Coursera Capstone Project

The Battle of Neighborhoods - Final Report (Week 1 and 2)

Coursera Capstone - REPORT

Content

Introduction Section :

1.1 Discussion of the "backgroung situation" leading to the problem at hand:

1.2 Problem to be resolved

1.3 Audience for this project.

Data Section:

2.1 Data of Current Situation (current residence place)

2.2 Data required to resolve the problem

2.3 Data sources and data manipulation

Methodology section :

3.1 Process steps and strategy to resolve the problem

3.2 Data Science Methods, machine learing, mapping tools and exploratory data analysis.

Results section :

Discussion section :

Conclusion section :

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 South 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 on 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. For returning the number of Asian restaurants in the vicinity of each neighbourhood, we will be utilizing Foursquare API, more specifically, its explore function.

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

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 population across each neighbourhood was calculated by dividing the population of the Southeast Asians demographic by the total population of each neighbourhood

Methodology section :

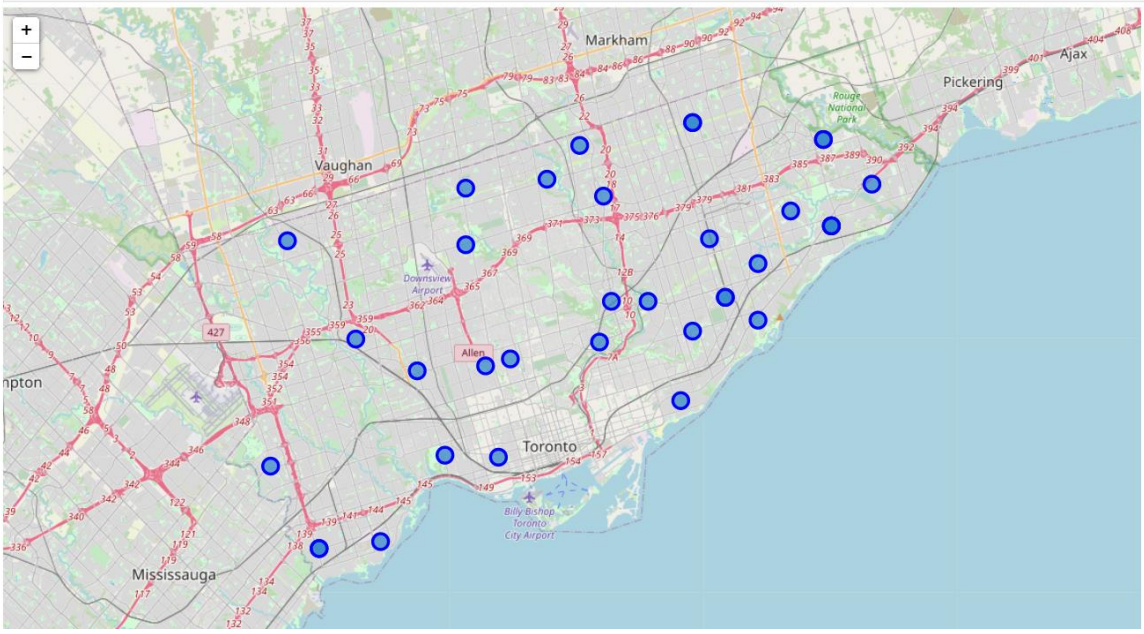
For analysing and predicting location to go to final result we need some important information about data we need to analyse. To analysis data we need to convert it in required format as done below

Out[10]:

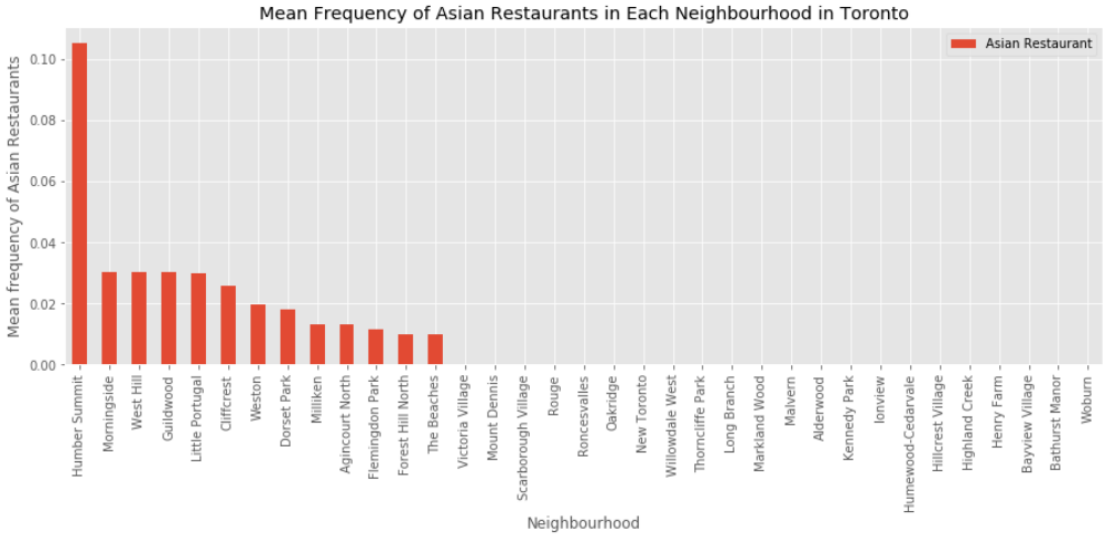
	Neighbourhood	Latitude	Longitude	Pop 20 - 29 years	Household Income	% South Asian
0	Thorncliffe Park	43.705369	-79.349372	2715.0	38645.0	0.466411
1	Rouge	43.806686	-79.194353	6615.0	72784.0	0.433908
2	Woburn	43.770992	-79.216917	8320.0	47908.0	0.402823
3	Malvern	43.806686	-79.194353	6660.0	53425.0	0.398799
4	Highland Creek	43.784535	-79.160497	2045.0	87321.0	0.361373

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

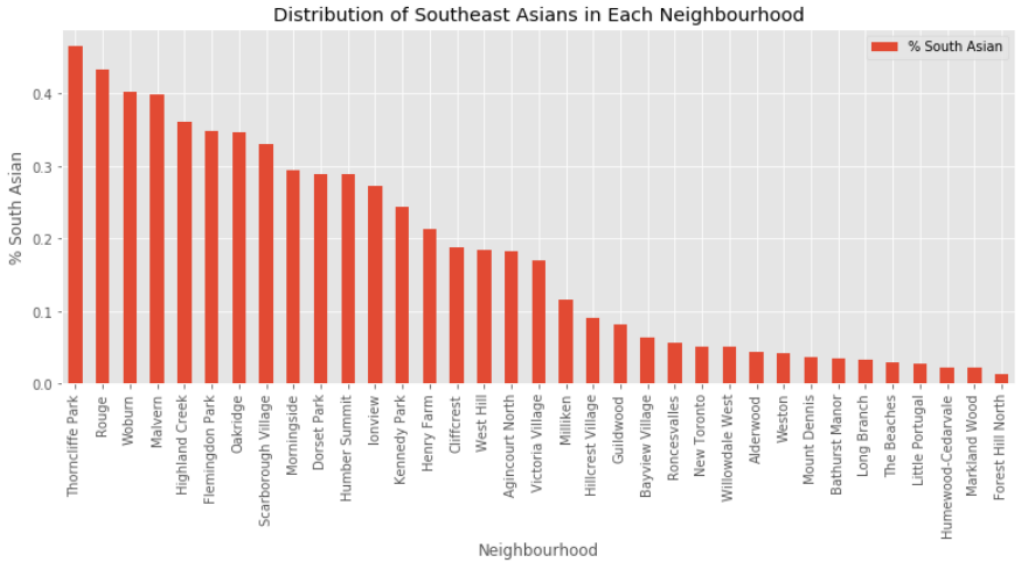
[12]:



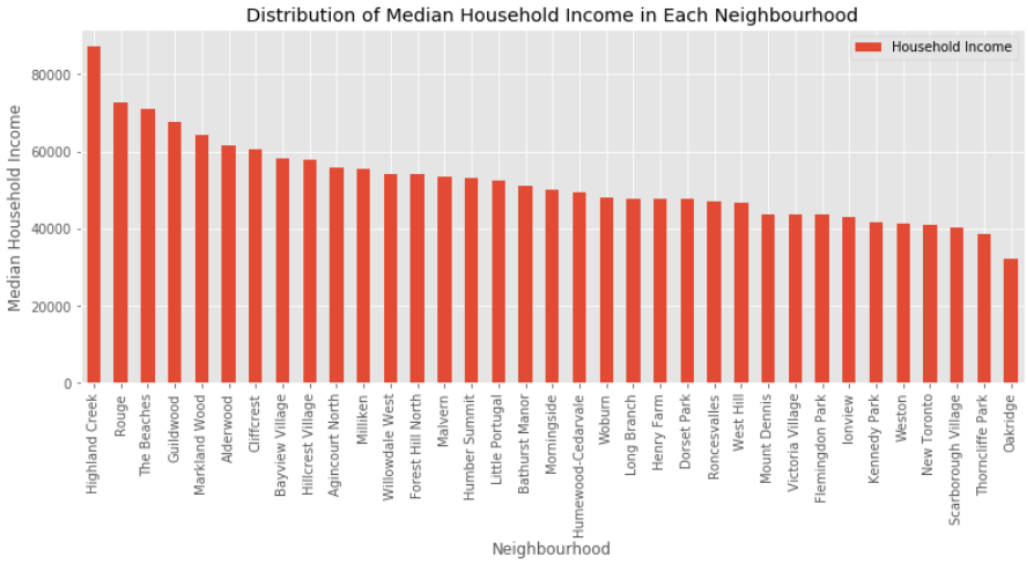
Using the Foursquare API's explore function, we could return the number of Asian restaurants located in each neighbourhood. By calculating the mean respectively, it can give us a better understanding of the frequency of occurrence in each neighbourhood. The argument for the use of frequency of Asian 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 operations in a neighbourhood that consists of a number of restaurants around the median value.



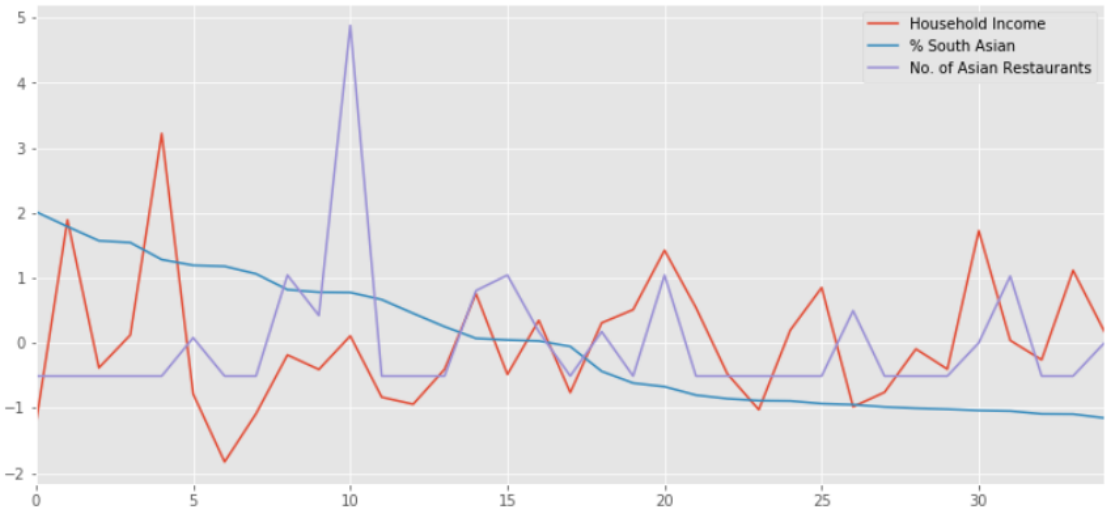
I hypothesize that there would too exist a linear relationship between the population of a specific ethnic group and the demand for its respective cultural cuisine. Hence, it would only be sound for our clients to carry out business operations in neighbourhoods that are relatively more densely populated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



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.

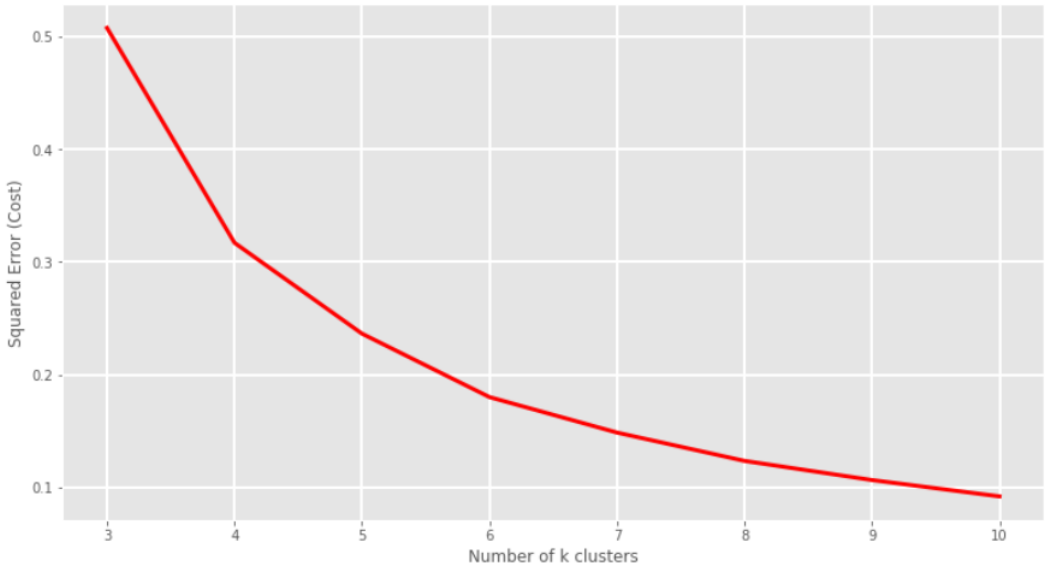


I look wheter any factors has direct relation ship with each other with the line plot mention below but it looks like the don’t have any linear relation with each other



From above graph it looks like there is no direct relationship between any factors mention above.

Now as the above factors don't have any direct correlation with each other it may possible due to different depending factors in different zone. That's why we will use k-means clustering algorithm to cluster location with similar characteristics so we need to find an optimum k value if k value is too big it will be to generalize if its value is too small it will have high error that's why it should be optimum. Before we fit the feature values into our model, we have to pre-assign the number of clusters the algorithm should label. 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



Here by using k=6 as optimum value we are going to make six no. of clusters.

Results:

By clustering algorithm, we get 6 clusters as given below they are group by clusters and by taking their mean value.

[30]:

Cluster Label	Latitude	Longitude	Pop 20 - 29 years	Household Income	% South Asian	Asian Restaurant
0	43.745658	-79.276954	3577.500000	43578.500000	0.331405	0.002968
1	43.702599	-79.432629	2098.000000	62473.200000	0.050023	0.002000
2	43.755066	-79.256359	3040.714286	55544.285714	0.153940	0.024745
3	43.756303	-79.565963	1740.000000	53272.000000	0.289143	0.105263
4	43.690869	-79.453492	2109.000000	47284.300000	0.051145	0.002961
5	43.795611	-79.177425	4330.000000	80052.500000	0.397641	0.000000

Discussion: -

[31]:

	Cluster Label	Neighbourhood	Latitude	Longitude	Pop 20 - 29 years	Household Income	% South Asian	Asian Restaurant
0	0	Thorncliffe Park	43.705369	-79.349372	2715.0	38645.0	0.466411	0.000000
2	0	Woburn	43.770992	-79.216917	8320.0	47908.0	0.402823	0.000000
3	0	Malvern	43.806686	-79.194353	6660.0	53425.0	0.398799	0.000000
5	0	Flemingdon Park	43.725900	-79.340923	3175.0	43511.0	0.348789	0.011494
6	0	Oakridge	43.711112	-79.284577	1645.0	32079.0	0.346696	0.000000
7	0	Scarborough Village	43.744734	-79.239476	2275.0	40181.0	0.330065	0.000000
9	0	Dorset Park	43.757410	-79.273304	3265.0	47630.0	0.289765	0.018182
11	0	Ionview	43.727929	-79.262029	1935.0	42971.0	0.273440	0.000000
12	0	Kennedy Park	43.727929	-79.262029	2225.0	41776.0	0.243240	0.000000
13	0	Henry Farm	43.778517	-79.346556	3560.0	47659.0	0.214018	0.000000

Cluster 0 :-

- **Low** Spending Power
- **High** Percentage of Target Customers
- **Low** Number of Competitors

[32]:

	Cluster Label	Neighbourhood	Latitude	Longitude	Pop 20 - 29 years	Household Income	% South Asian	Asian Restaurant
19	1	Hillcrest Village	43.803762	-79.363452	2195.0	57682.0	0.090056	0.00
21	1	Bayview Village	43.786947	-79.385975	3775.0	58028.0	0.063563	0.00
25	1	Alderwood	43.602414	-79.543484	1380.0	61402.0	0.044798	0.00
30	1	The Beaches	43.676357	-79.293031	2080.0	70957.0	0.029907	0.01
33	1	Markland Wood	43.643515	-79.577201	1060.0	64297.0	0.021793	0.00

Cluster 1: -

- **High** Spending Power
- **Low** Percentage of Target Customers
- **Low** Number of Competitors

[33]:

	Cluster Label	Neighbourhood	Latitude	Longitude	Pop 20 - 29 years	Household Income	% South Asian	Asian Restaurant
8	2	Morningside	43.763573	-79.188711	2845.0	50069.0	0.295331	0.030303
14	2	Cliffcrest	43.716316	-79.239476	1855.0	60384.0	0.188265	0.025641
15	2	West Hill	43.763573	-79.188711	3745.0	46803.0	0.184725	0.030303
16	2	Agincourt North	43.815252	-79.284577	4020.0	55893.0	0.182564	0.013333
18	2	Milliken	43.815252	-79.284577	3970.0	55464.0	0.115911	0.013333
20	2	Guildwood	43.763573	-79.188711	940.0	67678.0	0.082182	0.030303
31	2	Little Portugal	43.647927	-79.419750	3910.0	52519.0	0.028601	0.030000

Cluster 2 :-

- **Mid** Spending Power
- **Mid** Percentage of Target Customers
- **Mid** Number of Competitors

[34]:

	Cluster Label	Neighbourhood	Latitude	Longitude	Pop 20 - 29 years	Household Income	% South Asian	Asian Restaurant
10	3	Humber Summit	43.756303	-79.565963	1740.0	53272.0	0.289143	0.105263

Cluster 3 :-

- **Mid** Spending Power
- **High** Percentage of Target Customers
- **High** Number of Competitors

[35]:

	Cluster Label	Neighbourhood	Latitude	Longitude	Pop 20 - 29 years	Household Income	% South Asian	Asian Restaurant
17	4	Victoria Village	43.725882	-79.315572	2190.0	43743.0	0.170474	0.000000
22	4	Roncesvalles	43.648960	-79.456325	2100.0	46883.0	0.055763	0.000000
23	4	New Toronto	43.605647	-79.501321	1755.0	40859.0	0.051470	0.000000
24	4	Willowdale West	43.782736	-79.442259	3025.0	54226.0	0.050779	0.000000
26	4	Weston	43.706876	-79.518188	2570.0	41356.0	0.042519	0.019608
27	4	Mount Dennis	43.691116	-79.476013	1960.0	43790.0	0.037519	0.000000
28	4	Bathurst Manor	43.754328	-79.442259	1975.0	51076.0	0.034650	0.000000
29	4	Long Branch	43.602414	-79.543484	1520.0	47680.0	0.032725	0.000000
32	4	Humewood-Cedarvale	43.693781	-79.428191	2400.0	49252.0	0.022276	0.000000
34	4	Forest Hill North	43.696948	-79.411307	1595.0	53978.0	0.013275	0.010000

Cluster 4 :-

- **Low** Spending Power
- **Low** Percentage of Target Customers
- **Low** Number of Competitors

[36]:

	Cluster Label	Neighbourhood	Latitude	Longitude	Pop 20 - 29 years	Household Income	% South Asian	Asian Restaurant
1	5	Rouge	43.806686	-79.194353	6615.0	72784.0	0.433908	0.0
4	5	Highland Creek	43.784535	-79.160497	2045.0	87321.0	0.361373	0.0

Cluster 5 :-

- **Very High** Spending Power
- **High** Percentage of Target Customers
- **NO** Competitors

1. From different types of cluster mention above cluster 1 and cluster 4 has low South Asian population which are our target customer's which makes those clusters unfavourable for starting a restaurant.
2. Cluster 0 and Cluster 4 has low spending power which is also an important requirement which is not satisfied so Cluster 0 is also unfavourable
3. If compared Cluster 3 has High no of restaurant also has medium spending power and as competition is also high its hard to stay profitable in competition so Cluster 3 is not suggested to open a restaurant

In Cluster 2 and Cluster 5 if compare cluster 5 has high spending power with high no of South Asian people and low competition which makes Cluster 5 most favourable to start a new Restaurant. But with such high spending power there are no restaurant so its possible people are not interested in going to restaurant located here. So one needs to create a good environment, needs some publicity etc. to start a good profit form it.

Conclusion : -

In this study, I have labelled the neighbourhoods corresponding to their characteristics — spending power, percentage of target customers, and the number of competitors. The most promising group of neighbourhoods for opening an Asian Restaurant, with a niche in South Asian cuisine, appears to be Cluster 5 and Cluster 2

- In Cluster 2 , the medium spending power of the neighbourhoods in this cluster allows them to readily afford prices of the client's Asian restaurant menu.

- The average distribution of the percentage of target customers — the Southeast Asian demographic — indicates a relatively reasonable demand for the Asian cuisine.

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

- In Cluster 5, people have high spending power, no competition and high population but its also may possible that people are not interested in going to restaurant, as its only possibility client should try to start in cluster 5 location. As there is no competition it may result in to high profitability.