Decoding the Digital Divide:

Using Autoencoders to Uncover Latent Factors Driving Broadband Adoption Across U.S. ZIP Codes

Hari Narayanan

December 2025

Abstract

This study analyzes broadband adoption patterns across U.S. ZIP codes using American Community Survey (ACS) 2019-2023 data. By training a hybrid autoencoder on 15 demographic features, I identified two latent dimensions that exhibit perfectly monotonic relationships with broadband adoption rates—enabling deterministic ranking of ZIP codes by demographic risk without infrastructure data. These latent vectors capture distinct mechanisms: *rural composite disadvantage* (z6), where population size serves as a marker for bundled socioeconomic challenges, and *direct device poverty* (z3), where lack of computing devices directly predicts non-adoption. Together, these dimensions explain 53% of variance in adoption rates. The approach provides a first-pass triage mechanism for identifying at-risk ZIP codes where infrastructure data is unavailable or lagged.

1. Introduction

The digital divide remains one of the most persistent equity challenges in the United States. While policy discussions often focus on infrastructure availability, the question of *adoption* is equally critical. Even where broadband is available, significant portions of the population remain unconnected.

This study takes a novel approach: rather than modeling adoption as a function of known demographic variables directly, I use an autoencoder to learn compressed latent representations of ZIP code demographics, then examine which latent dimensions correlate most strongly with adoption outcomes. A linear or tree-based regression model would identify demographic predictors of adoption, but would not reveal whether those predictors operate as direct causes or as markers for underlying composite disadvantage. The autoencoder approach enables this distinction.

2. Data and Methodology

2.1 Data Sources

The analysis uses ZIP code-level data from the American Community Survey (ACS) 5-year estimates (2019-2023). Fifteen demographic features were selected, deliberately **excluding** the target variable (broadband adoption rate) to prevent data leakage.

2.2 Model Architecture

A **hybrid autoencoder** was trained to compress the 15-dimensional input space into a 6-dimensional latent space (z1 through z6). The architecture includes an encoder (Input → Hidden 64 → Latent 6), decoder (Latent 6 → Hidden 64 → Output), and prediction head (Latent 6 → Hidden 32 → Adoption Rate). *The prediction head regularizes the latent space toward adoption-relevant structure, preventing the encoder from learning purely reconstructive but policy-irrelevant representations.*

The model achieved **R^2 = 0.53** on the validation set, explaining 53% of variance in adoption rates.

2.3 Latent Selection Criteria

To identify which latent dimensions meaningfully relate to adoption, strict monotonicity criteria were applied: Spearman correlation ≥ |0.99|, zero reversals across deciles, and meaningful spread (>25 percentage points). Only **two latent dimensions (z3 and z6)** passed all criteria. The remaining four latents (z1, z2, z4, z5) failed due to insufficient spread, non-monotonicity, or both.

3. Results

3.1 Latent z6: Rural Composite Disadvantage

Latent z6 exhibits the strongest relationship with adoption, with a spread of **32.5 percentage points** between lowest and highest deciles (85% → 52% median adoption).

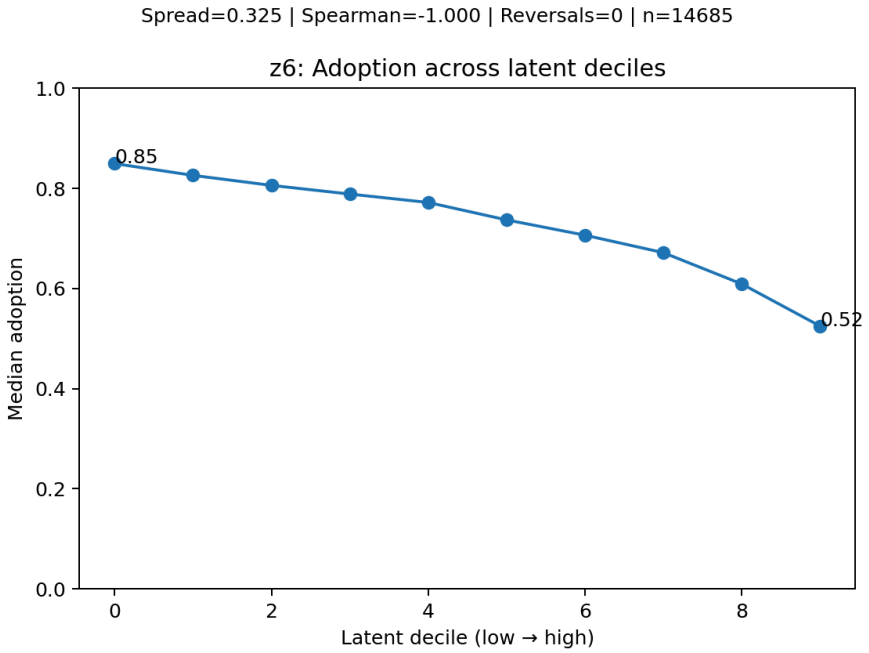

Figure 1: Broadband adoption rate across z6 latent deciles (n=14,685 ZIP codes). Perfect monotonic decrease (Spearman = -1.0).

What this means: Adoption drops below 70% starting at decile 7, **affecting approximately 4,400 ZIP codes** (deciles 7-9). These areas show compounding disadvantages: the top features correlated with z6 are total_households (0.59 dominance), total_population (0.52), pct_no_large_screen (0.49), and pct_bachelor_plus (0.45).

Key Insight: Population as Marker, Not Cause

While z6 correlates strongly with population size, the scatter plot below reveals that population itself has *no univariate relationship* with adoption—the trend line is essentially flat. The autoencoder has learned that a small population is a *marker* for a bundle of compounding disadvantages that collectively drive low adoption.

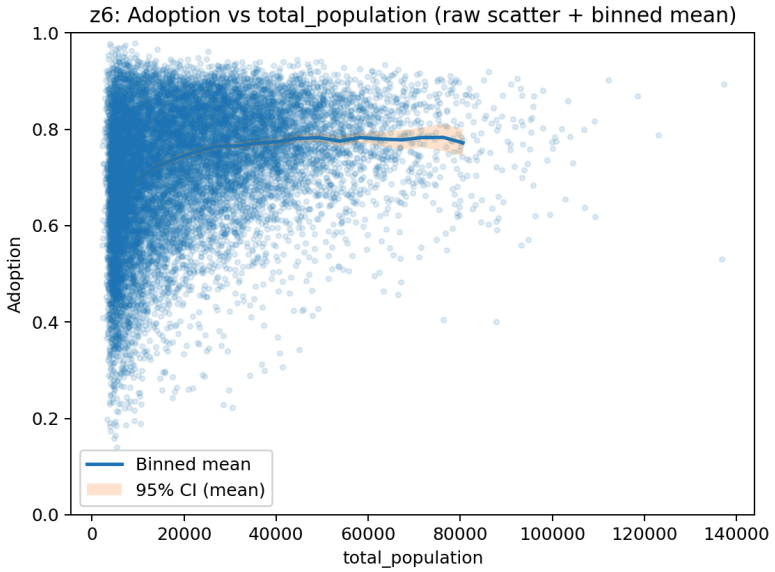

Figure 2: Adoption vs. total population shows no clear univariate relationship despite population being the top z6 correlate.

We verified that nonlinear transformations of population (log-scaling, polynomial terms) do not produce meaningful univariate relationships with adoption, confirming that the z6 signal is genuinely composite rather than a nonlinear population effect.

3.2 Latent z3: Direct Device Poverty

Latent z3 captures a purer socioeconomic dimension, with a spread of **26.7 percentage points** (87% → 61% median adoption).

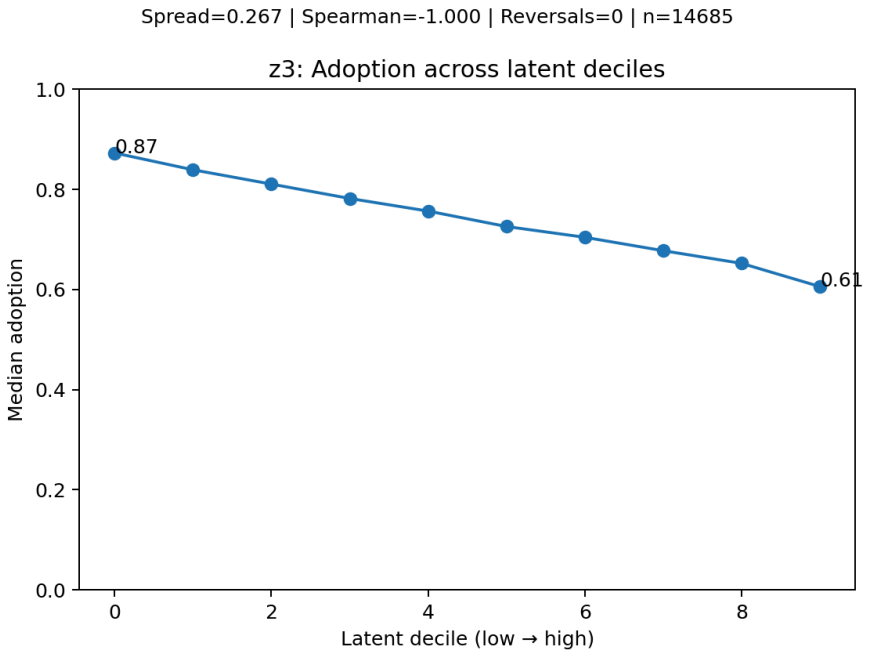

Figure 3: Broadband adoption rate across z3 latent deciles. Perfect monotonic decrease (Spearman = -1.0).

What this means: Adoption drops below 70% starting at decile 7, **affecting approximately 4,400 ZIP codes**. The top features correlated with z3 are pct_no_large_screen (0.60 dominance), median_income (0.58), pct_bachelor_plus (0.50), and poverty_rate (0.44). Unlike z6, these features show **strong univariate relationships** with adoption.

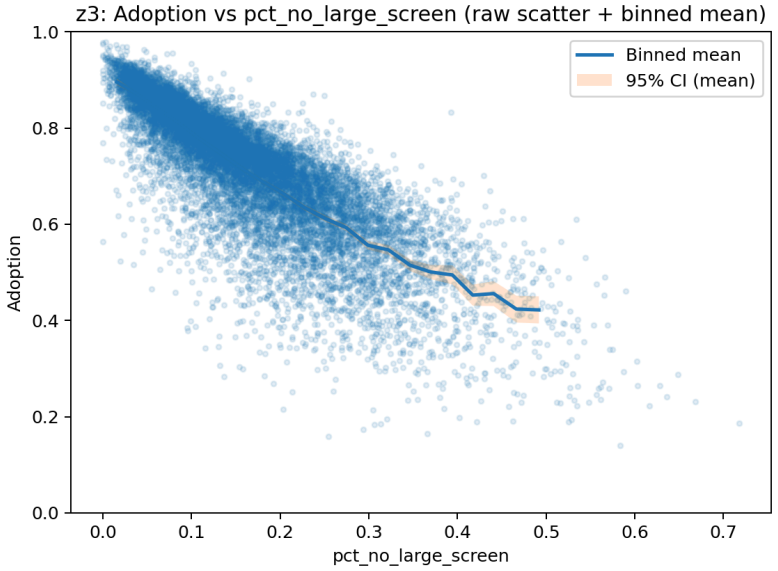

Figure 4: Strong negative relationship between device poverty (pct_no_large_screen) and adoption.

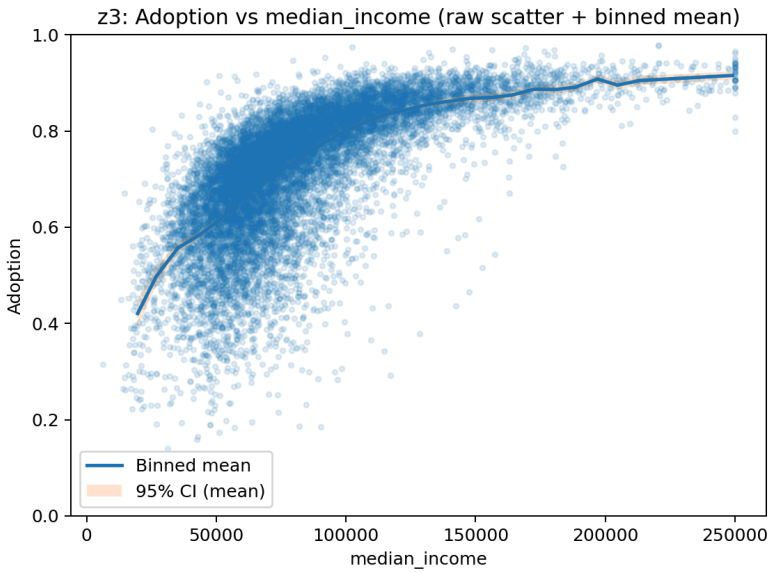

Figure 5: Clear positive relationship between median income and adoption, with diminishing returns above $100K.

This suggests z3 captures a more direct causal pathway: households without computing devices are less likely to subscribe to broadband, regardless of whether service is available.

3.3 Comparing the Two Dimensions

Aspect	z6 (Rural Composite)	z3 (Device Poverty)
Adoption Spread	32.5 pp (85%→52%)	26.7 pp (87%→61%)
ZIPs Below 70% Adoption	~4,400 (deciles 7-9)	~4,400 (deciles 7-9)
Top Feature	total_households (marker)	pct_no_large_screen (direct)
Univariate Signal	Weak (composite)	Strong (direct)
Policy Lever	Infrastructure + wraparound	Device subsidies, digital literacy

4. Discussion

4.1 Two Distinct Pathways

The autoencoder approach reveals something that traditional regression would obscure: the digital divide operates through at least two distinct mechanisms. A linear or tree-based model would identify population as an important predictor of adoption, but would not reveal that its predictive power disappears when decoupled from correlated socioeconomic and device-access factors.

The rural composite disadvantage pathway (z6) captures how small, isolated communities face compounding challenges. The fact that population size is the strongest correlate but has no direct relationship with adoption suggests that *infrastructure availability* (not captured in our features) may be the mediating factor.

The direct device poverty pathway (z3) is more straightforward: households without computing devices don't subscribe to broadband. The strong univariate relationships suggest that device access programs could have immediate impact on adoption.

4.2 Limitations

- **No infrastructure data:** The model cannot distinguish between "won't subscribe" and "can't subscribe."
- **Ecological fallacy:** ZIP code-level associations do not necessarily hold at the household level.
- **Validation set only:** Results are based on 14,685 ZIP codes in the validation set.

5. Conclusion

This analysis demonstrates that demographic features alone can explain over half the variance in broadband adoption across U.S. ZIP codes. The hybrid autoencoder approach identified two interpretable latent dimensions:

Rural composite disadvantage (z6): Adoption drops from 85% to 52% across the z6 spectrum, with approximately 4,400 ZIP codes falling below 70% adoption. Population size serves as a marker for bundled disadvantages but has no direct relationship with adoption.

Direct device poverty (z3): Adoption drops from 87% to 61%, with approximately 4,400 ZIP codes below 70% adoption. Device poverty (pct_no_large_screen) shows a strong direct relationship with adoption, suggesting device subsidies could have immediate impact.

Policy implication: These two pathways require different interventions. High-z6 areas need infrastructure investment paired with wraparound support. High-z3 areas may benefit more immediately from device subsidies and skills training. The methodology provides a **first-pass triage mechanism** for identifying at-risk ZIP codes when infrastructure data is unavailable.

Technical Appendix

Latent Selection Summary

Latent	Spread	Spearman	Reversals	Direction	Selected
z6	0.325	-1.000	0	decreasing	✓
z3	0.267	-1.000	0	decreasing	✓
z4	0.147	-0.988	1	decreasing	✗
z5	0.126	+0.576	1	increasing	✗
z1	0.098	+0.915	2	increasing	✗
z2	0.095	-0.661	4	decreasing	✗