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")](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json")

Here is what the first few lines of the first data file looks like:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

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.

Here is what the first few lines of the second data set look like:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop	4c537892fd2ea593cb077a28
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy	4d6af9426107f04dedeb297a
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop	4c783cef3badb1f7e4244b54
3	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy	5d5f5044d0ae1c0008f043c3
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop	4c25c212f1272d7f836385c5

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.

Here is what the first few lines of the third data set look like:

	Name	Venue ID	Venue Rating
0	Capri II Pizza	4d2cfa5cad25224bbbc5fb8f	6.8
1	Mario's Pizza	4c632f1cde1b2d7fed31e470	8.1
2	Kingsbridge Social Club	58935fd798f8aa7c14662653	9.5
3	Sam's Pizza	4bb114c4f964a520b9783ce3	8.8
4	Broadway Pizza & Pasta	4be72770910020a16f1ad514	7.4
5	Little Caesars Pizza	502bd9a6e4b0bea49203e0aa	6.6
6	Papa John's Pizza	5aa003f5b6eedb52c65bddb8	6.3
7	Domino's Pizza	4b4fbdb5f964a520811327e3	5.9
8	Acapella Gourmet Pizza & Restaurant	55906dbb498e4edbe4888785	NaN
9	Mama Maria's Pizza	4bc4f4bce58e9521483cc9e1	NaN

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.