bedford Park-Nortown	43.731486	-79.420227	0.04	123077	0.091453	0.895105	3
7	Lawrence Park North	43.730060	-79.403978	0.10	111730	0.156432	0.965035	3
2	Yonge-St.Clair	43.687859	-79.397871	0.13	114174	0.129709	0.972028	3

Fig. 20: Cluster with Label 3

Fig. 21: Clustering Results

- **Purple** Cluster 0
- **Pink** Cluster 1
- **Yellow** Cluster 2
- **Cyan** Cluster 3

Toronto 140 neighborhoods were clustered into 4 groups with labels from 0 to 3 based on the 4 features: Competition, Area Average Income, Target Customers, Safety. Each cluster was listed above along with any neighborhood inside.

V. Discussion

In the results section, all 4 clusters including their neighborhoods inside were listed and sorted by Competition(ascending), Area Average Income(descending), Target Customers(descending), and Safety(descending). Patterns and discovers were discussed in this section to provide insights on location selection for people who would like to start Asian cuisine restaurant business.

A. Cluster 0

- **LOW** Competition
- **LOW** Area Average Income
- **HIGH** Target Customers
- **MED** Safety

Cluster 0 contained neighborhoods that had very few Asian restaurants which meant not much competition in those regions. However, looking at the area average income, those neighborhoods would not likely to have a rich budget spending on food. In this case, if a new Asian restaurant built in there areas that would very likely had relative cheap but fine quality meals to attract customers as many neighborhoods here had an Asian population percentage around 30% to 50%, the highest area even reached 80% like 'Steeles', which was a lot. For safety, they tended to have moderate safety scores ranging from 70% to 90%.

Summary: Low risk and low-profit option as the restaurant very likely to have a stable income but hard to make huge money in a short period of time.

B. Cluster 1

- **LOW** Competition
- **MED** Area Average Income
- **LOW** Target Customers
- **HIGH** Safety

Cluster 1 also had a low competition score. The area average income was rated to mid-level, it was generally higher than neighborhoods in cluster 0. The difference showed in both the target customers and safety in which neighborhoods in cluster 1 had very few Asian populations and the safety scores were higher than those in cluster 0.

Summary: Medium risk and medium reward option as the new restaurant must have signatures to gain reputation in the neighborhood in order to attract non-Asian people to try the meals while more money earned compared to cluster 0 if succeed

C. Cluster 2

- **HIGH** Competition
- **MED** Area Average Income
- **HIGH** Target Customers
- **LOW** Safety

Neighborhoods in Toronto downtown regions, crowd with other Asian cuisine restaurants, lots of target customers, higher possibility of getting robbed as the safety scores were very low.

Summary: Not very recommended option as the rent fees were high in downtown and competition was intense there, especially for small startup restaurants

D. Cluster 3

- **LOW** Competition
- **HIGH** Area Average Income
- **LOW** Target Customers
- **HIGH** Safety

Cluster 3 as the last cluster had neighborhoods that had very high average incomes which meant customers were more likely to spend more money to get high-quality meals. Business here would not worry about getting robbed as the safety scores were high.

Summary: High risk and high reward option as restaurants with luxury food and nice environment would more likely to succeed here which meant a very high initial cost, but very high potential profit was waiting if succeed

VI. CONCLUSION

In this project, the initial problem statement was first brought onto the table as what were ideal locations to start up a new Asian cuisine restaurant. Then, a set of metrics were established to evaluate if a region was a good candidate or not, and that consisted of 4 features which were competition as it indicated how many similar Asian restaurants were in that area to compete with, area average income as it showed how rich the neighborhood was while higher-income areas had larger potential profit, target customers as it directly related to how many customers the restaurant most likely to receive, and lastly safety as how many crime cases happened in that area which may result in a difference in getting robbed or not. After gathering samples data points for each feature, a finalized dataset was constructed and ready for further analysis. In this case, the project used the K-Means clustering algorithm to group neighborhoods with similar properties and patterns together in order to find ideal candidates. Finally comparing and analyzing neighborhoods in each cluster, different strategies and insights were given to each group as they had their own advantages and disadvantages.

REFERENCES

[1] TorontoPoliceService. (2020, February 24). Robbery 2014 to 2019. Retrieved from Toronto Police Service Public Safety Data Portal: http://data.torontopolice.on.ca/datasets/robbery-2014-to-2019

[2] Social Development, F. A. (2020, June 1). About Neighbourhoods. Retrieved from Toronto Open Data: https://open.toronto.ca/dataset/neighbourhoods/

[3] Statistics Canada. 2017. Toronto, C [Census subdivision], Ontario and Canada [Country] (table). Census Profile. 2016 Census. Statistics Canada Catalogue no. 98-316-X2016001. Ottawa. Released November 29, 2017. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E (accessed June 3, 2020).

[4] FourSquare. (2020, June 2). FourSquare Place API Documentation. Retrieved from FourSquare Developer Portal: https://developer.foursquare.com/docs/places-api/