Socio-economic/Demographic Analysis for Marketing Trends - Report

Section 1: Classification Modelling

1.1 Data Loading and Initial Preprocessing

The classification pipeline began with a systematic ingestion and auditing phase to establish a high-fidelity dataset for predictive modeling.

1.1.1. Data Ingestion and Structural Audit

The Census Bureau dataset was loaded with explicit handling for missing value markers (e.g., "?", "NA", "NaN") to ensure they were correctly recognized as null entities during the initial parse.

- Initial Volume: 199,523 observations across 42 attributes.
- Feature Classification: Attributes were programmatically split into 8 numerical and 32 categorical features. This classification was achieved through a multi-step inference technique: coercing columns to numeric, calculating numeric coverage based on a 95% parsability threshold, and performing a cardinality analysis of unique ratios.
- Data Integrity: A total of 3,229 duplicate records were identified and removed, resulting in a final analytical set of 196,294 unique observations.

1.1.2. Strategic Missing Value Resolution

Missing values were managed using a structured, deterministic strategy to avoid the bias often introduced by simple mean or mode imputation:

- Mechanistic Attribution and Imputation: Analysis indicated that categorical variables exhibited varying types of missingness. Attributes like migration code were classified as MNAR (Missing Not At Random), as the absence of data likely represented a specific state (non-movers); these were assigned an "NA" category to preserve this "missing signal" for model training. Features such as Hispanic origin and country of birth were treated as MAR (Missing At Random) and resolved via deterministic mapping, using crosstabulation to impute values based on established demographic correlations. Remaining nulls with no structural pattern were classified as MCAR (Missing Completely At Random) and assigned a neutral "NA" label to maintain integrity without distorting distributions.
- Deterministic Mapping (hispanic origin): A conditional probability analysis was performed using crosstabulation and deterministic mapping between birth countries and Hispanic origin. For instance, individuals born in China or Taiwan were mapped to "All other" based on observed frequencies.
- Numerical Integrity: Both the target (income_binary) and weight columns were verified to have zero missing values and logically consistent ranges. All other Numeric columns had 0 missing values.

1.1.3. Feature Normalization and Type Casting

To prepare the data for Gradient Boosting (XGBoost), strict consistency in feature formatting was enforced:

- Whitespace Normalization: Techniques including astype(str) conversion and str.strip() were used to normalize categorical strings (notably in citizenship and major occupation code) to prevent redundant category creation.
- Target Encoding: The label was mapped from strings to a binary integer format using a dictionary map: 0 (Income <$50k) and 1 (Income >$50k).
- Optimization of Data Types: Categorical attributes were explicitly cast to the category dtype using a programmatic loop to reduce memory footprint and improve training efficiency. Numeric types were encoded to int64 and float64 dtypes.

1.2 Numeric Exploratory Data Analysis (EDA)

1.2.1. Dataset Context and Population Structure

The dataset includes 199,523 unweighted observations with a strongly imbalanced income split (~93.8% ≤50K, ~6.2% >50K). Because the >50K class is small, tail behavior can appear noisy in some plots; however, weighted analysis confirms the same relationships hold at the population level, indicating the patterns reflect true structure rather than sampling artifacts.

1.2.2. Workforce Participation as the Primary Income Driver

Weeks worked in year is the strongest numeric separator. The weighted median is 4 weeks (≤50K) versus 52 weeks (>50K), showing high-income individuals are overwhelmingly full-year participants. Distributions consistently show ≤50K contains a large mass at 0/intermittent work, making employment continuity a core driver of income stratification.

1.2.3. Investment Income as a High-Value Economic Signal

Capital gains and dividends are extremely zero-inflated and right-skewed; most individuals report 0, but non-zero cases are concentrated in the >50K group. Mean capital gains increase from $144 (≤50K) to $4,831 (>50K), with dividends showing a similar separation. Log/weighted plots confirm investment income captures "hidden" wealth signal tied to asset ownership.

1.2.4. Age as a Proxy for Economic Maturity

Age distributions shift older for high-income individuals: mean age rises from 33.7 (≤50K) to 46.3 (>50K). Weighted plots confirm >50K concentrates in mid-to-late career ranges, while the ≤50K group includes more young and non-working individuals, reinforcing age as a proxy for workforce maturity and stability.

1.2.5. Wage Structure and Sparse Reporting Effects

Wage per hour is structurally sparse (≥75% zeros), reflecting reporting/in-universe limitations rather than absence of earnings. When wage is present, log-weighted distributions show >50K has greater density in higher wage ranges, indicating wage intensity contributes to separation but must be interpreted alongside workforce participation.

1.2.6. Occupational Structure and Multivariate Drivers

Weighted occupational patterns show >50K individuals cluster in professional/managerial/specialized roles, while ≤50K is more represented in service/labor/non-workforce categories. Pairwise relationships indicate income is driven by interaction of weeks worked, investment income, age, and occupational positioning, not a single variable.

1.3 Categorical Exploratory Data Analysis (EDA)

1.3.1. Dataset Context and Structural Imbalance

The same imbalance persists categorically (199,523 total; 93.79% ≤50K, 6.21% >50K), so probability-based comparisons were emphasized over raw counts to avoid category-size distortion.

1.3.2. Education Level as the Primary Income Indicator

Education has the strongest categorical association with income probability. Advanced degrees show the highest >50K likelihood (Professional/Doctorate ~52–54%, Master's ~31%, Bachelor's ~20%), while below–high-school levels are typically <3%, and children show 0%. This supports education as a structural driver through access to skilled roles and stability.

1.3.3. Occupation, Industry, and Employment Structure Effects

Managerial/professional occupations show the highest >50K probabilities (~25–27%), while service/labor/farming roles are typically <6%; certain specialized occupation codes exceed 60%. Industries such as professional services, finance, and public administration show higher income representation than retail and agriculture. Employment structure reinforces this: self-employed incorporated (~35.4%) and government roles show higher income probability, while "Not in universe" is near-zero and reflects non-working segments.

1.3.4. Demographic and Structural Interpretation (Descriptive)

Sex-based differences are stable (males show higher >50K representation), reflecting workforce distribution and role composition. Race shows probability differences (e.g., higher representation for Asian/Pacific Islander and White relative to Black), but these patterns are descriptive and largely reflect interactions with education, occupation, and industry rather than causal effects. Structural categories such as Children and Not in universe represent workforce non-participation and must be interpreted accordingly. Because both sex categories have large sample sizes, this relationship is statistically stable and not driven by sampling variation.

1.3.5. Structural Categories and Non-Working Population Effects

Certain categorical groups represent structural non-participation rather than workforce roles. Categories such as "Children," "Not in universe," and certain rare occupation codes show extremely low-income probabilities due to structural absence from the workforce. These categories are important for segmentation but must be interpreted carefully, as they reflect demographic and structural positioning rather than economic productivity.

1.4 Grouped Exploratory Data Analysis (EDA)

1.4.1. Education as the Primary Structural Driver of Workforce and Wealth

Grouped analysis confirms that education strongly structures workforce participation, wealth accumulation, and income probability. Education shows the largest variation in numeric group means, with spreads of approximately 49.8 years in age, 6,749 in capital gains, and 42.8 weeks worked across education levels. Weighted summaries show that advanced degrees (Professional, Doctorate, Master's, Bachelor's) have significantly higher workforce participation (≈39–43 weighted mean weeks worked) compared to low education levels (≈20–25 weeks). These groups also show higher weighted capital gains and dividend income, confirming that education aligns with both stronger workforce attachment and greater wealth-generating potential.

1.4.2. Occupation, Industry, and Class of Worker Strongly Govern Workforce Attachment

Correlation ratio (η) analysis shows extremely strong structural relationships between workforce categories and weeks worked: major occupation (η ≈ 0.883), major industry (η ≈ 0.882), and class of worker (η ≈ 0.879). This confirms that employment classification directly governs workforce participation intensity. Managerial and professional occupations consistently show higher weeks worked and income probability, while structural categories such as "Not in universe" show near-zero workforce participation. Class-of-worker grouping further reveals that self-employed incorporated individuals and government workers show high workforce stability and elevated capital gains, confirming that employment structure influences both workforce continuity and economic capacity.

1.4.3. Capital Gains and Wealth Events Strongly Differentiate Income Across Workforce Groups

Zero-inflation grouped analysis confirms that capital gains events are disproportionately concentrated in high-income groups. Across major education levels such as Bachelor's, Master's, and High School, the >50K group shows approximately 3–6× higher non-zero capital gains rates compared to ≤50K individuals. Additionally, mean capital gains increase significantly in advanced education and professional workforce groups, confirming that investment-based wealth accumulation plays a critical role in income differentiation. In contrast, wage per hour shows structural reporting limitations, with many high-income professional and self-employed individuals reporting zero wage/hour, indicating that wage/hour is not a reliable standalone income indicator in grouped analysis.

1.4.4. Age, Career Progression, and Workforce Lifecycle Effects

Grouped age interactions confirm that income probability increases sharply with workforce maturity, peaking in the 46–55 age group, and declining beyond retirement ages. Education consistently shifts income probability upward within each age group, confirming that higher education accelerates career progression and income growth. Occupational interaction plots further show that managerial and professional occupations demonstrate increasing income probability with age, while non-working and structural categories remain low across all age groups. These patterns confirm that income growth reflects cumulative workforce experience and sustained employment participation.

1.4.5. Employment Status and Workforce Continuity as Direct Income Determinants

Employment grouping analysis confirms that workforce participation intensity directly determines income outcomes. Full-time employment categories show consistently high workforce participation (≈52 weeks worked median), while "Not in labor force" and structural non-working groups show near-zero workforce participation. Average weeks worked is consistently higher in the >50K group across nearly all worker classes, confirming that workforce continuity is one of the strongest structural predictors of income. Organizational structure analysis further shows that government and private sector employees operate in larger organizational environments (median employer size ≈4–6 workers), while self-employed individuals operate independently, reflecting differing economic structures and income pathways.

1.4.6. Gender and Occupational Interaction Effects on Income Probability

Grouped interaction analysis confirms persistent income probability differences between males and females across all workforce age groups, with the largest gap observed during peak career stages (46–55). Even within the same occupational groups, males show higher income probability, indicating structural differences in occupational positioning and workforce progression. Combined education and occupation grouping further confirms that the highest income probability combinations occur in advanced education and professional occupational roles, validating that income stratification is driven by the interaction of education, workforce participation, occupational structure, and wealth accumulation.

1.5 Industry Trends and Workforce Structure Analysis

Industry-level grouped analysis reveals clear structural differences in income probability driven by workforce stability, occupational composition, and education levels. Industries such as Professional services, Finance/Insurance/Real Estate, and public administration consistently show higher probability of earning >50K, supported by stronger workforce attachment with average weeks worked close to full-year employment levels, older workforce profiles, and higher concentration of advanced educational attainment (Bachelor's, Master's, and Professional degrees). These industries also exhibit higher average capital gains and dividend exposure, indicating greater presence of senior roles and wealth-generating positions. In contrast, industries such as Retail trade, Agriculture, and Personal services demonstrate significantly lower income probability, associated with lower average weeks worked, younger workforce composition, and lower education levels, reflecting reduced workforce stability and limited upward income progression. Moderately positioned industries, including Manufacturing, Construction, and Transportation, show mixed income outcomes due to variation in occupational roles, where technical and specialized positions demonstrate higher income probability than operational or labor-intensive roles.

Interaction analysis across industry, age, and workforce participation further confirms that income progression is strongly linked to sustained employment and career advancement within higher-skill industries. Industries with higher workforce continuity and skilled occupational composition demonstrate greater income probability and workforce stability, while labor-intensive industries show structural constraints due to intermittent employment patterns and lower skill requirements. These findings validate that industry-level income stratification is primarily driven by differences in workforce participation intensity, educational attainment, occupational specialization, and employment continuity rather than industry classification alone. This structural differentiation provides important insight for segmentation and workforce modeling by highlighting the role of industry as an indirect indicator of economic engagement and income potential.

1.6 Feature Engineering and Analysis

The feature engineering phase was designed to transform the raw census data into a highly predictive, low-noise "Hybrid" feature space. The objective was to capture complex socio-economic interactions while strictly controlling for dimensionality, redundancy, noise and information loss.

1.6.1. Initial Feature Space and Financial Transformations

The baseline feature set contained 41 predictors (numeric + high-cardinality categorical). EDA showed strong zero-inflation and right-skew in financial and work variables, so multiple representations were engineered for capital gains, capital losses, dividends, wage per hour, weeks worked, and investment income (raw, log1p, nonzero flags, and work_intensity) to identify the best signal. These candidates were evaluated using Random Forest feature importance + ROC-AUC and validated with XGBoost SHAP; the initial benchmark achieved RF ROC-AUC = 0.94917, and the corresponding boosted-model benchmark on the full feature set achieved XGBoost ROC-AUC = 0.954094.

Redundancy was removed by testing representation groups using importance consistency + ROC-AUC deltas. The retained numeric encodings were wage per hour (raw), capital gains_log1p, capital losses_log1p, dividends_log1p, investment_income, and weeks worked / work_intensity, while other variants were dropped. A separate wealth test confirmed that splitting components preserves signal: (capital gains_log1p + dividends_log1p) ROC-AUC = 0.949298 outperformed investment_income ROC-AUC = 0.948252. After numeric redundancy pruning, the benchmark improved to RF ROC-AUC = 0.94949.

Categorical redundancy was handled by grouping correlated granularities—occupation (major vs detailed), industry (major vs detailed), and household structure (two detailed fields)—and selecting the stronger member using the same importance + ROC-AUC criterion under high-cardinality encoding. Final pruning reduced features from 41 → 37 (Dropped: 1 numeric, 3 categorical) and improved both models: RF ROC-AUC = 0.949818, while XGBoost improved to ROC-AUC = 0.954502 on the pruned feature set.

1.6.2. Engineered Interaction Variables

To move beyond isolated demographic attributes, interaction features were engineered to better represent household structure and employment context while controlling redundancy and cardinality. The feature combination analysis focused on merging correlated columns that carried complementary ("synergistic") information so that signal was retained without inflating category cardinality, and to ensure important context variables were not missed in the final feature set.

- marital_tax_combo: Cross-feature combining marital status and tax filer status to capture household filing structure (e.g., joint filers) and implied dual-income/tax-liability dynamics.

- class_business_combo: Combination of class of worker and business ownership to separate incorporated self-employed profiles from wage-earners (a segment strongly associated with >50K outcomes).
- veteran_affiliation: A combined indicator created as: veteran_affiliation = veterans benefits + "__" + fill inc questionnaire for veteran's admin to merge two correlated veteran-related fields into a single stable categorical descriptor.

Cardinality checks and consistency validation were applied before adoption to confirm these combinations reduced redundancy without creating sparsity. The engineered interaction features were then re-evaluated using Random Forest and XGBoost (feature-importance agreement + ROC-AUC validation). On the pruned baseline feature set, performance was RF ROC-AUC = 0.949818 and XGBoost ROC-AUC = 0.954502. After introducing marital_tax_combo and class_business_combo, the score dropped slightly, to RF ROC-AUC = 0.949603 and XGBoost ROC-AUC = 0.953272, indicating the interactions added interpretability and reduced redundancy, but did not provide additional lift beyond the existing granular predictors under the current encoding.

Veteran-related fields were then compressed into veteran_affiliation to merge two correlated columns while controlling cardinality (4 resulting categories). In the compression test, replacing the original veteran columns with the combo feature produced near-identical ROC-AUC, showing minimal information loss: ROC-AUC = 0.951081 (augmented: originals + combo) versus ROC-AUC = 0.951058 (replaced: combo only). Finally, a compact modeling set was created by retaining the top ~30 ranked features (dropping ~10 low-importance fields such as year and enroll in edu inst last wk). Under this reduced set, performance shifted to RF ROC-AUC = 0.946936 and XGBoost ROC-AUC = 0.951081, reflecting the expected trade-off between dimensions and performance.

1.6.3. Feature encoding strategy and sparsity control

High-cardinality categoricals were explicitly treated to avoid one-hot sparsity and inflated compute. A threshold of >45 unique levels was used to route columns into advanced encoders, while the remaining categoricals stayed in One-Hot Encoding (OHE). Encoding ablation showed Target Encoding + OHE as the most stable choice across models:

- XGBoost (encoding ablation)
    - Target (High-card) + OHE: ROC-AUC 0.952000, LogLoss 0.119052
    - All OHE baseline: ROC-AUC 0.951141, LogLoss 0.119341
- Random Forest (encoding ablation)
    - Target (High-card) + OHE: ROC-AUC 0.946976, LogLoss 0.126491
    - All OHE baseline: ROC-AUC 0.946938, LogLoss 0.130045

Interpretation: Target encoding preserved predictive signal for high-cardinality fields while reducing OHE dimensionality and maintaining (or improving) ROC-AUC and calibration (LogLoss), making it the preferred production encoding.

1.7 Model Building and Evaluation

1.7.1. Compressed vs Hybrid vs Baseline feature sets

Three feature configurations were evaluated to balance information retention vs redundancy vs computability:

- Baseline (Original Raw Features): 40 predictors (8 numeric + 32 categorical) used as the reference configuration.
- Compressed (Top Features): 30-feature configuration produced after importance-driven pruning (used for efficiency and to reduce redundancy).
- Hybrid (Information-retaining wealth representation – 32 features): differed from the compressed representation by retaining the more informative investment components (capital gain, losses and dividends from stocks) instead of relying only on a single aggregated investment proxy; this improved separability because the capital components consistently surfaced as high-importance drivers.

1.7.2. Validation of feature usefulness (importance agreement + pruning logic)

Feature choices were validated using (1) Random Forest feature importance and (2) XGBoost SHAP to confirm agreement on top drivers before pruning. This ensured removal decisions were based on consistent signal, not a single-model artifact. After ranking, the compressed (top-30) set was formed and downstream evaluation confirmed competitive performance.

1.7.3. Model families and selection procedure

Three classifiers were evaluated under consistent preprocessing and weighting: Logistic Regression, Random Forest, and XGBoost. Model selection used GridSearchCV with ROC-AUC as the primary metric, and 5-fold cross-validation was reported for RF/XGBoost. PR-AUC was used as a supporting metric to reflect ranking quality under class imbalance.

1.7.4. Compressed (30-feature) benchmarking

GridSearchCV:

- XGBoost (Unbalanced): ROC-AUC 0.951504, PR-AUC 0.668048
- Random Forest (Unbalanced): ROC-AUC 0.947338, PR-AUC 0.652638

Inference: The compressed feature set maintained strong discrimination while reducing dimensionality. Balanced training shifted the operating point toward recall while keeping ROC-AUC nearly unchanged.

1.7.5. Baseline vs Hybrid performance and stability with cross-validation

Random Forest:

- Baseline CV ROC-AUC 0.9473 ± 0.0016 | Test Acc 0.9535, ROC-AUC 0.9502
- Hybrid CV ROC-AUC 0.9469 ± 0.0018 | Test Acc 0.9539, ROC-AUC 0.9495

XGBoost:

- Baseline CV ROC-AUC 0.9516 ± 0.0014 | Test Acc 0.9555, ROC-AUC 0.9547
- Hybrid CV ROC-AUC 0.9512 ± 0.0014 | Test Acc 0.9562, ROC-AUC 0.9547

Inference: Baseline and Hybrid achieved nearly identical ROC-AUC, while tight CV spreads indicate stable generalization rather than variance-driven gains. The final accepted model is the Hybrid model with the pruned features set (32 features) since it provides the perfect balance of maximum predictive power of the original model (40 features) and reduced dimensions closer to the compressed model (30 features).

1.7.6. Addressing Class Imbalance: Balanced vs. Unbalanced Models

To align the model output with varying strategic business objectives, two distinct XGBoost variants were trained:

- Unbalanced XGBoost: Trained without synthetic minority adjustments. This model optimized strictly for Precision, aggressively minimizing False Positives.
- Balanced XGBoost: Utilized the scale_pos_weight hyperparameter to mathematically penalize minority class misclassifications. This model dramatically increased Recall, capturing a much wider net of the high-income population.

1.7.7. Precision-Recall, Cumulative Gain Plot and ROC-AUC Evaluation

Because standard accuracy is a misleading metric on a 93% majority-class dataset, performance was evaluated via Precision-Recall (PR) dynamics:

- Precision–Recall curves were plotted to evaluate minority-class performance under strong class imbalance and to visualize the operating trade-off between capturing more >$50k cases (Recall) and keeping outreach efficiency high (Precision). The balanced training configuration pushed recall upward as intended, with a predictable precision drop. In parallel, cumulative gains / lift charts were generated to assess *how well the score ranks true >$50k individuals near the top*. The curves showed gains concentrated in the early deciles, indicating that the XGBoost Hybrid model is effective for top-percentile prioritization (e.g., targeting the highest-scoring segment first) and supports score-based targeting rather than a single fixed threshold.
- Extracting feature importances from the final XGBoost variants revealed a structural shift between setups. Under balanced training, education and detailed occupation recode increased in prominence, indicating that when optimized to recover a broader share of the affluent class, the model leans more heavily on career-structure signals and less on rare, extreme financial tail events.

1.7.8. Strategic Deployment Recommendations

The dual-model output provides actionable deployment flexibility:

- High-Precision Applications: For luxury retail marketing or high-cost direct mail campaigns where minimizing wasted ad spend (False Positives) is critical, the Unbalanced Model is the optimal choice.

- High-Recall Applications: For broad market-reach strategies, general economic research, or inclusive financial services where the goal is to identify all potential high-capacity consumers, the Balanced Model should be deployed to maximize the capture rate of the Total Addressable Market.
- Cross-Model Integration: A critical strategic decision involves integrating the segmentation model (section 2) as a high-weight feature in the Classification Model. Feeding the Cluster ID and profiles into the income prediction pipeline provides the socio-economic context required to distinguish between wage-driven and capital-driven earners, potentially reducing "False Negatives" in high-income targeting.

Section 2: Segmentation Modelling

2.1 Objective and Approach

The objective of this segmentation analysis was to identify meaningful population segments within the Census Bureau dataset by utilizing an unsupervised Machine Learning approach. By moving beyond arbitrary demographic partitions and grouping individuals based on multidimensional similarities in wealth, employment, and household structure, this analysis uncovers distinct profiles that reflect the real population structure. These results translate into actionable marketing personas, allowing for the optimization of product offerings and the precise allocation of marketing budgets based on the specific purchasing power and lifestyle of each identified group.

2.2 Dataset and Preprocessing

The analysis was performed on the Census Bureau dataset, in which each record represents an individual observation coupled with a population weight to ensure findings reflect the actual US population scale. To maintain consistency across the project's dual modelling tracks classification and segmentation, a unified data preparation pipeline was established.

2.2.1. Data Cleansing and Standardization: The data underwent rigorous cleaning, including the removal of duplicate records and the standardization of categorical variables through whitespace trimming and proper type casting. While the target income label was mapped to a binary variable (representing income above or below $50,000), it was excluded from the clustering phase to preserve the unsupervised nature of the segmentation.

2.2.2. Missing Value Strategy: Missing values were handled using a structured, deterministic approach rather than arbitrary imputation. Most missing categorical attributes were assigned an "NA" category to maintain data integrity. For the "Hispanic origin" variable, missing values were resolved by mapping birth countries to their most consistent historical demographic category, with unresolved cases categorized as "Unknown."

2.3 Feature Engineering

To optimize the segmentation model, the primary feature set was refined through a process of statistical evaluation and engineering. While the comprehensive technical analysis of these features is detailed in the classification section of this report, the selection for clustering was specifically driven by the need to capture socio-economic diversity without introducing mathematical noise.

2.3.1. Feature Refinement and Consolidation:

The original census variables were evaluated alongside several engineered features listed earlier (section 1.6) to better represent economic capacity and social status.

2.3.2. Analysis for Clustering Suitability:

Before the selection, the candidate features underwent a strict statistical audit:

- **Low Variance Analysis:** Features where over 90% of the population shared a single value were excluded, as they provide no discriminative power for grouping.
- **Multicollinearity Assessment:** Using Pearson correlation, highly correlated variables (correlation >0.75) were removed to prevent the double-counting of traits matrix, which can artificially inflate the importance of certain demographics during the distance calculation process.
- Variables with an excessive number of unique categories and high correlation scores were pruned to prevent sparse encoding (from One-Hot Encoding), which can lead to high-dimensional "sparsity" and an unintended bias.

This targeted selection process resulted in a high-fidelity feature space that effectively balances economic capacity, employment structure, and demographic identity.

Variables with near-zero variance, such as enroll in edu inst last wk, region of previous residence, and state of previous residence, were removed because their lack of diversity. Furthermore, high-cardinality features including country of birth self, country of birth father, and country of birth mother were excluded due to their high correlation with citizenship. Finally, raw financial variables such as capital gains, capital losses, and dividends from stocks were consolidated into a single engineered net_investment_income metric to resolve collinearity while preserving a holistic view of each individual's economic capacity.

2.3.3. Final Optimized Feature Set:

- **Numeric (4):** Age, number of persons worked for employer, net investment income, and weeks worked in year.
- **Categorical (13):** Major occupation, education, marital/tax status, major industry, race, hispanic origin, sex, employment status, household summary, family members under 18, citizenship, class-business combination, and veteran affiliation.

2.4 Model Selection and Training: K-Means Clustering

The K-Means Clustering algorithm was selected for its efficiency in partitioning large populations into distinct groups based on multidimensional feature similarity and scalability compared to hierarchical grouping. With over 150,000 observations, K-Means provides the computational efficiency required for a production environment, ensuring that the persona reports and profiles can be generated in minutes.

2.4.1. Transformation and Scaling:

- Numeric features were standardized using a StandardScaler (mean of 0, variance of 1) to ensure all features contributed equally to distance calculations, preventing high-magnitude variables (like investment income) from dominating smaller ones (like age). Signed Log Transformation (log1p) was applied to compress outliers without losing the directional impact of financial losses.
- Categorical features were transformed using One-Hot Encoding to convert them into a numerical format while preserving categorical relationships.

2.4.2. Incorporation of Population Weights: Crucially, the model incorporated the population weight column directly during training. This ensures that the clusters reflect the actual US population distribution rather than just the sample distribution, allowing each observation to contribute proportionally to its real-world representation.

2.5 Cluster Selection and Evaluation

The optimal number of clusters was determined by evaluating configurations from K=2 through K=12 using two mathematical metrics: Inertia (The Elbow Method): Measures within-cluster compactness. And Silhouette Score: Measures the separation between clusters.

Plot Interpretations:

- **The Left Plot (Inertia):** Displays a steep drop in inertia up to K=4, indicating that four clusters efficiently capture most of the population variance.
- **The Right Plot (Silhouette):** The score demonstrates a stabilization point at K=4 (approximately 0.25) before degrading at higher K values.
- **Conclusion:** K=4 was selected as the optimal hyperparameter, providing the best balance between statistical distinctiveness and business interpretability.

2.6 Cluster Profiles and Marketing Insights

2.6.1. Cluster profile 0: The Mass-Market Workforce (High Volume)

This segment represents 64.7 million people (23.6%) with an older demographic profile, characterized by a median age of 61. Workforce participation is minimal, with a median of zero weeks worked per year, and 91.5% are classified as “Not in universe” for career class, indicating limited active employment. Income probability is very low, with only 1.1% earning over $50k, and investment income is typically absent. Education levels are moderate, with high school graduates (34.8%) being the most common. Household structure is stable, dominated by householders (48.9%) and spouses (31.9%). Most individuals are adult household members, and under-18 family dependency is negligible in this segment.

Marketing Insight: This group has limited income growth and relies on fixed or predictable financial resources. Marketing should emphasize affordability, essential goods, and value-focused offerings such as discounts, store brands, and bundled household necessities.

2.6.2. Cluster profile 1: The Stable Professional Mid-Market (42.6% of population)

This is the largest segment, comprising 116.7 million people (42.6%) and representing the core working population. Workforce participation is strong, with a median of 52 weeks worked per year, and individuals are in their prime working age with a median age of 36. Income probability is moderate at 7.7% over $50k, and investment income is typically absent, indicating a primary reliance on employment income rather than capital wealth. Occupation distribution is broad, led by administrative support (15.1%), professional specialty (11.9%), sales (11.4%), and executive/managerial roles (10.7%). Education levels are moderate to high, including high school graduates (34.9%), some college (21.9%), and bachelor's degrees (14.4%). Stable family structures dominate, featuring householders (48.2%), spouses (26.5%), and a strong representation of married joint filers (52.1%).

Marketing Insight: This group represents the most stable and predictable consumer base. Marketing strategies should focus on convenience, subscription services, and household-oriented bundles to encourage repeat purchasing and long-term loyalty.

2.6.3. Cluster profile 2: The Affluent Capital-Class (8.8% of population)

This is the smallest segment (24.0 million people, 8.8%) but the most economically powerful. It has the highest income probability (33.4% >$50k) and significant investment income, with a median of $1,317 and a mean of $6,342, indicating both wage and capital-driven wealth. Workforce participation is strong with a median of 52 weeks worked, and individuals are in their peak earning years with a median age of 44. Occupation distribution is concentrated in high-skill leadership roles, including professional specialty (22.7%) and executive/managerial positions (20.9%). Education levels are significantly higher than other clusters, with 28.7% holding a bachelor's degree and 12.2% holding a master's degree. Households are stable and financially independent, characterized by householders (67.1%) and high married joint filing rates (65.1%).

Marketing Insight: This segment has the highest purchasing power and lowest price sensitivity. Marketing should emphasize premium products, quality, exclusivity, and convenience. This segment represents the highest revenue-per-customer opportunity.

2.6.4. Cluster profile 3: The Dependent Youth Segment (25.0% of population)

This segment represents 68.5 million people (25.0%) and is structurally distinct, consisting primarily of children and dependents. The median age is 8 years, and workforce participation is effectively zero. The education profile confirms this structure, with 86.9% classified as children. Occupation and industry fields are almost entirely "Not in universe," indicating no labor force participation. Income probability is 0.0%, and investment income is absent. Family composition is clearly defined within active households: 67.2% have both parents present, 24.6% have a mother only, and 3.5% have a father only. This confirms that the segment represents dependent members of larger family units.

Marketing Insight: This segment does not represent direct income earners but heavily influences household consumption. Marketing should target parents and households, focusing on family-oriented products, education-related goods, and essential household services.

2.7. Business Judgment Based on Segmentation Results

Premium Value Concentration Segment (Persona 2): Representing approximately 8.8% of the total weighted population, Persona 2 emerges as the highest-value segment due to its disproportionate share of Net Investment Income (Mean: $6,342) and the highest probability of earning above $50k (33.4%). This concentration of financial capacity and economic stability positions Persona 2 as a primary margin-generating segment with strong monetization potential. The elevated income profile and representation in skilled and leadership-aligned occupations indicate greater purchasing power and lower sensitivity to price fluctuations relative to other segments. Strategic prioritization of this segment through premium-tier offerings, exclusive service tiers, and high-value customer retention programs can maximize Customer Lifetime Value (CLV) while maintaining strong margin efficiency due to their higher spending potential.

Core Revenue Stabilization Segment (Persona 1): Accounting for nearly 43% of the Total Addressable Market (TAM), Persona 1 represents the core revenue-driving segment with the highest workforce participation and consistent employment intensity, as reflected in near-full employment coverage and stable weeks-worked distribution. This segment's large population share, and reliable income flow establish it as the primary contributor to sustained transaction volume and recurring economic activity. With moderate income probability and strong workforce integration, this segment offers high scalability potential for broad-reach engagement strategies focused on retention, accessibility, and operational efficiency. Investment in infrastructure and engagement models that prioritize consistency, availability, and long-term relationship stability will ensure sustained revenue throughput from this economically active segment.

Emerging Future-Value Segment (Persona 3): Persona 3 represents a future-value acquisition segment characterized by minimal direct income contribution but significant presence within structured household environments, with approximately 67.2% concentrated in dual-parent household structures. While current income and employment indicators remain low, this demographic positioning indicates strong long-term conversion potential into economically active segments. Early engagement with this group enables long-term lifecycle value development by establishing early brand presence and facilitating future transition into higher-income, workforce-integrated personas. This segment represents a strategic pipeline for future revenue expansion and sustained market growth.

Cost-Sensitive Sustainability Segment (Persona 0): Representing approximately 23.6% of the weighted population, Persona 0 is defined by an older demographic profile (Median Age: 61), lower workforce participation, and reduced income probability, indicating constrained purchasing capacity relative to other segments. This segment's economic profile suggests higher price sensitivity and lower discretionary spending potential, requiring a value-optimized engagement approach. Strategic emphasis on cost-efficient offerings, essential goods, and affordability-focused positioning ensures continued participation and market coverage without eroding profitability. By aligning resource allocation and product positioning with the economic characteristics of this segment, organizations can maintain volume coverage while preserving margin performance in higher-value segments.

References:

1. Gelman, A. (2004). Exploratory data analysis for complex models. Journal of Computational and Graphical Statistics, 13(4), 755–779.

2. Nicodemo, C. (2022). Exploratory data analysis on large data sets: The example of the Italian labour market (GLO Discussion Paper No. 1038). Global Labor Organization. https://www.econstor.eu/bitstream/10419/261928/1/1809720222.pdf

3. IPUMS CPS. (n.d.). Weights. IPUMS Current Population Survey, University of Minnesota. https://cps.ipums.org/cps/

4. Dataquest. (2025, May 14). Project tutorial: Customer segmentation using K-means clustering. Dataquest Blog. https://www.dataquest.io/blog/customer-segmentation-using-k-means-clustering/

5. Yin, H., Aryani, A., Petrie, S., Nambissan, A., Astudillo, A., & Cao, S. (2024). A rapid review of clustering algorithms (arXiv:2401.07389). arXiv. https://arxiv.org/abs/2401.07389

6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

8. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

9. Agarwal, A. (2024). Exploratory Data Analysis (EDA) for Banking and Finance. arXiv. https://arxiv.org/abs/2407.11976

10. Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

11. Kaggle. (n.d.). House prices: EDA, visualization & prediction [Kaggle Notebook]. https://www.kaggle.com/code/chapagain/house-prices-eda-visualization-prediction

12. Kaggle. (n.d.). House price EDA and modeling with Python [Kaggle Notebook]. https://www.kaggle.com/code/reidjohnson/house-price-eda-and-modeling-with-python

13. U.S. Census Bureau. (2026, January 8). Current Population Survey (CPS) documentation. U.S. Census Bureau. https://www.census.gov/programs-surveys/cps/technical-documentation.html