DC Block Clustering

Nicole Bills

Objective

Create a neighborhood explorer tool to help residents and visitors develop an understanding of Points of Interest in Washington, DC

Use Cases

- Plan a trip
- Select real estate
- Conduct city planning

Approach

Cluster Blocks based on Points of Interest from Open Data DC

- Schools
- Historical Sites
- Museums
- Monuments

- Fire Stations
- Police Stations
- Libraries
- Metro Stations

- Universities
- Hospitals
- Public Services
- Recreation Centers

Process

Clean &
Categorize

Cluster

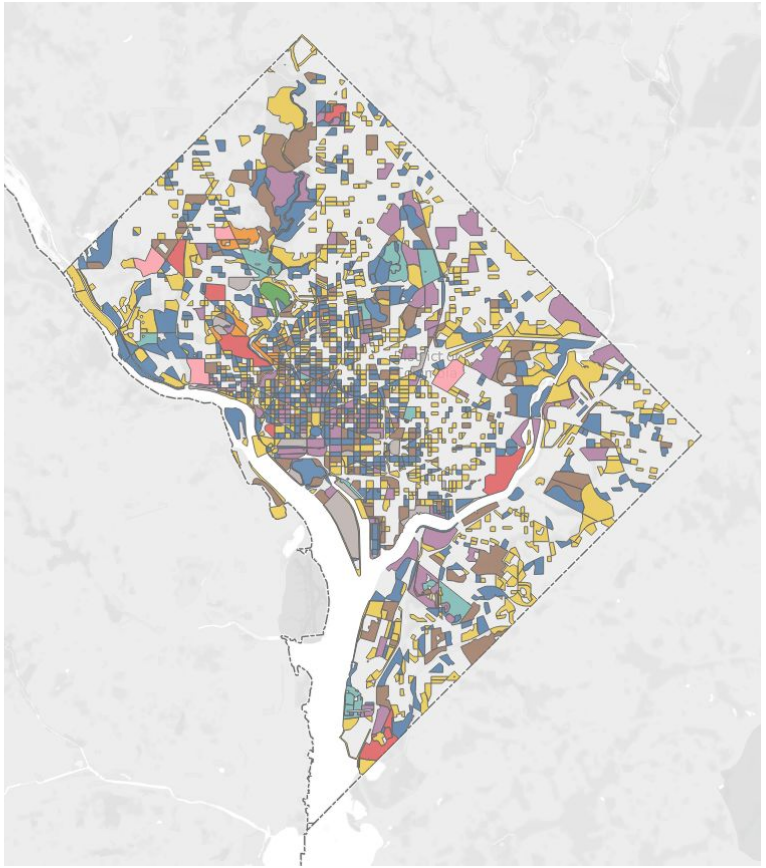
Score

Results

Cluster		Description
blue	:	convention centers, artsy, historic
orange	:	historic, packed with embassies
red	:	miscellaneous
teal	:	artsy, historic, civics
green	:	recreation
yellow	:	nature access
purple	:	civic, scholastic, public safety
pink	:	scholastic, packed with universities
brown	:	artsy, scholastic
gray	:	nature access, monuments

Results

Blocks of the same color have a similar composition in terms of Points of Interest



link

Next Steps

- Add commercial establishments in Yelp dataset
- Enable users to explore specific addresses
- Create photo explorer tool using Flickr API

Thank You!

Appendix

Process

Clean and Categorize

- Limit Points of Interest to public buildings, universities, and historic places
- Categorize based on function

Cluster

- Utilize machine learning to create ten clusters based on quantity of Points of Interest across categories

Score

- Rank clusters across five dimensions (Arts, Civics, History, Nature, and Recreation) based on percent composition within category

Scores

Cluster	Arts	Civics	History	Nature	Recreation
blue	3	3	4	3	4
orange	1	1	5	1	1
red	1	3	1	2	2
teal	5	4	5	2	2
green	2	1	1	1	5
yellow	2	2	3	5	3
purple	5	5	2	4	3
pink	3	1	2	3	1
brown	4	3	3	4	4
gray	4	1	4	5	5

Points of Interest

Baruch Bench of Inspiration

Samuel Hahnemann Memorial

Arena Stage

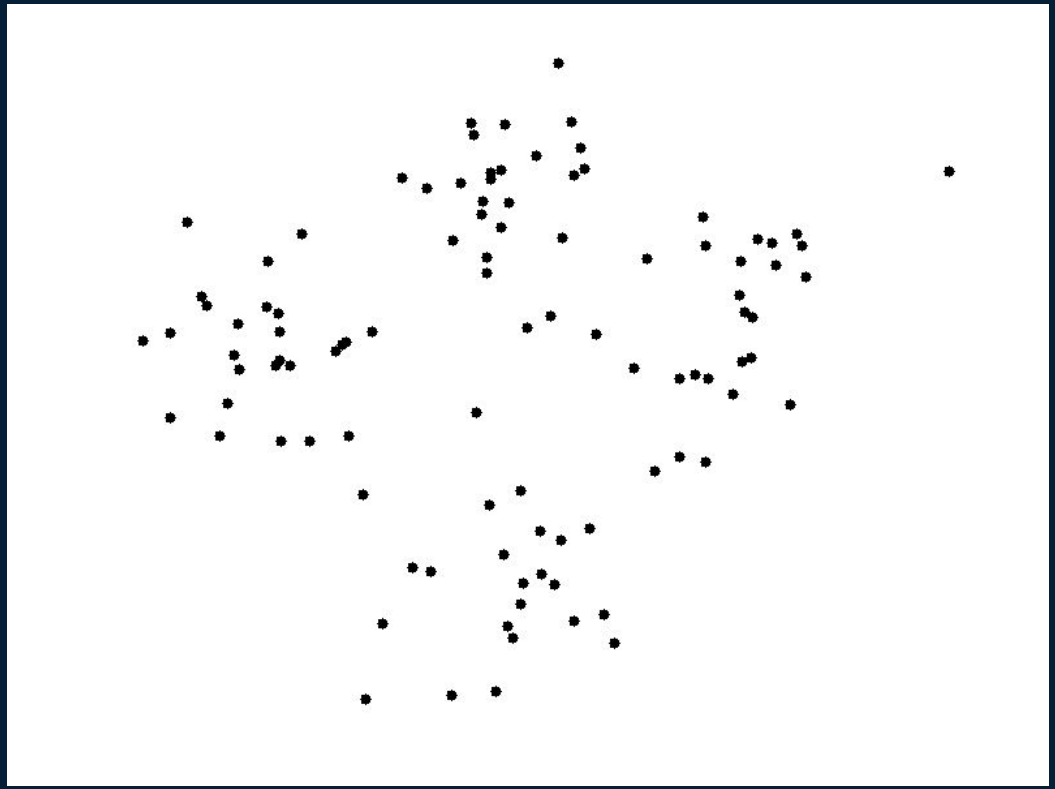
Japanese Lantern

Original Roosevelt Memorial

[k means]

Clustering

Unsupervised technique to create groups with high intra-group similarity / inter-group disimilarity



Source: Visuals and Animations by Andrey A. Shabalin, Ph.D.

[k means]

Clustering

Unsupervised technique to create groups with high intra-group similarity / inter-group disimilarity

1. Select k initial seeds
2. Assign each observation to a cluster to which it is "closest"
3. Recompute the cluster centroids
4. Reassign the observations to one of the clusters according to some rule
5. Stop if there is no reallocation

Silhouette Score

Higher score indicates better fit across different numbers of clusters