Prepared by Group 7: Mutsa Mungoshi

Structural Cost Burdens and SNAP

A Multivariate County-Level Analysis

IA 650: Data Mining

28 July 2025

Introduction

- The Supplemental Nutrition Assistance Program (SNAP) provides monthly food benefits to low-income households.
- Eligibility is based on income and household size
- Benefit levels follow federal standards (e.g., Thrifty Food Plan).
- Participation rates vary widely across counties.
- This study explores whether those differences reflect deeper cost structures.

Project Objectives

Analysis Phase

Examine how SNAP participation relates to structural cost burdens using descriptive statistics, correlation, ANOVA, and Chi-Square tests.

Strategy Development

Reduce dimensionality with PCA and apply both K-means and Hierarchical Clustering to identify and compare county-level cost profiles.

Implementation Plan

Visualize clusters geographically, evaluate group differences, and extract insights to inform geographically responsive policy design.

Data Sources and Variables

- Data Sources:
    - SNAP participation: U.S. Census SAIPE
    - Cost estimates: MIT Living Wage Calculator
    - Demographics & income: U.S. Census ACS
- Key Variables:
    - SNAP Rate — % of population receiving SNAP benefits
    - Median Family Income (in $/month)
    - Metro Status — metro vs. non-metro classification
    - Cost Components (in $/month):
        - Food, Housing, Transportation, Healthcare, Childcare, Taxes

Data Preprocessing

Aggregation
Combined cost and demographic data at the county (FIPS) level, averaging across household types.

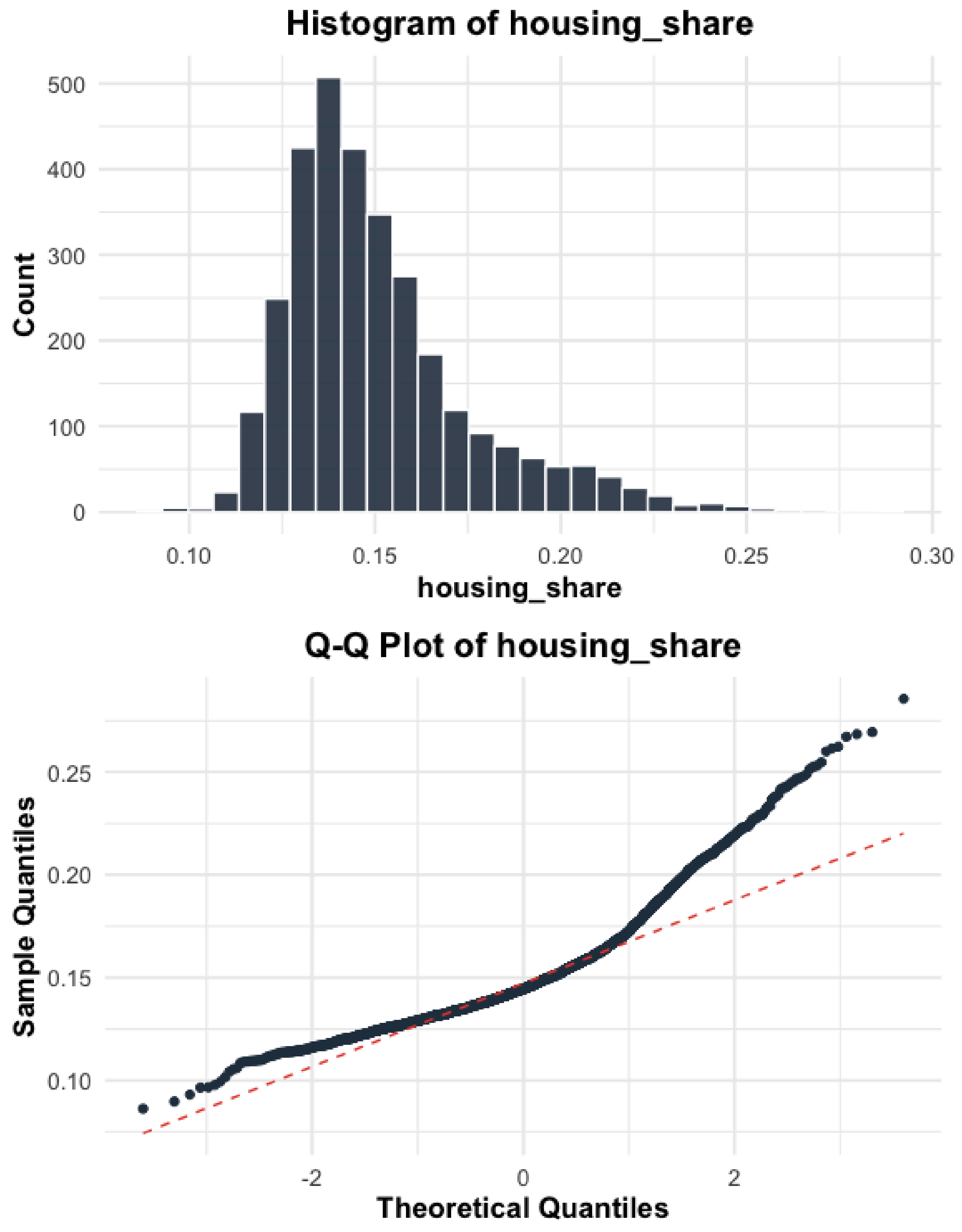
Data Cleaning
Removed missing and inconsistent values, and ensured alignment of SNAP, income, and cost entries by county

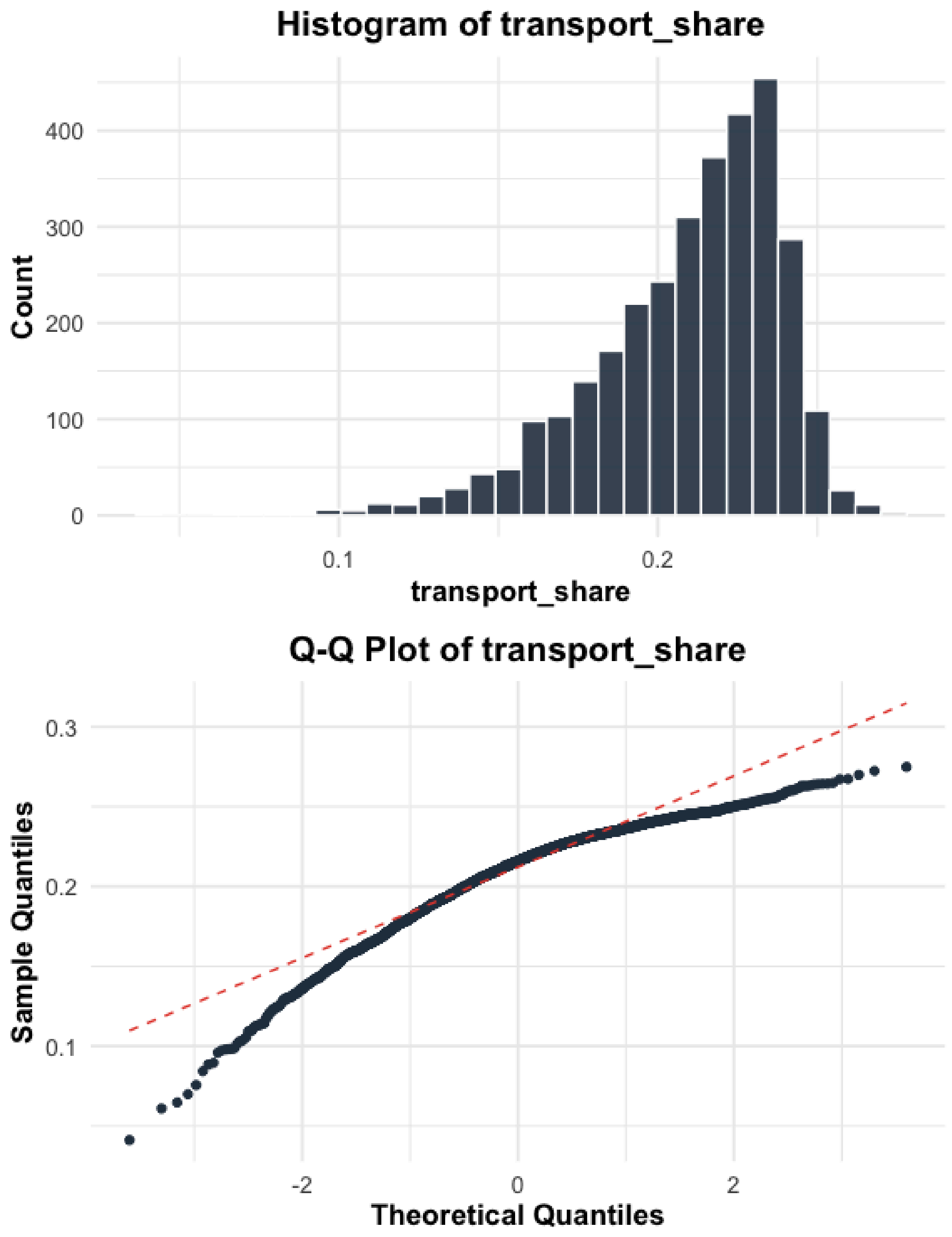
Scaling
Standardized all numeric predictors before PCA and clustering to ensure comparability.

Derived variables
Calculated cost shares (cost component / total monthly cost) as well as log-transformed skewed shares

Exploratory Data Analysis

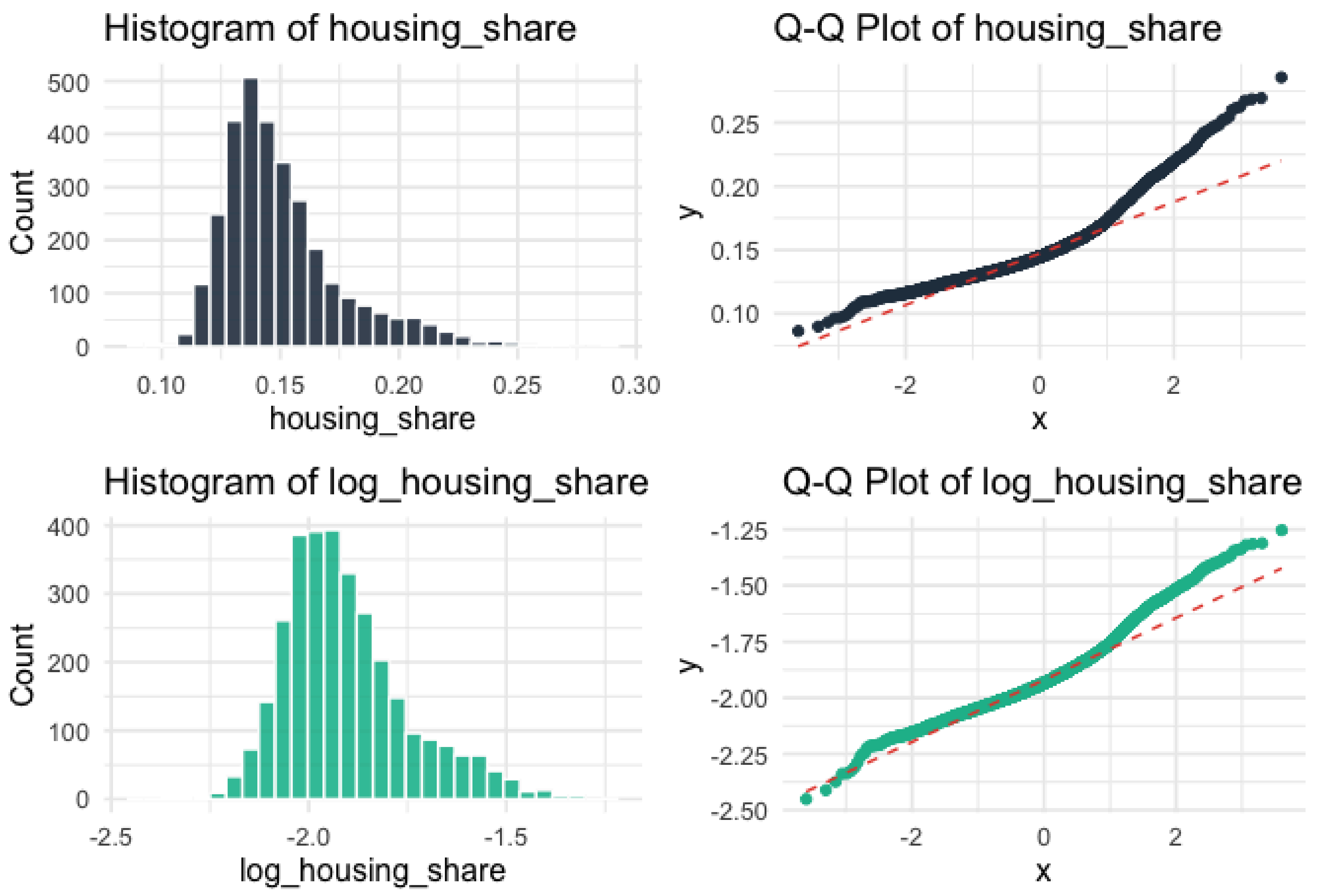
Example of two variables that we determined to be skewed, thus justifying the addition of log transformations for more reliable structure detection.





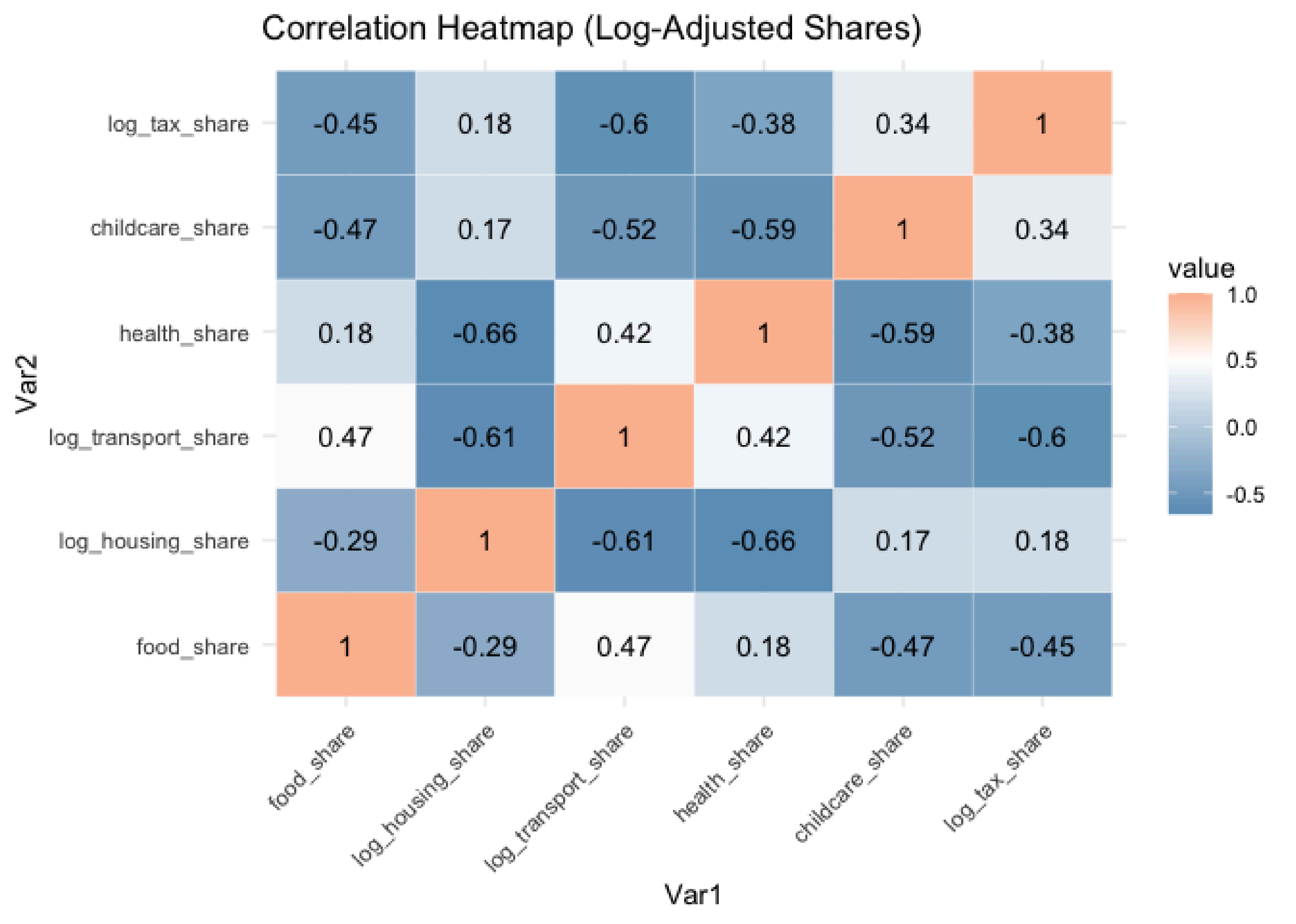
Exploratory Data Analysis (cont.)

Example of the effect of log transformations:



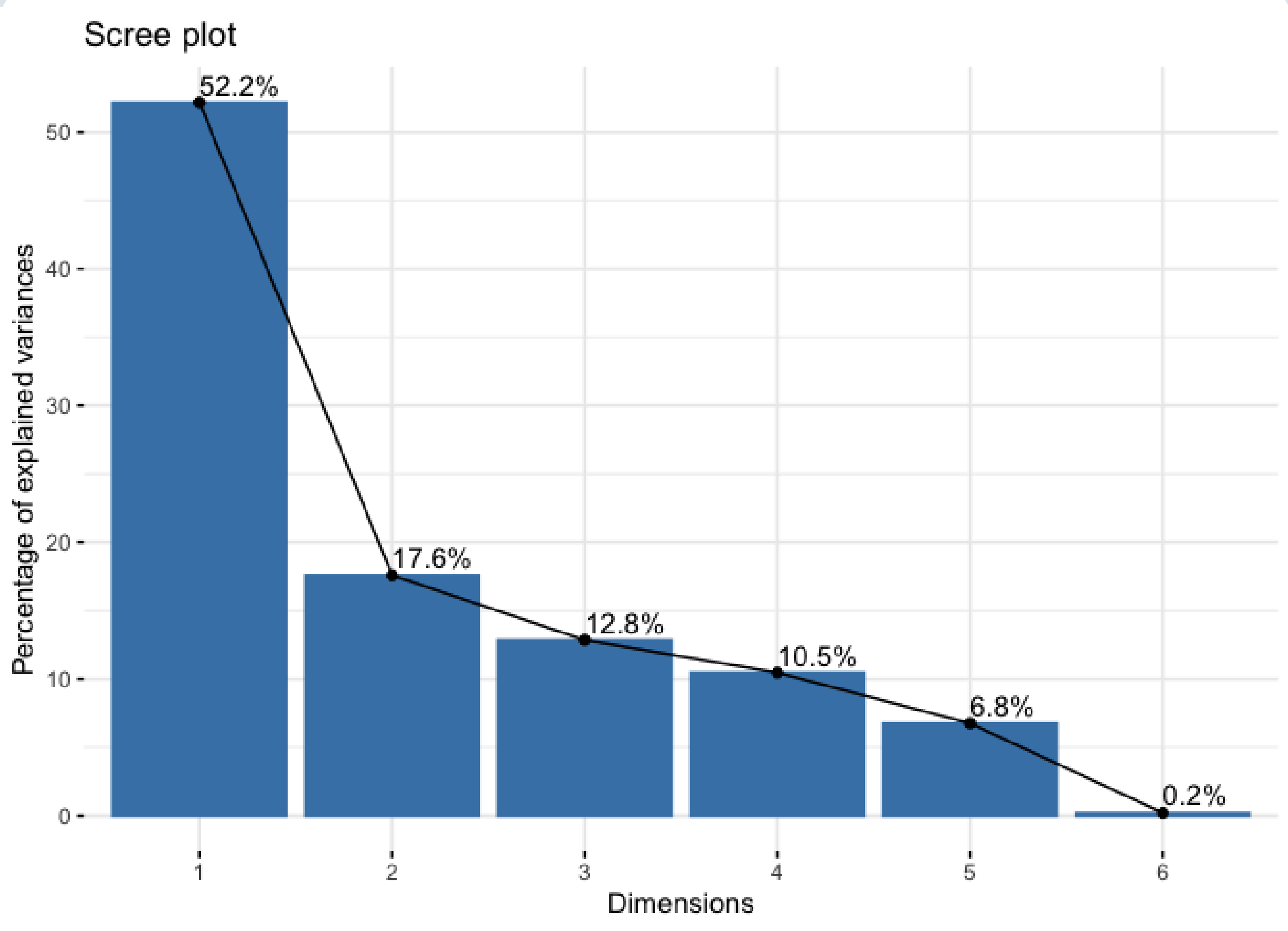
Correlation Analysis

- Displays Pearson correlations among log-adjusted cost categories
- Housing vs. transport: strong negative correlation (r = –0.61)
- Housing vs. health: strongest negative correlation (r = –0.66)
- Suggests cost tradeoffs across major categories
- Indicates multicollinearity, supporting PCA for dimension reduction



Principle Component Analysis

- PC1 (52.2% variance):
  - Captures overall cost burden; food/transport/health vs. housing/childcare/taxes
- PC2 (17.6% variance):
  - Represents a housing/food vs. health/taxes burden axis; Counties with high housing shares tend to have lower health burdens, and vice versa.
- PC3 (12.8% variance):
  - Represents a transport vs. childcare axis. Counties with high transport shares tend to have lower childcare shares, and vice versa.

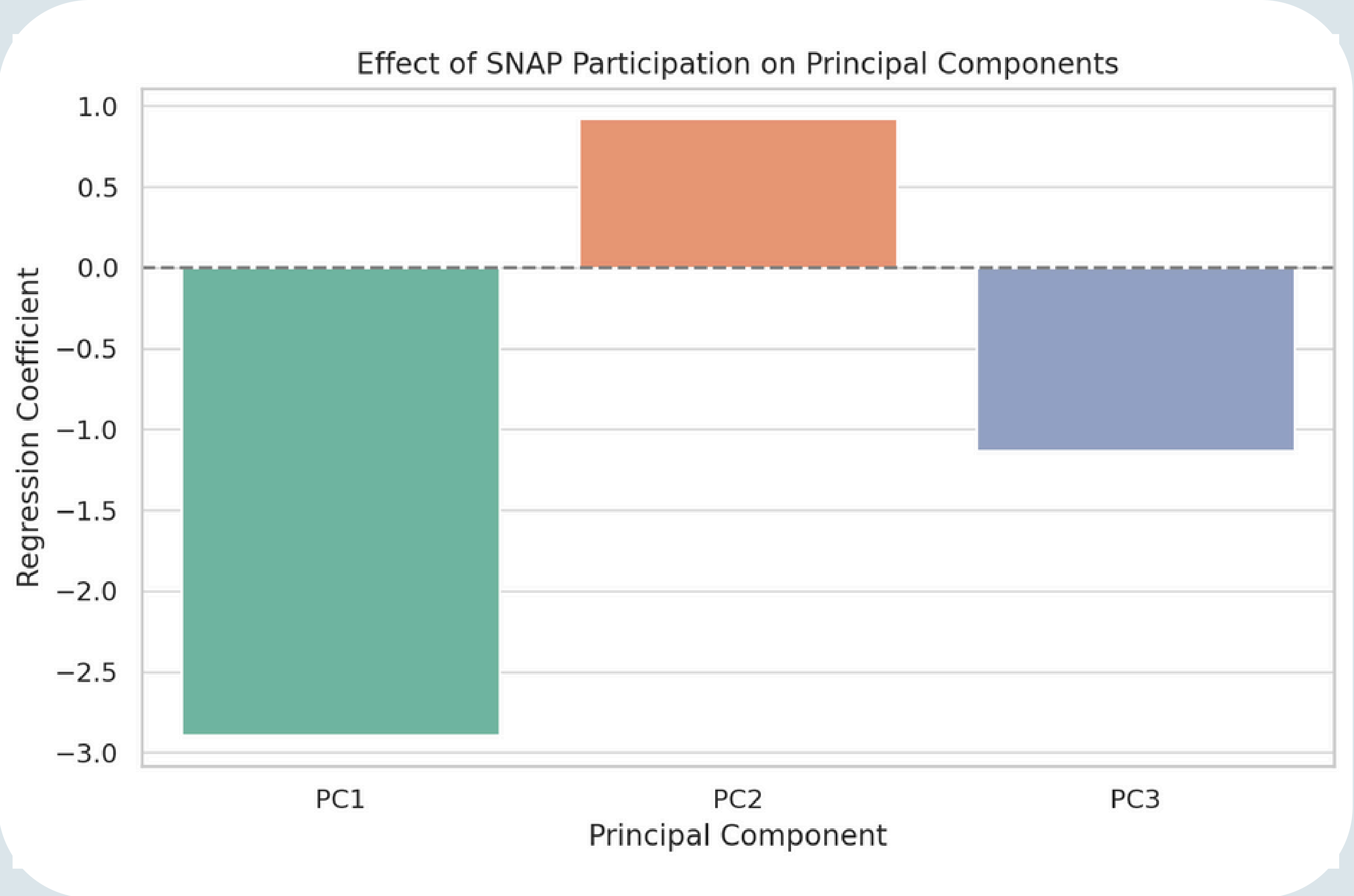


PCA Summary

Importance of components	**PC1**	**PC2**	**PC3**
Standard deviation	1.7691	1.0271	0,8779
Proportion of Variance	0.5216	0.1758	0.1285
Cumulative Proportion	0,5216	0,6974	0,8259
food	0.36	0.48	0.17
log_housing_share	−0.38	0.59	−0.40
log_transport_shhare	0.48	0.05	0.31
health_share	0.42	−0.47	−0.37
childcare_share	−0.41	−0.18	-0.72
log_tax_share	−0.38	−0.40	-0.23

Regression

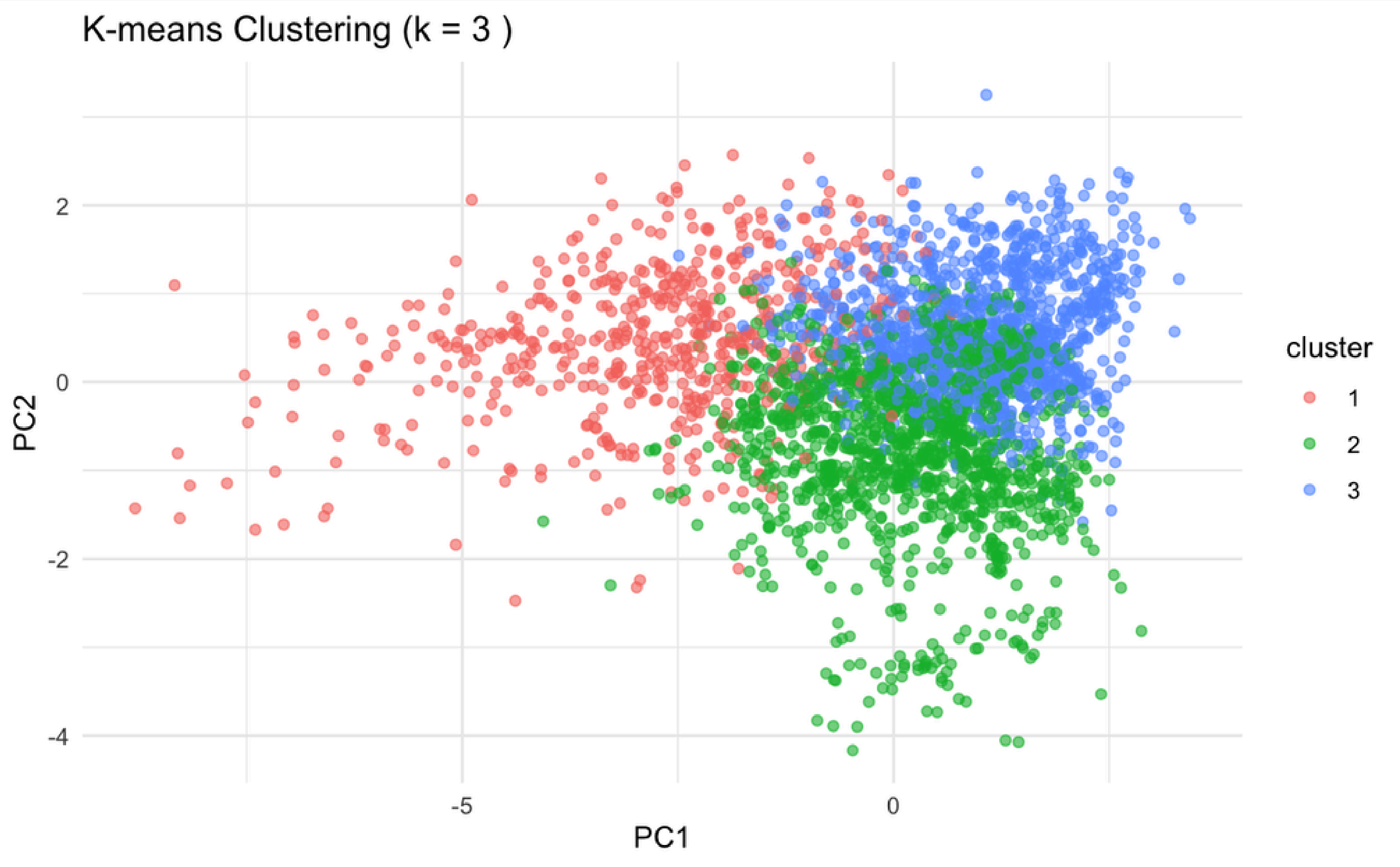
- PC1 (General Cost Burden):
  - SNAP participation is significantly higher in counties with elevated burdens from housing, childcare, and taxes, and lower shares of food and transportation.
  - Regression shows a strong negative association with SNAP (PC1 ~ SNAP: $\beta = -2.89$, $p < 0.001$).
- PC2 (Housing vs. Basic Necessities):
  - SNAP is more common in counties where housing costs dominate over food and health expenditures, reflecting tradeoffs made by low-income households.
  - Regression confirms a positive SNAP coefficient ($\beta = +0.92$, $p < 0.001$).
- PC3 (Childcare Burden Axis):
  - SNAP participation tends to be higher in counties with lower childcare shares, suggesting that other burdens—such as housing or taxes—may be more acute.
  - SNAP shows a significant negative effect ($\beta = -1.13$, $p < 0.001$).

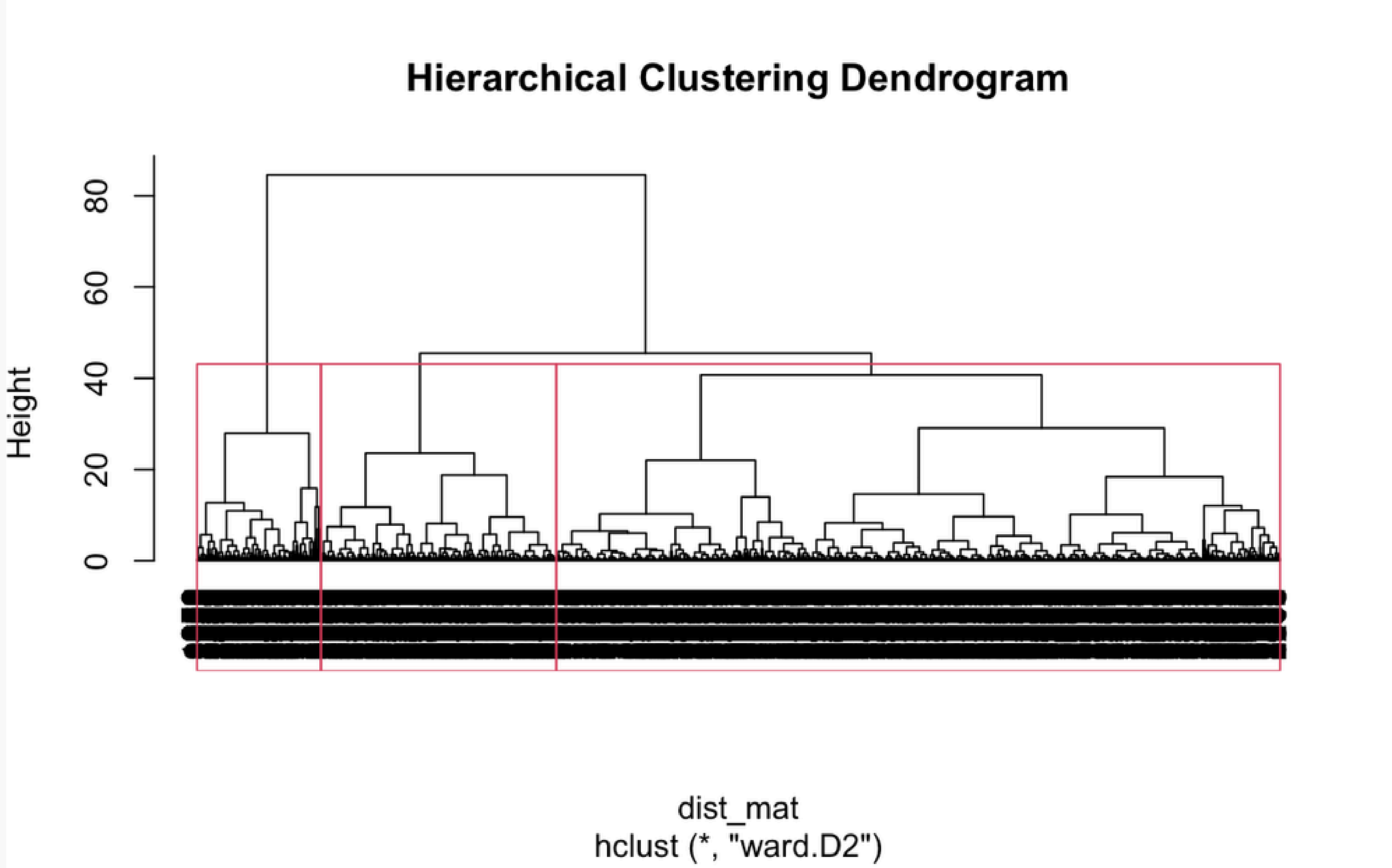


PC	SNAP Coefficient	Std. Error	t-statistic	p-value	R^2
PC1	-2.89	0.3	-9.66	< 0.000001	0.803
PC2	0.92	0.22	4.19	< 0.0001	0.688
PC3	-1.13	0.22	-5.22	< 0.000001	0.583

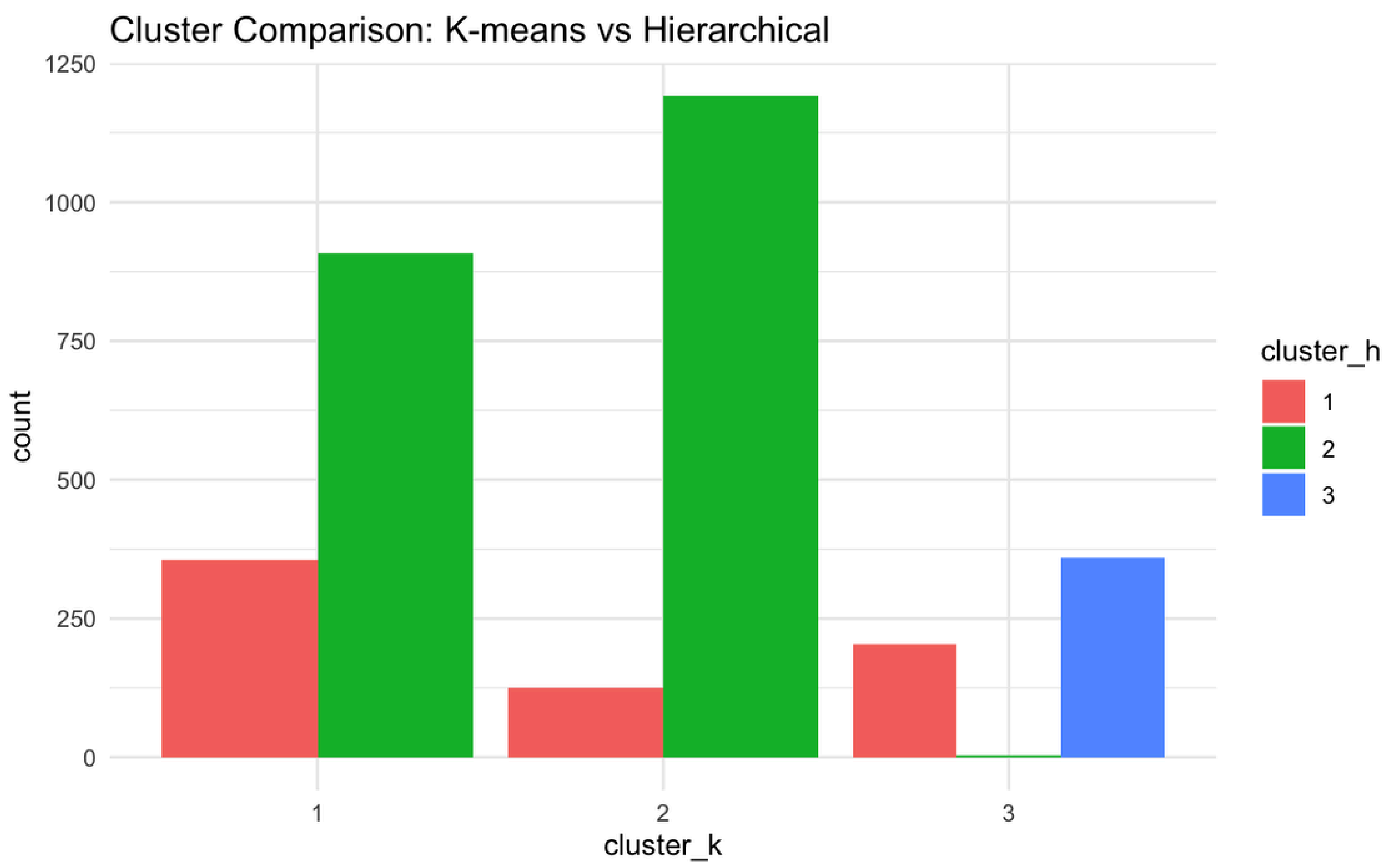
Clustering Techniques

- K-Means Clustering
  - Partitioned counties into cost structure groups based on log-adjusted shares
  - Optimal number of clusters selected via elbow method and silhouette scores
  - Captured counties with similar overall cost composition patterns
- Hierarchical Clustering (Complete Linkage)
  - Built a dendrogram to visualize nested similarities





Clustering Comparison

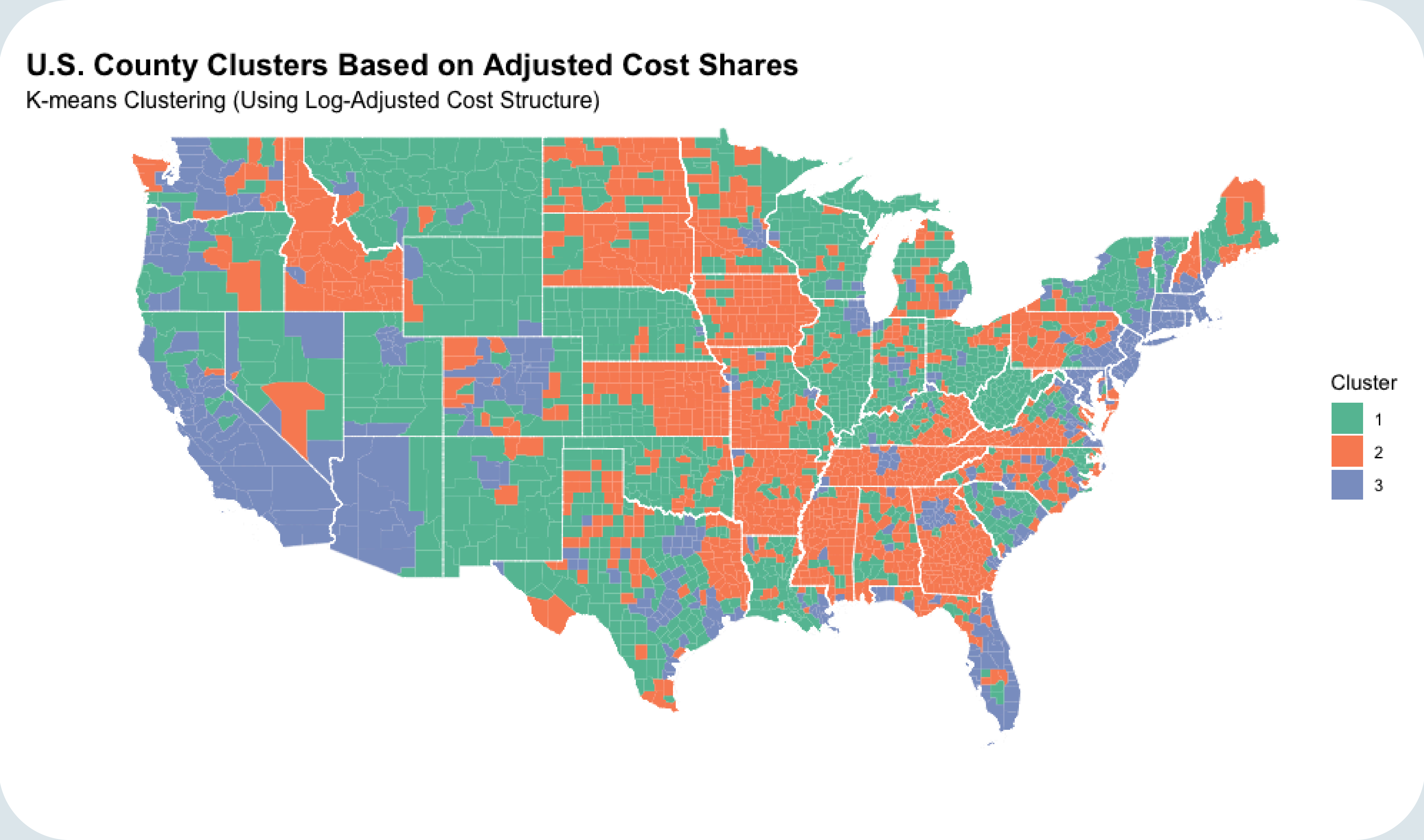


Why I Chose K-means

- K-means gave clearer, more interpretable clusters.
- Cluster patterns made more sense geographically.
- The results aligned better with known cost and SNAP patterns.
- Hierarchical split similar counties in confusing ways.
- K-means let me explore different k values more easily.

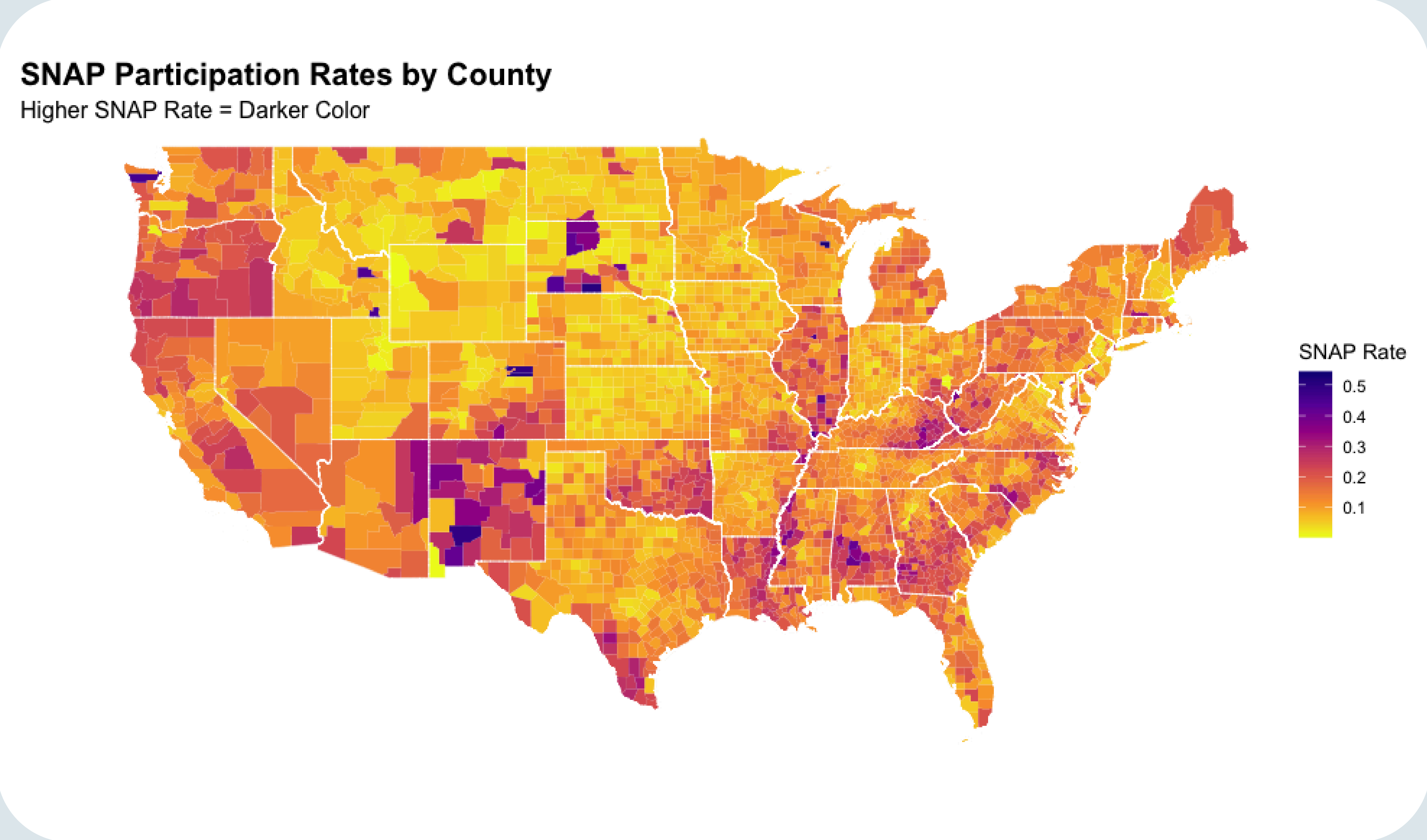
Observations

- Cluster 1 (Green): Balanced cost shares, mostly Midwest and Mountain West.
- Cluster 2 (Orange): Higher food/transport shares, common in South and rural areas.
- Cluster 3 (Purple/Blue): High housing/tax burden, concentrated in coastal and metro areas.
- Highlights geographic cost structure variation even after adjusting for total spending.
- Supports idea that place matters—similar incomes can mean different burdens by region.

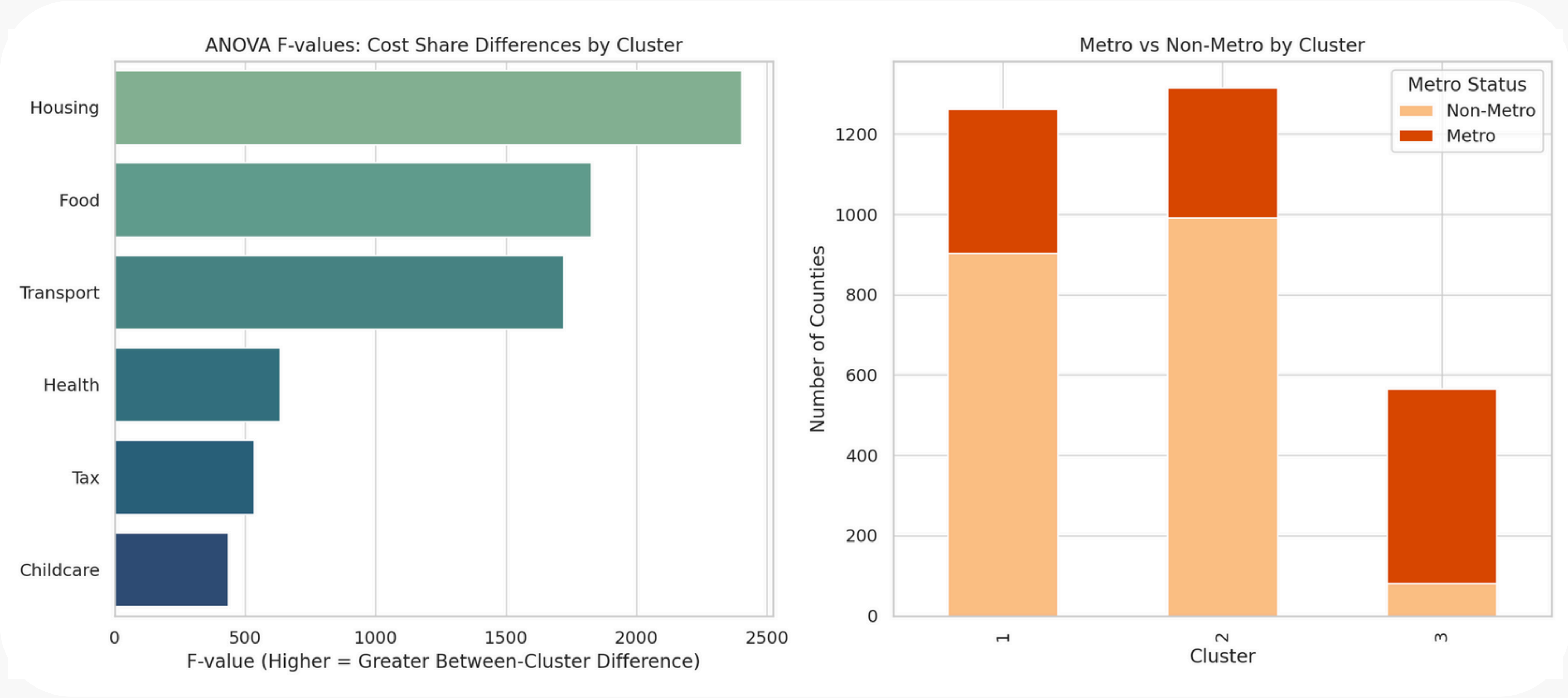


Observations (cont.)

- Higher SNAP Rate = Darker Color
- Darkest areas (e.g., Deep South, Appalachia, Southwest) show highest SNAP usage.
- High rates often overlap with persistent poverty or rural isolation.
- Lower rates seen in Midwest, Plains, and suburban coastal regions.
- Highlights how SNAP participation is unevenly distributed, reflecting both need and policy access.
- Provides spatial context for analyzing how SNAP relates to regional cost burdens.



ANOVA and Chi-squared test results



ANOVA Results:

- All cost categories vary significantly across clusters ($p < 0.001$).
- Housing and transport show the largest between-cluster differences (highest F-values).
- Food, health, childcare, and tax shares also differ, though less sharply.

Chi-Square Test:

- Strong association between metro status and cluster assignment ($p < 2.2e-16$).
- Cluster 3: Mostly metro counties.
- Clusters 1 & 2: Predominantly non-metro.

Statistical Techniques Applied

- Principal Component Analysis (PCA):
  - Reduced six correlated cost share variables into three uncorrelated principal components representing key cost structure dimensions.
- Multiple Linear Regression:
  - Modeled each PCA component as a function of SNAP rate, income, metro status, and state effects to explore how cost structures predict SNAP participation.
- K-Means & Hierarchical Clustering:
  - Grouped counties into cost-type clusters based on PCA scores; compared clustering results and validated group separation using ANOVA and Chi-square tests.
- ANOVA (Analysis of Variance):
  - Tested for significant differences in cost share distributions across clusters (all $p < 0.001$), confirming distinct cost profiles.
- Chi-Square Test:
  - Assessed association between metro status and cluster membership ($\chi^2 = 699.24$, $p < 0.001$), indicating clustering aligns with metro-rural divide.

Conclusion

- SNAP participation is linked not only to income but to structural cost burdens—especially housing, transport, and childcare.
- PCA uncovered interpretable dimensions of cost variation, enabling cleaner analysis and pattern discovery.
- K-means clustering revealed three distinct county types, differing by metro status and cost profiles.
- ANOVA confirmed that these clusters meaningfully differ across all cost categories ($p < 0.001$).
- Chi-square tests show cost structure clusters align with urban-rural divides.
- Policy implication: Effective support must consider local cost structure, not just income level—SNAP works best when paired with housing, transport, and childcare assistance tailored to regional burdens.
- Important caveat: Findings reflect correlations, not causation — SNAP may respond to cost burdens or co-occur with them, but direct effects are not established.

Future Scope

Temporal Analysis
Incorporate time-series or panel data to track changes in SNAP and cost burdens over time.

Program Interaction Effect
Examine how SNAP intersects with other safety nets like Medicaid or housing subsidies.

Finer Cost Granularity
Disaggregate cost categories further (e.g., rent vs. utilities, public vs. private childcare).

Predictive Modeling
Build models to forecast SNAP demand based on local cost structures and economic shocks.

Acknowledgements

Sumona Mundol

Professor

Naveen Ramachandra Reddy

Assistant Professor

Thank you