Article

Unveiling the Dynamics of Residential Energy Consumption: A Quantitative Study of Demographic and Personality Influences in Singapore Using Machine Learning Approaches

Jovan Chew [1,2], Anurag Sharma [2,*], Dhivya Sampath Kumar [1], Wenjie Zhang [3], Nandini Anant [4] and Jiaxin Dong [1]

1   Cluster of Engineering, Singapore Institute of Technology, Singapore 138683, Singapore
2   Electrical Power Engineering, Newcastle University in Singapore, Singapore 567739, Singapore
3   Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China
4   Agency for Science, Technology and Research (A*STAR), Singapore 138632, Singapore
*   Correspondence: anurag.sharma@newcastle.ac.uk

Abstract: In the pursuit of instigating a progressive transition towards a more sustainable future, policy officials all over the world are fervently advocating the use of energy conservation techniques targeted at residential customers. Keeping this in mind, a quantitative study was conducted in this work using the data from Singapore, which aims to investigate the relationships between a resident's pattern of energy utilisation and numerous demographic parameters as well as personality attributes. Moreover, the study was conducted with existing machine learning and data analytics approaches, including k-prototype unsupervised learning and statistical hypothesis tests. The obtained results denote a persuasive correlation between the consumption behaviour of the consumer for different appliances and factors such as income, energy knowledge, usage frequency, personality, etc. For instance, there is a higher probability of a consumer acting frugally and sparingly if they believe their energy consumption is insignificant. These findings can help policymakers identify the appropriate target populations for raising energy awareness in Singapore.

Keywords: data analytics; energy consumption behaviours; energy management; personality attributes; residential demand

1. Introduction

Global warming and climate change are among the most urgent challenges of our era. These challenges, inextricably connected with human energy utilisation and greenhouse gas emissions, demand immediate attention. An important portion of the world's total energy consumption is used by households; therefore, understanding how residential customers use energy will greatly improve energy efficiency and encourage energy savings to lower the risk of climate change. In the past few years, researchers have found that the energy consumption of an occupant can be influenced by several variables, including economic, social, environmental, climatic, and user-related features [1]. Furthermore, a preliminary analysis of associations among the Big Five personality characteristics, demographics, and energy-saving behaviours was conducted by [2], which reveals that individuals characterised by higher extraversion tend to exhibit fewer energy-saving efforts. Moreover, income was shown to be inversely connected with energy savings. Particularly, people with lower incomes may be more conscious of their energy usage [3,4] and more likely to make energy-saving decisions to save money [5]. Additionally, conscientiousness has a positive correlation with energy-saving behaviours [6]. However, there is very little empirical research on the interactions or relationships between energy consumption and household behavioural changes. Current research has focused on the relationship between

interpersonal factors [7], intrapersonal factors [8,9], and other factors [10–12] and household energy consumption behaviour. A key finding is that where individuals reside affects their views on energy problems, including how concerned they are about climate change and how they view energy difficulties.

In addition, a survey on participants' opinions of "most effective savings" revealed that 55.2% considered curtailment (e.g., turning off lights in unoccupied rooms) as the primary method of energy conservation, although it saves a relatively small amount of energy [13]. Only 11.7% cited efficiency (e.g., using efficient light bulbs) [14]. On average, participants overestimated the energy used and money saved for 15 sample activities by a factor of 2.8. According to [15], behavioural interventions can potentially reduce carbon emissions from residential energy usage by 20% over 10 years, equivalent to a 7.4% reduction in total US carbon emissions. A well-planned programme to raise public awareness of energy usage is likely to have a positive impact, as shown by [16], where higher numeracy scores and pro-environmental behaviours were linked to better energy perception. Additionally, energy social science is crucial for fieldwork investigations to enhance effectiveness and develop efficient energy conservation strategies.

An Individual's energy consumption behaviour is influenced by both behavioural and psychological variables, as well as objective and subjective factors. Objective elements include income levels, socioeconomic status, home features, and family size [17]. Subjective factors pertain to individuals' intentions and awareness. Household energy consumption, driven by various behavioural intervention strategies, is challenging to predict due to the unpredictability of human behaviour.

The residential sector accounts for approximately 14.9% of Singapore's total energy consumption, according to the Energy Market Authority (EMA) of Singapore [18]. However, there is a lack of studies on demographics and energy consumption behaviours in Singapore. This study explores the relationship between demographic characteristics, personality traits, and household energy behaviours in Singapore. To understand the relationship between energy characteristics and demographics relevant to home energy consumption, correlations are predicted using data analytics, machine learning, and statistical testing. The original contributions of this work can be outlined as follows:

- This study represents the first attempt to investigate how behavioural changes and demographic factors (e.g., income and housing type) interact to affect energy consumption among Singaporean households.
- Systematic and sophisticated data-driven analysis methods are employed to investigate the energy usage behaviours of residential consumers. This includes data management, energy calculation, a bottom-up approach, and machine learning methods to analyse the consumption behaviour of various household types in Singapore.
- Insights from this work could help policymakers make informed decisions about energy conservation. For example, there is a higher probability of consumers acting frugally if they believe their energy consumption is insignificant.

The remainder of the paper is organised as follows: Section 2 describes the methodology, including data collection and processing methods. The findings are presented and discussed in Section 3. Finally, the paper is concluded in Section 4.

2. Methodology

This study began with a review of the literature on domestic energy consumption, including the appliances used, demographic variables, interpersonal behaviours that affect energy consumption, and the relationships between these variables. Figure 1 illustrates the step-by-step flow of the proposed research methodology, and a detailed description of the steps is provided next.

2.1. Preliminary Study

Analysing the feasibility of a research study before conducting the main study is extremely important for achieving high-quality results. In other words, the pilot study

is essential to improving the questionnaire's quality and efficiency. The pilot study also gauged how well participants understood the survey questions. It is noted that while developing the questions, different response data types, such as nominal, ordinal, interval, and dichotomous scales, as well as equal-size intervals, were considered. After reviewing the inputs based on the pilot survey preliminary responses, the questions were revised to improve clarity and relevance. More specifically, changes were made to correct any confusing or unclear language, additional items were added, or redundancies were removed based on data trends and participant feedback. Thereafter, the revised questionnaire was tested with a small group of participants to ensure the changes made had the desired effect and were effective in data collection.
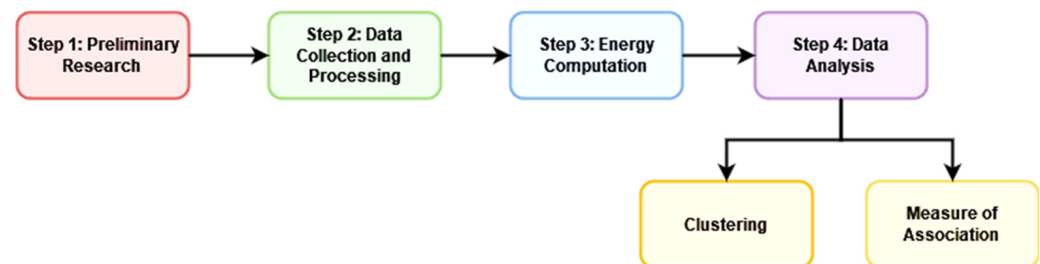
Figure 1. Research methodology.

2.2. Data Collection and Processing

An online quantitative survey was used to collect information on energy consumption behaviours in Singapore because there is very little open-access data available on this topic. Participants were asked three sets of questions. The first set focused on demographic variables (termed "demographic records"), while the other two related to appliance usage and frequency and residents' energy use (termed "behaviour records"). Demographic records included things like region, housing type, income per capita, and family size. Data on appliance usage included the types of appliances present in participants' households, including the oven, microwave, air conditioner, and other large-load appliances. The third set of questions on behaviours captured information such as energy habits and behaviours of participants, such as switching off the lights when a room is unoccupied, and the energy flexibility of different appliances. Table 1 provides an overview of the information collected in the online questionnaire. In this research, a convenience sampling method was used and hence data was collected from a population that is readily available and willing to participate [19]. The survey's participants are not limited to just members in easy accessibility and were specifically broadcasted to reach a wider range of demographics through social media platforms of friends, acquaintances, and their family members and friends. In terms of random sampling, the success rate was much higher instead because of the reduced bias and being more of a random sampling. It is noted that the success of any research is dependent on the representativeness of the sample to the target population. In this study, for a sample size of 151, the confidence level was determined to be at 90% with a margin error of 6.7%. These sets of questions were particularly chosen for reasons such as computing energy consumption based on demographics, appliance usage and behaviours, and analysing demographics and energy behaviour factors to determine the relationship towards energy consumption. For instance, one of the appliance record questions states: "On average, how many hours a day does your family spend watching the TV?". This aids in the computation of television appliance energy consumption for the household and provides information on the frequency of the TV. A behaviour record question states: "Do you switch off the lights when a room is unoccupied?". The user's decision, whether yes or no, would dictate their energy behaviour action with the lighting appliance. The dependent variable of interest is energy consumption (termed as the variable of "total_energy").

Table 1. Overview of survey questions.

Question Type	Information Collected	Number of Questions
Demographics	Housing type; number of family members; family members with different age; income per capita; region; energy consumption from bills	9
Appliances used and their frequency	Water heater; air conditioning; television; lighting; microwave; oven; washing machine	15
Energy habits and behaviours	Personality; switching off/on lights; using task lighting; unplug appliances; switch on aircon when feeling warm; keeping doors and windows closed when aircon switched on; opening the fridge for a longer time; fridge overload; energy thrifty; awareness; perception; flexibility of microwave; flexibility of aircon; flexibility of oven; flexibility of aircon; flexibility of washing machine; flexibility of water heater	16

2.3. Energy Computation

It is crucial to calculate the average energy consumption of various families to ascertain the participants' knowledge of their energy use. The household monthly energy consumption for a period of 12 months was obtained from the electricity provider along with the energy data obtained from the OEM for a more detailed analysis of the loading profile in different periods. These are accurate readings from the electricity billing and loading profiles, which reflect the electricity usage.

An HDB 5-room flat (author's home) was considered to identify and quantify various existing base and peak loads from the loading profile to help better understand the appliances used and the contributory factors of appliances towards energy consumption [20]. The selected appliances for this research were pivoted at the major household loads, which were stated in the questionnaire. They were mostly taken into consideration for the computation as they are the common house appliances of a modern household that reflect a user's daily behaviour and contribute largely to household energy consumption.

A bottom-up approach shown in Figure 2 was used to calculate the household energy consumption by building up the total load for each type of house and considering an estimated piece of a home appliance for the average monthly energy consumption in a statistically average manner. In addition, open data obtained was the average monthly household electricity consumption by dwelling type (in kWh) from the EMA's energy statistics from 2005 to 2020 [21]. Specifically, these are the average monthly household electricity consumption of different housing types in Singapore, such as 4-room flats, 5-room/executive flats, condominiums, and private housing. This data provides an accurate reference foe monthly energy consumption by dwelling type for computed energy consumption.
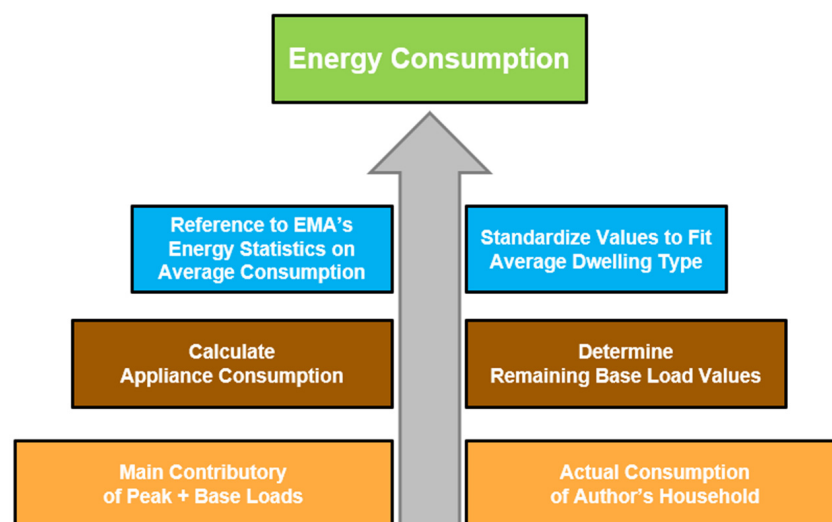
Figure 2. Bottom-up approach for energy computation.

The average consumption will be verified along with the actual electricity consumption taken from the electricity bills over a span of 12 months. The loads that are specified in the questionnaire were calculated according to the household's daily activity. For example, "On average, how many hours a day does your family spend watching TV?". This piece of information helps in estimating the television appliance energy consumption in kilowatt-hours. Similarly, for the other appliances, the energy consumption was estimated. Once the calculated base value was verified with the energy profiling and accurate billing for a few profiles and is referenced to the EMA's energy statistics data and average monthly electricity consumption data of dwelling housing types to gauge the calculated base value, these are the remaining base loads of different housing types that are complex to quantify and determine—Wi-Fi routers, rice cookers, laptop chargers, printers, and more. This is because every household has different quantities and types of appliances; some would have a printer or air purifier, while others do not. The base value would be standardised, which fits the most specified average consumption of dwelling types. Participants' energy consumption will be estimated as a result. Moreover, since the calculated energy consumption is an estimate, a range/interval of actual consumption was supplied in the questionnaire, as well as a factor of up to the $\pm 20\%$ threshold of the interval.

2.4. Data Analysis (Clustering)

The quantitative research data used in this study were not time-series data but rather average energy consumption data, where each participant's income level, housing type, appliance usage, etc. were the only variables that were unique to them. As a result, the strategy adopted was to distinguish between the group identities for the various participants' energy behaviours and cluster those behaviours according to how closely or how disproportionally they connect to their energy habits.

The k-Prototype cluster algorithm was selected due to its simplicity and its ability to provide the functionality of both the K-Means and K-Modes techniques for mixed data types as compared to other algorithms [22–24]. K-Prototype is based on a set of n observations, $X = \{X1, X2, \ldots, Xn\}$. $Xi = \{Xi1, Xi2, \ldots, Xim\}$ is consisted of *m* attributes; *mr* is the numerical attributes, *mc* is the categorical attributes, and $m = mr + mc$. The objective of this clustering is to divide n observations into *k* different clusters, $C = \{C1, C2, \ldots, Ck\}$ where *Ci* represents the i-th cluster centre. The distance $d(Xi, Cj)$ between *Xi* and *Cj* can be calculated:

$$d(Xi,\ Cj) = dr(Xi,\ Cj) + \gamma dc(Xi,\ Cj) \tag{1}$$

In addition, two cluster criteria, the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC), were selected and applied as cluster solutions. Both criteria estimate the quality of the number of parameters or the number of clusters using a penalty term. They were selected to determine how appropriately the *k* or parameter value fits the data set without overfitting or underfitting.

$$\text{AIC} = 2k - 2\ln(L) \tag{2}$$

BIC is similar in nature to the AIC; however, BIC penalises more heavily on parameter complexity.

$$\text{BIC} = \ln(n)k - 2\ln(L) \tag{3}$$

where, *L* is the maximum value of the likelihood function of the parameter, *n* is the number of data points, and *k* is the number of parameters or clusters to be estimated.

Moreover, another criterion method, the Silhouette Coefficient [25], was used to ascertain the reliability and validity of the cluster solutions. This determines whether the number of clusters was optimal and served as a synthetic indicator to assess the overall quality of the cluster. In this case, the overall quality of the K-Prototype was measured by computing the silhouette coefficients of each point to measure the similarity of each point on its cluster (this is the measure of cohesion) as compared to other clusters (this is the measure of separation). The metric for the silhouette scoring function was set to the default "Euclidean"

distance measurement, which is appropriate for continuous data. However, the inclusion of discrete data may pose a barrier to "Euclidean" distance measurement due to the "curse of dimensionality" [26]. To address this, sequence matcher, a class in the Python module "difflib", was used to compare similarities and locate matching elements. This analyses and evaluates how similar an observation's energy behaviour is to that of other observations in the same cluster. This method was used to evaluate performance by looking at the highest average percentage of matched elements across all clusters. To validate the scores, the average silhouette score was used to test numerical factor performance, the default metric 'Euclidean' for distance measurement, and a sequence matcher to measure similarities elements for categorical factors independently. The silhouette coefficient formula for each point is given by the following:

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), \ a(i))} \tag{4}$$

The points were summed and averaged to determine the silhouette score, $S(i)$.

$$S(i) = \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{5}$$

where, $a(i)$ is the average distance of the point i to all other points within its cluster, and $b(i)$ is the average distance of the point i with all the points in the nearest cluster to its cluster [25].

2.5. Data Analysis (Measure of Associations)

Descriptive and inferential statistics were used to identify the association between the different variables such as demographics, behaviours, personality traits, and appliance frequency and usage. These are performed using two statistical hypothesis tests: the Chi-square test of independence [27] and one-way analysis of variance (ANOVA) [28]. The Chi-square test of independence is a non-parametric, free test used to examine the relationship between two or more categorical variables. The ANOVA is a parametric test used to assess whether the differences in the averages of three or more groups are statistically significant. The descriptive statistic test is suitable for both nominal and ordinal types of categorical data. The observed and predicted actual frequencies for a categorical variable count were compared to determine if there is a disparity or relationship associated such that one value of the variable can be used to predict the other variable when there is no sampling error in the data population. Variables with a statistically significant difference are subjected to extra analysis for post-hoc testing to identify where the differences truly exist or which categories are contributing to these differences.

The chi-square value can be obtained using the following:

$$x^2 = \sum \frac{(Oi - Ei)^2}{Ei} \tag{6}$$

$$E = \frac{\text{total rows} * \text{total columns}}{\text{number of samples}} \tag{7}$$

where, Oi denotes the observed values, Ei is the expected values and x^2 is the Chi-square to determine and E is the sum of all expected values.

$$df = (\text{row} - 1) * (\text{column} - 1) \tag{8}$$

Calculating the degree of freedom, df and Chi-square value (x^2) will determine the p-value from the chi probability. In the event of multiple simultaneous tests of multiple

subgroups for testing, it is necessary to control the family-wise error rate (FWER) which is given as follows:

$$\text{FWER} = 1 - (1 - \alpha)^k \tag{9}$$

where, k is the number of tests performed. α is usually 0.05. Bonferroni Correction method is used to prevent a higher probability of incorrectly rejecting the true null hypothesis, which is the chance of finding false positive result (type 1 error).

$$\text{Bonferroni} - \text{corrected } \alpha = \frac{\text{original } \alpha}{k} \tag{10}$$

Bonferroni correction divides the per analysis alpha rate, α by number of statistical analyses performed, k. k can be determined with the product of number of rows with the columns. If any of the observed p-value after dividing k is less than 0.05, it is declared to be statistically significant. The new calculated α according to Bonferroni correction, is applied to the FWER formula to determine the new FWER in percent.

Under post-hoc testing, the adjusted standardized residuals were computed between the observed and expected values. Calculating the residuals and identifying cells with difference in residuals would indicate significant difference between observed and expected values for a cell. The larger the residual, the greater the contribution of the cell to the magnitude of the resulting chi-square obtained value. As stated in [29], "a cell-by-cell comparison of observed and estimated expected frequencies helps us to better understand the nature of the evidence" and the discrepancy is larger in cells with significant residuals than we would assume if the variables were independent.

$$\text{Adjusted residual} = \frac{(O - E)}{\sqrt{E * \frac{(1 - \text{RowMarginal})}{n} * \frac{(1 - \text{ColumnMarginal})}{n}}} \tag{11}$$

The row marginal refers to the total rows of the cells and the column marginal refers to the total columns of the cells whereas n refers to the total number of cases across all cells. The adjusted residual's denominator in (11) is the estimated standard error instead of the estimated standard deviation of the residual.

The second hypothesis test used here was ANOVA. The ANOVA test assesses whether the averages of more than two groups are statistically different from each other. This is appropriate for comparing averages of categorical with numerical variables. As a parametric test, it is critical to ensure the assumption checks are met to determine if the sample comes from a Gaussian distribution if the data set conforms to the assumption of homogeneity of variance for an unbiased F-statistic and whether a null hypothesis is being wrongly rejected. One assumption check performed was using a graphical method, QQ plot to examine whether two data sets are from a Gaussian distribution. The Gaussian distribution is a probability distribution that is exclusively determined by two data set parameters: the sample mean and standard deviation. Because of this property of the distribution, the analysis can be made easier and any variable with a Gaussian distribution may be projected with better accuracy. Equations (12)–(15) and Table 2 show the indices used to determine the value of p.

$$SSB = \sum_{j=1}^{p} \left(\overline{X}j - \overline{X} \right)^2 \tag{12}$$

$$SSW = \sum_{j=1}^{p} \sum_{i=1}^{nj} \left(Xij - \overline{X}j \right)^2 \tag{13}$$

$$dfB = k - 1 \tag{14}$$

$$dfW = n - k, \tag{15}$$

where nj is the number of observations in the *j*th group, Xj is the mean value in the *j*th group, $\overline{X}$ is the overall mean value, Xij denotes the *i*th observation in the *j*th group, n is the total number of observations and k is the number of groups. The X^2 value represents the amount of difference between the observed and expected values if there were no association in the population.

Table 2. ANOVA Test to determine p.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic	p
Between Groups	SSB	dfB	$MSB = \frac{SSB}{dfB}$	$F = \frac{MSB}{MSW}$	$P(F)$
Within Groups	SSW	dfw	$MSW = \frac{SSW}{dfw}$		
Total	SST	dfT			

3. Main Empirical Findings

The main empirical findings from this research are discussed here.

3.1. Cluster Selection

K-Prototype unsupervised learning algorithm was applied to the proposed data to cluster energy behaviours. Before clustering, PCA was used as a dimensionality reduction and visualization technique. The data was plotted in 2-D plane along the *x*-axis of PC1 and *y*-axis of PC2. PC1 captures the most variance as compared to PC2. In other words, data points would more likely be considered a cluster along PC1 axis instead of PC2 axis of equal distances, and the closer the distance, the more correlated the data points share amongst one another, this tells us more about how correlated data points are positioned when they are close together. The PCA plot before the clustering technique was applied as shown in Figure 3.
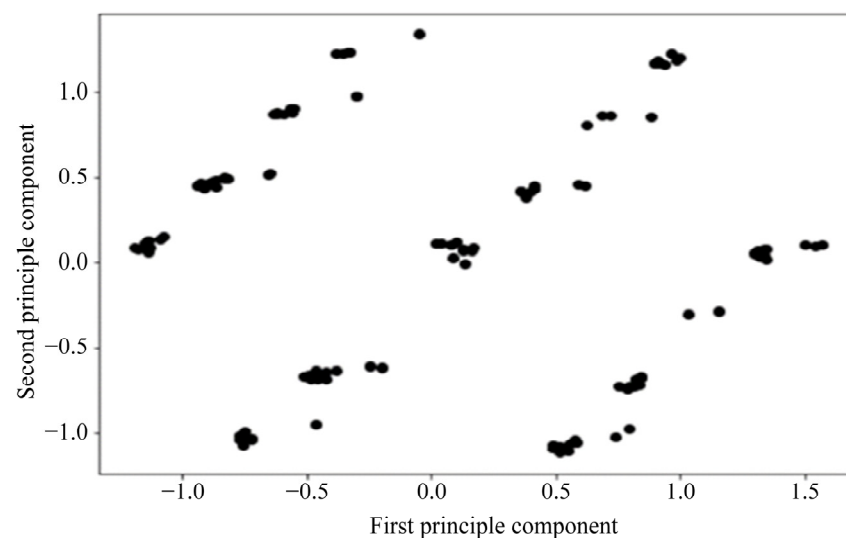
Figure 3. Plot before k-prototype cluster.

Observing the PCA plot (see Figure 3), the data points are positioned in a partitional clustering manner, which is very suitable using the K-Prototype as compared to hierarchical and density-based clustering. The proposed approach was to perform the clustering criterion using both AIC and BIC. This is because AIC frequently makes the mistake of selecting the number of parameters that is too large, whereas BIC frequently makes the mistake of selecting the number of parameters that is too small. When compared to AIC, BIC has a more consistent estimation. Underfitting is the most common error when there are fewer parameters, hence criteria with lower underfitting rates, such as AIC, frequently appear to be preferable [30]. Similarly, BIC would be preferred for larger parameters with

the common error of overfitting. In this case, BIC was used to note the minimum number and AIC to note the maximum number; the recommendation is to use both the AIC and BIC to select a point at which both criteria agree and overlap. The minimum point for BIC was found to be around 24, where the trend stopped decreasing and began to increase as the cluster number increased. However, there was no minimum point for AIC. As shown in Figure 4, the optimal number of clusters was identified to be between 12 and 15 using the elbow method. The AIC and BIC indicate that the ideal clusters are between 12 and 15. Furthermore, the highest silhouette index score of 0.2017 was achieved for 15 clusters validating that there are 15 distinct energy behaviours observed from the dataset as shown in Figure 5. At 15 clusters, sequence matcher also yields the highest accuracy at 0.89 as shown in Table 3.
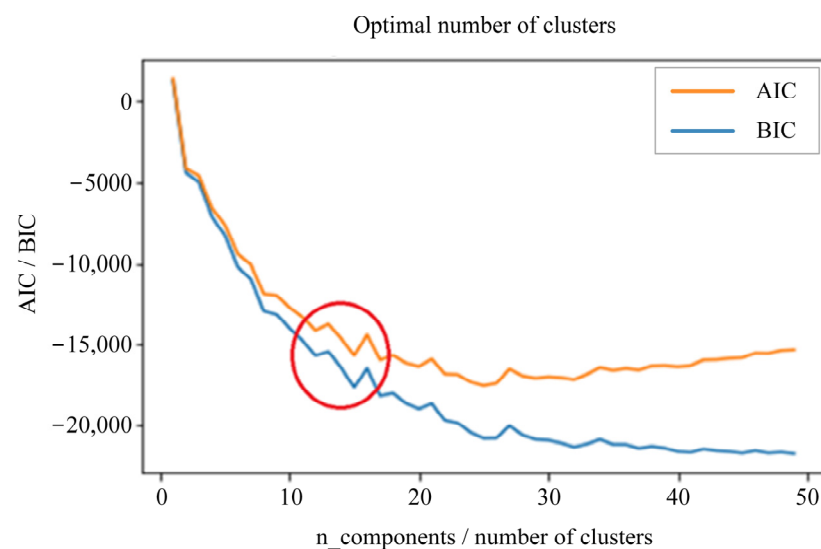

Figure 4. Trends of AIC and BIC. The red circle highlights the optimal number of clusters to be between 12 and 15.

Table 3. Scores for sequence matcher.

Cluster	Sequence Match Score
2	0.7889
3	0.7185
4	0.7324
5	0.7739
6	0.7793
7	0.8091
8	0.8368
9	0.8315
10	0.8656
11	0.8763
12	0.8816
13	0.8720
14	0.8816
15	0.8901
16	0.8784
17	0.8720
18	0.8742
19	0.8795

Overall, the results appeared to be satisfactory, with clusters developing when scatter points were close together. Nonetheless, certain clusters emerge in few spots along the data scatterplot, indicating that either the data is inaccurate, or the cluster missed some

information. A supposition is that the numerical values have a bigger variance and there is a need to consider both factors of categorical and numerical values. While categorical data are mostly matched in terms of behaviour and normalization was performed, numerical data may have bigger variance differences, resulting in some clustering misses. The higher the cluster, the more unique the behaviour, but fewer groupings per cluster preclude analysis. The result may be improved if the data set is greater, allowing for better classification, higher accuracy, and more conclusive conclusions. However, because the clusters have a slight deviation, a range of thresholds for reduced bias in the findings should be considered.
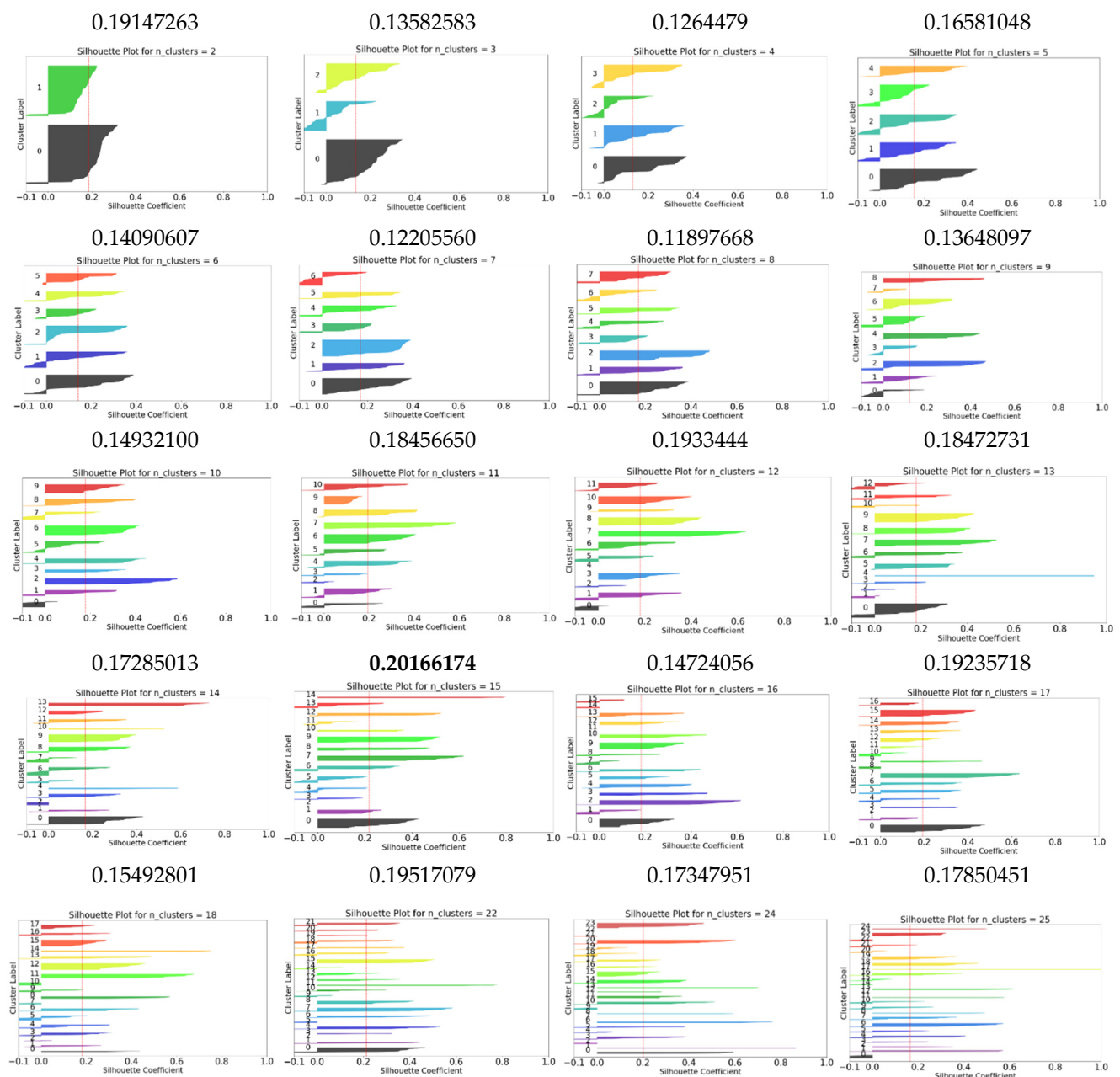
Figure 5. Silhouette Plots of n_clusters [2,25] with their respective average silhouette scores.

3.2. Statistical Hypothesis Test

The chi-square test of independence was performed for various variables, excluding those where the sample data did not meet the assumption check (i.e., expected count greater than 5) or failed the Bonferroni correction, regardless of statistical significance. The ANOVA test was used to compare averages of numerical and categorical variables. Findings from

the chi-square test were compared with the dependent variable of interest, household energy consumption, as both hypothesis tests may yield related results.

Results are presented only for variables that met the assumption checks and showed statistically significant differences. Categories with insufficient observations, such as 2-room flats (1 observation) and landed properties (4 observations), were excluded. The goal of the analysis is to identify specific characteristics related to energy consumption and to investigate variables of interest, particularly those with compelling results in the correlation analysis.

Table 4 shows the variables with a significant difference from one another. The chi^2 value represents the amount of difference between the observed and expected values if there were no association in the population. Using the Bonferroni correction for FWER to control the type I error, the result was evaluated and displayed in a table form. It was discovered that because there were numerous groups to compare for two variables, some of the expected counts were lower for certain groups and this violates the assumption. Even though there can be significant differences in correlation, the result may be ambiguous and inaccurate. This problem can be solved by combining some categories or by increasing the sample size.

Table 4. Significance results.

Variable 1	Variable 2	chi^2	p	Result
b_thrifty	b_perceive	21	0.00032	HSD
b_lights_task	b_flexi_lights	7.57	0.00592	SD
b_lights_pres	b_flexi_lights	3.82	0.050540	NSD
aircon_frequency	b_flexi_aircon	26.09	0.000009	HSD
b_aircon_temp	aircon_frequency	18.94	0.00028	HSD
microwave_frequency	b_flexi_microwave	27.94	0.000038	HSD
b_flexi_microwave	b_flexi_oven	16.63	0.000046	HSD
bills	aircon_frequency	16.43	0.000924	HSD
bills	b_aircon_temp	4.28	0.038664	SD
bills	b_flexi_aircon	8.49	0.003572	HSD
b_perceive	actual_energy	38.41	0.000006	HSD
b_awareness	personality	6.11	0.036120	SD
b_awareness	pci	6.01	0.111085	NSD

NSD: No significant difference; SD: Significant difference, correlation is statistically significant at $p < 0.05$; HSD: Highly significant difference, correlation is statistically highly significant at $p < 0.001$.

3.3. Inferences

The F-score is a statistical significance metric; a higher F-score indicates a stronger relationship between the variables. The critical value is a threshold used to determine whether a statistical result is significant, based on the chosen significance level ($p < 0.05$). It represents the minimum value that the F-score must exceed for the result to be considered statistically significant. Similarly, a larger chi-square value indicates a stronger relationship between the variables, determined by comparing the observed and expected frequencies of events to identify significant differences. Following the summarization of the results from both hypothesis tests, conjectures were made about the outcomes, which are detailed in Tables 4–6. Table 5 presents the results of the post-hoc test for multiple variables, where 'Obs' stands for observed, 'Exp' stands for expected, and 'Adj. Res' stands for adjusted residual.

Table 5. Significance of Post-hoc Test. (Part of the Table is also present in the Appendix A).

			b_perceive	High Usage
b_thrifty	No	Obs		11
		Exp		5.1
		Adj. Res		3.2
	Yes	Obs		4
		Exp		9.3
		Adj. Res		−2.5

Table 5. *Cont.*

			b_perceive	High Usage
actual_energy	<300 kWh	Obs		3
		Exp		7.9
		Adj. Res		−2.5
	>300 kWh	Obs		15
		Exp		10.1
		Adj. Res		2.5

			b_flexi_lights	
			Non-Flexible	Flexible
b_thrifty	No	Obs	66	24
		Exp	59.8	30.2
		Adj. Res	2.4	−2.4
	Yes	Obs	23	21
		Exp	29.2	14.8
		Adj. Res	−2.4	2.4

			aircon_frequency	
			Not in use/NA	Frequent
b_flexi_aircon	Non-Flexible	Obs	4	47
		Exp	11.3	33.4
		Adj. Res	−3.3	4.7
	Flexible	Obs	20	24
		Exp	12.7	37.6
		Adj. Res	3.3	−4.7
b_aircon_temp	Yes	Obs	5	50
		Exp	14.3	42.4
		Adj. Res	−4.3	2.7
	No	Obs	19	21
		Exp	9.7	28.6
		Adj. Res	4.3	−2.7
bills	Did check	Obs	20	18
		Exp	11.8	28
		Adj. Res	3.4	−3.4
	Did not check	Obs	10	53
		Exp	18.2	43
		Adj. Res	−3.4	3.4

Table 6. Significance results with *total_energy* as the target dependent variable.

Variable 1	F-Score	Critical Value	p (Two-Tailed)
house_type	15.6104	2.66794	2.69×10^{-12} **
pci	6.6347	3.21900	0.000173 **
b_aircon_temp	8.34173	5.14136	0.009644 *
b_flexi_aircon	17.5138	5.14136	0.000136 **
bills	6.6048	5.14136	0.020563 *
b_awareness	0.34853	5.14136	0.434947
b_thrifty	0.01176	3.78709	0.98831
b_perceive	1.83826	3.78709	0.162886
b_lights_task	0.17495	5.13212	0.676385
b_flexi_lights	0.24728	5.13212	0.619764
b_flexi_microwave	1.49678	5.13212	0.223193
actual_energy	11.7771	2.8792	2.88×10^{-8} **
personality	0.06840	5.13212	0.794059

* Correlation is statistically significant at $p < 0.05$. ** Correlation is statistically highly significant at $p < 0.001$.

3.3.1. Thriftiness—Perceptiveness (or Perception)

Users who indicated non-thrifty behavior were more likely to perceive high energy consumption and less likely to perceive moderate or low usage. This is based on present behavior rather than continuously changing behavior. When users perceive their energy consumption as low, they are more likely to engage in thrifty and economical behavior.

3.3.2. Flexibility of Lighting Appliance—Perceptive

When users perceive a lighting appliance as non-flexible, they are more likely to avoid using task lighting. Conversely, when users perceive a lighting appliance as flexible, they are more likely to adopt task lighting behavior. It was also found that the flexibility of lighting appliances has no relationship with the behavior of switching off lights when a room is unoccupied.

3.3.3. Awareness Behaviour vs. Personality

There was no significant difference between 'b_awareness' and 'personality' after applying multiple simultaneous tests with Bonferroni adjustment. Whether a household is primarily introverted or extroverted shows no correlation with energy consumption behavior or awareness.

3.3.4. Flexibility of Aircon vs. Frequency Use of Aircon

Users who are flexible with their air conditioning use are more likely to limit their usage compared to the average. Those who use air conditioning frequently are more likely to be rigid in their usage. In other words, the flexibility or rigidity of a user regarding a specific appliance reveals a lot about their frequency and pattern of usage. If an air conditioner is perceived as inflexible by the user, it is more likely to be used frequently rather than occasionally.

3.3.5. Switching on Aircon When Warm vs. Frequency Use of Aircon

Users who are rigid in their use of air conditioning are more likely to turn it on when they feel hot, placing a higher value on comfort and showing less inclination to change their behavior towards appliance use. Therefore, knowing whether a user uses an air conditioner frequently, infrequently, or never provides insight into their flexibility or inflexibility in using the appliance and whether they prioritize comfort in hot weather.

3.3.6. Awareness on One's Electricity Bill vs. Frequency + Behaviour of Aircon Appliance vs. Other Appliances + Different Appliances vs. Energy Consumption

The variable 'bills' is observed to have a significant association with air conditioning appliances. Being energy-conscious shows a higher positive correlation with the frequency and behavior of air conditioning use compared to other appliances such as microwaves, lighting, and washing machines. Air conditioning is widely regarded as the primary appliance for energy conservation or curtailment. If users see a need to reduce their energy consumption, they are likely to modify their air conditioning usage first, before considering other appliances.

3.3.7. Flexibility vs. Frequency vs. Behaviour of Appliance

The versatility of an appliance is closely related to its usage and frequency, as well as its behaviour. For example, if a user indicates that they are not flexible about watching television, the frequency with which they use the appliance is generally longer in hours.

3.3.8. Awareness Behaviour vs. Income per Capita

An intriguing finding was that the variable 'b awareness' had no correlation with the variable 'pci', which is the level of income. This was required for the study because logically, there should a relationship between the two variables. This is because, logically, the lower the level of income, the more likely an individual will experience financial constraints and

will be more mindful of one's electricity usage to limit or reduce it to save money. One potential explanation for this is limited data involved. Further study was conducted in the data analysis subsection.

3.3.9. Awareness Behaviour vs. Energy Consumption

Few variables from ANOVA test were found to have failed to reject the null hypothesis, which states that the group means from the independent variables have no significant difference towards the target dependent variable. A one-way between subjects ANOVA was conducted to compare the effect of (IV) Awareness Behaviour on (DV) Energy Consumption in two different groups of condition: Having energy awareness and not having energy awareness. There was a not a significant effect of IV on DV at the $p < 0.05$ level for the two conditions [$F(1, 94) = 0.34853$, $p = 0.434947$].

When it comes to 'b awareness', there is no association with total energy usage when a user is aware of his or her own energy consumption level, whether it is high, moderate, or low. This is especially unexpected given that plausibly, as well as in other studies, there is a link between awareness and energy use. If the user is more aware of energy conservation strategies or is aware of energy usage, there is a better chance that energy will be conserved, whether the user is an environmentalist or intends to save money and energy. Similarly, 'b_aircon_temp' should be associated with total energy consumption, as it has previously been determined that the behaviour of air conditioning appliances is correlated with frequency, flexibility, and usage, which in turn influences the total energy consumption of a household.

Some of the inferences that need to be reassessed with respect to Sections 3.3.8 and 3.3.9 are as follows:

- There is no direct relationship between the variable 'b_awareness' and 'total energy'.
- It is possible that the data displayed is based on average, rather than changes in behaviour per participant, because each participant has different behaviours and factors such as demographics, which cannot be directly compared.
- The relationship may not be based on a direct relationship, but on the interaction of multiple features, resulting in an indirect relationship. A supposition is factors such as demographics, behaviours, how thrifty the family is, or how they perceive one's energy consumption will change how energy awareness impacts the total energy consumption.
- If there is a relationship between the two variables, this can be investigated further in the next two subsections.

Some of the variables can be further studied to determine relationship and there were also those without a direct relationship with each other, this type of association can be difficult to quantify. These may be related to the unique behaviour of an individual, or there may be multiple factors and features involved which led to an indirect relationship as part of the conjectures discussed in the previous subsection. The ANOVA test results had indicated that there was no significant difference between the group means of energy awareness towards total energy. However, from a logical point of view, this would be somewhat related. A possibility could be that the data displays based on average, and not based on changes in behaviours for participants. In other words, every participant has different forms of behaviour and factors contributing to differences such as demographics.

3.4. Energy Awareness

It is assumed that when actual energy consumption was accurately stated by the participant in the questionnaire and matched with the computed total energy consumption, the user is aware and has known about their energy consumption and its level of usage. However, in case of a mismatch with total energy consumption, the user may not be aware or may have given erroneous answers in the questionnaire, thus they are most likely labelled as unaware. The variable for matching value is termed as 'bills'.

With this information, a heat map was created as shown in Figure 6, and the results show that having energy awareness has a modest correlation coefficient of 0.2, whereas not having energy awareness resulted in a result of -0.032. Similarly, both *p*-values were observed to have a drastic difference. The hypothetical assumption is that consumers who are aware of their energy consumption will make a conscientious effort to reduce their energy consumption as seen with the positive correlation with energy awareness. Considering 'bills' and 'b awareness' are interrelated; it was established that being truly energy conscious entails two criteria: knowing one's electricity bill and being aware of one's degree of energy usage.
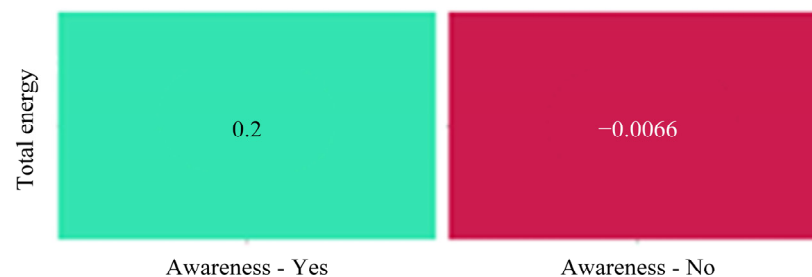
Figure 6. Heat map for matching and mismatching awareness.

Figure 7 depicts a multiple box plot describing how being conscious or oblivious of energy affects how one views their energy usage in relation to total energy consumption. In terms of median points, people who are aware appear to expend less energy than those who are not aware. For all participants living in a 4-room flat, the average value of all users who are oblivious of energy usage is 507.675 kWh, while the average value of all users who are aware of energy usage is 421.535 kWh. For all participants living in a 5-room flat, the average value of all users who are oblivious of energy usage is 532.64 kWh, while the average value of all users who are aware of energy usage is 432.285 kWh. Data from 4-room flats and 5-room flats were chosen because they had the most data collected of any house type. In general, there was a reduction in energy consumption when users were aware of their energy consumption as opposed to when they were not aware of their energy consumption.

According to observations in Figure 7, there is significantly reduced total energy consumption when there is energy awareness. When comparing energy awareness to no energy awareness, there were fewer fluctuations, outliers, and the interquartile range was more condensed across all PCI levels except PCI > $4000, where the difference was negligible, indicating that PCI > $4000 has little or no influence in decision making. When comparing no energy awareness to energy awareness for PCI of $1000 or less, a significant deviation was found; this can be interpreted as users with energy awareness and a PCI of $1000 or less being more motivated and impactful in decision-making and making a deliberate attempt to cut consumption.

It has been determined that having energy awareness generally led to a user reducing energy consumption conscientiously. However, this is not always the case because the correlation coefficient was not substantial as observed from the above results. Energy awareness when coupled with other factors addressed indicates that a specific variable may be a group relationship, or an indirect correlation might indicate whether one would conscientiously reduce consumption. If a user is thrifty, he or she is labelled as "Normal", and would most likely make a conscientious effort to reduce usage when approached with energy awareness.

If a user has energy awareness and perceives his or her energy level to be low, he or she may take action to limit consumption because one is more inclined to be thrifty. Similarly, if a user thinks his or her energy level is high, he or she is unlikely to be thrifty. As seen in the preceding analysis, an additional level of PCI factor would be considered in that case, as PCI influences decision-making in energy reduction, particularly if PCI is less

than $1000, there is a very high likelihood that the user would make a conscientious effort to reduce consumption regardless of circumstances.

The previously formed cluster behaviours, as well as demographic characteristics such as house type, income level, and total family members, were chosen and grouped to produce a data frame for analysis. This effectively groups the individual and ensures that comparison and demographics are not biased. Each of the variables, perception, energy thriftiness, and awareness, was looped through the data to check the validity of the hypothesis assumptions. The purpose is to validate the assumptions made herein based on changes in participant behaviour to see if energy consumption is influenced. An abundance of data was required for better analysis, however just a handful were accessible and selected for this test.
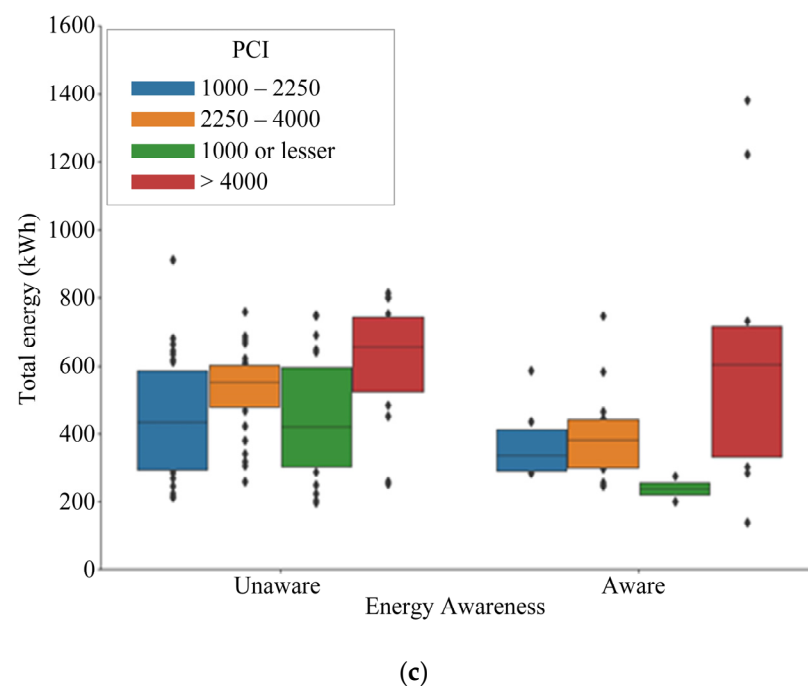

(**a**)

(**b**)

Figure 7. *Cont.*

(**c**)
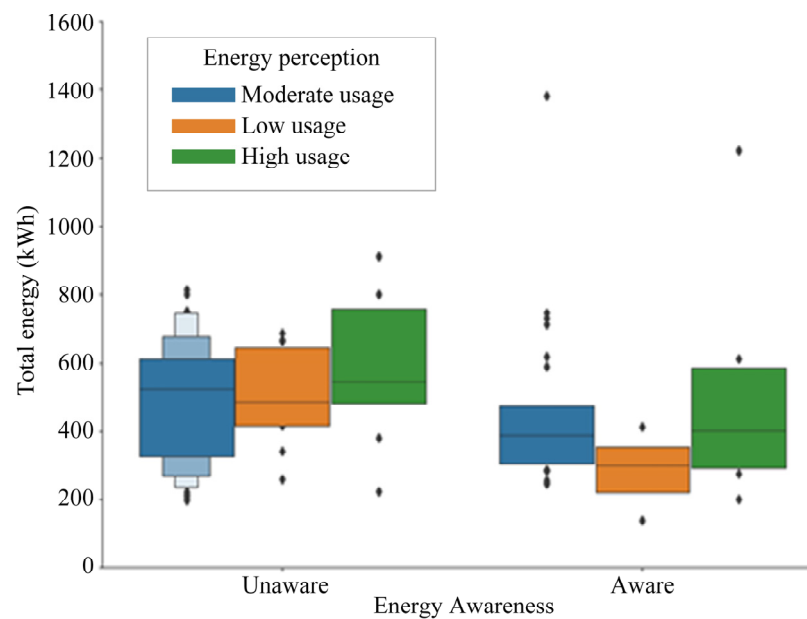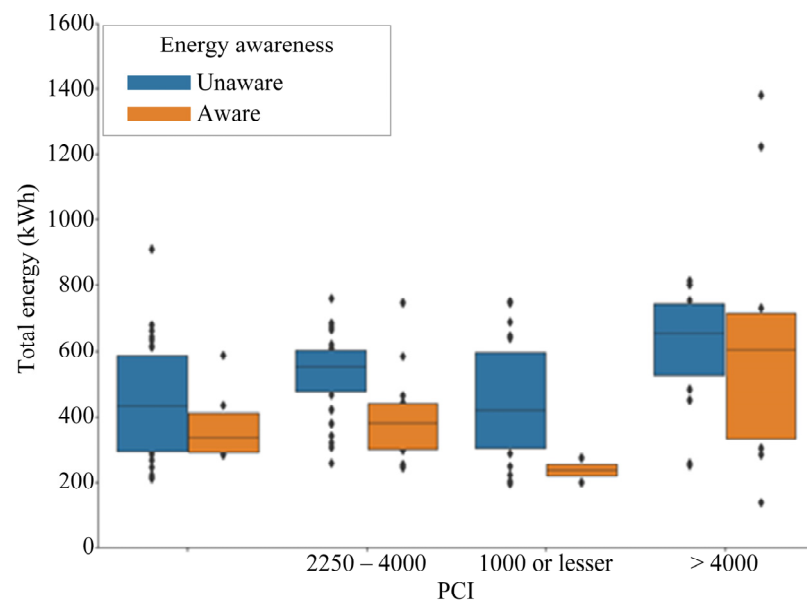
Figure 7. Energy awareness vs. PCI vs. energy consumption. (**a**) Plot 1, (**b**) Plot 2, (**c**) Plot 3.

Based on the findings shown in Table 7, the hypothesis assumptions were correct. Based on the probability of statistical assumption, observation 4 with thriftiness "NS" should be "N". The final observation produced an ambiguous result due to the involvement of income level, perception, and thriftiness variables. The perception is high usage, which suggests they are unlikely to be thrifty, and when a factor of PCI > $4000 is considered, the result is a "NAT." In any case, more data meant more conclusive results for this finding. A decision tree and ANN multi-classification algorithms can be employed for future training and prediction tasks for this type of behaviour or lifestyle, as well as the research of other energy consumption behaviours in Singapore's demographics context, such as education level and region of stay.

Table 7. Data comparison test for validation of assumptions.

Demographics	Perception	Thriftiness	Before Energy Awareness	After Energy Awareness	Reduction	Comments
4-Room Flat, PCI > 4000	MU	Y	N	Y	Y	"Nom"
5-Room Flat, PCI 1000–2250	MU	Y	N	Y	Y	"Nom"
5-Room Flat, PCI 2250–4000	MU	Y	N	Y	Y	"Nom"
Condominium PCI > 4000	HU	NS	N	Y	N	"NAT"

MU: Moderate usage, HU: High usage. Y: Yes; N: No; NS: Not Sure. Nom: Normal; NAT: No Action Taken.

Furthermore, the concept of "Diffusion of Innovation" [31] is relevant here. It explains why some people adopt a particular behaviour more quickly than others. In general, the population is divided into five groups: innovators (2.5%), early adopters (12.5%), early majority (34%), late majority (34%), and laggards (16%). Early adopters and innovators are more responsive to changes in the importance of energy conservation, whether it is related to their savings or pro-environmental behaviour for the wellness of the climate. Laggards are a small minority of the population who are resistant to change because they may be secluded and have little exposure to facts and media. They are traditional and fixated with the present; they can be of any age but are more common in the medium to senior age range. In this situation, a fraction of these people may be oblivious of energy conservation knowledge or even be unsure of their own electricity information, resulting in "NAT" where they are unconvinced by variables attributing to changes.

3.5. Energy Perception

The lower quartile, interquartile range, and upper quartile were calculated using the EMA's energy statistics of monthly energy consumption for dwelling type from 2005 to 2020 as shown in Table 8. The lower quartile denotes quartile 1 of the 25th percentile, which is the data's lowest 25%. The upper quartile corresponds to quartile 3 of the 75th percentile, which is the data's highest 25%. Finally, the difference between the upper and lower quartiles is represented by the middle 50th percentile, also known as the interquartile range. This was adopted as a true reference for energy perception, with a value less than the lower quartile representing a low consumption range, a value in the interquartile range representing a moderate usage range, and a value greater than the upper quartile representing a high usage range. For instance, if a user has an energy consumption of less than 144 kWh and perceives it as low consumption, he or she is correct because it falls within the lower quartile range.

Table 8. EMA's energy statistic with Q1, Q2 and Q3.

House Type	Min	Max	Avg	Q1	Q2	Q3
1-Room/2-Room	123.8	215.5	155.3	144.0	154.9	164.5
3-Room	232.9	360.3	275.6	260.1	277.4	291.5
4-Room	314.1	507.0	380.2	355.7	382.4	403.6
5-Room/Executive Flat	380.8	628.6	465.1	435.9	468.8	494.9
Condominium/Private Housing	435.4	796.0	625.0	556.5	625.4	690.9
Landed Property	1025.1	1503.0	1197.9	1135.5	1201.7	1258.1

Q1: Lower Quartile; Q2: Middle Quartile; Q3: Higher Quartile.

It was discovered that of the 142 participants (after data processing), whether they were energy aware or not, 26 accurately perceived their energy consumption level, whereas 116 incorrectly perceived their energy usage level. This equates to a perception accuracy of 18.3 percent. The reason for this accuracy is that each user's perception is ultimately unique and is dependent on their own perception, which is notably related to whether they are regarded as thrifty or not in their energy use. If they are regarded as thrifty, the perception should be low, as previously concluded based on the hypothesis findings. Without knowing the true perception range, individuals perceive whether they are regarded as thrifty or not in terms of energy use.

When assessing whether an individual has energy awareness and categorizing them into two groups—one with energy awareness and one without—the results showed that being energy aware delivered a higher accuracy of energy perception. Only 69 of the 142 individuals are valid when both components of energy awareness are considered. The results suggest that the percentage of people who are energy unaware accurately perceive energy usage is 10%, whereas the percentage of people who are energy aware is 30.8%. This represents a threefold increase in perceptual accuracy. Consider when you are an adult and you have just received your first electricity bill for a specific month, and then subsequently for other months, you will notice some months may have higher or lower consumption based on their level of usage, which gradually forms a psychological effect and memory on the level of usage. As a result, you gradually develop a level of energy awareness, which has the potential to reduce energy use while also allowing for more accurate energy perception about their consumption.

4. Conclusions

In this study, we proposed to develop an effective data analytical methodology for analyzing residential customers' energy consumption behaviours. Considering that mixed data is realistic in many real-world applications, the K-Prototype unsupervised cluster algorithm was adopted to estimate and cluster participants' energy behaviours. This requires determining cluster configuration using the Akaike Information Criterion and the Bayesian Information Criterion as cluster criteria. To assess the validity of the best cluster configuration, the silhouette coefficient and sequence matcher were employed to determine the highest accuracy score for individual performance, equating to the best score for mixed data type optimal cluster. The scores were determined to be accurate and correspond to the result

determined from using both AIC and BIC as recommended choices to select using analytical and statistical hypothesis techniques improved the performance of a single generic predictor, which is crucial for forecasting energy consumption. A trained ANN can forecast energy consumption with a high degree of accuracy with directly associated variables. Changes in behaviour would result in changes in energy consumption. Furthermore, variables can have a causal relationship even if they do not have a statistically significant difference.

Results demonstrate the interrelation between energy perception, thriftiness, income level, and awareness. Individuals are more likely to consciously reduce energy consumption when they recognize the importance of minimizing it, with personal income (PCI) playing a significant role Understanding a user's flexibility with an appliance provides insight into their usage patterns, frequency, perspective, and behaviour on the appliance.

This study has successfully developed an effective data analytical methodology for analyzing residential customers' energy consumption behaviors using the K-Prototype unsupervised clustering algorithm. However, several limitations must be acknowledged. Firstly, reliance on self-reported survey data may introduce biases and inaccuracies. Secondly, while the sample size is adequate for this preliminary analysis, it may not fully represent the diverse demographics of Singapore's residential sector. Thirdly, the use of mixed data types in clustering can challenge the uniformity and consistency of results across different variables.

Despite these limitations, the study provides valuable insights into the influence of energy consumption behaviour, serving as an initial step towards effective energy management and reduced greenhouse gas emissions. For instance, recognizing that users are more likely to adjust their air conditioning usage in response to energy awareness campaigns can guide the development of policies that target such appliances for energy-saving interventions.

Moreover, the study highlights the importance of incorporating behavioral insights into energy policy. Policymakers can leverage these findings to promote behavioral changes through public awareness campaigns, incentives for energy-efficient appliances, and feedback mechanisms that make energy consumption more transparent to consumers. This approach can support national goals for energy efficiency and greenhouse gas emission reduction.

Additionally, the findings can inform the creation of demand-side management strategies tailored to the specific needs and behaviors of different demographic groups. By understanding the causal relationships between variables such as income level, thriftiness, and energy awareness, policymakers can develop nuanced policies that address the unique challenges faced by various population segments.

Appendix A.

Appendix A.1. Variables List

Table A1. List of key variables and definitions.

Variable	Description
actual_energy	Actual energy consumption indicated by participants. i.e., 300–450 kWh
aircon_frequency	How frequently is the aircon being used by the participants?
b_aircon_temp	Would participants switch on the air conditioner if they are feeling warm?

Table A1. *Cont.*

Variable	Description
b_awareness	Participants are aware of the level of energy consumption used by them, whether is low, moderate or high usage.
b_flexi_aircon	Does the participant consider aircon appliances as flexible?
b_flexi_lights	Does the participant consider lighting appliances as flexible?
b_flexi_microwave	Does the participant consider microwave appliances as flexible?
b_flexi_oven	Does the participant consider oven appliances as flexible?
b_flexi_wm	Does the participant consider washing machine appliances as flexible?
b_fridge_load	If the participant's home fridge is overloaded (90–100% packed)
b_lights_pres	Would a participant switch off the lights when a room is unoccupied?
b_lights_task	Would a participant use task lighting instead of ceiling lights i.e., using a desk lamp light instead of ceiling light when studying?
b_more_energy_now	If the participant's family has used more electricity during the pandemic as compared to before the pandemic?
b_perceive	Users' perception of energy usage, whether is low, moderate or high usage.
b_thrifty	Is the participant's family mindful when using electricity and not splurging?
base_energy_app	Base energy value calculated for remaining base loads.
bills	Participants have indicated their bills correctly to match energy consumption. In other words, they have either been already aware of their electricity bills or they have made the effort to check their electricity bills and indicated accurately in the survey which previously, they were not aware of.
fam_total	Total family members living in the same household.
house_type	Type of housing. I.e., 4-Room Flat
pci	Participant's family income per capita.
personality	Does the participant consider their family mainly to be more introverted or extroverted?
region	Region of stay. I.e., East area, north area, central area.
shower_frequency	The average showering time for participants.
total_energy	Total energy consumption of participant's household.
tv_frequency	What is the participant's family's average hours of watching the television daily?
Type I Error	Rejecting the null hypothesis (Ho) when in fact it is true.
Type II Error	Failing to reject the null hypothesis (Ho) when in fact it is false.

Appendix A.2. Supplementary Information for Table 5 Significance of Post-Hoc Test

Table A2. Significance of Post-hoc Test (remaining results).

| | | | Bills | |
			Did Check	Did Not Check
b_aircon_temp	Yes	Obs	21	59
		Exp	27.7	52.3
		Adj. Res	−2.5	2.5
	No	Obs	25	28
		Exp	18.3	34.7
		Adj. Res	2.5	−2.5
b_flexi_aircon	Non-Flexible	Obs	16	47
		Exp	24.2	38.8
		Adj. Res	−2.9	2.9
	Flexible	Obs	35	35
		Exp	26.8	43.2
		Adj. Res	2.9	−2.9
			b_flexi_oven	
			Flexible	Non-Flexible
b_flexi_microwave	Flexible	Obs	85	11
		Exp	76	20
		Adj. Res	4.1	−4.1
	Non-Flexible	Obs	25	18
		Exp	34	9
		Adj. Res	−4.1	4.1

References

1. Yu, Z.; Fung, B.C.; Haghighat, F.; Yoshino, H.; Morofsky, E. A systematic procedure to study the influence of occupant behaviour on building energy consumption. *Energy Build.* **2011**, *43*, 1409–1417. [CrossRef]
2. Schweiker, M.; Hawighorst, M.; Wagner, A. The influence of personality traits on occupant behavioural patterns. *Energy Build.* **2016**, *131*, 63–75. [CrossRef]
3. Abrahamse, W.; Steg, L.; Vlek, C.; Rothengatter, T. The effect of tailored information, goal setting, and tailored feedback on household energy use, energy-related behaviours, and behavioural antecedents. *J. Environ. Psychol.* **2007**, *27*, 265–276. [CrossRef]

4.   Poortinga, W.; Steg, L.; Vlek, C.; Wiersma, G. Household preferences for energy-saving measures: A conjoint analysis. *J. Econ. Psychol.* **2003**, *24*, 49–64. [CrossRef]
5.   Gill, Z.M.; Tierney, M.J.; Pegg, I.M.; Allan, N. Low-energy dwellings: The contribution of behaviours to actual performance. *Build. Res. Inf.* **2010**, *38*, 491–508. [CrossRef]
6.   Shen, M.; Lu, Y.; Tan, K.Y. Big Five Personality Traits, Demographics and Energy Conservation Behaviour: A Preliminary Study of Their Associations in Singapore. *Energy Procedia* **2019**, *158*, 3458–3463. [CrossRef]
7.   Jain, R.K.; Gulbinas, R.; Taylor, J.E.; Culligan, P.J. Can social influence drive energy savings? Detecting the impact of social influence on the energy consumption behaviour of networked users exposed to normative eco-feedback. *Energy Build.* **2013**, *66*, 119–127. [CrossRef]
8.   Wang, Z.; Zhang, B.; Yin, J.; Zhang, Y. Determinants and policy implications for household electricity-saving behaviour: Evidence from Beijing, China. *Energy Policy* **2011**, *39*, 3550–3557. [CrossRef]
9.   Yang, S.; Shipworth, M.; Huebner, G. His, hers or both's? The role of male and female's attitudes in explaining their home energy use behaviours. *Energy Build.* **2015**, *96*, 140–148. [CrossRef]
10. Revell, K. Estimating the environmental impact of home energy visits and extent of behaviour change. *Energy Policy* **2014**, *73*, 461–470. [CrossRef]
11. Valkila, N.; Saari, A. Attitude–behaviour gap in energy issues: Case study of three different Finnish residential areas. *Energy Sustain. Dev.* **2013**, *17*, 24–34. [CrossRef]
12. Loock, C.M.; Landwehr, J.; Staake, T.; Fleisch, E.; Pentland, A.S. The influence of reference frame and population density on the effectiveness of social normative feedback on electricity consumption. In Proceedings of the 33rd International Conference on Information Systems (ICIS), Orlando, FL, USA, 16–19 December 2012.
13. Kempton, W.; Harris, C.; Keith, J.; Weihl, J. Do Consumers Know "What Works" in Energy Conservation? In *Families and the Energy Transition*; John, B., David, A.S., Marvin, B.S., Eds.; Routledge: London, UK, 1985.
14. Attari, S.Z.; DeKay, M.L.; Davidson, C.I.; De Bruin, W.B. Public perceptions of energy consumption and savings. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 16054–16059. [CrossRef]
15. Dietz, T.; Gardner, G.T.; Gilligan, J.; Stern, P.C.; Vandenbergh, M.P. Household actions can provide a behavioural wedge to rapidly reduce US carbon emissions. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 18452–18456. [CrossRef]
16. Lesic, V.; De Bruin, W.B.; Davis, M.C.; Krishnamurti, T.; Azevedo, I.M. Consumers' perceptions of energy use and energy savings: A literature review. *Environ. Res. Lett.* **2018**, *13*, 033004. [CrossRef]
17. Zhou, K.; Yang, S. Understanding household energy consumption behaviour: The contribution of energy big data analytics. *Renew. Sustain. Energy Rev.* **2016**, *56*, 810–819. [CrossRef]
18. Energy Market Authority (EMA). *Singapore Energy Statistics 2019*; EMA: Singapore, 2019.
19. Farrokhi, F.; Mahmoudi-Hamidabad, A. Rethinking convenience sampling: Defining quality criteria. *Theory Pract. Lang. Stud.* **2012**, *2*, 784–792. [CrossRef]
20. Chuan, L.; Ukil, A. Modeling and validation of electrical load profiling in residential buildings in Singapore. *IEEE Trans. Power Syst.* **2014**, *30*, 2800–2809. [CrossRef]
21. Energy Market Authority (EMA). *Average Monthly Household Electricity Consumption by Dwelling Type*; EMA: Singapore, 2020.
22. Kim, B. A fast K-prototypes algorithm using partial distance computation. *Symmetry* **2017**, *9*, 58. [CrossRef]
23. Madhuri, R.; Murty, M.R.; Murthy, J.V.R.; Reddy, P.P.; Satapathy, S.C. Cluster analysis on different data sets using K-modes and K-prototype algorithms. In *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II*; Springer: Cham, Switzerland, 2014; pp. 137–144.
24. Jia, Z.; Song, L. Weighted k-prototypes clustering algorithm based on the hybrid dissimilarity coefficient. *Math. Probl. Eng.* **2020**, *2020*, 5143797. [CrossRef]
25. Zhou, H.B.; Gao, J.T. Automatic method for determining cluster number based on silhouette coefficient. In *Advanced Materials Research*; Trans Tech Publications Ltd.: Stafa-Zurich, Switzerland, 2014; Volume 951, pp. 227–230.
26. Bengio, S.; Bengio, Y. Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Trans. Neural Netw.* **2000**, *11*, 550–557. [CrossRef]
27. McHugh, M.L. The chi-square test of independence. *Biochem. Medica* **2013**, *23*, 143–149. [CrossRef] [PubMed]
28. Kim, T.K. Understanding one-way ANOVA using conceptual figures. *Korean J. Anesthesiol.* **2017**, *70*, 22. [CrossRef] [PubMed]
29. Agresti, A. *An Introduction to Categorical Data Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
30. Dziak, J.J.; Coffman, D.L.; Lanza, S.T.; Li, R.; Jermiin, L.S. Sensitivity and specificity of information criteria. *Brief. Bioinform.* **2020**, *21*, 553–565. [CrossRef]
31. Kaminski, J. Diffusion of innovation theory. *Can. J. Nurs. Inform.* **2011**, *6*, 1–6.