Emil Safarov

data analyst

portfolio

Emil Safarov

Data Analytics | Agile Practitioner

Santa Monica, CA

14+ years experience in HR and Organizational Strategy, Marketing across public and private industries

I've turned my passion for asking 'Why?' into a profession—structuring unstructured analytical skills into meaningful data-driven insights.

Education

- **Master's Degree in Management**

Public Administration Academy, Baku, Azerbaijan (2020–2023)

- **Master's Degree in Diplomacy**

Baku Slavic University, Baku, Azerbaijan (2010–2013)

- **Bachelor's Degree in International Relations**

Azerbaijan State Economic University, Baku, Azerbaijan (2005–2009)

Technical certification

- **PMI Agile Certified Practitioner (PMI-ACP®)**

Project Management Institute (PMI) (since 07/2023 – active 07/2027)

- **Career Foundry Bootcamp - Data Analytics Certificate**

Berlin, Germany (01/2024 – 05/2025)

- **Santa Monica College - Data Analytics Certificate**

Santa Monica College, Santa Monica, CA (02/2024 – expected 06/2025)

- **Santa Monica College - Data Science Certificate**

Santa Monica College, Santa Monica, CA (02/2024 – expected 10/2025)

Technical skills

Certified Agile Practitioner (PMI-ACP)
Expertise in Python, SQL, R, Tableau, Excel, UNIX

Education based projects

Modeling Healthcare Access in U.S. Counties

Exploring Socioeconomic and Environmental Determinants through Advanced Analytics

GameCo Marketing Plan Analysis

Data-Driven Insights for Budget Optimization

Optimizing Medical Staffing During Influenza Season

A Data-Driven Approach to Reducing Flu-Related Mortality

Instacart Grocery Basket Analysis

A Data-Driven Approach to Customer Segmentation & Ad Optimization

Rockbuster Stealth Data Analysis Project

Leveraging Data Analytics Online Rental Strategy

Work based projects

Data Collection for Baku City Master Plan

Building the Foundation: Data-driven urban planning

Creative Industries Cluster in Baku

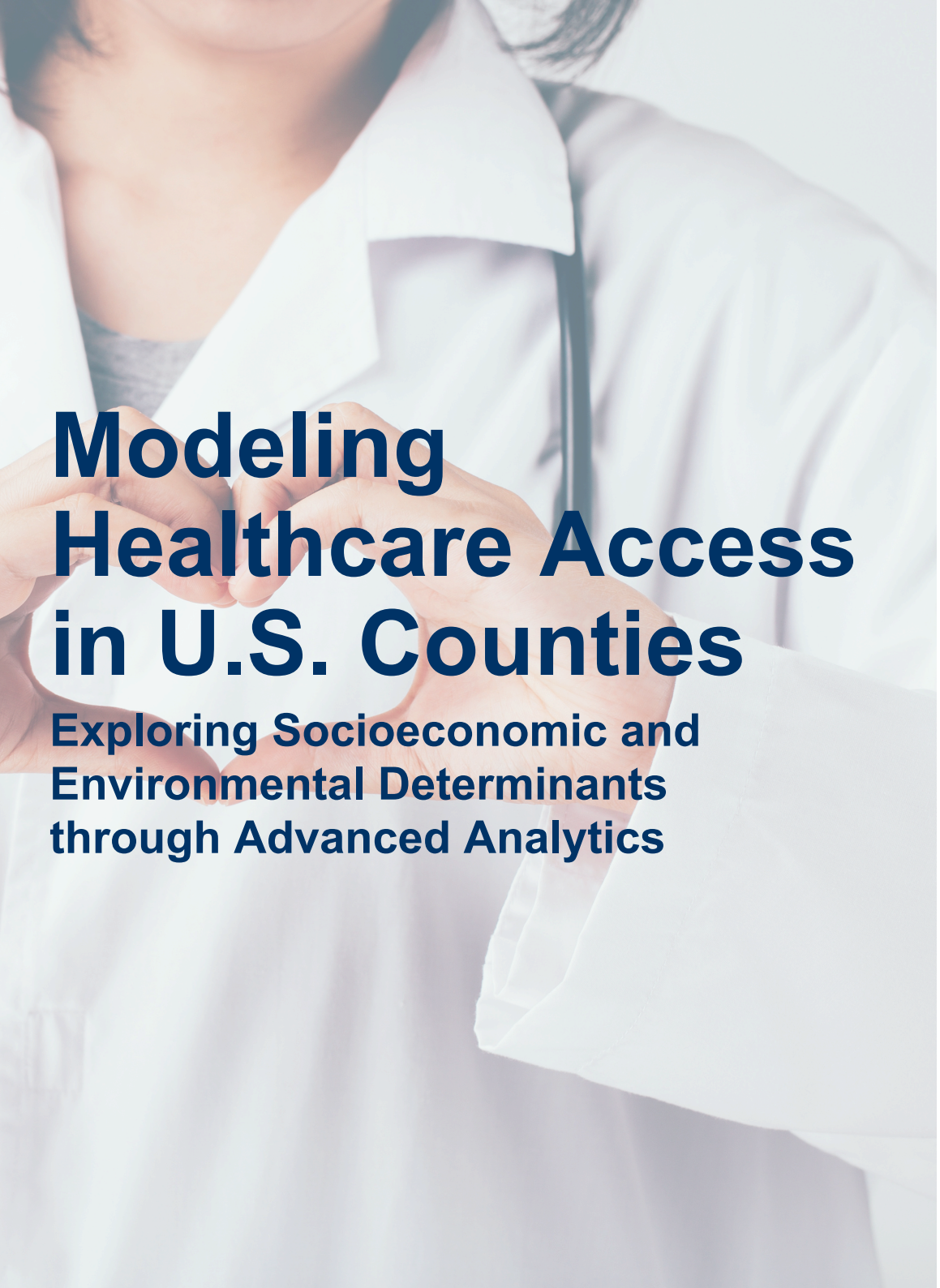
Concept on Transforming Baku's Creative Sector through Urban Regeneration

Baku City Illumination Project

A Data-Driven Approach to Sustainable Urban Lighting

Agile Transformation & Organizational Change initiative

Implementing Scalable Change for Organizational Excellence



Modeling Healthcare Access in U.S. Counties

Exploring Socioeconomic and Environmental Determinants through Advanced Analytics

Problem Overview:

Best Counties to Live in the U.S.: A Data-Driven Approach to Healthcare Access and Quality of Life

This project explores county-level patterns in healthcare access, affordability, and quality of life across the United States. It was developed as part of a broader portfolio initiative to identify the most suitable counties for four key demographic groups: young families, first-generation immigrants, recent graduates, and empty nesters. These groups face unique constraints related to income, housing, healthcare, and community integration, and this project seeks to provide actionable, data-informed insights tailored to their needs.

This project carried out as my fully independent as Advanced Analytics module of the Data Analyst Bootcamp, from data sourcing and cleaning to modeling and visualization, all components were designed, executed, and presented by me to demonstrate both technical skill and storytelling through data.

Data Sourceing & Preparation

Data sourcing and preparation accounted for nearly three-quarters of the total project effort. The resulting clean dataset was designed not only to power this analysis, but also to serve as a foundation for future projects under the Best Counties to Live in the U.S. portfolio. A downloadable version of the dataset will be published to my personal Kaggle repository to promote transparency and encourage reuse.

To ensure a comprehensive and geographically consistent analysis, eight authoritative datasets were sourced from publicly available U.S. federal and non-profit repositories. The data spans 2019 to 2024 and offers broad coverage across socioeconomic, health, environmental, and housing indicators at the county level.

All datasets were merged using FIPS codes to maintain geographic integrity across all 3,143 U.S. counties. The final master dataset includes over 500 variables, making it well-suited for advanced spatial, statistical, and machine learning analyses.

Preprocessing & Standardization
The datasets underwent detailed preprocessing, including:

- De-duplication, renaming, and column standardization
- State-level mean imputation for missing values (by year and variable)
- Z-score normalization to compare indicators on a common scale
- Removal of unused or sensitive fields, such as margin-of-error columns

Where applicable, multi-year averages were computed to create more stable indicators. Demographic fields were consolidated into standard federal age brackets, and the Simpson Diversity Index was calculated to measure racial and ethnic diversity.
This project was developed primarily using the Health & Social Metrics dataset from the County Health Rankings & Roadmaps (UW + RWJF, 2020–2024), which served as the backbone of the analysis.

Full data sources and the cleaned dataset are available in my GitHub repository and will also be published on my Kaggle profile.

Research questions

This project seeks to uncover the key socioeconomic and environmental variables that explain disparities in healthcare access across U.S. counties. Using a curated dataset of over 500 county-level indicators, the analysis was guided by a set of questions rooted in both equity and practical decision-making.

Core Research Questions

1. Which socioeconomic and lifestyle factors most strongly predict access to healthcare at the county level?
2. Do income levels and access to exercise opportunities have a measurable impact on overall healthcare availability?
3. Are there county clusters that consistently represent underserved populations in terms of both economic and healthcare indicators?
4. Can counties be segmented into meaningful groups to inform targeted health policy or relocation decisions?

H_0, H_1 Hypotheses

The exploratory analysis and early correlation patterns led to the formulation of the following hypotheses:

H_0 (Null Hypothesis): There is no statistically significant relationship between Median Income, Access to Exercise Opportunities, or Food Insecurity and the county-level Healthcare Access Score.

H_1 (Alternative Hypothesis): Median Income and Access to Exercise Opportunities are positively associated with Healthcare Access, while Food Insecurity is negatively associated.

These questions were especially **framed around** the needs of:

- Young families seeking long-term healthcare infrastructure
- First-generation immigrants requiring public service access
- Recent graduates interested in affordability and coverage
- Empty nesters prioritizing stability and proximity to care

Analytical Approach & Methods

To address the research questions and evaluate the hypotheses, I used a multi-method analytical framework .

Double Exploratory Data Analysis

Initial data exploration involved descriptive statistics and pairwise correlations between variables of interest. A correlation matrix was produced to evaluate both within-group and cross-group relationships. The analysis revealed strong internal correlations (e.g., between child poverty and food insecurity), but generally weak cross-group associations, validating the need to build a composite healthcare score and test combined effects.

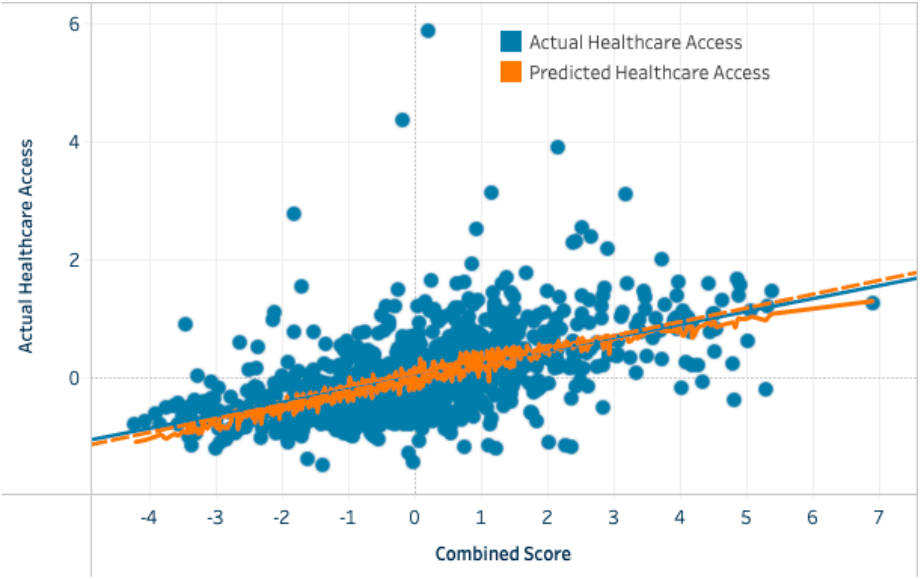
	Diversity Index	Exercise Access	Food Insecurity	Healthcare Access Score	Income Inequality	Median Income	Unemployment	Uninsured
Diversity Index	100%	18%	22%	10%	33%	6%	22%	39%
Exercise Access	18%	100%	-31%	55%	-12%	46%	1%	-26%
Food Insecurity	22%	-31%	100%	-31%	60%	-73%	55%	41%
Healthcare Access Score	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%

Sample of a correlation matrix between the Unified Healthcare Access Score and socioeconomic variables, developed during the second EDA.

Regression Modeling

A series of linear regression models were developed to test the strength and significance of key predictors, including Median Income, Exercise Access, and Food Insecurity. Four variations of the model were tested to identify the best-performing combination based on R-squared and MSE values.

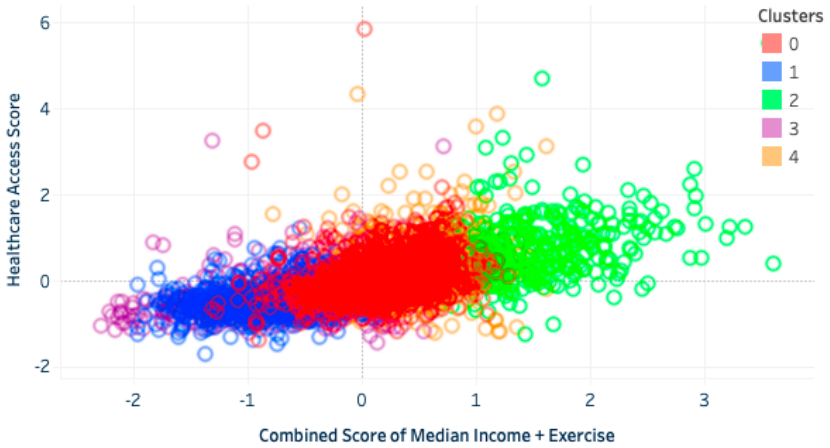
Model 2, using only Median Income and Exercise Access, achieved the best performance, by reducing MSE from 0.42 to 0.36 and nearly doubled the explained variance, with the R^2 improving from 0.16 to 0.29.



These results highlight that while income is a key predictor, the presence of community-level health infrastructure, plays a critical complementary role .

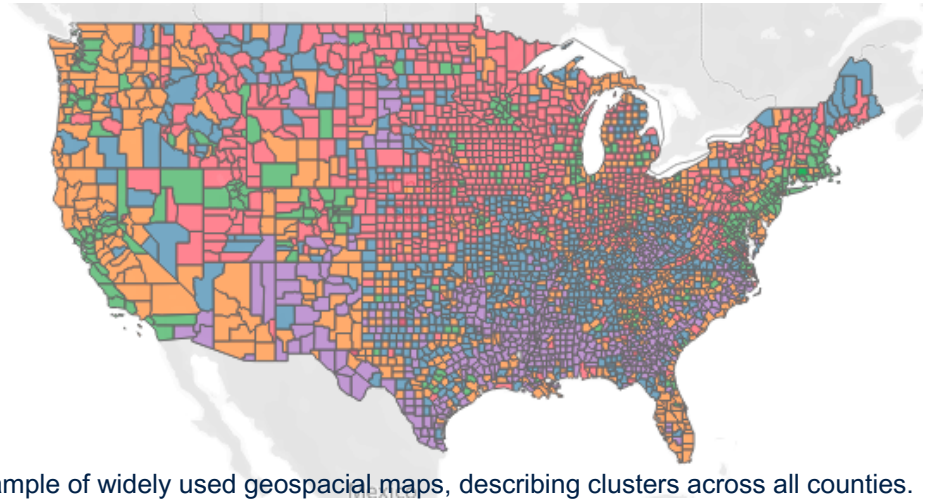
Clustering Analysis

To identify distinct groups of counties with similar profiles, K-Means clustering was applied using scaled inputs of Median Income and Exercise Access. The elbow method supported the selection of five clusters, each representing unique socio-health patterns across U.S. counties.



Geospatial Visualization

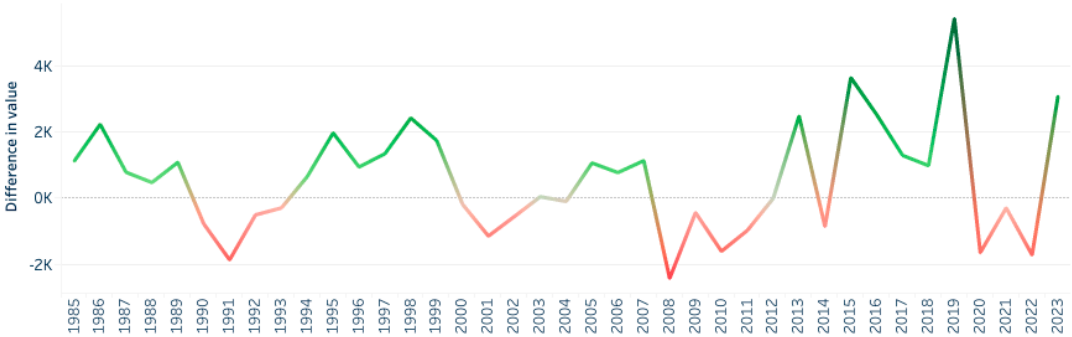
Geographic consistency was essential to understanding access disparities. Using Python and Tableau, geospatial analysis was conducted with choropleth maps that visualized composite healthcare access scores, as well as uninsured rates and income levels at the county level.



Sample of widely used geospacial maps, describing clusters across all counties.

Time-Series Exploration

As a complementary perspective, real median household income from 1985 to 2023 was analyzed to identify economic shocks that may explain structural dips in healthcare access. The analysis highlighted sharp income drops in 1991, 2009, and 2020—periods associated with national recessions and public health disruptions.



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

Key findings:

The analysis produced a range of actionable insights that help explain how access to healthcare varies across U.S. counties. These results not only validate the original hypotheses but also offer deeper clarity into structural inequalities and community strengths.

Statistical Highlights

Median Income and Access to Exercise Opportunities were the strongest predictors of healthcare access, together explaining 29% of the variation in the composite Healthcare Access Score (Model 2).

Food Insecurity, while conceptually important, did not improve model performance when added to income and exercise access variables.

Cluster-Based Findings

Five distinct clusters emerged through K-Means clustering, each representing unique patterns in healthcare access and socioeconomic status.

Red Cluster: Underserved counties with high poverty and limited access
Blue Cluster: Balanced but slightly struggling regions
Green Cluster: Prosperous and well-served communities
Purple Cluster: Unstable, working-age populations with mixed access
Orange Cluster: Modest-income areas with high uninsured rates

Economic Trends & Income Shocks

Median income time series analysis (1985–2023) revealed all sharp downturns aligned with national crises: 1991 (early recession), 2009 (Great Recession), and 2020 (COVID-19). These dips suggest that household-level income volatility may indirectly reduce healthcare access, particularly in vulnerable counties. And the most importantly together, these insights validate key predictors of access while introducing new layers of interpretability through clustering and historical trends.

Limitations

While this project offers meaningful insights into healthcare access at the county level, several limitations should be acknowledged:

Statistical Limitations

The regression model explains only 29% of the variance in healthcare access, suggesting many unmeasured factors may be influencing outcomes.
A linear model was used; more complex relationships may exist but were not captured through this method.
Results may not generalize beyond the U.S. county-level context or to specific populations without further adaptation.

Data Quality & Coverage

The analysis relies heavily on American Community Survey (ACS) and County Health Rankings data, both of which carry an estimated ±12% margin of error, particularly in smaller counties.
Some variables (e.g., insurance rates, preventive screenings) may reflect lagging indicators or vary in how consistently they are reported.
Data smoothing (e.g., 5-year averages) improves stability but may obscure short-term shocks or emerging disparities.

Variable Constraints

Only a subset of relevant indicators (13 core variables) were selected for modeling to maintain interpretability and data consistency.
Variables like healthcare provider density, transportation access, or state policy differences were not included but could enhance future models.
Despite these constraints, the analysis offers a scalable framework and a transparent methodology that can be extended, refined, or replicated in future studies based on the clean and preposesesed dataset.

Tools & Technologies

pandas, NumPy, matplotlib, seaborn, scikit-learn

Tableau: Interactive dashboards and mapping visualizations

Links to the project

GitHub link of research

Tableau link of the project

print("hey, World!")

print("let’s talk.")

e.z.safarov@gmail.com

www.linkedin.com/in/ezsafarov

https://github.com/ezsafarov