Unsupervised Analysis of Colorado's 2020 Census Data

Identifying Demographic Patterns through Clustering

August 30, 2024

Introduction

Problem Statement:

- What demographic patterns exist within Colorado based on the 2020 Census data?
- How can these patterns inform policy and resource allocation?

Data Overview

- Description: Overview of the 2020 US Census dataset.
- Key Features: Geographic Identifiers, Demographic Information, Housing Data, and more.
- Source: Kaggle cleaned version of US Census website.

High-Level Summary of the 2020 US Census Dataset

Overview:

The 2020 US Census dataset is an extensive and comprehensive dataset encompassing the 24th decennial United States Census. Released publicly on August 12, 2021, the dataset provides a detailed snapshot of the demographic, geographic, and social attributes of the US population across all 50 states. The dataset includes approximately 12 GB of data and covers a wide array of variables that are essential for analyzing population trends, demographic shifts, and geographic distributions.

Content:

- Geographic Identifiers: GEOID, GEOCODE, STATE (FIPS), COUNTY (FIPS), Census Tracts, and more.
- Demographic Information: Population counts by race, ethnicity, and combinations of racial groups.
- Housing Data: Housing unit counts, vacancy status, and types of housing structures.
- Legislative and Political Boundaries: Congressional districts, state legislative districts, and voting districts.
- Urban and Rural Designations: Urban areas, rural designations, and urban growth areas.
- Institutional Data: Counts of institutionalized populations, such as those in correctional facilities or nursing homes.
- Regional Classifications: Metropolitan and micropolitan areas, New England city and town areas, and regions.

Acknowledgements:

The dataset has been sourced from the US Census website and prepared for use on Kaggle by cleaning and organizing the data for accessibility, with no changes made to the original data.

Source:

https://www.kaggle.com/datasets/zusmani/us-census-2020

Methods

Machine Learning Approaches:

K-Means Clustering: Identifying distinct demographic groups.

PCA (Principal Component Analysis): Reducing dimensionality to visualize clusters.

t-SNE (t-distributed Stochastic Neighbor Embedding): Visualizing high-dimensional data in 2D.

Results

Clustering Analysis:

Visualization of clusters showing distinct demographic patterns across Colorado.

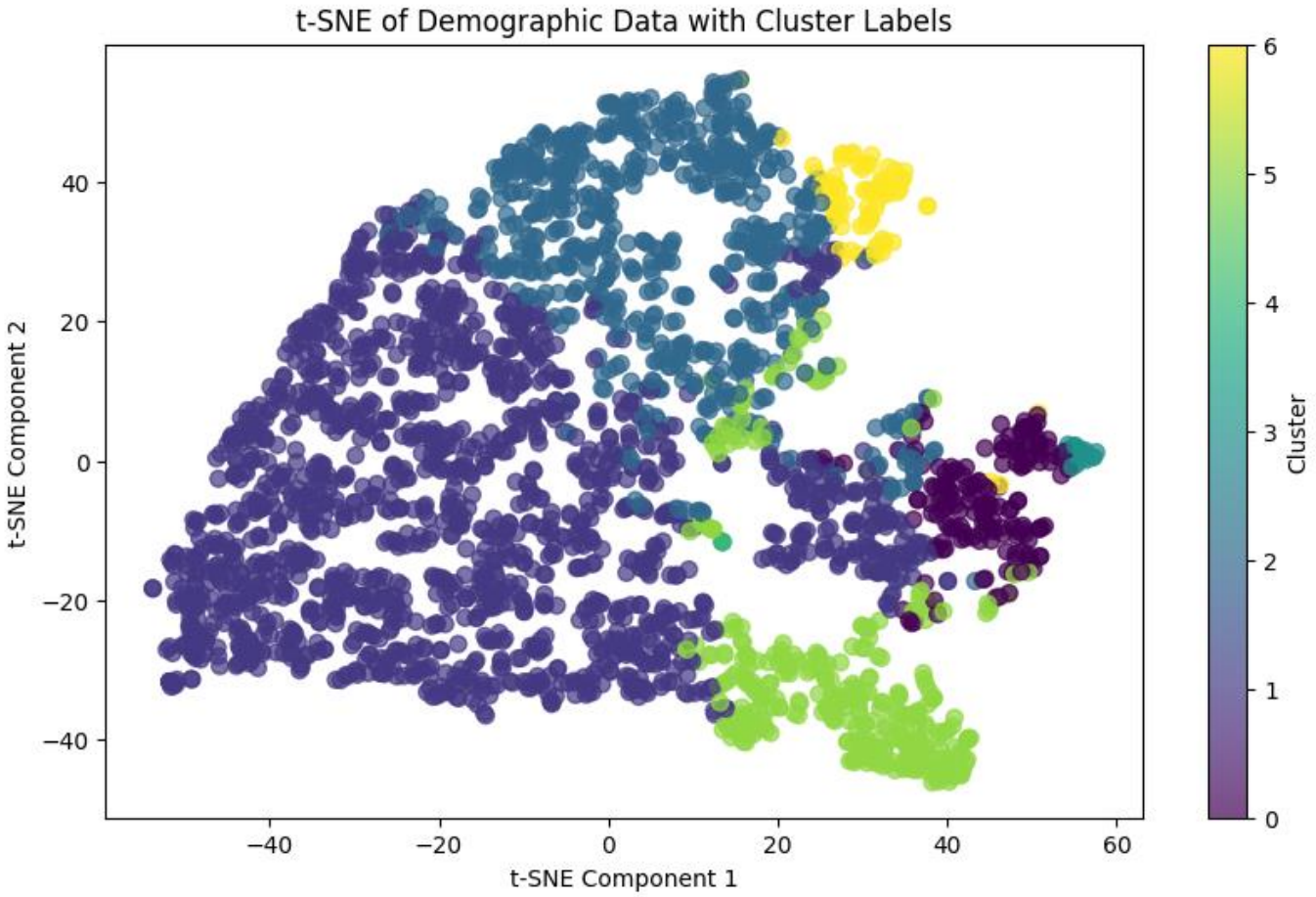
Interpretation of clusters: For example, clusters with a high concentration of Latino population vs. more diverse clusters.

Visualizations

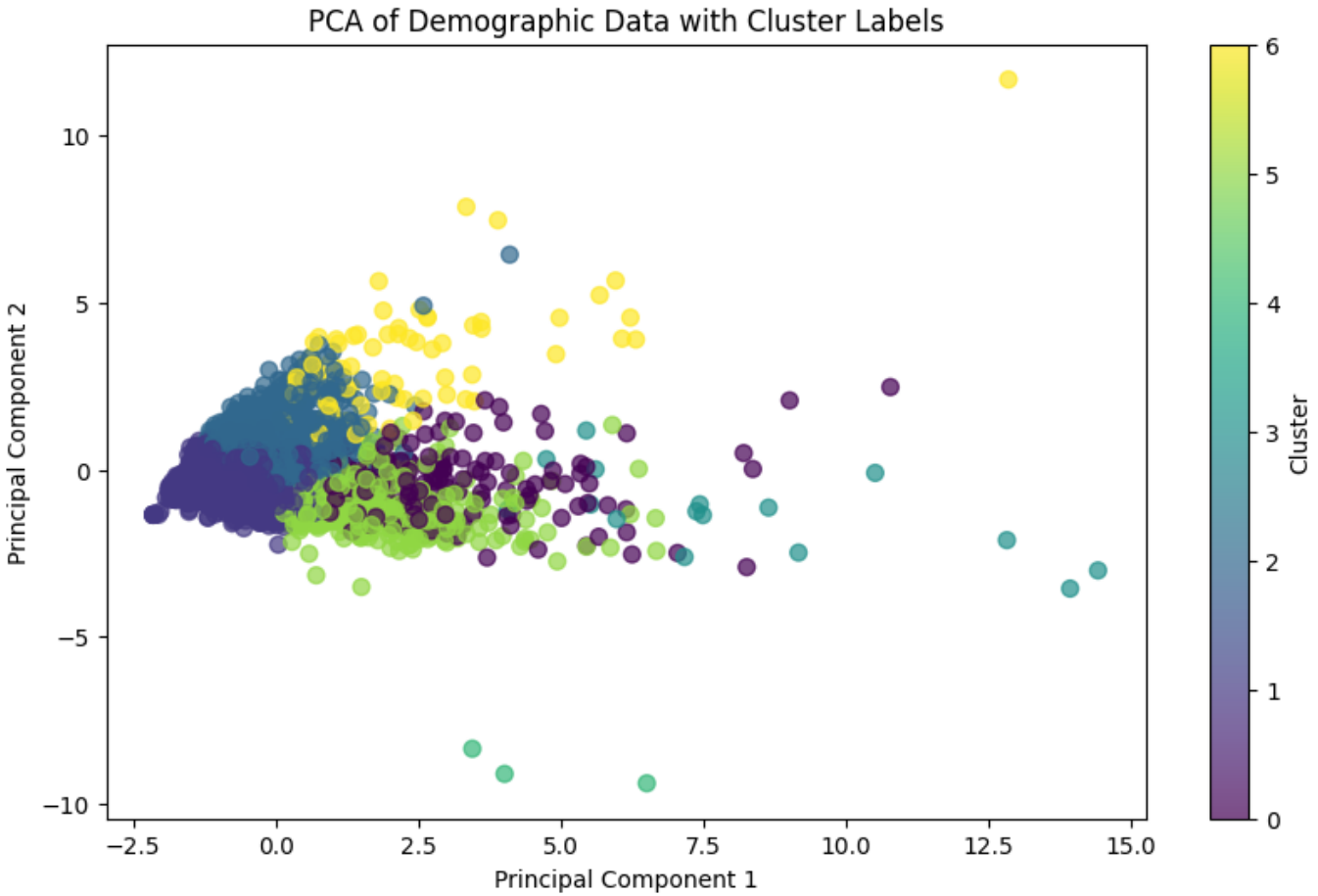
PCA and t-SNE Visualizations:

- Show how different clusters are well-separated or overlapping.
- Discuss the implications of these visualizations.

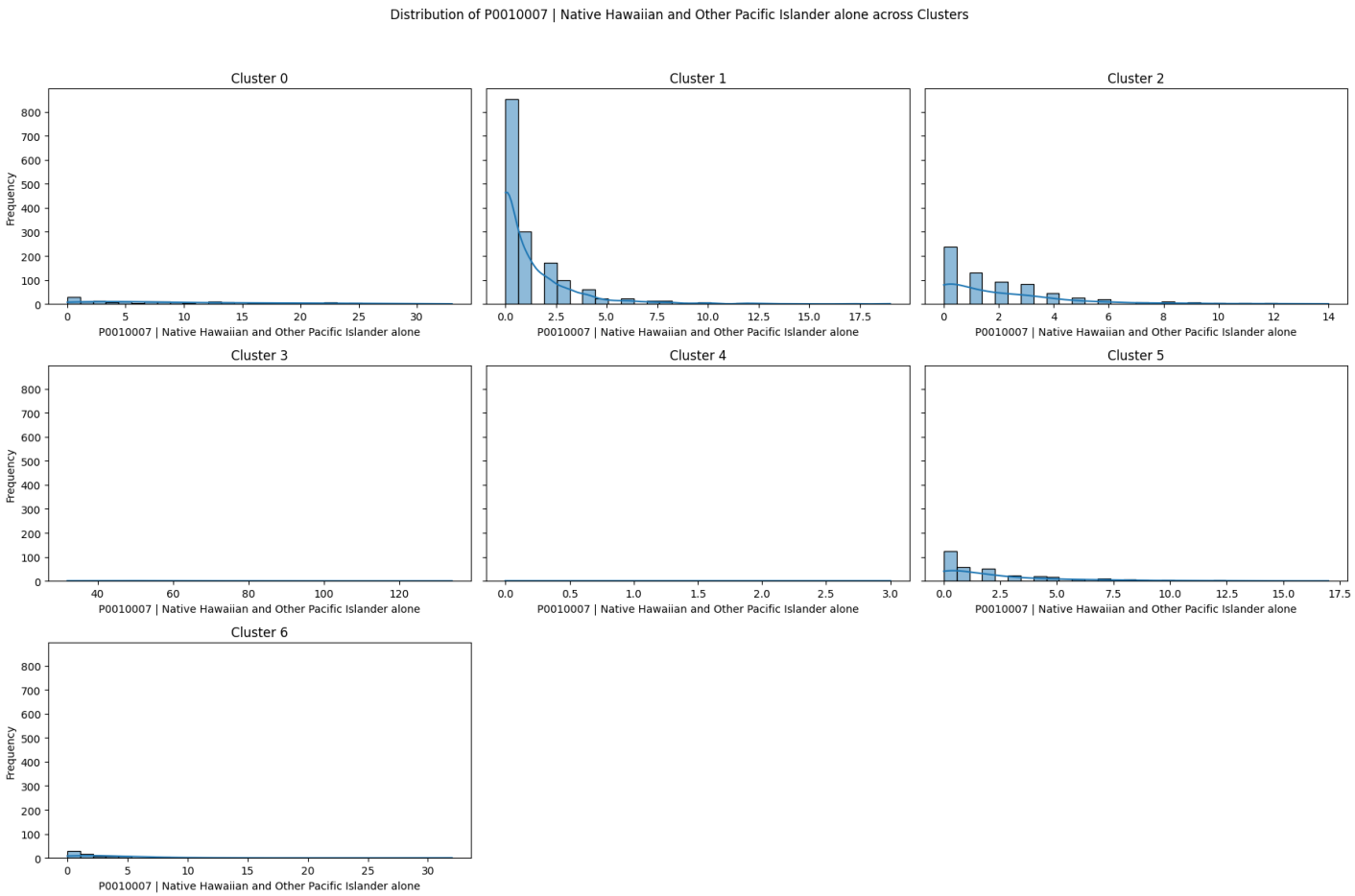
t-SNE, Colorado, 2020 Census



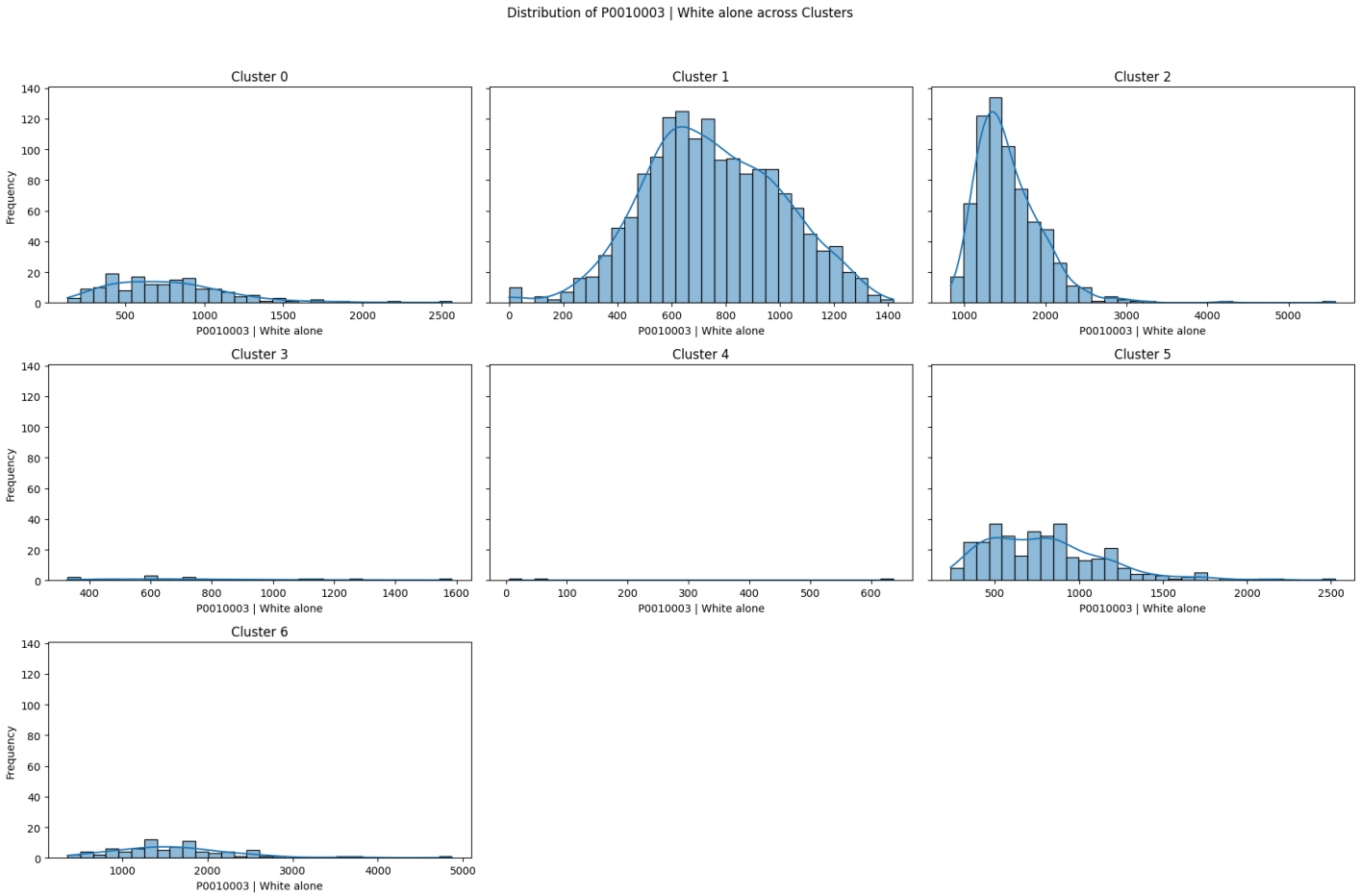
PCA, Colorado, 2020 Census



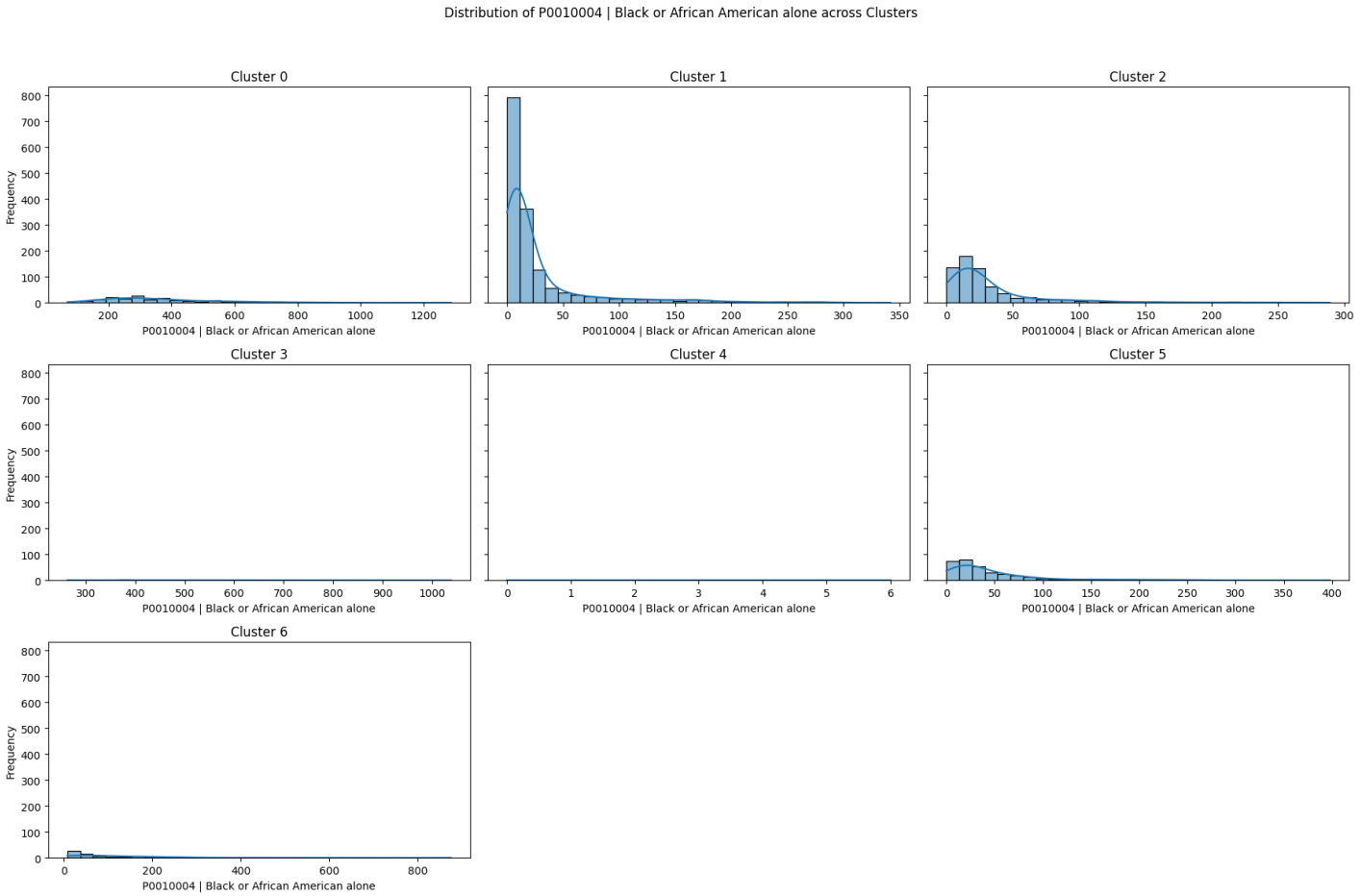
Native Hawaiian and Other Pacific Islander alone - across Clusters



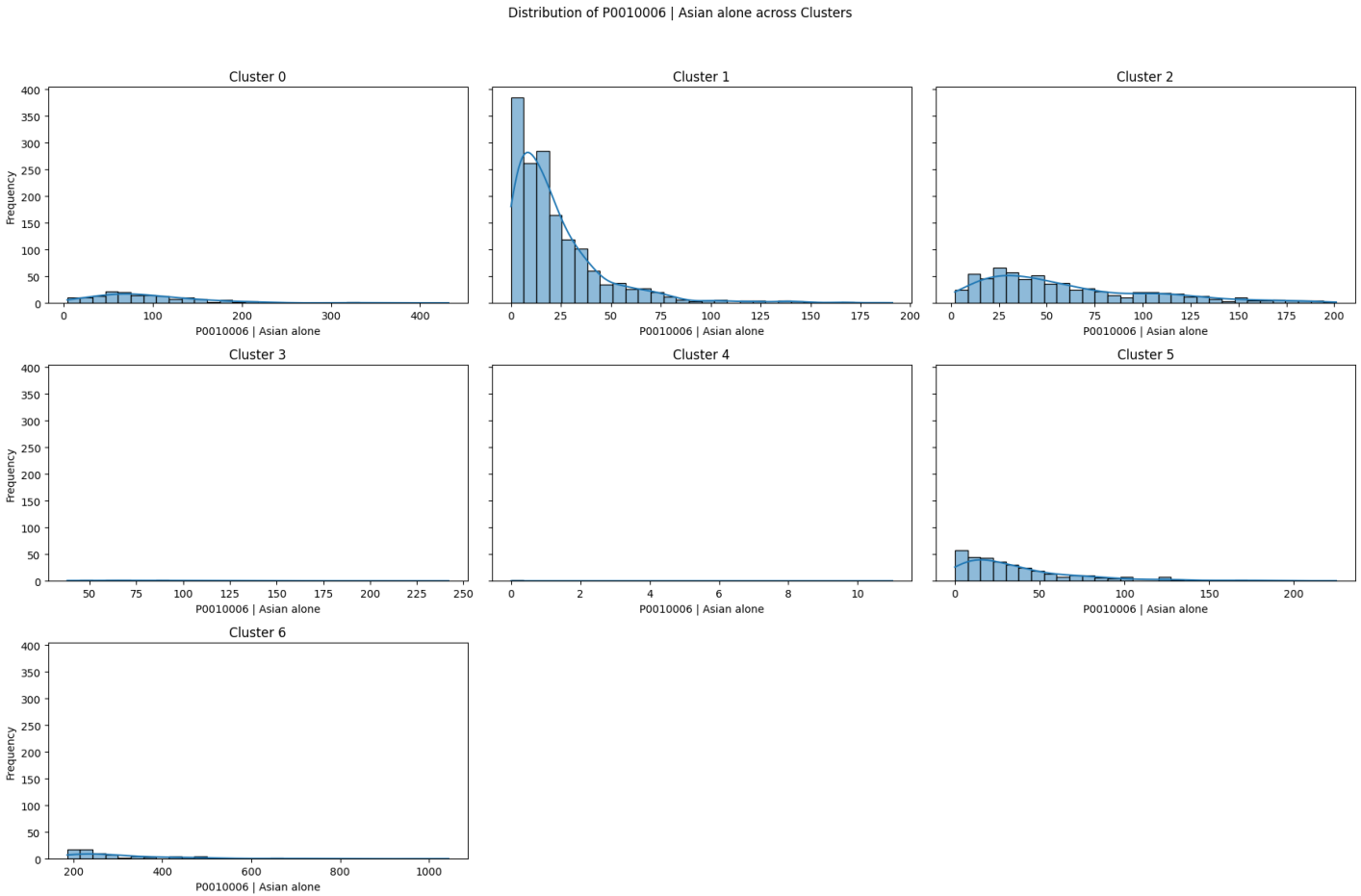
White alone - across Clusters



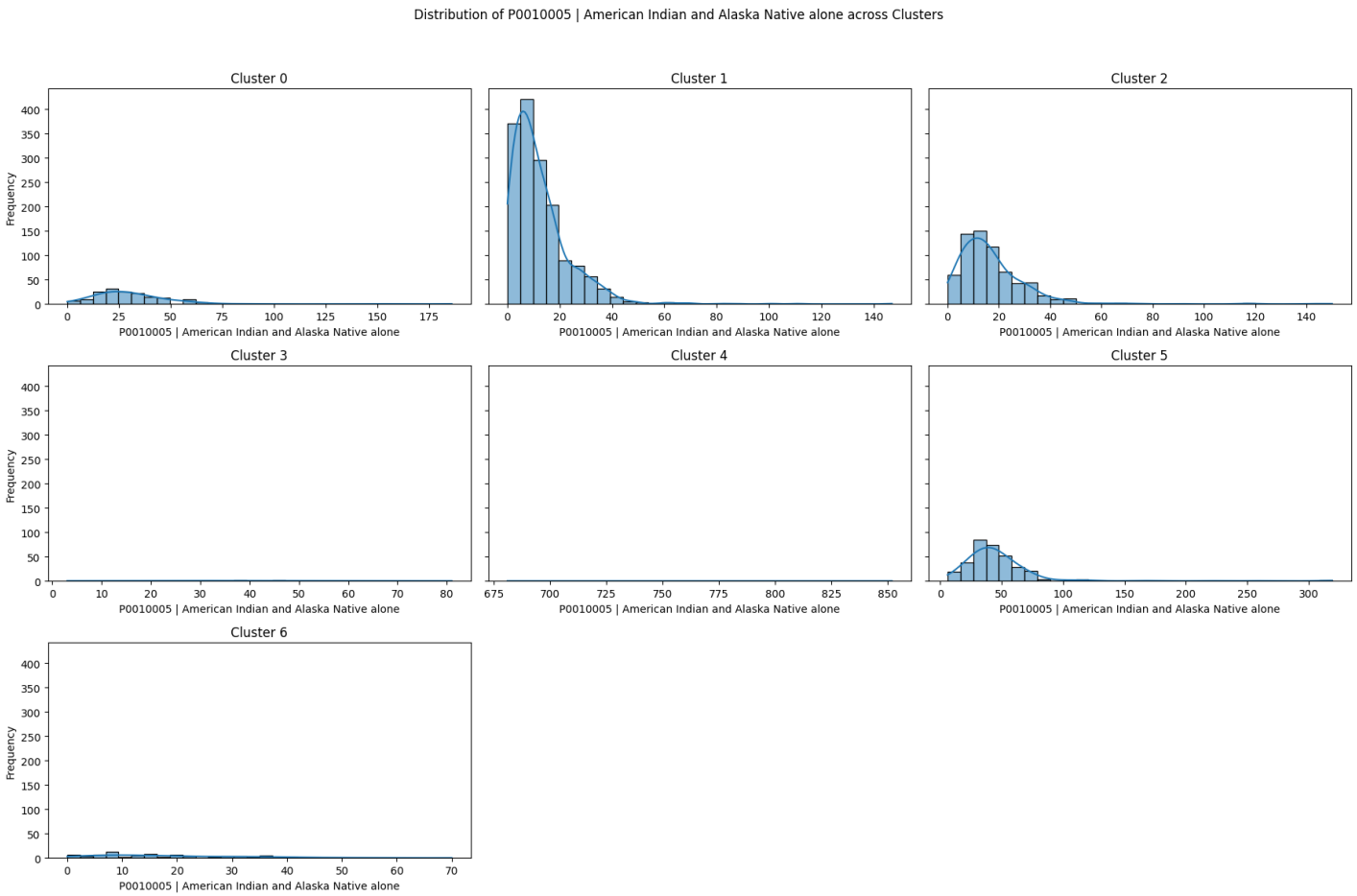
Black or African American alone - across Clusters



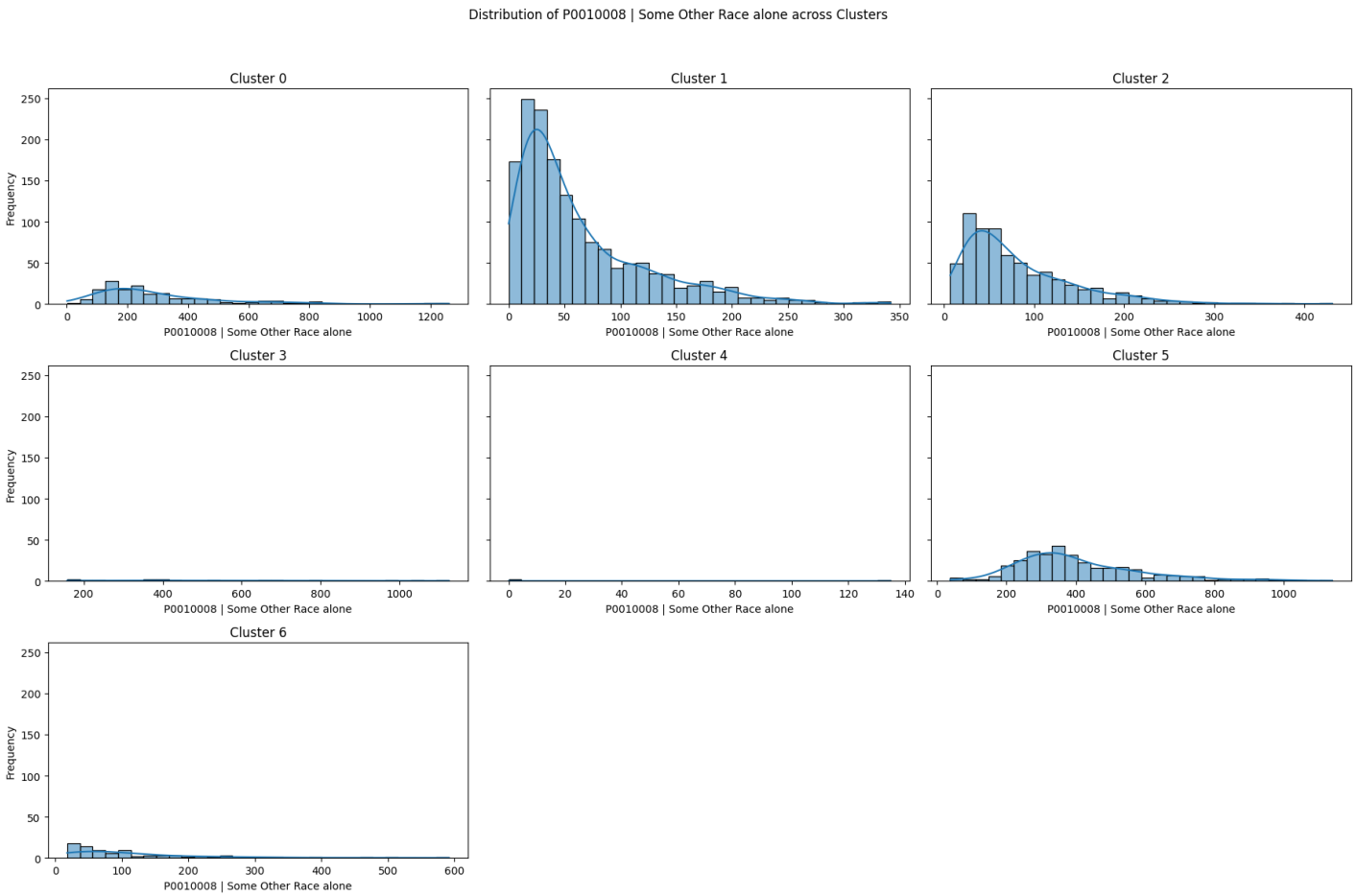
Asian alone - across Clusters



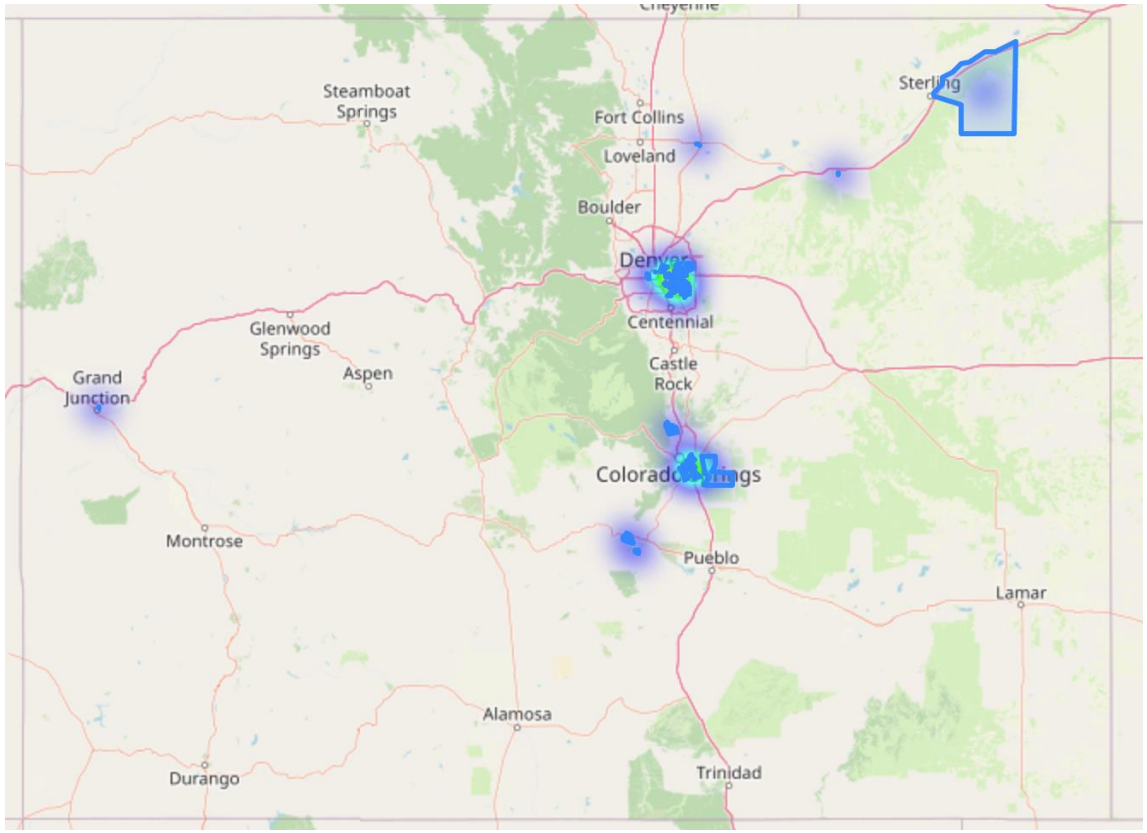
American Indian and Alaska Native alone - across Clusters



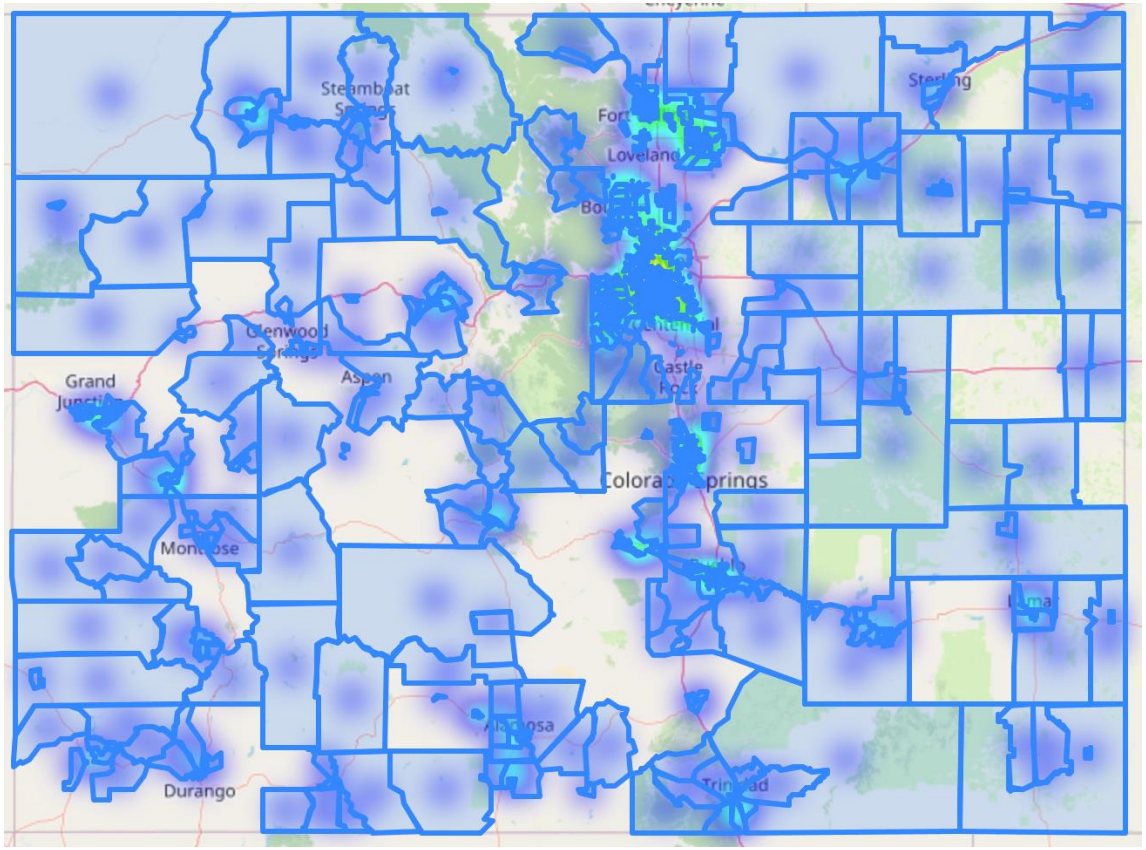
Some Other Race alone - across Clusters



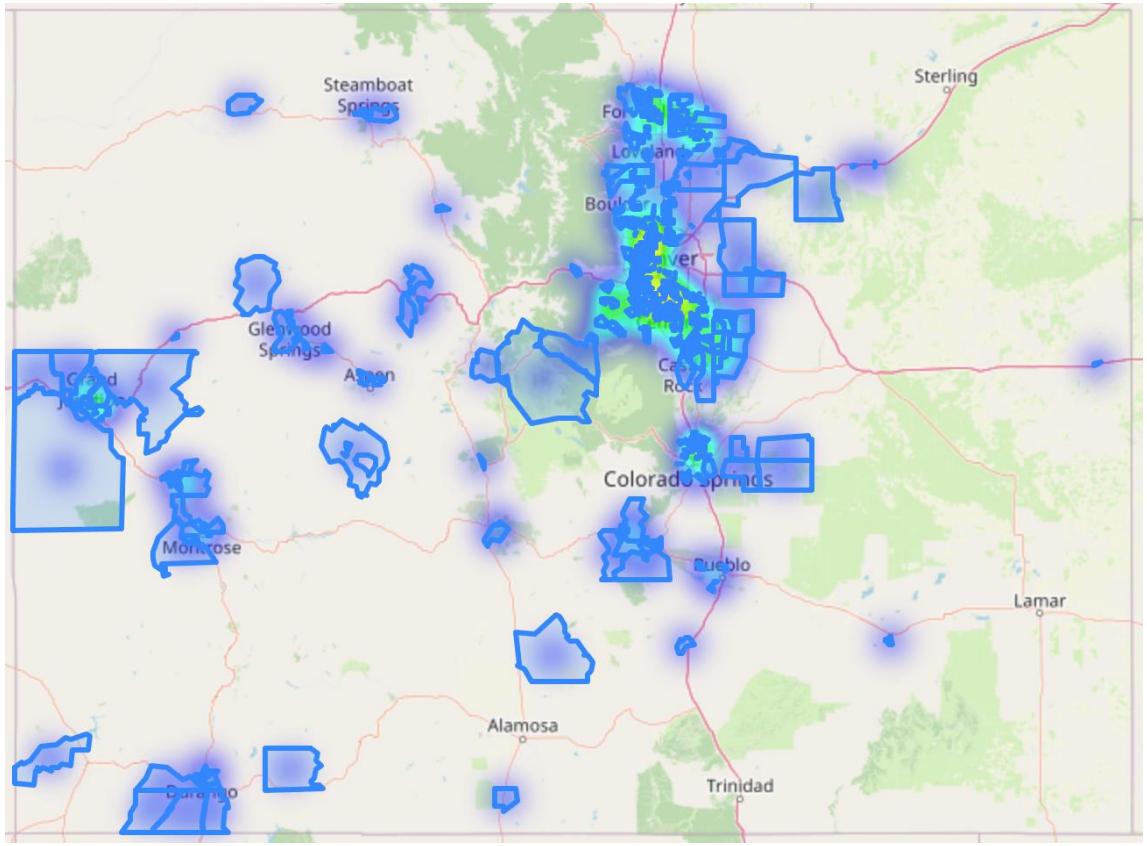
Cluster 0 – geospatial view



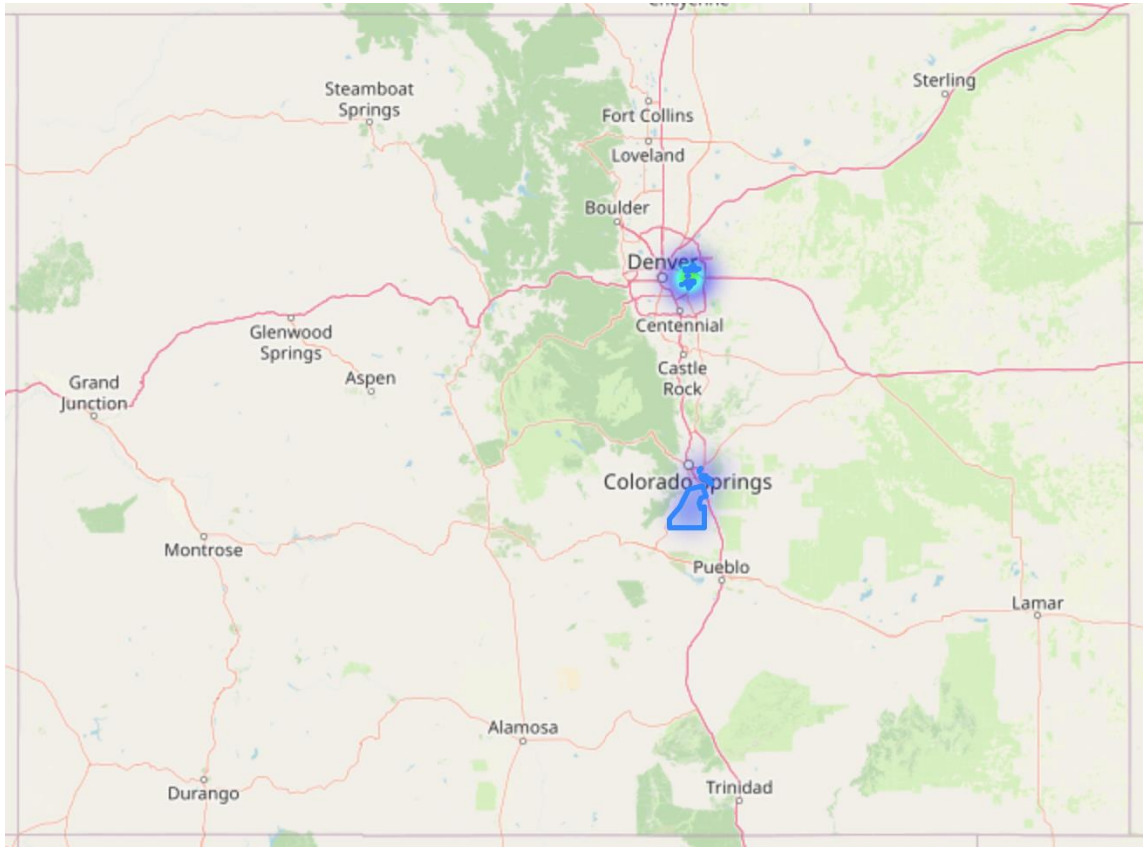
Cluster 1 – geospatial view



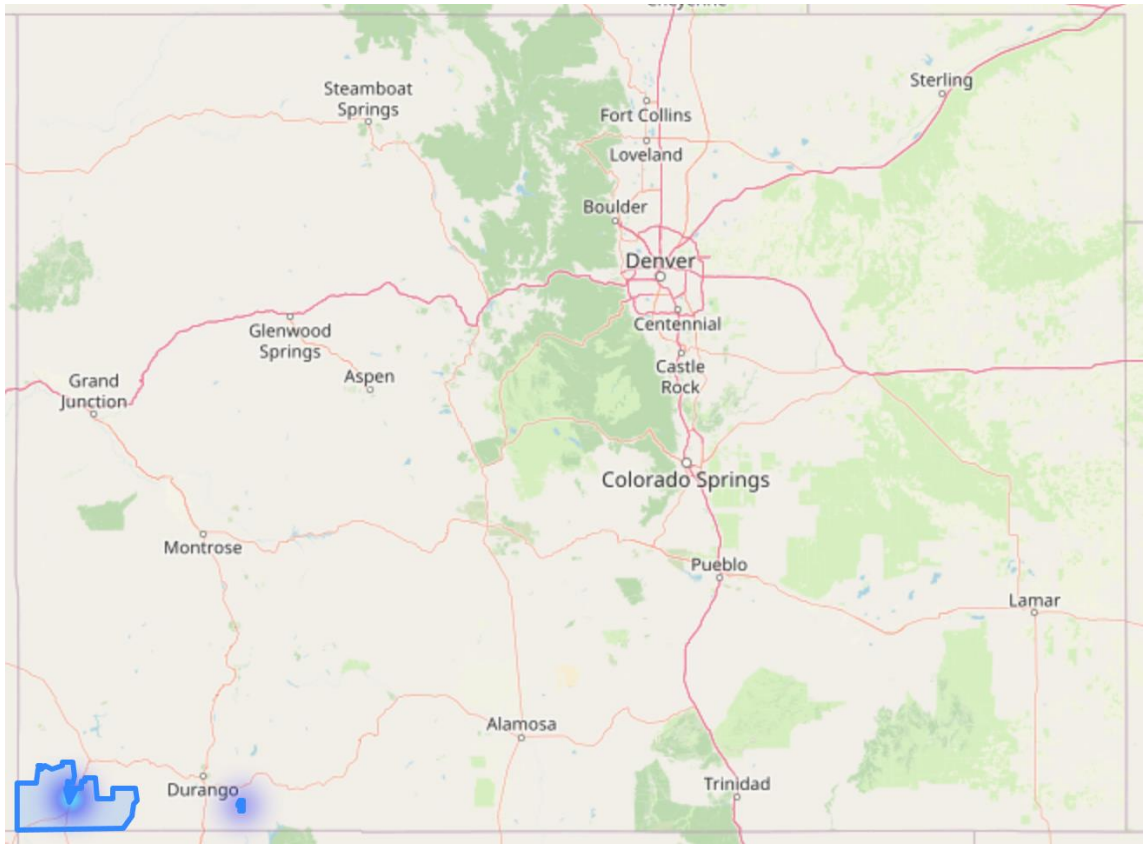
Cluster 2 – geospatial view



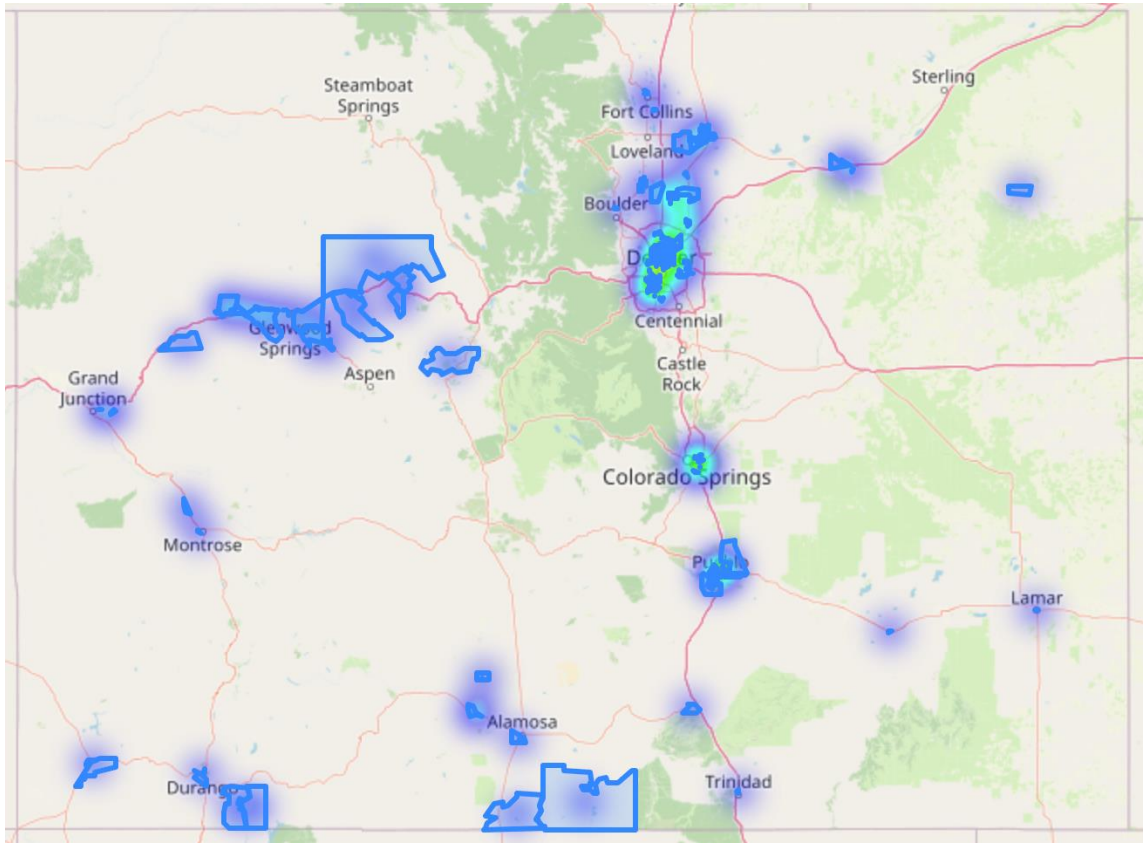
Cluster 3 – geospatial view



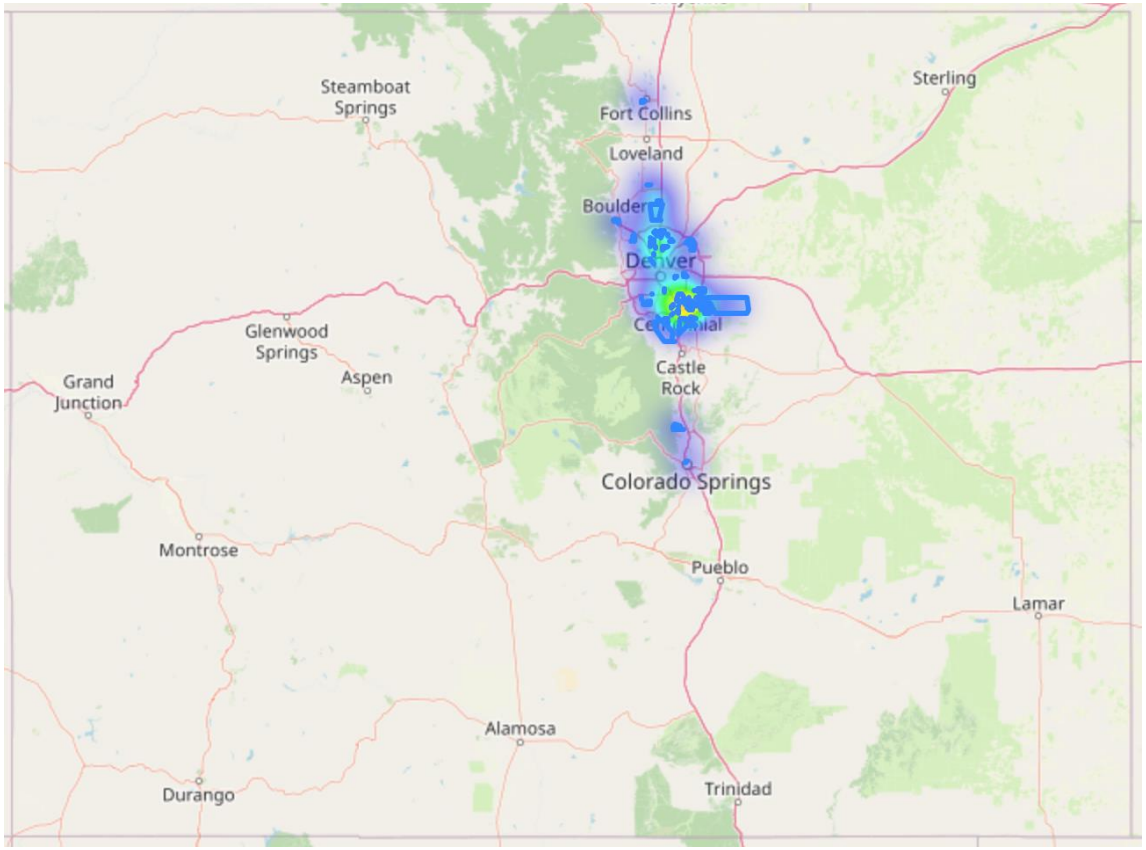
Cluster 4 – geospatial view



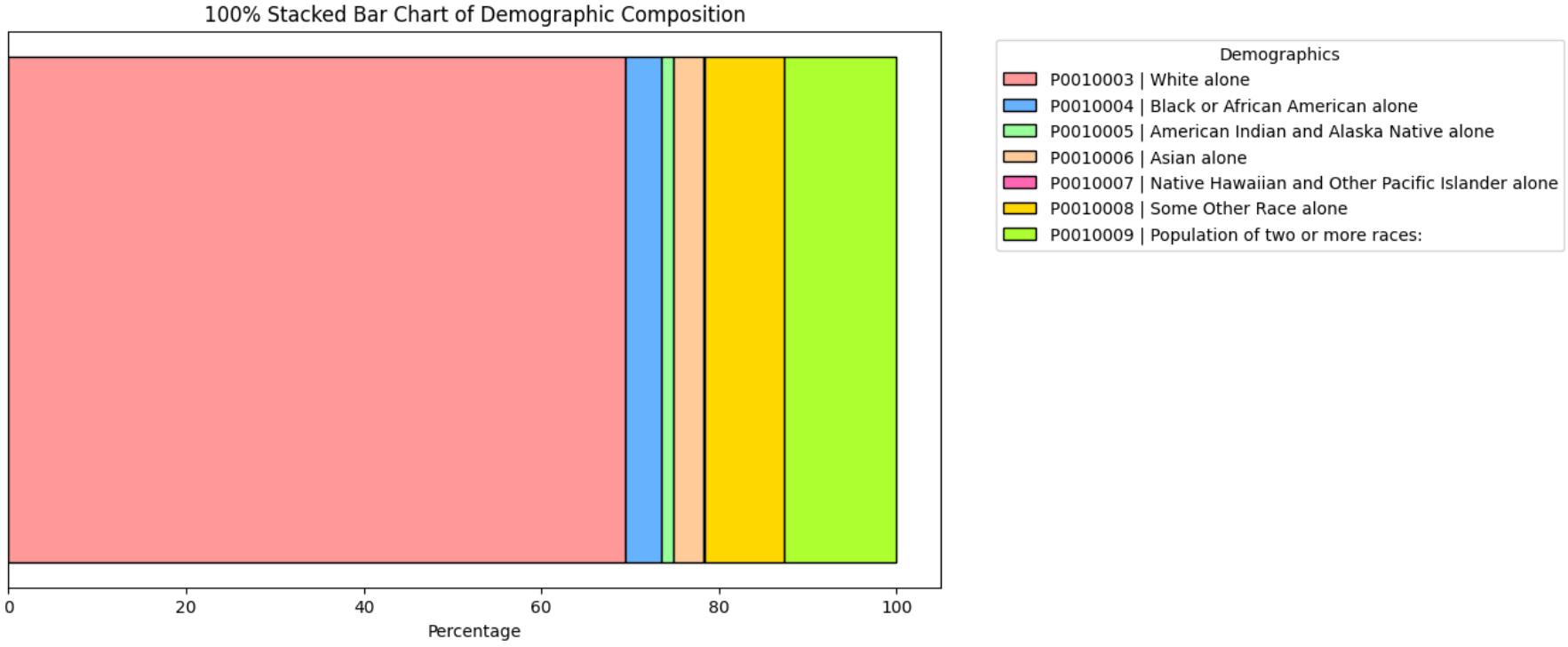
Cluster 5 – geospatial view



Cluster 6 – geospatial view



Record count - % of data



Conclusion

- **Key Findings:**
  - Identification of clear demographic segments within Colorado.
  - Potential applications: Targeted policy interventions, community planning, business strategy.

Future Work

Next Steps:

Explore additional clustering methods like DBSCAN or Gaussian Mixture Models.

Incorporate more socio-economic variables to refine analysis.

Thank You

Thank you for your attention!

Questions?

- **Public GitHub Link:**

https://github.com/LEBLAPI1/ColoradoCensus2020_ClusteringAnalysis

- **Google Colab Link:**

https://drive.google.com/file/d/1rtdCJG4GQ9eRJipRS2S9g2SbbNrFa-Jn