Report for Capstone Project, “The Battle of Neighborhoods”

Applied Data Science Capstone

1/2/2021

2. Data

This project draws on three sources of data. The first source gives the locations of all the neighborhoods in New York City. The second identifies, for each neighborhood, what venues belong to that neighborhood. The third contains the ratings for each venue.

The first data set, containing the neighborhood location data, has been supplied by the Coursera instructor for the purposes of completing a lab assignment. They originally came from a publicly available source such as Wikipedia. The data file contains the neighborhood name (unique identifier), borough, latitude, and longitude for each of the 306 neighborhoods in New York City. This data file is located at: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json")

The second data set, which identifies what venues belong to each neighborhood, is downloaded from Foursquare, which is a local search-and-discovery app developed by Foursquare Labs. Foursquare Labs is a privately-held technology company with approximately 400 employees. Foursquare City Guide has over 50 million users and has been operating since 2009. It provides information on millions of venues around the world. This data set contains the venue name, latitude, longitude, category, and venue ID (unique identifier) for 10,159 venues in NYC. It is downloaded using an “explore” endpoint, which is a non-premium API request.

The third data set, which contains user-provided ratings for the pizza venues in New York City, is also downloaded from Foursquare. It contains the name, venue ID (unique identifier), and average user rating for the 440 pizza venues in New York City. Because not all pizza venues have user ratings on Foursquare, the dataset contains 84 missing ratings values, leaving 356 valid data points. This dataset is downloaded using a “details” endpoint, which is a premium API request.

To formulate our business recommendations based on this data, we merge the three data files. First, the pizza venues are assigned to neighborhoods using the latitude and longitude coordinates. Second, the ratings are merged with the venue data using the venue ID as the unique identifier.