

IBM_capstone_Battle_Report

July 7, 2020

1 IBM Applied Data Science Capstone

1.1 The Battle of the Neighborhoods - Report

1.2 Introduction: Business Problem

1.3 Introduction

Opening a restaurant may be one of the best options to have a generous income. It is true that it may be a risky business adventure, but with the right planning and organization, a restaurant is one of the most successful businesses out there. Our lifestyle has changed and that leads society to need more and more food services. Today's jobs have us so busy that consumers have neither the time nor the inclination to prepare their own meals. They want to taste a delicious meal without worrying about cooking it, they want to eat in a healthy way without having to scrub the dishes afterwards.

Some areas have unusual restrictive ordinances that can affect your business, such as a time limit per day for freight trucks. It is also necessary to analyze the volume of people who pass through the restaurant's location in the opening hours and the percentage of them who might want to enter it. Finally, it is also extremely important to analyze the competition from nearby restaurants, especially if you're going to place it in a crowded area.

In this project, we will evaluate the feasibility of opening up a restaurant in Toronto. Toronto is the provincial capital of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. The purpose of the project is to define which areas would be more profitable to open a restaurant.

1.4 Data

One of the main aspects that we have to consider for opening up a restaurant is location. In fact, location is paramount when it comes to define the success (or failure) of a food business. Several factors should be taken into account on this regard, including the following questions:

1. How can localization contribute to your sales volume?
2. Is it easy to get to your business?
3. What restrictive ordinances are there in the area?
4. What is the traffic density?
5. Is there easy to park in the area?

Thus, two main characteristics that the target location should have are that it would need to be closer as much as possible to the center of the city and it would have to be located in an area with a low restaurant density. To solve this issue we will analyze the following data:

1. Neighborhoods of Toronto
2. Restaurants per neighborhood

To accomplish our goal, we will use the Foursquare API and Wikipedia's list of postal codes of Canada. Data analysis will be carried out using Python version 3.7 running on a Jupyter Lab environment.

1.5 Methodology

The main moto of this project is to find best location to open a new restaurant in Toronto, Canada based on competition in different locality and their population. So, to do this I have used 2 different data sets available as mentioned above. Those 2 data set contains Locality information of Toronto, different age group of people in the people, population.

To solve the problem I am going to use "K-Means Clustering Algorithm". K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are the centroids of the K clusters, which can be used to label new data. Labels for the training data (each data point is assigned to a single cluster). Also, I will be utilizing different maps in-order to give a clear vision to the target audience.

Steps we took for the analysis:

1. Collected required data: location and type (category) of every restaurant within our lat and lng. We have also the type of resutaurants in particular locality.
2. Explored the restaurant density across different areas of Toronto - we will use K- mean to identify a few promising areas close to center with low number of restaurants and their type.
3. Explored the most promising areas and within those create clusters of locations that meet some basic requirements established in discussion with stakeholders: we will take into consideration locations with less restaurants in radius of 500 meters, We will present map of all such locations but also create clusters (using k-means clustering) of those locations to xplore neighborhood.

1.6 Analysis

1. **Data identification, capturing and cleaning:** search & identify the relavant data source and capture it, here we are using wikipedia to get data about Toronto, Canada. Then we remove all the redundant value (data cleaning). Then we combine neighborhood similar Bronx. Now the data is clean and ready to use.
2. **Combining different data source and sorting neighborhood based on Longitude and latitude:** we combined the neighborhood dataset with postal address and dataset with Latitude & Longitude and save them it separate data frame.The resultant data frame with contain details aabout Postal code, Brough, Neighborhood, Latitude & Longitude.Then visualize it using folium map.

3. **Explore the Toronto's neighborhoods:** Firstly, we explored all the neighborhoods in the city of Toronto, using the Latitude & Longitude data, using Foursquare API to get the Restaurant venues available in Toronto. Explore the unique categories in the neighborhood. Filter the Venues details for all possible 'Restaurants'. Find each neighborhood along with the top most common venues. Identify the top 10 venues for each neighborhood.
4. **Clustering:** with an assumption of 5 clusters, use K-Cluster algorithm to come up with 5 different clusters in Toronto with similar set of Venues. Explore each cluster and determine the discriminating venue categories that distinguish each cluster. Identify the clusters & Boroughs/Neighborhoods with maximum number restaurants and their types.

1.7 Results and discussion

Our analysis shows that although there is a great number of restaurants in Toronto, there are pockets of low restaurant density fairly close to city center. We have 4 boroughs and 74 neighborhoods inside geographical coordinate of 43.653963, -79.387207. Based on our initial assumption of the cluster with maximum number of restaurants will have the best possibility to have a new restaurant due to the need in the area. Based on the resultant clusters it looks like Cluster 1 and Cluster 5 have higher number of restaurants than rest of the clusters.

It is entirely possible that there is a very good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

1.8 Conclusion

Purpose of this project was to identify areas in Toronto with low number of restaurants in order to aid stakeholders in narrowing down the search for optimal location for a new restaurant. By calculating restaurant density distribution from Foursquare data we have first identified general boroughs that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.