# Final Report: Income Classification and Customer Segmentation for Retail Marketing

**Client: Walmart (Retail Marketing) - JPMC Take-Home**
**Prepared by: Subramaniya Siva T S**
**Date: September 17, 2025**

## 1) Executive Summary

This report delivers two analytical assets built from the 1994–1995 CPS-derived dataset (199,523 records with 40 demographic and employment attributes, a survey weight, and a binary income label):

- **Income Classification:** A supervised model that predicts whether an individual's income exceeds $50,000. The pipeline applies robust cleaning, targeted feature engineering (for example, education ordinal, investment income flags, work-attachment features), and column-wise preprocessing with imputation, log transforms for skewed monetary variables, scaling, and one-hot encoding. Several model families are evaluated under weighted metrics to respect survey design; gradient boosting variants consistently perform best on PR-AUC. The solution also supports precision-targeted thresholding (for example, $\geq 0.70$ precision) for budget-efficient marketing.
- **Marketing Segmentation:** An unsupervised clustering (KMeans) using the cleaned feature set to produce six interpretable segments, following the notebook's configuration. Each segment is profiled with weighted share, weighted $>$ $50,000 rate, top categorical characteristics, and numeric means to support creative, channel, and offer strategies.

**Key outcomes:**

- The classifier enables audience prioritization by score thresholds aligned to desired precision (fewer wasted impressions at initial rollout; recall can be expanded later).
- The six-segment solution surfaces distinct groups (for example, non-workers or teens, prime-age full-time workers, older cohorts, professional or educated clusters), each with clear messaging and assortment implications.
- Both assets are fully reproducible: command-line interfaces, per-run logging, saved figures and artifacts, fixed random seeds, and survey-weighted evaluation.

## 2) Data & Business Context

### Data Source

The dataset used for this project comes from the 1994–1995 Current Population Survey (CPS), conducted by the U.S. Census Bureau. It is a weighted sample that reflects the broader U.S.

population through stratified sampling. Each record represents one or more individuals in the general population, with the survey weight indicating how many people that record corresponds to.

### Data Structure

- **Size:** ∼199,500 observations.
- **Features:** 40 demographic and employment-related variables, including age, education, industry, occupation, marital status, work history, and income-related variables such as wages, capital gains or losses, and dividends.
- **Label:** Binary indicator showing whether income is $\leq$ \$50,000 or $>$ \$50,000.
- **Weight:** A continuous value used to scale results so they represent the population distribution, not just the sample.

### Important Considerations

- **Imbalance:** A relatively small share of the population earns more than \$50,000. This imbalance matters both statistically (skewed labels) and commercially (a small but high-value target group).
- **Survey Weighting:** Any analysis or model ignoring the weights would misrepresent the population. For example, a subgroup with few sampled cases but large weights could represent millions of people.

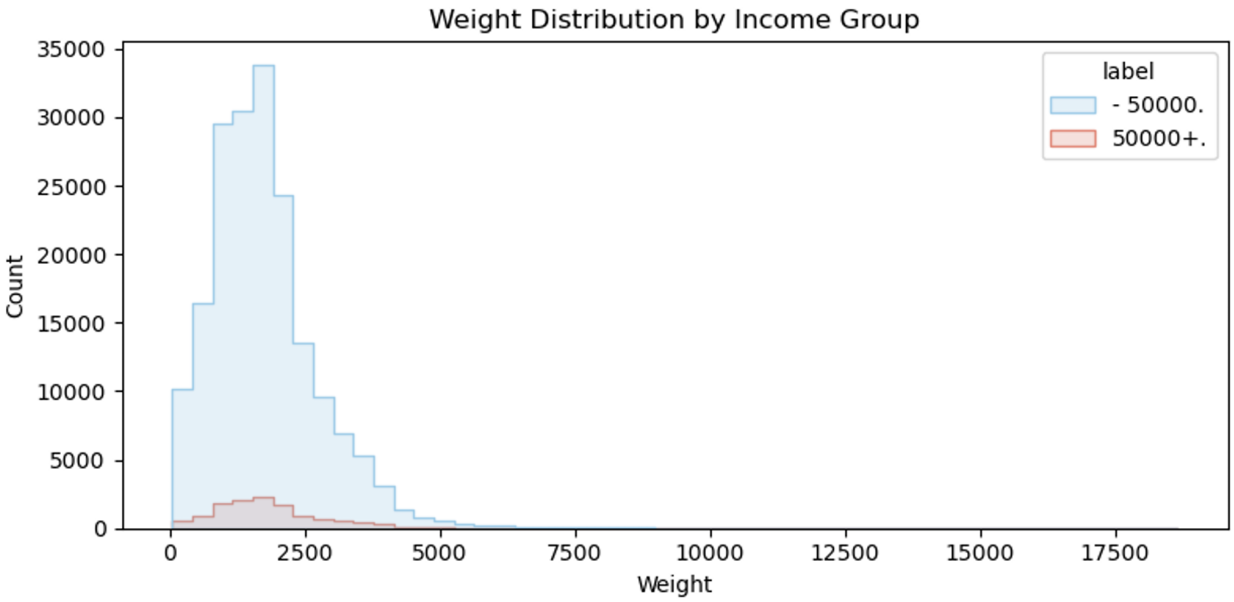



Figure 1: Weighted income label distribution.

## 3) Exploratory Data Analysis (EDA)

### 3.1 Data quality and coverage

**Sample and scope.** The dataset contains ∼199,000 survey records from CPS 1994–1995, each with an associated survey weight. All descriptive statistics in this section are reported as weighted population estimates.

**Label imbalance.** Only ∼6.4% of the weighted population earns $>$ \$50,000; ∼93.6% earns $\leq$ \$50,000. This imbalance influences model choice and evaluation (PR-AUC preferred; precision-targeted thresholds recommended).

**Missingness.** Columns are largely complete. The only notable missingness is in **hispanic origin** (∼0.44%), which is imputed (most-frequent) and retained as categorical.

**Implications for modeling.** Imbalance favors precision-first launch thresholds. Low missingness allows simple, robust imputation.

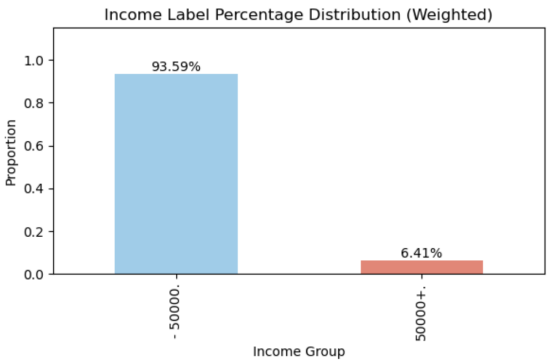



Figure 2: Weighted label distribution.

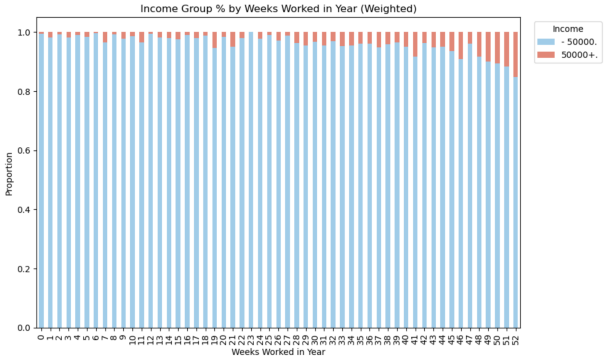
### 3.2 Labor-force attachment as the strongest macro signal

**Weeks worked.** The population is bimodal: many report 0 or 52 weeks. The $>$ \$50,000 share concentrates among full-year workers and is near zero among non-workers.

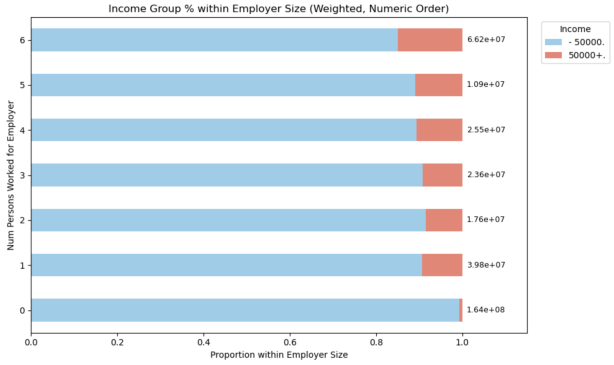
**Full or part-time status.** Full-time workers display much higher $>$ \$50,000 rates than part-time workers or those not in the labor force.

**Employer size.** Larger employer-size codes correspond to higher $>$ \$50,000 shares, consistent with higher-paying roles in bigger organizations.

**Implications.** Labor-force attachment should be explicitly encoded (for example, `worked_any_weeks`, `full_year_52w`). These variables also provide clear segment narratives.




(a) Income proportion by weeks worked.




(b) Income proportion by employer size.

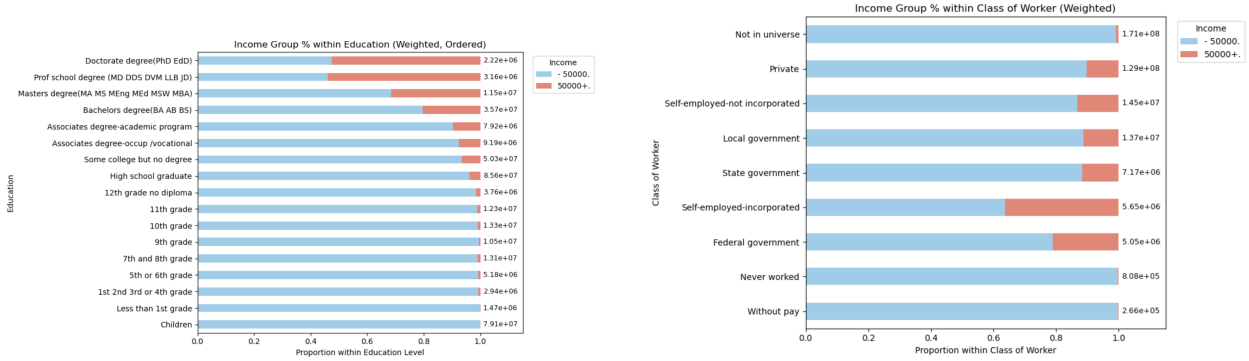
Figure 3: Income proportion by weeks worked and employer size.

### 3.3 Human capital and sector (education, age, class of worker)

**Education.** Monotonic lift: higher attainment → higher > $50,000 share; bachelor's and above are disproportionately > $50,000.

**Age.** > $50,000 peaks at ages 30–50, falls after ∼60; ages ≤25 are overwhelmingly ≤ $50,000. For linear models: use an ordinal education feature and coarse age bins.

**Class of worker.** Private and NIU dominate counts; within workers, *Self-employed (incorporated)* and *Federal government* show higher > $50,000 rates, while *Never worked* and *Without pay* are almost entirely ≤ $50,000. Useful for segment positioning (value vs. premium; channel).




(a) Income proportion by education. (b) Income proportion by class of worker.

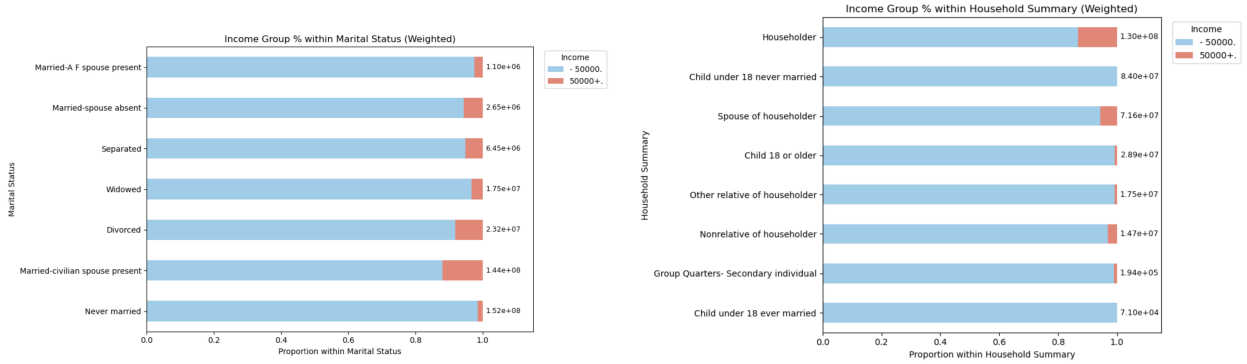
Figure 4: Human capital and sector patterns relevant to income and marketing segmentation.

### 3.4 Household structure and marital status

**Marital status.** The largest groups are *Never married* and *Married (civilian spouse present)*. The > $50,000 share is highest in the latter; *Divorced* and *Widowed* skew to ≤ $50,000.

**Household summary.** Householders and spouses contain most higher-income cases; children and nonrelatives are overwhelmingly ≤ $50,000.

**Implications.** Household role provides interpretable hooks for creative and offer strategy.




(a) Income proportion by marital status. (b) Income proportion by household summary.

Figure 5: Income proportion by marital status and household summary.

### 3.5 What matters most (for models and marketing)

**Strongest signals:** Labor-force attachment, education, age, and class of worker consistently explain income variation.

## 4) Data Cleaning and Feature Engineering

### 4.1 Goals and principles

The cleaning pipeline preserves population representativeness and maximizes signal for tabular models while keeping transformations simple and reproducible.

### 4.2 Standardization and type fixes

- **CPS code normalization.** Binary or ternary CPS flags (for example, *own business or self employed*, *veterans benefits*) are mapped to {`Yes`, `No`, `NIU`} and treated as categorical.
- **"NIU" recoding.** Code columns with `0` meaning "not in universe" (for example, detailed industry or occupation recodes) are converted to strings, ensuring categorical interpretation.
- **Year.** The *year* column is treated as a categorical indicator (`1994` or `1995`).

### 4.3 Label and weights

- **Target.** The label is mapped to $y \in \{0, 1\}$ with $1 = (> \$50{,}000)$.
- **Survey weight.** The provided sampling weight $w$ is carried through EDA and model evaluation; training uses a reweighted scheme to balance classes, while validation uses the original $w$ to reflect population prevalence.

### 4.4 Feature engineering (compact and high-signal)

The following derived features are added to capture patterns observed in EDA without overfitting:

- **Education (ordinal):** `education_ord` encodes ordered attainment (Children $\rightarrow$ Doctorate). Motivated by the monotonic lift in $> \$50{,}000$.
- **Work attachment:** `worked_any_weeks`$= 1[\text{weeks} > 0]$, `full_year_52w`$= 1[\text{weeks} = 52]$.
- **Investment income:** `invest_income`$=$ dividends+capital gains$-$capital losses; flag `has_invest_income`$=$ $1[\text{invest\_income} > 0]$.
- **Wage flag:** `has_wage`$= 1[\text{wage per hour} > 0]$.
- **Student status:** `is_student`$= 1[\text{enroll in edu inst last wk} \in \{\text{High school}, \text{College}\}]$.

### 4.5 Preprocessing graph (used by both classifier and segmentation)

- **Skewed numeric** (*wage*, *capital gains or losses*, *dividends*): median imputation → non-negativity clip → $\log(1+x)$ → robust scaling.
- **Other numeric** (for example, *age* plus residual numerics): median imputation; robust or standard scaling (no scaling for inherently discrete counts such as *weeks worked* or `education_ord`).
- **Categorical:** most-frequent imputation → one-hot encoding with `drop=if_binary`; unknowns are handled as `ignore` at inference.

## 5) Model Architectures and Training

### 5.1 Classification model

**Architecture.** The pipeline is: *clean* → *feature engineering* (education ordinal, work-attachment, wage or investment flags, student) → *ColumnTransformer* → estimator. Multiple estimators are supported (only those installed are run):

- **SGD Logistic** (`loss=log_loss`, modest $L_2$; strong baseline).
- **Random Forest** (300 trees, max depth 12, leaf size 5; non-linear, interaction capture).
- **XGBoost** (500 trees, depth 6, learning rate 0.05, subsample 0.8, hist tree method).
- **LightGBM** (700 trees, 31 leaves, learning rate 0.05, subsample and colsample 0.8).
- **CatBoost** (800 iterations, depth 6, learning rate 0.05; ordered boosting).

Tree models that require dense input (for example, Random Forest, CatBoost) receive a dense matrix; others consume sparse output from one-hot encoding directly.

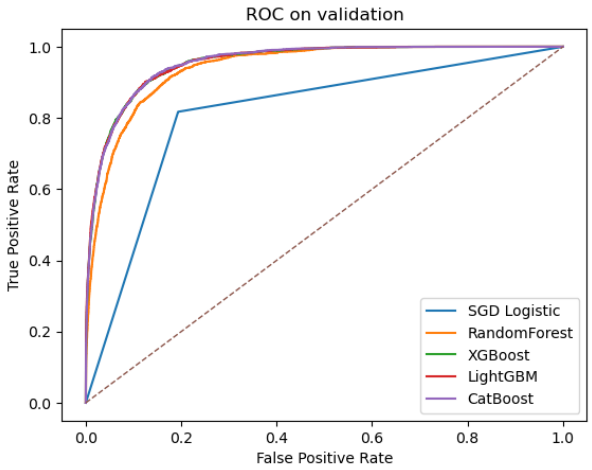
**Training split and weighting.** Data are split once into train and validation with a stratified 80/20 holdout (`random_state=42`). The CPS survey weight $w_i$ is carried throughout. To counter class imbalance during training, positives receive a multiplicative factor

$$\text{pos\_factor} = \frac{\sum_{i:y_i=0} w_i}{\sum_{i:y_i=1} w_i}, \quad \tilde{w}_i = \begin{cases} w_i \cdot \text{pos\_factor}, & y_i = 1 \\ w_i, & y_i = 0 \end{cases}$$

and the estimator is fit with $\tilde{w}_i$. Validation always uses the original $w_i$ to reflect population prevalence.

**Objective and metrics.** Each estimator optimizes its native objective (for example, logistic loss or gradient-boosting loss) under sample weights.

**Evaluation setup.** Validation uses the original CPS survey weights to reflect population prevalence. Metrics emphasized: (i) weighted ROC-AUC, and (ii) operating thresholds chosen from the PR curve. For activation, a precision-targeted threshold (for example, $\geq 0.70$) trades recall for budget efficiency.




(a) Weighted ROC curves.

| Model | ROC-AUC | AP |
|---|---|---|
| LightGBM | 0.955 | 0.697 |
| XGBoost | 0.955 | 0.690 |
| CatBoost | 0.956 | 0.688 |
| Random Forest | 0.941 | 0.595 |
| SGD Logistic | 0.812 | 0.196 |

(b) Validation performance (weighted).

Figure 6: Classifier performance on the validation split under survey weights. The ROC curve (left) shows separation, while the table (right) summarizes ROC-AUC and average precision.

**Findings.** Gradient boosting delivers the strongest ROC-AUC and a more favorable precision-recall frontier at actionable regions. Linear (SGD or logistic) provides a fast, interpretable baseline; Random Forest performs competitively but lacks calibrated probabilities without post-processing.

**Recommended operating point.** Adopt the boosted model with a precision-targeted threshold (for example, $\geq 0.70$) for initial rollout. As ROI stabilizes, relax the threshold to increase reach while monitoring PR-AUC and cost per acquisition.

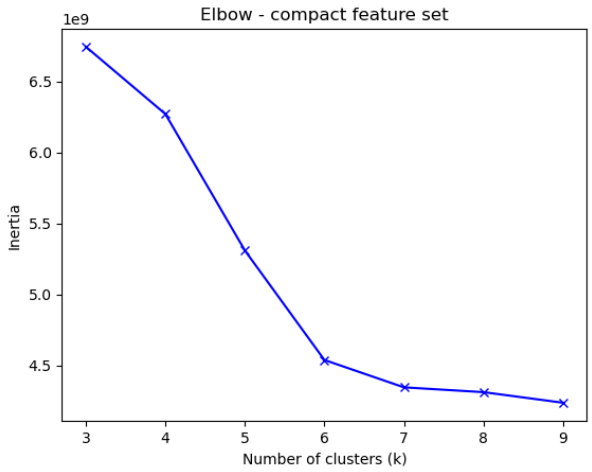
## 6) Segmentation: Results and Marketing Use

### 6.1 Model and $k$ selection

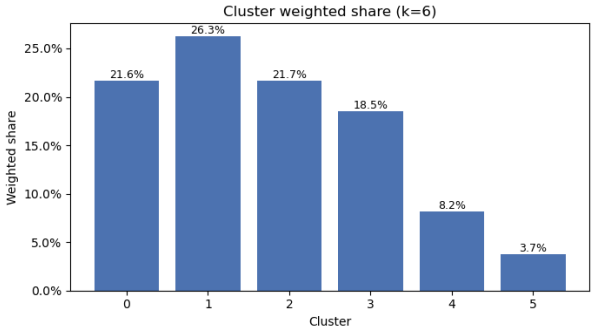
KMeans was fit on the cleaned feature space (numeric with $\log(1 + x)$ where needed; categoricals via one-hot encoding), using CPS survey weights as `sample_weight`. The elbow (Fig. 7a) bends at $k \approx 6$; $k = 6$ is used for a good fit and interpretability trade-off.

### 6.2 Population shares and high-level contrasts

Six segments are well-sized for activation (Fig. 7b); differences are driven by age, marital status, and labor-force attachment. Sex and race are broadly balanced across clusters.

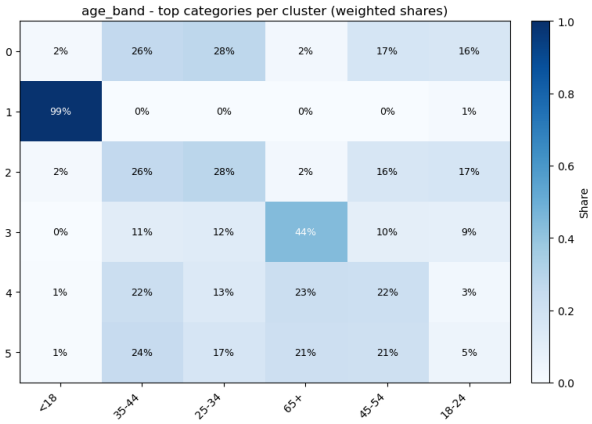



(a) Elbow diagnostic for $k \in [3, 9]$; optimal $k = 6$.

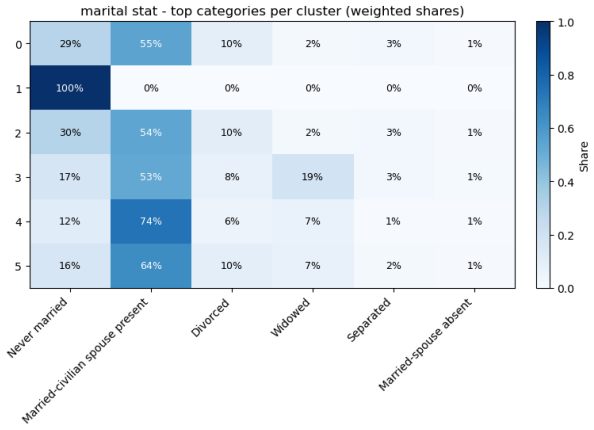



(b) Cluster shares: C0 (21.6%), C1 (26.3%), C2 (21.7%), C3 (18.5%), C4 (8.2%), C5 (3.7%).

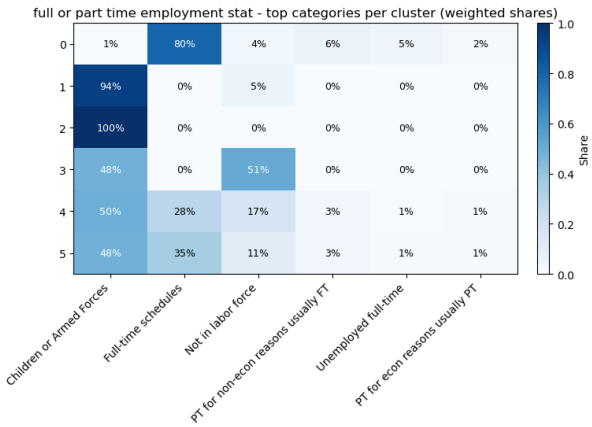
Figure 7: Segmentation diagnostics: elbow method selection and resulting cluster proportions.



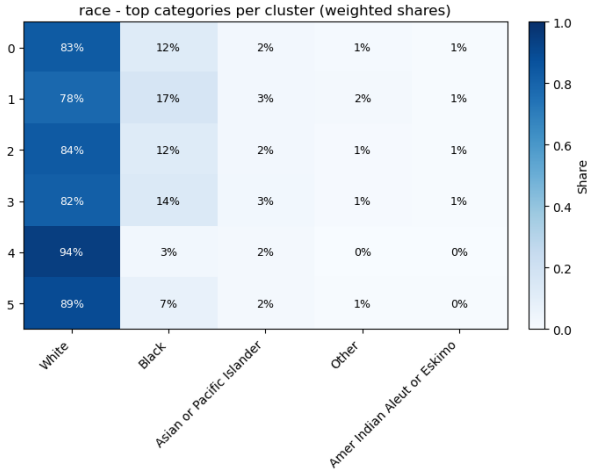

(a) Age bands (weighted).




(b) Marital status (weighted).




(c) Full or part-time and labor-force.




(d) Race.

Figure 8: Key drivers of separation: age, marital status, labor-force attachment, and race.

### 6.3 Personas and actions

**C0: Full-time schedules.** Dominated by full-time status. *Use:* convenience and time-savers (curbside, quick meals), premium essentials; app or email.
**C1: Children or teens (largest).** Nearly all under "Children" (Fig. 8). *Use:* message to guardians; back-to-school, family bundles, value.
**C2: Armed forces heavy.** Concentrated in "Children or Armed Forces". *Use:* service-friendly offers (pickup or shipping), durable goods, financial services.
**C3: Older adults.** Peak at 65+; lower labor-force attachment. *Use:* pharmacy or health, home, value guarantees; CRM or direct mail.
**C4: Married, civilian spouse present (high).** Strong "married, civilian spouse present" signal and **higher weighted positive rate**. *Use:* **priority target**. Premium grocery, household multi-packs, electronics; app, loyalty, search.
**C5: Small mixed cohort.** Balanced on sex and race; mixed marital and work patterns. *Use:* test creative variants.

### 6.4 Fairness note

Figures indicate race and sex are similar across clusters, suggesting the segmentation is not trivially splitting on sensitive attributes. Production activation should still include fairness checks and policy review.

## 7) How to Run the Code

### Prerequisites

- Python 3.9+.
- From project root: `pip install -r requirements.txt`.
- Data files in `data/`: `census-bureau.data` and `census-bureau.columns`.

### Basic usage (run from `src/`)

```
cd src
# Classification (train + eval, saves models/plots/metrics)
python classifier.py --save_models

# Segmentation (k=6 by default, elbow computed)
python segment.py --save_models
```

### Common options

```
# Choose specific classifier models
```

```
python classifier.py --models sgd,rf

# Precision-targeted thresholding for XGBoost
python classifier.py --models xgb --target_precision 0.80

# Force number of clusters
python segment.py --k 8
```

### Outputs (per run)

Created under `outputs/logs/run_YYYYMMDD_HHMMSS/`:

- `classifier.log` or `segment.log`, and `run.json`.
- `metrics.json` (classifier), `segments.csv` and `summary.json` (segmentation).
- `figs/` (ROC or PR, elbow, cluster share, heatmaps, numeric means).

Saved models (when `--save_models`): `models/preprocessor.pkl`, `models/model_<Name>.pkl`, `models/seg_preprocessor.pkl`, `models/seg_kmeans_k<k>.pkl`.

### Path overrides (optional)

```
python classifier.py --data_path ../data/census-bureau.data \
  --columns_path ../data/census-bureau.columns \
  --outputs_dir ../outputs --models_dir ../models \
  --run_name demo --save_models
```

## References


[1] U.S. Census Bureau (1994–1995). *Current Population Survey (CPS) Technical Documentation*. Available at `https://www2.census.gov/programs-surveys/cps/methodology/CPS-Tech-Paper-77.pdf`.

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., *et al.* (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[3] 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). doi: 10.1145/2939672.2939785