IBM Applied Data Science Through Coursera - Capstone

The Battle of Neighborhoods - Week 4 Report

Find the Best Neighborhood in Chicago to Open a Boutique Donut Shop

Loren Krokowski - June 2020

Introduction / Business Problem

Being able to foresee this growing trend, a client capitalized on the industry growth by opening his first specialty donut shop in Atlanta, Georgia approximately 10 years ago. This client continued to add additional locations in major metropolitan cities across the United States at a rate of 3 locations per year. Chicago is the next remaining large city that this client is looking to expand operations. It costs over $1,000,000 in starting capital to secure the ideal location, purchase and set up the equipment, hire and train employees, and begin a substantial marketing campaign. Picking the ideal neighborhood is the most critical step in this process because it is the costliest to change at a later date.

Even though the client already has existing locations nationwide, he partners with a local investor who will manage day to day operations and together they focus on creating a local presence where the donut shop has more of a family-owned mom-and-pop feel and not that of a national chain. This focus is strengthened by the locations being in urban areas with more densely populated neighborhoods, higher foot traffic during all times of the day and a mix of business and residential as opposed to suburban outskirts where the national chains reside and foot traffic is nonexistent.

The client has already identified the following parameters that have proven to be successful in locating the most ideal neighborhoods to open their existing locations:

- A large volume of nearby coffee shops - The client's previous research and experience has shown that his target audience is generally the same demographic as those that frequent coffee shops. A larger than average coffee shop presence indicates there will be a larger than average target audience for donuts.
- A competing donut shop - Contrary to most people's belief to open a new retail location far away from existing competition, the client believes it is more beneficial for all donut shops when a handful exist in the same neighborhood. This

helps to reinforce to the community that donuts themselves are popular and desired and revenue for all donut shops can increase.

- Higher than average population density – Communities with higher population densities that also contain foot traffic at varying times of the day generate a consistent ongoing target audience.

The goal of this project is to analyze communities in Chicago, develop an understanding of prominent retail and demographics for each community, and ultimately identify a short list of possibilities for the client to choose from. At that point it will be up to the client to visit each neighborhood, research leasing options, and make the final decision. Our work to narrow his options to just a handful of communities will save him weeks of time by not having to research and drive all over town himself, not to mention the fact that our analysis will be based on hard data to back our selections.

Gather Data

In order to perform this analysis, the project will take place in a Jupyter Notebook running Python programming language and will rely heavily on pandas dataframes. Below is a highlight of the core processes involved:

a. The Wikipedia page https://en.wikipedia.org/wiki/Community_areas_in_Chicago provides a table of 77 communities in the city of Chicago, including population density. This data will be gathered into a pandas dataframe using the read_html function. The dataframe will be cleaned up to remove unnecessary columns and make the column headers more relevant.
b. The dataframe will be updated to include latitude and longitude coordinates for each community which will then be used in the next step. This will involve using Geopy geocoders through Nominatim to update our dataframe for each community. The updated dataframe will not need to be cleaned more to only add columns for latitude & longitude, removing any additional data brought in by the Geopy process.

c. The Foursquare API will be used to gather venue data for each of the communities based on their coordinates as derived above. Venue data will be limited to a yet to be defined radius to secure the best volume of data.
d. The venue data retrieved from Foursquare will be aggregated to produce a table showing the most common venues for each community.
e. A clustering analysis will be run using k-means to determine the discriminating venue categories that differentiate each cluster.
f. A new dataframe will be created highlighting only those communities where coffee shop is in one of the topmost common venues. Our client is specifically using a high frequency of nearby coffee shops as the single most important factor in determine community location.
g. An analysis of the population density based on the reduced list of high frequency coffee shops will be performed to further differentiate and reduce the list

The end result of this analysis will be to determine the most qualified community in Chicago for our client to open a new donut shop. Of course there will be additional variables that the client will need to take into consideration, but the core data of nearby venues that fit the retail demographic coupled with above average population density will give the client tools to begin his search that would have otherwise taken weeks or legwork to produce.