Segmenting and Clustering the City of Surrey based on the Restaurants and Popular Venues

Preethi Sriram

January 15, 2020

1. Introduction

1.1 Background

The City of Surrey is located in British Columbia, Canada and is a popular locality for south Asian immigrants. The Community is flourishing with new businesses opening up all the time. There are so many popular South Asian restaurants in this community. On careful observation, it is obvious that it is not just about the cuisine but the location that makes them preferable. It is to be noted that some restaurants had to be shut down because of the lack of business despite having the same cuisine and ambience that other popular restaurants offer. What makes some businesses successful and others not so much? The location and the venues around may have an impact on the customer base. If one were to go for a nice dinner on a cold night, they would prefer to go somewhere that has easy parking and some popular stores around so they can get some chores done along with their dinner. If the restaurant is around a noisy and not so clean fish market, one may avoid even though they have the best dim sums in town.

1.2 Problem Statement

A Businessman is looking to open a fine dining south asian restaurant in the flourishing community of Surrey, British Columbia. He is aware of the strong presence of South Asian communities in this area but is that alone enough to convince him to open up an Asian Restaurant? He needs to know what kind of restaurants are already popular in this community. This information along with the most popular venues in this community will help him reach a sound business decision. To solve this Business Problem, we need a dataset of the restaurants that are present in this community. We need to use Foursquare api to find the popular venues around the restaurants.

1.3 Interest

Some of the questions that this project can help solve are :

1. What kind of restaurants are operating in this community?

2. What are the popular venues around these restaurants where people prefer to go?

3. What is the potential business around these popular venues?

4. Clustering the community based on the restaurants and popular venues so the stakeholder gets an idea of what is the demographic of each venue

2. Data Acquisition and Cleaning

2.1 Data Sources

The data we need is a dataset of restaurants operating in this community of Surrey, British Columbia. I found the open data of City of Surrey quite useful because they had the list of restaurants licensed by Fraser Health. The dataset contains the name of restaurants, their addresses and their coordinates. We also need Foursquare developer access to make requests to view venue data.

1. City of Surrey open data

https://data.surrey.ca/dataset/restaurants

2. Foursquare APi to make requests

https://foursquare.com

Based on the results from foursquare venue data, we can cluster the community based on the restaurants and popular venues. This clustering enables us to view the community in terms of a stake holder's point of view.

2.2 Data Cleaning

The Dataset had seven columns and 1346 rows. The attributes like the 'tracking number' are irrelevant to our project and hence was dropped. The attribute city had the same value 'Surrey' for all entries and the attribute 'Type' had the same value 'Restaurant' for all the entries and hence

these are irrelevant to our project as well. Hence our final dataframe has the following attributes: ‘Restaurant Name’, ‘Address’, ‘Latitude’ and ‘Longitude’. No other changes were made to the dataset.

3. Methodology

With the cleaned dataset, we now start exploring. We start by exploring the first row of our dataframe or the first restaurant in the dataset. We need to know what other venues are popular around this particular restaurant. We need Foursquare developer access to have credentials like the client id and client secret. They are a part of the request URL along with query type, version and the coordinates. The URL is set and then passed as a parameter of the get request to the foursquare. The request returns three venues located around this restaurant one of which is this restaurant itself which makes sense considering the fact that this is a popular sushi place. We proceed to retrieve the nearby venues for the rest of the restaurants in the dataset. Now we get a dataframe of all the nearby venues and we use one hot encoding to get the dummies for all nearby venues. They are then grouped by the restaurant name to avoid the redundancy and we sort the dataframe according to the most common venues for each restaurant.

The machine learning algorithm that is used to segment and cluster the resulting dataframe is ‘K means Clustering’. We set the clusters to 5 and fit the model on our dataframe.

4. Results

The resulting clusters have five clusters each clustering the community based on the nearby venues. The first cluster is all the restaurants around yoga clubs and sandwich places. The restaurants around fish markets are grouped in another cluster. This is useful because no one wants to dine in a fine restaurant around fish markets. Looking at the clusters, it is evident that it’s a good idea to open the restaurant around any restaurant in the first cluster as they are in a busy and happening community around yoga centers and coffee shops.

5. Discussion

The clustering algorithm that we use here is ‘K means’ and it does a fine job in clustering the community based off the restaurants and popular venues. If we were to cluster only based on the restaurants present in the community, it would have yielded a result that has not much useful

insights. The use of foursquare api to retrieve the venues nearby has boosted our search for the perfect spot. This gives us an insight into the customer base that's likely to visit each community. A potential restaurateur would not just base his restaurant's location simply out of an algorithm and it makes sense that he would visit the potential locations physically before deciding on it. The algorithm gives a much more useful insight into how the communities can be grouped and can provide the businessman with much more valuable choices.

6. Conclusion

The results of the clustering algorithm proves that the first cluster which is grouped around yoga clubs and cafes would prove to be a better choice for the businessman. I conclude this project with the observation that the restaurant would be successful if opened around the pizza places in this community of Nordel way, fraser way and 100 avenue.