**CEJST Project Report**

**Kassidi Knight**  
**Data 499: Data Science Capstone (Fall 2024)**

## Interactive Dashboard Link

<https://public.tableau.com/app/profile/kassidi.knight/viz/CEJST_Dashboard/Finalv1Dashboard?publish=yes>

## Website Link

<https://kassidimariah.github.io/Kassidis_Quality_Of_Life_Heatmaps/>

## Abstract

This project examines the Climate and Economic Justice Screening Tool (CEJST) dataset to analyze environmental and demographic quality-of-life metrics across the United States. By leveraging Python libraries like GeoPandas and Folium, the project maps trends at the census tract and county levels. Unique aggregations and visualizations are employed to uncover disparities, emphasizing their significance for policy-making and resource allocation in underprivileged areas.

## Introduction

Environmental justice is critical in addressing the intersection of ecological degradation and social inequality. This project seeks to analyze and visualize CEJST metrics, focusing on environmental burdens (e.g., flood risks, PM2.5 levels) and socioeconomic challenges (e.g., poverty levels). The insights aim to support policymakers, researchers, and activists in identifying areas needing intervention. Unlike static reporting, this work combines data aggregation and geospatial mapping for dynamic and interactive insights, setting it apart from traditional studies.

## Previous Work

Previous research on environmental justice has utilized datasets like the CEJST to analyze demographic disparities. For example, studies have highlighted the disproportionate environmental burdens faced by communities of color. This project especially focused on expanding the work done by the Climate and Economic Justice Screening Tool (CEJST) produced by the federal government. While existing work often employs descriptive statistics, this project integrates geospatial analysis to provide interactive, visual insights. Additionally, by using Python and open-source tools, this project enables reproducibility and flexibility for future work.

## Dataset

The CEJST dataset includes various environmental and demographic metrics at the census tract level, aggregated to counties for this project. Key attributes analyzed include:

* **Environmental**: Flood risk, PM2.5 air quality levels, energy burden.
* **Socioeconomic**: Poverty rates, minority population percentages.

The dataset was obtained from federal sources and required preprocessing to handle missing values, normalize formats, and aggregate data. Challenges included handling large file sizes and aligning shapefile geometries with tabular data for mapping.

## Technical Approach

The analysis used geospatial and statistical techniques to derive insights:

* **GeoPandas and Folium** for mapping and visualization.
* Aggregation methods like mean and sum to compare results for boolean indicators.
* Data cleaning techniques to harmonize census tract data with county-level shapefiles.
* Statistical summaries and trend analyses to interpret the metrics.

## Implementation

Python was the primary tool, with key libraries including GeoPandas for geospatial data manipulation, Folium for interactive maps, and Pandas for data wrangling. Key implementation steps included:

1. Preprocessing and merging CEJST data with shapefiles using common keys.
2. Aggregating metrics (e.g., mean poverty rates, sum of at-risk properties) at the county level.
3. Visualizing metrics using interactive heatmaps, customized for user-friendly exploration.
4. Iterative testing and optimization to handle large datasets efficiently.

## Results

Key findings include:

* **Environmental Burdens**: Counties with high flood risks often correlate with lower socioeconomic indicators.
* **Poverty Metrics**: Aggregated poverty rates reveal disparities across states, with significant clusters in historically underprivileged regions.
* **Interactive Maps**: Custom visualizations allow stakeholders to explore metrics dynamically, uncovering granular patterns not visible in static reports.

Sample outputs include heatmaps and choropleth maps demonstrating these trends, with color gradients indicating metric intensity.

## Conclusion and Future Work

This project successfully analyzed and visualized CEJST data to highlight environmental and demographic inequalities. The interactive maps provide a scalable foundation for decision-making. Future directions include integrating more datasets (e.g., health outcomes) and refining visualizations for enhanced accessibility.

# **Regression Analysis**

## Dataset Overview

**Granularity**: Census tract level.

## Variables:

1. **Demographics**:
   * Percentages of racial groups (e.g., American Indian/Alaska Native, Asian, Black, White, Hispanic/Latino).
2. **Environmental Metrics**:
   * PM2.5 levels in the air.
   * Energy burden.
3. **Health Metrics**:
   * Prevalence of coronary heart disease, asthma, and diabetes.
4. **Poverty Metrics**:
   * Percentage of individuals below the poverty line.

## Methods

**Data Preprocessing**

* Addressed missing values through imputation or exclusion.
* Scaled continuous variables for comparability.
* Derived poverty metrics from raw census tract data and aggregated them.

## Regression Models

1. **Regression Model 1**: PM2.5 and Demographics
   * **Formula**: PM2.5 ~ Demographic percentages (e.g., Percent White, Percent Black)
   * **Key Insights**:
     + Higher percentages of **White**, **Black**, **Asian**, **American Indian**, and **Hispanic/Latino** populations are significantly associated with **lower PM2.5 levels**.
     + **Native Hawaiian or Pacific Islander** population showed no significant association.
     + **R-squared**: 0.223 (22.3% of the variance in PM2.5 is explained by demographics).
2. **Regression Model 2**: PM2.5, Total Population, and Energy Burden
   * **Formula**: PM2.5 ~ Total Population + Energy Burden
   * **Key Insights**:
     + Higher **population density** increases PM2.5 levels.
     + **Energy burden** has a negative relationship with PM2.5, suggesting areas with higher burdens may also exhibit lower pollution, possibly due to socioeconomic factors.
     + **R-squared**: 0.009 (minimal explanatory power).
3. **Regression Model 3**: Coronary Heart Disease and Demographics
   * **Formula**: Coronary Heart Disease ~ Demographic percentages
   * **Key Insights**:
     + Higher percentages of **Native Hawaiian**, **American Indian**, **Black**, and **Hispanic** populations are strongly associated with increased coronary heart disease prevalence.
     + **R-squared**: 0.165 (16.5% of variance explained).

## Correlation Analysis

* **Variables**: Total Population, PM2.5, and Energy Burden
* **Findings**:
  + Positive but weak correlation between **PM2.5 and Total Population** (r=0.075).
  + Negative correlation between **Energy Burden and PM2.5** (r=-0.061).
  + Weak negative correlation between **Energy Burden and Total Population** (r=-0.090).

## Results

**Regression Model 1 (PM2.5 and Demographics)**

* **Significant Predictors**:
  + **Percent White**: Negative association (-6.16 units).
  + **Percent Black**: Negative association (-5.02 units).
  + **Percent Asian**: Negative association (-0.82 units).
  + **Percent Hispanic/Latino**: Negative association (-4.13 units).
  + **Percent American Indian**: Strongest negative association (-9.03 units).
  + **Percent Other Races**: Positive association (+2.33 units).
* **Interpretation**:
  + Diverse racial demographics correlate with reduced PM2.5 levels, except "Other Races" (positive correlation).

**Regression Model 2 (PM2.5, Total Population, and Energy Burden)**

* **Significant Predictors**:
  + **Total Population**: Positive association.
  + **Energy Burden**: Negative association.
* **Interpretation**:
  + High-density areas experience elevated PM2.5, while energy burden correlates inversely.

**Regression Model 3 (Coronary Heart Disease and Demographics)**

* **Significant Predictors**:
  + **Native Hawaiian**: Highest positive correlation (+1482.93 cases/1% increase).
  + **American Indian**: (+1231.96 cases/1% increase).
  + **Black or African American**: (+894.34 cases/1% increase).
  + **Hispanic/Latino**: (+662.91 cases/1% increase).
  + **White**: (+796.37 cases/1% increase).
  + **Other Races**: Negative correlation (-63.70 cases/1% increase).
* **Interpretation**:
  + Certain racial demographics (Native Hawaiian, American Indian) are strongly linked to coronary heart disease prevalence.

**Correlation Heatmap**

* Highlights weak relationships among key variables (Total Population, PM2.5, and Energy Burden).

## Visualizations

1. **Regression Coefficients**: Bar charts illustrating the strength of predictors.
2. **Correlation Heatmap**: Visualizing relationships between predictors.
3. **Residual Plots**: Checking model assumptions.

## Key Takeaways

1. **Environmental Health**:
   * PM2.5 levels are influenced by demographics and total population.
   * Energy burden exhibits an inverse relationship, suggesting socioeconomic factors play a role.
2. **Health Disparities**:
   * Native Hawaiian and American Indian populations exhibit the highest risks for coronary heart disease.
3. **Policy Implications**:
   * Addressing health and environmental disparities requires tailored interventions based on demographic and socioeconomic characteristics.
4. **Limitations**:
   * Low R-squared values highlight the need for additional predictors.
   * Non-normal residuals and multicollinearity require further exploration.

## Future Work

1. **Include More Predictors**:
   * Land use, traffic, and industrial factors.
2. **Time-Series Analysis**:
   * Investigate trends over time.
3. **Policy Recommendations**:
   * Develop actionable plans based on demographic-environmental interactions.

# **Poverty Analysis**

## Key Findings

## Racial Demographics and Poverty

**Analysis revealed disparities in poverty levels across racial groups. By calculating average poverty percentages for each racial category, it became evident that minority populations, particularly Hispanic or Latino and Black or African American groups, faced higher poverty rates.**

**Key Statistics:**

* **Mean poverty rate across all counties: 34.56%**
* **Median poverty rate: 34.5%**

**Using bar charts, racial compositions were visualized to highlight population disparities. A subsequent analysis of raw population counts across racial groups showed that, while White populations are numerically dominant, minority groups disproportionately experience poverty.**

**Visualization:**

* **Bar Chart 1: Displays average percentages of racial/ethnic groups across counties.**
* **Bar Chart 2: Shows total population counts for each racial/ethnic group.**

**Both visualizations reinforced the disproportionate burden of poverty on minority populations.**

**Environmental Risks and Correlation Analysis**

**A core component of the analysis involved examining the correlation between environmental risks (e.g., fire and flood risk) and demographic compositions. Using a heatmap of the correlation matrix, the following observations emerged:**

* **Fire risk exhibited a weak positive correlation with Hispanic or Latino populations (0.12) and American Indian/Alaska Native populations (0.08).**
* **Flood risk similarly showed a weak positive correlation with Hispanic or Latino populations (0.11).**
* **White populations had a weak negative correlation with both fire (-0.07) and flood risks (-0.12).**

**While the correlations were not strong, they pointed to potential systemic vulnerabilities in areas with higher minority populations. For example, geographic and infrastructural factors may contribute to higher risks in these communities.**

**Visualization:**

* **Heatmap: Illustrated the correlation matrix, providing a clear depiction of relationships between racial demographics and environmental risks.**

**Composite Poverty Score: A Multi-Dimensional Metric**

**A composite "Poverty Score" was created by combining multiple poverty-related indicators into a single metric. This approach allowed for a more comprehensive understanding of socio-economic vulnerability. The indicators included:**

* **Percentages of individuals below the poverty line.**
* **Environmental risks (e.g., fire, flood).**
* **Health and linguistic isolation metrics.**

**Using MinMax scaling, each variable was normalized before aggregation. The final Poverty Score was calculated as the average of all contributing metrics.**

**Mapping Poverty Scores**

**Visualizing Poverty Scores on geographic maps provided actionable insights:**

* **Census Tract-Level Map: Displayed Poverty Scores with granular detail, highlighting high-risk zones.**
* **County-Level Map: Aggregated data to provide a broader view of regional disparities.**

**The interactive maps revealed clusters of high Poverty Scores in areas with overlapping socio-economic and environmental vulnerabilities. For example, counties in the Southeast exhibited consistently high scores, suggesting systemic issues that warrant targeted intervention.**

## Federal Poverty Line Indicators

**A screenshot of a table

Description automatically generated**

## Percentage of People Below 100% Federal Poverty Line

This metric represents individuals living at or below the federally defined poverty threshold.

* **Top Counties with Extreme Poverty**:
  + **Mellette County, South Dakota**: Leads with a staggering **98%** of residents below the poverty line.
  + **Corson County, South Dakota** and **Ziebach County, South Dakota** follow closely, with poverty rates of **96.5%** and **96.0%**, respectively.
  + **Clinch County, Georgia**, at **95.5%**, reflects the same economic vulnerabilities seen in rural regions.
* **Geographic Concentration**:
  + Rural counties in **South Dakota**, **Mississippi**, and **Kentucky** are heavily represented, reflecting systemic challenges in access to resources, employment, and infrastructure.
* **Implications**:
  + South Dakota consistently shows extreme poverty levels, with systemic issues likely tied to underinvestment in indigenous and rural communities.
  + Other rural counties across the southern U.S. face similar economic hardship, emphasizing the need for targeted policy interventions.

## Percentage of People Below 200% Federal Poverty Line

This metric expands the poverty threshold to include individuals earning up to twice the federal poverty level, offering a broader perspective on economic insecurity.

* **Top Counties with Widespread Economic Hardship**:
  + **Issaquena County, Mississippi**: Tops the list with **68%** of its population below 200% of the poverty line.
  + **Quitman County, Mississippi** (**66.67%**) and **Keweenaw County, Michigan** (**66.5%**) also report alarmingly high levels of economic insecurity.
* **Regional Observations**:
  + Counties in **Mississippi**, **Louisiana**, and **Texas** dominate this list, reflecting persistent poverty in the Southeast and rural Midwest.
  + Several counties overlap between the two charts, such as **Mellette County, SD** and **Ziebach County, SD**, indicating deep and persistent poverty.
* **Insights**:
  + The 200% threshold reveals that economic hardship extends well beyond those classified as impoverished under the stricter 100% threshold. A significant portion of these populations may struggle to afford basic needs despite being above the official poverty line.

## Key Takeaways

1. **Severe Economic Disparities**:
   * Counties with the highest poverty rates (e.g., Mellette and Corson in South Dakota) reflect extreme economic distress, warranting urgent intervention.
   * Mississippi emerges as a critical focus area, with multiple counties reporting widespread poverty at both thresholds.
2. **Systemic Challenges**:
   * High poverty rates in rural counties suggest systemic barriers such as limited access to education, healthcare, and employment opportunities.
   * Indigenous communities, particularly in South Dakota, face compounded challenges, as evidenced by consistently high poverty rates.
3. **Policy Recommendations**:
   * **Targeted Economic Development**: Invest in job creation, education, and infrastructure in rural and indigenous communities.
   * **Social Safety Nets**: Expand federal and state programs to support low-income families, particularly in regions with high poverty rates.
   * **Holistic Support**: Address healthcare access and housing affordability, which are closely tied to poverty and economic security.

## Insights and Discussion - Disparities in Poverty

**The analysis showed clear disparities in poverty levels among racial groups. Minority populations consistently faced higher poverty rates, suggesting entrenched systemic inequalities. These findings align with existing research on racial wealth gaps and unequal access to resources.**

## Environmental Risks

**While the correlations between environmental risks and racial compositions were weak, they highlight potential areas for further study. Geographic and infrastructural factors likely play a role in exposing certain populations to higher risks, but additional data (e.g., housing quality, access to emergency services) would be necessary to draw stronger conclusions.**

## Use of Composite Metrics

**The creation of the Poverty Score demonstrated the utility of composite metrics in simplifying complex datasets. By integrating multiple indicators, the score provided a holistic view of socio-economic vulnerability, facilitating comparisons across regions.**

## Policy Implications

**The findings from this analysis have several implications for policymakers:**

1. **Targeted Interventions: Counties with high Poverty Scores should be prioritized for resource allocation, including social programs, environmental remediation, and infrastructure improvements.**
2. **Racial Equity: Policies should address systemic inequities, ensuring that minority populations receive equitable access to resources and opportunities.**
3. **Data-Driven Decision-Making: Interactive maps and composite metrics can serve as tools for policymakers, enabling evidence-based decisions.**

# **Health Analysis**

## Top Counties with High Diabetes Rates

The bar chart (see Figure 1) illustrates the five counties with the highest rates of diagnosed diabetes. Notably:

* **Greene County** has the highest diabetes rate, closely followed by **Clay County** and **Perry County**.
* These counties exhibit a diabetes prevalence significantly above the average, highlighting a potential focus area for public health interventions.

The clustering of counties with high diabetes rates suggests underlying socio-economic or environmental factors influencing health outcomes. For instance, these areas may have limited access to healthcare or face higher environmental stressors.

## Distribution of Health Scores

The histogram of health scores (see Figure 2) shows a relatively normal distribution, with most counties falling between **0.3 and 0.4**. Key takeaways include:

* A small subset of counties exhibits scores higher than 0.5, indicating exceptionally high health risks.
* Counties with low scores (below 0.25) are relatively rare, suggesting widespread moderate health challenges.

The distribution underscores the importance of prioritizing counties at the higher end of the spectrum for resource allocation.

## Analysis of Health Metrics

The composite Health Score offers a unified view of chronic disease burdens across counties. By averaging scaled metrics for asthma, diabetes, and coronary heart disease, the score highlights areas with compounded vulnerabilities. Key findings include:

* Counties with high health scores also tend to rank high in specific conditions like diabetes or asthma.
* Geographic clustering of high health scores aligns with broader socio-economic trends in the state.

## Asthma Prevalence by County

A screenshot of a graph

Description automatically generated

The bar chart displays the average current asthma prevalence among adults aged 18 years and older across various counties in the United States. The visualization highlights the top counties with the highest asthma rates, underscoring significant health disparities.

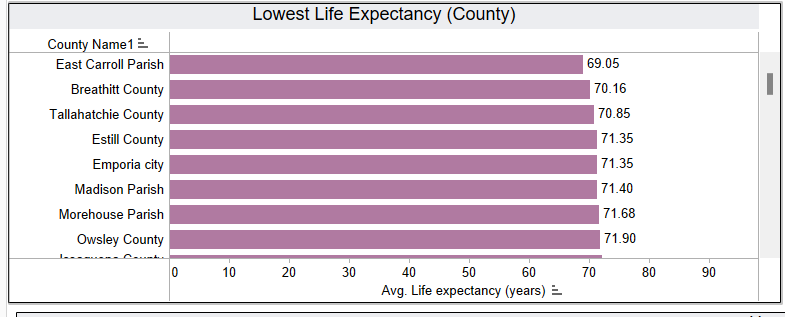
## Key Observations:

1. **Todd County, South Dakota** leads with the highest asthma prevalence at 1,375 cases per 100,000 adults. This indicates a severe burden of respiratory health issues within this community.
2. **McDowell County, West Virginia** and **Apache County, Arizona** follow closely, with rates of 1,337.5 and 1,327.4 cases per 100,000 adults, respectively.
3. Several counties, including **Bethel Census Area, Alaska** and **Menominee County, Wisconsin**, report asthma prevalence exceeding 1,300 cases per 100,000 adults, reflecting a consistent trend of high prevalence in certain rural and underprivileged areas.
4. **Geographic Disparities**:
   * Counties in states like **Mississippi** (e.g., Holmes County, Humphreys County, and Quitman County) and **Montana** (e.g., Big Horn County, Glacier County) are repeatedly represented, indicating regional clusters of elevated asthma prevalence.
   * Native populations in areas such as the **Northwest Arctic Borough, Alaska**, also show high asthma rates, possibly linked to environmental and healthcare access challenges.

## Implications:

1. **Environmental Factors**: High asthma prevalence in certain regions may correlate with poor air quality, exposure to environmental pollutants, or inadequate housing conditions.
2. **Healthcare Access**: Many of the counties listed are rural or economically disadvantaged, potentially limiting residents' access to healthcare and preventive interventions for asthma management.

## Counties with the Lowest Life Expectancy



## Lowest Life Expectancy:

* **East Carroll Parish, Louisiana** reports the lowest average life expectancy at **69.05 years**, significantly below the national average.

## Regional Patterns:

* Most counties with low life expectancy, such as **Tallahatchie County (70.85 years)** and **Breathitt County (70.16 years)**, are located in the Southern United States, particularly in economically disadvantaged or rural areas.

## Underlying Factors:

* Lower life expectancy is often linked to systemic issues, including:
  + **Limited access to healthcare.**
  + **High poverty rates.**
  + **Environmental risks** (e.g., poor air quality, lead exposure).
  + **Chronic diseases** such as diabetes and heart disease.

## Policy Recommendations

* Counties with lower life expectancy require targeted public health interventions, such as:
  + Expanding healthcare access.
  + Addressing socioeconomic inequities.
  + Implementing preventative care programs for chronic conditions.

## Coronary Heart Disease and Other Health Risks

In addition to diabetes, coronary heart disease is a significant concern in several counties. Mapping these metrics highlights geographic patterns of vulnerability:

* Counties with high rates of heart disease often overlap with high diabetes rates, pointing to potential co-morbidities.
* Environmental factors, such as air quality and access to healthy foods, may play a role in these patterns.

## Policy Recommendations

Based on the analysis, several policy recommendations emerge:

1. **Targeted Health Interventions**: Focus on counties with high health scores, such as Greene and Perry, to improve access to healthcare and preventive measures.
2. **Chronic Disease Management Programs**: Develop initiatives to address diabetes and coronary heart disease, emphasizing education and lifestyle interventions.
3. **Environmental Health Policies**: Address environmental factors contributing to chronic diseases, such as air quality and access to recreational spaces.

# **Education Analysis**

## Top Counties with Highest Education Scores

The bar chart above highlights the top five counties with the highest education scores. These scores were calculated by normalizing and combining critical indicators such as:

* **Percent of individuals age 25 or older with less than a high school degree.**
* **Percent of residents not currently enrolled in higher education.**
* **Presence of abandoned mines.**

## Key Observations:

1. **Wayne County** leads with the highest Education Score, followed by **Clay County** and **Wolfe County**.
2. Counties in the top five exhibit elevated educational vulnerabilities, such as high percentages of residents with minimal formal education and a lack of higher education enrollment.
3. These areas also show potential correlations with environmental challenges, such as abandoned mines, which further compound educational and economic challenges.

## Distribution of Education Scores

The histogram of education scores provides a broad view of how counties fare in terms of educational outcomes:

* Most counties cluster around mid-range education scores, reflecting moderate educational attainment challenges.
* A small subset of counties falls in the lower range of scores, indicating severe educational disparities and vulnerabilities.
* The tail-end of the distribution includes counties with higher scores, which could reflect robust educational resources or better socioeconomic conditions.

This distribution suggests that while many counties face moderate challenges, there is a critical need to address the extreme outliers.

## Methodology

## 1. Data Preparation

The analysis began by merging education metrics with geographic shapefiles for Tennessee, Kentucky, and West Virginia. This ensured spatial alignment and enabled mapping of education metrics at the county level. Key preprocessing steps included:

* Filtering geographic data for the target states.
* Normalizing education-related indicators using **MinMaxScaler** for comparability.
* Handling missing data and ensuring consistency across metrics.

## 2. Composite Education Score

The **Education Score** was calculated by averaging normalized indicators. Equal weights were assigned to all metrics:

* High school completion rates.
* Higher education enrollment rates.
* Presence of abandoned mines as an environmental challenge.

## 3. Data Aggregation

Counties were aggregated by:

* Demographic and education metrics (average calculations across census tracts).
* Geometric aggregation to merge spatial data.

## Key Findings

**Disparities in Educational Attainment**

1. **Counties with Low Education Scores**:
   * These counties typically exhibit high dropout rates, minimal higher education participation, and environmental challenges like abandoned mines.
   * Rural areas dominate this group, reflecting systemic barriers to educational access and infrastructure.
2. **Geographic Clustering**:
   * High-scoring counties often appear in urban or semi-urban areas, where resources such as schools and community programs are more accessible.
   * Low-scoring counties are concentrated in economically disadvantaged regions with environmental degradation.

## Policy Recommendations

1. **Educational Access Programs**:
   * Targeted interventions to increase high school graduation rates in counties with low scores.
   * Scholarships and outreach programs to encourage higher education enrollment.
2. **Infrastructure Development**:
   * Invest in rural education infrastructure, including school buildings, technology, and transportation.
3. **Community Engagement**:
   * Develop local initiatives to address educational and environmental challenges, including cleaning abandoned mines and providing community learning resources.
4. **Data-Driven Monitoring**:
   * Utilize the composite Education Score to track progress over time and allocate resources effectively.

# **Environmental Analysis**

1. **Environmental Data**:
   * Energy burden and air quality metrics (e.g., PM2.5).
   * Risk indicators like proximity to hazardous waste sites.
   * Housing metrics like pre-1960s lead paint indicators.
2. **Poverty Data**: Metrics reflecting socioeconomic challenges aggregated into a "Poverty Score."
3. **CEJST Shapefile**: Geospatial data at the census tract level.

## Data Preprocessing

1. Imported environmental and poverty datasets and merged them with shapefile data for Tennessee counties.
2. Converted relevant columns to numeric types for calculations and aggregations.
3. Aggregated energy burden metrics by race to calculate mean burdens at the county level.

## Analysis Methods

1. **Environmental Score**: Calculated by averaging normalized environmental indicators (energy burden, PM2.5, etc.).
2. **KMeans Clustering**: Grouped counties into clusters based on mean energy burdens for various racial groups to identify patterns.
3. **Interactive Maps**:
   * Displayed environmental scores, PM2.5 levels, and poverty scores using layered maps.
   * Enabled toggling between layers for dynamic exploration.

## Results

1. **KMeans Clustering**:
   * Scatterplot shows clusters of counties with distinct energy burdens for racial groups, highlighting disparities.
   * Clusters suggest higher burdens for certain racial groups in specific regions.
2. **Environmental Score Map**:
   * Interactive map visualizes the aggregated environmental score for each Tennessee county.
   * Counties with higher scores indicate greater environmental challenges.
3. **Poverty Map**:
   * Displays poverty scores by county using a red-gradient choropleth.
   * Counties with higher scores align with those facing significant environmental burdens, indicating overlapping challenges.

## Key Insights

1. Counties with high energy burdens often coincide with high poverty scores, revealing a double burden of economic and environmental challenges.
2. Disparities in energy burdens by racial groups are evident, with Black and Hispanic populations facing higher burdens in specific regions.
3. KMeans clustering effectively highlights patterns, allowing policymakers to target interventions in the most affected clusters.

## Visualizations

1. **KMeans Clustering Plot**:
   * Showcases clusters based on energy burdens for racial groups.
   * Useful for identifying counties needing targeted assistance.
2. **Environmental Score Map**:
   * Illustrates geographic distribution of environmental challenges.
3. **Poverty Score Map**:
   * Overlayed with environmental data for a comprehensive view.

## Analysis of Energy Burden vs. Low Median Income

## Key Observations

1. **Positive Correlation**:
   * The regression line demonstrates a positive relationship between low median income and energy burden.
   * As the percentage of households with low median income increases, the average energy burden also rises, indicating that economically disadvantaged areas face higher energy costs relative to their income.
2. **Regression Equation**:
   * **Avg. Energy Burden = 0.0503193 \* (Low Median Household Income Percentile) + 0.93051**
   * The slope (**0.0503193**) suggests that for each percentile increase in low median household income, the average energy burden increases by approximately **0.05 units**.
3. **Model Fit**:
   * **R-Squared = 0.34818**:
     + Approximately **34.8%** of the variability in average energy burden is explained by the low median income percentile.
     + While this indicates a moderate relationship, other factors not included in the model likely contribute to energy burden disparities.
   * **P-Value < 0.0001**:
     + The relationship between income and energy burden is statistically significant, meaning the observed correlation is unlikely due to chance.

## Interpretation

1. **Energy Burden Inequality**:
   * Communities with lower median household income face disproportionately higher energy burdens. This suggests systemic inequities in energy access and affordability.
2. **Socioeconomic Implications**:
   * High energy burdens in low-income areas may exacerbate financial strain, forcing families to allocate a significant portion of their income to utilities. This could reduce their ability to afford other necessities such as food, healthcare, and education.
3. **Geographic Patterns**:
   * Although not explicitly mapped in this visualization, such disparities often align with historically underserved regions, particularly rural areas and minority-dense urban neighborhoods.

## Policy Implications

1. **Energy Assistance Programs**:
   * Expand subsidies or assistance programs targeting high-burden, low-income households to alleviate financial strain.
   * Provide grants or tax incentives for energy-efficient upgrades in low-income housing.
2. **Renewable Energy Access**:
   * Encourage investments in renewable energy solutions, such as solar panel installations, in economically disadvantaged communities to reduce energy costs.
3. **Infrastructure Improvements**:
   * Upgrade infrastructure in low-income areas to increase energy efficiency and reduce waste, ultimately lowering energy costs for residents.
4. **Data-Informed Interventions**:
   * Use models like this to identify areas with the highest energy burden and allocate resources accordingly.

## Lead Paint Risk Analysis: Homes Built Before 1960

A screenshot of a table with numbers and names

Description automatically generated

## Top Counties with High Proportions of Pre-1960 Homes

1. **St. Louis City, Missouri**:
   * Leads the list with **74.97%** of homes built before 1960, indicating a substantial risk of lead exposure.
2. **Covington City, Virginia**, and **Baltimore City, Maryland**:
   * Follow closely with **72.50%** and **72.34%**, respectively, reflecting a severe need for intervention.
3. **Kings County, New York**:
   * At **69.50%**, highlights the risks present in densely populated urban centers.
4. **Kansas Counties**:
   * Osborne, Republic, and Cloud Counties in Kansas show consistently high percentages (67–68%), suggesting rural areas are not exempt from this risk.

## Health Impacts of Lead Exposure

* **Cognitive and Neurological Effects**:
  + Lead exposure is particularly harmful to young children, as it can impair brain development, leading to reduced IQ, learning disabilities, and behavioral issues.
  + For adults, chronic exposure can cause memory loss, difficulty concentrating, and cognitive decline.
* **Physical Health Risks**:
  + Lead poisoning can result in anemia, kidney damage, and high blood pressure.
  + Pregnant women exposed to lead are at increased risk of miscarriage, stillbirth, or delivering babies with developmental issues.
* **Generational Impact**:
  + Children exposed to lead during critical developmental periods may face lifelong challenges, perpetuating cycles of poverty and poor health in affected communities.

## Geographic and Socioeconomic Implications

1. **Urban Centers**:
   * Cities such as St. Louis, Baltimore, and Kings County represent urban areas with older housing stock, likely linked to historical underinvestment in housing renewal and maintenance.
2. **Rural Regions**:
   * Counties in Kansas, Pennsylvania, and Nebraska demonstrate that rural communities are also at high risk, potentially due to limited resources for home renovations and lead abatement programs.
3. **Equity Concerns**:
   * Vulnerable populations, including low-income families and communities of color, are disproportionately impacted due to the affordability of older homes and limited access to remediation efforts.

## Policy Recommendations

1. **Lead Paint Abatement Programs**:
   * Increase funding for lead remediation in high-risk areas, prioritizing homes with children under six and pregnant women.
   * Expand incentives for landlords and homeowners to replace lead-based paint and pipes.
2. **Public Health Campaigns**:
   * Raise awareness about the risks of lead exposure and provide resources for identifying and mitigating hazards in older homes.
   * Offer free or subsidized testing for lead in homes and blood tests for residents in high-risk areas.
3. **Targeted Interventions in Urban Centers**:
   * Focus on cities like St. Louis, Baltimore, and Kings County, where dense populations amplify the public health impact of lead exposure.
4. **Support for Rural Communities**:
   * Ensure rural counties receive adequate funding and technical assistance for lead paint remediation efforts.

Final thoughts: Clearly both urban and rural areas face significant challenges in mitigating lead exposure risks, making a comprehensive national strategy essential.

## Future Work

1. **Correlation Analysis**: Quantify the relationship between environmental and poverty scores.
2. **Policy Recommendations**: Use insights to propose interventions.
3. **Temporal Analysis**: Explore changes over time to evaluate policy impacts.

## References

**U.S. Council on Environmental Quality.** (n.d.). *Climate and Economic Justice Screening Tool.* Retrieved from <https://screeningtool.geoplatform.gov/en/#5.96/36.576/-82.65>