COURSERA CAPSTONE PROJECT

IBM Applied Data Science

French Patisserie in Toronto, ON, Canada

by: Jamie Chen

October, 2020

TABLE OF CONTENTS

INTRODUCTION/BUSINESS PROBLEM

Bakery industry is a $3.5 Billion USD (2017) industry in Canada. It is expected to continue to grow by a compound annualized rate of 3.4%. While it is a completive market with moderate growth, with passion and skill, it is a great industry to be in.

Celeste has just finished culinary school and graduated top of her class from Le Cordon Bleu in Paris, France. She is going back home to Toronto, Canada. Her proud parents have secured funding to invest in her dream French patisserie. Her ideal location would be in a place that has other businesses around but not too many other bakeries in downtown Toronto area.

DATA

- List of postal codse, neighborhoods, and burough's in Toronto.
- Latitude and longitide coordinates of those neighborhoods.
- Foursquare location date which contains venues in each of the neighborhoods.
- This data will be used to perform clustering to determine if a certain neighborhood is suitable for Celeste's business.

METHODOLOGY

To solve Celeste's problem, we will use the Foursquare location data and postal-code list data from Wikipedia and Geospatial Coordinates data in combination.

We firstly scraped the postal-code list from URL: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. The list from Wikipedia contains Canadian Postal code with the beginning of M, and the corresponding borough and neighborhood.

We will then acquire the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we combine the postal codes list table with Geospatial coordinates which includes each neighborhoods' longitude and latitude. After these processes have been done, we begin to explore the neighborhoods in Toronto by using Foursquare location data. For instance, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters.

With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be found from all the returned venues. Then we will analyse each neighborhood by grouping the rows with "Downtown Toronto" and taking the mean of the frequency of occurrence of each venue category. Since we are analyzing the "Bakery" data, we will filter the "Bakery" as venue category for the neighborhoods.

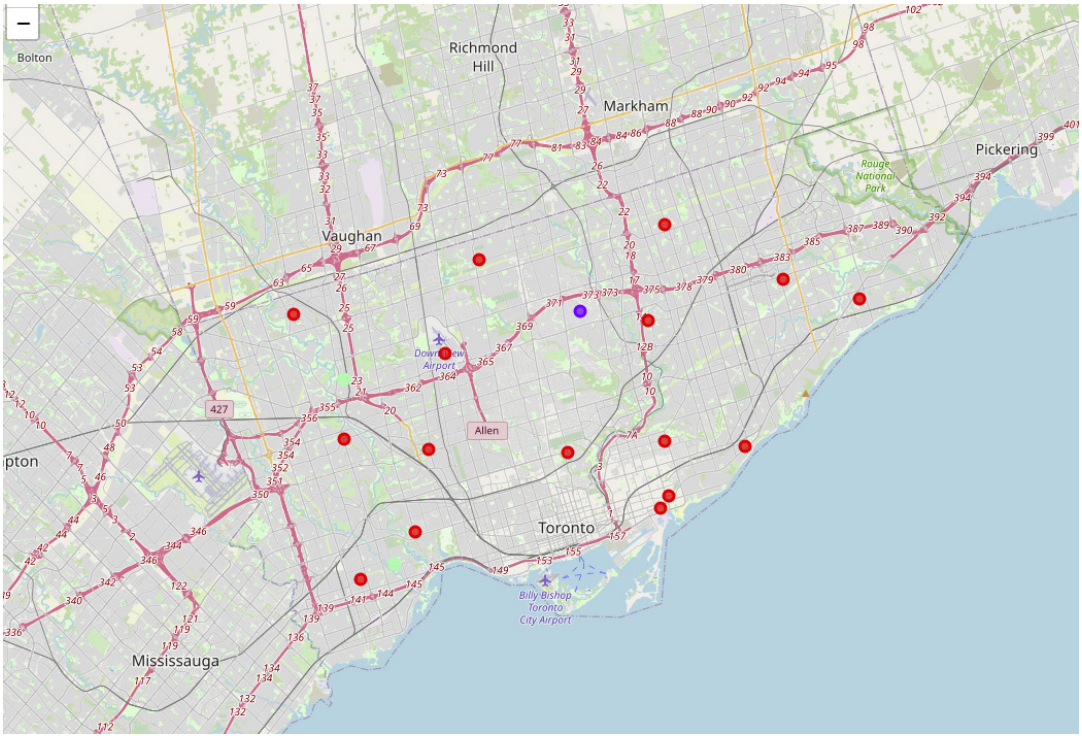
Then according to the data frame, we can figure out the appropriate neighborhood for Celeste to open her patisserie without too many competitors.

RESULTS

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence of “Bakery”:

- Cluster 0: Neighborhoods with very small number of bakery
- Cluster 1: Neighborhoods with relatively small number of bakery
- Cluster 2: Neighborhoods with relatively higher concentration of bakery

The results of the clustering are visualized in the map below with cluster 0 in red cluster 1 in purple and cluster 2 in green.



DISCUSSION

According to our observation from resulting map above, most of the bakeries are concentrated in the northeast and west areas of Toronto with the higher number in cluster 2 and moderate number in cluster 1, also, cluster 0 has very low number of bakeries in the neighborhoods. This represents a great opportunity for Celeste to open a new patisserie in the cluster 0 or cluster 1 areas with relatively small competition, and Celeste is advised to avoid neighborhoods in cluster 2 which has a higher number of bakeries.

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 machine learning by clustering the data into 3 clusters based on their similarity and lastly providing recommendations to the relevant stakeholders. By looking at the location data and Foursquare data we explored, we could find out that the neighborhoods where a higher number of bakeries already exsit, mainly Queen's Park and Ontario Provicial Governement.

We also took at a look at the neighborhood that has other dessert shops and coffee shops. There are higher number of dessert shops in Christie, St. James Town, and Cabbagetown, and a number of coffee shops in Church and Wellesley, Kensington Market, Chinatown, and Grange Park.

Considering that Celeste is trying to avoid opening the patisserie in a neighborhood that also has too many other bakeries, but needing to be where there are other business indicating a viable business location, it would seem that perhaps Christie, St. James Town, and Cabbagetown, could be good possibilies for a new patisserie since there are not many bakeries in these areas but the existence of dessert shops could mean a popular destination for this type of business.