ANALYZING THE VARIOUS REGIONS IN NEW BRUNSWICK FOR STARTING A CONVENIENCE STORE

APPLIED DATA SCIENCE CAPSTONE PROJECT

BY: PRUTHVI BRAHMBHATT

INTRODUCTION

- New Brunswick is one of the 13 provinces in Canada and is part of the three Maritime provinces and included as one of the four Atlantic provinces
- In recent years, an increasing level of attention has been provided in attracting and retaining customers and businesses to the province.
- For this reason, I have chosen to focus on New Brunswick to understand and explore potential hidden opportunities using the Foursquare API data and machine learning techniques
- 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.

DATA ACQUISITION

- The data used in this project are
  - Postal code and location information via Wikipedia

    Link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_E
  - Latitudinal and Longitudinal data via *urrlib* and *Nominatim* modules in python
  - Venues and venue information within 1.5k from each location using Foursquare API
- Multiple empty entries exist within these datasets, which were removed as part of the data cleansing process

METHODOLOGY

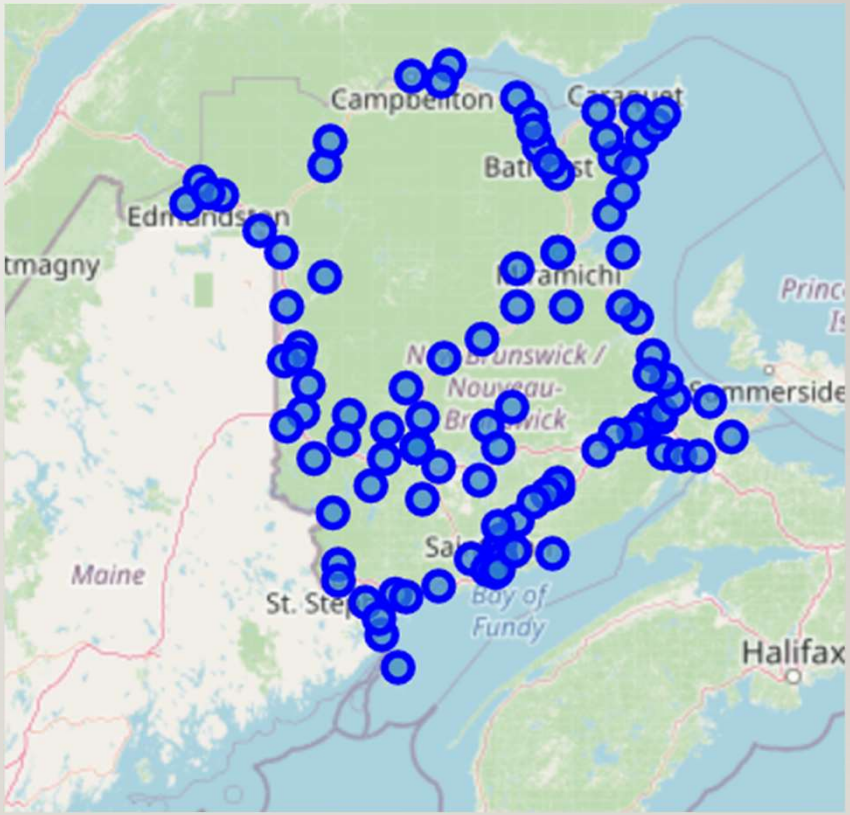
- Constructing the dataset
- Exploring the dataset
- Employing the K-Means Clustering technique

CONSTRUCTING THE DATASET

- Two datasets were formed in our project
  - *Nb_data:* Consists of postal code and location information merged with latitude and longitude information
  - *Final_venues:* Consists of venues and information in each of the location
- To obtain the postal code and location information, web scraping tools and the BeautifulSoup module were employed on the link.
- We observe that there are 110 locations (without empty entries) considered in New Brunswick, with 1004 venues in total.
- There were 137 unique venue categories found.

EXPLORING THE DATASET

- We obtained the map of New Brunswick with the locations marked, to observe our data.



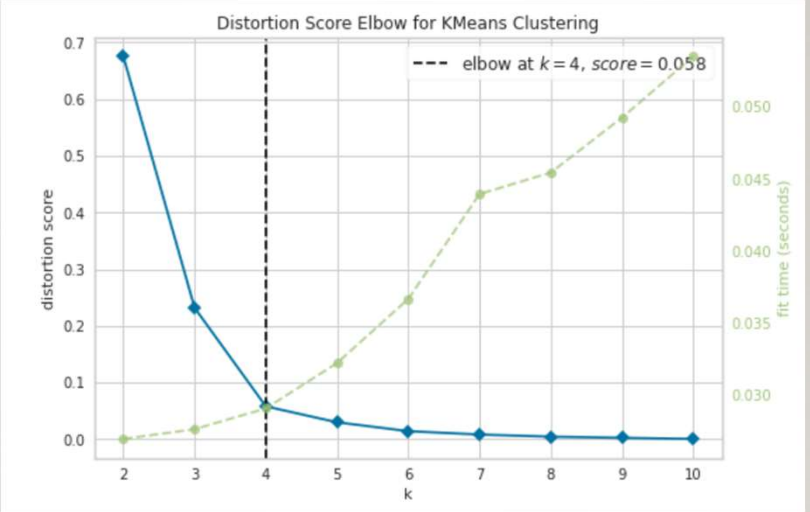
EMPLOYING K-MEANS CLUSTERING

- Before we move to the clustering technique, we create a dataset that can be used for clustering purposes
- On the side is a sample of the dataset to illustrate the purpose. Here, *Convenience Store* represents the average score of convenience store occurrences in the area

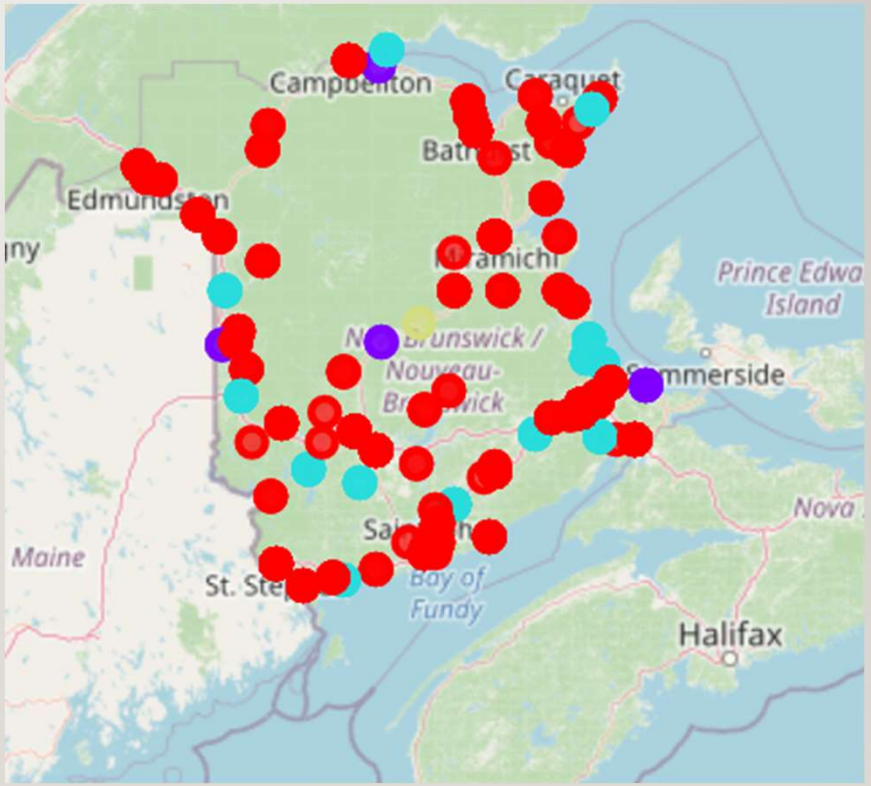
	Location	Convenience Store
0	Allardville	0.0
1	Apohaqui	0.0
2	Baie-Sainte-Anne	0.0
3	Balmoral	0.5
4	Bath	0.0

EMPLOYING K-MEANS CLUSTERING

- We use the Distortion Score Elbow technique to obtain the optimal number of clusters required.
- From the plot, we can observe that the optimal number of clusters (K) is 4.
- Now we have prepared the necessary inputs, we perform the clustering technique and obtain our clusters

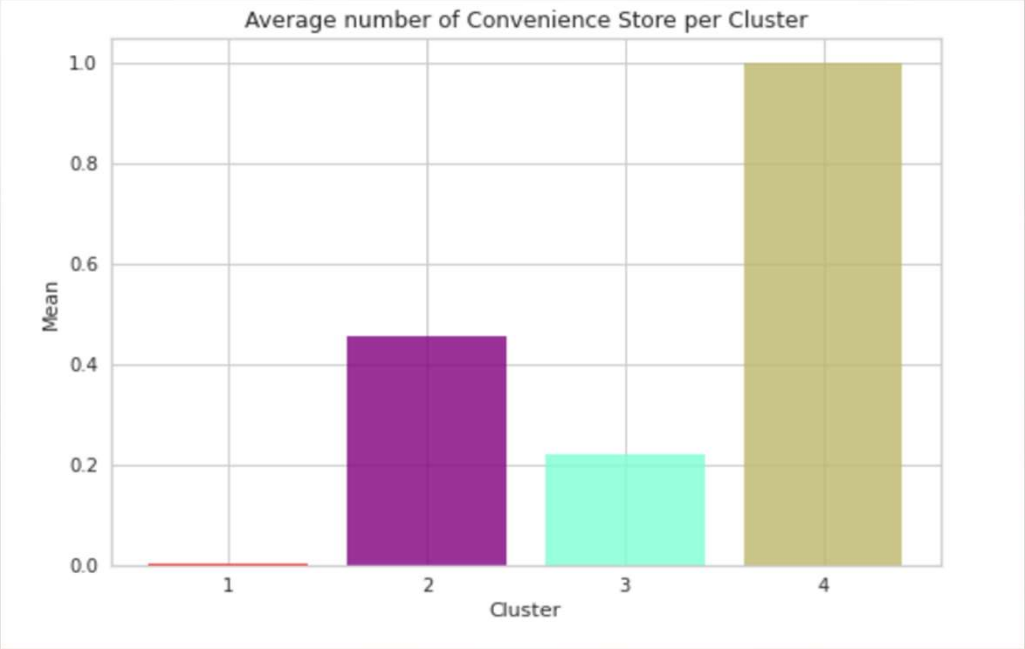


CLUSTERED MAP OF NEW BRUNSWICK



RESULTS

- The plot on the side plots the score of average occurrences of convenience stores in each cluster
- As can be observed, Cluster 4 shows the highest occurrence per cluster, while Cluster 1 shows the lowest.



DISCUSSION

- A score lower than 1 from the previous plot indicates that there is a lower number of venues located, which would indicate a lower demand of services in the location
- Given this, opening a convenience store in a location within Cluster 1 would be ideal, 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 seems to have a less thriving demand, with no other venues listed, indicating little room for growth.
- Locations in Cluster 3 and Cluster 4 are also potential regions where bringing in a convenience store seems ideal.

CONCLUSION

- To conclude , our analysis indicates 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.
- Doaktown , the only location in Cluster 4, seems to be a possible location to avoid opening a convenience store.
- 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, which requires further access to Foursquare.
  - 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.