USING MACHINE LEARNING TO FIND LOCATIONS TO OPEN A BAKERY IN TORONTO

JUNE 3, 2020/ IBM APPLIED DATA SCIENCE CAPSTONE

SUBMITTED BY MEENU CHOUDHARY

Introduction

The aim of capstone project is to develop a battle of neighborhood for opening a Bakery in Toronto. The ideas is to find a neighborhood where there are no or not many Bakery. This will provide a great insight to entrepreneurs looking for opening a Bakery business opportunity in Toronto. Bakery is one of the basic need of people. The results of this project will add great value in the decision making for entrepreneurs or existing bakery owners who wants to expand to other neighborhood.

Business Problem

The objective is to find the most suitable location for the entrepreneur or existing bakery owner to open a new bakery in Toronto, Canada. Using clustering by k-means of machine learning methods in data science, this project aims to provide insight to answer the business question: Which location is the most suitable for opening a bakery by an entrepreneur looking for business opportunity or existing bakery owner looking for expansion?

Target Audience

The entrepreneur who is looking for business opportunity or the existing bakery owner who wants to expand to new location.

Data

To solve this problem, I will need below data:

- List of neighborhoods in Toronto, Canada.

	Postalcode	Borough	Neighborhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

- Latitude and Longitude of these neighborhoods.

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

- Venue data related to Bakery. This will help us find the neighborhoods that are most suitable to open a Bakery.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Berczy Park	56	56	56	56	56	56
Brockton, Parkdale Village, Exhibition Place	22	22	22	22	22	22
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	16	16	16	16	16	16
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	17	17	17	17	17	17
Central Bay Street	65	65	65	65	65	65
Christie	17	17	17	17	17	17
Church and Wellesley	78	78	78	78	78	78
Commerce Court, Victoria Hotel	100	100	100	100	100	100
Davisville	32	32	32	32	32	32
Davisville North	9	9	9	9	9	9
Dufferin, Dovercourt Village	13	13	13	13	13	13
First Canadian Place, Underground city	100	100	100	100	100	100
Forest Hill North & West, Forest Hill Road Park	4	4	4	4	4	4
Garden District, Ryerson	100	100	100	100	100	100

Extracting the data

- Scrapping of Toronto neighborhoods via Wikipedia.
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package or using a csv file.
- Using Foursquare API to get a venue data related to these neighborhoods.

Methodology

After defining the business problem, the next step is to gather data which contains neighborhood information of Toronto, Canada. The data is scrapped from Wikipedia page("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M") using pandas html table scraping method. The BeautifulSoup package in Python is used.

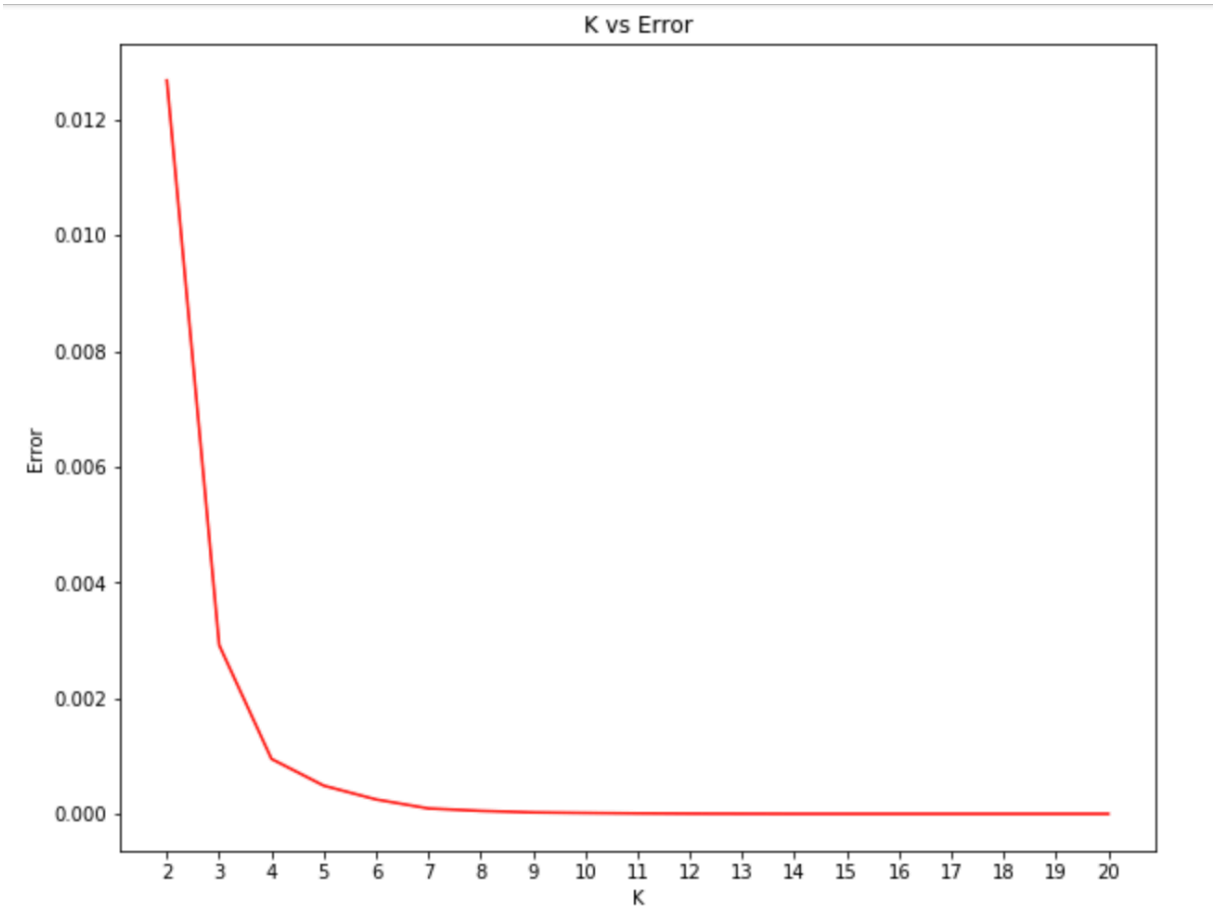
The data contains postal code, borough and neighborhood information. But the location data is also needed for the analysis. The location data can be imported using Geocoder package in Python or else the csv file can be used if available.

After location information of neighborhood, the next step is to find the venues under 1 km from Foursquare using credentials. Credentials can be obtained by creating an account on Foursquare. The location coordinates is utilized by Foursquare to pull the list of venues near these neighborhoods. The visualization map is analyzed using Folium package to verify whether these are correct coordinates.

From Foursquare, the names, categories, latitude and longitude information of the venues can be pulled. The analysis how many unique categories in each venues and grouping them by taking mean values on the frequency of occurrence of each venue category.

After that look for a specific category (in this project, it is Bakery) at each venue. Then, perform the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centeriods, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for business problem in hand. The no. of cluster can be random value or the elbow plot can be used to get the no. of cluster value. In this project, the neighborhoods are clustered in 4 based on elbow method.

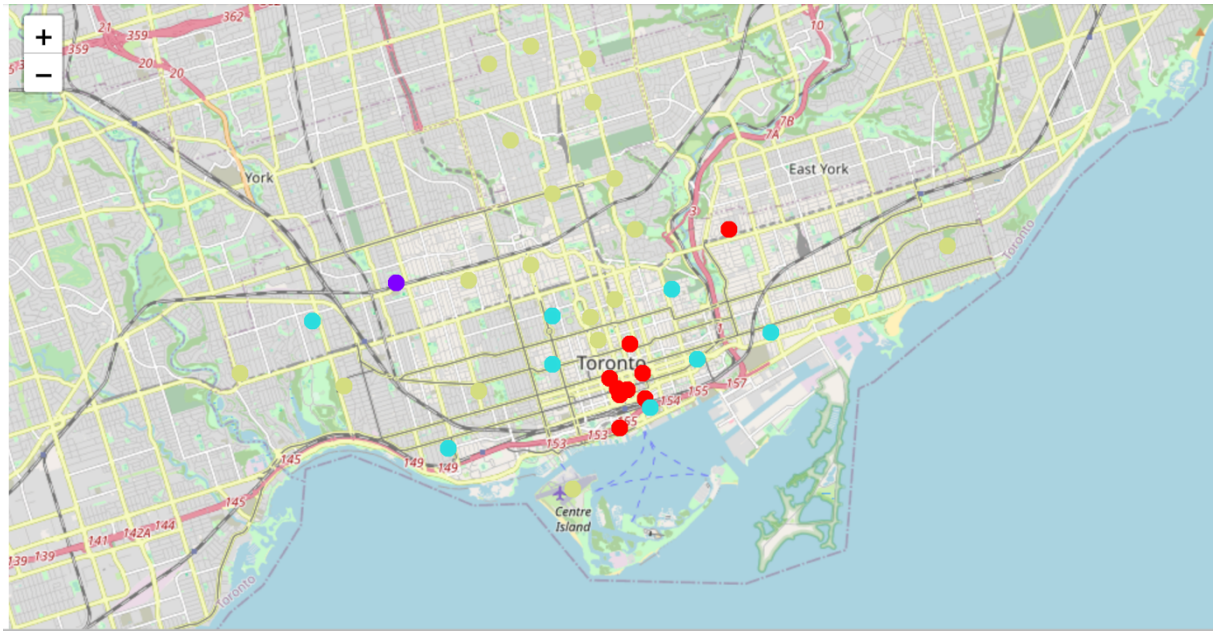


The analysis on clusters will give many insights such as number of Bakery in each clusters along with the neighborhood information. Based on the concentration of

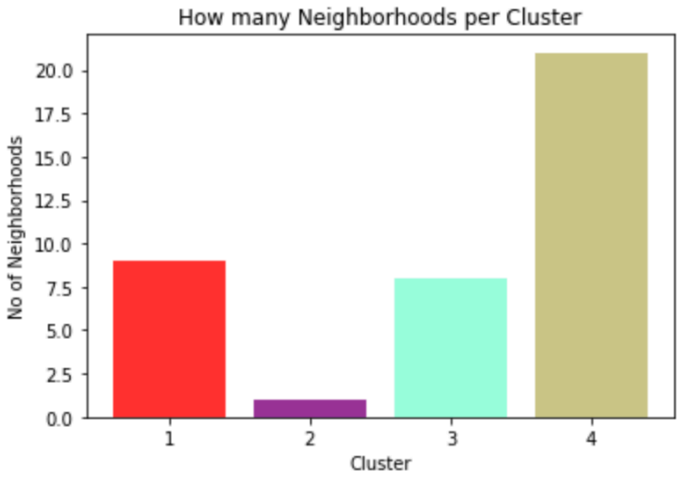
clusters, the suitable location can be proposed to entrepreneur or existing bakery owner. The neighborhood in the cluster with minimum mean value could be the best location for opening a new bakery.

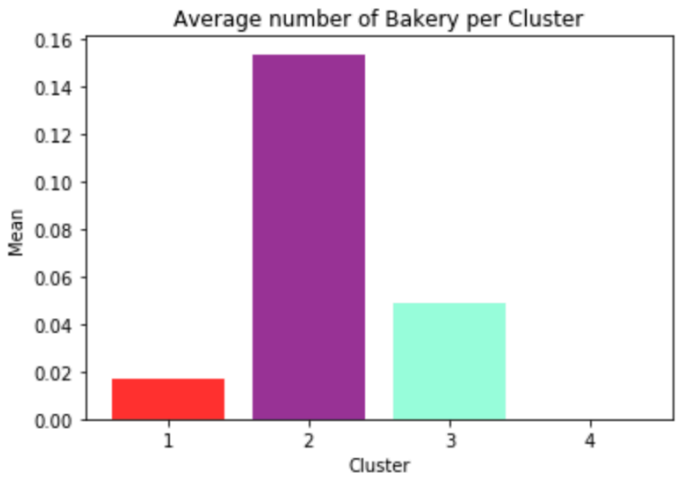
Results

Clusters



The results from k-means clustering show that Toronto neighborhoods can be clustered into 4 clusters based on number of Bakery located in each neighborhood:





- Cluster 0: 9 Neighborhoods with less no. of Bakery
- Cluster 1: 1 Neighborhoods with highest number of Bakery
- Cluster 2: 8 Neighborhoods with of Bakery
- Cluster 3: 21 Neighborhoods with zero Bakery

The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in purple color ,Cluster 2 in bright green color and Cluster 3 in light green color.

Recommendations

There is no bakery in cluster 4 where there are highest number of neighborhood. Therefore, this project recommends the entrepreneur or existing bakery owner to open a bakery in these locations with no competition.

Limitations and Suggestions for Future Research

In this project, the consideration of one factor is accounted: the occurrence / existence of Bakery in each neighborhood. There are many other factors such as population density, income of residents, rent could influence the decision to open a new bakery. The detail analysis will take more time and might not be possible to include in this project. Future research can take other consideration into account for more accurate analysis.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendation to decision maker.

References

List of neighborhoods in
Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

All codes for this project can be found here:
https://github.com/MeenuChoudhary/Coursera_Capstone/blob/master/Capstone%20Assignment_Bakery.ipynb