IBM – Coursera

Data Science Professional

Machine learning for Optimal Placement of New Coffee Shops in Toronto City

Final report

Rasoul Arabjamaloei 2019

Table of content:

List of Figures:

I. Introduction:

This report is prepared for the capstone project as a requirement to complete the coursera data science professional certificate. The capstone project is the 9th course in the series developed by IBM and presented by coursera. This report is the first one of the two reports on the capstone project. In this document the introduction/problem description section is provided. The objective of this project is to leverage the data science skills and the FourSquare location data to explore the possible solutions for a problem in hand.

The main objective in this project is to explore the Toronto neighborhoods, find the population of these neighborhoods and the surrounding venues, and use the clustering method to understand how the number of venues and population correlate and use this relationship to propose a method for optimized placement of new venues to improve its success chance. The hypothesis is that the more populated areas with less restaurants are a better target place for opening a new restaurant.

The idea comes from the simple supply-demand analysis that is performed in any business plan development process. Although the population density in Toronto is not stationary but dynamic and changes during different hours of the day, the population density is a sure number that determines the minimum number of possible visitors or shoppers in the neighborhood.

Therefore, this study would show that if the supply-demand for shoppers and buyers population in Toronto neighborhoods is balanced or not. It could be followed by another project to investigate what could be the possible reasons if the supply-demand is underbalanced or over-balanced.

The target audience for this report are:

- The investors who are looking to find the best possible investing options in opening new venues including restaurants in Toronto.
- The data scientists who are going to follow this methodology for other cities in the world.
- Restaurant chain owners who are looking to make a balance in supply-demand in their city of choice.
- The data science students who are looking for ideas to work on for their future project. Recommendations for future works would be provided at the last part of this report.

II. Data description:

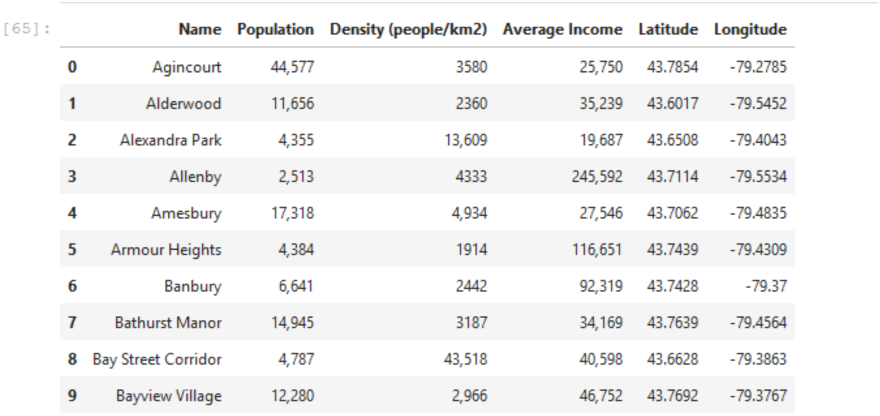
Toronto is the heart of the Canada’s economy. Being a major economy center of America, Toronto has been explored by many investors to find the best opportunities in this area. Therefore, it is not surprising to see that there’s very good data sets of Toronto neighborhoods specifications on the web.

FourSquare API is used to extract the list of venues and their types. The first step in data extraction is to extract the Toronto neighborhoods list and population from the following Wikipedia page: https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

We can extract the name of the neighborhoods from the given page and select the required data.

For each neighborhood, the name, population, population density and average income is extracted and ordered in a tabular format. At the next step the coordinates (latitude and longitude) are found by using geocoders library. Then the geospatial location is sent to FourSquare API. Using the “explore” endpoint, a list of surrounding venues in a pre-defined radius is returned by FourSqure. The occurrence of each venue type in neighborhoods would then be counted and one hot encoding is applied to turn each venue type into a column with their occurrence as the value.

The neighborhoods population, population density, average income and geospatial coordinates are placed in a table format dataframe as shown by figure 1.



[65]:

	Name	Population	Density (people/km2)	Average Income	Latitude	Longitude
0	Agincourt	44,577	3580	25,750	43.7854	-79.2785
1	Alderwood	11,656	2360	35,239	43.6017	-79.5452
2	Alexandra Park	4,355	13,609	19,687	43.6508	-79.4043
3	Allenby	2,513	4333	245,592	43.7114	-79.5534
4	Amesbury	17,318	4,934	27,546	43.7062	-79.4835
5	Armour Heights	4,384	1914	116,651	43.7439	-79.4309
6	Banbury	6,641	2442	92,319	43.7428	-79.37
7	Bathurst Manor	14,945	3187	34,169	43.7639	-79.4564
8	Bay Street Corridor	4,787	43,518	40,598	43.6628	-79.3863
9	Bayview Village	12,280	2,966	46,752	43.7692	-79.3767

Figure 1: Neighborhoods dataset

Each row represents a neighborhood dataset and each column is the properties of that neighborhood. The dataset has 5 features and 174 samples.

III. Methodology:

As geocoders didn’t return the coordinates of some of the neighborhoods, the rows corresponding to these neighborhoods were dropped. Figure 2 shows the remaining neighborhoods that make 155 sets of data.

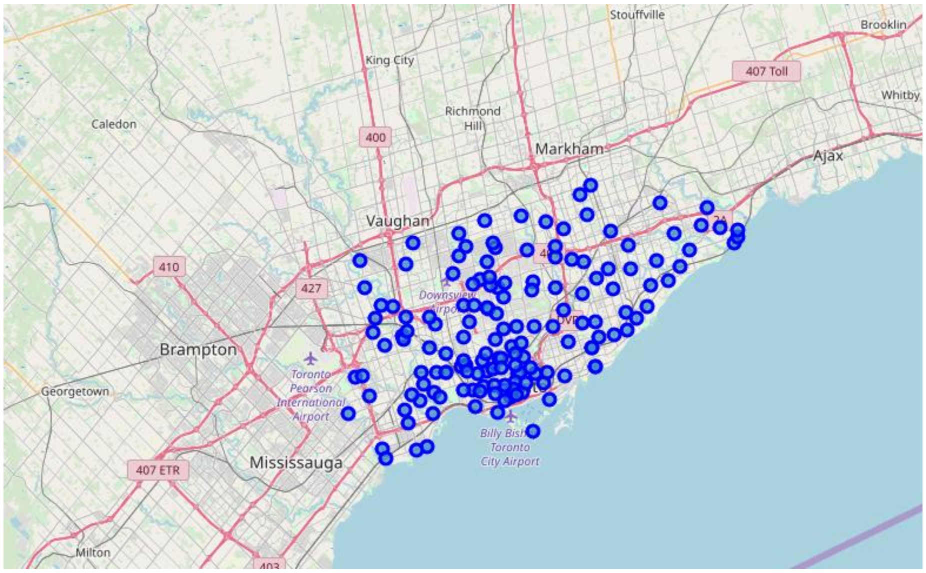
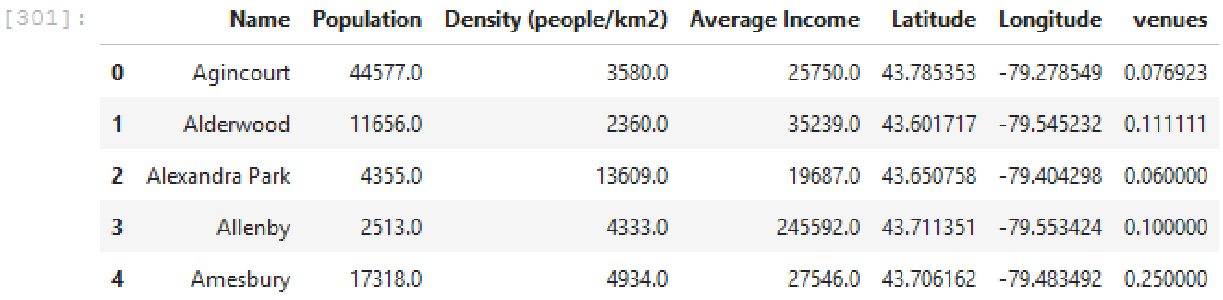


Figure 2: The map

A function was built to make API calls to FourSquare and get the top 50 venues and their types in the distance of 500 meters from the neighborhoods geospatial coordinates.

The venues were grouped by their types and neighborhoods and a column of neighborhood names were added to them. In this project I decided to work on coffee shops and build a model to find the optimized locations for opening a new coffee place. The same method can be applied to find the best location for any other types of venues.

For the purpose of our study the number of coffee shops and cafes were added together and then inserted to the dataframe shown in figure 1. Figure 3 shows the final dataframe.

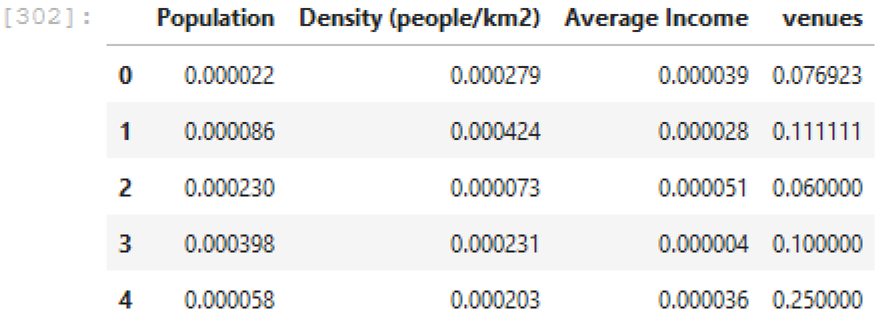


[301]:

	Name	Population	Density (people/km2)	Average Income	Latitude	Longitude	venues
0	Agincourt	44577.0	3580.0	25750.0	43.785353	-79.278549	0.076923
1	Alderwood	11656.0	2360.0	35239.0	43.601717	-79.545232	0.111111
2	Alexandra Park	4355.0	13609.0	19687.0	43.650758	-79.404298	0.060000
3	Allenby	2513.0	4333.0	245592.0	43.711351	-79.553424	0.100000
4	Amesbury	17318.0	4934.0	27546.0	43.706162	-79.483492	0.250000

Figure 3: Final dataframe

The Population, Population density and Average income variables were converted to 1/variable so that they correlate directly with the venues mean. A new dataframe was built by taking the new Population, Population density and Average income variables and the mean value of venues. Figure 4 shows this dataframe. This dataframe was used for k-mean clustering.



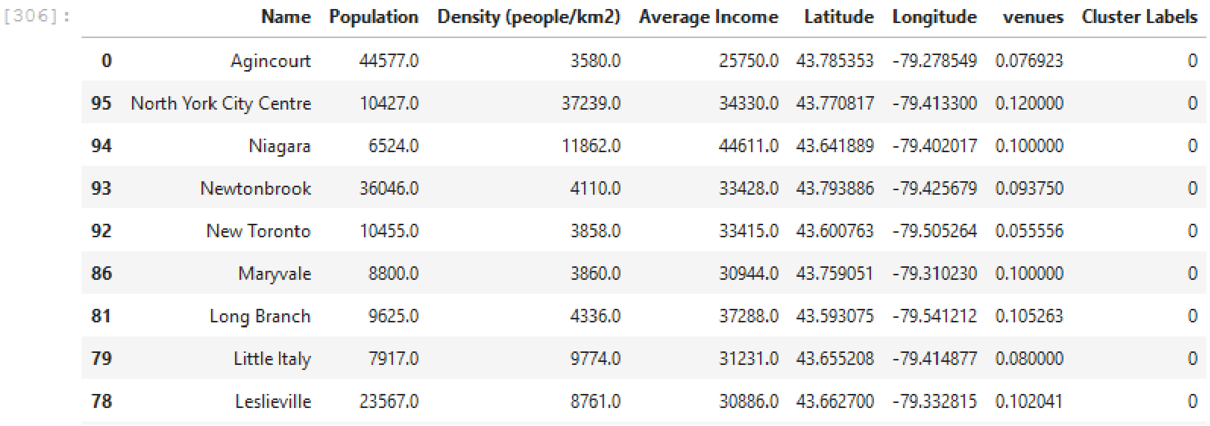
[302]:

	Population	Density (people/km2)	Average Income	venues
0	0.000022	0.000279	0.000039	0.076923
1	0.000086	0.000424	0.000028	0.111111
2	0.000230	0.000073	0.000051	0.060000
3	0.000398	0.000231	0.000004	0.100000
4	0.000058	0.000203	0.000036	0.250000

Figure 4: Data for k-mean clustering

IV. Results:

K-Mean clustering was applied and the neighborhoods were divided to 5 clusters. The investment for coffee shops at the neighborhoods that fall into the same clusters has the same potential for success. The cluster labels of each neighborhood was found and a new dataframe was built by adding the cluster labels to the dataframe of figure 3. Figure 5 represents this new dataframe.



[306]:

	Name	Population	Density (people/km2)	Average Income	Latitude	Longitude	venues	Cluster Labels
0	Agincourt	44577.0	3580.0	25750.0	43.785353	-79.278549	0.076923	0
95	North York City Centre	10427.0	37239.0	34330.0	43.770817	-79.413300	0.120000	0
94	Niagara	6524.0	11862.0	44611.0	43.641889	-79.402017	0.100000	0
93	Newtonbrook	36046.0	4110.0	33428.0	43.793886	-79.425679	0.093750	0
92	New Toronto	10455.0	3858.0	33415.0	43.600763	-79.505264	0.055556	0
86	Maryvale	8800.0	3860.0	30944.0	43.759051	-79.310230	0.100000	0
81	Long Branch	9625.0	4336.0	37288.0	43.593075	-79.541212	0.105263	0
79	Little Italy	7917.0	9774.0	31231.0	43.655208	-79.414877	0.080000	0
78	Leslieville	23567.0	8761.0	30886.0	43.662700	-79.332815	0.102041	0

Figure 5: The neighborhoods dataframe with the cluster labels

The folium library was used to plot the localization of the neighborhoods with cluster labels. Figure 5 shows the distribution of these clusters on the map.

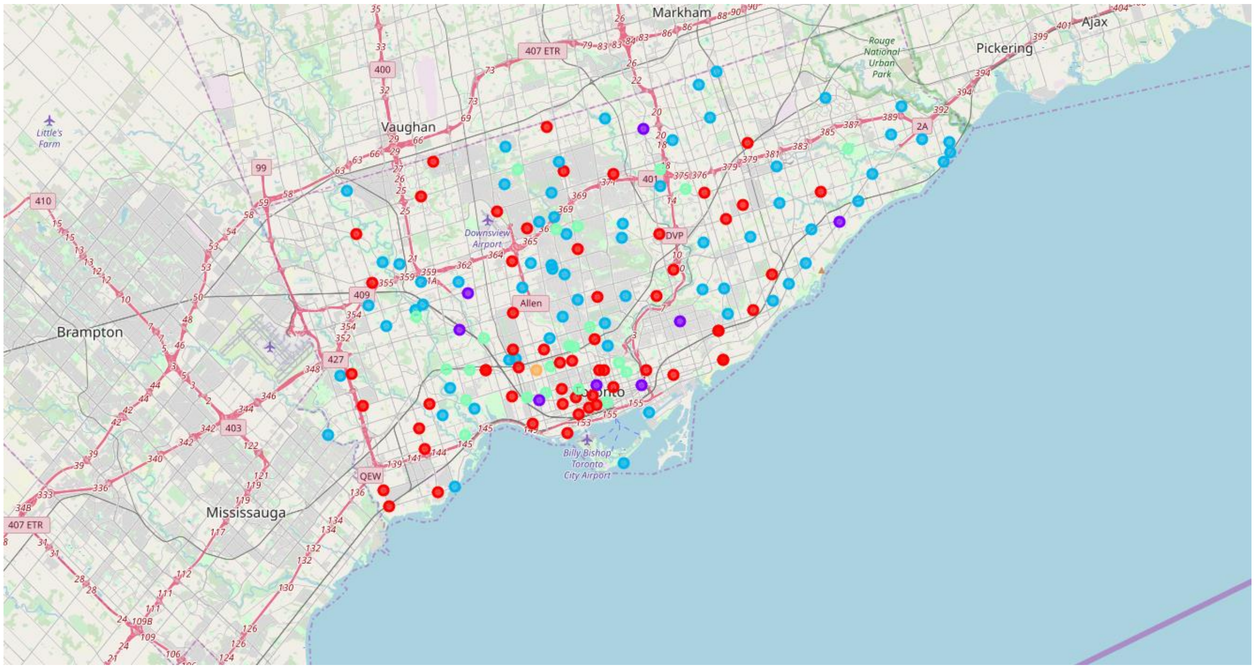


Figure 6: Localization of the neighborhoods with their cluster labels as colurs.

V. Discussion:

To understand what these clusters means, the scatter plot tool was implemented. The cluster labels were plotted versus Population density, Population and Average income as shown by figures 7, 8 and 9.

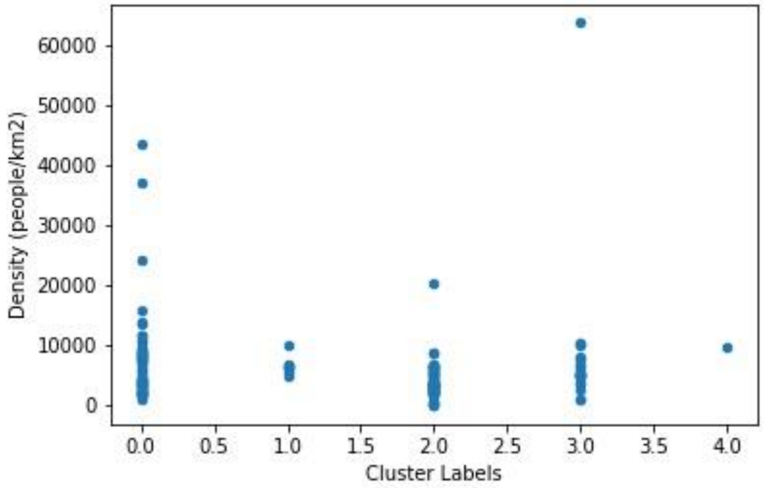


Figure 7: Cluster labels versus Population density

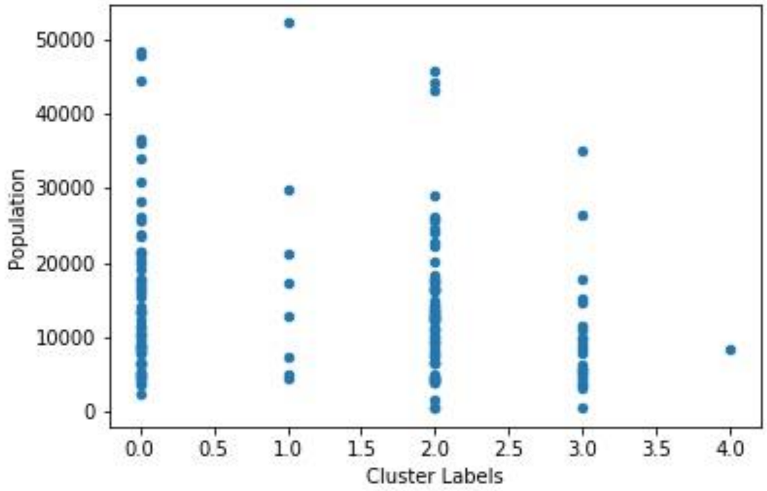


Figure 8: Cluster labels versus Population

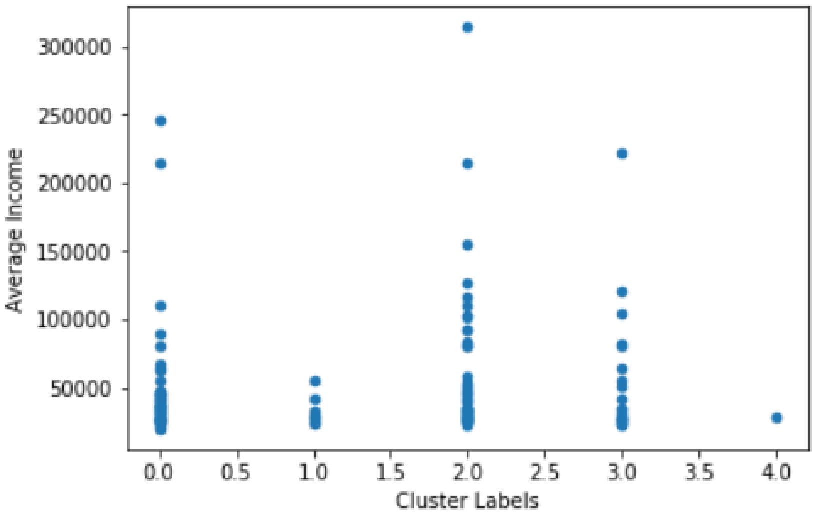


Figure 9: Cluster labels versus Average income

As it is observed in figures 7-9, there is no clear relationships between cluster labels and Population density, Population and Average income. Therefore we can deduce that all of these variables have been important in clustering. To understand the effect of combination of these variables, a new variable was defined by dividing the venues by Population and Average income and the new variable was plotted versus cluster label as shown by figure 10.

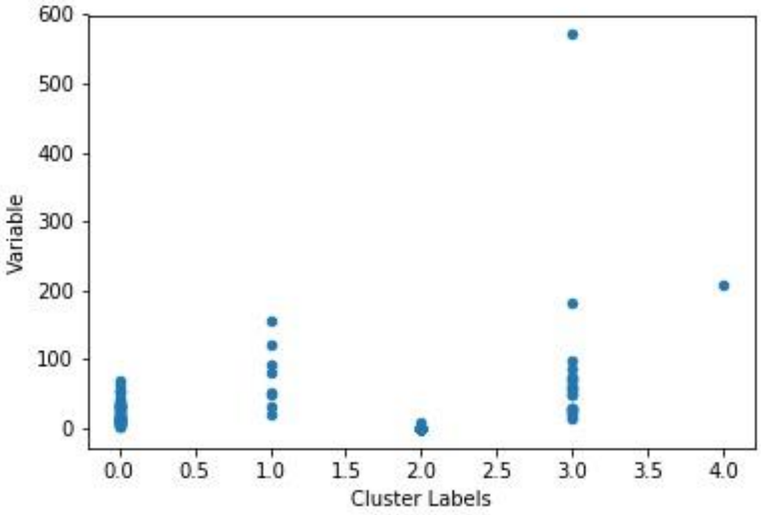


Figure 10: The new variable versus cluster labels

It is observed that the new variable correlates better with the cluster labels. The lower the variable means the better potential for openning new venues. Therefore the best places to open new coffee shops from best to worst are 2, 0, 3, 1 and 4 approximately. However it is obvious that the neighborhoods in cluster 2 have the best potential for new venues.

VI. Conclusions and recommendations:

In this project the Foursquare API tool was used in combination with geocoders, folium, Scikit Learn and matplotlib libraries to find the best locations to open a new coffee shop in the city of Toronto. The hypothesis is that the neighborhoods with higher population, higher population density, higher average income and lower existing venues have the best potential for the success rate of a new coffee shop. The results showed that k-mean clustering method could make distinct neighborhood clusters based on the given information.

This methodology can be used to optimize the process of finding best places for opening other new venues.