Using K-means to Find Optimal Restaurant Location

Introduction

In this project, I will use machine learning to give a recommendation to a franchise owner of an Indian restaurant chain on the optimal location to open a new branch in Toronto, Canada. Canada, and especially Toronto, has been known as a top location for Indians to reside. In fact, among all the Indians in Canada, approximately 51% live in the greater Toronto area. This means that having representation of Indian culture is crucial. In addition to that, having an Indian restaurant present in a place with a high density population of Indians will be successful.

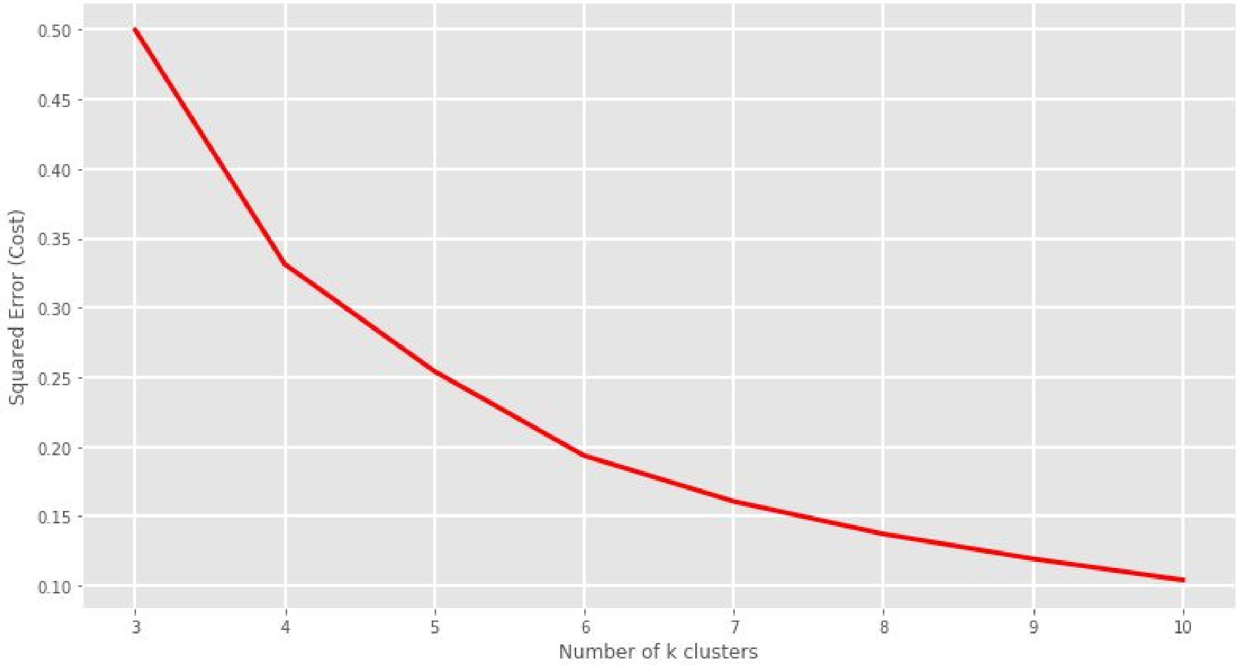
Data

The data I will be using is a table that consists of the different South Asian populations in a given postal code. From here, I will use clustering to find a cluster that has a high density of South Asians and will conclude that that will be the optimal location to house a new branch. This will also, obviously, require location and geographical data, as well as density of existing Indian restaurants nearby an area. This will be done using the Foursquare API.

Methodology

On a high level, the methodology used in order to complete this project is as follows. Find areas with a high Indian population and a low density of existing Indian restaurants. Using this information, a conclusion as per where a restaurant should be placed next can be determined. The core tech used in this project is K-means clustering. K means is a un-supervised machine learning clustering algorithm that clusters a group of data points into similar categories. The definition of “similar” can be calculated in many different ways, for example, the euclidean distance between two data points.

When using K-means clustering, one thing that is up to us is the value of ‘k’ or the number of clusters we want to group our items into. The overall goal of clustering is to minimize the intra-group distance and maximize the outer-group distance, meaning we want to form groups with items that are similar with each other, but distinct from other points. The value of ‘k’ can be determined by running the k-means algorithm multiple times with different values of K. From here, we can pick a value that best fits, or minimizes the error. However, one thing to look out for is, inherently, with a larger number of clusters there will be less error. This is why it is important to look out for the ‘elbow point’. An example of the elbow point can be seen below.

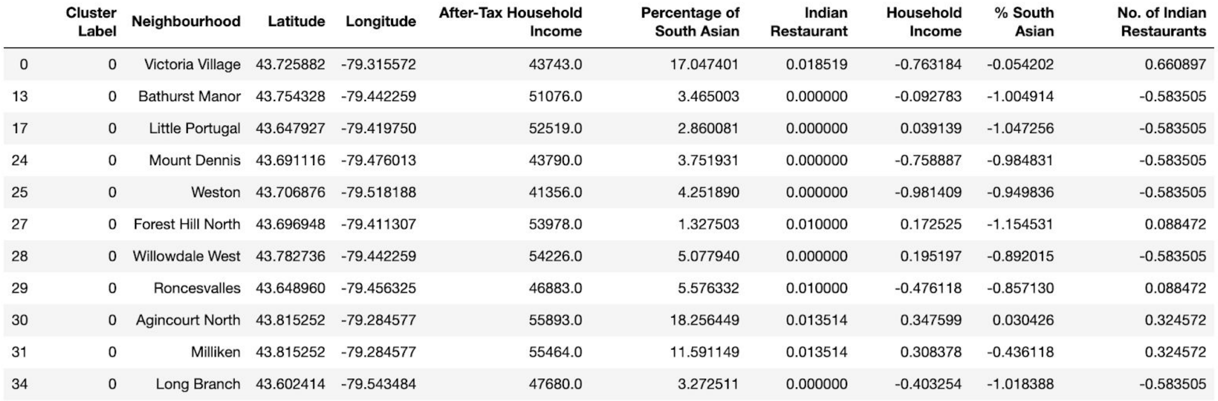


From the graph above, it can be seen that the best number of clusters to pick is 6. From here, we can use the K-means clustering algorithm to find 6 different clusters. Analyzing the clusters will find us the best 'cluster' or location to open a new indian restaurant.

Results

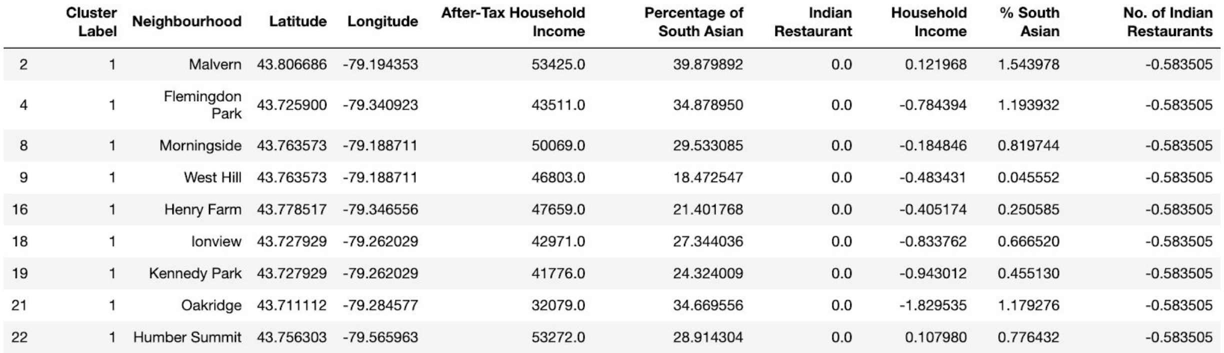
After analyzing the 6 different clusters, these are the items that are in each cluster.

Cluster one:



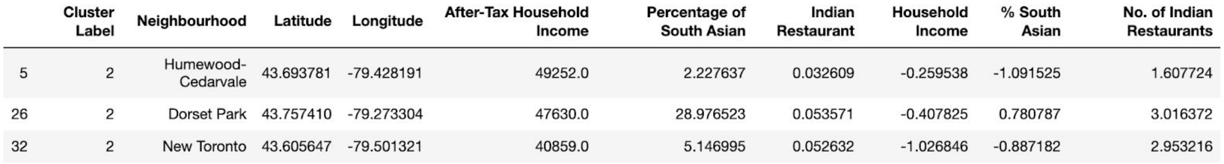
	Cluster Label	Neighbourhood	Latitude	Longitude	After-Tax Household Income	Percentage of South Asian	Indian Restaurant	Household Income	% South Asian	No. of Indian Restaurants
0	0	Victoria Village	43.725882	-79.315572	43743.0	17.047401	0.018519	-0.763184	-0.054202	0.660897
13	0	Bathurst Manor	43.754328	-79.442259	51076.0	3.465003	0.000000	-0.092783	-1.004914	-0.583505
17	0	Little Portugal	43.647927	-79.419750	52519.0	2.860081	0.000000	0.039139	-1.047256	-0.583505
24	0	Mount Dennis	43.691116	-79.476013	43790.0	3.751931	0.000000	-0.758887	-0.984831	-0.583505
25	0	Weston	43.706876	-79.518188	41356.0	4.251890	0.000000	-0.981409	-0.949836	-0.583505
27	0	Forest Hill North	43.696948	-79.411307	53978.0	1.327503	0.010000	0.172525	-1.154531	0.088472
28	0	Willowdale West	43.782736	-79.442259	54226.0	5.077940	0.000000	0.195197	-0.892015	-0.583505
29	0	Roncesvalles	43.648960	-79.456325	46883.0	5.576332	0.010000	-0.476118	-0.857130	0.088472
30	0	Agincourt North	43.815252	-79.284577	55893.0	18.256449	0.013514	0.347599	0.030426	0.324572
31	0	Milliken	43.815252	-79.284577	55464.0	11.591149	0.013514	0.308378	-0.436118	0.324572
34	0	Long Branch	43.602414	-79.543484	47680.0	3.272511	0.000000	-0.403254	-1.018388	-0.583505

Cluster two:



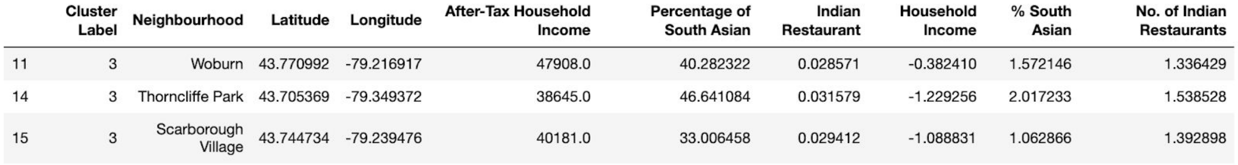
	Cluster Label	Neighbourhood	Latitude	Longitude	After-Tax Household Income	Percentage of South Asian	Indian Restaurant	Household Income	% South Asian	No. of Indian Restaurants
2	1	Malvern	43.806686	-79.194353	53425.0	39.879892	0.0	0.121968	1.543978	-0.583505
4	1	Flemingdon Park	43.725900	-79.340923	43511.0	34.878950	0.0	-0.784394	1.193932	-0.583505
8	1	Morningside	43.763573	-79.188711	50069.0	29.533085	0.0	-0.184846	0.819744	-0.583505
9	1	West Hill	43.763573	-79.188711	46803.0	18.472547	0.0	-0.483431	0.045552	-0.583505
16	1	Henry Farm	43.778517	-79.346556	47659.0	21.401768	0.0	-0.405174	0.250585	-0.583505
18	1	Ionview	43.727929	-79.262029	42971.0	27.344036	0.0	-0.833762	0.666520	-0.583505
19	1	Kennedy Park	43.727929	-79.262029	41776.0	24.324009	0.0	-0.943012	0.455130	-0.583505
21	1	Oakridge	43.711112	-79.284577	32079.0	34.669556	0.0	-1.829535	1.179276	-0.583505
22	1	Humber Summit	43.756303	-79.565963	53272.0	28.914304	0.0	0.107980	0.776432	-0.583505

Cluster three:



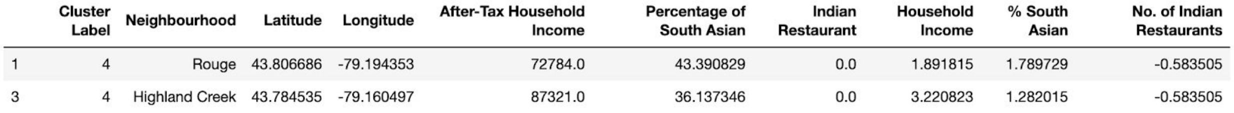
	Cluster Label	Neighbourhood	Latitude	Longitude	After-Tax Household Income	Percentage of South Asian	Indian Restaurant	Household Income	% South Asian	No. of Indian Restaurants
5	2	Humewood-Cedarvale	43.693781	-79.428191	49252.0	2.227637	0.032609	-0.259538	-1.091525	1.607724
26	2	Dorset Park	43.757410	-79.273304	47630.0	28.976523	0.053571	-0.407825	0.780787	3.016372
32	2	New Toronto	43.605647	-79.501321	40859.0	5.146995	0.052632	-1.026846	-0.887182	2.953216

Cluster four:



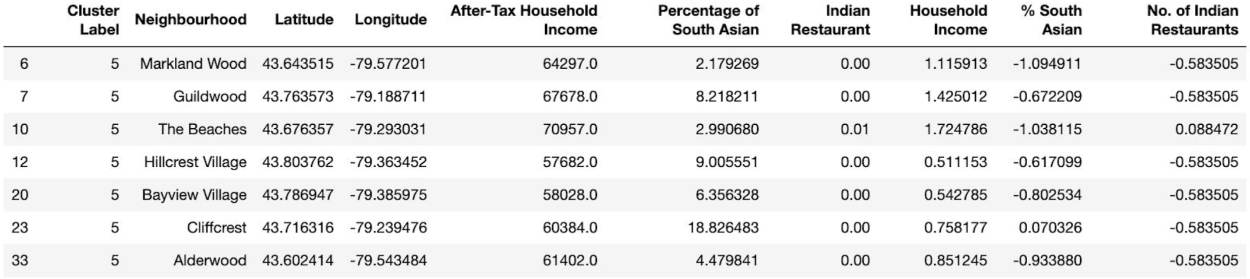
	Cluster Label	Neighbourhood	Latitude	Longitude	After-Tax Household Income	Percentage of South Asian	Indian Restaurant	Household Income	% South Asian	No. of Indian Restaurants
11	3	Woburn	43.770992	-79.216917	47908.0	40.282322	0.028571	-0.382410	1.572146	1.336429
14	3	Thorncliffe Park	43.705369	-79.349372	38645.0	46.641084	0.031579	-1.229256	2.017233	1.538528
15	3	Scarborough Village	43.744734	-79.239476	40181.0	33.006458	0.029412	-1.088831	1.062866	1.392898

Cluster five:



	Cluster Label	Neighbourhood	Latitude	Longitude	After-Tax Household Income	Percentage of South Asian	Indian Restaurant	Household Income	% South Asian	No. of Indian Restaurants
1	4	Rouge	43.806686	-79.194353	72784.0	43.390829	0.0	1.891815	1.789729	-0.583505
3	4	Highland Creek	43.784535	-79.160497	87321.0	36.137346	0.0	3.220823	1.282015	-0.583505

Cluster six:



	Cluster Label	Neighbourhood	Latitude	Longitude	After-Tax Household Income	Percentage of South Asian	Indian Restaurant	Household Income	% South Asian	No. of Indian Restaurants
6	5	Markland Wood	43.643515	-79.577201	64297.0	2.179269	0.00	1.115913	-1.094911	-0.583505
7	5	Guildwood	43.763573	-79.188711	67678.0	8.218211	0.00	1.425012	-0.672209	-0.583505
10	5	The Beaches	43.676357	-79.293031	70957.0	2.990680	0.01	1.724786	-1.038115	0.088472
12	5	Hillcrest Village	43.803762	-79.363452	57682.0	9.005551	0.00	0.511153	-0.617099	-0.583505
20	5	Bayview Village	43.786947	-79.385975	58028.0	6.356328	0.00	0.542785	-0.802534	-0.583505
23	5	Cliffcrest	43.716316	-79.239476	60384.0	18.826483	0.00	0.758177	0.070326	-0.583505
33	5	Alderwood	43.602414	-79.543484	61402.0	4.479841	0.00	0.851245	-0.933880	-0.583505

From analyzing these results, we find that a restaurant placed in cluster five would be most appropriate. The reason for this is because these two areas have a high percentage of South Asian population as well as a low number of already existing Indian Restaurants.

Discussion

Through the K-means clustering algorithm, we determined a cluster that has the optimum values of South Asian population as well as number of existing Indian Restaurants. Based off these results, it would make sense to open a restaurant in this general cluster or area, which consists of the neighborhoods of Rogue and Highland Creek. One obvious improvement that can be made to this project is using a neural network instead of a K means clustering algorithm. The neural network will be more complex and will yield better results. In addition to this, distinctions among Indian restaurants can be further studied. For example, we can look into different kinds of Indian restaurants such as South Indian food, and North Indian food. Similar to this project if we find a dataset on locations with high South Indian population and high North Indian population, we can place these restaurants in those locations accordingly.

Conclusion

In this project, I started out with a goal that aimed to determine the optimum location to open a new Indian restaurant in the city of Toronto, Canada. I chose to do this project based in Toronto due to the high indian population that is present there. Using unsupervised machine learning algorithms we determined the optimal location to place this restaurant by clustering census and geographical data as well as exploring areas with a lot of existing indian restaurants.