Analyzing the various regions in New Brunswick for Starting a Convenience Store

Applied Data Science Capstone Project

By: Pruthvi Brahmbhatt

Table of Contents

1. Introduction

New Brunswick is one of the 13 provinces in Canada, part of the three Maritime provinces and included as one of the four Atlantic provinces. In recent years, the province has attracted a critical mass of businesses in sectors such as advanced manufacturing, cybersecurity, and digital health, among others. This would invariably attract further participants, both customers and other businesses alike, leading to great opportunities and realization of potential.

Thus, I have chosen to focus on New Brunswick to understand and explore potential opportunities that can be obtained using machine learning techniques and the Foursquare API. Given that there is an expected growth in population and economy through the development of New Brunswick, I plan to focus on finding a suitable location to operate a convenience store. The goal of this project is to use the Foursquare location data and K-Means clustering technique of the venue information to determine the ideal location in New Brunswick to open a convenience store.

Although the analysis in this report may be specifically set to those individuals wishing to open a convenience store, a similar analysis can be performed on any other type of venue as well, thereby including any individual wishing to identify potential opportunities within our target audience.

2. Data Acquisition

As performed in the previous assignments, I have obtained the postal code information and the associated Forward Sortation Area (FSA) via Wikipedia and, after standard editing practices, used the *urllib* module to obtain the location data, i.e., the latitudinal and longitudinal data, for each of the FSAs obtained. We make use of FSAs as a substitute for major centres since they are more structurally ideal to cover most of the active regions in New Brunswick.

After this, we use the *Nominatim* module to obtain the location data of New Brunswick. Finally, we make us of the Foursquare API to obtain the frequently visited venues in proximity to each of the FSA location, with which we perform our analysis.

It is important to note that there are empty elements within the initial Wikipedia data, where postal codes have not yet been assigned to any region. There were also other empty information regarding venues for certain regions as well. I have removed these rows from the dataset as and when encountered.

3. Methodology

Constructing the Dataset

I first obtained the link for the Wikipedia page containing our required information (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_E), after which I employed the web scraping tools and the *BeautifulSoup* module to obtain the Wikipedia html page and access the table within, with which we obtain our first dataset *nb_data*. The *nb_data* dataset contains the Postal Code and associated FSA location information.

	index	Postal Code	Location
0	0	E1A	Moncton East (Dieppe)
1	1	E2A	Bathurst
2	2	E3A	Fredericton North
3	3	E4A	Chipman
4	4	E5A	Moores Mills
...	...	...	...
175	175	E5Z	Not assigned
176	176	E6Z	Not assigned
177	177	E7Z	Not assigned
178	178	E8Z	Not assigned
179	179	E9Z	Not assigned

180 rows × 3 columns

Figure 1: nb_data dataframe

After removing the rows that have not been assigned a location yet, I moved on to obtaining the longitudinal and latitudinal data of each of the locations in our dataset and merge this new information to *nb_data.*

	level_0	Postal Code	Location	Latitude	Longitude
0	0	E1A	Moncton East (Dieppe)	46.11647835	-64.67556069377879
1	1	E2A	Bathurst	47.626529	-65.654297
2	2	E3A	Fredericton North	45.947959350000005	-66.65336235897576
3	3	E4A	Chipman	46.1778309	-65.8774732
4	4	E5A	Moores Mills	45.2863562	-67.2888138
...	...	...	...	...	...
105	153	E1X	Tracadie-Sheila	47.5146399	-64.90893552417663
106	156	E4X	St-Louis-de-Kent	46.7377084	-64.9719715
107	164	E3Y	Grand Falls Northeast	47.0463119	-67.7393601
108	165	E4Y	Rogersville	46.73286	-65.4255381
109	174	E4Z	Petitcodiac	45.9361658	-65.1759490204391

110 rows × 5 columns

Figure 2:nb_data Dataframe after merge

We can see here that there are 110 locations considered in the end from our dataset. The final stage of our data construction phase involves including the list of venues within a 1.5km radius at each of the locations present in the dataset. We name this new dataset as *final_venues*.

	Location	Location Latitude	Location Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Moncton East (Dieppe)	46.11647835	-64.67556069377879	TriStar Mercedes-Benz Moncton	46.123225	-64.683409	Auto Dealership
1	Moncton East (Dieppe)	46.11647835	-64.67556069377879	Greater Moncton International Airport (YQM) (G...	46.116169	-64.688426	Airport
2	Moncton East (Dieppe)	46.11647835	-64.67556069377879	Moncton Airport Security checkpoint	46.116188	-64.688283	Airport Service
3	Moncton East (Dieppe)	46.11647835	-64.67556069377879	Post Security Lounge	46.116221	-64.688389	Airport Lounge
4	Moncton East (Dieppe)	46.11647835	-64.67556069377879	Avis Car Rental	46.116349	-64.688412	Rental Car Location
...	...	...	...	...	...	...	...
999	Rogersville	46.73286	-65.4255381	Pizza Delight	46.736983	-65.430956	Pizza Place
1000	Petitcodiac	45.9361658	-65.1759490204391	Tim Hortons	45.931528	-65.168804	Coffee Shop
1001	Petitcodiac	45.9361658	-65.1759490204391	Foodland - Petitcodiac	45.931855	-65.175998	Grocery Store
1002	Petitcodiac	45.9361658	-65.1759490204391	Circle K	45.932447	-65.168782	Convenience Store
1003	Petitcodiac	45.9361658	-65.1759490204391	Shell	45.931752	-65.168671	Gas Station

1004 rows × 7 columns

Figure 3: final_venues Dataframe

We can observe from here, and within the code, that there are 1004 venues listed in total, of which there are 137 unique categories of venues. Now that our required datasets have been constructed (*nb_data* and *final_venues*), we move on to exploring the dataset.

Exploring the Dataset

As part of the exploration phase, I produced a map of New Brunswick, marked with the locations in our dataset, as shown below.

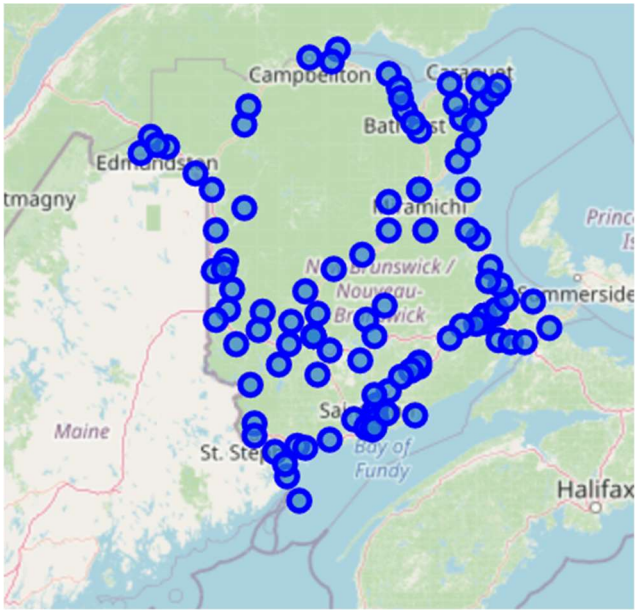


Figure 4: Map of New Brunswick marked with locations

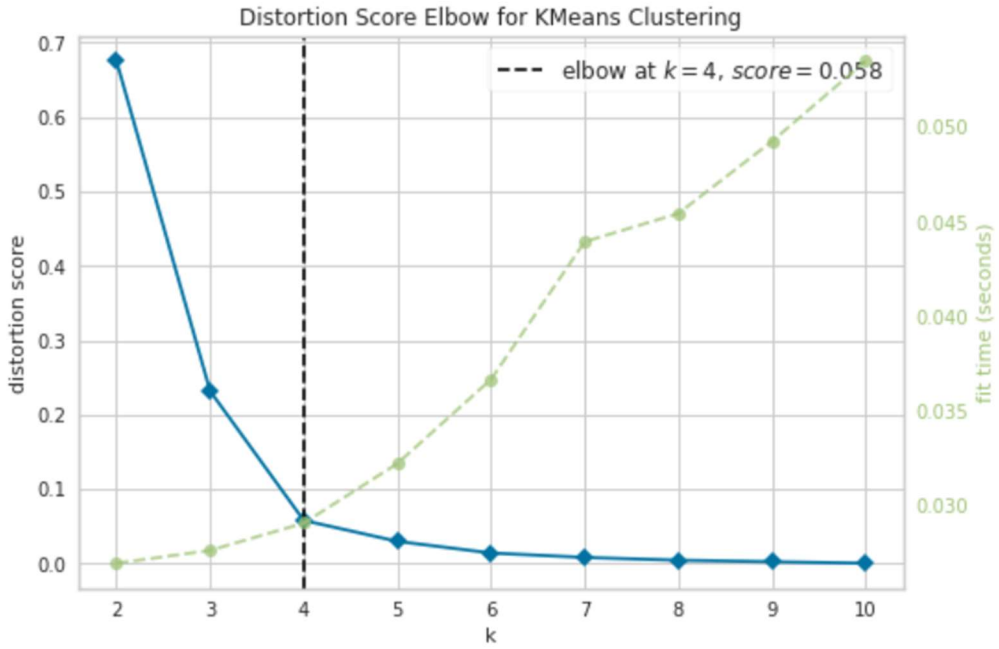
Employing K-Means Clustering

Before we begin the employing the technique, we need to create the data that forms the basis of the clustering pattern. Since our focus is on convenience stores specifically, we move to calculating the average occurrence of convenience stores per location.

First, we perform one-hot encoding to convert categorical data into numerical data, with which we group by location and obtain the mean of the occurrences of convenience stores. The resulting dataset would look as follows, with Convenience Store column referring to the mean occurrence:

	Location	Convenience Store
0	Allardville	0.0
1	Apohaqui	0.0
2	Baie-Sainte-Anne	0.0
3	Balmoral	0.5
4	Bath	0.0

After this, we move to finding the optimal K value to form our clusters. In other words, we attempt to find the optimal number of clusters required to cluster our data. We perform this by finding the Distortion Score Elbow, a plot of which is rendered belo



We can see from our plot above that the optimal number of clusters required is 4, with a score of 0.058.

We apply the K-Means clustering and plot the map with the clustered marked in colour.

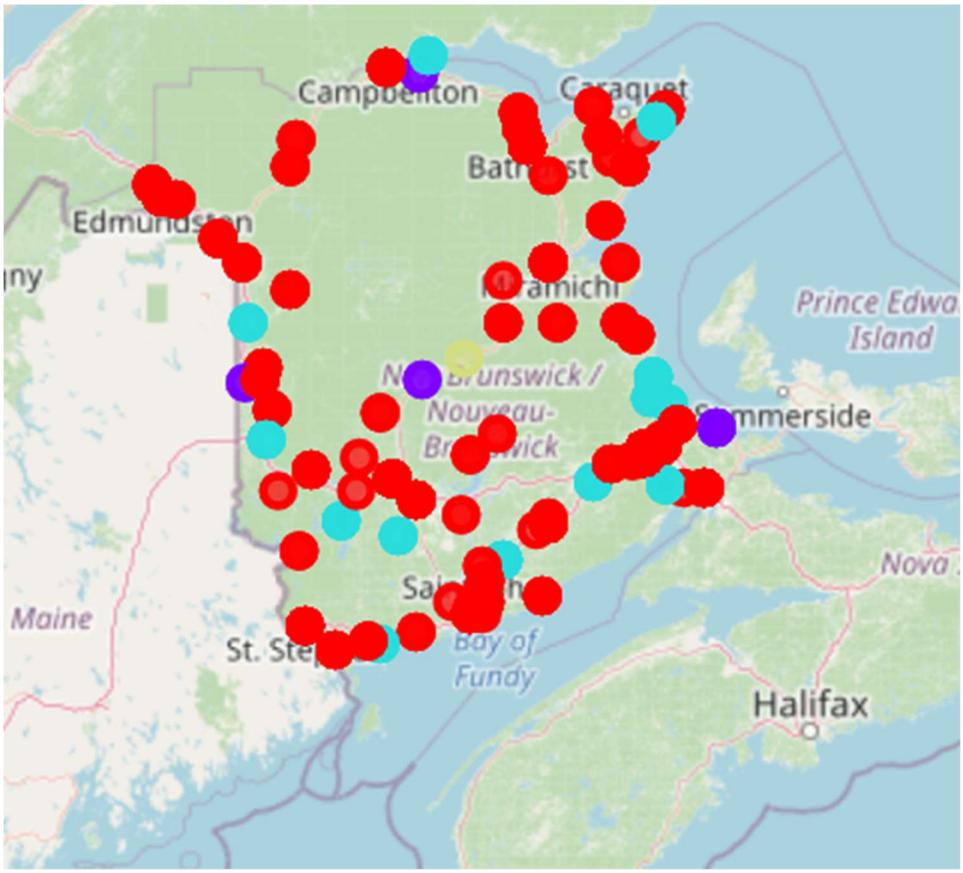


Figure 5: Map of New Brunswick with mapped clusters

4. Results

We obtain the following clusters

Cluster 1:

	Location	Convenience Store	Cluster Labels	Venue	Venue Latitude	Venue Longitude	Venue Category
49	Moncton Central	0.031250	0	Circle K	46.085569	-64.807102	Convenience Store
51	Moncton Northwest	0.031250	0	Circle K	46.085569	-64.807102	Convenience Store
52	Moncton West	0.031250	0	Circle K	46.085569	-64.807102	Convenience Store
90	Sussex	0.083333	0	Circle K	45.722621	-65.506368	Convenience Store

Cluster 2:

	index	Location	Convenience Store	Cluster Labels	Venue	Venue Latitude	Venue Longitude	Venue Category
0	3	Balmoral	0.5	1	Resto Chez Madie	47.973981	-66.440265	Restaurant
1	3	Balmoral	0.5	1	Depanneur Yatout	47.975966	-66.432701	Convenience Store
2	8	Boiestown	0.5	1	Circle K	46.456452	-66.417791	Convenience Store
3	8	Boiestown	0.5	1	Carr's Computer Service.	46.454861	-66.426926	Electronics Store
4	13	Cap-Pelé	0.4	1	Bel-Air Take Out	46.214923	-64.267940	Seafood Restaurant
5	13	Cap-Pelé	0.4	1	PJC Jean Coutu Santé-Beauté	46.217005	-64.282261	Pharmacy
6	13	Cap-Pelé	0.4	1	Circle K	46.215295	-64.269818	Convenience Store
7	13	Cap-Pelé	0.4	1	Doiron Market	46.214729	-64.269123	Convenience Store
8	13	Cap-Pelé	0.4	1	Fred's Bakery	46.215764	-64.291854	Bakery
9	14	Centreville	0.5	1	Grandma's	46.431306	-67.708973	Burger Joint
10	14	Centreville	0.5	1	M & D	46.433310	-67.700442	Convenience Store

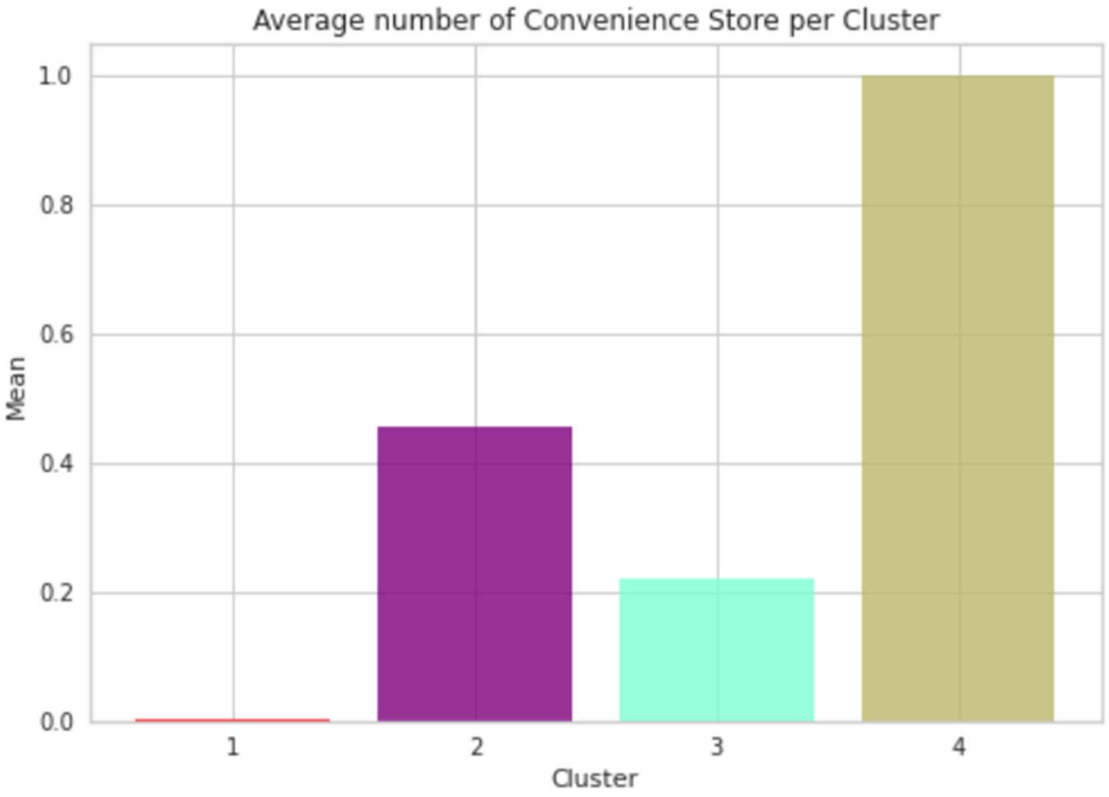
Cluster 3:

	index	Location	Convenience Store	Cluster Labels	Venue	Venue Latitude	Venue Longitude	Venue Category
0	9	Bouctouche	0.142857	2	Tim Hortons	46.466165	-64.731514	Coffee Shop
1	9	Bouctouche	0.142857	2	Forum de Bouctouche	46.472280	-64.721771	Hockey Arena
2	9	Bouctouche	0.142857	2	dixie lee	46.471701	-64.721616	Snack Place
3	9	Bouctouche	0.142857	2	Centre James K Irving	46.475277	-64.719850	Hockey Arena
4	9	Bouctouche	0.142857	2	Circle K	46.469035	-64.727952	Convenience Store
5	9	Bouctouche	0.142857	2	Irving arboretum	46.480613	-64.713317	Park
6	9	Bouctouche	0.142857	2	Co-Op Bouctouche	46.464003	-64.735503	Grocery Store
7	16	Cocagne	0.333333	2	Cocagne Marina	46.334442	-64.623896	Harbor / Marina
8	16	Cocagne	0.333333	2	Cocagne Variety	46.340042	-64.619507	Convenience Store
9	16	Cocagne	0.333333	2	Arena Cocagne	46.335390	-64.627404	Hockey Arena
10	18	Dalhousie	0.250000	2	Tim Hortons	48.066179	-66.374769	Coffee Shop
11	18	Dalhousie	0.250000	2	Days Inn	48.065309	-66.373026	Hotel

Cluster 4

	Location	Convenience Store	Cluster Labels	Venue	Venue Latitude	Venue Longitude	Venue Category
19	Doaktown	1.0	3	B & L Convenience Store	46.569961	-66.103779	Convenience Store

We also plot the measure of average number of convenience stores per cluster below



5. Discussion

The plot above shows the average number of convenience stores per cluster. We can observe that Cluster 4 has the highest average number of convenience stores, followed by Cluster 2, Cluster 3 and finally, Cluster 1. In other words, Cluster 4 has the highest number of convenience stores per location, whereas Cluster 1 has the lowest number of convenience stores per location.

This result indicates that opening a convenience store in a location within Cluster 1 would be apt, given that there is a booming market of multiple venue categories, signifying a bustling market open with opportunities. On the other hand, the location within Cluster 4 already has a convenience store available at any given location, with no other venues listed, indicating little room for growth.

Since any average lower than 1 indicates that there is a lesser number of convenience stores compared to the number of locations, Cluster 3 and Cluster 4 are also potential collections of locations where bringing in a convenience store seems apt.

6. Conclusion

Based on our analysis, we find that opening a convenience store in any location within Cluster 1 has potential for growth, due to a relatively well running economy in the area. Cluster 2 and Cluster 3 are other potential list of candidates to investigate, while Cluster 4 can be avoided.

Although this analysis attempts to answer the question of where a convenience store could be opened for great future growth, further analysis can be performed with a variety of datasets to provide a more clear and accurate description. One suggestion could be to include user information from Foursquare API to show the number of visits to these venues and reviews of the place. This unfortunately requires the upgraded feature and thus we restrict ourselves to the above analysis for now.

Another issue that can be tackled to gain a better picture, is to obtain neighborhood information in each of our locations. Looking into more detailed datasets could provide a better picture overall.