IBM Applied Data Science Capstone

Week #4 – Sections 1 and 2

“Opening a New Restaurant in the Toronto Area”

1. Introduction / Business Problem

1.1 Pre-Introduction

This research paper is my capstone project (course #9) for the IBM Data Science Professional Certificate. Part of the requirements are to publish a blog summarizing the results, hence this article. This is my fourth or fifth data science project, but first one published to the public – got give kudos to this IBM certification and the way it was organized such that writing an article is part of the final.

1.2 Introduction

Poke restaurants are an innovative and healthy Hawaiian cuisine that are seeing growing demand. [1] Poke “build-your-own” bowl flexibility allows diners to choose from multiple bases[2] (salad, rice or quinoa), multiple protein sources[2] (salmon, tuna, octopus, red snapper, etc.), and multiple toppings[2] (10+ sauces and oils, seeds, onions, cucumbers, crab-salad, seaweed, dices mangoes or oranges, and much more). Poke bowls generally offer superior nutrition and taste relative to other fast food options[3].

1.3 Business Problem

The objective of this research project is to analyze and select the best location in Seattle for our “hypothetical” client to open a new Poke Restaurant. S/he already owns and operates one Poke restaurant just North of Seattle and is looking to expand by opening a second somewhere in Seattle.

Searching for the optimal location is challenging because it is not as simple as finding a geographic gap where there is not yet a Poke restaurant. It is more complex in that potential customers aren’t interested in driving out to an isolated neighborhood with one Poke restaurant; but rather they prefer frequenting common clusters of restaurants(see “clustering” game theory[4]…not to be confused with “clustering” algorithms[5]) However, caution must be taken to carefully select the best location, because there is risk that the Poke market is already oversaturated[6] with 68 Poke restaurants currently in Seattle as per a Google Map search (via jump to last-in-list).[7]

To address this business problem, research will be done using data science methodology and machine learning techniques such as clustering to find and rank the best locations.

2. Data

2.1 Data Needs

To solve the business problem above, we will need the following data:

1. List of neighborhoods in the Greater Seattle area from which to select the best locations.
2. Latitude and longitude coordinate of those neighborhoods. This is necessary to plot the map and tie in venue data.
3. Venue data, particularly restaurant data that is both with, and without Poke restaurants to perform clustering and find desirable gaps in service while simultaneously grouping with similar restaurants.
4. Ultimately, we want the following factors to be built into our decision model:
   - Size in Square Miles (used to calculate density)
   - Density of **non**-Poke restaurants in the neighborhood (higher is better)
   - Density of Poke restaurants in the neighborhood (lower is better)
   - Population density of neighborhood (higher is better)
   - Renter vs. Owner Occupied in the neighborhood (renters are better)
   - Urban Village Type ("Urban Center" beats "Hub Urban" beats "Residential Urban" beats "Manufacturing Urban")
   - Distance of neighborhood to city center (shorter is better)
   - Median income for Neighborhood residents (higher is better, Poke is higher priced)
   - Median age (younger is better)
   - Foursquare will add details like sales, tips, restaurants by type, "like" ratings, and much more

2.2 Data Sources

The City of Seattle OpenData[8] site has a list of 42 Seattle neighborhoods, including size and demographics such as population, ethnicity, percent home ownership and more. The data is readily downloadable as a CSV file to form our master list of neighborhoods from which to choose. The Python Geocoder package will be used to establish the latitude and longitude coordinates of each neighborhood.

Another data source is the Foursquare API[9] to fetch venue data for the neighborhoods. Foursquare is one of the largest venue databases having over 105 million[10] global points of interest and 125,000[11] developers building location-aware experiences with the Foursquare API. For this research, we will focus on the venue category of Restaurant, and in particular any having "Poke" in the name in order to help us solve the business problem. Once the data has been sourced, it must then undergo data cleaning, data wrangling, and machine learning (K-means clustering) as finally visualization using

Python’s Folium library to map results.