Battle of the Neighborhoods: Where should an entrepreneur open a new yoga studio in Washington, DC?

Prepared by Robin Bloodworth on 9/29/21

Introduction

A young business entrepreneur is interested in opening a yoga studio in Washington, DC. They are new to the area and aren't sure which neighborhood would be the best fit to open their yoga studio. They want to make sure the yoga studio is located in a neighborhood that is hip and trendy enough that the yoga studio would have plenty of customers, but they also don't want to open it in a neighborhood that already has a lot of yoga studios, since the competition here might be too high for a new yoga studio to succeed.

This project will use Foursquare API data to explore and cluster neighborhoods in Washington DC based on the top 5 venue types located in each neighborhood. The results of this exploration and clustering analysis will then be used to select the top neighborhoods for the entrepreneur to consider opening their new yoga studio.

Data Sources

To conduct a neighborhood analysis of Washington, DC, I used two main data sources. The first is a JSON file from the DC government website containing all 132 neighborhoods in DC with their corresponding latitudes and longitudes, located on this page (https://opendata.dc.gov/datasets/DCGIS::neighborhood-labels/about). An example of the features contained in this dataset is shown below:


```
{'type': 'Feature',
 'properties': {'OBJECTID': 1,
  'GIS_ID': 'nhood_050',
  'NAME': 'Fort Stanton',
  'WEB_URL': 'http://NeighborhoodAction.dc.gov',
  'LABEL_NAME': 'Fort Stanton',
  'DATELASTMODIFIED': '2003-04-10T00:00:00Z'},
 'geometry': {'type': 'Point',
  'coordinates': [-76.98034770695811, 38.85565773097726]}}
```

Next, I used Foursquare API location data to extract and geocode venues in each of the neighborhoods in Washington, DC. Foursquare API location data includes many different characteristics of venues, including venue type, venue name, and latitude and longitude of the venues. See screen shot below for example:

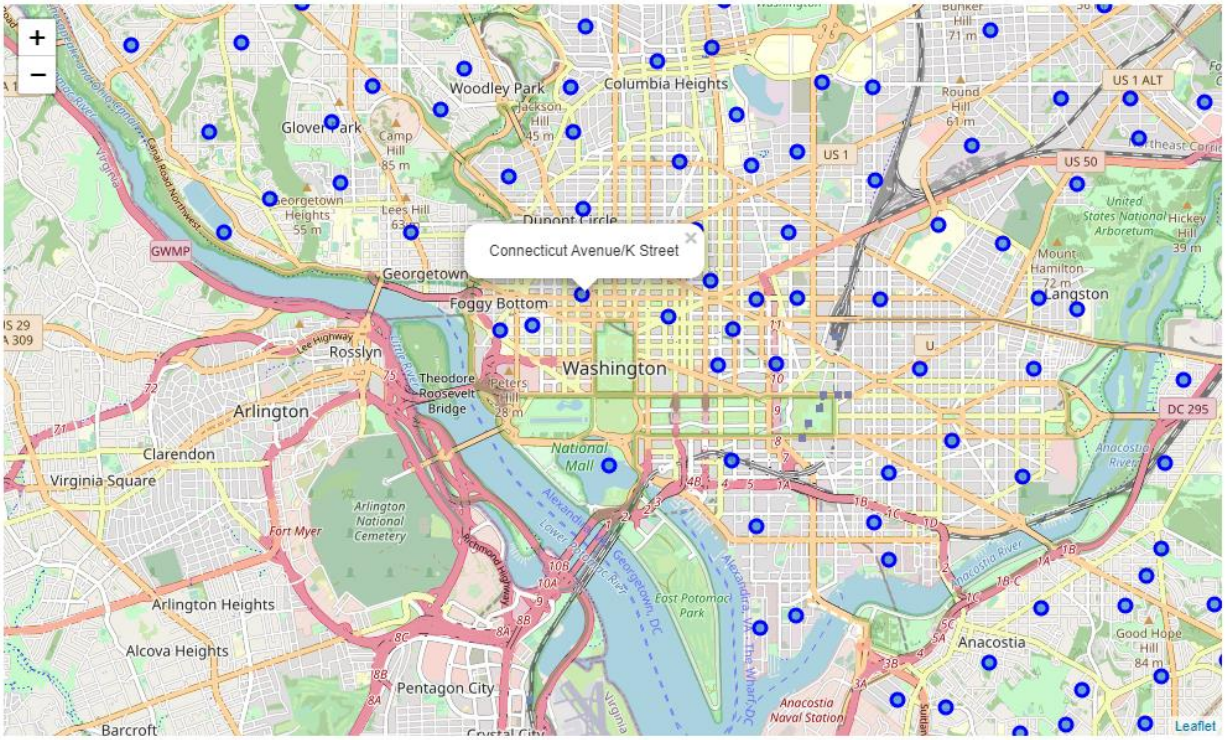
	Venue	Venue Latitude	Venue Longitude	Venue Category
	Anacostia Community Museum	38.856728	-76.976899	Museum
	Fort Stanton Recreation Center	38.857118	-76.978816	Recreation Center
	Fort Stanton Park	38.857541	-76.978266	Park
	Stanton Road SE & Suitland Parkway SE	38.853278	-76.983289	Intersection
	Douglass Community Recreation Center	38.852218	-76.977411	Park

Methods

First, I uploaded the JSON file of Washington, DC Neighborhoods from the DC city government website into my Jupyter notebook and examined the features to understand the unique composition of the JSON file. I then created a pandas dataframe with the columns “Neighborhood,” “Latitude,” and “Longitude” and imported the corresponding fields from the JSON file into the pandas dataframe and examined the first few rows and the shape to ensure proper transfer.

	Neighborhood	Latitude	Longitude
0	Fort Stanton	38.855658	-76.980348
1	Congress Heights	38.841077	-76.997950
2	Washington Highlands	38.830237	-76.995636
3	Bellevue	38.826952	-77.009271
4	Knox Hill/Buena Vista	38.853688	-76.967660

Next, I used geolocator to obtain the latitude and longitude coordinates of Washington, DC, and I used those coordinates to load a folium map of DC. I then added the neighborhoods to the map using folium using circles with blue popups with the neighborhood's name (see screenshot below).



I queried foursquare for a list of nearby venues for each neighborhood including venue name, category, latitude, and longitude, and imported these into a pandas dataframe, and I checked the first few rows and shape to ensure proper transfer.

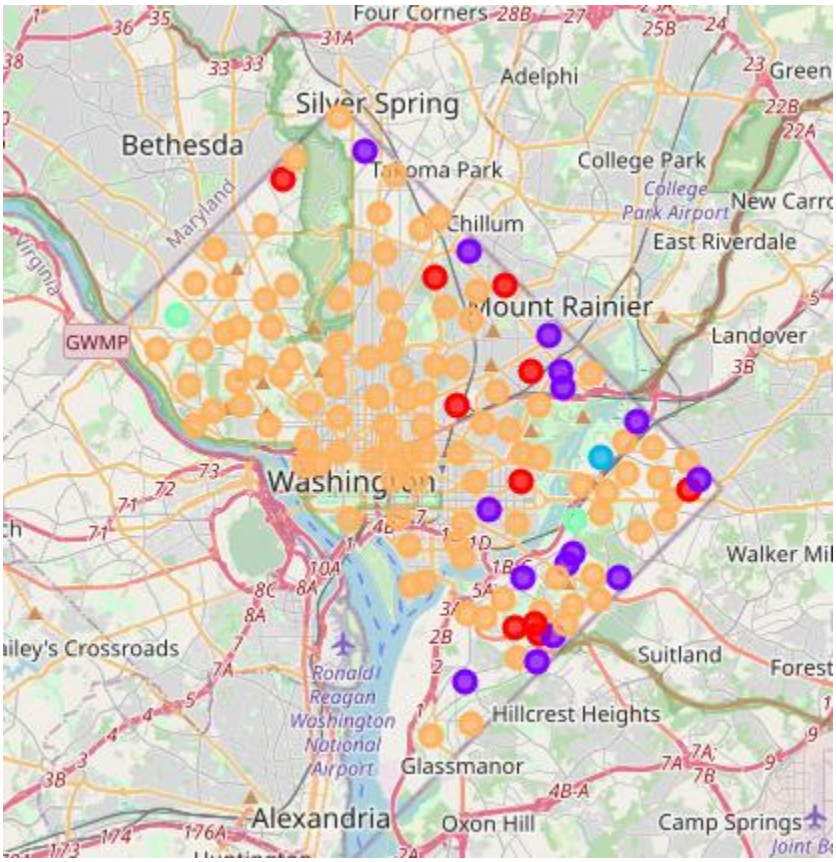
Next, I grouped the venues by neighborhood to total how many venues were in each neighborhood and calculated the total number of unique venue categories.

I then used one hot encoding to designate dummy variables for each venue category for each venue attached to each neighborhood, which I then used to calculate the mean number of each venue type for each neighborhood.

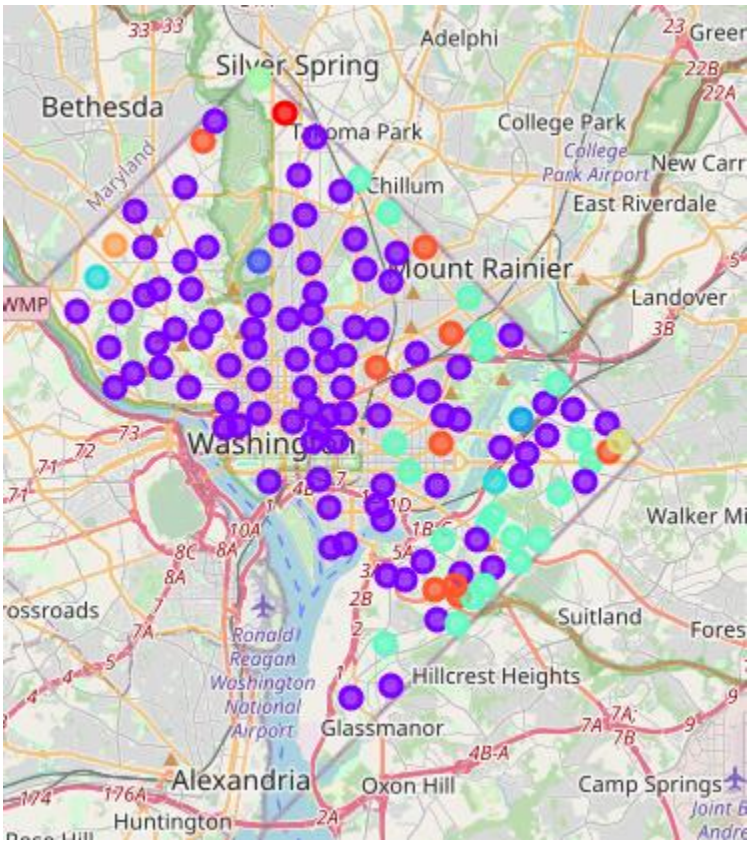
I then calculated the top 10 venues in each neighborhood, printed them to inspect, and then imported them into a pandas dataframe (see screenshot below).

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
16th Street Heights	Bus Stop	Coffee Shop	Food	Pizza Place	Bed & Breakfast
Adams Morgan	Bar	Ice Cream Shop	Italian Restaurant	Asian Restaurant	Spa
American University Park	Italian Restaurant	BBQ Joint	Photography Lab	Pet Store	Pet Service
Arboretum	Ice Cream Shop	Nightclub	Fast Food Restaurant	Botanical Garden	Chinese Restaurant
Barnaby Woods	Field	Gym / Fitness Center	BBQ Joint	Park	Performing Arts Venue
Barry Farm	Bus Stop	Rental Car Location	Intersection	Basketball Court	Metro Station
Bellevue	Pizza Place	Shoe Repair	Playground	Chinese Restaurant	ATM
Benning	Fried Chicken Joint	Discount Store	Liquor Store	Donut Shop	Shopping Mall
Benning Ridge	Burger Joint	Convenience Store	Shipping Store	Insurance Office	Record Shop
Bloomingdale	Bus Stop	Italian Restaurant	Liquor Store	Dog Run	Dive Bar
Brentwood	Sports Club	Sandwich Place	Organic Grocery	Grocery Store	Food Truck

At this point the data were ready to be analyzed, so I clustered the neighborhoods using k-means clustering, with number of clusters to be 5. However, this gave me some lopsided clusters, with one cluster seeming to have the vast majority of the venues (see screenshot below).



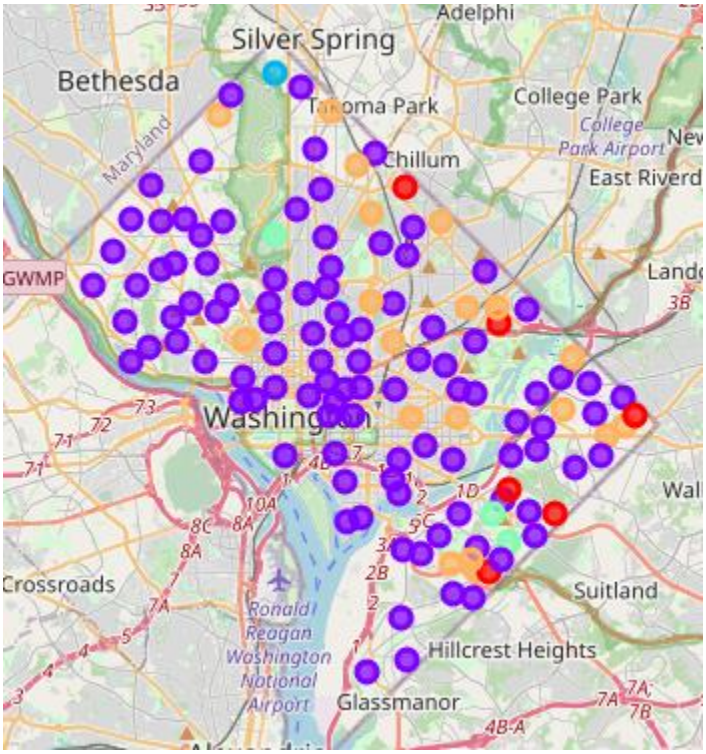
Then I tried 10 clusters, which didn't seem to improve the issue (see screen shot below), so I went back to 5.



Finally, I examined the top 5 venues in the neighborhoods of each cluster to decide which cluster would be the best cluster to put a yoga studio. Once I decided on which cluster to use, I dropped all neighborhoods with any of the following in their top 5 venues (gym, gym/fitness center, sports club, dance studio, recreation center). Finally, I kept all neighborhoods with a restaurant in their top venue category. This left me with a list of neighborhoods in the most trendy cluster of neighborhoods that did not have too many gyms, but that were "hip" enough by having a restaurant as their top venue.

Results

There are 132 neighborhoods in Washington, DC. These neighborhoods contain a total of 2,643 venues, which fall into 313 categories. Neighborhoods contained anywhere from 1 to 100 neighborhoods. See below the final screen shot of clusters:



See the following table for frequencies of neighborhoods in each cluster, which ranged from 1 to 101.

Cluster	Number of Neighborhoods
Cluster 0	6
Cluster 1	101
Cluster 2	1
Cluster 3	3
Cluster 4	19

Cluster 0 had liquor stores as the top venue in all neighborhoods, a gym in the top 5, and more businesses the leisure activities/restaurants, so this cluster didn't seem to have the hip neighborhoods in it that would be ideal for a yoga studio.

Cluster 1 definitely had several neighborhoods with gyms in the top five venues, but it also had several neighborhoods with several restaurants and/or leisure activities that imply the neighborhoods are hip.

Cluster 2's neighborhood did not seem trendy. It did have a restaurants in the 4th and 5th most common venues, but the top three were bus stop, ATM, and repair shop. This indicates to me that the neighborhood is small and not very trendy.

Cluster 3 had a couple of gyms in the top 5, and again did not have as many restaurants as businesses, so it seems less hip and trendy.

Finally, Cluster 4 was similar to cluster 1 in that it had several neighborhoods with gyms, and even a yoga studio, in top 5 venues, but the other neighborhoods seemed to be more commercial and business focused than Cluster 1.

Based on examining the top 5 venues in each neighborhood by cluster, I decided that Cluster 1 would be the best place to open a yoga studio.

Starting with the 101 neighborhoods in cluster one, I then dropped all neighborhoods with any type of gym or fitness type center in the top 5 venues, which left me with 84 neighborhoods. Finally, I kept only neighborhoods with restaurants in the top venue, which left me with 14 neighborhoods. Any of these 14 neighborhoods seem hip enough to open up a yoga studio. See the screen shot below:

Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
American University Park	38.947612	-77.090250	1	Italian Restaurant	BBQ Joint	Pet Store	Pet Café	Peruvian Restaurant
Brightwood	38.966379	-77.026874	1	Chinese Restaurant	Mexican Restaurant	Pizza Place	Southern / Soul Food Restaurant	Gas Station
Burrville	38.900243	-76.921802	1	American Restaurant	Seafood Restaurant	Fish & Chips Shop	Chinese Restaurant	ATM
Cardozo/Shaw	38.917168	-77.027550	1	New American Restaurant	Bar	Coffee Shop	American Restaurant	Pizza Place
Chinatown	38.899151	-77.020135	1	American Restaurant	Café	Hotel	Cocktail Bar	Italian Restaurant
Cleveland Park	38.936098	-77.064402	1	Mexican Restaurant	Xinjiang Restaurant	Steakhouse	Indian Restaurant	Thai Restaurant
Dupont Circle	38.912128	-77.040984	1	Thai Restaurant	Italian Restaurant	Spa	Pizza Place	Greek Restaurant
Georgetown	38.909556	-77.064796	1	American Restaurant	Coffee Shop	Pizza Place	Dessert Shop	Vietnamese Restaurant
Historic Anacostia	38.863186	-76.984678	1	American Restaurant	Convenience Store	Coffee Shop	Fast Food Restaurant	Comfort Food Restaurant
Mount Pleasant	38.931741	-77.040656	1	Latin American Restaurant	Grocery Store	Café	Thai Restaurant	Liquor Store
Palisades	38.929808	-77.103520	1	American Restaurant	Shipping Store	Deli / Bodega	Coffee Shop	Convenience Store
Penn Quarter	38.895370	-77.022181	1	American Restaurant	Art Museum	Theater	Salad Place	Indian Restaurant
Shepherd Park	38.982980	-77.032126	1	Ethiopian Restaurant	Department Store	Women's Store	Liquor Store	Boutique
Trinidad	38.908462	-76.982737	1	Asian Restaurant	American Restaurant	Deli / Bodega	Breakfast Spot	Moving Target

Discussion

Using geocoded neighborhoods data from an open source page from the government of Washington, DC and venue location and attributes data from Foursquare API, I was able to use K-means Clustering machine learning to identify the best neighborhoods to open up a yoga studio in Washington, DC that did not have too many gyms already, but were also hip and trendy enough where a yoga studio would do well. Based on this spatial analysis, an entrepreneur could then investigate other market factors of the neighborhoods (e.g. rent prices, demographics, etc.) to further zero in on the best location that worked well for them.

Conclusion

This analysis shows how open data sources and data science methods can be used to solve real world problems.