Segmenting the Community Areas of Chicago for offsite ATM installations

Shagul Hameed Meeran

05 September 2019

1. Introduction

1.1 Background

Banks predominantly start with locating ATM in their branches. That establishes the primary ATM (Core) network. Off-site ATMs are installed at places where foot-traffic is high.

Typically, a Retail Bank's marketing department would research areas of interest where the bank should have its off-site ATM installed. Such research is grouped by various parameters, of which the following are key parameters:

- Estimated traffic per day
- Location popularity
- Possible venues that generate cash transactions (e.g. Restaurants, Groceries)
- Service Routes (for Security company, is the ATM within a route or away from it)
- Other ATMs in the vicinity
- Leasing space issues / costs
- Marketing potential
- Communication & Power infrastructure
- Threat scale for vandalism and/or crime
- Competition Factor (if there are multiple ATMs in the same location)
- Vestibule Preferences

All these parameters are taken into account and prioritized and then the banks then take the next step in trying to find out the operating costs of such an ATM

1.2 Problem

Though there are some mechanisms to find out most of the above said parameters, it requires field work predominantly for the factors such as location popularity, other ATMs and available venues mix that promote cash transactions. This field work is meant to be time consuming and costly and also different banks would conduct the same exercise for the same locations.
If this field work is accompanied by some data from machine learning that provides the nature of the neighborhood, its popularity and the available venues mix, then

the field work can be reduced a lot and focused only on the key possible areas and neighborhoods.
This machine learning should be able to segment the neighborhoods into different clusters based on the types of venues and popularity and other interested factors.

1.3 Interest

All commercial banks that needs to plan to open its offsite ATM locations or extending its existing ones would be able to use this segmentation to take some informed decisions. Also it gives back the information about the possible competition from other banks in every locations based on their existing ATM networks.

2. Data acquisition and cleaning

2.1 Data Sources

The primary list of community areas of Chicago can be web scrapped from the Wikipedia
https://en.wikipedia.org/wiki/Community_areas_in_Chicago

There are the following additional information that would be obtained from the city of Chicago Portal

1. Grocery Stores
2. Sidewalk Café Permits

Along with the latitude and longitude of each community areas of Chicago would be provided with an external CSV file.

All the venues details for each community areas would be extracted from Foursquare API.

2.2 Data Cleaning

The community areas has to be web scrapped and stored in a CSV file. These data will have community area numbers. Using the community area numbers, the additional data extracted from Chicago Data portal (grocery stores, sidewalk cafes) will be combined together to form a single data frame.

To this data frame, which contains the above set of information, the geocoding CSV file will be read and merged together. The final resultant data set will be

Community Area No	Community Area	Grocery Stores Count	Sidewalk Café Count	Latitude	Longitude

With this data set, the set of venues which are located in each of these community areas would be extracted using the Foursquare API passing the latitude and longitude of each community area.

Also, not all venues are considered. The following filters will be applied to the venue list. The idea to prioritize the venues that promotes cash transactions.

1. Specific venues on arts & entertainment (Arcade, Art Gallery, Bowling Alley, Casino, Concert Hall, Historic Site, Memorial Site, Movie Theater, Museum – all types, Stadiums – all types, Theme Park, Zoo)
2. All type of restaurants, bakery, bistro, cafeteria, cafes, coffee shop, diner, etc.
3. All type of bars and nightclubs
4. All type of medical institutions
5. All grocery and small stores

All these venues count will be merged for each community areas and finally the data will be grouped together into the following columns.

1. Entertainment
2. Tourism
3. Food
4. Nightlife
5. Health
6. Retail

Community Area	Latitude	Longitude	Entertainment	Tourism	Food	Night Life	Health	Retail

3. Exploratory Data Analysis

<To be added>

4. Segmentation with Clustering

<To be added>

5. Conclusions

<To be added>

6. Future directions

<To be added>