IBM Data Science Capstone

Battle of the Neighborhoods -Washington DC

Problem

Washington, DC is an affluent and dynamic city with a thriving restaurant scene. I would like to open a new restaurant, and based on the neighborhood data scraped from the Foursquare API. There are many areas that are overserved, and some that are underserved, and this analysis will determine which areas could be good sites for a new restaurant.

Using these data, I used the following approach:

1. List the Washington, DC neighborhoods
2. Cluster the neighborhoods using K-means clustering based on Foursquare data
3. Determine the number of venues in each cluster
4. Contrast that with the number of users in each cluster
5. Determine which neighborhood is underrepresented and would best support a restaurant

Data Sources

Using Python machine learning, a list of neighborhoods in Washington DC will be clustered using K-means clustering. Each neighborhood has its restaurants, and using this dataset, I will determine which locations would be best to put this restaurant.

The criteria need also to be a popular, and densely populated neighborhood, while being under-represented by restaurants.

The restaurant data will come from the Foursquare API.

The DC neighborhood data will come from the following sources:

- Open Data DC https://opendata.dc.gov/datasets/ where the list of neighborhoods and their locations will be scraped.
- DC.gov office of GIS services https://octo.dc.gov/service/dc-gis-services where additional location data will be retrieved

This is an API that allows scraping for analysis. This will provide a list of neighborhoods, and the neighborhood latitude and longitude. Using this data, I will cluster venues by neighborhood, and determine which neighborhoods are underserved by venue type and population density. From that analysis, I will determine where I need to establish my new restaurant.

List of Neighborhoods

A list of neighborhoods was scraped from https://opendata.arcgis.com/. This data contained both the list of neighborhoods as well as the latitude and longitude of each neighborhood.

Table 1

	X	Y	OBJECTID	GIS_ID	NAME	WEB_URL	LABEL_NAME	DATELASTMODIFIED
0	-76.980348	38.855658	1	nhood_050	Fort Stanton	http://NeighborhoodAction.dc.gov	Fort Stanton	2003/04/10 00:00:00+00
1	-76.997950	38.841077	2	nhood_031	Congress Heights	http://NeighborhoodAction.dc.gov	Congress Heights	2003/04/10 00:00:00+00
2	-76.995636	38.830237	3	nhood_123	Washington Highlands	http://NeighborhoodAction.dc.gov	Washington Highlands	2003/04/10 00:00:00+00
3	-77.009271	38.826952	4	nhood_008	Bellevue	http://NeighborhoodAction.dc.gov	Bellevue	2003/04/10 00:00:00+00
4	-76.967660	38.853688	5	nhood_073	Knox Hill/Buena Vista	http://NeighborhoodAction.dc.gov	Knox Hill/Buena Vista	2003/04/10 00:00:00+00

Cleaned List

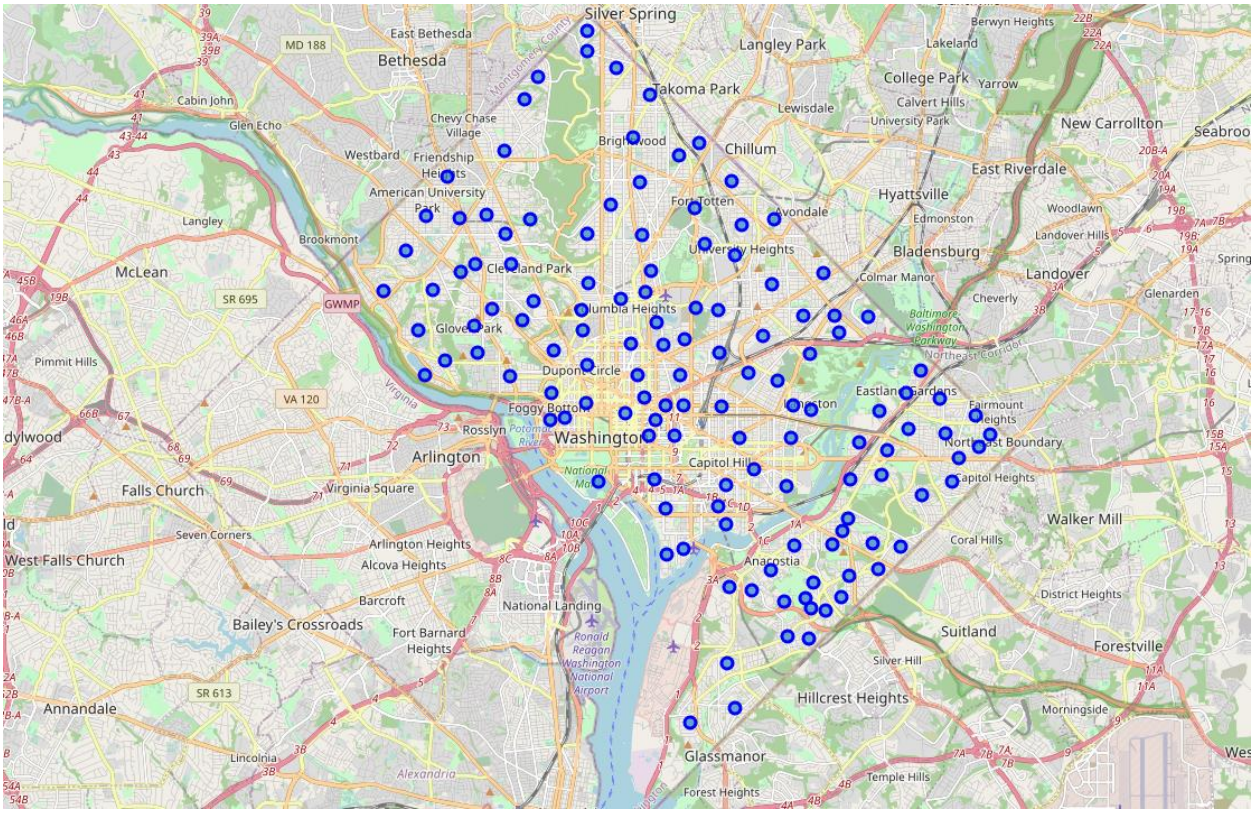
The data was cleaned and only the latitude and longitude remained. The top 5 neighborhoods are listed below.

Table 2

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

Map of DC Neighborhoods

Using these data, map of Washington DC neighborhoods was generated. There are 131 individual neighborhoods in this list. The generated map was compared with a Google map of Washington DC, and every neighborhood was in the correct location.



FourSquare Data Set

The FourSquare site was scraped for a list of DC venues. Each venue was listed, as well as the type of venue and the latitude and longitude of each venue.

Table 3

	name	categories	lat	lng
0	Washington Monument	Monument / Landmark	38.889401	-77.035244
1	National Museum of African American History an...	History Museum	38.891171	-77.032818
2	World War II Memorial	Monument / Landmark	38.889377	-77.040516
3	The Hay-Adams	Hotel	38.900510	-77.036885
4	Renwick Gallery	Art Museum	38.898962	-77.039189

Venues in each Neighborhood

Then the number of venues in each neighborhood was generated.

Table 4

Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category
16th Street Heights	16	16	16	16
Adams Morgan	61	61	61	61
American University Park	2	2	2	2
Arboretum	16	16	16	16
Barnaby Woods	4	4	4	4
…	…	…	…	…
West End	49	49	49	49
Woodland	4	4	4	4
Woodland-Normanstone	5	5	5	5
Woodley Park	23	23	23	23
Woodridge	5	5	5	5

Neighborhood Cluster

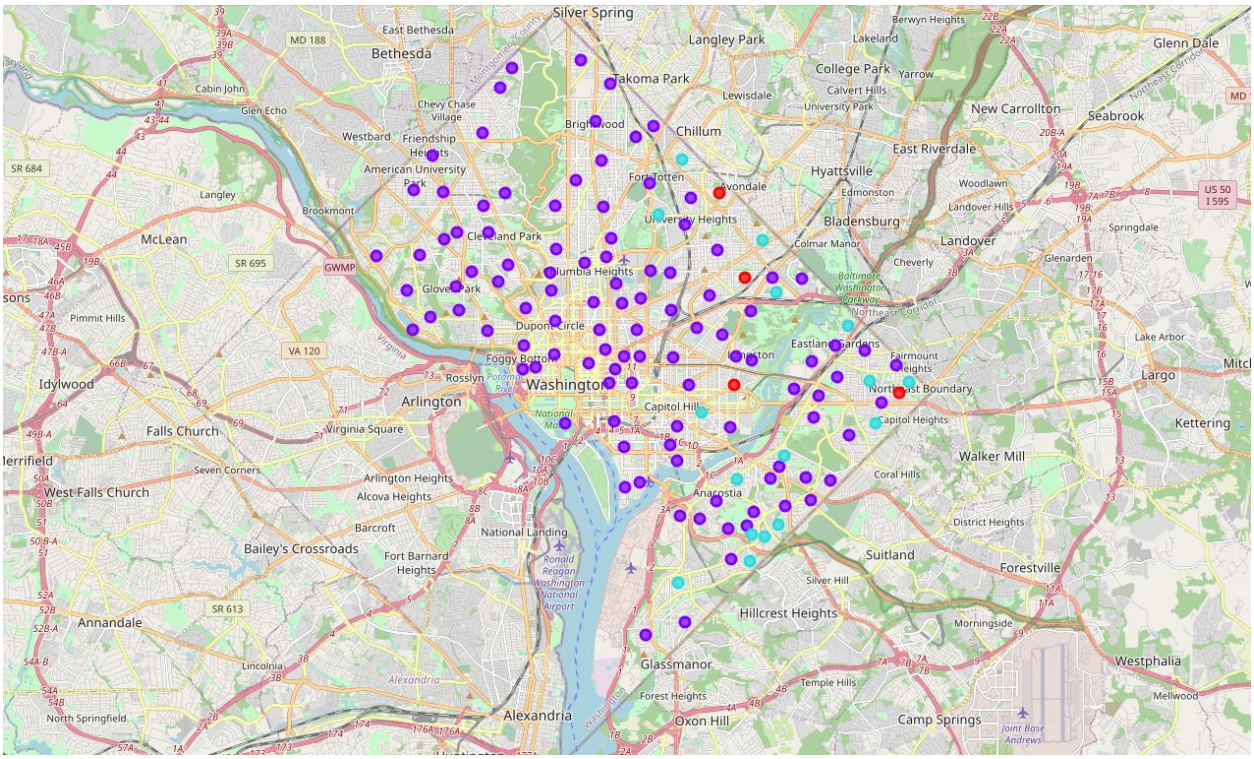
The neighborhoods were clustered into 4 groups using k-means clustering based on the fact there are 4 quadrants to DC (NW, SW, NE, and SE) and the top 10 venues for each neighborhood were listed. Each neighborhood was merged to the latitude and longitude

Table 5

Longitude	**Latitude**	**Neighborhood**	**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**	**7th Most Common Venue**	**8th Most Common Venue**	**9th Most Common Venue**	**10th Most Common Venue**
-76.980348	38.855658	Fort Stanton	1.0	Park	Museum	Intersection	American Restaurant	Dog Run	Art Gallery	Flower Shop	Fast Food Restaurant	Field	Filipino Restaurant
-76.997950	38.841077	Congress Heights	2.0	Liquor Store	Ice Cream Shop	Deli / Bodega	Health & Beauty Service	Convenience Store	Road	Tennis Court	American Restaurant	Intersection	Fried Chicken Joint
-76.995636	38.830237	Washington Highlands	1.0	Grocery Store	Liquor Store	Asian Restaurant	Seafood Restaurant	Basketball Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Fish Market
-77.009271	38.826952	Bellevue	1.0	Baseball Field	Pizza Place	Shoe Repair	Playground	Basketball Court	Exhibit	Eye Doctor	Falafel Restaurant	Farmers Market	Fast Food Restaurant
-76.967660	38.853688	Knox Hill/Buena Vista	2.0	Liquor Store	Grocery Store	Convenience Store	Fish Market	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop

Map of DC Clusters

A map of the clusters was then generated. This map shows 4 clusters, with the majority of neighborhoods winding up in cluster 1, in purple. There are also markers for cluster 0 (red), cluster 2 (blue), and cluster 3 (green).



Results and Conclusions

From the map above, cluster 1 (purple) is well represented by restaurants and other venues, so I will not put a restaurant there.

Likewise, Cluster 0 (red) and cluster 3 (green) appear to have plenty of restaurants.

Cluster 2 (blue) is underrepresented by restaurants. I will open my restaurant in one of the neighborhoods in cluster 2. I will use the data I generated to further examine the neighborhoods listed in the table below to determine where I will put my restaurant.

Table 6

Neighborhood	**1st Most Common Venue**	**2nd Most Common Venue**	**3rd Most Common Venue**	**4th Most Common Venue**	**5th Most Common Venue**	**6th Most Common Venue**	**7th Most Common Venue**	**8th Most Common Venue**	**9th Most Common Venue**	**10th Most Common Venue**
Congress Heights	Liquor Store	Ice Cream Shop	Deli / Bodega	Health & Beauty Service	Convenience Store	Road	Tennis Court	American Restaurant	Intersection	Fried Chicken Joint
Knox Hill/Buena Vista	Liquor Store	Grocery Store	Convenience Store	Fish Market	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop
Shipley	Convenience Store	Dance Studio	Performing Arts Venue	Liquor Store	Chinese Restaurant	Wings Joint	Food Service	Food & Drink Shop	Food	Food Truck
Garfield Heights	Convenience Store	Bus Stop	Park	Wings Joint	Yoga Studio	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop
Twining	Liquor Store	Bike Rental / Bike Share	Restaurant	Pharmacy	Convenience Store	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant
Fairlawn	Liquor Store	Fried Chicken Joint	Sandwich Place	Shop & Service	Deli / Bodega	Hostel	Hospital	Hotel	Exhibit	Eye Doctor
Pleasant Hill	Sandwich Place	Dance Studio	Bus Stop	Chinese Restaurant	Liquor Store	Gym	Discount Store	Fish Market	Falafel Restaurant	Farmers Market