Battle of Neighborhoods

London, Washington

BY
Luciano Guerra Domínguez

Problem Statement

- I have to move to Washington DC from my actual neighborhood in London.
- We have an store in London with great success.
- There is business option to open a new store in Washington with similar typology as our London store.
- How can we minimize the risk of opening a new store in Washington and have no success?

Objective

- Collecting Neighborhood's top trending venues using Foursquare API(Beautiful Soup, http request)
- Forming neighborhood clusters based on venue categories using unsupervised k-mean clustering algorithm(sklearn)
- Identifying and understanding the similarities and differences between two chosen neighborhoods to retrieve more insights and to conclude with ease which neighborhood wins over other.
- Detect the closets neighborhood to my University to place our store.

Python packages and Dependencies:

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- Geopy – To retrieve Location Data
- Matplotlib – Python Plotting Module
- Sklearn – Python machine learning Library
- Folium – Map rendering Library

Work flow

- Web Scraping and Data Wrangling
- Top Trending Places Extraction and Clustering
- Decision Making based on the clustered neighborhoods and distance between neighborhoods and my target place (Gerogetown University)

Web Scraping and Data Wrangling

London Neighborhood

	Code	Area_name	Inner/_Outer_London	GLA_Population_Estimate_2017	GLA_Household_Estimate_2017
0	E09000001	City of London	Inner London	8800	5326
1	E09000002	Barking and Dagenham	Outer London	209000	78188
2	E09000003	Barnet	Outer London	389600	151423
3	E09000004	Bexley	Outer London	244300	97736
4	E09000005	Brent	Outer London	332100	121048

Washington Neighborhood

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

Compact Dataset

Venues Extraction using Four Square API and Clustering

Four Square API Calls to Collect Neighborhood Venue Category and LAT/LNG

	name	categories	lat	lng
0	Il Corvo	Italian Restaurant	47.602522	-122.331952
1	Biscuit B*tch	Breakfast Spot	47.603237	-122.332010
2	Columbia Tower Club	Social Club	47.604507	-122.330484
3	Juicy Cafe	Café	47.604329	-122.330958
4	Tat's Delicatessen	Sandwich Place	47.601901	-122.332423

One Hot Encoding to Convert Labels into Numbers

	Neighborhood	Yoga Studio	ATM	Accessories Store	Advertising Agency	Airport
0	Alki Beach	0.00	0.0	0.0	0.025	0.0
1	Bainbridge Island	0.00	0.0	0.0	0.000	0.0
2	Ballard	0.01	0.0	0.0	0.000	0.0
3	Bellevue	0.00	0.0	0.0	0.000	0.0
4	Belltown	0.00	0.0	0.0	0.000	0.0

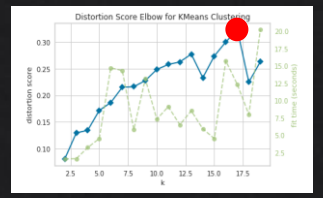
Venues Grouped by Neighborhood 229 Unique Venues

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Alki Beach	Park	Ice Cream Shop	Coffee Shop	Thai Restaurant
1	Bainbridge Island	Tree	Women's Store	Fabric Shop	Forest
2	Ballard	Coffee Shop	Brewery	Mexican Restaurant	Sandwich Place
3	Bellevue	Indian Restaurant	Coffee Shop	Other Repair Shop	Spa
4	Belltown	Bar	Coffee Shop	Gym	Sushi Restaurant

K-Means Clustering

Silhouette Score

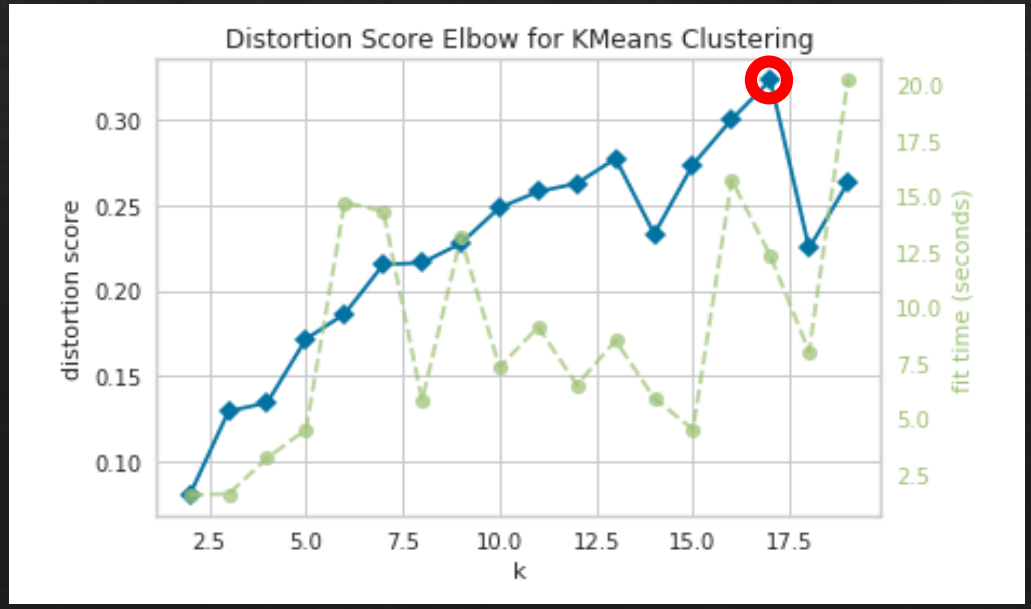
Elbow Method



- Cluster 1
- Cluster 2
- Cluster 3

Elbow Criterion Method

By default, the scoring parameter metric is set to distortion, which computes the sum of squared distances from each point to its assigned center. However, two other metrics can also be used with the KElbowVisualizer – silhouette and calinski_harabaz. The silhouette score calculates the mean Silhouette Coefficient of all samples, while the calinski_harabaz score computes the ratio of dispersion between and within clusters.



sklearn.metrics.silhouette_score

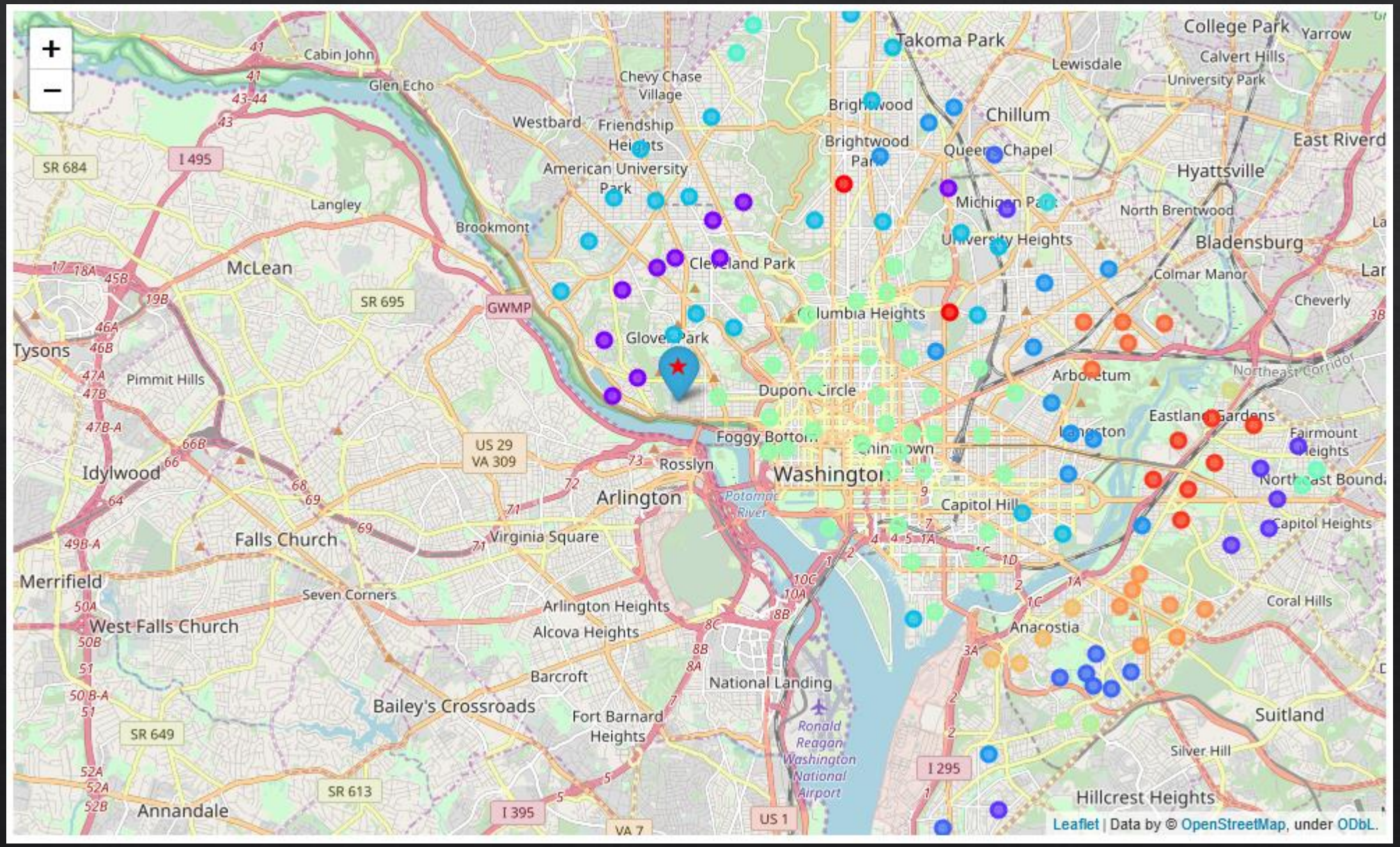
The Silhouette Coefficient is calculated using the

mean intra-cluster distance (a) and
the mean nearest-cluster distance (b) for each sample.

The Silhouette Coefficient for a sample is (b - a) / max(a, b).

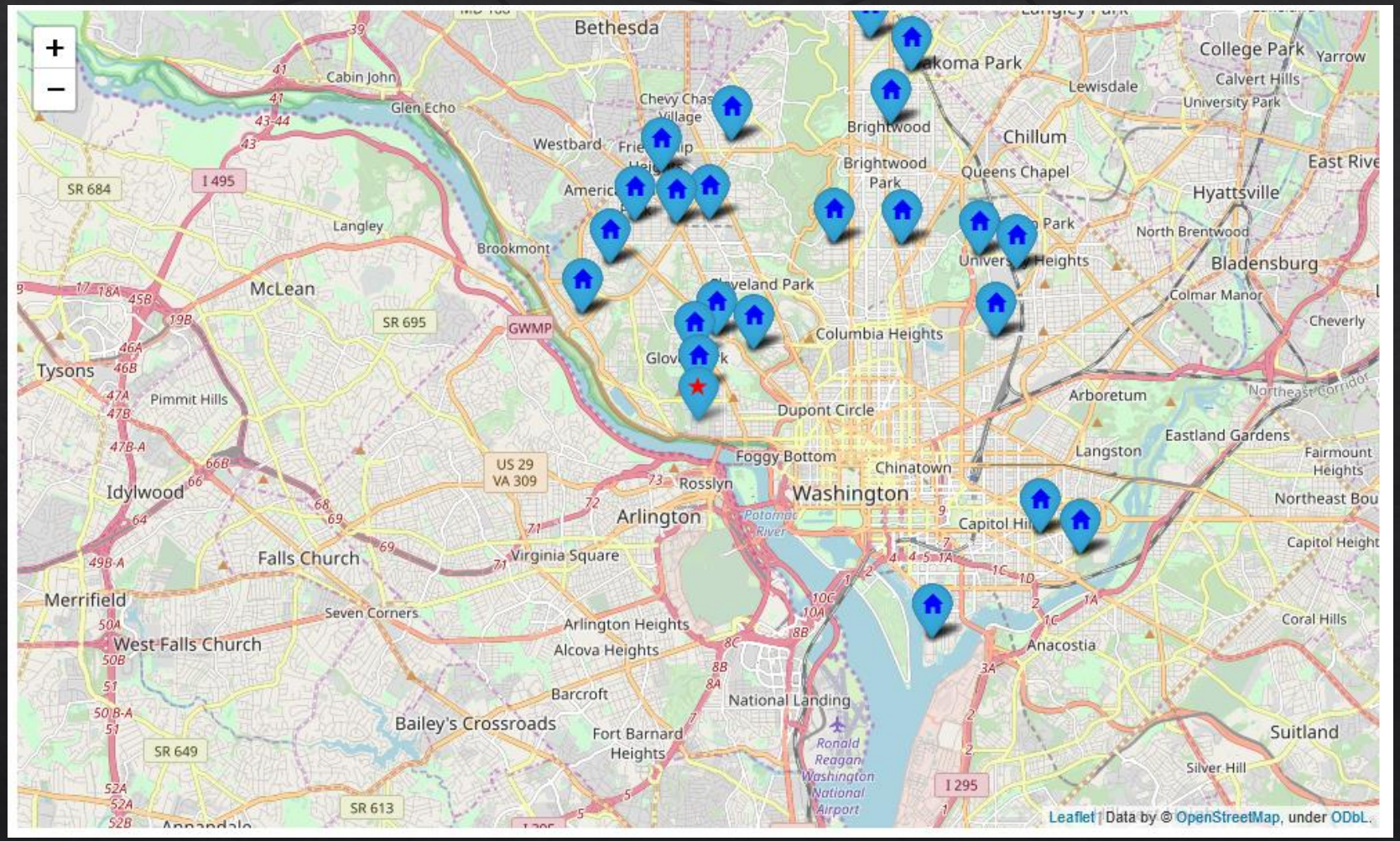
The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.

Neighborhood Clusters

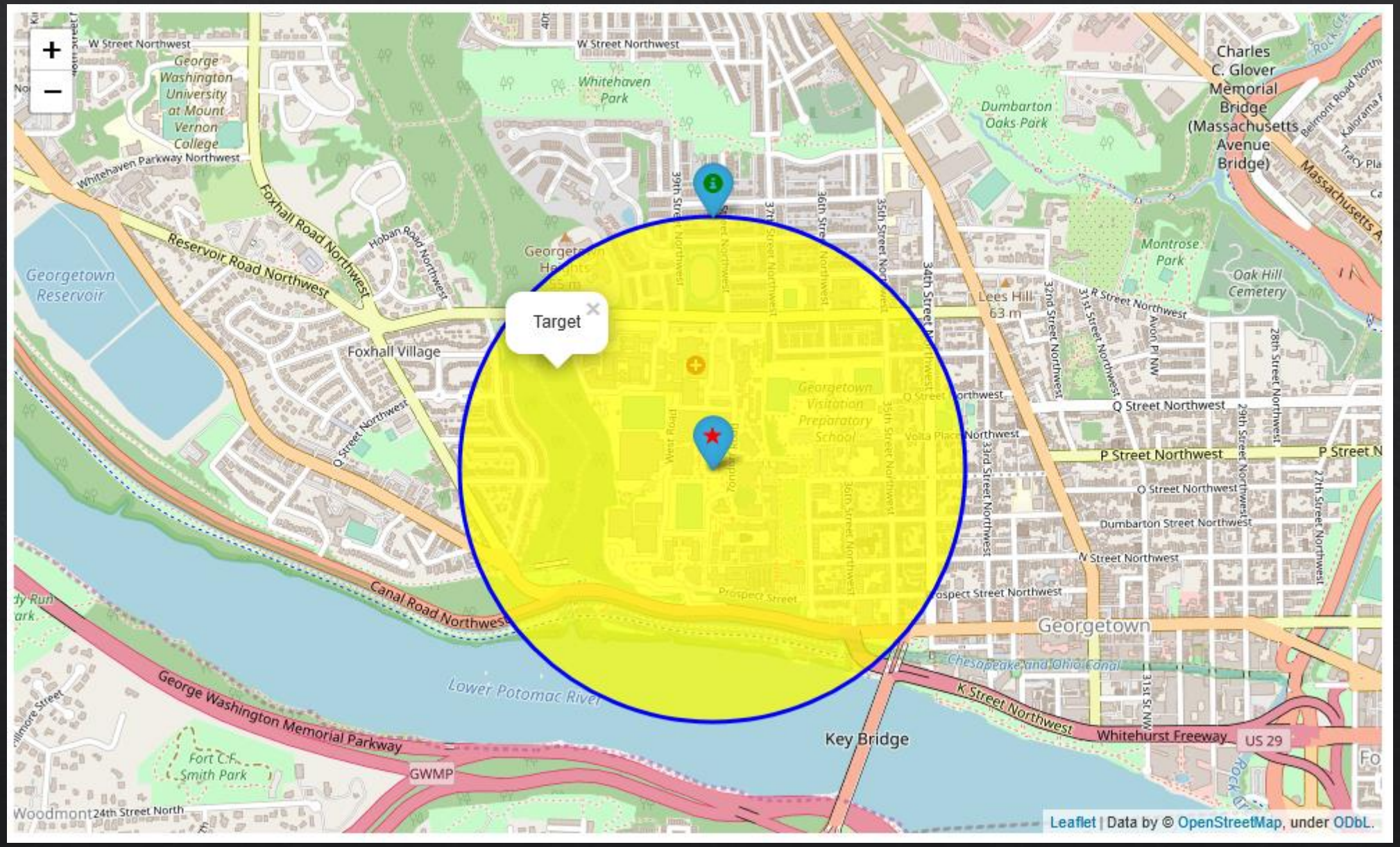


Decision Making

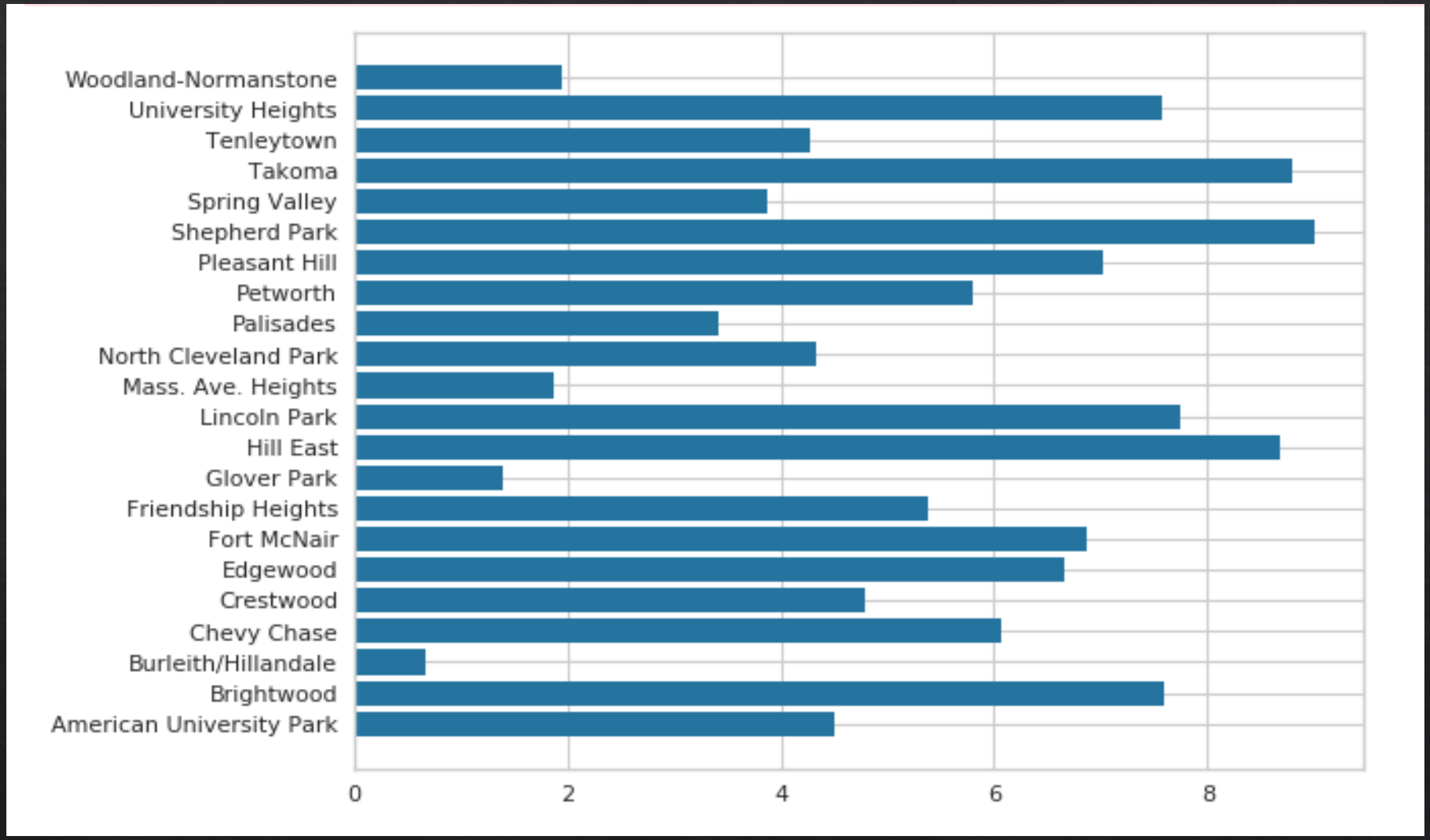
Clusters like London Neighborhood



Distance Comparation



Distances


Woodland-Normanstone
University Heights
Tenleytown
Takoma
Spring Valley
Shepherd Park
Pleasant Hill
Petworth
Palisades
North Cleveland Park
Mass. Ave. Heights
Lincoln Park
Hill East
Glover Park
Friendship Heights
Fort McNair
Edgewood
Crestwood
Chevy Chase
Burleith/Hillandale
Brightwood
American University Park
0
2
4
6
8

conclusion

- This Analysis concludes that the best neighborhood in Washington to place our store is:
  - Burletih/ Hillande

	City	Distance	Latitude	Longitude	Neighborhood	Cluster	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
85	Washington	0.67316	38.915003	-77.074566	Burleith/Hillandale	5	Pizza Place	Coffee Shop	Sandwich Place	Park	Mexican Restaurant	Italian Restaurant	Vietnamese Restaurant	Trail	Bar	Bagel Shop

Thank you!