Coursera Applied Data Science Capstone course

Week 5 assignment

by

Catherine Tam

April 25, 2019

Introduction and Business Problem

A client is looking to open a coffee shop serving gourmet tea, coffee and desserts in the Greater Toronto Area (GTA). She needs some recommendations on where to open her business. Given the exquisite quality of food and drinks served and to cover the costs, she is looking for an affluent neighbourhood where people are willing to splurge. The population targeted is also the younger age group under 40 years old who tends to value such lifestyle and experience. The client is aware of potential competitions. Therefore, she expects to pick a location that has less competition allowing her to build the client base at the beginning. There are over 100 neighbourhoods in the GTA and she wonders which areas are worth looking at so she can conduct further analysis before deciding on a location.

Data

The following is a description of all the data included in the analysis:

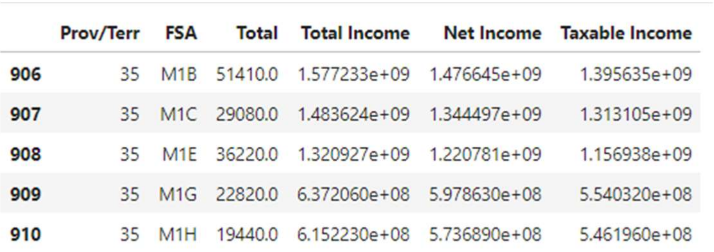
a) Income

The Canada Revenue Agency (CRA) administers tax laws for the Government of Canada and for most provinces and territories, and administers various social and economic benefit and incentive programs delivered through the tax system. The CRA has published the 2017 edition of tables based on Forward Sortation Area (FSA) (first three digits of the postal codes) summarizing the most recent 2015 tax year assessment or reassessment information on its website. The CRA uses the taxfiler's mailing address and postal code as it appears on the T1 Income Tax and Benefit Return to determine the FSA as of December 31, 2015.

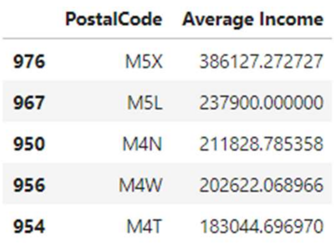
The income classes presented in the tables are based on the total income assessed (including employment income, pension income, investment income, self-employment income, social benefit payments and other income) and the number of tax filers.

A csv file was obtained from the CRA link at https://www.canada.ca/content/dam/cra-arc/prog-policy/stats/individual-tax-stats-fsa/2015-tax-year/tbl1a-en.csv After data was loaded into a dataframe, only rows of FSA starting with the letter "M" indicating Toronto was retained. Average Income for each postal code was then calculated based on Total Income divided by the Total (the number of tax filers).

The client has indicated a specific population with high income is targeted for her business. Therefore, neighbourhood average income information has been included in the analysis.



	Prov/Terr	FSA	Total	Total Income	Net Income	Taxable Income
906	35	M1B	51410.0	1.577233e+09	1.476645e+09	1.395635e+09
907	35	M1C	29080.0	1.483624e+09	1.344497e+09	1.313105e+09
908	35	M1E	36220.0	1.320927e+09	1.220781e+09	1.156938e+09
909	35	M1G	22820.0	6.372060e+08	5.978630e+08	5.540320e+08
910	35	M1H	19440.0	6.152230e+08	5.736890e+08	5.461960e+08

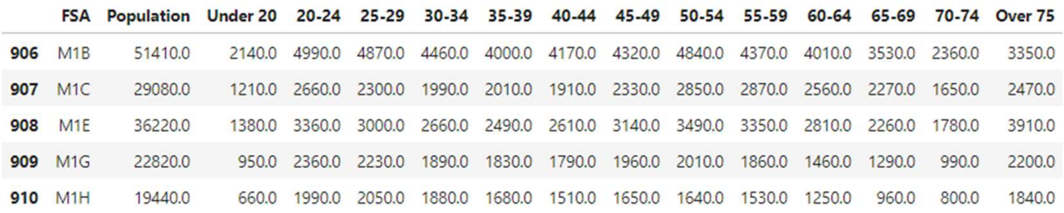


	PostalCode	Average Income
976	M5X	386127.272727
967	M5L	237900.000000
950	M4N	211828.785358
956	M4W	202622.068966
954	M4T	183044.696970

b) Age

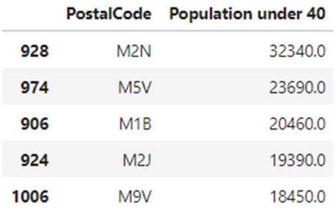
The CRA has indicated that the age of the taxfiler is determined from the reported year of birth on the Income Tax and Benefit Return. For individuals who did not report a year of birth, their age is imputed by the CRA for statistical completeness.

A csv file was obtained from the CRA link at https://www.canada.ca/content/dam/cra-arc/prog-policy/stats/individual-tax-stats-fsa/2015-tax-year/tbl2-en.csv After data was loaded into a dataframe, only rows of FSA starting with the letter "M" indicating Toronto was retained.



	FSA	Population	Under 20	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	Over 75
906	M1B	51410.0	2140.0	4990.0	4870.0	4460.0	4000.0	4170.0	4320.0	4840.0	4370.0	4010.0	3530.0	2360.0	3350.0
907	M1C	29080.0	1210.0	2660.0	2300.0	1990.0	2010.0	1910.0	2330.0	2850.0	2870.0	2560.0	2270.0	1650.0	2470.0
908	M1E	36220.0	1380.0	3360.0	3000.0	2660.0	2490.0	2610.0	3140.0	3490.0	3350.0	2810.0	2260.0	1780.0	3910.0
909	M1G	22820.0	950.0	2360.0	2230.0	1890.0	1830.0	1790.0	1960.0	2010.0	1860.0	1460.0	1290.0	990.0	2200.0
910	M1H	19440.0	660.0	1990.0	2050.0	1880.0	1680.0	1510.0	1650.0	1640.0	1530.0	1250.0	960.0	800.0	1840.0

Population under 40 was calculated by adding up all the relevant age groups.



	PostalCode	Population under 40
928	M2N	32340.0
974	M5V	23690.0
906	M1B	20460.0
924	M2J	19390.0
1006	M9V	18450.0

The client has indicated a specific age population is targeted for her business. Therefore, population under 40 for each neighbourhood has been included in the analysis.

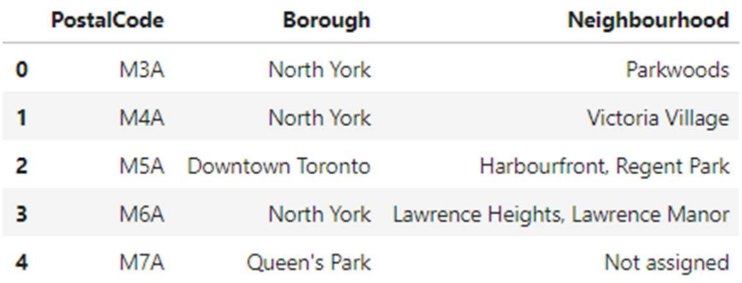
It should be noted that data for both income and age taken from the CRA website is only a subset of the true population in each neighbourhood since not every individual may be a tax filer. However, the dataset is still a good representation of the population describing its characteristics, and has value for the analysis.

The information as of December 31, 2015 may seem a little dated as well. However, income and age characteristics of neighbourhoods are not expected to change drastically year over year.

c) Neighbourhood names

A Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M containing a table with postal code, borough and neighbourhood in Toronto was used.

- Data was scraped and transformed into a dataframe.

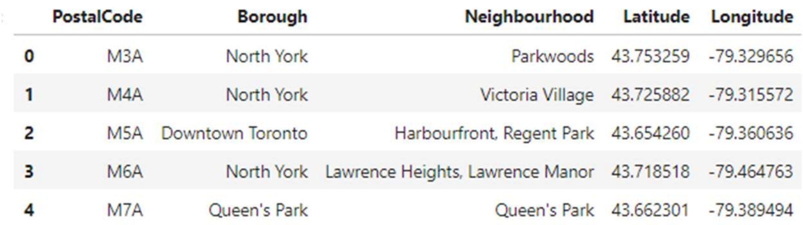


	PostalCode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront, Regent Park
3	M6A	North York	Lawrence Heights, Lawrence Manor
4	M7A	Queen's Park	Not assigned

- Only postal codes with an assigned borough was used.
- If more than one neighborhood exists for a postal code, they are combined into one row with the neighborhoods separated by a comma (e.g. the neighbourhood for M5A is Harbourfront, Regent Park).
- If a postal code has a borough but no neighbourhood assigned, the neighborhood will be the same as the borough (e.g. Queen’s Park).

d) Geospatial data

A csv file from http://cocl.us/Geospatial_data containing latitude and longitude data was used to create the following dataframe.

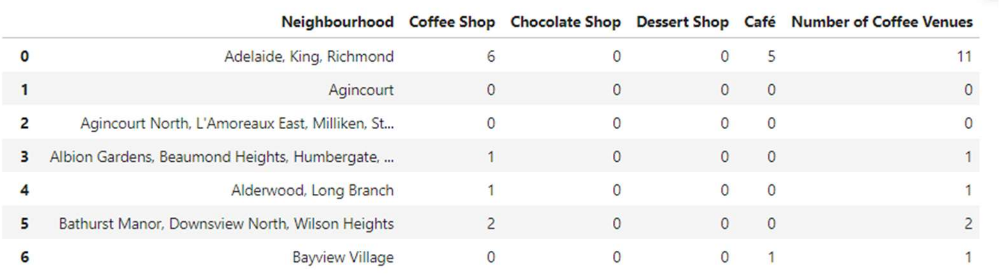


	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront, Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494

This data will be used for the purposes of calls to the Foursquare API as well as data visualization later.

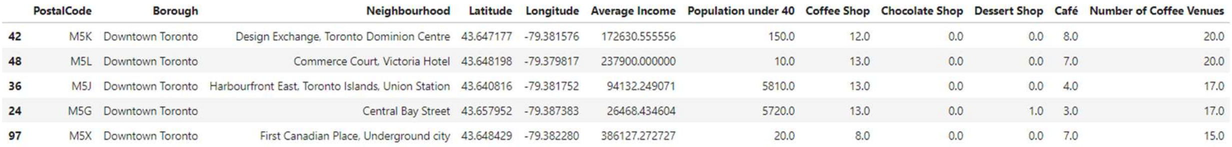
e) Existing Coffee venues

In order to understand the existing venues in each neighbourhood to gauge direct competition with the business, calls to the Foursquare API was made to search for all venues. The focus is then on particular venue categories of interest including Coffee Shop, Chocolate Shop, Dessert Shop and Café.



	Neighbourhood	Coffee Shop	Chocolate Shop	Dessert Shop	Café	Number of Coffee Venues
0	Adelaide, King, Richmond	6	0	0	5	11
1	Agincourt	0	0	0	0	0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0	0	0	0	0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	1	0	0	0	1
4	Alderwood, Long Branch	1	0	0	0	1
5	Bathurst Manor, Downsview North, Wilson Heights	2	0	0	0	2
6	Bayview Village	0	0	0	1	1

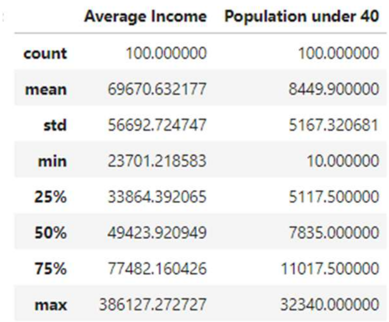
Finally, all data described above including Postal Code, Neighbourhood, Average Income, Population under 40, Latitude, Longitude, Number of Coffee Venues was merged into a central dataframe as follows, ready for next steps of the analysis.



	PostalCode	Borough	Neighbourhood	Latitude	Longitude	Average Income	Population under 40	Coffee Shop	Chocolate Shop	Dessert Shop	Café	Number of Coffee Venues
42	M5K	Downtown Toronto	Design Exchange, Toronto Dominion Centre	43.647177	-79.381576	172630.555556	150.0	12.0	0.0	0.0	8.0	20.0
48	M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817	237900.000000	10.0	13.0	0.0	0.0	7.0	20.0
36	M5J	Downtown Toronto	Harbourfront East, Toronto Islands, Union Station	43.640816	-79.381752	94132.249071	5810.0	13.0	0.0	0.0	4.0	17.0
24	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	26468.434604	5720.0	13.0	0.0	1.0	3.0	17.0
97	M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-79.382280	386127.272727	20.0	8.0	0.0	0.0	7.0	15.0

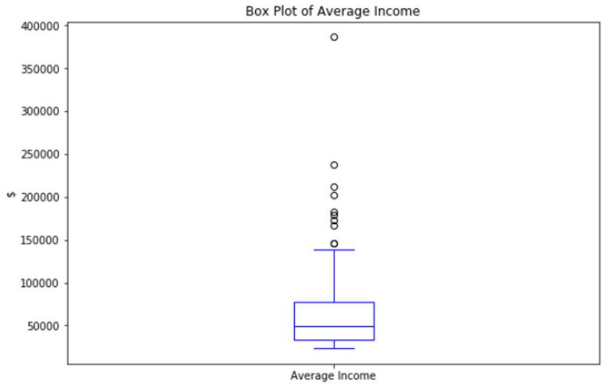
Methodology

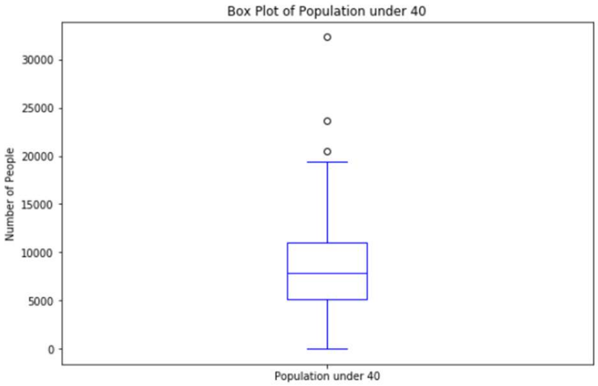
To further understand the data, especially around Average Income and Population under 40, exploratory data analysis was performed. The following is a descriptive statistics summary of the 100 neighbourhoods. Mean average income is $69k and population under 40 is 8450.



	Average Income	Population under 40
count	100.000000	100.000000
mean	69670.632177	8449.900000
std	56692.724747	5167.320681
min	23701.218583	10.000000
25%	33864.392065	5117.500000
50%	49423.920949	7835.000000
75%	77482.160426	11017.500000
max	386127.272727	32340.000000

Box plots further describe the distribution of the data.

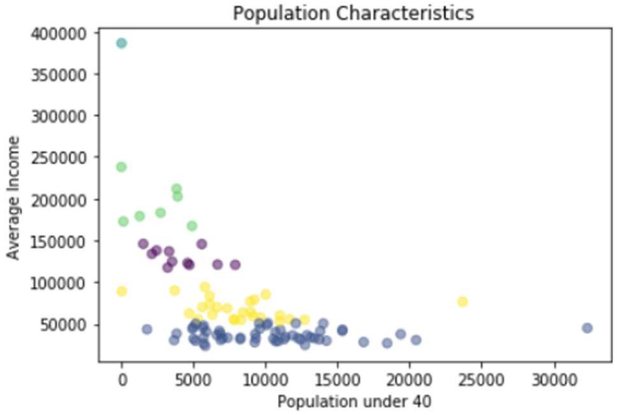




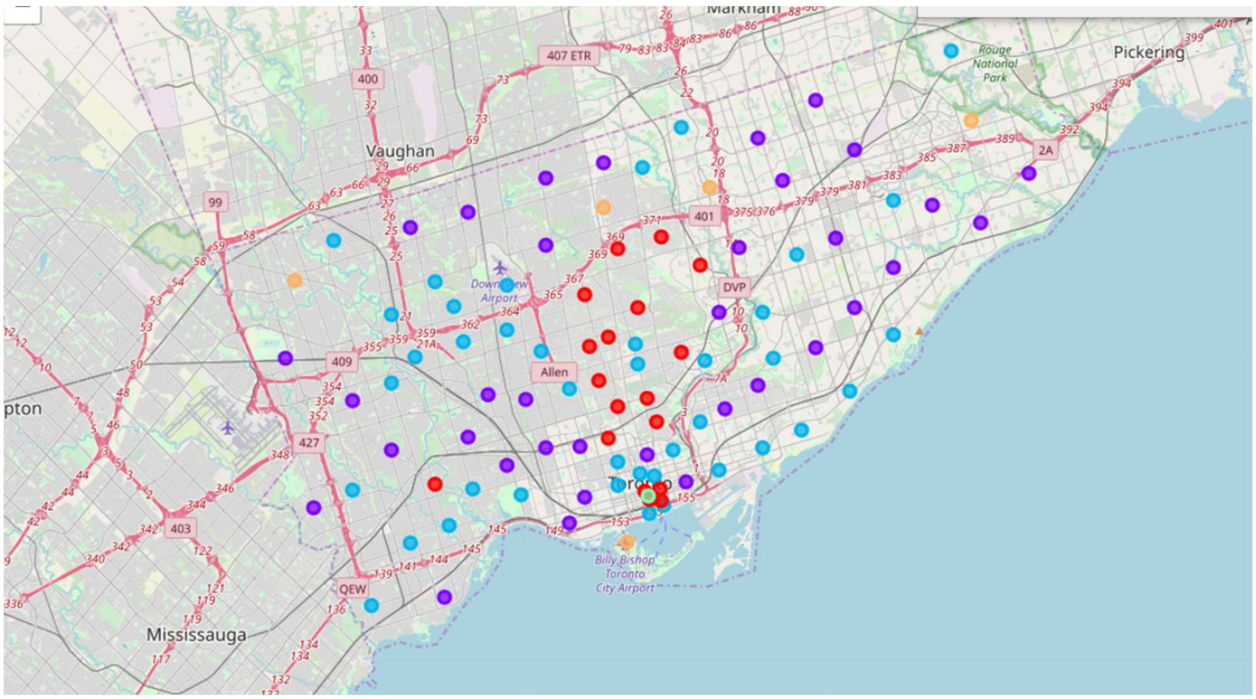
K-means was used for clustering to quickly discover insights from the unlabeled neighbourhood data based on Average Income and Population under 40. After the data is normalized over standard deviation, the number of clusters is set to 5 for fitting.

Results

Results from K-means return the following clustering results *in colour*.

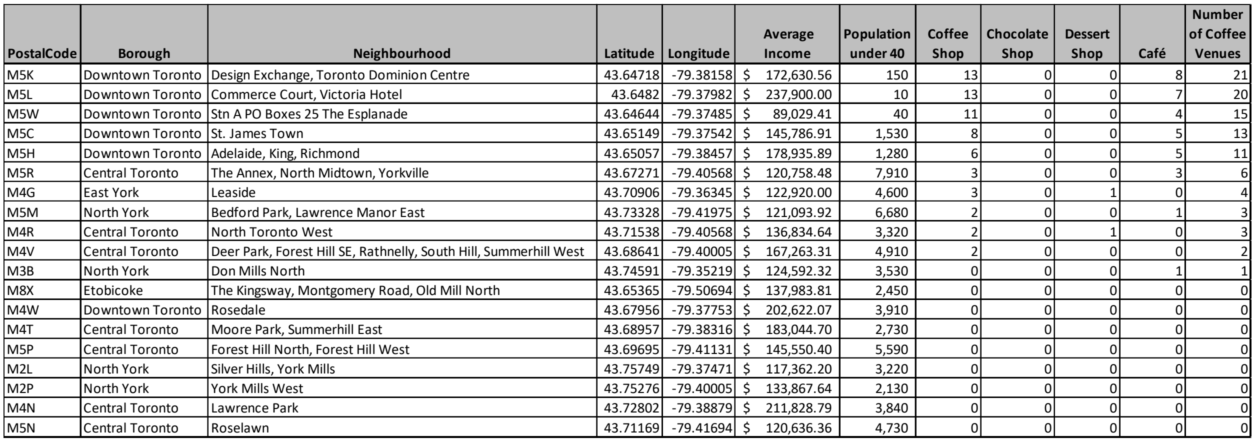


A map generated through folium allows further visualization of the results.



Discussion

The two-variable K-means classification analysis has segmented the neighbourhoods into 5 clusters. Since neighbourhoods with higher average income are targeted for the business, neighbourhoods in Cluster 1 as shown below would likely represent some good candidates for considerations.



PostalCode	Borough	Neighbourhood	Latitude	Longitude	Average Income	Population under 40	Coffee Shop	Chocolate Shop	Dessert Shop	Café	Number of Coffee Venues
M5K	Downtown Toronto	Design Exchange, Toronto Dominion Centre	43.64718	-79.38158	$ 172,630.56	150	13	0	0	8	21
M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.6482	-79.37982	$ 237,900.00	10	13	0	0	7	20
M5W	Downtown Toronto	Stn A PO Boxes 25 The Esplanade	43.64644	-79.37485	$ 89,029.41	40	11	0	0	4	15
M5C	Downtown Toronto	St. James Town	43.65149	-79.37542	$ 145,786.91	1,530	8	0	0	5	13
M5H	Downtown Toronto	Adelaide, King, Richmond	43.65057	-79.38457	$ 178,935.89	1,280	6	0	0	5	11
M5R	Central Toronto	The Annex, North Midtown, Yorkville	43.67271	-79.40568	$ 120,758.48	7,910	3	0	0	3	6
M4G	East York	Leaside	43.70906	-79.36345	$ 122,920.00	4,600	3	0	1	0	4
M5M	North York	Bedford Park, Lawrence Manor East	43.73328	-79.41975	$ 121,093.92	6,680	2	0	0	1	3
M4R	Central Toronto	North Toronto West	43.71538	-79.40568	$ 136,834.64	3,320	2	0	1	0	3
M4V	Central Toronto	Deer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West	43.68641	-79.40005	$ 167,263.31	4,910	2	0	0	0	2
M3B	North York	Don Mills North	43.74591	-79.35219	$ 124,592.32	3,530	0	0	0	1	1
M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.65365	-79.50694	$ 137,983.81	2,450	0	0	0	0	0
M4W	Downtown Toronto	Rosedale	43.67956	-79.37753	$ 202,622.07	3,910	0	0	0	0	0
M4T	Central Toronto	Moore Park, Summerhill East	43.68957	-79.38316	$ 183,044.70	2,730	0	0	0	0	0
M5P	Central Toronto	Forest Hill North, Forest Hill West	43.69695	-79.41131	$ 145,550.40	5,590	0	0	0	0	0
M2L	North York	Silver Hills, York Mills	43.75749	-79.37471	$ 117,362.20	3,220	0	0	0	0	0
M2P	North York	York Mills West	43.75276	-79.40005	$ 133,867.64	2,130	0	0	0	0	0
M4N	Central Toronto	Lawrence Park	43.72802	-79.38879	$ 211,828.79	3,840	0	0	0	0	0
M5N	Central Toronto	Roselawn	43.71169	-79.41694	$ 120,636.36	4,730	0	0	0	0	0

There is also a decent Population under 40 in most of these neighbourhoods giving a good potential customer base. However, it should be noted that a few neighbourhoods, especially those in the Downtown Toronto area, seem to have a number of existing coffee-related venues already, pointing to potential competitions. The client may want to avoid these neighbourhoods as a result.

Other clusters contain neighbourhoods that fit to a lesser extent the criteria of the business. One of the clusters has actually been captured as an outlier and should be ignored.

Conclusion

Using a combination of statistics and machine learning tools, we are able to address the client's problem and arrive at a list of suggested neighbourhoods that are suitable location candidates for her future gourmet coffee shop business.