Coverage Gaps: Uncovering the Underserved in Life Insurance

Paul Hayden LaPiana

July 2025

Abstract

This study investigates widespread underinsurance in the U.S. life insurance market, revealing counterintuitive patterns that challenge conventional industry assumptions. By integrating de-identified MIB application data with the FINRA National Financial Capability Study (NFCS) and the Federal Reserve Survey of Consumer Finances (SCF), the analysis constructs an enriched dataset capturing demographic, behavioral, and financial variables. A tiered data harmonization and merging strategy enabled precise alignment across sources. Further exploratory analysis revealed strong geographic disparities in financial behavior and insurance adequacy, with the Northeast and Northwest consistently outperforming the Southeast across multiple indicators. Additionally, financial education emerged as a key differentiator, significantly associated with higher-value policy ownership and stronger overall financial wellness.

Dimensionality reduction using UMAP and clustering with K-Means revealed three distinct market groups: Affluent Self-Reliant Optimizers (41.6% of market, $96k average income), Financially Stressed Protectors (37.9%, $59k income), and a Low-Engagement Middle Ground (20.5%, $73k income) characterized by systematic non-response to financial literacy questions and avoidance of financial planning conversations.

Surprisingly, the highest-earning segment demonstrated the lowest insurance adequacy score (0.240), while the financially constrained group achieved the best adequacy ratio (0.393), contradicting expectations that income drives coverage levels. Although all segments fell below the 1.0 adequacy threshold, a substantial 65% performance gap exists between the most and least covered groups. The identification of a significant low-engagement population validates the critical role of financial education in driving insurance coverage decisions and reveals a substantial market opportunity for educational intervention strategies.

XGBoost feature importance analysis identified behavioral and psychological factors—including financial confidence, risk attitudes, and life circumstances—as more predictive of coverage decisions than income alone. Statistical analysis confirmed significant intergroup differences across financial, behavioral, and geographic dimensions. These findings demonstrate that traditional income-based targeting systematically misses both the psychological drivers of insurance adequacy and the engagement barriers that prevent 20.5% of the market from participating in financial planning conversations, creating opportunities for behavioral-based segmentation and targeted financial education strategies that could transform industry acquisition and retention approaches.

1 Introduction

Life insurance is a critical component of financial security, yet underinsurance remains widespread in the United States. Traditional industry approaches to understanding coverage patterns rely heavily on broad demographic categories that may oversimplify the complex psychological and behavioral factors actually driving insurance decisions. While these demographic bins provide convenient targeting frameworks, they often fail to capture the nuanced interplay of financial literacy, risk attitudes, family circumstances, geographical location, and decision-making behaviors that determine whether individuals seek adequate protection.

Problem Definition

This study employs advanced data integration and segmentation techniques to identify naturally occurring population groups with distinct insurance adequacy patterns and behavioral characteristics. By combin-

ing proprietary MIB application data with comprehensive financial and behavioral datasets, I construct an empirical framework for measuring insurance adequacy and apply dimensionality reduction and clustering analysis to reveal underlying market segments. The objective is to provide insurance decision-makers with actionable insights for developing targeted acquisition strategies, product designs, and engagement approaches based on actual customer behaviors rather than demographic assumptions.

2 Scientific Question

Can integrated feature engineering, dimensionality reduction, and unsupervised clustering reveal distinct behavioral market segments that traditional demographic targeting systematically misses, and do ensemble methods identify psychological and geographic factors as stronger predictors of insurance adequacy than income-based approaches?
This study takes a bottom-up approach to identifying underserved market segments, departing from the conventional top-down methodology that begins by defining static demographic categories. By allowing the data to drive segmentation, this method enables the discovery of nuanced relationships and shared traits across groups traditionally treated as distinct.

2.1 Hypotheses:

1. Financial literacy, financial stress, and insurance coverage will vary significantly by geographic region, indicating systemic inequalities.
2. Exposure to financial education will positively correlate with higher financial literacy scores and insurance adequacy.
3. Dimensionality reduction followed by clustering will uncover statistically distinct market segments with unique demographic, financial, and behavioral characteristics.
4. Respondents with higher financial stress scores will be significantly more likely to be underinsured, regardless of income level.
5. Key features identified by ensemble methods will explain more variance in coverage adequacy than income alone.

2.2 Process Overview:

1. Conduct exploratory data analysis on MIB data and third-party datasets to identify key features for enrichment.
2. Clean and preprocess supplemental datasets to ensure compatibility with MIB data.
3. Merge datasets using a tiered combination of demographic and financial variables.
4. Perform dimensionality reduction using UMAP with hyperparameter tuning and cross-validation.
5. Apply K-Means clustering and determine the optimal number of clusters.
6. Use ANOVA to confirm that resulting clusters are statistically distinct.
7. Apply Random Forest to identify key features driving insurance coverage and cluster differentiation.

3 Data

To conduct this research, I will utilize data from **MIB, FINRA, and the Federal Reserve**. The MIB dataset captures de-identified insurance application data that captures age bands, gender, face amount bands, postal code, and application count. I will enrich this information by integrating key features from both publicly available datasets. The FINRA dataset is the **National Financial Capability Study (NFCS)**, a nationwide survey encompassing responses from over 25,000 U.S. adults, designed to assess their overall financial capability. This dataset is highly comprehensive, **featuring 27,118 observations and 168 categorical features**. The Federal Reserve data includes the **Federal Reserve Survey of Consumer Finances (SCF)**, a highly detailed survey capturing extensive financial information from respondents, including income, net worth, assets, liabilities, credit usage, and other significant financial factors. The dataset utilized for this analysis is from 2022 and **consists of 22,975 observations across 5,473 coded features**. Both of the 3rd party datasets include coded columns and multiple choice answers.

4 Exploratory Data Analysis and Preprocessing

This analysis integrates three datasets:

1. **MIB Proprietary Data**
2. **FINRA National Financial Capability Study**
3. **Federal Reserve Survey of Consumer Finances**

While the MIB proprietary dataset serves as the primary source, the other two datasets were included to supplement key variables related to income, financial literacy, family structure, insurance coverage, and financial stress. Merging these datasets enables the creation of a robust, enriched dataset that captures statistically likely demographic, financial, and behavioral profiles.

MIB Proprietary Data

MIB collects de-identified data to support risk management across the life insurance industry, helping member companies detect fraud, verify application details, and improve underwriting. The dataset used in this analysis includes **5,448,458 observations across 8 de-identified features**, all sourced from 2022. A full summary of the features is shown in Table 1.

Feature	Description
`holding_company`	Insurance carrier (undisclosed)
`year_name`	Year that the observation was sourced (2022)
`month_name`	Month that the observation was sourced
`gender`	Gender of the observation
`residence_postal_code`	Zip code; where the observation was sourced
`age_band`	Industry standard age band that encompasses the observation
`face_amount_band`	Life insurance face amount band that encompasses the observation
`inquiry_count`	Number of life insurance applications submitted by the observation

Table 1: Summary of MIB Dataset Features

To better understand the composition of insurance applicants, we begin with a high-level exploratory data analysis (EDA) of the MIB dataset. While confidentiality requirements limit disclosure of granular details, we present two key distributions—age band and face amount band—to illustrate relevant population patterns.

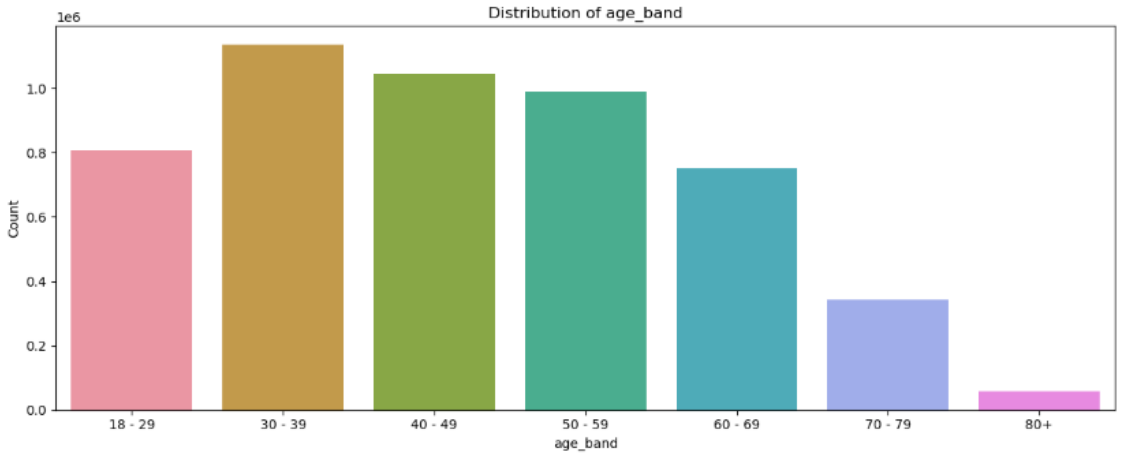


Figure 1: Distribution of age band in MIB data

Figure 1 shows that the majority of insurance applicants fall within the 30–49 age range. This aligns with typical life milestones, such as starting a family or purchasing a home—when individuals are more likely to seek life insurance coverage. The tapering distribution in older age bands reflects expected attrition due to mortality.

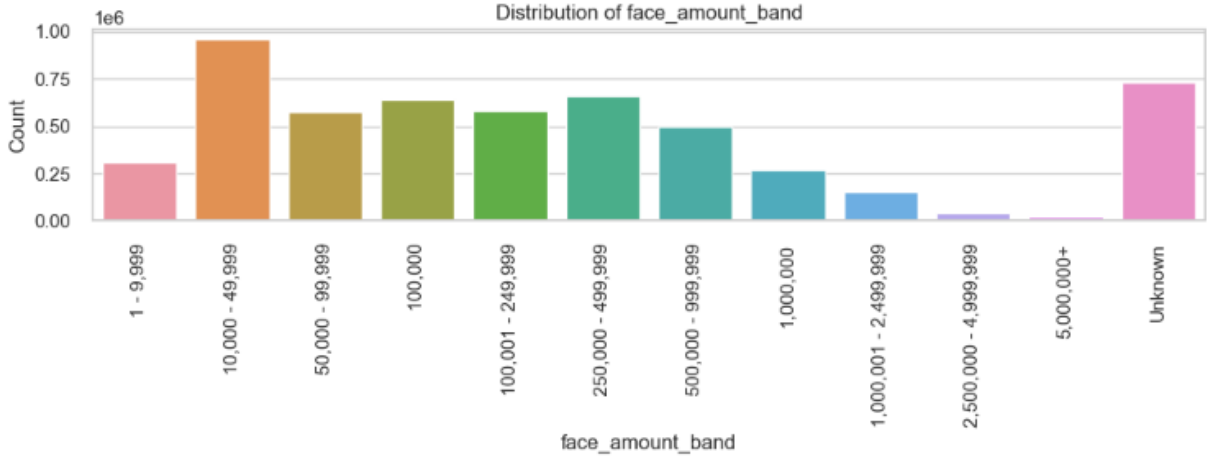


Figure 2: Distribution of face amount band in MIB data

Figure 2 displays the distribution of life insurance face amount bands. The distribution is roughly normal but skewed slightly right, with most individuals opting for policies between $10,000 and $49,000. This trend suggests a potential gap in financial literacy: many individuals may be underestimating how much coverage is needed. Industry guidelines often recommend coverage equal to 10 years of income, yet most applicants appear to be selecting significantly lower face amounts—potentially due to cost concerns, lack of education, or minimal employer-provided policies.

National Financial Capability Study (NFCS)

The **National Financial Capability Study (NFCS)**, conducted by FINRA, is designed to assess financial literacy and financial stress among more than 25,000 Americans. Due to the survey's multiple-choice format, numeric responses often correspond to categorical options (e.g., "never," "once," "more than once"). The dataset also includes rich demographic information. A sample of relevant variables is shown below:

Feature	Description
A50B	Respondent age/gender bins
A6	Marital status
A7	Living arrangements
STATEQ	State
A5_2015	Education level
A8_2021	Approximate annual income
M20	Was financial education offered by a school or college you attended

Table 2: Summary of Key NFCS Features

In addition to demographic variables, the NFCS captures a wide range of behavioral and attitudinal financial data, including approximate household income, self-assessed financial knowledge, and general financial habits. The survey also features a short financial literacy quiz and several questions aimed at assessing financial stress. Using the quiz responses, I created a proxy variable for actual financial knowledge, which I then analyzed in relation to key demographic factors to uncover meaningful patterns.

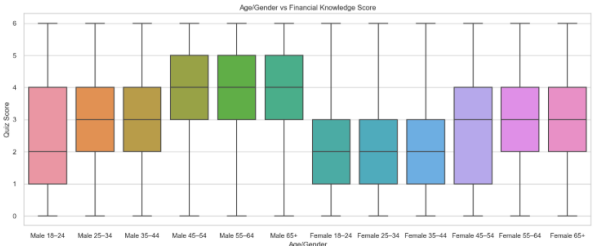


Figure 3: *
(a) Financial Literacy by Age Band and Gender

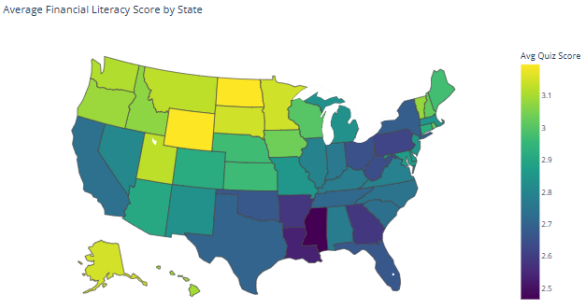


Figure 4: *
(b) Financial Literacy by State

Figure 5: Financial Literacy by Demographic Differences

These visualizations reveal several notable trends. Figure 3 shows that financial literacy generally increases with age, which aligns with greater financial exposure over time. More strikingly, it also highlights consistently lower quiz scores for women across all age groups—a pattern that may reflect systemic, cultural, or societal disparities. This observation directly supports the project's focus on underserved populations, suggesting that gender may be a meaningful differentiator in financial education needs.

Figure 4 highlights geographic variation in financial literacy, with lower scores concentrated in the Southeast and higher scores in the Northwest. While this underscores the role of regional context, aggregated census region scores appear nearly identical on average, indicating that local or state-level factors may drive these differences.

The NFCS also asks respondents whether they have received formal financial education, offering a valuable lens into how such exposure correlates with financial knowledge.

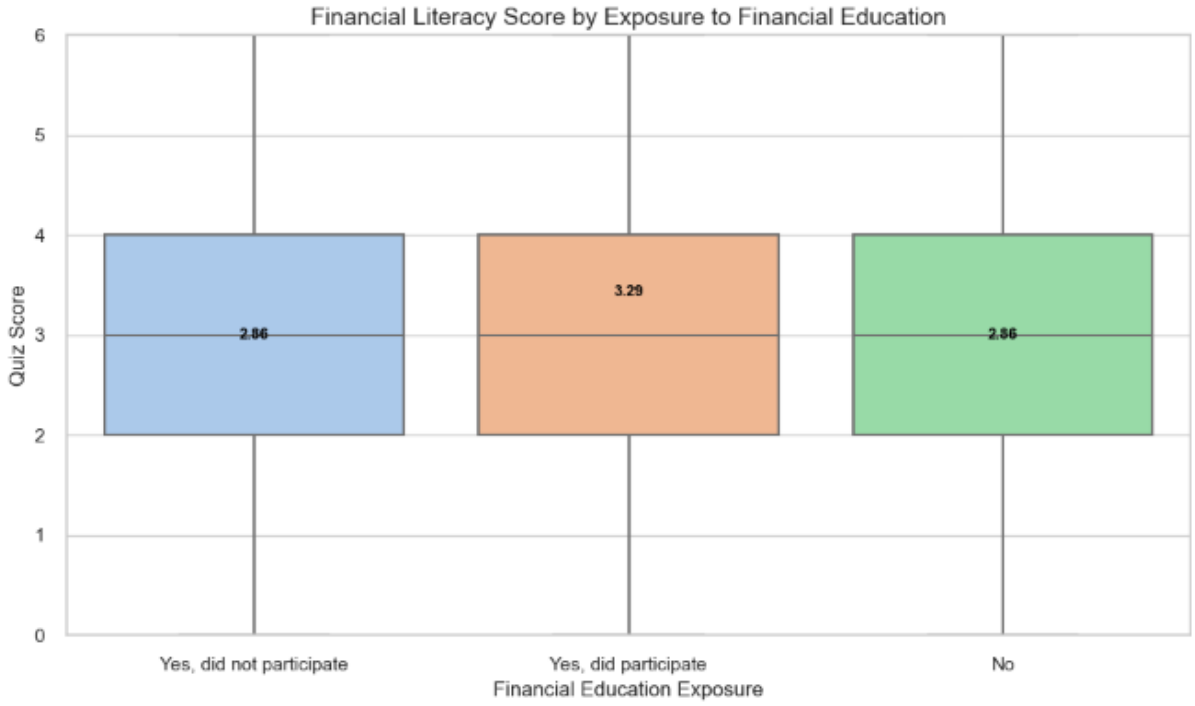


Figure 6: Financial Literacy by Exposure to Financial Education

As shown in Figure 6, respondents who received financial education reported significantly higher quiz scores (**3.29**) compared to those who had not (**2.86**). This finding supports the hypothesis that even limited formal education can meaningfully improve financial literacy outcomes.

In addition to literacy, the NFCS provides insight into various forms of financial stress. I grouped these questions into six thematic categories—**Medical Stress**, **Debt Stress**, **Housing Stress**, **Student Loan Stress**, **Parental Relief**, and **General Financial Stress**—and computed scores within each based on respondent indicators. To consolidate these scores into a single overall financial stress measure, I evaluated four aggregation strategies:

1. **Equal Weighting:** All features contribute equally to the final score.
2. **Frequency-Based Weighting:** Features with fewer missing values receive higher weights.
3. **Variance-Based Weighting:** Features with higher variance are weighted more heavily.
4. **Correlation-Adjusted Weighting:** Features that are less redundant are weighted more heavily.

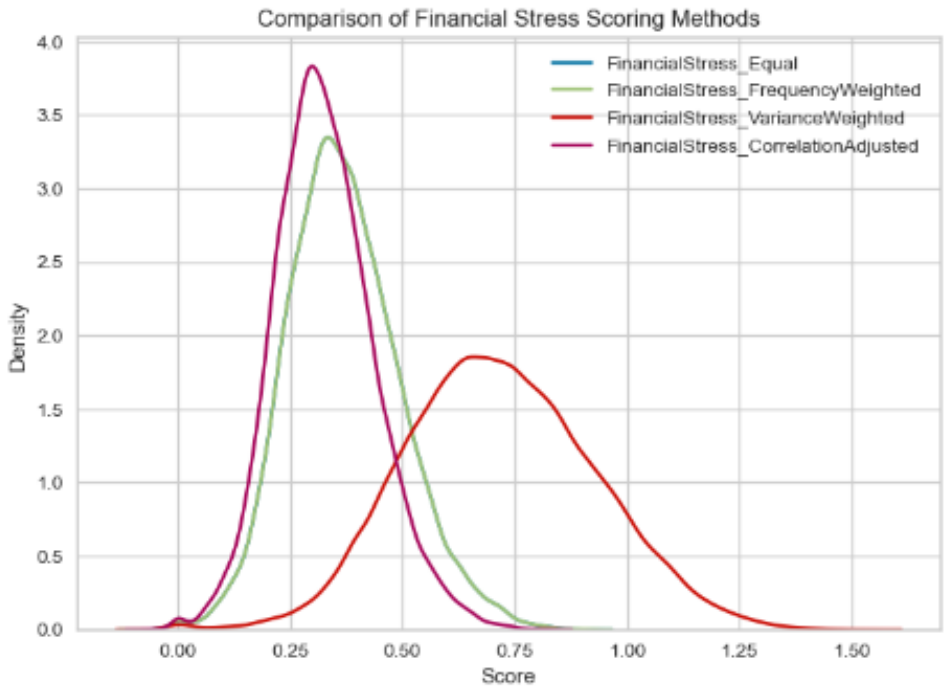


Figure 7: Methods of Financial Stress Score Calculation

As illustrated in Figure 7, the equal and frequency-based methods produced similar right-skewed distributions. The correlation-adjusted method resulted in an even stronger skew, while the variance-weighted approach produced a more normal distribution. This not only aids dimensionality reduction and clustering but also captures a broader range of financial stress experiences. Given these advantages, I selected the **variance-weighted score** as the primary proxy for financial stress in subsequent analysis.

Survey of Consumer Finances (SCF)

The **Federal Reserve Survey of Consumer Finances (SCF)** offers a comprehensive snapshot of U.S. household finances, capturing detailed information on income, assets, liabilities, net worth, credit behavior, and other core financial indicators. This analysis uses the 2022 wave of the SCF, which includes **22,975 observations across 5,473 coded variables**. Like the NFCS, many responses are categorical and encoded numerically. Fortunately, the SCF is accompanied by a detailed codebook, which allowed me to efficiently identify variables relevant to this study. A sample of key variables is shown below in Table 3.

Feature	Description
`x14`	Respondent age
`x8203`	Marital status
`x5931`	Education level
`x5729`	Total income before taxes
`x6670`	Employment status
`x4003`	Current face value of term life policies

Table 3: Summary of Key SCF Features

Unlike the previous datasets, the SCF contains direct, high-quality information related to life insurance coverage. It captures whether respondents or their family members hold life or term insurance policies and includes the total face value of these policies. This data fills a critical gap in the earlier datasets and serves as a cornerstone for the insurance adequacy analysis in this project.

As previously mentioned, a widely accepted benchmark for adequate life insurance coverage is a policy that replaces at least ten years of income and offsets major debts. Using this standard, I engineered a binary variable to identify underinsured individuals.

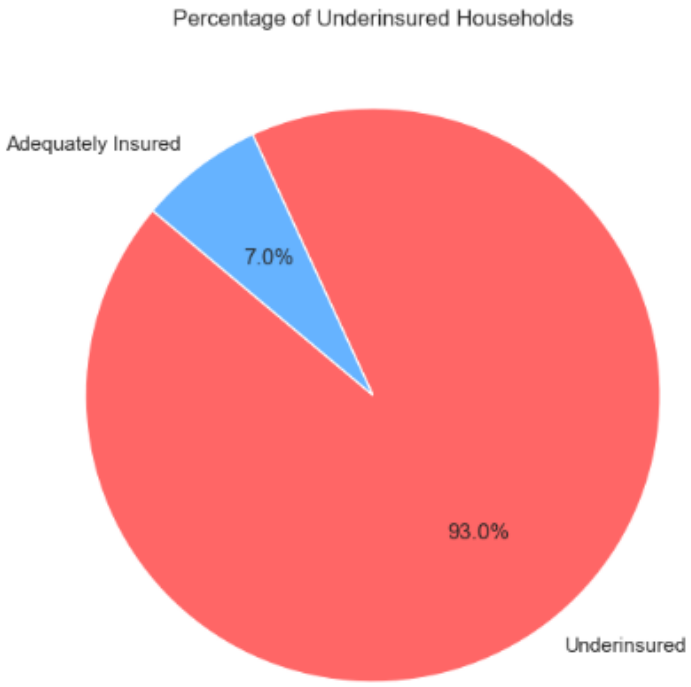


Figure 8: Underinsurance Rate Among U.S. Adults

As shown in Figure 8, **93%** of respondents were classified as underinsured. This striking figure points to widespread coverage inadequacy, potentially driven by gaps in financial literacy, product knowledge, or barriers in the insurance acquisition process. Accurately identifying and tracking these underinsured individuals is essential for informing targeted interventions and product design.

To evaluate whether the industry-standard 10-year income replacement benchmark is overly conservative – or if U.S. adults are genuinely that dramatically underinsured – I expanded the analysis to include 5-year and 8-year present value benchmarks.

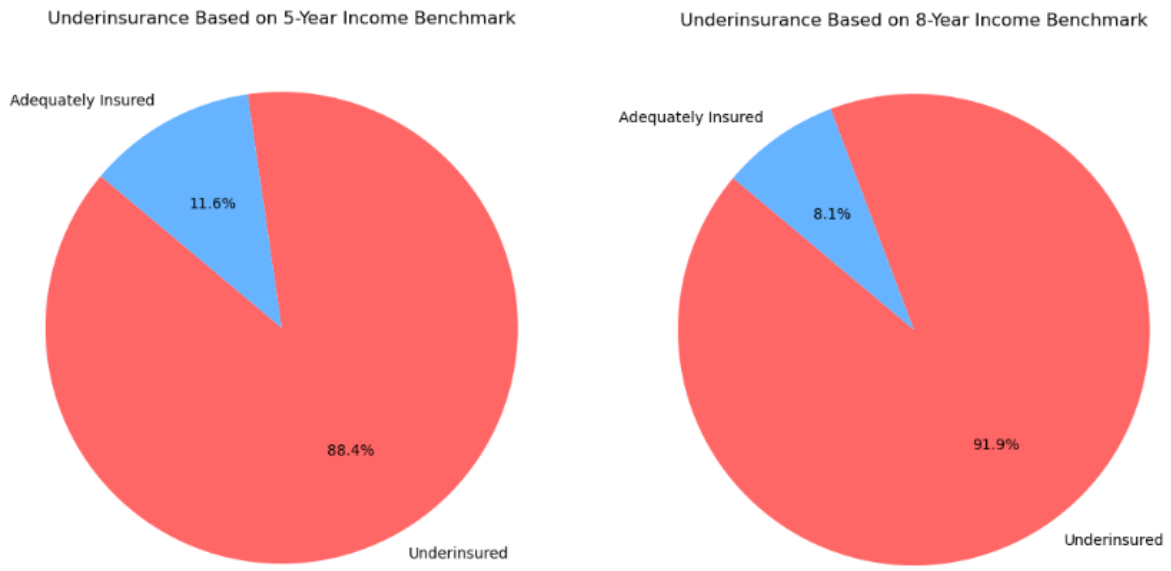


Figure 9: Underinsurance Rates Among U.S. Adults Using 5-Year and 8-Year Benchmarks

The findings are staggering. Not only are most Americans falling short of long-term insurance adequacy—they're also unprepared for short- and medium-term scenarios. When applying a 5-year benchmark, just **11.6%** of U.S. adults meet the adequacy threshold. That number drops to a mere **8.1%** under the 8-year benchmark.
These results underscore a systemic and widespread underinsurance crisis, one that extends far beyond conservative industry assumptions and reflects a broader structural gap in financial protection.

During exploratory analysis, I also identified a major class imbalance in the SCF income distribution. Over **21%** of respondents fell into the "$300k or more" income category—a figure that starkly contrasts with national statistics, which place that share closer to 1%. To correct this distortion, I implemented a rebalancing function designed to align SCF income distributions with those of the NFCS, which more closely reflect the U.S. population. The rebalancing procedure involved the following steps:

1. Verifies all target income categories are present.
2. Randomly undersamples overrepresented income brackets.
3. Synthetically oversamples underrepresented groups by adding noise to numeric fields while preserving structure.
4. Preserves original datatypes.

The procedure was successful, resulting in a more realistic and usable income distribution for downstream modeling. Table 4 shows the distribution before and after rebalancing:

Income Label	Original Distribution	Rebalanced Distribution
Less than $15k	6.34%	12.27%
15k – 25k	7.97%	10.85%
25k – 35k	7.98%	10.76%
35k – 50k	10.94%	14.19%
50k – 75k	12.76%	18.47%
75k – 100k	8.83%	13.17%
100k – 150k	11.64%	12.79%
150k – 200k	5.68%	4.47%
200k – 300k	6.52%	2.06%
$300k or more	21.36%	0.98%

Table 4: Income Distribution Before and After Rebalancing

Figure 12 visually demonstrates the impact of the rebalancing process.

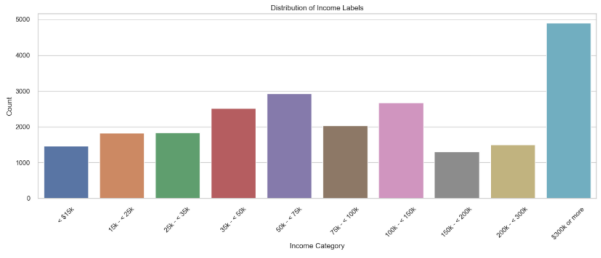


Figure 10: *
(a) Distribution of Income Label before Rebalancing

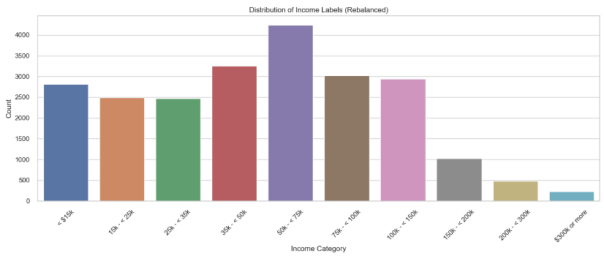


Figure 11: *
(b) Distribution of Income Label after Rebalancing

Figure 12: Change in Income Label Distribution

This rebalancing process played a pivotal role in strengthening the overall validity of my analysis. By aligning the dataset with real-world distributions, I ensured that the insights generated are grounded in population-level reality and better suited for actionable application.

5 Methodology

The objective of this analysis is to uncover latent groups of underinsured individuals and examine the shared traits that define them. The first step was to harmonize key features across the three datasets to ensure accurate and meaningful merges.

5.1 Data Harmonization:

Harmonized Feature	NFCS Transformation	SCF Transformation
age_band_harmonized	Probabilistic mapping to MIB bands	Convert raw age to band
gender_harmonized	Map 1/2 to Male/Female	Map 1/2 to Male/Female
income_band_harmonized	Map category to midpoint, then bin	Bin continuous income into ranges
education_harmonized	Collapse 7 categories to 4	Collapse 15 categories to 4
employment_harmonized	Collapse 8 categories to 5	Collapse 12 categories to 5
risk_tolerance_harmonized	Clean 1-10, filter missing	Normalize 0-10 to 1-10
financial_knowledge_harmonized	Rescale 1-7 to 1-10	Normalize 0-10 to 1-10
has_credit_harmonized	Map 1/2 to True/False	Convert limit to binary
home_ownership_harmonized	Map 1/2 to True/False	Map 1/5 to True/False
face_amount_band_harmonized	None	Bin numeric face values

Table 5: Data Harmonization

5.2 Two-Phased Tiered Merge:

These harmonization steps ensured consistency across datasets and allowed for clean, structured merges based on shared features.

Now that the data is aligned correctly, I can begin my merge. To try and create the most accurate enriched data possible, I used a two phase tiered merging strategy. This allowed me to match observations based on different combinations of demographic and financial data. The matching tiers are summarized in the table below.

Matching Tier	Phase 1	Phase 2
Tier 1	Age + Gender + Income + Education + Employment + Home Ownership	Age + Face Amount Band + Gender + State
Tier 2	Age + Gender + Income + Education + Employment	Age + Face Amount Band + State
Tier 3	Age + Gender + Income + Education	Age + Face Amount Band + Gender
Tier 4	Age + Gender + Income	Age + Face Amount Band
Tier 5	Age + Gender	Age + State + Gender
Tier 6	Age	Age + State
Tier 7	No Constraints	Age
Tier 8	N/A	No Constraints

Table 6: Two-Phase Tiered Merge

Phase 1: Merging NFCS and SCF: The first phase of the tiered merging strategy focused on integrating the NFCS and SCF datasets to build a comprehensive supplemental dataset. This dataset combines behavioral, demographic, and financial variables to serve as a robust enrichment layer for MIB data. Matching was performed using the tiered system of demographic and financial features, shown above in Table 6. The merge between NFCS and SCF was highly successful. Table 7 summarizes the number of matches achieved at each tier.

Matching Tiers	Number of Matches	Percentage of Matches
Tier 1	18,434	80.3%
Tier 2	125	0.5%
Tier 3	2,515	10.9%
Tier 4	1,893	8.2%
Tier 5	0	0.0%
Tier 6	0	0.0%
Tier 7	3	0.0%

Table 7: Two-Phase Tiered Merge

Over 90% of observations were matched within the first three tiers, with near-total matching by Tier 4. This resulted in a robust enriched dataset for the next phase of analysis.

Phase 2: Merging Enriched Data with MIB: The second phase involved merging the enriched supplemental dataset onto a stratified sample of 100,000 observations from the primary MIB dataset. Matching tiers in this phase focused on demographic features, geographic information, and insurance-specific attributes like face amount bands.

Matching Tiers	Number of Matches	Percentage of Matches
Tier 1	17,332	17.3%
Tier 2	15,768	15.8%
Tier 3	58,521	58.5%
Tier 4	2,449	2.5%
Tier 5	4	0.0%
Tier 6	3	0.0%
Tier 7	0	0.0%
Tier 8	5,876	5.9%

Table 8: Two-Phase Tiered Merge

This phase was also highly successful, with over 90% of MIB observations matched within the first three tiers. The result is a deeply enriched, well-aligned dataset positioned for robust segmentation and clustering. To validate the success of the merge process, I examined the distribution of several key demographic variables.

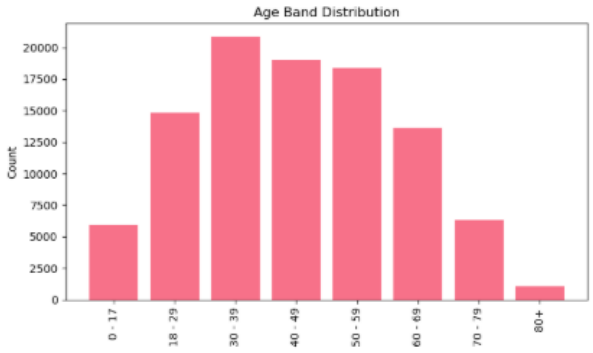


Figure 13: *
(a) Post-Merge Distribution of Age

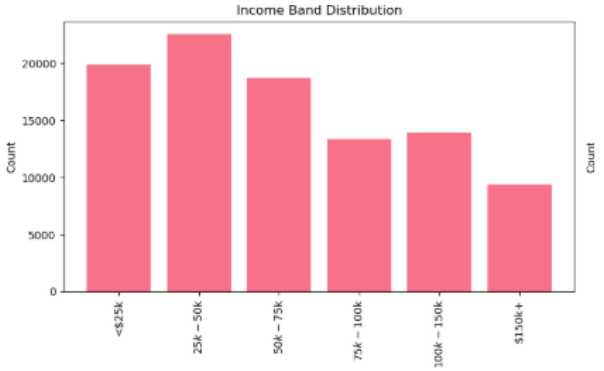


Figure 14: *
(b) Post-Merge Distribution of Income

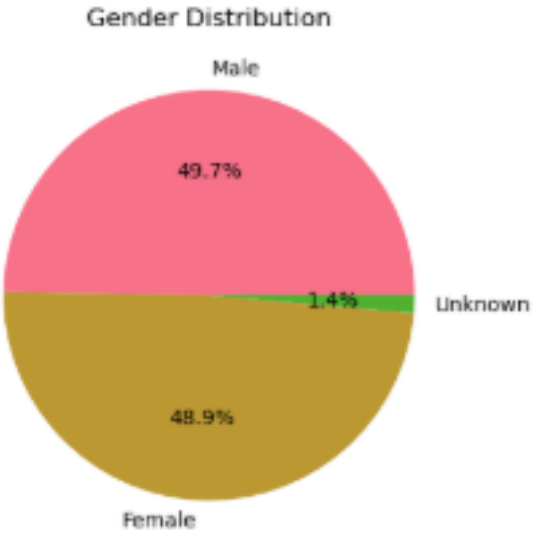
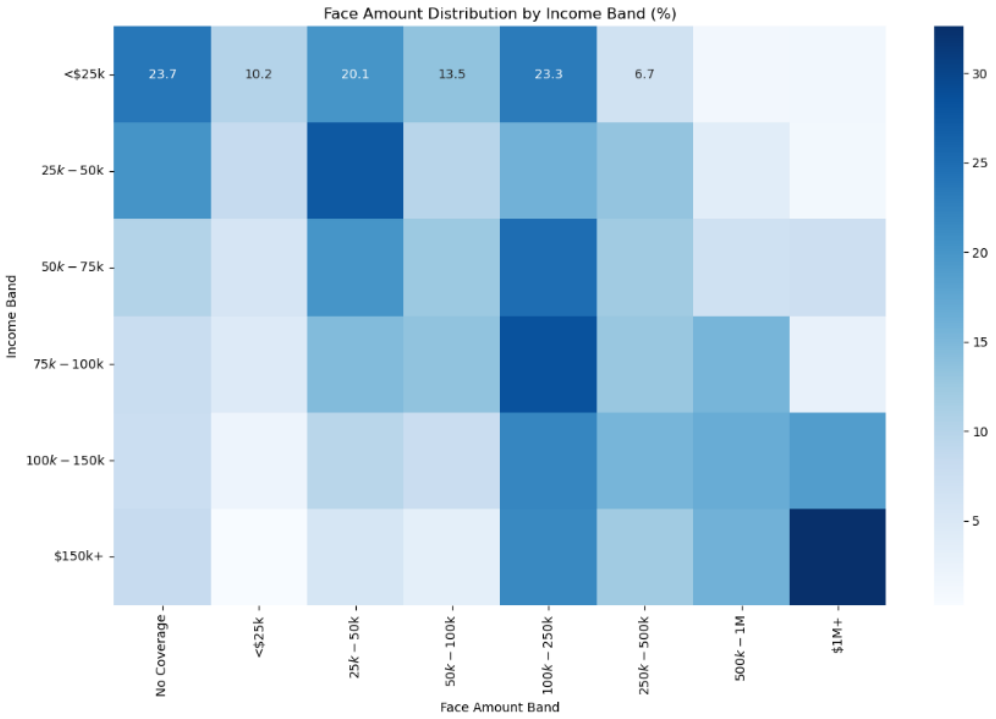


Figure 15: *
(c) Post-Merge Distribution of Gender

As shown above, the age, income, and gender distributions align closely with expected population trends, confirming that the merged dataset retains strong demographic integrity.

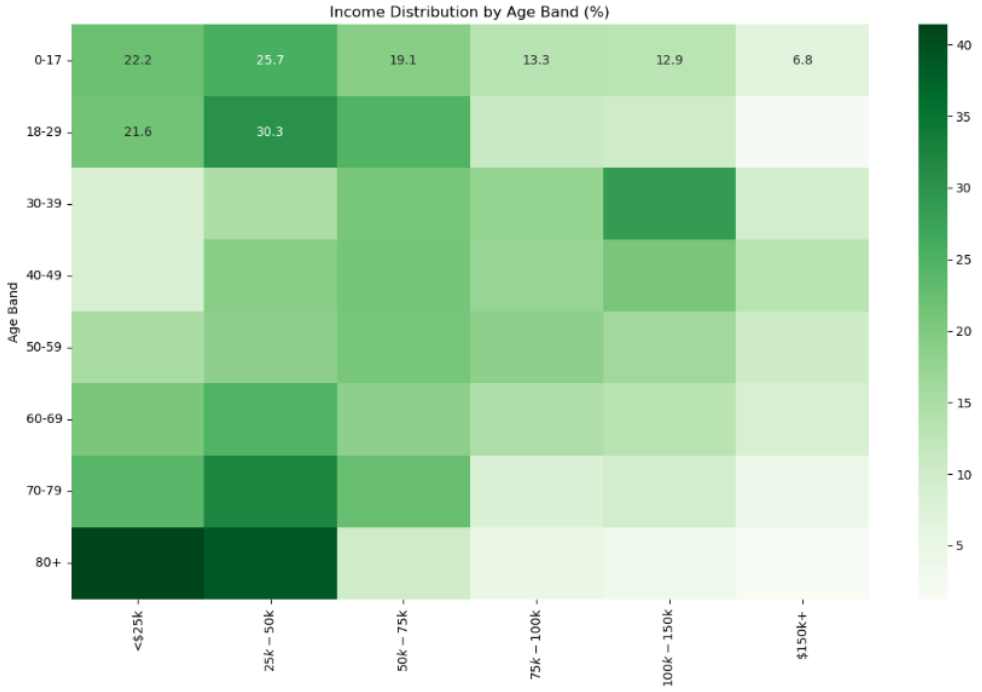
Merged Dataset EDA:

With the data validated, I began exploratory analysis of the merged dataset. I first examined how life insurance face amounts vary by income band.



This visualization reveals a clear upward trend: as income increases, so does the likelihood of holding higher-value life insurance. The positive correlation between income band and face amount band highlights how financial capacity influences insurance coverage levels.

Next, I explored the relationship between income and age.



This plot exhibits a "checkmark" pattern. Income generally rises with age until peaking in the 40–49 age band, after which it declines. This reflects common career and retirement trajectories, with income growth slowing or reversing as individuals exit peak earning years. Further, it gives us three clear marketing lanes: early engagement in the 20s, growth in the 30s, and the prime insurance window in the 40s. Each requires different messaging, but the timing is predictable and actionable.

Building on earlier findings that financial literacy varies by state, I broadened the analysis to examine key financial characteristics across geographic regions.

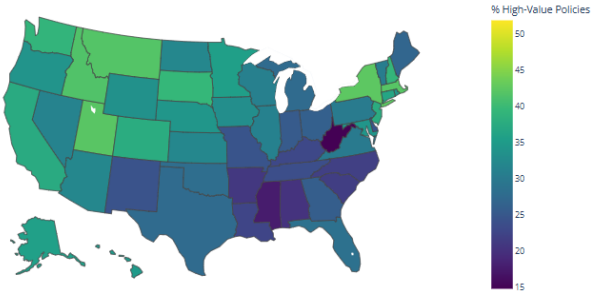


Figure 16: *
(a) High-Value Life Insurance Policies by State ($500k+)

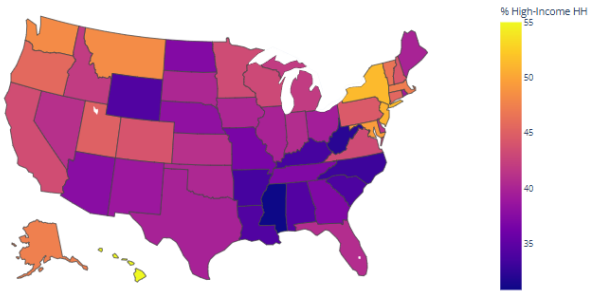


Figure 17: *
(b) High-Income Households by State ($100k+)

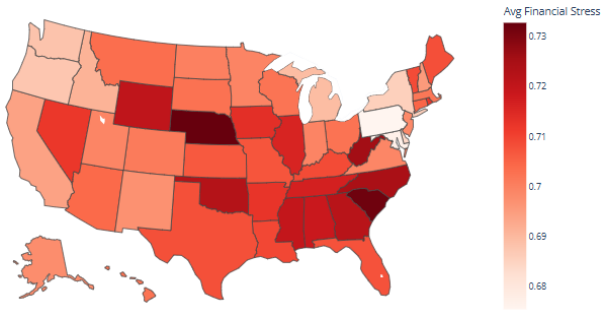


Figure 18: *
(c) Average Financial Stress by State (Higher = More Stress)

Similar to the earlier geographic breakdown of financial literacy, distinct regional patterns emerge. States in the Northwest and Northeast exhibit higher concentrations of high-value life insurance policies (those greater than or equal to $500k), with standouts including New York, Massachusetts, New Jersey, Connecticut, and New Hampshire. In contrast, states in the Southeast display markedly lower rates.

The distribution of high-income households follows a similar pattern. The highest concentrations are found in the Northeast and Northwest, particularly in New York, New Jersey, Maryland, and Washington. Hawaii also stands out—likely due to luxury real estate ownership. Meanwhile, the South, especially the Southeast, has a much lower share of high-income households.

The financial stress map further supports these disparities. States like Georgia, Mississippi, Alabama, and the Carolinas report elevated average stress scores. Conversely, the most financially stable states—such as Oregon, Washington, California, New York, and Maryland—cluster in the Northeast and Northwest. While states like California and New York may seem financially stressed on the surface, they also contain concentrated pockets of wealth, adding important nuance to the story—especially for large, demographically diverse states like California.

To further examine how financial and insurance variables relate at the state level, I constructed a correlation heatmap capturing pairwise relationships between key metrics.

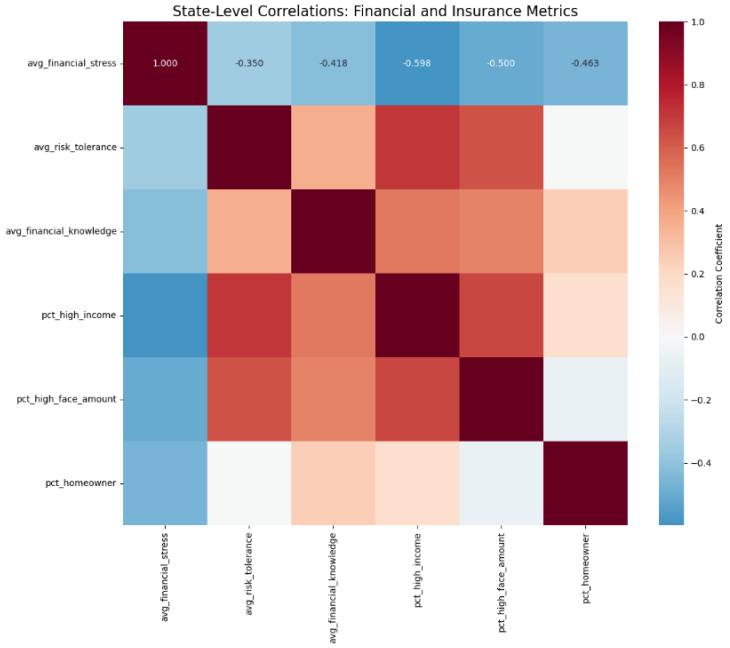


Figure 19: State-Level Correlations: Financial and Insurance Metrics

The heatmap reveals several meaningful patterns. Most notably, **average financial stress is negatively correlated with all other metrics**, with the strongest inverse relationship observed with the percentage of high-income households. This is consistent with expectations—states with more high-earning households tend to report lower average financial stress.

Financial knowledge shows weak to moderate positive correlations with several indicators: average risk tolerance, percentage of high-income households, percentage of high face-value insurance policies, and percentage of homeowners. These associations suggest that states with higher levels of financial understanding also tend to exhibit stronger financial footing across other dimensions.

The percentage of high-income households also correlates positively with multiple features: it is strongly correlated with risk tolerance, moderately with financial knowledge and high face-value policies, and weakly with homeownership. Together, these relationships suggest that state-level financial health indicators often move in tandem—better-informed, higher-earning populations also tend to hold stronger insurance coverage and demonstrate more financial confidence.

Additionally, the percentage of high face-value life insurance policies is moderately correlated with average risk tolerance, average financial knowledge, and the percentage of high-income households. This reflects the inherent complexity of life insurance as a financial product; understanding and purchasing higher-value policies often requires both a baseline level of financial literacy and a degree of risk tolerance to appreciate their long-term benefits. The correlation with high-income households is particularly intuitive, as individuals with greater financial means are more likely to require, and be able to afford, larger coverage amounts.

It's important to note that these findings represent **correlations at the state level**, not causality. For example, while financial knowledge and high-income prevalence are positively associated, this does not imply that one directly causes the other. Instead, these relationships likely reflect deeper structural and regional factors, such as differences in education access, financial culture, and local economic conditions.

To consolidate these findings, I developed a composite **Financial Wellness Score** that combines financial stress, risk tolerance, and financial knowledge. This offers a holistic picture of state-level financial health.

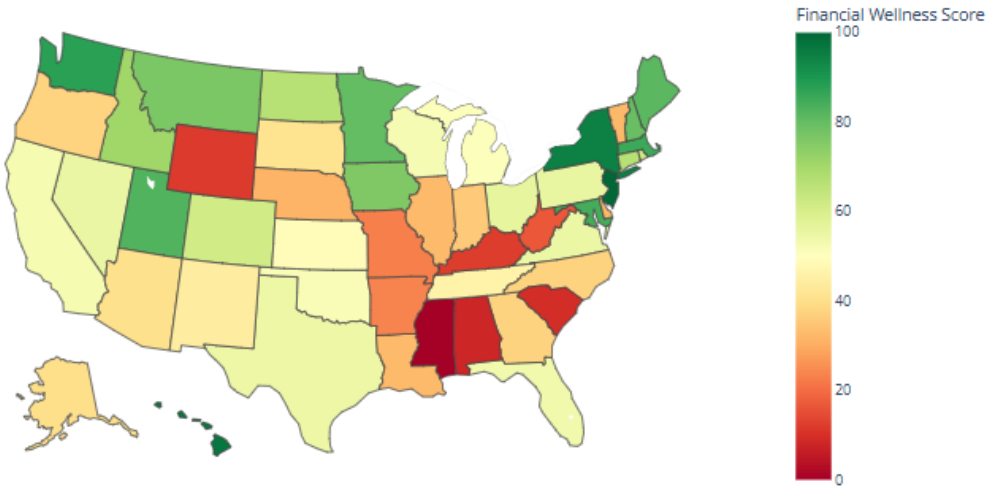


Figure 20: Financial Wellness Score by State

The resulting map highlights states with the strongest overall financial wellness—New York, New Jersey, Maryland, Massachusetts, Oregon, Utah, New Hampshire, Minnesota, and Montana—as well as those with the weakest scores, including Mississippi, Alabama, South Carolina, Kentucky, West Virginia, and Wyoming.

Finally, I created a proxy measure for **Insurance Market Strength** by averaging the percentage of high-value policies and high-income households in each state.

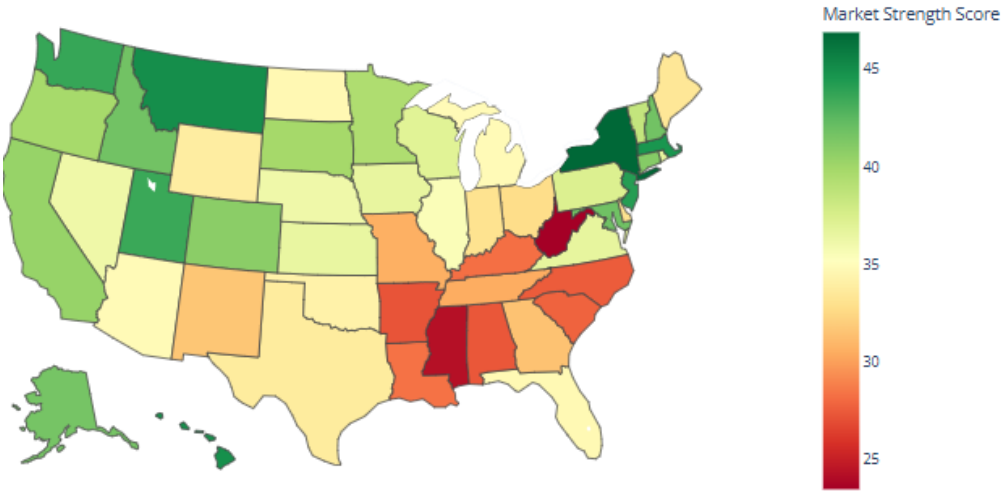


Figure 21: Insurance Market Strength by State

This measure paints a similar picture: the insurance market is strongest in the Northeast and Northwest, and weakest in the Southeast and parts of Appalachia, particularly West Virginia.
Taken together, these insights reveal stark disparities in financial preparedness, behavior, and access across U.S. states, pointing to deeper structural, economic, and cultural divides that must be addressed by policymakers, insurers, and financial educators alike.

Ultimately, this state-level correlation analysis reinforces prior findings from earlier visualizations and more advanced modeling. It confirms that geographic location is a strong contextual factor in financial behavior and insurance adequacy, and that clusters of high or low financial wellness often co-occur across multiple dimensions.

Lastly, I tested how financial education correlates with insurance adequacy. Specifically, I examined whether individuals who received financial education tend to hold higher-value life insurance policies.

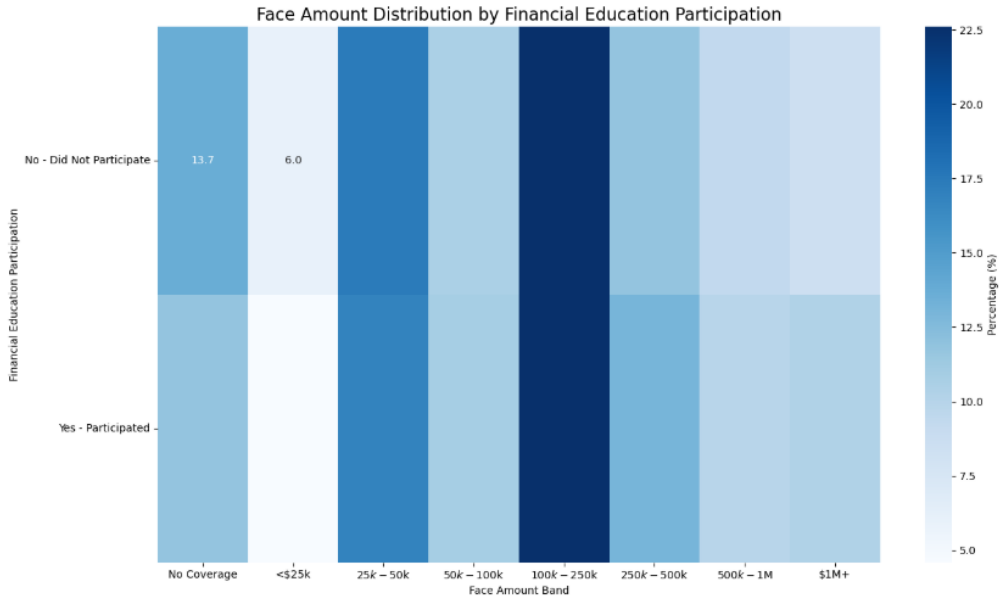


Figure 22: Distribution of Face Amounts by Financial Education Participation

The heatmap above highlights a clear divergence in insurance coverage based on financial education participation. Individuals who did *not* receive financial education show a higher density in the lower face

amount bands, particularly below $250k. In contrast, those who *did* participate in financial education exhibit more density across higher face amount bands, indicating broader coverage. Interestingly, both groups converge around the $100k – $250k range, suggesting this may be a natural breakpoint in coverage adequacy. However, individuals without financial education are far more likely to fall at or below this threshold.

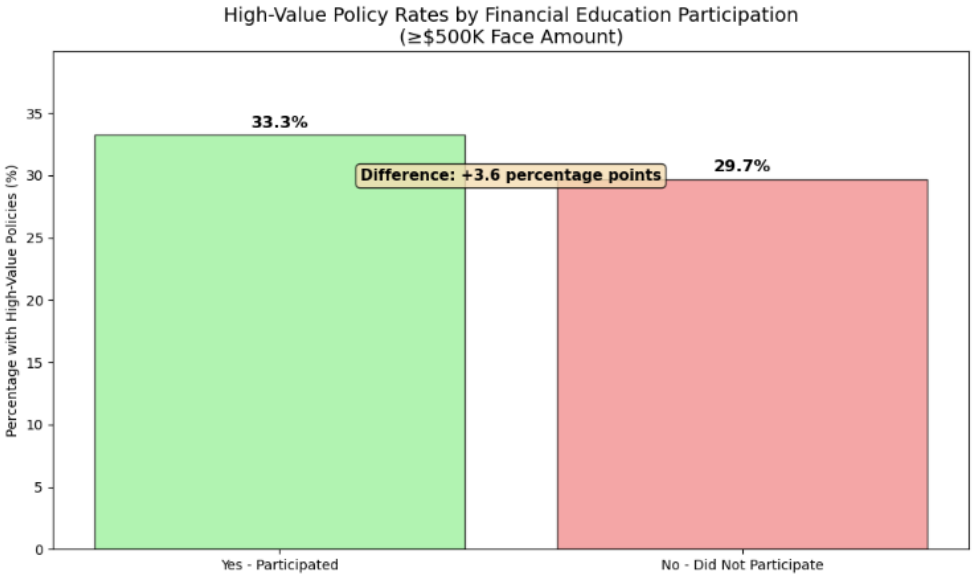


Figure 23: Rate of High-Value Policies by Financial Education Participation

To quantify the difference, I calculated the percentage of high-value policies, defined here as policies with a face value of $500k or more. Among individuals who had received financial education, this rate was **3.6% higher** than among those who had not. The difference is statistically significant, with a p-value of **less than 0.00001**. Moreover, the percentage of high-income households (those earning $100k or more annually) was also notably higher, by **7.03%**, among the financially educated group.

These results reinforce earlier findings: financial education is positively associated with stronger insurance coverage and higher income. While correlation does not imply causation, the consistent pattern across multiple variables and methods underscores the importance of early and ongoing financial education in shaping long-term financial outcomes.

5.3 UMAP and K-Means:

With the enriched dataset fully prepared, the next step was dimensionality reduction to project the high-dimensional data into a more tractable format for clustering. For this, I selected **UMAP (Uniform Manifold Approximation and Projection)** due to its ability to preserve both local and global data structure while remaining computationally efficient.

To ensure the strongest possible embedding, I built a robust UMAP pipeline with the support of Anthropic's Claude Sonnet 4. This pipeline incorporated hyperparameter tuning and cross-validation, testing 60 combinations across three key parameters: number of neighbors, minimum distance, and distance metric. For each configuration, four quality metrics were calculated to evaluate embedding performance:

Quality Metric	Description
`Trustworthiness`	Measures if points that are close in the embedding were actually close in the original high-dimensional plane
`Continuity`	Measures if points that were true neighbors in the original space remain close in the embedding
`Global Structure Preservation`	Uses Spearman correlation of pairwise distances to measure how well the relative ordering of distances between all point pairs is preserved
`Clustering Quality`	Best silhouette score across different cluster numbers

Table 9: Summary of UMAP Quality Metrics

Each of the four scores was averaged into a composite metric across cross-validation folds. The UMAP configuration with the highest mean composite score across cross-validation folds was selected as the final projection.

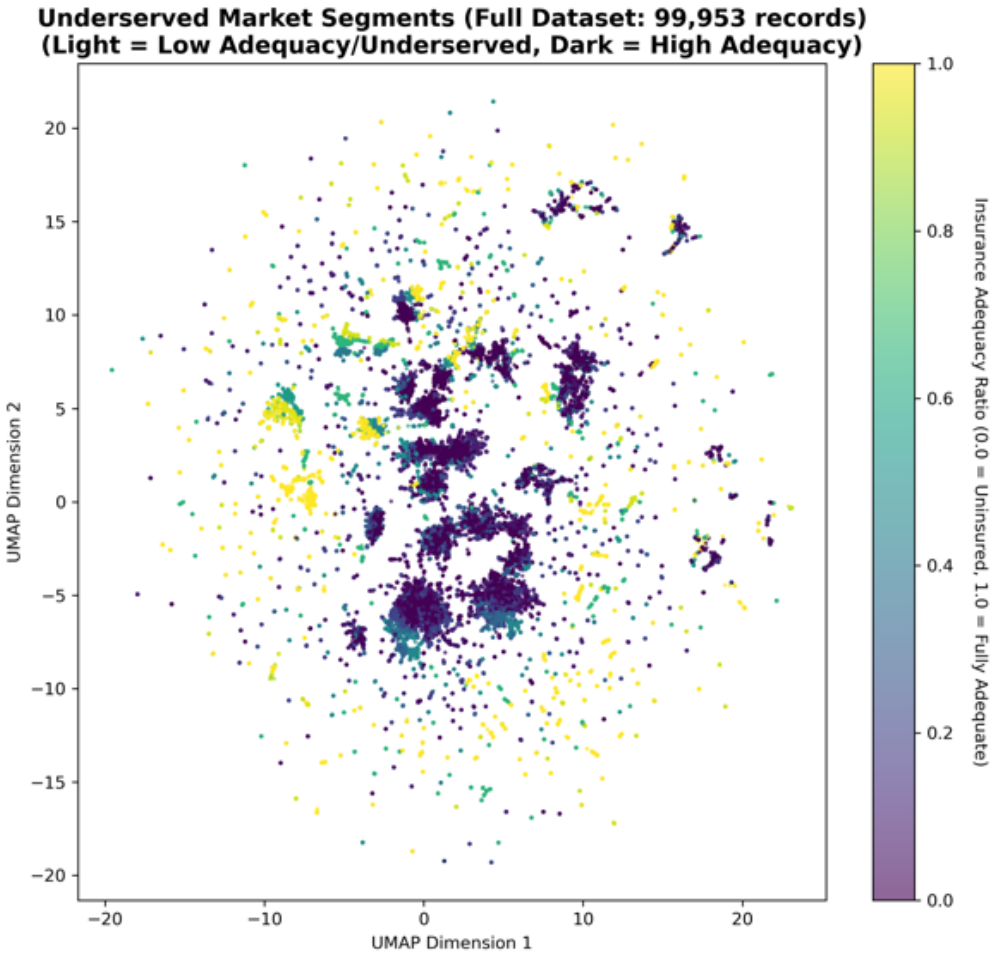


Figure 24: Optimal UMAP

As seen in Figure 24, individuals who are uninsured or critically underinsured tend to cluster near the center of the graph, while smaller pockets of adequately insured individuals appear on the periphery. This structure supports the hypothesis that underinsurance is widespread and potentially concentrated among demographically similar groups.

With a meaningful two-dimensional embedding in place, I applied **K-Means clustering** to identify

latent groups within the data. Reducing dimensionality beforehand ensures that Euclidean distance (used by K-Means) remains a reliable metric for measuring similarity. To determine the optimal number of clusters, I used silhouette scoring, selecting the number of centroids that maximized clustering quality. The algorithm identified **three clusters** as the optimal configuration.

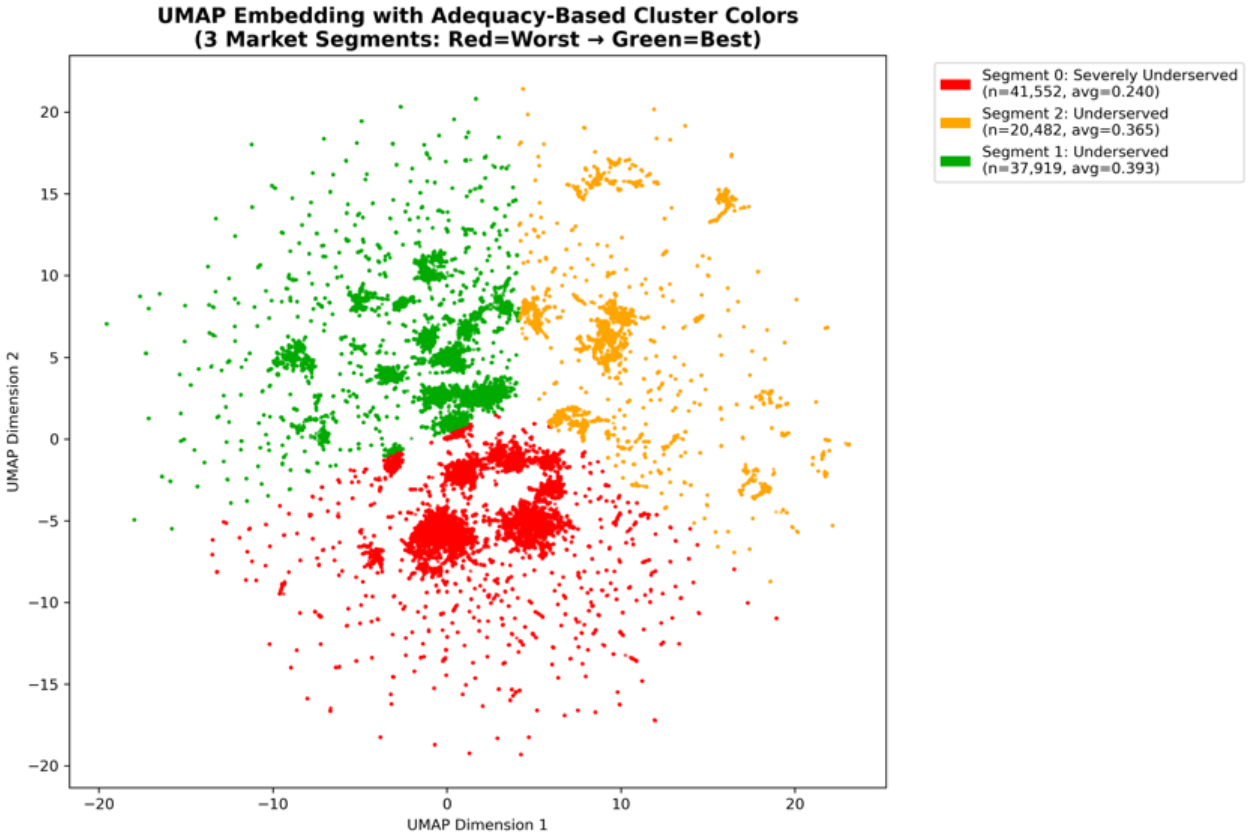


Figure 25: Optimal K-Means

Figure 25 illustrates the three identified clusters, each colored by its average insurance adequacy score. The clustering reveals clear differentiation in insurance adequacy levels across segments, with detailed segment characteristics presented in the Results section. This sets the stage for deeper analysis into the behavioral and demographic traits that define each group.

5.4 ANOVA:

Following cluster identification, I employed **Analysis of Variance (ANOVA)** to statistically validate segment differentiation. ANOVA tests whether there are significant differences between cluster means across all features, ensuring the segments represent distinct market groups rather than arbitrary data divisions.

To develop actionable business personas, I calculated mean values for all features within each cluster to compare inter-cluster differences. This approach enables identification of the most discriminating characteristics for each segment, providing the foundation for targeted business strategies.
The persona development process involved:

- Statistical validation through ANOVA testing across 239 features
- Identification of the most differentiating variables using effect size analysis
- Integration of demographic, financial, and behavioral characteristics

- Creation of actionable segment profiles for business application

This methodology ensures that identified segments are both statistically robust and business-relevant.

5.5 Feature Importance:

The final step in my analysis involved identifying the most influential features driving both insurance adequacy and cluster differentiation. To accomplish this, I developed an advanced machine learning pipeline with the help of Anthropic's Claude Sonnet 4. The pipeline independently runs both a **Random Forest** and **XGBoost** model on the same dataset, evaluates each using cross-validated accuracy and AUC, aggregates feature importance rankings, and generates side-by-side visual comparisons. It also includes automated detection for potential data leakage by flagging suspicious patterns and correlations.

Model	Insurance Adequacy	Market Segments
Random Forest	87.6%	96.9%
XGBoost	87.5%	96.4%

Table 10: Ensemble Method Performance

Both models achieved high and realistic performance levels, avoiding the telltale signs of overfitting. While Random Forest slightly outperformed XGBoost on both tasks, I ultimately selected **XGBoost** for final interpretation due to its ability to uncover a more diverse and insightful range of important features. Specifically, it produced an insurance adequacy accuracy of **87.5%** and a market segmentation accuracy of **96.4%**.

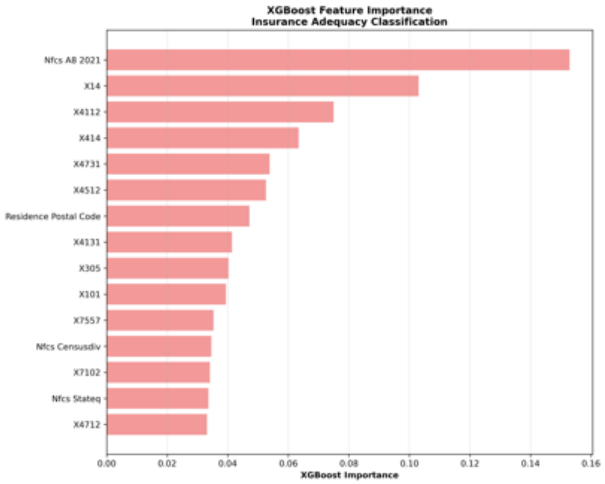


Figure 26: *
(a) Important Features for Insurance Adequacy

Figure 27: *
(b) Important Features for Segment Classification

Figure 28: Significant Features for Insurance Adequacy and Segment Prediction

As shown in Figure 28, XGBoost identified a rich mix of demographic, financial, behavioral, and geographic variables contributing to prediction quality. Feature importance was derived from a composite measure that combines two complementary methods:

1. **Built-in XGBoost Feature Importance:** This native attribute measures importance based on **frequency** (how often a feature is used to split nodes across all trees) and **gain** (the improvement in the objective function when using that feature to split).
2. **SHAP (Shapley Additive ExPlanations) Values:** Offers a more nuanced view by quantifying each feature's marginal contribution to individual predictions, accounting for non-linearity and feature interactions.

This hybrid approach provides a deeper understanding of the true drivers behind model predictions. While frequency and gain-based metrics capture commonly used split features, SHAP values elevate those with high explanatory power that may not appear frequently but significantly influence outcomes. This dual perspective is likely what gave XGBoost its advantage in surfacing less obvious, yet highly relevant, features—ultimately providing a more complete picture of the behavioral and demographic patterns that define insurance adequacy and cluster membership.

6 Results

6.1 Hypothesis Evaluation:

Hypothesis #	Statement	Result
H1	Financial metrics vary across regions, indicating systemic differences.	True
H2	Financial education correlates with higher insurance adequacy.	True
H3	Dimensionality reduction and clustering will reveal distinct market segments.	True
H4	Higher financial stress is associated with underinsurance.	False
H5	Behavioral and psychological factors are strong predictors of insurance coverage and should be more heavily emphasized in industry models and segmentation strategies.	True

Table 11: Evaluation of Key Hypotheses

This project validated four out of five initial hypotheses, each offering key insights that shaped the modeling approach. Regional differences (H1) were confirmed, with the Northeast and Northwest showing strong financial and insurance health, while the Southeast lagged across nearly all indicators. Participation in financial education (H2) was associated with a meaningful upward shift in coverage distribution; those without education showed greater density in lower face amount bands, while those with education were more likely to hold higher coverage. Unsupervised clustering (H3) proved effective, as UMAP and K-Means revealed three statistically distinct market segments. However, H4 was disproven: the most financially stressed group had the highest insurance adequacy, likely due to lower income thresholds making the 10× benchmark more attainable. Finally, H5 was supported, as XGBoost identified both demographic and behavioral traits - like financial confidence and decision-making style - as key predictors, reinforcing the value of psychologically informed segmentation strategies.

6.2 Market Segment Analysis:

The clustering analysis identified two distinct behavioral customer segments representing approximately 80% of the market, plus a third group characterized by survey non-engagement patterns that requires separate analytical treatment.

- **Segment 0: Affluent Self-Reliant Optimizers (41.6%):** Highest average income (**$96k**), lowest insurance coverage (**$224k**), lowest insurance adequacy ratio (**0.240**), highest financial knowledge scores (**3.34**), lowest confidence gap (**2.01**) indicating realistic self-assessment, low mortgage-to-home ratio (**14.8%**), highest risk tolerance (**4.68**), fewest credit cards (**2.63**), fewest financially dependent children (**3.8**), highest education level (**4.87**), and lowest financial stress (**0.681**). **Highly educated, financially sophisticated professionals who choose alternative wealth protection strategies over traditional life insurance.**
- **Segment 1: Financially Stressed Protectors (37.9%):** Lowest average income (**$59k**), moderate insurance coverage (**$380k**), highest insurance adequacy ratio (**0.393**), lowest financial knowledge

scores, highest mortgage-to-home ratio (**32%**), more credit cards (**3.18**), highest number of financially dependent children, lowest risk tolerance, lowest education level (**4.23**), and highest financial stress (**0.739**). **Risk-averse families under financial pressure who prioritize protection despite economic constraints.**

- **Segment 2: Low-Engagement Middle Ground (20.5%):** This segment is characterized by high rates of survey non-response ("Don't Know" and "Prefer not to say" responses), moderate income (**$73k**), mid-level education (**4.595**), moderate financial stress (**0.686**), and moderate number of financially dependent children (**4.23**). Their consistent "middle ground" positioning across key demographics may contribute to financial planning uncertainty—lacking both the clear protection urgency of highly stressed families and the confident sophistication of affluent self-reliant individuals. **Represents individuals avoiding financial complexity due to unclear decision drivers rather than distinct behavioral preferences.**

Primary Market Insights from Behavioral Segments (0 & 1):

The contrast between segments reveals a profound paradox in insurance behavior. **Segment 0 (Affluent Self-Reliant Optimizers)** possesses every advantage—highest income, education, financial knowledge, and risk tolerance—yet demonstrates the most critical insurance gap (**0.240 adequacy**). Their profile suggests deliberate choice rather than oversight: with few dependents, low financial stress, and sophisticated financial understanding, they may be employing alternative wealth protection strategies such as self-insurance through investments, trusts, or other financial instruments.

Segment 1 (Financially Stressed Protectors) represents the opposite extreme: despite facing the highest financial stress, supporting the most dependents, and having the lowest financial knowledge and income, they achieve the highest insurance adequacy ratio (**0.393**). Their risk-averse nature, combined with high dependent burden and financial vulnerability, drives responsible insurance behavior that prioritizes protection over other financial goals. This group's willingness to purchase coverage despite economic constraints demonstrates strong risk awareness and family-focused financial decision-making.

The behavioral differences extend beyond income to fundamental risk philosophies. Segment 0's high risk tolerance allows them to forgo traditional insurance protection, confident in their ability to self-insure or recover from financial setbacks. Segment 1's low risk tolerance, combined with high dependent obligations, makes insurance a necessity rather than an option, leading to more adequate coverage despite limited resources.

The distribution of the two primary behavioral segments (**79.5% combined**) reveals important market dynamics driven by clear decision catalysts at the extremes. The **20.5% Low-Engagement Middle Ground** group's moderate positioning across all key variables—income, education, financial stress, and family obligations—may paradoxically create decision paralysis rather than clear motivation. Unlike the extremes that generate obvious decision drivers (high sophistication enabling self-reliance vs. high stress demanding protection), this middle group lacks compelling catalysts in either direction, leading to financial planning avoidance.

Notably, all segments fall well below the adequacy threshold of 1.0, reinforcing that underinsurance spans across all demographic profiles. However, the 65% gap between the highest and lowest adequacy scores (0.393 vs 0.240) indicates that clear situational drivers produce more decisive insurance behaviors than moderate, balanced circumstances.

Methodological Note: The identification of the low-engagement group (Segment 2) through clustering analysis provides valuable insight into survey response patterns and highlights the importance of targeted engagement strategies for different customer types. Future analysis would benefit from separate modeling approaches for engaged versus non-engaged customer groups.

These findings fundamentally challenge conventional assumptions about the relationship between income and insurance adequacy. Financial capacity enables choice, but clear situational drivers—whether sophisticated confidence or urgent family protection needs—determine decisive action. The "middle ground" paradox reveals that moderate circumstances across multiple dimensions may actually inhibit decision-making, creating a substantial market segment that avoids financial planning due to unclear priorities rather than lack of need. The presence of distinct adequacy patterns driven by extreme rather than moderate circumstances provides a foundation for targeted intervention strategies that address both the psychological drivers of protection decisions and the decision paralysis created by balanced but unclear situations.

6.3 Feature Importance Analysis:

The XGBoost analysis revealed striking diversity in the features that drive insurance decision-making, spanning demographic, financial, behavioral, and geographic dimensions. This comprehensive feature set demonstrates that insurance adequacy cannot be predicted through simplistic models, but rather requires understanding the complex interplay of individual circumstances and attitudes.

6.3.1 Insurance Adequacy Prediction:

XGBoost identified fifteen key features that most accurately predict whether an individual maintains adequate life insurance coverage. These features reveal the multifaceted nature of insurance decision-making, encompassing not just financial capacity but also behavioral tendencies and environmental factors.

Feature	Description	Original Dataset
A8 2021	Approximate household income	NFCS
x14	Respondent's age	SCF
x4112	Amt earned before taxes (not self-employed)	SCF
x414	Total credit limit	SCF
x4731	Spouse's pre-tax income from owned business	SCF
x4512	Number of years respondent has worked full time	SCF
Residence Postal Code	5-digit ZIP code	MIB
x4131	Respondent's pre-tax income from owned business	SCF
x305	Attitude towards credit	SCF
x101	Number of people living in household	SCF
x7557	Willingness to take financial risk	SCF
CENSUSDIV	Census division	NFCS
x7102	Financial information Source #2	SCF
STATEQ	State	NFCS
x4172	Amt earned before taxes (self-employed)	SCF

Table 12: Top Features for Insurance Adequacy

The feature importance rankings reveal several critical insights for insurance adequacy prediction. **Income dominates the top predictors**, appearing in multiple forms including household income (A8 2021), employment earnings (x4112, x4172), and business income (x4731, x4131). This reinforces that financial capacity remains the primary constraint to adequate coverage. **Age emerges as the second most critical factor**, validating industry observations about declining insurance uptake among younger demographics and highlighting the importance of life-stage targeting.

Notably, **behavioral and attitudinal factors** prove equally important as pure demographics. Features like attitude towards credit (x305) and willingness to take financial risk (x7557) demonstrate that individual psychology significantly influences insurance decisions, suggesting that behavioral segmentation may be as valuable as traditional demographic approaches. The prominence of **geographic variables** at multiple levels (ZIP code, state, and census division) underscores the powerful role of local market conditions, regulatory environments, and cultural factors in shaping insurance adequacy patterns.

6.3.2 Market Segmentation Prediction:

The features that best differentiate between market segments reveal a nuanced picture where income dominance coexists with strong collective influence from life circumstances and family structure.

Feature	Description	Original Dataset
A8 2021	Approximate household income	NFCS
A6	Marital status	NFCS
A7	Living arrangements	NFCS
M4	Self-assessed financial knowledge	NFCS
F1	Credit card ownership	NFCS
x14	Respondent's age	SCF
A7A	Marital status (general)	NFCS
x4112	Amt earned before taxes (not self-employed)	SCF
EA 1	Home ownership	NFCS
x414	Total credit limit	SCF
x101	Number of people living in household	SCF
FinancialStress	Engineered metric for financial stress	NFCS
A11	Number of financially dependent children	NFCS
A10	Spouse's employment status	NFCS
x4512	Number of years respondent has worked full time	SCF

Table 13: Top Features for Market Segmentation

The segmentation features reveal that **household income (A8 2021) emerges as the single most dominant predictor**, far exceeding other individual variables in importance. However, **family structure variables collectively demonstrate powerful combined influence** on segment differentiation. Marital status (A6, A7A), living arrangements (A7), dependent children (A11), and spouse employment (A10) together create a constellation of family dynamics that fundamentally shape financial priorities and insurance needs. **Financial sophistication and stress levels** also prove critical, with self-assessed financial knowledge (M4) and engineered financial stress metrics distinguishing how different segments approach financial planning.

This pattern differs notably from insurance adequacy prediction, where multiple income sources dominated the top predictors. For segmentation, **income leads as a single factor, but family circumstances collectively rival its importance**. Variables like home ownership (EA 1), work experience (x4512), and household composition (x101) reinforce that life stage stability and family obligations strongly influence segment membership. This finding validates the earlier observation that while income capacity enables choice, **family dynamics and life circumstances determine which segment individuals belong to**, explaining why high earners (Segment 0) can still be severely underinsured based on their life stage and risk attitudes rather than earning capacity alone.

6.3.3 Feature Diversity Implications:

The breadth of important features across both prediction tasks demonstrates that effective insurance strategies must address multiple dimensions simultaneously. Traditional approaches focusing solely on income and age miss critical behavioral, geographic, and life-circumstance factors that significantly influence both adequacy levels and segment membership. This feature diversity provides the foundation for developing comprehensive, multi-dimensional approaches to market segmentation and product targeting.

7 Recommendations

7.1 Segment 0: Affluent Self-Reliant Optimizers

Strategic Approach: Sophisticated Alternative Wealth Protection

Core Messaging: Your self-reliance strategy is working—let's optimize it. Insurance as strategic diversification, not traditional protection.

7.1.1 Product Strategy

- Lead with **sophisticated permanent life products** as tax-advantaged investment vehicles
- Offer **variable universal life** with premium investment options and flexibility
- Present **business succession and estate planning** solutions leveraging life insurance
- Emphasize **tax optimization strategies** rather than basic protection needs

7.1.2 Sales Approach

- **Respect their sophistication**—position insurance as portfolio diversification tool
- **Advanced financial modeling** showing insurance within broader wealth strategies
- **Challenge their assumptions** about self-insurance efficiency vs. leveraged protection
- **Peer influence** from other sophisticated investors who use insurance strategically

7.1.3 Marketing Channels

- **Wealth management partnerships** and private banking relationships
- **Estate planning attorney networks** and tax advisor referrals
- **Executive forums** and high-net-worth professional events
- **Social media thought leadership content** on advanced insurance strategies

7.1.4 Key Success Metrics

- Premium volume per policy (targeting sophisticated, high-value products)
- Cross-sell success with investment and estate planning services
- Advisor referral network development and conversion rates

7.2 Segment 1: Financially Stressed Protectors

Strategic Approach: Maximize Protection Within Constrained Budgets

Core Messaging: Your family-first instincts are exactly right. Let's make sure that protection fits your budget today and tomorrow.

7.2.1 Product Strategy

- Focus on **maximum coverage term life** with guaranteed level premiums
- Offer **group/employer-sponsored** products with automatic payroll deduction
- Develop **family protection bundles** combining spouse and children coverage efficiently
- Create **premium holiday options** for financial hardship periods

7.2.2 Sales Approach

- **Validate their priorities**—recognize they're already making smart protection choices
- **Budget-conscious planning** with clear affordability guidelines
- **Life event optimization**—timing increases with income growth or family changes
- **Simplified underwriting** to reduce barriers and speed up protection implementation

7.2.3 Marketing Channels

- **Employer benefits partnerships** and workplace financial wellness programs
- **Community organization outreach** through schools, religious institutions, unions
- **Mobile-first, simplified** application and service experiences
- **Family-focused social campaigns** emphasizing responsible protection behavior

7.2.4 Key Success Metrics

- Policy persistency rates (critical for financially stressed customers)
- Coverage-to-income ratios (measuring protection adequacy improvements)
- Customer satisfaction with service and claims experience

7.3 Segment 2: Low-Engagement Middle Ground

Strategic Approach: Building Financial Confidence Through Education

Core Messaging: Making financial protection simple, accessible, and pressure-free. Let's start with what matters most to you.

7.3.1 Engagement Strategy

- Lead with **financial education workshops** focused on basic concepts
- Offer **no-pressure consultation sessions** with educational focus
- Develop **simple, transparent product explanations** without jargon
- Create **step-by-step decision-making tools** to build confidence

7.3.2 Product Strategy

- Start with **basic term life insurance** with clear, simple benefits
- Offer **online application processes** with minimal complexity
- Provide **flexible coverage amounts** starting with affordable options
- Develop **bundled packages** that simplify decision-making

7.3.3 Communication Approach

- **Build trust first** - focus on education over sales
- **Use multiple touchpoints** - email, text, phone with opt-out options
- **Leverage social proof** - testimonials from similar customers
- **Gamify learning** - interactive tools and progress tracking
- **Address survey fatigue** - shorter, more engaging information gathering

7.3.4 Marketing Channels

- **Digital-first approach** with self-service options
- **Community partnerships** through local organizations
- **Workplace benefits education** sessions
- **Social media campaigns** with educational content
- **Referral programs** from current customers

7.3.5 Key Success Metrics

- Engagement rate improvement (survey completion, seminar attendance)
- Conversion from education to consultation
- Customer progression through the sales funnel
- Financial literacy score improvements

7.4 Priority Allocation

7.4.1 Immediate Focus (Next 6 Months):

1. **Segment 0** - Highest ROI opportunity with biggest coverage gaps and financial sophistication
2. **Segment 1** - Strong insurance behavior despite constraints, immediate conversion potential

7.4.2 Medium-term Development (6-18 months):

1. **Segment 2** - Long-term engagement strategy requiring educational foundation and trust-building

7.4.3 Strategic Rationale:

The low-engagement group represents a significant opportunity that requires a fundamentally different approach. Rather than attempting immediate product sales, the strategy focuses on **building financial confidence and engagement** as prerequisites to insurance discussions. This segment's 20.5% market share, combined with moderate income levels ($73k), suggests substantial volume potential once engagement barriers are addressed.

Success with this group requires patience and a customer education investment, but the payoff could be substantial given their underrepresentation in traditional insurance marketing. The approach acknowledges that their non-response patterns likely stem from **financial complexity avoidance** rather than lack of need, requiring trust-building and simplification strategies.

8 Summary

Conventional wisdom suggests that higher incomes naturally translate to more adequate insurance coverage, yet my analysis reveals the opposite. When data is enriched and properly segmented, **the highest-earning group ($96k) demonstrates the lowest insurance adequacy score (0.240)**, exposing a fundamental flaw in industry assumptions. While universal underinsurance affects all identified segments, with none achieving the 1.0 adequacy threshold, the 65% performance gap between the most and least covered segments reveals substantial differentiation in protection behaviors that transcends simple income metrics.

Traditional income-based targeting fails because it ignores the behavioral, psychological, and geographic factors that actually drive insurance decisions. Critically, income remains a powerful predictor, but operates counter to industry assumptions with higher earners demonstrating systematically lower coverage adequacy. Financial literacy, confidence gaps, life circumstances, and regional influences prove far more predictive of

coverage adequacy than earning power alone. The analysis further reveals that **20.5% of the market actively disengages from financial planning conversations**, creating a substantial opportunity for educational intervention strategies. This finding validates the critical role of financial education in driving insurance coverage and highlights the practical challenge of reaching customers who avoid financial complexity.

The precision targeting and behavioral segmentation methods demonstrated in this analysis offer insurance decision-makers a clear path to transform their acquisition strategies. Rather than chasing demographics, the industry can now target the psychological and behavioral drivers that actually determine coverage decisions while developing specialized engagement strategies for financially disengaged customers. The identification of two distinct behavioral segments representing 79.5% of the market, combined with a systematic approach to the low-engagement group, represents a fundamental market disruption opportunity—one that rewards companies sophisticated enough to move beyond traditional assumptions toward data-driven customer understanding and targeted financial education initiatives.

9 Limitations

Generalizability:

This analysis leverages nationally representative datasets (NFCS and SCF) integrated with industry-scale MIB data representing a stratified sample of 100,000 actual insurance applications out of over 5.4 million. The demographic rebalancing procedures ensured alignment with U.S. population distributions, supporting broad generalizability to the American insurance market. However, findings may be most applicable to markets with similar regulatory environments and insurance industry structures.

Methodological Choices:

My selection of the variance-adjusted financial stress measure, while chosen for normality properties to support clustering algorithms, represents one of several possible approaches. Alternative measures exhibited right-skewed distributions that might have yielded different segmentation patterns. A more comprehensive analysis comparing multiple stress measures throughout the entire analytical pipeline could provide additional validation of segment stability.

Data Integration Complexity:

The merge process between the three datasets, while executed successfully using multiple matching criteria and validation checks, represents a complex undertaking. Although the matching procedures produced reasonable coverage and passed standard validation tests, more sophisticated integration approaches might capture additional nuanced relationships between the datasets.

Data Quality and Survey Response Patterns:

The analysis relies heavily on third-party survey data, which introduces inherent response bias concerns. Participants may have provided socially desirable responses or had varying interpretations of financial concepts, potentially affecting the accuracy of self-reported financial behaviors and attitudes.

Notably, the clustering analysis revealed that one segment (20.5% of respondents) demonstrated systematic non-response patterns to financial literacy questions, frequently selecting "Don't Know" or "Prefer not to say" options. While this pattern provided valuable insights into market engagement levels and validated the importance of financial education initiatives, it also highlights the challenge of analyzing customers who actively avoid financial complexity. Future research would benefit from alternative data collection methods for this population, such as behavioral observation rather than self-reported surveys.

10 References

10.1 Data Sources:

- **FINRA National Financial Capability Study:** [Link]
- **Federal Reserve Survey of Consumer Finances:** [Link]

10.2 Literature Survey:

- **Uniform Manifold Approximation and Projection (UMAP) and its Variants: Tutorial and Survey:** [Link]
- **Clustering with UMAP: Why and How Connectivity Matters:** [Link]
- **Analysis of Variance (ANOVA) Comparing Means of more than Two Groups:** [Link]
- **Methodology and Application of One-Way ANOVA:** [Link]
- **Random Forests:** [Link]

10.3 Artificial Intelligence:

- **Claude Sonnet 4 - Anthropic**