

Coursera Capstone

IBM Applied Data Science Capstone

Opening a new Restaurant in Calgary, Alberta, Canada

By Jimmy Zhu

August 2020

Problem Statement

Keeping a restaurant running requires a continuous stream of eaters placing orders for the food that the restaurant offers. The quality of the food, marketing plan, and competitive pricing are all factors that will drive that traffic but location is also a key decision for the restaurant owner to make which can allow the restaurant to flourish. The more restaurants in a given neighborhood, the greater the odds that restaurant traffic may diminish. Hence, the business question is what neighborhoods are underserved by restaurants and what are some of the restaurants that already have a presence in that area. This question can apply for brand new restaurant owners who are looking to enter a market or for restaurant owners to expand into regions with similar competitive landscape.

Data

Location data will be webscraped from Wikipedia identifying neighborhoods belonging to the same postal code within Calgary, Alberta, Canada. Each neighborhood will be passed into the Foursquare API to retrieve data of the venues within that neighborhood. There will be a maximum of 100 venues within 500m of the latitude and longitude of each respective neighborhood. The data retrieved from Foursquare includes the venue name, venue latitude, venue longitude, and venue category. The venue category describes the category that a venue belongs to, such as Sports Bar for 'Shark Club Sports Bar & Grill'. Venues returned belong to a large variety of venues from banks to hotels to restaurants so an additional column titled 'General Venue Categories' will be added. This new column will generalize the venue category into their appropriate business sectors such as Banks, Food, Entertainment, Health and Fitness, Hospitality, Retail, and Services. Now each venue has a General Venue Category assigned to it so we can compare businesses which are more related to one another. In this project, the venues belonging to 'FOOD' will be isolated and grouped together by the neighborhood. Within a pandas dataframe, each unique postal code and the corresponding neighborhood(s) belonging to that postal code will have a single row with the quantity of restaurants within those neighborhood(s) and all the names of each of those restaurants. K-means clustering will be applied to the dataset based on the quantity of restaurants in each neighborhood(s) so that five clusters are defined to distinguish areas that have similar quantities of restaurants. An interactive map showing Calgary will include markers color coded by the cluster they belong to with a popup which identifies the name of the neighborhood(s) belonging to a unique postal code, the cluster label, and the restaurants that are in that area. Using this, restaurant owners can make decisions of where to start up their restaurant.