JPM Take Home Assignment

Niloofar Fadavi

February 16, 2026

1 Understanding the Data Collection Process

This report presents a comprehensive analysis of the Current Population Survey (CPS) dataset for income prediction modeling. The dataset contains 5,000 observations with 40 features, including demographic characteristics, employment information, migration patterns, and financial data. Understanding the data collection methodology and feature distributions is essential for building robust predictive models that accurately represent the U.S. civilian non-institutional population. In this section, we summarize the main concepts behind how the dataset was collected. It worth it to have an understanding of how the data is collected before building a machine learning model. For example, it this data set it was not clear for me why we have a column called weight. The following information makes it clear. So we can decide with more care about the machine learning techniques we choose.

The dataset used in this project comes from the Current Population Survey (CPS), which is conducted by the U.S. Census Bureau and the Bureau of Labor Statistics. CPS is a probability-based survey designed to represent the civilian non-institutional population of the United States. Rather than surveying every individual, CPS uses a multi-stage sampling strategy to efficiently and reliably collect data.
The entire United States is divided into 1,987 Primary Sampling Units (PSUs). Each PSU consists of a county or a group of neighboring counties within a state. Metropolitan areas often form their own PSUs, while smaller counties are combined to ensure diversity and efficient data collection. Within each state, PSUs are grouped into strata. A stratum may contain one PSU (called a self-representing stratum) or multiple PSUs that are similar in characteristics such as unemployment rate, industry composition, and wage levels. If a stratum contains only one PSU, that PSU is automatically selected. If a stratum contains multiple PSUs, one PSU is selected randomly using a probability proportional to size (PPS) method. PPS means that PSUs with larger populations have a higher chance of being selected into the sample. For example, within a stratum:

- A PSU with a population of 50,000 has twice the probability of being selected compared to a PSU with a population of 25,000.

This approach ensures that larger population areas are adequately represented while still allowing smaller areas to be included in the sample. After a PSU is selected, households within that PSU are grouped into clusters based on geographic proximity. A systematic sampling method is then used to select a subset of households for interviews. This clustering approach reduces travel costs and improves operational efficiency while maintaining statistical validity.

Because not all PSUs and households have the same probability of being selected, each observation in the dataset is assigned a weight. The sampling weight represents how many people in the real population a single record corresponds to. In general, the weight is the inverse of the probability of selection:

$$\text{Weight} = \frac{1}{P(\text{selection})}$$

The presence of sampling weights and probability-based selection affects how we interpret the dataset and how models should be evaluated.

2 Initial Data Audit and Structure

We must first understand the 40 features that will be used to train the machine learning model. The data has to be preprocessed before modeling. This includes identifying the data type of each feature, examining distributions and summary statistics, and checking for missing values. So, in this section, we show the information we obtained regarding the features and prepare the require information for the prepossessing phase of the data.

The target variable `label` has the following distribution:

Table 1: Target Variable Distribution

Class	**Count**	**Percentage**
- 50000.	4,714	94.28%
50000+.	286	5.72%

Remark 1: The dataset exhibits severe class imbalance with a ratio of approximately 16:1. This will require special handling during modeling (e.g., stratified sampling, class weights, upper/under sampling, or appropriate evaluation metrics like precision-recall curves).

2.1 Numeric Feature Review

Table 2 presents a comprehensive statistical summary of all numeric features in the dataset, including demographic variables, time-based measures, and financial variables.

Table 2: Numeric Features: Complete Statistical Summary

Feature	**Type**	**Min**	**Max**	**Mean**	**Missing**	**Missing %**
Demographic & Time Variables						
age	int64	0	90	34.48	0	0.00%
weeks worked in year	int64	0	52	23.94	0	0.00%
year	int64	94	95	94.50	0	0.00%
Employment Variables						
num persons worked for employer	int64	0	6	2.03	0	0.00%
Financial Variables (Zero-Inflated)						
wage per hour	int64	0	8,000	63.25	0	0.00%
capital gains	int64	0	99,999	449.74	0	0.00%
capital losses	int64	0	3,770	39.30	0	0.00%
dividends from stocks	int64	0	99,999	167.02	0	0.00%

Several features are stored as integers but do not represent true numeric magnitudes. Instead, they are coded category identifiers. Treating them as continuous variables would introduce false ordering and distance relationships into the model. Therefore, these fields are recast as categorical features before encoding. Table 3 lists these variables and their basic properties.

We assume that the coded categories have no intrinsic ordering and must be handled as nominal variables. **Remark 2)** For preprocessing, these features are encoded using categorical encoding methods rather than numeric scaling.

2.2 Categorical Features Analysis

The dataset contains 28 features stored as categorical/text data (dtype: str), plus the 4 integer-coded categorical features discussed previously. These features represent demographic characteristics, employment information, geographic data, and household composition. Table 4 presents a comprehensive summary organized by cardinality level.

Table 3: Integer-Coded Categorical Features

Feature	Type	Unique	Missing	Missing %
detailed industry recode	int64	49	0	0.00%
detailed occupation recode	int64	46	0	0.00%
own business or self employed	int64	3	0	0.00%
veterans benefits	int64	3	0	0.00%

Remark 3) As shown in the table, these categorical features must be converted to a proper numeric format before they can be used in machine learning models.

3 Data Wrangling

3.1 Missing Data

Three different missing-value patterns are observed in the dataset. Each pattern is handled differently to avoid introducing bias or artificial information into the model.

3.1.1 Complete Features (24 features)

These variables contain no missing values and can proceed directly to encoding and transformation steps. No imputation is required.

3.1.2 Low Missingness (5 features)

These variables contain a small fraction of missing values (below about 3%), which is consistent with survey nonresponse or recording gaps:

- `hispanic origin`: 13 missing (0.26%)
- `state of previous residence`: 24 missing (0.48%)
- `country of birth self`: 87 missing (1.74%)
- `country of birth mother`: 146 missing (2.92%)
- `country of birth father`: 154 missing (3.08%)

Handling strategy: Because this dataset comes from a survey, some answers are missing, especially for more personal questions. People may skip questions for privacy or other personal reasons. This means the missing values are not always random and may be related to important characteristics of the person. If we delete all rows with missing values, we lose real observations and may remove an important subgroup from the data. If we fill the missing values with something like the most common category, we add made-up information and change the true distribution of the data. Both actions can bias the model. For example, the feature `country of birth`, some people may choose not to answer. This choice itself may be related to income or social factors. If we drop these rows, we lose that group. If we replace the missing value with the most common country, we insert incorrect data. Instead, we create a separate category called **"Unknown"** for missing values and keep those records. This keeps all observations, avoids adding false information, and allows the model to learn whether nonresponse itself carries useful signal.

3.1.3 Structured Missingness (4 features)

A subset of migration-related variables shows a highly consistent missingness pattern, with approximately 50% of values missing in each field. These features are: `migration code-change in msa`, , `migration code-change in reg` , `migration code-move within reg` , `migration prev res in sunbelt`.
The nearly identical missing rate across all four variables indicates that the missingness is not due to data

Table 4: Categorical Features: Complete Summary

Feature	Unique	Missing	Missing %
Binary Features (2 categories)			
sex	2	0	0.00%
Low Cardinality (3–9 categories)			
enroll in edu inst last wk	3	0	0.00%
live in this house 1 year ago	3	0	0.00%
member of a labor union	3	0	0.00%
migration prev res in sunbelt	3	2,509	50.18%
fill inc questionnaire for veteran's admin	3	0	0.00%
citizenship	5	0	0.00%
race	5	0	0.00%
family members under 18	5	0	0.00%
reason for unemployment	6	0	0.00%
region of previous residence	6	0	0.00%
tax filer stat	6	0	0.00%
marital stat	7	0	0.00%
detailed household summary in household	7	0	0.00%
full or part time employment stat	8	0	0.00%
migration code-change in reg	8	2,509	50.18%
class of worker	9	0	0.00%
hispanic origin	9	13	0.26%
migration code-change in msa	9	2,509	50.18%
migration code-move within reg	9	2,509	50.18%
Medium Cardinality (10–20 categories)			
major occupation code	14	0	0.00%
education*	17	0	0.00%
High Cardinality (21+ categories)			
major industry code	23	0	0.00%
detailed household and family stat	26	0	0.00%
country of birth father	41	154	3.08%
country of birth self	41	87	1.74%
country of birth mother	42	146	2.92%
state of previous residence	49	24	0.48%

*Ordinal categorical variable with natural ordering from "Less than 1st grade" to "Doctorate degree"

quality problems or random reporting gaps. Instead, it reflects the survey design itself. These questions are only asked for respondents who reported a residential move within the last year. Because the missingness mechanism is systematic and rule-based, standard missing-data treatments such as row deletion or statistical imputation are not appropriate.

Handling strategy: All missing values in these variables are recoded as an explicit categorical level labeled **"Not Applicable"**. This preserves the survey logic, retains all observations, and allows models to learn whether migration-question applicability itself is predictive.

3.2 Categorical Encoding Strategy

3.2.1 Ordinal Features

Most categorical features are **nominal** (no intrinsic ordering). However, one feature has natural ordering: **education** (17 categories. This feature is ordinal because the categories follow a natural progression from lowest to highest education level. Unlike nominal features, the order carries real meaning and should be preserved during preprocessing.

Handling strategy: Because this variable has a true order, it is encoded with integers instead of one-hot encoding. Each category is mapped once to an increasing numeric level so that larger numbers correspond to higher educational attainment. This keeps the ranking information. Based on our current domain understanding, we defined the following mapping for encoding the ordered education levels. The assigned values preserve the correct ranking between categories but do not necessarily represent exact proportional differ-

ences. In future work, this mapping could be refined by consulting subject-matter experts to choose more representative numeric spacing between levels.

```
education_map = {
    "Children": 0, "Less than 1st grade": 1, "1st-4th grade": 2,
    "5th-6th grade": 3, "7th-8th grade": 4, "9th grade": 5,
    "10th grade": 6, "11th grade": 7, "12th grade no diploma": 8,
    "High school graduate": 9, "Some college": 10,
    "Associate's (vocational)": 11, "Associate's (academic)": 12,
    "Bachelor's degree": 13, "Master's degree": 14,
    "Professional degree": 15, "Doctorate degree": 16
}
```

For more information of how exactly the mapping is done, please refer to the code.

3.2.2 Preprocessing Task: Categorical Encoding Strategy

1. Binary Features (2 categories)
Features: `sex`
Strategy: Label encoding (0/1) or one-hot encoding (creates 1 binary column).
2. Low Cardinality Features (3–9 categories) These variables contain only a small number of distinct categories. Expanding them does not significantly increase dimensionality. Table 5 lists all low-cardinality categorical features along with their cardinality.

Table 5: Low Cardinality Categorical Features (3–9 categories)

#	Feature	Cardinality
1	enroll in edu inst last wk	3
2	live in this house 1 year ago	3
3	member of a labor union	3
4	migration prev res in sunbelt	3
5	fill inc questionnaire for veteran's admin	3
6	own business or self employed*	3
7	veterans benefits*	3
8	citizenship	5
9	race	5
10	family members under 18	5
11	reason for unemployment	6
12	region of previous residence	6
13	tax filer stat	6
14	marital stat	7
15	detailed household summary in household	7
16	full or part time employment stat	8
17	migration code-change in reg	8
18	class of worker	9
19	hispanic origin	9
20	migration code-change in msa	9
21	migration code-move within reg	9

*Integer-coded categorical features (stored as int64 but treated as nominal categories).

Strategy Each feature with k categories is converted into $k-1$ binary columns (dropping one category to avoid multicollinearity). For example, `race` with 5 categories becomes 4 binary indicator columns, with the reference category (e.g., White) implicitly represented when all indicator columns are zero. This encoding preserves the categorical nature of the variables without imposing false ordering or magnitude relationships between categories.

3. Medium Cardinality Features (10–20 categories) There is only one feature in this category. One-hot encoding is still feasible but begins to expand the feature space more noticeably. Table 6 lists the only medium-cardinality categorical feature along with its cardinality.

Table 6: Medium Cardinality Categorical Features (10–20 categories)

#	Feature	Cardinality
1	major occupation code	14

*Ordinal categorical feature (already handled with ordinal encoding in Section 3.2.1).

Strategy This feature is one-hot encoded, creating 13 binary indicator columns.

4. High Cardinality Features (21+ categories) These variables contain a large number of distinct categories. Direct one-hot encoding would create an excessively wide feature matrix, leading to high dimensionality, increased sparsity, and greater risk of overfitting. Therefore, alternative encoding strategies are required to reduce dimensional expansion while preserving categorical information. Table 7 lists all high-cardinality categorical features along with their cardinality. If these 8 features were one-hot encoded

Table 7: High Cardinality Categorical Features (21+ categories)

#	Feature	Cardinality
1	major industry code	23
2	detailed household and family stat	26
3	country of birth self	41
4	country of birth father	41
5	country of birth mother	42
6	detailed occupation recode*	46
7	detailed industry recode*	49
8	state of previous residence	49

*Integer-coded categorical features (stored as int64 but treated as nominal categories).

directly, they would generate approximately 317 new binary columns (sum of cardinalities minus 8 reference categories). This would create a highly sparse feature matrix (most values would be zero) and introduce risk of overfitting, especially with limited training data (5,000 observations).

Strategy To control dimensionality while preserving categorical information, we apply **frequency encoding**. This approach replaces each category with its relative frequency (proportion) in the training dataset. Categories that appear frequently receive higher values, while rare categories receive lower values. Basically, we are trying to give the model a summary of the data instead of the membership to each of the classes themselves. Frequency encoding keeps the feature as a single numeric column while still transferring useful information about the category distribution. After encoding, the original categorical column is replaced by a single numeric column containing frequency values between 0 and 1.

4 Methodology

This section describes the machine learning approach used to predict whether an individual's income exceeds $50,000 per year. The methodology follows a systematic pipeline from data splitting through model evaluation, with specific attention to the severe class imbalance and the presence of outliers in the financial features.

4.1 Train-Test Split Strategy

Due to the severe class imbalance (94.28% negative class, 5.72% positive class), a **stratified train-test split** is employed to ensure both training and test sets maintain the same class proportions. This prevents

the test set from being unrepresentative of the overall population distribution and ensures reliable evaluation metrics.

The data is split as follows:

- **Training set**: 80% (4,000 observations)
- **Test set**: 20% (1,000 observations)

The test set is held out entirely and used only for final model evaluation after all development decisions are complete. No information from the test set is used during model training, hyperparameter tuning, or feature scaling to prevent data leakage.

4.2 Handling Class Imbalance: Oversampling

The dataset is strongly imbalanced, with a class ratio of approximately 16.5:1 (4,714 negative vs. 286 positive examples). Without correction, a model can achieve high accuracy by predicting only the majority class, while failing to detect high-income individuals.

Approach Used Here: Random Oversampling of the Minority Class: In this study, class imbalance is handled using **random oversampling** of the minority class during training. In each training fold, the minority class is randomly duplicated to match the majority class size (1:1 ratio). This forces the model to learn decision boundaries using a balanced training signal without discarding potentially useful majority-class examples. (I didn't use undersampling because it would leave too few observations compared to the number of features, making it harder for the model to learn properly.)

Other valid imbalance-handling methods also exist and can be evaluated in future work:

- Class weights: Increase the loss penalty for minority-class errors without changing the sample counts.
- Undersampling: Reduce majority samples to balance classes, though this discards potentially useful data.
- SMOTE: Generate synthetic minority samples rather than simple duplication.
- Hybrid methods: Combine undersampling and oversampling (e.g., SMOTE + Tomek links).

Model performance can change depending on which strategy is used. In future experiments, these alternative methods can be tested and compared to determine which approach gives the best precision–recall tradeoff for this dataset.

4.3 Scaling

I used RobustScaler to scale the numerical features but other approaches can be applied and assessed too.

4.4 Model Selection and Cross-Validation

Three well-known classification models are evaluated in this study: Logistic Regression, Random Forest, and Support Vector Machine (SVM). These models were selected because they represent three different learning families (linear, tree-based, and margin-based) and are widely used as strong baselines in tabular classification problems. For each model, a grid of hyperparameters is defined, and the best configuration is selected using cross-validation. Grid search with stratified k-fold cross-validation is applied so that class proportions remain consistent across folds. This reduces selection bias and produces a more reliable estimate of model performance. Model selection depends on the evaluation metric. Two common metrics for imbalanced classification are **precision** and **recall**.

Precision measures how many predicted positive cases are actually positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision means that when the model predicts income above \$50K, it is usually correct.

Recall measures how many actual positive cases are successfully detected:

$$\text{Recall} = \frac{TP}{TP + FN}$$

High recall means that most individuals with income above $50K are identified by the model.
There is usually a tradeoff between precision and recall. The preferred metric depends on the project objective and error cost. The key question is which error is more harmful: falsely labeling someone as high-income (false positive) or missing a truly high-income individual (false negative). This is a management and domain decision and should be aligned with business goals. In this work, precision is treated as the primary metric. Therefore, hyperparameter tuning is performed using cross-validation scores based on precision. Recall and accuracy are still reported for all models to provide a complete performance view. The table below summarizes precision and recall for the three methods on the test set. Table 8 summarizes the test set performance for all three models. The choice of best model cannot be determined by a single metric alone

Table 8: Model Performance with Random Undersampling (Test Set, n=1,000)

Model	**Accuracy**	**Precision**	**Recall**
Logistic Regression	0.919	0.372	0.614
Random Forest	0.915	0.379	**0.772**
SVM	**0.947**	**0.643**	0.158

and should be made in consultation with the business unit to understand the specific priorities and use case requirements. Due to the severe class imbalance (94.28% negative class), accuracy is not a reliable metric for model comparison. A naive model that always predicts the majority class (income $\leq$ $50,000) would achieve 94% accuracy while providing no predictive value. Therefore, model selection must focus on precision and recall metrics for the positive class, which directly measure performance on the minority class of interest.
The model selection process should follow these steps:

1. **Identify the business objective**: What is the model being used for?
2. **Quantify error costs**: What is the financial or operational cost of a false positive vs. a false negative?
3. **Determine priority**: Is precision more important, recall more important, or are both equally important?
4. **Select the appropriate model**:
   - If precision is critical $\rightarrow$ Choose SVM
   - If recall is critical $\rightarrow$ Choose Random Forest
   - If balanced performance is desired $\rightarrow$ Choose Random Forest (Class Weights) or Logistic Regression
5. **Fine-tune decision threshold**: If needed, adjust the classification threshold (default 0.5) on a validation set to achieve the desired precision-recall balance

5 Customer Segmentation Analysis

Segmentation divides the customer base into homogeneous groups that share similar characteristics, enabling the development of tailored marketing strategies for each segment. The preprocessed dataset contains 135 features after encoding all categorical variables. While this high-dimensional representation is beneficial for supervised learning tasks (income prediction), it presents significant challenges for unsupervised clustering, Distance-based clustering algorithms like K-Means perform poorly in high-dimensional spaces because distances between points become less meaningful as dimensions increase and more importantly it will be difficult to describe and interpret for the business unit. Thus, we selected 12 features across four categories:

1. **Demographics**: age, education level (ordinal), sex
2. **Economic Status**: capital gains, capital losses, dividends from stocks
3. **Employment**: weeks worked per year, number of persons worked for employer
4. **Derived Indicators**:
    - Marital status (married vs. not married)
    - Employment status (full-time vs. other)
    - Presence of capital income (any capital gains or dividends)
    - High education (Bachelor's degree or higher)

This feature set balances dimensionality reduction (from 135 to 12 features) with interpretability, enabling clear segment profiling for marketing purposes.

5.1 Determining Optimal Number of Clusters

To identify the appropriate number of customer segments, we evaluated K-Means clustering with different values of K (number of clusters) ranging from 2 to 10. Three complementary metrics were used to assess clustering quality:

- **Elbow Method (Inertia)**: Measures within-cluster compactness. Lower values indicate tighter clusters. The "elbow point" where the rate of decrease slows suggests an optimal K.
- **Silhouette Score**: Measures how well each point fits within its assigned cluster compared to other clusters. Values range from -1 to 1, with higher values indicating better-defined, more separated clusters.
- **Davies-Bouldin Index**: Measures the average similarity between each cluster and its most similar neighboring cluster. Lower values indicate better clustering with greater separation between clusters.

Table 9 presents the clustering quality metrics for different values of K.

Table 9: Clustering Quality Metrics by Number of Clusters

K	Silhouette Score	Davies-Bouldin Index	Inertia
2	0.261	1.602	47,050
3	0.257	1.656	41,873
4	0.278	1.406	37,163
5	0.292	1.341	32,657
6	0.308	1.331	28,975
7	0.324	1.233	25,446
8	0.328	1.164	22,378
9	0.331	1.127	20,375
10	**0.350**	**1.034**	18,848

Decision Point Consultation: Our clustering quality metrics show continuous improvement as K increases from 2 to 10, with no plateau or clear optimal point (even elbow is not found). This pattern suggests that even larger values of K (beyond 10) would likely yield further statistical improvements. However, selecting the number of segments is not purely a statistical optimization problem.

Customer segmentation serves a business purpose: enabling marketing teams to develop and execute targeted strategies for distinct customer groups. While K=10 achieves the best statistical metrics, managing 10 separate marketing campaigns simultaneously exceeds the practical capacity of most organizations. Each additional segment requires dedicated resources for strategy development, content creation, channel management, and performance tracking. In practice, marketing teams effectively manage 4-6 customer segments.

Beyond this range, segments become difficult to differentiate, strategies overlap, and execution quality suffers due to resource constraints.

We select **K=4** as a pragmatic balance between statistical quality and operational feasibility.

5.2 Segment Profiles

Table 10 presents the demographic, economic, and employment characteristics of the four identified segments.

Table 10: Customer Segment Profiles

Characteristic	**Seg 0**	**Seg 1**	**Seg 2**	**Seg 3**
Segment Name	Youth/Family	Seniors /Retirees	Working Professionals	Premium Educated
Size (customers)	1,565	1,030	2,305	100
Size (%)	31.3%	20.6%	46.1%	2.0%
Demographics				
Average Age (years)	11	60	39	45
Male (%)	50%	32%	51%	62%
Married (%)	0%	65%	64%	54%
Bachelor's+ (%)	0%	10%	25%	37%
Economic Status				
Income ¿50K (%)	0.0%	1.4%	10.6%	27.0%
Has Capital Income (%)	1%	15%	21%	21%
Avg Capital Gains ($)	0	161	904	0
Avg Dividends ($)	2	358	184	406
Employment				
Full-time (%)	1%	2%	42%	34%
Avg Weeks Worked	1.4	3.2	47.8	40.1

5.2.1 Segment 0: Youth/Family Segment (31.3%)

This segment comprises children and young dependents with an average age of 11 years. No members earn over $50,000 or work full-time, indicating they are minors included in the household survey. This segment represents the family context of working adults.

5.2.2 Segment 1: Seniors/Retirees (20.6%)

Characterized by an average age of 60 years and predominantly female (68%), this segment consists of retirees and pre-retirees. Only 2% work full-time, and 65% are married. While income from employment is minimal (1.4% earn over $50,000), 15% have investment income, suggesting reliance on retirement savings and fixed-income products.

5.2.3 Segment 2: Working Professionals (46.1%)

The largest segment, comprising 46% of the population, consists of mid-career working adults (average age 39). Notable characteristics include 42% full-time employment, balanced gender distribution, and 64% married. Education levels are moderate (25% Bachelor's+), and 10.6% earn over $50,000. This segment shows significant investment activity (21% have capital income, average capital gains of $904).

5.2.4 Segment 3: Premium Educated Segment (2.0%)

A small but high-value segment of 100 customers averaging 45 years old. This group shows the highest income levels (27% earn over $50,000) and education (37% Bachelor's+). While capital gains are zero,

average dividends of $406 indicate dividend-focused investment strategies. This segment represents affluent, educated professionals with substantial purchasing power.

5.3 Marketing Recommendations

Based on segment characteristics, we recommend the following marketing strategies:

Segment 0 - Youth/Family Segment:

- **Strategy**: Family-oriented marketing targeting parents
- **Products**: Educational services, children's products, family packages
- **Channels**: Parents (email, social media), schools, family events
- **Messaging**: Family-friendly, educational value, safety, trust

Segment 1 - Seniors/Retirees:

- **Strategy**: Retirement and security-focused offerings
- **Products**: Retirement planning, health insurance, travel packages, fixed-income products
- **Channels**: Traditional mail, email, phone, community centers
- **Messaging**: Security, comfort, peace of mind, reliability
- **Cross-sell**: Investment advisory, tax optimization, estate planning, healthcare products

Segment 2 - Working Professionals:

- **Strategy**: Convenience and value for busy professionals
- **Products**: Investment products, workplace benefits, convenience services
- **Channels**: Email, mobile apps, social media, online platforms
- **Messaging**: Convenience, efficiency, work-life balance, value
- **Cross-sell**: Investment advisory, professional development, workplace benefits, tech products

Segment 3 - Premium Educated:

- **Strategy**: Premium, personalized service
- **Products**: Wealth management, premium services, luxury goods, exclusive offerings
- **Channels**: Personalized email, direct mail, VIP events, dedicated account managers
- **Messaging**: Exclusivity, quality, status, premium experience
- **Cross-sell**: Comprehensive wealth management, tax optimization, premium tech products

The segmentation model provides a foundation for targeted marketing execution. Recommended next steps include:

1. **Validation**: Test segment definitions with marketing team to ensure alignment with organizational capacity
2. **Campaign Development**: Create segment-specific marketing campaigns with tailored messaging and offers
3. **Performance Tracking**: Establish metrics to measure segment-specific campaign performance and customer lifetime value

4. **Refinement**: Periodically re-cluster as customer base evolves and business priorities shift
5. **Personalization**: Use segment membership to personalize website content, email campaigns, and product recommendations

The segmentation analysis demonstrates how unsupervised learning can transform a large, undifferentiated customer base into actionable marketing segments, enabling more efficient resource allocation and improved customer engagement.