Executive Summary

This report presents a comprehensive end-to-end data science analysis of the Adult Income Census dataset, covering exploratory data analysis (EDA), feature engineering, supervised income classification, and unsupervised population segmentation. The dataset represents a large-scale demographic survey with mixed numerical and categorical variables and includes survey weights that reflect real-world population proportions. The analysis was conducted with a strong emphasis on data quality, interpretability, scalability, and population-representative modeling.

The study first establishes a clean and reliable analytical foundation through systematic data inspection, missing value handling, and feature transformations. Multiple classification models—Logistic Regression, Random Forest, and CatBoost—were developed and evaluated under severe class imbalance. CatBoost demonstrated the strongest overall performance due to its native handling of categorical features and robustness to imbalance.

For segmentation, both unweighted and weighted clustering approaches were explored. Weighted K-Means was selected as the primary segmentation model, enabling clusters to reflect true population prevalence rather than sample density. The number of clusters was selected using the Davies–Bouldin Index (DBI), with K=5 chosen as a balanced and interpretable solution. Cluster interpretation was further enhanced using Random Forest–based explainability and SHAP values. Overall, the results provide actionable insights into income drivers and population subgroups while maintaining methodological rigor and scalability.

Exploratory Data Analysis (EDA) Report – Adult Income Census Data

1. Overview of the Dataset

- Dataset Size: ~185,000 rows and 42 columns.
- Data Types: Mixed types including integers, floating-point numbers, and categorical values.
- Objective: Understand the structure, quality, and characteristics of the data in order to prepare it for further analysis and modeling.

2. Data Inspection

- Columns and Data Types: All 42 columns were examined to understand which columns are numerical and which are categorical.
- Categorical Variables Misrepresented as Numeric: Some columns, although stored as integers, actually represent categories, such as:
  - Major industrial code
  - Major occupational code
  - Veteran benefits
  - Business ownership / self-employment
  - Year (only 1994 and 1995)
- Action: These columns were recast as categorical to reflect their true nature.

3. Missing Values and Null Handling

- Initial Check: Only one column (Hispanic origin) had null values (~800 missing entries).
- Handling: Replaced missing values in Hispanic origin with "Do not know" to maintain information integrity.

4. Inconsistent and Improper Values

- Observation: Many columns contained the ? character representing unknown or missing information.
- Scope: ? was present in approximately 8 columns, affecting ~100,000 rows (~50% of data).
- Investigation: Each affected column was individually inspected to see if an existing value already represented unknown/missing information.
    - One column had "Not identifiable" as an existing representation of unknown values.
- Handling:
    - Replaced ? with "Not identifiable" for the column that already had this representation.
    - Replaced all other ? values in the remaining columns with "Unknown".

5. Unique Values Inspection

- Objective: Identify misspellings, inconsistencies, or improper values in categorical columns.
- Approach: Examined unique values of all 42 columns to ensure that categories were consistent and no erroneous labels were present.

6. Numerical Feature Analysis

6.1 Distribution Inspection

- Plotted distributions of all continuous numerical variables.
- Observation:
    - Except for age, all numerical columns were heavily skewed, either left-skewed or right-skewed.
    - Many numerical columns (e.g., capital gains, capital losses, integer employment-related columns) contained a high proportion of zero values.

6.2 Transformation of Skewed Variables

- Logarithmic transformation was applied to skewed numerical columns (except age) to reduce skewness and approximate symmetry.
- Age was not transformed to preserve its natural distribution and interpretability.

6.3 Handling Special Cases

- Investigated rows with age = 0, representing newborns/children.
- Checked consistency across other columns (education, class of worker, marital status, veteran benefits, employment status).
- Findings:
    - Most columns for age = 0 were consistent (children-related categories, “not in universe” for employment).
    - Exception: 7 rows had inconsistent marital status (e.g., “Married civilian spouse”) for children.

- Action: Dropped the 7 inconsistent rows to reduce noise.

7. Correlation Analysis

- Objective: Identify highly correlated numerical features to avoid multicollinearity in supervised learning models.
- Method: Correlation heatmap for all numerical columns.
- Observation:
    - Number of weeks worked and number of employees under the individual were highly correlated, suggesting potential redundancy.
- Action: Considered feature selection/merging for highly correlated columns in downstream modeling.

8. Categorical Feature Analysis

8.1 Proportion with Respect to Income Labels

- Examined categorical variables such as education, age, race, Hispanic origin to see their distribution across income groups (>50K vs <50K).
- Observations:
    - Higher education levels (e.g., PhD, Masters) correlated with higher probability of income >50K.
    - Lower education levels had a higher proportion of income <50K.
    - Weighted distribution confirmed imbalanced income labels:
        - ~93% of weighted samples earned <50K
        - ~8% earned >50K

8.2 Reduction of High Cardinality Features

- Education Column: Initially had 17 distinct levels.
- Action: Grouped similar levels into meaningful categories:
    - Grades 1–10 → “Primary/Secondary”
    - Grades 11–12 → “High School”
    - Associate, Bachelor’s, Master’s → consolidated into higher education groups
- Result: Reduced number of levels from 17 → 11–12 categories, making analysis and modeling more tractable.

Feature Engineering and Encoding – Adult Income Census Data

1. Education Feature Consolidation

- Original education column had 17 distinct levels.

- Action: Grouped similar education levels into broader categories:
    - Grades 1–10 → “Kindergarten to 10th grade”
    - Grades 11–12 → “High School”
    - Associate, Bachelor’s, Master’s → combined into higher education groups
- Result: Reduced distinct levels from 17 → 11–12, simplifying analysis and modeling.

2. Country of Birth Features

- Columns analyzed: country of birth self, country of birth mother, country of birth father.
- Objective: Determine if all three columns contribute to income prediction.
- Method: Incremental modeling with area under the curve (AUC) for weighted samples.
    - Result: Each column added significant predictive power (>2% AUC); all retained.
- High Cardinality Handling:
    - 51 distinct countries per column → ~150 potential one-hot columns.
    - Weighted frequency analysis showed most countries <1% of population.
    - Action: Retained top countries (United States, Mexico, China, Italy); all others encoded as "Other".

3. Household and Family Structure Features

- Columns analyzed: detailed household summary in household, detailed household and family stat.
- Observations:
    - detailed household and family stat contained hierarchical and redundant information.
    - Most categories were combinations of keywords like: householder, spouse, child, grandchild, other.
- Actions:
    - Extracted key keywords to reduce cardinality from 51 → 8–10 categories.
    - Determined detailed household summary in household is sufficient; dropped detailed household and family stat.

4. Migration Features

- Columns analyzed: migration code-change in MSA, migration code move within the region, code in region.
- Objective: Reduce dimensionality while preserving predictive power.
- Method: AUC analysis for weighted samples.
    - All three migration-related columns contributed significant predictive power.

- Action: Retained all three columns.

5. Region of Previous Residence

- Columns: region of previous residence, region of previous states.
- Observations:
    - Many distinct combinations between state and region columns.
- Action: Combined the two columns into a single feature region of previous residence to reduce dimensionality and simplify encoding.

6. Education Encoding

- Education is an ordinal feature.
- Action: Applied label encoding to preserve ranking (e.g., higher education → higher encoded value).

7. Other Categorical and Numerical Features

- All other categorical features were one-hot encoded.
- Numerical features were passed through unchanged.
- Applied weighted logistic regression to account for class imbalance.

Model Selection and Evaluation – Adult Income Census Data

1. Logistic Regression

1.1 Model Choice and Rationale

- Selection Reason: Logistic regression was chosen as the base classification model due to:
    - Its interpretability and simplicity.
    - Ability to handle sample weights, which is crucial given the highly imbalanced income classes (~93% <50K, ~8% >50K).
    - Efficient computation for moderate-sized datasets and high-dimensional feature spaces.
- Solver Choice:
    - liblinear solver was used because:
        - It is computationally efficient for small-to-medium datasets.
        - Supports L1 (LASSO) and L2 regularization, which is important for high-dimensional datasets (~280–305 features after one-hot encoding).
        - saga and other solvers were computationally expensive for this dataset size.

1.2 Regularization and Feature Selection

- LASSO (L1) Regularization:

    - Applied to reduce feature dimensionality and remove uninformative predictors.
    - Benefits:
        - Automatically assigns zero weight to irrelevant features.
        - Retains only important predictors, improving model interpretability, precision, and recall.
- Hyperparameter Tuning:
    - Used Optuna for hyperparameter optimization.
    - Parameters tuned:
        - Penalty: L1, L2, or Elastic Net
        - Regularization strength (lambda)

1.3 Model Performance

- Evaluation Metrics (Weighted Logistic Regression):
    - Accuracy: moderate due to class imbalance.
    - Precision and Recall:
        - Minority class (>50K income) recall was low (~0.38), indicating the model struggles to correctly identify high-income samples.
    - Area Under Curve (AUC): provided overall class separation but highlighted imbalance sensitivity.

1.4 Feature Interpretation

- Important Predictors:
    - Industry code and geographical region
    - Gender and age
- Indicates that demographic and employment-related features are significant in predicting income.

2. Random Forest

2.1 Model Choice and Rationale

- Logistic regression has limitations in capturing non-linear relationships and interactions between features, especially in high-dimensional, mixed-type data.
- Random Forest was chosen as a second model due to:
    - Ability to model complex, non-linear patterns.
    - Ensemble approach: Combines multiple decision trees to improve predictive performance and reduce overfitting.
    - Robust to imbalanced datasets: Each tree focuses on a subset of features, reducing bias toward dominant classes.

    - Handles mixed feature types (numerical + categorical) without extensive preprocessing.

2.2 Key Model Mechanics

- At each tree node, a subset of features (e.g., 5) is considered for splitting:
    - Reduces multicollinearity and ensures diverse decision paths.
- Ensemble averaging ensures more stable predictions and better recall for minority classes.
- Categorical variables were passed as-is (without manual encoding) to allow the model to naturally handle category splits.

2.3 Hyperparameter Considerations

- Number of trees (n_estimators): sufficient to stabilize predictions without excessive computation.
- Max depth and feature subset (max_features) tuned to balance bias and variance.

2.4 Model Advantages for Imbalanced Data

- Random Forest reduces sensitivity to class imbalance via:
    - Random feature sampling at each split.
    - Bootstrapping ensures diversity among trees.
    - Combined vote of all trees improves recall for minority class (>50K income).

3. CatBoost Modeling

3.1 Model Choice and Rationale

- CatBoost was selected because it is specifically designed to handle categorical features natively, eliminating the need for manual one-hot encoding.
- Advantages for this dataset:
    - Efficient handling of high-cardinality categorical features, like education, country of birth, and household relationships.
    - Supports weighted samples, which is essential due to the class imbalance (~93% <50K, ~8% >50K).
    - Reduces risk of target leakage compared to naive encoding methods.
    - Handles non-linear relationships and feature interactions automatically.

3.2. Categorical Feature Treatment

- Automatic Encoding: CatBoost internally converts categorical strings to numeric representations using ordered target statistics.
- No manual preprocessing required for categories such as:
    - education (ordinal)
    - country of birth self/mother/father (grouped rare countries into “Other”)

  - detailed household summary
  - marital status, occupation, industry code
- Maintains interpretability of categorical splits within the model while efficiently reducing dimensionality.

3.3 Model Mechanics and Hyperparameters

- Decision Trees Ensemble: CatBoost builds an ensemble of oblivious decision trees, which are symmetric trees reducing overfitting.
- Key Hyperparameters Tuned:
  - iterations (number of trees)
  - depth (tree depth)
  - learning_rate (step size for boosting)
  - l2_leaf_reg (L2 regularization)
  - border_count (for categorical feature binning)
- Hyperparameters optimized using weighted cross-validation to balance class importance and improve minority class recall.

3.4. Advantages for Imbalanced Data

- CatBoost naturally handles class imbalance by allowing sample weights and incorporating gradient boosting with careful leaf estimation.
- Ensemble approach improves recall for minority class (>50K income) compared to logistic regression.
- Captures complex interactions between categorical and numerical features automatically.

3.5. Feature Importance and Interpretation

- CatBoost provides feature importance scores, highlighting the most predictive categorical and numerical variables.
- Key insights observed:
  - Education level, industry code, and region of residence significantly affect income prediction.
  - Demographics (gender, race, age) continue to play a strong role in predicting income levels.
- Allows ranking features based on gain, prediction contribution, or interaction effects.

Conclusion: Based on confusion matrix, AUC Cat Boost model performed slightly better than random forest and logistic regression model.

Selected Model Architectures For Customer Segmentation

K-Means (Baseline Architecture)

K-Means was selected as the baseline clustering model due to its simplicity, scalability, and widespread industry adoption. It provides a useful reference point to understand how ignoring population weights affects cluster formation.

Role in the project:

- Establish a non-weighted benchmark
- Quantify bias introduced by sample imbalance
- Serve as a diagnostic baseline rather than a final solution

Weighted K-Means (Primary Architecture)

Weighted K-Means was chosen as the primary model architecture because it directly incorporates survey weights into centroid updates and objective optimization. This ensures that each observation contributes proportionally to its representation in the population.

Reasons for selection:

- Produces population-representative clusters
- Maintains scalability for large datasets
- Preserves centroid-based interpretability
- Aligns model output with business and policy objectives

To apply K-Means clustering, categorical variables must first be encoded into numerical form. In our dataset, more than 280 features are categorical, which results in a very high-dimensional feature space when one-hot encoding is applied. If distance calculations such as Euclidean or Manhattan distance are performed directly on dense one-hot vectors, clustering becomes both computationally expensive and prone to misuse, as each iteration requires repeated distance calculations across 200,000 records and hundreds of dimensions. To mitigate this, categorical features are represented using sparse matrix formats, which significantly reduce computational and memory overhead by avoiding calculations on zero values. This makes centroid updates and distance computations tractable at scale. However, a known limitation remains: while encoded categorical values are treated as equally spaced in the feature space, they may carry different semantic meanings and relationships that are not captured by geometric distance alone. As a result, K-Means may not fully reflect semantic similarity between categories, which is a recognized trade-off of distance-based clustering on encoded categorical data.

Alternative clustering methods were evaluated but found unsuitable for this problem. DBSCAN was not selected because density estimation becomes unreliable in high-dimensional sparse spaces, leading to unstable cluster formation and excessive sensitivity to parameter choices such as eps and min_samples. Spectral clustering was excluded due to its reliance on pairwise similarity matrix construction and eigen-decomposition, which introduces quadratic memory complexity and makes it infeasible for a dataset of this size. These limitations are directly tied to computational requirements rather than modeling preference. Given these constraints, K-Means—when combined with sparse representations, population weighting, and post-hoc interpretability—provides the most practical balance between scalability, computational efficiency, and business interpretability for large-scale census segmentation.

While non-distance-based segmentation methods such as rule-based tree models and latent class analysis were considered, each presents practical limitations when applied to large-scale, weighted census data. Tree-based methods provide strong interpretability but lack stability and compactness, while probabilistic latent models do not scale efficiently to high-dimensional, sparse feature spaces. As a result, distance-based clustering—specifically weighted K-Means combined with sparse representations and post-hoc explainability—remains the most suitable and defensible segmentation approach for this use case

Selecting the number of clusters (K) is the primary hyperparameter decision in K-Means clustering. Traditional selection techniques such as the Elbow method and Silhouette score were evaluated but found to be less suitable for this dataset. The Elbow method relies on changes in within-cluster inertia, which tends to decrease smoothly in high-dimensional, sparsely encoded feature spaces without exhibiting a clear inflection point, making it difficult to identify an optimal K. Similarly, Silhouette scores depend on point-to-centroid and inter-cluster distances that become unstable and less interpretable when applied to mixed-type data dominated by one-hot encoded categorical variables. In contrast, the Davies–Bouldin Index (DBI), which evaluates the ratio of within-cluster dispersion to between-cluster separation, provides a more robust relative measure of cluster compactness and separation under these conditions. DBI scores were computed across a range of candidate K values, and the selected number of clusters (in our case K = 5) corresponds to the configuration with the lowest DBI score, indicating improved cluster separation while avoiding over-segmentation.

Although the lowest Davies–Bouldin Index was observed at K=2, this configuration was not selected due to insufficient segmentation granularity for population analysis. DBI values increased sharply between K=2 and K=3 and showed relatively gradual variation beyond K=4, indicating diminishing improvements in cluster separation with higher K values. Based on this stabilization pattern, K=5 was selected as a balanced solution that provides meaningful segmentation granularity while avoiding over-fragmentation. This choice reflects a trade-off between statistical compactness and business interpretability rather than strict optimization of the DBI metric.

Clusters were learned using unweighted K-Means, and survey weights were applied post-clustering to estimate real-world population representation. While the unweighted solution suggests relatively balanced clusters at the population level (with clusters 0, 1, 2, and 4 each representing ~15–18% of the population and cluster 3 at ~27.5%), this balance does not hold when clustering is performed using population weights.

In the weighted K-Means solution, clusters 0, 1, and 2 emerge as the dominant population segments, with Cluster 1 alone representing the largest share (33.3%), while cluster 3's population share drops substantially (~16%) and Cluster 4, which represents only ~2.0% of the total population, indicating a small but distinct group. This divergence indicates that unweighted K-Means captures patterns driven by data density, whereas weighted K-Means reveals segments that are truly prevalent in the underlying population.

K - prototypes uses a matching-based distance, which allows for clearer attribution of feature contributions to each cluster compared to centroid-based distances used in weighted K-Means. Using this approach, Cluster 1 represents the smallest population segment (~2%), while Cluster 2 is the most prevalent (~30.4%), with Clusters 0, 3, and 4 forming mid-sized segments ranging between ~18% and ~23%.

For each cluster, weighted averages of feature values were computed, and feature contributions were summarized based on their average squared deviation from the cluster centroid. Features with the largest aggregated contributions were treated as descriptively dominant within each cluster. Across all clusters for weighted k means, features fall into five interpretable dimensions and we also observed some overlapping cases between clusters.

1. Mobility & Residential Stability : Migration codes, residence 1 year ago, sunbelt indicators
2. Labor Market Attachment :Weeks worked, wage per hour, full/part-time status, class of worker
3. Demographics & Household Composition: Age, sex, marital status, household summary, family members under 18
4. Occupation & Industry Structure: Major/detailed occupation and industry recodes
5. Income & Capital Signals :Dividends, capital gains/losses, veterans benefits

To understand the drivers of cluster assignment, a Random Forest classifier was trained on the original features to predict cluster membership, and SHAP values were computed to quantify feature importance. The resulting SHAP plots for both weighted K-Means and K-Prototypes indicate that features such as education, age, tax filer status, household composition, migration history, and weeks worked consistently contribute the most to cluster assignments across groups.