
Modeling Healthcare Access in U.S. Counties:
A Multivariate Analysis of Socioeconomic and Environmental Factors

A datadriven portfolio project exploring equitable care through demographic and socioeconomic indicators

Analyzing Determinants of Care Availability

With rising housing costs and uneven healthcare services, it's increasingly important to understand what drives access to care. This analysis explores how key socio-economic and environmental factors—including income, access to exercise, and food insecurity—shape healthcare availability across U.S. counties.

Centering Demographics in Access Analysis

While this project was developed independently as a portfolio piece, the analysis offers valuable findings that extend beyond its original scope. Though grounded in four key demographic profiles—young families, first-generation immigrants, recent graduates, and empty nesters—it addresses broader issues of equitable access to essential services faced by many communities.

Data Exploration and Modeling

This analysis provides insights to guide data-driven decisions about healthcare accessibility. It was conducted through exploratory data analysis and the application of machine learning techniques—including regression, clustering, and time series modeling—on scaled county-level socioeconomic and environmental indicators.

Five-Year Snapshot of U.S. County Data

The datasets reflect aggregated observations from the most recent five-year period (2020–2024), providing a current snapshot that can serve as a reliable reference for the next 2 to 3 years—assuming no major economic or public health disruptions. The analysis covers all counties across the 50 U.S. states and the District of Columbia.

Built from Trusted Public Health Sources

The final dataset was constructed by integrating eight distinct sources, primarily drawing from County Health Rankings (2020–2024), a dataset developed by the University of Wisconsin Population Health Institute in collaboration with the Robert Wood Johnson Foundation. This resource includes 45 key county-level metrics and aggregates national data on health outcomes and social determinants from reputable agencies such as the CDC, CMS, and the U.S. Census Bureau.

First Insights from EDA on Healthcare Access

Feature Selection Strategy

This project began with Exploratory Data Analysis (EDA) to identify key drivers of healthcare access. From over 50 cleaned features, 13 mean-level variables were selected based on logical relevance and analytical value. All variables represent **five-year (2020–2024) means**, aggregated at the **county level across 3,143 U.S. counties** in all 50 states and the District of Columbia.

These indicators are grouped into three thematic categories:

Healthcare Access and Resources:
Uninsured Rate, Primary Care Physicians Rate, Preventable Hospital Stays, Mammography Screening, Flu Vaccination
Social and Economic Predictors:
Median Household Income, Child Poverty Rate, Unemployment Rate, Income Inequality, Racial/Ethnic Diversity Index
Environmental and Lifestyle Factors:
Access to Exercise Opportunities, Food Insecurity Rate

Insurance Impact

The **Uninsured Rate** showed a weak negative correlation, averaging **−0.30** across five healthcare-related features. The strongest relationship emerged with **Mammography Screening Rates**, showing a moderate negative correlation of **−0.46**. This suggests that higher uninsured rates—particularly among age-specific populations—are closely linked to reduced access to preventive care for women.

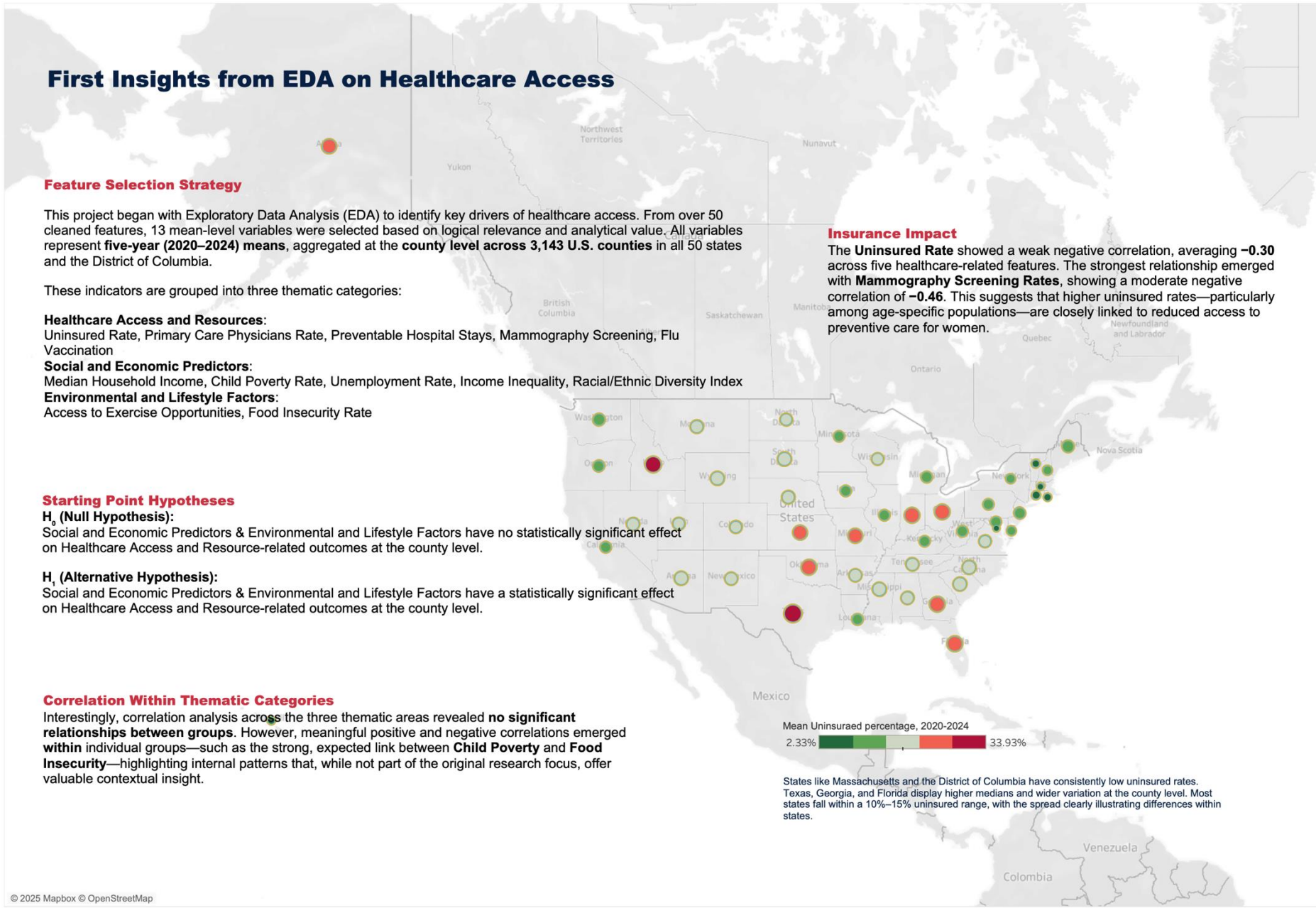
Starting Point Hypotheses

H_0 (Null Hypothesis):
Social and Economic Predictors & Environmental and Lifestyle Factors have no statistically significant effect on Healthcare Access and Resource-related outcomes at the county level.

H_1 (Alternative Hypothesis):
Social and Economic Predictors & Environmental and Lifestyle Factors have a statistically significant effect on Healthcare Access and Resource-related outcomes at the county level.

Correlation Within Thematic Categories

Interestingly, correlation analysis across the three thematic areas revealed **no significant relationships between groups**. However, meaningful positive and negative correlations emerged **within** individual groups—such as the strong, expected link between **Child Poverty** and **Food Insecurity**—highlighting internal patterns that, while not part of the original research focus, offer valuable contextual insight.



Second EDA: Rejecting the Null Hypothesis

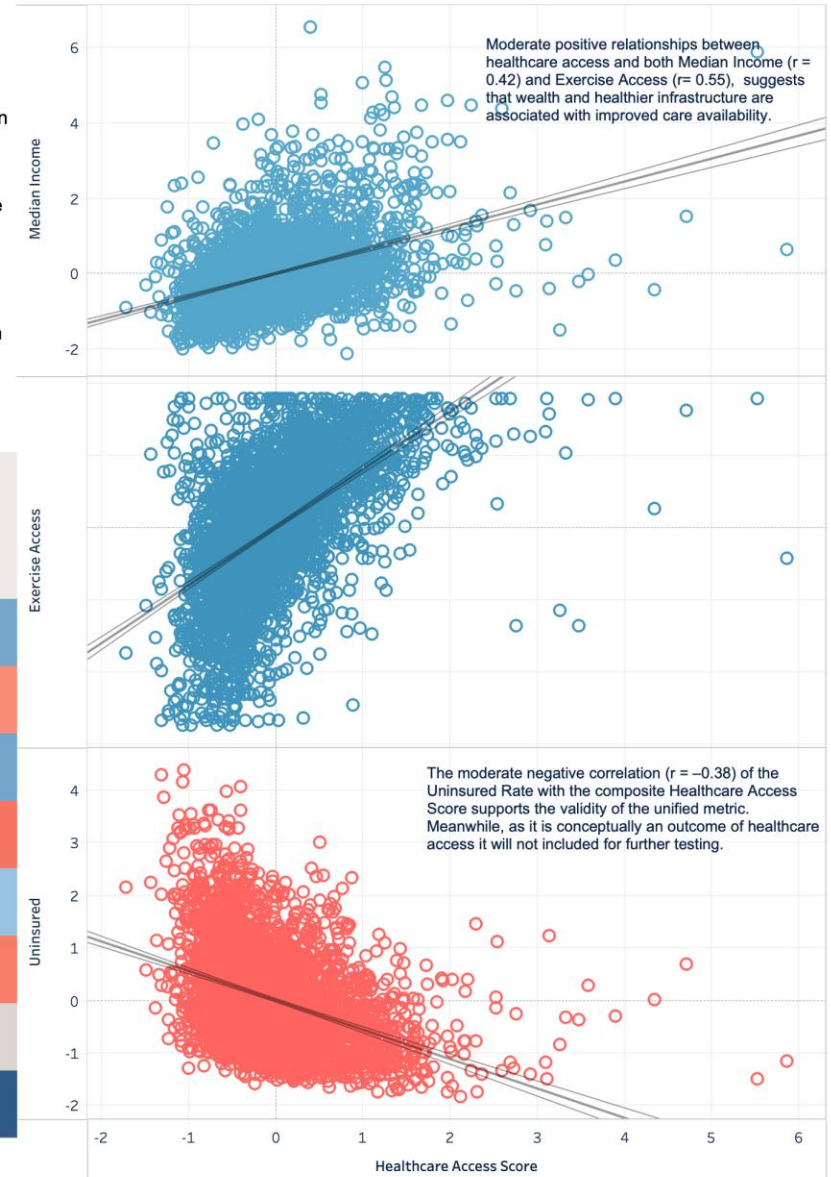
Pivoting Toward a Unified Healthcare Metric

Following the initial EDA—where each feature was examined individually—most predictors showed a median correlation below 10% with the target variable, indicating weak linear relationships and limited standalone predictive power. This prompted a methodological pivot: rather than analyzing variables in isolation, health-related indicators were combined into a single, unified **Healthcare Access Score** to better reflect multidimensional access. This composite measure proved effective, yielding stronger and more interpretable patterns in subsequent analysis.

Weighting Strategy for the Healthcare Access Score

Each standardized (z-scored) indicator in the Healthcare Access Score was weighted based on public health priorities outlined by the World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), and Agency for Healthcare Research and Quality (AHRQ). Core services like primary care (weight = 0.30) and mental health (0.25) were prioritized more heavily due to their foundational role in access to care.

	Diversity Index	Exercise Access	Food Insecurity	Healthcare Access Score	Income Inequality	Median Income	Unemployment	Uninsured
Diversity Index	100%	18%	22%	10%	33%	6%	22%	39%
Exercise Access	18%	100%	-31%	55%	-12%	46%	1%	-26%
Food Insecurity	22%	-31%	100%	-31%	60%	-73%	55%	41%
Healthcare Access Score	10%	55%	-31%	100%	-2%	42%	-7%	-39%
Income Inequality	33%	-12%	60%	-2%	100%	-43%	37%	18%
Median Income	6%	46%	-73%	42%	-43%	100%	-30%	-34%
Unemployment	22%	1%	55%	-7%	37%	-30%	100%	0%
Uninsured	39%	-26%	41%	-39%	18%	-34%	0%	100%



Confirming Patterns Before Machine Learning

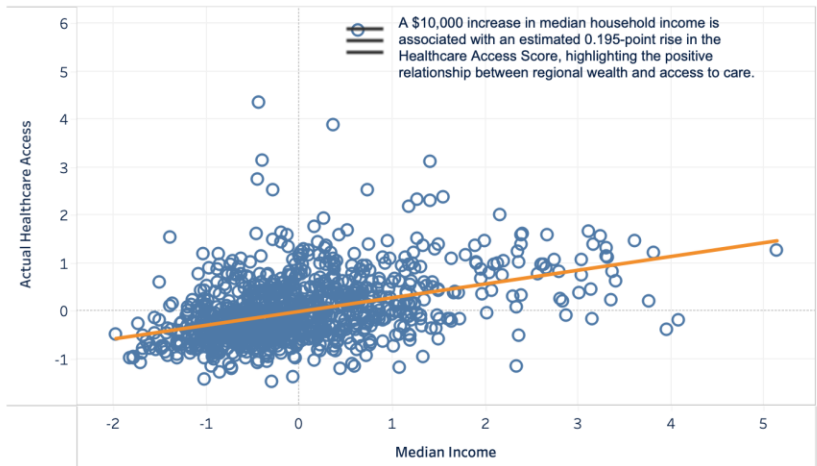
Following two rounds of exploratory data analysis, the results provided preliminary support for the proposed hypotheses. Based on these findings, the analysis proceeds with statistical testing through machine learning methods.

H_0 (Null Hypothesis):
There is no statistically significant relationship between Median Income, Access to Exercise Opportunities, or Food Insecurity and overall Healthcare Access.
H_1 (Alternative Hypothesis):
Median Income and Exercise Access are positively associated with Healthcare Access, while Food Insecurity exhibits a moderate negative relationship.

Supervised linear regression

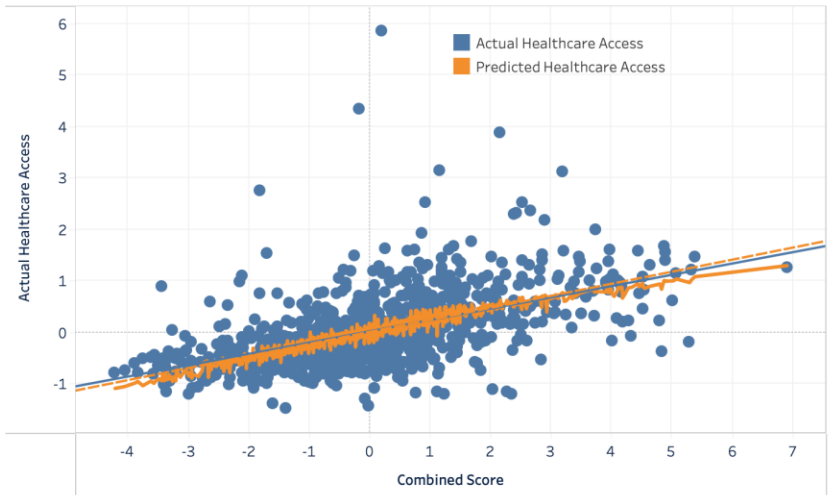
To avoid possible multicollinearity factors, only Median Income, Exercise Access, and Food Insecurity were chosen to add to the linear regression model with Healthcare Access Score, examining their individual and combined relationship with the y-value at a time. During the testing linear regression 4 models were developed and the starting point was between Healthcare Access Score and Median Income.

Model 1



The predictions were moderately off on average, and only 16% of the variation in the Healthcare Access Score was explained by income alone, which was not sufficient to accurately model the outcome.

Model 2



Among the four models tested, Model 2—combining Median Income with Exercise Access—demonstrated the strongest predictive performance. It reduced the Mean Squared Error (MSE) from 0.42 to 0.36 and nearly doubled the explained variance, with the R^2 improving from 0.16 to 0.29. These results highlight that while income is a key predictor, the presence of community-level health infrastructure, such as access to exercise opportunities, plays a critical complementary role in shaping healthcare access across regions. Other 2 models didn't increase the performance significantly.

Clustering for Pattern Discovery in Healthcare Access

The elbow method suggested an optimal cluster count of five. Notably, in the second and most statistically satisfying regression model, clustering revealed a clear linear structure when plotted against the Healthcare Access Score, enhancing interpretability and highlighting the spatial distribution of the clusters.

To explore patterns beyond supervised learning, K-Means clustering was applied using the two key predictors from the second linear regression model — Median Income and Exercise Access — as the foundation for unsupervised segmentation.

Red Cluster – High Risk, Underserved
The Red Cluster represents the most vulnerable counties in the dataset. These areas are marked by deep socioeconomic challenges, including widespread child poverty, high unemployment, and elevated levels of food insecurity and income inequality. Residents in these counties typically experience the lowest median incomes and severely limited access to healthcare services. This cluster reflects communities in urgent need of targeted support and policy intervention.

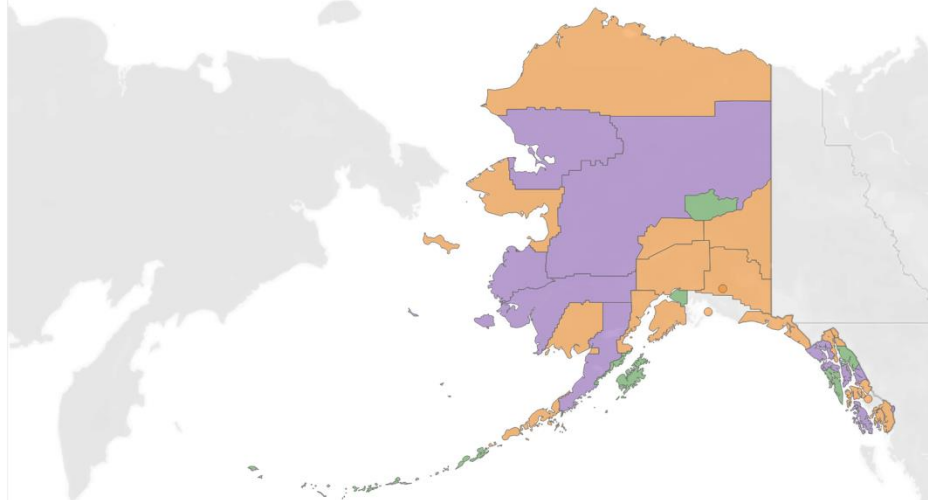
Blue Cluster – Balanced, Slightly Struggling
The Blue Cluster captures counties that fall near the average across most indicators. These areas show slightly lower income levels and healthcare access, but don't display severe disadvantages. They may represent transitional or mid-tier communities—neither thriving nor in crisis—making them a useful reference point for evaluating broader trends.

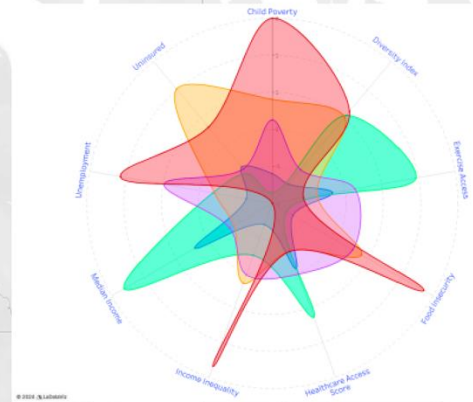
Green Cluster – Prosperous & Well-Served
The Green Cluster represents high-performing counties across the United States. These areas stand out for their strong socioeconomic and health profiles, with notably high median incomes, excellent healthcare access, and widespread exercise opportunities. Additionally, they exhibit low rates of food insecurity, child poverty, and uninsured individuals. Overall, this cluster reflects well-resourced communities with strong infrastructure and access to essential services—making them clear leaders in terms of community well-being.

Purple Cluster – Working-Age Unstable
The Purple Cluster is made up of counties with a mixed profile. While not facing the extremes seen in the Red Cluster, these areas still struggle with pockets of economic instability. They tend to have higher-than-average unemployment and moderately low income levels. These communities may represent younger or more diverse working-class populations navigating financial uncertainty, with uneven access to resources.

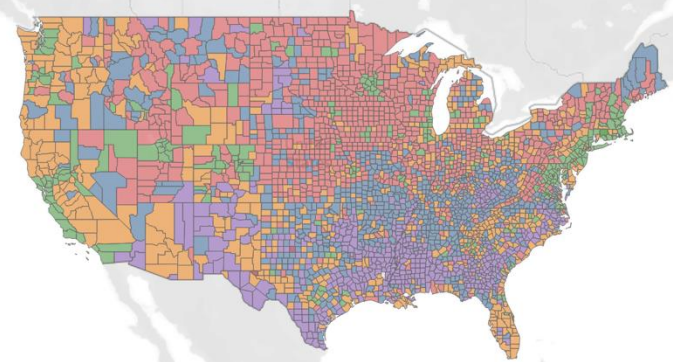
Orange Cluster – Underinsured but Not Poor
The Orange Cluster includes counties where residents may not be economically deprived but still face serious gaps in healthcare access. These areas are defined by high uninsured rates and below-average healthcare access, despite having relatively modest income levels. The population may include working individuals without employer-sponsored insurance, possibly reflecting communities with large informal labor sectors or high immigrant populations.

Healthcare Access in America: Top Counties and Cluster Insights





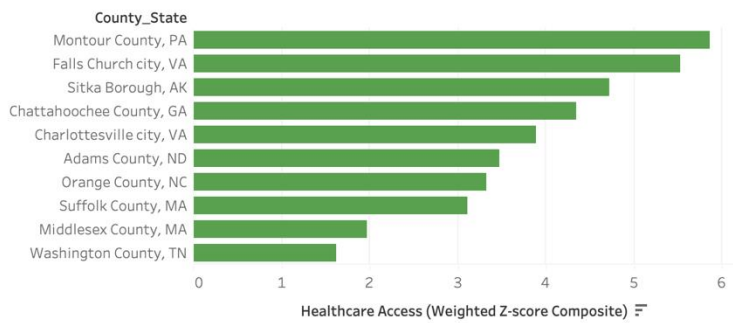
Clusters
0
1
2
3
4



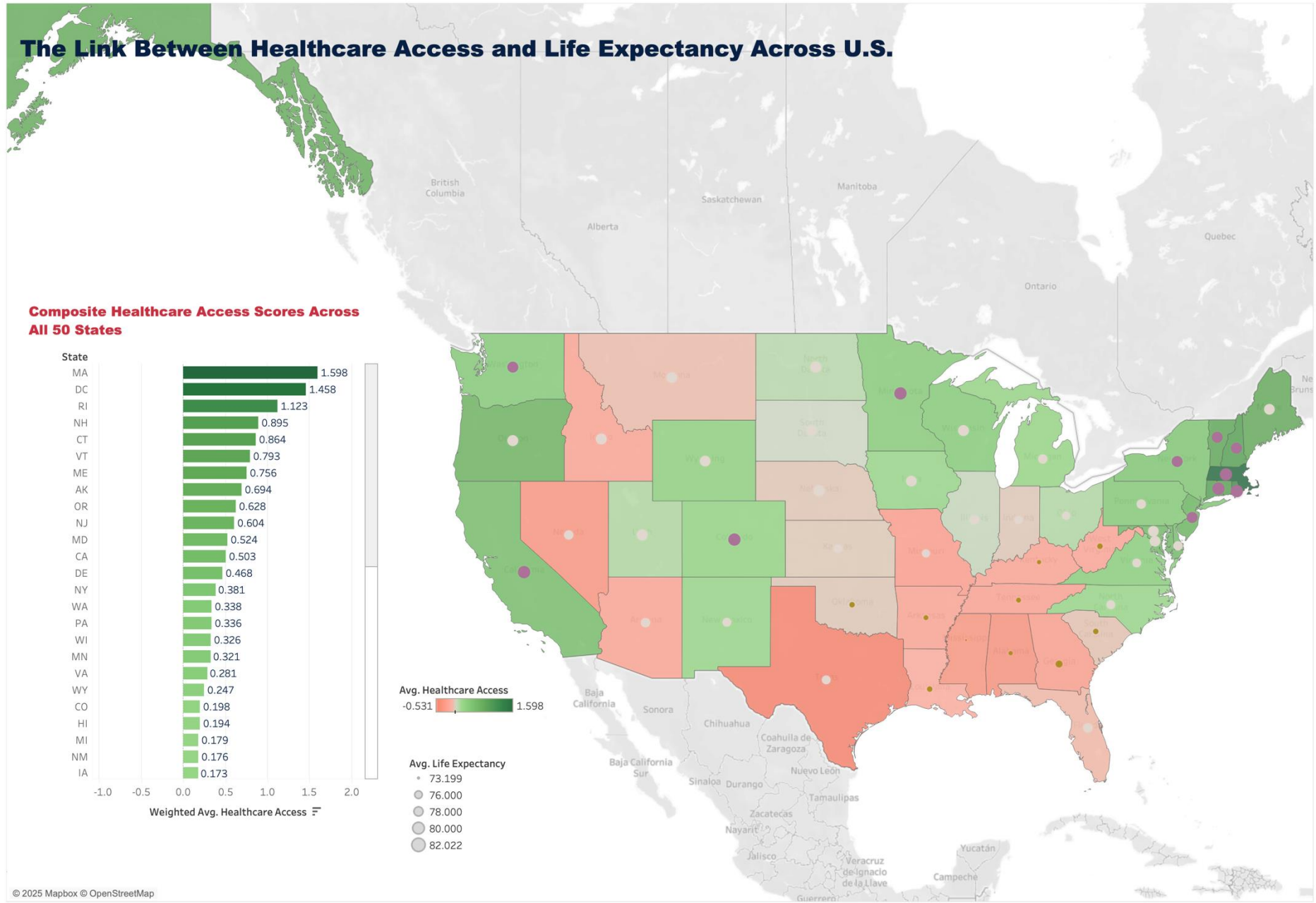
Mexico

Venezuela

The top 10 counties in the United States with the highest Healthcare Access Scores

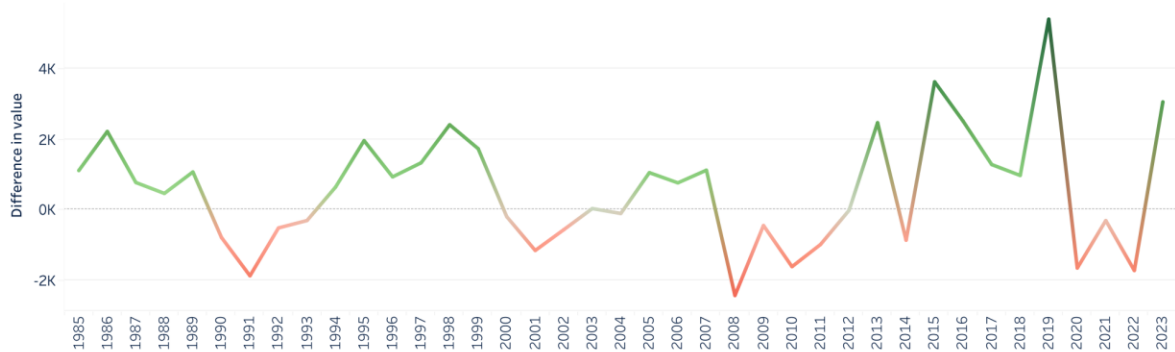


The Link Between Healthcare Access and Life Expectancy Across U.S.



Explaining Healthcare Access Through Income Dynamics

Change in Median Household Income (1985–2023), sourced from FRED



Income Volatility and Its Implications for Access to Care

This differenced time series plot of Median Income highlights year-over-year changes and reveals three sharp declines—around 1991, 2009, and 2020—which align with known economic downturns: the early '90s recession, the Great Recession, and the COVID-19 crisis. Each of these drops signals a potential contraction in household financial stability. In the context of this research, such sudden income declines could correspond to reduced healthcare access, particularly in vulnerable counties, as lower income levels often constrain insurance coverage, preventive care utilization, and local health infrastructure funding.

Conclusion and Rejecting the Null

This analysis demonstrates that healthcare access across U.S. counties is shaped by intersecting socioeconomic and environmental factors. By creating a unified Healthcare Access Score and applying regression, correlation, and clustering methods, the study uncovered meaningful patterns among over 3,000 counties.

Key predictors—**Median Income**, **Exercise Access**, and **Food Insecurity**—emerged as influential, though limited in predictive power. While clustering revealed five distinct county profiles and helped surface regional disparities, the regression model explains only **29% of the variance** in healthcare access. This underscores the complexity of access and the need for deeper, multidimensional data to capture the full picture.

As the analysis is based on American Community Survey (ACS) data, users should consider an approximate **±12% average margin of error** across the selected variables—particularly in smaller or rural counties where variability may be higher.

Despite this limitation, the analysis offers insights that extend beyond its original scope. While it was designed around four key demographic groups—**young families, first-generation immigrants, recent graduates, and empty nesters**—its relevance is broader. It highlights real differences in healthcare infrastructure and access that can inform personal decision-making, public health strategies, and policy development.

For these groups in particular, the findings provide a practical reference point for identifying **counties and states with strong access to care**, spanning five core aspects of healthcare availability. These locations can be clearly visualized through **green-coded zones** in maps and cluster profiles, signaling communities with robust systems in place to support long-term health and well-being.

As a final conclusion, the analysis supports **rejection of the null hypothesis** and affirms the **alternative hypothesis**: **Median Income** and **Exercise Access** are positively associated with **Healthcare Access**. The findings consistently show that counties with higher levels of income and greater access to exercise opportunities tend to have stronger healthcare access scores. This reinforces the importance of socioeconomic and lifestyle factors in shaping equitable access to care across the United States.