Report

Prayas Dataset

E467: Group Assignment 2

Group Number - 11

Margaux Massol	- 25V0014
Prajwal Nayak	- 22B4246
Sarvadnya Purkar	- 22B4232
Somisetty Ruchita	- 21D170041

1.0 Executive Summary

This report presents a comprehensive, multi-layered analysis of household energy consumption data from the eMARC dataset. The primary objective was to move beyond conventional, surface-level metrics to develop a deep, actionable understanding of consumer behavior. By integrating static profiling using Gaussian Mixture Models, temporal analysis of daily consumption patterns, and advanced time-series clustering techniques, this project successfully segmented households into distinct, behaviorally-defined personas.

A pivotal discovery was the identification of a statistically significant behavioral shift in the majority of households, strongly correlated with the 2020 COVID-19 lockdowns. This event reshaped typical consumption rhythms, creating a "new normal" of high, consistent baseload energy use.

Furthermore, the analysis successfully identified future-focused personas based on their long-term energy trajectories ("The Escalators," "The Stable," and "The Converters"). The culmination of this research is a strategic framework for driving real-world change through hyper-targeted advice, data-driven peer benchmarking, and a quantitative method for measuring the ROI of future conservation initiatives. This report transforms raw data into a strategic intelligence tool for proactive, data-informed energy management.

2.0 Foundational Analysis: Establishing Core Personas

The initial phase focused on creating rich, static profiles to understand the fundamental characteristics of each household, addressing the "who" and "how much" of energy consumption.

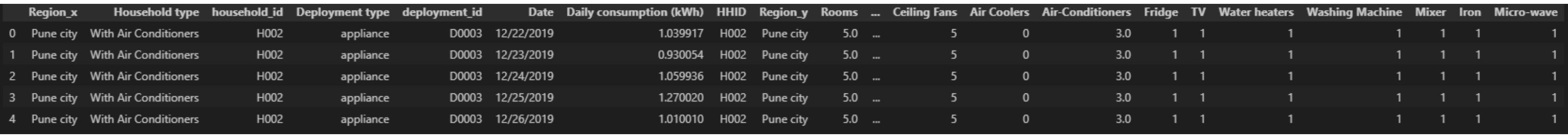
2.1 The Critical Role of Exploratory Data Analysis (EDA)

Before any advanced modeling or clustering could be performed, a thorough **Exploratory Data Analysis (EDA)** phase was conducted to ensure the quality, integrity, and readiness of the data. This foundational step was crucial for understanding the underlying structure of the dataset, identifying potential issues, and uncovering initial patterns that would guide the subsequent, more complex stages of the analysis. The EDA process was methodical and involved several key stages, each providing vital insights.

2.2 Data Loading, Merging, and Initial Inspection

The analysis began with two separate datasets: a time-series file containing daily energy consumption readings and a survey file detailing the static characteristics of each household. The first critical step was to merge these two sources into a single, cohesive DataFrame. This was achieved by joining them on the `household_id` key, creating a unified dataset where each daily consumption entry was enriched with the corresponding household's survey data, such as its area, number of occupants, and appliance ownership.

Initial inspections were then performed using functions like `.info()` and `.describe()` to get a preliminary overview of the data's structure, identify the data types of each column, and check for obvious missing values.



	Region_x	Household type	household_id	Deployment type	deployment_id	Date	Daily consumption (kWh)	HHID	Region_y	Rooms	...	Ceiling Fans	Air Coolers	Air-Conditioners	Fridge	TV	Water heaters	Washing Machine	Mixer	Iron	Micro-wave
0	Pune city	With Air Conditioners	H002	appliance	D0003	12/22/2019	1.039917	H002	Pune city	5.0	...	5	0	3.0	1	1	1	1	1	1	1
1	Pune city	With Air Conditioners	H002	appliance	D0003	12/23/2019	0.930054	H002	Pune city	5.0	...	5	0	3.0	1	1	1	1	1	1	1
2	Pune city	With Air Conditioners	H002	appliance	D0003	12/24/2019	1.059936	H002	Pune city	5.0	...	5	0	3.0	1	1	1	1	1	1	1
3	Pune city	With Air Conditioners	H002	appliance	D0003	12/25/2019	1.270020	H002	Pune city	5.0	...	5	0	3.0	1	1	1	1	1	1	1
4	Pune city	With Air Conditioners	H002	appliance	D0003	12/26/2019	1.010010	H002	Pune city	5.0	...	5	0	3.0	1	1	1	1	1	1	1

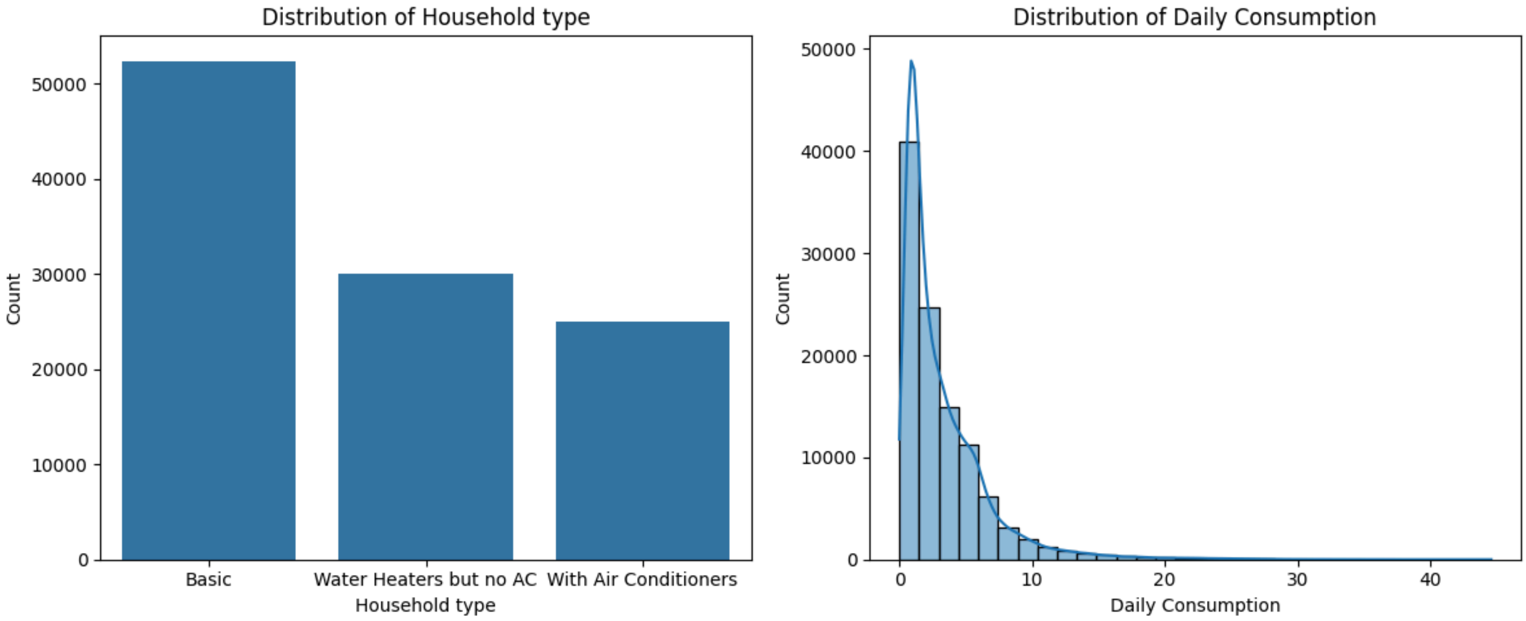
Data Cleaning and Preprocessing

The raw, merged dataset required significant cleaning to prepare it for robust analysis. The most critical task was the conversion of the 'Date' column from a generic object (string) type into a proper **datetime format**. This was an essential step that enabled all subsequent time-series operations, including chronological plotting, time-based filtering (e.g., separating pre- and post-lockdown periods), and the calculation of trend-based features. We also systematically checked for missing values and handled them appropriately to prevent errors during modeling. Basic outlier detection was performed on the consumption data to ensure that extreme, anomalous readings did not unduly skew the analysis. This meticulous cleaning process ensured the reliability and accuracy of all subsequent findings.

Univariate Analysis: Understanding Key Variables

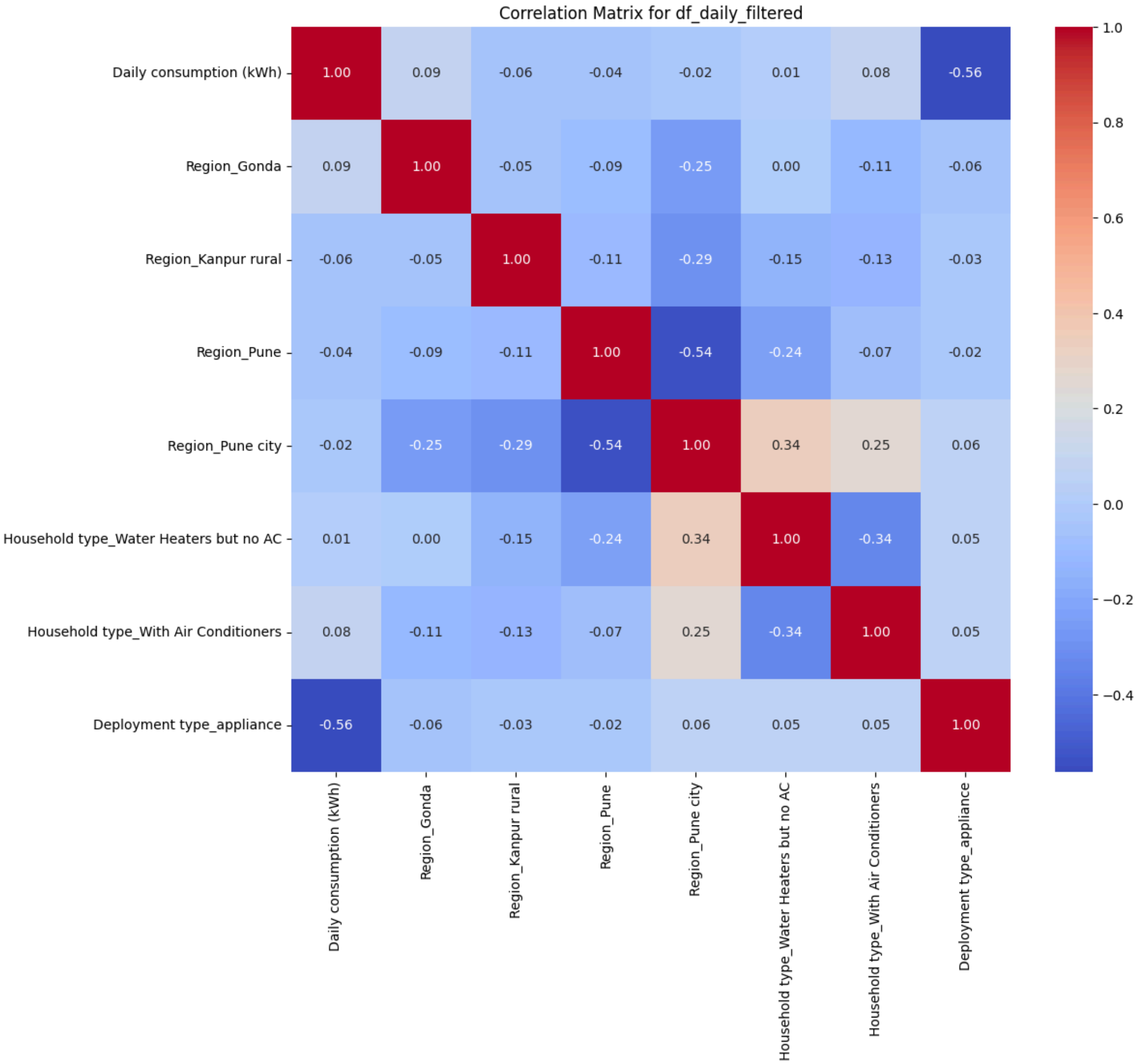
Univariate analysis involved examining individual variables to understand their distributions. This step helps to grasp the baseline characteristics of the dataset.

- **Consumption Distribution:** A histogram of 'Daily consumption (kWh)' was generated. This revealed a right-skewed distribution, indicating that while most households had relatively low-to-moderate daily consumption, a smaller number of households had significantly higher consumption, pulling the average up.
- **Household Composition:** A bar chart for the categorical 'Household type' variable showed the frequency of each category ('With Air Conditioners', 'Water Heaters but no AC', 'Basic'). This visual quickly established the overall appliance landscape of the households in the study.



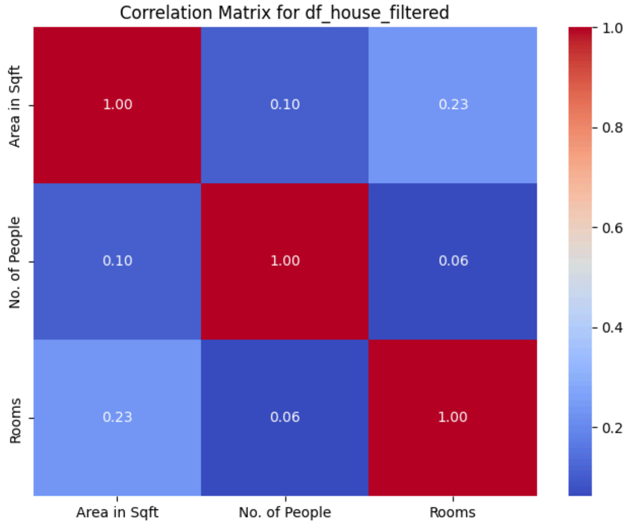
Bivariate Analysis: Uncovering Initial Relationships

With a grasp of the individual variables, bivariate analysis explored the relationships between pairs of variables to uncover initial patterns. This was where the first powerful insights began to emerge.



- **Consumption vs. Appliance Ownership:** We used box plots to visually compare the distribution of 'Daily consumption (kWh)' across the different 'Household type' categories. This plot provided the first clear visual evidence that households **with major appliances like air conditioners had a statistically different consumption profile**, showing a higher median and wider range of consumption.
- **Consumption vs. Household Size:** Scatter plots were employed to investigate the relationship between consumption, home area, and the number of occupants. These plots revealed a positive, albeit noisy, correlation, suggesting that larger homes with more people generally tended to use more energy.

This exploratory phase was instrumental in forming the initial hypotheses that were later tested and confirmed with more advanced statistical methods. The patterns uncovered here provided the direct rationale for the feature engineering that followed, confirming that metrics like EUI (normalizing consumption by area) would be a valuable and insightful feature for the final clustering models.



2.2 Methodology: Feature Engineering and Clustering

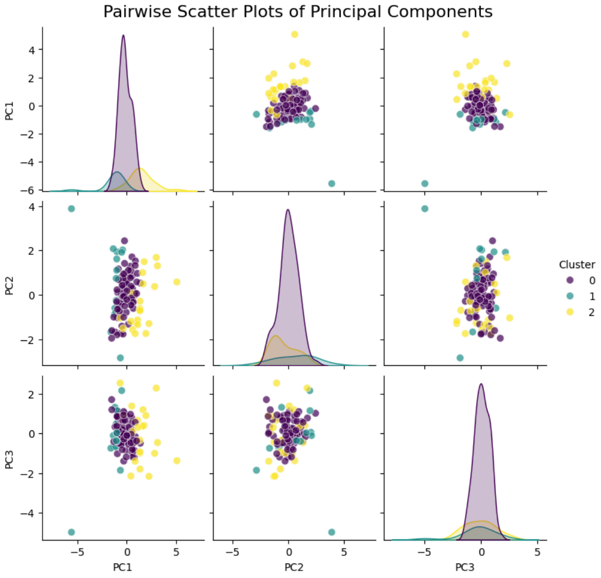
After an initial data cleaning and validation process where daily consumption data was merged with household survey data, we engineered several key features to capture the nuances of energy behavior:

- **Energy Use Intensity (EUI):** Consumption normalized by home area (kWh/Sqft) to measure efficiency.
- **Per-Capita EUI:** EUI further normalized by the number of occupants to understand density effects.
- **Consumption Volatility:** The standard deviation of daily use, serving as a proxy for the predictability of a household's routine.
- **Seasonal Ratio:** The ratio of mean summer consumption to mean winter consumption, quantifying weather dependency.
- **Weekend-Weekday Ratio:** Capturing the difference in weekly lifestyle patterns.

Using this enriched feature set, we employed a **Gaussian Mixture Model (GMM)** for clustering. This probabilistic model was chosen over simpler methods like K-Means due to its ability to handle overlapping clusters and provide a more flexible and nuanced segmentation.

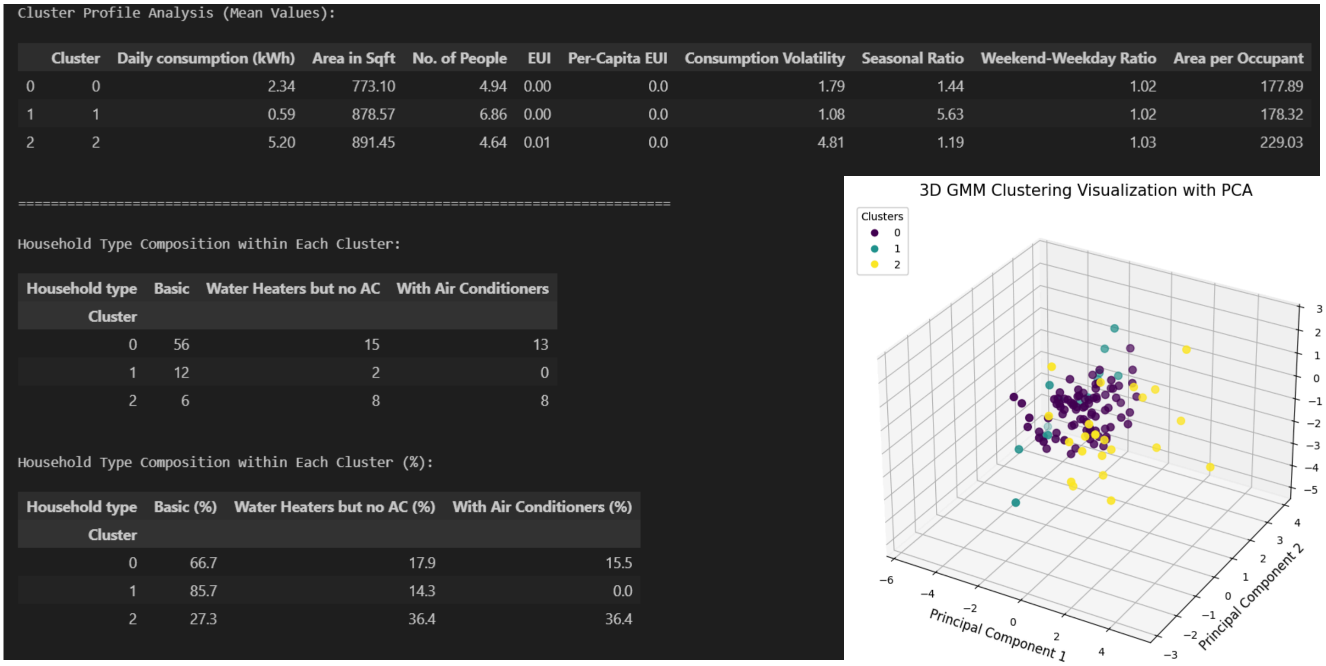
We also performed K-Mean clustering using only partial features (EUI, Consumption Volatility and Seasonal Ratio) to compare results and household profiles obtained when using simpler algorithms.

2.3 GMM Results: Three Foundational Personas



The GMM clustering algorithm identified three robust and distinct personas:

- **The High Baseline Consumers:** This group (22 households) is defined by the **highest average daily consumption (5.20 kWh)** and high volatility. Their energy use is consistently high year-round, not driven by seasons, pointing to a lifestyle heavily reliant on major appliances. Composition analysis confirmed they have the highest concentration of homes with both water heaters and air conditioners (36.4% each).
- **The Steady Spenders:** The largest cohort (84 households), representing the "average" household with **moderate and predictable energy consumption (2.34 kWh)**. This group forms the stable core of the user base.
- **The Minimalist Users:** A small but highly efficient group (14 households) with **exceptionally low daily consumption (0.59 kWh)**. Composition analysis shows they are predominantly 'Basic' households (85.7%) with no major energy-intensive appliances.



```
Cluster Profile Analysis (Mean Values):
```

	Cluster	Daily consumption (kWh)	Area in Sqft	No. of People	EUI	Per-Capita EUI	Consumption Volatility	Seasonal Ratio	Weekend-Weekday Ratio	Area per Occupant
0	0	2.34	773.10	4.94	0.00	0.0	1.79	1.44	1.02	177.89
1	1	0.59	878.57	6.86	0.00	0.0	1.08	5.63	1.02	178.32
2	2	5.20	891.45	4.64	0.01	0.0	4.81	1.19	1.03	229.03

```
================================================================================

Household Type Composition within Each Cluster:
```

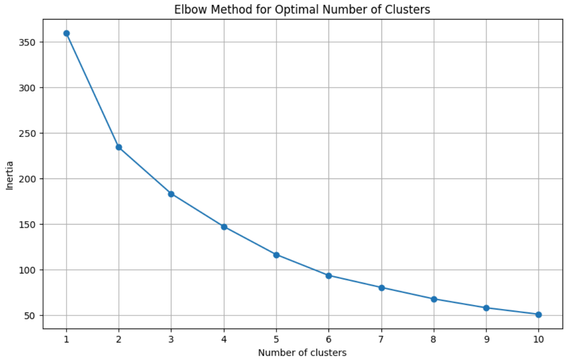
Household type	Basic	Water Heaters but no AC	With Air Conditioners
Cluster			
0	56	15	13
1	12	2	0
2	6	8	8

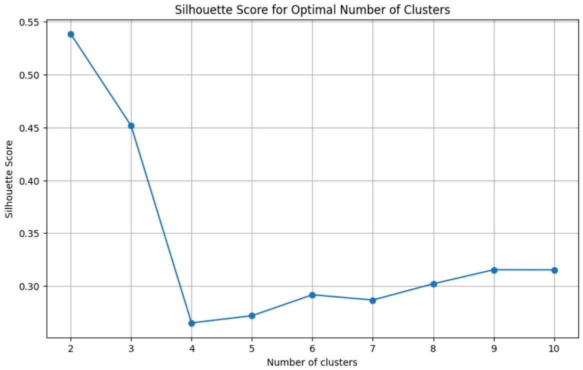
```
Household Type Composition within Each Cluster (%):
```

Household type	Basic (%)	Water Heaters but no AC (%)	With Air Conditioners (%)
Cluster			
0	66.7	17.9	15.5
1	85.7	14.3	0.0
2	27.3	36.4	36.4

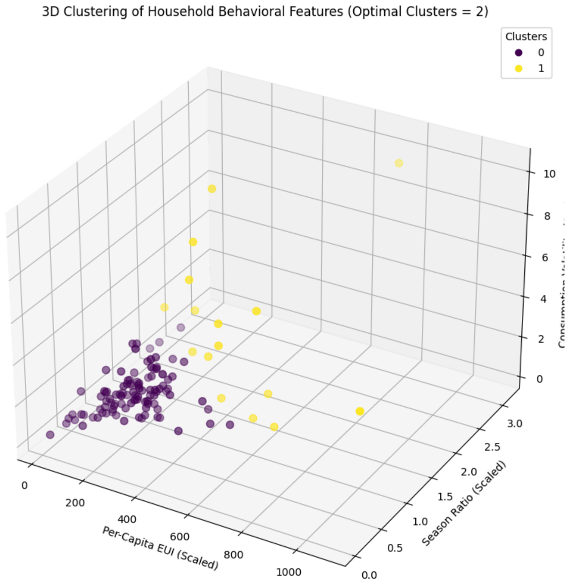
2.4. K-Means results : Two Behavioral Profiles

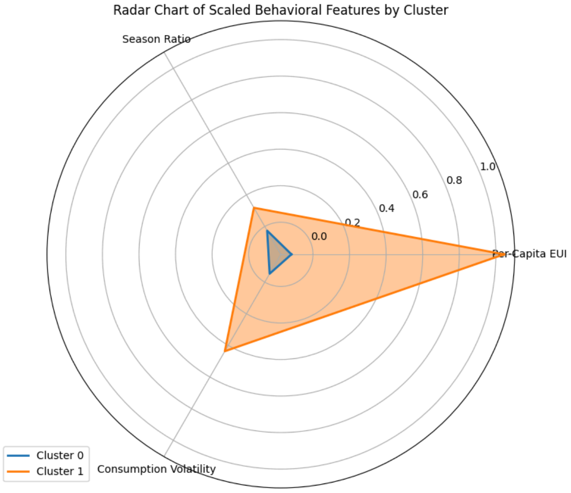
After standardizing EUI, Seasonal Ratio, and Consumption Volatility features using z-score scaling. K-Means clustering was applied across a range of cluster numbers to determine the optimal number of clusters through both the Elbow Method and Silhouette Scores. Both test results revealed that two clusters provided a clear separation between low and high consumption behaviors.





Cluster assignments were then visualized in three-dimensional feature space, highlighting distinct behavioral groupings based on EUI and consumption volatility.

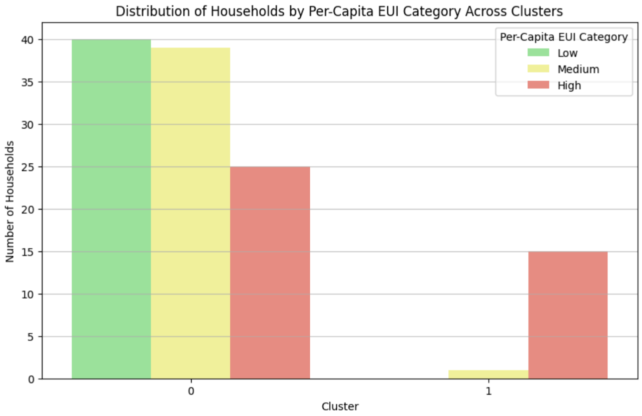




- Similarly to our previous **Steady Spenders** cluster, Cluster 0 consists of a **104 energy-stable, low-consuming** households.
- Cluster 1 captures **16 high-consumption, behaviorally dynamic households**. Their energy use is more variable, with moderate seasonal dependency and higher overall demand.

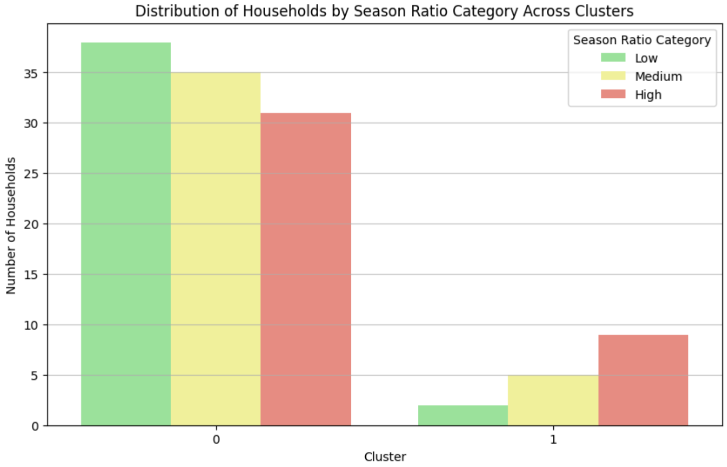
When diving into deeper cluster profile type analysis, several clear behavioral patterns emerge across per-capita EUI, seasonal ratio, and consumption volatility categories. These insights not only validate the effectiveness of the clustering but also highlight actionable pathways for targeted interventions.

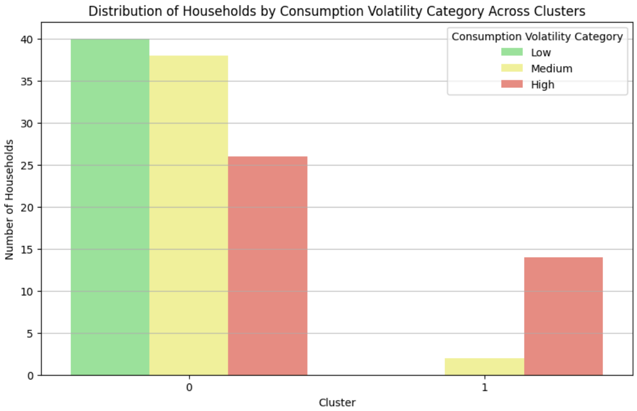
The first set of charts breaks down households by EUI category within each cluster.



Cluster 0 contains the vast majority of low and medium EUI households, with roughly 40 households in each category. There is also a notable number of high-EUI households in Cluster 0, indicating that even within the “efficient” group, a subset of users has higher individual energy intensity. Cluster 1, in contrast, is dominated by high-EUI households, with very few medium and almost no low users. This separation confirms that the clustering process effectively differentiated between lower-intensity and high-intensity users.

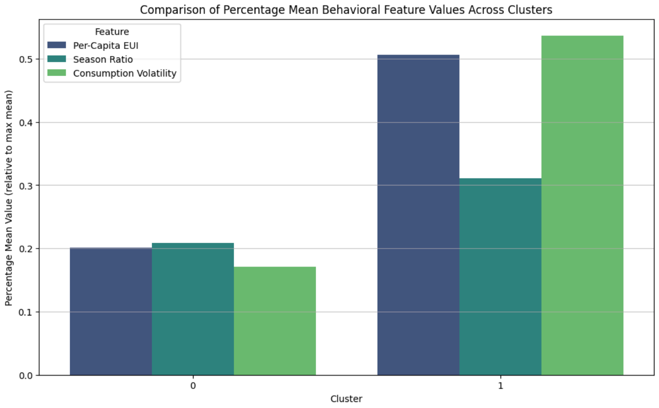
The seasonal ratio distribution tells a complementary story. Cluster 0 is again composed largely of low and medium ratio households, suggesting relatively stable or moderately seasonal consumption patterns. A smaller but significant group with high seasonality also exists in this cluster, which may represent households using cooling or heating seasonally but still maintaining overall lower intensity. Cluster 1 shows a greater concentration of high seasonality households, reinforcing the idea that this group’s consumption is more strongly driven by seasonal peaks—most likely due to air conditioning or similar appliances.





When examining consumption volatility, Cluster 0 again dominates the low and medium categories, while Cluster 1 is largely populated by high-volatility households. This indicates that households in Cluster 1 not only consume more energy but also do so in a less stable manner, with greater fluctuations over time.

These categorical distributions align closely with the mean behavioral feature comparison. Cluster 1's mean per-capita EUI and volatility values are more than double those of Cluster 0. The seasonal ratio difference is less pronounced but still significant, suggesting that Cluster 1 households are simultaneously high-intensity and peak-driven, whereas Cluster 0 households are lower intensity with more balanced demand profiles.



The final chart examines household type distribution within each cluster. Cluster 0 is dominated by Basic households, as well as a large share of water-heater and AC households, indicating that even appliance-owning households can maintain efficient, low-volatility profiles through better load management. In contrast, Cluster 1 contains a higher proportion of appliance-heavy households, particularly those with water heaters and air conditioners. This composition supports the interpretation that appliance ownership—especially cooling—plays a major role in driving both seasonal peaks and volatility.

Taken together, these visualizations paint a coherent picture of three distinct household behavioral personas - independent of cluster and closer to the GMM clustering results :

- **Efficient and Stable (Cluster 0 – Low/Medium EUI, Low Volatility)**: Predominantly Basic households, with some appliance owners who manage their load effectively. Policy efforts here should focus on **maintaining performance** and **reinforcing best practices**, for example through community recognition and behavioral nudges.

- **Seasonal Peak Spenders (High Season Ratio & Volatility)**: A smaller group within Cluster 1 characterized by **erratic, peak-driven demand**, likely due to seasonal cooling or heating. These households are prime candidates for **demand-side management** measures—such as promoting efficient air conditioning systems, offering programmable thermostats, or providing real-time consumption feedback.

- **Structurally Inefficient (High EUI, Lower Volatility)**: Households that consume more energy consistently throughout the year, likely due to **inefficient appliances**, poor insulation, or continuous standby loads. For this group, **appliance replacement programs** and **energy audits** would yield the greatest impact.

This analysis demonstrates that clustering goes beyond descriptive segmentation. It provides targeted, evidence-based strategies for different household groups: sustaining efficiency among low users, managing peaks among seasonal spenders, and addressing structural waste among steady high users. These distinctions are critical for designing effective, personalized energy interventions.

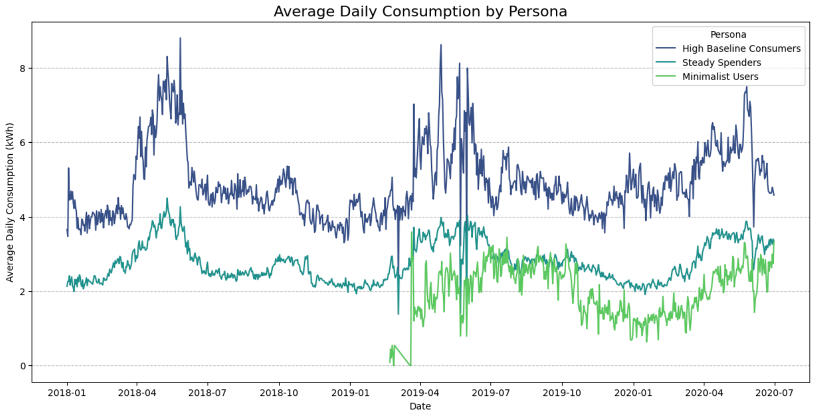
While the K-Means clustering provides a useful initial segmentation of households into two broad behavioral groups, the resulting clusters are relatively coarse and lack the level of profile differentiation needed for more nuanced policy design. In particular, they fail to fully capture the diversity of seasonal and volatility-driven usage patterns observed in the dataset. By contrast, the Gaussian Mixture Model (GMM) offers greater flexibility in identifying clusters with overlapping boundaries and varying shapes, enabling more behaviorally descriptive groupings. Therefore, in the next stage of the analysis, we proceed with GMM-based clustering to extract richer household personas and to better inform targeted intervention strategies.

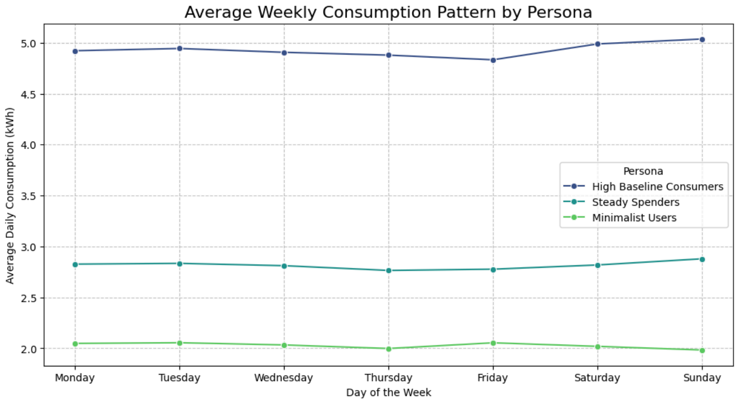
3.0 Temporal & Behavioral Analysis: Rhythms and Real-World Events

This phase added a time-series dimension to our static profiles, focusing on how and when energy is consumed.

3.1 Visualizing Persona Behavior

Plotting the average daily consumption for each persona over the entire time period visually confirmed our initial findings. The "High Baseline Consumers" exhibited a consistently elevated energy use, while the "Steady Spenders" showed a more moderate, seasonal curve, and the "Minimalists" remained flat near zero.





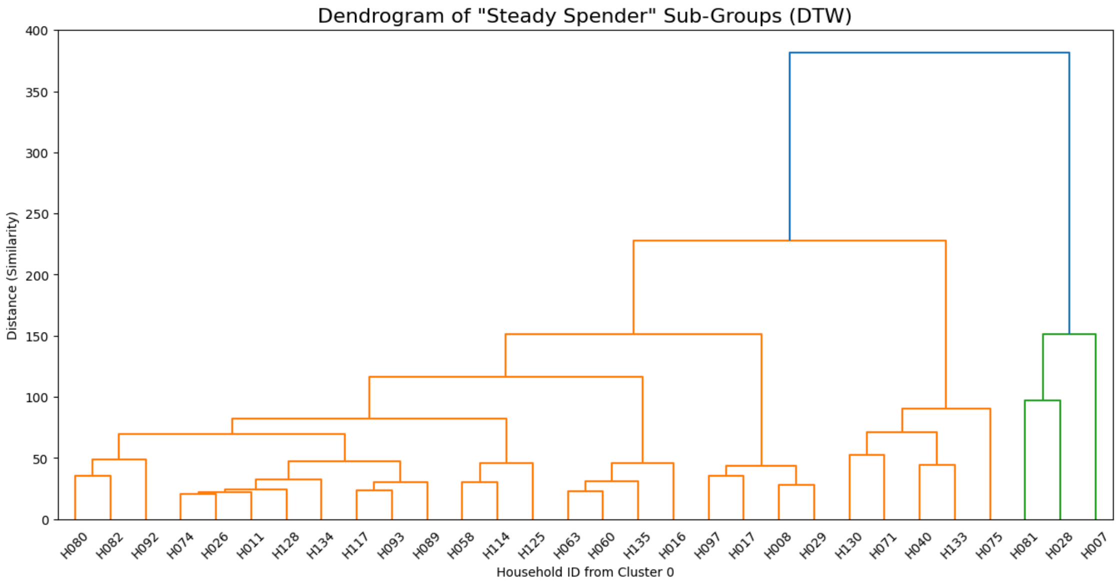
3.2 Deep Dive: Uncovering the COVID-19 Lockdown Effect

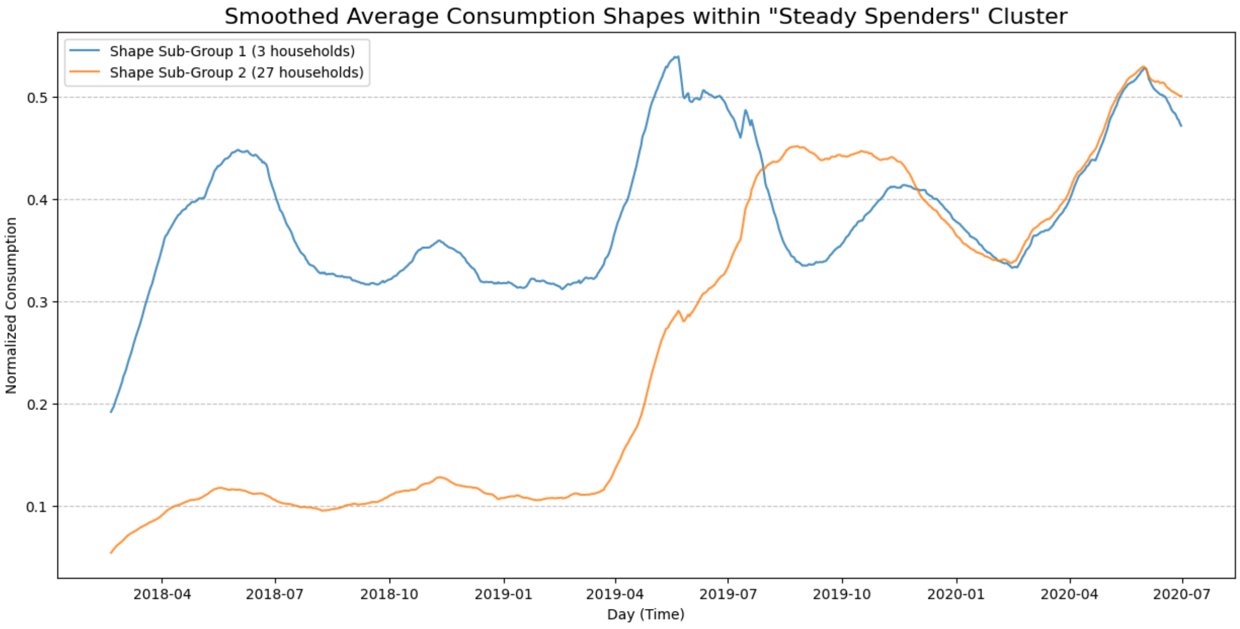
The most significant finding emerged from a two-stage clustering analysis. We isolated the largest persona, the "Steady Spenders," and applied **Dynamic Time Warping (DTW)**—a technique that clusters time series based on *shape* similarity rather than magnitude—to uncover hidden lifestyle patterns.

This revealed that the group was not homogenous but composed of two distinct sub-groups whose behavior was profoundly shaped by the 2020 pandemic:

- **The "Consistent Responders" (27 of 30 in the sample):** This dominant majority displayed an abnormally flat, consistently high consumption pattern throughout 2020. This is a clear behavioral signature of a work-from-home lifestyle, where constant appliance use (computers, lighting, cooking) created a new, high baseload that masked typical summer seasonality.
- **The "Seasonal Outliers" (3 of 30 in the sample):** This tiny minority maintained a traditional, weather-driven consumption pattern with a clear summer peak, serving as a control group against the lockdown-affected majority.

This hypothesis was **statistically validated** using an independent t-test comparing consumption before and after February 2020 for the main group. The result was a **p-value of 1.04e-31**, providing overwhelming evidence that the observed shift in consumption was a real, statistically significant phenomenon.





4.0 Predictive Analysis: Identifying Future Consumption Trajectories

The final analytical phase shifted focus from past behavior to future trends, aiming to identify which households were on a path of increasing or decreasing consumption.

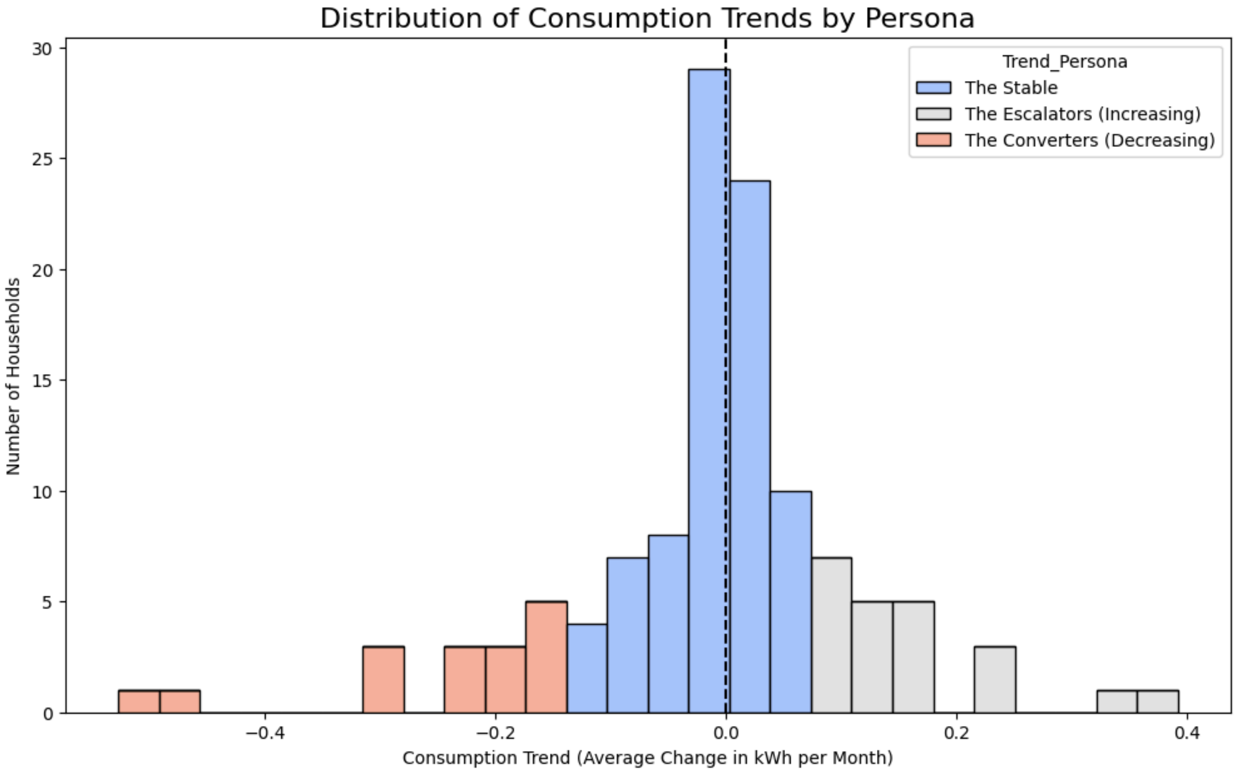
4.1 Methodology: Clustering by Trend

We fitted a linear regression model to the year-long consumption data for every household. The **slope of this line** was extracted as a single, powerful feature representing the household's "Consumption Trend." We then used K-Means clustering on these slope values to segment the population.

4.2 Results: Future-Focused Personas

This analysis identified three critical, forward-looking groups:

- **The Escalators (22 Households):** A high-priority group whose energy consumption showed a statistically significant **increasing trend** over the year.
- **The Stable (82 Households):** The majority of households, whose long-term consumption was consistent and predictable.
- **The Converters (16 Households):** A "success story" group that demonstrated a **decreasing trend** in energy use, providing a potential model for successful conservation.



5.0 In-Depth Profiles and Strategic Recommendations

5.1 The High Baseline Consumers

This persona represents a critical segment of high-energy users whose behavior is defined by consistency rather than seasonality. Their average daily consumption of 5.20 kWh is the highest among all groups, and their usage remains elevated throughout the year. This pattern strongly indicates that their primary energy drivers are not weather-dependent activities like cooling, but rather a large and constant baseload demand. This is likely attributable to a combination of factors, including the use of older, less efficient major appliances (such as refrigerators or water heaters), a high number of electronic devices, or lifestyle habits that involve continuous energy use.

The key insight for this group is that their path to savings lies in addressing this constant energy drain. Therefore, the most effective actionable strategy is to target them with recommendations for a comprehensive **appliance audit**. This intervention is precisely tailored to their consumption profile. The goal is to guide them in systematically identifying the specific appliances contributing most to their high baseload. Communications should focus on the long-term financial benefits of upgrading to modern, energy-efficient models, thereby offering a direct and impactful solution to their specific energy consumption pattern.

5.2 The "Steady Spenders": A Story of Two Sub-Groups

Our analysis revealed that the largest "average" user group is, in fact, not a monolith. Deeper time-series clustering showed a clear split, with behaviors largely defined by their response to the 2020 COVID-19 lockdowns.

5.2.1 Sub-Group A: The "Consistent Responders" (The Majority)

These households represent the "new normal" that emerged during the pandemic era. Their consumption profile is abnormally flat and consistently high, a statistically significant shift confirmed by our analysis. Their traditional seasonality, which would typically show a summer peak, was effectively erased by a new, high-energy daily routine. This is the clear behavioral fingerprint of widespread work-from-home and remote learning, involving the constant use of computers, monitors, lighting, and kitchen appliances throughout the day.

The actionable strategy for this group must acknowledge this new reality. Engaging them with advice tailored to **work-from-home habits** is crucial. Recommendations should focus on practical solutions for their environment, such as promoting energy-efficient home office setups, educating them on the impact of "phantom power" from always-on electronics, and suggesting the use of smart power strips to easily manage and de-energize workstations when not in use. This approach addresses their specific, pandemic-induced consumption pattern directly.

5.2.2 Sub-Group B: The "Seasonal Outliers" (The Minority)

These three households served as a crucial control group within the dataset. They maintained a traditional, weather-driven consumption pattern with a clear summer peak, seemingly unaffected by the lockdown trends that reshaped the behavior of the majority. Their energy use is predictable and directly tied to the seasons, making their primary energy driver easy to identify. Our analysis quantified that an average air conditioning unit adds approximately 2 kWh to a daily bill, providing a specific leverage point.

Given this clear pattern, the actionable strategy is to provide them with highly targeted advice on **efficient cooling**. Recommendations should be specific and impactful, focusing on thermostat management (e.g., programming for higher temperatures when away), improving home insulation to reduce thermal loss, and supplementing AC use with ceiling or portable fans to mitigate their primary energy cost driver.

5.3 The Minimalist Users

This group represents the model of energy efficiency. With extremely low and consistent consumption, they are the most energy-conscious users in the dataset. They are already achieving the desired outcome of low energy use. Therefore, the strategic goal for this persona is not reduction, but rather **positive reinforcement and long-term retention**. It is critical to acknowledge their excellent behavior to ensure they remain engaged and don't inadvertently shift into higher consumption brackets in the future.

The actionable strategy involves several components. First, actively praise their low energy use by acknowledging them as "Energy Champions" in communications, potentially coupled with small rewards or inclusion in prize drawings. Second, provide value through low-cost or no-cost tips that optimize their already-efficient lifestyle, such as advice on phantom power or efficient lighting. Finally, focus on "future-proofing" their efficiency by providing guides and information to help them choose the most energy-efficient options when their existing appliances eventually need replacing.

5.4 The Future-Focused Personas (Long-Term Trajectories)

These personas cut across all other groups and are used for proactive, strategic management based on each household's long-term energy journey.

- **The Escalators**: These 22 households are on a clear upward trend in energy use and represent a future risk for both themselves (higher bills) and the grid (increased demand). The appropriate action is to **engage them proactively**. This involves sending early-warning alerts about their rising consumption and offering a complimentary energy audit *before* their bills become a significant problem. This strategy shifts from reactive problem-solving to proactive intervention.
- **The Converters**: These 16 households are a clear success story, as they are actively reducing their energy use over time. The action here is twofold: first, **congratulate them** to reinforce their positive behavior and build customer loyalty. Second, treat them as a valuable source of information. They could be surveyed to understand what specific actions led to their success (e.g., a new appliance, a specific behavioral change), providing insights that can be used to craft more effective recommendations for other groups.
- **The Stable**: This predictable majority, whose consumption is neither increasing nor decreasing, forms the core of the user base. The action for this group is to **monitor their trends** and continue providing standard, persona-based advice (e.g., based on whether they are a "High Baseline" or "Steady Spender"). The primary goal is to ensure they remain stable and do not transition into the "Escalators" category over time.

6.0 Strategic Framework for Actionable Outcomes

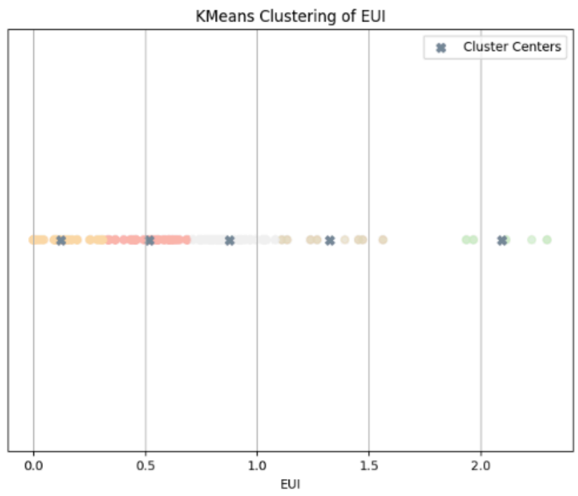
The culmination of this analysis is a strategic framework for driving real-world change by combining all layers of insight into targeted interventions.

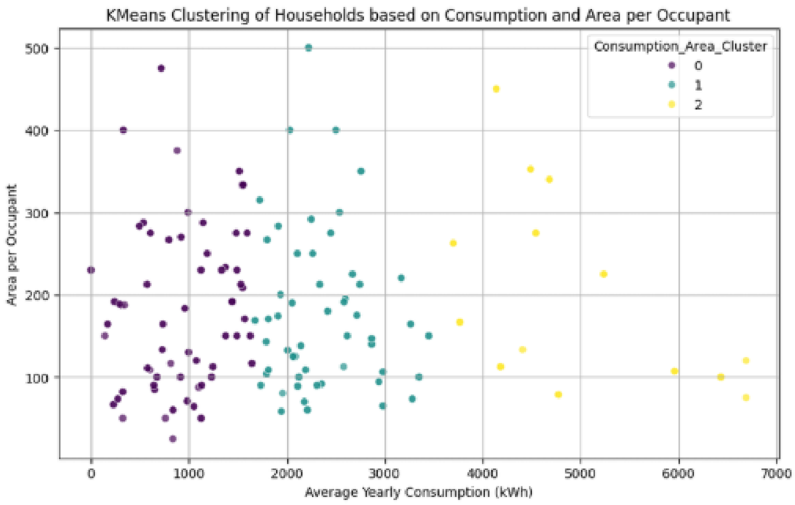
- **Personalized Advice**: We can now move beyond generic tips. An "Escalator" gets a proactive alert about their rising trend; a "High Baseline Consumer" gets advice on appliance efficiency; a member of the "Consistent Responders" (the COVID group) gets tips for managing work-from-home energy use; and a "Minimalist User" receives positive reinforcement.

- **Data-Driven Benchmarking**: Users can be motivated by comparing their consumption not to a vague "average," but to their true, behaviorally-defined peer group, making the comparison more relevant and impactful.
- **Measuring Policy Impact**: The trend-based personas ("Escalators," "Converters") serve as direct Key Performance Indicators (KPIs). The success of any conservation program can be precisely measured by its ability to transition households from the "Escalator" category to the "Stable" or "Converter" categories over time, providing a clear return on investment.

7.0 Feedback and Peer Learning

We initially looked at the data given and decided to use clustering on 2 dimensions : Area of occupant and average yearly consumption, although there seems to be some clusters formed, they are not good enough, so before we finalised to use this, we decided to work on incorporating other factors too. Also as told in the feedback, we also used graph clustering using DTW as a parameter for analysing results and we got to see some variety of clusters.





8.0 Conclusion

This multi-layered analysis successfully transformed a raw dataset into a sophisticated intelligence system. By integrating static profiling, temporal analysis, and predictive trend modeling, we have created a holistic and deeply contextualized understanding of household energy behavior. The resulting framework enables a strategic, proactive, and measurable approach to energy management, empowering targeted interventions that can drive meaningful and lasting conservation.