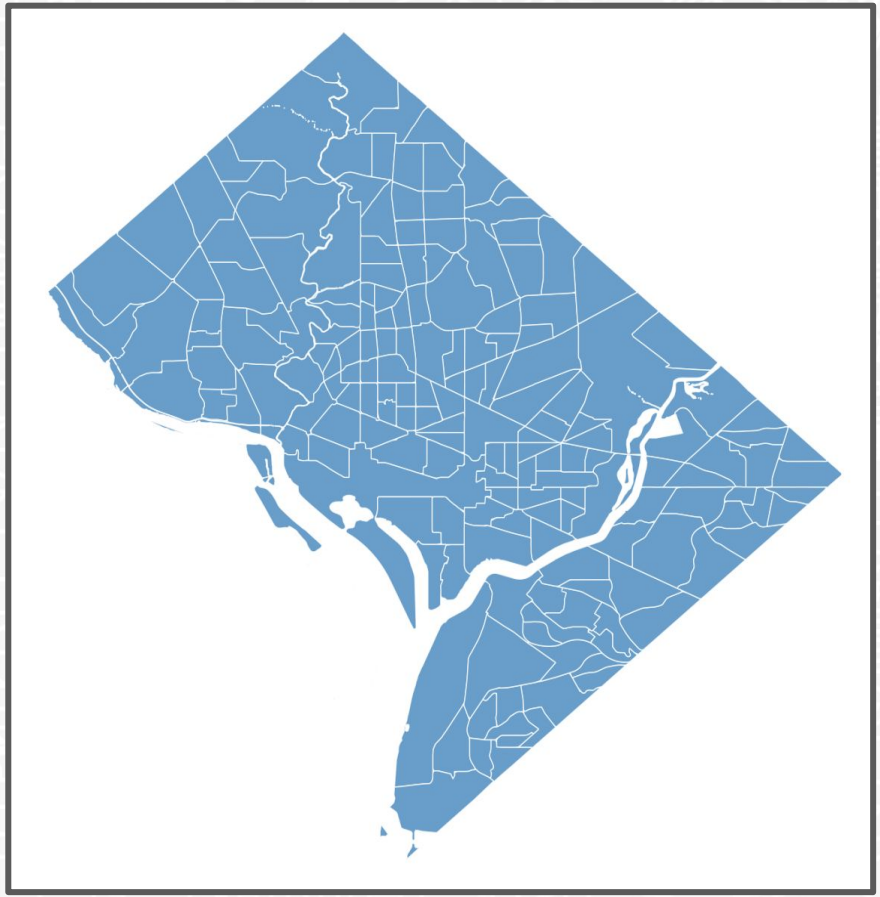
DC Geodemograhpic Segmentation

Applying Machine Learning to Cluster DC Neighborhoods

Eric R. Schultz

Overview:

- The Problem
- The Data
- The Method
- The Results
- Conclusions and Recommendations



Problem:

Where and to what extent is DC similar and dissimilar across a host of socio-demographic variables?

How can we measure groups or clusters of likeness across those variables?

Data:

- 2018 American Community Survey
- 12 sociodemographic variables
- 179 DC Census Tracts

AMERICAN COMMUNITY SURVEY

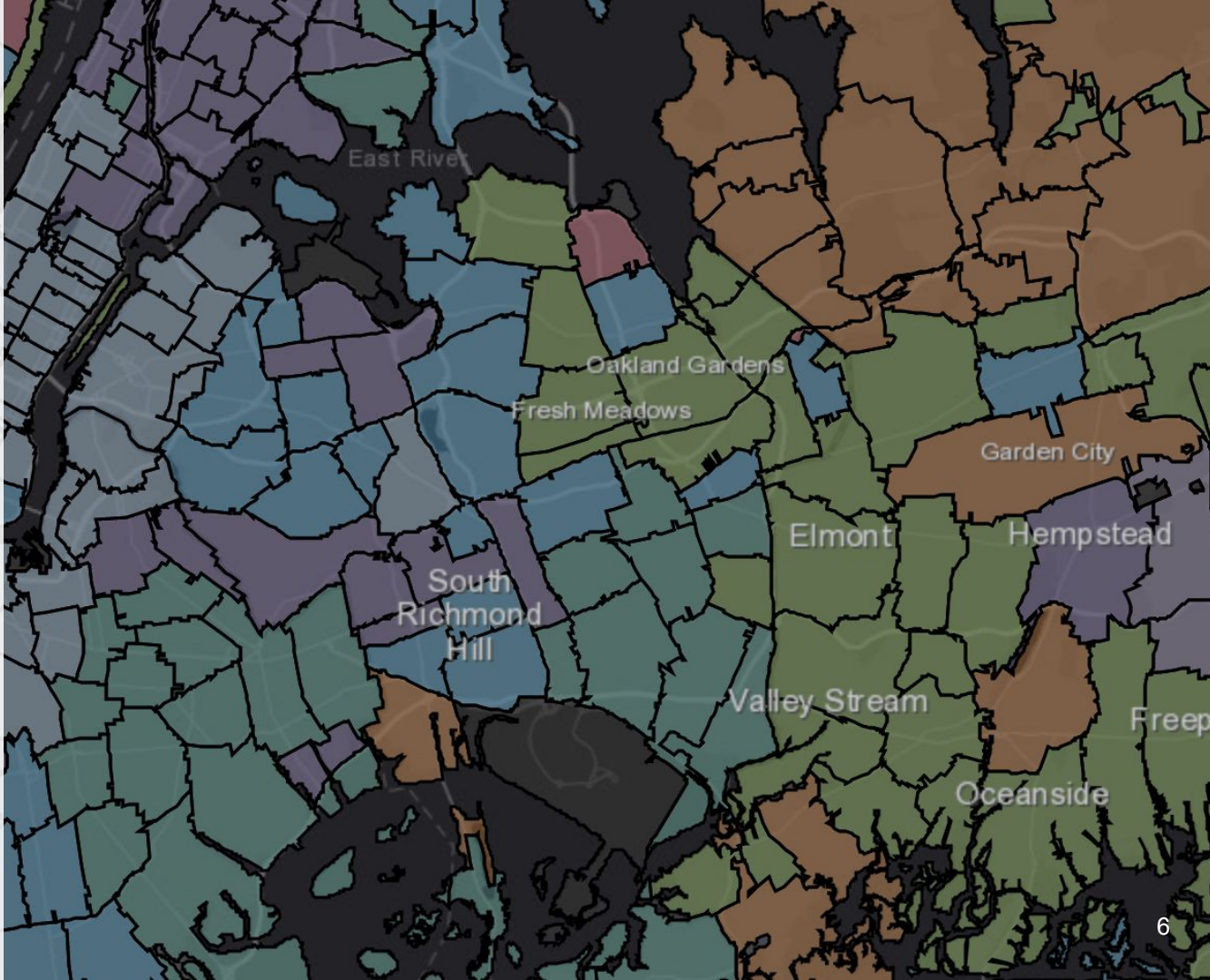
U.S. CENSUS BUREAU

Open Data DC

Methods

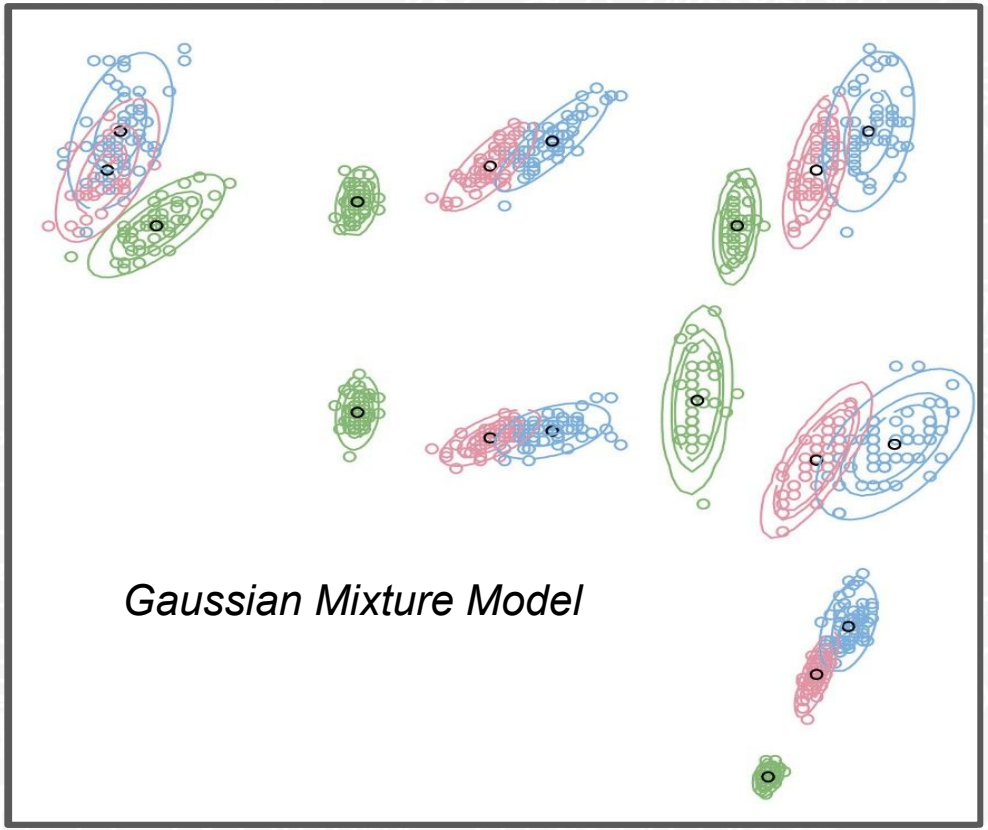
Methods:

- Geodemographic segmentation
- Clusters generated using Gaussian Mixture Model (GMM)



Gaussian Mixtures:

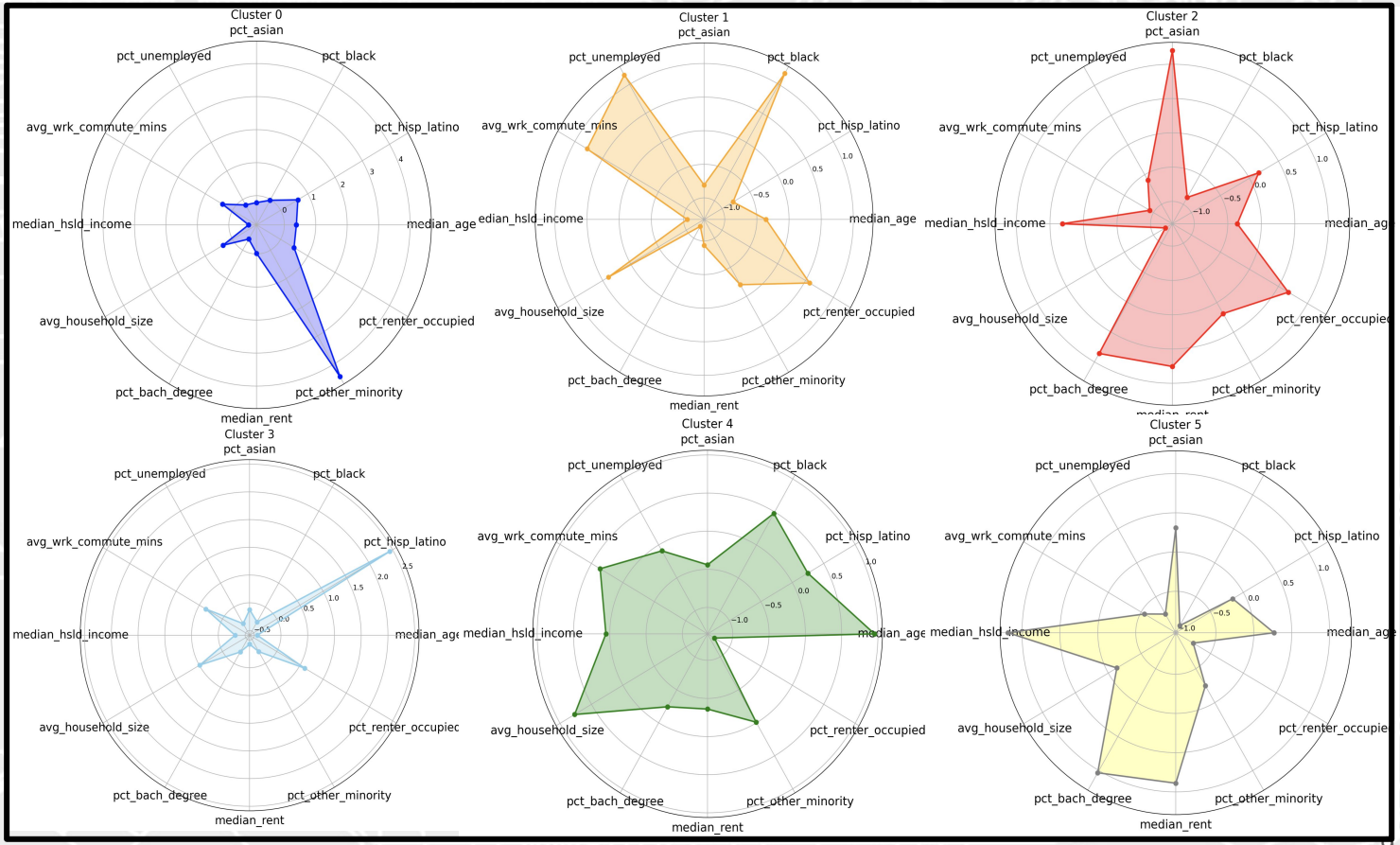
- Member of “soft” classifier algorithms
- Suited for representing subpopulations.
- Optimal Clusters = 6
- Silhouette Score = +.3



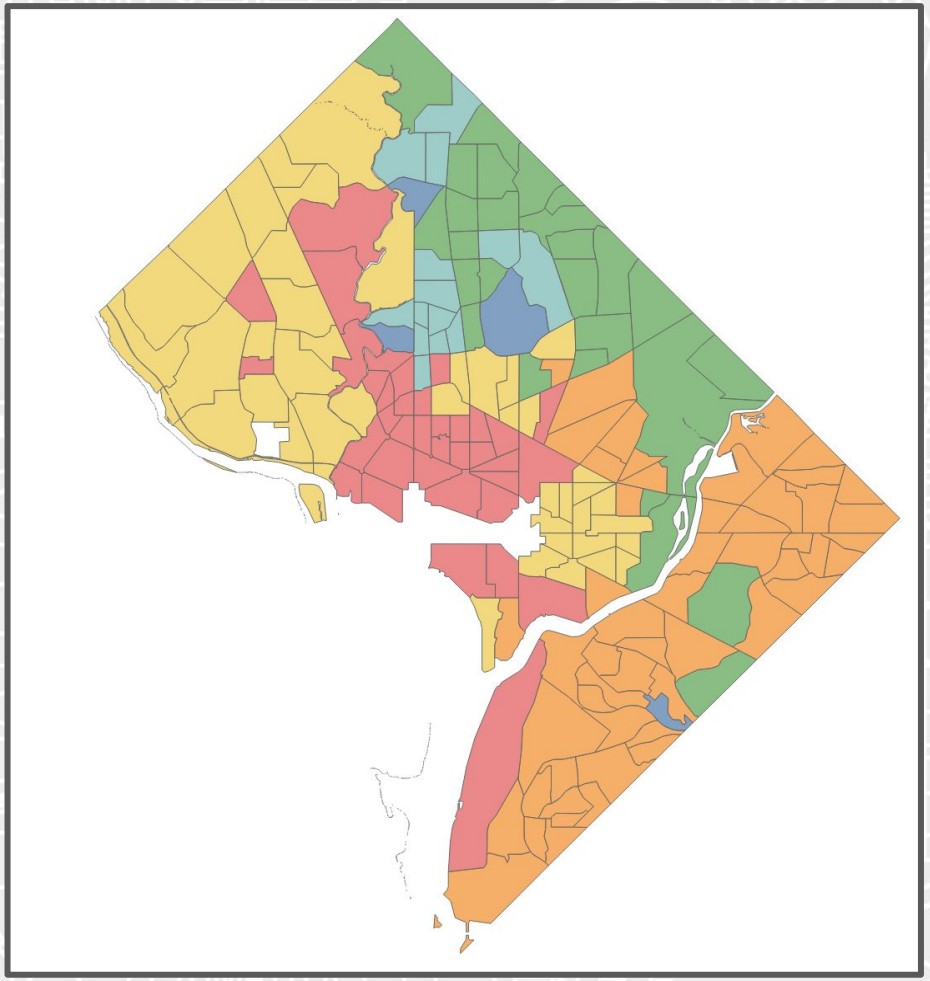
Gaussian Mixture Model

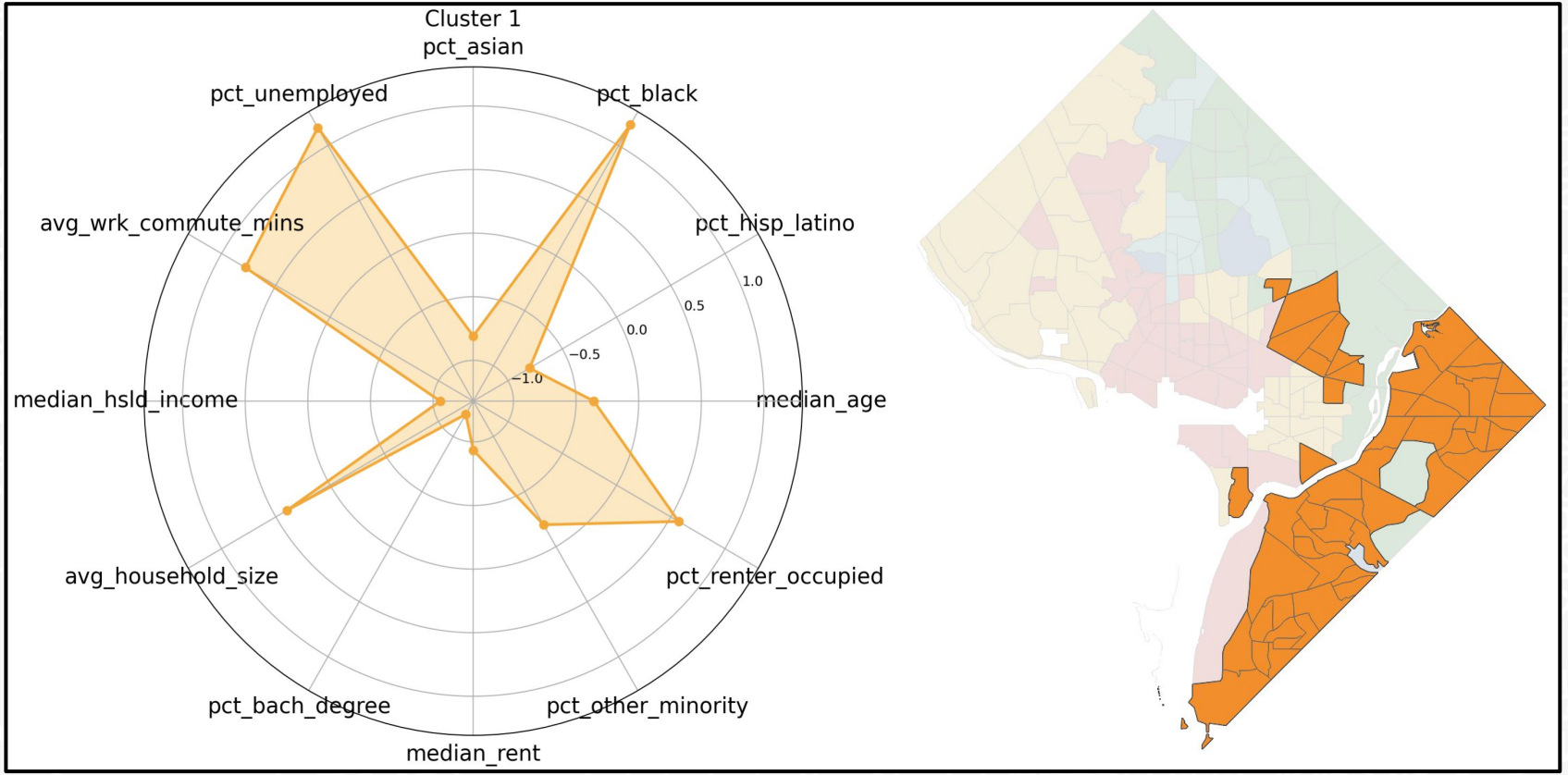
Results

Results:

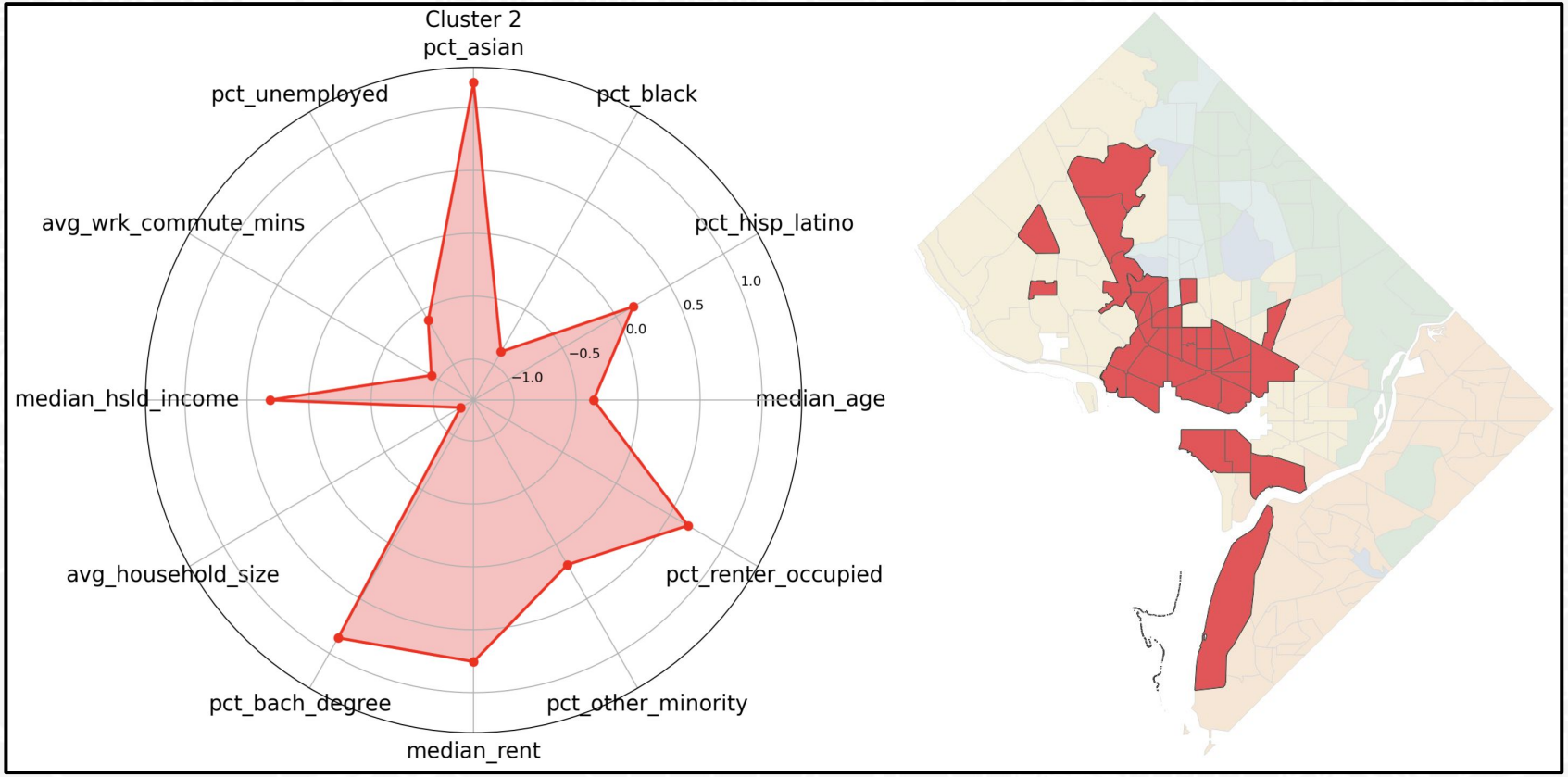

Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
pct_asian
pct_black
pct_hisp_latino
median_age
pct_renter_occupied
pct_other_minority
median_rent
pct_bach_degree
avg_household_size
median_hsld_income
avg_wrk_commute_mins
pct_unemployed

Clusters Mapped to Census Tracts:

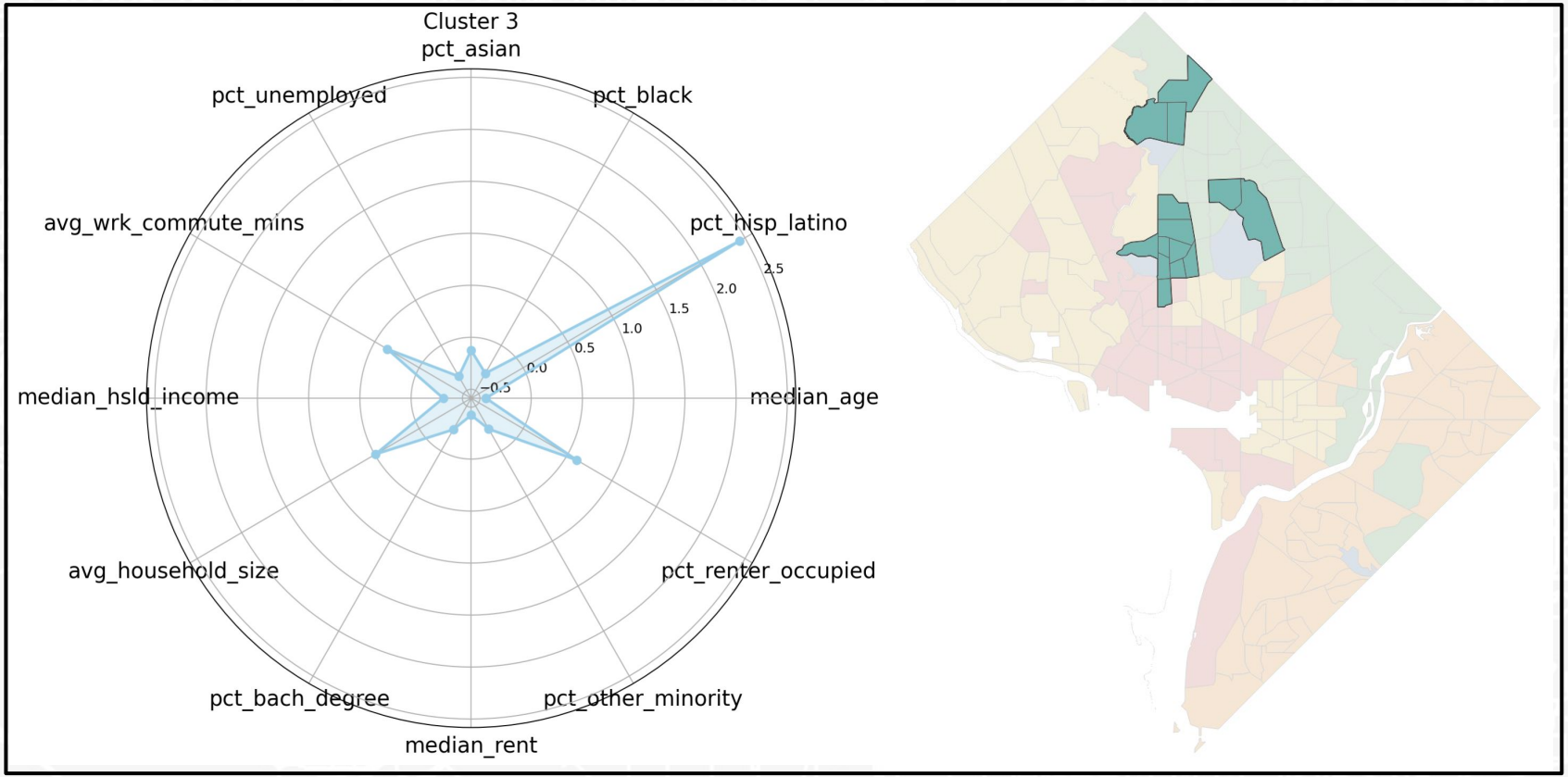



Cluster 1
pct_asian
pct_black
pct_hisp_latino
median_age
pct_renter_occupied
pct_other_minority
median_rent
pct_bach_degree
avg_household_size
median_hsld_income
avg_wrk_commute_mins
pct_unemployed
1.0
0.5
0.0
−0.5
−1.0

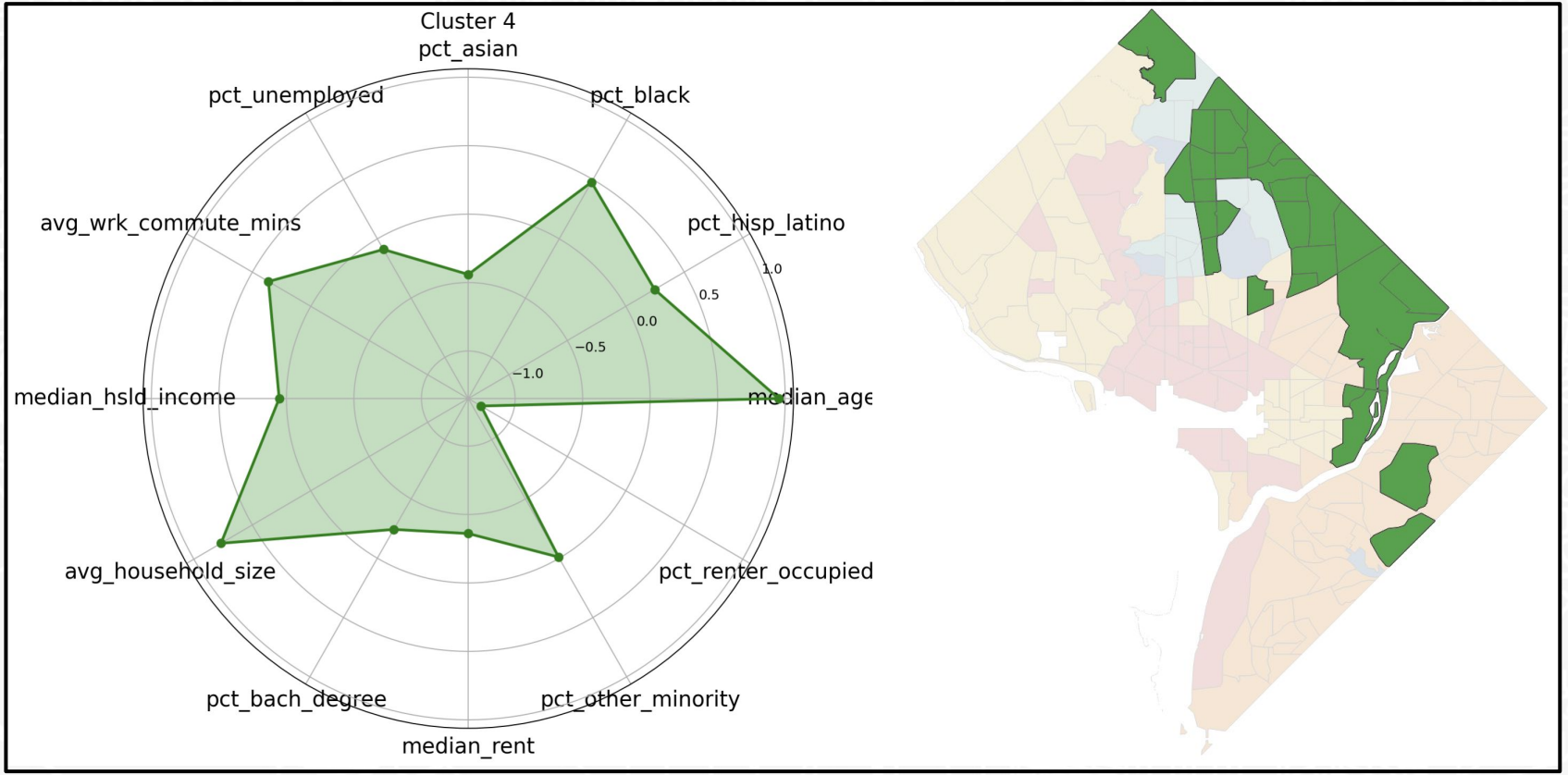
Cluster One


Cluster 2
pct_asian
pct_black
pct_hisp_latino
median_age
pct_renter_occupied
pct_other_minority
median_rent
pct_bach_degree
avg_household_size
median_hsld_income
avg_wrk_commute_mins
pct_unemployed
1.0
0.5
0.0
−0.5
−1.0

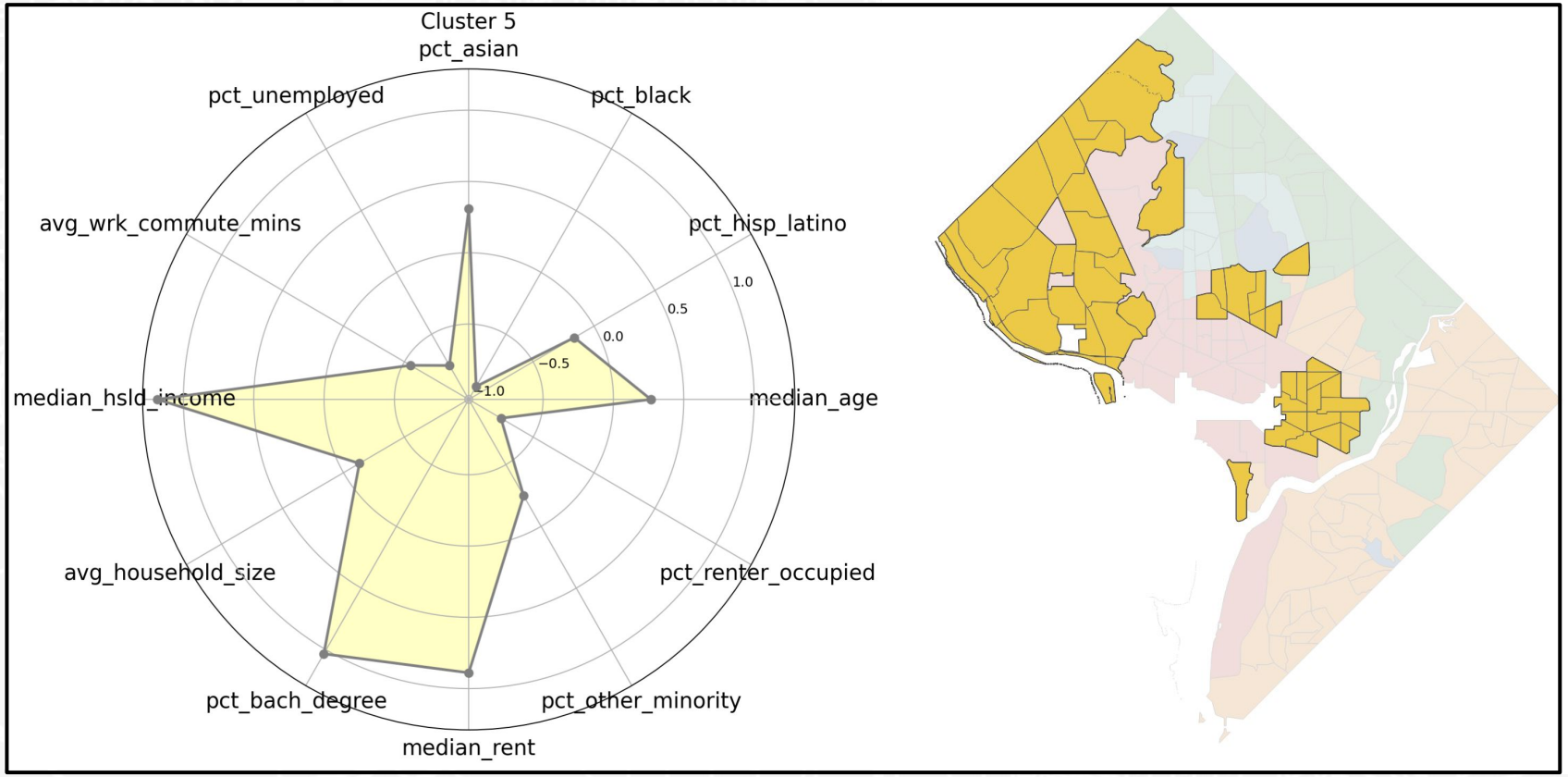
Cluster Two



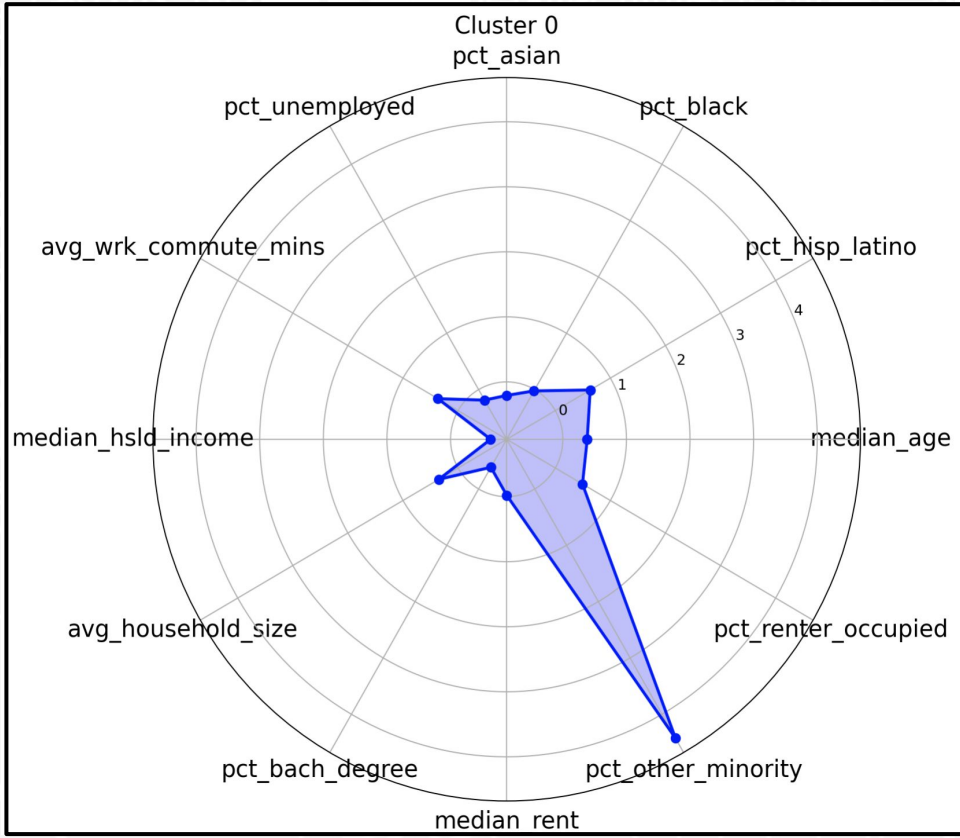
Cluster Three

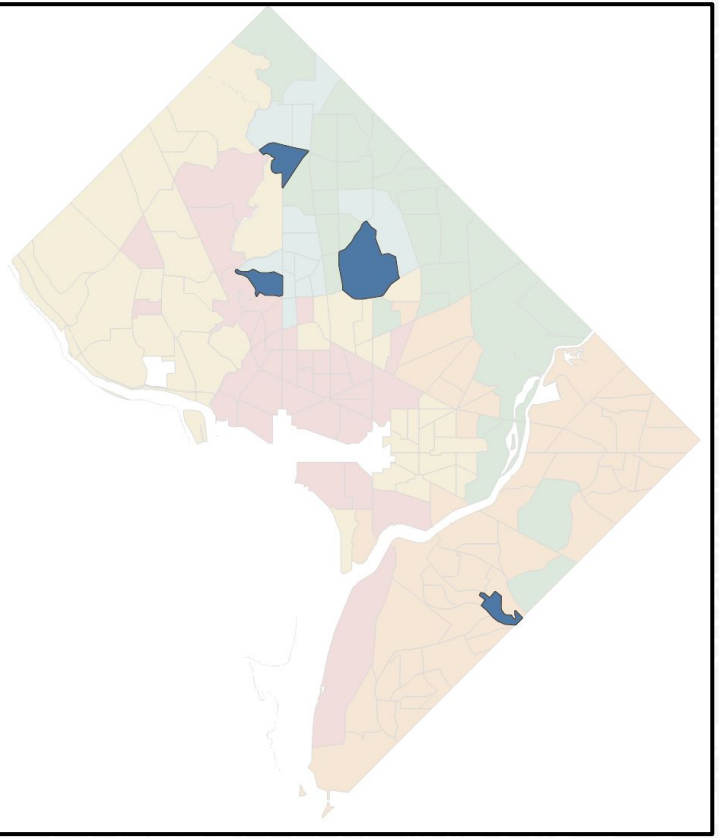


Cluster Four



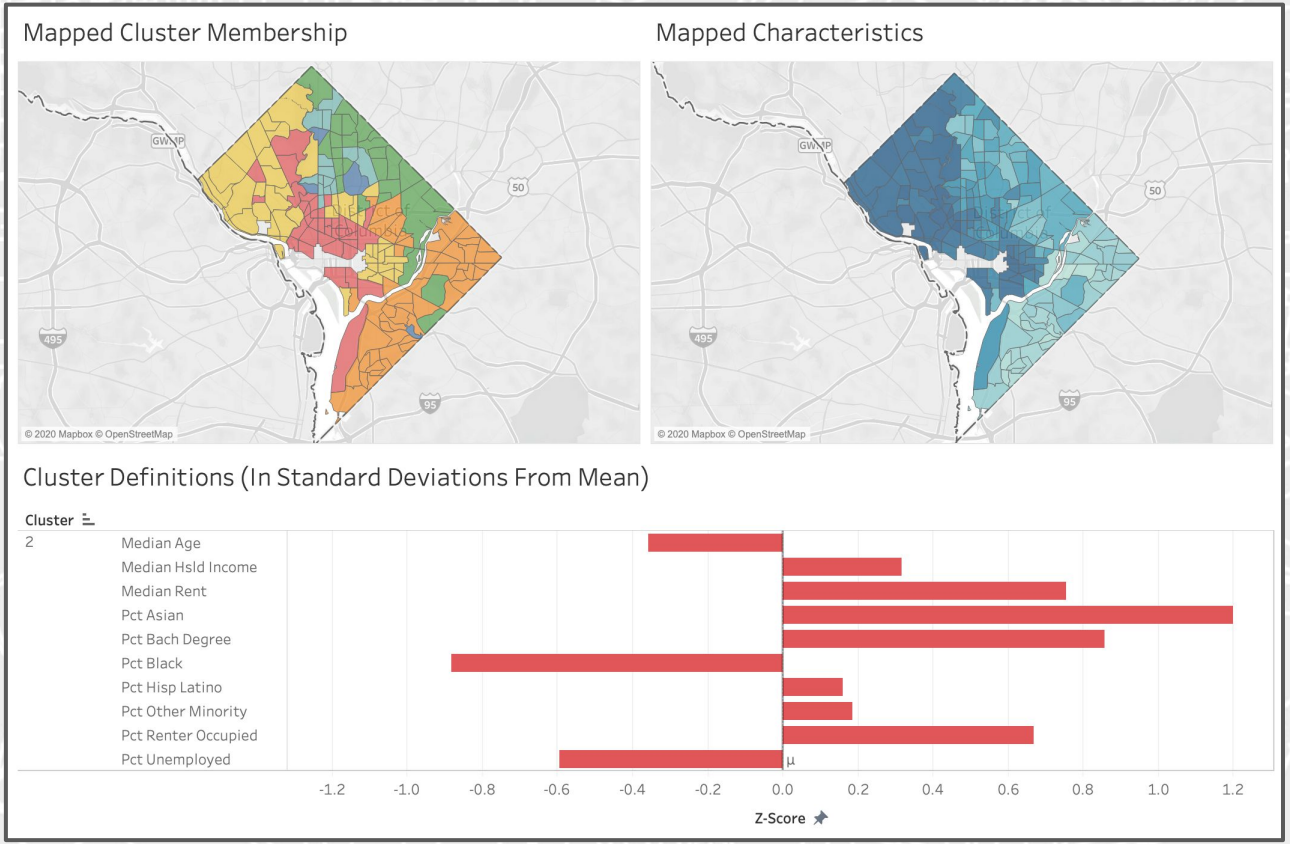
Cluster Five





Cluster Zero

Interactive Demo:



Conclusions and Recommendations

Conclusions and Recommendations:

1. At the census-tract level, DC is geographically segregated. This analysis offers a fairly bleak reality of DC.

2. The spatial outliers are worth further exploration.

3. Consider running dimensionality reduction on a broader range of variables.

4. Consider Hierarchical Clustering techniques.

5. Consider using a smaller unit of analysis, like 2020 Census Blocks (will be available late next year)

Questions