APPLIED DATA SCIENCE CAPSTONE PROJECT

The Battle of Neighborhoods

Wendy Kambestad

09/09/2019

Problem Description

My problem centers on opening a small music entertainment club within some neighborhood in San Francisco. The club would not only offer evening entertainment but would ideally be used to offer some music classes and instruction. I am seeking to establish the business in a thriving neighborhood with residents who would attend events and where such venue types are not already abundant.

Background

I assume that information a new business owner wants in order to wisely decide location includes basic city affordability and resident demographics.

- Ability to afford entertainment. A middle-class population has some cash for discretionary spending.
- Reasonably priced homes. A business owner may ideally want to live within the community where his/her business is located.
- Population volume. There must be a good size adult population within the community to patronize the club.
- Crime rate. Obviously, opening a club in a low crime area would be ideal.

Another major decision factor is whether music club venues already exist in these communities. Establishing a new club where there is less competition makes sense.

Data Requirements and Collection

My plan to include basic demographics was hindered by lack of freely available data. San Francisco data for home prices, population, and income was not readily available in granular enough detail for neighborhood or district analysis. While I found this information within Federal Census available to the public, it was at the city level only. Below data sources were used.

1. San Francisco Police District Geodata

San Francisco geo coordinates data, which was used in an earlier machine learning exercise, was imported to provide boundaries needed to visualize crime incidents. (See below.)

2. Crime Data by Police District

Crime data for 2018-present found at https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783 was used.

3. San Francisco Neighborhoods

Because I could not find any basic San Francisco neighborhood dataset, I compiled my own Excel file using information from a Wikipedia page, https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco.
I added geographical coordinates with the help of Wikipedia and GoogleMaps and using a visually gauged center of the neighborhood unless Wikipedia provided coordinates. Some neighborhoods are larger than others and I realize this method of creating a location dataset is imperfect. Finally, I associated each neighborhood with a police district.

4. Venue Data

Search - > Music Specific Venues

Foursquare venue data was collected for all neighborhoods using the Search Venue API with a list of category ids. Categories searched included music and performing arts venues.

Concert Hall: 5032792091d4c4b30a586d5c
Music Venue: 4bf58dd8d48988d1e5931735
Jazz Club: 4bf58dd8d48988d1e7931735
Piano Bar: 4bf58dd8d48988d1e8931735
Rock Club: 4bf58dd8d48988d1e9931735

Venue categories exist in a hierarchy structure. The Concert Hall category is at a higher level than the other categories, and seemed to contain some categories that were not music related. After getting actual venue results I examined and removed venues that were outside of the search needs.

Explore -> Venues

Venue data was collected using the Explore Venue API for neighborhoods that did not have any music related venues.

Methodology

The basic methodology used was to first run the 115 San Francisco neighborhoods through a routine that captured Foursquare music related venues within 500 meters. In order to determine those neighborhoods currently lacking music related venues, I merged the musical neighborhoods back with the full list of 115. 84 neighborhoods had nearby access to music venues, leaving 31 that did not. I then ran the 31 non-music venue neighborhoods again through the Foursquare API, this time using the explore option to find whatever venues could be found.

How do these non-music venue neighborhoods compare to each other? Would certain types of neighborhoods be better candidates than others for a new music venue? I used K-Means clustering to be able to compare the neighborhoods. Each neighborhood was analyzed for top frequency of the resulting venue categories. Using K-Means clustering of those neighborhoods, I could attempt to segment and compare future clients.

Results

The concentration of music venues appears to be highest in districts with highest incidents of crime. Of all 115 San Francisco neighborhoods analyzed, 84 had nearby music venues and 31 had none. Through neighborhood segmentation and clustering, 3 different profiles were suggested. One neighborhood had no venues at all, which requires more research to understand the underlying reason.

- Cluster 0 (5 neighborhoods) has a Playground as a top venue. Top venues also include schools, transportation stops, stores, and dog runs. “Family”
- Cluster 1 (20 neighborhoods) has a wide variety of stores, shops, ethnic restaurants, entertainment, and activity venues. “Eclectic”
- Cluster 2 (5 neighborhoods) is dominated by outdoor venues including trails, parks, gun range, and to a much lesser extent, food-related venues. “Open Spaces”.

From the surface, Cluster 1 “Eclectic” appears to be a good profile for a new music business, and there is a good variety of business already. Cluster 0 “Family” could also be an appropriate profile, especially considering my goal of not only providing music entertainment but education as well. Cluster 2 “Open Spaces” does not have a profile that appears very business friendly and I would drop this neighborhood cluster from consideration.

Cluster 0 and 1 have plenty of neighborhoods outside of the higher crime districts.

Discussion

Results would be more viable with resident demographic and detailed city information. Additionally, I’m certain the method of creating neighborhood and coordinate data could be improved and would lead to a more precise outcome.

Conclusion

This was a useful final project, not so much for the results of a hypothetical problem but simply for the practice of various skills covered in the IBM Data Science course. I undertook this 9-course track because I wanted to have a feel for what data science involves, and the only way I can really do that is by getting hands-on with some of the tools and techniques. Mission accomplished! This has been a great course and totally worth my time, effort, and money.