Unpacking Socioeconomic Inequality: Clustering Filipino Households Based on Income and Expenditure Pattern

Blix D. Foryasen
College of Computing and Information Technologies
National University – Philippines
Manila, Philippines
foryasenbd@students.national-u.edu.ph

Jerico C. Lim
College of Computing and Information Technologies
National University – Philippines
Manila, Philippines
limjc@students.national-u.edu.ph

Jessy Cassandra M. Mapanao
College of Computing and Information Technologies
National University – Philippines
Manila, Philippines
mapanaojm@students.national-u.edu.ph

***Abstract*—This study analyzes socioeconomic disparities among Filipino households using K-Means clustering on PCA-transformed data. Three clusters emerged: (1) low-income urban households, (2) entrepreneurial and agricultural households, and (3) financially stable working households. K-Means was the most effective model, validated using the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores. The findings inform financial literacy programs, credit access, and policy interventions to promote financial inclusion.**

***Index Terms*—Socioeconomic disparity, K-Means clustering, PCA, income and expenditure patterns, Filipino households**

I. Introduction

Income inequality has remained a persistent issue in the Philippines despite a declining GINI Index from 47.7% in 2000 to 42.3% in 2018. Various household characteristics–such as their heads' education level, occupations, employment sectors, and geographic location–are the most significant contributors to income disparity [1]. While poverty reduction efforts, including increased employment growth outside of agriculture, have lowered the poverty rate from 49.2% in 1985 to 16.7% in 2018, gaps remain stark. The top 1% of earners constitute 17% of the national income, while the bottom 50% capture only 14% [1]. Thus, the Philippines has one of the highest rates of income inequality in East Asia, with a GINI Index of 42.3 in 2018 [2].

Addressing these gaps requires understanding how Filipino households allocate their income between essential and non-essential expenditures, savings, and debt. This study aims to analyze Filipino household income and expenditure behaviors using clustering algorithms to identify distinct financial behavior groups. By segmenting households based on their spending priorities, household head characteristics, income levels, and savings and debt behavior, the research provides valuable insights to inform targeted financial and policy interventions to enhance economic resilience and financial inclusion. Policymakers and financial institutions can utilize the findings to develop data-driven policies and financial strategies tailored to specific household clusters, fostering long-term economic stability in the Philippines.

II. Review of Related Literature

Filipino household finance is shaped by socioeconomic, behavioral, and macroeconomic factors, as evidenced by various studies. In the Philippines, data sources such as the Family Income and Expenditure Survey (FIES) offer valuable insights into household income, expenditures, savings, and debt behavior [3]. In contrast, the CFS, conducted by the Bangko Sentral ng Pilipinas (BSP), places greater emphasis on household assets, liabilities, savings, and borrowing behavior, capturing a more comprehensive picture of financial health [4]. However, discrepancies between survey data and national accounts may distort poverty and inequality estimates, highlighting the need for improved data collection protocols [5]. Moreover, persistent income inequality, particularly between younger and older households, further compounds these challenges [6]. Rising out-of-pocket health expenditures place additional financial strain on vulnerable families, particularly in regions like Mindanao [7]. Behavioral factors such as financial literacy, emotions, and materialism significantly contribute to over-indebtedness, while family presence and entrepreneurial income positively influence savings and perceived financial performance [8; 9; 10]. These findings underscore the importance of enhanced financial literacy programs, targeted safety nets, and policy reforms to foster inclusive financial stability.

Household consumption patterns evolve alongside economic growth, as wealthier households diversify their spending across a wider array of goods and services, leading to increasingly niche consumption preferences, particularly for luxury items, while spending on necessities remains relatively consistent across countries [23]. This diversification process is closely linked to Engel's Law, which states that as income rises, the proportion of income spent on food decreases, allowing more resources to be allocated to non-food essentials and discretionary goods, although the COVID-19 pandemic highlighted the vulnerability of low-income households to economic shocks and the critical role of adaptive social protection systems in mitigating poverty risks and supporting upward mobility [24; 25].

In the Philippine context, classifying household socioeconomic status poses unique challenges due to evolving digital inclusion trends and cultural influences [11]. Naviamos and Niguidula (2020) successfully applied SVM to classify households into poor and non-poor categories, demonstrating the potential of machine learning in improving poverty identification [12]. Internationally, wealth indices serve as alternative measures to income and consumption data for assessing socioeconomic status, which can be particularly useful for global comparisons [13].

Clustering techniques play a crucial role in analyzing household data to uncover spending patterns, enabling policymakers, economists, and researchers to gain deeper insights into socioeconomic behavior. Various clustering algorithms, including K-means, hierarchical clustering, and DBSCAN, offer distinct advantages when applied to socioeconomic datasets. Studies have shown that robust deep K-means models improve clustering accuracy in high-dimensional

data, DBSCAN is effective at identifying non-convex clusters and handling noise, and hierarchical clustering methods offer multi-resolution results and reveal cluster organization, though they can be time-consuming or inaccurate [14; 15; 16]. Meanwhile, utilizing the Gaussian Mixture Models (GMMs) can also further enhance clustering analysis by providing probabilistic cluster memberships, accommodating clusters of varying sizes and shapes, and demonstrating greater flexibility across diverse datasets, particularly when data scaling varies [17; 18]

Each algorithm's performance varies depending on data complexity, structure, and dimensionality, emphasizing the importance of algorithm selection in household expenditure research. By exploring clustering techniques with rich household data, such as the Family Income and Expenditure Survey (FIES) from the Philippine Statistics Authority, it can help develop comprehensive socioeconomic profiles to inform data-driven policy decisions.

III. METHODOLOGY

The project investigates income and expenditure patterns for segmenting Filipino households into clusters with distinguishing characteristics. Unsupervised machine learning techniques, namely k-means, agglomerative clustering, and Gaussian Mixture Modeling, aided in determining unique clusters of Filipino households. Beforehand, the data went through a feature engineering pipeline to preprocess the data for clustering. Each process is outlined and expounded as follows:

A. Data Collection

The dataset was uploaded by a user named Francis Paul Flores in 2018 at Kaggle, an online data science platform and community for data scientists and machine learning practitioners. The data involved a sample from the 2018 Family Income and Expenditure Survey (FIES), spearheaded by the Philippine Statistics Authority. The survey is undertaken every three years through two household visits, once every 6 months, to provide data on household income and expenditure. Surveyors collected how a household allocates funds on various expenditure items, sources of income, asset ownership, household composition, house characteristics, living conditions of households, household head demographics, and related information affecting income and expenditure levels and patterns of households from different regions in the Philippines.

The dataset contains 41544 rows and 60 selected features from the 2018 FIES, which covered about 180,000 households. Each row represents one household from one of the 17 regions in the Philippines, described through individual expenditure allocations, total household income and sources of income, number of assets per type, household living conditions, family compositions, and household head characteristics.

B. Data Pre-processing

To ensure data consistency, quality, and readability, several data cleaning, transformation, normalization, and feature extraction techniques were employed to analyze household characteristics, ensuring the data was interpretable for training a clustering model and clustering analysis. With the dataset containing selected features from the FIES data lacking additional context that may aid deeper analysis and insights, the pre-processing pipeline is outlined as follows:

Initial Sanity Check

The dataset contained 7536 rows of null values for both household head occupation and class of worker. Additionally, 1580 rows of null values were detected for toilet facilities. Given the significant amount of null values and all the features are categorical, such null entries were converted to "Unknown."

Exploratory Data Analysis

Histograms and boxplots proved vital in visualizing the distribution of values for numerical features, while count plots offer if data is well distributed on a certain categorical feature. Every individual numerical and categorical feature was iterated to verify skewness and uniformity.

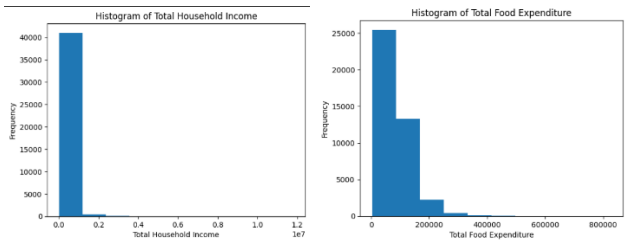


Fig. 1. Histogram of Total Household Income and Total Food Expenditure Indicating Positive Skewness

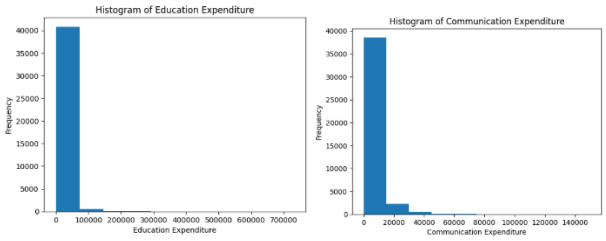


Fig. 2. Histogram of Education and Communication Expenditure Indicating Zero-Inflated Distribution

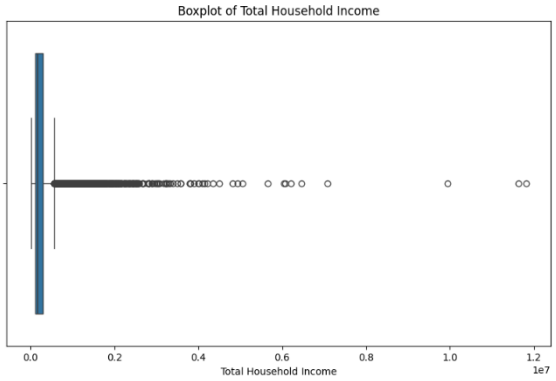


Fig. 3. Boxplot of Total Household Income Showing Extreme Values

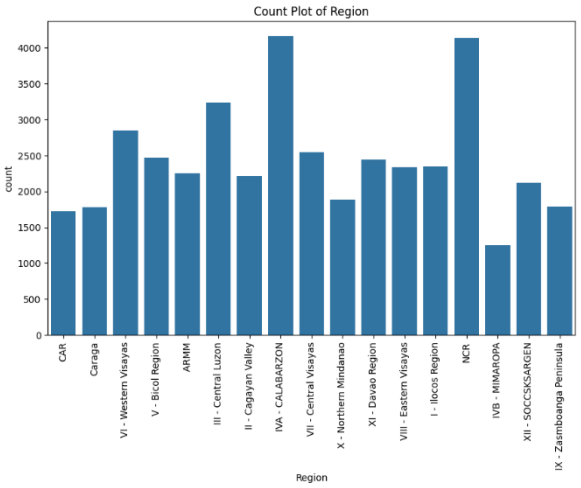


Fig. 4. Count plot of Regions shows lacking representations of households per region

Upon review of the histograms of all numerical columns, numerical features show positive skewness and are zero-inflated. Thus, there exists a need for outlier handling to reduce extreme values that misrepresent the population. Transformation and standardization should be applied to handle positive skewness and zero-inflated distribution. Also, households from regions lacking representation may not adequately represent the income and spending behavior of households from their area.

Feature Creation

$$Non - EssentialSpending = \sum Non - EssentialExpenditureItems$$

Formula 1. Summation Formula for Non-Essential Expenditures

$$EssentialSpending = \sum EssentialExpenditureItems$$

Formula 2. Summation Formula for Essential Expenditures

$$TotalExpenditure = EssentialSpending + Non - EssentialSpending$$

Formula 3. Summation Formula for Total Expenditures

New Features were derived from the original features to capture more relevant information. Non-essential spending and essential spending provide nuances in household spending behaviors. Restaurant and hotels, alcoholic beverages, education, transportation expenditure, tobacco, clothing, footwear and other wear expenditure, medical care expenditure, communication expenditure, miscellaneous goods and services expenditure, special occasions expenditure, and crop farming and gardening expenses fall under non-essential spending. In contrast, total food, housing and water, medical care, and communication expenditures cover essential spending [19]. Essential and non-essential spending accounts for the household's total expenditures.

$$Debt/SavingsAmount = TotalIncome + TotalExpenditure$$

Formula 4. Debt-Savings-Income-Expenditure Balance Formula

The dataset does not provide data on debts or savings incurred by each household. Such information can be collected by taking the difference between the total household income and the newly created total expenditure. A positive difference indicates the amount the household saved, while a negative difference shows debts incurred to cover overspending.

$$FoodItemRatio = \frac{FoodItemExpenditure}{TotalFoodExpenditure}$$

Formula 5. Food Item Expenditure Ratio Formula

$$SpecificExpenditureItemRatio = \frac{SpecificExpenditure}{TotalExpenditure}$$

Formula 6. Formula of Proportion of Specific Expenditure to Total Expenditure

$$DebtToIncomeRatio = \frac{DebtAmount}{TotalIncome}$$

Formula 7. Debt-to-Income Ratio Formula

$$ExpenditureToIncomeRatio = \frac{TotalExpenditure}{TotalIncome}$$

Formula 8. Income Utilization Ratio Formula

$$EntrepreneurialIncomeRatio = \frac{AmountFromEntrepreneurialActivities}{TotalIncome}$$

Formula 9. Entrepreneurial Income Proportion Formula

Total food expenditures encompass all individual expenditures, specifically for meat, vegetables, fish and marine products, bread and cereal, rice, fruits, and vegetables. To understand the specific financial dietary behavior of each household, individual food ratios to total food expenditure were computed. Similar to total food expenditure, ratios to the total household expenditure for food, housing and water, alcoholic beverages, tobacco, clothing and footwear, miscellaneous goods and services, special occasions, restaurant and hotels, crop farming and gardening, medical care, education, communication, transportation, and imputed house rental, were computed. Debt and Savings relative to Income indicate financial stress and surplus, respectively, while expenditure to income ratio captures the overall financial behavior of one household. The entrepreneurial income ratio indicates how much of the total income comes from business-related activities. Ratios normalize financial behaviors across households, allowing for a more meaningful comparison independent of absolute income and expenditure levels.

$$PerCapitaIncome = \frac{TotalIncome}{NumberOfFamilyMembers}$$

Formula 10. Per Capita Income Formula

$$PerCapitaExpenditure = \frac{TotalIncome}{NumberOfFamiluyMembers}$$

Formula 11. Per Capita Expenditure Formula

A household with ₱50,000 income and 10 members may not exhibit the same financial behaviors as a household earning the same amount but with two members. To uncover disparities in the cost of living and economic stress, per capita income and per capita expenditure were computed, allowing normalized total household income dependent on family size and how much each household member contributes to or relies on spending.

To add more information about the number of unemployed, the total number of employed members is subtracted from the total number of family members. Additionally, members under 5 years and 5 to 17 years old were grouped into one to indicate the number of kids within the household.

Categorical Feature Transformation

Table 1. Number of Household Heads under a Specific Major Occupation Group

Major Occupation Group	Count
Skilled agricultural, forestry and fishery workers	9042
Uncategorized	7536
Elementary occupations	7177
Managers	6963
Service and sales workers	3555
Craft and related trades workers	3085
Plant and machine operators and assemblers	2172
Professionals	971
Clerical Support workers	564
Technicians and associate professional	391
Armed forces occupations	88

Household Head Occupation contains 378 unique values that, when encoded, may confuse model training. To simplify the number of unique values, specific occupations were mapped according to the 2012 Philippine Standard Occupation Classification (PSOC), which the PSA uses to classify occupations into different occupational groups of the working population [20]. Specific occupations fall under one of 10 PSOC major groups, while the uncategorized group contains the handled missing null values.

Household Head Highest Grade Completed consists of 46 unique values that could be classified based on their highest educational attainment. Thus, each unique value is mapped into preschool, elementary non-graduate, elementary graduate, high school non-graduate, high school graduate, college non-graduate, college graduate, or post-baccalaureate, further ordinal encoded from 0 to 7, respectively.

To gain a grasp of the type of employment and the source of funds for household heads, worker classes were simplified into Formal, Informal, Entrepreneurial, and Unpaid Employment. A household head is considered in Formal Employment if they work for the government, government corporation, or a private establishment. Informal Employment comes from workers in private households or working in a family-operated farm or business. Self-employed and Employer household heads are considered under entrepreneurial. Workers without pay from any field, even in a family-operated business or farm, are considered Unpaid Workers.

Outlier Detection and Removal

As observed from EDA, there exist households with extreme household incomes that do not represent the entirety of the population. Interquartile Range (IQR) and Isolation Forest were used to handle and detect outliers and anomalies.

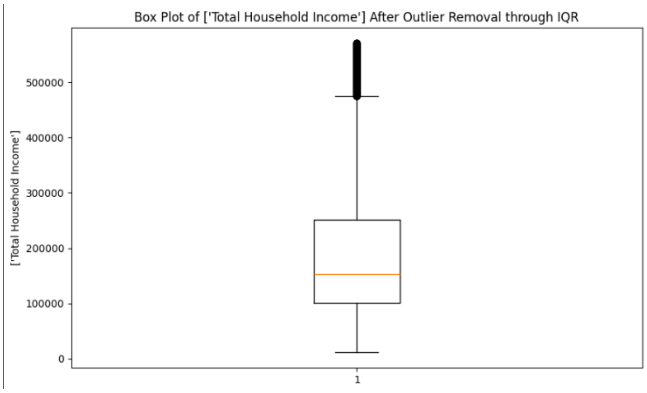


Fig. 5. Reduction of Outliers through IQR

Interquartile Range limited total household income within the bounds of the 25th Percentile and 3rd Percentile with an outlier step of 1.5. Any value beyond the bounds was considered an outlier and omitted from the dataset. After removing 3151 outliers, the household income ranged from 11,285 to 570,492 pesos with a mean income of 189885.70 pesos for all 38393 non-outlier households. The distribution of values shows a considerable improvement from the original dataset in limiting extreme values.

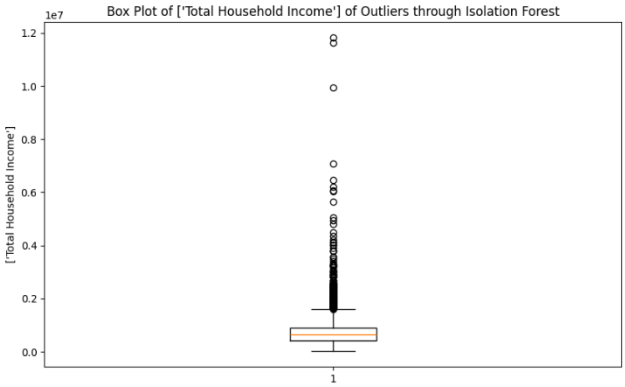


Fig. 6. Reduction of Outliers through Isolation Forest

Isolation Forest isolates outliers by checking all dataset features, unlike IQR. Isolation forest assumes that outliers are data points isolated when plotted into decision trees. After removing 4155 outliers, the household income ranged from 11,285 to 2,317,700 pesos, with a mean income of 191324.30 pesos for all 37389 non-outlier households. Visual analysis of the boxplot indicates the presence of extreme income values from the extracted non-outliers. Given Isolation Forest's multivariate approach of outlier detection, specific households were omitted as they are outliers in some of the features present in the dataset.

To compare the results of the outlier detection algorithms implemented, they are retained as separate dataframes alongside the original dataframe to be passed into various feature selection algorithms and clustering models.

Numerical Feature Transformation and Normalization

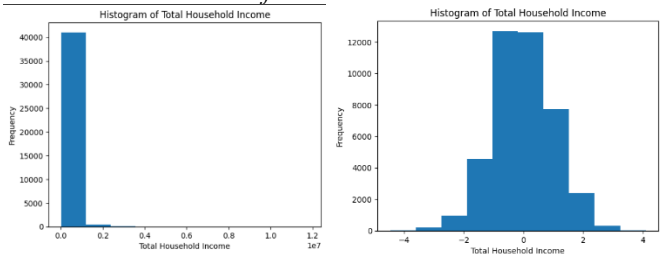


Fig. 7. Before and After Transformation and Normalization of Total Household Income

Through EDA, most numerical features had positive skewness and were zero-inflated. This indicates that most households often lie in one specific income range or do not spend much on a particular expenditure item. Yeo-Johnson Power Transformation aids in transforming right-skewed data to a normal distribution. After transformation, numerical data is standardized to ensure comparability between values.

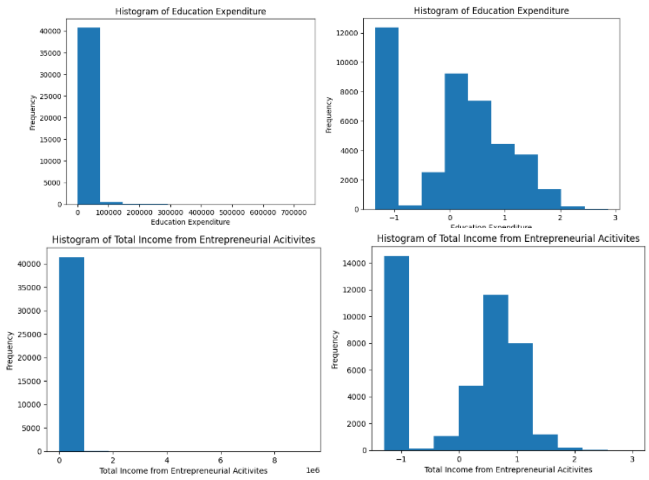


Fig. 8. Education Expenditure and Total Income From Entrepreneurial activities remains zero-inflated even after Transformation and Normalization

Feature Selection

Given the focus on spending behaviors and patterns of Filipino Households, features indicating asset ownership (e.g., number of cars, number of refrigerators, number of airconditioners), living situation (e.g., Main Source of Water Supply, Electricity), and house characteristics (e.g., Type of Roof, Type of Wall) are excluded among the features. The absolute spending power of a household for each expenditure item is omitted, while the ratios and the aggregates on spending and income are kept. Households with higher income naturally have better absolute spending capability than those of low income. A high-income family may spend 50,000 on food, while a low-income family spends 5,000. The absolute spending difference is substantial, but if both allocated 20% of their income to food, they share the same spending behavior. Moreover, keeping the individual expenditure items alongside income and expenditure aggregates promotes multicollinearity, something ratios help to improve model interpretability.

Three feature selection techniques were employed to identify the most informative variables. Recursive feature elimination with cross-validation (RFEVC) was applied to handle multicollinearity through Ridge Regression with a minimum selection threshold of five features and a cross-validation of 5. The final number of selected features varied across datasets, resulting in 45 features for the original dataset, 52 for the IQR-filtered dataset, and 44 for the isolation forest-filtered dataset. Mutual information (MI) was used to select features based on information gain, with the number of features determined by the corresponding RFEVC outputs. Furthermore, a random forest regressor (RFR) was trained using GridSearchCV with a cross-validation of five and a specified hyperparameter grid.

```
param_grid = {
    'n_estimators': [100, 200, 500],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2],
}

grid_search = GridSearchCV(
    estimator=model_rf,
    param_grid=param_grid,
    cv=5,  # 5-fold cross-validation
    n_jobs=-1,
    verbose=2
)
```

Fig. 9. Parameter Grid for Random Forest Regressor

The model, trained with total household income as the target variable and a Root Mean Squared Error (RMSE) of 0.008, identified total expenditure, remaining income, debt-to-income ratio, and expenditure-to-income ratio as the most important features across all datasets.

Table 2. Datasets for PCA and Model Training

Index	Preprocessing Step Taken
0	Original Dataset
1	Removed Outliers through IQR
2	Removed Outliers through Isolation Forest
3	Original Dataset with RFEVC Features
4	Original Dataset with Random Forest Features
5	Original Datae with Mutual Information Features
6	IQR Dataset with RFEVC Features
7	IQR Dataset with Random Forest Features
8	IQR Dataset with Mutual Information Features
9	Isolation Forest Dataset with RFEVC Features
10	Isolation Forest Dataset with Random Forest Features
11	Isolation Forest Dataset with Mutual Information Features
12	Original Dataset with Combined RFEVC and Mutual Information Features
13	IQR Dataset with Combined RFEVC and Mutual Information Features

Each feature selection method was applied to the three dataset variants, creating multiple datasets. Combined variants involve common features found between MI and RFEVC, which returned 45, 52, and 46 features, respectively.

Dimensionality Reduction

Principal component analysis (PCA) was applied to each dataset to facilitate visualization and enhance clustering performance, generating additional variants where the transformed data was at least 90% retained. Subsequently, datasets successfully projected into 2D or 3D spaces were used for clustering model training, ensuring that the final clustering models operated on datasets that retained meaningful variance while reducing dimensional noise.

C. *Experimental Setup*

Google Colab was used as the primary IDE for the end-to-end methodology centered around Python-based data science and machine learning frameworks, primarily leveraging scikit-learn (v1.6.1) for clustering models, feature selection, and evaluation metrics. Data preprocessing and transformation

were handled using NumPy (v1.26.4) and Pandas (v2.2.2), while visualization was supported by Matplotlib (v3.10.0) and Seaborn (v0.13.2). Additional statistical computations and hierarchical clustering analyses were performed using SciPy (v1.13.1). Visual Studio Code served as a backup for running local hyperparameter optimization. To match the libraries used in Google Colab, a new Python environment containing the primary libraries was initialized.

D. Algorithm

K-Means Clustering

K-means clustering partitions data into k distinct, non-overlapping clusters where each data point is specifically assigned to the cluster with the nearest centroid. Through iteration, points are assigned to the nearest centroid, which is iteratively updated based on the mean of points to minimize the within-cluster variance. K-means excels at discovering simple, spherical clusters and is computationally efficient, making it suitable for large datasets. However, it requires predefining the number of clusters and is sensitive to outliers and the initial placement of centroids.

$$J(v) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left((\|x_i - v_j\|)^2 \right)$$

Formula 12. K-Means Clustering Formula [21]

Where:

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Agglomerative Clustering

Agglomerative Clustering is an unsupervised hierarchical clustering algorithm that builds a hierarchy of clusters from the bottom up. Each data point is treated as a single cluster and is successively merged until all points belong to a single cluster based on a linkage criterion. A hierarchical representation of the data through dendrograms makes Agglomerative Clustering efficient, especially when the number of clusters remains unknown, offering flexibility in exploring different cluster granularities. However, larger datasets could make the algorithm computationally expensive and sensitive to noise and outliers.

$$d_{12} = min_{i,j}\ d(X_i, Y_j)$$

Formula 13. Single Linkage Formula [22]

This refers to the distance between the nearest points of the two clusters.

$$d_{12} = max_{i,j}\ d(X_i, Y_j)$$

Formula 14. Complete Linkage Formula [22]

This refers to the distance between the members that are the farthest from each other (the most dissimilar) in the two clusters.

$$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l}\ d(X_i, Y_j)$$

Formula 15. Average Linkage Formula [22]

This method calculates the distances between all pairs of points across two clusters and takes the average of these distances. It is also known as UPGMA — Unweighted Pair Group Method with Arithmetic Mean.

$$d_{12} =\ d(\bar{x}, \bar{y})$$

Formula 16. Centroid Method Formula [22]

This entails determining the mean vector location for each cluster and measuring the distance between their centroids.

Gaussian Mixture Models

Gaussian Mixture Models (GMMs) assume data points are generated from a mixture of several Gaussian distributions, each representing a cluster. GMMs estimate the mean, covariance, and mixing coefficients using the Expectation-Maximization (EM) algorithm, allowing for soft assignments of data points to clusters based on their probabilities. Thus, this method allows capturing complex, non-spherical cluster shapes through a probabilistic framework for clustering, offering flexibility in modeling data distributions. However, it can be sensitive to initialization, which requires careful selection of the number of components.

E. Training Procedure

The training procedure involved several steps to ensure the robustness of the clustering models. First, feature selection was conducted, where three different techniques—RFEVC, Mutual Information, and Random Forest Regressor—were applied to the datasets. The selected features were then used to construct multiple versions of the dataset, each undergoing standardization and transformation before clustering analysis. PCA was applied to facilitate visualization and reduce dimensionality while preserving variance.

For hyperparameter tuning, K-Means Clustering relied on WCSS/SSE plots to determine the optimal K, while Agglomerative Clustering leveraged dendrogram analysis. GMM's components were optimized using AIC and BIC scores through hyperparameter tuning through a parameter grid covering 1 to 7 n-components and full, spherical, tied, and diag covariance types. The final clustering models were trained separately on each dataset variation, ensuring performance was evaluated across different preprocessing strategies, including outlier handling and dimensionality-reduced data.

F. Evaluation Metrics

The Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Score were employed to assess clustering quality. The Silhouette Score measured the cohesion and separation of clusters, evaluating how well individual data points fit within their assigned cluster. A higher silhouette score proves that clusters are well-separated.

The Calinski-Harabasz Index assessed the compactness of clusters relative to their separation, with higher values indicating better-defined clusters. It provides a global measure of compactness and separation, with a higher value indicating that data points are more spread out between clusters than within them.

The Davies-Bouldin Score provided an additional measure of cluster compactness and similarity, where lower values signified more distinct clusters. Such a metric ensures that clusters are not overly similar to one another.

Performance comparisons were conducted by establishing K-Means as the baseline model, given its well-defined clusters in PCA plots. Evaluation metric scores from Agglomerative Clustering and GMM were then benchmarked against K-Means to determine the most appropriate clustering technique for household segmentation. The optimal model was identified based on a combination of quantitative evaluation metrics and visual assessment of cluster separation in 2D and 3D PCA scatter plots.

G. *Comparison of Clustering Algorithms*

Three unsupervised learning algorithms—K-Means Clustering, Agglomerative Clustering, and Gaussian Mixture Models (GMM)—were utilized for clustering analysis. Each clustering model underwent hyperparameter tuning to optimize performance.

K-Means Clustering was chosen as the baseline model due to its efficiency and interpretability, particularly when paired with PCA visualizations. The optimal number of clusters K was determined through the Elbow Method with Within-Cluster Sum of Squares (WCSS/SSE).

Agglomerative Clustering, a hierarchical method, allowed insight into the nested structure of socioeconomic clusters. Through visual analysis, the number of clusters was inferred from dendrogram analysis, identifying significant hierarchical splits.

Gaussian Mixture Models (GMM) provided a probabilistic clustering approach, allowing households to have soft memberships across multiple socioeconomic segments. GMM's number of components was fine-tuned based on Bayesian Information Criterion (BIC) to balance model complexity and performance.

These algorithms were selected due to their complementary strengths in cluster identification. K-Means and Agglomerative Clustering provided interpretable and structured clusters, while GMM captured probabilistic cluster distributions, offering flexibility in analyzing household economic segments. Upon comparison of models, K-Means performed slightly better in clustering all three dataset treatments across all three metrics and significantly better in runtime and computational costs.

IV. RESULTS AND DISCUSSION

This section outlines the results of the clustering analysis implemented on the dataset. The features underwent various data treatments, as outlined in the methodology, before fitting into the clustering models.

A. *Dimensionality Reduction*

The dataset was subject to Principal Component Analysis to reduce dimensionality while preserving the important features. The explained variance ratio was analyzed using the cumulative variance to determine the best number of principal components.

Table 3. Comparison of Principal Component Explained Variance across chosen Dataframes

Dataframe	Principal Component	Explained Variance
Dataframe 4	PC1	66.62%
	PC2	25.09%
Dataframe 7	PC1	66.75%
	PC2	25.00%
Dataframe 10	PC1	67.55%
	PC2	25.24%

The results indicate that all dataframes applied by random forest regression can be reduced and visualized into two principal components. The first principal component (PC1) explains approximately 55%, 53%, and 55% of the variance, while the second principal component (PC2) explains approximately 37%, 39%, and 38% of the variance, respectively. The most contributing feature for PC1 in dataframes 4, 7, and 10, were remaining income, debt to income ratio, and remaining income, respectively. All dataframes share total expenditure as the most contributing feature for PC2. The first two principal components consistently capture a significant portion of the variance across different dataframes, highlighting their importance in reducing dimensionality while retaining most of the data's variability.

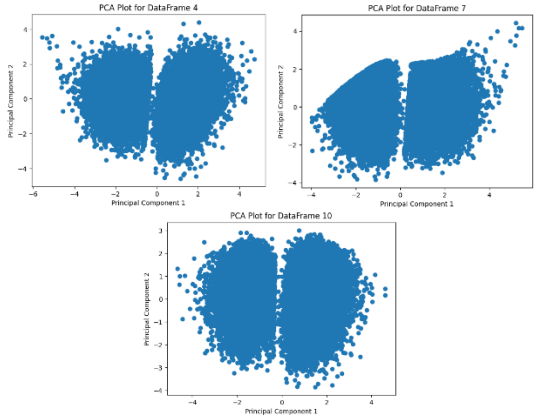


Fig. 10. PCA Plots of Chosen Dataframes

Through visual analysis, two clusters of data for clustering analysis are identified. Considering varying income levels of households from low-income to high-income, further analysis of sub-clusters within the two major clusters could provide specific characteristics from the two major clusters of households.

B. *Silhouette Plots*

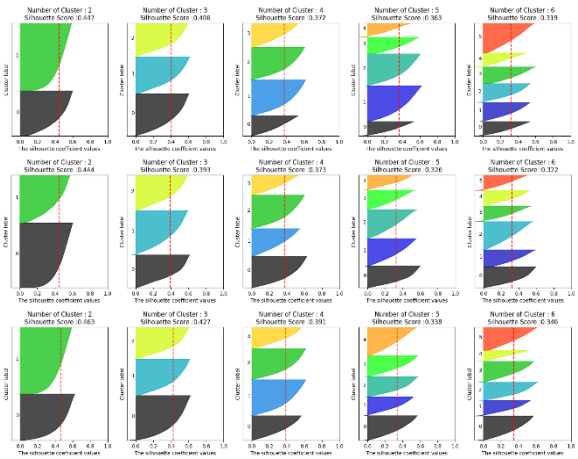
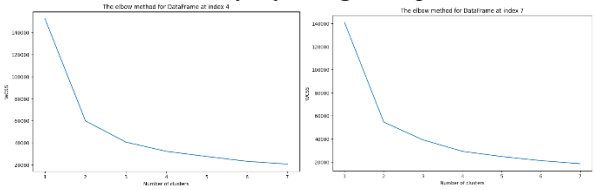


Fig. 11. Silhouette Plots of Chosen Dataframes

The chosen dataframes reveal that the optimal number of clusters is two, with Dataframe 10 achieving the highest silhouette score of 0.463, indicating well-defined clusters. Dataframes 4 and 7 also show relatively high silhouette scores for two clusters, 0.447 and 0.444 respectively, suggesting effective clustering for socioeconomic analysis.

C. *Elbow Method Plots for finding the Optimal K Value*



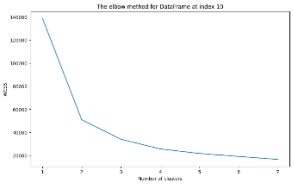


Fig. 12. Elbow Plots of the Chosen Dataframes

Visual analysis of the elbow plots from all dataframes show an elbow at 2 and 3 clusters, with 2 being the most prominent. The elbow plot coincides with the projected number of clusters in the PCA Plots of all three dataframes.

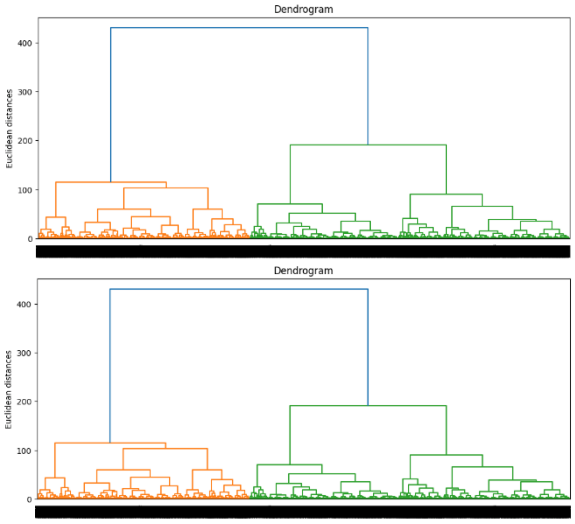
D. K Means Model

Table 4. K-Means Clustering Evaluation Results Across Dataframes

Dataframe	No. of Clusters	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
4	3	0.4386	0.9755	41327.1032
	4	0.3892	0.9962	51601.3132
	5	0.3332	1.0417	47010.5687
	6	0.3427	0.9513	46347.2546
7	3	0.4511	0.7726	49924.3615
	4	0.3919	0.9782	48834.6165
	5	0.3943	0.9088	45050.1700
	6	0.3448	0.9817	43458.8986
10	3	0.4858	0.7523	57635.8301
	4	0.4126	0.9102	55016.1963
	5	0.3871	0.9536	46997.1416
	6	0.3318	1.0167	46790.3755

The evaluation results of K-Means clustering across different dataframes show that for each dataframe, clustering into 3 clusters consistently achieved the highest Silhouette Score and the lowest Davies-Bouldin Score, indicating well-separated and compact clusters. This suggests that 3 clusters are likely the optimal number for segmenting the data across the evaluated dataframes.

E. Agglomerative Clustering Model Dendograms



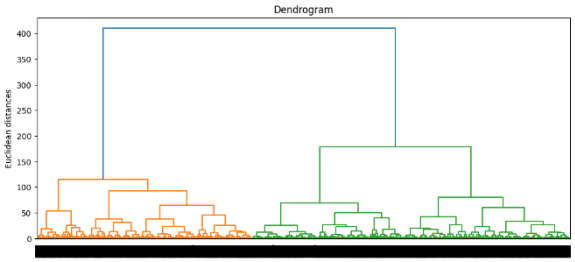


Fig. 13. Dendograms of the Chosen Dataframes

All three dendograms project three distinct vertical lines (branches) that reach relatively high on the Euclidean distance scale before merging. This indicates the presence of three major clusters within the data. The blue cluster merges with the orange and green clusters at a much higher distance than the orange and green merge with each other. This indicates that the blue cluster is significantly more dissimilar to the other two clusters than they are to each other. Green clusters show more distinct sub-clusters than the orange cluster. This finding further supports the exploration of a third cluster within the dataset.

F. Agglomerative Clustering Model

Table 5. Agglomerative Clustering Evaluation Results Across Dataframes

Dataframe	No. of Clusters	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
4	3	0.4470	0.7697	55450.1309
	4	0.3695	0.9380	46504.7423
	5	0.3720	0.9757	42983.4659
	6	0.3270	1.0324	41084.8095
7	3	0.4279	0.7948	47058.5817
	4	0.3604	1.0094	44347.9187
	5	0.3467	0.9686	39862.2908
	6	0.2857	1.0896	37463.9270
10	3	0.4547	0.7634	50420.1437
	4	0.3744	0.9911	44473.8352
	5	0.3739	0.9683	41495.2394
	6	0.2922	1.1032	38226.2154

The results of Agglomerative Clustering across dataframes show that 3 clusters generally achieved higher Silhouette Scores and relatively lower Davies-Bouldin Scores compared to other cluster counts, indicating better-defined clusters. This suggests that, similar to K-Means, 3 clusters may be the optimal configuration for partitioning the data across the evaluated dataframes.

G. Gaussian Mixture Model

Table 6. Gaussian Mixture Model Evaluation Results Across Dataframes

Dataframe	No. of Clusters	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
4	3	0.4824	0.8981	49535.8690
	4	0.3580	0.9386	39027.7944
	5	0.2956	1.0909	39293.8438
	6	0.3069	0.9937	38636.9717
7	3	0.4564	0.9491	41822.9293
	4	0.3853	0.9796	47714.1186
	5	0.3331	0.9859	40718.1100
	6	0.3253	1.0838	36807.3379

10	3	0.4824	0.7607	56442.6006
	4	0.4035	0.9305	52538.3154
	5	0.3773	0.9660	44057.3738
	6	0.3197	1.0558	42255.5286

The evaluation results for the Gaussian Mixture Model across dataframes show that 3 clusters consistently achieved the highest Silhouette Scores and the lowest Davies-Bouldin Scores, indicating better cluster separation and compactness. This implies that, across all tested dataframes, the GMM algorithm also suggests that 3 clusters are the optimal choice for segmenting the data.

H. Comparison of Models

Table 7. Comparison of Clustering Algorithms Across Dataframes Using Multiple Evaluation Metrics

Dataframe	No. of Clusters	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
Dataframe 4				
K-Means	3	0.4386	0.9755	41327.1032
Agglomerative Clustering	3	0.4470	0.7697	55450.1309
Gaussian Mixture Model	3	0.4824	0.8981	49535.8690
Dataframe 7				
K-Means	3	0.4511	0.7726	49924.3615
Agglomerative Clustering	3	0.4279	0.7948	47058.5817
Gaussian Mixture Model	3	0.4564	0.9491	41822.9293
Dataframe 10				
K-Means	3	0.4858	0.7523	57635.8301
Agglomerative Clustering	3	0.4547	0.7634	50420.1437
Gaussian Mixture Model	3	0.4824	0.7607	56442.6006

The comparison of clustering algorithms across dataframes shows that the Gaussian Mixture Model (GMM) consistently achieved the highest Silhouette Scores, indicating better-defined and more cohesive clusters. K-Means performed competitively, particularly in Dataframe 10, with the highest Calinski-Harabasz Score, reflecting good cluster separation. Overall, GMM appears to be the most effective algorithm across the dataframes, with K-Means as a strong alternative depending on the specific evaluation metric prioritized.

I. Cluster Profiling

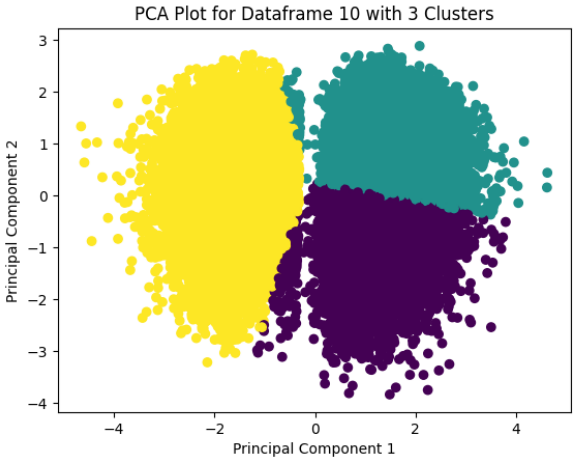


Fig. 14. Mapped Clusters through K-Means Clustering (k=3) in PCA Plot

Three clusters of family households have been clustered through K-Means. The clusters will be analyzed based on their family composition, household head characteristics, income sources, expenditure patterns, and savings and debt behavior to uncover disparities among Filipino Households and provide possible interventions for each cluster.

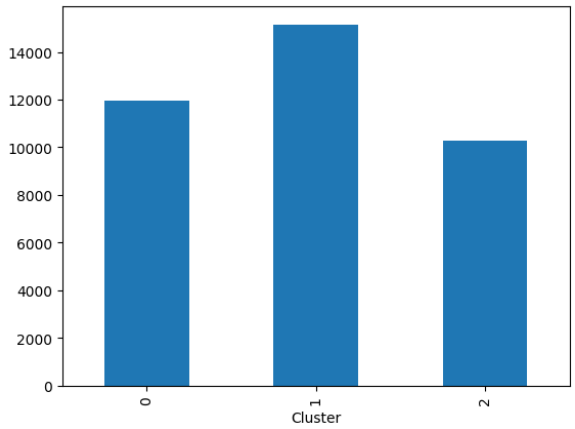
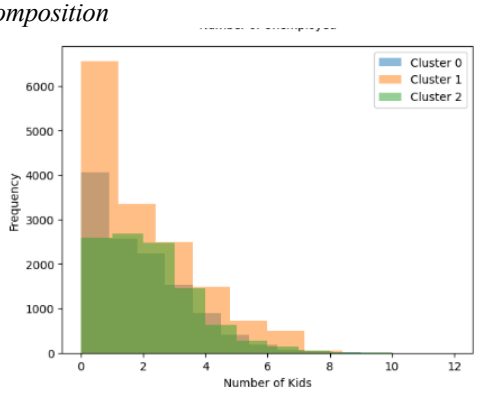


Fig. 15. Distribution of Households per cluster

Cluster 1 has the highest number of households, with 15148, while Clusters 0 and 2 consist of 11960 and 10281 households, respectively. Distinguishing characteristics are laid out as follows.

Family Composition



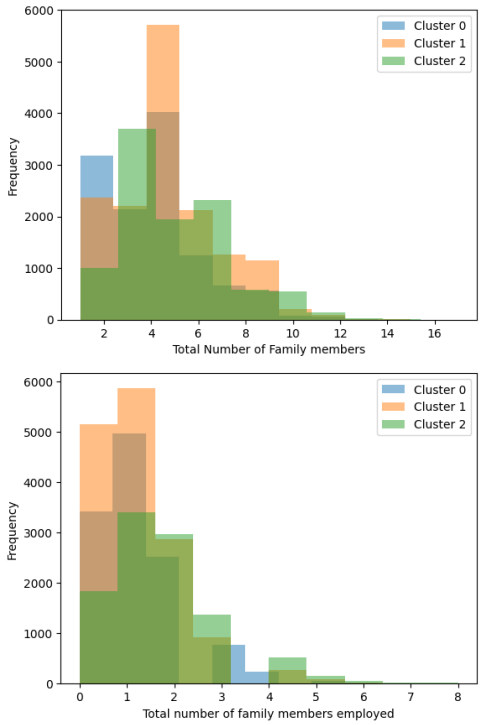


Fig. 16. Histograms of Family Composition Columns

Cluster 1 consists of large households with mostly kids unsuitable for the labor force, thus indicating high unemployment. Despite having the largest family size, households from Cluster 2 involve multiple working adults and have the highest employment among all the three clusters. Households from Cluster 0 are relatively small with moderate employment and could suggest economic stability.

Household Head Characteristics

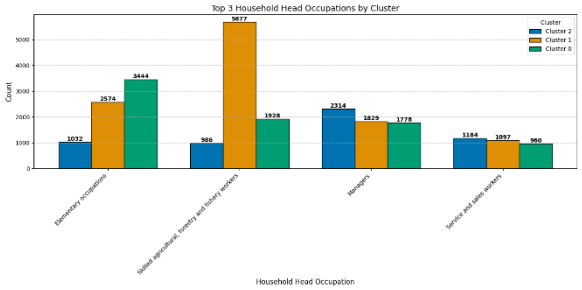


Fig. 17. Top three Major Occupation Groups of Clusters

Excluding uncategorized individuals, elementary occupations, skilled agricultural, forestry, and fishery workers, and managers appear in all three clusters as the most prominent occupation groups. To understand the differences between each major occupation group, it is essential to examine the specific job occupations of the household heads.

Table 8. Top three Jobs of each Clusters

Cluster	Occupation	Count
0	Farmhands and laborers	1817
	General managers/managing proprietors in transportation, storage and communications	654
	Inland and coastal waters fishermen	581
1	Rice Farmers	1976
	Farmhands and laborers	1350
	Corn farmers	1287
2	General managers/managing proprietors in wholesale and retail trade	710
	General managers/managing proprietors in transportation, storage and communications	614
	Car, taxi and van drivers	402

Farmers and laborers encompass Cluster 1, which aligns with the agricultural and elementary occupations group. Household heads from Cluster 0 thrive through foraging as fishermen, farmhands, or laborers. Household heads from Cluster 2 take managerial positions or work stable jobs in their respective fields.

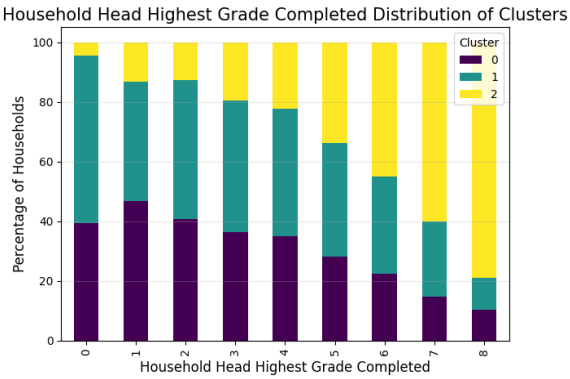


Fig. 18. Distribution of Household Head Highest Grade Completed

Education is often a strong indicator of socioeconomic status. As observed, most Household Heads from Clusters 0 and 1 do not proceed to college (6), possibly due to the nature of their jobs as farmers, fishermen, or laborers. Cluster 2 progressed into higher levels of education as compared to clusters 0 to 1, indicating socioeconomic disparities.

Fig. 19. Stacked Bar Chart of Household Head Class of Worker

Household Heads from Cluster 1 engage more in entrepreneurial activities or are self-employed, possibly in agribusiness, to sustain the needs of their families. Formal Employment is well spread out through all clusters, with Household Heads from Cluster 0 taking more Formal Employment jobs compared to Clusters 1 and 2 despite gaps in educational attainment.

Statistical Analysis and Significance Testing

Given that expenditure and income ratios are generally not distributed, the Kruskal-Wallis H test was conducted to assess whether significant differences exist between household clusters in expenditure and income patterns, followed by post hoc multiple testing corrections and effect size estimation. This test is a non-parametric alternative to the one-way ANOVA and is appropriate when comparing more than two independent groups—in this case, the identified three household clusters.

Kruskal-Wallis test was performed for each expenditure category to determine whether statistically significant differences existed among the clusters. The test produces an H-statistic, which quantifies the variance in rankings between groups, and an associated p-value, indicating whether these differences are unlikely chance. Moreover, a Bonferroni correction was applied to adjust for multiple comparisons. This method controls for the family-wise error rate by adjusting p-values, ensuring that statistically significant findings are not artifacts of conducting multiple hypothesis tests.

While p-values indicate statistical significance, they do not convey the magnitude of the differences observed. To provide a measure of practical significance, epsilon-squared was computed for each expenditure category, which quantifies the proportion of variance in the dependent variable (expenditure ratios) that can be explained by the independent variable (household cluster membership). Effect sizes were interpreted using conventional guidelines:

- ES < 0.02 - Very small
- 0.02 <= ES < 0.13 – Small
- 0.13 <= ES < 0.26 – Medium
- ES >= 0.26 - Large

The results of the Kruskal-Wallis tests indicate that expenditure patterns significantly differ among household clusters for most categories.

Table 9. P-values and epsilon-squared values of Kruskal-Wallis Tests

Feature	p-value	Epsilon-squared
Meat Ratio	0.000	0.098177
Fish Ratio	3.934e-241	0.029740
Fruit Ratio	5.985e-03	0.000402
Vegetables Ratio	5.984e-03	0.000402
Bread and Cereals Ratio	0.000	0.136722
Other Bread and Cereals Ratio	4.157e-36	0.004486
Food Ratio	0.000e+00	0.073267
Housing and water Ratio	0.000e+00	0.076148
Alcoholic Beverages Ratio	6.892e-08	0.001011
Tobacco Ratio	2.146e-50	0.006246
Clothing and Footwear Ratio	1.790e-253	0.031260
Miscellaneous Goods and Services Ratio	0.000e+00	0.112667
Special Occasions Ratio	0.000e+00	0.044799
Restaurant and hotels Ratio	0.000e+00	0.079056
Crop Farming and Gardening Ratio:	0.000e+00	0.150167
Medical Care Ratio	9.280e-218	0.026861
Education Ratio	1.182e-230	0.028449
Communication Ratio:	0.000e+00	0.149033
Transportation Ratio	0.000e+00	0.059308
House Rental Ratio	0.000e+00	0.047525
Expenditure to Income Ratio	0.000e+00	0.724697
Per Capita Income	0.000e+00	0.285684
Per Capita Expenditure	0.000e+00	0.203632
Essential Spending to Expenditure	0.000e+00	0.168645
NonEssential Spending to Expenditure	0.000e+00	0.168645
Debt to Income Ratio	0.000e+00	0.908197
Savings to Income Ratio	0.000e+00	0.763971
Total Household Income	0.000e+00	0.497661
Total Income from Entrepreneurial Acitivites	2.527e-153	0.018828
Entrepreneurship Income to Total Income Ratio:	0.000e+00	0.038328

Highly significant differences ($p < 0.001$, large effect sizes) were observed in Expenditure to Income Ratio, Per Capita Income, Total Household Income, Debt to Income Ratio, and Savings to Income Ratio. These large effect sizes suggest substantial disparities in how different clusters allocate their financial resources, particularly concerning expenditure, debt, and savings behaviors.

Statistically significant differences with medium effect sizes were noted in the Bread and Cereals Ratio, Crop Farming and Gardening Ratio, Communication Ratio, Per Capita Expenditure, NonEssential Spending to Expenditure, and Essential Spending to Expenditure, highlighting meaningful variations in spending habits between clusters.

More minor yet significant effects ($p < 0.05$, small effect sizes) were detected in categories including Fish Ratio, Food Ratio, Housing and water Ratio, Clothing and Footwear Ratio, Miscellaneous Goods and Services Ratio, Special Occasions Ratio, Restaurant and hotels Ratio, Medical Care Ratio, Education Ratio, and Transportation Ratio. While these differences are statistically significant, their effect sizes indicate that the practical impact of these variations is relatively modest, Entrepreneurship Income to Total Income Ratio.

Overall, the results demonstrate that household clusters exhibit significantly different expenditure behaviors across multiple categories, with some differences being particularly pronounced in areas related to financial management, food consumption, and essential versus non-essential spending.

Income Sources

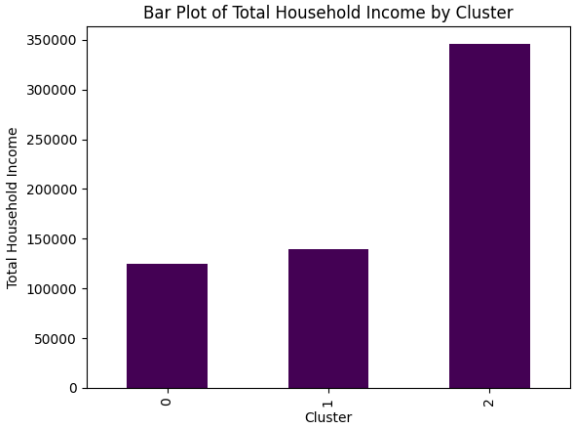


Fig. 20. Bar Plot of Average Total Household Income of Households in each Cluster

Average annual income from Households in Clusters 0 and 1 fail in comparison to households belonging to Cluster 2. On average, households from Clusters 0 and 1 earn around 140,000 pesos annually, which is about 200,000 pesos difference from households in Cluster 2. This can be attributed to the requirements for securing stable jobs in the Philippine job market, which rely heavily on educational attainment as a bare minimum. Given that most household heads from Clusters 0 and 1 do not finish a degree, limited opportunities are present when securing a job. Moreover, such findings show that jobs in the agricultural sector, especially farmers, earn way less than those in the industrial sector.

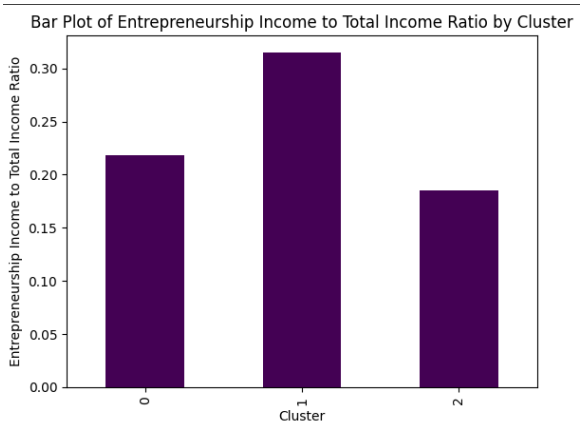


Fig. 21. Average Household Proporiton of Income from Entrepreneurial Activities

Given that most of the household heads from Cluster 1 engage in business, most of their income is sourced from their entrepreneurial activities. An observable difference of at least 10% exists between Cluster 0 or Cluster 2 in terms of the share of entrepreneurial income on the total household income. It is worth noting, however, that most of these households are farmers who earn way less than those from other industries in the Philippines, as observed from the discrepancy between the raw annual income.

Expenditure Priorities (Essential and Non-Essential Spending)

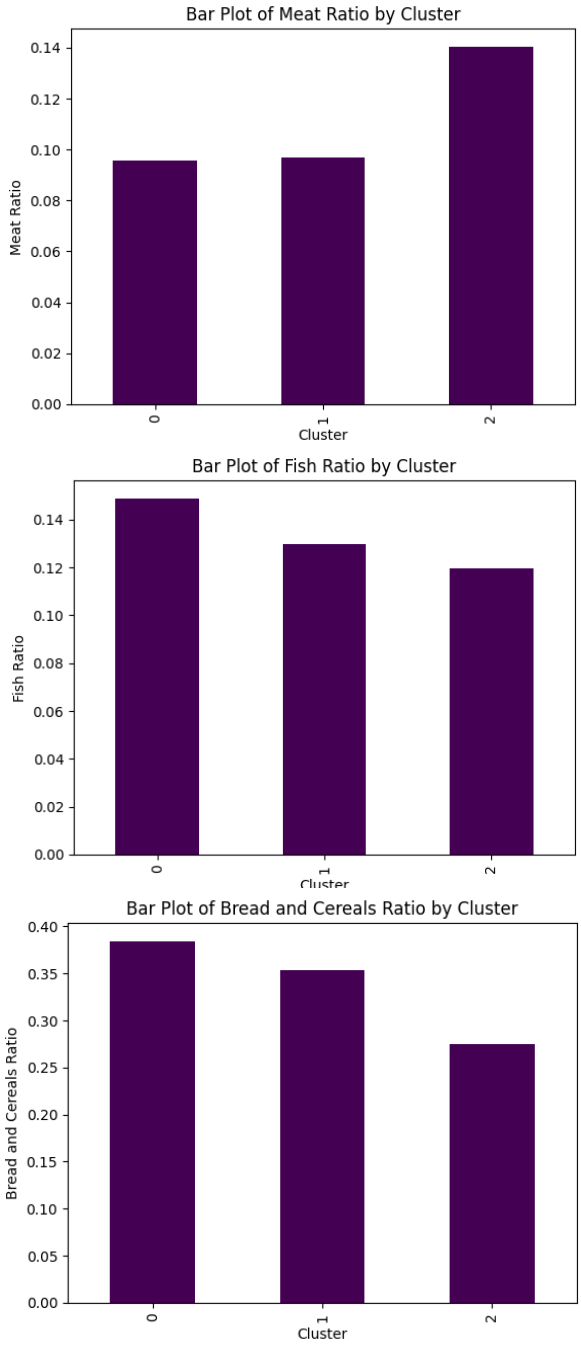


Fig. 22. Average Proportions of Meat, Fish, Bread and Cereals relative to Households' Food Expenditure

Households from Cluster 2 prefer Meat over Fish compared to Clusters 0 and 1, who like the latter. Fish preference for Cluster 0 could come from the proportion of fishermen within the cluster. Interestingly, Cluster 2 Households spend less on bread and cereals than those from Clusters 0 and 1. Expenditure on bread and cereals involves rice, which aligns with agricultural households from Clusters 0 and 1.

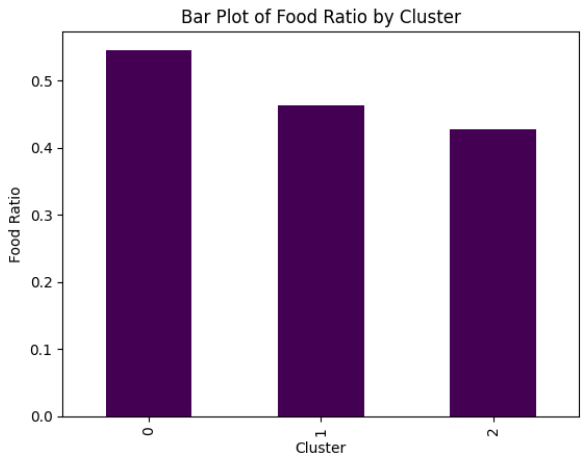


Fig. 23. Average Food Allocation for each Cluster relative to Households' Total Expenditure

Households from Cluster 0 put more emphasis on essential food expenditure than Cluster 2. Engel's law states that the amount spent on food decreases as household income increases [26]. Thus, there lies a possibility that expenditure is reallocated to other expenditure items for Cluster 2 households.

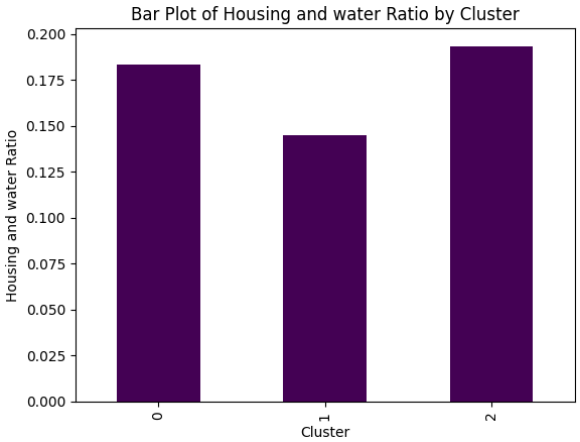


Fig. 24. Average Household Allocation on Housing and Water Expenditure Relative to Household's Total Expenditure

Cluster 1 has a noticeably lower expenditure allocation on "Housing and Water Ratio" than Clusters 0 and 2, indicating lower housing costs or water consumption. Cluster 2 consists of large households, which may incur more significant water spending and a larger area for accommodation than households from clusters 0 and 1. Households from Cluster 1 allocate relatively the same as Cluster 2, indicating the living conditions of low-income households thriving in areas with high costs of living in the Philippines.

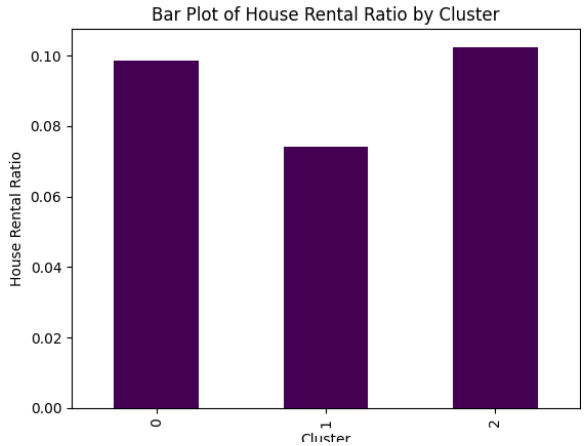


Fig. 25. Average Household Allocation on Rental Payments Relative to Total Expenditure

Households from Clusters 0 and 2 allocated more into rental payments than Cluster 1 households. Such a finding supports the similarity of the two clusters in Housing and Water expenditure allocation. Compared to Cluster 2 households, Cluster 0 households are low-income households that rent and pay for water in urban areas, thus increasing financial stress for such households.

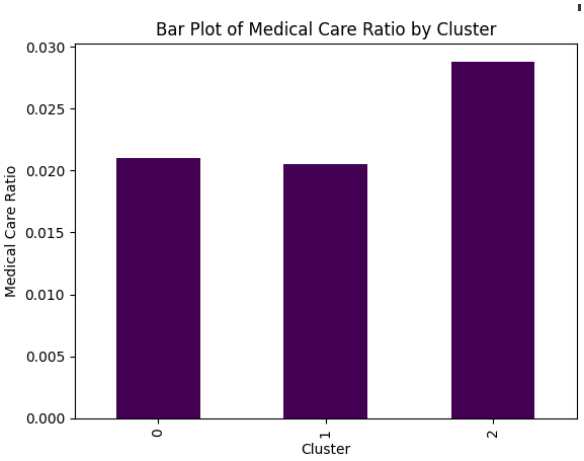


Fig. 26. Average Medical Expense Allocation Relative to Total Expenditure

Cluster 2 allocates a percent more on medical expenses compared to both Cluster 0 and Cluster 1. Given the raw income capabilities of Cluster 2, a one percent difference shows a significant disparity in healthcare access between high-income and low-income families.

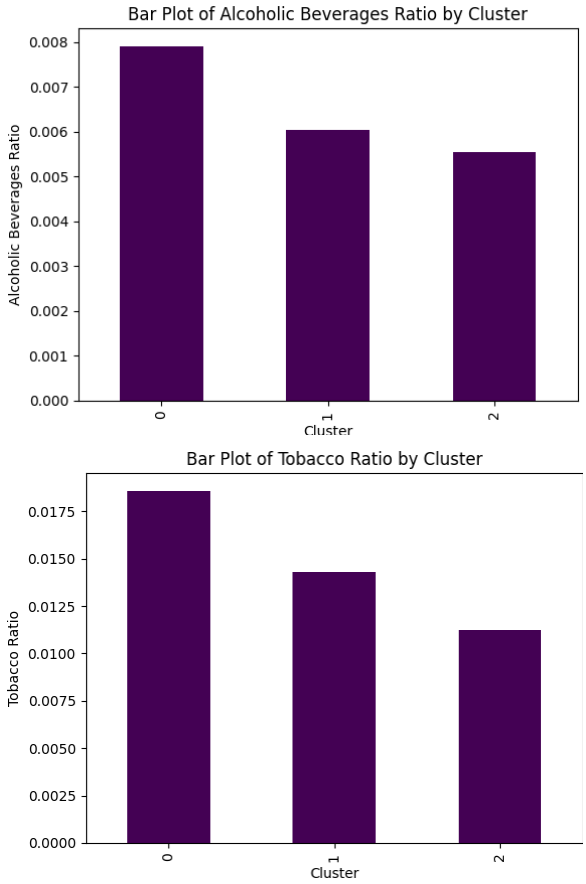
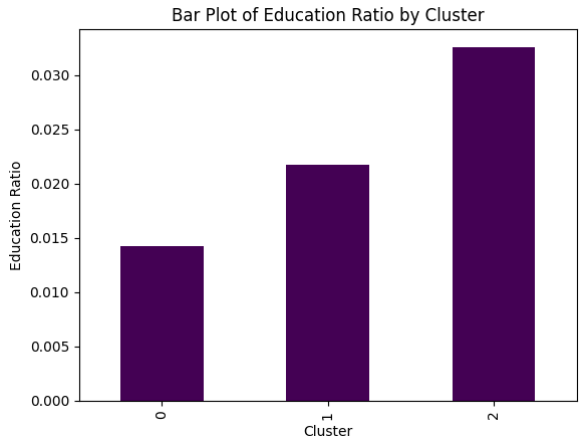


Fig. 27. Average Household Allocation on Vices Relative to Household's Total Expenditure

Households from Cluster 0 show significant spending on tobacco and alcoholic products as compared to Clusters 1 and 2. This raises health concerns regarding low-income households, especially the inequality in medical care access between high-income and low-income households



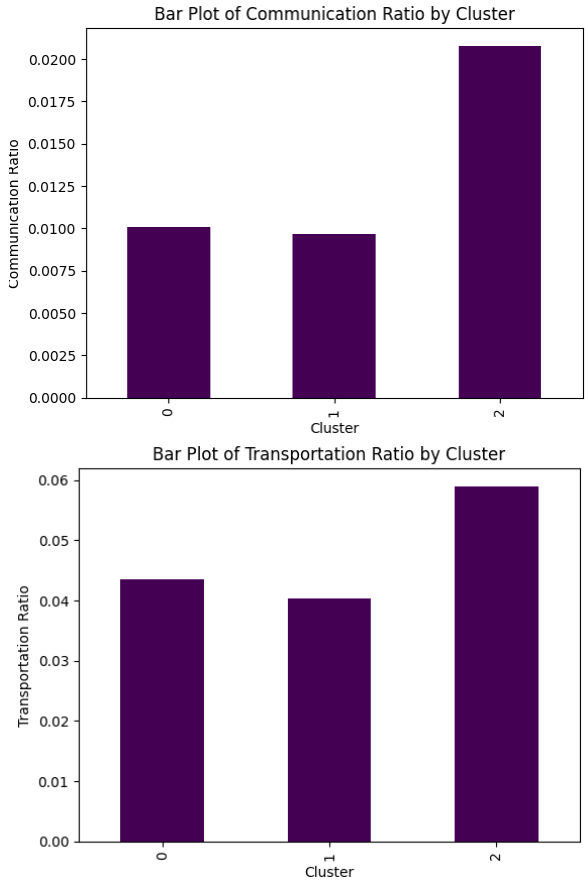


Fig. 28. Average Household Allocation on Education, Communication, and Transportation Relative to Total Expenditure

All 3 clusters of Filipino households allocate the same proportion to education, communication, and Transportation. Cluster 2 allocates at least a percent more than the latter two. Likewise, with healthcare, access to quality education, information, and transportation avenues remains despite similar spending behavior on the three aspects. Gaps in learning, awareness, and accessibility exist between high-income and low-income households.

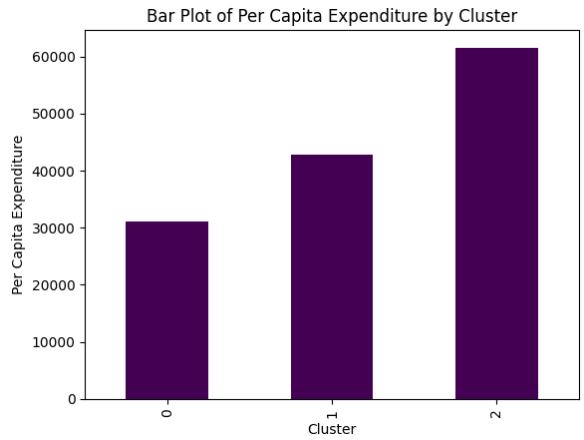


Fig. 29. Spending Contribution and Reliance of Each Household Member

Cluster 2 Households incur the highest spending contribution and reliance among the three clusters, possibly affected by high income or household size. However, concerns

are raised for Cluster 1 households with large family sizes that rely heavily on spending despite low annual household income.

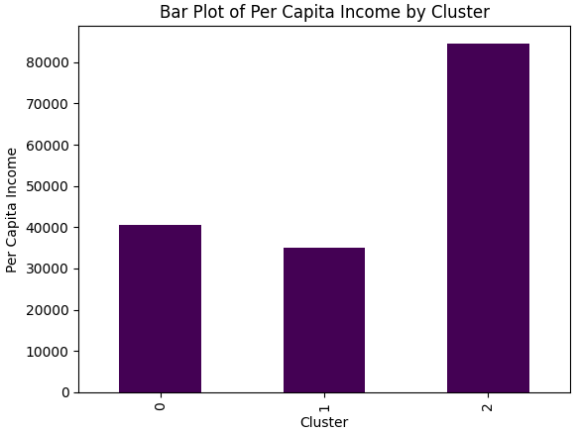


Fig.30. Average Income per Household Member

Both household members from Clusters 0 and 2 contribute more than their spending contribution, indicating potential savings or surplus. Despite their high spending reliance, cluster 2 households save money for future investments. However, households from cluster 2 spend way over their spending capability, possibly incurring debt or sourcing income outside their income capabilities.

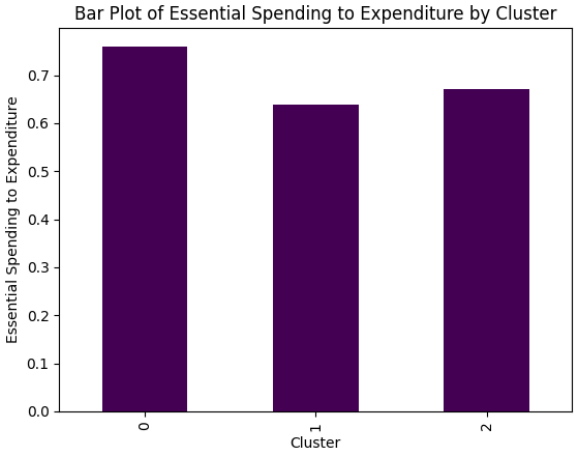


Fig. 31. Average Household Allocation on Essential Expenditure Items Relative to Total Expenditure

Low-income households from Cluster 0 allocate the most among the three clusters on essential spending items, which include housing and water, food, medical care, and communication. Such behavior could be attributed to the high cost of living in the area they live in as renters. Cluster 1, however, ranks the lowest among the three clusters of households on essential spending allocation. This could be attributed to the nature of their occupation as farmers or business people.

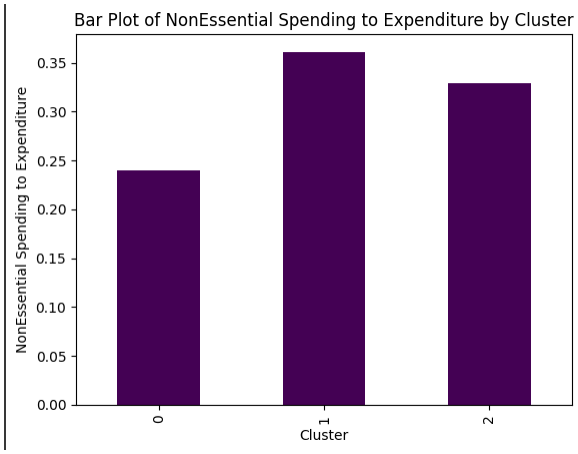


Fig. 32. Average Non-Essential Spending

Among the three clusters, households from Cluster 1 allocate the most money to non-essential spending. Despite having more income, households from Cluster 2 rank below Cluster 1, mostly low-income households with a large family size. This could be attributed to the nature of their occupation as farmers or business people.

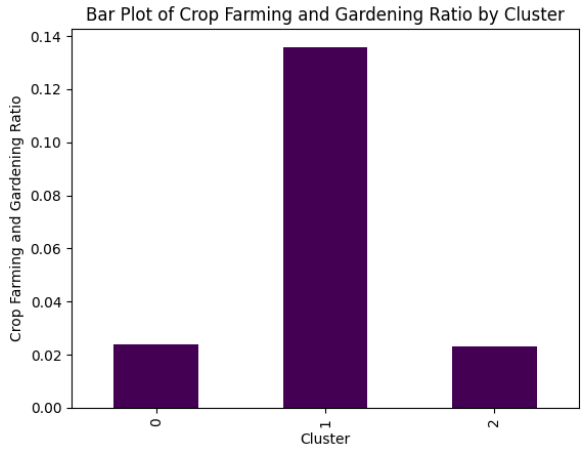


Fig. 33. Average Household Allocation on Crop Farming and Gardening Relative to Total Expenditure

As initially suspected, agricultural households from Cluster 2 were allocated the highest in their line of work due to its nature. Cluster 0 and Cluster 2 spent the least, possibly due to living outside farm areas and gardening as a hobby. Issues remain among farmers and businesses that invest but earn less in return, especially when the principal is costly.

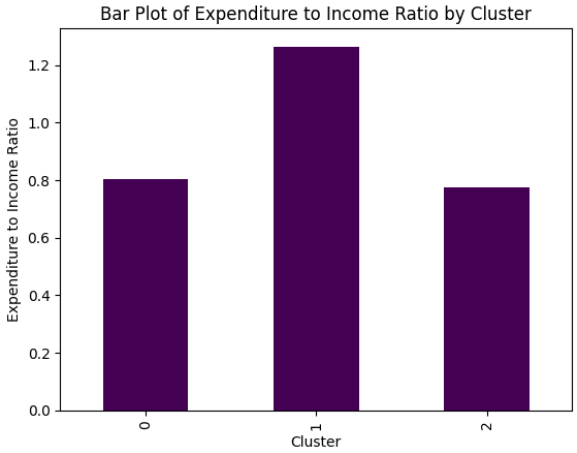


Fig. 34. Average Expenditure to Income among Households across different Clusters

Thus, households from Cluster 2, especially those in the agricultural sector, spend at least 20% above their spending capabilities. Such a finding raises the need for financial policies to aid households from Cluster 2 to make ends meet and save money for future investments and economic growth.

Savings and Debt Behavior

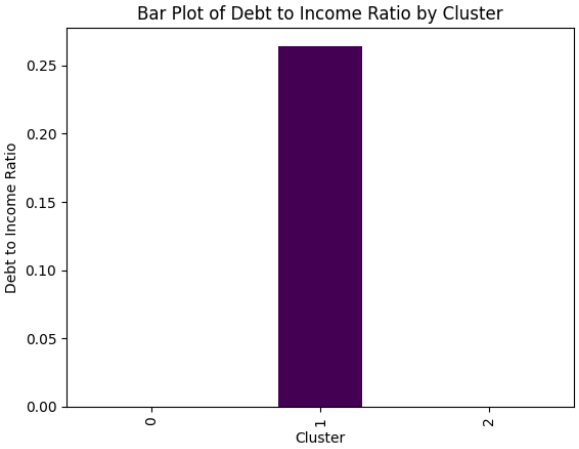


Fig. 35. Average Amount Borrowed Relative to Household Income

Households from Clusters 0 and 2 manage their money well to incur possible savings that may be allocated to future investments. In contrast, Cluster 2 households borrow money from external sources to cover their daily essentials and initial principal on their job occupations and businesses, indicating the need for inclusive and sustainable financial policies.

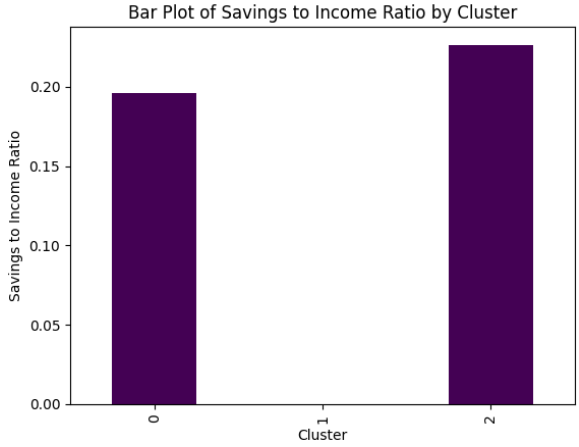


Fig. 36. Average Amount Saved Relative to Household Income

Clusters 0 and 2 households allocate effectively and are financially literate on their expenditures to ensure savings from their household income. Savings are vital for low-income households from Cluster 0 to maintain financial security and economic stability.

J. Distinguishable Characteristics from each cluster

Cluster 0: Urban Low-Income Households

This cluster represents small, low-income households that experience moderate employment. Many household heads work as fishermen, farmhands, or laborers, typically in informal or low-wage jobs. Education levels are relatively low, with most household heads not pursuing higher education.

Despite financial constraints, Cluster 0 households prioritize essential spending such as food, medical care, and housing. A significant portion of their income is allocated to rental payments and utility costs, suggesting they live in urban or semi-urban areas with a high cost of living. Their food preferences lean toward fish, possibly influenced by their occupations in fishing and foraging.

However, a concerning aspect of this cluster is the high expenditure on tobacco and alcohol, raising public health concerns. However, they also demonstrate financial literacy, effectively managing their expenses to save money. Their financial behavior highlights the importance of economic stability in low-income urban settings, where saving and budgeting are crucial for survival. Policies supporting affordable housing, medical care access, and financial literacy programs could benefit them.

Cluster 1: Struggling Agricultural Families

This cluster consists of large households with many dependents yet limited employment opportunities. Most household heads are engaged in agricultural work or informal businesses, such as farming and small-scale entrepreneurship. Due to their low educational attainment, they often face limited job opportunities outside the agricultural sector.

Their annual income is significantly lower than that of other clusters, making entrepreneurial activities a major source of livelihood. Despite their reliance on business income, their earnings remain low, indicating structural challenges in the agricultural sector. Households in this cluster allocate the least to essential spending and spend more on non-essential items, possibly due to the fluctuating nature of their income streams.

Cluster 1 also has the lowest spending on housing and

utilities, suggesting they live in areas with lower living costs or own their homes, possibly in rural areas. However, despite their financial struggles, these households tend to spend beyond their income capabilities, potentially incurring debt to cover daily needs and business operations. This highlights the need for financial policies and support for agribusiness sustainability. Providing subsidies, fair market access, and financial assistance prevent these households from incurring long-term debt cycles.

Cluster 2: Financially Stable Working Families

Cluster 2 consists of large, financially stable households with multiple working adults. Unlike Cluster 1, where large family sizes indicate economic strain, Cluster 2 households benefit from high employment rates and stable income sources. Many household heads hold managerial or professional positions, with higher education levels contributing to job security.

These households have the highest annual income among all three clusters, earning about 200,000 pesos more than Clusters 0 and 1 on average. Their financial capacity allows them to allocate spending efficiently across multiple categories, including medical care, education, and savings. They also exhibit lower reliance on staple food expenditures (e.g., bread and cereals) and a greater preference for meat over fish, possibly due to dietary shifts associated with higher income levels.

Despite their high earning capacity, Cluster 2 households demonstrate responsible financial management, saving money for future investments while meeting their spending needs. However, some households may still overspend beyond their means, potentially incurring debt to sustain their lifestyle. This raises concerns about financial planning and long-term sustainability for certain high-income families.

V. CONCLUSION

This study analyzed the socioeconomic disparities among Filipino households by examining their income and expenditure patterns. Using unsupervised machine learning techniques such as K-Means on PCA-transformed data, the research successfully identified three distinct household clusters, each exhibiting unique financial behaviors.

The findings suggest that Filipino households adopt varying financial strategies based on household head characteristics, income levels, expenditure priorities, and savings and debt behaviors. Significant disparities in expenditure allocation and financial management were observed among the clusters. Cluster 0 consists of low-income urban Filipino households prioritizing essential spending but struggling with financial security. Cluster 1 comprises individuals engaged in business or agriculture, with income primarily derived from entrepreneurial activities. This cluster also exhibits the highest non-essential spending, reflecting their lifestyle as farmers or business owners. Cluster 2 includes financially stable working households benefitting from higher education, stable employment, and better savings habits. The clustering results were validated using the Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score, with K-Means emerging as the most effective clustering model. The application of unsupervised learning enabled the discovery of hidden patterns in financial behavior that traditional analyses may overlook.

Beyond statistical significance, these insights can guide financial institutions and policymakers in developing targeted interventions for household segments. The findings emphasize the need for financial literacy programs, improved access to credit, and policies supporting agricultural households. Addressing these disparities through data-driven policy interventions can promote financial inclusion and long-term economic stability in the Philippines. Future research can enhance these findings by incorporating additional socioeconomic factors, exploring regional disparities, and utilizing longitudinal data to track financial behaviors over time.

REFERENCES

[1] Belghith Nadia, Belhaj Hassine Fernandez, Francine Claire Chang, David, Clarissa Crisostomo. “Overcoming Poverty and Inequality in the Philippines : Past, Present, and Prospects for the Future.” *World Bank*, documents.worldbank.org/curated/en/099325011232224571.

[2] Mangaluz, Jean. “PH Ranks 15th in World Bank’s Income Inequality Report | Inquirer News.” *INQUIRER.net*, 25 Nov. 2022, newsinfo.inquirer.net/1697382/fwd-world-bank-ph-ranks-15th-among-63-countries-for-income-inequality.

[3] “Consumer Finance Survey,” Bangko Sentral Ng Pilipinas. https://www.bsp.gov.ph/SitePages/MediaAndResearch/ConsumerFinanceSurvey.aspx

[4] “Family Income and Expenditure | Philippine Statistics Authority.” https://rsso11.psa.gov.ph/family-income-and-expenditure

[5] Albert, Asis, and J. Vizmanos, “National Accounts and Household Survey Estimates of household expenditures: Why do they differ and why should we be concerned?,” 2017. https://www.pids.gov.ph/publication/discussion-papers/national-accounts-and-household-survey-estimates-of-household-expenditures-why-do-they-differ-and-why-should-we-be-concerned

[6] M. R. V. & W. W. & Z. Z. Zhu, “Sources of inequality in the Philippines: Insights from stochastic dominance tests for richness and poorness,” ideas.repec.org, 2020, [Online]. Available: https://ideas.repec.org/a/bla/worlde/v43y2020i10p2650-2673.html

[7] Antonio, “Catastrophic Expenditure for Health in the Philippines,” Acta Medica Philippina, 2022, [Online]. Available: https://actamedicaphilippina.upm.edu.ph/index.php/acta/article/view/6190

[8] P. Biyanwila and N. Anuradha, “Behavioral Factors Affecting Household Over-Indebtedness: A Literature review,” Nov. 28, 2022. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4287738

[9] A. Santiago, S. Pandey, and Ma. T. Manalac, “Family presence, family firm reputation and perceived financial performance: Empirical evidence from the Philippines,” Journal of Family Business Strategy, vol. 10, no. 1, pp. 49–56, Feb. 2019, doi: 10.1016/j.jfbs.2019.02.002.

[10] Agustina. Tan-Cruz PhD, R. C. Castro PhD, J. B. C. Tan Dba, and E. D. Cruz PhD, “Family savings of the working population in Mindanao,” Mar. 01, 2020.

https://repository.umindanao.edu.ph/handle/20.500.14045/678
[11] C. S. Uy-Tioco, "'Good enough' access: digital inclusion, social stratification, and the reinforcement of class in the Philippines," Communication Research and Practice, vol. 5, no. 2, pp. 156–171, Apr. 2019, doi: 10.1080/22041451.2019.1601492.
[12] M. P. Naviamos and J. D. Niguidula, "A Study on Determining Household Poverty Status," ACM Digital Library, pp. 79–84, Jan. 2020, doi: 10.1145/3378936.3378969.
[13] M. J. P. Poirier, K. A. Grépin, and M. Grignon, "Approaches and Alternatives to the Wealth Index to Measure Socioeconomic Status Using survey Data: A critical interpretive synthesis," Social Indicators Research, vol. 148, no. 1, pp. 1–46, Sep. 2019, doi: 10.1007/s11205-019-02187-9.
[14] Huang, Kang, Xu, and Liu, "Robust deep k-means: An effective and simple method for data clustering," Science Direct, 2021, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0031320321001837
[15] Z. Xing and W. Zhao, "Block-Diagonal guided DBSCAN clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 11, pp. 5709–5722, May 2024, doi: 10.1109/tkde.2024.3401075.
[16] W.-B. Xie, Y.-L. Lee, C. Wang, D.-B. Chen, and T. Zhou, "Hierarchical clustering supported by reciprocal nearest neighbors," Information Sciences, vol. 527, pp. 279–292, Apr. 2020, doi: 10.1016/j.ins.2020.04.016.
[17] L. Scrucca, "Identifying connected components in Gaussian finite mixture models for clustering," Computational Statistics & Data Analysis, vol. 93, pp. 5–17, Jan. 2015, doi: 10.1016/j.csda.2015.01.006.
[18] C. Ellis, "When to use gaussian mixture models - Crunching the Data," Crunching the Data, Jun. 08, 2022. https://crunchingthedata.com/when-to-use-gaussian-mixture-models/
[19] "2018 Consumer Finance Survey: A Snapshot of Filipino Household Finances," 2018. [Online]. Available: https://www.bsp.gov.ph/Media_And_Research/Consumer%20Finance%20Survey/CFS_2018.pdf
[20] "2012 Philippine Standard Occupational Classification." https://psa.gov.ph/classification/psoc/technical-notes
[21] "Data Clustering Algorithms - k-means clustering algorithm." https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
[22] "14.4 - Agglomerative Hierarchical Clustering," PennState Ebberly College of Science. https://online.stat.psu.edu/stat505/lesson/14/14.4
[23] A. Chai, E. Stepanova, and A. Moneta, "Quantifying expenditure hierarchies and the expansion of global consumption diversity," Science Direct, 2023, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167268123002457
[24] Albert, Abrigo, Quimba, and Vizmanos, "Poverty, the Middle Class, and Income Distribution amid COVID-19," Philippine Institute of Development Studies, 2020, [Online]. Available: https://pidswebs.pids.gov.ph/CDN/PUBLICATIONS/pidsdps2022.pdf
[25] H. Ritchie, "Engel's Law: Richer people spend more money on food, but it makes up a smaller share of their income," Our World in Data, Jan. 19, 2023. https://ourworldindata.org/engels-law-food-spending
[26] A. Hayes, "Engel's Law, Curve, and Coefficient Explained," Investopedia, Aug. 07, 2024. https://www.investopedia.com/terms/e/engels-law.asp
[27] "Filipino family income and expenditure," Kaggle, Oct. 05, 2017. https://www.kaggle.com/datasets/grosvenpaul/family-income-and-expenditure/data