BA820 – Project M3

Project Title: Childcare Cost Analysis and County Segmentation

Section and Team Number: B1 Team 4

Team Members: Khoi Thai, Hamza Tariq, Tanmay Yenge, Tanish Puneeth

1. Integrated Problem Framing and Updated Questions

After Milestone 2, our team realized that despite our analysis of the same childcare data, we were looking at similarity and structure from different methodological angles. While some analyses centered on affordability normalization and representation sensitivity, others centered on multivariate profile formation and segmentation via dimensionality reduction.

After collaborative discussion, we narrowed our analytical aim and formulated the following research question for Milestone 3:

How do structurally defined county segments differ in childcare affordability, and within each segment, which counties exhibit anomalous behavior relative to structurally similar peers?

This reformulation represents a shift from similarity analysis to structured segmentation and then anomaly detection conditional on segmentation. Instead of comparing counties across a single dimension, we seek to find structurally similar groups and then assess deviations within those groups. This multi-level approach combines representation design, dimensionality reduction, clustering, and anomaly detection into a single, unified unsupervised modeling process.

2. Recap of Individual M2 Contributions

In Milestone 2, each team member undertook an individual analytical inquiry with the common childcare dataset.

Hamza built an anomaly detection system using PCA and K-means to detect which counties had a different childcare burden compared to similar counties. This showed that differences in affordability cannot be attributed to income or price alone.

Khoi built a PCA-based clustering model to classify counties based on labor market structure and childcare burden, showing that some labor structures are systematically more burdened regardless of similarity.

Tanmay's analysis showed that price alone is not sufficient to capture financial burden and that counties with similar income and childcare prices can be vastly different in terms of burden. Also, that patterns of similarity change with representation, indicating that affordability needs to be a key component of structural modeling.

Tanish used K-means to group counties into four affordability types, finding about 3% face an extreme childcare cost crisis that should be policy priorities.

Each of these methods gave useful but incomplete information. Segmentation showed structural grouping, anomaly detection showed extreme deviations, and burden construction explained how affordability could be measured. However, these methods were not yet integrated into a single framework.

3. Integration Strategy and Synergy Effort

Milestone 3 combines representation design, segmentation, and anomaly detection into a coherent analytical sequence.

What Was Reused

- The childcare burden metric from M2 became the central affordability measure.
- The clustering framework from Khoi's work was retained to define structural segments.
- The distance based anomaly detection method from Hamza's work was reused.

What Was Modified

- In M2, anomaly detection was applied globally. In M3, anomaly detection is performed within each cluster, making deviations context dependent.
- Segmentation is now explicitly tied to affordability interpretation rather than general similarity.
- The dataset was restricted to 2018 so that each county appears once, preventing temporal duplication from influencing cluster structure.

What Was Discarded

- Purely visual similarity comparisons without segmentation were deprioritized.
- Global anomaly detection without structural conditioning was replaced with cluster-conditional detection.
- Extremely broad feature sets were narrowed to focus on affordability, income, labor-force participation, unemployment, and occupational structure.

Assumptions Challenged

1. **Assumption:** High childcare price alone implies high stress.
   How it was Challenged: Representation via burden showed that affordability must account for income.

2. **Assumption:** Anomalies can be identified meaningfully without structural context.
   How it was Challenged: Conditional anomaly detection produced more interpretable results.

3. **Assumption:** More clusters necessarily improve segmentation.
   How it was Challenged: Silhouette analysis demonstrated strongest separation at k=2.

We can combine segmentation with anomaly detection to ensure that we shift away to actually useful classification where a business can learn which counties have an unexpectedly high or low childcare affordability as compared to structurally similar ones, instead of relying on falsely claimed national or regional averages. It allows employers, childcare providers and investors to identify the areas where affordability is a localized issue that can be addressed by one of the

following factors: supply constraints, pricing inefficiency or labor-market mismatch, instead of structural constraints only. This means that organizations will be able to focus subsidies, benefits, pricing decisions and expansion decisions in greater accuracy, avoiding wasteful investment in areas where results are structurally anticipated and maximizing ROI by stepping in where results are not doing as well as fundamental predicts.

4. Integrated Analysis and Results

4.1 Structural Segment Differences in Affordability

Cluster profiling reveals meaningful differences in economic and childcare structure:

Metric	Cluster 0	Cluster 1
Mean Household Income	$43,865	$60,821
Mean Infant Center Cost	$133.69	$192.19
Mean Affordability Burden	0.161	0.166

Counties in Cluster 1 have significantly high costs of childcare and greater household incomes. They are however, also slightly burdened by the affordability (0.166 vs. 0.161). This implies that higher childcare expenditure in affluent counties partially neutralizes income gains, and maintaining affordability strain.

Cluster 0 is a more loosely comparable share of counties (1,202 vs. 1,158 in Cluster 1), with lower income and lower costs of average childcare, although with a slightly lower affordability burden on average.

In terms of decision-making, the results suggest that an income-based targeting will only lack any meaningful affordability risk. Any success of the interventions should be segment-specific: cost-containing measures should produce greater returns in high-income, high-cost counties, whereas income support or wage-based subsidies can be more effective in low-income markets.

4.2 Anomalies Within Structural Segments

Within-cluster anomaly detection revealed distinct dispersion patterns:

Cluster 0

- Mean centroid distance: 2.308

Cluster 1

- Mean centroid distance: 2.228

Both clusters exhibit comparable dispersion, with Cluster 0 showing a slightly higher mean centroid distance (2.308 vs. 2.228), suggesting relatively balanced structural homogeneity across the two segments.

Notable anomalous counties include:

- **Loving County, TX** (Cluster 0), exhibiting the highest deviation from its cluster centroid, reflecting its extremely small and atypical rural character.
- **Bronx County, NY** (Cluster 1), reflecting the high childcare costs and urban density characteristic of large metropolitan areas.
- **Issaquena County, MS** and **Oglala Lakota County, SD** (Cluster 0), reflecting extreme rural or economically atypical structural conditions.

These results demonstrate that substantial variation exists even within structurally similar county segments, implying that uniform policy or investment strategies risk misallocating resources. By evaluating anomalies relative to comparable peer groups rather than global averages, the analysis produces more context-aware insights that better support targeted interventions, efficient budget allocation, and higher-impact policy design.

5. Insights Gained Through Integration

The integrated modeling process produced several key insights:

1. **Representation matters:** By directly adding affordability burden to the featured set enhanced business interpretability of clusters, with the segments being based on actual household pressure and not statistical proximity.
2. **Structural segmentation adds decision value:** Clustering revealed distinct county profiles that appear continuous in raw data, enabling more meaningful peer comparisons and targeted strategies instead of one-size-fits-all solutions.
3. **PCA strengthened model reliability:** Dimensionality reduction reduced multicollinearity and clarified the dominant structural drivers of childcare burden, improving cluster stability and confidence in downstream insights.
4. **Peer-based anomaly detection is more actionable**: The analysis of anomalies in structurally similar groups when compared to global outlier detection would reveal performance disparities and aid in making more accurate priorities and interventions.
5. **End-to-end collaboration was critical:** Coordinating representation choices, dimensionality reduction, segmentation, and interpretation ensured analytical rigor translated into insights that support resource allocation, policy design, and operational decision-making.

Collaboration was essential in balancing representation sensitivity, dimensionality reduction, segmentation, and interpretability.

6. Limitations, Open Questions, and Next Steps

While we are happy with what we have achieved in M3, there are definitely limitations to our analysis and questions that remain unanswered.

Single year: We used 2018 to avoid confusion, but we miss how affordability changes over time. Missing data: About 32% of counties did not have childcare cost data and were left out. Limited features: We focused on economic and labor variables, missing policy factors, demographics, and supply dynamics. Next steps:

1. Extend clustering to multiple years to test stability.
2. Examine unusual counties to understand specific conditions.
3. Include policy and demographic variables.
4. Conduct sensitivity analysis on clustering methods and thresholds.
5. Explore patterns in missing data.

Appendix

Shared GitHub Repository

Repository Link: https://github.com/TanmayYenge/ba820-project

Files Structure:

Contribution Table

Team Member	M2 Contributions	M3 Integration Role
Khoi	Developed a PCA-based clustering framework to group counties by labor-market structure and childcare burden, revealing that certain labor profiles are systematically more burdened even among otherwise similar counties.	Led integration effort; implemented PCA for structural interpretation; created visualization pipeline; primary coder for integrated notebook; synthesized findings into coherent narrative
Tanish	Used K-means to group counties into four affordability types, finding about 3% face an extreme childcare cost crisis that should be policy priorities.	Provided input on feature selection decisions; helped identify redundancies in exploratory approaches; contributed to discussion of which M2 analyses to prioritize for integration
Tanmay	Showed that similarity patterns shift when representation changes, suggesting that affordability should be central to structural modeling.	Tested sensitivity to alternative k values; helped validate that k=2 was a robust choice; contributed to data preprocessing standardization; assisted with result interpretation and documentation and making of the report.
Hamza	Developed an anomaly detection framework using PCA and K-means to identify counties where childcare burden deviated from similar counties. This demonstrated that affordability differences	Developed a cluster-conditional anomaly detection framework, interpreted analytical findings, and prepared supporting methodology to ensure seamless integration.

	cannot be explained by income or cost alone.	

Note: The integration was primarily led by Khoi and Hamza, who combined the M2 segmentation work with new anomaly detection methods. Tanish and Tanmay's M2 exploratory analyses informed feature selection and validated findings, but their work was not directly incorporated into the integrated pipeline to maintain narrative coherence. All team members participated in discussions about the integration strategy and the interpretation of results.

Process Overview

Analytical Pipeline:

1. **Data Preparation** → Filter to 2018, compute burden metric, drop missing values, create one observation per county
2. **Feature Selection** → Select core economic variables (burden, income, costs, unemployment, labor participation)
3. **Standardization** → Scale features for distance-based methods
4. **PCA** → Understand structural dimensions and visualize county space
5. **KMeans Clustering** → Test k=2 to k=10, select k=2 based on silhouette scores
6. **Cluster Profiling** → Compute average characteristics for each segment
7. **Anomaly Detection** → Calculate distance to cluster centroid, flag top 1% within each cluster
8. **Interpretation** → Analyze patterns, compare to M2, synthesize insights

This pipeline integrates structural segmentation with context-aware anomaly detection to provide a comprehensive view of childcare affordability across U.S. counties.

Use of Generative AI Tool

On the notebook, we consult ChatGPT to guide us on generating python code blocks based on the steps we discuss as well as giving feed backs on our interpretation of the result.

https://chatgpt.com/share/6994e7f9-eb24-8013-babe-269ae1a363bd

In the report,

After writing and combining findings into the report, I used ChatGPT to act as a tutor and grade our team's work according to the instructions given. We were given an 88/100 and were suggested a few things to change but that was sort of changing the narrative so we did not really follow everything that ChatGPT said but just changed some headings and fixed grammar errors.

https://chatgpt.com/share/6994fe4f-8b38-8002-8918-d9ee1e74dea9