Applied Data Science Capstone Project

for

IBM Data Science Professional Certificate

January 20, 2021

John Koplimae

Table of Contents

Introduction

There is something special about a local coffee shop. The intimate environment, the sense of community, a comfortable space to work from, and of course a great cup of joe. Throughout the city of Toronto there are many, if not hundreds, of coffee shops. So in this competitive market of small independent coffee shops, where might a prospective proprietor open their new coffee shop?

Business Problem

Opening a coffee shop, like any small business, comes with hurdles and challenges, some being unique to this specific type of business.
This report shall look specifically at the city of Toronto and its neighbourhoods. With many coffee shops currently open around the city and some neighbourhoods having a concentration of shops, knowing which neighbourhood to open a shop in might feel like a daunting task. Within the scope of this analysis we look to help remove this hurdle for a potential owner/operator by figuring out where a good location, geographically, to open a new coffee shop would be.

Data

For this analysis, data from several sources shall be used. This data shall include:

- Neighbourhood information:
    - Name
    - Location (latitude and longitude)
    - Population
    - Number of Coffee shops per
- Coffee shop information:
    - Name and location

Sources of data are:

- Open Data Portal for the CIty of Toronto - https://open.toronto.ca/
    - Using the data files (.csv) for Neighbourhood Boundaries and Profiles.
    - Profiles - https://open.toronto.ca/dataset/neighbourhood-profiles/

- Foursquare API
    - To acquire the venue data for coffee shops in these neighbourhoods and the number of shops.

The data will be used to determine the population per neighbourhood and how many coffee shops are in each neighbourhood. The idea being that a neighbourhood with few coffee shops but a high population would be the ideal location for a new shop to open.

Methodology

For this report we shall need information on the neighbourhoods in Toronto such as name, location(latitude and longitude), and the population of each. This data can be imported as a .csv file from the Open Data Portal for the CIty of Toronto website. Once imported this file shall be turned into a dataframe which may be cleaned and organized further that will best allow for an accurate evaluation of the data.
To clean the dataframe all rows and columns that did not contain data pertaining to neighbourhood name, geographical coordinates and population were removed. The population data was then reformatted by removing the comma and changing the type so that it could be used for graphing and plotting The below graph visually displays the population of each neighbourhood of Toronto in descending order.

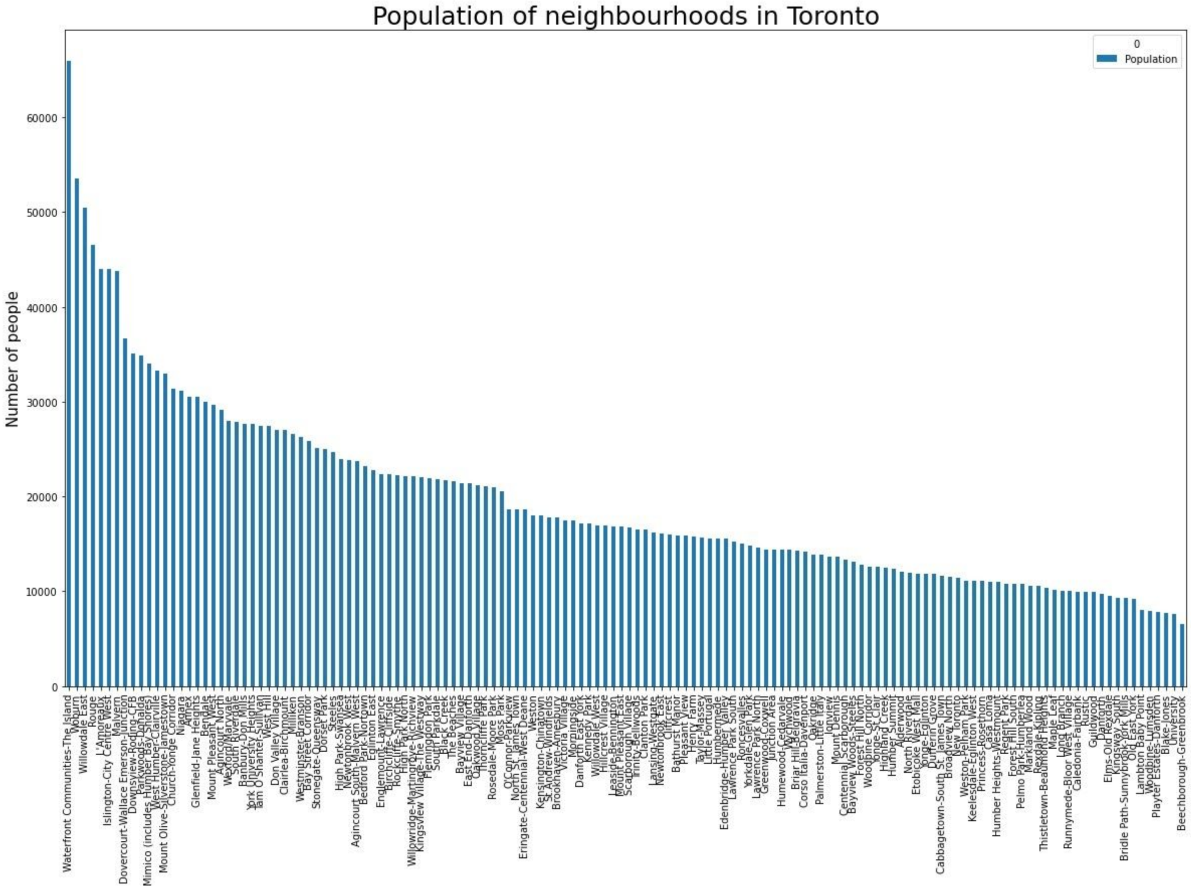


Fig. 1

Following this the neighbourhoods of Toronto were plotted on a map for easy reference to their location and distribution across The City of Toronto.

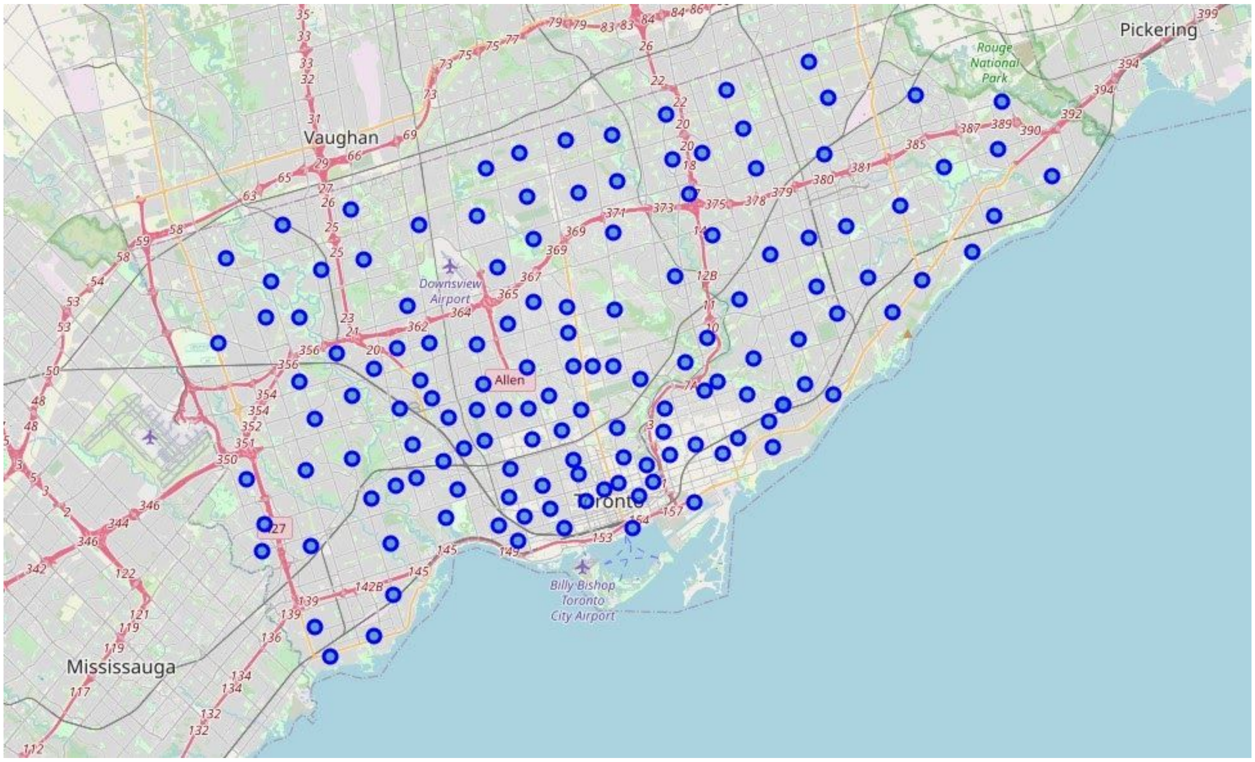


Fig. 2

The venue data was the next source needed. To get this, a call to the Foursquare API was sent to retrieve the 100 venues within 1000m radius of the geographical coordinates for each neighbourhood. A new dataframe was created with this venue information, it was then filtered for coffee shops only in each neighbourhood. Both the coffee shop and population data are normalized to make interpretation of their magnitude possible. With this done the K-means clustering algorithm is applied to the dataset to determine the clusters. Five clusters were used to segment the data. A new map of Toronto was plotted with the neighbourhoods shown in their assigned clusters to help with a visual assessment of the clustering and location.

Note: Cluster #4 is light teal in colour, which makes it a little hard to see all the points in the cluster.

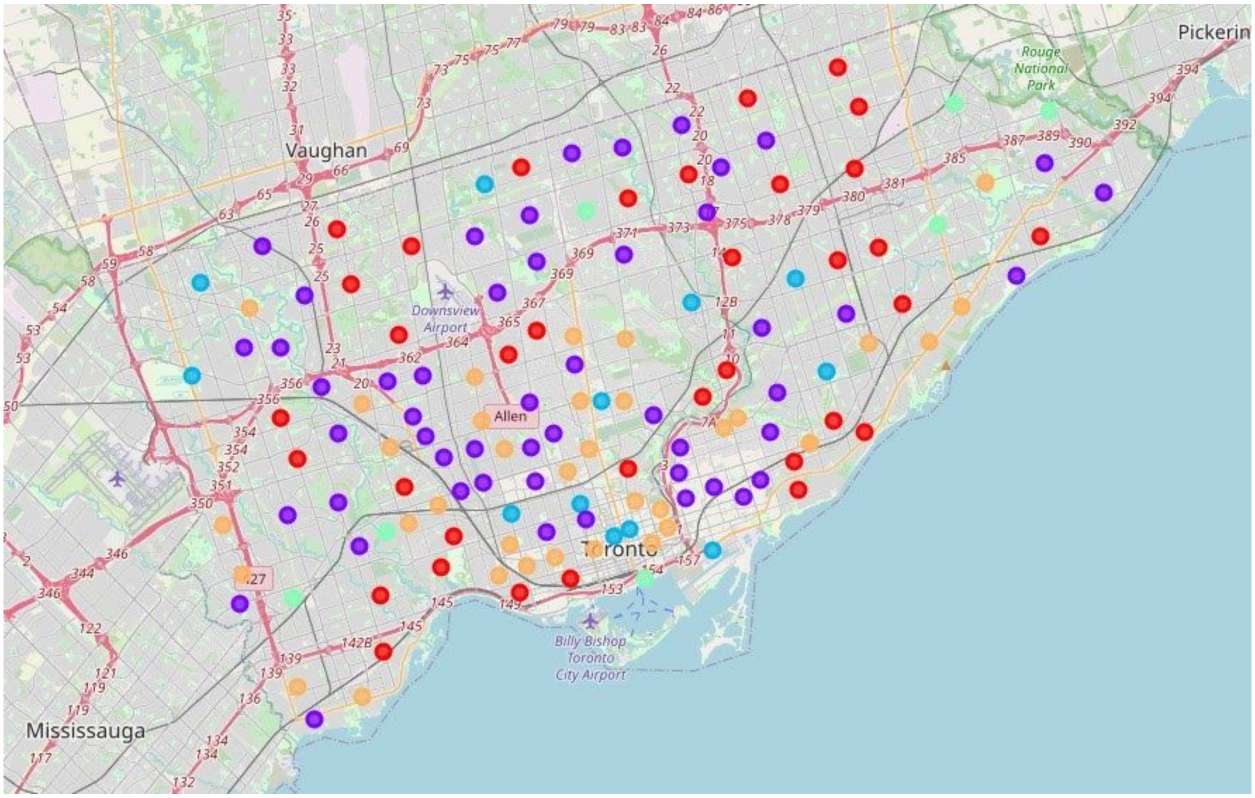


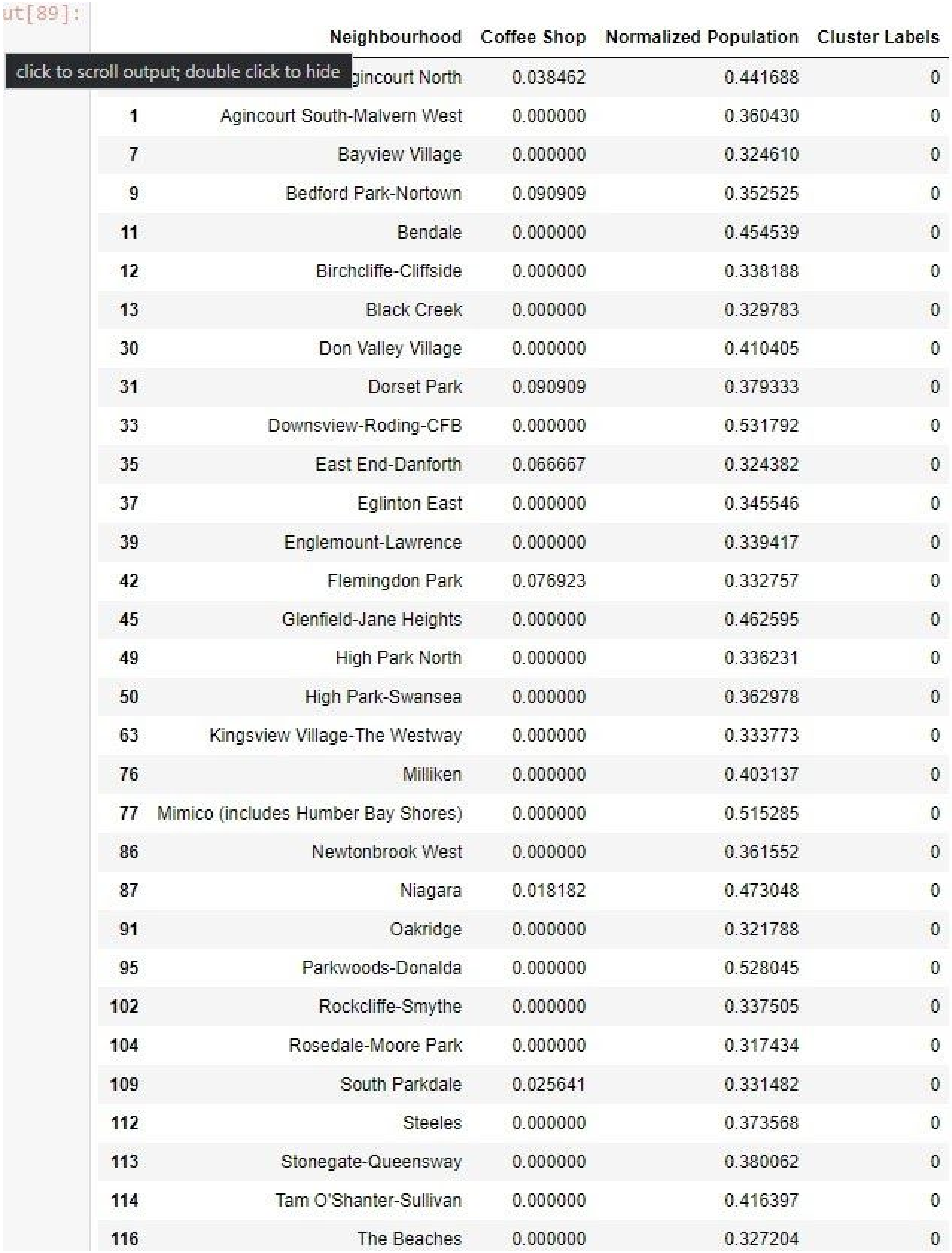
Fig. 3

Results

A total of five clusters were created to sort the neighbourhoods, the coffee shops in the neighbourhoods and the population of each neighbourhood. After reviewing the information of each cluster they can best be described as stated in the below table.

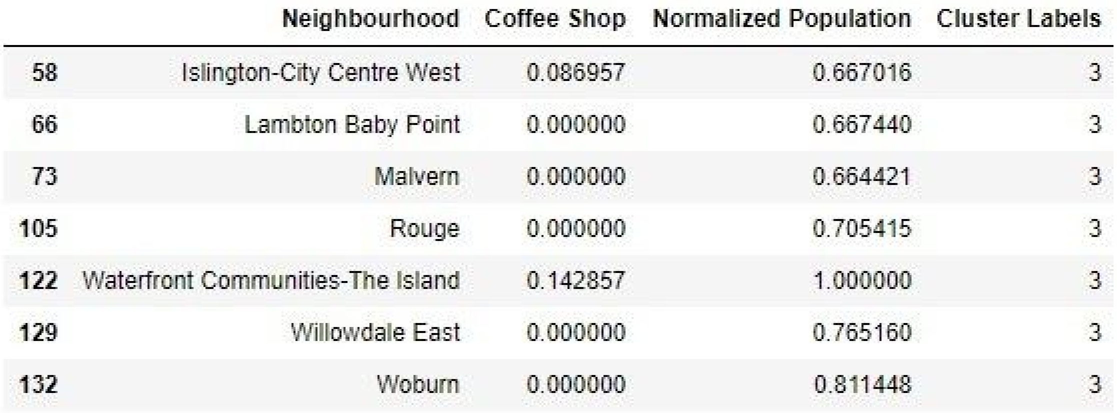
	Coffee Shop Frequency	**Population**
Cluster #1	Low to none	medium
Cluster #2	Low to none	Low
Cluster #3	Medium	Medium
Cluster #4	Low	High
Cluster #5	Medium	Medium - Low

From reviewing the map and the above table Clusters #1 and #4 are of interest for the purpose of this assessment. Cluster #1 is marked with red and cluster #4 is marked with light teal colour on the map of Toronto (Fig. 3). The clustered data for each cluster is as follows:



	Neighbourhood	Coffee Shop	Normalized Population	Cluster Labels
	gincourt North	0.038462	0.441688	0
1	Agincourt South-Malvern West	0.000000	0.360430	0
7	Bayview Village	0.000000	0.324610	0
9	Bedford Park-Nortown	0.090909	0.352525	0
11	Bendale	0.000000	0.454539	0
12	Birchcliffe-Cliffside	0.000000	0.338188	0
13	Black Creek	0.000000	0.329783	0
30	Don Valley Village	0.000000	0.410405	0
31	Dorset Park	0.090909	0.379333	0
33	Downsview-Roding-CFB	0.000000	0.531792	0
35	East End-Danforth	0.066667	0.324382	0
37	Eglinton East	0.000000	0.345546	0
39	Englemount-Lawrence	0.000000	0.339417	0
42	Flemingdon Park	0.076923	0.332757	0
45	Glenfield-Jane Heights	0.000000	0.462595	0
49	High Park North	0.000000	0.336231	0
50	High Park-Swansea	0.000000	0.362978	0
63	Kingsview Village-The Westway	0.000000	0.333773	0
76	Milliken	0.000000	0.403137	0
77	Mimico (includes Humber Bay Shores)	0.000000	0.515285	0
86	Newtonbrook West	0.000000	0.361552	0
87	Niagara	0.018182	0.473048	0
91	Oakridge	0.000000	0.321788	0
95	Parkwoods-Donalda	0.000000	0.528045	0
102	Rockcliffe-Smythe	0.000000	0.337505	0
104	Rosedale-Moore Park	0.000000	0.317434	0
109	South Parkdale	0.025641	0.331482	0
112	Steeles	0.000000	0.373568	0
113	Stonegate-Queensway	0.000000	0.380062	0
114	Tam O'Shanter-Sullivan	0.000000	0.416397	0
116	The Beaches	0.000000	0.327204	0

Fig. 4 - Cluster #1, missing 4 rows at the end do to length of cluster.



	Neighbourhood	Coffee Shop	Normalized Population	Cluster Labels
58	Islington-City Centre West	0.086957	0.667016	3
66	Lambton Baby Point	0.000000	0.667440	3
73	Malvern	0.000000	0.664421	3
105	Rouge	0.000000	0.705415	3
122	Waterfront Communities-The Island	0.142857	1.000000	3
129	Willowdale East	0.000000	0.765160	3
132	Woburn	0.000000	0.811448	3

Fig. 5 - Cluster #4

Discussion

There are a couple of observations and limitations that have been noted during this investigation of the neighbourhoods. The Foursquare API returned 1971 venues, but with a search criteria of 100 venues to return per neighbourhood and 141 neighbourhoods, this seems like quite a bit less returned than expected. This may be due to a couple factors, firstly the 1000m radius might not give enough area to return venue samples in the neighbourhoods that are more residential with larger property size or areas with farmland and warehouses/factories. These areas would typically be in the outer surrounding area of a city and in particular the City of Toronto. But increasing the radius too much would then impact the data returned on neighbourhoods closer to the center of the city with smaller area and possible overlap then. In future it might be of greater benefit to investigate the city centre neighbourhoods and the further out neighbourhoods separately to give a better idea of venues, in this case coffee shops, by modifying the radius according to area of each neighbourhood.

The other factor would be the limit of 100 venues returned by the personal free account for the Foursquare API. With this not being an issue for the larger neighbourhoods currently, it might have only returned a portion of the coffee shops in the city center neighbourhoods. Thus only giving a partial picture of how many shops are in each neighbourhood. This is something that can only be fixed by paying for an account which would return a greater number of venues per API call.

Conclusion

In conclusion, with the data that was obtained, cleaned and observed, it would be recommended to look at cluster #4 with its low number of coffee shops and high level of population for its neighbourhoods, would be the best area to open a new coffee shop. Cluster #1 would be a close second for the lower population level. One thing to keep in mind would be the location of the neighbourhood in regards to its position in Toronto, as one that is closer to the city centre in the cluster would serve as the more ideal location.