Final Project Write-Up

*\*Note: Much of the information in the write-up is already available on the project website\**

**Introduction**

This project aims to guide policymakers in identifying areas of Pennsylvania where resources can be allocated more equitably and effectively. By examining demographic, economic, and environmental data, the goal is to uncover patterns and clusters of counties that highlight disparities and areas of need. The results of this study will provide evidence-based insights to inform resource allocation and policy interventions, particularly in addressing poverty, unemployment, and access to environmental resources like parks.

**Research Questions**

1. Are demographic, economic, and environmental factors spatially correlated across the state of Pennsylvania?
2. How can the identification of similar county clusters help policymakers equitably allocate resources?

**Data**

To answer these research questions, we integrated data from several sources:

1. [American Community Survey](https://www.census.gov/programs-surveys/acs) (ACS): Using the Census Bureau’s ACS 5-year estimates API, we gathered key demographic and economic indicators for all counties in Pennsylvania. These include:
   1. Total population
   2. Median household income
   3. Per capita income
   4. Educational attainment (high school graduates, bachelor’s degree holders)
   5. Poverty rate
   6. Median housing value
   7. Racial demographics (White, Black, Hispanic populations)
   8. Geospatial Data: We obtained county boundary geometries for Pennsylvania from a GeoJSON file and reprojected the data to Pennsylvania State Plane (EPSG: 2272) for spatial analysis.
2. [Park Data](https://www.pasda.psu.edu/uci/DataSummary.aspx?dataset=307): Data on local park coverage by county was sourced from the Pennsylvania Department of Conservation and Natural Resources (DCNR). Using park acreage, we calculated the percentage of local park area relative to each county’s total area.
3. [Unemployment Data](https://hdpulse.nimhd.nih.gov/data-portal/social/table?age=916&age_options=age16_1&demo=00012&demo_options=workforce_2&race=00&race_options=raceall_1&sex=0&sex_options=sex_3&socialtopic=090&socialtopic_options=social_6&statefips=42&statefips_options=area_states&function=getTable&path=social): Unemployment rates were extracted from the National Institute on Minority Health and Health Disparities.

**Data Preparation and Cleaning**

Merging Datasets: The ACS data was merged with geospatial county boundaries using unique identifiers. Additional park and unemployment data were integrated using county names. Derived Variables: Several new variables were calculated to enhance the analysis:

1. Percent White, Black, Hispanic populations: Calculated as proportions of the total population.
2. Educational attainment percentages: Calculated for both high school graduates and bachelor’s degree holders.
3. Percent Local Park: Calculated as the proportion of county area covered by local parks.
4. Handling Missing Values: Missing values were handled by imputing medians or zeros for specific variables to ensure compatibility for clustering and other analyses.

**Analysis Approach**

Correlation Analysis: A correlation matrix was generated to explore relationships between demographic, economic, and environmental variables.

Spatial Visualization: Maps were created to visualize spatial patterns in key variables like unemployment rates, poverty levels, and local park coverage.

Clustering Analysis: Using K-means clustering, counties were grouped based on socioeconomic, demographic, and environmental characteristics. This helped identify clusters of counties with similar needs or advantages.

Socioeconomic Index: A composite index was calculated combining median household income, bachelor’s degree attainment, and poverty rate to rank counties based on overall socioeconomic well-being.

**Goal and Application**

The results of this analysis will:

1. Highlight disparities across Pennsylvania counties in income, education, environmental access, and unemployment.
2. Identify clusters of counties with shared characteristics to allow targeted policymaking.
3. Serve as a resource for state and local governments to allocate resources more equitably, focusing on underserved communities.

**General Conclusions**

* Cluster 1 was made up of poorer, rural counties that struggled with unemployment and access to public green spaces.
* Cluster 2 was concentrated in the southeastern portion of the state, nearby Philadelphia, and was made up of economically thriving counties with a high education attainment level.
* Cluster 3 was Philadelphia County alone, known for its low household income and limited educational attainment.
* Finally, Cluster 4 was made up of counties that were moderately rural and had a moderate socio-economic status, with slightly above average household income and moderate educational attainment.
* This indicated that Clusters 1 and 4 should be prioritized for state and federal resources to address entrenched economic disparities.
* Parks in Cluster 3, while abundant, should be maintained through state support to ensure accessibility and functionality.
* Further, thriving counties in Cluster 2 should be used as a model for economic development and park infrastructure improvements in struggling areas.
* Finally, ensure that policies for struggling clusters account for the geography of the area. For example, Cluster 1 is struggling and is rural while Cluster 4 is a high-poverty area but urban.
* Final conclusions:
  + Clusters 1 and 4 are clear areas of need, with high poverty and unemployment rates, calling for targeted interventions. However, one cluster is rural while the other is urban.
  + Clusters 2 and 3 are more stable but would benefit from investments in green infrastructure and educational initiatives to maintain and enhance their resilience.