Socio-Economic Factors and U.S. County-Level Housing Prices

Exploring how income, poverty, education, and unemployment relate to housing values across U.S. counties using machine learning and statistical clustering.

Research Objectives and Methodology

Primary Goal

Link county-level median home values with key socio-economic characteristics including income, poverty rates, educational attainment and unemployment figures

Analytical Approach

Employ correlation analysis, K-means clustering, outlier detection and tree-based feature importance models to reveal underlying patterns

Geographic Focus

Examine regional inequalities and distinctive market characteristics across 3,052 US counties with complete data coverage

Data Sources: Detailed Overview

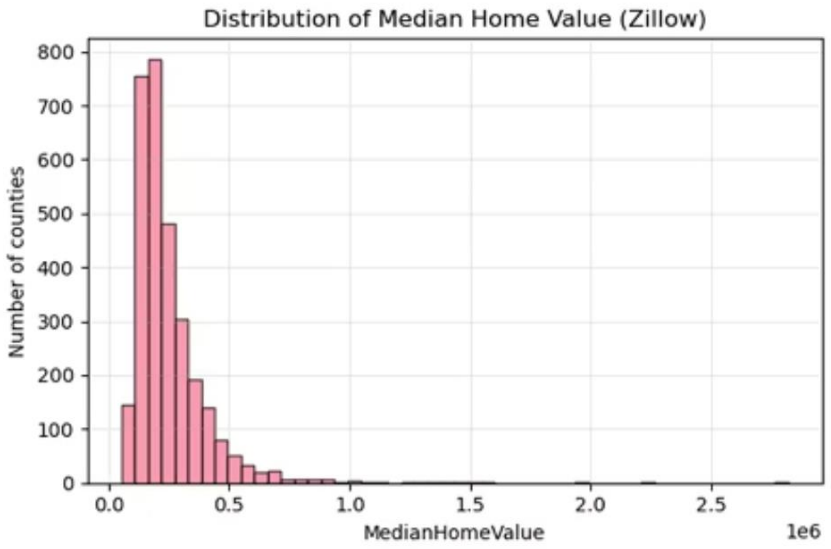
Data Source	*Description*	*Processing Approach*	*Purpose*
Zillow Home Value Index (3,073 rows)	*County-level median home values including RegionName, State, and FIPS codes*	*Downloaded CSV from Zillow research portal; cleaned FIPS codes using .str.zfill() to create uniform 5-digit identifiers; filtered for latest available date*	*Analyze housing price distribution and serve as primary dependent variable*
US Census Bureau (ACS API) (3,222 rows)	*Socio-economic data from American Community Survey including median income, population, poverty statistics, and educational attainment*	*Accessed via Census API with authentication; decoded variable names (B19013_001E, B15003_022E, etc.); calculated derived metrics (poverty rate, college-educated percentage); standardized FIPS codes*	*Examine correlations between housing prices and socio-economic factors*
Bureau of Labor Statistics (LAUS) (3,225 rows)	*County-level unemployment data , including unemployment rate and labor force statistics*	*Downloaded Excel file from BLS website; cleaned column names and filtered for 2022 data; standardized state and county FIPS codes; merged with other datasets using FIPS keys*	*Investigate relationship between labor market conditions and housing markets*

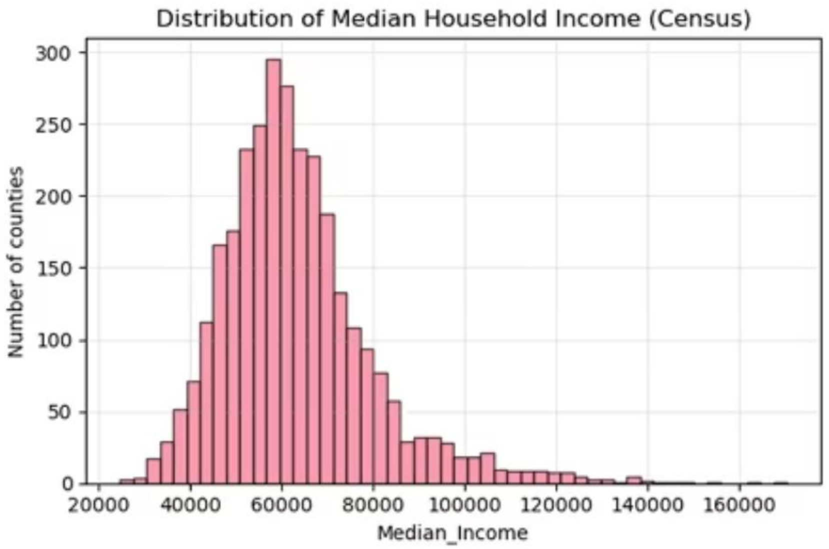
After merging datasets by county FIPS codes and cleaning missing values, the final analytical dataset comprises 3,052 counties with complete information across all variables.

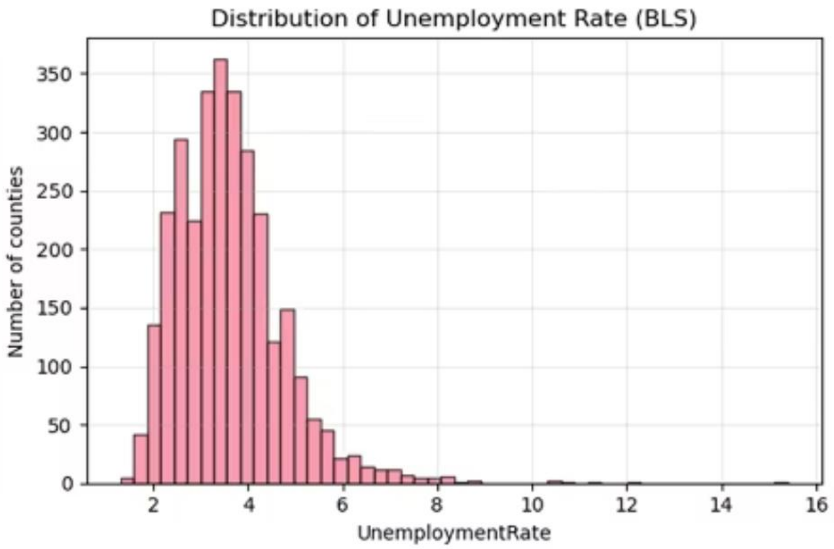
Descriptive Overview of US Counties in 2022

Descriptive statistics reveal extraordinary diversity in housing markets and socio-economic conditions across America's 3,052 counties.

$248K	$63.4K	13.7%	3.6%
Mean Home Value	**Mean Income**	**Mean Poverty Rate**	**Mean Unemployment**
Range: $51K – $2.82M (highly right-skewed distribution)	*Range: $24.6K – $170.5K across counties*	*Range: 1.7% – 43.2% showing stark disparities*	*Range: 1.3% – 15.4% across labour markets*


Distribution of Median Home Value (Zillow)


Distribution of Median Household Income (Census)


Distribution of Unemployment Rate (BLS)

Clear Relationships Between Housing and Socio-Economic Factors

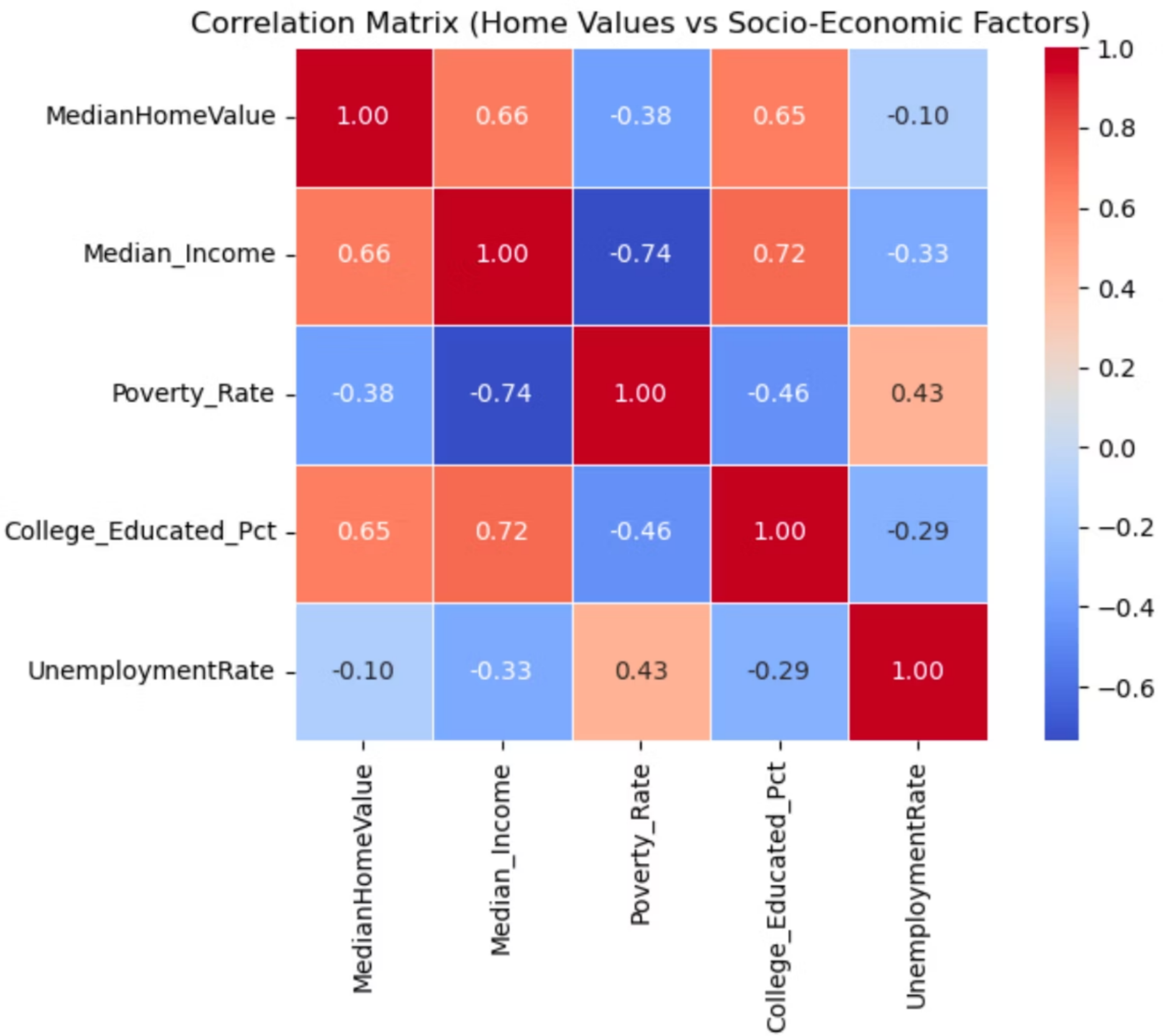
Correlations with Median Home Value

- *Median home value has a strong positive correlation with median household income (r ≈ 0.66) and the share of college-educated adults (r ≈ 0.65).*

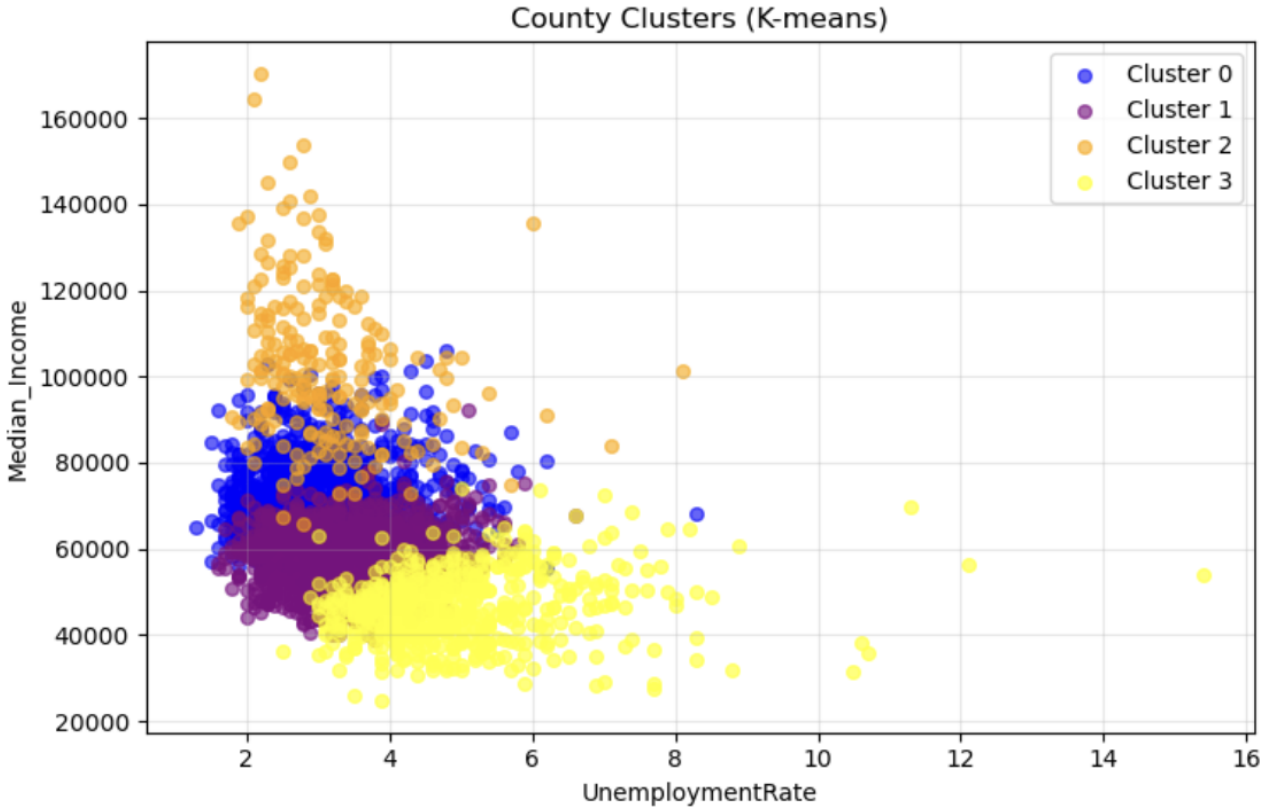
- *Poverty rate shows a moderate negative correlation with home values (r ≈ −0.38).*

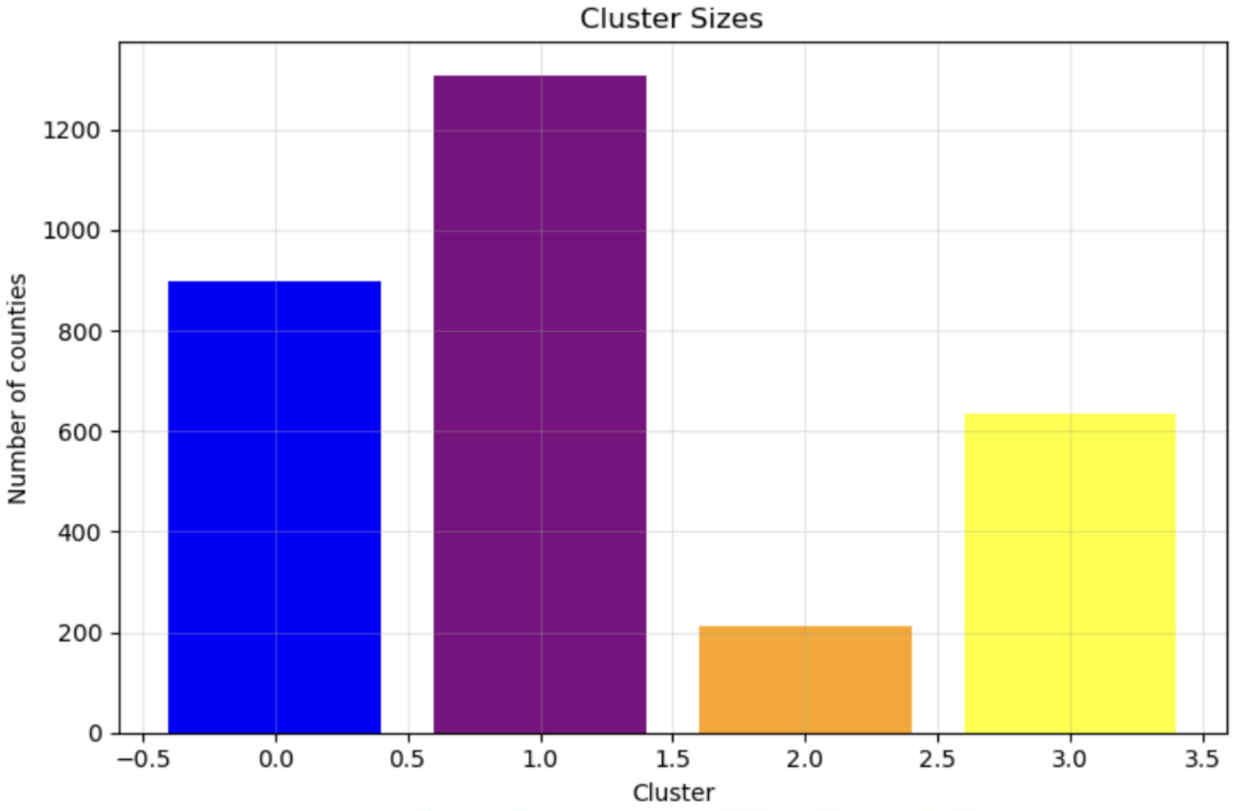
- *Unemployment rate has only a weak negative correlation (r ≈ −0.10).*

Overall, richer and more educated counties tend to have much higher home values, while unemployment alone explains very little.

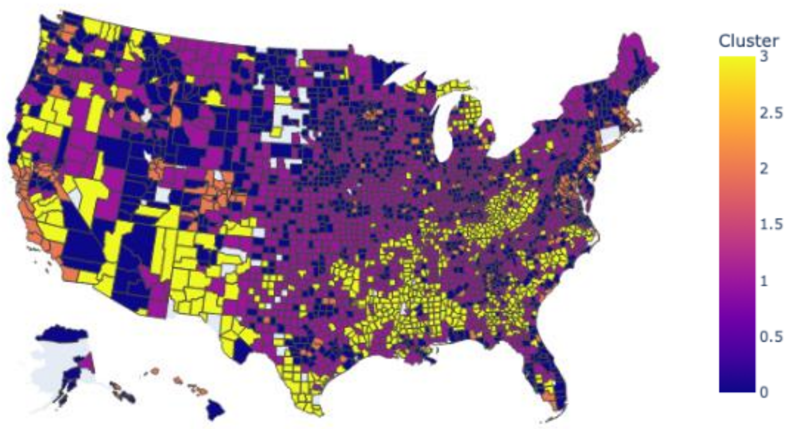


Four Distinct Socio-Economic Profiles (K-means)



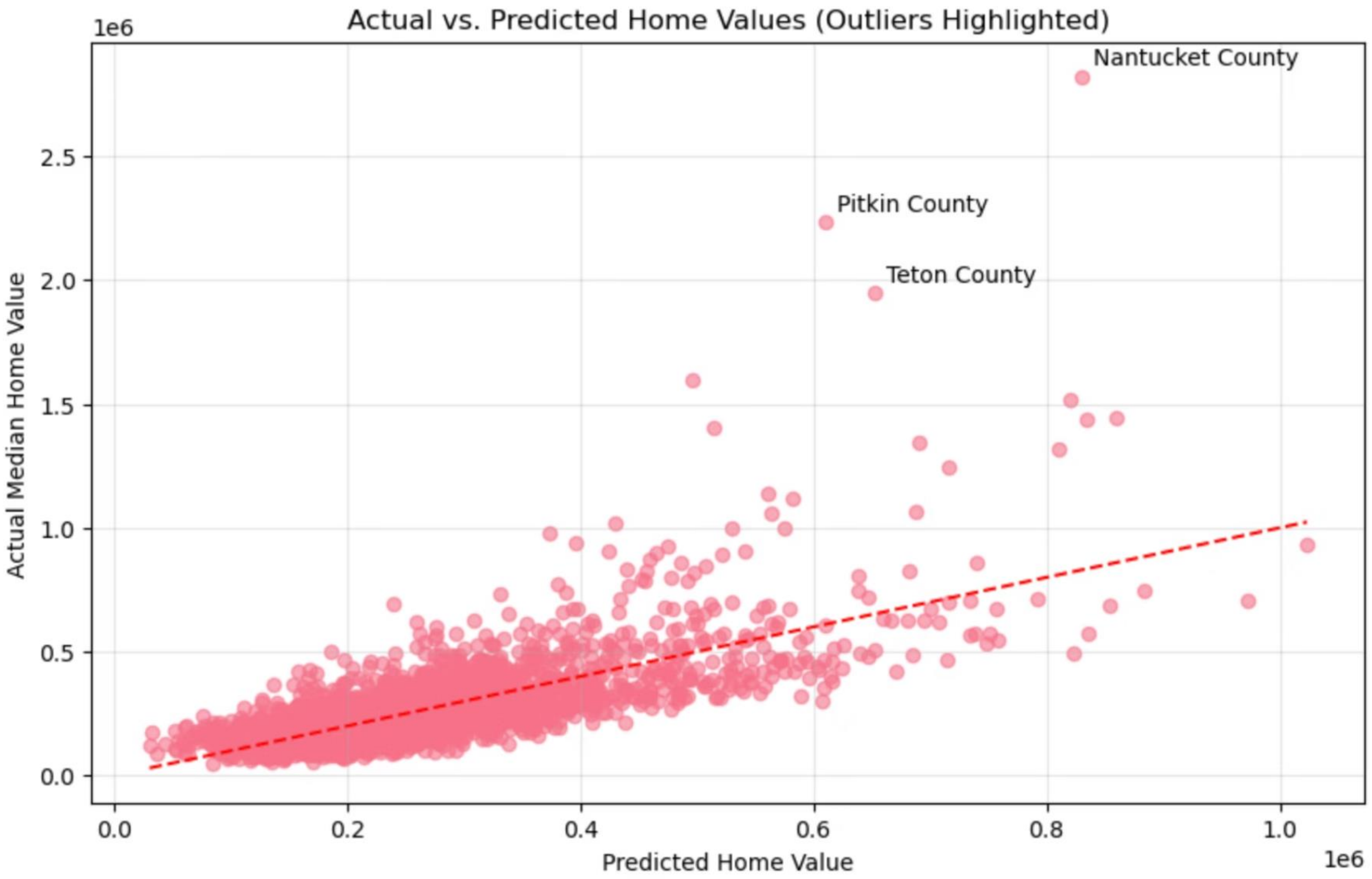


County Clusters: Housing & Socio-Economic Profiles



- *Cluster 0 - Prosperous Counties (898 counties):*
  - *Above average income ($72K), good home values ($292K)*
  - *Moderate education (29% college+), low poverty (9.6%)*
- *Cluster 1 - Average Counties (1,308 counties):*
  - *Middle income ($58K), moderate home values ($179K)*
  - *Average education (19% college+), typical poverty (13.5%)*
- *Cluster 2 - Affluent Counties (212 counties):*
  - *High income ($100K+), expensive homes ($545K+)*
  - *Highly educated (45% college+), low poverty (7%)*
- *Cluster 3 - Struggling Counties (634 counties):*
  - *Low income ($47K), affordable homes ($150K)*
  - *Low education (15% college+), high poverty (20.6%)*

Market Exceptions and Key Insights



Top Overpriced Counties:

- Nantucket, MA: +$1.99M above predicted
- Pitkin, CO (Aspen): +$1.62M above predicted
- Teton, WY: +$1.29M above predicted

Key Conclusions:

- Income and education are the strongest predictors of county-level home values.
- There is clear economic stratification in the US housing market, with four distinct county profiles.
- Some local markets, especially tourist and resort areas, follow “different rules” and are priced far above socio-economic fundamentals.
- Regional disparities in both housing affordability and socio-economic conditions are clearly visible.

Challenges and Lessons Learnt

Data Access and Cleaning

- *BLS county unemployment file required manual download due to 403 errors*
- *Careful FIPS code formatting and alignment across Zillow, Census, and BLS sources*
- *Substantial data engineering effort to merge disparate public datasets*

Statistical Challenges

- *Managing highly skewed distributions, especially for home values*
- *Handling outliers whilst preserving meaningful variation*
- *Balancing statistical rigour with interpretability*

Methodological Decisions

- *Selecting optimal number of clusters through elbow method and domain judgement*
- *Iterating on cluster interpretation to ensure meaningful, actionable insights*
- *Combining statistical analysis with careful qualitative assessment*

This project demonstrated the importance of robust data engineering when working with real public data and the value of combining multiple analytical approaches to understand complex socio-economic relationships.

Thank You!

Questions & Discussion