The Battle of the Neighbourhoods

Sunil Rao
Mar 8 · 9 min read

IBM/Coursera Applied Data Science Capstone Project by Sunil Rao

Introduction: Business Problem

India is the seventh-largest country by area, the second-most populous country, and the most populous democracy in the world. India has very rich cultural heritage dating back to 6500 BCE. Since India achieved its independence in 1947, the country is making a constant progress in all areas. Indians have traveled across the world and have taken their culture to those countries. Indian diaspora is actively engaged, contributing significantly to local development and relationship with India. Many Indians travel to different places across. Indian cuisine is now popular all over.

Canada is home to the world's tenth largest Indian diaspora. The Indo-Canadian population according to the Census 2016 is 1,374,710 (3.9%). Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. 10.4% of Indian diaspora in Canada, live in the City of Toronto.

There are many Indian restaurants in the City of Toronto to cater to people of South Asian origin, travelers from India and other South Asian countries and people who like Indian Cuisine. As per research by Restaurants

Canada, In 2020, commercial food service sales will improve slightly with a 4.0% nominal increase, and by 2021, combined commercial and non-commercial food service sales are forecast to surpass the $100-billion mark.

Many factors are required to run successful restaurants — a good location, consistent clientele, great chef, great menu, comfortable décor, good customer service, and such.

The objective of this project is to determine suitable locations in Toronto to start a new Indian Restaurant using Machine learning utilizing census, demographic data and location, type information of other businesses.

Sources:

- Wikipedia website (https://en.wikipedia.org/wiki/Indo-Canadians),
- Restaurants Canada website (https://www.restaurantscanada.org/resources/foodservice-industry-forecast/)

Data

The following data sources will be utilized in finding suitable neighbourhood in the City of Toronto.

The Neighbourhood data from
https://open.toronto.ca/dataset/neighbourhoods/
The Neighbourhood Profiles data from
https://open.toronto.ca/dataset/neighbourhood-profiles/
Venue data related to restaurants in the neighbourhoods using Foursquare API https://foursquare.com/

The Neighbourhood data consists of the boundaries of City of Toronto Neighbourhoods. The data is packaged as GeoJSON file with projection as WGS84. The data includes features such as AREA_CODE, AREA_NAME, LATITUDE, LONGITUDE, and geometry. This is used to get a list of neighbourhoods in the City of Toronto and display neighbourhood boundaries in geo-map.

The Neighbourhood Profiles data consists of demographic, social and economic characteristics of the people and households in each neighbourhood. The data includes census, age and sex, families and households, language, immigration and internal migration, ethno-cultural diversity, Aboriginal peoples, housing, education, income, and labour. For the purpose of this project, the following data is used — Population, Average Income.

Foursquare — a location technology platform, provides data related to location and venues across the world. The Foursquare API is used to get a list of Restaurants in each neighbourhood in the City of Toronto. The geo-location data from Neighbourhood data is used to query venues using the Foursquare API. The venue data includes Venue Category, Venue Name, Venue Latitude, and Venue Longitude.

The data from different sources is analyzed to understand the key characteristics (i.e. exploratory data analysis). The demographic data (Population and Average Income) and venues data is merged to create a unified data set. The unified data set is further analyzed using “K-Nearest Neighbor” machine learning algorithm to classify neighbourhoods into different sets based on similarity. The sets are analyzed to find out suitable neighbourhoods to start a new Indian restaurant.

Methodology

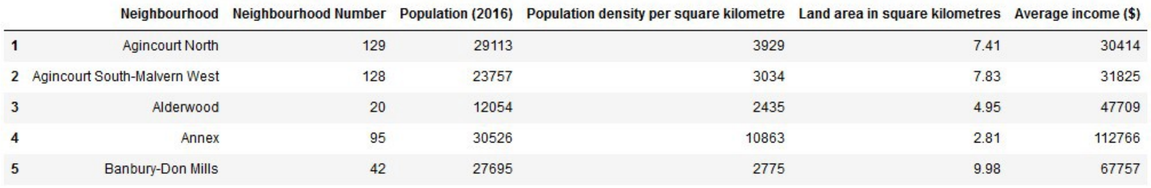
Get Demographic Data of Neighbourhoods in City of Toronto

The City Government of Toronto has published the data about the city collected by its agencies in the Open Data Portal (https://open.toronto.ca/). Through the open data portal, the City government hopes to the City more transparent, accountable, participatory and accessible.

From the Open Data Portal, download the Neighbourhood Profiles data-set (https://open.toronto.ca/dataset/neighbourhood-profiles/). The Neighbourhood Profiles provide a portrait of the demographic, social and economic characteristics of the people and households in each City of Toronto neighbourhood. The data is based on tabulations of 2016 Census of Population data from Statistics Canada. The data-set is in Comma Separated Values format.

From the data-set, extract the relevant fields for this study:

- Neighbourhood (Name)
- Neighbourhood Number
- Population (2016)
- Population density per square kilometre
- Land area in square kilometres
- Average income ($)



	Neighbourhood	Neighbourhood Number	Population (2016)	Population density per square kilometre	Land area in square kilometres	Average income ($)
1	Agincourt North	129	29113	3929	7.41	30414
2	Agincourt South-Malvern West	128	23757	3034	7.83	31825
3	Alderwood	20	12054	2435	4.95	47709
4	Annex	95	30526	10863	2.81	112766
5	Banbury-Don Mills	42	27695	2775	9.98	67757

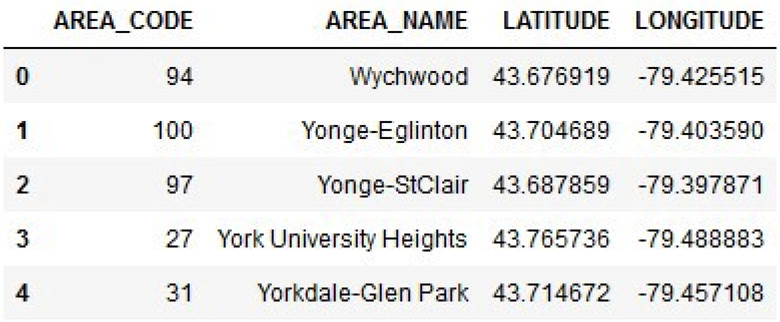
Neighbourhood Profiles Data

Get Boundaries of City of Toronto Neighbourhoods

From the Open Data Portal, download the Neighbourhood data-set (https://open.toronto.ca/dataset/neighbourhoods/). The Neighbourhoods data-set provide the boundary of each Neighbourhood of the City of Toronto. This data can be used to display on a map. The data-set is in GeoJSON format with WGS84 project format.

From the data-set, extract the relevant fields for this study:

- AREA_LONG_CODE
- AREA_NAME
- LONGITUDE
- LATITUDE
- geometry

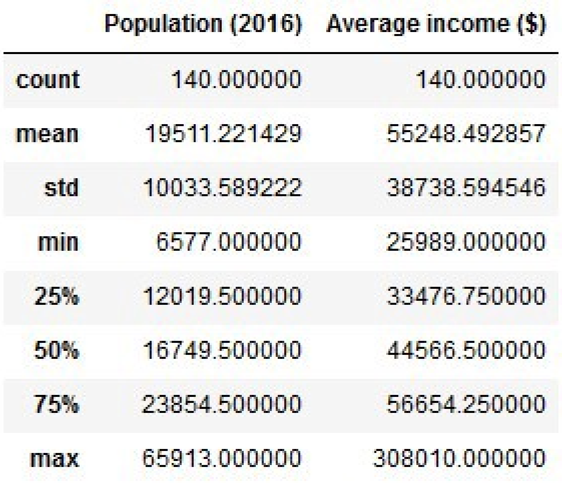


	AREA_CODE	AREA_NAME	LATITUDE	LONGITUDE
0	94	Wychwood	43.676919	-79.425515
1	100	Yonge-Eglinton	43.704689	-79.403590
2	97	Yonge-StClair	43.687859	-79.397871
3	27	York University Heights	43.765736	-79.488883
4	31	Yorkdale-Glen Park	43.714672	-79.457108

Neighbourhood Boundaries Data

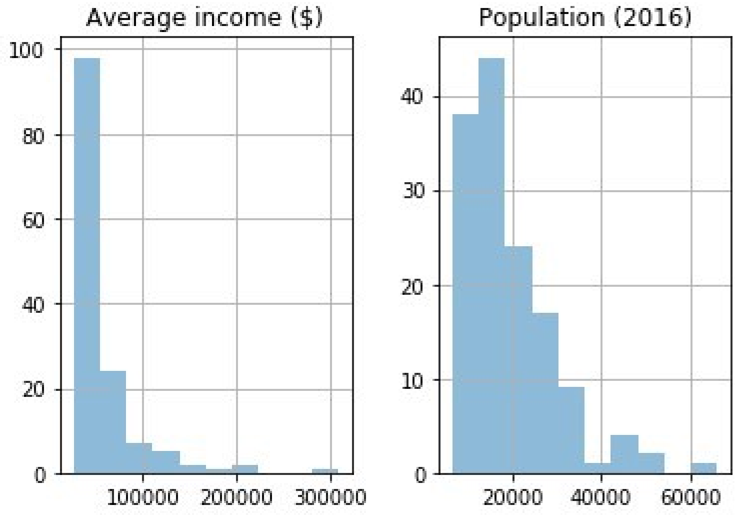
Exploratory Data Analysis

Explore the data through descriptive statistics and histogram plots to understand the data. Through exploratory data analysis, we find the number of data points, mean and dispersion of each attribute. Through histogram, we find the number / distribution of neighbourhoods based on the parameters -Population and Average Income.



	Population (2016)	Average income ($)
count	140.000000	140.000000
mean	19511.221429	55248.492857
std	10033.589222	38738.594546
min	6577.000000	25989.000000
25%	12019.500000	33476.750000
50%	16749.500000	44566.500000
75%	23854.500000	56654.250000
max	65913.000000	308010.000000

Descriptive Statistics



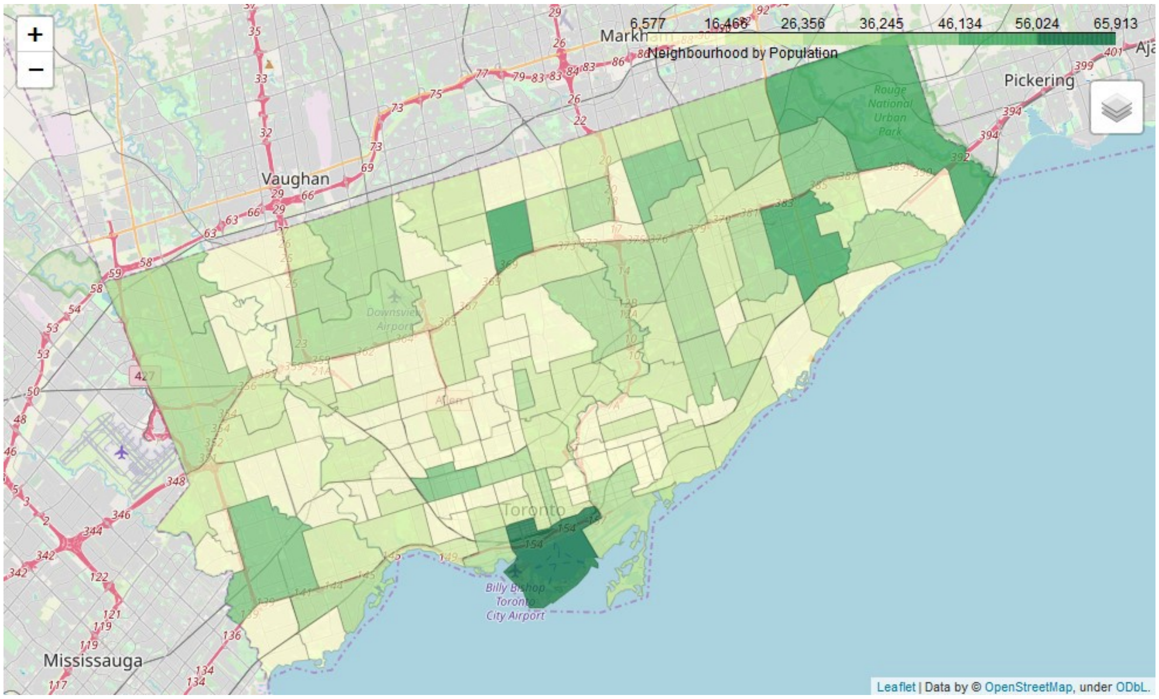
Histogram

From the Descriptive Statistics, we know there are 140 neighbourhoods in City of Toronto. There are no missing data.

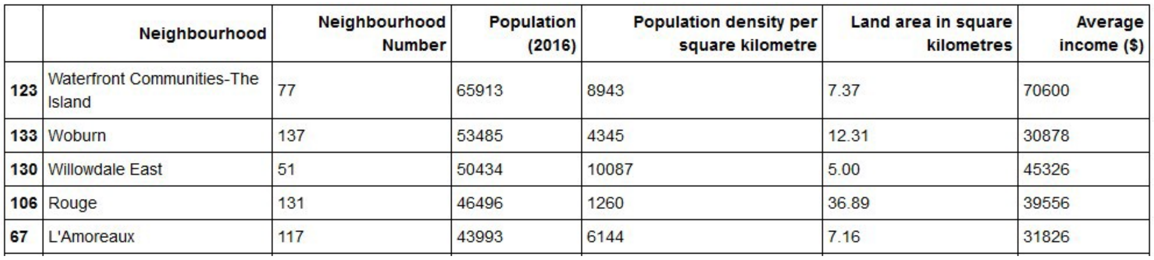
The histogram of population and average income show the Pareto Distribution — a right skewed distribution that has long tail. This means the Population and Average Income are not distributed equally across the city.

Data Visualization

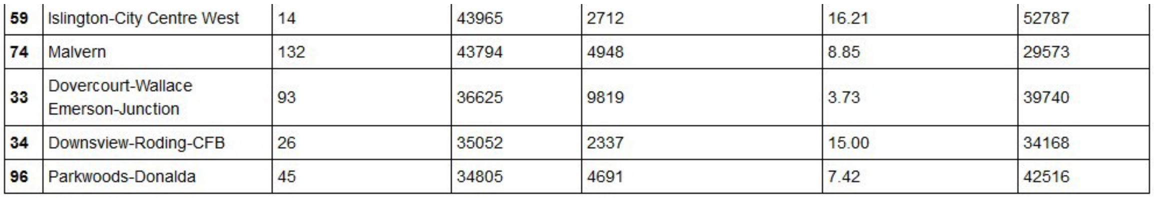
Display the color coded neighbourhoods on a map to know the spatial layout based on parameters — Population and Average Income. Display and analyze characteristics of the Top 10 neighbourhoods for each parameter.


6,577
16,466
26,356
36,245
46,134
56,024
65,913
Neighbourhood by Population
Markh
Pickering
Vaughan
Rouge National Urban Park
Downsview Airport
Toronto
Billy Bishop Toronto City Airport
Mississauga
Leaflet | Data by © OpenStreetMap, under ODbL.

Neighbourhoods Map by Population

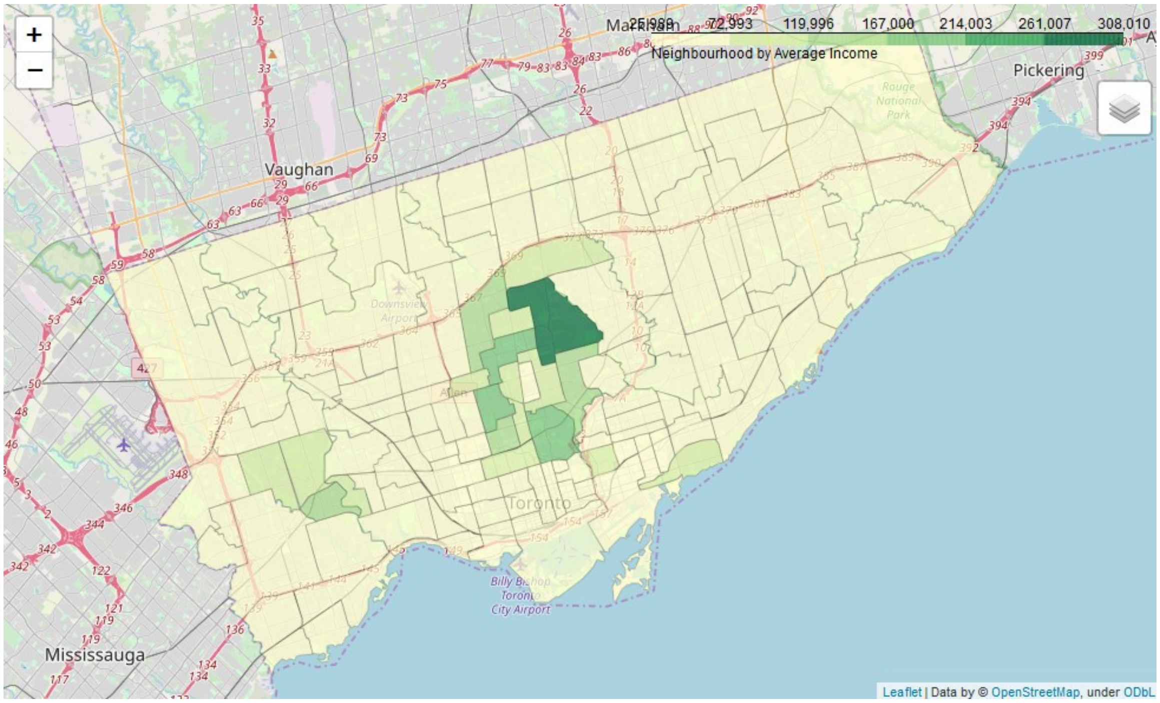


	Neighbourhood	Neighbourhood Number	Population (2016)	Population density per square kilometre	Land area in square kilometres	Average income ($)
123	Waterfront Communities-The Island	77	65913	8943	7.37	70600
133	Woburn	137	53485	4345	12.31	30878
130	Willowdale East	51	50434	10087	5.00	45326
106	Rouge	131	46496	1260	36.89	39556
67	L'Amoreaux	117	43993	6144	7.16	31826

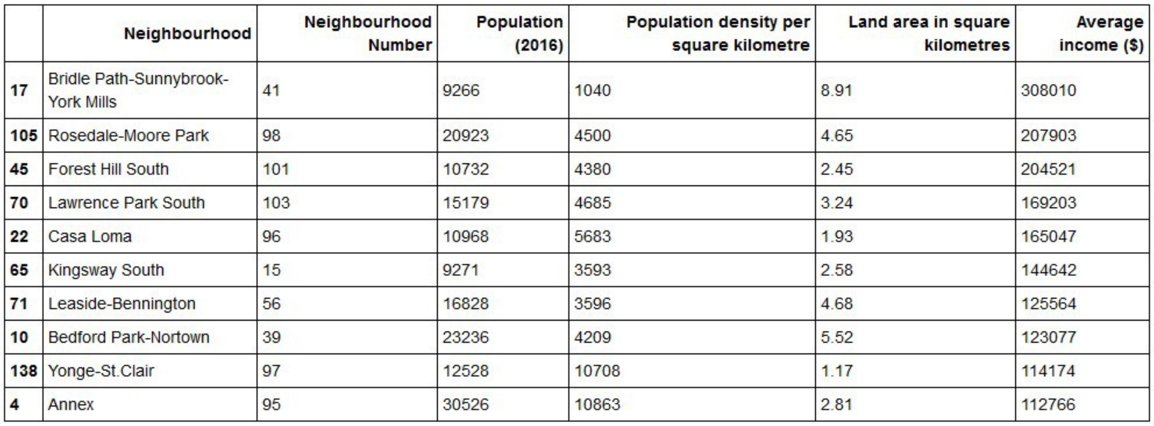


59	Islington-City Centre West	14	43965	2712	16.21	52787
74	Malvern	132	43794	4948	8.85	29573
33	Dovercourt-Wallace Emerson-Junction	93	36625	9819	3.73	39740
34	Downsview-Roding-CFB	26	35052	2337	15.00	34168
96	Parkwoods-Donalda	45	34805	4691	7.42	42516

Top 10 Neighbourhoods by Population



Neighbourhoods Map by Average Income

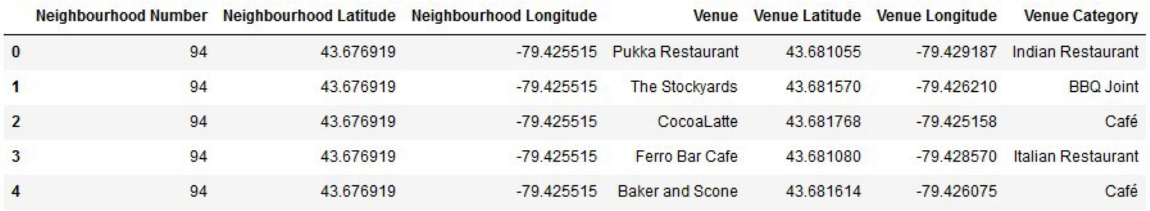


	Neighbourhood	Neighbourhood Number	Population (2016)	Population density per square kilometre	Land area in square kilometres	Average income ($)
17	Bridle Path-Sunnybrook-York Mills	41	9266	1040	8.91	308010
105	Rosedale-Moore Park	98	20923	4500	4.65	207903
45	Forest Hill South	101	10732	4380	2.45	204521
70	Lawrence Park South	103	15179	4685	3.24	169203
22	Casa Loma	96	10968	5683	1.93	165047
65	Kingsway South	15	9271	3593	2.58	144642
71	Leaside-Bennington	56	16828	3596	4.68	125564
10	Bedford Park-Nortown	39	23236	4209	5.52	123077
138	Yonge-St.Clair	97	12528	10708	1.17	114174
4	Annex	95	30526	10863	2.81	112766

Top 10 Neighbourhoods by Average Income

Get the Restaurants in the City of Toronto

Use Foursqaure API to get the venues of interest (Restaurants) in the City of Toronto. The Foursquare API explore() is used to query the venues. The parameters for each search are — Top Level Category 'Food', latitude-longitude of neighbourhood and radius to search. The API returns the list of venues with attributes — Venue Name, Venue Category, Venue Latitude, Venue Longitude. The Venue data of Neighbourhoods is accumulated into a data-set. Duplicate venues are removed to get the final list of Restaurants.



	Neighbourhood Number	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	94	43.676919	-79.425515	Pukka Restaurant	43.681055	-79.429187	Indian Restaurant
1	94	43.676919	-79.425515	The Stockyards	43.681570	-79.426210	BBQ Joint
2	94	43.676919	-79.425515	CocoaLatte	43.681768	-79.425158	Café
3	94	43.676919	-79.425515	Ferro Bar Cafe	43.681080	-79.428570	Italian Restaurant
4	94	43.676919	-79.425515	Baker and Scone	43.681614	-79.426075	Café

Restaurants in the City of Toronto

Exploring the data returned by Foursquare, we find that there are 2832 Restaurants in The City of Toronto. And there are 115 unique categories of Restaurants. We also note that there are 71 Indian Restaurants.

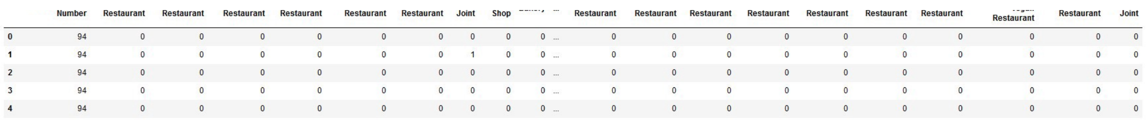
Analysis

We will use Machine Learning algorithm to classify the neighbourhoods based on similarity given the demographics data (Population and Average Income) and Venue Categories. we will use K-Nearest Neighbor(KNN).

Pre-Processing

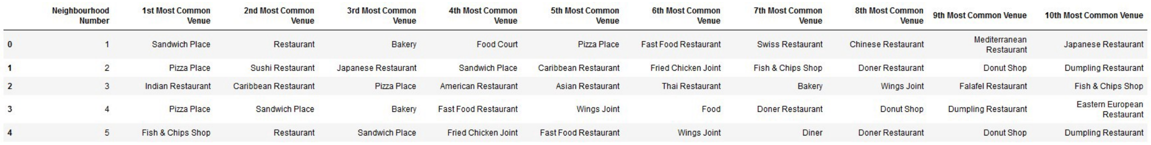
The parameter "Venue Category" is categorical data. It cannot be used directly in machine learning. Use one hot encoding to convert to a suitable format for processing. The data is grouped by neighbourhood and sorted to get the top 10 venues by neighbourhood.

Neighbourhood	Afghan	African	American	Arepa	Argentinian	Asian	BBQ	Bagel	Bakery	...	Tapas	Tex-Mex	Thai	Theme	Tibetan	Turkish	Udon	Vegetarian / Vegan	Vietnamese	Wings



	Number	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Joint	Shop		...	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Restaurant	Joint
0	94	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	94	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0
2	94	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	94	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	94	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

One Hot Encoding



	Neighbourhood Number	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	1	Sandwich Place	Restaurant	Bakery	Food Court	Pizza Place	Fast Food Restaurant	Swiss Restaurant	Chinese Restaurant	Mediterranean Restaurant	Japanese Restaurant
1	2	Pizza Place	Sushi Restaurant	Japanese Restaurant	Sandwich Place	Caribbean Restaurant	Fried Chicken Joint	Fish & Chips Shop	Doner Restaurant	Donut Shop	Dumpling Restaurant
2	3	Indian Restaurant	Caribbean Restaurant	Pizza Place	American Restaurant	Asian Restaurant	Thai Restaurant	Bakery	Wings Joint	Falafel Restaurant	Fish & Chips Shop
3	4	Pizza Place	Sandwich Place	Bakery	Fast Food Restaurant	Wings Joint	Food	Doner Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant
4	5	Fish & Chips Shop	Restaurant	Sandwich Place	Fried Chicken Joint	Fast Food Restaurant	Wings Joint	Diner	Doner Restaurant	Donut Shop	Dumpling Restaurant

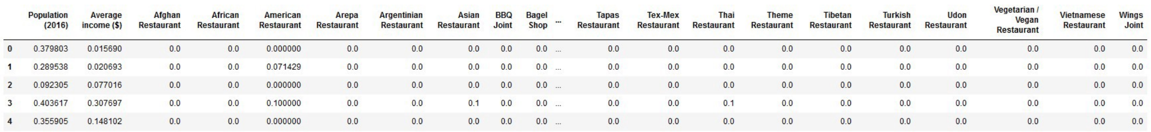
Top 10 Restaurants by Neighbourhood

Feature selection/extraction

Select the parameters to be used for machine leaning classification. The parameters are normalized to remove bias of one/more parameters. Post normalization, each parameter will have range of 0 to 1.

The parameters selected for classification are:

- Population
- Average Income
- Restaurant Type

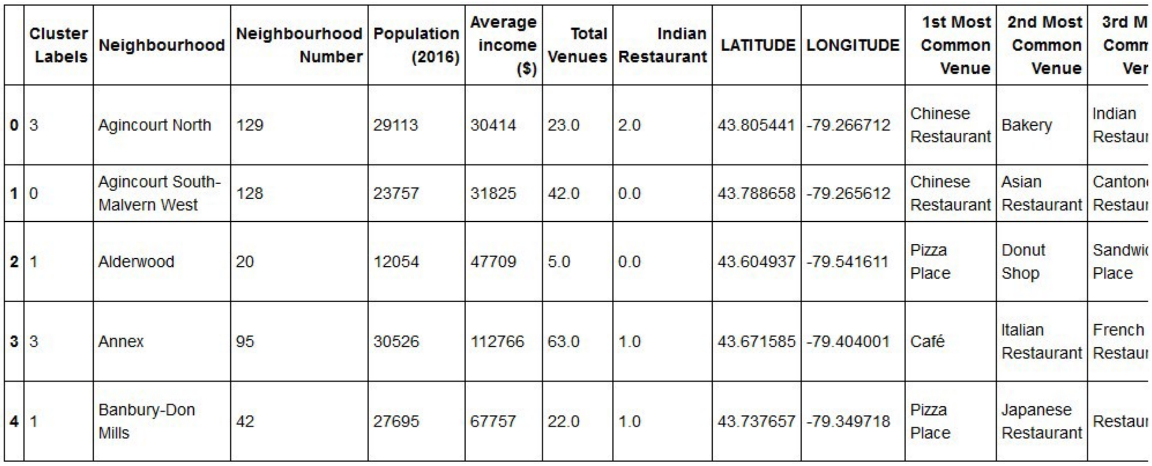


	Population (2016)	Average income ($)	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	...	Tapas Restaurant	Tex-Mex Restaurant	Thai Restaurant	Theme Restaurant	Tibetan Restaurant	Turkish Restaurant	Udon Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Wings Joint
0	0.379803	0.015690	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.289538	0.020693	0.0	0.0	0.071429	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.092305	0.077016	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.403617	0.307697	0.0	0.0	0.100000	0.0	0.0	0.1	0.0	0.0	...	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.355905	0.148102	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Feature Selection

Classification

We will use K-Nearest Neighbor(KNN) algorithm to classify each neighbourhood into 4 clusters. By trial, 4 is chosen as the best value for k to build the model.

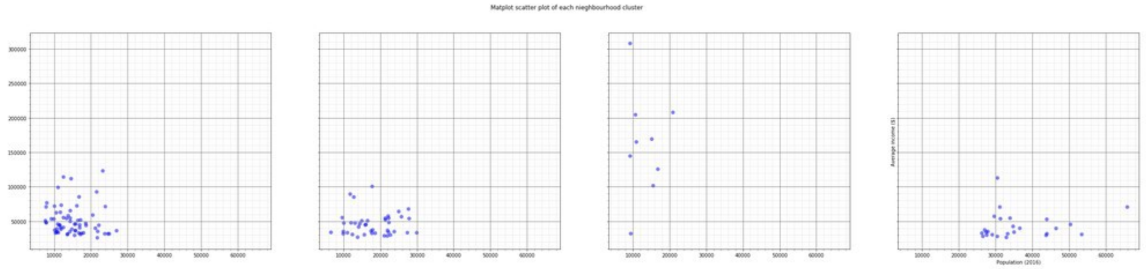


	Cluster Labels	Neighbourhood	Neighbourhood Number	Population (2016)	Average income ($)	Total Venues	Indian Restaurant	LATITUDE	LONGITUDE	1st Most Common Venue	2nd Most Common Venue	3rd M Comm Ver
0	3	Agincourt North	129	29113	30414	23.0	2.0	43.805441	-79.266712	Chinese Restaurant	Bakery	Indian Restau
1	0	Agincourt South-Malvern West	128	23757	31825	42.0	0.0	43.788658	-79.265612	Chinese Restaurant	Asian Restaurant	Canton Restau
2	1	Alderwood	20	12054	47709	5.0	0.0	43.604937	-79.541611	Pizza Place	Donut Shop	Sandwi Place
3	3	Annex	95	30526	112766	63.0	1.0	43.671585	-79.404001	Café	Italian Restaurant	French Restau
4	1	Banbury-Don Mills	42	27695	67757	22.0	1.0	43.737657	-79.349718	Pizza Place	Japanese Restaurant	Restau

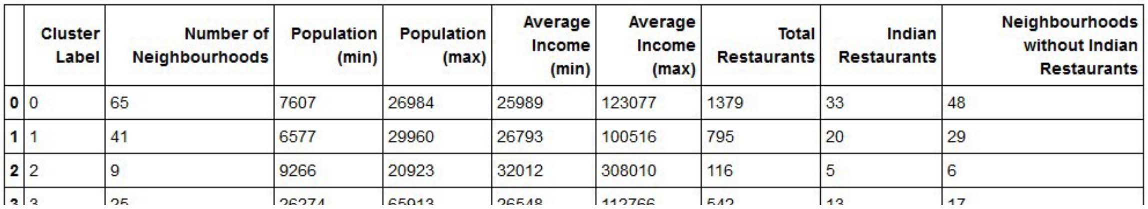
Classification of Neighbourhoods

Examine Clusters

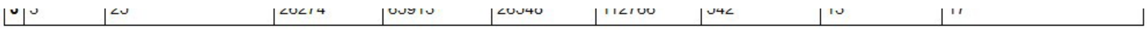
Let us examine the clusters to find out the characteristics of Neighbourhoods in them. First, we will plot each cluster based on demographics data i.e. Population and Average Income. We will get descriptive statistics of each cluster. We will then visualize the Neighbourhoods in each cluster on a map to understand the spatial relationships. These different representations helps us to understand the classification done by the KNN algorithm and identify key characteristics.



Scatter Plot of Each Cluster

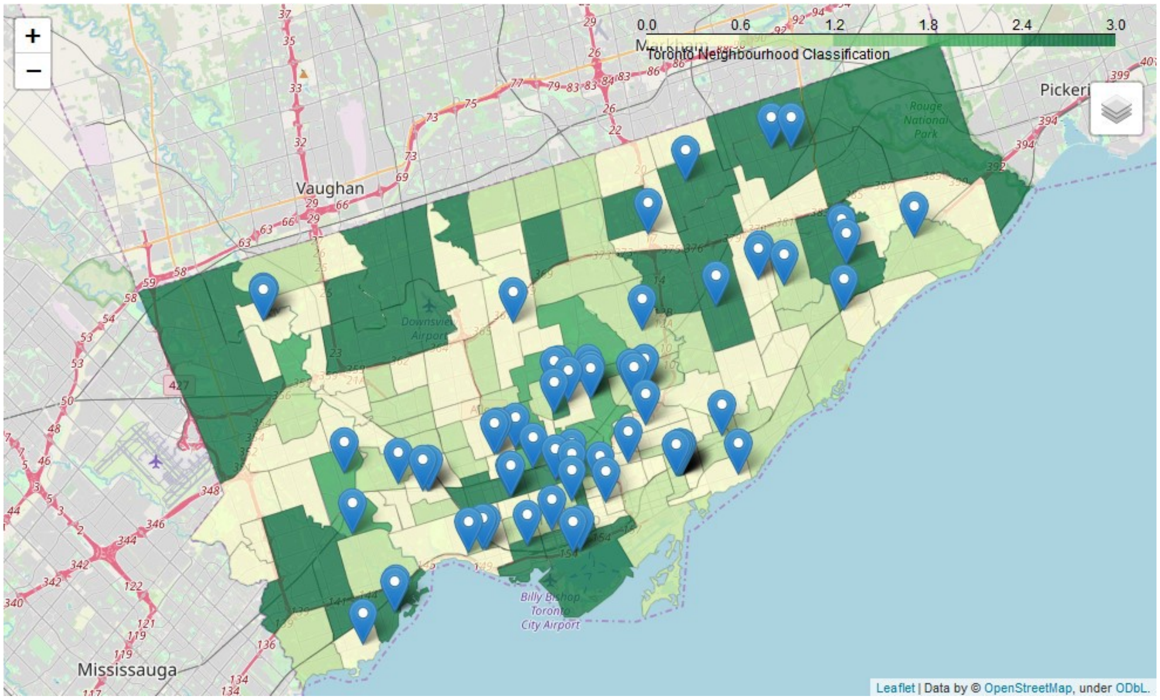


	Cluster Label	Number of Neighbourhoods	Population (min)	Population (max)	Average Income (min)	Average Income (max)	Total Restaurants	Indian Restaurants	Neighbourhoods without Indian Restaurants
0	0	65	7607	26984	25989	123077	1379	33	48
1	1	41	6577	29960	26793	100516	795	20	29
2	2	9	9266	20923	32012	308010	116	5	6
3	3	25	26274	65913	26548	112766	542	13	17



3	3	25	26274	65915	26548	112766	542	15	17

Descriptive Statistics of Each Cluster



+
−
0.0
0.6
1.2
1.8
2.4
3.0
Toronto Neighbourhood Classification
Vaughan
Pickeri
Rouge National Park
Downsview Airport
Billy Bishop Toronto City Airport
Mississauga
Leaflet | Data by © OpenStreetMap, under ODbL.

Neighbhourhood Map by Classification and Indian Restaurants

Characteristics of Neighbourhoods in Clusters

Examining the scatter plot and descriptive statistics, we can assume the characteristics of each cluster as follows:

- Cluster 0: Middle-High Income Neighbourhoods, situated around the central region, has highest number of Restaurants
- Cluster 1: Middle-High Income Neighbourhoods, situated around the central region, with moderate number of Restaurants
- Cluster 2: This consists of predominantly high income neighbourhoods, situated in the central region.

- Cluster 3: Highly populous Neighbourhoods with low-mid income levels, with moderate number of Restaurants

Results and Discussion

The K-Nearest Neighbor machine learning algorithm classified each neighbourhoods of the City of Toronto into 4 clusters based on the similarities given the demographic data (Population, Average Income and the type/number of Restaurants in each neighbourhood. Based on the classification and the number of Indian Restaurants in each cluster and neighbourhood, a prospective entrepreneur can narrow down to the best possible locations to open an Indian Restaurant. After analyzing the classification and presence of other Indian Restaurants, the recommendation are as follows:

- Cluster 0: Middle-High Income Neighbourhoods, situated around the central region, has highest number of Restaurants.
  This cluster has the maximum number of neighbourhoods and highest number of Restaurants. There are 39 Indian Restaurants in this cluster spread across in 20 neighbourhoods. with highest concentration of Indian and other Restaurants, the competition would be high. But there is still an opportunity for a new 4-Star fine dining Indian Restaurant in the 48 neighbourhoods which currently do not have an Indian Restaurant.
- Cluster 1: Middle-High Income Neighbourhoods, situated around the central region, with moderate number of Restaurants.
  This cluster of neighbourhoods is like Cluster 0 but with moderate number of Restaurants. This makes it a more favorable choice of location for a new Indian Restaurants. A 3-Star or 4-star fine dining Indian Restaurant.

- Cluster 2: This consists of predominantly high income neighbourhoods, situated in the central region.
  There are 5 Indian Restaurants in this cluster with 3 in the neighbourhood Leaside-Bennington. There are 5 neighbourhoods that are potential locations for opening a new 5-Star, very fine dining Indian Restaurant to cater to the high-end clientele in the neighbourhood.
- Cluster 3: Highly populous Neighbourhoods with low-mid income levels, with moderate number of Restaurants.
  This cluster of neighbourhoods is suitable for a 3-Star or 4-star fine dining Indian Restaurant in 17 neighbourhoods that currently do not have an Indian Restaurant.

Conclusion

The availability of data and improvements in Data Science and Machine Learning is helping in analysis and decision making. In the current study, the data about demographic profiles of neighbourhoods of the City of Toronto and Restaurants data from Foursquare are used in classifying the neighbourhoods in to different clusters using K-Nearest Neighbor algorithm. The clustering helps an entrepreneur in identifying suitable locations to open an Indian Restaurants. It also provides clues to the type of clientele for the restaurant. Armed with this insight the entrepreneur can further investigate other aspects required to open and run a successful business. The analysis and classification can be further be improved by using additional data points such as proximity to major attractions, distance from similar restaurants, ethnicity of the population and tourists, etc.

The clustering algorithm classified the neighbourhoods into 4 clusters. Exploratory analysis of clusters revealed characteristics of neighbourhoods in each cluster. This insight helped in narrowing down to the type of Restaurant that possibly will flourish in the neighbourhood.

For some of the neighbourhoods it was difficult to identify the rationale for the classification. There needs to be more research in the area of explainable machine learning and artificial intelligence.

Data Science Capstone Project