Finding Hidden Communities in WA Census Data: What 20-Somethings Can Teach Us About Housing Trends.

By: Doung Le, Eloho Okoloko, and Surya Kailash, Ramesh May 27, 2025

Introduction

Washington State is changing fast, and so are the people who live here. To gain an understanding of the Washington population, we accessed the 2023 census data provided by the Integrated Public Use Microdata Series population database (IPUMS). The original dataset had over 10 million records and 38 variables, including responses to questions regarding the household, e.g. annual property insurance and taxes, annual utilities costs, number of families, couples and parents in the households, total family income and socioeconomic index,.. We narrowed our research scope to a more manageable subset with the following criteria: adults between the ages of 20 and 30 who are renting homes with at least two bedrooms and earn more than $40,000 annually.

Our goal was to find patterns within this group that aren't immediately obvious from just looking at rows and columns. Who lives with roommates? Who's likely to live alone? Who might be struggling to afford rent? To do that, we turned to unsupervised machine learning. That means we didn't tell the computer what kind of people we were looking for, we let the data speak for itself.

We used the following tools:

- PCA (Principal Component Analysis) to reduce the data to something we could visualize and understand
- K-Means Clustering to group similar individuals together
- Hierarchical Clustering to explore different ways of grouping the same population
- Matrix Completion to help fill in gaps in the dataset

A Quick Tour of the Algorithms

PCA (Principal Component Analysis)
PCA reduces high-dimensional data into fewer dimensions (we used 2) while preserving as much variability as possible. This makes it easier to visualize and cluster the data. It's powered by a technique called SVD (Singular Value Decomposition). PCA helps us simplify complex

data, but it is not easy to interpret as each new axis (or principal component) is a mix of the original variables.

K-Means Clustering

K-Means groups individuals into k clusters based on how similar their features are. We tested multiple values of k (from 2 to 7), and used:

- Elbow method (to assess how much "within-group variance" remains)
- Silhouette score (to measure how distinct each group is)

Hierarchical Clustering

This builds a tree-like structure (a dendrogram) showing how data points can be merged into clusters at different thresholds. Linkage methods like ward, average, complete, and single are visualized to compare structures. This method, although beneficial in the way that it does not require a pre-specified number of clusters, is very computationally expensive and sensitive to noise.

Matrix Completion

Matrix completion allows us to estimate missing values based on patterns in the existing data. To test the effectiveness of this approach, we simulated missing entries in a complete dataset and applied PCA-based matrix completion. The method begins by filling missing values with column means, then iteratively applies Singular Value Decomposition (SVD) to reconstruct the matrix using its top principal components. This continues until the reconstruction error on observed values stabilizes, resulting in a completed matrix suitable for downstream models like PCA or clustering without losing rows.

How We Got Our Hands Dirty With the Data

Preprocessing

The dataset was first standardized to ensure all features contributed equally to the clustering process. This is essential as PCA and hierarchical clustering are sensitive to the scale of input variables.

K-means Clustering

Since the dataset had many variables, we used Principal Component Analysis (PCA) to reduce the dimensions and reveal major trends. We retained 2 principal components to simplify visualization and clustering. Next, we applied K-Means clustering to the PCA-transformed data. To choose the best number of clusters (k), we tested values from 2 to 10 and compared:

- Inertia (within-cluster variation) for the elbow method
- Silhouette Score to measure how well-separated the clusters were

Hierarchical Clustering

Due to the computational cost of hierarchical clustering on large datasets, we randomly sampled 2000 data points from the standardized dataset to generate dendrograms and evaluate different clustering structures. For each linkage method, we generated a dendrogram, truncated to the top 5 levels and cut the dendrogram at k = 3. We evaluated four linkage methods:

- Single: maximum pairwise distance
- Complete: minimum pairwise distance
- Average: mean pairwise distance
- Ward: minimizes within-cluster variance

Matrix Completion

While the original dataset did contain missing values, we preprocessed it by subsetting relevant columns and dropping incomplete rows. This left us with a complete dataset suitable for our application. To evaluate the effectiveness of PCA-based matrix completion, we simulated missing values by randomly masking a subset of entries, allowing us to assess the quality of imputation against known ground truth.

We initialized missing entries with column means and applied iterative low-rank approximation using Singular Value Decomposition (SVD) with the top 5 principal components. At each iteration, we reconstructed the matrix and updated the imputed values accordingly. Convergence was measured using the mean squared sum of error (MSS) on the observed (non-missing) entries.

The algorithm converged after 6 iterations, with the MSE stabilizing at 0.484 and the relative error dropping below the convergence threshold of 1e-7. To evaluate imputation quality, we calculated the correlation between the imputed and true values (before masking), which was 0.61, suggesting a moderately strong recovery of the original data structure.

The Patterns We Discovered

K-means Clustering

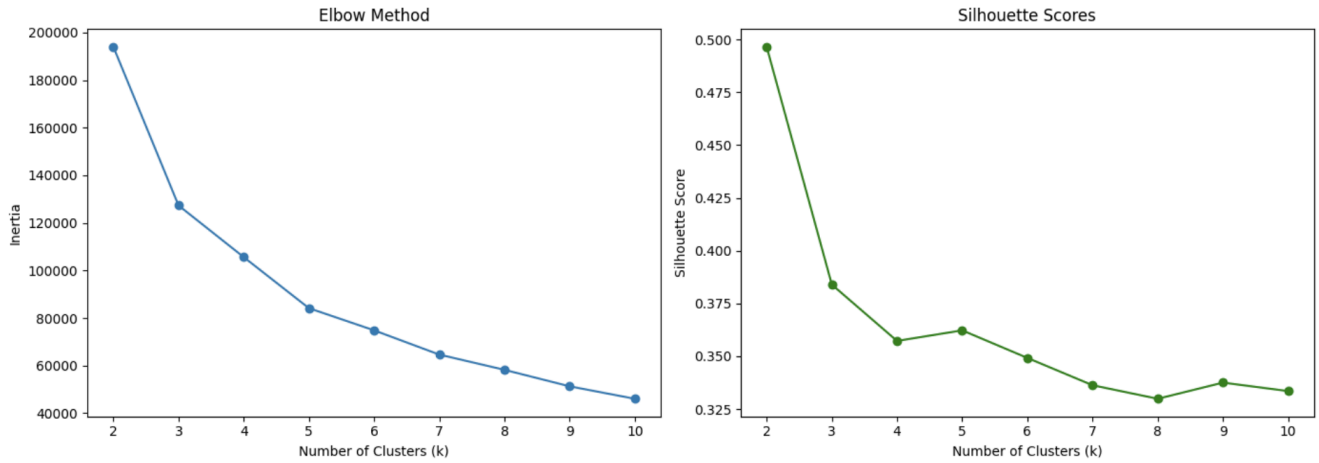

Figure 1. *Finding the Right Number of Groups (Elbow Method) and Measuring Cluster Quality (Silhouette Scores)*

Based on both elbow and silhouette plots, k=3 emerged as a reasonable balance between simplicity and clustering quality. While other values of k slightly increased the silhouette score, the separation and interpretability of three clusters were clearer in visual plots.

Feature	**Correlation**
Family Size	+0.4189
Number of Mothers in Household	+0.3915
Number of Fathers in Household	+0.3893
Number of Children in Household	+0.3310
Number of Housing Units in Structure	-0.2949

Table 1. *Top 5 features contributing to PCA 1*

Our analysis shows that PC1 captures variation in household composition and housing density. Features with the highest positive correlation, such as family size, number of mothers and fathers in the household, and number of children. This indicates that PC1 increases with larger, multigenerational families. In contrast, the number of housing units in a structure has a negative correlation, suggesting that higher PC1 values are associated with households living in single-unit or low-density housing rather than large apartment complexes. In essence, PC1 separates large family households from smaller or individual-based households, particularly those residing in more compact, multi-unit housing. This component highlights an underlying structural divide in living arrangements that may reflect socioeconomic, cultural, or generational housing patterns.

Feature	*Correlation*
Age	*+0.4348*
Year of Birth	*-0.4348*
Total Family Income	*+0.3952*
Total Personal Income	*+0.3343*
Number of Families in Household	*-0.2271*

Table 2. *Top 5 features contributing to PCA 2*

PC2 captures variation related to age, income, and household complexity. Age shows the strongest positive correlation, while year of birth is negatively correlated—confirming that higher PC2 values correspond to older individuals. Both total family and personal income are also

positively correlated with PC2, suggesting that these older individuals tend to be wealthier. In contrast, the number of families in a household is negatively correlated, indicating that those with higher PC2 scores are more likely to live in single-family or simpler living arrangements. In essence, PC2 distinguishes older, wealthier individuals from younger or lower-income individuals, particularly those in shared or multi-family households. This component may reflect differences in life stage, economic stability, and housing independence.

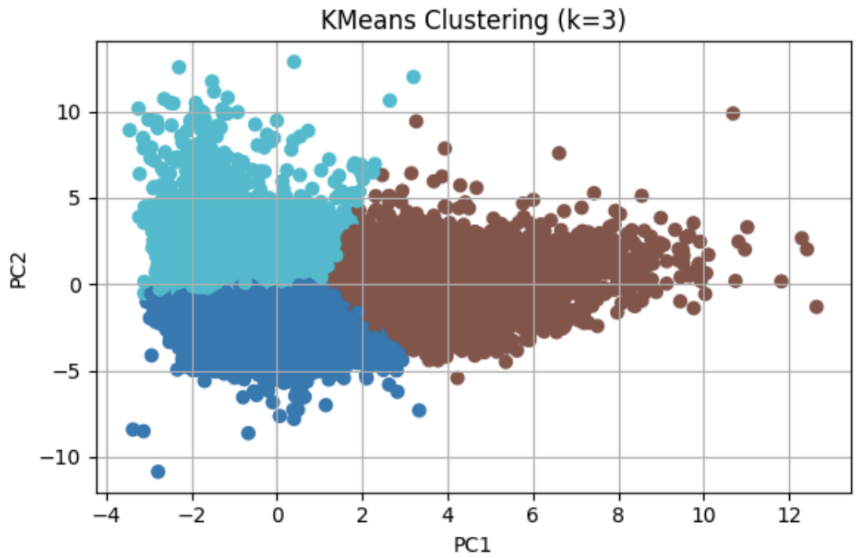

Figure 2. *Clustering with K-Means (k=3)*

PC1 (X-axis): Primarily reflects household structure and size. High values are associated with large families, more children, and single-family homes. Low values suggest smaller or single households, often in apartments.

PC2 (Y-axis): Separates by age and income. High values indicate older individuals with higher income; low values represent younger individuals with fewer resources.

Cluster Color	**Region on PCA Plot**	**Description**	**Label**
Brown	Far right (High PC1)	Large family households, multiple adults & kids	Cluster A
Blue	Bottom left (Low PC1 & PC2)	Young renters, lower income, likely students	Cluster B
Cyan	Top left (Low PC1, High PC2)	Wealthy, older professionals, smaller households	Cluster C

Table 3. *Cluster interpretations (k=3)*

Hierarchical Clustering

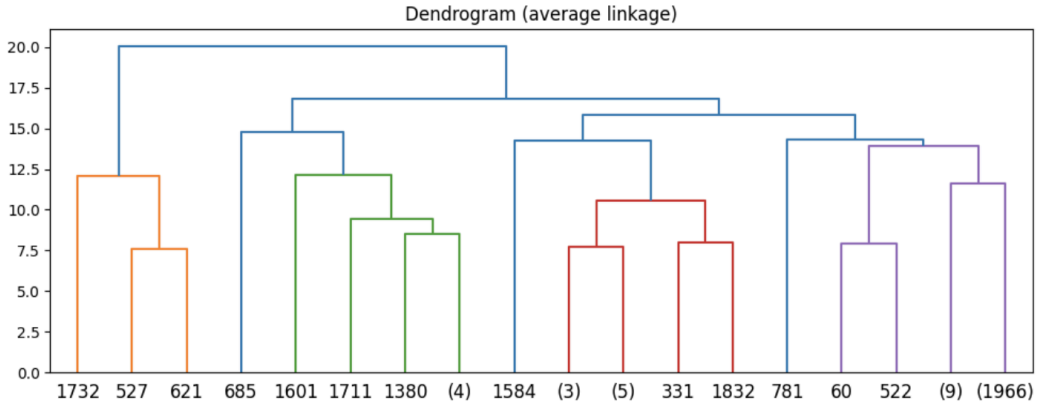

Figure 3. *Hierarchy of Clusters (Average Linkage Dendrogram)*

Linkage Method	**Silhouette Score**
Single	0.5797
Complete	0.5217
Average	0.5863
Ward	0.2052

Table 4. *Silhouette Score Table for each linkage method*

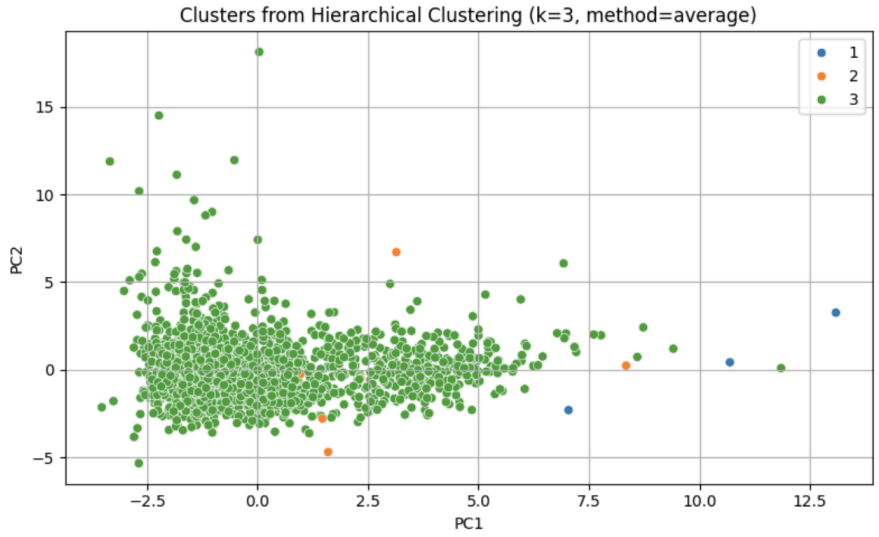

Figure 4. *Groups from Hierarchical Clustering (Average Linkage, k = 3)*

In average linkage, cluster 3 dominates almost everything. This suggests that most of the sampled data is being grouped into a single cluster. This defeats the purpose of clustering. Cluster 1 and 2 are small and scattered, indicating outliers and not genuine clusters. The slicing threshold might be too high, or the dendrogram had no strong splits beyond the main blob.

Although the silhouette score suggested a decent clustering result, the plot showed a lack of meaningful separation. This discrepancy highlights a limitation of hierarchical clustering for this dataset. As a result, we concluded that K-Means clustering may be more appropriate, especially when used on PCA-transformed data with further tuning (via elbow method and silhouette analysis) to determine the optimal number of clusters.

Matrix Completion

Iteration	**Mean Squared Sum of Error(MSS)**	**Relevant Error**
1	0.484041	4.640476e-01
2	0.483991	5.582803e-05
3	0.483984	7.283264e-06
4	0.483983	1.344908e-06
5	0.483983	2.727519e-07
6	0.483983	5.660008e-08

Table 5. *PCA/SVD-Based Matrix Completion Iteration Summary*

The algorithm converged rapidly, reaching stability within 6 iterations. The mean squared sum of error (MSS) across observed entries decreased steadily, starting from 0.4840 and stabilizing at 0.48398. The relative error fell below the convergence threshold (1e-7) by the sixth iteration, indicating that further updates to the imputed values would yield minimal improvement. The figure below illustrates the convergence trajectory of both the MSE and relative error over successive iterations on a log scale.

To assess the accuracy of the imputations, we compared the imputed values against the true values (prior to masking). This resulted in a correlation coefficient of 0.61, suggesting a moderately strong agreement and effective recovery of the underlying data structure.

Why This Matters and What's Next

Our analysis sought to uncover meaningful clusters among renters in Washington aged 20–30 with moderate income and 2+ bedrooms, using a combination of PCA, KMeans, and

hierarchical clustering, followed by matrix completion to simulate instances of missing data. The findings from multiple modeling approaches offer a rich perspective on renter demographics and behavior patterns while also revealing the challenges of unsupervised learning on real-world census data.

Key Takeaways & Interpretation

Principal Component Analysis (PCA) reduced the high-dimensional demographic dataset into two interpretable axes:

- PC1 reflected family size and living arrangement, separating large, possibly multigenerational households from smaller or apartment-based living.
- PC2 captured age and socioeconomic status, contrasting older, wealthier individuals with younger, lower-income residents.

KMeans Clustering (k=3) performed well on the PCA-reduced data (silhouette score ≈ 0.58), identifying three intuitive and distinct renter groups:

- Large family households (suburban, multi-generational)
- Young renters with lower income
- Older, wealthier individuals in smaller households.

Hierarchical clustering with average linkage offered the highest silhouette score (~0.5863) and valuable insight into the nested relationships among clusters, though its dendrogram was harder to interpret and computationally heavier. Ward linkage, while often useful, underperformed here (silhouette ≈ 0.2052), indicating poorly separated clusters in this specific dataset.

These findings can inform housing policy and urban planning, particularly around zoning, affordable housing programs, or targeted support for distinct renter demographics. For example, it could help in identifying clusters with high family size but low income could prioritize subsidy efforts, or recognizing areas with older, wealthier renters may highlight demand for senior-friendly urban development. While our models have limitations, like imperfect cluster separability and assumptions, they offer actionable insight and demonstrate how unsupervised learning techniques can extract meaningful structure from messy, real-world social data.

To evaluate the feasibility of imputing missing data in large demographic datasets, we applied PCA-based matrix completion. Although the original dataset contained some missing values, we filtered to a complete subset and then simulated missingness to test the model's ability to recover true values.

The algorithm successfully reconstructed the missing entries through a low-rank approximation, using only the first five principal components. It converged quickly within six iterations, achieving a stable mean squared error of 0.484 and a correlation of 0.61 between the imputed and original values. This moderate-to-strong agreement suggests that principal components captured much of the essential structure of the data, even when some values were masked.

These results highlight the potential of matrix completion techniques to salvage incomplete data without discarding entire rows or columns, a common problem in social science and demographic datasets where missingness is frequent. Effective imputation can support more inclusive analyses and reduce the bias introduced by listwise deletion.

Future Model Improvements

To enhance our results, future work could use a data-driven method for selecting the number of PCA components based on explained variance. Clustering could be improved by exploring scalable alternatives like DBSCAN or GMM, and by using methods that handle mixed data types more effectively, such as k-prototypes or Gower distance. For matrix completion, more advanced techniques like regularized matrix factorization or autoencoders may provide better accuracy, especially with more missing data.

Reference

Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Grace Cooper, Julia A. Rivera Drew, Stephanie Richards, Renae Rodgers, Jonathan Schroeder, and Kari C.W. Williams. IPUMS USA: Version 16.0 [dataset]. Minneapolis, MN: IPUMS, 2025. https://doi.org/10.18128/D010.V16.0