Choose Where to Open a New Fitness Center in Chicago

Part I: Introduction

According to National Center for Health Statistics (note), 39.7% of adults aged 20 and over in the US have obesity. As the obesity issue becomes worse over the last decade among Americans, there is a strong demand for gym and fitness facilities, especially in the cities with large number of population.

Chicago is the third largest city in the US and well-known for its beautiful downtown skyline as well as it long windy winter. Its cold weather starts in December and stretches all the way to March or even April. In cold and windy days, people are more likely to stay indoor when doing exercise, and gym and fitness facility is a popular business in Chicago. This report will try to analyze and compare different datasets and help business owner find best possible location for opening a new fitness center in the city of Chicago.

Part II: About the Data

1. Data Source

This report will mainly use three datasets for analysis and discussion, including venue data in Chicago area from Foursquare, the demographics and community data from the Illinois State's data portal website, as well as the geospatial data of Chicago communities from Wikipedia and City of Chicago.

The venue data obtained via Foursquare API will be used to show the current available gym and fitness facilities with their coordinates, i.e. the 'market supply' in Chicago. The geospatial data from Wikipedia mainly provides the coordinates of the community centroids. These coordinates provide the key information when using the explore function to request venue data from Foursquare API based on a location point. The geographic boundaries data from City of Chicago is used to decide if venues are inside one community or not, as well as for the visualization.

The demographic data from 2013-2017 American Community Survey that contains the population, age, employment and other information by communities in Chicago. The venue and demographic data together will be used to analyze which area of Chicago may need more potential gym and fitness facility services, i.e. the market demand.

2. Data Preparation

   a. Venue data

For the venue data, Foursquare provides all types of venue information. Because we want to mostly find out the market demand and the market need for gym and fitness service market. We will not focus on venues other than gym and fitness related services.

When requesting data from Foursquare, the searching radius of each community's centroid is set to 2500 meters since this distance can reach to all the communities' boundaries. Since the radius causes communities next to each other having overlapped venues, the duplicates are dropped from the search results. The original result has 2237 venues. With duplicates removed, there are 1086 venues left.

Foursquare sets up a specific catalog 'Gym / Fitness Center' category for them with 12 the sub categories. They are boxing gym, climbing gym, cycle studio, gym pool, gymnastics gym, gym, martial arts dojo, outdoor gym, pilates studio, track, weight loss center and yoga studio. Because this report's main purpose to provide suggestion for opening fitness center with large space and more functions, the specialized gym, studio and outdoor venue are not direct competitors for fitness centers. This report will divide the venue data into two groups for further analysis. The venues will be treated as fitness center if they fall in 'gym' category or if they are not categorized as specialized venue above. After renamed and sorted, the venue data has each venue's name, venue category, coordinates and the community it belongs to.

b. Geospatial data

To prepare the geospatial data for Chicago's communities, Beautiful Soup function is used to retrieve the data from the Wikipedia web page ('en.wikipedia.org/wiki/Community_areas_in_Chicago#List_of_community_areascontains'). The page has all the 77 communities' names and overview, and each community also has a link that redirects to its own page where contains the coordinates of the community centroid. After cleaned and sorted, this dataset has community name, centroid coordinates, as well as area and total population.

The geographic boundary data of all the 77 communities is downloaded as GEOJSON file from City of Chicago's website and does not need to be cleaned. It will be used to filter the venues not belonging to each community when calculating the total venue numbers.

c. Demographic data

There are 83 main features with 233 labels in in the original dataset. In order to make the analysis efficient and consistent, the following steps are performed on the dataset.

Step 1. Main features that are unrelated to people in community are dropped.

Step 2. After the first step, main features that are dependent to each other are grouped, making the remaining data into 10 main features.

1) Population (Grouped), including Age Cohorts, Household Size;
2) Household Type;
3) Economic Status (Grouped), including Employment Status, Household Income, Monthly Housing & Transportation Cost as a Percentage of Household Income;
4) Commute (Grouped), including Mode of Travel to Work, Vehicles Available;
5) Educational Attainment;
6) Housing & Tenure;
7) Housing (Grouped), including Housing Type, Housing Size, Housing Age;

8) Race and Ethnicity;
9) Nativity;
10) Language Spoken at Home.

Step 3. Remove main features that are not suitable for analysis, including Race and Ethnicity, Nativity, Language Spoken at Home. After this step, there are 7 main features left.

Step 4. Select the sub features that represents the grouped feature.

1. Population (Grouped): keeps only Age Cohorts;
2. Household type;
3. Economic status (Grouped), keeps only Household Income;
4. Commute (Grouped), keeps only Mode of Travel to Work;
5. Educational Attainment;
6. Housing & Tenure;
7. Housing (Grouped), keeps only Housing Type;

Step 5. Group and select one label that can represent for each feature.

1. Population (Age Cohorts): This original feature has 6 labels, including Under 19, 20-34, 35-49, 50-64, 65-74 and Over 85. Because the active groups will be kept, including 20-34, 35-49 and 50-64, which will be combined as 'Active Group'. The 'Active Group' will be used to represent Population feature.
2. Household type: This original feature has Family, Single parent with Child, Non-Family. The Family and Single parent with Child will be grouped into 'Family'. The percentage of 'Family' in total of Family and Non-Family will be used to represent this feature.
3. Economic status (Household Income): The median Household Income for each community will be used for the further analysis.
4. Commute (Mode of Travel to Work): this original feature has 7 labels, including Work at Home, Total Commuters, Drove Alone, Carpool, Transit, Walk or Bike, Other. Since the major difference is about close to work or not. 'Work at Home' and 'Walk or Bike' will be grouped together as 'Close to Work'. The percentage of 'Close to Work' in total group number will be used to represent this feature.
5. Educational Attainment: this original feature has 6 labels, including Less than High School, High School Diploma or Higher, Some College, Associate's Degree, Bachelor's Degree or Higher and Graduate or Professional Degree. The higher education will be grouped together as 'Higher Education', including Some College, Associate's Degree, Bachelor's Degree or Higher and Graduate or Professional Degree. The percentage of 'Higher Education' in the total of higher Education and non-higher education will be used to represent this feature.
6. Housing & Tenure: this feature has 4 labels, including Occupied Housing Units, Owner-Occupied, Renter-Occupied, Vacant Housing Units. The percentage of Owner-Occupied in total of Occupied Housing Units will be used to represent this feature.
7. Housing (Housing Type): Housing type has 5 labels, including Single Family, Detached, Single Family, Attached, 2 Units, 3 or 4 Units, 5 or more Units. Because the most distinguished housing type from others is 5 or more Units, this type will be calculated as the percentage of total number to represent this feature.

After the above 5 steps, the demographic data is ready to be analyzed.

Now all the datasets are cleaned with unnecessary columns and information removed from the data. In order to merge the data or comparison, community names are all capitalized to keep the same format between different datasets.