

FINAL REPORT: US FOOD PROJECT – ASSISTANCE NEED ANALYSIS AND MODELING

I. Project Overview

The goal of the US FOOD project based on the Assistance module that I'm working on is to analyze, identify and PRIORITIZE the distribution and effectiveness of aid programs across regions and model a **Need Score** derived from assistance coverage and participation indicators, enriched with **socioeconomic and demographic data** at the regional level.

. The project involves:

1. **Data preparation**
2. **Regression analysis (Notebook A)**
3. **Classification modeling (Notebook B)**
4. **Clustering analysis (Notebook C)**

The final objectives are :

- Identify the **main drivers** of the Need Score
- Evaluate whether **socioeconomic indicators alone** can predict need
- Compare **multiple machine learning models**
- Segment regions using **unsupervised learning**

2. Data Description

2.1 Data Sources

- **Assistance dataset**
 - Coverage rate
 - Participation rate
 - Benefit per capita
 - Need score (target variable)
- **Socioeconomic dataset**
 - Income, poverty indicators
 - Demographic composition
 - Rurality and population change

2.2 Data Integration

- Datasets merged by `ID_Region` (assistance and socio-economics)
- Socioeconomic indicators aggregated at the regional level

Data Preparation

File in the attachment on blackboard :Assistance data preparation.ipynb

Datasets used:

- DIM_Assistance_CLEAN.xlsx – cleaned assistance data
- DIM_Socioeconomics_CLEAN.xlsx – socioeconomic indicators with cleaned data from my team member sarah .

Steps performed:

1. Data Loading

- Connected to SQL Server (DW_US_FOOD) using **ODBC (pyodbc)**.
 - Extracted core and dimensional tables:
 - **Dim_Assistance** (main analytical dataset)
 - **Dim_Region** (geographical context)
 - **Dim_Date** (temporal dimension)
 - Imported data into Pandas DataFrames for analysis.
-

2. Data Exploration & Quality Assessment

- Inspected dataset structure:
 - Number of rows and columns
 - Data types and variable definitions
 - Performed descriptive analysis:
 - Summary statistics (mean, median, min, max, quartiles)
 - Distribution inspection for numerical variables
 - Data quality checks:
 - Duplicate row detection
 - Missing value analysis per column
 - Generated automated exploratory reports using **ydata_profiling** to:
 - Identify skewness and outliers
 - Detect correlations and redundant features
 - Highlight data consistency issues
-

3. Feature Selection & Column Reduction

- Removed non-informative or redundant variables:
 - **ID_Assistance** (technical identifier)
 - **Coverage_change** (derived and highly correlated)
 - **Policy_access_score** (low analytical contribution)

- Retained **ID_Region** due to its relevance for spatial analysis.
-

4. Data Cleaning

- Handled invalid placeholder values:
 - Replaced values such as -8888, -888800 with NaN
 - Outlier treatment:
 - **IQR method** applied to:
 - Coverage_rate
 - Participation_rate
 - **Winsorization (1%)** applied to:
 - Benefit_per_capita (to preserve extreme but valid values)
-

5. Missing Value Treatment

- Applied **K-Nearest Neighbors (KNN) imputation**:
 - Number of neighbors: $k = 5$
 - Variables imputed:
 - Coverage_rate
 - Participation_rate
 - Benefit_per_capita
 - Ensured numerical coherence across imputed values.
-

6. Data Transformation

- Applied **logarithmic transformation (log1p)** to:
 - Benefit_per_capita (to reduce right skewness)
 - Normalized key numerical variables using **Min-Max Scaling**:
 - Coverage_rate
 - Participation_rate
 - Benefit_per_capita
-

7. Feature Engineering – Need Score

- Constructed a composite **need_score** using scaled variables:
 - Low coverage increases need: $(1 - \text{Coverage_rate})$
 - Low participation increases need: $(1 - \text{Participation_rate})$
 - High benefit intensity contributes to need
- Final formula:
 - `need_score = average( low_coverage , low_participation , high_benefit )`

- Interpretation:
 - Higher values indicate greater unmet assistance need.

8. Final Validation

- Removed remaining duplicate observations.
 - Verified final dataset integrity:
 - No missing values
 - Consistent variable ranges
 - Conducted visual validation:
 - Histograms for distributions
 - Boxplots for outlier confirmation
 - Correlation heatmap
 - Temporal trends by year
-

9. Data Export

- Final cleaned and engineered dataset saved as:
 - **DIM_Assistance_CLEAN.xlsx**

Outcome:

- Dataset prepared for:
 - Regression modeling
 - Classification
 - Clustering

Regression Analysis (Notebook A)

File in the attachment on blackboard :Notebook A Regression.ipynb

3.1 Objective

Predict the **Need Score** using supervised regression models and identify the **most influential features**.

3.2 Feature Selection Strategy

An initial pool of socioeconomic and demographic features was evaluated using **tree-based feature importance**.

When focusing **only on socioeconomic variables**, the strongest predictors were:

Socio-economics Feature	Importance
Income	0.18
Poverty_rate	0.16
Child_poverty	0.15

3.4 Models Evaluated

- Random Forest Regressor
- XGBoost Regressor
- LightGBM Regressor

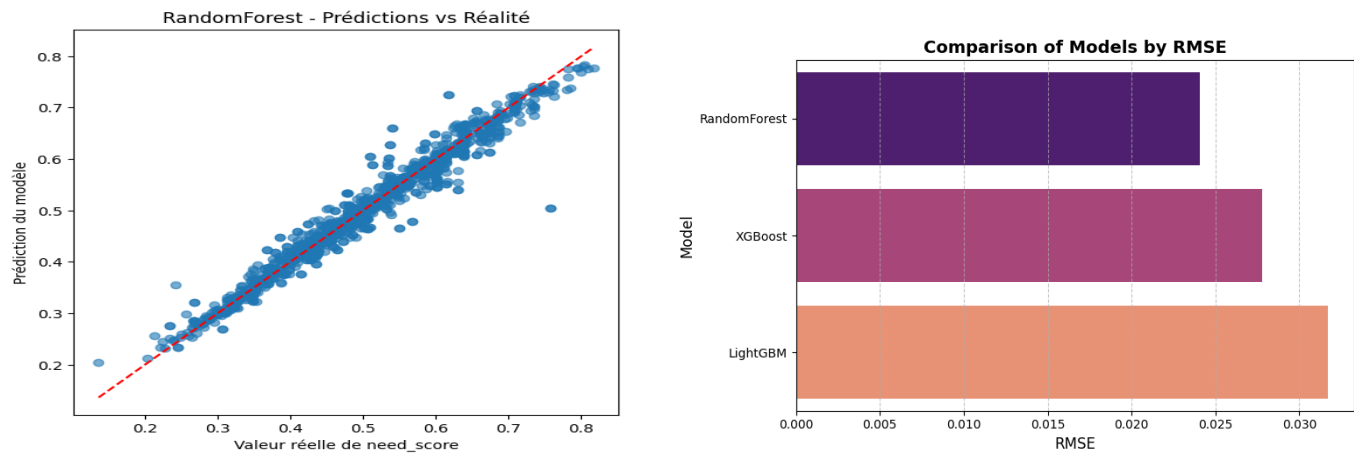
3.5 Model Comparison

Models were evaluated using:

- R^2
- RMSE
- MAE

A comparative table and bar-chart visualizations were used to assess performance.

	Model	R2	RMSE	MAE
0	RandomForest	0.961418	0.024045	0.014744
1	XGBoost	0.948459	0.027792	0.018929
2	LightGBM	0.933013	0.031684	0.022044



4. Notebook B – Classification Analysis

4.1 Objective

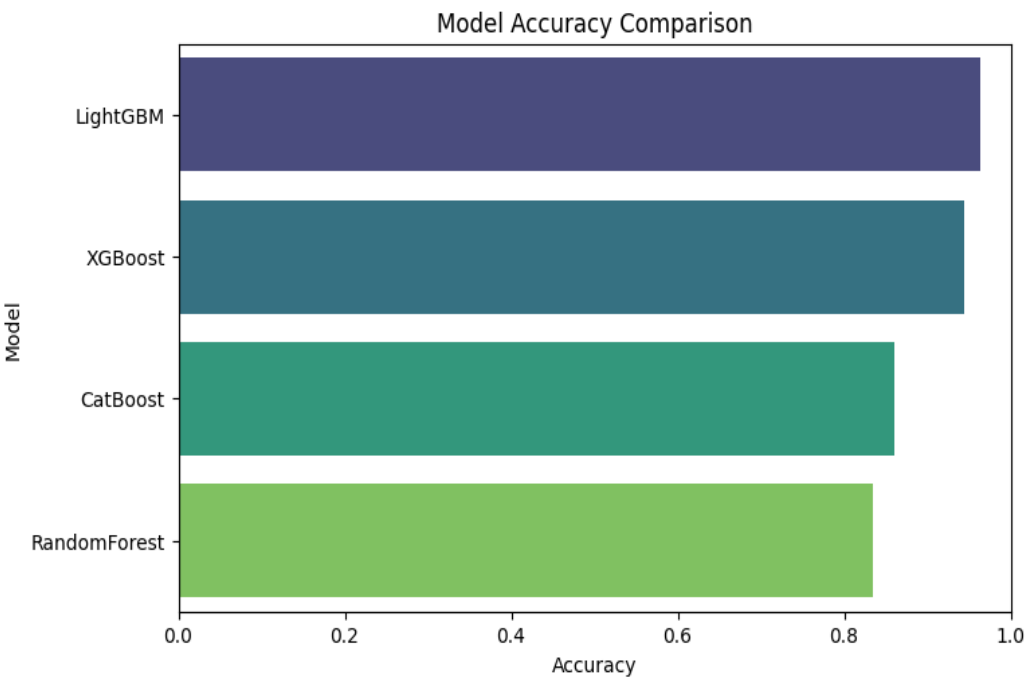
Classify regions into **need-level categories** based on socioeconomic and assistance indicators.

4.2 Models Evaluated

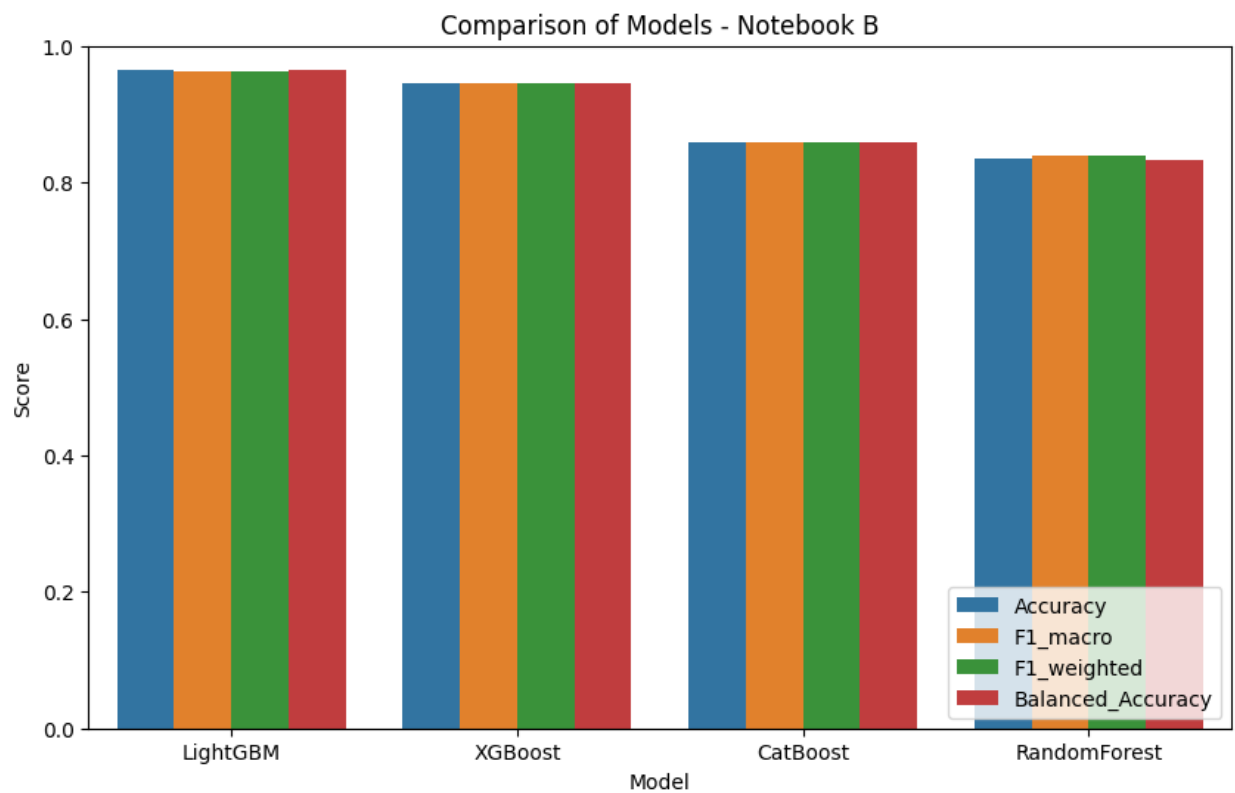
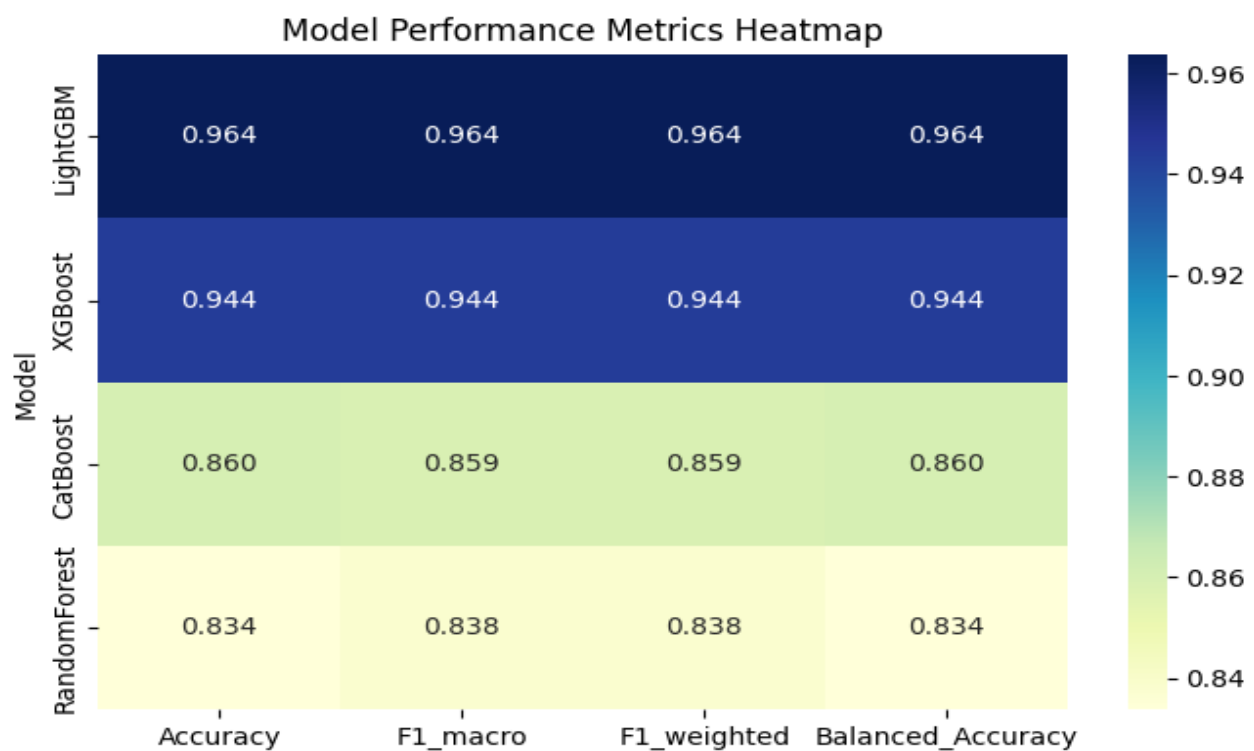
- Random Forest
- XGBoost
- LightGBM

4.3 Evaluation Metrics

- Accuracy
- F1-score
- Confusion matrices



	Accuracy	F1_macro	F1_weighted	Balanced_Accuracy
LightGBM	0.963710	0.963644	0.963651	0.963700
XGBoost	0.944355	0.944233	0.944244	0.944333
CatBoost	0.859677	0.859064	0.859085	0.859656
RandomForest	0.833871	0.838243	0.838213	0.833737



5. Notebook C – Unsupervised Learning (Clustering)

Objective

- Identify structural patterns in assistance distribution across regions.
- Detect **underserved regions** where socio-economic vulnerability is not aligned with aid coverage.
- Apply **unsupervised learning** to avoid predefined assumptions.
- Use **need_score strictly for interpretation**, not for clustering (to prevent circular reasoning and bias).

Feature Selection Strategy

- **Clustering features (inputs):**
 - Coverage_rate
 - Income
 - Poverty_rate
 - **Interpretation-only features (excluded from clustering):**
 - need_score
 - Benefit_per_capita
 - **Rationale:**
 - Focus clustering on **observable structural indicators**
 - Preserve *need_score* as an **external validation metric**
-

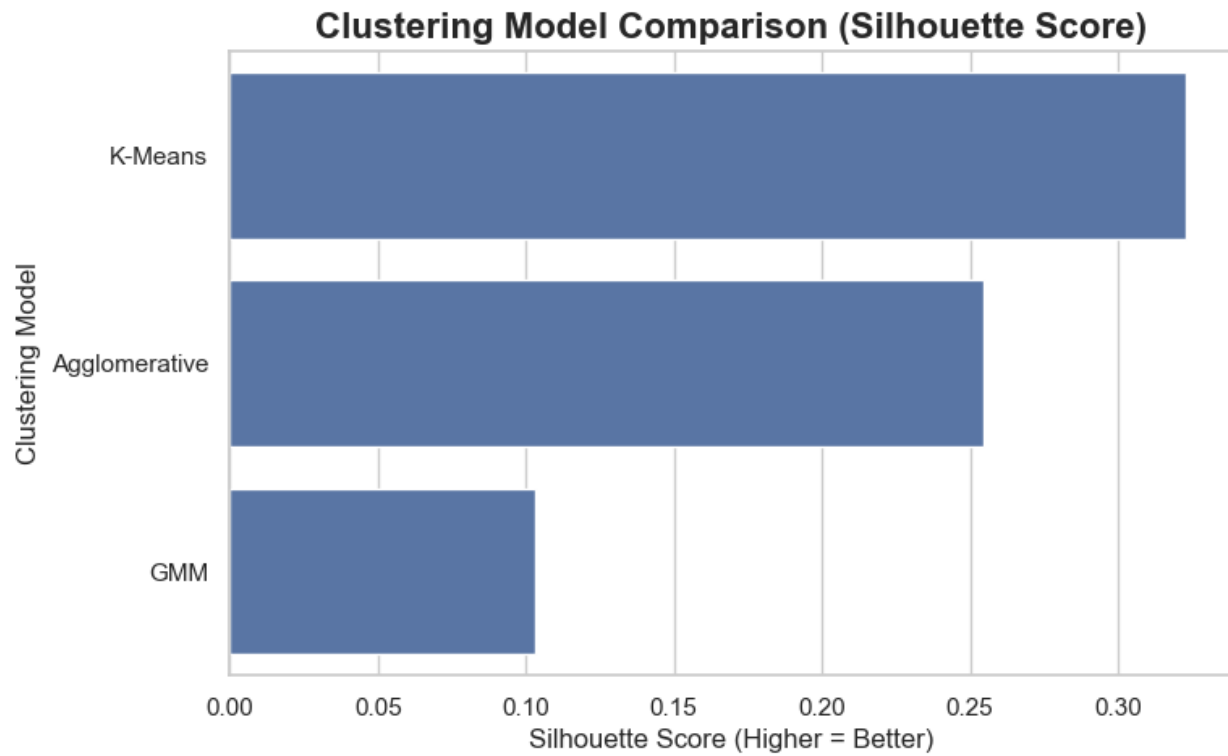
Data Preprocessing

- Applied **StandardScaler**:
 - Ensures equal contribution of features
 - Prevents dominance of income scale
 - Dimensionality reduction using **PCA**:
 - Retained **95% of explained variance**
 - Reduced noise and multicollinearity
 - Improved clustering stability and visualization
-

Clustering Methods Evaluated

- **K-Means**
 - Primary clustering method
 - Evaluated for $k = 2$ to 8
- **Agglomerative Clustering (Ward linkage)**

- Hierarchical validation
- **Gaussian Mixture Model (GMM)**
 - Soft clustering with probabilistic confidence



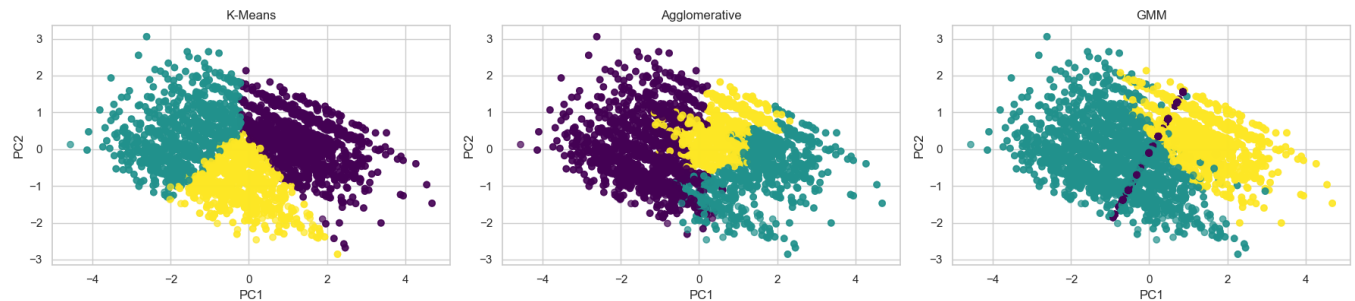
Model Selection and Validation

- Evaluation metrics:
 - **Silhouette Score** (cluster cohesion and separation)
 - **Davies–Bouldin Index** (cluster compactness)

	k	Silhouette	Davies_Bouldin
0	2	0.372715	1.040909
1	3	0.322883	1.090286
2	4	0.329603	1.004514
3	5	0.316501	1.060323
4	6	0.308580	1.061015
5	7	0.292722	1.109250

	k	Silhouette	Davies_Bouldin
6	8	0.281027	1.098036

Comparison of Clustering Models in PCA Space



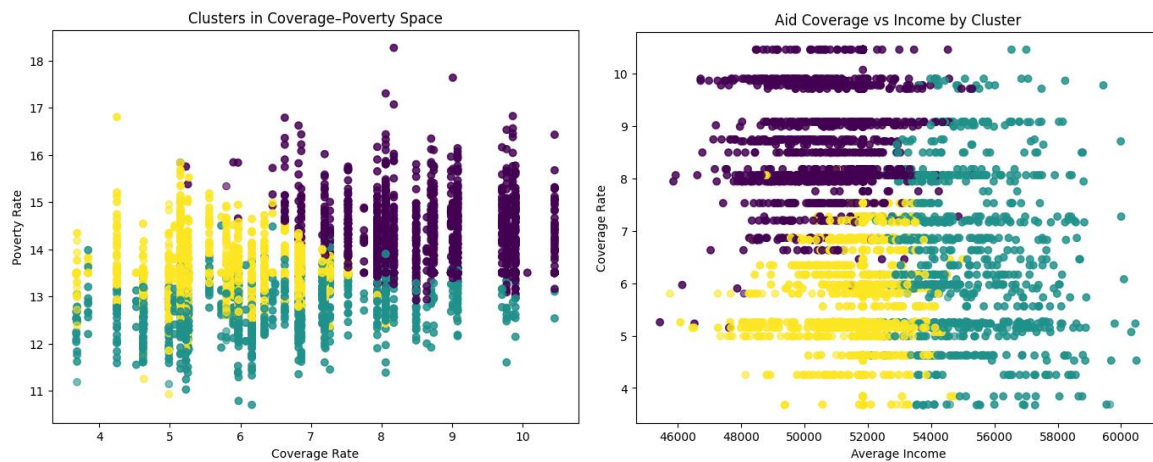
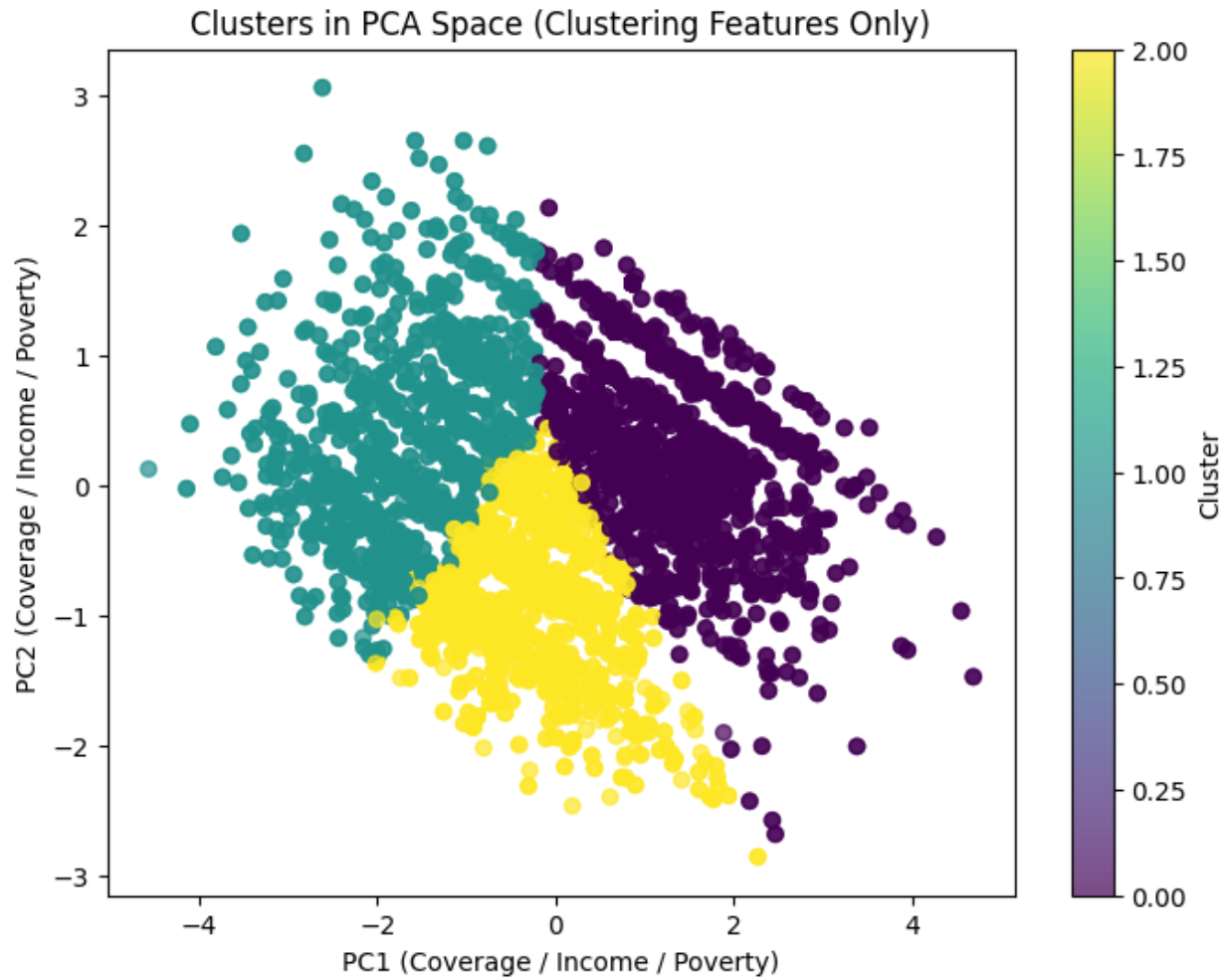
- Results:
 - $k = 3$ provided the best balance between:
 - Interpretability
 - Cluster separation
 - Policy relevance
- K-Means selected as the **reference clustering model**.

Cluster Profiling (K-Means)

Average characteristics per cluster:

- Coverage rate
- Income level
- Poverty rate
- Benefit per capita
- Need score (interpretation only)

This profiling enabled a **policy-oriented interpretation** of each cluster.



Cluster Interpretation

Cluster 0 — Low Need / Adequately Supported

- Characteristics:
 - Moderate to high coverage

- Average income
 - Moderate poverty
 - Benefits aligned with need
 - Interpretation:
 - Assistance distribution is broadly consistent with socio-economic conditions
 - Policy implication:
 - Maintain current support
 - Monitor trends over time
-

Cluster 1 — Medium Need / Underserved (Priority)

- Characteristics:
 - Low coverage rate
 - Low income
 - High poverty
 - Lower benefit per capita
 - Elevated need score
 - Interpretation:
 - Structural mismatch between vulnerability and aid allocation
 - Policy implication:
 - **Top priority for intervention**
 - Reallocate or increase assistance coverage
-

Cluster 2 — High Support / Well-Supported

- Characteristics:
 - High coverage rate
 - Higher income
 - Lower poverty
 - Higher benefit per capita
 - Lower need score
 - Interpretation:
 - Aid reaches these regions effectively
 - Policy implication:
 - Continue support
 - Potential resource rebalancing toward underserved regions
-

Visualization Insights

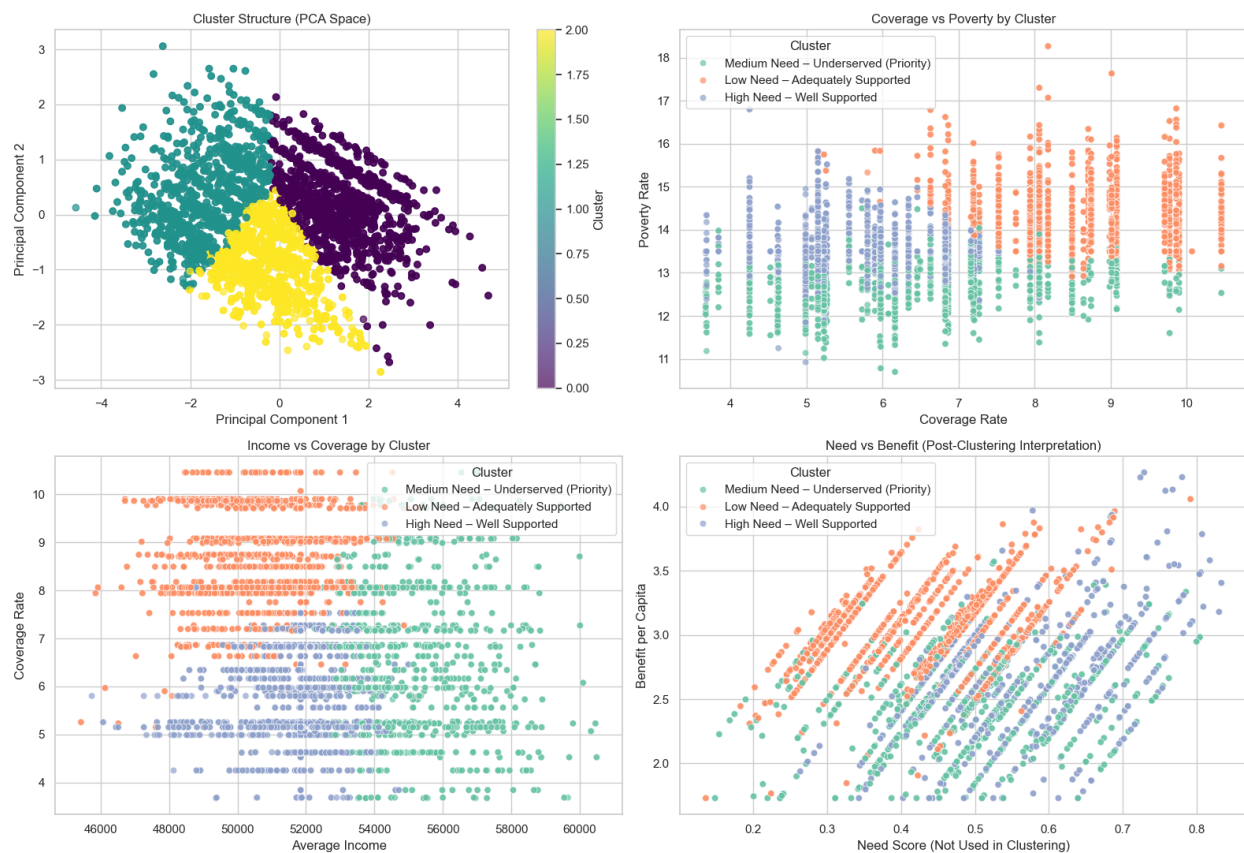
- Coverage vs Poverty plots:
 - Clear separation of vulnerable underserved regions
- PCA cluster maps:

- Strong cluster structure using assistance and socio-economic indicators only
- Need vs Benefit projection:
 - Confirms that **underserved clusters have high need but low benefit**
 - Validates exclusion of need score from clustering

Key Findings

- One cluster consistently shows:
 - High poverty
 - Low income
 - Insufficient coverage
- This group represents the **highest-impact policy target**.

Unsupervised Clustering Summary — Aid Distribution & Socioeconomic Context



This figure synthesizes the clustering results by jointly visualizing cluster structure, socio-economic vulnerability, aid coverage, and unmet need. It highlights a distinct group of regions with high poverty and need but insufficient coverage, supporting targeted policy intervention.

- Unsupervised clustering successfully identified **hidden structural inequities** in aid allocation.
- The methodology provides:

- Data-driven prioritization
- Transparent decision support
- Strong alignment with policy use cases
- Results support targeted reallocation strategies to improve equity and effectiveness of assistance programs.

Final CONCLUSION :

In conclusion, this US Food Assistance project, leveraging the power of Machine Learning has successfully transformed complex food security data into clear, actionable insights.

We've moved beyond surface-level analysis to **pinpoint *who* needs assistance and *where***, paving the way for smarter, more targeted policy decisions. The goal of delivering resources to those most in need is now closer than ever.