Battle of the Neighborhoods

Lance Dugger

April 21, 2020

Problem Statement

My friend is a Masters student at the University of Pittsburgh. He has been offered a job at the university post graduation and is interested in purchasing a home. He has asked me to help him decide in which neighborhood he will purchase a home. He enjoys biking and being outside, so he is interested in being in close proximity to a park. He also wants to be able to bike to work. He is in his late twenties and still goes out to socialize frequently on the weekends so he wants the neighborhood to have a lively social scene. I will be examining neighborhood data in Pittsburgh to give him the top three neighborhoods he should consider based on his priorities.

Data Description

I will be using data from the 2010 Pittsburgh Census. The dataset can be found here: https://catalog.data.gov/dataset/pgh-snap/resource/82d3188d-e5b9-4a47-b588-75dd468a926e. This data provided all the neighborhood names from Pittsburgh as well as crime rate, median house pricing, and other neighborhood statistics. I found the latitude and longitude of each neighborhood on wikipedia and manually entered them into an excel document. The latitude and longitude information were used in conjuntion with the Foursquare API to determine the most common venues in each neighborhood and cluster the neighborhoods based on those venues.

Methodology

The data was uploaded as an excel document, and was then converted to a pandas dataframe. The data was 78 rows X 3 columns, rows representing the 78 distinct neighborhoods and the columns representing Neighborhood, Latitude, and Longitude, respectively. It is important to note that this format was not directly from the raw data, and had been manipulated in excel prior to upload.

	Neighborhood	Lat	Long
0	Allegheny Center	40.453	-80.005
1	Allegheny West	40.452	-80.016
2	Allentown	40.421	-79.994
3	Arlington	40.415	-79.970
4	Banksville	40.412	-80.039

Figure: Initial Dataframe

Methodology cont.

The data was then used to create a map of the Pittsburgh area, showing markers at the geographical center of each neighborhood. The map was generated using the Folium package.

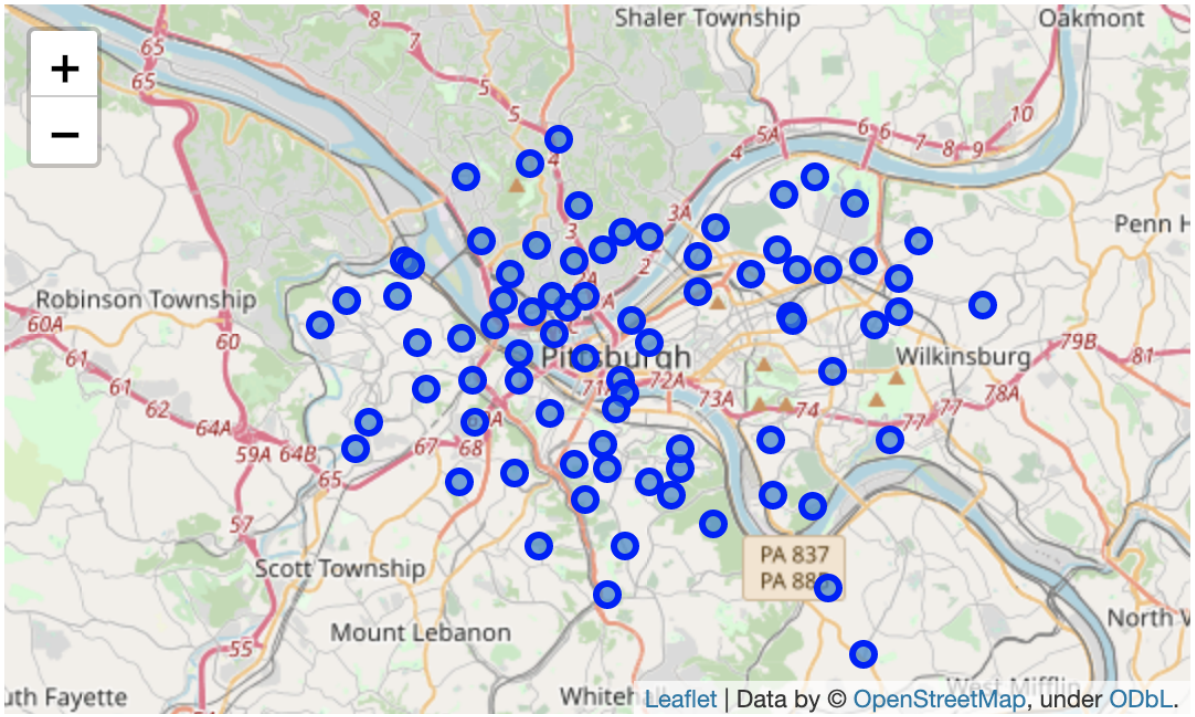

Figure: Pittsburgh Neighborhood Map

Methodology cont.

A function was defined to pass the latitude and longitude of each neighborhood to the FourSquare API and return a list of venues within 500 meters of the neighborhoods center, as defined by the latitude and longitude values. These values were assigned to a dataframe named venues.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Allegheny Center	40.453	-80.005	Children's Museum of Pittsburgh	40.452793	-80.006569	Museum
1	Allegheny Center	40.453	-80.005	Federal Galley	40.451605	-80.006045	Food Court
2	Allegheny Center	40.453	-80.005	Park House	40.453284	-80.001504	Bar
3	Allegheny Center	40.453	-80.005	El Burro	40.455860	-80.006689	Mexican Restaurant
4	Allegheny Center	40.453	-80.005	Bistro To Go	40.453450	-80.000995	Deli / Bodega

Figure: Venues Dataframe

Methodology cont.

Onehot encoding was used to assign a binary value to each of the venues in their respective neighborhoods and then those values were summed and grouped by neighborhood. Finally, a function was defined to return the most common venues in each neighborhood, and those were stored in a new dataframe. KMeans Clustering was used to determine which neighborhoods were similar to eachother, and then the individual clusters were examined to determine which neighborhood would best suit my friend's needs.

Results

Cluster 3 was selected as the best neighborhood for my friend to moved based on the following parameters: large amount of parks, breweries and bars. The figure below shows the first five neighborhoods in Cluster 3.

	Neighborhood	Lat	Long	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Allegheny Center	40.453	-80.005	2.0	Deli / Bodega	Park	Exhibit	Zoo	Bar	Café
1	Allegheny West	40.452	-80.016	2.0	American Restaurant	Pub	Sandwich Place	Fast Food Restaurant	Food Truck	Thai Restaurant
2	Allentown	40.421	-79.994	2.0	Discount Store	Chinese Restaurant	Vegetarian / Vegan Restaurant	Italian Restaurant	Coffee Shop	Caribbean Restaurant
4	Banksville	40.412	-80.039	2.0	Park	Pool	Pizza Place	Zoo	Food Truck	Food Service
6	Beltzhoover	40.416	-80.003	2.0	Park	Cosmetics Shop	Health & Beauty Service	Tennis Court	Zoo	Food Truck

Figure: Cluster 3 Dataframe

Results cont.

The neighborhoods in Cluster 3 were then mapped to show their proximity to the University of Pittsburgh. After the map was generated, the following neighborhoods were chosen for their proximity: Squirrel Hill, Shadyside, Oakland, Bloomfield, Polish Hill, Strip District, Hill District, and Central Business District.

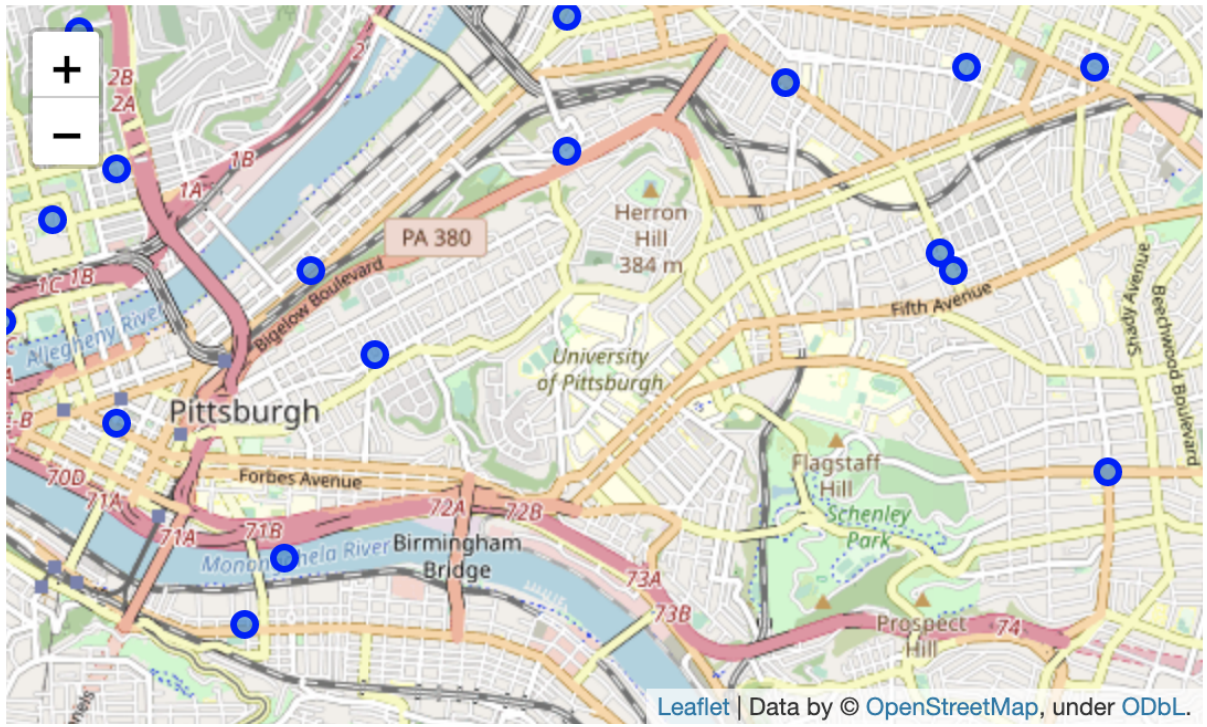

Figure: Cluster 3 Map

Observations

After reviewing both the map and the venue dataframe it is worth noting several issues with the data. Although proximity to a park was one of the top priorities, using the parks category to sort the data may have not been the best method. The method used looked for the amount of individual parks, as opposed to finding one larger park that might be more enjoyable for riding a bicycle. After viewing the map, it is clear that Squirrel Hill is actually the best neighborhood for bicycling because it is in close proximity to Schenley Park, which is one of the largest parks in the city and has a significant amount of bike trails. It is also worth noting that breweries are under-represented compared to bars, based on how broad the category bar is for classifying venues.

Conclusion

After reviewing the results I have determined that the best neighborhood for my friend to move to is Squirrel Hill based on it's proximity to University of Pittsburgh, it's proximity to Schenley Park, and it's proximity to the neighborhood of Oakland, which has a vibrant nightlife