Clustering neighborhoods in New York City as recipient groups of healthcare service

Sol Kim

June 15, 2021

1. Introduction

1.1 Background

The government not only focuses on improving medical services to citizens but also promotes citizen's health by providing various services such as expanding public areas for outdoor activities or providing complementary medical checkups for disease prevention. The accessibility of healthcare and medical services is crucial to maintain a physically active lifestyle and promptly receive medical treatment when needed. A living environment has a significant influence on individual health status.

Since New York city has different demographic characteristics among boroughs and its neighborhoods such as GDP and density rate, several demographic factors can be considered as highly related to the development of infrastructure and healthcare and fitness-related business growth. Therefore, it is advantageous for the government to accurately understand the physical environment of its neighborhoods including accessibility of health-related facilities and financial conditions so that they can have a strategic guideline of healthcare policies.

1.2 Problem

This project assesses the living environment of neighborhoods based on the population density, income, and accessibility of medical and fitness-related facilities. The project aims to cluster neighborhoods into several recipient groups of healthcare service based on the data analysis, particularly the neighborhoods of Manhattan, Brooklyn, and Staten Island.

1.3 Interest

This project will give useful information not only to government health agencies but also private healthcare service companies because it helps them understand the needs of potential customers depending on where they live and suggest better services or products.

2. Data acquisition and cleaning

2.1. Data sources

The location data of neighborhoods of New York City is found in the dataset provided by the IBM course and location data and category list of medical centers and fitness-related facilities can be obtained using Foursquare API. The demographics such as GDP, median income, poverty rate, and density rate can be

found in Wikipedia which specifies different sources of the data. However, the demographic data is based on the unit of boroughs, not neighborhoods, so those values have to be assigned the same to all neighborhoods within the corresponding borough, which might not be an accurate determinant but can be enough to show a general picture of the financial condition of neighborhoods.

- The location data of neighborhoods : https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json
- Foursquare API category ID : https://developer.foursquare.com/docs/build-with-foursquare/categories/
- Demographics : https://en.wikipedia.org/wiki/Demographics_of_New_York_City

2.2 Data cleaning

Data that might contribute to determining segmentation were combined into one table.

First, the neighborhood coverage had to be narrowed down after experiencing difficulties of repeatedly getting search results from Foursquare APL for every neighborhood for the reason of limit API calls of Foursquare for the basic service user. Only three boroughs out of five were used for this project data and they were selected by its discrete value of GDP-the Staten Island has low GDP, the Brooklyn, middle and the Manhattan, high) because later it can also show if there is any relation between the accessibility of medical and sports facilities and their GDP. However, it should be repeated with other boroughs to confirm it. The total number of neighborhoods of three boroughs was 173.

Second, using Foursquare category id, nearby venue data of medical centers and outdoor and sports facilities were gathered. Since the result from the API has a lot of venues not in the category, I filtered the result again so that it can only include venues specified in the category id. As there were not a lot of missing values in medical and sports venues, those rows were dropped from the table. Then the number of medical venues and sports venues was included in the features.

Next, GDP, median income, density rate, and the poverty rate of the borough were included in the table. As I mentioned, these data were only available for the unit of the borough. Consequently, it is unavoidable to give inputs that lead to categorizing neighborhoods into three boroughs. However, the result can clearly show if there are other clusters within the boroughs, which are determined by other variables.

2.3 Feature selection

After data cleaning, there were 173 samples (neighborhoods) and 6 features in the data.