Opening of new restaurant in the neighborhood.

Sneha Videkar

May 15, 2020

1 Introduction

Yelp [1][2] is a company which publishes crowd-sourced reviews about businesses using the company website.
Yelp dataset captures business data from all around the world. Mainly this site captures reviews related to business of various categories like Dentist, Restaurant, etc. Along with reviews, it lists location details of every business in Yelp database.

1.1 Business Problem

Success of new business depends on many factors. Location is the key factor which should be considered while opening new business. For this purpose, location attributes present in the Yelp dataset[4] will help find a place to start a new business.
This project aim to help investors/entrepreneur to find an appropriate location for their prospective restaurant business. With the help of Yelp Business dataset and machine learning techniques like clustering, this project will locate top 10 most common restaurant businesses in a neighborhood. This will guide investors while opening a new restaurant business at particular location.

1.2 Target Audience

This project is mainly helpful for the investors/entrepreneur who are looking for an appropriate location for their prospective restaurant business. Also, with the help of exploratory analysis from this project can help restaurant business owners to identify potential competitors.

2 Data Description

2.1 Data acquisition

For this problem below dataset will be used:

1. Business.csv will be used from Yelp Dataset. This dataset is taken from Kaggel website[3] and Yelp dataset website[4]. Figure 1 gives an idea about few records from the Business.csv dataset.

   Table.1 lists all features associated with every business entity.

business_id	name	neighborhood	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	categories
kCoE3jvEtg6UVz5SOD3GVw	"BDJ Realty"	Summerlin	"2620 Regatta Dr, Ste 102"	Las Vegas	NV	89128	36.207430	-115.268460	4.0	5	1	Real Estate Services;Real Estate;Home Services...
xcgFnd-MwkZeO5G2HQ0gAQ	"T & T Bakery and Cafe"	Markham Village	"35 Main Street N"	Markham	ON	L3P 1X3	43.875177	-79.260153	4.0	38	1	Bakeries;Bagels;Food
fNMVV_ZX7CJSDWQGdOM8Nw	"Showmars Government Center"	Uptown	"600 E 4th St"	Charlotte	NC	28202	35.221647	-80.839345	3.5	7	1	Restaurants;American (Traditional)
l09JfMeQ6ynYs5MCJtrcmQ	"Alize Catering"	Yonge and Eglinton	"2459 Yonge St"	Toronto	ON	M4P 2H6	43.711399	-79.399339	3.0	12	0	Italian;French;Restaurants
lHYiCS-y8AFjUitv6MGpxg	"Starbucks"	Liberty Village	"85 Hanna Avenue"	Toronto	ON	M6K 3S3	43.639863	-79.419533	4.0	21	1	Food;Coffee & Tea

Figure 1: Overview of Business.csv

Table 1: Business data column list

Data columns	Type
business_id	object
name	object
neighborhood	object
address	object
city	object
state	object
postal_code	object
latitude	float64
longitude	float64
stars	float64
review_count	int64
is_open	int64
categories	object

2. Forsquare API [5] will be used to get the most common venues for above mentioned business and in Toronto area. Only near by venues of type restaurants will be considered for further analysis to achieve the project goal.

2.2 Data Preprocessing

Yelp Business data was loaded in the IBM Watson Studio and accordingly csv file was read in dataframe called businessDF.

Yelp business data frame has captured various categories of the business data. In this project, business with "Restaurant" as a category are considered for modelling purpose.

There are in total 67741 rows in Yelp Business dataset after pre-processing.

3 Methodology

Exploratory analysis was performed to get better understanding of the data. Yelp dataset tries to captures details about different businesses. This is highlighted in the Figure 2. Here, Worcloud library was used as data visualization

technique.

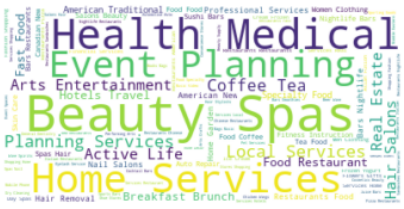

Figure 2: Different business categories

This project has focused on business with category as Restaurants. The data frame restaurantBusinessDF captures only restaurant related business details. Average ratings of all the restaurants were compared using bar plot. From the Figure 3, we can identify city namely Mont-Royal and Bellevue has achieved highest ratings for Restaurant business.

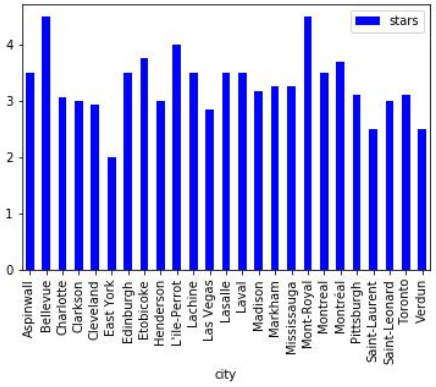

Figure 3: Average star rating for restaurants per city

With help of Forsquare API [5] and Forsquare developer account, neighborhood area of restaurant businesses derived from the Yelp Business data was explored. Area with radius of 500 meters was explored to get nearby venues. Different categories of the venues data was acquired with help of Forsqaure exploration technique. As main aim of this project to guide investors to find better location for their potential restaurant business, I have consider only nearby restaurant venues. Using various function of pandas dataframe, top 10 common restaurants in each neighborhood were identified.

Machine learning technique like clustering was used in this project. The top common restaurants from each area were grouped using clustering. K-means clustering was used to cluster the neighborhood in 3 clusters. This helped us to identify, which neighborhood has highest number of restaurants of similar kind. Based on these details, this project will try to find better location for the potential restaurant business.

4 Results

For better understanding, the result of clustering is shown in figure 4 for the Toronto area. From the map it is clear that, neighborhood is divided in 3 different cluster.
Cluster 0 : is represented with red colour and has moderate number of restaurants.
Cluster 1 : is represented with purple colour and has highest number of restaurants.
Cluster 2 : is represented with green colour and has few restaurants.

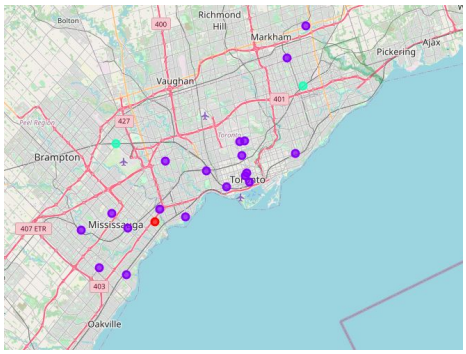

Figure 4: Cluster wise data of restaurant businesses in Toronto

5 Discussion

From figure 4 and figure 5, we can say most of the restaurants are in the neighborhood which belong to cluster 1. In total, 84 restaurants are in cluster 1. With this information, we can advise new business owners to avoid neighborhood area which belongs to cluster 1 while opening their prospective restaurant's business. On the other hand Cluster 0 and 2 has a handful of restaurants and it gives an excellent opportunity to start a new business.

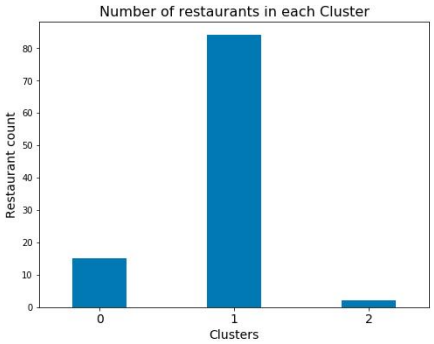

Figure 5: Restaurants assigned to each cluster

Further cluster analysis provides detailed information on type of top 10 restaurant in each cluster along with neighborhood area in which they live. With this information, the business owner is notified with potential competitors in the neighborhood. Cluster 1 has restaurants of type Italian, Indian, Korean etc. Upon comparison of cluster 0 with cluster 1, we can say that more fast-food restaurants are listed in a neighborhood which belongs to cluster 0. So if an investor is looking to open Fast-food restaurant then neighborhood which belongs to cluster 1 or 2 will be of good choice rather than cluster 0.

6 Conclusion and future work

With this project, we can guide investors to find an appropriate location for their prospective restaurant business. Machine learning and visualization techniques gives better information to the investors. Interactive GUI could be future work for this project.

References

[1] https://en.wikipedia.org/wiki/Yelp

[2] https://www.yelp.com/about

[3] https://www.kaggle.com

[4] https://www.yelp.com/dataset/documentation/main

[5] https://foursquare.com/